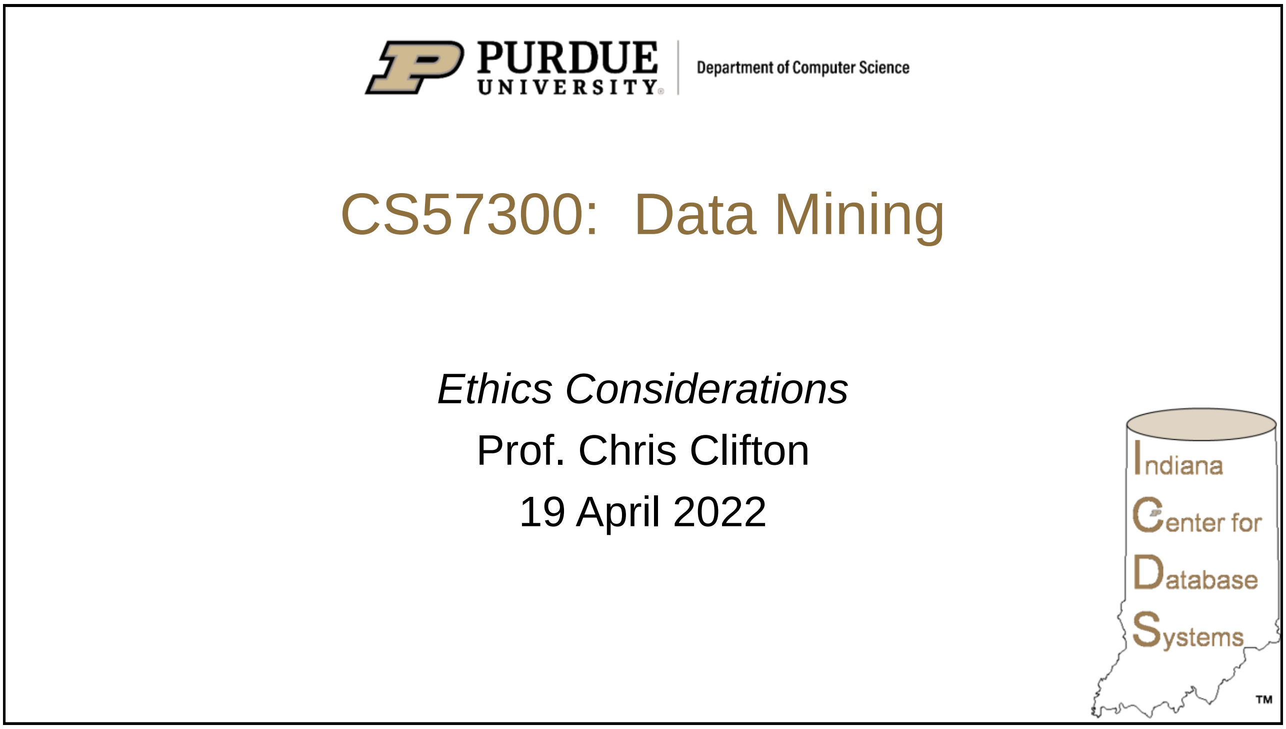


PURDUE UNIVERSITY | Department of Computer Science

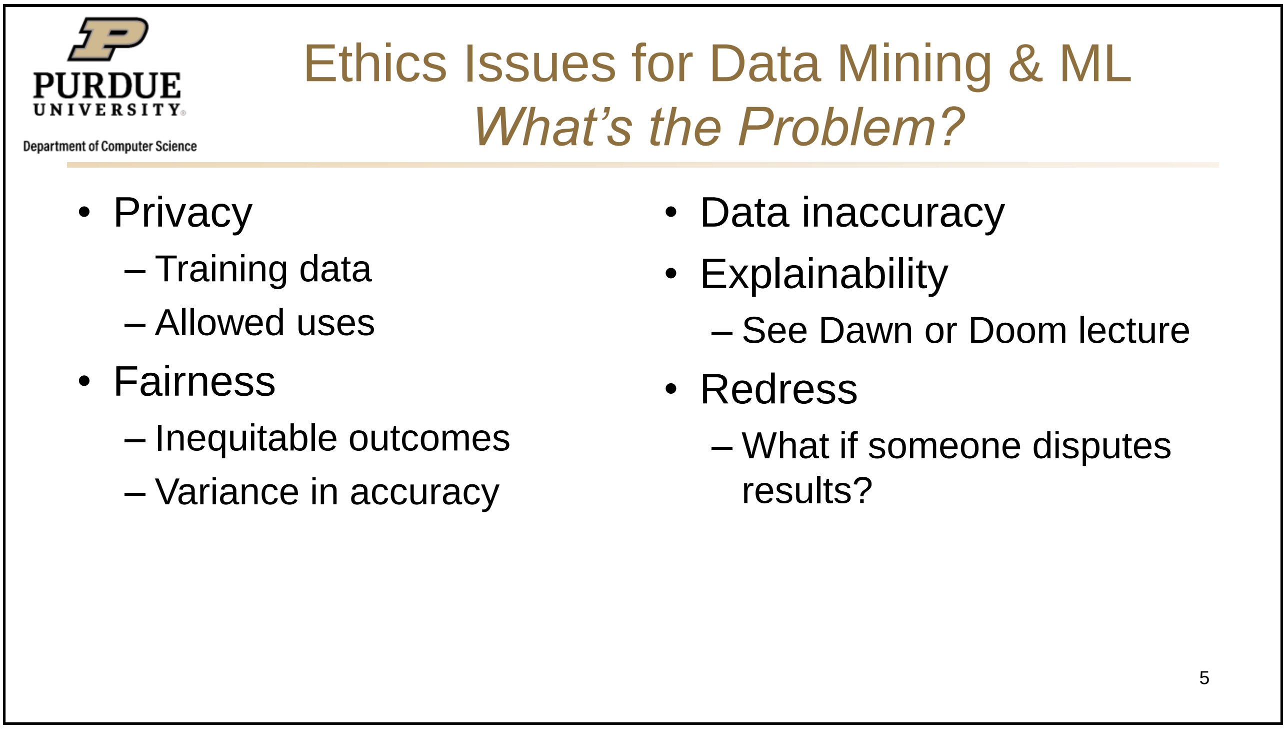
CS57300: Data Mining

Ethics Considerations

Prof. Chris Clifton

19 April 2022

Indiana Center for Database Systems ™



PURDUE UNIVERSITY

Department of Computer Science

Ethics Issues for Data Mining & ML *What's the Problem?*

- Privacy
  - Training data
  - Allowed uses
- Fairness
  - Inequitable outcomes
  - Variance in accuracy
- Data inaccuracy
- Explainability
  - See Dawn or Doom lecture
- Redress
  - What if someone disputes results?

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

“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

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.

The AOL Awakening

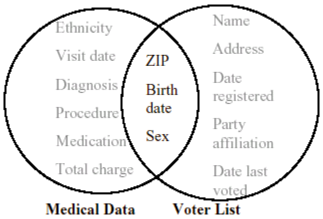
- In Aug 2006, AOL released its customers web searches for research studies
- 20 Million unique queries of 650K unique users
- <user-i
- NY Tim n individual from the queries
  - Queries included “60 single men” “landscapers in Lilburn, Ga”
  - Many more queries contained enough information to uniquely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

AOL fired its CTO over this issue;

Two researchers were forced out

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

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

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.

