Intro to Big Data: Crash Course Statistics #38

Crash Course: Statistics
https://youtube.com/watch?v=vku2Bw7Vkfs
https://nerdfighteria.info/v/vku2Bw7Vkfs

Hi, I'm Adriene Hill, and welcome back to Crash Course Statistics.

You may have seen an ad before this video, or maybe there's one on the twitter feed you're scrolling through right now. Those ads are great examples of how "Big Data" is used.

They're often chosen just for you based on the sites you've been to, your sex, approximate age, where you live and a bunch of other variables. That data is part of HUGE--GIGANTICALLY HUGE--amount of data about you and everyone else. Almost every time you click, or don't click an ad, that data gets stored somewhere.

Every time you watch a YouTube video like this one, YouTube keeps a record of it. Even some toothbrushes and water bottles collect data on your everyday habits. Data sets include the clicks of everyone who has ever been on Amazon every like and comment on every Instagram picture every purchase you make with a credit card every show you stream on Netflix and how long you watch.

With 7.5 billion people on the planet, lots of data is created every second. I mean pretty much just by existing, you're creating data. So much data that we call it…"Big Data." INTRO In the days before smartphones, laptops, and personal computers, data was hard to come by.

It took a lot of time and effort to record measurements, and store them. Often, data from the United States Census--which takes place every 10 years--would take almost 10 years to collect and put together. Computers have helped shorten the time it takes to collect, summarize, and store data but as our power to collect and analyze data increases we just make more and more of it.

The term "Big Data," in the way we use it today, is usually credited to John Mashey. In the 1990s, he used the term to describe data that is so big and complex, that commonly used tools to work with data, everything from collecting to interpreting, just can't handle it. Your phone records your location, the apps you use--and how long you use them--and all those apps that you use are each collecting their own data on you.

That's why Stub Hub won't stop pinging me about Taylor Swift concert tickets. They KNOW me. The Coca Cola Company collects data from tons of places, including those soft drink machines that let you add a variety of additional flavors to your regular soda of choice.

That's the reason we now have Cherry Sprite! Enough people were choosing that combination and Coke had the data to prove it, so they put it in cans. We've created an interconnected world that's sometimes referred to as the "Internet of Things".

Consider the network of "smart" devices that collect data and can potentially communicate with each other everything from your refrigerator to your car to your watch to your lights. Scientists have even rigged some SPINACH plants to be able to wirelessly send emails about their surroundings. Even when you visit a ski resort, they're collecting data.

They may give you a scannable RFID pass, allowing automated ski lift access. Plus, the resort employees will know where you are while you ski. And an app will give you all kinds of stats, like how many days you've skied and your vertical distance.

The whole point of Big Data is that there's too much of it to wrap our heads around. So let's take one tiny aspect of it: Facebook likes. For years, those likes seemed pretty useless.

I don't care if you like The Godfather or Starbucks or Beyoncé. Everybody likes those things. But, that information is more revealing than you might think.

In 2013, the Proceedings of the National Academy of Sciences published a study out of the Psychometrics Centre at Cambridge University. The participants were around 58,500 Facebook users who took a personality survey on the researchers' app. Then, they requested permission to view the users' "likes." They found, "Individual traits and attributes can be predicted to a high degree of accuracy based on records of users' Likes." So liking "Thunderstorms," "Science," and "Curly Fries" were signs that someone was highly intelligent.

Liking "Wu-Tang Clan," "Shaq," and "Being Confused After Waking Up From Naps" pointed towards someone being a heterosexual man. A person's interest in Hello Kitty led to a surprisingly detailed prediction. The paper claimed, "Users who liked the 'Hello Kitty' brand tended to be high on Openness and low on 'Conscientiousness,' 'Agreeableness,' and 'Emotional Stability.' They were also more likely to have Democratic political views and to be of African-American origin, predominantly Christian, and slightly below average age." This is a tiny piece of the puzzle that can give you a sense of Big Data in action.

If a little bit of information about a person can actually reveal a lot, then multiply that by the tons of other data they're producing each day. Then, that data gets used. Facebook itself sorts people into categories, like political views.

In 2016, the New York Times reported, "Even if you do not like any candidates' pages, if most of the people who like the same pages that you do -- such as Ben and Jerry's ice cream -- identify as liberal, then Facebook might classify you as one, too." (That's just for the U. S., by the way. We don't know what they're gathering about people's views in other countries.) Categories like this allow advertisers on Facebook to select very specific criteria and send ads to the exact groups of people that they want to see them.

For example, a Bloomberg analysis of 2016 U. S. presidential campaign finances noted that the Trump campaign chose particular groups of Hillary Clinton supporters to see anti-Clinton ads on social media, trying to make them less likely to vote. Between May and July of 2018, the Planned Parenthood Federation of America was second to The Trump Make America Great Again Committee in Facebook political ad spending in the U.

