Lecture Notes

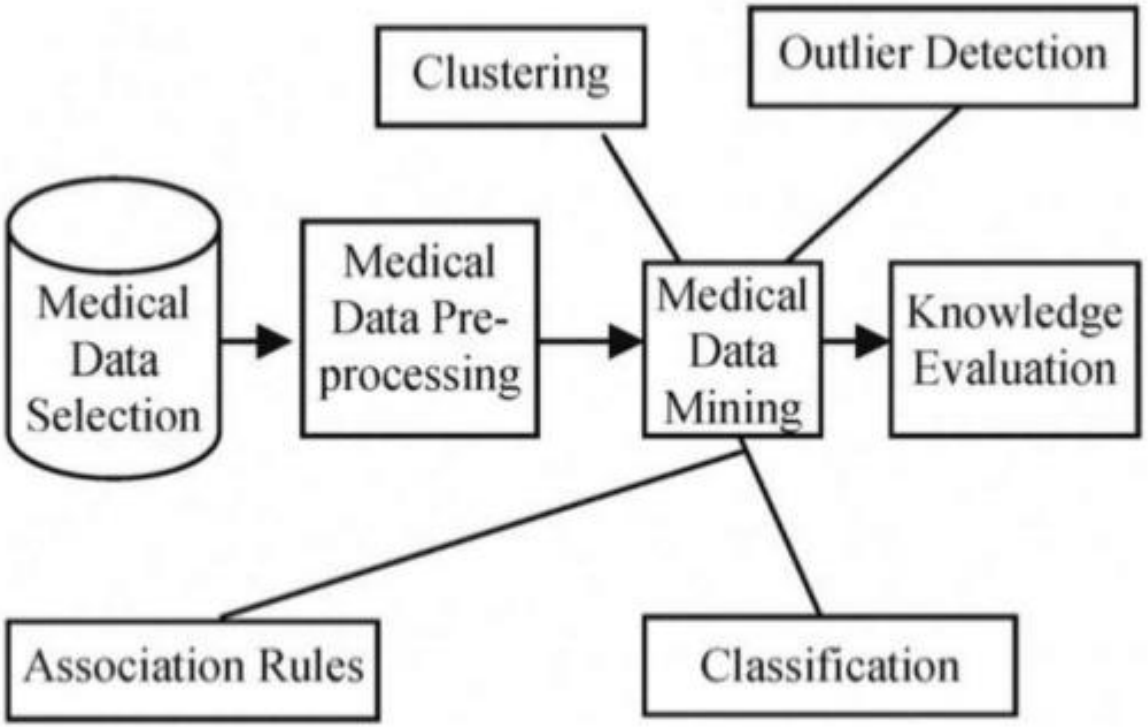
Big Data in Medical Informatics

Week 10:

Data Analytics in Medicine – Part 1

RWTH AACHEN UNIVERSITY

Data Analytics in Medicine



A Critical Examination of Data Mining Classification Techniques in Medical Image Databases

Stages of Healthcare Data Mining

1. Data Acquisition: Collecting data from various sources.
2. Data Preprocessing: operations applied to the data to prepare it for further analysis:
   - data cleaning to filter out noisy data elements and to cope with missing values,
   - data normalization to cope with heterogeneous sources, temporal alignment, and data formatting.
3. Data Transformation: operations for representing the data appropriately and selecting specific features
   - feature extraction and selection

Stages of Healthcare Data Mining

4. Modeling: Applying knowledge discovery algorithms to identify patterns in the data.
   - anomaly detection to identify statistically deviant data,
   - association rules to find dependencies and correlations in the data,
   - clustering models to group data elements according to various notions of similarity,
   - classification models to group data elements into predefined classes,
   - regression models to fit mathematical functions to data,
   - summarization models to summarize or compress data into interesting pieces of information.
5. Evaluation: operations for evaluation and interpretation of the results of the modeling process.

Application of Data Mining In Medicine

- What is the fundamental question in medicine ?

Application of Data Mining In Medicine

- What is the fundamental question in medicine ?
  - Decide on the intervention/treatment that is most appropriate and effective for a particular patient

Application of Data Mining In Medicine

- What is the fundamental question in medicine ?
  - Decide on the intervention/treatment that is most appropriate and effective for a particular patient
  - How to find an answer ?
    - By asking more questions

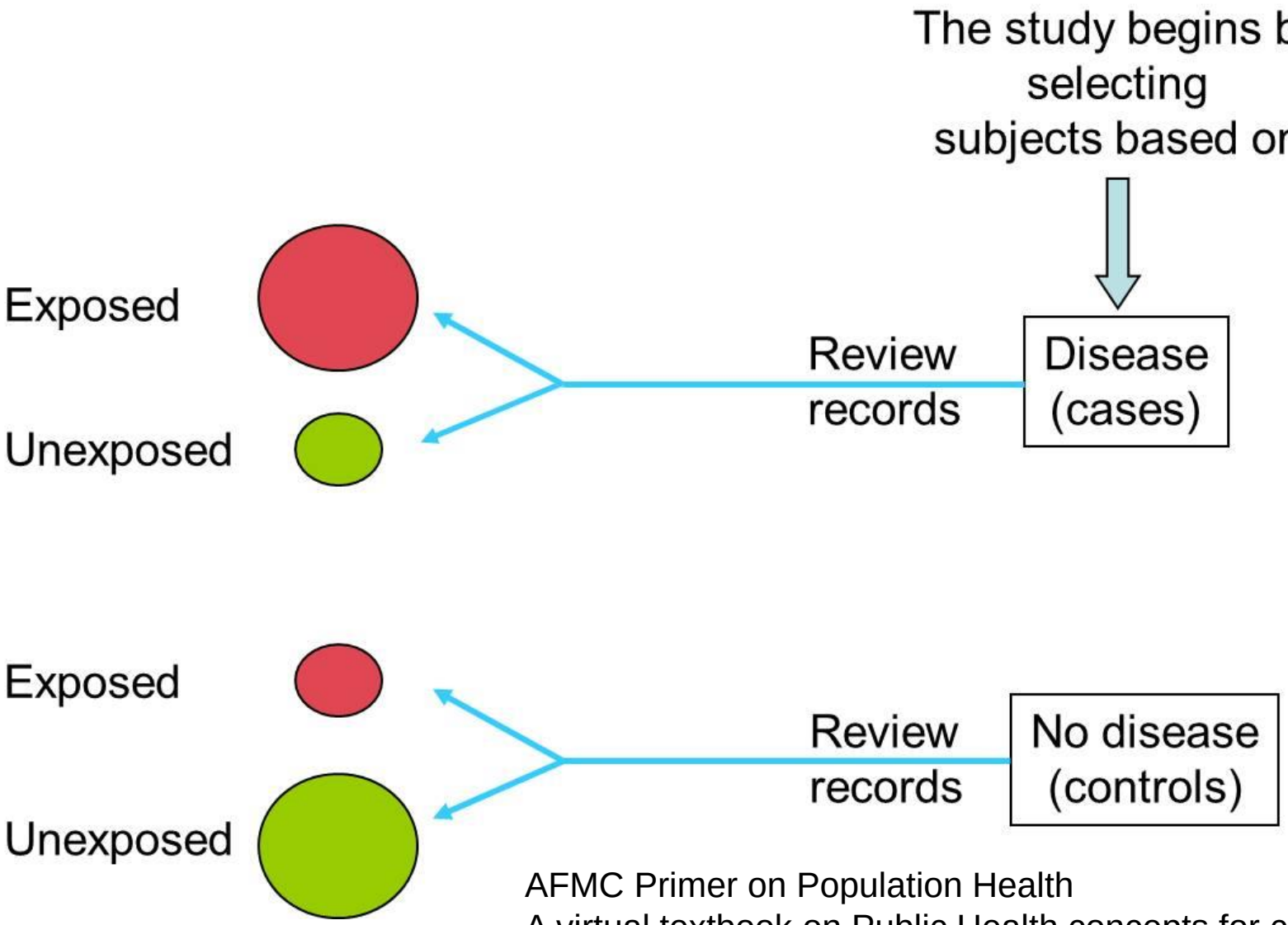
Application of Data Mining In Medicine

- Start with the most basic epidemiological questions:
  - How many patients at a particular time are affected by this condition?
  - How many new cases do we discover each year?
  - What are the symptoms of the disease?
  - What is the natural history of the disease,

    i.e. what are the precursors and consequences of this condition?

Application of Data Mining In Medicine

- Then ask more complex questions ans test them
  - assemble a cohort (group) of patients,
    - some of whom are extremely likely to have the diseases (cases) and
    - other who most likely do not (controls).
  - What is an intervention?
    - often drug therapies or surgeries,
    - but can also include recommendations for life style changes and/or patient education.



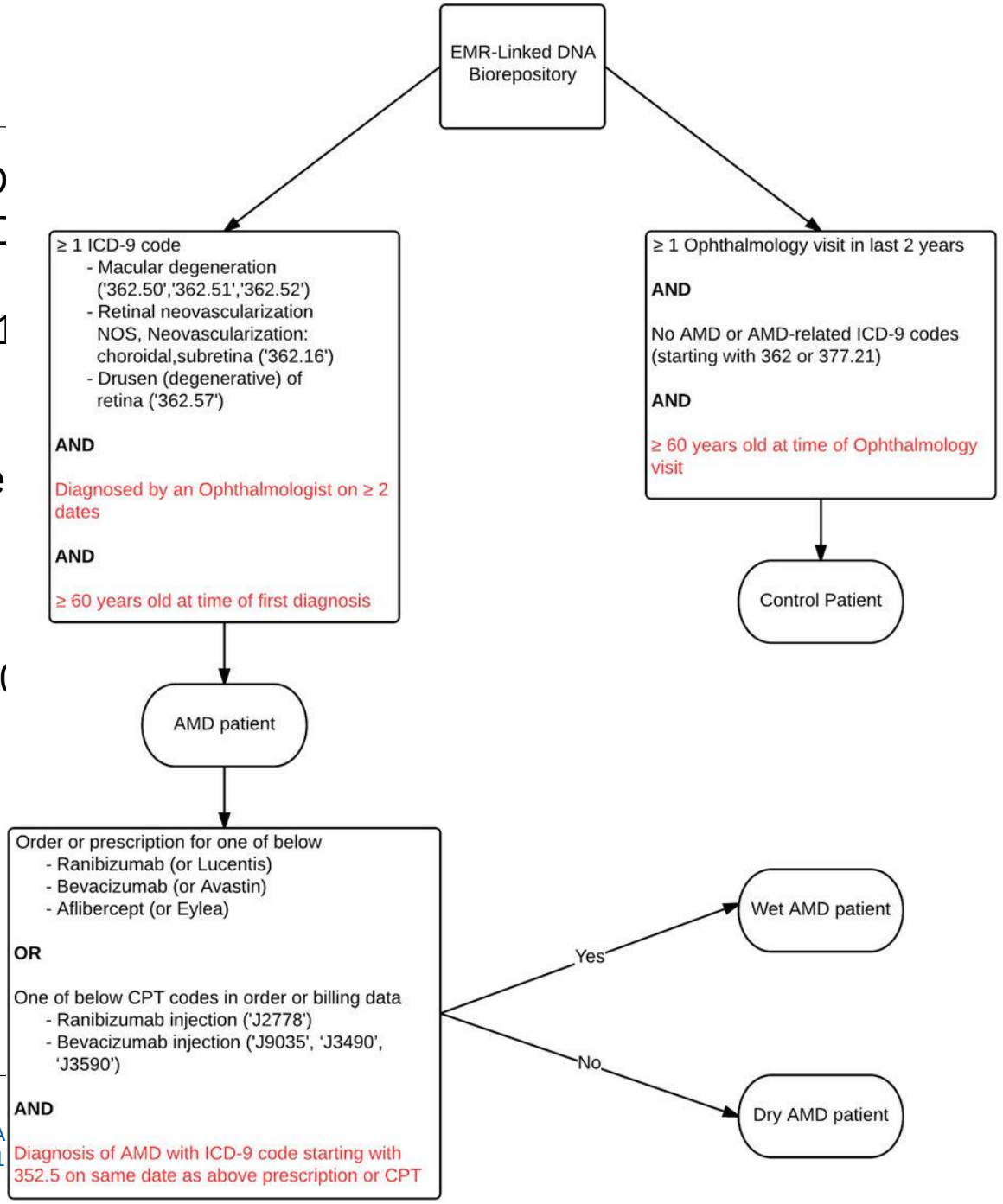
AFMC Primer on Population Health
A virtual textbook on Public Health concepts for clinicians.

