Data and Text Mining

Data representation and manipulation I

prof. dr. Bojan Cestnik
Temida d.o.o. & Jozef Stefan Institute
Ljubljana
bojan.cestnik@temida.si

Contents

- **Introduction**
- Basic Data Mining process
- Data kinds and formats
- ER diagram
- Data exploration
- Data preparation

- Examples

Study guide and rules for IKT2

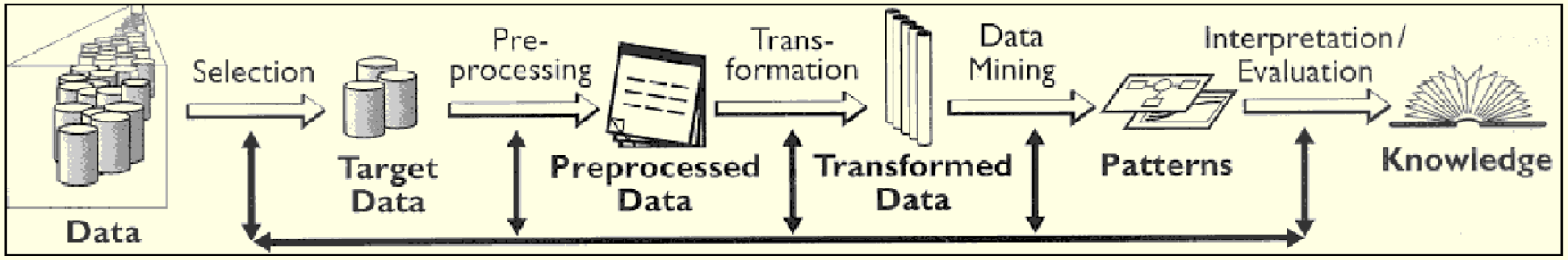
- Lecture schedule
  - Wednesday, 20. 11. 2024 15:00 - 19:00
- Web page: www.temida.si/~bojan/MPS/
- Literature for study
- Collaboration during lectures
- Exam

Basic Data Mining process

- **Input:** transaction data table, relational database, text documents, web pages
- **Goal:** construct a classification model, find interesting patterns in data, etc.

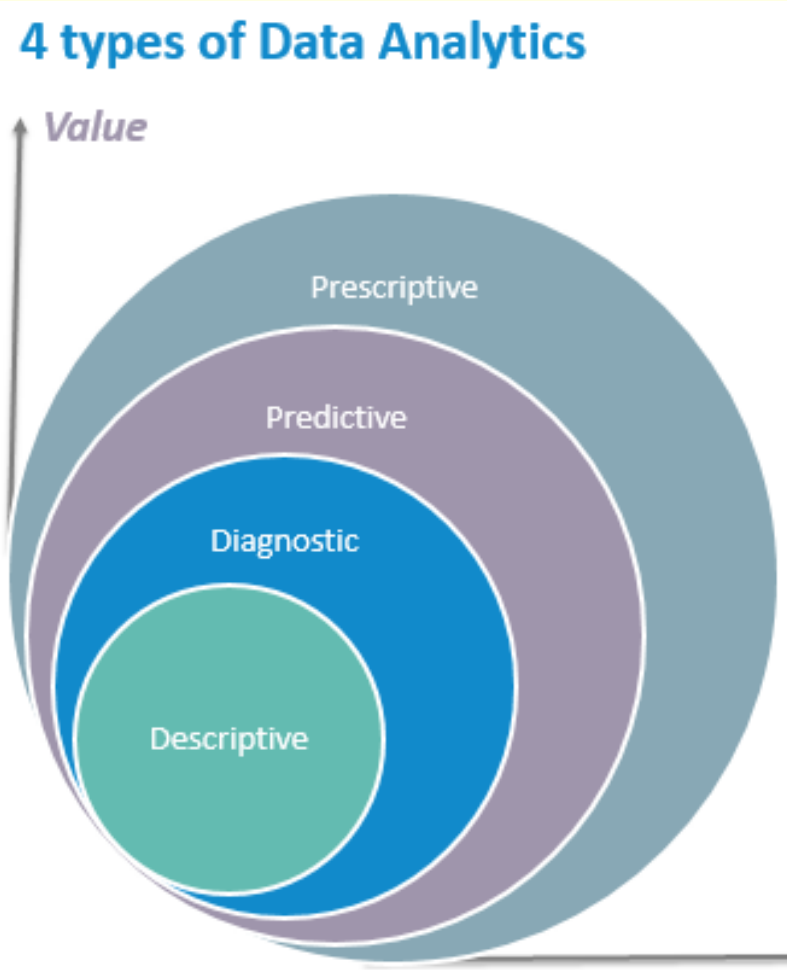
- Your turn - Q11: % of data preprocessing?
- https://www.qtvity.com/index_eng.php

KDD process



- KDD (Knowledge Discovery in Databases) process involves several steps
  - Data preparation
  - Data mining
  - Evaluation and use of discovered patterns
- Data Mining is the key step
  - Only 15%-25% of the entire KDD process

Types of Data Analytics



What is the data telling you?

Descriptive: ***What's happening in my business?***

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: ***Why is it happening?***

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: ***What's likely to happen?***

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: ***What do I need to do?***

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

Principa
www.principa.co.za

https://www.kdnuggets.com/2017/07/4-types-data-analytics.html

Data kinds and formats

- Kinds of data:
  - Descriptive tables: instances, attributes, classes
  - Texts: documents, paragraphs, sentences, words
  - Multimedia: pictures, music, movies
  - …
- Data formats:
  - Relational databases
  - .xls: Excel table format
  - .csv: comma-separated file
  - .arff: attribute-relation file format (Weka)
  - …

