Karlstad 2016

Big Data Privacy & Anonymisation,

Vicenç Torra

August, 2016

School of Informatics, University of Skövde, Sweden

Outline

Outline

- Anonymization (masking methods)
- and big data
- Data provenance and privacy

Outline

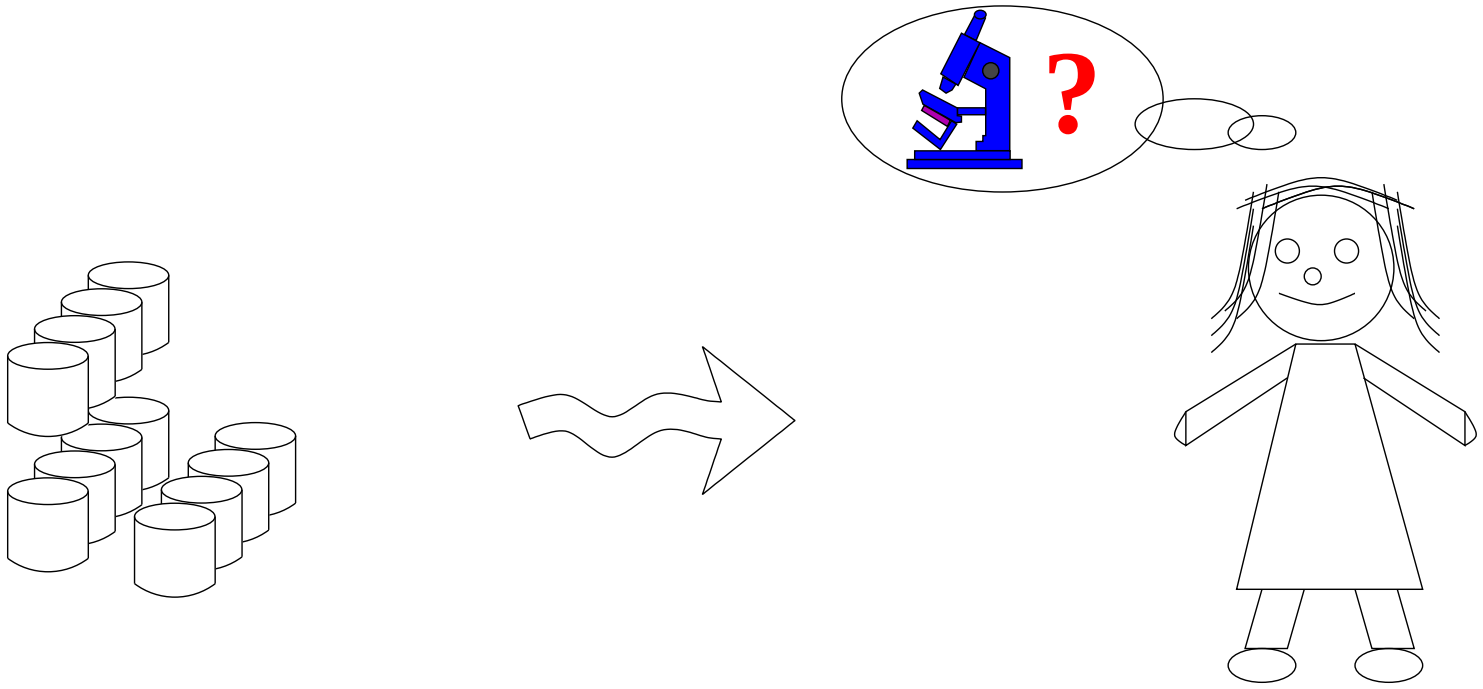
Introduction

Introduction

Introduction

Classification w.r.t. our knowledge on the analysis of a third party

- Data-driven or general purpose (*analysis not known*)
  → anonymization methods / masking methods
- Computation-driven or specific purpose (*analysis known*)
  → cryptographic protocols, differential privacy
- Result-driven (*analysis known: protection of its results*)



Introduction

Classification w.r.t. our knowledge on the analysis of a third party

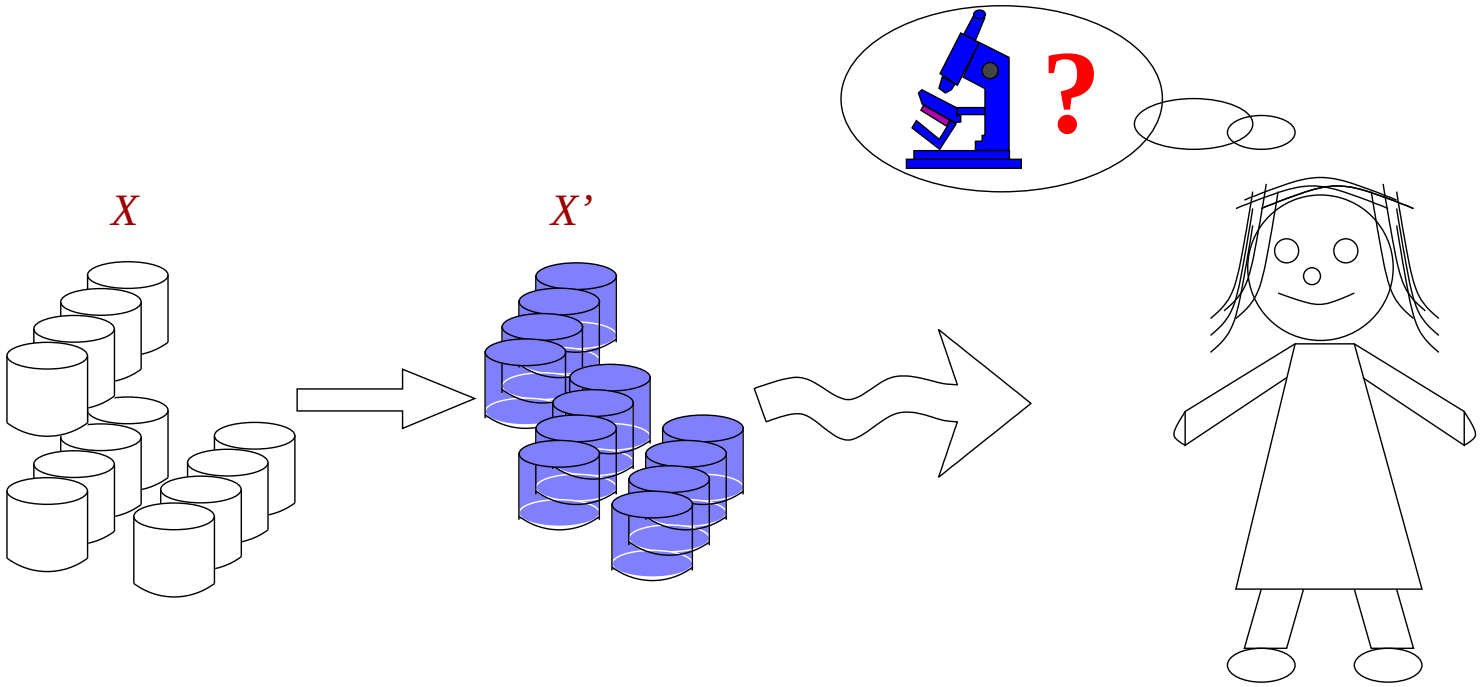
- Data-driven or general purpose (*analysis not known*)
  $\rightarrow$ anonymization methods / masking methods
  - Example: blood glucose level prediction for diabetes
    need the data, but unclear the type of model to be built
    neural networks, suport vector machines, decision trees, ...

