UNIVERSITY OF MANNHEIM

Data Mining

Introduction to Data Mining

Hallo

- **Prof. Dr. Christian Bizer**
- Professor for Information Systems
- Research Interests:
  - Data and Web Mining
  - Web Data Integration
  - Data Web Technologies
- Room: B6 - B1.15
- eMail: christian.bizer@uni-mannheim.de

Hallo

- **M. Sc. Wi-Inf. Alexander Brinkmann**
- Graduate Research Associate
- Research Interests:
  - Data Search using Deep Learning
  - Product Data Categorization
- Room: B6, 26, C 1.04
- eMail: alexander.brinkmann@uni-mannheim.de

- Will teach one of the exercise groups and will supervise student projects.

Hallo

- **M. Sc. Wi-Inf. Keti Korini**
- Graduate Research Associate
- Research Interests:
  - Schema Mapping
  - Table Annotation using Deep Learning
- Room: B6, 26, C 1.03
- eMail: kkorini@uni-mannheim.de
- Will teach one of the exercise groups and will supervise student projects.

Hallo

- **M. Sc. Wi-Inf. Ralph Peeters**
- Graduate Research Associate
- Research Interests:
  - Entity Matching using Deep Learning
  - Product Data Integration
- Room: B6, 26, C 1.04
- eMail: ralph.peeters@uni-mannheim.de

- Will teach one of the exercise groups and will supervise student projects.

Course Organisation

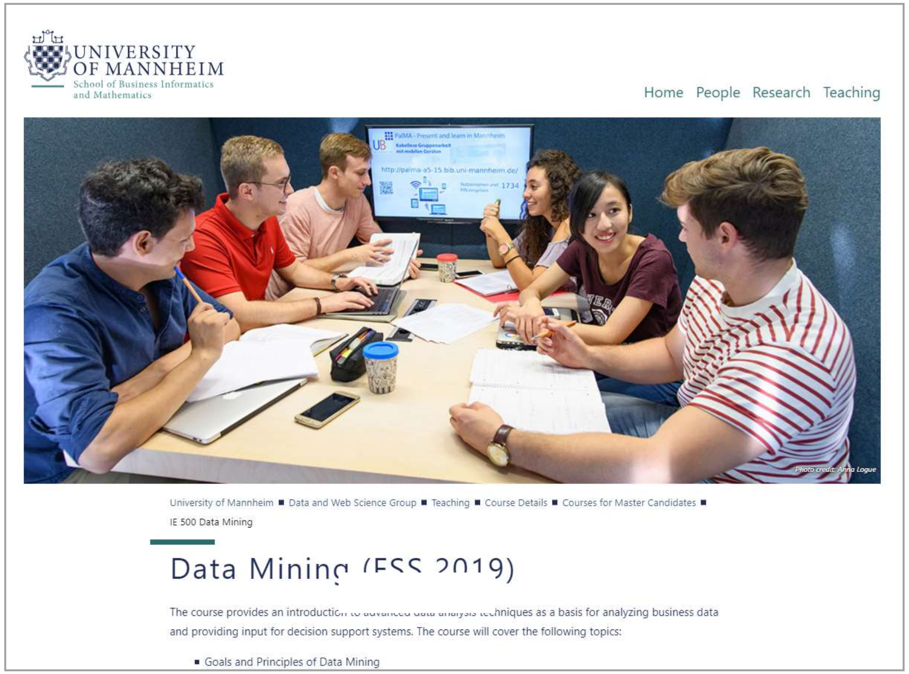
- Lecture
  - introduces the principal methods of data mining
  - discusses how to evaluate the learned models
  - presents practical examples of data mining applications
- Exercise Groups
  - students experiment with the learned methods using Python
- Project Work
  - teams of six students realize a data mining project
  - teams may choose their own data sets and tasks
    (in addition, I will propose some suitable data sets and tasks)
  - teams write a 10 page summary about their project and present the results
- Grading
  - 75% written exam, 20% project report, 5% presentation of project results

Course Organisation

– Course Webpage
  - provides up-to-date information, lecture slides, and exercise material
  - https://www.uni-mannheim.de/dws/teaching/course-details/courses-for-master-candidates/ie-500-data-mining/

– Mailing List, Discussion Forum
  - ILIAS eLearning System, https://ilias.uni-mannheim.de/

– Time and Location
  - **Lecture:**
    - Wednesday, 10.15 - 11.45, A5 B1.44
  - **Exercise:**
    - Thursday, 10:15 - 11:45, Keti Korini
      Room A 104 (B6 , Bauteil A)
    - Thursday, 12:00 - 13:30, Alex Brinkmann
      Room A 104 (B6 , Bauteil A)
    - Thursday, 13:45 - 15:15, Ralph Peeters
      Room A 104 (B6 , Bauteil A)



University of Mannheim ■ Data and Web Science Group ■ Teaching ■ Course Details ■ Courses for Master Candidates ■ IE 500 Data Mining

Data Mining (FSS 2019)

The course provides an introduction to advanced data analysis techniques as a basis for analyzing business data and providing input for decision support systems. The course will cover the following topics:

- Goals and Principles of Data Mining

Lecture Contents

1. Introduction to Data Mining	What is Data Mining? Tasks and Applications The Data Mining Process
2. Cluster Analysis	K-means Clustering, Density-based Clustering, Hierarchical Clustering, Proximity Measures
3. Classification	Nearest Neighbor, Decision Trees and Forests, Rule Learning, Naïve Bayes, SVMs, Neural Networks, Model Evaluation, Hyperparameter Selection
4. Regression	Linear Regression, Nearest Neighbor Regression, Regression Trees, Time Series
5. Text Mining	Preprocessing Text, Feature Generation, Feature Selection
6. Association Analysis	Frequent Item Set Generation, Rule Generation, Interestingness Measures

Schedule

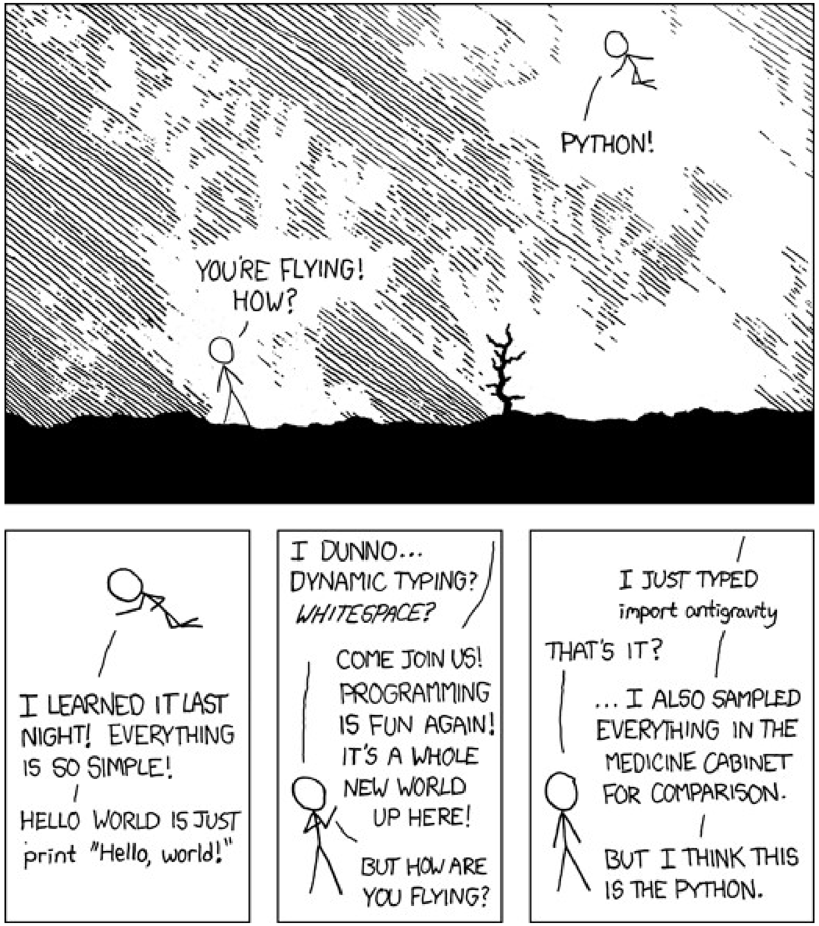
Week	Wednesday	Thursday
14.02.2024	Lecture: Introduction to Data Mining Tutorial: Introduction to Python	Exercise: Preprocessing/Visualization
21.02.2024	Lecture: Cluster Analysis	Exercise: Cluster Analysis
28.02.2024	Lecture: Classification 1	Exercise: Classification
06.03.2024	Lecture: Classification 2	Exercise: Classification
13.03.2024	Lecture: Classification 3	Exercise: Classification
20.03.2024	Lecture: Regression	Exercise: Regression
	- Easter Break -	
10.04.2024	Lecture: Text Mining	Exercise: Text Mining
17.04.2024	Introduction to the student projects and group formation	Preparation of project outline
22.04.2024	Submission of project outlines (Deadline: 23:59)	
24.04.2024	Lecture: Association Analysis	Exercise: Association Analysis
25.04.2024	Feedback on project outlines	
30.04.2024	Project Work	Feedback on demand
08.05.2024	Project Work	Feedback on demand
15.05.2024	Project Work	Feedback on demand
17.05.2024	Submission of project reports (Deadline: 23:59)	
22.05.2024	Presentation of project results	

Deadlines

- Submission of project proposal
  - Thursday, April 22nd, 23:59
- Submission of final project report
  - Sunday, May 17th, 23:59
- Project presentations
  - Wednesday May 22nd
  - everyone has to attend the presentations

Introduction to Python

- Today, 14.02 at 15:30-17:00 in Room B6, A101
- Topics:
  - Setup of environment (Anaconda, Jupyter Notebooks)
  - Python Intro / Design Goals
  - Basic programming concepts in Python
- Support
  - Help with environment setup
  - Q&A
- Material
  - Tutorial slides available on website
  - Try to install Anaconda before the tutorial


YOU'RE FLYING! HOW?
PYTHON!
I LEARNED IT LAST NIGHT! EVERYTHING IS SO SIMPLE!
HELLO WORLD IS JUST print "Hello, world!"
I DUNNO... DYNAMIC TYPING? WHITESPACE?
COME JOIN US! PROGRAMMING IS FUN AGAIN! IT'S A WHOLE NEW WORLD UP HERE!
BUT HOW ARE YOU FLYING?
I JUST TYPED import antigravity
THAT'S IT?
... I ALSO SAMPLED EVERYTHING IN THE MEDICINE CABINET FOR COMPARISON.
BUT I THINK THIS IS THE PYTHON.

