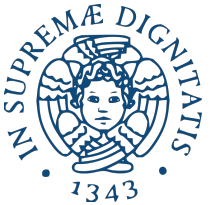
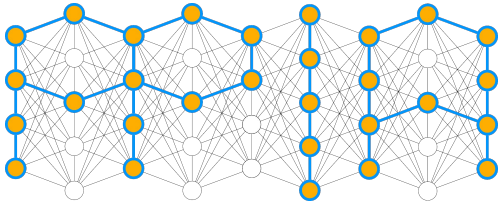
UNIVERSITÀ
DEGLI STUDI
DI PADOVA





MICRODATA PROTECTION

PRIVACY PRESERVING INFORMATION ACCESS

PhD in Information Engineering
A.Y. 2025/2026

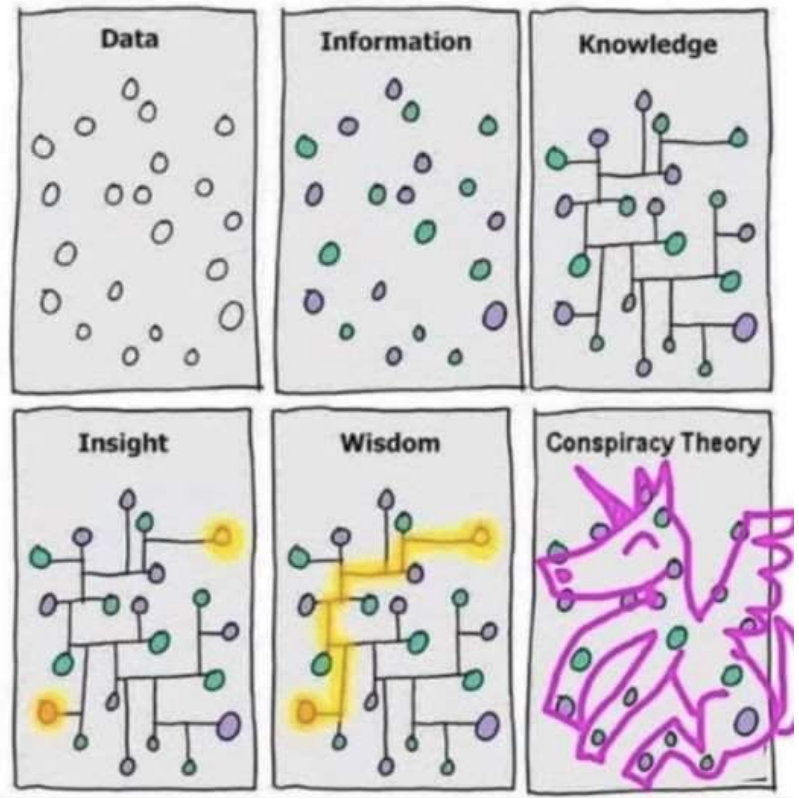
GUGLIELMO FAGGIOLI
Intelligent Interactive Information Access (IIIA) Hub
Department of Information Engineering
University of Padua

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DATA

Data: information, especially facts or numbers, collected to be examined, considered and used to help with making decisions.

Data make sense only if considered together.



DATA

Microdata: pieces of information at the level of individual respondents. data about a single respondent.

Each record refers to a specific person/patient/user/corporation (an entity that needs its privacy protected).

Each record has several attributes.

TYPES OF ATTRIBUTES

- **Identifiers**: attributes that allow to (almost) univocally identify the person

TYPES OF ATTRIBUTES

- **Identifiers**: attributes that allow to (almost) univocally identify the person
  SSN, name, address

TYPES OF ATTRIBUTES

- **Identifiers**: attributes that allow to (almost) univocally identify the person

- **Quasi-identifiers**: attributes that, in isolation, do not provide much information but, if combined, might identify the person

TYPES OF ATTRIBUTES

- **Identifiers**: attributes that allow to (almost) univocally identify the person

- **Quasi-identifiers**: attributes that, in isolation, do not provide much information but, if combined, might identify the person -
  Sex, DoB, residence ZIP, marital status

TYPES OF ATTRIBUTES

- **Identifiers**: attributes that allow to (almost) univocally identify the person
- **Quasi-identifiers**: attributes that, in isolation, do not provide much information but, if combined, might identify the person
- **Confidential attributes**: attributes that concerns the personal sphere of the respondent

TYPES OF ATTRIBUTES

- **Identifiers**: attributes that allow to (almost) univocally identify the person

- **Quasi-identifiers**: attributes that, in isolation, do not provide much information but, if combined, might identify the person

- **Confidential attributes**: attributes that concerns the personal sphere of the respondent
  Medical conditions, political opinions, religious beliefs

TYPES OF ATTRIBUTES

- **Identifiers**: attributes that allow to (almost) univocally identify the person
- **Quasi-identifiers**: attributes that, in isolation, do not provide much information but, if combined, might identify the person
- **Confidential attributes**: attributes that concern the personal sphere of the respondent
- Non-confidential attribute: anything that (**at current time**) cannot be used to disclose users' identity or sensible information.

MICRODATA

SSN	Name	Race	DoB	Sex	ZIP	Marital Status	Disease
1230044954330	Yu Zhong	asian	64/04/12	F	35138	divorced	hypertension
1230082394331	Tao Jiang	asian	64/09/13	F	35142	divorced	obesity
2934944954322	Tang Hanying	asian	64/04/15	F	35148	married	chest pain
1230004449530	Djimon Igwe	black	63/03/15	M	35138	married	obesity
3823893498549	Ashton Katy	black	63/02/18	M	35138	married	short breath
7938458593247	Alyssa Bryce	black	64/09/27	F	35137	single	short breath
9584935832839	Jean Harmon	white	64/09/27	F	35137	single	obesity
7852438634549	Caitlyn Em	white	64/09/27	F	35141	single	chest pain
3582347528778	Louise Wayland	white	64/09/27	F	35141	widow	short breath

