دوره جامع تکنولوژی داده (DataTech)

مبانی علم داده

۱ مبانی علم داده
۲ برنامه‌نویسی در علم داده با پایتون
۳ یادگیری ماشین و یادگیری عمیق
٤ حکمرانی داده
۵ مدیریت پایگاه داده
٦ متن‌کاوی و وب‌کاوی
۷ کلان داده
۸ هوش تجاری
۹ مباحث ویژه در علم داده

فناپ FANAP

وزارت علوم، تحقیقات و فناوری
دانشگاه خاتم
Khatam University

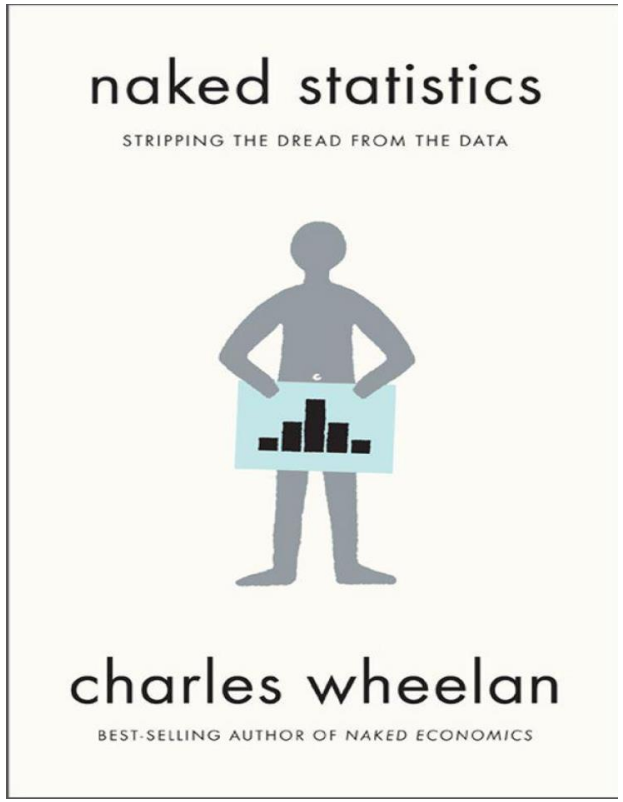
معرفی مدرس

احسان اخترکاوان

- دکتری تخصصی مهندسی کامپیوتر
- عضو هیات علمی دانشگاه و استادیار پایه 7

پل های ارتباطی: e.akhtarkavan@khatam.ac.ir

مراجع درس



Naked statistics

Stripping the Dread from the Data

CHARLES WHEELAN

سرفصل مطالب

- مقدمه - آمار توصيفي
- همبستگي
- مباني احتمال-مشكلات احتمال
- اهميت داده
- تئوري حد مركزي-استنتاج آماري
- تحليل رگرسيون و مشكلات آن

What is Data Science?

علم داده بين رشته اى است و هر كسي از ظن خويش يار آن مي شود.

- از ديد مهندسي كامپيوتر:

= افرادي كه تكنولوژي هايي توسعه مي دهند كه مي تواند از داده ياد بگيرد

+ كساني داده دارند

+ كساني كه مشكل يا مسئله دارند

- از ديد متخصصي آمار و كامپيوتر:

= ساختن ابزارهايي براي حل مسئله ها به كمك داده

= ساختن ابزارهايي براي آشكار كردن الگوها، پيش بيني رخدادها، و فهم دنيا از طريق اكتشاف داده ها

نيازمند هدايت علم كامپيوتر و بهينه سازي و آمار به مسيرهاي جديد

- از ديد آمارزيستي:

= يك خط توليد است كه به ما اجازه مي دهد اطلاعات را از داده استخراج كنيم، شامل فعاليتهايي مانند جمع آوري داده و كدينگ داده، تجميع و پردازش داده، تحليل داده و نهايتا متصورسازي داده

What is Data Science?

- از ديد خبرنگاري:

= يك رشته بسيار وسيع با ظرفيت است كه هم مي تواند به دانشكده هاي مختلف آگاهي بخشي كند و هم از آنها آگاهي كسب كند

آگاهي كه با ارزش ها و اخلاق رشته هاي مختلف منطبق مي شود.

- از ديد رياضيات كاربردي:

= با تفكر آماري مي توانيم آنچه در گذشته متوجه نمي شديم را متوجه بشويم. چيزهايي كه بخشي از زندگي ما بوده اند ولي نمي دانستيم كه حوزه هاي پر داده اي هستند. مانند ورزش. با اين ديد جديد مي توانيم درك خود را در آن حوزه با كمك ديتاست تغيير دهيم

علم داده بر خلاف آنچه تصور مي شود تازه متولد شده است و افراد در حال ايجاد تعادل ميان ابزارهاي مدرن، مشاهداتي كه دهه ها وجود داشته اند، و رياضيات كه خيلي وقت است در دسترس بوده اند براي توليد علم داده هستند.

What is Data Science?

- از ديد علوم كامپيوتر: علم داده تقاطع سه علم:

1. كامپيوتر براي پيدا كردن الگوريتم هايي كه به مجموعه داده هاي خيلي بزرگ اعمال بشوند، فرضيه هاي زيادي را بررسي كنند و پاسخ ها را پيدا كنند
2. آمار يا تفكر استنتاجي براي حل مسائل معكوس كه بايد حل كنيم تا بتوانيم پديده هايي را كه در واقع آن داده ها توليد كرده اند بشناسيم
3. تخصص حوزه اي كه در حال بررسي آن هستيم تا بتوانيم سوالات درستي بپرسيم و چيزهايي را كه براي پيشرفت آن حوزه مهم است را پيدا كنيم

- از ديد مهندسي MBA

1. ما در حال انجام يك تغييرالگوي رفتاري هستيم (paradigm shift) به كمك علم داده هستيم
2. تاكنون اينگونه بوده كه به كامپيوتر مي گفتيم چه بكند و كامپيوتر انجام مي داد ولي الان وارد دوره اي مي شويم كه قرار است كامپيوتر ها ياد بگيرند، پيش بيني كنند و همكاري كنند.

What is Data Science?

- از ديد مهندسي بيوانفورماتيك

1. علم داده تقاطع قدرتمند متدهايي از علم كامپيوتر و متدهايي از آمار است كه به يك حوزه خاص اعمال مي گردد.
2. بخصوص متدهاي آماري كه توسط كامپيوترها مقياس بزرگي پيدا مي كنند و به حوزه خاص اعمال مي گردد

- از ديد مهندسي صنايع

1. علم داده يك حوزه بين رشته اي است كه از آمار، علم كامپيوتر و تكنولوژي هاي مديريت ديتست هاي بزرگ براي حل مشكلات يك حوزه خاص استفاده مي كند

What is Data Science?

- از ديد مهندسي عمران

1. علم داده تقاطع آمار و احتمال و علم كامپيوتر با يك كاربرد است
2. علم داده جريان داده را مي گيرد و با آمار و احتمال و كامپيوتر آن را بررسي مي كند و به يك كاربرد خاص اعمال مي كند.

چه سوالاتي را پاسخ مي دهد؟

- What questions can Data Science answer?
- Why is there an explosion of data?
- What role does data visualization play in Data Science?
- How did you become interested in Data Science?
- What do you predict will happen in Data Science in 5 years?
- What skills does a Data Scientist need to be successful?
- What should a non-Data Scientist know about Data Science?

Statistics

- Statistics can be used to explain everything from DNA testing to the idiocy of playing the lottery.
- Statistics can help us identify the factors associated with diseases like cancer and heart disease;
- It can help us spot cheating on standardized tests.
- Statistics can even help you win on game shows.

- The paradox of statistics is that they are everywhere—presidential polls—but the discipline itself has a reputation for being uninteresting and inaccessible.

Statistics

- Many statistics classes are overly laden with math and jargon.
- Believe me, the technical details are crucial—but it's just Greek if you don't understand the intuition.
- Meanwhile, the course is long on examples to convince you that there are great reasons to learn this stuff.
- *Statistics can be really interesting, and most of it isn't that difficult.*
- The key is to separate the important ideas from the mysterious technical details that can get in the way.
- No. Statistics rarely offers a single "right" way of doing anything.
- It provides meaningful information in an easily accessible way.

Example : Sum of infinite series

- An infinite series is a pattern of numbers that goes on forever, such as 1 + ½ + ¼ + ⅛ . . . continues to infinity.
- Imagine that you have positioned yourself exactly 2 m from a wall.
- Now move half the distance to that wall (1 m), so that you are left standing 1m away.
- From 1m away, move half the distance to the wall once again (½ m).
- And from ½ m away, do it again (move ¼ m). Then do it again (move ⅛ m). And so on.
- But you will never hit the wall, because by definition each move takes you only half the remaining distance.

Warning !!!!!!!

- Statistics is like a high-caliber weapon: helpful when used correctly and potentially disastrous in the wrong hands.
- Statistics can be *overly accessible* in the sense that anyone with data and a computer can do sophisticated statistical procedures with a few keystrokes.
- The problem is that if the data are poor, or if the statistical techniques are used improperly, the conclusions can be wildly misleading and even potentially dangerous.
- Example: ***People Who Take Short Breaks at Work Are Far More Likely to Die of Cancer.***

Data and Statistics

- “Data is merely the raw material of knowledge.”
- Statistics is the most powerful tool we have for using information to some meaningful end, whether that is identifying underrated baseball players or paying teachers more fairly.

- Swedish mathematician and writer Andrejs Dunkels: It’s easy to lie with statistics, but it’s hard to tell the truth without them

Disparate questions:

- How can we catch schools that are cheating on their standardized tests?
- How does Netflix know what kind of movies you like?
- How can we figure out what substances or behaviors cause cancer, given that we cannot conduct cancer-causing experiments on humans?
- Does praying for surgical patients improve their outcomes?
- Is there really an economic benefit to getting a degree from a highly selective college or university?
- What is causing the rising incidence of autism?

Statistics and DATA

- **Here is a quick tour of how statistics can bring meaning to raw data.**

1. Description and Comparison
2. Inference
3. Identifying Important Relationships (Statistical Detective Work)
4. Assessing Risk and Other Probability-Related Events
5. *Regression Analysis*

1 . Description and Comparison

- It’s a nice tool for making a quick comparison between things.
- We use numbers everywhere in life, to summarize information.
- The GPA is a handy tool for assessing their academic potential.
- Someone who has a 3.7 GPA is clearly a stronger student than someone at the same school with a 2.5 GPA.
- That makes it a nice descriptive statistic.
- It’s easy to calculate, it’s easy to understand, and it’s easy to compare across students.

But it's not perfect

- The GPA does not reflect the difficulty of the courses that different students may have taken.
- How can we compare a student with a 3.4 GPA in classes that appear to be relatively non-challenging and a student with a 2.9 GPA who has taken calculus, physics, and other tough subjects?

- GPA "fix." A honors class = 5 , A non-honors = 4
- For a student taking a lot of honors classes, any A in a non-honors course, such as gym or health education, would actually pull the GPA down, even though it is impossible to do better than an A in those classes.

Was that insane (Silly)?

- An overreliance on any descriptive statistic can lead to misleading conclusions, or cause undesirable behavior.
- Descriptive statistics exist to simplify, which always implies some loss of nuance or detail.
- Does it simplify things? Yes, that is both the strength and the weakness of any descriptive statistic.

Example: Gini index

- Gini index is a standard tool in economics for measuring income inequality.
- The Gini index measures how evenly wealth (or income) is shared within a country on a scale from zero to one.
- The Gini index has no intrinsic meaning; it's a tool for comparison.

Last reports	2007	1997
U.S.'s: 0.37, Sweden: 0.29. Canada: 0.34. China: 0.38. Brazil: 0.53 South Africa: 0.63	U.S.: 0.45, Sweden: 0.23. Canada: 0.32. China: 0.42. Brazil: 0.54. South Africa: 0.65	U.S.: 0.41, Sweden: 0.25. Canada: 0.32. China: 0.42. Brazil: - South Africa:-

2. Inference

- One key function of statistics is to use the data we have to make informed conjectures about larger questions for which we do not have full information.
- In short, we can use data from the **“known world”** to make informed inferences about the “unknown world.”

Counting homeless people

- It is **expensive** and **logistically difficult** to count the homeless population in a large metropolitan area.
- One important statistical practice is sampling, which is the process of gathering data for a small area, say, a handful of census regions, and then using those data to make an informed judgment, or inference, about the homeless population for the city as a whole.
- Sampling requires far less resources than trying to count an entire population;
- done properly, it can be every bit as accurate.

A political poll is one form of sampling.

- A research organization will attempt to contact a sample of households that are broadly representative of the larger population and ask them their views about a particular issue or candidate.
- This is obviously much cheaper and faster than trying to contact every household in an entire state or country.
- The Gallup in US and ISPA in Iran, calculate a methodologically sound poll of more than 1,000 households will produce roughly the same results as a poll that attempted to contact every household in America.

3. Assessing Risk and Other Probability-Related Events

- Many businesses must assess the risks associated with assorted adverse outcomes.
- Casinos make money in the long run—always.
- That does not mean that they are making money at any given moment.
- The whole gambling industry is built on games of chance, meaning that the outcome of any particular roll of the dice or turn of the card is uncertain.

Assessing Risk

- They cannot make those risks go away entirely, just as a casino cannot guarantee that you won't win every hand of blackjack that you play.
- However, any business facing uncertainty can manage these risks by engineering processes so that the probability of an adverse outcome, anything from an environmental catastrophe to a defective product, becomes acceptably low.

Wall Street mistakes

- Wall Street firms will often evaluate the risks posed to their portfolios under different scenarios, with each of those scenarios weighted based on its probability.
- The financial crisis of 2008 was precipitated in part by a series of market events that had been deemed extremely unlikely, as if every player in a casino drew blackjack all night.
- These Wall Street models were flawed and that the data they used to assess the underlying risks were too limited, but the point here is that any model to deal with risk must have probability as its foundation.

Insurance industry

- When individuals and firms cannot make unacceptable risks go away, they seek protection in other ways.
- The entire insurance industry is built upon charging customers to protect them against some adverse outcome, such as a car crash or a house fire.
- The insurance industry does not make money by eliminating these events; cars crash and houses burn every day.
- Sometimes cars even crash into houses, causing them to burn.

Insurance industry

- Instead, the insurance industry makes money by **charging premiums that are more than sufficient to pay for the expected payouts from car crashes and house fires.**
- (The insurance company may also try to lower its expected payouts by encouraging safe driving, fences around swimming pools, installation of smoke detectors in every bedroom, and so on.)

catch cheats

- For example, the company will flag exams at a school or test site on which the number of ***identical wrong answers is highly unlikely***, usually a pattern that would happen by chance less than one time in a million.
- The mathematical logic stems from the fact that we cannot learn much when a large group of students all answer a question correctly.
- That's what they are supposed to do; they could be cheating, or they could be smart.
- But when those same test takers get an answer wrong, they should not all consistently have ***the same wrong answer***.
- If they do, it suggests that they are copying from one another.

What are the Odds ?

- The company also looks for exams in which a test taker does significantly better on hard questions than on easy questions **(suggesting that he/she had answers in advance)** and

- for exams on which the number of "**wrong to right**" erasures is significantly higher than the number of "right to wrong" erasures **(suggesting that a teacher or administrator changed the answer sheets after the test)**.

Limitations of probability

- A large group of test takers might have the same wrong answers by coincidence; in fact, the more schools we evaluate, the more likely it is that we will observe such patterns just as a matter of chance.
- A statistical anomaly does not prove wrongdoing.
- **Probability is one weapon in an arsenal that requires good judgment.**

The limitations of using probability

- **Delma Kinney**, a 54-old Atlanta man, won $1 million in an instant lottery game in 2008 and then another $1 million in an instant game in 2011.
- The probability of that happening to the same person is somewhere in the range of 1 in 25 trillion.
- We cannot arrest Mr. Kinney for fraud on the basis of that calculation alone.

- (though we might inquire whether he has any relatives who work for the state lottery).

4. Identifying Important Relationships (Statistical Detective Work)

- Does smoking cigarettes cause cancer?

- The scientific method dictates that if we are testing a scientific hypothesis, we should conduct a controlled experiment in which the variable of interest (e.g., smoking) is the only thing that differs between the experimental group and the control group.

- If we observe a marked difference in some outcome between the two groups (e.g., lung cancer), we can safely infer that the variable of interest is what caused that outcome.

We can’t do experiment on humans.

- If our working hypothesis is that smoking causes cancer, it would be unethical to assign recent college graduates to two groups, smokers and nonsmokers, and then see who has cancer at the twentieth reunion.
- (We can conduct controlled experiments on humans when our hypothesis is that a new drug or treatment may improve their health;
- we cannot knowingly expose human subjects when we expect an adverse outcome.)*

We cannot do experiment on humans.

- No. Smokers and nonsmokers are likely to be different in ways other than their smoking behavior.
- For example, smokers may be more likely to have other habits, such as drinking heavily or eating badly, that cause adverse health outcomes.
- We would also have a serious problem with the data on which we are basing our analysis.
- Smokers who have become seriously ill with cancer are less likely to attend the twentieth reunion.
- The further the class gets from graduation, say, a fortieth or a fiftieth reunion, the more serious this bias will be.

DNA

- As a result, statistics is a lot like good detective work.
- DNA from a cigarette butt, teeth marks on an apple, a single fiber from a car floor mat—and then ***use the evidence to catch a violent criminal***.
- The experts do not have the conventional evidence used to find the bad guy, such as an eyewitness or a surveillance videotape.
- So they turn to scientific inference instead.

5. Regression Analysis

- Regression analysis is the tool that enables researchers **to isolate a relationship between two variables**, such as smoking and cancer, while **holding constant (or "controlling for")** the effects of other important variables, such as diet, exercise, weight, and so on.

- ***Eating a bran muffin reduces chances of getting colon cancer?***

Bran muffin & colon cancer

- Researchers will gather detailed information on thousands of people, including how frequently they eat bran muffins, and then use regression analysis to do two crucial things:

(1) quantify the association observed between eating bran muffins and contracting colon cancer; and

(2) quantify the likelihood that the association between bran muffins and a lower rate of colon cancer observed in this study is merely a coincidence—rather than a meaningful insight about the relationship between diet and health.

Statistically significant?

- It means that the analysis has uncovered an association between two variables that is not likely to be the product of chance alone.

Why? What is the point?

- Well, that exposes one of the **limitations of regression analysis**.
- We can isolate a strong association between two variables by using statistical analysis,
- but we cannot necessarily explain why that relationship exists, and in some cases, we cannot know for certain that the relationship is causal, **meaning that a change in one variable is really causing a change in the other**.
- The point is not to do math, or to dazzle friends and colleagues with advanced statistical techniques.
- The point is to learn things that inform our lives.

Lies, Damned Lies, & Statistics

- Even in the best of circumstances, statistical analysis **rarely unveils "the truth."**
- We are usually building a inferred case based on **imperfect** data.
- As a result, there are numerous reasons that intellectually honest individuals may disagree about statistical results or their implications.
- At the most basic level, we may disagree on the question that is being answered.
- Sports enthusiasts will be arguing for all eternity over ***"the best baseball player ever"*** because there is no objective definition of ***"best."***

More socially significant questions

- Fancy descriptive statistics can inform this question, but they will never answer it definitively.
- What is happening to the economic health of the **middle class**?
- That answer depends on how one defines both "**middle class**" and "**economic health**."
- Even when we can conduct large, controlled experiments on human beings, they are neither **easy** nor **cheap**.
- Researchers did a large-scale study on whether or not prayer reduces postsurgical complications. ***That study cost $2.4 million***.

Neither **easy** nor **cheap**

- "You go to war with the army you have—not the army you might want or wish to have at a later time."
- We conduct statistical analysis using the best data and methodologies and resources available.
- The approach is not like addition or long division, in which the correct technique yields the "**right**" answer and a computer is always more precise and less fallible than a human.
- **Statistical analysis is more like good detective work.**

Controversial results

- Smart and honest people will often disagree about what the data are trying to tell us.
- But who says that everyone using statistics is smart or honest?
- The book "***How to Lie with Statistics***, which was first published in 1954 and has sold over a million copies.

what is the point of learning statistics?

- The reality is that you *can* lie with statistics.
- Or you can make inadvertent (unplanned) errors.
- In either case, the mathematical precision attached to statistical analysis can dress up some serious nonsense.
- BUT:
  1. To summarize huge quantities of data.
  2. To make better decisions.
  3. To answer important social questions.

what is the point of learning statistics?

4. To recognize patterns that can refine how we do everything from selling diapers to catching criminals.
5. To catch cheaters and prosecute criminals.
6. To evaluate the effectiveness of policies, programs, drugs, medical procedures, and other innovations.
7. And to spot the scoundrels who use these powerful tools for nefarious ends.

Central tendency

- The first descriptive task is often to find some measure of the “middle” of a set of data, or what statisticians might describe as its “central tendency.”
- The most basic measure of the “**middle**” of a distribution is the mean, or average.
- The mean, or average, turns out to have some problems in that regard, namely, that it is prone to distortion by “outliers,” which are observations that lie farther from the center.

Descriptive Statistics

- Let us ponder for a moment two seemingly unrelated questions:

(1) What is happening to the economic health of the middle class?

(2) Who was the greatest baseball player of all time?

- The first question is profoundly important. (presidential campaigns)
- The second question is trivial (in the literal sense of the word), but baseball enthusiasts can argue about it endlessly.

Common points?

- What the two questions have in common is that they can be used to illustrate the strengths and limitations of descriptive statistics, **which are the numbers and calculations we use to summarize raw data.**
- If I want to demonstrate that Derek Jeter is a great baseball player, I can sit you down and describe every at bat in every Major League game that he's played.
- That would be raw data, and it would take a while to digest, given that Jeter has played seventeen seasons with the New York Yankees and taken **9,868** at bats.

On the other hand

- Or I can just tell you that at the end of the 2011 season Derek Jeter had a career batting average of **0.313**.
- That is a descriptive statistic, or a "summary statistic."
- The batting average is a gross simplification of Jeter's seventeen seasons.
- It is easy to understand, elegant in its simplicity—and limited in what it can tell us.

Economic descriptive statistics

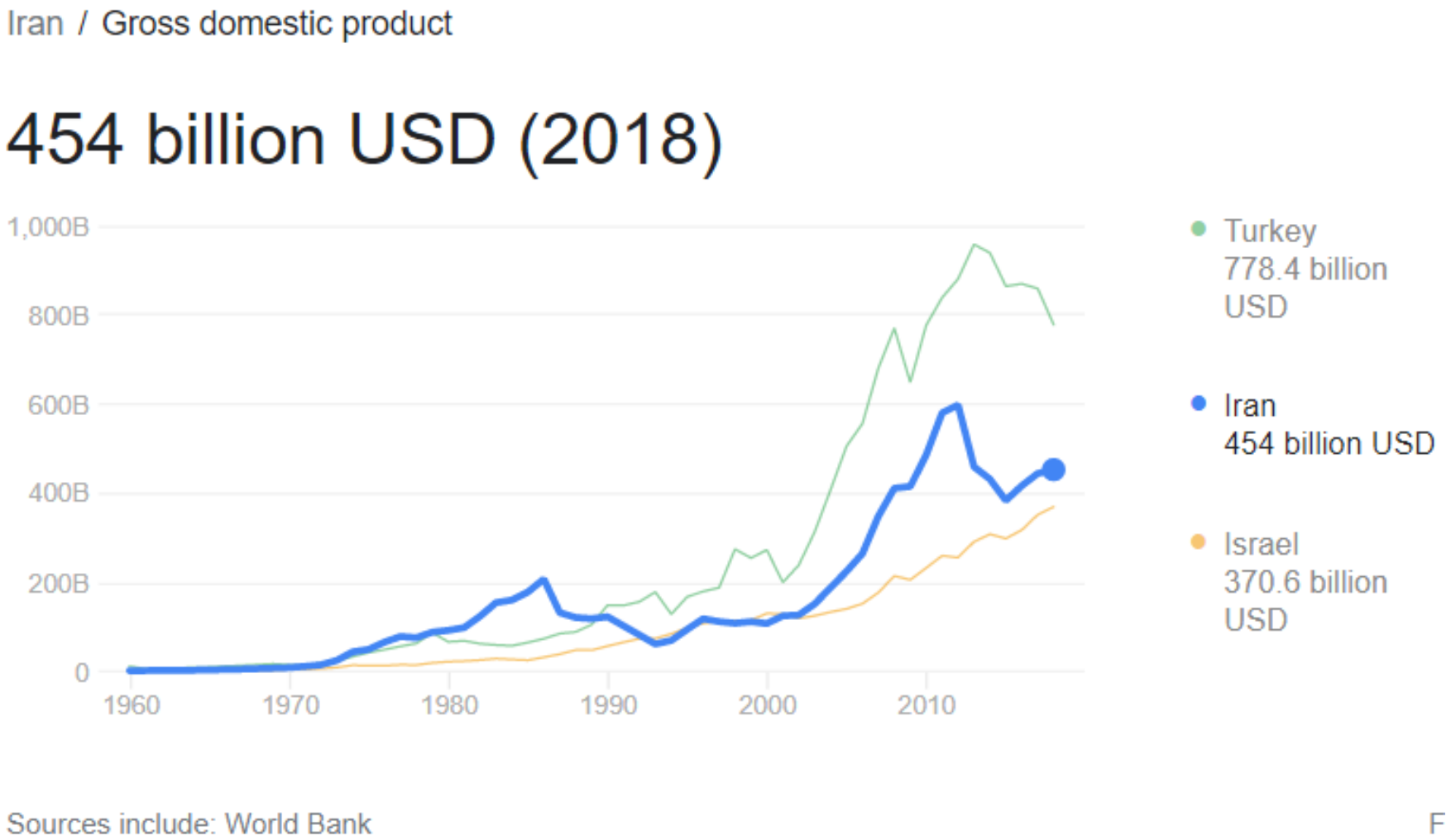
- Are the people we define as middle class getting richer, poorer, or just running in place?
- A reasonable answer—though by no means the “right” answer—would be to calculate the change in per capita income over the course of a generation, which is roughly thirty years.
- What descriptive statistics they would use to assess the economic well-being of a typical citizen.

Per capita income

- Per capita income is a simple average: total income divided by the size of the population.
- By that measure, average income in the US climbed from **$7,787** in **1980** to **$26,487** in **2010** and **($53,820** in **2018**).
- GDP in Iran is expected to reach 410.00 USD Billion by the end of 2020, according to Trading Economics global macro models and analysts expectations.
- In the long-term, the Iran GDP is projected to trend around 415.00 USD Billion in 2021 and 420.00 USD Billion in 2022, according to our econometric models.

Per capita income

Per capita income



Warning !!!!!!! problem.

- My quick calculation is technically correct and yet totally wrong in terms of the question I set out to answer.
- To begin with, the figures above are **not adjusted for inflation**.
- (A per capita income of $7,787 in 1980 is equal to about $19,600 when converted to 2010 dollars.)
- **That's a relatively quick fix.**
- The bigger problem is that the average income in America is not equal to the income of the average American.
- Let's unpack that clever little phrase.

Summarize information

- We perform calculations that reduce a complex array of data into a handful of numbers that describe those data, just as we might encapsulate a complex, complicated Olympic gymnastics performance with one number: 9.8.
- The good news is that these descriptive statistics give us a **manageable and meaningful summary of the underlying phenomenon.**
- The bad news is that any simplification invites abuse. Descriptive statistics can be like online dating profiles: **technically accurate and yet pretty darn misleading.**

Or maybe not.

- To get your mind around this concept, imagine that ten guys are sitting; each of these guys earns $35,000 a year, which makes the mean annual income for the group $35,000.
- Bill Gates walks in, with an annual income of $1 billion.
- The new mean : $91 million
- The sensitivity of the mean to outliers is why we should not gauge the economic health of the middle class by looking at per capita income.
- Because there has been explosive growth in incomes at the top end of the distribution—CEOs, hedge fund managers, and athletes
- The average income in the U.S. could be heavily skewed by the mega-rich.

The median

- The median is the point that divides a distribution in half, meaning that half of the observations lie above the median and half lie below.
- If we return to the bar example, the median annual income for the ten guys originally sitting in the bar is $35,000.
- When Bill Gates walks in, the median annual income for the eleven of them is still $35,000.
- The income of the guy sitting on the sixth stool represents the median income for the group.
- If Warren Buffett comes in and sits down on the twelfth stool next to Bill Gates, the median still does not change

Mean or median??????????

- Neither the median nor the mean is hard to calculate; **the key is determining which measure of the "middle" is more accurate in a particular situation** (a phenomenon that is easily exploited).
- Meanwhile, the median has some useful relatives.

Quartiles and Deciles

- As we've already discussed, the median divides a distribution in half.
- The distribution can be further divided into quarters, or quartiles.
- The 1st quartile consists of the bottom 25% of the observations;
- The 2nd quartile consists of the next 25% of the observations; and so on.
- Or the distribution can be divided into deciles, each with 10% of the observations.

Percentiles

- We can go even further and divide the distribution into hundredths, or percentiles.
- Each percentile represents 1 % of the distribution, so that the 1st percentile represents the bottom 1 % of the distribution and the 99th percentile represents the top 1 % of the distribution.
- The benefit of these kinds of descriptive statistics is that they describe where a particular observation lies compared with everyone else.

Quartiles/Deciles/Percentiles

- If I tell you that your child scored in the 3rd percentile on a reading comprehension test, you should know immediately that the family should be logging more time at the library.
- You don't need to know anything about the test itself, or the number of questions that your child got correct.
- The percentile score provides a ranking of your child's score relative to that of all the other test takers.

Relative vs. Absolute figures

- An “absolute” score, number, or figure has some intrinsic meaning.
- Absolute figures can usually be interpreted without any context or additional information.
- When I tell you that I shot 83, you don’t need to know what other golfers shot that day in order to evaluate my performance.
- If I place ninth in the golf tournament, that is a relative statistic.
- A “relative” value or figure has meaning only in comparison to something else, or in some broader context, such as compared with the eight golfers who shot better than I did.
- Most standardized tests produce results that have meaning only as a relative statistic.

Relative vs. Absolute figures

- If I tell you that a third grader in an elementary school scored 43 out of 60 on the mathematics, that absolute score doesn't have much meaning.
- But when I convert it to a percentile—meaning that I put that raw score into a distribution with the math scores for all other third graders—then it acquires a great deal of meaning.
- If 43 correct answers falls into the 83rd percentile, then this student is doing better than most of his peers statewide.
- If he's in the 8th percentile, then he's really struggling.
- **In this case, the percentile (the relative score) is more meaningful than the number of correct answers (the absolute score).**

Standard deviation

- Another statistic that can help us describe what might otherwise be a jumble of numbers is the standard deviation, which is a measure of how dispersed the data are from their mean.
- In other words, how spread out are the observations?
- Suppose I collected data on the weights of 250 people on an airplane, and I also collected the weights of a sample of 250 qualifiers for Marathon.
- **Now assume that the mean weight for both groups is roughly the same, say 155 pounds**.
- 320-pound football players, the tiny screaming infant across the row and the six-year-old kid.

Standard deviation

- On the basis of the descriptive tools introduced so far, the weights of the airline passengers and the marathoners are nearly identical.
- *But they're not.*
- Yes, the weights of the two groups have roughly the same "middle," but the airline passengers have far more dispersion around that midpoint, meaning that their weights are spread farther from the midpoint.
- The marathon runners look like they all weigh the same amount, while the airline passengers have some tiny people and some large people.

Standard deviation

- The weights of the airline passengers are “more spread out,” which is an important attribute when it comes to describing the weights of these two groups.
- The standard deviation is the descriptive statistic that allows us to assign a single number to this dispersion around the mean.

Standard deviation example

- Suppose you do blood test, and a few days later her assistant leaves a message on your answering machine to inform you that your HCb2 count (a fictitious blood chemical) is 134.
- You rush to the Internet and discover that the mean HCb2 count for a person your age is 122 (and the median is about the same).
- Holy crap! If you're like me, you would finally draft a will.
- When you call the doctor's office back, the physician's assistant informs you that your count is within the normal range.
- But how could that be? ***"My count is 12 points higher than average!".***

What the heck does that mean?

- "The standard deviation for the HCb2 count is 18," the technician informs you curtly.
- There is natural variation in the HCb2 count, as there is with most biological phenomena (e.g., height).
- While the mean count might be 122, plenty of healthy people have counts that are higher or lower.
- The danger arises only when the HCb2 count **gets excessively high or low.**
- So how do we figure out what **"excessively"** means in this context?

Excessively?????

- As we've already noted, the standard deviation is a measure of dispersion, meaning that it reflects how tightly the observations cluster around the mean.
- For many typical distributions of data, a high proportion of the observations lie within one standard deviation of the mean
- (meaning that they are in the range from one standard deviation below the mean to one standard deviation above the mean).
- Yes, your count is 12 above the mean, but that's less than one standard deviation.
- Of course, far fewer observations lie two standard deviations from the mean, and fewer still lie three or four standard deviations away.

What are the Odds ?

- Some distributions are more dispersed than others.
- Hence, the standard deviation of the weights of the 250 airline passengers will be higher than the standard deviation of the weights of the 250 marathon runners.
- A frequency distribution with the weights of the airline passengers would literally be fatter (more spread out) than a frequency distribution of the weights of the marathon runners.
- **Once we know the mean and standard deviation for any collection of data, we have some serious intellectual power.**

What are the Odds ?

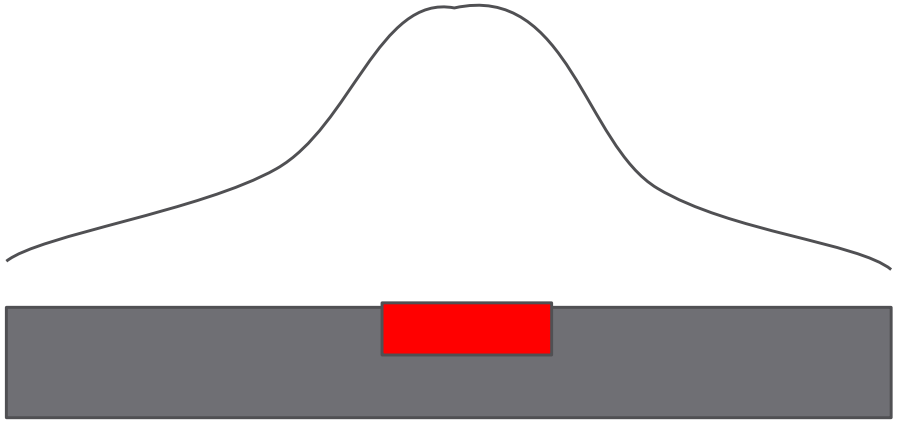
- For example, suppose I tell you that the mean score on the math test is 500 with a standard deviation of 100.
- The bulk of students taking the test will be within one standard deviation of the mean, or between 400 and 600.
- How many students do you think score 720 or higher?
- Probably not very many, since that is more than two standard deviations above the mean.

The normal distribution.

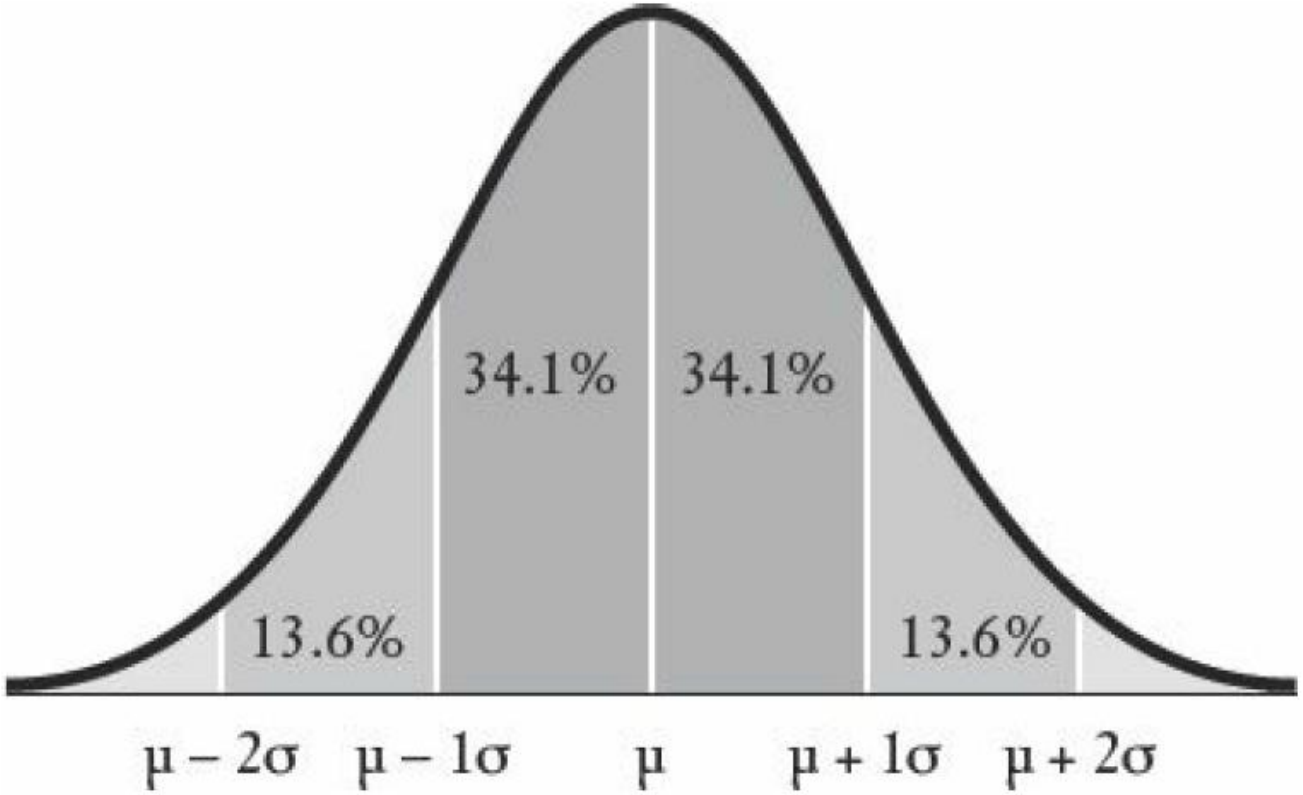
- Data that are distributed normally are symmetrical around their mean in a bell shape that will look familiar to you.
- The normal distribution describes many common phenomena.
- Imagine a frequency distribution describing popcorn popping on a stove top.
- Some kernels start to pop early, maybe one or two pops per second; after ten or fifteen seconds, the kernels are exploding madly.
- Then gradually the number of kernels popping per second fades away at roughly the same rate at which the popping began.

standard deviation

- Each test is specifically designed to produce a normal distribution of scores with mean 500 and standard deviation of 100.
- According to the *Wall Street Journal*, Americans even tend to park in a normal distribution at shopping malls; most cars park directly opposite the mall entrance—the "peak" of the normal curve—with "tails" of cars going off to the right and left of the entrance.



The normal distribution



- The mean is the middle line which is often represented by the Greek letter μ.
- The standard deviation is often represented by the Greek letter σ.
- Each band represents one standard deviation.

The normal distribution

- The beauty of the normal comes from the fact that we know by definition exactly what proportion of the observations in a normal distribution lie within
- one standard deviation of the mean (68.2 percent),
- within two standard deviations of the mean (95.4 percent),
- within three standard deviations (99.7 percent), and so on.
- This may sound like trivia. In fact, it is the foundation on which much of statistics is built.

To compare two figures or quantities

- I’m one inch taller than my brother; today’s temperature is nine degrees above the historical average for this date; and so on.
- Those comparisons make sense because most of us recognize the scale of the units involved.
- One inch does not amount to much when it comes to a person’s height, so you can infer that my brother and I are roughly the same height.
- Conversely, nine degrees is a significant temperature deviation in just about any climate at any time of year, so nine degrees above average makes for a day that is much hotter than usual.

To compare two figures or quantities

- But suppose that I told you that Granola Cereal A contains 31 milligrams more sodium than Granola Cereal B.
- Unless you know an awful lot about sodium (and the serving sizes for granola cereal), that statement is not going to be particularly informative.
- Or what if I told you that my cousin Al earned $53,000 less this year than last year? Should we be worried about Al?
- Or is he a hedge fund manager for whom $53,000 is a rounding error in his annual compensation?

we're missing context

- In both the sodium and the income examples, we're missing context.
- The easiest way to give meaning to these relative comparisons is by using percentages.
- It *would* mean something if I told you that Granola Bar A has 50 % more sodium than Granola Bar B, or that Uncle Al's income fell 47 % last year.
- Measuring change as a percentage gives us some sense of scale.

Percentages are potentially confusing or even deceptive.

- Assume that a department store is selling a dress for $100.
- The assistant manager marks down all merchandise by 25 percent.
- But then that assistant manager is fired and the new assistant manager raises all prices by 25 percent.
- What is the final price of the dress? actually $93.75.
- The assistant store manager cuts the price of the $100 dress by 25 % or $25, the price down to $75.
- The price be raised 25 % or $18.75, which is how the final price ends up at $93.75 (and not $100).

percentage change

- The point is that a **percentage change** always gives the value of some figure *relative to something else*.
- Therefore, we had better understand what that something else is.
- I once invested some money in a company.
- Since it was a private venture, there were no requirements as to what information had to be provided to shareholders.
- A number of years went by without any information on the fate of my investment; the funder was fairly tight-lipped on the subject.
- Finally I received a letter in the mail informing me that the firm's profits were **46 % higher** than the year before.

percentage change

- There was no information on the size of those profits in absolute terms, meaning that I still had absolutely no idea how my investment was performing.
- Suppose that last year the firm earned 27 cents—essentially nothing.
- This year the firm earned 39 cents—also essentially nothing.
- Yet the company's profits grew from 27 cents to 39 cents, which is technically a 46 % increase.

Why?

- Obviously the shareholder letter would have been more of a disappointment if it pointed out that the firm's cumulative profits over two years were less than the cost of a cup of coffee.
- To be fair to my roommate, he eventually sold the company for hundreds of millions of dollars, earning me a 100 % return on my investment.
- (Since you have no idea how much I invested, you also have no idea how much money I made—which reinforces my point here very nicely!)

Percentage change vs. a change in percentage points

- Percentage change must not be confused with a change in percentage points.
- Rates are often expressed in percentages.
- The sales tax rate in Illinois is 6.75 percent.
- I pay my agent 15 % of my book royalties.
- These rates are collected against some quantity, such as income in the case of the income tax rate.
- Obviously the rates can go up or down; less intuitively, the ***changes*** in the rates can be described in much dissimilar ways.

Percentage change vs. a change in percentage points

- The best example of this was a recent change in the Illinois personal income tax, which was raised from 3 % to 5 percent.
- There are two ways to express this tax change, both of which are technically accurate.
- One says, (correctly) that the state income tax *rate* was increased by *2 percentage points* (from 3 % to 5 percent).
- The other says (also correctly) that the state income tax had been raised by *67 percent*.

What is the point?

- The absolute change in the tax rate; (from 3 % to 5 percent). (D)
- The percentage change in the tax burden. (*67 percent) (GOP)*
- As noted, both descriptions are technically correct,

- though I would argue that the Republican description more accurately conveys the impact of the tax change, since what I'm going to have to pay to the government—the amount that I care about, as opposed to the way it is calculated—really has gone up by 67 percent.

Index

- Many phenomena defy perfect description with a single statistic.
- The advantage of any index is that it consolidates lots of complex information into a single number.
- We can then rank things that otherwise defy simple comparison—anything from quarterbacks to colleges to beauty pageant contestants.
- Alas, the disadvantage of any index is that it consolidates lots of complex information into a single number.
- There are countless ways to do that; each has the potential to produce a different outcome.

UN Human Development Index

- Any index is highly sensitive to the descriptive statistics that are combined together to build it, and to the weight given to each of those components.
- As a result, indices range from useful but imperfect tools to complete fakes.
- An example of the former is the United Nations Human Development Index, or HDI.
- The HDI was created as a measure of economic well-being that is broader than income alone.
- The HDI uses income as one of its components but also includes measures of life expectancy and educational attainment.

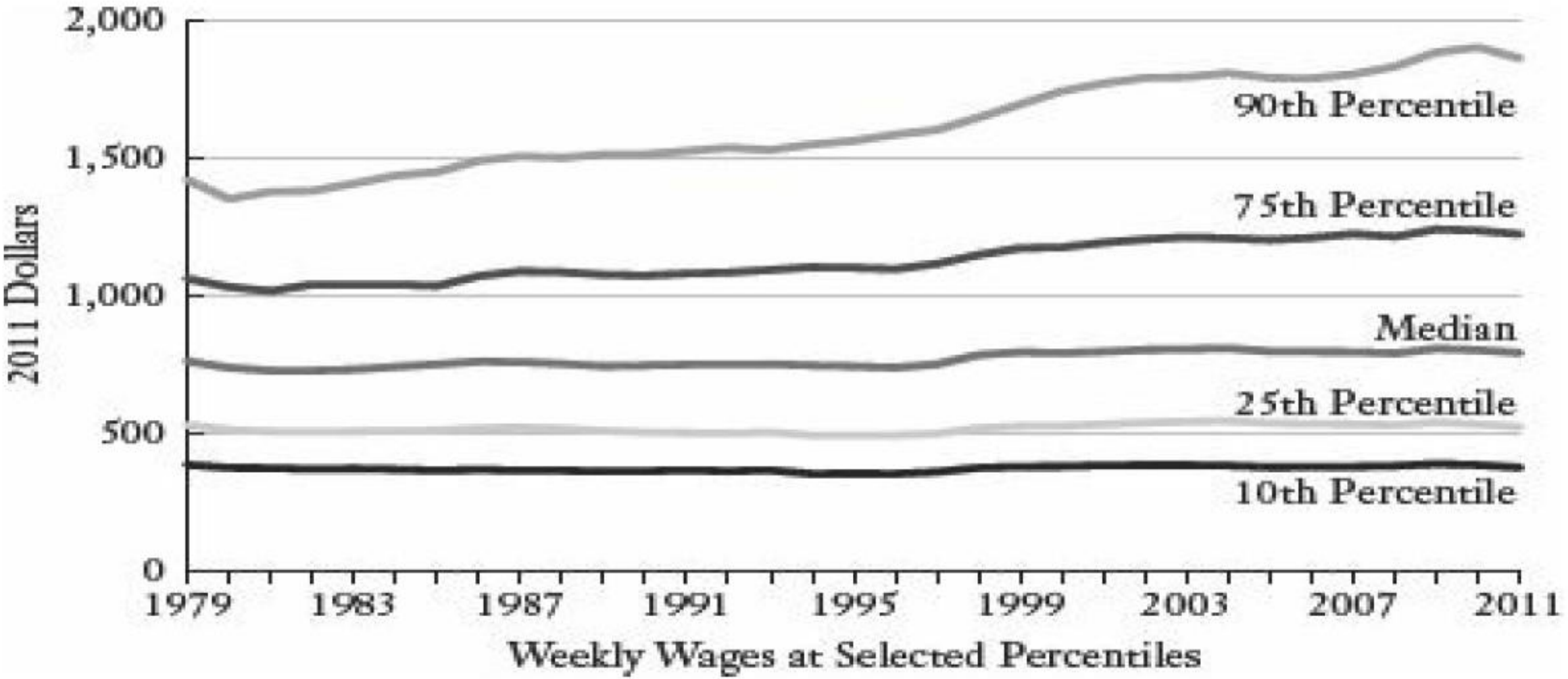
Human Development Index

- The US ranks eleventh in the world in terms of per capita economic output (behind several oil-rich nations like Qatar, Brunei, and Kuwait) but fourth in the world in human development.
- It's true that the HDI rankings would change slightly if the component parts of the index were reconfigured, but no reasonable change is going to make Zimbabwe zoom up the rankings past Norway.
- The HDI provides a handy and reasonably accurate snapshot of living standards around the globe.

American wages

- Having said all that, here is a graph of American wages over the past three decades.
- I've also added the 90th percentile to illustrate changes in the wages for middle-class workers compared over this time frame to those workers at the top of the distribution.

American wages



Deceptive Description

"He's got a great personality!" and other true but grossly misleading statements

He’s got a great personality

- To anyone who has ever planned dating, the phrase “he’s got a great personality” usually sets off alarm bells, not because the description is necessarily wrong, but for what it may *not* reveal, such as the fact that the guy has a prison record or that his divorce is “not entirely final.”
- We don’t doubt that this guy has a great personality;
- we are suspicious that a true statement, **the great personality**, is being used to mask or obscure other information in a way that is seriously misleading (assuming that most of us would prefer not to marry ex-felons who are still married).
- The statement is not a lie per se, meaning that it wouldn’t get you convicted of perjury (lie), but it still could be so inaccurate as to be untruthful.

On the other hand

- And so it is with statistics.
- Although the field of statistics is rooted in mathematics, and ***mathematics is exact***, the use of statistics to describe complex phenomena is ***not exact***.
- That leaves plenty of room for shading the truth.
- Mark Twain famously remarked that there are three kinds of lies: **lies, damned lies, and statistics**.*
- As the last chapter explained, most phenomena that we care about can be described in multiple ways.

multiple ways of describing the same thing

- Once there are multiple ways of describing the same thing
- (e.g., “he’s got a great personality” *or*
- “he was convicted of securities fraud”),
- the descriptive statistics that we choose to use (or not to use) will have a profound impact on the impression that we leave.
- Someone with evil motives can use perfectly good facts and figures to support entirely disputable or illegitimate conclusions.

“precision” != “accuracy”.

- Accuracy is a measure of whether a figure is broadly consistent with the truth
- Precision is to use exact numbers: 3.14
- If an answer is accurate, then more precision is usually better.
- But no amount of precision can make up for inaccuracy.
- In fact, precision can mask inaccuracy by giving us a false sense of certainty, either inadvertently or quite deliberately.

Examples

- Standard yardage markers
  - Even the most precise measurements or calculations should be checked against common sense.
- McCarthy
  - McCarthy waved in the air a piece of paper and declared, “I have here in my hand a list of 205 Communist senators.
  - It turns out that the paper had no names on it at all, but the specificity of the charge gave it credibility, despite the fact that it was a bald-faced lie.

Value at risk

- Many of the Wall Street risk management models prior to the 2008 financial crisis were quite precise.
- The concept of “value at risk” allowed firms to quantify with precision the amount of the firm’s capital that could be lost under different scenarios.
- The problem was that the super sophisticated models were the equivalent of setting my range finder to meters rather than to yards.
- The math was complex and mysterious.
- The answers it produced were reassuringly precise.

Warning !!!!!!!

- But the assumptions about what might happen to global markets that were embedded in the models were just plain wrong, making the conclusions wholly inaccurate in ways that destabilized not only Wall Street but the entire global economy.

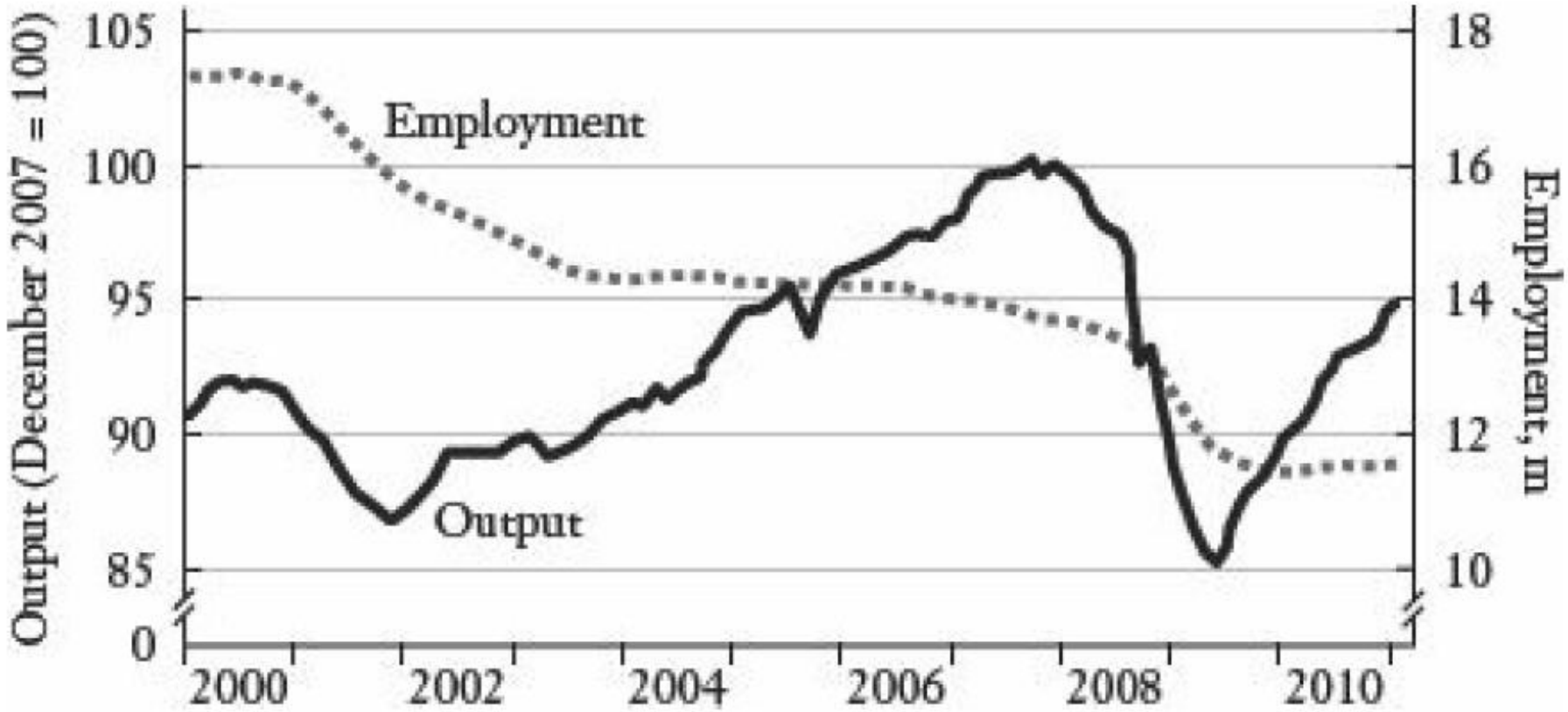
- Even the most precise and accurate descriptive statistics can suffer from a more fundamental problem: **a lack of clarity over what exactly we are trying to define, describe, or explain**.

How healthy is American manufacturing?

- American manufacturing jobs are being lost in huge numbers to China, India, and other low-wage countries.
- High-tech manufacturing still thrives in the U.S. and that America remains one of the world's top exporters of manufactured goods.
- Which is it?
- This would appear to be a case in which sound analysis of good data could reconcile these competing narratives.
- Is U.S. manufacturing profitable and globally competitive, or is it shrinking in the face of intense foreign competition? Both.

The Rustbel recovery

- The British news magazine the *Economist* reconciled the two seemingly contradictory views of U.S. manufacturing with the following graph.



Example

- Together, these two stories—rising manufacturing output and falling employment—tell the complete story.
- Manufacturing in the United States has grown steadily more productive, meaning that factories are producing more output with fewer workers.
- But there are a lot fewer manufacturing jobs, which is terrible news for the displaced workers who depended on those wages.
- The answer depends on how one chooses to define health: output or employment?

Unit of analysis

- Politician A (the challenger): “Our schools are getting worse! **60%** of our schools had lower test scores this year than last year.”
- Politician B (the incumbent): “Our schools are getting better! **80%** of our students had higher test scores this year than last year.”

- It’s entirely possible for most of the students to be improving and most of the schools to be getting worse.
- **if the students showing improvement happen to be in very big schools**.

Unit of analysis

- Politician A (a populist): “Our economy is in the crapper! **30** states had falling incomes last year.”
- Politician B (more of an elitist): “Our economy is showing appreciable gains: **70%** of Americans had rising incomes last year.”

- What I would infer from those statements is that the biggest states have the healthiest economies: New York, California, Texas, Illinois, and so on.
- The thirty states with falling average incomes are likely to be much smaller: Vermont, North Dakota, Rhode Island, and so on.

Globalization good or bad?

- The key lesson is to pay attention to the unit of analysis.
- Who/what is being described, and is that different from the “who” or “what” being described by someone else?
- Is globalization making income inequality around the planet better or worse?
- globalization has merely worsened existing income inequalities;
- richer countries in 1980 (as measured by GDP per capita) tended to grow faster between 1980 and 2000 than poorer countries.

Poor countries or poor people ?

- The rich countries just got richer, suggesting that trade, outsourcing, foreign investment, and the other components of “globalization” are merely tools for the developed world to extend its economic hegemony.

- Down with globalization!
- Down with globalization!

- As the *Economist* points out, “If you consider people, not countries, global inequality is falling rapidly.”
- We don’t care about poor countries; *we care about poor people*.
- And a high proportion of the world’s poor people happen to live in China and India.

Poor countries or poor people ?

- Both countries are huge (with a population over a billion); each was relatively poor in 1980.
- Not only have China and India grown rapidly over the past several decades, but they have done so in large part because of their increased economic integration with the rest of the world.
- They are “rapid globalizers,” as the *Economist* has described them.
- Given that our goal is to improve human misery, it makes no sense to give China (population 1.3 billion) the same weight as Mauritius (population 1.3 million) when examining the effects of globalization on the poor.

AT&T and Verizon

- Verizon launched an aggressive advertising campaign touting the geographic coverage of its network;
- paltry geographic coverage of the AT&T network
- “AT&T covers 97 % of Americans.”

- Note the use of the word “Americans” rather than “America.”
- the key to good cell service is having a network in place where callers actually live and work, not necessarily where they go camping.

Mean or median?

- The mean is a simple average: the sum of the observations divided by the number of observations. (The mean of 3, 4, 5, 6, and 102 is 24.)
- The median is the midpoint of the distribution; half of the observations lie above the median and half lie below. (The median of 3, 4, 5, 6, and 102 is 5.)

- If, for some reason, I would like to describe this group of numbers in a way that **makes it look big**, I will focus on the **mean**.
- If I want to make it **look smaller**, I will cite the **median**.

Bush Tax cut

- While pushing the plan, the administration pointed out that 92 million Americans would receive an average tax reduction of over $1,000 ($1,083 to be precise).
- But was that summary of the tax cut accurate?
- According to the *New York Times*, "The data don't lie, but some of them are mum."
- Would 92 million Americans be getting a tax cut? Yes.
- Would most of those people be getting a tax cut of around $1,000? No.
- The median tax cut was less than $100.

Huh?

- A relatively small number of extremely wealthy individuals were eligible for very large tax cuts; these big numbers skew the mean, making the average tax cut look bigger than what most Americans would likely receive.
- The median is not sensitive to outliers, and, in this case, is probably a more accurate description of how the tax cuts affected the typical household.
- Of course, the median can also do its share of dissembling ***because it is not sensitive to outliers.***

Mean or median?

- Suppose that you have a potentially fatal illness.
- The good news is that a new drug has been developed that might be effective.
- The drawback is that it's extremely expensive and has many unpleasant side effects. "
- But does it work?" you ask.
- The doctor informs you that the new drug increases the median life expectancy among patients with your disease by **two weeks**.
- That is hardly encouraging news; the drug may not be worth the cost and unpleasantness.

Mean or median?

- Yet the median may be a horribly misleading statistic in this case.
- Your insurance company refuses to pay for the treatment; it has a pretty good case on the basis of the median life expectancy figures.
- Suppose that many patients do not respond to the new treatment but that some large number of patients, **say 30 or 40 percent, are cured entirely.**
- This success would not show up in the median (though the mean life expectancy of those taking the drug would look very impressive).
- ***In this case, the outliers—those who take the drug and live for a long time—would be highly relevant to your decision***.

Example: Stephen Jay Gould

- Stephen Jay Gould was diagnosed with a form of cancer that had a median survival time of **eight months**;
- he died of a different and unrelated kind of cancer twenty years later.
- Gould subsequently wrote a famous article called "**The Median Isn't the Message,"** in which he argued that his scientific knowledge of statistics saved him from the erroneous conclusion that he would necessarily be dead in eight months.
- The definition of the median tells us that half the patients will live at least eight months—and possibly much, much longer than that.
- The mortality distribution is "right-skewed," which is more than a technicality if you happen to have the disease.

Mean or median

- In this example, the defining characteristic of the median—that it does not weight observations on the basis of *how far* they lie from the midpoint, only on whether they lie above or below—turns out to be its weakness.
- **In contrast, the mean *is* affected by dispersion.**
- From the standpoint of accuracy, the median versus mean question revolves around whether the outliers in a distribution distort what is being described or are instead an important part of the message.
- Of course, nothing says that you must choose the median or the mean.
- Any comprehensive statistical analysis would likely present both.

“apples and oranges” comparisons.

- Suppose you are trying to compare the price of a hotel room in London with the price of a hotel room in Paris.
- Your child reports back that hotel rooms in Paris are more expensive, around 180 a night; a comparable room in London is 150 a night.

Inflation

- A dollar today is not the same as a dollar 60 years ago; it buys much less.
- Because of inflation, something that cost $1 in 1950 would cost $9.37 in 2011.
- As a result, any monetary comparison between 1950 and 2011 without adjusting for changes in the value of the dollar would be less accurate than comparing figures in euros and pounds—*since the euro and the pound are closer to each other in value than a 1950 dollar is to a 2011 dollar.*
- This is such an important phenomenon that economists have terms to denote whether figures have been adjusted for inflation or not.

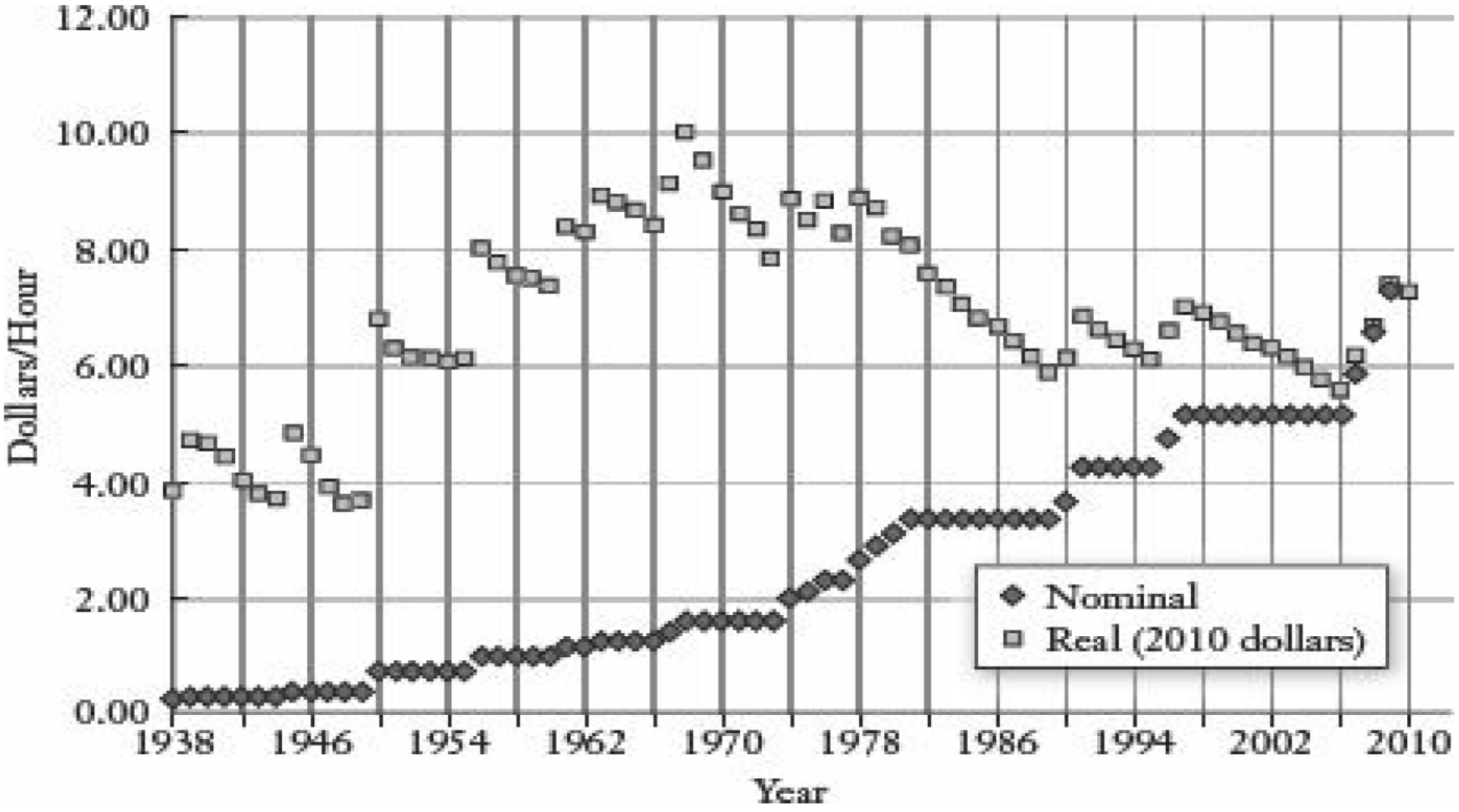
Nominal vs. *Real* figures

- *Nominal* figures are not adjusted for inflation.
- If we spent $10 million on a program in 1970 and $40 million on the same program in 2011, *the federal commitment to that program has actually gone down*.
- Yes, spending has gone up in nominal terms, but that does not reflect the changing value of the dollars being spent.
- One 1970 dollar is equal to$5.83 in 2011; the government would need to spend $58.3 million in 2011 to provide support comparable to the $10 million it was spending in 1970.

Nominal figures vs. *Real* figures

- *Real* figures, on the other hand, are adjusted for inflation.
- Check out the following graph of the U.S. federal minimum wage, which plots both the nominal value of the minimum wage and its real purchasing power in 2010 dollars.
- The federal minimum wage is all about the number on the check, not what that number can buy.

Nominal figures vs. *Real* figures



Hollywood studios box office

- What were the top five highest-grossing films (domestic) of all time as of 2011?
- *1. Avatar (2009)*
- *2. Titanic (1997)*
- *3. The Dark Knight* (2008)
- *4. Star Wars Episode IV* (1977)
- *5. Shrek 2* (2004)
- Hollywood likes to make each blockbuster look bigger and more successful than the last.

Hollywood studios box office

- So what are the top grossing films in the U.S. of all time, *adjusted for inflation*?
- 1. *Gone with the Wind* (1939)
- 2. *Star Wars Episode IV* (1977)
- 3. *The Sound of Music* (1965)
- 4. *E.T.* (1982)
- 5. *The Ten Commandments* (1956)

In real terms, *Avatar* falls to number 14; *Shrek 2* falls all the way to 31st.

What are the Odds ?

- We often use percentages to express these changes because they give us a sense of scale and context.
- Percentages don't lie—but they can exaggerate.
- The portion of taxes in supporting for the **Suburban Cook County Tuberculosis Sanitarium District** was slated to rise by **527%!**
- However, this change would cost me less than a good turkey sandwich.
- The **Tuberculosis Sanitarium District** deals with roughly a hundred cases a year; it is not a large or expensive organization.
- The *Chicago Sun-Times* pointed out that for the typical homeowner, the tax bill would go from **$1.15 to $6.**

from a low base

- Researchers will sometimes qualify a growth figure by pointing out that it is "**from a low base**," meaning that any increase is going to look large by comparison.
- Suppose the secretary of defense reports that defense spending will grow only 4 % this year. Great news!
- Not really, given that the Defense Dep. budget is nearly $700 billion.
- Four % of $700 billion is $28 billion, that seemingly paltry 4 % increase in the defense budget is ***more than the entire NASA budget and about the same as the budgets of the Labor and Treasury Departments combined.***

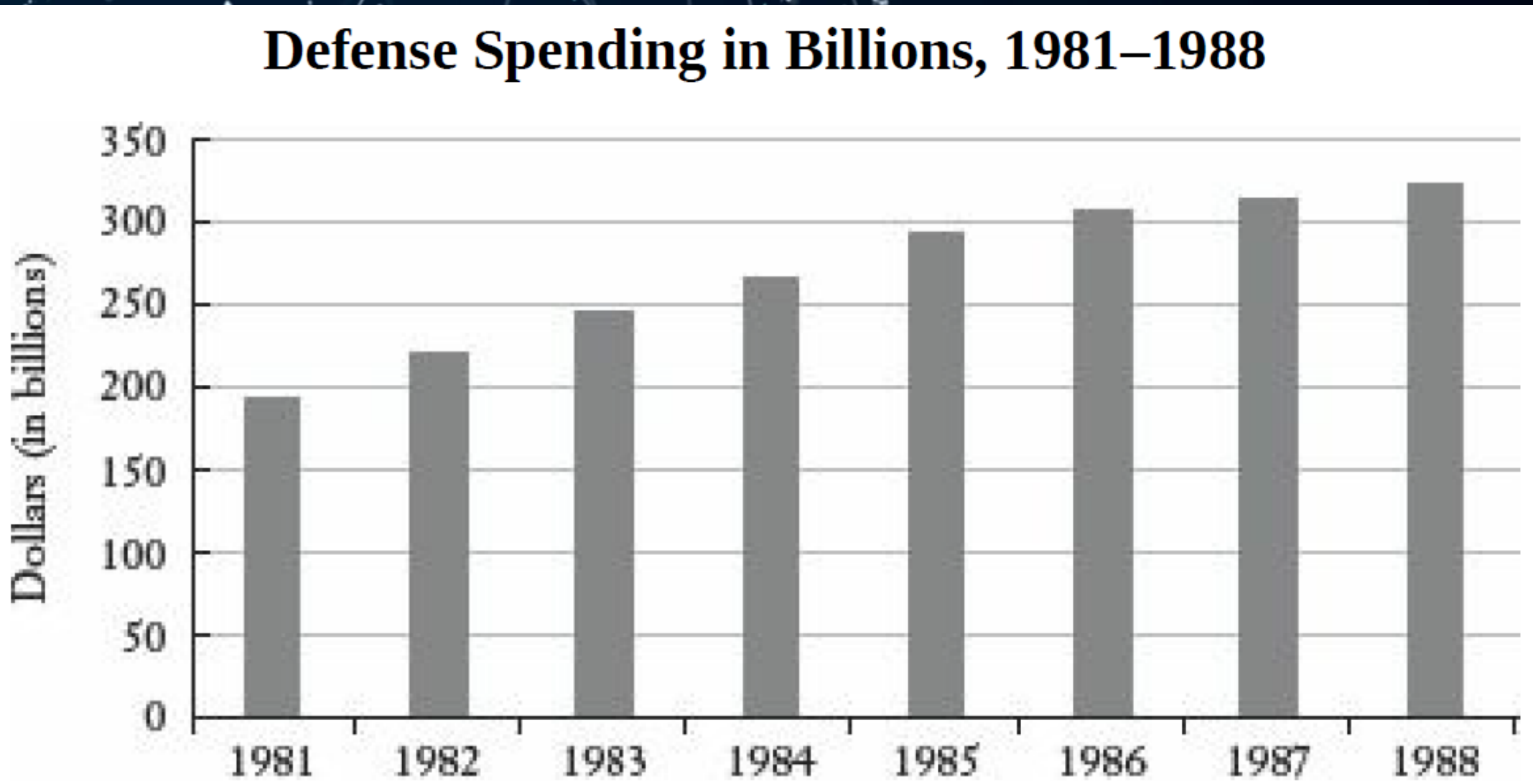
The normal distribution.

- In a similar vein, your boss might point out that as a matter of fairness, every employee will be getting the same raise this year, 10%.
- What a magnanimous gesture—except that if your boss makes $1 million and you make $50,000, his raise will be $100,000 and yours will be $5,000.
- The statement "***everyone will get the same 10% raise this year***" just sounds so much better than "***my raise will be twenty times bigger than yours***."
- Both are true in this case.

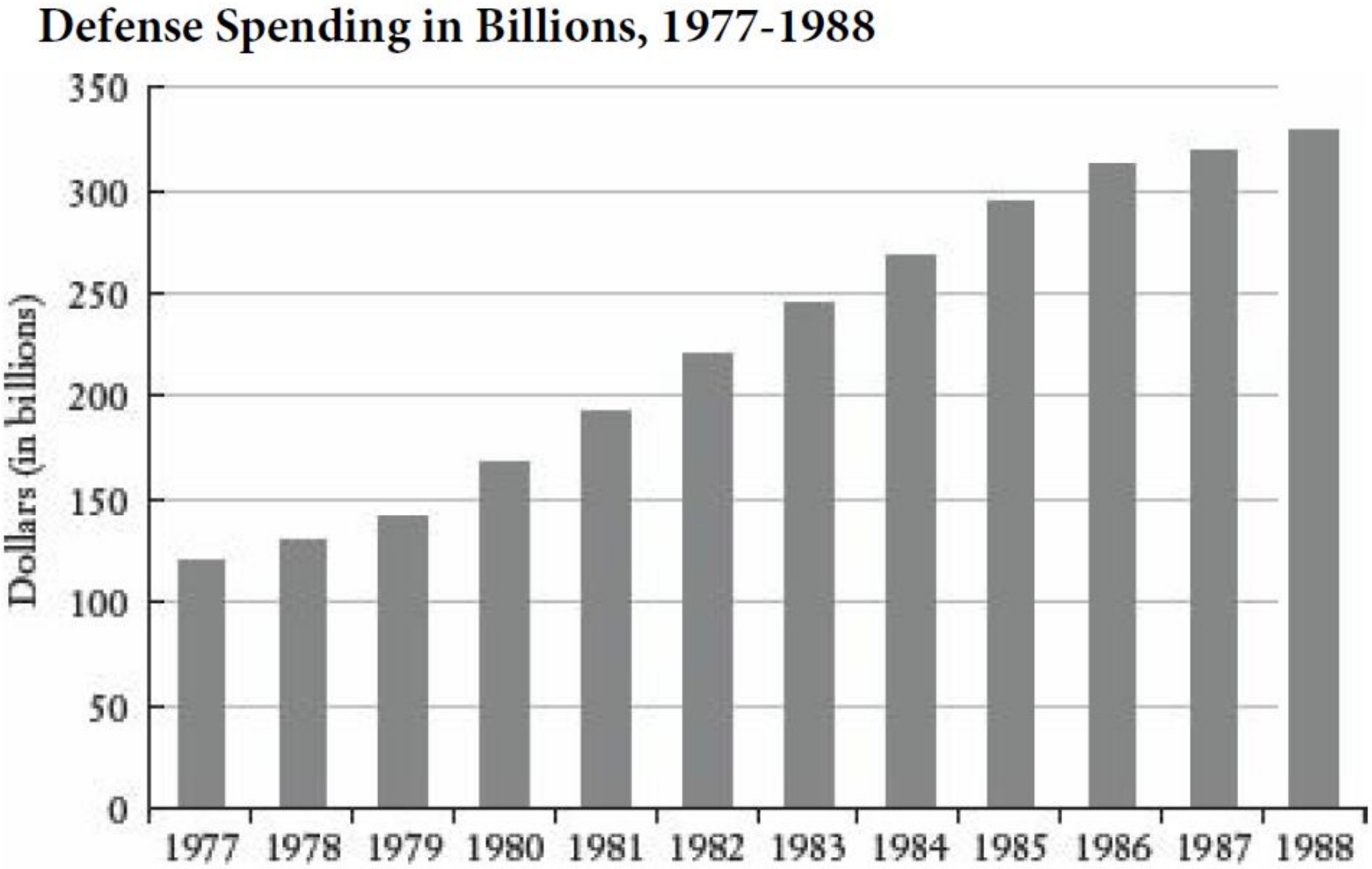
changing over time

- Any comparison of a quantity changing over time must have a start point and an end point.
- One can sometimes manipulate those points in ways that affect the message.

changing over time



changing over time



Measurement

- "You can't manage what you can't measure." True.

- *But you had better be darn sure that what you are measuring is really what you are trying to manage.*
- The most common measure of quality for both schools and teachers is test scores.
- If students are achieving impressive scores on a well-conceived standardized test, then presumably the teacher and school are doing a fine job. Wrong.
- Any evaluation of teachers or schools that is based solely on test scores will present a dangerously inaccurate picture.

selective enrollment schools

- Students who live in affluent, highly educated communities are going to test well from the moment their parents drop them off at school on the first day of kindergarten.
- What we need is some measure of “value-added” at the school level, or even at the classroom level.
- Students with different abilities or backgrounds may also learn *at different rates*.
- Researchers are working to develop statistical techniques that measure instructional quality in ways that account appropriately for different student backgrounds and abilities.

selective enrollment schools

- One of the most important admissions criteria is standardized test scores.
- So let's summarize:
- (1) these schools are being recognized as "excellent" for having students with high test scores;
- (2) to get into such a school, one must have high test scores.

- This is the logical equivalent of giving an award to the basketball team for doing such an excellent job of producing tall students.

Selective enrollment schools

- Even if you have a solid indicator of what you are trying to measure and manage, the challenges are not over.
- The good news is that “managing by statistics” can change the underlying behavior of the person or institution being managed for the better.
- Statistics measure the outcomes that matter; incentives give us a reason to improve those outcomes.
- Or, in some cases, just to make the statistics look better.
- That’s the bad news

"moving away" rather than dropping out.

- Here is another example: "It's never a good day when *60 Minutes* shows up at your door."
- Dan Rather and the *60 Minutes II* crew made a trip to Houston and found that the manipulation of statistics was far more impressive than the educational improvement.
- High schools routinely classified students **who quit high school as transferring to another school returning to their native country, or leaving to pursue a General Equivalency Diploma (GED).**

“moving away” rather than dropping out.

- None of which count as dropping out in the official statistics.
- Houston reported a citywide dropout rate of 1.5 % in the year that was examined;
- *60 Minutes* calculated that the true dropout rate was between 25 and 50 percent.
- Another way to improve test scores is to prevent the worst students from taking the test.
- If the scores of the lowest-performing students are eliminated, the average test score for the school or district will go up, even if all the rest of the students show no improvement at all.

Don’t let them take the exam

- Prevent the worst students from taking the test
- In Texas, the statewide achievement test is given in tenth grade.
- There was evidence that Houston schools were trying to keep the weakest students from reaching tenth grade.
- In one particularly egregious example, a student spent three years in ninth grade and then was promoted straight to eleventh grade—a deviously clever way of keeping a weak student from taking a tenth-grade benchmark exam without forcing him to drop out (which would have showed up on a different statistic).

Rod Paige

- It's not clear that Rod Paige was complicit in this statistical trickery during his tenure as Houston superintendent; however, he did implement a rigorous accountability program that gave cash bonuses to principals who met their dropout and test score goals and that fired or demoted principals who failed to meet their targets.
- Principals definitely responded to the incentives; that's the larger lesson.
- But you had better be darn certain that the folks being evaluated can't make themselves look better (statistically) in ways that are not consistent with the goal at hand.

Rod Paige

- It's not clear that Rod Paige was complicit in this statistical trickery during his tenure as Houston superintendent; however, he did implement a rigorous accountability program that gave cash bonuses to principals who met their dropout and test score goals and that fired or demoted principals who failed to meet their targets.
- Principals definitely responded to the incentives; that's the larger lesson.
- But you had better be darn certain that the folks being evaluated can't make themselves look better (statistically) in ways that are not consistent with the goal at hand.

Scorecards

- The state introduced “scorecards” that evaluate the mortality rates for the patients of cardiologists performing coronary angioplasty, a common treatment for heart disease.
- This seems like a perfectly reasonable and helpful use of descriptive statistics.
- The proportion of a cardiologist’s patients who die in surgery is an important thing to know, and it makes sense for the government to collect and broadcast such data since individual consumers would not otherwise have access to it.
- So is this a good policy? **The easiest way for a doctor to improve his mortality rate is by refusing to operate on the sickest patients.**

Scorecards

- 83 % of the cardiologists surveyed said that, because of the public mortality statistics, some patients who might benefit from angioplasty might not receive the procedure;
- 79 % of the doctors said that some of their personal medical decisions had been influenced by the knowledge that mortality data are collected and made public.

- A statistical index has all the potential pitfalls of any descriptive statistic—plus the distortions introduced by combining multiple indicators into a single number.

Judgment

- The benefit of the USNWR rankings is that they provide lots of information about thousands of schools in a simple and accessible way.
- "Each indicator is assigned a weight (expressed as a percentage) based on our judgments about which measures of quality matter most."
- Judgment is one thing; arbitrariness is another.
- He cites a questionnaire sent out by a former chief justice of the Michigan Supreme Court to roughly one hundred lawyers asking them to rank ten law schools in order of quality.
- Penn State's was one of the law schools on the list; the lawyers ranked it near the middle.
- *At the time, Penn State did not have a law school.*

People love easy answers.

- What is the best place? Number 1
- The overall lesson of this chapter is that statistical malfeasance has very little to do with bad math. If anything, impressive calculations can obscure nefarious motives.
- The fact that you've calculated the mean correctly will not alter the fact that the median is a more accurate indicator.
- Judgment and integrity turn out to be surprisingly important.
- A detailed knowledge of statistics does not deter wrongdoing any more than a detailed knowledge of the law averts criminal behavior.
- With both statistics and crime, the bad guys often know exactly what they're doing!

CHAPTER 4: Correlation

How does Netflix know what movies I like?

Correlation

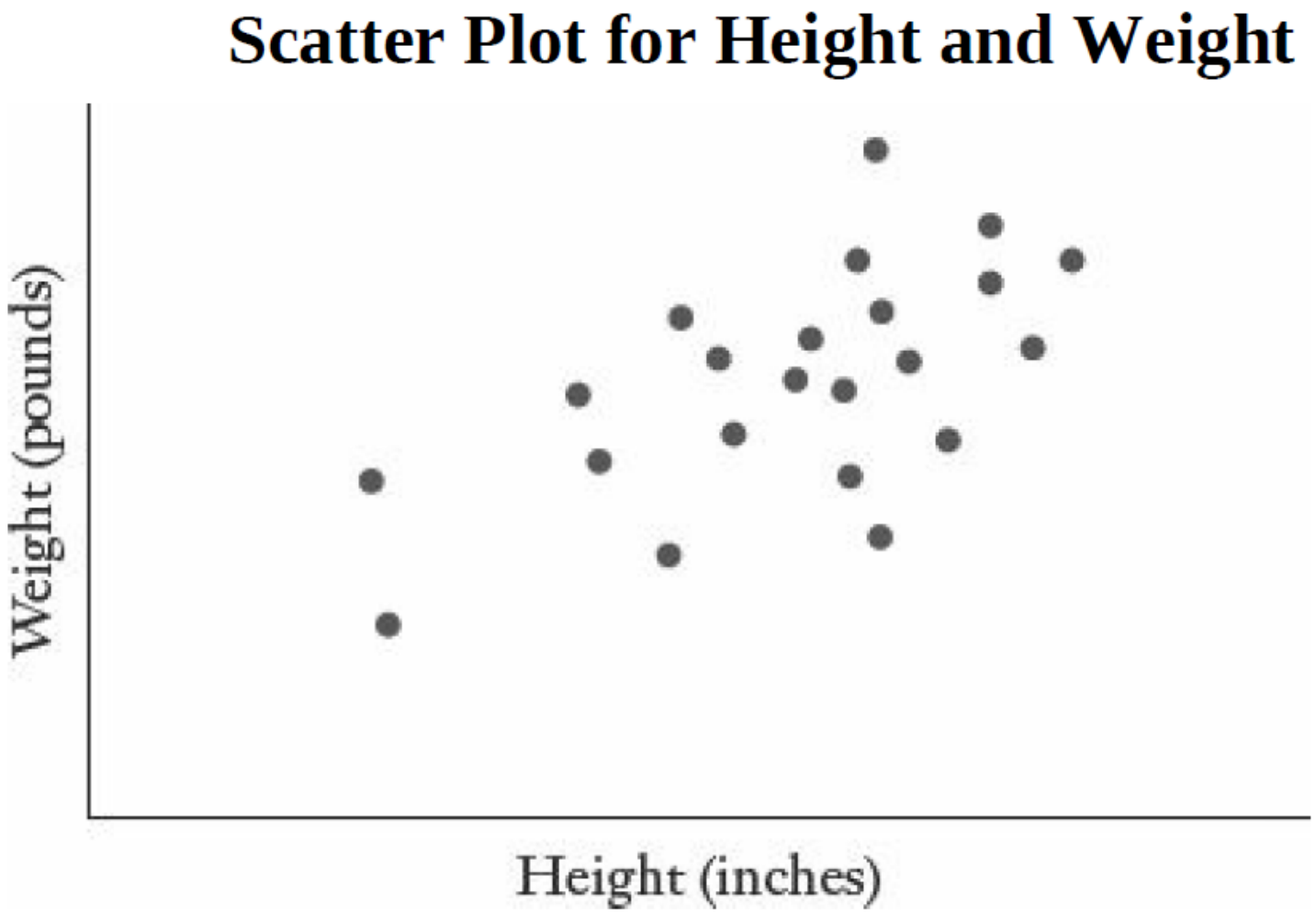
- The Netflix recommendations that I've watched in the past have been terrific.
- And when a film is recommended that I've already seen, it's typically one I've really enjoyed.
- it's all based on correlation.
- Correlation measures the degree to which two phenomena are related to one another.
- For example, there is a correlation between summer temperatures and ice cream sales.
- When one goes up, so does the other.

On the other hand

- Two variables are positively correlated if a change in one is associated with a change in the other in the same direction, such as the relationship between height and weight.
- A correlation is negative if a positive change in one variable is associated with a negative change in the other, such as the relationship between exercise and weight.
- The tricky thing about these kinds of associations is that not every observation fits the pattern.
- Sometimes short people weigh more than tall people.

Height and weight

- Sometimes people who don't exercise are skinnier than people who exercise all the time.
- Still, there is a meaningful relationship between height and weight, and between exercise and weight.



The correlation coefficient

- Instead, the power of correlation as a statistical tool is that we can encapsulate an association between two variables in a single descriptive statistic: the correlation coefficient.
- First, it is a single number ranging from –1 to 1.
- A correlation of 1, often described as perfect correlation, means that every change in one variable is associated with an equivalent change in the other variable in the same direction.
- A correlation of –1, or perfect negative correlation, means that every change in one variable is associated with an equivalent change in the other variable in the opposite direction.
- The closer the correlation is to 1 or –1, the stronger the association.

The correlation coefficient

- A correlation of 0 (or close to it) means that the variables have no meaningful association with one another, such as the relationship between shoe size and SAT scores.
- The second attractive feature of the correlation coefficient is that it has no units attached to it.
- The correlation coefficient does a seemingly miraculous thing: It collapses a complex mess of data measured in different units (like our scatter plots of height and weight) into a single, elegant descriptive statistic.

The correlation coefficient

- 1.Calculates the mean and standard deviation for both variables.
- 2. Converts all the data so that each observation is represented by its distance (in standard deviations) from the mean.
- 3. Here I'll wave my hands and let the computer do the work.

- When individuals in the sample are tall, say, 1.5 or 2 standard deviations above the mean, what do their weights tend to be *as measured in standard deviations from the mean for weight*?
- And when individuals are near to the mean in terms of height, what are their weights as measured in standard units?

strong correlation

- If the distance from the mean for one variable tends to be broadly consistent with distance from the mean for the other variable then we would expect a **strong positive correlation**.
- If distance from the mean for one variable tends to correspond to a similar distance from the mean for the second variable *in the other direction then we would expect a* **strong negative correlation**.

- Scholastic Aptitude Test, is a standardized exam made up of three sections: math, reading, and writing.

Sat EXAMPLE

- The purpose of the test is to measure academic ability and predict college performance.
- Isn't that what high school is for?
- Why is a four hour test so important when college admissions officers have access to *four years* of high school grades?
- ***High school grades are an imperfect descriptive statistic.***
- A student who gets average grades while taking a tough schedule of math and science classes may have more academic ability and potential than a student at the same school with better grades in less challenging classes.

But is SAT a good measure of ability?

- If we want a metric that can be compared easily across students, we could also have all high school seniors run the 100 yard dash, which is cheaper and easier than administering the SAT.
- The problem, of course, is that performance in the 100 yard dash is uncorrelated with college performance.
- ***It's easy to get the data; they just won't tell us anything meaningful.***
- The SAT does a reasonably good job of predicting first year college grades.
- The College Board publishes the relevant correlations.

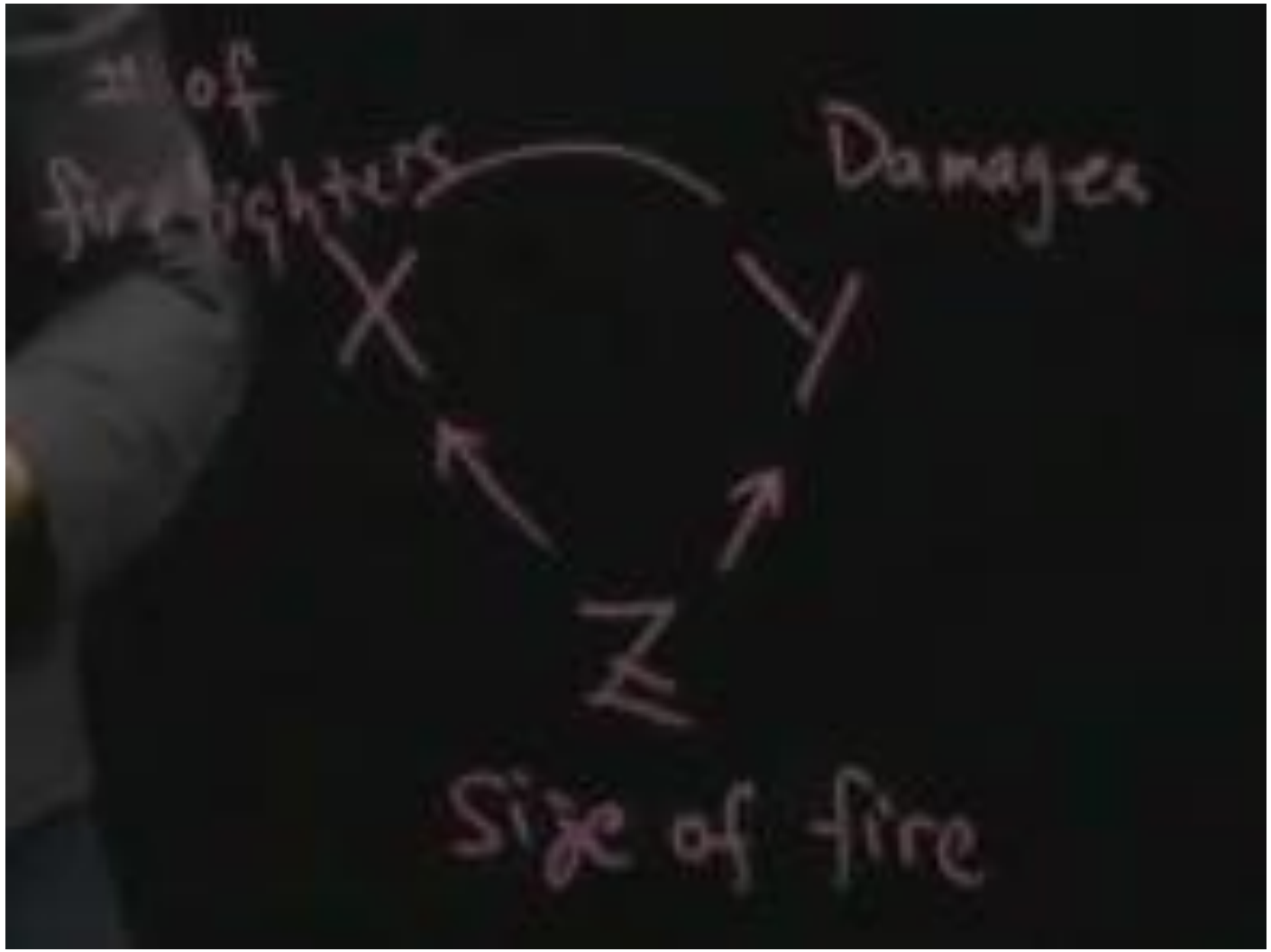
But is SAT a good measure of ability?

- On a scale of 0 (no correlation at all) to 1 (perfect correlation),
- The correlation between high school grade point average and first-year college grade point average is 0.56.
- (To put that in perspective, the correlation between height and weight for adult men in the United States is about 0.4.)
- That would seem to argue for ditching the SAT, as the test does not seem to do any better at predicting college performance than high school grades.

correlation does not imply causation

- A positive or negative association between two variables does not necessarily mean that a change in one of the variables is causing the change in the other.
- For example, a likely positive correlation between a student's SAT scores and the number of televisions that his family owns.
- This does not mean that overeager parents can boost their children's test scores by buying an extra five TVs for the house.
- Nor does it likely mean that watching lots of TV is good for academic achievement.
- The most logical explanation for such a correlation would be that highly educated parents can afford a lot of TVs and tend to have children who test better than average.
- Both the TVs and the test scores are likely caused by a third variable, which is parental education.

correlation does not imply causation



That's the "big picture."

- The actual methodology is much more complex.
- In fact, Netflix launched a contest in 2006 in which members of the public were invited to design a mechanism that improved on existing Netflix recommendations by at least 10 % (meaning that the system was 10 % more accurate in predicting how a customer would rate a film after seeing it).
- The winner would get $1,000,000.
- Every individual or team that registered for the contest was given "training data" consisting of more than 100 million ratings of 18,000 films by 480,000 Netflix customers.

Netflix

- A separate set of 2.8 million ratings was “withheld,” meaning that Netflix knew how the customers rated these films but the contest participants did not. The competitors were judged on how well their algorithms predicted the actual customer reviews for these withheld films.
- Over three years, thousands of teams from over 180 countries submitted proposals.
- There were two requirements for entry.
- First, the winner had to license the algorithm to Netflix.
- And second, the winner had to “describe to the world how you did it and why it works.”

discuss association between variables

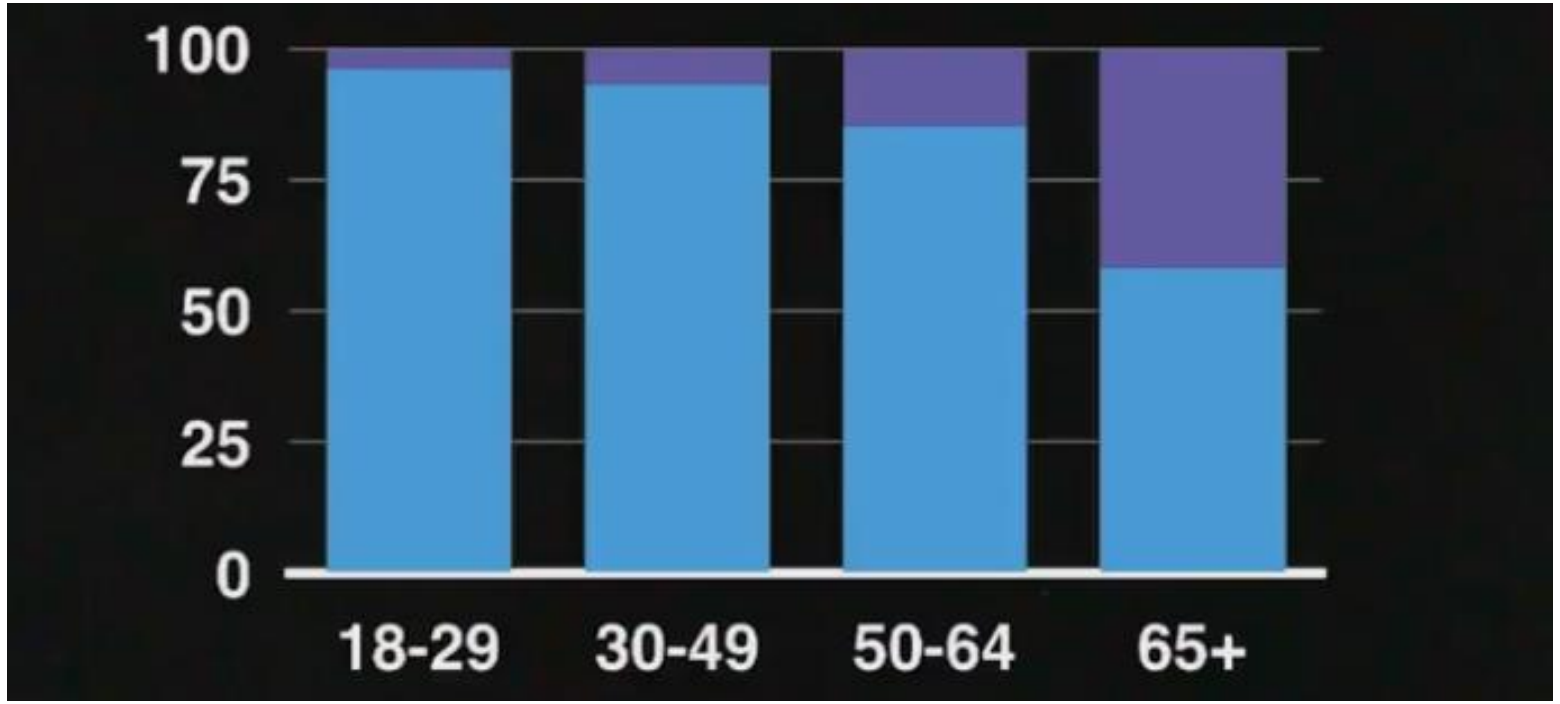
- Association is defined as when you observe certain values of one variable are observed more frequently, more often, with certain values of another variable.