Data sources example

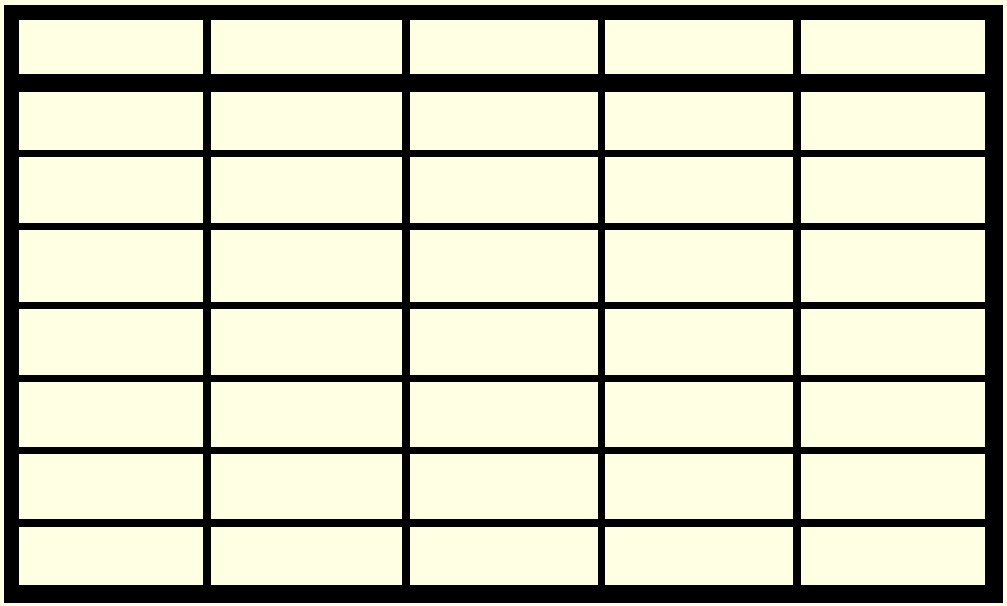
- Local telephone company:
  - When the call was placed, who called, how long the call lasted, etc.
- Catalog company:
  - Items ordered, time and duration of calls, promotion response, credit card used, shipping method, etc.
- Credit card processor:
  - Transaction date, amount charged, approval code, vendor number, etc.
- Credit card issuer:
  - Billing record, interest rate, available credit update, etc.
- Package carrier:
  - Zip code, value of package, time stamp at truck, time stamp at sorting center, etc.

Tables I

- Single table: instances, attributes, classes

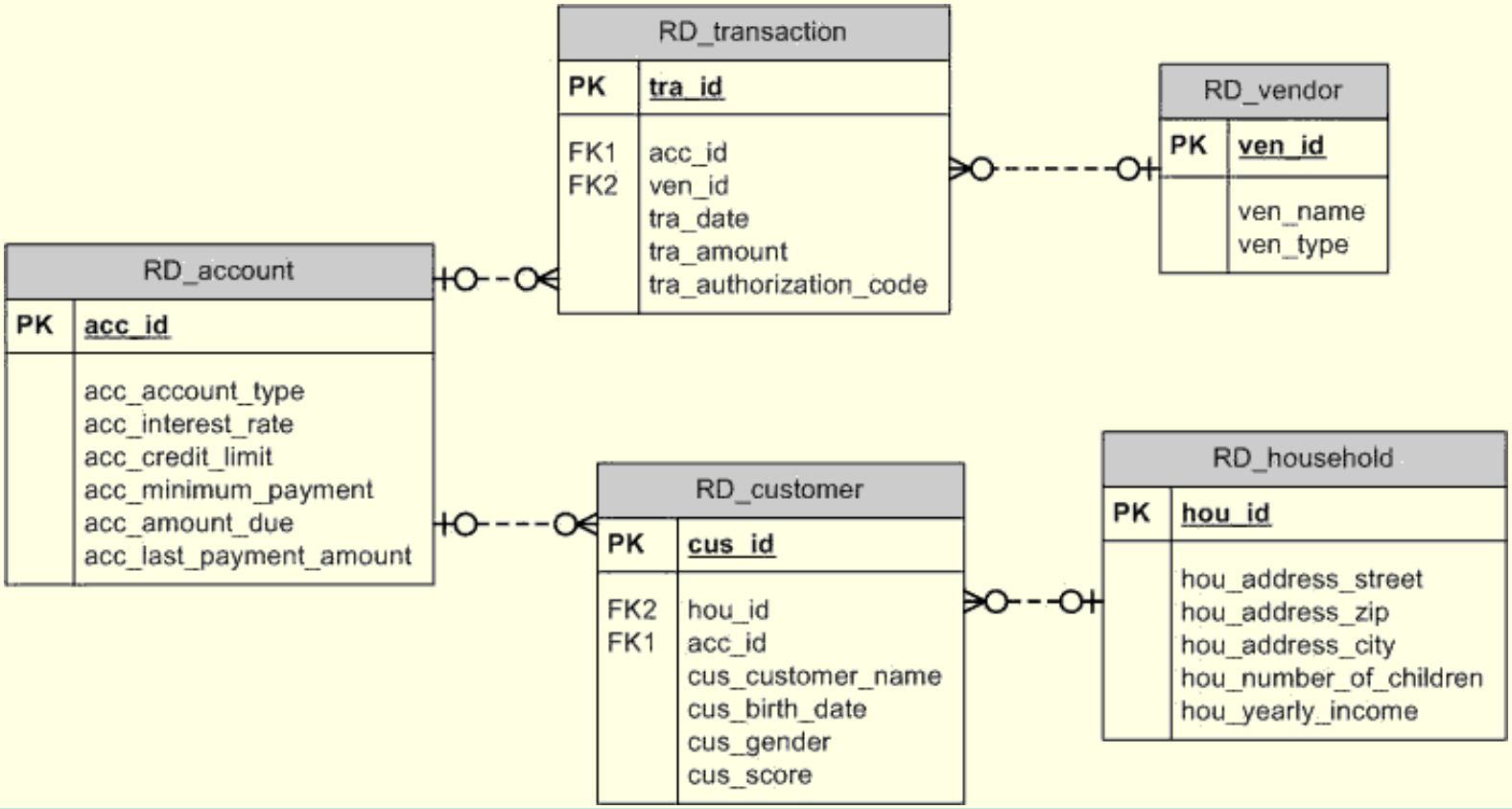
attributes class

instances



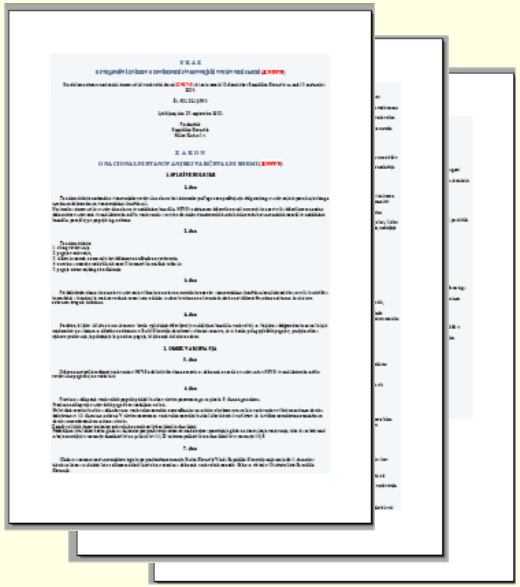
Tables II

- Many tables: relations, ER diagram



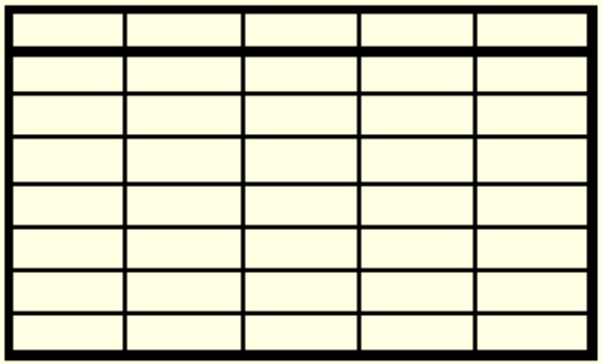
Texts I

- Documents, web pages, etc.
- Transformations: lemmatization, stop-words, named entities, etc.
- Bag-of-words representation