PURDUE UNIVERSITY
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

PURDUE UNIVERSITY
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*

PURDUE UNIVERSITY
Department of Computer Science

What We Can Do

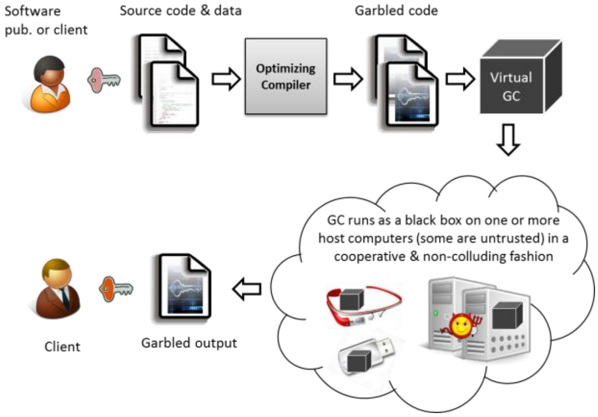
- Encryption
  - Reduce risk to minimal levels when data not in use
- Anonymization
  - Produce usable data that is hard to link to individuals
- Noise addition
  - Usable data where any link to individuals (or information we surmise about individuals) is guaranteed to be uncertain/suspect

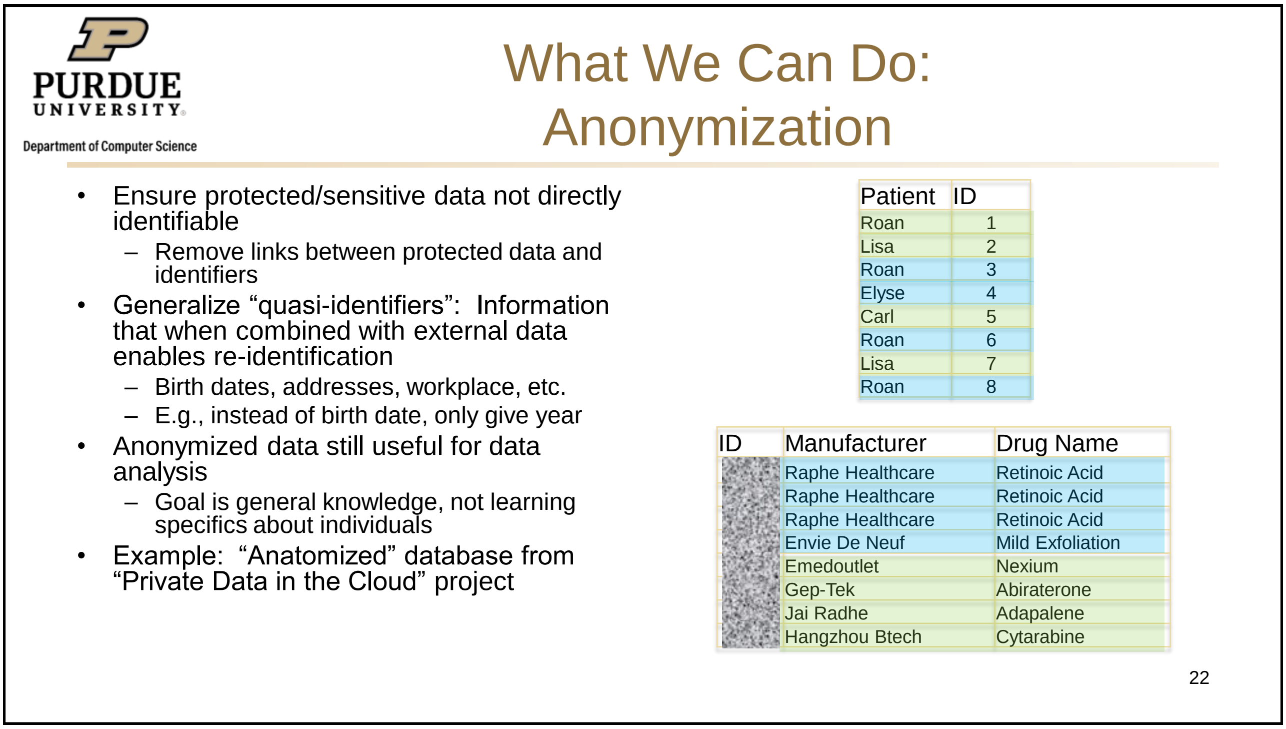
PURDUE UNIVERSITY
Department of Computer Science

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



PURDUE UNIVERSITY

Department of Computer Science

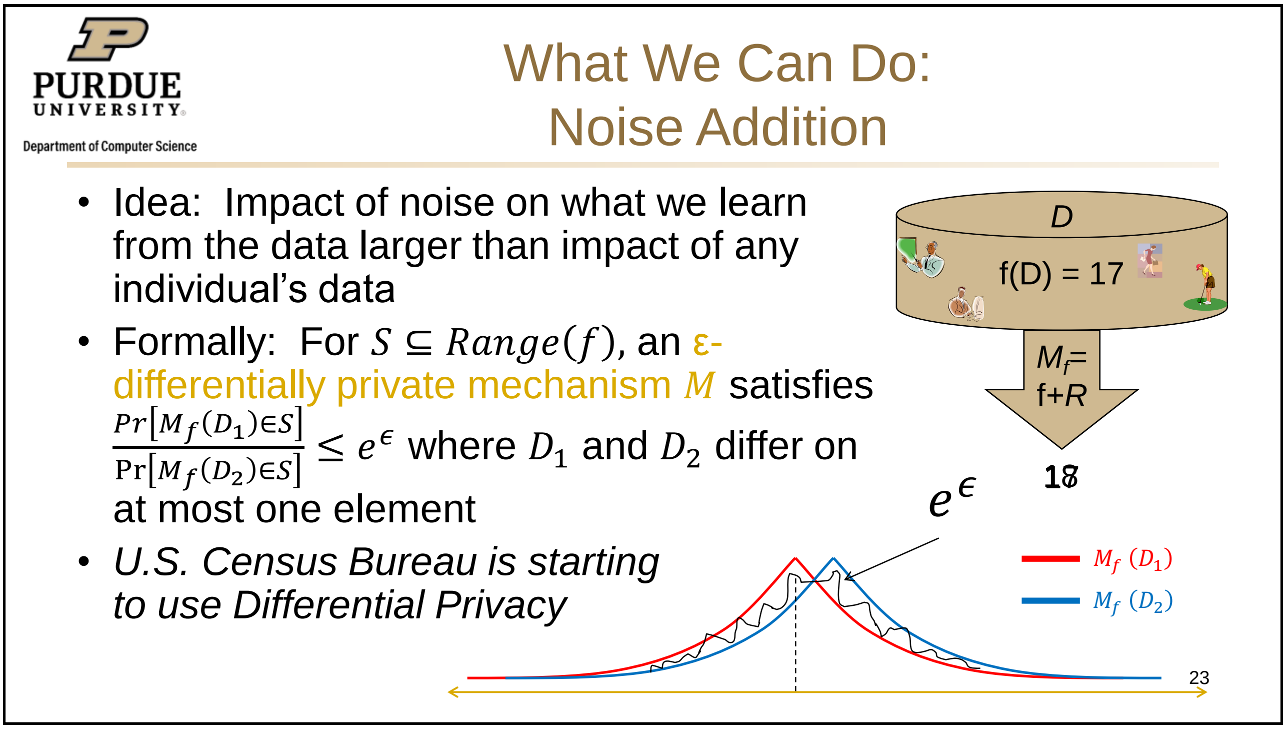
What We Can Do: Anonymization

- 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

22



PURDUE UNIVERSITY

Department of Computer Science

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 $\frac{Pr[M_f(D_1) \in S]}{\Pr[M_f(D_2) \in S]} \leq e^{\epsilon}$ where D_1 and D_2 differ on at most one element
- *U.S. Census Bureau is starting to use Differential Privacy*

23

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

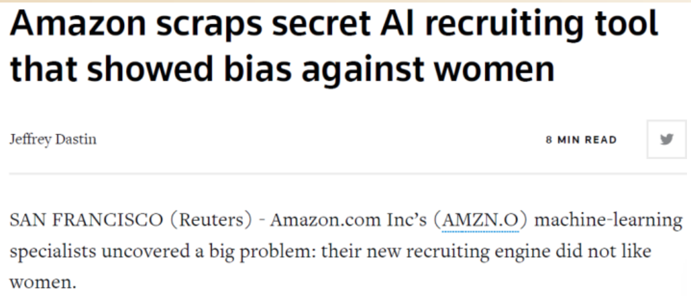
Ethics Issues for Data Mining & ML
What's the Problem?

- Privacy
  - – Training data
  - – Allowed uses
- Fairness
  - – Inequitable outcomes
  - – Variance in accuracy

- Data inaccuracy
- Explainability
  - – See Dawn or Doom lecture
- Redress
  - – What if someone disputes results?

PURDUE UNIVERSITY
Department of Computer Science

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



Amazon scraps secret AI recruiting tool that showed bias against women

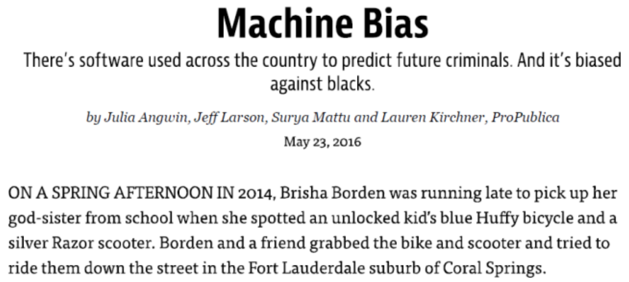
Jeffrey Dastin 8 MIN READ

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*

PURDUE UNIVERSITY
Department of Computer Science

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.

