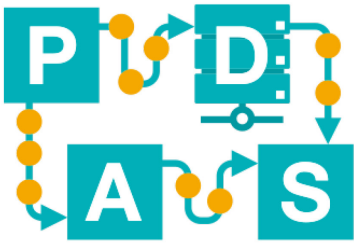
Responsible Data Science (2/2)

Lecture 21

IDS-L21



Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Outline of Today's Lecture

- Terminology
- Cryptography
- Privacy-Preserving Techniques
  - Randomization
  - Group-Based Anonymization
    - k-Anonymity
    - l-diversity
    - t-closeness
  - Distributed Privacy Preservation
  - Privacy Loss via Results (e.g. models, rules).
- Responsible Process Mining

Terminology

MOOC_666


fairness vs accuracy
confidentiality vs accuracy
confidentiality vs transparency
confidentiality vs fairness

Concepts

Privacy

Privacy is defined as the right of an individual to control who has access to his/her individual information.

Confidentiality

Confidentiality refers to an agreement about maintenance and who has access to the classified/sensitive data.

Anonymity

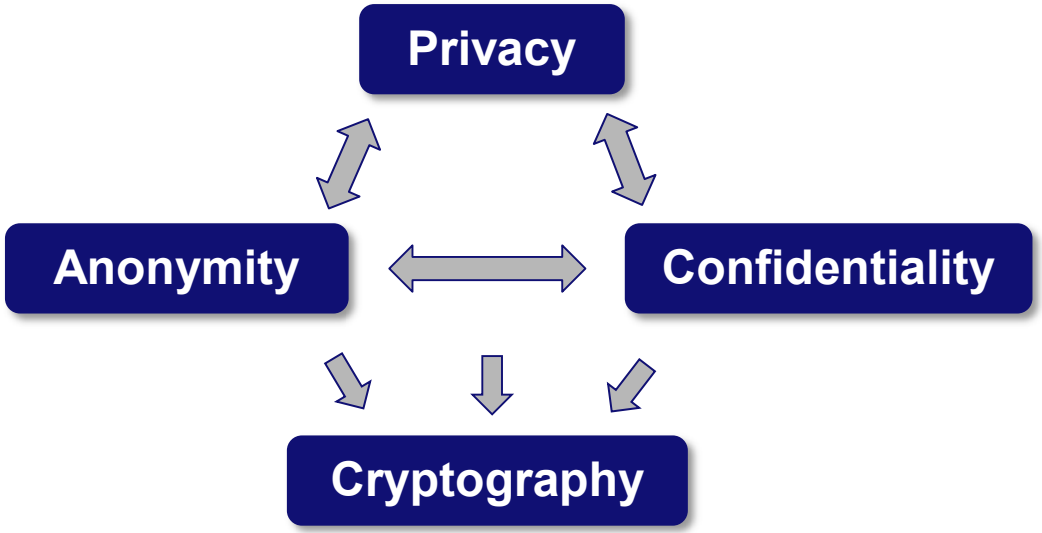
Data is anonymous when the individual described is not known and can not be identified.

Cryptography

Cryptography is about constructing and analyzing protocols that prevent third parties or the public from reading private/confidential information.

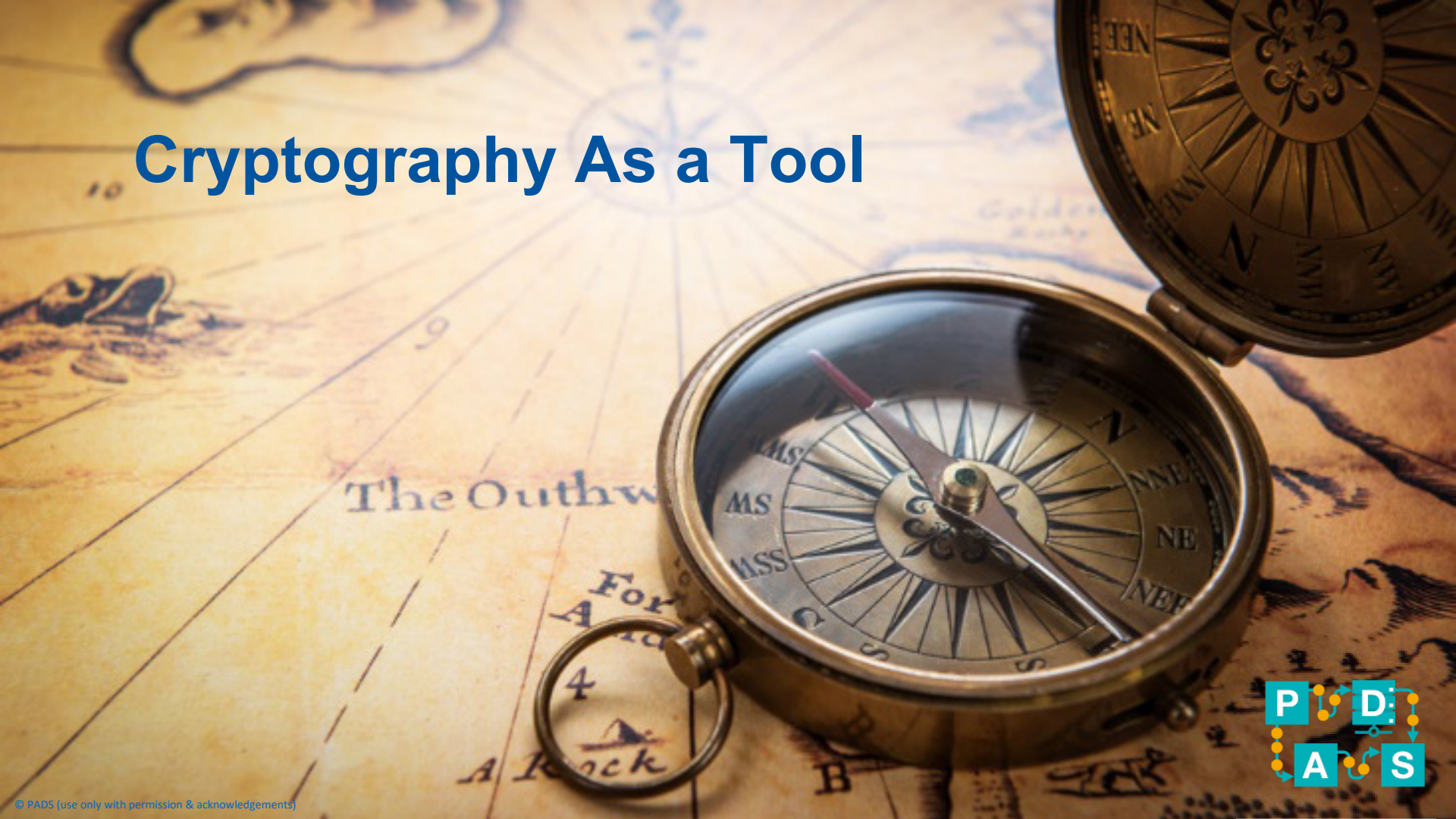
Concepts

- Often these concepts are used interchangeably
- Anonymization techniques can be used to provide privacy, but they are not the same concept
- Cryptography is a tool which can be used for providing privacy, confidentiality, and anonymity.



Chair of Process and Data Science

Cryptography As a Tool



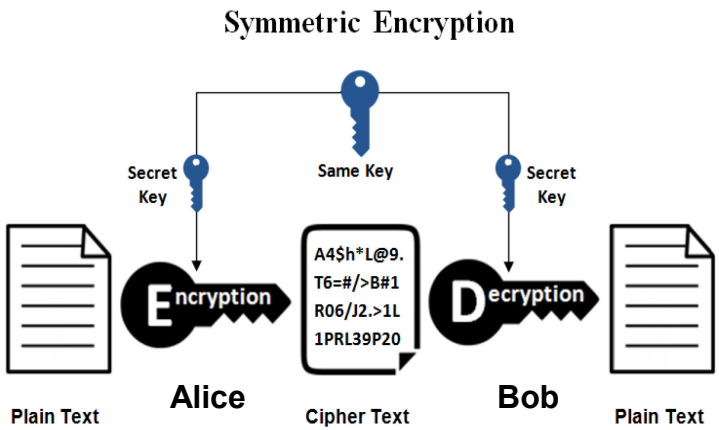
Cryptosystem

- Cryptosystem is a suite of cryptographic algorithms needed to implement a particular security service.
  - Cryptography is not just using some keys to convert a readable data to unreadable data.
  - Cryptographic algorithms should be selected and used for a particular purpose.

- There are different kinds of cryptosystems:
  - Symmetric cryptosystem
  - Asymmetric cryptosystem
  - Deterministic cryptosystem
  - Probabilistic cryptosystem
  - Homomorphic cryptosystem
  - Etc.

Cryptosystem Symmetric, e.g., AES, DES, etc.

- In symmetric cryptosystems, the same secret key is used to encrypt and decrypt a message.
  - Data manipulation in symmetric systems is fast.
  - Advanced Encryption Standard (AES) is a symmetric encryption algorithm.

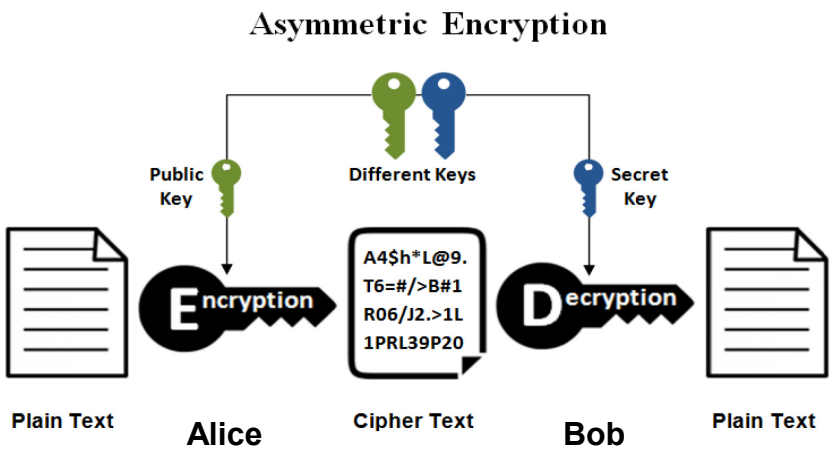


Both Alice and Bob need to agree on a key

Chair of Process and Data Science

Cryptosystem Asymmetric, e.g., RSA

- **Asymmetric** cryptosystems use a **public key** to encrypt a message and a **private key** to decrypt it or vice versa.
  - Security of the public key is not required because it is publicly available and can be passed over the internet.
  - It enhances the security of communication.
- **Rivest-Shamir-Adleman (RSA)** is a well-known asymmetric encryption algorithm.