Application of Data Mining In Medicine

- First problem: cohort identification
- How to define case and controls?
  - Phenotyping algorithms
    - Used to characterize the disease in terms of patient characteristics observable from the EHR data
    - either hand-crafted or machine learned
  - high-throughput clinical phenotyping (HTCP) algorithms that apply specific inclusion and exclusion criteria to clinical data, available through the EMR, could generate a large cohort of potentially eligible study subjects

Phenotyping Algorithms

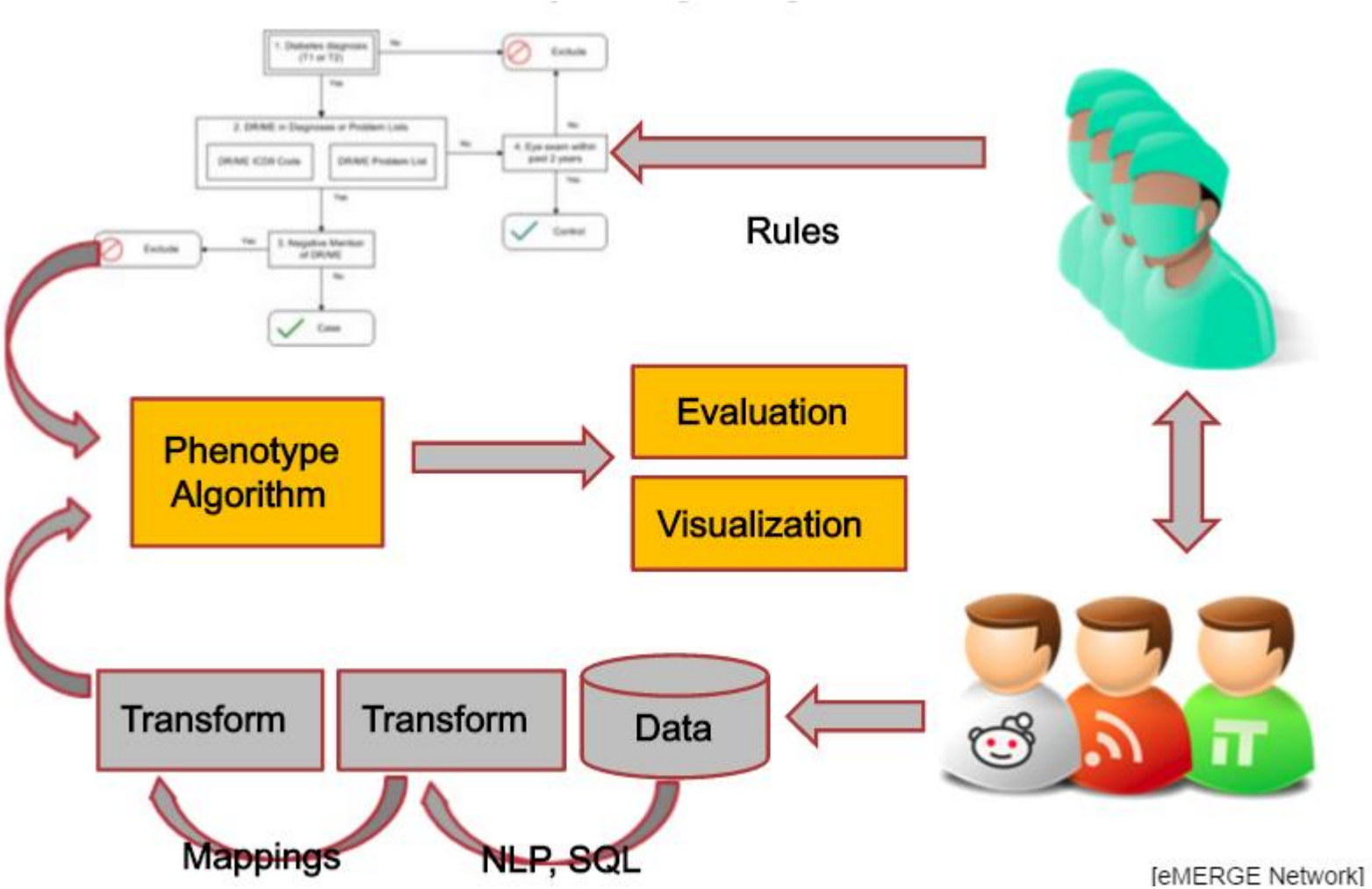
the algorithm to identify all AMD cases using the criterion of AMD ICD-9 codes entered by an ophthalmologist (362.50, 362.51 362.52, 362.16, 362.57). To classify cases as “wet” AMD cases within this population, we additionally required a current procedural terminology (CPT) code (J2778: ranibizumab injection, J9035, J3490 or J359(bevacizumab injection), or an order or prescription for ranibizumab , bevacizumab, or aflibercept.



Simonett, Joseph M., et al. "A validated phenotyping algorithm for genetic association studies in age-related macular degeneration." *Scientific reports* 5 (2015).

RWTH Informatik 5 | Ahornstr. 55 D-52056 A
Tel +49/241/8021501 | Fax +49/241/8022321

EHR Driven Phenotyping Algorithms



Strategic Health IT Advanced Research Projects (SHARP) Area 4: Secondary Use of EHR Data Project 3: High-Throughput Phenotyping Jyotishman Pathak, PhD.

Application of Data Mining In Medicine

- Once the cohort is defined, data can be collect, process and analyzed
  - relevant known or potential predictors can be collected and
  - predictive models can be build, such as
    - predict the risk of disease, e.g. probability of developing the condition in 5 years (risk prediction)
    - investigate which predictors are relevant (biomarker discovery, risk factor discovery) in developing the outcome.

- Dissemination of knowledge: clinical practice guidelines.

Application Areas of Data Mining in Medicine

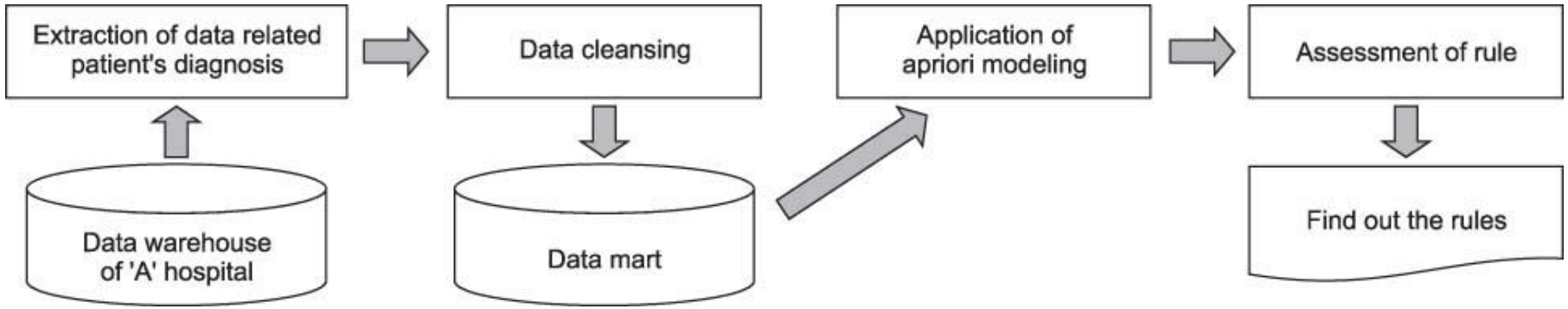
- Area: Understanding the Natural History of Disease
  - questions:
    - Is the condition or disease serious?
    - Are large numbers of patients involved?
    - What are the societal implications of the disease?
    - Has the disease been studied before?
  - Possible data analytic focuses:
    - The prevalence of the disease,
    - comorbidity analysis,
    - the incidence of the disease (patient medical trajectories).

Application Areas of Data Mining in Medicine

- Understanding the Natural History of Disease : Comorbidity analysis
  - the process of exploring and analyzing relationships between frequently co-occurring diseases.
  - Such as , patients suffering from type 2 diabetes mellitus (T2DM) often also suffer from hypertension, hyperlipidemia and impaired fasting glucose (IFG).
  - Some diseases occur in clusters and it is desirable to treat them simultaneously

Application Areas of Data Mining in Medicine

- Using the association rule mining framework, Shin et al. explored the comorbidities associated with hypertension such as non-insulin dependent diabetes mellitus, cerebral infection and chronic renal failure



A Mi Shin, In Hee Lee, Gyeong Ho Lee, Hee Joon Park, Hyung Seop Park, Kyung Il Yoon, Jung Jeung Lee, and Yoon Nyun Kim. Diagnostic analysis of patients with essential hypertension using association rule mining. Healthcare informatics research, 16(2):77{81, June 2010.

Application Areas of Data Mining in Medicine

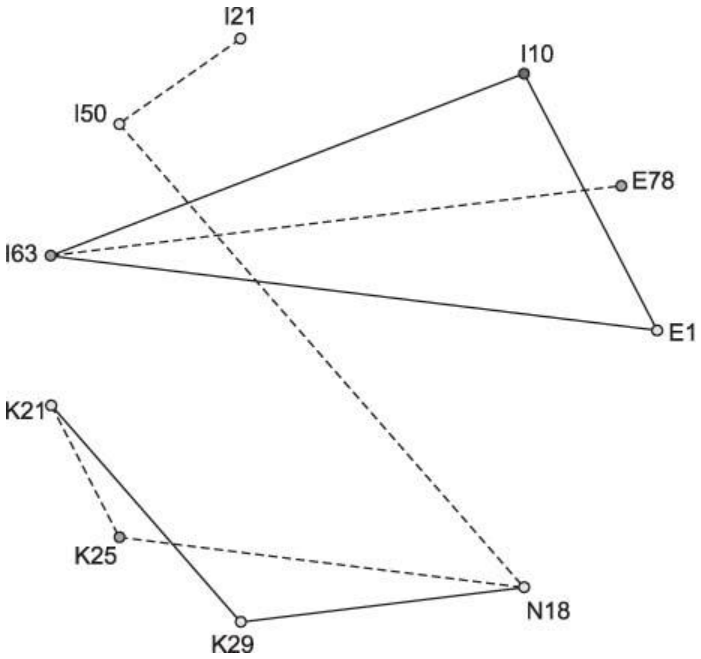
Antecedent	Consequent	Support (%)	Confidence (%)	Lift
E11	I10	35.15	100.00	1.000
I10	E11	35.15	35.15	1.000
I63	I10	21.21	100.00	1.000
I10	I63	21.21	21.21	1.000
I10, I63	E11	7.91	37.31	1.062
I10, E11	I63	7.91	22.49	1.061
I10, E11	I20	5.54	15.75	1.079
I10, E11	N18	5.52	15.69	1.545

8 association rules are extracted and the used threshold values were as follows: support, ≥5%; and confidence, ≥15%.

$$\text{Support (\%)} = \frac{\text{Number of disease } A \cap B}{\text{Total number of disease}}$$

$$\text{Confidence (\%)} = \frac{\text{Number of disease } A \cap B}{\text{Number of disease } A}$$

$$\text{Lift} = \frac{\text{Number of disease } A \cap B \times \text{Total number of disease}}{\text{Number of disease } A \times \text{Number of disease } B}$$



Association graph by using web node. E11: non-insulin-dependent diabetes mellitus, E87: other disorders of fluid, electrolytes and acid-base balance, I10: essential hypertension, I21: acute myocardial infarction, I50: heart failure, I63: cerebral infarction, K21: gastrooesophageal reflux disease, N18: chronic renal failure.

