Developing Data Products

SC4125

Module 8

Data Governance

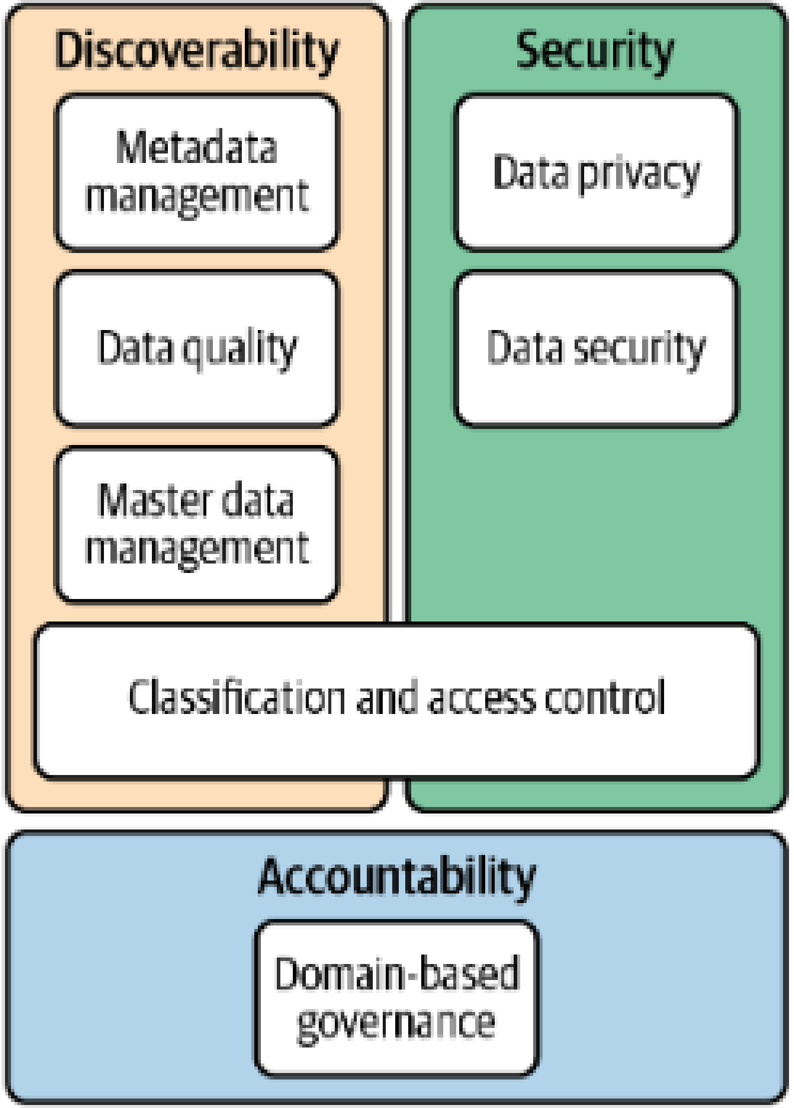
Teaching material

Accompanying teaching material

These slides are accompanied with a jupyter notebook and an html deck of slides.

The big picture

- **Data governance** adds to stakeholders' trust in the data —specifically, in how that data is collected, analyzed, published, or used.
- **Discoverability:** Make technical metadata, lineage information, and a business glossary readily available. Business critical data needs to be correct and complete. Master data management to guarantee that data is finely classified ensuring appropriate protection against inadvertent or malicious changes or leakage.
- **Security:** Depending on the nature of business and data - regulatory compliance, management of sensitive data (e.g., personally identifiable information, business intelligence and assets), data security and exfiltration prevention.
- **Accountability:** Provide an operating model for ownership and accountability around boundaries of data domains.



