

CMPT 733 – Big Data Programming II

Responsible Data Science

Instructor Steven Bergner

Course website <https://coursys.sfu.ca/2024sp-cmpt-733-g1/pages>

Slides by Jiannan Wang

Data scientists have a lot of power

A lot of data

A lot of data-driven
decisions

A lot of ML/Stats
methods

Whether Tom can get admitted by a university

Whether Tom can get an offer from a company

Whether Tom can get a loan from a bank

Whether Tom can express his option on a website

Whether Tom can be treated properly in a hospital

...

What is a right decision?

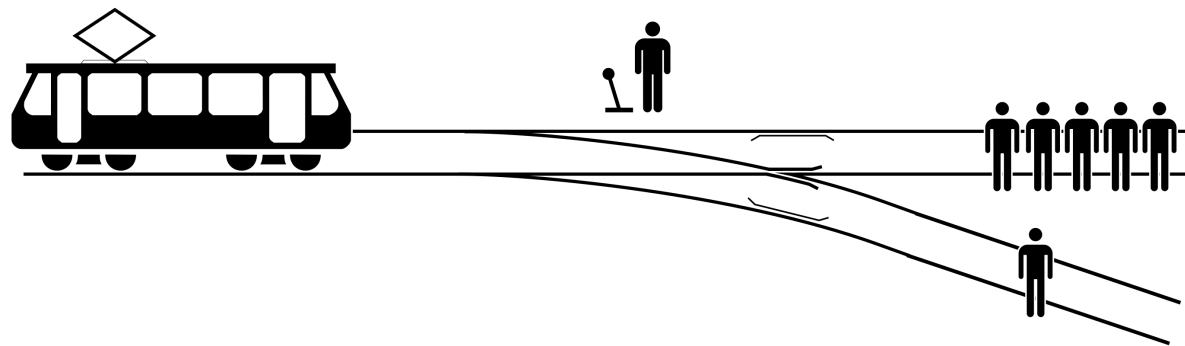
EASY



or



HARD



Women less likely to be shown ads for high-paid jobs on Google, study shows

Automated testing and analysis of company's advertising system reveals male job seekers are shown far more adverts for high-paying executive jobs



“ One experiment showed that Google displayed adverts for a career coaching service for “\$200k+” executive jobs 1,852 times to the male group and only 318 times to the female group. Another experiment, in July 2014, showed a similar trend but was not statistically significant. ”

Amazon scraps a secret A.I. recruiting tool that showed bias against women

PUBLISHED WED, OCT 10 2018•6:15 AM EDT | UPDATED THU, OCT 11 2018•2:25 PM EDT

- Amazon.com's machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.
- The team had been building computer programs since 2014 to review job applicants' resumes with the aim of mechanizing the search for top talent, five people familiar with the effort told Reuters.
- The company's experimental hiring tool used artificial intelligence to give job candidates scores ranging from one to five stars — much like shoppers rate products on Amazon, some of the people said.



The New York Times

Many Facial-Recognition Systems Are Biased, Says U.S. Study

Algorithms falsely identified African-American and Asian faces 10 to 100 times more than Caucasian faces, researchers for the National Institute of Standards and Technology found.

