Data Mining

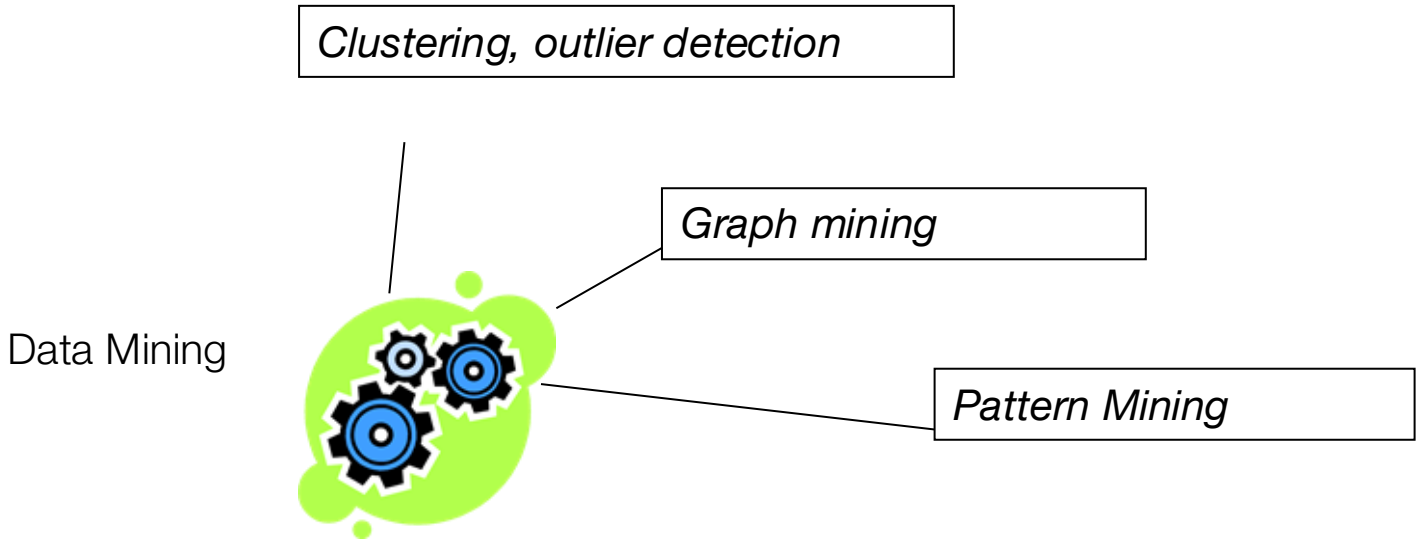
Davide Mottin

Learning goals for today

- What is Data Mining?
- Why am I having this course?
- Basic Data Mining terminology, premiliaries and tasks
- Statistics for exploratory data analysis

Course overview

- Data Mining
- Basic techniques: exploratory statistics
- Clustering and outlier detection
- Graph mining
- Sequence and pattern mining



Course goals

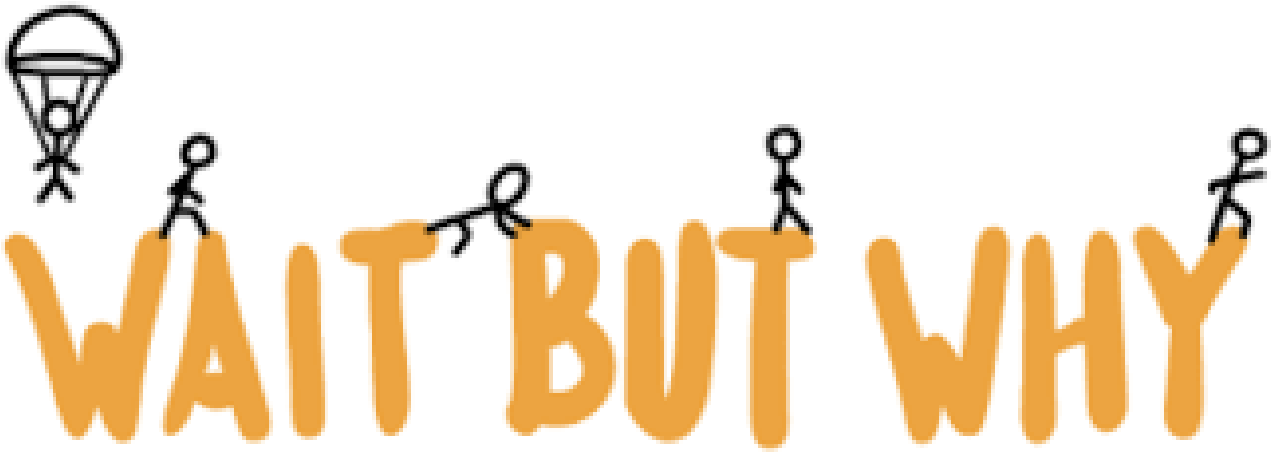
- **Identify** suitable data mining algorithms to use in specific scenarios and data distributions.
- **Describe** and list core data mining concepts and algorithms.
- **Evaluate**, **analyse and explain** the results of algorithms on real data.
- **Compare** and **assess** the advantages and disadvantages of different data mining approaches w.r.t. application requirements.
- **Discuss** and **theorize** on correctness, soundness and runtime properties of data mining algorithms.
- **Generalize** data mining algorithms to different data and different application requirements.

Prerequisites

- Basic **computer science** and **programming**
- Basic **probability theory** and **linear algebra**
- **Machine learning**

Mentimeter warm-up

- What do you expect from this course?



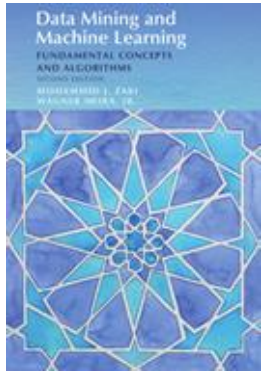
Join at menti.com | use code 4686 4638

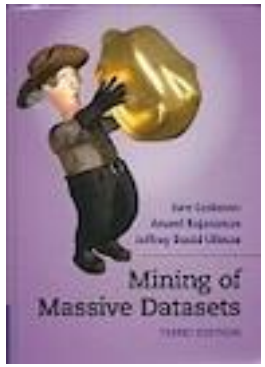
Course organization

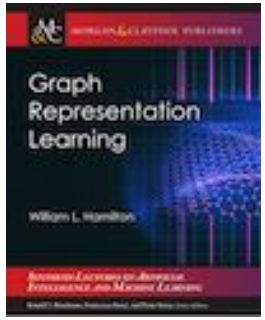
- **Lectures**
  - **Instructors**: Davide Mottin (davide@cs.au.dk)
  - Mentimeter for ensuring the pace / material is adequate
  - Other methods will be used and experimented during the course
  - Tuesday kl. 14-17 , (5123-111)
- **Weekly exercises** (**not graded** useful for exam preparation and feedback)
  - **TAs:** Connor Pugh, Piechen Liu, Tianyi Hu
  - Programming exercises in python, theory questions + project feedback
  - Two groups:
    - Thursday kl. 10-12 5335-192
    - Thursday kl. 10-12 5335-184
- **Project** (**graded, groups of 2-4 people**) 

  - Clustering, graph mining, pattern mining
  - Deadlines announced on brightspace
- 10 ECTS = ca. 15 hours per week
- **PhD edition:** talk to me!

Topics & Material







- **Topics**
  - Introduction to Data Mining
  - Clustering & Outlier detection
  - Graph mining
  - Pattern mining
- **Literature:**
  - Zaki/Meira: Data Mining and Analysis
    - Download: https://dataminingbook.info/book_html/
  - Leskovec, Rajaraman, Ullman: Mining of Massive Datasets
    - Download: http://www.mmds.org/#ver30
  - Hamilton, W.L. Graph representation learning.
    - Download: https://www.cs.mcgill.ca/~wlh/grl_book/files/GRL_Book.pdf
  - Further reading material will be posted in brightspace (articles etc.)

Roadmap

Clustering

- Representative-based
- Density-based
- Hierarchical and Subspace
- Outlier detection

Graph Mining

- Spectral Theory and clustering
- Community Detection
- Link Analysis
- Similarities and Graph Embeddings
- Graph Convolutional Networks

Pattern mining

- Frequent subgraph mining
- Frequent Items and Association Rules
- Text mining

Icons made by Flat Icons, Freepik from www.flaticon.com

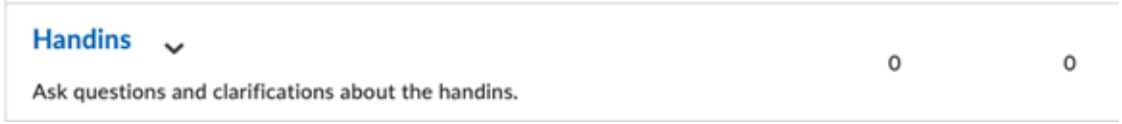
Exam logistic

- **Course project** (30%) 

  - One single submission (mid May)
  - 2 Mandatory Checkpoints + 1 project description
- **Oral exam** (70%)
  - A number of topics covering the three modules
    - more details toward the end of the course
  - 20 minutes total:
    - 10 minutes presentation
    - 8 minutes questions
    - 2 minutes assessment and decision

Project

- Follows the three parts of the course
- Groups of 2-4 people
- Roughly 140 hours **per student**
- No group partner?
  - Post a message in the discussion board



- **Deadline for group enrollment:**
  - **February 3**
- **After deadline** **automatic assignment** to empty groups

Project

- **Two tracks:**
  - Track A: Standard Analysis Track
  - Track B: Research-Oriented Track
- Schedule:
  - All deadlines are at 13.59 (before the lecture)

Weeks	Assignment	Deadline
Week 1	Group Formation	03/02
Week 2-3	Dataset and task selection	17/02
Week 4-6	C1: Clustering and Outlier detection	10/03
Week 7-11	C2: Graph Mining	14/04
Week 12-15	Final: Pattern Mining	15/05

Project

- **Two tracks:**
  - Track A: Standard Analysis Track
  - Track B: Research-Oriented Track
- Schedule:
  - All deadlines are at 13.59 (before the lecture)

Weeks	Assignment	Deadline
Week 1	Group Formation	03/02
Week 2-3	Dataset and task selection	17/02
Week 4-6	C1: Clustering and Outlier detection	10/03
Week 7-11	C2: Graph Mining	14/04
Week 12-15	Final: Pattern Mining	15/05

Track A: Standard analysis

- **Focus**
  - Correct application of data mining algorithms from the course
  - Clear documentation and interpretation of results
- **Expectation**
  - Sound methodology
  - Honest discussion of limitations
  - Coherent reasoning across project stages

Track B: Research-oriented

- Select **one algorithm** from a curated list of research papers provided by the instructors or ask for an additional method.
- Implement the algorithm (fully or partially, possibly simplified).
- Study its behavior by:
  - replicating a key result, **or**
  - comparing it with a course method, **or**
  - exposing a limitation or failure case.