MICRODATA: IDENTIFIERS

SSN	Name	Race	DoB	Sex	ZIP	Marital Status	Disease
1230044954330	Yu Zhong	asian	64/04/12	F	35138	divorced	hypertension
1230082394331	Tao Jiang	asian	64/09/13	F	35142	divorced	obesity
2934944954322	Tang Hanying	asian	64/04/15	F	35148	married	chest pain
1230004449530	Djimon Igwe	black	63/03/15	M	35138	married	obesity
3823893498549	Ashton Katy	black	63/02/18	M	35138	married	short breath
7938458593247	Alyssa Bryce	black	64/09/27	F	35137	single	short breath
9584935832839	Jean Harmon	white	64/09/27	F	35137	single	obesity
7852438634549	Caitlyn Em	white	64/09/27	F	35141	single	chest pain
3582347528778	Louise Wayland	white	64/09/27	F	35141	widow	short breath

MICRODATA: QUASI IDENTIFIERS

SSN	Name	Race	DoB	Sex	ZIP	Marital Status	Disease
1230044954330	Yu Zhong	asian	64/04/12	F	35138	divorced	hypertension
1230082394331	Tao Jiang	asian	64/09/13	F	35142	divorced	obesity
2934944954322	Tang Hanying	asian	64/04/15	F	35148	married	chest pain
1230004449530	Djimon Igwe	black	63/03/15	M	35138	married	obesity
3823893498549	Ashton Katy	black	63/02/18	M	35138	married	short breath
7938458593247	Alyssa Bryce	black	64/09/27	F	35137	single	short breath
9584935832839	Jean Harmon	white	64/09/27	F	35137	single	obesity
7852438634549	Caitlyn Em	white	64/09/27	F	35141	single	chest pain
3582347528778	Louise Wayland	white	64/09/27	F	35141	widow	short breath

MICRODATA: CONFIDENTIAL ATTRIBUTES

SSN	Name	Race	DoB	Sex	ZIP	Marital Status	Disease
1230044954330	Yu Zhong	asian	64/04/12	F	35138	divorced	hypertension
1230082394331	Tao Jiang	asian	64/09/13	F	35142	divorced	obesity
2934944954322	Tang Hanying	asian	64/04/15	F	35148	married	chest pain
1230004449530	Djimon Igwe	black	63/03/15	M	35138	married	obesity
3823893498549	Ashton Katy	black	63/02/18	M	35138	married	short breath
7938458593247	Alyssa Bryce	black	64/09/27	F	35137	single	short breath
9584935832839	Jean Harmon	white	64/09/27	F	35137	single	obesity
7852438634549	Caitlyn Em	white	64/09/27	F	35141	single	chest pain
3582347528778	Louise Wayland	white	64/09/27	F	35141	widow	short breath

SECURITY VS PRIVACY

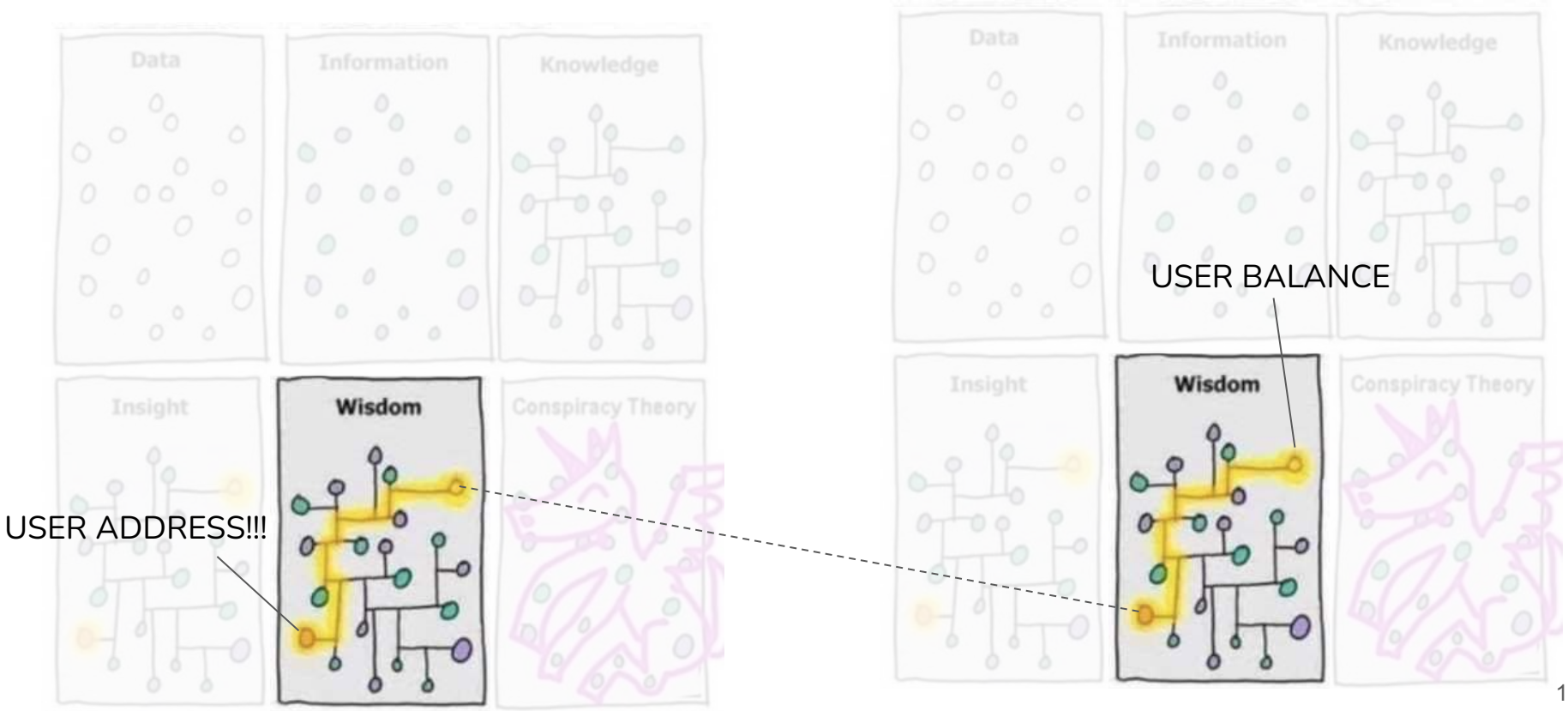
Security → Protecting the data from unauthorized accesses, theft or corruption.

