NUS | Computing
National University of Singapore

CS5228: Knowledge Discovery and Data Mining

Lecture 1 — Introduction & Overview

Outline

- **Course Logistics**
- Overview
  - What is Knowledge Discovery / Data Mining?
  - Common Data Mining tasks
  - Types of data & data representations
- Data preparation
  - Data quality
  - Exploratory Data Analysis (EDA)
  - Data preprocessing
- Summary

Course Logistics

- Lectures & Tutorials
  - Friday, LT17: 6.30-8.30 pm / 8.30-9.30 pm
  - Physical classes (all recorded)
  - Announcements & materials on Canvas

- Where to ask questions
  - Canvas discussion (you are also strongly encouraged to answer questions!)
  - Email to teaching team (for private concerns or sensitive question, e.g., about an assignment)

Teaching Team

Lecturer

Christian von der Weth
chris@comp.nus.edu.sg

TA

Xu Zhengtao
xuzhengtao@u.nus.edu

TA

Sahajpreet Singh
sahajpreet.singh@u.nus.edu

TA

Chai Zenghao
zenghaochai@u.nus.edu

TA

Lu Jiahao
jiahao.lu@u.nus.edu

Assessments

- 3 assignments (11% each)
  - Programming tasks + theoretical questions (Python)
  - Discussions allowed, but code and answers must be submitted individually
- Quiz in the last lecture (11% each)
  - MCQ/MRQ Canvas quiz
  - Open-book but no Internet (screen recording will be required)
- Midterm (22%)
  - MCQ/MRQ + essay questions using Examplify
  - Open-book but blocked Internet
- Project (34%)
  - Group project (~4 students per group, more details after enrollment is complete)
  - Kaggle InClass Competition

Lesson Plan (tentative deadlines; check Canvas!)

Release dates for assignments (Fri);
submission deadline 3 weeks later (Thu)

Week	Date	Topics	Important Dates
1	Aug 15	Introduction	
2	Aug 23	Clustering I	
3	Aug 29	Clustering II	A1
4	Sep 05	Association Rule Mining	
5	Sep 12	Regression & Classification I	
6	Sep 19	Regression & Classification II	A2
Recess	Sep 26	No class	
7	Oct 03	**Midterm Exam** (pending confirmation)	Midterm (Weeks 1-6)
8	Oct 10	Regression & Classification III	
9	Oct 17	Recommender Systems	
10	Oct 24	Graph Mining	A4
11	Oct 31	Dimensionality Reduction	
12	Nov 07	Data Stream Mining	
13	Nov 14	Review & Outlook + **Quiz**	Quiz 2 (Weeks 7-12)

Course Policies

- Zero-Tolerance for Plagiarism
  - Students will be reported to University for disciplinary action for plagiarism/cheating offence
  - Offenders will receive F grade for the module (for any assessment with 10%+ weight!!!)
  - Assignments: discussion allowed but each students must submit their individual solutions

- Resources
  - https://www.comp.nus.edu.sg/cug/plagiarism/

Course Policies

- AI use in class
  - Generally allowed for ideation, brainstorming, self-learning, improve writing
  - Take-home assignments: AI tools permitted but need to be acknowledged
  - Exams (midterm, quiz): AI tools not permitted incl. locally installed tools (e.g., open LLMs)

- Resources
  - https://libguides.nus.edu.sg/new2nus/acadintegrity
    (see the "Guidelines on the Use of AI Tools For Academic Work" tab)
  - https://myportal.nus.edu.sg/studentportal/student-discipline/all/docs/NUS-Plagiarism-Policy.pdf

Course Policies

- Right Infringements on NUS Course Materials

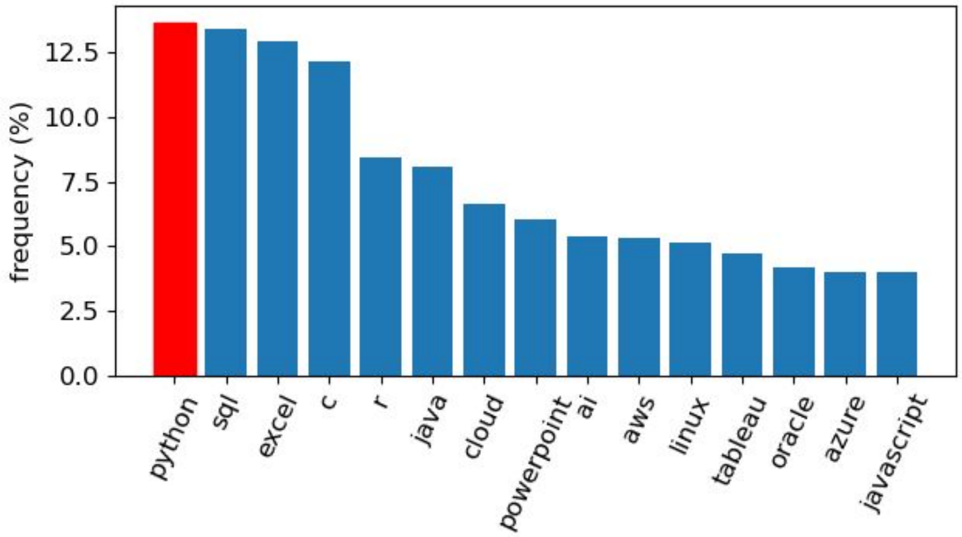
All course participants (including permitted guest students) who have access to the course materials on LumiNUS or any approved platforms by NUS for delivery of NUS modules are not allowed to re-distribute the contents in any forms to third parties without the explicit consent from the module instructors or authorized NUS officials.

What You Need

- Programming environment: Python + Jupyter
  - All implementation tasks will be in Python
  - Assignments will include Jupyter notebooks
  - Supplementary Jupyter notebooks for hands-on practice
- Common packages for data science
  - NumPy
  - pandas
  - NetworkX
  - scikit-learn

Why Python?

- Analysis of job descriptions
  - 15k+ job offers from JobStreet
    (data analyst, data engineer, data scientist)
  - Quick-&-dirty keyword extraction
  - ...but check for yourself! :)



Learning Outcomes

- Fundamental knowledge about concepts & algorithms in data mining
  - Nature of data: data representations, data and attribute types
  - Common data mining tasks and important algorithms (with their strengths and weaknesses)
  - Problems, risks & ethical issues of "unrestrained" data mining

A1-3
midterm
quiz

- Perform data mining tasks for new applications in practice
  - Given a dataset and task, select appropriate techniques to solve the task
  - Justify design and implementation decisions
  - Interpret results and assess limitations

project

References

- Textbooks (useful but not required)
  - J. Leskovec, A. Rajaraman, J. Ullman: *Mining Massive Datasets*
    (online version available at: http://www.mmds.org/)
  - P. Tan, M. Steinbach, A. Karpatne, V. Kumar: *Introduction to Data Mining*
  - More in Canvas Readings

- Jupyter Notebooks on Github (recommended but not required)
  - CS5228 repository
  - SELENE repository

Outline

- Course Logistics
- **Overview**
  - What is Knowledge Discovery / Data Mining?
  - Common Data Mining tasks
  - Types of data & data representations
- Data preparation
  - Data quality
  - Exploratory Data Analysis (EDA)
  - Data preprocessing
- Summary

What is Knowledge Discovery & Data Mining

*"The **non-trivial** extraction of implicit, previously unknown, and potentially **useful** information from data."*

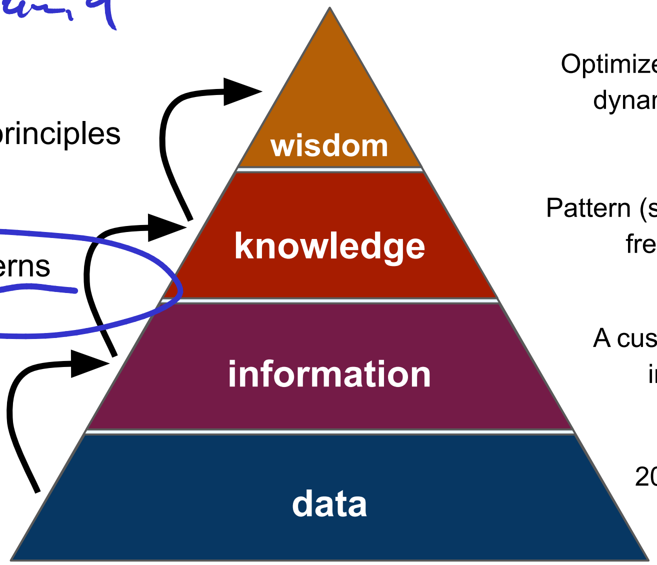
(Frawley, Piatetsky-Shapiro, Matheus; 1991)

DIKW pyramid

Understanding principles

CS5228 Understanding patterns

Understanding relationships



Optimize product order and item placements dynamic pricing, bundled promotions, etc.

