

CMPT 733

Responsible Data Science

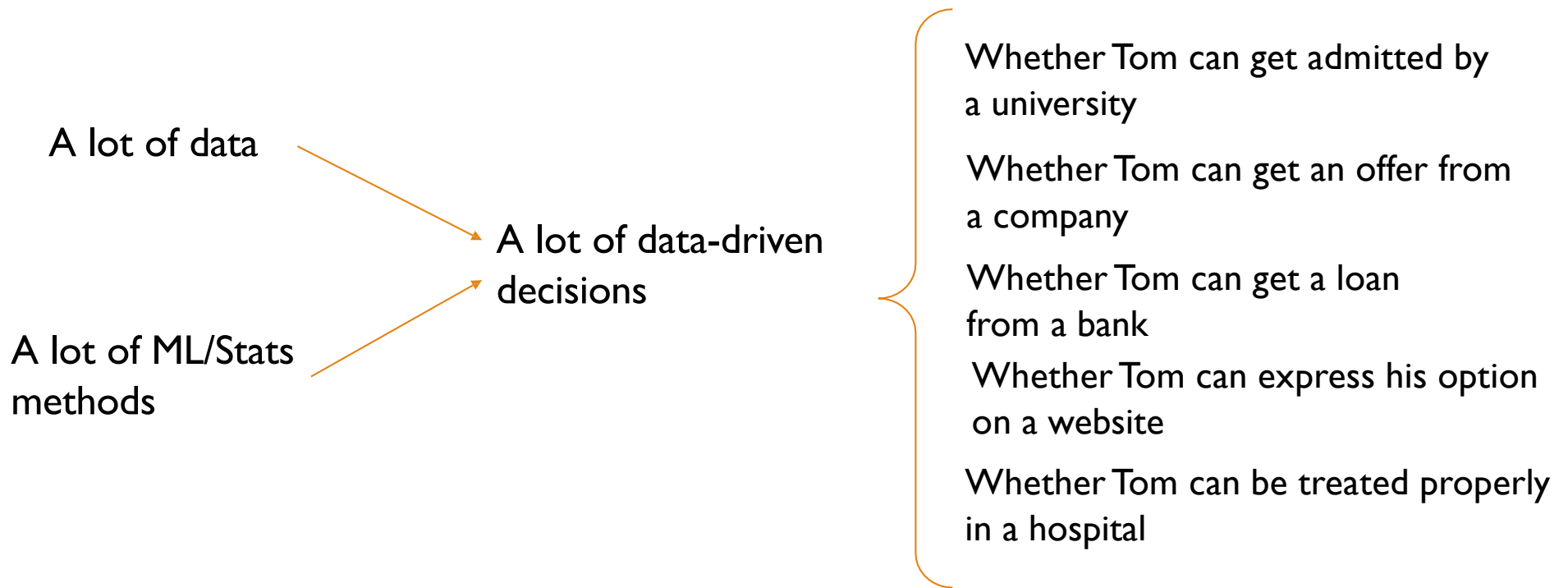
Instructor

Zhengjie Miao

Course website

<https://coursys.sfu.ca/2025sp-cmpt-733-gl/pages/>

Data scientists have a lot of power

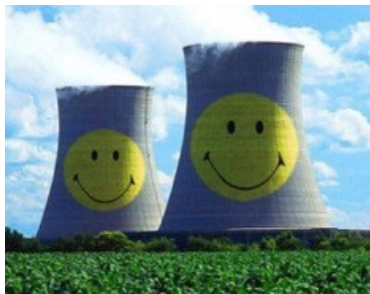


Some quotes

- Information is “news, facts, or knowledge” – Cambridge Dictionary
- “Knowledge is power” – Francis Bacon
- “with great power comes great responsibility” – Uncle Ben

What is a right decision?

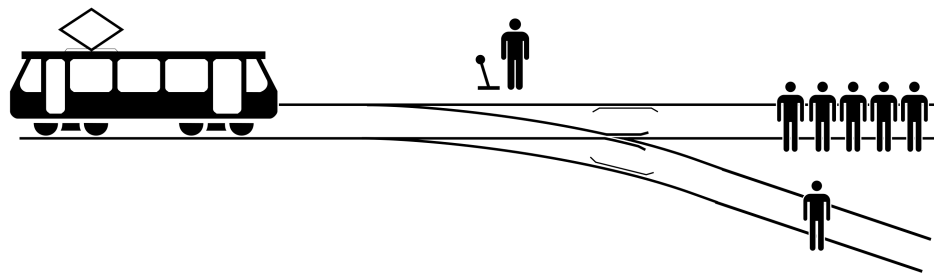
EASY



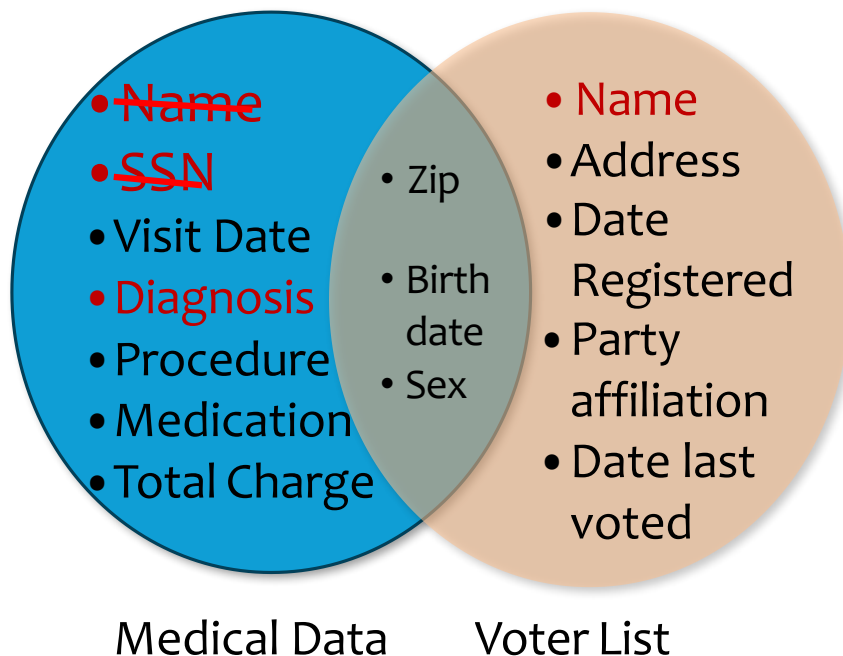
or



HARD



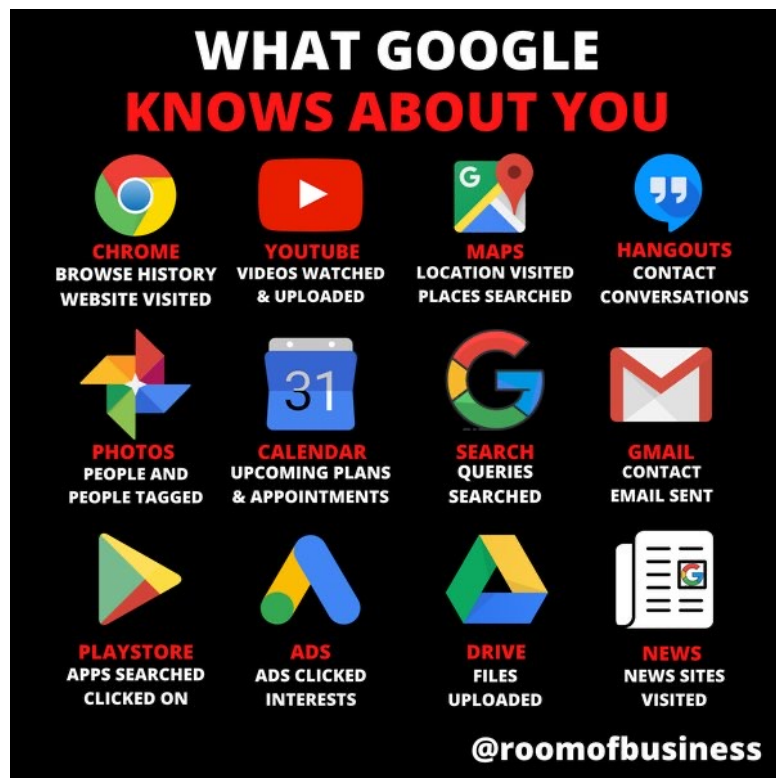
Data → responsibility



- The Massachusetts Governor Privacy Breach [Sweeney IJUFKS 2002]
 - 87 % of US population uniquely identified using ZipCode, Birth Date, and Sex.
 - Governor of MA uniquely identified

Name linked to Diagnosis

Data → responsibility



- Google knows you more than yourself

Data → responsibility

Facebook political microtargeting at center of GDPR complaints in Germany

Natasha Lomas @riptari / 2:00 AM EDT • March 21, 2023

 Comment



In its latest piece of strategic litigation, the precision-punching European privacy rights campaign group noyb has used data donated by users of the 'Who Targets me' browser extension, which analyzes political microtargeting on Facebook, to build a case against every political party in Germany — for what it alleges is unlawful processing of voters' personal data via Facebook's adtech platform during the 2021 federal elections.

<https://techcrunch.com/2023/03/20/facebook-political-ads-germany-gdpr-complaints-noyb/>

Data → responsibility



In 2016, a team of journalists from ProPublica constructed a dataset of more than 7000 individuals arrested in Broward County, Florida between 2013 and 2014 in order to analyze the efficacy of COMPAS.

In addition, they collected data on future arrests for these defendants through the end of March 2016.

“<..> was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.”

Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Data → responsibility



World Business Markets Breakingviews Video More

- ““Everyone wanted this holy grail,” one of the people said. “They literally wanted it to be an engine where I’m going to give you 100 resumes, it will spit out the top five, and we’ll hire those.”
- But by 2015, the company realized its new system was not rating candidates for software developer jobs and other technical posts in a gender-neutral way.
- That is because Amazon’s computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry.”

RETAIL OCTOBER 10, 2018 / 4:04 PM / UPDATED 4 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

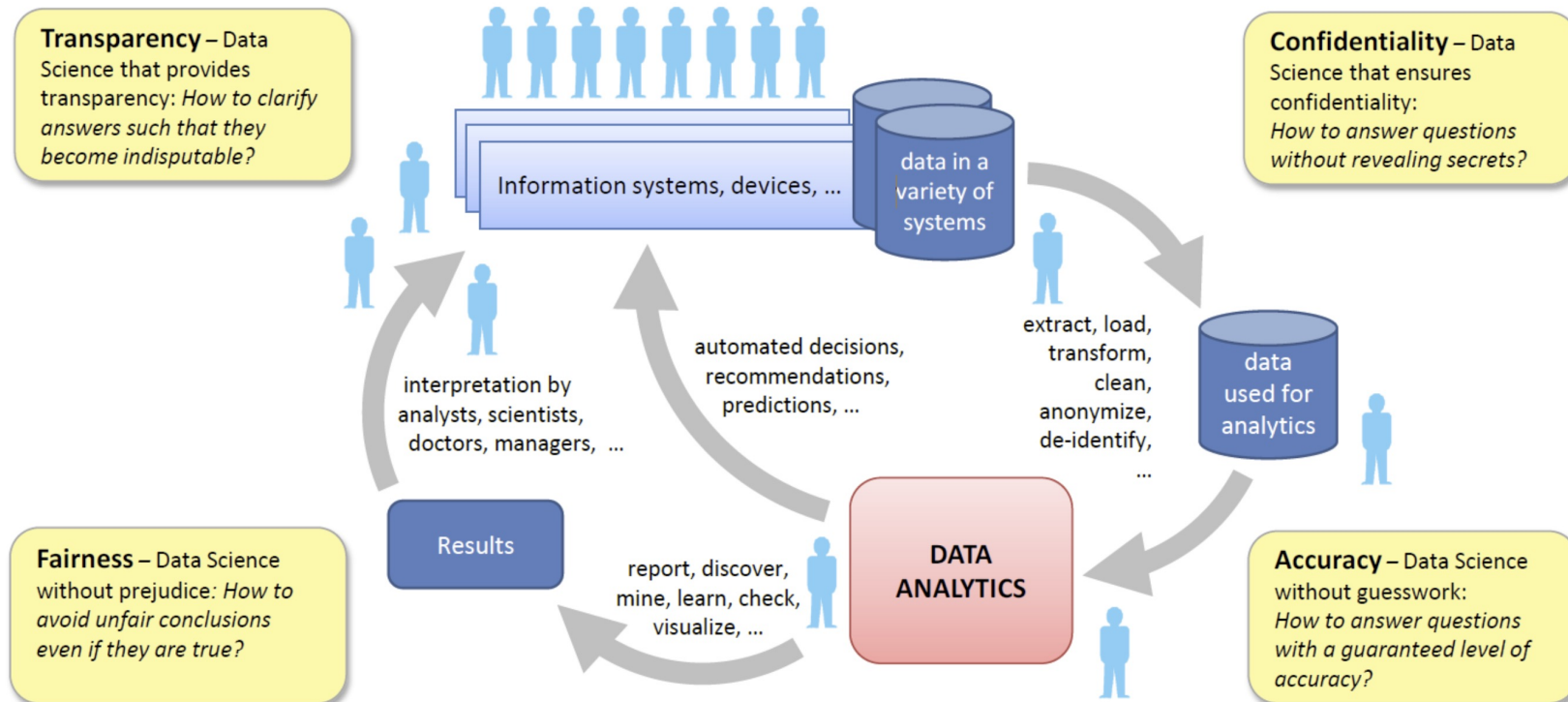
8 MIN READ



SAN FRANCISCO (Reuters) - Amazon.com Inc’s [AMZN.O](#) machine-learning

<https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-%20showed-bias-against-women-idUSKCN1MK08G/>

Data Science Ethics



<https://redasci.org/>

Algorithmic Fairness

Overcoming Racial Bias In AI Systems And Startlingly Even In AI Self-Driving Cars

Racial bias in a medical algorithm favors white patients over sicker black patients

AI expert calls for end to UK use of 'racially biased' algorithms

AI Bias Could Put Women's Lives At Risk - A Challenge For Regulators

Gender bias in AI: building fairer algorithms

Bias in AI: A problem recognized but still unresolved

Amazon, Apple, Google, IBM, and Microsoft worse at transcribing black people's voices than white people's with AI voice recognition, study finds

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

When It Comes to Gorillas, Google Photos Remains Blind

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

The Week in Tech: Algorithmic Bias Is Bad. Uncovering It Is Good.