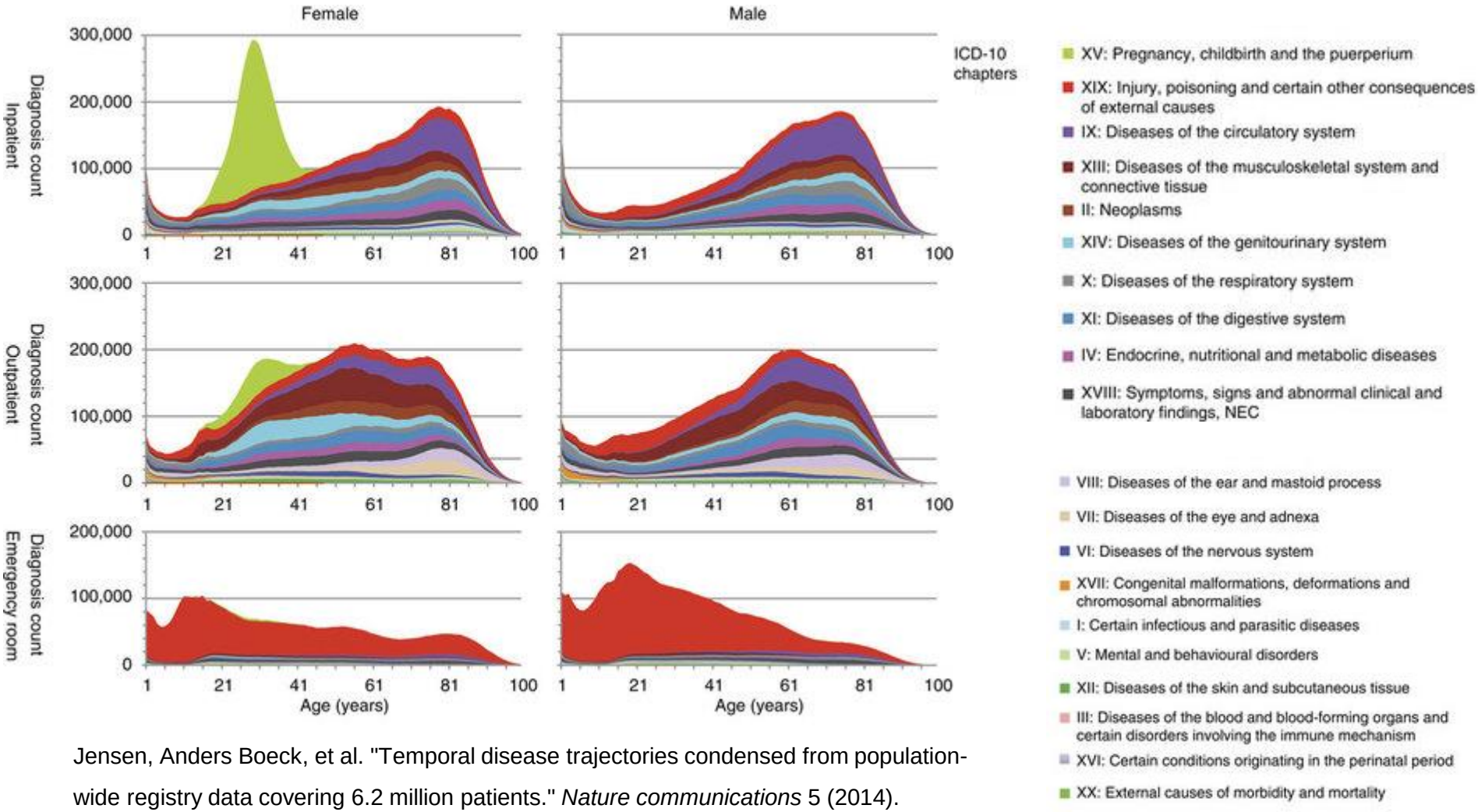
Application Areas of Data Mining in Medicine

- Understanding the Natural History of Disease : the patient's medical trajectory prediction
  - medical state of a patient can be represented using laboratories test results, diagnosis codes or medication information
  - progression of a patient's medical state over time is known as the patient's medical trajectory
  - Examples: the progression of the patient from a healthy state through conditions like hypertension, hyperlipidemia, impaired fasting glucose (IFG), type 2 diabetes mellitus and eventually towards diabetes complications (e.g. amputation, severe paralysis or death)

Application Areas of Data Mining in Medicine

- Jensen et al. have explored temporal disease progression patterns in data from an electronic health record registry which covers the entire population of Denmark.
- Findings demonstrate how these trajectories have predictive potential and might be the basis for predicting the next probable step in disease progression.

Jensen, Anders Boeck, et al. "Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients." *Nature communications* 5 (2014).

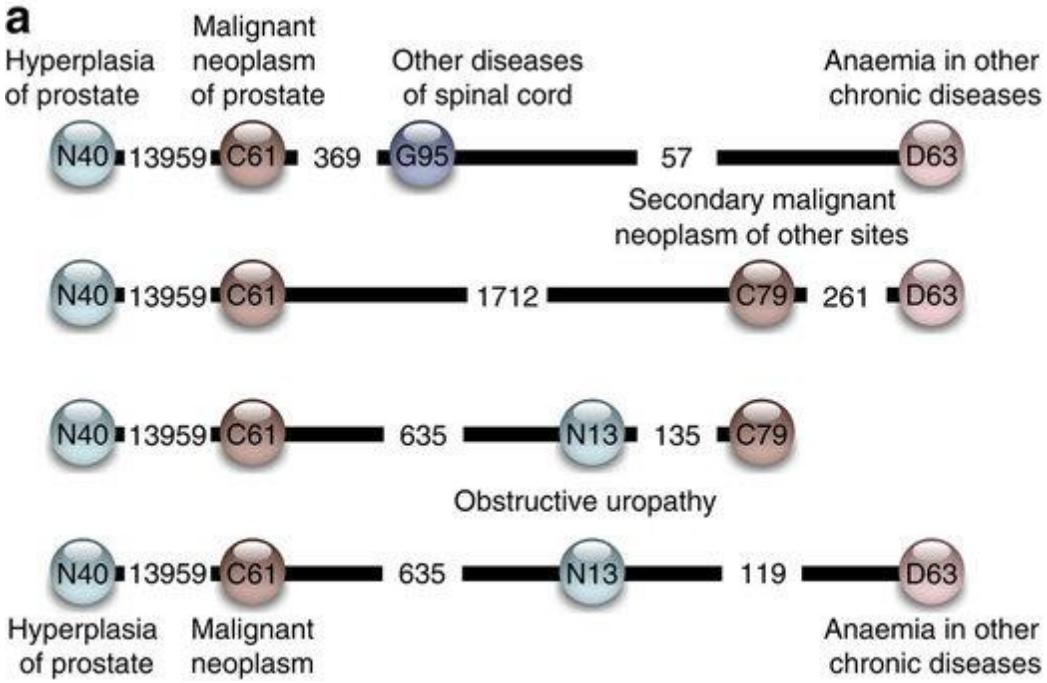


Jensen, Anders Boeck, et al. "Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients." *Nature communications* 5 (2014).

Application Areas of Data Mining in Medicine

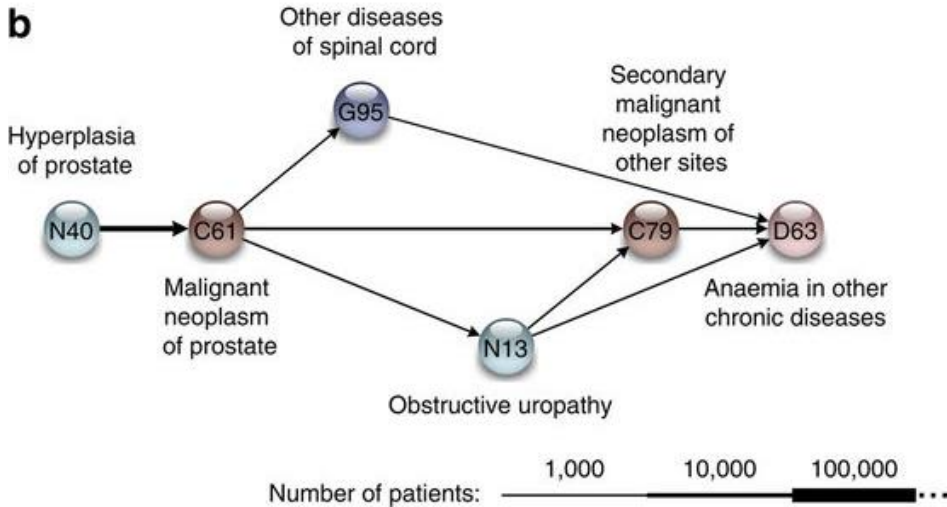
- Temporal co-morbidity analysis as basis for trajectories:
  - Searched temporal correlations among pairs of diagnoses
  - performed a cohort study where exposed patients who had a specific pair of diagnoses were matched with comparison patients with same age, gender and type of hospital encounter.
  - Using this cohort, they identified 1171 significant trajectories
- Clustering trajectories reveals disease development patterns
  - clustered the trajectories based on which diagnoses they shared.
  - These trajectories were then clustered using key diagnosis codes such as chronic obstructive pulmonary disease (COPD) and gout.
  - As a similarity measure between diagnosis pairs: the Jaccard Index.

Application Areas of Data Mining in Medicine



The figure illustrates the transition from trajectories to a trajectory cluster. Each circle represents a diagnosis and is labelled with the corresponding ICD-10 code. The colours represent different ICD-10 chapters. The temporal diagnosis progression goes from left to right.

(**a**) All trajectories that contribute to the prostate-cancer cluster. The number of patients, who follow the trajectory until a given diagnosis, is given in the edges. (**b**) The prostate cancer trajectory cluster that represents all the trajectories. The width of the edges corresponds to the number of patients with the directed diagnosis pair from the full population. The cluster describes a normal progression from having hyperplesia of prostate diagnosed to having prostate cancer, cancer metastasis and



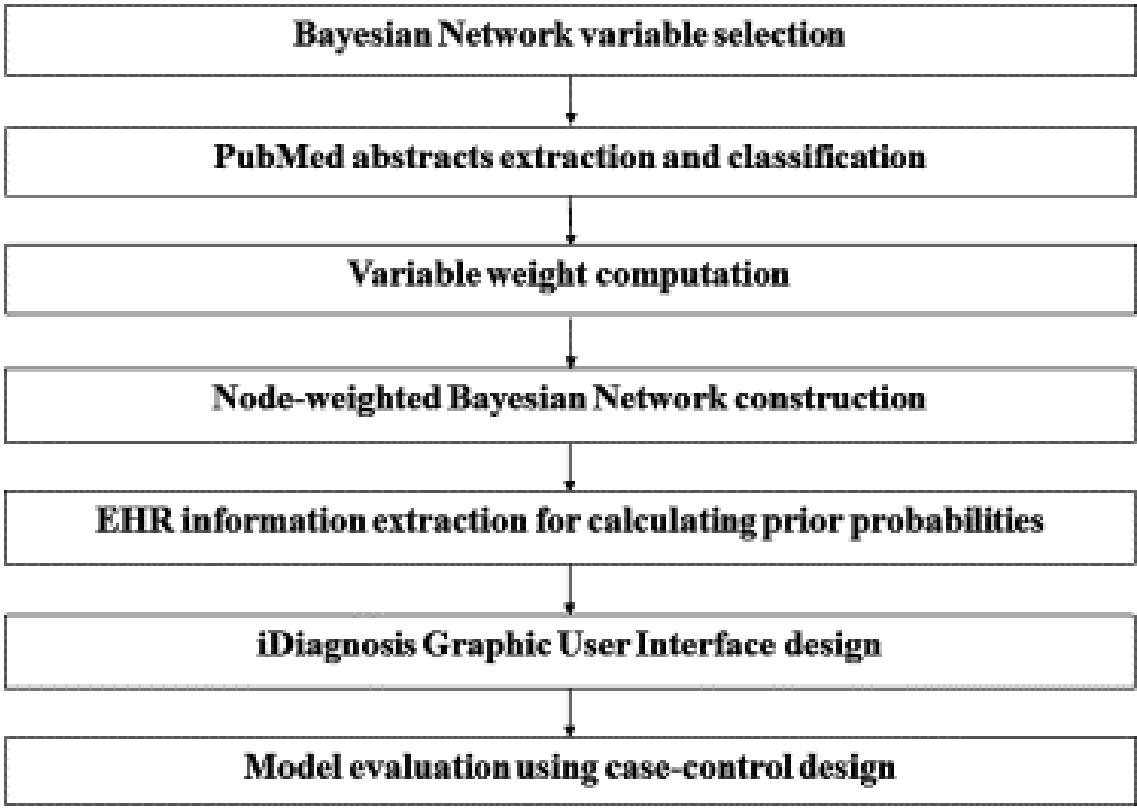
Application Areas of Data Mining in Medicine

- Risk Prediction/Biomarker Discovery
  - constructing predictive models to assess the patient's risk and progression from a patient's current medical state to a medical state associated with potentially advanced medical complications.
- Predicting the next complication
  - Future complications arising due to patient's current medical condition
- Quantifying the effect of Intervention
  - Data mining techniques such as association rule mining have been used to measure the effect of interventions.
- Adverse Event Detection
  - detecting any untoward medical occurrence caused by mismanagement of patient health. Such medical errors might arise due to accidental surgical practices, drug reactions or the use of outdated medical guidelines.

