CIS399: The Science of Data Ethics	Lecture 22: April 9, 2019
Instructor: Michael Kearns & Ani Nenkova	Scribes: Haroni Amare & John O'Farrell

1 System Data

In looking at high level data, we have spent a lot of time looking at securing personal data. In looking to secure personal data, we largely focus on guaranteeing that the data being used in these studies will not harm the people who choose to donate their data. Differential privacy is very useful when it comes to preserving data because people feel more secure about giving up information if they know that they will not have to face risks for giving up their data. However, in addition to looking at personal data, it is also important to look at system data. Analysis of system data focuses on making sure that the data used in the study is accurate and representative of the topic at hand.

An example where data can be unrepresentative is when people have to self report their data, as this can lead to inaccurate data. For example, if a person is asked to recall what their diet has been for the past few days and they are only going off of memory, then there is a chance that they will forget about specifics while self-reporting, leading to inaccurate data. Furthermore, some aspects of data can be difficult to self-report on, due to stigmas surrounding their answers. For example, a study was conducted in which subjects were hired randomly and asked to self -report their opinions on vaccines. The subjects were randomly assigned the type of survey, where some surveys said that the flu shot was mandatory, others just asked if the participant got a flu shot. Going against the expectation, the survey found certain groups far more represented in the study than they should be.

Sometimes the personal data is the data that the system is trained on, resulting in an overlap between system and personal data. For example, autocorrect relies on data that is entered by the user, which means that the data involved can be both very private, requiring privacy protection, and potentially more representative of a person that other methods of collecting data.

2 Privacy on the Internet

Another concern about differential privacy is the risk it poses to users who can unwittingly have their data stolen. As recently as two weeks ago, a sandal emerged in which facebook was caught actively buying data from apps in order to harvest users information. The amount of information stolen from apps has attempted to be measured in studies. One such study used a multitude of random android apps, in which it tried all combinations of 5 letter beginnings and harvested the autocomplete options. The study used each query along with autocomplete to use static analysis to see if there are any other domains where information is being sent to. By looking at these domains, which lead to certain software or companies, one is able to get a sense of where else data in an app is being sent to. The results from this study can be show in Figure 1. The conclusions shows that an incredible amount of apps have connections to other software, such as google, facebook and twitter.

Super genre	Number of apps	Med	Q1	Q3	> 10	none
News	26281	7	4	11	29.9%	6.5%
Family	8930	7	4	11	28.3%	7.2%
Games & Entertainment	291952	6	4	10	24.5%	7.3%
Art & Photography	27593	6	4	10	16.8%	3.6%
Music	65099	6	4	8	13.5%	4.1%
Health & Lifestyle	163837	5	3	8	15.4%	9%
Communication & Social	39637	5	2	8	16.2%	13.4%
Education	79730	5	2	8	13.3%	11.9%
Productivity & Tools	265297	5	2	8	11.9%	13.5%

Figure 1: Logging where apps send information. Almost 30 percent of news apps send information to over 10 different domains, while only 6.5 percent do not send any information

However, it is possible that these pings to different domains is not directly autocorrect data being sent to these sites. For example, apps that use either a share button or a log in through Google or Facebook page would also transmit data. As such, these percentages work more as an upper bound, as there is no guarantee that all of this information is being directly connected, but if is very likely that there is some data that is being collected and harvested.

Other analysis from the study gives more insight into the domains where data is sent to. One of the other breakdowns of the data found that 88% of apps that are sending datahave a reference to a domain in Google, 43% to Facebook, 34% to Twitter, and 26% to Verizon. In addition, figure 2 details the locations of the domains that these apps are sending their information to. Both of these charts help illustrate how data is transmitted from one app in ways that the user might not necessarily be aware of.

Country	# Apps	% Apps
US	865369	90.2%
China	48451	5.1%
Norway	30674	31.2%
Russia	24889	2.6%
Germany	24773	2.6%
Signapore	19323	2.0%
UK	14451	1.5%
Austria	4754	0.5%
South Korea	3366	0.4%
Japan	1801	0.2%

Figure 2: Locations of the domains that are receiving information

As the internet ecosystem starts moving toward more control by platforms such as amazon, IOS or google, there is a chance the privacy could actually become a competitive advantage. With privacy scandals abound, as more services start to make a living on these platforms, there is a chance that these platforms can start to compete when it comes to digital privacy, as platforms would prefer to not have a reputation as data thieves. However, there are still privacy scandals to this day, as some of these platforms seem willing to compromise users privacy for access to more data.

