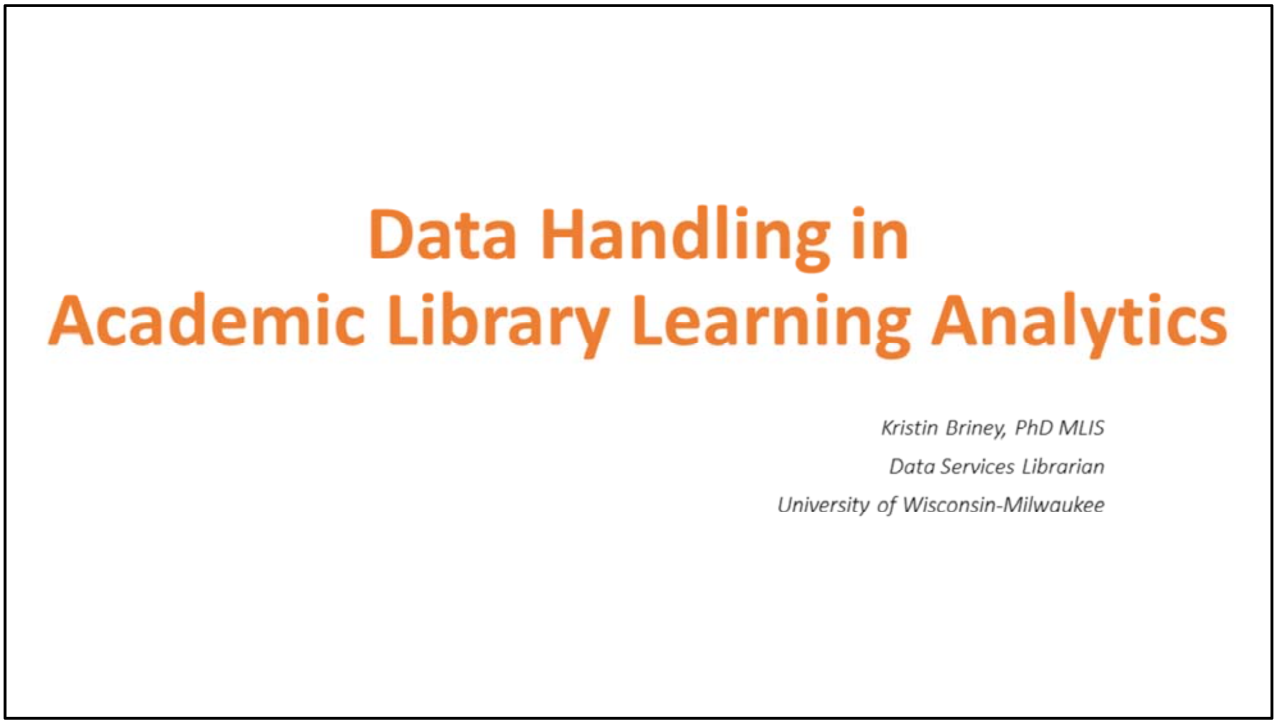
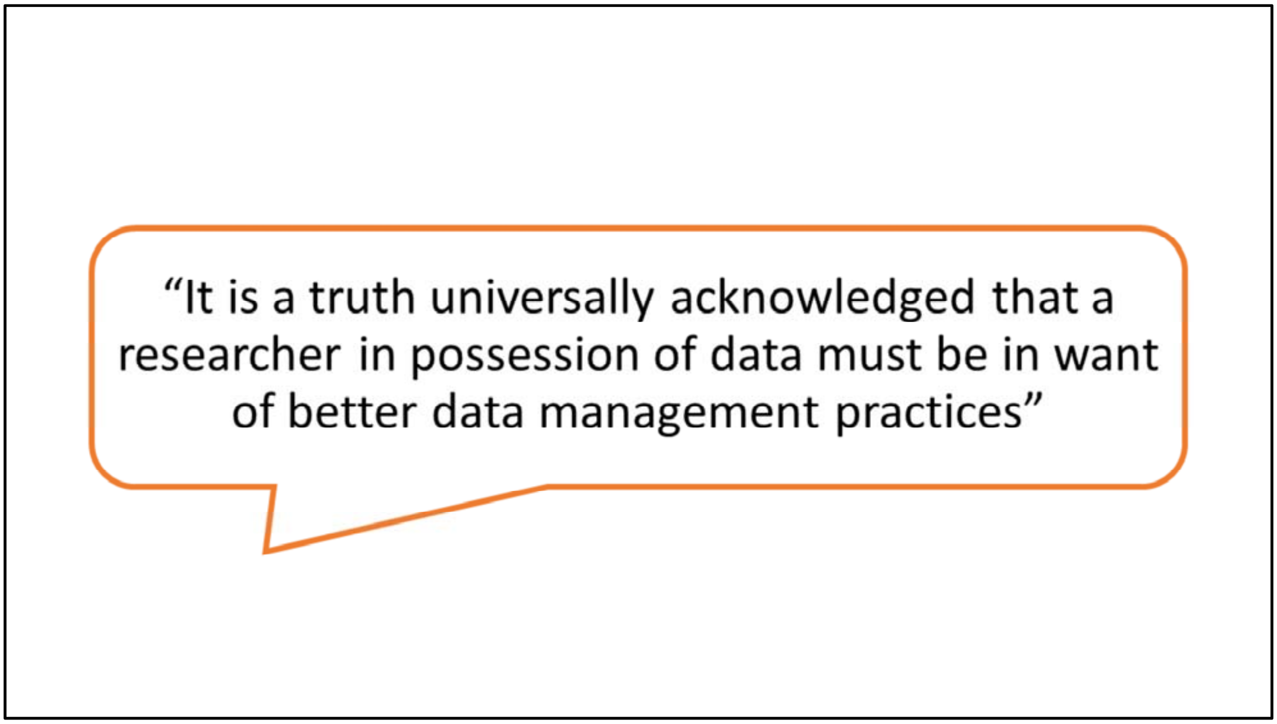


Data Handling in Academic Library Learning Analytics

Kristin Briney, PhD MLIS
Data Services Librarian
University of Wisconsin-Milwaukee

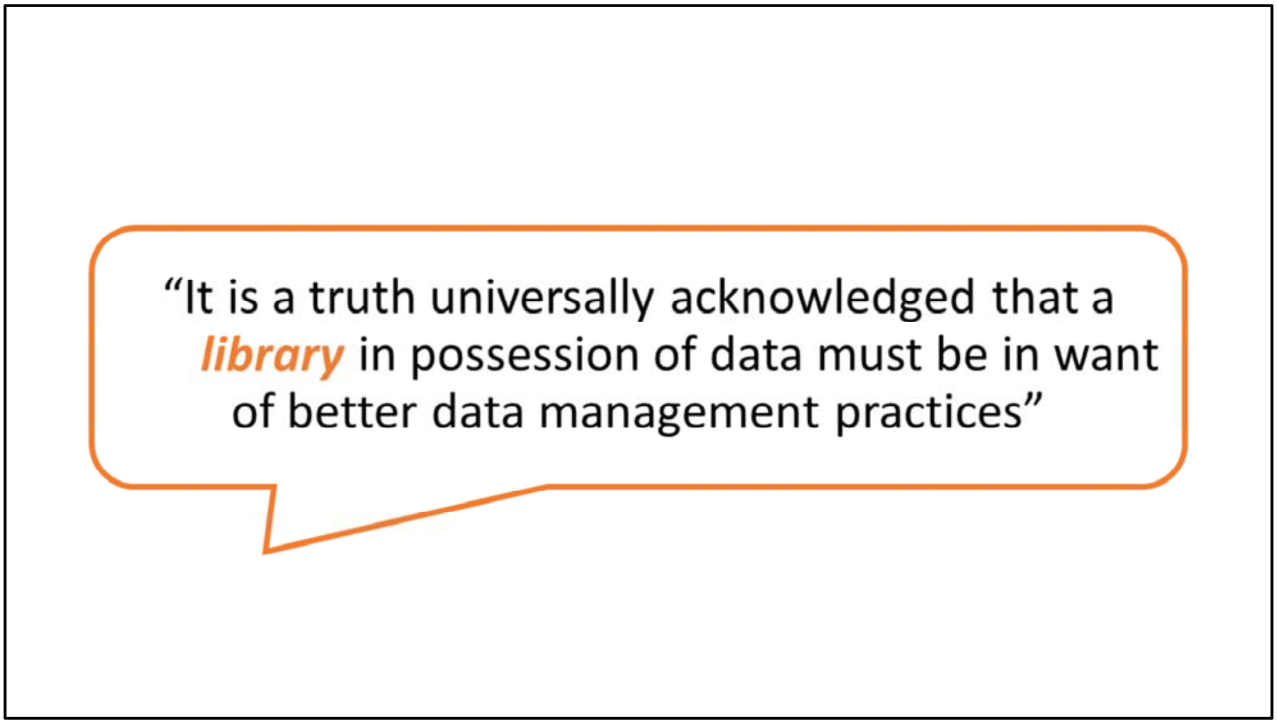
I’m very excited to be here today to talk about how academic libraries are handling data in learning analytics projects.



"It is a truth universally acknowledged that a researcher in possession of data must be in want of better data management practices"

Before I get into the content of my talk, I want to give you a sense of how I approach data management by butchering a quote from Jane Austen: "it is a truth universally acknowledged that a researcher in possession of data must be in want of better data management practices."

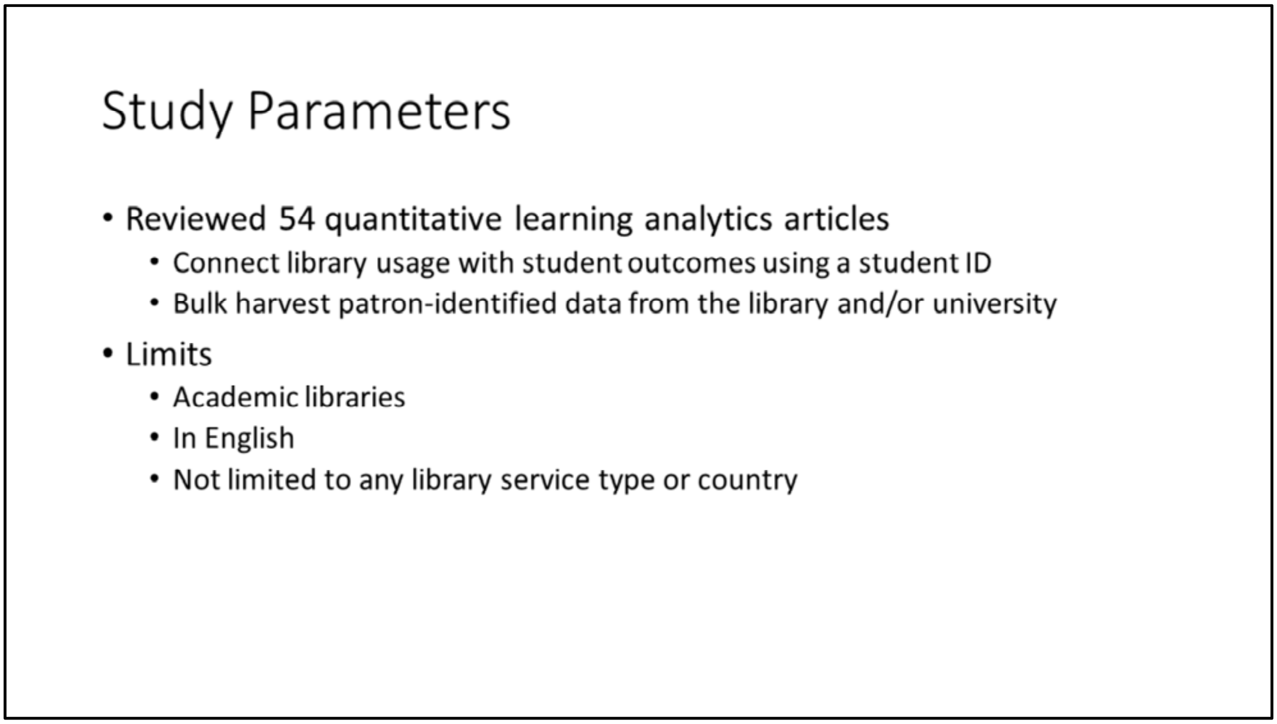
I think that if you talk to a lot of data librarians, like myself, we all have a sense that most researchers need help with managing their data. It's not something that researchers are routinely taught, which is part of the work that we do as data librarians.