Artificial Intelligence has a gender bias problem – just ask Siri

The Best Algorithms Struggle to Recognize Black Faces Equally

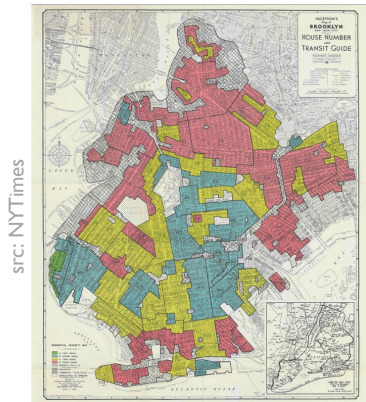
US government tests find even top-performing facial recognition systems misidentify blacks at rates five to 10 times higher than they do whites.

source: <https://towardsdatascience.com/algorithm-bias-in-artificial-intelligence-needs-to-be-discussed-and-addressed-8d369d675a70>

Algorithmic Fairness

- Algorithm bias is the lack of fairness that emerges from the output of a computer system
- Fairness is typically defined in terms of **invariance** of algorithmic decisions to variables that considered as sensitive
- Examples of sensitive variables: gender, ethnicity, sexual orientation, disability, etc.

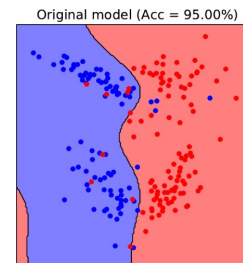
What are the sources of bias?



Historical bias in training data

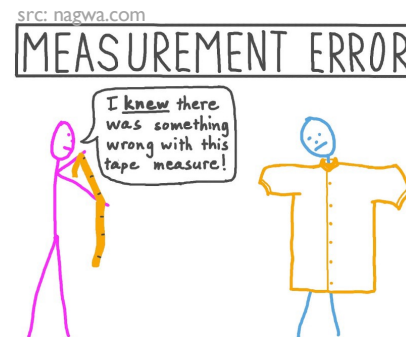


Selection bias



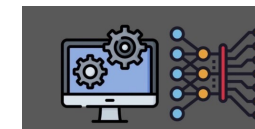
Adversarial data attacks

src: <https://labs.f-secure.com>



Data integration

src: nagwa.com



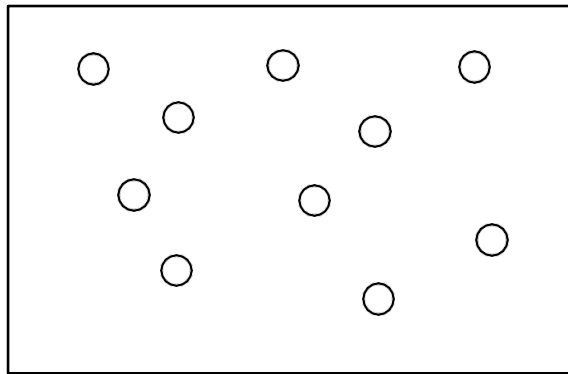
Model design choices

Hooker, Sara. "Moving beyond "algorithmic bias is a data problem"." Patterns 2.4 (2021): 100241.

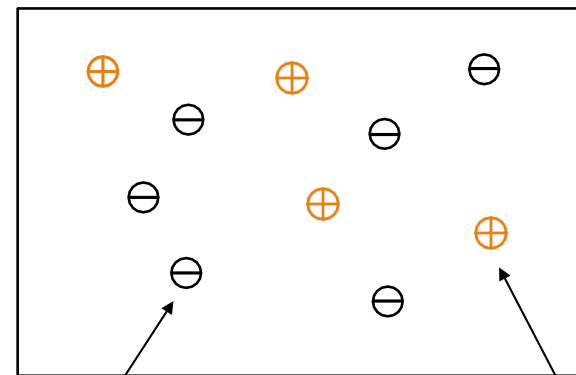
How to formalize and measure “Fairness”?

Fairness in Machine Learning

Is my model fair?



Admit 40% students to MPCS



Not
admit

Admit

Recap: Image Classification

- We'd like to learn a cat classifier, which is a function f from the input space to a class
 - In this example, input space = {pictures}, represented as a vector x of pixel values
 - $\text{class} \in \{0, 1\}$
- Ideally,

$$\bullet f \left(\begin{array}{c} \text{[Image of a cat]} \end{array} \right) = 1.$$

Fair Classification

- Intuitive question of fairness: are people being treated equally?
- Is our classifier working as well for Persian cats as Russian Blue?



- What is “treated equally?”

Fair Classification

- What does it mean to be fair in binary classification?

- It is unfair to classify



as a cat,

- but classify



as not a cat

Individual Definitions

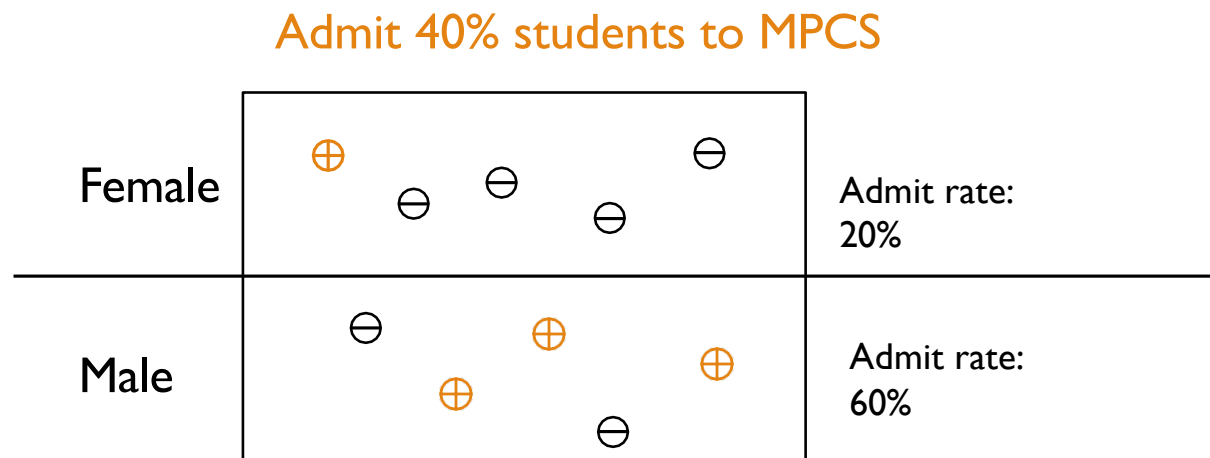
- Fairness Through Awareness [DHP+12]:
 - Similar individuals should be treated similarly
 - Require distance metrics for both data points and classification results.
 - Add the fairness constraint when minimizing the loss function:
 - For all x, y , $\text{Dist}(f(x), f(y)) \leq \text{dist}(x, y)$

Individual Definitions

- Only fair *ex ante*
- Distance metrics are hard to find
- Discriminate groups
 - A “fair” model could reject all students in one group

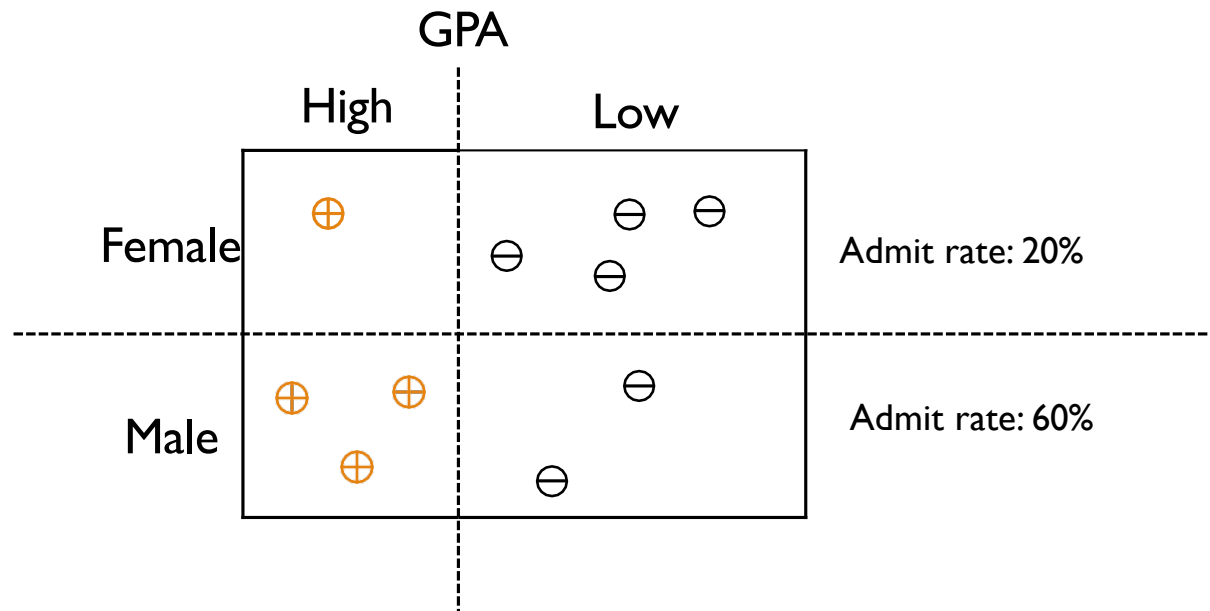
Fairness in Machine Learning

Female and male applicants are treated differently



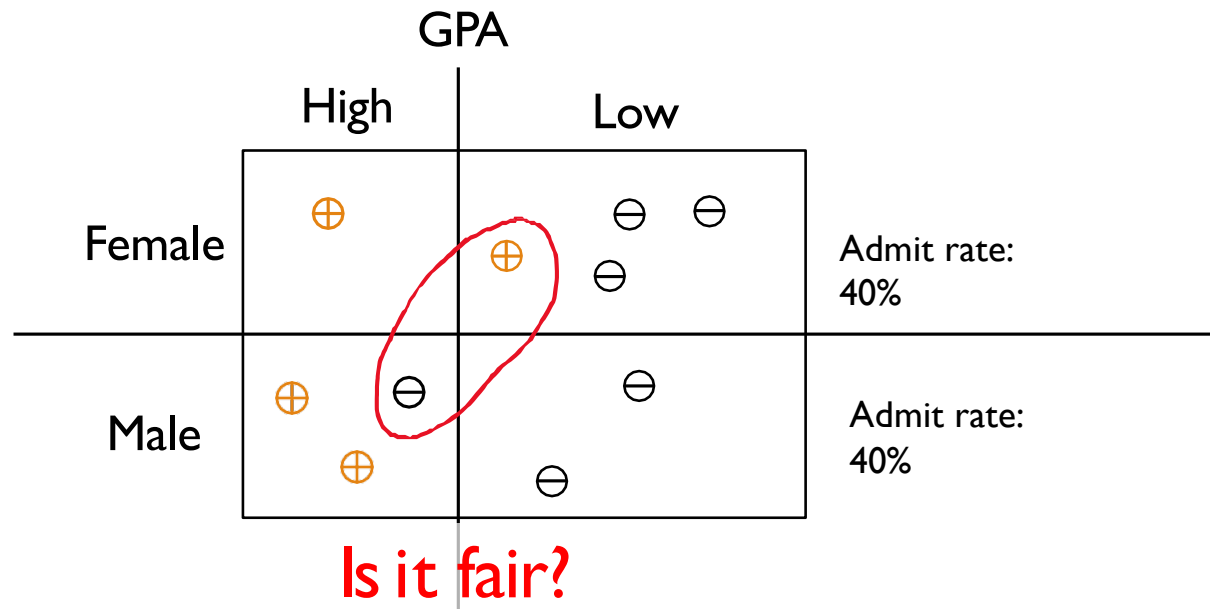
Fairness in Machine Learning

How to make my model fair?



