PURDUE UNIVERSITY | Department of Computer Science

CS57100: Artificial Intelligence
Ethics and AI

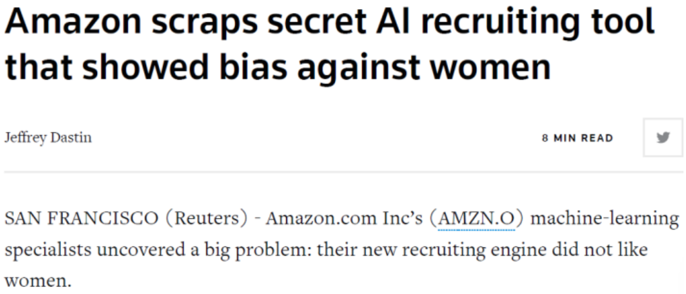
Prof. Chris Clifton

11 November 2022

Indiana Center for Database Systems™

PURDUE UNIVERSITY
Department of Computer Science

What's all the fuss?
(Dastin '18)



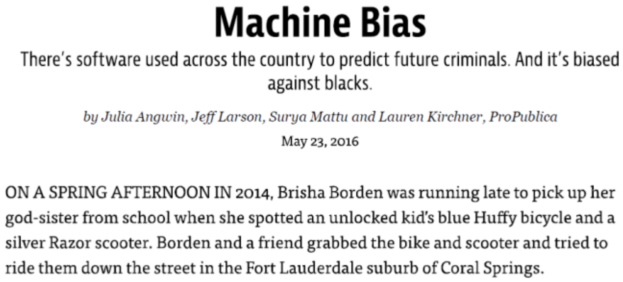
Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

- Resume screening tool
  - Trained on prior applications
  - Demonstrated bias toward male applicants
  - Manual avoidance of "obvious" discriminatory words
- *Scrapped for fear of remaining biases*

What's all the fuss?
(Angwin, Larson, Mattu, Kirchner '16)



Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

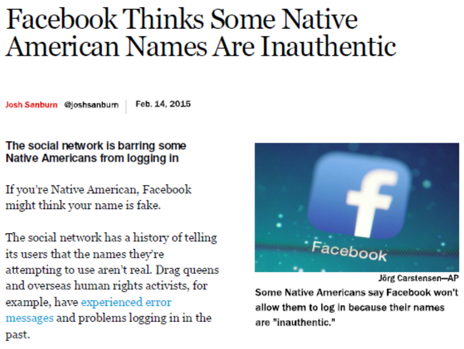
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

- Similar cases lead to different outcomes
  - Minor theft (shoplifting, stealing a bike)
  - Black offender predicted as more likely to commit future crime than white
  - *Despite white offender having criminal record!*
- Statistical analysis suggests this is common

What's all the fuss?
(Sanburn '15)



Facebook Thinks Some Native American Names Are Inauthentic

Josh Sanburn @joshsanburn | Feb. 14, 2015

The social network is barring some Native Americans from logging in

If you're Native American, Facebook might think your name is fake.

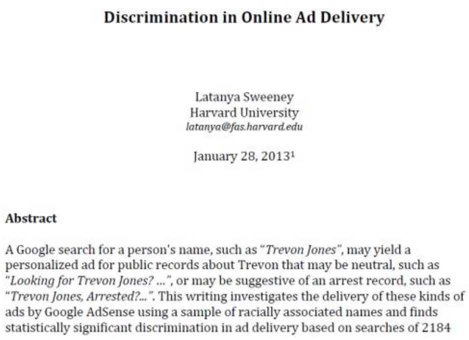
The social network has a history of telling its users that the names they're attempting to use aren't real. Drag queens and overseas human rights activists, for example, have experienced error messages and problems logging in in the past.

Jörg Carstensen—AP

Some Native Americans say Facebook won't allow them to log in because their names are "inauthentic."

- Ms. Lone Elk (and others) required to provide identification to use Facebook
  - Viewed as potential violation of "real name" policy
- No such barriers for "dominant majority"

What's all the fuss?
(Sweeney '13)



Discrimination in Online Ad Delivery

Latanya Sweeney
Harvard University
latanya@fas.harvard.edu

January 28, 2013[1]

Abstract

A Google search for a person's name, such as "*Trevon Jones*", may yield a personalized ad for public records about Trevon that may be neutral, such as "*Looking for Trevon Jones? …*", or may be suggestive of an arrest record, such as "*Trevon Jones, Arrested?...*". This writing investigates the delivery of these kinds of ads by Google AdSense using a sample of racially associated names and finds statistically significant discrimination in ad delivery based on searches of 2184

- Blacks and whites see different ads on the internet
  - *Even if race not part of the profile*
- Sweeney found that first names typically associated with blacks and whites lead to different ads
  - Otherwise identical profiles and histories

What's all the fuss?
(Datta,Tschantz, and Datta '15)

DE GRUYTER OPEN   Proceedings on Privacy Enhancing Technologies 2015; 2015 (1):92–112