By Natasha Singer and Cade Metz

Dec. 19, 2019



Data Science Ethics

Informed Consent

Data Ownership

Privacy

Data Validity

Algorithmic Fairness

Data Science Ethics

★★★★★ 4.8 536 ratings



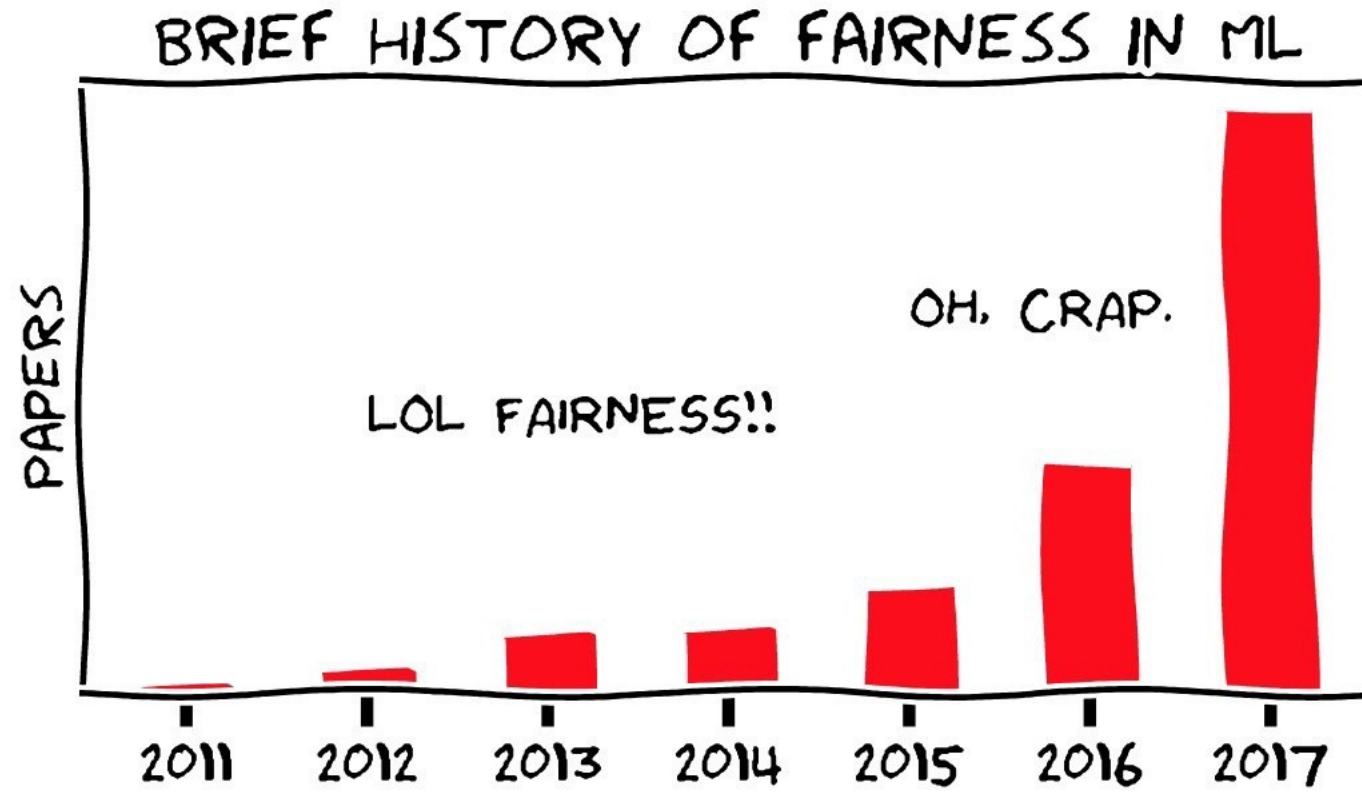
H.V. Jagadish

<https://www.coursera.org/learn/data-science-ethics/>

DS-GA 3001.009: Special Topics in Data Science: Responsible Data Science

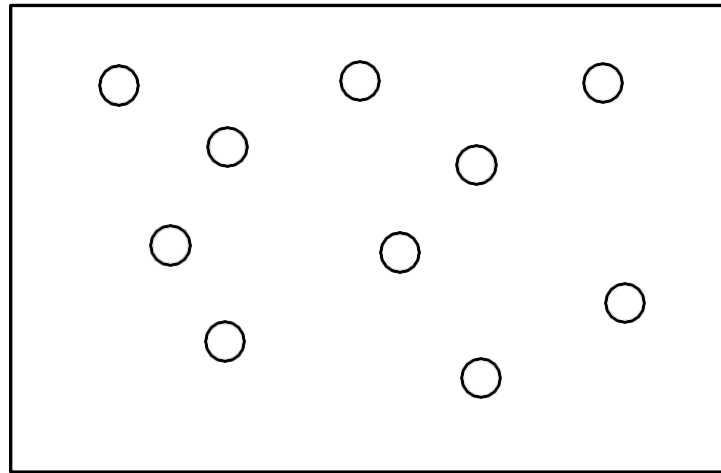
<https://dataresponsibly.github.io/courses/spring19/>

Fairness in Machine Learning

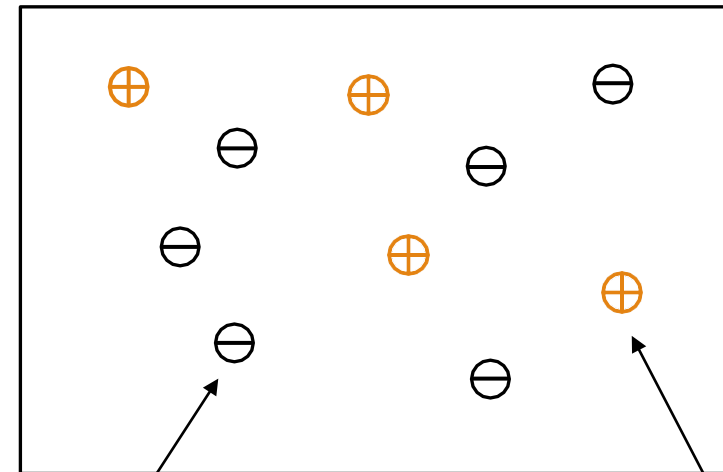


Fairness in Machine Learning

Is my model fair?



