# Video #1:

## Slide #1:

This is PDAT 625G Big Data Security and Ethics -   
 Module 3: Privacy in the Big Data Age.

## Slide #2:

This module walks through a couple of different topics, one is what is privacy anyway, then we'll think about how big data, in particular, makes privacy into a much more complicated issue than it would be otherwise.

Then we're going to pivot into the general idea of why we care about privacy, both as consumers and as people who take care of other people's data, how do we protect our own data and what can we do with the data that we have in our care.

### Slide #3:

Australian Roger Clarke was one of the first data privacy advocate, as early as 1984, before anyone used the term "data scientist." In 1997, he made a list of "four privacies" that we should be concerned with.

Since that time, people have added to the list, expanded it and some would say there's actually seven privacies. I've actually stuck those underneath the original four so that we can talk about them.

The first one is "Privacy of Person" This is the oldest of the privacies, the one we think about regarding physical search warrants and ones "personal space." To violate this, even a police officer needs a warrant. You generally are thought to have autonomy over your own person. The same is true of your house and other spaces.

In data times, we expand this to the idea of secret GPS tracking, and your image.

If you go into a bank, they will record you, but we expect them to be careful with that video. This has changed over time, and we know that legal things are still catching up.

There's generally been a sense that you aren't private when you're in public, which sounds silly when I say it like that. I'm making this video at my house and putting up on YouTube, so I can't then be mad when someone knows about my painting.   
It's nice, and not private.

The second privacy, which again comes from an ancient idea is the idea that your own communication, your own correspondence is private. Ben Franklin, you might remember, was the first Postmaster General, under President Washington, and you might think that's a lame job for someone so famous. To be fair, he was a generation older than those guys in Hamilton (he's not even IN Hamilton), so maybe he was too old to be Secretary of State. But, he didn't take it as a cushy job - he really thought this idea that the US mail needed to be kept sacred and the postal service needed to have this impartial role, the idea that your mail would never be searched or read and that even who you sent it to would not be a matter of public record. This was expanded in the early 20th century, that they needed a warrant to tap your phones.

Taking that into the Internet age, your emails should be private. Gmail famously uses "double-encryption," so it's very hard to read your mail, but if I send someone a private messenger in Facebook, is that really private? Certainly not if I post on their page, since that's the point. If I send someone a text message, a DM, how safe is it?

What about my work email "@truman.edu" Can Truman read any of my email if they want? Do they need a reason? That's tricky.

The third privacy is the privacy of personal behavior and action, which grows out of the idea that you have the right to your own thoughts. You go to jail for actions, but not for your thoughts, and as a rule, we think you're free to think whatever you'd like. This is related to freedom of speech, and it's important to.

Privacy of Association is a trickier one, and Roger Clarke didn't write about this in 1997. Are your associations private? Can you connect with your priest without anyone even knowing you are part of that religion? Again, not if you go to the public church space. Can I have a secret friend? Not if they are my Facebook friend? LinkedIn is built around the idea that our connections are public, and you can directly connect to my associates without talking to me first.

Privacy of Personal Experience is one that isn't protected very well these days, the idea that what you did should be private unless you've chosen to share it. Librarians zealously guard what books you've read, but does Amazon?

Last and not least is the idea of "Privacy of Data," the idea that the data that you've generated should be private, unless you make it public. Roger Clarke certainly thinks that companies like Amazon and Netflix have a duty to protect your data. Of course, you clicked the box, so you allow them to do some things, but still, this shouldn't be a blanket waiver.

## Slide # 4

How does the expansion of Big data make privacy even more complicated than Roger Clarke imagined in 1996?

Of course, it's big. So, there's a lot, and it's harder to be careful.

There are three kinds of predictions that we think big data does -- the article you read (or will read) is going to go into those in more detail.

1) Consequential predictions, the ones where we try to anticipate consequences.

Right, so a doctor says, "If we don't operate now, this bad thing might happen. Of course, if we do operate, this other bad thing might happen" It's consequential. "Here's why you should quit smoking" didn't need big data, but we can make very specific predictions about consequences from these massive data sets we have.

