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
- Phone: +49 621 181 2677
- eMail: chris@informatik.uni-mannheim.de

Hallo

- **M. Sc. Wi-Inf. Anna Primpeli**
- Graduate Research Associate
- Research Interests:
  - Semantic Annotations in Web Pages
  - Active Learning for Identity Resolution
  - Product Data Integration
- Room: B6, 26, C 1.04
- eMail: anna@informatik.uni-mannheim.de

- Will teach the RapidMiner exercises 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@informatik.uni-mannheim.de

- Will teach one Python exercise group and will supervise student projects.

Hallo

- **M. Sc. Wi-Inf. Alexander Brinkmann**
- Graduate Research Associate
- Research Interests:
  - Data Search using Deep Learning
  - Product Data Categorization
- Room: B6, 26, C 1.03
- eMail: alex.brinkmann@informatik.uni-mannheim.de

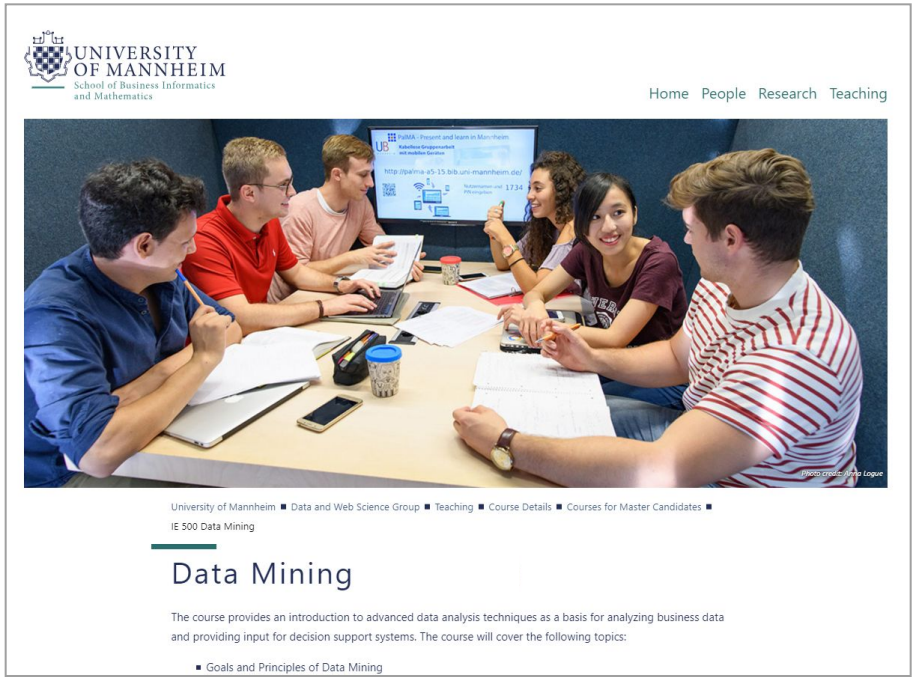
- Will teach one Python exercise group and will supervise student projects.

Course Organisation

- Lecture
  - introduces the principle methods of data mining
  - discusses how to evaluate generated models
  - presents practical examples of data mining applications from the corporate and Web context
- Exercise Groups
  - students experiment with the methods using RapidMiner or 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/
- Solutions to the Exercises
  - ILIAS eLearning System, https://ilias.uni-mannheim.de/
- Time and Location
  - Lecture:
    - Wednesday, 10.15 - 11.45, A5, **ZOOM 04**
  - Exercise:
    - Thursday, 10.15 - 11.45
      Room **ZOOM 13**(RapidMiner, Anna)
    - Thursday, 12.00 - 13.30,
      Room **ZOOM 16**(Python, Ralph)
    - Thursday, 13.45 - 15.15,
      Room **ZOOM 15**(Python, Alex)



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, RapidMiner Text Extension
6. Association Analysis	Frequent Item Set Generation, Rule Generation, Interestingness Measures