"It is a truth universally acknowledged that a *library* in possession of data must be in want of better data management practices"

What I specifically want to discuss today, however, is that "it is a truth universally acknowledged that a library in possession of data must be in want of better data management practices."

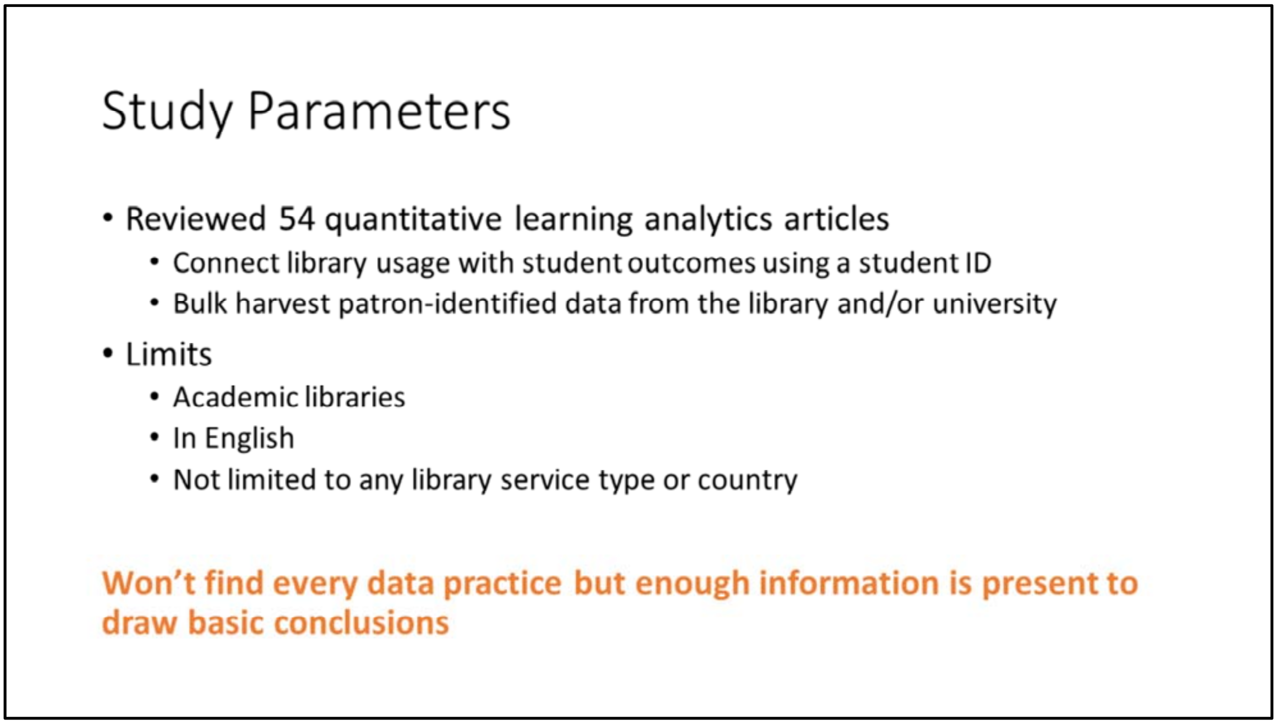
I really don't believe that libraries are any different when it comes to doing research such as for learning analytics. And that's what I want to talk about today.



Study Parameters

- Reviewed 54 quantitative learning analytics articles
  - Connect library usage with student outcomes using a student ID
  - Bulk harvest patron-identified data from the library and/or university
- Limits
  - Academic libraries
  - In English
  - Not limited to any library service type or country

This talk represents a research project that I finished earlier this year looking at 54 learning analytics articles from academic libraries. For inclusion in my research, these studies had to correlate library usage with student outcomes, connecting the data at the individual level using something like a student ID number. There were no limitations on country or library service, although I can’t read anything other than English.

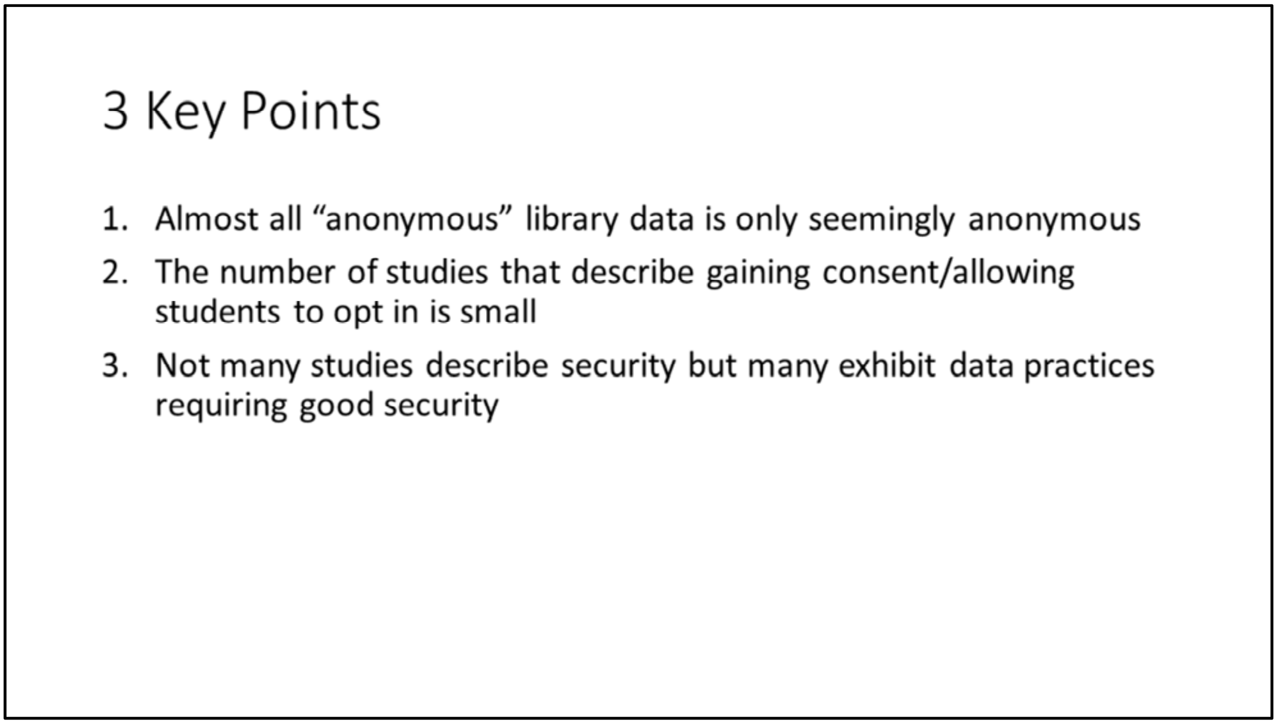


Study Parameters

- Reviewed 54 quantitative learning analytics articles
  - Connect library usage with student outcomes using a student ID
  - Bulk harvest patron-identified data from the library and/or university
- Limits
  - Academic libraries
  - In English
  - Not limited to any library service type or country

Won't find every data practice but enough information is present to draw basic conclusions

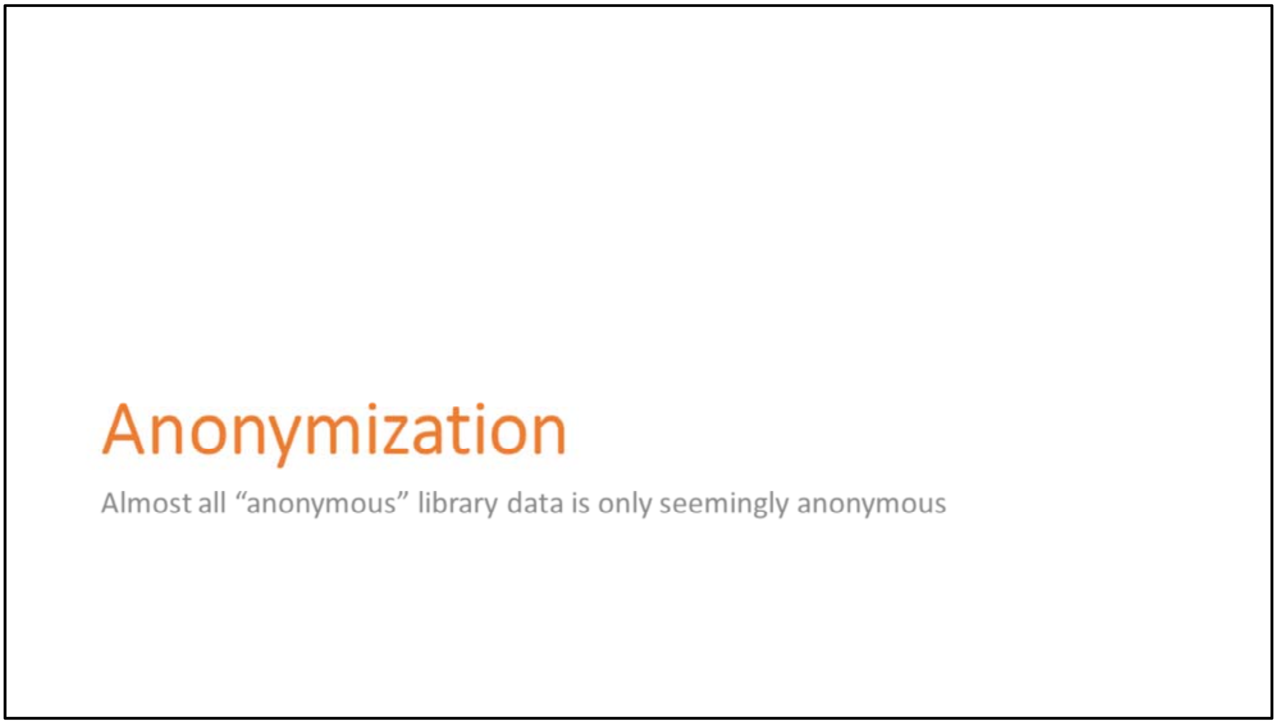
I do want to add that this research method has limitations in that I'm not going to see everyone's data handling practices in a publication; it's just the wrong venue for that. But I hope to show you that there is enough information there to draw some general conclusions which can be a starting point for improving our practices.