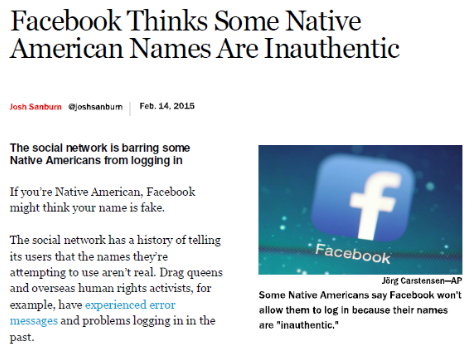
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

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.

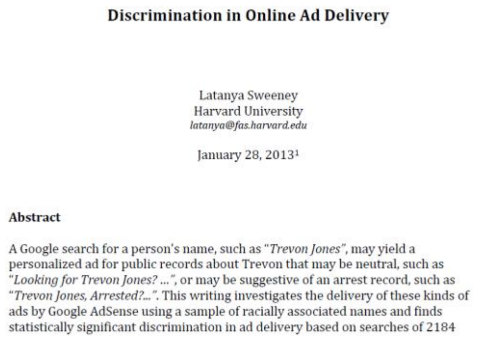
The social network has a history of telling its users that the names they're attempting to use aren't real. Drag queens and overseas human rights activists, for example, have experienced error messages and problems logging in in the past.

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]

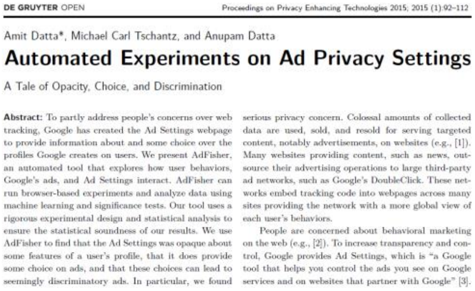
Abstract

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)?

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's 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

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?

Credit Scoring using Decision Trees *(with Abhishek Sharma)*

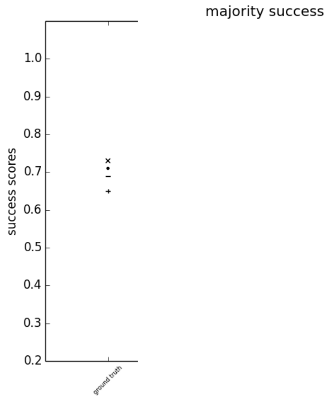
- Experiment in Fairness using Statlog (German Credit Data) Data Set

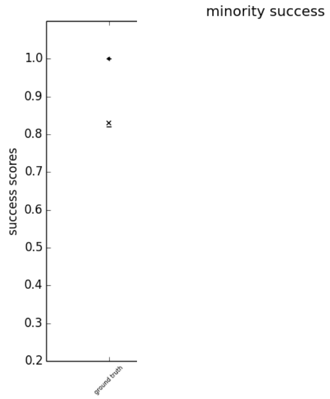
  Data made available by Professor Dr. Hans Hofmann, Universität Hamburg via the UCI Machine Learning Repository

- Learn a decision tree from historical decisions
  - Data about credit applications
  - Decision made
    - *Better training data would be if loan was repaid…*
- Decision tree: model used to make future decisions
  - Goal is to make similar decisions to historical data

PURDUE
UNIVERSITY
Department of Computer Science

Credit Dataset: Majority vs. Minority Positive Decisions

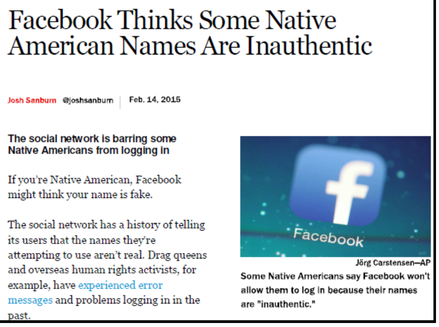




PURDUE
UNIVERSITY
Department of Computer Science

Why is Machine Learning Introducing Bias?

- Key idea: ML typically optimizes for overall accuracy
- What is going on?
  - – Distinct models that work best for majority, minority
  - – Optimizing for global accuracy (revenue, …) selects model that works for majority
- Accurate / effective model for majority
  - – But a bad model for the minority



Facebook Thinks Some Native American Names Are Inauthentic

Josh Sanburn @joshsanburn | Feb. 14, 2015

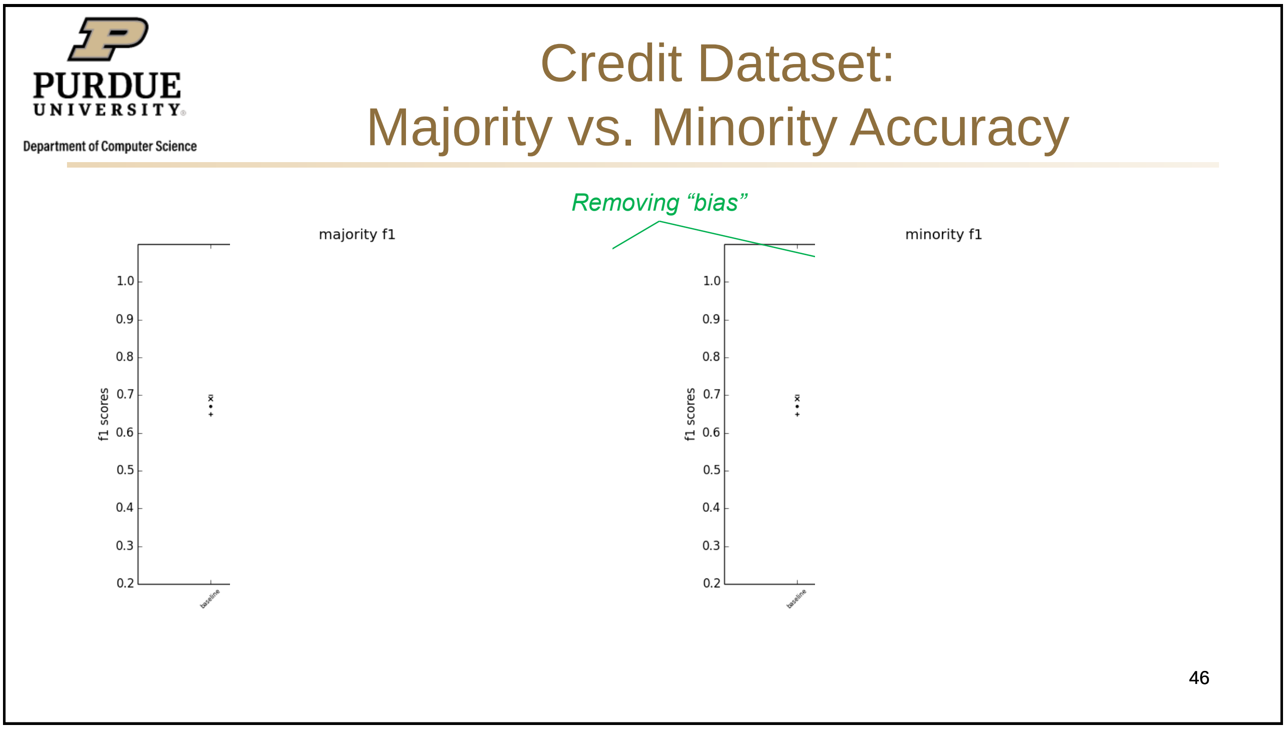
The social network is barring some Native Americans from logging in

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."