Application Areas of Data Mining in Medicine

Example : Predicting the next complication

- Zhao et al. proposed a novel method which combines PubMed knowledge and EHRs to develop a weighted Bayesian Network Inference (BNI) model for pancreatic cancer prediction.

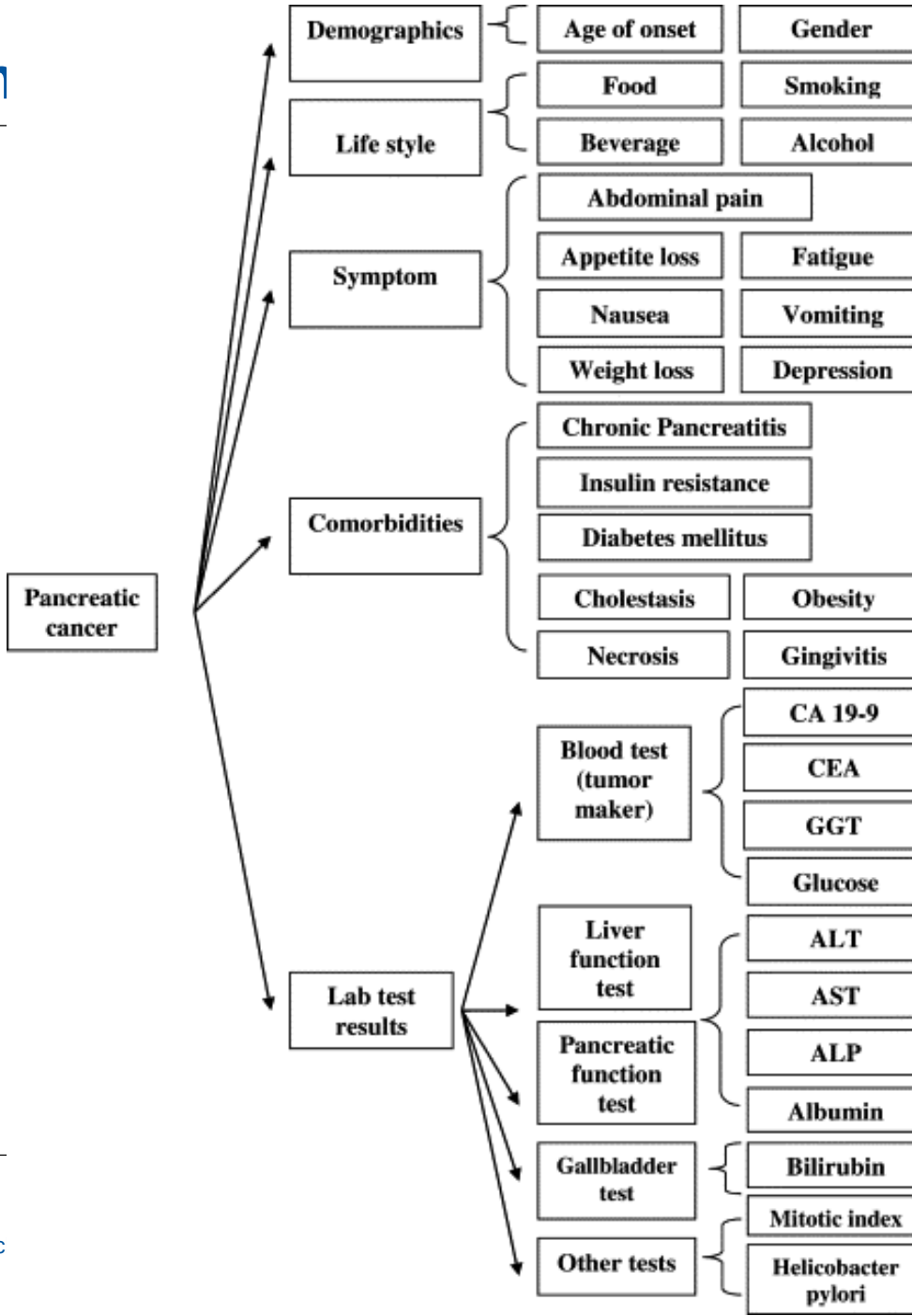


Zhao, Di, and Chunhua Weng. "Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction." *Journal of biomedical informatics* 44.5 (2011): 859-868.

Application Areas of Data Min

- They identified 31 variables associated with pancreatic cancer by aggregating the results from a PubMed review

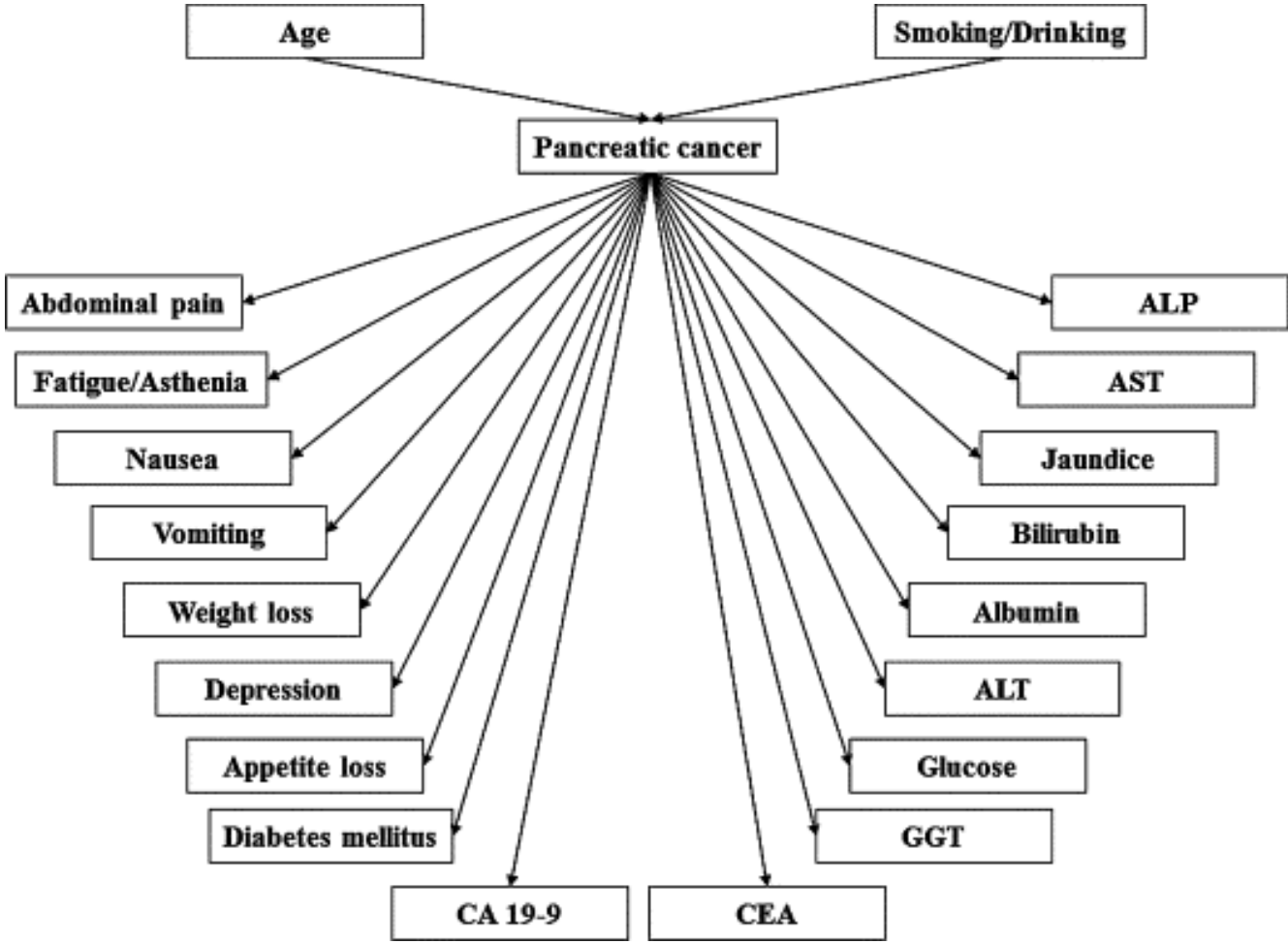
- Each risk factor is treated as a binary variable without considering the severity, degree, accumulative length, or other quantitative information of the risk factor. The value “true” represents the presence of a factor and the value “false” represents the absence of a risk factor.



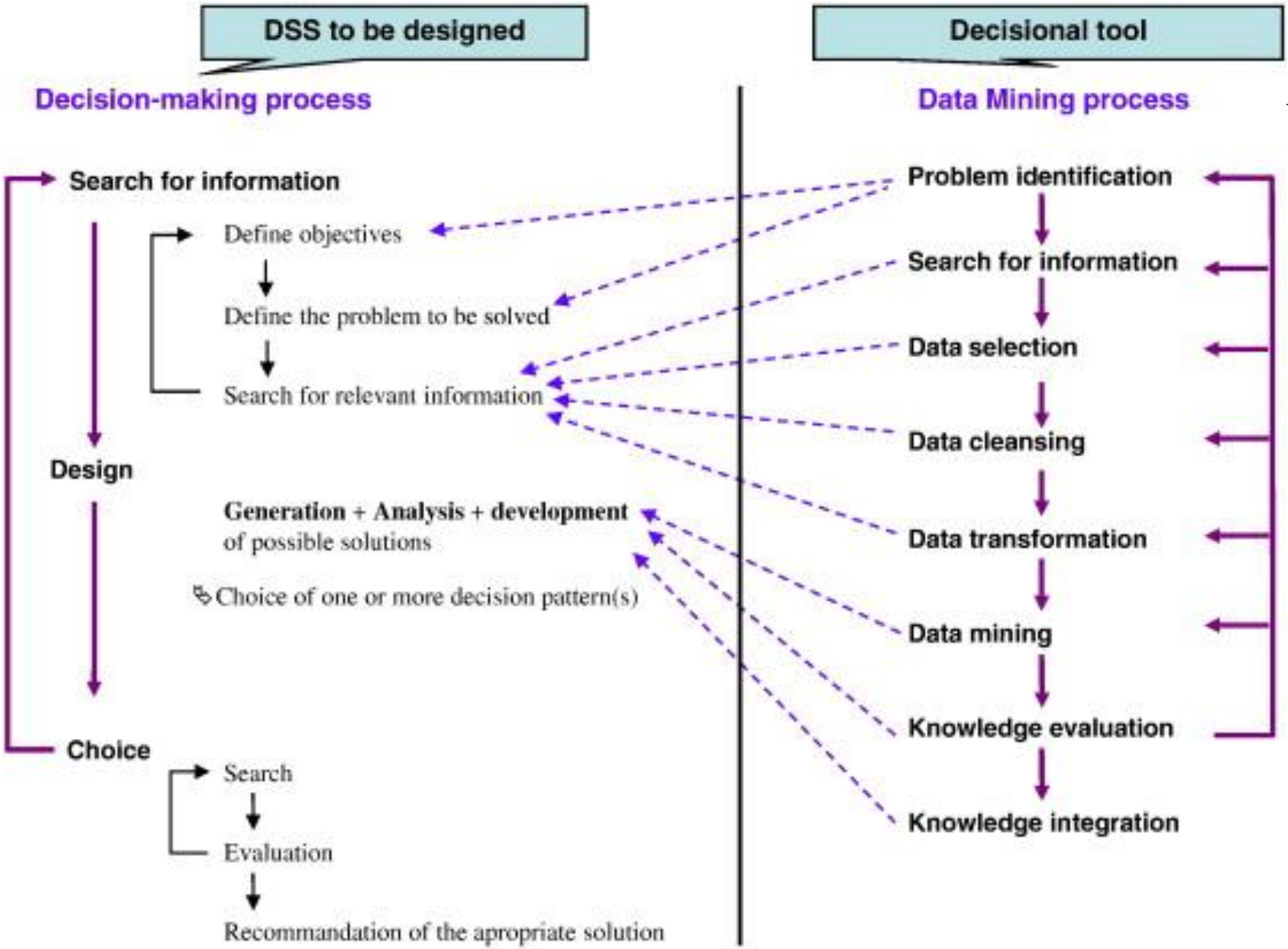
RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aac

