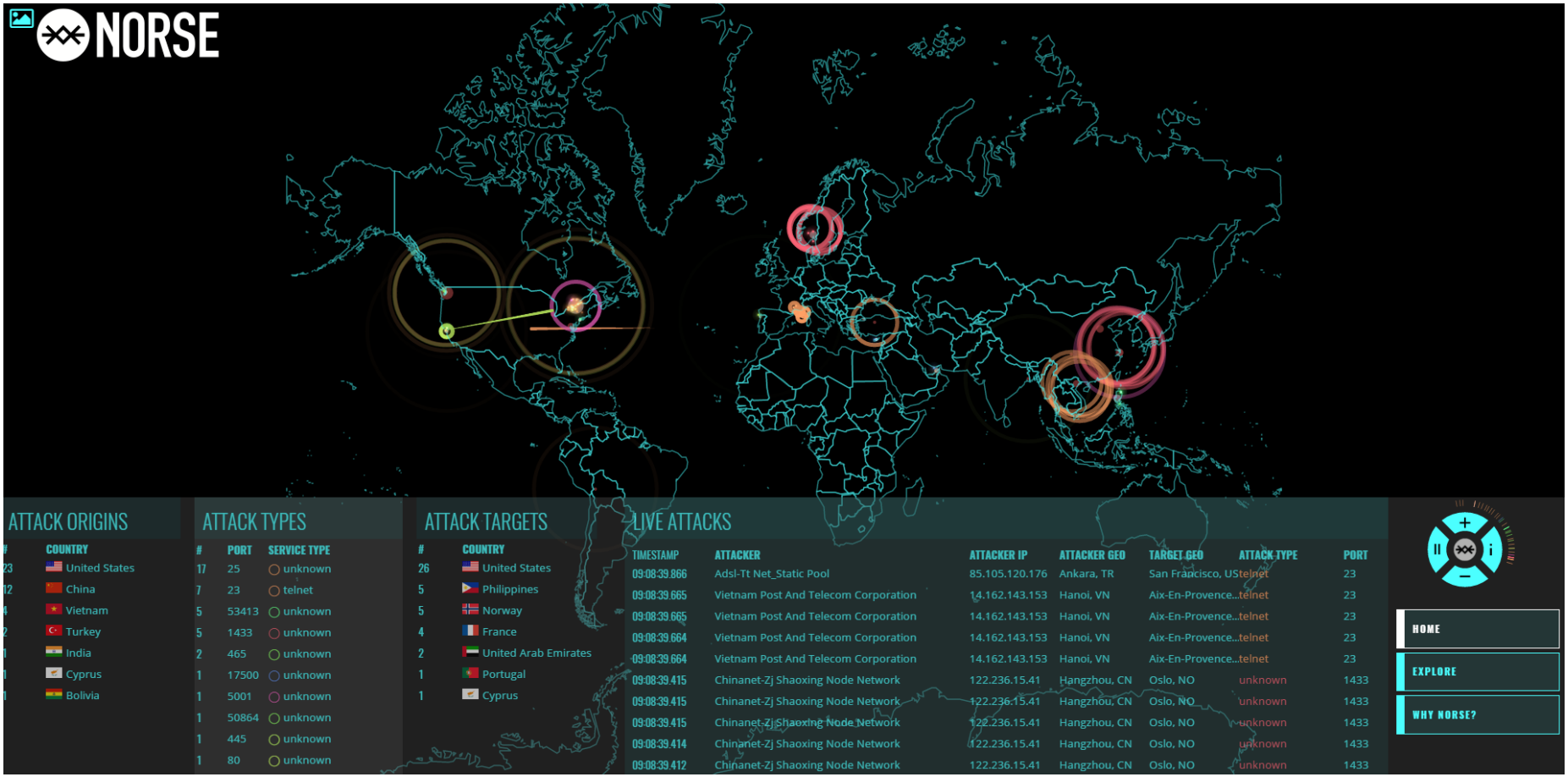
Business Intelligence & Process Modelling

Frank Takes

Universiteit Leiden

Lecture 3 — BI & Descriptive Analytics

IP Viking map



`http://map.norsecorp.com`

Recap

- **Business Intelligence**: anything that aims at providing actionable information that can be used to support business decision making
  - Business Intelligence
  - **Visual Analytics**
  - **Descriptive Analytics**
  - Predictive Analytics
- Process Modelling (April and May)

Universiteit Leiden

Visual Analytics
("last week's leftovers" or: "how it's not done")

Visualization

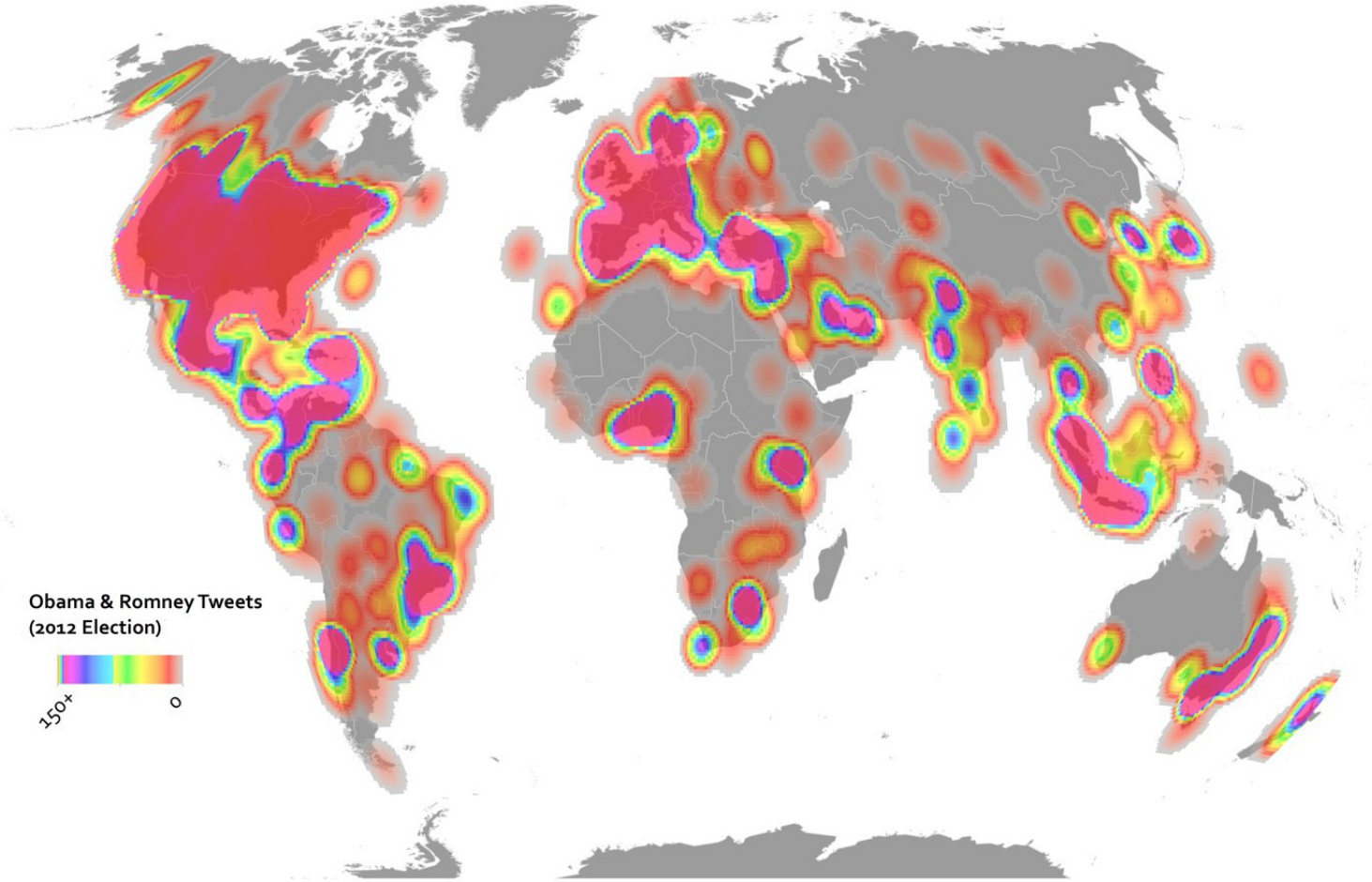
- **Visualization**: mapping data properties to visual attributes
- Good visualization: "proper" mapping of data attributes to visual attributes and properly "balancing" the number of data properties and visual attributes used

Visualization

- **Visualization**: mapping data properties to visual attributes
- Good visualization: "proper" mapping of data attributes to visual attributes and properly "balancing" the number of data properties and visual attributes used
- **Bad** visualization:
  - False data input
  - Misleading visual attributes
  - Abusing human background knowledge

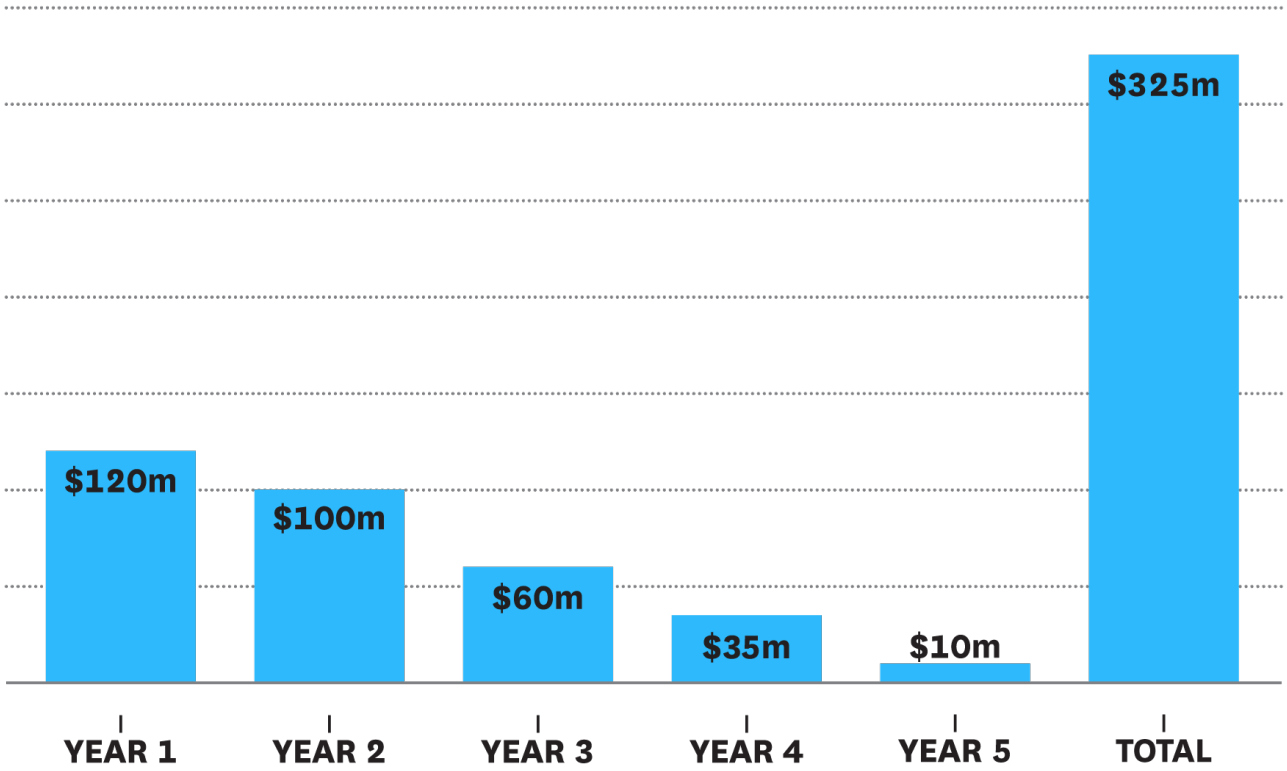
“Unbiased” data

Rainbow colors



http://poynter.org/uncategorized/224413

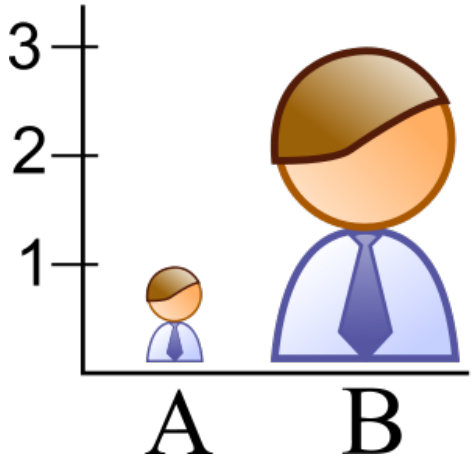
Parts and sums



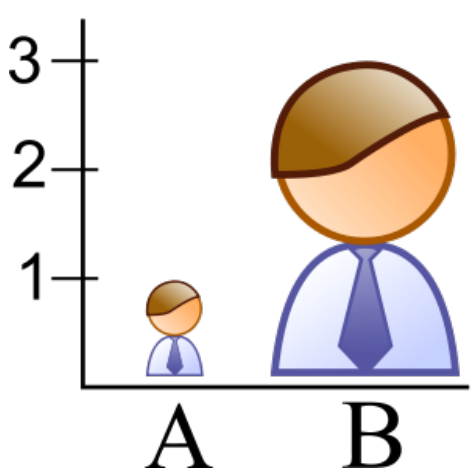
https://hbr.org/2014/12/vision-statement-how-to-lie-with-charts

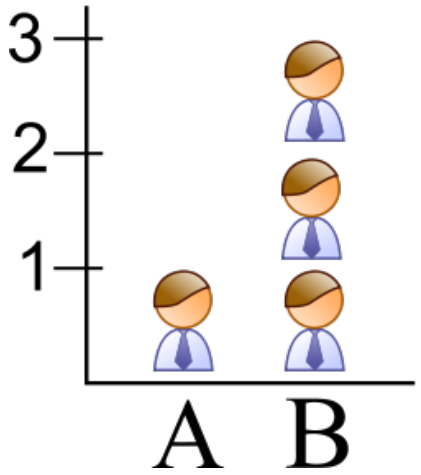
2D bars and icons

2D bars explained

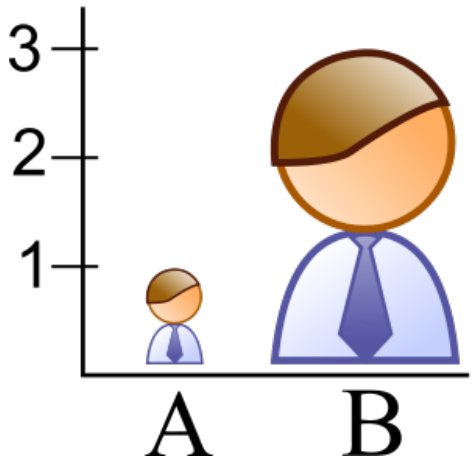


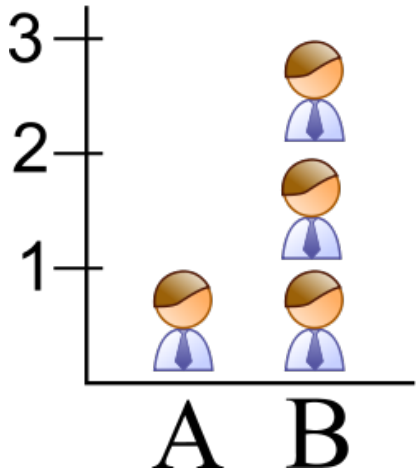
2D bars explained

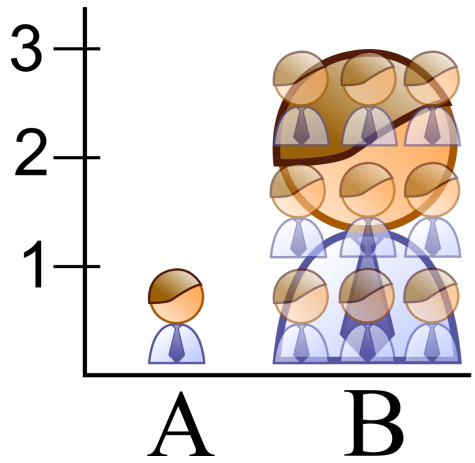




2D bars explained

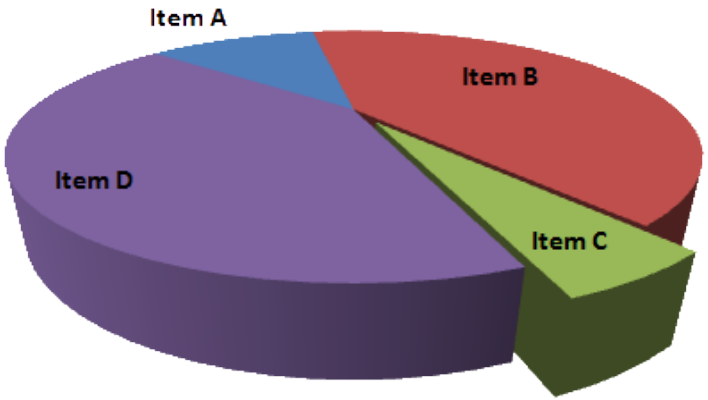




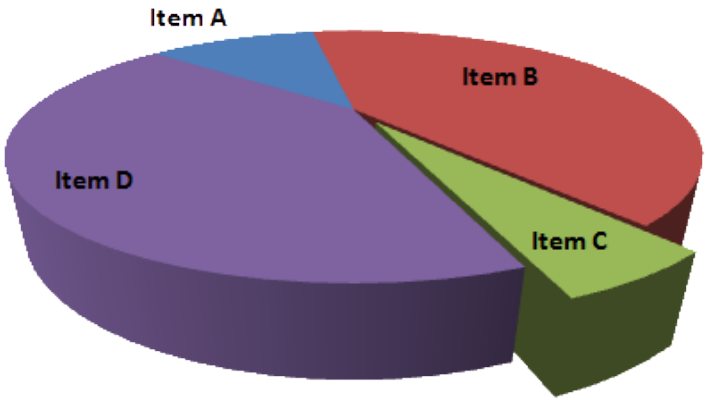


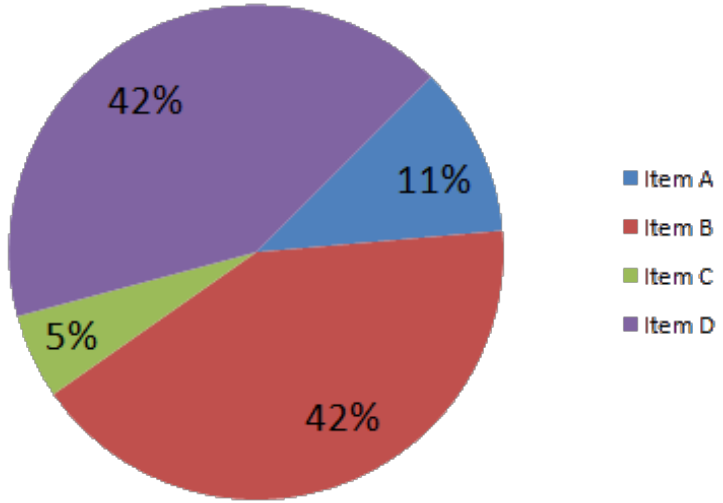
http://en.wikipedia.org/wiki/Misleading_graph