Anonymization: Masking methods

Anonymization: Masking methods

Anonymization: Masking methods

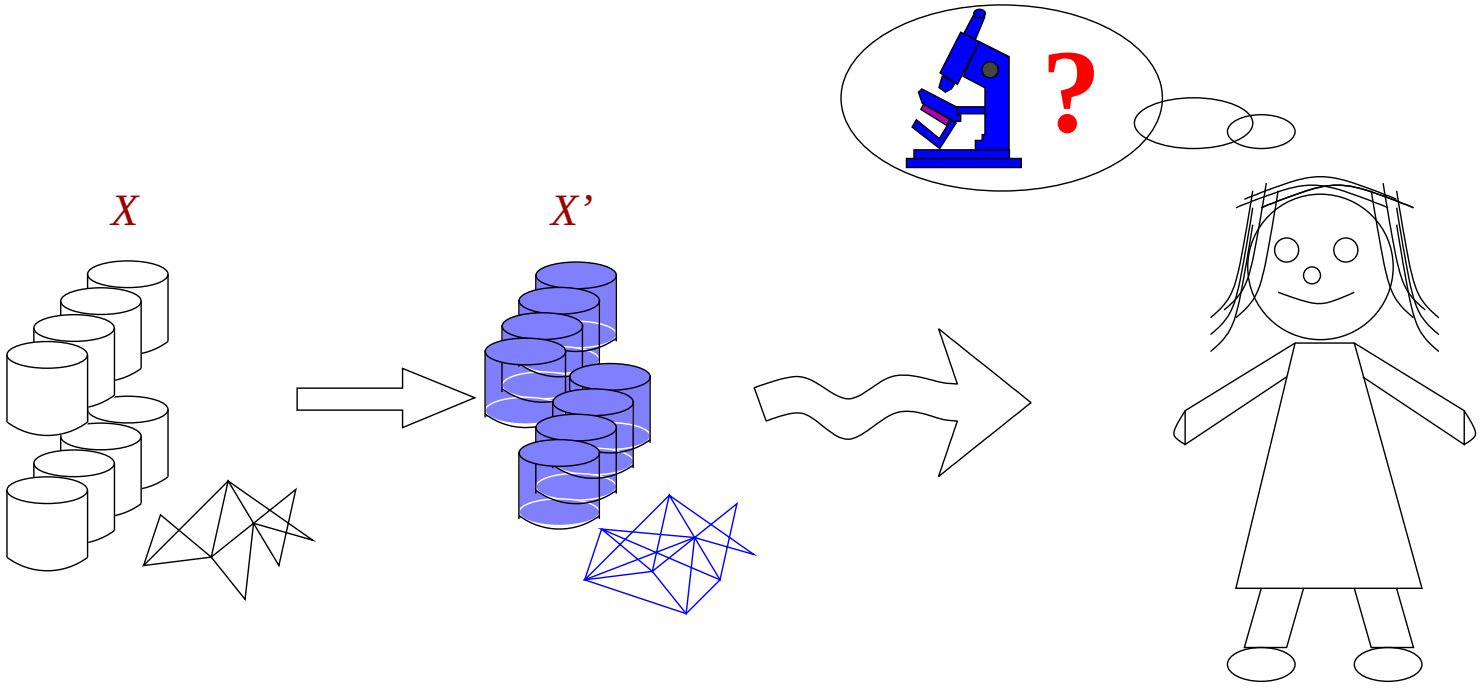
Anonymization/masking method: Given a data file X compute a file X' with data of *less quality*.