Security concerns data access and it is a fundamental part to achieve privacy.

Privacy Protection → Protecting the information contained in the data or inferable from it.

Privacy concerns raise when we release (part of) the data or we try to obtain value.

SECURITY VS PRIVACY



SECURITY VS PRIVACY

A mail telling me that my anagraphic data has been stolen

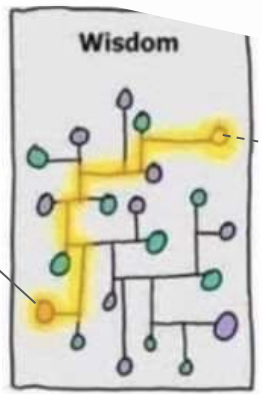
enel

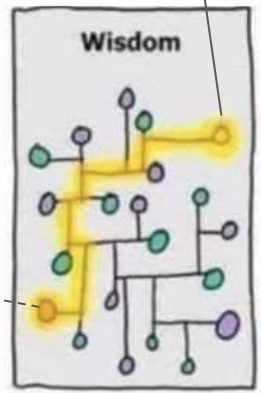
Buongiorno,

La informiamo che, in data ** *** ****, la società Network Contacts SpA – di cui Enel Energia si avvale per lo svolgimento di diversi servizi – ha subito un **attacco informatico che ha riguardato i dati personali di alcuni clienti** di Enel Energia, tra cui i Suoi (dati anagrafici (nome, cognome, codice fiscale, partita iva), indirizzo postale, numero di telefono, indirizzo e-mail).

USER BALANCE

USER ADDRESS!!!





RELEASING THE DATA

Depending on the setting, it might be necessary (e.g., statistics), desirable (e.g., marketing), or suggested (e.g., research) to release some data.

Data can be released in:

- Aggregated form → **macrodata**
- API → **statistical databases**
- Original form → **microdata**

MACRODATA, STATISTICAL DATABASES AND MICRODATA

Macrodata: data released in **aggregated** form.

Weaknesses:

MACRODATA, STATISTICAL DATABASES AND MICRODATA

Macrodata: data released in **aggregated** form.

Weaknesses:

- information for rare subjects can be inferred also from aggregated data
- Low flexibility and information availability

Possible protection techniques:

- selective obfuscation of sensible cells

MACRODATA, STATISTICAL DATABASES AND MICRODATA

Macrodata: data released in **aggregated** form.

Statistical databases: databases that accepts only **distributional queries**.

Weaknesses:

MACRODATA, STATISTICAL DATABASES AND MICRODATA

Macrodata: data released in **aggregated** form.

Statistical databases: databases that accepts only **distributional queries**.

Weaknesses:

- **Reconstruction** and **differencing attacks** are still possible
- Low flexibility and information availability

Possible protection techniques:

- Limit the queries or perturbate the results - Differential privacy

MACRODATA, STATISTICAL DATABASES AND MICRODATA

Macrodata: data released in **aggregated** form.

Statistical databases: databases that accepts only **distributional queries**.

Weaknesses:

- Reconstruction and differencing attacks are still possible
- **Low flexibility and information availability**

MACRODATA, STATISTICAL DATABASES AND MICRODATA

Macrodata: data released in **aggregated** form.

Statistical databases: databases that accepts only **distributional queries**.

Weaknesses:

- **Low flexibility and information availability**

Microdata: information about **single** respondents.

MACRODATA

Macrodata is, by definition, released under the form of **tables**.

Three main categories of tables:

- Counts

	Meat	Vegetables	Fish	Eggs	Tot
M	4	5	0	2	11
F	2	5	4	3	14
Tot	6	10	4	5	25

respondents indicating their favourite food

MACRODATA

Macrodata is, by definition, released under the form of **tables**.

Three main categories of tables:

- Frequencies

	Meat	Vegetables	Fish	Eggs	Tot
M	0.16	0.2	0	0.08	0.44
F	0.08	0.2	0.16	0.12	0.56
Tot	0.24	0.4	0.16	0.2	1

proportion of respondents indicating their favourite food

MACRODATA

Macrodata is, by definition, released under the form of **tables**.

Three main categories of tables:

- Magnitudes: statistics about a quantity of interest

	Meat	Vegetables	Fish	Eggs	Tot
M	3.4	10.7	1.3	2.6	18
F	3.9	11.2	1.5	1.4	18
Tot	7.3	21.9	2.8	4	36

average number of portions per week

SENSITIVE CELLS

Sensitive cells are cells that provide too much information by identifying few subjects.

the **threshold rule** allows identifying sensitive cells: according to the type of data and quantities considered, we decide a threshold that identifies what cells are sensitive.

For example, if we look at our count table, we might say that 2 is our threshold.

SENSITIVE CELLS: COUNTS AND FREQUENCIES

The following techniques can be used to protect sensitive cells:

- cell suppression

  Remove cells containing values below the threshold (**primary suppression**).

  If the values can still be inferred (or estimated), remove also other cells with linear programming (**secondary suppression**).

SENSITIVE CELLS: COUNTS AND FREQUENCIES

The following techniques can be used to protect sensitive cells:

- rounding

  round a value to the closest multiple of a chosen **base number**

values			sum
32	24	6	62

values			sum
~~32~~30	~~24~~20	~~6~~10	~~62~~60

SENSITIVE CELLS: COUNTS AND FREQUENCIES

The following techniques can be used to protect sensitive cells:

- roll up categories

  reduce the dimension of the table by **joining** categories - this increases the number of subjects in each (macro-)category.

values			sum
32	24	6	62

values		sum
32	30	62

SENSITIVE CELLS: COUNTS AND FREQUENCIES

The following techniques can be used to protect sensitive cells:

- sampling

  Instead of using **census** data carry out a **survey** and **sample** subjects.