Pattern (shopping behavior): many customers frequently buy milk and cereal together

A customer bought bread, cereal and milk in FairPrice NUH on August 15, 2023

2023-08-15T15:05:30Z

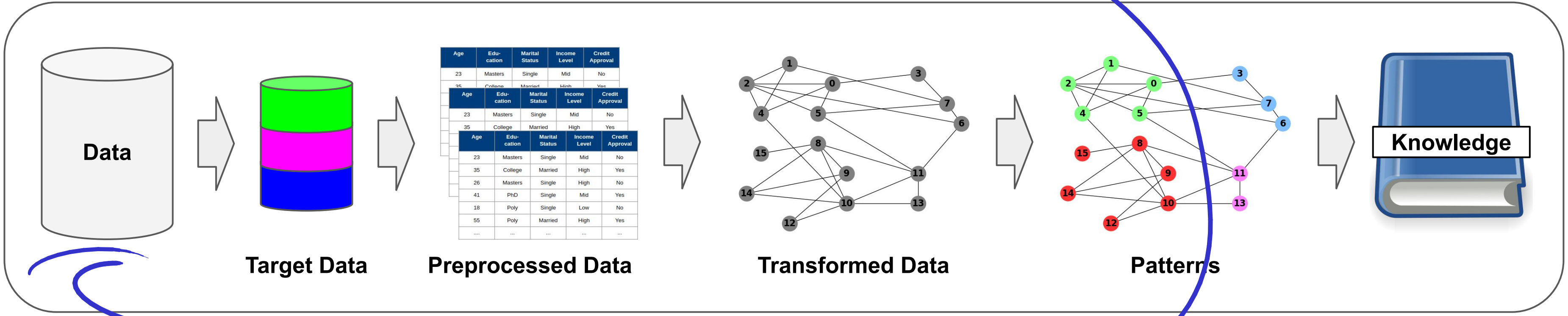
1.2933, 103.7844

From Data to Knowledge

Data Selection
- Identify relevant data to solve a given task

Data Transformation
- Convert data into suitable representation

Postprocessing
- Visualization
- Interpretation
- Understanding
- ...



Data Preprocessing
- Handling missing data
- Duplicate elimination
- Feature selection
- Normalization
- ...

Data Mining
- Clustering
- Classification
- Regression
- Associations
- Correlations
- ...

Adapted from: [From Data Mining to Knowledge Discovery in Databases](#) (Fayyad et al., 1996)

What is NOT Knowledge Discovery & Data Mining?

- Trivial extraction of information/patterns from data
  - Looking up a phone number in phone directory
  - Dividing students based on their degree course
  - Calculating the total sales of a company
- Data analysis not yielding patterns (i.e., new information)
  - Monitoring a patient's heart rate for abnormalities
  - Querying a Web search engine

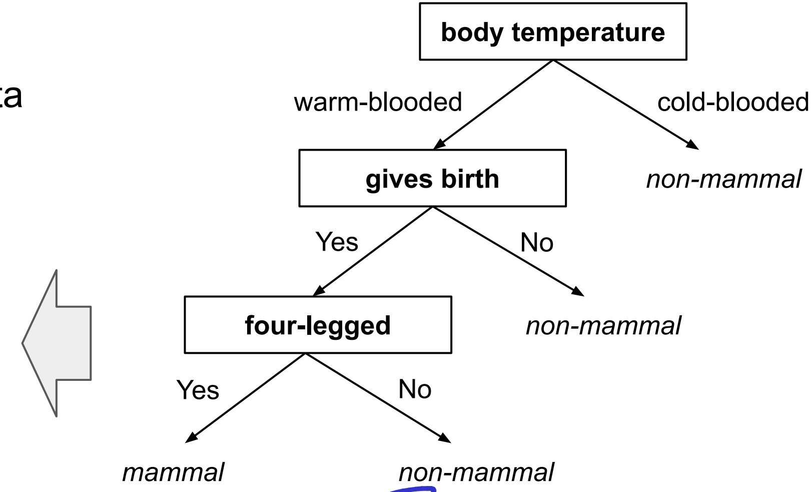
What Makes a Pattern Useful or Meaningful?

"If you torture the data long enough, it will confess to anything"

(Ronald Coase; 1981 — slightly paraphrased)

- Main goal: **Generalizability**
  - Patterns should remain accurate over unseen data
  - Common causes: small and/or biased data

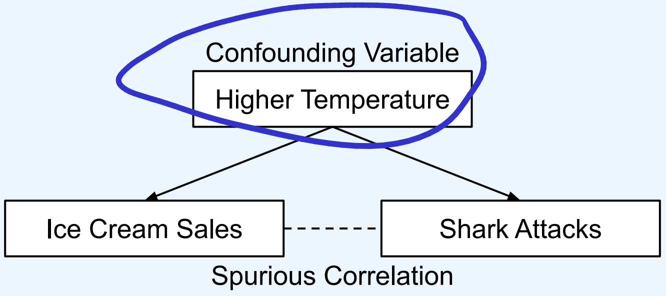
But what about humans and platypuses, etc.?



There is Always Some Pattern in Your Data (even in random data)

- Bizarre and Surprising Insights
  - *"Female-named hurricanes kill more people than male hurricanes."*
  - *"Users of Chrome and Firefox browsers make better employees."*
  - *"Shark attacks increase when ice cream sales increase"*
  - *"Music taste predicts political affiliation."*
  - *"A job promotion can lead to quitting."*
  - *"Vegetarians miss fewer flights."*
  - *"Smart people like curly fries."*
  - *"Higher status, less polite."*

Important: Patterns indicate correlations, but correlation does not imply causation!



Source: Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die (Eric Siegel, 2013)

Spotting "Shady" Patterns — Reality Check

- What is the (perceived) difference between the 2 statements below?
  - In the context of identifying and/or assessing patterns

"The higher the sales of ice cream, the higher the number of shark attacks."

vs.

"The higher the concentration of anti-mullerian hormone, the lower the concentration of follicle-stimulating hormone."

Note: "This doesn't make sense!" is rarely a good argument.

Data Mining Gone Wrong

"Your scientists were so preoccupied with whether they could, they didn't stop to think if they should."

(Ian Malcolm; Jurassic Park, 1991)

Malaysian Bar troubled over judges using AI for sentencing

Applicant pre-selection by algorithm: How to convince an AI of yourself

Algorithms are deciding who gets the first vaccines. Should we trust them?

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Millions of black people affected by racial bias in health-care algorithms

The computers rejecting your job application

Quick Quiz

What is (arguably) **NOT**
a "proper" Data Mining task?