Schedule

Bold = session takes place live via ZOOM

Week	Wednesday	Thursday
3.03.2021	**Introduction to Data Mining Introduction to Python**	**Exercise Preprocessing/Visualization**
10.03.2021	Lecture Clustering	**Exercise Clustering**
17.03.2021	Lecture Classification 1	**Exercise Classification**
24.03.2021	Lecture Classification 2	**Exercise Classification**
14.04.2021	Lecture Classification 3	**Exercise Classification**
21.04.2021	Video Lecture Regression	**Exercise Regression**
28.04.2021	Video Lecture Text Mining	**Exercise Text Mining**
5.05.2021	Video Lecture Association Analysis	**Exercise Association Analysis**
12.05.2021	**Introduction to the Student Projects and Group Formation**	Preparation of Project Outlines
19.05.2021	**Feedback on Project Outlines**	Project Work
26.05.2021	Project Work	**Feedback on demand**
2.06.2021	**Feedback on demand**	Project Work
13.06.2021	Submission of project report	Preparation of presentation
16.06.2021	**Presentation of project results**	**Presentation of project results**

Extra tutorial session: Introduction to Python

For all students which are not familiar with Python and Jupyter Notebooks, Ralph and Alex will offer **one** additional tutorial session.

When? → Today, Wednesday 03.03 at 15:30-17:00

Where? → Online, WIM-ZOOM Room 13

https://uni-mannheim.zoom.us/j/5630194600?pwd=ZWM2MlovRkM2RnJXQ2tIeU0zcUV3Zz09

The slides and notebooks used for the exercise will be uploaded in ILIAS and the webpage of the course.

Deadlines

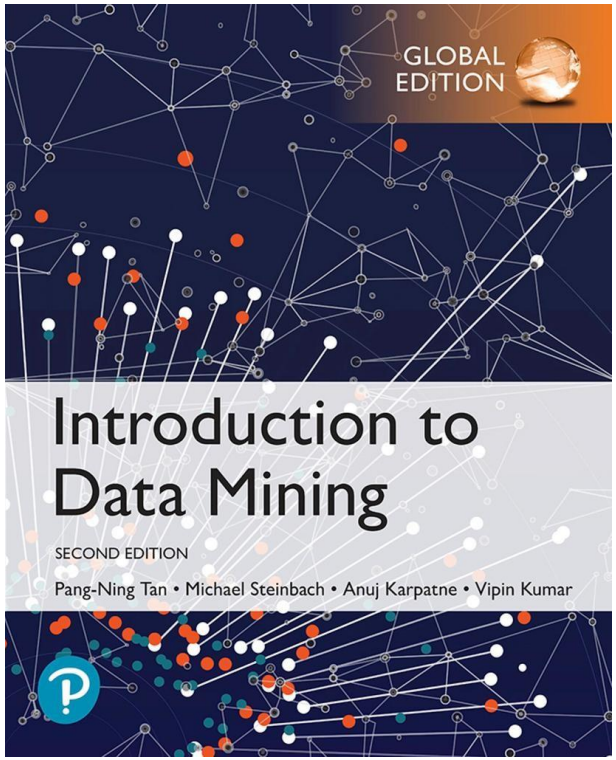
- Submission of project proposal
  - Sunday, May 16th, 23:59
- Submission of final project report
  - Sunday, June 13th, 23:59
- Project presentations
  - Wednesday June 16th, Thursday, June 17th
  - everyone has to attend the presentations

Final Exam

- Date and Time: tbd
- Room: tba
- Duration: 60 minutes
- Structure: 6 open questions that
  - Goal is to check whether you have understood the lecture content
    - we try to cover all major chapters of the lecture: clustering, classification, regression, association analysis, text mining
  - Require you to describe the ideas behind algorithms and methods
    - often: How do methods react to special pattern 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

