IMMUTA

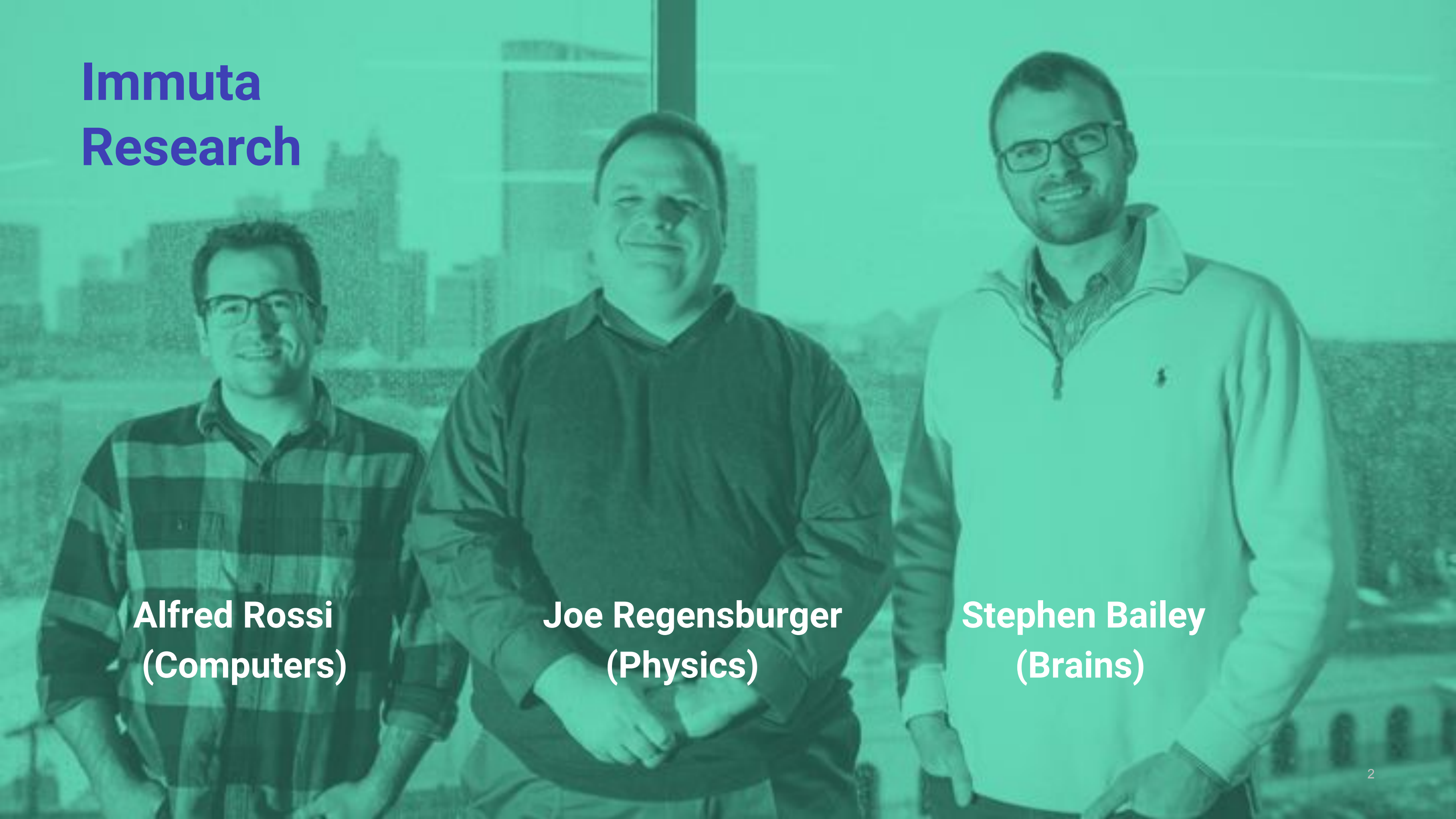
The Data Scientist's Guide to Preserving Privacy

Columbus Data Science MeetUp

Stephen Bailey* Ph.D., Alfred Rossi* Ph.D., Joe Regensburger Ph.D.

March 20, 2019

Immuta Research



Alfred Rossi
(Computers)

Joe Regensburger
(Physics)

Stephen Bailey
(Brains)

A (very comprehensive) innovation timeline

2000s

2010s

March 20, 2019

Generating and Storing Data

Expanding digital footprint turned business activity, consumer behavior, social lives, etc. into data.

Deriving Value From Data

Democratized tools, machine learning and interactive analytics made data-driven decision-making possible and valuable.

Responsibly Leveraging Data

Increased risk associated with data utilization will generate new legal, ethical, and management tools and standards.

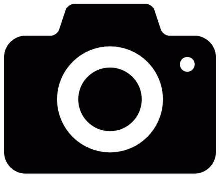
e.g.,https://ai.google/education/responsible-ai-practices

Technologies of publicity

The expectation of privacy -- and its violation by others -- is a distinctly modern problem.

Philosophizing about privacy went mainstream in the late 1890s, fueled by technological advances that made it much easier to know and be known:

- High speed photography
- Tabloids
- Telephones and wiretapping



… now the right to life has come to mean the right to enjoy life -- the right to be let alone; ... and the term “property” has grown to comprise every form of possession -- intangible as well as tangible.

Louis Brandeis and Samuel Warren
The Right to Privacy, 1890

HARVARD LAW REVIEW.

VOL. IV. DECEMBER 15, 1890. NO. 5.

THE RIGHT TO PRIVACY.

"It could be done only on principles of private justice, moral fitness, and public convenience, which, when applied to a new subject, make common law without a precedent; much more when received and approved by usage."

WILLES, J., in Millar v. Taylor, 4 Burr. 2303, 2312.

THAT the individual shall have full protection in person and in property is a principle as old as the common law; but it has been found necessary from time to time to define anew the exact nature and extent of such protection. Political, social, and economic changes entail the recognition of new rights, and the common law, in its eternal youth, grows to meet the demands of society. Thus, in very early times, the law gave a remedy only for physical interference with life and property, for trespasses *vi et armis*. Then the "right to life" served only to protect the subject from battery in its various forms; liberty meant freedom from actual restraint; and the right to property secured to the individual his lands and his cattle. Later, there came a recognition of man's spiritual nature, of his feelings and his intellect. Gradually the scope of these legal rights broadened; and now the right to life has come to mean the right to enjoy life,—the right to be let alone; the right to liberty secures the exercise of extensive civil privileges; and the term "property" has grown to comprise every form of possession—intangible, as well as tangible.

Thus, with the recognition of the legal value of sensations, the protection against actual bodily injury was extended to prohibit mere attempts to do such injury; that is, the putting another in

What is privacy, from a legal perspective?

HIPAA (2003)

A covered entity is permitted to use **protected health information** from which certain specified direct identifiers of individuals and their relatives, household members, and employers have been removed.

GDPR (2016)

Data subjects are the ultimate owners of their data and retain rights over that data, including:

- right to erasure
- right to be informed
- right to restrict processing
- right to data portability

https://www.hhs.gov/sites/default/files/privacysummary.pdf
https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/

What is privacy, from a data perspective?

id	name	birth_date	total_sales
...	...	...	...

No data

id	name	birth_date	total_sales
839	lorem	01-01-1970	0.00
783	ipsum	01-01-1970	0.00
12	dolor	01-01-1970	0.00
1034	sit	01-01-1970	0.00
8314	amet	01-01-1970	0.00
...	...	...	...

Random data

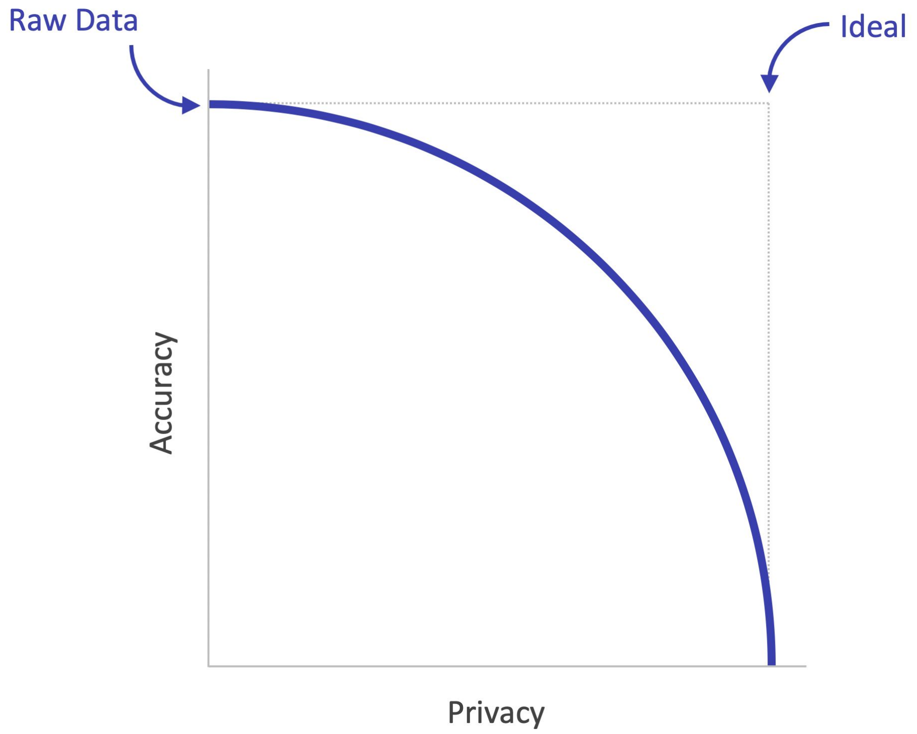
What is privacy, from a data scientist perspective?