	Use Internet	Do not use
18-29	48	2
30-49	93	7
50-64	85	5
65+	29	21

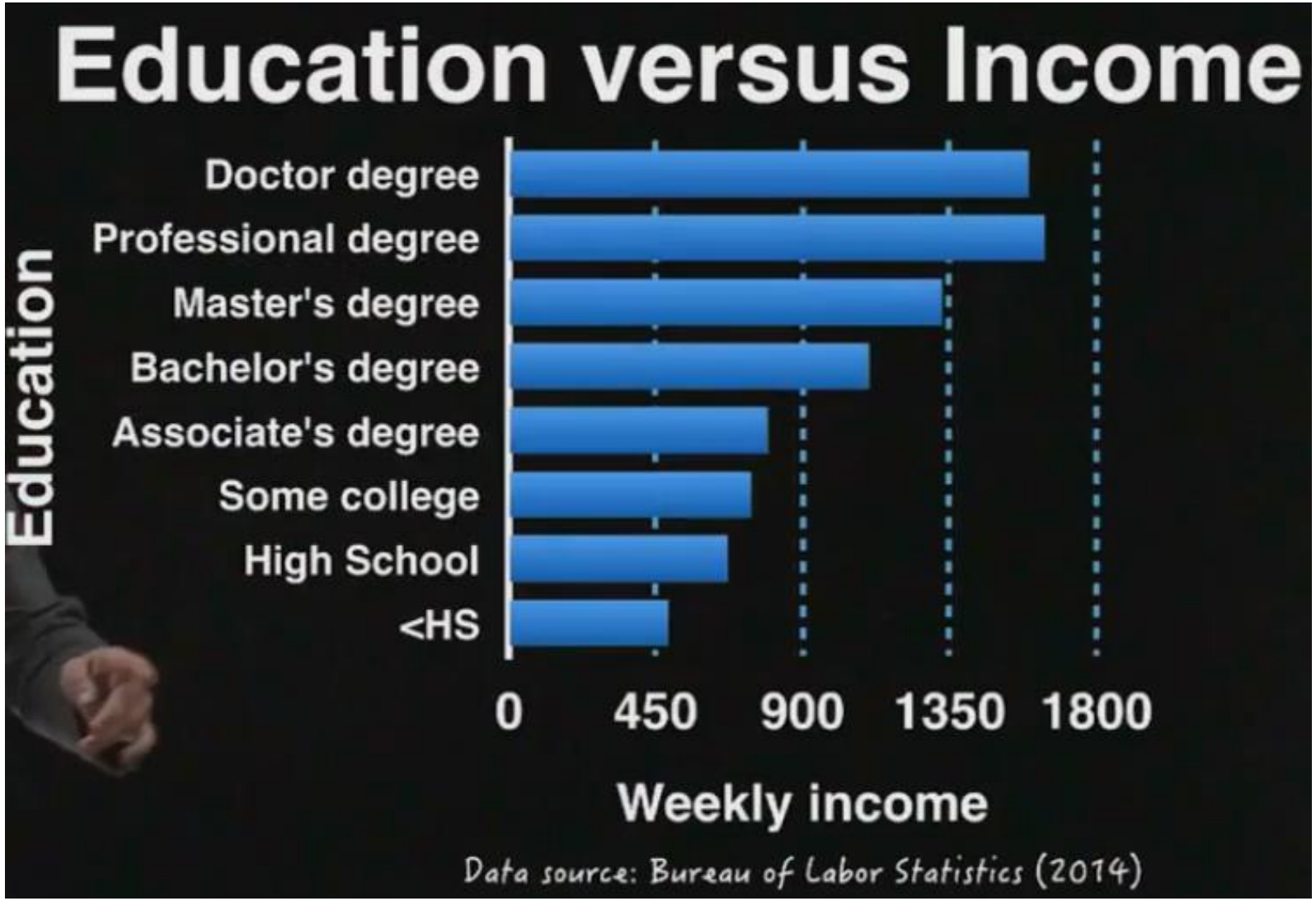
Hypothetical example based on findings from Pew Research Center on Internet Use (August 2012)

discuss association between variables

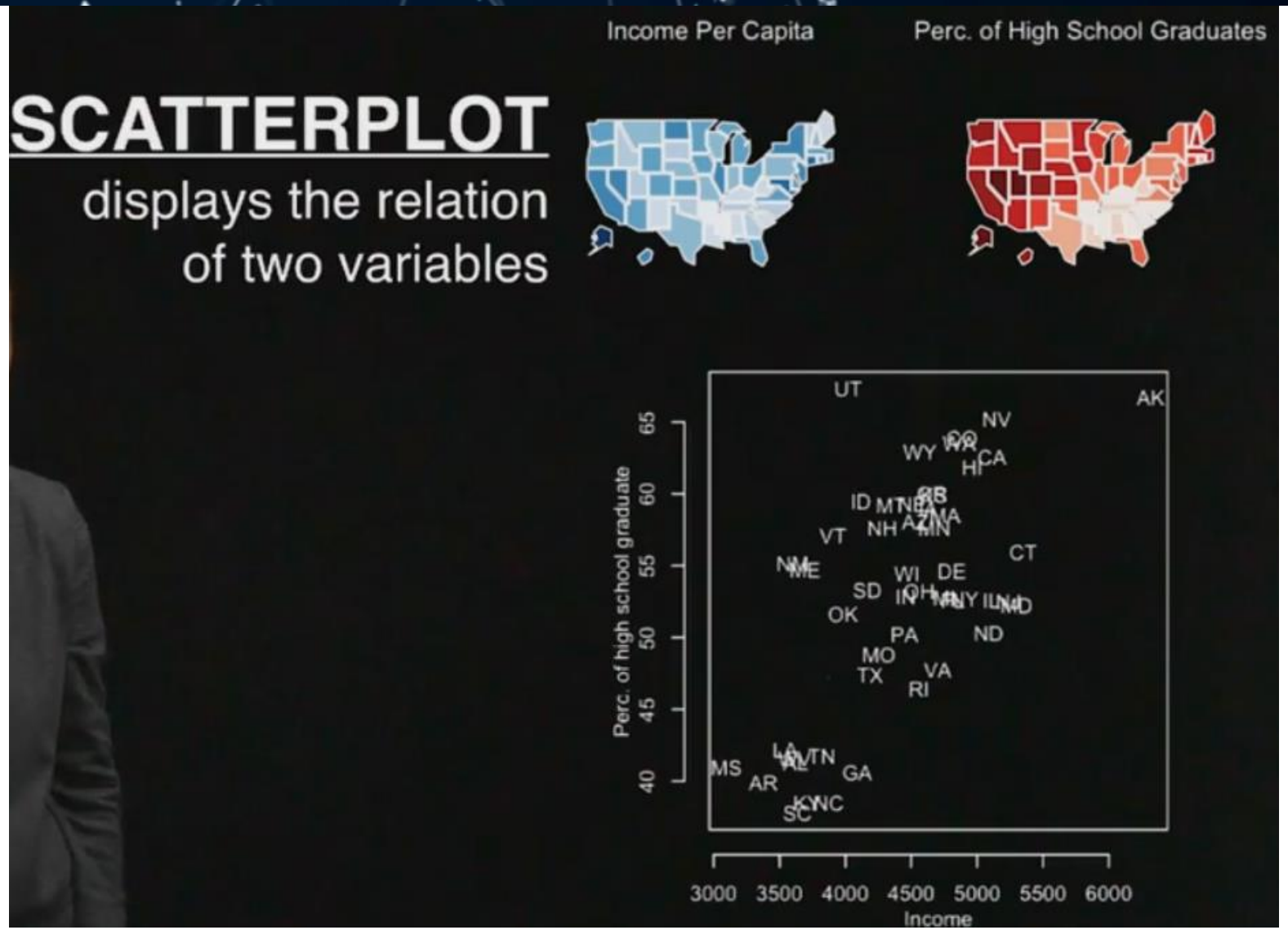
- Association is defined as when you observe certain values of one variable are observed more frequently, more often, with certain values of another variable.



Numerical Data Association



Numerical Data Association-Scatterplot



measure the association

- There are several ways we can compute numerical statistics to measure the association between age and internet use in this example.
- We can first report the **difference in proportions**, such as the difference between 96% and 58%.
- Second, we can compute something called **relative risks**, which work as such.
- It's related to the destination of probability, especially conditional probability.

$$P_1 = P(internet|young) = 98\%$$
$$P_2 = P(internet|senior) = 58\%$$
$$Relative\ risk = P_1 / P_2 = 1.689$$

Odds ratio

- The third way of summarizing this in statistics is something called odds ratio.
- ODDS=chance or likelihood

$$Odds\ ratio = \frac{\frac{P_1}{1-P_1}}{\frac{P_2}{1-P_2}}$$

And this statistic calculates the ratio of the odds of having the internet among the young adults and among the senior adults. Therefore, it is called observation.

Basic Probability

- *Don't buy the extended warranty on your $99 printer*

Risky marketing campaign Schlitz

- At halftime of the Super Bowl, in front of 100 million people around the world, the company broadcast a live taste test pitting Schlitz Beer against a key competitor, Michelob.
- Bolder yet, the company did not pick random beer drinkers to evaluate the two beers; *it picked 100 Michelob drinkers*.
- This was the culmination (peak) of a campaign that had run throughout the NFL playoffs.
- There were five live television taste tests in all, each of which had 100 consumers of a competing brand (Budweiser, Miller, or Michelob) conduct a blind taste test between their supposed favorite beer and Schlitz.
- Each of the beer taste-offs was promoted aggressively, just like the playoff game during which it would be held (e.g., “Watch Schlitz v. Bud, Live during the AFC Playoffs”).

On the other hand

- The marketing message was clear: Even beer drinkers who think they like another brand will prefer Schlitz in a blind taste test.
- For the Super Bowl spot, Schlitz even hired a former NFL referee to oversee the test.
- Given the risky nature of conducting blind taste tests in front of huge audiences on live TV, one can assume that Schlitz produced a spectacularly delicious beer, right?
- Not necessarily. Schlitz needed only a mediocre beer and a solid grasp of statistics to know that this would almost certainly work out in its favor.
- Most beers in the Schlitz category taste about the same; ironically, that is exactly the fact that this advertising campaign exploited.

essentially a coin flip

- Assume that the typical beer drinker cannot tell Schlitz from Budweiser from Michelob from Miller.
- In that case, a blind taste test between any two of the beers is essentially a coin flip.
- On average, half the taste testers will pick Schlitz, and half will pick the beer it is "challenging."
- This fact alone would probably *not* make a particularly effective advertising campaign.
- And Schlitz absolutely, positively would not want to do this test among its own loyal customers; roughly half of these Schlitz drinkers would pick the competing beer.
- It looks bad when the beer drinkers supposedly most committed to your brand choose a competitor in a blind taste test—which is exactly what Schlitz was trying to do to its competitors.

- Schlitz did something cleverer.
- The genius of the campaign was conducting the taste test exclusively among beer drinkers who stated that they preferred a competing beer.
- If the blind taste test is really just a coin flip, then roughly half of the Budweiser or Miller or Michelob drinkers will end up picking Schlitz.
- That makes Schlitz look really good.
- *Half of all Bud drinkers like Schlitz better!*

bad luck in the taste test would be a disaster

- Michelob drinker will pick Schlitz 50 % of the time,
- what if the 100 Michelob drinkers taking the test turn out to be quirky (odd)? Yes, the blind taste test is the equivalent of a coin toss, but what if most of the tasters chose Michelob *just by chance*?
- After all, if we lined up the same 100 guys and asked them to flip a coin, it's entirely possible that they would flip 85 or 90 tails.
- That kind of bad luck in the taste test would be a disaster for the Schlitz brand (not to mention a waste of the $1.7 million for the live television coverage).

Statistics to the rescue!

- If there were some kind of statistics superhero and unveiled the details of what statisticians call a ***binomial experiment (***or a ***Bernoulli trial).***
- The key characteristics of a binomial experiment are that we have a fixed number of trials (e.g., 100 taste testers), each with two possible outcomes (Schlitz or Michelob), and the probability of "success" is the same in each trial.
- Assuming the probability of picking one beer or the other is 50 percent,
- Defining "success" as a tester picking Schlitz.
- Assume that all the "trials" are independent,

EXAMPLE

- With only this information, a statistical superhero can calculate the probability of all the different outcomes for the 100 trials, such as 52 Schlitz and 48 Michelob or 31 Schlitz and 69 Michelob.
- Those of us who are not statistical superheroes can use a computer to do the same thing.
- The chances of all 100 taste testers picking Michelob were 1 in 1,267,650,600,228,229,401,496,703,205,376.
- There was probably a bigger chance that all of the testers would be killed at halftime by an asteroid.

risky gambit

- Let's assume that Schlitz would have been pleased if at least 40 of the 100 tasters picked Schlitz—an impressive number given that all of the men taking the live blind taste test had declared to be Michelob drinkers.
- An outcome *at least that good* was highly likely.
- If the taste test is really like a flip of the coin, then basic probability tells us that there was a 98 % chance that at least 40 of the tasters would pick Schlitz, and an 86 % chance that at least 45 of the tasters would.
- In theory, this wasn't a very risky gambit at all.

So what happened to Schlitz?

- At halftime of the 1981 Super Bowl, exactly 50 % of the Michelob drinkers chose Schlitz in the blind taste test.
- There are two important lessons here:
  1. probability is a remarkably powerful tool, and
  2. many leading beers in the 1980s were indistinguishable from one another.
- This chapter will focus primarily on the first lesson.

Probability

- Probability is the study of events and outcomes involving an element of uncertainty.
- Examples:
  1. Investing in the stock market or
  2. flipping a coin.
  3. Flipping a coin four times in a row
- The last one involves additional layers of uncertainty, because each of the four flips can result in a head or a tail.
- If you flip a coin four times in a row, I cannot know the outcome in advance with certainty (nor can you).

Many events have known probabilities

- Yet I *can* determine in advance that some outcomes (two heads, two tails) are more likely than others (four heads).
- In fact, if you can understand why the probability of flipping four heads in a row with a fair coin is 1 in 16, you can (with some work) understand everything from how the insurance industry works .

- Many events have known probabilities
- Other events have probabilities that can be **inferred on the basis of past data.**

Fly rather than drive

- For example, the Australian Transport Safety Board published a report quantifying the fatality risks for different modes of transport.
- Despite widespread fear of flying, the risks associated with commercial air travel are tiny.
- Australia hasn't had a commercial air fatality since the 1960s, so the fatality rate per 100 million kilometers traveled is essentially zero.
- The rate for drivers is .5 fatalities per 100 million kilometers traveled.
- The really impressive number is for motorcycles—be an organ donor.
- The fatality rate is **thirty-five** times higher for motorcycles than for cars.

That's the "big picture."

- In September of 2011, a 6.5-ton NASA satellite was plummeting to earth and was expected to break apart once it hit the earth's atmosphere.
- What were the chances of being struck by the debris?
- Should I have kept the kids home from school?
- The rocket scientists at NASA estimated that the probability of any individual person's being hit by a part of the falling satellite was 1 in 21 trillion.
- In the end, the satellite did break apart on reentry, but scientists aren't entirely certain where all the pieces ended up.
- No one reported being hurt.

What is *likely to happen?*

- Probabilities do not tell us what will happen for sure; they tell us what is *likely to happen* and what is *less likely to happen*.
- Sensible people can make use of these kinds of numbers in business and life.
- For example, when you hear on the radio that a satellite is plummeting to earth, you should not race home on your motorcycle to warn the family.

Freakonomics

- When it comes to risk, our fears do not always track with what the numbers tell us we should be afraid of.
- One of the striking findings from *Freakonomics*, by Steve Levitt and Stephen Dubner, was that swimming pools in the backyard are far more dangerous than guns in the closet.
- Levitt and Dubner calculate that a child under ten is one hundred times more likely to die in a swimming pool than from a gun accident.

Freakonomics

- An intriguing paper by three Cornell researchers, found that thousands of Americans may have died since the September 11 **attacks *because they were afraid to fly***.
- We will never know the true risks associated with terrorism; we do know that driving is dangerous.
- When more Americans opted to drive rather than to fly after 9/11, there were an estimated 344 additional traffic deaths per month in October, November, and December of 2001 (taking into account the average number of fatalities and other factors that typically contribute to road accidents, such as weather).

What likely happened

- Probability can also sometimes tell us *after the fact* what likely happened and what likely did not happen—as in the case of DNA analysis.
- When the technicians find a trace of saliva on an apple core near a murder victim, that saliva does not have the murderer's name on it, even when viewed under a powerful microscope by a very attractive technician.
- Instead, the saliva (hair/skin/bone fragment) will contain a DNA segment.
- Each DNA segment in turn has regions, or **loci**, that can vary from individual to individual (**except for identical twins, who share the same DNA**).

DNA sample is a “match,”

- When the medical examiner reports that a DNA sample is a “match,” that’s only part of what the prosecution has to prove.
- Yes, the loci tested on the DNA sample from the crime scene must match the loci on the DNA sample taken from the suspect.
- However, the prosecutors must also prove that the match between the two DNA samples is not merely a coincidence.
- **More than 99 % of all DNA is identical among all humans**.
- the more loci that can be tested, and the more natural genetic variation there is in each of those loci, the more certain the match becomes.
- Or, to put it a bit differently, the less likely it becomes that the DNA sample will match more than one person

DNA sample is a “match,”

- You should see three things.
  - First, anything less than a full match of the entire genome leaves some room for uncertainty.
  - Second, the more “loci” that can be tested, the less uncertainty remains. When researchers have unlimited time and resources, the typical process involves testing thirteen different loci.
  - And third, context matters.
- This match would be extremely compelling if you also happened to be caught speeding away from the crime scene with the victim’s credit cards in your pocket.

DNA sample is a “match,”

- The intellectual pushback was instigated when a crime lab analyst in Arizona running tests with the state’s DNA database discovered two unrelated felons whose DNA matched at nine loci;
- according to the FBI, the chances of a nine-loci match between two unrelated persons are 1 in 113 billion.
- **Subsequent searches of other DNA databases turned up more than a thousand human pairs with genetic matches at nine loci or more**.

Joint probability

- The probability of two independent events' *both* happening is the product of their respective probabilities.
- In other words, the probability of Event A happening *and* Event B happening is the probability of Event A multiplied by the probability of Event B.
- This formula is applicable only if the events are independent, meaning that the outcome of one has no effect on the outcome of another.
- For example, the probability that you throw heads on the first flip does not change the likelihood of your throwing heads on the second flip.

outcome A *or* outcome B

- one event happens *or* another event happens: outcome A *or* outcome B (again assuming that they are independent).
- In this case, the probability of getting A or B consists of the sum of their individual probabilities: the probability of A *plus* the probability of B.

expected value

- The expected value or payoff from some event, say purchasing a lottery ticket, is the sum of all the different outcomes, each weighted by its probability and payoff.
- Suppose you are invited to play a game in which you roll a single die.
- The payoff to this game is $1 if you roll a 1; $2 if you roll a 2; $3 if you roll a 3; and so on.
- What is the expected value for a single roll of the die?
- Each possible outcome has a 1/6 probability, so the expected value is:

$$\tfrac{1}{6}(\$1) + \tfrac{1}{6}(\$2) + \tfrac{1}{6}(\$3) + \tfrac{1}{6}(\$4) + \tfrac{1}{6}(\$5) + \tfrac{1}{6}(\$6) = \tfrac{21}{6}, \text{ or } \$3.50.$$

expected value

- In fact, the expected value turns out to be extremely powerful because it can tell you whether a particular event is “fair,” given its price and expected outcome.
- Suppose you have the chance to play the above game for $3 a throw.
- Does it make sense to play? Yes, because the expected value of the outcome ($3.50) is higher than the cost of playing ($3.00).
- This does not guarantee that you will make money by playing once, but it does help clarify which risks are worth taking and which are not.

Do not buy lottery tickets

- In Illinois, the probabilities associated with the various possible payoffs for the game are printed on the back of each ticket.
- I purchased a $1 instant ticket.
- On the back—in tiny, tiny print—are the chances of winning different cash prizes, or a free new ticket:
- 1 in 10 (free ticket);
- 1 in 15 ($2);
- 1 in 42.86 ($4);
- 1 in 75 ($5); and so on up to the
- 1 in 40,000 chance of winning $1,000.

crucial lessons

- I calculated the expected payout for my instant ticket by adding up each possible cash prize weighted by its probability.
- It turns out that my $1 lottery ticket has an expected payout of roughly $0.56, making it an absolutely miserable way to spend $1.
- As luck would have it, I won $2.
- This is one of the crucial lessons of probability.
  - Good decisions can turn out badly
  - Bad decisions can still turn out well, at least in the short run
- But probability wins in the end.

The law of large numbers

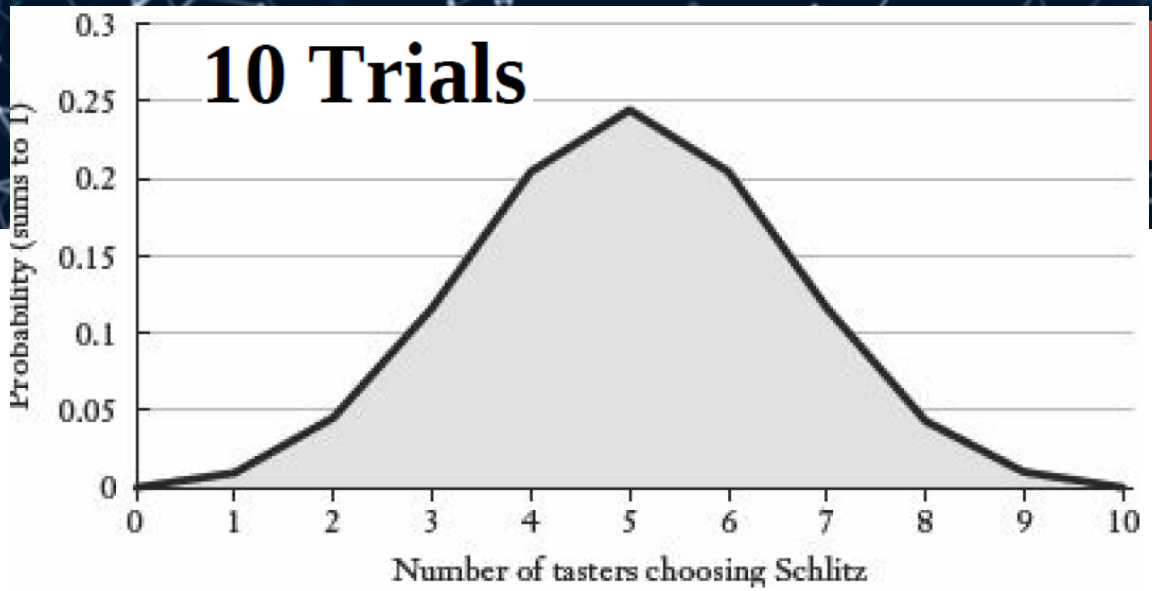
- An important theorem known as the law of large numbers tells us that as the number of trials increases, the average of the outcomes will get closer and closer to its expected value.
- Yes, I won $2 playing the lotto today.
- And I might win $2 again tomorrow.
- But if I buy thousands of $1 lottery tickets, each with an expected payout of $.56, then it becomes a near mathematical certainty that I will lose money.
- By the time I've spent $1 million on tickets, I'm going to end up with something strikingly close to $560,000.

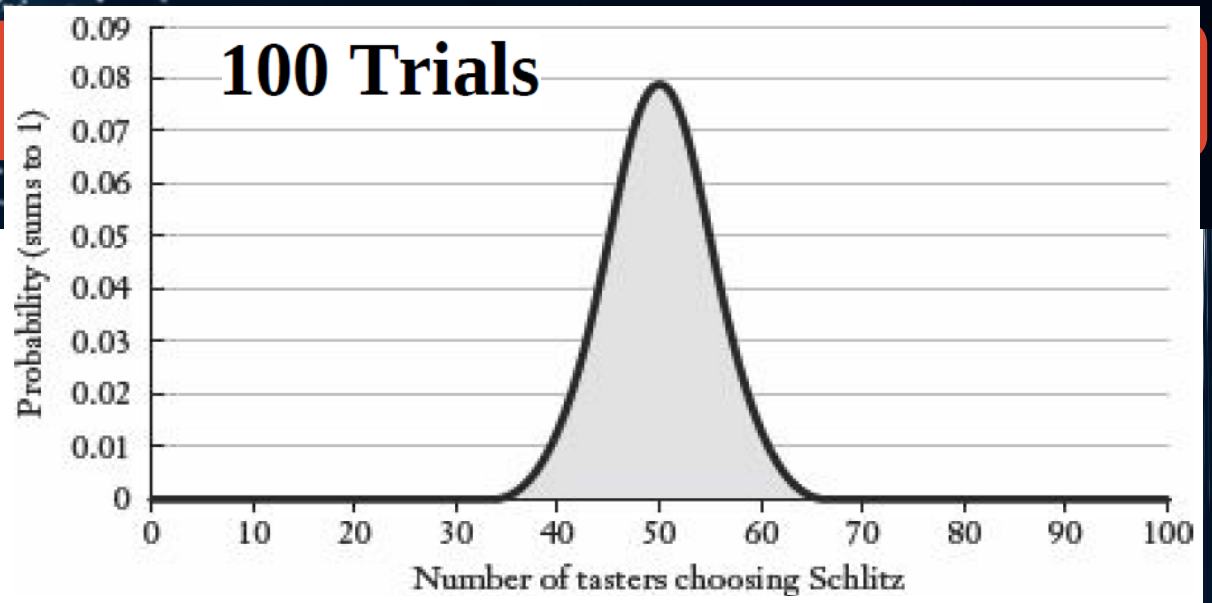
The law of large numbers

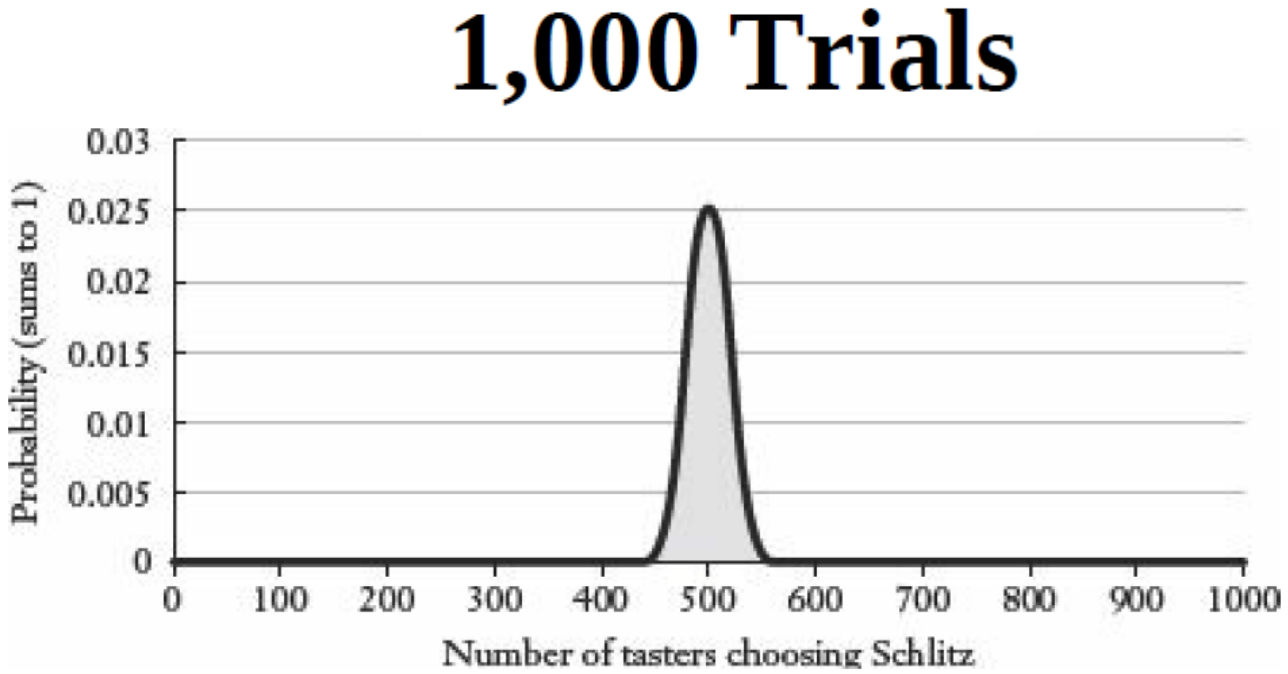
- The law of large numbers explains why casinos always make money in the long run.
- The probabilities associated with all casino games favor the house.
- If enough bets are wagered (ventured) over a long enough time, the casino will be certain to win more than it loses.
- The law of large numbers also demonstrates why Schlitz was much better off doing 100 blind taste tests at halftime of the Super Bowl rather than just 10.
- Check out the “probability density functions” for a Schlitz type of test with 10, 100, and 1,000 trials.

Probability density function

- A PDF simply plots the different outcomes along the x-axis and the expected probability of each outcome on the y-axis;
- the weighted probabilities—each outcome multiplied by its expected frequency—will add up to 1.)
- Again I'm assuming that the taste test is just like a coin flip and each tester has a 0.5 probability of choosing Schlitz.
- As you can see below, the expected outcome converges around 50 % of tasters' choosing Schlitz as the number of tasters gets larger.
- At the same time, the probability of getting an outcome that deviates sharply from 50 % falls sharply as the number of trials gets large.


10 Trials
Probability (sums to 1)
0.3
0.25
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8 9 10
Number of tasters choosing Schlitz


100 Trials
Probability (sums to 1)
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0 10 20 30 40 50 60 70 80 90 100
Number of tasters choosing Schlitz


1,000 Trials
Probability (sums to 1)
0.03
0.025
0.02
0.015
0.01
0.005
0
0 100 200 300 400 500 600 700 800 900 1000
Number of tasters choosing Schlitz

Probability density function

- I specified earlier that Schlitz executives would be happy if 40 % or more of the Michelob drinkers chose Schlitz in the blind taste test.
- The figures below reflect the probability of getting that outcome as the number of tasters gets larger:

10 blind taste testers: 0.83

100 blind taste testers: 0.98

1,000 blind taste testers: 0.9999999999

1,000,000 blind taste testers: 1

Bayes' formula

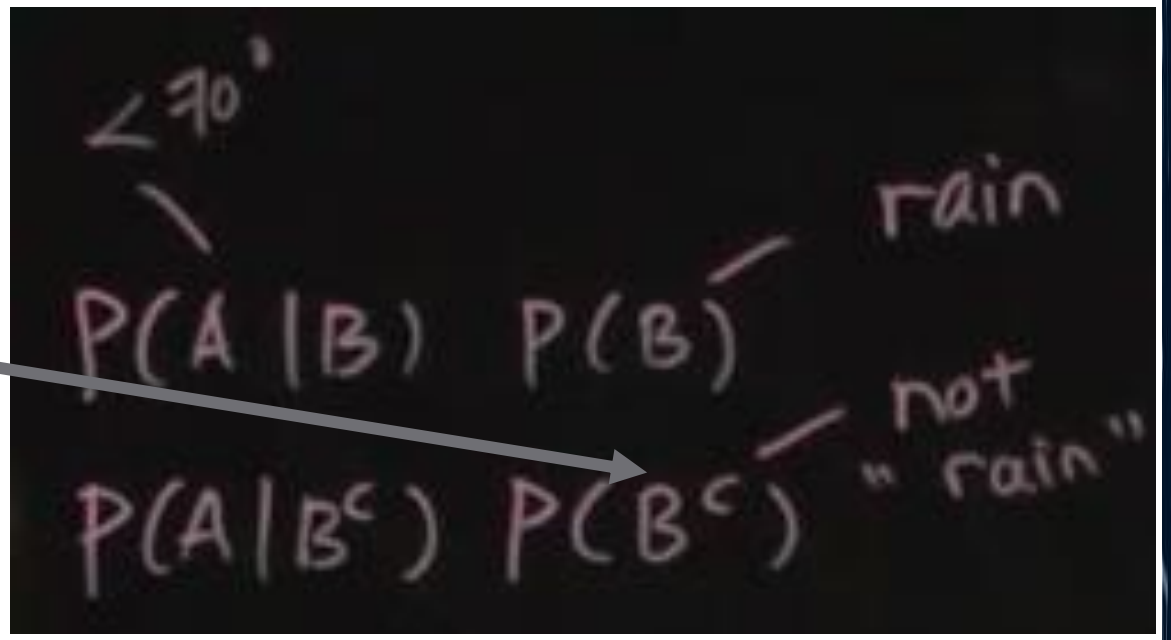
- Another interesting probability result concerning conditional probability is Bayes' formula.
- Bayes' formula takes advantage of the conditional probability of one event given the outcome of another event, to do a reverse probability calculation.
- It works like this.
- you have probability of A given B, **P(A|B)**
- and also you have probability of B. **P(B)**
- At the same time, you have probability of B complement, **P(Bc)**

Bayes' formula

- Remember this is the not B event, is exactly the complement of event B.
- So whatever outcomes are in B will not be in this is not B event.
- The not B event contains all the outcomes that are not contained in the definition of B.
- So you also have A given B complement. **P(A|Bc)**

Bayes' formula

- So this will be similar as saying, let's consider B being whether it rained or not.
- Then this will be if it didn't rain, not rain event.
- And A will be, if it was below 70 degrees for the temperature, and for the same true event.

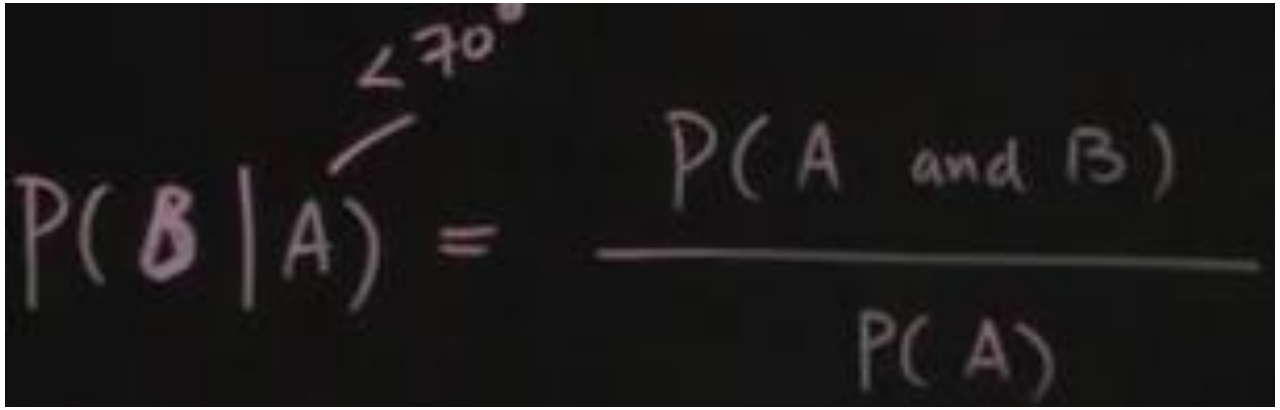


Bayes' formula

- So using Bayes' formula, what we can get is the probability of B given A.
- From the past historical data on weather, I know how many days it rained, and I know how many days it didn't rain.
- For the rainy days, I know the temperature from the past.
- That's how many times when it rained the temperature dropped below 70 degrees,
- and how many times when it didn't rain that it also dropped under 70 degrees.

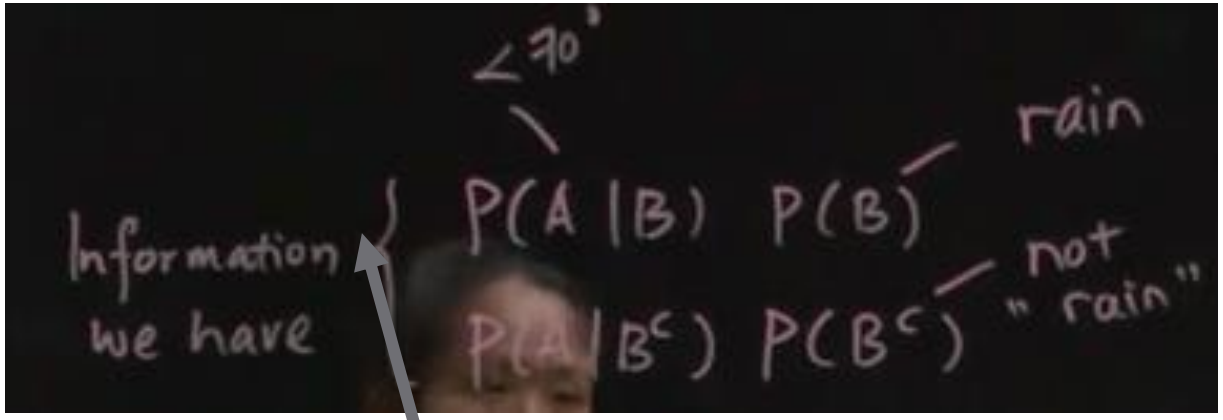
Bayes' formula

- So now, assuming that for a particular day, we observed the temperature-- say we're having a meeting in a room where we have no window to observe outside.
- But from our cell phones, we know the current temperature is very low.
- knowing a low temperature, for a particular period of the year-- what is the conditional probability of raining outside?
- So this, by definition is the probability of A and B divided by probability of A.

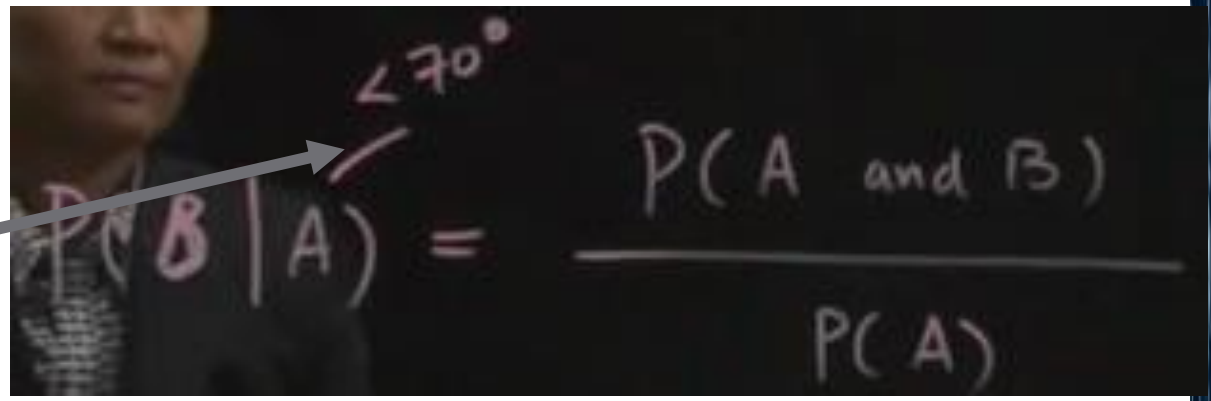


Bayes' formula

- And you can see that from this knowledge-- so this part is our resources, this is the information we have.

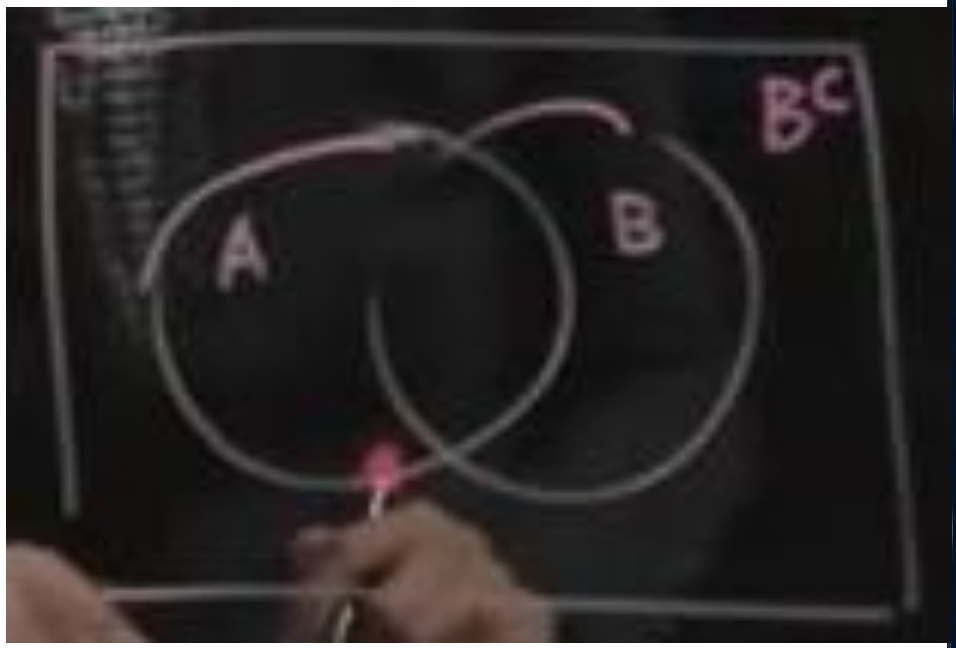


- How do we use this information we have to derive the two ingredients that are required to derive this conditional probability?



Bayes' formula

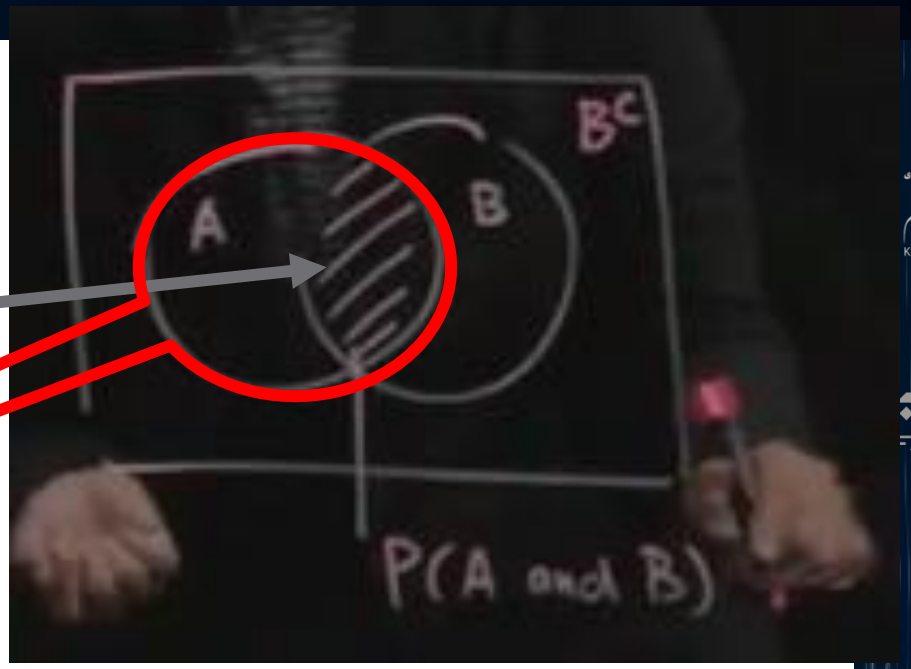
- So here, I will draw the Venn diagram again.
- Here you have A and you have B.
- One thing that we didn't discuss previously, is when you have B, the diagram is indicating that the rectangular area is the sample space, the circular area for B contains all the outcome belonging to the inside.

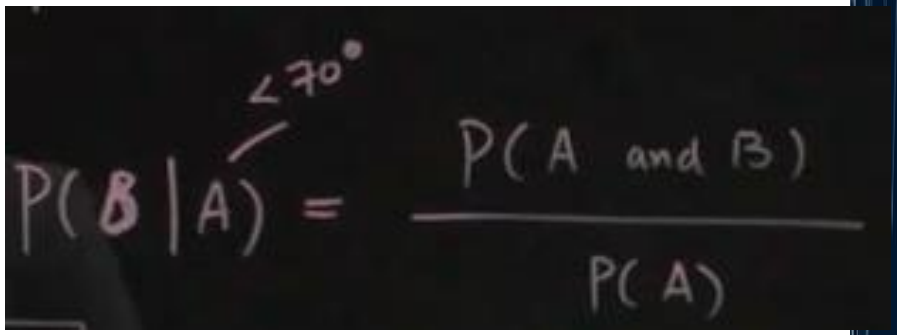


- So therefore on the outside is the not B event.

Bayes' formula

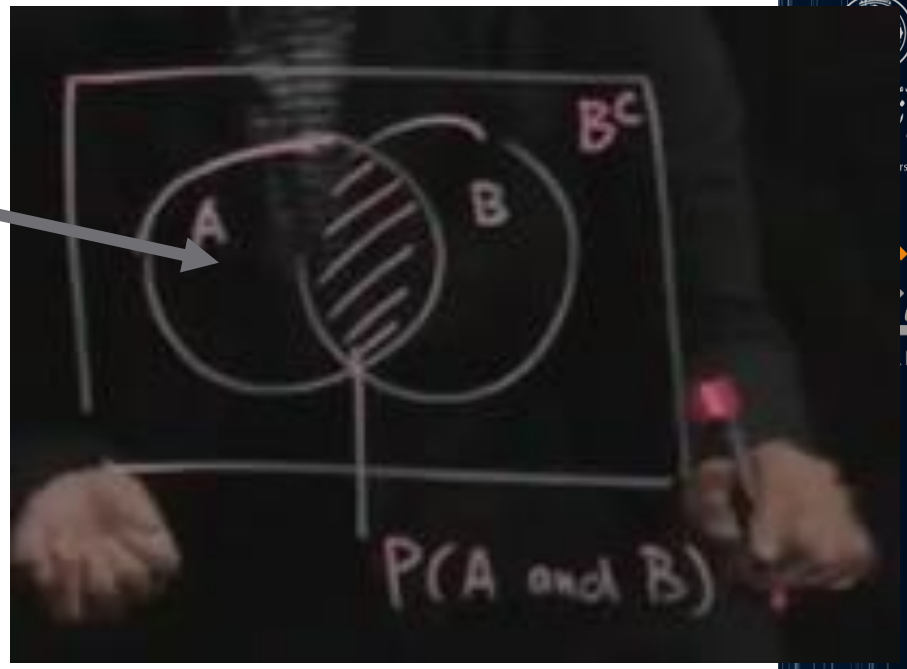
- So the B complement is the complement area to B within this rectangular area.
- So let's think about this area here.
- So this is the probability of A and B.
- So this is where we need for the numerator.
- For the denominator, we need the total area of the shape.





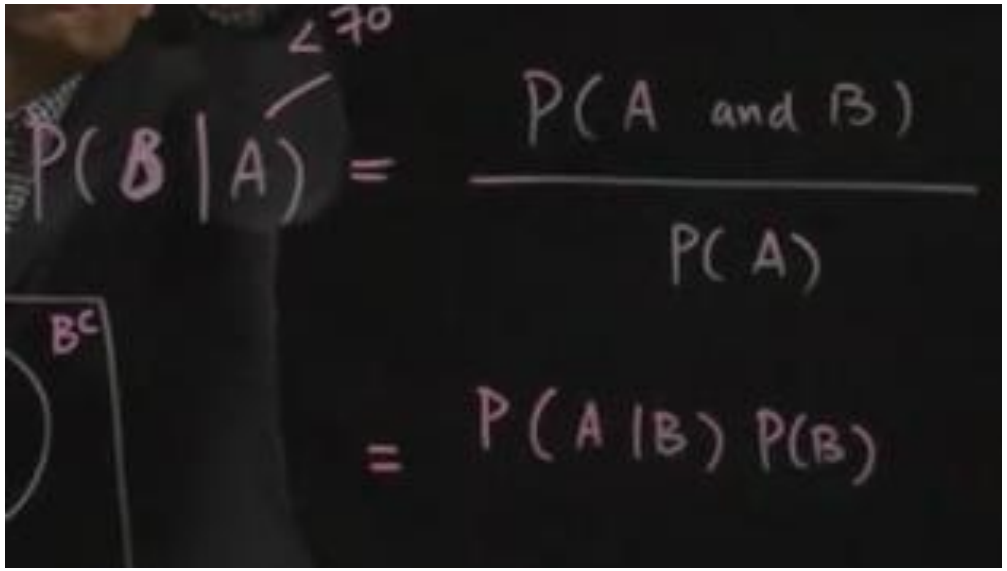
Bayes' formula

- So we know part of it already, back here.
- How do we derive this part?
- This part, it's also intersection.
- So this is A and the not B event. OK?
- It's the intersection of this circular area corresponding to A, with the B complement, the area outside B.
- So, here I use these dots to indicate this area.



Bayes' formula

- So, to get this quantity, we need to know the probability of A, and the probability of A and B.
- The numerator is really easy using this.
- So I will give you probably A given B, times probability of B, using directly the information already here.

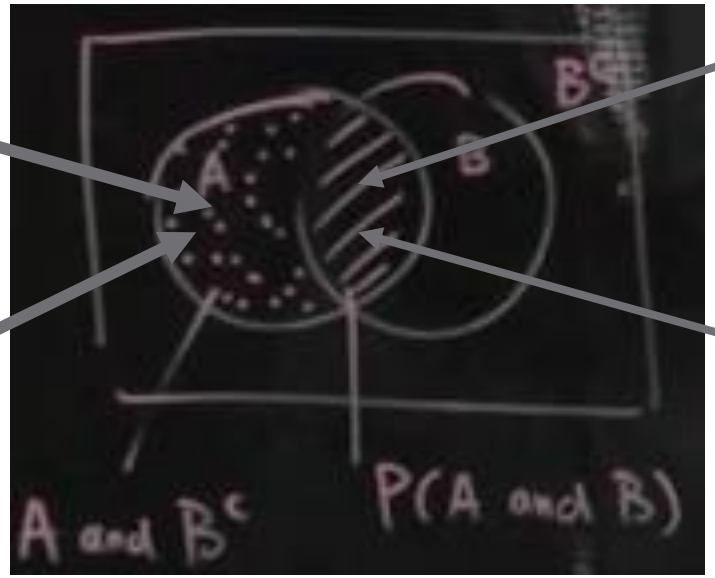


Bayes' formula

- So for the denominator, we need both the size of this area and this area.

$$P(A|B^c) = \frac{P(A \text{ and } B^c)}{P(B^c)}$$

$$P(A \text{ and } B^c) = P(A|B^c)\ P(B^c)$$



$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(A \text{ and } B) = P(A|B)\ P(B)$$

- The information we have provides us the way of probability of A given B, times probability of B, in this area.
- And probability of A given B complement, times the probability of B complement, the other area.

Bayes' formula

- Even though we do not have the information to give us directly what the total size of A, the information we have gives the probability of A in parts through conditional probabilities.
- So combining the two parts together, we gather the total probability of A.

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{P(A|B)\,P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

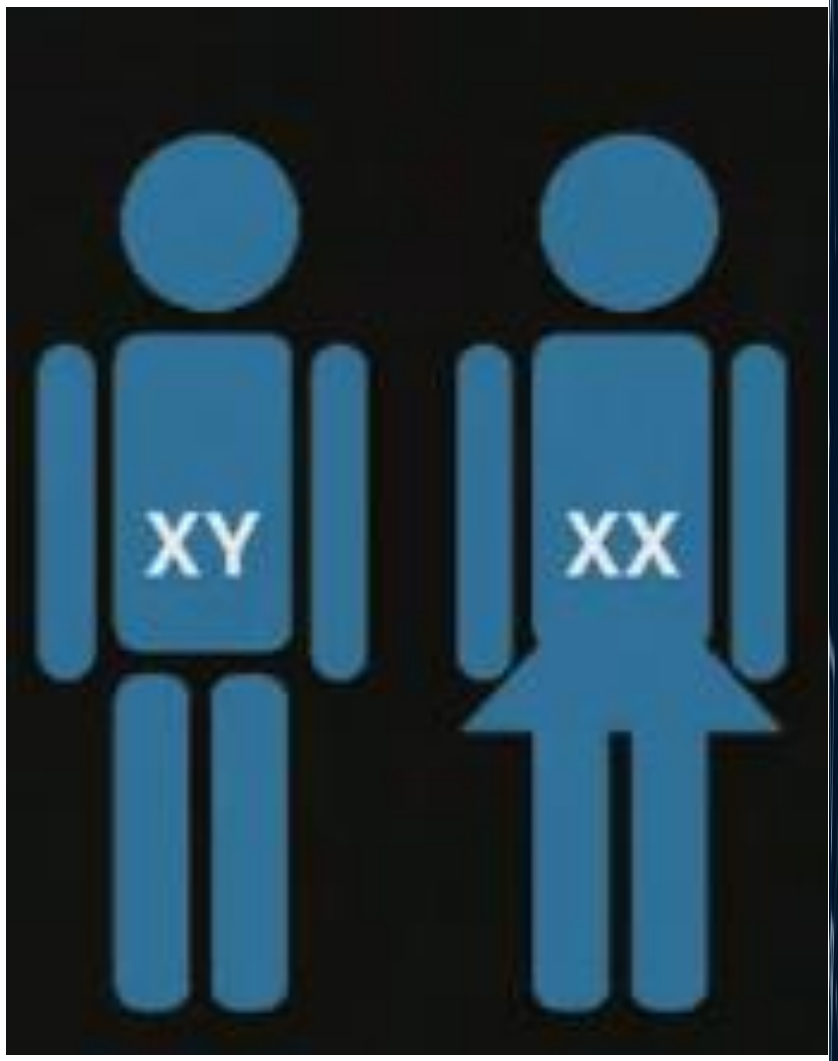
Bayes' formula

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{P(A|B)\,P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

- So this formula is called Bayes' formula.
- It takes advantage of information on A, condition on B, and the marginal information or the individual information of event B.
- And do a reverse calculation for us to derive what is the information about B we can derive when we have knowledge about A.

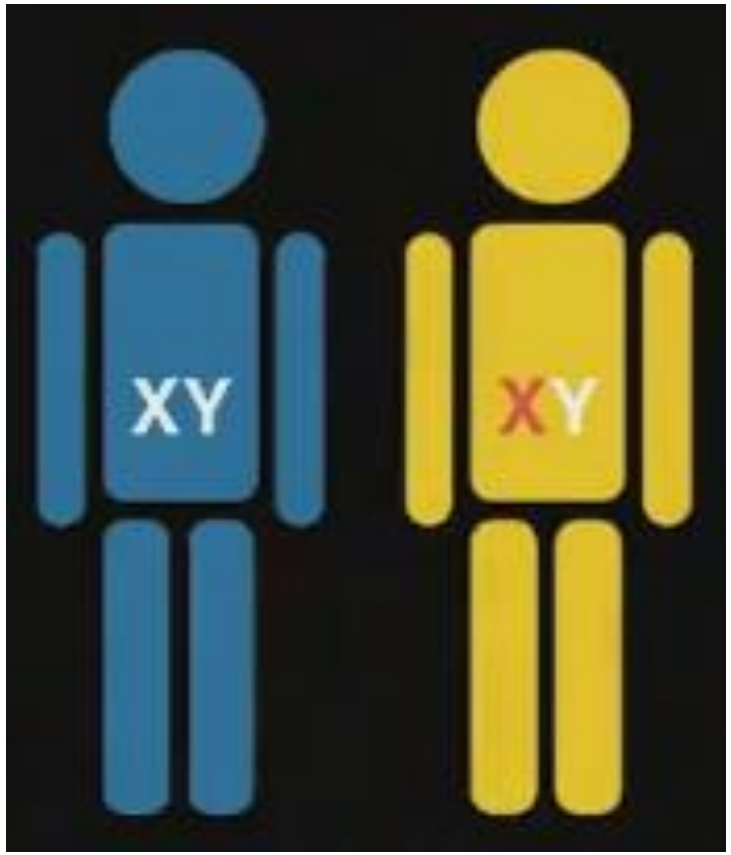
Hemophilia

- Let's now consider another probability example using the disease Hemophilia. This is a hypothetical example.
- Hemophilia is a genetic disease that can pass from one generation to the next, and it is known to be x linked.
- The human males have an x chromosome and a y chromosome, and the females have two x chromosomes.
- This gene is x linked.



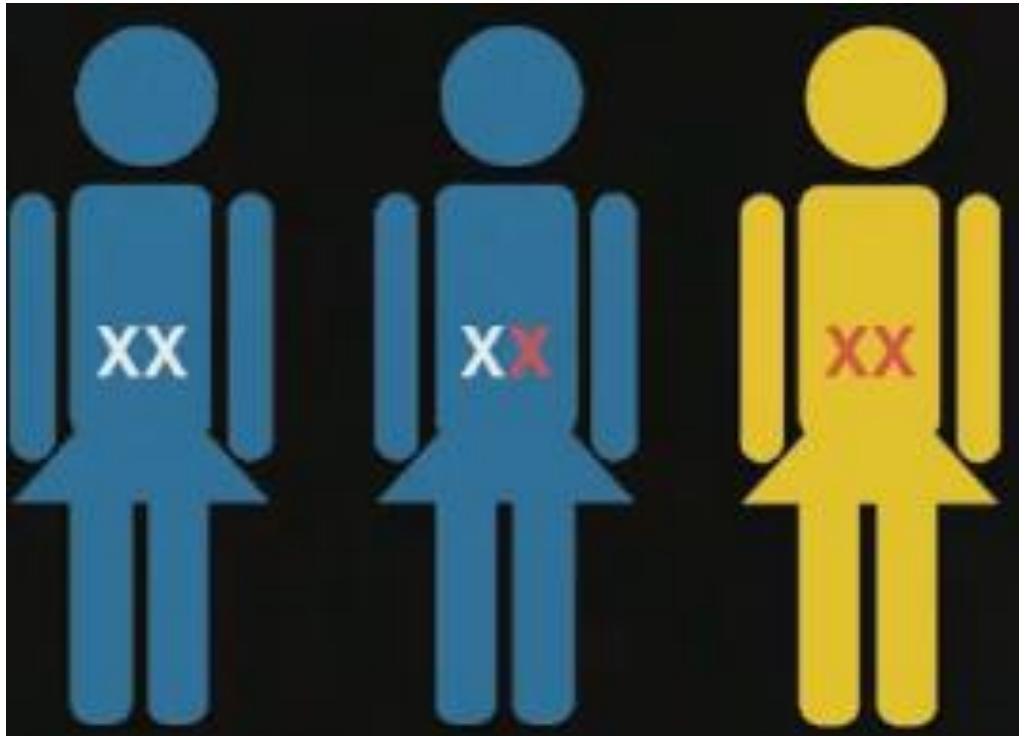
Hemophilia

- So, for the male population, if you have a normal x chromosome, then you will not be affected.
- If you have x chromosome carrying the disease variant of this gene, you will be affected.



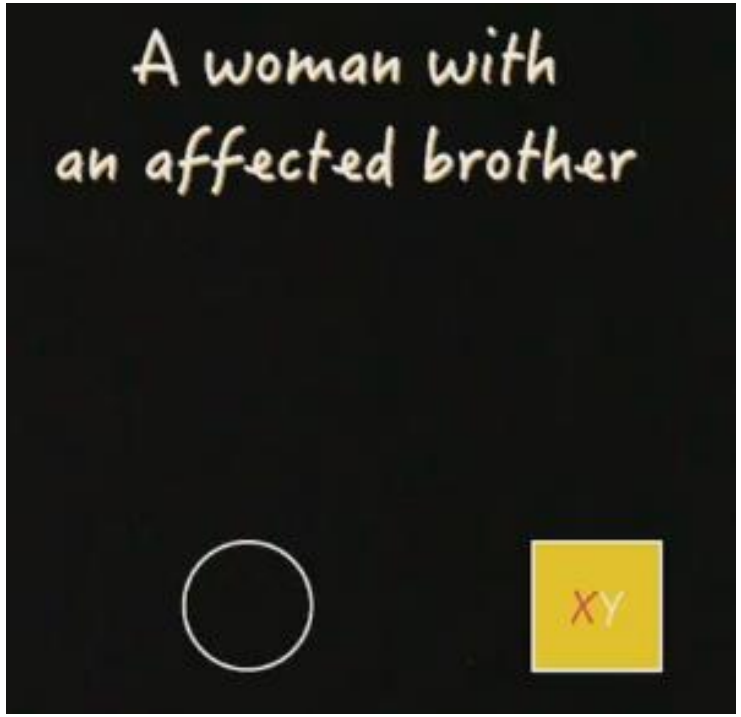
Hemophilia

- In the females, because you have two copies, if you have two normal copies of the gene, you will not be affected.
- If you only have one copy of normal genes and one copy of the disease gene, you also will not be affected.
- It is only when you have two copies of the disease variant of the gene that you will be affected by the disease.
- So this sets up the background for our example.



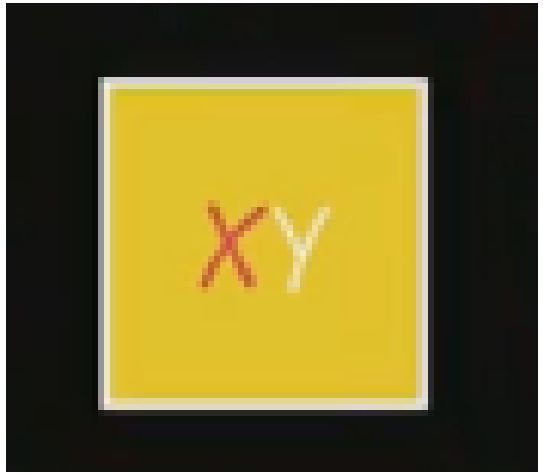
Hemophilia

- So in this example, let's consider a woman with an affected brother.
- So we use circles to indicate a female, and a square to indicate male.



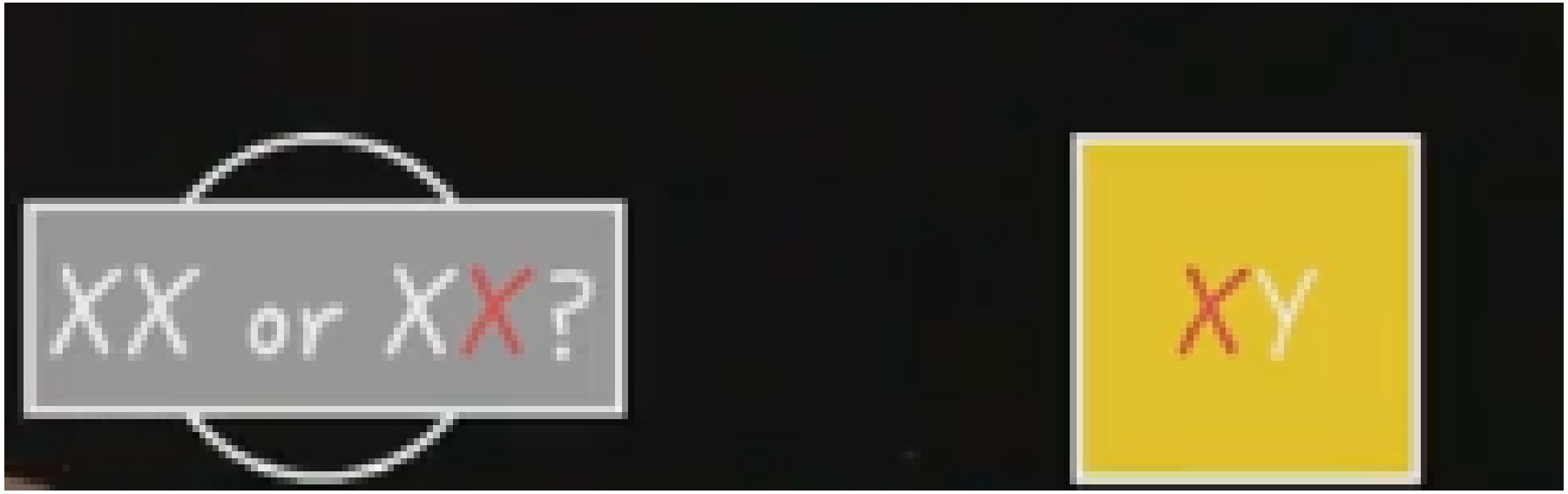
Hemophilia

- And here for this woman, that we know that has a brother that is affected.
- Because for a male subject, in order to get affected you only need one copy of the x chromosome.
- The one and only x chromosome is carrying the affected gene.



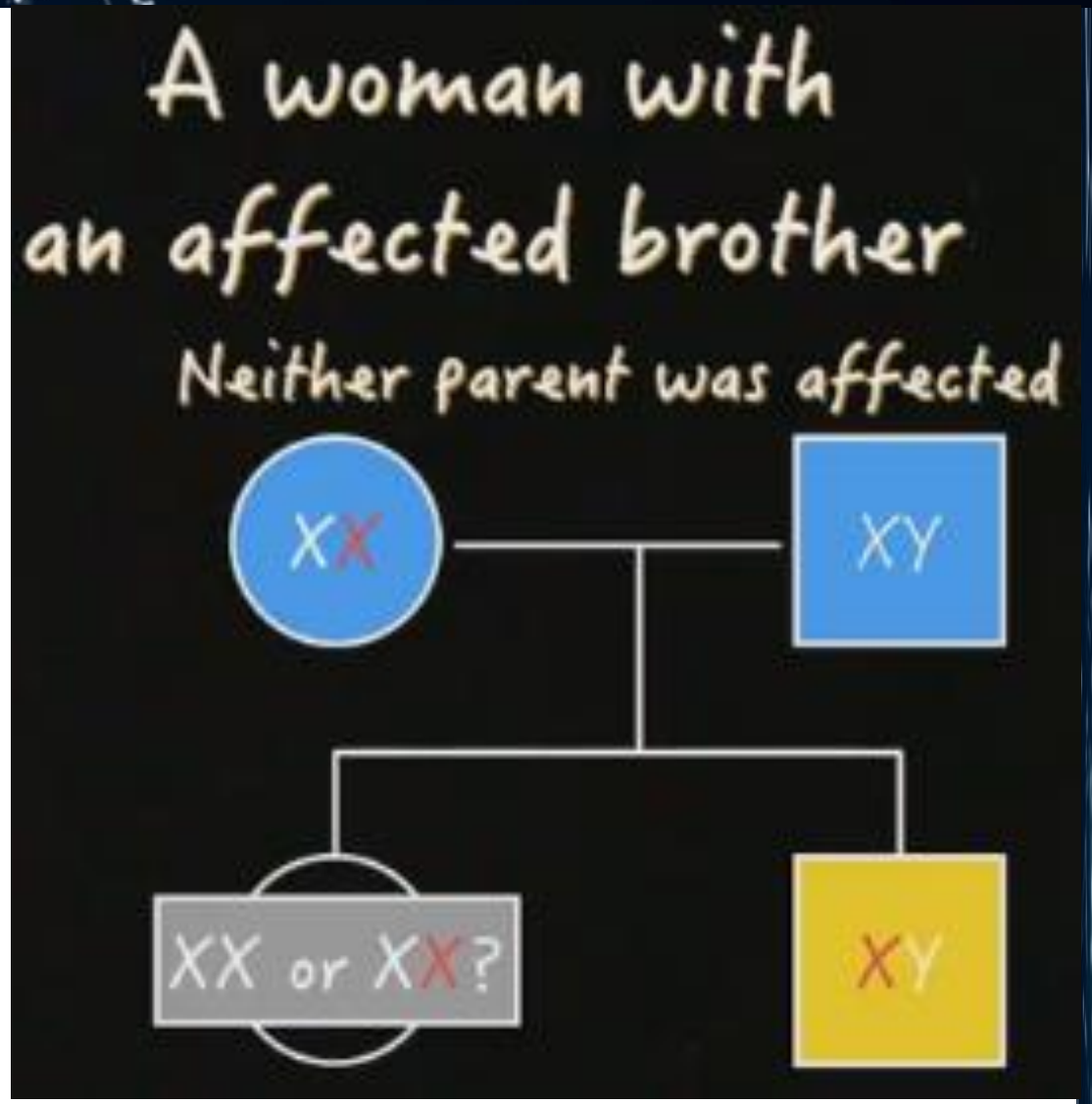
Hemophilia

- So then the question is, for this woman what is her possible genetic information on the x chromosome?
- Does she have two normal copies, or one normal copy and one affected copy?
- And then, we also know neither parent was affected.



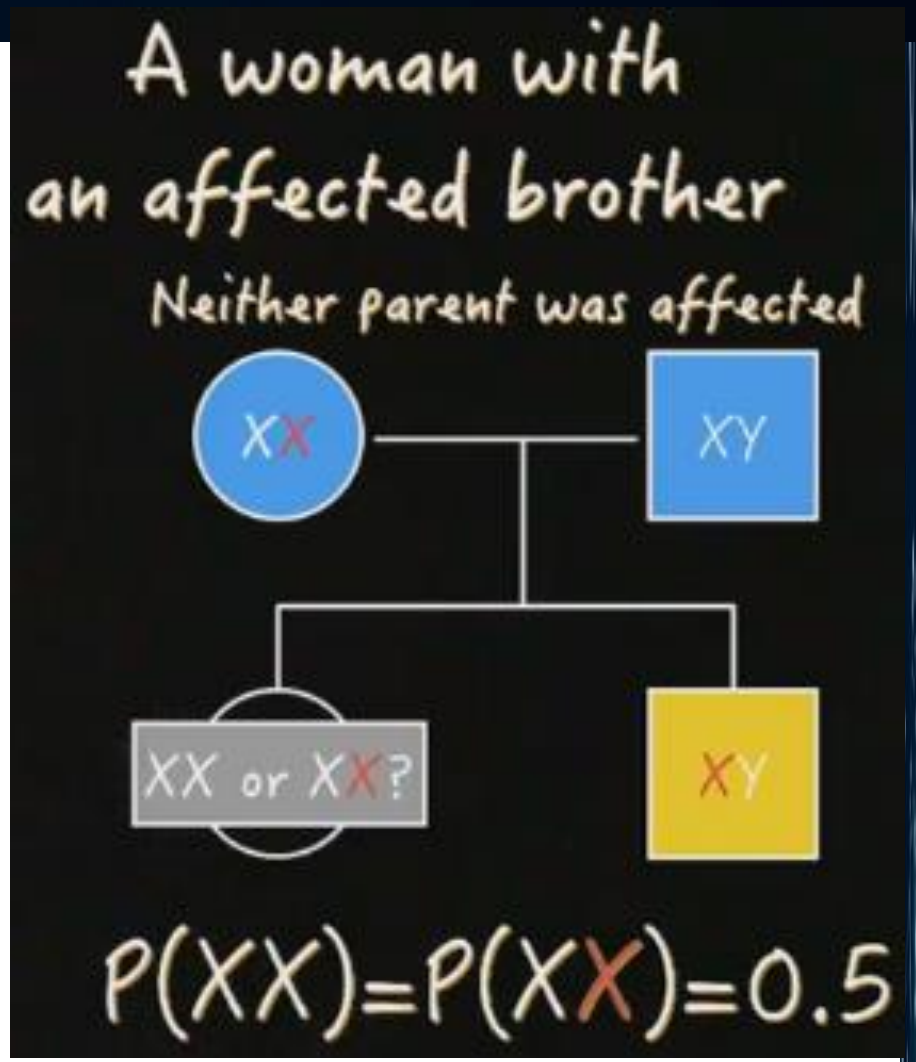
Hemophilia

- So because the woman has a brother who has an affected copy, then we know that one of the parents has to have an affected copy.
- And because the father was not affected, we know that the father has a normal copy of x and y.
- Because the mother was not affected, then she must have a normal copy of x in addition to an affected copy of x, because she has an affected son.



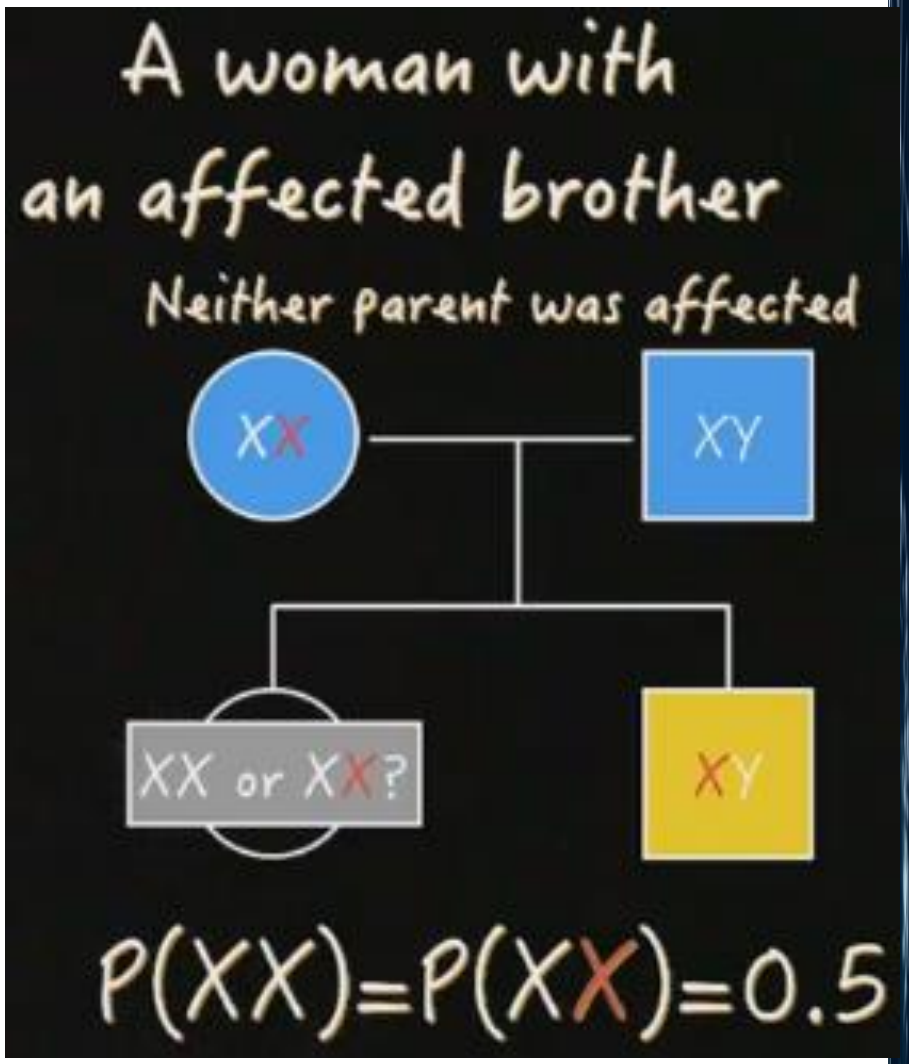
Hemophilia

- So because of these, and in genetics there's a random transmission of information from parent to their children.
- And this woman, she could have received a red copy from her mother, or a normal copy from her mother.
- And the only copy of x she can receive from her father is a normal copy.



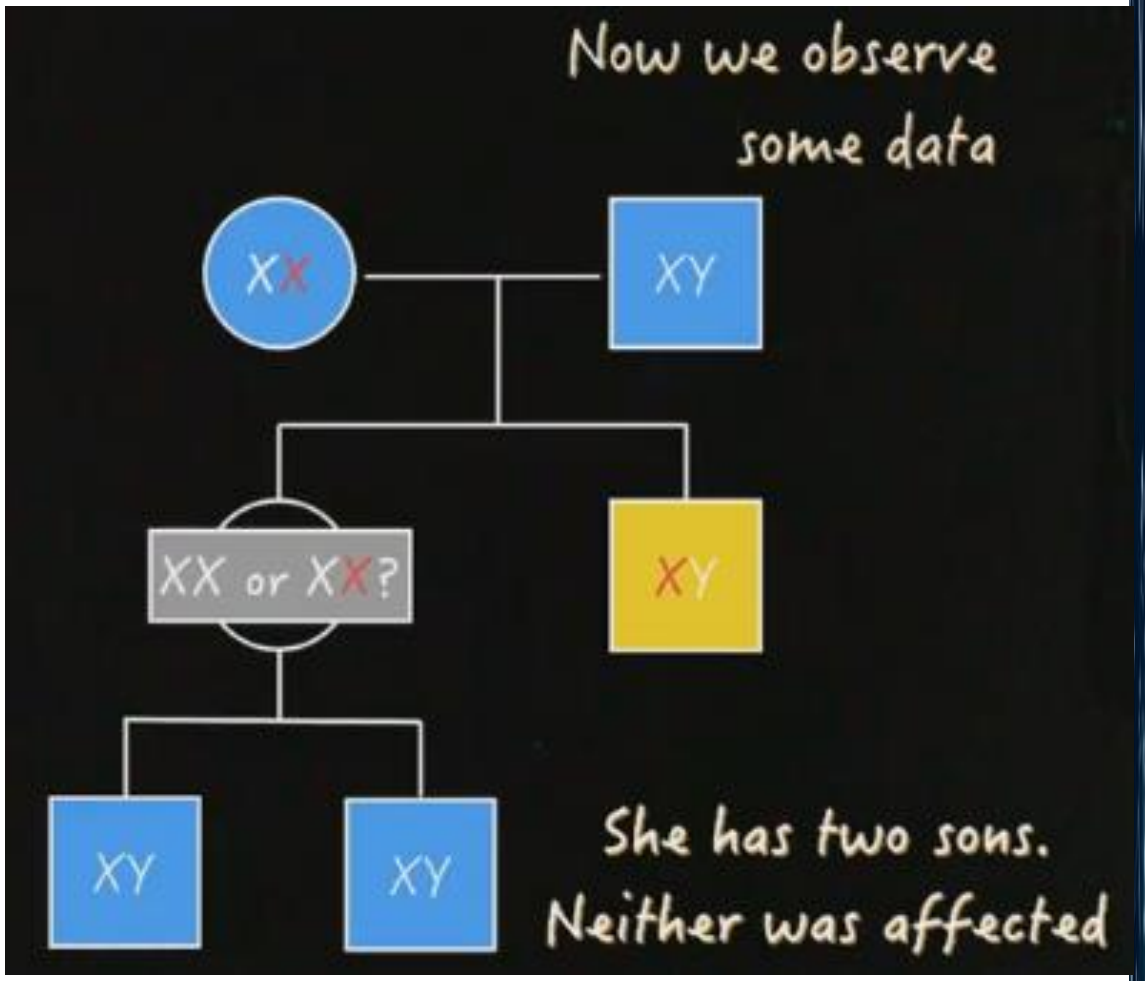
Hemophilia

- So her mother can transmit the normal copy or the affected the copy, the red one, equally likely.
- So we know that the probability for this woman in question, the probability of two normal copies or the probability of one normal, one red, is half and half.
- It depends on which copy her mother transmitted to her.



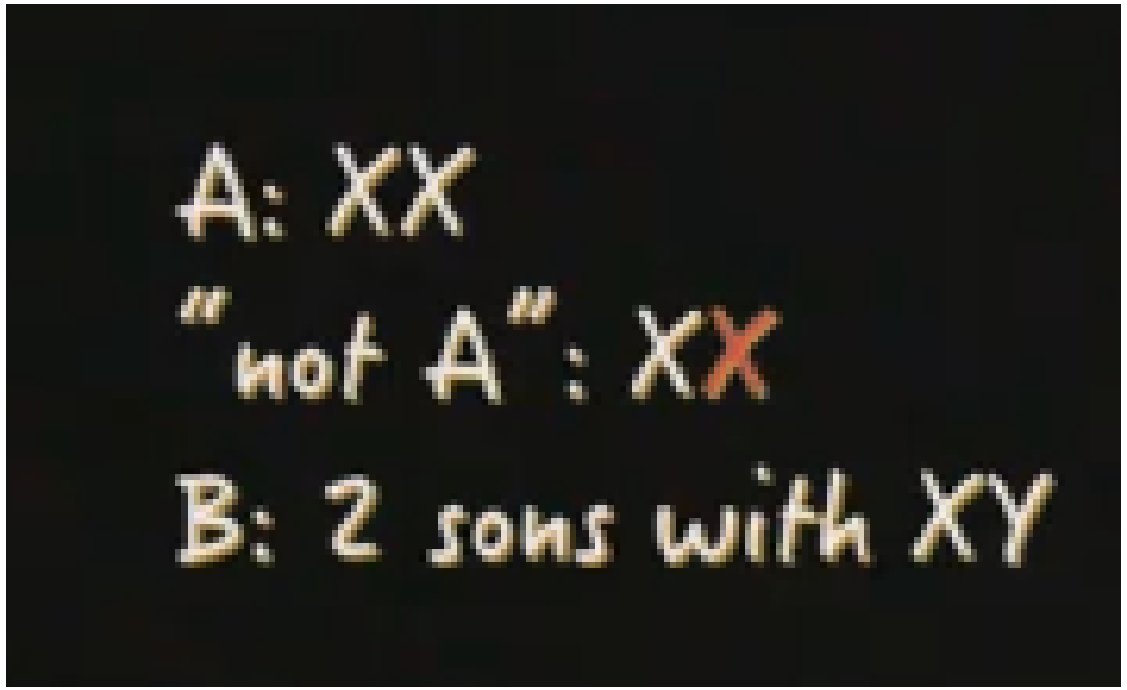
Hemophilia

- So now let's observe some data, more information to carry out the probability.
- This woman in question has two sons.
- Neither was affected.
- So we know that this woman can transmit a copy of x to her sons, and neither of them were affected.



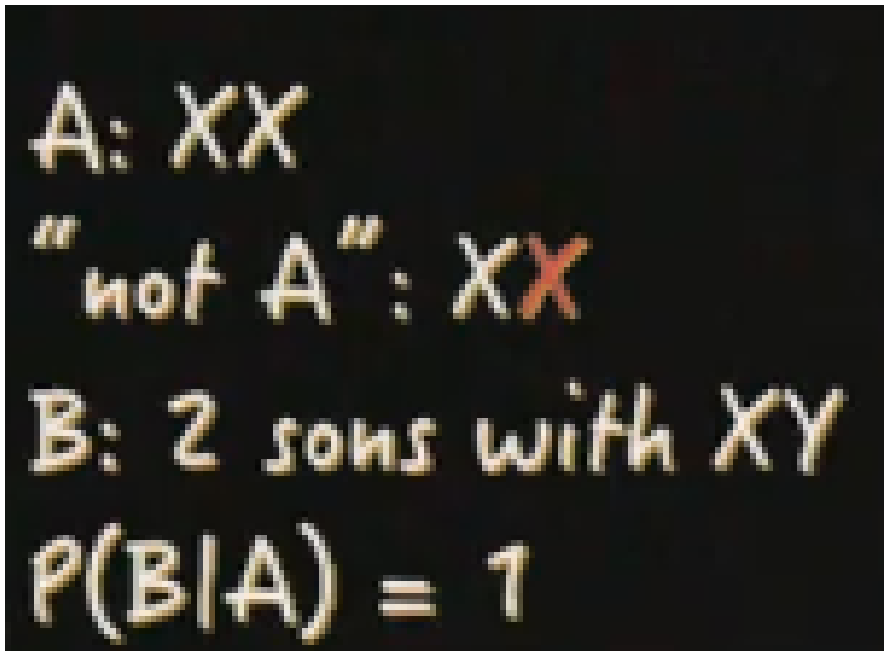
Hemophilia

- So event A is this woman has two normal copies of the chromosome.
- And not A event is she has one normal copy and one disease copy.
- Event B would be, both her sons have a normal copy of x and the y.



Hemophilia

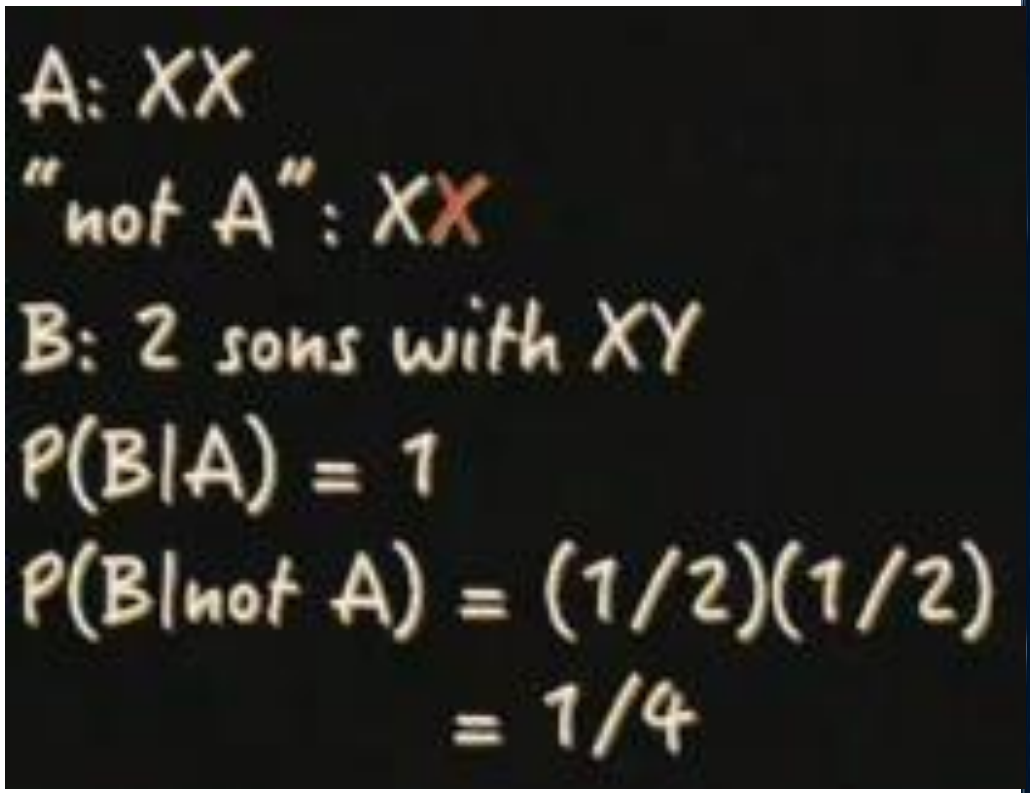
- So because she can transmit either copy of her chromosome equally likely, if she has no copy of the disease chromosome(XX), the probability for her to have two unaffected boys is 100%.



- Because she has no disease gene to transmit.

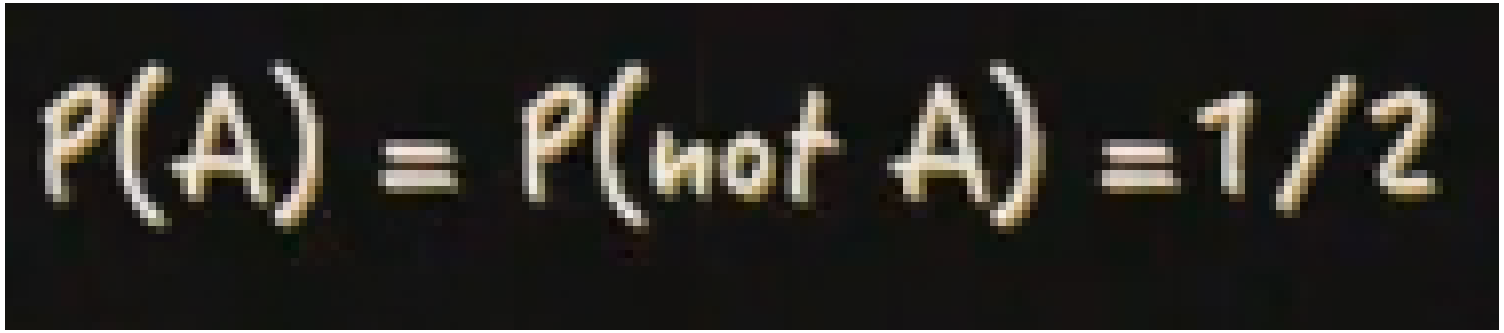
Hemophilia

- But if she is actually-- in a not A event-- if she actually has an affected gene, that she has a 50% chance to transmit to her son.
- Then, we have 1/2 chance for the first son to be unaffected, and 1/2 for the second son to be affected.
- So this is the conditional probability.



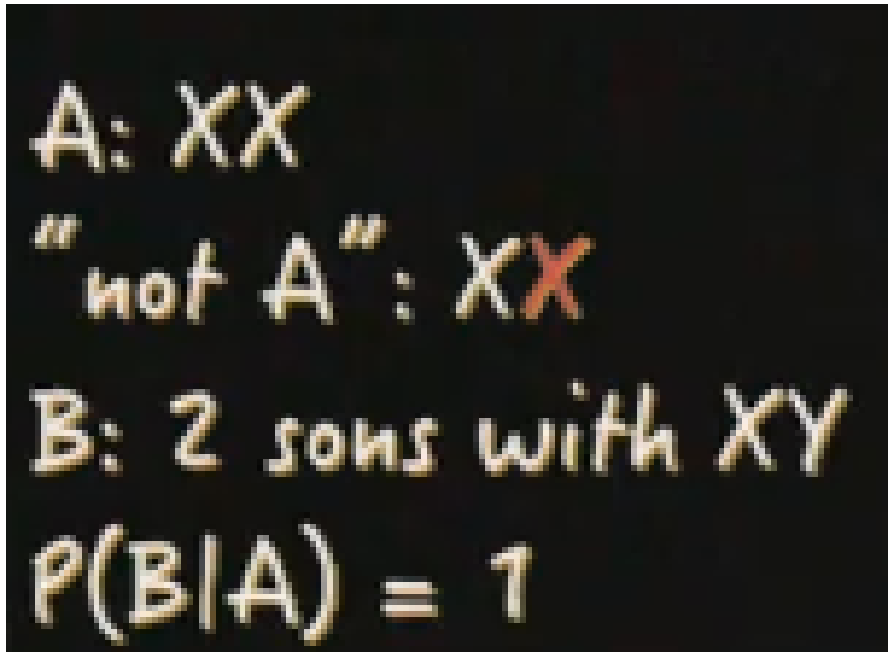
Hemophilia

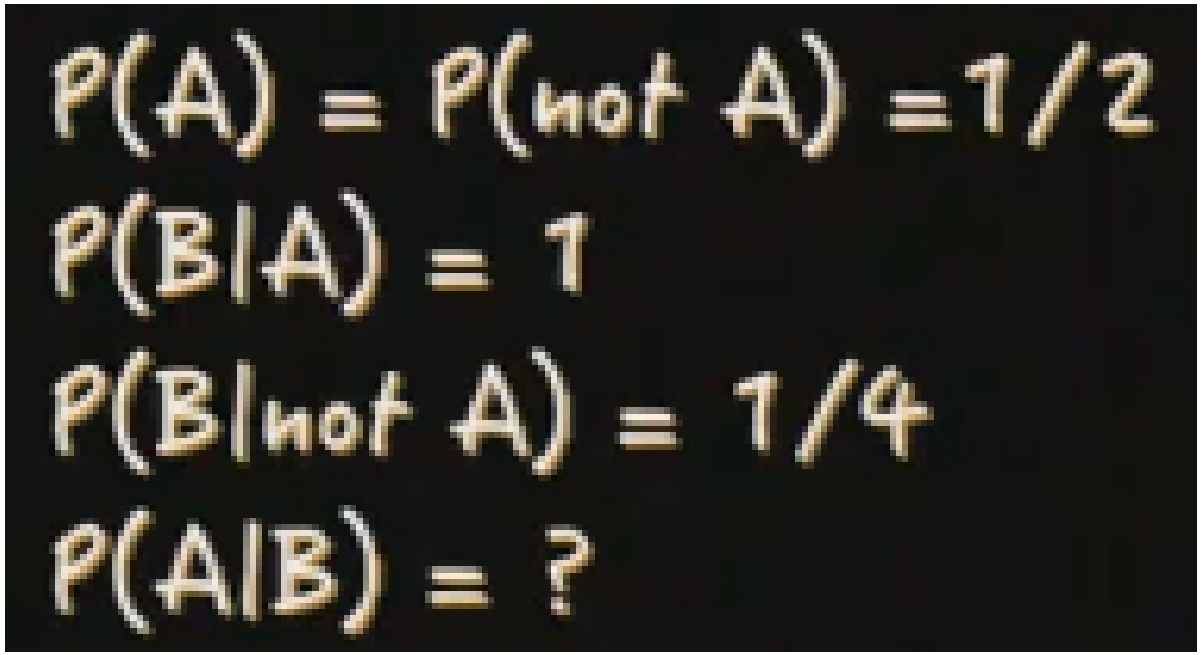
- Now we want to use Bayes' formula to figure out, given the information we have about her children, can we now have an updated evaluation of her chance of having either two normal copies or one normal copy and one disease copy?
- So we know that, without any knowledge of her children, the probability of A and the probability of not A is 1/2 and 1/2, from one of the previous slides.



Hemophilia

- If from A, the probability of having two unaffected sons is 1.
- If from not A, the probability of two unaffected sons is 1/4.





Hemophilia

- Putting all this together, using Bayes' formula, we now want to find out what is the probability of A given B.
- And using the Bayes' formula-- I am just going to copy down the Bayes' formula.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(Ac)}$$

- A we define as she has two normal copies of the chromosome.
- And not A is when she has the disease copy of x chromosome.

Hemophilia

- And here, we know that this part is 1/2, and this part is 1.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(Ac)} = \frac{1/2}{1/2 + 1/8} = \frac{4}{5}$$

1 1/2 1 1/2 1/4 1/2

Hemophilia

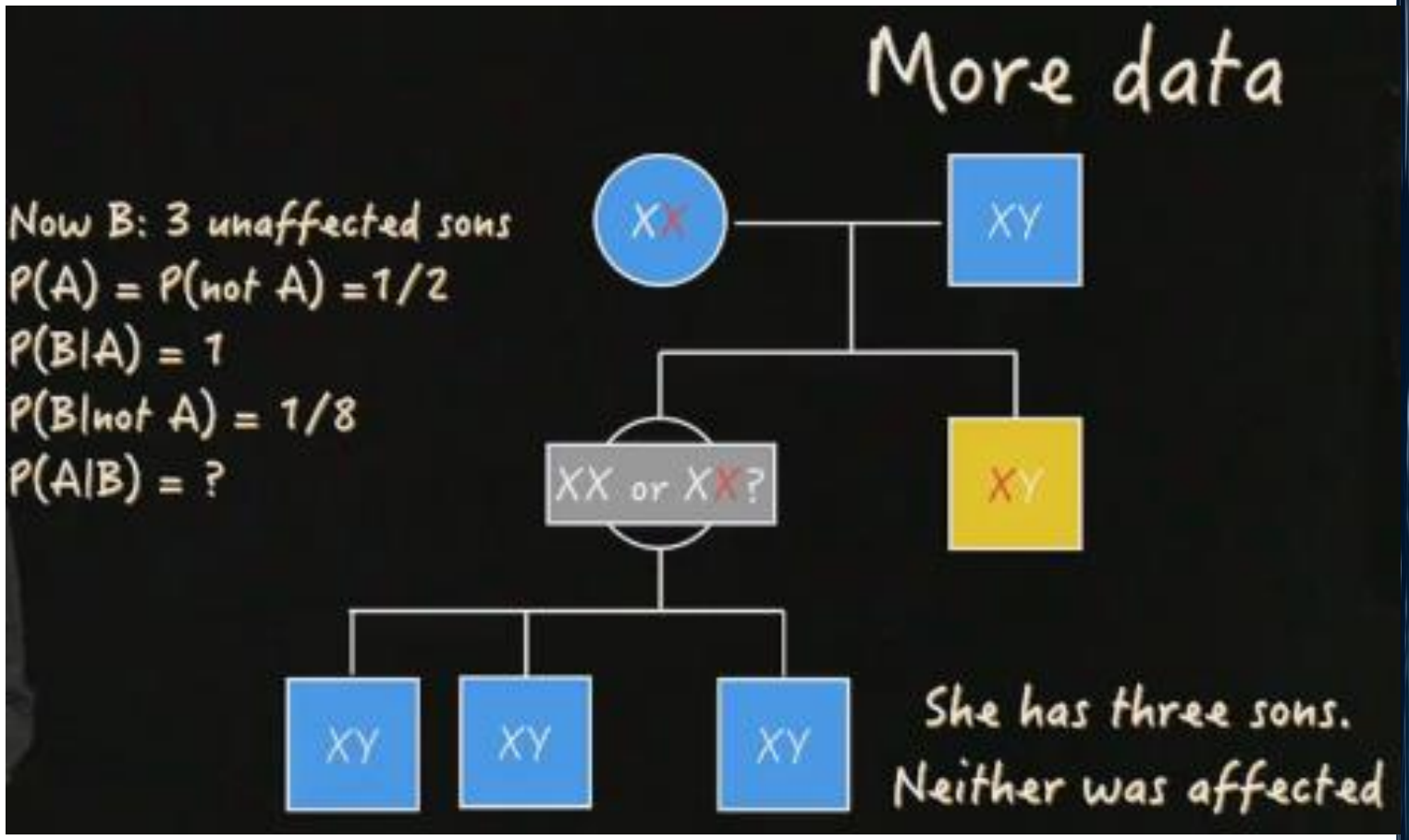
- before any knowledge about her children or her sons, we have the information is about half and half.
- We're not sure whether she has the disease chromosome or not.
- After observing that she has two sons, neither were affected,
- then we carry out the Bayes' formula to look at, conditioned on the information we have about the two boys she has,
- what is the chance that she actually has two normal copies of the gene?
- And then you can see after this computation, it has become for 4/5, which is 80%.

Hemophilia

- And this is a higher number than 1/2, previously we had.
- So meaning that, she's more likely to have two normal copies rather than one normal copy and one disease copy.
- And the reason for that, is the event we observed the disease status of her two sons provide additional information on what genetic information she carries.
- So this is how probability theory helps us evaluate the likeliness of a certain possibility in your data.

Hemophilia

- So now, we consider more data.
- Let's say, now we have additional observations taken.
- She actually has three sons.
- None of them are affected.