Ideal for group students strongly motivated in data mining that want to go a step forward

Project - Selection

- Choose **one dataset** from the list provided by the instructors **or** bring your own
  - If you bring your own, you need approval as the dataset might not be large or well-structured enough
- Define a short **analysis goal or task** (e.g., structure discovery, anomaly detection, explanation).
- Submit the description of the dataset and task through Brightspace
- The chosen dataset should be significant:
  - Around 10,000 rows
  - Contain ideally text and numeric values
  - Possibility to extract patterns (to be discussed)

Project - Checkpoints

- The project includes **two mandatory checkpoints**, aligned with the three parts of the course.
- Checkpoints:
  - are **mandatory**,
  - receive **written feedback only**,
  - do **not** receive a numeric grade.
- You are expected to **address the feedback** in your final submission.

Project – Final submission

- Submit
  - the ipynb python notebook file
  - the pdf export of the python notebook
  - A declaration of use of Generative AI (see Brightspace)
- The final submission is **graded**

- Generative AI is:
  - Allowed for
    - refinement of text
    - Suggestions on possible analyses / tasks
    - Clarifications about algorithms
  - Not allowed for
    - Copy pasting entirely generated text
    - Generate the implementation

Exercises: goals and requirements

1. Deepen your understanding of the material
2. Apply algorithms *in practice*
3. Prepare you for the exam
4. Getting feedback on the project

Three types of questions

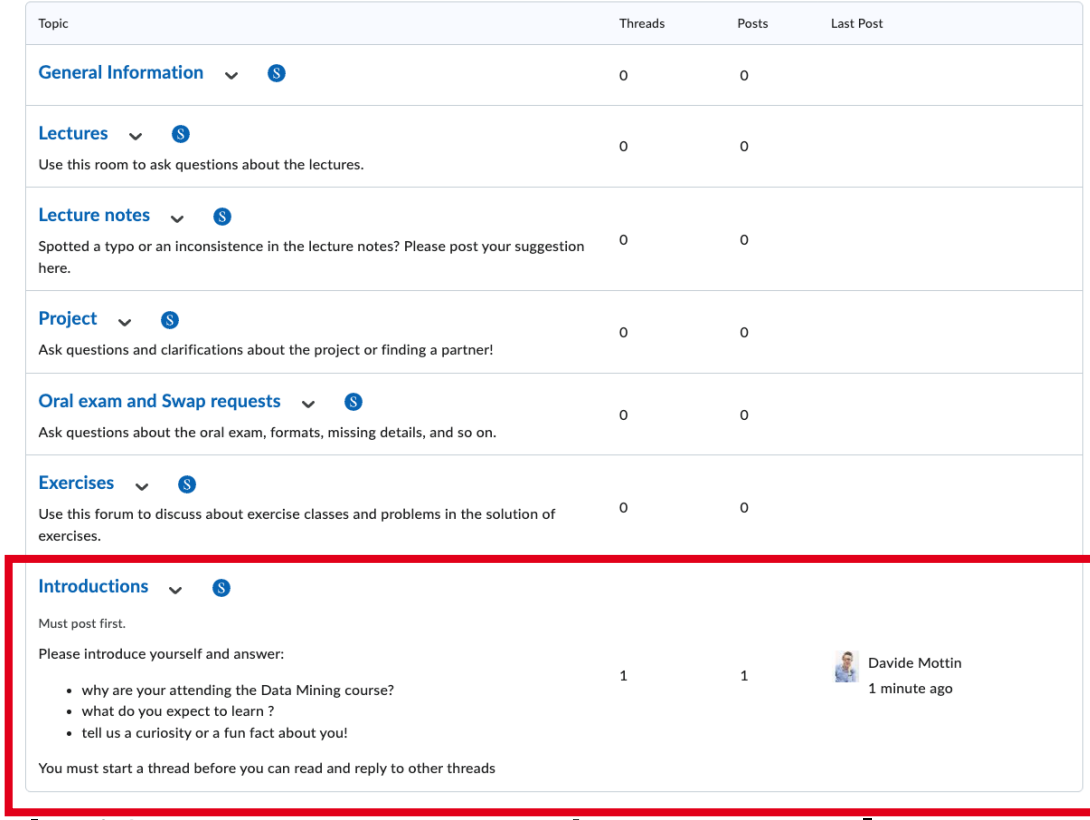
- **Knowledge questions** to warm up
- **Skills questions** to apply the material and consider new aspects
- **Implementation questions**: implement the material, often with a twist: we provide python notebooks
  - This is **NOT** a python course, i.e., python is only a tool for us
- **Published on Tuesdays**

This week Exercises

- Only one room 5335-184
- Basic exercises
  - Recommended for students struggling with python
  - Small feedback on possible groups

Comunication Channels

- Brightspace discussion



Topic	Threads	Posts	Last Post
General Information	0	0	
Lectures Use this room to ask questions about the lectures.	0	0	
Lecture notes Spotted a typo or an inconsistence in the lecture notes? Please post your suggestion here.	0	0	
Project Ask questions and clarifications about the project or finding a partner!	0	0	
Oral exam and Swap requests Ask questions about the oral exam, formats, missing details, and so on.	0	0	
Exercises Use this forum to discuss about exercise classes and problems in the solution of exercises.	0	0	
Introductions Must post first. Please introduce yourself and answer: • why are your attending the Data Mining course? • what do you expect to learn ? • tell us a curiosity or a fun fact about you! You must start a thread before you can read and reply to other threads	1	1	Davide Mottin 1 minute ago

- **Emails:** only if necessary and **extremely** urgent!

Feedback

- The course is a work in progress:
  - Any **feedback** is appreciated
  - Any **comments** on slides and clarity as well
  - There might be some mistake here and there (but I will do my best to polish the material)
  - **Ask questions** if you don't understand something. Better a question in class than a doubt during the exam!

(There's) no such thing as a stupid question

a.k.a. do not be afraid to ask something trivial.
Someone else might have the same question

Intro to Data Mining

Data is a golden mine

- Data explosion problem
  - Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories
- We are drowning in data, but starving for knowledge!
- **Solution:** Data warehousing and data mining
  - Data Warehousing and on-line analytical processing (OLAP)
  - Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

Data contains value and knowledge

Evolution of Database Technology

- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.) and application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s—2000s:
  - Data mining and data warehousing, multimedia databases, and Web databases

What is Data Mining?

- **Given lots of data**
- **Discover patterns and models that are:**
  - **Valid:** hold on new data with some certainty
  - **Useful:** should be possible to act on the item
  - **Unexpected:** non-obvious to the system
  - **Understandable:** humans should be able to interpret the pattern
- What is NOT data mining?
  - (Deductive) query processing.
  - Expert systems or small ML/statistical programs

Data Mining

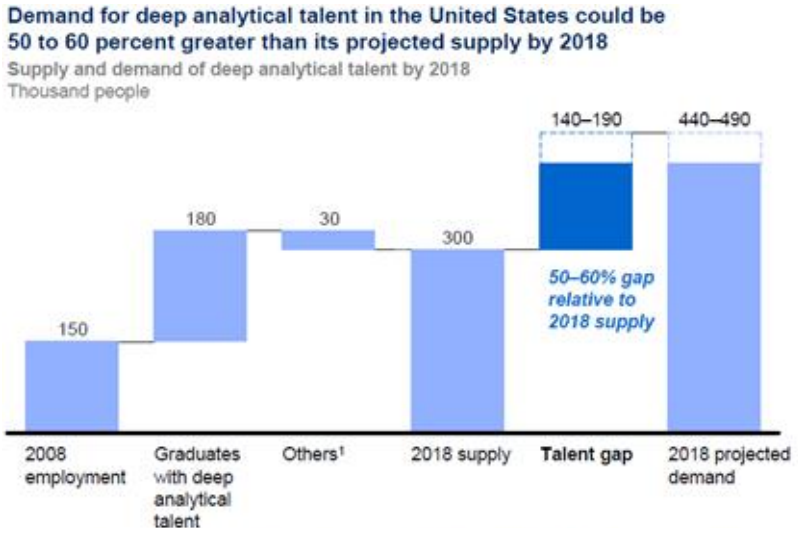
- To extract the knowledge data needs to be
  - Stored → Systems
  - Managed → Databases and Data Management
  - And ANALYZED → this class

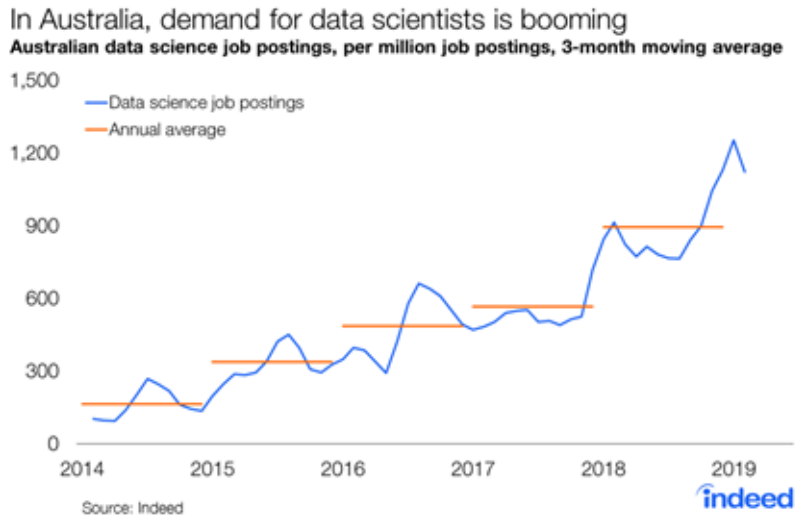
Data Mining ≈ Big Data ≈
Predictive Analytics ≈ Data Science

Data Science is needed everywhere

- Data scientists are needed throughout Denmark; the largest proportion of job are sought after in the main capital area (40% in the municipality of Copenhagen).
- The job market is looking for English-speaking employees, as evidenced by the fact that 66.3% of the job postings are in English.
- Data science competencies are needed in many sectors, i.e., the public sector, companies and universities.