In practice, privacy is a continuum

To preserve privacy, organizations have to make the data less closely resemble the raw data (or full data).

Moving along this curve, data become more robust against certain types of privacy risks.

The actual trade-off is highly coupled with analytical context.



NO:
How much data can I get?

YES:
How much information do I need?

Privacy controls

Establishing some terminology

Dataset

Attributes

Records

id	name	birth_date	income
1056	Leslie Knope	04-16-1973	58,850.23
1057	Ron Swanson	09-21-1964	87,500.12
1058	April Ludgate	11-16-1992	28,043.56
1059	Jerry Gurgich	02-14-1955	1,064.75
1060	Andy Dwyer	06-20-1990	21,320.00
...	...	...	...

Records may represent people or instances of activity which should be protected from disclosure.

Attribute types

- Identifier
- Quasi-identifier
- Sensitive

What are we safeguarding?

Identity Disclosure

Identify that a record corresponds to an individual.

Attribute Disclosure

Reveal or closely estimate the value of an attribute.

Participation Disclosure

Identify that an individual contributed to some analytical product.

Dataset Linkability

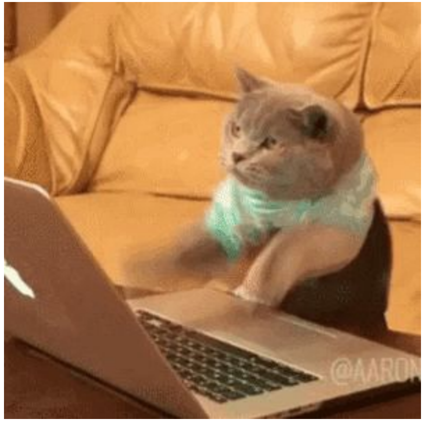
Identify groups of two or more related records.

Privacy controls

To protect against attacks, think like an attacker

Anticipate and do not underestimate. Adopting a “worst-case” attitude ensures that you won’t underestimate an attacker. Imagine your adversary has...

Time, energy & resources

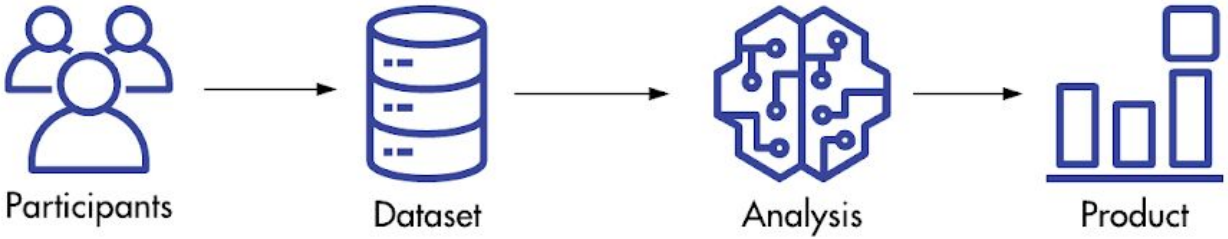


Ample external knowledge

Privacy controls

Data protection by design

Private information can leak throughout the data science workflow, and each step allows for different privacy-preserving techniques.



How can we accomplish this?

Organizations have adopted a variety of practices in the pursuit of privacy.

De-identification

Replace identifying or quasi-identifying attributes with substitute information, a la HIPAA

k-Anonymization

Suppress or generalize information in such a way that it can no longer be traced to an individual record

Differential Privacy

Formally limit the ability of an attacker to reason about analysis input from observing the output.

Privacy-preserving approaches

De-identification

Defining de-identification

De-identification uses masking to replace identifying attributes, while preserving structure.

Masking can be generalized as :

$$m' = f_s(m)$$

There are many masking techniques available, and in general, most privacy-preserving techniques can be considered a form of masking.

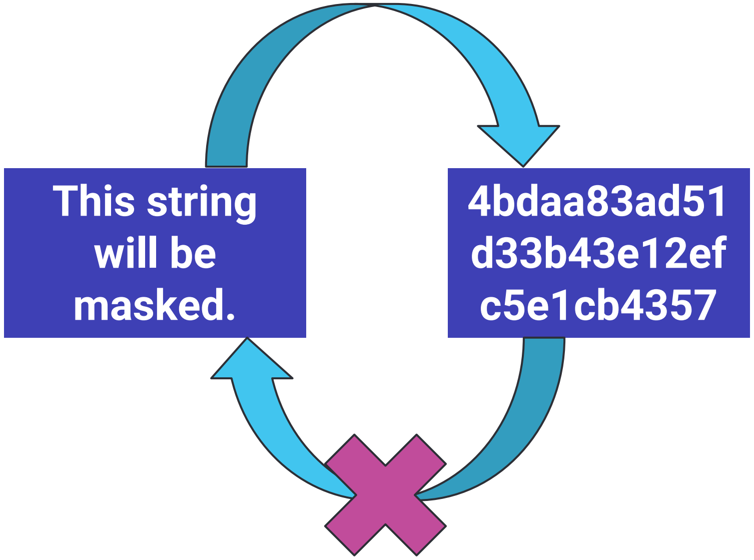
Methods of Masking

Original Value
Leslie Knope
Ron Swanson
April Ludgate
Jerry Gurgich
Andy Dwyer

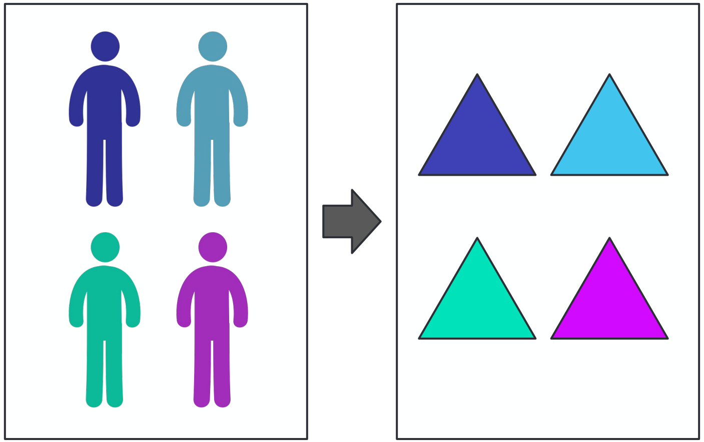
Arbitrary Rule	Regular Expression	Tokenization	Cryptographic Hashing
Person A	Leslie	TGVzbGll...	62a50607e...
Person B	Ron	Um9uIFN...	7413d7386...
Person C	April	QXByaW...	1ff4f635da...
Person D	Jerry	SmVycnk...	3745f5186...
Person E	Andy	QW5keSB...	5e7af7c25...

What makes a good mask?

Hard to invert



Preserves some useful structure



What attributes should be masked?

Identifying attributes

Uniquely identifying information, including full name, SSN, phone number, address, fingerprints, and photographs.

Quasi-identifying attributes

Attributes which can be leveraged to uniquely identify an individual. Can be combined to identify an individual.

- IP Address
- Zip code
- Ethnicity

HIPAA lists 18 types of "personally identifiable information" but the list may be much longer in the current data climate!

https://www.luc.edu/its/aboutits/itspoliciesguidelines/hipaainformation/18hipaaidentifiers/

Implementation notes

Cryptographic hashing

A strong masking process is **cryptographic hashing**.

1. Prepend or append the original value with a “salt”.
2. Mask with a strong hash function, such as SHA-128 or RIPEMD.

Implementation notes

Attribute masking

Hashing can be easily implemented in many different languages and dialects.