SENSITIVE CELLS: COUNTS AND FREQUENCIES

The following techniques can be used to protect sensitive cells:

- Controlled tabular adjustment function (CTA)

  replace lower values with the threshold one and, using linear programming, adjust the others to maintain equal the sum.

eg:

values			sum
32	24	6	62

values			sum
~~32~~30	~~24~~22	~~6~~10	62

SENSITIVE CELLS: MAGNITUDES

The following techniques can be used to identify sensitive **magnitude** cells:

- (n, k)-rule

  Are considered sensitive cells for which less than n subjects have contributed at least for the k% of the cell's total value.

  (1, 50)-rule considers sensitive cells where a single user contributed more than the 50% of the value of the cell.

SENSITIVE CELLS: MAGNITUDES (N, K)-RULE (EXAMPLE)

(3-30) rule

	Meat	Vegetables	Fish	Eggs	Tot
M	4	5	0	2	11
F	2	5	4	3	14
Tot	6	10	4	5	25

SENSITIVE CELLS: MAGNITUDES (N, K)-RULE (EXAMPLE)

(3-30) rule

sensitive: 2 (2<3) individuals contributed to each cell, each with 50% (50>30) of the value of the cell

	Meat	Vegetables	Fish	Eggs	Tot
M	4	5	0	2	11
F	2	5	4	3	14
Tot	6	10	4	5	25

SENSITIVE CELLS: MAGNITUDES (N, K)-RULE (EXAMPLE)

(3-30) rule

non-sensitive: 3 (3=3) individuals contributed to each cell, each with 33% (33>30) of the value of the cell.

	Meat	Vegetables	Fish	Eggs	Tot
M	4	5	0	2	11
F	2	5	4	3	14
Tot	6	10	4	5	25

SENSITIVE CELLS: MAGNITUDES

The following techniques can be used to identify sensitive **magnitude** cells:

- p-percentage

  $t - v_1 - v_2 < p/100 \cdot v_1$

  where t is the value of the cell, v_1 and v_2 are respectively the biggest and second biggest values that contribute to the value of the cell, while p is the security parameter ($p<100$).

  if the inequality is true, t is sensitive (easy to estimate v_1).

SENSITIVE CELLS: MAGNITUDES

The techniques applied to protect counts and frequencies can also be used to protect magnitude tables:

- Cell suppression
- Rounding
- Roll-up categories
- Sampling
- CTA

MICRODATA: RISKS

- highly visible tuples: **outliers** are often most vulnerable subjects (rare diseases, peculiarities), but also the most exposed to identification risks.

- it is possible to **join** microdata with external sources to gain identify subjects (re-identification attack)

- **several common attributes** between microdata and external sources

MICRODATA: NATURAL DEFENSES

- microdata are (often) only a **subset** of the whole population: it is hard to be sure that the subject is among the considered data
- data is often **not up-to-date**: typically, few years pass between the collection of the data and the release
- microdata contains **natural noise** that makes it hard to anchor it to other sources
- microdata and external sources might report data in **different forms**

TYPES OF DISCLOSURES

- **Identity disclosure:** reconstruction of the individual identity with a combination of their attributes (SSN, name, address)

- **Attribute disclosure:** a combination of indirect identifiers allows inferring a given attribute value

- **Inferential disclosure:** information can be inferred with high probability from statistical properties of the released data.

SANITIZING THE DATA

Sanitizing is the process of removing explicit identifiers:

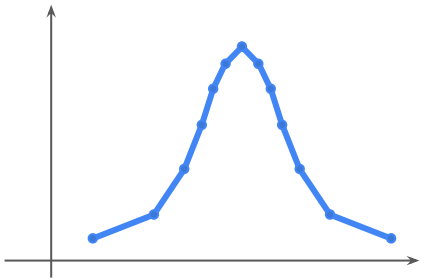
- names
- addresses
- identifying codes (SSN, CF, etc)
- phone numbers

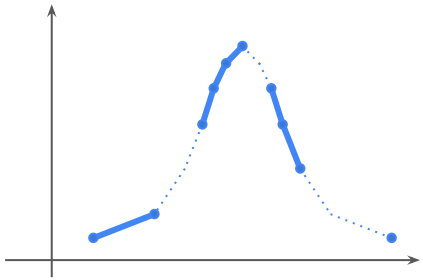
MICRODATA: DESIDERATA

Protect user privacy: modify, change, hide, transform the data so that it is impossible to identify who the respondent is.

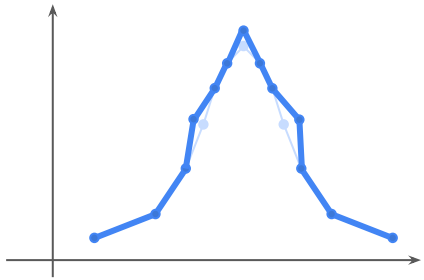
Maintain key statistical properties: if data are so scrambled that it is impossible to identify any pattern, then they are as good as white noise.