3D pies



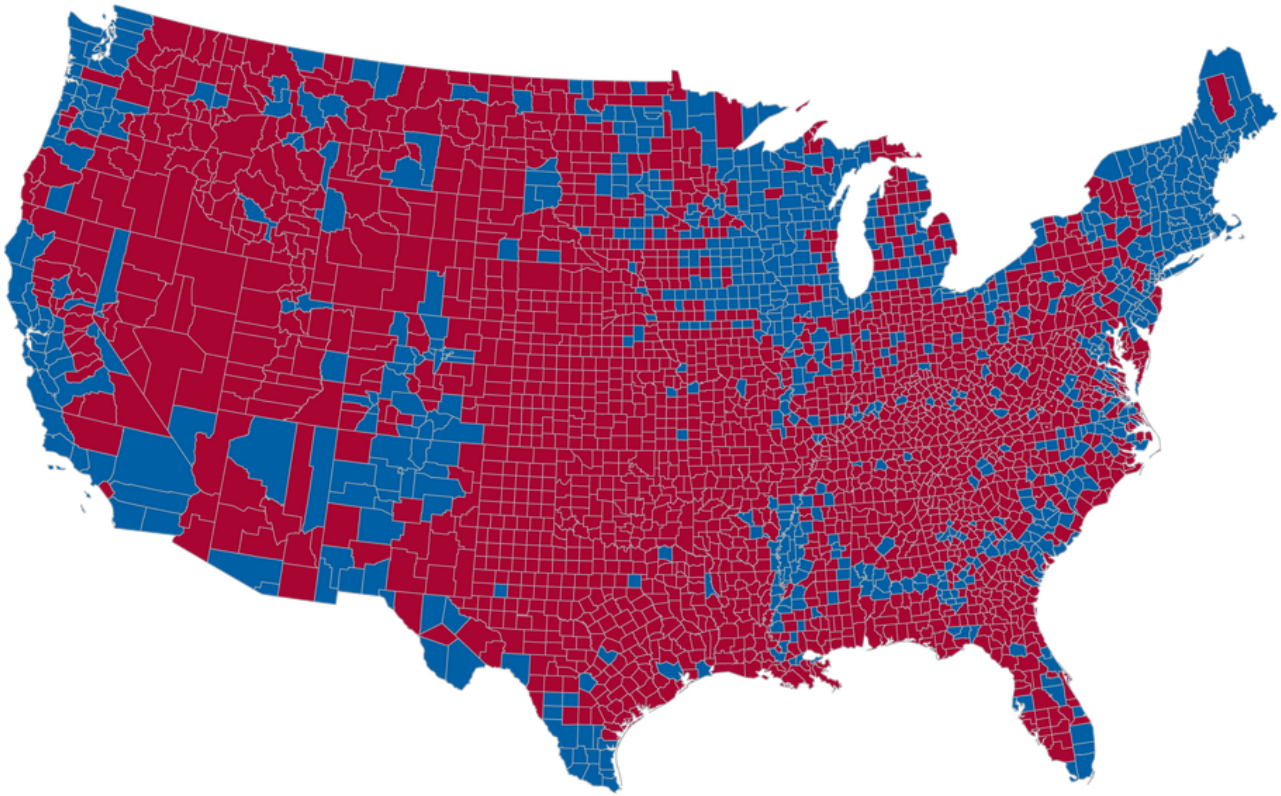
3D pies





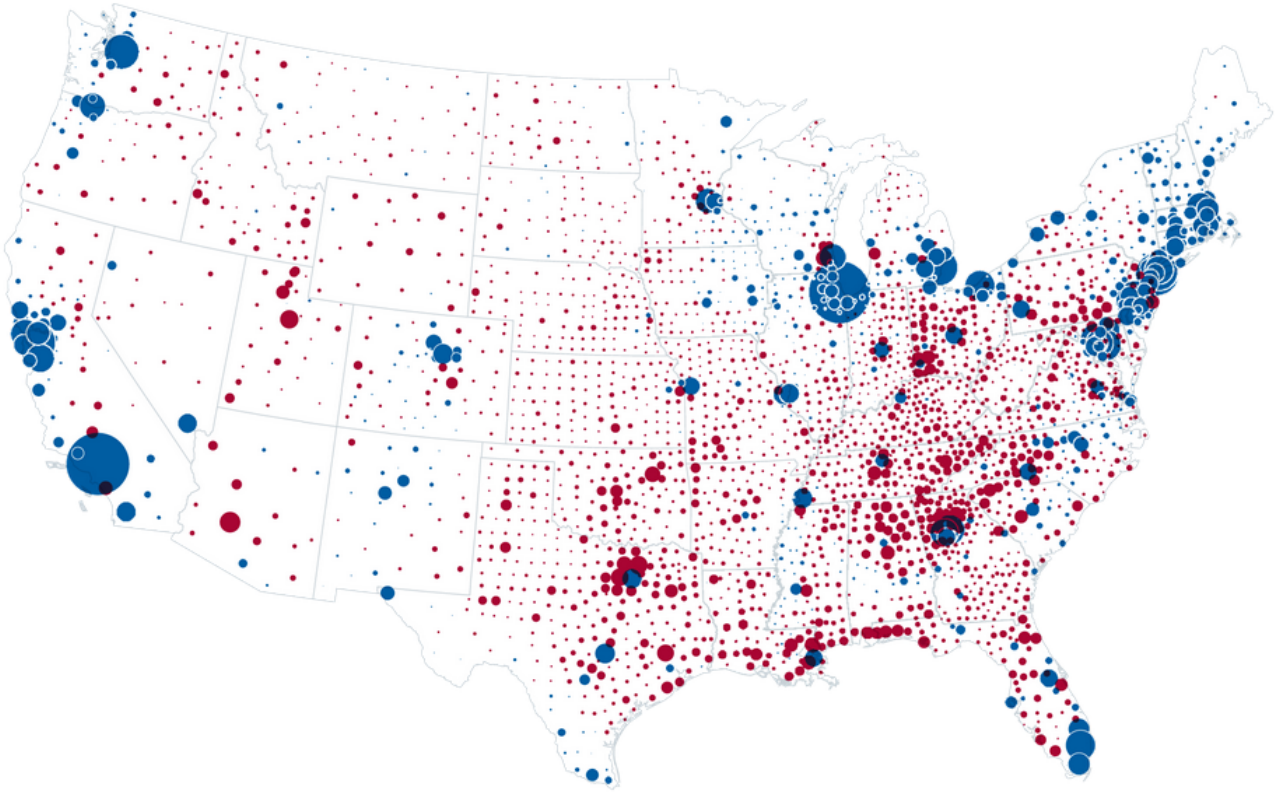
http://en.wikipedia.org/wiki/Misleading_graph

Color-coding geographic regions



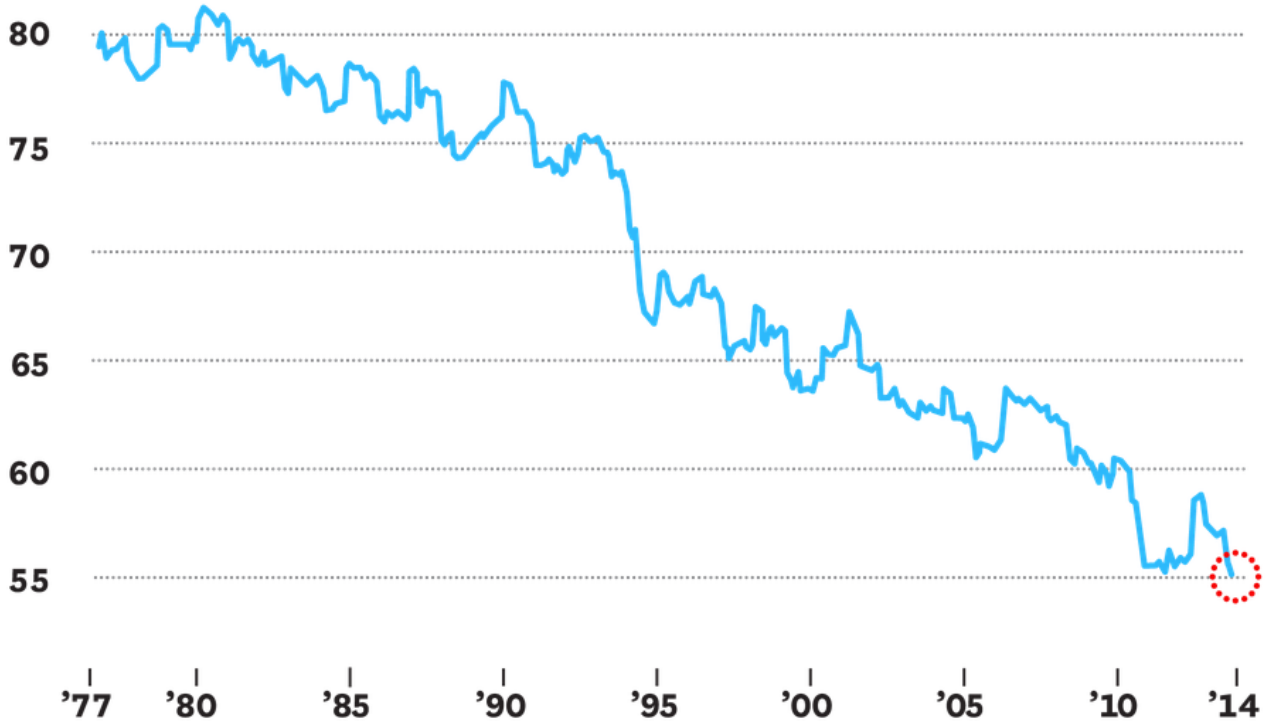
https://hbr.org/2014/12/vision-statement-how-to-lie-with-charts

Color-coding geographic regions



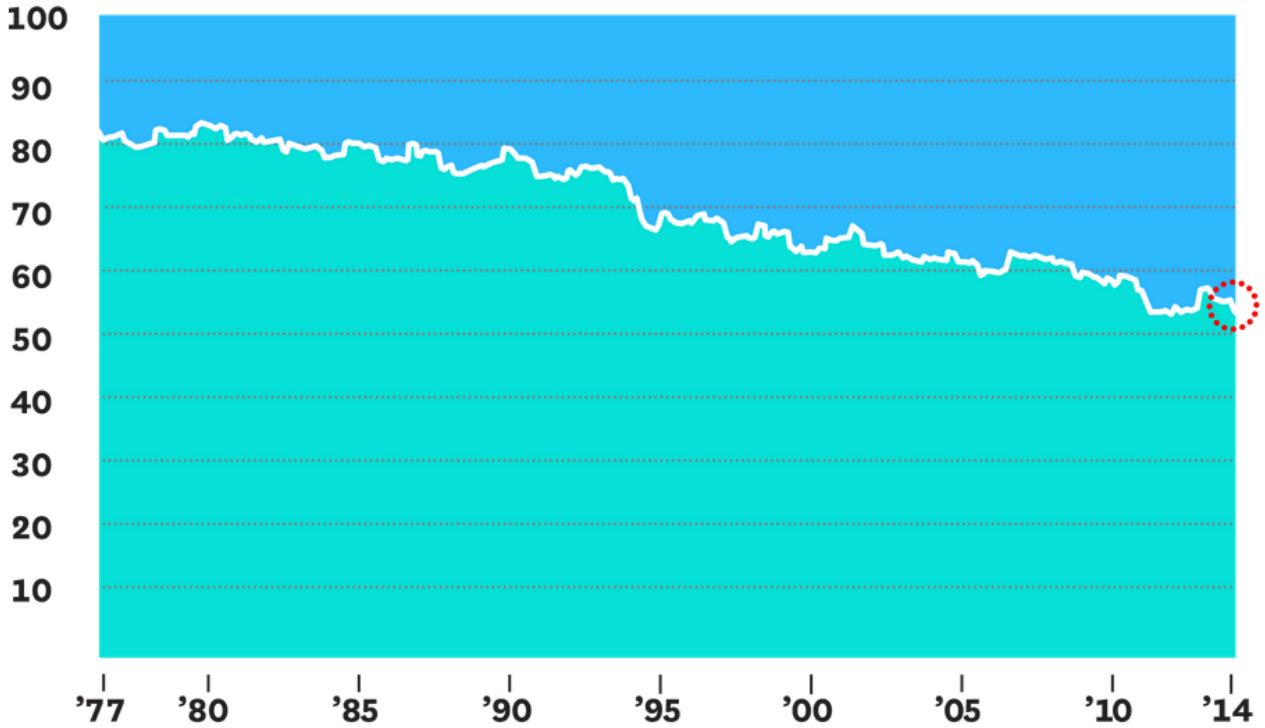
https://hbr.org/2014/12/vision-statement-how-to-lie-with-charts

Axis ranges



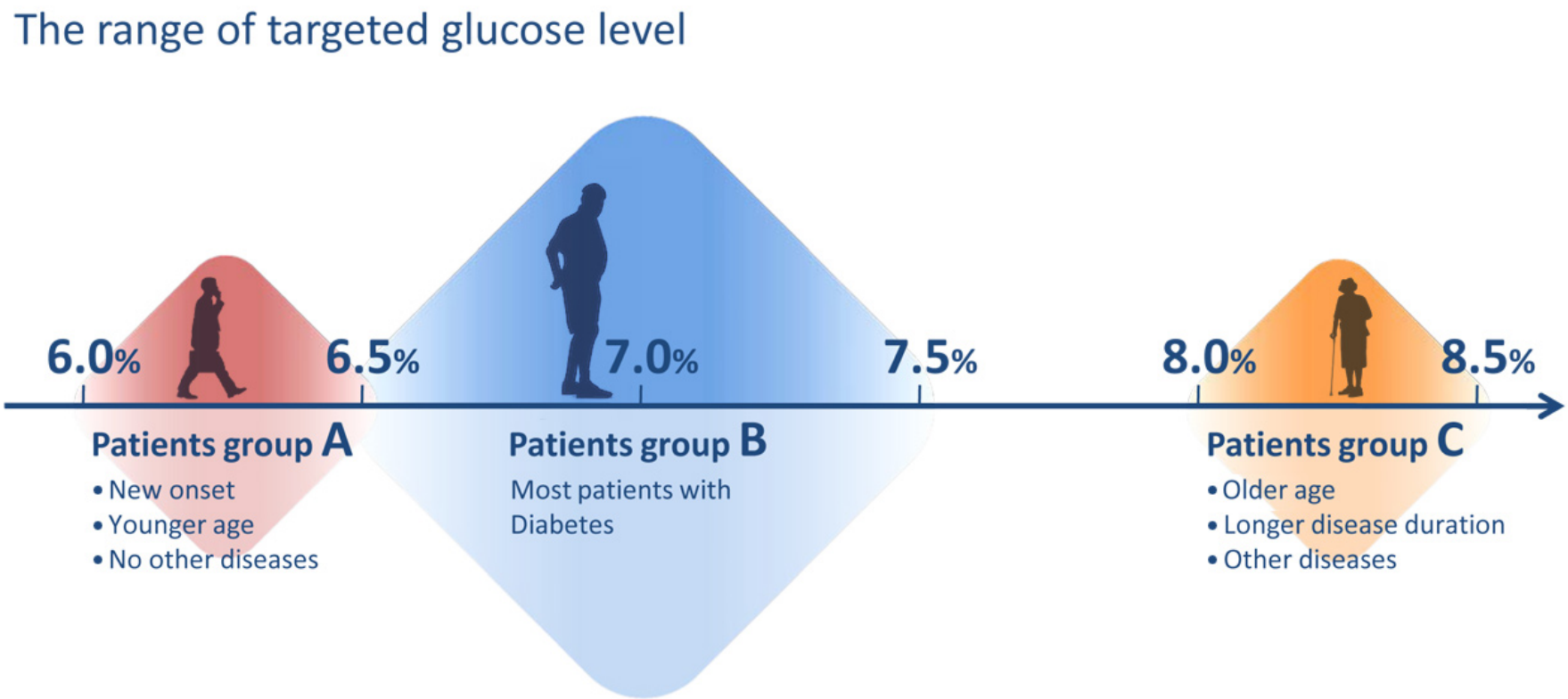
https://hbr.org/2014/12/vision-statement-how-to-lie-with-charts