Amit Datta*, Michael Carl Tschantz, and Anupam Datta

Automated Experiments on Ad Privacy Settings

A Tale of Opacity, Choice, and Discrimination

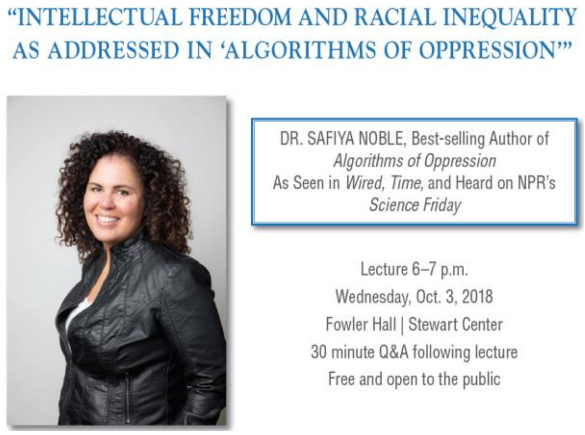
Abstract: To partly address people's concerns over web tracking, Google has created the Ad Settings webpage to provide information about and some choice over the profiles Google creates on users. We present AdFisher, an automated tool that explores how user behaviors, Google's ads, and Ad Settings interact. AdFisher can run browser-based experiments and analyze data using machine learning and significance tests. Our tool uses a rigorous experimental design and statistical analysis to ensure the statistical soundness of our results. We use AdFisher to find that the Ad Settings was opaque about some features of a user's profile, that it does provide some choice on ads, and that these choices can lead to seemingly discriminatory ads. In particular, we found

serious privacy concern. Colossal amounts of collected data are used, sold, and resold for serving targeted content, notably advertisements, on websites (e.g., [1]). Many websites providing content, such as news, outsource their advertising operations to large third-party ad networks, such as Google's DoubleClick. These networks embed tracking code into webpages across many sites providing the network with a more global view of each user's behaviors.

People are concerned about behavioral marketing on the web (e.g., [2]). To increase transparency and control, Google provides Ad Settings, which is "a Google tool that helps you control the ads you see on Google services and on websites that partner with Google" [3].

- Study of impact of different ad privacy settings
- Disclosing Gender resulted in fewer ads for high-paying jobs

And it isn't just CS people who notice



"INTELLECTUAL FREEDOM AND RACIAL INEQUALITY AS ADDRESSED IN 'ALGORITHMS OF OPPRESSION'"

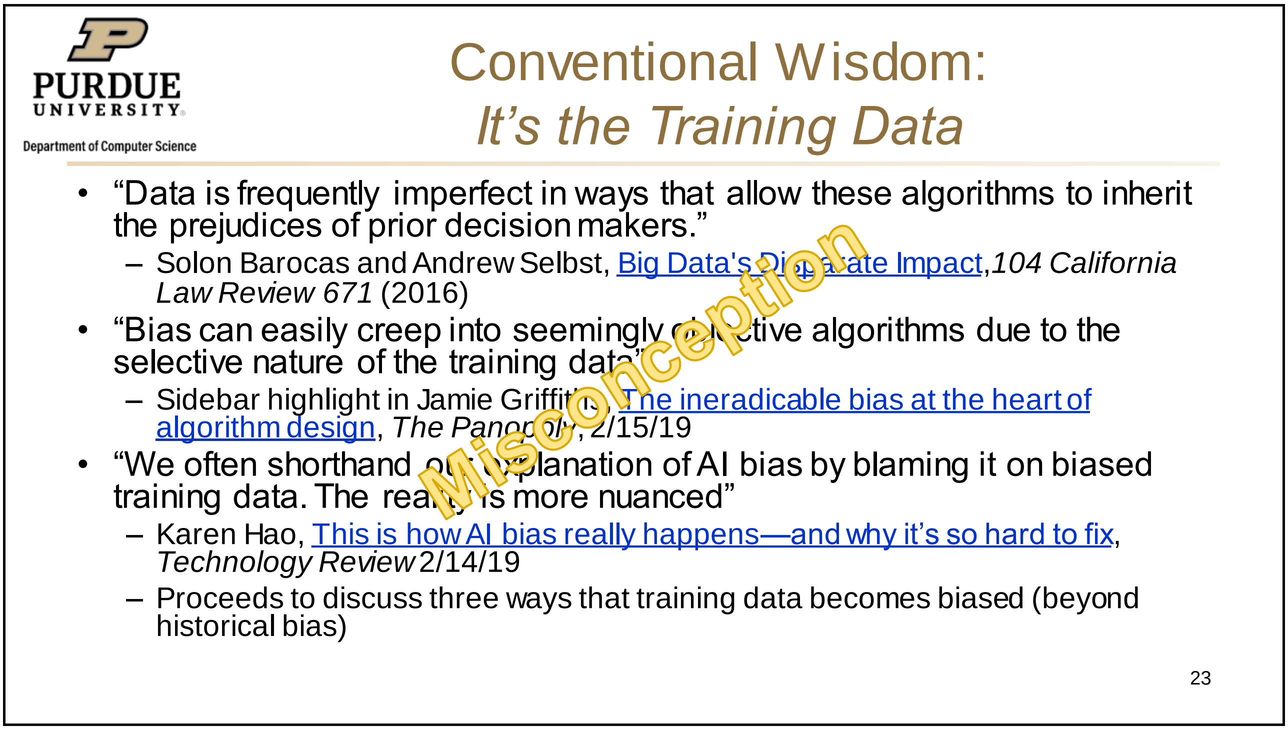
DR. SAFIYA NOBLE, Best-selling Author of *Algorithms of Oppression*
As Seen in *Wired, Time*, and Heard on NPR's *Science Friday*

Lecture 6–7 p.m.
Wednesday, Oct. 3, 2018
Fowler Hall | Stewart Center
30 minute Q&A following lecture
Free and open to the public

- **In an increasingly automated world, what IF AI tools punish the poor?**
- Feb. 13, 2019
  Fowler Hall
  Purdue U.