Other forms of masking include:

- Reversible masking
- Format-preserving encryption
- Locality sensitive hashing

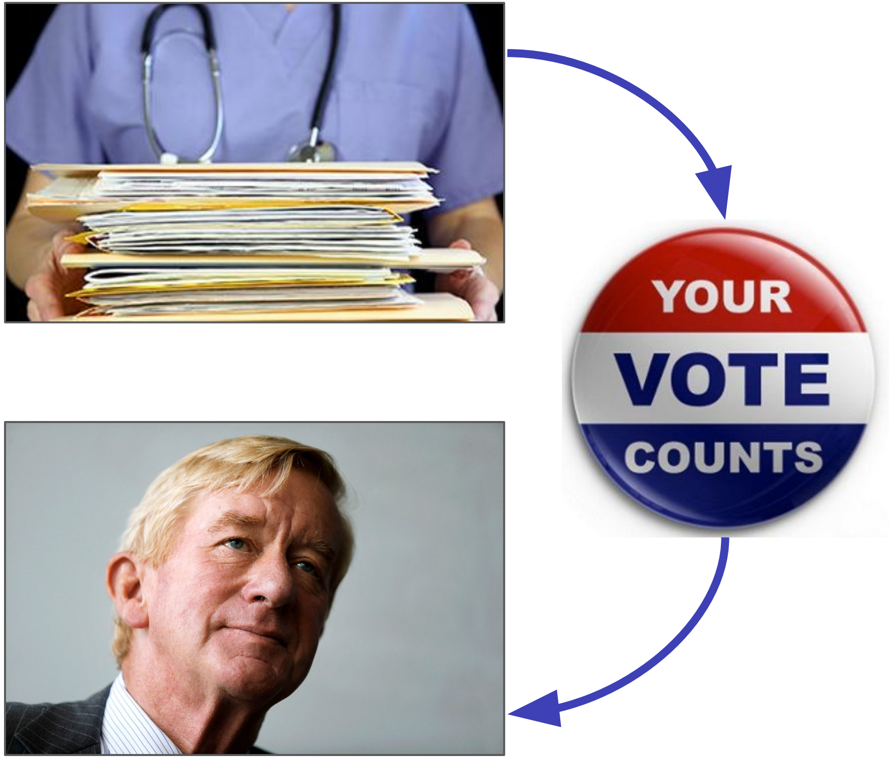
```
import hashlib
import pandas

def mask_value(s, salt):
    salted_bytes = (salt + s).encode()
    h = hashlib.sha1(salted_bytes)
    return h.message_digest()

df = pandas.read_sql_table(tab, conn)
id_cols = [“name”, “address”]

df[id_cols].apply(mask_value)
       “name”     |      “address”
    “0c3e5870…”   | “5f594b62b8a1dn0Q1…”
    “4bdaa83a…”   | “d51d33b43e12eYip2…”
```

Attacking a de-identified dataset

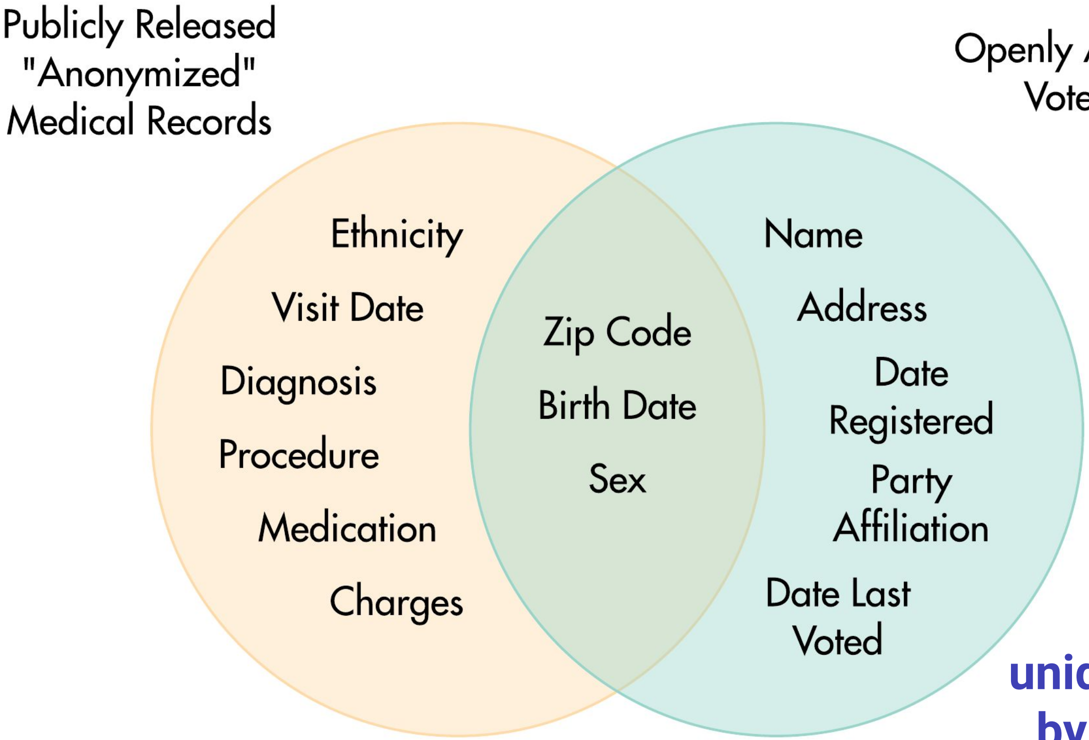
In 1997, Massachusetts General Hospital released about 15,000 medical records.

To protect privacy, **they masked identifying attributes**, including patient name and address.

Harvard researcher Latanya Sweeney “linked” publicly available datasets to identify the records of then-governor Bill Weld.



Attacking a de-identified dataset



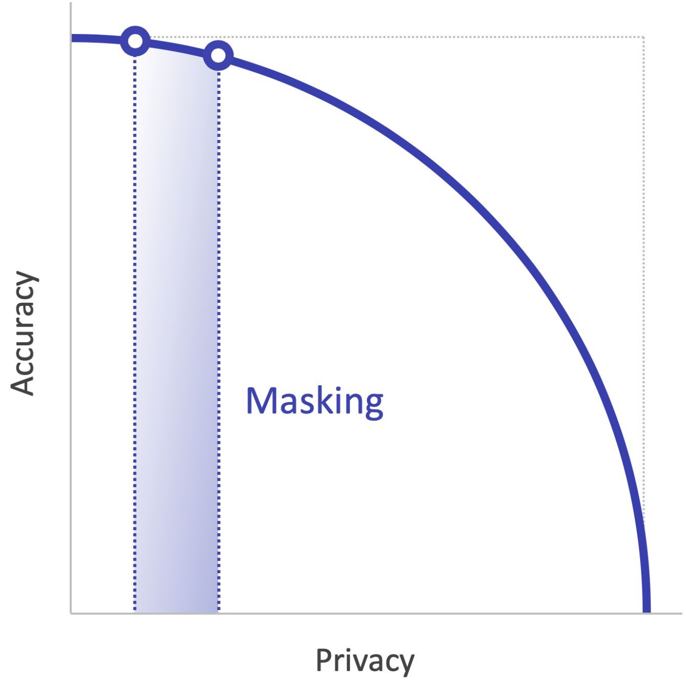
87% of the population is uniquely identifiable by Zip Code, Birth Date and Sex!

When to use de-identification

De-identification is commonbecause it preserves structure of the raw data while obfuscating semantic content.

However, it provides *fragile* privacy protections, especially in the age of massive open datasets.

Recommendation:
Masking should be used liberally, especially as first line of defense in protecting sensitive data attributes.



Disclaimer: Chart is for illustration purposes only. The actual trade-off is coupled tightly to context.

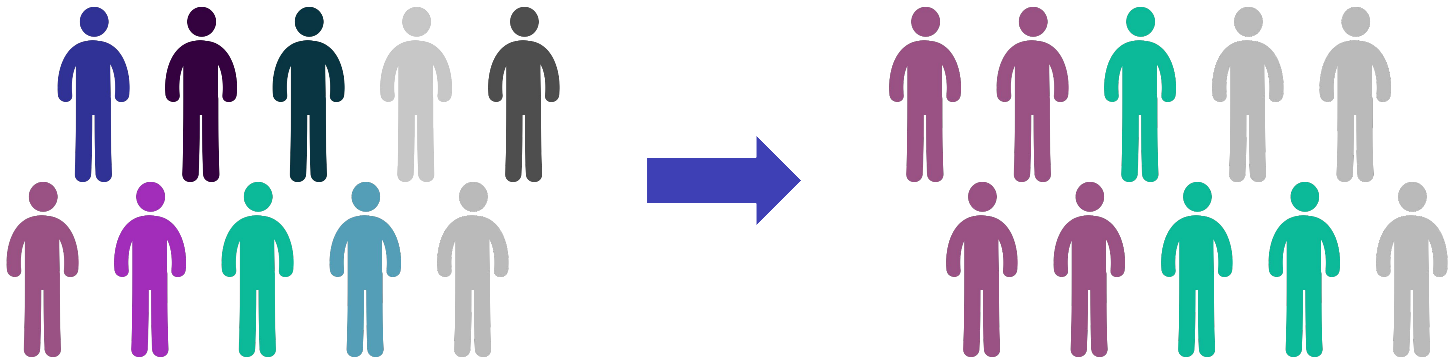
Privacy-preserving approaches

k-Anonymization

Defining *k*-anonymity

An individual cannot be distinguished from at least *k*-1 others in the dataset.

The idea is to limit the discriminating power of quasi-identifiers.