documents

words



Texts II

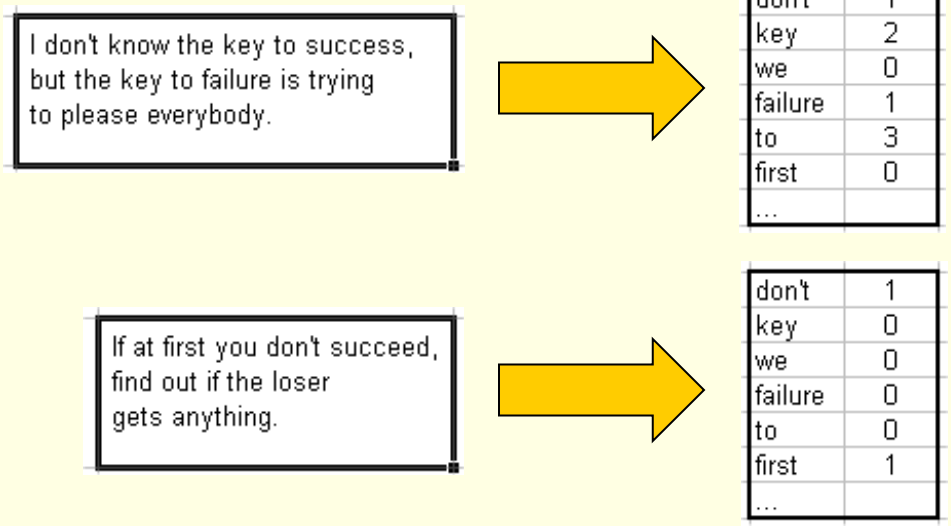
- Areas of text processing
  - Semantic web – Knowledge representation and Reasoning
  - Information retrieval – Search in DB
  - Natural language processing – Computational linguistics
  - Text mining – Data analysis

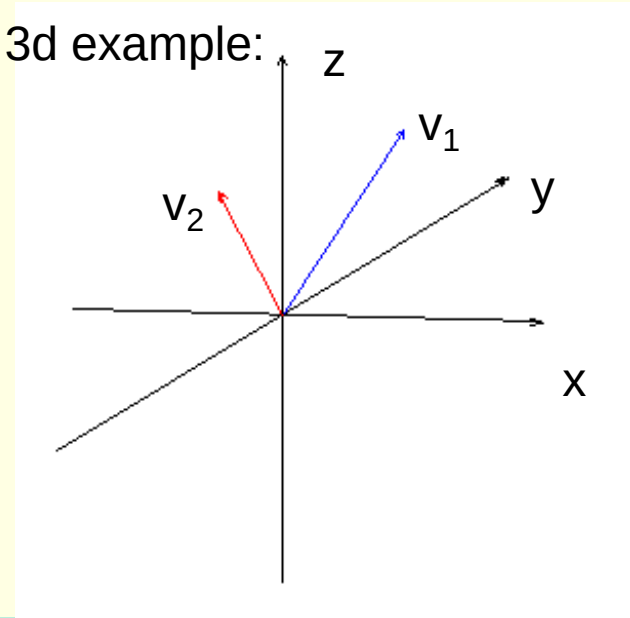
Texts III

- TFIDF measure for word relevance
  - (Term Frequency * Inverse Document Frequency)
  - Term Frequency: word frequency in a particular document (paragraph)
  - Inverse Document Frequency: how infrequent a word is in the collection of all documents (paragraphs)

Texts IV – document similarity

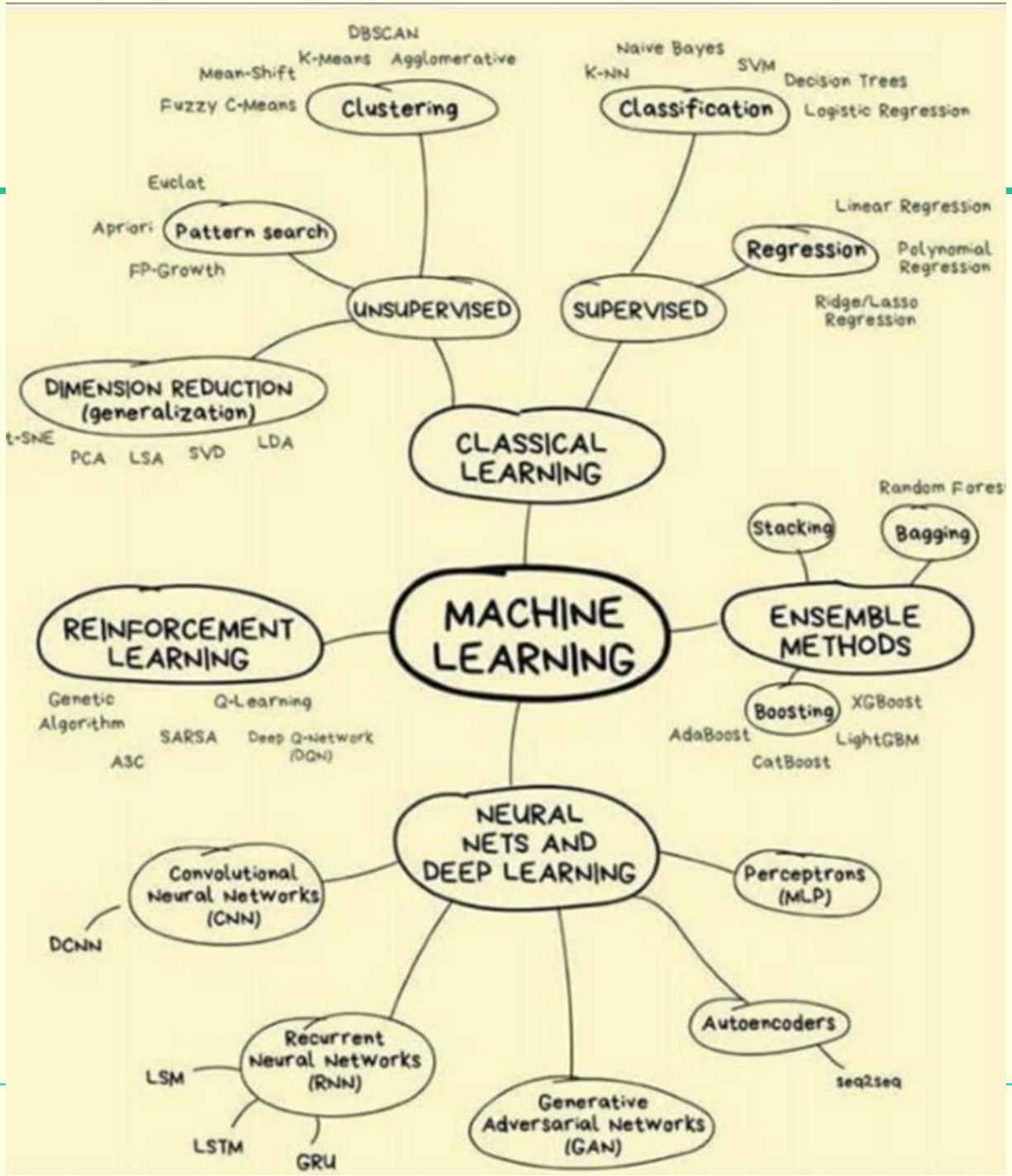
- Ideal: semantic similarity
- Practical: statistical similarity
  - Representation of documents as vectors
  - Cosine similarity between documents (embeddings)