2) A second kind, called "preferential predictions" is the idea that we want to find things that you would "prefer" so Amazon netflix spotify all those things are trying to predict what your preferential PICs would be the third and the one.

Big data can go deeper, by connecting pieces we didn't ever know about. If you subscribe to this magazine, and if you purchase this thing from Amazon and watch this show on Netflix, then, we won't send you ads for this candidate, because you're never goint to vote for him anyway.

3) That leads us into the third kind, that, I think, raises the most worry is the idea of a "preemptive prediction." It tries to get out ahead of you, and it tries to "change the future," maybe even limit the options that you see, to steer you in a good direction. But, that means that you might not even get to know what choice you didn't get tomake.

It could be preventative, like the "no fly list," that identifies people for special screening or even excluding from flying altogether, based on the idea that you might be a terrorist. Not even "probable cause," but "possible cause" that keeping those folks off of the airplane is a good idea.

In a job interview process, the keyword and other screening done by the algorithm tries to narrow down the applicant pool before a live human even touches it.

These three kinds of predictions can all make us worry, but especially the third, since we won't even know what we won't know.

Another complication is that, we know that your social security number, your bank account number, your password are all ways to get lots of information about you, but Quasi-Identifiers are sets of public data that can be put together to figure things out as well.

In 2013, Harvard Professor Latanya Sweeney found that for a sample of people she studied, if she had their gender, birthdate and zip code, she could exactly identify 87% of americans exactly. I can only imagine that it's gotten easier since then.

So, even beyond your "truly" private things, we should be trying to keep as much as we can private.

The rest of the module will focus what we can do to take care of the data we are working with, what principles and techniques will help us keep private things private and secure.

# Video #2:

## Slide #5

What can we do about it?

I think you've already looked at the acm code of ethics and the Statistical Code of Ethics - they are actually very similar.

The Harvard Business Review actually has a broad idea of what data ethics might look like, and the data science association is a newer group that also thinks about this.

All of these documents are great to ponder, what we stewards of data, how we do the best we can.

How can we act in a way that is responsible, ethical and achieves the goals of what we want to do with the data? If you look across all of those documents, I've boiled it down to five ideas, and we can walk through those together in the next slides.

Do note that we are not priests, we are not lawyers, we are not doctors or counselors.

We do not have any privacy protections.

But we do want to make sure we're careful not to say what we shouldn't.

Let's talk about these five ideas in turn, with a special eye to the last two, confidentiality and anonymity.

## Slide #6

The first three aren't really the focus of this module, but I do want to mention them here, the idea of being careful and intentional with our data.

We aren't just haphazardly throwing things together, but we have a plan.

Physical security as part of that. Do you keep your data in a locked room?

Do you have a good password? Any clue at all?

There's an idea that computer scientists talk about, "privacy through obscurity."

Oh, the data is on a thumb drive but it's in my pocket and so no one would think to look at my pocket or it's in the glove box of my car or something like that well that's not good.

The idea that it is somewhere physically secure.

Picking a good password -- again this class isn't about that, you know it's important.

Keeping data offline can also help. So often, we keep everything in the cloud, but maybe we shouldn't keep raw data there. Now, in practice, sites like Google Cloud are as secure as my computer, but now my password becomes even more important.

There certainly are cases where keeping the data on a thumb drive and only plugging it into your computer when you need it makes more sense, along with turning off your computer when you aren't using it. There certainly are industries and companies that have a separate offline computer where secure analysis is done, or where the "intranet" is kept separate from the "internet"

Now for most of what we do, you don't need to have James Bond level security with Retinal scans and three passwords and physical keys, but even thinking about what kind of data you have might help you be more careful to protect other people's privacy.

A second key point is being aware and open about weaknesses in the data we've talked about in our other data science classes. Data is never perfect, and we always build in assumptions about what's missing. We do that for our conclusions, to be sure, but we also do it in our data cleaning. Often we don't even know what we don't know.

I'm a data collection person by training, and I'm always an advocate for being more thoughtful in how we collect the data, in order to make it easier to store the data securely (and of course analyze it).

