Anonymization

Dr. Balázs Pejó

www.crysys.hu

Agenda

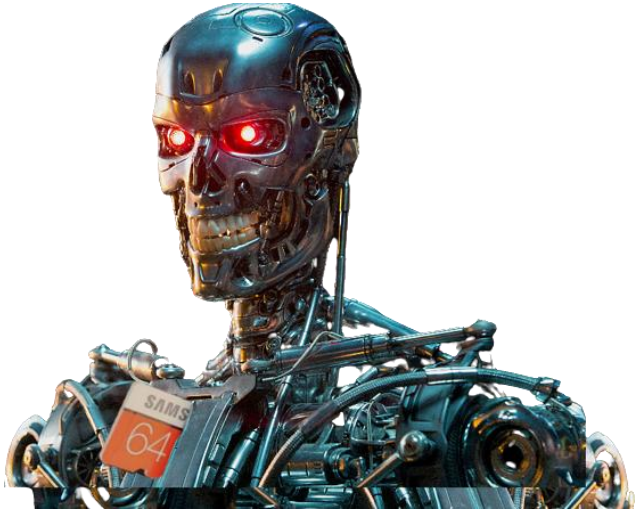
- Dark Patterns
- Tracking
- GDPR
- Deidentification
- Machine Learning
- Anonymization
- Cryptography

- Naïve Solutions
  - Attribute Types
  - Anonymization Primitives
- K-Anonymity
  - Methods
- L-Diversity & T-Closeness
- K-Map & D-Presence
- Tools
- Synthetic Data
  - Generative Models

Recap

Recap (ML)

- Machine Learning has its own security and privacy vulnerabilities.
  - Adversarial Examples
  - Membership Inference Attack

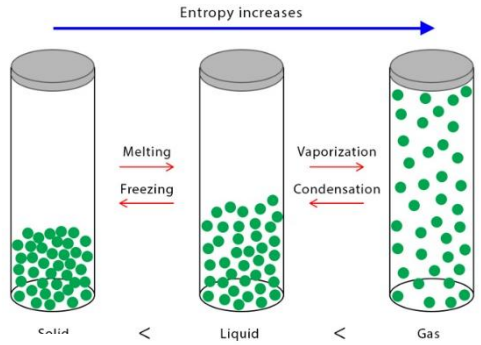


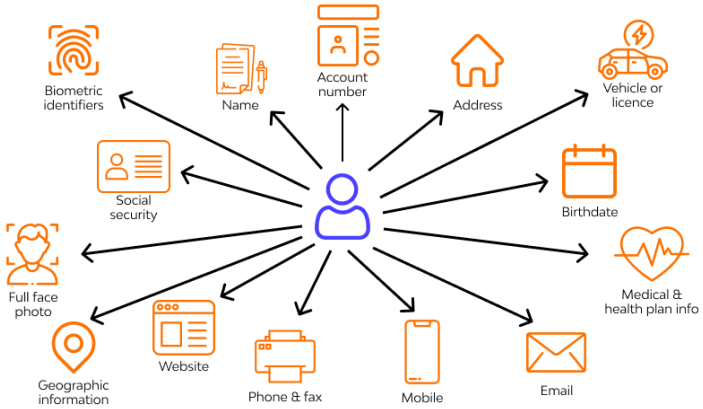
- Machine Learning can be used to deanonymize datasets.

- Machine Learning can be used for anonymization as well.
  - Synthetic Data

Recap (DeID)

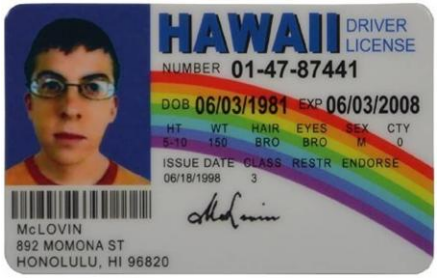
- GDPR: if a de-anonymization attack is plausible and reasonably successful, the data is personal (even if it is encrypted).
- Entropy decreases fast due to the curse of dimensionality.
  - More information decreases the anonymity set.
- Too many too accurately statistics could expose the entire database.
- Query auditing can be on-line (prevention) or off-line
- Personally Identifiable Information (PII) are often not necessary at all for identification.





Attribute Types

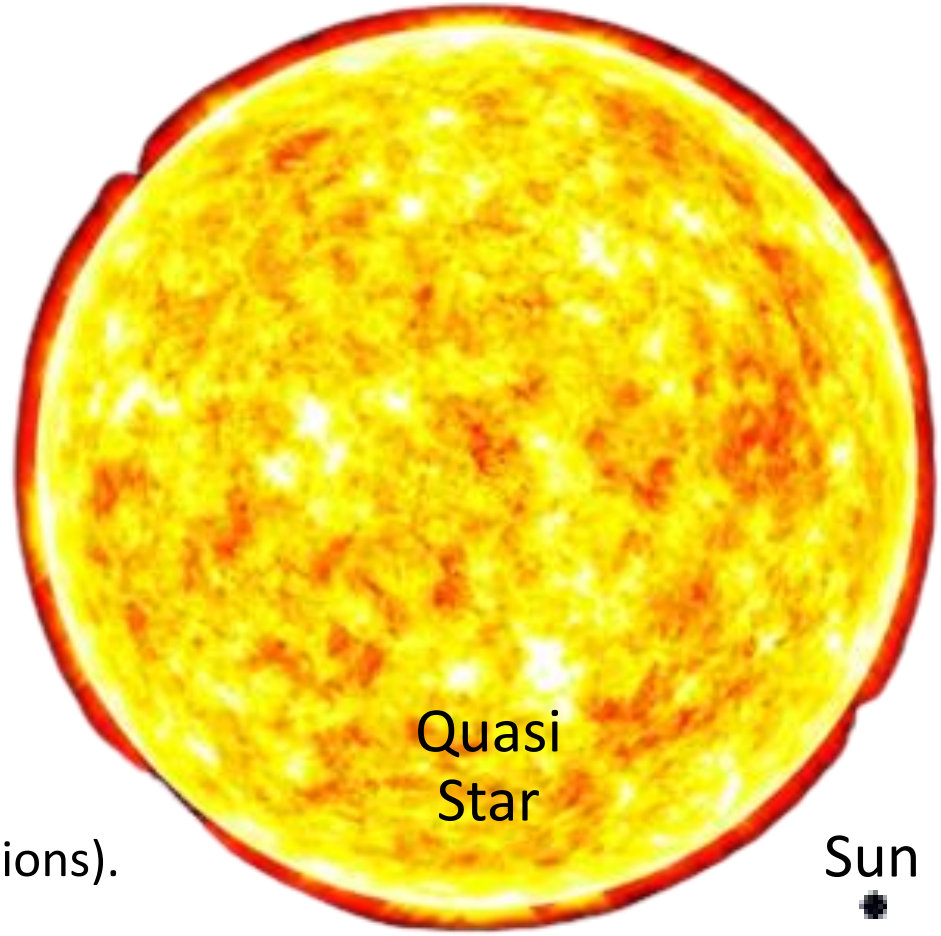
- In terms of privacy, four type of attributes can be distinguished.



- Identifying: attribute which can uniquely identify a person.
  - E.g., name, phone number, social number, etc.
- Quasi Identifying: attribute that can uniquely identify a person if combined with other QI attributes.
  - E.g., sex, profession, age, etc.
- Sensitive: attributes including private information and must be kept private, but necessary for the analysis.
  - E.g., salary, diagnosis, etc.
- Insensitive: attributes that can be made public.
  - Depending on the context, almost anything can be a QI.