ChatGPT as Your Personal Tutor

We use ChatGPT in the exercises to

1. discuss suitable methods and parameter settings for different use cases
2. generate and debug Python code for experimenting with the methods
3. generate multiple-choice and open questions for self-assessment

Source: New York Times

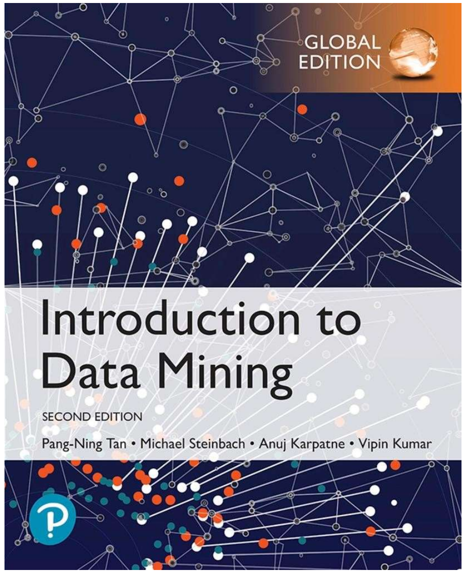
OpenAI

Final Exam

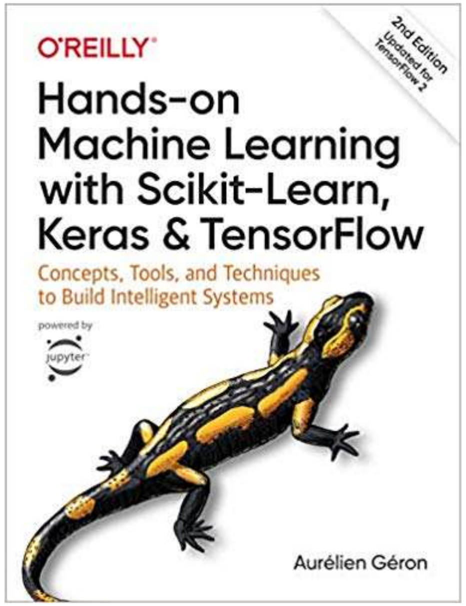
- Date and Time: Thursday, 13th June 2024
- Duration: 60 minutes
- Structure: 6 open questions that
  - Check whether you have understood the content of the lecture
    - we try to cover all major chapters of the lecture: cluster analysis, classification, regression, association analysis, text mining
  - Require you to describe the ideas behind algorithms and methods
    - often: How do methods react to special patterns in the data?
  - Might require you to do some simple calculations for which
    - you need to know the most relevant formulas
    - you do not need a calculator
- There is **no second exam** as the course is offered every semester.
  - If you fail the exam, you also need to redo the project. So better prepare well.

Textbooks for the Course

Pang-Ning Tan, Michael Steinbach, Vipin Kumar:
Introduction to Data Mining. 2nd Edition.
Pearson / Addison Wesley.

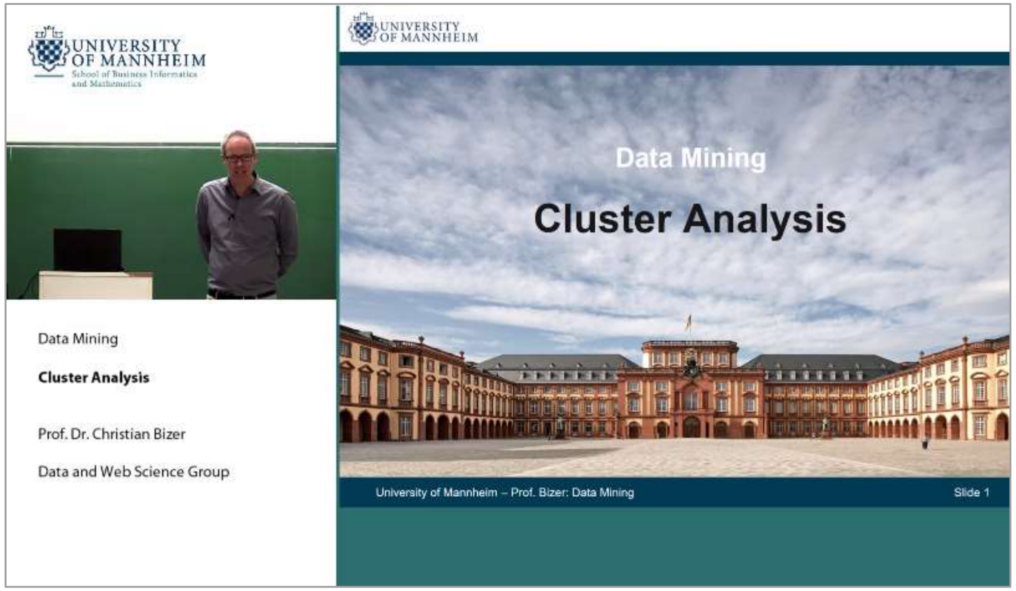

GLOBAL EDITION
Introduction to Data Mining
SECOND EDITION
Pang-Ning Tan • Michael Steinbach • Anuj Karpatne • Vipin Kumar

Aurélien Géron:
Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow.
2^{nd} or 3^{rd} Edition, O'Reilly, 2019 or 2022


O'REILLY
2nd Edition
Updated for TensorFlow 2
Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow
Concepts, Tools, and Techniques to Build Intelligent Systems
powered by jupyter
Aurélien Géron

Videos and Screencasts

- **Lecture Videos** by Heiko Paulheim and Christian Bizer
  - https://www.uni-mannheim.de/dws/teaching/lecture-videos/#c101905
- **Screencasts** for the Exercises by Ralph Peeters
  - https://www.uni-mannheim.de/dws/teaching/lecture-videos/#c101906



Data Mining

Cluster Analysis

Prof. Dr. Christian Bizer

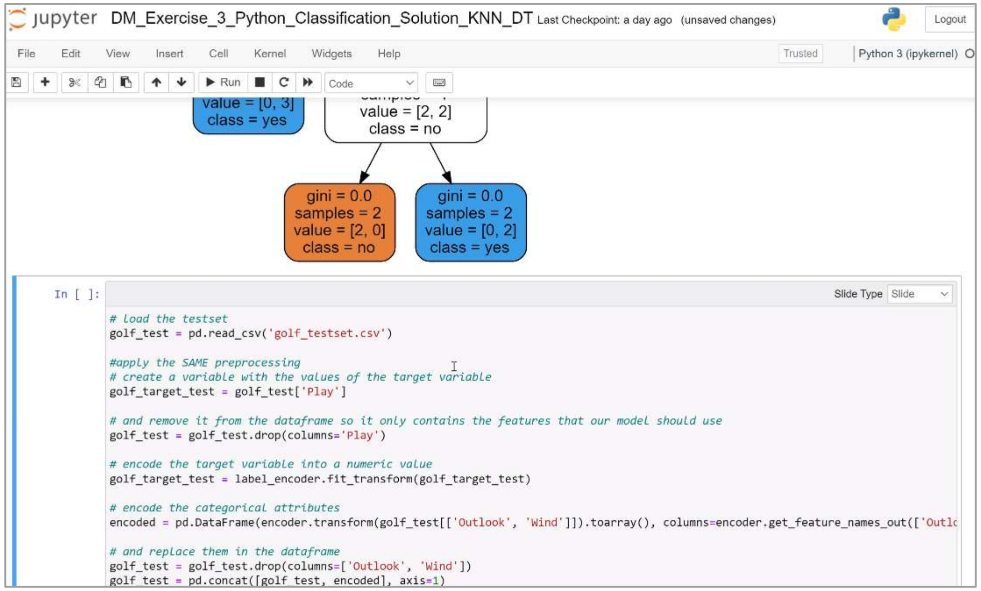
Data and Web Science Group

Data Mining

Cluster Analysis

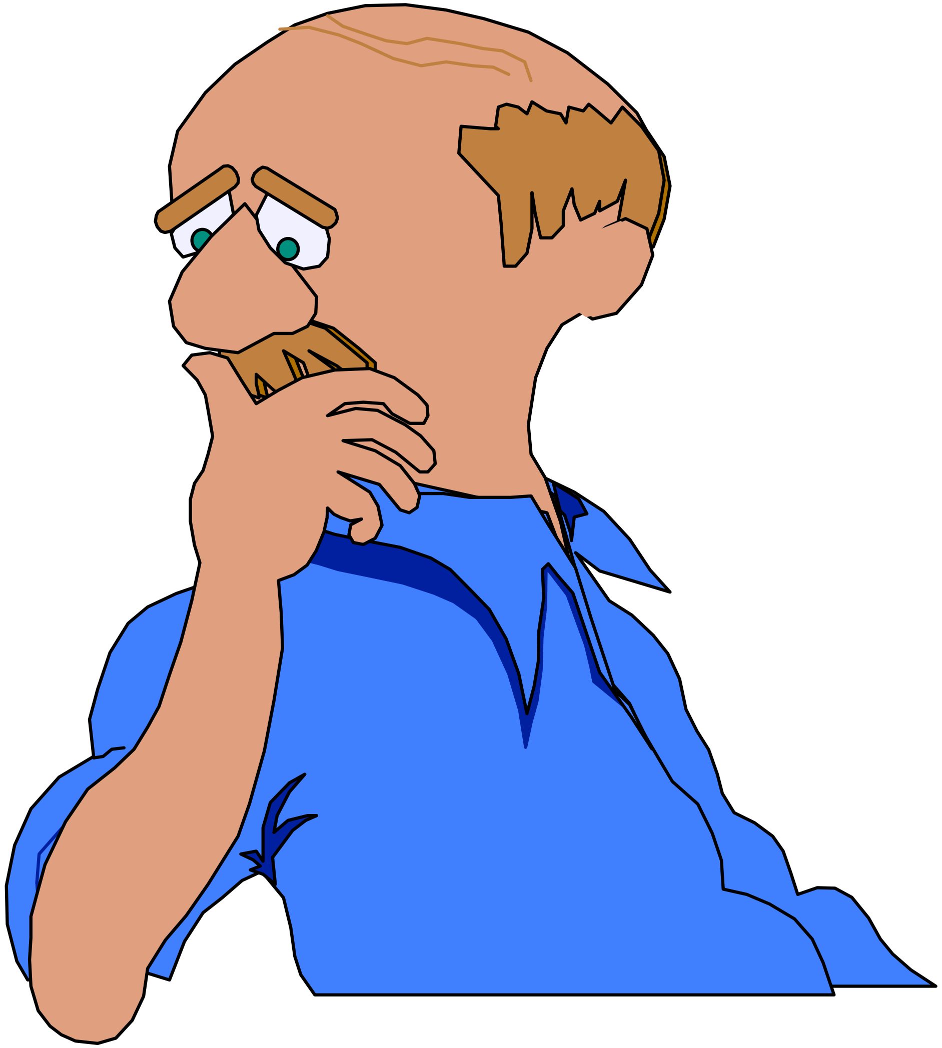
University of Mannheim – Prof. Bizer: Data Mining