Admit 40% students to MPCS

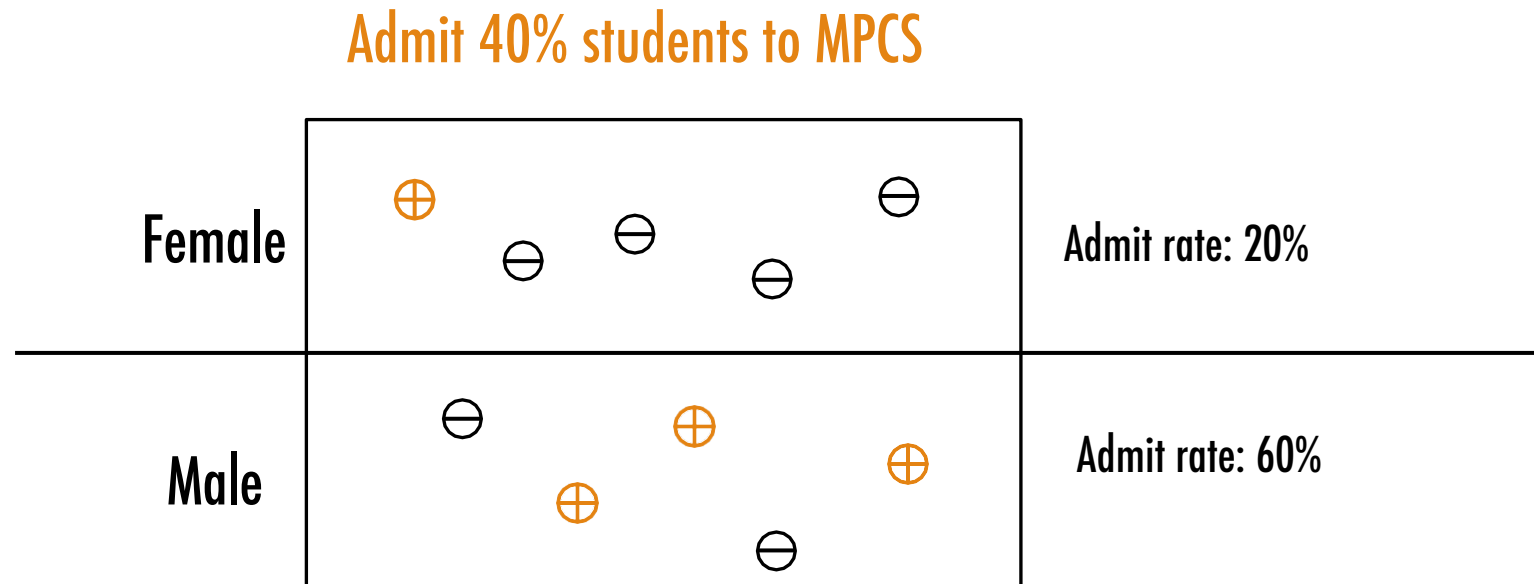


Not admit

Admit

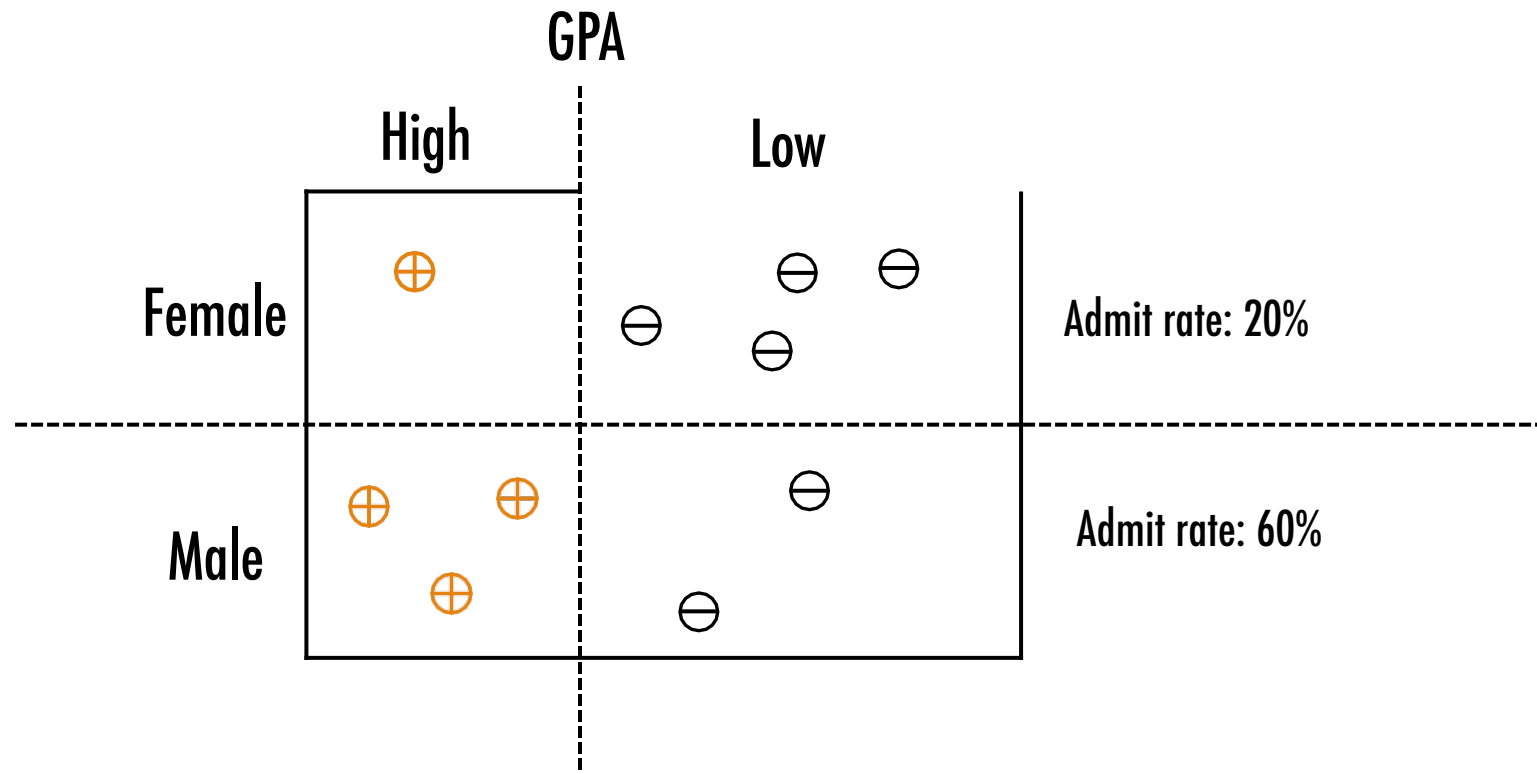
Fairness in Machine Learning

Female and male applicants are treated differently



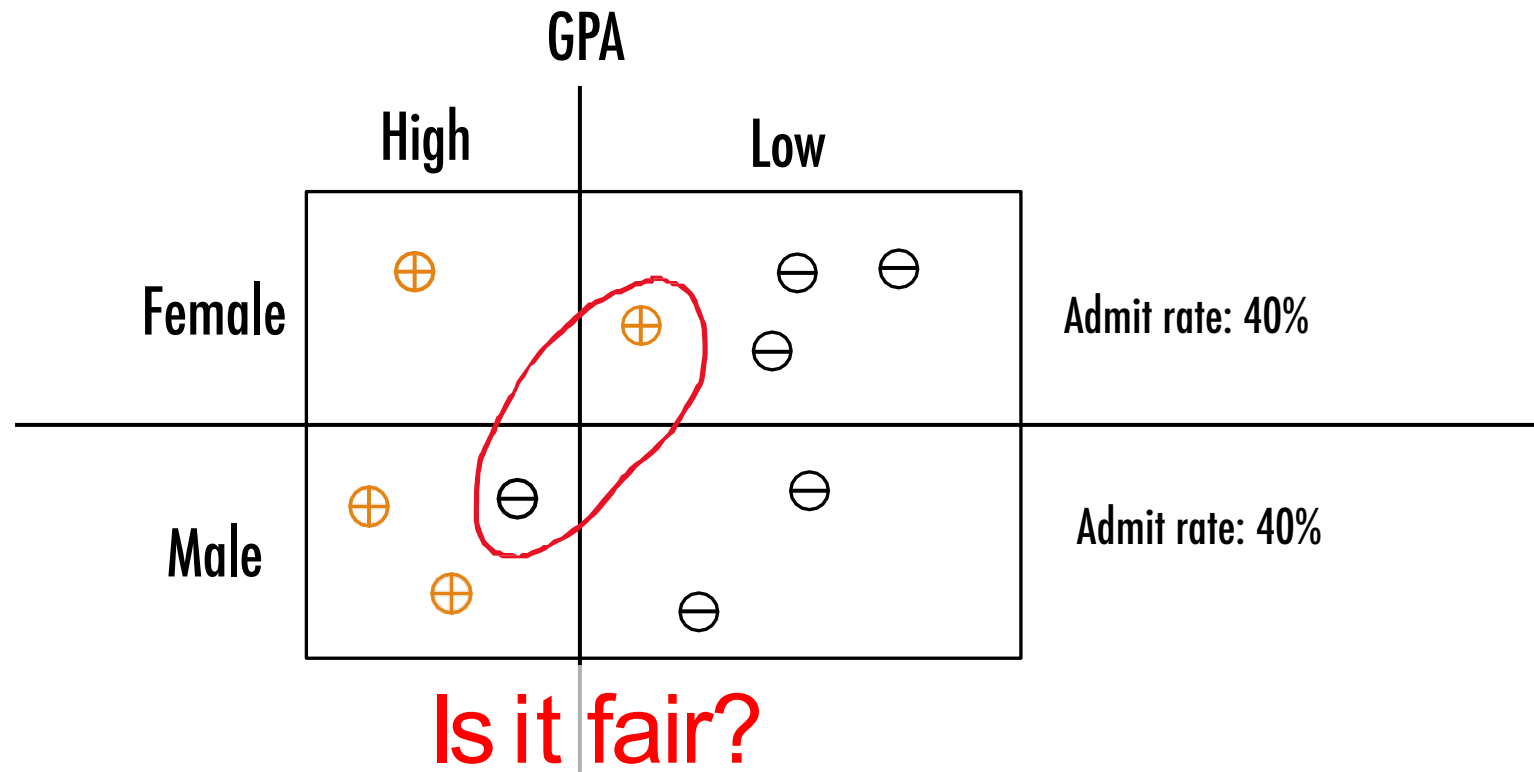
Fairness in Machine Learning