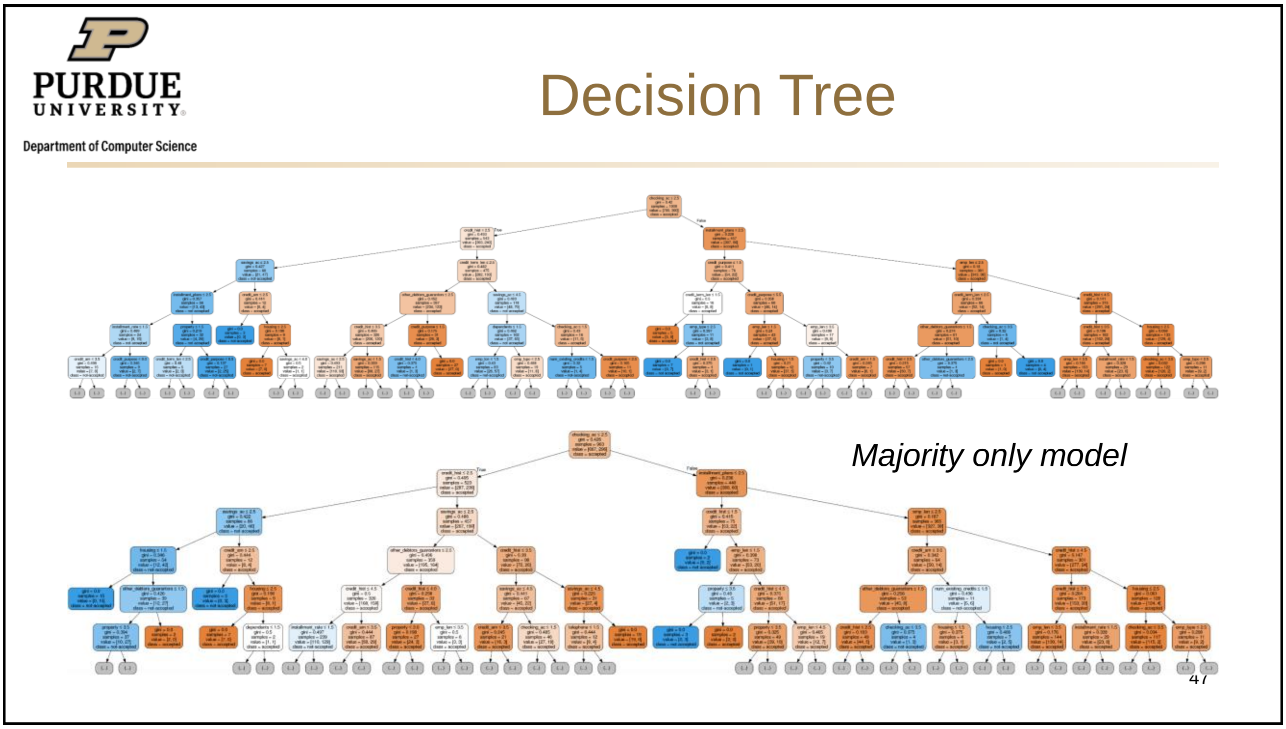


PURDUE UNIVERSITY

Department of Computer Science

Credit Dataset: Majority vs. Minority Accuracy

Removing "bias"



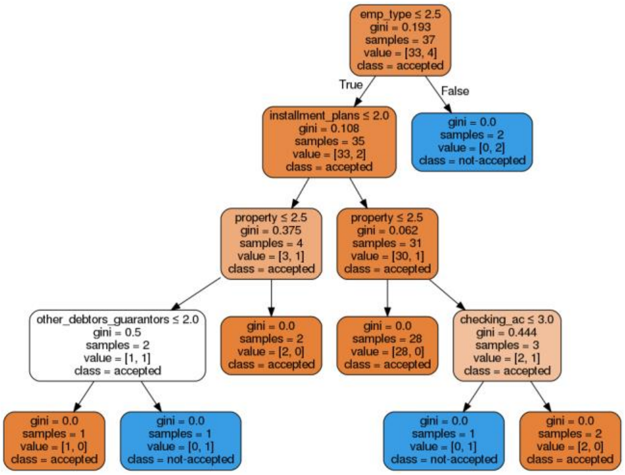
PURDUE UNIVERSITY

Department of Computer Science

Decision Tree

Majority only model

Decision Tree: Minority Only Model


emp_type ≤ 2.5 gini = 0.193 samples = 37 value = [33, 4] class = accepted
True
False
installment_plans ≤ 2.0 gini = 0.108 samples = 35 value = [33, 2] class = accepted
gini = 0.0 samples = 2 value = [0, 2] class = not-accepted
property ≤ 2.5 gini = 0.375 samples = 4 value = [3, 1] class = accepted
property ≤ 2.5 gini = 0.062 samples = 31 value = [30, 1] class = accepted
other_debtors_guarantors ≤ 2.0 gini = 0.5 samples = 2 value = [1, 1] class = accepted
gini = 0.0 samples = 2 value = [2, 0] class = accepted
gini = 0.0 samples = 28 value = [28, 0] class = accepted
checking_ac ≤ 3.0 gini = 0.444 samples = 3 value = [2, 1] class = accepted
gini = 0.0 samples = 1 value = [1, 0] class = accepted
gini = 0.0 samples = 1 value = [0, 1] class = not-accepted
gini = 0.0 samples = 1 value = [0, 1] class = not-accepted
gini = 0.0 samples = 2 value = [2, 0] class = accepted

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

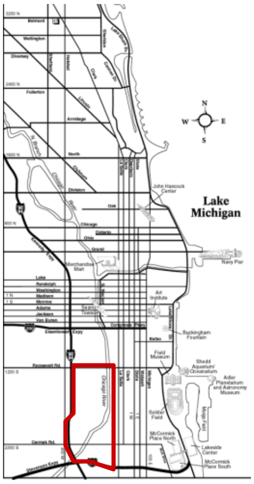
PURDUE UNIVERSITY

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?



PURDUE UNIVERSITY

Department of Computer Science

Balance Training Data

- What if we get rid of majority/minority?
  (with Murat Kantarcioglu and Yan Zhou, UT Dallas)
- Augment training data with synthetic data
  - Generated to be similar to real data
- Synthetic data skewed to eliminate disparity in training data
  - Balance sizes of privileged/unprivileged groups
  - Balance positive/negative outcomes between groups

Is Unbiased Training Data Enough?

- Rakin Haider: ML bias from unbiased data
- Assumptions:
  - Training data correct
  - Privileged and unprivileged groups of same size
  - Positive outcome probability same for both groups
- Difference
  - Different optimal models for the two groups
  - Optimal model for privileged group is higher accuracy

Result: Biased Outcome

- Resource-scarce environment (e.g., selective college admissions): Optimal accuracy global model favors privileged class
  - This wasn’t true in the training data
- Analysis based on Bayesian model
  - Presumably “good” practical ML will do the same
  - Demonstrated on a variety of real-world classifiers
    - *Including some explicitly designed to reduce bias*
- Reflects a type of **Systemic Bias**

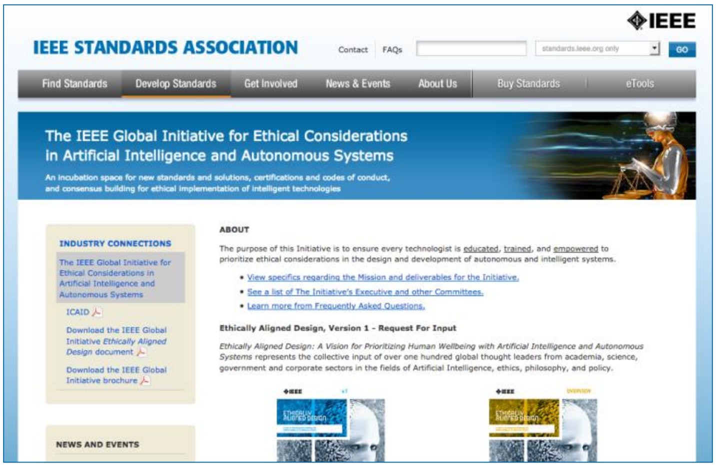
PURDUE
UNIVERSITY

Department of Computer Science

Ideas for the Future

- Tests for Bias?
  - Or perhaps just *potential* bias?
  - ethicstoolkit.ai
- Fundamental changes in machine learning?
  - Objective functions other than accuracy
- IEEE-SA P7003: Standard for Algorithmic Bias Considerations
  - *Work in Progress*
- Understand distinction between Bias and Personalization (supported by the Mellon Foundation):
  - What determines if a recommendation is “Biased” or “Personalized”
  - Explored Participatory Design to elicit issues
  - *Joint work with Kendall Roark (Data Ethicist, Purdue Libraries) and Daniel Kelly (Purdue Philosophy Dept.)*

What do we do about it?
Standards and Best Practices


IEEE
IEEE STANDARDS ASSOCIATION
Contact
FAQs
standards.ieee.org only
GO
Find Standards
Develop Standards
Get Involved
News & Events
About Us
Buy Standards
eTools
INDUSTRY CONNECTIONS
The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems
ICAID
Download the IEEE Global Initiative *Ethically Aligned Design* document
Download the IEEE Global Initiative brochure
NEWS AND EVENTS

The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems

An incubation space for new standards and solutions, certifications and codes of conduct, and consensus building for ethical implementation of intelligent technologies

ABOUT

The purpose of this Initiative is to ensure every technologist is educated, trained, and empowered to prioritize ethical considerations in the design and development of autonomous and intelligent systems.