(given a dataset of supermarket transactions)

A Finding the largest sets of products most frequently bought together

B Finding groups of similar users based on the buying behavior

C (✓) Finding all purchases of a bundled promotion (i.e., multiple items)

D ✓ Finding the products most frequently bought on weekends after 6pm

Outline

- Course Logistics
- **Overview**
  - What is Knowledge Discovery / Data Mining?
  - **Common Data Mining tasks**
  - Types of data & data representations
- Data preparation
  - Data quality
  - Exploratory Data Analysis (EDA)
  - Data preprocessing
- Summary

Methods — Association Rules

- Input: transactional data
  - Transaction: data record with set of items
  - Set of items are from a fixed collection
- Pattern: Association rules
  - Rules predicting the occurrence of items based on the occurrence of other items

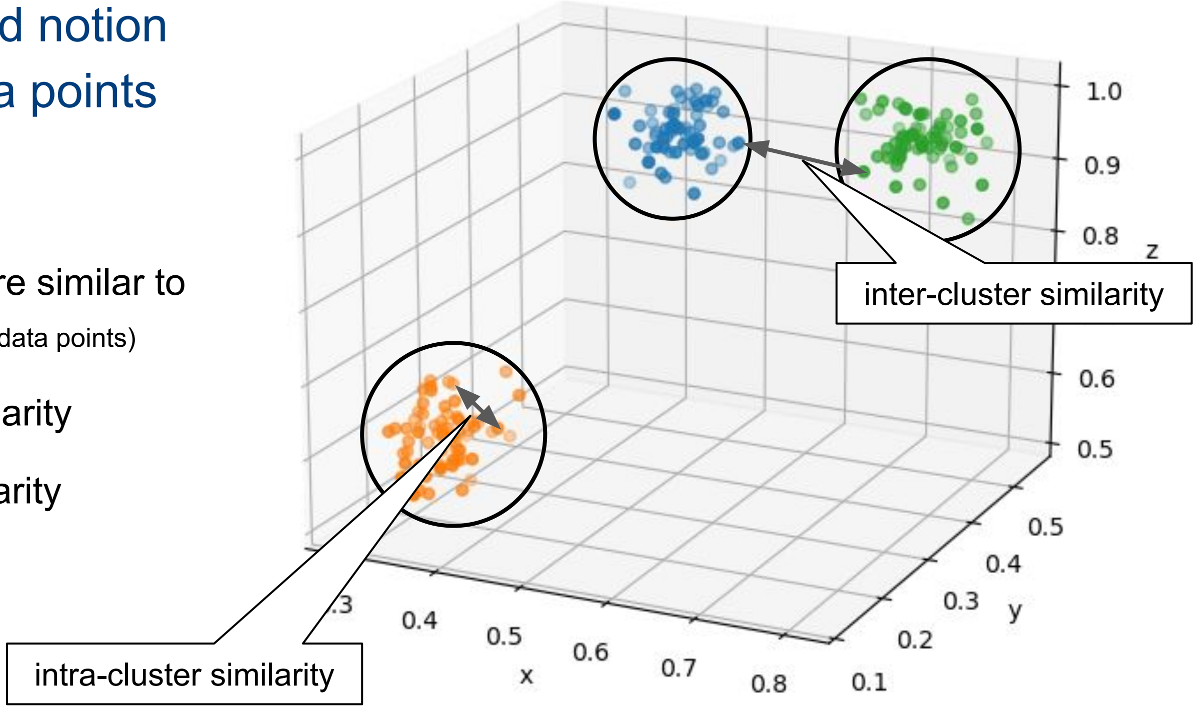
ID	Items
1	covid-19, anosmia, cough, fatigue
2	flu, anosmia, headache
3	covid-19, anosmia, headache, fatigue, fever
4	covid-19, flu, anosmia, fatigue
5	flu, depression, fatigue, fever, headache
...	

{anosmia, fatigue} ➜ {covid-19}

Methods — Clustering

- Input: Data & well-defined notion of similarity between data points

- Pattern: Clusters
  - Groups of data points that are similar to each other (compared to the other data points)
  - Maximize **intra-cluster** similarity
  - Minimize **inter-cluster** similarity

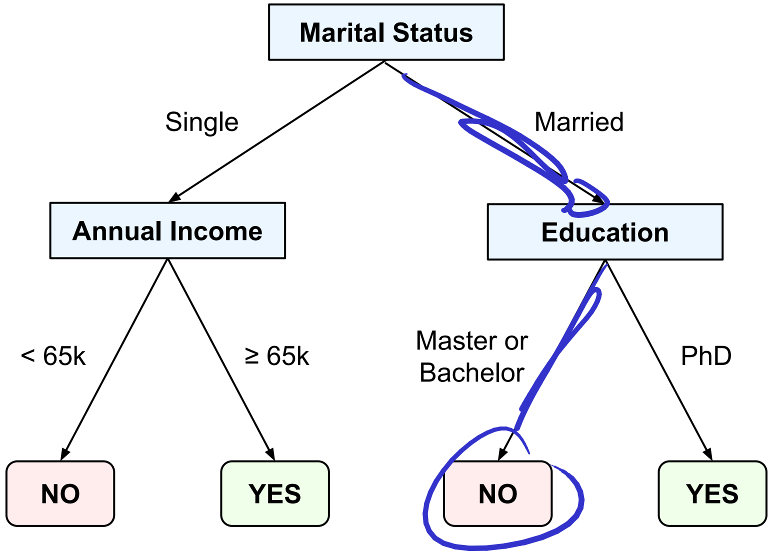


Methods — Classification

- Input: Dataset with multiple attributes

- Pattern: **Categorical** value of an attribute as function of other attribute values
  - K-Nearest Neighbor, Decision Trees, Linear Classification, etc.

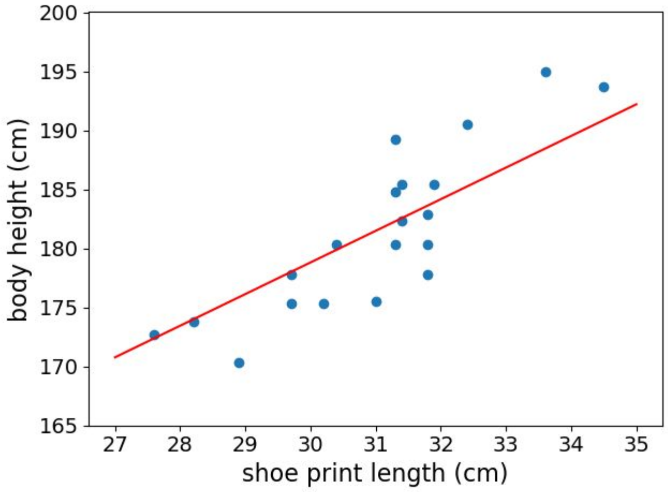
Age	Edu-cation	Marital Status	Annual Income	Credit Default
23	Masters	Single	75k	Yes
35	Bachelor	Married	50k	No
26	Masters	Single	70k	Yes
41	PhD	Single	95k	Yes
18	Bachelor	Single	40k	No
55	Master	Married	85k	No
30	Bachelor	Single	60k	No
35	PhD	Married	60k	Yes
28	PhD	Married	65k	Yes



Methods — Regression

- Input: Dataset with multiple attributes

- Pattern: **Numerical** value of an attribute as function of other attribute values
  - K-Nearest Neighbor, Regression Trees, Linear Regression, etc.



Question: *"What is the expected height of a person that leaves a shoe print of size 32.2cm?"*

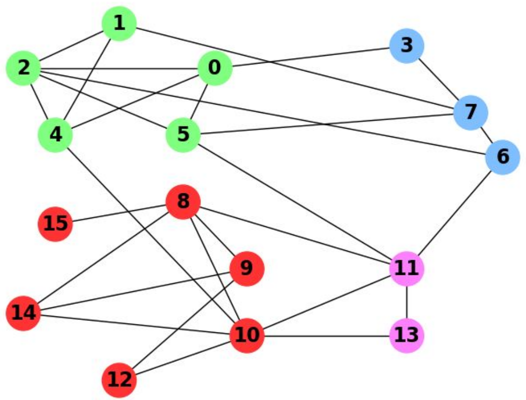
Answer: ?