Slide 1



```
# load the testset
golf_test = pd.read_csv('golf_testset.csv')

#apply the SAME preprocessing
# create a variable with the values of the target variable
golf_target_test = golf_test['Play']

# and remove it from the dataframe so it only contains the features that our model should use
golf_test = golf_test.drop(columns='Play')

# encode the target variable into a numeric value
golf_target_test = label_encoder.fit_transform(golf_target_test)

# encode the categorical attributes
encoded = pd.DataFrame(encoder.transform(golf_test[['Outlook', 'Wind']]).toarray(), columns=encoder.get_feature_names_out(['Outlo

# and replace them in the dataframe
golf_test = golf_test.drop(columns=['Outlook', 'Wind'])
golf_test = pd.concat([golf_test, encoded], axis=1)
```

Questions?



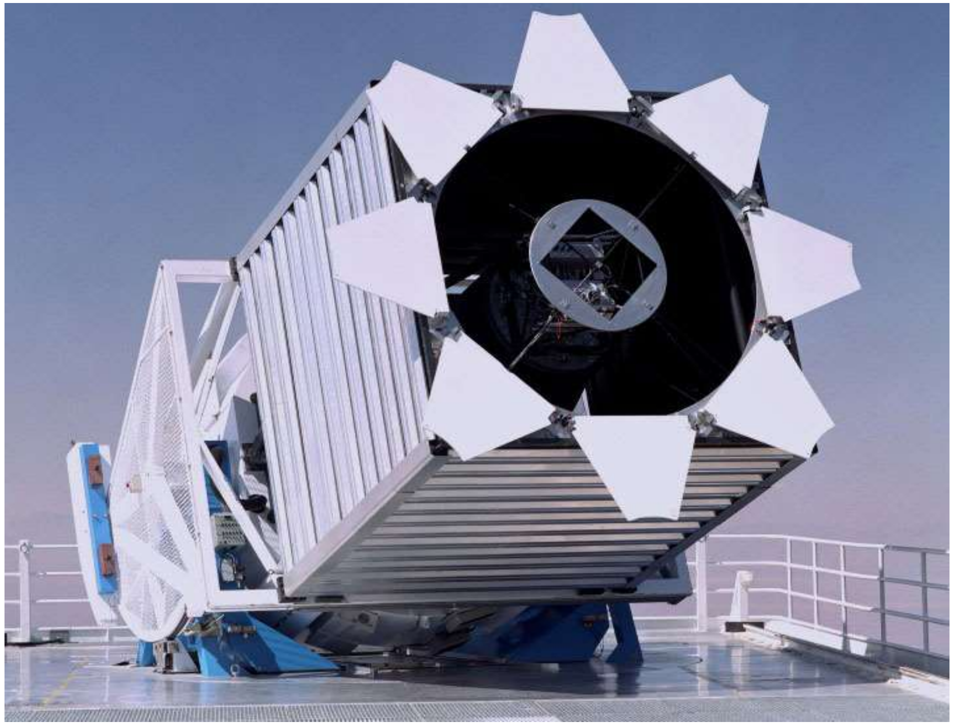
Outline: Introduction to Data Mining

1. What is Data Mining?
2. Tasks and Applications
3. The Data Mining Process

1. What is Data Mining?

- Large quantities of data are collected about all aspects of our lives
- This data contains interesting patterns
- Data Mining helps us to
  1. discover these patterns and
  2. use them for decision making across all areas of society, including
     - Business and industry
     - Science and engineering
     - Medicine and biotech
     - Government
     - Individuals

“We are Drowning in Data...”



Sloan Digital Sky Survey
≈ 200 GB/day
≈ 73 TB/year

Predict

- Type of sky object: Star or galaxy?

“We are Drowning in Data...”

US Library of Congress

≈ 235 TB archived

≈ 40 Wikipedias

arXiv Preprint Server

> 2 million papers

Discover

- Topic distributions*
- Citation networks

Train

- Large Language Models

* Lansdall-Welfare, et al.: Content analysis of 150 years of British periodicals. PNSA, 2017.

“We are Drowning in Data...”

Facebook

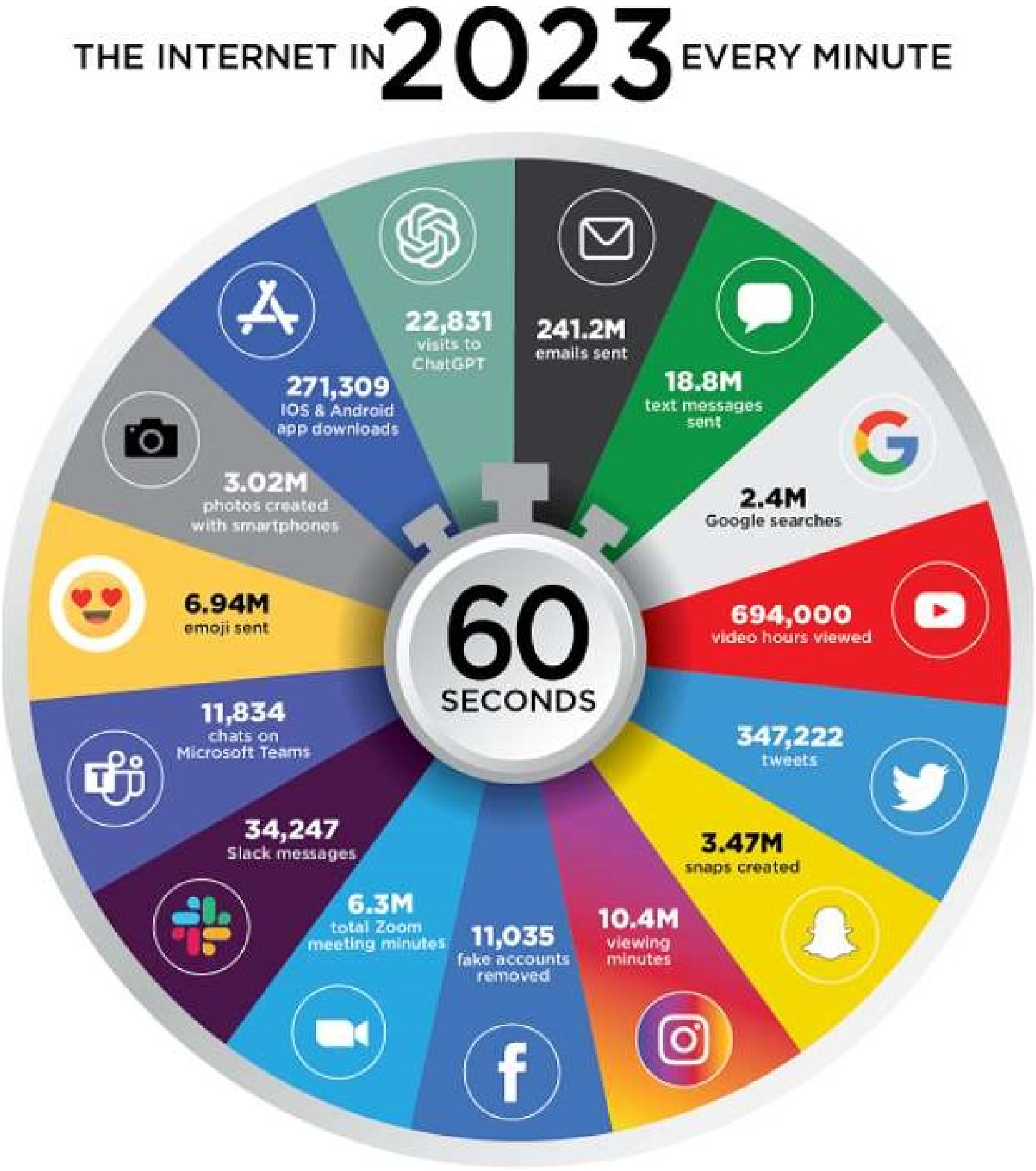
- 4 Petabyte of new data generated every day
- over 300 Petabyte in Facebook‘s data warehouse

Predict

- Interests and behavior of over one billion people

https://www.brandwatch.com/blog/facebook-statistics/
http://www.technologyreview.com/featuredstory/428150/what-facebook-knows/

“We are Drowning in Data...”



Predict

- Interests and behavior of mankind

“We are Drowning in Data...”

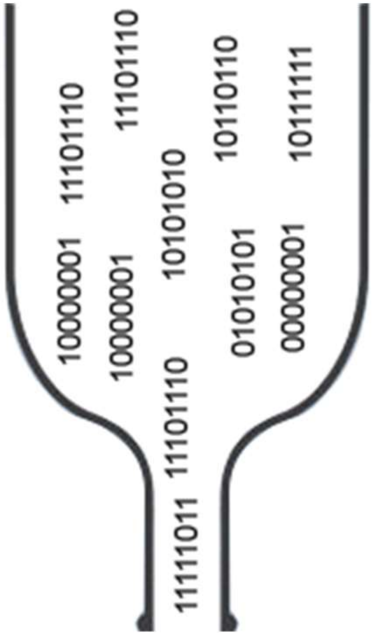
Law enforcement agencies collect unknown amounts of data from various sources

- Cell phone calls
- Location data
- Web browsing behavior
- Credit card transactions
- Online profiles (Facebook)
- …

Predict

- Terrorist or not?
- Social Credit

“We are Drowning in Data ... but starving for knowledge!”



← Amount of data that is collected

← Amount of data that can be looked at by humans

We are interested in the patterns, not the data itself!

Data Mining methods help us to

- discover interesting patterns in large quantities of data
- take decisions based on the patterns

Definitions of Data Mining

- Definitions

Exploration & analysis, of large quantities of data in order to discover meaningful patterns.