Axis ranges



https://hbr.org/2014/12/vision-statement-how-to-lie-with-charts

Who understands?



http://www.multimension.com/project/upgrading-clinical-infographics/

Universiteit Leiden

Data Mining in a BI context

Overview

- Data warehouse
- Data preparation
- Data mining theory recap
- Data mining case studies
- Data mining evaluation techniques

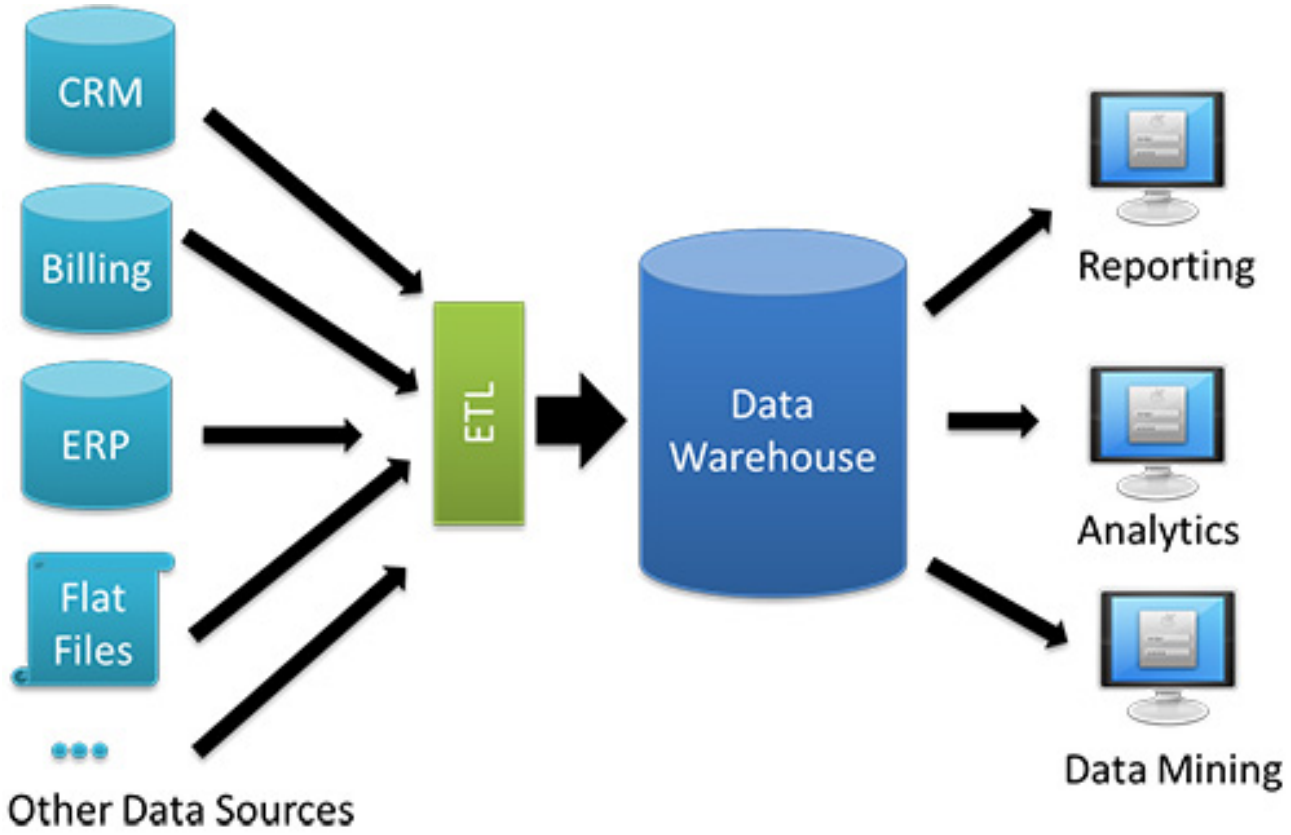
Data warehouse

- **Data warehouse**: a copy of transaction data specifically structured for query and analysis (R. Kimball)
- **Data warehouse**: a system used for reporting and data analysis (Wikipedia)
- **Data warehouse**: a subject oriented, integrated, nonvolatile, timestamped collection of data designed to support management's decision support needs (B. Inmon)

Data warehouse data

- In a data warehouse, data is organized around **subjects** (whereas information systems are organized around applications)
- Data is collected from heterogeneous sources and may already be aggregated (for example from an ERP or CRM system)
- Data is timestamped
- Data is nonvolatile

Data warehouse



http://savis.vn/

Transactional system vs. Data warehouse

Universiteit Leiden

Transactional System

- Holds current data
- Detailed data
- Volatile data
- High transaction frequency
- Oriented on daily operations
- Support for daily decisions
- Many operational users
- Availability very important
- Data storage focus

Data warehouse

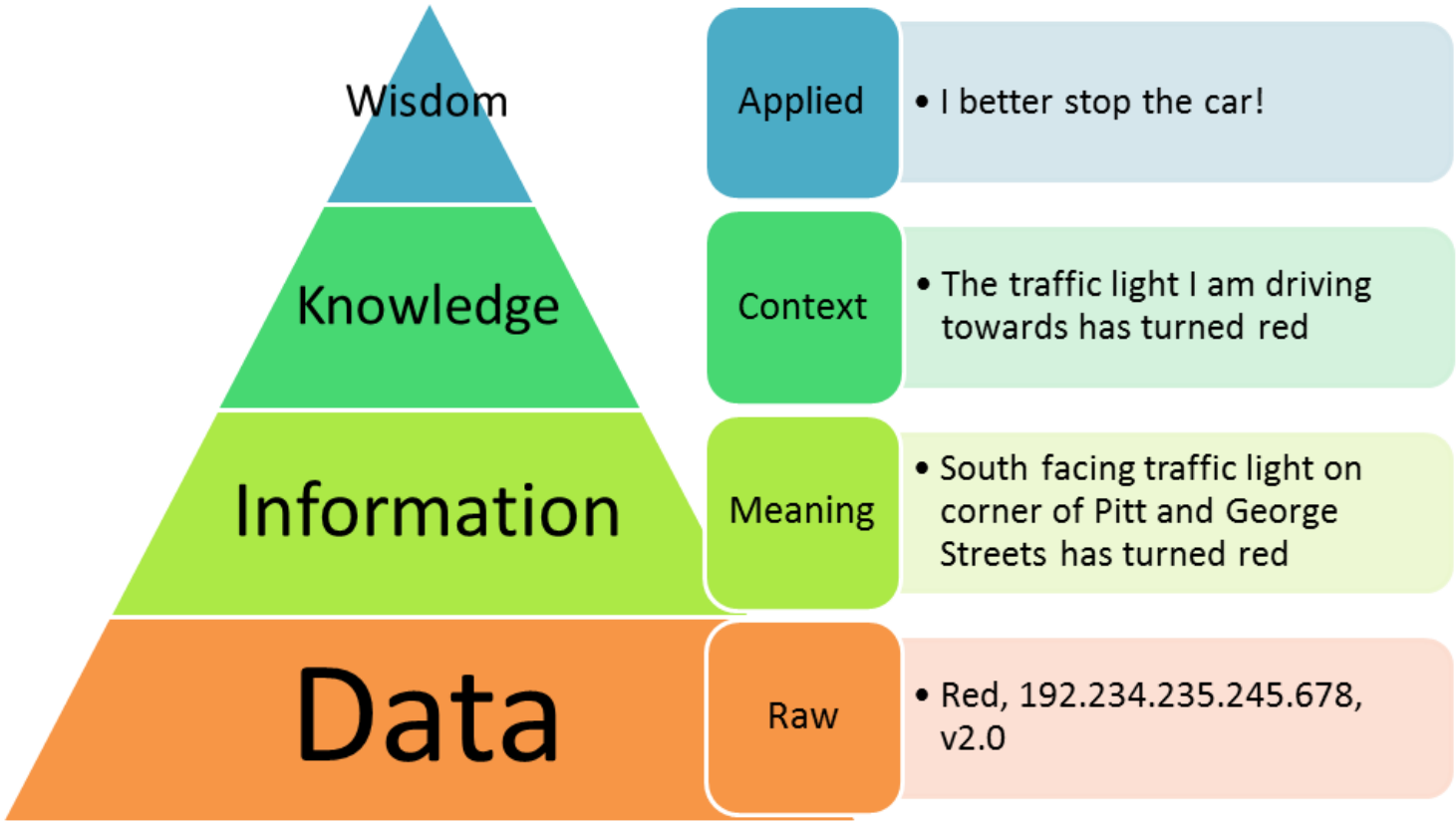
- Current and historic data
- Detailed and aggregated data
- Nonvolatile data
- Medium-low frequency
- Oriented on data analysis
- Support for strategic decisions
- Few decision-making users
- Availability not so important
- Information acquisition focus

`https://www.fer.unizg.hr/` (Business Intelligence)

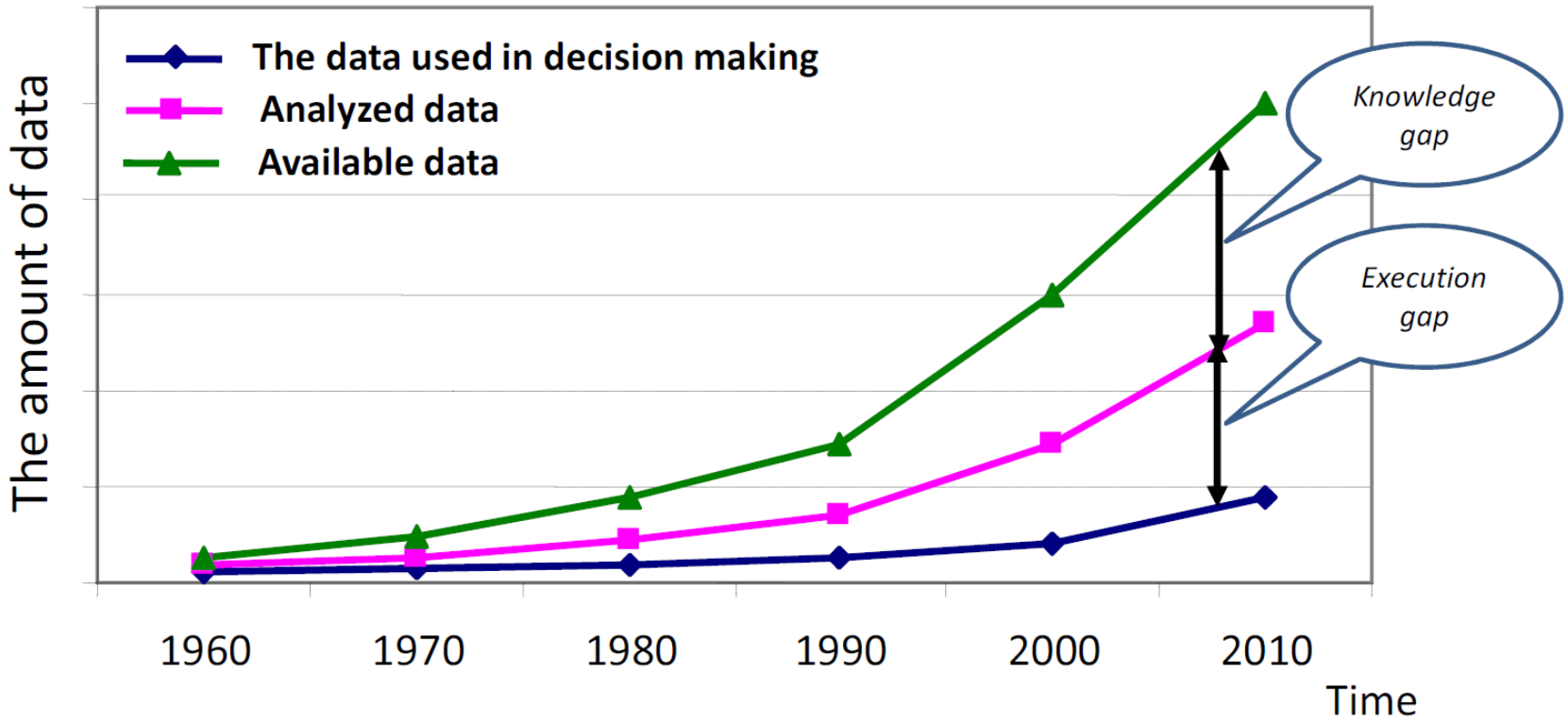
Data mining

- **Data mining**: the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems (Wikipedia)
- **Data mining**: the practice of examining large pre-existing databases in order to generate new information (Oxford)
- **Data mining**: knowledge discovery from data (or information) in an automated way (DIKW pyramid)

DIKW Pyramid



DIKW Gaps

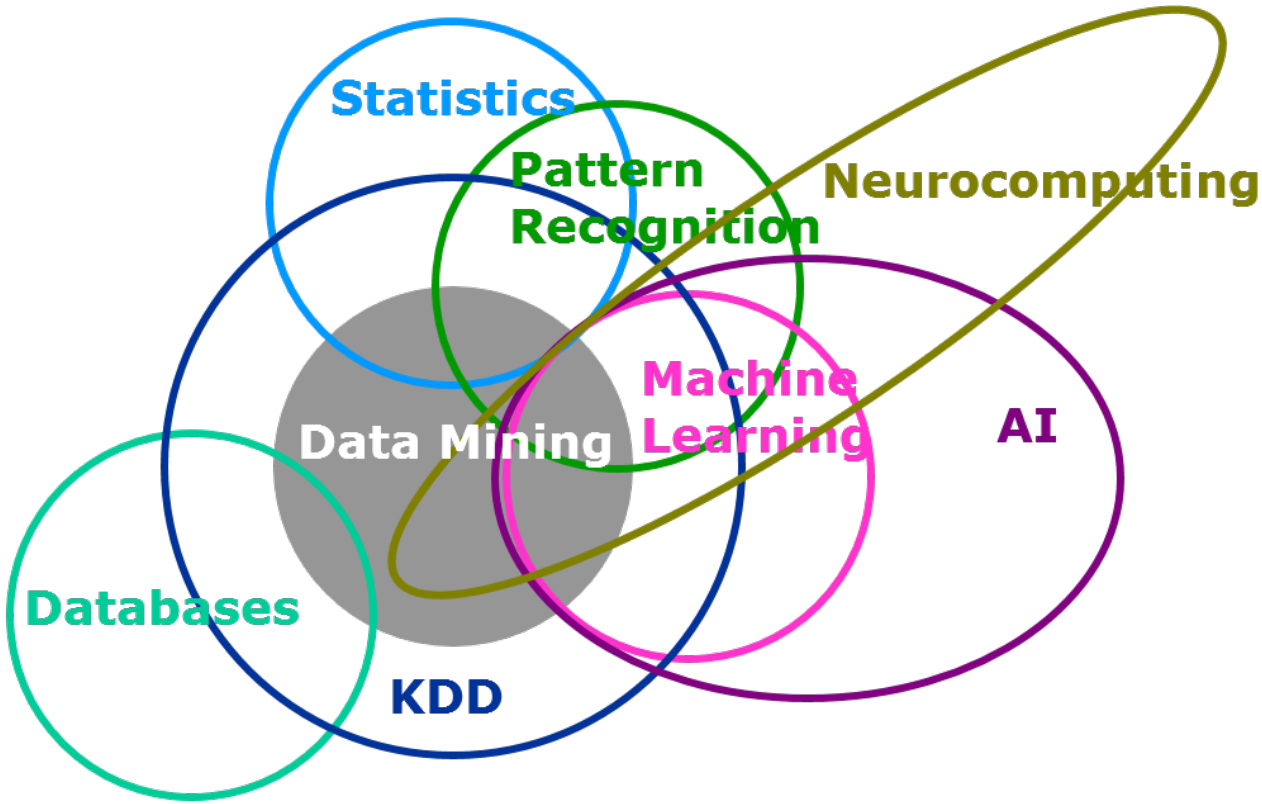


ZPR FER Zagreb - Business Intelligence 20113

Data mining . . .

- KDD: Knowledge Discovery in Databases
- Data archeology
- Information harvesting
- Knowledge extraction
- Machine learning
- Big data techniques?
- Data science?
- Business intelligence?