MICRODATA PROTECTION: MASKING



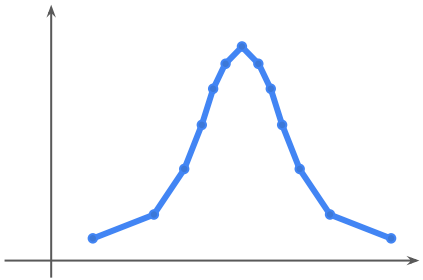


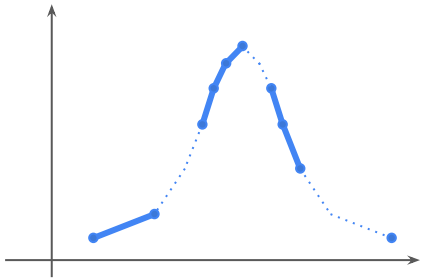
non perturbative



perturbative

MICRODATA PROTECTION: MASKING



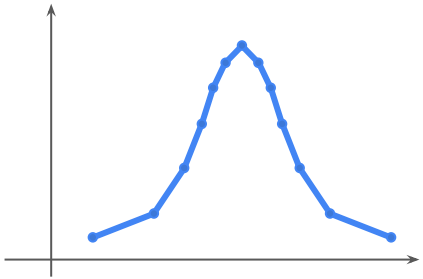


non perturbative

Eliminate points from the original microdata

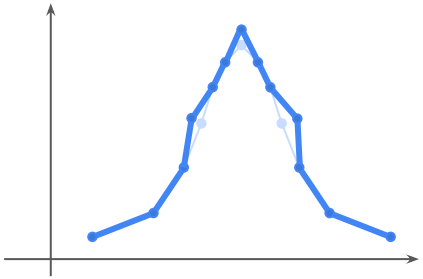
perturbative

MICRODATA PROTECTION: MASKING



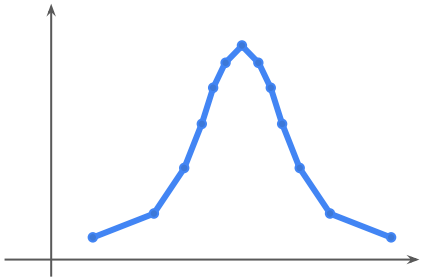
Remove unique combinations and/or introduce new combinations of attributes

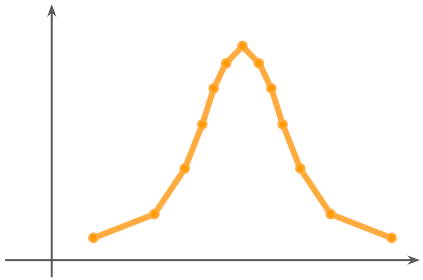
non perturbative



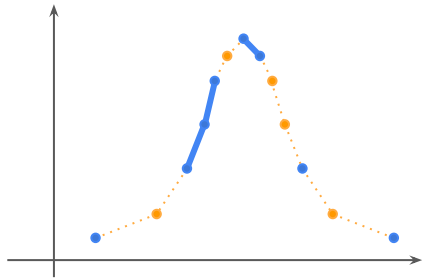
perturbative

MICRODATA PROTECTION: SYNTHETIC DATA GENERATION



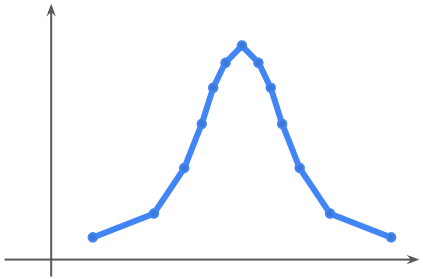


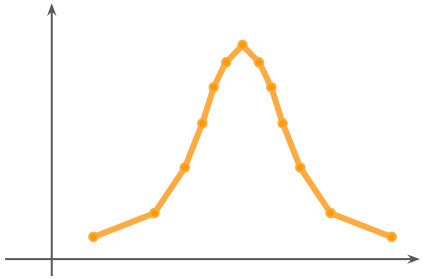
fully synthetic



perturbative

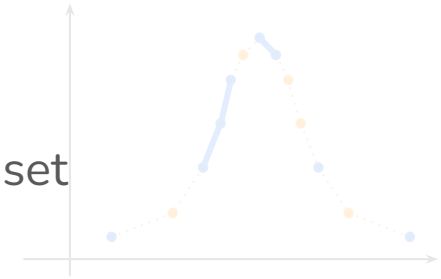
MICRODATA PROTECTION: SYNTHETIC DATA GENERATION





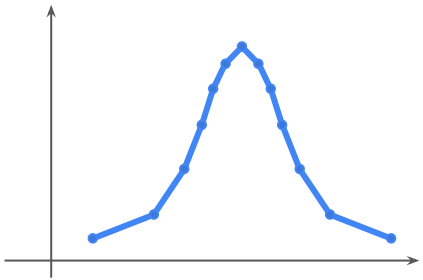
fully synthetic

Generate a new dataset

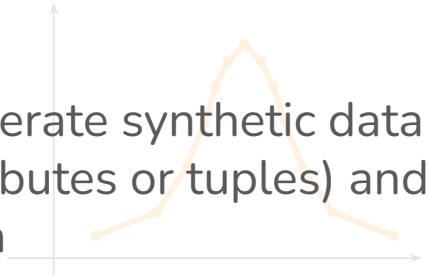


perturbative

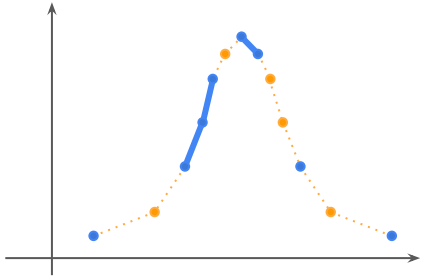
MICRODATA PROTECTION: SYNTHETIC DATA GENERATION



Generate synthetic data (e.g., for some attributes or tuples) and combine it with real data



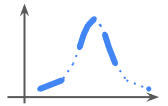
fully synthetic



perturbative

MASKING

SAMPLING



Remove some observations to decrease the risk of identification.

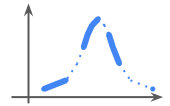
Less subjects means also **lower probability** that a subject is on our dataset.

“Just a few” policy.

CATEGORICAL: YES

CONTINUOUS: YES

LOCAL SUPPRESSION



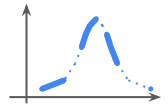
Replace attributes or cells that contribute to disclosure risk with "**missing value**".

Similar to what we did for macrodata. More on this in the next lectures.

CATEGORICAL: YES

CONTINUOUS: YES

GLOBAL RECODING


non perturbative

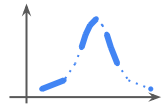
Partition the domain of an attribute into **disjoint intervals** (often with equal size) and replace a value with the label of the interval it falls in.

What happens to the extremities?

CATEGORICAL: ONLY ORDERED

CONTINUOUS: YES

TOP-CODING



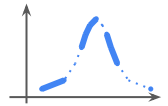
A specific instance of the global coding. Convert all the values above a certain threshold with a “top-code”.

For example, replace all people who are taller than 1.9m, with the attribute “>1.9m”.

CATEGORICAL: ONLY ORDERED

CONTINUOUS: YES

BOTTOM-CODING



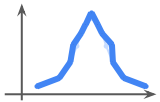
A specific instance of the global coding. Convert all the values above a certain threshold with a “bottom-code”.