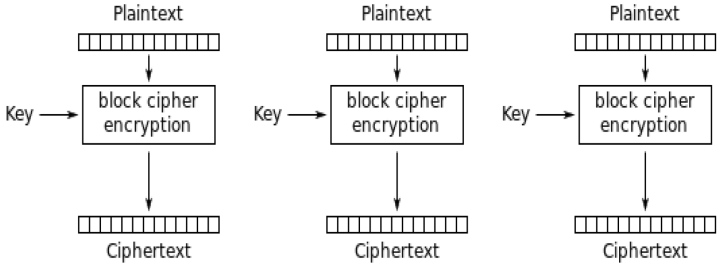
- Both keys are of course related.
- Alice can encrypt the plain text using Bob's public key.
- Bob can decrypt using his private key.
- He does not need to share this key.

Chair of Process and Data Science

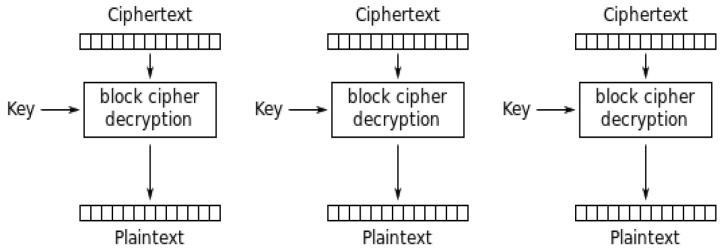
Cryptosystem Deterministic, e.g., AES-ECB

- A deterministic cryptosystem is a cryptosystem which always produces the same cipher text for a given plaintext and key, even over separate executions of the encryption algorithm.
  - It is used when the differences need to be preserved.
  - AES-ECB (Electronic Code Book) is a deterministic encryption algorithm.
  - AES will only encrypt 128 bit of data, but if we want to encrypt whole messages we need to choose a block mode.
    - The simplest block mode is Electronic Codebook or ECB.

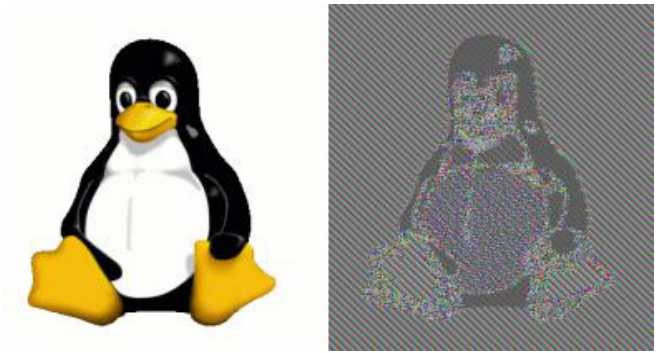
Cryptosystem Deterministic



Electronic Codebook (ECB) mode encryption



Electronic Codebook (ECB) mode decryption



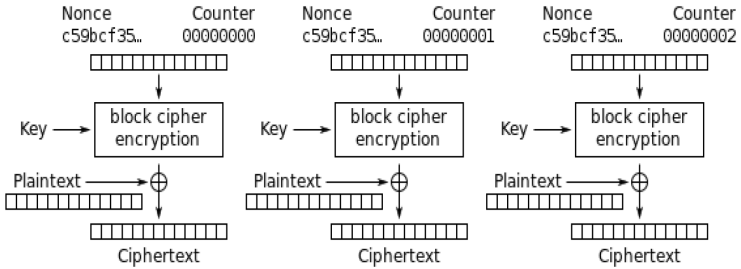
Since identical blocks are encrypted in the same way, one can still see patterns.

Chair of Process and Data Science

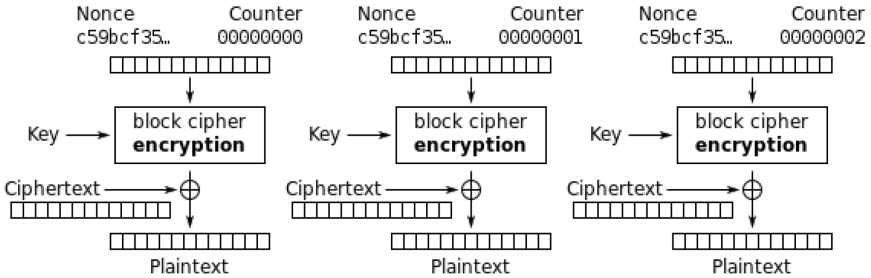
Cryptosystem Probabilistic, e.g., AES-CTR

- A probabilistic cryptosystem as opposed to deterministic cryptosystem is a cryptosystem which uses randomness in an encryption algorithm so that when encrypting the same plaintext several times it will produce different ciphertexts.
  - It is used when the differences need to be hidden
  - AES-CTR (Counter Mode) is a deterministic encryption algorithm, but:
    - Every block is encrypted with the key, a nonce (for initialization), and the counter value.
    - Since the counter of each block is different, the ciphertexts are different (even when plaintext are the same).

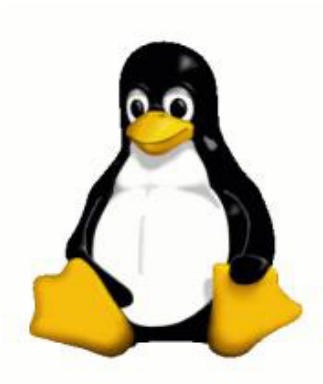
Cryptosystem Probabilistic

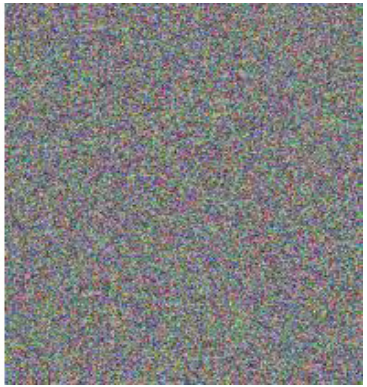


Counter (CTR) mode encryption



Counter (CTR) mode decryption

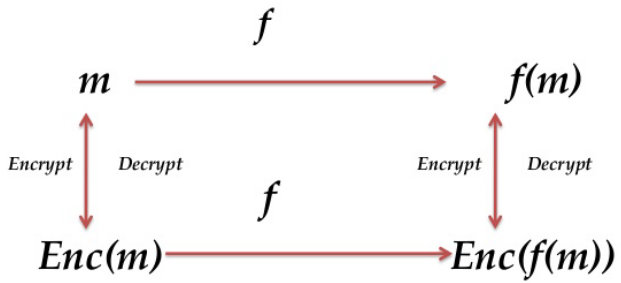


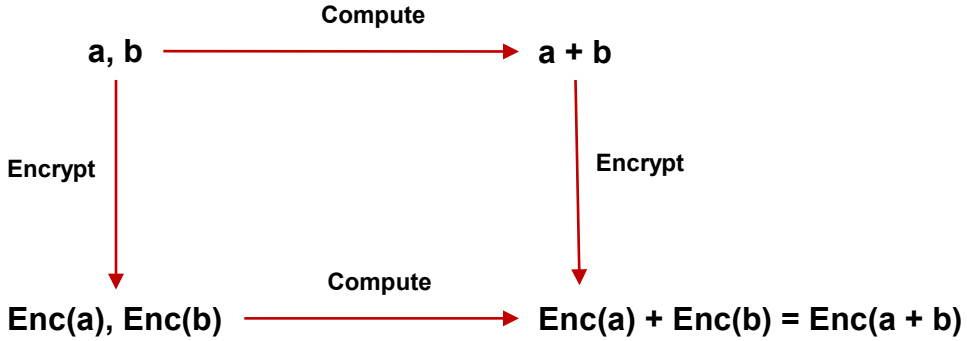


Chair of Process and Data Science

Cryptosystem Homomorphic, e.g., Pallier

- **Homomorphic encryption** is a form of encryption that allows computation on ciphertexts without decryption.
  - When it supports *arbitrary computation* on ciphertexts, it is called fully homomorphic
  - When it supports just some specific computations, it is called partially homomorphic
  - For example, Paillier is a partially homomorphic cryptosystem.