Data mining



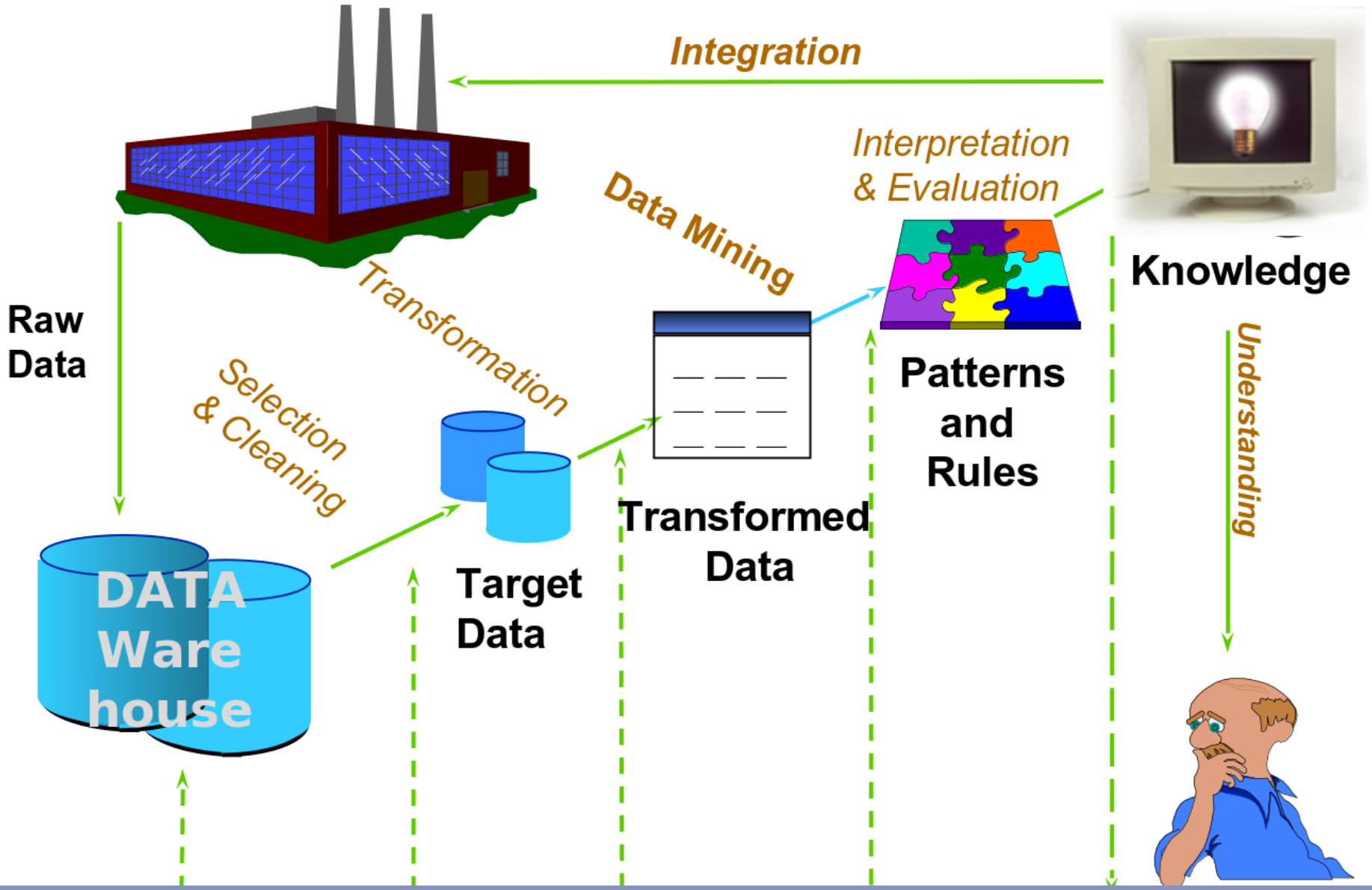
http://blogs.sas.com/content/subconsciousmusings/2014/08/22

KDD

- **Knowledge Discovery in Data** is the
  - non-trivial process of identifying
  - valid,
  - novel,
  - potentially useful
  - and ultimately understandable

  patterns in data.
- Fayyad et al., *Advances in knowledge discovery and data mining*, MIT press, 1996.

KDD


Integration
Interpretation
& Evaluation
Data Mining
Knowledge
Raw
Data
Transformation
Selection
& Cleaning
Patterns
and
Rules
Understanding
Transformed
Data
Target
Data
DATA
Ware
house

Why data mining now?

- Data flood / data explosion
- Cloud computing power
- Cheap storage
- Algorithms have matured
- Software is available
- Competition is killing

Data mining in businesses

- Process management
- Market basket analysis
- Marketing
- Customer loyalty
- Fraud detection
- Trend analysis

Data mining in practice

1. Learn about the problem domain
2. Data selection
3. Data, cleaning, preprocessing and reduction
4. Data mining
5. Interpretation of information
6. Apply knowledge in domain

Data preprocessing

- Sampling
- Normalization
- Missing data
- Data conflicts
- Duplicate data
- Ambiguity in data

Guidelines for successful data mining

- The data must be available
- The data must be relevant, adequate and clean
- There must be a well-defined problem
- The problem should not be solvable by means of ordinary query or OLAP tools
- The results must be actionable

Successful data mining in businesses

- Use a small team with a strong internal integration and a loose management style
- Carry out a small pilot project before a major data mining project
- Identify a clear problem owner responsible for the project, e.g., from sales or marketing
- Try to realize a positive return on investment within 6 to 12 months
- Have top management back the project up

Universiteit Leiden

Break?

Data attribute types

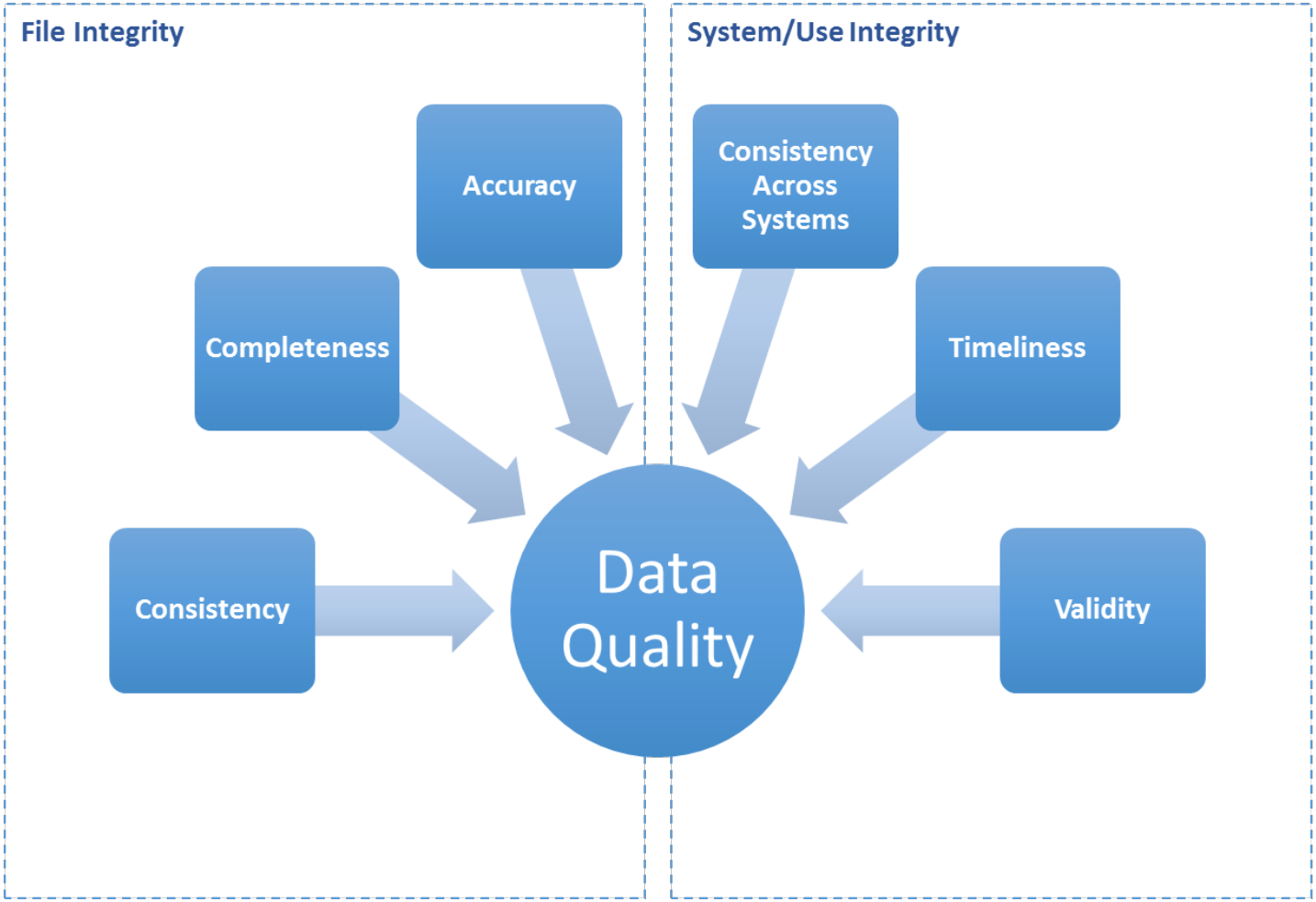
- **Categorical** attributes: discrete
  - **Nominal** attribute: has no logical ordering
    (e.g., colors or names)
  - **Ordinal** attribute: has ordering
    (e.g.: bad, OK, good, perfect)
- **Numerical** attributes continuous
  (e.g., 4.815m and EUR 162 342)

Data quality

- Accuracy
- Completeness
- Consistency (uniformity)
- Validity
- Timeliness

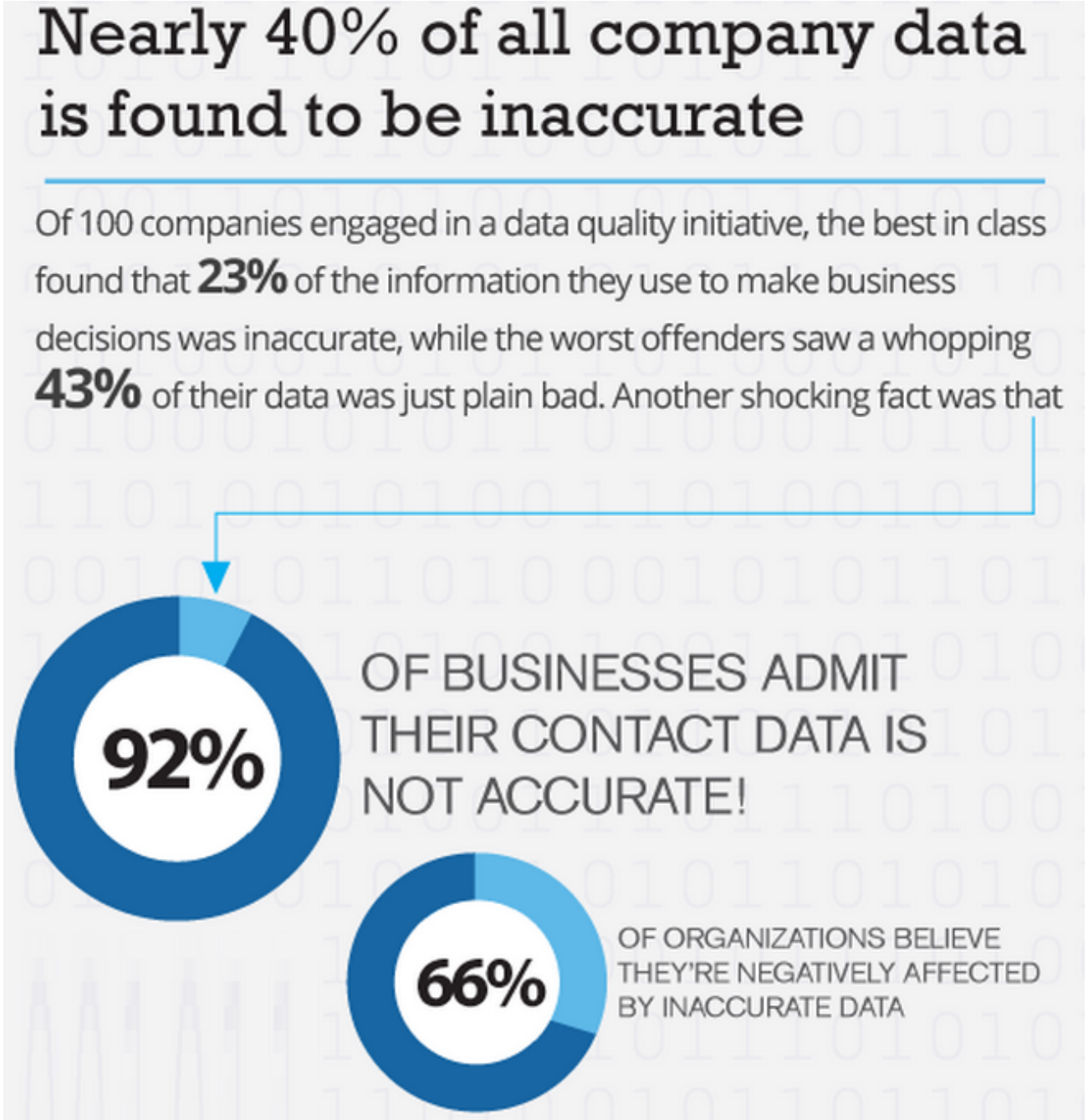
- **Data cleaning**, data cleansing, data scrubbing, . . .

Data quality



http://www.hicxsolutions.com/supplier-management-programmes/

http://halobi.com/wp-content/uploads/data-quality-infographic.png



http://halobi.com/wp-content/uploads/data-quality-infographic.png

Universiteit Leiden

Example: Corporate data quality

Data quality

- ORBIS database (Bureau van Dijk, `http://orbis.bvdinfo.com`)
- Aggregates data from Chambers of Commerce across the world
- Snapshot from September 2015

Data quality

- ORBIS database (Bureau van Dijk, `http://orbis.bvdinfo.com`)
- Aggregates data from Chambers of Commerce across the world
- Snapshot from September 2015
- Extracted all firms (including meta-data such as operating revenue, employees, assets and market capitalization)

Data quality

- ORBIS database (Bureau van Dijk, `http://orbis.bvdinfo.com`)
- Aggregates data from Chambers of Commerce across the world
- Snapshot from September 2015
- Extracted all firms (including meta-data such as operating revenue, employees, assets and market capitalization)
- 140,087,471 firms found.

Data quality

- ORBIS database (Bureau van Dijk, `http://orbis.bvdinfo.com`)
- Aggregates data from Chambers of Commerce across the world
- Snapshot from September 2015
- Extracted all firms (including meta-data such as operating revenue, employees, assets and market capitalization)
- 140,087,471 firms found. Is that all?

Observed data

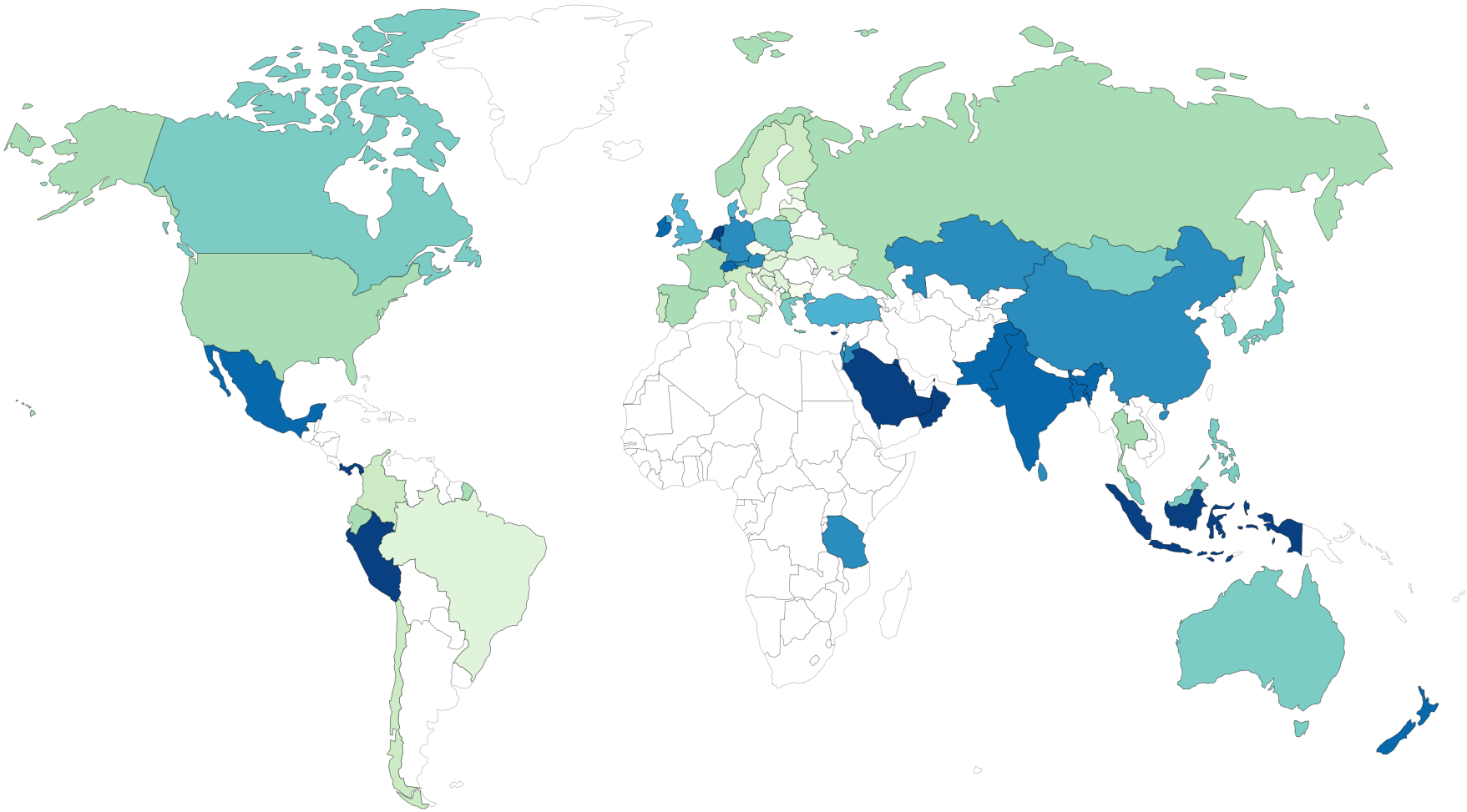


Figure : Observed average revenue per country (darker is more)