Source: Data Governance - The Definitive Guide by Eryurek et al

Example

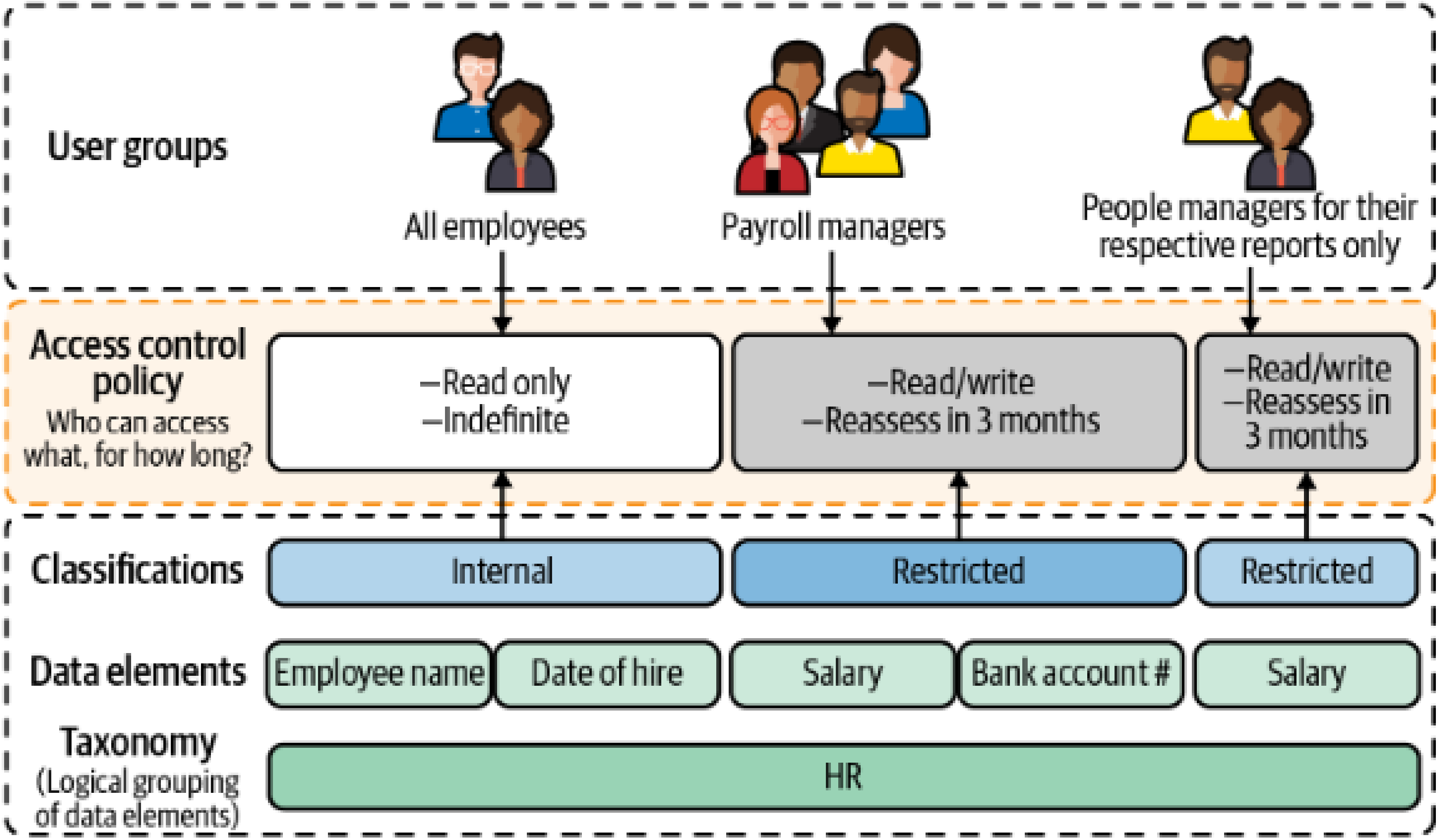


Image source: Data Governance - The Definitive Guide by Eryurek et al

A few data governance considerations

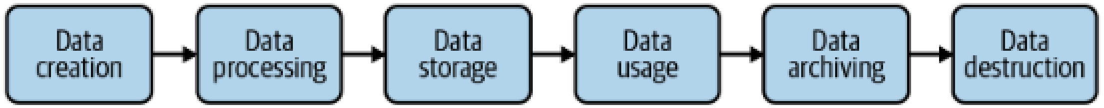
Access control

Need to know: Mandatory + discretionary access control. Per-use-case policies.

Defense in depth: Apply multiple layers of defense.

Compliance

Regulatory obligations



Store data in narrow slices

Don't store all data together, and instead segregate by purpose. Data lifecycle (retention/deletion).

Backup

Periodically check restoration capability

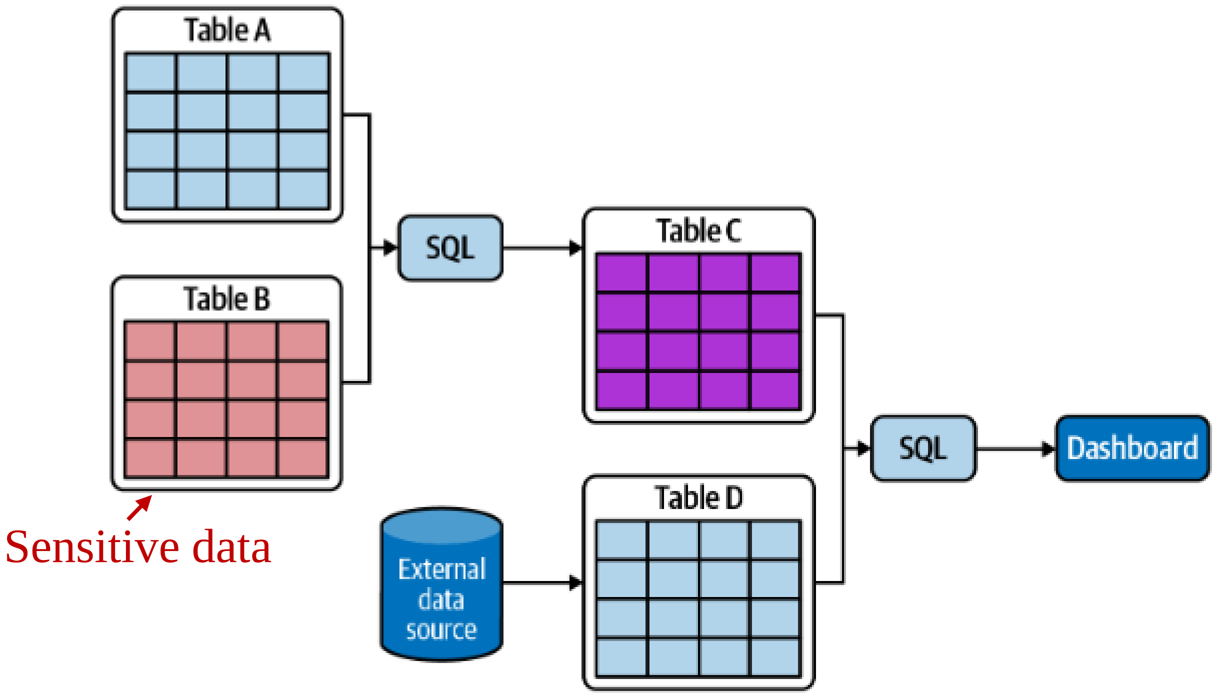
Audit

Log access and changes, carry out periodic checks (also of the infra)

Version & quality control

Lineage/provenance, Meta-data management, Cataloguing

Lineage workflow



Use cases:

Quality control: How do I know if the data used is trusted, and not coming from less trusted systems without manual oversight?