Methods — Graph Mining

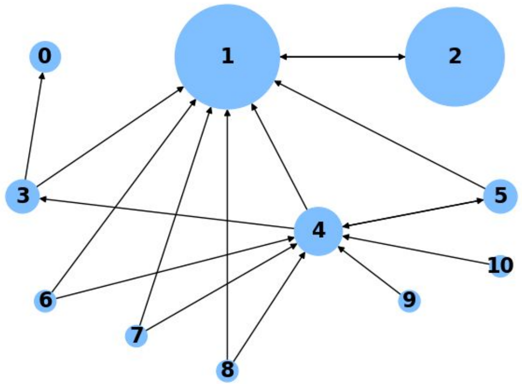
- Input: *G = (V, E)*
  - Set of vertices (or nodes) V (data points)
  - Set of edges E (relationship between data points)

- Patterns based on graph structure, e.g.:

Finding communities of nodes



Finding "important" nodes



Methods — Recommender Systems

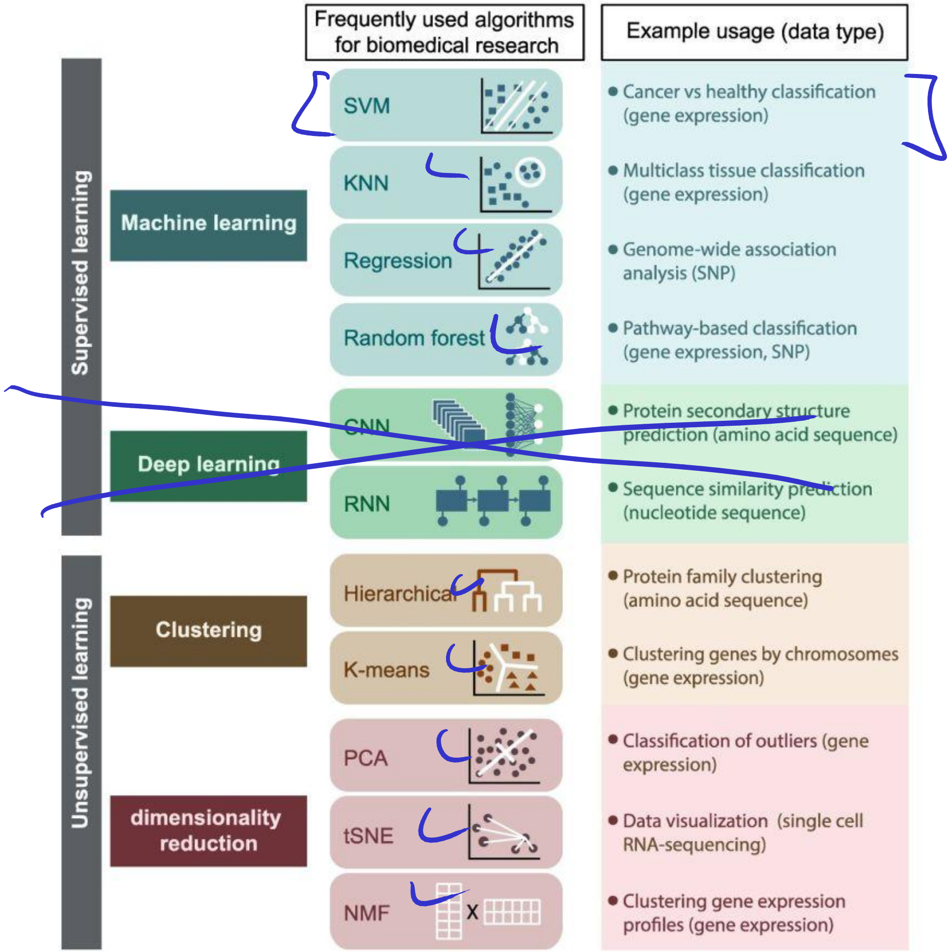
- Input: User-rated items
  (e.g., movies rated by viewers)
  - How would Bob rate the movie "Heat"?
  - Should "Heat" be recommended to Bob?

	Clueless	Heat	Jarhead	Big	Rocky
Alice	2	4	5	0	1
Bob	1	???	4	0	2
Claire	1	0	4	3	0
Dave	5	1	2	0	5
Erin	1	5	3	0	3

- Patterns based on similarities to predict missing values
  - Exploiting features of items
  - Exploiting similarities between users or items

Data Mining in Practice

- Example: Biomedical Research
  - Set of important data mining algorithms
  - Relevant for many other fields
  - Many covered here in CS5228
    (main exception: no deep learning)

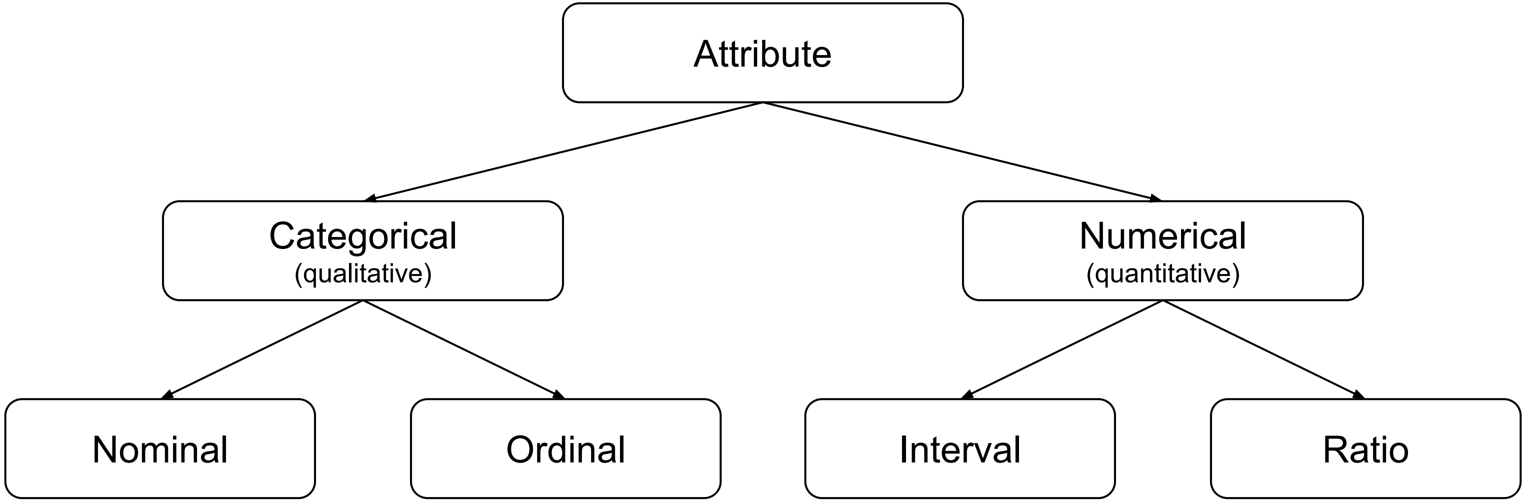


Source: Some LinkedIn post

Outline

- Course Logistics
- Overview
  - What is Knowledge Discovery / Data Mining?
  - Common Data Mining tasks
  - **Types of data & data representations**
- Data preparation
  - Data quality
  - Exploratory Data Analysis (EDA)
  - Data preprocessing
- Summary

Types of Attributes



Nominal

- Values are only labels
- Operations:
  =, ≠
- Examples: sex (m/f), eye color, zip code

Ordinal

- Values are labels with a meaningful order
- Operations:
  =, ≠, <, >
- Examples: (street numbers), education level

Interval

- Values are measurements with a meaningful distance
- Operations:
  =, ≠, <, >, +, -
- Examples: body temperature in °C, calendar dates

ilk

Ratio

- Values are measurements with a meaningful ratio
- Operations:
  =, ≠, <, >, +, -, *, /
- Examples: age, weight, income, blood pressure

Types of Data

CS5228

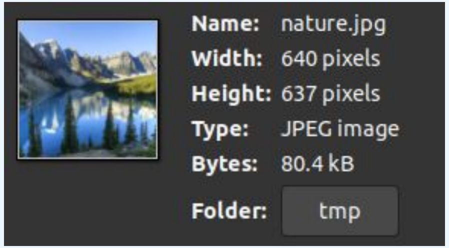
(Well-)Structured Data

- Highly organized: adheres to predefined data model
- Each object has the same fixed set of attributes
- Easy to search, aggregate, manipulate, analyze data
- Examples: Relational databases, spreadsheets

Age	Edu-cation	Marital Status	Income Level	Credit Approval
23	Masters	Single	Mid	No
35	College	Married	High	Yes
26	Masters	Single	High	No
41	PhD	Single	Mid	Yes
18	Poly	Single	Low	No
55	Poly	Married	High	Yes
....	...	...	...	...