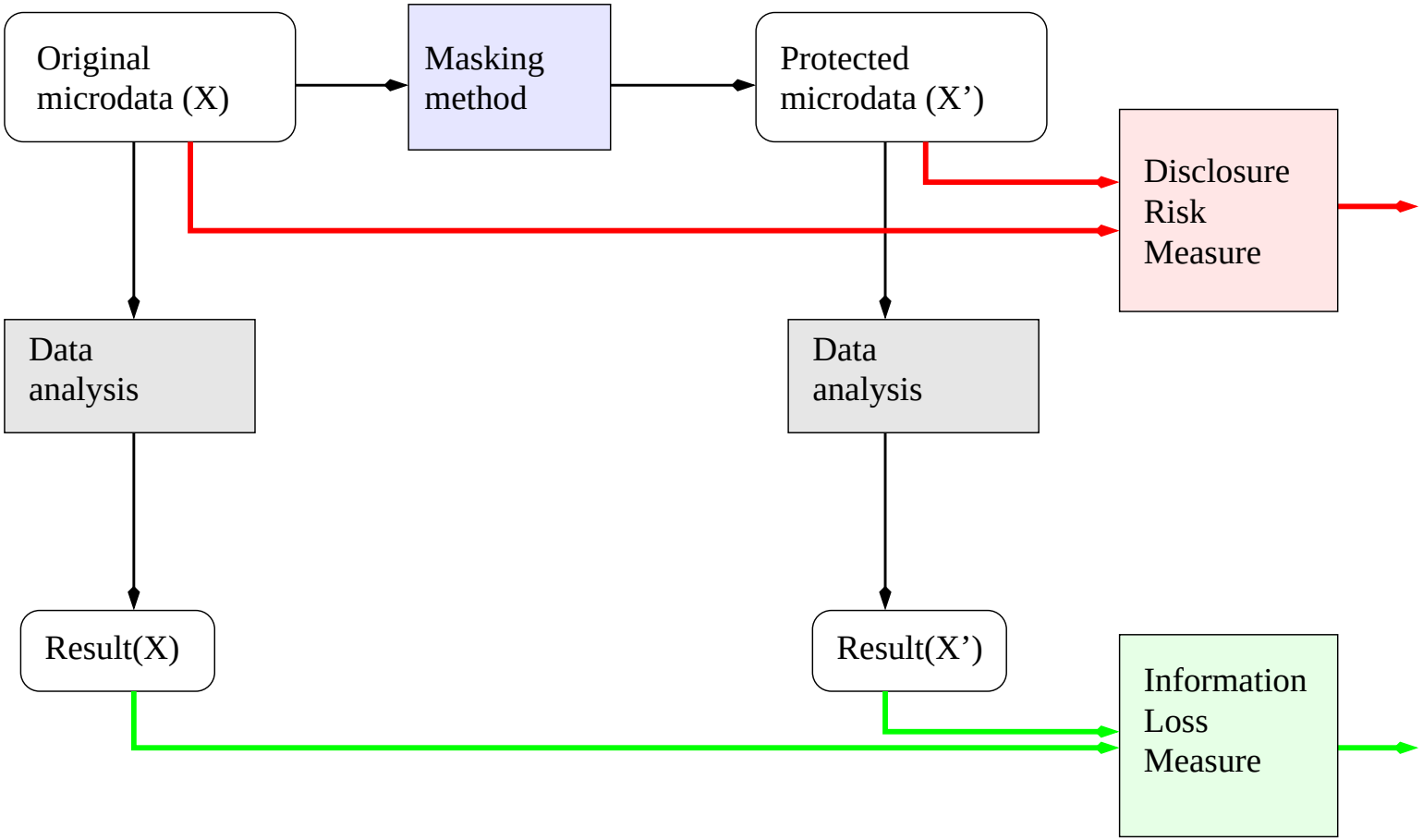
Masking methods

Approach valid for different types of data

- Databases, documents, search logs, social networks, . . .
  (also masking taking into account semantics: wordnet, ODP)



Research questions



Masking methods

Masking methods. (anonymization methods)

Masking methods

Masking methods. (anonymization methods)

- Perturbative. (less quality=erroneous data)
  E.g. noise addition/multiplication, microaggregation, rank swapping

Masking methods

Masking methods. (anonymization methods)

- Perturbative. (less quality=erroneous data)
  E.g. noise addition/multiplication, microaggregation, rank swapping
- Non-perturbative. (less quality=less detail)
  E.g. generalization, suppression

Masking methods

Masking methods. (anonymization methods)

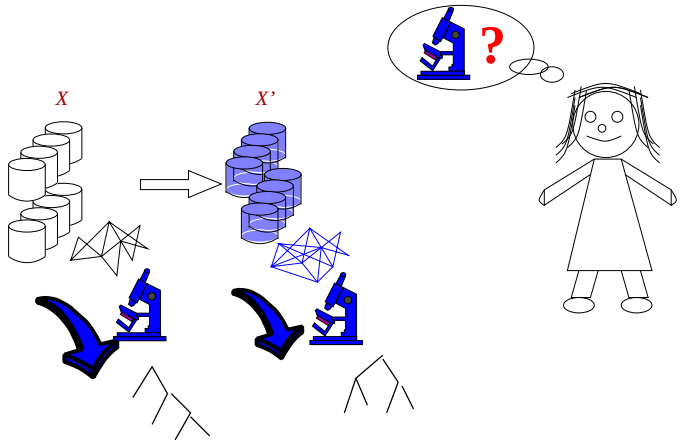
- Perturbative. (less quality=erroneous data)
  E.g. noise addition/multiplication, microaggregation, rank swapping
- Non-perturbative. (less quality=less detail)
  E.g. generalization, suppression
- Synthetic data generators. (less quality=not real data)
  E.g. (i) model from the data; (ii) generate data from model

Information loss

Information loss measures. Compare X and X' w.r.t. analysis (f)

$$IL_f(X, X') = divergence(f(X), f(X'))$$

- f: generic vs. specific (data uses)
  - Statistics
  - Machine learning: Clustering and classification
    For example, classification using decision trees
  - ... specific measures for graphs



Disclosure risk

Disclosure risk.

- Identity disclosure vs. Attribute disclosure
  - Attribute disclosure: (e.g. learn about Alice's salary)
    - ⋆ Increase knowledge about an attribute of an individual
  - Identity disclosure: (e.g. find Alice in the database)
    - ⋆ Find/identify an individual in a masked file

Within artificial intelligence, some attribute disclosure is expected.

Disclosure risk assesment

Disclosure risk.

- Identity disclosure vs. Attribute disclosure
- Boolean vs. quantitative measures
  threshold vs. risk measurement

Disclosure risk assesment

Disclosure risk.

- Identity disclosure vs. Attribute disclosure
- Boolean vs. quantitative measures
  threshold vs. risk measurement
  (minimize information loss vs. multiobjetive optimization IL/DR)

Disclosure risk assesment

Disclosure risk.

- Identity disclosure vs. Attribute disclosure
- Boolean vs. quantitative measures
  threshold vs. risk measurement
  (minimize information loss vs. multiobjetive optimization IL/DR)

