Big Data Problems: Crash Course Statistics #39

Crash Course: Statistics
https://youtube.com/watch?v=Im3GkAYUivE
https://nerdfighteria.info/v/Im3GkAYUivE

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

In the last episode, we talked about the value of Big Data. But, as Big Data (and the statistics we do with it) permeate more areas of our lives, there are also new problems that come up.

How can we learn from useful data while still keeping it safe and private? When there's SO much data that we have to rely on algorithms to manage it can we trust those algorithms? Today, we're going to have a discussion about the potential downsides of Big Data in our lives and some possible solutions.

INTRO Let's start with a Thought Bubble. This story comes out of a collaboration between the University of Washington and the University of California Irvine. The team wanted to create an algorithm that could take an image and determine whether it was of a husky or a wolf.

To do that, they trained the algorithm with a bunch of pictures. These images are a great example of "BIG DATA" that we CAN wrap our heads around. Pictures are generally made up of millions of tiny pixels.

And each pixel is made up of three colors, red, green, and blue. So three values per pixel is a lot of data. The algorithm ended up doing pretty well.

The team may have anticipated that it would recognize the animals' different body types, facial features, or body placement. But it turned out that it wasn't focusing on the animals' appearances at all. It was mainly looking at snow.

In the data used to create the algorithm, the research team inadvertently included many photos of wolves in the snow. Huskies were often pictured without any snow around. The algorithm picked up on that and glommed onto it as an easy way to tell if something was a husky or wolf.

Once the researchers learned this, they did an experiment, feeding the algorithm new images that had been digitally altered. According to one of them, Dr. Sameer Singh, "When we hid the wolf in the image and sent it across, the network would still predict that it was a wolf, but when we hid the snow, it would not be able to predict that it was a wolf anymore." The algorithm just learned from what it was given.

And based on the data it was trained with, snow meant the image was way more likely to be a wolf. Even more than adorable wolf-i-ness. Thanks, Thought Bubble.

That algorithm brings us to our first concern with Big Data: bias. The defining characteristic of "Big Data" is that it's big. It's too big for a lot of the usual programs we use to look at data. In fact it's sometimes even too big for us to comprehend. And when huge amounts of data are used to create algorithms, we can inadvertently introduce bias. Like the wolf and snow problem Other algorithms could do similar things, but with higher stakes. Ones used to determine mortgage and insurance rates, or assess the risk someone will do something illegal in the future, might pick up on things like race or other minority statuses.

And this is real. Judges in the U. S. use risk assessment programs while making sentencing decisions. A commonly-used one is COMPAS, which was created by the company Equivant. It basically gives a score of how likely a person is to commit another crime within two years. In 2016, ProPublica published an investigation of COMPAS. They looked at the scores of 7,000 people who had been arrested Broward County, Florida. The scores were compared with whether those people actually ended up committing crimes again within two years.

In addition to other concerning revelations, ProPublica found, "The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants. [And] white defendants were mislabeled as low risk more often than black defendants." Equivant, then called Northpointe, did disagree with these findings. But, the company's founder, Tim Brennan, also claimed that in order to make scores as accurate as possible, certain factors had to be included that could correlate with race. ProPublica cited the examples "poverty, joblessness and social marginalization." So we can't consider ourselves "safe" just because we think that our data is neutral.

We should also look for ways to make sure that the data we use to create our algorithm is as representative as possible. If we wanted to build an algorithm that predicted the success of CEOs, and we ONLY gave it examples of males who succeeded and females who failed then our algorithm will have a bias. We have to supply it with good, unbiased data. Males who succeeded and failed, and females that succeeded and failed.

In the tech world, you'll often hear the phrase "Garbage In, Garbage Out" which means that bad input will lead to bad output. You can't put biased data into an algorithm, and expect an unbiased output.

It can be hard to determine what kind of data will lead to biased decision making especially considering most of these algorithms are proprietary. In the Equivant example I mentioned earlier, the company wouldn't reveal the details of the algorithm used for COMPAS to ProPublica for that exact reason. It's also hard to figure out exactly what an algorithm is doing from the time we give it raw data, to the time it gives us an output, or decision.

With the methods we've talked about in this series, like regression, it's easy to see which variables they consider important. But other Big Data methods, like neural networks, are often way less forthcoming with the "reasoning" behind their outputs. While we can't always tell what algorithms are doing, some researchers have made other algorithms that can act as a sort of translator to turn the complex calculations of another algorithm into something humans can understand. The more humans can understand what an algorithm is doing, the more opportunities we have to recognize biased data and the resulting decisions.