Semi-Structured Data

- No rigid data model: mix of structured & unstructured data
- Data exchange formats: XML, JSON, CSV
- Tagged unstructured data (e.g., photo + date/time, location, exposure, resolution, flash, etc)



Unstructured Data

- No fixed data model
- Requires more advanced data analysis techniques
- Examples: images, videos, audio, text, social media

0.0	0.0	5.0	13.0	9.0	1.0	0.0	0.0
0.0	0.0	13.0	15.0	10.0	15.0	5.0	0.0
0.0	3.0	15.0	2.0	0.0	11.0	8.0	0.0
0.0	4.0	12.0	0.0	0.0	8.0	8.0	0.0
0.0	5.0	8.0	0.0	0.0	9.0	8.0	0.0
0.0	4.0	11.0	0.0	1.0	12.0	7.0	0.0
0.0	2.0	14.0	5.0	10.0	12.0	0.0	0.0
0.0	0.0	6.0	13.0	10.0	0.0	0.0	0.0

Types of Data Representations — Record Data

Data matrix: collection records; each record consisting of a fixed set of attributes

Age	Edu-cation	Marital Status	Annual Income	Credit Approval
23	Masters	Single	75k	Yes
35	Bachelor	Married	50k	No
26	Masters	Single	70k	Yes
41	PhD	Single	95k	Yes
18	Bachelor	Single	40k	No
55	Master	Married	85k	No
30	Bachelor	Single	60k	No
35	PhD	Married	60k	Yes
28	PhD	Married	65k	Yes

Transaction data: collection records; each record involves a set of items

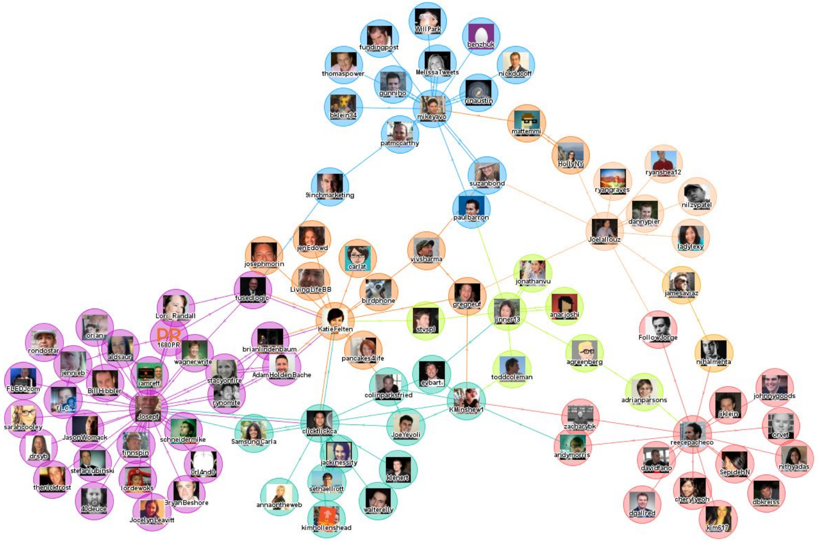
ID	Items
1	covid-19, anosmia, cough, fatigue
2	flu, anosmia, headache
3	covid-19, anosmia, headache, fatigue, fever
4	covid-19, flu, anosmia, fatigue
5	flu, depression, fatigue, fever, headache
...	

Types of Data Representations — Graph Data

Example: traffic data

Source: https://www.lta.gov.sg/

Example: social network data



Source: http://touchgraph.com/

Types of Data Representations — Ordered Data

Example: stock prices (sequence of data points)

Source: https://sg.finance.yahoo.com

Quick Quiz

What type of attribute is **Annual Income**?

ID	Age	Edu-cation	Marital Status	Annual Income	Credit Approval
101	23	Masters	Single	75k	Yes
102	35	Bachelor	Married	50k	No
103	26	Masters	Single	70k	Yes
104	41	PhD	Single	95k	Yes
105	18	Bachelor	Single	40k	No
...	...	...	...	...	...

A Nominal

B Ordinal

C Interval

D ✓ Ratio

Quick Quiz

What type of attribute is **Education**?

ID	Age	Edu-cation	Marital Status	Annual Income	Credit Approval
101	23	Masters	Single	75k	Yes
102	35	Bachelor	Married	50k	No
103	26	Masters	Single	70k	Yes
104	41	PhD	Single	95k	Yes
105	18	Bachelor	Single	40k	No
...	...	...	...	...	...

A Nominal

B ✓ Ordinal

C Interval

D Ratio

Quick Quiz

What type of attribute is **ID**?

ID	Age	Edu-cation	Marital Status	Annual Income	Credit Approval
101	23	Masters	Single	75k	Yes
102	35	Bachelor	Married	50k	No
103	26	Masters	Single	70k	Yes
104	41	PhD	Single	95k	Yes
105	18	Bachelor	Single	40k	No
...	...	...	...	...	...

A Nominal

B Ordinal

C Interval

D Ratio

Outline

- Course Logistics
- Overview
  - What is Knowledge Discovery / Data Mining?
  - Common Data Mining tasks
  - Types of data & data representations
- **Data preparation**
  - **Data quality**
  - Exploratory Data Analysis (EDA)
  - Data preprocessing
- Summary

Data Quality — Noise

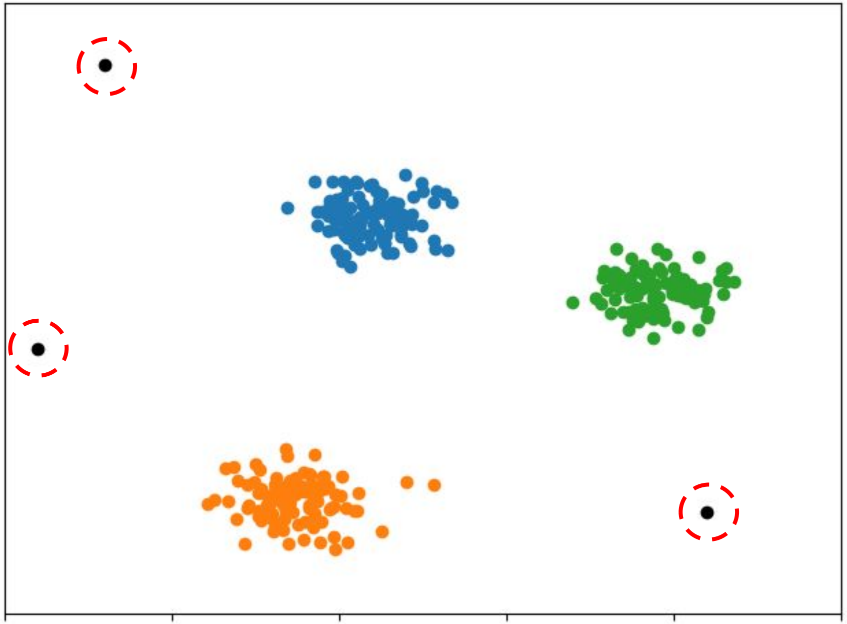
- Data = true signal + **noise**
  - Sensor readings from faulty devices
    (also intrinsic noise or external influences)
  - Errors during data entry
    (by humans or machines)
  - Errors during data transmission
  - Inconsistencies in data formats
    (e.g., iso time vs unix time, DD/MM/YYYY vs. MM/DD/YYYY)
  - Inconsistencies in conventions
    (e.g., meters vs. miles, meters vs. centimeters)

Data Quality — Outliers

- Outlier: Data point with attribute values considerably different from other points