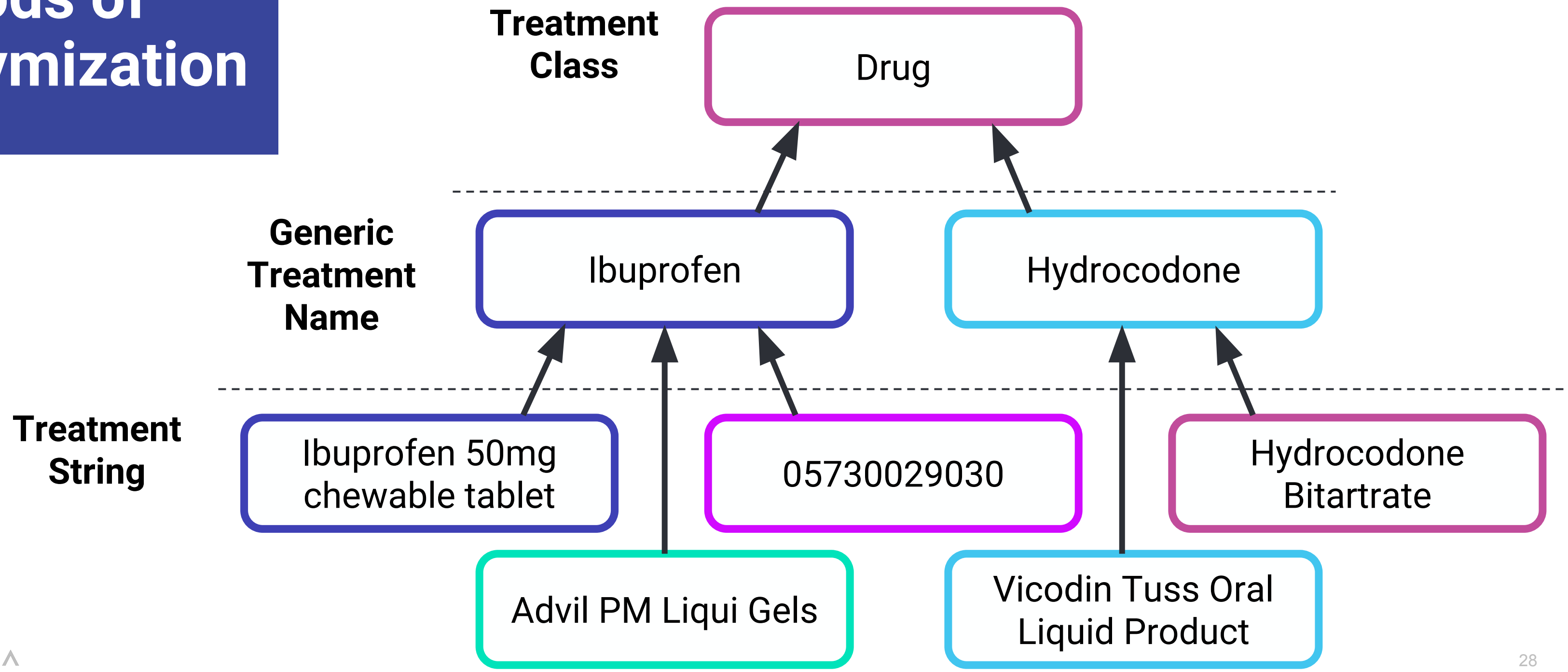
10 distinct hues

3 distinct hues

Methods of Anonymization

Generalization

Reduce specificity by grouping similar values together.



Methods of Anonymization

Suppression

Reduce cardinality of dataset by removing attributes.

8 distinct records

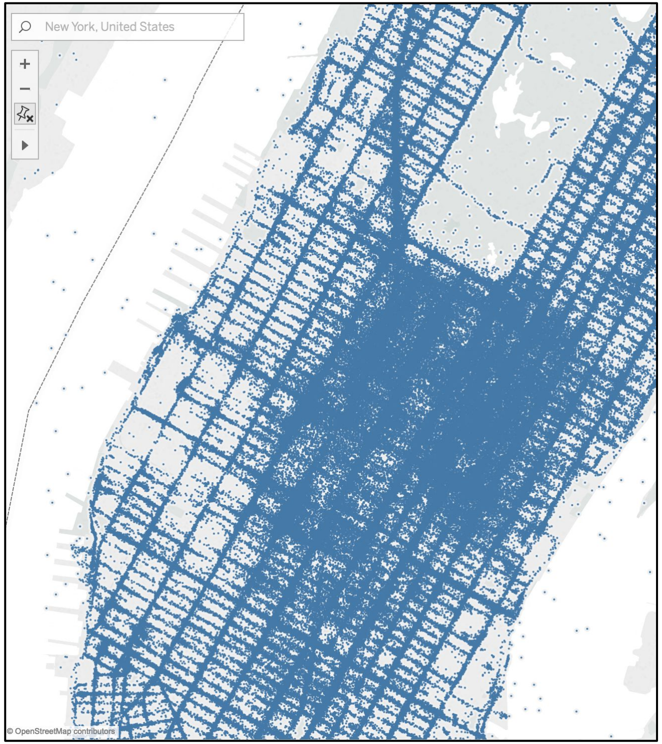
Name	Sex
Leslie Knope	F
Ron Swanson	M
April Ludgate	F
Jerry Gurgich	M
Andy Dwyer	M
Tom Haverford	M
Donna Meagle	F
Ann Perkins	F

2 distinct records

Name	Sex
	F
	M
	F
	M
	M
	M
	F
	F

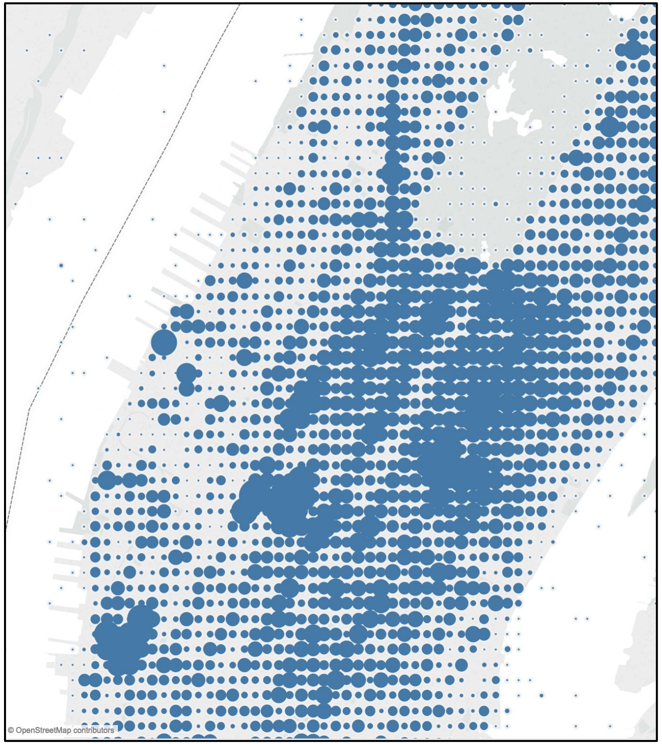
Anonymity in action

Bucketing GPS coordinates



Round to nearest 0.005°

Delete locations w/ fewer than *k* values



Implementation notes

Outliers can make anonymization difficult

It may be helpful to restrict analysis scope to achieve anonymity, either by removing outliers or setting minimum / maximum caps (e.g., "> 100")

k=1

Location	Date
Austin	01-16-2019
Dallas	07-15-2018
San Antonio	05-06-2005
San Francisco	02-01-2019
Los Angeles	12-26-2018
Columbus	02-24-2018
Cleveland	04-13-2019

Generalize to State & Decade

k=1

Location	Date
Texas	01-01-2010
Texas	01-01-2010
Texas	01-01-2000
California	01-01-2010
California	01-01-2010
Ohio	01-01-2010
Ohio	01-01-2010

Remove outlier

k=2

Location	Date
Texas	01-01-2010
Texas	01-01-2010
California	01-01-2010
California	01-01-2010
Ohio	01-01-2010
Ohio	01-01-2010

k-anonymization

To evaluate anonymity levels:

- Suppress identifiers and generalize quasi-identifiers
- Group by quasi-identifiers and count records in each bucket
- The minimum is *k*.

To accomplish anonymization, there are multiple algorithms available -- but not many are implemented in common frameworks.

```
import pandas

df = pandas.read_sql_table(tab, conn)
qid_cols = [“gender”, “city”]

df.groupby(qid_cols).size().min()
    {
        (“m”, “plain city”): 4,
        (“f”, “columbus”): 34,
        (“m”, “columbus”): 21,
        (“f”, “dayton”): 53,
        ...
    }
```

Not tested, but a possible implementation:
https://github.com/alessiovierti/python-datafly

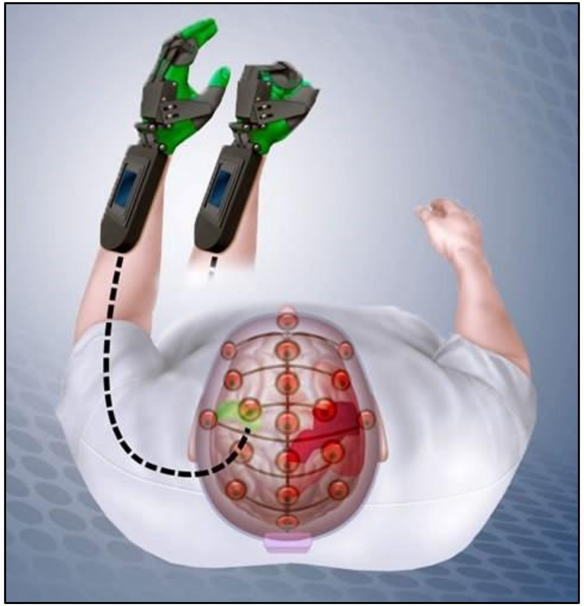
Attacking a *k*-anonymized dataset

An example scenario:

Testing patients -- some are amputees -- as part of a neuro-prosthetics trial at the nearby Veteran's Affairs Hospital.