Machine learning

- Developing algorithms and models that enable machines to learn from data
- Key concepts:
  - **Supervised learning:** ML algorithms learn **from labeled data** (input-output pairs) to make predictions about new, unseen data
  - **Unsupervised learning:** ML algorithms learn **from unlabeled data** to identify patterns, relationships, or structures in the data
  - **Reinforcement learning:** ML algorithms learn by **interacting with an environment** and receiving feedback as rewards or penalties
  - **Transfer learning:** ML algorithms focus on **storing knowledge gained** while solving one problem and applying it to a different but related problem


MACHINE LEARNING
CLASSICAL LEARNING
UNSUPERVISED
SUPERVISED
Clustering
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Classification
Naive Bayes
SVM
K-NN
Decision Trees
Logistic Regression
Pattern search
Euclat
Apriori
FP-Growth
Regression
Linear Regression
Polynomial Regression
Ridge/Lasso Regression
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
REINFORCEMENT LEARNING
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C
ENSEMBLE METHODS
Stacking
Bagging
Random Forest
Boosting
XGBoost
AdaBoost
LightGBM
CatBoost
NEURAL NETS AND DEEP LEARNING
Convolutional Neural Networks (CNN)
DCNN
Perceptrons (MLP)
Recurrent Neural Networks (RNN)
LSM
LSTM
GRU
Autoencoders
seq2seq
Generative Adversarial Networks (GAN)

Multimedia: music I

- Finding the right attributes to describe different pieces of music
- Data preparation and pre-processing
- The need for special tools for data preparation

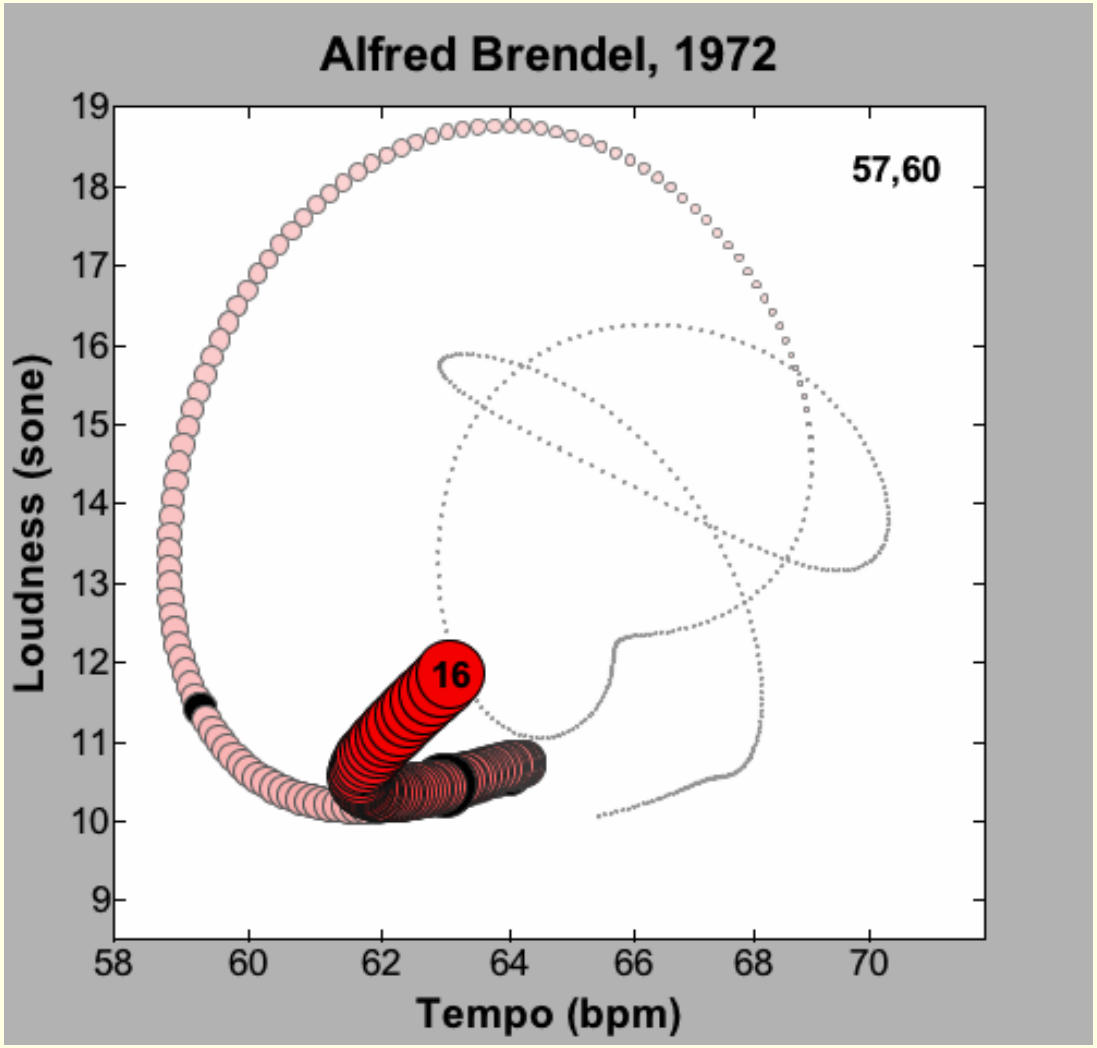
Multimedia: music II

- Mozart - Piano Sonata 13 - KV 333:
  - 1: https://www.youtube.com/watch?v=h-CM7cNb_Dk
  - 2a: https://www.youtube.com/watch?v=9lLaHB7eGKc
  - 2b: https://www.youtube.com/watch?v=pq-f6qVeZwM
  - 2c: https://www.youtube.com/watch?v=h_dOL3XD2DE
- Your turn - Q12: What differentiates the two piano performances?