- Case 1: Outliers are noise
  - Negatively interfere with data analysis
  - (Try to) remove outliers and/or use methods less prone to outliers

- Case 2: Outliers are targets
  (the goals is to find rare/strange/odd data points)
  - Credit card fraud
  - Intrusion detection



Data Quality — Missing Values

- Common causes
  - Attribute values not collected
    (e.g., broken sensor, person refused to report age)
  - Attributes not applicable in all cases
    (e.g., no income information for children)

- Handling missing values
  - Remove data points with missing values
  - Remove attributes with missing values
    (not all attributes are always equally important)
  - (Try to) fill in missing values
    (e.g., average temperature readings of nearby sensors)

Age	Edu-cation	Marital Status	Annual Income	Credit Default
23	Masters	Single	75k	Yes
N/A	Bachelor	Married	N/A	No
26	Masters	Single	70k	Yes
41	PhD	Single	95k	Yes
18	Bachelor	Single	40k	No
55	Master	Married	N/A	No
30	Bachelor	Single	N/A	No
35	PhD	Married	60k	Yes
N/A	PhD	Married	65k	Yes

Data Quality — Duplicates

- Duplicates: Data points referring to the same object/entity
  (e.g., two records in a database refer to the same real-world person)
  - Exact duplicates: data points have the same attribute values
  - Near duplicates: data points (slightly) differ in their attribute values
    (e.g., same person with the same phone number but in different formats)

- Task: Duplicate Elimination
  - Relatively easy for exact duplicates
  - Generally very difficult for near duplicates

Note: Duplicates are a major issue when merging data from multiples heterogeneous sources. Due to its complexity, duplicate elimination is beyond the scope of this lecture

Outline

- Course Logistics

- Overview
  - What is Knowledge Discovery / Data Mining?
  - Common Data Mining tasks
  - Types of data & data representations

- **Data preparation**
  - Data quality
  - **Exploratory Data Analysis (EDA)**
  - Data preprocessing

- Summary

Exploratory Data Analysis (EDA)

- EDA — getting to know your data (trough basic transformation and visualization)
  - Assess data quality
  - Basic sanity checks
  - Get first insights into data
  - Formulate new questions

No formal process with strict rules!

Running example:
Cardiovascular Disease Dataset
(modified to make some points)

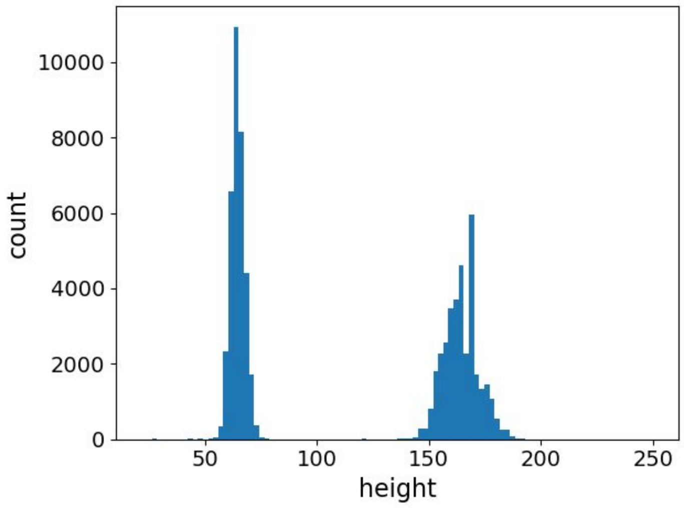
	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0

Source: [Cardiovascular Disease dataset]

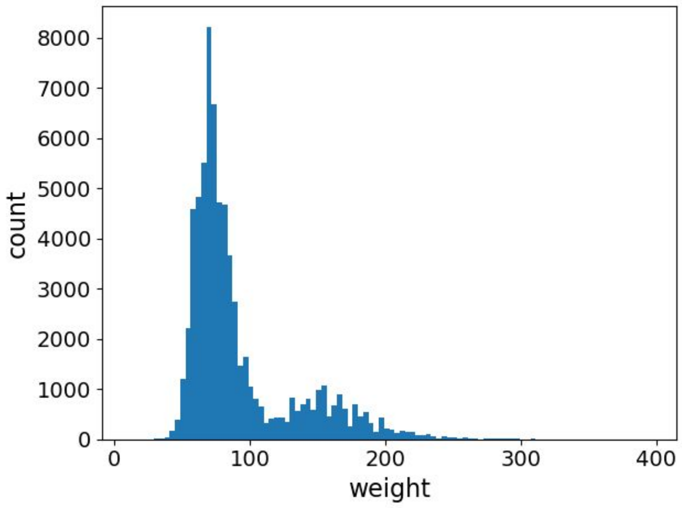
EDA — Identifying Noise

- Using histograms to inspect distribution of data values

Noise in the height values
- 50% measured in inches
- 50% measured in centimeters



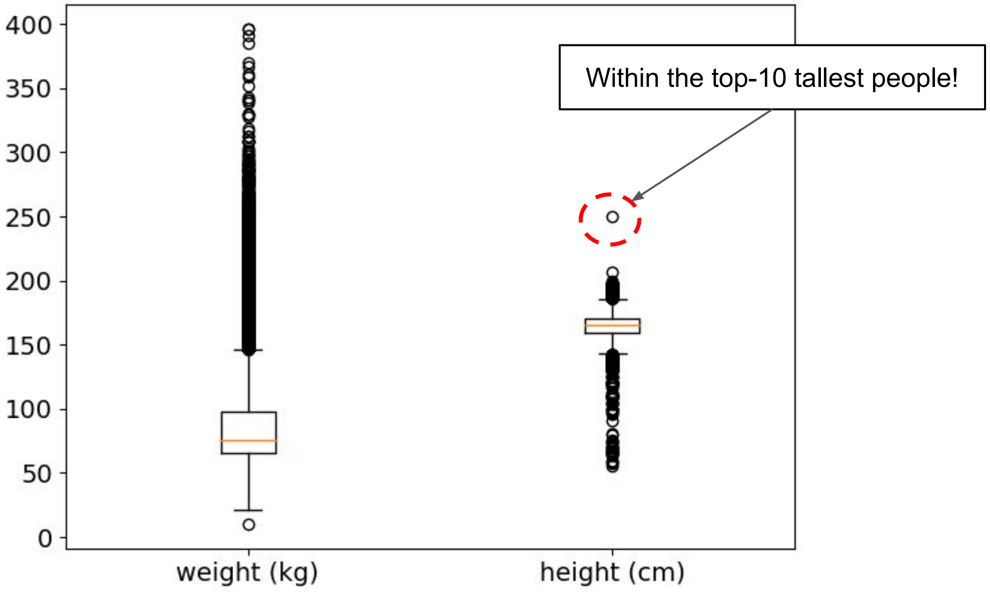
Noise in the weight values
- 80% measured in kilograms
- 20% measured in pounds