You collect some clinical and psychological measures, including IQ, to compare performance throughout the trial.

You want to *k*-anonymize these data and share with a collaborator.



Attacking a *k*-anonymized dataset

To 3-anonymize the dataset, the researchers grouped on **two quasi-identifiers**.

They also generalized a sensitive attribute, **IQ**, to be less specific

Age	Limbs	IQ	Mobility
...	...	...	...
• 40	• 2	Medium	45
• 50	• 3	Low	52
• 50	• 3	Low	34
• 50	• 3	Low	41
• 40	• 4	High	23
• 40	• 4	Medium	46
...	...	...	...

Despite the dataset being 3-anonymous, the authors have leaked that any individual in their 50's with 3 limbs has a Low IQ score -- not a flattering result!

Extending *k*-anonymization: ℓ-diversity

ℓ-diversity adds an additional constraint to anonymization: each group must contain at least ℓ "well-represented" values.

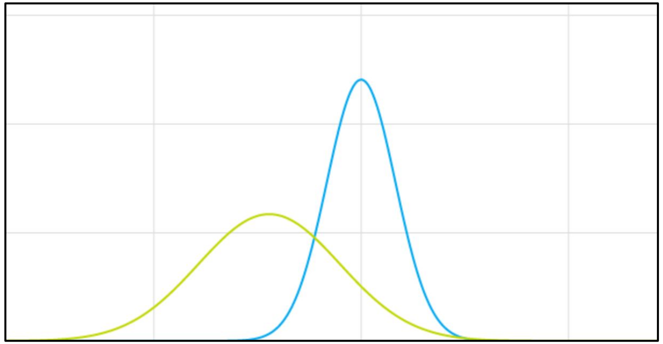
However, *probabilistic* inference attacks may still be possible.

- All but one person in the (50 y/o + 3 limbs) group has a mobility score lower than 42!

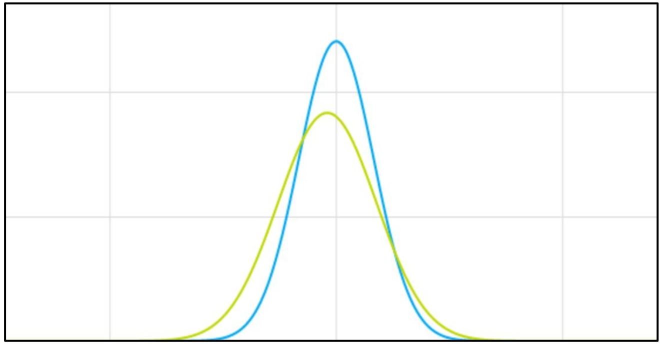
Age	Limbs	IQ	Mobility
...	...	...	...
• 40	• 2	Medium	75
• 50	• 3	**Low**	23
• 50	• 3	**Low**	40
• 50	• 3	**Medium**	15
• 50	• 3	**Medium**	28
• 50	• 3	**Medium**	37
• 50	• 3	**Medium**	42
• 50	• 3	**High**	108
• 40	• 4	Medium	83
...	...	...	...

Extending *k*-anonymization: *t*-closeness

We can circumvent probabilistic inferences by ensuring that within-group distributions closely match the global distributions.



ℓ-diversity leakage: subgroup distribution is lower than population



t-closeness restriction: subgroup distribution must be *t*-close to pop.

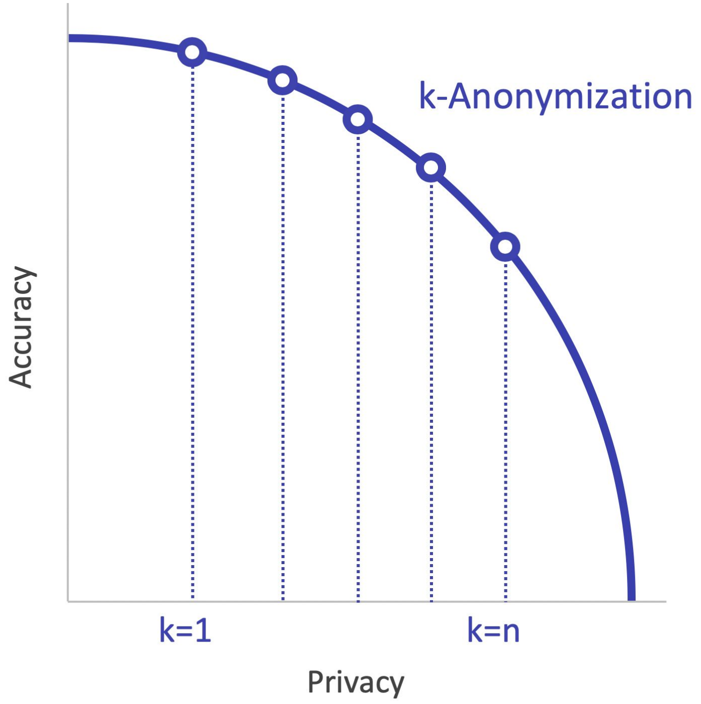
The rub: Once your data are *k*-anonymized and *t*-close… what's left?

When to implement *k*-anonymization

By limiting the resolution of quasi-identifiers, it promotes a “safety in groups” privacy and makes it **hard to link** records with certainty.

However, it is still possible to identify sensitive attributes that are homogeneous within a group. At high *k*, there may also be high utility loss.

Recommendation:
k-anonymization is a useful but manual way to address data linkage issues. *k*-anonymization may be especially useful for applications that do not require analytical utility, e.g. searchable databases.



Disclaimer: Chart is for illustration purposes only. The actual trade-off is coupled tightly to context.

Privacy-preserving approaches

Differential privacy

Motivating differential privacy

All privacy-preserving techniques diminish the utility of the dataset, but they have a second disadvantage: they lack a mathematical backbone.

Differential privacy asks:

How can I mathematically guarantee that an attacker is limited in their ability to make inferences about individual rows in the input?

Motivating differential privacy

Limit the attacker's ability to make confident inferences by injecting noise into the analysis!

Motivating differential privacy (another way)

An individual should be able to deny their participation in the database

D1

Name	Sex	Age	Secret
Pam Beasley	F	31	Y
Stanley Hudson	M	47	N
Erin Hannon	F	23	Y
Dwight Schrute	M	45	N
Andy Bernard	M	24	Y
Jim Halpert	M	28	N
Holly Flax	F	39	Y

By adding noise, we want to make it unclear if a result originated from **D1** or **D2**.

D2

Name	Sex	Age	Secret
Pam Beasley	F	31	Y
Stanley Hudson	M	47	N
Erin Hannon	F	23	Y
Dwight Schrute	M	45	N
Andy Bernard	M	24	Y
Jim Halpert	M	28	N

Does Holly have a secret? I know she's older...

Attacker

Methods of Differential Privacy

Many methods exist -- a differentially private average is shown here.

```
SELECT AVG(Age)
WHERE Sex = 'F' AND
Secret = 'Y'
```

Name	Sex	Age	Secret
Pam Beasley	F	31	Y
Erin Hannon	F	23	Y
Holly Flax	F	39	Y

DP Average

Mix noise into analysis

29.3 DP Result

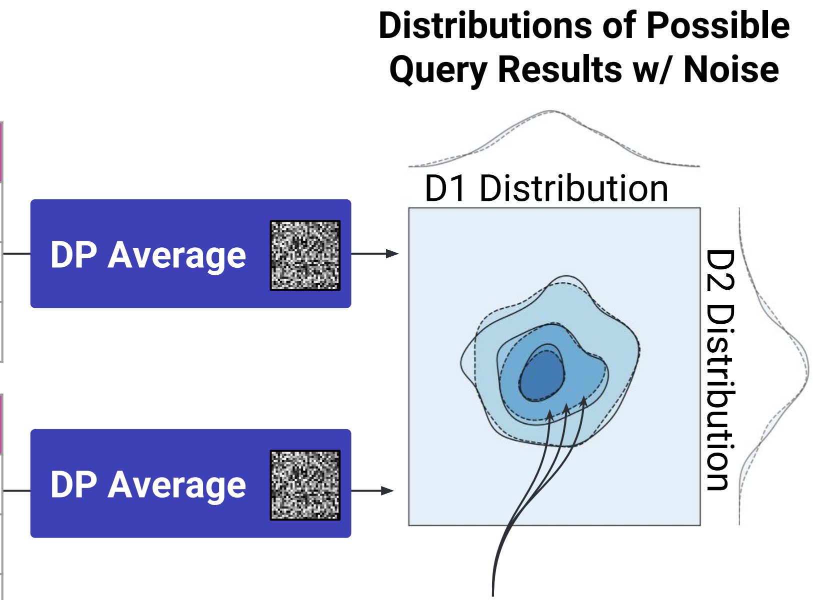
Quantifying likelihood of dataset origin