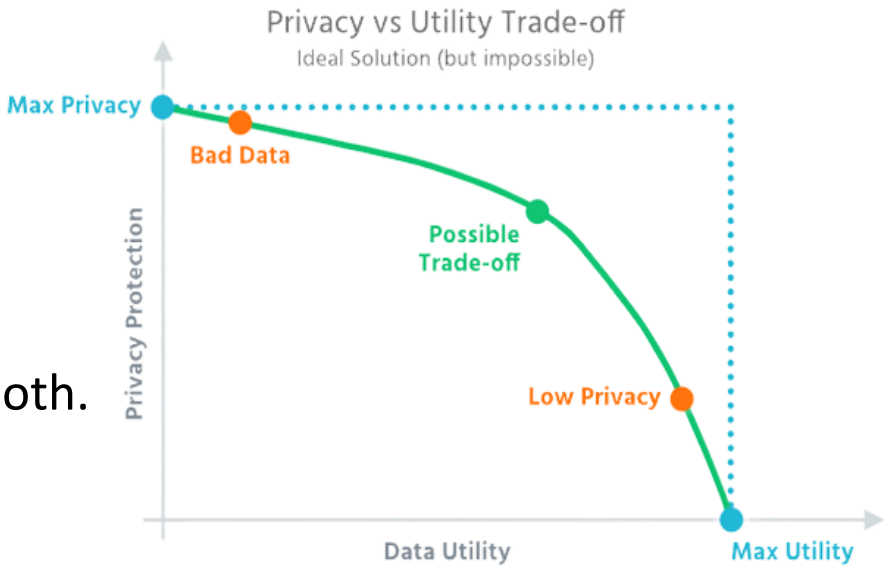
Quasi Identifiers

- Information which is not necessarily sensitive by themselves, but they might be used in a reidentification attack.
- There is no universal list of quasi-identifiers, it always depends on the attack model.
  - Some data types are almost always QIs (ZIP code, age, gender).
  - Other depend on the context (like timestamps, medical conditions).



Deanonimization

- The past decades have shown us that perfect and usable anonymization does not exist.
  - Data can be either useful or perfectly anonymous but never both.



- No anonymization procedure can guarantee privacy.
  - German Federal Data Protection Act (1974) defined anonymization as: *[…] alteration of personal data […] so that the individual information about personal or factual circumstances can […] only be attributed with a disproportionately large expenditure of time, cost and labor.*
- Factual anonymization requires a disproportionately large effort on the part of an attacker.
  - Disproportionality was not defined precisely, i.e., it is a legal grey area.

Anonymization

- Depending on the jurisdiction, different standards are applied.
- In Europe, the Article 29 Working Party's Opinion 05/14 says: Data is anonymized when three things are impossible:
  - Singling out: the possibility to isolate some records in the dataset.
  - Linkability: linking of data points of an individual to create a larger profile.
  - Inference: the ability to deduce one attribute from another attributes.

- Data are anonymised if all identifying elements (aka quasi-identifiers) have been eliminated.
  - No element may be left in the information which could, by exercising reasonable effort, serve to re-identify the person(s) concerned.
- If the data have been successfully anonymized, they are no longer personal data (i.e., can be shared, etc.).

Naïve solutions

- Pseudonymization : randomize identifiers
  - New York taxi database: randomization was done in a naive way reverse-engineering the license plates from pseudonyms was possible.
- Masking : remove identifiers altogether

Name	Date of birth	ZIP code	Visit date	Visit reason
(REDACTED)	1987-04-02	14217	2022-03-30	Flu

  - Quasi identifying attributes remains in the data: 87% of the US population could potentially be identified by their ZIP code combined with gender and date of birth.

Researchers spotlight the lie of 'anonymous' data

TechCrunch

The Guardian

New York taxi details can be extracted from anonymised data, researchers say

Rule-based techniques

- Obfuscation : adding some random perturbation to individual values.
- Generalization: making some attributes less granular.
- Suppression: remove records with rare values.

Name	Date of birth	ZIP code	Visit date	Visit reason
(REDACTED)	1987-04-**17**	14***	2022-03-30	Flu

- Even privacy notions from the scientific literature can fail to protect sensitive data.
- Aggregate the data
  - Reconstruction Attack: it is often possible to combine multiple statistics and retrieve individual records.

Visit date	Visit reason	Count	Average age
2022-03-30	Flu	5	32
2022-04-17	Broken arm	1	52
2022-06-24	Check-up	13	37
...	...	...	...


RESEARCH ARTICLE | COMPUTER SCIENCES

Confidence-ranked reconstruction of census microdata from published statistics

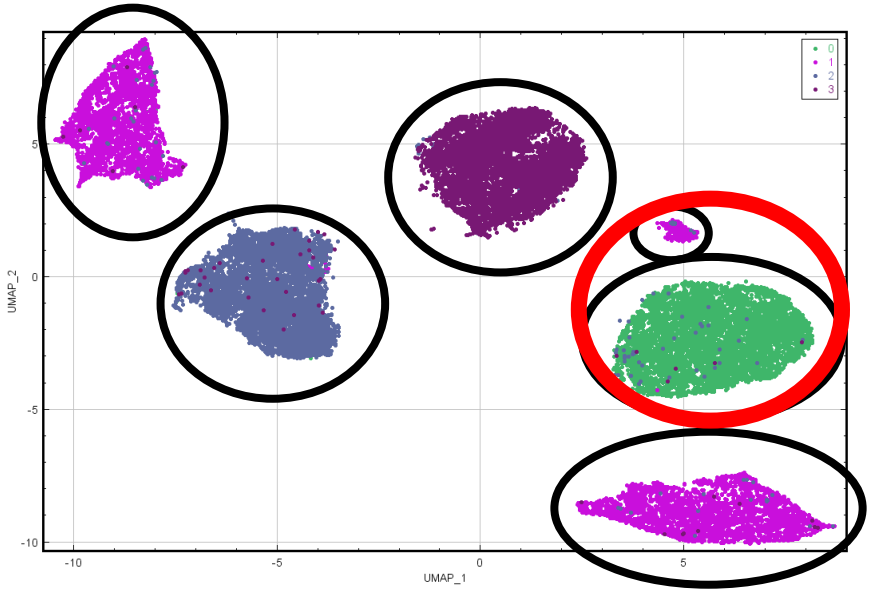
Travis Dick, Cynthia Dwork, Michael Kearns, +3, and Zhiwei Steven Wu Authors Info & Affiliations

Anonymization Primitives

- Pseudonymization / Masking / Suppression / Generalization
- Obfuscation: adding noise (with zero mean) to the attribute value, average value can still be computed fairly accurately.
- Clustering: in the aggregated data all values must be above a threshold (if below, merge two categories).
- Thresholding: values above/below a value are truncated.

Amazon Salaries

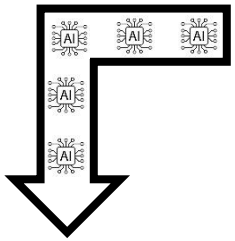
- John S. – 51K
- Steven L. – 48K
- Marie W. – 76K
- Jeff B. – 100K (~~924568256K~~)

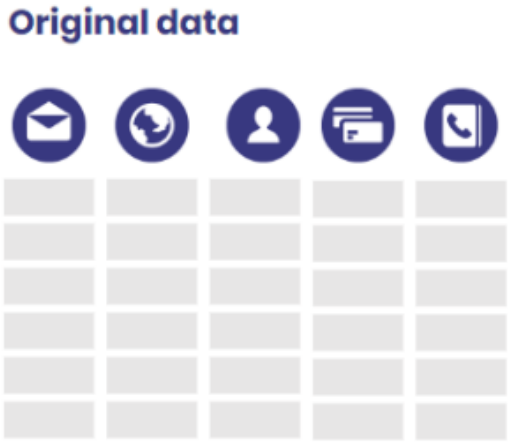


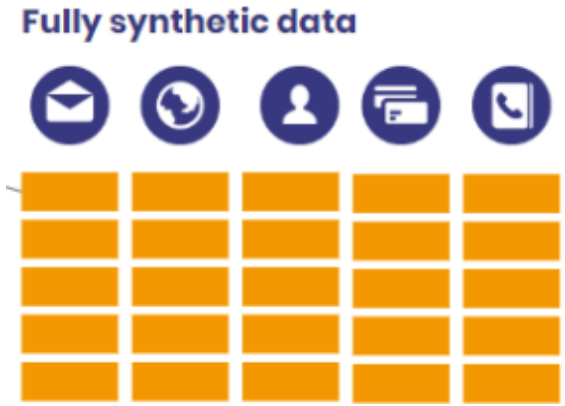
Even More Primitives

- Sampling: remove each record with probability p > 0.
- Rounding
- Data swapping: the attribute values of unique records are swapped.
- Imputation (synthetic data generation): build a model to generalize from the underlying distribution.