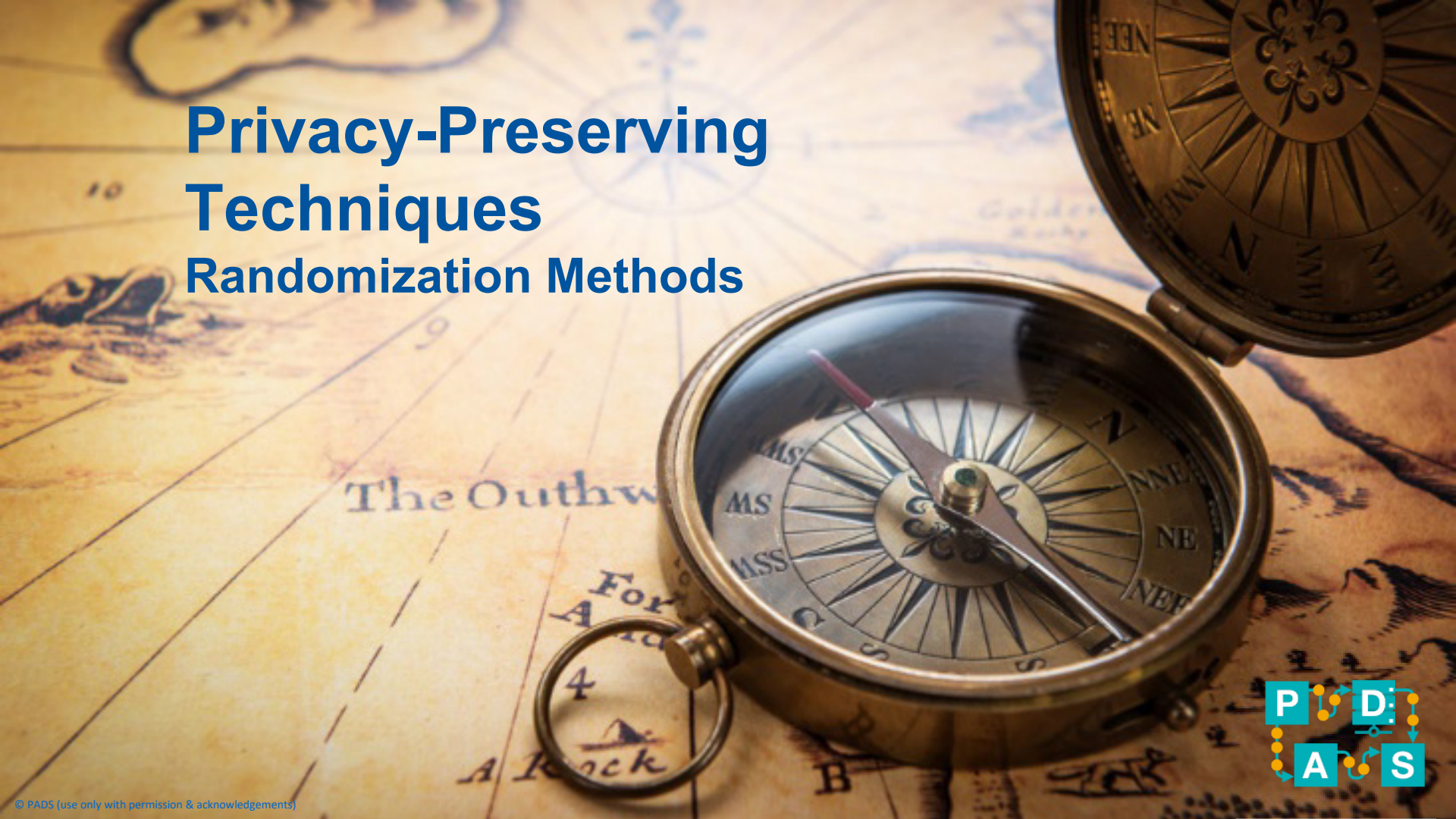
Chair of Process and Data Science

Cryptography Conclusions

- A cryptographic algorithm should be chosen with respect to the use case (goals and possible threats).
- Picking an inappropriate cryptographic algorithm can lead to more computations without any improvement in security.
- Cryptographic algorithms can be combined to provide pragmatic solutions when data sets are large and should be kept confidential.
  - For example, most of the time a symmetric cryptosystem is used for encrypting large data (because it is faster) and a symmetric-key is encrypted by an asymmetric cryptosystem in order to exchange it securely.

Privacy-Preserving Techniques

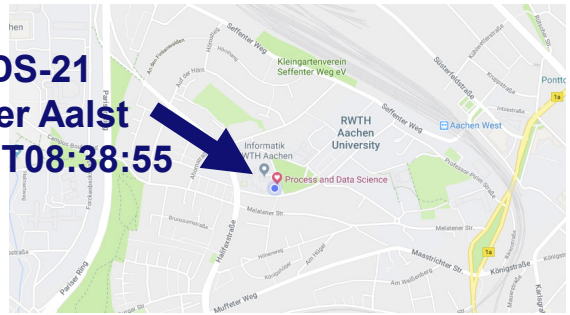
Randomization Methods



Privacy-Preserving

- In recent years, data mining and machine learning showed to be a possible threat to privacy due to the widespread proliferation of electronic data maintained by corporations.
- Privacy-preserving techniques use some form of transformation on the data in order to protect confidential information.
- Typically, such methods reduce the granularity of representation in order to reduce the privacy. This reduction in granularity results in some loss of effectiveness of data management or mining algorithms.

Lecture IDS-21
Wil van der Aalst
10-1-2019T08:38:55



Teaching
Professor
January, 2019

Chair of Process and Data Science

Privacy-Preserving Techniques

- **Privacy-preserving techniques can be categorized as follows:**
  - **Randomization**
  - **Group based anonymization**
  - **Distributed privacy-preserving**

Privacy-Preservation Using Randomization

Additive Noise

- Consider a set of data records denoted by

$$X = \{x_1, x_2, \ldots, x_n\}$$

- For record $x_i \in X$, we add noise component which is drawn from the probability distribution.
- These noise components are drawn independently, and they are denoted by

$$Y = \{y_1, y_2, \ldots, y_n\}$$

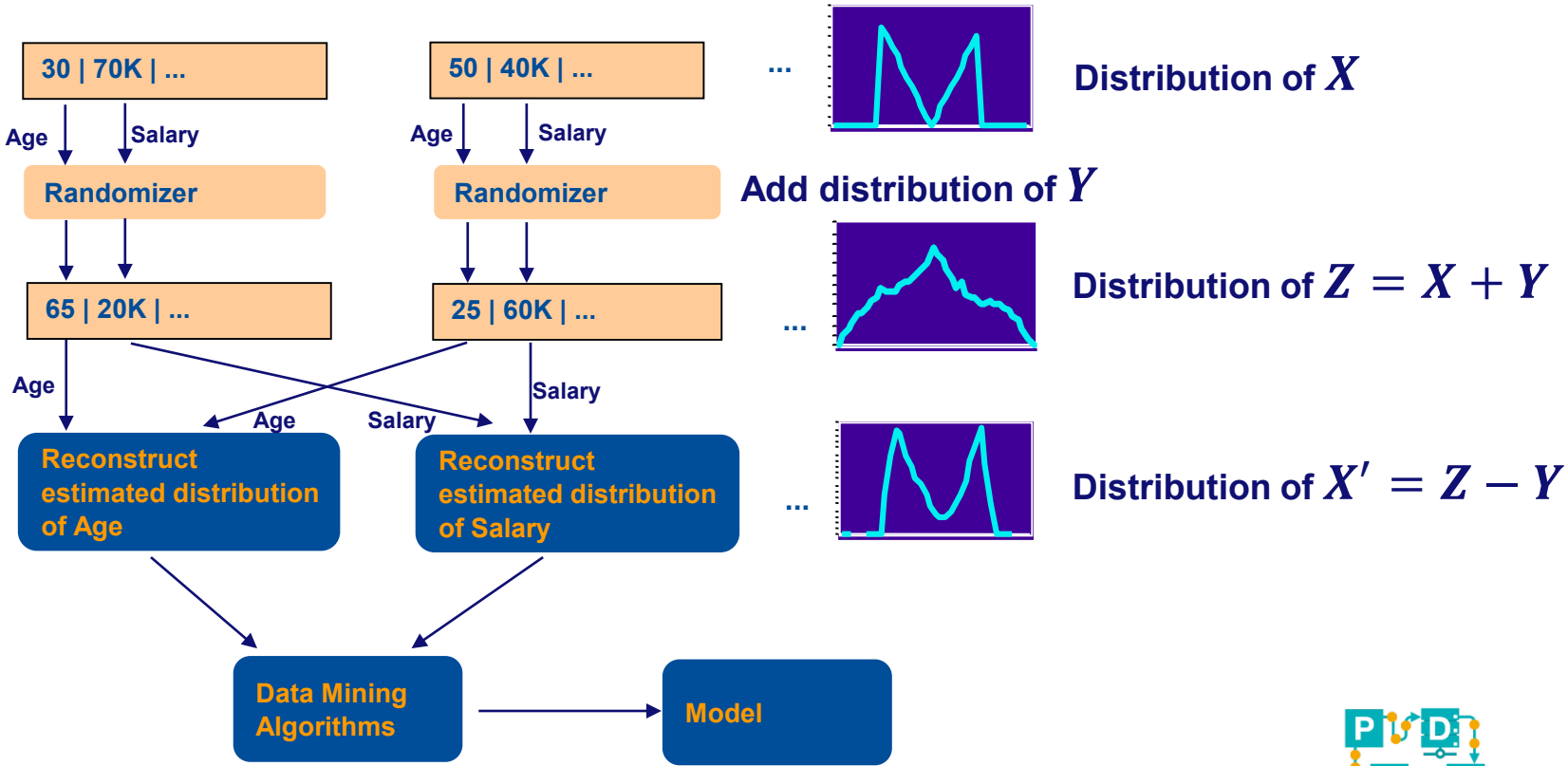
- The new set of distorted records are denoted by:

$$Z = \{z_1 = x_1 + y_1, z_2 = x_2 + y_2, \ldots, z_n = x_n + y_n\}$$

- It is assumed that the variance of the added noise is large enough, so that the original record values cannot be guessed easily from the distorted data.
- The original records cannot be recovered, but the distribution of the original records can be estimated.
- Next to additive noise $(Z = X + Y)$ one can use multiplicative noise $(Z = X \cdot Y)$.