D1

Name	Sex	Age	Secret
Pam Beasley	F	31	Y
Erin Hannon	F	23	Y
Holly Flax	F	39	Y

D2

Name	Sex	Age	Secret
Pam Beasley	F	31	Y
Erin Hannon	F	23	Y



`SELECT AVG(Age) WHERE Sex='F' AND Secret='Y'`	31.2

Deep dive

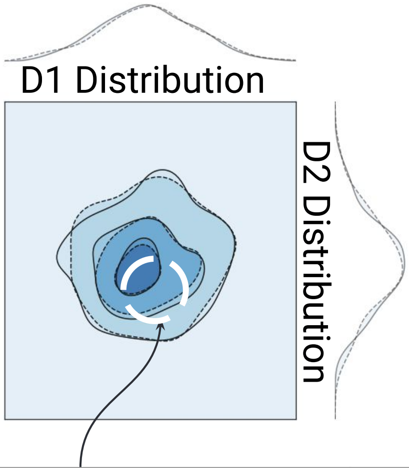
Quantifying likelihood of dataset origin

The attacker knows that Holly would bring the average age of the group above 31, but cannot tell if she contributed since the result could due to noise…

Can we reconcile? What's the relative likelihood of ending up in S, if database is **D1** vs **D2**?

$$\frac{\Pr[A(D_1) \in S]}{\Pr[A(D_2) \in S]}$$

Distributions of Possible Query Results w/ Noise



`SELECT AVG(Age) WHERE Sex='F' AND Secret='Y'`	28.7, 29.3, 31.2, …

Deep dive

Quantifying likelihood of dataset origin

Differential privacy makes the requirement that the likelihood ratio of a given result is bounded:

$$e^{-\varepsilon} \leq \frac{\Pr[A(D_1) \in S]}{\Pr[A(D_2) \in S]} \leq e^{\varepsilon}$$

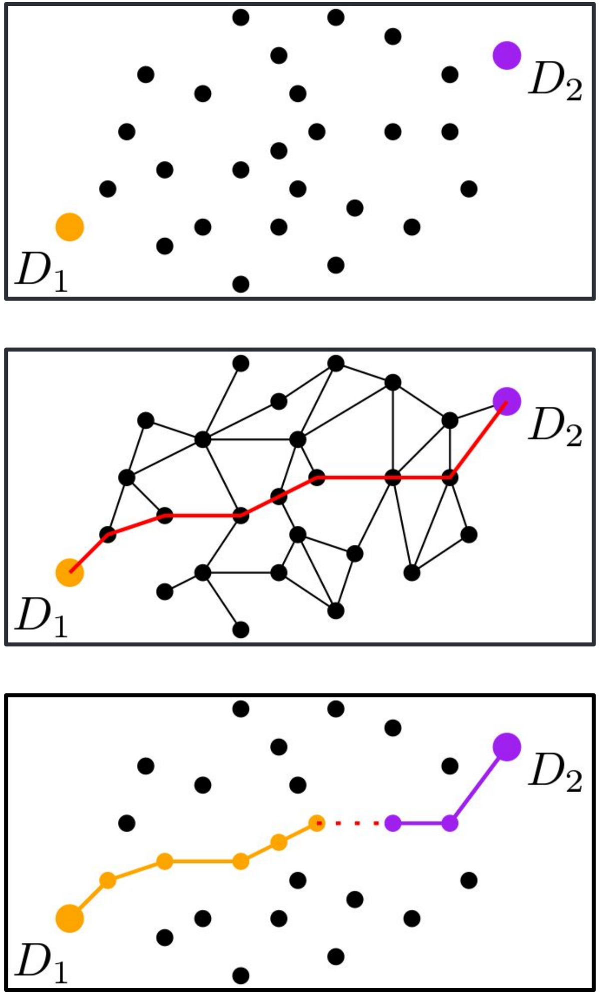
When **ε** is high, the likelihood ratio is somewhere in [0, ∞]; you often **CAN** tell the difference between some **D1** & **D2**.

When **ε** is small, the likelihood ratio is in [1-ε, 1+ε]; you **CANNOT** tell the difference between any **D1** & **D2** with high statistical confidence.

Why is randomization necessary?

To be useful, the analysis should:

- Not always return the same thing
- Universally guarantee privacy:
    - Should prevent the attacker from distinguishing all pairs of databases that differ by one record.



Implementation notes

Differential privacy

How do you build privacy mechanisms that have this property?

This depends both on the desired level of protection and the type of analysis you're performing.

A sampling of techniques with proven implementations:

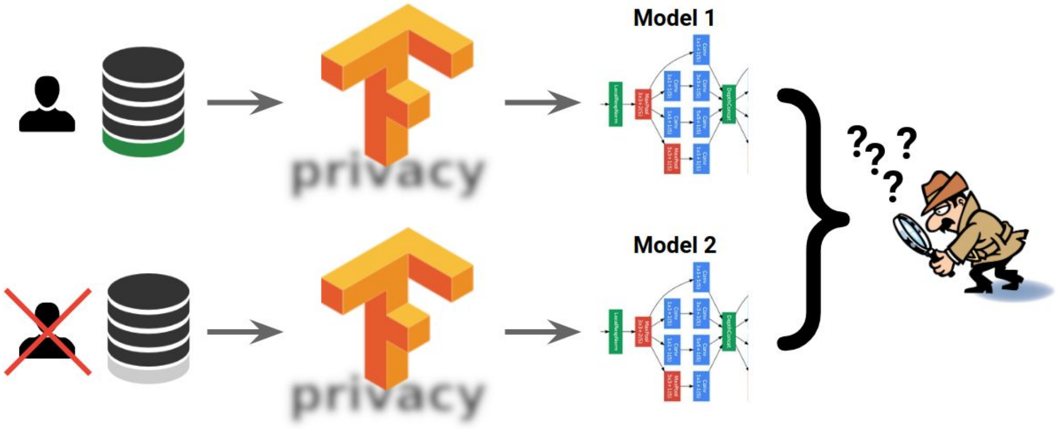
- Descriptive statistics
- Regression models
- Gaussian random projection for dimensionality reduction
- Minimum spanning trees
- Learning decision tree models
- Stochastic gradient descent with differentially private updates (for training models)

Implementation notes

Differentially private AI / ML

In principle, most analytical techniques are able to be made differentially private, such that *adjacent models* would not significantly differ.

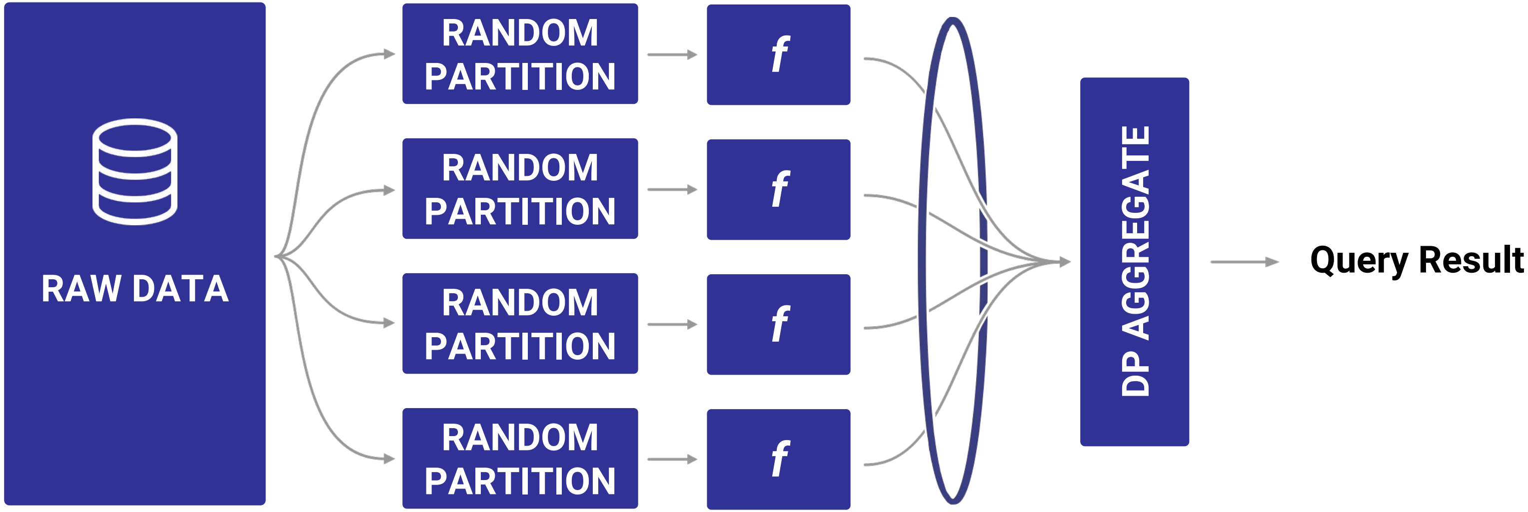
One benefit is a regularization effect on the data: by making the model immune to effects of individuals, they are more likely to generalize across subjects.