Audit: Show me all the sensitive data within our data warehouse, and what systems use sensitive data?

Compliance: I need to report and audit all systems that process PII.

Source: Data Governance - The Definitive Guide by Eryurek et al

Data governance as a data scientist

- Several issues of immediate concern:
  - Lineage & provenance
  - Version and data quality control
  - Enabling/enhancing privacy for analytics

Privacy models

- Pseudonymization
- k-anonymity (and friends)
- Differential privacy

Pseudonymization

Original Purchase Table

Shop	User ID	Time	Price	Price Bin
	7abc1a23	09/23	$97.30	$49 - $146
	7abc1a23	09/23	$15.13	$5 - $16
	3092fc10	09/23	$43.78	$16 - $49
	7abc1a23	09/23	$4.33	$2 - $5
	4c7af72a	09/23	$12.29	$5 - $16
	89c0829c	09/24	$3.66	$2 - $5
	7abc1a23	09/24	$35.81	$16 - $49

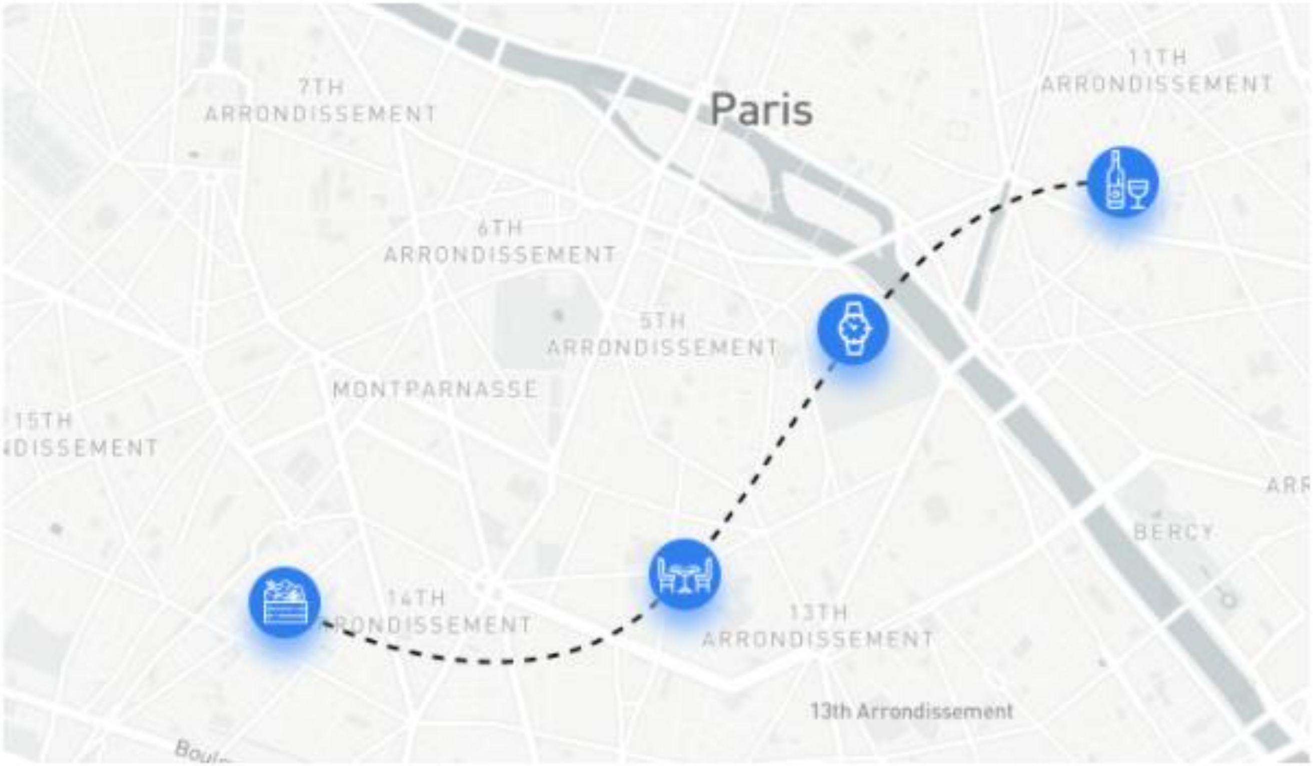
Source: https://mosaiceffect.com/

Pseudonymization

Mosaic effect: different factors can be used in conjunction with one another to determine if an individual is identifiable.