What are the reasons?

- Discrimination intentionally programmed into the system?
  - Let's hope not
- Historical bias in the training data?
  - May explain some, but not all
- Insensitivity on the part of developers?
  - Maybe
- Or perhaps we don't know (yet)?



PURDUE UNIVERSITY

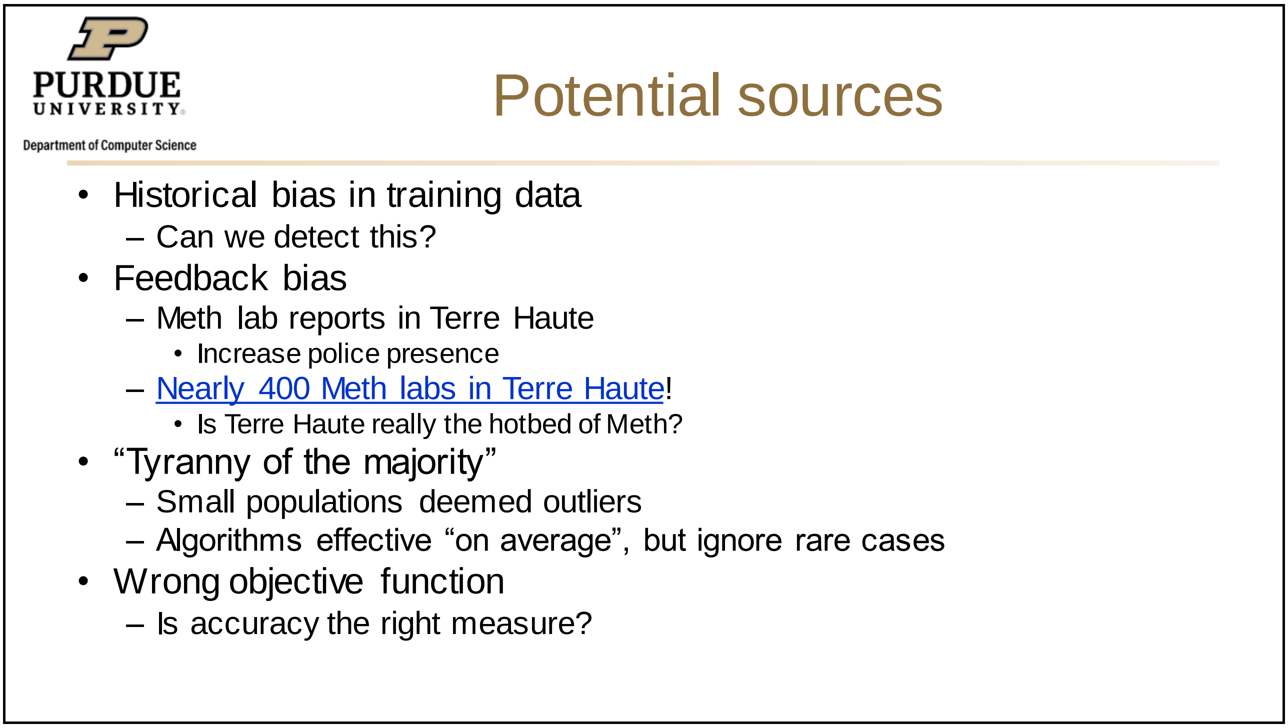
Department of Computer Science

Conventional Wisdom: *It’s the Training Data*

- “Data is frequently imperfect in ways that allow these algorithms to inherit the prejudices of prior decision makers.”
  - Solon Barocas and Andrew Selbst, Big Data's Disparate Impact, *104 California Law Review 671* (2016)
- “Bias can easily creep into seemingly objective algorithms due to the selective nature of the training data”
  - Sidebar highlight in Jamie Griffith, The ineradicable bias at the heart of algorithm design, *The Panoply*, 2/15/19
- “We often shorthand our explanation of AI bias by blaming it on biased training data. The reality is more nuanced”
  - Karen Hao, This is how AI bias really happens—and why it’s so hard to fix, *Technology Review* 2/14/19
  - Proceeds to discuss three ways that training data becomes biased (beyond historical bias)

Misconception

23



PURDUE UNIVERSITY

Department of Computer Science

Potential sources

- Historical bias in training data
  - Can we detect this?
- Feedback bias
  - Meth lab reports in Terre Haute
    - Increase police presence
  - Nearly 400 Meth labs in Terre Haute!
    - Is Terre Haute really the hotbed of Meth?
- “Tyranny of the majority”
  - Small populations deemed outliers
  - Algorithms effective “on average”, but ignore rare cases
- Wrong objective function
  - Is accuracy the right measure?

So Where Is the Problem?

- We can show that some machine learning techniques should *reduce* bias from that in the training data
  - So why do we have so many examples of biased ML?
- **It isn't just the training data!**

Myth: Machines are Unbiased

- Machine Learning can *introduce* bias against minority groups
  - Even when the minority is *advantaged*

25

What can we do?