For example, replace all people who are shorter than 1.7m, with the attribute “<1.7m”.

CATEGORICAL: ONLY ORDERED

CONTINUOUS: YES

ROUNDING



Similar to the global coding, but instead of replacing with the label of the interval, use a point in it.

given a step size b and k points p_1, ..., p_k, each interval I_i is defined as follows:

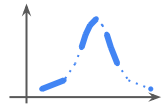
$$I_i = \begin{cases} \left[0,\ p_1 + \frac{b}{2}\right], & \text{if } i = 0 \\ \left[p_i + \frac{b}{2}, p_{i+1} + \frac{b}{2}\right) & \text{if } 0 < i < k \\ \left[p_k,\ \max\right] & \text{if } i = k - 1 \end{cases}$$

and p_i is the label of the i-*th* interval.

CATEGORICAL: NO

CONTINUOUS: YES

GENERALIZATION



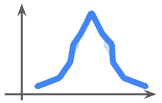
Using a **generalization hierarchy**, replace attributes with a generalization. For example, instead of using the full date of birth, use only the year or the month, or use only part of the zip code.

More on this in the next lectures.

CATEGORICAL: YES

CONTINUOUS: VIA GLOBAL RECORDING

RESAMPLING



values
10
18
20
8
11
14

S1	S2	S3
11	18	11
8	8	14
20	20	18
18	10	10
14	10	11
14	11	20

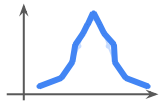
S1	S2	S3
8	8	10
11	10	11
14	10	11
14	11	14
18	18	18
20	20	20

Resample with replacement the column that you wish to obfuscate multiple times, and sort the values within each sample.

CATEGORICAL: NO

CONTINUOUS: YES

RESAMPLING



values
10
18
20
8
11
14

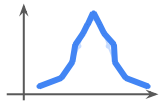
S1	S2	S3
11	18	11
8	8	14
20	20	18
18	10	10
14	10	11
14	11	20

S1	S2	S3	values
8	8	10	8.7
11	10	11	10.7
14	10	11	11.7
14	11	14	13
18	18	18	18
20	20	20	20

CATEGORICAL: NO

CONTINUOUS: YES

RESAMPLING



values
10
18
20
8
11
14

S1	S2	S3
11	18	11
8	8	14
20	20	18
18	10	10
14	10	11
14	11	20

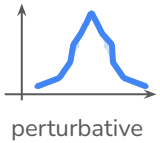
S1	S2	S3
8	8	10
11	10	11
14	10	11
14	11	14
18	18	18
20	20	20

values
8.7
10.7
11.7
13
18
20

CATEGORICAL: NO

CONTINUOUS: YES

LOSSY COMPRESSION



Treat the **table as an image** and use a lossy compression algorithm (e.g., jpeg).

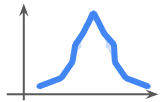
The challenging aspect is to convert values into “colors”.

The **compression rate** corresponds to the **obfuscation parameter**.

CATEGORICAL: NO

CONTINUOUS: YES

PRAM



Post RAndomized Method

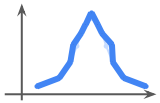
	meat	fish	eggs	vegetables
meat				
fish				
eggs				
vegetables				

Randomly change the label, using a - appositely chosen - distribution of probability over the others.

CATEGORICAL: YES

CONTINUOUS: NO

PRAM



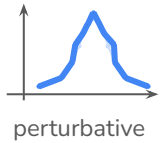
Post RAndomized Method

Randomly change the label, using a - appositely chosen - distribution of probability over the others.

CATEGORICAL: YES

CONTINUOUS: NO

PRAM



Post RAndomized Method

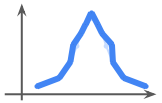
	meat	fish	eggs	vegetables
meat	0.6	0.25	0.1	0.05
fish	0.3	0.4	0.22	0.08
eggs	0.15	0.15	0.4	0.3
vegetables	0.05	0.1	0.28	0.57

Randomly change the label, using a - appositely chosen - distribution of probability over the others.

CATEGORICAL: YES

CONTINUOUS: NO

RANDOM NOISE



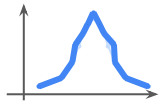
Additive noise: add noise sampled from a given distribution. Typically not sufficient - needs to be combined with linear (continuous) or non-linear (categorical) transformations.

Uncorrelated vs correlated additive noise: in the second case, draw the noise by also taking into account the co-variance of the attributes.

CATEGORICAL: NO

CONTINUOUS: YES

SWAPPING



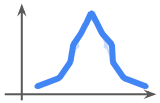
Randomly swap the values for the sensitive attributes.

This technique grants a high level of privacy but do not preserve statistical properties on subdomains (i.e., groupings over non-confidential attributes).

CATEGORICAL: YES

CONTINUOUS: YES

RANK SWAPPING

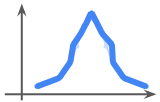


Swap values among tuples presenting, at most, **a distance of p** with p obfuscation parameter.

CATEGORICAL: YES

CONTINUOUS: YES

MICRO-AGGREGATION/BLURRING



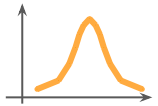
- **Cluster tuples** based on the maximal similarity criteria.
- compute the **mean** of the sensitive parameter for each cluster of tuples.
- for each tuple in the cluster, replace the sensitive value with the the mean computed in the previous step

CATEGORICAL: YES

CONTINUOUS: YES

SYNTHETIC DATA GENERATION

CHOLESKY DECOMPOSITION

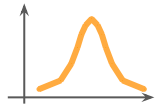


compute the **Cholesky decomposition** of the covariance matrix for the data ($C = U^T \times U$) and multiply U by R, a matrix of the same dimension of the original data containing random values such that its covariance matrix is the identity.