  Amazon Salaries
  - John S. – 50K
  - Steven L. – ~~50K~~ 80K
  - Marie W. – ~~80K~~ 50K







Name	Department
[redacted]	[redacted]
[redacted]	[redacted]
Margret	Biology
[redacted]	[redacted]
Davis	Comp. Sci.
Sheryl	Biology

Ad Hoc Methods

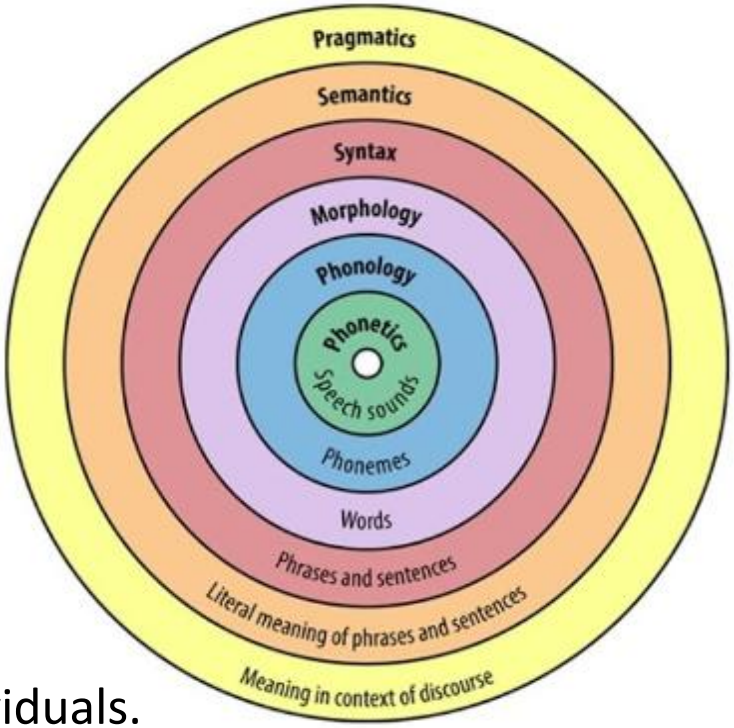
- These primitives are often applied and combined in a non-principled 'ad hoc' manner until some known re-identification attacks fail.
  - There is no guarantee that other, new attack will also fail in the future.

- We first need to define what anonymized data means exactly.
  - It should never be the resistance against some known attacks.
- Then, apply the primitives according to a well-defined (anonymization) algorithm so that the result can be formally proven to satisfy the privacy definition.
  - This principled approach provides guarantee against any attack with respect to the privacy definition.

- The question is whether the privacy definition is reasonable?

Privacy Models

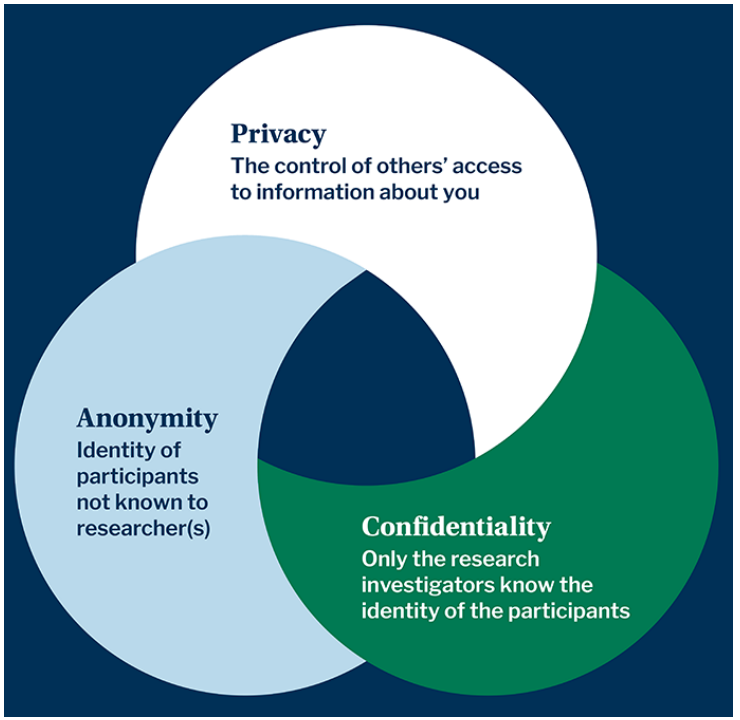
- Perfect (Absolute) Privacy
  - Anything that can be learnt about an individual from a database should also be learnable without accessing the database.
  - Impossible if we want the anonymized data to be useful!
- Syntactic privacy models
  - Impose syntactic conditions on the released data, e.g., each record should occur at least K times.
  - Adversarial background knowledge is always pre-defined.
  - Not necessarily prevent information leakage.



- Semantic privacy models
  - Controls information leakage about individuals.
  - Adversarial background knowledge is not necessarily limited.

Definitions

- We need more rigorous (formal) definition of what anonymity mean!
- How to use the anonymization primitives to get a well-anonymized data and to preserve the utility?
- (Syntactic) Identity disclosure
  - K-anonymity & K-Map
- (Syntactic) Attribute disclosure
  - L-diversity & T-closeness
  - D-presence
- Semantic models
  - Differential Privacy

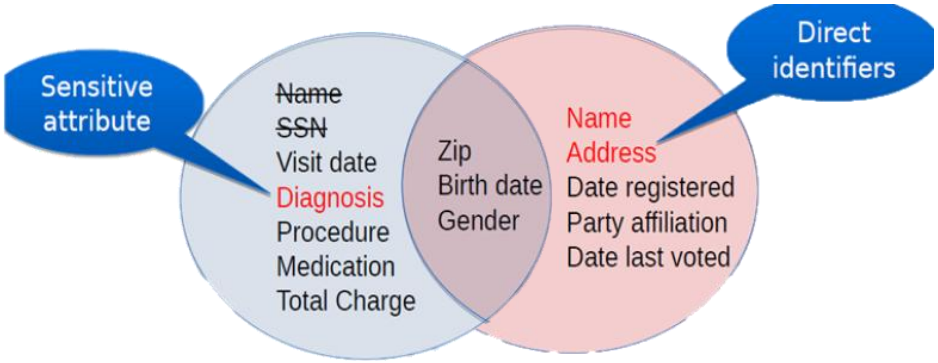


  - N-confusion
  - M-invariance
  - P-uncertainty
  - (h,k,p)-coherence
- and many more …

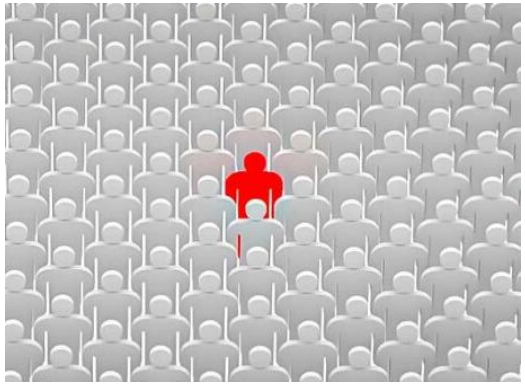
K-Anonymity

Latanya Sweeney

- De-identified the Governor of Massachusetts by matching hospitalization records with voter registration records.



- Recipe
  - The hospital data contained demographic information that could be used to distinguish between different records.
  - A secondary database was available to figure out the demographic information about the target.
  - The target was in both datasets.
  - The demographic information of the target was unique within both datasets: only one record had the demographic values of the governor.