EDA — Identifying Noise / Outliers

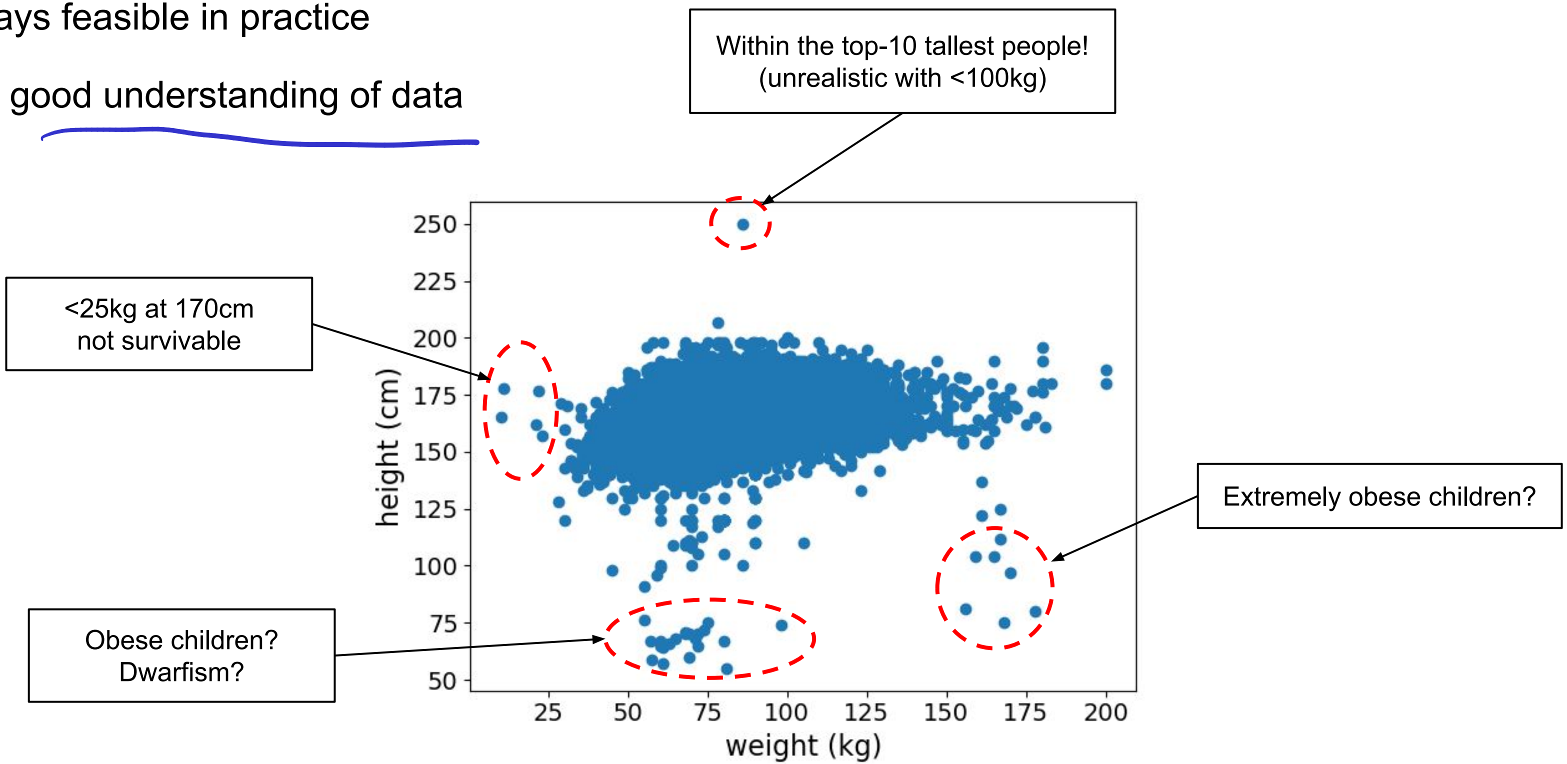
- Box plots to inspect distribution of attribute values
  - Make outliers explicit

Note: Not all outliers are "bad" or considered noise. For example, a CEO's salary is typically much higher than the one of the average employee. Whether it should be removed depends on the goal of the analysis



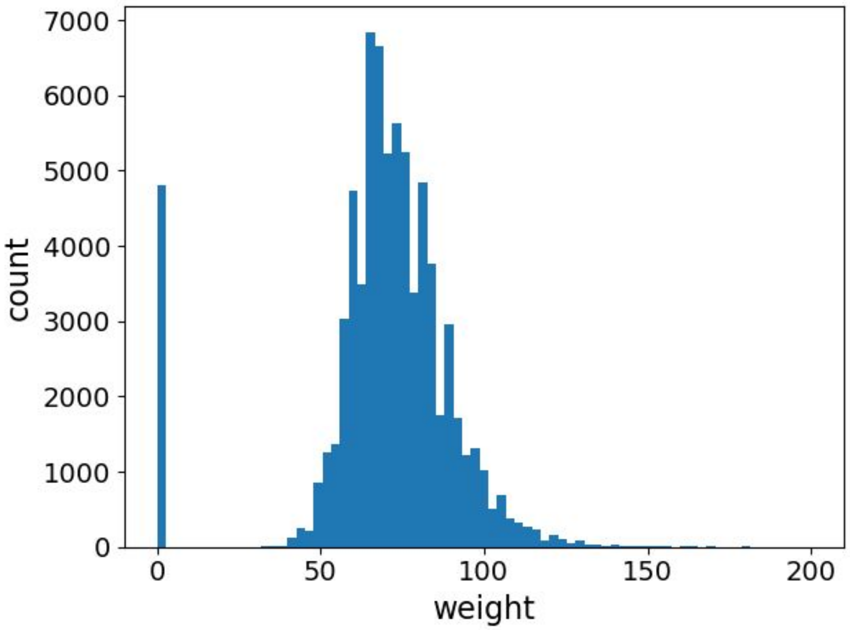
EDA — Identifying Noise / Outliers

- Using scatter plot to inspect correlations
  - Not always feasible in practice
  - Require good understanding of data



EDA — Missing Values

- Example: Default value (0) if people did not disclose weight
  - Can already negatively affect simple analysis such as calculating means/averages



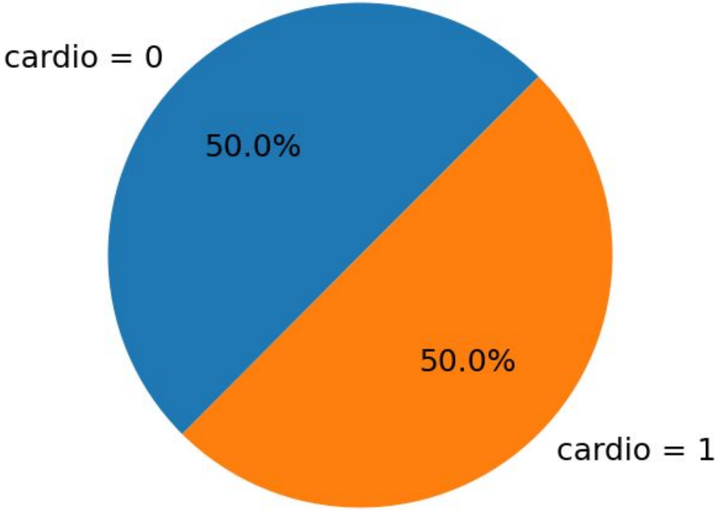
EDA — Attribute Types

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0

- Looks numerical but is categorical (ordinal)
  (1: normal, 2: above normal, 3: well above normal)
- Usually part of the documentation of dataset
- Interpretation requires good understanding of the data
  ➜ Generally impossible for automated methods

EDA — Distribution of Class Labels

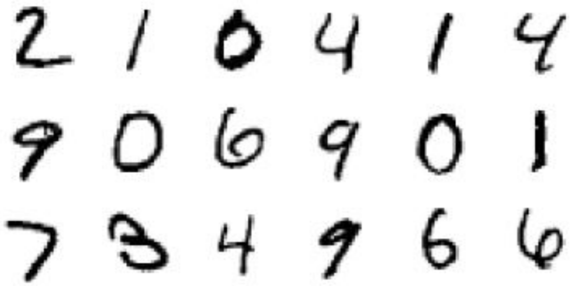
- Classification tasks generally benefit from balanced datasets
  - Balanced = all classes are (almost) equally represented
  - Distribution of classes also affects evaluation of found patterns



EDA — Visualizing High-Dimensional Data

- Visualization using dimensionality reduction techniques (here: t-SNE)

- MNIST Dataset
  - 60k handwritten digits 0, 1, 2, …, 9
    (~6k samples for class)
  - 28×28 pixels ➜ 784 features
    (integer grayscale valueS 0..255)



EDA — Unstructured Data (just some intuitions)

- Plain text
  - Language, (size of) vocabulary
  - Formal vs. informal text (e.g., social media content with slang, emoticons, emojis)
- Images/videos
  - Dimensions and resolutions
  - Color spaces
- Audio
  - Sampling rate and frequency range
  - Types of recording (e.g., voice vs. music)

Quick Quiz

Which of the statements on the right is **True**?