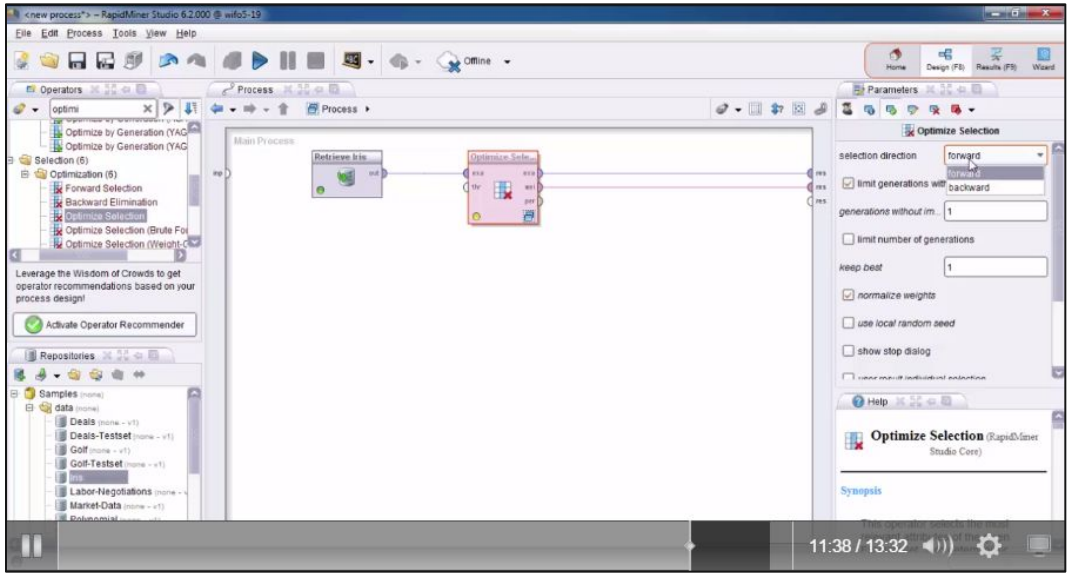
Text Book for the Course

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



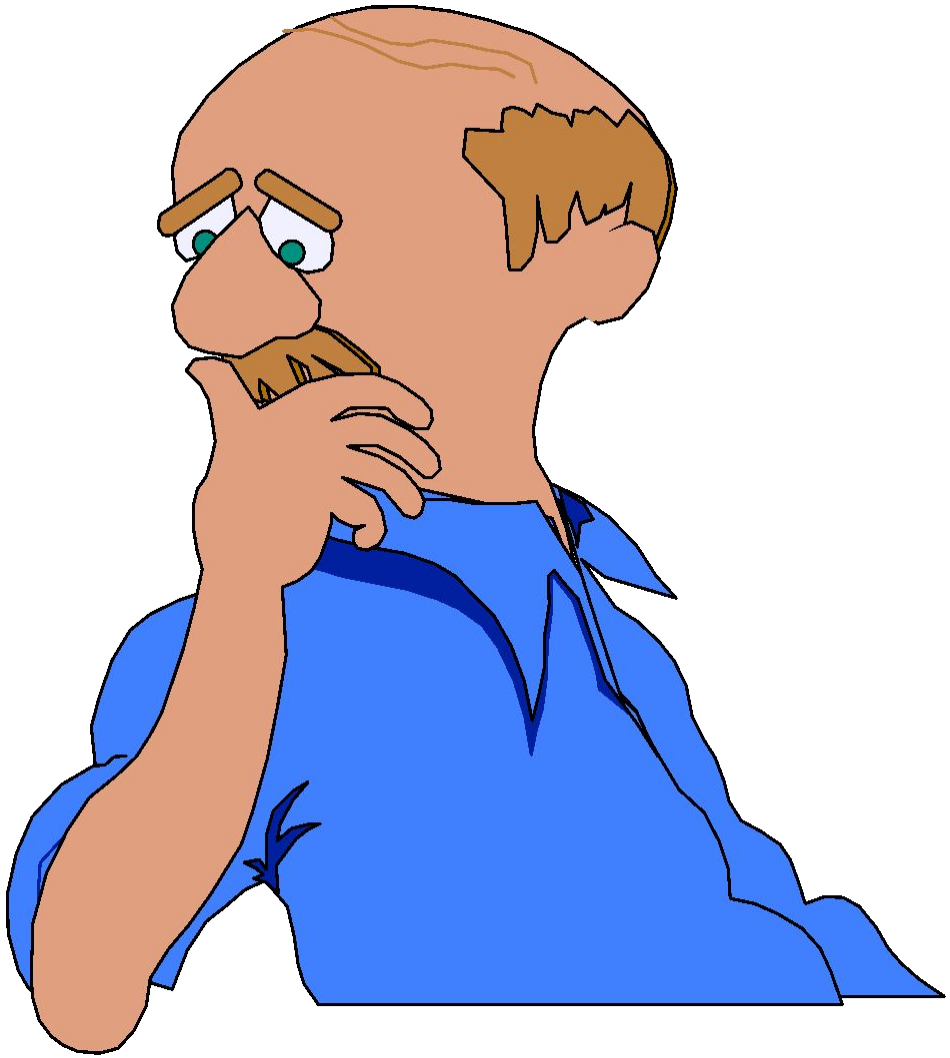
Lecture Videos and Screencasts

1. Video recordings of the lectures from FSS 2020
2. Step-by-step introduction to relevant RapidMiner features
3. Step-by-step solutions of the RapidMiner exercises



http://dws.informatik.uni-mannheim.de/en/teaching/lecture-videos/

Questions?