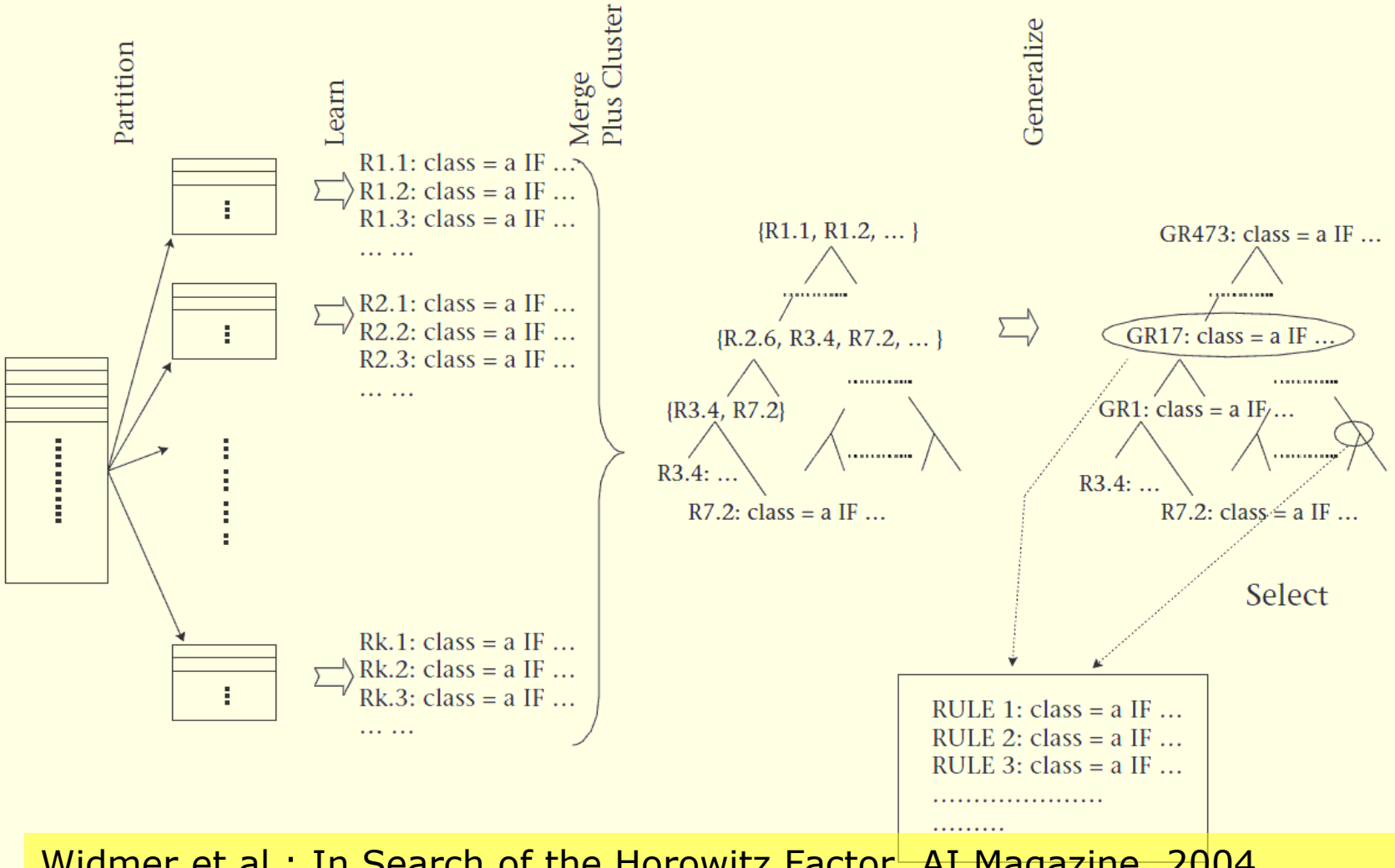
Multimedia: music III

- Music performance visualization (Widmer et al., 2004)
- Different players have different ways of building expression in music
- Subtle changes in beat level tempo versus loudness for each note played are measured
- Visual representation in tempo-loudness space as a trajectory is called performance worm