Analysis of the Danish Data Science Job Market, Barbara Plank, 2018





Applications

- Database analysis and decision support
  - Market analysis and management
    - target marketing, customer relation management, market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and management
- Other Applications
  - Text mining (news group, email, documents) and Web analysis.
  - Intelligent query answering

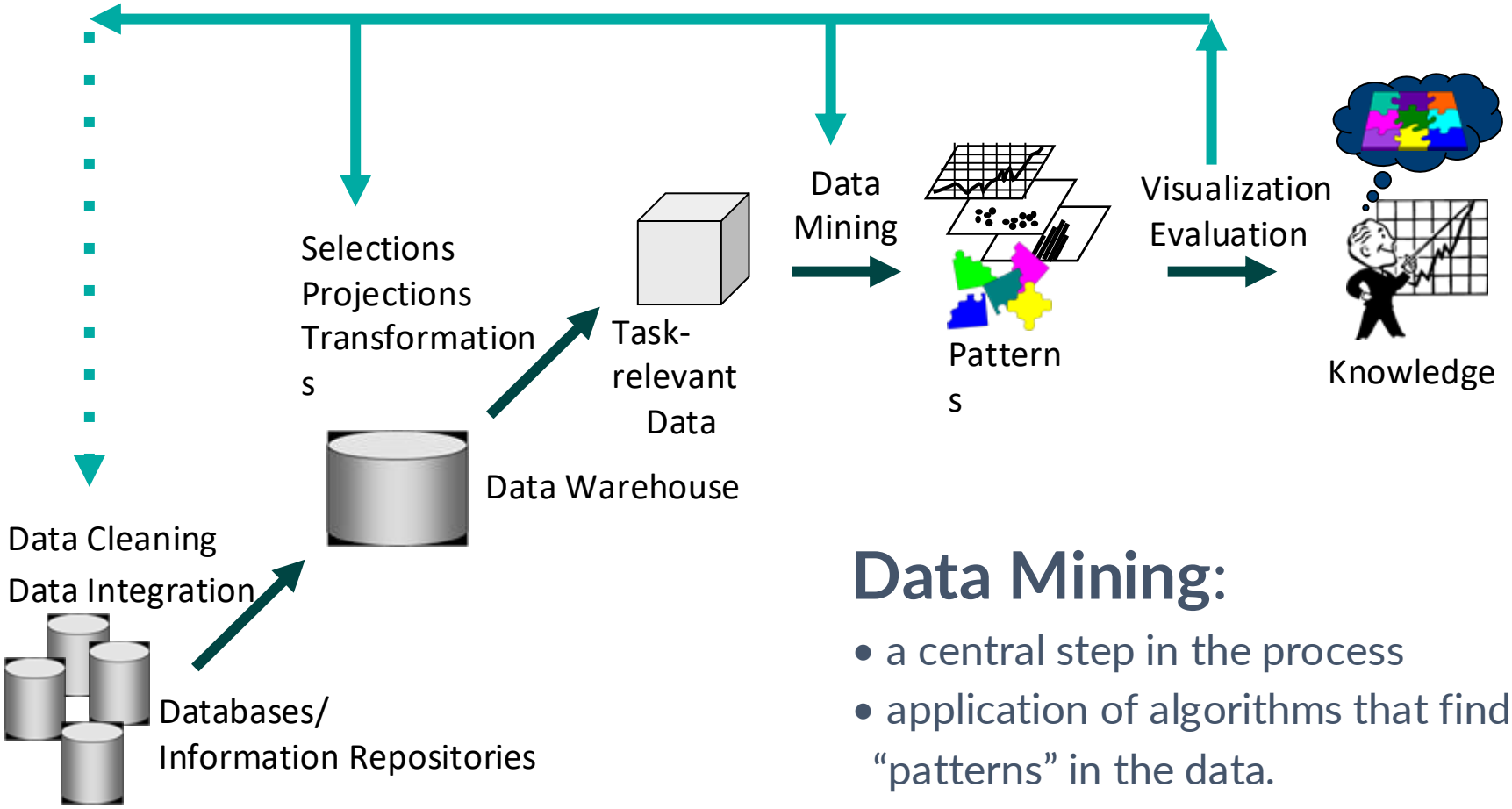
Market Analysis and Management (1)

- Where are the data sources for analysis?
  - Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies
- Target marketing
  - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
- Determine customer purchasing patterns over time
  - Conversion of a single to a joint bank account: marriage, etc.
- Cross-market analysis
  - Associations/co-relations between product sales
  - Prediction based on the association information

Market Analysis and Management (2)

- Customer profiling
  - data mining can tell you what types of customers buy what products (clustering or classification)
- Identifying customer requirements
  - identifying the best products for different customers
  - use prediction to find what factors will attract new customers
- Provide summary information
  - various multidimensional summary reports
  - statistical summary information (data central tendency and variation of the data)

The KDD (knowledge discovery in databases) process

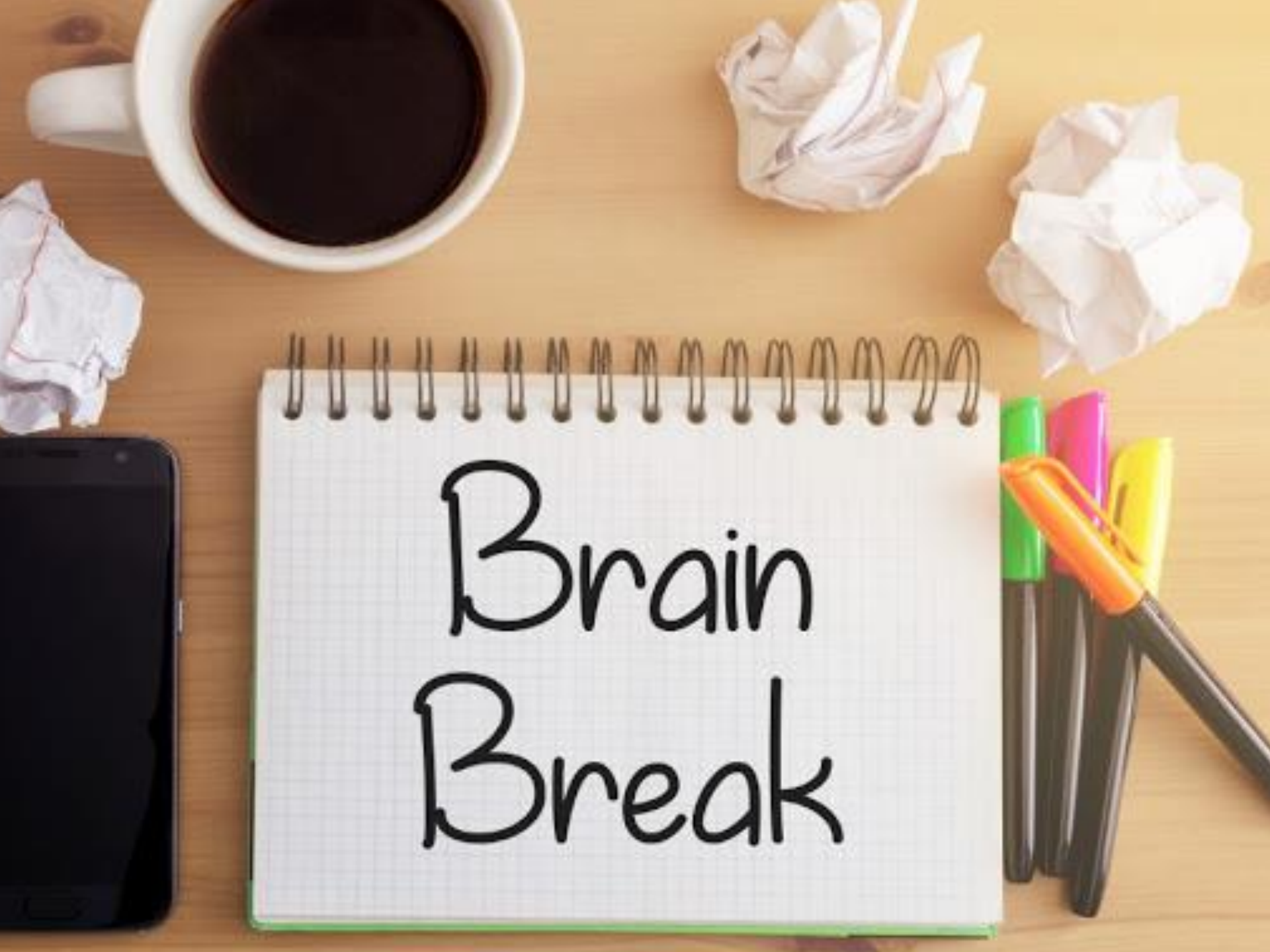


Data Mining:

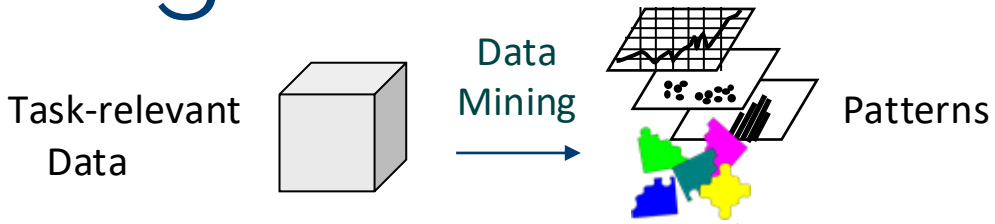
- a central step in the process
- application of algorithms that find "patterns" in the data.