3 Key Points

1. Almost all “anonymous” library data is only seemingly anonymous
2. The number of studies that describe gaining consent/allowing students to opt in is small
3. Not many studies describe security but many exhibit data practices requiring good security

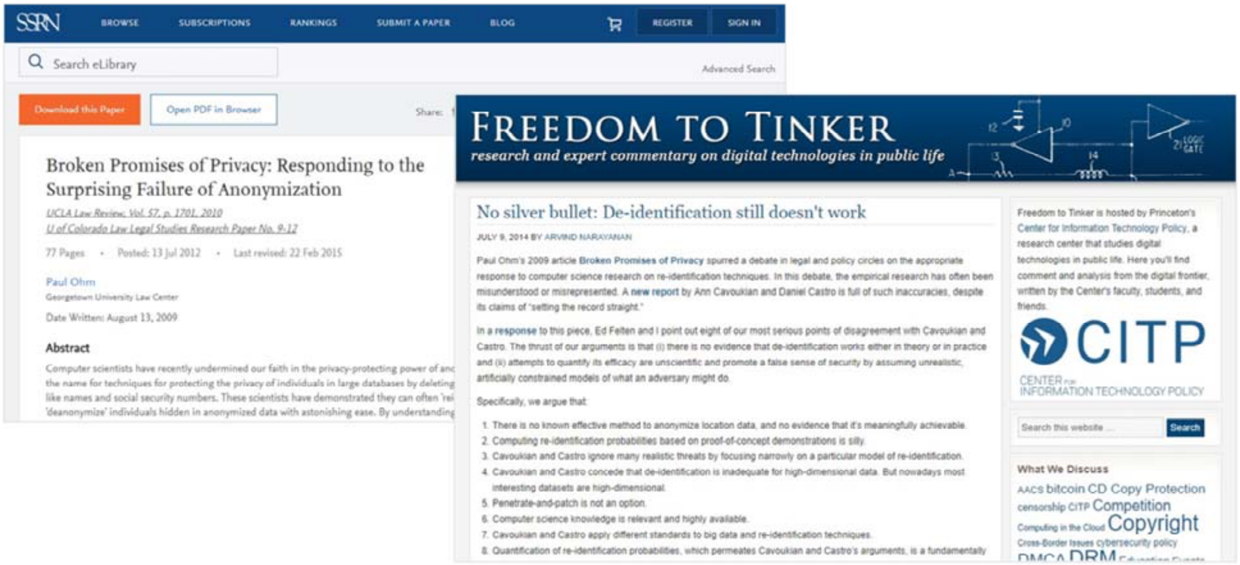
Due to having a limited amount of time, I’m going to focus on three key things in this presentation. First, almost all “anonymous” learning analytics data isn’t actually anonymous. Second, the number of studies that describe gaining consent and/or allowing students to opt in is pathetically small. And third, very few studies describe their security practices but many more exhibit data practices that require good security.



Anonymization

Almost all “anonymous” library data is only seemingly anonymous

So let’s start with the first topic: anonymization.



Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization

UCLA Law Review, Vol. 57, p. 1701, 2010
U of Colorado Law Legal Studies Research Paper No. 9-12

77 Pages • Posted: 13 Jul 2012 • Last revised: 22 Feb 2015

Paul Ohm
Georgetown University Law Center

Date Written: August 13, 2009

Abstract

Computer scientists have recently undermined our faith in the privacy-protecting power of anc the name for techniques for protecting the privacy of individuals in large databases by deleting like names and social security numbers. These scientists have demonstrated they can often 'rei 'deanonymize' individuals hidden in anonymized data with astonishing ease. By understanding

FREEDOM TO TINKER
research and expert commentary on digital technologies in public life

No silver bullet: De-identification still doesn't work

JULY 9, 2014 BY ARVIND NARAYANAN

Paul Ohm's 2009 article Broken Promises of Privacy spurred a debate in legal and policy circles on the appropriate response to computer science research on re-identification techniques. In this debate, the empirical research has often been misunderstood or misrepresented. A new report by Ann Cavoukian and Daniel Castro is full of such inaccuracies, despite its claims of "setting the record straight."

In a response to this piece, Ed Felten and I point out eight of our most serious points of disagreement with Cavoukian and Castro. The thrust of our arguments is that (i) there is no evidence that de-identification works either in theory or in practice and (ii) attempts to quantify its efficacy are unscientific and promote a false sense of security by assuming unrealistic, artificially constrained models of what an adversary might do.

Specifically, we argue that:

1. There is no known effective method to anonymize location data, and no evidence that it's meaningfully achievable.
2. Computing re-identification probabilities based on proof-of-concept demonstrations is silly.
3. Cavoukian and Castro ignore many realistic threats by focusing narrowly on a particular model of re-identification.
4. Cavoukian and Castro concede that de-identification is inadequate for high-dimensional data. But nowadays most interesting datasets are high-dimensional.
5. Penetrate-and-patch is not an option.
6. Computer science knowledge is relevant and highly available.
7. Cavoukian and Castro apply different standards to big data and re-identification techniques.
8. Quantification of re-identification probabilities, which permeates Cavoukian and Castro's arguments, is a fundamentally

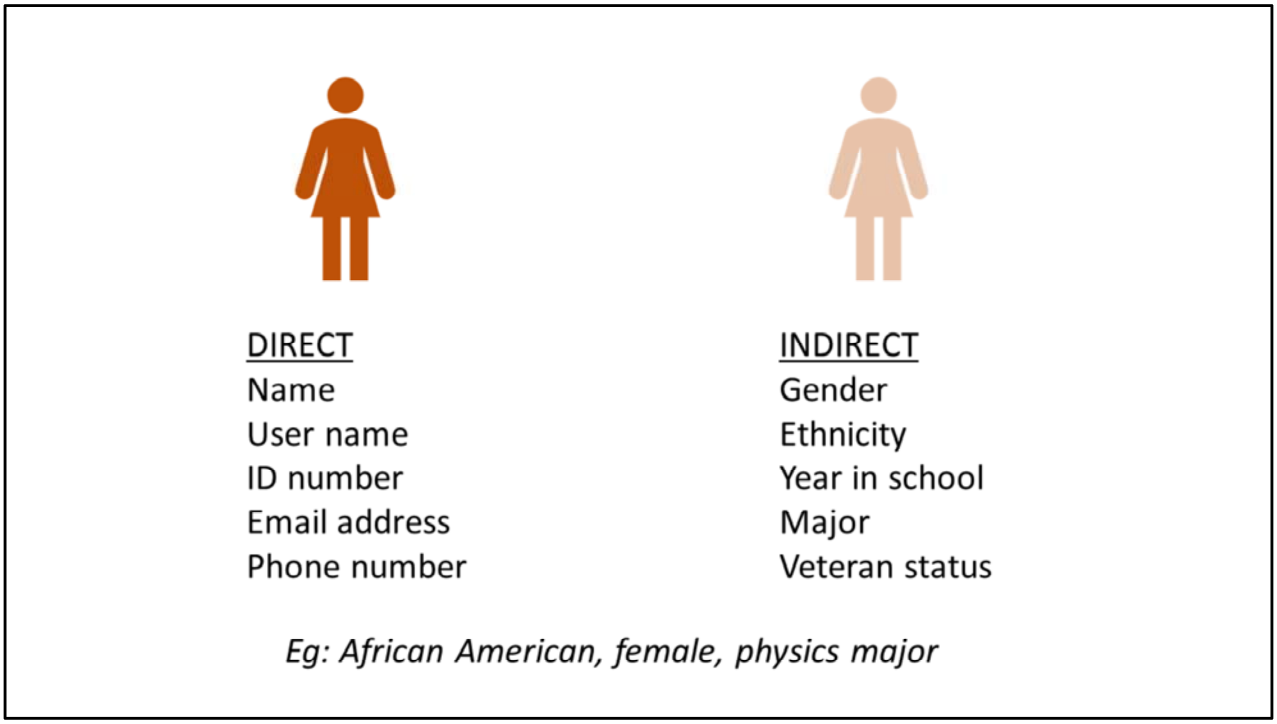
Freedom to Tinker is hosted by Princeton's Center for Information Technology Policy, a research center that studies digital technologies in public life. Here you'll find comment and analysis from the digital frontier, written by the Center's faculty, students, and friends.

CITP
CENTER FOR INFORMATION TECHNOLOGY POLICY

Search this website ...
Search
What We Discuss
AACS bitcoin CD Copy Protection censorship CITP Competition Computing in the Cloud Copyright Cross-Border Issues cybersecurity policy DMCA DRM

Ohm, Paul. 2009. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006.
Narayanan, Arvind. 2014. "No Silver Bullet: De-Identification Still Doesn't Work." *Freedom to Tinker*. https://freedom-to-tinker.com/blog/randomwalker/no-silver-bullet-de-identification-still-doesnt-work/.

I won't be able to get into all of the details of anonymizing data in this session but I want you know that it is difficult. And there is a small but growing voice within the expert community that says that it's actually impossible to anonymize any data that's about people. Individuals have been re-identified by only 4 credit card purchases and with the abundance of existing data, true anonymization is becoming increasingly impossible.



DIRECT
Name
User name
ID number
Email address
Phone number

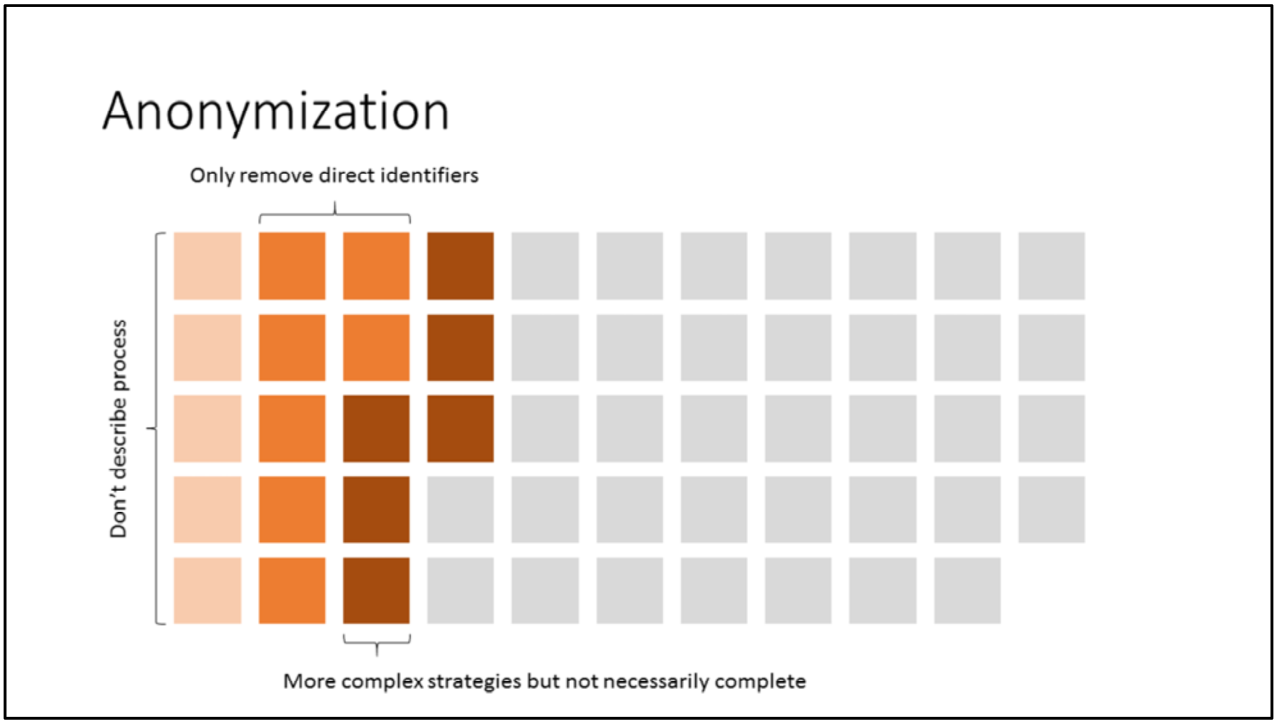
INDIRECT
Gender
Ethnicity
Year in school
Major
Veteran status

Eg: African American, female, physics major

To help you understand the difficulty of anonymization, I want to explain that there are multiple types of data: direct identifiers, indirect identifiers, and behavioral data (which I'm not going to talk about today).

Direct identifiers are single pieces of data that tell you exactly who I am. They are things like name, telephone number, ID number, email, etc. Indirect identifiers don't tell you my identity on their own but can be very powerful in combination. I always like to give the example of African American female physics major, because if that combination shows up in your dataset, you've probably just identified someone! Unfortunately, the most identifiable people in our data are often those already marginalized.

When we anonymize data, we need to be concerned with both direct and indirect identifiers.