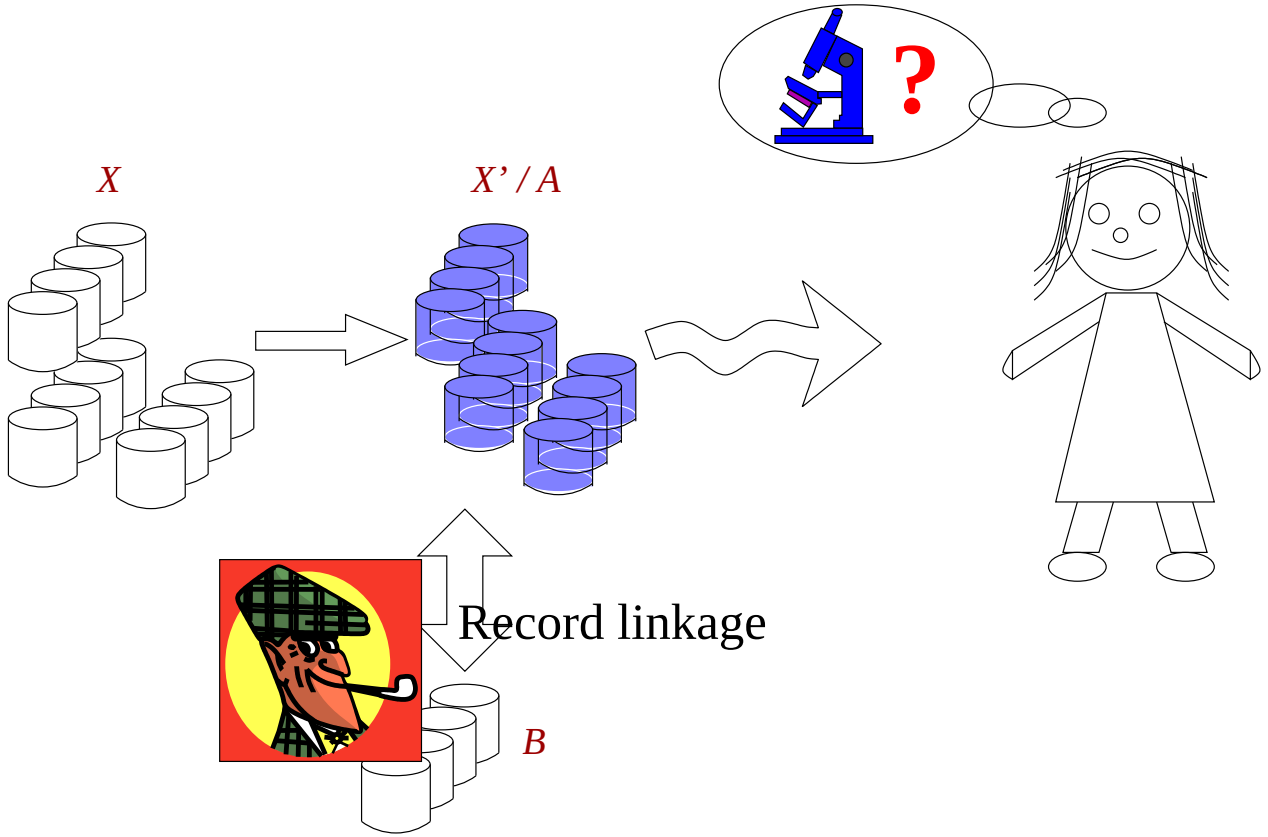
Examples. Privacy models / disclosure risk measures

	Attribute disclosure	Identity disclosure
Boolean	Differential privacy	k−Anonymity
Quantitative	Interval disclosure	Re−identification (record linkage) Uniqueness

Disclosure risk assesment

A scenario for identity disclosure: Reidentification

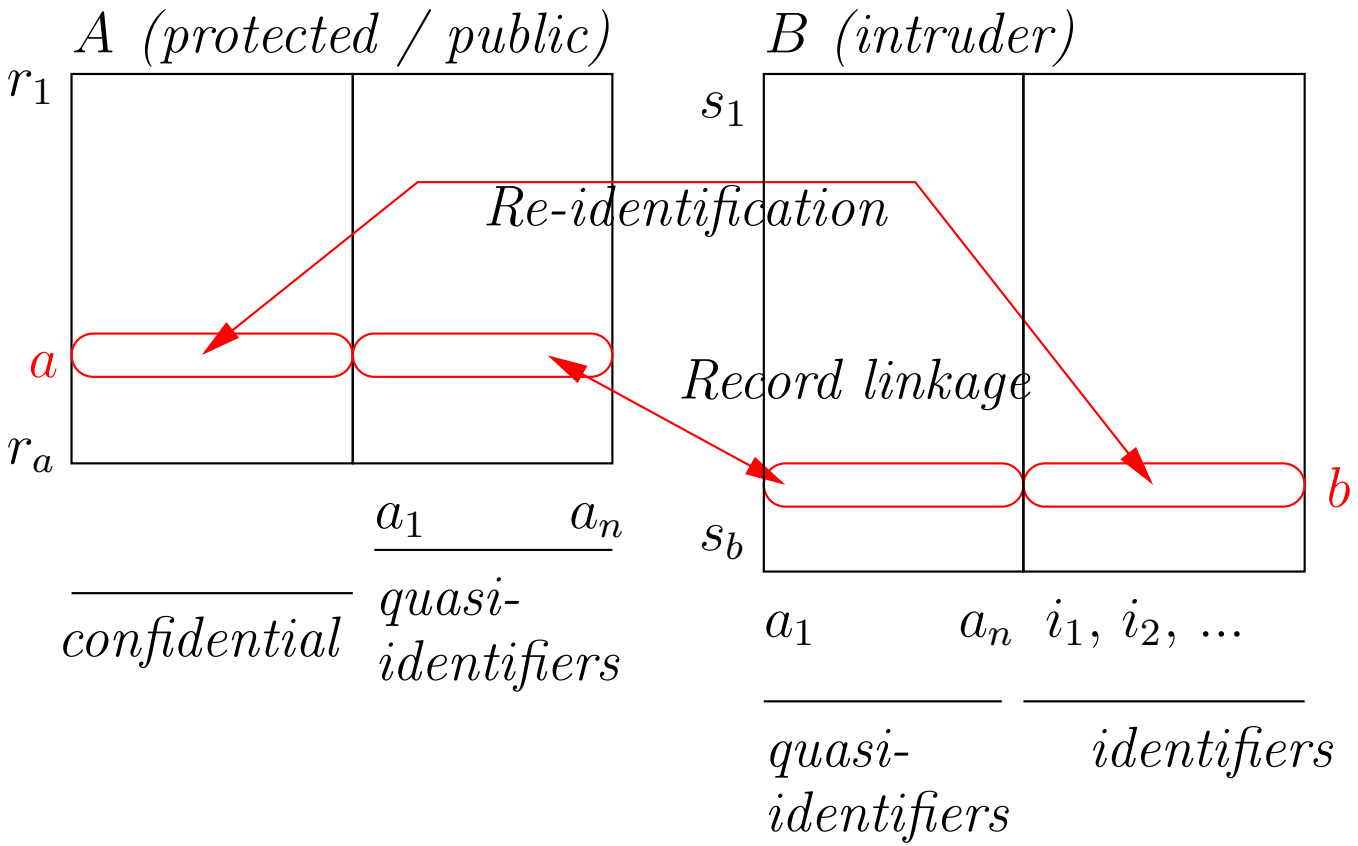
- A: File with the protected data set
- B: File with the data from the intruder (subset of original X)



Disclosure risk assesment

A scenario for identity disclosure: Reidentification

- A: File with the protected data set
- B: File with the data from the intruder (subset of original X)



Disclosure risk assesment

A scenario for identity disclosure. Reidentification

- Flexible scenario. Different assumptions on what available
  E.g., only partial information on individuals/characteristics
- Worst-case scenario for disclosure risk assessment
  (upper bound of disclosure risk)