Privacy-Preservation Using Randomization

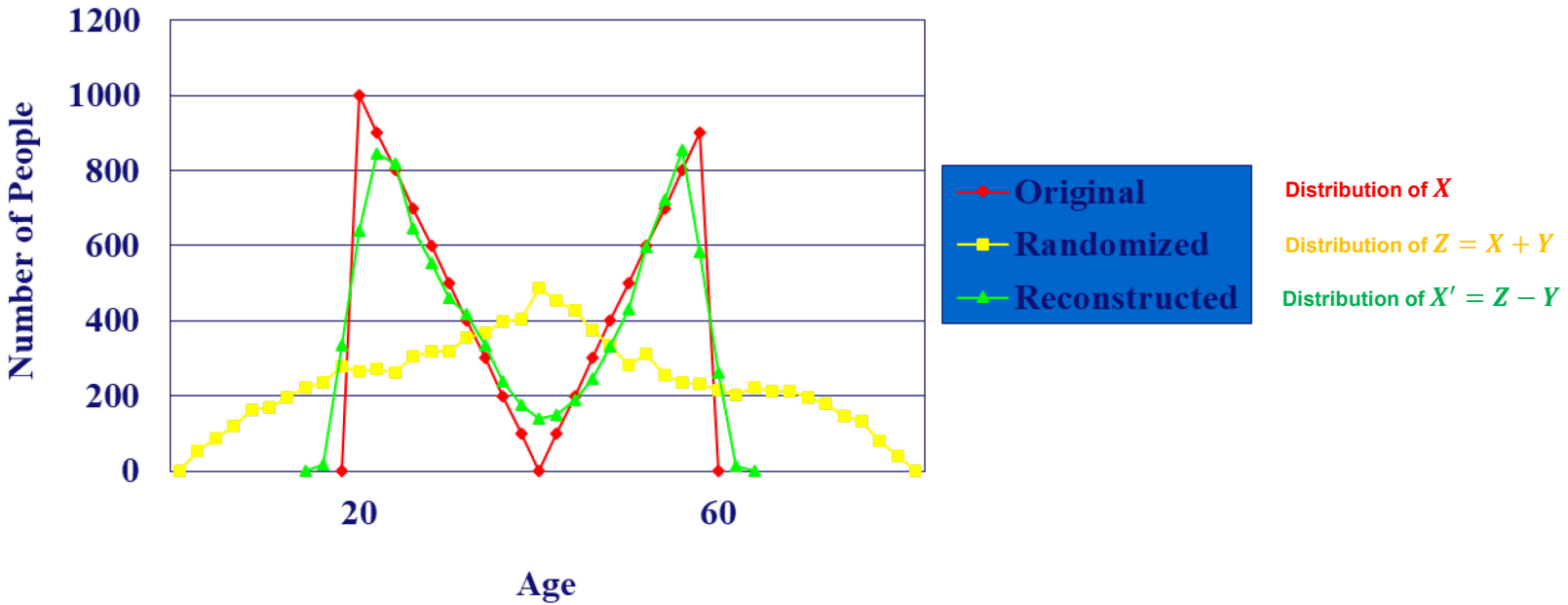
Additive Noise



Chair of Process and Data Science

Privacy-Preservation Using Randomization

Additive Noise



Privacy-Preservation Using Randomization

Additive Noise

- **Strengths**
  - It is relatively simple, and does not require knowledge of the distribution of other records in the data.
    - Therefore, the randomization method can be implemented at *data collection time*.

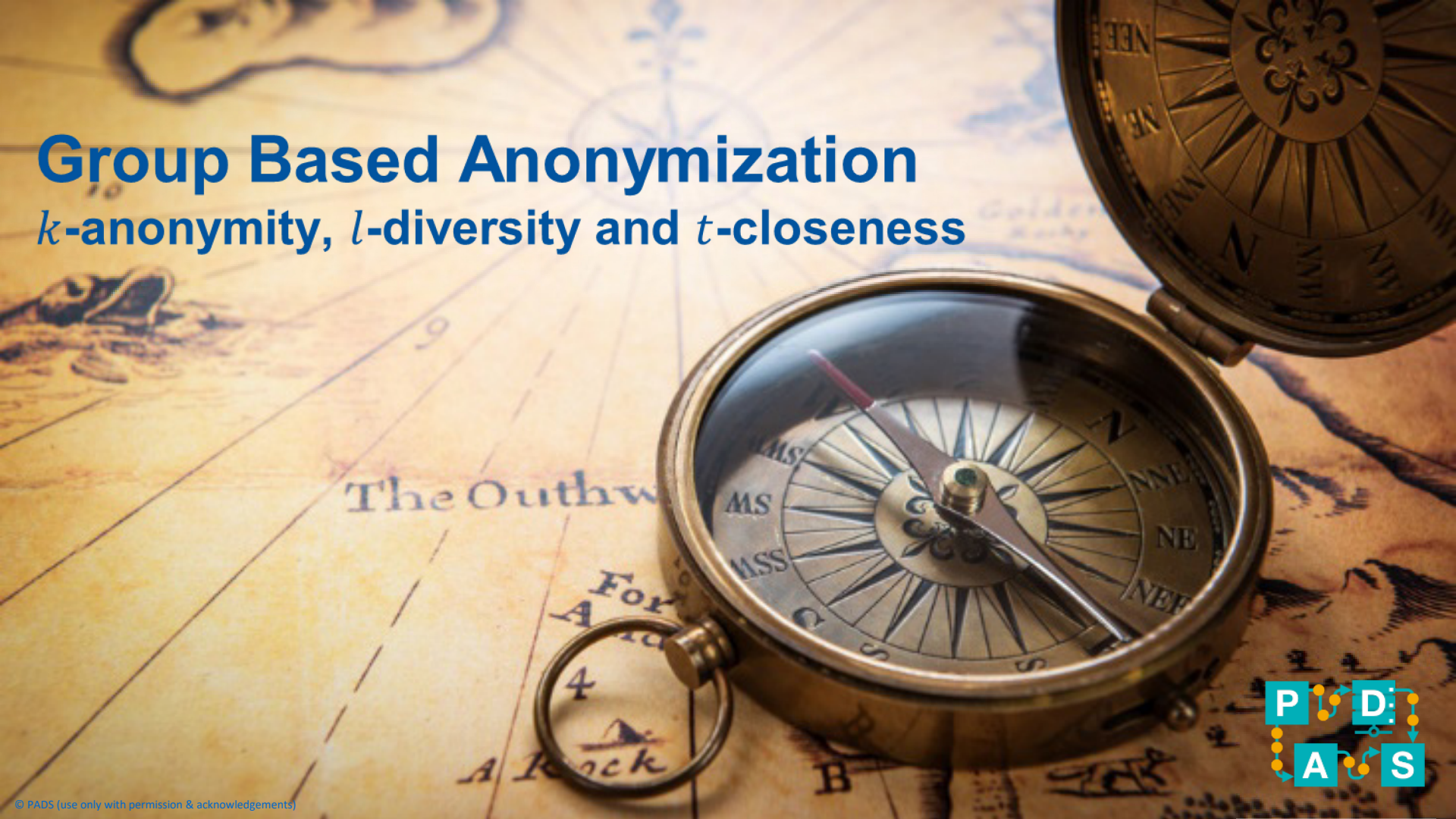
- **Weaknesses**
  - It treats all records equally irrespective of their local density.
    - Therefore, outlier records are more susceptible to adversarial attacks as compared to records in more dense regions in the data.
  - The correlation structure in the original data can be estimated fairly accurately (in larger data sets) even after noise addition.
    - Once the broad correlation structure in the data has been determined, one can then try to remove the noise in the data in such a way that it fits the aggregate correlation structure of the data.

Privacy-Preservation Using Randomization

- Other techniques for the randomization
  - Multiplicative Noise
    - The most common method of randomization is that of additive perturbations. However, multiplicative perturbations can also be used to good effect for privacy-preserving data mining.
  - Data Swapping
    - In data swapping, the values across different records are swapped in order to perform the privacy-preservation (instead of addition or multiplication).

Group Based Anonymization

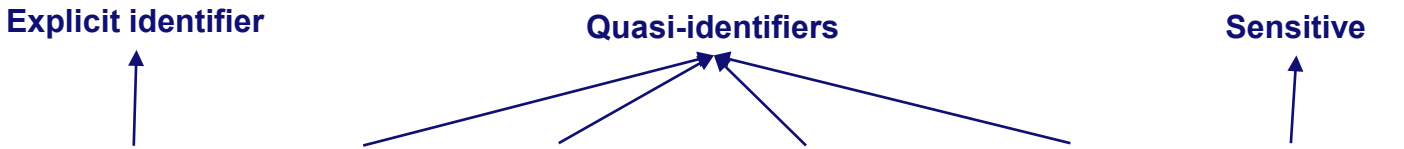
k-anonymity, l-diversity and t-closeness



Group Based Anonymization

- **Attributes can be divided into the following three categories:**
  - **Explicit identifiers**
    - **Attributes that clearly identify individuals**
      - **E.g., Social Security Number, Address, and Name, …**
  - **Quasi-identifiers**
    - **Attributes whose values when taken together can potentially identify an individual**
      - **E.g., Zip-code, Birthdate, Position, and Gender.**
  - **Sensitive attributes**
    - **Attributes that are considered sensitive, such as Disease and Salary**

Group Based Anonymization



Name	Age	Gender	State of domicile	Religion	Disease
Ramsha	22	Female	Tamil Nadu	Hindu	Gastric ulcer
Yadu	24	Female	Kerala	Hindu	Gastritis
Salima	25	Female	Tamil Nadu	Muslim	Flu
Sunny	25	Male	Karnataka	Parsi	Bronchitis
Joan	24	Female	Kerala	Christian	Heart Disease
Bahuksana	23	Male	Karnataka	Buddhist	Bronchitis
Rambha	19	Male	Kerala	Hindu	Flu
Kishor	24	Male	Karnataka	Hindu	Gastric ulcer
Johnson	17	Male	Kerala	Christian	Cancer
John	19	Male	Kerala	Christian	Flu

Group Based Anonymization

- When data is released (makes public), it is necessary to prevent the sensitive information of the individuals from being disclosed.
- Two types of information disclosure:
  - **Identity disclosure**
    - Identity disclosure occurs when an individual can be linked to a particular record in the released table.
  - **Attribute disclosure**
    - Attribute disclosure occurs when new information about some individuals is revealed, i.e., the released data makes it possible to infer the characteristics of an individual more accurately than it would be possible before the data release (e.g., salary or disease).
- Identity disclosure often leads to attribute disclosure.