Hemophilia

- So here, we still have the same information without any data observed on her children.
- And we have this information, that if she had normal copies, and of course none of her children would be affected.
- And for, not A event, the probability for her to have three normal sons becomes 1/8.
- So, if this is the case, now as an exercise, can you figure out what would be the probability of A given B?

Hemophilia

- And here, we know that this part is 1/2, and this part is 1.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(Ac)} = \frac{\frac{1}{2}}{\frac{1}{2} + \frac{1}{16}} = \frac{\frac{1}{2}}{\frac{9}{16}} = \frac{8}{9} = 0.88$$

1 1/2 1 1/2 1/8 1/2

A warranty is just a form of insurance

- The entire insurance industry is built on probability.
- When you insure anything, you are contracting to receive some specified payoff in the event of a clearly defined unforeseen event.
- The key idea is that in exchange for a regular and predictable payment, you have transferred to the insurance company the risk of having your car stolen, crushed, or even totaled by your own bad driving.
- Why are these companies willing to assume such risks? Because they will earn large profits in the long run if they price their premiums correctly.
- Obviously some cars insured will get stolen.
- Others will get totaled when their owners drive over a fire hydrant,…

"expected loss"

- But most cars insured will be just fine. To make money, the insurance company **need only collect more in premiums than it pays out in claims.**
- And to do that, the firm must have a solid grasp of what is known in industry jargon as the "expected loss" on every policy.
- This is exactly the same concept as expected value, only with an insurance twist.
- If your car is insured for $40,000, and the chances of its getting stolen in any given year are 1 in 1,000, then the annual expected loss on your car is $40.
- The annual premium for the theft portion of the coverage must be *>$40*.

"expected loss"

- At that point, the insurance company becomes just like the casino or the lottery.
- Yes, there will be payouts, but over the long run what comes in will be more than what goes out.
- As a consumer, you should recognize that insurance *will not* save you money in the long run.
- What it *will* do is prevent ***some unacceptably high loss***, such as replacing a $40,000 car that was stolen or a $350,000 house that burned down.

“bad bet”

- Buying insurance is a “bad bet” from a statistical standpoint since you will pay the insurance company, on average, more than you get back.
- Yet it can still be a sensible tool for protecting against outcomes that would otherwise **wreck your life**.
- Oddly, someone as rich as **Warren Buffett** can save money by not purchasing car insurance, homeowner’s insurance, or even health insurance because he can afford whatever bad things might happen to him.

“expected loss”

- Which finally brings us back to your $99 printer!
- We’ll assume that you’ve just picked out the perfect new laser printer at Best Buy.
- When you reach the checkout counter, the sales assistant will offer you a series of extended warranty options.
- For another $25 or $50, Best Buy will fix or replace the printer should it break in the next year or two.
- On average, you’ll pay more for the extended warranty than you would to repair the printer.

“expected loss”

- you should immediately be able to guess all of the following:
- (1) Best Buy is a for-profit business that seeks to maximize profits.
- (2) The sales assistant is eager for you to buy the extended warranty.
- (3) From numbers 1 and 2, we can infer that the cost of the warranty to you is greater than the expected cost of fixing or repairing the printer for Best Buy.
- If this were not the case, Best Buy would not be so aggressive in trying to sell it to you.
- (4) If your $99 printer breaks and you have to pay out of pocket to fix or replace it, this will not meaningfully change your life.

“expected loss”

- The broader lesson—and one of the core lessons of personal finance—is that
- you should always insure yourself against any adverse **events** that you cannot comfortably afford to withstand.
- You should skip buying insurance on everything else.

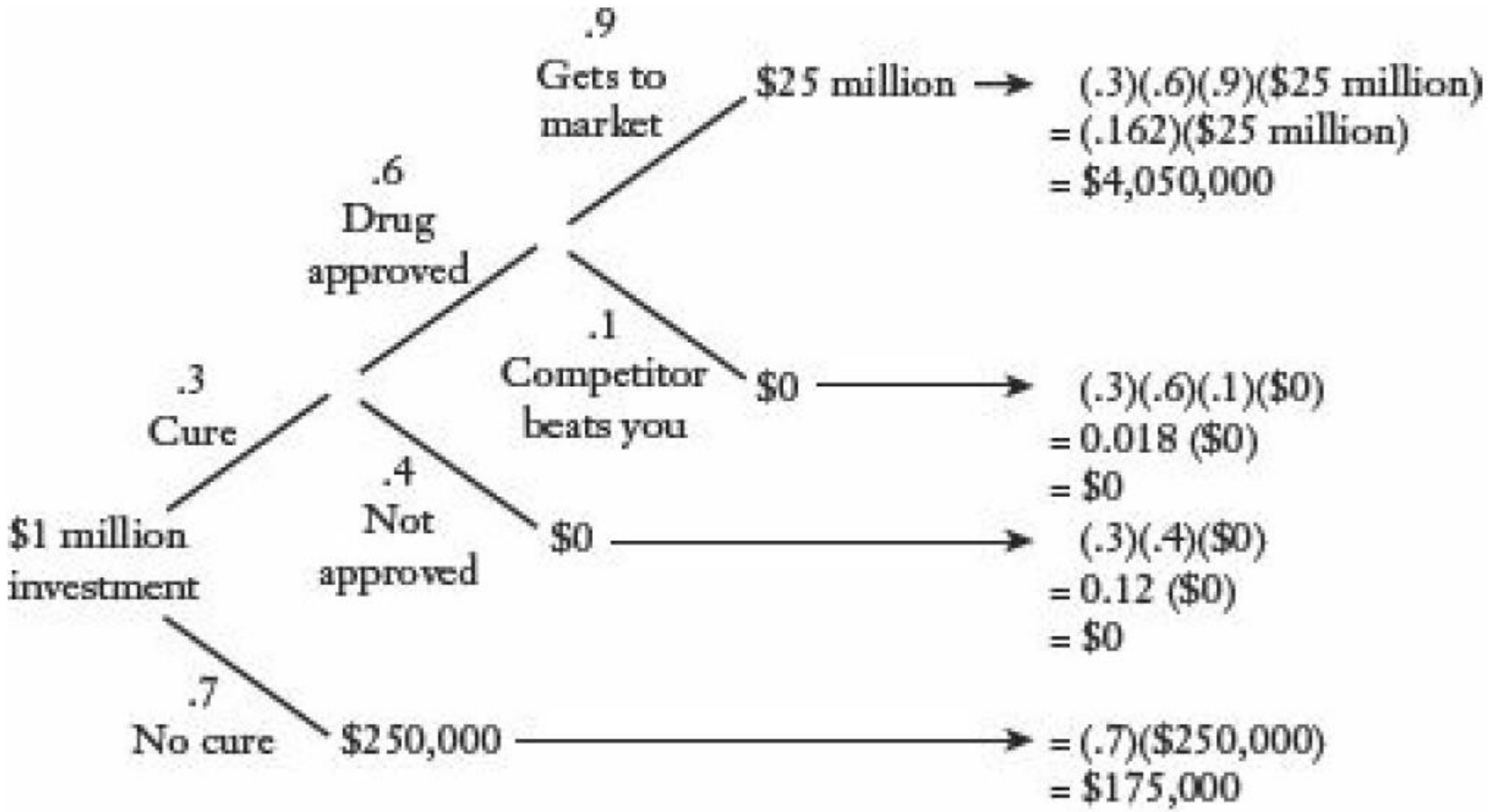
Invest or not?

- Suppose a friend of yours has asked you to invest $1 million in a research venture examining a new cure for male pattern baldness.
- You would probably ask what the likelihood of success will be;
- you'll get a complicated answer.
- This is a research project, so there is only a **30% chance** that the team will discover a cure that works.
- If the team does not find a cure, you will get **$250,000** of your investment back, as those funds will have been reserved for taking the drug to market (testing, marketing, etc.)

Invest or not?

- Even if the researchers are successful, there is only a **60% chance** that the U.S. FDA will approve the new miracle baldness cure as safe for use on humans.
- • Even then, if the drug is safe and effective, there is a **10% chance** that a competitor will come to market with a better drug at about the same time, wiping out any potential profits.
- • If everything goes well—the drug is safe, effective, and unchallenged by competitors—then the best estimate on the return on your investment is **$25 million**.
- This particular opportunity has an attractive expected value.

Invest or not?



Expected payoff = $4,050,000 + $0 + $0 + $175,000
= $4,225,000

Invest or not?

- The weighted payoff is $4.225 million.
- On the other hand, the most likely outcome, is that the company will not discover a cure for baldness and you will get only $250,000 back.
- Your appetite for this investment might depend on your risk profile.
- The law of large numbers suggests that an investment firm, or a rich individual like Warren Buffet, should seek out hundreds of opportunities like this with uncertain outcomes but attractive expected returns.
- Some will work; many won't.
- On average, these investors will make a lot of money, just like an insurance company or a casino.

HIV/AIDS test for everybody or not?

- If the expected payoff is in your favor, more trials are always better.
- Suppose we can test for some rare disease with a high degree of accuracy.
- let's assume the disease affects 1 of every 100,000 adults and the test is 99.9999 % accurate.
- The test never generates a false negative;
- however, roughly 1 in 10,000 tests conducted on a healthy person will generate a false positive
- The striking outcome here is that despite the impressive accuracy of the test, **most of the people who test positive will not have the disease**.

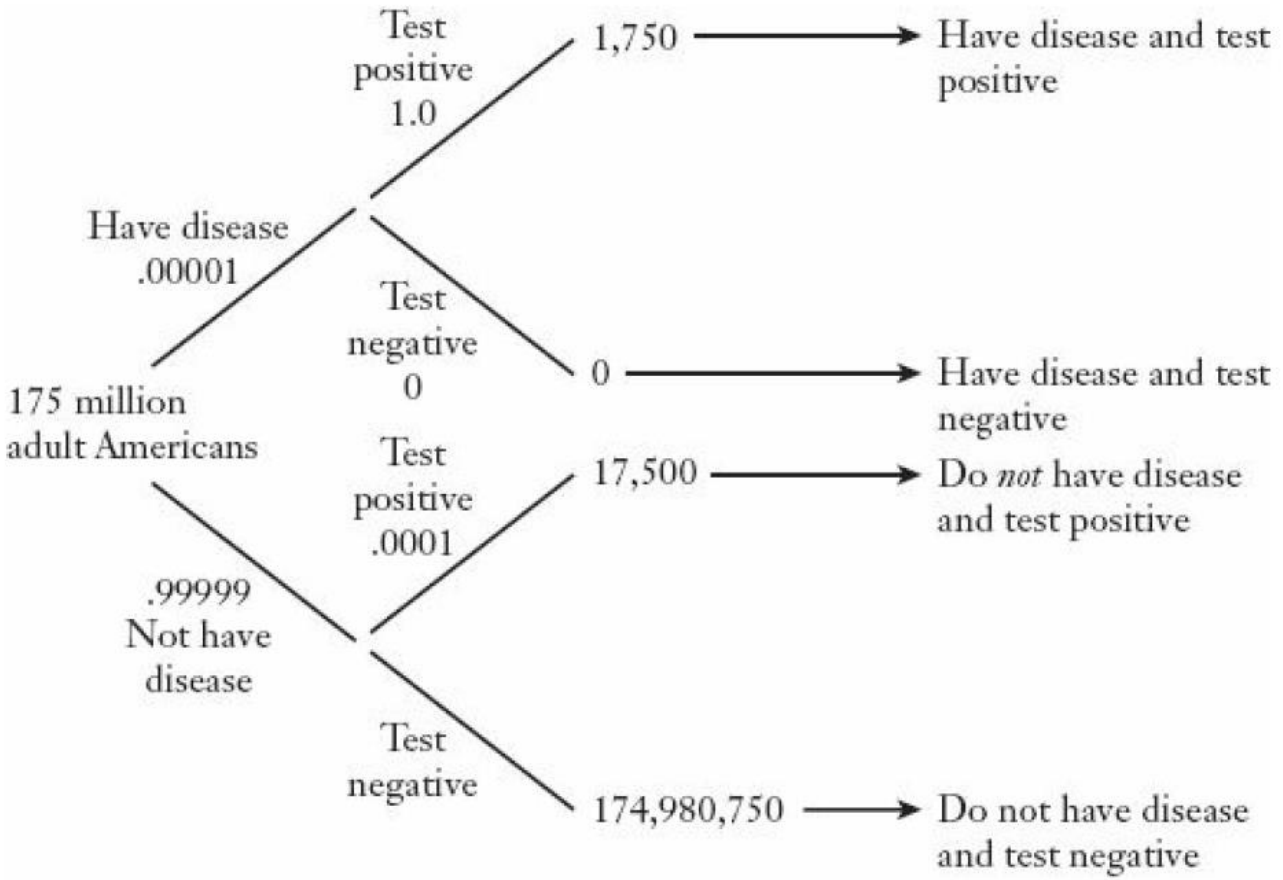
HIV/AIDS test for everybody or not?

- Suppose we can test for some rare disease with a high degree of accuracy.
- let's assume the disease affects 1 of every 100,000 adults and the test is 99.9999 % accurate.
- The test never generates a false negative;
- however, roughly 1 in 10,000 tests conducted on a healthy person will generate a false positive
- The striking outcome here is that despite the impressive accuracy of the test, **most of the people who test positive will not have the disease**.

Do HIV/AIDS test for everybody or not?

- This will generate enormous anxiety among those who falsely test positive; it can also waste finite health care resources on follow-up tests and treatment.
- Only 1,750 adults have the disease. They all test positive.
- Over 174 million adults do not have the disease.
- Of this healthy group who are tested, 99.9999 get the correct result that they do not have the disease.
- Only 0.0001 get a false positive.
- But 0.0001 of 174 million is still a big number.
- In fact, 17,500 people will, on average, get false positives.

Widespread Screening for a Rare Disease



$$\frac{\text{People with disease}}{\text{Those told they have the disease}} = \frac{1{,}750}{1{,}750 + 17{,}500} = \frac{1{,}750}{19{,}250} = .09 = 9\%$$

HIV/AIDS test for everybody or not?

- Let's look at what that means.
- A total of 19,250 people are notified that they have the disease; only 9 % of them are actually sick!
- And that's with a test that has a very low rate of false positives.
- In the case of a disease like HIV/AIDS, public health officials will often recommend that the resources available be used to screen the populations at highest risk, such as gay men or intravenous drug users.

Securities and Exchange Commission (SEC)

- SEC the government agency responsible for enforcing federal laws related to securities trading, uses a similar methodology for catching inside traders.
- The SEC uses powerful computers to scrutinize hundreds of millions of stock trades and look for suspicious activity, such as a big purchase of shares in a company just before a takeover is announced, or the dumping of shares just before a company announces disappointing earnings.
- Of course, smart investors are always trying to anticipate good and bad news and to devise perfectly legal strategies that consistently beat the market.
- Being a good investor does not necessarily make one a criminal.

Minority Report

- Tom Cruise plays a "pre-crime" detective who is part of a bureau that uses technology to predict crimes *before* they're committed.
- Well folks, that's not science fiction anymore.
- In 2011, the *New York Times* ran the following headline: "Sending the Police before There's a Crime."
- The story described how detectives were dispatched to a parking garage in downtown Santa Cruz by a computer program that predicted that there was a high likelihood of burglaries from cars at that location on that day.
- Police subsequently arrested two women peering into car windows.
- One had outstanding arrest warrants; the other was carrying illegal drugs.

Predictive policing

- “Predictive policing” is part of a broader movement called predictive analytics.
- Crime will always involve an element of uncertainty, as will determining who is going to crash his car or default on his mortgage.
- Probability helps us navigate those risks.
- And information refines our understanding of the relevant probabilities.
- Businesses facing uncertainty have always sought to quantify their risks.

Predictive policing

- Lenders request things like income verification and a credit score.
- Yet these blunt credit instruments are starting to feel like the prediction equivalent of a caveman's stone tools.
- The confluence of huge amounts of digital data and cheap computing power has generated fascinating insights into human behavior.
- Insurance officials correctly describe their business as the "transfer of risk"—and so they had better understand the risks being transferred to them.

Predictive policing

- The credit card companies are at the forefront of this kind of analysis, both because they are aware of so much data on our spending habits and because their business model depends so heavily on finding customers who are just barely a good credit risk.
- (The customers who are the best credit risks tend to be money losers because they pay their bills in full each month;
- the customers who carry large balances at high interest rates are the ones who generate fat profits—as long as they don't default.)
- One of the most intriguing studies of who is likely to pay a bill and who is likely to walk away was generated by J. P. Martin, "a math-loving executive" at Canadian Tire.

Predictive policing

- When Martin analyzed the data, he discovered that what customers purchased was a remarkably precise predictor of their subsequent payment behavior when used in conjunction with traditional tools like income and credit history.
- “People who bought cheap, generic automotive oil were much more likely to miss a credit-card payment than someone who got the expensive, name brand stuff.
- People who bought carbon-monoxide monitors for their homes or those little felt pads that stop chair legs from scratching the floor almost never missed payments.
- Anyone who purchased a chrome-skull car accessory or a ‘Mega Thruster Exhaust System’ was pretty likely to miss paying his bill eventually.”

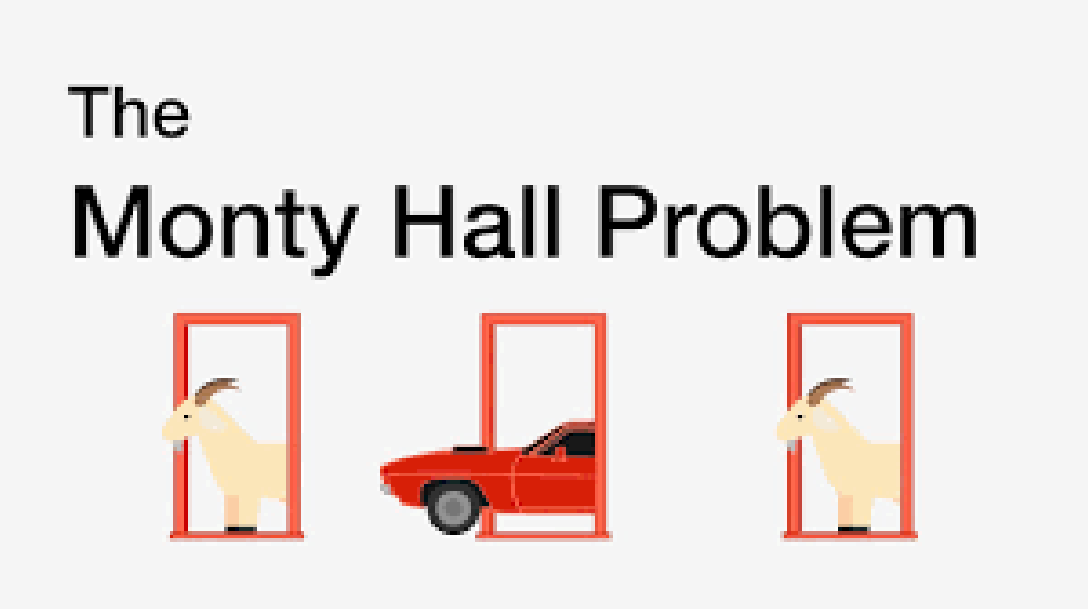
Probability

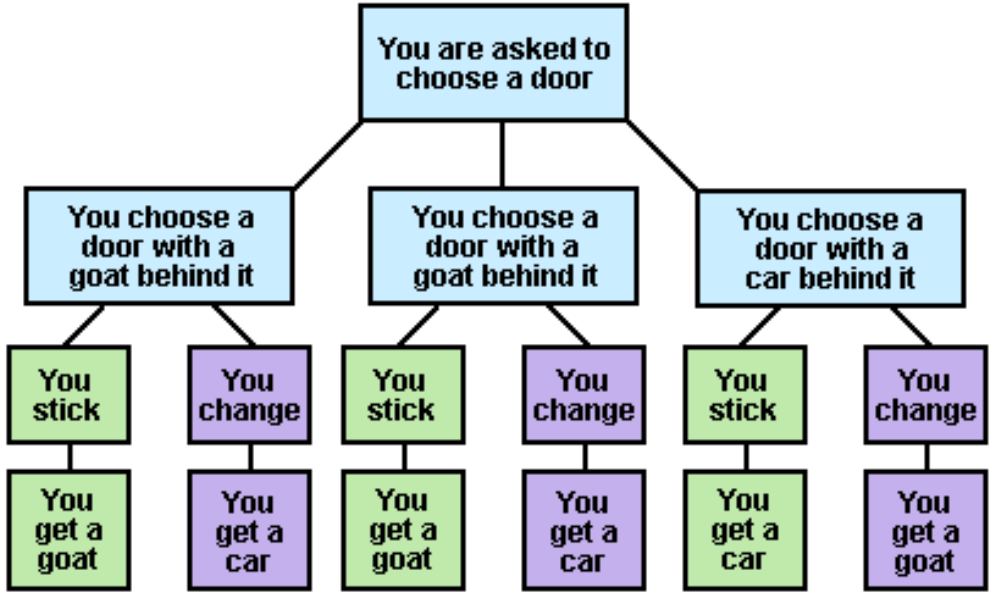
- Probability gives us tools for dealing with life's uncertainties.
- You shouldn't play the lottery.
- You should invest in the stock market if you have a long investment horizon (because stocks typically have the best long-term returns).
- You should buy insurance for some things, but not others.
- Probability can even help you maximize your winnings on game shows.
- That said (or written), probability is not deterministic.
- No, you shouldn't buy a lottery ticket—but you still might win money if you do.
- And yes, probability can help us catch cheaters and criminals—but when used inappropriately it can also send innocent people to jail.

RESEARCH
Khatam University
FANAP

The Monty Hall Problem

- The "Monty Hall problem" is a famous probability-related conundrum (problem) faced by participants on the game show *Let's Make a Deal*, which premiered in the United States in 1963 and is still running in some markets around the world.





Problems with Probability

- *How overconfident math geeks nearly destroyed the global financial system*

Value at Risk model

- Statistics cannot be any smarter than the people who use them.
- And in some cases, they can make smart people do dumb things.
- One of the most irresponsible uses of statistics in recent memory involved the mechanism for gauging risk on Wall Street prior to the 2008 financial crisis.
- At that time, firms throughout the financial industry used a common barometer of risk, the Value at Risk model, or VaR.

VaR

- In theory, VaR combined the elegance of an indicator (collapsing lots of information into a single number) with the power of probability (attaching an expected gain or loss to each of the firm's assets or trading positions).
- The model assumed that there is a range of possible outcomes for every one of the firm's investments.
- •For example, if the firm owns General Electric stock, the value of those shares can go up or down.
- •When the VaR is being calculated for some short period of time, say, one week, the most likely outcome is that the shares will have roughly the same value at the end of that stretch as they had at the beginning.

VaR

- There is a smaller chance that the shares may rise or fall by 10%.
- And an even smaller chance that they may rise or fall 25%, and so on.
- On the basis of past data for market movements, the firm's quantitative experts (often called "quants" or "rich nerds") could assign a dollar figure, say $13 million, that represented the maximum that the firm could lose on that position over the time period being examined, with 99 % probability.
- In other words, 99 times out of 100 the firm would not lose more than $13 million on a particular trading position;
- **1 time out of 100, it would.**

VaR

- Prior to the financial crisis of 2008, firms trusted the VaR model to quantify their overall risk.
- If a single trader had 923 different open positions (investments that could move up or down in value), each of those investments could be evaluated as described above for the General Electric stock;
- **from there, the trader's total portfolio risk could be calculated**.

VaR

- The formula even took into account the correlations among different positions.
- For example, if two investments had expected returns that were negatively correlated, a loss in one would likely have been offset by a gain in the other, making the two investments together less risky than either one separately.
- Overall, the head of the trading desk would know that bond trader Bob Smith has a 24-hour VaR of $19 million, again with 99 % probability.
- The most that Bob Smith could lose over the next 24 hours would be $19 million, 99 times out of 100.

VaR

- Then, even better, the aggregate risk for the firm could be calculated at any point in time by taking the same basic process one step further.
- The underlying mathematical mechanics are obviously fabulously complicated, as firms had a dizzying array of investments in different currencies,
- with different amounts of leverage (the amount of money that was borrowed to make the investment), trading in markets with different degrees of liquidity, and so on.
- Despite all that, the firm's managers apparently had a precise measure of the magnitude of the risk that the firm had taken on at any moment in time.

VaR

- “VaR’s great appeal, and its great selling point to people who do not happen to be quants, is that it expresses risk as a single number, a dollar figure, no less.”
- At J. P. Morgan, where the VaR model was developed and refined, the daily VaR calculation was known as the “4:15 report” because it would be on the desks of top executives every afternoon at 4:15, just after the American financial markets had closed for the day.
- VaR has been called “potentially catastrophic,” “a fraud,” and many other things not fit for a family book about statistics like this one.

VaR

- In particular, the model has been blamed for the start and severity of the financial crisis.
- The primary critique of VaR is that the underlying risks associated with financial markets are not as predictable as a coin flip or even a blind taste test between two beers.
- **The false precision embedded in the models created a false sense of security**.
- The VaR was like a faulty speedometer, which is arguably worse than no speedometer at all.
- If you place too much faith in the broken speedometer, you will be unconscious to other signs that your speed is unsafe.
- In contrast, if there is no speedometer at all, you have no choice but to look around for clues as to how fast you are really going.

two huge problems

- There were two huge problems with the risk profiles encapsulated by the VaR models.
- First, the underlying probabilities on which the models were built were based on past market movements; however, in financial markets (unlike beer tasting), the future does not necessarily look like the past.
- There was no intellectual justification for assuming that the market movements from 1980 to 2005 were the best predictor of market movements after 2005.
- In some ways, this failure of imagination resembles the military's periodic mistaken assumption that the next war will look like the last one.

EXAMPLE

- In the 1990s and early 2000s, commercial banks were using lending models for home mortgages that assigned zero probability to large declines in housing prices.
- Housing prices had never before fallen as far and as fast as they did beginning in 2007.
- But that's what happened.

Alan Greenspan:

- “The whole intellectual edifice, however, collapsed in the summer of [2007] because the data input into the risk management models generally covered only the past two decades, a period of euphoria (joy).
- Had instead the models been fitted more appropriately to historic periods of stress, capital requirements would have been much higher and the financial world would be in far better shape, in my judgment.”

tail risk

- Second, even if the underlying data could accurately predict future risk, the 99 % assurance offered by the VaR model was dangerously useless, *because it's* ***the 1 % that is going to really mess you up***.
- Hedge fund manager David Einhorn explained, "This is like an air bag that works all the time, except when you have a car accident."
- If a firm has a Value at Risk of$500 million, that can be interpreted to mean that the firm has a 99% chance of losing no more than $500 million over the time period specified.

tail risk

- Well, hello, that also means that the firm has a 1 % chance of losing more than $500 million—much, much more under some circumstances.
- In fact, the models had nothing to say about how bad that 1 % scenario might turn out to be.
- Very little attention was devoted to the “tail risk,” the small risk (named for the tail of the distribution) of some catastrophic outcome.

Hank Paulson:

- Many firms assumed they could raise cash in a pinch by selling assets.
- But during a crisis, every other firm needs cash, too, so all are trying to sell the same kinds of assets.
- It's the risk management equivalent of saying, "I don't need to stock up on water because if there is a natural disaster, I'll just go to the supermarket and buy some."
- Of course, after an asteroid hits your town, fifty thousand other people are also trying to buy water; by the time you get to the supermarket, the windows are broken and the shelves are empty.

The Impact of the Highly Improbable and a scathing critic of VaR:

- "The greatest risks are never the ones you can see and measure, but the ones you can't see and therefore can never measure.
- The ones that seem so far outside the boundary of normal probability that you can't imagine they could happen in your lifetime—even though, of course, they do happen, more often than you care to realize."
- The statistical self-confidence at commercial banks and on Wall Street ultimately contributed to the most severe global financial contraction since the Great Depression.

3 fundamental errors of Wall street guys

- **First**, they confused precision with accuracy. (my golf range finder): exact and wrong.
- The false precision led Wall Street executives to believe that they had risk on a leash (in control) when in fact they did not.
- **Second**, the estimates of the underlying probabilities were wrong.
- As Alan Greenspan pointed out in testimony quoted earlier in the chapter, the relatively peaceful and prosperous decades before 2005 should not have been used to create probability distributions for what might happen in the markets in the succeeding decades.
- This is the equivalent of walking into a casino and thinking that you will win at roulette 62 % of the time because that's what happened last time you went gambling.
- It would be a long, expensive evening.

3 fundamental errors of Wall street guys

- Third, firms neglected their “tail risk.”
- The VaR models predicted what would happen 99 times out of 100.
- That’s the way probability works.
- **Unlikely things happen.**
- In fact, over a long enough period of time, they are not even that unlikely.
- People get hit by lightning all the time.
- My mother has had three holes in one.

statistics as a powerful weapon

- In sticking with the “statistics as a powerful weapon” metaphor, I will paraphrase the gun rights lobby: Probability doesn’t make mistakes; people using probability make mistakes.

- The balance of this chapter will catalog some of the most common probability-related errors, misunderstandings, and ethical dilemmas.

Assuming events are independent when they are not.

- The probability of a jet engine's failing for any reason during is not independent.
- Sudden infant death syndrome (SIDS),
- Meadow's Law—the idea that one infant death is a tragedy, two are suspicious and three are murder.

- Kathleen Folbigg: Genetics may help free Australian woman ...
- https://edition.cnn.com › 2021/03/19 › australia › kathlee...
- Esfand 30, 1399 AP — **Genetics** may free a woman **convicted of killing** her 4 **babies** and help other parents explain the unexplainable.

Not understanding when events ARE independent.

- People looking longingly at the dice or cards and declaring that they are “due.”
- The belief otherwise is sometimes called “the gambler’s fallacy.”
- In fact, if you flip a fair coin 1,000,000 times and get 1,000,000 heads in a row, the probability of getting tails on the next flip is still 1/2.
- A player who makes a shot is more likely to hit the next shot than a player who has just missed.

Perception vs. empirical reality

- “no evidence for a positive correlation between the outcomes of successive shots.”
- The players hit an average of 48 % of their field goals after hitting their last shot and 47 % after missing.
- For 14 of 26 players, the correlation between making one shot and then making the next was negative.
- Only one player showed a significant positive correlation between one shot and the next.
- 91% of basketball fans surveyed at Stanford and Cornell agreed with the statement that a player has a better chance of making his next shot after making his last 2 or 3 shots than he does after missing his last 2/3 shots.

Perception vs. empirical reality

- The difference between the perception and the empirical reality.
- “People’s intuitive conceptions of randomness depart systematically from the laws of chance.”
- We see patterns where none may really exist.
- Like cancer clusters.
- Some statistically unlikely number of people in a particular area have contracted a rare form of cancer.

Clusters happen.

- It must be the water, or the local power plant, or the cell phone tower.
- Of course, any one of those things might really be causing adverse health outcomes. (statistics can identify such causal relationships.)
- But this cluster of cases may also be the product of pure chance, even when the number of cases appears highly improbable.
- Yes, the probability that five people in the same school or workplace will contract the same rare form of leukemia may be one in a million, *but there are millions of schools and workplaces*.

what likely happened / didn't happen

- The chance of winning the lotto may be 1 in 20 million, but none of us is surprised when *someone* wins, because millions of tickets have been sold.

- **Richard Lustig** is an American man who came to prominence for winning relatively large prizes in seven state-sponsored lottery games from 1993 to 2010. His prizes totaled over $1 million. He wrote *Learn How To Increase Your Chances of Winning the Lottery.*

- An Idaho **man** can now call himself a **lotto winner six times** over. Bryan Moss, from Meridian, **won** a $250,000 prize from a Crossword scratch game on Thursday, the Idaho **Lottery** said in a press release, marking his sixth ever **lottery win**.

what likely happened / didn't happen

- I ask everyone in the class to take out a coin and stand up……
- "How did you do it?"
- "What are the best training exercises for flipping so many tails in a row?"
- "Is there a special diet that helped you pull off this impressive accomplishment?"
- When we see an anomalous event like that out of context, however, we assume that something besides randomness must be responsible.

The prosecutor's fallacy

- Suppose you hear testimony in court to the following effect:
- (1) a DNA sample found at the scene of a crime matches a sample taken from the defendant; and
- (2) there is only one chance in a million that the sample recovered at the scene of the crime would match anyone's besides the defendant.
- On the basis of that evidence, would you vote to convict?

- The prosecutor's fallacy (a mistaken belief) occurs when the context surrounding statistical evidence is neglected.

The prosecutor's fallacy

- Here are two scenarios, each of which could explain the DNA evidence being used to prosecute the defendant.
- *Defendant 1*: This defendant, a rejected lover of the victim, was arrested three blocks from the crime scene carrying the murder weapon.
- After he was arrested, the court compelled him to offer a DNA sample, which matched a sample taken from a hair found at the scene of the crime.

The prosecutor's fallacy

- *Defendant 2*: This defendant was convicted of a similar crime in a different state several years ago.
- As a result of that conviction, his DNA was included in a national DNA database of over a million violent felons.
- The DNA sample taken from the hair found at the scene of the crime was run through that database and matched to this individual, who has no known association with the victim.
- *Because the chances of finding a coincidental one in a million match are relatively high if you run the sample through a database with samples from a million people.*

Reversion (or regression) to the mean

- *Sports Illustrated* jinx (bad luck) of being on the cover of a magazine.
- One explanation is that being on the cover of the magazine has some adverse effect on subsequent performance.
- The more statistically sound explanation is that teams and athletes appear on its cover after some anomalously good stretch (such as a twenty-game winning streak) and that their subsequent performance merely reverts back to what is normal, or the mean.
- This is the phenomenon known as reversion to the mean.
- Probability tells us that any outlier is likely to be followed by outcomes that are more consistent with the long-term average.

“Best Managers’s underperformance?

- When CEOs achieve “superstar” status, they get distracted by their new prominence.
- They write their memoirs.
- They are invited to sit on outside boards.
- They begin searching for trophy spouses.
- “Our results suggest that media-induced superstar culture leads to behavioral distortions beyond mere mean reversion.”
- In other words, when a CEO appears on the cover of Businessweek, sell the stock.

Statistical discrimination.

- When is it okay to act on the basis of what probability tells us is likely to happen, and when is it not okay?
- insurance companies may not charge different rates to men and women, because it violates the EU's principle of equal treatment.
- To insurers, however, gender-based premiums aren't discrimination; they're just statistics.
- Men typically pay more for auto insurance because they crash more.
- Women pay more for annuities (pension) because they live longer.

Statistical discrimination.

- Obviously many women crash more than many men, and many men live longer than many women.
- But, as explained in the last chapter, insurance companies don't care about that.
- They care only about what happens on average, because if they get that right, the firm will make money.
- The interesting thing about the EU policy is that the authorities are not pretending that gender is unrelated to the risks being insured;
- They are simply declaring that disparate rates based on sex are unacceptable.*

outcome A *or* outcome B

- Probability can lead us to some interesting but upsetting places in this regard.
- How should we react when our probability-based models tell us that methamphetamine smugglers from Mexico are most likely to be Hispanic men aged between eighteen and thirty and driving red pickup trucks between 9:00 p.m. and midnight when we also know that the vast majority of Hispanic men who fit that profile are not smuggling methamphetamine?

- Yes, that is just basic.
- But it is also statistics with social implications.
- If we want to catch violent criminals and terrorists and drug smugglers and other individuals with the potential to do enormous harm, then we ought to use every tool at our disposal.

philosophical and legal question

- Probability can be one of those tools.
- It would be naïve to think that gender, age, race, ethnicity, religion, and country of origin collectively tell us nothing about anything related to law enforcement.
- Is it okay to discriminate if the data tell us that we'll be right far more often than we're wrong?
- How much of that is acceptable?
- If we can build a model that identifies drug smugglers correctly 80 out of 100 times, what happens to the poor souls in the 20 percent—*because our model is going to harass (offend) them over and over and over again*.
- For all the elegance and precision of probability, there is no substitute for thinking about what calculations we are doing and why we are doing them.

The Importance of Data

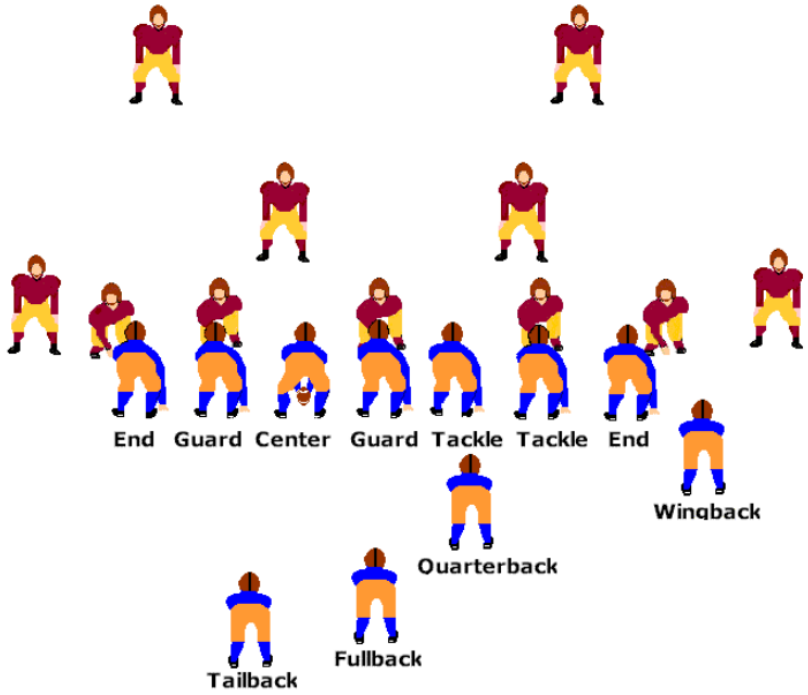
- *"Garbage in, garbage out"*

Data

- In the spring of 2012, researchers published a striking finding in the esteemed journal *Science*.
- According to this cutting-edge research, when male fruit flies are rejected repeatedly by female fruit flies, they drown their sorrows in alcohol.
- The *New York Times* described the study in a front page article: “They were young males on the make, and they struck out not once, not twice, but a dozen times with a group of attractive females hovering nearby.
- So they did what so many men do after being repeatedly rejected: they got drunk, using alcohol as a balm for unfulfilled desire.”1

Data

- Data are to statistics what a good offensive line is to a star quarterback.
- In front of every star quarterback is a good group of blockers.
- They usually don't get much credit.
- But without them, you won't ever see a star quarterback.
- Most statistics books assume that you are using good data, just as a cookbook assumes that you are not buying rancid meat and rotten vegetables.
- But even the finest recipe isn't going to salvage a meal that begins with spoiled ingredients.



Data

- So it is with statistics; no amount of fancy analysis can make up for fundamentally flawed data.
- Hence the expression “garbage in, garbage out.”
- Data deserve respect, just like offensive linemen.

Data

- We generally ask our data to do one of three things.
- First, we may demand a data sample that is representative of some larger group or population.
- Because inferences made from reasonably large, properly drawn samples can be every bit as accurate as attempting to elicit **(obtain)** the same information from the entire population.
- Statistics books almost always illustrate this point by drawing colored marbles out of an urn (jug).

Data

- telephone random dialer has potential flaws.
- Some individuals (particularly low-income persons) may not have a telephone.
- Others (particularly high-income persons) may be more prone to screen calls and choose not to answer.

- **If you've stirred your soup adequately, a single spoonful can tell you how the whole pot tastes.**

Sampling

- (1) A representative sample is a fabulously important thing, for it opens the door to some of the most powerful tools that statistics has to offer.
- (2) Getting a good sample is harder than it looks.
- (3) Many of the most **egregious statistical assertions** are caused by good statistical methods applied to bad samples, not the opposite.
- (4) Size matters, and bigger is better. it should be intuitive that a larger sample will help to smooth away any freak variation.

Sampling

- One crucial warning is that a bigger sample will not make up for errors in its composition, or “bias.”
- A bad sample is a bad sample.
- No supercomputer or fancy formula is going to rescue the validity of your national presidential poll if the respondents are drawn only from a telephone survey of Washington, D.C., residents.
- The residents of Washington, D.C., don’t vote like the rest of America; calling 100,000 D.C. residents rather than 1,000 is not going to fix that fundamental problem with your poll.
- **In fact, a large, biased sample is arguably worse than a small, biased sample because it will give a false sense of confidence regarding the results.**

Source of comparison.

- The second thing we often ask of data is that they provide some source of comparison.
  - Is a new medicine more effective than the current treatment?
  - Are ex-convicts who receive job training less likely to return to prison than ex-convicts who do not receive such training?
  - Do students who attend charter schools perform better than similar students who attend regular public schools?
- In these cases, the goal is to find two groups of subjects who are broadly similar except for the application of whatever “treatment” we care about.

Source of comparison.

- In a social science context, the word “treatment” is broad enough to encompass anything ….
- As with any other application of the scientific method, we are trying to isolate the impact of *one specific intervention or attribute*.
- This was the genius of the any experiment.
- The researchers must figure out a way to create a control group and a “treatment” group; the subsequent difference in their behaviors can then be attributed to whether they were treated or not.
- In the physical and biological sciences, creating treatment and control groups is relatively straightforward.

Source of comparison.

- Chemists can make small variations from test tube to test tube and then study the difference in outcomes.
- Biologists can do the same thing with their petri dishes.
- Even most animal testing is simpler than trying to get fruit flies to drink alcohol.
- We can have one group of rats exercise regularly on a treadmill and then compare their mental acuity in a maze with the performance of another group of rats that didn't exercise.
- But when humans become involved, things grow more complicated.

RESEARCH
Khatam University
FANAP

Source of comparison.

- Sound statistical analysis often requires a treatment and a control group, yet we cannot force people to do the things that we make laboratory rats do. (And many people do not like making even the lab rats do these things.)
- Do repeated concussions **(hits)** cause serious neurological problems later in life?
- This is a really important question.
- The future of football hangs on the answer.
- Yet it is a question that cannot be answered with experiments on humans.
- So unless and until we can teach fruit flies to wear helmets and run the spread offense, we have to find other ways to study the long-term impact of head trauma.
- One recurring research challenge with human subjects is creating treatment and control groups that differ *only* in that one group is getting the treatment and the other is not.

Source of comparison.

- For this reason, the **"gold standard"** of research is randomization, a process by which human subjects are randomly assigned to either the treatment or the control group.
- We do not assume that all the experimental subjects are identical.
- Instead, probability becomes our friend (once again), and we assume that randomization will evenly divide all relevant characteristics between the two groups—both the characteristics we can observe, like race or income, but also confounding characteristics that we cannot measure or had not considered, such as perseverance or faith

Just because

- To quote my teenage daughter, “Just because.”
- We sometimes have no specific idea what we will do with the information—but we suspect it will come in handy at some point.
- This is similar to a crime scene detective who demands that all possible evidence be captured so that it can be sorted later for clues.
- Some of this evidence will prove useful, some will not. If we knew exactly what would be useful, we probably would not need to be doing the investigation in the first place.

Just because

- You probably know that smoking and obesity are risk factors for heart disease.
- You probably don't know that a long-running study of the residents of Framingham, Massachusetts, helped to clarify those relationships.
- Framingham is a suburban town of some 67,000 people about twenty miles west of Boston.
- To non-researchers, it is best known as a suburb of Boston.
- To researchers, Framingham is best known as the home of the Framingham Heart Study, one of the most successful and influential longitudinal studies in the history of modern science.

Just because

- A longitudinal study collects information on a large group of subjects at many different points in time, such as once every two years.
- The same participants may be interviewed periodically for ten, twenty, or even fifty years after they enter the study, creating a remarkably rich source of information.
- In the case of the Framingham study, researchers gathered information on 5,209 adult residents of Framingham in 1948: height, weight, blood pressure, educational background, family structure, diet, smoking behavior, drug use, and so on.
- Most important, researchers have gathered follow-up data *from the same participants* ever since (and also data on their offspring, to examine genetic factors related to heart disease).

Framingham Data

male	age	education	Current Smoker	Cigs PerDay	BP Meds	Prevalent Stroke	Prevalent Hyp	diabetes	Tot Chol	Sys BP	Dia BP	BMI	Heart Rate	glucose	Ten Year CHD
1	39	4	0	0	0	0	0	0	195	106	70	26.97	80	77	0
0	46	2	0	0	0	0	0	0	250	121	81	28.73	95	76	0
1	48	1	1	20	0	0	0	0	245	127.5	80	25.34	75	70	0
0	61	3	1	30	0	0	1	0	225	150	95	28.58	65	103	1
0	46	3	1	23	0	0	0	0	285	130	84	23.1	85	85	0
0	43	2	0	0	0	0	1	0	228	180	110	30.3	77	99	0
0	63	1	0	0	0	0	0	0	205	138	71	33.11	60	85	1
0	45	2	1	20	0	0	0	0	313	100	71	21.68	79	78	0
1	52	1	0	0	0	0	1	0	260	141.5	89	26.36	76	79	0

Just because

- The Framingham data have been used to produce over two thousand academic articles since 1950, including nearly a thousand between 2000 and 2009.
- These studies have produced findings crucial to our understanding of cardiovascular disease, many of which we now take for granted (presume):
  1. cigarette smoking increases the risk of heart disease (1960);
  2. physical activity reduces the risk of heart disease and obesity increases it (1967);
  3. high blood pressure increases the risk of stroke (1970);
  4. high levels of HDL cholesterol (henceforth known as the “good cholesterol”) reduce the risk of death (1988);
  5. individuals with parents and siblings who have cardiovascular disease are at significantly higher risk of the same (2004 and 2005).

Types Data

- **Categorical**: the values represent different categories for the individuals; do not have arithmetical meaning.
- **Quantitative:** the values represent numerical quantities that can be ordered and averaged.
- **Ordinal**: the values represent ordered categories; such as “how often do you exercise?”—Everyday, frequently, sometimes, rarely, never.

Individual	Gender	Height	Education
1	Male	70	College
2	Female	68	College
3	Male	69	High School

Displaying categorial variable

- • For categorical variable, we summarize the data using the counts of observed occurrences of each value.
- • Alternatively, we can use percentage or proportion.

Individual	Gender	Height	Education
1	Male	70	College
2	Female	68	College
3	Male	69	High School

Displaying quantitative variable

- **For quantitative variables, we also summarize the data using the counts of observed occurrences of values.**
- **Different from categorical variables, we may count occurrences within intervals rather than individual values.**
- **We also use percentage or proportion.**
- Area of each bar (The height of the bar multiplied by the width of the interval) = the proportion of death in the corresponding age interval.

Longitudinal data

- Longitudinal data sets are the research equivalent of a Ferrari.
- The data are particularly valuable when it comes to exploring causal relationships that may take years or decades to unfold.
- For example, the **Perry Preschool Study** began in the late 1960s with a group of 123 African American 3-4-yearolds from poor families.
- The participating children were randomly assigned into a group that received an intensive preschool program and a comparison group that did not.
- **Researchers then measured various outcomes for both groups *for the next forty years.***

Early childhood education

- The results make a compelling case for the benefits of early childhood education.
  1. The students who received the intensive preschool experience had higher IQs at age five.
  2. They were more likely to graduate from high school.
  3. They had higher earnings at age forty.
  4. In contrast, the participants who did not receive the preschool program were significantly more likely to have been arrested five or more times by age forty.

cross-sectional data set

- **Not surprisingly, we can't always have the Ferrari.**
- **The research equivalent** of a Toyota is a cross-sectional data set, which is a collection of data gathered at a single point in time.
- For example, if epidemiologists are searching for the cause of a new disease (or an outbreak of an old one), they may gather data from all those afflicted in hopes of finding a pattern that leads to the source.
  1. What have they eaten?
  2. Where have they traveled?
  3. What else do they have in common?
- Researchers may also gather data from individuals who are not afflicted by the disease **to highlight contrasts** between the two groups.

The third reason we collect data is

- Behind every important study there are good data that made the analysis possible.
- And behind every bad study . . . well, read on.
- People often speak about "lying with statistics."
- I would argue that some of the most egregious statistical mistakes involve *lying with data*; the statistical analysis is fine, but the data on which the calculations are performed are bogus or inappropriate.
- Here are some common examples of "garbage in, garbage out."

Problems with

- ***Selection bias***
- ***Publication bias***
- ***Recall bias***
- ***Healthy user bias***

Problems with data-***Selection bias***

- "Nixon couldn't have won. I don't know anyone who voted for him."
- As polls with good samples get larger, they get better, since the margin of error shrinks.
- As polls with bad samples get larger, the pile of garbage just gets bigger and smellier.
- *Literary Digest* predicted that Landon would beat Roosevelt with 57 % of the popular vote.
- In fact, Roosevelt won in a landslide, with 60 % of the popular vote and 46 of 48 states in the electoral college.
- The *Literary Digest* sample was "garbage in": the magazine's subscribers were wealthier than average Americans, and therefore more likely to vote Rep, as were households with telephones and cars in 1936.

Problems with data-*Selection bias*

- We can end up with the same basic problem when we compare outcomes between a treatment and a control group if the mechanism for sorting individuals into one group or the other is not random.
- Consider a recent finding in the medical literature on the side effects of treatment for prostate cancer.
- There are three common treatments for prostate cancer:
  - surgical removal of the prostate;
  - radiation therapy;
  - or brachytherapy (which involves implanting radioactive “seeds” near the cancer)

Problems with data-Selection bias

- A related source of bias, known as self-selection bias, will arise whenever individuals volunteer to be in a treatment group.
- For example, prisoners who volunteer for a drug treatment group are different from other prisoners *because they have volunteered to be in a drug treatment program*.
- If the participants in this program are more likely to stay out of prison after release than other prisoners, that's great—but it tells us absolutely nothing about the value of the drug treatment program.
- These former inmates may have changed their lives because the program helped them kick drugs.

Problems with data-Selection bias

- Or they may have changed their lives because of other factors that also happened to make them more likely to volunteer for a drug treatment program (such as having a really strong desire not to go back to prison).
- We cannot separate the causal impact of one (the drug treatment program) from the other (being the kind of person who volunteers for a drug treatment program).

Problems with data: Publication bias

- Positive findings are more likely to be published than negative findings, which can skew the results that we see.
- Suppose you have just conducted a rigorous, longitudinal study in which you find conclusively that **playing video games *does not* prevent colon cancer.**
- You've followed a representative sample of 100,000 volunteers for twenty years; those participants who spend hours playing video games have roughly the same incidence of colon cancer as the participants who do not play video games at all.
- We'll assume your methodology is perfect.

Problems with data: Publication bias

- Which prestigious medical journal is going to publish your results?
- None, for two reasons.
  - First, there is no strong scientific reason to believe that playing video games has any impact on colon cancer, so it is not obvious why you were doing this study.
  - Second, and more relevant here, the fact that something *does not* prevent cancer is not a particularly interesting finding. After all, most things *don't* prevent cancer.
- Negative findings are not especially attractive, in medicine or elsewhere.
- The net effect is to distort the research that we see, or do not see.

Problems with data: Publication bias

- Suppose that one of your classmates has conducted a different longitudinal study.
- **She finds that people who spend a lot of time playing video games *do* have a lower incidence of colon cancer.**
- *Now that is interesting!*
- That is exactly the kind of finding that would catch the attention of a medical journal, the popular press, bloggers, and video game makers.

Problems with data: Publication bias

- *New York Times* article on the publication bias surrounding drugs for treating depression:
- "The makers of antidepressants like Prozac and Paxil never published the results of about a third of the drug trials that they conducted to win government approval, misleading doctors and consumers about the drugs' true effectiveness."
- It turns out that 94 % of studies with positive findings on the effectiveness of these drugs were published, while only 14 % of the studies with nonpositive results were published.

Problems with data: Publication bias

- For patients dealing with depression, this is a big deal.
- When all the studies are included, the antidepressants are better than a placebo by only "**a modest margin**."
- To combat this problem, medical journals now typically require that any study be registered at the beginning of the project if it is to be eligible for publication later on.

Problems with data: **Publication bias**

- This gives the editors some evidence on the ratio of positive to non-positive findings.
- If 100 studies are registered that propose to examine ***the effect of skateboarding on heart disease***, and only one is ultimately submitted for publication with positive findings, the editors can infer that the other studies had non-positive findings (or they can at least investigate this possibility).

Problems with data: ***Recall bias***

- Memory is a fascinating thing—though not always a great source of good data.

Problems with data: ***Healthy user bias.***

- People who take vitamins regularly are likely to be
- healthy—*because they are the kind of people who take vitamins regularly!*

The Central Limit Theorem

- *The Lebron James of statistics*

Extraordinary power to generalize

- At times, statistics seems almost like magic.
- We are able to draw sweeping and powerful conclusions from relatively little data.
- Somehow we can gain meaningful insight into a presidential election by calling a mere one thousand American voters.
- We can test a hundred chicken breasts for salmonella at a poultry processing plant and conclude from that sample alone that the entire plant is safe or unsafe.
- *Where does this Extraordinary power to generalize come from?*

Extraordinary power to generalize

- Much of it comes from the central limit theorem, which is the Lebron James of statistics.
- The central limit theorem is the “power source” for many of the statistical activities that involve using a sample to make inferences about a large population (like a poll, or a test for salmonella).
- These kinds of inferences may seem mystical;
- In fact, they are just a combination of two tools that we’ve already explored: probability and **proper sampling**.
- Before plunging into the mechanics of the central limit theorem, here is an example to give you the general intuition.

Extraordinary power to generalize

- Suppose you live in a city that is hosting a marathon.
- Runners from all over the world will be competing, which means that many of them do not speak English.
- The logistics of the race require that runners check in on the morning of the race, after which they are randomly assigned to buses to take them to the starting line.
- Unfortunately one of the buses gets lost on the way to the race.
- (no one has a cell phone and that the driver does not have a GPS navigation device)
- As a civic leader in this city, you join the search team.

Extraordinary power to generalize

- As luck would have it, you stumble upon (face) a broken-down bus near your home with a large group of unhappy international passengers, none of whom speaks English.
- This must be the missing bus!
- Except you have one lingering doubt . . . the passengers on this bus are, well, **very large**.
- Based on a quick glance, you compute that the average weight for this group of passengers has got to be over 220 pounds.
- There is no way that a random group of marathon runners could all be this heavy.
- You radio your message to search headquarters: "**I think it's the wrong bus. Keep looking**."

Extraordinary power to generalize

- Further analysis confirms your initial impression.
- When a translator arrives, you discover that this disabled bus was headed to the **International Festival of Sausage**, which is also being hosted by your city on the same weekend. Congratulations.
- If you can grasp how someone who takes a quick look at the weights of passengers on a bus can infer that they are probably not on their way to the starting line of a marathon, **then you now understand the basic idea of the central limit theorem**.
- The rest is just details.
- And if you understand the central limit theorem, most forms of statistical inference will seem relatively intuitive.

Extraordinary power to generalize

- The core principle underlying the central limit theorem is that a large, properly drawn sample will resemble the population from which it is drawn.
- Obviously there will be variation from sample to sample, but the probability that any sample will deviate massively from the underlying population is very low.
- Lots of big people run marathons; there are likely to be hundreds of people who weigh over 200 pounds in any given race.
- But the majority of marathon runners are relatively thin.
- Thus, the likelihood that so many of the largest runners were randomly assigned to the same bus is very, very low.

Extraordinary power to generalize

- You could conclude with a reasonable degree of confidence that this was not the missing marathon bus.
- Yes, you could have been wrong, but probability tells us that most of the time you would have been right.
- That's the basic intuition behind the central limit theorem.
- When we add some statistical bells and whistles, we can quantify the likelihood that you will be right or wrong.

Extraordinary power to generalize

- For example, we might calculate that in a marathon field of 10,000 runners with a mean weight of 155 pounds, there is less than a 1 in 100 chance that a random sample of 60 of those runners (our lost bus) would have a mean weight of 220 pounds or more.
- The central limit theorem enables us to make the following inferences, all of which will be explored in greater depth in the next chapter.

RESEARCH
Khatam University
FANAP

central limit theorem

1. If we have detailed information about some population, then we can make powerful inferences about any properly drawn sample from that population.
2. If we have detailed information about a properly drawn sample (mean, stddev), we can make strikingly accurate inferences about the population from which that sample was drawn.
3. If we have data describing a particular sample, and data on a particular population, we can infer whether or not that sample is consistent with a sample that is likely to be drawn from that population.
4. If we know the underlying characteristics of two samples, we can infer whether or not both samples were likely drawn from the same population.

Example

- For example, assume that a school principal has detailed information on the standardized test scores for all the students in his school (mean, stddev).
- That is the relevant population.
- Now assume that a bureaucrat from the school district will be arriving next week to give a similar standardized test to 100 randomly selected students.
- **The performance of those 100 students, the sample, will be used to evaluate the performance of the school overall.**
- How much confidence can the principal have that the performance of those randomly chosen 100 students will accurately reflect how the entire student body has been performing on similar standardized tests? **Quite a bit**.
- **According to the central limit theorem, the average test score for the random sample of 100 students will not typically deviate sharply from the average test score for the whole school.**

RESEARCH

Example

- This is essentially working in the opposite direction from the example above, putting ourselves in the shoes of the school district bureaucrat who is evaluating various schools in the district.
- Unlike the school principal, this bureaucrat does not have (or does not trust) the standardized test score data that the principal has for all the students in a particular school, which is the relevant population.
- Instead, he will be administering a similar test of his own to a random sample of 100 students in each school.

central limit theorem

- Can this administrator be reasonably certain that the overall performance of any given school can be evaluated fairly based on the test scores of a sample of just 100 students from that school?
- Yes.
- The central limit theorem tells us that a large sample will not typically deviate sharply from its underlying population

central limit theorem

- Can this administrator be reasonably certain that the overall performance of any given school can be evaluated fairly based on the test scores of a sample of just 100 students from that school? Yes.

- The central limit theorem tells us that a large sample will not typically deviate sharply from its underlying population.

- A methodologically sound poll of 1,200 Americans can tell us a great deal about how the entire country is thinking.

Example

- We know the mean weight for the marathoners.
- And we know the mean weight for the passengers on the broken-down bus.
- The central limit theorem enables us to calculate the probability that a particular sample (people on the bus) was drawn from a given population (the marathon field).
- If that probability is low, then we can conclude with a high degree of confidence that the sample was not drawn from the population in question
- (e.g., the people on this bus really don't look like a group of marathon runners headed to the starting line).

Extraordinary power to generalize

- We now know that a marathon is going on in the city, as well as the International Festival of Sausage.
- Assume that both groups have thousands of participants, and that both groups are operating buses, all loaded with random samples of either marathon runners or sausage enthusiasts.
- Further assume that two buses collide.
- In your capacity as a civic leader, you arrive on the scene and are tasked with determining whether or not both buses were headed to the same event (sausage festival or marathon)

Extraordinary power to generalize

- Suppose that the average weight of the passengers on one bus is 157 pounds, with a standard deviation of 11 pounds
- (meaning that a high proportion of the passengers weigh between 146 pounds and 168 pounds).
- Now suppose that the passengers on the second bus have a mean weight of 211 pounds with a standard deviation of 21 pounds
- (meaning that a high proportion of the passengers weigh between 190 pounds and 232 pounds).
- Does it seem likely that the passengers on those two buses were randomly drawn from the same population?
- you can also see that the variation in weights *between* the two buses is very large compared with the variation in weights *within* each bus.

Extraordinary power to generalize

- These passengers have a mean weight that is twenty-one pounds higher than the mean weight for the marathon field, something that should only occur by chance less than 1 time in 100.
- As a result, I can reject the hypothesis that this is a missing marathon bus with 99 % confidence—meaning I should expect my inference to be correct 99 times out of 100.
- And yes, probability suggests that on average I'll be *wrong* 1 time in 100.

Extraordinary power to generalize

- According to the central limit theorem, the sample means for any population will be distributed roughly as a normal distribution around the population mean.

1. Suppose we have a population, like our marathon field, and we are interested in the weights of its members.

Any sample of runners, such as each bus of sixty runners, will have a mean.

2. If we take repeated samples, such as picking random groups of sixty runners from the field over and over, then each of those samples will have its own mean weight.

These are the sample means.

Extraordinary power to generalize

- 3. Most of the sample means will be very close to the population mean.
- Some will be a little higher.
- Some will be a little lower.
- Just as a matter of chance, a very few will be significantly higher than the population mean, and a very few will be significantly lower.

- **4. The central limit theorem tells us that the sample means will be distributed roughly as a normal distribution around the population mean.**

Extraordinary power to generalize

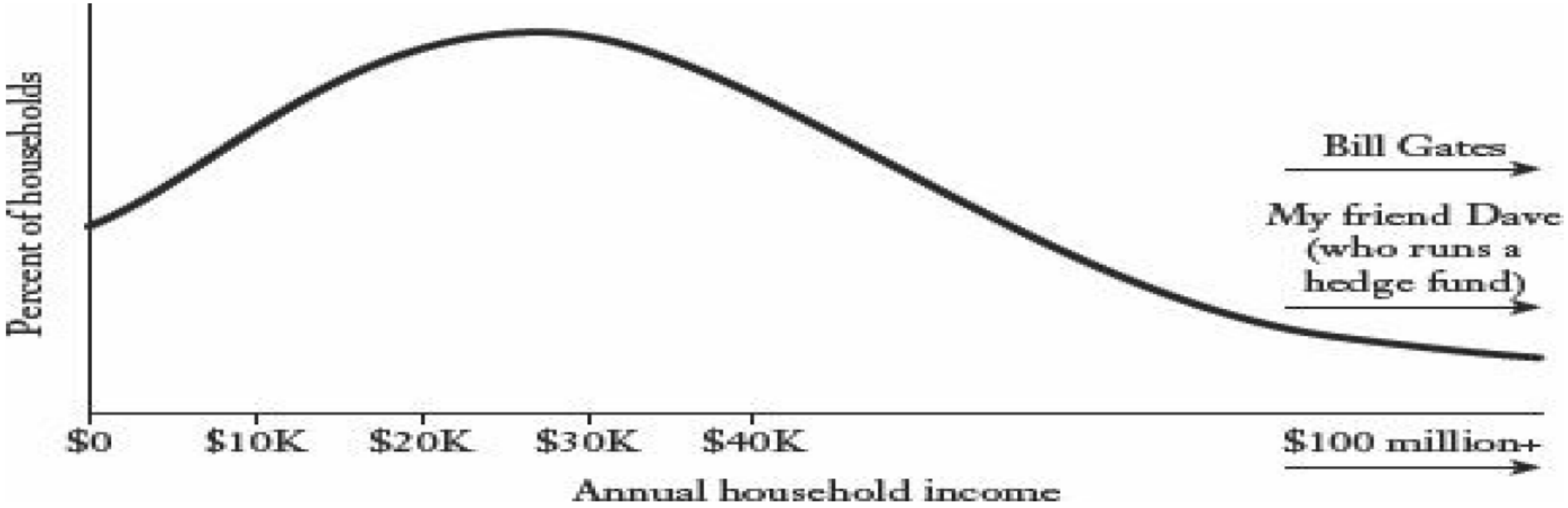
- 68% of the observations lie within **1** standard deviation of the mean,
- 95% lie within **2** standard deviations, and so on.

4. All of this will be true no matter what the distribution of the underlying population looks like.

The population from which the samples are being drawn does not have to have a normal distribution in order for the sample means to be distributed normally.

Extraordinary power to generalize

- Let's think about some real data, say, the household income distribution in the United States.
- Household income is not distributed normally in America; instead, it tends to be skewed to the right.
- No household can earn less than $0 in a given year, so that must be the lower bound for the distribution.
- Meanwhile, a small group of households can earn staggeringly large annual incomes—hundreds of millions or even billions of dollars in some cases.
- **As a result, we would expect the distribution of household incomes to have a long right tail**—something like this:



The median household income in is roughly $51,900;

The mean household income is $70,900.[1]

(People like Bill Gates pull the mean household income to the right)

Extraordinary power to generalize

- Now suppose we take a random sample of 1,000 U.S. households and gather information on annual household income.
- On the basis of the information above, and the central limit theorem, what can we infer about this sample?

- **Quite a lot, it turns out.**

- First of all, our best guess for what the mean of any sample will be is the mean of the population from which it's drawn.
- The whole point of a representative sample is that it looks like the underlying population.

Extraordinary power to generalize

- A properly drawn sample will, on average, look like America.
- There will be hedge fund managers and homeless people and police officers and everyone else—all roughly in proportion to their frequency in the population.
- Therefore, we would expect the mean household income for a representative sample of 1,000 American households to be about $70,900.
- Will it be exactly that?
- **No. But it shouldn't be wildly different either.**

Extraordinary power to generalize

- If we took multiple samples of 1,000 households, we would expect the different sample means to cluster around the population mean, $70,900.
- We would expect some means to be higher, and some to be lower.
- Might we get a sample of 1,000 households with a mean household income of $427,000?
- Sure, that's possible—but highly unlikely.
- (Remember, our sampling methodology is sound (healthy);

Extraordinary power to generalize

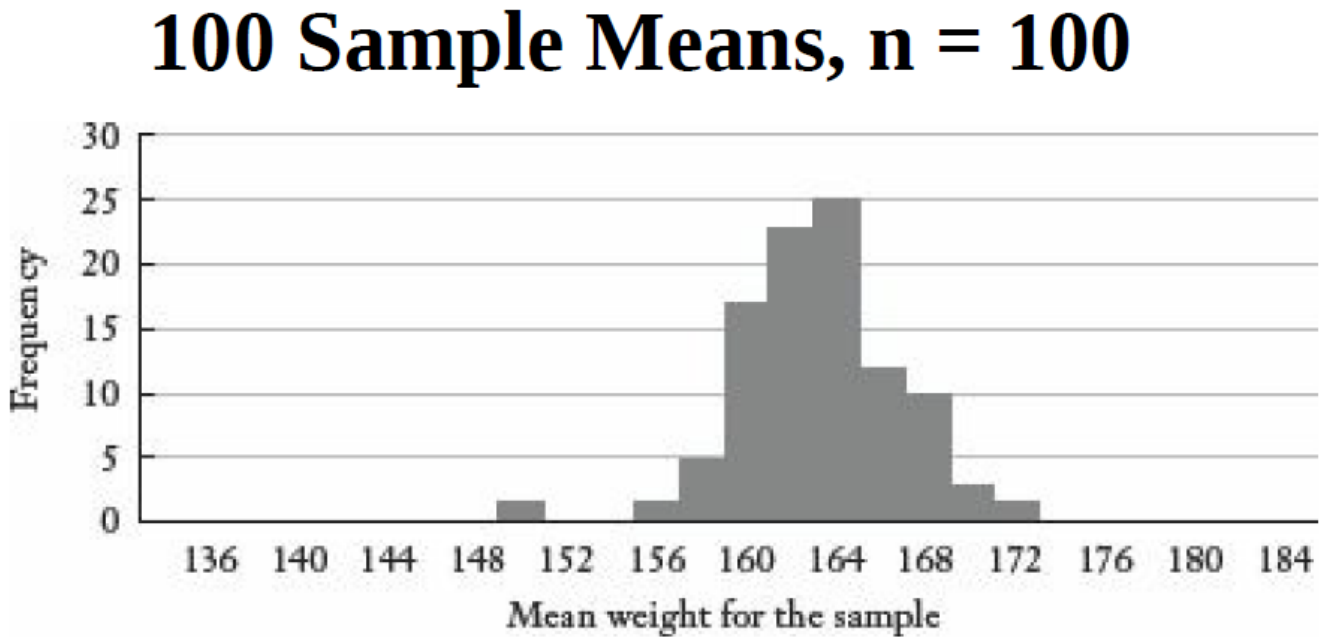
- we are not conducting a survey in the parking lot of the Greenwich Country Club.)
- It's also highly unlikely that a proper sample of 1,000 American households would have a mean income of $8,000.
- That's all just basic logic.
- The central limit theorem enables us to go one step further by describing the expected distribution of those different sample means as they cluster around the population mean.
- Specifically, the sample means will form a normal distribution around the population mean, which in this case is $70,900.

Extraordinary power to generalize

- Remember, the shape of the underlying population doesn't matter.
- The household income distribution in the United States is plenty skewed, *but the distribution of the sample means will not be skewed*.
- If we were to take 100 different samples, each with 1,000 households, and plotted the frequency of our results, we would expect those sample means to form the familiar "bell-shaped" distribution around $70,900.
- The larger the number of samples, the more closely the distribution will approximate the normal distribution.
- And the larger the size of each sample, the tighter that distribution will be.

Americans' Changing Lives

- The University of Michigan conducts a longitudinal study called Americans' Changing Lives, which consists of detailed observations on several thousand American adults, including their weights.
- The weight distribution is skewed slightly right, because it's biologically easier to be 100 pounds overweight than it is to be 100 pounds underweight.
- The mean weight for all adults in the study is 162 pounds.



Americans' Changing Lives

- The larger the sample size and the more samples taken, the more closely the distribution of sample means will approximate the normal curve.
- (As a rule of thumb, the sample size must be at least 30 for the central limit theorem to hold true.)
- This makes sense.
- A larger sample is less likely to be affected by random variation.
- A sample of 2 can be highly skewed by 1 particularly large or small person.
- In contrast, a sample of 500 will not be unduly affected by a few particularly large or small people.

RESEARCH

Americans' Changing Lives

- We are now very close to making all of our statistical dreams come true!
- The sample means are distributed roughly as a normal curve, as described above.
- The power of a normal distribution derives from the fact that
- we know roughly what proportion of observations will lie within one standard deviation above or below the mean (68 percent);
- what proportion of observations will lie within two standard deviations above or below the mean (95 percent); and so on.

The standard error

- To quantify that intuition—to be able to say that this inference will be correct 95 % of the time, or 99 percent, or 99.9 percent—we need just one more technical concept: the standard error.
- The standard error measures the dispersion of the sample means.
- How tightly do we expect the sample means to cluster around the population mean?
- There is some potential confusion here, as we have now introduced two different measures of dispersion: the standard deviation and the standard error.

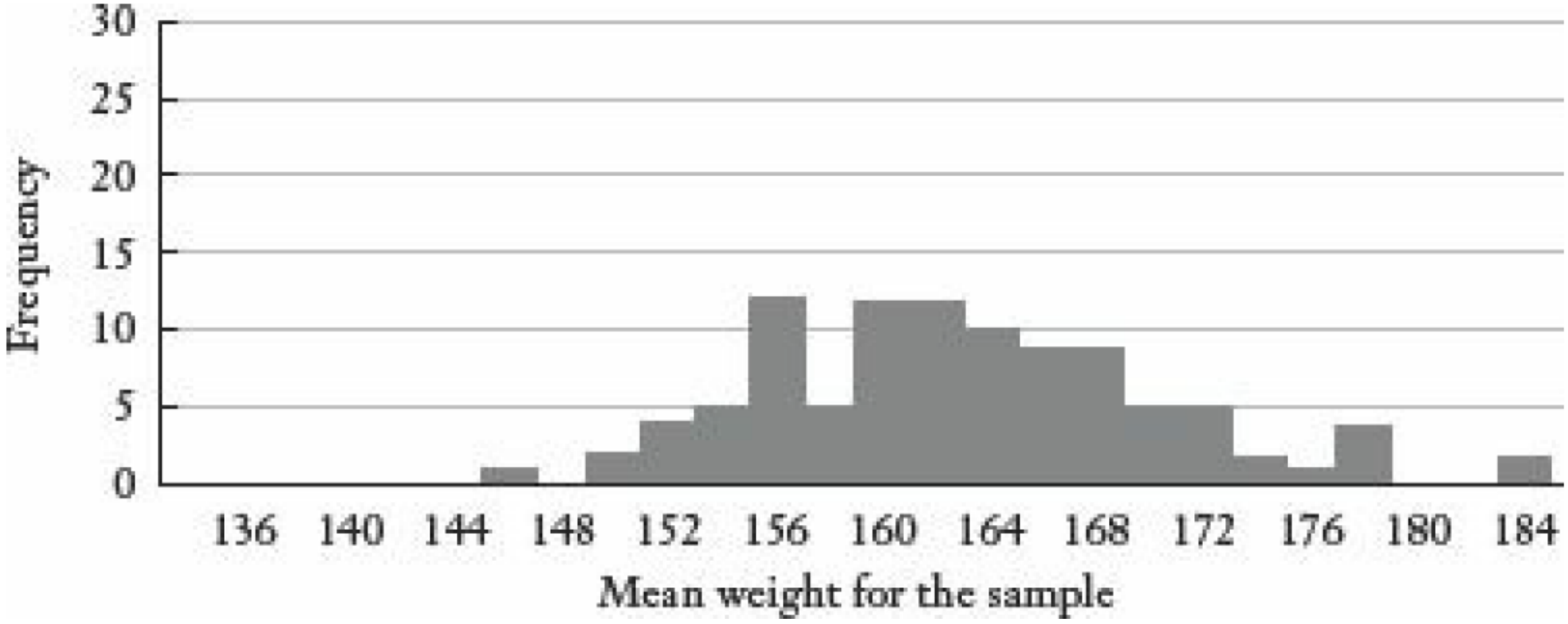
the standard error

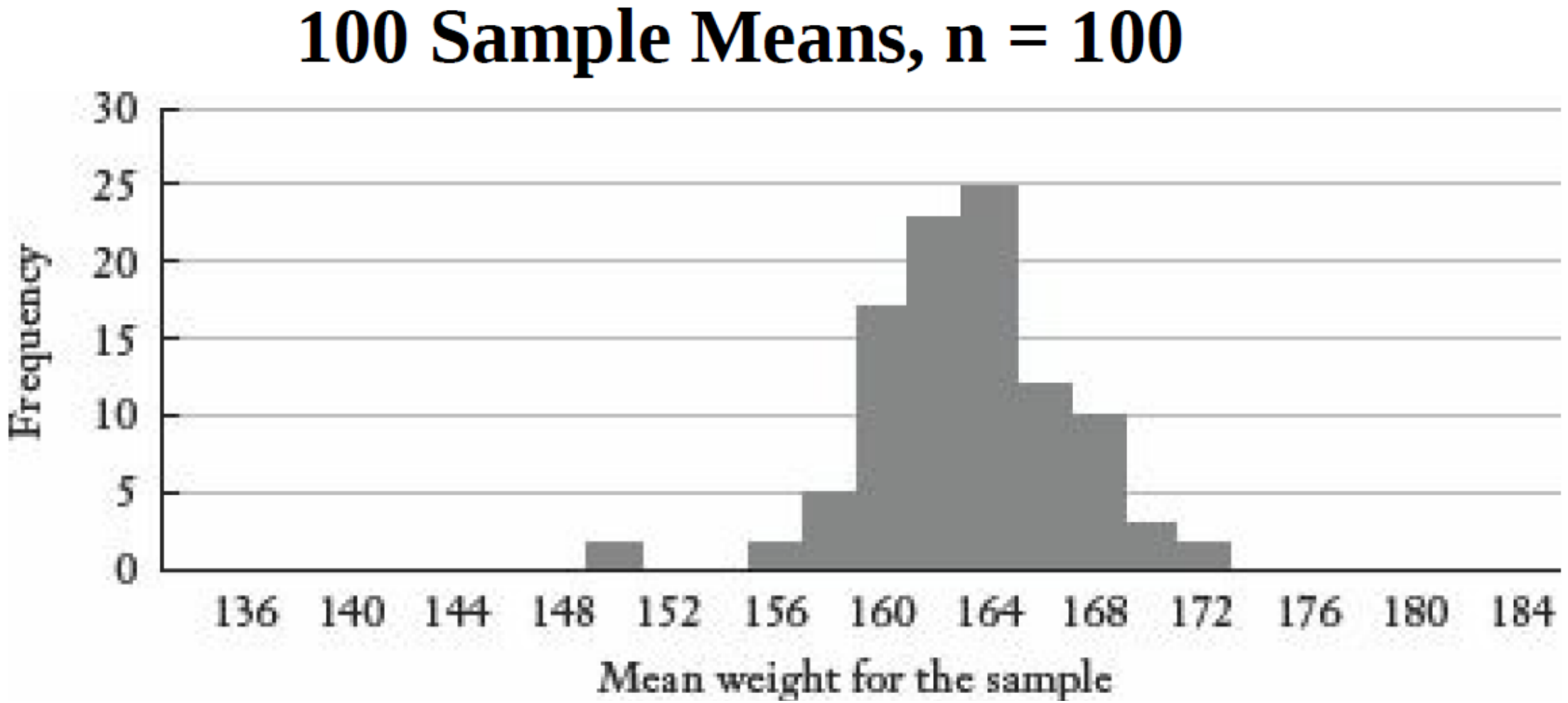
- Here is what you need to remember to keep them straight:
- 1. The standard deviation measures dispersion in the underlying population.
- In this case, it might measure the dispersion of the weights of all the participants in the Framingham Heart Study, or the dispersion around the mean for the entire marathon field.
- 2. The standard error measures the dispersion of *the sample means*.
- **If we draw repeated samples of 100 participants from the Framingham Heart Study, what will the dispersion of those sample means look like? Exercise**

the standard error

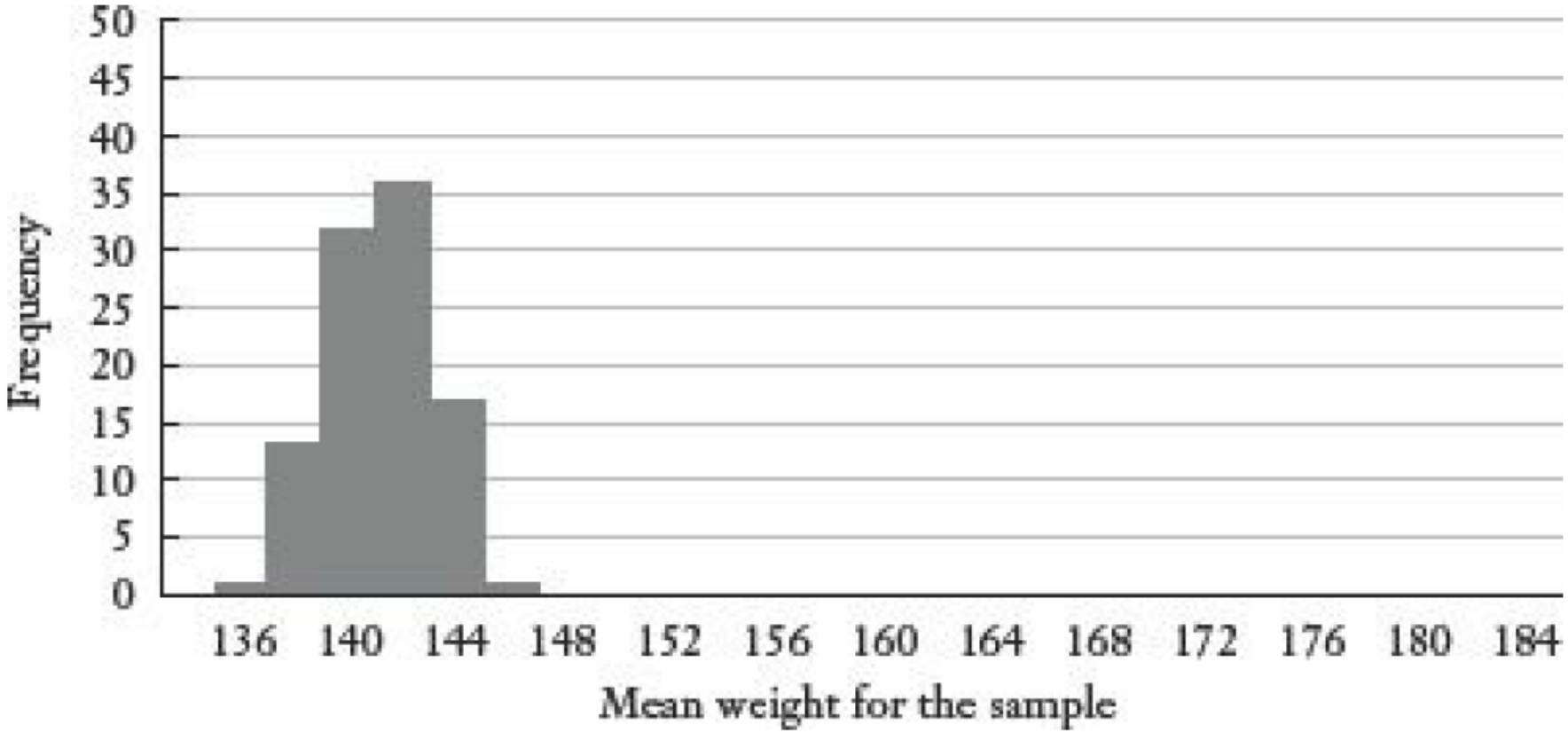
- 3. Here is what ties the two concepts together: The standard error is the standard deviation of the sample means!
- • A large standard error means that the sample means are spread out widely around the population mean;
- • a small standard error means that they are clustered relatively tightly.
- Here are three real examples from the Changing Lives data.

100 Sample Means, n = 20




100 Sample Means, n = 100
Frequency
30
25
20
15
10
5
0
136
140
144
148
152
156
160
164
168
172
176
180
184
Mean weight for the sample

Female Population Only/100 Sample Means, n = 100



the standard error

- The second distribution, which has a larger sample size, is more tightly clustered around the mean than the first distribution.
- The larger sample size makes it less likely that a sample mean will deviate sharply from the population mean.
- The final set of sample means is drawn only from a subset of the population, women in the study.
- Since the weights of women in the data set are less diffuse/spread than the weights of all persons in the population, it stands to reason that the weights of samples drawn just from the women would be less dispersed than samples drawn from the whole Changing Lives population.

the standard error

- (These samples are also clustered around a slightly different population mean, since the mean weight for all females in the Changing Lives study is different from the mean weight for the entire population in the study.)

- The pattern that you saw above holds true in general.
- Sample means will cluster more tightly around the population mean as the size of each sample gets larger (e.g., samples of 100 rather than 30).

the standard error

- And the sample means will cluster *less tightly* around the population mean when the underlying population is more spread out (e.g., sample means for the entire population were more dispersed than the sample means for just the females in the study).
- If you've followed the logic this far, then the formula for the standard error follows naturally:

$$std\ error = \frac{s}{\sqrt{n}}$$

- where s is the standard deviation of the population from which the sample is drawn, and n is the size of the sample.

the standard error

- • The standard error will be large when the standard deviation of the underlying distribution is large.
- • A large sample drawn from a highly dispersed population is also likely to be highly dispersed;
- • a large sample from a population clustered tightly around the mean is also likely to be clustered tightly around the mean.

$$std\ error = \frac{s}{\sqrt{n}}$$

the standard error

- If we are still looking at weight, we would expect the standard error for a sample drawn from the entire Changing Lives population to be larger than the standard error for a sample drawn only from the men in their twenties.
- **This is why the standard deviation (s) is in the numerator.**
- Similarly, we would expect the standard error to get smaller as the sample size gets larger, since large samples are less prone to distortion by extreme outliers.
- **This is why the sample size (n) is in the denominator.**
- **(The reason we take the square root of n will be left for a more advanced text; the basic relationship is what's important here.)**

example:

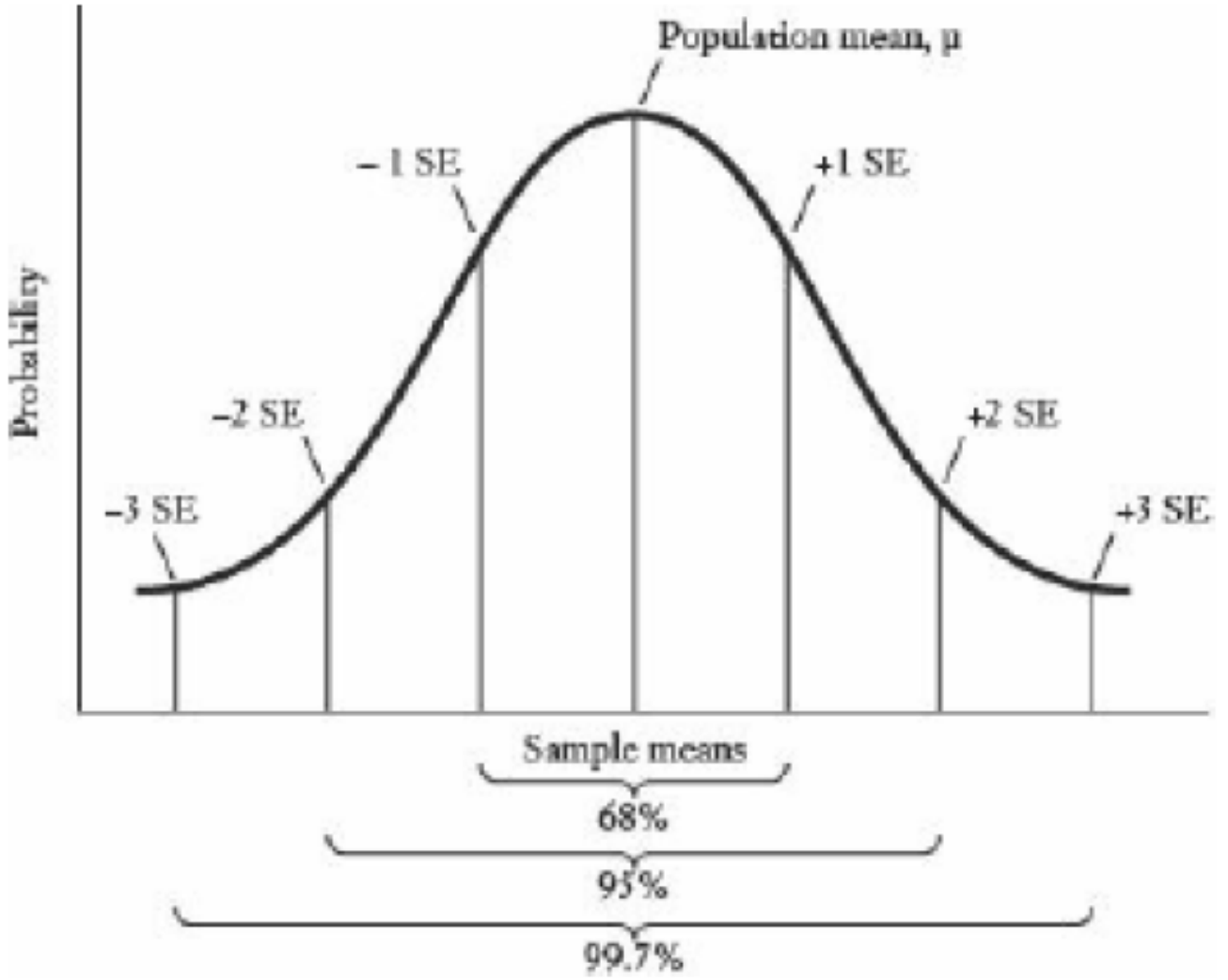
- In the case of the Changing Lives data, **we actually know the standard deviation of the population**; often that is not the case.
- For large samples, we can assume that the standard deviation of the sample is reasonably close to the standard deviation of the population.*
- Finally, we have arrived at the payoff for all of this.
- Because the sample means are distributed normally (thanks to the central limit theorem), we can harness the power of the normal curve.

$$std\ error = \frac{s}{\sqrt{n}}$$

example:

- We expect that roughly
- • **68** % of all sample means will lie within **ONE** standard error of the population mean;
- • **95** % of the sample means will lie within **two** standard errors of the population mean;
- • and **99.7** % of the sample means will lie within **three** standard errors of the population mean.

Frequency Distribution of Sample Means



example:

- you see a crashed bus on the side of the road.
- Apparently the bus swerved to miss a wild fox crossing the road, and all of the passengers are unconscious but not seriously hurt.
- Paramedics on the scene inform you that the mean weight of the 62 passengers on the bus is 194 pounds.
- Fortunately you know the mean weight 162 and standard deviation 36 for the entire Changing Lives population, you have a working knowledge of the central limit theorem.
- From that information, we can calculate the standard error for a 62-person sample (the number of unconscious passengers on the bus):

$$std\ error = \frac{36}{\sqrt{62}} = 4.6$$

example:

- The difference between the sample mean (194 pounds) and the population mean (162 pounds) is 32 pounds>6*SE, or well more than three standard errors.
- We know from the central limit theorem that 99.7 % of all sample means will lie within three standard errors of the population mean.
- **That makes it extremely unlikely that this bus represents a random group of Changing Lives participants.**
- In your duty as a civic leader, you call the study officials to tell them that this is probably not their missing bus, only now you can offer statistical evidence, rather than just “a guess.”
- You report to the Changing Lives folks that you can reject the possibility that this is the missing bus at the 99.7 % confidence level.

example:

- Your analysis is further confirmed when paramedics conduct blood tests on the bus passengers and discover that the mean cholesterol level for the bus load of passengers is **five standard errors** above the mean cholesterol level for the Changing Lives study participants.
- That suggests, correctly it later turns out, that the unconscious passengers are involved with the Festival of Sausage.
- I've tried to stick with the basics in this chapter.
- You should note that for the central limit theorem to apply, the sample sizes need to be relatively large (over 30 as a rule of thumb).

example:

- We also need a relatively large sample if we are going to assume that the standard deviation of the sample is roughly the same as the standard deviation of the population from which it is drawn.
- There are plenty of statistical fixes that can be applied when these conditions are not met—but that is all frosting on the cake (and maybe even sprinkles on the frosting on the cake).

The “big picture”

- 1. If you draw large, random samples from any population, the means of those samples will be distributed normally around the population mean (regardless of what the distribution of the underlying population looks like).
- 2. Most **sample means** will lie reasonably close to the **population mean**; the standard error is what defines “reasonably close.”
- 3. The central limit theorem tells us the probability that a sample mean will lie within a certain distance of the population mean.
- It is relatively unlikely that a sample mean will lie more than two standard errors from the population mean, and extremely unlikely that it will lie three or more standard errors from the population mean.

The “big picture”

- **4. The less likely it is that an outcome has been observed by chance, the more confident we can be in surmising (inferring) that some other factor is in play.**
- That’s pretty much what statistical inference is about.
- The central limit theorem is what makes most of it possible.
- And until Lebron James wins as many NBA championships as Michael Jordan (six), the central limit theorem will be far more impressive than he is.

Inference

- *Why my statistics professor thought I might have cheated*

Example

- “You really did much better on the final than you did on the midterm.”
- This was not a congratulatory visit during which I was recognized for finally doing serious work in the class.
- There was an implicit accusation (though not an explicit one) in his summons; the expectation was that I would explain why I did so much better on the final exam than the midterm.

Example, why????????

- I'm more sympathetic to his line of thinking.
- In nearly every course I've taught, there is a striking degree of correlation between a student's performance on the midterm and on the final.
- It *is* highly unusual for a student to score below average on the midterm and then near the top of the class on the final.

Example, why????????

- Believe it or not, this anecdote embodies much of what you need to know about statistical inference, including both its strengths and its potential weaknesses.
- *Statistics cannot prove anything with certainty*.
- **Instead, the power of statistical inference derives from observing some pattern or outcome and then using probability to determine the most likely explanation for that outcome.**

Example 2

- Suppose a strange gambler arrives in town and offers you a wager:
- He wins $1,000 if he rolls a six with a single die;
- you win $500 if he rolls anything else—a pretty good bet from your standpoint.
- He then proceeds to roll ten sixes in a row, taking $10,000 from you.
  - One possible explanation is that he was lucky.
  - An alternative explanation is that he cheated somehow.
- The probability of rolling ten sixes in a row with a fair die is roughly 1 in 60 million.
- You can't prove that he cheated, but you ought at least to inspect the die.

Example 2

- Of course, the most likely explanation is not always the right explanation.
- Extremely rare things happen.
- Linda Cooper has been struck by lightning four times.
- (The probability of getting hit by lightning=just once as 1 in 600,000.)
- Linda Cooper's insurance company cannot deny her coverage simply because her injuries are statistically improbable.

- But an unlikely pattern is just an unlikely pattern unless it is corroborated/ confirmed by additional evidence.

Example 3

- For now, we should appreciate that statistical inference uses data to address important questions.
  - Is a new drug effective in treating heart disease?
  - Do cell phones cause brain cancer?
- Please note that **I'm not claiming that statistics can answer** these kinds of questions clearly ;
- instead, **inference tells us what is likely, and what is unlikely.**

Example 3

- Researchers cannot prove that a new drug is effective in treating heart disease, even when they have data from a carefully controlled clinical trial.
- After all, it is entirely possible that there will be random variation in the outcomes of patients in the treatment and control groups that are unrelated to the new drug.
- If 53 out of 100 patients taking the new heart disease medication showed marked improvement compared with 49 patients out of 100 receiving a placebo, we would not immediately conclude that the new medication is effective.
- This is an outcome that can easily be explained by chance variation between the two groups rather than by the new drug.

Example 3

- But suppose instead that 91 out of 100 patients receiving the new drug show marked improvement, compared with 49 out of 100 patients in the control group.
- It is still possible that this impressive result is unrelated to the new drug;
- the patients in the treatment group may be particularly lucky or resilient.
- *But that is now a much less likely explanation*.

taking the new heart disease medication --- receiving a placebo

53 out of 100 --- 49 patients out of 100

91 out of 100 --- 49 out of 100

easily be explained by chance --- lucky/resilient? **less likely**

Example 3

- In the formal language of statistical inference, researchers would likely conclude the following:
- (1) If the experimental drug has no effect, we would rarely see this amount of variation in outcomes between those groups
- (2) It is therefore highly improbable (UNLIKELY) that the drug has no positive effect.
- (3) The alternative—and more likely—explanation for the pattern of data observed is that the experimental drug has a positive effect.

Statistical inference

- Statistical inference is the process by which the data speak to us, enabling us to draw meaningful conclusions.
- This is the payoff!
- The point of statistics is not to do complicated mathematical calculations; the point is to gain insight into meaningful social phenomena.
- Statistical inference is really just the marriage of two concepts that we've already discussed: data and probability (with a little help from the central limit theorem).

Statistical inference, **central limit theorem**

- **All of the examples will assume that we are working with large, properly drawn samples, it means that the central limit theorem applies, and that the mean and standard deviation for any sample will be roughly the same as the mean and standard deviation for the population from which it is drawn.**
- Both of these things make our calculations easier.

hypothesis testing

- statistics alone cannot *prove* anything; instead, we use statistical inference to accept or reject explanations on the basis of their relative likelihood.
- Any statistical inference begins with an implicit or explicit **null hypothesis.**
- This is our starting assumption, which will be rejected or not on the basis of subsequent statistical analysis.
- If we reject the **null hypothesis**, then we typically accept some alternative hypothesis that is more consistent with the data observed.
- As a matter of logic, the **alternative hypothesis** is a conclusion that must be true if we can reject the null hypothesis.

Courtroom

- Null hypothesis: **the defendant is innocent**.
- Alternative hypothesis: **the defendant is guilty**.

- The job of the prosecution is to persuade the judge or jury to reject that assumption **(innocent)** and accept the alternative hypothesis, which is that the defendant is guilty.

New drug

- Null hypothesis: This new experimental drug is no more effective at preventing malaria than a placebo.
- Alternative hypothesis: This new experimental drug can help to prevent malaria.
- The data: At the end of some period of time, the group receiving the experimental drug has far fewer cases of malaria than the control group.
- This would be an extremely unlikely outcome if the experimental drug had no medical impact.
- As a result, we ***reject the null hypothesis*** **(**no impact**)**, and we ***accept the alternative hypothesis***: This new experimental drug can help to prevent malaria.

Substance abuse treatment

- The null hypothesis and alternative hypothesis are logical complements.
- If one is true, the other is not true.
- Or, if we reject one statement, we must accept the other.
- Null hypothesis: Substance abuse treatment for prisoners does not reduce their re-arrest rate after leaving prison.
- Alternative hypothesis: Substance abuse treatment for prisoners will make them less likely to be rearrested after they are released.
- The (hypothetical) data: Prisoners were randomly assigned into two groups; the “treatment” group received substance abuse treatment and the control group did not.

Substance abuse treatment

- At the end of five years, both groups have similar rearrest rates.
- In this case, we *cannot reject* the null hypothesis.
- **WE ACCEPT** that substance abuse treatment is not an effective tool for keeping ex-offenders from returning to prison.
- **It may seem counterintuitive, but researchers often create a null hypothesis in hopes of being able to reject it**.
- In both of the examples above, a research "success" (finding a new malaria drug or reducing recidivism involved rejecting the null hypothesis.
- The data made that possible in only one of the cases (the malaria drug).

Guilty beyond a reasonable doubt

- In a courtroom, the threshold for rejecting the presumption of innocence is the qualitative assessment that the defendant is “guilty beyond a reasonable doubt.”
- The judge or jury is left to define what exactly that means.
- Statistics harnesses the same basic idea, but “guilty beyond a reasonable doubt” is defined quantitatively instead.
- Researchers typically ask, If the null hypothesis is true, how likely is it that we would observe this pattern of data by chance?

guilty beyond a reasonable doubt

- Medical researchers might ask, If this experimental drug has no effect on heart disease (our null hypothesis), how likely is it that 91 out of 100 patients getting the drug would show improvement compared with only 49 out of 100 patients getting a placebo?
- If the data suggest that the null hypothesis is extremely unlikely—as in this medical example—then we must reject it and accept the alternative hypothesis (that the drug is effective in treating heart disease).

A high number of “wrong-to-right” erasures.

- Null hypothesis: the standardized test scores are legitimate and any irregular patterns of erasures are merely a product of chance.
- We certainly do not want to be punishing students or administrators because an unusually high proportion “wrong-to-right” erasures
- But “unusually high” cannot describe what was happening in Atlanta.
- Some classrooms had answer sheets on which the number of wrong-to-right erasures were **20** to **50** standard deviations above the state norm.
- (remember that most observations in a distribution typically fall within two standard deviations of the mean.)

a high number of “wrong-to-right” erasures.