Disclosure risk assesment

A scenario for identity disclosure. Reidentification

- Flexible scenario. Different assumptions on what available
  E.g., only partial information on individuals/characteristics
- Worst-case scenario for disclosure risk assessment
  (upper bound of disclosure risk)
  - Maximum information

Disclosure risk assesment

A scenario for identity disclosure. Reidentification

- Flexible scenario. Different assumptions on what available
  E.g., only partial information on individuals/characteristics
- Worst-case scenario for disclosure risk assessment
  (upper bound of disclosure risk)
  - Maximum information
  - Most effective reidentification method

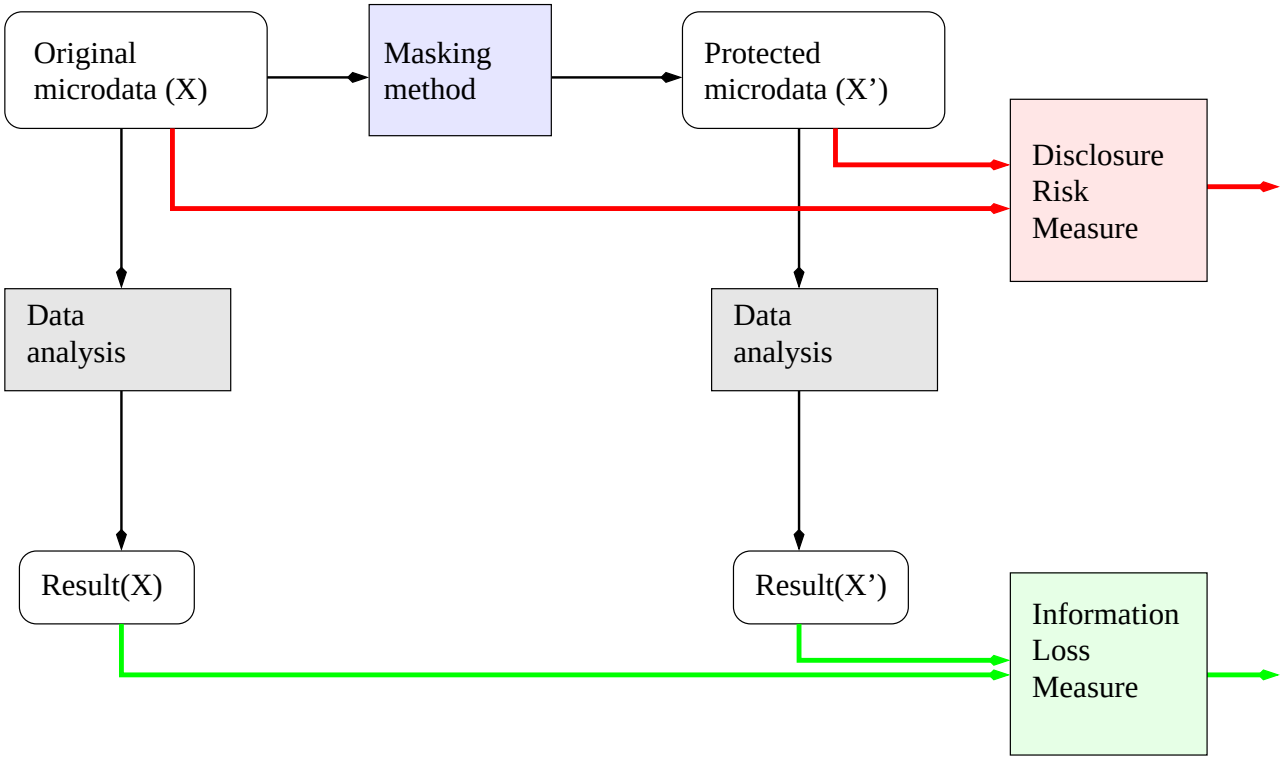
Disclosure risk assesment

A scenario for identity disclosure. Reidentification

- Flexible scenario. Different assumptions on what available
  E.g., only partial information on individuals/characteristics
- Worst-case scenario for disclosure risk assessment
  (upper bound of disclosure risk)
  - Maximum information: Use original file to attack
  - Most effective reidentification method: Use ML
    Use information on the masking method (transparency)

Anonymization: summary

Summary.