- View specifics regarding the Mission and deliverables for the Initiative.
- See a list of The Initiative's Executive and other Committees.
- Learn more from Frequently Asked Questions.

Ethically Aligned Design, Version 1 - Request For Input

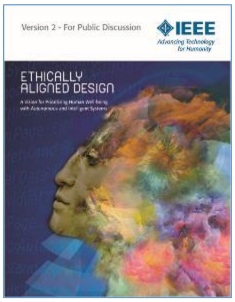
Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems represents the collective input of over one hundred global thought leaders from academia, science, government and corporate sectors in the fields of Artificial Intelligence, ethics, philosophy, and policy.

IEEE STANDARDS ASSOCIATION IEEE

57

Ethically Aligned Design

A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems



Version 2

- Launched December 2017 as a Request for Input
- Created by over 250 Global A/IS & Ethics professionals, in a bottom up, transparent, open and increasingly globally inclusive process
- Incorporates over 200 pages of feedback from public RFI and new Working Groups from China, Japan, Korea and more
- Thirteen Committees / Sections
- Contains **over one hundred twenty** key Issues and Candidate Recommendations

https://ethicsinaction.ieee.org/

IEEE STANDARDS ASSOCIATION IEEE

IEEE P70xx Standards Projects

IEEE P7000: Model Process for Addressing Ethical Concerns During System Design

IEEE P7001: Transparency of Autonomous Systems

IEEE P7002: Data Privacy Process

IEEE P7003: Algorithmic Bias Considerations

IEEE P7004: Child and Student Data Governance

IEEE P7005: Employer Data Governance

IEEE P7006: Personal Data AI Agent Working Group

IEEE P7007: Ontological Standard for Ethically Driven Robotics and Automation

IEEE P7008: Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems

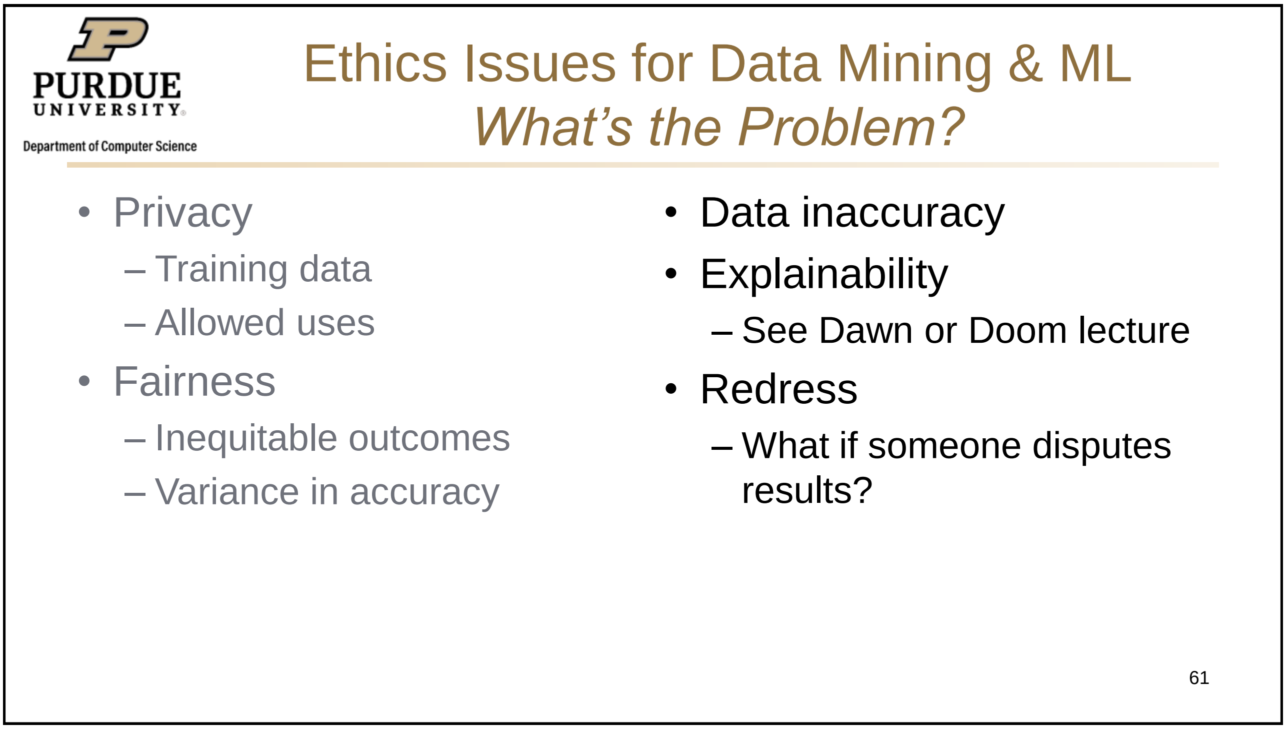
IEEE P7009: Fail-Safe Design of Autonomous and Semi-Autonomous Systems

IEEE P7010: Wellbeing Metrics Standard for Ethical AI and Autonomous Systems

IEEE P7011: Process of Identifying and Rating the Trustworthiness of News Sources

IEEE P7012: Standard for Machines Readable Personal Privacy Terms

IEEE STANDARDS ASSOCIATION IEEE

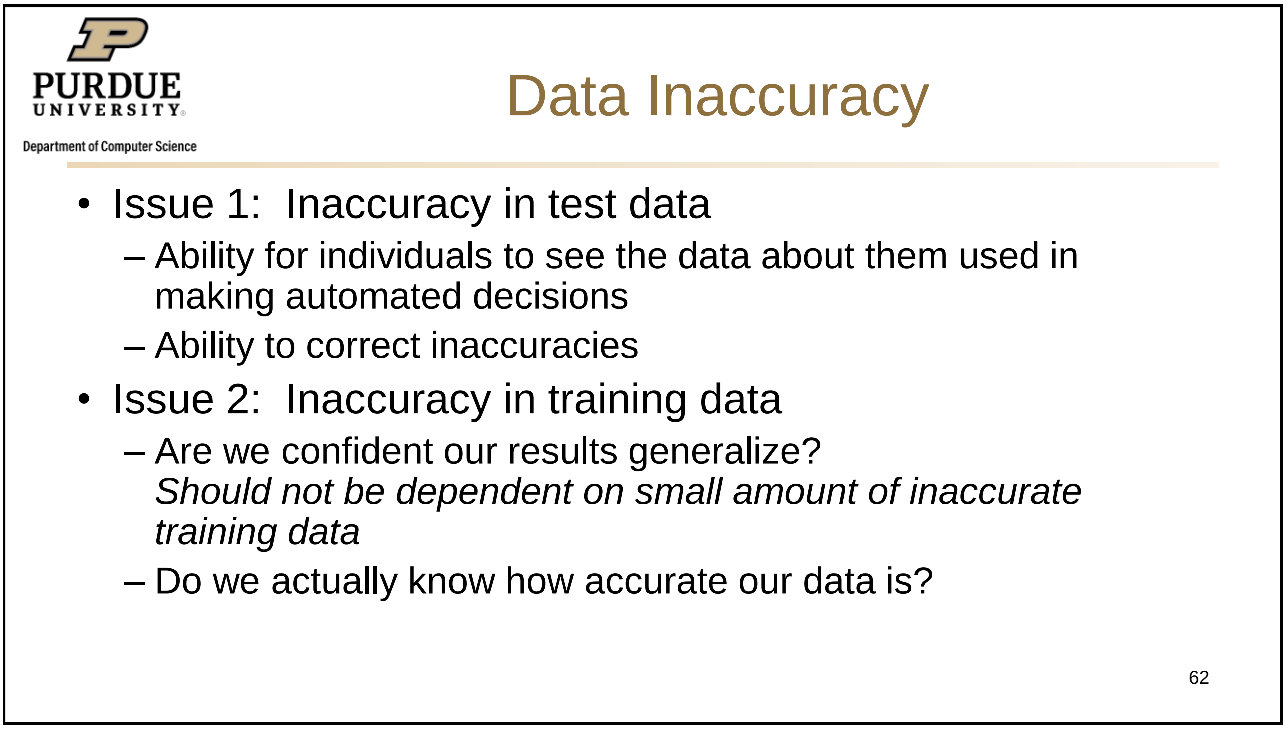


PURDUE UNIVERSITY
Department of Computer Science

Ethics Issues for Data Mining & ML
What's the Problem?