- Detect discriminatory outcomes from machine learning
  - [Pedreschi08, Pedreschi09, Luong11, Ruggieri11]
- Relabel training samples
  - [Kamiran09, Zliobaite11, Kamiran11]
- Adjust scoring functions
  - [Calders10, Kamiran10]
- statistical parity
  - [Dwork12, Zemel13]

Myth: We Just Need Statistical Equality

PURDUE UNIVERSITY

Department of Computer Science

Multiple Measures: *Disparate Treatment vs. Disparate Impact*

- Disparate treatment: Individuals from different groups treated differently
  - Otherwise identical individuals have different outcome based only on group membership
- Disparate impact: Outcomes different between different groups
  - No individuals are “the same”
  - Different outcomes for different groups, even if some other explanation
- Prior work largely relies on *using* special categories
  - This can qualify as disparate treatment

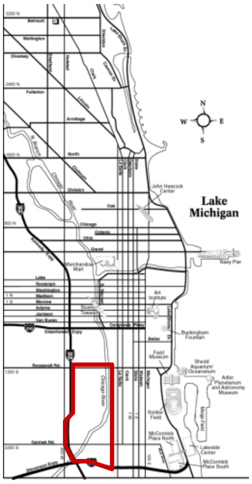
PURDUE UNIVERSITY

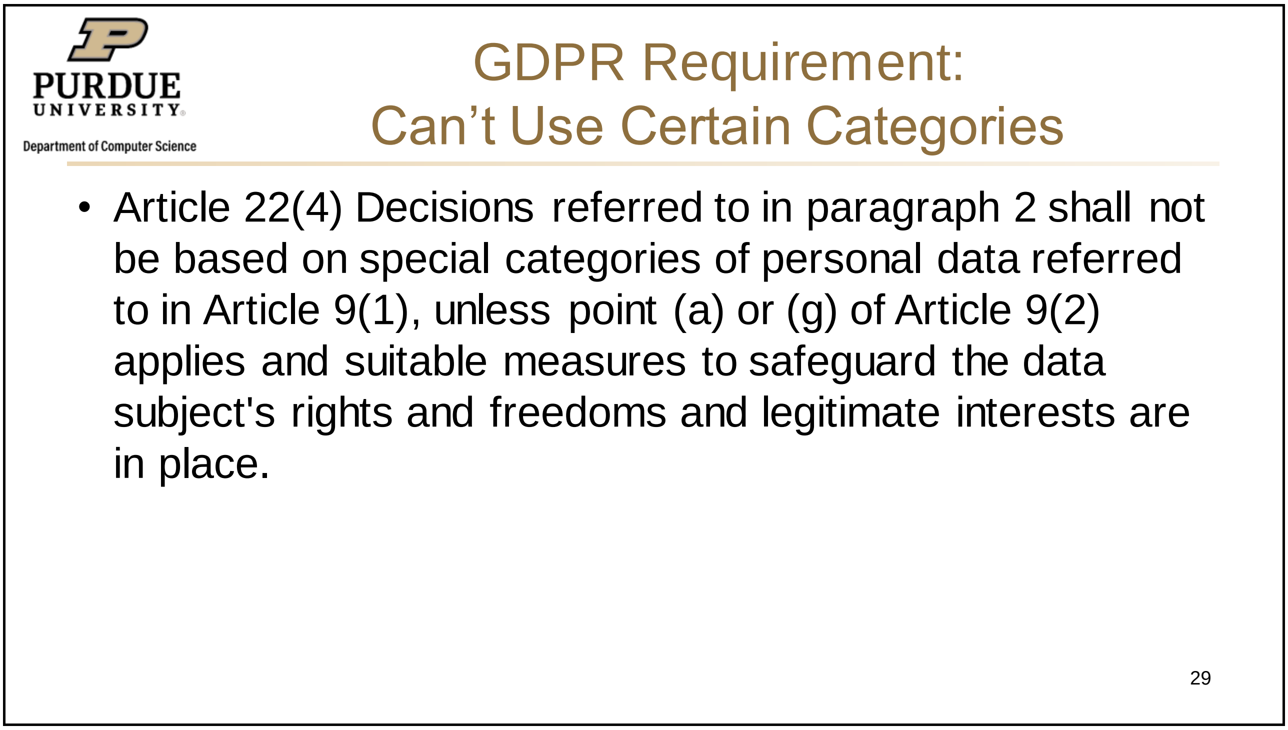
Department of Computer Science

Why Disparate Impact?