Group Based Anonymization

- The goal is to limit the (identity/attribute) disclosure risk to an acceptable level by anonymization.
- The first step of anonymization is to remove explicit identifiers.
- However, this is not enough, as an adversary may already know the quasi-identifier values of some individuals in the table (e.g. female professor below 30, RWTH employee living in Schleiden).
- Therefore, the following techniques have been introduced:
  - k-anonymity
  - l-diversity
  - t-closeness

Group Based Anonymization: k-anonymity

- A common anonymization approach is generalization, which replaces quasi-identifier values with values that are less-specific but semantically consistent.
  - As a result, more records will have the same set of quasi-identifier values.
- We define an equivalence class of an anonymized table to be a set of records that have the same values for the quasi-identifiers.
- k-anonymity requires that each equivalence class contains at least k records.

Group Based Anonymization: k-anonymity

	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

1-anonymity

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

4-anonymity requires that each equivalence class contains at least 4 records.

Group Based Anonymization: k-anonymity

There are two common methods for achieving k-anonymity for some value of k.

- **Suppression**: In this method, certain values of the attributes are replaced by an asterisk '*'. All or some values of a column may be replaced by '*'.
- **Generalization**: In this method, individual values of attributes are replaced by a broader category.
  - For example, the value '19' of the attribute 'Age' may be replaced by ' ≤ 20', the value '23' by '20 < Age ≤ 25' , etc.

Group Based Anonymization: k-anonymity

	ZIP Code	Age	Disease
1	47677	29	Heart Disease
2	47602	22	Heart Disease
3	47678	27	Heart Disease
4	47905	43	Flu
5	47909	52	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
8	47673	36	Cancer
9	47607	32	Cancer

Quasi-identifiers

Sensitive attribute

Equivalence class
With **k=3** records

	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

(partial) suppression

generalization

(partial) suppression

If a table satisfies k-anonymity for some value k, then anyone who knows only the quasi-identifier values of one individual cannot identify the record corresponding to that individual with confidence greater than 1/k.

Group Based Anonymization: k-anonymity

- While k-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure.
- K-anonymity focuses only on quasi-identifiers (QIDs) such that each QID tuple occurs in at least k records for a dataset with k-anonymity, and sensitive attributes are not considered.

	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

Homogeneity attack: The sensitive attribute value for all the records of this equivalence class are the same (**Heart Disease**)

Background knowledge attack: Suppose that, by knowing Carl's age and zip code, Alice can conclude that Carl corresponds to a record in this equivalence class. Furthermore, suppose that Alice knows that Carl has very low risk for heart disease.
This background knowledge enables Alice to conclude that Carl most likely has cancer.

l-diversity has been introduced to address this issue

Group Based Anonymization: l-diversity

- An equivalence class is said to have l-diversity, if there are at least l "well-represented" values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity.

- There are different interpretations of term "well-represented":
  - Distinct l-diversity
  - Entropy l-diversity
  - Recursive (c,l)-diversity

Group Based Anonymization: l-diversity

- **Distinct l-diversity:**
  - **The simplest understanding of "well represented" would be to ensure there are at least l distinct values for the sensitive attribute in each equivalence class.**
  - **Distinct l-diversity does not prevent probabilistic inference attacks (background knowledge attack).**
    - **This motivated the development of the entropy l-diversity and recursive (c,l)-diversity.**

Distinct 3-diverse

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	Heart Disease
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

Three different values

Group Based Anonymization: l-diversity

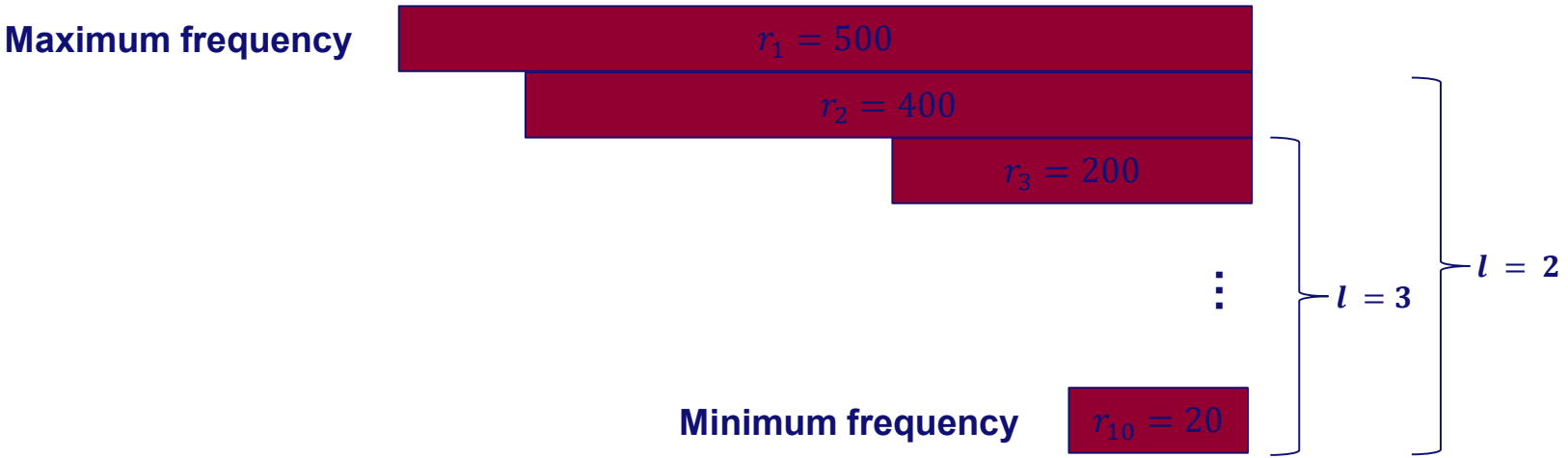
- **Entropy l-diversity.**
  - **The entropy of an equivalence class E is defined to be**
    - $Entropy(E) = -\sum_{s \in S} p(E,s) log(p(E,s))$
    - **In which S is the domain of the sensitive attribute, and $p(E,s)$ is the fraction of records in E that have sensitive value s.**
  - **A table is said to have entropy l-diversity if for every equivalence class E, $Entropy(E) \geq log(l)$. (This corresponds to l values equally distributed.)**
  - **In order to have entropy l-diversity for each equivalence class, the entropy of the entire table must be at least $\log(l)$.**
  - **Sometimes this is too restrictive, as the entropy of the entire table may be low if a few values are very common.**
    - **This leads to the recursive (c,l)-diversity.**

See the earlier definition of entropy in the context of information gain.

Group Based Anonymization: l-diversity

- **Recursive (c,l)-diversity:**
  - It makes sure that the most frequent value does not appear too frequently, and the less frequent values do not appear too rarely.
  - Let $s_1, s_2, \dots, s_m$ be the possible values of the sensitive attribute S in an equivalence class. Assume that we sort the counts of possible values in descending order and name the elements of the resulting sequence $r_1, r_2, \dots, r_m$.
  - An equivalence class is (c,l)-diverse if $r_1 < c(r_l + r_{l+1} + \cdots + r_m)$ for some user-specific constant c.
  - A table is said to have recursive (c,l)-diversity if all of its equivalence classes have recursive (c,l)-diversity.

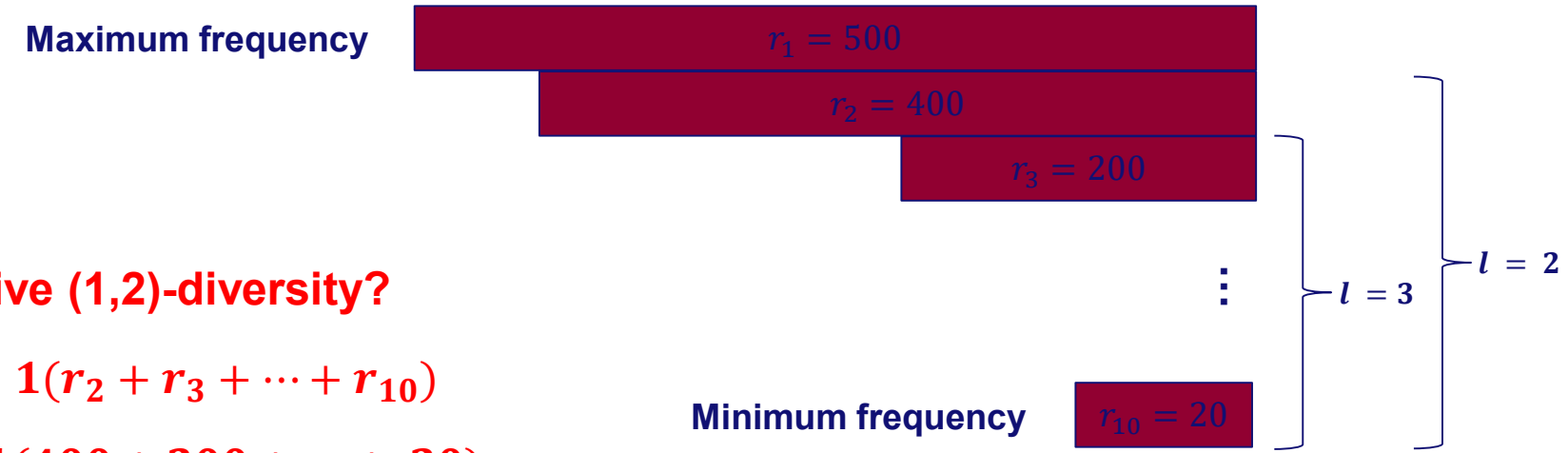
Group Based Anonymization: l-diversity



(c,l)-diverse if $r_1 < c(r_l + r_{l+1} + \cdots + r_m)$

Group Based Anonymization: l-diversity

(c,l)-diverse if $r_1 < c(r_l + r_{l+1} + \cdots + r_m)$



Recursive (1,2)-diversity?