- Privacy
  - Training data
  - Allowed uses
- Fairness
  - Inequitable outcomes
  - Variance in accuracy
- Data inaccuracy
- Explainability
  - See Dawn or Doom lecture
- Redress
  - What if someone disputes results?

61



PURDUE UNIVERSITY
Department of Computer Science

Data Inaccuracy

- Issue 1: Inaccuracy in test data
  - Ability for individuals to see the data about them used in making automated decisions
  - Ability to correct inaccuracies
- Issue 2: Inaccuracy in training data
  - Are we confident our results generalize?
    Should not be dependent on small amount of inaccurate training data
  - Do we actually know how accurate our data is?

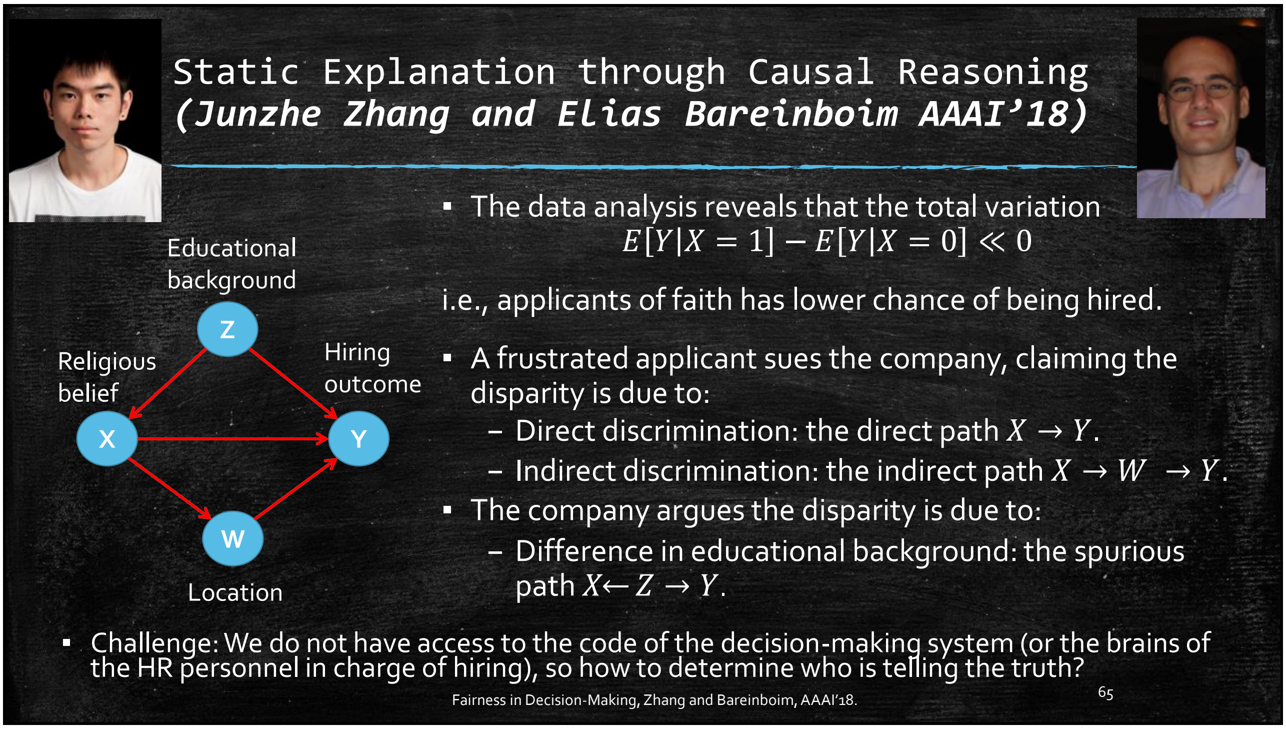
62

Ethics Issues for Data Mining & ML
What's the Problem?

- Privacy
  - Training data
  - Allowed uses
- Fairness
  - Inequitable outcomes
  - Variance in accuracy
- Data inaccuracy
- Explainability
- Redress
  - What if someone disputes results?

First step: Transparency

- Analyze and explain the model
  - Very difficult
  - Likely only understandable to technology and domain experts
- Analyze and explain a decision
  - Input data analysis
  - Static explanation
  - Design/Code review and statistical analysis
  - Sensitivity analysis
  - Reverse-engineering the model

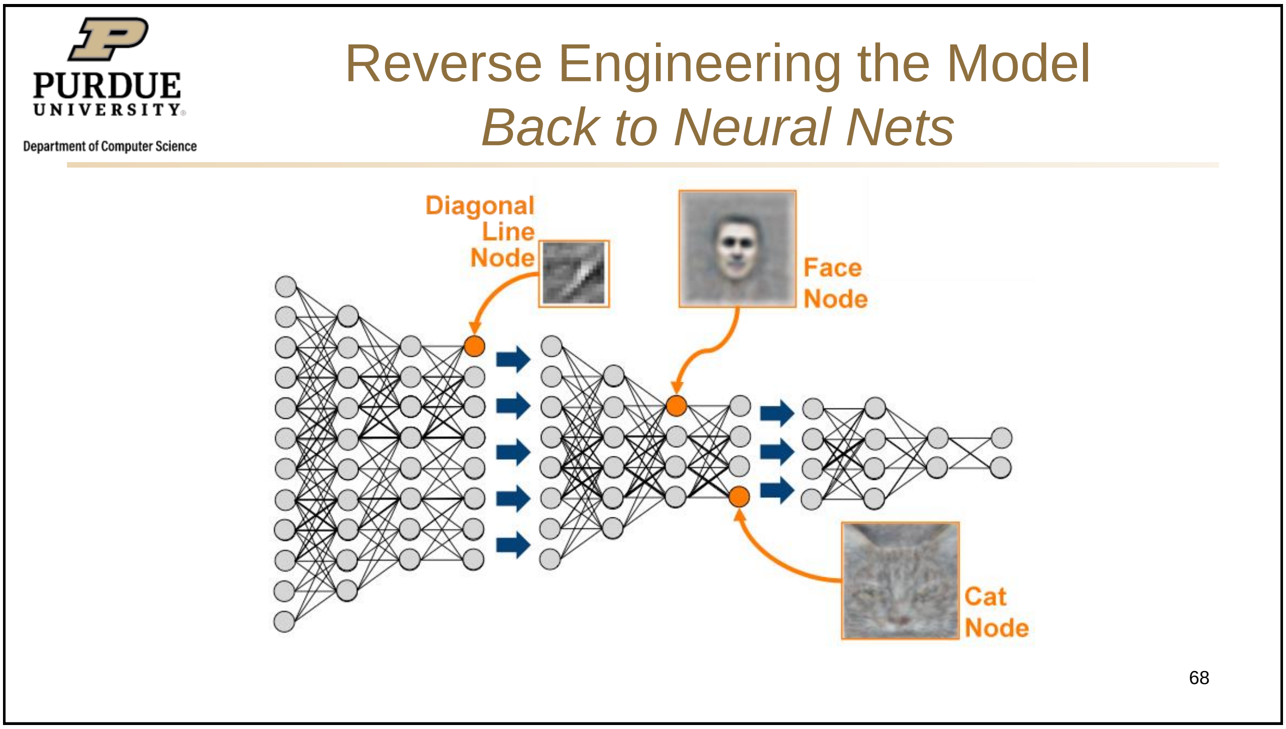


Static Explanation through Causal Reasoning *(Junzhe Zhang and Elias Bareinboim AAAI'18)*

- The data analysis reveals that the total variation
$$E[Y|X=1]-E[Y|X=0] \ll 0$$
i.e., applicants of faith has lower chance of being hired.
- A frustrated applicant sues the company, claiming the disparity is due to:
  - Direct discrimination: the direct path $X \to Y$.
  - Indirect discrimination: the indirect path $X \to W \to Y$.
- The company argues the disparity is due to:
  - Difference in educational background: the spurious path $X \leftarrow Z \to Y$.
- Challenge: We do not have access to the code of the decision-making system (or the brains of the HR personnel in charge of hiring), so how to determine who is telling the truth?