Original Purchase Table

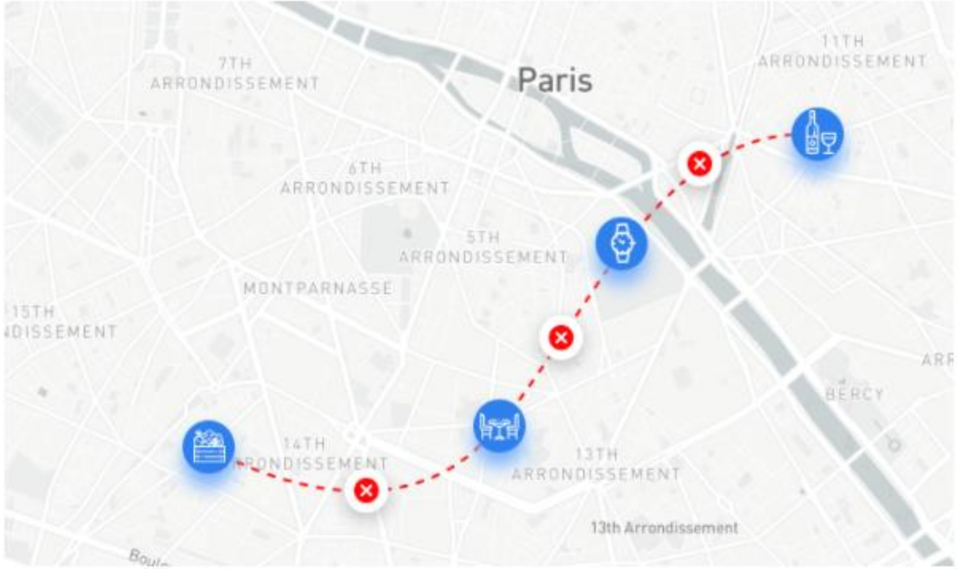
Shop	User ID	Time	Price	Price Bin
	7abc1a23	09/23	$97.30	$49 - $146
	7abc1a23	09/23	$15.13	$5 - $16
	3092fc10	09/23	$43.78	$16 - $49
	7abc1a23	09/23	$4.33	$2 - $5
	4c7af72a	09/23	$12.29	$5 - $16
	89c0829c	09/24	$3.66	$2 - $5
	7abc1a23	09/24	$35.81	$16 - $49



Source: https://mosaiceffect.com/

Pseudonymization

GDPR compliant Pseudonymisation requires that personal data must be transformed so that the identity of individuals cannot be discovered by linkage attacks. To achieve GDPR compliant Pseudonymisation, the practice of tokenization can be expanded to use dynamically-generated tokens applied to both direct and indirect identifiers. data about the Pseudonym used to obscure the activities of User ID “7abc1a23” is retained, but it is made available only to authorised parties under controlled conditions - it is not revealed to the outside world.



Pseudonymised Purchase Table

Shop	User ID	Time	Price	Price Bin	Pseudonymised
	67c0Gt11	09/23	$97.30	$49 - $146	✓
	54&#3216	09/23	$15.13	$5 - $16	✓
	3092fc10	09/23	$43.78	$16 - $49	
	DeTym321	09/23	$4.33	$2 - $5	✓
	4c7af72a	09/23	$12.29	$5 - $16	
	89c0829c	09/24	$3.66	$2 - $5	
	HHyargLM	09/24	$35.81	$16 - $49	✓

Additional Information

Time	Pseudonym	User ID
09/23	67c0Gt11	7abc1a23
09/23	54&#3216	7abc1a23
09/23	DeTym321	7abc1a23
09/24	HHyargLM	7abc1a23

Source: https://mosaiceffect.com/

Pseudonymization

- From a purely technical point of view, Pseudonymization, even with GDPR’s stringency, is nevertheless wishful thinking in terms of privacy.
  - e.g., in case the “additional information” table is revealed through a breach!
    - with the definition of differential privacy, we shall see more on the implication of even a single record

k-anonymity

Publishing microdata

- Identifier(s): Attribute(s) in data record that uniquely identifies an individual in a population.
- Quasi-identifier(s): set of non-sensitive attributes, which, if linked with external data may uniquely identify at least one individual in the population
- Sensitive attributes

Voter Registration Data

Name	Age	Sex	Zipcode
Ahmed	25	Male	53711
Brooke	28	Female	55410
Casey	31	Female	90210
Dave	19	Male	02174
Evelyn	40	Female	02237

Patient Data

Age	Sex	Zipcode	Disease
25	Male	53711	Flu
25	Female	53712	Hepatitis
26	Male	53711	Brochitis
27	Male	53710	Broken Arm
27	Female	53712	AIDS
28	Male	53711	Hang Nail

Example from Mondrian Multidimensional K-Anonymity by LeFevre et al

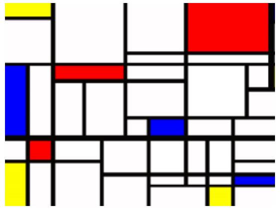
k-anonymity

Table T is k-anonymous with respect to attributes X1, ..., Xd if every unique tuple (x1, ..., xd) in the (multiset) projection of T on X1, ..., Xd occurs at least k times (forming equivalence classes).