https://link.medium.com/T80XeIHfQU

Implementation notes

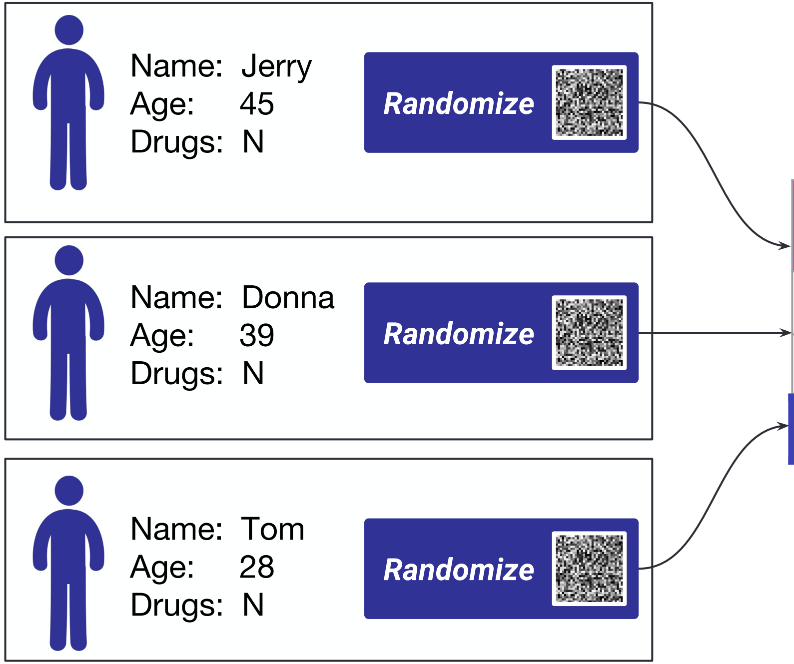
Sample and aggregate



Local differential privacy

Motivation:

Randomize record for plausible deniability of content (not participation)



Name	Age	Drug Use
Jerry	47	N
Donna	36	N
Tom	29	Y

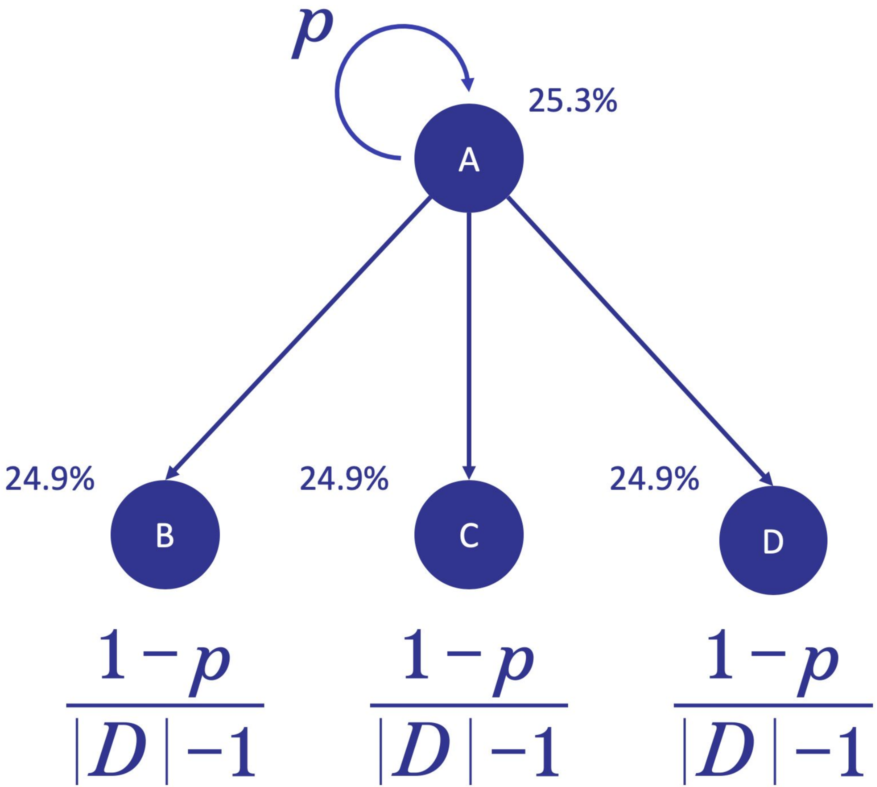
Implementation notes

Local differential privacy

We can use Randomized Response to "scramble" (mask) categorical values.

The result is local DP so long as the output only weakly depends on the input.

$$p = \frac{e^{\varepsilon}}{e^{\varepsilon} + |\mathcal{D}| - 1}$$



Attacking differential privacy

If an attacker is able to make unlimited differentially private queries, then they may be able to reconstruct properties of the underlying dataset.

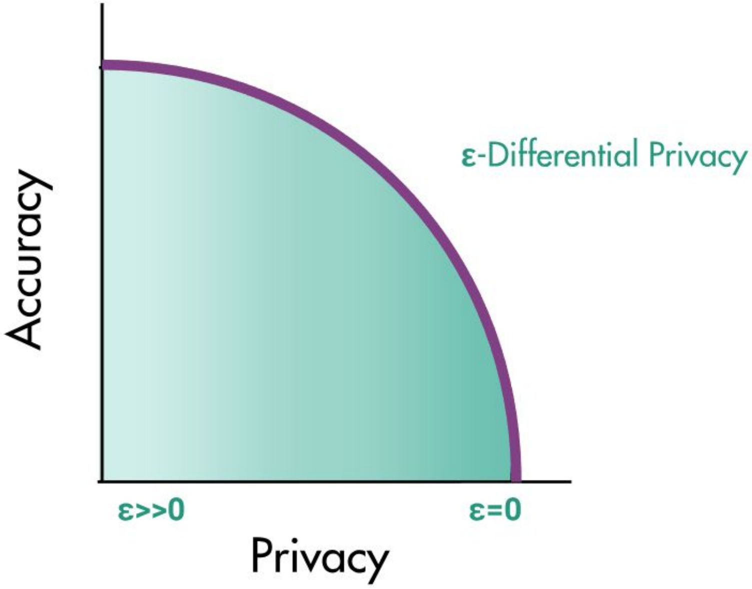
To combat this, many implementations employ a “privacy budget”, which restricts the number of queries that can be made against a dataset.

When to implement differential privacy

Differential privacy provides formal guarantees of “plausible deniability” for an individual’s data, at low levels of ε.

This requires adding (often significant) amounts of noise to the data, as well as a more customized and tightly monitored analytical process.

Recommendation:
Use differential privacy when data is extremely sensitive and aggregate analysis can survive the introduction of noise.



Disclaimer: Chart is for illustration purposes only. The actual trade-off is coupled tightly to context.

Privacy-preserving approaches

Privacy in practice

Review

Privacy-preserving techniques

We have covered three techniques today….

De-identification

Replace identifying or quasi-identifying attributes with substitute information, a la HIPAA

k-Anonymization

Suppress or generalize information in such a way that it can no longer be traced to an individual record

Differential Privacy

Formally limit the ability of an attacker to reason about analysis input from observing the output.

Review

Privacy-preserving concepts

De-identification
Obfuscate attributes

Dataset

f

Attributes

Records

id	name	birth_date	income
1056		01-01-1973	58,850.23
1057		01-01-1964	87,500.12
1058		01-01-1992	28,043.56
1059		01-01-1955	1,064.75
1060		01-01-1990	21,320.00
...	...	...	...

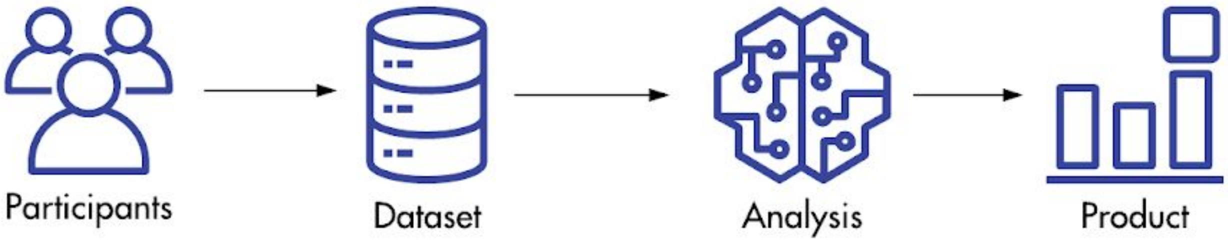
Differential privacy
Inject noise into analysis of dataset

***k*-Anonymization**
Reduce signal in records

Workflow

Data protection by design

Private information can leak throughout the data science workflow, and each step allows for different privacy-preserving techniques.