K-Anonymity

- A dataset is said to be K-Anonymous if every combination of values for demographic columns in the dataset appears at least for K different records.
  - An attacker knows the demographic information of their target, but then this will be linked to k different individuals, so it will be impossible to select which one.

ZIP code	age
4217	34
1742	77
1743	77
4217	34

ZIP code	age
4217	34
1742	34
4217	77
1742	77

ZIP code	age
4217	34
4217	34
1742	77
1742	77

Selecting K

- Regulations do not include specific parameter values in laws or guidelines, since there is no convincing argument to be made for a given choice.
  - The level of risk depends on many more fuzzy parameters (e.g., how valuable the data is, how bad would a privacy incident be, etc.).
- Data owners do not know how to choose the parameter either, so they usually buy the services of a privacy consultant.
  - The consultant does not know either what is the "good" choice.
  - But they usually have more practical experience of what are common values in the industry for similar levels of risk.

Methods

- Generalization: replace specific quasi-identifiers with less specific values until get K identical records.
  - Global: easier to do data analysis
  - Local: keep more utility at the cost of complex representation.
- Suppression: if generalization causes too much information loss, delete cells or records or attributes.
  - Good if there are outliers!

ZIP code	age
4217	34
4217	34
1742	34
1742	31

ZIP code	age
4217	30-34
4217	30-34
1742	30-34
1742	30-34

ZIP code	age
4217	34
4217	34
1742	30-34
1742	30-34

ZIP code	age
4217	34
4217	39
1742	75
1691	77
9755	13

ZIP code	age
4217	30-39
4217	30-39
1000-1999	75-79
1000-1999	75-79

Heterogeneity

- Anonymity groups are created from similar records, i.e., it needs a similarity measure.
- Heterogeneous attributes
  - Very different semantic (e.g., Age, ZIP code, Sex) which cannot be compared.
  - What is more similar; two records that have the same age or live in the same location?
  - Via K-d tree the data can be partitioned with the constraint that each partition must have at least k records.

- Non-heterogeneous attributes
  - The same semantic (e.g., location, movie rating, etc.).
  - Possible to use a single metric between records (e.g., Earth Mover Distance).
  - Any standard clustering technique with the constraint that each cluster must have at least k records.

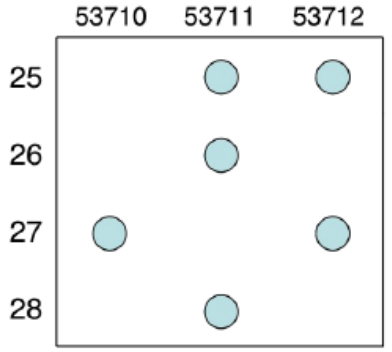
Heterogeneous

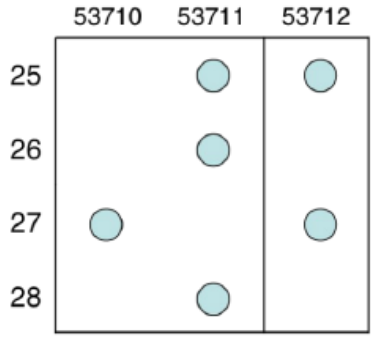
- Mondrian (K-d Tree Partitioning)
  - Given: dataset, k
  - Assumption: there is a total ordering on every attribute domain.

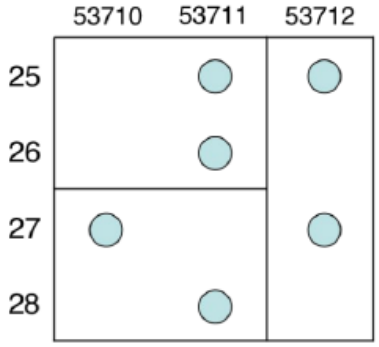
1. Choose a dimension (attribute) in the data.
   - Each Quasi Identifier corresponds to a dimension.
2. Find the median and split the data accordingly.
3. Repeat 1 & 2 until k-anonymity is not satisfied.

Age	ZIP
23	53711
25	53712
26	53711
27	53710
27	53712
28	53711

Age	ZIP
[25 – 26]	53711
[25 – 27]	53712
[25 – 26]	53711
[27 – 28]	[53710 – 53711]
[25 – 27]	53712
[27 – 28]	[53710 – 53711]







Non-Heterogeneous

1. Pick a pivot record randomly and find its K-1 nearest records. This will be an anonymity group.
2. Release the average of these records (aka cluster center) within a group instead of the individual records.
3. Repeat until all records are clustered.

- If there are many attributes (high-dimensional data), the anonymized data can be very distorted.
  - Typical for set-valued data such as location data, transactional data, etc. (i.e., all modern datasets).

- Curse of dimensionality: in high dimension, records are different, i.e., the distance to the nearest and farthest neighbors is almost the same.
  - Clustering will result in a very inaccurate data.

L-Diversity & T-Closeness

Flaws of K-Anonymity

- K-Anonymity is the oldest privacy definition, and it is simple to understand, so it has been adopted by many communities (e.g., healthcare) for their data anonymization needs.
- When an attacker successfully reidentifies someone in a dataset, it's not necessarily a privacy issue.
  - The Voter Registry is already public with identifiers in it.

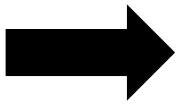
- The leak of sensitive information associated to one given individual is the problem, not the reidentification itself.

- K-Anonymity does not capture this.
  - The definition just prevents you from knowing the real identity of an anonymized record.

Information Leakage

- An attacker might figure out private information about someone, without reidentifying their record. Even if the dataset is K-Anonymous.
  - Name is obviously identifying; must be removed completely.
  - ZIP code and age are quasi-identifiers; they can help identify someone, but reducing their precision might prevent this.
  - Diagnostic is sensitive, but typically considered a secret, so non-identifying.
- Even though the dataset is 2-anonymous, an attacker (knowing Dan's demographic information) could still figure out his diagnosis.

name	ZIP code	age	diagnostic
Alice	4217	34	Common flu
Bob	4212	39	Healthy
Camille	4732	39	Otitis
Dan	4743	23	Otitis



ZIP code	age	diagnostic
421*	30-39	Common flu
421*	30-39	Healthy
47**	20-39	Otitis
47**	20-39	Otitis

L-Diversity

- If all users with the same quasi-identifier tuples (i.e., buckets) have the same sensitive values, that could lead to private information leakage.
  - Solution: imposing some diversity in the sensitive values associated to the same (generalized) tuple.

- L-Diversity: it builds on the definition of K-Anonymity.
  - States that each bucket must have at least L distinct sensitive values.
  - L-Diversity implies L-Anonymity (as each bucket should contain at least L users).
- An attacker is unable to know which records corresponds to Dan. Moreover, the sensitive value stays also private.

ZIP code	age	diagnostic
4***	20-39	Common flu
4***	39	Healthy
4***	39	Otitis
4***	20-39	Otitis

Flaws of L-Diversity

- What's the key idea behind L-Diversity?
  - If the attacker has uncertainty over the sensitive value, then private info leakage is avoided.
- Suppose the attacker knows that their target has ZIP code 4235 and age 25.
  - The target's record is one of the first two rows.
  - The attacker learns that its target either has AIDS, or hepatitis B.
- Not sure which, but still can infer that the target has a sexually transmitted infection.

ZIP code	age	diagnostic
42**	20-29	AIDS
42**	20-29	Hepatitis B
17**	30-39	Otitis
17**	30-39	Healthy

Probabilistic Information Leakage

- Same attack as before: targeting someone with ZIP code 4235 and age 25.
  - An attacker cannot learn the diagnostic for certain.
  - But they can get a strong suspicion that the target has lupus: 9 out of 10 records share this!
  - An insurance company might increase the premium because of such a suspicion.

ZIP code	age	diagnostic
42**	20-29	Lupus
42**	20-29	Lupus
42**	20-29	Lupus
42**	20-29	Lupus
42**	20-29	Lupus
42**	20-29	Lupus
42**	20-29	Healthy
42**	20-29	Lupus
42**	20-29	Lupus
42**	20-29	Lupus

- How to protect against this type of attack?
  - Requiring that sensitive attributes are diverse is not enough.