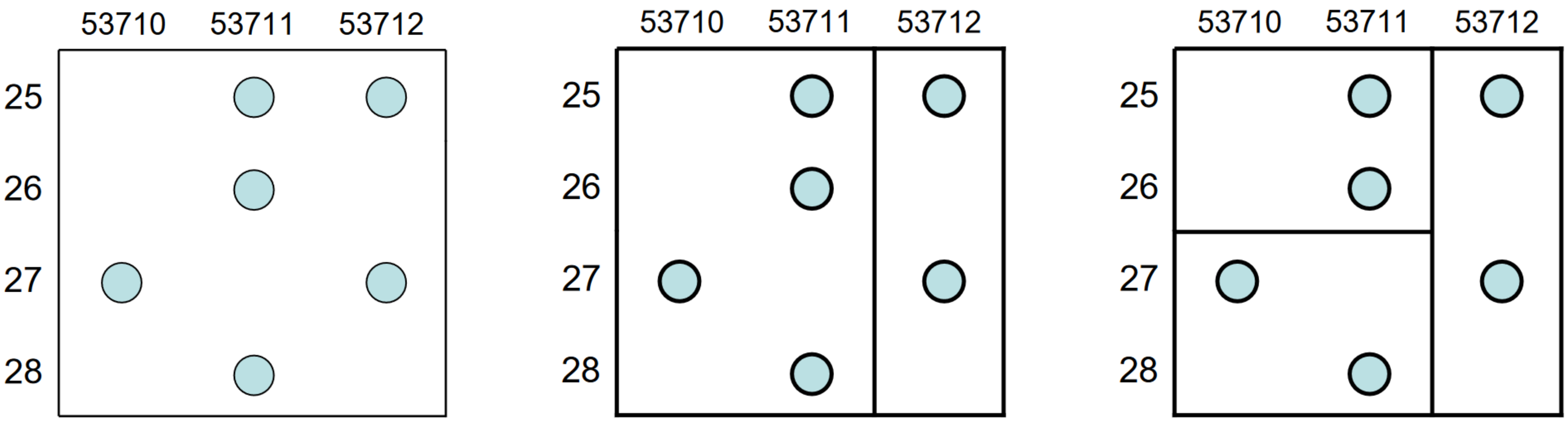
- Identifier(s): Attribute(s) in data record that uniquely identifies an individual in a population.
- Quasi-identifier(s): set of non-sensitive attributes, which, if linked with external data may uniquely identify at least one individual in the population
- Sensitive attributes

Age	Sex	Zipcode	Disease
[25-26]	Male	53711	Flu
[25-27]	Female	53712	Hepatitis
[25-26]	Male	53711	Brochitis
[27-28]	Male	[53710-53711]	Broken Arm
[25-27]	Female	53712	AIDS
[27-28]	Male	[53710-53711]	Hang Nail

k-anonymity: The Mondrian greedy partitioning algorithm



Named after
Piet Mondrian



Example from Mondrian Multidimensional k-Anonymity by LeFevre et al

l-diversity

- Homogeneity Attack
- Background knowledge attack

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	$*$	Heart Disease
2	130**	< 30	$*$	Heart Disease
3	130**	< 30	$*$	Viral Infection
4	130**	< 30	$*$	Viral Infection
5	1485*	≥ 40	$*$	Cancer
6	1485*	≥ 40	$*$	Heart Disease
7	1485*	≥ 40	$*$	Viral Infection
8	1485*	≥ 40	$*$	Viral Infection
9	130**	$3*$	$*$	Cancer
10	130**	$3*$	$*$	Cancer
11	130**	$3*$	$*$	Cancer
12	130**	$3*$	$*$	Cancer

Example from ℓ-Diversity: Privacy Beyond k-Anonymity by Machanavajjhala et al

Progenies of privacy models …

l-diversity is neither necessary nor sufficient to prevent attribute disclosure.

There are many other variations, e.g., t-closeness, m-invariance, δ-disclosure, etc.

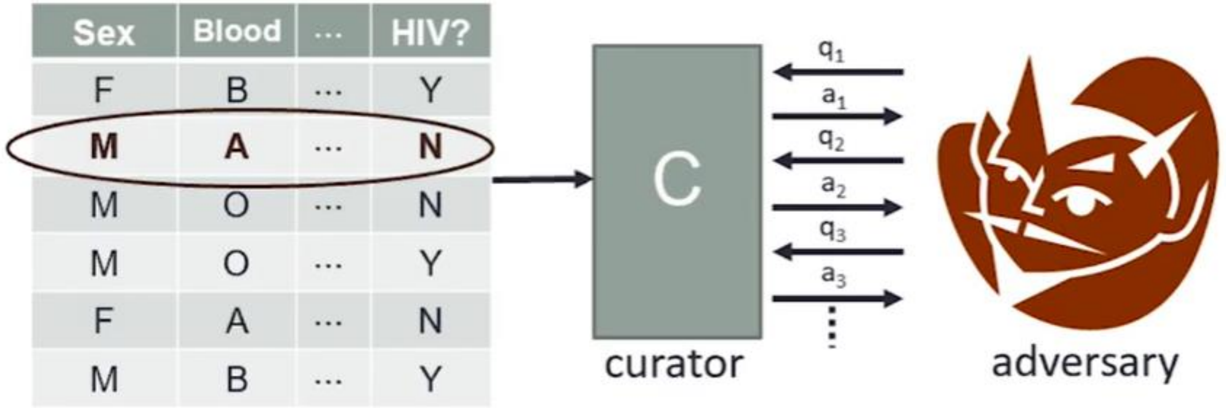
Still no perfect or quantifiable privacy guarantee!