Application Areas of Data Mining in Medicine

- the weighted BNI model significantly outperformed the conventional BNI and two other classifiers (k-Nearest Neighbor and Support Vector Machine)

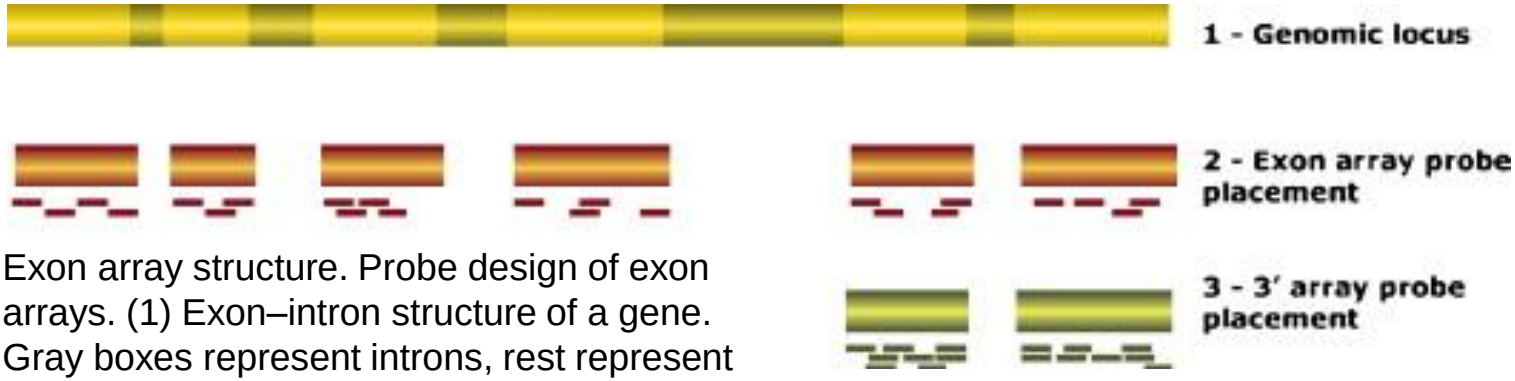


The topology of the Bayesian Network for predicting pancreatic cancer.



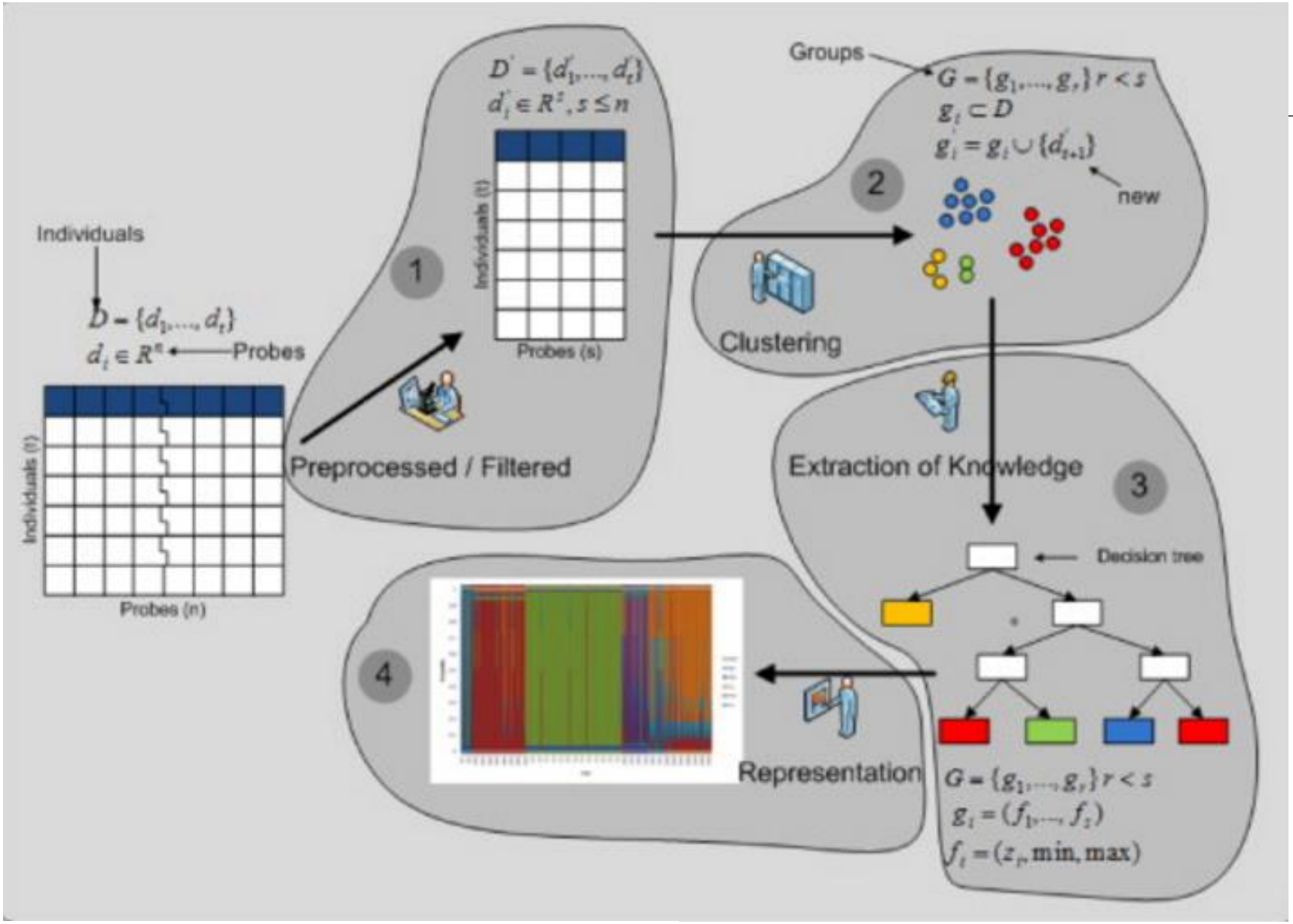
Ayed, B. M., Ltifi, H., Kolski,C. & Alimi, A. (2010) A usercentered approach for the design & implementation of KDD-based DSS: A case study in the healthcare domain. Decision Support Systems, 50, 64- 78.

- Example study: Model of experts for decision support in the diagnosis of leukemia patients -
- Leukemia is a type of blood cancer that results from an abnormal functioning of the bone marrow
- Four most important types of leukemia are: acute and chronic myeloid leukemia (AML, CML), and acute and chronic lymphocytic leukemia (ALL, CLL)
- Microarrays have been successfully tested in identifying leukemia prognoses.



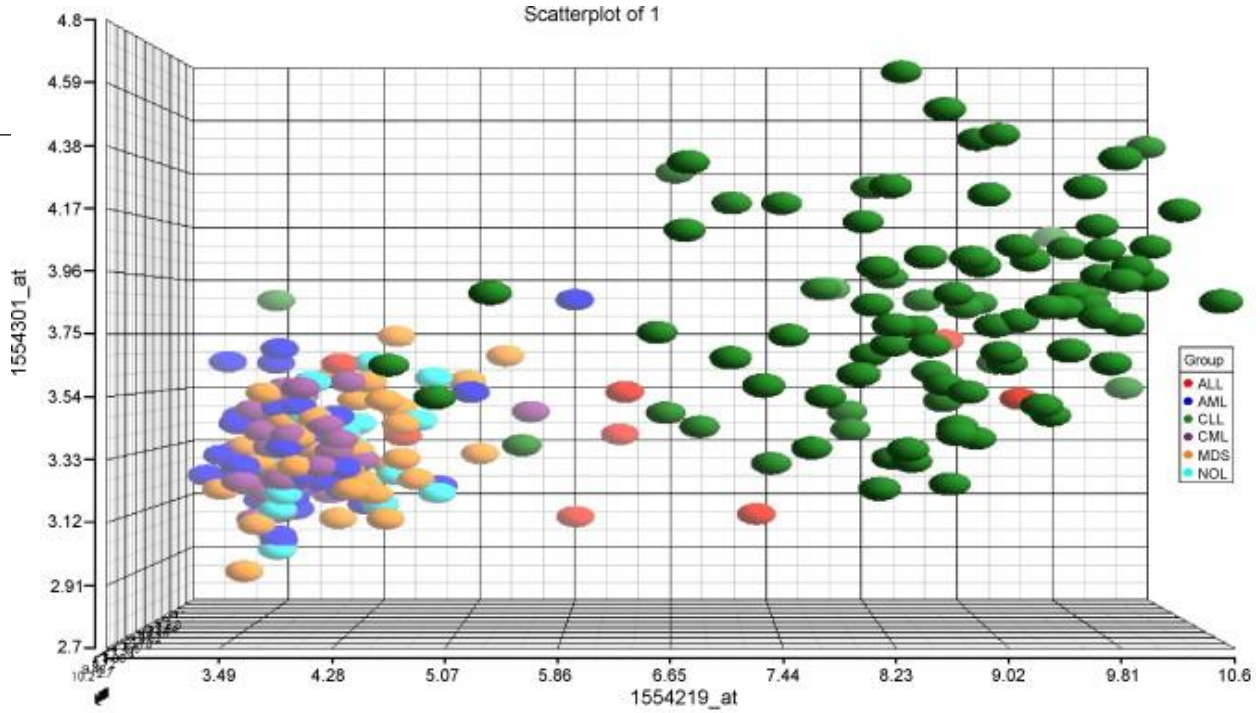
Exon array structure. Probe design of exon arrays. (1) Exon–intron structure of a gene. Gray boxes represent introns, rest represent exons. Introns are not drawn to scale. (2) Probe design of exon arrays. Four probes target each putative exon. (3) Probe design of 3′expression arrays. Probe target the 3′end of mRNA sequence..