Big Data and Anonymization

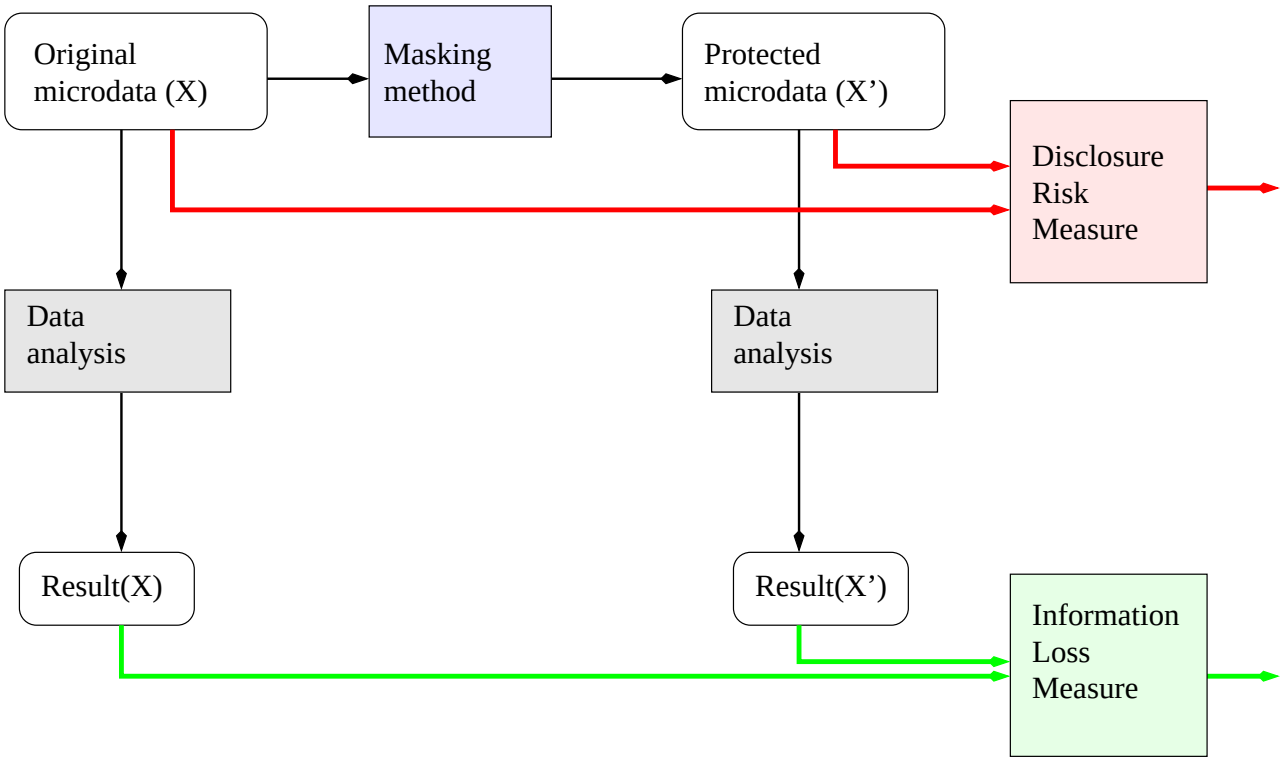
Big Data

Big Data. Definitions based on 3Vs (or 4Vs, 5Vs, etc)

- Volume. Huge amounts of data
  Facebook generates 4 new petabytes of data per day (oct. 2014)
- Velocity. Real time streams of data flowing from diverse resources.
  Either from sensors or from internet (from e-commerce or social media)
- Variety. Data from a vast range of systems and sensors, in different formats and datatypes
  Including (unstructured) text, logs and video

Big Data and Privacy

Big data and privacy. Similar discussion? yes (quick answer)



Big Data and Privacy

Big data and privacy. What's new with big data?

Big Data and Privacy

Big data and privacy. What's new with big data?

new level for privacy risk

Big Data and Privacy

Big data and privacy. What's new with big data?

new level for privacy risk

Difficulties.

- Lack of control and transparency. Who has my data?
  Data from sensors and cameras, screening posts in social networks, analysis of web searches, tracking cookies, data brokers

Big Data and Privacy

Big data and privacy. What's new with big data?

new level for privacy risk

Difficulties.

- Lack of control and transparency. Who has my data?
  Data from sensors and cameras, screening posts in social networks, analysis of web searches, tracking cookies, data brokers
- Data reusability
  Big data analytics' main goal: use data for new purposes

Big Data and Privacy

Big data and privacy. What's new with big data?

new level for privacy risk

Difficulties.

- Lack of control and transparency. Who has my data?
  Data from sensors and cameras, screening posts in social networks, analysis of web searches, tracking cookies, data brokers
- Data reusability
  Big data analytics' main goal: use data for new purposes
- Data inference and re-identification
  Linking databases increase the risk of identification
  Effective inference algorithms: inference of sensitive attributes

Big Data and Privacy

Claims or research issues. Issue #1.

Technology should help people to know what others know/infer about them.

- Effective (ML/DM) inference algorithms can infer sexual orientation, and political or religious affiliation[1].
- It is *useless* that we protect sensitive information, without protecting what permits to infer sensitive information.

[1] Kosinski, M., Stillwell, D., Graepel, T. (2013) Private traits and attributes are predictable from digital records of human behavior, PNAS. "The model's accuracy was lowest (60%) when inferring whether users' parents stayed together or separated before users were 21 y old".

Big Data and Privacy

Claims or research issues. Issue #2.

Databases should be anonymized/masked in origin.

- There already exist masking methods causing low information loss (and low disclosure risk).
- On machine learning (ML) and data mining (DM) algorithms
  - ML algorithms are resistant to errors.
  - Not all data is equally important for ML algorithms.
  - Big data mining algorithms do not always use all data (sampling).
  - Preprocessing in ML, dimensionality reduction, sampling, etc. should be combined with masking methods/can exploit the results of masking methods.
  - Study: ML and DM algorithms that lead to good models on masked data. Masked data can be seen as causing noise, but also as dimensionality reduction or noise reduction.

So, we do not really need all raw data for ML and DM algorithms.

Big Data and Privacy

Claims or research issues. Issue #3.

Anonymization needs to provide controlled linkability.

- Linkability is a basic requirement for big data
- How to ensure some level of linkability between databases while ensuring privacy?
- E.g., linkability at group level in k-anonymity.

Big Data and Privacy

Claims or research issues. Issue #4.

Privacy models need to be composable

- Composability. Given several data sets with privacy guarantees, their *combination* also satisfies the privacy guarantee.
- Results for differential privacy (positive) and k-anonymity (negative)

Big Data and Privacy

Claims or research issues. Issue #5.

User privacy should be in place (decentralized anonymity)

- Users anonymize their data in origin.
- Anonymized data is transferred to the data collector (or to the service provider)
- No need to trust the data collector
- Local anonymization and collaborative anonymization

Big Data and Privacy

Claims or research issues. Issue #6.

Need to deal with **big** data

- Large volumes of data
- Dynamic data
- Streaming data

Big Data and Privacy

Claims or research issues. Issue #6.1.

Need to deal with **big** data: Large volumes of data

- Efficient algorithms are being developed for data of high dimension. They include masking methods, IL and DR measures. E.g.,
  - Standard databases: microaggregation
  - Graphs and social networks: random noise, generalization, microaggregation
  - Location privacy

Big Data and Privacy

Claims or research issues. Issue #6.2.

Need to deal with **big** data: Dynamic data

- Data changes with respect to time, and data needs to be published regularly.
- Independent application of e.g. k-anonymity fails[2]
- Specific algorithms are being developed.

[2] In a class a single student born in February, at least 2 in the other months. Different releases can disclosure that there is a student born in February. From [ST, IJUFKS 2012]

Big Data and Privacy

Claims or research issues. Issue #6.3.

Need to deal with **big** data: Streaming data

- Difficulties due to the incompleteness of the information
- Methods based on sliding windows.

Data provenance

Data provenance

Privacy rights

- The right to amend and the right to be forgotten

Privacy rights

- The right to amend and the right to be forgotten
- Data provenance to implement these rights

Data provenance

- is the technology that permits to have the history of the data.

Advantages (beyond helping privacy *problems*)

- Improve data quality, permits accountability, and help users to assess the validity and trust of the information.

Provenance structures

- Annotations on the data
  - (38.2, Doctor Jekyll, 3/August/2016)
  - (180, mean(C1:C15), 4/August/2016)

Provenance structures

- Annotations on the data
  - (38.2, Doctor Jekyll, 3/August/2016)
  - (180, mean(C1:C15), 4/August/2016)
- They can be quite complex
  - data elements integrating several data elements, integration of several sources, application of complex models obtained from other data.