- Mortgage Redlining
  - Racial discrimination in home loans prohibited in US
  - Banks drew lines around high risk neighborhoods!!!
  - These were often minority neighborhoods
  - Result: Discrimination (redlining outlawed)

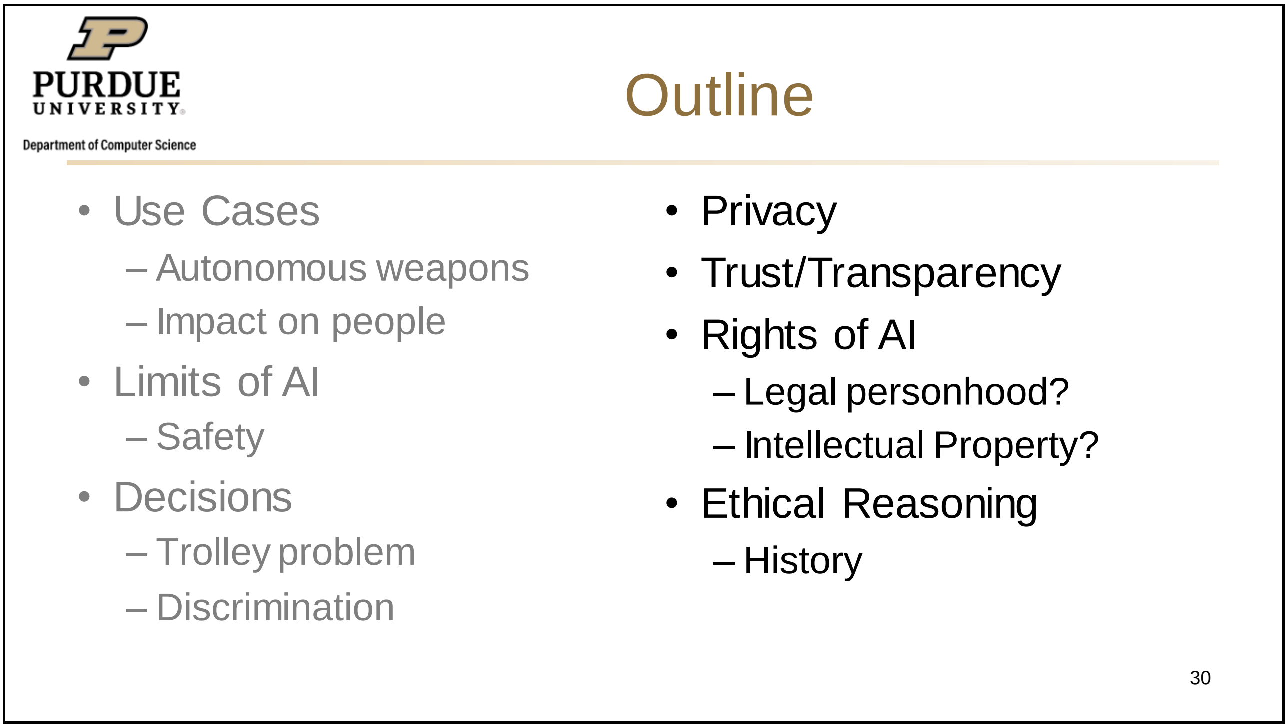
  What about data mining that “singles out” minorities?





GDPR Requirement: Can’t Use Certain Categories

- Article 22(4) Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.



Outline

- Use Cases
  - Autonomous weapons
  - Impact on people
- Limits of AI
  - Safety
- Decisions
  - Trolley problem
  - Discrimination
- Privacy
- Trust/Transparency
- Rights of AI
  - Legal personhood?
  - Intellectual Property?
- Ethical Reasoning
  - History

What is Privacy?

- “The right to be let alone” - *Warren & Brandeis, 4 Harvard L.R. 193 (Dec. 15, 1890)*
  - My information protected so it doesn’t adversely affect me in the future
- Control over data
  - My information used only in ways I approve
- Issues:
  - Disclosure / sharing
  - Approved use
  - Recourse

Data Privacy: The Goal

- Protect the Individual
  - “Everyone has the right to the protection of personal data concerning him or her. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified.” – Charter of Fundamental Rights of the European Union
- Challenges: What do we mean by
  - “concerning” an individual
  - Protection
  - Consent
  - Access / rectified

European Commission

PURDUE UNIVERSITY
Department of Computer Science

“Obvious” answers

- Concerning an individual
  - Has your name/address/other identifying information
- Protection
  - Only used/accessed in expected, intended, authorized ways
- Consent
  - You know and agree to what is done with the data
- Access/Rectify
  - You can see the data and correct errors

PURDUE UNIVERSITY
Department of Computer Science

Consent

- When you apply for a (job, grad school, …), do you consent to that data being used with an ML model to decide if you should be accepted?
  - Amazon tried it: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
- What about having your data used as training data to make decisions about others?
  - *Ungraded assignment (post-midterm): Read the terms of service and privacy policy of Facebook or some other social media you use, and think about this question.*

“Obvious” answers

- Concerning an individual
  - Has your name/address/other identifying information
- Protection
  - Only used/accessed in expected, intended, authorized ways
- Consent
  - You know and agree to what is done with the data
- Access/Rectify
  - You can see the data and correct errors

Concerning an Individual: IC 24-4.9-2-10

Sec. 10. "Personal information" means:

(1) a Social Security number that is not encrypted or redacted; or

(2) an individual's first and last names, or first initial and last name, and one (1) or more of the following data elements that are not encrypted or redacted:

(A) A driver's license number.

(B) A state identification card number.

(C) A credit card number.

(D) A financial account number or debit card number in combination with a security code, password, or access code that would permit access to the person's account.