Completeness per size category

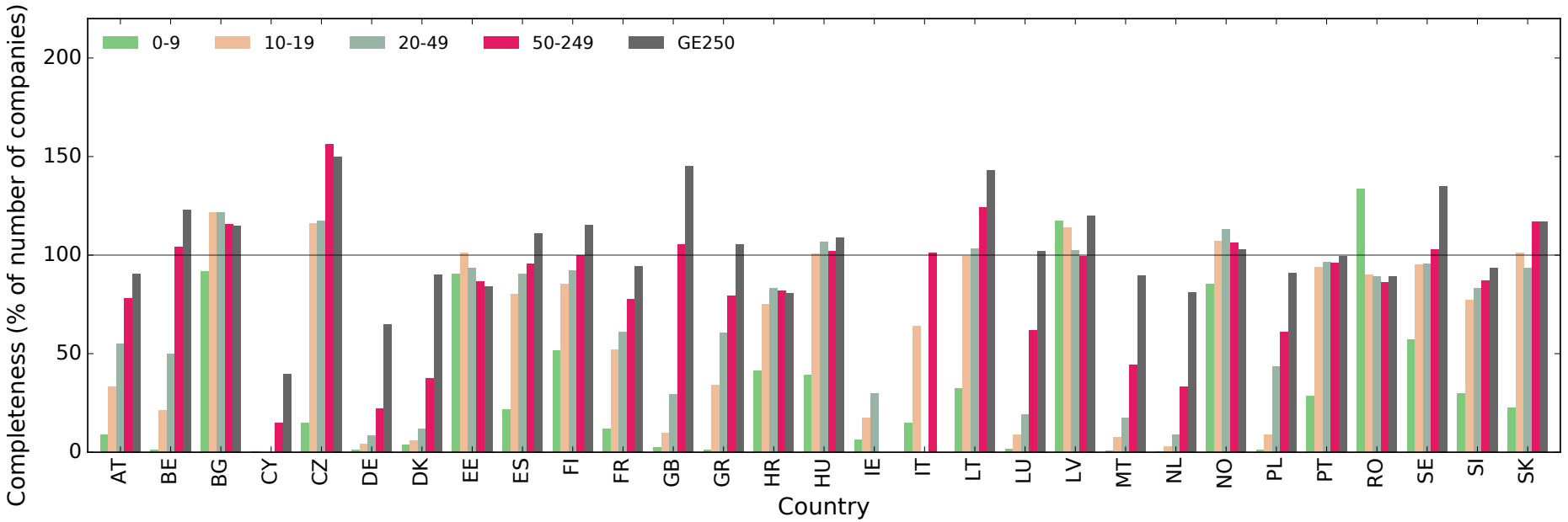


Figure : Percentage of companies present, segmented by number of employees.

Assessing completeness

- Lognormal distribution for firm revenue in a country
- Idea: fix distribution scale based on known high quality countries
- Estimate mean revenue for each country using World Bank indicators

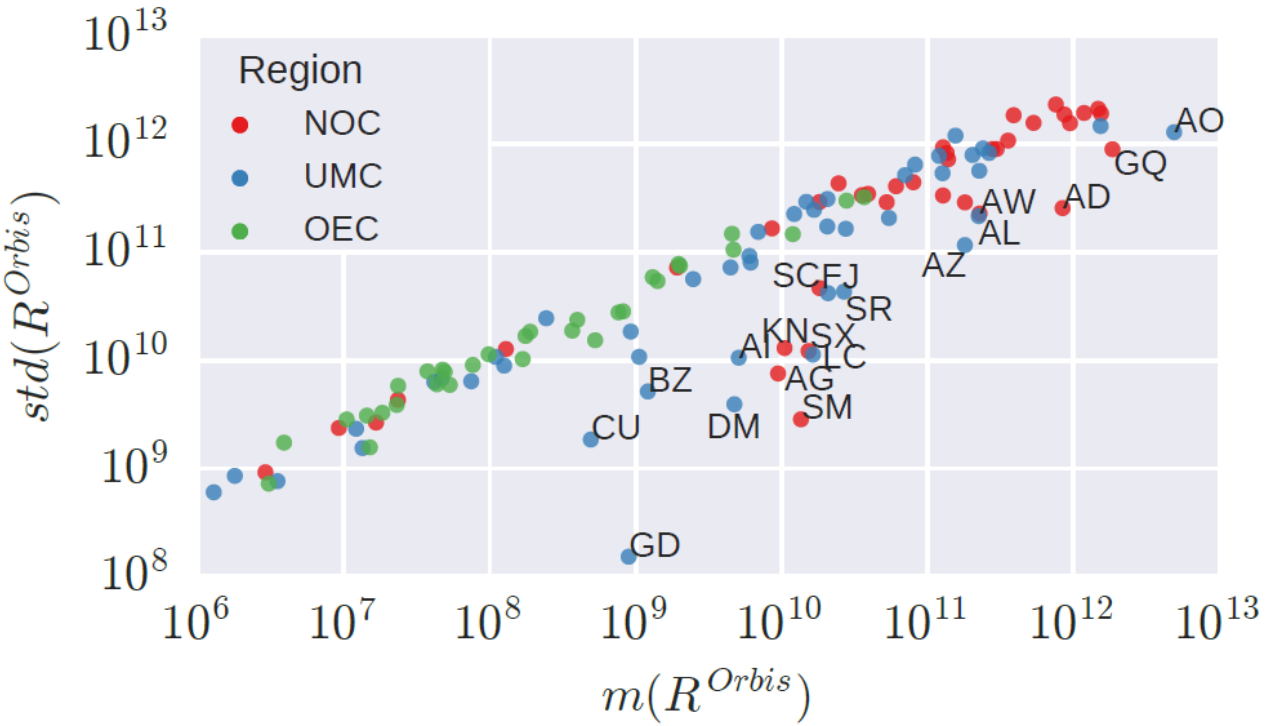
Assessing completeness

- Lognormal distribution for firm revenue in a country
- Idea: fix distribution scale based on known high quality countries
- Estimate mean revenue for each country using World Bank indicators
- Result: `GDP per capita` $\sim$ `Mean revenue`

Assessing completeness

- Lognormal distribution for firm revenue in a country
- Idea: fix distribution scale based on known high quality countries
- Estimate mean revenue for each country using World Bank indicators
- Result: `GDP per capita` $\sim$ `Mean revenue`
- `Mean revenue` $\sim$ Distribution location
- Assess completeness by comparing observed average revenue with estimated average revenue

Mean vs. standard deviation



Understanding average revenue

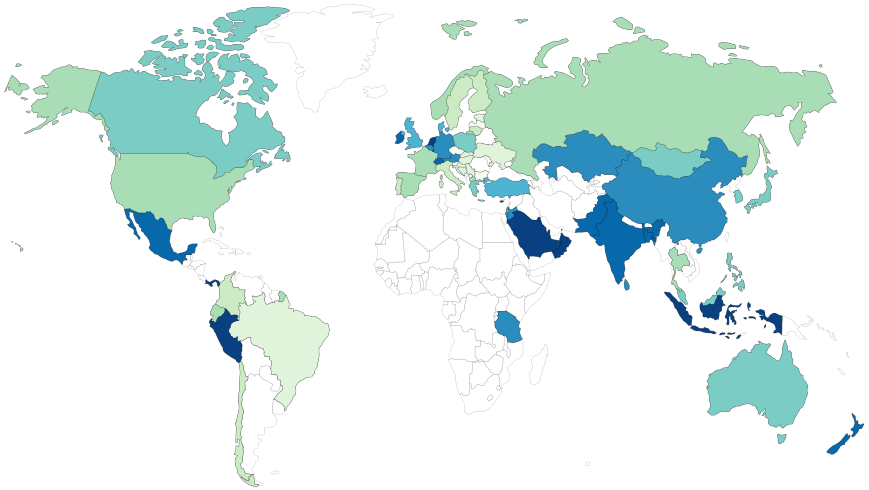


Figure : Observed average revenue

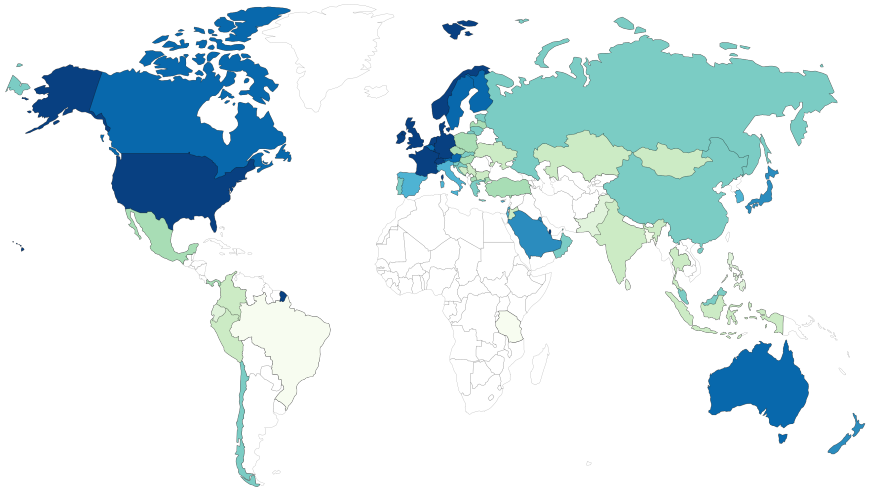
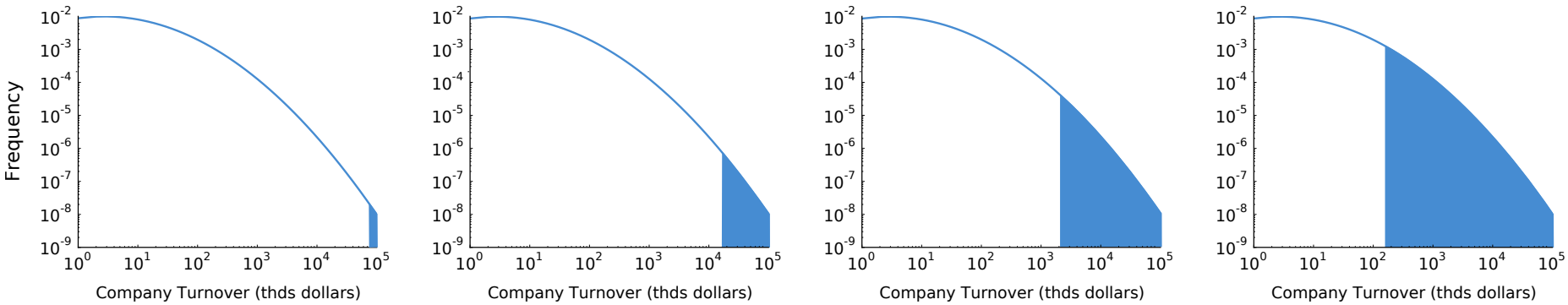
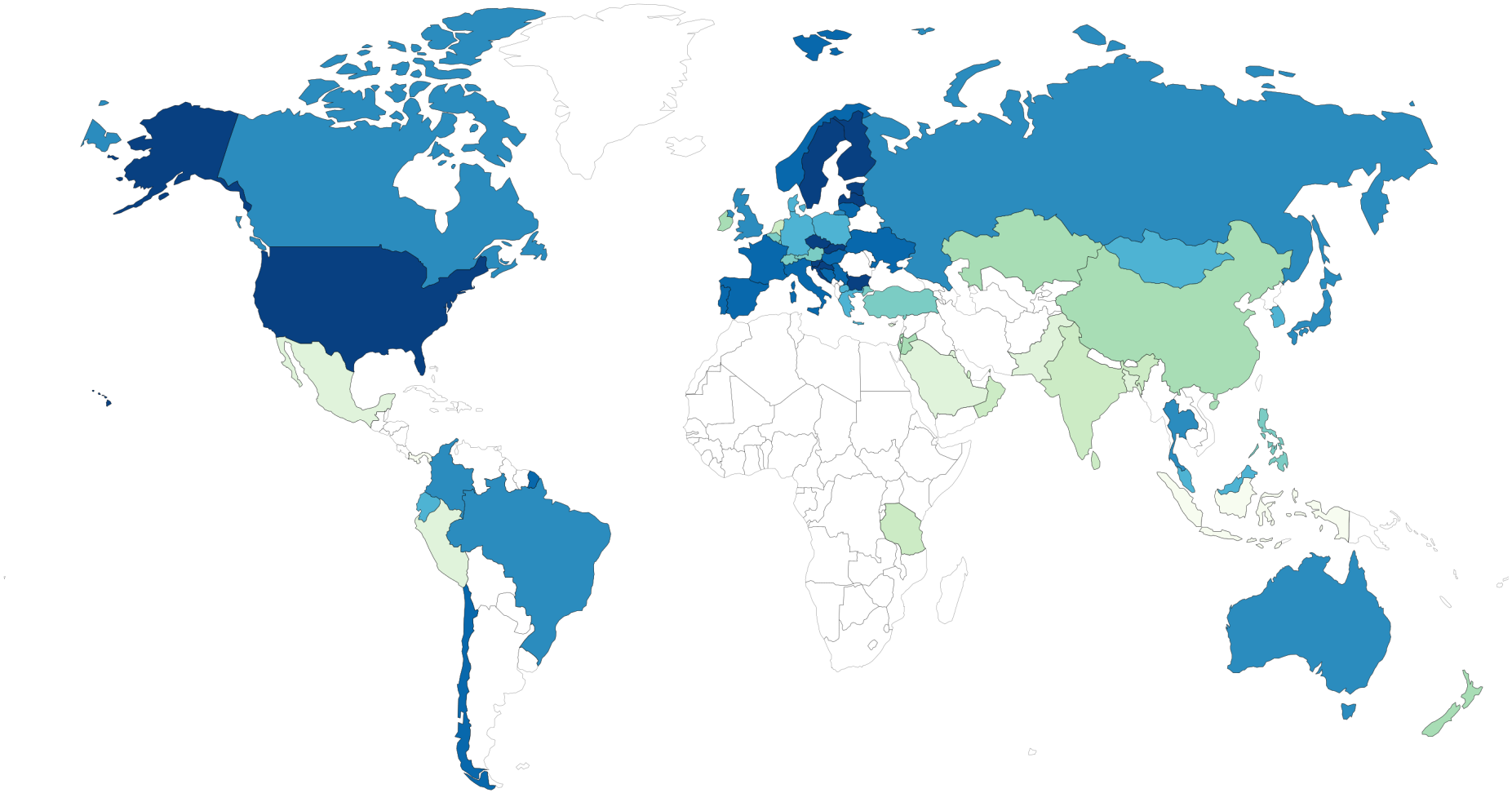


Figure : Estimated average revenue

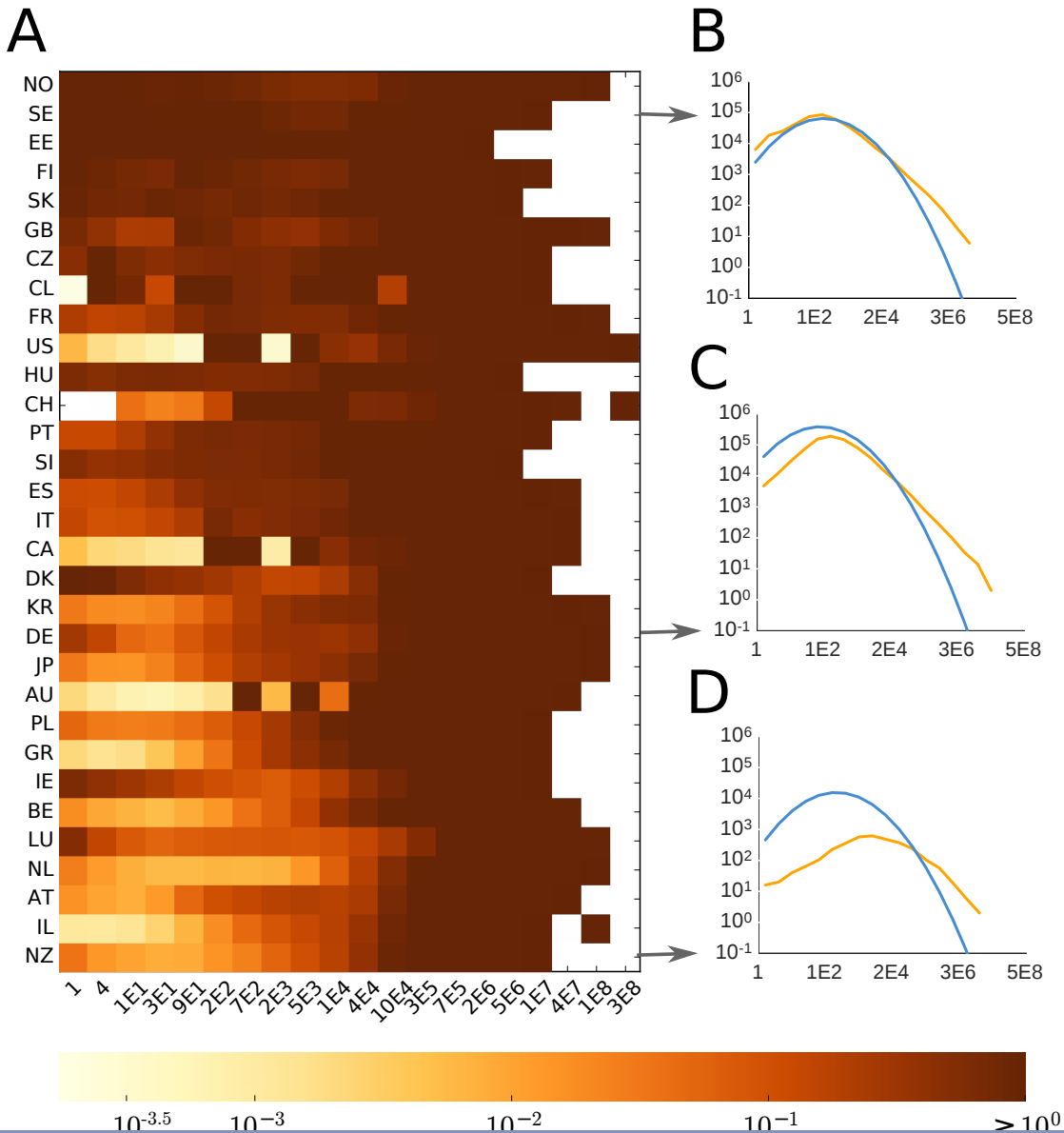
Low average in rich countries



Real completeness



Completeness per country



Categories of techniques

- Machine learning
  - Supervised learning: learning on labeled data
  - Semi-supervised learning: partially labeled data
  - **Unsupervised learning**: leaning/mining on unlabeled data
  - Reinforcement learning: agents learning to act in an environment

