Data, Methods, Scenarios

Mining Massive Datasets

Prof. Carlos Castillo

Topic 02

Universitat Pompeu Fabra Barcelona

Main Sources

- Data Mining, The Textbook (2015) by Charu Aggarwal (Chapter 1) + slides by Lijun Zhang
- Mining of Massive Datasets, 2nd edition (2014) by Leskovec et al. (Chapter 1)
- Data Mining Concepts and Techniques, 3rd edition (2011) by Han et al. (Chapters 1-2)

Contents

- Types of data
- Types of problem
- Example scenarios
- Major challenges

Data types

Nondependency / Dependency

- Nondependency oriented data can be structured so items are separate
  - Relational data, text data
- Dependency oriented data includes relationships between items
  - Graphs, time series

Mixed attribute data

- Most attributes we will deal with are numerical, they quantify something
- Sometimes attributes are categorical
  - Categorical
    - Example: elephant, tiger, moose, …
  - Binary (two categories)
    - Example: present, absent
  - Ordinal (two or more categories that can be naturally sorted)
    - Example: low, medium, high
- Real-world datasets include both types

Binary attributes, sets, dummy vars.

- Every binary attribute can be used as a marker of belonging to a set and viceversa

Name	Age	Gender	Race	ZIP Code
John S.	45	M	African American	05139
Manyona L.	31	F	Native American	10598
Sayani A.	11	F	East Indian	10547
Jack M.	56	M	Caucasian	10562
Wei L.	63	M	Asian	90210

- **One-hot encoding**: every categorical attribute taking one of k values can be encoded as k “dummy” binary attributes

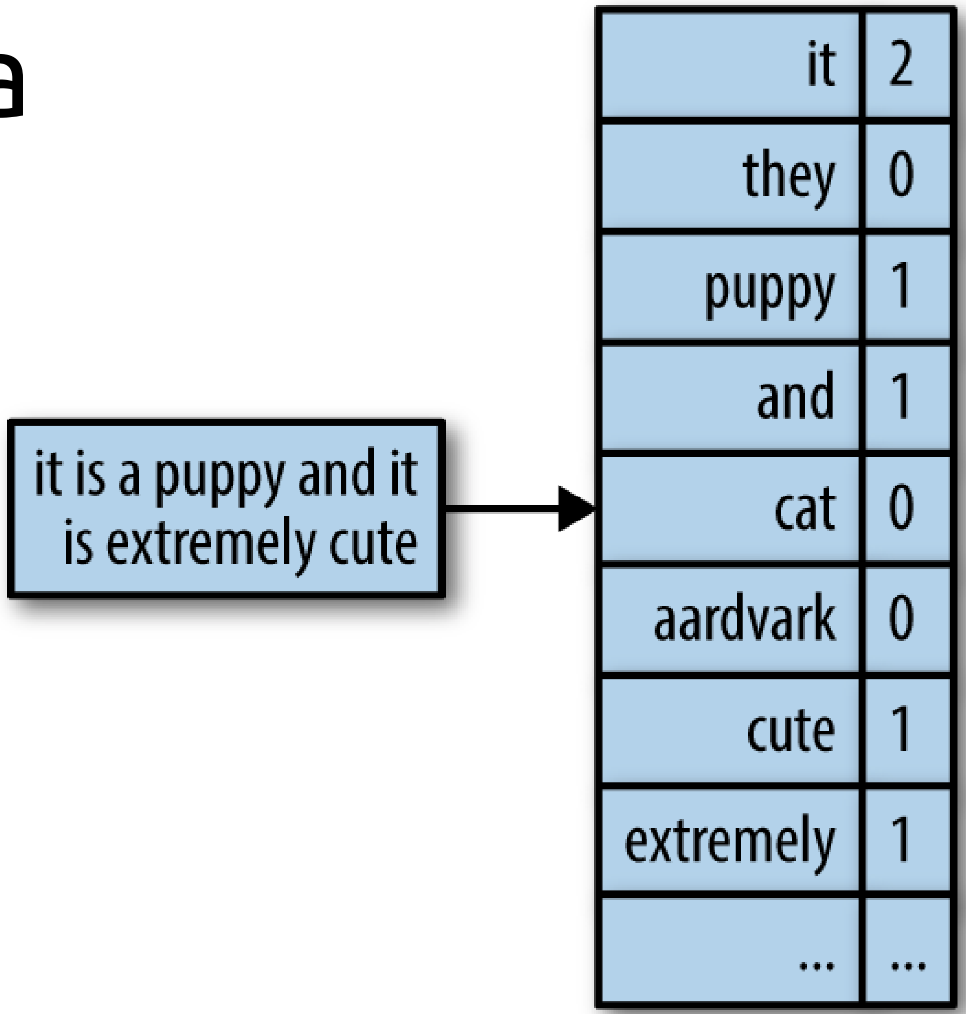
Question

- Suppose you encode race and gender using one-hot encoding. How many columns will your new dataset have?