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



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>

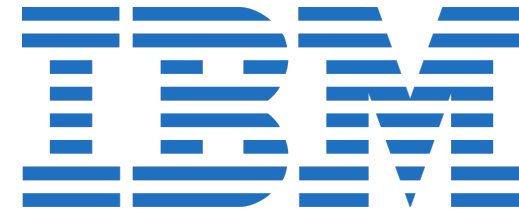
Datasets

Toolbox

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

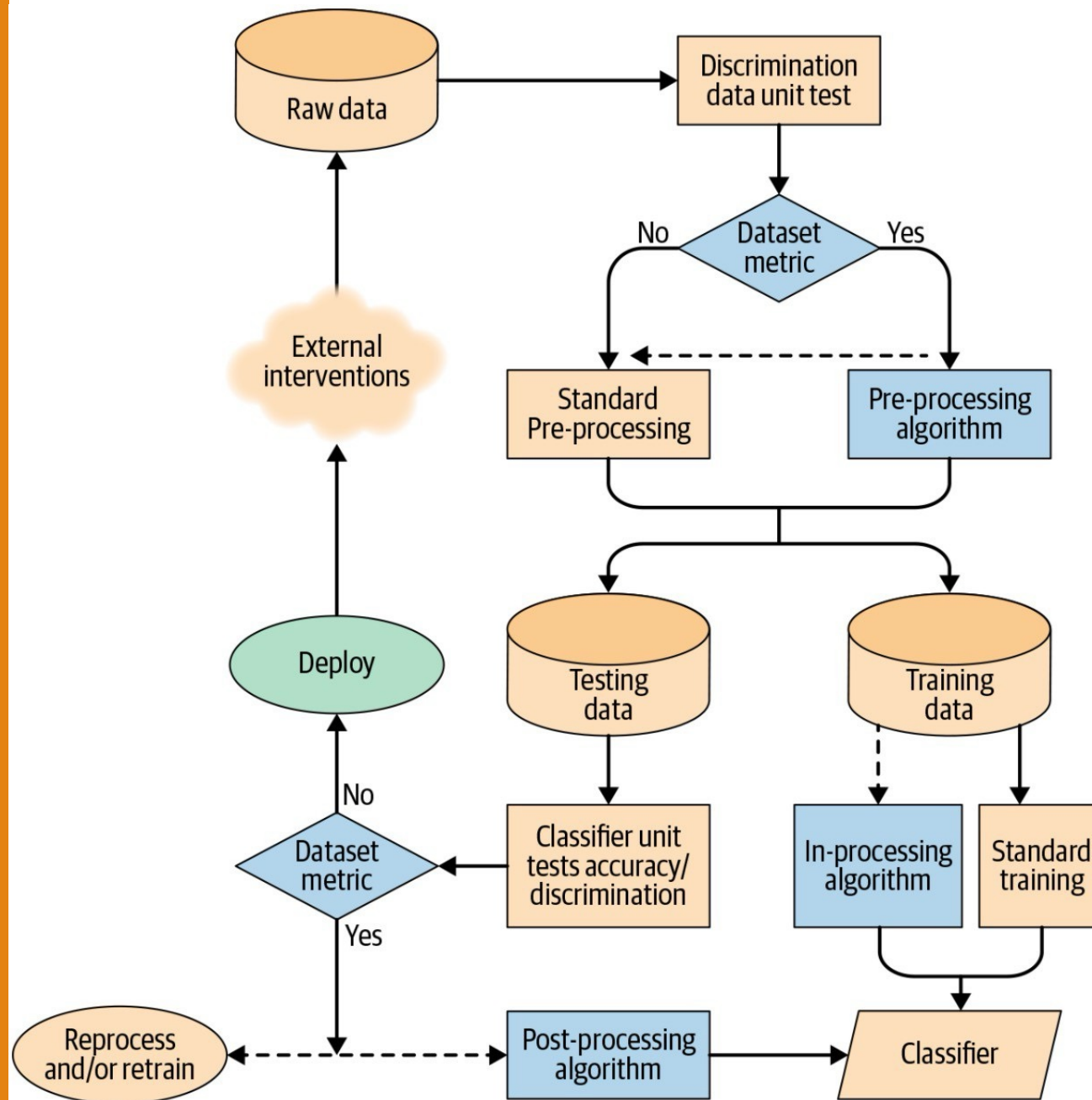
Guidance

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.