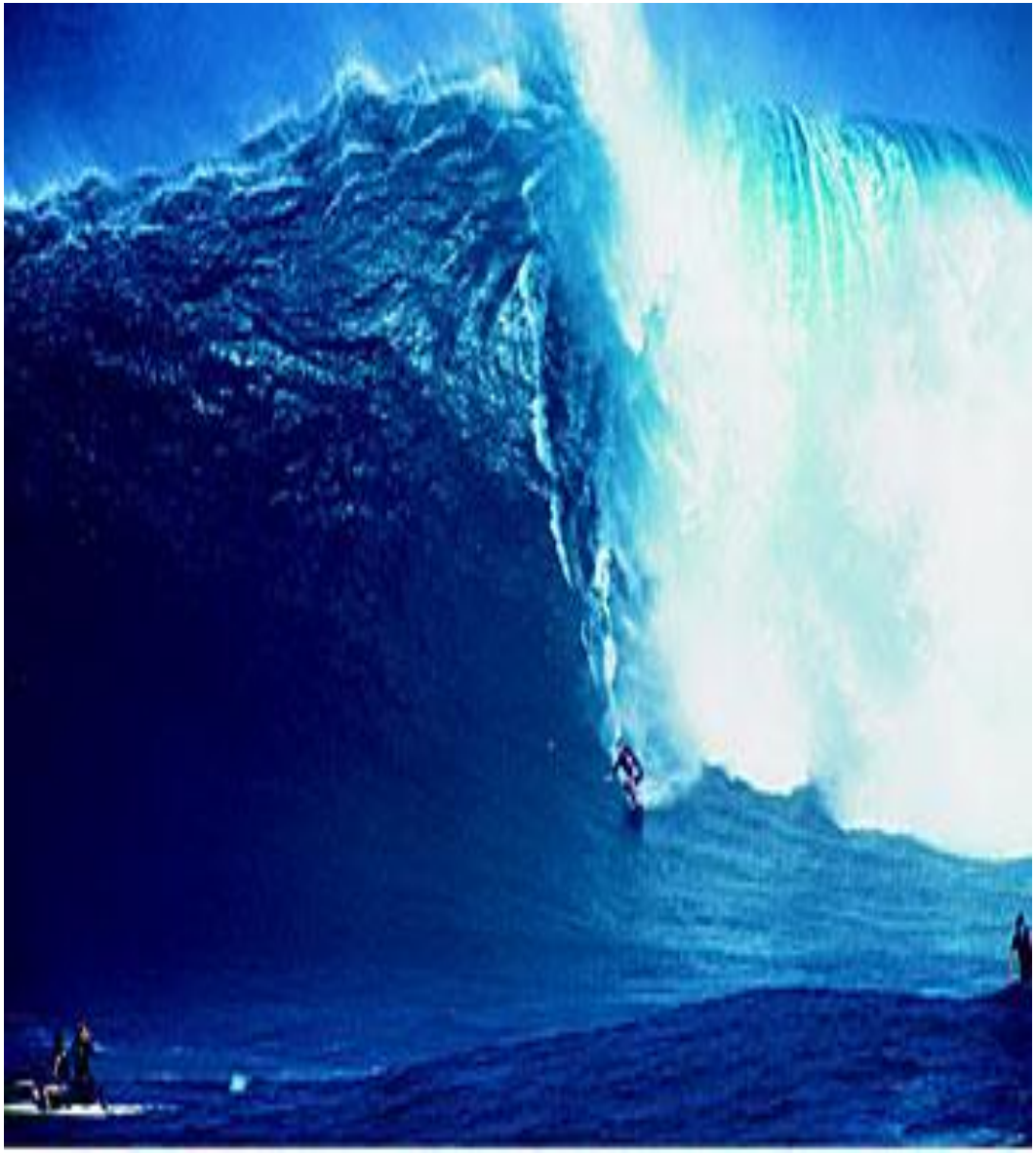


Outline: Introduction to Data Mining

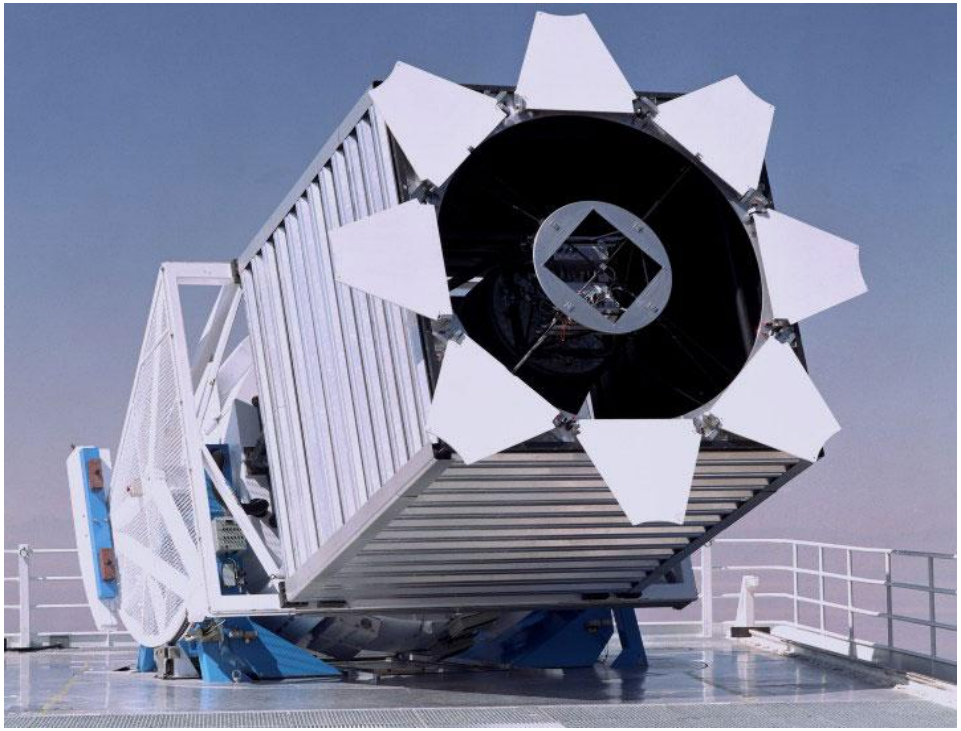
1. What is Data Mining?
2. Tasks and Applications
3. The Data Mining Process
4. Data Mining Software