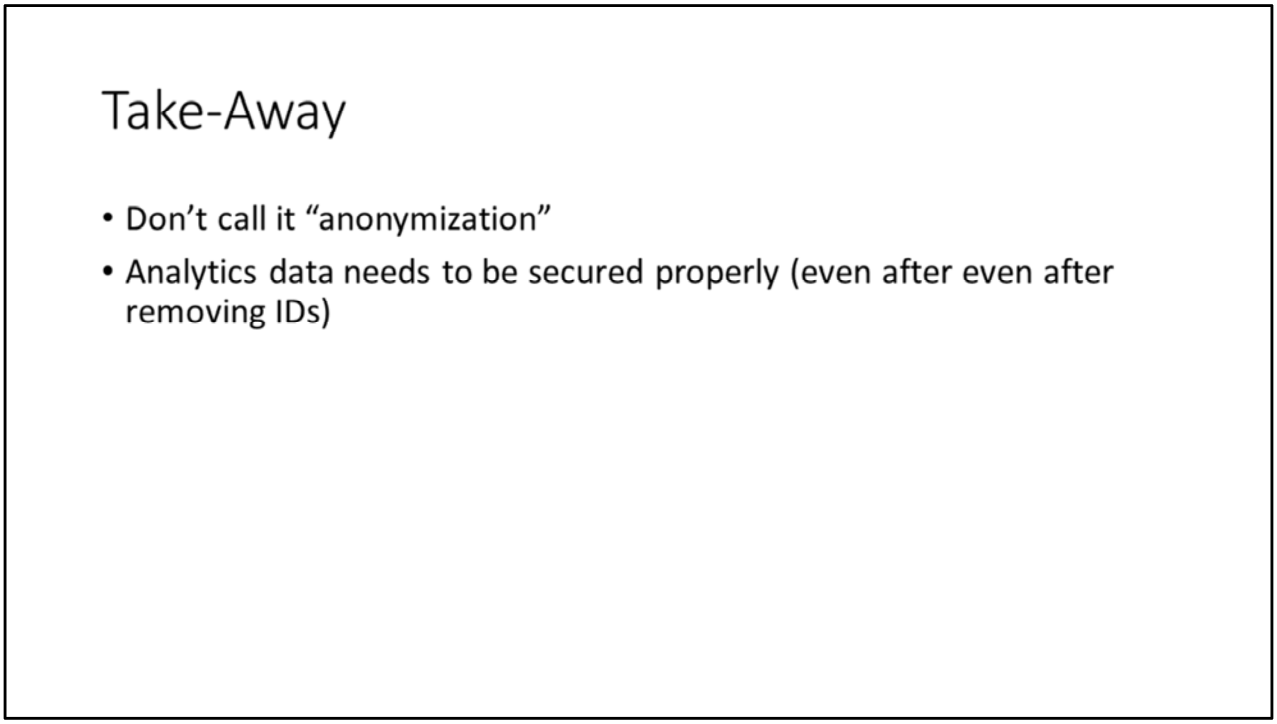


Anonymization

So let's look at what libraries are actually doing. Of the 54 studies I looked at, 18 talked about anonymizing their data.

Of those 18: 5 said their data was anonymous but didn't tell me how it happened; 7 described their anonymization process as only removing direct identifiers; and the remaining 6 used further data suppression methods that don't necessarily equate with full anonymization. There's one dataset in that last group that I think might stand up to more robust re-identification attempts, largely because they removed a lot of data.

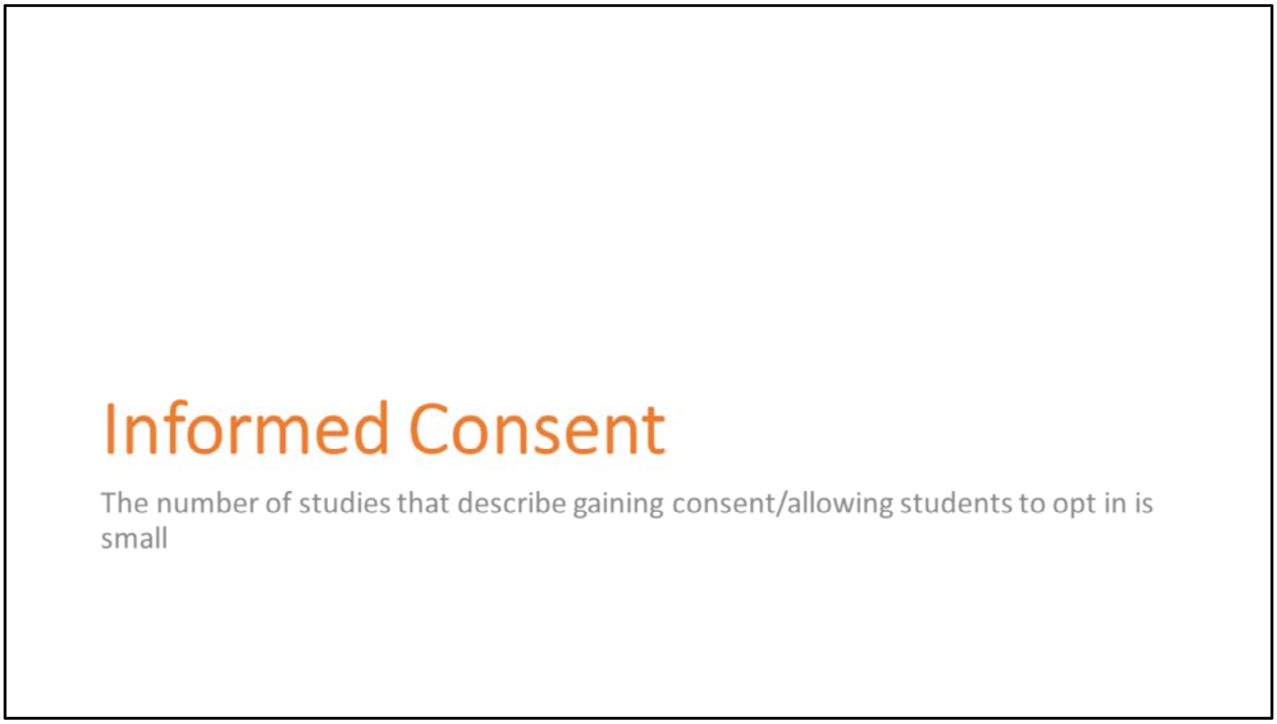
Taken as a whole, this presents a picture of libraries not actually anonymizing their data.



Take-Away

- Don’t call it “anonymization”
- Analytics data needs to be secured properly (even after even after removing IDs)

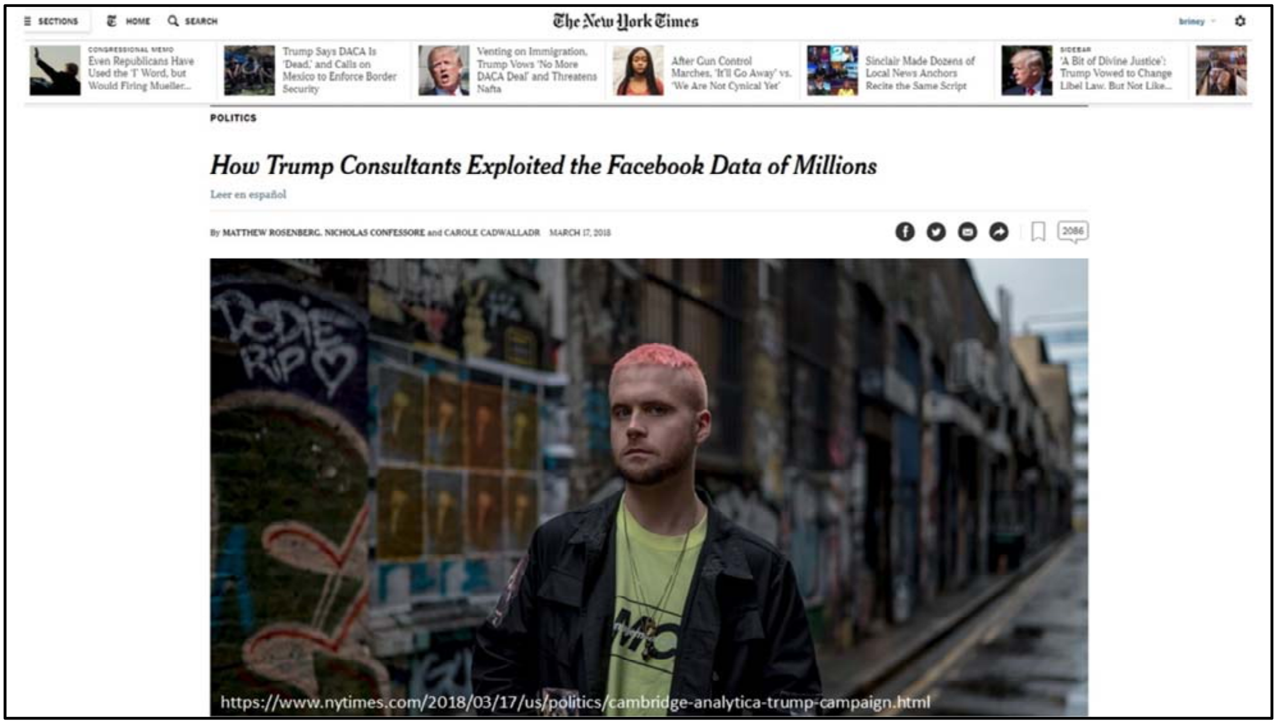
So what can we take away from this? First, let’s stop calling what we’re doing “anonymization.” Second, we need to secure that data. My biggest concern is that a library thinks their data is anonymous when it actually isn’t and they don’t protect it properly. We can keep using existing methods of removing identifying information, but even with the best anonymization procedures, it’s still good practice to secure data about people.



Informed Consent

The number of studies that describe gaining consent/allowing students to opt in is small

Okay, let’s move on to our second topic: informed consent.



CONGRESSIONAL MEMO
Even Republicans Have Used the 'I' Word, but Would Firing Mueller...

Trump Says DACA Is 'Dead,' and Calls on Mexico to Enforce Border Security

Venting on Immigration, Trump Vows 'No More DACA Deal' and Threatens Nafta

After Gun Control Marches, 'It'll Go Away' vs. 'We Are Not Cynical Yet'

Sinclair Made Dozens of Local News Anchors Recite the Same Script

SIDEBAR
'A Bit of Divine Justice': Trump Vowed to Change Libel Law. But Not Like...

POLITICS

How Trump Consultants Exploited the Facebook Data of Millions

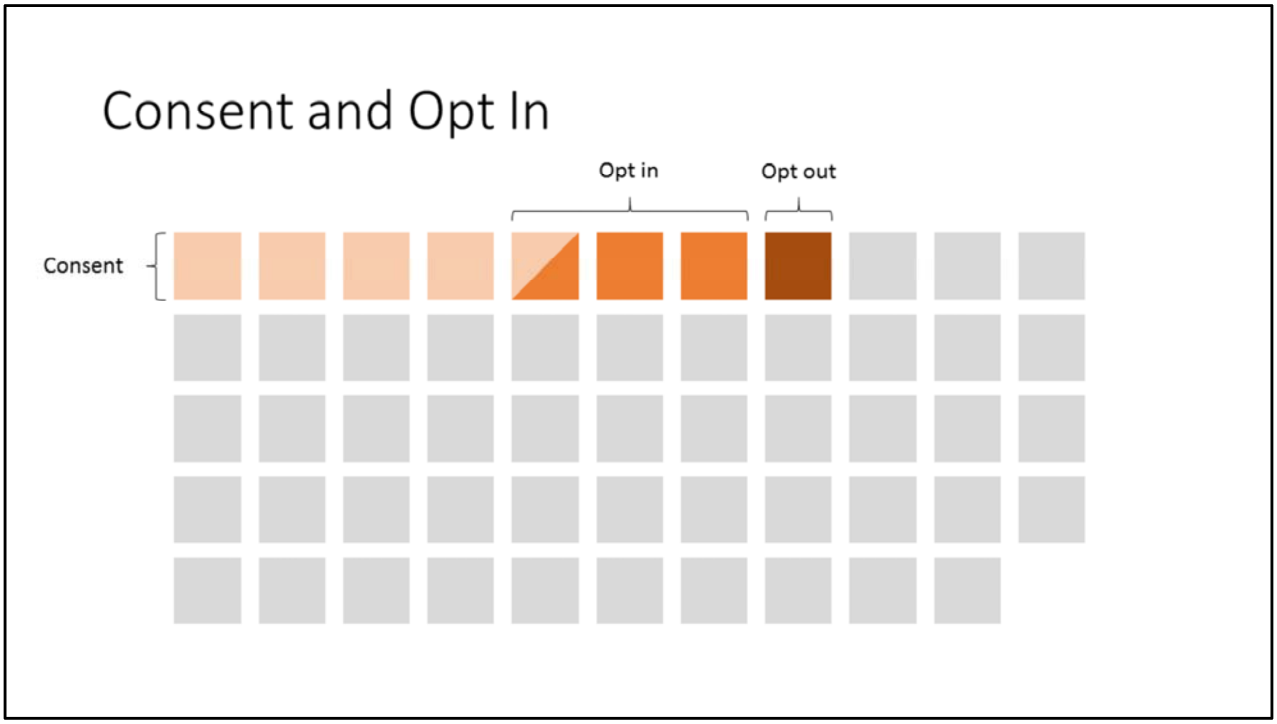
Leer en español

By MATTHEW ROSENBERG, NICHOLAS CONFESSORE and CAROLE CADWALLADR MARCH 17, 2018

https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html

I know Abigail already mentioned Cambridge Analytica but I think it's a great example about why we feel consent is necessary. We want to have a say over what people can and can't do with our data.