Fairness in Decision-Making, Zhang and Bareinboim, AAAI'18.

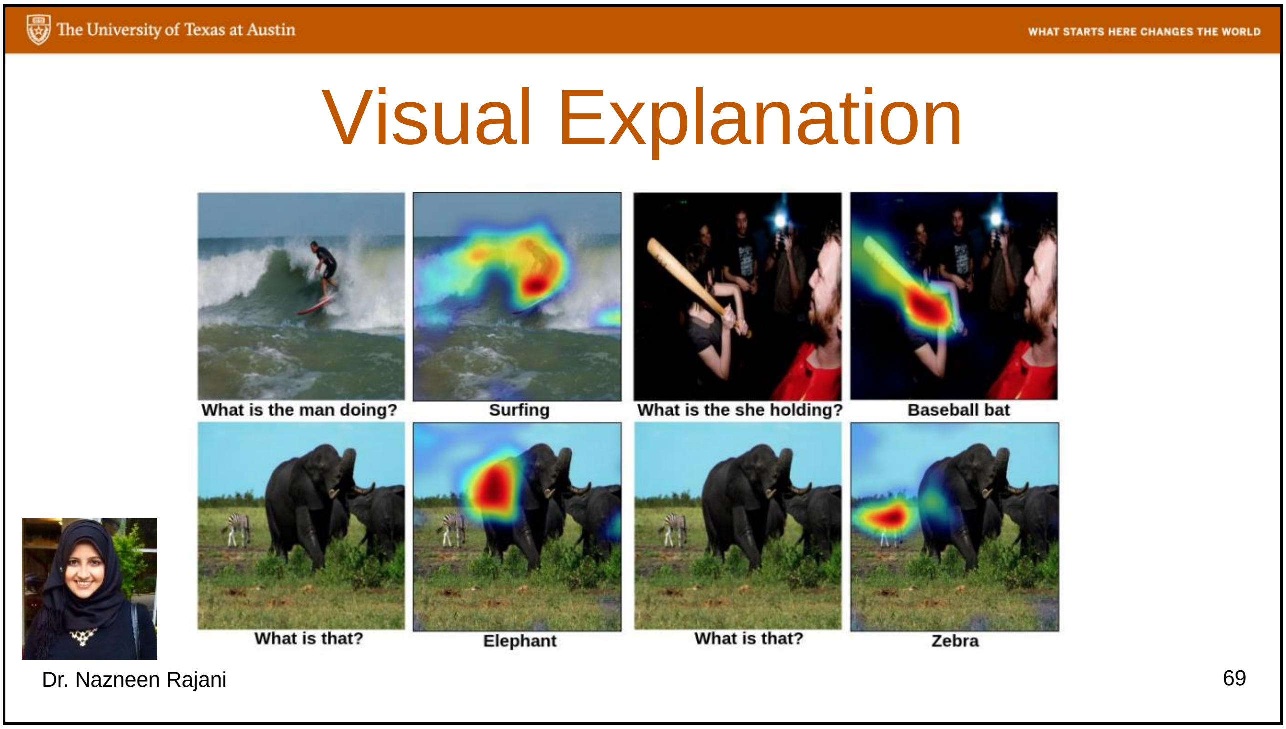
65



PURDUE UNIVERSITY
Department of Computer Science

Reverse Engineering the Model *Back to Neural Nets*

68



The University of Texas at Austin

WHAT STARTS HERE CHANGES THE WORLD

Visual Explanation

What is the man doing?

Surfing

What is the she holding?

Baseball bat

What is that?

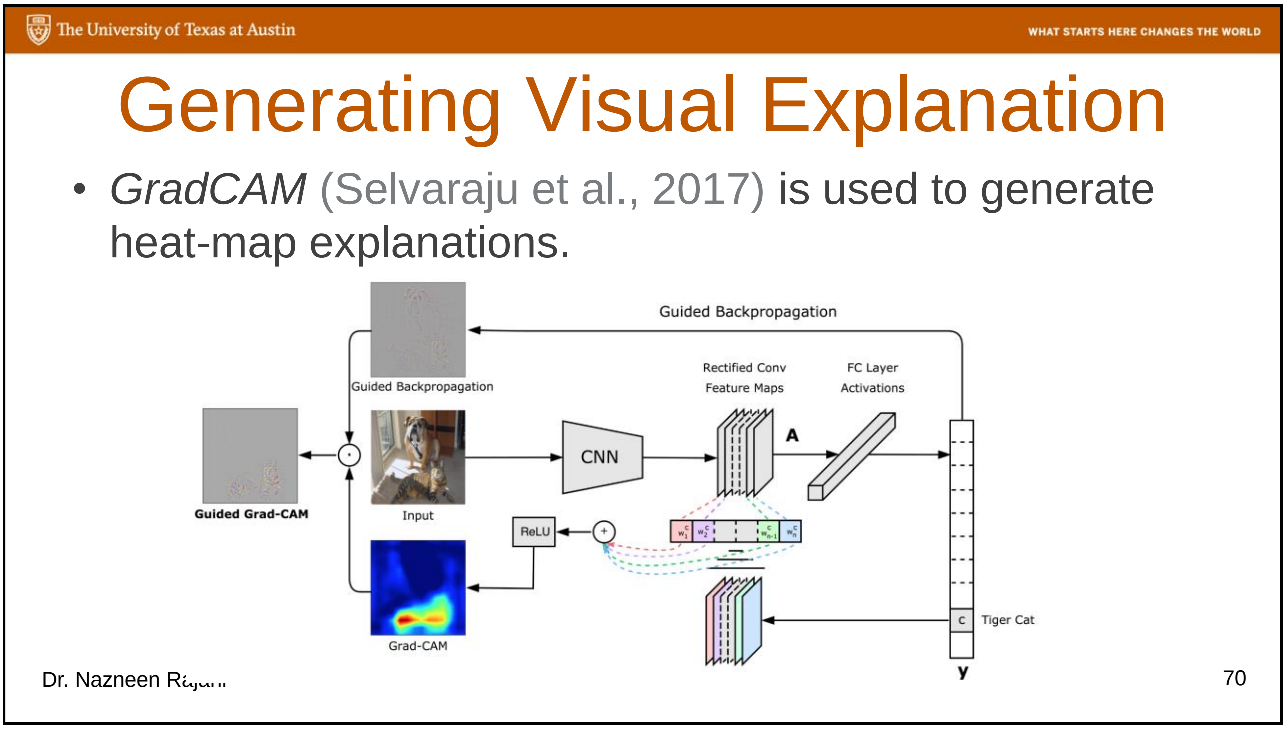
Elephant

What is that?

Zebra

Dr. Nazneen Rajani

69



The University of Texas at Austin

WHAT STARTS HERE CHANGES THE WORLD

Generating Visual Explanation

- *GradCAM* (Selvaraju et al., 2017) is used to generate heat-map explanations.

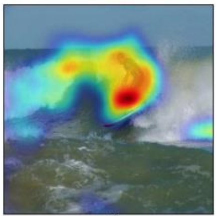
Dr. Nazneen Rajani

70

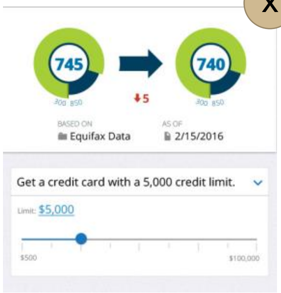
PURDUE UNIVERSITY
Department of Computer Science

Are Explanations Accurate?

- Do these explanations really capture how decisions are made?
  - Sensitivity Analysis, Causal Reasoning
    - Explain outcome, not process
  - Heat maps
    - maybe?
- But does it matter?



Surfing



745 → 740

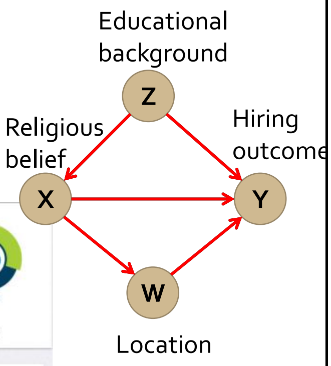
300 850 ↓5 300 850

BASED ON Equifax Data AS OF 2/15/2016

Get a credit card with a 5,000 credit limit.