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



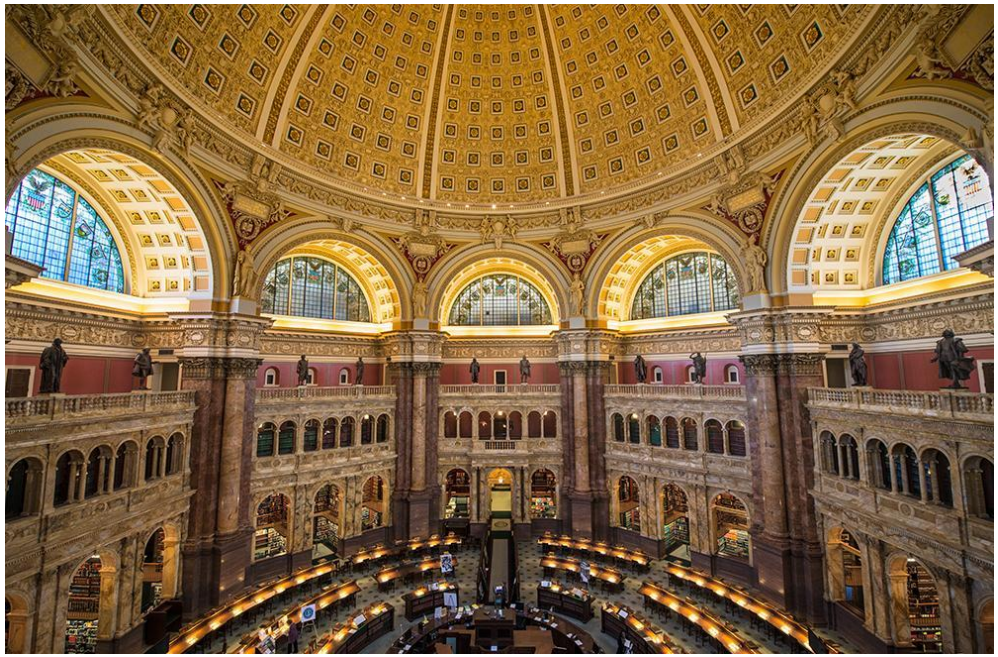
“We are Drowning in Data...”



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

Discover
- Topic distributions
- Historic trends*
- Citation networks

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

“We are Drowning in Data...”