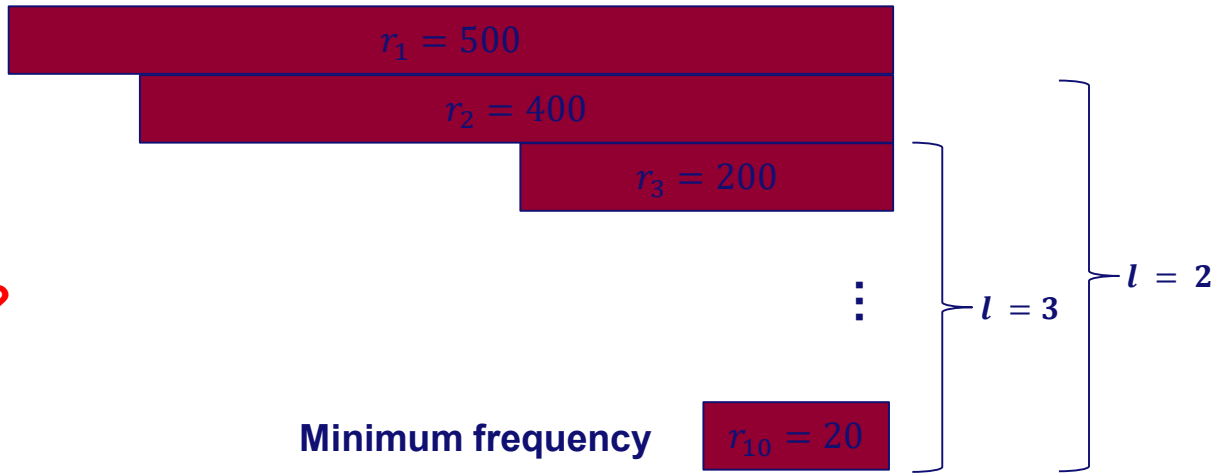
If: $r_1 < 1(r_2 + r_3 + \cdots + r_{10})$

$500 < 1(400 + 200 + \ldots + 20)$

Hence, (1,2)-diversity does hold.

Group Based Anonymization: l-diversity

(c,l)-diverse if $r_1 < c(r_l + r_{l+1} + \cdots + r_m)$



Recursive (5,10)-diversity?

If: $r_1 < 5(r_{10})$

But $500 \not< 5(20) = 100$

Hence, (5,10)-diversity does not hold.

Limitations of l-diversity

l-diversity may be difficult and unnecessary to achieve.

- **Suppose that the original data has only one sensitive attribute which is the test result for a particular virus. It takes two values: positive and negative.**
- **Further suppose that there are 10000 records, with 99% of them being negative, and only 1% being positive.**
- **Distinct l-diversity: The are only two different values, so we can never do better than distinct 2-diversity.**
- **Entropy l-diversity: The entropy will be close to 0, so we can never do better than distinct 1-diversity.**
- **Recursive (c,l)-diversity: Similar issues, need very high c value.**

Limitations of l-diversity

l-diversity is insufficient to prevent attribute disclosure.

- **Similarity Attack: When the sensitive attribute values in an equivalence class are distinct but semantically similar, an adversary can learn important information.**

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

One knows that Bob's record corresponds to one of the first three records, then one knows that Bob's salary is in the range [3K–5K] and can infer that Bob's salary is relatively low.

Knowing that Bob's record belongs to the first equivalence class enables one to conclude that Bob has some stomach-related problems

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

This leakage of sensitive information occurs because l-diversity does not take into account the semantical closeness of these values

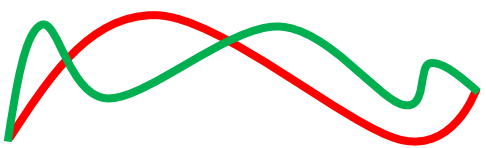
Group Based Anonymization: t-closeness

- **An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness.**

$$Distance(DC, DT) \leq \boldsymbol{t}$$

- DC: Distribution of a sensitive attribute in the class
- DT: Distribution of the attribute in the whole table

- **Distance measure is supposed to reflect the sematic distance among values.**
  - **Earth Mover's Distance (EMD) is used to this purpose.**



Chair of Process and Data Science

Other Types of Group Based Anonymization

- **Personalized privacy-preservation**
  - Since not all individuals or entities are equally concerned about their privacy, we may wish to treat the records in a given data set very differently for anonymization purposes.

- **Utility-based privacy-preservation**
  - The process of privacy-preservation leads to loss of information for data mining purposes. This loss of information can also be considered a loss of utility for data mining purposes.
  - The main idea is to anonymize the data in such a way that it remains useful for particular kinds of data mining or database applications.
  - E.g. the best analysis possible given a “privacy budget”.

Distributed Privacy-Preserving

(just the concept)

Distributed Privacy-Preserving

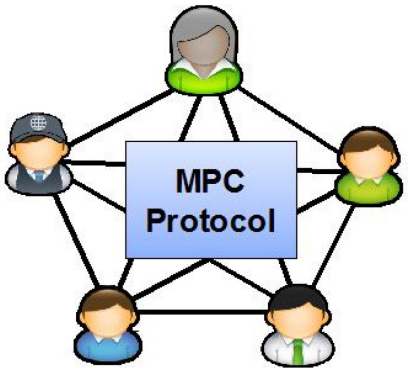
- **The key goal in most distributed methods for privacy-preserving data mining is to allow computation of useful aggregate statistics over the entire data set without compromising the privacy of the individual data sets within the different participants.**

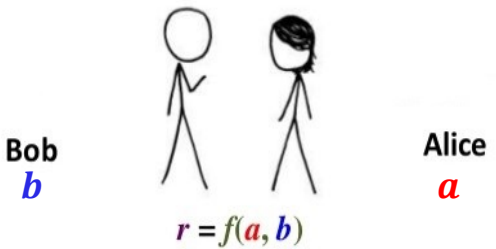
- **The problem of distributed privacy-preserving data mining overlaps closely with a field in cryptography for determining secure Multi-Party Computations (MPC).**

Distributed Privacy-Preserving

- Alice and Bob may have two inputs x and y respectively, and may wish to both compute the function $f(x, y)$ without revealing x or y to each other. This problem can also be generalized across k parties by designing the k argument function $\mathbf{h}(x_1, x_2, \ldots, x_k)$.

- In order to compute $f(x, y)$ or $\mathbf{h}(x_1, x_2, \ldots, x_k)$, a protocol has to be designed for exchanging information in such way that the function is computed without compromising privacy.





Chair of Process and Data Science

Distributed Privacy-Preserving

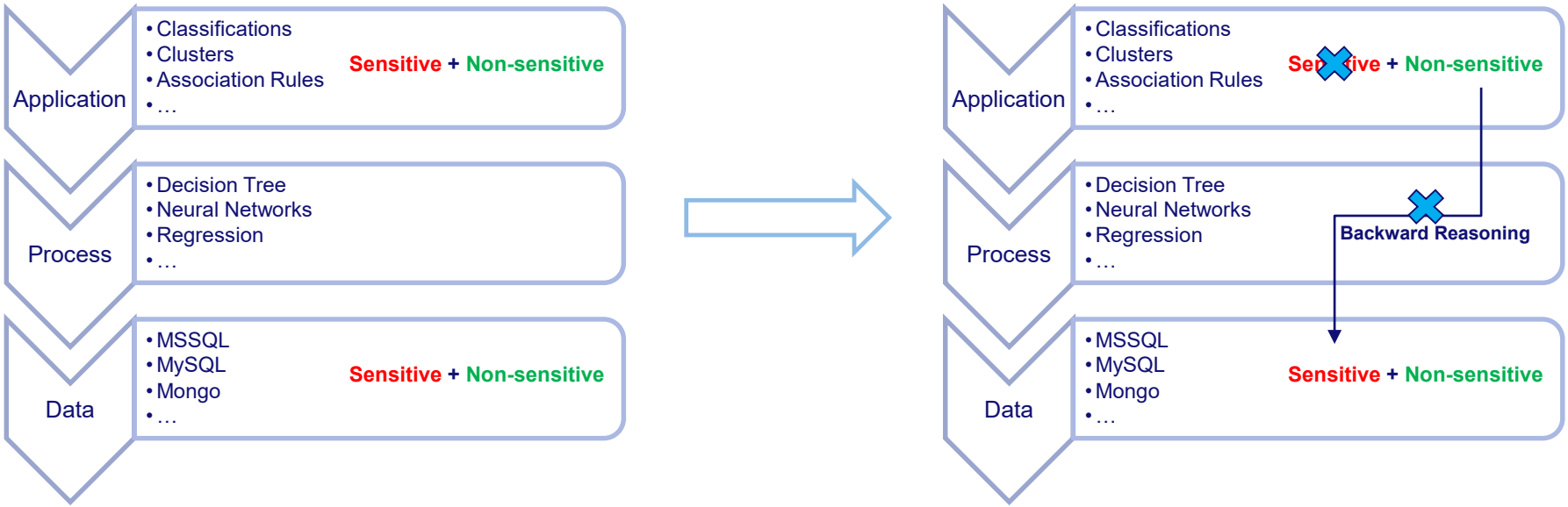
- The data sets may either be horizontally partitioned or be vertically partitioned.
  - In **horizontally** partitioned data sets, the individual records are spread out across multiple entities, each of which have the same set of attributes.
  - In **vertical** partitioning, the individual entities may have different attributes (or views) of the same set of records.
- Both kinds of partitioning pose different challenges to the problem of distributed privacy-preserving data mining.

Privacy-Loss Through Analysis Results

Privacy-Loss Through Analysis Results

- In many cases, the output of applications can be used by an adversary in order to make significant inferences about the behavior of the underlying data.
  - For example, in cases, where commercial data needs to be shared, the association rules may represent sensitive information for target-marketing purposes, which needs to be protected from inference.
- The key goal here is to prevent adversaries from making inferences from the end results of data science results.

Privacy-Preserving of Application Results



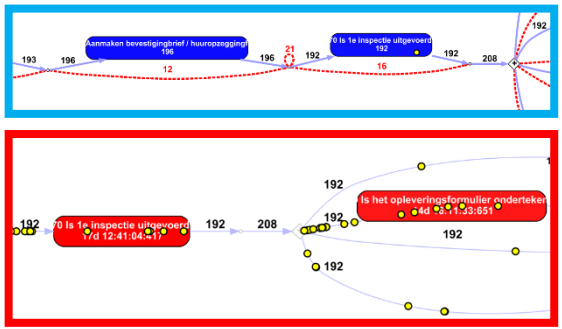
Chair of Process and Data Science

Responsible Process Mining


process mining will make things better, faster, more efficient, more effective, cheaper, …
…, but with great power comes great responsibility!!