AIF360 Algorithms

Pre-processing

- Reweighing
- 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	+

Reweightings

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  $C$ 
```

F. Kamiran 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

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

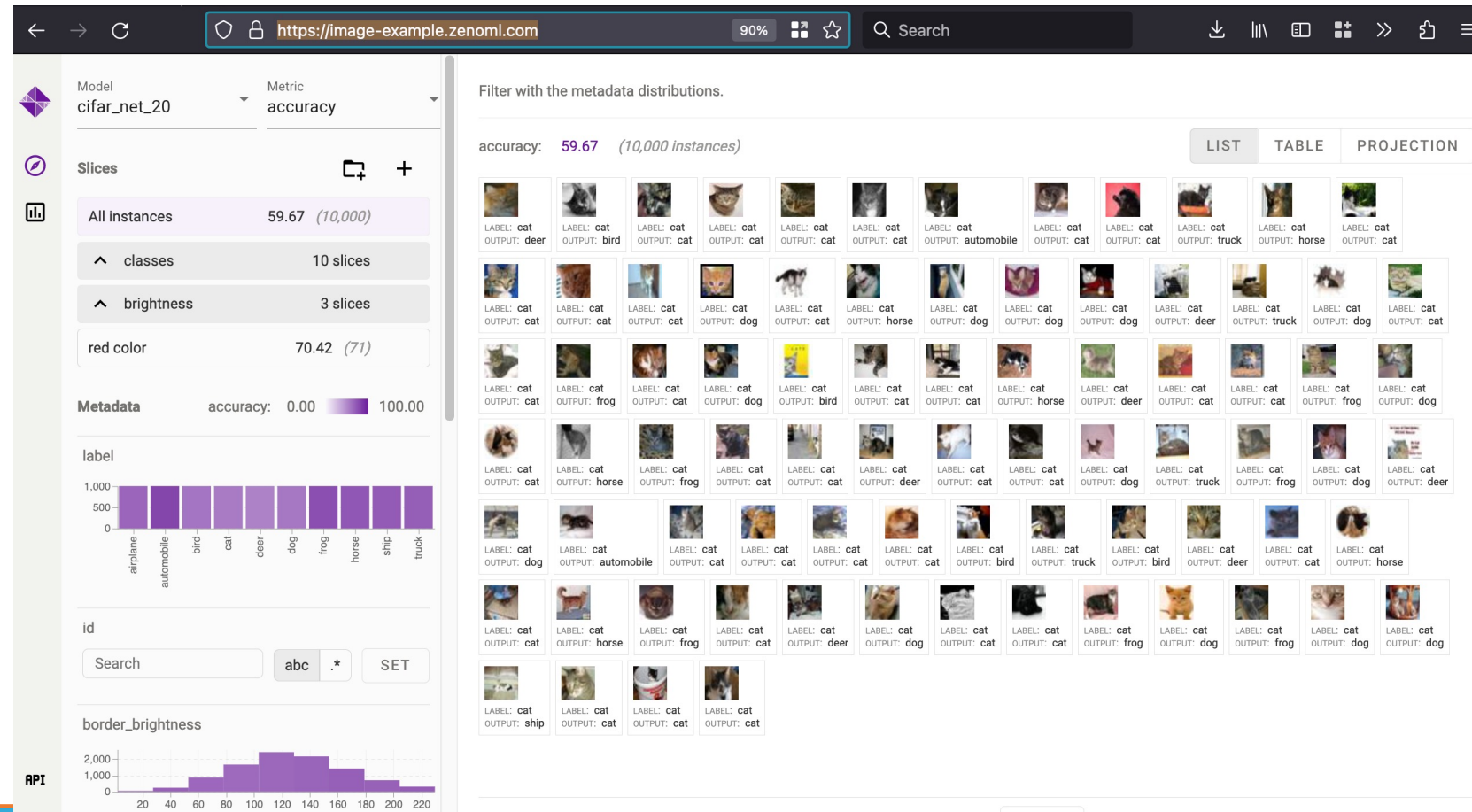
$$\frac{5 \times 4}{10 \times 1} = 2$$

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

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

AI Data Management & Eval

- <https://hub.zenoml.com/home>



Conclusion

Big Picture

- Why responsible data science?
- Data science ethics

Fairness

- Equality vs Equity
- AIF360

Reweighting



or



CMPT 733 – Big Data Programming II

Privacy Enhancing Technologies

Instructor Steven Bergner

Course website <https://coursys.sfu.ca/2024sp-cmpt-733-g1/pages/>

Slides by Ricardo Silva Carvalho | SFU's Big Data Hub

GOALS

- Familiarize with the **principles** of privacy preservation.
- Understand the goal and applicability of commonly used privacy tools.