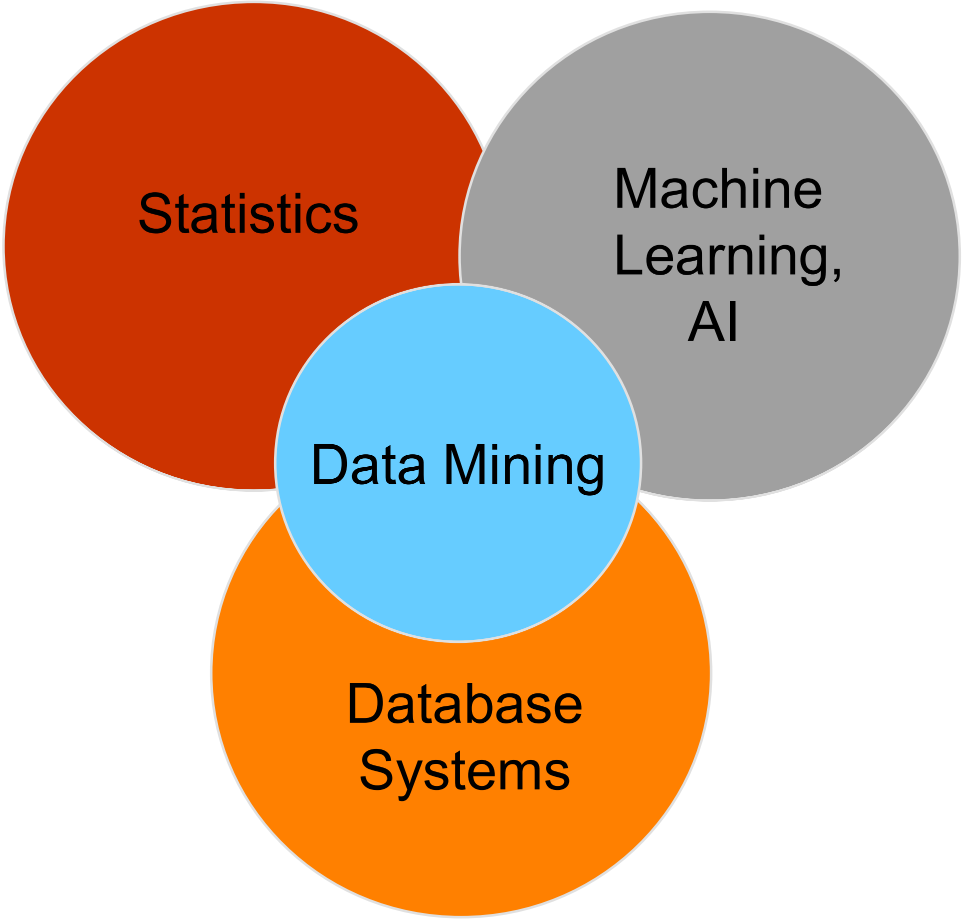
Non-trivial extraction of
- **implicit,**
- **previously unknown, and**
- **potentially useful**

information from data.

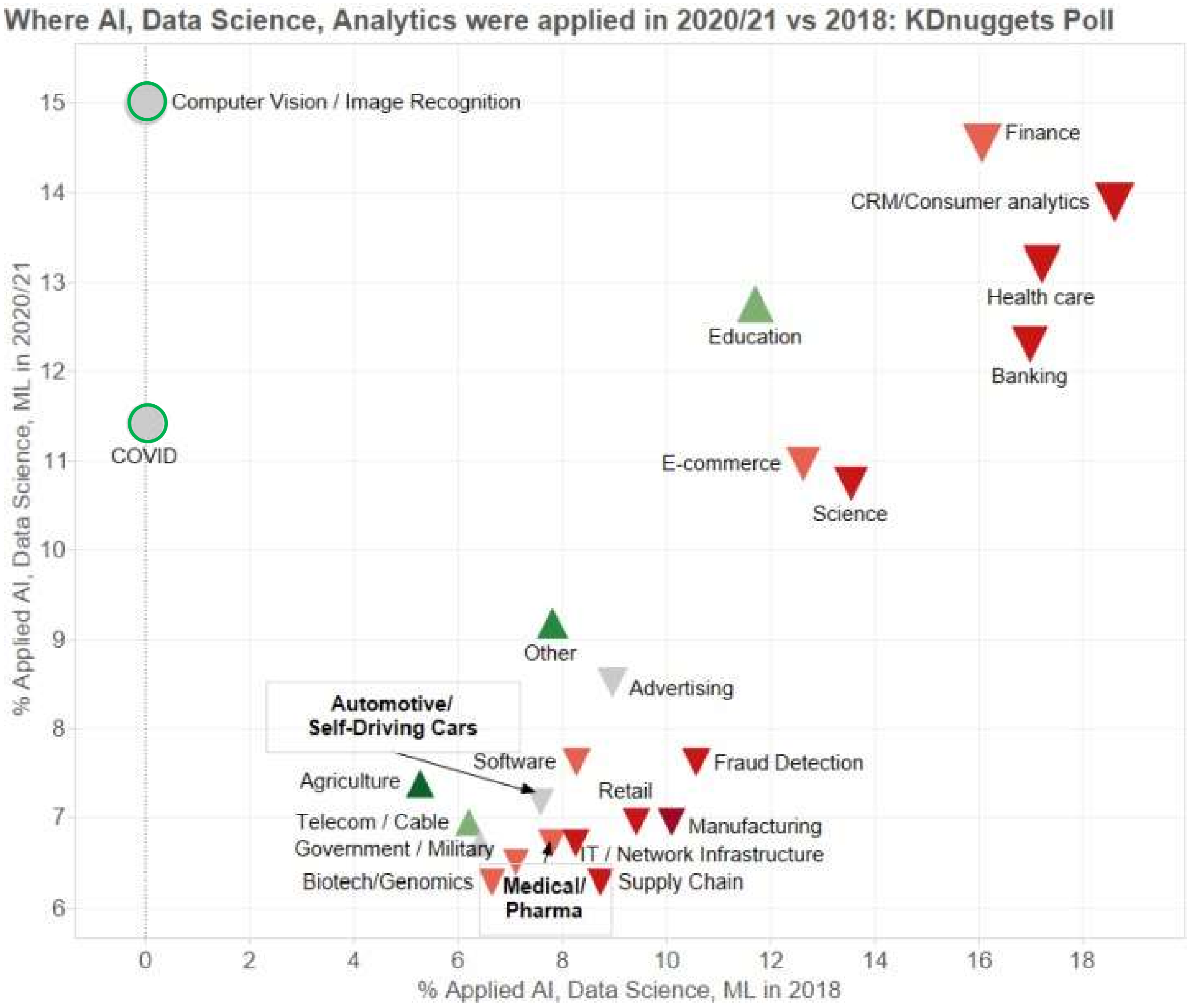
- Data Mining methods
  1. detect interesting patterns in large quantities of data
  2. **support** human decision making by providing such patterns
  3. **predict** the outcome of a future observation based on the patterns

Origins of Data Mining

- Data Mining combines ideas from statistics, machine learning, artificial intelligence, and database systems
- Tries to overcome shortcomings of traditional techniques concerning
  - large amount of data
  - high dimensionality of data
  - heterogeneous and complex nature of data
  - explorative analysis beyond hypothesize-and-test paradigm



Survey on Data Mining Application Fields



Source: KDnuggets online poll, 447 (2021) and 435 (2018) participants
https://www.kdnuggets.com/2021/06/poll-where-analytics-data-science-ml-applied.html

2. Tasks and Applications

- Descriptive Tasks
  - Goal: Find patterns in the data.
  - Example: W*hich products are often bought together?*

- Predictive Tasks
  - Goal: Predict unknown values of a variable
    - given observations (e.g., from the past)
  - Example: W*ill a person click a online advertisement?*
    - given her browsing history

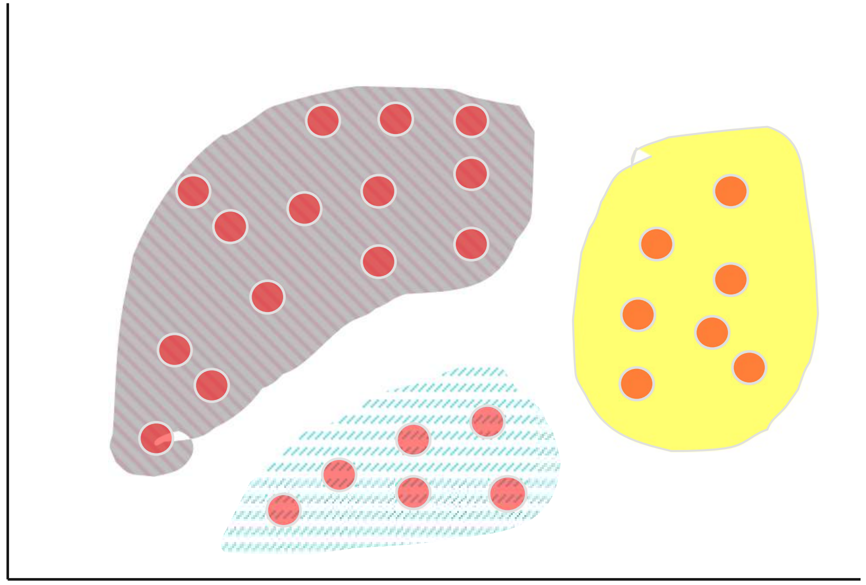
- Machine Learning Terminology
  - descriptive = unsupervised
  - predictive = supervised

Data Mining Tasks

1. Cluster Analysis [Descriptive]
2. Classification [Predictive]
3. Regression [Predictive]
4. Association Analysis [Descriptive]

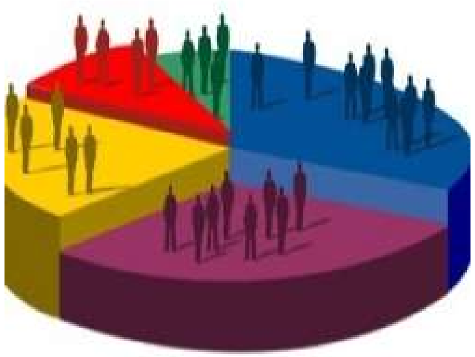
2.1 Cluster Analysis: Definition

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find groups such that
  - data points in one group are more similar to one another
  - data points in separate groups are less similar to one another
- Similarity Measures
  - Euclidean distance if attributes are continuous
  - other task-specific similarity measures
- Goals
  1. intra-cluster distances are minimized
  2. inter-cluster distances are maximized