Universiteit Leiden

Unsupervised learning

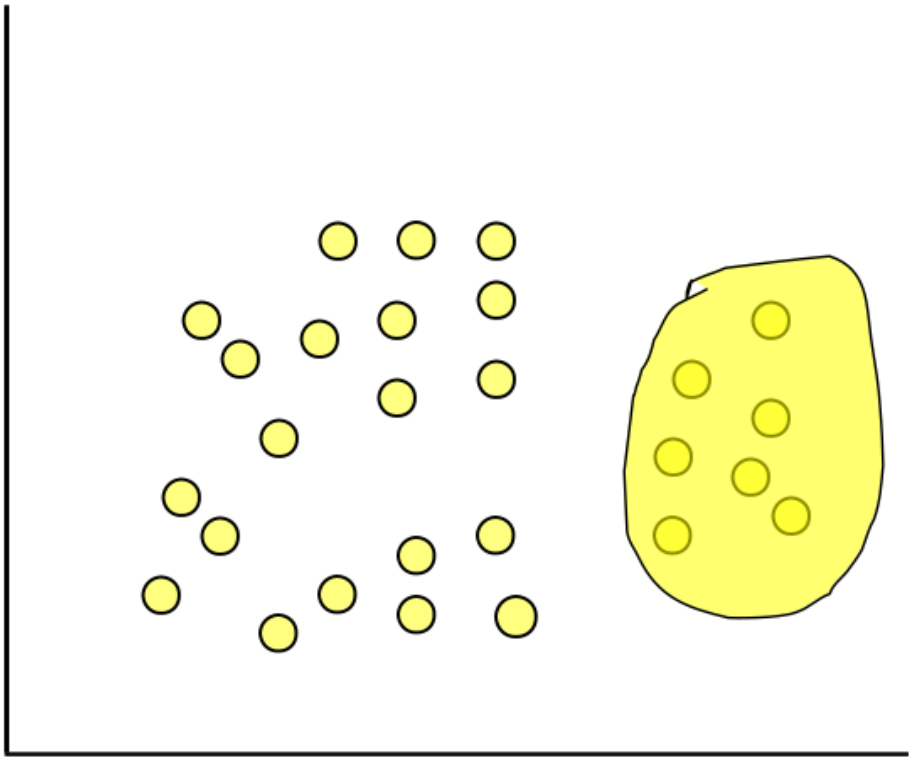
Categories of techniques

- **Unsupervised learning: leaning/mining on unlabeled data**
- Supervised learning: learning on labeled data
- Semi-supervised learning: partially labeled data
- Reinforcement learning: agents learning to act in an environment

Unsupervised learning

- **Clustering**
- Anomaly detection
- Pattern recognition
- Data summarization

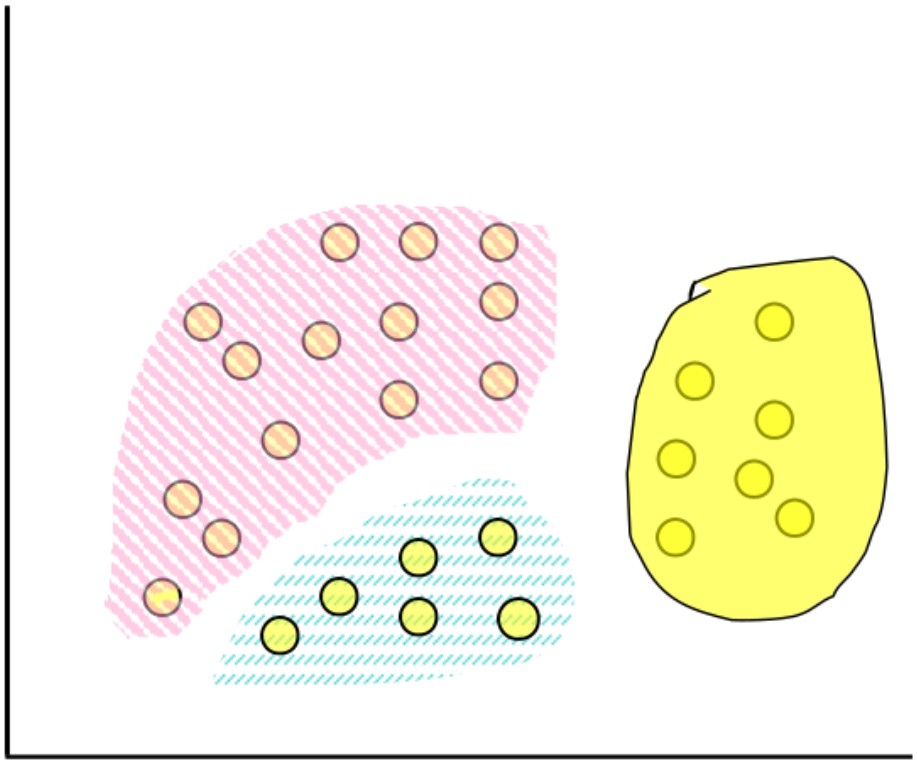
Clustering



Clustering

- Data is unlabeled
- Label data: grouping

Clustering



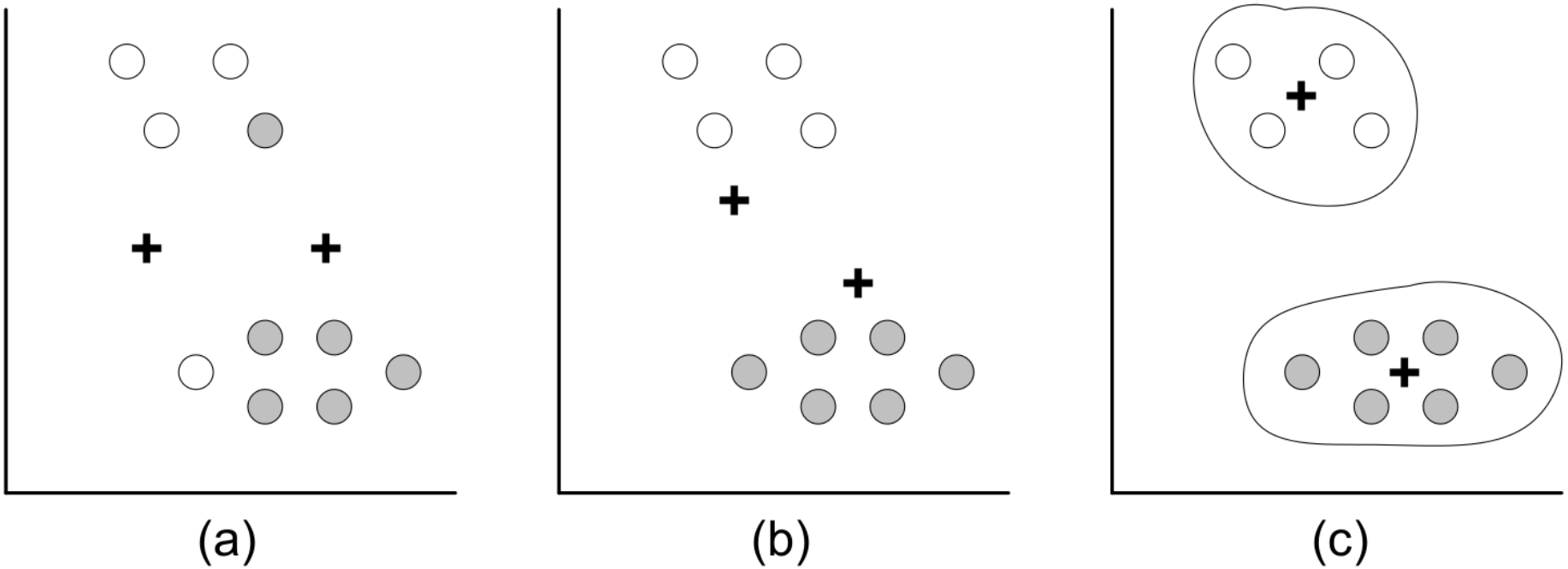
Clustering

- Data is unlabeled
- Label data: grouping
- Grouping based on similar attributes: relatively close "neighbors" in n-dimensional space

k-means Clustering

1. k **means** are randomly placed
2. k clusters are created by assigning each observation to the nearest mean (according to some **distance notion**)
3. the **centroid** of each cluster becomes the new mean
4. steps 1–3 are repeated until convergence

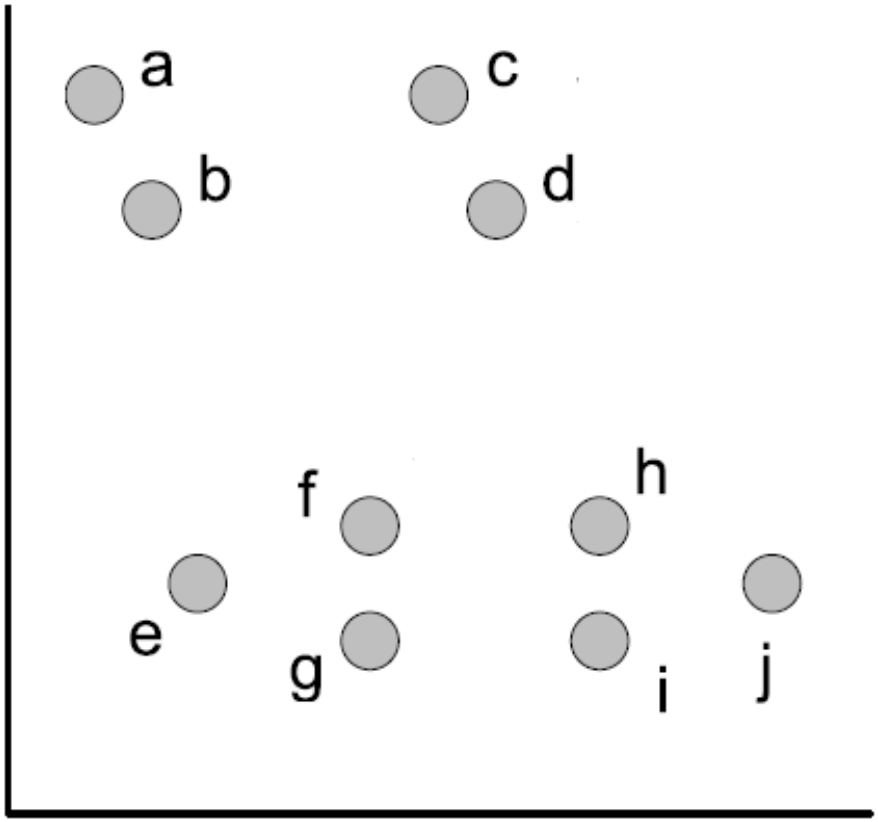
k-means Clustering



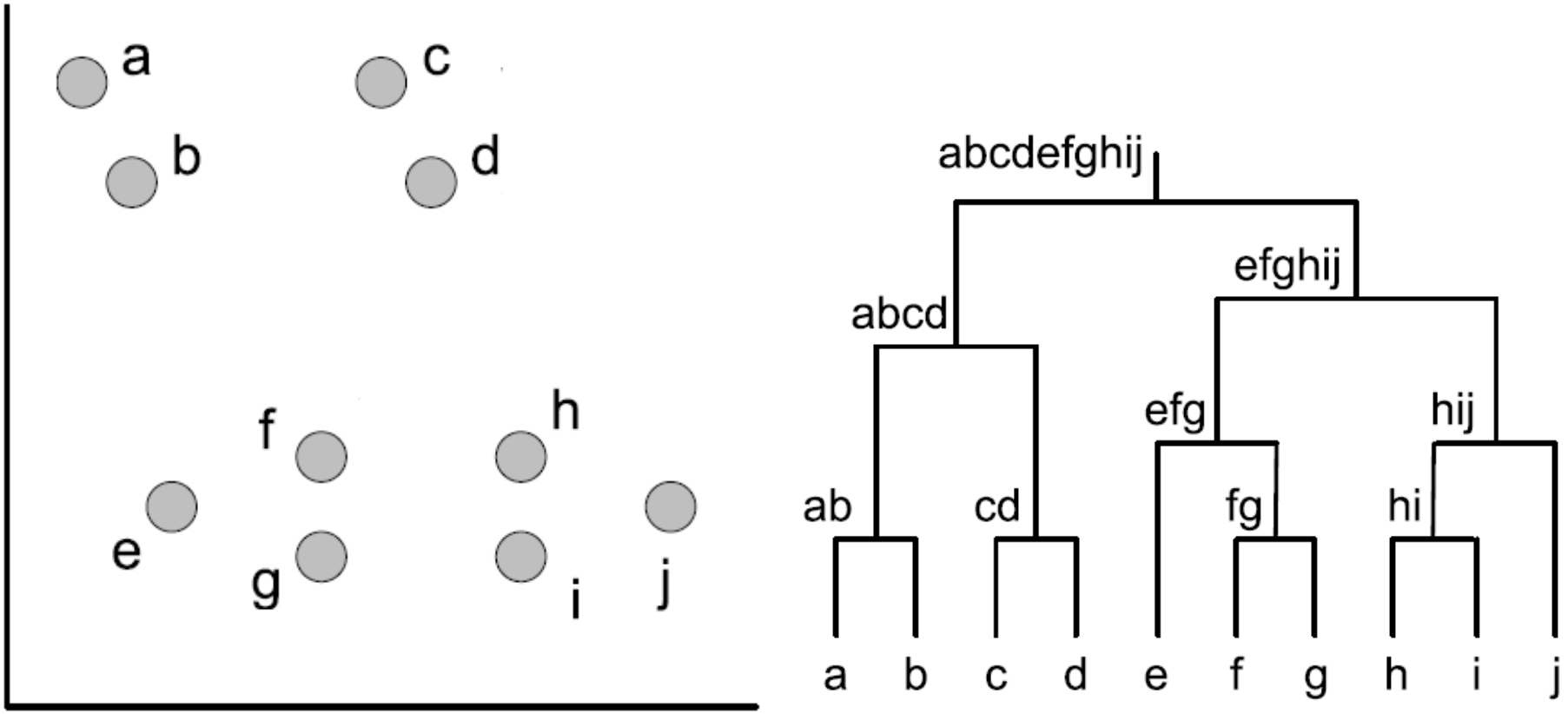
Hierarchical clustering

1. Define a distance function between objects
2. Assign each object to its own cluster
3. Merge the two nearest clusters (based on distance between its objects) into one cluster
4. Until there is only one cluster, go to **3**
5. Pick a level in the resulting dendogram as the preferred method of clustering