Multimedia: music IV

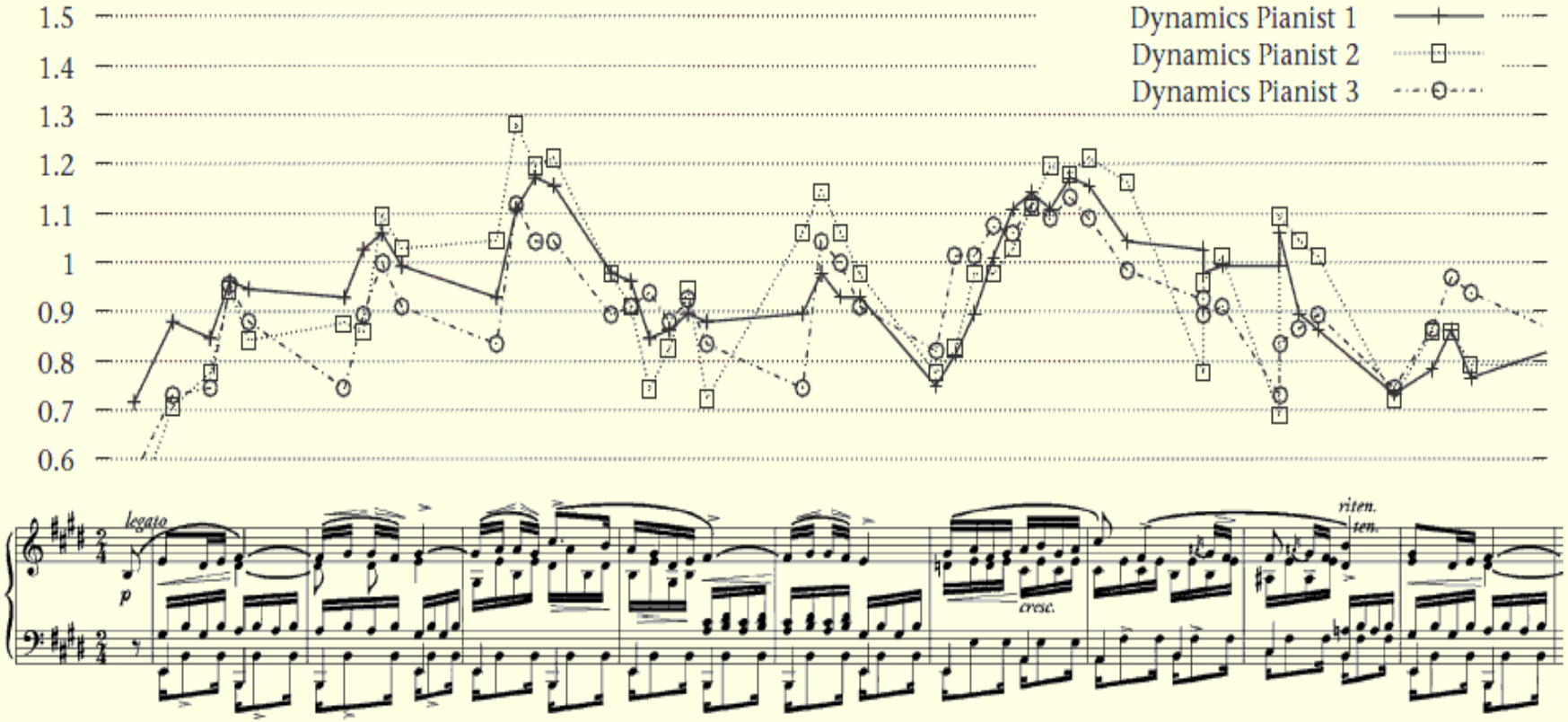


Multimedia: music V



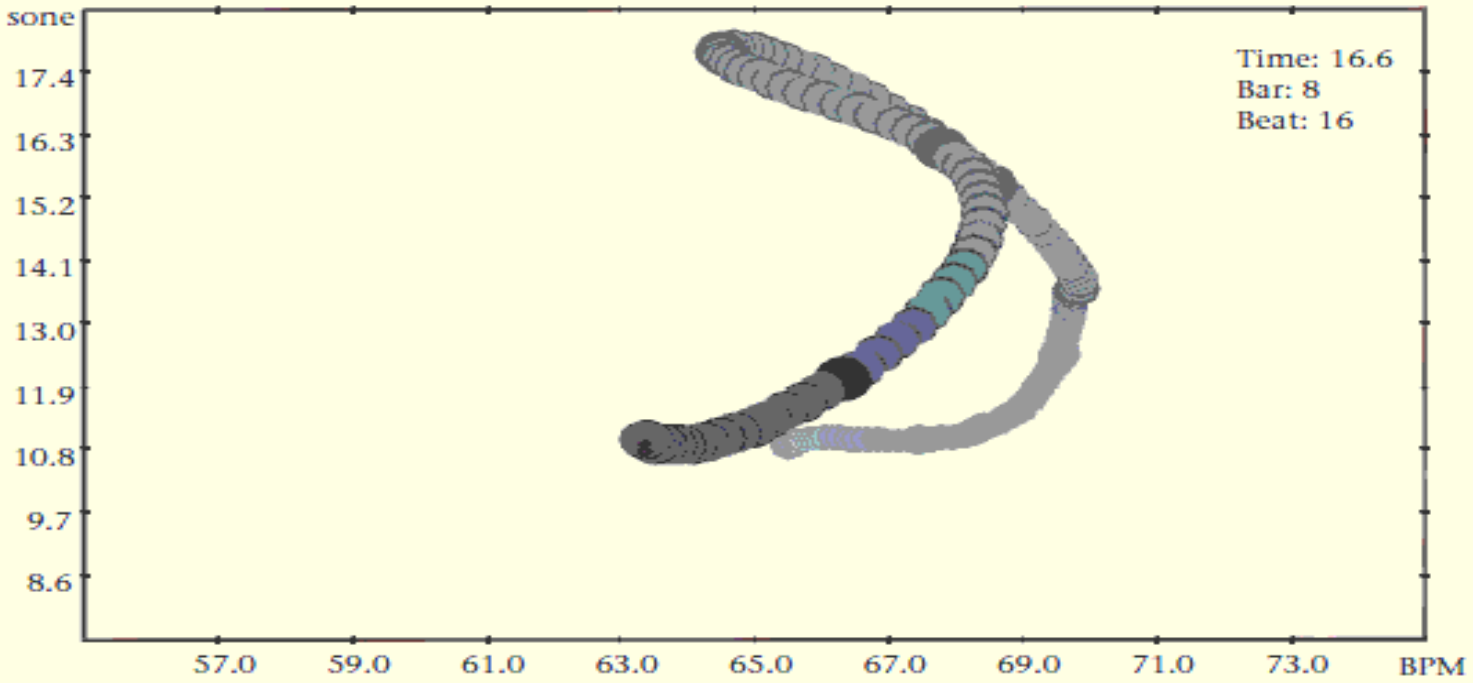
Widmer et al.: In Search of the Horowitz Factor, AI Magazine, 2004