During the break …

- Start thinking to your group
- Imagine an interesting task you could solve
- Do you have a dataset? What analysis could you run?


Brain
Break

Basic Data Mining Tasks



- Clustering
- Association Rules (Frequent Pattern Mining)
- Outlier Detection
- Other methods
  - Concept Characterization and Discrimination
  - Sequential patterns
  - Trends and analysis of changes
  - Methods for special data types, e.g., spatial data mining, web mining
  - …
- Classification (mostly Machine Learning)

Clustering

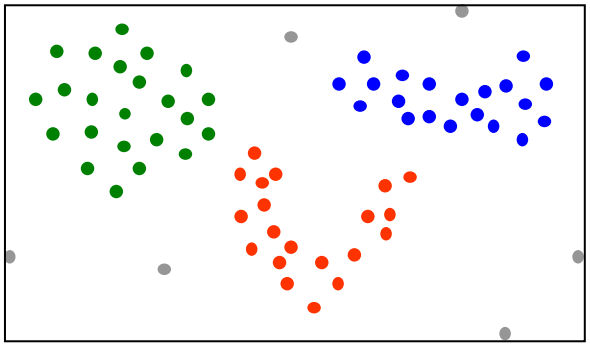
Class labels are unknown

Task: Group objects into sub-groups (clusters)

- Similarity function (or dissimilarity fct. = distance) to measure similarity between objects
- Objective: "maximize" intra-class similarity and "minimize" interclass similarity

Applications:

- Customer profiling/segmentation
- Document or image collections
- Web access patterns
- . . .

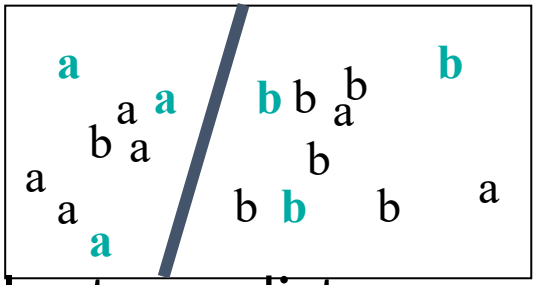


Classification

Class labels are known

Task: Find models/functions/rules based on input examples that

- describe and distinguish classes
- predict class membership for “new” objects



Applications

- Classify gene expression values for tissue samples to predict disease type and suggest best possible treatment
- Automatic assignment of categories to large sets of newly observed celestial objects
- Predict unknown or missing values (? KDD pre-processing step)
- . . .

Supervised vs. Unsupervised Learning

- Supervised learning (classification)
  - Supervision: training data (observations, measurements, etc.) with labels = class of the observations
  - Data is classified based on the training set
- Unsupervised learning (clustering)
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

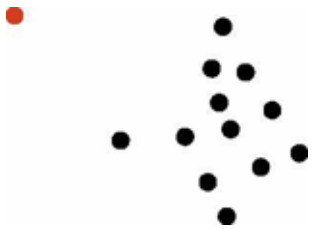
Outlier detection

Hawkins Definition of Outliers

An object that deviates so much from the rest of the data set as to arouse suspicion that it was generated by a different mechanism

Applications:

- Credit card fraud detection
- Telecom fraud detection
- Customer segmentation
- Medical analysis



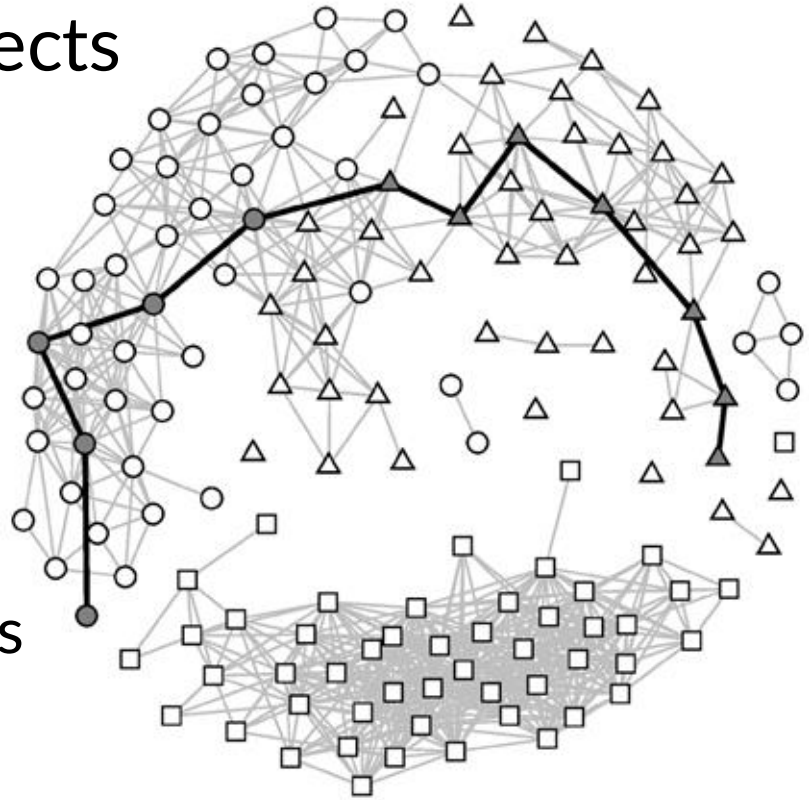
Here: mostly unsupervised; can also be supervised (class imbalance) and semi-supervised (discussed later)

Graph mining

Graph: structure that models connections between data objects

- E.g. friends in a social network
- Protein connections
- Map routes
- ....

- Data Mining on graphs finds additional types of patterns
  - communities in social networks
  - missing links
  - frequent structures
  - ...



Association Rules

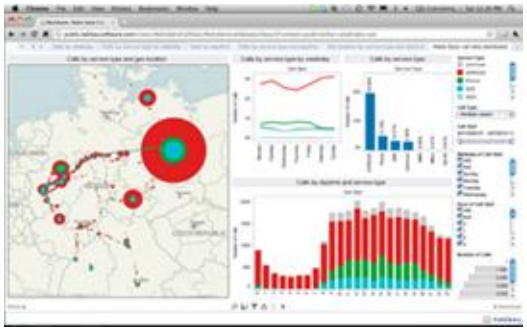
Association rule mining

Find frequent patterns, associations, correlations, or causal structures among sets of items in transactions or other information repositories

- **Rule form:** "Body ⇒ Head [support, confidence]"
- **Applications:** market-basket analysis, cross-marketing, catalog design
  - Which items are bought together frequently?
- **Transaction database**
  - {butter, bread, milk, sugar}; {butter, flour, milk, sugar}; {butter, eggs, milk, salt}; {eggs}; {butter, flour, milk, salt, sugar}
- **Associations:**
  - buys(butter) ? buys(milk), buys(butter) ? buys(sugar)

Exploratory data analysis

Traditional



On data

- **Traditional exploration:** observational studies, monitoring
  - Scientific method
- **Data exploration:** explore data attributes individually or extract key characteristics of a data sample

Data Exploration

Cleaning and profiling

Visualization

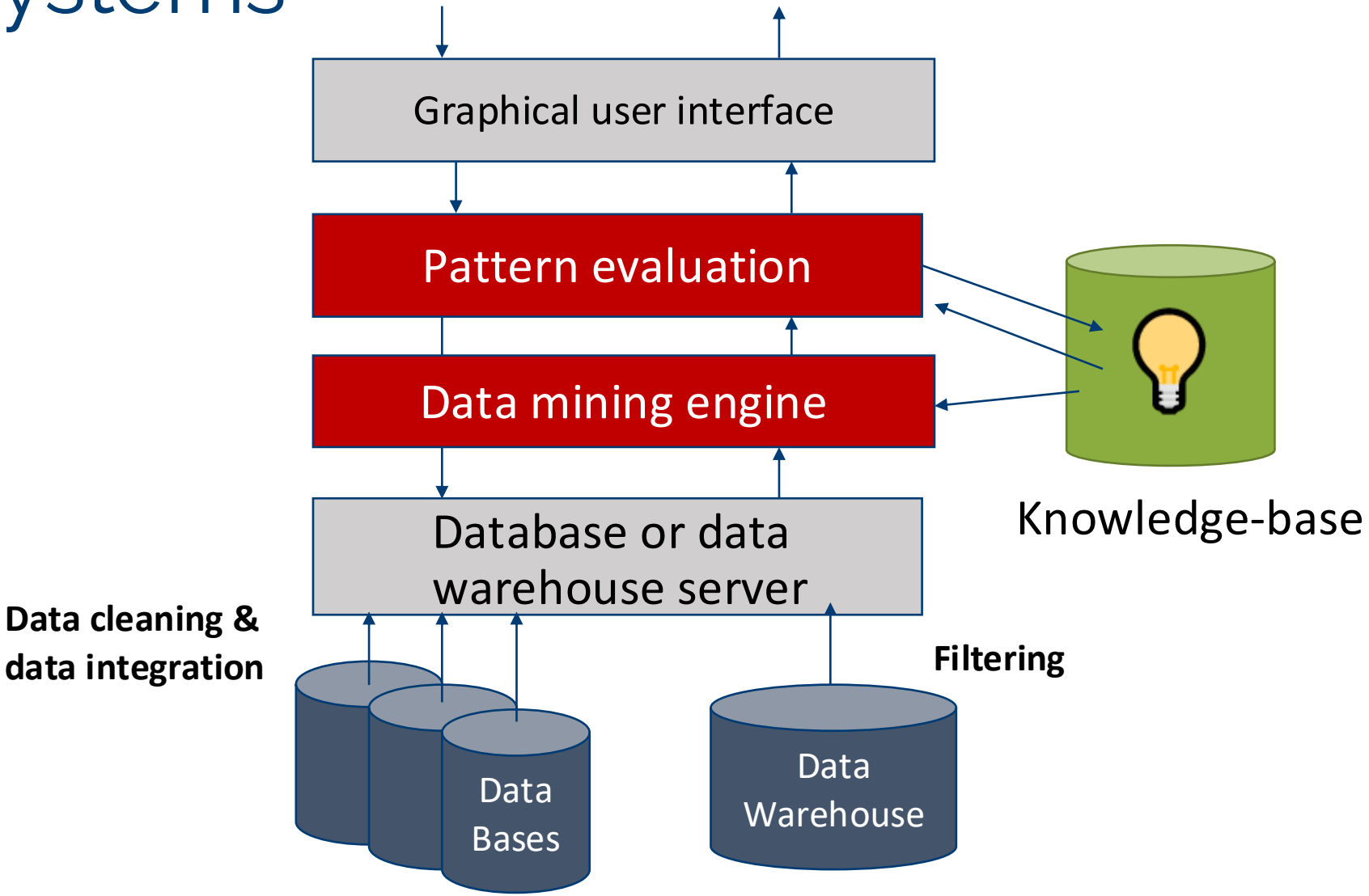
Mining

Check
Data Visualization Course @AU

- Large part of data mining is **data pre-processing** ☹
  - Clean data
  - Select relevant data, profile
- **Visualization** is crucial for human understanding, feedback and iterative data mining steps