Sometimes, we don't even have any control over how the data is collected, it's collected by someone else but if there are ways to design the system even just a little bit better, we can see a big improvement.

The third point is the idea of being transparent. If we're clear how we kept our data, then we don't have to worry about accusations that someone has fiddled with it or tampered with it. We know who has touched it and who has cleaned it, not just for reproducibility, but for privacy.

Can our methods be public even when our data is private? For analysis, we know it can be, that's the whole point of data reproducibility. Can we share a "fake" data set to show how the analysis was done, without relaxing our privacy standards? A baby data set of just a few data points that would show somebody how your data set works and how your script, your reproducible R code so that they see what's going on.

These are three of the five points and again we're going over them very quickly because they're not really the focus of this module but certainly you don't want to forget them as we do.

## Slide #7

Next up is the idea of being confidential.

All of those various principles that I pointed to a couple slides ago, point to this idea that we have a duty to do our best to keep data secure and private.

Again we don't have the legal protections that lawyers and priests and counselors and those kind of folks have, but we do have to do our best.

Technically, we can talk about the idea of making pseudonyms which the process we call pseudonymization, which I always have trouble saying.

As you could imagine it's pseudo-anonymous because typically you have a key that you can go back and figure things out right so just taking everyone's name and replacing it with a number.

Typically, when we make a code for those kind of identifying information we do keep them somewhere safe. Sometimes, though we throw it away and that would certainly make it a little bit more anonymous but typically, we have the key somewhere behind it, so we can go back if we need to.

Now, I can tell you, for instance, when we do pseudonyms for things other than identifying information, for instance, what if, instead of having male and female.

You just list one, two, and three instead of male and female, and other/prefer not to say. At Truman, students can easily figure out which one is which, our campus is about 60% female and 40% male, with just a handful of people identifying with a non-binary gender.

In survey responses, we also know that, overall, women tend to have higher response rates than male. So if you just looked for the bigger bucket it's pretty easy to tell who are women, and the smallest bucket are those who did not identify a binary gender.

Of course, if you're doing something regarding race, because we have a good idea of the proportions of different races, not just at Truman but in Missouri or in the United States or in a workforce. Anonymizing those doesn't really help all that much there.

Still, you can imagine cases were making pseudonyms for divisions or different parts of a company or different product lines, there are cases where that would be useful, but normally when we think of pseudonyms it's really about identifying information.

If you do go with a system that is going to be reversible, meaning that you, you know, keep the columns of pseudonym so you can go back one to the other you do you want to keep that in a safe place sometimes people will do. A formula, a key, or something. There's lots of ways to do that digitally.

Ideally, you do it in a way that regular people can't easily figure out, but especially if you need to keep track of it, you keep track of the algorithm so that you could go backwards later.

## Slide #8

Now a technique that falls somewhere between confidentiality and anonymity is what we call dynamic data masking, where we limit what each person in the pipeline can see.

And typically we get an impartial person or just the person who does the data management who can can see all of the data. You could imagine an eye doctor which has a variety of information about you - often they have your full medical history, not just your vision data. We know the data science has a big role and companies like Cerner and Express Scripts here in Missouri are working on this kind of secure sharing of medical records.

So, the receptionist who sets up your appointment, greets you at the office, and has you fill out paperwork. She doesn't see your actual medical information. She really sees that you have an appointment today, she might have the forms that you fill out because she puts them in a file before they get entered. She might even do the data entry for those but she may be can't pull that data out once it goes in or see your other records.

The vision tech the person who does your initial screening, in some cases, that person can write the data they collect, but can only see a little bit of your data - the information they need to do their job.

For instance, if you're diabetic or you have problems with glaucoma or those kind of things, you do want the tech to know, but can you keep it as secure as you can.

Even the doctor doesn't see the whole picture, typically the doctor can see all of your medical information, but they don't see the financial parts. This is good for a couple reasons you don't want to get less good care, just because you owe a few dollars.

Or you don't want the doctor to pester you about it, but even from the doctor's perspective it's a distraction, and so having too much information when you're in the room with a patient. It slows you down and might keep you from making good quick decisions.