Name	Age	Gender	Race	ZIP Code
John S.	45	M	African American	05139
Manyona L.	31	F	Native American	10598
Sayani A.	11	F	East Indian	10547
Jack M.	56	M	Caucasian	10562
Wei L.	63	M	Asian	90210

Textual data

- Text be represented as:
  - As a string
  - As a set of binary variables, one for each word in the dictionary, with value True iff the word belongs to the text (the “bag-of-words” model)
  - As a set of numerical variables indicating number of occurrences (the “vector space” model)



http://uc-r.github.io/creating-text-features

Time series data

- Contextual attributes
  - Timestamps, sequence number, …
- Behavioral attributes
  - Readings of a sensor, value of the variable, …

- *Multivariate* time series data has multiple behavioral attributes

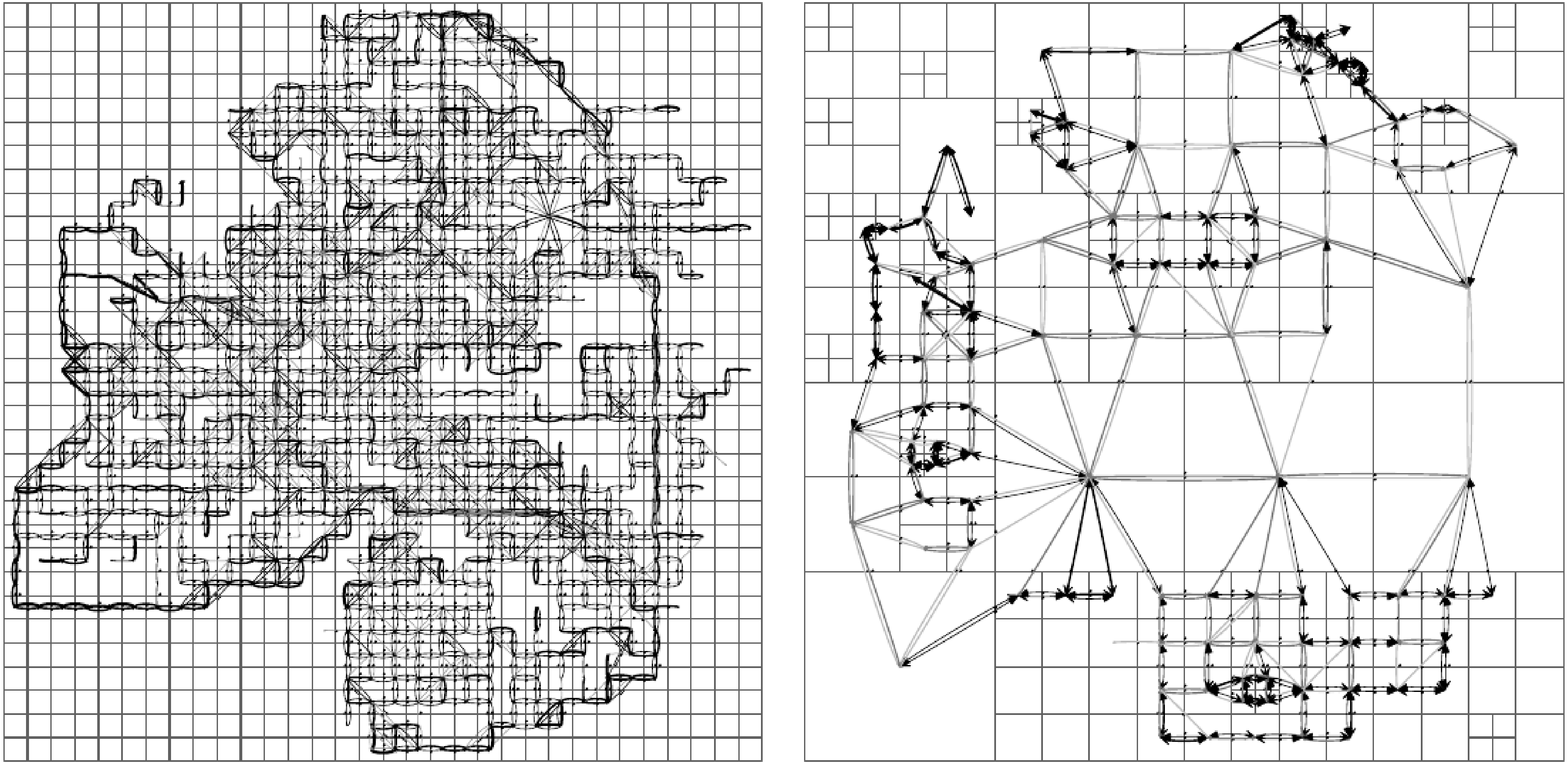
Spatial data

- Two (lat/long) or three (lat/long/elevation) spatial attributes
- Represented by images
- Remote sensing data

Spatiotemporal data

- Spatial and temporal attributes are contextual
  - Example: sea surface temperature
- Temporal attribute is contextual, spatial attribute is behavioral
  - Example: trajectories

Example: trajectory data aggregation



Bonchi, F., Castillo, C., Donato, D., & Gionis, A. (2009). Taxonomy-driven lumping for sequence mining. Data Mining and Knowledge Discovery, 19(2), 227-244.

Problem types

Data mining methods try to find relationships

- **Between columns**