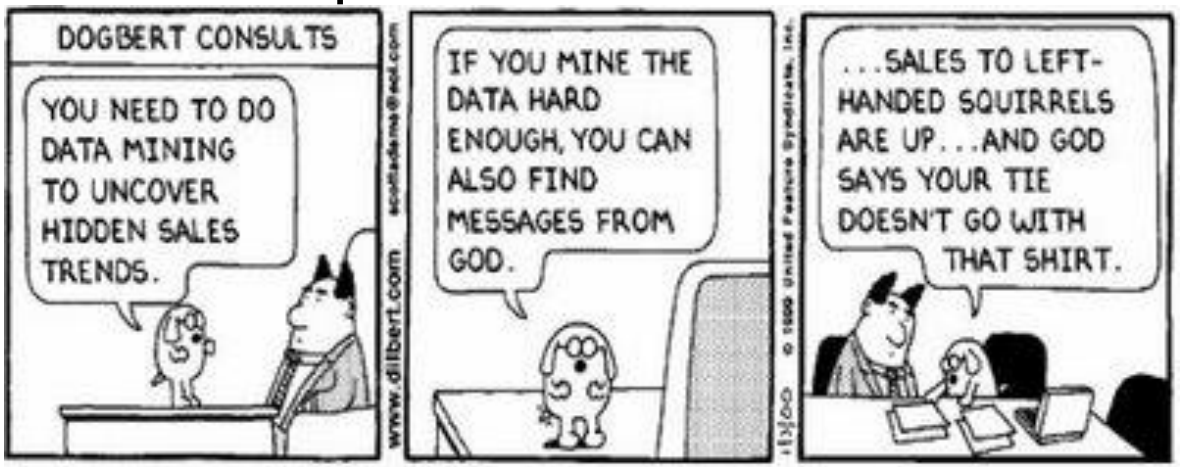
In explorative analysis, user plays a central role (human-in-the-loop)

Architecture of Typical Data Mining Systems



Meaningfulness of Analytic Answers

- **A risk with "Data mining" is that an analyst can "discover" patterns that are meaningless**
- Statisticians call it **Bonferroni's principle**:
  - Roughly, if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap



Are All the "Discovered" Patterns Interesting?

Thousands of patterns: Not all of them are interesting ☹

- Suggested approach: Human-centered, query-based, focused mining

Interestingness measures ☺

A pattern is interesting if it is easily understood by humans, valid on **test** data, potentially useful, novel, or validates some hypothesis that a user seeks to confirm

Objective vs. subjective interestingness measures

- **Objective**: based on statistics and structures of patterns, e.g., support, confidence, etc.
- **Subjective**: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

Conferences and Journals on Data Mining

- KDD Conferences
  - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
  - ACM Web Search and Data Mining (WSDM)
  - The Web Conference (TheWebConf)
  - SIAM Data Mining Conf. (SDM)
  - (IEEE) Int. Conf. on Data Mining (ICDM)
  - Conf. on Principles and practices of Knowledge Discovery and Data Mining (PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
- Other related conferences
  - ACM SIGMOD
  - VLDB
  - (IEEE) ICDE
  - WWW, SIGIR
  - ICML, CVPR, NeurIPS,…
- Journals
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD

Sources: DBLP, CiteSeer, Google Scholar

Data Mining: On What Kind of Data?

- Relational databases
- Data warehouses
- Transactional databases
- Online data (Monitoring, streaming,…)
- Graph data
- Advanced DB and information repositories
  - Object-oriented and object-relational databases
  - Spatial databases
  - Time-series data and temporal data
  - Text databases and multimedia databases
  - Heterogeneous and legacy databases
  - WWW

Statistics for exploratory data analysis

A.k.a. statistics and notions you should have already heard about.

An example dataset

Iris setosa

Iris versicolor

Iris virginica

Four attributes: Sepal length, Sepal width, Petal length, Petal width

Selecting two attributes for visualization using PCA (two principle components). availab on textbook home page

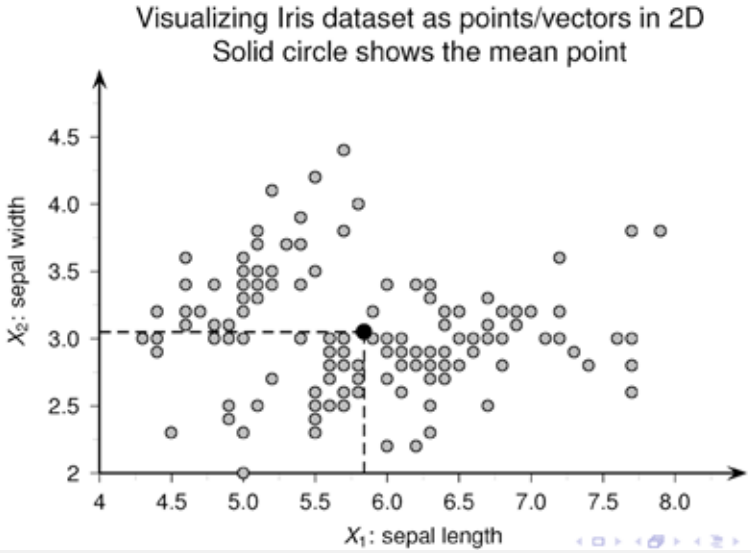
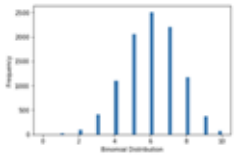


Image sources: Wikipedia

$X: \mathcal{O} \rightarrow \mathbb{R}$ random variable

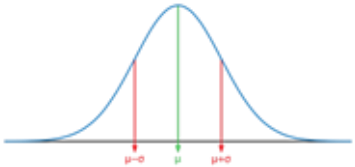
Random variables



X discrete

Probability mass function

- $f(x) = P(X = x)$
- $f(x) \geq 0$
- $\sum_x f(x) = 1$



X continuous

Probability density function

- $P(X \in [a, b]) = \int_a^b f(x)dx$
- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)dx = 1$

Is $f(x) \leq 1$?

NO! $f(x) = \frac{P(X \in [x-\varepsilon, x+\varepsilon])}{2\varepsilon}$

Cumulative distribution function

$$F(x) = P(X \leq x) = \sum_{u \leq x} f(u)$$

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du$$

Quantile function: minimum value with probability q

$$F^{-1}(q) = \min\{x | F(x) \geq q\}$$

Mean

- Random variable X
- **Mean:** arithmetic average of the values of X
- Location or central tendency of distribution of X

If X is discrete

$$\mu = E[X] = \sum_x xf(x)$$

$f(x)$ is X probability distribution

If X is continuous

$$\mu = E[X] = \int_{-\infty}^{+\infty} x\, f(x)dx$$

$f(x)$ is X probability density function

Sample mean

The *sample mean* is a statistic, i.e., a function $\hat{\mu}: \{x_1, \ldots, x_n\} \rightarrow \mathbb{R}$, that is the average value of x_i

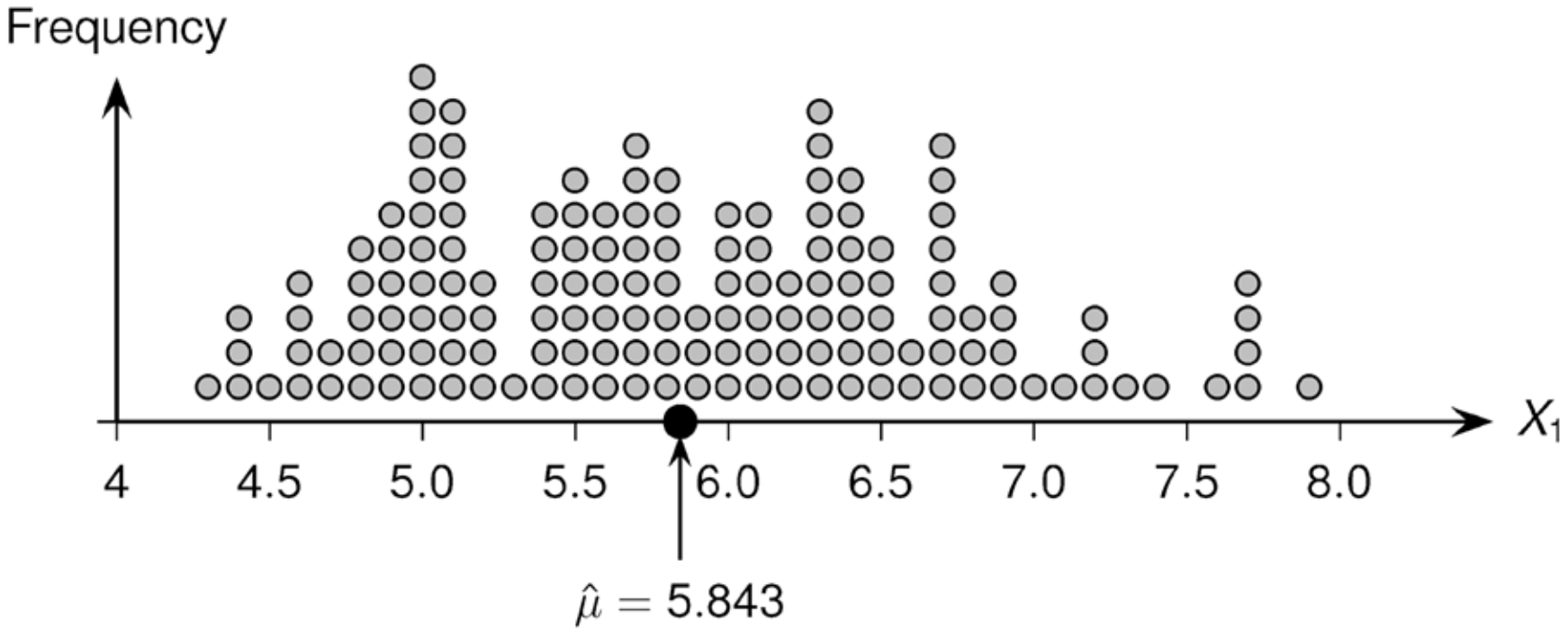
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- The sample mean is an **estimator** for the **unknown** mean μ
- A generic estimator $\hat{\theta}$ is unbiased for parameter θ if $E[\hat{\theta}] = \theta$ for every θ.
  - e.g., the sample mean is unbaiased
- **A statistic is robust** if it is not affected by extreme values