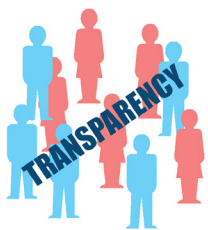
Responsible Process Mining

Who to blame?

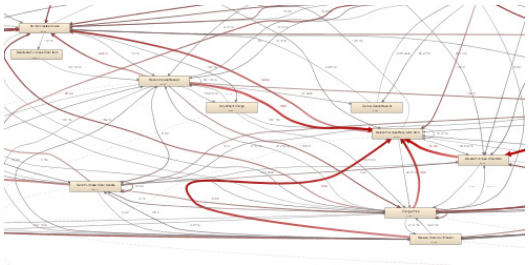


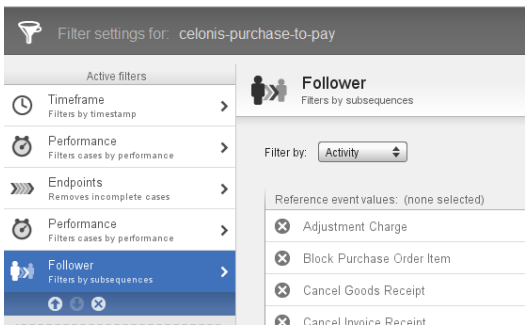
case id	event id	properties				
		timestamp	activity	resource	cost	…
1	35654423	30-12-2010:11.02	register request	Pete	50	…
	35654424	31-12-2010:10.06	examine thoroughly	Sue	400	…
	35654425	05-01-2011:15.12	check ticket	Mike	100	…
	35654426	06-01-2011:11.18	decide	Sara	200	…
	35654427	07-01-2011:14.24	reject request	Pete	200	…
2	35654483	30-12-2010:11.32	register request	Mike	50	…
	35654485	30-12-2010:12.12	check ticket	Mike	100	…
	35654487	30-12-2010:14.16	examine casually	Pete	400	…
	35654488	05-01-2011:11.22	decide	Sara	200	…
	35654489	08-01-2011:12.05	pay compensation	Ellen	200	…
3	35654521	30-12-2010:14.32	register request	Pete	50	…
	35654522	30-12-2010:15.06	examine casually	Mike	400	…
	35654524	30-12-2010:16.34	check ticket	Ellen	100	…

Event data are super sensitive



Always a model is returned





Filter until you get what you want

Chair of Process and Data Science

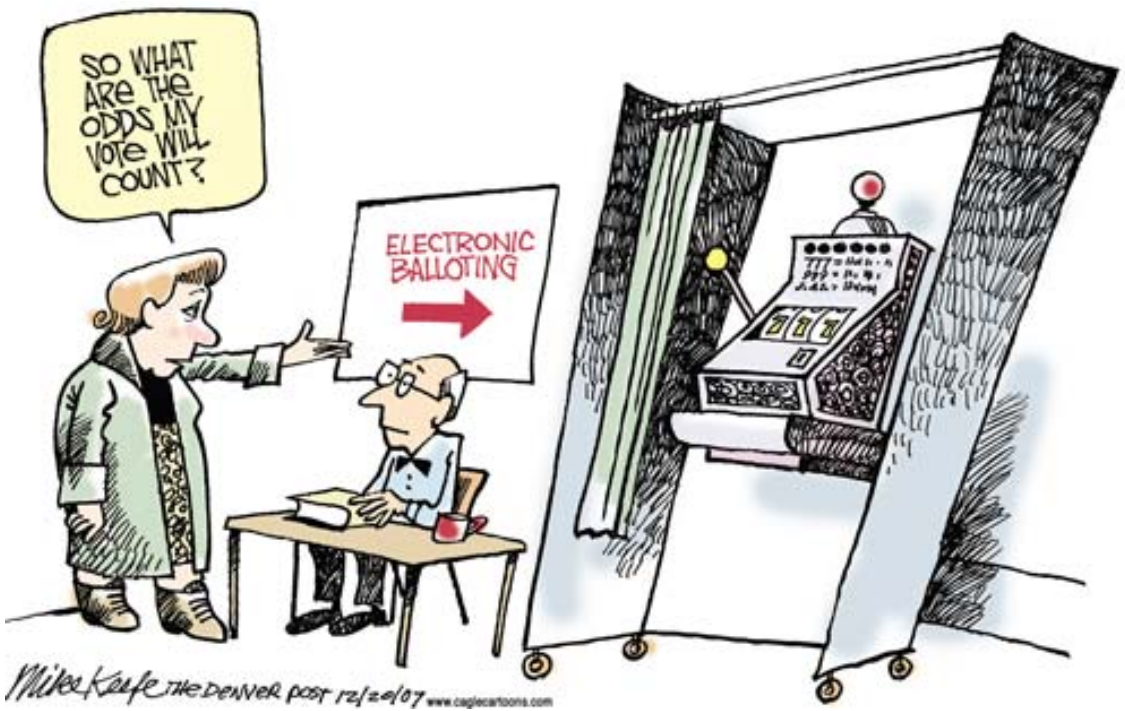
Confidentiality in Process Mining (work by Majid Rafiei)

- On the one hand, recent breakthroughs in process mining resulted in powerful techniques, encouraging organizations and business owners to improve their processes through process mining.

- On the other hand, there are great concerns about the use of highly sensitive event data.

The goal is to find an approach that allows us to hide confidential information in a controlled manner while ensuring that the desired process mining results can still be obtained.

Like e-voting

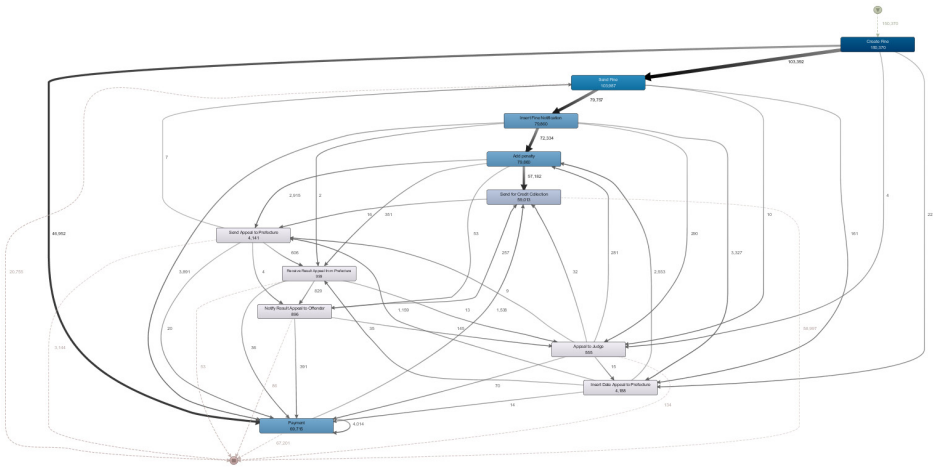


- Ensure that each event “counts”.
- Able to verify the correctness of the result (e.g., process model).
- Not revealing sensitive information (i.e., individual events).

Chair of Process and Data Science

Process Mining on Encrypted Event Data

- **Break correlations, e.g., [John:abcd, Pete:acbd, Mary:aed, Sue:abcd, Oscar:acbd], becomes [>a^5, d>^{5},ab^2,bc^2,cd^2, ac^2,cb^2,bd^2, ae,ed].**
- **Can still produce proper process models:**



Chair of Process and Data Science

Confidentiality in Process Mining

- Suppose that the following table is part of event log about surgeries in a hospital.
- Further assume that all the attributes are encrypted by a deterministic encryption method (monotonic).
- Given background knowledge about surgeons and approximate cost, is there any data leakage?

Case ID	Activity	Resource	Timestamp	Cost
1ab	Abc1dfg	0fgh14	123	5000
2cd	Chf5jkl	024sdfk	125	6000
3ty	215sfs0	.543s1s	254	3500
1tu	2154@3	3242s2	248	2000
1za	321$22	02315d	157	5500

Confidentiality in Process Mining

Some attacks

- **Cost analysis**
  - **The most expensive or cheapest costs along with background knowledge can be used to infer sensitive information.**
  - **For example, one can find out that the most expensive surgery was a “brain surgery” by “Dr. John” on date “10/01/2018”, and the patient name was “Jack”.**
- **Frequency Mining**
  - **The most/less frequent activities along with background knowledge can be used to infer sensitive information.**
- **Exploring Order/Position of Activities**
- **…**

Case ID	Activity	Resource	Timestamp	Cost
1ab	Abc1dfg	0fgh14	123	5000
2cd	Chf5jkl	024sdfk	125	6000
3ty	215sfs0	.543s1s	254	3500
1tu	2154@3	3242s2	248	2000
1za	321$22	02315d	157	5500