  - Find associations, correlations, …
  - If there is *one* key column: classification, prediction, …
- **Between rows**
  - Find clusters
  - Detect outliers

Example:
Association pattern mining

- Sparse binary databases representing, e.g., items a person is interested in

$$\begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \in \{0,1\}^{5\times 4}$$

- The relative frequency of a pattern is its support

Frequent Patterns	Support
$\{2,3\}$	3/5
$\{1,4\}$	2/5

https://cs.nju.edu.cn/zlj/Course/DM_15.html

Association pattern mining (cont.)

- Given a binary $n \times d$ data matrix D,
  - determine all subsets of columns such that all the values in these columns take on the value True for at least a fraction *min_support* of the rows in the matrix.
- The relative frequency of a pattern is referred to as its support

Association pattern mining (cont.)

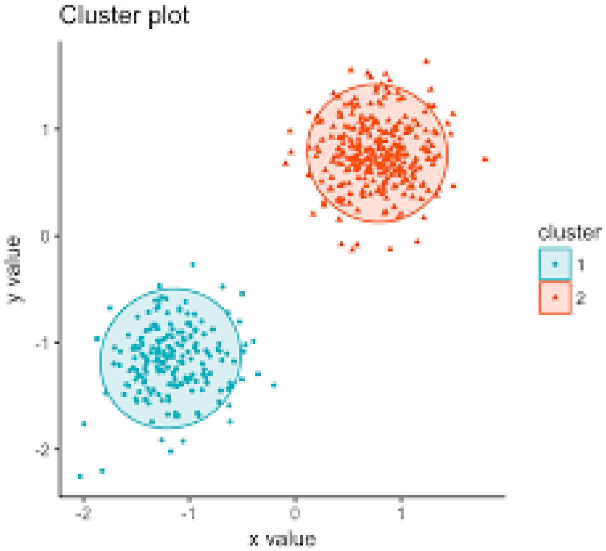
- The confidence of a rule A→B is
  - support(A U B) / support(A)
- Example:
  - { Chips, Olives } → { Beer }

Exercise

- The confidence of a rule A→B is
  - support(A U B) / support(A)
- Suppose
  - 10 people buy only Chips and Beer
  - 20 people buy only Chips and Olives
  - 30 people buy only Olives and Beer
  - 40 people buy all three: Chips, Olives, and Beer.
- What is the confidence of the rule { Chips, Olives } → { Beer } ?

Answer in Nearpod Collaborate
https://nearpod.com/student/
Code to be given during class