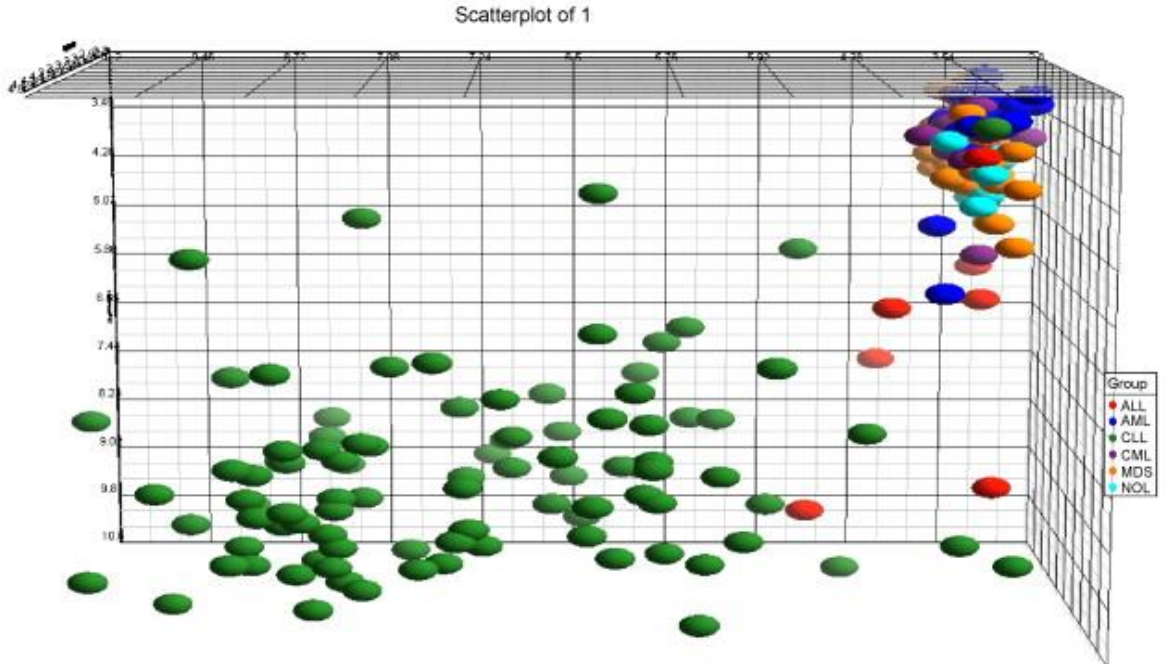
Corchado, J. M., De Paz, J. F., Rodriguez, S. & Bajo, J. (2009) Model of experts for decision support in the diagnosis of leukemia patients. Artificial Intelligence in Medicine, 46, 3, 179-200



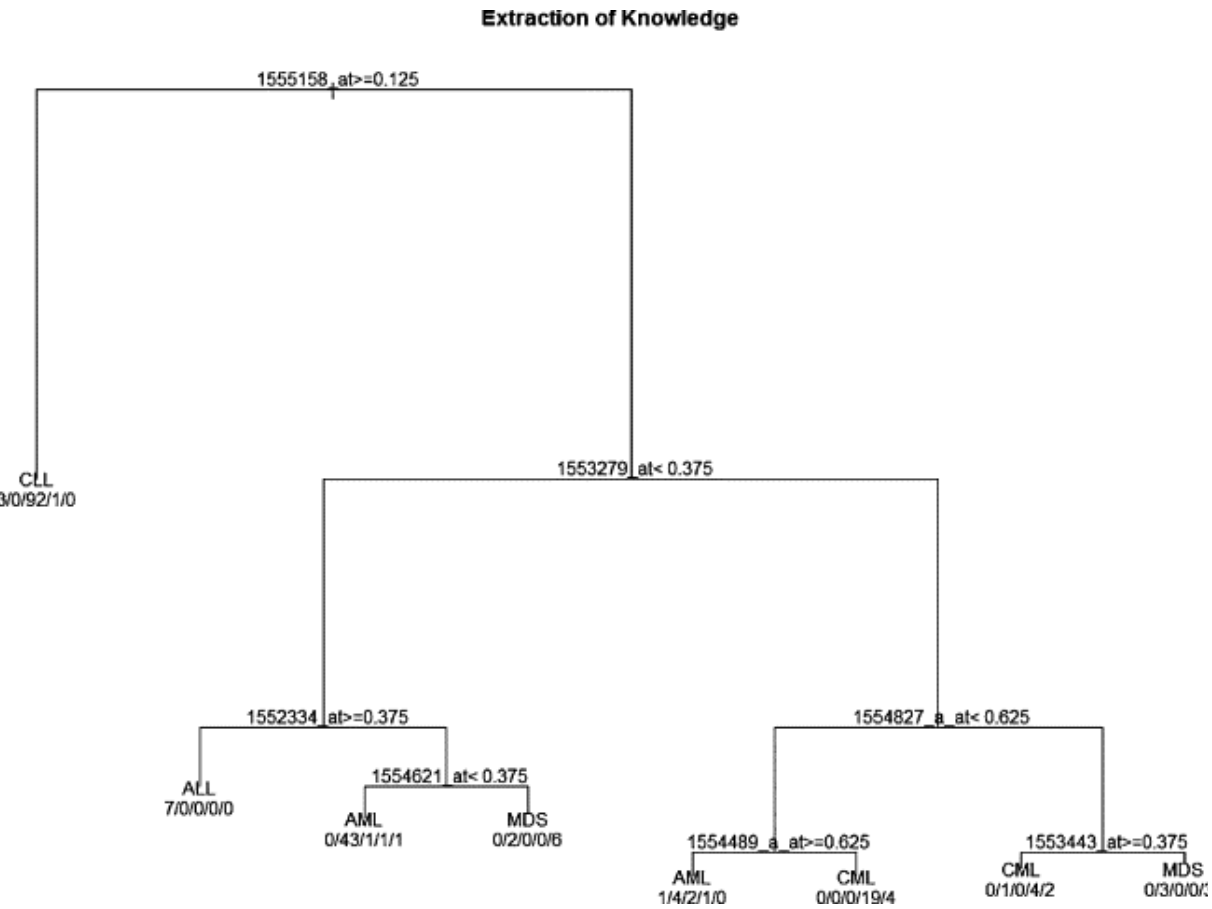
Corchado, J. M., De Paz, J. F., Rodriguez, S. & Bajo, J. (2009) Model of experts for decision support in the diagnosis of leukemia patients. Artificial Intelligence in Medicine, 46, 3, 179-200

- After the pre-processing and filtering the next step in the analysis process is to perform the clustering of individuals based on their proximity according to their probes.
- Classification CLL from the most important probes extracted by the CART algorithm. Each of the axes represents one of these probes extracted by CART for the classification of the CLL group.





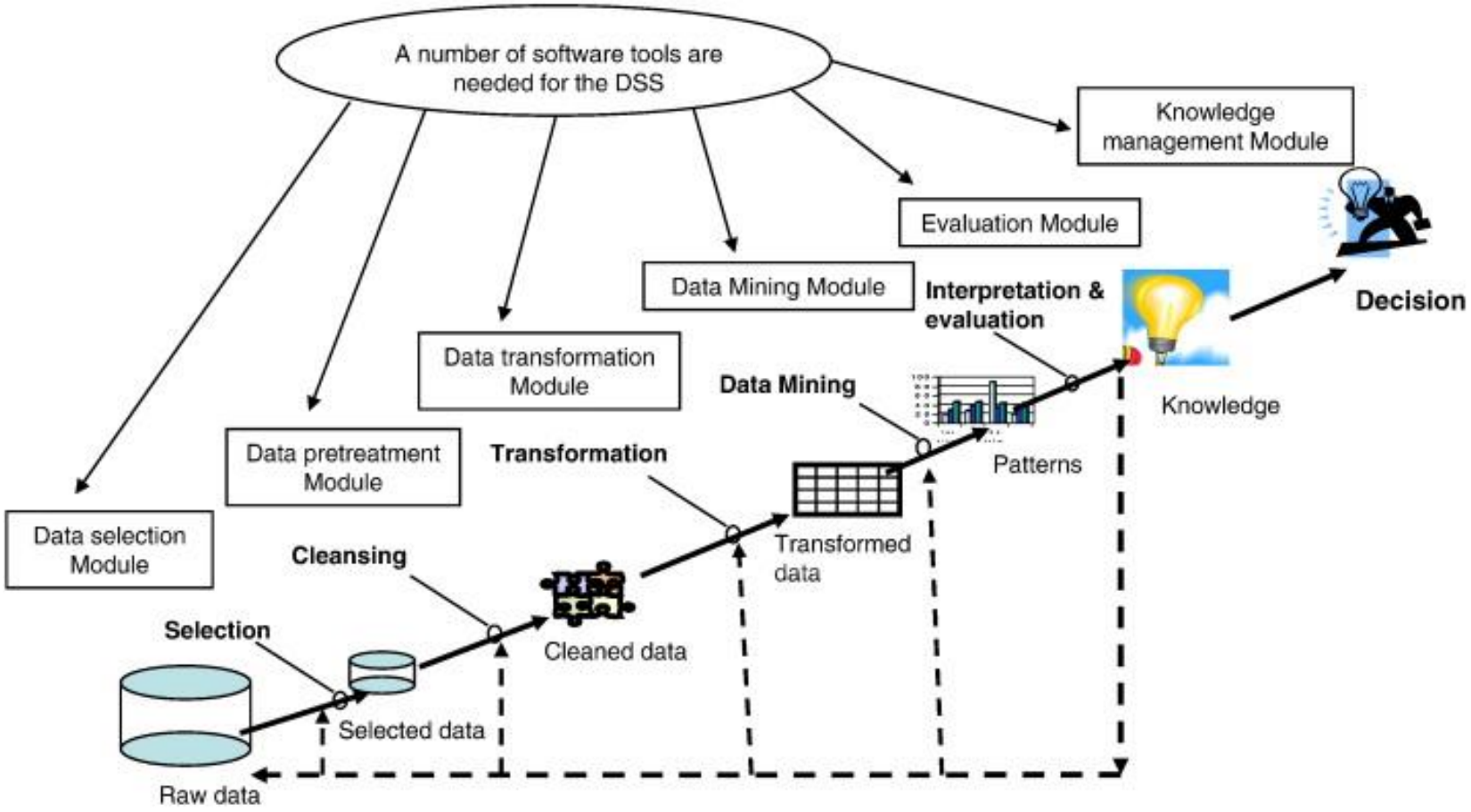
RWTH Informatik 5 | Ahornstr. 55 D-
Tel +49/241/8021501 | Fax +49/241/8



- ALL. This is a type of cancer of the blood and bone marrow caused by an abnormal proliferation of lymphocytes.
- AML. This is a type of cancer in the bone marrow characterized by the proliferation of myeloblasts, red blood cells or abnormal platelets.
- CLL. This is a type of cancer characterized by a proliferation of lymphocytes in the bone marrow.
- CML. This is caused by a proliferation of white blood cells in the bone marrow.
- MDS (Myelodysplastic Syndromes). This refers to a group of diseases of the blood and bone marrow in which the bone marrow does not produce a sufficient amount of healthy cells. This can progress to acute leukemia.
- NOL (Normal). No leukemias.

CART is a non-parametric statistical method for extraction of knowledge in classifications.

RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms



Ayed, B. M., Ltifi, H., Kolski,C. & Alimi, A. (2010) A usercentered approach for the design & implementation of KDD-based DSS: A case study in the healthcare domain. Decision Support Systems, 50, 64- 78.

Data Preprocessing

- The medical data is generally collected for patient-care purposes and research is only a secondary consideration - create problems for the data mining tools and techniques.

– Patient records collected for diagnosis and prognosis are characterized by
  – their incompleteness (missing parameter values),
  – incorrectness (systematic or random noise in the data),
  – sparseness (few or non-representable patient records), and
  – inexactness (inappropriate selection of parameters for the given task).

– Genetic data such as microarray gene expression data, microarrays often miss to produce data for a considerable amount of genes, therefore missing values imputation is required

Data Preprocessing

- Data Cleaning
  - Real-word data tend to be incomplete, noisy, and inconsistent.
  - Data cleansing routines attempt to
    - fill in missing values,
    - smooth out noise,
    - identify outliers,
    - correct inconsistencies, and
    - improve data to address quality issues.