Robustness

Join at menti.com | use code 4686 4638

Sample mean for iris sepal length



Median

- The median of random variable is the value m

$$P(X \leq m) \geq \frac{1}{2} \text{ and } P(X \geq m) \geq \frac{1}{2}$$

- Is the middle-most value:
  - half of the values of X are less and half more than m
- **Alternative:** The median is the value m for which

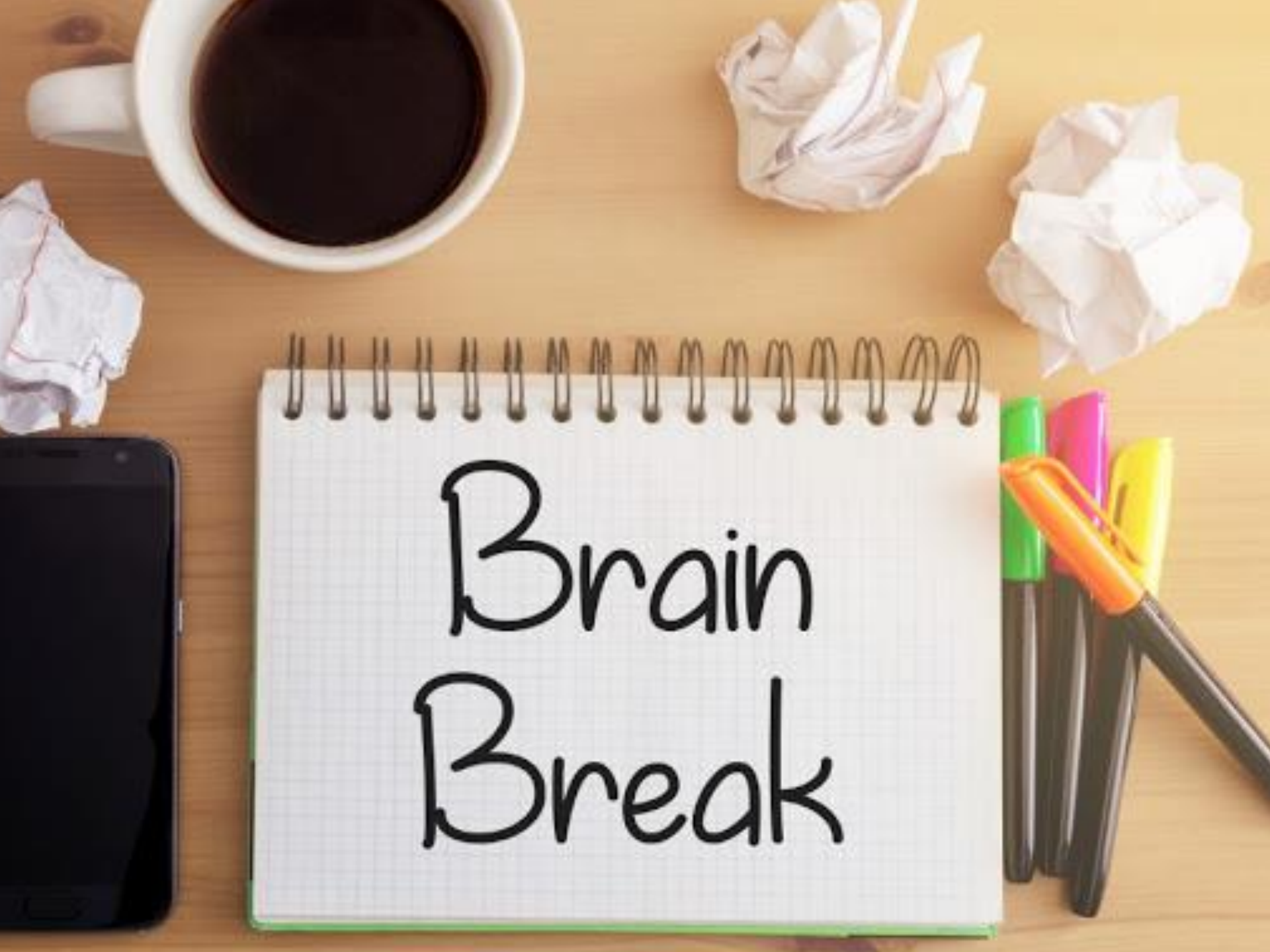
$$F(m) = 0.5 \text{ or } F^{-1}(0.5) = m$$

- The median is robust
  - It is not affected by extreme values

Mode

- Random variable X
- The value in which $f(x)$ attains maximum value
- The sample mode is the value for which the empirical probability attains its maximum

$$\text{mode}(X) = \arg\max_x \hat{f}(x)$$


Brain
Break

Covariance

- The covariance between two attributes X_1, X_2 is a measure of the association of the linear dependence

$$\sigma_{12} = E\left[(X_1 - \mu_1)(X_2 - \mu_2)\right] = E\left[X_1 X_2\right] - E[X_1][X_2]$$

If X_1 and X_2 are independent, then

$$E\left[X_1 X_2\right] - E[X_1]E[X_2] \Rightarrow \sigma_{12} = 0$$

The sample covariance is

$$\hat{\sigma}_{12} = \frac{1}{n}\sum_{i=1}^{n}(X_1 - \hat{\mu}_1)(X_2 - \hat{\mu}_2)$$

Correlation

- Correlation is the standardized covariance
- Obtained through normalization with the standard deviation of each variable

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

What is the sample correlation?

Geometric interpretation of covariance and correlation

If we denote with Z_1 and Z_2 the attribute vectors centered in 0:

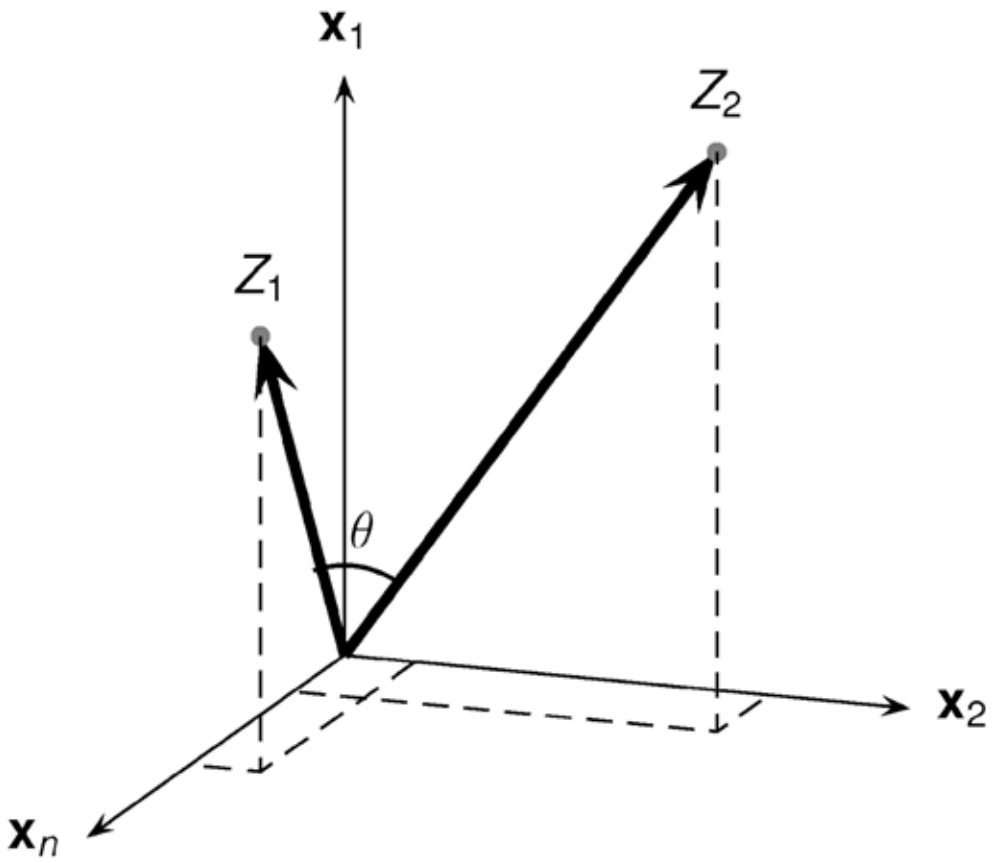
$$Z_1 = X_1 - \mathbf{1} \cdot \hat{\mu}_1 = \begin{pmatrix} x_{11} - \hat{\mu}_1 \\ x_{21} - \hat{\mu}_1 \\ \vdots \\ x_{n1} - \hat{\mu}_1 \end{pmatrix} \qquad Z_2 = X_2 - \mathbf{1} \cdot \hat{\mu}_2 = \begin{pmatrix} x_{12} - \hat{\mu}_2 \\ x_{22} - \hat{\mu}_2 \\ \vdots \\ x_{n2} - \hat{\mu}_2 \end{pmatrix}$$

The sample covariance is $\hat{\sigma}_{12} = \frac{Z_1^T Z_2}{n}$ and the sample correlation is ???