A different perspective on privacy

Publishing statistics, or supporting an interactive statistical database

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

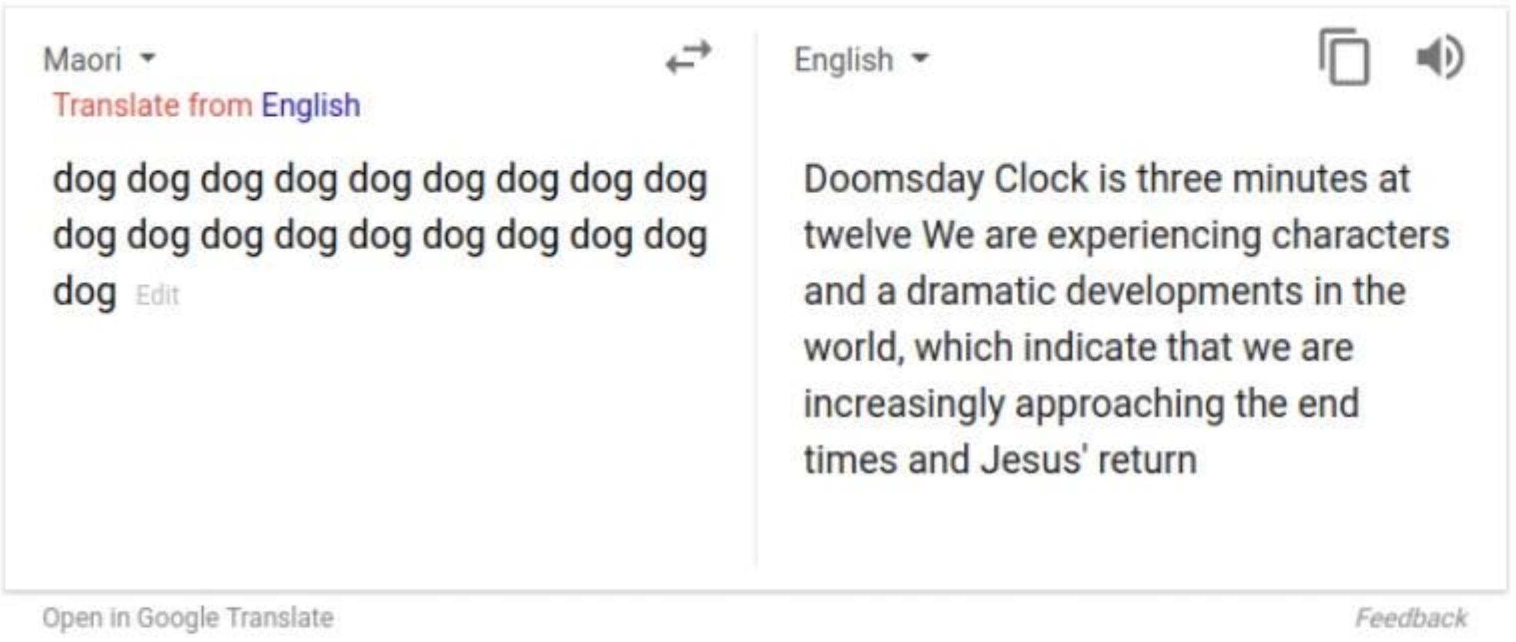
		AGE		
STATISTIC	**GROUP**	**COUNT**	**MEDIAN**	**MEAN**
1A	total population	7	30	38
2A	female	4	30	33.5
2B	male	3	30	44
2C	black or African American	4	51	48.5
2D	white	3	24	24
3A	single adults	(D)	(D)	(D)
3B	married adults	4	51	54
4A	black or African American female	3	36	36.7
4B	black or African American 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)



Sources: Table - Understanding Database Reconstruction Attacks on Public Data by Garfinkel et al
Figure - The Science Behind WhiteNoise: Differential Privacy talk by S. Vadhan

Models and memories

Example: The neural network used for training may inadvertently spill out actual data instances it was trained with!

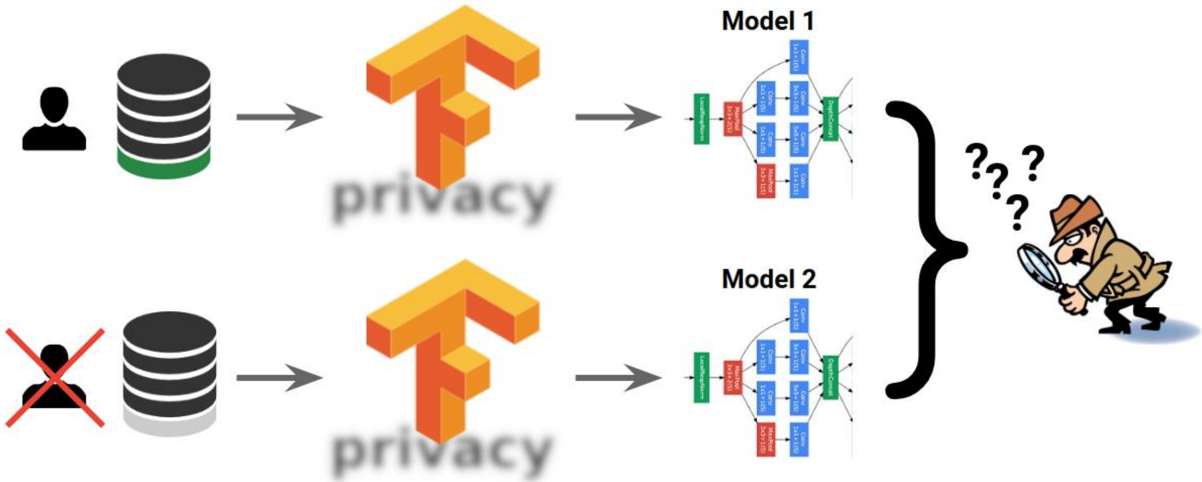


Cue: dyslexic, agnostic, insomniac joke?