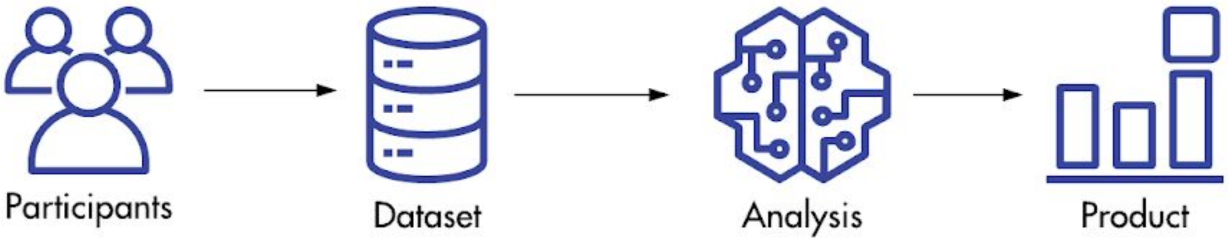


Workflow

Data protection by design

Privacy- Preserving Techniques

- De-identification
- k-Anonymization
- Differential Privacy

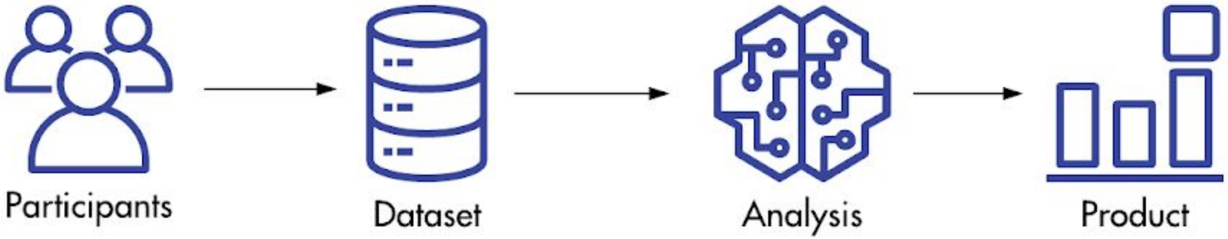


Workflow

Data protection by design

Privacy- Preserving Techniques

- De-identification
- k-Anonymization
- Differential Privacy

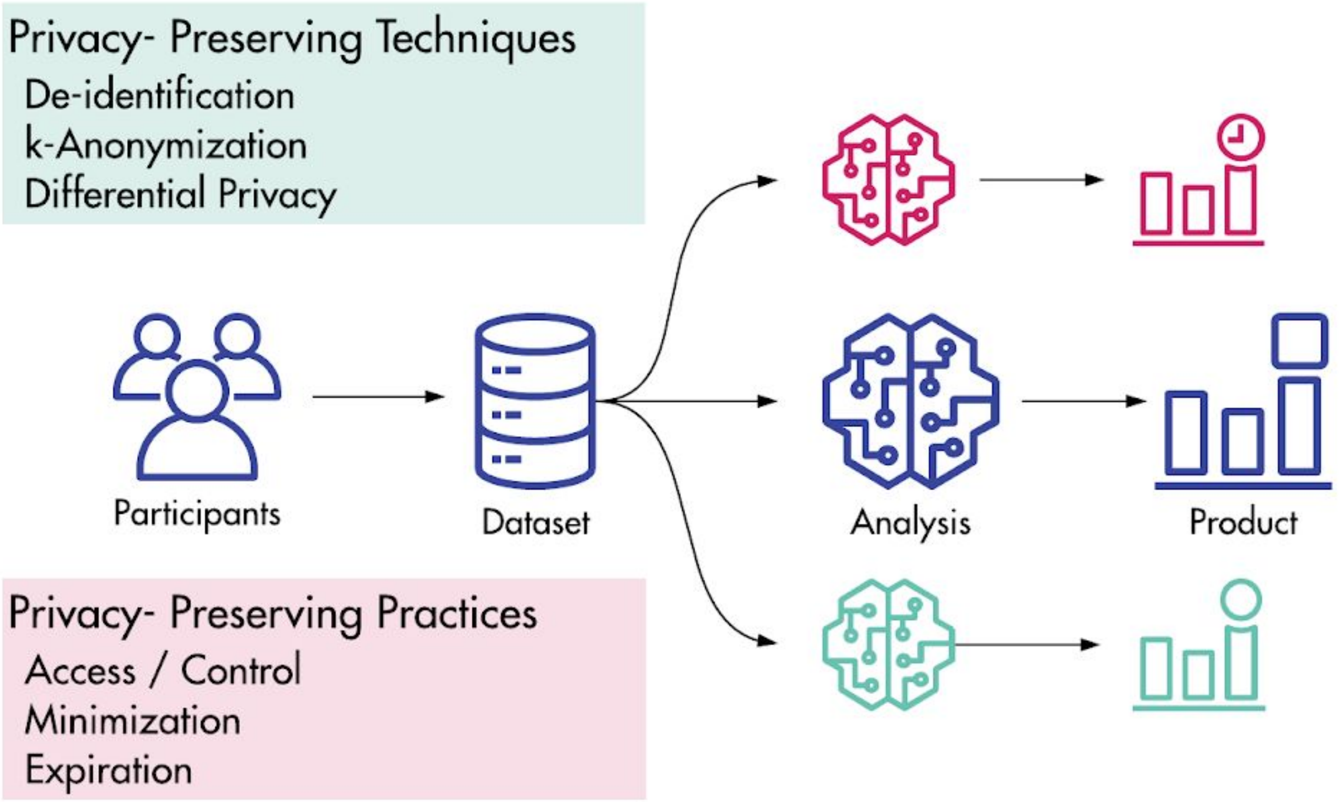


Privacy- Preserving Practices

- Access / Control
- Minimization
- Expiration

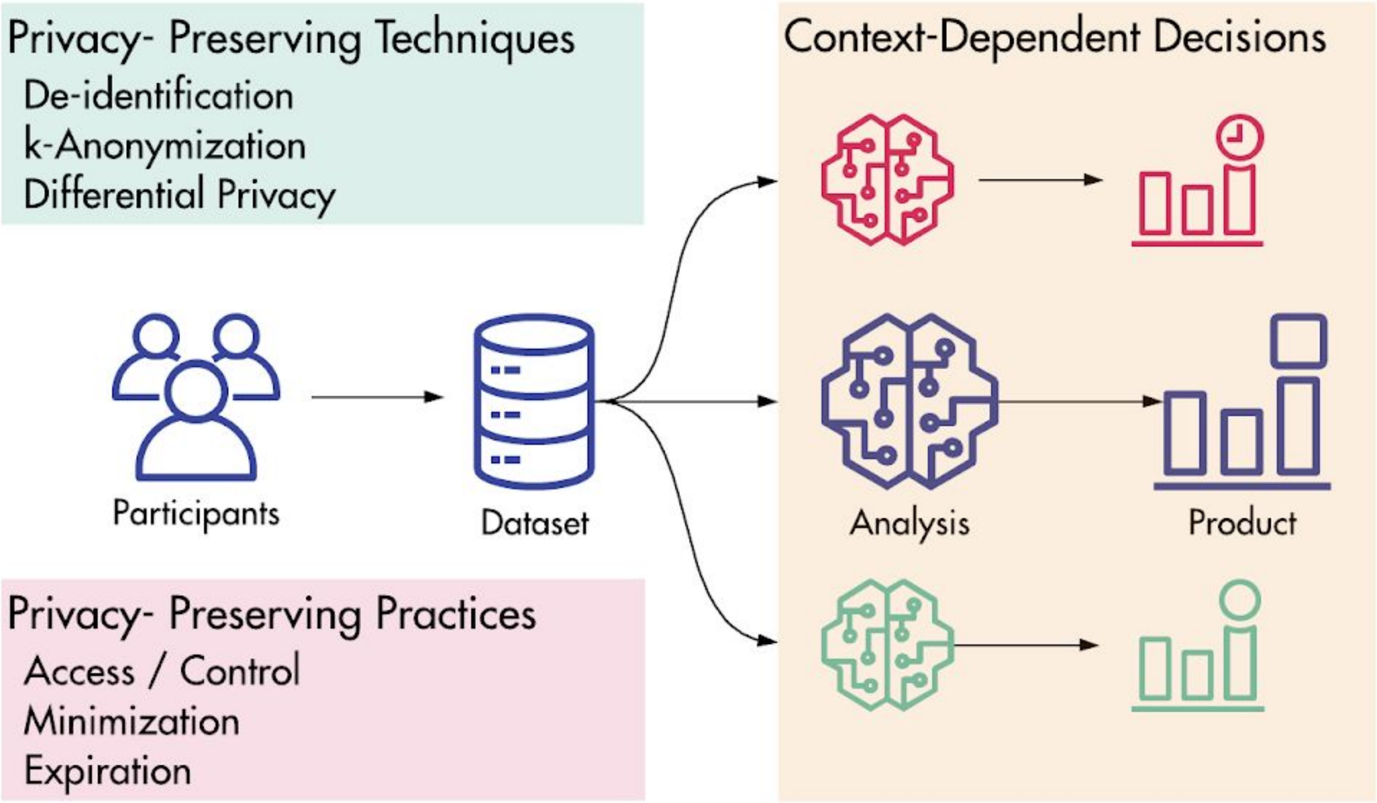
Workflow

Data protection by design



Workflow

Data protection by design



Data is the pollution problem of the information age, and protecting privacy is the environmental challenge.

— Bruce Schneier

IMMUTA

The Data Scientist's Guide to Preserving Privacy

Columbus Data Science MeetUp

Stephen Bailey*, Alfred Rossi*, Joe Regensburger

March 20, 2019