- So how likely was it to erase massive numbers of wrong answers and replace them with correct answers just as a matter of chance?
- The official described the probability of the Atlanta pattern occurring without cheating as roughly equal to the chance of having 70,000 people show up for a football game who all happen to be over 213cm tall.

- **Could it happen? Yes.**
- **Is it likely? Not so much.**

a high number of “wrong-to-right” erasures.

- • *Atlanta officials could not prove that cheating was going on.*
- • They could reject the null hypothesis that the results were legitimate.
- And they could do so with a “high degree of confidence,” meaning that the observed pattern was nearly impossible among normal test takers.
- They therefore explicitly accepted the alternative hypothesis, which is that something fishy was going on.
- Subsequent investigation did in fact uncover the “smoking erasers.”
- a group of teachers who held a weekend pizza party during which they went through exam sheets and changed students’ answers.

a high number of “wrong-to-right” erasures.

- But how implausible/unlikley does the null hypothesis have to be before we can reject it and invite some alternative explanation?
- One of the most common thresholds that researchers use for rejecting a null hypothesis is 5 percent, **or** in decimal form: .05.

significance level

- This probability is known as a significance level, and it represents the upper bound for the likelihood of observing some pattern of data if the null hypothesis were true.
- Stick with me for a moment, because it's not really that complicated.
- Let's think about a significance level of .05.
- We can reject a null hypothesis at the .05 level if there is less than a 5% chance of getting an outcome at least as extreme as what we've observed if the null hypothesis were true.

hijacked bus

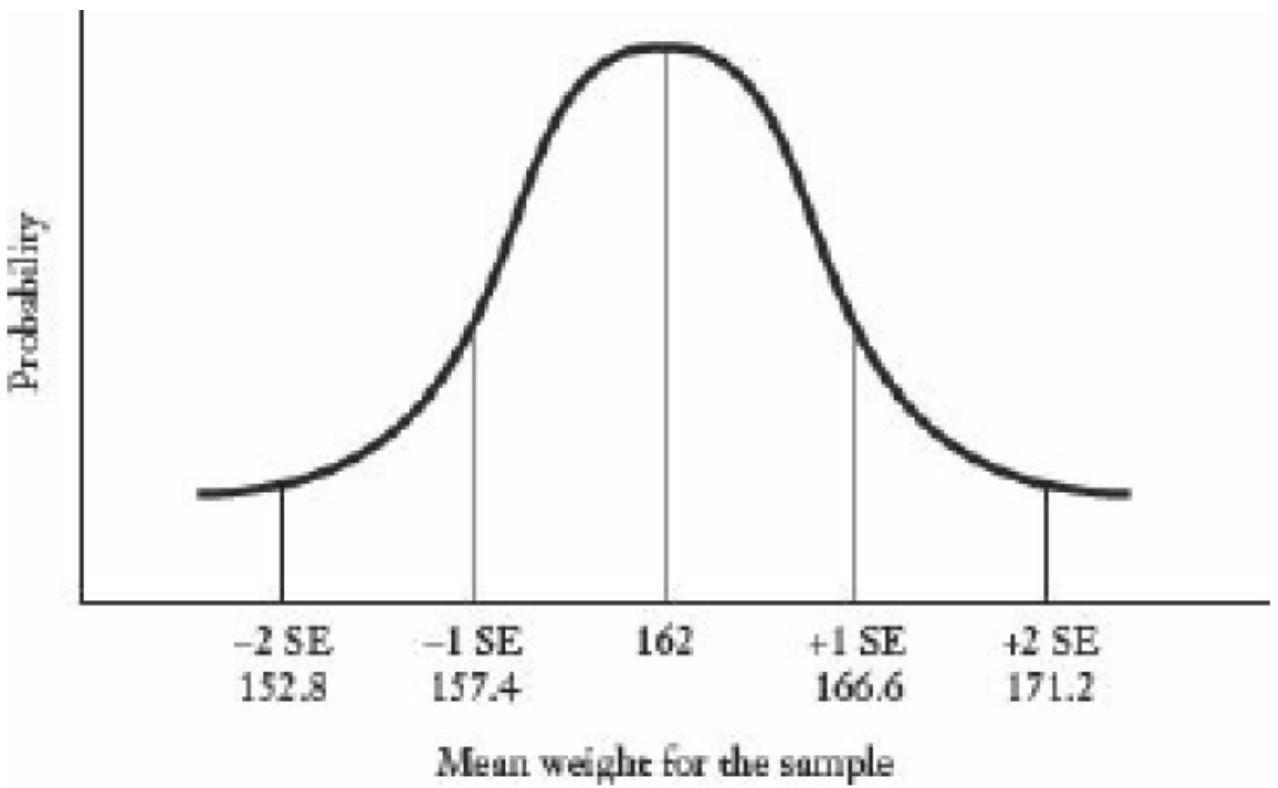
- We would expect roughly 95% of all 60-person samples drawn from the Changing Lives participants to have a mean weight within two SE of the population mean, or roughly between 153 pounds and 171 pounds.*

- **95% of the sample weights to be 162-2*4.6 to 162+2*4.6**

- Conversely, only 5 times out of 100 would a sample of 60 persons randomly drawn from the Changing Lives participants have a mean weight that is greater than 171 pounds or less than 153 pounds.

hijacked bus

- If the mean weight of the 60 passengers on the hijacked bus is above 171 or below 153, then you **will reject the null hypothesis** that the bus contains Changing Lives participants,
- **accept the alternative hypothesis** that the bus contains 60 people headed somewhere else, and await further orders.



The mean weight for this 60-person sample is 136 pounds, which falls **more than two SE** below the mean.

hijacked bus

- you can reject the null hypothesis that this bus contains a random sample of 60 Changing Lives study participants at the .05 significance level.
- This means
- (1) the mean weight on the bus falls into a range that we would expect to observe only 5 times in 100 if the null hypothesis were true and this were really a bus full of Changing Lives passengers;
- (2) you can reject the null hypothesis at the .05 significance level; and

hijacked bus

- (3) on average, 95 times out of 100 you will have correctly rejected the null hypothesis, and 5 times out of 100 you will be wrong, meaning that you have concluded that this is not a bus of Changing Lives participants, when in fact it is.

- This sample of Changing Lives folks just happens to have a mean weight that is particularly high or low relative to the mean for the study participants overall.

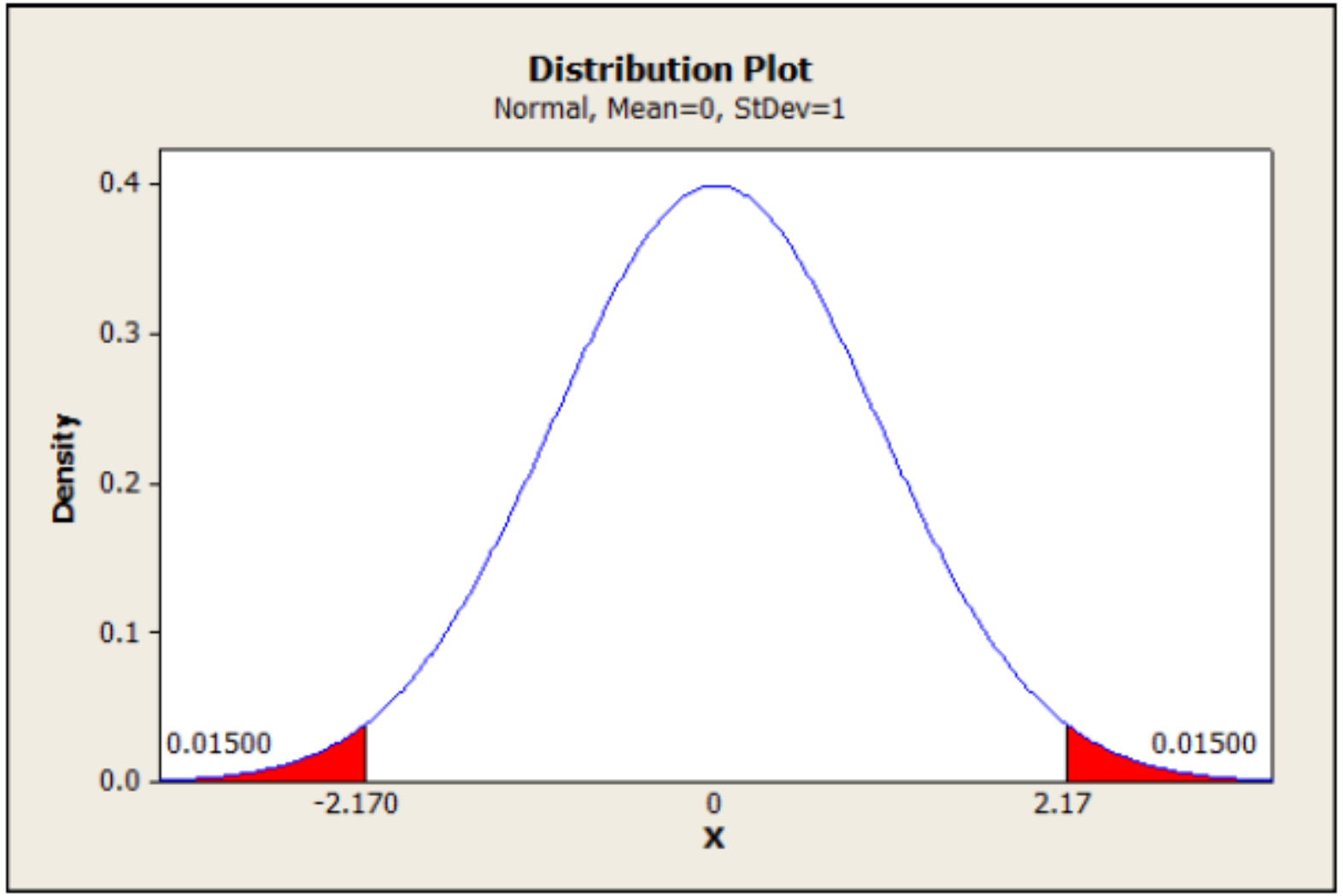
p-value

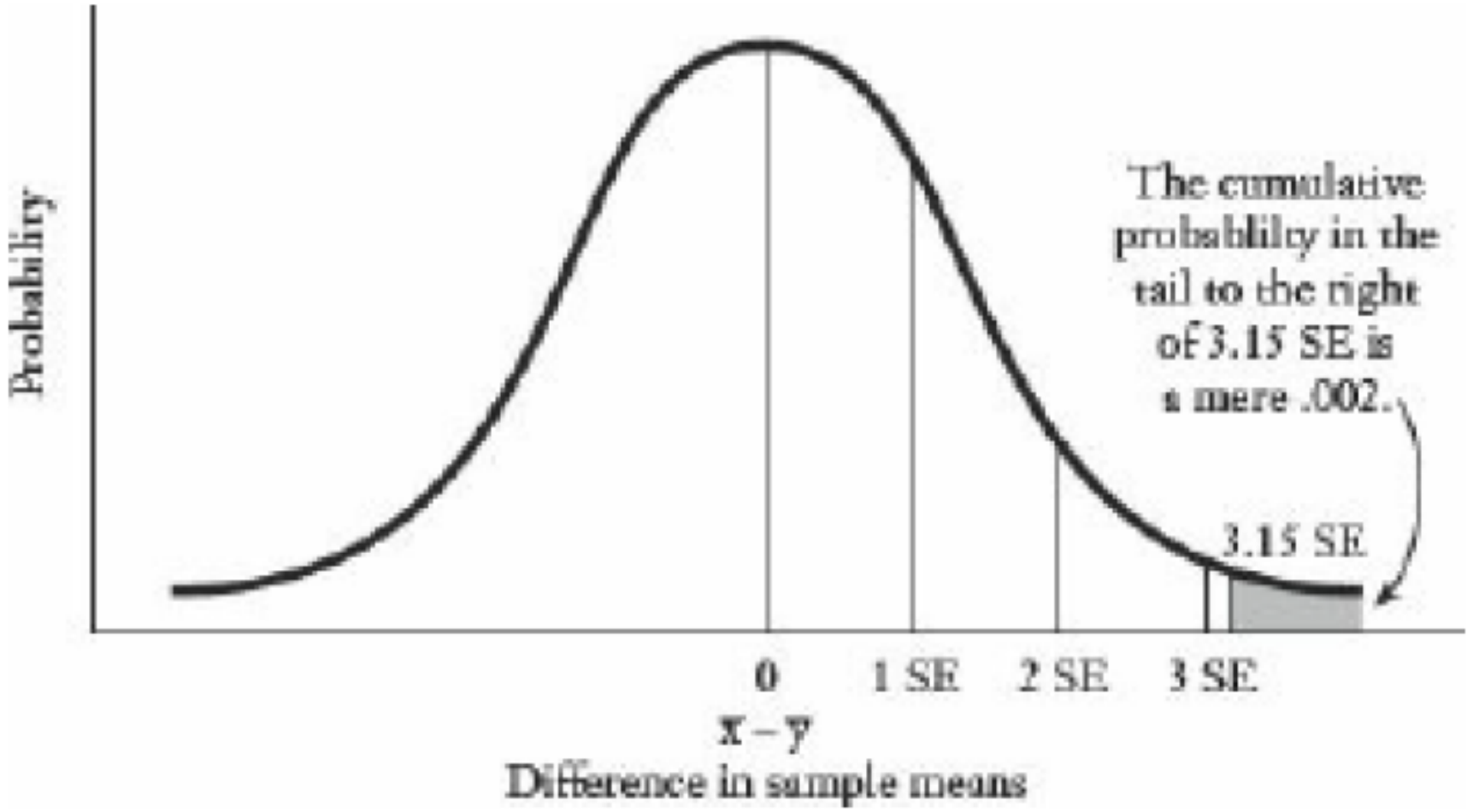
- The p-value is the specific probability of getting a result at least as extreme as the one you've observed if the null hypothesis is true.

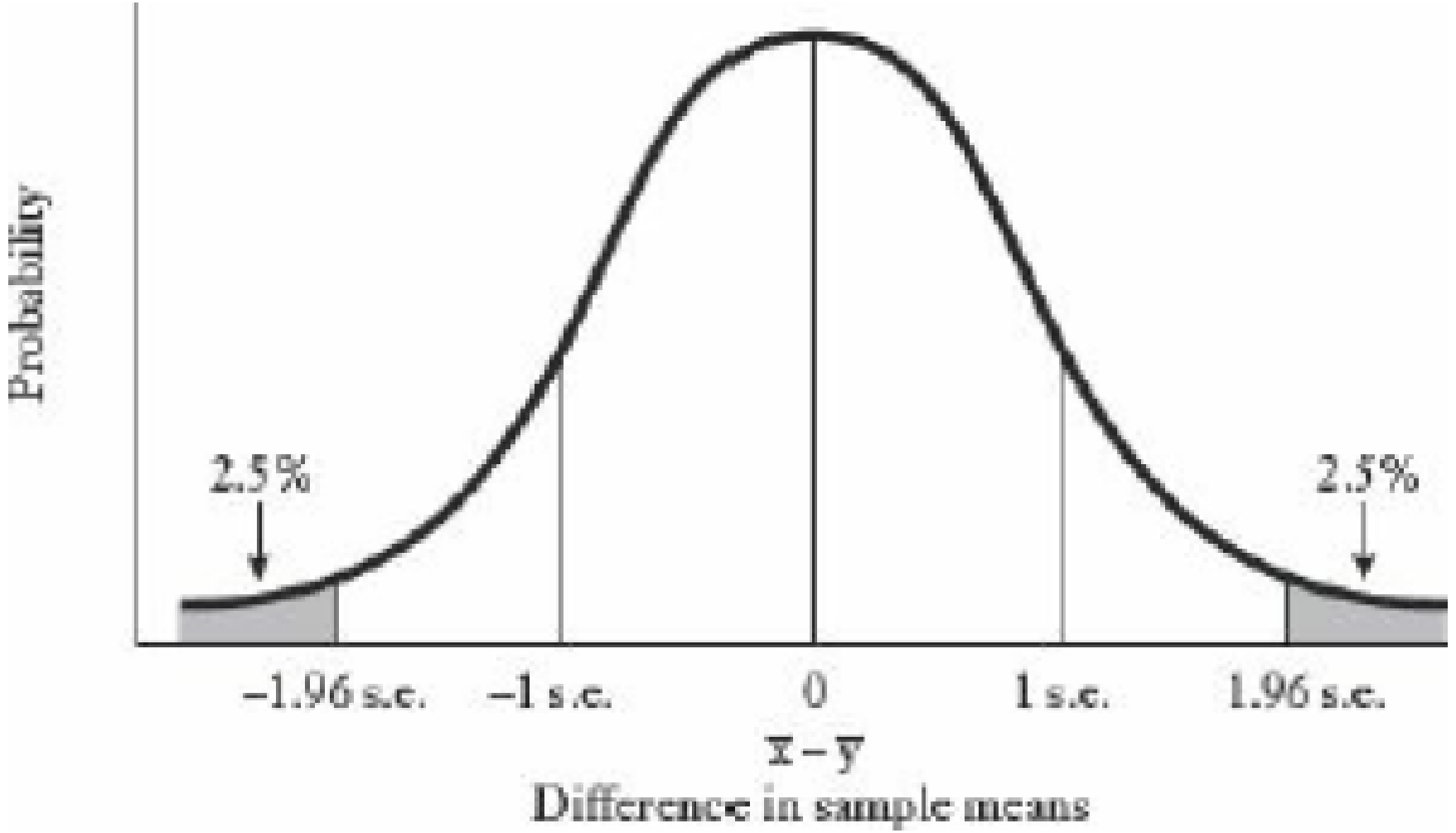
- The mean weight for the passengers on this bus is 136, which is 5.7 SE below the mean for the Changing Lives study participants.
- The probability of getting a result at least that extreme if this really were a sample of Changing Lives participants is less than .0001. ($p<.0001$.)

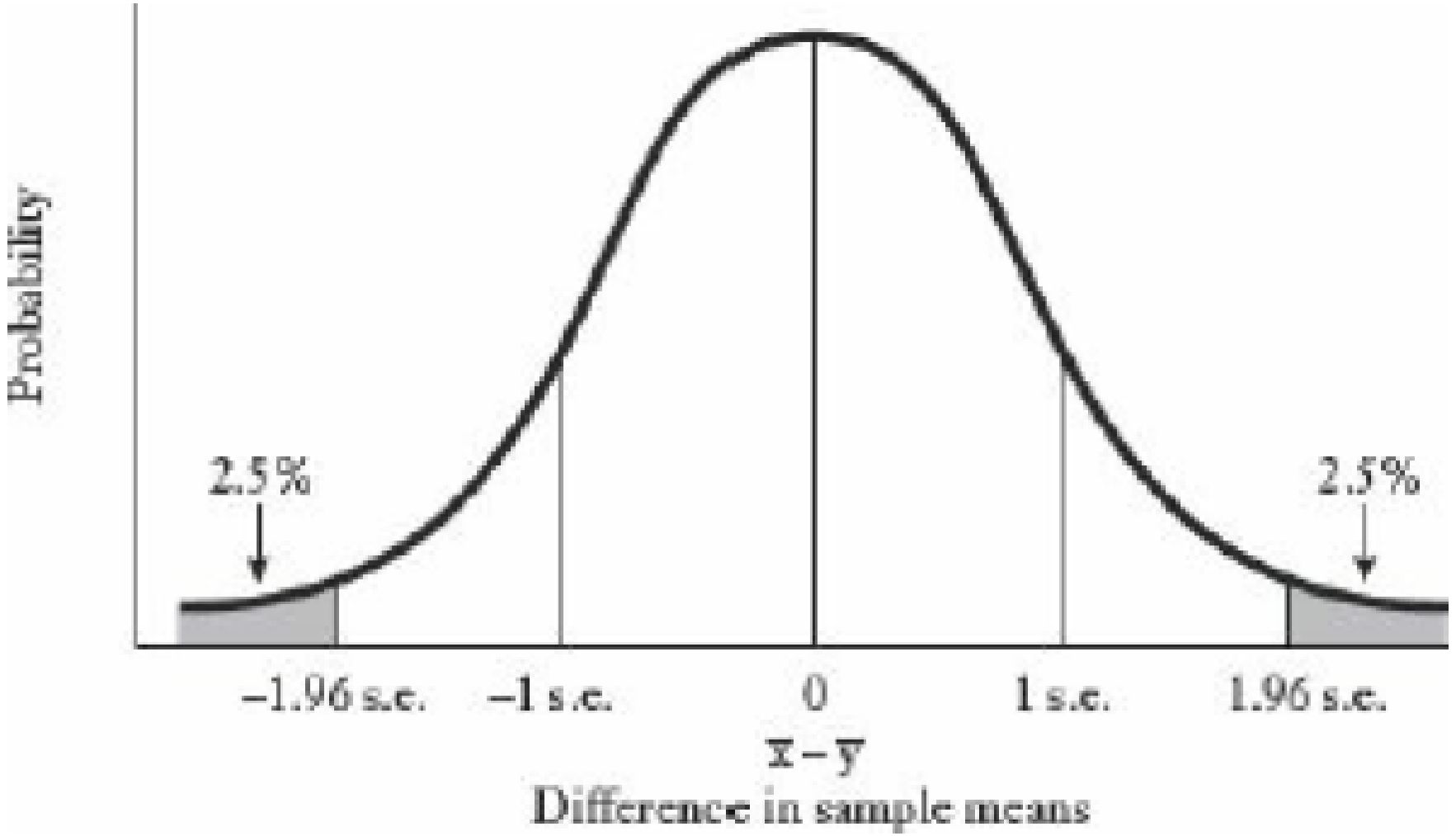
p-value

- If the .05 significance level seems somewhat arbitrary, that's because it is.
- There is no single standardized statistical threshold for rejecting a null hypothesis.
- **Both .01 and .1 are also reasonably common thresholds for doing the kind of analysis described above.**


Distribution Plot
Normal, Mean=0, StDev=1
Density
0.4
0.3
0.2
0.1
0.0
0.01500
0.01500
-2.170
0
2.17
X


Probability
The cumulative probability in the tail to the right of 3.15 SE is a mere .002.
3.15 SE
0
1 SE
2 SE
3 SE
x − y
Difference in sample means


Probability
2.5%
2.5%
−1.96 s.e.
−1 s.e.
0
1 s.e.
1.96 s.e.
x̄ − ȳ
Difference in sample means


Probability
2.5%
2.5%
−1.96 s.e.
−1 s.e.
0
1 s.e.
1.96 s.e.
x̄ − ȳ
Difference in sample means

p-value

- Obviously rejecting the null hypothesis at the 0.01 level carries more statistical weight than rejecting the null hypothesis at the 0.1 level.
- • 0.01 level: meaning that there is less than a 1 in 100 chance of observing a result in this range if the null hypothesis were true
- • 0.1 level: meaning that there is less than a 1 in 10 chance of observing this result if the null hypothesis were true.
- For now, the important point is that when we can reject a null hypothesis at some reasonable significance level, the results are said to be “statistically significant.”

Here is what that means in real life.

- When you read in the newspaper that people who eat twenty bran muffins a day have lower rates of colon cancer than people who don't eat huge amounts of bran, the underlying academic research probably looked something like this:
- (1) In some large data set, researchers determined that individuals who ate at least twenty bran muffins a day had a lower incidence of colon cancer than individuals who did not report eating much bran.
- (2) The researchers' null hypothesis was that eating bran muffins has no impact on colon cancer.

Here is what that means in real life.

- (3) The disparity (a great difference) in colon cancer outcomes between those who ate lots of bran muffins and those who didn't could not easily be explained by chance alone.
- (4) More specifically, if eating bran muffins has no true association with colon cancer, the probability of getting such a wide gap in cancer incidence between bran eaters and non–bran eaters by chance alone is lower than some threshold, such as 0.05.

Here is what that means in real life.

- (5) The academic paper probably contains a conclusion that says something along these lines: "**We find a statistically significant association between daily consumption of twenty or more bran muffins and a reduced incidence of colon cancer. These results are significant at the .05 level**."
- The headline is probably more direct and interesting: "20 Bran Muffins a Day Help Keep Colon Cancer Away."
- However, that newspaper headline, while much more interesting to read than the academic paper, may also be introducing a serious inaccuracy.

Here is what that means in real life.

- The study does not actually claim that eating bran muffins lowers an individual's risk of getting colon cancer;
- it merely shows a negative correlation between the consumption of bran muffins and the incidence of colon cancer in one large data set.
- This statistical association is not sufficient to prove that the bran muffins *cause* the improved health outcome.
- After all, the kind of people who eat bran muffins (particularly twenty a day!) may do lots of other things that lower their cancer risk, such as eating less red meat, exercising regularly, getting screened for cancer, and so on. (The "healthy user bias")

correlation != causation

- The study does not actually claim that eating bran muffins lowers an individual's risk of getting colon cancer;
- it merely shows a negative correlation between the consumption of bran muffins and the incidence of colon cancer in one large data set.
- This statistical association is not sufficient to prove that the bran muffins *cause* the improved health outcome.
- After all, the kind of people who eat bran muffins (particularly twenty a day!) may do lots of other things that lower their cancer risk, such as eating less red meat, exercising regularly, getting screened for cancer, and so on. (The "healthy user bias")

correlation != causation

- I should also point out that statistical significance says nothing about the *size* of the association.
- People who eat lots of bran muffins may have a lower incidence of colon cancer—but how much lower?
- The difference in colon cancer rates for bran muffin eaters and non–bran muffin eaters may be trivial;
- the finding of statistical significance means only that the observed effect, however tiny, is not likely to be a coincidence.

Eating a banana before the SAT

- Suppose you stumble across a well-designed study that has found a statistically significant positive relationship between eating a banana before the SAT and achieving a higher score on the math portion of the test.
- One of the first questions you want to ask is, How big is this effect?
- It could easily be .9 points; on a test with a mean score of 500, that is not a life-changing figure.
- Meanwhile, a finding that there is "no statistically significant association" between two variables means that any relationship between the two variables can reasonably be explained by chance alone.

software products

- According to the article, Carnegie Mellon University sells a software program called Cognitive Tutor with this bold claim:
- “Revolutionary Math Curricula. Revolutionary Results.”
- Yet an assessment of Cognitive Tutor conducted by the U.S. Department of Education concluded that the product had “no discernible/visible effects” on the test scores of high school students.
- (The *Times* suggested that the appropriate marketing campaign should be “Undistinguished Math Curricula. Unproven Results.”)
- In fact, a study of ten software products designed to teach skills such as math or reading found that nine of them “did not have statistically significant effects on test scores.”

"Link in Autism, Brain Size."

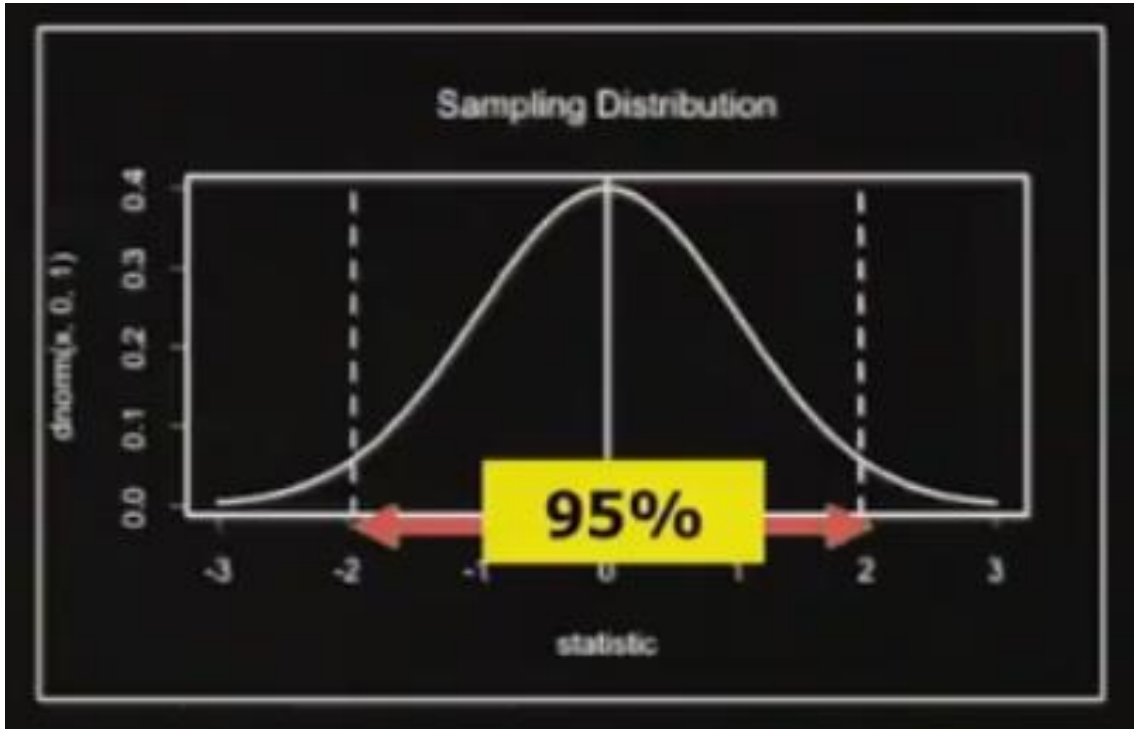
- This is an important breakthrough, as the causes of autism spectrum disorder remain elusive.
- The first sentence of the *Wall Street Journal* story, which summarized a paper published in the *Archives of General Psychiatry*, reports, "Children with autism have larger brains than children without the disorder, and the growth appears to occur before age 2, according to a new study released on Monday."
- On the basis of brain imaging conducted on 59 children with autism and 38 children without autism, researchers reported that children with autism have brains that are up to 10% larger than those of children of the same age without autism.

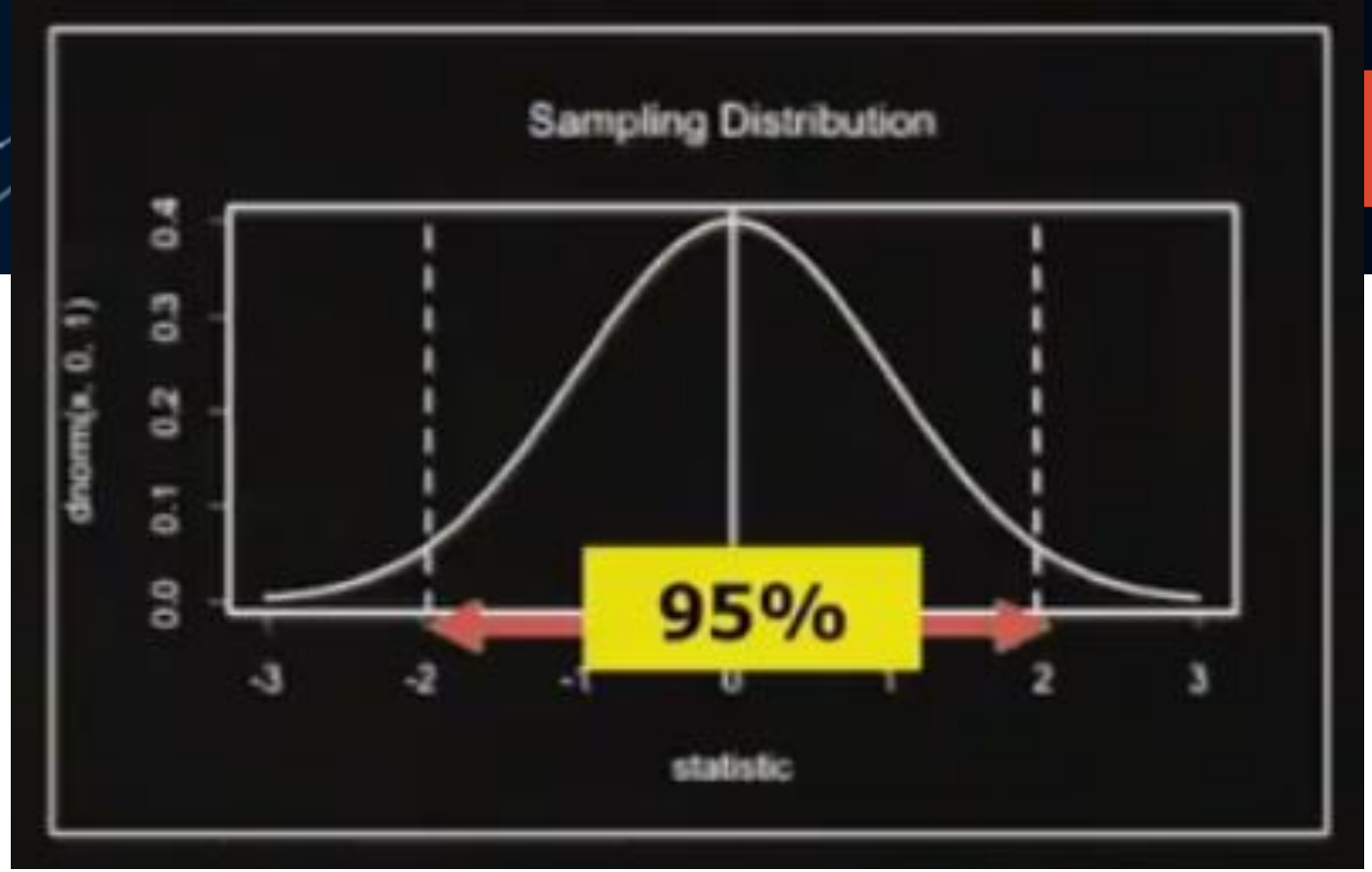
“Link in Autism, Brain Size.”

- Here is the relevant medical question: Is there a physiological difference in the brains of young children who have autism spectrum disorder?
- If so, this insight might lead to a better understanding of what causes the disorder and how it can be treated or prevented.
- And here is the relevant statistical question:
- Can researchers make sweeping inferences about autism spectrum disorder in general that are based on a study of a seemingly small group of children with autism (59) and an even smaller control group (38)—a mere 97 subjects in all?

Confidence Intervals

- How do we use sample distribution in inference?
- For confidence interval, is we want to use the variability measured in the sampling distribution to extend a point estimate, such as sample mean, into an interval estimate.





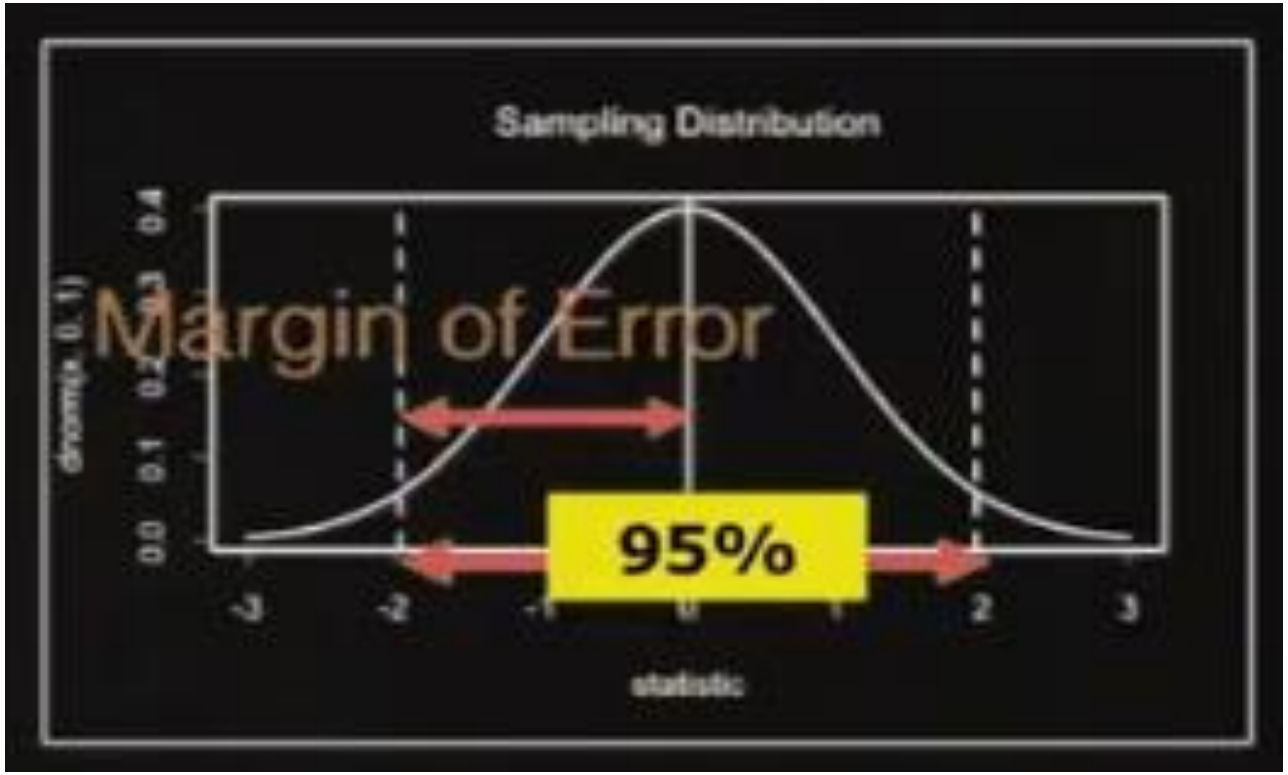
- **we say the population mean or the population quantity of interest is estimated to be within an interval instead of equal to a point.**
- By extending from a point estimate to an interval estimate, our chance of covering the true value of the parameter, or the quantity in the population we're interested in, improves. This is how we use that.

Confidence Intervals

- So assuming you actually know the sampling distribution--which is rarely the case in practice
- -- and you want to create a region where most of the sample estimates fall into–
- so you want to describe the variability from the sampling distribution, the behavior of the sample estimate according to the distribution model using one length, and this length corresponding to 95% of chance.
- The width of this length corresponding to a notion that we have discussed in our statistical thinking practice called margin of error.

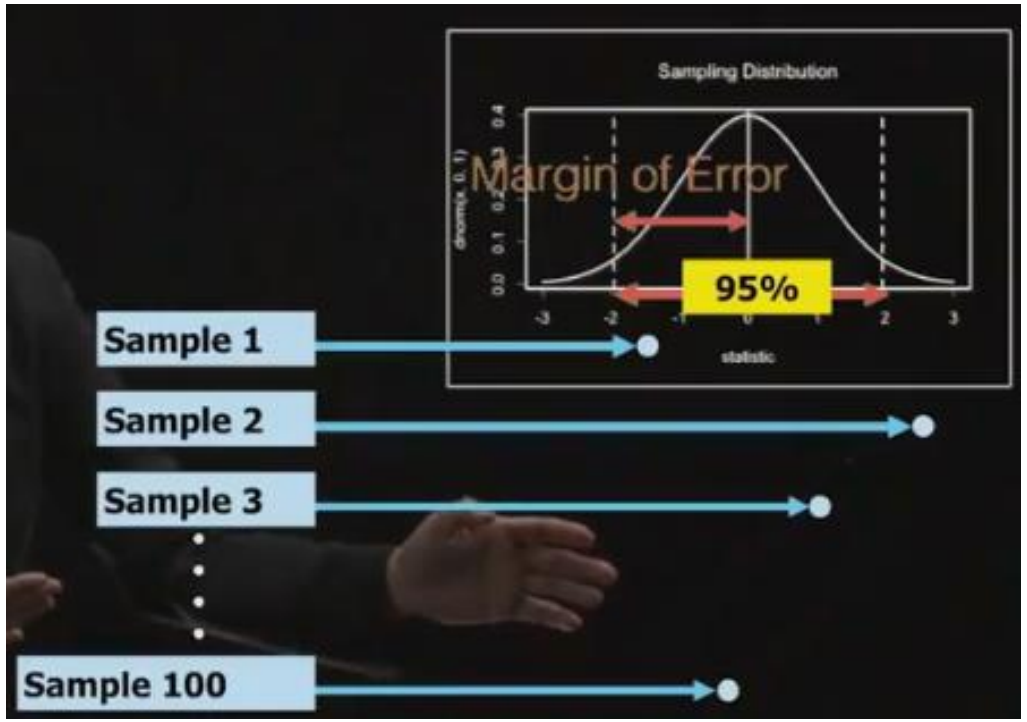
Confidence Intervals

- So the margin of error is a length equal to the half of an interval—that is, given the center of variability, you'll go this much to the left and this much to the right, you contain about 95% of the viability in the sample estimates.



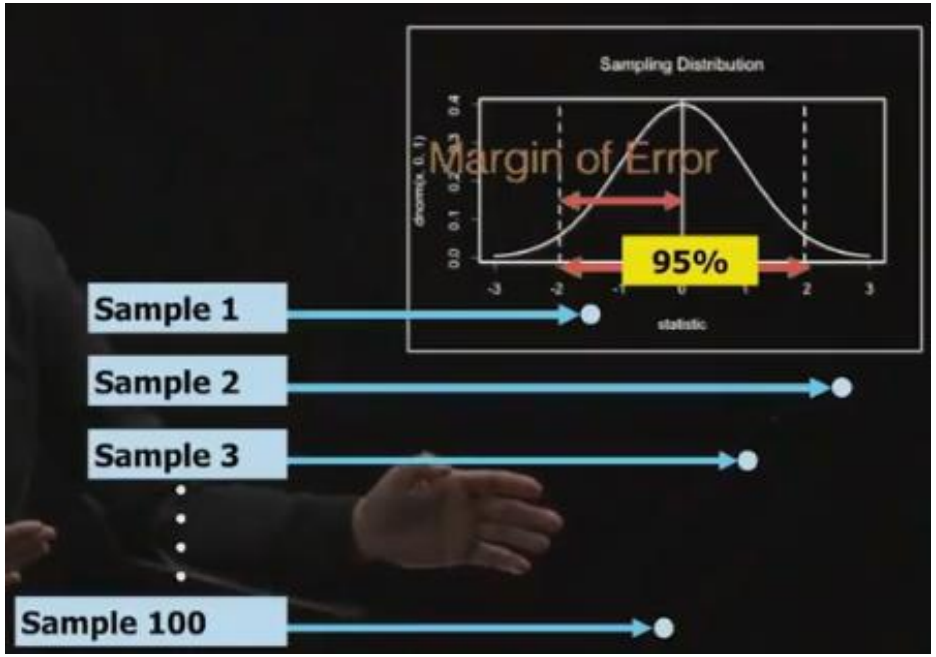
Confidence Intervals

- That means if you draw different samples-- one, two, three, all the way to, say, 100 samples--most of the sample estimates derived from these random samples should give a numerical summary of the statistic of question within this range of variability, the central 95% interval.



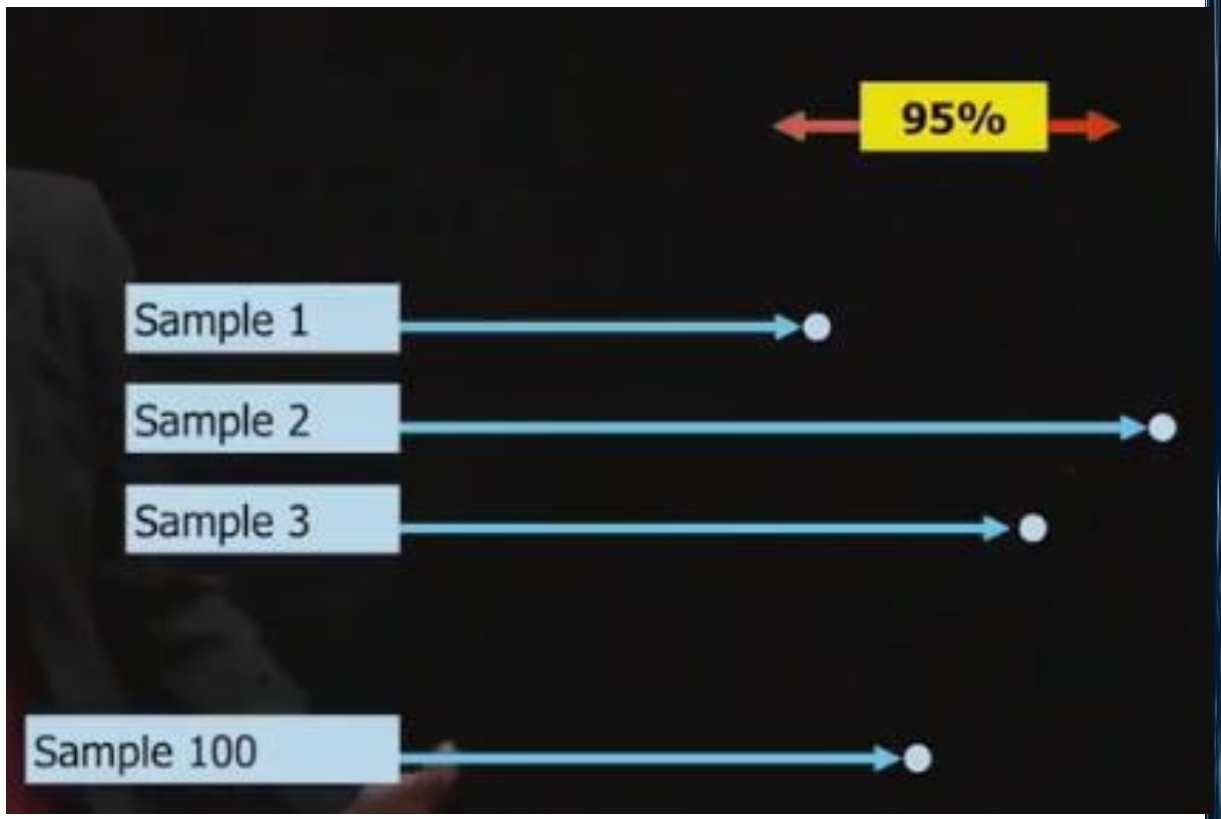
Confidence Intervals

- This is a probability interval because this is a true sampling distribution.
- So this is assuming that we actually know the sampling distribution, what are the expected behaviors of the sample estimates if you have many of them?



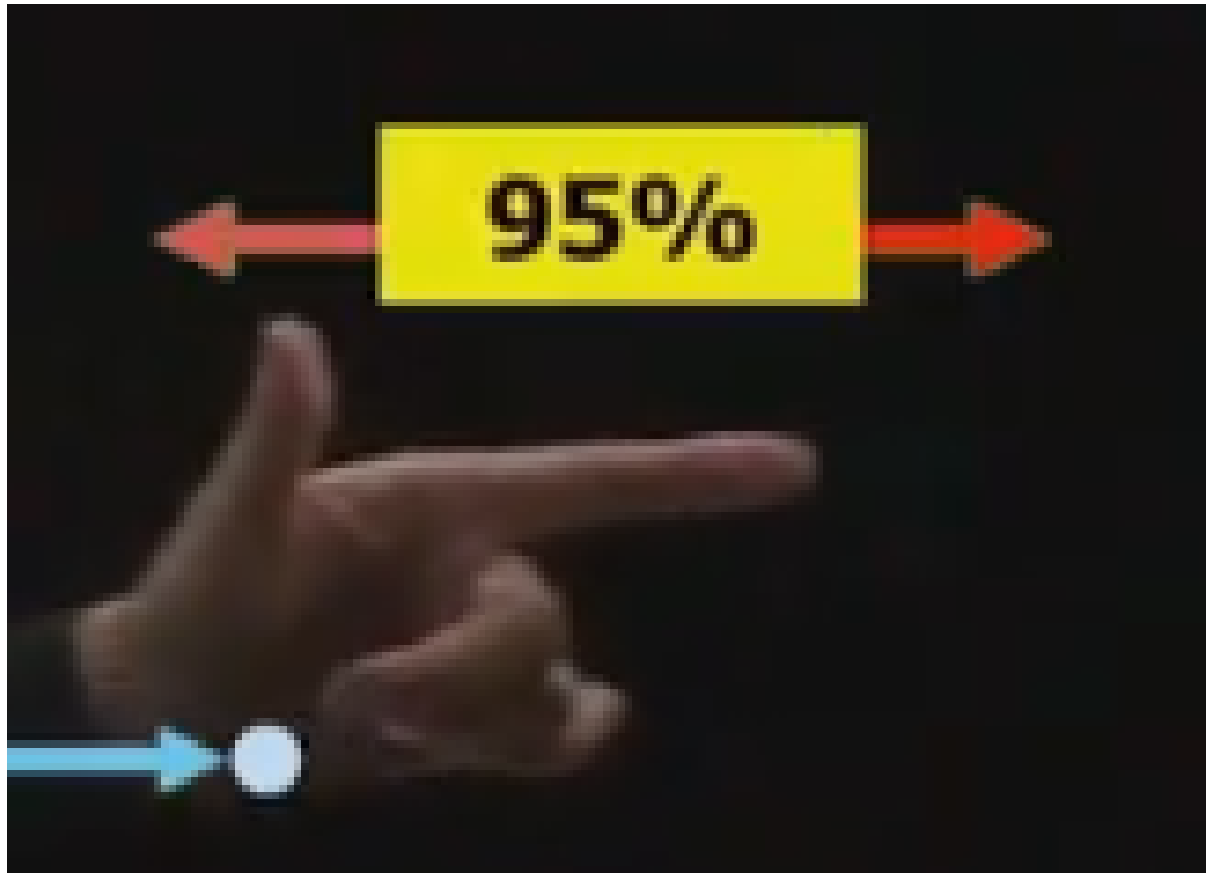
Confidence Intervals

- Of course, in practice you do not have knowledge of the sampling distribution, so that disappears.
- Given some mathematical tools, assumptions, and probability calculation, most of the time we can estimate this width.
- So the width of this interval, probability interval that if **centered at the truth** should contain 95% of the sample estimates.

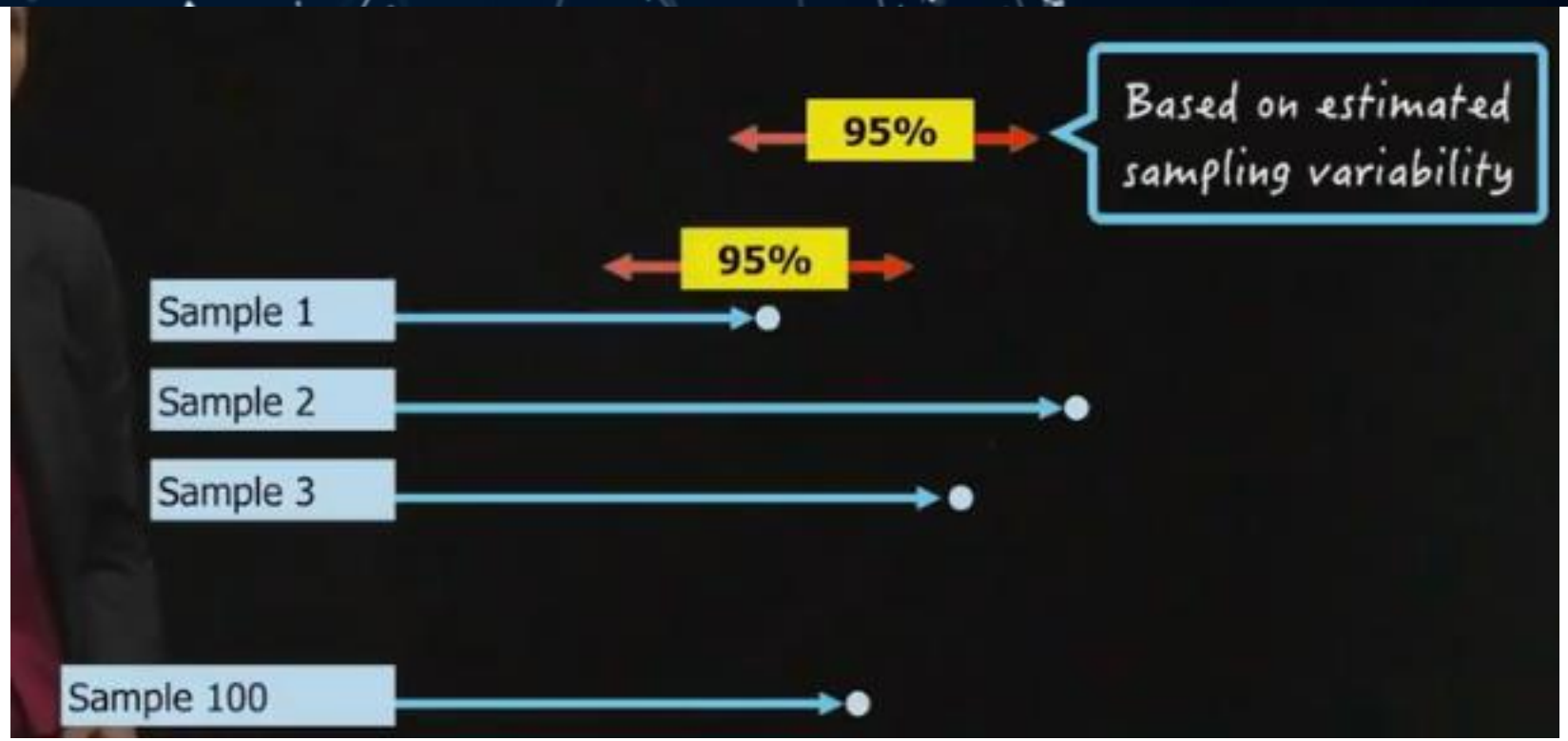


Confidence Intervals

- Therefore, based on our information from the sample real data or certain assumptions, we derive this quantity, this width, and apply it to individual sample estimates.

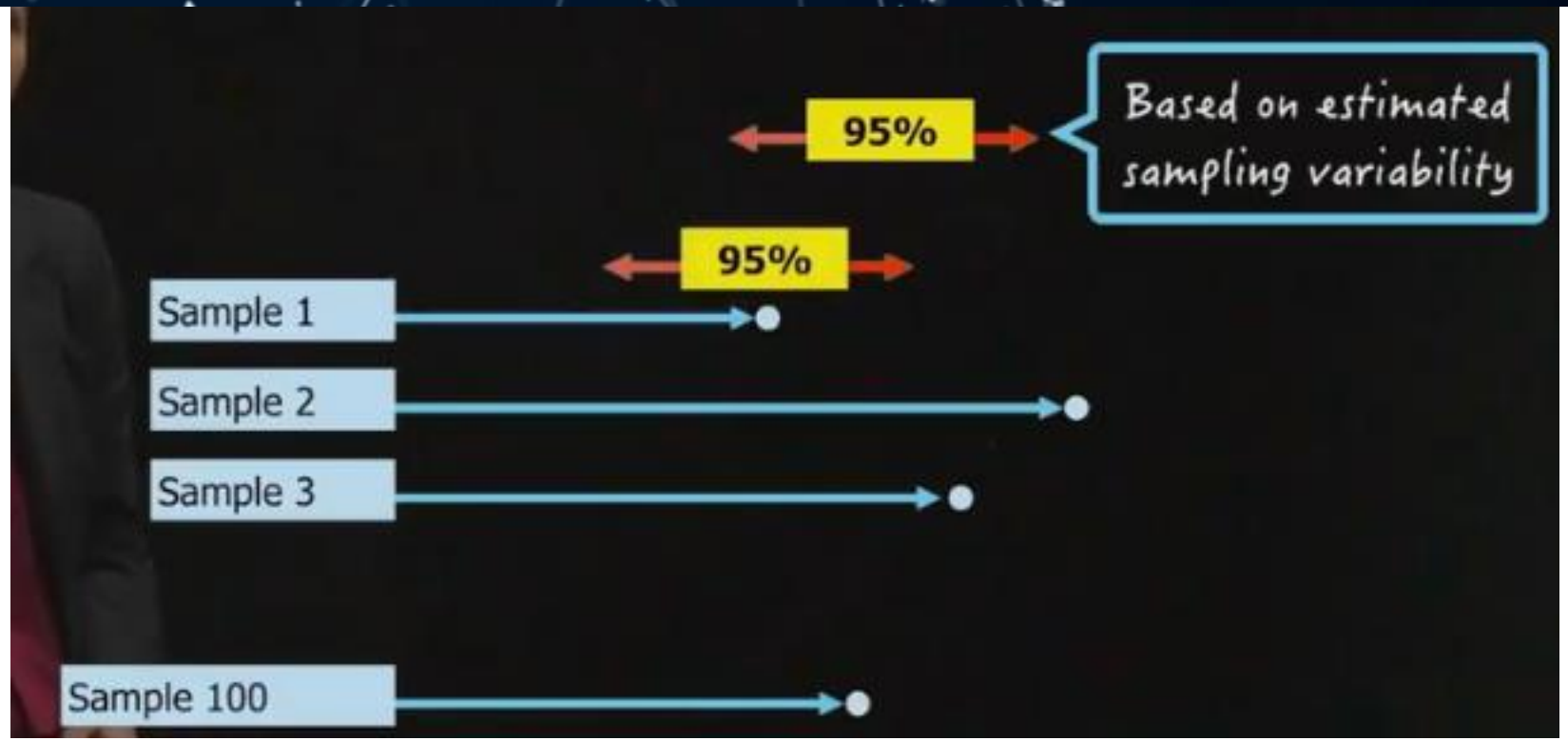


Confidence Intervals



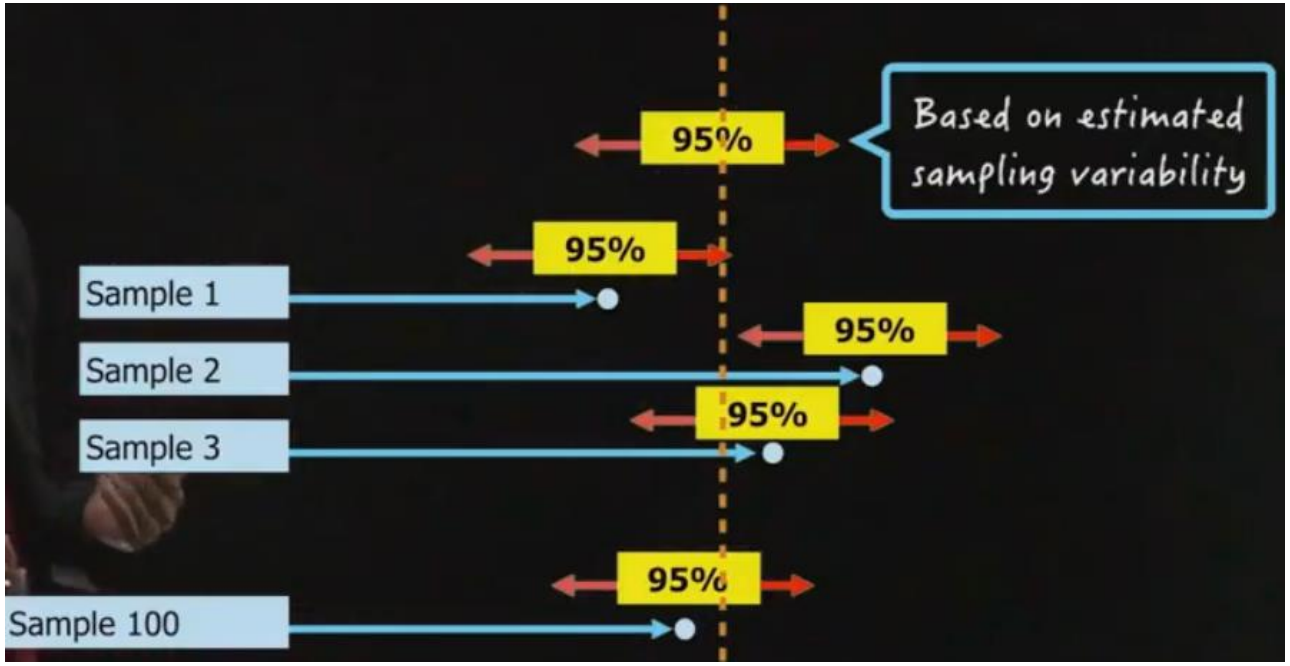
- It means: I take the sample estimates and you extend the sample estimates to an interval using the length with a width derived on some estimation of sampling variability.

Confidence Intervals



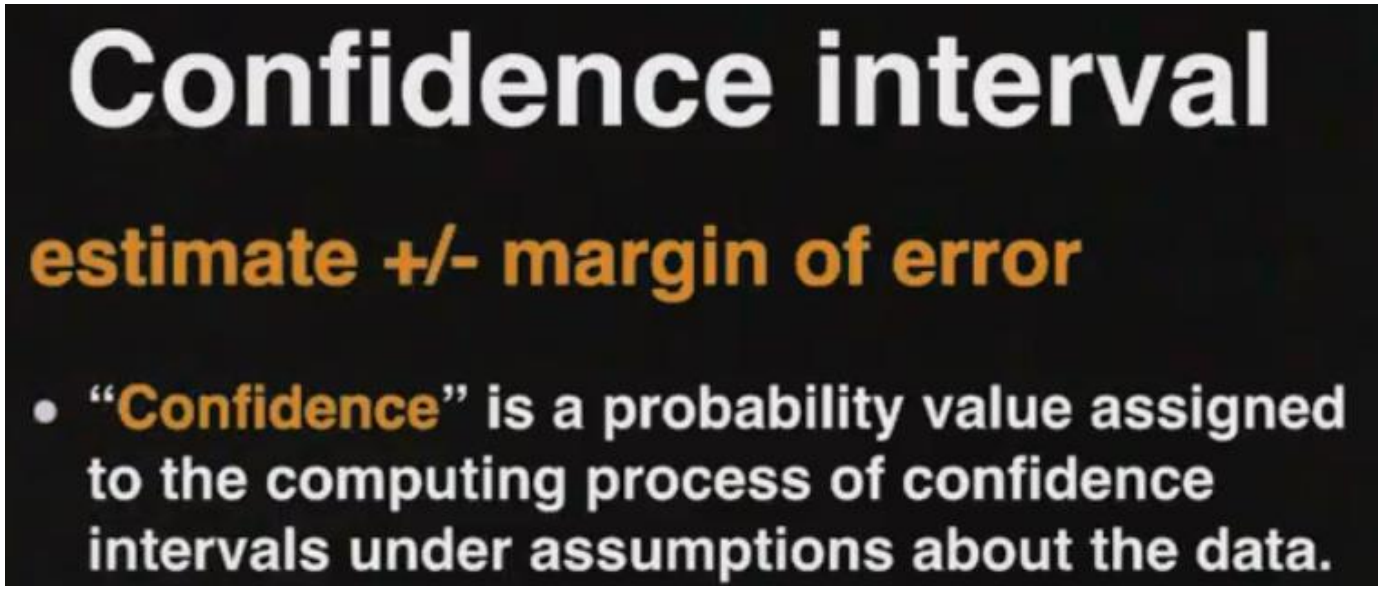
- you estimate the extent of sampling variability using information from real data and you apply that width centered at the sample estimates to create an interval estimate.

Confidence Intervals



- And if you have a lot of samples and repeat this process many, many times, according to our probability discussion on the previous slide as what the meaning of this length from the true sampling distribution, there should be a 95% chance that this individual interval created using individual sample estimates should cover the invisible truth.

Confidence Intervals



- Therefore in practice, we need to consider ways to reduce margin of error.

factors affecting margin of error

- The factors affecting margin of error include

1. the **overall variability** in the population the higher the variability in the population, the wider the margin of error tend to be
2. the **confidence level** desired by our inference
   the higher confidence you want in your results, the wider the interval tend to be.
3. The third factor is **sample size**. The larger the sample size, the more information you have in the sample, the narrower the confidence interval tends to be.

"Link in Autism, Brain Size."

- The answer is yes.
- The researchers concluded that the probability of observing the differences in total brain size that they found in their two samples would be a mere 2 in 1,000 ($p = 0.002$) if there is in fact no real difference in brain size between children with and without autism spectrum disorder in the overall population.
- First, you should recognize that each group of children, the 59 with autism and the 38 without autism, constitutes a reasonably large sample drawn from their respective populations—all children with and without autism spectrum disorder.
- The samples are large enough that the central limit will apply.

a quick tour of the central limit theorem

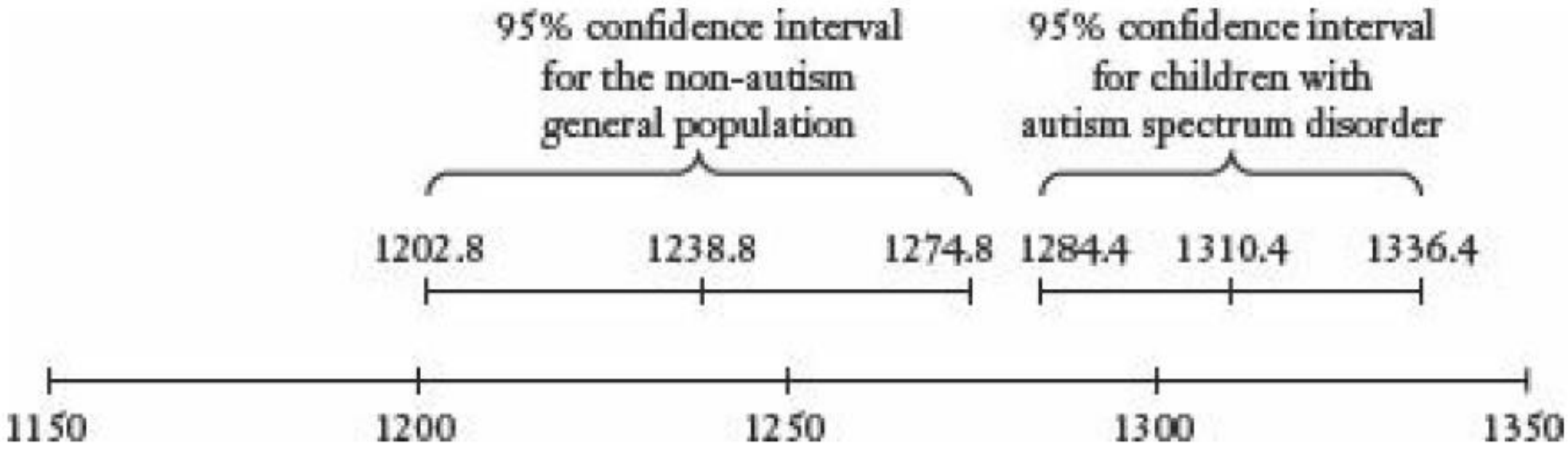
- The central limit theorem tells us:
- (1) the sample means for any population will be distributed roughly as a normal distribution around the true population mean;
- (2) we would expect the sample mean and the sample standard deviation to be roughly equal to the mean and standard deviation for the population from which it is drawn; and
- (3) roughly 68% of sample means will lie within one standard error of the population mean, roughly 95% will lie within two SE of the population mean, and so on.

a quick tour of the central limit theorem

- In less technical language, this all means that any sample should look a lot like the population from which it is drawn;
- while every sample will be different, it would be relatively rare for the mean of a properly drawn sample to deviate by a huge amount from the mean for the relevant underlying population.
- Similarly, we would also expect two samples drawn from the same population to look a lot like each other.
- Or, to think about the situation somewhat differently, if we have two samples that have extremely dissimilar means, the most likely explanation is that they came from different populations.

a quick tour of the central limit theorem

- In less technical language, this all means that any sample should look a lot like the population from which it is drawn;
- while every sample will be different, it would be relatively rare for the mean of a properly drawn sample to deviate by a huge amount from the mean for the relevant underlying population.
- Similarly, we would also expect two samples drawn from the same population to look a lot like each other.
- Or, to think about the situation somewhat differently, if we have two samples that have extremely dissimilar means, the most likely explanation is that they came from different populations.


95% confidence interval
for the non-autism
general population
95% confidence interval
for children with
autism spectrum disorder
1202.8
1238.8
1274.8
1284.4
1310.4
1336.4
1150
1200
1250
1300
1350

The brain study

- The researchers' null hypothesis was presumably that there are no anatomical differences in the brains of children with and without autism.
- The alternative hypothesis is that the brains of children with autism spectrum disorder are fundamentally different.
- Such a finding would still leave lots of questions, but it would point to a direction for further inquiry.

The brain study

- In this study, the children with autism spectrum disorder had a mean brain volume of 1310.4 cm^3;
- the children in the control group had a mean brain volume of 1238.8 cm^3.
- Thus, the difference in average brain volume between the two groups is 71.6 cm^3.
- How likely would this result be if in fact there were no difference in average brain size in the general population between children who have autism spectrum disorder and children who do not?

The brain study

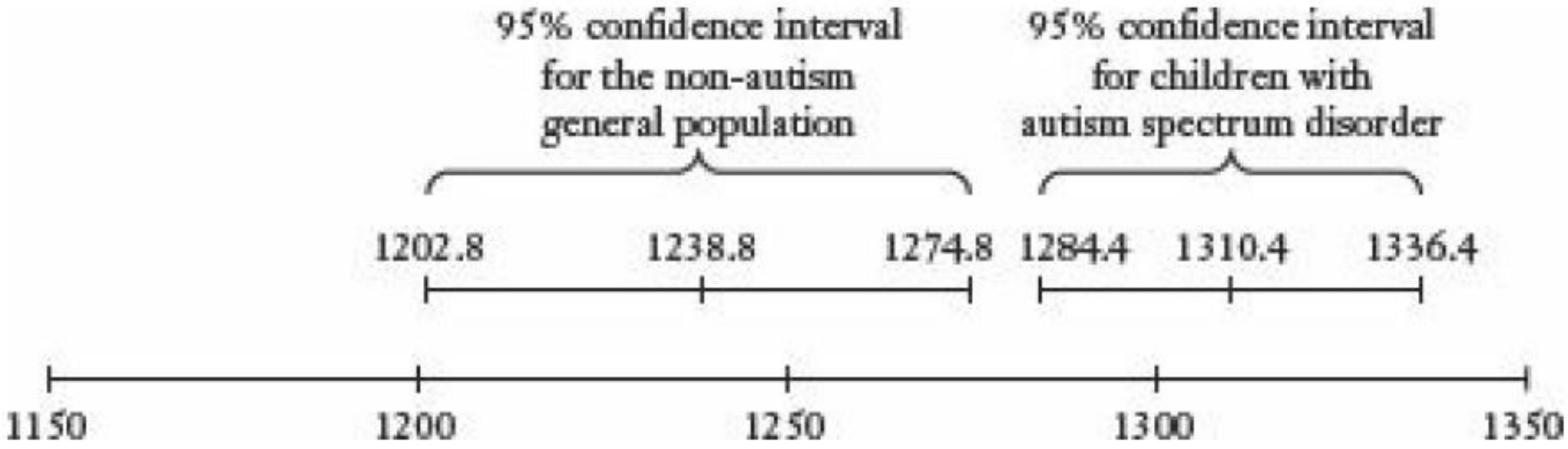
- The standard error for the total brain volume of the 59 children in the autism spectrum disorder sample is 13 cm^3;
- the standard error for the total brain volume of the 38 children in the control group is 18 cm^3.
- You will recall that the central limit theorem tells us that for 95 samples out of 100, the sample mean is going to lie within two SE of the true population mean, in one direction or the other.
- As a result, we can infer from our sample that 95 times out of 100 the interval of 1310.4 cm^3 ± 26 will contain the average brain volume for *all* children with autism spectrum disorder.
- This expression is called a confidence interval.

The brain study

- This expression is called a confidence interval.
- We can say with 95% confidence that the range 1284.4 to 1336.4 cm^3 contains the average total brain volume for children in the general population with autism spectrum disorder.
- Using the same methodology, we can say with 95% confidence that the interval of 1238.8 ± 36, or between 1202.8 and 1274.8 cm^3, will include the average brain volume for children in the general population who do not have autism spectrum disorder.

The brain study

- The *lower bound* of our 95% confidence interval for the average brain size of children with autism in the general population (1284.4 cm^3) is still higher than the *upper bound* for the 95% confidence interval for the average brain size for young children in the population without autism (1274.8 cm^3), as the following diagram illustrates.



The brain study

- This is the first clue that there may be an underlying anatomical difference in the brains of young children with autism spectrum disorder.
- Still, it's just a clue.
- All of these inferences are based on data from fewer than 100 children.
- Maybe we just have wacky samples.
- One final statistical procedure can bring all this to realization.
- We can calculate the exact probability of observing a difference of means at least this large (1310.4 cm^3 versus 1238.8 cm^3) if there is really no difference in brain size between children with autism spectrum and all other children in the general population.
- We can find a p-value for the observed difference in means.

The brain study

- If we draw two large samples from the same population, we would expect them to have very similar means.
- In fact, our best guess is that they will have identical means.
- For example, if I were to select 100 players from the NBA and they had an average height of 6 feet 7 inches, then I would expect another random sample of 100 players from the NBA to have a mean height close to 6 feet 7 inches.
- Okay, maybe the two samples would be 1 inch or 2 apart.
- But it's less likely that the means of the two samples will be 4 inches apart—and even less likely that there will be a difference of 6 or 8 inches.

The brain study

- It turns out that we can calculate a standard error for the difference between two sample means;
- This standard error gives us a measure of the dispersion we can expect, on average, when we subtract one sample mean from the other.
- (As noted earlier, the formula is in the chapter appendix.)
- The important thing is that we can use this standard error to calculate the probability that two samples come from the same population.

Here is how it works:

- 1. If two samples are drawn from the same population, our best guess for the difference between their means is zero.
- 2. The central limit theorem tells us that in repeated samples, the *difference between the two means* will be distributed roughly as a normal distribution.
- 3. If the two samples really have come from the same population, then in roughly 68 cases out of 100, the difference between the two sample means will be within one standard error of zero.

Here is how it works:

- And in roughly 95 cases out of 100, the difference between the two sample means will be within two SE of zero.
- And in 99.7 cases out of 100, the difference will be within three SE of zero—which turns out to be what motivates the conclusion in the autism research paper that we started with.
- As noted earlier, the difference in the mean brain size between the sample of children with autism spectrum disorder and the control group is 71.6 cm^3.

Here is how it works:

- The standard error on that difference is 22.7, meaning that the difference in means between the two samples is more than three standard errors from zero; we would expect an outcome this extreme (or more so) only 2 times in 1,000 if these samples are drawn from an identical population.
- In the paper published in the *Archives of General Psychiatry*, the authors report a p-value of .002, as I mentioned earlier.
- Now you know where it came from!

significant pitfalls

- A Cornell professor explicitly proposed a null hypothesis, conducted an experiment to test his null hypothesis, and then rejected the null hypothesis at the .05 significance on the basis of the experimental results.
- The result was uproar, in scientific circles as well as mainstream media outlets like the *New York Times*.
- The researcher in question was testing humans' capacity to exercise extrasensory perception, or ESP.
- The null hypothesis was that ESP does not exist; the alternative hypothesis was that humans do have telepathic powers.
- To study this question, the researcher recruited a large sample of participants to examine two "curtains" posted on a computer screen.

significant pitfalls

- A software program randomly put an X photo behind one curtain or the other.
- In repeated trials, study participants were able to pick the curtain with the X photo behind it 53 % of the time, whereas probability says they would be right only 50 % of the time.
- Because of the large sample size, the researcher was able to reject the null hypothesis that extrasensory perception does not exist and accept instead the alternative hypothesis that extrasensory perception can enable individuals to sense future events.
- The decision to publish the paper was widely criticized on the grounds that a single statistically significant event can easily be a product of chance, especially when there is no other evidence corroborating or even explaining the finding.

significant pitfalls

- The *New York Times* summarized the critiques:
- “Claims that defy almost every law of science are by definition extraordinary and thus require extraordinary evidence.
- Neglecting to take this into account —as conventional social science analyses do—makes many findings look far more significant than they really are.”
- One answer to this kind of nonsense would appear to be a more rigorous threshold for defining statistical significance, such as .001.*
- But that creates problems of its own.
- Choosing an appropriate significance level involves an inherent trade-off.

significant pitfalls

- If our burden of proof for rejecting the null hypothesis is too low (e.g., .1), we are going to find ourselves periodically rejecting the null hypothesis when in fact it is true (as I suspect was the case with the ESP study).
- In statistical jargon, this is known as a Type I error.
- Consider the example of an American courtroom, where the null hypothesis is that a defendant is not guilty and the threshold for rejecting that null hypothesis is "guilty beyond a reasonable doubt."
- Suppose we were to relax that threshold to something like "a strong hunch that the guy did it."
- This is going to ensure that more criminals go to jail—and also more innocent people.

Well, 1 in 10 is not exactly wildly unlikely.

- Consider this challenge in the context of approving a new cancer drug.
- In a statistical context, this is the equivalent of having a relatively low significance level, such as 0.1.
- For every ten drugs that we approve with this relatively low burden of statistical proof, one of them does not actually work and showed promising results in the trial just by chance.
- **A Type I error involves wrongly rejecting a null hypothesis**.
- Though the terminology is somewhat counterintuitive, this is also known as a "**false positive**."

Well, 1 in 10 is not exactly wildly unlikely.

- Here is one way to reconcile the jargon.
- When you go to the doctor to get tested for some disease, the null hypothesis is that you do not have that disease.
- If the lab results can be used to reject the null hypothesis, then you are said to test positive.
- And if you test positive but are not really sick, then it's a false positive.

Well, 1 in 10 is not exactly wildly unlikely.

- In any case, the lower our statistical burden for rejecting the null hypothesis, the more likely it is to happen.
- Obviously we would prefer not to approve ineffective cancer drugs, or send innocent defendants to prison.
- But there is a tension here.
- The higher the threshold for rejecting the null hypothesis, the more likely it is that **we will fail to reject a null hypothesis that ought to be rejected**.
- If we require five eyewitnesses in order to convict every criminal defendant, then a lot of guilty defendants are wrongly going to be set free. (Of course, fewer innocents will go to prison.)

Well, 1 in 10 is not exactly wildly unlikely.

- If we adopt a .001 significance level in the clinical trials for all new cancer drugs, then we will indeed minimize the approval of ineffective drugs. (There is only a 1 in 1,000 chance of wrongly rejecting the null hypothesis that the drug is no more effective than a placebo.)
- Yet now we introduce **the risk of not approving many effective drugs** because we have set the bar for approval so high.
- This is known as a Type II error, or **false negative**.

Well, 1 in 10 is not exactly wildly unlikely.

- Which kind of error is worse?
- That depends on the circumstances.
- The most important point is that you recognize the trade-off.
- There is no statistical “free lunch.”

Type I vs. Type II errors: Spam filters.

- The null hypothesis is that any particular e-mail message is *not* spam.
- Your spam filter looks for clues that can be used to reject that null hypothesis for any particular e-mail.
- A Type I error would involve screening out an e-mail message that is not actually spam (a false positive).
- A Type II error would involve letting spam through the filter into your inbox (a false negative).

Type I vs. Type II errors: Spam filters.

- Given the costs of missing an important e-mail relative to the costs of getting the occasional message about herbal vitamins, most people would probably err on the side of allowing Type II errors.
- An optimally designed spam filter should require a relatively high degree of certainty before rejecting the null hypothesis that an incoming e-mail is legitimate and blocking it.
- We have numerous tests for the early detection of cancer, such as mammograms (breast cancer), the PSA test (prostate cancer), and even full-body MRI scans for anything else that might look suspicious.
- The null hypothesis for anyone undergoing this kind of screening is that no cancer is present.

Type I vs. Type II errors: Screening for cancer.

- The screening is used to reject this null hypothesis if the results are suspicious.
- The assumption has always been that a Type I error (a false positive that turns out to be nothing) is far preferable to a Type II error (a false negative that misses a cancer diagnosis).
- Historically, cancer screening has been the opposite of the spam example.
- Doctors and patients are willing to tolerate a fair number of Type I errors (false positives) in order to avoid the possibility of a Type II error (missing a cancer diagnosis).
- More recently, health policy experts have begun to challenge this view

Type I vs. Type II errors: Screening for cancer.

- We have numerous tests for the early detection of cancer, such as mammograms (breast cancer), the PSA test (prostate cancer), and even full-body MRI scans for anything else that might look suspicious.
- The null hypothesis for anyone undergoing this kind of screening is that no cancer is present.
- The screening is used to reject this null hypothesis if the results are suspicious.
- The assumption has always been that a Type I error (a false positive that turns out to be nothing) is far preferable to a Type II error (a false negative that misses a cancer diagnosis).

Type I vs. Type II errors: Screening for cancer.

- Historically, cancer screening has been the opposite of the spam example.
- Doctors and patients are willing to tolerate a fair number of Type I errors (false positives) in order to avoid the possibility of a Type II error (missing a cancer diagnosis).
- More recently, health policy experts have begun to challenge this view because of the high costs and serious side effects associated with false positives.
- Neither a Type I nor a Type II error is acceptable in this situation, which is why society continues to debate about the appropriate balance between fighting terrorism and protecting civil liberties.

Type I vs. Type II errors: Capturing terrorists

- The null hypothesis is that an individual is not a terrorist.
- As in a regular criminal context, we do not want to commit a Type I error and send innocent people to Guantánamo Bay.
- Yet in a world with weapons of mass destruction, letting even a single terrorist go free (a Type II error) can be literally catastrophic.
- This is why—whether you approve of it or not—the US is holding suspected terrorists at Guantánamo Bay on the basis of less evidence than might be required to convict them in a regular criminal court.

Type I vs. Type II errors: Capturing terrorists

- Statistical inference is not magic, nor is it infallible (dependable), but it is an extraordinary tool for making sense of the world.
- We can gain great insight into many life phenomena just by determining the most likely explanation.
- Most of us do this all the time (e.g., “I think that college student passed out on the floor surrounded by beer cans has had too much to drink” rather than “I think that college student passed out on the floor surrounded by beer cans has been poisoned by terrorists”).
- Statistical inference merely formalizes the process.

standard error for a difference of means

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

→ numerator yields the size of the difference in means

→ denominator yields the standard error for a difference in mean between two samples

- where $\bar{x}$= mean for sample x
- $\bar{y}$= mean for sample y
- s_x= standard deviation for sample x
- s_y= standard deviation for sample y
- n_y = number of observations in sample x
- n_y = number of observations in sample y

Polling

- *How we know that 64 % of Americans support the death penalty (with a sampling error ± 3 percent)*

Political polls

- NYT : “A deep sense of anxiety and doubt about the future hangs over the nation.”
- Here is a snapshot of what Americans had to say in the fall of 2011
- A shocking 89% said that they distrust government to do the right thing, the highest level of distrust ever recorded.
- 2/3 said that wealth should be more evenly distributed in the country.
- 43% said that they generally agreed with the views of the Occupy Wall Street movement.
- A slightly higher percentage, 46 %, said that the views of the people involved in the Occupy Wall Street movement “generally reflect the views of most Americans.”

Political polls

- 64% approved of Obama's handling of his job as president—and an identical 46% disapproved of his job performance.
- A mere 9% approved of the way Congress was handling its job.
- Even though the presidential primaries would begin in only two months, roughly 80 % of Republican primary voters said "it was still too early to tell whom they will support."

Political polls

- The power of polling stems from the same source as our previous sampling examples: the central limit theorem.
- If we take a large, representative sample of American voters (or any other group), we can reasonably assume that our sample will look a lot like the population from which it is drawn.
- If exactly half of American adults disapprove of SOMETHNIG, then our best guess about the attitudes of a representative sample of 1,000 Americans is that about half of them will disapprove of that SOMETHNIG.

RESEARCH

margin of error

- When you read that a poll has a “margin of error” of ± 3 percent, this is really just the same kind of 95 % confidence interval that we calculated in the last chapter.
- Our “95 % confidence” means that if we conducted 100 different polls on samples drawn from the same population, we would expect the answers we get from our sample in 95 of those polls to be within 3 percentage points in one direction or the other of the population’s **true sentiment.**
- In the context of the job approval question NYT poll, we can be 95 % confident that the true proportion of all Americans who disapprove of President Obama’s job rating lies in the range of 46 % ± 3 percent, or between 43 % and 49 percent.

margin of error

- For any properly drawn random sample, the standard error is equal to **sqrt(p(1-p)/n)** where p is the proportion of respondents expressing a particular view, (1 – p) is the proportion of respondents expressing a different **view.**
- The standard error also tends to be smaller when p and (1 – p) are far apart.
- For example, the standard error will be smaller for a poll in which 95 % of respondents express a certain view than for a poll in which opinions tend to split 50-50.
- This is just math, since (.05)(.95)=.047, while (.5)(.5)=.25; a smaller number in the numerator of the formula leads to a smaller standard error.

“exit poll”

- 500 representative voters on election day finds that 53% voted for X; 45% voted for Y;
- and 2 % supported Z.
- If we use X as our proportion of interest, the standard error for this exit poll would be

$$\sqrt{(.53)(1-.53)/500} = \sqrt{(.53)(.47)/500} = \sqrt{.25/500} = \sqrt{.0005} = .02236.$$

- For simplicity, we’ll round the standard error for this exit poll to .02.

“exit poll”

- Remember, the standard error gives us a sense of how often we can expect our sample proportion (the exit poll) to lie reasonably close to the true population proportion (the election outcome).
- We know that roughly 68 % of the time we can expect the sample proportion—the 53% who voted for X—to be within one standard error of the true final tally.
- As a result, you tell your producer “with 68% confidence” that your sample, which shows X getting 53% of the vote ± 2%, or between 51 and 55 percent, is X’s true tally.

“exit poll”

- Meanwhile, the same exit poll shows that the Y has received 45% of the vote.
- If we assume that the vote tally for the X candidate has the same standard error (a simplification), we can say with 68% confidence Y's tally is 45% of the vote ± 2%, between 43 and 47.

X (Republican) 53%

Y (Democrat) 45%

Z (Independent) 2%

(Margin of Error 2%)

“exit poll”

- 68 times out of 100 your exit poll results will be within one standard error of the true election outcome, your producer, points out the obvious math—**32 times out of 100 your exit poll *will not be* within one standard error of the true election outcome**.
- Then what?

“exit poll”

- You explain that there are two possibilities:
- (1) the X candidate could have received *even more* votes than your poll predicted, in which case you still will have called the election correctly. Or
- (2) there is a reasonably high probability that the Y candidate has received far more votes than your poll has reported.
- She screams, “How can we be sure that we have the right result?”
- you cannot be certain of any result until all of the votes are counted.
- However, you can offer a 95% confidence interval instead.
- In this case, your graphic will be wrong, on average, only 5 times out of 100.

“exit poll”

- The only way the station can be more confident of its polling results is by broadening the “margin of error".
- You show your boss the new fancy graphic:

X (Republican) 53%
Y (Democrat) 45%
Z (Independent) 2%
(Margin of Error 4%)

“exit poll”

- We know from the central limit theorem that roughly 95 % of sample proportions will lie within *two standard errors* of the true population proportion (which is 4% in this case).
- Therefore, if we want to be more confident of our polling results, we have to be less ambitious in what we are predicting.
- As the above graphic illustrates, at the 95 % confidence level, the television station can announce that the X has earned 53 % of the vote ±4 percent, or **49% to 57%** of the votes cast.
- Meanwhile, the Y has earned 45 % ± 4 percent, or **41% to 49%** of the votes cast.

- *The miracle elixir*

Can stress on the job kill you? Yes.

- There is compelling evidence that rigors on the job can lead to premature death, particularly of heart disease.
- But it's not the kind of stress you are probably imagining.
- CEOs are at significantly *less* risk than their secretaries, who dutifully answer the phone and perform other tasks as instructed.
- How?
- It turns out that the most dangerous kind of job stress stems from having "low control" over one's responsibilities.

Can stress on the job kill you? Yes.

- Several studies of thousands of British civil servants have found that workers who have little control over their jobs have a significantly higher mortality rate than other workers in the civil service with more decision making authority.
- According to this research, it is not the stress associated with major responsibilities that will kill you;
- **it is the stress associated with being told what to do while having little say in how or when it gets done.**

Can stress on the job kill you? Yes.

- The relevant question is how researchers can possibly come to such a conclusion.
- Clearly this cannot be a randomized experiment.
- We cannot arbitrarily assign human beings to different jobs, force them to work in those jobs for many years, and then measure who dies at the highest rate.
- **Instead, researchers have collected detailed longitudinal data on thousands of individuals in the British civil service**;
- these data can be analyzed to identify meaningful associations, such as the connection between “low control” jobs and coronary heart disease.

association

- A simple association is not enough to conclude that certain kinds of jobs are bad for your health.
- If we merely observe that low-ranking workers have higher rates of heart disease, our results would be confounded by other factors.
- For example,
- • *we would expect low-level workers to have less education*
- • *They may be more likely to smoke*
- • *They may have had less healthy childhoods,*
- • *Or their lower pay may limit their access to health care.*
- And so on.

association

- The point is that any study simply comparing health outcomes across a large group of British workers will not really tell us much.
- Other sources of variation in the data are likely to obscure the relationship that we care about.
- Is **“low job control”** really causing heart disease?
- Or is it some combination of **other factors that happen to be shared by people with low job con**trol, in which case we may be completely missing the real public health threat.
- Regression analysis is the statistical tool that helps us deal with this challenge.

Regression analysis

- regression analysis allows us to quantify the ***relationship between a particular variable and an outcome*** that we care about while ***controlling for other factors.***
- In other words, we can isolate the effect of one variable, such as having a certain kind of job, while holding the effects of other variables constant.
- The Whitehall studies used regression analysis to measure the health impacts of low job control among people who are similar in other ways, such as smoking behavior.
- (Low-level workers do in fact smoke more than their superiors;)

Regression analysis

- Most of the studies that you read about in the newspaper are based on regression analysis.
- When researchers conclude that children who spend a lot of time in day care are more prone to behavioral problems in elementary school than children who spend that time at home, **the study has not randomly assigned thousands of infants either to day care or to home care with a parent.**
- Nor has the study simply compared the elementary school behavior of children who had different early childhood experiences without recognizing that these populations are likely to be different in other fundamental ways.

Regression analysis

- Different families make different child care decisions *because they are different*.
- Some households have two parents present; some don't.
- Some have two parents working; some don’t.
- Some households are wealthier or more educated than others.
- All of these things affect child care decisions, *and they affect how children in those families will perform in elementary school*.
- When done properly, regression analysis can help us estimate the effects of day care apart from other things that affect young children: family income, family structure, parental education, and so on.

Regression analysis **when done properly**

- Now, there are two key phrases in that last sentence.
- The first is "**when done properly**."
- Given adequate data and access to a personal computer, a six-year-old could use a basic statistics program to generate regression results.
- Personal computing has made the mechanics of regression analysis almost effortless.
- The problem is that the mechanics of regression analysis are not the hard part; the hard part is determining **which variables** ought to be considered in the analysis and **how that can best be done**.
- Regression analysis is like one of those fancy power tools.
- It is relatively easy to use, but hard to use well—and potentially dangerous when used improperly.

RESEARCH
Khatam University
FANAP

help us estimate

- The second important phrase above is "**help us estimate**."
- Our child care study does not give us a "right" answer for the relationship between day care and subsequent school performance.
- Instead, it quantifies the relationship observed *for a particular group of children over a particular stretch of time*.
- Can we draw conclusions that might apply to the broader population?
- Yes, but we will have the same limitations and qualifications as we do with any other kind of inference.

Regression analysis **Sample**

- First, our sample has to be representative of the population that we care about.
- A study of 2,000 young children in Sweden will not tell us much about the best policies for early childhood education in rural Mexico.
- And second, there will be variation from sample to sample.
- If we do multiple studies of children and child care, each study will produce slightly different findings, even if the methodologies are all sound and similar.

Regression analysis is similar to polling.

- The good news is that **if we have a large representative sample and solid methodology, the relationship we observe for our sample data is not likely to deviate wildly from the true relationship for the whole population.**
- If 10,000 people who exercise 3 or more times a week have sharply lower rates of cardiovascular disease than 10,000 people who don't exercise (but are similar in all other important respects),
- then the chances are pretty good that we will see a similar association between exercise and cardiovascular health for the broader population.
- That's why we do these studies.
- (The point is not to tell those non exercisers who are sick at the end of the study that they should have exercised.)

RESEARCH

Regression analysis is similar to polling.

- **The bad news is that we are not proving definitively that exercise prevents heart disease.**
- **We are instead rejecting the null hypothesis that exercise has no association with heart disease**, on the basis of some statistical threshold that was chosen before the study was conducted.
- Specifically, the authors of the study would report that if exercise is unrelated to cardiovascular health, the likelihood of observing such a marked difference in heart disease between the exercisers and non exercisers in this large sample would be less than 5 in 100, or below some other threshold for statistical significance.

Regression analysis is similar to polling.

- Suppose that this particular study compared a large group of individuals who play squash regularly with those of an equal-sized group who get no exercise at all.
- Playing squash does provide a good cardiovascular workout.
- However, we also know that squash players tend to be affluent enough to belong to clubs with squash courts.
- Wealthy individuals may have great access to health care, which can also improve cardiovascular health.
- If our analysis is sloppy (untidy), we may attribute health benefits to playing squash when in fact the real benefit comes from being wealthy enough to play squash (in which case playing polo would also be associated with better heart health, even though the horse is doing most of the work).

Regression analysis is similar to polling.

- Or perhaps causality goes the other direction.
- Could having a healthy heart “cause” exercise? Yes.
- • Individuals who are sick, particularly those who have some primary form of heart disease, will find it much harder to exercise.
- • They will certainly be less likely to play squash regularly.
- Again, if the analysis is sloppy or oversimplified, the claim that exercise is good for your health may simply reflect the fact that people who start out unhealthy find it hard to exercise.
- In this case, playing squash doesn’t make anyone healthier; it merely separates the healthy from the unhealthy.

Regression analysis

- There are so many potential regression pitfalls that I've devoted the next chapter to the most egregious errors.
- For now, we'll focus on what can go right.
- Regression analysis has the amazing capacity to isolate a statistical relationship that we care about, such as that between job control and heart disease, while taking into account other factors that might confuse the relationship.

Regression analysis

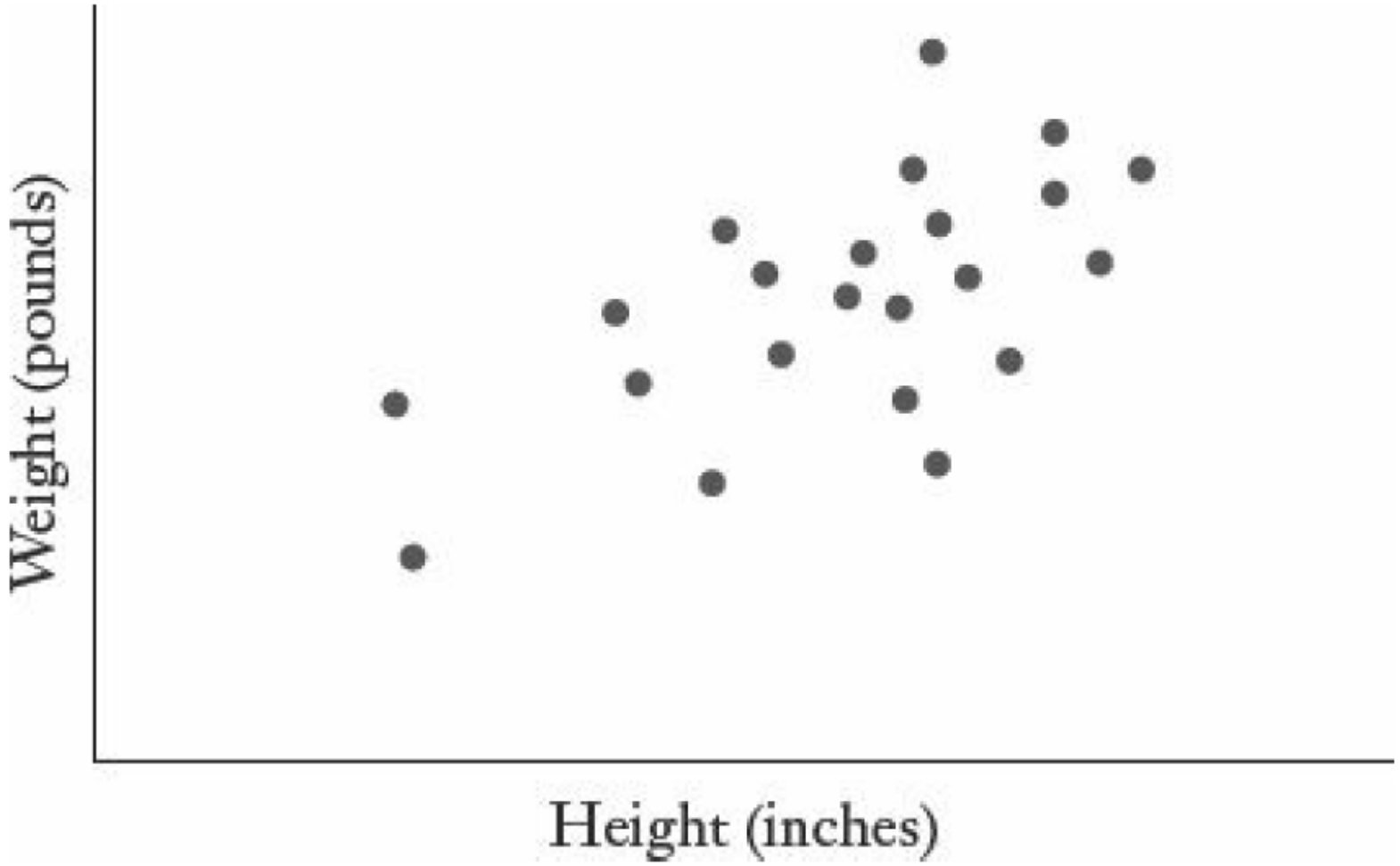
- How exactly does this work?
- If we know that low-level British civil servants smoke more than their superiors, how can we distinguish which part of their poor cardiovascular health is due to their low-level jobs, and which part is due to the smoking?
- These two factors seem inseparably intertwined.
- Regression analysis (done properly!) can untangle them.
- To explain the intuition, I need to begin with the basic idea that underlies all forms of regression analysis—from the simplest statistical relationships to the complex models cobbled together by Nobel Prize winners.