S.. A Planned Parenthood spokesperson told the New York Times, "Running ads on Facebook is a targeted and cost-effective way to reach both our 2.4 million patients and 12 million supporters." They use location targeting, so they can be specific about their resources in a given area. The spokesperson also noted that they run negative political advertisements about the Trump-Pence administration.

And the political implications go beyond that. Another researcher at Cambridge University, Aleksandr Kogan, used a similar method and quiz app to that study I mentioned earlier. That helped the political consulting film Cambridge Analytica get data on up to 87 million Facebook users.

There's a good chance that Big Data has positively impacted your life. Perhaps you saved some money on your grocery bill by using coupons that were tailored to your shopping habits. Or you got to buy that Cherry Sprite in a can.

Big Data is used to personalize medicine, to predict which baseball players a team should recruit, and to create driverless cars. You're also using Big Data every time you use Google Maps. If you have your location enabled on your phone, information about your

location and speed is constantly being sent back to Google.

That information alone isn't super useful to anyone. BUT, countless people around you are also using Google Maps. So, Google has a TON of data about where people are and how fast they're moving.

Because they've been doing this for a while, they also know what traffic SHOULD look like based on things like the day of the week, what time it is, even holidays. So, with all their data, they can then tell you whether there's a lot of traffic on a particular road. In 2013, Google acquired the app Waze, which gave them even MORE data to work with.

Waze users tell the app when they see traffic and accidents. So your Google Maps app uses that, too. It also keeps track of your personal history, which is how it can prepare you for your specific morning commute.

The system City Brain, which was implemented in Hangzhou, China starting in 2016, takes this concept one step further. The goal of City Brain is to minimize traffic in the city. And like Google Maps, it's also run by a company: a huge retailer called Alibaba.

The difference is: they have the help of local government as well. So, the City Brain A-I system gets data in ways similar to Google Maps. But, they also have access to information from the transportation bureau and city surveillance cameras.

Alibaba claimed they were able to increase traffic speed by 15% in an area where they had been given the power to control over 100 intersections. And it's a two way street. (Pause for ungodly amounts of laughter.) The city also uses their access to this information to see where accidents have occurred, to get directions for emergency vehicles, and to determine areas that need infrastructure changes. In 2018, it was announced that City Brain was being implemented in a second city: Kuala Lumpur, Malaysia.

Of course, I don't expect you to be unquestionably psyched about all of this. I'm not. Not everyone wants private companies to know where they are.

And we're going to talk about privacy concerns in depth next week. But let's move onto another use of Big Data in the Thought Bubble. Netflix uses "Big Data" to improve your entertainment experience.

To give recommendations, Netflix's algorithm learns from an endless stream of data on clicks, watch time, if you like movies starring Matt Damon. It might learn that people who watch Queer Eye, are more likely to enjoy The Great British Bake Off, and that people who binge watch tend to like shows with more Seasons available. It's also why you might get weirdly specific category recommendations like "Lovable Losers" and "TV Shows about Friendship." But Netflix doesn't stop there.

Big Data also influences the image you'll see for a show or movie. For example, here are some of the images you might be shown for the Netflix show, Stranger Things. Netflix uses all the data at its disposal to decide which image you'll see.

Since the Title and Image are your first exposure to the content, choosing a picture that's attractive to you can affect your decision to watch it. Take, for example, the movie Good Will Hunting. This post from the Netflix Tech Blog shows how your past viewing habits can influence which image you get5.

If you're an avid romance watcher, you might be more drawn to a picture of Matt Damon and Minnie Driver kissing. But, if you watch a bunch of comedies, Robin Williams might be enough to convince you to watch. You wouldn't have known he was in the movie if you had been shown the other image, and you may never what Ben Affleck's Boston accent sounds like.

Just kidding. He does it in every movie..you'd know. Using the HUGE amount of data at its disposal allows Netflix to make YOUR viewing experience better.

How do you like them apples. Thanks, Thought Bubble. And Big Data can do much more than convince us to watch a movie.

We could be able to better personalize medicines by sequencing a patient's genome, and predicting which medicine will have the fewest side effects. Or which treatment is least likely to interact with an existing heart condition. Big Data is here to stay.

It lets us do things like use machines to recognize the faces of criminals based on security footage, or make sure that Amazon Warehouses are stocked so that you can get a video game for your niece in time for her birthday. And you're creating it right this second. YouTube knows you made it to the end of the video or at least nearly the end.

But the complexity and sheer amount of data that's being collected can present some problems. In the next episode, we'll talk about a few different ways we can overcome or at least manage some of them. Thanks for watching, I'll see you next time.