Fairness in Machine Learning

How to make my model fair?



Two notions of fairness

Equality

Giving everyone the same thing



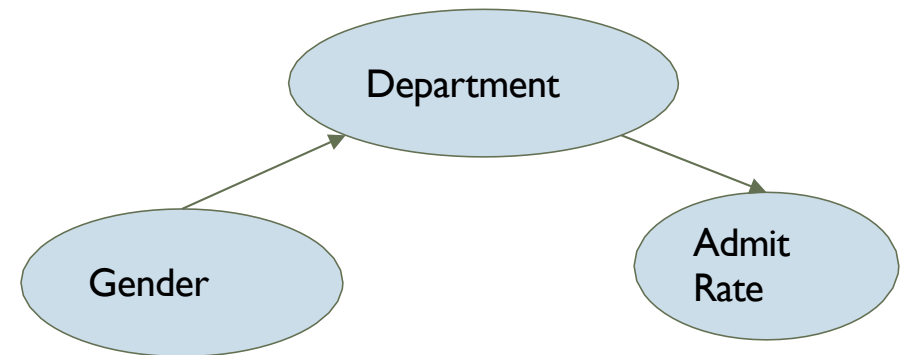
Equity

Giving everyone access to the same opportunity

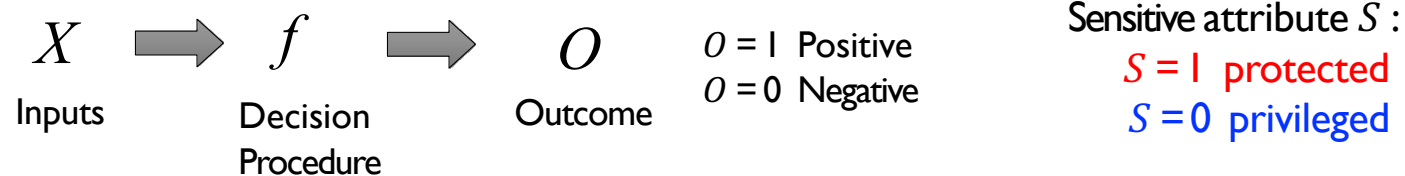


RECAP: Fairness in college admission

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%
Total	2691	45%	1835	30%



Fair Classification: Think about some intuitive definitions of “fairness”



X: Features and qualifications: age, hobbies, test scores, grades, etc.

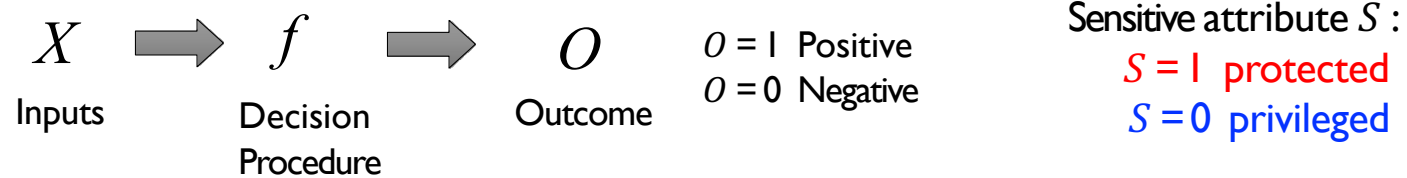
S: Gender

O: Admission Decisions

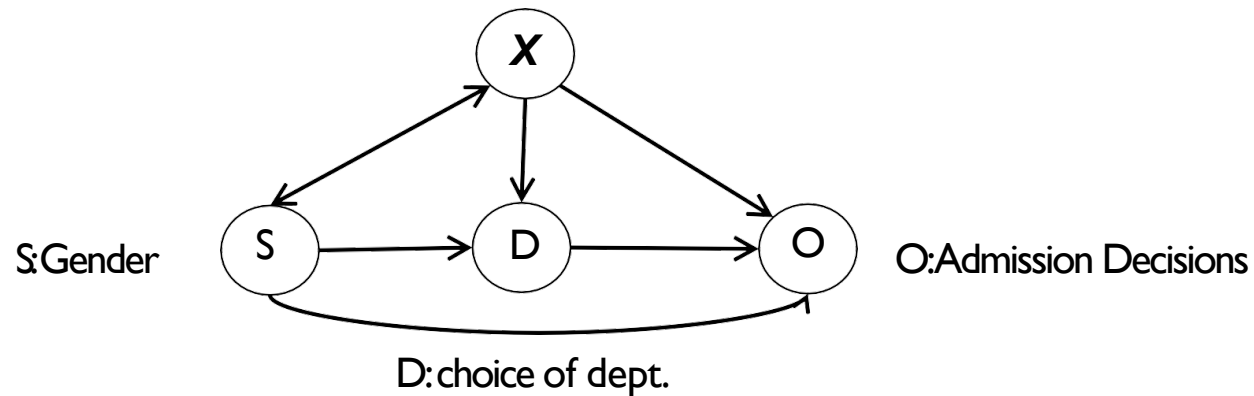
Other factors:

- **D** = department they applied to
- **Y** = Whether they successfully graduate if they are admitted

Fair Classification



X : Features and qualifications: age, hobbies, test scores, grades, etc.



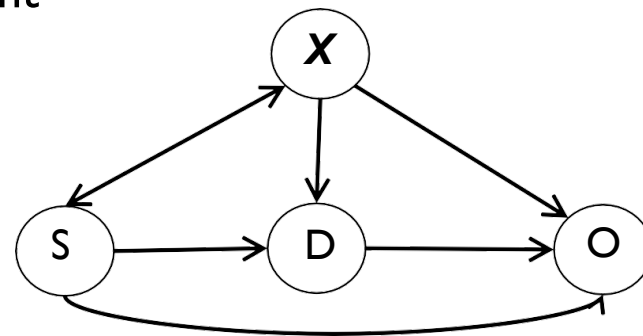
Group Fairness

- Demographic Parity
 - a.k.a. Statistical Parity or Benchmarking
 - $\mathbb{P}(O=1 | S=1) = \mathbb{P}(O=1 | S=0)$

Same fraction of admitted males and females

S and O should be marginally independent

$$O \perp\!\!\!\perp S$$



Group Fairness

- We say that a classifier f has **disparate impact** (DI) of τ ($0 < \tau < 1$) if

$$\frac{\mathbb{P}(O = 1 \mid S = 1)}{\mathbb{P}(O = 1 \mid S = 0)} \leq \tau$$

- The protected class is positively classified less than τ times as often as the unprotected class.
- A disparate impact ratio below **0.8 (80%)** typically suggests potential unfairness or discrimination (legal metric used in practice)
- **For example, if the fraction of admitted women is less than 80% of the fraction of admitted men, it signals potential disparate impact**

Group Fairness

- Demographic Parity
 - a.k.a. Statistical Parity or Benchmarking
 - $\mathbb{P}(O=1 | S=1) = \mathbb{P}(O=1 | S=0)$

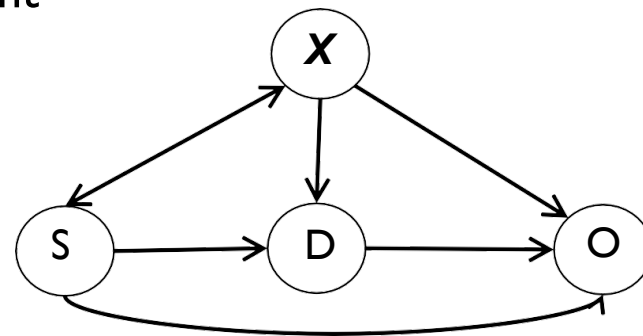
Same fraction of admitted men and women

S and O should be marginally independent

$$O \perp\!\!\!\perp S$$

Can it be ensured if decision are not based on S?
(Fairness through Blindness/unawareness)

Other issues?



Group Fairness

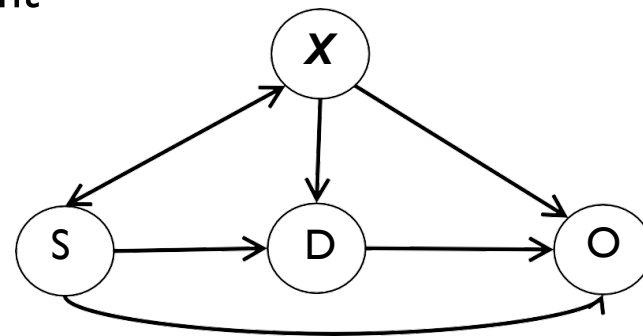
- Demographic Parity
 - a.k.a. Statistical Parity or Benchmarking
 - $\mathbb{P}(O=1 | S=1) = \mathbb{P}(O=1 | S=0)$

Same fraction of admitted men and women

S and O should be marginally independent

$$O \perp\!\!\!\perp S$$

Suppose it happens that one of the S has very high quality applications than the other, or applied to a highly competitive department



Group Fairness

- Conditional Statistical Parity

- For any $A=a$

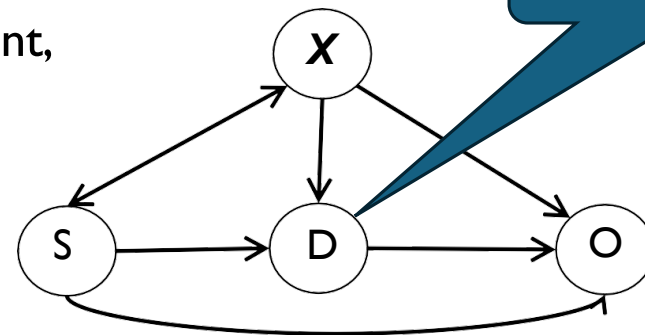
Admissible attributes

- $\mathbb{P}\{O=1 | S=1, A=a\} = \mathbb{P}\{O=1 | S=0, A=a\}$

Same fraction of admitted men and women, in each department

S and O should be marginally independent, conditioned on D

$$O \perp\!\!\!\perp S \mid D$$



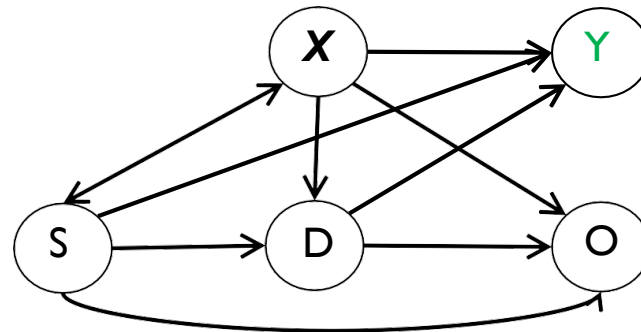
Group Fairness

Equal opportunity, true positive rate equality across protected groups

- True Positive Parity: $\mathbb{P}\{O=1 \mid S=1, Y=1\} = \mathbb{P}\{O=1 \mid S=0, Y=1\}$

$$O \perp\!\!\!\perp S \mid Y$$

Among those applicant who (do not) graduate, the rate of admitted students should be independent of applicants' gender.



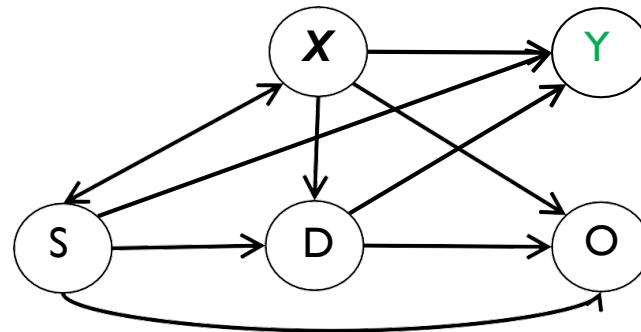
Group Fairness

Equalized odds, procedure accuracy (error rate) equality and disparate mistreatment

- False Negative Parity: $\mathbb{P}\{O=0|S=1,Y=1\}=\mathbb{P}\{O=0|S=0,Y=1\}$
- False Positive Parity: $\mathbb{P}\{O=1|S=1,Y=0\}=\mathbb{P}\{O=1|S=0,Y=0\}$

$$O \perp\!\!\!\perp S \mid Y$$

Among those applicant who (do not) graduate, the rate of admitted students should be independent of applicants' gender.