Of course, there are financial people who can see the financial information, and they can see which procedures, since that's what they bill, but they can't see your test results or they can't see your actual medical documents.

That leaves only the tech people who can see it all.

In practice, they don't, because they don't actually care - it's not what they need to see, but we have to trust them.

But, do we keep track of that person, how could you make sure that person is secure. Do we log when they do look at the files? Are they the same person who manages that log? If you remember Jurassic Park, the tech guy, the guy from Seinfeld, messes up the whole place.

Now, you can imagine fancier systems where you need multiple people, multiple passwords or multiple steps in order to see everything. So maybe the central tech guy can't see your raw data, he sees some encrypted version of it.

We sometimes talk about nuclear keys like from all those old movies, where the two keys had to be in sync and turn their keys together, so that you could launch the torpedo or nuclear missile or whatever.

And there are systems that work that way they're a little more complicated and again depending what you're studying that might be a system that works for you.

Certainly, the gold standard is to have true anonymity. Statisticians, we tend to think about this in reporting information, but computer scientists and computer engineers think about the raw data itself, and that's the next video in this module.

# Video #3 –

## Slide 9

The ideal solution is to have anonymity where we can't share raw data even if we wanted to. One simple form of that which we call differential privacy, that is, different people have different levels of privacy.

This is what I do as a statistician when I report my results. so even if i've seen the individual case names and I might know which person goes with which data. But, my report doesn't, and my charts just give overall summararies. Nowhere does it say "this one is Bob" and if I did it correctly, you can't figure it out either.

So, it's confidential to me, but anonymous to everyone else. Combined with having a separate person anonymize the data, like we described on the last video, it's probably the easiest way to have good privacy.

If you want to think about what you're sharing in an intentional way, we use an idea called "K anonymity."

This is where you literally go in and change the data set before you run the analysis, to be sure that no one can figure it out. K just means the lowest number identified - it doesn't have to be too big, just 5 or 10 is plenty. We call it K anonymity because K gives us sort of the dividing line answering "how anonymous is it?"

One technique used is suppression, where you just delete columns with individual answers that are never duplicated. You, you delete the column with the names, with open-ended responses, that kind of thing. You could replace it with an ID, but typically, you don't even do that.

If you are doing open-ended text analysis, you don't throw them away, but you put them someplace totally separate where no one can combine them again. This does limit analysis, since you now can't say, "This comment was from a female respondent," although you could pull selected variables with the text - that's less private, of course, but if your main interest is about gender, maybe you need to do that.

There's also the idea of generalization. This is where you combine attributes rather than showing raw data. You give kind of a range so, for instance, instead of having an exact date of birth, you just give a range, 18-25 years old. Remember that date of birth was one variable that Tanya Sweeney could use to find out so much about you.

Instead of looking at major at Truman, we look by school or we look by kind of major, so we look at STEM or we look at humanities majors. Maybe you split customers up by region, by kind of purchasing, or by some other group characteristic, even though you often think of it more specifically "in the field."

Now, people in the Spanish program want to see what Spanish majors look like and sales reps want to know about their exact clients, so maybe you do have the data both ways somewhere, but in the data that's analyzed and certainly what's publicly available or even the data that shared with other an analysts you put in this K-anonymity.

We have to watch how we generalize, because sometimes you have to smoosh so many things in order to get to K, it becomes a little bit absurd. Just making a big category called "other" is easy, but often not very useful. You can imagine at Truman, we have a handful of "big majors," Business, Psychology, Biology, Health, whatever, and if K is too large (say 50 graduates per year), we'd end up lumping everything else into a huge "other category," containing Nursing, Statistics, French, and whatever else. Not very handy. Humanities, Other STEM, or whatever, would be much more useful, but it takes a bit of a human touch to pull that off.

K anonymity is not perfect, certainly someone with some background knowledge might still be able to figure out what's going on. That idea, you know the black woman physics major even if you smash that into the black woman stem major.