You will see that in the exercises!

Geometric interpretation of covariance and correlation

Covariance: $\hat{\sigma}_{12} = \frac{Z_1^T Z_2}{n}$



Covariance Matrix

- The covariance matrix contains variances and covariances of any set of attributes.
- For two attributes X_1, X_2 the covariance matrix is

$$\boldsymbol{\Sigma} = E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

- Since $\sigma_{12} = \sigma_{21}, \boldsymbol{\Sigma}$ is symmetric
- The total variance is

$$var(\{X_1, X_2\}) = tr(\boldsymbol{\Sigma}) = \sigma_1^2 + \sigma_2^2$$

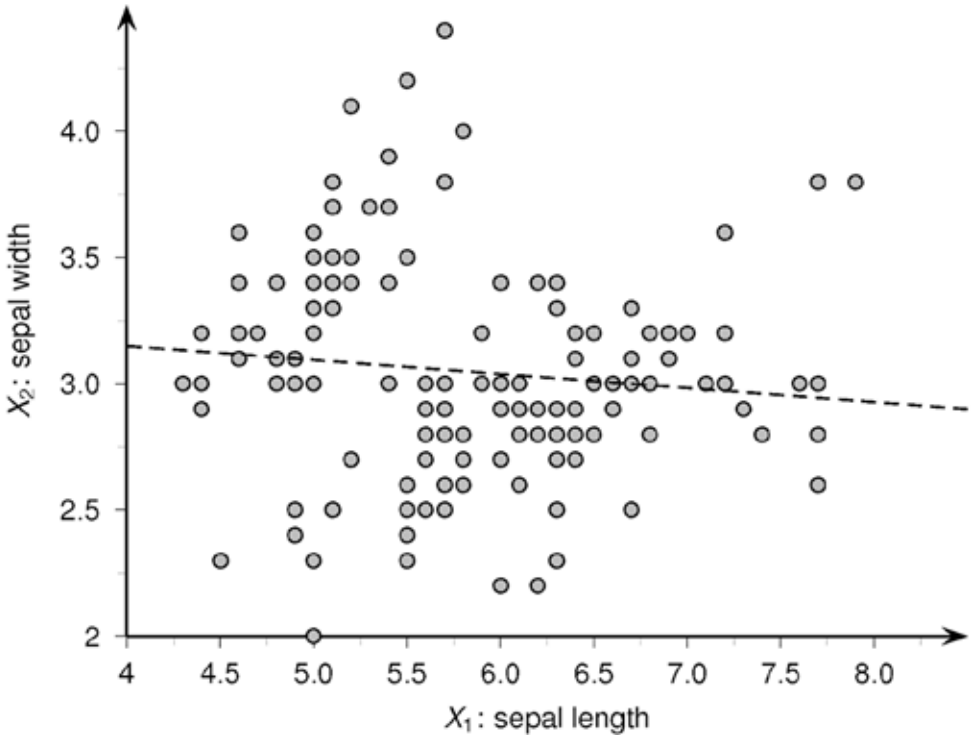
so $tr(\Sigma) \geq 0$.

- The generalized variance is

$$|\boldsymbol{\Sigma}| = \det(\boldsymbol{\Sigma}) = \sigma_1^2\sigma_2^2 - \sigma_{12}^2 = (1 - \rho_{12}^2)\sigma_1^2\sigma_2^2$$

Since $|\rho_{12}| \leq 1$, then $\det(\boldsymbol{\Sigma}) \geq 0$

Correlation sepal width/length



The sample mean is

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} 5.843 \\ 3.054 \end{pmatrix}$$

The sample covariance matrix is

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix}$$

The sample correlation is

$$\hat{\rho}_{12} = \frac{-0.039}{\sqrt{0.681 \cdot 0.187}} = -0.109$$

Data Normalization

If the attributes are in different scales (e.g. km, cm), than normalization is necessary.

Range Normalization

- X attribute, $x_1, \ldots, x_n$ dataset of points (sample).
- Range normalization scales the values by the range (max - min) in the range [0,1]

$$x_i' = \frac{x_i - \min_i\{x_i\}}{\max_i\{x_i\} - \min_i\{x_i\}}$$

Standard Score normalization (z-normalization)

- Each value is replaced by its z-score:

$$x_i' = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

- The new attribute has mean 0 and standard deviation 1

Normalization example

$\mathbf{x}_i$	Age (X_1)	Income (X_2)
$\mathbf{x}_1$	12	300
$\mathbf{x}_2$	14	500
$\mathbf{x}_3$	18	1000
$\mathbf{x}_4$	23	2000
$\mathbf{x}_5$	27	3500
$\mathbf{x}_6$	28	4000
$\mathbf{x}_7$	34	4300
$\mathbf{x}_8$	37	6000
$\mathbf{x}_9$	39	2500
$\mathbf{x}_{10}$	40	2700

Since `Income` is much larger, it dominates `Age`. The sample range for `Age` is $\hat{r} = 40 - 12 = 28$, whereas for `Income` it is $2700 - 300 = 2400$. For range normalization, the point $\mathbf{x}_2 = (14, 500)$ is scaled to

$$\mathbf{x}_2' = \left(\frac{14 - 12}{28}, \frac{500 - 300}{2400}\right) = (0.071, 0.035)$$

For z-normalization, we have

	Age	Income
$\hat{\mu}$	27.2	2680
$\hat{\sigma}$	9.77	1726.15

Thus, $\mathbf{x}_2 = (14, 500)$ is scaled to

$$\mathbf{x}_2' = \left(\frac{14 - 27.2}{9.77}, \frac{500 - 2680}{1726.15}\right) = (-1.35, -1.26)$$

Attribute Dependence: Contingency Analysis

The contingency table for X_1, X_2 is the matrix of observed counts n_{ij} for value i in X_1 and value j in X_2

$$\mathbf{N}_{12} = n \cdot \hat{\mathbf{P}}_{12} = \begin{pmatrix} n_{11} & n_{12} & \cdots & n_{1m_2} \\ n_{21} & n_{22} & \cdots & n_{2m_2} \\ \vdots & \vdots & \ddots & \vdots \\ n_{m_1 1} & n_{m_1 2} & \cdots & n_{m_1 m_2} \end{pmatrix}$$

$\hat{\mathbf{P}}_{12}$ is the empirical joint Probability Mass Function (PMF) for X_1 and X_2.
The contingency table is augmented with row and column marginal counts

$$\mathrm{N}_1 = n \cdot \hat{\boldsymbol{p}}_1 = \begin{pmatrix} n_1^1 \\ \vdots \\ n_{m_1}^1 \end{pmatrix}, \qquad \mathrm{N}_2 = n \cdot \hat{\boldsymbol{p}}_2 = \begin{pmatrix} n_1^2 \\ \vdots \\ n_{m_2}^1 \end{pmatrix}$$

$\mathrm{N}_1, \mathrm{N}_2$ have multinomial distribution with parameters $\hat{\boldsymbol{p}}_1, \hat{\boldsymbol{p}}_2$.
Also $\mathbf{N}_{12}$ is multinomial with parameters $\{\hat{P}_{ij}\}$

Contingency table: sepal length vs sepal width

Sepal length (X_1)	Sepal width (X_2)			
		Short a_{21}	Medium a_{22}	Long a_{23}	Row Counts
	Very Short (a_{11})	7	33	5	$n_1^1 = 45$
	Short (a_{12})	24	18	8	$n_2^1 = 50$
	Long (a_{13})	13	30	0	$n_3^1 = 43$
	Very Long (a_{14})	3	7	2	$n_4^1 = 12$
	Column Counts	$n_1^2 = 47$	$n_2^2 = 88$	$n_3^2 = 15$	$n = 150$

Chi-Square (χ^2) Independence Test

- X_1 and X_2 are independent
- Their joint PMF is $\hat{p}_{ij} = \hat{p}_i^1 \cdot \hat{p}_j^2$
- The expected frequency for each pair of values is

$$e_{ij} = n \cdot \hat{p}_{ij} = \cdots = \frac{n_i^1 n_j^2}{n}$$

- The χ^2 statistic quantifies the difference between observed and expected counts

$$\chi^2 = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

What is the probability of obtaining such a avalue?

The sampling distribution for χ^2 follows the chi-squared density function.

$$f(x|q) = \frac{1}{2^{q/2}\Gamma(q/2)} x^{\frac{q}{2}-1} e^{-\frac{x}{2}}$$

Where $q = (m_1 - 1)(m_2 - 1)$ is the degree of freedom.

χ^2 In the example

	Expected Counts	X_2		
		Short (a_{21})	Medium (a_{22})	Short (a_{23})
X_1	Very Short (a_{11})	14.1	26.4	4.5
	Short (a_{12})	15.67	29.33	5.0
	Long (a_{13})	13.47	25.23	4.3
	Very Long (a_{14})	3.76	7.04	1.2

	Observed Counts	X_2		
		Short (a_{21})	Medium (a_{22})	Long (a_{23})
	Very Short (a_{11})	7	33	5
	Short (a_{12})	24	18	8
	Long (a_{13})	13	30	0
	Very Long (a_{14})	3	7	2

The chi-squared statistic value is $\chi^2 = 21.8$.