CATEGORICAL: NO

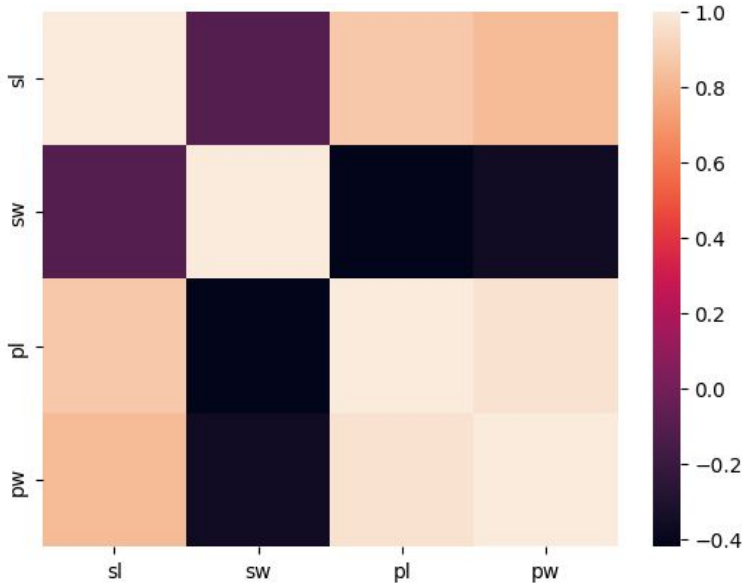
CONTINUOUS: YES

CHOLESKY DECOMPOSITION



Covariance matrix:

how much two variables change together (if one is big, the other is big too)

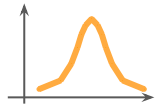


$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

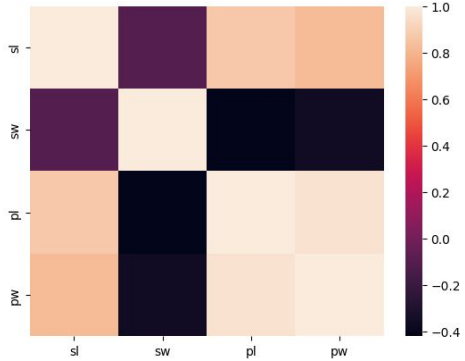
CATEGORICAL: NO

CONTINUOUS: YES

CHOLESKY DECOMPOSITION

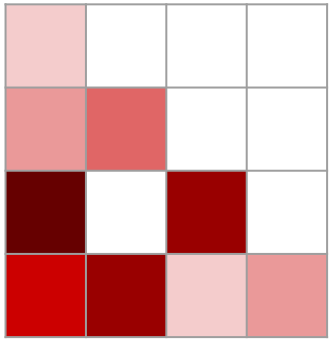


cov matrix



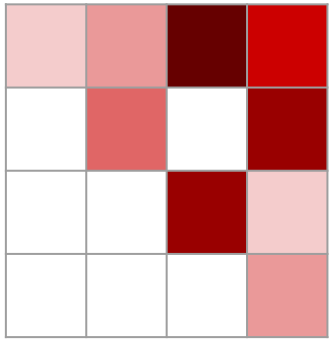
=

U



X

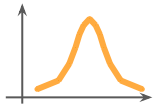
U^T



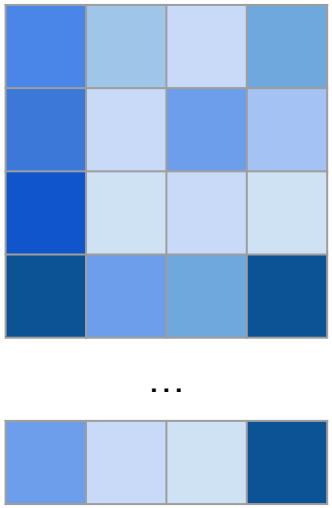
CATEGORICAL: NO

CONTINUOUS: YES

CHOLESKY DECOMPOSITION

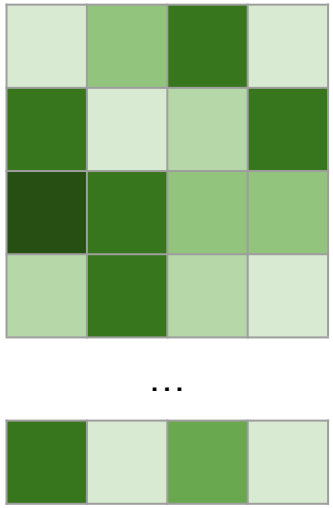


Synthetic dataset



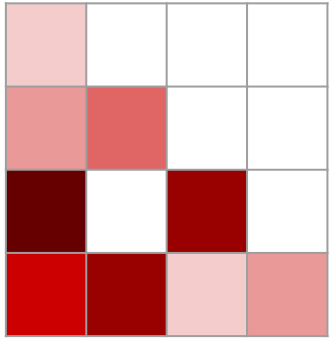
=

R



X

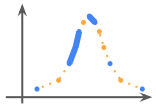
U



CATEGORICAL: NO

CONTINUOUS: YES

RANDOM RESPONSE



Technique widely adopted in social sciences that falls under the **differential privacy** framework.

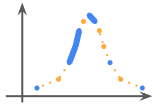
Ask to the individual to **answer to a question among a set of questions (many non sensitive, one sensitive)**, without indicating what question they decided. By knowing the distributions of the answers to non sensitive questions, we can **infer the distribution also about the sensitive one**.

More on this on next lectures

CATEGORICAL: YES

CONTINUOUS: NO

BLANK AND IMPUTE



Blank out the sensitive attributes for a random subset of tuples.

Replace the values on the cells that have been blanked out with an imputation (e.g., the mean or the most common category over highly similar tuples).

CATEGORICAL: YES

CONTINUOUS: YES

MEASURING THE CONFIDENTIALITY

DISCLOSURE RISK: UNIQUENESS

Population uniqueness:

Probability that a combination of attributes is unique in the population.

$$\Pr(PU) = \sum_j \frac{I(F_j = 1)}{N}$$

DISCLOSURE RISK: UNIQUENESS

Population uniqueness:

Probability that a combination of attributes is unique in the population.

$$\Pr(PU) = \sum_{j} \frac{I(F_j = 1)}{N}$$

1 if number of individuals with unique combination of attributes is 1, 0 else

Unique combinations of (selected) attributes

Dimension of the population

DISCLOSURE RISK: UNIQUENESS

Sample uniqueness:

Probability that a combination of attributes that is a unique in the population, is also a unique in the sample.

$$\Pr(PU \mid SU) = \sum_j \frac{I(f_j = 1 \wedge F_j = 1)}{I(f_j = 1)}$$

DISCLOSURE RISK: RECORD LINKAGE

The malicious adversary links a user from your dataset to another dataset, according to some common attributes (join), to expand their knowledge on the user.