RESEARCH
Khatam University
FANAP

Regression analysis

- At its core, regression analysis seeks to find the “best fit” for a linear relationship between two variables.
- A simple example is the relationship between height and weight.
- People who are taller tend to weigh more—though that is obviously not always the case.
- If we were to plot the heights and weights of a group of graduate students, you might recall what it looked like from Chapter 4:

Scatter Plot for Height and Weight



Ordinary Least Squares

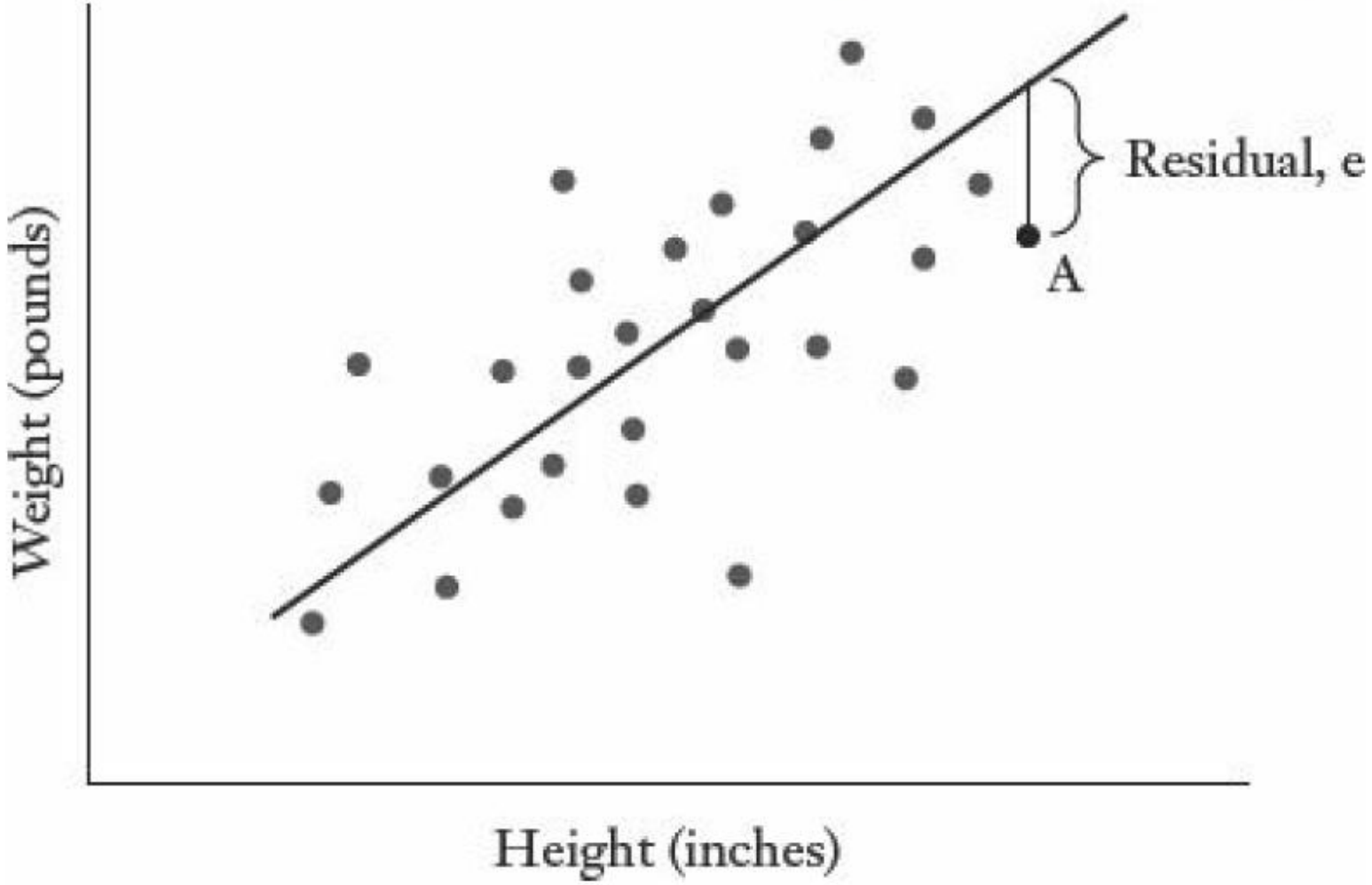
- If you were asked to describe the pattern, you might say something along the lines of “Weight seems to increase with height.”
- This is not a terribly insightful or specific statement.
- Regression analysis enables us to go one step further and “fit a line” that best describes a linear relationship between the two variables.
- Many possible lines are broadly consistent with the height and weight data.
- But how do we know which is the *best* line for these data?
- In fact, how exactly would we define “best”?
- Regression analysis typically uses a methodology called ordinary least squares, or OLS.

RESEARCH

Ordinary Least Squares

- OLS fits the line that minimizes the sum of the squared residuals.
- That's not as awfully complicated as it sounds.
- Each observation in our height and weight data set has a residual, which is its vertical distance from the regression line, except for those observations that lie directly on the line, for which the residual equals zero.
- (On the diagram below, the residual is marked for a hypothetical person A.)

Line of Best Fit for Height and Weight



Ordinary Least Squares

- It should be intuitive that the larger the sum of residuals overall, the worse the fit of the line.
- The only non intuitive twist with OLS is that the formula takes the *square* of each residual before adding them all up (which increases the weight given to observations that lie particularly far from the regression line, or the "outliers").
- Ordinary least squares "fits" the line that minimizes the sum of the squared residuals, as illustrated below.

Ordinary Least Squares

- If the technical details have given you a headache, you can be forgiven for just grasping at the bottom line, which is that OLS gives us the best description of a linear relationship between two variables.
- The result is not only a line but, as you may recall from high school geometry, an equation describing that line.
- This is known as the regression equation, and it takes the following form: $y = a + bx$,
- where y is weight in pounds;
- a is the y-intercept of the line (the value for y when x = 0);
- b is the slope of the line; and x is height in inches.

Ordinary Least Squares

- The slope of the line we've fitted, b, describes the "best" linear relationship between height and weight for this sample, as defined by OLS.
- The regression line certainly does not describe every observation in the data set perfectly.
- But it is the best description we can find for what is clearly a meaningful relationship between height and weight.
- It also means that every observation can be explained as

$$\mathbf{WEIGHT = a + b(HEIGHT) + e,}$$

- where e is a "residual" that catches the variation in weight for each individual that is not explained by height.

Ordinary Least Squares

- Finally, it means that our best guess for the weight of any person in the data set would be a + b(HEIGHT).
- Even though most observations do not lie exactly on the regression line, the residual still has an expected value of zero since any person in our sample is just as likely to weigh more than the regression equation predicts as he is to weigh less.

Ordinary Least Squares

- Let's look at some real height and weight data from the Changing Lives study, though I should first clarify some basic terminology.
- The variable that is being explained—weight—is known as the dependent variable (because it depends on other factors).
- The variables that we are using to explain our dependent variable are known as explanatory variables since they explain the outcome that we care about.(independent variables or control variables.)
- Let's start by using height to explain weight among the participants; later we'll add other potential explanatory factors.

Ordinary Least Squares

- There are 3,537 adult participants in the study, number of observations, or n = 3,537.
- When we run a simple regression on the data with weight as the dependent variable and height as the only explanatory variable, we get the following results:

WEIGHT = –135 + (4.5) × HEIGHT IN INCHES

- **a = –135**. This is the y-intercept, which has no particular meaning on its own.
- This figure is also known as the constant, because it is the starting point for calculating the weight of all observations in the study.

Ordinary Least Squares

WEIGHT = –135 + (4.5) × HEIGHT IN INCHES

- **a = –135**. This is the y-intercept, which has no particular meaning on its own.
- This figure is also known as the constant, because it is the starting point for calculating the weight of all observations in the study.
- **b = 4.5.** Our estimate for b, 4.5, is known as a regression coefficient, or in statistics jargon, "the coefficient on height," because it gives us the best estimate of the relationship between height and weight among the Changing Lives participants.
- The regression coefficient has a convenient interpretation: a **one-unit** increase in the independent variable (height) is associated with an increase of **4.5 units** in the dependent variable (weight).

Ordinary Least Squares

- For our data sample, this means that a 1-inch increase in height is associated with a 4.5 pound increase in weight.
- Thus, if we had no other information, our best guess for the weight of a person who is 70 inches tall in the Changing Lives study would be –135 + 4.5 (70) = 180 pounds.
- This is our payoff, as we have now quantified the best fit for the linear relationship between height and weight for the Changing Lives participants.
- The same basic tools can be used to explore more complex relationships and more socially significant questions.

sign, size, and significance.

- For any regression coefficient, you will generally be interested in three things: sign, size, and significance.
- ***Sign***. The sign (positive or negative) on the coefficient for an independent variable tells us the direction of its association with the dependent variable.
- In the simple case above, the coefficient on height is positive.
- Taller people tend to weigh more.
- Some relationships will work in the other direction.
- I would expect the association between exercise and weight to be negative.
- Running more is associated with weighing less.

sign, size, and significance.

- ***Size***. How big is the observed effect between the independent variable and the dependent variable?
- Is it of a magnitude that matters? In an explanation of why some people weigh more than others, height is clearly an important factor.
- In other studies, we may find an explanatory variable that has a statistically significant impact on our outcome of interest—meaning that the observed effect is not likely to be a product of chance—but that effect may be so small as to be trivial or *socially insignificant*.

sign, size, and significance.

- For example, suppose that we are examining determinants of income.
- Why do some people make more money than others?
- The explanatory variables are likely to be things like education, years of work experience, and so on.
- In a large data set, researchers might also find that people with whiter teeth earn $86 more per year than other workers, ceteris paribus.
- ("Ceteris paribus" comes from the Latin meaning "other things being equal.")

sign, size, and significance.

- The positive and statistically significant coefficient on the “white teeth” variable assumes that the individuals being compared are similar in other respects: same education, same work experience, and so on.
- Our statistical analysis has demonstrated that whiter teeth are associated with $86 in additional annual income per year and that this finding is not likely to be a mere coincidence.
- This means
- (1) we’ve rejected the null hypothesis that really white teeth have no association with income with a high degree of confidence; and
- (2) if we analyze other data samples, we are likely to find a similar relationship between good-looking teeth and higher income.

sign, size, and significance.

- So what? We've found a statistically significant result, but not one that is particularly meaningful.
- To begin with, $86 per year is not a life-changing sum of money.
- From a public policy standpoint, $86 is also probably less than it would cost to whiten an individual's teeth every year, so we can't even recommend that young workers make such an investment.
- And, although I'm getting a chapter ahead of myself, I'd also be worried about some serious methodological problems.

sign, size, and significance.

- For example, having perfect teeth may be associated with other personality traits that explain the earnings advantage; the earnings effect may be caused by the kind of people who care about their teeth, not the teeth themselves.
- For now, the point is that we should take note of the size of the association that we observe between the explanatory variable and our outcome of interest.

sign, size, and significance.

- ***Significance.*** Is the observed result an aberration based on a quirky sample of data, or does it reflect a meaningful association that is likely to be observed for the population as a whole?
- This is the same basic question that we have been asking for the last several chapters.
- In the context of height and weight, do we think that we would observe a similar positive association in other samples that are representative of the population?
- Our regression coefficient is based on an observed relationship between height and weight for a particular sample of data.
- If we were to test another large sample of data, we would almost certainly get a slightly different association between height and weight and therefore a different coefficient.

RESEARCH
Khatam University
FANAP

sign, size, and significance.

- The relationship between height and weight observed in the Whitehall data is likely to be different from the relationship observed between height and weight for the participants in the Changing Lives study.
- However, we know from the central limit theorem that the mean for a large, properly drawn sample will not typically deviate wildly from the mean for the population as a whole.
- Similarly, we can assume that the observed relationship between variables like height and weight will not typically bounce around wildly from sample to sample, assuming that these samples are large and properly drawn from the same population.

sign, size, and significance.

- Think about the intuition: It's highly unlikely (though still possible) that we would find that every inch of height is associated with 4.5 additional pounds among the Changing Lives participants but that there is no association between height and weight in a different representative sample of 3,000 adult Americans.
- This should give you the first clue of how we'll test whether our regression results are statistically significant or not.
- As with polling and other forms of inference, we can calculate a standard error for the regression coefficient.

sign, size, and significance.

- The standard error is a measure of the likely dispersion we would observe in the coefficient if we were to conduct the regression analysis on repeated samples drawn from the same population.
- • a different sample of 3,000 Americans, we might find that each inch of height is associated with 4.3 pounds.
- • for another sample of 3,000 Americans, we might find that each inch is associated with 5.2 pounds.
- Once again, the normal distribution is our friend.
- For large samples of data, we can assume that our various coefficients will be distributed normally around the “true” association between height and weight in the American adult population.

sign, size, and significance.

- On that assumption, we can calculate a standard error for the regression coefficient that gives us a sense of how much dispersion we should expect in the coefficients from sample to sample.
- However, I must warn that when we are working with a small sample of data—such as a group of 20 adults rather than the 3,000+ persons the normal distribution is no longer willing to be our friend.
- Specifically, if we repeatedly conduct regression analysis on different small samples, we can no longer assume that our various coefficients will be distributed normally around the "true" association between height and weight in the American adult population.

t-distribution

- Instead, our coefficients will still be distributed around the “true” association between height and weight for the American adult population in what is known as a t-distribution.
- Nothing else changes; any basic statistical software package will easily manage the additional complexity associated with using the t-distributions.
- For this reason, the t-distribution will be explained in greater detail in the chapter appendix.

why the standard error matters?

- Sticking with large samples for now (and the normal distribution), the most important thing to understand is why the standard error matters.
- As with polling and other forms of inference, we expect that more than half of our observed regression coefficients will lie within one standard error of the true population parameter.
- Roughly 95 % will lie within two standard errors. And so on.
- With that, we're just about home, because now we can do a little hypothesis testing.
- Once we have a coefficient and a standard error, we can test the null hypothesis that there is in fact no relationship between the explanatory variable and the dependent variable (meaning that the true association between the two variables in the population is zero).

RESEARCH
Khatam University
FANAP

why the standard error matters?

- In our simple height and weight example, we can test how likely it is that we would find in our Changing Lives sample that every inch of height is associated with 4.5 pounds if there is really no association between height and weight in the general population.
- I've run the regression by using a basic statistics program; the standard error on the height coefficient is .13.
- This means that if we were to do this analysis repeatedly—say with 100 different samples—then we would expect our observed regression coefficient to be within 2 SE of the true population parameter roughly 95 times out of 100.

why the standard error matters?

- We can therefore express our results in two different but related ways.
- • First, we can build a 95% confidence interval.
- We can say that 95 times out of 100, we expect our confidence interval, which is 4.5 ± .26, to contain the true population parameter.
- This is the range between 4.24 and 4.76.
- there is only a 5 % chance that we are wrongly rejecting the null hypothesis.

why the standard error matters?

- We can therefore express our results in two different but related ways.
- • Second, we can see that our 95 %confidence interval (4.24 and 4.76) for the true association between height and weight does not include zero.
- Thus, we can reject the null hypothesis that there is *no association* between height and weight for the general population at the 95 % confidence level.
- • This result can also be expressed as being statistically significant at the .05 level;
- there is only a 5 % chance that we are wrongly rejecting the null hypothesis.

why the standard error matters?

- In fact, our results are even more extreme than that.
- The standard error (.13) is extremely low relative to the size of the coefficient (4.5).
- One rough rule of thumb is that the coefficient is likely to be statistically significant when the coefficient is at least twice the size of the standard error.*
- A statistics package also calculates a p-value, which is .000 in this case, meaning that there is essentially zero chance of getting an outcome as extreme as what we've observed (or more so) if there is no true association between height and weight in the general population.

why the standard error matters?

- Remember, we have not *proved* that taller people weigh more in the general population;
- we have merely shown that our results for the Changing Lives sample would be highly anomalous if that were not the case.

the R^2

- one other statistic of note: the R2, which is a measure of the total amount of variation explained by the regression equation.
- We know that we have a broad variation in weight in our Changing Lives sample.
- Many of the persons in the sample weigh more than the mean for the group overall; many weigh less.
- The R^2 tells us how much of that variation around the mean is associated with differences in height alone.
- The answer in our case is .25, or 25 percent.

It's time to take off the training wheels!

- The more significant point may be that 75 % of the variation in weight for our sample remains unexplained.
- There are clearly factors other than height that might help us understand the weights of the Changing Lives participants.
- This is where things get more interesting.
- As I've promised, regression analysis allows us to unravel complex relationships in which multiple factors affect some outcome that we care about, such as income, or test scores, or heart disease.
- When we include multiple variables in the regression equation, the analysis gives us an estimate of the linear association between *each* explanatory variable and the dependent variable while holding other dependent variables constant, or "controlling for" these other factors.

RESEARCH

It's time to take off the training wheels!

- Let's stick with weight for a while.
- We've found an association between height and weight; we know there are other factors that can help to explain weight (age, sex, diet, exercise …).
- Regression analysis (often called multiple regression analysis or multivariate regression analysis) will give us a coefficient for each explanatory variable included in the regression equation.
- In other words, *among people who are the same sex and height*, what is the relationship between age and weight?

adding one more variable to the equation

- Once we have more than one explanatory variable, we can no longer plot these data in two dimensions.
- Yet the basic methodology is the same as in our simple height and weight example.
- As we add explanatory variables, a statistical package will calculate the regression coefficients that minimize the total sum of the squared residuals for the regression equation.
- Let's work with the Changing Lives data for now; We can start by adding one more variable to the equation that explains the weights of the Changing Lives participants: age.

adding one more variable to the equation

- When we run the regression including both height and age as explanatory variables for weight, here is what we get.

WEIGHT = –145 + 4.6 × (HEIGHT in INCHES**) + .1 × (AGE IN** YEARS**)**

- The coefficient on age is 0.1.
- That can be interpreted to mean that every additional year in age is associated with 0.1 additional pounds in weight, *holding height constant*.
- For any group of people who are the same height, on average those who are ten years older will weigh one pound more.

- This is not a huge effect, but it's consistent with what we tend to see in life.
- **The coefficient is significant at the .05 level**.
- You may have noticed that the coefficient on height has increased slightly.
- Once age is in our regression, we have a more refined understanding of the relationship between height and weight.
- Among people who are the same age in our sample, or "holding age constant," every additional inch in height is associated with 4.6 pounds in weight.

Let's add one more variable: sex.

- This will be slightly different because sex can only take on two possibilities, male or female.
- How does one put M or F into a regression? The answer is that we use what is called a binary variable, or dummy variable.
- In our data set, we enter a 1 for those participants who are female and a 0 for those who are male.
- The sex coefficient can then be interpreted as the effect on weight of being female, ceteris paribus. The coefficient is –4.8, which is not surprising.
- **We can interpret that to mean that for individuals who are the same height and age, women typically weigh 4.8 pounds less than men.**

RESEARCH

Let's add one more variable: sex.

- We know that women tend to be shorter than men, but our coefficient takes this into account since we have already controlled for height.
- What we have isolated here is the effect of being female.
- The new regression becomes:

WEIGHT = –118 + 4.3 × (HEIGHT IN INCHES) + .12 (AGE IN YEARS) – 4.8 (IF SEX IS FEMALE)

- Our best estimate of the weight of a 53-year-old woman who is 65 inches is: –118 + 4.3 (65) + .12 (53) – 4.8 = 163 pounds.
- And our best guess for a 35-year-old male who is 6 feet 3 inches is –118 + 4.3 (75) + .12 (35) = 209 pounds.
- We skip the last term in our regression result (–4.8) since this person is not female.

Let's add one more variable: sex.

- Now we can start to test things that are more interesting and less predictable.
- What about education?
- How might that affect weight?
- I would hypothesize that better-educated individuals are more health conscious and therefore will weigh less, ceteris paribus.
- We also haven't tested any measure of exercise; I assume that, holding other factors constant, the people in the sample who get more exercise will weigh less.

Let's add one more variable: **poverty**.

- What about **poverty**? Does being low-income in America have effects that manifest themselves in weight?
- The Changing Lives study asks whether the participants are receiving food stamps, which is a good measure of poverty in America.
- Finally, I'm interested in **race**. We know that people of color have different life experiences in the US *because of their race*.
- There are cultural and residential factors associated with race in America that have implications for weight.
- Many cities are still characterized by a high degree of racial segregation African Americans might be more likely than other residents to live in "food deserts," which are areas with limited access to grocery stores that carry fruits, vegetables, and other fresh produce.

Let's add one more variable: **race**.

- We can use regression analysis to separate out the independent effect of each of the potential explanatory factors described above.
- For example, we can isolate the association between race and weight, holding constant other socioeconomic factors like educational background and poverty.
- *Among people who are high school graduates and eligible for food stamps, what is the statistical association between weight and being black?*

one more variable: **race.**

- At this point, our regression equation is so long that it would be cumbersome to print the results in their entirety here.
- Academic papers typically insert large tables that summarize the results of various regression equations.
- I have included a table with the complete results of this regression equation in the appendix to this chapter.
- In the meantime, here are the highlights of what happens when we add education, exercise, poverty (as measured by receiving food stamps), and race to our equation.

Let's add one more variable: **race**.

- All of our original variables (height, age, and sex) are still significant.
- The coefficients change little as we add explanatory variables.
- All of our new variables are statistically significant at the .05 level.
- The R2 on the regression has climbed from .25 to .29.
- (Remember, an R2 of zero means that our regression equation does no better than the mean at predicting the weight of any individual in the sample;
- an R2 of 1 means that the regression equation perfectly predicts the weight of every person in the sample.)
- A lot of the variation in weight across individuals remains unexplained.

RESEARCH

Education and exercise.

- *Education turns out to be negatively associated with weight, as I had hypothesized.*
- Among participants in the Changing Lives study, each year of education is associated with –1.3 pounds.
- Not surprisingly, **exercise is also negatively associated with weight**.
- The Changing Lives study includes an index that evaluates each participant in the study on his or her level of physical activity.
- Those individuals who are in the bottom quintile of physical activity weigh, on average, 4.5 pounds more than other adults in the sample, ceteris paribus.
- Those in the bottom quintile f or physical activity weigh, on average, nearly 9 pounds more than adults in the top quintile for physical activity.

poverty.

- *Individuals receiving food stamps (the proxy for poverty in this regression) are heavier than other adults.*
- Food stamp recipients weigh an average of 5.6 pounds more than other Changing Lives participants, ceteris paribus.

race.

- The race variable turns out to be particularly interesting.
- Even after one controls for all the other variables described up to this point, race still matters a lot when it comes to explaining weight.
- The non-Hispanic black adults in the Changing Lives sample weigh, on average, roughly 10 pounds more than the other adults in the sample.
- Ten pounds is a lot of weight, both in absolute terms and compared with the effects of the other explanatory variables in the regression equation.
- This is not a quirk of the data.
- The p-value on the dummy variable for non-Hispanic blacks is .000 and the 95 % confidence interval stretches from 7.7 pounds to 16.1 pounds.

RESEARCH
Khatam University
FANAP

variable: **race**.

- What is going on? The honest answer is that I have no idea.
- Let me reiterate a point that was buried earlier in a footnote: I'm just playing around with data here to illustrate how regression analysis works.
- The analytics presented here are to true academic research what street hockey is to the NHL.
- If this were a real research project, there would be weeks or months of follow-on analysis to probe this finding.

variable: **race**.

- What I can say is that I have demonstrated why multiple regression analysis is the best tool we have for finding meaningful patterns in large, complex data sets.
- We started with a ridiculously banal exercise: quantifying the relationship between height and weight. Before long, we were knee-deep in issues with real social significance.
- In that vein, I can offer you a real study that used regression analysis to probe a socially significant issue: gender discrimination in the workplace.
- The curious thing about discrimination is that it's hard to observe directly.

more variable: **race**.

- No employer ever states explicitly that someone is being paid less because of his or her race or gender or that someone has not been hired for discriminatory reasons (which would presumably leave the person in a different job with a lower salary).
- Instead, what we observe are gaps in pay by race and gender that may be the result of discrimination: whites earn more than blacks;
- men earn more than women; and so on.

more variable: **race**.

- The methodological challenge is that these observed gaps may also be the result of underlying differences in workers that have nothing to do with workplace discrimination, such as the fact that women tend to choose more part-time work.
- How much of the wage gap is due to factors associated with productivity on the job, and how much of the gap, if any, is due to labor force discrimination?
- No one can claim that this is a trivial question.

more variable: **race**.

- Regression analysis can help us answer it.
- However, our methodology will be slightly more roundabout than it was with our analysis explaining weight.
- Since we cannot measure discrimination directly, we will examine other factors that traditionally explain wages, such as education, experience, occupational field, and so on.
- The case for discrimination is circumstantial (indirect):
- If a significant wage gap remains after controlling for other factors that typically explain wages, then discrimination is a likely culprit **(guilty)**.

more variable: **race**.

- The larger the unexplained portion of any wage gap, the more suspicious we should be.
- As an example, let's look at a paper by three economists examining the wage trajectories of a sample of roughly 2,500 men and women who graduated with MBAs from the Booth School of Business at the University of Chicago.
- Upon graduation, male and female graduates have very similar average starting salaries: $130,000 for men and $115,000 for women.
- After ten years in the workforce, however, a huge gap has opened up; women on average are earning a striking 45 % less than their male classmates: $243,000 versus $442,000.

more variable: **race**.

- In a broader sample of more than 18,000 MBA graduates who entered the workforce between 1990 and 2006, being female is associated with 29 % lower earnings.
- What is happening to women once they enter the labor force?
- According to the authors of the study, discrimination is *not* a likely explanation for most of the gap.
- The gender wage gap fades away as the authors add more explanatory variables to the analysis.

one more variable: race.

- For example, men take more finance classes in the MBA program and graduate with higher grade point averages.
- When these data are included as control variables in the regression equation, the unexplained portion of the gap in male-female earnings drops to 19 percent.
- When variables are added to the equation to account for post-MBA work experience, particularly spells out of the labor force, the unexplained portion of the male-female wage gap drops to 9 percent.
- And when explanatory variables are added for other work characteristics, such as employer type and hours worked, the unexplained portion of the gender wage gap falls to under 4 percent.

one more variable: **race.**

- For workers who have been in the labor force more than ten years, the authors can ultimately explain all but 1 % of the gender wage gap with factors unrelated to discrimination on the job.*
- They conclude, “We identify three proximate reasons for the large and rising gender gap in earnings:
- differences in training prior to MBA graduation;
- differences in career interruptions; and differences in weekly hours.
- These three determinants can explain the bulk of gender differences across the years following MBA completion.”

one more variable: **race**.

- I hope that I've convinced you of the value of multiple regression analysis, particularly the research insights that stem from being able to isolate the effect of one explanatory variable while controlling for other confounding factors.
- I have not yet provided an intuitive explanation for how this statistical "miracle elixir" works.
- When we use regression analysis to evaluate the relationship between education and weight, how does a statistical package control for factors like height, sex, age, and income when we know that our Changing Lives participants are *not* identical in these other respects?

one more variable: **race**.

- To get your mind around how we can isolate the effect on weight of a single variable, say, education, imagine the following situation.
- Assume that all of the Changing Lives participants are gathered in one place.
- Now assume that the men and women are separated.
- And then assume that both the men and the women are further divided by height.
- There will be a room of six-foot tall men.
- Next door, there will be a room of 6-foot 1-inch men, and so on for both genders.

one more variable: **race**.

- If we have enough participants in our study, we can further subdivide each of those rooms by income.
- Eventually we will have lots of rooms, each of which contains individuals who are identical in all respects *except for education and weight, which are the two variables we care about*.
- There would be a room of 45-year-old 5-foot 5-inch men who earn $30,000 to $40,000 a year.
- Next door would be all the 45-year-old 5-foot 5-inch women who earn $30,000 to $40,000 a year. And so on (and on and on).

one more variable: **race**.

- There will still be some variation in weight in each room; people who are the same sex and height and have the same income will still weigh different amounts—though presumably there will be much less variation in weight in each room than there is for the overall sample.
- Our goal now is to see how much of the remaining variation in weight in each room can be explained by education.
- In other words, what is the best linear relationship between education and weight in each room?

one more variable: **race**.

- The final challenge, however, is that we do not want different coefficients in each “room.”
- The whole point of this exercise is to calculate a single coefficient that best expresses the relationship between education and weight for the entire sample, while holding other factors constant.
- What we would like to calculate is the single coefficient for education that we can use *in every room* to minimize the sum of the squared residuals for all of the rooms combined.
- What coefficient for education minimizes the square of the unexplained weight for every individual across all the rooms?

one more variable: race.

- That becomes our regression coefficient because it is the best explanation of the linear relationship between education and weight for this sample when we hold sex, height, and income constant.
- As an aside, you can see why large data sets are so useful.
- They allow us to control for many factors while still having many observations in each “room.”
- Obviously a computer can do all of this in a split second without herding thousands of people into different rooms.

one more variable: **race**.

- Let's finish the chapter where we started, with the connection between stress on the job and coronary heart disease.
- The Whitehall studies of British civil servants sought to measure the association between grade of employment and death from coronary heart disease over the ensuing years.
- One of the early studies followed 17,530 civil servants for seven and a half years.
- The authors concluded, "Men in the lower employment grades were shorter, heavier for their height, had higher blood pressure, higher plasma glucose, smoked more, and reported less leisure-time physical activity than men in the higher grades.

RESEARCH

one more variable: **race.**

- Yet when allowance was made for the influence on mortality of all of these factors plus plasma cholesterol, the inverse association between grade of employment and [coronary heart disease] mortality was still strong."
- The "allowance" they refer to for these other known risk factors is done by means of regression analysis.*
- The study demonstrates that holding other health factors constant (including height, which is a decent proxy for early childhood health and nutrition), working in a "low grade" job can literally kill you.

one more variable: **race**.

- Skepticism is always a good first response.
- I wrote at the outset/start of the chapter that “low-control” jobs are bad for your health.
- That may or may not be synonymous with being low on the administrative totem pole.
- A follow-up study using a second sample of 10,308 British civil servants sought to drill down on this distinction.
- The workers were once again divided into administrative grades—high, intermediate, and low—only this time the participants were also given a fifteen-item questionnaire that evaluated their level of “decision latitude or control.”

one more variable: **race**.

- These included questions such as “Do you have a choice in deciding how you do your job?” and categorical responses (ranging from “never” to “often”) to statements such as “I can decide when to take a break.”
- The researchers found that the “low-control” workers were at significantly higher risk of developing coronary heart disease over the course of the study than “high control” workers.
- Yet researchers also found that workers with hard job demands were at no greater risk of developing heart disease, nor were workers who reported low levels of social support on the job. Lack of control seems to be the killer, literally.

one more variable: **race.**

- The Whitehall studies have two characteristics typically associated with strong research.
- First, the results have been replicated elsewhere.
- In the public health literature, the “low-control” idea evolved into a term known as “job strain,” which characterizes jobs with “high psychological workload demands” and “low decision latitude.”
- Between 1981 and 1993, thirty-six studies were published on the subject; most found a significant positive association between job strain and heart disease.

one more variable: **race.**

- Second, researchers sought and found corroborating/confirming biological evidence to explain the mechanism by which this particular kind of stress on the job causes poor health.
- Work conditions that involve rigorous demands but low control can cause physiological responses (such as the release of stress-related hormones) that increase the risk of heart disease over the long run.
- Even animal research plays a role; low-status monkeys 6and baboons (who bear some resemblance to civil servants at the bottom of fthe authority chain) have physiological differences from their high-status peers that put them at greater cardiovascular risk.

one more variable: **race**.

- All else equal, it's better not to be a low-status baboon, which is a point I try to make to my children as often as possible, particularly my son.
- The larger message here is that regression analysis is arguably the most important tool that researchers have for finding meaningful patterns in large data sets.
- We typically cannot do controlled experiments to learn about job discrimination or factors that cause heart disease.
- Our insights into these socially significant issues and many others come from the statistical tools covered in this chapter.

RESEARCH

one more variable: race.

- In fact, it would not be an exaggeration to say that a high proportion of all important research done in the social sciences over the past half century (particularly since the advent of cheap computing power) draws on regression analysis.
- Regression analysis supersizes the scientific method; we are healthier, safer, and better informed as a result.
- So what could possibly go wrong with this powerful and impressive tool?
- Read on.

Common Regression Mistakes

The mandatory warning label

Here is one of the most important things to remember when doing research that involves regression analysis: Try not to kill anyone. You can even put a

little Post-it note on your computer monitor:

“Do not kill people with your research.”

Because some very smart people have inadvertently violated that rule.

Nurses' Health Study

- Beginning in the 1990s, the medical establishment agreed around the idea that older women should take estrogen supplements to protect against heart disease, osteoporosis, and other conditions associated with menopause.
- By 2001, some 15 million women were being prescribed estrogen in the belief that it would make them healthier. Why?
- Because research at the time—using the basic methodology—suggested this was a sensible medical strategy.
- In particular, a longitudinal study of 122,000 women (the Nurses' Health Study) found a negative association between estrogen supplements and heart attacks.

Nurses' Health Study

- Women taking estrogen had one-third as many heart attacks as women who were not taking estrogen.
- This was not a couple of teenagers using dad's computer to run regression equations.
- The Nurses' Health Study is run by the Harvard Medical School and the Harvard School of Public Health.

hormone replacement therapy

- Meanwhile, scientists and physicians offered a medical theory for why hormone supplements might be beneficial for female health.
- A woman's ovaries produce less estrogen as she ages; if estrogen is important to the body, then making up for this deficit in old age could be protective of a woman's long-term health.
- Hence the name of the treatment: hormone replacement therapy.
- Some researchers even began to suggest that older men should receive an estrogen boost.

hormone replacement therapy

- And then, while millions of women were being prescribed hormone replacement therapy, estrogen was subjected to the most rigorous form of scientific study: clinical trials.
- Rather than searching a large data set like the Nurses' Health Study for statistical associations that may or may not be causal, a clinical trial consists of a controlled experiment.
- One sample is given a treatment, such as hormone replacement;
- Another sample is given a placebo.
- Clinical trials showed that women taking estrogen had a higher incidence of heart disease, stroke, blood clots, breast cancer, and other adverse health outcomes.

Clinical trials

- Estrogen supplements did have some benefits, but those benefits were far outweighed by other risks.
- Beginning in 2002, doctors were advised not to prescribe estrogen for their aging female patients.
- The New York Times Magazine asked a delicate but socially significant question: How many women died prematurely or suffered strokes or breast cancer because they were taking a pill that their doctors had prescribed to keep them healthy?
- The answer: "A reasonable estimate would be tens of thousands."

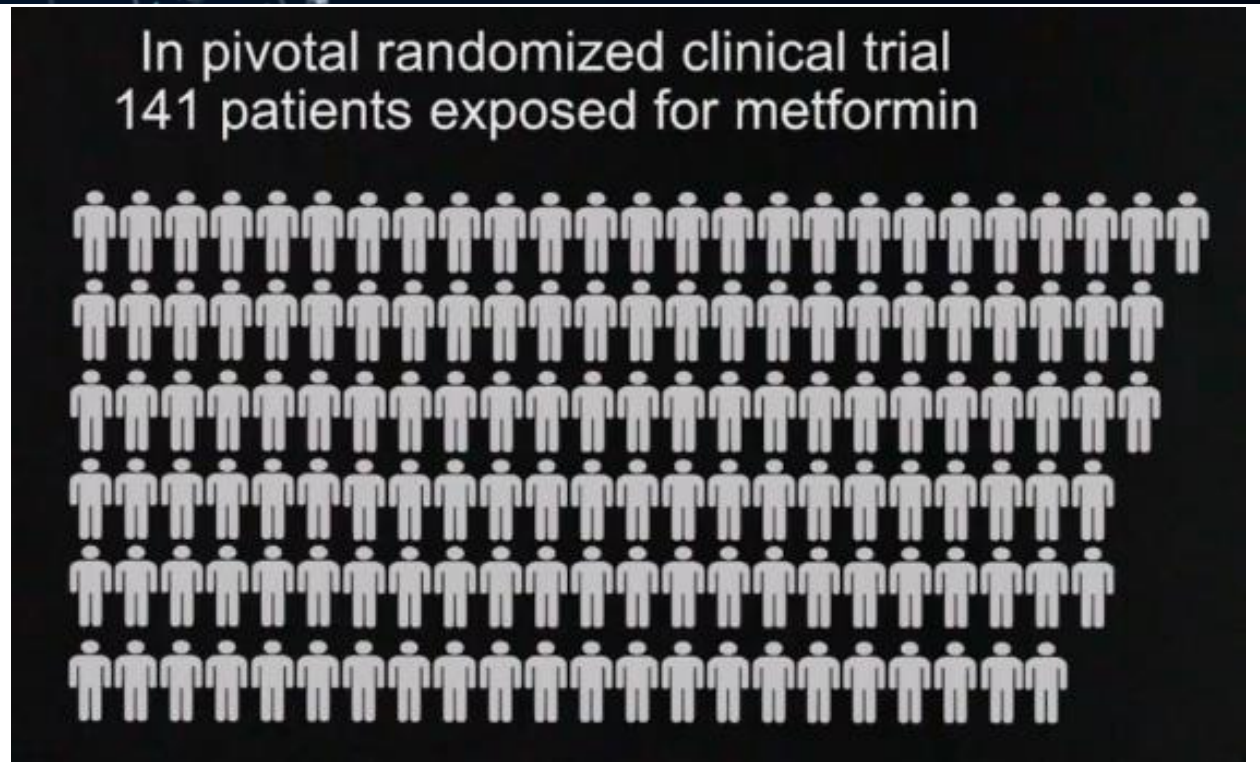
Clinical trials

- Regression analysis is the hydrogen bomb of the statistics arsenal.
- Every person with a personal computer and a large data set can be a researcher in his or her own home or cubicle.
- What could possibly go wrong? All kinds of things.
- Regression analysis provides precise answers to complicated questions.
- These answers may or may not be accurate.
- In the wrong hands, regression analysis will yield results that are misleading or just plain wrong.

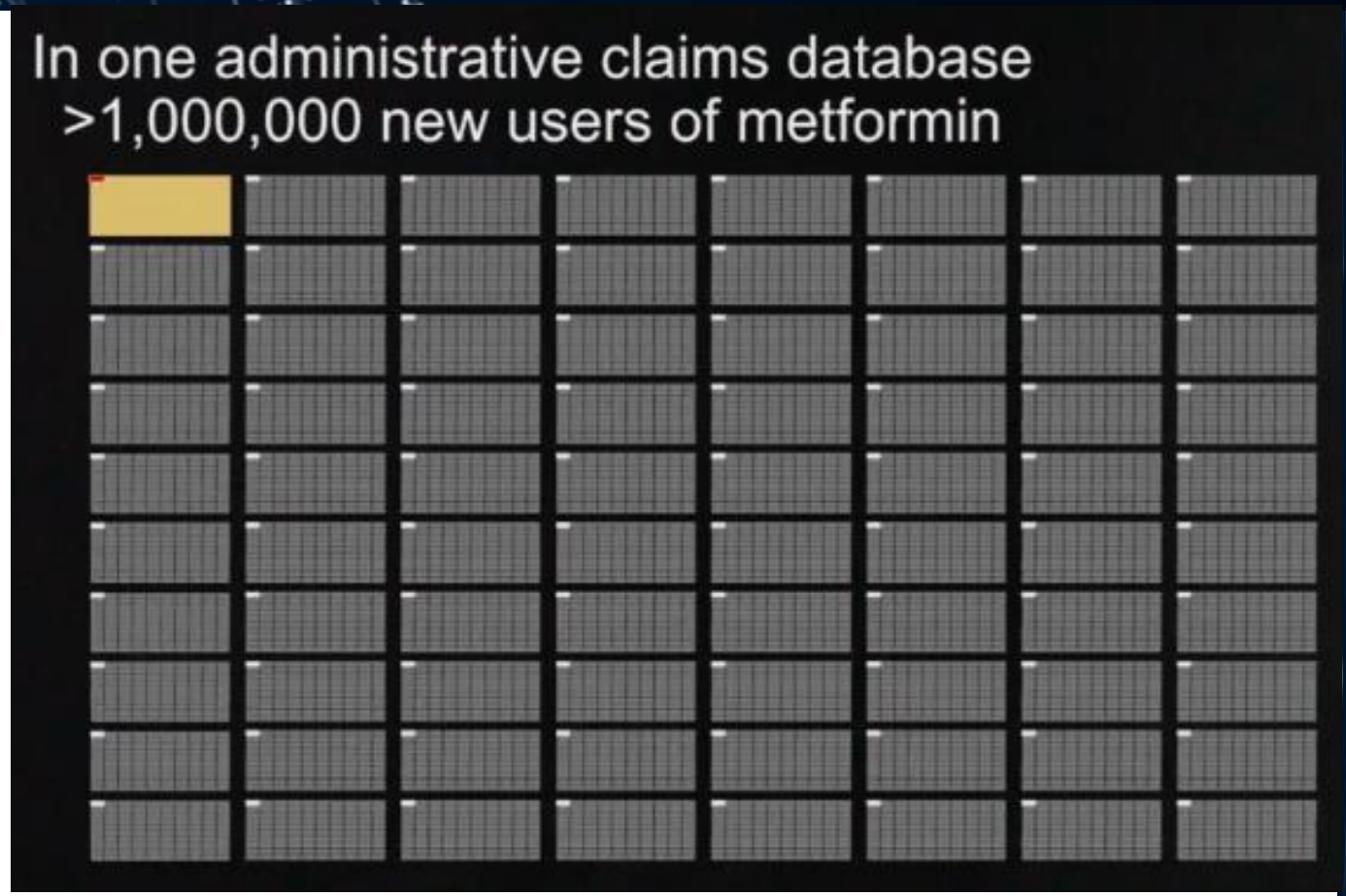
Observational Health Studies

- Very large scale health care data has become available to researchers.
- And there's a lot of excitement about the potential for these data to improve the way we deliver health care and understand the effects of health care interventions.
- Just to illustrate the change that's taken place, consider a drug, that drug called Metformin.
- So Metformin is a widely used medication for diabetes.
- Metformin has been on the market for several decades.
- And when the drug was approved some decades ago by the Food and Drug Administration, there was a pivotal, so-called pivotal randomized clinical trial that was the primary evidence that led people to the drug.

So a concern is suppose there is a **somewhat rare side effect** that is associated with Metformin,
some bad thing that happens to a small percentage of the patients who take Metformin.
Well, there's a very good chance that it was never seen in that randomized trial, because there were only 141 patients.



Flash forward to today, we now have large scale patient level databases, health care databases, that we as researchers have been using, in which, the particular databases we use, there are actually over a million new users of Metformin.

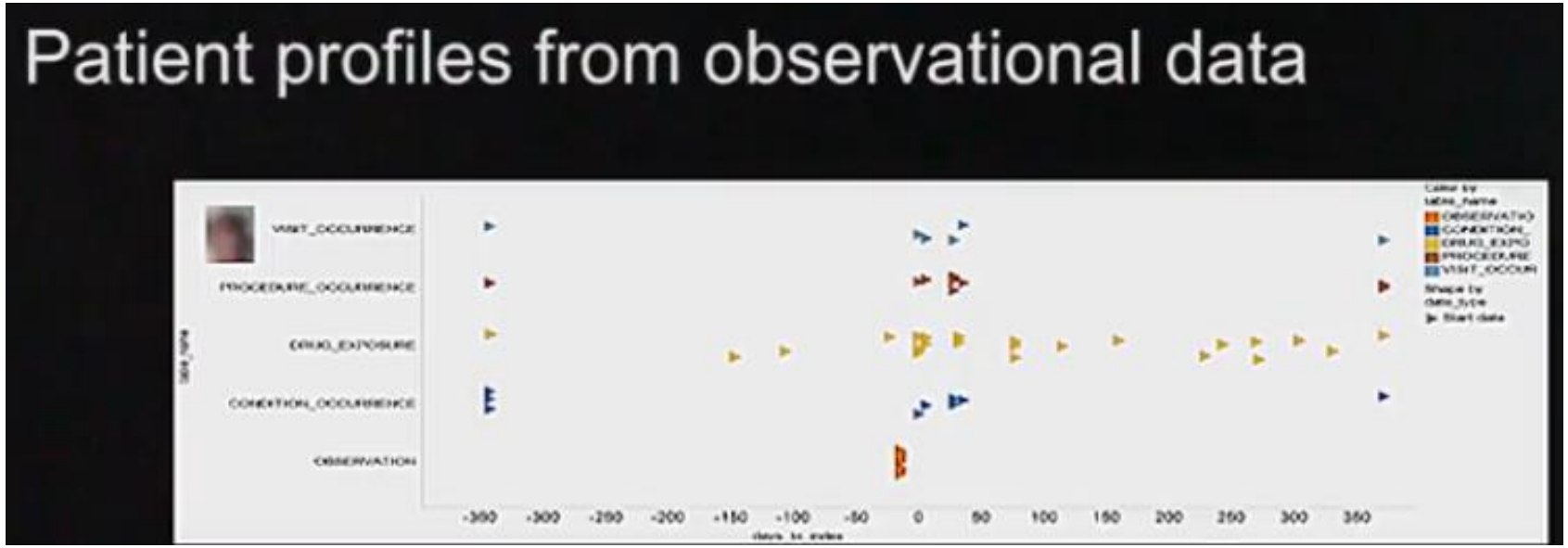


Observational Health Studies

- These are observational data, meaning these are not the product of designed experiments.
- Rather, these data are gathered as a byproduct of health care practice.
- And some of the databases in particular, the ones I'm mostly going to focus on, are so-called **administrative claims databases**.
- So these are generated because of financial transactions in the health care system.
- There are also electronic health record systems which record primary medical records.

Observational Health Studies

- But in any event, the databases that I'm going to focus on here say there are over a million users of Metformin.
- So the idea, excitement here and the idea and the hope is that with that much data, with that many users of Metformin, surely we can figure out what good things and bad things this drug is causing.
- We would like to do causal inference using these databases to infer downstream effects, good and bad, of Metformin. So that's the idea.

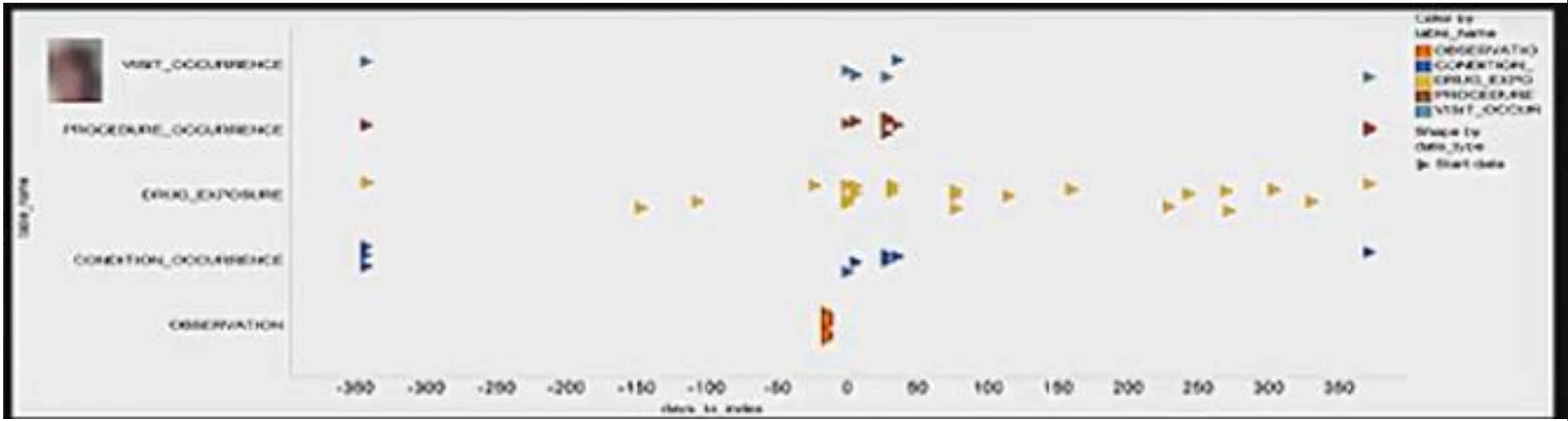


Observational Health Studies

- Now let me show you some examples of what these data look like in a sort of cartoon version.
- These are real data. These are not real patients whose pictures are on the slides. But this is a typical data record for a particular patient.
- what you see on the horizontal axis is time.

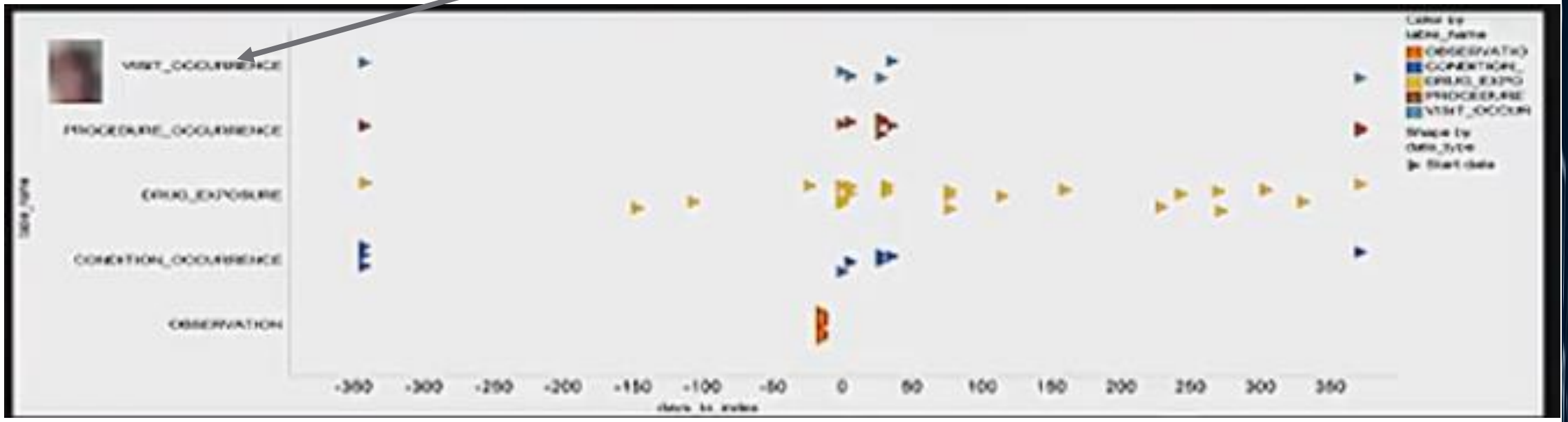
Observational Health Studies

- I'm showing you here about 2 years' worth of data for this particular patient.



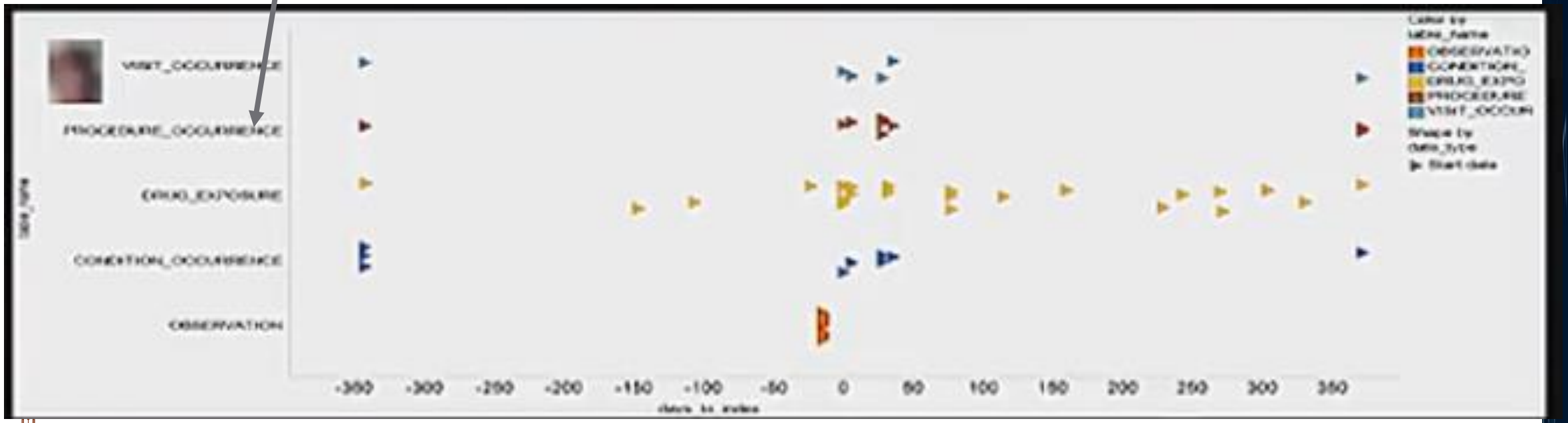
Observational Health Studies

- And then there are different kinds of data that we're showing here for this patient.
- So on the top are visits that this patient made, be that to primary care doctor or an emergency room or an in patient visit or whatever.



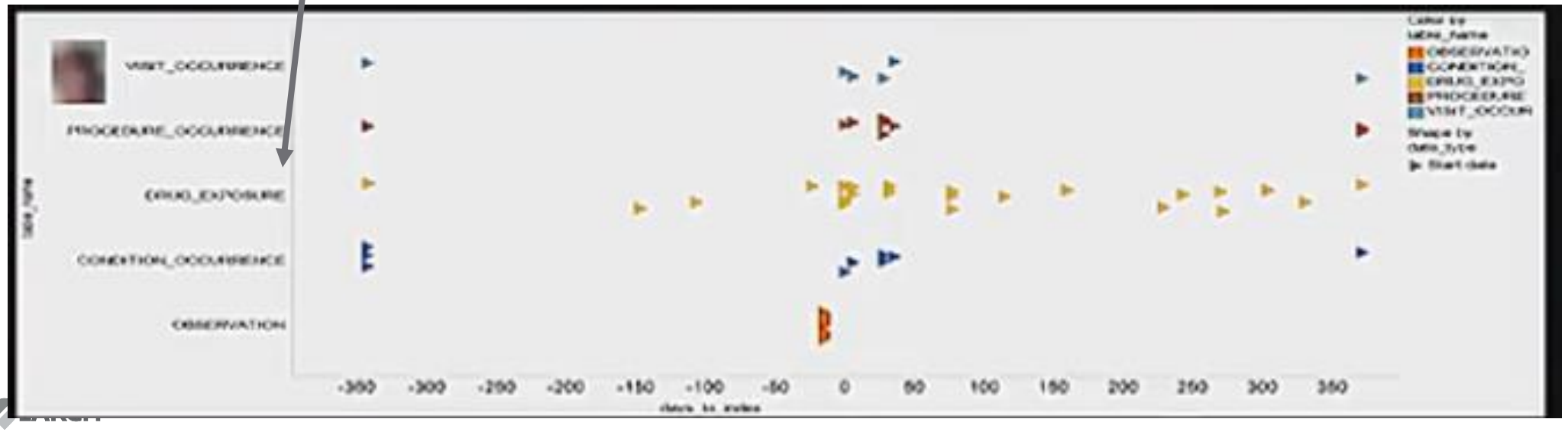
Observational Health Studies

- These are procedures this patient had at different points in time.
- And we know what each of those procedures are.



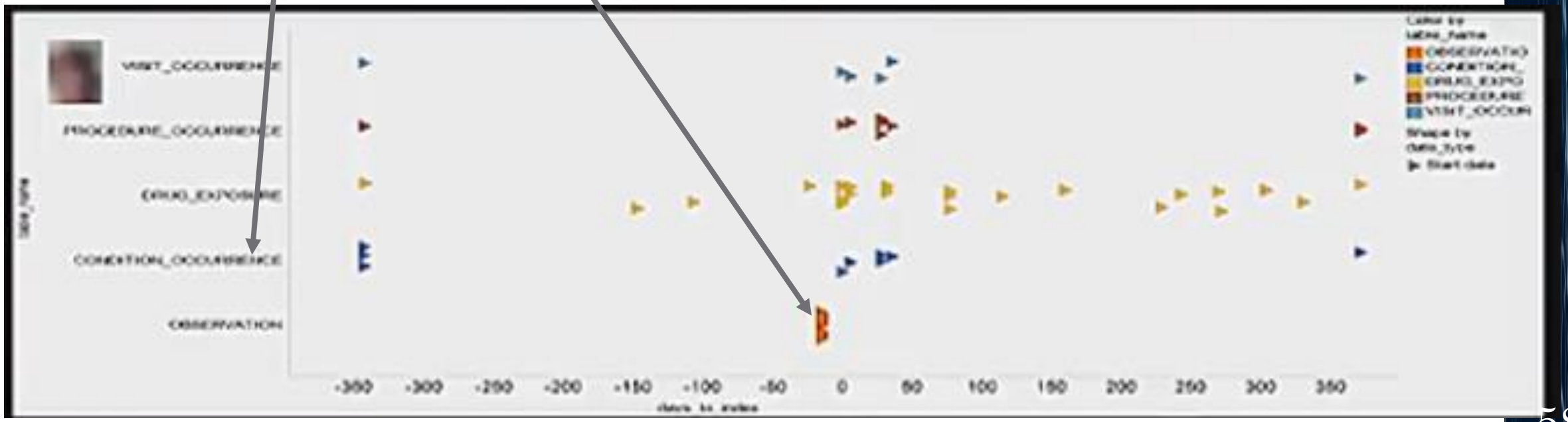
Observational Health Studies

- And these are **drug exposures** where we have a notion of a drug year.
- So we know, for this patient, when they were prescribed a certain drug and for the duration of that prescription.
- And if they repeat the prescription, we can construct these eras during which we believe the patient was consuming a drug.



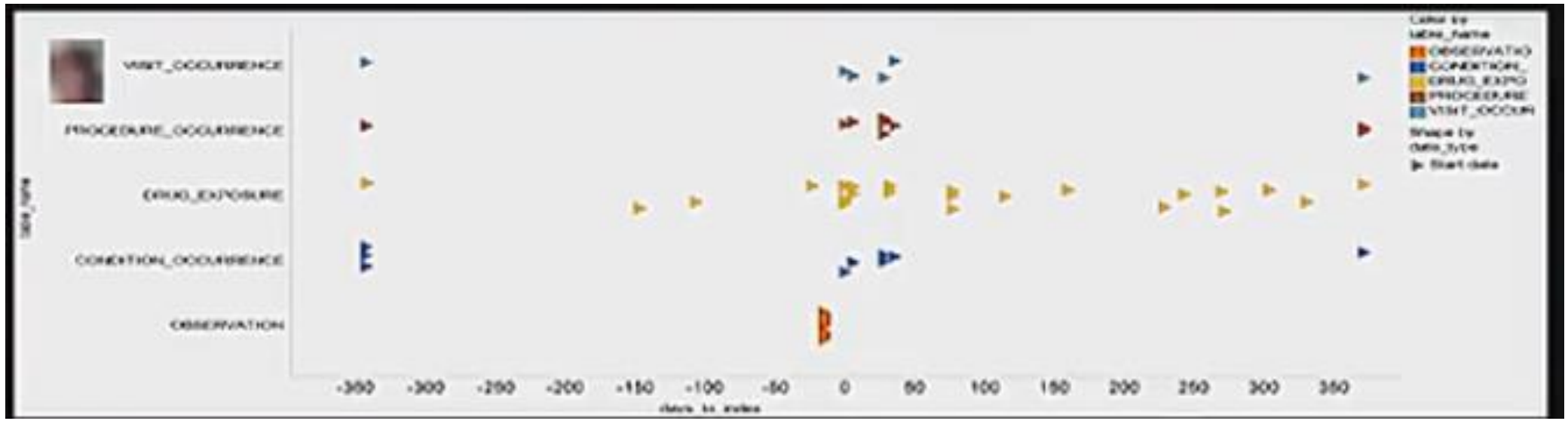
Observational Health Studies

- We also have **conditions** recorded for this patient.
- So these are **diagnosis** that this patient had.
- And finally we have observations. These are basically lab values.
- And for this patient we only have one set of lab values.



Observational Health Studies

- you can see from this patient's record, obviously this patient was having an encounter with the health care system at the beginning of the two year period.
- And there's quite a lot of activity in the middle of the period.
- And then the patient is taking some drugs, essentially, throughout the period. So that's what these data look like.



Observational Health Studies

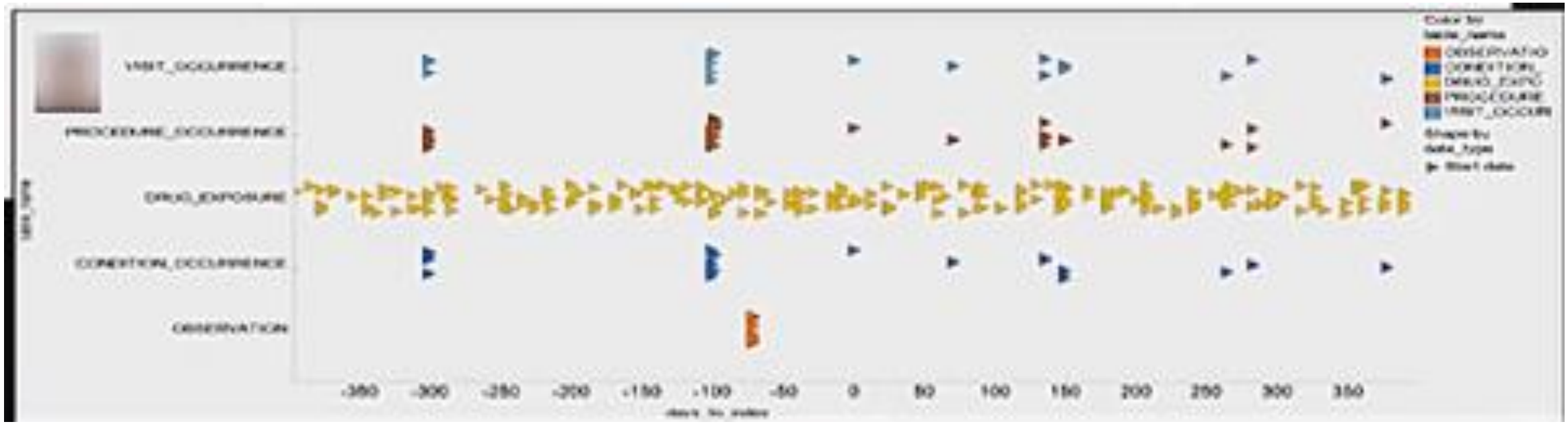
- you can see from this patient's record, obviously this patient was having an encounter with the health care system at the beginning of the two year period.
- And there's quite a lot of activity in the middle of the period.
- And then the patient is taking some drugs, essentially, throughout the period. So that's what these data look like.
- So in the case with Metformin, I can hand you one million of these, of objects that look like this.
- And the task that we are engaged in, we and hundreds or thousands of researchers around the world, is can we use these data to infer the effects of Metformin?

Observational Health Studies

- That's a pretty standard, typical patient and a modest amount of interaction with the health care system.
- Here's a different patient. Again, these are real data.
- This patient has rather a lot more interactions with the health care system.

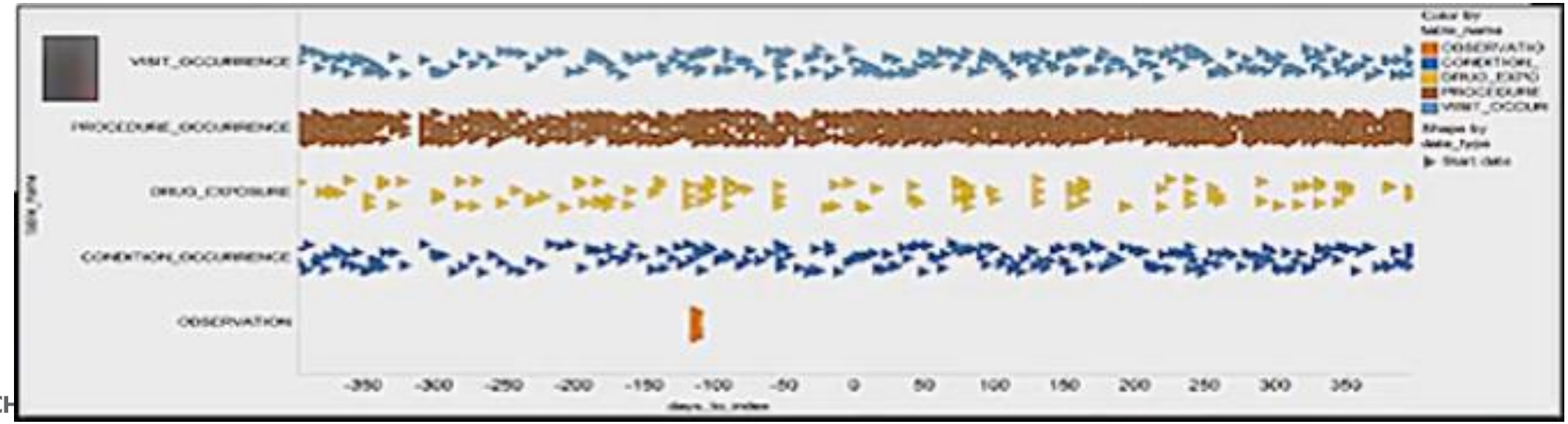
Observational Health Studies

- And in particular, as you can see, this patient is taking quite a few drugs throughout the two year period in this particular case.



Observational Health Studies

- Here's yet another patient.
- And alas, there are actually many patients like this in these databases.
- This is a patient who obviously has had many interactions with the health care system, lots and lots of visits, many procedures, many drugs, and lots of different diagnosis that this patient has had. So the point of this is, and this might sound simple.



Observational Health Studies

- I give you the data for a million patients.
- And you're just trying to figure out what Metformin causes.
- But just glancing at these records, I think it's sort of intuitive that maybe this isn't so simple.
- So look at this third patient here.
- Supposing one of those drugs is Metformin, well, and let's say, a potential side effect of Metformin, let's say, it's a concern about some form of cancer, is one of those conditions that appears later on in this patient.

Observational Health Studies

- Well, it's going to be pretty tricky to figure out was it the Metformin that caused that?
- Or was it something else, some other drug that caused that?
- Or was it something not drug related at all?
- So that's the challenge here is: can we come up with methods, statistical and machine learning methods, to figure out real, basically to distinguish between real causes and spurious associations.
- That's the objective here.
- Now as a matter of fact, many researchers are conducting these kinds of studies in these kinds of databases.
- And medical journals are full of studies using these databases.

RESEARCH

Observational Health Studies

- Here's a typical one.
- This is an article that appeared a few years ago in the British Medical Journal, a very prestigious medical journal.
- And this is a very standard type of paper.
- Pick up any issue of any major medical journal.
- And you will find papers just like this one.

BMJ

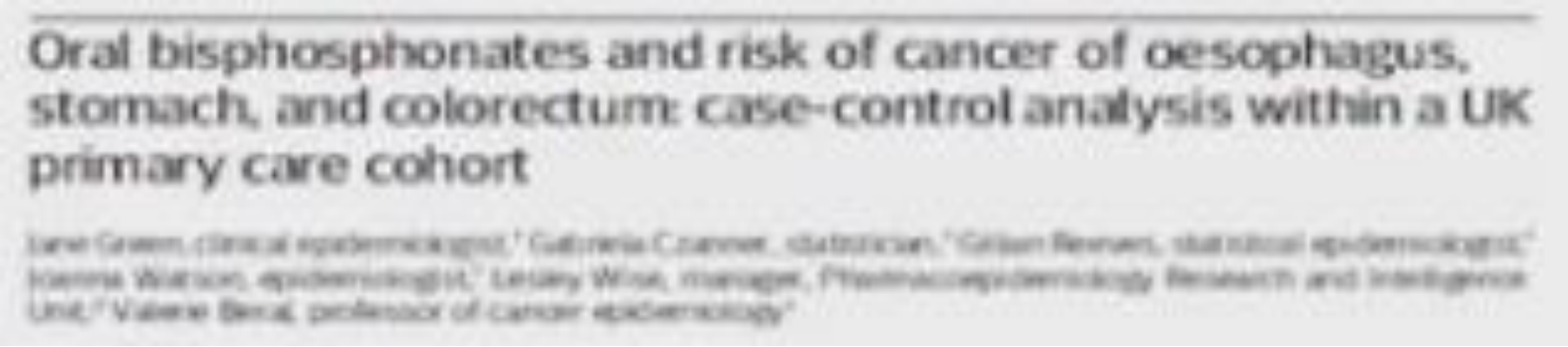
RESEARCH

Oral bisphosphonates and risk of cancer of oesophagus, stomach, and colorectum: case-control analysis within a UK primary care cohort

1. *BMJ* 2010;341:c4444.

Observational Health Studies

- So this particular paper concerns a class of drugs called oral bisphosphonates.
- These are drugs that are prescribed to patients with osteoporosis.
- And so brand names that you might have seen ads for are drugs like Fosamax and Reclast and so on. There are several of them.
- This paper is about oral bisphosphonates, this class of drugs, and the risk of cancer of the esophagus I'm going to focus on here.


Oral bisphosphonates and risk of cancer of oesophagus, stomach, and colorectum: case-control analysis within a UK primary care cohort

Observational Health Studies

- And so they're going to use, in this case, it's an electronic health record database.
- They're going to use a large scale database like the cartoons we were just looking at to try and make inference about a **potential causal association between oral bisphosphonates and cancer of the esophagus سرطان مري.**
- **Bisphosphonates PREVENT THE LOSS OF BONE DENSITY**
- Basically we want to know, do these drugs cause cancer of the esophagus?
- In this particular case, this is actually September 2010, this is the conclusion that they reached.

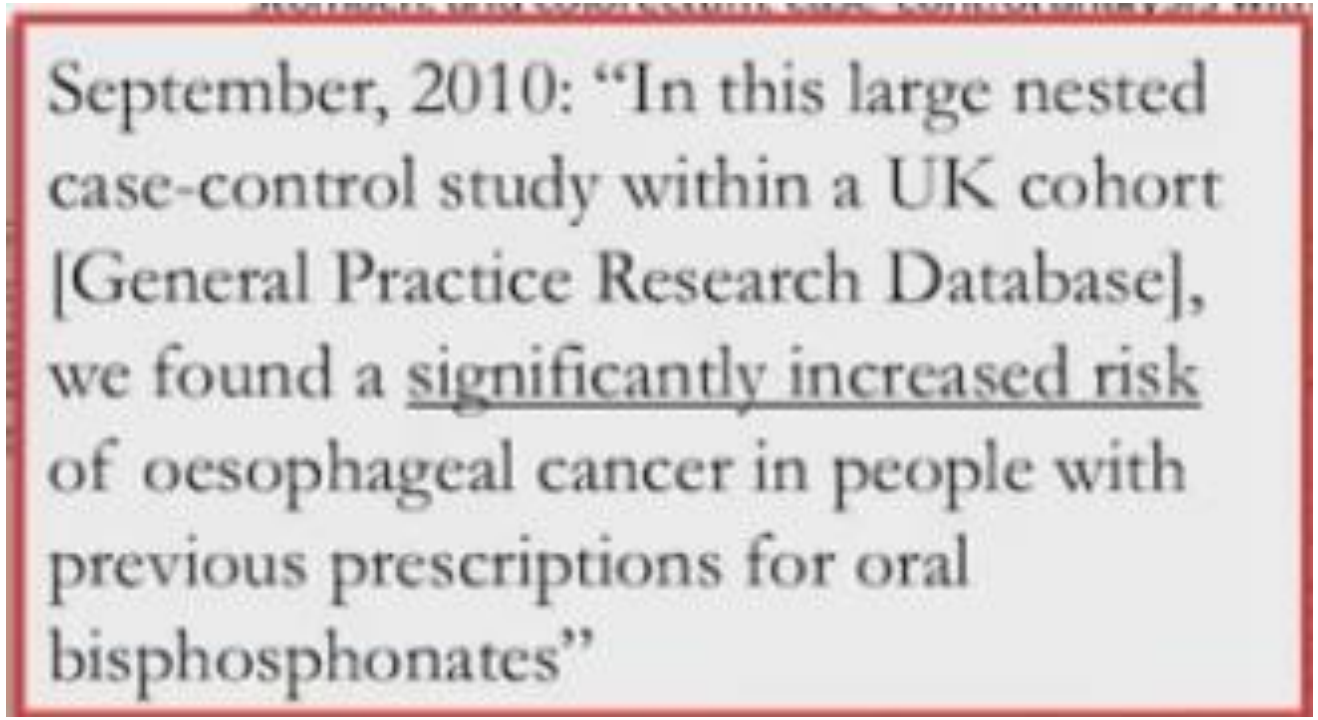
Observational Health Studies

- In this large, nested case control study,
  so there they're describing the design
  of the study that these authors conducted in the database,
- it's a very standard type of study, in this
  large, nested case control study within
  a UK cohort, United Kingdom-based data,

September, 2010: "In this large nested case-control study within a UK cohort [General Practice Research Database], we found a significantly increased risk of oesophageal cancer in people with previous prescriptions for oral bisphosphonates"

Observational Health Studies

- it's a database called the **General Practice Research Database**.
- The authors concluded, **we found a significantly increased risk of esophageal cancer in people with previous prescriptions for oral bisphosphonates**.


September, 2010: "In this large nested case-control study within a UK cohort [General Practice Research Database], we found a significantly increased risk of oesophageal cancer in people with previous prescriptions for oral bisphosphonates"

Observational Health Studies

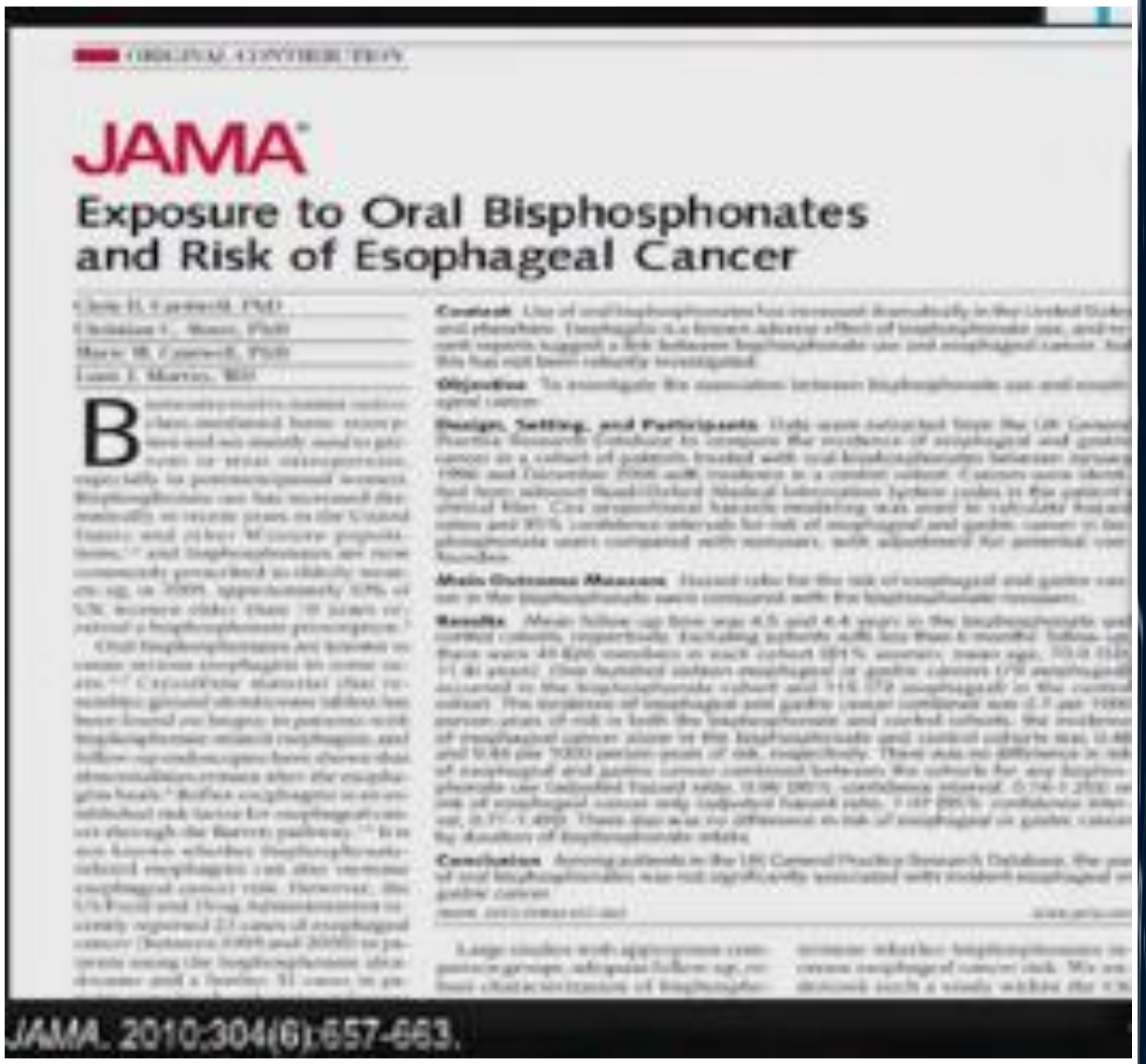
- Now you'll notice that they're not using the word cause in there.
- They're just saying they found a significantly increased risk.
- But in actual fact, what we care about, of course, is **causation**.
- We want to know is it these drugs that caused the esophageal cancer?
- And alas, unhappily, these authors have concluded that yes.
- Apparently these drugs do.
- They found a significant increase to the risk of esophageal cancer.
- I should perhaps put this in context.
- These authors are conducting this study for a reason.

Observational Health Studies

- There is a concern with these drugs that they might cause esophageal cancer.
- This concern has arisen from many different sources.
- But there are a number of cases that have been reported in the literature.
- There are also so-called spontaneous report databases where we have found reports of patients on bisphosphonates subsequently having diagnosis of esophageal cancer.
- So this study, there's a context to this study.
- It's not completely out of the blue.
- In any event, large electronic health record database focusing on a particular causal association, bisphosphonates and esophageal cancer, and concluding that there is a risk. That's September 2010.

Observational Health Studies

- And right about the same time, actually the month before, in JAMA, the Journal of the American Medical Association, a paper appeared with a very similar title, "Exposure to Oral Bisphosphonates and the Risk of Esophageal Cancer".
- So these authors basically studying the same issue

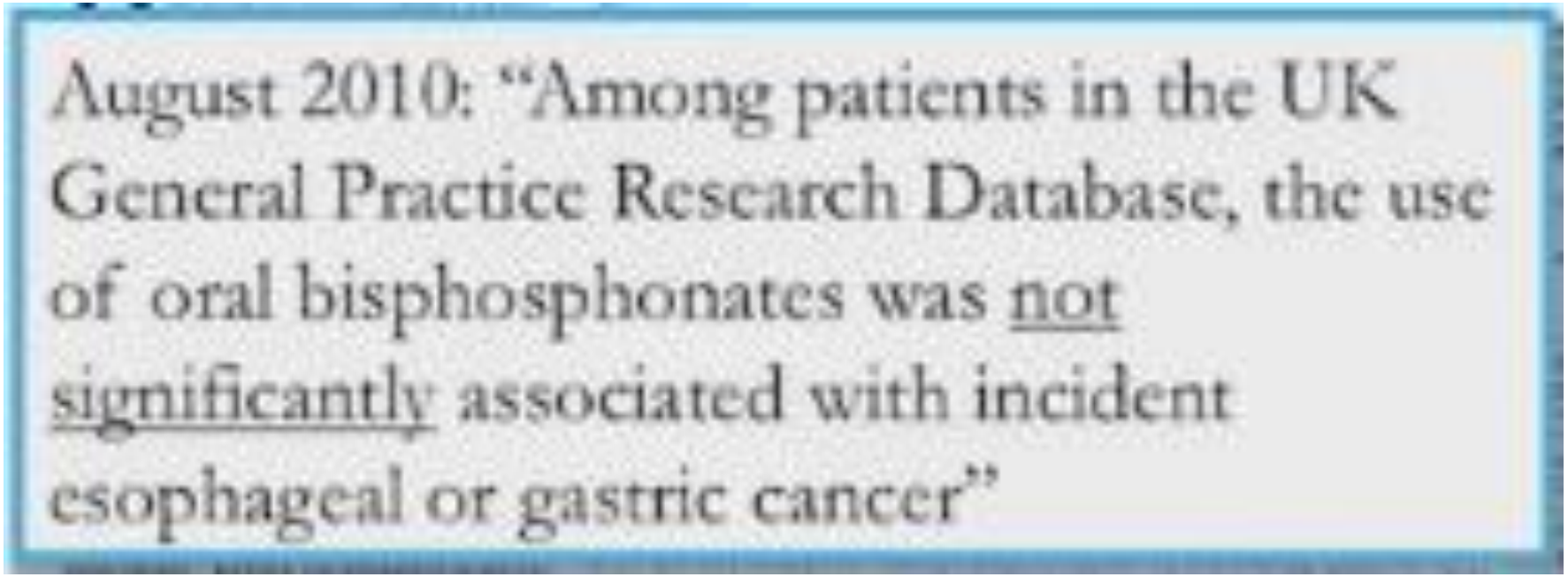

JAMA

Exposure to Oral Bisphosphonates and Risk of Esophageal Cancer

JAMA. 2010;304(6):657-663.

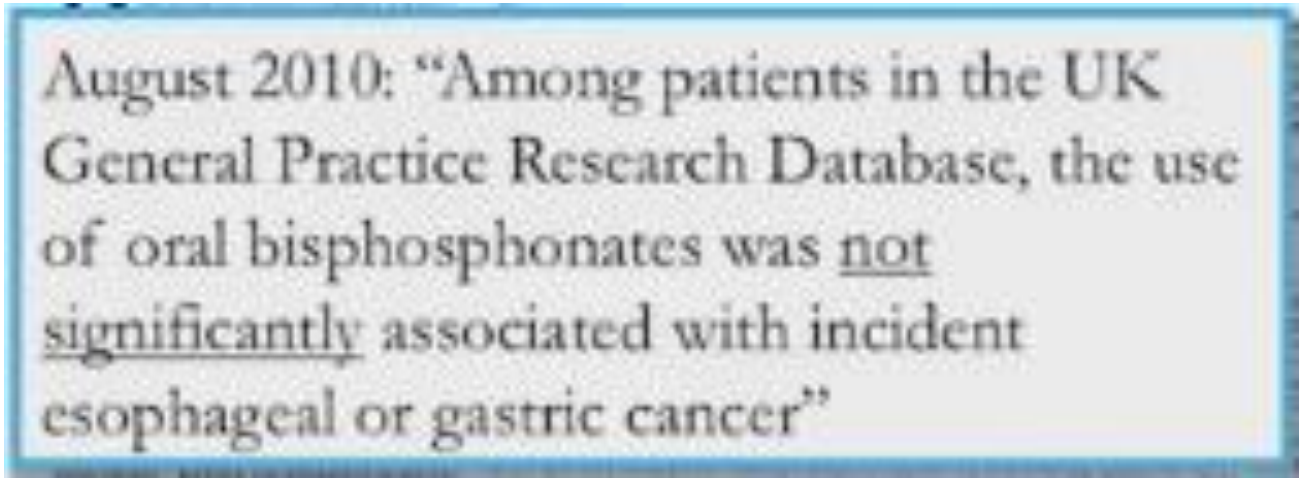
Observational Health Studies

- And here's the conclusion that they drew.
- Among patients in the UK General Practice Research Database, the use of oral bisphosponates was not significantly associated with incident esophageal or gastric cancer.



Observational Health Studies

- So the opposite conclusion in other words.
- One paper concludes that **there is an association**. There's a significantly increased risk. Watch out. We better be careful.
- The other paper concluding that there **is not a significant association** between them. Relax. We don't have to worry about it.



Observational Health Studies

- The kicker here is, as you can see from the papers, both sets of authors used the same database.
- So two well-regarded sets of researchers and analysts analyzing the same data, focusing on the same issue, and arriving at opposite conclusions.
- So this is obviously quite a concern.
- And one wonders if there's some fundamental flaw in the way that we're conducting these studies that could lead to this kind of a mess, basically, where we've got conflicting conclusions, in this case, using the same data.

What is the quality of the current evidence from observational analyses?

August 2010: "Among patients in the UK General Practice Research Database, the use of oral bisphosphonates was not significantly associated with incident esophageal or gastric cancer"

September, 2010: "In this large nested case-control study within a UK cohort [General Practice Research Database], we found a significantly increased risk of oesophageal cancer in people with previous prescriptions for oral bisphosphonates"

2. *JAMA*. 2010;304(6):657-663.

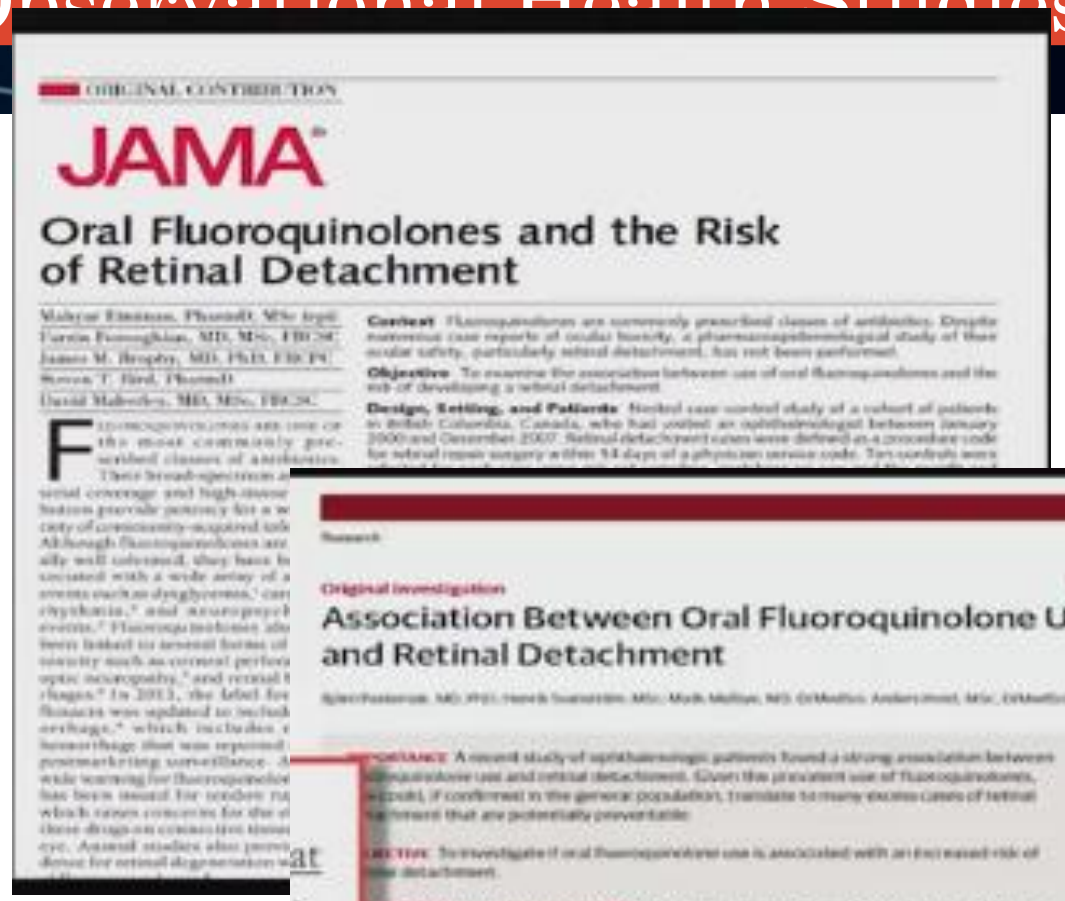
1. *BMJ* 2010;341:c4444.

Observational Health Studies

- So one might think that I chose this example because it's unique in the history of science.
- And I'm just using it to make a point.
- And in actual fact, this is a very common phenomenon.
- So I just show a set.
- Very quickly, I've run through a set of other examples of the same phenomenon.
- And these are three or four examples out of dozens or perhaps hundreds that I could have chosen to make the point here.

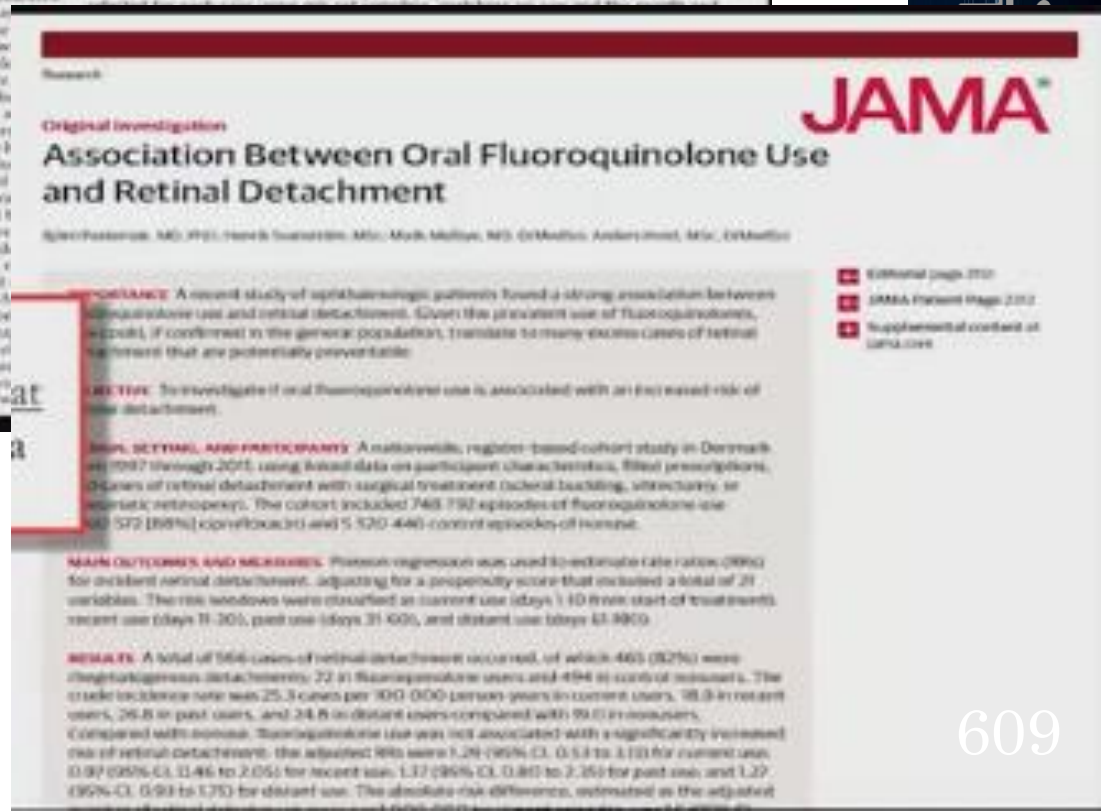
Observational Health Studies

- Here's a paper in JAMA, very prestigious journal, oral fluoroquinolones, this is a class of antibiotics, and the risk of retinal detachment.
- Same basic kind of issue.
- We want to know do fluoroquinolones cause retinal detachment.
- And these authors used a large scale observational health care database to examine this issue.
- Paper published in August 2012.


ORIGINAL CONTRIBUTION

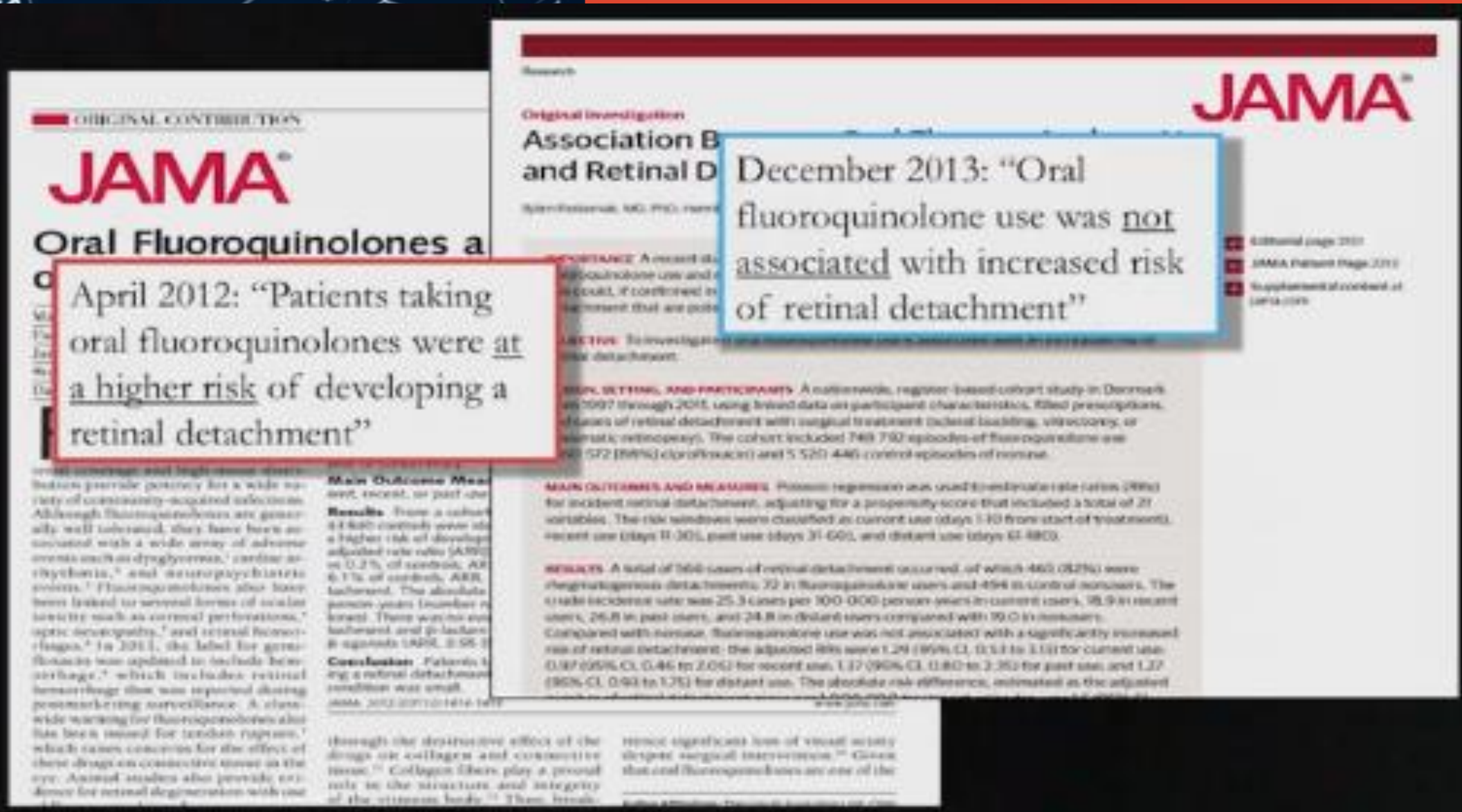
JAMA

Oral Fluoroquinolones and the Risk of Retinal Detachment


JAMA

Original Investigation

Association Between Oral Fluoroquinolone Use and Retinal Detachment



3. *JAMA*. 2012;307(13):1414-1419.

4. *JAMA*. 2013;310(20):2184-2190.

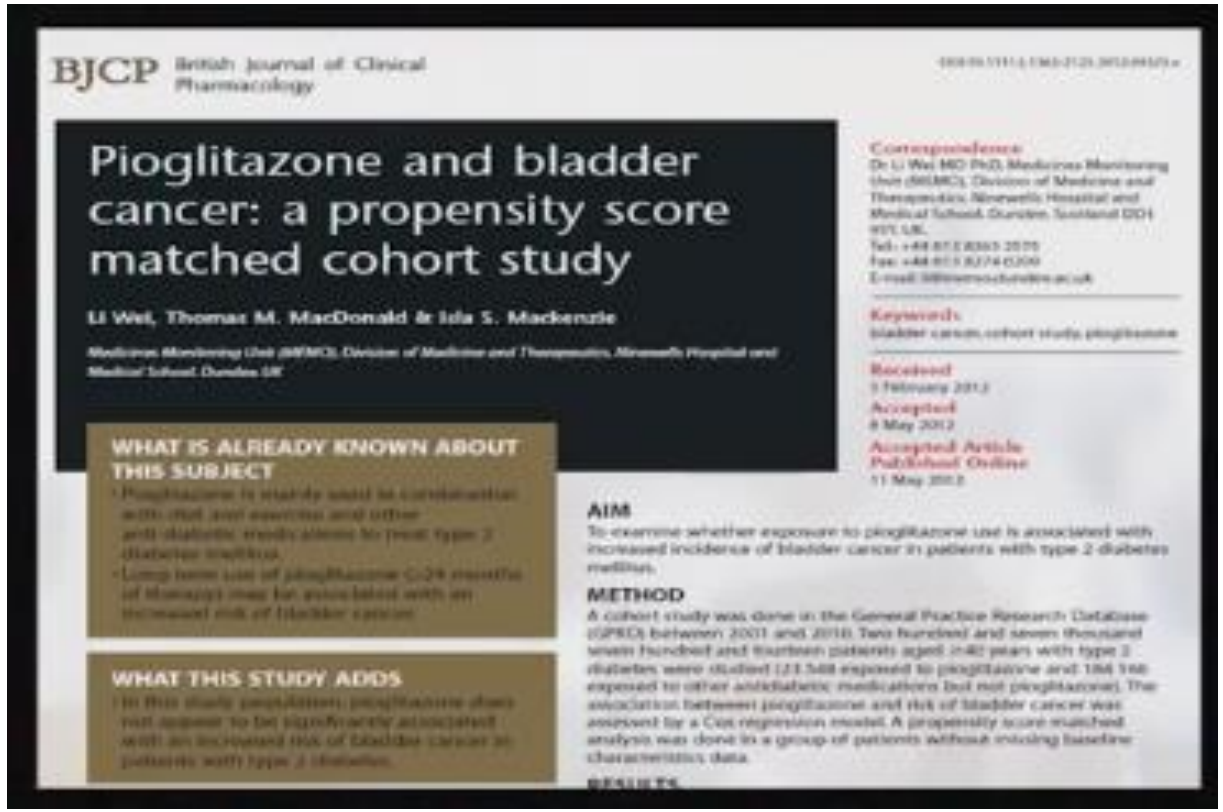
Observational Health Studies

- And they concluded patients taking oral fluoroquinolones were **at higher risk** of developing retinal attachment.
- So these authors are saying there is an association.
- Same journal, sometime later, same question, different conclusion.
- Oral fluoroquinolone use **was not associated** with increased risk of retinal detachment.
- I will go on.