- Identify and select the appropriate privacy technologies for a given practical scenario.



Image by [Engin Akyurt](#)

Topics Today

- Overview of privacy preserving technologies
- Previous attempts at privacy and possible attacks.



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

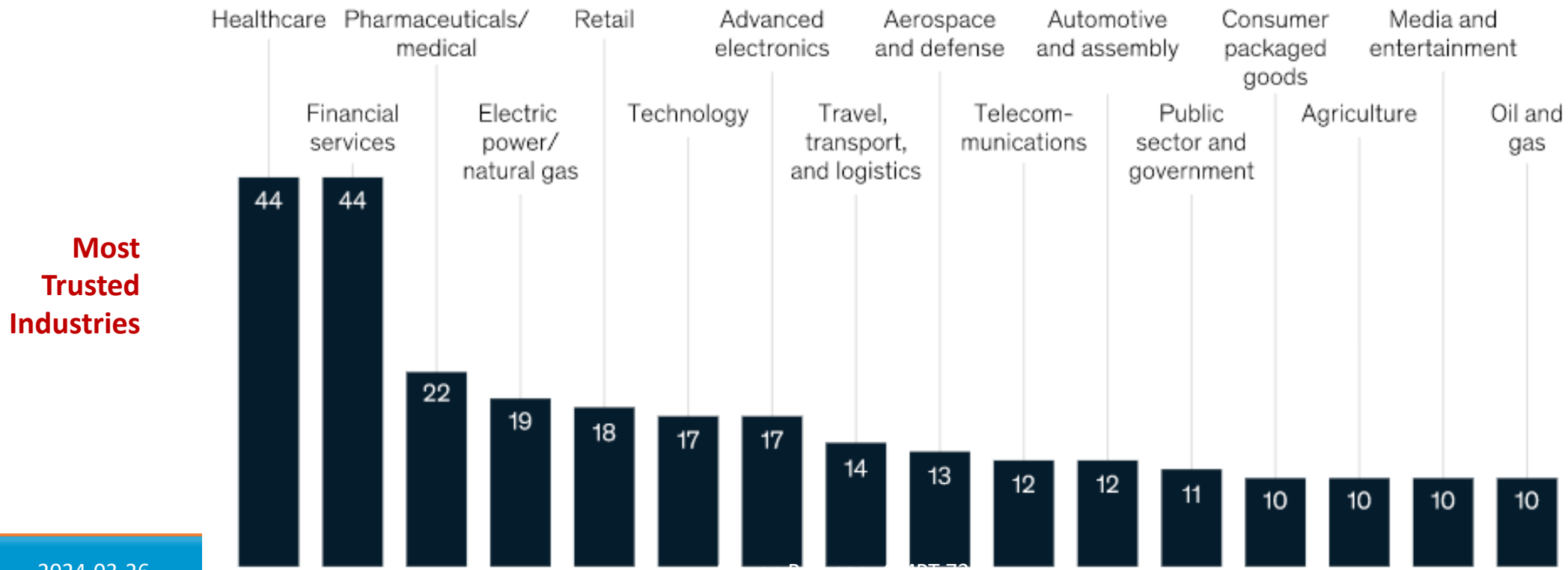
- Goal: Reduce chances of identification
- Sensitive data (*and subject*)
 - Health records
 - Search logs
 - Location data
 - Private conversation
- Disclosure can be harmful
- Data = leverage



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

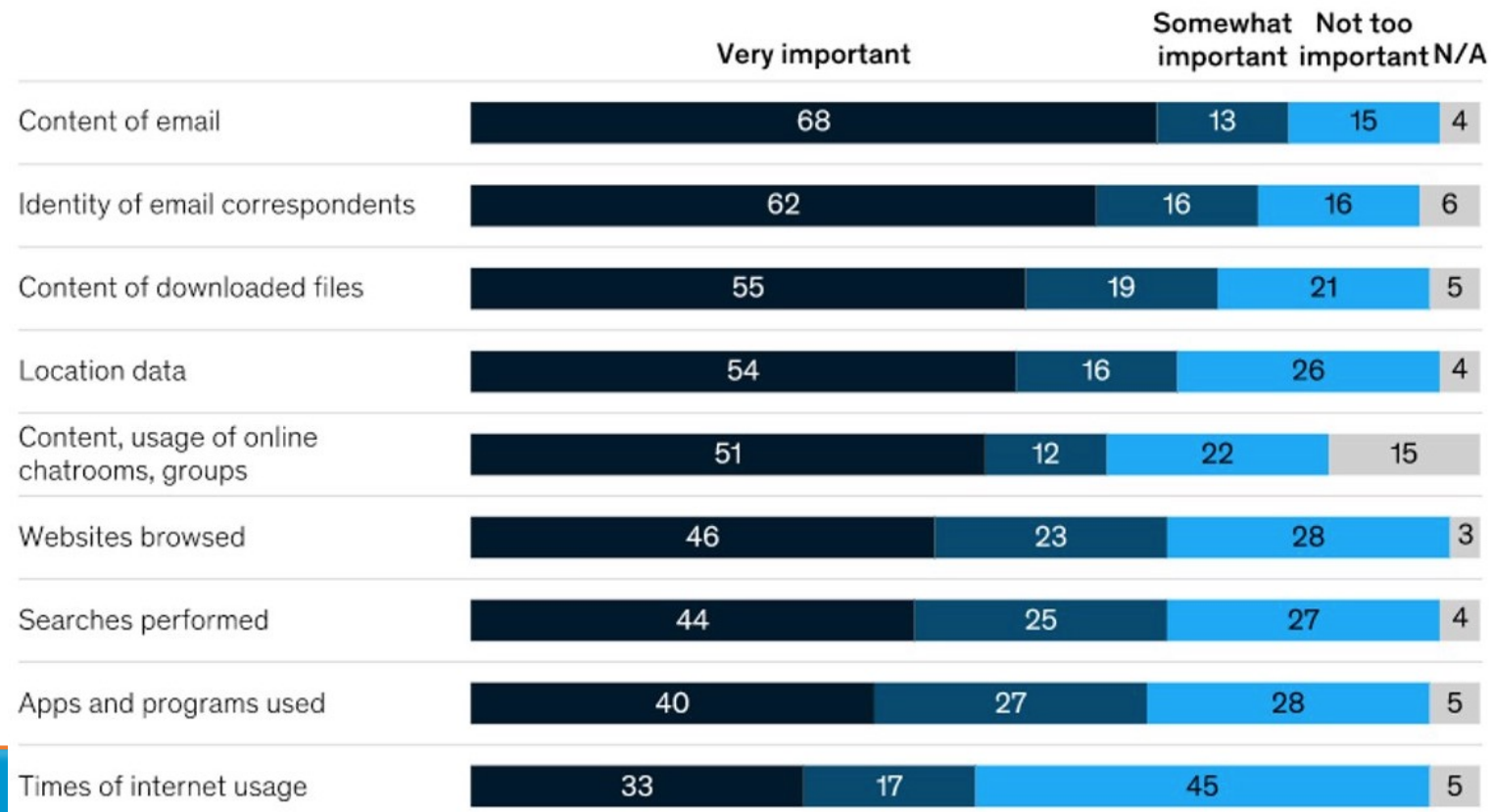


Image by [Pixabay](#)

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

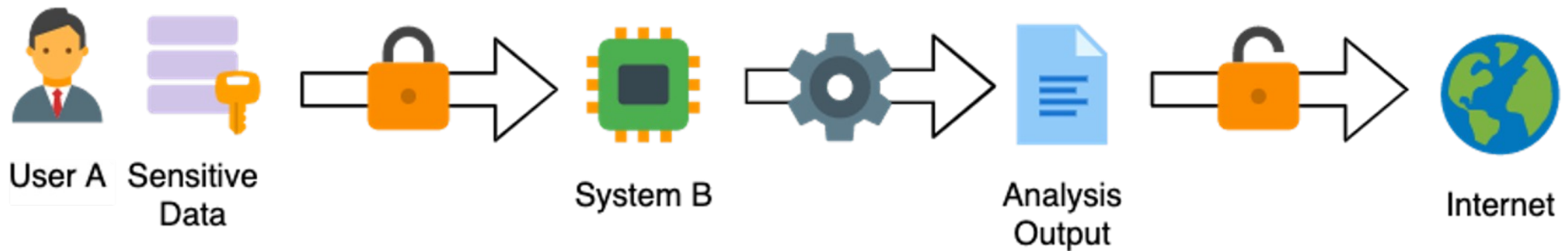


Image by [Fauxels](#)

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



Image by [Angela Roma](#)

- **Truly** anonymized data is no longer subject to GDPR

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.

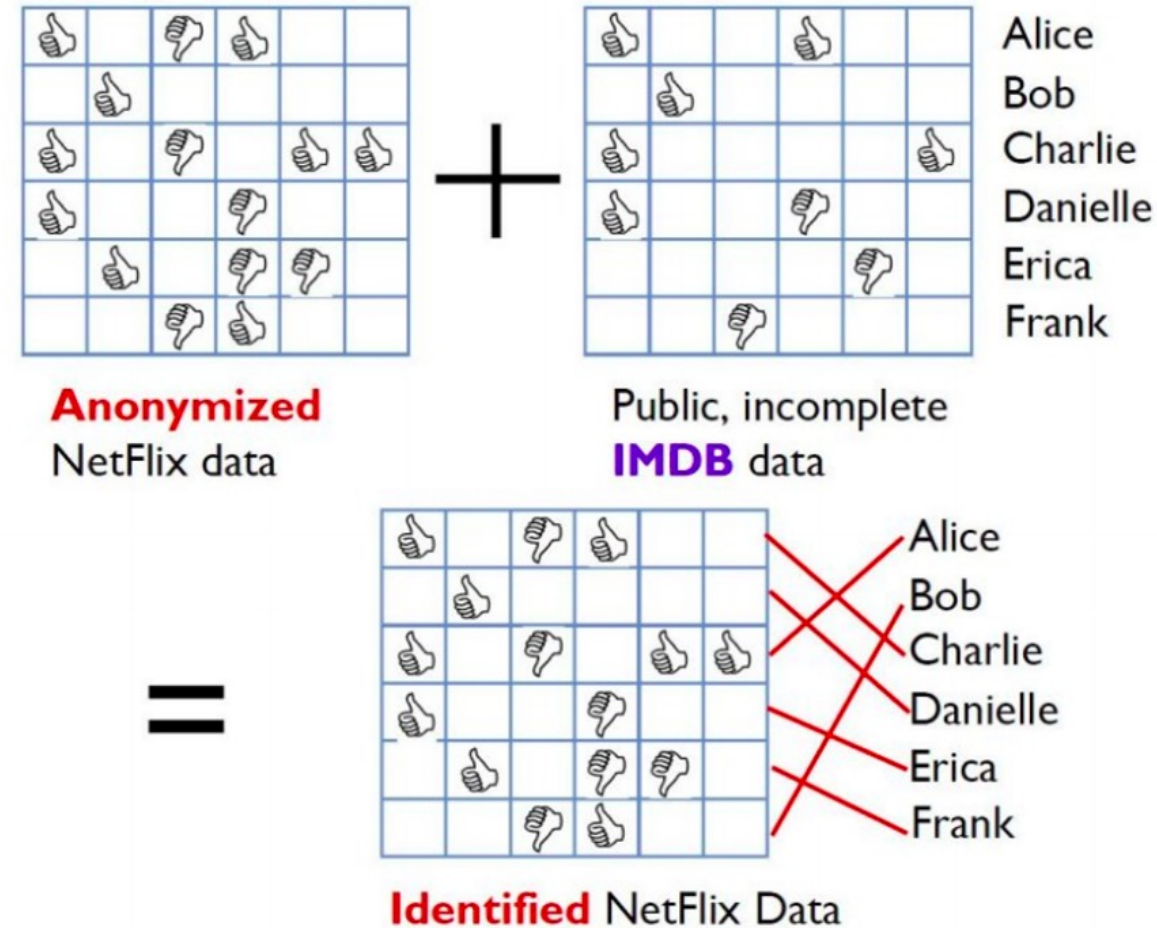


Image by: "[How to break anonymity of the Netflix Prize dataset](#)", A. Narayanan, V. Shmatikov, 2008.