I've even heard that consent and allowing students to opt in are what makes libraries different than Facebook and Google.



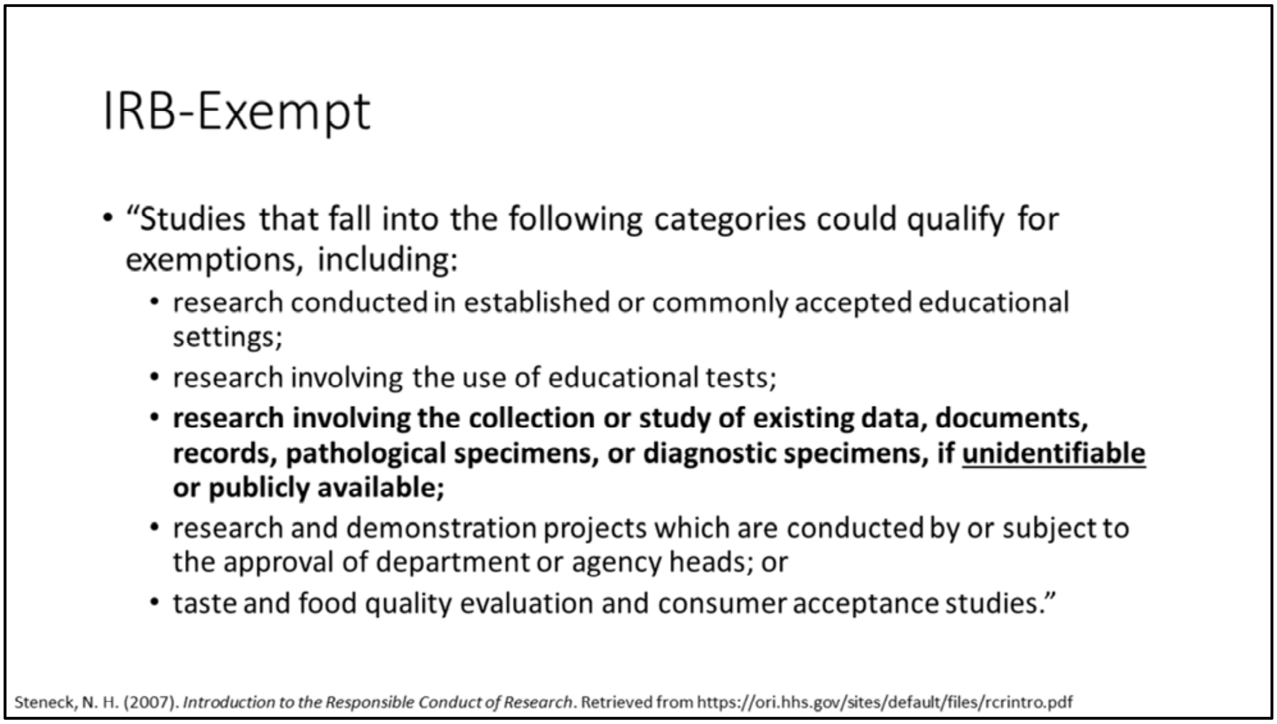
Consent and Opt In

Unfortunately, the data doesn’t show that.

Only 5 of the 54 students explicitly talk about getting students’ consent. And the major of those are problematic because they are two-part studies where they do analytics and something like a focus group. The articles talk quite clearly about getting consent for the focus group but there is no mention of consent for the mass data harvesting and analytics.

The model I like better are the 3 studies that only perform analytics on students who opted into participating. The overlapping square is a study by Angie Thorpe, et al. at IU-Kokomo and I really like their consent model. They trained all of their front-line staff on the consent form and process and only did analytics on the students who elected to participate. It’s worth noting that they only had 70-some students opt in. I know that IU-Kokomo is small, but it’s not that small! Many students will say “no thank you” if given the choice.

There was also one study that allowed students to opt out.



IRB-Exempt

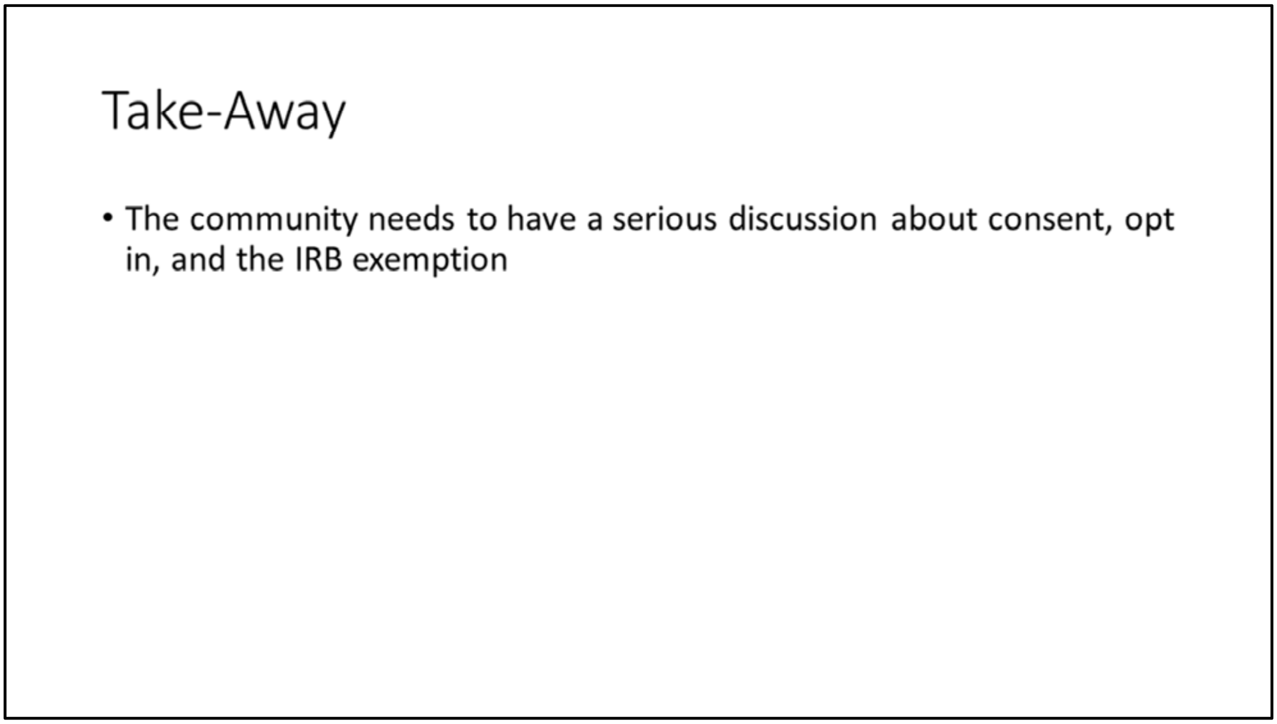
- “Studies that fall into the following categories could qualify for exemptions, including:
  - research conducted in established or commonly accepted educational settings;
  - research involving the use of educational tests;
  - **research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if unidentifiable or publicly available;**
  - research and demonstration projects which are conducted by or subject to the approval of department or agency heads; or
  - taste and food quality evaluation and consumer acceptance studies.”

Steneck, N. H. (2007). *Introduction to the Responsible Conduct of Research*. Retrieved from https://ori.hhs.gov/sites/default/files/rcrintro.pdf

The number of studies gaining consent for analytics is small but we also have to take into account IRB requirements. All of this research was conducted at academic institutions, where human subjects research is governed by the Institutional Review Board, or IRB. IRB requires protections such as consent for human subject research with a few exceptions. I’ve listed them here and highlighted the one applies here.

Libraries are allowed to do research using existing data without having to do the full human subject protocols, such as consent. Do note the last part of that requirement, however. The requirement is that the data either be public (which it’s not) or unidentifiable. I just spent several minutes talking to you about how our data is not anonymous.

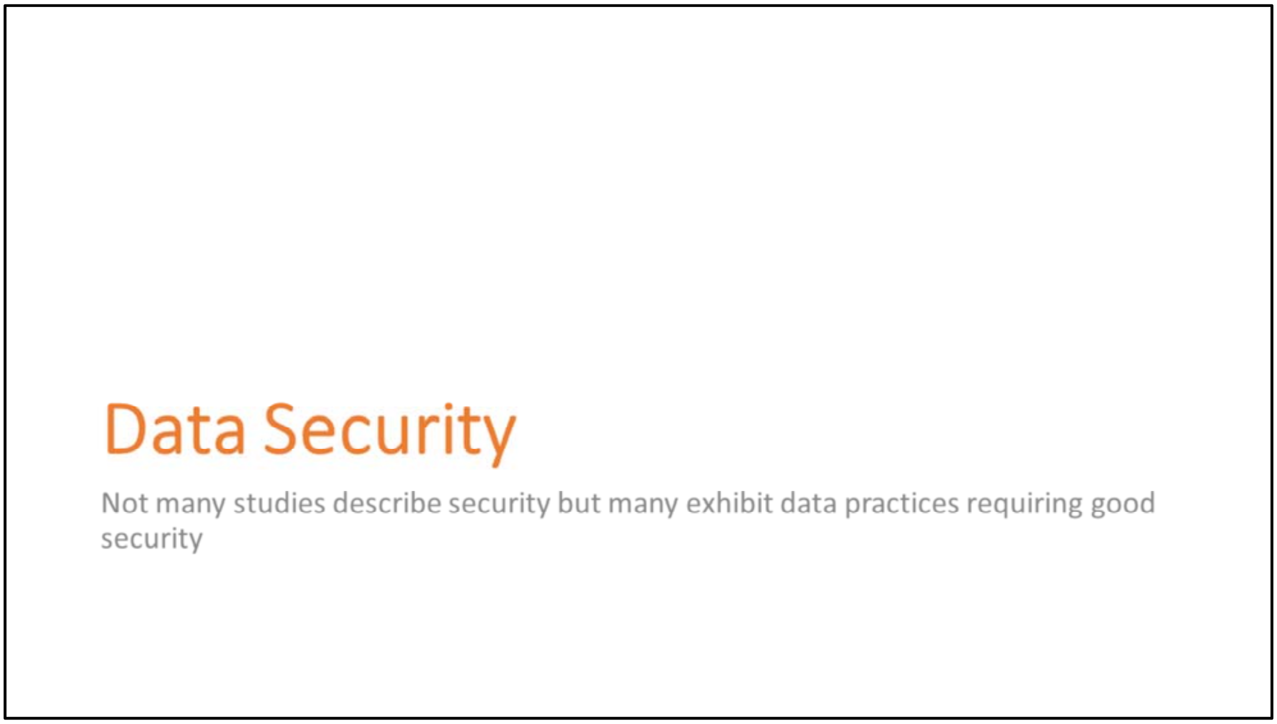
So I have a problem with the validity of the exemption in many of the studies we’re talking about.