Facebook

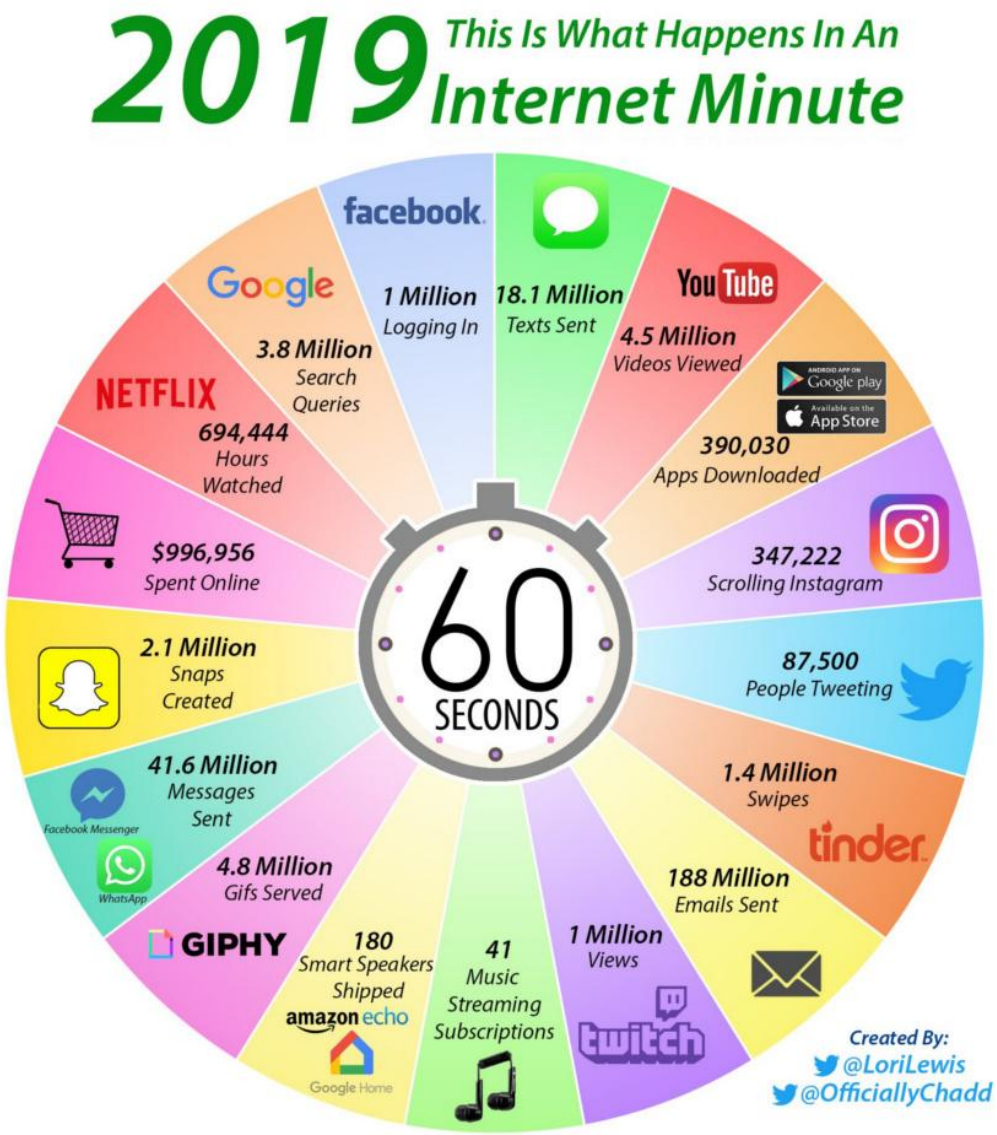
- 4 Petabyte of new data generated every day
- over 300 Petabyte in Facebook‘s data warehouse

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

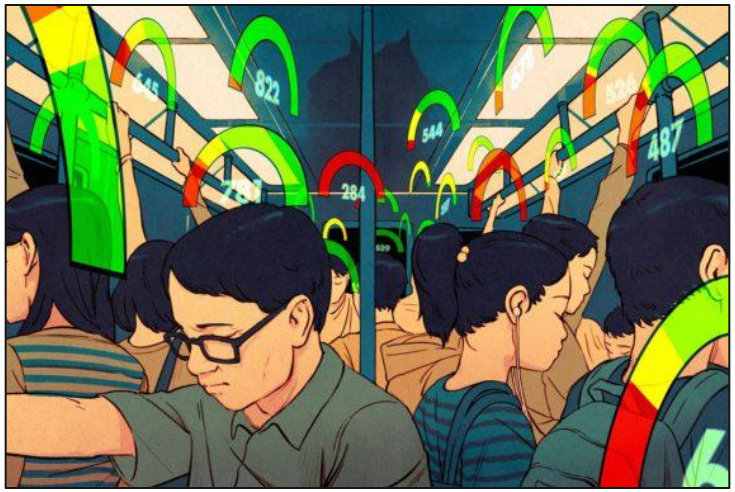
“We are Drowning in Data...”