Data Preprocessing

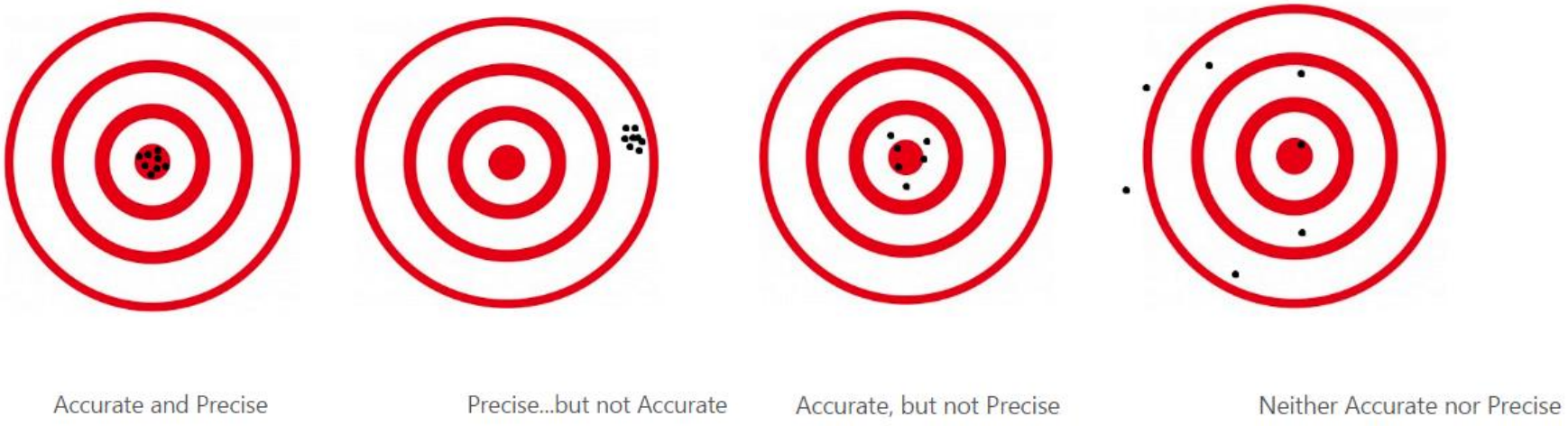
- *Data Quality Dimensions*
  - Data quality = fitness for use

- Specific quality dimensions : accuracy, completeness, currency and consistency
  - Accuracy : a measure of the distance between the data value and the value which is considered correct.
  - Completeness: measures “the extent to which data are of sufficient breadth, depth and scope for the task at hand”
    - The percentage of null values in a column table is an example of (in)completeness.
  - Currency: measures how promptly data are updated .
  - Consistency: refer to the violation of semantic rules defined over (a set of) data items

Data Preprocessing

Accuracy vs Precision

- **accuracy** describes the **difference between** the measurement and the part's actual value,
- **Precision** describes the variation you see when you measure the same part repeatedly with the same device



http://blog.minitab.com/blog/real-world-quality-improvement/accuracy-vs-precision-whats-the-difference

Data Preprocessing

- Data Integration
  - the problem of combining data residing at different sources, and providing the user with a unified view of these data
  - Schema Mapping is the detection of equivalent schema elements in different sources to turn the data into a common representation.
- Data Fusion
  - the last step of a data integration proces
  - the data coming from heterogeneous sources are combined and fused into a single representation,
  - duplicate and inconsistencies in the data are resolved.

Data Preprocessing

Example Case (Boselli, Roberto, et al, 2014):

We want investigate the relationships between the working conditions and some illnesses (especially mental illness)

We have two data source: a medical registry and a labour administrative archive.

medical registry: stores data about drug prescriptions used to treat mental illnesses.

labour administrative archive: records job start and cessation dates

- Example domain rule: by the country law and common practice: an employer can't have more than one full time contract active at the same time.

Data Preprocessing

- An example of data describing the working career of a person

Event #	Event Type	Employer-ID	Date
01	Part Time Start	Firm 1	12^{th}/01/2010
02	Part Time Cessation	Firm 1	31^{st}/03/2011
03	Full Time Start	Firm 2	1^{st}/04/2011
04	Full Time Start	Firm 3	1^{st}/10/2012
05	Full Time Cessation	Firm 3	1^{st}/06/2013
...	...	...	...

- Can you see any problem ?

Data Preprocessing

- An example of data describing the working career of a person

Event #	Event Type	Employer-ID	Date
01	Part Time Start	Firm 1	12^{th}/01/2010
02	Part Time Cessation	Firm 1	31^{st}/03/2011
03	Full Time Start	Firm 2	1^{st}/04/2011
04	Full Time Start	Firm 3	1^{st}/10/2012
05	Full Time Cessation	Firm 3	1^{st}/06/2013
...	...	...	...

- It is inconsistent with respect to the semantic (domain rule) just introduced: a Full Time Cessation is missing for the contract started by *event 03*

- *Typical data cleaning solution:*
  - *add a Full Time Cessation event in a date between event 03 and event 04.*

Data Preprocessing

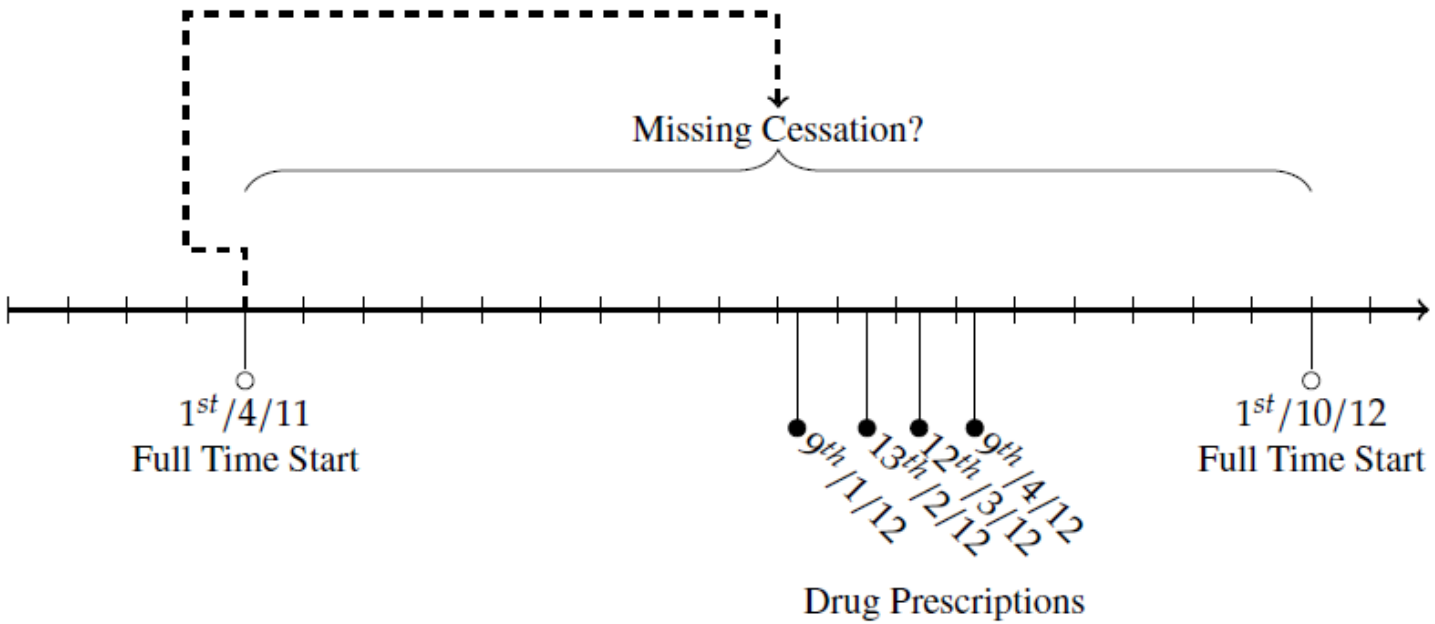
- the corrections performed on a single data source can affect the other archives.
- the labour archive content just introduced will be integrated with data describing drug prescriptions for mental illness.

Event #	Drug Description	Quantity	Date
01	Drug X	...	9^{th}/01/2012
02	Drug X	...	13^{th}/02/2012
03	Drug X	...	12^{th}/03/2012
04	Drug X	...	9^{th}/04/2012
...	...	...	...

- drug prescriptions of the same person whose career has been described
- the person started receiving prescriptions from 9th/01/2012.

Data Preprocessing

- Let us suppose a researcher is investigating the relationships (if any) between unemployment and mental illness.
- In such a case, the Full Time Cessation event to be added by the cleansing procedure is important.
- Depending on the date chosen, the person may be considered receiving the treatment as a worker or as an unemployed



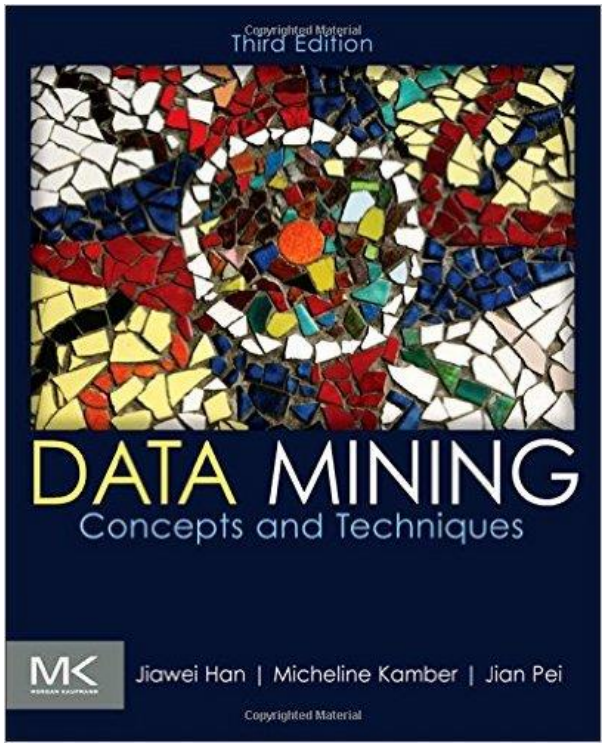
Data Preprocessing

- How to start:
  - Define your quality dimensions
  - Define possible sources of inconsistencies – due to the data integration
  - Explore the characteristics of your data set to understand your cohort
  - Carefully design your solutions for data integration and cleaning , address the possible problems which may occur

Data Preprocessing

Data preprocessing techniques :

- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation



Refer to Data Mining Books:

Jiawei Han, Micheline Kamber,

and Jian Pei, Data Mining: Concepts and Techniques,

3rd edition, Morgan Kaufmann, 2011. (1st ed., 2000) (2nd ed., 2006)

http://hanj.cs.illinois.edu/cs412/bk3/02.pdf

http://hanj.cs.illinois.edu/cs412/bk3/03.pdf

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):
  - Weighted arithmetic mean:
  - Trimmed mean: chopping extreme values
- Median: A holistic measure
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula:

$$\bar{x} = \frac{1}{n}\sum_{i-1}^{n} x_i$$

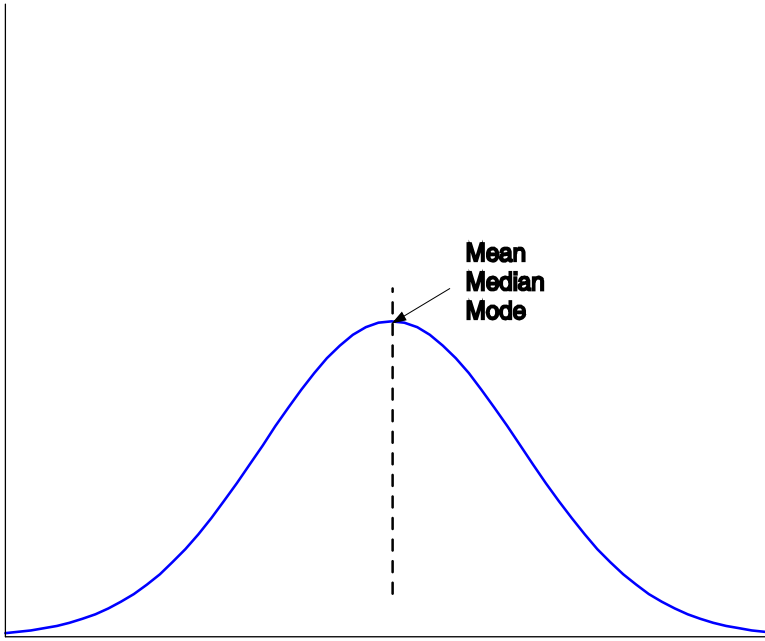
$$\bar{x} = \frac{\sum_{i-1}^{n} w_i x_i}{\sum_{i-1}^{n} w_i}$$