The number of degrees of freedom are

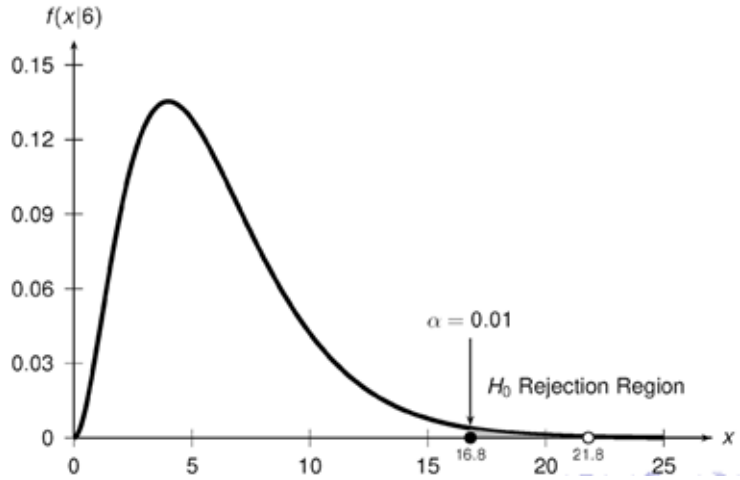
$$q = (m_1 - 1) \cdot (m_2 - 1) = 3 \cdot 2 = 6$$

p-value of Independence test

The p-value of a statistic θ (in our example χ^2) is:

the probability that a particular statistical measure, such as the mean or standard deviation, of an **assumed probability distribution** will be greater than or equal to observed results

In practice: fix a α, e.g. 0.01 . If the p-$value(z) \leq \alpha$ we can reject the null hypothesis (e.g., the two variables are independent). The p-value can be interpreted only under the null-hypothesis being true.



Link: Common misuse of p-values

Multiway contingeny analysis

Details!

Given $\mathbf{X} = (X_1, X_2, \cdots, X_d)^T$. The chi-squared statistic is given as

$$\chi^2 = \sum_{\mathbf{i}} \frac{(n_{\mathbf{i}} - e_{\mathbf{i}})^2}{e_{\mathbf{i}}} = \sum_{i_1=1}^{m_1} \sum_{i_2=1}^{m_2} \cdots \sum_{i_d=1}^{m_d} \frac{(n_{i_1,i_2,\ldots,i_d} - e_{i_1,i_2,\ldots,i_d})^2}{e_{i_1,i_2,\ldots,i_d}}$$

Under the null hypothesis, that attributes are independent, the expected number of occurrences of the symbol tuple $(a_{1i_1}, a_{2i_2}, \ldots, a_{di_d})$ is given as

$$e_{\mathbf{i}} = n \cdot \hat{p}_{\mathbf{i}} = n \cdot \prod_{j=1}^{d} \hat{p}_{i_j}^{j} = \frac{n_{i_1}^1 n_{i_2}^2 \ldots n_{i_d}^d}{n^{d-1}}$$

The total number of degrees of freedom for the chi-squared distribution is given as

$$\begin{aligned} q &= \prod_{i=1}^{d} |dom(X_i)| - \sum_{i=1}^{d} |dom(X_i)| + (d-1) \\ &= \left(\prod_{i=1}^{d} m_i\right) - \left(\sum_{i=1}^{d} m_i\right) + d - 1 \end{aligned}$$

Distances and similarities

d dimensions, s common values

- Eucledian distance

$$\delta(x_i, x_j) = \|x_i - x_j\| = \sqrt{x_i^\top x_i - 2x_i x_j + x_j^\top x_j} = \sqrt{2(d-s)}$$

In binary vectors

- Hamming distance

$$\delta_H(x_i, x_j) = d - s$$

- Cosine similarity of an angle θ among vectors:

$$\cos\theta = \frac{x_i^\top x_j}{\|x_i\|\|x_j\|} = \frac{s}{d}$$

- Jaccard Coefficient

$$J(x_i, x_j) = \frac{s}{2(d-s)+s} = \frac{s}{2d-s}$$

Measuring Similarity

- Similarity is often measured with a distance function $dist$
  - Small $dist(x, y)$: x and y are more similar
  - Large $dist(x, y)$: x and y are less similar
- **Properties** of a distance function
  1. $\text{dist}(x, y) \geq 0$ (positive semidefinite)
  2. $\text{dist}(x, y) = 0 \Leftrightarrow x = y$ (definite)
  3. $\text{dist}(x, y) = \text{dist}(y, x)$ (symmetry)
  4. $\text{dist}(x, z) \leq \text{dist}(x, y) + \text{dist}(y, z)$ (triangle inequality)
- **With** *triangle inequality* dist is a metric
- **Without** *definite condition* dist is a pseudometric
- Definition of distance function **application dependent**

Metrics

Minkowski distances

- For standardized numerical attributes, i.e., vectors $x = (x_1, ..., x_d)$ and $y = (y_1, ..., y_d)$ from a d-dimensional vector space:
  - General L_p-Metric (Minkowski-Distance) $$d_p(x,y) = \sqrt[p]{\sum_{i=1}^{d} |x_i - y_i|^P}$$
  - $p = 2$: Euclidean Distance (cf. Pythagoras) $$d_2(x,y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$$
  - $p = 1$: Manhattan-Distance (city block) $$d_1(x,y) = \sum_{i=1}^{d} |x_i - y_i|$$
  - $p \to \infty$: Maximum-Metric $$d_\infty(x,y) = \max\{|x_i - y_i|, 1 \le i \le d\}$$

Discretization

- Converts numeric into categorical attributes
- Also called binning
- **Equal-width Intervals**
  - Partition into k bins of equal width w

$$w = \frac{x_{max} - x_{min}}{k}$$

  - The ith interval boundary is

$$v_i = x_{min} + iw, \text{for } i = 1, \dots, k-1$$

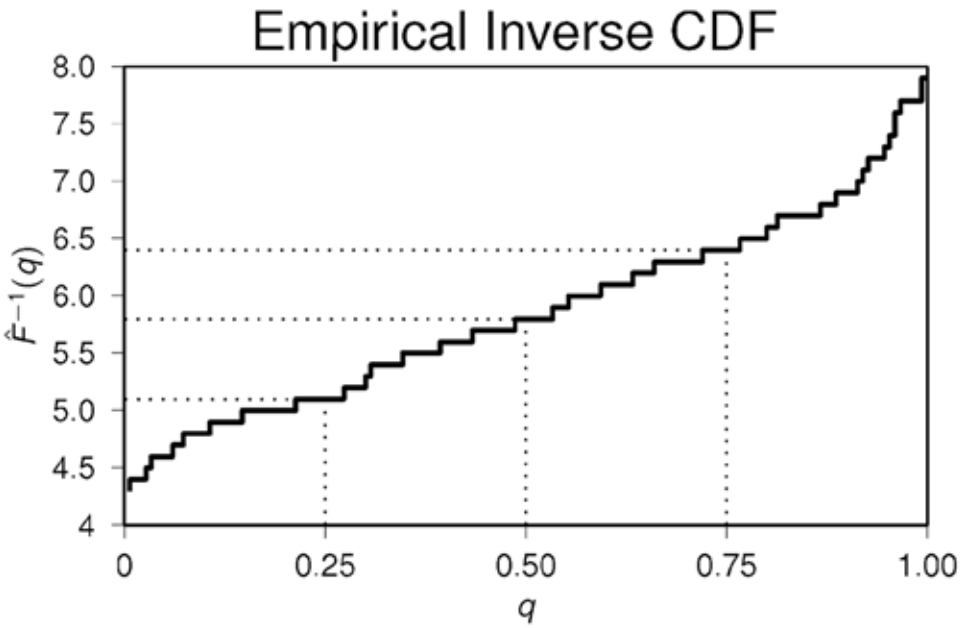
- **Equal-Frequency (equi-depth) intervals:**
  - Divide the range X into intervals with the same number of points computed from the inverse cumulative distribution function

$$\hat{F}^{-1}(q) = \min\{x | P(X \leq x) \geq q\}$$

  - Each interval contain 1/k of the probability mass – The boundaries are

$$v_i = \hat{F}^{-1}\left(\frac{i}{k}\right) \text{ for } i = 1, \dots, k-1$$

Equal-Frequency discretization: sepal length (4 bins)



Quartile values:

$$\hat{F}^{-}1(0.25) = 5.1$$

$$\hat{F}^{-}1(0.5) = 5.8$$

$$\hat{F}^{-}1(0.75) = 6.4$$

Range: $[4.3, 7.9]$

Bin	Width	Count
$[4.3, 5.1]$	0.8	$n_1 = 41$
$(5.1, 5.8]$	0.7	$n_2 = 39$
$(5.8, 6.4]$	0.6	$n_3 = 35$
$(6.4, 7.9]$	1.5	$n_4 = 35$

Summary

- **Data Mining**
  - Find novel, interesting patterns in data
  - Unsupervised approaches
  - In this course: exploratory, clustering, outlier detection, graph mining and frequent pattern mining
- **Exploratory Data Mining**
  - Characterize the data (often statistically)
  - Univariate, bivariate, multivariate
- **Data preprocessing**
  - Normalization, standardization

- **Next week**: representative-based clustering

What did you (not) grasp today?

What was today's lecture about?

- Data Mining is defined as…
- …. is not considered Data Mining.
- The major difference to Machine Learning is…
- Important characteristics for exploratory data mining are…

Acknowledgements

- Slides for book Data Mining and Analysis by Mohammed J Zaki and Wagner Meira Jr
- Slides for book Data Mining: Concepts and Techniques, 2nd ed. by Jiawei Han and Micheline Kamber
- Slides for course on Knowledge Discovery in Databases by Jörg Sander and Martin Ester
- Slides for course Data Mining Algorithms by Thomas Seidl
- Slides for course Advanced Data Mining Algorithms by Ira Assent

References

- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.
- M. H. Dunham. Data Mining: Introductory and Advanced Topics. Prentice Hall, 2003.
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.
- T. Imielinski and H. Mannila. A database perspective on knowledge discovery. Communications of the ACM, 39:58-64, 1996.
- G. Piatetsky-Shapiro, U. Fayyad, and P. Smith. From data mining to knowledge discovery: An overview. In U.M. Fayyad, et al. (eds.), Advances in Knowledge Discovery and Data Mining, 1-35. AAAI/MIT Press, 1996.
- G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991.
- D. Hand, H. Mannila, P. Smyth. Principles of Data Mining. MIT Press, 2001.