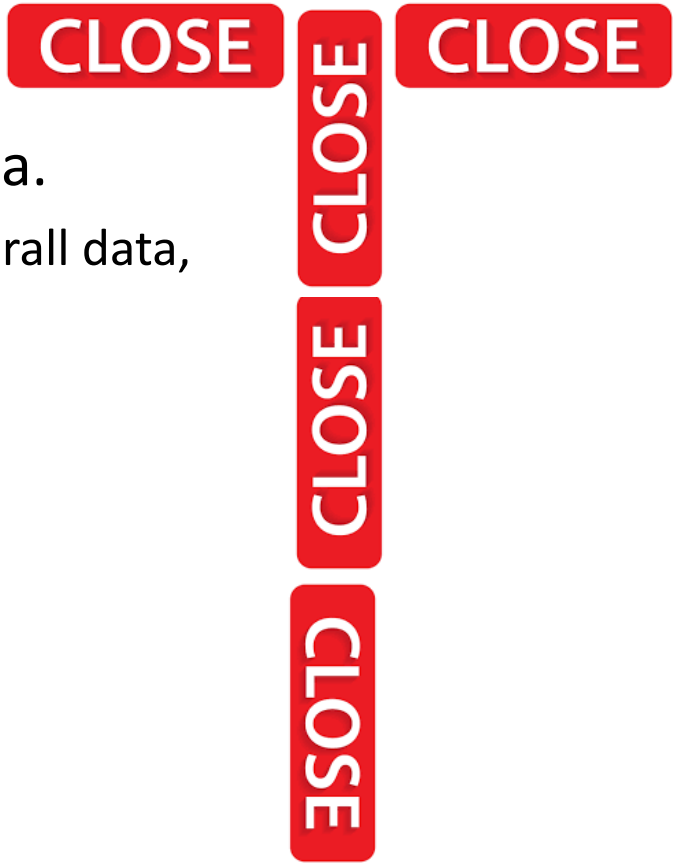
The Guardian

Sun 12 Nov 2023 08.00 CET

Private UK health data donated for medical research shared with insurance companies

T-Closeness

- T-Closeness require that the distribution of sensitive attributes is roughly the same as the rest of the data.
  - If 40% of the records are "healthy" in the overall data, then each bucket must also have roughly 40% of "healthy" records.
  - This way, the attacker's knowledge cannot change too much from the baseline.



- This kind of information is what we were interested in the first place.
  - Completely destroys the utility of the data.

Comparison

- Algorithmically, L-Diversity and K-Anonymity are very similar.
  - The basic blocks are generalization and suppression.
- Choosing K for K-Anonymity is difficult, but selecting L for L-Diversity is not better.
  - No official guideline or regulation will help you choose.
- Although it is strictly stronger, it is hardly ever used.
  - Utility loss: in a study, researchers find that applying 3-diversity was worst than using 100-anonymity.
  - Privacy gain: the advantage is unclear, it is a "fix" for one of K-Anonymity's most obvious flaws.
- In security, simply patching bugs one after the other is not a great defense mechanism.
  - T-Closeness is just another “fix” with limited benefits and heavy costs.

Composability

- Hospital **A** publishes a 4-anonym dataset.
- Hospital **B** publishes a 6-anonym dataset.
- They are independent releases! (Which is very common today.)
- The attacker know Charle, who is 26 years old, and lives in ZIP code 13011.
  - Also knows that he has visited both hospitals.

- There is only one common disease with these quasi-identifiers!

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

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<35	*	AIDS
2	130**	<35	*	Tuberculosis
3	130**	<35	*	Flu
4	130**	<35	*	Tuberculosis
5	130**	<35	*	Cancer
6	130**	<35	*	Cancer
7	130**	≥35	*	Cancer
8	130**	≥35	*	Cancer
9	130**	≥35	*	Cancer
10	130**	≥35	*	Tuberculosis
11	130**	≥35	*	Viral Infection
12	130**	≥35	*	Viral Infection

K-Map & D-Presence

Case Study

- A doctor studies human sexual behavior.
  - Run a survey with only few volunteers (40).

- Data is sensitive, blatant sharing with other researchers is not an option.
  - Based on the attributes, ZIP code and age are the Quasi Identifiers.
- Using 10-Anonymity results in buckets like $20 \leq age \leq 50$.
  - The utility is quite destroyed as it bundles together very different ages.
- Is this necessary? Who are the attackers?

  - The researchers whom the data is shared possibly unknown the subjects.

  - None of these people have background information about who is in the dataset.
  - Thus, besides distinguish between different records the attacker also have to find the real identity of the records

- This attacker has significantly weaker capabilities than for K-Anonymity!

K-Map

ZIP code	age
85535	79
60629	42

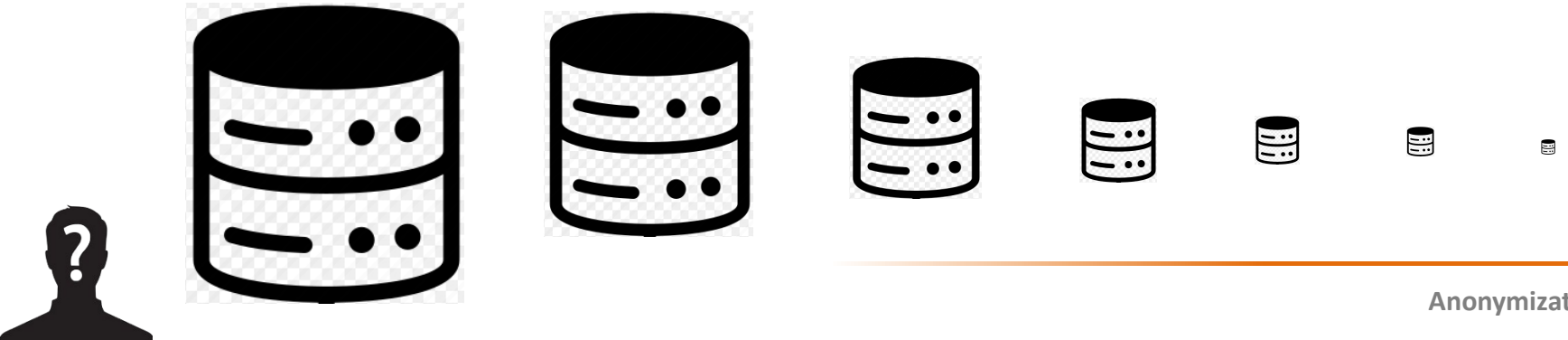
- At first glance, the amount of information for this two individuals seems to be the same.
  - 85535 corresponds to a rural area with approximately 20 people. Probably only one is 79.
  - 60629 corresponds to a city with more than 100,000 people. Thousand are 42.
- According to K-Anonymity, both rows might be unique.
- It make sense to weaken the adversary by assuming it does not know whether their target is in the dataset.

- We compare rows with all other individuals sharing the same values in a larger dataset (called Reidentification Dataset).
  - For instance, "everybody living in the US“.
- Once the QI and the Reidentification Dataset are determined, the definition is straightforward:
  - A dataset satisfies K-Map, if every combination appears at least K times in the reidentification dataset.

ZIP code	age
85***	79
60629	42

Strength

- When choosing the QI and RD, one must think hard at what an attacker could do.
  - K-Map is suitable for attackers with limited resources (e.g., voter files).
  - For more powerful attackers K-Anonymity is a safer bet.
- K-Map weakness comes from the assumption that the attacker has zero information about their target.
  - Except that they live in the US.
- To make the attack model stronger, it is possible to use smaller Reidentification Database (aka better background knowledge).



Use Case

- A different doctor studies a treatment of a particular chronic disease.
  - Run a clinical trial.

ZIP code	age
85535	10
85535	12
85535	13
85535	13
85535	16
85535	43

- Which information is sensitive exactly?
  - For the survey, the answers of each participant.
  - For the clinical study, being in the dataset is the sensitive information.