Even if you get it to K=5 anonymity, you can fill figure it out when you look across several variables (k-anonymity is done on each variable separately). So if we say oh it's the non-traditional student who's a woman of color, even though each of those have K anonymity for each, someone could look at both of those variables and figure out who the person is exactly.

There's also the problem of homogeneity, which is the idea that we can look at the patterns and figure out what's going on. Of course, that's the purpose of an analysis, but you have to watch that you don't accidentally limit the privacy.

## Slide #10

Going to the next level, we can think about epsilon differential privacy. If you've had higher math classes, you may have seen that we use epsilon as the Greek Letter to represent changes that are so small they don't matter (as opposed to delta, which we use in Calculus to mean "large enough that it does matter." Mathematician joke: we would have preferred the epsilon variant of COVID to delta, since it would have been a much smaller change, but we didn't even notice it.)

With epsilon differential privacy, you literally alter the data, but in such small increments that it doesn't really matter. It definitely confounds the snooping algorithms that try to identify people, but it could confuse us, if we aren't careful.

Imagine, we have everyone's weight in a health database, and we just use a random number generator to add or subtract a little bit, maybe with mean = 0 and standard Deviation 3. So, many people will have changes in their weight from 0-3 pounds, and about 1/3 will be altered by more than 3 pounds, but in a random direction. About than 3 per thousand will be altered by 10 pounds.

Data isn't exactly right or we're going to add a little bit to your birthday, so will make you up to two months older two months younger, that would be enough to foil Tanya Sweeney and her algorithm, since MANY people have the same birth "season." The summaries will still be correct and any deviations from the correct answer are going to be random and hopefully small enough that they don't change your results, but the raw data is not actually the raw data anymore.

This is not always good when the raw data is important, so you'd never do this on a 'live' medical record - you can imagine telling Doctors that you added extra variation in, just for fun, even beyond what you normally see in how data like weight or blood pressure change over the course of a day anyway. They'd hate it.

But I can imagine having Dr. Thatcher doing it to that data we had before we gave it to our consultants to do their analysis.

Taking this a step further, "L" diversity and exponential mechanisms are even more complicated ways that you can anonymize the data in a way that it can't go backwards very easily, if at all.

## Slide # 11:

Now, we can take this even a step further, to use hash functions.

Now, in general, hash functions are how databases store information efficiently. They are used for everything, and this class isn't about that. But, one kind of hash function, cryptographic hash functions, make a true kind of "one-way" storage. Once you've encoded it, it's almost impossible to figure it out, other than by brute force "guess and check," and most systems protect against that.

Passwords are almost always stored this way, and when you hear about data breaches, they'll sometimes say that passwords were hacked, but only in cryptographic hash form, so they aren't that worried about it.

Cryptographic hashes are sometimes thought to be "too much" for normal data storage, since we can't see what we have, but they can certainly be useful. There are also techniques in between, where you can use several keys (the submarine missile launch metaphor again) to unlock the code. There are different kinds, and I think we'll revisit this later in the class, but this is definitely a thing where some corners of data science go much more deeply into this than others.

## Slide #12

Certainly, as you get deeper into this topic, you know, this is a just a one week module in an eight week course, but as you get deeper into this topic, you can certainly find cooler and really interesting ways in order to make your data truly truly anonymous.

The last thing in this module I just want to mention, we always have to come back to what the practical concerns are. If we do use an anonymous technique, if we do use anything beyond data masking, like epsilon differential privacy, or cryptographic hashes, we're actually changing the raw data, and you can't go back. Maybe you use encrypted data for your day-to-day work, but once a year, you make a "raw backup" and file it somewhere. Maybe you decide to keep a "real" copy of the data on a flash drive in the back of your desk drawer, in the cookie jar, or filed at the lawyer's office, but now that it exists, you have a weakness.

Once you throw data away once you change it, you can go back so if you're really going to change the original you better mean it. We statisticians joke that we err on the side of hoarding because we want to keep all the data, because we can imagine doing lots of kinds of cool post-analytics, meta-analysis, down the road.

You have to balance what you're actually doing with the data, what you might do in the future, and what could happen if you lose it. For some things, keeping the real data is totally fine, and these safeguards would really be hurdles.