Provenance structures

- Annotations on the data
  - (38.2, Doctor Jekyll, 3/August/2016)
  - (180, mean(C1:C15), 4/August/2016)
- They can be quite complex
  - data elements integrating several data elements, integration of several sources, application of complex models obtained from other data.
- They can be quite large
  - they may duplicate (or more) the size of a database.

Provenance structures

- Annotations on the data
  - (38.2, Doctor Jekyll, 3/August/2016)
  - (180, mean(C1:C15), 4/August/2016)
- They can be quite complex
  - data elements integrating several data elements, integration of several sources, application of complex models obtained from other data.
- They can be quite large
  - they may duplicate (or more) the size of a database.
- They are sensitive and should not be forged
  - Who and when modified a data element may be confidential.

Provenance representation

- Fine grained vs coarse grained data provenance

Provenance representation

- Fine grained vs coarse grained data provenance
- Where and why provenance
  - Where provenance: the origin of the data
  - Why provenance: the process that generated the data

Provenance representation

- Fine grained vs coarse grained data provenance
- Where and why provenance
  - Where provenance: the origin of the data
  - Why provenance: the process that generated the data
- Chains and graphs
  - Chains: application of sequential processes
  - Graphs: more flexible, data from the same source is combined after different processing

Requirements for data provenance (processing) (or difficulties)

- Completeness.
  - All actions represented
- Efficiency.
  - Fine-grained provenance duplicates database size.
    Algorithms should be efficient.
- Not yet fully standardized (... less for big data)

Provenance and privacy

Privacy

- Privacy and security on the data provenance structures
- Privacy for sensitive data provenance
- Privacy beyond privacy of data provenance

Provenance and privacy

Privacy and security on the data provenance structures .

(Secure data provenance). Requirements

- Distributed.
  - Databases flow through untrusted environments.
- Integrity.
  - Nobody can forge provenance data
- Availability.
  - Auditors should be able to access provenance information in a secure, fast and reliable manner
- Privacy and confidentiality.
  - Avoid disclosure. Only authorized users can access the information.

Provenance and privacy

Privacy issues. Privacy of data provenance

- Secure data provenance, to ensure distributed approach, integrity, availability, and privacy. Cryptographic approaches and access control mechanisms.
- Privacy for sensitive data provenance: Anonymization, to release *one-shot* data provenance.

Provenance and privacy

Other privacy issues.

(Data provenance privacy issues beyond privacy of data provenance)

- Deletion/amendment may require the reconsideration of inferences.

Provenance and privacy

Other privacy issues.

(Data provenance privacy issues beyond privacy of data provenance)

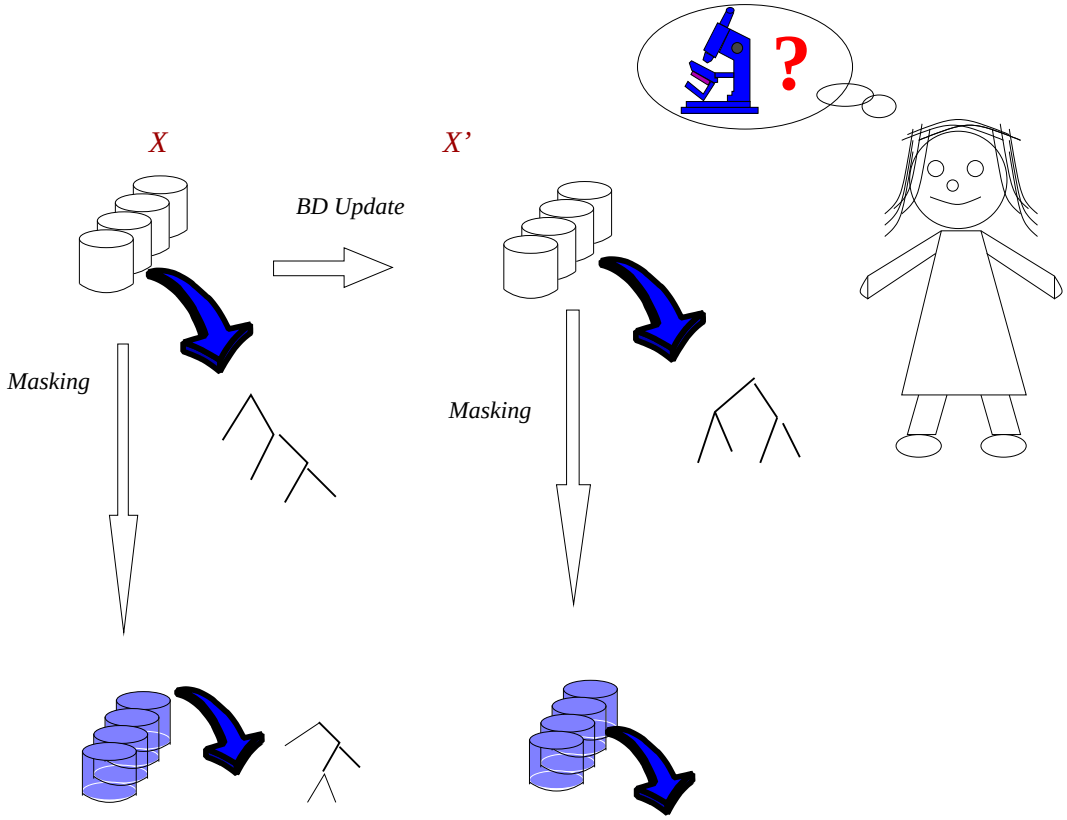
- Deletion/amendment may require the reconsideration of inferences.
  inferences = machine learning models (decision trees)

Provenance and privacy

Other privacy issues.