- Attacker knows that who lives in ZIP code 85535.
  - 5-5 people have ages between [10, 19] and [20, 29].
  - 10-10 people have ages between [30, 39] and [40, 49].
  - 20 people are 50 or older.

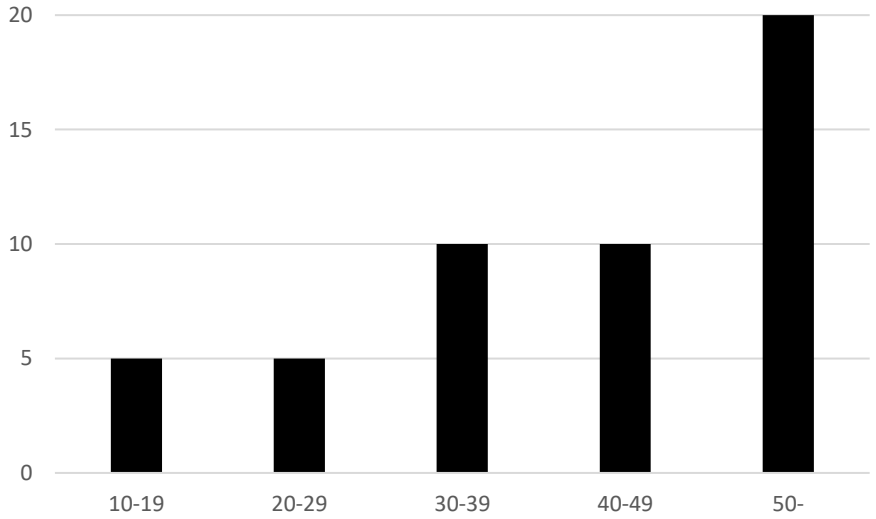
ZIP code	age
85535	10-19
85535	10-19
85535	10-19
85535	10-19
85535	10-19
85535	40-49

- Sharing the data using 5-Map.

Flaws of K-Map

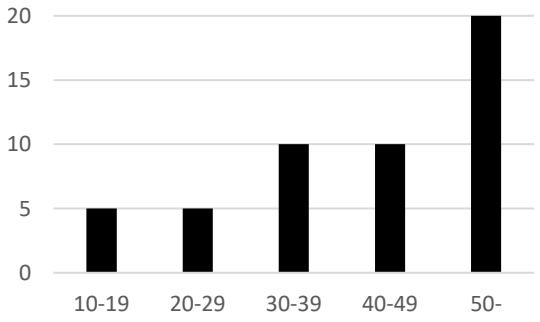
- An attacker knows that there are 5 people aged between 10 and 19 in ZIP code 85535.
  - As all of them are part of the dataset, the attacker learns sensitive individual information without re-identification.
- For each combination of quasi-identifier attributes:
  - K-Anonymity counts the number of records in the dataset.
  - K-Map counts the number of records in the larger population.
- Inclusion probability
  - The ratio $\frac{K_{Anon}}{K_{Map}}$ for the records (85535,[10,19]) is 5/5 = 1.
  - The ratio $\frac{K_{Anon}}{K_{Map}}$ for the records (85535,[40,49]) is 1/10=0.1.

ZIP code	age
85535	10-19
85535	10-19
85535	10-19
85535	10-19
85535	10-19
85535	40-49



D-Presence

- K_{Map} is always higher than K_{Anon}, so their ratio is at most 1.
  - For certain attributes, K_{Anon} and K_{Map} could be equal.
- In D-Presence D is the highest ratio between those two numbers.
  - The highest ratio measures the worst case.
  - D = 1 gives zero guarantee.
- The ratio $\frac{K_{Anon}}{K_{Map}}$ for the records (85535,[10,39]) is 5/20 = 0.25.
- The ratio $\frac{K_{Anon}}{K_{Map}}$ for the records (85535,[40,49]) is 1/10=0.1.



Anon 1

Definition

Relations

¼-Presence

10-Map

ZIP code	age
85535	10-39
85535	10-39
85535	10-39
85535	10-39
85535	10-39
85535	40-49

1-Presence

5-Map

ZIP code	age
85535	10-19
85535	10-19
85535	10-19
85535	10-19
85535	10-19
85535	40-49

Anonymizer Tools

Open & Closed Source Tools

Anon 2

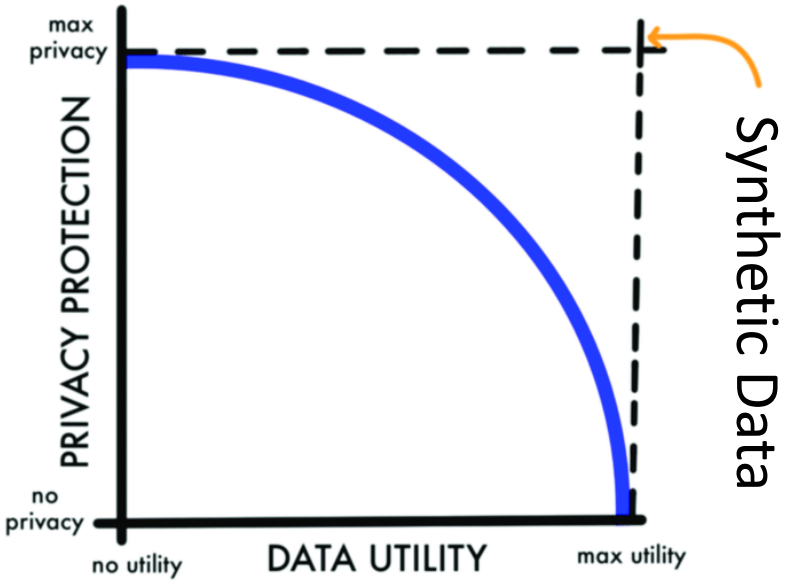
Tool Capabilities

- Open Source
  - Amnestia
  - Anonimatron
  - ARX
  - ProPrivacy
  - SDCTools
- Closed Source
  - AirCloak
  - Privacy Analytics Eclipse

Synthetic Data

Privacy Mirage

- Records created by statistically modeling the original data and then using those models to generate new data values that reproduce the original data's statistical properties.
- Improves data utility
  - Data Augmentation: the ability to enlarge the dataset by generating new data, i.e., include synthetic data to complement the original data.
  - The more complex the deep learning model is, the higher the number of parameters that needs to be trained, therefore, the more data is needed.



- Improves privacy protecytion
  - Privacy Preservation: the ability to replace the dataset by generating new data, i.e., removing the sensitive data and using only synthetic data.

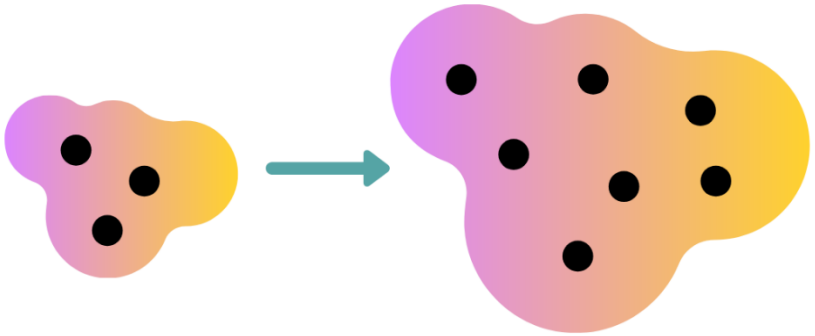
Problem with Utility

- Utility: it is hard to obtain high quality, balanced datasets with labels.
- Metrics to evaluate data generation techniques is crucial.
  - They play a vital role in guiding the development and comparison of different generative models.
- Data Augmentation is measured directly by measuring the model's performance, i.e., whether it increased or decreased when using the original data along the generated samples.
- There is no clear consensus on which are the appropriate sets of metrics that are needed to be used to test if the synthetic data is of use.