Take-Away

- The community needs to have a serious discussion about consent, opt in, and the IRB exemption

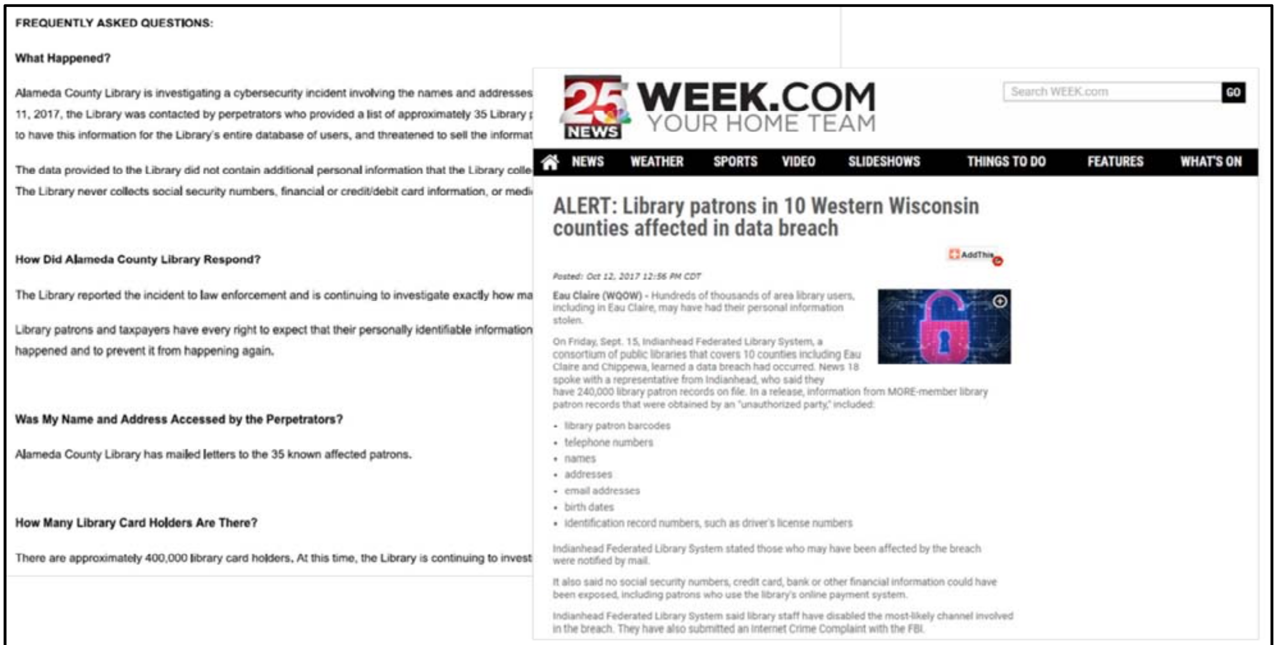
I don’t have any take aways other than we need to talk about this more. I have the data and I have concerns and I would love to hear more from the library community and experts on consent.



Data Security

Not many studies describe security but many exhibit data practices requiring good security

Let’s wrap things up with our third topic: data security.



FREQUENTLY ASKED QUESTIONS:

What Happened?

Alameda County Library is investigating a cybersecurity incident involving the names and addresses
11, 2017, the Library was contacted by perpetrators who provided a list of approximately 35 Library
to have this information for the Library's entire database of users, and threatened to sell the informat

The data provided to the Library did not contain additional personal information that the Library colle
The Library never collects social security numbers, financial or credit/debit card information, or medi

How Did Alameda County Library Respond?

The Library reported the incident to law enforcement and is continuing to investigate exactly how ma

Library patrons and taxpayers have every right to expect that their personally identifiable information
happened and to prevent it from happening again.

Was My Name and Address Accessed by the Perpetrators?

Alameda County Library has mailed letters to the 35 known affected patrons.

How Many Library Card Holders Are There?

There are approximately 400,000 library card holders. At this time, the Library is continuing to invest

ALERT: Library patrons in 10 Western Wisconsin counties affected in data breach

Posted: Oct 12, 2017 12:56 PM CDT

Eau Claire (WQOW) - Hundreds of thousands of area library users, including in Eau Claire, may have had their personal information stolen.

On Friday, Sept. 15, Indianhead Federated Library System, a consortium of public libraries that covers 10 counties including Eau Claire and Chippewa, learned a data breach had occurred. News 18 spoke with a representative from Indianhead, who said they have 240,000 library patron records on file. In a release, information from MORE-member library patron records that were obtained by an "unauthorized party," included:

- library patron barcodes
- telephone numbers
- names
- addresses
- email addresses
- birth dates
- identification record numbers, such as driver's license numbers

Indianhead Federated Library System stated those who may have been affected by the breach were notified by mail.

It also said no social security numbers, credit card, bank or other financial information could have been exposed, including patrons who use the library's online payment system.

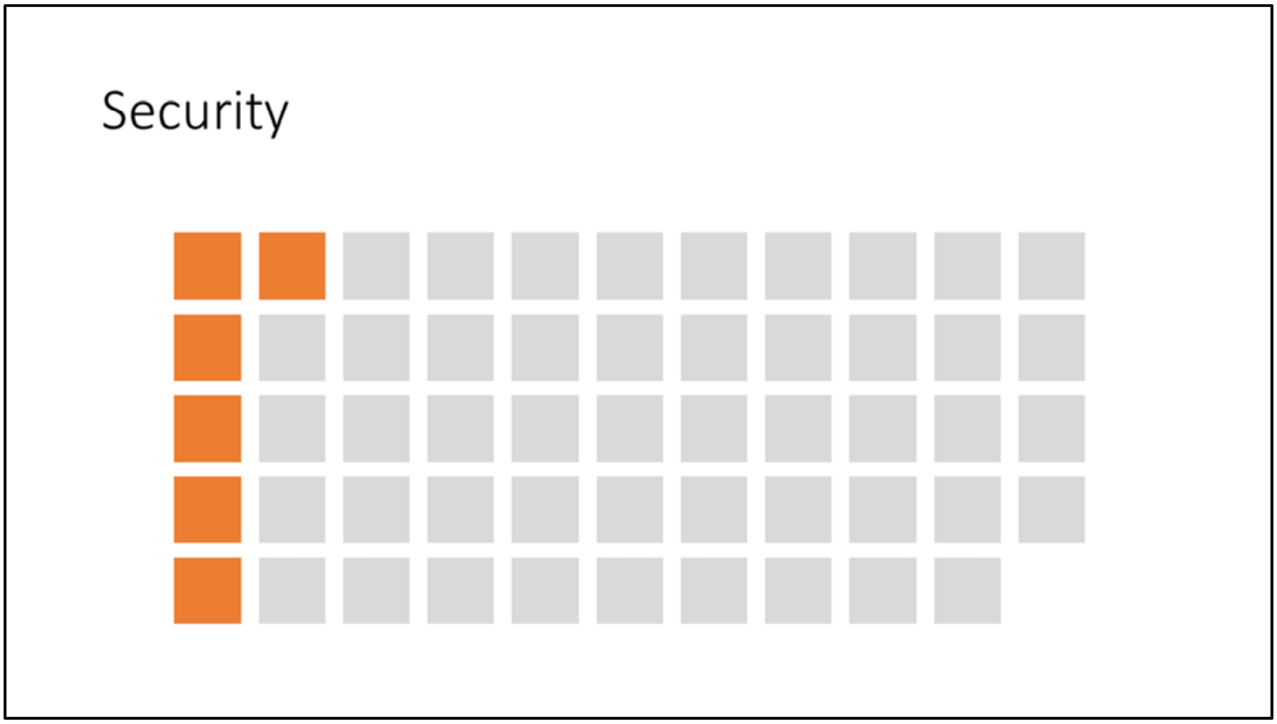
Indianhead Federated Library System said library staff have disabled the most-likely channel involved in the breach. They have also submitted an Internet Crime Complaint with the FBI.

Alameda County Library. "Frequently Asked Questions," 2017. https://www.aclibrary.org/content/frequently-asked-questions.
Week.com. "ALERT: Library Patrons in 10 Western Wisconsin Counties Affected." *News 25*. October 12, 2017.
http://www.week.com/story/36583282/2017/10/Thursday/alert-library-patrons-in-10-western-wisconsin-counties-affected-in-data-breach.

We hear a lot about data security in the news; every week it seems like another company has experienced a data breach.

And it's not just corporations. Universities and even public libraries have been breached. In 2017, there were data breaches at 2 public libraries: one in California and one in my home state of Wisconsin. These breaches ended up being more damaging than they might have needed to be because the libraries were holding driver's license numbers and full date-of-birth – two data points that often allow you to unlock other sensitive information.

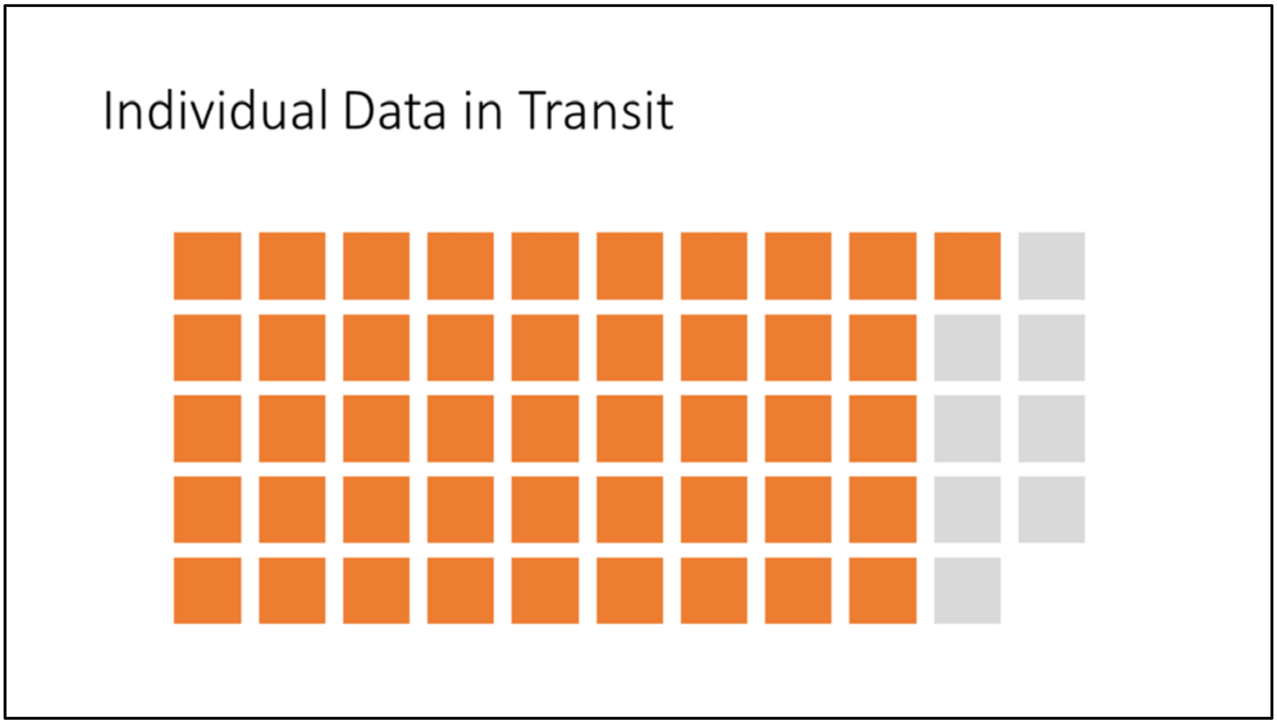
So let's look at what academic libraries are doing in this area.



Security

Unfortunately, very few libraries talk about their security set up. This doesn’t surprise me, as a published article is not the common format for discussing data security.