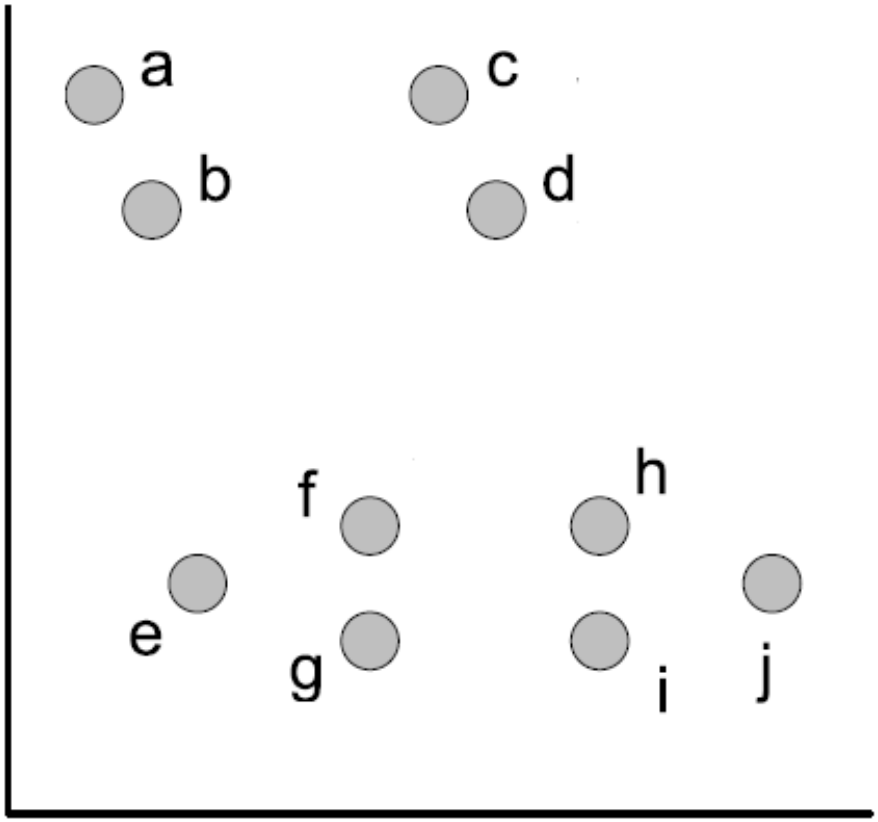
Hierarchical clustering

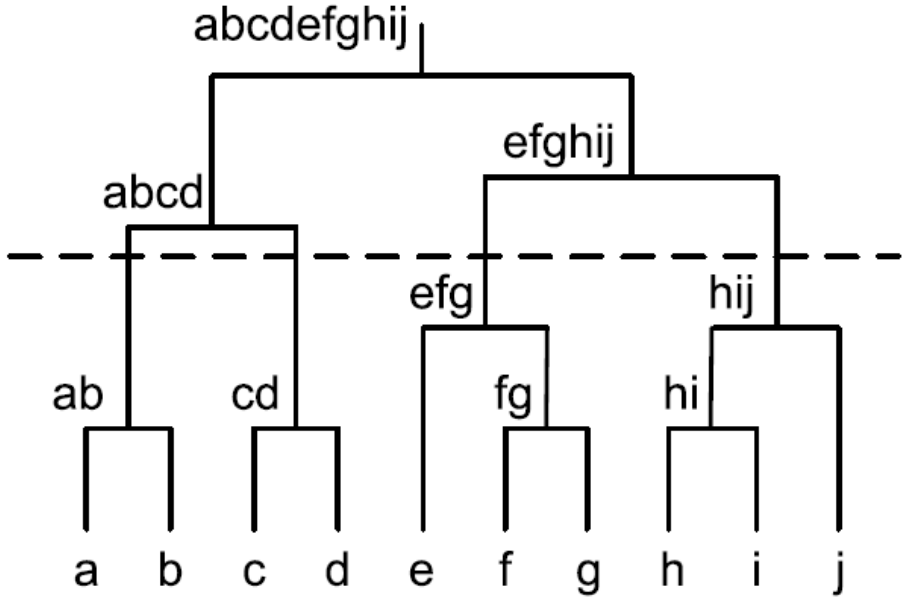


Hierarchical clustering

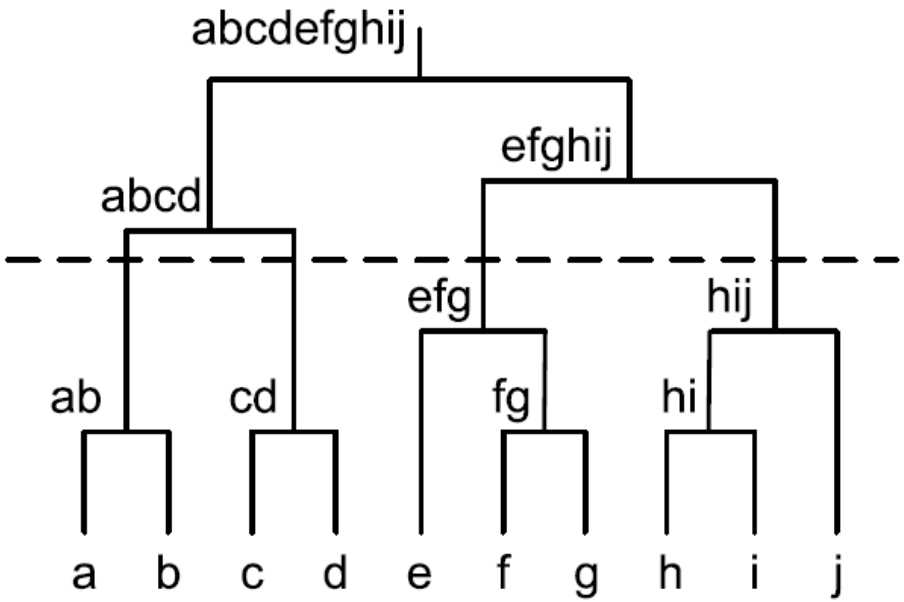


Hierarchical clustering

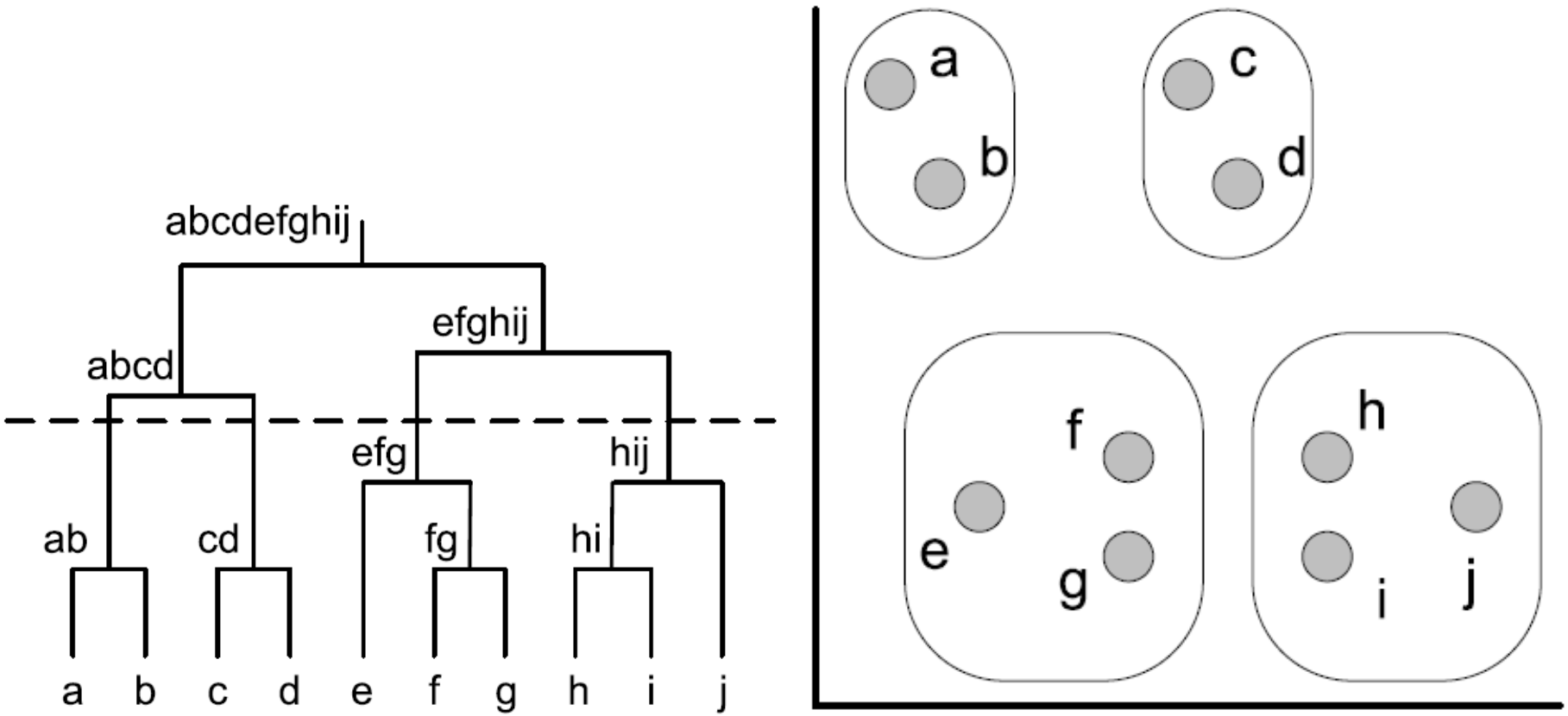




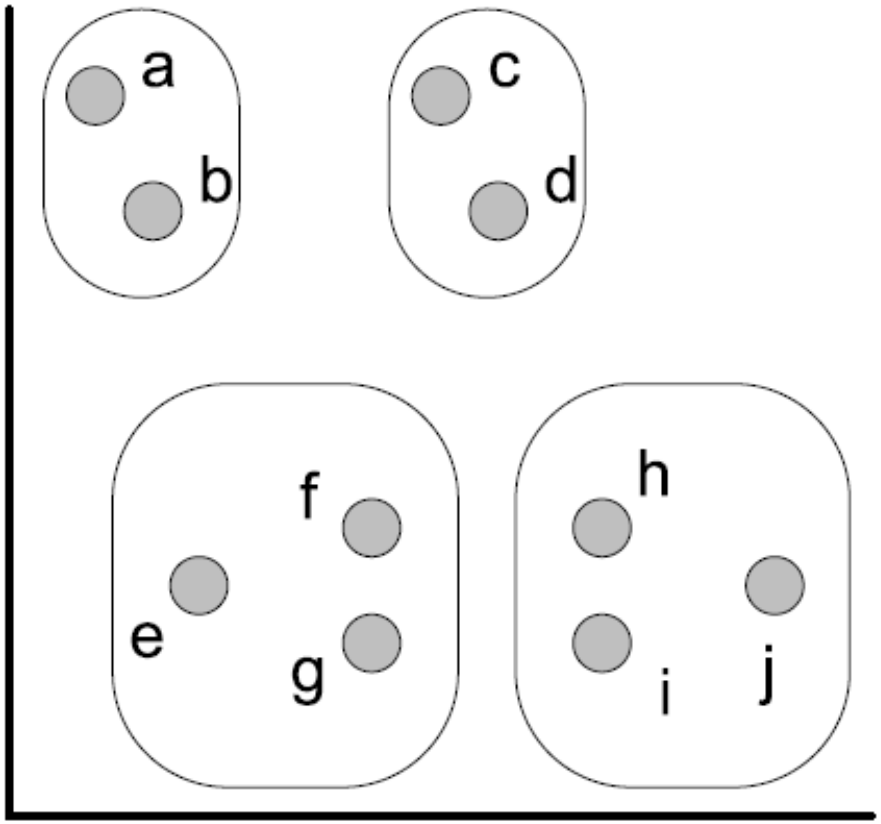
Hierarchical clustering



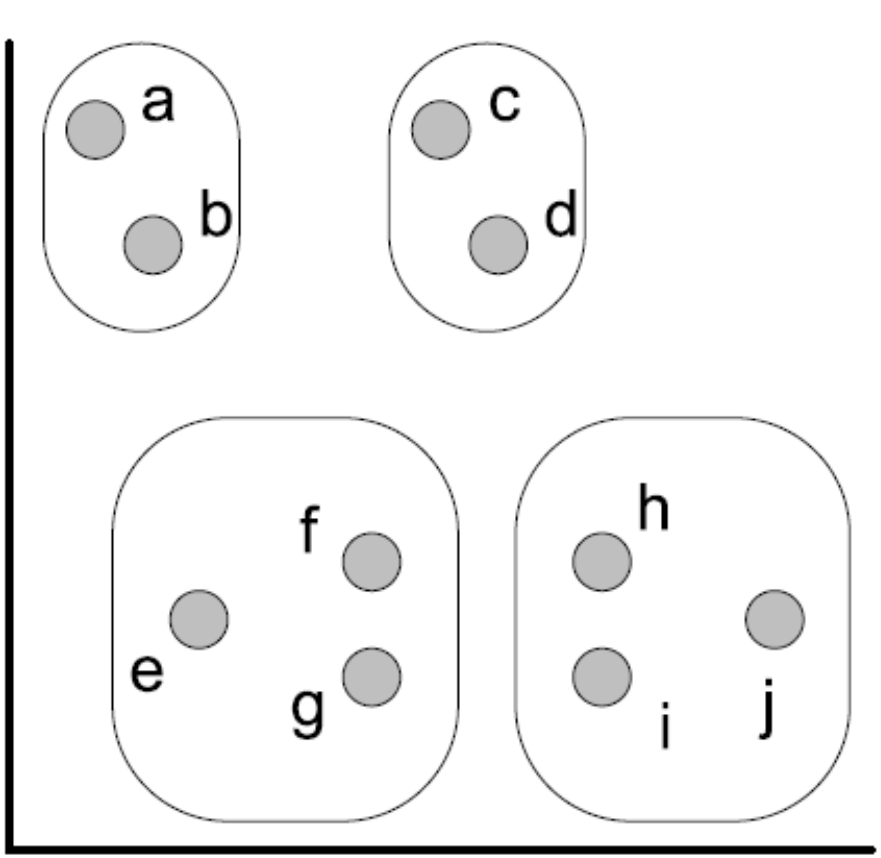
Hierarchical clustering

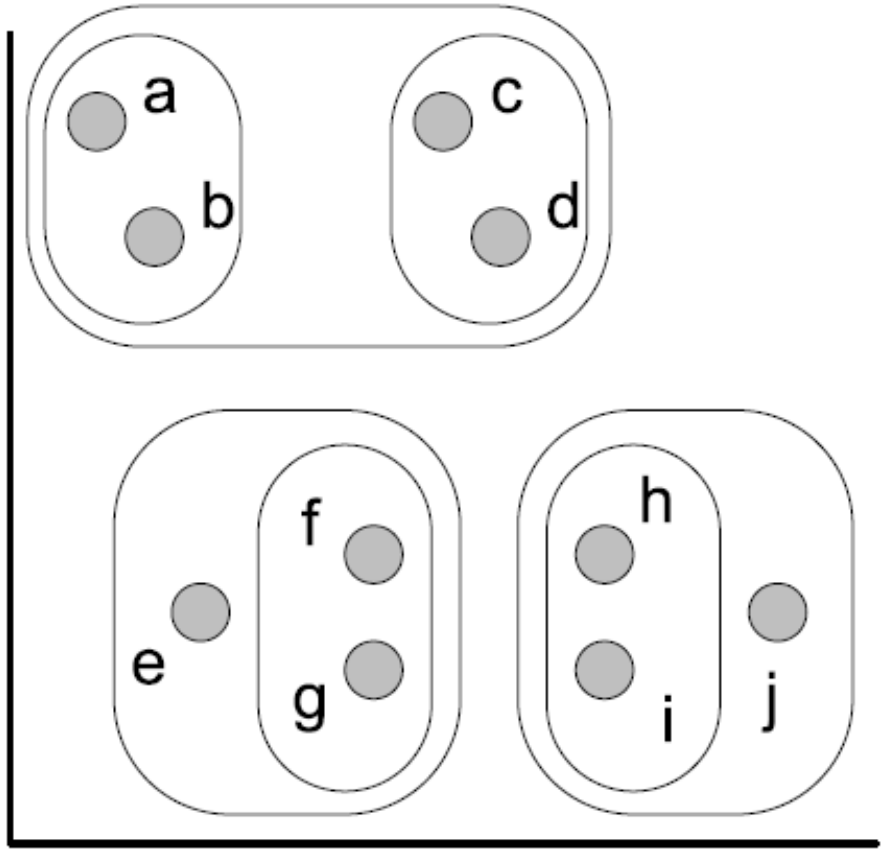


Hierarchical clustering

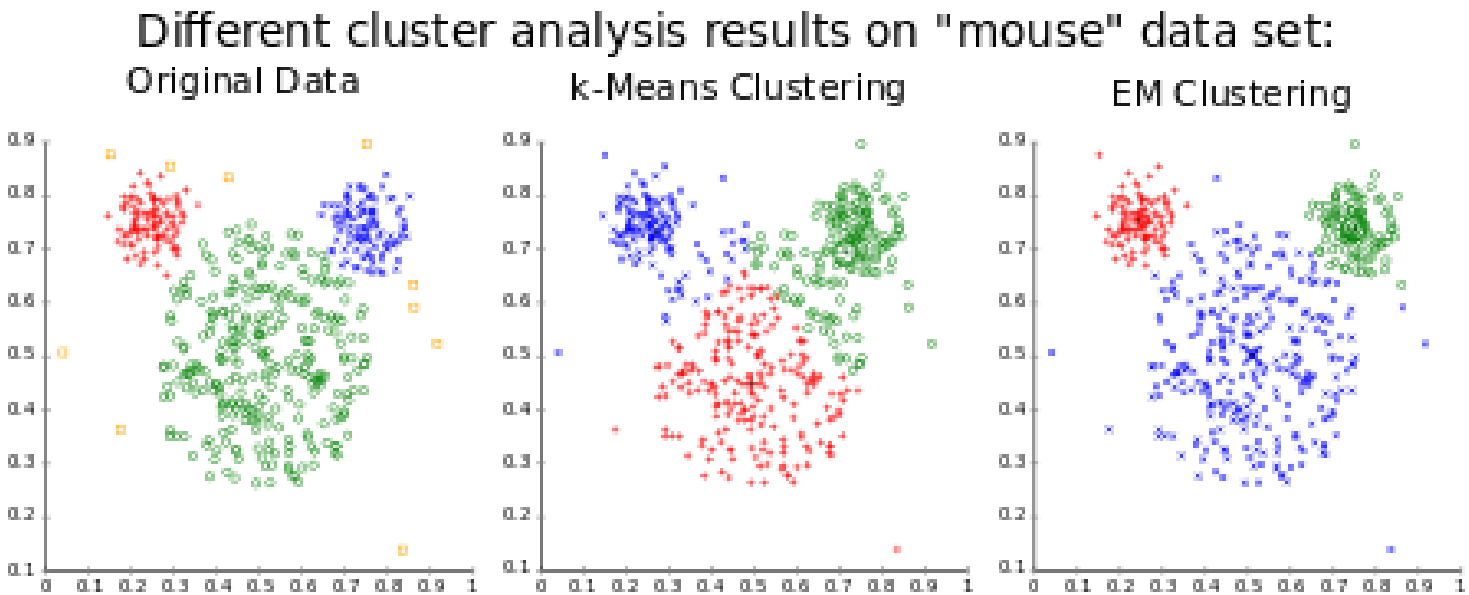


Hierarchical clustering





Clustering validation



Expectation-Maximization (EN) clustering: `https://en.wikipedia.org/wiki/Expectation-maximization_algorithm`