Problem with Privacy

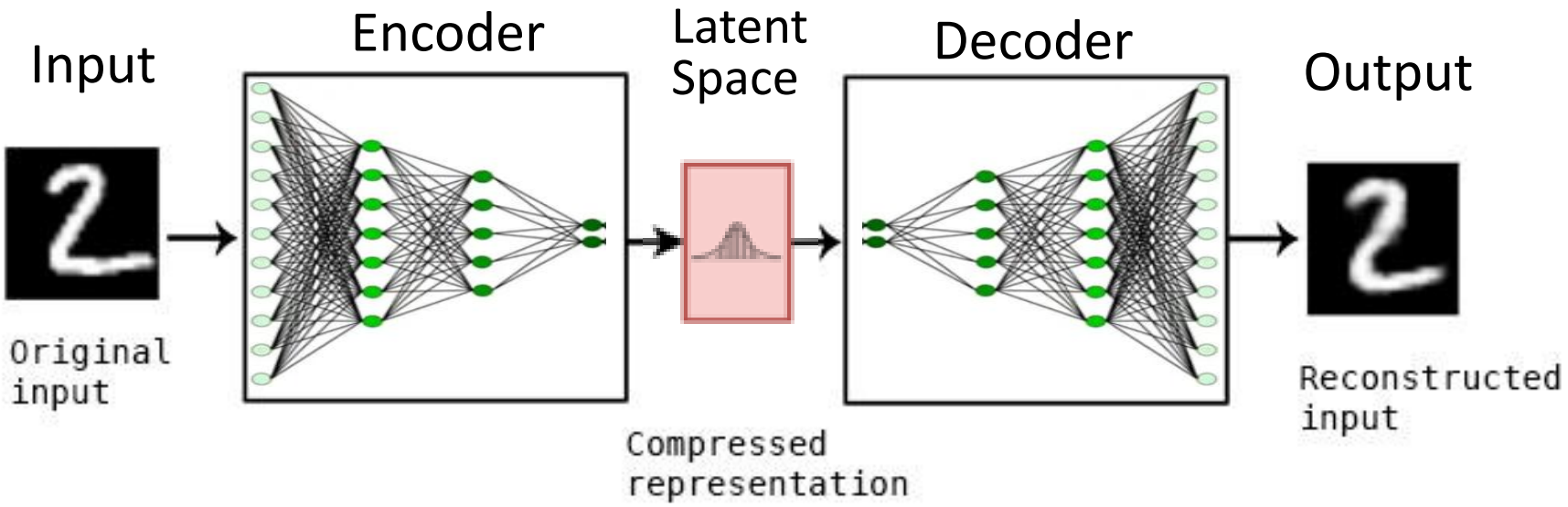
- Privacy: personal information can also be contained in synthetic, i.e., artificially generated, data.

- Modern algorithms for creating synthetic data train machine learning models on the original data and then create new data sets from it, sometimes on demand.
  - This means that personal data that has been integrated into such a model can potentially also be found in the synthetic data.
- Model Types
  - Variational AutoEncoders
  - Generative Adversarial Networks
  - Diffusion Models



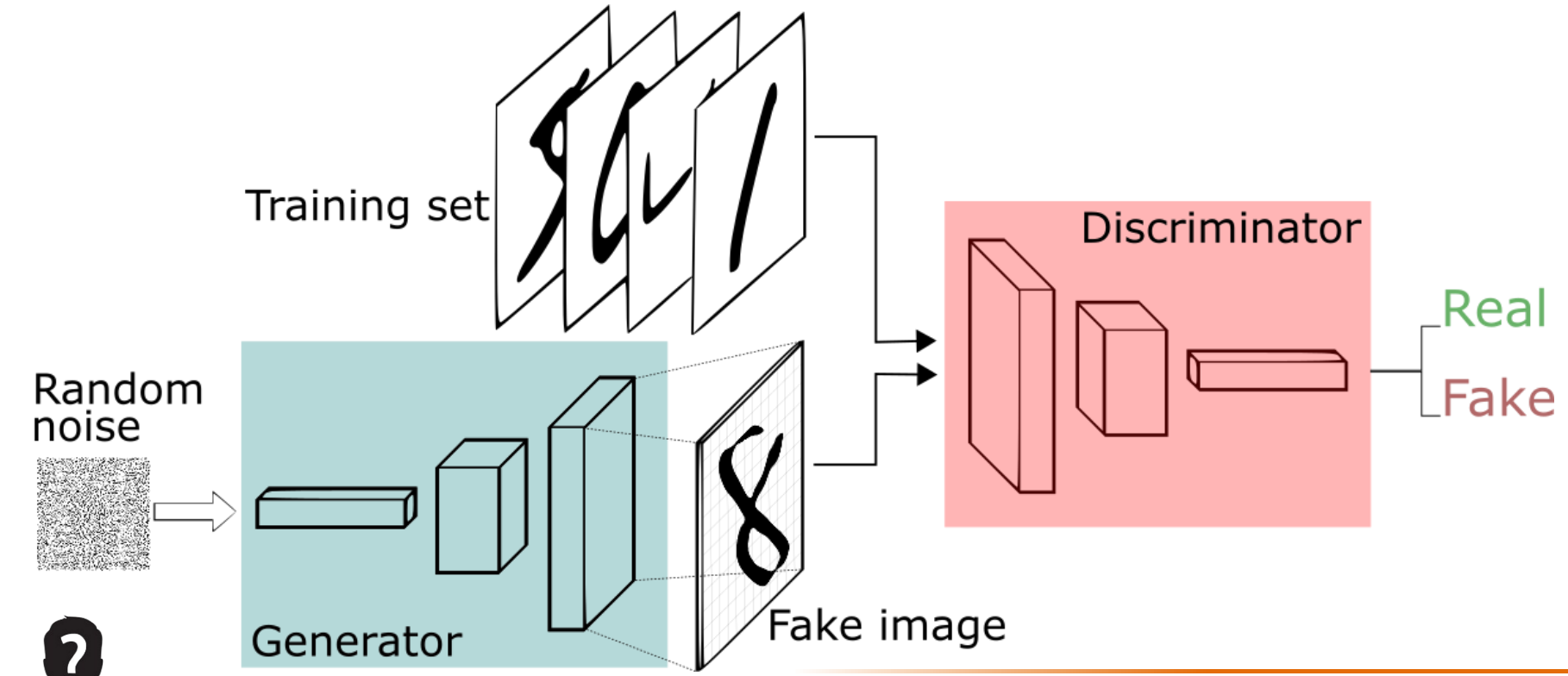
VAE

- Variational AutoEncoders
  - An autoencoder learns an efficient representation (encoding) for a set of data, typically for dimensionality reduction.
  - VAE is an autoencoder whose encodings distribution is regularized during the training in order to ensure that its latent space has good properties allowing us to generate some new data.



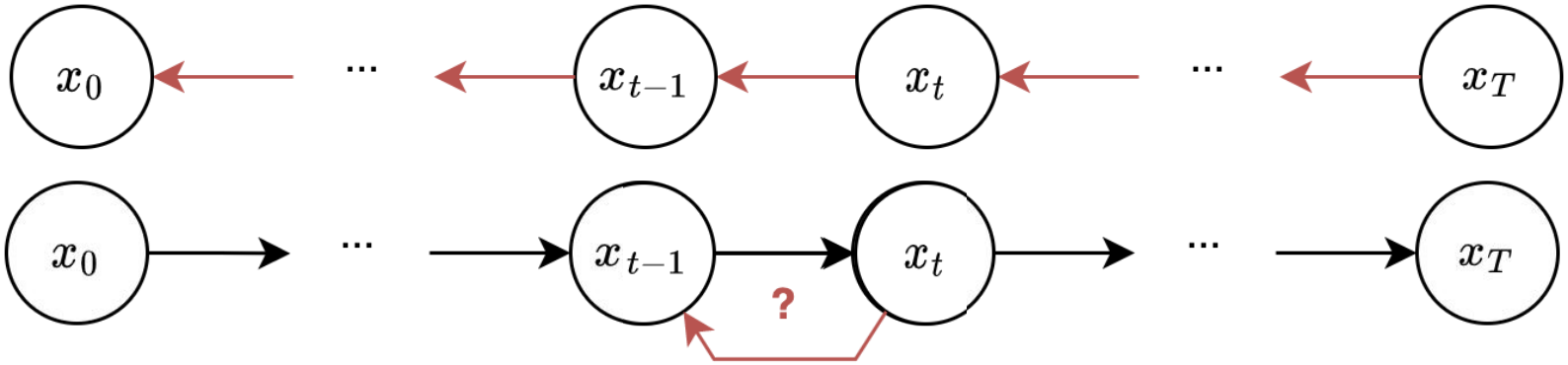
GAN

- Generative Adversarial Network consist of two neural networks:
  - The generator learns to generate synthetic data samples.
  - The discriminator learns to distinguish between real and generated data.
- They are trained in an adversarial manner, continually competing.