PURDUE UNIVERSITY

Department of Computer Science

The AOL Awakening

- In Aug 2006, AOL released its customers web searches for research studies
- 20 Million unique queries of 650K unique users
- <user-i
- NY Tim n individu from the queries
  - Queries included “60 single men” “landscapers in Lilburn, Ga”
  - Many more queries contained enough information to uniquely identify the person
- *And it keeps going (Netflix, NYC Taxi, …)*

AOL fired its CTO over this issue;

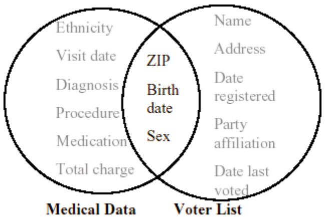
Two researchers were forced out

PURDUE UNIVERSITY

Department of Computer Science

Re-identifying “anonymous” data (Sweeney ’01)

- 37 US states mandate collection of information
- Dr. Sweeney purchased the voter registration list for Cambridge Massachusetts
  - 54,805 people
- 69% unique on postal code and birth date
- 87% US-wide with all three



- Solution: k-anonymity
  - Any combination of values appears at least k times
- Developed systems that guarantee k-anonymity
  - Minimize distortion of results

PURDUE UNIVERSITY

Department of Computer Science

Quiz: Indiana Breach Disclosure Law IC 24-4.9-2-10

Suppose someone in the Dean’s office downloaded student information (unencrypted) onto a USB to give to the registrar, and then the USB key disappeared. Which of the following information on the USB key would be considered “Personal Information” and trigger Indiana’s Breach Disclosure law:

A. Student name, address, and unpaid parking violations

B. Student name, address, and photo

C. Student name and Purdue ID number

D. Student name, address, email, telephone, date of birth, and last four digits of social security number

PURDUE UNIVERSITY

Department of Computer Science

Redaction: IC 24-4.9-2-11

(a) Data are redacted for purposes of this article if the data have been altered or truncated so that not more than the last four (4) digits of:

(1) a driver's license number;

(2) a state identification number; or

(3) an account number;

is accessible as part of personal information.

(b) For purposes of this article, personal information is "redacted" if the personal information has been altered or truncated so that not more than five (5) digits of a Social Security number are accessible as part of personal information.

Department of Computer Science

Anonymity: The Goal

- Prevent Disclosure of Personal Information
  - GDPR: 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly
  - Qatar Law 13 of 2016: Personal Data: Data belonging to an Individual with specified or reasonably specifiable identity whether through such Personal Data or through combining the same with any other data
  - *But still use the data where appropriate!*
- Problem: It can't be done!
  - "Perfect" privacy requires zero utility (e.g., the data must be encrypted.)
  - As soon as we can use the data (e.g., decrypt), it is at risk

Department of Computer Science

Why Perfect Privacy is Impossible *(Dwork, McSherry, Nissim, and Smith '06)*

- Background Knowledge
  - Adversary may already know a lot
  - Whatever we provide (even de-identified or anonymized data) may add to that knowledge
- It may just take that "last bit of knowledge" to give the adversary the ability to violate privacy
  - *We can formally prove 1 bit may be too much*