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

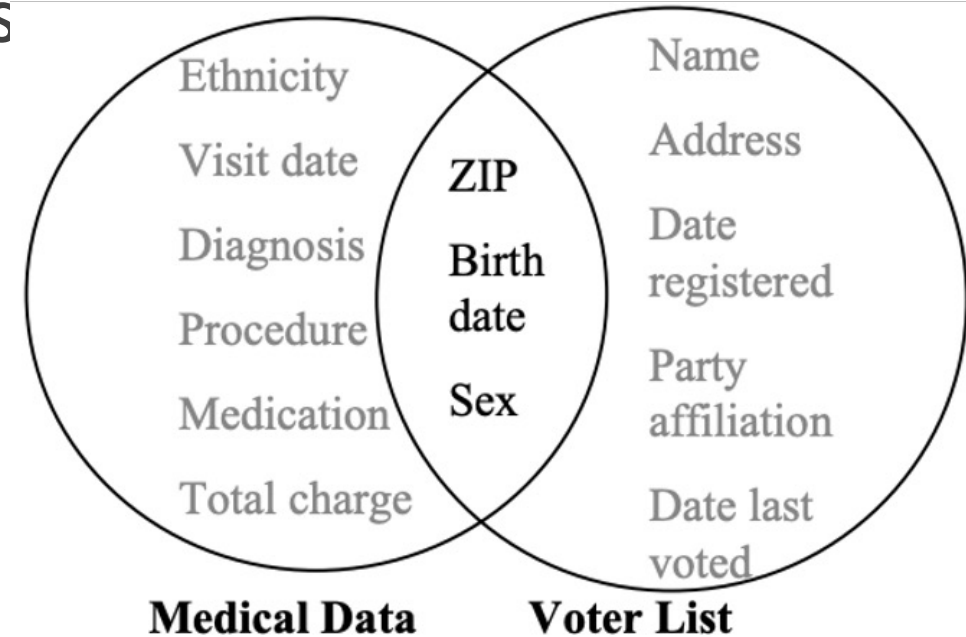


Image by: “Matching known patients to health records in Washington State Data”, L. Sweeney, 2013.

World's Biggest Data Breaches & Hacks

World's Biggest Data Breaches & Hacks

Selected events over 30,000 records

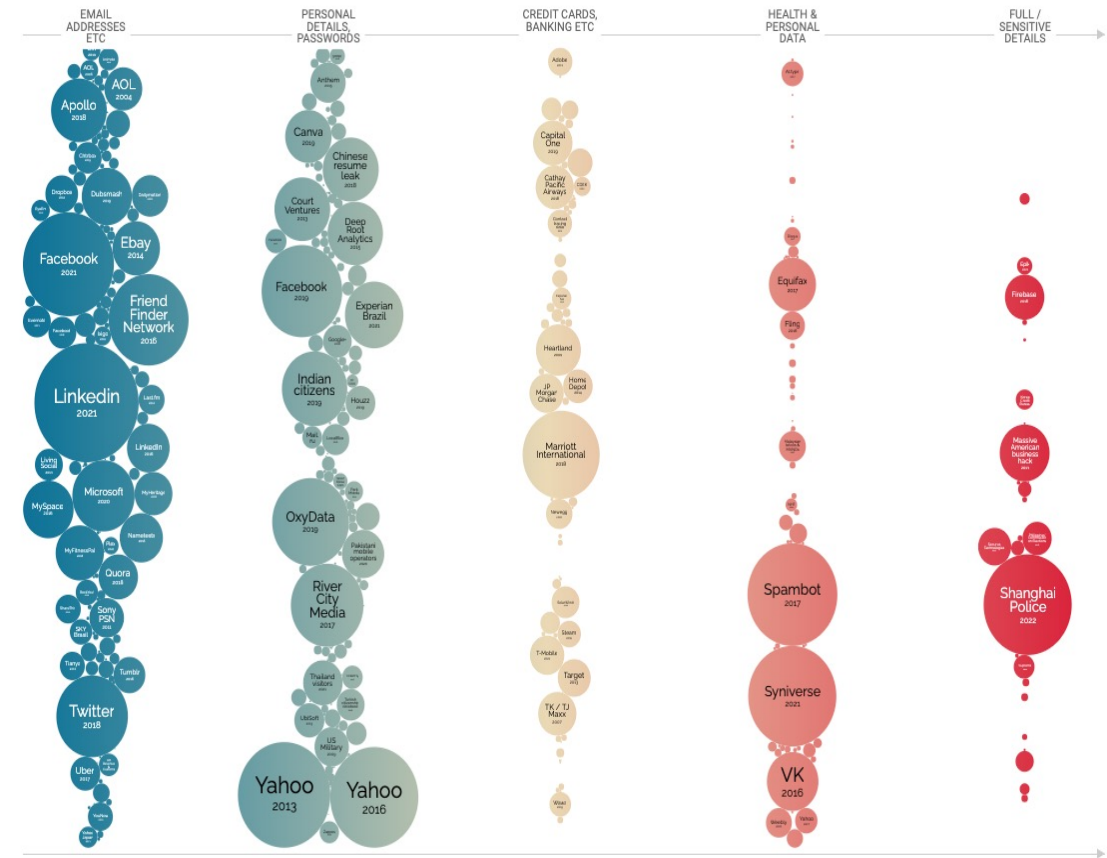
UPDATED: Sep 2022

size: records lost

interesting story



Data Breaches by data sensitivity



**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.

Examples of successful anonymization

- The following 2015 paper shows 4 examples:
 - Anonymising and sharing individual patient data, by Khaled El Emam, Sam Rodgers, and Bradley Malin
 - The GlaxoSmithKline trials repository,^{42 43} which now has multiple pharmaceutical companies using it to manage the data request process and share data (www.clinicalstudydatarequest.com)
 - The Data Sphere project, a consortium of pharmaceutical companies, sharing data from the control arm of oncology trials^{44 45}
 - The Yale University Open Data Access project, which is initially making trial data from Medtronic available^{46 47}