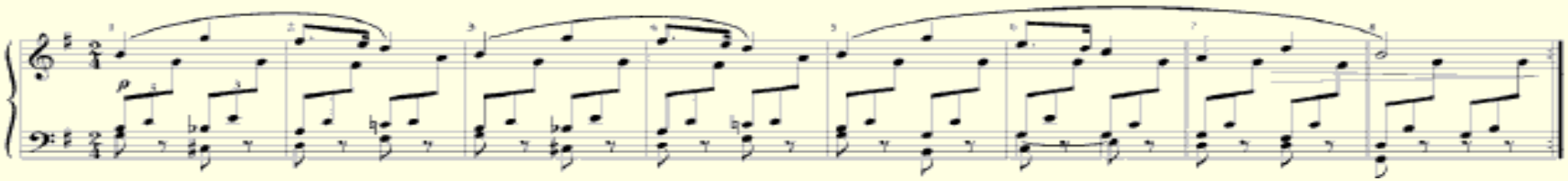
Multimedia: music VI



Dynamics curves comparison

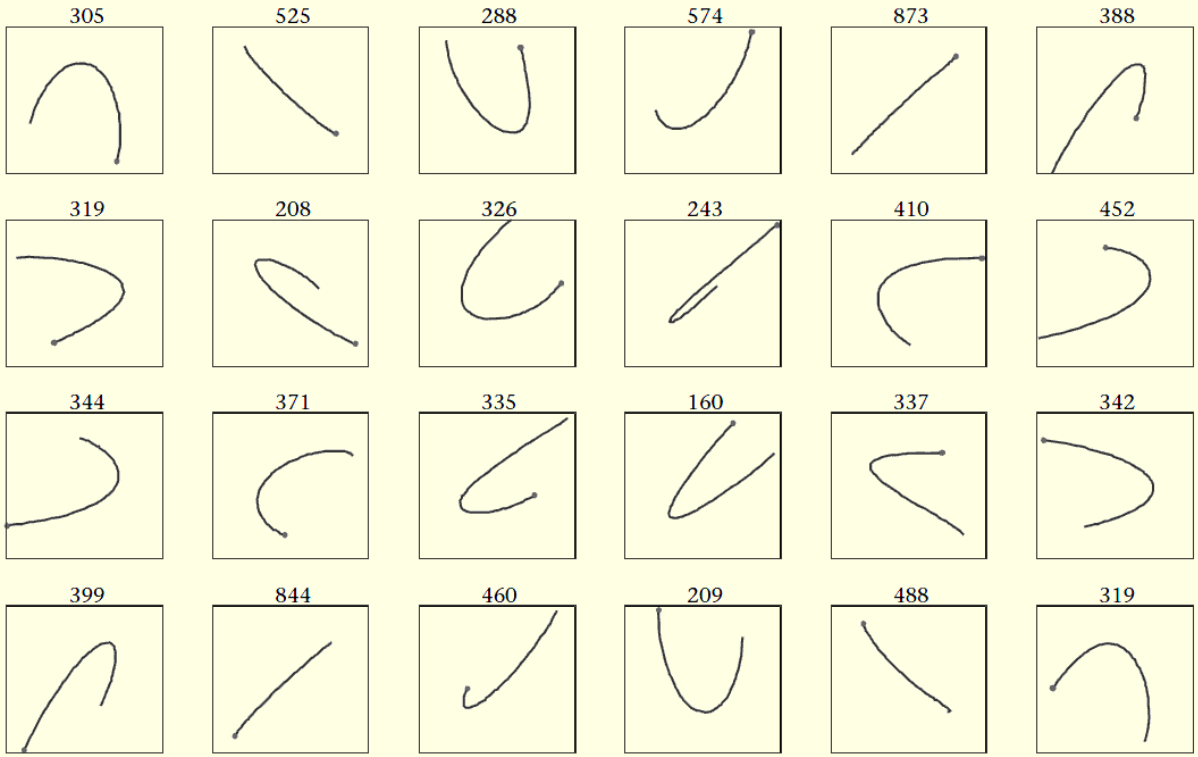
Multimedia: music VII





- Tempo / loudness performance curve

Multimedia: music VIII



- Mozart performance "alphabet"
- http://www.cp.jku.at/projects/yqx/

- Your turn - Q13: Key success factor?

Approaches to data gathering

- Problem definition
  - Class variable (dependent variable)
  - Attributes and values (independent variables)
- (1) Manual table construction
- (2) Generation from existing database

- (3) Combination of (1) and (2)

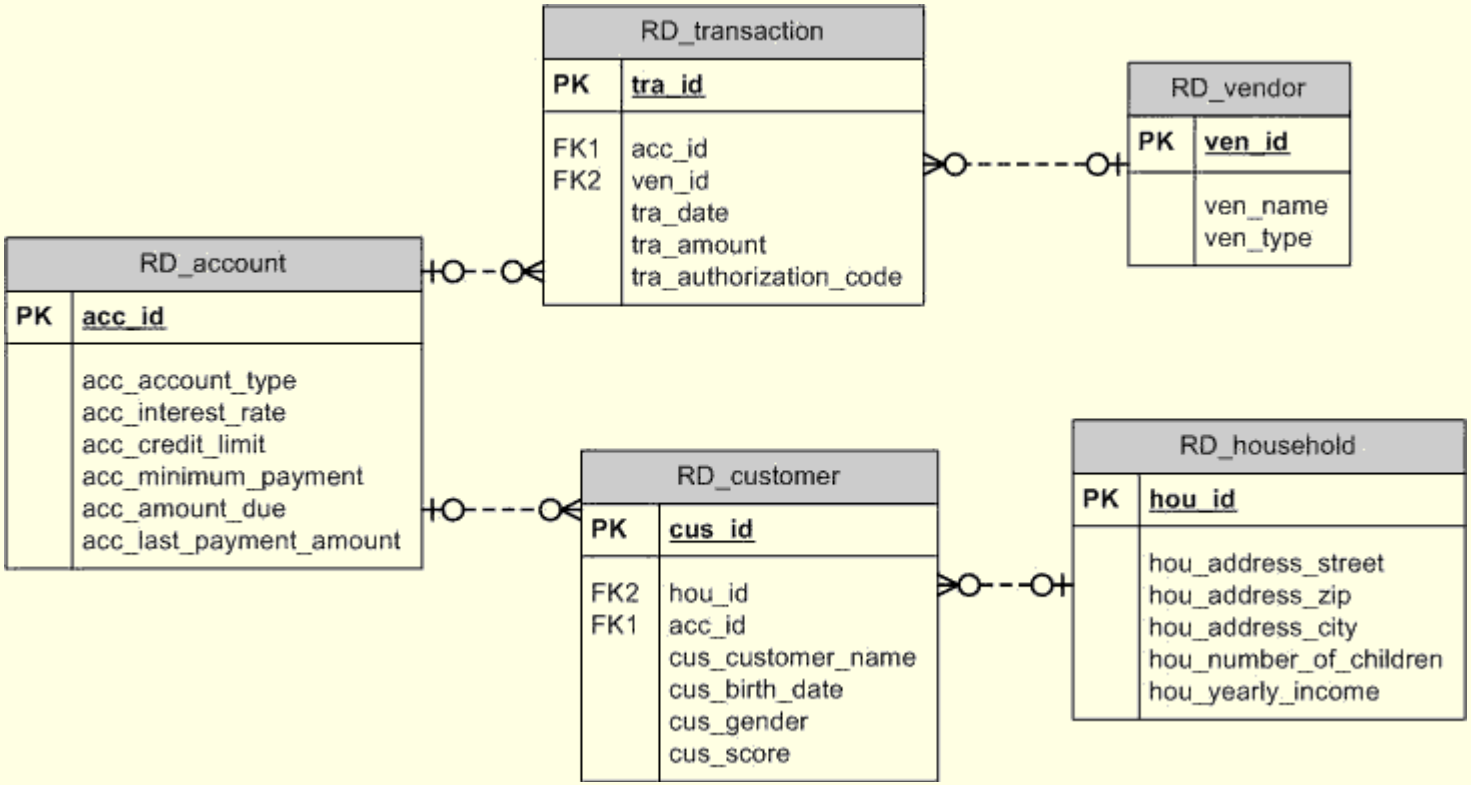
Models of the real world

- Real world: objects (entities), properties (attributes), relations
- Models: abstractions from the real world

- Data model: ERD diagram
- Conceptual data model – semantic view
- Logical data model – business view
- Physical data model – performance view

ER diagram

- Entities, attributes, relations



Entity = Table

- Rows: instances
- Columns: attributes, class

Attr1	Attr2	Attr3	…	AttrN	Class
v11	v12	v13	…	v1N	c1
v21	v22	v23	…	v2N	c2
v31	v32	v33	,,,	v3N	c3
…	…	…	…	…	…
vM1	vM2	vM3	…	vMN	cM

SQL

- Queries for ERD model
- Operations:
  - Data exploration
  - Data transformation
- Examples in MySQL and R

Data exploration

- What are the values in each column?
  - Columns with (almost) only one value
  - Columns with unique values
- What unexpected values are in each column?
- Are there any data format irregularities, such as time stamps missing hours and minutes or names being both upper- and lowercase?
- What relationships are there between columns?
- What are frequencies of values in columns and do these frequencies make sense?