(Data provenance privacy issues beyond privacy of data provenance)

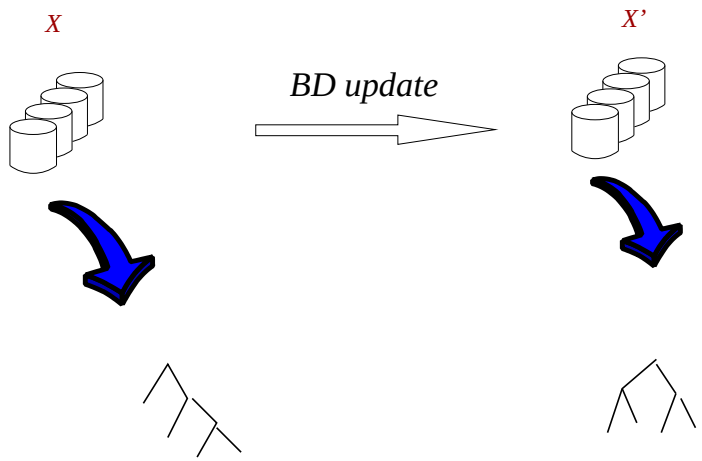
- Deletion/amendment may require the reconsideration of inferences. inferences = machine learning models (decision trees)



$M(X) = M(X')$ (in provenance) vs. $M(X)(y) = M(X')(y)$ (in IL)

Provenance and privacy

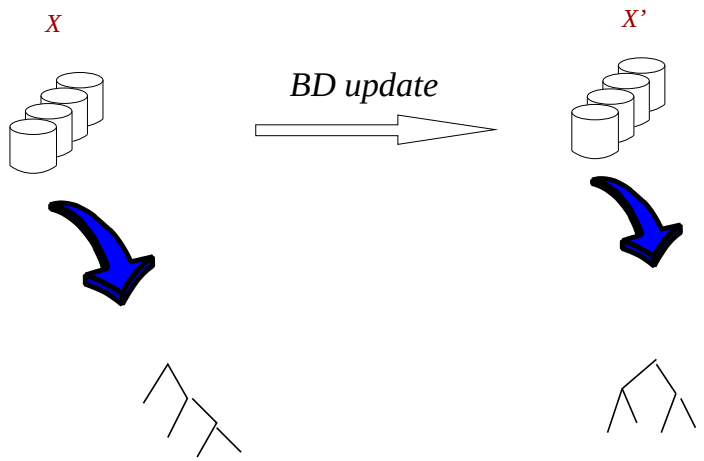
Other privacy issues. Data mining & provenance



- Should we annul/nullify a model G learnt from a dataset when some records are deleted/amended? Decisions should be revoked?

Provenance and privacy

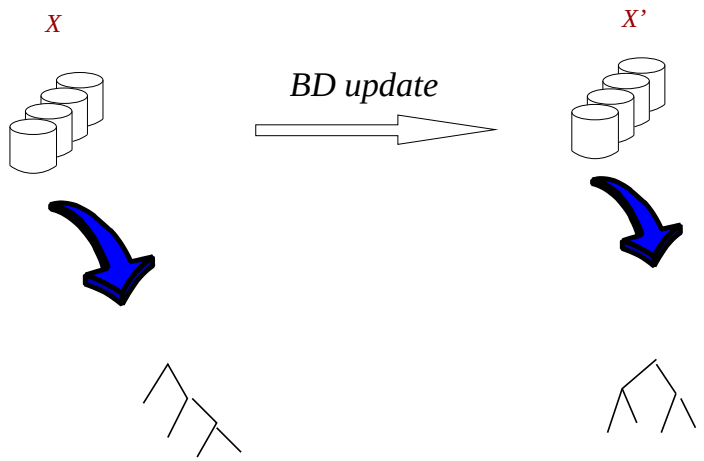
Other privacy issues. Data mining & provenance



- Should we annul/nullify a model G learnt from a dataset when some records are deleted/amended? Decisions should be revoked?
  e.g. G=decision tree (mortgage denied/accepted)
  μ=remove (all) people with salary between [15000,20000] EUR.

Provenance and privacy

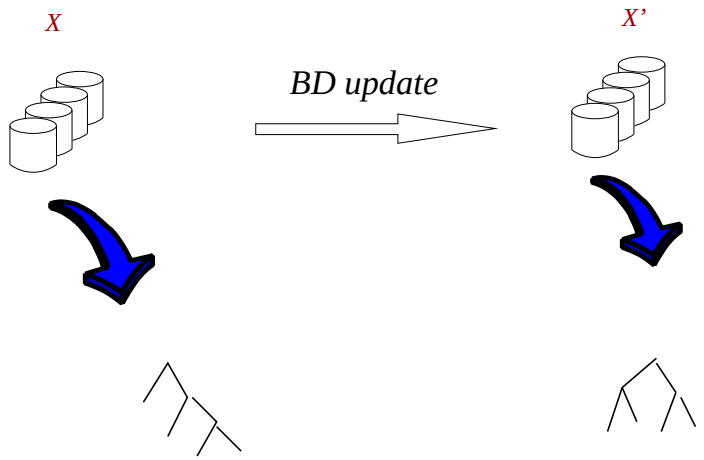
Other privacy issues. Data mining & provenance



- Should we annul/nullify a model G learnt from a dataset when some records are deleted/amended? Decisions should be revoked?
  e.g. G=decision tree (mortgage denied/accepted)
  μ=remove (all) people with salary between [15000,20000] EUR.
- Given two (different) models G and G' extracted from the files, do they guarantee privacy on the modifications (μ)?

Provenance and privacy

Other privacy issues. Data mining & provenance



- Should we annul/nullify a model G learnt from a dataset when some records are deleted/amended? Decisions should be revoked?
  e.g. G=decision tree (mortgage denied/accepted)
  μ=remove (all) people with salary between [15000,20000] EUR.
- Given two (different) models G and G' extracted from the files, do they guarantee privacy on the modifications (μ)?
  e.g., intruder has G and G', can infer μ?

Some research lines

Some research lines

- Disclosure risk and transparency
  i.e., intruder knows how data has been protected
- Synthetic data for graphs (social networks)
- Provenance and data privacy

Summary

Summary

- Anonymization and big data
- Some lines of research related to big data and data provenance
  - Technology to help users to know what others can infer from them
  - Methods so that databases can be anonymized at origin
  - Methodology for controlled linkability
  - Composability of privacy models
  - Decentralized anonymity
  - Efficient algorithms for big data
  - Secure data provenance and anonymization methods for provenance
  - Interaction between data privacy and data provenance

Thank you

References

Related references.

- G. D'Acquisto, J. Domingo-Ferrer, P. Kikiras, V. Torra, Y.-A. de Montjoye, A. Bourka, Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics, ENISA: European Union Agency for Network and Information Security, December 2015.
- V. Torra, G. Navarro-Arribas, Integral privacy, manuscript.
- D. Abril, G. Navarro-Arribas, V. Torra, Supervised Learning Using a Symmetric Bilinear Form for Record Linkage, Information Fusion 26 (2015) 144-153.
- J. Herranz, S. Matwin, J. Nin, V. Torra, V. (2010) Classifying data from protected statistical datasets, Computers & Security 29:8 875-890
- K. Stokes, V. Torra, Multiple releases of k-anonymous data sets and k-anonymous relational databases, Int. J. of Unc. Fuzziness and Knowledge Based Systems, 20:6 (2012) 839-853.