Observational Health Studies

- Pioglitazone and bladder cancer.
- This is an important public health issue.

May 2012: "In this study population, pioglitazone does not appear to be significantly associated with an increased risk of bladder cancer in patients with type 2 diabetes."



BJCP British Journal of Clinical Pharmacology

Pioglitazone and bladder cancer: a propensity score matched cohort study

Li Wei, Thomas M. MacDonald & Isla S. Mackenzie

WHAT IS ALREADY KNOWN ABOUT THIS SUBJECT

WHAT THIS STUDY ADDS

AIM

METHOD

RESULTS

Observational Health Studies

- Pioglitazone is a widely used as diabetes medication.
- And some evidence has emerged that these drugs might cause, that this drug might cause bladder cancer.
- This paper appeared in the British Journal of Clinical Pharmacology.

May 2012: "The use of pioglitazone is associated with an increased risk of incident bladder cancer among people with type 2 diabetes."

BJCP British Journal of Clinical Pharmacology

Pioglitazone and bladde

May 2012: "In this study population, pioglitazone does not appear to be significantly associated with an increased risk of bladder cancer in patients with type 2 diabetes."

WHAT THIS STUDY ADDS

5. BJCP 75: 254–259.

BMJ

EARCH

The use of pioglitazone and the risk of bladder cancer in people with type 2 diabetes: nested case-control study

OPEN ACCESS

Laurent Azoulay *assistant professor*[1,2], Hui Yin *statistician*[1], Kristian B Filion *assistant professor*[1,2], Jonathan Assayag *graduate student*[1], Agnieszka Majdan *endocrinologist*[3], Michael N Pollak *oncologist and professor*[2], Samy Suissa *professor*[1]

METHOD

6. BMJ 2012;344:e3645.

Observational Health Studies

- Cardwell CR, Abnet CC, Cantwell MM, Murray LJ. Exposure to Oral Bisphosphonates and Risk of Esophageal Cancer. *JAMA*. 2010;304(6):657-663. http://jamanetwork.com/journals/jama/fullarticle/186382 .
- Green Jane, Czanner Gabriela, Reeves Gillian,Watson Joanna, Wise Lesley, Beral Valerie et al. Oral bisphosphonates and risk of cancer of oesophagus, stomach, and colorectum: case-control analysis within a UK primary care cohort *BMJ* 2010; 341 :c4444 doi: http://dx.doi.org/10.1136/bmj.c4444
- Etminan M, Forooghian F, Brophy JM, Bird ST, Maberley D. Oral Fluoroquinolones and the Risk of Retinal Detachment. *JAMA*.2012;307(13):1414-1419. http://jamanetwork.com/journals/jama/fullarticle/1148331

Observational Health Studies

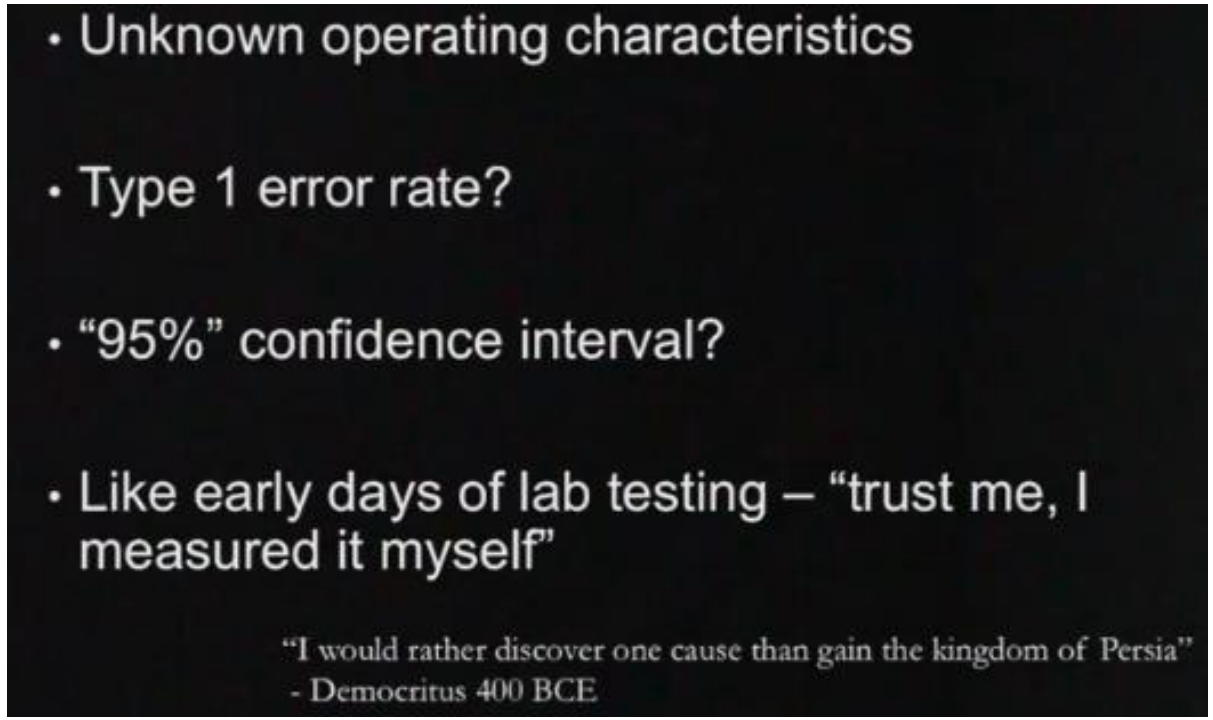
- Pasternak B, Svanström H, Melbye M, Hviid A. Association Between Oral Fluoroquinolone Use and Retinal Detachment. *JAMA*. 2013;310(20):2184-2190. http://jamanetwork.com/journals/jama/fullarticle/1785461
- Wei, L., MacDonald, T. M. and Mackenzie, I. S. (2013), Pioglitazone and bladder cancer: a propensity score matched cohort study. British Journal of Clinical Pharmacology, 75: 254–259. http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2125.2012.04325.x/abstract
- Azoulay Laurent, Yin Hui, Filion Kristian B, AssayagJonathan, Majdan Agnieszka, Pollak Michael N et al.The use of pioglitazone and the risk of bladder cancer in people with type 2 diabetes: nested case-control study BMJ 2012; 344 :e3645 doi: http://dx.doi.org/10.1136/bmj.e3645

Observational Health Studies

- I could go on and on.
- I'm just showing you a few examples.
- The point here is: there is a problem here with the way we're conducting these studies.
- We conduct the studies. And we proceed as if we're getting valid and correct causal inferences from these studies.
- But in fact, they're problematic in that the results are, in general, all over the place.
- This is very common to see conflicting results, particularly in different databases.

Observational Health Studies

- But also, as I've just shown, actually different analysts analyzing the same data and arriving at different conclusions.
- I would claim that, in fact, the operating characteristics of these kinds of studies are in fact unknown.
- And let me just drill down on that a little bit.



Observational Health Studies

- So in each of those papers that I just went through quickly, the centerpiece of the results section of the **paper is an estimate of the effect of the medication on the outcome in question**, bisphosphonates and esophageal cancer (سرطان مری), let's say.
- Typically, many of these papers focus in on something called a relative risk.
- A relative risk is basically, what's the relative risk of esophageal cancer on bisphosphonates as against not on bisphosphonates, for example.
- So the papers produce a relative risk estimate.
- It might be 1.4., It might be 2., It might be 0.8., 1 would be neutral.

Observational Health Studies

- 1 would say the risk is the same on the drugs as off the drugs.
- But crucially, these papers also produce something called a 95% confidence interval.
- They produce an interval around that estimate that captures our uncertainty about the true effect estimate.
- **So if the 95% percent confidence interval does not include 1, conventionally we say that the result is statistically significant.**
- So if you have a relative risk of, say, 2. That's the point estimate.
- And the interval goes from 1.5 to 2.8, that interval does not include 1.
- **We would say it's statistically significant.**

Observational Health Studies

- In those papers I just showed you, that's when they said we found a significantly increased risk, that's what's going on.
- They have a 95% interval that does not include 1.
- Well, that interval, the idea is that it is constructed in a way that 95% of the time when you calculate an interval like this, **it should contain the truth**, the actual true relative risk associated with bisphosphonates and esophageal cancer, which we don't know what that is, but it exists.
- So the idea is that 95% of the time, the interval includes the truth.
- And 5% of the time it doesn't.

Observational Health Studies

- And that gives a sense, 95% being a fairly high probability, we're reasonably confident in our findings if the interval does not contain 1.
- Unfortunately, however though, we have absolutely no idea how often in practice when we do these studies, **we have no idea how often these intervals actually contain the truth**, contain the true underlying relative risk.
- We don't know if it's 100%.
- We don't know if it's 95%, which would be great.
- And we don't know if it's 50%.
- And there's quite a difference between those.

Observational Health Studies

- So if in fact, those intervals only contain the truth 50% of the time, that's problematic.
- Then we're basically saying we have not learned very much from the study.
- We don't have very high quality evidence.
- Similarly with the type 1 error rate, this is a technical term, it's to do with basically how often do you conclude **there is a significant finding, when in fact there's nothing going on, sort of a false positive, if you will.**
- Again, typically we conduct these kinds of studies to have a-- we typically target a 5% type 1 error rate.

Observational Health Studies

- We were prepared to tolerate, in many of these settings, a 5% false positive rate.
- Could be 1%. It could be 10%. But it's often people fixate on 5%.
- And we would like it to be the case that when we conduct studies like this, that the **false positive rate is indeed controlled at 5%.**
- In actual fact, we have no idea what the type 1 error rate is.
- Maybe it's 5%. Maybe it's 50%. Maybe it's 1%.
- We simply don't know.

Observational Health Studies

- So this situation is rather scary, not least because caregivers, regulators, pharma companies, and so on around the world are making health care decisions on the basis of these studies,
- the oral bisphosphonates study that showed an increased risk of esophageal cancer, that found its way onto the front page of major newspapers.
- So I have no doubt that it impacted health care practice and that there are patients who stopped taking those drugs.
- And that's a little scary, given that we actually have no idea what the reliability is of that study and its finding.

Observational Health Studies

- There's a colleague of mine at the FDA who likens (compares) this situation to the early days of laboratory testing.
- So 30 or 40 years ago, if you wanted to measure a white blood cell count or a cholesterol level, you took a blood sample.
- And you handed it to your local lab technician.
- And they went to their lab. They did something.
- They came back sometime later. And they gave you a number. 26.
- You have no idea what the operating characteristics are.
- You have no idea if that was likely to be within 10% of the correct answer, 100% percent, 50% of the correct answer, no idea.

Observational Health Studies

- These days, flash forward, laboratories around the country and around the world are highly regulated.
- They routinely test positive controls and negative controls.
- And we have a very good understanding of the operating characteristics of common laboratory test measurements.
- We've a very good understanding of the margin of error associated with some of those laboratory measurements.
- That's where we've got to get to with observational studies.
- The current situation is very unsatisfactory where we just have no idea what the operating characteristics are.

Observational Health Studies

- We need to conduct research to establish what the operating characteristics are.
- And of course, what we'd really like to do is to conduct research to develop methods and algorithms and so on that have good operating characteristics.
- I just highlight the quote there from Democritus 400 BCE.
- This is all about discovering causes.
- And Democritus said, **"I would rather discover one cause than gain the kingdom of Persia".**
- And that's a very perceptive comment.
- Discovering causes from data is very hard, we have discovered.

Clinical trials

- And, as the estrogen example illustrates, even in the right hands this powerful statistical tool can send us speeding dangerously in the wrong direction.
- The balance of this chapter will explain the most common regression "mistakes."
- I put "mistakes" in quotation marks, because, as with all other kinds of statistical analysis, clever people can knowingly exploit these methodological points to nefarious ends.
- Here is a "Top Seven" list of the most common abuses of an otherwise extraordinary tool.

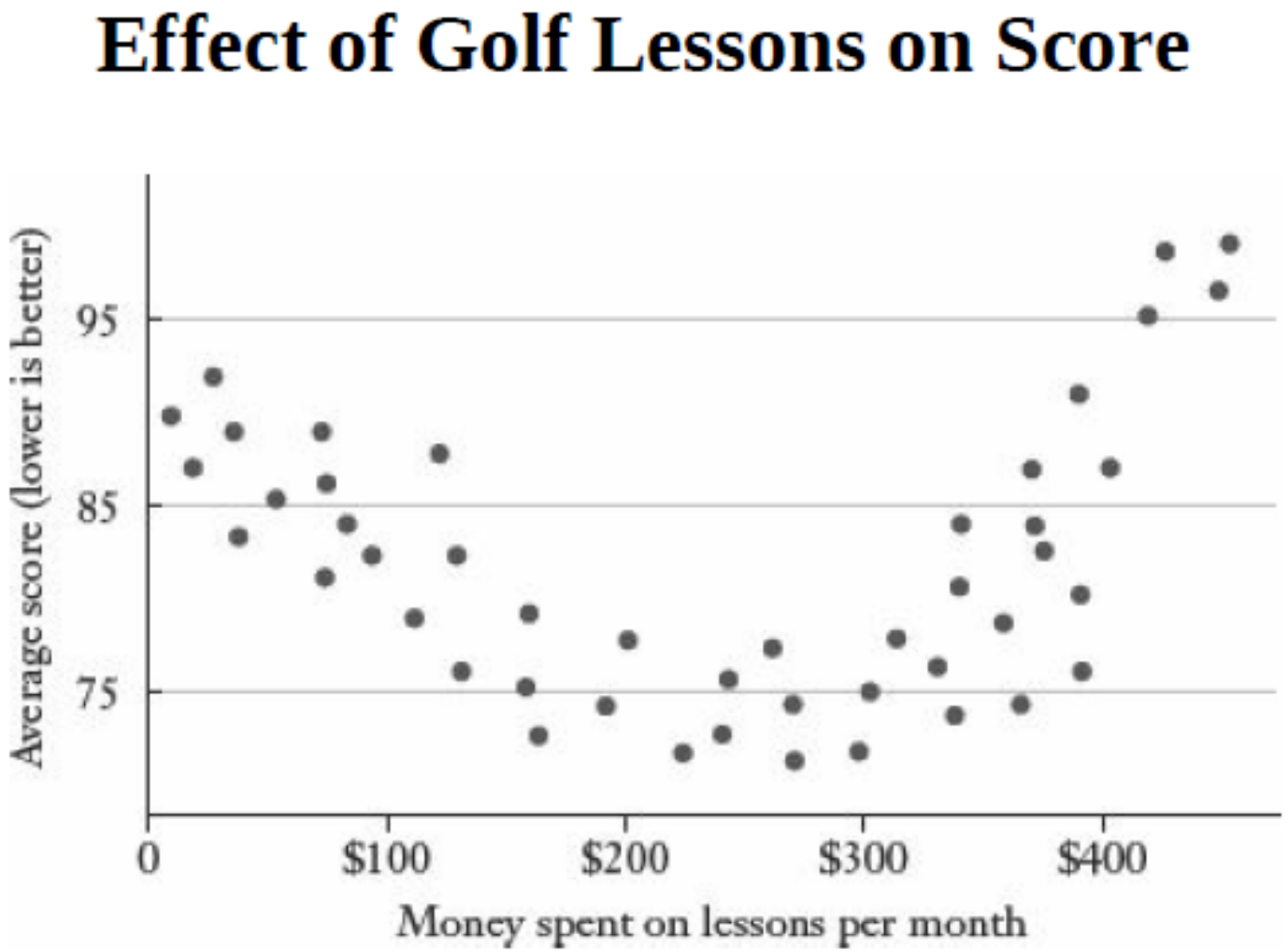
Using regression to analyze a nonlinear relationship

- Have you ever read the warning label on a hair dryer—the part that cautions, Do Not Use in the Bath Tub?
- And you think to yourself, “What kind of moron uses a hair dryer in the bath tub?”
- It’s an electrical appliance; you don’t use electrical appliances around water.
- They’re not designed for that. If regression analysis had a similar warning label, it would say, **Do Not Use When There Is Not a Linear Association between the Variables That You Are Analyzing**.

Using regression to analyze a nonlinear relationship

- Remember, a regression coefficient describes the slope of the "line of best fit" for the data; a line that is not straight will have a different slope in different places.
- As an example, consider the following hypothetical relationship between the number of golf lessons that I take during a month (an explanatory variable) and my average score for an eighteen-hole round during that month (the dependent variable).
- As you can see from the scatter plot, there is no consistent linear relationship.

Using regression to analyze a nonlinear relationship



Using regression to analyze a nonlinear relationship

- There is a pattern, but it cannot be easily described with a single straight line.
- The first few golf lessons appear to bring my score down rapidly.
- There is a negative association between lessons and my scores for this stretch; the slope is negative.
- More lessons yield lower scores (which is good in golf).
- But then when I reach the point where I'm spending $200-$300 a month on lessons, the lessons do not seem to have much effect at all.
- There is no clear association over this stretch between additional instruction and my golf scores; the slope is zero.

Using regression to analyze a nonlinear relationship

- And finally, the lessons appear to become counterproductive.
- Once I'm spending $300 a month on instruction, incremental lessons are associated with higher scores; the slope is positive over this stretch.
- The most important point here is that we cannot accurately summarize the relationship between lessons and scores with a single coefficient.
- The best interpretation of the pattern described above is that golf lessons have several different linear relationships with my scores.
- You can see that; a statistics package will not.
- If you feed these data into a regression equation, the computer will give you a single coefficient.

Using regression to analyze a nonlinear relationship

- That coefficient will not accurately reflect the true relationship between the variables of interest.
- The results you get will be the statistical equivalent of using a hair dryer in the bath tub.
- Regression analysis is meant to be used when the relationship between variables is linear.
- A textbook or an advanced course in statistics will walk you through the other core assumptions underlying regression analysis.
- As with any other tool, the further one deviates from its intended use, the less effective, or even potentially dangerous, it's going to be.

RESEARCH
Khatam University
FANAP

Correlation does not equal causation

- Regression analysis can only demonstrate an association between two variables.
- As I have mentioned before, we cannot prove with statistics alone that a change in one variable is causing a change in the other.
- In fact, a sloppy regression equation can produce a large and statistically significant association between two variables that have nothing to do with one another.
- Suppose we were searching for potential causes for the rising rate of autism in the United States over the last two decades.

RESEARCH

Correlation does not equal causation

- Our dependent variable—the outcome we are seeking to explain—would be some measure of the incidence of the autism by year, such as the number of diagnosed cases for every 1,000 children of a certain age.
- If we were to include annual per capita income in China as an explanatory variable, we would almost certainly find **a positive and statistically significant association between rising incomes in China and rising autism rates in the U.S. over the past twenty years**.
- Why? Because they both have been rising sharply over the same period.

Correlation does not equal causation

- Yet I highly doubt that a sharp recession in China would reduce the autism rate in the United States.
- (To be fair, if I observed a strong relationship between rapid economic growth in China and autism rates in China alone, I might begin to search for some environmental factor related to economic growth, such as industrial pollution, that might explain the association.)
- The kind of false association between two variables that I have just illustrated is just one example of a more general phenomenon known as spurious (false) causation.
- There are several other ways in which an association between A and B can be wrongly interpreted.

Reverse causality

- A statistical association between A and B does not prove that A causes B.
- In fact, it's entirely plausible that B is causing A.
- I mentioned to this possibility earlier in the golf lesson example.
- Suppose that when I build a complex model to explain my golf scores, the variable for golf lessons is consistently associated with worse scores.
- The more lessons I take, the worse I shoot!
- One explanation is that I have a really, really bad golf instructor.

Reverse causality

- A more plausible explanation is that I tend to take more lessons when I'm playing poorly; bad golf is causing more lessons, not the other way around.
- (There are some simple methodological fixes to a problem of this nature. For example, I might include golf lessons in one month as an explanatory variable for golf scores in the next month.)
- As noted earlier in the chapter, causality may go in both directions.

Reverse causality

- Suppose you do some research demonstrating that states that spend more money on K–12 education have higher rates of economic growth than states that spend less on K–12 education.
- A positive and significant association between these two variables does not provide any insight into which direction the relationship happens to run.
- Investments in K–12 education could cause economic growth.
- On the other hand, states that have strong economies can afford to spend more on K–12 education, so the strong economy could be causing the education spending.

Reverse causality

- Or, education spending could boost economic growth, which makes possible additional education spending—the causality could be going in both ways.
- The point is that we should not use explanatory variables that might be affected by the outcome that we are trying to explain, or else the results will become hopelessly tangled.
- For example, it would be inappropriate to use the unemployment rate in a regression equation explaining GDP growth, since unemployment is clearly affected by the rate of GDP growth.

Reverse causality

- Or, to think of it another way, a regression analysis finding that lowering unemployment will boost GDP growth is a silly and meaningless finding, since boosting GDP growth is usually required in order to reduce unemployment.
- We should have reason to believe that our explanatory variables affect the dependent variable, and not the other way around.

Omitted variable bias.

- You should be skeptical the next time you see a huge headline proclaiming, “Golfers More Prone to Heart Disease, Cancer, and Arthritis!”
- I would not be surprised if golfers have a higher incidence of all of those diseases than nongolfers; I also suspect that golf is probably good for your health because it provides socialization and modest exercise.
- How can I reconcile those two statements? Very easily.
- Any study that attempts to measure the effects of playing golf on health must control properly for age.

Omitted variable bias.

- In general, people play more golf when they get older, particularly in retirement.
- Any analysis that leaves out age as an explanatory variable is going to miss the fact that golfers, on average, will be older than nongolfers.
- Golf isn't killing people; old age is killing people, and they happen to enjoy playing golf while it does.
- I suspect that when age is inserted into the regression analysis as a control variable, we will get a different outcome.
- Among people who are the same age, golf may be mildly preventive of serious illnesses.
- That's a pretty big difference.

RESEARCH
Khatam University
FANAP

Omitted variable bias.

- In this example, age is an important “omitted variable.”
- When we leave age out of a regression equation explaining heart disease or some other adverse health outcome, the “playing golf” variable takes on two explanatory roles rather than just one.
- It tells us the effect of playing golf on heart disease, and it tells us the effect of being old on heart disease (since golfers tend to be older than the rest of the population).
- In the statistics lingo, we would say that the golf variable is “picking up” the effect of age.
- The problem is that these two effects are commingled (mixed).

Omitted variable bias.

- At best, our results are a jumbled mess.
- At worst, we wrongly assume that golf is bad for your health, when in fact the opposite is likely to be true.
- Regression results will be misleading and inaccurate if the regression equation leaves out an important explanatory variable, particularly if other variables in the equation "pick up" that effect.

Omitted variable bias.

- Suppose we are trying to explain school quality.
- This is an important outcome to understand: What makes good schools?
- Our dependent variable—the quantifiable measure of quality—would most likely be test scores.
- We would almost certainly examine school spending as one explanatory variable in hopes of quantifying the relationship between spending and test scores.
- There are many potentially significant omitted variables here, but the crucial one is parental education.

Omitted variable bias.

- Well-educated families tend to live in affluent areas that spend a lot of money on their schools;
- such families also tend to have children who score well on tests (and poor families are more likely to have students who struggle).
- If we do not have some measure of the socioeconomic status of the student body as a control variable, our regression results will probably show a large positive association between school spending and test scores—
- when in fact, those results may be a function of the kind of students who are walking in the school door, not the money that is being spent in the building.

Omitted variable bias.

- I remember a college professor's pointing out that SAT scores are highly correlated with the number of cars that a family owns.
- He implied that the SAT was therefore an unfair and inappropriate tool for college admissions.
- The SAT has its flaws but the correlation between scores and family cars is not the one that concerns me most.
- I do not worry much that rich families can get their kids into college by purchasing three extra automobiles.
- The number of cars in a family's garage is a proxy for their income, education, and other measures of socioeconomic status.

Omitted variable bias.

- The fact that wealthy kids do better on the SAT than poor kids is not news. (As noted earlier, the mean SAT critical reading score for students from families with a household income over $200,000 is 134 points higher than the mean score for students in households with income below $20,000.)4
- The bigger concern should be whether or not the SAT is “coach-able”
- How much can students improve their scores by taking private SAT prep classes?
- Wealthy families clearly are better able to send their children to test prep classes.

Omitted variable bias.

- Any causal improvement between these classes and SAT scores would favor students from wealthy families relative to more disadvantaged students of equal abilities (who presumably also could have raised their scores with a prep class but never had that opportunity).

Highly correlated explanatory variables (multicollinearity).

- If a regression equation includes two or more explanatory variables that are highly correlated with each other, the analysis will not necessarily be able to discern the true relationship between each of those variables and the outcome that we are trying to explain.
- An example will make this clearer.
- Assume we are trying to gauge the effect of illegal drug use on SAT scores.
- Specifically, we have data on whether the participants in our study have ever used cocaine and also on whether they have ever used heroin.
- (We would presumably have many other control variables as well.)

Highly correlated explanatory variables (multicollinearity).

- What is the impact of cocaine use on SAT scores, holding other factors constant, including heroin use?
- And what is the impact of heroin use on SAT scores, controlling for cocaine use and other factors?
- The coefficients on heroin and cocaine use might not actually be able to tell us that.
- The methodological challenge is that people who have used heroin have likely also used cocaine.
- If we put both variables in the equation, we will have very few individuals who have used one drug but not the other, which leaves us very little variation in the data with which to calculate their independent effects.

RESEARCH

Highly correlated explanatory variables (multicollinearity).

- Think back for a moment to the mental imagery used to explain regression analysis in the last chapter.
- We divide our data sample into different “rooms” in which each observation is identical except for one variable, which then allows us to isolate the effect of that variable while controlling for other potential confounding factors.
- We may have 692 individuals in our sample who have used both cocaine and heroin.
- However, we may have only 3 individuals who have used cocaine but not heroin and 2 individuals who have used heroin and not cocaine.

Highly correlated explanatory variables (multicollinearity).

- Any inference about the independent effect of just one drug or the other is going to be based on these tiny samples.
- We are unlikely to get meaningful coefficients on either the cocaine or the heroin variable;
- we may also obscure the larger and more important relationship between SAT scores and using either one of these drugs.
- When two explanatory variables are highly correlated, researchers will usually use one or the other in the regression equation, or they may create some kind of composite variable, such as “used cocaine or heroin.”

Highly correlated explanatory variables (multicollinearity).

- For example, when researchers want to control for a student's overall socioeconomic background, they may include variables for both "mother's education" and "father's education," since this inclusion provides important insight into the educational background of the household.
- However, if the goal of the regression analysis is to isolate the effect of either a mother's or a father's education, then putting both variables into the equation is more likely to confuse the issue than to clarify it.
- The correlation between a husband's and a wife's educational attainments is so high that we cannot depend on regression analysis to give us coefficients that meaningfully isolate the effect of either parent's education (just as it is hard to separate the impact of cocaine use from the impact of heroin use).

Extrapolating beyond the data

- Regression analysis, like all forms of statistical inference, is designed to offer us insights into the world around us.
- We seek patterns that will hold true for the larger population.
- However, our results are valid only for a population that is similar to the sample on which the analysis has been done.
- In the last chapter, I created a regression equation to predict weight based on a number of independent variables.
- The R2 of my final model was 0.29, which means that it did a decent job of explaining the variation in weight for a large sample of individuals—all of whom happened to be adults.

RESEARCH
Khatam University
FANAP

Extrapolating beyond the data

- So what happens if we use our regression equation to predict the likely weight of a newborn?
- Let's try it. My daughter was 21 inches when she was born.
- We'll say that her age at birth was zero; she had no education and did not exercise.
- She was white and female.
- The regression equation based on the Changing Lives data predicts that her weight at birth should have been **negative 19.6 pounds**. (She weighed 8½ pounds.)

Extrapolating beyond the data

- The authors of one of the Whitehall studies referred to in the last chapter were strikingly explicit in drawing their narrow conclusion:
- "*Low control in the work environment is associated with an increased risk of future coronary heart disease among men and women employed in government offices*".

Data mining (too many variables).

- If omitting important variables is a potential problem, then presumably adding as many explanatory variables as possible to a regression equation must be the solution. Nope.
- Your results can be compromised if you include too many variables, particularly extraneous explanatory variables with no theoretical justification.
- For example, one should not design a research strategy built around the following premise:
- Since we don't know what causes autism, we should put as many potential explanatory variables as possible in the regression equation just to see what might turn up as statistically significant; then maybe we'll get some answers.

Data mining (too many variables).

- If you put enough junk variables in a regression equation, one of them is bound to meet the threshold for statistical significance just by chance.
- The further danger is that junk variables are not always easily recognized as such.
- Clever researchers can always build a theory after the fact for why some curious variable that is really just nonsense turns up as statistically significant.

Data mining (too many variables).

- To make this point, I often do the same coin flipping exercise that I explained during the probability discussion.
- In a class of forty students or so, I'll have each student flip a coin.
- Any student who flips tails is eliminated; the rest flip again.
- In the second round, those who flip tails are once again eliminated.
- I continue the rounds of flipping until one student has flipped five or six heads in a row.
- You may recall some of the silly follow-up questions: "What's your secret? Is it in the wrist? Can you teach us to flip heads all the time? Maybe it's that Harvard sweatshirt you're wearing."

RESEARCH

Data mining (too many variables).

- Obviously the string of heads is just luck; the students have all watched it happen.
- However, that is not necessarily how the result could or would be interpreted in a scientific context.
- The probability of flipping five heads in a row is 1/32, or .03.
- This is comfortably below the .05 threshold we typically use to reject a null hypothesis.

Data mining (too many variables).

- Our null hypothesis in this case is that **the student has no special talent for flipping heads; the lucky string of heads.**
- (which is bound to happen for at least one student when I start with a large group) allows us to reject the null hypothesis and adopt the alternative hypothesis: **This student has a special ability to flip heads.**
- After he has achieved this impressive feat, we can study him for clues about his flipping success—his flipping form, his athletic training, his extraordinary concentration while the coin is in the air, and so on.
- And it is all nonsense.
- This phenomenon can plague even legitimate research.

Data mining (too many variables).

- The accepted convention is to reject a null hypothesis when we observe something that would happen by chance only 1 in 20 times or less if the null hypothesis were true.
- Of course, if we conduct 20 studies, or if we include 20 junk variables in a single regression equation, then on average we will get 1 bogus statistically significant finding.
- The NYT Magazine captured this tension wonderfully in a quotation from Richard Peto, a medical statistician and epidemiologist: "**Epidemiology is so beautiful and provides such an important perspective on human life and death, but an incredible amount of rubbish is published.**

Data mining (too many variables).

- Even the results of clinical trials, which are usually randomized experiments and therefore the gold standard of medical research, should be viewed with some skepticism.
- In 2011, the Wall Street Journal ran a frontpage story on what it described as one of the "dirty little secrets" of medical research:
- **"Most results, including those that appear in top-flight peer-reviewed journals, can't be reproduced."**
- One reason for this "dirty little secret" is the positive publication bias described in Chapter 7.

Data mining (too many variables).

- If researchers and medical journals pay attention to positive findings and ignore negative findings, then they may well publish the one study that finds a drug effective and ignore the nineteen in which it has no effect.
  - Some clinical trials may also have small samples (such as for a rare diseases), which magnifies the chances that random variation in the data will get more attention than it deserves.
  - On top of that, researchers may have some conscious or unconscious bias, either because of a strongly held prior belief or because a positive finding would be better for their career. (No one ever gets rich or famous by proving what doesn't cure cancer.)

Data mining (too many variables).

- For all of these reasons, a shocking amount of expert research turns out to be wrong.
- John Ioannidis, a Greek doctor and epidemiologist, examined forty-nine studies published in three prominent medical journals.8
- Each study had been cited in the medical literature at least a thousand times.
- Yet roughly one-third of the research was subsequently refuted (disproved) by later work.
- (For example, some of the studies he examined promoted estrogen replacement therapy.)

Data mining (too many variables).

- Dr. Ioannidis estimates that roughly half of the scientific papers published will eventually turn out to be wrong.
- His research was published in the Journal of the American Medical Association, one of the journals in which the articles he studied had appeared.
- This does create a certain mind-bending irony: If Dr. Ioannidis's research is correct, then there is a good chance that his research is wrong.

Data mining (too many variables).

- Regression analysis is still an awesome statistical tool.
- (Okay, perhaps my description of it as a “miracle elixir” in the last chapter was a little hyperbolic.)
- Regression analysis enables us to find key patterns in large data sets, and those patterns are often the key to important research in medicine and the social sciences.
- Statistics gives us objective standards for evaluating these patterns.
- When used properly, regression analysis is an important part of the scientific method.
- Consider this chapter to be the mandatory warning
- label.

Data mining (too many variables).

- All of the assorted specific warnings on that label can be boiled down to two key lessons.
- First, designing a good regression equation—figuring out what variables should be examined and where the data should come from—is more important than the underlying statistical calculations.
- This process is referred to as estimating the equation, or specifying a good regression equation.
- The best researchers are the ones who can think logically about what variables ought to be included in a regression equation, what might be missing, and how the eventual results can and should be interpreted.

Data mining (too many variables).

- Second, like most other statistical inference, regression analysis builds only a indirect case.
- An association between two variables is like a fingerprint at the scene of the crime.
- It points us in the right direction, but it's rarely enough to convict.
- Any regression analysis needs a theoretical underpinning: Why are the explanatory variables in the equation?
- What phenomena from other disciplines can explain the observed results?

Data mining (too many variables).

- For instance, why do we think that wearing purple shoes would boost performance on the math portion of the SAT or that eating popcorn can help prevent prostate cancer?
- The results need to be replicated, or at least consistent with other findings.
- Even a miracle elixir won't work when not taken as directed.

program evaluation

- Welcome to program evaluation, which is the process by which we seek to measure the causal effect of some intervention—anything from a new cancer drug to a job placement program for high school dropouts.
- Or putting more police officers on the street.
- The intervention that we care about is typically called the “treatment,” though that word is used more expansively in a statistical context than in normal parlance.
- A treatment can be a literal treatment, as in some kind of medical intervention, or it can be something like attending college or receiving job training upon release from prison.

RESEARCH

program evaluation

- The point is that we are seeking to isolate the effect of that single factor; ideally we would like to know how the group receiving that treatment fares compared with some other group whose members are identical in all other respects but for the treatment.
- The important takeaway is that we can answer tricky but socially meaningful questions we just have to be clever about it.
- Here are some of the most common approaches for isolating a treatment effect.

program evaluation

- Randomized, controlled experiments: The most straightforward way to create a treatment and control group is to—wait for it—create a treatment and control group.
- Natural experiment: Not everybody has millions of dollars lying around to create a large, randomized trial.
- Nonequivalent control: Sometimes the best available option for studying a treatment effect is to create nonrandomized treatment and control groups.
- Difference in differences: One of the best ways to observe cause and effect is to do something and then see what happens.

OMOP Experiments

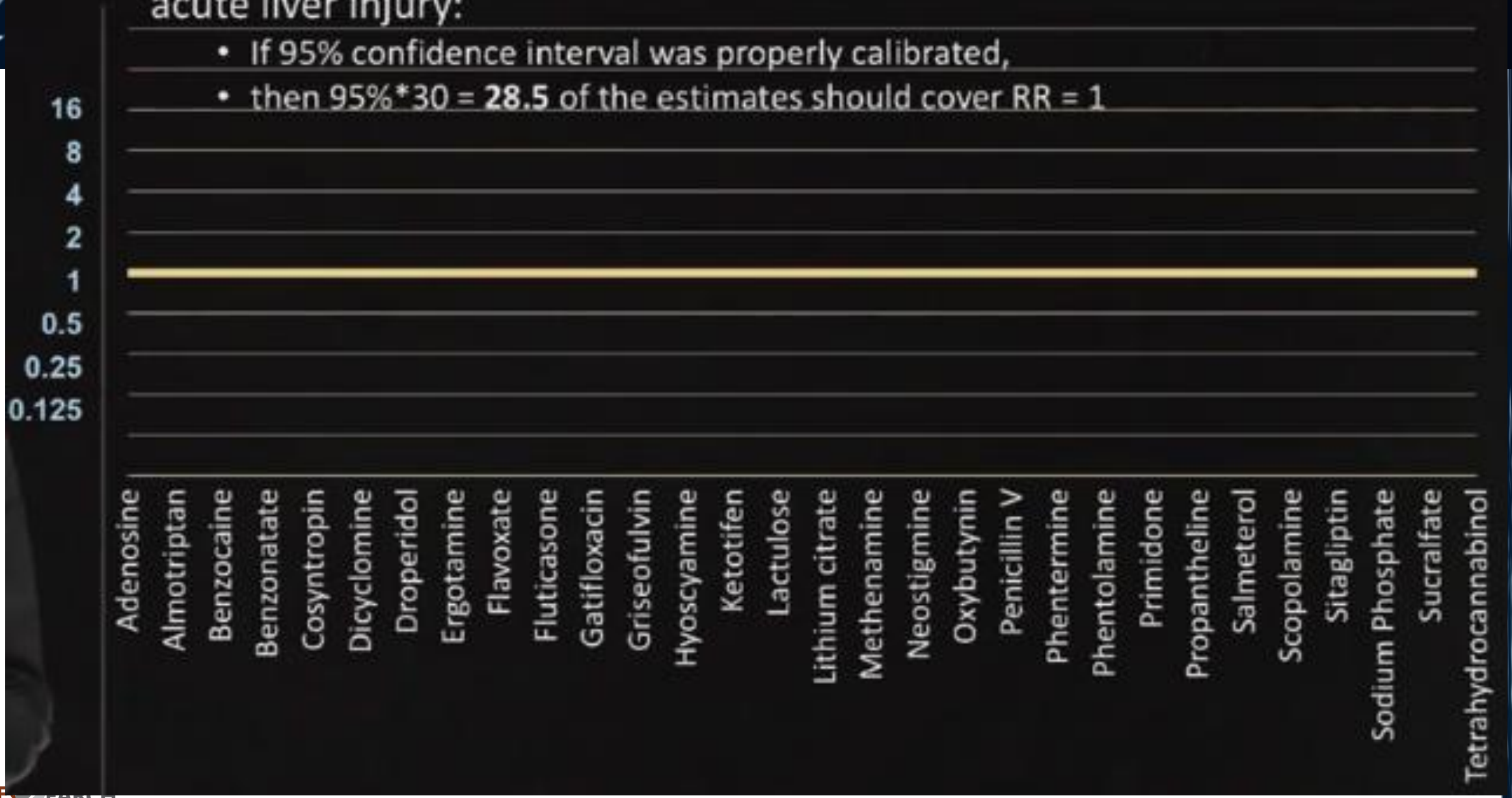
- Observational Medical Outcomes Partnership OMOP project lasted for several years.
- Three important lessons that we can learn from the OMOP project.
- So the first lesson I'd like to focus on concerns operating characteristics.
- The operating characteristics of observational studies in big health care databases are unknown to a large extent.
- And in particular, when we consider statistical artifacts like confidence intervals and p-values, we don't have much understanding of whether the operating characteristics are close to the nominal characteristics.

OMOP project

- Let's focus on the confidence intervals.
- Let's think about 95% confidence intervals.
- They did the following, and they did this over and over in many different contexts.
- They took a particular epidemiological design, in this case, the cohort design, which is a very widely used design and very common in the literature.
- They applied it in a particular database. This is a Medicare database.
- And they applied for **30 negative controls** for acute liver injury.
- So we talked about negative controls previously.

Applying the cohort design to MDCR against 30 **negative controls** for acute liver injury:

- If 95% confidence interval was properly calibrated,
- then 95%*30 = **28.5** of the estimates should cover RR = 1



- So these are drug outcome pairs where we believe there's no association.
- So here we have 30 different drugs along the horizontal axis.
- And we believe through exhaustive literature searches and other means, we're fairly sure that none of these drugs cause acute liver injury.
- So therefore, the true relative risk for each of these drugs and acute liver injury is one.
- So this vertical axis here is on the scale of relative risk.
- Here's one unity.

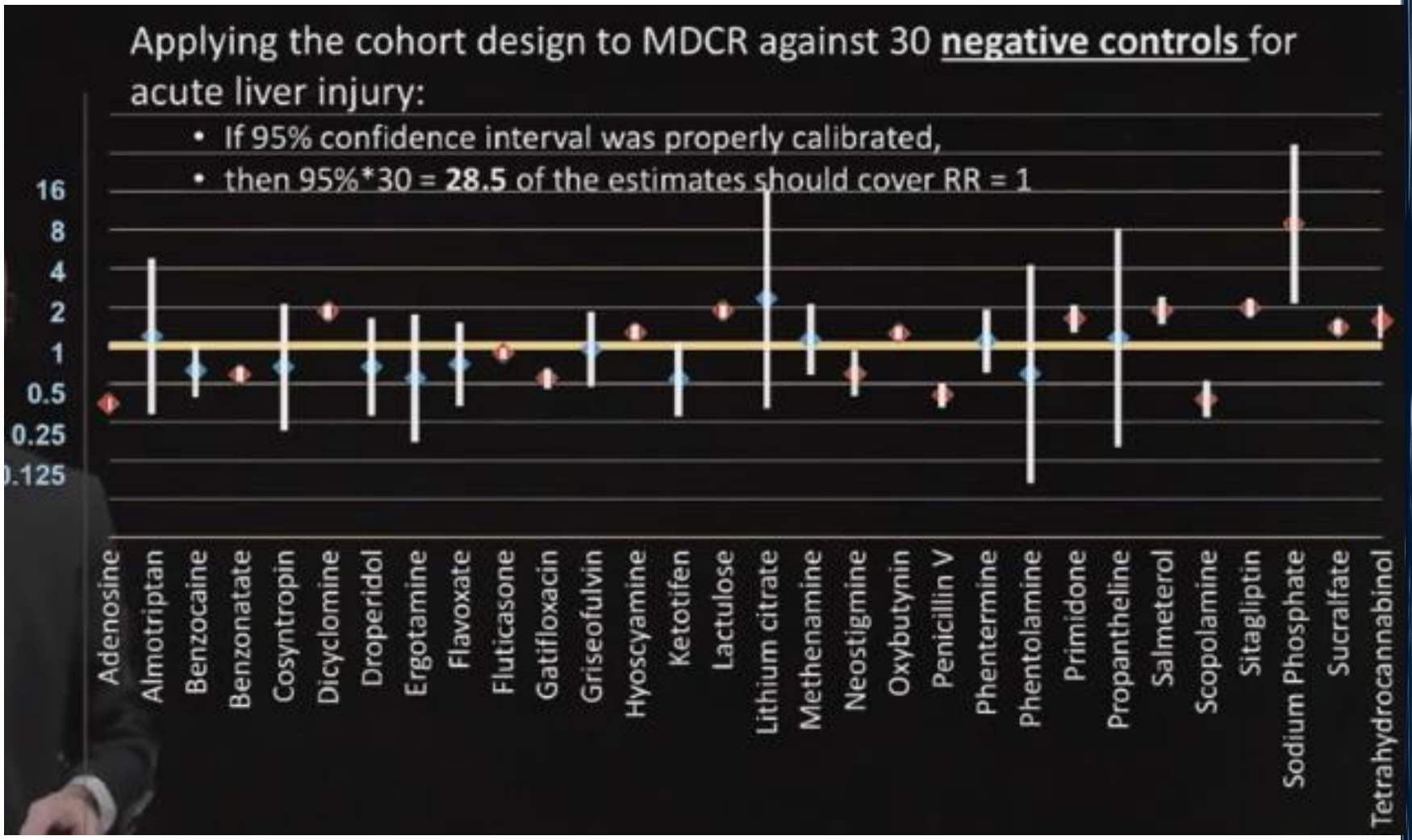
OMOP project

- And so we believe that the true relative risk for every one of these drugs with respect to acute liver injury is one.
- They neither increase the risk of acute liver injury nor decrease the risk of acute liver injury.
- So therefore, if we compute a 95% confidence interval for each one of these drugs with respect to acute liver injury, it should be the case--this is the whole idea of a 95% confidence interval.
- It should be the case that 95% of the 30 intervals contain one, and 5% of them do not.

- Now, as it happens here, 95% is not an integer. It's 28.5.
- So we would expect, when we compute our 30 confidence intervals, that about 28 or 29 of them should contain one, and a smaller number of them, one or two of them, should not contain one.
- And it might not be terribly surprising if it was 27 of the intervals or 30 of the intervals.
- But it should be in and around 28 and 1/2 of the intervals should contain one.

OMOP project

- What actually happened when we did this, when we ran this particular design set up in a particular way in this database for acute liver injury?
- But here are the 30 confidence intervals.



OMOP project

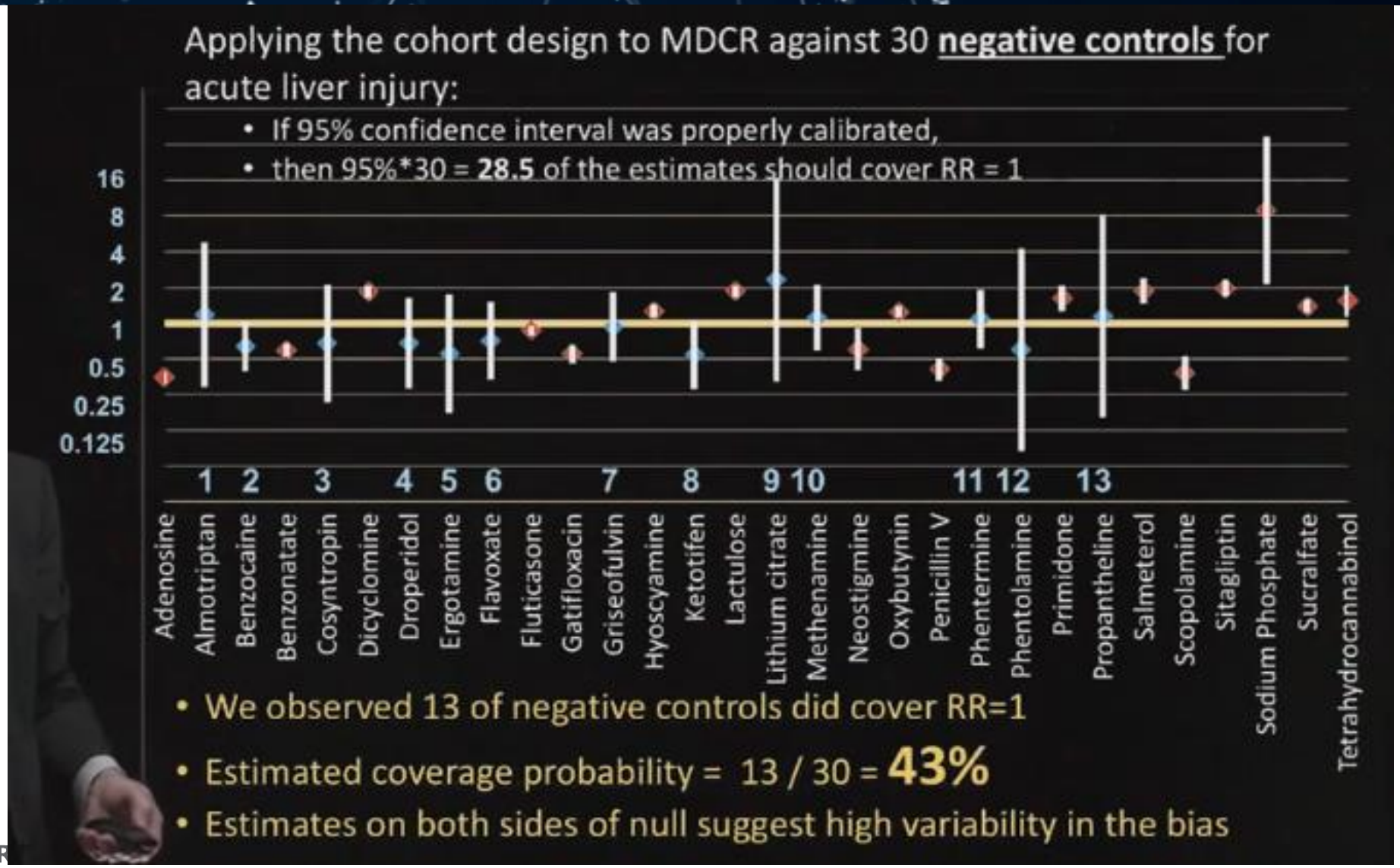
- So each confidence-- let's look at this one here, for example.
- So this confidence interval is a 95% confidence interval for Almotriptan-- that's a drug-- and acute liver injury.
- And as you can see, the point estimate here is just bigger than one.
- The confidence interval goes from something bigger than 4, all the way down to something like 0.4.
- And indeed, this interval straddles (includes) one.
- So as we discussed, 95% of the intervals, 28 or 29 of the intervals, should look like that one.
- They should contain one.

OMOP project

- And a small number of the intervals should look like this one.
- This one's a very narrow interval here.
- This confidence interval, 95% confidence interval, does not contain one.
- It's actually nowhere near one.
- It's all below about 0.5.
- And the reason why some of these intervals are very narrow and some of them are quite wide is because these drugs vary quite significantly in terms of their prevalence, how widely used they are.

OMOP project

- And therefore, in this database, this Medicare database, for some of these drugs we have quite a lot of information, quite a lot of data, quite a large number of patients taking these drugs.
- And hence, we get very precise confidence intervals.
- And for some of the drugs, they're not so widely used.
- So we don't have so much data and you get much wider confidence intervals.


Applying the cohort design to MDCR against 30 negative controls for acute liver injury:
• If 95% confidence interval was properly calibrated,
• then 95%*30 = 28.5 of the estimates should cover RR = 1
16
8
4
2
1
0.5
0.25
0.125
1 2 3 4 5 6 7 8 9 10 11 12 13
Adenosine
Almotriptan
Benzocaine
Benzonatate
Cosyntropin
Dicyclomine
Droperidol
Ergotamine
Flavoxate
Fluticasone
Gatifloxacin
Griseofulvin
Hyoscyamine
Ketotifen
Lactulose
Lithium citrate
Methenamine
Neostigmine
Oxybutynin
Penicillin V
Phentermine
Phentolamine
Primidone
Propantheline
Salmeterol
Scopolamine
Sitagliptin
Sodium Phosphate
Sucralfate
Tetrahydrocannabinol
• We observed 13 of negative controls did cover RR=1
• Estimated coverage probability = 13 / 30 = 43%
• Estimates on both sides of null suggest high variability in the bias

OMOP project

- But anyway, the main point here is to do with operating characteristics.
- What fraction of these confidence intervals actually contain one.
- It's that it should be around about 28. Let's count them all.
- So as we count across here, what we find is that 13 of the confidence intervals contain one.
- And correspondingly, 17 of them do not contain one.
- So therefore, the estimated coverage probability in this context is actually 43%.
- This is not good. That number is meant to be 95%.

OMOP project

- So instead of 95% of the intervals including the truth, one in this case, it's only 43%.
- Now, this particular analysis is a particular design, a particular database, a particular outcome.

OMOP project

- We ran this exact same experiment for many different epidemiological designs on many databases for many different outcomes.
- And the picture is typically like this one.
- So typically, the coverage probability is in the 40% to 60% range.
- This is very problematic.
- It's nowhere near the nominal characteristics.
- Or put another way, when we do a study, an observational study in a health care database, and we produce a 95% confidence interval, we think we're fairly sure in some sense that that interval contains the true relative risk.

OMOP project

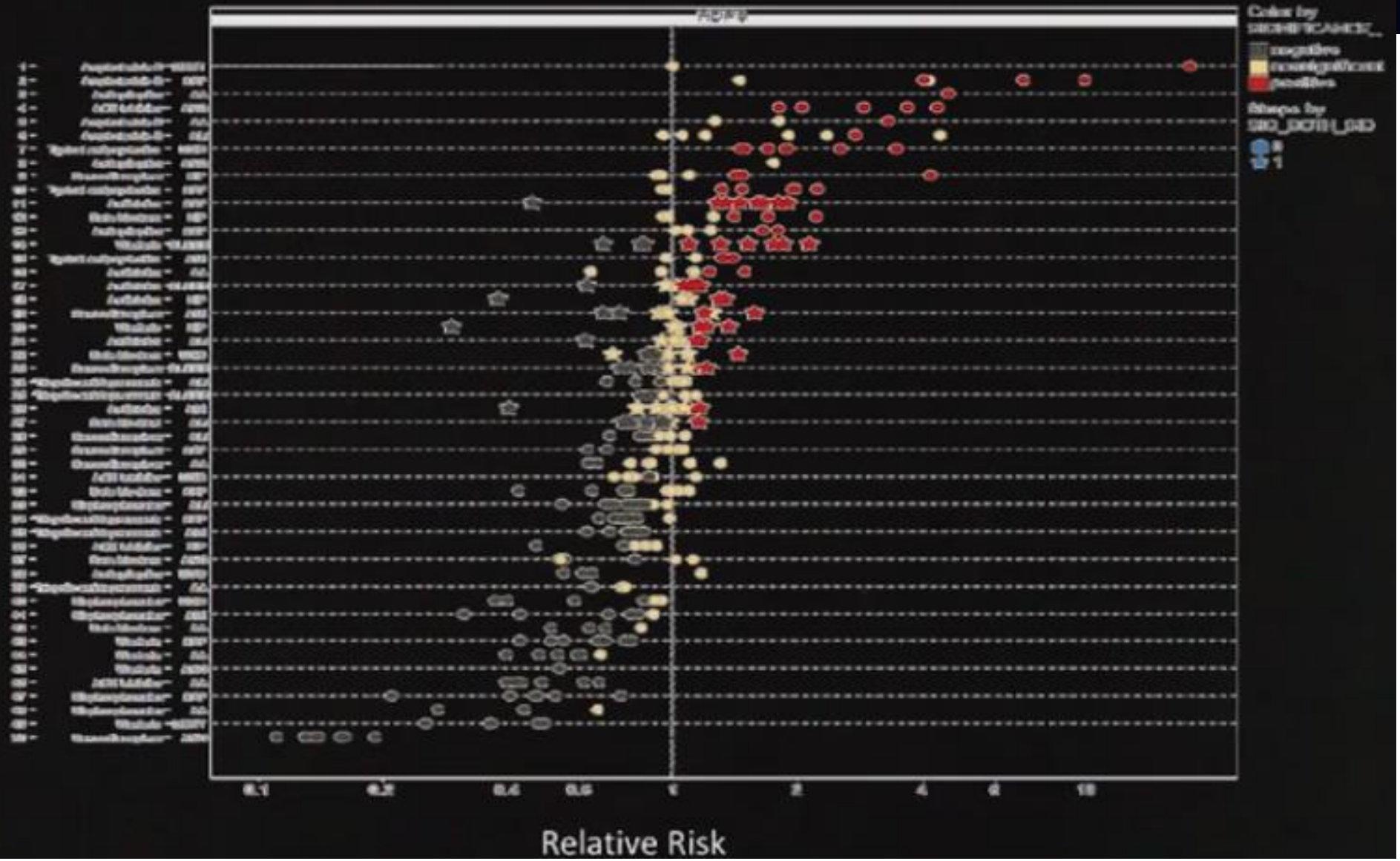
- But in fact, the chance-- it should be 95% chance that it contains the truth.
- The actual chance that it contains the truth is somewhere more like 50%.
- So this is quite problematic in terms of decision making.
- A second lesson that we learned is that database heterogeneity is perhaps a much more significant factor than we thought it.
- So let me explain.
- So we did the following.

OMOP project

- We took, in this particular case, 50 drug outcome pairs.
- These are listed here on the-- the font is quite small.
- But these are, if you will, like 50 studies.
- 50 different drug outcome pairs that we studied.
- We then took a particular analysis.
- We took several particular analyses.
- But in this case here, we took one analysis and we applied that analysis to each of these 50 drug outcome pairs in up to 10 different databases.

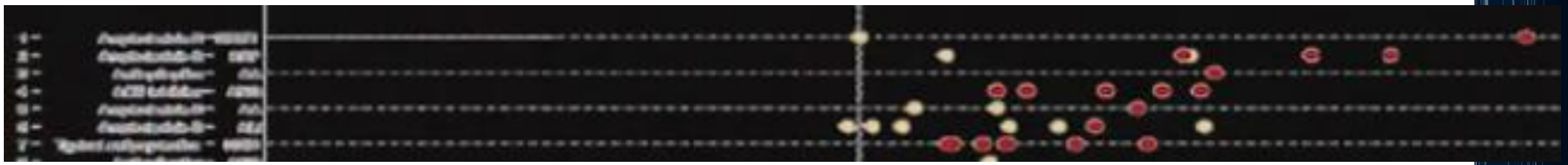
Database heterogeneity

Holding analysis constant, different data may yield different estimates



OMOP project

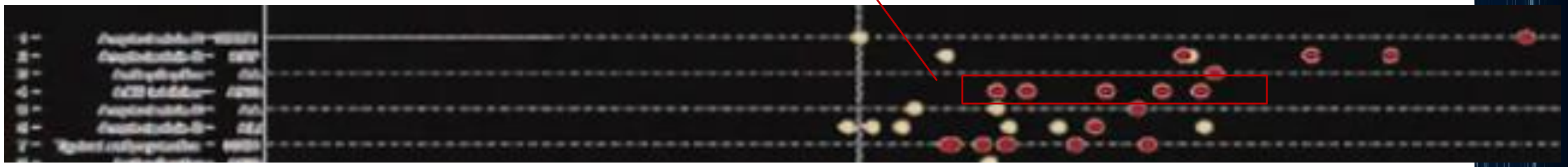
- So what we're looking at here, the dots in this picture are-- so here for example, this is Amphotericin B and aplastic anemia.
- And we ran a particular analysis in several different databases.
- And what we see is that the estimate from each of the databases in which we ran this study is the point estimate is bigger than one.
- And some of these, the yellow ones are not statistically significant.
- And the red one over here is statistically significant.
- So in this particular case, we ran the analysis on three different databases.



OMOP project

And the good news in this context, or let's look at the one directly above it, is there's a qualitative agreement amongst the different analyses.

So in this particular case, if you run this analysis on five different databases, all five of them say that there is a positive effect.



The estimated relative risk is bigger than one.

And all of them say that it's statistically significant.

OMOP project

- However, if you cast your eye up and down this picture, across all these drug outcome pairs, you can see that in many instances,
- you can go from a statistically significant result in one direction, less than one, **showing a protective effect**,
- to a statistically significant effect in the other direction, **suggesting a toxic effect**, depending on which database you use.

Color by
Shapes by
• When applying a propensity score adjusted new user cohort design to 10 databases for 53 drug-outcome pairs:
• 43% of these pairs had substantial heterogeneity in their findings (I^2 > 75%) where pooling would not be advisable
• 21% of pairs had at least 1 source with significant positive effect and at least 1 source with significant negative effect
Relative Risk

OMOP project

- Now, in this particular case, it's a propensity (tendency) score adjusted new user cohort design, which is probably the most widely used design in the health care literature.
- For this particular analysis, what we find is 21% of the pairs had at least one source, one database with a significant positive effect and one database with a significant negative effect.
- So depending on, if you will, you can choose a database to get whatever answer you want. This is not mathematically surprising.
- These databases represent different populations, different patients.

OMOP project

- But if you look at a typical observation study published in the health care literature, what you will find is they choose a database, the Kaiser database, or the GPRD, or whatever.

- They choose a database without comment and show the results of the analysis on that database.
- What this is showing, and it's not just specific to this particular analysis, across a variety of analytic methods,

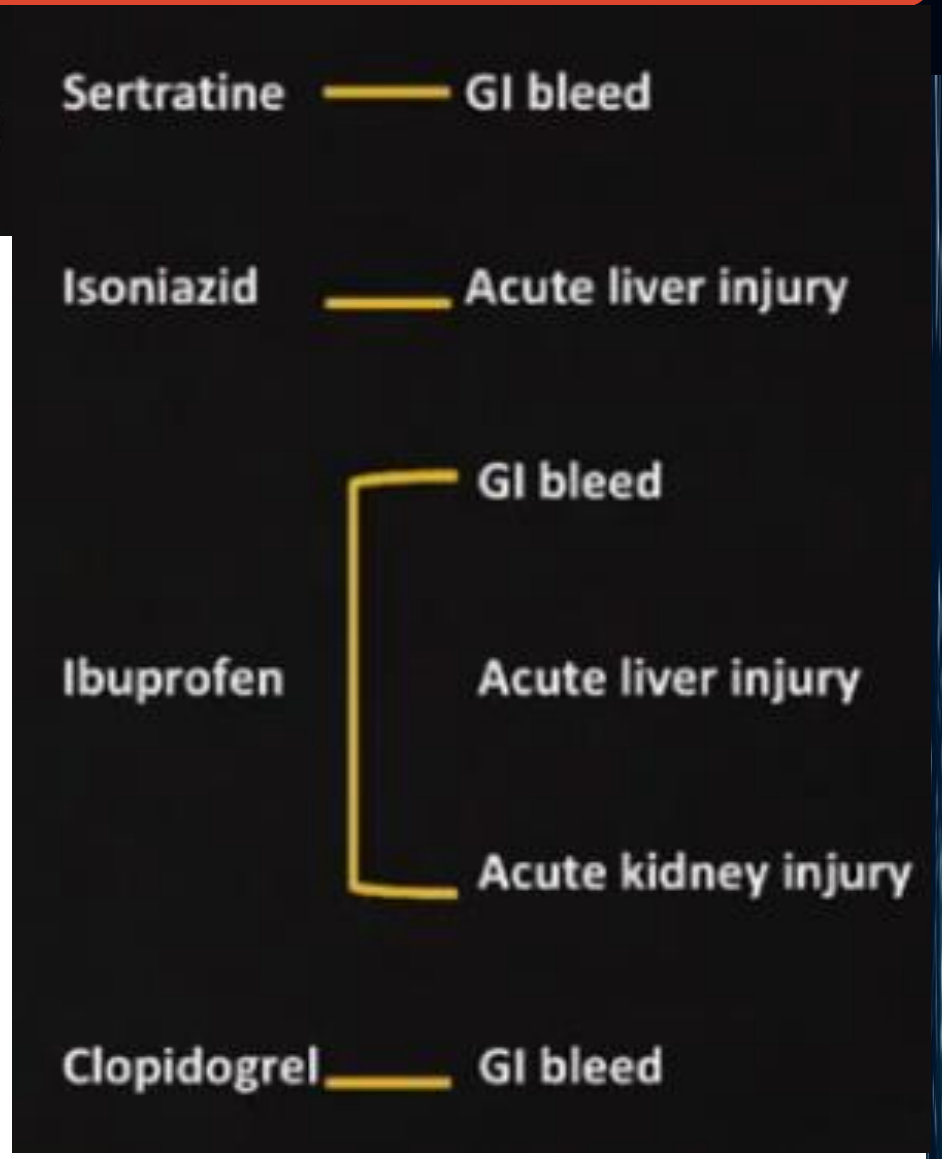
OMOP project

- what we find is that there's between typically in 20% to 40% percent of the cases that we looked at, you can go from statistically significant in one direction to statistically significant in the other direction, depending on which database you pick.
- So put another way, when you pick up a paper that's done an analysis in a database, what the reader generally doesn't know is, had those authors used a different database, they might well have gotten a different answer, which is just quite problematic.

Parameter sensitivity

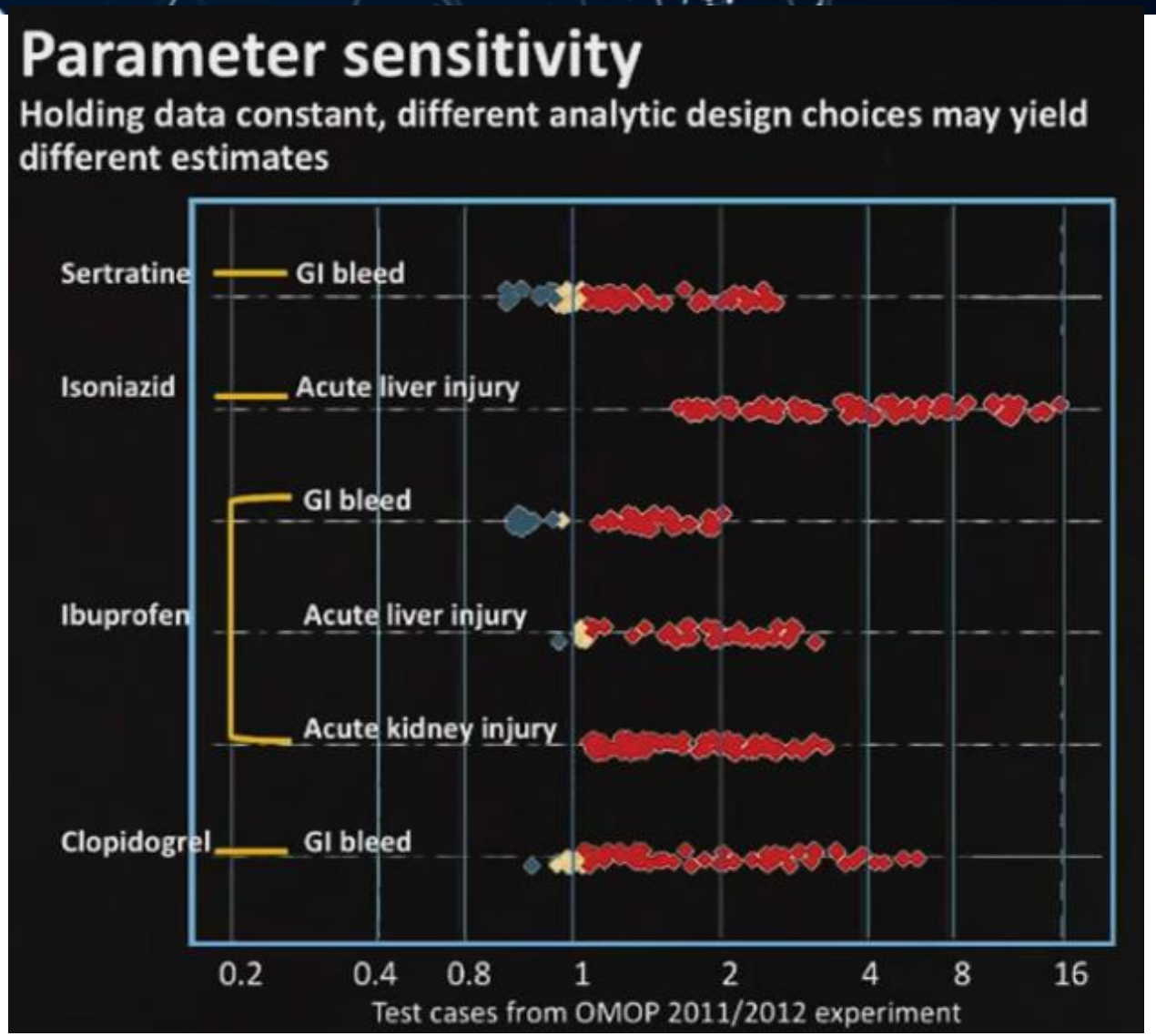
Holding data constant, different analytic design choices may yield different estimates

- The third lesson that I'll just highlight briefly concerns parameter sensitivity.
- So when you conduct these kinds of observational studies, there are various design choices that one makes.
- So for example, I'm showing here an analysis of Sertraline, which is a drug, and GI bleeding.



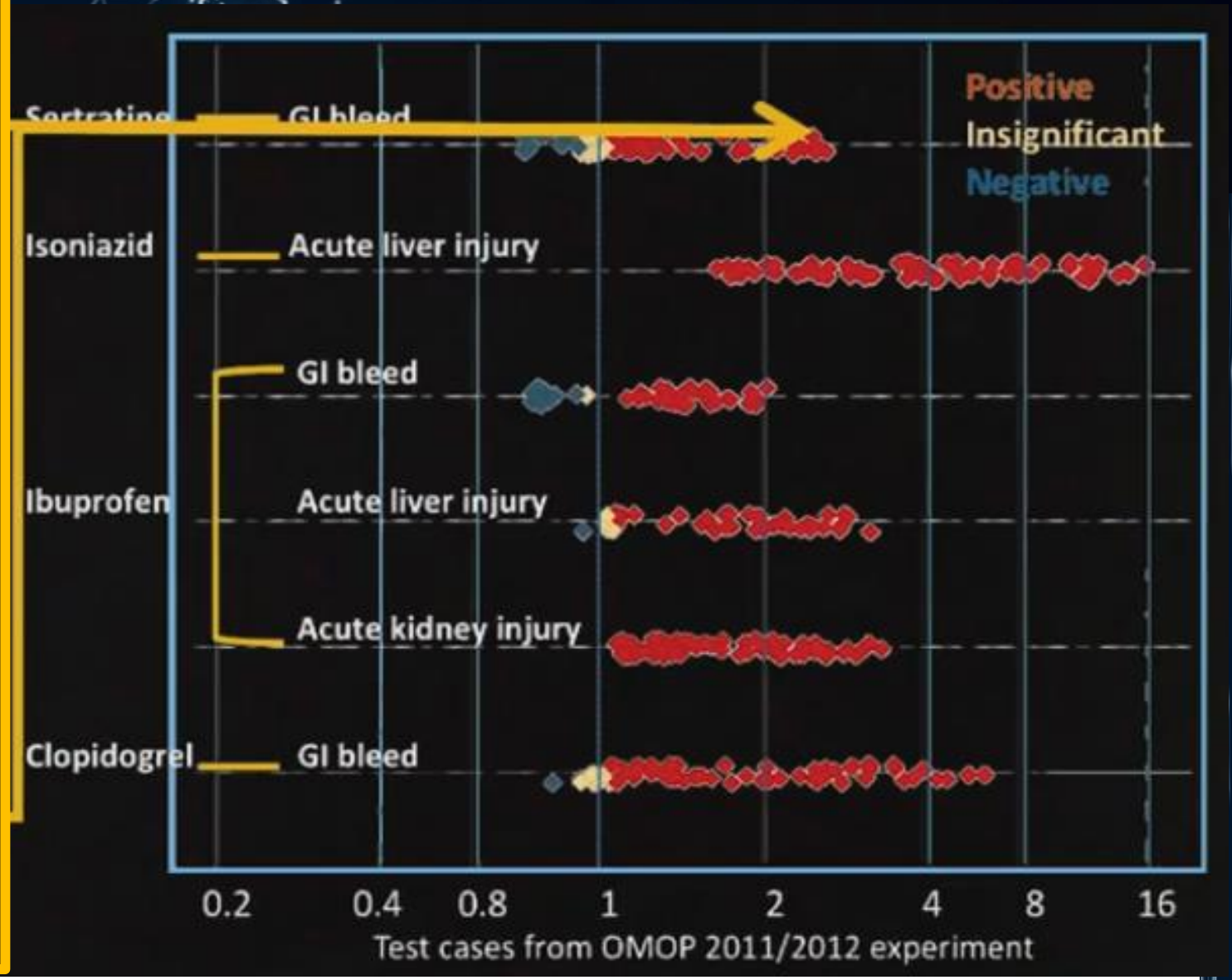
Parameter sensitivity

Holding data constant, different analytic design choices may yield different estimates



Sertraline-GI Bleed:
RR=2.45(2.06-2.92)

- Controls per case: up to controls per case
- Required observation time prior to outcome: 180d
- Time-at-risk: 30d from exposure start
- Include index date in time-at-risk: No
- Case-control matching strategy: Age and Sex
- Nesting within indicated population: No
- Exposures to include: First occurrence
- Metric: Odds ratio with Mantel Haenszel adjustment by age and gender



- This particular analysis yields a relative risk of 2.45.
- And you can see the 95% confidence interval.

OMOP project

- So for that particular analysis, these are some of the design choices that correspond, that go into making up that-- to producing that 2.45 and that 95% confidence interval.
- This particular case, it's a case control design.
- So the kinds of choices that an analyst has to make when implementing a study like this, you have to choose the number of controls per case.

Sertraline-GI Bleed: RR=2.45(2.06-2.92)

- **Controls per case: up to controls per case**
- **Required observation time prior to outcome: 180d**
- **Time-at-risk: 30d from exposure start**
- **Include index date in time-at-risk: No**
- **Case-control matching strategy: Age and Sex**
- **Nesting within indicated population: No**
- **Exposures to include: First occurrence**
- **Metric: Odds ratio with Mantel Haenszel adjustment by age and gender**

OMOP project

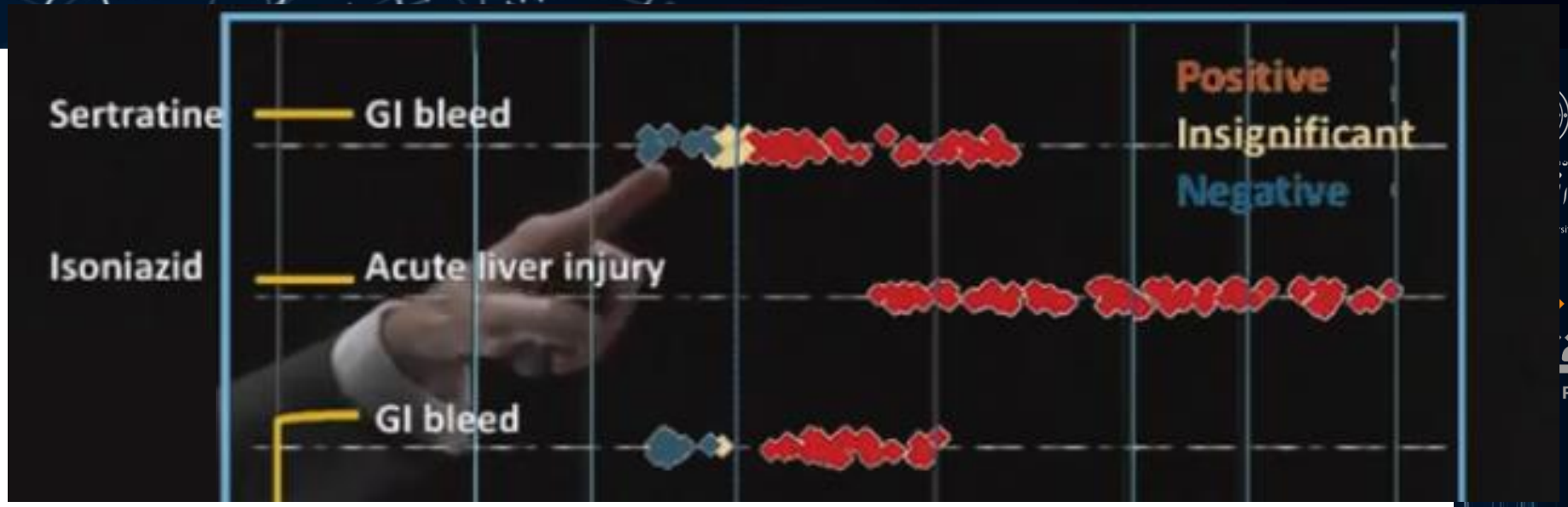
- You have to define a clean period prior to the outcome.
- You have to define the time at risk.
- In this particular case, it would be 30 days from the start of exposure to the drug.
- And so on.

Sertraline-GI Bleed: RR=2.45(2.06-2.92)

- **Controls per case: up to controls per case**
- **Required observation time prior to outcome: 180d**
- **Time-at-risk: 30d from exposure start**
- **Include index date in time-at-risk: No**
- **Case-control matching strategy: Age and Sex**
- **Nesting within indicated population: No**
- **Exposures to include: First occurrence**
- **Metric: Odds ratio with Mantel Haenszel adjustment by age and gender**

OMOP project

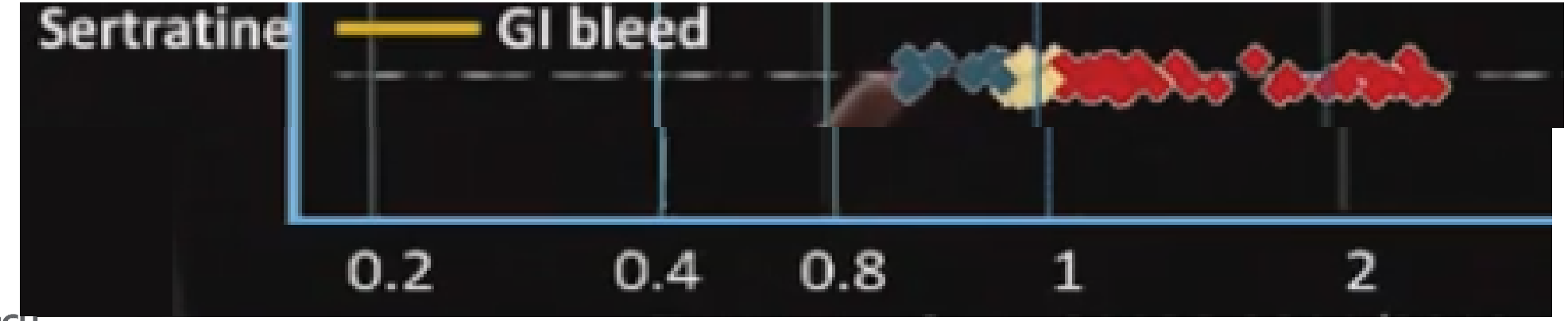
- So there are a variety of choices that an analyst has to make.
- These are quite subjective in nature.
- There's no obvious right choice for any of these particular design choices.



- So what we're showing in this picture here is each dot corresponds a variant on those design choices that I just showed you.

OMOP project

- So in this particular case-- this is Sertraline and GI bleeding-- what we see is, depending on the choice that you make, you can go from a large, positive, statistically significant estimate, bigger than two, to a statistically significant protective effect, or negative effect, an estimate less than one.
- And basically, everything in between, depending on the particular choices you make.



OMOP project

- This is unfortunate.
- We would prefer if our analyses were robust to these choices so that no matter what choices you made within reason, you would get qualitatively similar findings.
- That is not the case.
- And this is not specific to Sertraline and GI bleeding.
- Across many drug outcome pairs and many types of analytic approaches, this is what we found, is that there is a lot of sensitivity to choices for which different analysts will, in general, make different choices.

Graphs Are Comparisons

- In this model, we'll discuss exploratory data analysis and visualization.
- **Exploratory data analysis** refers to display of data-- or more generally, display of any numerical information--in a way that can allow us to discover patterns that we did not expect to see.
- **Visualization**, more generally, refers to the techniques that we use to see data or to see the numerical information.
- Exploratory data analysis involves three challenges.
- The first is how to graph the data, how to display the numerical information in a way that's clear.

RESEARCH
FANAP
Khatam University

Graphs Are Comparisons

- There's been a lot of research on this.
- A lot of techniques have been developed that can allow us to display a lot of information in a small space.
- Why is it important to display information in a small space?
- If we can display information clearly and concisely in a small space, we can make several graphs, as well demonstrate.
- And as a researcher, as a scientist or an engineer, you can look at these several graphs and look at patterns, look at changes in a way that you wouldn't be able to see if you were merely looking at a table of numbers or looking at a very large chart that had very little information.

Graphs Are Comparisons

- Dense display of information allows you to see more and make more comparisons.
- So the first challenge of exploratory data analysis is displaying the data in high resolution and in an effective way, moving beyond pie charts and bar charts to graphs that can
- allow more information to be displayed in a smaller space.
- The second challenge of exploratory data analysis is interpreting your results.
- As I said, the goal of EDA is to see patterns, to discover things in the data that you were not expecting to see.

Graphs Are Comparisons

- So we have to train ourselves.
- We have to train your eyes, train our brains to be able to read graphs and pull out what we weren't expecting.
- And then, after that, to go back to our data and inspect the individual data points,
- or inspect the trends in order to understand why what we saw happen and why it was unexpected.
- The final challenge of exploratory data analysis is modeling.
- As I said, EDA involves finding things that were unexpected.
- So what was expected?
- What's expected is based on your statistical model.

RESEARCH
Khatam University
FANAP

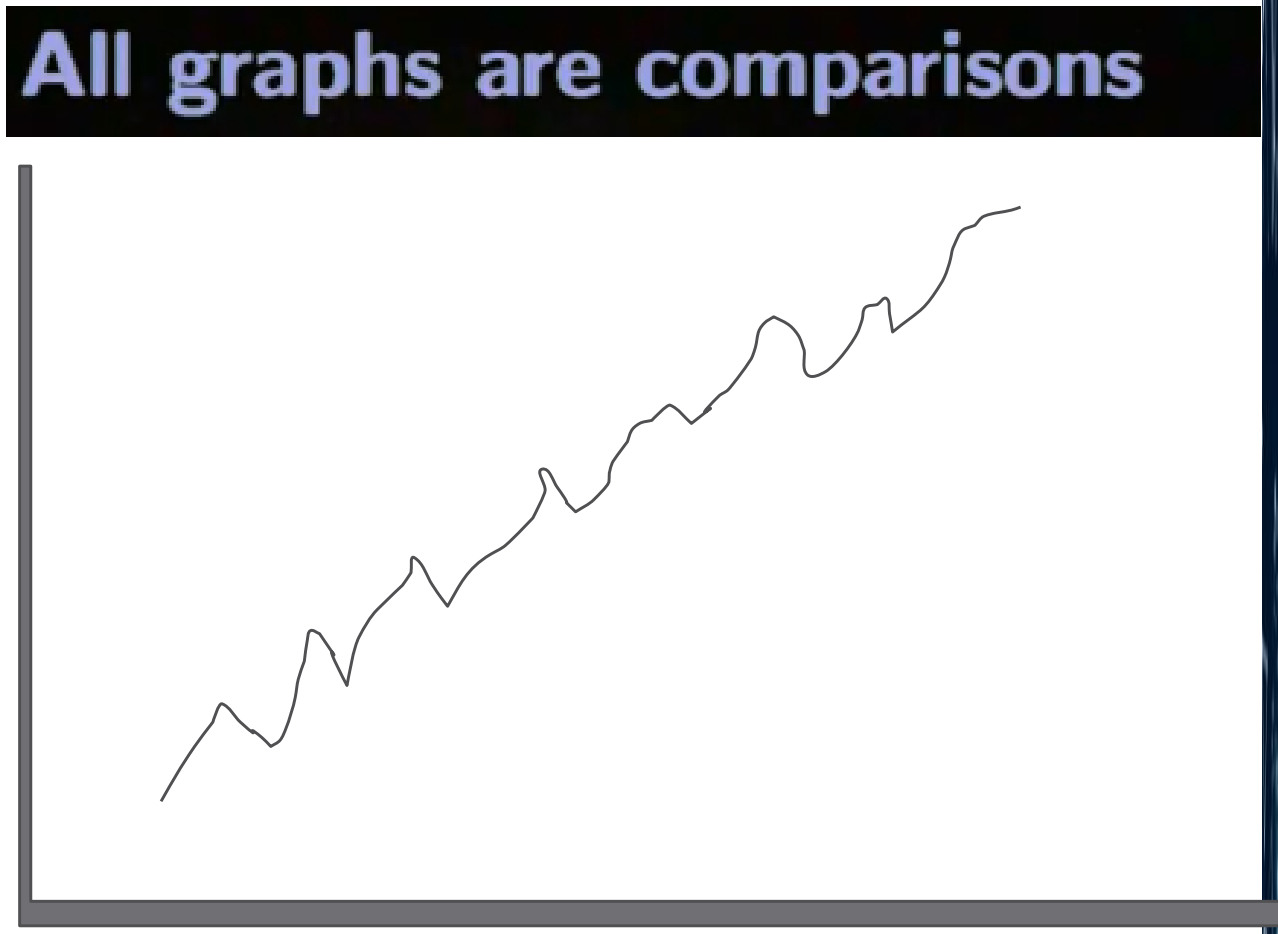
Graphs Are Comparisons

- Statistical models, such as regression, can be used to summarize what we know about our data.
- The better our model is, the more we learn when, through exploratory data analysis, we find something surprising.
- What is a graph?
- What is a data display?
- I'm going to make the claim that every data display is a comparison, either explicit or implicit.
- So let's show some examples.

RESEARCH

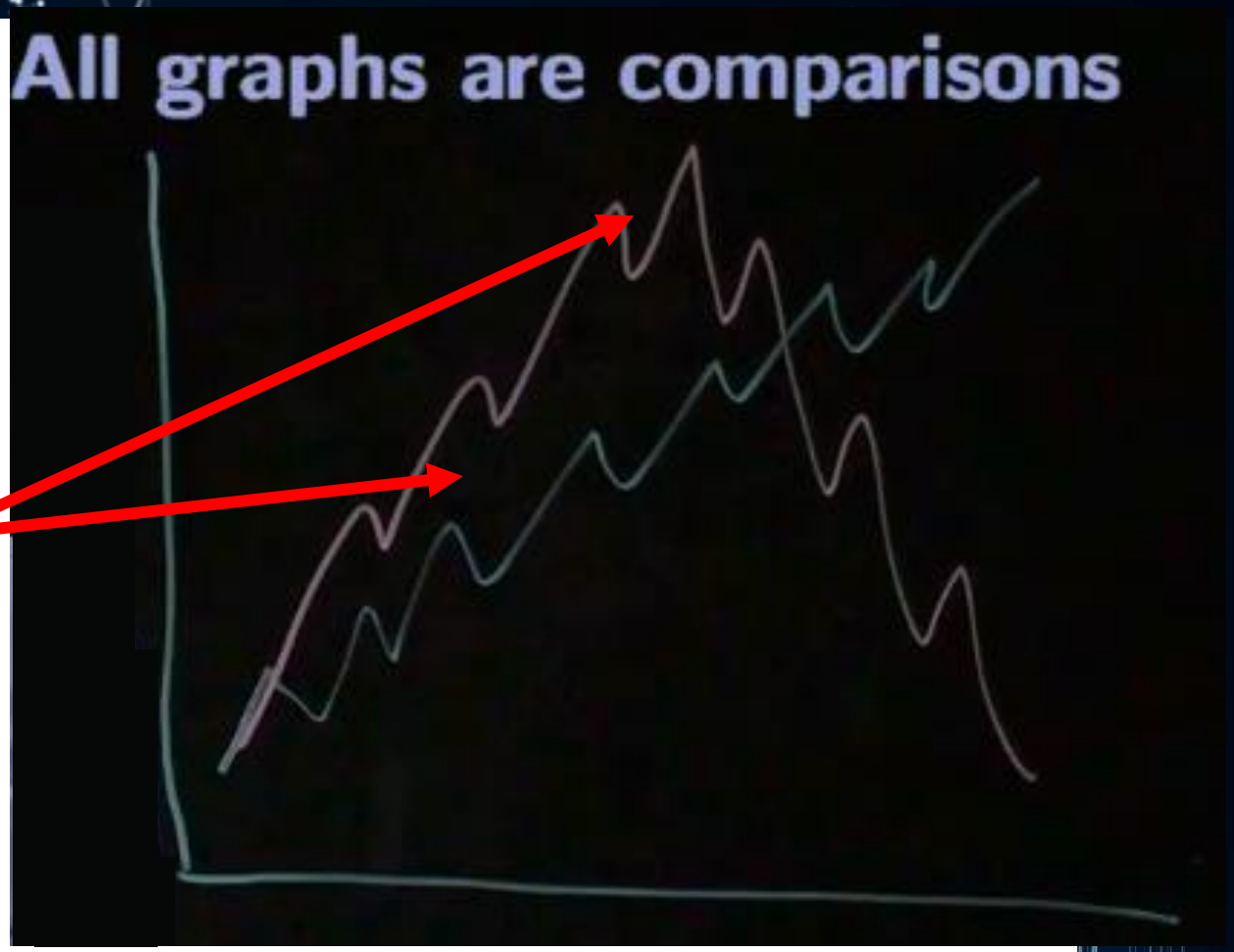
Graphs Are Comparisons

- What about a simple time series going like this?
- That's a comparison.
- What do I mean by comparison?
- It's only one line.
- It's an implicit comparison to a horizontal line.
- It's going up.



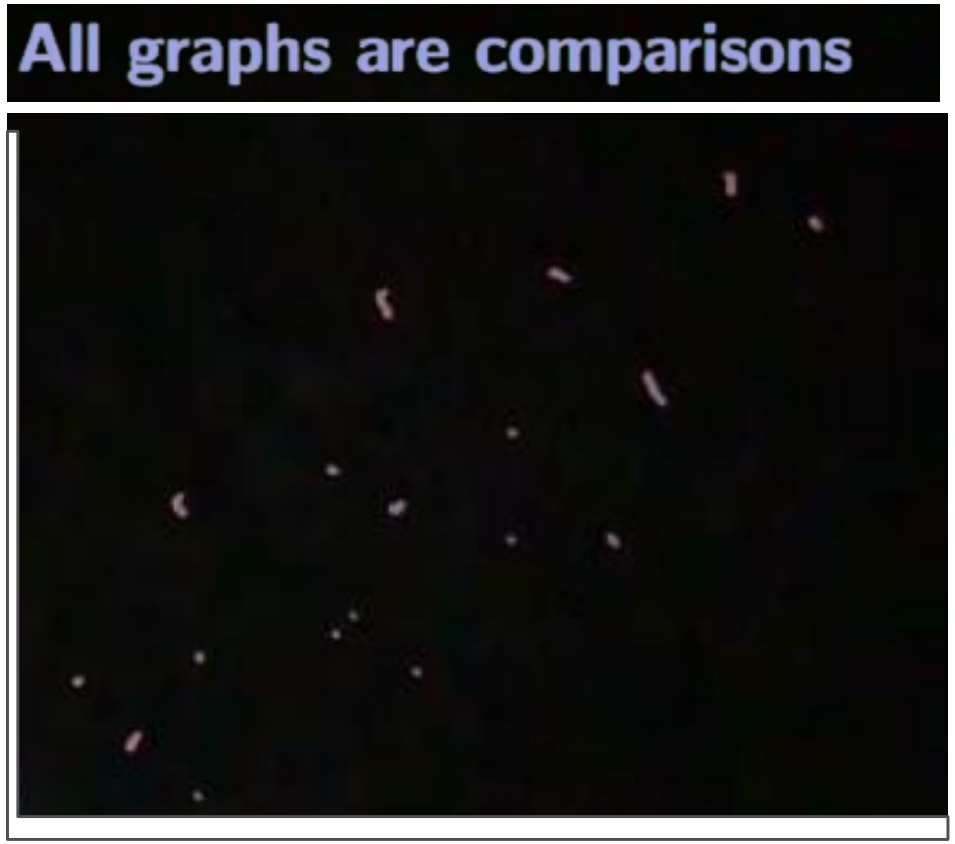
Graphs Are Comparisons

- Here's another time series.
- That's a comparison, too.
- I would say it's an implicit comparison to a line that continues going up.
- It's a comparison to a monotonic increase.
- And we see a change.
- Sometimes the comparison is more clear.



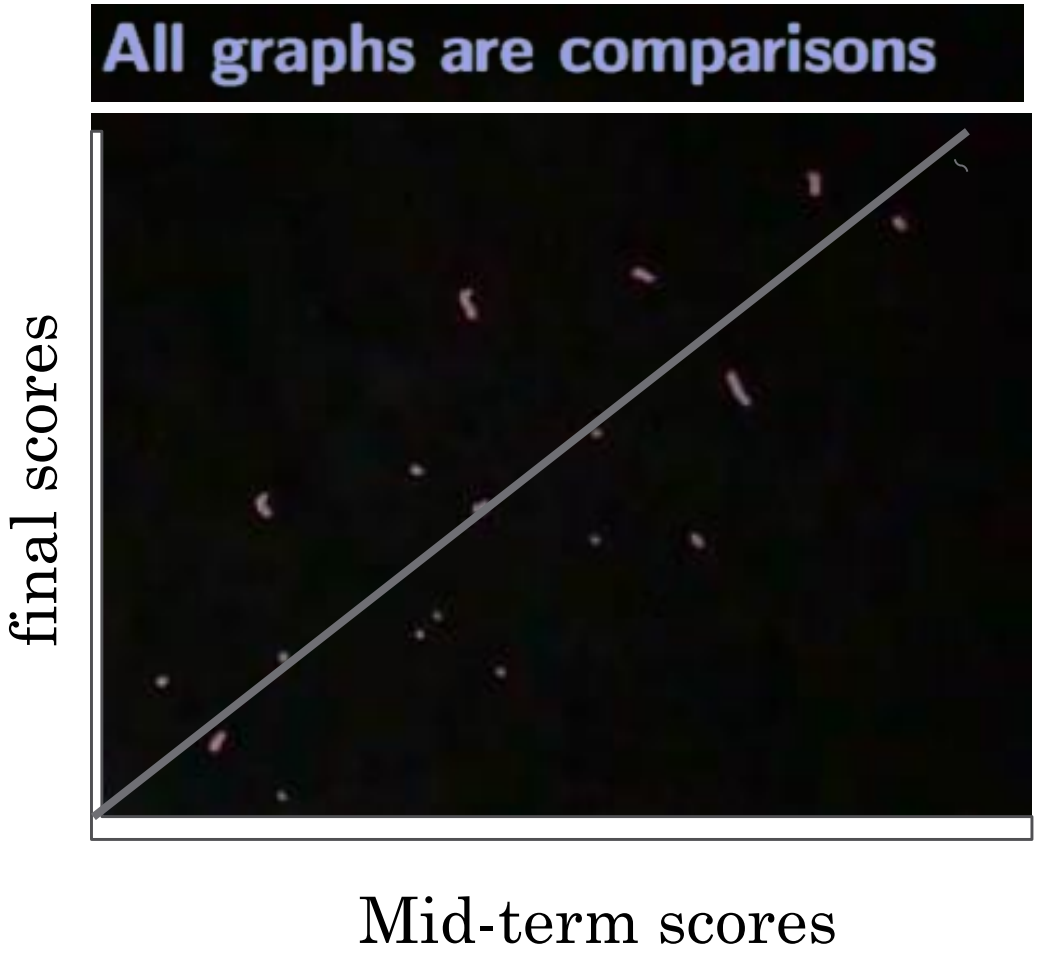
Graphs Are Comparisons

- Suppose we have test scores for students.
- We have test scores on the midterm and the final.
- And we have some data.
- the scores go between 0 and 100.
- By making a scatterplot here, we're
- comparing the different students.
- We're seeing that they did differently.
- They didn't all get the same score.



Graphs Are Comparisons

- We can draw a dotted line at 0 and we
- can compare to no change.
- It looks like some of the students did better,
- some did worse.
- Maybe these are grades for students in one class.
- Then I can show you grades for students in another class.



Graphs Are Comparisons

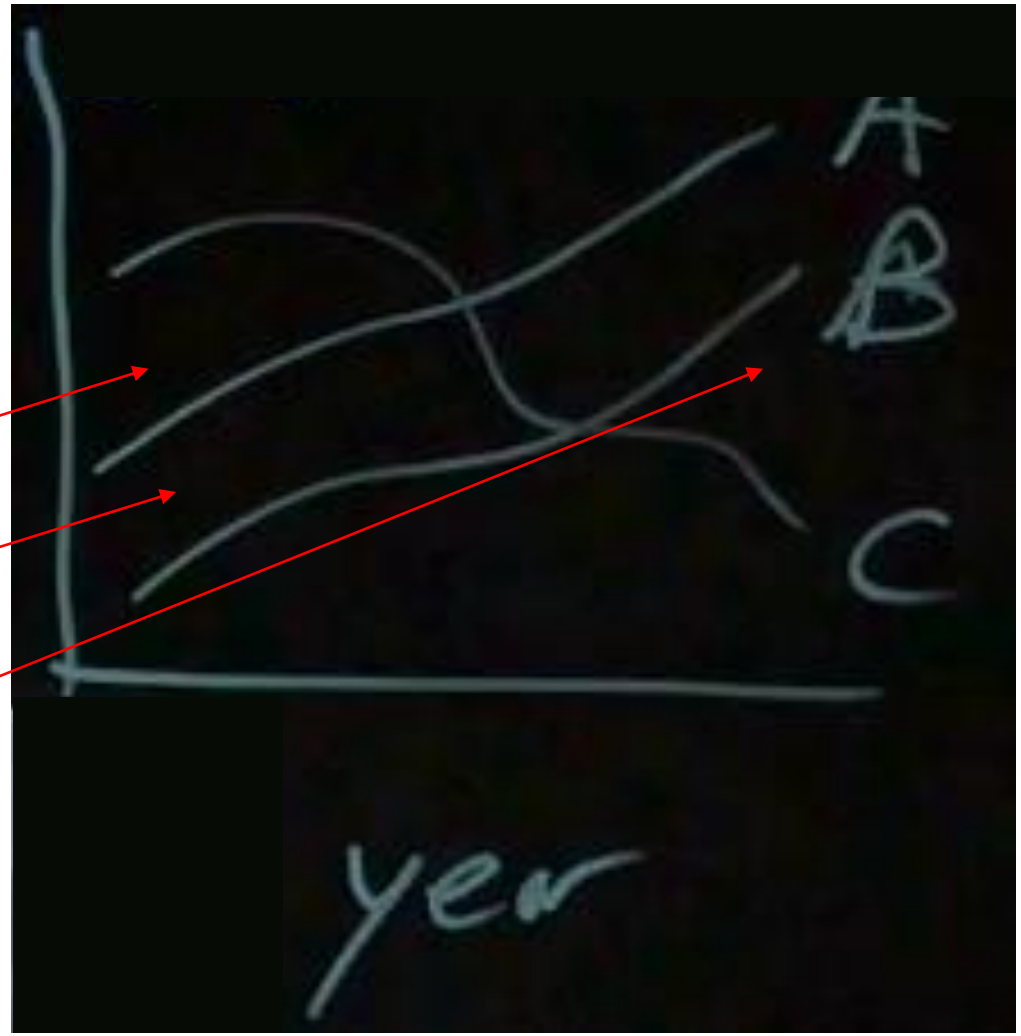
- We can draw a dotted line at 0 and we
- can compare to no change.
- It looks like some of the students did better,
- some did worse.
- Maybe these are grades for students in one class.