Some believe that search and social media websites that use algorithms to affect your experience based on your data should be required to release more information about that algorithm -- how it works and what it's doing. That's called algorithmic transparency. Privacy is another big concern in the Age of Big Data. There's all kinds of personal data about you that you might not want people to know. There are your entertainment choices, like what you're reading and how many times you've watched Fuller House or listened to "I Like It". And your school or work--emails, cloud services, web browsing. Even your basic information like your location, your step count, or heart rate are tracked by your various smart devices. Companies like 23andMe or Ancestry.com might even have your genetic code. Maybe you spend a lot of time at a place with security cameras. There are a lot of questions when thinking about privacy: Who has access to all that information? What are they doing with it? Who are they sharing it with? And what assumptions are they making about us with the data they have?

In 2018, The European Union implemented a law--the General Data Protection Regulation, or GDPR for short--that addresses a lot of the privacy concerns people have with the use of Big Data. It requires companies that deal with Big Data to be more transparent

about what they're collecting and who can see it. And it might be one of the reasons you got a LOT of emails about updated Privacy Policies…back in May of 2018.

The U. S. has The Children's Online Privacy Protection Act, which went into effect in 2000. It's intended to protect the privacy of children under the age of thirteen. The Act basically requires websites and apps to get parental approval for the personal information it might collect from kids. And using that information for targeted ads is not allowed. In 2018, a study of about 6,000 children's apps was published in the journal Proceedings on Privacy Enhancing Technologies. It found that about 57% of them were "potentially violating COPPA." Examples of violations included "sharing of personal information without applying reasonable security measures," "potential sharing [of] persistent identifiers with third parties for prohibited purposes," and "[sharing] location or contact information without consent." Later that year, the attorney general of New Mexico filed a lawsuit against an app maker for violating COPPA. Privacy laws have been around for a long time all over the world. But as they pertain to Big Data, a lot of this stuff is new, we're still figuring it out.

At the same time, when universities, hospitals, and other organizations share data, we learn a lot. It can be useful. A health organization's survey on risky behaviors, like drug use, could have incredibly valuable results to researchers and policy makers. So, we can try to make it so that data can't be easily connected to the specific person it came from. The obvious first step is to not include people's names, or other unique, identifying personal information. But that may not be enough. If someone has a rare disease, simply knowing the city where they live might be enough to figure out who they are. One option to combat this issue is to make sure that there are at least 2 or more subjects that have the same characteristics. This is called k-anonymity. K is the number of subjects who share the exact same characteristics. If there are two people with that disease from that city, we have 2-anonymity because there were 2 subjects with the same characteristics. In our dataset, these two subjects are indistinguishable from each other, which helps keep the data private. And the larger k is, on average, the better.

Outside of medical research, there are debates about what companies should be expected to keep private. DNA companies, for example. In 2018, Joseph James DeAngelo was arrested as the suspected Golden State Killer. Investigators found DeAngelo because they had DNA from a crime scene, which they uploaded to a public, online genealogy database called GEDmatch. The database doesn't collect DNA, but lets people upload profiles. So, the investigators were able to connect DeAngelo's DNA with other relatives and figure out who he was from there. Even though he hadn't uploaded anything to the site personally, the information his relatives had submitted was enough. GEDmatch does have rules about whose DNA you can upload to the site, like you can upload your own or someone else's with permission. Their site policy also currently states that DNA can be uploaded if it was "obtained and authorized by law enforcement to either: (1) identify a perpetrator of a violent crime against another individual; or (2) identify remains of a deceased individual." And the revelation that cases could get solved this way has led to questions of how private companies with DNA databases should be keeping their data. Currently, we don't know how often cases get solved like this. Although a spokesperson for 23andMe told the New York Times that they've received "a handful of inquiries over the course of 11 years" from law enforcement. He claimed that data was never handed over.

Criminal investigations aside, it's regular practice for both 23andMe and AncestryDNA to share data with medical researchers. Though participants can opt in or out. And in 2018, it was announced that the pharmaceutical company GlaxoSmithKline invested $300 million in 23andMe for drug development with the company's resources.

We have privacy laws in the U. S., but a lot of this is still ambiguous.

When it comes to your personal privacy, the best thing you can do is try to be as informed as possible about what's happening to your data when you put it out there. And all the data out there means there's a lot of information that can be stolen. And yes, better technology DOES allow for more protections like encryption but it also exposes our data to wider scale breeches. Hackers are after your personal information--- information that can be used to set up lines of credit-- like when Equifax was hacked in 2017. They also want your photos (remember the iCloud in hack in 2014) Your indiscretions… (Ashley Madison) Your email addresses… (Yahoo was hacked back in 2013 and 2014) Your business files… (remember the Sony studios hack after "The Interview") Hackers have no qualms about cutting off your access to play FIFA--like when the Playstation network shut down after an attack in 2011. Companies and institutions like these that collect our data have responsibility to protect it. But just how much responsibility and what happens when they don't. We're still figuring that out. We don't want to let our excitement about Big Data to outpace our caution. We don't want to be like the scientists in Jurassic

Park: so preoccupied with whether we could, and not stopping to think about whether we should. As a society we need to think about and implement solutions to the problems big data creates. We want to use for good not not good. Thanks for watching. I'll see you next time.