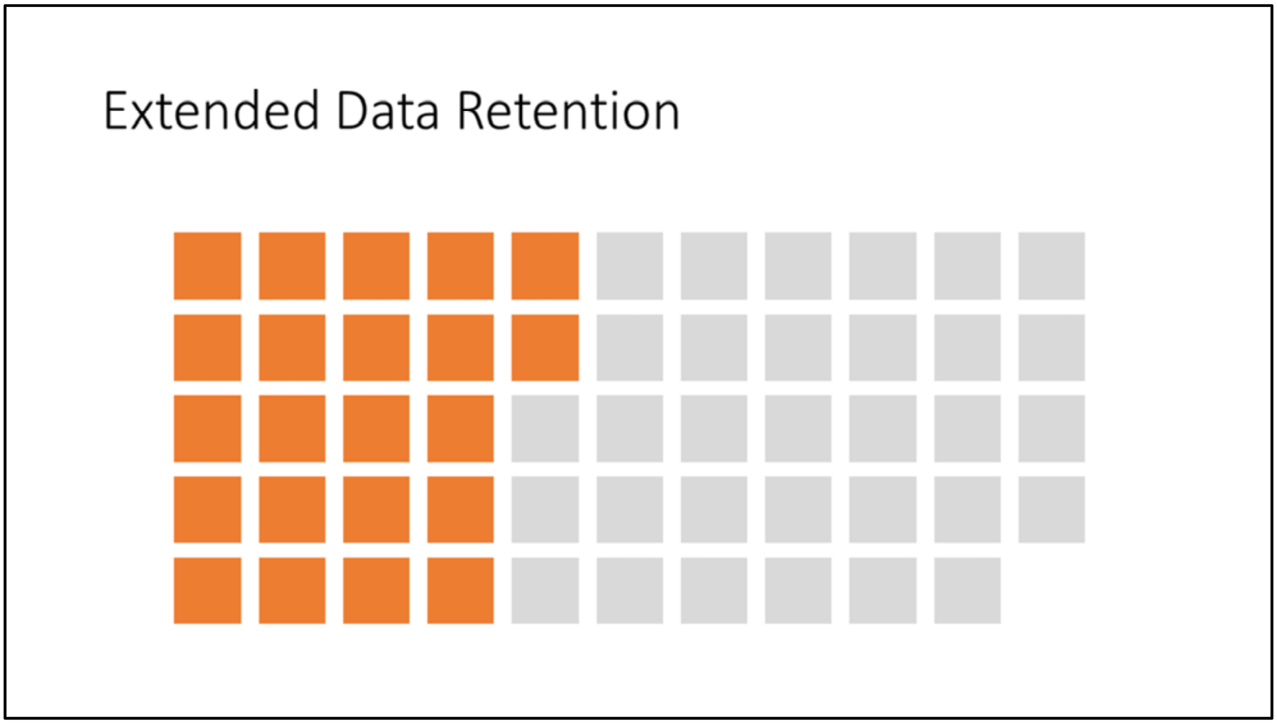
Of the articles that discuss security, many say that their campus IT experts are running their analytics server. That’s great! More of this please!



Individual Data in Transit

That said, there are a number of studies that raise security concerns. For example, most of the studies I looked at implied transferring individual data between the library and the university. (Remember I said that these studies correlate library usage and student outcomes, connecting on sometime like an individual ID number, so that data has to get between the library and university somehow.)

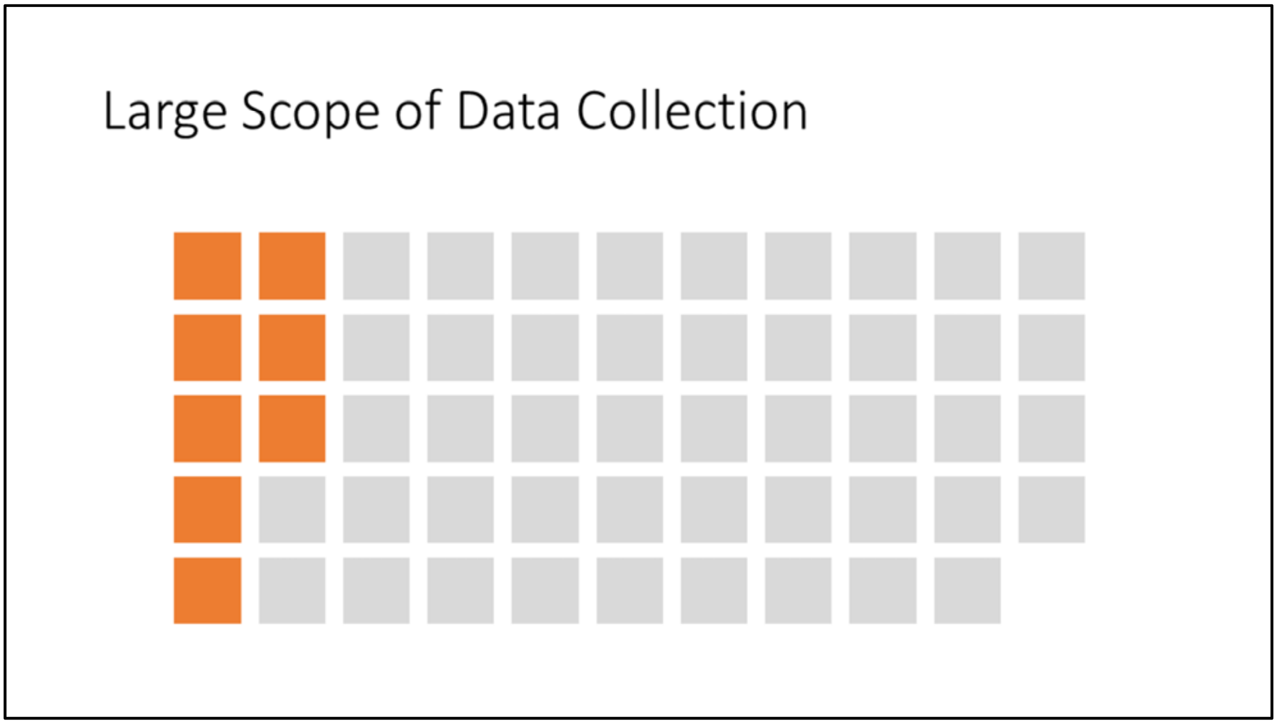
This transfer could be happening securely but almost no one is talking about this and data-in-transit is an easy place for sensitive information to get lost.



Extended Data Retention

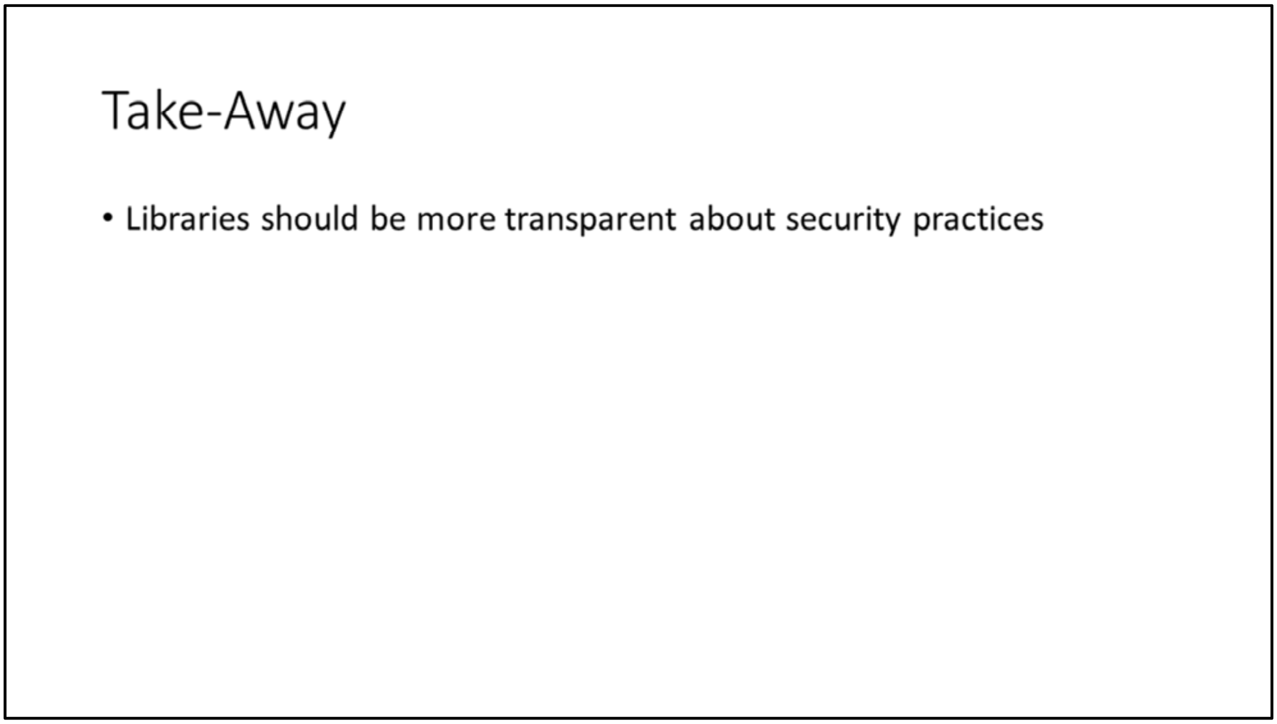
Another security concerns is that we’re holding on to data for long periods of time. 22 studies hold data for 2, 3, 5, 10, or more years.

A security rule of thumb is that the best way to secure data is to not have it at all. So the longer we’re holding data, the more concerned we need to be about security.



Large Scope of Data Collection

And what we’re holding is really sensitive. I found a several studies that are collecting 10 or more data points on a student. Another study by Kogut found that almost half of library learning analytics articles look at GPA, a sensitive data point. I also found 5 studies that examine socioeconomic status or proxy for such. This data absolutely requires good security.



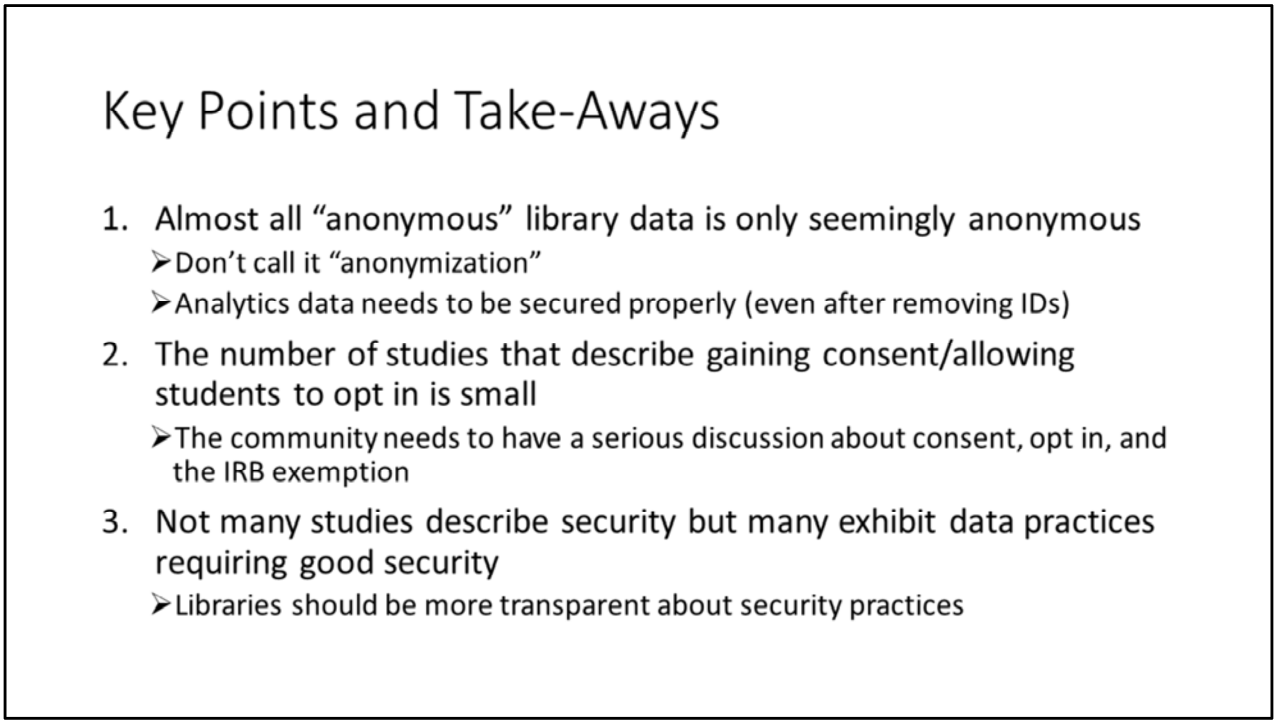
Take-Away

- Libraries should be more transparent about security practices

The take away is that I would love to hear more about how we’re securing this data. Let’s definitely call out and publish examples of people doing it well.