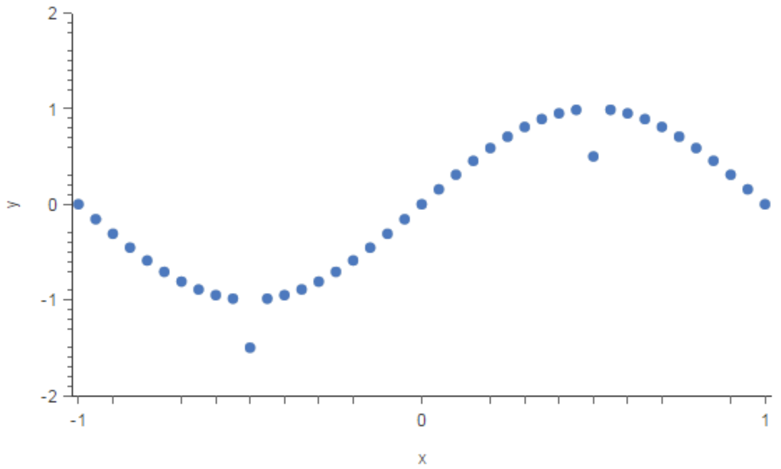
Clustering



- Partition records/rows in a way that
  - elements in the same partition are similar
  - elements in different partitions are different
- *But what does it mean to be similar? How many sets? Can a record/row belong to two sets? To zero sets? ...*
- Applications:
  - Segmentation, summarization, ...
  - Sometimes a step in a larger DM algorithm

Image credit: http://www.sthda.com/english/articles/tag/pam-clustering/

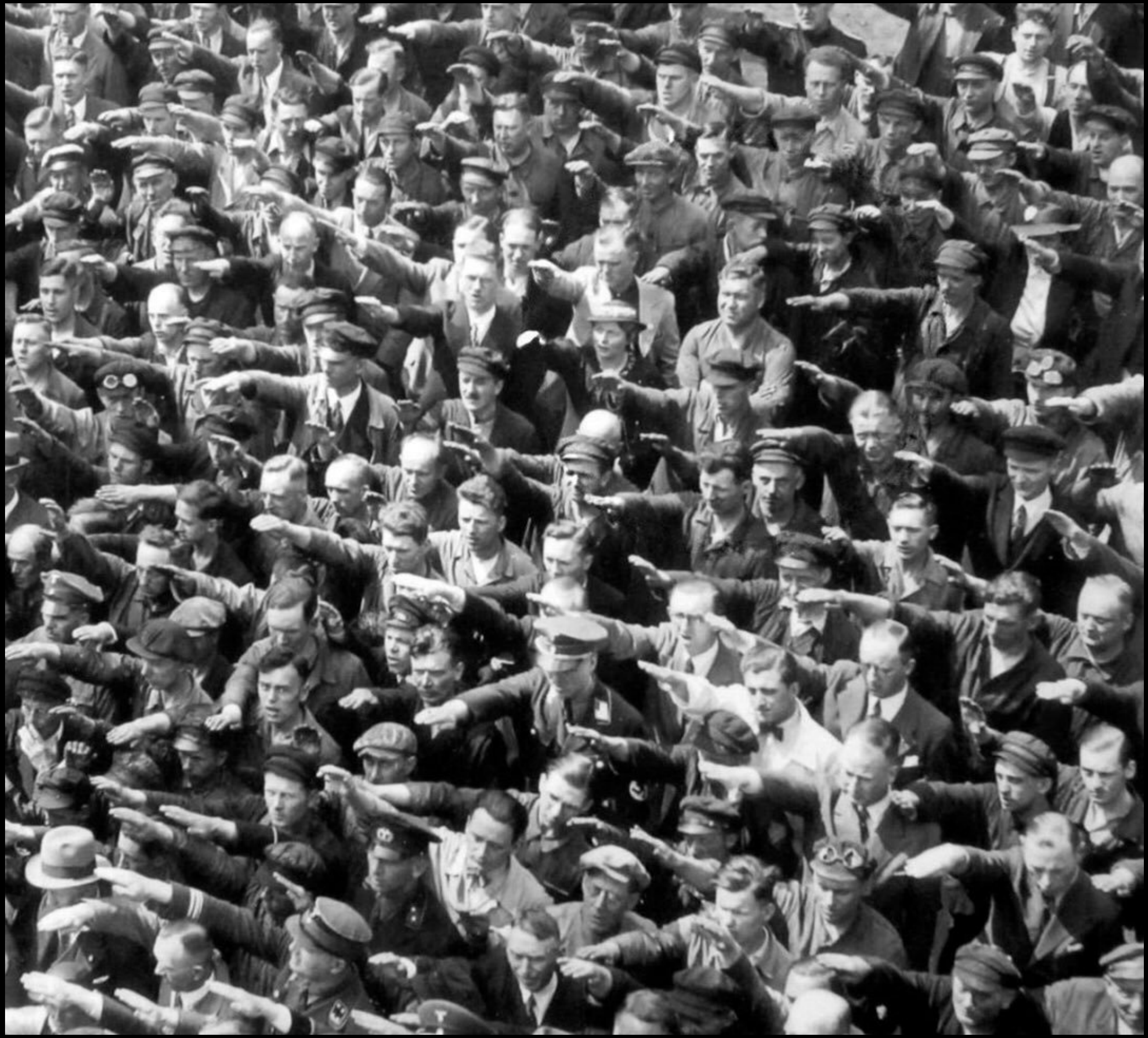
Outlier detection



- Given a database, find records/rows that are different from the rest of the database
- *But what does it mean to be different? How many can be different? How different should they be?*
- Applications:
  - Intrusion detection, credit card fraud, interesting sensor events, medical diagnosis, …