Diffusion

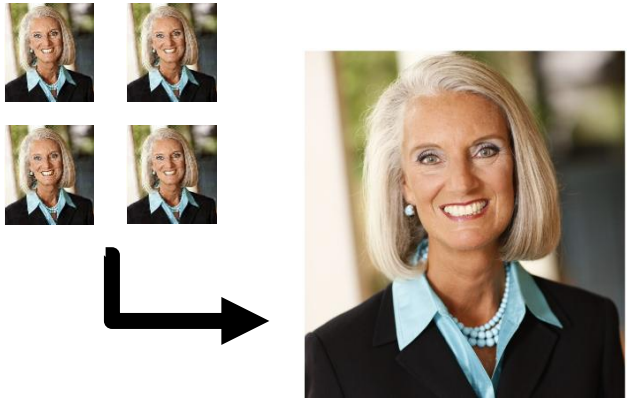
- In each step, a random noise is added to the data.
- The model learns to remove nose in a single step.
- Data Generation: iteratively ‘removing noise’ from an initial random data.



Attack: Extraction

- Data extraction: recover data x from the training dataset.

- Attacker generates many data;
  if some are close, it means there
  is a similar instance in the training data.



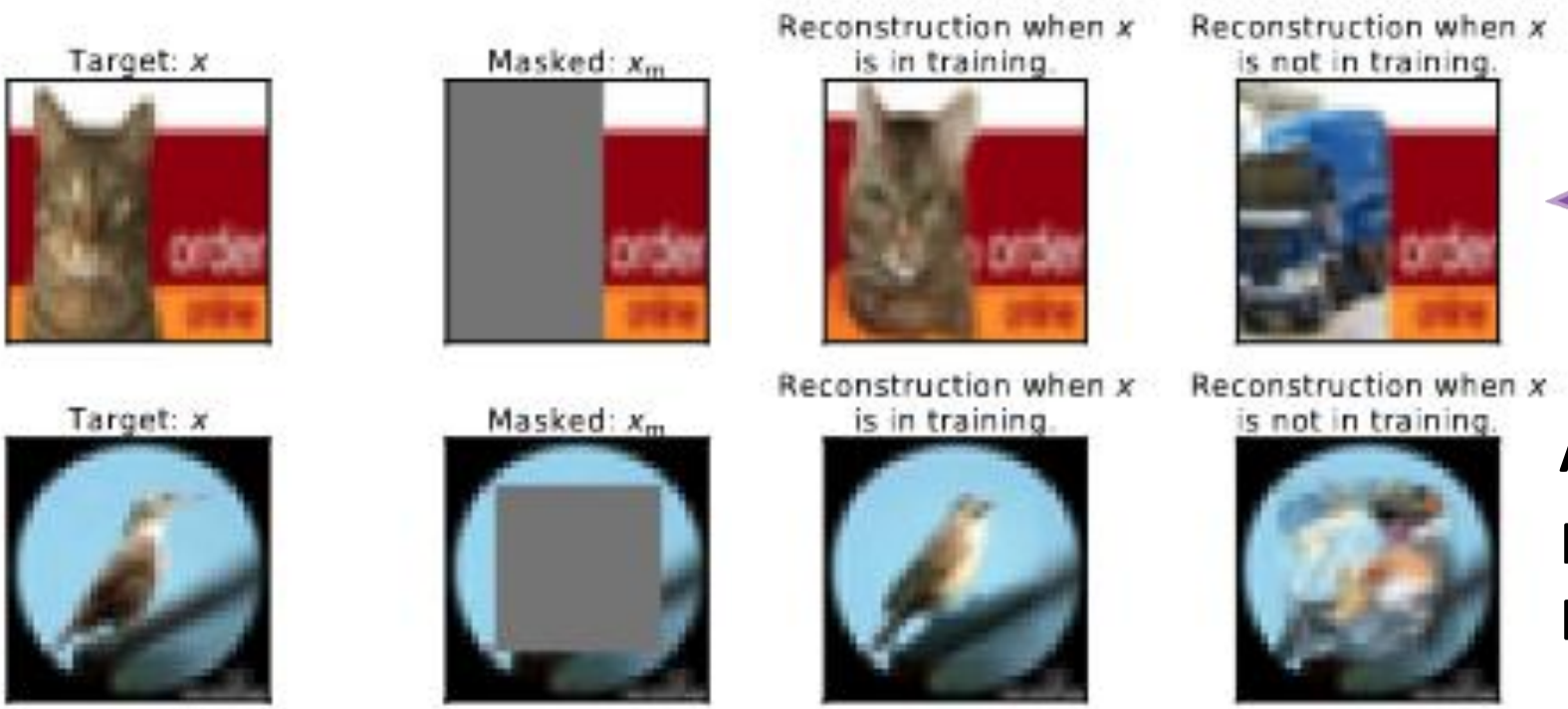
Original:

Generated:

	Parameters	Memorization rate
Stable Diffusion	890 Million	93 out of 175 Million
Imagen	**2 Billion**	**23 out of 5 Million**

Attack: Reconstruction

- Data reconstruction: reconstruct x from partial knowledge of it.
  - Partial knowledge improves extraction results.



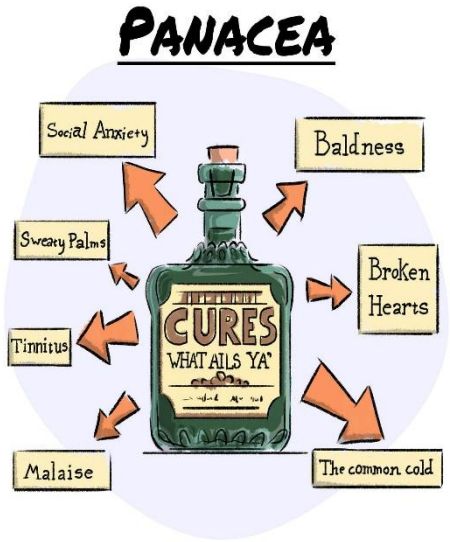
Anon 3

Newer Example

- Do large-scale models work by generating novel output, or do they just copy and interpolate between individual training examples?

Take Away

- Using anonymity primitives in an AdHoc manner is not sufficient.
  - Determining the privacy model (e.g., the capabilities of the attacker) is crucial.
- Utility: if there are many attributes (high-dimensional data), the anonymized data can be very distorted (curse of dimensionality).
- Privacy: provided guarantee is often too weak.
  - Does not compose.
  - Breaks with not considered background knowledge.
- K-Anonymity is wide-spread, but not perfect.
  - L-Diversity & T-Closeness are merely naïve fixes.
  - K-Map & D-Presence are weaker, and hard to use.
- Synthetic data generation is not a panacea.



Control Questions

- List three synthetic data generation architectures and explain at least one in more detail!
- Explain the main primitives used for K-Anonymity and a few techniques to achieve them when the data is (non-)heterogeneous!
- Name six anonymization primitives with examples!

References

- K-anonymity: A model for protecting privacy
- L-diversity: Privacy beyond k-Anonymity
- T-closeness: Privacy Beyond k-Anonymity and l-Diversity
- Achieving k-anonymity privacy protection using generalization and suppression
- δ-presence without complete world knowledge
- Extracting training data from diffusion models
- Ted Blog