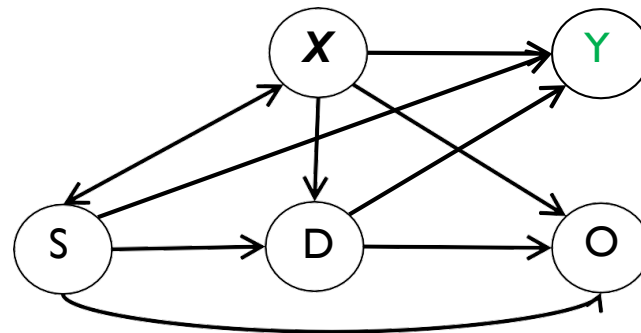
Group Fairness

Predictive Parity, Outcome Test or Test-fairness or Calibration

- The same predicted positive value (PPV)
 - $\mathbb{P}\{Y=1 \mid S=1, O=1\} = \mathbb{P}\{Y=1 \mid S=0, O=1\}$
 - $\mathbb{P}\{Y=1 \mid S=1, O=0\} = \mathbb{P}\{Y=1 \mid S=0, O=0\}$

$$Y \perp\!\!\!\perp S \mid O$$

Among those applicant that are admitted,
the rate of those who attain college degree
should be the same for men and women



Quick summary

Metric	Conditional on	Example Use-Case
Demographic Parity	None	Strictly equal outcomes, regardless of qualifications
Conditional Statistical Parity	Legitimate factors/ Admissible attributes	Allow legitimate differences
Equal Opportunity	True positive cases	Equal true positive rate across groups, , qualifications matter
Equalized Odds	True outcomes (labels)	Equal accuracy (error rates) across groups, , qualifications matter

Toolkits

- <https://github.com/fairlearn/fairlearn>



- <https://github.com/Trusted-AI/AIF360>



<https://github.com/tensorflow/fairness-indicators>



AIF360

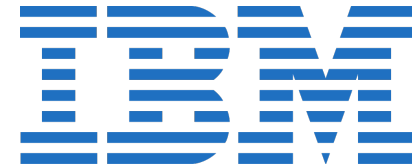
<https://github.com/Trusted-AI/AIF360>

Datasets Toolbox

- Fairness metrics (30+)
- Fairness metric explanations
- Bias mitigation algorithms (9+)

Guidance

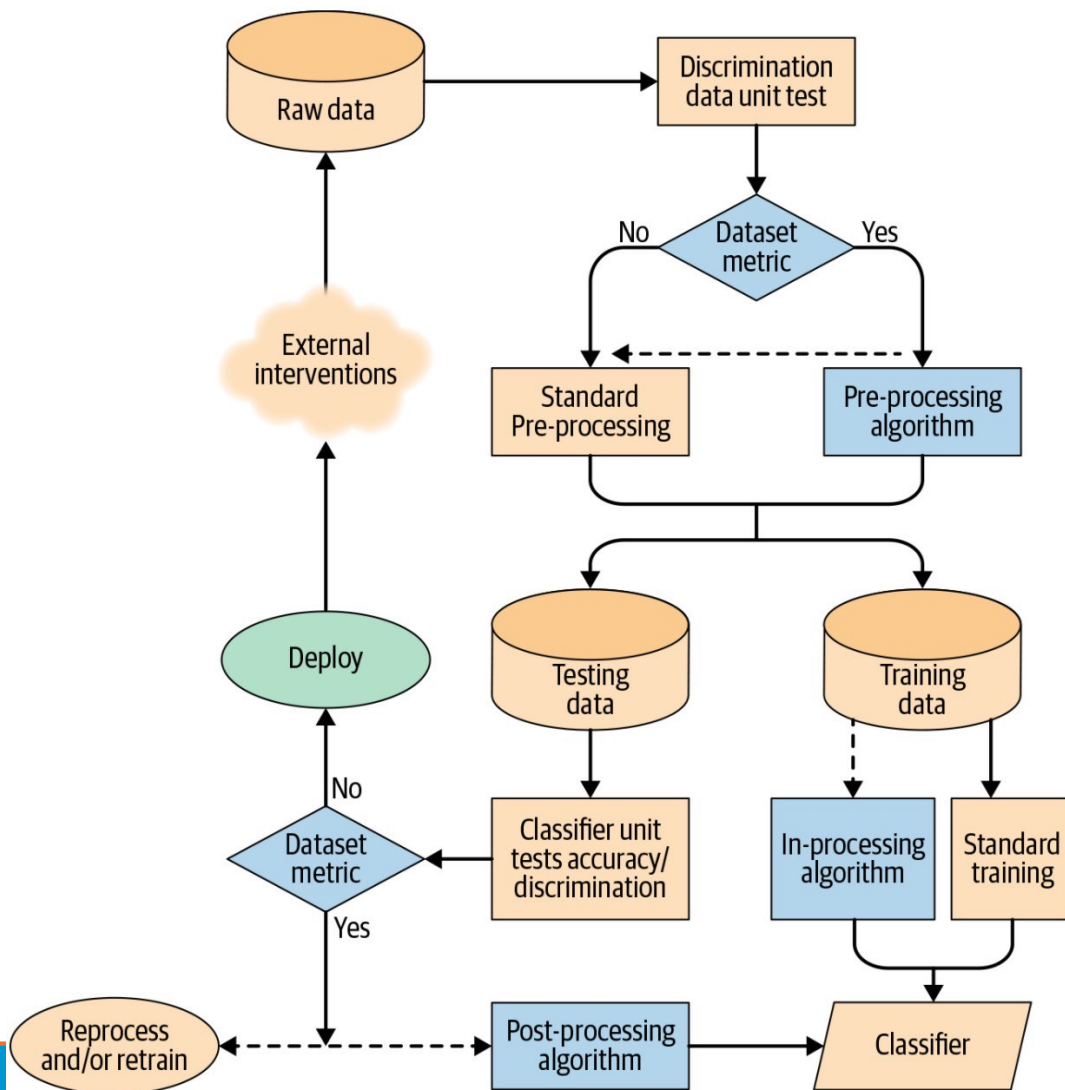
Industry-specific tutorials



Bias In the Machine Learning Pipeline

AI Fairness by Trisha Mahoney, Kush R. Varshney, and Michael Hind Copyright © 2020 O'Reilly Media. All rights reserved.

2025-03-25



AIF360 Algorithms

Pre-processing

- Reweighting
- Disparate Impact Remover
- Learning Fair Representations
- Optimized Preprocessing

In-processing

- Calibrated Equality of Odds
- Equality of Odds
- Reject Option Classification

Post-processing

- ART Classifier
- Prejudice Remover
- Post-processing

Reweighting

Modify the weights of different training examples such that

$P(\text{admit} \mid \text{Sex} = \text{'Female'})$

$=$

$P(\text{admit} \mid \text{Sex} = \text{'Male'})$

Sex	Ethnicity	Highest degree	Job type	Class
M	Native	H. school	Board	+
M	Native	Univ.	Board	+
M	Native	H. school	Board	+
M	Non-nat.	H. school	Healthcare	+
M	Non-nat.	Univ.	Healthcare	—
F	Non-nat.	Univ.	Education	—
F	Native	H. school	Education	—
F	Native	None	Healthcare	+
F	Non-nat.	Univ.	Education	—
F	Native	H. school	Board	+

Reweighting

Algorithm 3: Reweighting

Input: $(D, S, Class)$

Output: Classifier learned on reweighed D

```
1: for  $s \in \{F, M\}$  do  
2:   for  $c \in \{-, +\}$  do  
3:     Let  $W(s, c) := \frac{|\{X \in D \mid X(S) = s\}| \times |\{X \in D \mid X(Class) = c\}|}{|D| \times |\{X \in D \mid X(Class) = c \text{ and } X(S) = s\}|}$   
4:   end for  
5: end for  
6:  $D_W := \{\}$   
7: for  $X$  in  $D$  do  
8:   Add  $(X, W(X(S), X(Class)))$  to  $D_W$   
9: end for  
10: Train a classifier  $C$  on training set  $D_W$ , taking into account the weights  
11: return Classifier  $C$ 
```

EKamiran and T.Calders, "Data Preprocessing Techniques for Classification without Discrimination," Knowledge and Information Systems, 2012
(<https://link.springer.com/content/pdf/10.1007%2Fs10115-011-0463-8.pdf>)

Reweighting - Example

Sex	Ethnicity	Highest degree	Job type	Cl.	Weight
M	Native	H. school	Board	+	0.75
M	Native	Univ.	Board	+	0.75
M	Native	H. school	Board	+	0.75
M	Non-nat.	H. school	Healthcare	+	0.75
M	Non-nat.	Univ.	Healthcare	—	2
F	Non-nat.	Univ.	Education	—	0.67
F	Native	H. school	Education	—	0.67
F	Native	None	Healthcare	+	1.5
F	Non-nat.	Univ.	Education	—	0.67
F	Native	H. school	Board	+	1.5

$$\frac{5 \times 6}{10 \times 4} = 0.75$$

$$\frac{5 \times 4}{10 \times 1} = 2$$

$$\frac{5 \times 4}{10 \times 3} = 0.67$$

$$\frac{5 \times 6}{10 \times 2} = 1.5$$

AIF360 Example

```
## import dataset
privileged_groups = [{'sex': 1}]
unprivileged_groups = [{'sex': 0}]
dataset_orig = load_preproc_data_adult(['sex'])

all_metrics = ["Statistical parity difference",
               "Average odds difference",
               "Equal opportunity difference"]
```

```
# Get the dataset and split into train and test
dataset_orig_train, dataset_orig_vt = dataset_orig.split([0.7], shuffle=True)
dataset_orig_valid, dataset_orig_test = dataset_orig_vt.split([0.5], shuffle=True)
```

✓ 0.0s

Difference in mean outcomes between unprivileged and privileged groups = -0.190244

✓ 0.0s

46

```

from aif360.algorithms.preprocessing.reweighing import Reweighing

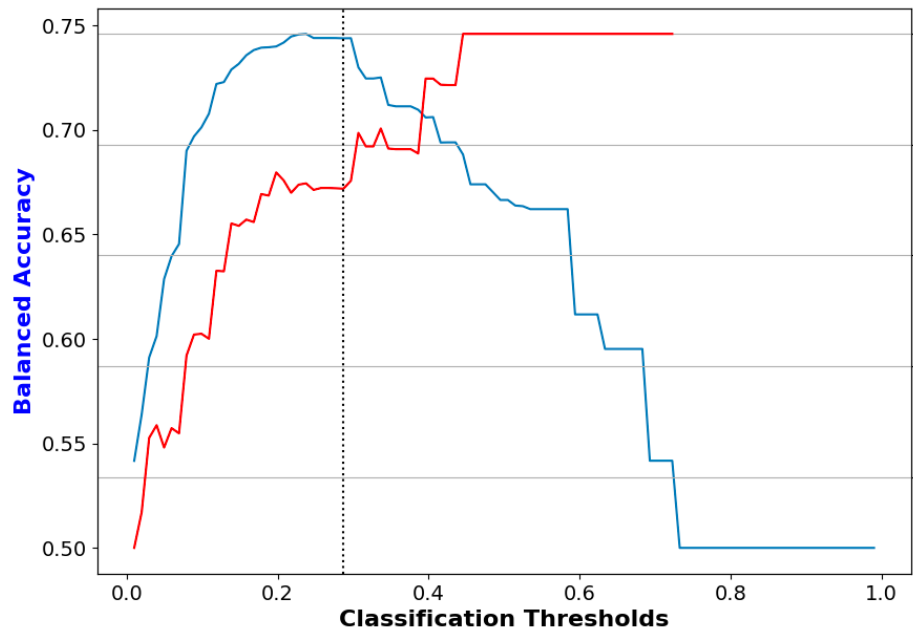
RW = Reweighing(unprivileged_groups=unprivileged_groups,
|         |         |         | privileged_groups=privileged_groups)
RW.fit(dataset_orig_train)
dataset_transf_train = RW.transform(dataset_orig_train)