Image credits: https://www.kdnuggets.com/2017/01/3-methods-deal-outliers.html

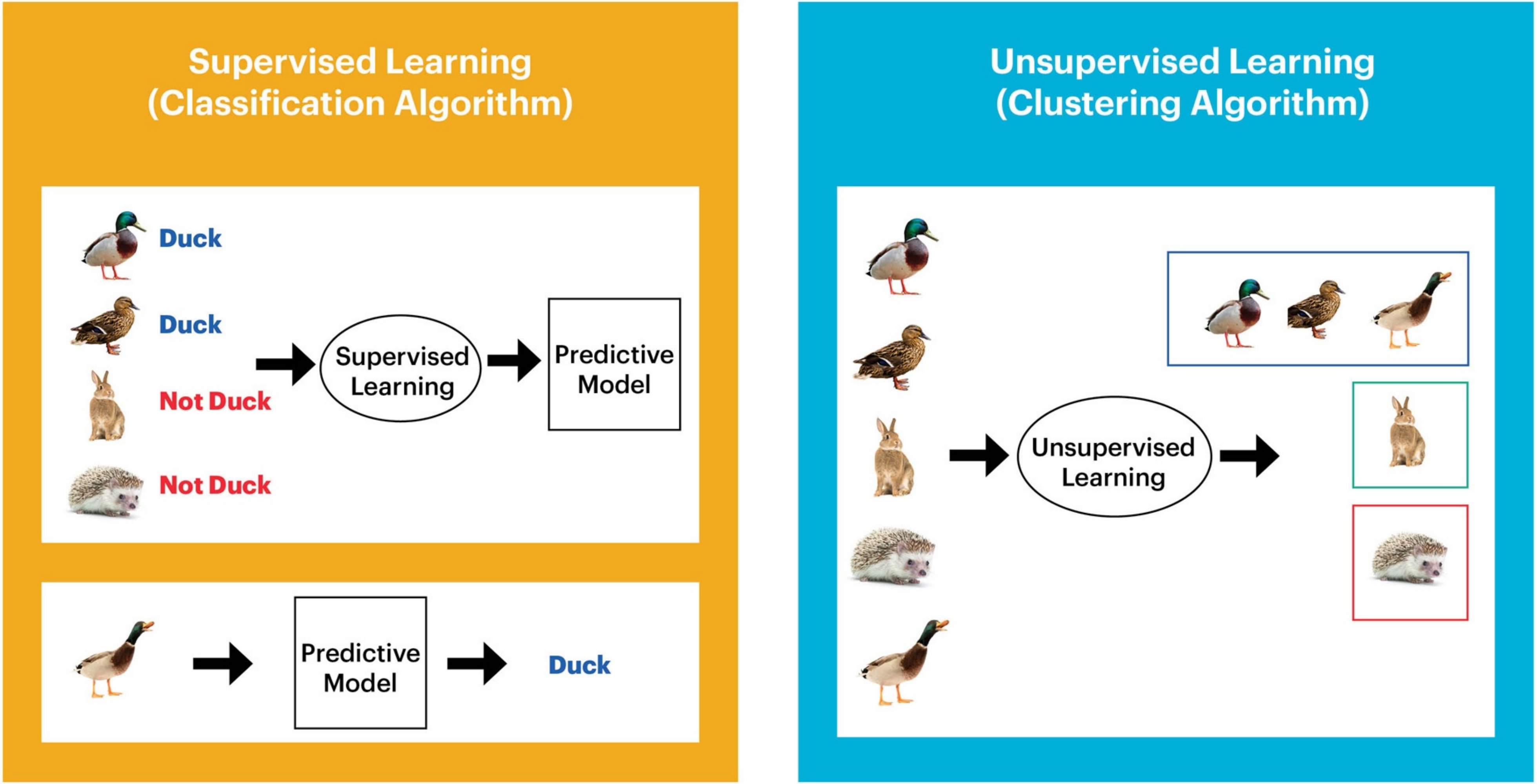
Outlier detection (cont.)