Some fields have specific rules about what you have to keep for the lawyers in escrow or the accountants, or an in-house notary or just the "team black" person on your own team, and you can be sure that you want to do that, but other fields have specific expectations about privacy, and I'm probably not the person to tell you where your data falls on that spectrum.

Now, we know that James Bond and Jason Bourne know how to use those weaknesses, but it's much more likely that Ned from Jurassic Park is going to be your downfall than a man with a golden gun.

Our Student consulting center CASE, had the evaluation sub-contract on a multi-million dollar oral health grant at AT Still University on the other side of Kirksville, as they were planning to open a new dental school that would help serve rural patients in the region. As part of that, we had surveys from thousands of patients, including actual medical records. On our project, we had a lead consultant Gita Garth, who worked with a student team to do all of the analysis. We had a different team who scanned in the bubble sheets with Dr. Scott Thatcher who got the raw data. They were our "team black" - Scott's role was to anonymize the data for statistical consultants, and assign it a new ID number so that they never saw any actual information, but could do the things they needed to do. He never looked at the raw data to do any analysis, other than what he needed to do to make sure it was cleaned correctly. He kept the key so he could go back if we ever needed it, but I don't think he ever did.

The weak spot in that system was Dr. Thatcher, so all we can do is trust him. Now, this was dental data, not the DNA to dinosaurs, so maybe it was OK. The student consultants were the public face of the project, with Gita Garth. In the end, it worked out great, and it was our largest and most successful project to date.

One phrase we use in consulting contracts is "Willing and competent to testify thereto." Meaning that I know what I'm doing, enough that if I did have to go to court, I could, and everything would be laid out in the necessary level of detail.

Can you have it all? No. Know everything we do as a compromise and we do the best we can, with what we got. Add in the demand for speed, especially with big data analytics, add in the decentralization of data, maybe the data doesn't even live on your central server where you can keep it safe -- it lives out at each of your branch offices or factories.

Maybe you only combine it once a month when you're doing your monthly report or your quarterly report or whatever, so you have to think about how to keep the data secure at the various locations, working with the IT team. Still we need to think about those things, so we end where we began, with the idea that thinking as much as we can, ahead of time being intentional, so that we can get the results that we need.

## Other Module Materials

Course Learning Objectives: A successful student will:

* • identify vocabulary and issues important to Big data security,
* • describe algorithmic decision making and its benefits and challenges,
* • explore techniques and methods for keeping big data secure,
* • critique cases where big data security was an issue, and
* • discuss legal issues for big data.

Module 3: Privacy in the Big Data age - original module topics

o Social costs of big data

* Prediction and pre-emption
  + <http://www.datascienceassn.org/sites/default/files/Prediction%2C%20Preemption%2C%20Presumption.pdf>
* Ethical way of using big data

o Data masking

* Anonymizing
* Pseudonymizing
  + Data Masking
  + Quasi-Identifiers
* Dynamic Data Masking - it's all there, but not for YOU
* K-Anonymity
* Randomized Response

o Differential privacy methods

* Laplace Mechanism
* Epsilon-differential privacy
* l-diversity
* Exponential mechanism

# Sources:

Four (or five or seven) privacies:

Roger Clarke 1997 and revised: <http://www.rogerclarke.com/DV/Intro.html#Mis>

Seven Privacies: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.827.5418&rep=rep1&type=pdf>

ISACA Infographic: <http://chiefit.me/wp-content/uploads/2017/01/Privacy-Infographic_res_eng_0117.pdf>

Harvard Business School Principles: <https://online.hbs.edu/blog/post/data-ethics>

ACM Code of Ethics and Professional Conduct

AmStat Ethical Guidelines for Statistical Practice

Assignment: Read this: <http://psychbrief.com/anonymous-data-r/>

<https://www.storytellingwithdata.com/blog/2020/4/30/five-ways-to-anonymize-your-data>

<https://github.com/EmilHvitfeldt/gganonymize>

<https://cran.r-project.org/web/packages/synthpop/vignettes/synthpop.pdf>