```

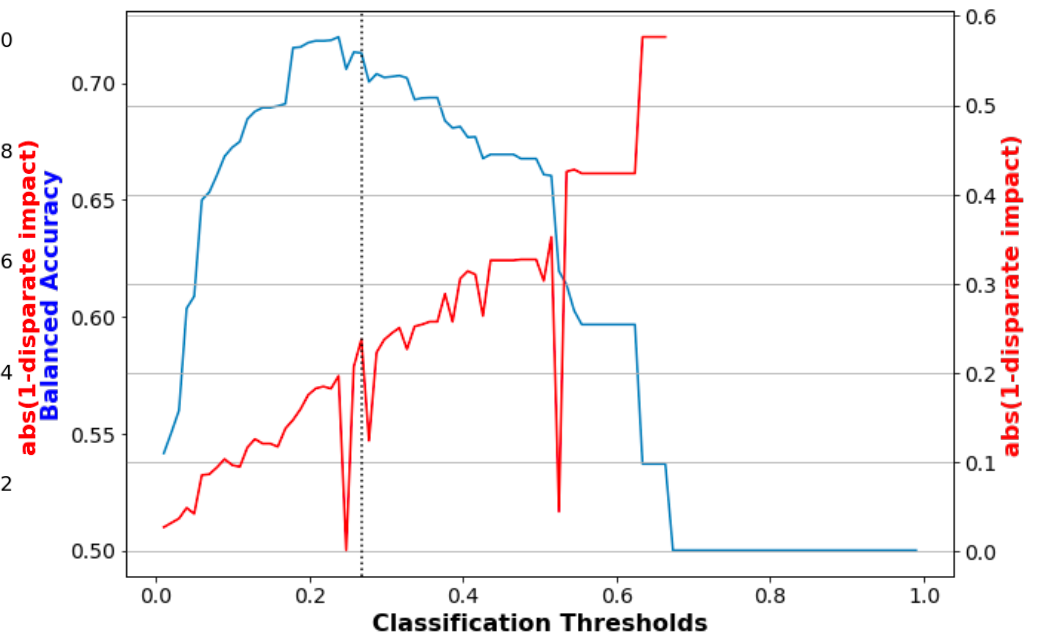
dataset_transf_train

✓ 0.0s

instance weights features					labels	
		protected attribute				
		race	sex	Age		
instance names					instance names	
391	1.090119	0.0	1.0		391	0.0
1899	1.090119	0.0	1.0		1899	0.0
24506	0.790634	1.0	1.0		24506	1.0
32816	1.090119	1.0	1.0		32816	0.0
47892	1.090119	1.0	1.0		47892	0.0



classifier trained with original training data



classifier trained with reweighted training data

```
from aif360.explainers import MetricTextExplainer
text_expl_orig = MetricTextExplainer(metric_orig_train)
text_expl_tran = MetricTextExplainer(metric_transf_train)
print(text_expl_orig.disparate_impact())
print(text_expl_tran.disparate_impact())
```

✓ 0.0s

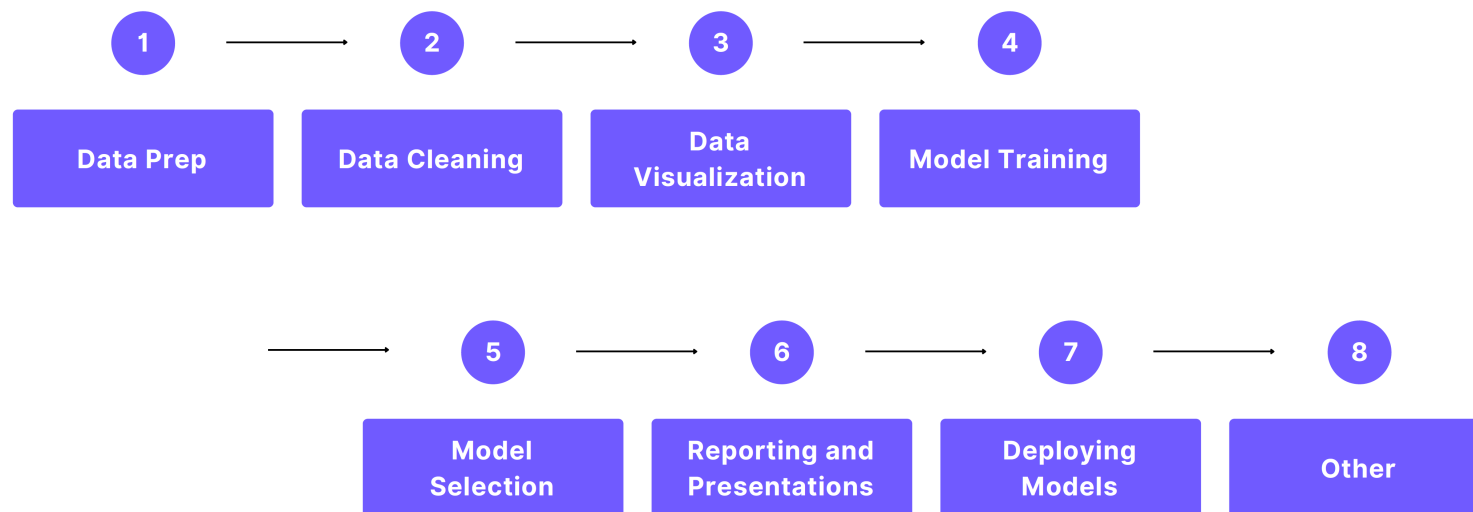
Disparate impact (probability of favorable outcome for unprivileged instances \square probability of favorable outcome for privileged instances): 0.3677778370794947

Disparate impact (probability of favorable outcome for unprivileged instances \angle probability of favorable outcome for privileged instances): 1.0

Responsible Data Management

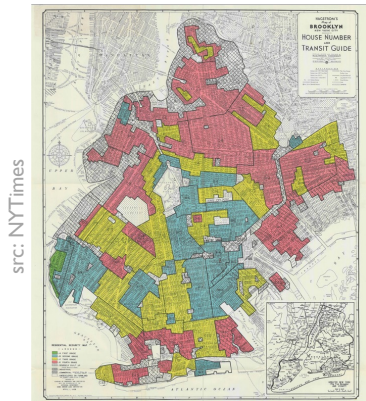
Recap: Data Science Pipeline

Thinking about your current role, what tasks are most time consuming? (Responses ranked from most to least time consuming)



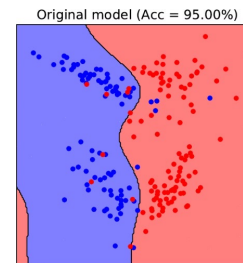
n=1,071

What are the sources of bias?



Historical bias in training data

src: openai.com

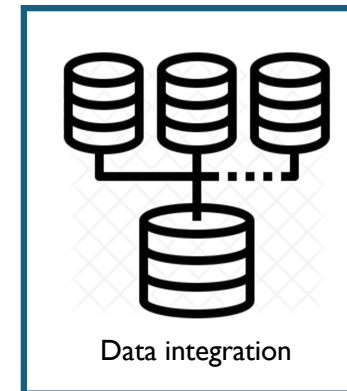
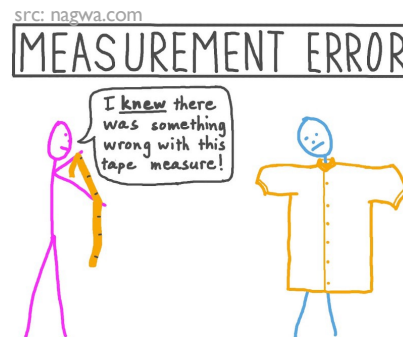


src: <https://labs.f-secure.com>

Adversarial data attacks



Selection bias



src: nagwa.com



Model design choices

Hooker, Sara. "Moving beyond "algorithmic bias is a data problem"." Patterns 2.4 (2021): 100241.

AI Needs Good Data

- ML depends on data and data-driven algorithms are only as good as the data they work with
- Data usually obtained through multiple sources.
 - Goes through significant process of cleaning and integration
- Responsible AI requires in the Responsible data pipeline

Data Cleaning

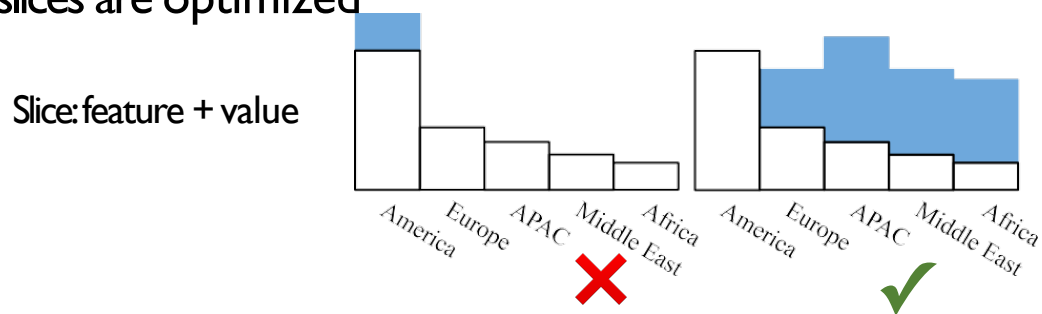
- Always important, even more critical for responsible AI
 - Incomplete and incorrect data typically hurt minorities, further increasing the data bias in such cases.
- Example
 - Two groups (minority and majority); a small portion belong to the minority
 - A simple task: compute average
 - An incorrect majority value does not significantly impact the average
 - An incorrect minority value may significantly skew the average

Missing values resolution issue

- Downstream approaches to resolve missing (and unclear) values can further increase bias in data
- Example (two resolution strategies)
 1. rows with missing values are removed
 - removing a minority row further decreases the data coverage
 2. missing values are replaced with the column average
 - the average value is mostly affected by majorities
 - bias further increased

Data Acquisition

- Slice Discovery
 - Identifying problematic slices of data that cause bias
 - Selectively acquiring the right amount of data for problematic slices
 - possibly different amounts of data per slice s.t. accuracy and fairness on all slices are optimized



Conclusion

Big Picture

- Why responsible data science?
- Data science ethics

Fairness

- Fairness measures in ML
- AIF360

Reweighting

Responsible Data Management



or



Sources

- FairML book, chapter 4: <https://fairmlbook.org/>
- Slides by Jiannan Wang in CMPT 733
- Slides by Sudeepa Roy in CompSci 590 Spring'23, Duke University
- Fatemeh Nargesian, Abolfazl Asudeh, and H. V. Jagadish, “Responsible Data Integration: Next-generation Challenges.” SIGMOD'22 Tutorial

CMPT 733

Privacy Enhancing Technologies

Instructor

Zhengjie Miao

Course website

<https://coursys.sfu.ca/2025sp-cmpt-733-gl/pages/>

Slides by

Ricardo Silva Carvalho | SFU's Big Data Hub

Topics Today

- Overview of privacy preserving technologies
- Previous attempts at privacy and possible attacks
- Understand the goal and applicability of commonly used privacy tools.



Image by [Engin Akyurt](#)

WHAT DO WE MEAN BY PRIVACY?

PROTECT PERSONAL DATA

- According to GDPR, "personal data" means:

Any information relating to an "identified or identifiable natural person ('data subject')", which is:

- One who can be identified, directly or indirectly, in particular by reference to:
 - an identifier such as a name, an identification number, location data, an online identifier or to
 - one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.