Conclusion

All right, time to wrap up.



Key Points and Take-Aways

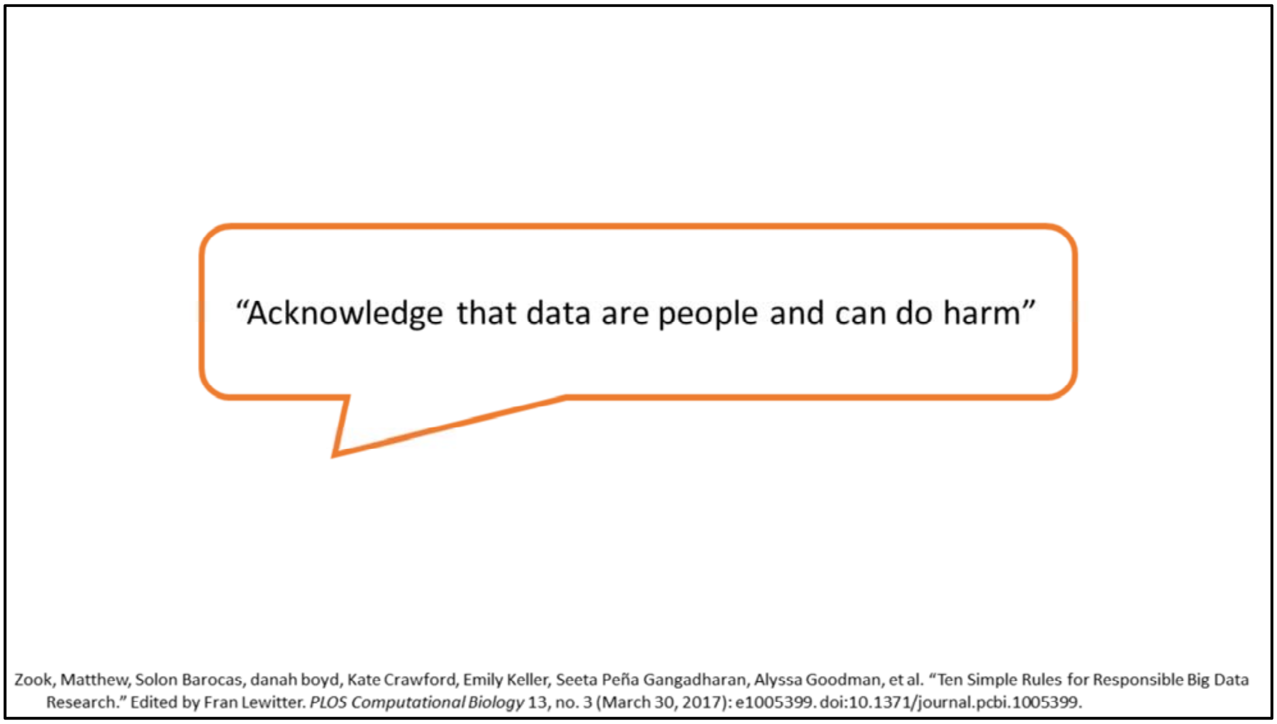
1. Almost all “anonymous” library data is only seemingly anonymous
   - ➢Don’t call it “anonymization”
   - ➢Analytics data needs to be secured properly (even after removing IDs)
2. The number of studies that describe gaining consent/allowing students to opt in is small
   - ➢The community needs to have a serious discussion about consent, opt in, and the IRB exemption
3. Not many studies describe security but many exhibit data practices requiring good security
   - ➢Libraries should be more transparent about security practices

I talked about 3 things today:

First, most analytics data isn’t actually anonymous. So let’s stop calling it “anonymous” and protect it better.

Second, the number of studies that give students a choice to participate is really small. Coupled with the IRB exemption and the fact that our data is not “unidentifiable”, means that we need to be talking about this.

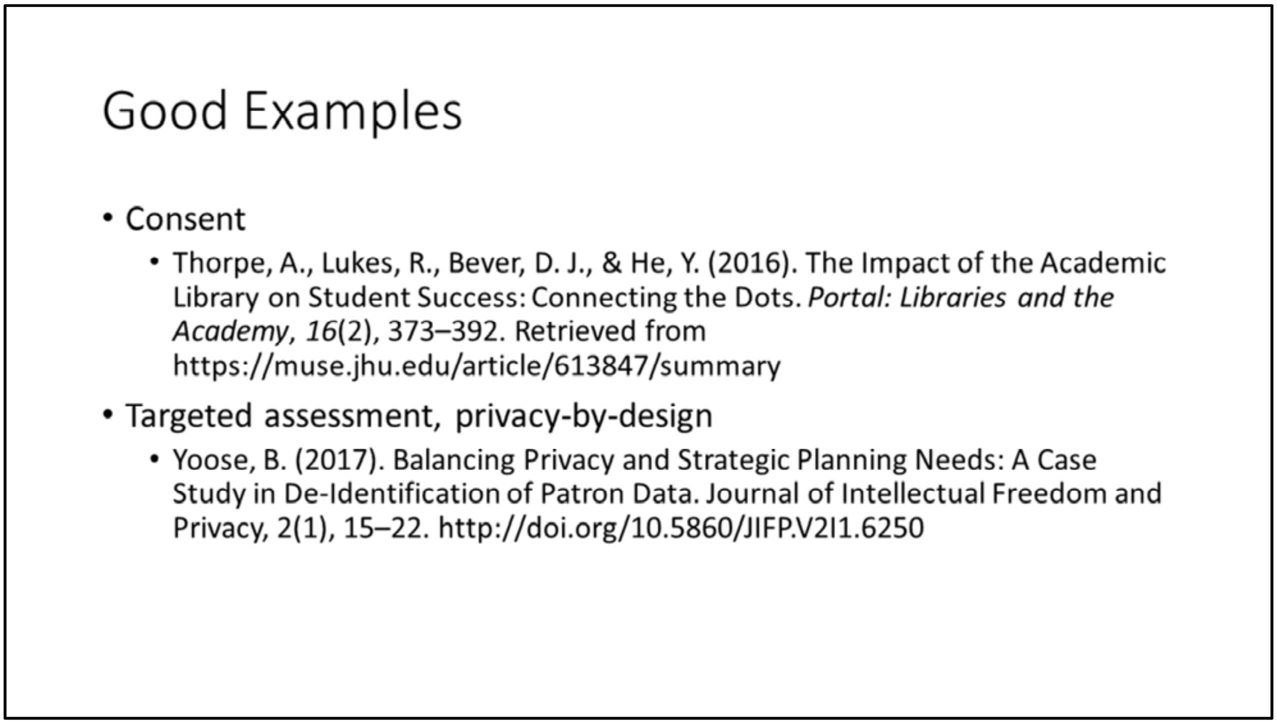
Third, all of the sensitive data we’re collecting and keeping requires good security and I would love to see more examples of people doing this well.



"Acknowledge that data are people and can do harm"

Zook, Matthew, Solon Barocas, danah boyd, Kate Crawford, Emily Keller, Seeta Peña Gangadharan, Alyssa Goodman, et al. "Ten Simple Rules for Responsible Big Data Research." Edited by Fran Lewitter. *PLOS Computational Biology* 13, no. 3 (March 30, 2017): e1005399. doi:10.1371/journal.pcbi.1005399.

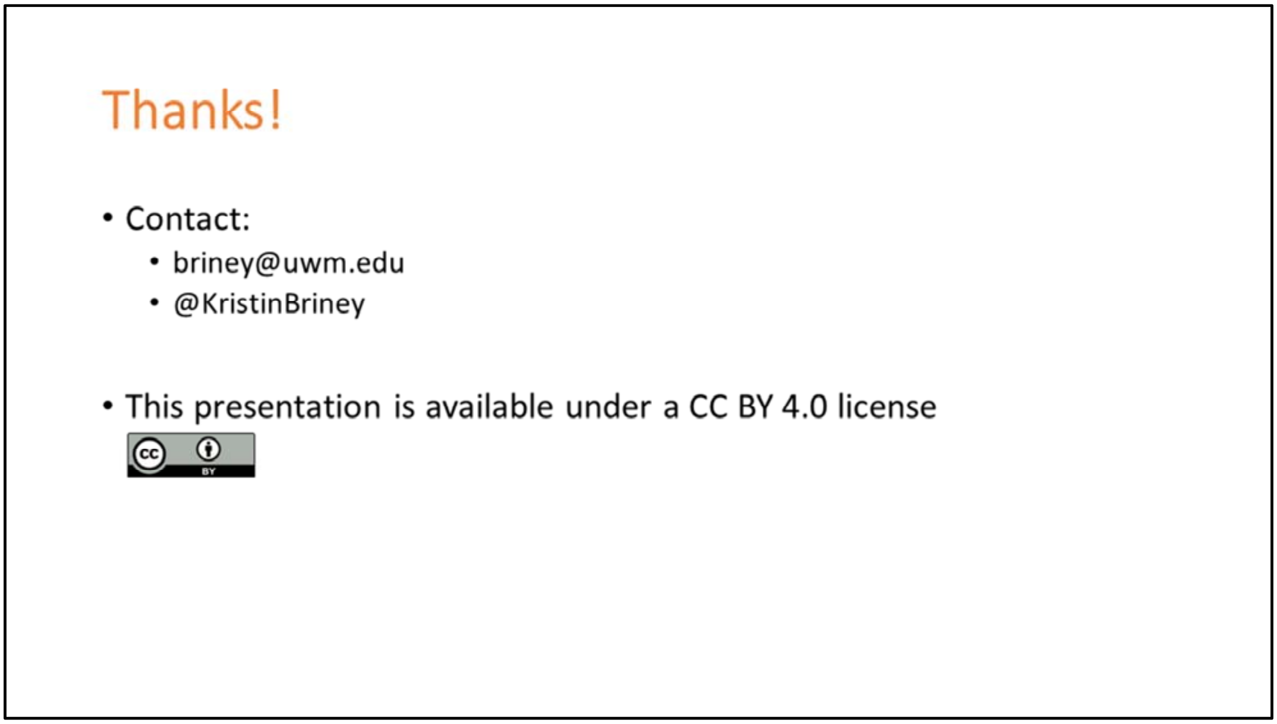
I want to wrap up with another quote, this one from an article by Zook, et al. called "Ten Simple Rules for Responsible Big Data Research". We need to "acknowledge that data are people and can do harm." I think that if we can use this as a guiding principle, we'll start being more thoughtful about how we handle learning analytics data and see the need for using multiple layers of protection around this type of data.



Good Examples

- Consent
  - Thorpe, A., Lukes, R., Bever, D. J., & He, Y. (2016). The Impact of the Academic Library on Student Success: Connecting the Dots. *Portal: Libraries and the Academy*, *16*(2), 373–392. Retrieved from https://muse.jhu.edu/article/613847/summary
- Targeted assessment, privacy-by-design
  - Yoose, B. (2017). Balancing Privacy and Strategic Planning Needs: A Case Study in De-Identification of Patron Data. Journal of Intellectual Freedom and Privacy, 2(1), 15–22. http://doi.org/10.5860/JIFP.V2I1.6250

Finally, I have two good examples to share with you. First is the one I cited earlier by Angie Thorpe, et al. This is the best example of consent and opt in I've found in all of these studies. Second is a paper by Becky Yoose on balancing privacy and assessment. It's hands-down the best example of doing analytics in a privacy and security-conscious way and I would dearly love to see many more papers like this.



Thanks!

- Contact:
  - briney@uwm.edu
  - @KristinBriney

Thank you!