Limit: $5,000

$500 $100,000



PURDUE UNIVERSITY
Department of Computer Science

Emotional vs. Rational Decision-Making

- Humans have been shown to be emotional in their decision making
  - fMRI analysis of how decisions are made *(De Martino, Kumaran, Seymour, Dolan, Science 2006)*
- We rationalize our decisions
  - Explanations justify why we the decisions are good, not how we make them
- Is this good enough for explaining AI?
  - *Does this qualify as making ethical decisions?*

PURDUE UNIVERSITY
Department of Computer Science

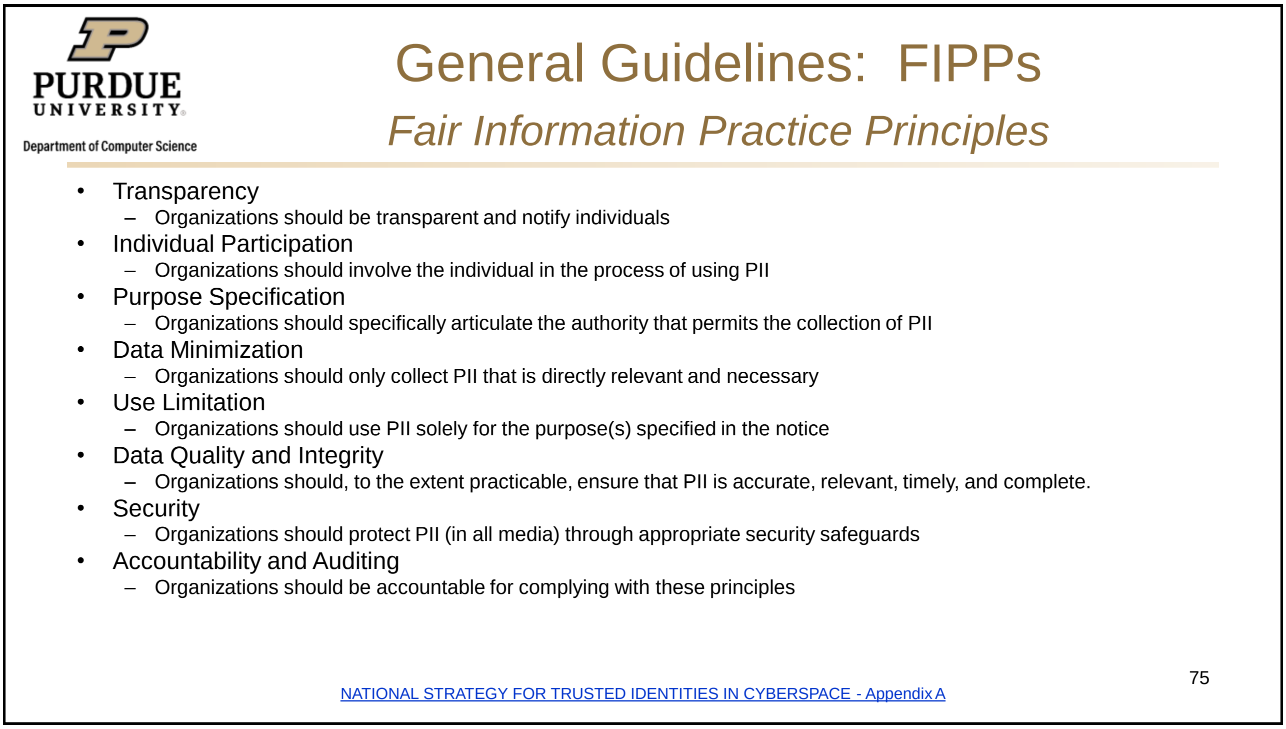
Ethics Issues for Data Mining & ML
What's the Problem?

- Privacy
  - Training data
  - Allowed uses
- Fairness
  - Inequitable outcomes
  - Variance in accuracy
- Data inaccuracy
- Explainability
- Redress
  - What if someone disputes results?

PURDUE UNIVERSITY
Department of Computer Science

(Not?) Understanding AI

- We may not fully understand how AI does what it does
  - We want AI to solve hard problems!
  - *If we can't solve the problem, should we expect to understand how AI does it?*
- We do want reasons why AI is doing the right thing
  - *We're figuring out how to do this*
- We just need to make sure AI does the right thing

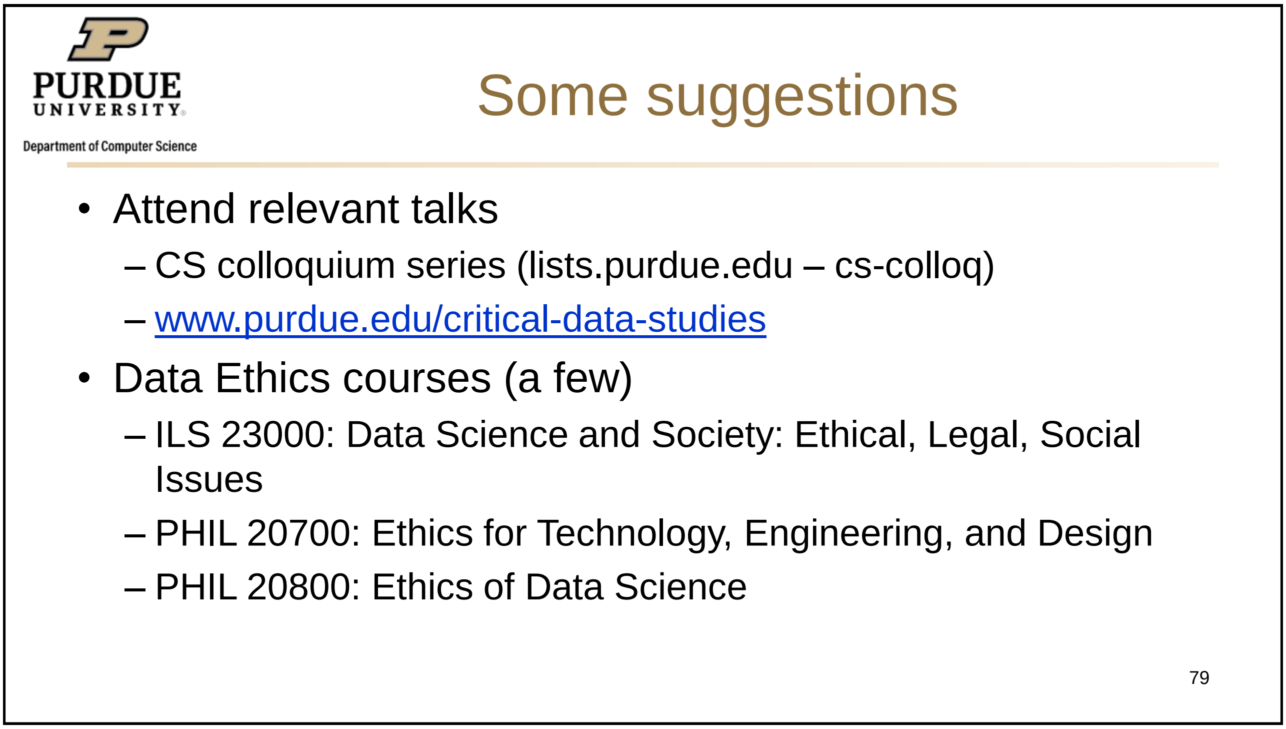


General Guidelines: FIPPs

Fair Information Practice Principles

- Transparency
  - Organizations should be transparent and notify individuals
- Individual Participation
  - Organizations should involve the individual in the process of using PII
- Purpose Specification
  - Organizations should specifically articulate the authority that permits the collection of PII
- Data Minimization
  - Organizations should only collect PII that is directly relevant and necessary
- Use Limitation
  - Organizations should use PII solely for the purpose(s) specified in the notice
- Data Quality and Integrity
  - Organizations should, to the extent practicable, ensure that PII is accurate, relevant, timely, and complete.
- Security
  - Organizations should protect PII (in all media) through appropriate security safeguards
- Accountability and Auditing
  - Organizations should be accountable for complying with these principles

NATIONAL STRATEGY FOR TRUSTED IDENTITIES IN CYBERSPACE - Appendix A



Some suggestions

- Attend relevant talks
  - CS colloquium series (lists.purdue.edu – cs-colloq)
  - www.purdue.edu/critical-data-studies
- Data Ethics courses (a few)
  - ILS 23000: Data Science and Society: Ethical, Legal, Social Issues
  - PHIL 20700: Ethics for Technology, Engineering, and Design
  - PHIL 20800: Ethics of Data Science