Graphs Are Comparisons

- Then I can show you grades for students in another class.
- They're different, too.
- I can draw a regression line, in which case
- I'm comparing this line to this line, and I'm also comparing the points to the lines.

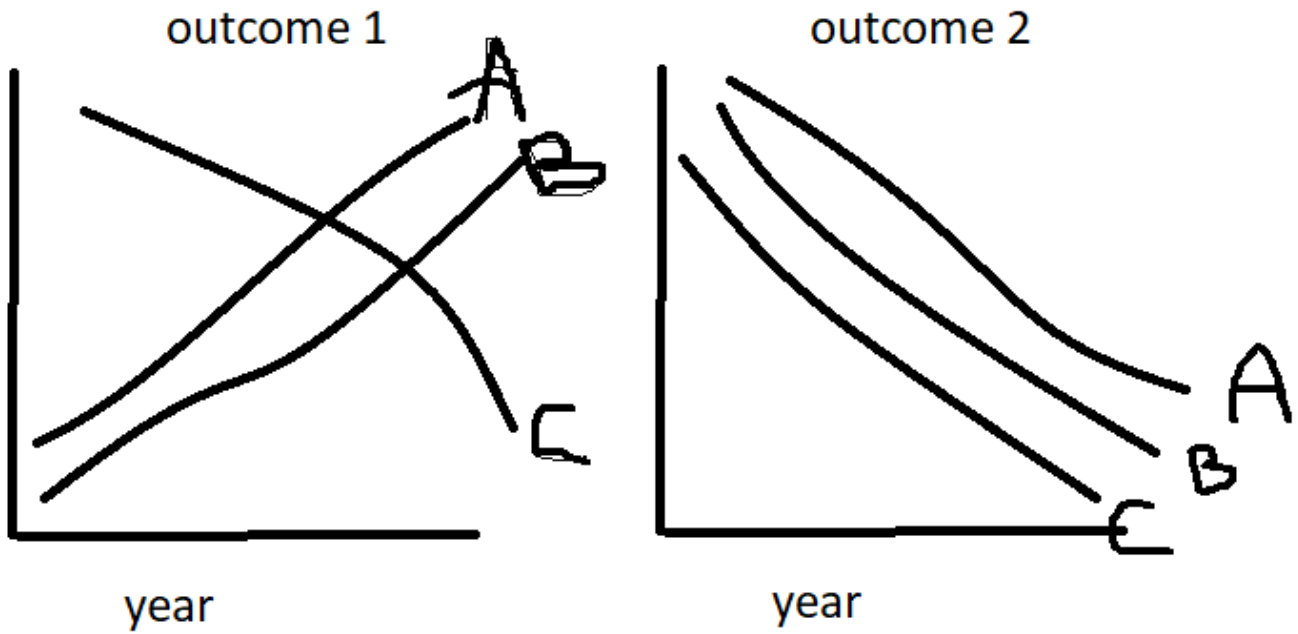
Graphs Are Comparisons

- Or perhaps we have trends in different countries.
- So here, we have time.
- Here, we have some outcome.
- And we see that in one country it was doing this.
- In another country we are seeing this-- B.
- And in another country we were seeing this.



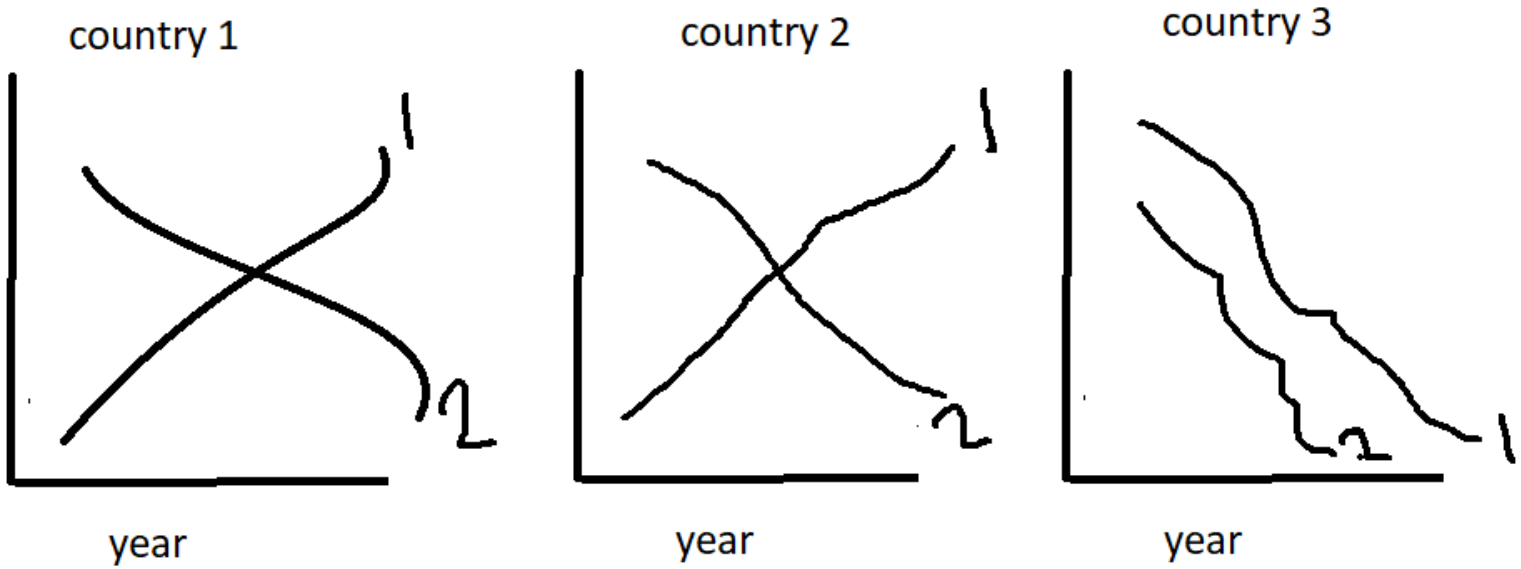
Graphs Are Comparisons

- And we can do further comparisons.
- So maybe this is outcome 1.
- And this is outcome 2.
- And in outcome 2 we have a different trend.
- A does this, B does this, and C does that.



- This paragraph allows us to compare
- what's going on in the different countries for one outcome and the other.
- Alternatively, we could display it a different way.

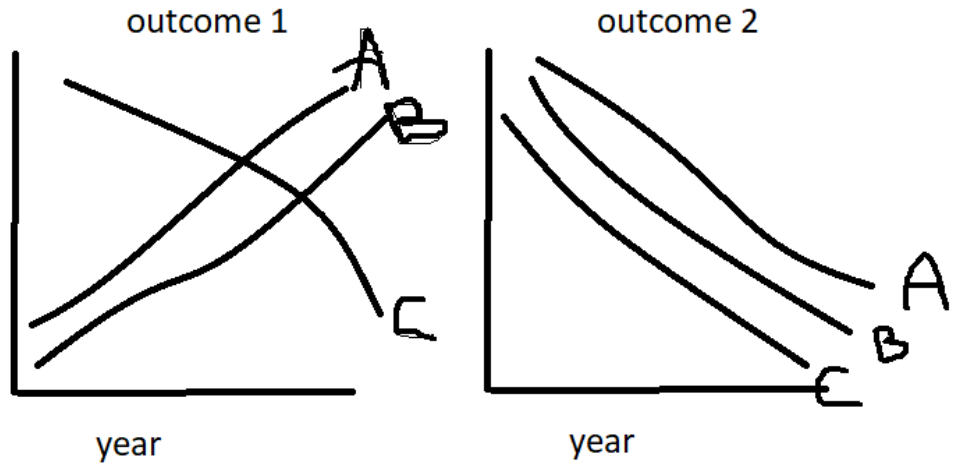
Graphs Are Comparisons

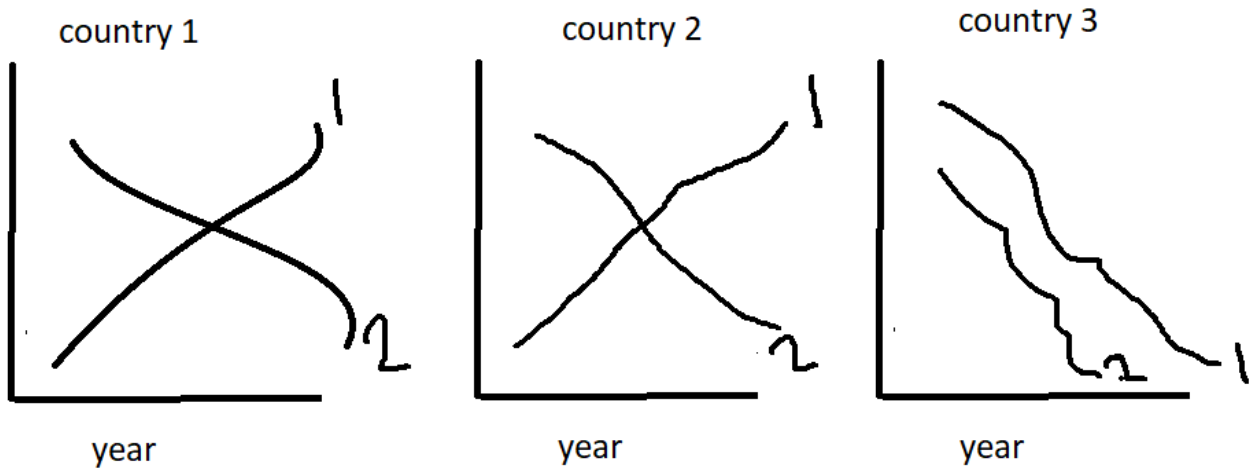


- We could have graphs for the separate countries.
- So we could show country 1, country A, country B, and country C. And then we could display outcome 1.
- So I'm trying to read this off here.
- Which of these displays is better?

Graphs Are Comparisons

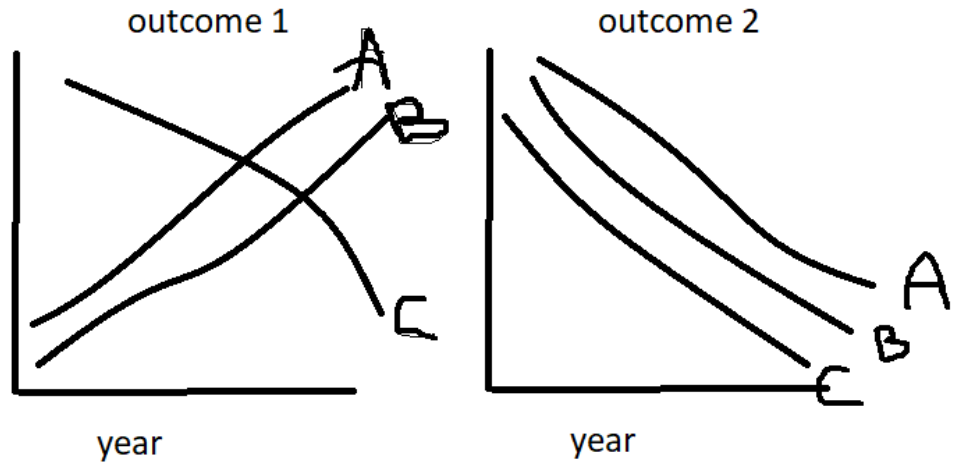
- It depends on our purpose.
- If our goal is to compare the countries directly, it probably makes sense to look at one outcome and then the other.
- On the other hand, if our goal is to compare the two outcomes within each country,
- then this sort of displayed will be more effective.

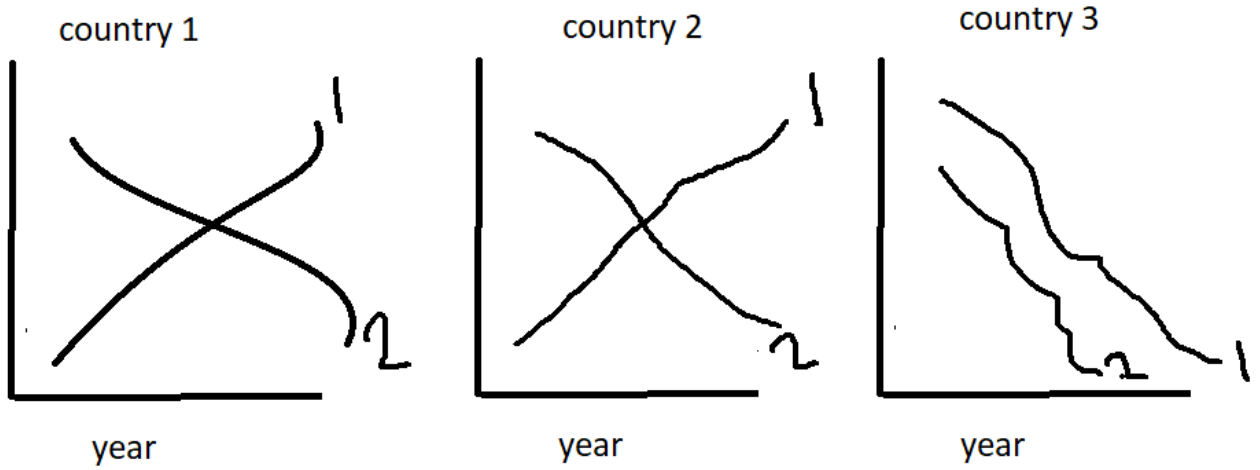




Graphs Are Comparisons

- I'm empathizing graphs as comparisons for two reasons.
- First, when interpreting a graph, when reading a graph, we want to think in terms of comparisons.
- Sometimes it's obvious.
- We're comparing A to B or A to C.
  We're comparing 1 to 2.
- Other times it's more subtle.
- We see a graph with a single line, and it's implicitly being compared to what we are expecting to see.

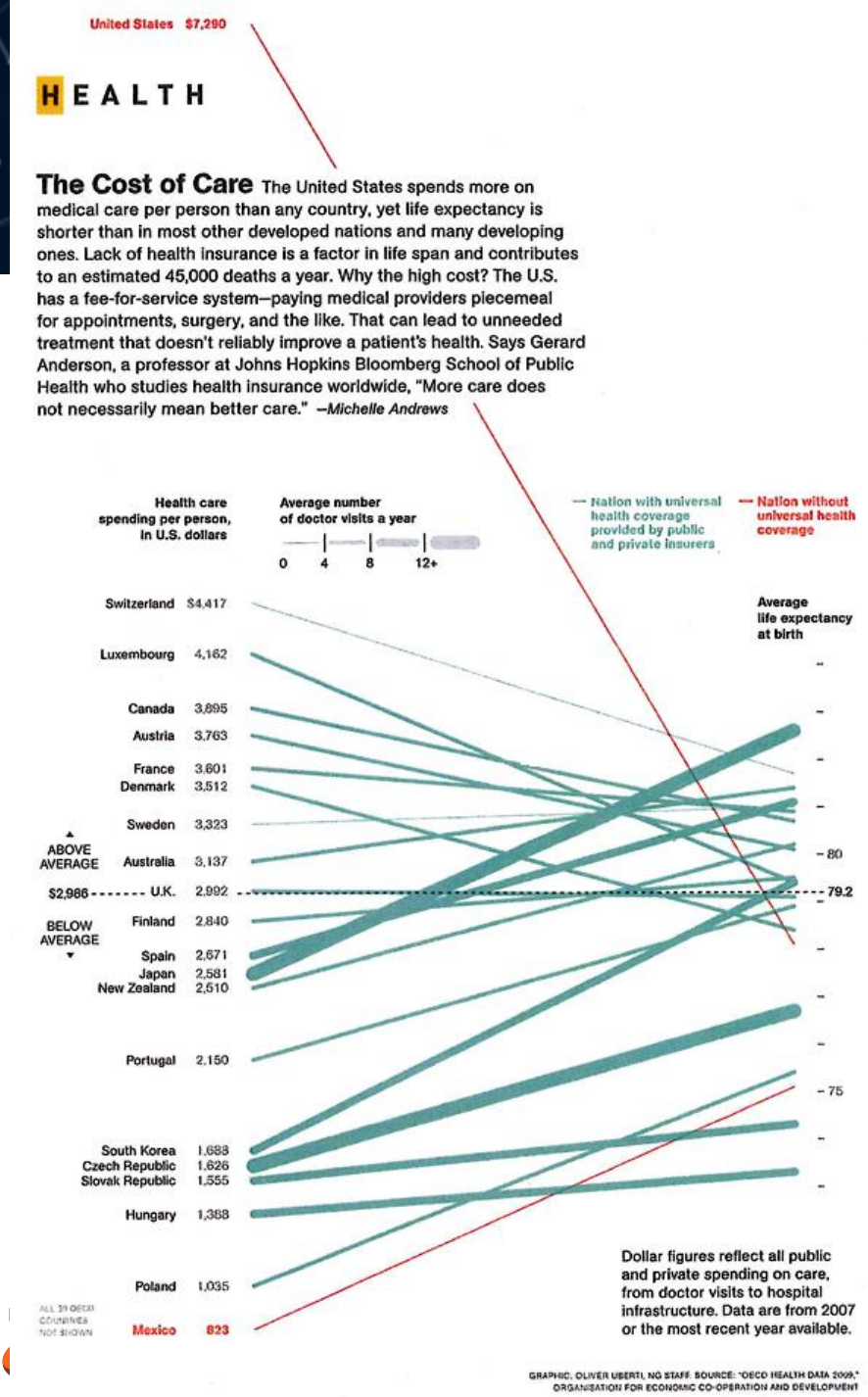




Graphs Are Comparisons

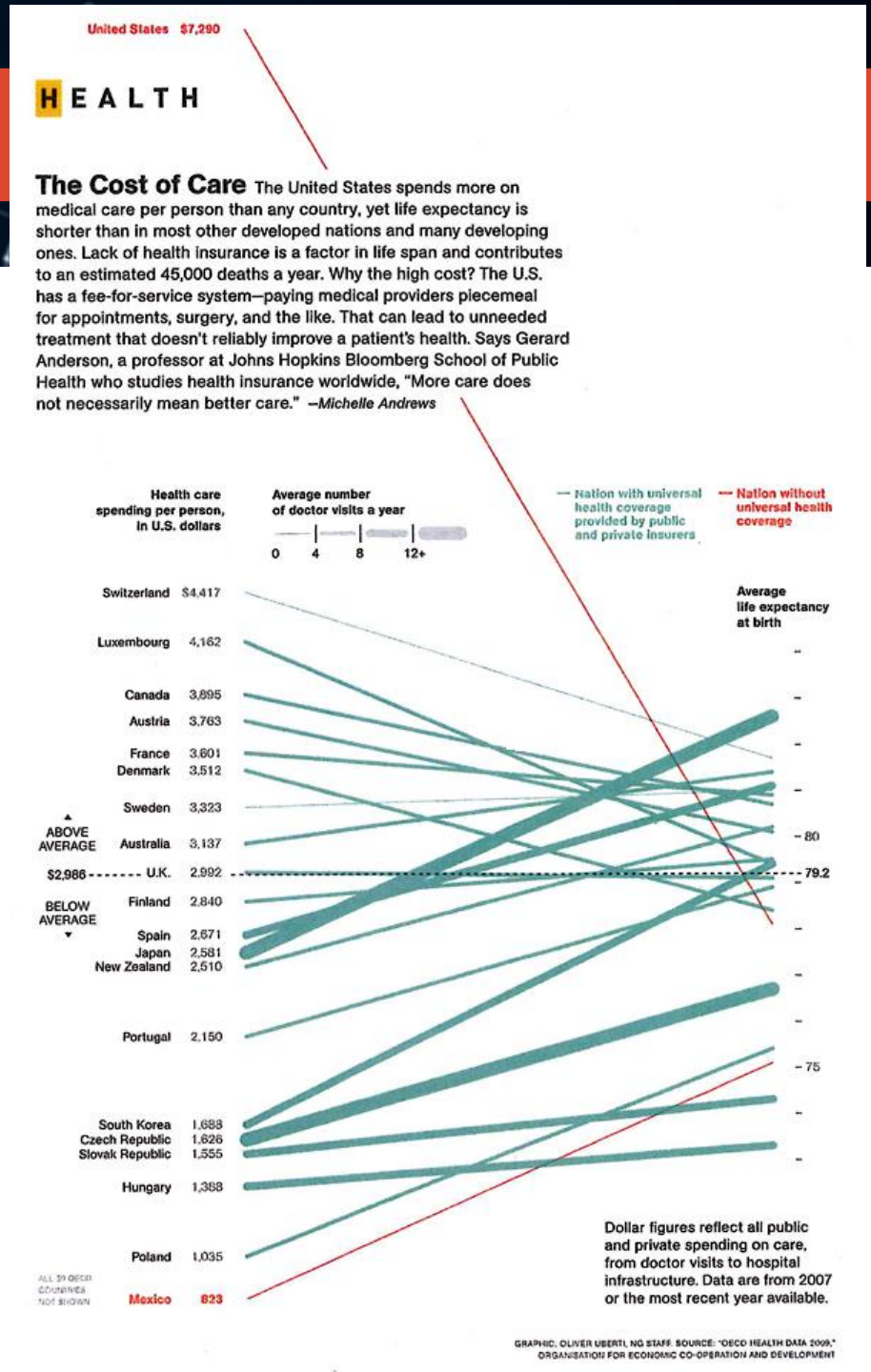
- So that's one reason to understand
- that graphs are comparisons, in order to better read graphs.
- The second reason I'm emphasizing that graphs are comparisons is because when you make a graph,
- you can be more effective if you understand what comparisons you're particularly interested in.

Graphs Are Comparisons

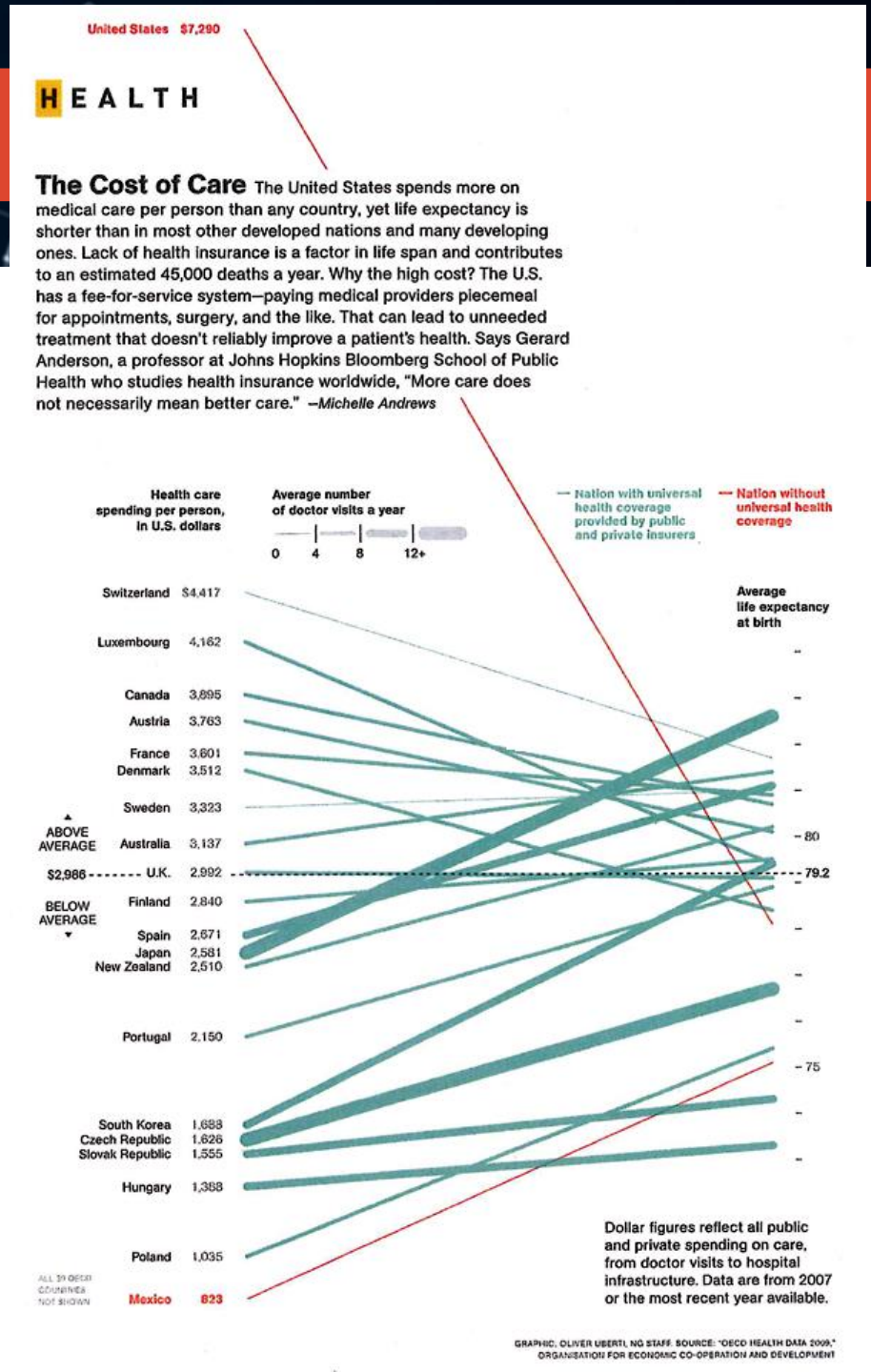


- I'll now demonstrate these principles in the context of several examples.
- This graph, which went viral on the web a few years ago, displays health care spending and life expectancy in several countries.

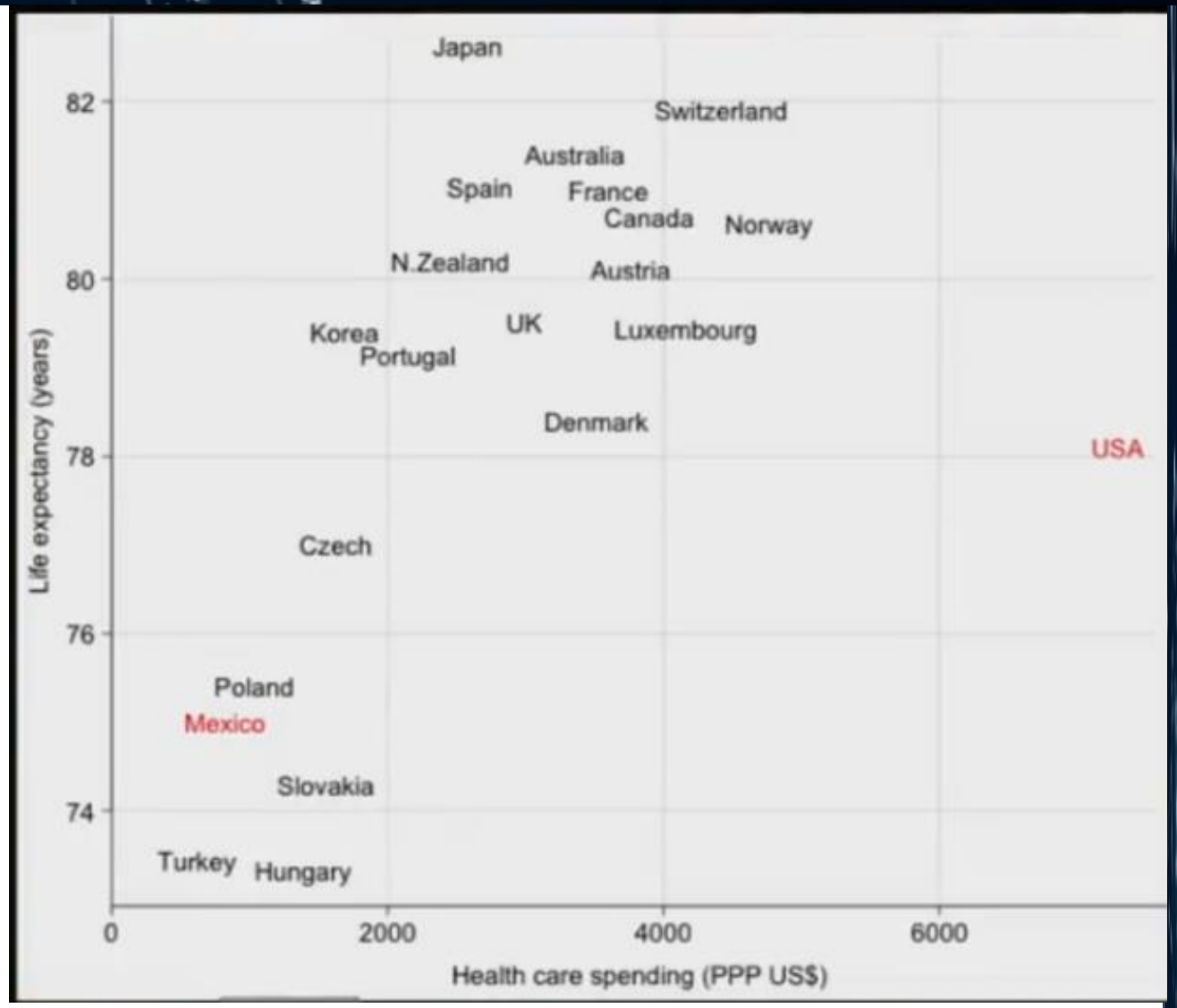
- The left axis shows health spending.
- The right axis shows life expectancy.
- And each line shows a country.
- The widths of the line also convey information, which I'm not going to get into here.
- What really jumps out in this graph is that red line at the top, the United States, where health care spending is about twice as much as it is in any of the other countries, while life expectancy is right in the middle of the pack.



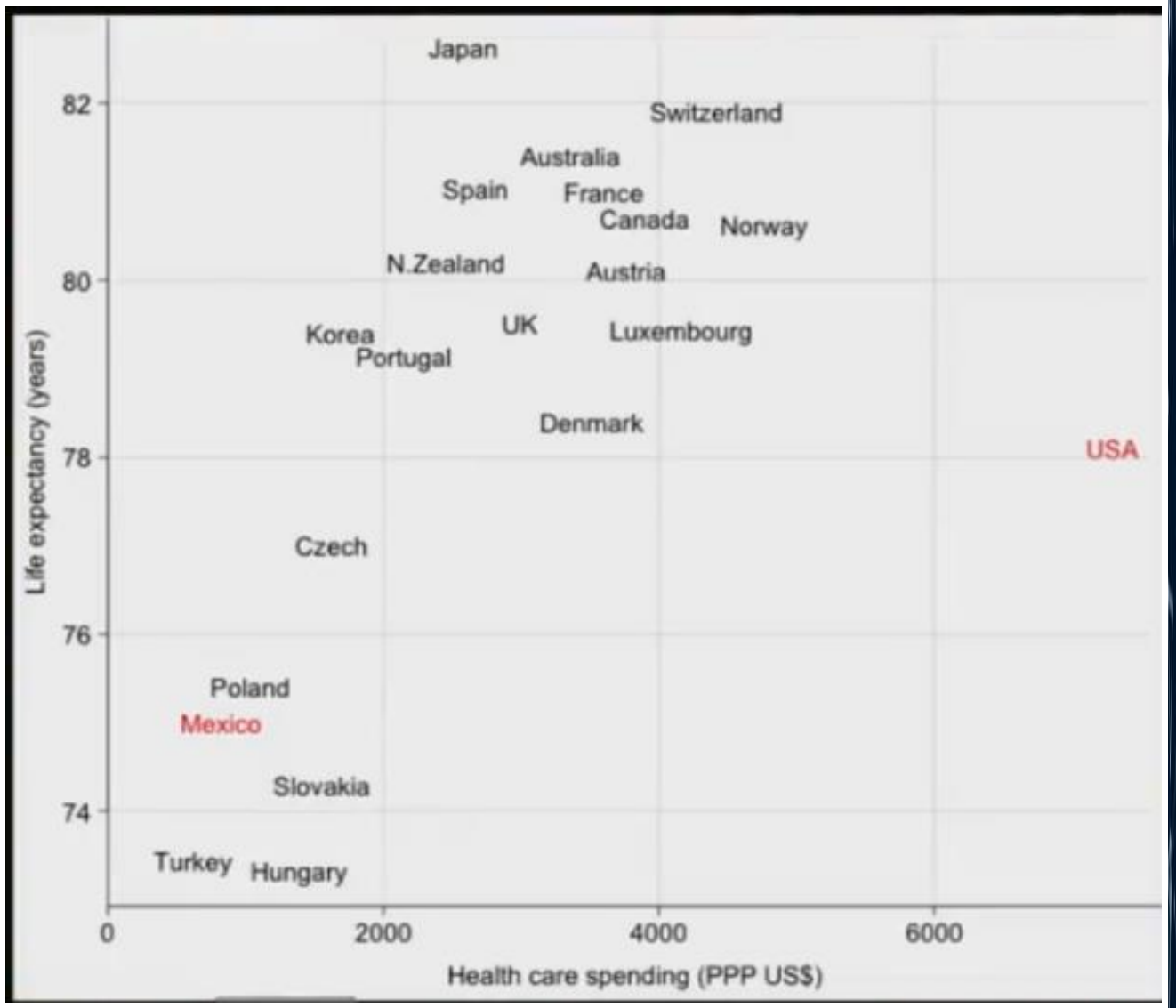
- So remember-- graphs as comparisons-- what's being compared here? The US to everywhere else.
- However, effective as this graph is at making its political point, from a statistical standpoint, we can do better.
- So now I'm going to re-express these data as a scatterplot.
- For every country, we're displaying two numbers, health care spending and life expectancy.



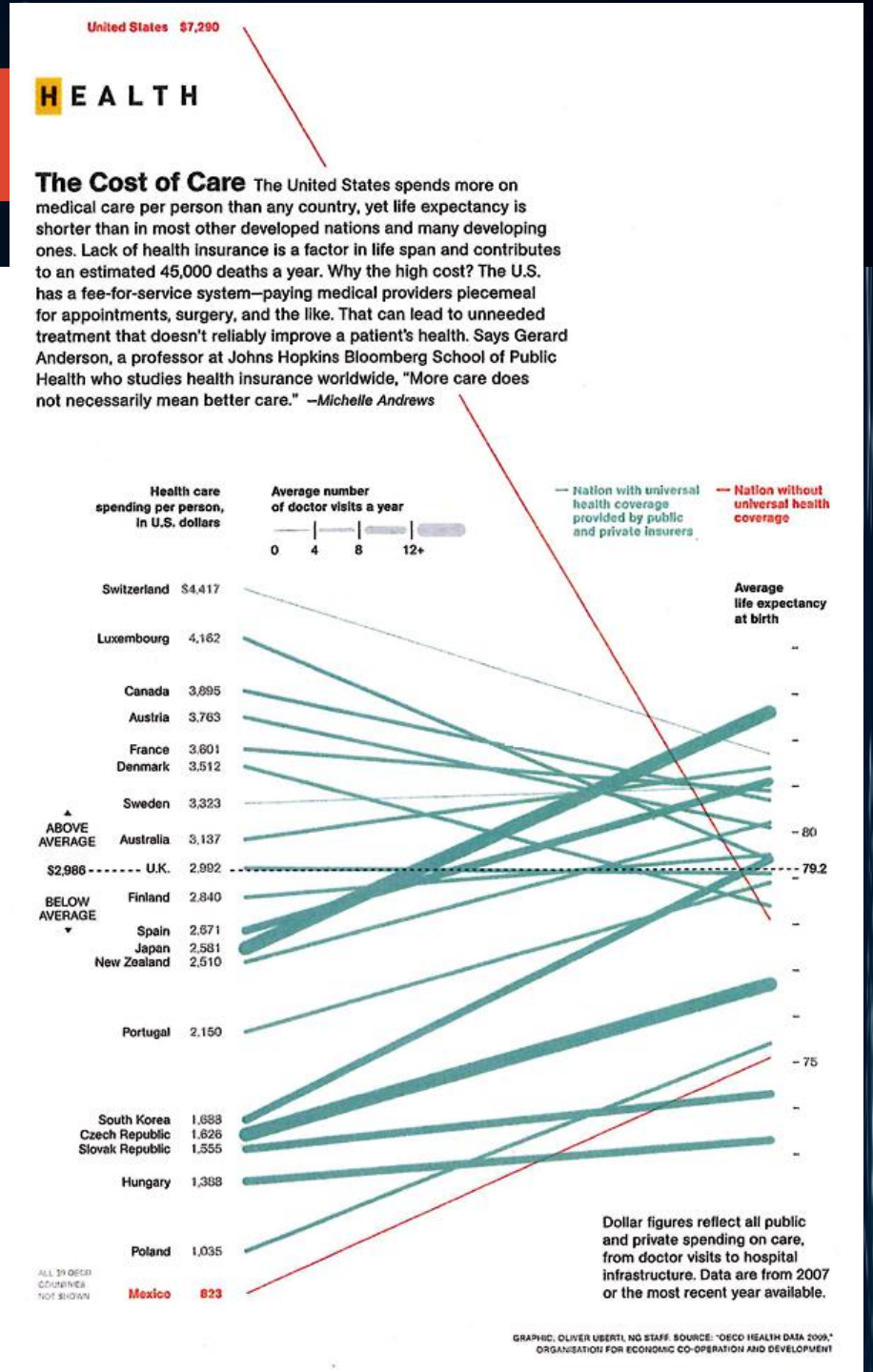
- So let's plot them that way.
- Again, the US is in red.
- Mexico is also in red. It was in red in the original graphic, at the bottom.
- Again, the US stands out.
- We're spending about twice as much as everybody else with similar life expectancy as these other countries.
- But we learn a lot more from the scatterplot.



- In particular, when you take away the US, you see a positive correlation between health care spending and life expectancy.
- In addition, we can easily identify the individual countries.
- We can see where Denmark is, we can see Japan with its highest life expectancy, we can see the lower income countries on the lower left that have slightly lower life expectancy, and of course, less health care spending.

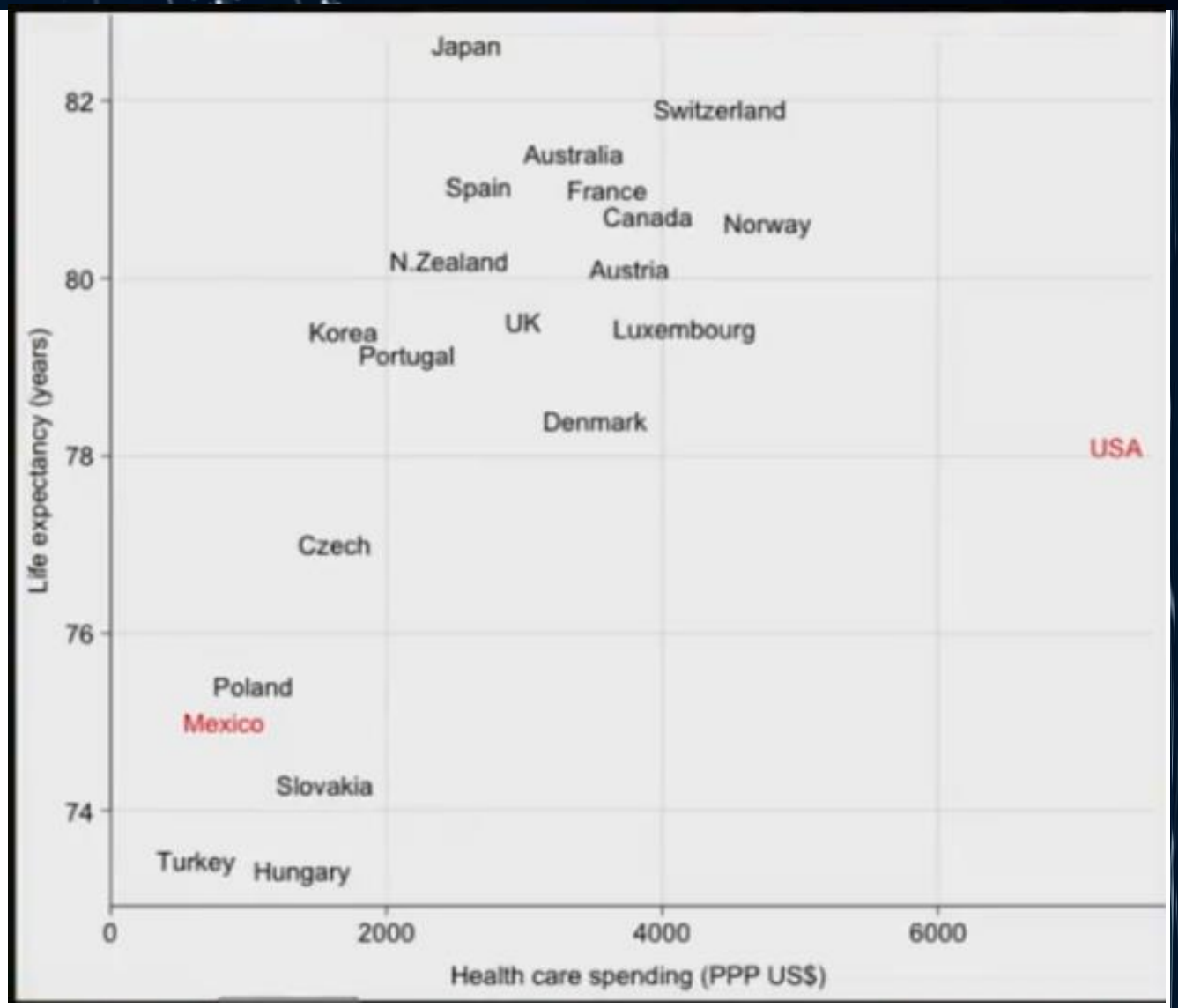


- If we go back to the original graph, we can't see that sort of detail.
- We can't make those comparisons.
- Yes, if you're careful, you can read off the data in each country.
- The data are all there.
- You can go to the country names on the left side of the graph and read off the numbers.
- But there's no way to see a general pattern.
- Yes, we see the US spends much more than other countries, but beyond that, we can't really see the general trend in the data.

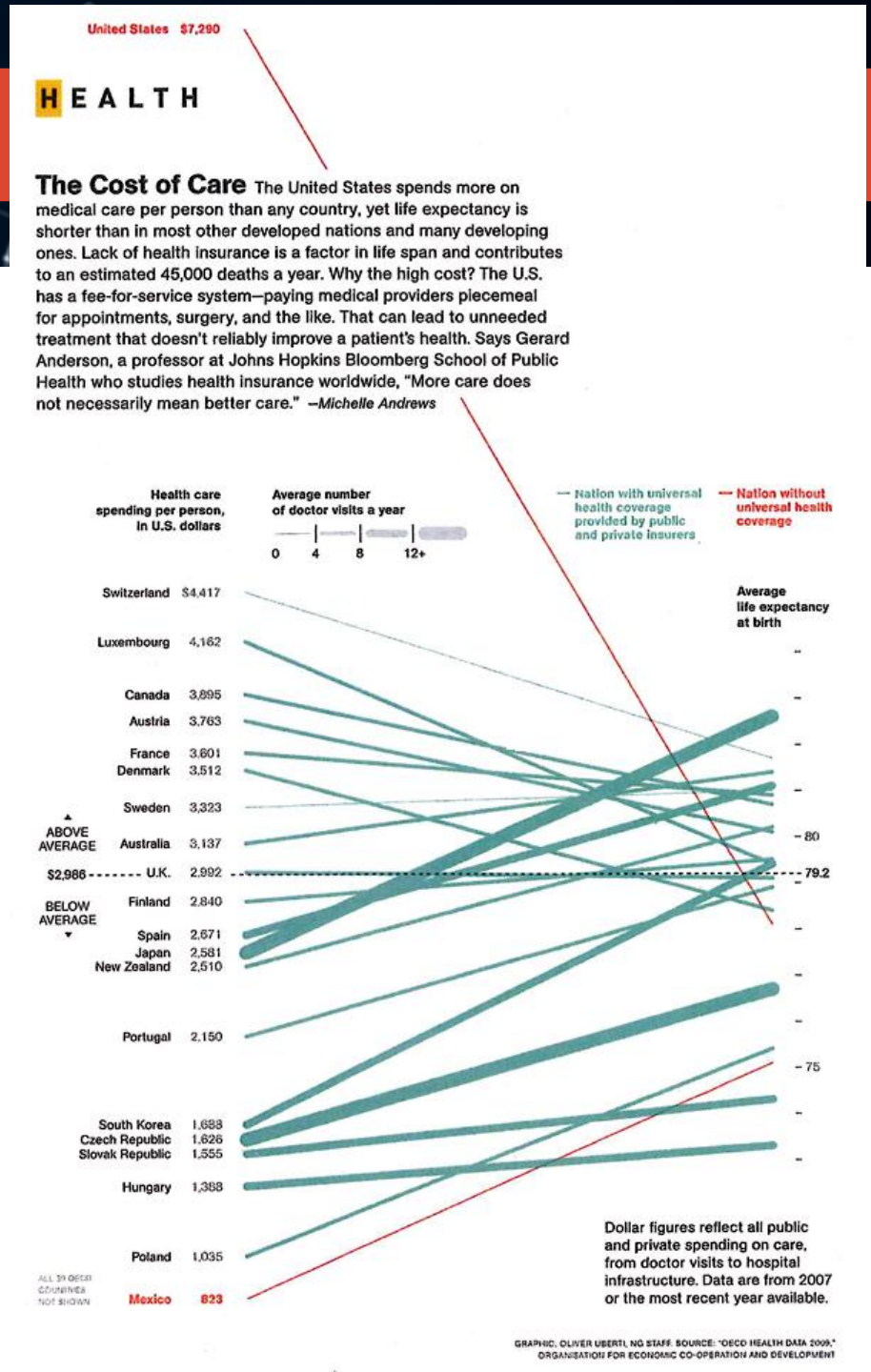


RESEARCH

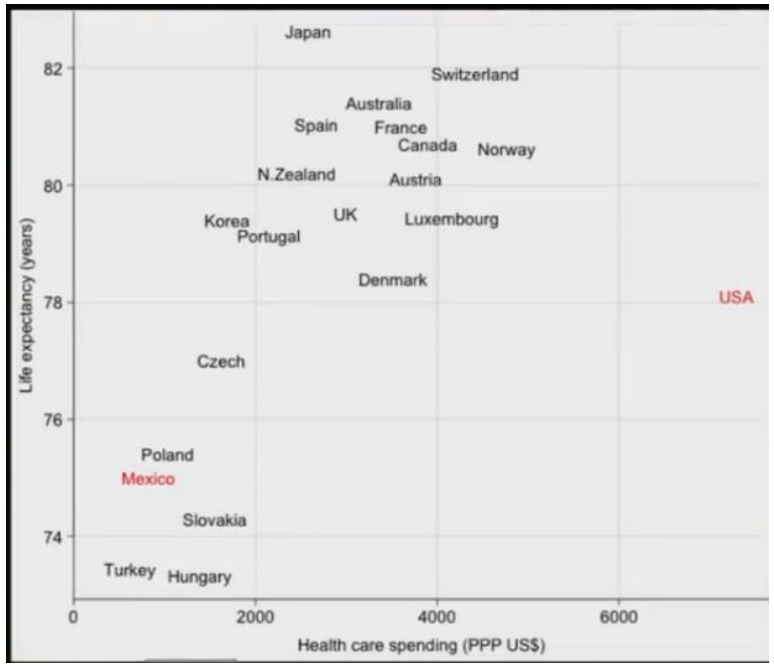
- Once we display it this way, we can see the positive correlation in the countries outside the US, which makes the US look special and different in a slightly different way.
- This is not to say that the original graph was bad or useless.

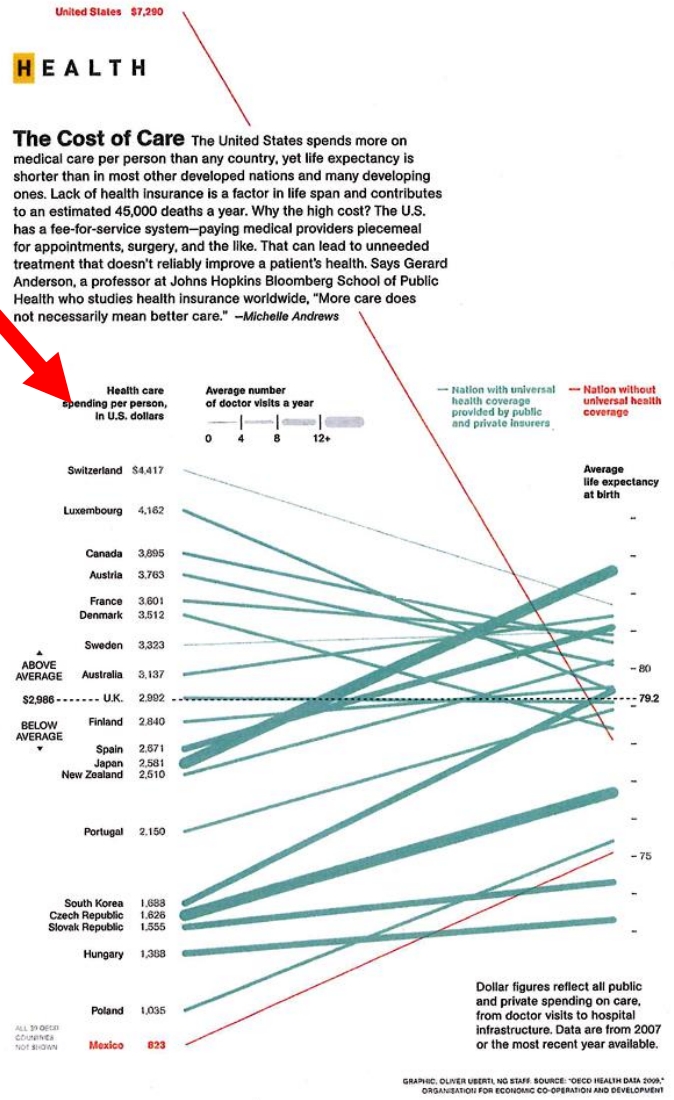


- The original graph was very attractive.
- It very dramatically showed the difference.
- There's something very striking about a red line that actually goes outside the border of the graph.
- And I don't criticize the people for making this graph.

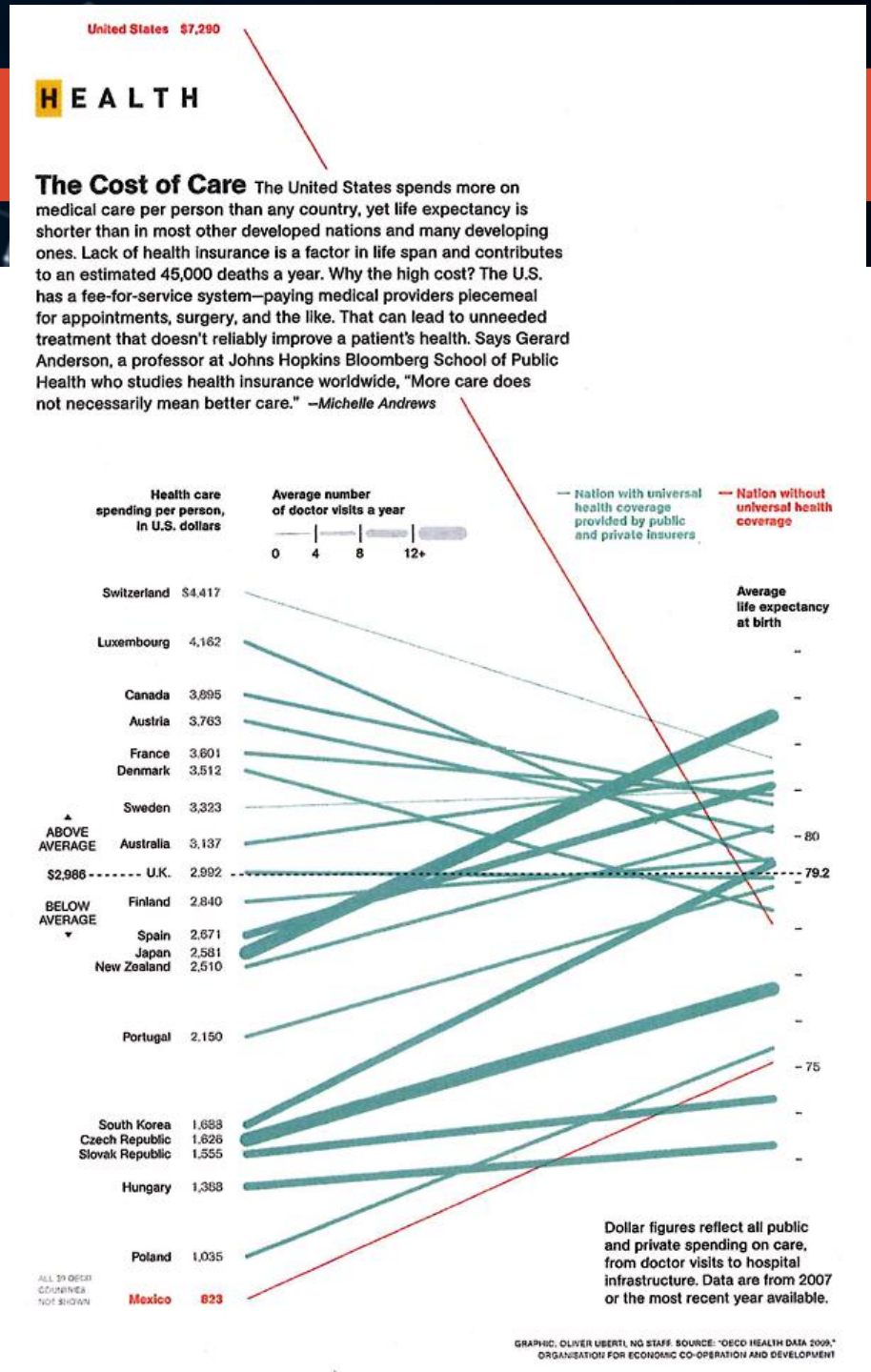


- What I think would be ideal is, if you have a graph like this on the web that gathers people's attention, you can click on that graph and get this graph, which has more information.

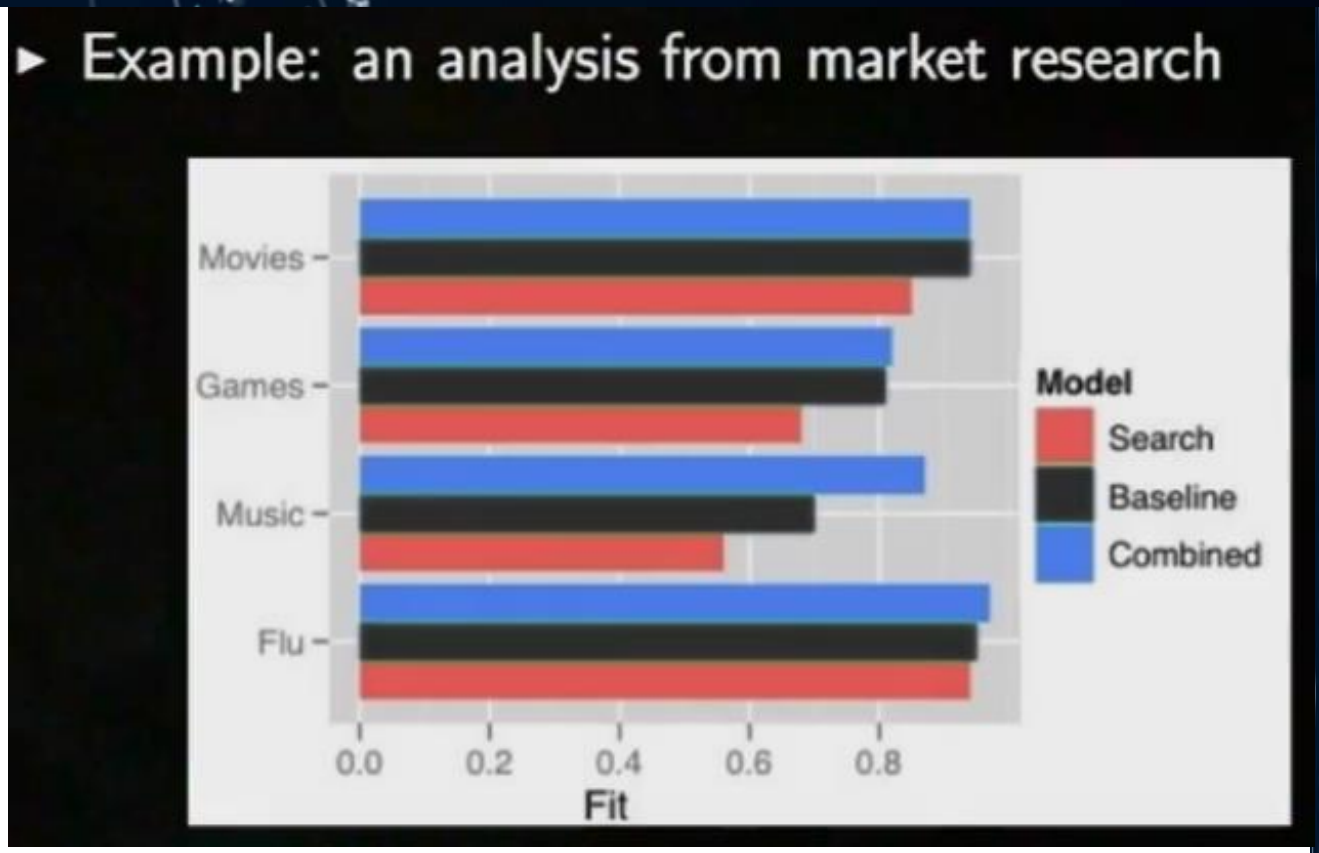




- And then maybe click again and get to a spreadsheet with the data so people can analyze it themselves.
- For our next example, I'm going to focus on how re-expression of data allows certain comparisons to be made more apparent.

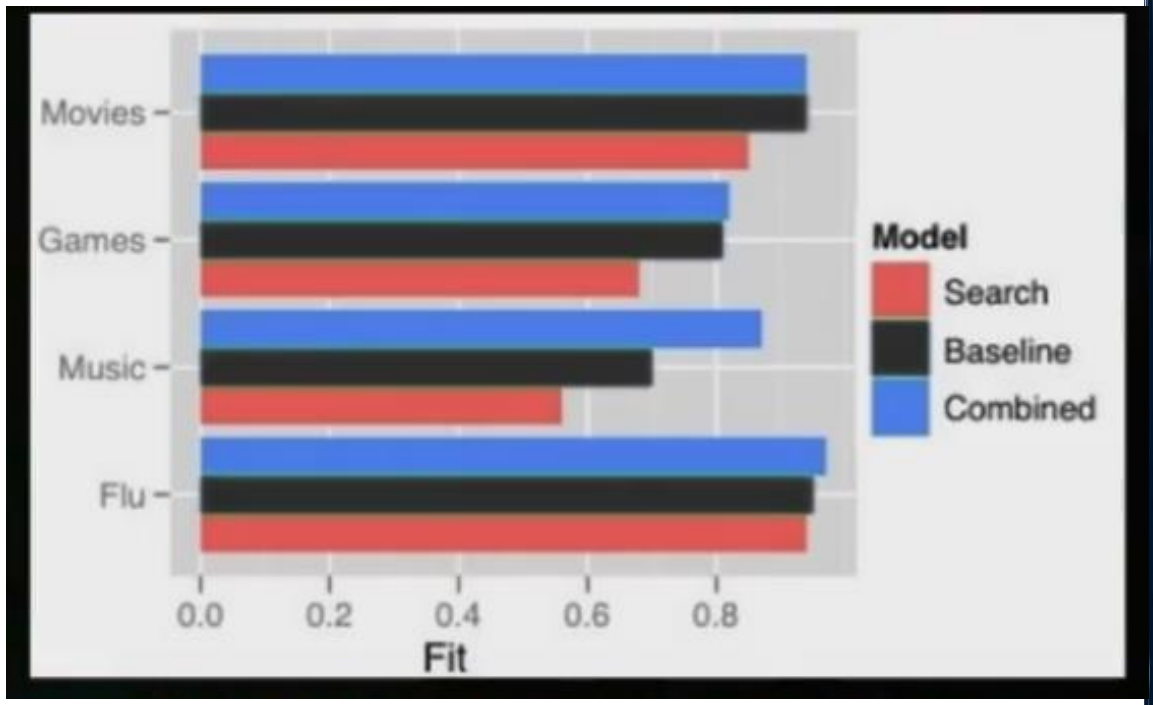


- The example comes from market research.
- My colleagues fit three different models to four different data sets, and then they made this graph to display the predictive performance of the models in the different data sets.
- And we can see, what does this allow us to do?



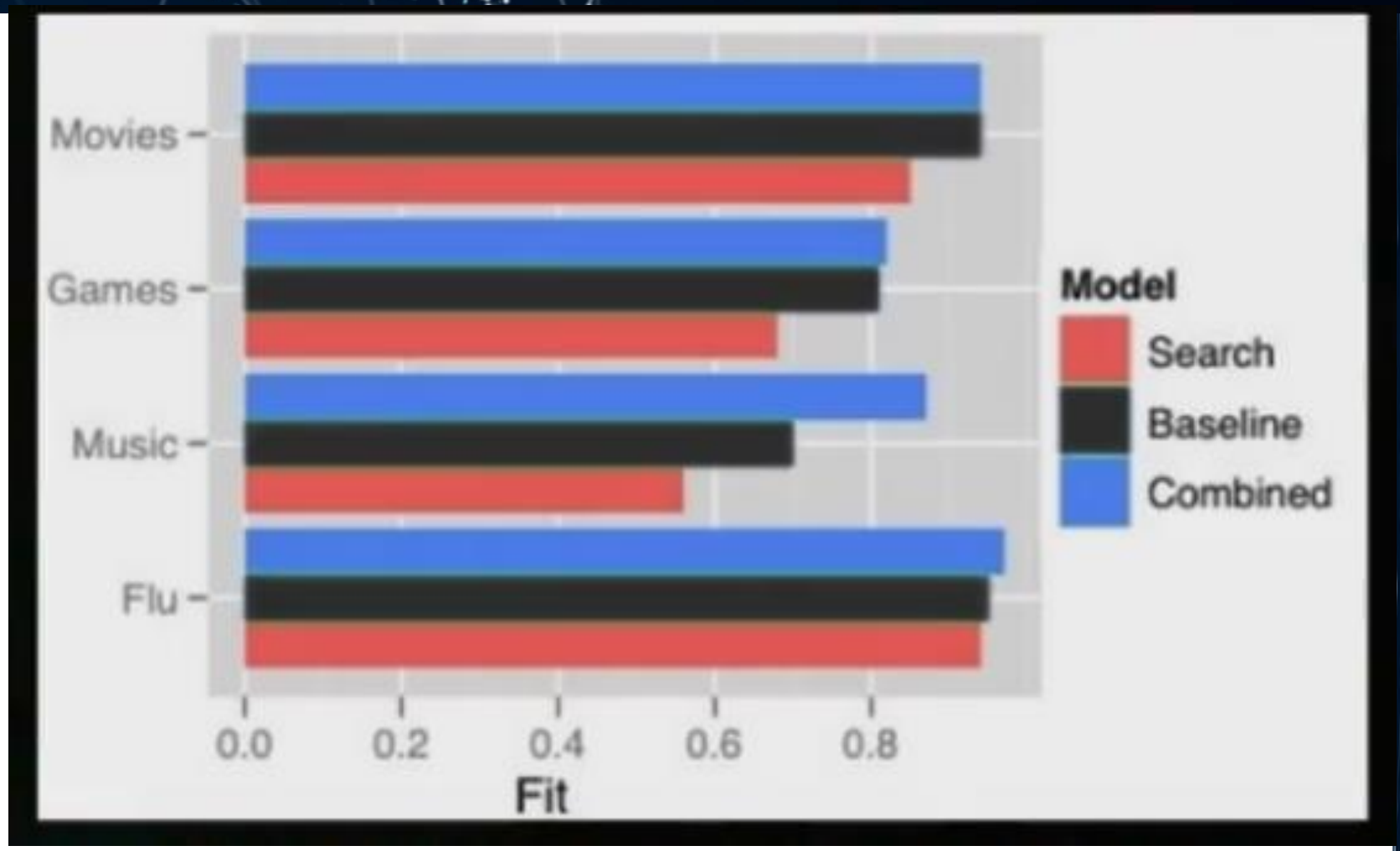
- We can compare the three models on each data set.

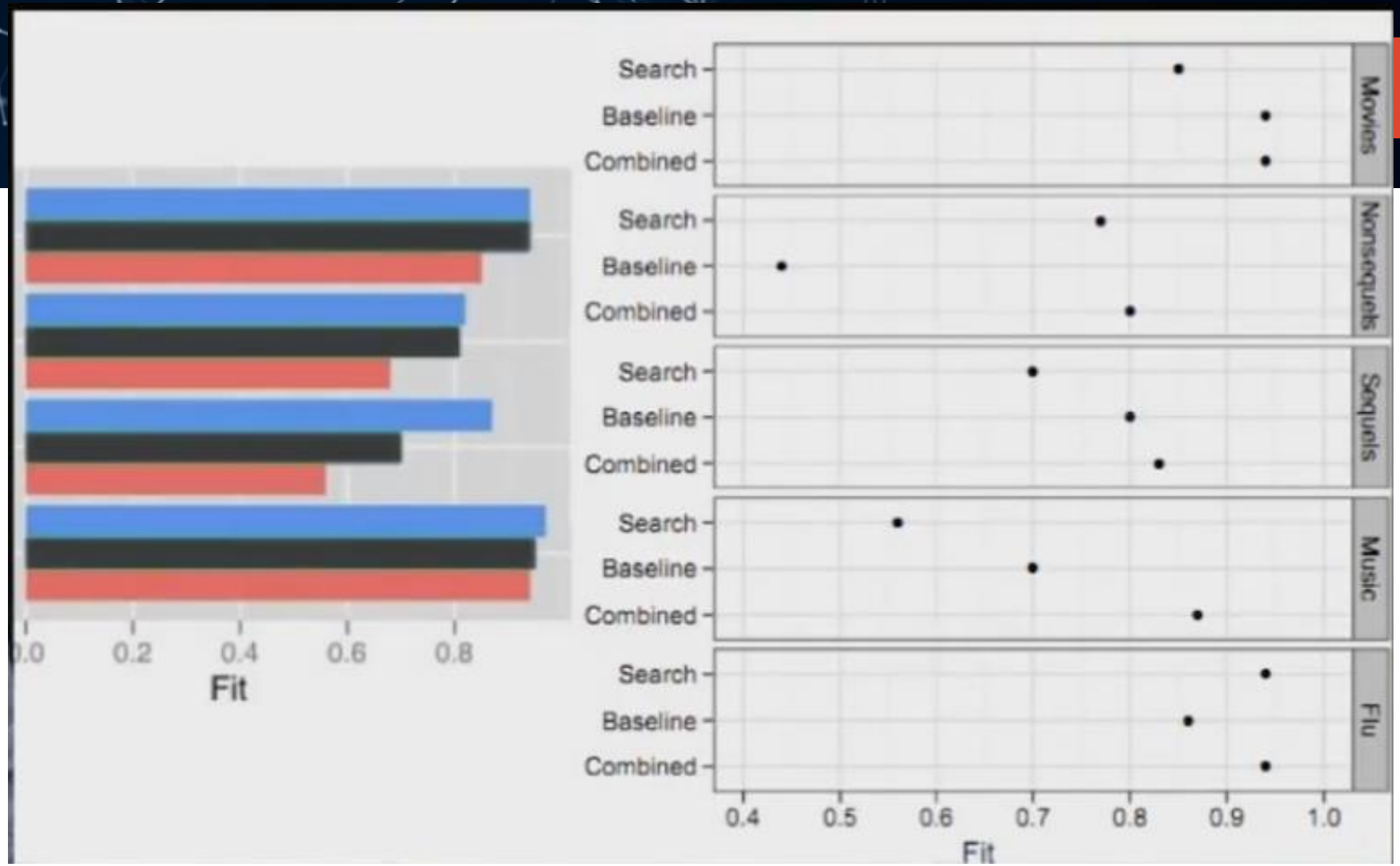
- So for each data set, we can see which model performed better.
- And it looks like, in general, the blue model, the combined model, had the highest fit.
- It performed better.
- Which sort of makes sense.
- I don't actually know the details, but it sounds like a combined model should do better than others.



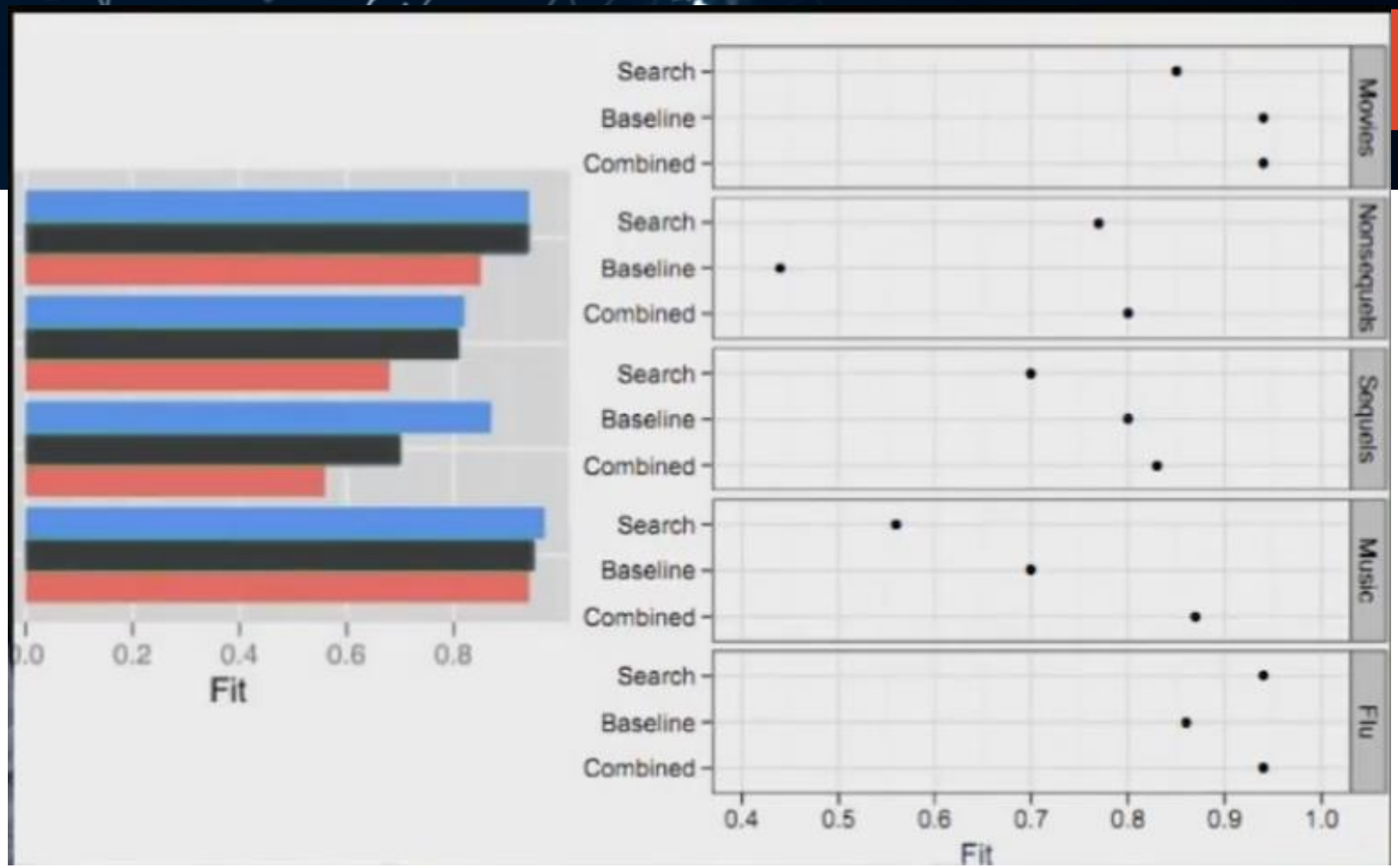
So this graph very effectively shows the comparison within each group.

- However, it doesn't allow us to compare the four different sets very clearly.
- I expressed this concern to my colleagues, and they replotted the data.

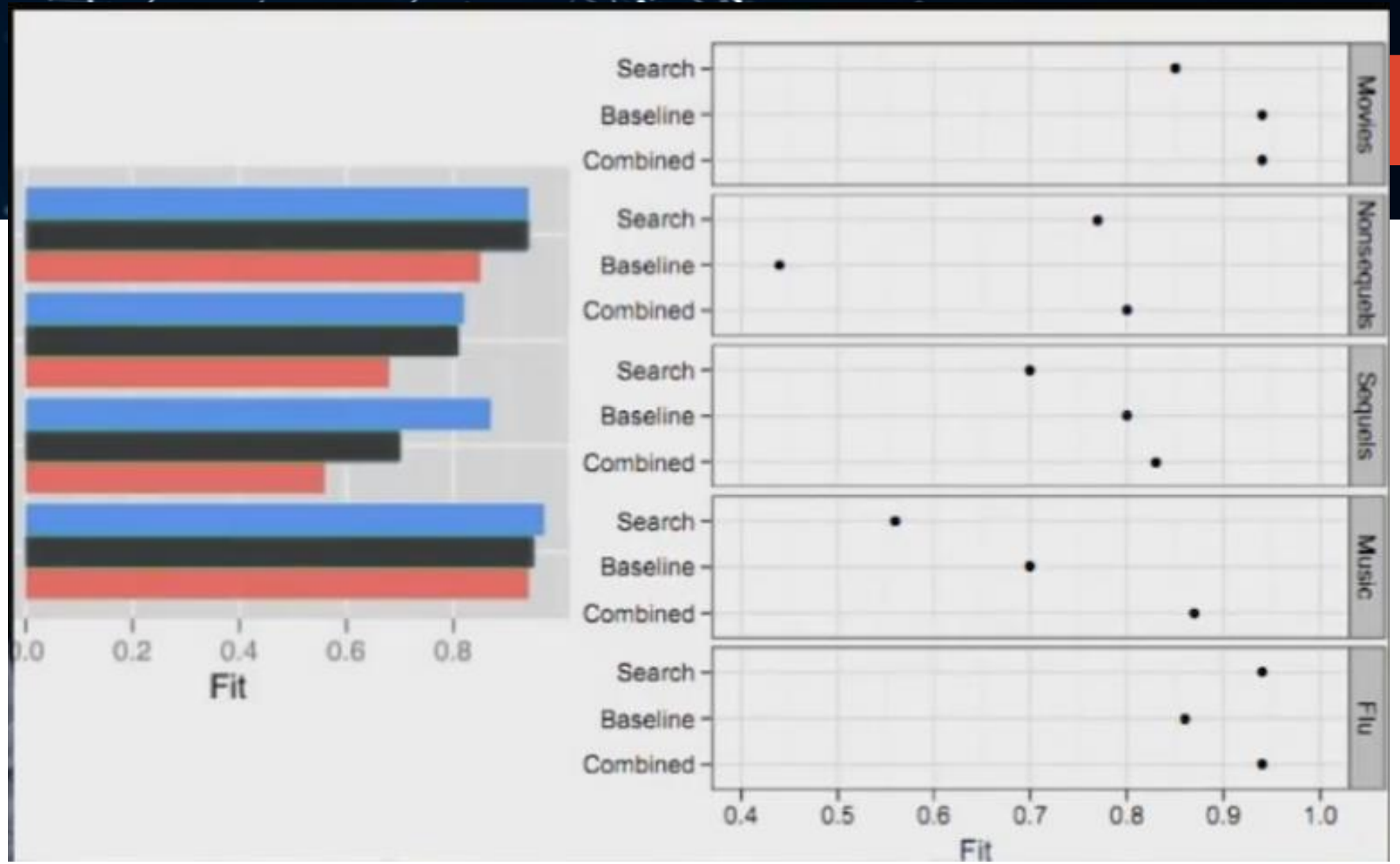




- The original graph is on the left and the replot (re-expression) is on the right.

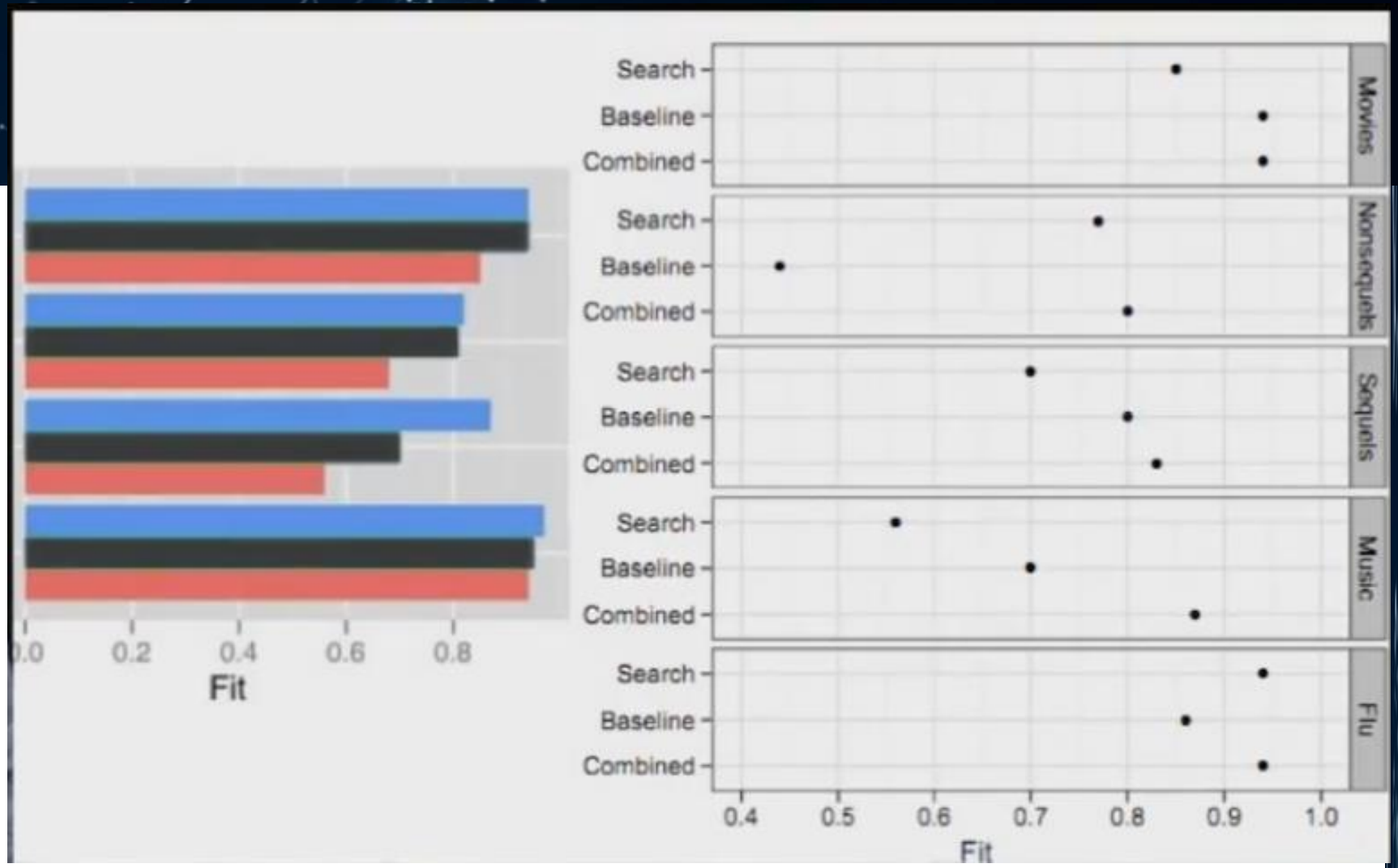


- They also added a fifth data when doing this.
- They were hoping that the re-expression was what I was looking for, that I would be able to make the comparisons that I wanted.



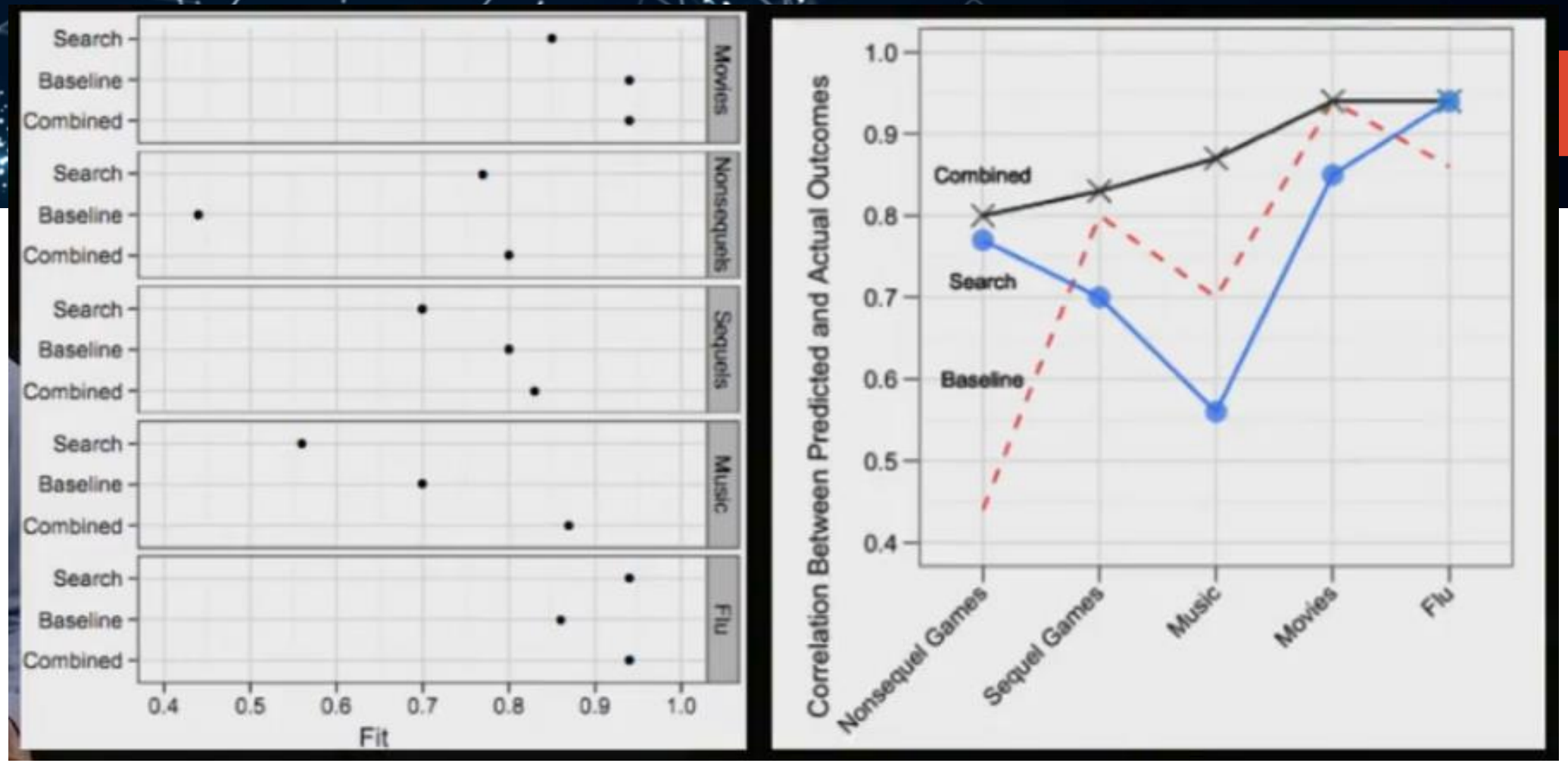
- But actually, I have the same trouble with the graph on the right as I do with the one on the left, which is that I can make comparisons within each of the models but it's very hard for me to make comparisons between models.

- I find that it's hard to get the big picture of where the combined model is doing particularly well compared to the search and baseline.

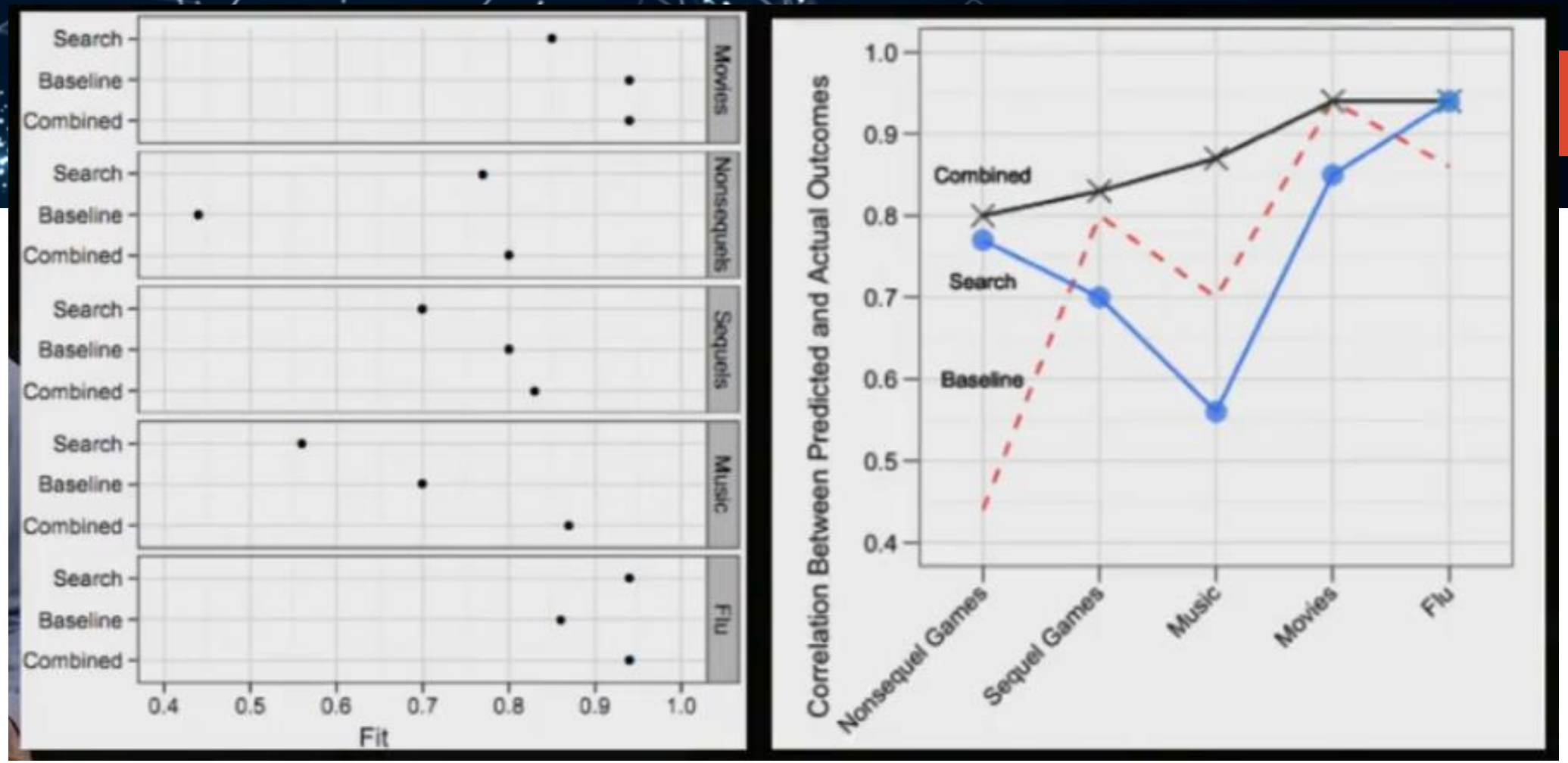


- But this was interesting to me, because I had explained my concerns to my colleagues, and they made a new graph.

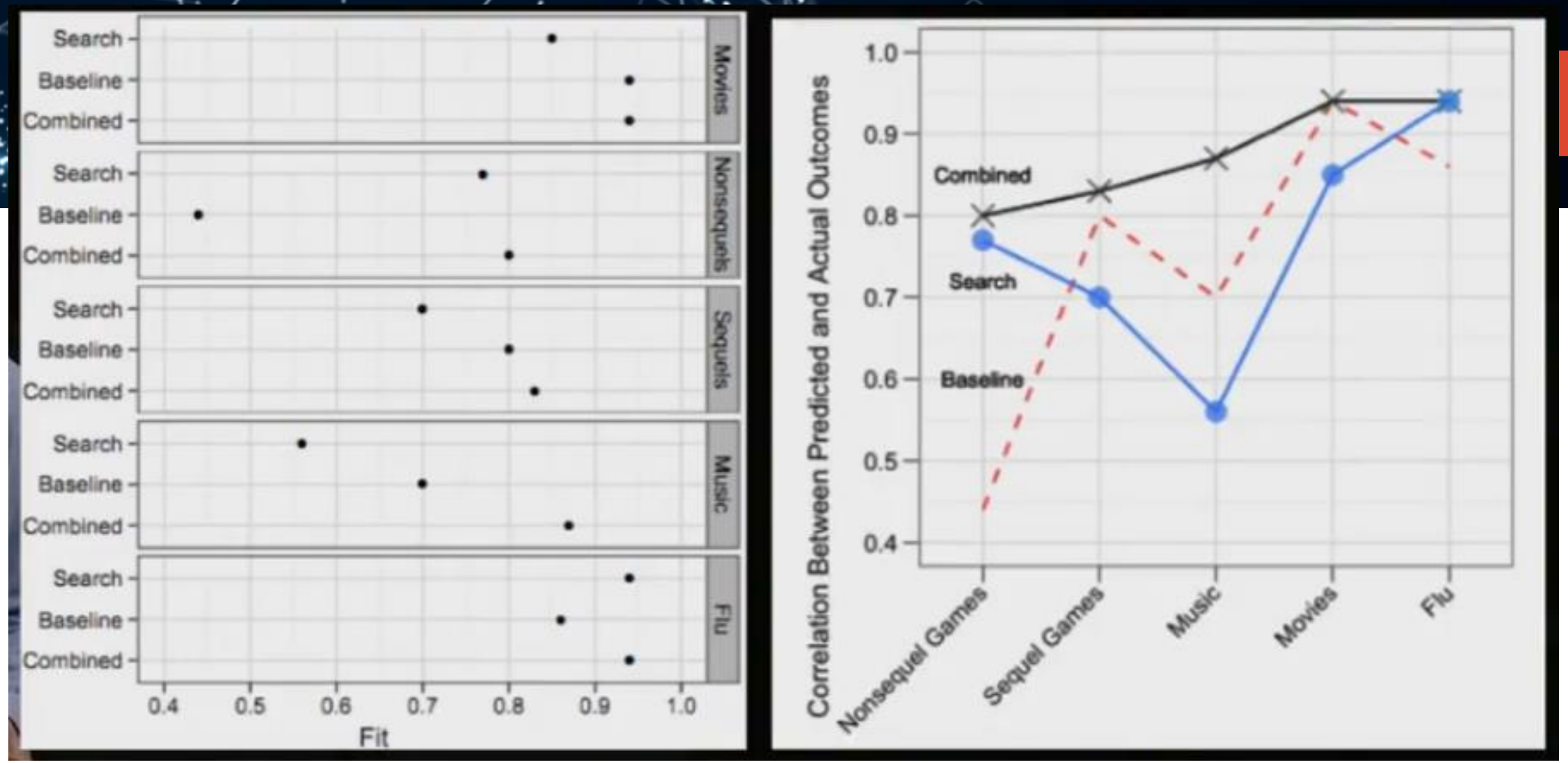
RESEARCH



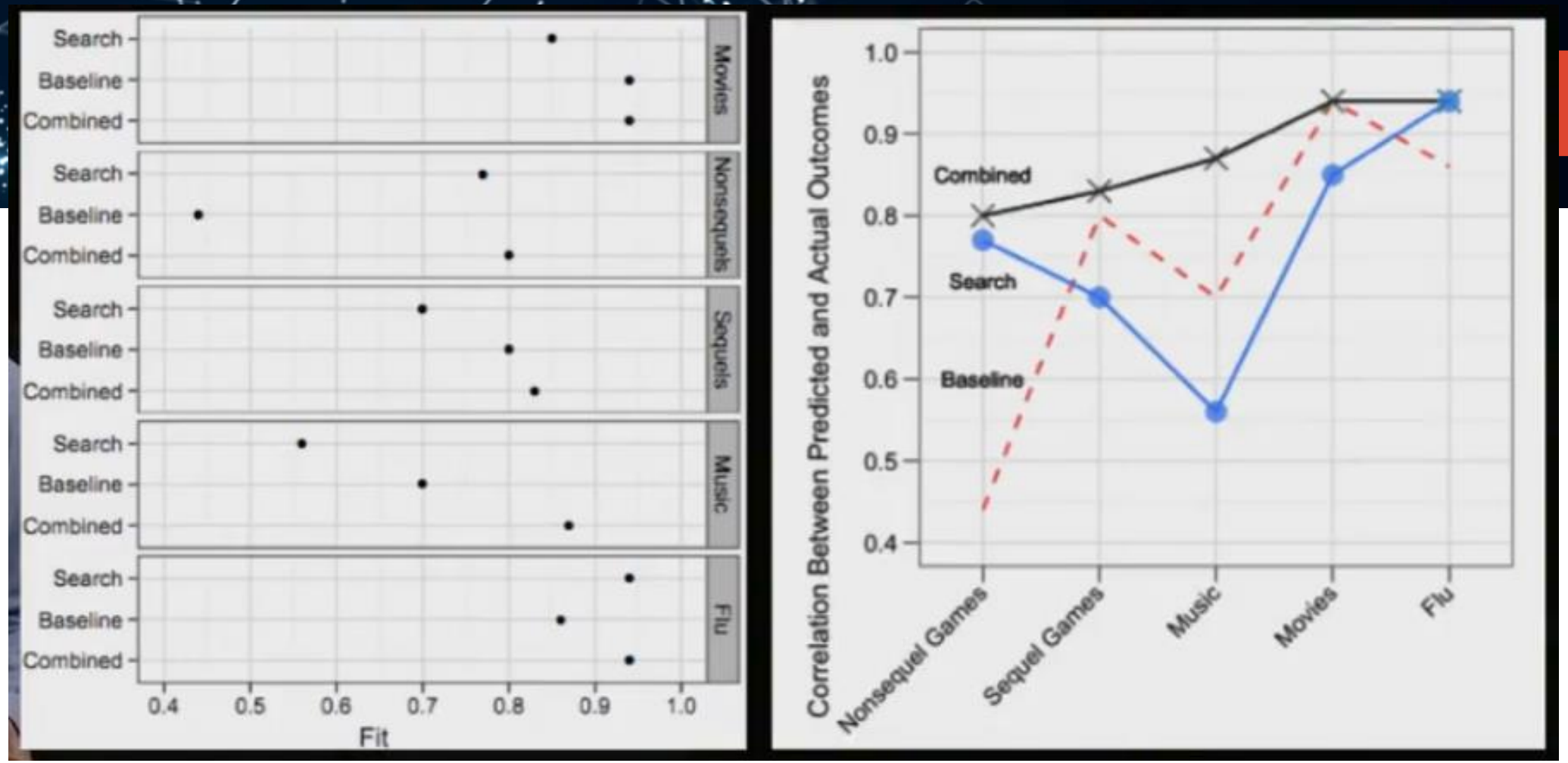
- But it still wasn't what I was looking for, which made me feel that I didn't have a good enough language to express my concerns.
- In any case, they redrew the data one more time in a way that it turned out I liked a lot, and that's the graph on the right.



- So again, they're comparing three models, search, baseline, and the combined model on five data sets.
- The differences is, in this graph, I can not only make the comparison for each data set, I can also see the big picture.