- Result
  - A descriptive grouping of data points

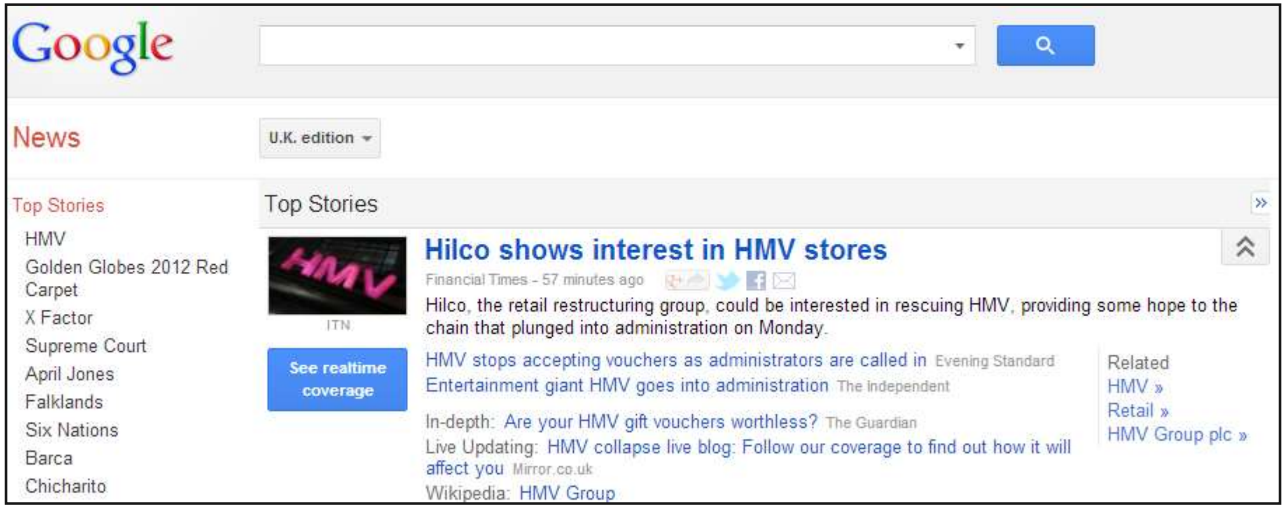
Cluster Analysis: Application 1



- Application area: Market segmentation
- Goal: Find groups of similar customers
  - where a group may be conceived as a marketing target to be reached with a distinct marketing mix
- Approach:
  1. collect information about customers
  2. find clusters of similar customers
  3. measure the clustering quality by observing buying patterns after targeting customers with distinct marketing mixes

Cluster Analysis: Application 2

- Application area: Document Clustering
- Goal: Find groups of documents that are similar to each other based on terms appearing in them
- Approach
  1. identify frequently occurring terms in each document
  2. form a similarity measure based on the frequencies of different terms
- Application Example: Grouping of articles in Google News



Hilco shows interest in HMV stores

Financial Times - 57 minutes ago

Hilco, the retail restructuring group, could be interested in rescuing HMV, providing some hope to the chain that plunged into administration on Monday.

HMV stops accepting vouchers as administrators are called in Evening Standard

Entertainment giant HMV goes into administration The Independent

In-depth: Are your HMV gift vouchers worthless? The Guardian

Live Updating: HMV collapse live blog: Follow our coverage to find out how it will affect you Mirror.co.uk

Wikipedia: HMV Group

2.2 Classification: Definition

- Goal: Previously unseen records should be assigned a class from a given set of classes as accurately as possible.

- Approach:
- Given a collection of records (*training set*)
  - each record contains a set of *attributes*
  - one attribute is the *class attribute (label)* that should be predicted
- Find a *model* for predicting the class attribute as a function of the values of other attributes

Classification: Example

- Training set:

"tree"

"tree"

"tree"

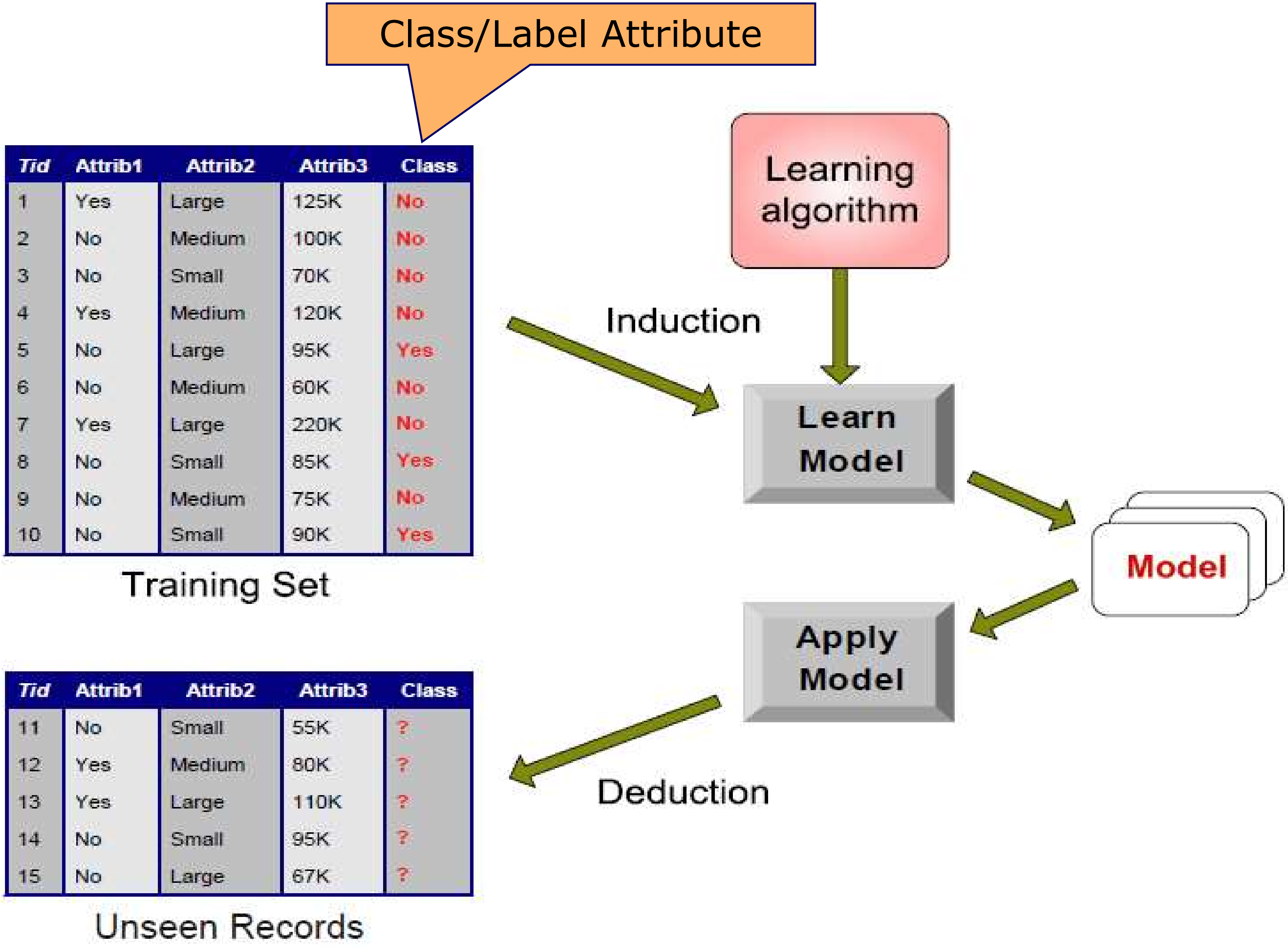
"not a tree"

"not a tree"

"not a tree"

- Learned model: "Trees are big, green plants without wheels."

Classification: Workflow



Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Unseen Records

Classification: Application 1

- Application area: Fraud Detection
- Goal: Predict fraudulent cases in credit card transactions.
- Approach:
  1. Use credit card transactions and information about account-holders as attributes
     - When and where does a customer buy? What does he buy?
     - How often he pays on time? etc.
  2. Label past transactions as fraud or fair transactions This forms the class attribute
  3. Learn a model for the class attribute from the transactions
  4. Use this model to detect fraud by observing credit card transactions on an account

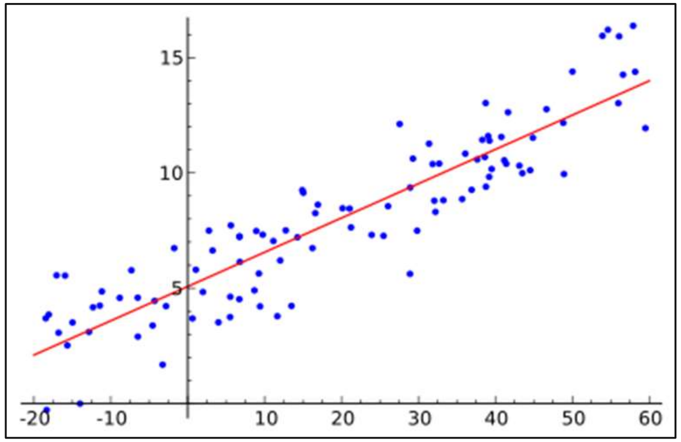
Classification: Application 2

- Application area: Direct Marketing

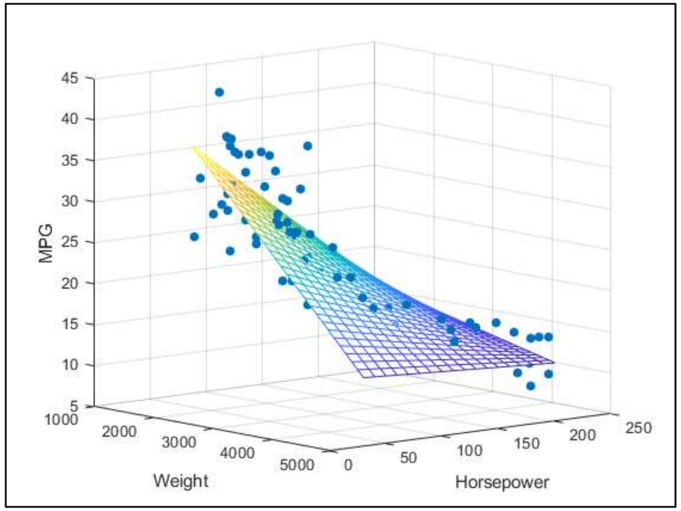
- Goal: Reduce cost of a mailing campaign by targeting only the set of consumers that likely to buy a new product
- Approach:
  1. Use data from a campaign introducing a similar product in the past
     - we know which customers decided to buy and which decided otherwise
     - this {buy, don’t buy} decision forms the class attribute
  2. Collect various demographic, lifestyle, and company-interaction related information about the customers
     - age, profession, location, income, marriage status, visits, logins, etc.
  3. Use this information to learn a classification model
  4. Apply model to decide which consumers to target

2.3 Regression

- Predict a value of a continuous variable based on the values of other variables, assuming a linear or nonlinear model of dependency



- Examples:
  - Predicting the price of a house or car
  - Predicting sales amounts of new product based on advertising expenditure
  - Predicting miles per gallon (MPG) of a car as a function of its weight and horsepower
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.