Image by [Angela Roma](#)

PROTECT PERSONAL DATA

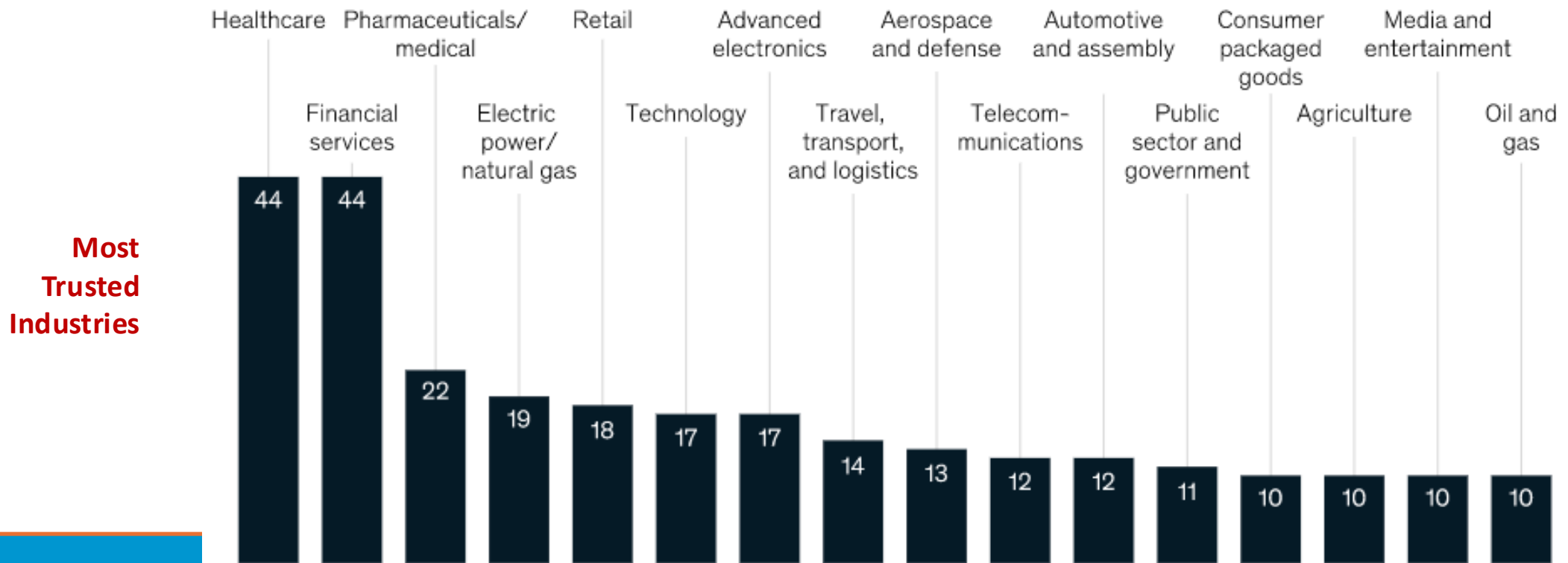
- Goal: Reduce chances of identification
- Sensitive data (and subject)
 - Health records
 - Search logs
 - Location data
 - Private conversation
- Disclosure can be harmful
- Data = leverage



Image by [Travis Saylor](#)

PRIVACY AWARENESS

- 87% would not do business with a company if they had concerns about its security practices. Source: [McKinsey's Survey, North America, 2020](#)



PRIVACY IS NOT JUST ABOUT SENSITIVE DATA

- Depends on the parties involved
 - Appropriate consent
- How the data will be shared?
- Examples
 - Our medical data
 - Facial recognition
 - Location tracking
- Data subjects in control of their data

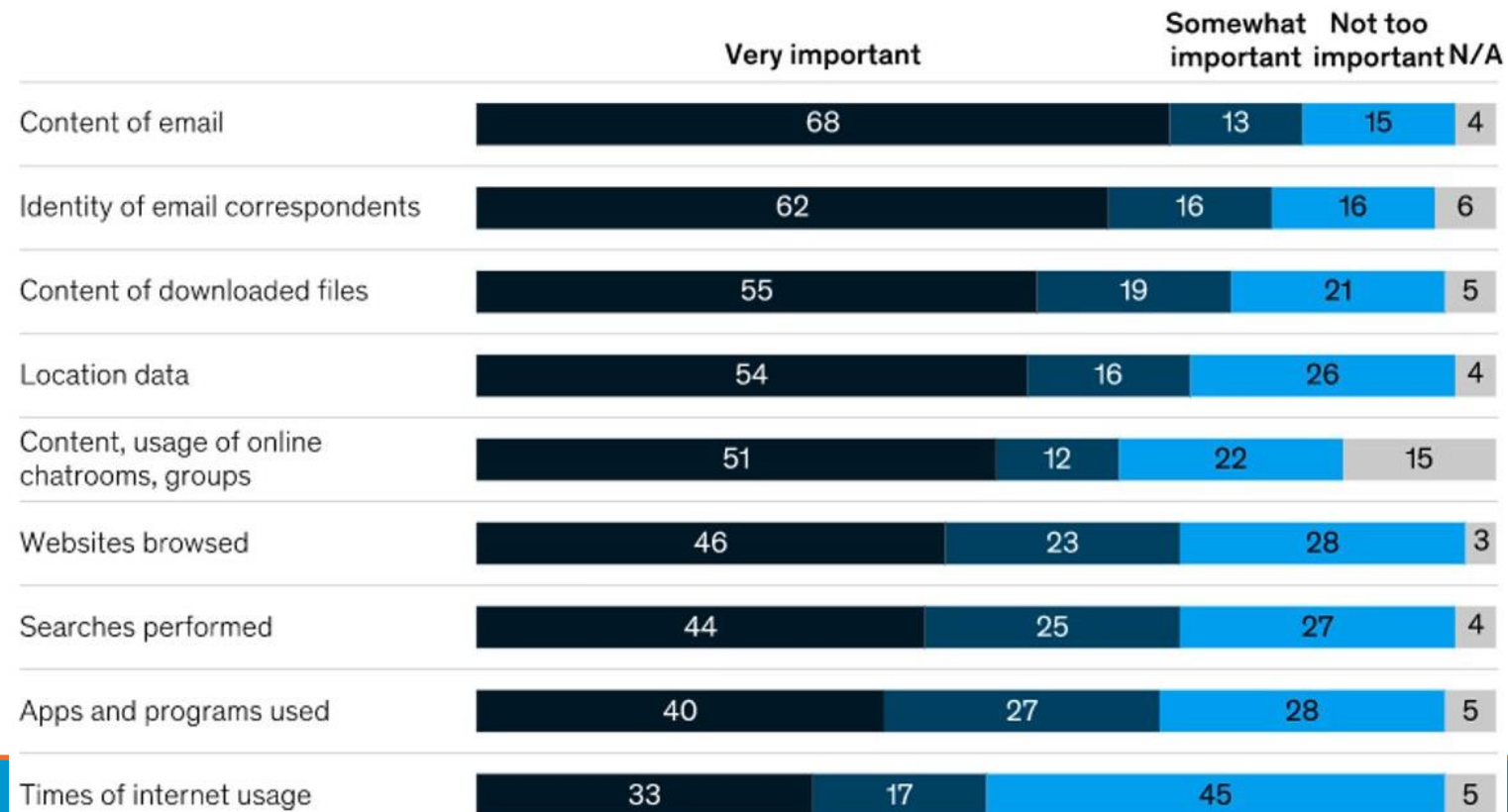


Image by [Pixabay](#)

PRIVACY AWARENESS

- 87% would not do business with a company if they had concerns about its security practices. Source: [McKinsey's Survey, North America, 2020](#)

Importance
by type of
digital data



SHARING SENSITIVE DATA CAN BE BENEFITIAL

- Academic research
- Policy making
- Searching for terrorists
- Drug trials
- Market research
- Large-scale crisis

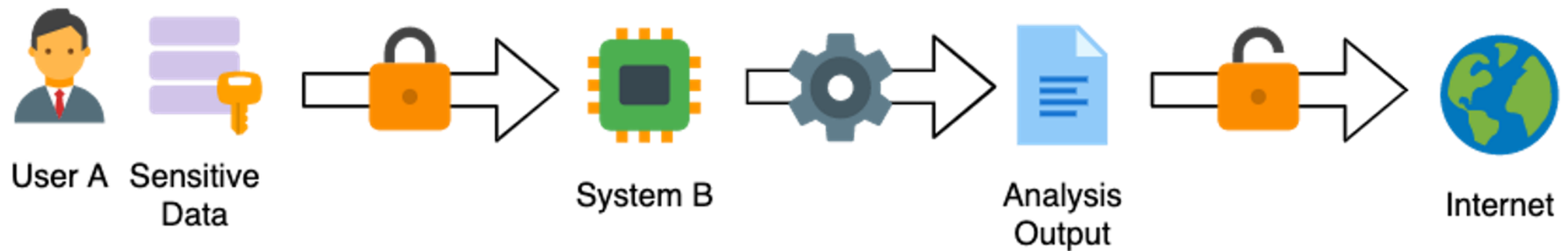


Image by [Fauxels](#)

HOW CAN WE ENABLE THE USE
OF SENSITIVE DATA,
WHILE PROTECTING THE PRIVACY
OF THE DATA SUBJECTS?

PRIVACY IS DIFFERENT FROM SECURITY

- Limit knowledge vs Limit access



PRIVACY: INPUT vs OUTPUT

- Input
 - Trusted Curator
 - Secure Enclaves
 - Encryption
- Output
 - Anonymization
 - Differential Privacy
 - Synthetic Data



Image by [Oleksandr Pidvalnyi](#)

PRIVACY ENHANCING TECHNOLOGIES

- Anonymization
- Differential Privacy
- Synthetic Data
- Homomorphic Encryption
- Secure Multi-Party Computation
- Federated Learning



Image by [Pexels](#)

ANONYMIZATION

ANONYMIZED DATA

- General Data Protection Regulation (GDPR) defines "anonymized data":

“information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.”



Image by [Christian Gonzalez](#)

ANONYMIZED DATA

- [GDPR WP29](#)
- Data is anonymized when three things are impossible
 - the “singling out” of an individual,
 - the linking of data points of an individual to create a larger profile (“linkability”)
 - and the ability to deduce one attribute from another attribute (“inference”).



Image by [Christian Gonzalez](#)

ANONYMIZATION

- From IAPP's Guide:
 - Anonymization techniques basically reduce the "identifiability" of one or more individuals from the original dataset to a level acceptable by the organization's risk portfolio
- Goal:
 - Reduce chances of identification
 - Personable Identifiable Information (PII)



Image by [Christian Gonzalez](#)

ANONYMIZATION

- IAPP Glossary of Privacy Terms:
 - "The process in which individually identifiable data is altered in such a way that it "no longer can be"* related back to a given individual"
- *has a negligible chance to be



Image by [Christian Gonzalez](#)

Personal data and Anonymization

1. Direct identifiers
 - Name, Passport number
 2. Indirect or Quasi-identifiers
 - Gender, zip code, birthdate
 3. Sensitive identifiers
 - Diagnosis, Browser log
- Truly anonymized data is no longer subject to GDPR



Image by [Angela Roma](#)

How to Anonymize personal data?

- In general, involve complex analysis
- How to assess "identifiability"?
 - Requires subject-matter experts
 - Example: Medical data usually requires someone with sufficient healthcare knowledge to assess how unique (i.e., how identifiable) a record is



Image by [Pexels](#)

How to Anonymize personal data?

- Techniques have specific purpose
 - Usually, we combine multiple techniques
- Hard to assess risk of disclosure
- Typical trade-off between:
 - Data quality
 - Level of de-identification



Image by [Pexels](#)

SOME ANONYMIZATION APPROACHES

- Pseudonymization
- Suppression
- Masking
- Generalization
- Swapping
- Perturbation
- Aggregation
- K-anonymity



Image by [Miguel Padriñán](#)

Pseudonymization

- Replacing identifying data with pseudonyms
- Use-case: When original values are securely kept but can be retrieved and linked back to the pseudonym
- Still is "personal data" according to GDPR

Person	Age	Gender
Charlie	29	F
Bob	34	M
Alice	55	F



Person	Age	Gender
24572	29	F
84625	34	M
45342	55	F




Identity Table

Pseudonym	Person
24572	Charlie
84625	Bob
45342	Alice

Suppression

- Attribute suppression
 - Use-case: When attribute cannot be suitably anonymized
 - A "derived" attribute may be a better option
- Record suppression
 - Use-case: When the row is an outlier

Student	Teacher	Score
Alice	Rachel	88
Bob	Rachel	92
Charlie	John	89
Donald	John	79



Teacher	Score
Rachel	88
Rachel	92
John	89
John	79

Masking

- Changing characters by a constant symbol
- Use-case: When hiding part of a string is "enough"

Zip code	Order Price	Quantity
993831	\$1040	4
880012	\$509	2
770344	\$839	3



Zip code	Order Price	Quantity
99xxxx	\$1040	4
88xxxx	\$509	2
77xxxx	\$839	3

Generalization

- Reduce precision: create larger categories, ranges
- Use-case: Generalized values can still be useful

Person	Age	Address
383745	24	369 East Street
827459	45	1047 Pinetree Road
925870	30	770 Tampa Avenue
498544	37	291 Lloyd Street
147402	64	107 Stone Road



Person	Age	Address
383745	21-30	East Street
827459	41-50	Pinetree Road
925870	21-30	Tampa Avenue
498544	31-40	Lloyd Street
147402	>60	Stone Road

Swapping

- Rearranging attribute data
- Use-cases: When there is no need for analysis of relationships between attributes at the record level.