 - The Immport Immunology Database and Analysis Portal⁴⁸

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?

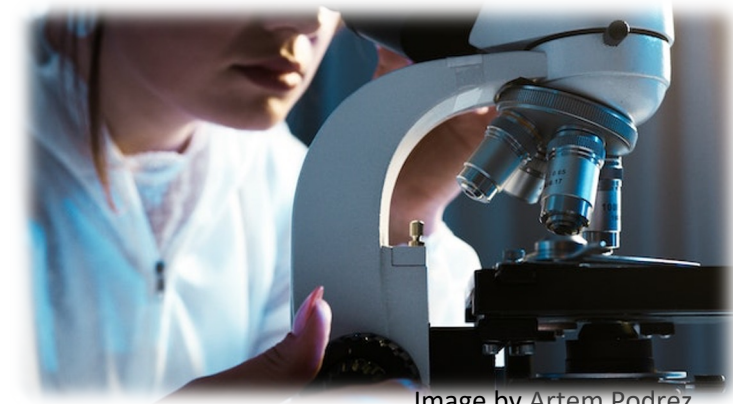


Image by [Artem Podr  z](#)

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**.

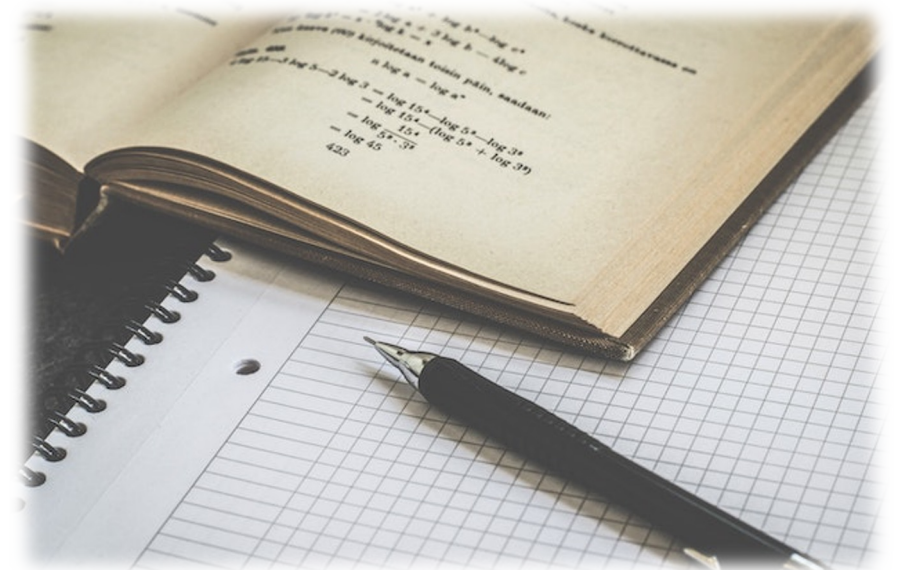


Image by [Lum3n](#)

Formal Privacy definition

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

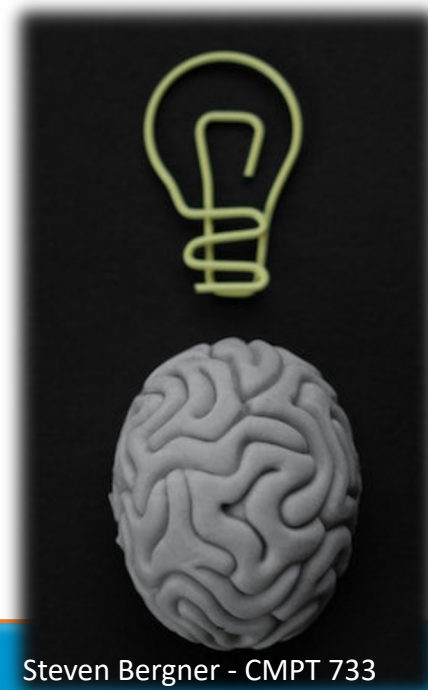


Image by E
[katerina](#)
[Bolovtsova](#)

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?

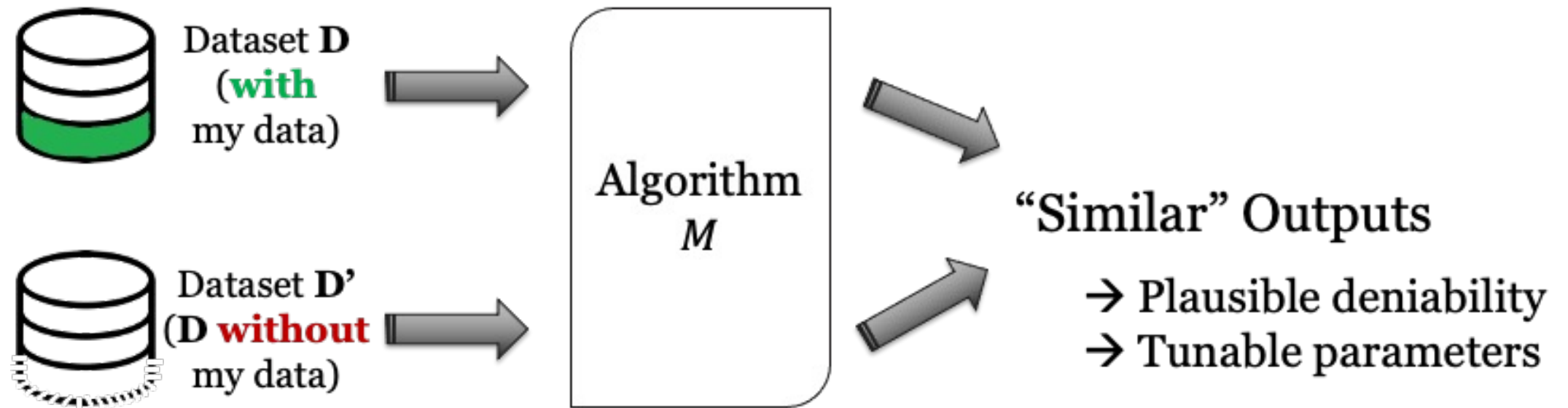
- Quote from [Dwork and Roth, 2014]:

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

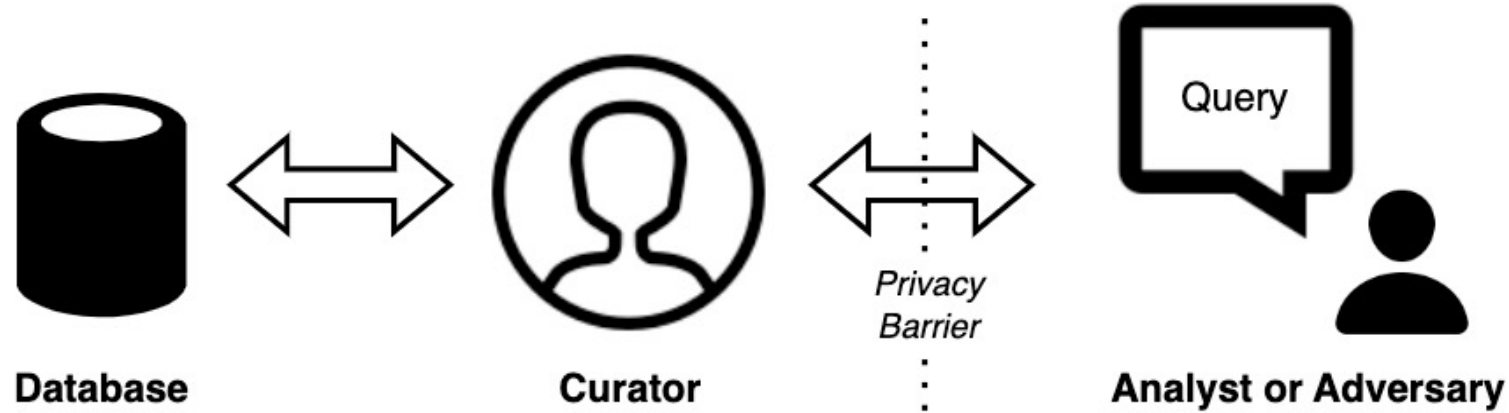
*“You will **not** be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, **no matter what other studies, data sets, or information sources, are available.**”*

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?

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

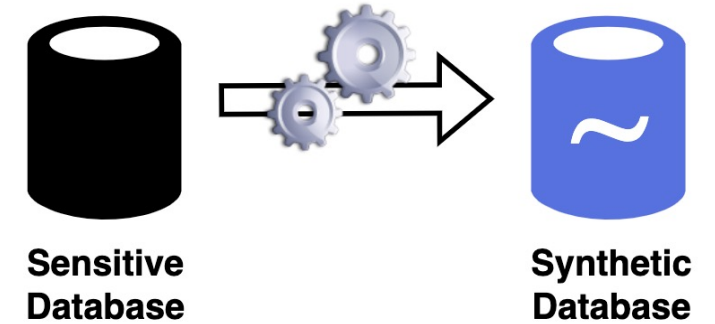
DP in summary

- [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