A Outliers are always noise and need to be removed before an analysis

B As long as my class labels are balanced, I will get good results

C Boxplots are often insufficient to identify all outliers in a dataset

D If attribute values show a weird distribution, ~~I know~~ something is off

Something might be

Outline

- Course Logistics
- Overview
  - What is Knowledge Discovery / Data Mining?
  - Common Data Mining tasks
  - Types of data & data representations
- **Data preparation**
  - Data quality
  - Exploratory Data Analysis (EDA)
  - **Data preprocessing**
- Summary

Data Preprocessing

- Main purposes
  - Improve data quality ("*Garbage in, garbage out!*")
  - Generate valid input for data mining algorithms
  - Remove complexity from data to ease analysis

- Core preprocessing task
  - Data cleaning
  - Data reduction
  - Data transformation
  - Data discretization

Data Cleaning

- Improve data quality
  - Remove or fill missing values
  - Identify and remove outliers
    (if outliers are not the goal of the analysis)
  - Identify and remove/merge duplicates
  - Correct errors and inconsistencies
    (e.g., convert inches to centimeters)

Non-trivial tasks and typically very application-specific

Data Reduction

- Reducing the number of data points
  - Sampling — select subset of data points (typically random or stratified sampling)
  - Commonly used for preliminary analysis or when the data size is extremely large
- Reducing the number of attributes
  - Removing irrelevant attributes (e.g., ids or ethically questionable attributes such as religion, sexual orientation, etc.)
  - Dimensionality reduction — mapping the data into a lower-dimensional space (PCA, LDA, t-SNE, etc.)
- Reducing the number of attribute values (form of noise removal)
  - Aggregation or generalization
  - Binning with smoothing

Reducing the Number of Attribute Values — Examples

- Aggregation
  - Moving up concept hierarchy of numerical attributes (e.g., from days to years)
  - Generalization for categorical attributes

	id	age	gender	height
0	0	18393	2	168
1	1	20228	1	156
2	2	18857	1	165
3	3	17623	2	169
4	4	17474	1	156

	id	age	gender	height
0	0	50.0	2	168
1	1	55.0	1	156
2	2	51.0	1	165
3	3	48.0	2	169
4	4	47.0	1	156

- Binning and smoothing
  - Sort by attribute value (e.g., height)
  - Split data into bins of equal sizes
  - Replace each value with bin mean (the means are also rounded in this example)

55	57	59	60	64	65	65	66	67	67	67	68	68	70	70	70	...

55 57 59 60 64	65 65 66 67 67	67 68 68 70 70	70 ...

59 59 59 59 59	66 66 66 66 66	69 69 69 69 69	72 ...

Data Transformation

- Some data reduction techniques also transform the data
  - Dimensionality reduction, aggregation/generalization, binning, etc.

- Attribute construction
  - Add or replace attribute inferred from existing attributes
  - Example: weight, volume ➜ density

- Normalization
  - Scaling attribute values to value into a specified range (e.g., [0,1])
  - Standardization: scaling by using mean and standard deviation

Normalization — Examples

[0, 1]

Min-max normalization

$$x_i^{weight} = \frac{x_i^{weight} - min(x^{weight})}{max(x^{weight}) - min(x^{weight})}$$

weight
62.0
85.0
64.0
82.0
56.0

weight
0.273684
0.394737
0.284211
0.378947
0.242105

weight
-0.847867
0.749826
-0.708937
0.541431
-1.264657

$$x_i^{weight} = \frac{x_i^{weight} - \mu^{weight}}{\sigma^{weight}}$$

Standardization
(z-score normalization)

Data Discretization

- Converting continuous attributes into ordinal attributes
  - Some algorithms accept only categorical attributes
  - Convert a regression task to a classification task
- Example: Convert weight to a weight category
  - Many existing discretization methods
  - Here: discretization using 3 user-defined bins

	id	age	gender	weight
0	66283	14461	1	86.0
1	4780	14740	1	57.0
2	34457	21090	1	88.0
3	83116	15869	2	60.0
4	70356	20687	1	103.0

→

	id	age	gender	weight	weight_bin	weight_class
0	66283	14461	1	86.0	(70, 90]	average
1	4780	14740	1	57.0	(0, 70]	light
2	34457	21090	1	88.0	(70, 90]	average
3	83116	15869	2	60.0	(0, 70]	light
4	70356	20687	1	103.0	(90, 150]	heavy

One-Hot Encoding

- Converting categorical attributes into numerical attributes
  - Converting categorical attributes into a series of binary attributes 0/1
  - Allows the application of any methods for numerical features on categorical attributes

weight_class

- Example

ord.

	id	weight_class
0	66598	light
1	88878	average
2	80363	heavy
3	52546	light
4	17715	average

nominal

	id	weight_class_light	weight_class_average	weight_class_heavy
0	66598	1	0	0
1	88878	0	1	0
2	80363	0	0	1
3	52546	1	0	0
4	17715	0	1	0

Quick Quiz

Which attributes are generally(!) not relevant for the analysis and **SHOULD be removed**?

ID	Age	Edu-cation	Marital Status	Annual Income	Email	Credit Default
101	23	Masters	Single	75k	alice@...	Yes
102	35	Bachelor	Married	50k	bob@...	No
103	26	Masters	Single	70k	claire@...	Yes
104	41	PhD	Single	95k	dave@...	Yes
105	18	Bachelor	Single	40k	erin@...	No
106	24	Masters	Single	65k	fred@...	Yes

A ID + Email

B Age + Email

C ID + Education

D ID + Marital Status

Quick Quiz — Side Note

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Journal of Research in Personality 42 (2008) 1116–1122

JOURNAL OF RESEARCH IN PERSONALITY

www.elsevier.com/locate/jrp

Brief Report

How extraverted is honey.bunny77@hotmail.de? Inferring personality from e-mail addresses

Mitja D. Back *, Stefan C. Schmukle, Boris Egloff

Department of Psychology, University of Leipzig, Seeburgstr. 14-20, 04103 Leipzig, Germany

Available online 29 February 2008

Quick Quiz

Which attributes are arguably not relevant or "problematic" and **SHOULD be removed**?

Age	Religion	Edu-cation	Has Account	Annual Income	Zodiac Sign	Credit Approval
23	Buddhist	Masters	Yes	75k	Leo	Yes
35	Buddhist	Bachelor	Yes	50k	Gemini	No
26	Muslim	Masters	Yes	70k	Libra	Yes
41	Christian	PhD	Yes	95k	Leo	Yes
18	Buddhist	Bachelor	Yes	40k	Virgo	No
24	Muslim	Masters	Yes	65k	Aries	Yes

A Religion + Education + Zodiac Sign

B (✓) Religion + Zodiac Sign + Has Account

C Religion + Zodiac Sign

D Has Account + Zodiac Sign

Quick Quiz — Side Note

CNA Insider

Does a job seeker's horoscope matter? For some companies, the answer is yes

There are companies that turn to unconventional methods like astrology, tarot reading and numerology to help guide hiring decisions. What beliefs are these practices grounded in, and how legitimate are they?

Outline

- Course Logistics
- Overview
  - What is Knowledge Discovery / Data Mining?
  - Common Data Mining tasks
  - Types of data & data representations
- Data preparation
  - Data quality
  - Exploratory Data Analysis (EDA)
  - Data preprocessing
- **Summary**

Summary

- Course Logistics

- Core Concepts
  - What is (not) Data Mining?
  - Knowledge discovery process
  - Overview to common tasks

Data ➜ Knowledge

- Data preparation
  - Types of data and data quality
  - Exploratory data analysis (EDA)
  - Data preprocessing

Know your data & clean your data!

Solutions to Quick Quizzes

- Slide 22: D
- Slide 37: D
- Slide 38: B (A also OK)
- Slide 39: A (in general)
- Slide 55: C
- Slide 65: A
- Slide 67: B (C also OK)