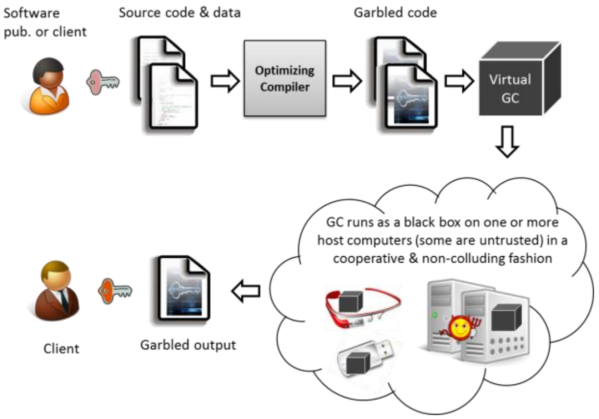
What We Can Do

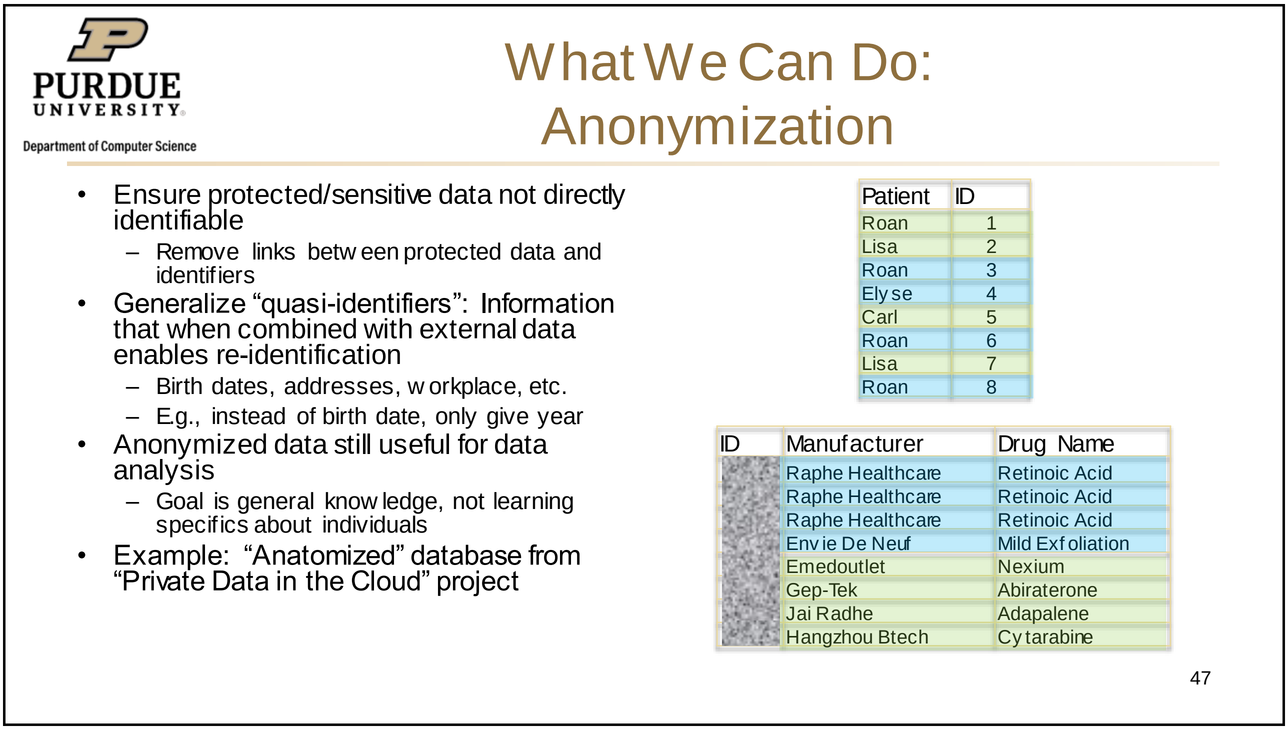
- Encryption
  - Reduce risk to minimal levels when data not in use
- Anonymization
  - Produce usable data that is hard to link to individuals
- Noise addition
  - Usable data where any link to individuals (or information we surmise about individuals) is guaranteed to be uncertain/suspect

What We Can Do: Encryption

- Goal: Reduce risk to minimal levels when data not in use
- Encrypted Computation
  - Process the data while it is encrypted
  - Decrypt final output: Generalized, non-individual results
- Basic tools
  - Homomorphic Encryption, Commutative Encryption, Order Preserving Encryption
- Research Prototypes can accomplish many data processing and analysis tasks using these tools
  - Garbled Computing: Compute without revealing either the data or the program

- Garbled Computing.


Software pub. or client
Source code & data
Garbled code
Optimizing Compiler
Virtual GC
GC runs as a black box on one or more host computers (some are untrusted) in a cooperative & non-colluding fashion
Client
Garbled output

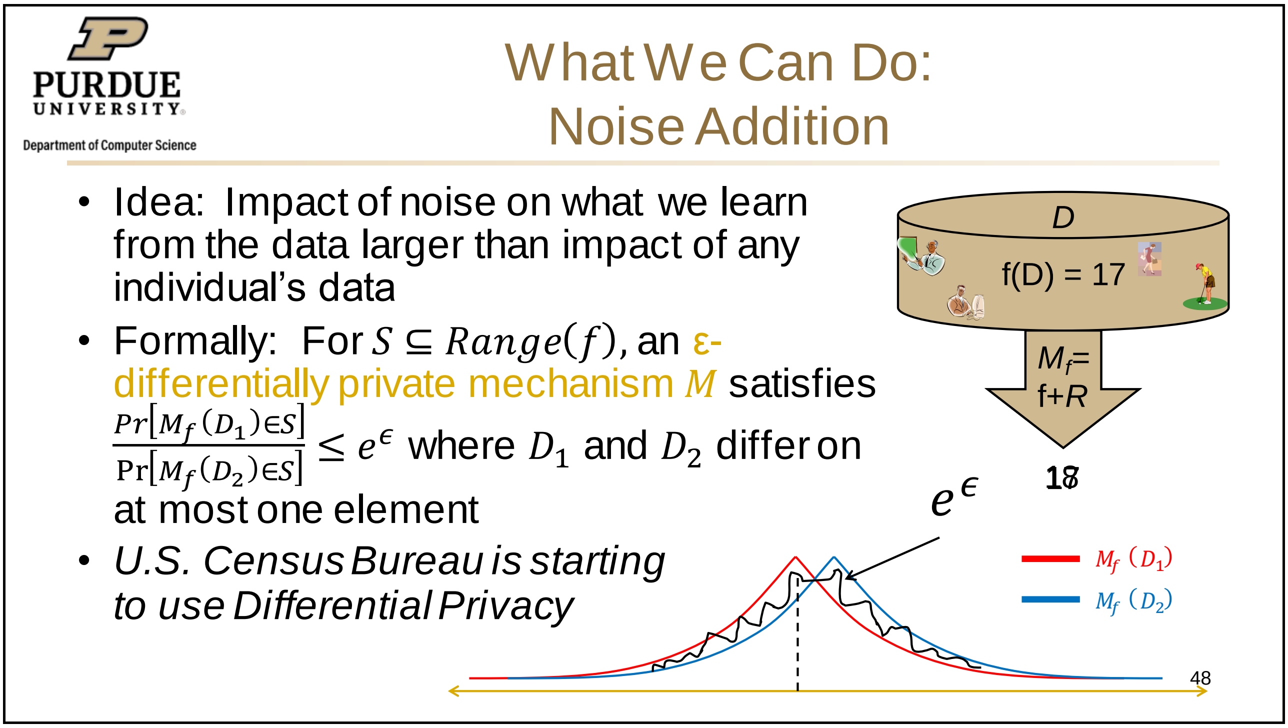


What We Can Do: Anonymization

- Ensure protected/sensitive data not directly identifiable
  - Remove links between protected data and identifiers
- Generalize "quasi-identifiers": Information that when combined with external data enables re-identification
  - Birth dates, addresses, workplace, etc.
  - E.g., instead of birth date, only give year
- Anonymized data still useful for data analysis
  - Goal is general knowledge, not learning specifics about individuals
- Example: "Anatomized" database from "Private Data in the Cloud" project