- Difference to classification: The predicted attribute is continuous, while classification is used to predict nominal attributes (e.g. *yes/no*)

2.4 Association Analysis: Definition

- Given a set of records each of which contain some number of items from a given collection
- discover frequent itemsets and produce association rules which will predict occurrence of an item based on occurrences of other items

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Frequent Itemsets

{Diaper, Milk, Beer}

{Milk, Coke}

Association Rules

{Diaper, Milk} --> {Beer}

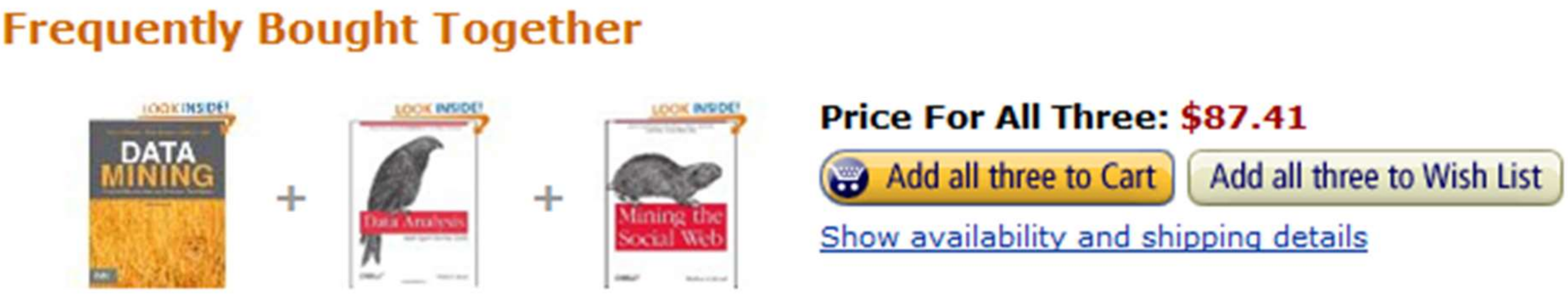
{Milk} --> {Coke}

Association Rule Discovery: Applications 1

- Application area: Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items
  - A classic rule and its implications:
    - if a customer buys diapers and milk, then he is likely to buy beer as well
    - so, don't be surprised if you find six-packs stacked next to diapers!
    - promote diapers to boost beer sales
    - if selling diapers is discontinued, this will affect beer sales as well

- Application area: Sales Promotion

amazon.com



Frequently Bought Together

Price For All Three: $87.41

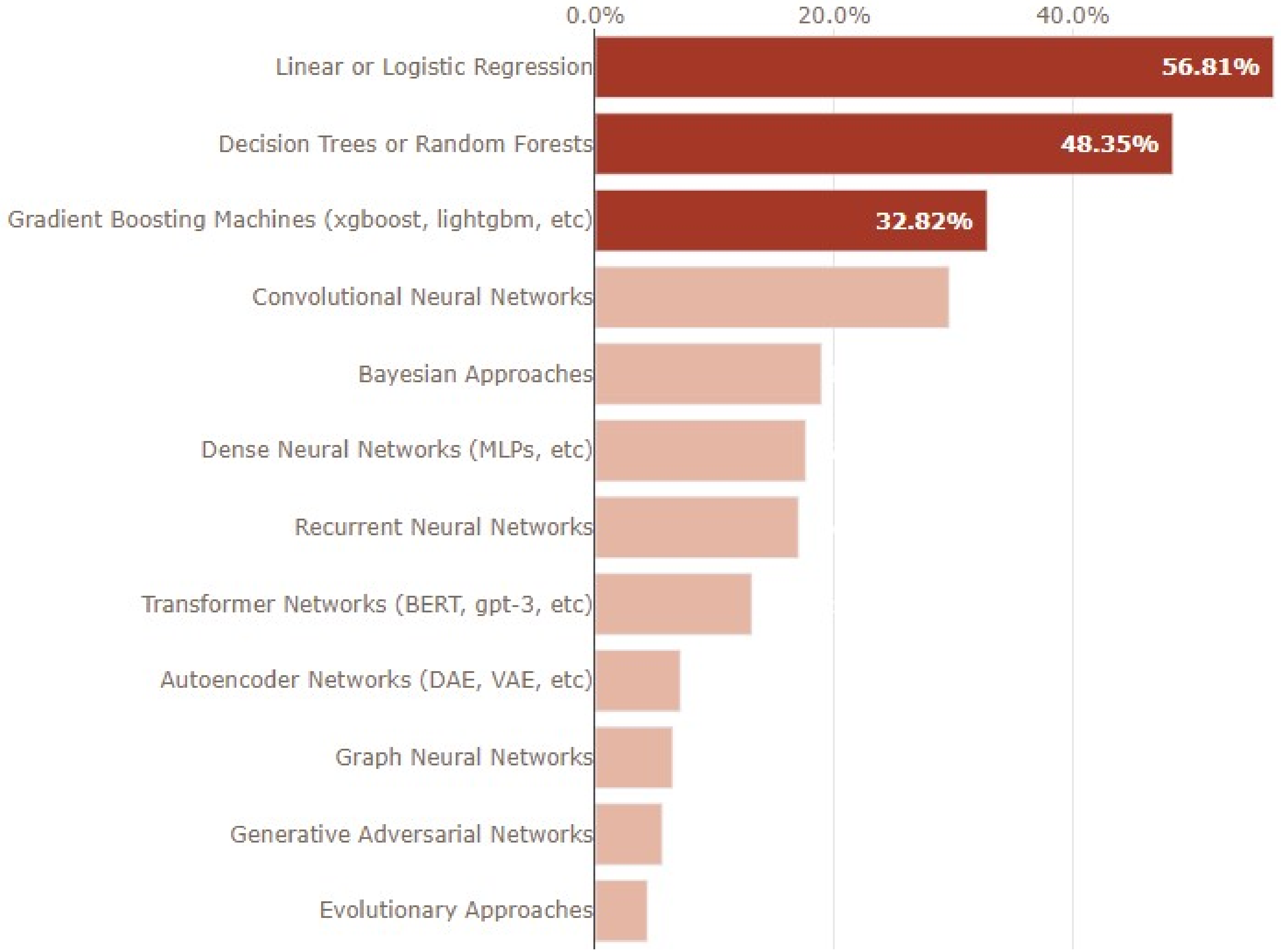
Association Rule Discovery: Application 2

- Application area:
  Inventory Management

- Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households

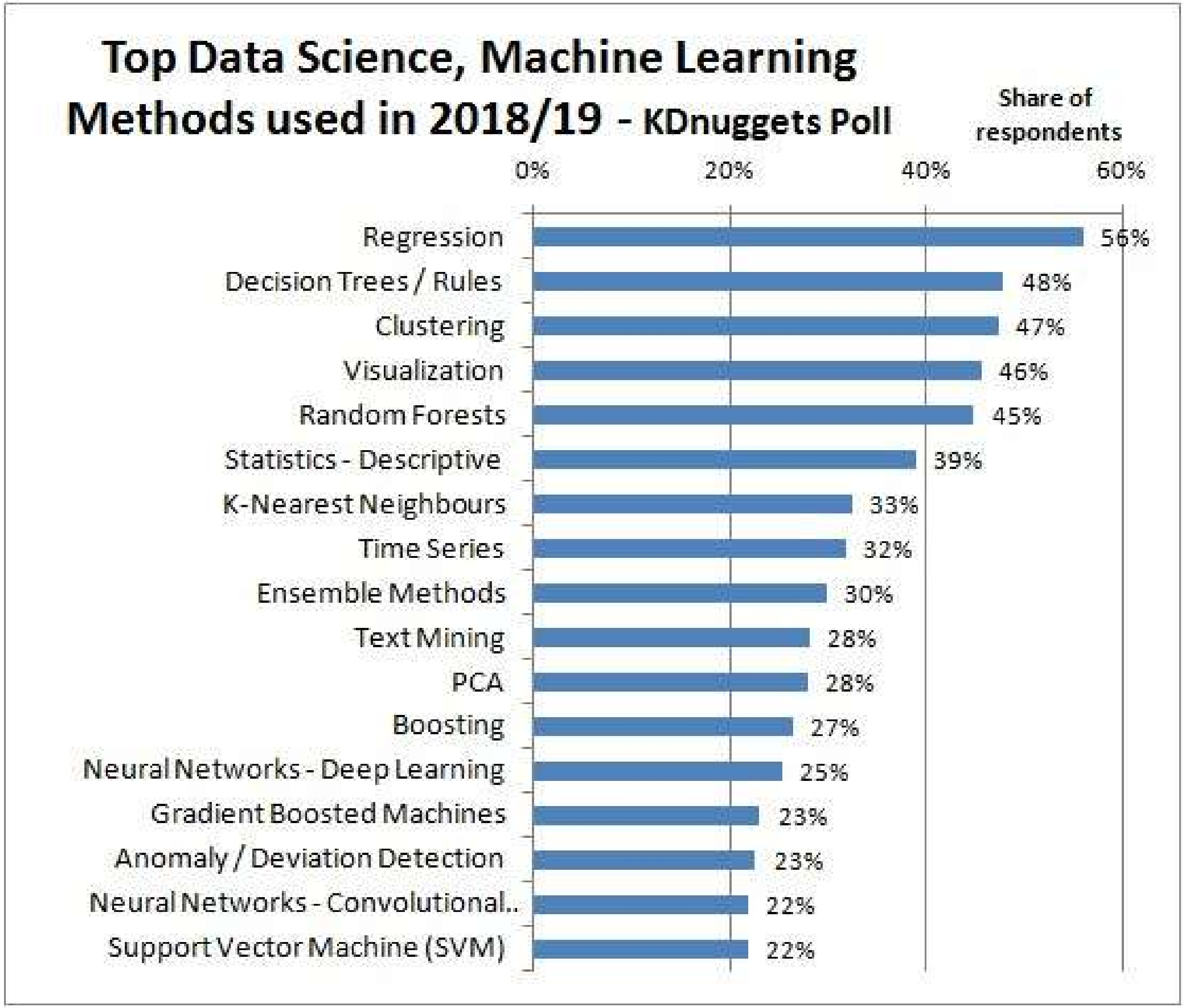
- Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns

Which Methods are Used in Practice?