Summary for one column I

- The number of distinct values in the column
- Minimum and maximum values
- An example of the most common value (called the mode in statistics)
- An example of the least common value (called the antimode)
- Frequency of the minimum and maximum values

Summary for one column II

- Frequency of the mode and antimode
- Number of values that occur only one time
- Number of modes (because the most common value is not necessarily unique)
- Number of antimodes

Basic statistical concepts

- The Null Hypothesis
- Confidence (versus probability)
- Normal Distribution

Data preparation I

- Dataflow operations:
  - Read
  - Output
  - Select (chooses the columns for the output; each column is either equal to input column or a function of some input columns)
  - Filter (removes rows based on the values in one or more columns; each input row either is or is not in the output table)
  - Append (appends columns to an existing table)

Data preparation II

- Dataflow operations:
  - Union (appends equally headed rows to an existing result)
  - Aggregate (groups columns together based on a common key; all the responding rows are summarized in a single output row)
  - Lookup (joining small tables)
  - Join (matches rows in two tables; for every matching pair a new row is created in the output)
  - Sort

Data types

- Numeric
  - Categorical
  - Rank
  - Interval
  - True numerics
- Date and time
- String

- Your turn - Q14: Tools?

Derived variables I

- During preprocessing or processing?
- Often contain very similar information
- Examples:
  - weight / height ^2
  - debt / earnings
  - population / area
  - credit limit – balance
- Difference, ratio?
- Summarizations
- Extracting features from single columns
  - Date, time

- Your turn - Q15: The role of derived variables?

Derived variables II

- Example with adding features to Neural Networks (TensorFlow):
  - https://developers.google.com/machine-learning/crash-course/introduction-to-neural-networks/playground-exercises
- Even with Neural Nets, some amount of feature engineering is often needed to achieve best performance
- The role of LLM (e.g. ChatGPT) in data preprocessing

Data sampling

- Selecting the right level of granularity
- Depends on the data types
  - Categorical
  - Rank
  - Interval
  - True numerics
- Sometimes we have to take what we have and do the best with it

- Your turn - Q16: Why is data sampling important?

Data variability

- How much data is enough?
  - How many rows?
  - How many columns?
  - How many bytes?
  - How much history?
- Selecting the right sample size:
  - https://www.surveysystem.com/sscalc.htm
- Random sampling
- Beware of biased samples

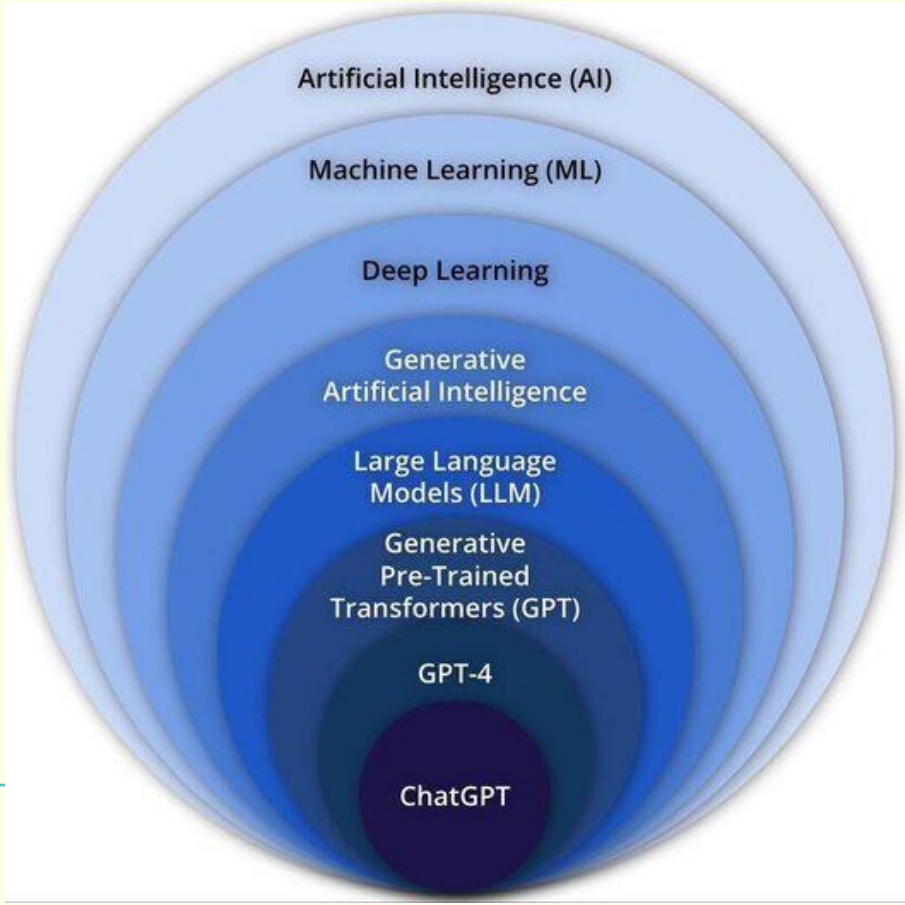
Confidence vs. probability

- Statistical measures
- Stratified sampling techniques
- Example: variables gender and age in questionnaires
- Handling outliers
  - Do nothing
  - Filter the rows
  - Ignore the column
  - Replace the outlying values
  - Bin values into ranges
- Handling missing data

- Your turn - Q17: What is p-value?

Data exploration with ChatGPT

- Queries and prompts for data tables
- Operations:
  - Data exploration
  - Data transformation
- Examples in ChatGPT



Data exploration with R

- From „kaggle" challenge ASHRAE – Great Energy Predictor III: https://www.kaggle.com/c/ashrae-energy-prediction/overview
- Datasets:
  - Train
  - Building
  - Weather
- Task: construct a model to predict energy consumption
- Assignment: Preprocess the ASHRAE data

Overview

- DM algorithms want data in table format
- Data comes from warehouses , data marts, OLAP systems, external sources, etc.
- Data has to be transformed into a DM format: aggregations, joins
- Useful column types: categories, ranks, intervals, true numerics
- The art of DM: creating derived variables