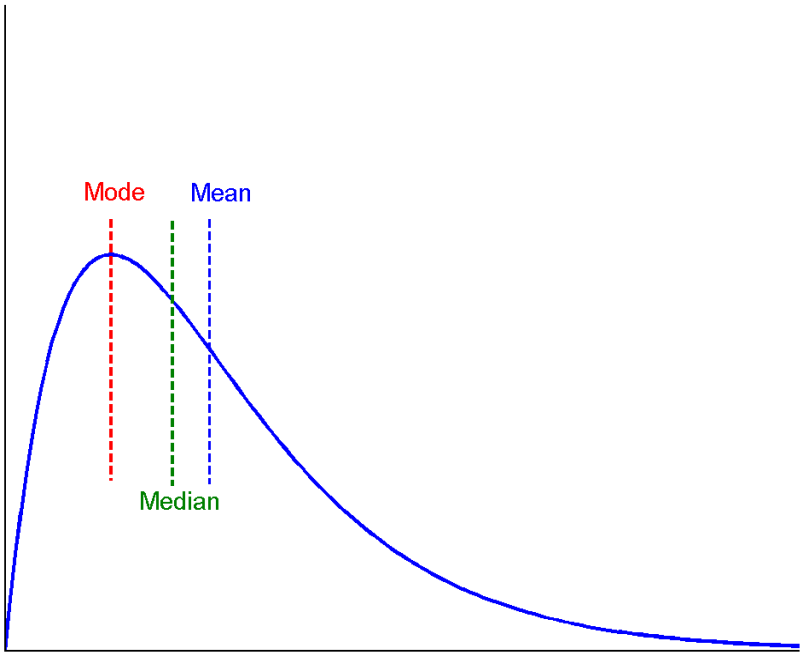
$$median = L_1 + \left(\frac{n/2 - (\sum f)l}{f_{median}}\right)c$$

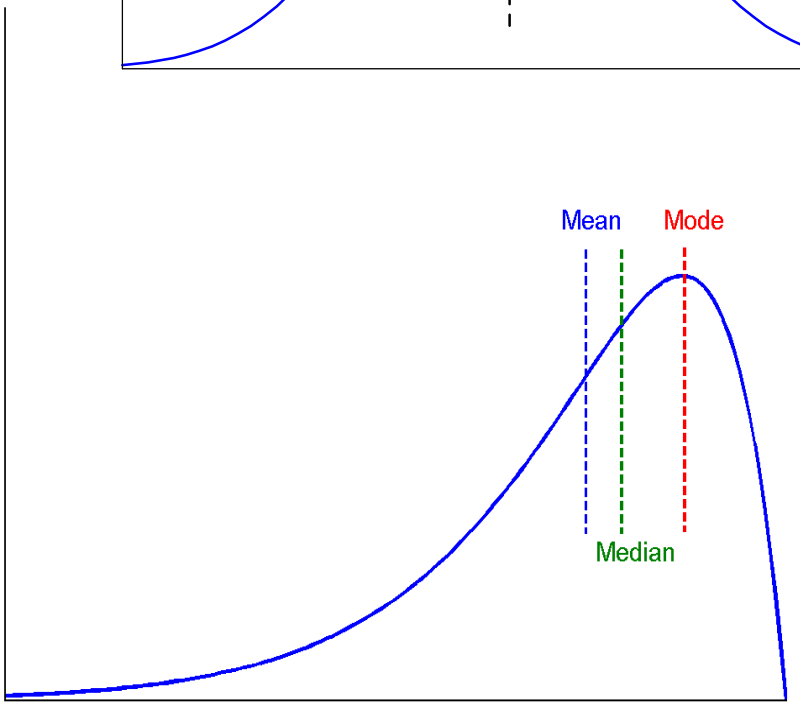
$$mean - mode = 3 \times (mean - median)$$

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data







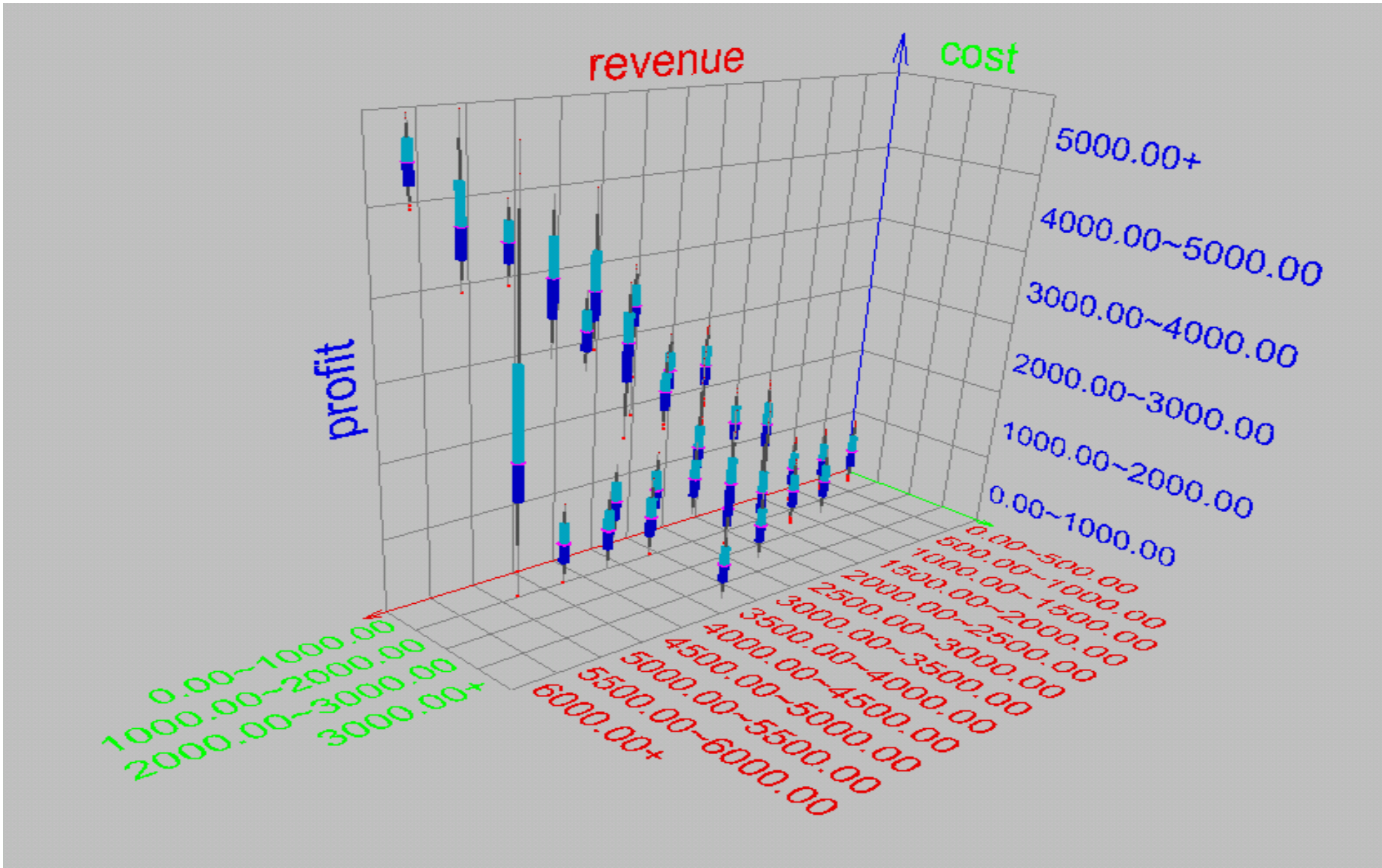
Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)
  - Inter-quartile range: IQR = $Q_3 - Q_1$
  - Five number summary: min, Q_1, M, Q_3, max
  - Boxplot: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually
  - Outlier: usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation (*sample: s, population: σ)*
  - Variance: (algebraic, scalable computation)

$$s^2 = \frac{1}{n-1}\sum_{i-1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i-1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i-1}^{n} x_i\right)^2\right]$$

  - Standard deviation *s (or σ)* is the square root of variance s^2 *(or* σ^2*)*

Visualization of Data Dispersion: Boxplot Analysis



Data Transformation: Normalization

- Min-max normalization: to [new_min$_A$, new_max$_A$]

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A.$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716.$

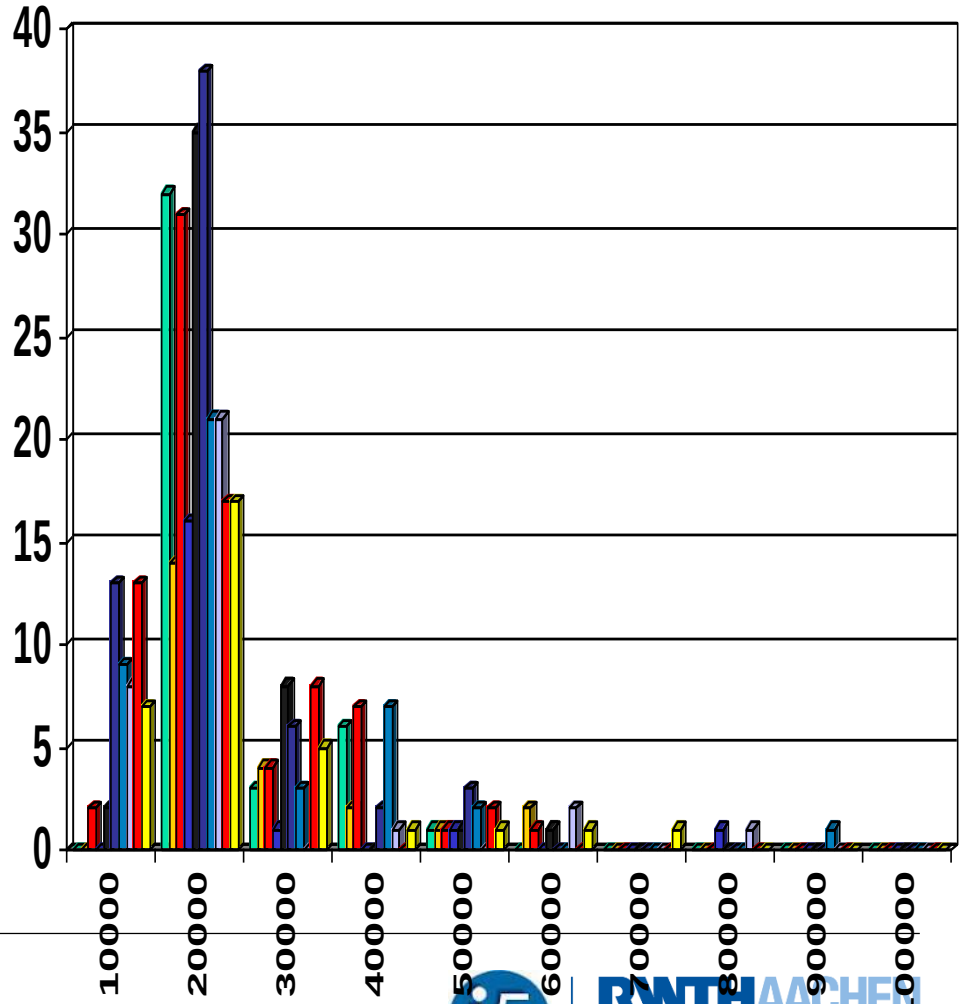
  - Z-score normalization (standardization) (μ: mean, σ: standard deviation)

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A},$$

  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225.$

Data Reduction Method : Histograms

- Divide data into buckets and store average (sum) for each bucket
- Partitioning rules:
  - Equal-width: equal bucket range
  - Equal-frequency (or equal-depth)
  - V-optimal: with the least *histogram variance* (weighted sum of the original values that each bucket represents)
  - MaxDiff: set bucket boundary between each pair for pairs have the β–1 largest differences



RWTH Informatik 5 | Ahornstr. 55 D-52056 Aachen Germany |
Tel +49/241/8021501 | Fax +49/241/8022321 | http://dbis.rwth-aachen.de/cms

References

- Mining Electronic Health Records (EHR): A Survey, Pranjul Yadav, Michael Steinbach, Vipin Kumar, Gyorgy Simon. Minneapolis, MN 55455-0159 USA. October 13, 2015
- Corchado, J. M., De Paz, J. F., Rodriguez, S. & Bajo, J. (2009) Model of experts for decision support in the diagnosis of leukemia patients. Artificial Intelligence in Medicine, 46, 3, 179-200
- Boselli, Roberto, et al. "A policy-based cleansing and integration framework for labour and healthcare data." *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*. Springer Berlin Heidelberg, 2014. 141-168.