Sources: Vice.com

Differential privacy

Publishing statistics, or supporting an interactive model



- Can't learn anything new about a specific individual in the DB?
  What about learning something generic which is still true about the individual? E.g., Salary range for SCSE grads.

- This is not to be considered as a privacy compromise (definitional convention/choice): If we can learn the same thing about the individual, even if the specific individual were to be replaced by another random member of the population.

- Disentangle learning about the population as a whole versus learning about an individual.

Cynthia Dwork's introductory talk on The definition of differential privacy
Image source: TensorFlow Blog

Differential privacy

Publishing statistics, or supporting an interactive model

The outcome of an analysis is (almost) the same, irrespective of whether an individual is included or not included in the dataset.

- Can't learn anything new about a specific individual in the DB?
  What about learning something generic which is still true about the individual? E.g., Salary range for SCSE grads.
- This is not to be considered as a privacy compromise (definitional convention/choice): If we can learn the same thing about the individual, even if the specific individual were to be replaced by another random member of the population.
- Disentangle learning about the population as a whole versus learning about an individual.

Cynthia Dwork's introductory talk on The definition of differential privacy

Differential privacy

Formal definition

A randomized mechanism $M: D \rightarrow R$ satisfies (ε, δ)-differential privacy if for any two adjacent datasets $X, X' \in D$ and for any measurable subset of outputs $Y \subseteq R$ it holds that*:

$$\Pr[M(X) \in Y] \leq e^{\varepsilon} \Pr[M(X') \in Y] + \delta$$

* With the term δ, a weaker form of differential privacy is achieved (than without it). The original definition did not have this term.

- Adjacent dataset: Datasets which are different only by presence/absence of one sample of data.

  Variation: An individual data sample is replaced by another individual sample (2ε), the entire set of samples from one user is present/absent.

* Definition as used in TensorFlow privacy technical paper.

Randomized response

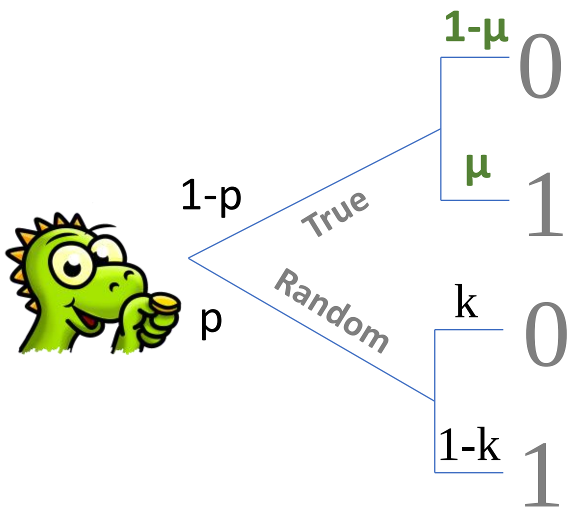
Mechanism predates the advent of the notion of differential privacy and associated formal treatment, and provides **refutability**

Local differential privacy is a model with the added restriction that even if an adversary has access to the personal responses of an individual in the database, that adversary will still be unable to learn too much about the user's personal data, in contrast to **global differential privacy** that incorporates a **central aggregator with access to the raw data**.

Text adapted from Wikipedia

Do you have attribute A? answer is **boolean 0/1**

Consider that A actually happens, i.e., 1 with a **probability μ**



$\Pr(1) = (1-p)\,\mu + p(1-k)$

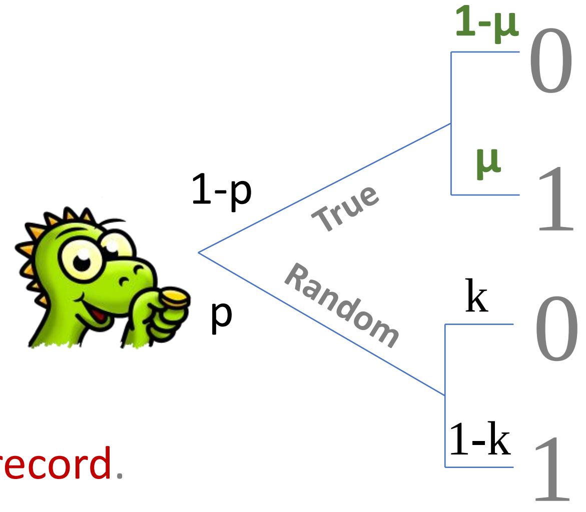
If there are N_1 yes responses out of N responses, then we can estimate **μ'**

$$\mu' = \frac{\frac{N_1}{N} - p(1-k)}{1-p}$$

Estimating ϵ

Consider X and X' as our db before/after adding data for a specific individual.

Do you have attribute X?



Consider query **K()** returns the count of **yes** in the DB.
Assume that we know the response of this query from before adding the new record.

- If the **actual answer is 0**, the probability that the value of **K()** is unchanged after adding the data is:
  1-p + pk
- If the **actual answer is 1**, the probability that the value of **K()** is unchanged after adding the data is:
  pk

We can establish **two inequalities**:

$$1 \leq e^{\epsilon} \cdot p \cdot k$$
$$1 \leq e^{\epsilon} \cdot (1 - \mathrm{p} + p \cdot k)$$

We can then determine, based on an **upper bound** for ϵ:

$$\epsilon = -ln(p \cdot k)$$

Note: One may want to support other data – non-binary, non-tabular – and carry out other kinds of operations than just counting. The nature of noise introduced, and the data/computational primitives that can be constructed over such "noisy data" varies. While some versatile approaches exist, it is still an active area of research, with several open challenges.