- **Deterministic**: look exactly for the attributes in another dataset.
  Disadvantage: relevance of the attributes is ignored.

DISCLOSURE RISK: RECORD LINKAGE

The malicious adversary links a user from your dataset to another dataset, according to some common attributes, to expand their knowledge on the user.

- **Probabilistic**: all the tuples in the two datasets are compared - to each pair, a probability p that the two tuples represent a match is computed.

DISCLOSURE RISK: RECORD LINKAGE

The malicious adversary links a user from your dataset to another dataset, according to some common attributes, to expand their knowledge on the user.

- **Distance-based**: each tuple in the first dataset is matched to the most similar one in the second one, according to a specific **similarity measure** *f* that can also take into consideration the **importance** of each attribute

DISCLOSURE RISK: RECORD LINKAGE

How to use it at your advantage:

you don’t know which database will have your adversary, but you can consider **different (vertical) partitions of your data**, and **measure the risk** according to various record linkage approaches.

How often is the record linkage strategy capable of correctly identify (or do it with high probability) the record, given a subset of attributes?

INFORMATION LOSS: CONTINUOUS DATA

compute the statistics of interest on both original and protected data and compute an error measure:

- Mean Square Error (MSE)
- Mean Absolute Error (MAE)
- Mean Variation

INFORMATION LOSS: CATEGORICAL DATA

- Direct comparison:
  - if categories are not ordered count how many tuples have changed category compared to the original data
  - if categories are ordered, count how many categories of distance are present between the original category and the new one

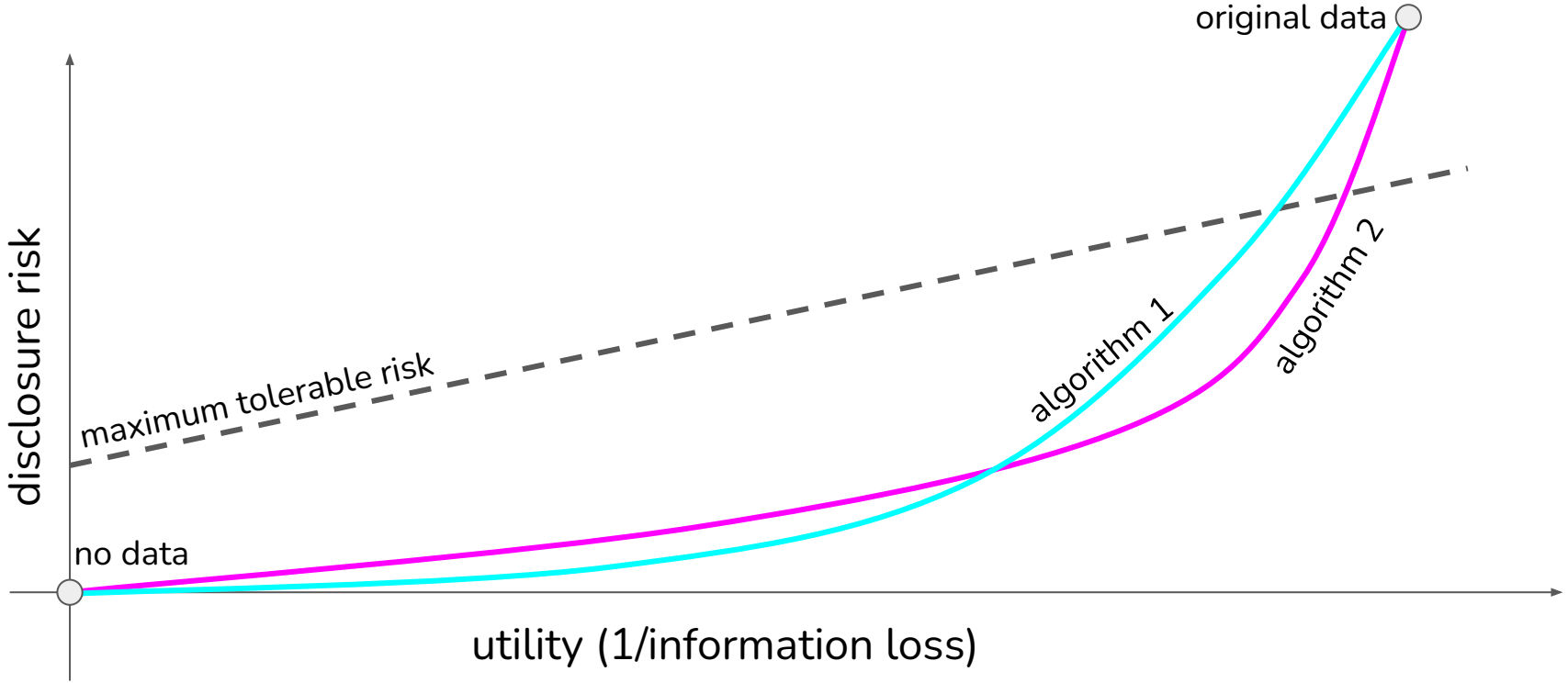
- contingency tables:
  - compare contingency tables for the original and the protected data.

- entropy:
  - compute the Shannon entropy between the original and protected data.

INFORMATION LOSS: MACHINE LEARNING

Measure the decrease in performance (prediction accuracy, precision, recall, ecc…) between the performance on original and protected microdata.

MEASURING THE TRADE OFF: R-U CONFIDENTIALITY MAP



REFERENCES

Slides based on:

Ciriani, V., De Capitani di Vimercati, S., Foresti, S., Samarati, P. (2007). Microdata Protection. In: Yu, T., Jajodia, S. (eds) Secure Data Management in Decentralized Systems. Advances in Information Security, vol 33. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-27696-0_9