Confidentiality in Process Mining

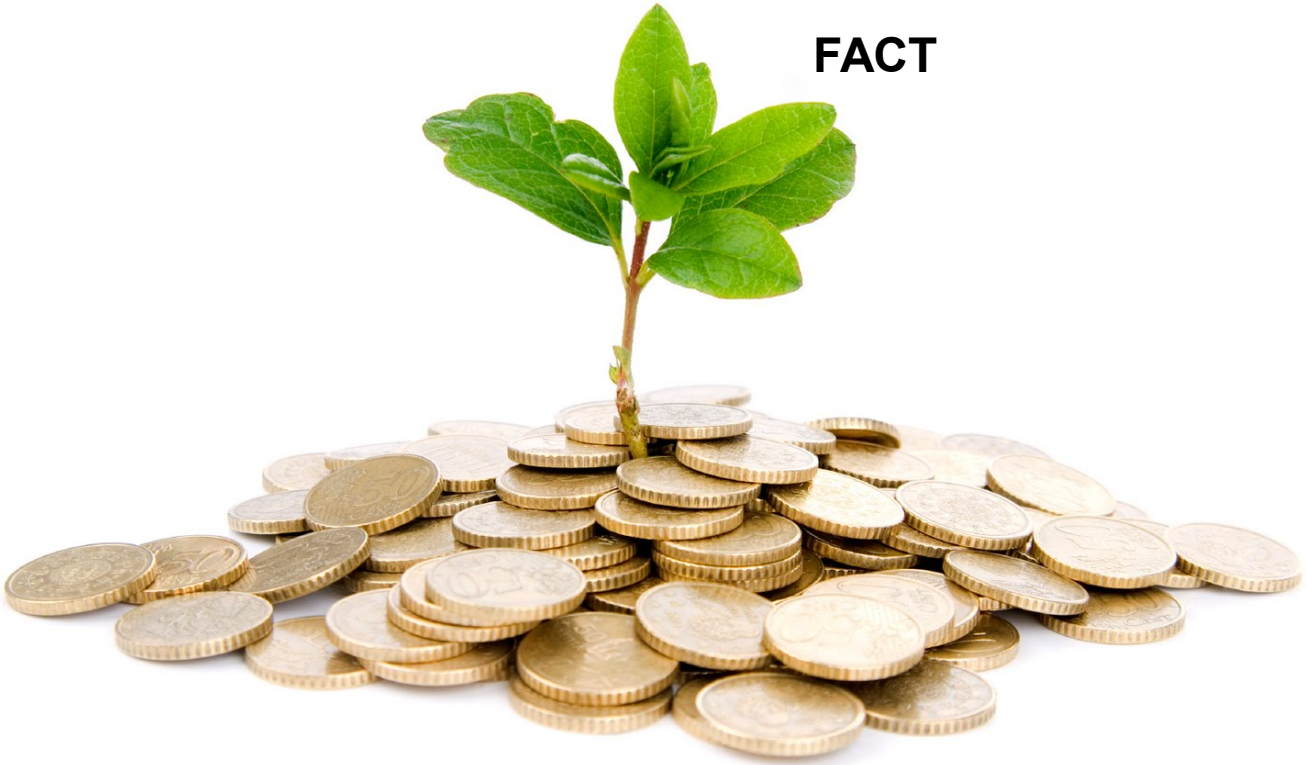
Assume the event data are encrypted safely, there are other challenges:

- **Encrypted Results**
  - **Since results are encrypted, the data analyst is not able to interpret the results (without decryption).**
- **Impossibility of Accuracy Evaluation**
  - **How can we make sure that a result of the encrypted event log is the same as the result of the plain event log? (without decryption)**
- **Cryptography is a resource consuming activity, and decryption is even much more expensive (time, energy, etc.) than encryption.**

Responsible Data Science implies handling tradeoffs subjectively (i.e., by human judgment)!

Conclusion

Responsible Data Science



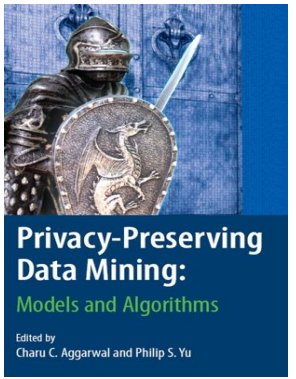
Responsible Data Science

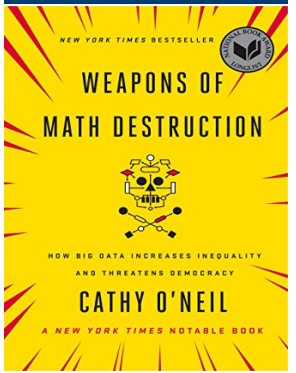
Next to giving an overview of the field (RDS/FACT), we focused on:

- **Fairness**
  - **How to measure?**
  - **How to ensure fairness (e.g., making a decision tree fair)?**
- **Confidentiality**
  - **Cryptography**
  - **Randomization**
  - **Anonymization (k-anonymity, l-diversity, t-closeness, etc.)**
  - **Distribution**
  - **Results**
- **Example: Process Mining**

Relevant Literature

- van der Aalst, Wil MP. "Responsible data science: using event data in a “people friendly” manner." *International Conference on Enterprise Information Systems*. Springer, Cham, 2016.
- Pedreshi, Dino, Salvatore Ruggieri, and Franco Turini. "Discrimination-aware data mining." *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008.
- Kamiran, Faisal, Toon Calders, and Mykola Pechenizkiy. "Discrimination aware decision tree learning." *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 2010.
- van der Aalst, Wil MP. "Responsible data science." *Business and Information Systems Engineering*. Springer, Cham, 2017.
- Kashid, Asmita, Vrushali Kulkarni, and Ruhi Patankar. "Discrimination-aware data mining: a survey." *International Journal of Data Science* 2.1 (2017): 70-84.
- Sweeney, Latanya. "k-anonymity: A model for protecting privacy." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002): 557-570.
- Machanavajjhala, Ashwin, et al. “L-Diversity: Privacy Beyond k-Anonymity." *22nd International Conference on Data Engineering (ICDE'06)*, IEEE, 2006.
- Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity." *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 2007.
- Aggarwal, Charu C., and S. Yu Philip. "A general survey of privacy-preserving data mining models and algorithms." *Privacy-preserving data mining*. Springer, Boston, MA, 2008. 11-52.
- Agrawal, Rakesh, and Ramakrishnan Srikant. *Privacy-preserving data mining*. Vol. 29. No. 2. ACM, 2000.
- Saygin, Yücel, Vassilios S. Verykios, and Chris Clifton. "Using unknowns to prevent discovery of association rules." *ACM Sigmod Record* 30.4 (2001): 45-54.
- Evfimievski, Alexandre, et al. "Privacy preserving mining of association rules." *Information Systems* 29.4 (2004): 343-364.
- Oliveira, Stanley RM, Osmar R. Zaiane, and Yücel Saygin. "Secure association rule sharing." *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, Berlin, Heidelberg, 2004.
- Lindell, Yehuda, and Benny Pinkas. "Privacy preserving data mining." *Annual International Cryptology Conference*. Springer, Berlin, Heidelberg, 2000.
- Verykios, Vassilios S., et al. "State-of-the-art in privacy preserving data mining." *ACM Sigmod Record* 33.1 (2004): 50-57.





Chair of Process and Data Science

#	Lecture	date	day
Lecture 1	Introduction	10/10/2018	Wednesday
Lecture 2	Crash Course in Python	11/10/2018	Thursday
Instruction 1	*Python*	*12/10/2018*	*Friday*
Lecture 3	Basic data visualisation/exploration	15/10/2018	Monday
Lecture 4	Decision trees	16/10/2018	Tuesday
Instruction 2	*Decision trees and data visualization*	*17/10/2018*	*Wednesday*
Lecture 5	Regression	18/10/2018	Thursday
Lecture 6	Support vector machines	19/10/2018	Friday
Instruction 3	*Regression and support vector machi*	*22/10/2018*	*Monday*
Lecture 7	Neural networks (1/2)	23/10/2018	Tuesday
Instruction 4	*Neural networks and supervised learr*	*24/10/2018*	*Wednesday*
Lecture 8	Neural networks (2/2)	25/10/2018	Thursday
Lecture 9	Evaluation of supervised learning prob	26/10/2018	Friday
Instruction 5	*Neural networks and supervised learr*	*29/10/2018*	*Monday*
Lecture 10	Clustering	30/10/2018	Tuesday
Lecture 11	Frequent items sets	31/10/2018	Wednesday
Lecture 12	Association rules	01/11/2018	Thursday
Lecture 13	Sequence mining	02/11/2018	Friday
Instruction 6	*Clustering, frequent items sets, associ*	*05/11/2018*	*Monday*
Lecture 14	Process mining (unsupervised)	06/11/2018	Tuesday
Lecture 15	Process mining (supervised)	07/11/2018	Wednesday
Instruction 7	*Process mining and sequence mining*	*30/11/2018*	*Friday*
Lecture 16	Text mining (1/2)	05/12/2018	Wednesday
Instruction 8	*Text mining and process mining*	*06/12/2018*	*Thursday !!*
Lecture 17	Text mining (2/2)	12/12/2018	Wednesday
Lecture 18	Data preprocessing, data quality, binning, etc.	13/12/2018	Thursday
Lecture 19	Visual analytics & information visualization	19/12/2018	Wednesday
backup		20/12/2018	Thursday
Instruction 9	*Text mining, preprocessing and visualization*	*21/12/2018*	*Friday*
Lecture 20	Responsible data science (1/2)	09/01/2019	Wednesday
Lecture 21	Responsible data science (2/2)	10/01/2019	Thursday
Instruction 10	*Responsible data science*	*11/01/2019*	*Friday*
Lecture 22	Big data (1/2)	16/01/2019	Wednesday
Lecture 23	Big data (2/2)	17/01/2019	Thursday
Instruction 11	*Big data*	*18/01/2019*	*Friday*
Lecture 24	Closing	23/01/2019	Wednesday
backup		24/01/2019	Thursday
Instruction 12	*Example exam questions*	*25/01/2018*	*Friday*
backup		30/01/2019	Wednesday
backup		31/01/2019	Thursday
extra	*Question hour*	*01/02/2019*	*Friday*

Lecture 20	Responsible data science (1/2)	09/01/2019	Wednesday
Lecture 21	Responsible data science (2/2)	10/01/2019	Thursday
Instruction 10	*Responsible data science*	*11/01/2019*	*Friday*
Lecture 22	Big data (1/2)	16/01/2019	Wednesday
Lecture 23	Big data (2/2)	17/01/2019	Thursday
Instruction 11	*Big data*	*18/01/2019*	*Friday*
Lecture 24	Closing	23/01/2019	Wednesday
backup		24/01/2019	Thursday
Instruction 12	*Example exam questions*	*25/01/2018*	*Friday*
backup		30/01/2019	Wednesday
backup		31/01/2019	Thursday
extra	*Question hour*	*01/02/2019*	*Friday*