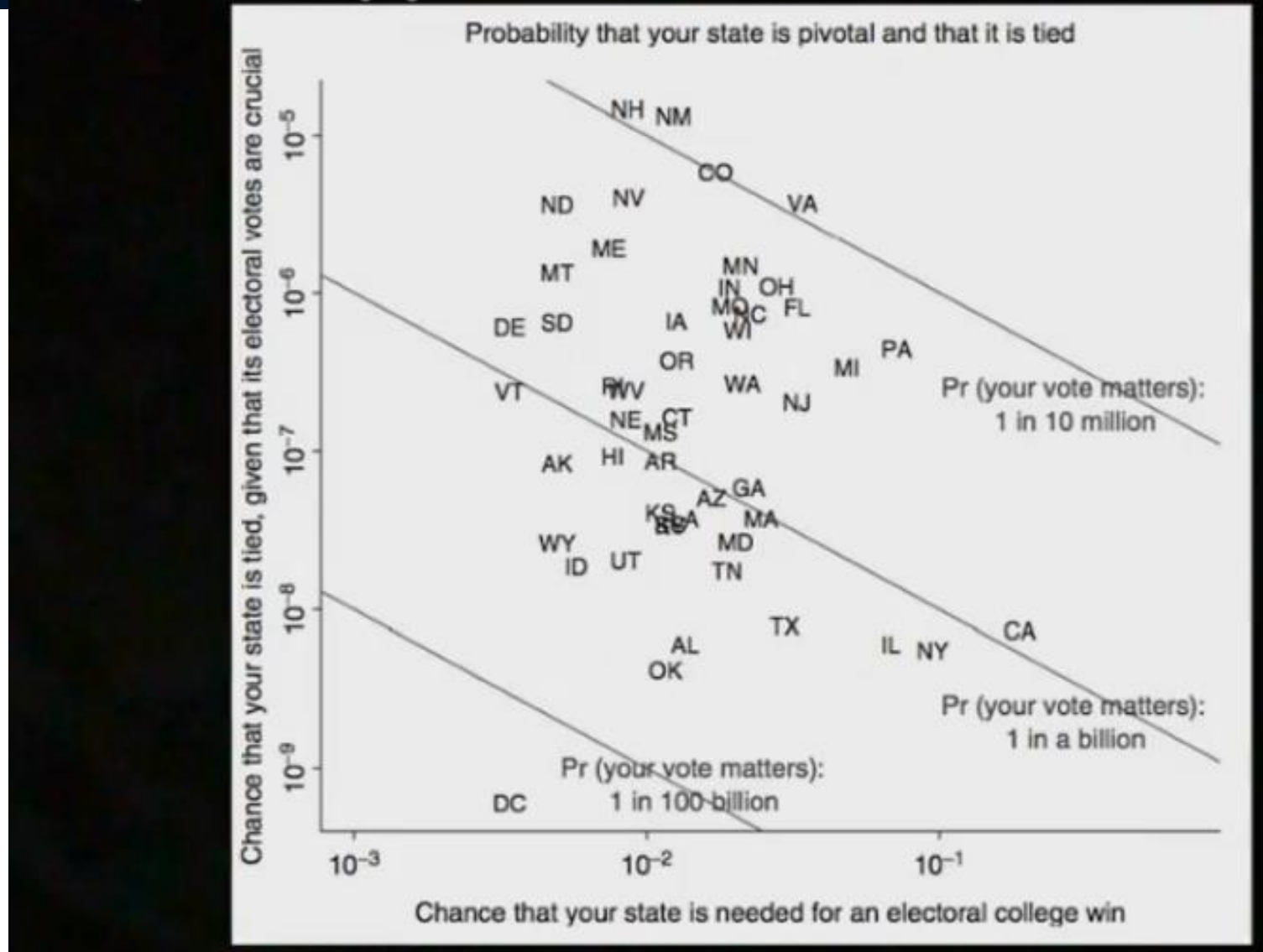
- I can see which sorts of data the search model is doing particularly poorly in and which data the baseline model is doing particularly poorly.
- It gives me a bird's eye view.



- Now, what's interesting to me is that's the same data.
- It takes the same amount of space on the page.
- And yet, the comparisons are much clearer.

Displaying estimates from a model

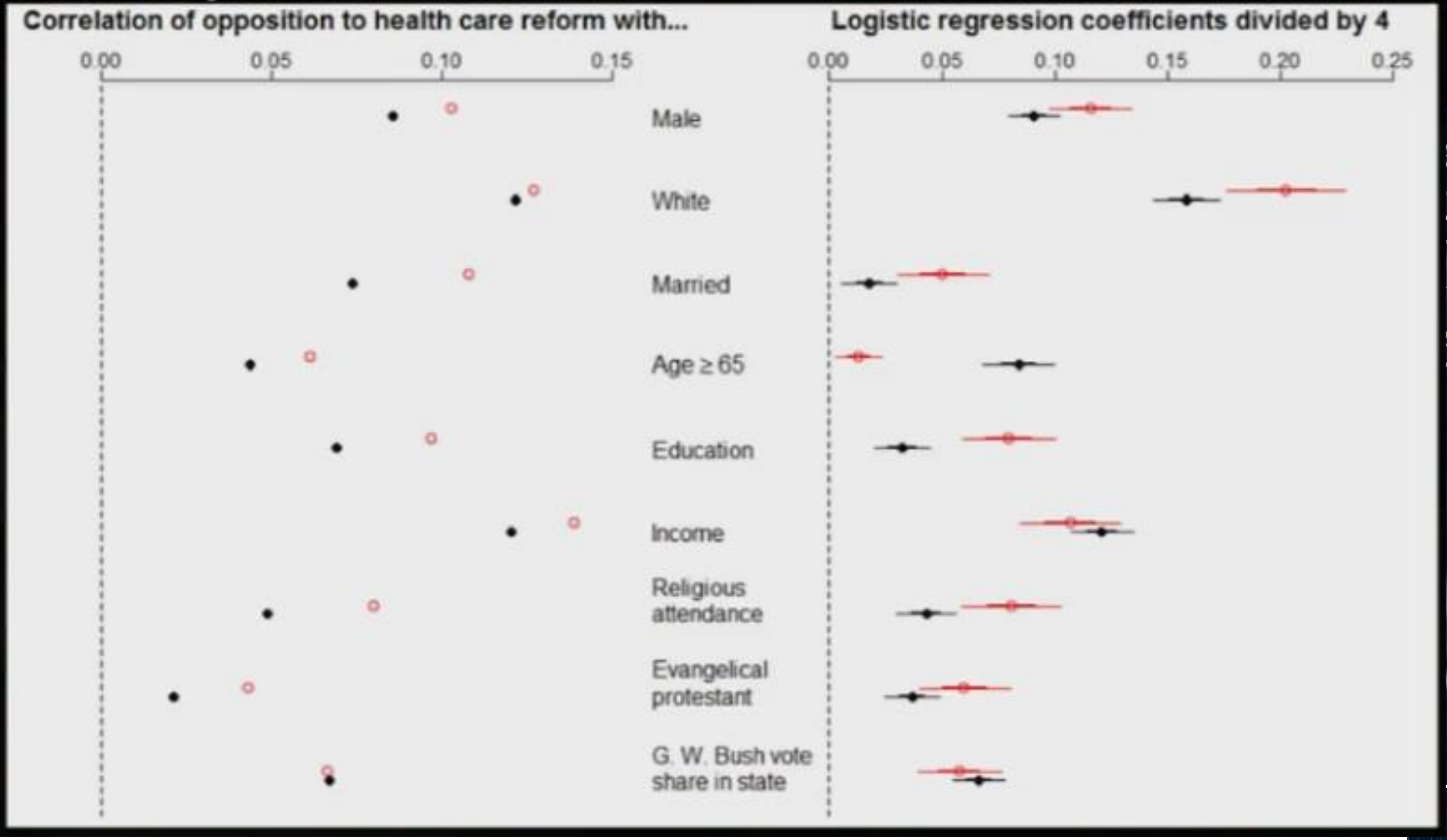
- Visualization is not just for raw data.
- We can also use graphics and data visualization tools to understand models that we fit.
- And here's an example.



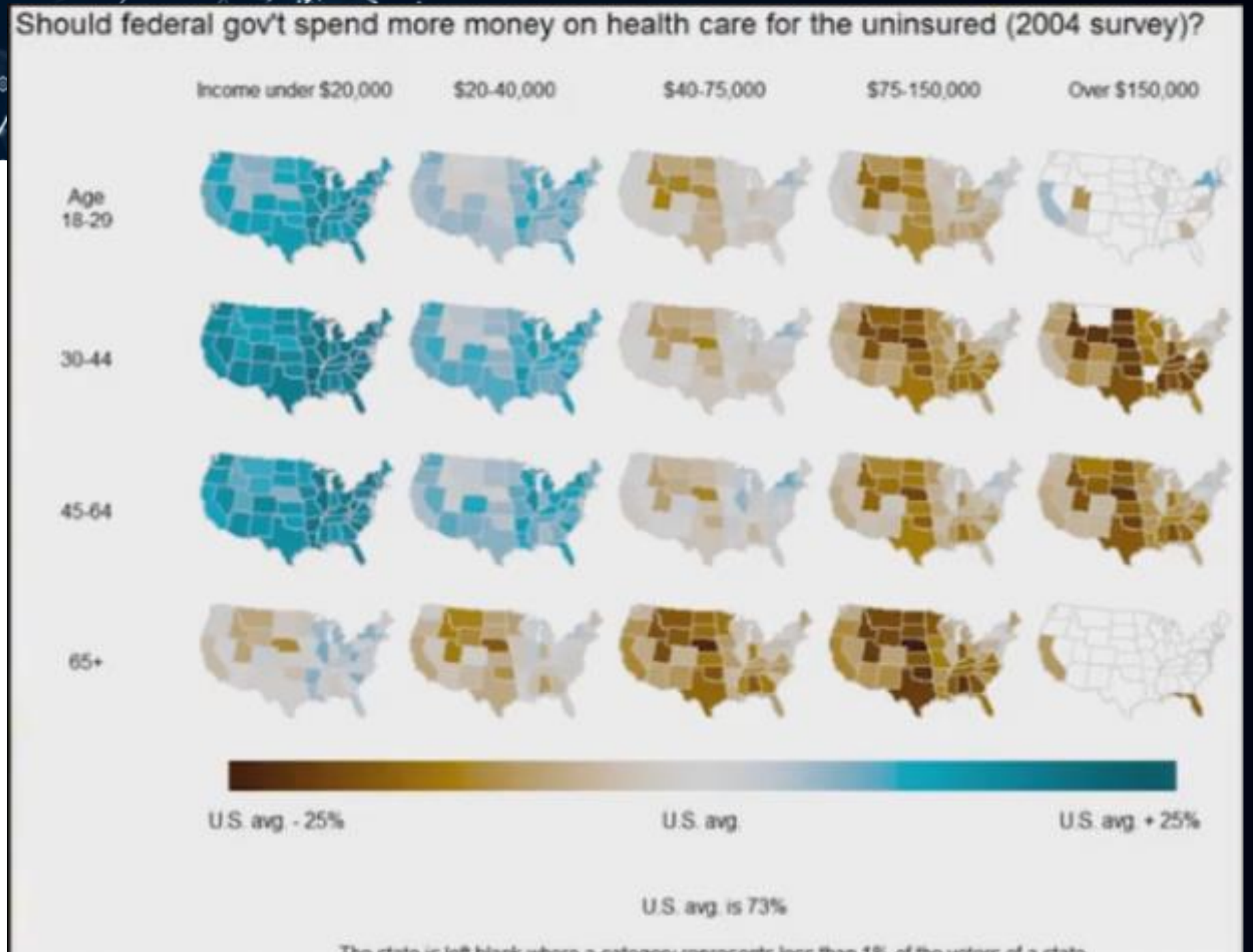
Graphs Are Comparisons

- I'll conclude our discussion of comparisons with two displays analysis of public opinion on health care reform based on survey data from a few years ago.
- We had data on a bunch of survey respondents, and the first thing we did was fit correlations and regressions.
- The graph on the left shows the correlation between opposition to health care reform and a bunch of predictors-- whether respondent is a man/white/married/ older/more educated/income/religious attendance/an evangelical Protestant, and the Republican support in your state that you lived in.

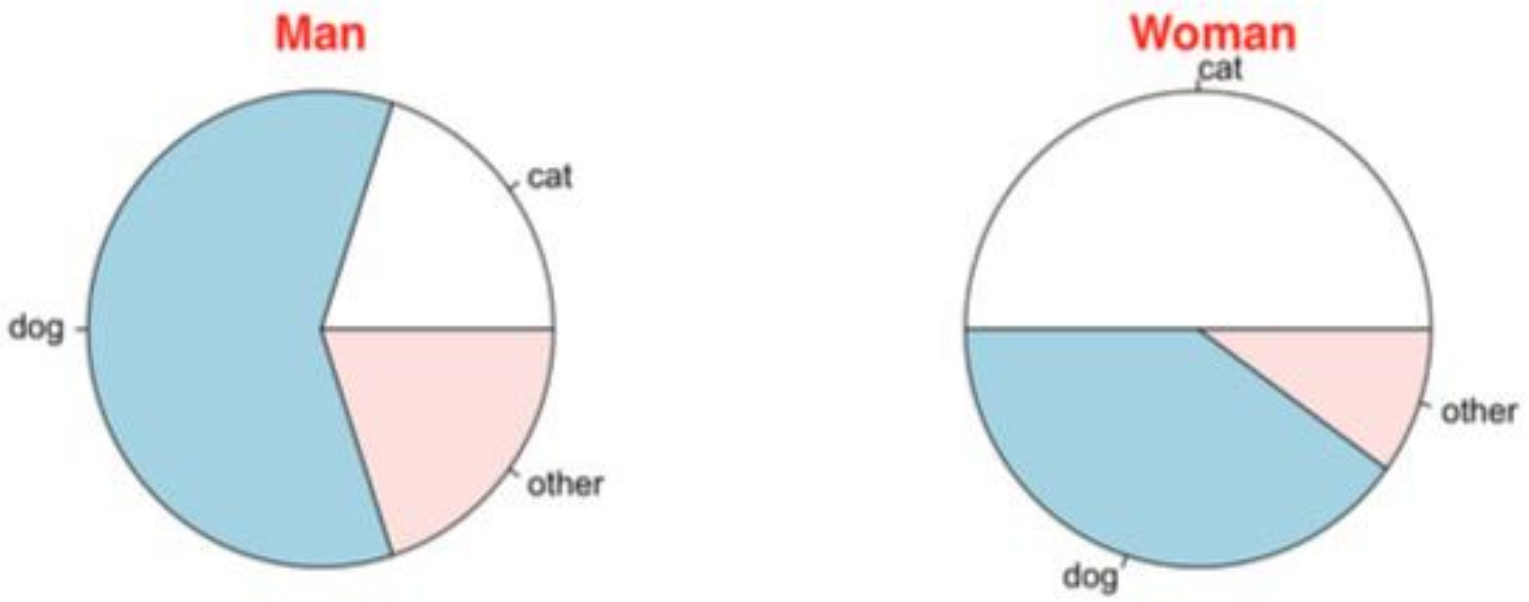
Public opinion on health care reform



- We can also break up the data by geography.
- This grid of maps shows our estimate of support for health care reform as a function of age, income, and what state you live in.
- These are not graphs of raw data.



- In summary, graphs or comparisons, whether you're making your own graph or reading a graph that somebody else makes, you should be aware of what comparisons are facilitated by the particular plotting tool that's been used and what comparisons are harder to see.



Pie chart

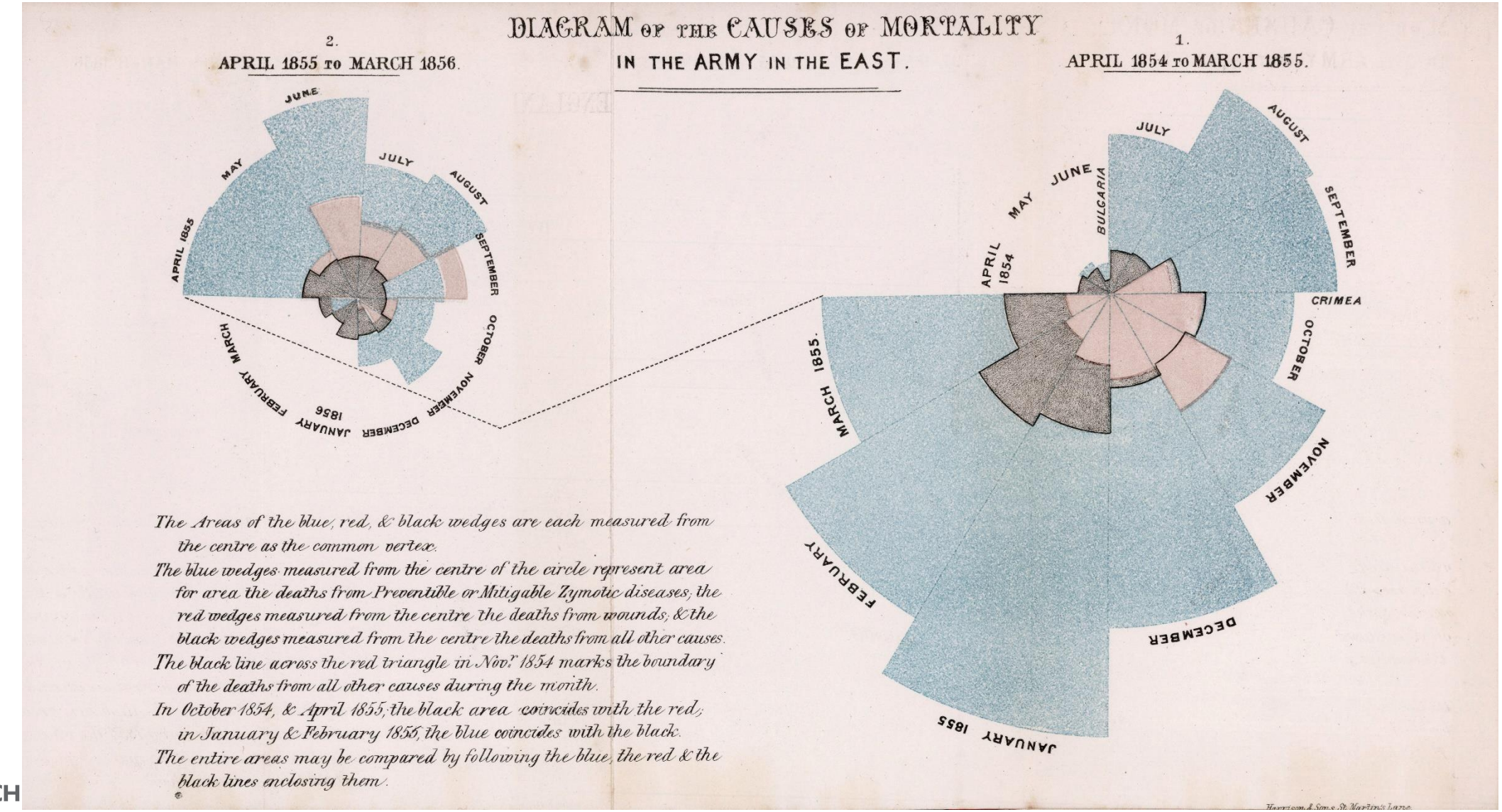
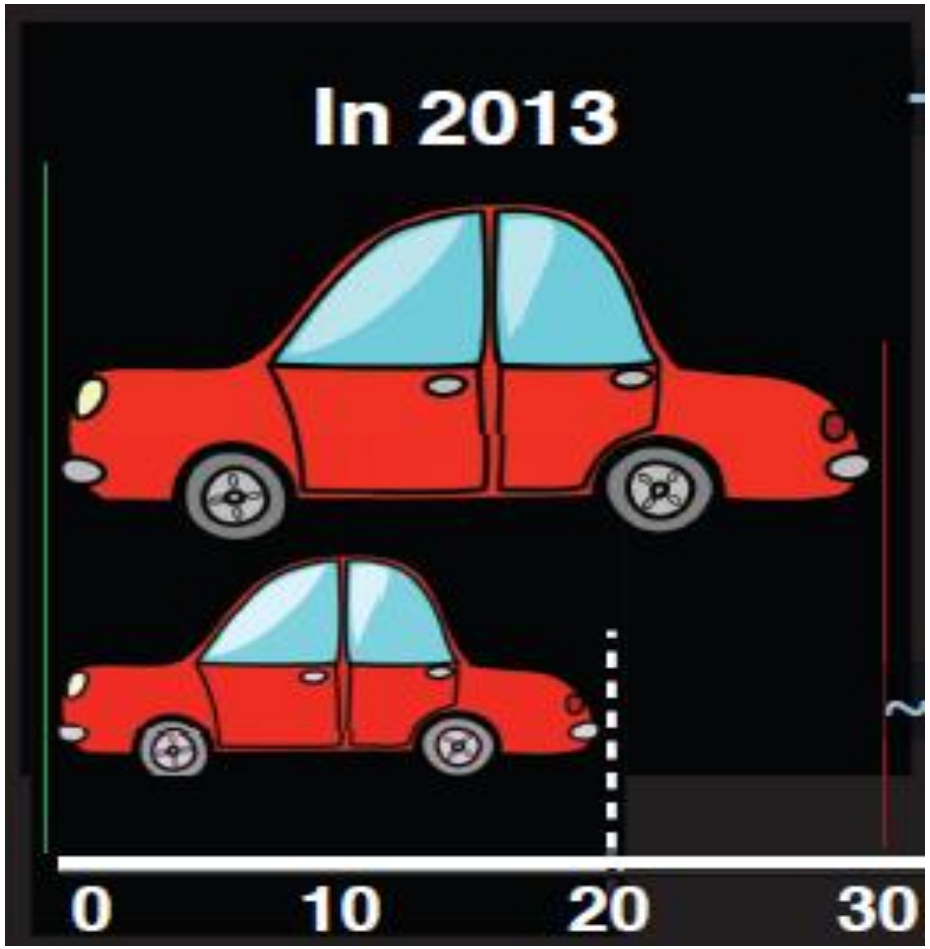


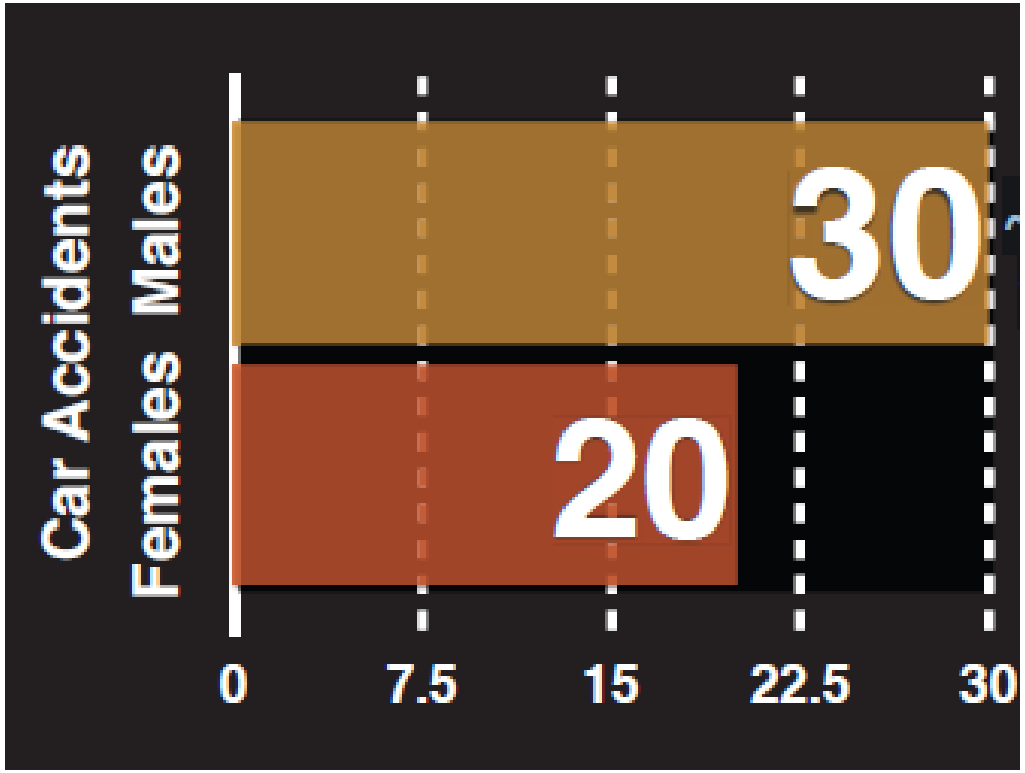
DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST.
2.
APRIL 1855 TO MARCH 1856.
1.
APRIL 1854 TO MARCH 1855.
JUNE
JULY
AUGUST
SEPTEMBER
OCTOBER
NOVEMBER
DECEMBER
JANUARY 1856
FEBRUARY
MARCH
APRIL 1855
MAY
BULGARIA
CRIMEA
APRIL 1854
MARCH 1855.
JANUARY 1855
The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.
The black line across the red triangle in Nov.r 1854 marks the boundary of the deaths from all other causes during the month.
In October 1854, & April 1855, the black area coincides with the red; in January & February 1855, the blue coincides with the black.
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

Area principle



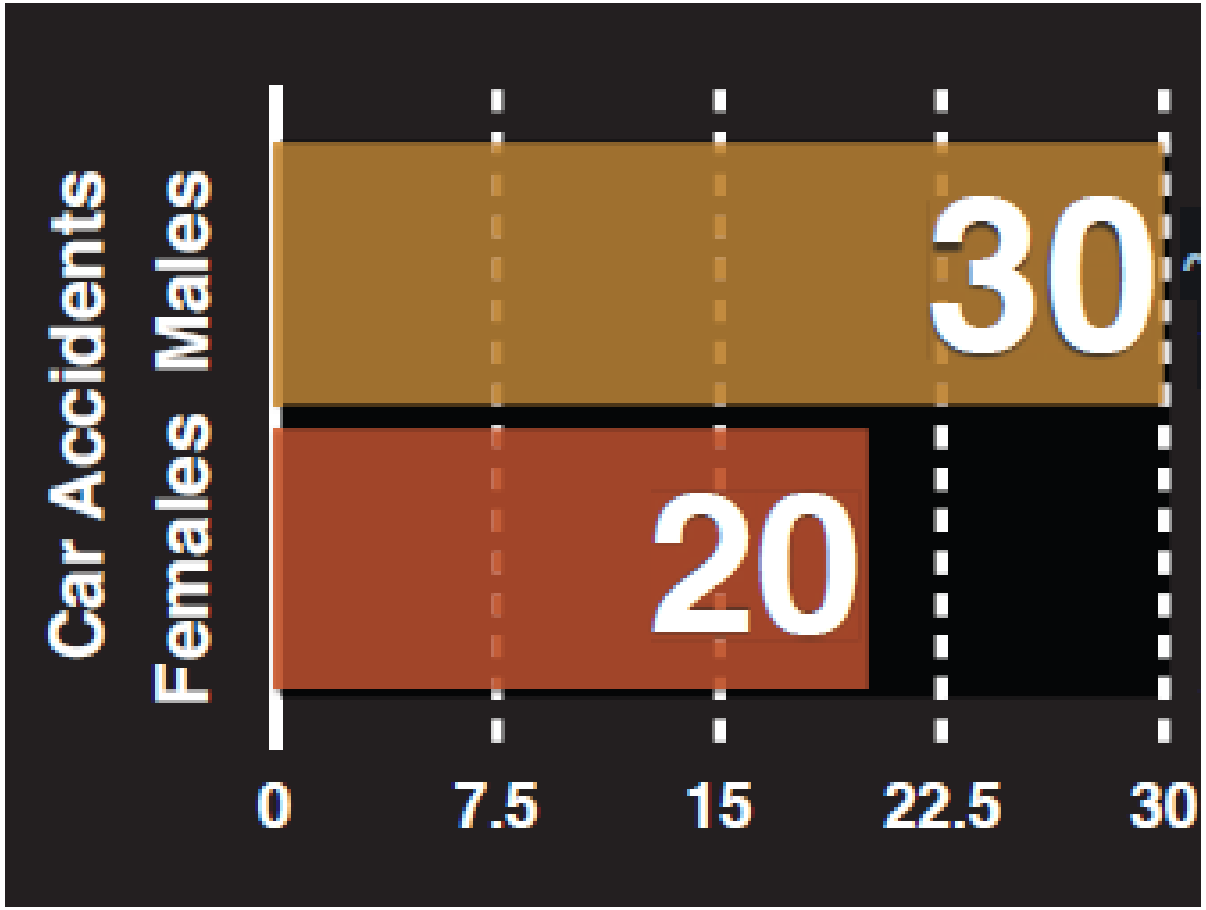
- The area principle
- — the size of the area correlates with the data summaries.
- ~30% of accidental deaths of males were due to automobile accidents.
- ~20% of accidental deaths of females were due to automobile accidents.

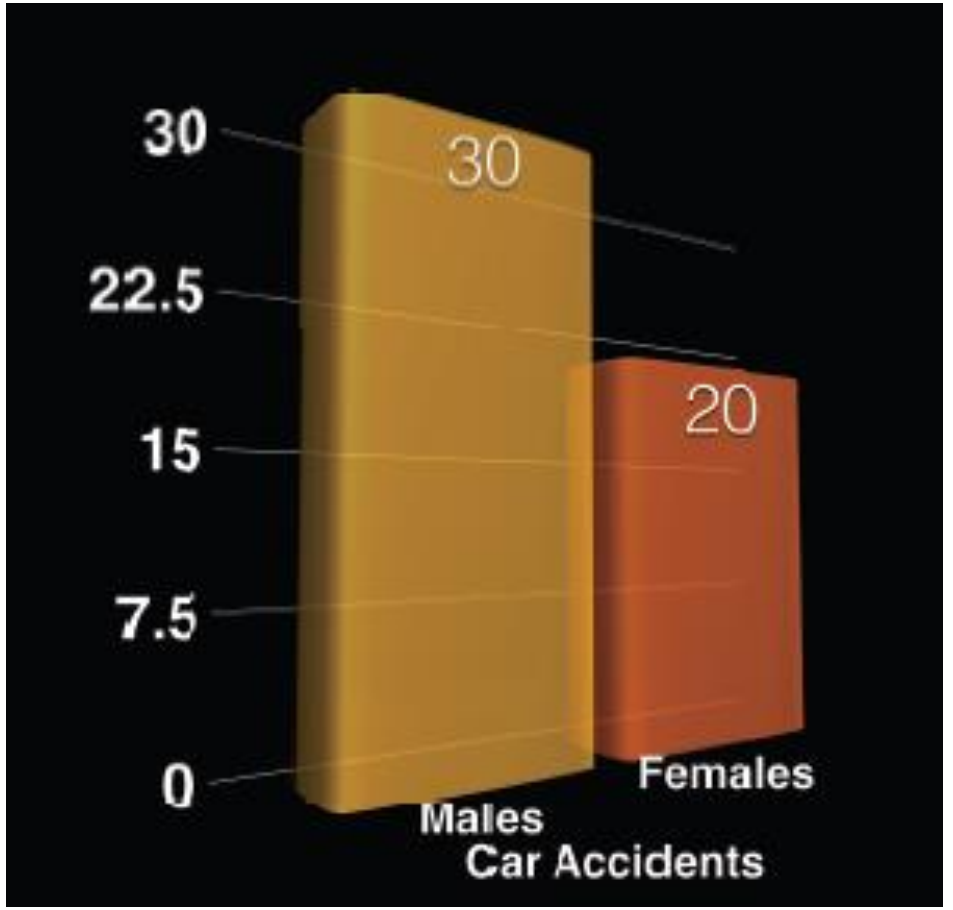
Area principle



- The area principle
- — the size of the area correlates with the data summaries.
- ~30% of accidental deaths of males were due to automobile accidents.
- ~20% of accidental deaths of females were due to automobile accidents.

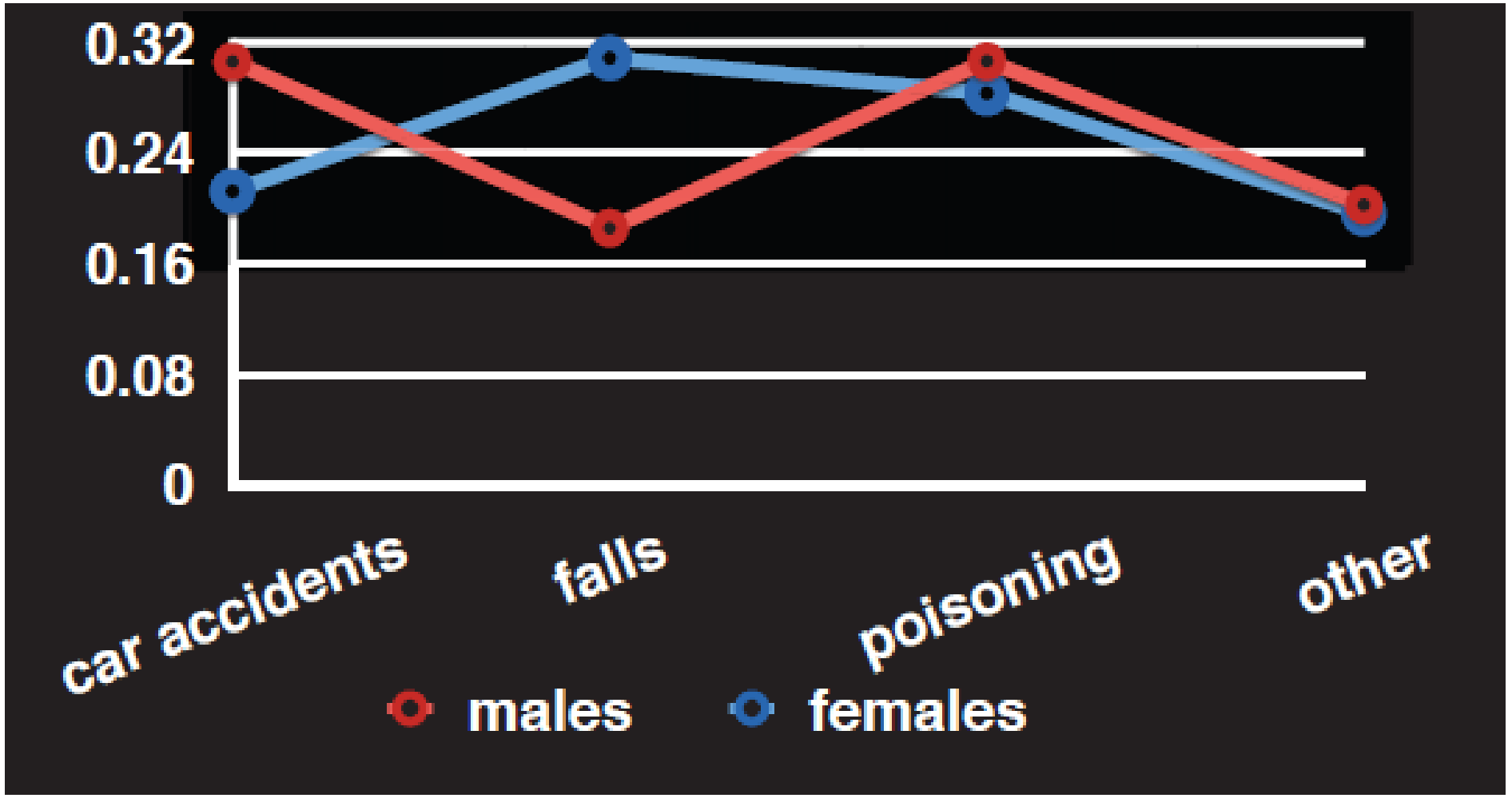
3D effects?





- **3D effects distort the visualization and often violate the area principle.**

Side by side comparison

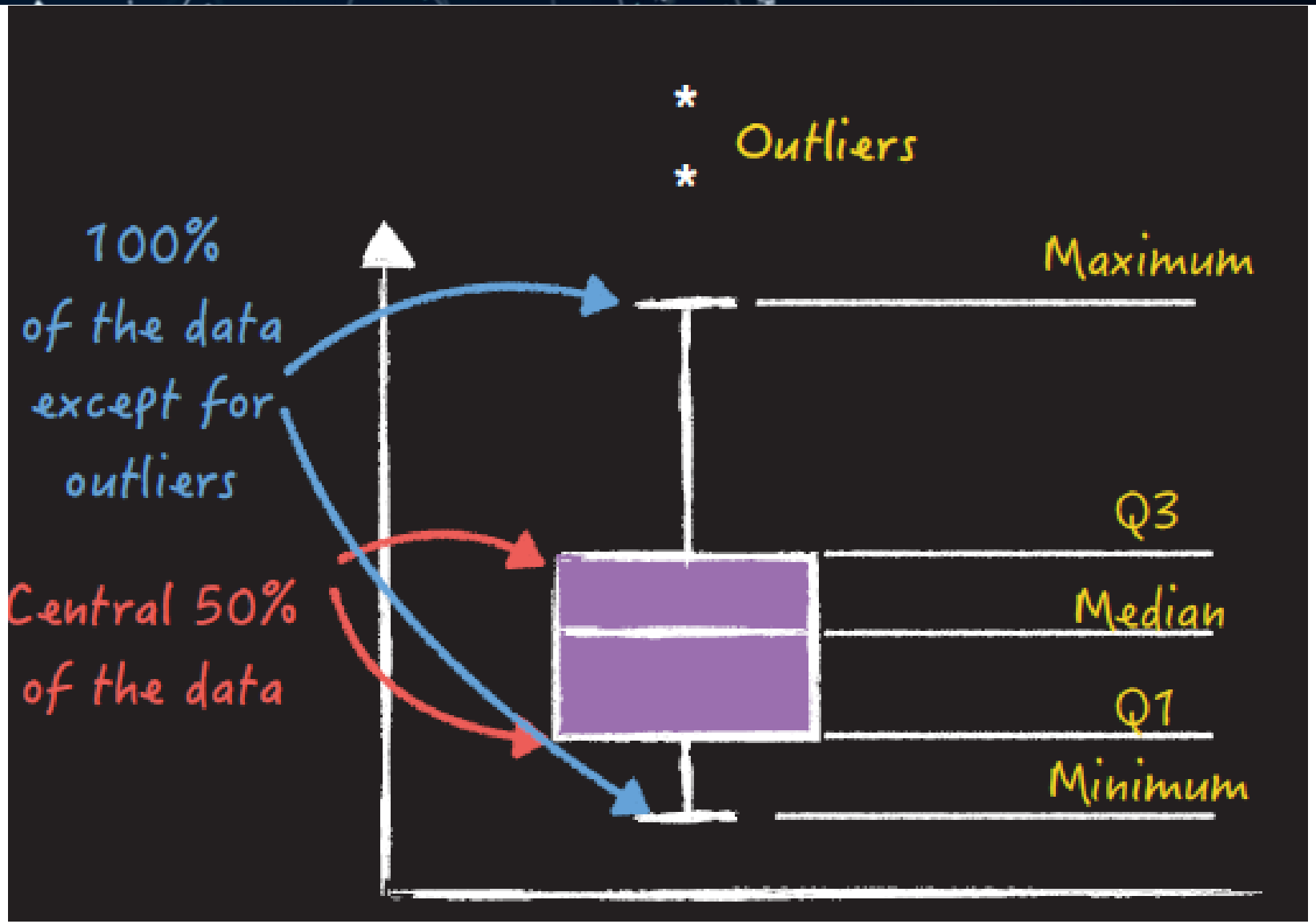


- Accidental Deaths in 2013 by gender and by different causes (proportions)

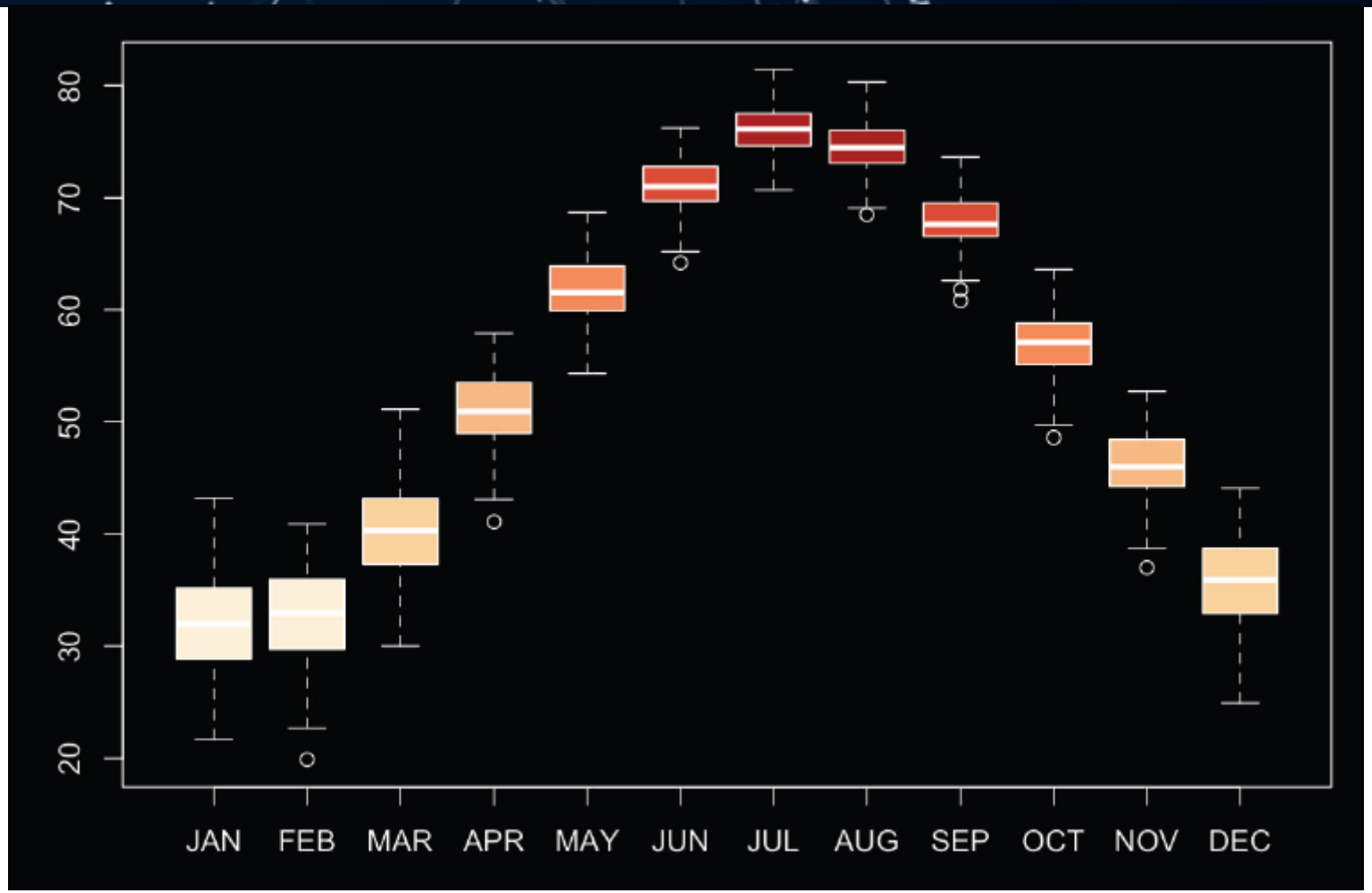
Create Meaningful visualization

- Accidental Deaths in 2013 by gender and by different causes (raw counts)

Box plot

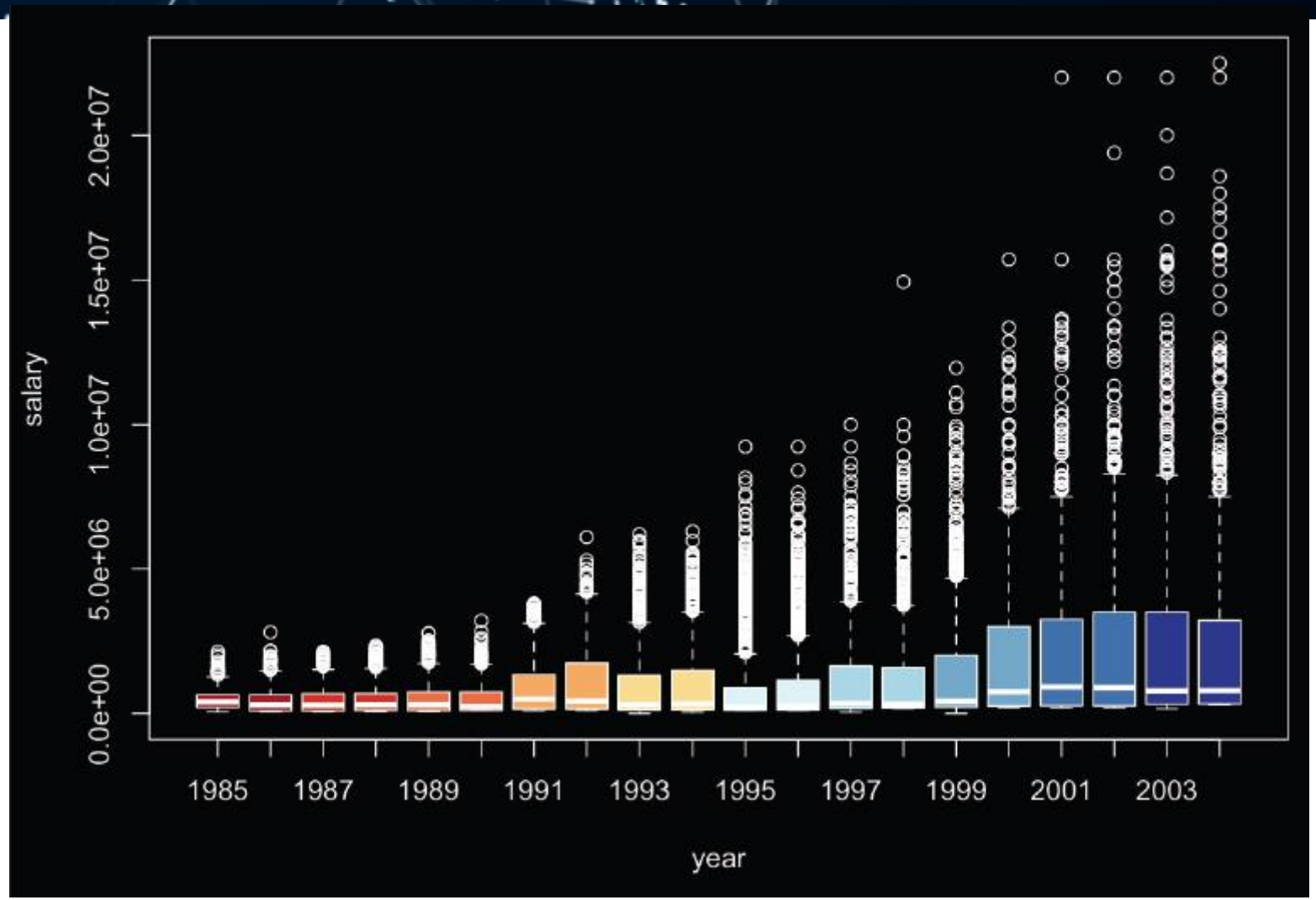

Outliers
100%
of the data
except for
outliers
Maximum
Q3
Median
Q1
Minimum
Central 50%
of the data

Box plot



NYC Monthly Average Temperature 1869-2012

Box plot

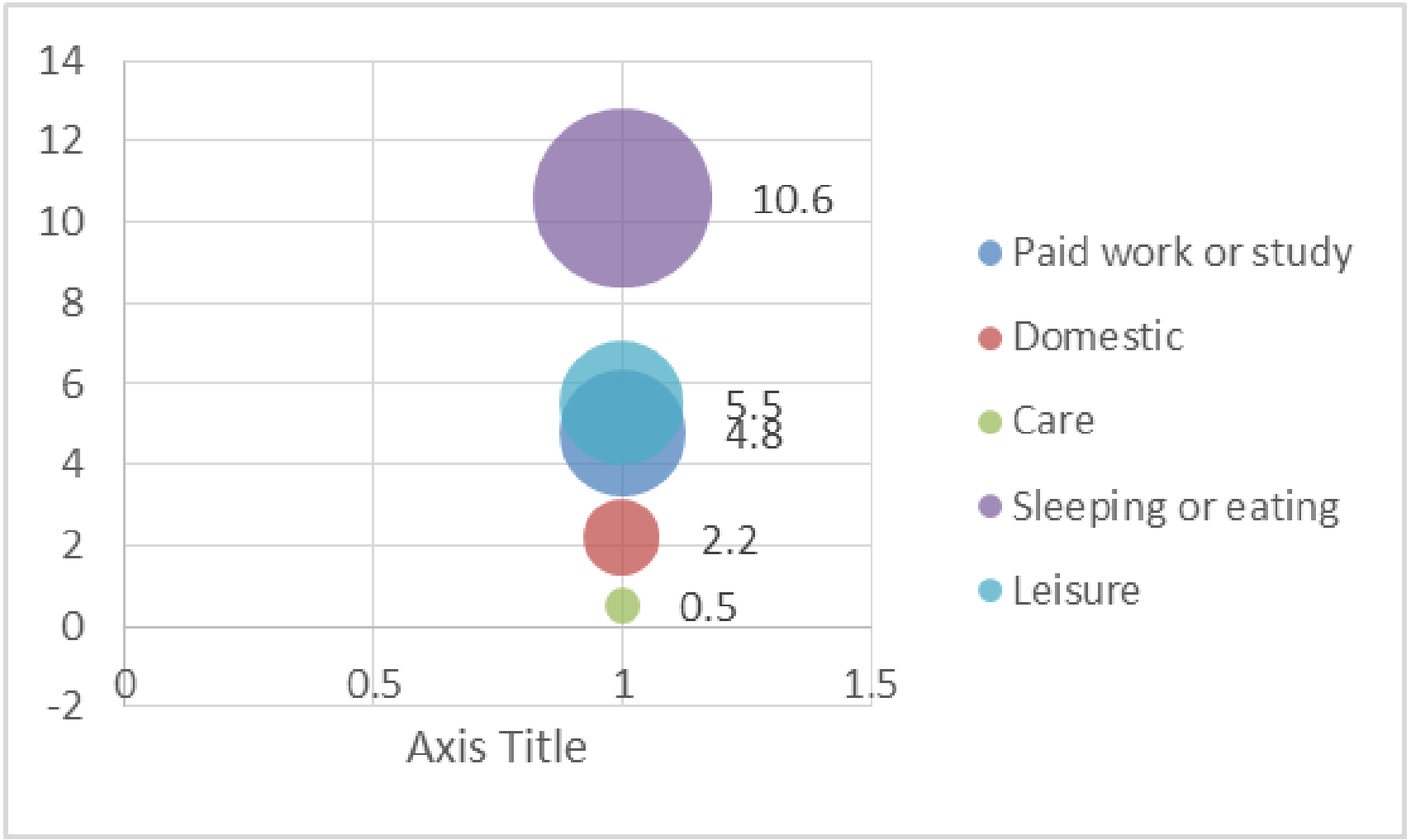


MLB Salary 1985-2004

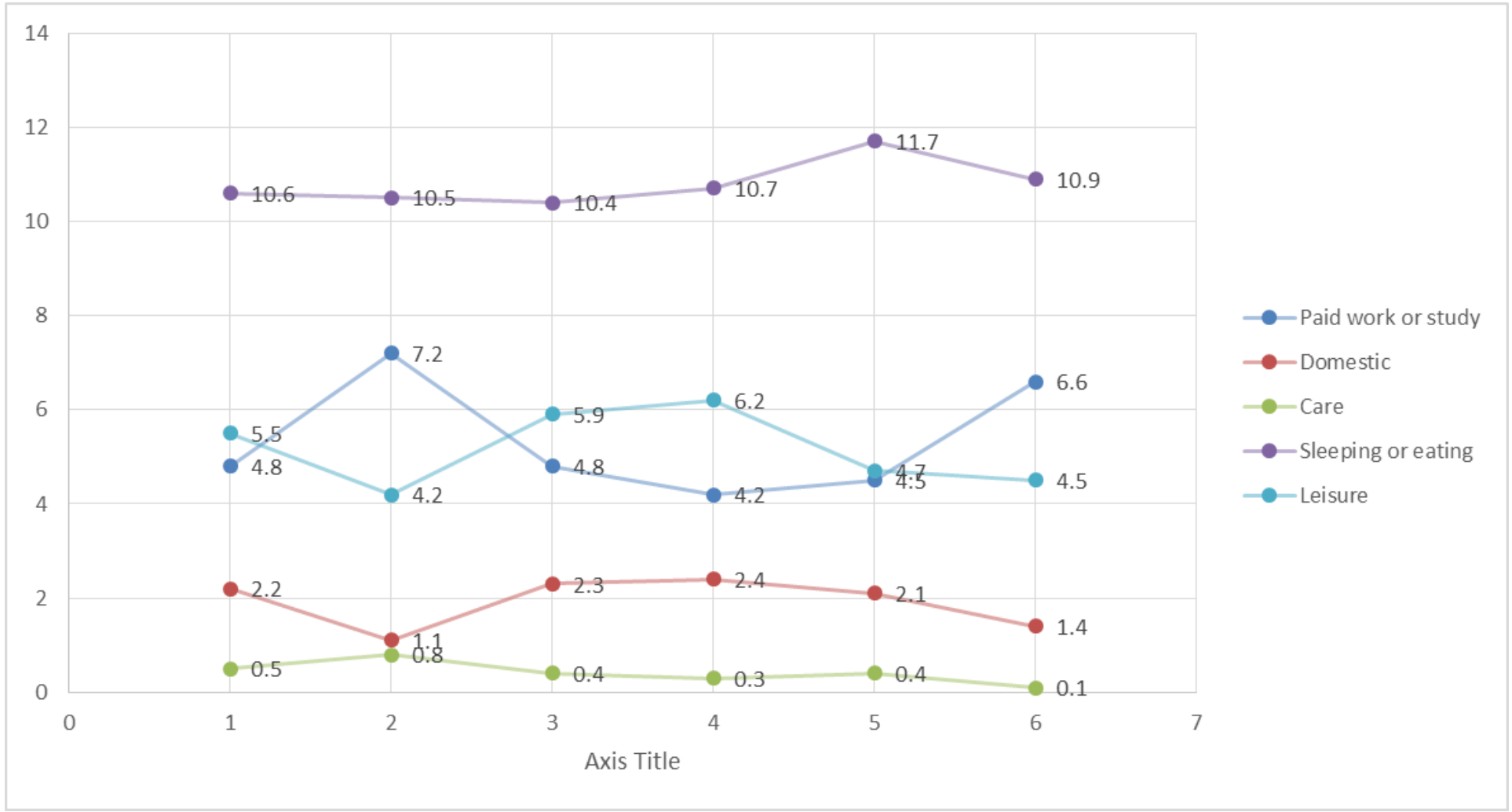
Data Table

	Paid work or study	Domestic	Care	Sleeping or eating	Leisure	Other
USA	4.8	2.2	0.5	10.6	5.5	0.3
Mexico	7.2	1.1	0.8	10.5	4.2	0.1
UK	4.8	2.3	0.4	10.4	5.9	0.2
Germany	4.2	2.4	0.3	10.7	6.2	0.3
France	4.5	2.1	0.4	11.7	4.7	0.6
Japan	6.6	1.4	0.1	10.9	4.5	0.5

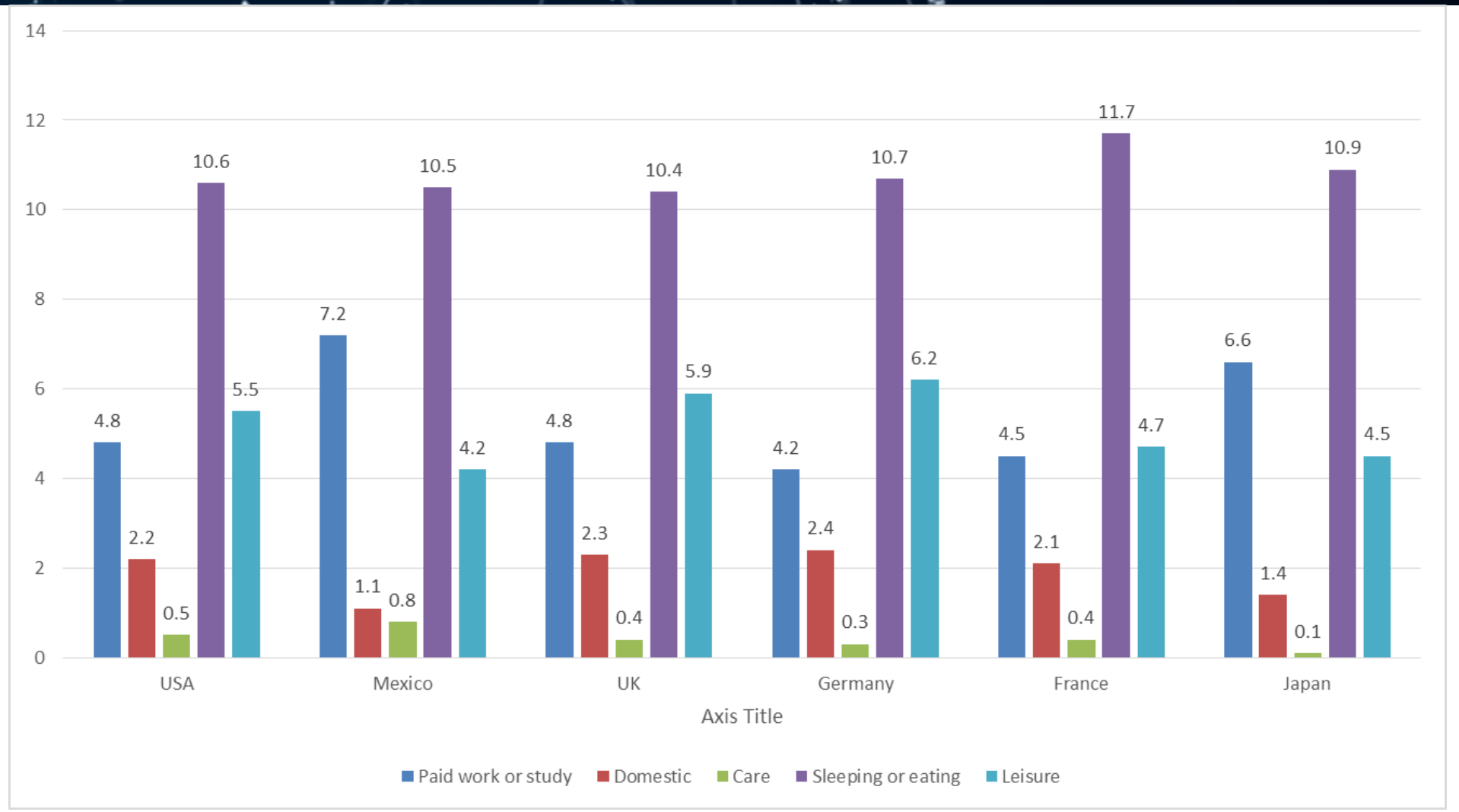
Ehsan Akhtarkavan



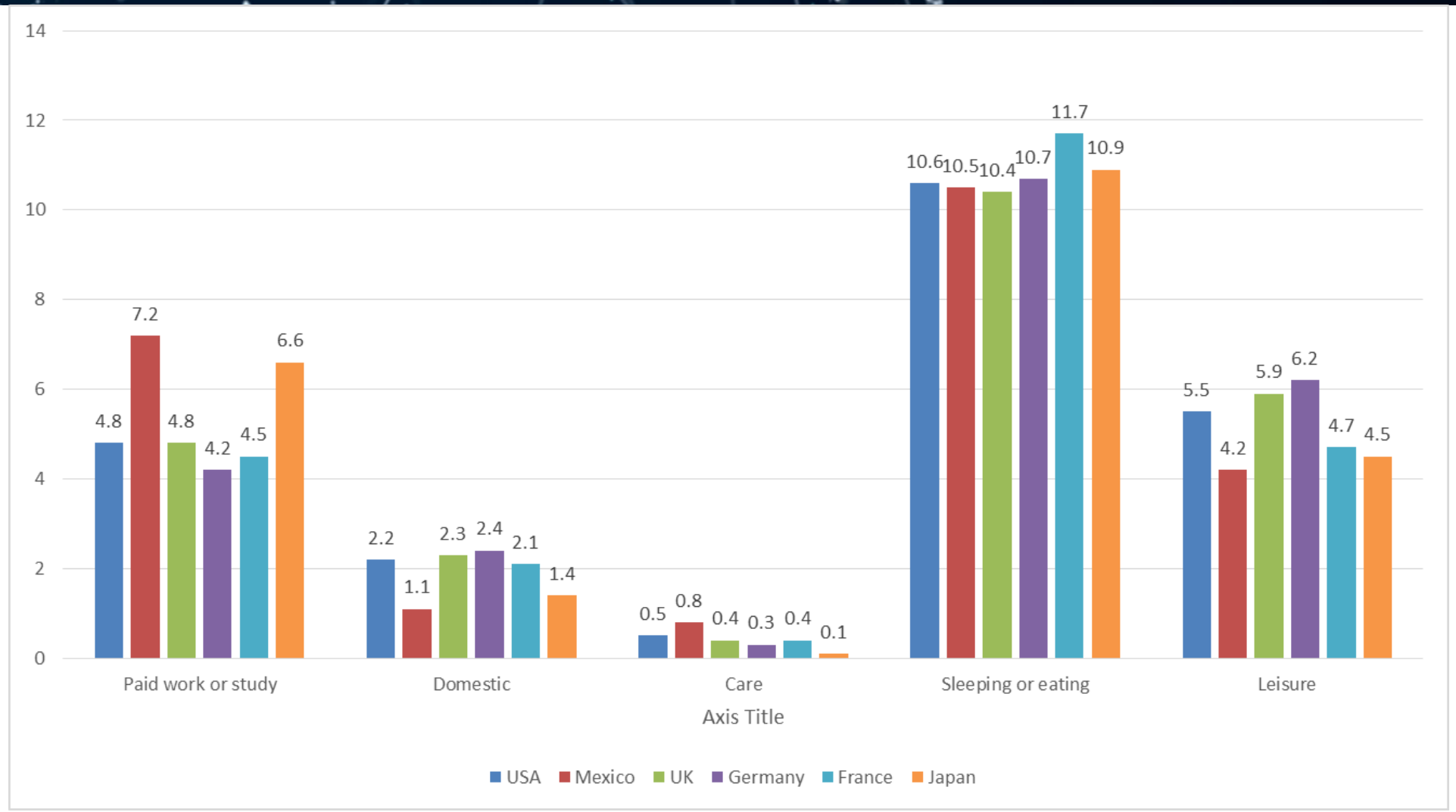
Ehsan Akhtarkavan



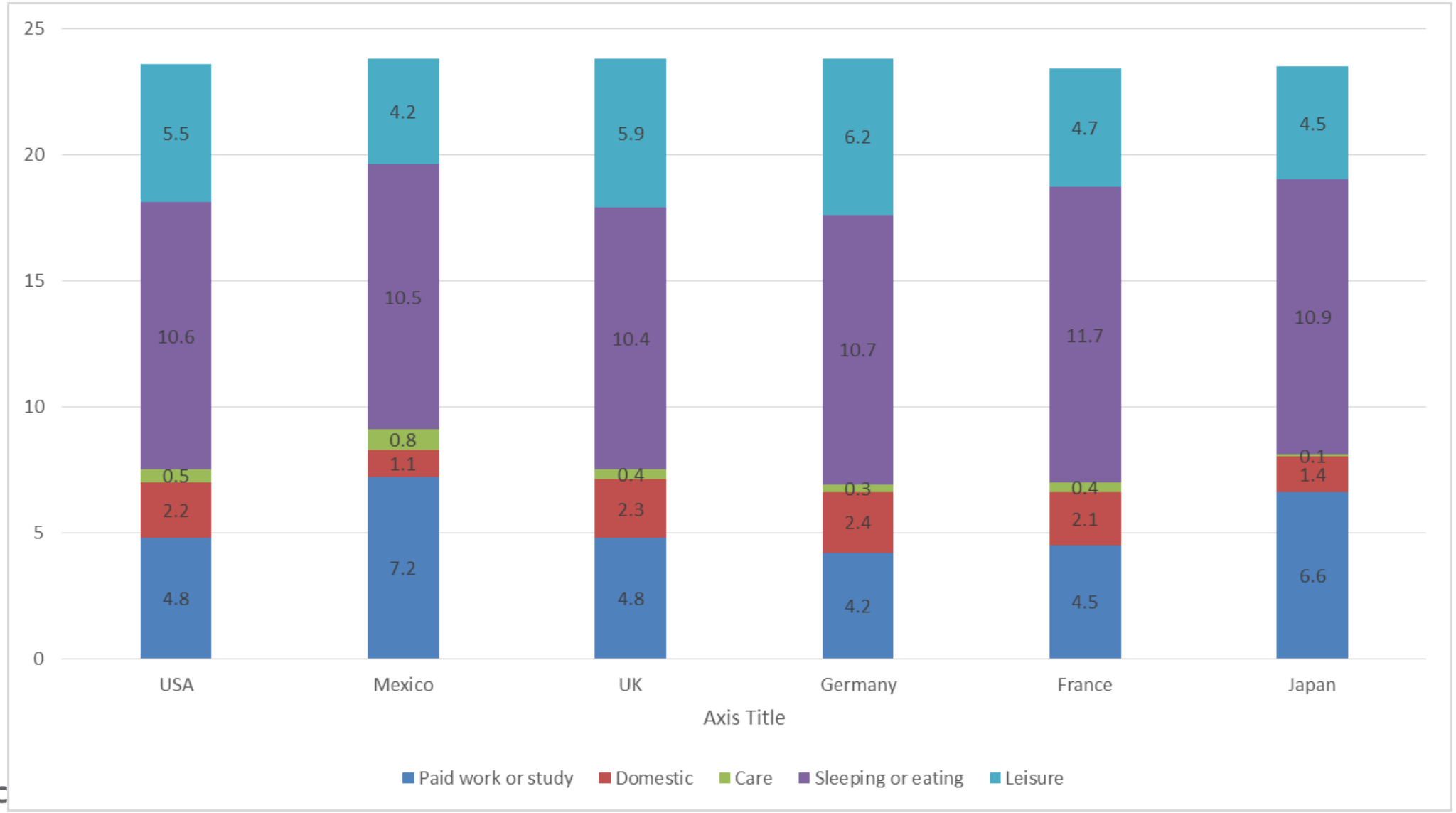
Ehsan Akhtarkavan



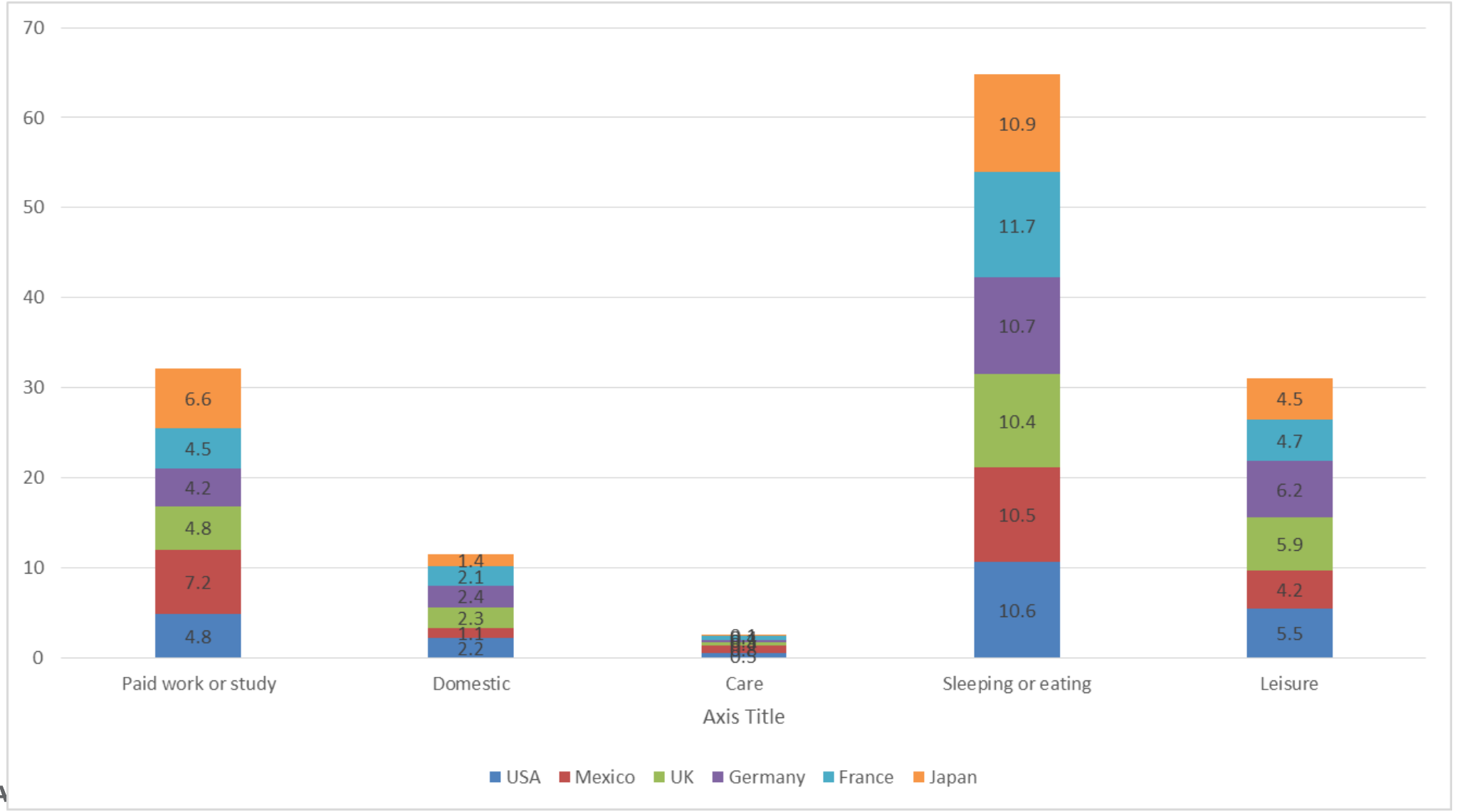
Ehsan Akhtarkavan



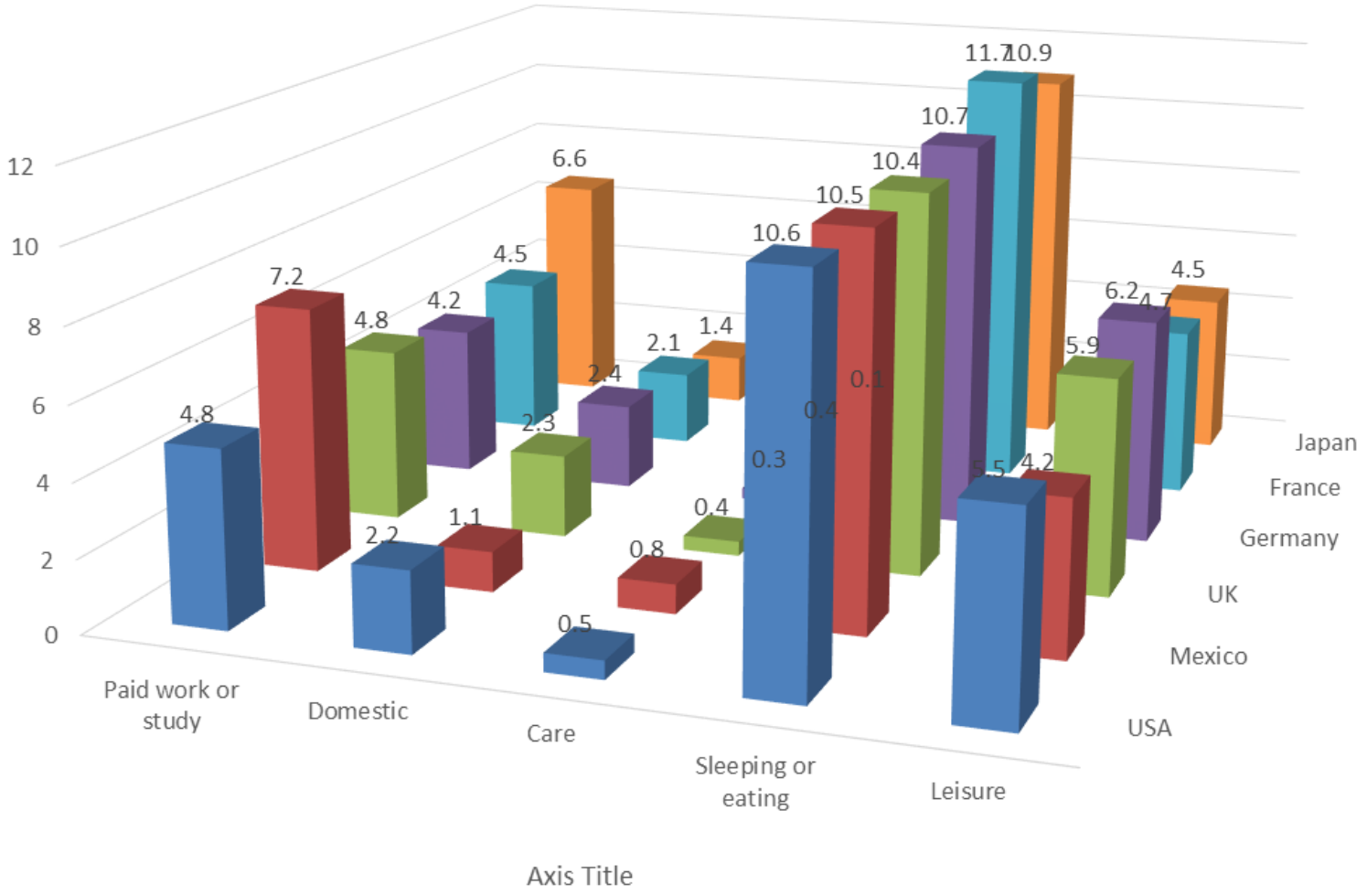
Ehsan Akhtarkavan



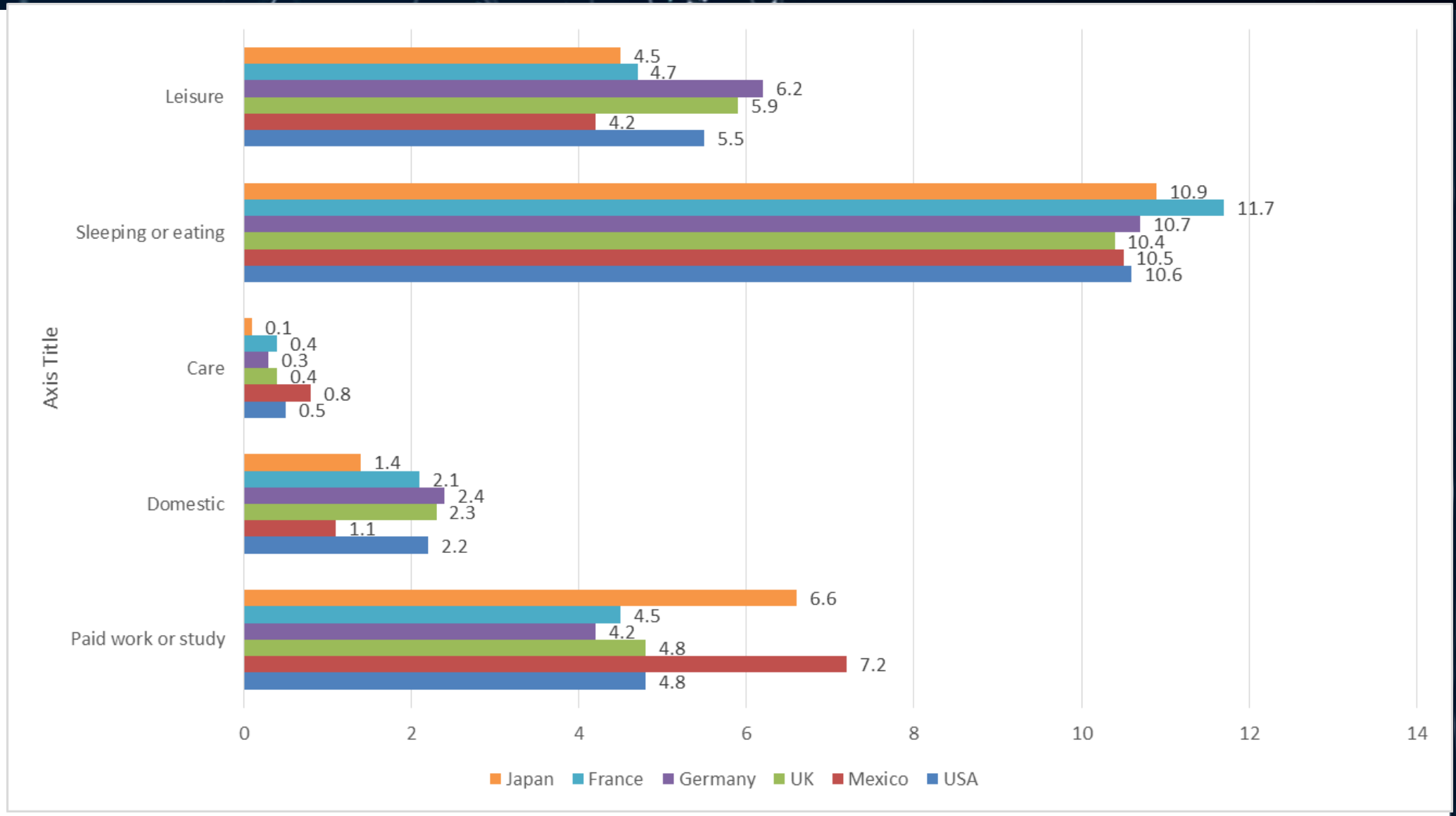
Ehsan Akhtarkavan



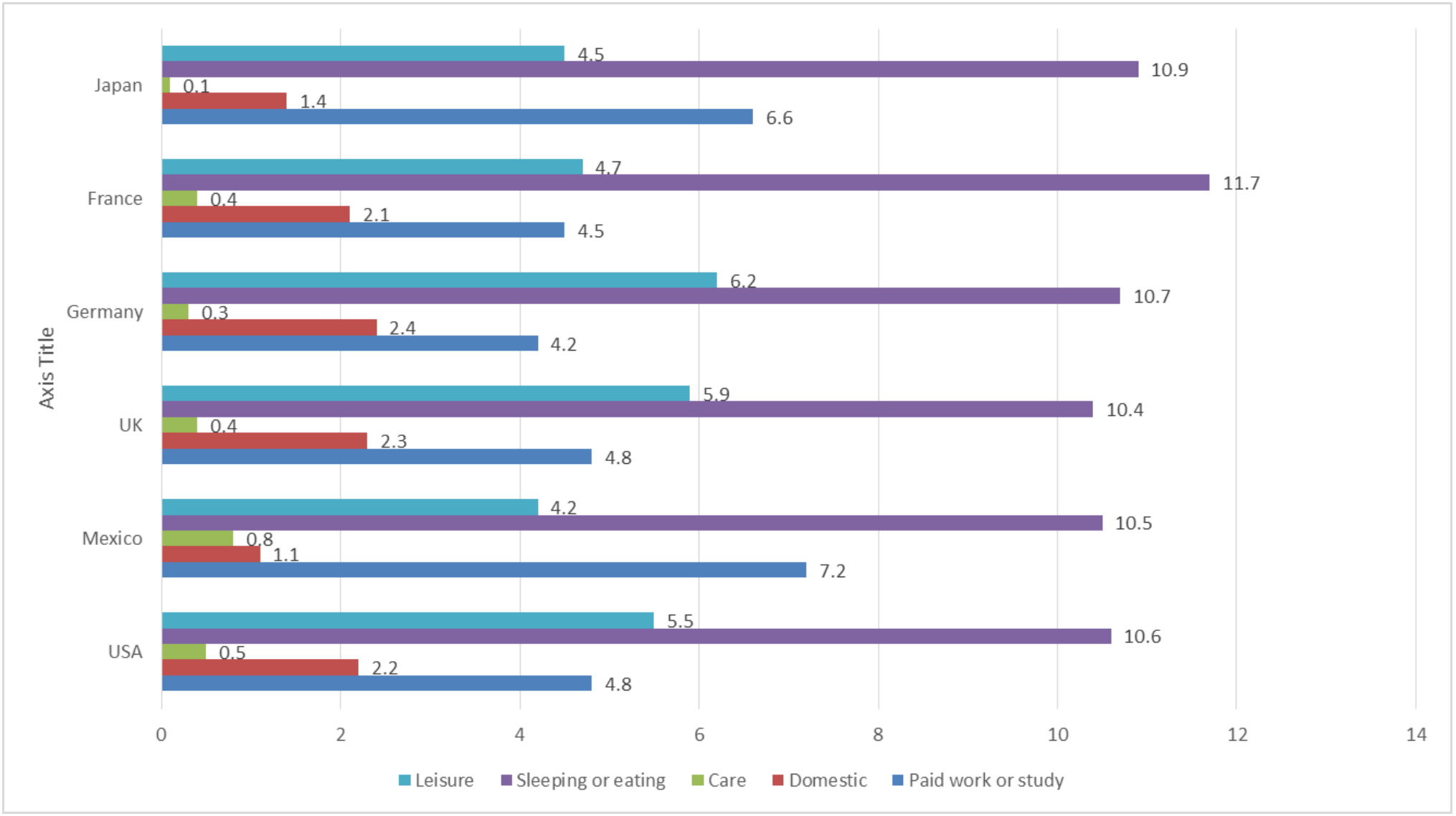
Ehsan Akhtarkavan



Ehsan Akhtarkavan



Ehsan Akhtarkavan

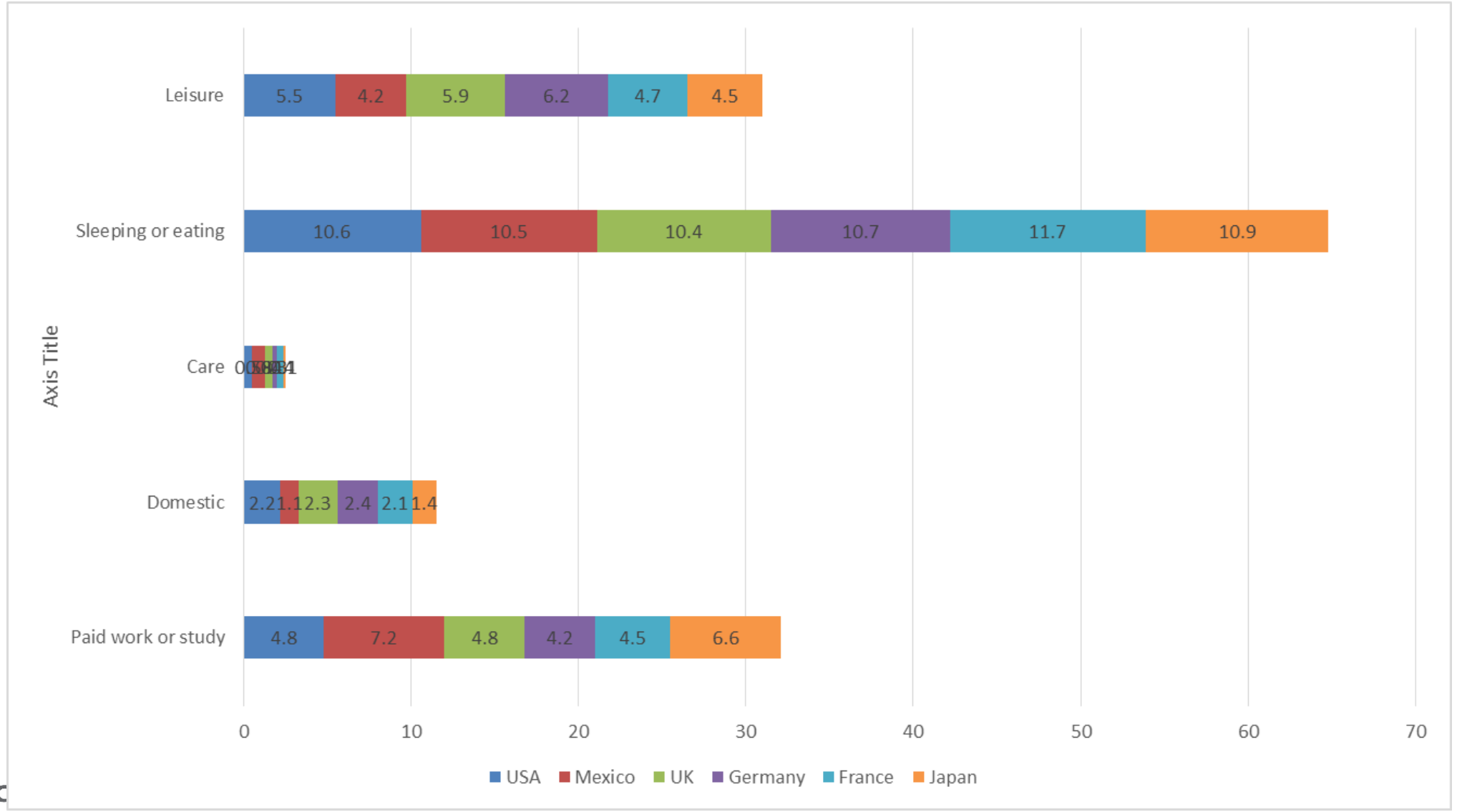


Ehsan Akhtarkavan

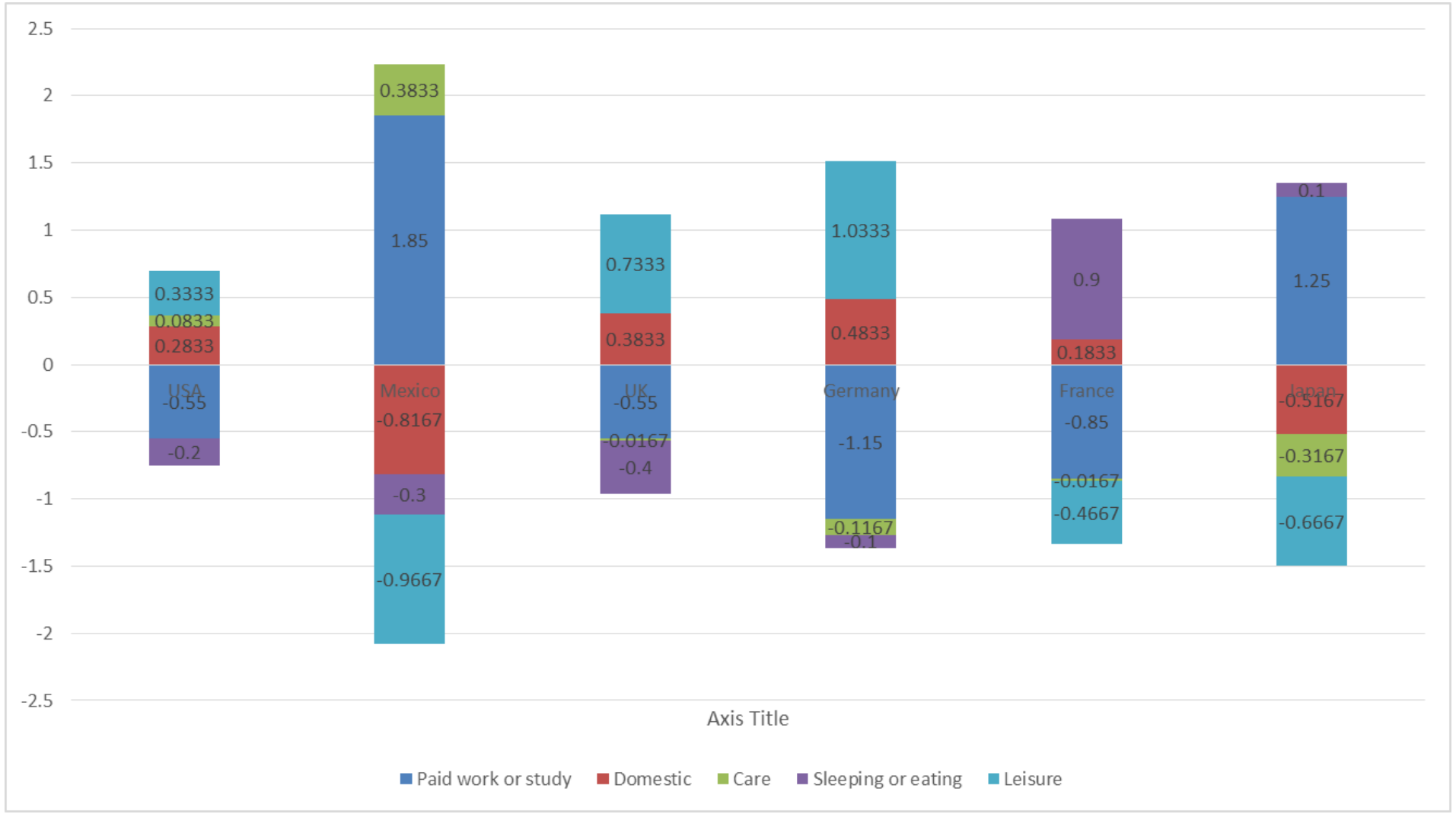
	Paid work or study	Domestic	Care	Sleeping or eating	Leisure
USA	-0.55	0.2833	0.0833	-0.2	0.3333
Mexico	1.85	-0.8167	0.3833	-0.3	-0.9667
UK	-0.55	0.3833	-0.0167	-0.4	0.7333
Germany	-1.15	0.4833	-0.1167	-0.1	1.0333
France	-0.85	0.1833	-0.0167	0.9	-0.4667
Japan	1.25	-0.5167	-0.3167	0.1	-0.6667

Value – mean(Values)

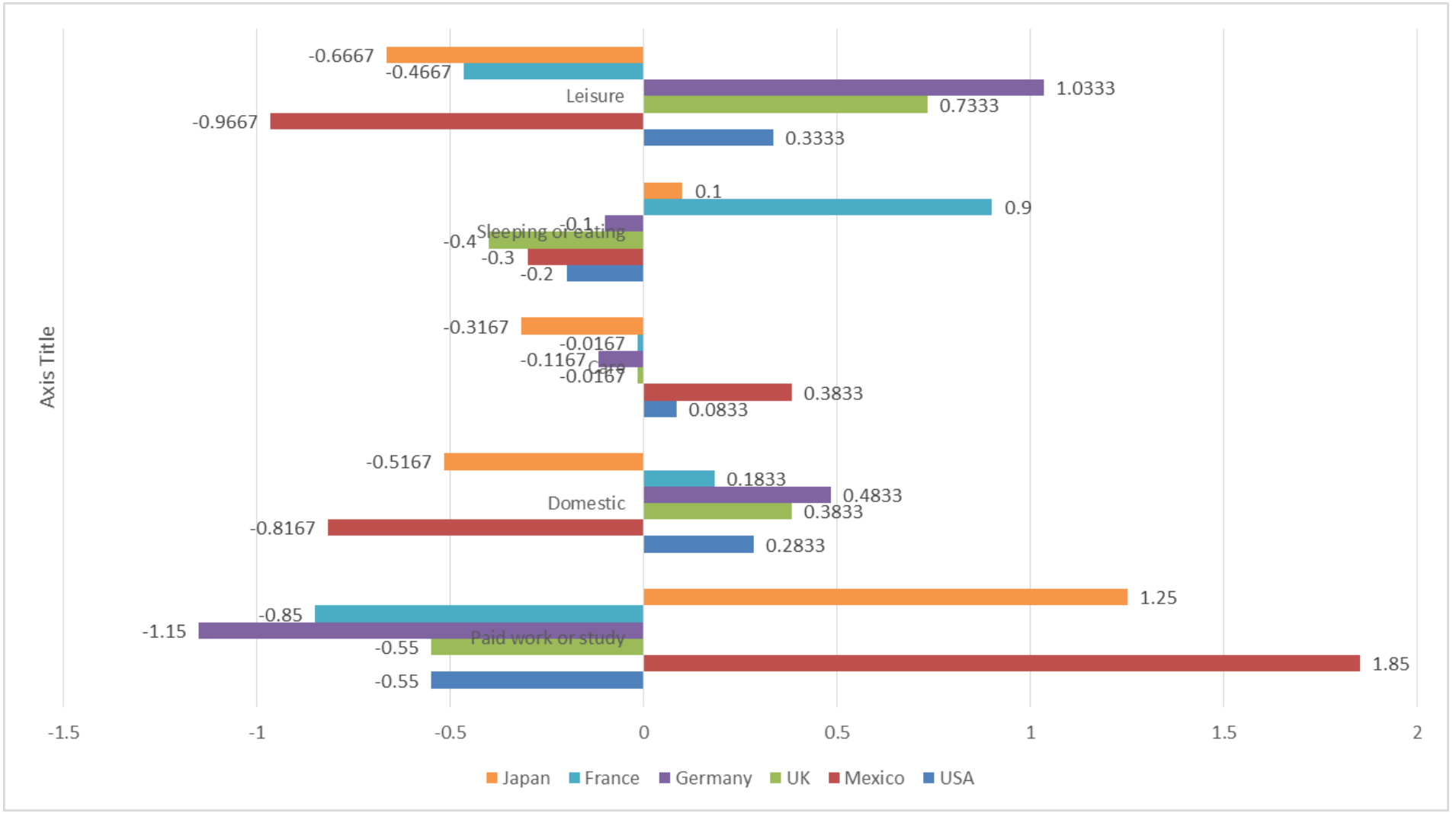
Ehsan Akhtarkavan



Ehsan Akhtarkavan

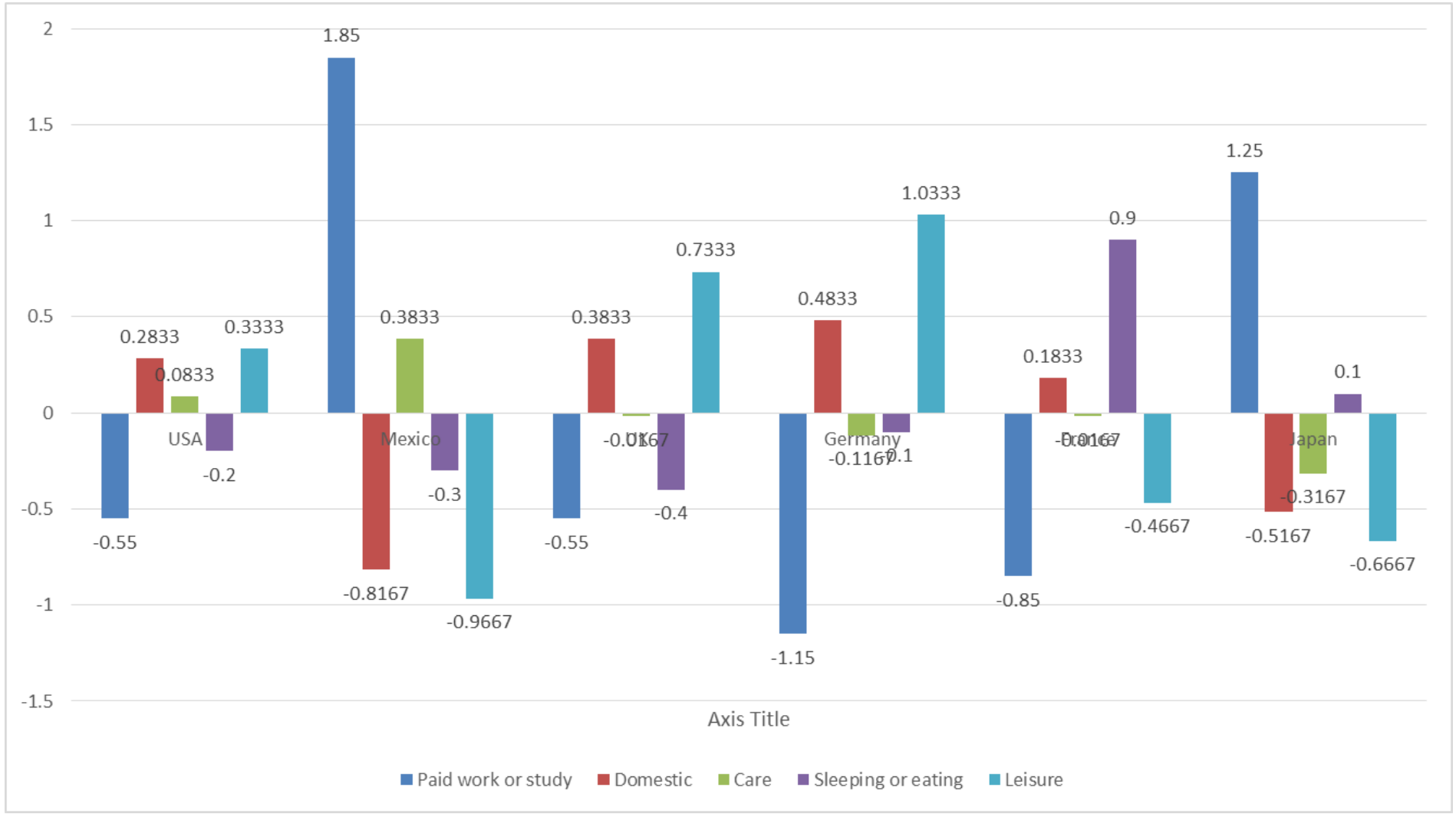


Ehsan Akhtarkavan



Ehsan Akhtarkavan

Ehsan Akhtarkavan



Ehsan Akhtarkavan

با تشکر ؟؟؟

فناپ
FANAP
وزارت علوم، تحقیقات و فناوری
دانشگاه خاتم
Khatam University