Job	Date of Birth	# Orders
Professor	20 Mar 1990	2
Salesman	10 May 1978	3
Nurse	22 Feb 1994	8
Lawyer	17 May 1985	5
Programmer	13 Dec 1982	1



Job	Date of Birth	# Orders
Salesman	13 Dec 1982	5
Nurse	17 May 1985	8
Lawyer	20 Mar 1990	3
Programmer	10 May 1978	1
Professor	22 Feb 1994	2

Perturbation

- Slightly modifying values, e.g., rounding or adding noise.
 - Base-x: rounding to the nearest multiple of x
- Use-case: When small changes are acceptable
- Example: base-5,3,3

Person	Height (cm)	Weight (kg)	Age
987352	161	50	30
292944	177	70	36
862833	158	46	20
134973	173	75	22
738937	169	82	44



Person	Height (cm)	Weight (kg)	Age
987352	160	51	30
292944	175	69	36
862833	160	45	21
134973	175	75	21
738937	170	81	42

Aggregation

- Summarize values
- Use-case: When aggregated data fulfills the purpose

Person	Income	Donation
854865	\$4000	\$200
376972	\$6000	\$300
198309	\$2000	\$100
736392	\$5000	\$300
282763	\$3000	\$300
743639	\$5000	\$700
937354	\$1000	\$100



Income (\$)	Nr. of donations	Sum of Donations (\$)
1000 - 2999	2	200
3000 - 4999	2	500
5000 - 6999	3	1300
TOTAL	7	2600

K-anonymity

- K-anonymity is a property of a dataset
 - A dataset is k-anonymous if quasi-identifiers for each person in the dataset are identical to at least k – 1 other people also in the dataset.
 - We compute the k-anonymity value based on one or more columns, or fields, of a dataset.

Zip code	Age
997356	34
990023	35
334863	77
330121	78



2-anonymous

Zip code	Age
990023	34
990023	34
330121	78
330121	78

K-anonymity

Age	Gender	Job	Orders
25	F	Lawyer	3
32	M	Salesman	8
20	F	Banker	2
49	F	Web Developer	11
21	F	Legal Assistant	9
34	M	Salesman	13
49	F	Programmer	5
27	F	Legal Assistant	3
33	F	Lawyer	8



Age	Gender	Job	Orders
21-30	F	Lawyer	3
31-40	M	Salesman	8
21-30	F	Banker	2
41-50	F	IT	11
21-30	F	Legal Assistant	9
31-40	M	Salesman	13
41-50	F	IT	5
21-30	F	Legal Assistant	3
21-30	F	Lawyer	8

"Orders" was considered as a non-identifier, without a need to further anonymize this attribute.

K-anonymity

- Issues:
 - The value of k is not indicative of protection level
 - There is no formal indication of how to choose k
- To choose k :
 - Understand risk of privacy incidents
 - Try out typical values (e.g., 5 to 15)
- K-anonymity is hard but still used, especially in healthcare

POSSIBLE ATTACKS

Linkage attacks

- Use auxiliary information (side knowledge) to re-identify individuals
- Example:

Name	Zip Code	Age	Gender	Salary
—	64***	31-40	M	60k
—	67***	41-50	M	70k
—	64***	41-50	F	80k
—	67***	31-40	F	50k
—	62***	21-30	M	40k

- Suppose you know a friend with:
 - Zip Code: 64152, Gender: F

Linkage attacks

- [[Sweeney, 2002](#)] reports that, from the 1990 U.S. Census they observed:
 - 87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on:
 - 5-digit Zip Code
 - Gender
 - Date of birth



Image by [San Fermin Pamplona](#)

Differencing Attacks

- Comparing two data points
 1. Group and subgroup
 - Total purchased per store/day
 - Total purchased per loyal program/store/day
 - Only you and another person used loyal program X in day Y
 2. Times t and $t+1$
 - Average salary of employees in 2020
 - Average salary of employees in 2021
 - Only you and another individual were hired



Image by [Markus Spiske](#)

Reconstruction Attacks

1. Define constraints
2. Look for valid values

Example **for 2B**:

- Ages A, B, C, e.g. $A \leq B \leq C$
- $B=30$
- $1 \leq A \leq B \leq C \leq 125$
- $(A+B+C)/3 = 44$

These constraints already leave us with only 30 possibilities of (A,B,C)

TABLE 1: **FICTIONAL STATISTICAL DATA FOR A FICTIONAL BLOCK**

STATISTIC	GROUP	AGE		
		COUNT	MEDIAN	MEAN
1A	total population	7	30	38
2A	female	4	30	33.5
2B	male	3	30	44
2C	black or African American	4	51	48.5
2D	white	3	24	24
3A	single adults	[D]	[D]	[D]
3B	married adults	4	51	54
4A	black or African American female	3	36	36.7
4B	black or African American male	[D]	[D]	[D]
4C	white male	[D]	[D]	[D]
4D	white female	[D]	[D]	[D]
5A	persons under 5 years	[D]	[D]	[D]
5B	persons under 18 years	[D]	[D]	[D]
5C	persons 64 years or over	[D]	[D]	[D]

Note: Married persons must be 15 or over

Database Reconstruction

- Seminal work: [[Dinur and Nissim, 2003](#)]
- [[Dwork and Roth, 2014](#)]:
 - "Fundamental Law of Information Recovery"
 - Giving overly accurate answers to too many questions will inevitably destroy privacy.
 - Overly accurate estimates of too many statistics will divulge the entire database, no matter how one attempts to blunt the attack by introducing inaccuracies.



Image by [Seven Storm](#)

ATTEMPTS AT PRIVACY

The Netflix Prize dataset

- Netflix Prize:
 - 10% of users
 - Average of 200 ratings/user
- Example of result:
 - An attacker who knows the subscriber's ratings on 2 movies and the dates has a 64% chance to completely identify the subscriber.
 - Goes to 80+% for unpopular movies.

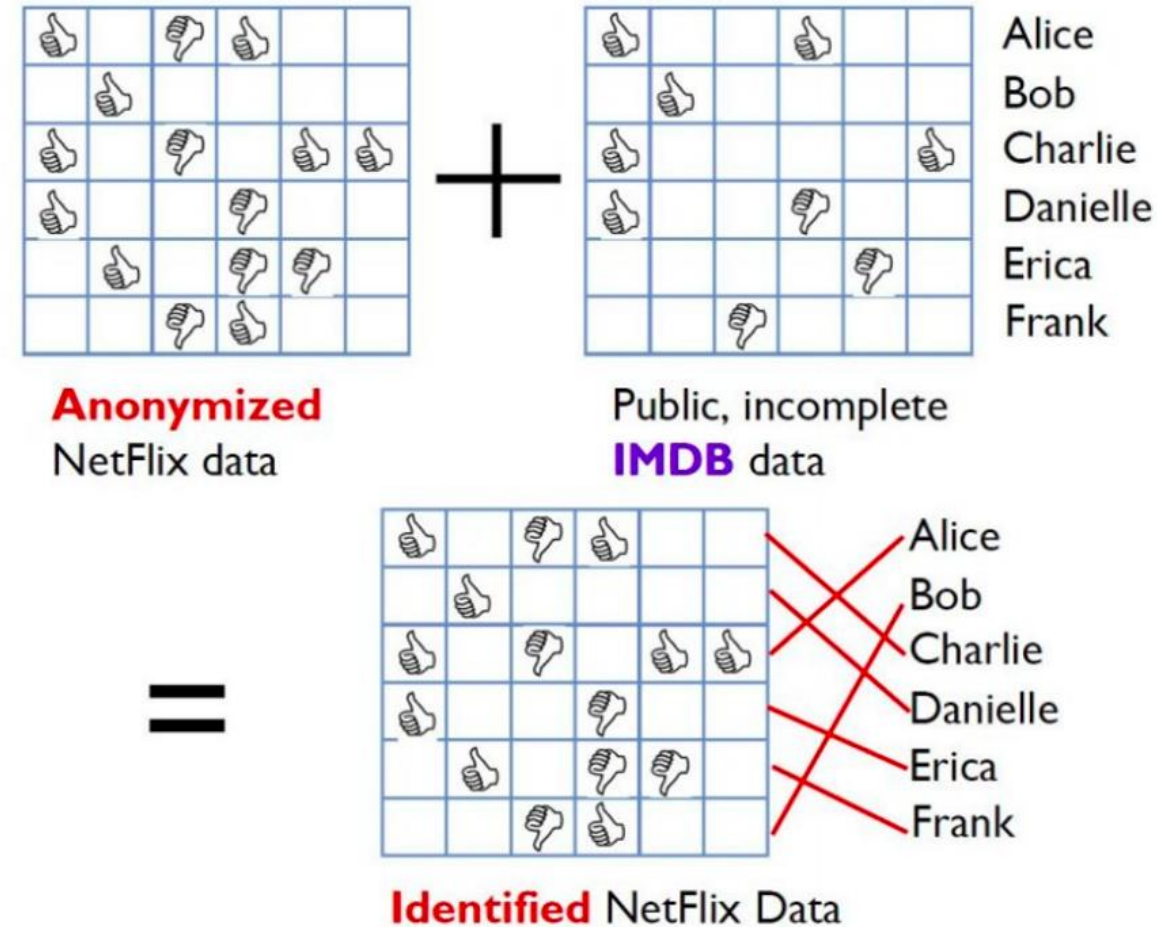


Image by: "How to break anonymity of the Netflix Prize

dataset", A. Narayanan, V. Shmatikov, 2008.

Massachusetts Group Insurance Commission

- Anonymized medical history of patients (all hospital visits, diagnosis, prescriptions)
- Latanya Sweeney
 - MIT Grad Student
 - Purchased Cambridge voter roll for \$20
 - Identified the medical information of William Weld, former governor of Massachusetts

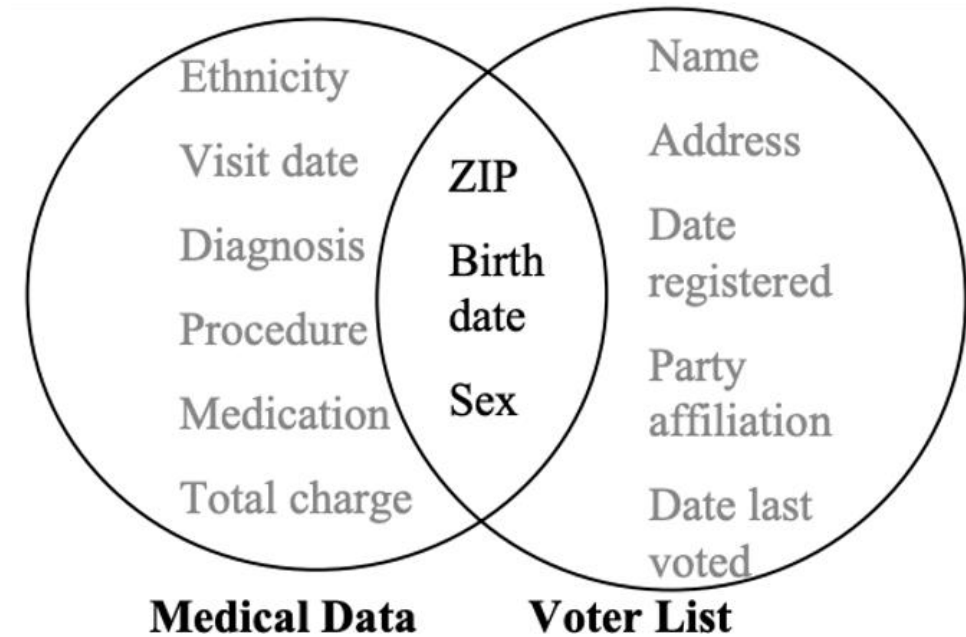


Image by: “[Matching known patients to health records in Washington State Data](#)”, L. Sweeney, 2013.