Hierarchical vs. k-means clustering

- Time complexity (linear vs. quadratic)
- Predefined number of clusters
- Influence of outliers
- Assumption of the presence of a hierarchical structure

Case: Anomalies in energy expenditure

Case: anomalies in energy expenditure

- BSc project J. Kalmeijer in cooperation with "Rijkswaterstaat"
- Total of 254 objects all over the Netherlands
- Energy expenditure over 3 years known for each object
- Measurements every 15 minutes:
  365 days $\times$ 24 hours $\times$ 4 measurements $\approx$ 35.000 yearly measurements

Rijkswaterstaat

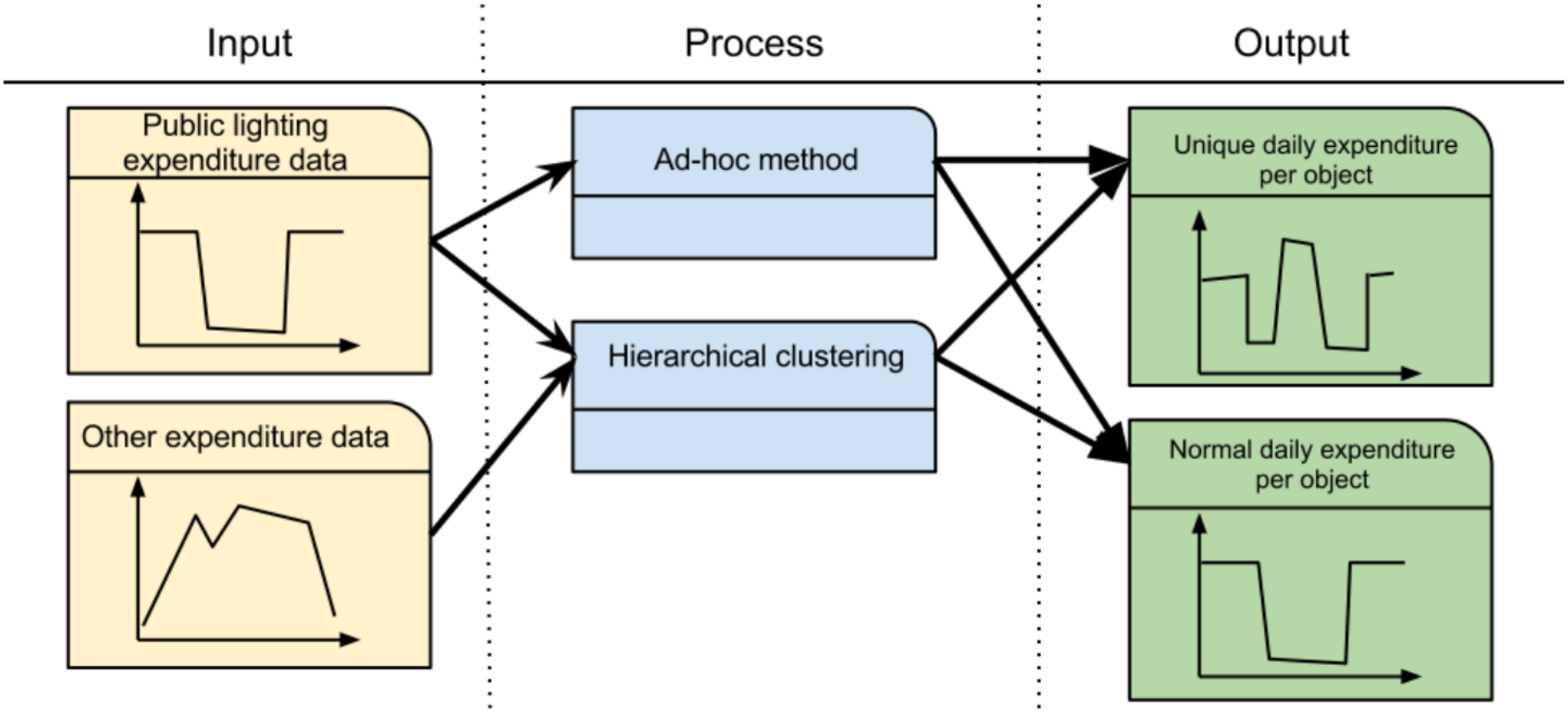
Objects

- Public lighting or traffic control
- Office
- Tunnel
- Radarpost
- Pumping station
- Floodgate or weir
- Traffic control center
- Bridge or dam
- Small building

Goal

- Clustering on all data:
  - Public lighting
  - All other objects
- Clustering to detect object groups
- Identify regular energy usage pattern of objects
- Objects are of different types
- **Detect anomalies in energy usage per object type**
- Data-driven!

Approach


Input
Process
Output
Public lighting expenditure data
Other expenditure data
Ad-hoc method
Hierarchical clustering
Unique daily expenditure per object
Normal daily expenditure per object

A clustering result

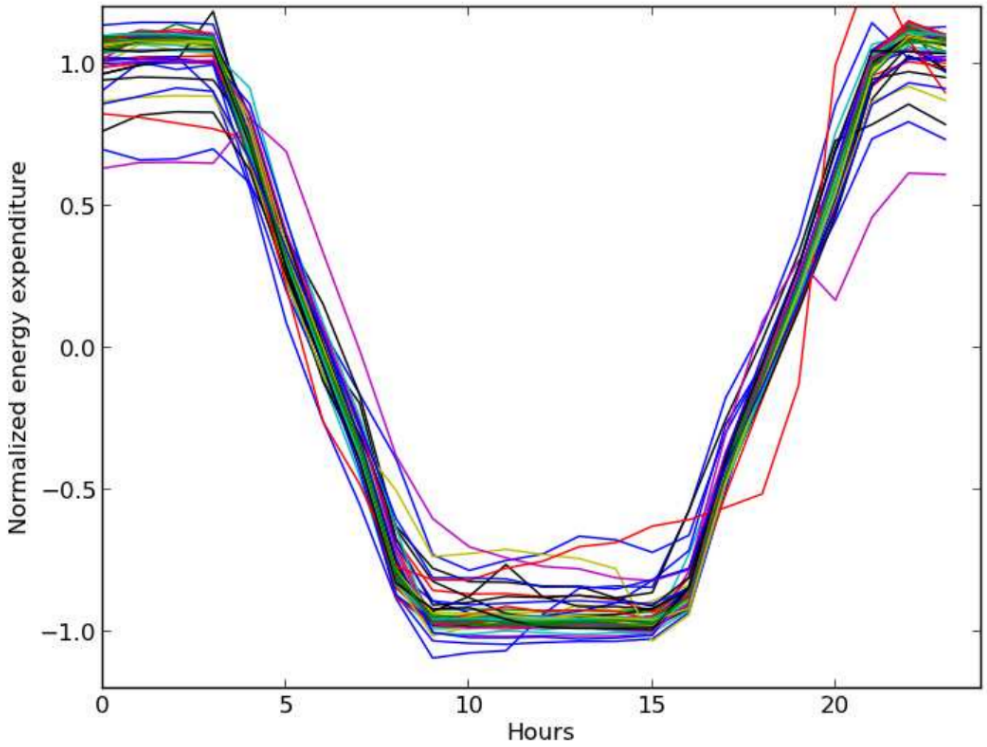


Figure : Public lighting objects

Anomaly detection results

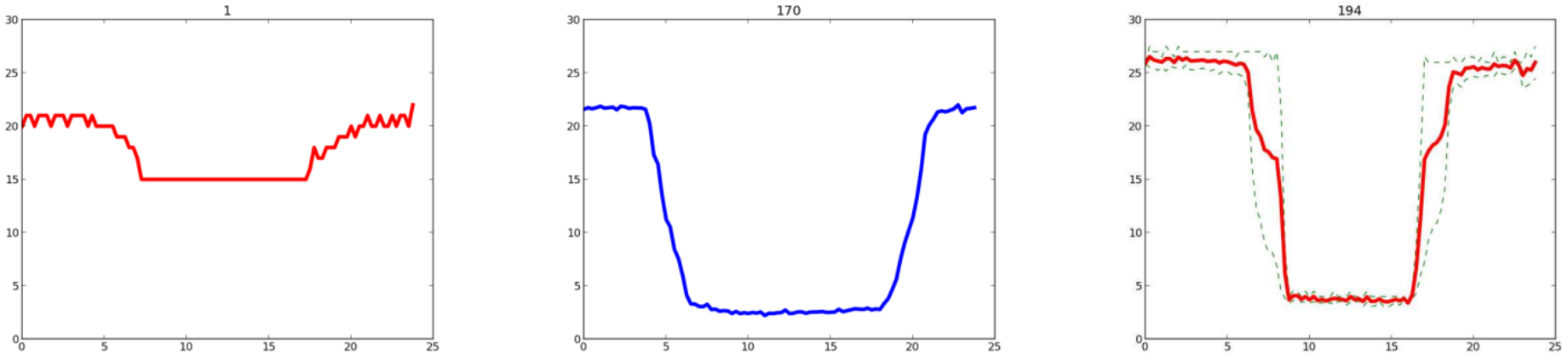


Figure : Outlier in seasonal behavior

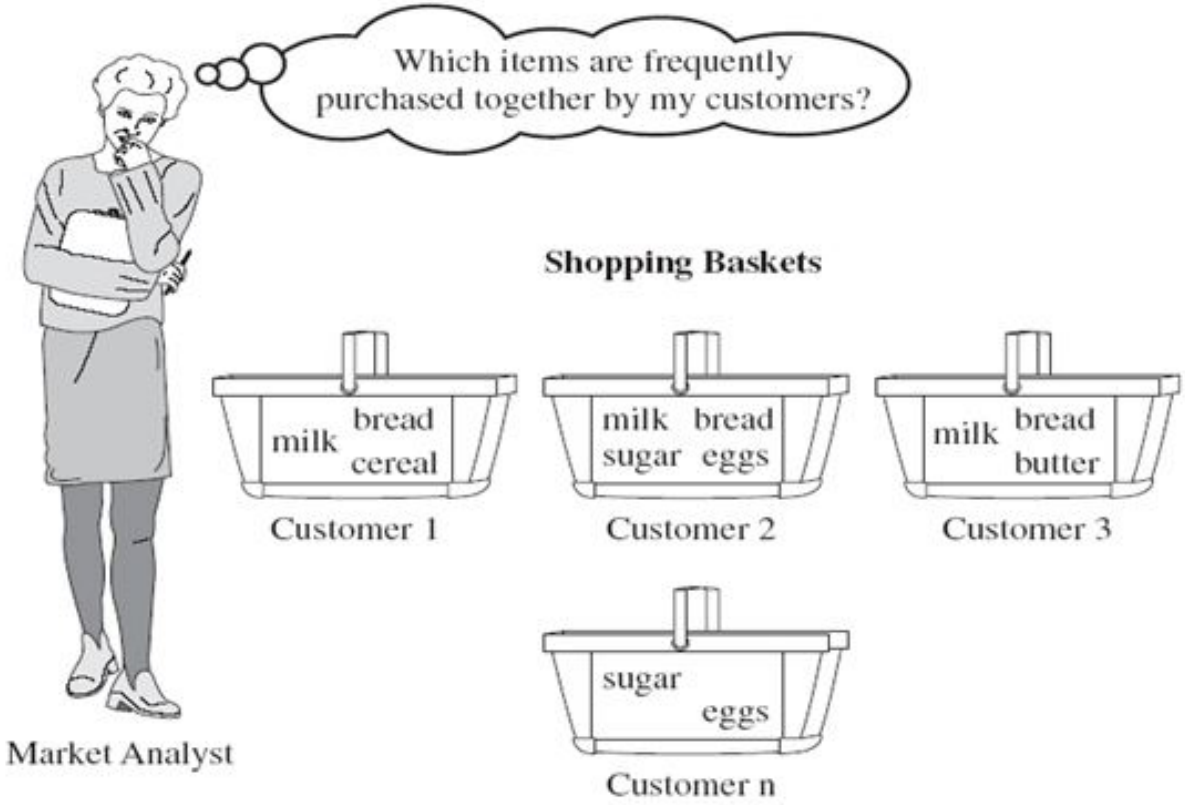
Project results and conclusion

- Objects clustered into types based on the data
- Some anomalies detected for various types of objects
- Correlations between weather and object (types) identified
- Data-driven insight!

Unsupervised learning

- Clustering
- Anomaly detection
- Pattern recognition
- **Data summarization**

Market basket analysis



Han & Kamber, *Data mining: Concepts and techniques*, 2006

Association

- X and Y are variables. There are N instances, of which N_X instances have variable X
- Derive rules of the form `IF(X) THEN Y` $X \Rightarrow Y$

$$support(X \Rightarrow Y) = N_{X \wedge Y}/N$$

$$confidence(X \Rightarrow Y) = N_{X \wedge Y}/N_X$$

$$lift(X \Rightarrow Y) = \frac{N_{X \wedge Y} N}{N_X N_Y}$$

Association

- X and Y are variables. There are N instances, of which N_X instances have variable X
- Derive rules of the form `IF(X) THEN Y` $X \Rightarrow Y$

$$support(X \Rightarrow Y) = N_{X \wedge Y}/N$$

$$confidence(X \Rightarrow Y) = N_{X \wedge Y}/N_X$$

$$lift(X \Rightarrow Y) = \frac{N_{X \wedge Y} N}{N_X N_Y}$$

- *support*:

Association

- X and Y are variables. There are N instances, of which N_X instances have variable X
- Derive rules of the form `IF(X) THEN Y` $X \Rightarrow Y$

$$support(X \Rightarrow Y) = N_{X \wedge Y}/N$$

$$confidence(X \Rightarrow Y) = N_{X \wedge Y}/N_X$$

$$lift(X \Rightarrow Y) = \frac{N_{X \wedge Y} N}{N_X N_Y}$$

- *support*: higher is better
- *confidence*:

Association

- X and Y are variables. There are N instances, of which N_X instances have variable X
- Derive rules of the form `IF(X) THEN Y` $X \Rightarrow Y$

$$support(X \Rightarrow Y) = N_{X \wedge Y}/N$$

$$confidence(X \Rightarrow Y) = N_{X \wedge Y}/N_X$$

$$lift(X \Rightarrow Y) = \frac{N_{X \wedge Y} N}{N_X N_Y}$$

- *support*: higher is better
- *confidence*: close to 1
- *lift*:

Association

- X and Y are variables. There are N instances, of which N_X instances have variable X
- Derive rules of the form `IF(X) THEN Y` $X \Rightarrow Y$

$$support(X \Rightarrow Y) = N_{X \wedge Y}/N$$

$$confidence(X \Rightarrow Y) = N_{X \wedge Y}/N_X$$

$$lift(X \Rightarrow Y) = \frac{N_{X \wedge Y} N}{N_X N_Y}$$

- *support*: higher is better
- *confidence*: close to 1
- *lift*: factors higher is better

Association rules

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

http://www.saedsayad.com

Unsupervised learning

- Clustering
- **Anomaly detection**
- Pattern recognition
- Data summarization

Anomaly detection

- **Supervised**: normal/outlier can be learned as a class attribute
- **Semi-supervised**: train on a labeled dataset, determine outliers in unlabeled data based on likelihood of a deviation
- **Unsupervised**: identify patterns (for example, using clustering) and then select small clusters or instances that do not logically fall in any of the large clusters

Assignment 1

- Gaming industry context
- Sales log spanning 4 years of sales
- Apply and compare BI techniques
- Inspect, visualize, aggregate, segment, score ...
- Deliverables:
  1. Web-based BI Dashboard
  2. Short assignment report in LaTeX

Assignment 1 — Hints

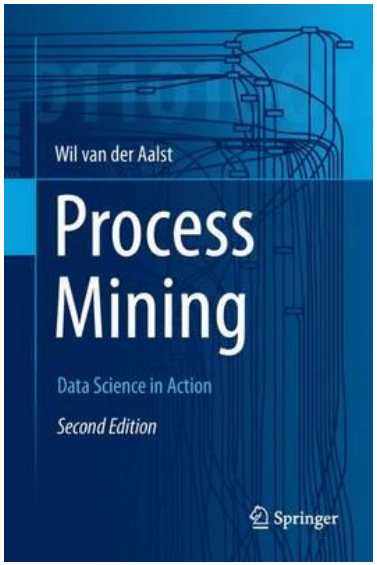
- Model: MySQL database containing the data
- View: HTML page using Javascript that reads JSON
- Controller: PHP outputs relevant data in JSON

Lab session February 23

- Make serious progress with Assignment 1
- Continue with dashboard and data integration
- Error reporting in PHP and other handy tricks: `http://liacs.leidenuniv.nl/ict`
- Start thinking about the BI questions
- Ask all relevant questions

Credits

Lecture partially based on (slides of the (previous edition of the)) course book: W. van der Aalst, *Process Mining: Data Science in Action*, 2nd edition, Springer, 2016.



Slides partially based on "From Data Mining to Knowledge Discovery: An Introduction" by Gregory Piatetsky-Shapiro (`KDnuggets.com`)