Patient	ID
Roan	1
Lisa	2
Roan	3
Elyse	4
Carl	5
Roan	6
Lisa	7
Roan	8

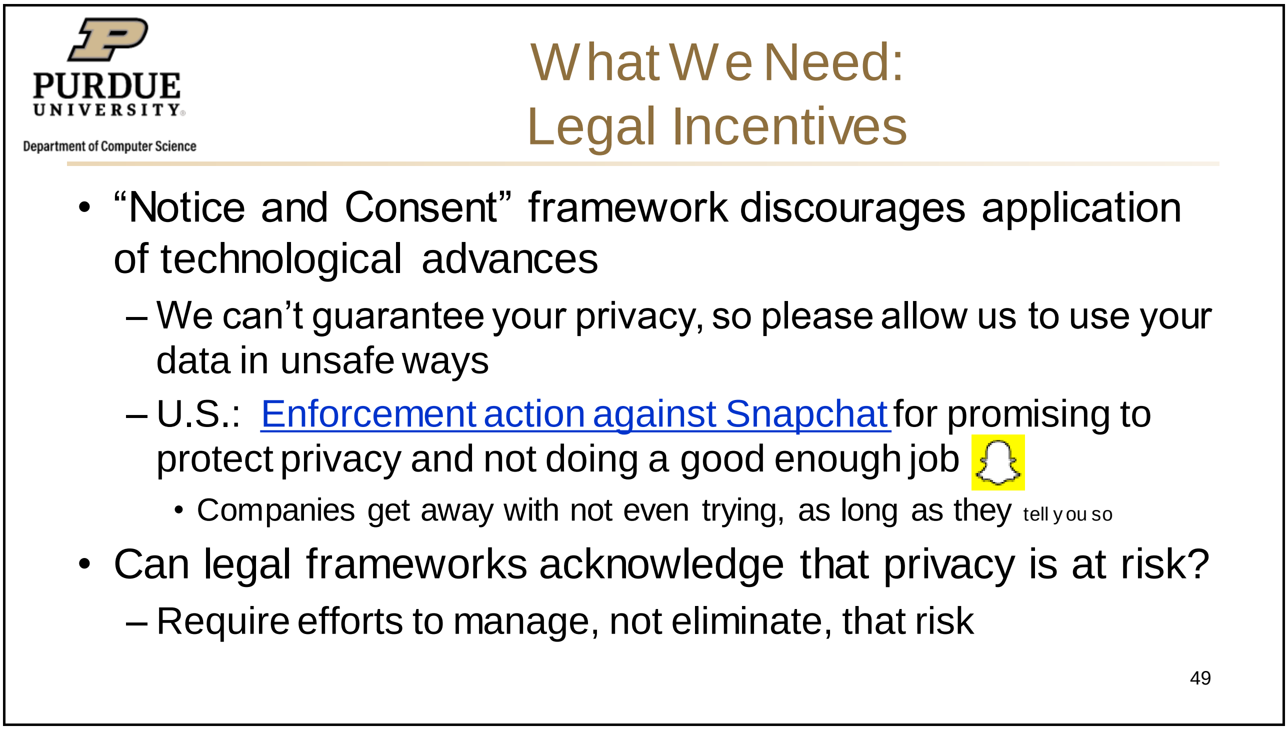
ID	Manufacturer	Drug Name
	Raphe Healthcare	Retinoic Acid
	Raphe Healthcare	Retinoic Acid
	Raphe Healthcare	Retinoic Acid
	Envie De Neuf	Mild Exfoliation
	Emedoutlet	Nexium
	Gep-Tek	Abiraterone
	Jai Radhe	Adapalene
	Hangzhou Btech	Cytarabine



What We Can Do: Noise Addition

- Idea: Impact of noise on what we learn from the data larger than impact of any individual's data
- Formally: For $S \subseteq Range(f)$, an ε-differentially private mechanism M satisfies $\frac{Pr[M_f(D_1) \in S]}{\Pr[M_f(D_2) \in S]} \le e^{\epsilon}$ where D_1 and D_2 differ on at most one element
- *U.S. Census Bureau is starting to use Differential Privacy*

D
f(D) = 17
M_f= f+R
18
e^{ϵ}
$M_f\ (D_1)$
$M_f\ (D_2)$



What We Need: Legal Incentives

- “Notice and Consent” framework discourages application of technological advances
  - We can’t guarantee your privacy, so please allow us to use your data in unsafe ways
  - U.S.: Enforcement action against Snapchat for promising to protect privacy and not doing a good enough job
    - Companies get away with not even trying, as long as they tell you so
- Can legal frameworks acknowledge that privacy is at risk?
  - Require efforts to manage, not eliminate, that risk