Some other approaches and tools for privacy-enhancing analytics

- Federated learning
  See more on [Federated learning with Tensor Flow]
- Secure multiparty computation
- Homomorphic computation primitives

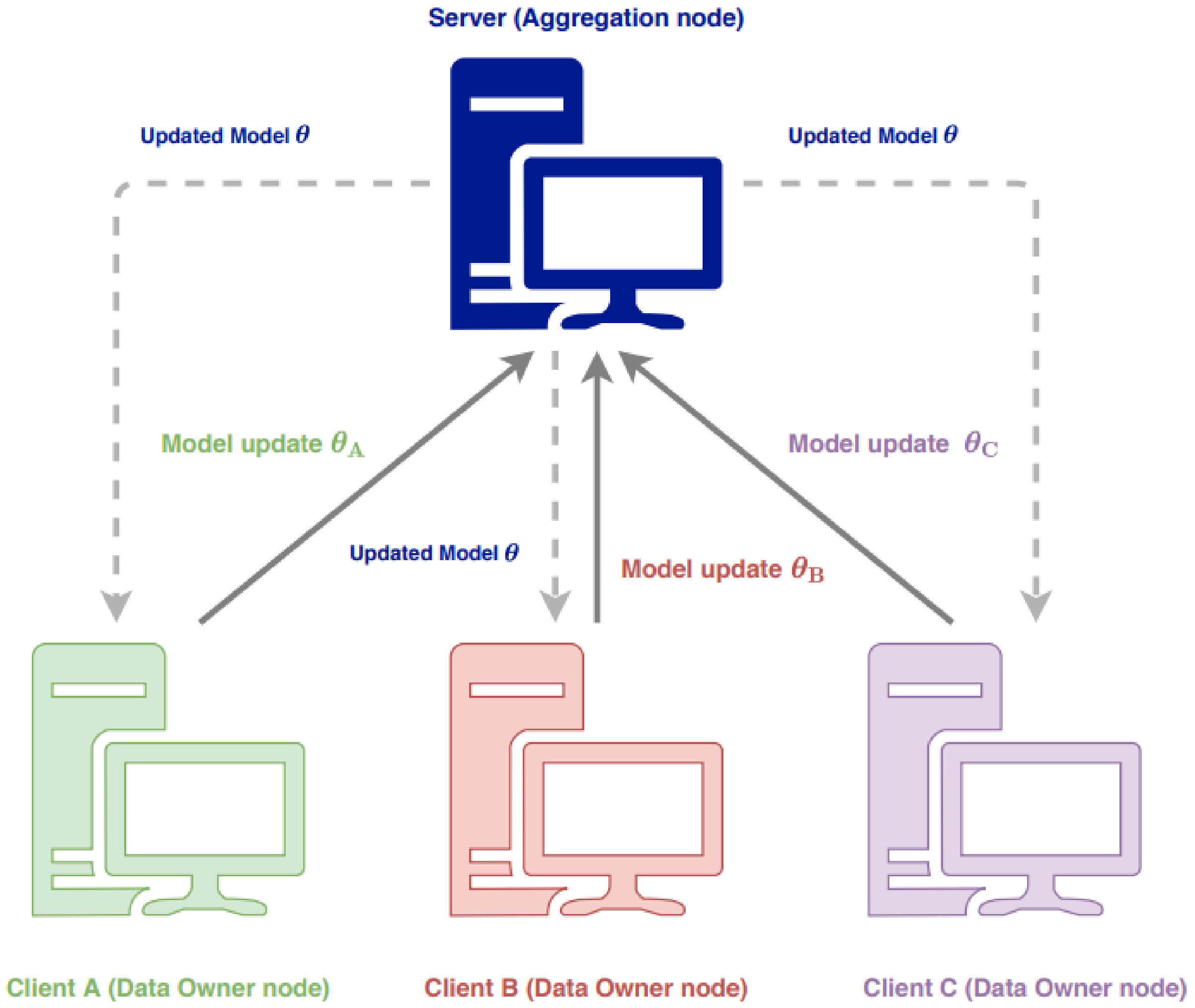


Image source: [Federated learning and Differential Privacy] article

Further practical considerations

- **Semantic relations among features,** e.g., ZIP codes are geographically clustered
  - Their roles in terms of utility
  - Information dependency/redundancy and impact on privacy

- **Use-case driven**

  \- Utility of specific features, e.g., adapt the heuristics to determine k-anonymous groups

- **Record suppression**

  \- It may sometimes be beneficial (in terms of utility) to remove some records altogether, rather than try to include every record and still form k-anonymous equivalence groups

This deck of slides is accompanied with Jupyter notebook for hands-on activities.

The notebook code/examples are based on/adapted from Kiprotect's tutorial on Data Privacy for Data Scientists by Andreas Dewes and Katherine Jarmul.

A few general purpose tools and libraries

for differential privacy and federated learning

https://github.com/opendp/

https://blog.tensorflow.org/2019/03/introducing-tensorflow-privacy-learning.html

https://www.tensorflow.org/federated/federated_learning

Disclaimer: There are capability limitations in terms of the kind of computations (and dependent analysis/algorithms) that can currently be supported.


"That's all Folks!"