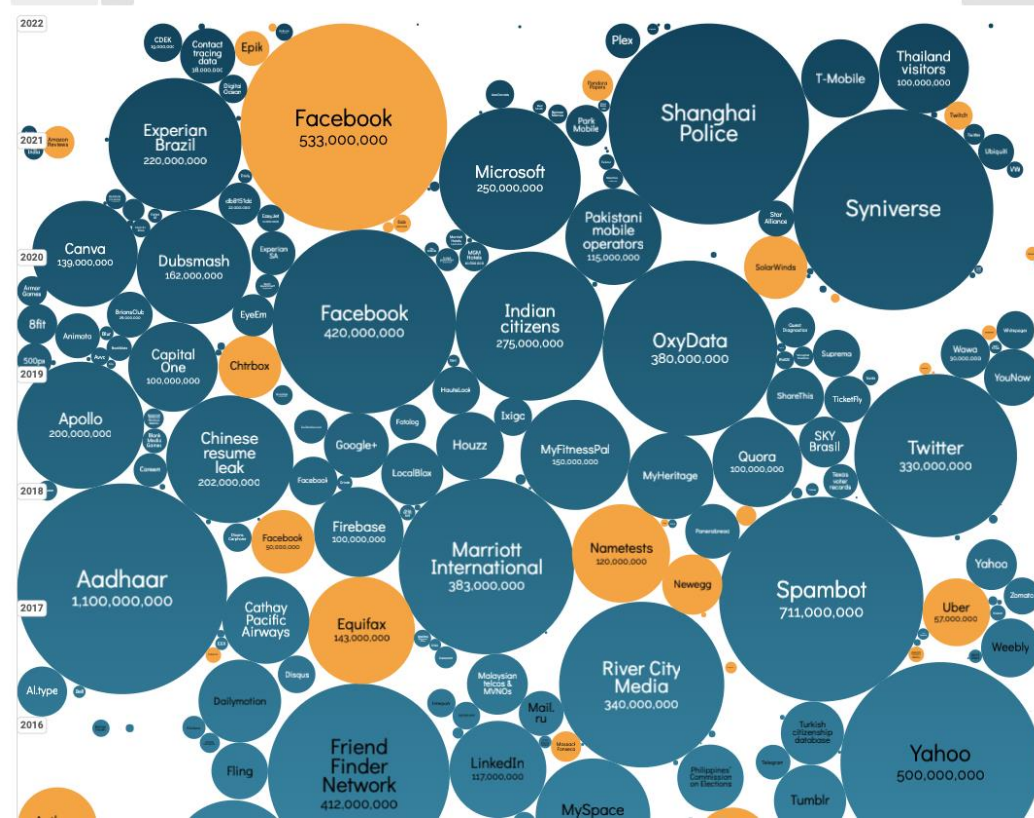
World's Biggest Data Breaches & Hacks

World's Biggest Data Breaches & Hacks

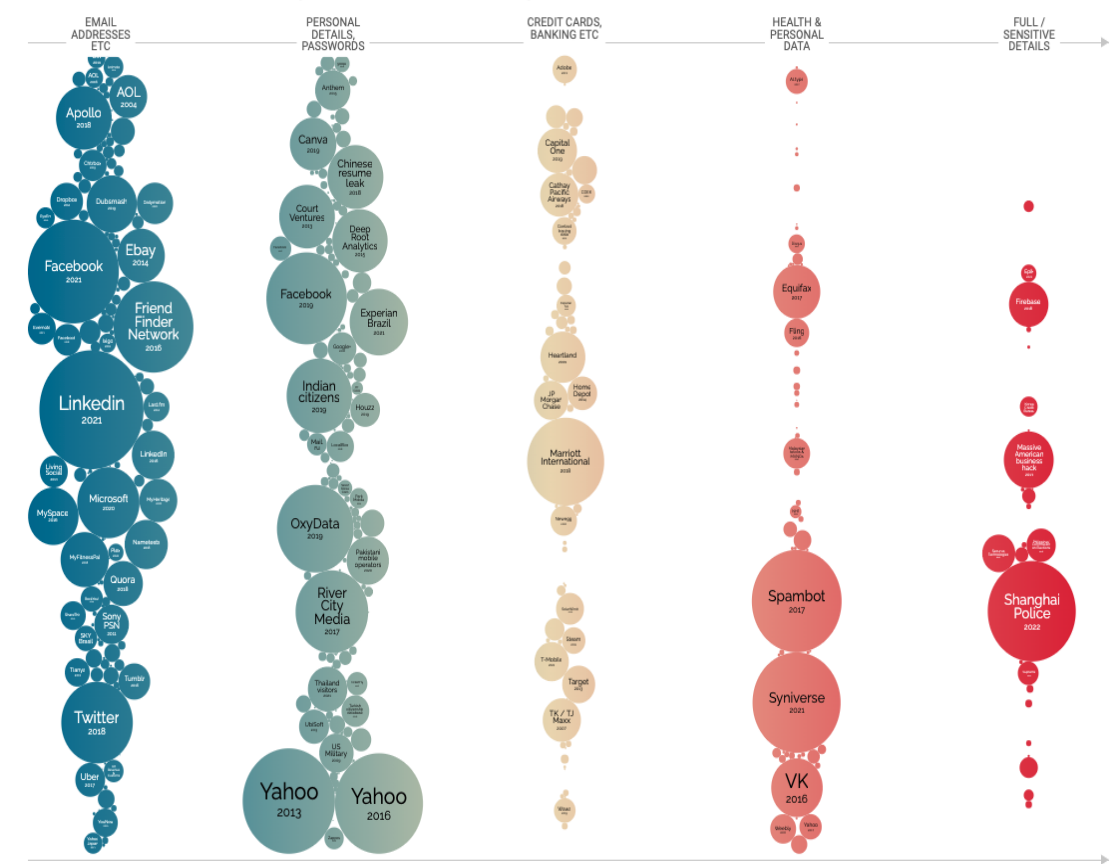
Selected events over 30,000 records

UPDATED: Sep 2022

size: records lost filter



Data Breaches by data sensitivity



<https://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/>

**SO ANONYMIZATION DOES NOT
WORK?**

Does Anonymization work?

- Anonymization can work if done properly
- Many previous fiascos had datasets mislabeled as anonymous
 - Mostly because of existing quasi-identifiers
 - Only removing direct identifiers is not enough!
- A systematic review of attacks of health data shows:
 - Only 2 out of 14 attacks were on datasets properly anonymized, with one of them having re-identification only of 2 out of 15,000.

Using Anonymization in practice

- We must be careful
 - High dimensional data is very challenging
- Subject-matter expert is essential
 - Double-check to remove quasi-identifiers
 - Usually, it will be tailored to one purpose
- Data does not live in isolation
 - What are other possible external datasets?



Image by [Artem Podrez](#)

Anonymization is hard and may not be enough

- The anonymization necessary may destroy utility
- High-dimensional data is essentially unique
- Privacy needs to be dealt very seriously



Image by [Pexels](#)

We need **FORMAL** privacy guarantees

- Anonymization techniques depend on the dataset
- What happens when the dataset we anonymized is updated?
- It's hard to define every nuance in a dataset to guarantee privacy



Image by [Pixabay](#)

HOW TO WRITE A FORMAL DEFINITION OF PRIVACY?

Formal Privacy definition

- What are we looking for?
- Ideal scenario:
 - If the **output of an algorithm** on a dataset containing my data **does not change** if I **remove** my data from that dataset then my privacy is fully protected.



Image by [Lum3n](#)

Formal Privacy definition

- Can we construct a useful algorithm which does not change a given output no matter who we remove from the dataset?



Image by E
katerina
ulovtsova

Formal Privacy definition

- Can we construct a useful algorithm which does not change a given output no matter who we remove from the dataset?

No!

- What can we do instead?
 - Offer a knob to tune Privacy vs Utility (accuracy)
 - Plausible deniability



INTRODUCTION TO DIFFERENTIAL PRIVACY - DP

Differential Privacy – What do we want?

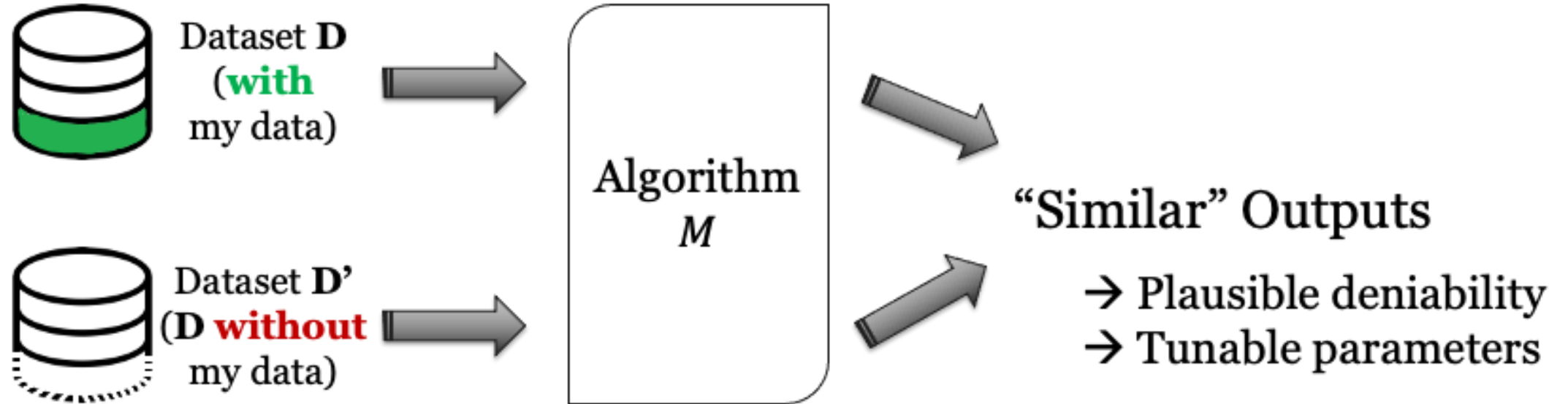
- Quote from [[Dwork and Roth, 2014](#)]:

Differential Privacy describes a promise,
made by a **data holder**, or curator,
to a **data subject**:

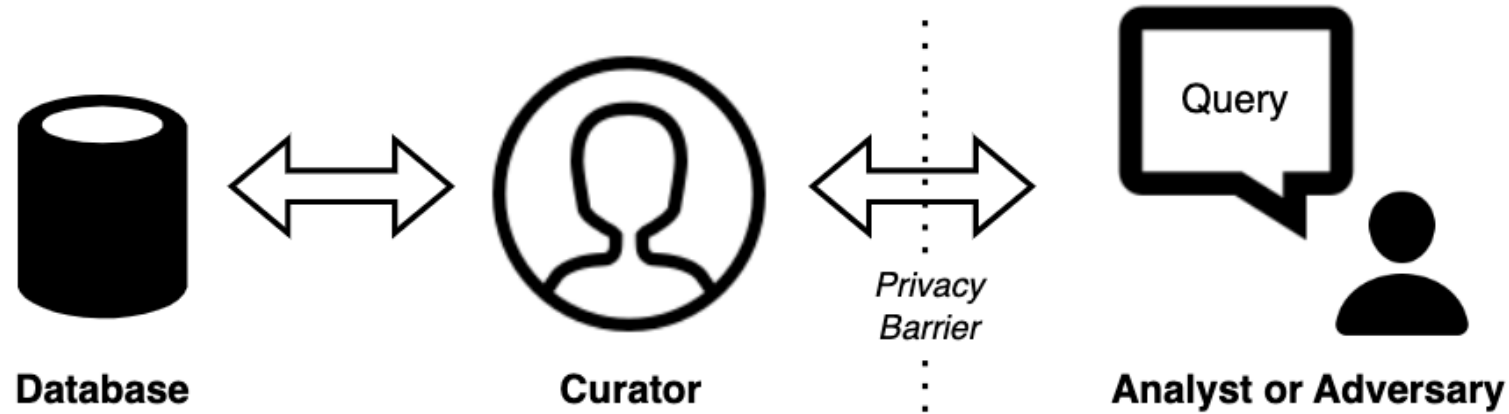
“You will **not** be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, **no matter what other studies, data sets, or information sources, are available.**”

Differential Privacy – Intuition

- If algorithm M is differentially private, then
for any individual data (e.g., my data) in any dataset D



Differential Privacy – In practice



- Too many accurate answers lead to reconstruction of data
- We will "add noise" to avoid that
 - How to set the noise?
 - ϵ -differential privacy: $\Pr[M(D) = o] \leq e^\epsilon \Pr[M(D') = o]$

Is DP the best choice for my problem?

Is DP the right tool for my problem?

- Designed for analyses that do not heavily depend on individual data
 - Is just one person likely to change the result?
- Analysis' results should be about the same if small changes in the data occur
- Examples
 - How aggregated do the results need to be?
 - Are you interested in outliers?



Image by [Pixabay](#)

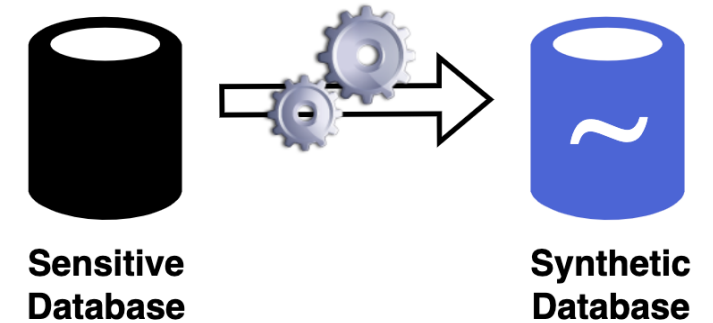
DP in summary

- [[Dwork and Roth, 2014](#)]:
- "Differential Privacy addresses the paradox of learning nothing about an individual while learning useful information about a population. It is a definition, not an algorithm."

SYNTHETIC DATA

Synthetic Data

- Create data that resembles the sensitive data while maintaining privacy
- Only useful if keeps similar utility to original data
 - What is the purpose?
- Synthetic data by default is **not** privacy preserving
 - Example: Membership Inference Attacks [[Shokri et. al, 2017](#)]
- To guarantee privacy, Differential Privacy can be used



CONTENTS

Introduction to PETS

Examples of:

- Anonymization
- K-anonymity
- Differential Privacy