3 Predictive Applications in Medicine

Now looking to other applications of machine learning technology, uses in medicine is important to look at as many of its problems discussed in class have risen in this context. We will now look at two specific examples in pancreatic cancer detection and prostate cancer treatment.

3.1 Pancreatic Cancer Detection Technology

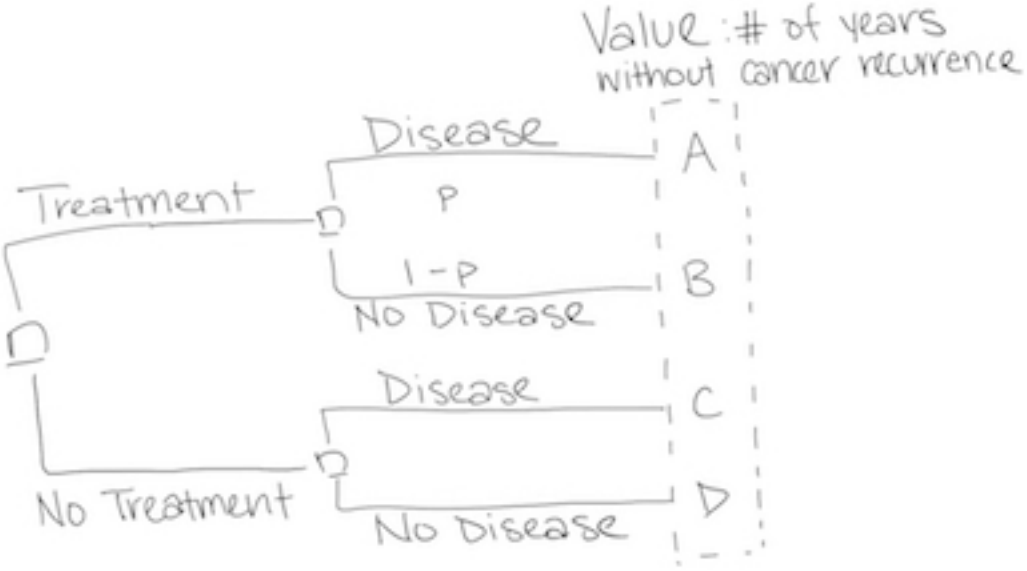
With pancreatic cancer, the earlier it is detected the higher chance the patient has to survive. So, to begin this study, the researchers first investigated early symptoms of the cancer. They continued by digging through search logs provided by other companies to see the number of people looking up those early symptoms, followed by later searches to see if they were diagnosed. They determined if they were diagnosed or not by seeing if their later searches included statements like: "just diagnosed with pancreatic cancer what to do next." Afterwards, the researchers presented a classifier which predicts searcher's demographics, location and risk factor. In this particular case, it was especially emphasized to minimize false positives.

The classifier gave a very low false positive rate and achieved a success rate of about 30%. This case was one where its invasive nature provided life-saving results, an example of the potential benefits of lowering privacy.

3.2 Prostate Cancer Study

Set Up - There are two types of operation: one which is more invasive, has more side effects but less chance of cancer recurrence, and another which is less invasive, has less side effects but higher chance of recurrence.

In order to treat the cancer, the doctors need to know how much it spread, which has to be found out during surgery. Below is a decision tree which outlines all possible outcomes:



Looking at the specific variables, p is the probability that the patient has the disease, and 1-p is the probability that they do not. A, B, C and D are the values of the four outcomes which are

represented in number of years the patient has without cancer recurrence.
We now define p_t as the probability of having the disease when both courses of action (treatment or no treatment) are equally valuable. This gives us the following equation:
$(p_t * A)(1 - p_t)(B) = (p_t * D) + (1 - p_t)(D)$
We can then manipulate the equation above to get us to the proportion of probability of disease vs no disease:

$$\begin{aligned}(p_t * A) - (p_t * C) &= D(1 - p_t) - B(1 - p_t) \\ p_t(A - C) &= (1 - p_t)(D - B) \\ \frac{A - C}{D - B} &= \frac{1 - p_t}{p_t}\end{aligned}$$

Here, A and D are the benefits and C and B are costs. More specifically, A (true positive) and D (true negative) are the benefits of taking the right course of action with the presence of the disease or lack thereof. B is the penalty of giving treatment when not needed (false positive), and C is the penalty for not giving treatment when needed (false negative).

In next lecture, we will look at curves of net benefit using true positive and false positive values.