August Landmesser in 1936

Data classification

- When data has a special feature known as a class label
- A model can learn from previous data to associate a record/row to a class label
- Applications:
  - Too many to list here :-)



https://blog.westerndigital.com/machine-learning-pipeline-object-storage/

Western Digital.

Tasks with complex data types

- Frequent temporal patterns
- Time series motifs
- Graph motifs
- Trajectory clusters
- Collective classification
- ...

Data types x Prototypical problems

Problem	Time series	Spatial	Sequence	Networks
Patterns	Motif-mining Periodic pattern	Colocation patterns	Sequential patterns Periodic Sequence	Structural patterns
	Trajectory patterns			
Clustering	Shape clusters	Spatial clusters	Sequence clusters	Community detection
	Trajectory clusters			
Outliers	Position outlier Shape outlier	Position outlier Shape outlier	Position outlier Combination outlier	Node outlier Linkage outlier Community outliers
	Trajectory outliers			
Classification	Position classification Shape classification	Position classification Shape classification	Position classification Sequence classification	Collective classification Graph classification
	Trajectory classification			

Data Mining, The Textbook (2015) by Charu Aggarwal

Example scenarios

Example scenario 1

- Place products in a store to maximize co-purchases of items frequently bought together
  - Input data: baskets
  - Output: similar pairs
  - Algorithm: frequent pattern mining

Example scenario 2

- Recommend movies to users in a video-on-demand platform
  - Input data: viewing history
  - Output: recommendations for a user
  - Simple algorithm: k nearest neighbors

Example scenario 3

- Help diagnose if an electrocardiogram is associated to a health problem
  - Input data: time series, possibly multi-dimensional
  - Output: binary label or risk score
  - Algorithms: outlier detection or classification

Example scenario 4

- Help a sysadmin determine if an intruder is trying or has accessed the network
  - Input data: time series of event records
  - Output: binary label or risk score
  - Algorithms: event detection

Are these data mining tasks?

A) Dividing the customers of a company by gender
B) Finding credit card scammers among customers of a company
C) Computing the total sales of a company
D) Sorting a student database by student identification number
E) Predicting the future stock price of a company using past records
F) Determine when a complex machine needs to be repaired
G) Extracting the frequencies of a sound wave

Answer in Nearpod Time to Climb
https://nearpod.com/student/
Code to be given during class

Major challenges

Methodological challenges

- Mining new kinds of knowledge
- Mining multidimensional data
- Fully utilizing the expertise of domain experts who know the data better
- Handling uncertainty, noise, incompleteness

User interaction challenges

- Allowing users to ask the questions that matter to them
- Performing interactive mining
- Presenting and visualizing data mining results

Efficiency and scalability

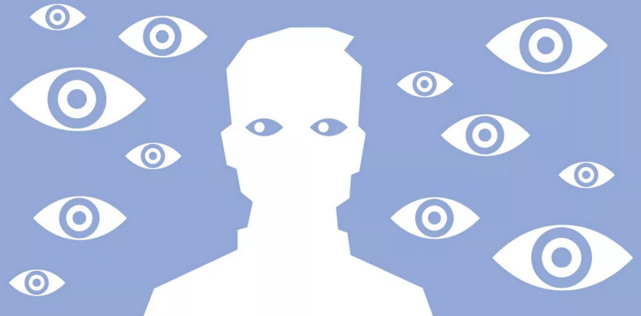
- Data cannot be stored in a single machine
  - Processing time of an algorithm might be exponential in input … or even polynomial, even with small degree: a process can become unreasonably slow very quickly
  - Streaming algorithms
  - Parallel/distributed mining algorithms

Diversity of database types

- Real databases are a complex mixture of very rich and diverse data types
- Mining dynamic, networked, global data repositories
  - Integrating from complementary sources

Data mining can be harmful

- Social impacts of data mining
  - Who wins? And more importantly, who loses?
- Privacy-preserving data mining
  - Avoid invisible, pervasive, invasive data mining



Summary

Things to remember

- Types of data
- Types of data mining methods
- Prototypical data mining scenarios
- Typical challenges of data mining

Exercises for this topic

- **Section 1.9 of Data Mining, The Textbook (2015) by Charu Aggarwal**
- Exercises 1.7 of Introduction to Data Mining, Second Edition (2019) by Tan et al.