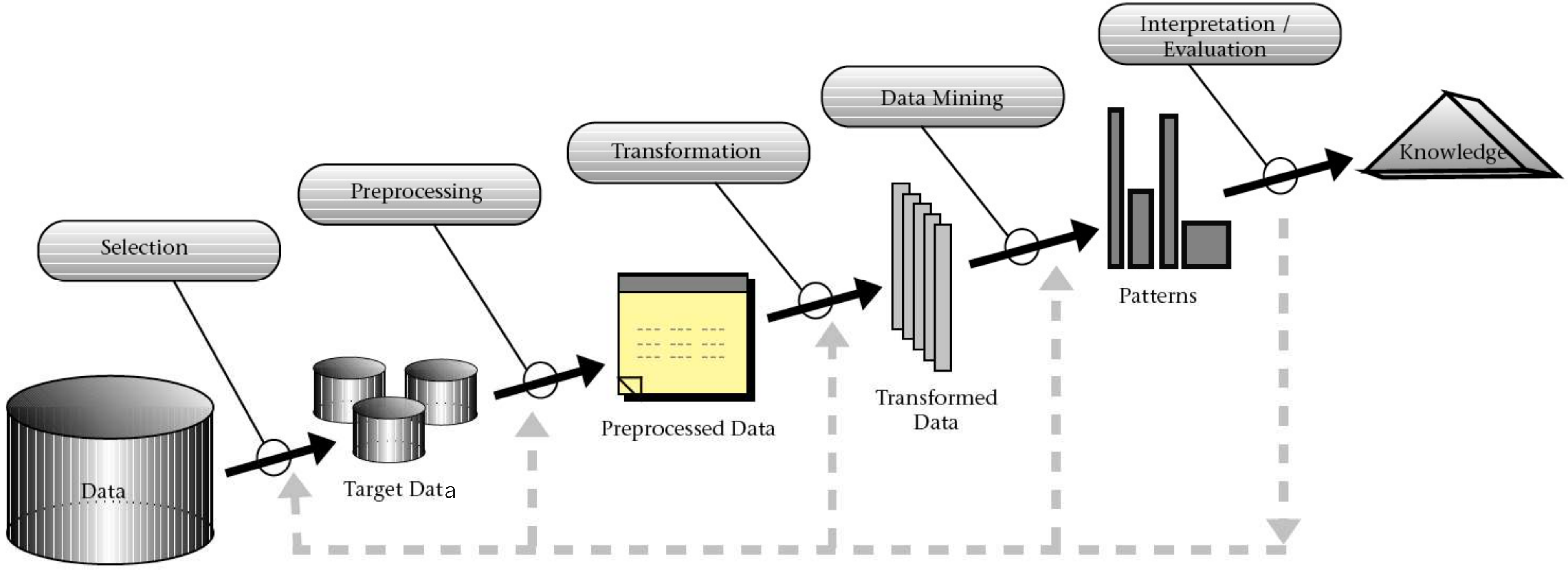
Source: Kaggle online poll 2022, 23,997 respondents,
https://www.kaggle.com/code/eraikako/data-science-and-mlops-landscape-in-industry

Which Methods are Used in Practice?



Source: KDnuggets online poll, 833 votes, question: methods used last year for real-world app?
https://www.kdnuggets.com/2019/04/top-data-science-machine-learning-methods-2018-2019.html

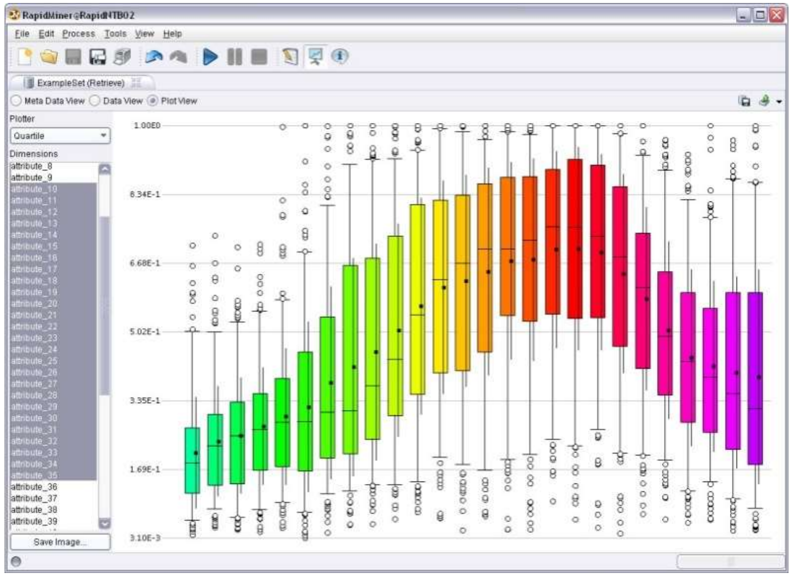
3. The Data Mining Process

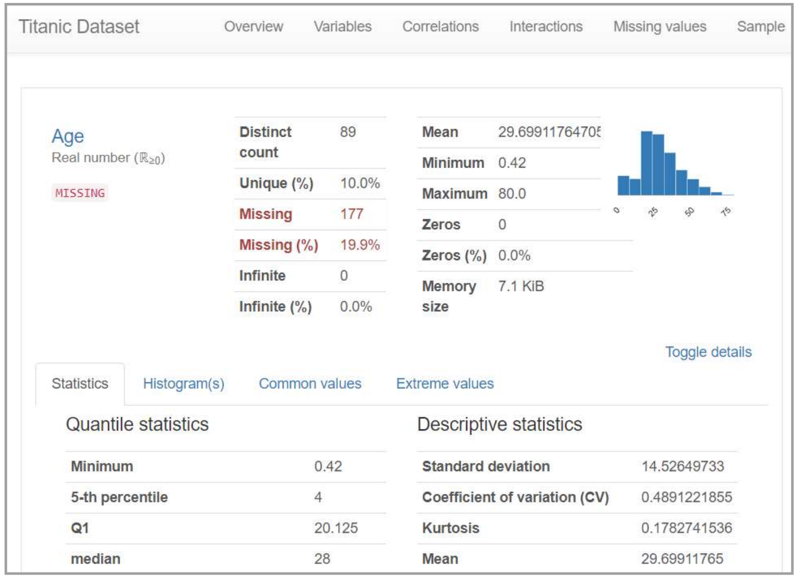


Source: Fayyad et al. (1996)

3.1 Selection and Exploration

- Selection
  - What data is potentially useful for the task at hand?
  - What data is available?
  - What do I know about the quality of the data?
- Exploration / Profiling
  - Get an initial understanding of the data
  - Calculate basic summarization statistics
  - Visualize the data
  - Identify data problems such as outliers, missing values, duplicate records




Titanic Dataset — Overview | Variables | Correlations | Interactions | Missing values | Sample

Age
Real number ($\mathbb{R}_{\geq 0}$)
MISSING

Distinct count	89	Mean	29.69911764705
Unique (%)	10.0%	Minimum	0.42
Missing	177	Maximum	80.0
Missing (%)	19.9%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	7.1 KiB

Toggle details

Statistics | Histogram(s) | Common values | Extreme values

Quantile statistics

Minimum	0.42
5-th percentile	4
Q1	20.125
median	28

Descriptive statistics

Standard deviation	14.52649733
Coefficient of variation (CV)	0.4891221855
Kurtosis	0.1782741536
Mean	29.69911765

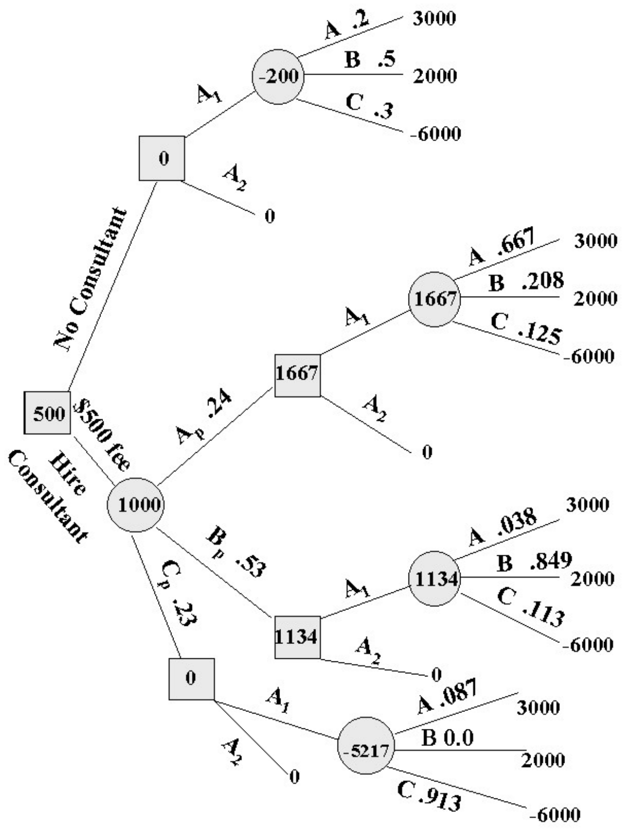
3.2 Preprocessing and Transformation

- Transform data into a representation that is suitable for the chosen data mining methods
  - scales of attributes (nominal, ordinal, numeric)
  - number of dimensions (represent relevant information using less attributes)
  - amount of data (determines hardware requirements)
- Methods
  - discretization and binarization
  - feature subset selection / dimensionality reduction
  - attribute transformation / text to term vector / embeddings
  - aggregation, sampling
  - integrate data from multiple sources
- Good data preparation is key to producing valid and reliable models
- Data integration and preparation is estimated to take 70-80% of the time and effort of a data mining project

3.3 Data Mining

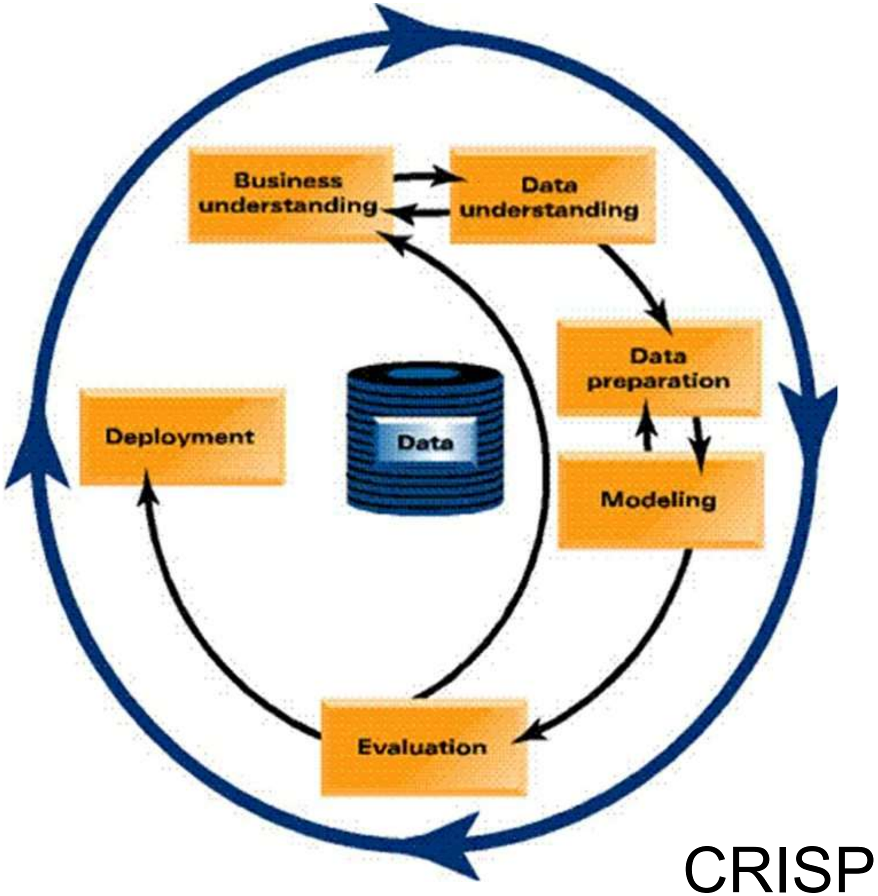
- Input: Preprocessed Data
- Output: Model / Patterns

1. Apply data mining method
2. Evaluate resulting model / patterns
3. **Iterate**
   - experiment with different hyperparameter settings
   - experiment with multiple alternative methods
   - improve preprocessing and feature generation
   - increase amount or quality of training data



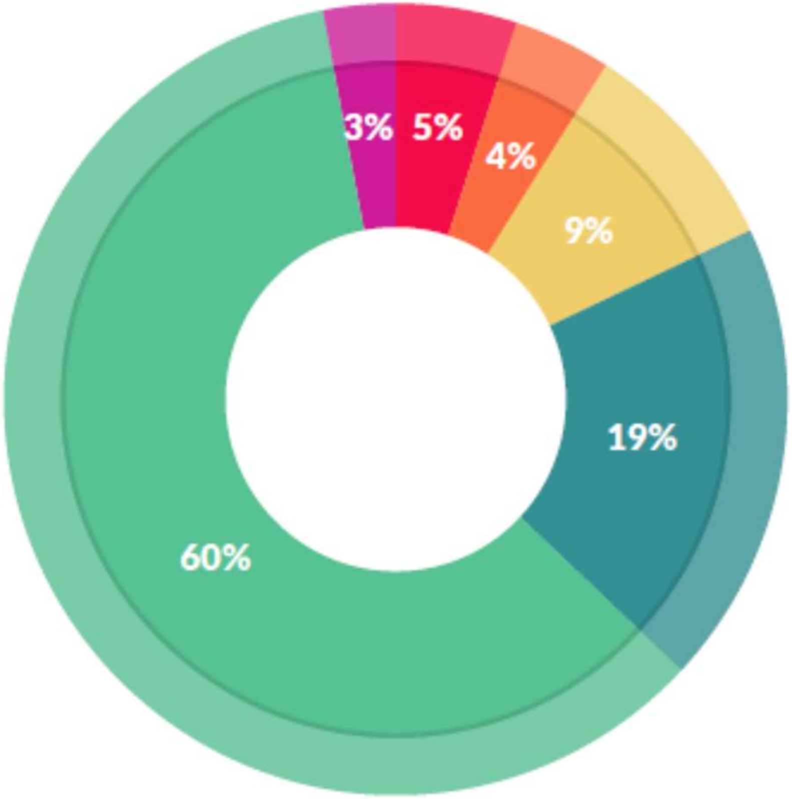
3.4 Deployment

- Use model in the business context
- Keep iterating in order to maintain and improve model



CRISP-DM Process Model

How Do Data Scientists Spend Their Days?



What data scientists spend the most time doing

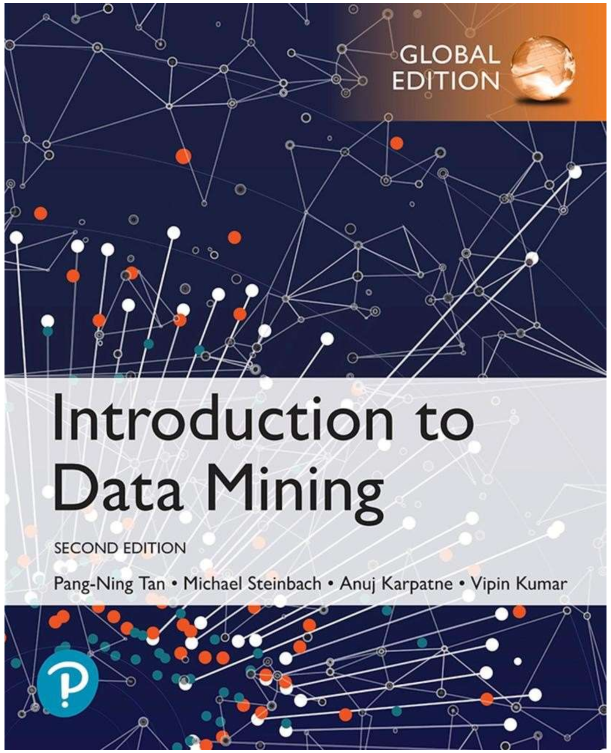
- ● *Building training sets: 3%*
- ● *Cleaning and organizing data: 60%*
- ● *Collecting data sets; 19%*
- ● *Mining data for patterns: 9%*
- ● *Refining algorithms: 4%*
- ● *Other: 5%*

Advertisement:
Thus, attend
IE670
Web Data
Integration ☺

Source: CrowdFlower Data Science Report 2016: http://visit.crowdflower.com/data-science-report.html

Literature for this Chapter

Pang-Ning Tan, Michael Steinbach, Vipin Kumar:
Introduction to Data Mining. 2nd Edition.
Pearson / Addison Wesley.

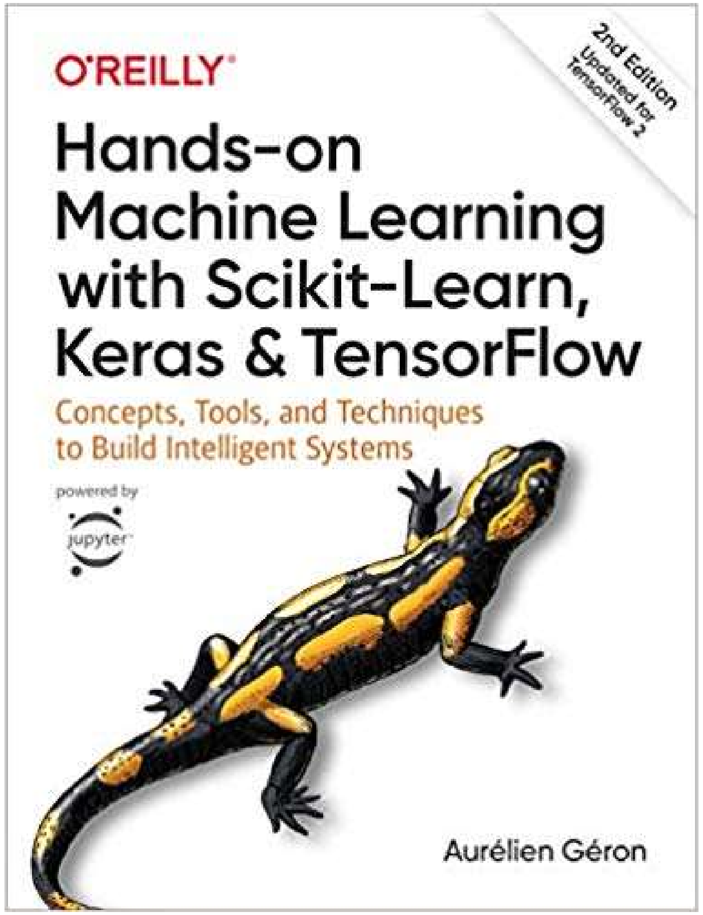

GLOBAL EDITION
Introduction to Data Mining
SECOND EDITION
Pang-Ning Tan • Michael Steinbach • Anuj Karpatne • Vipin Kumar

Chapter 1: Introduction

Chapter 2: Data

Literature – Python

1. **Scikit-learn Documentation:** https://scikit-learn.org/stable/user_guide.html
2. Aurélien Géron: **Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow.**
   2nd Edition, O'Reilly, 2019


O'REILLY
2nd Edition
Updated for TensorFlow 2
Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow
Concepts, Tools, and Techniques to Build Intelligent Systems
powered by
jupyter
Aurélien Géron

Thank you!

- Are there any questions?

- Next …
  1. install the **Anaconda Python** distribution
  2. attend the tutorial **Introduction to Python** today
  3. get an **OpenAI account** for using ChatGPT
  4. attend exercise **Preprocessing/Visualization** on Thursday
  5. attend the lecture **Cluster Analysis** next week
  6. …