Law enforcement agencies collect unknown amounts of data from various sources

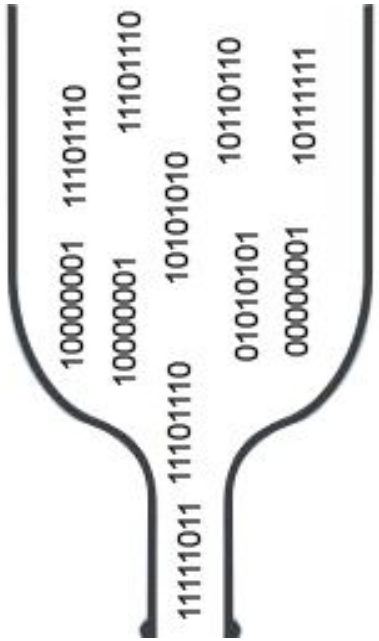
- Cell phone calls
- Location data
- Web browsing behavior
- Credit card transactions
- Online profiles (Facebook)
- …

Predict

- Terrorist or not?
- Trustworthiness



“...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

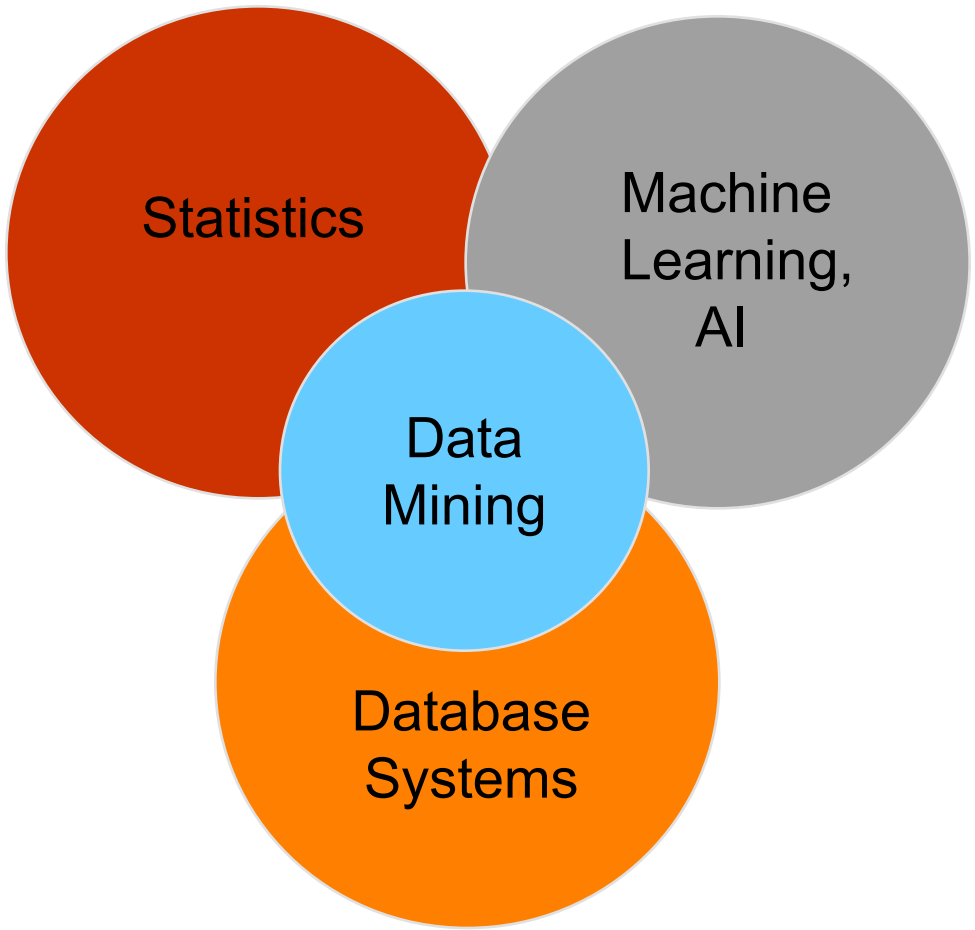
- **implicit,**
- **previously unknown, and**
- **potentially useful**

information from data.

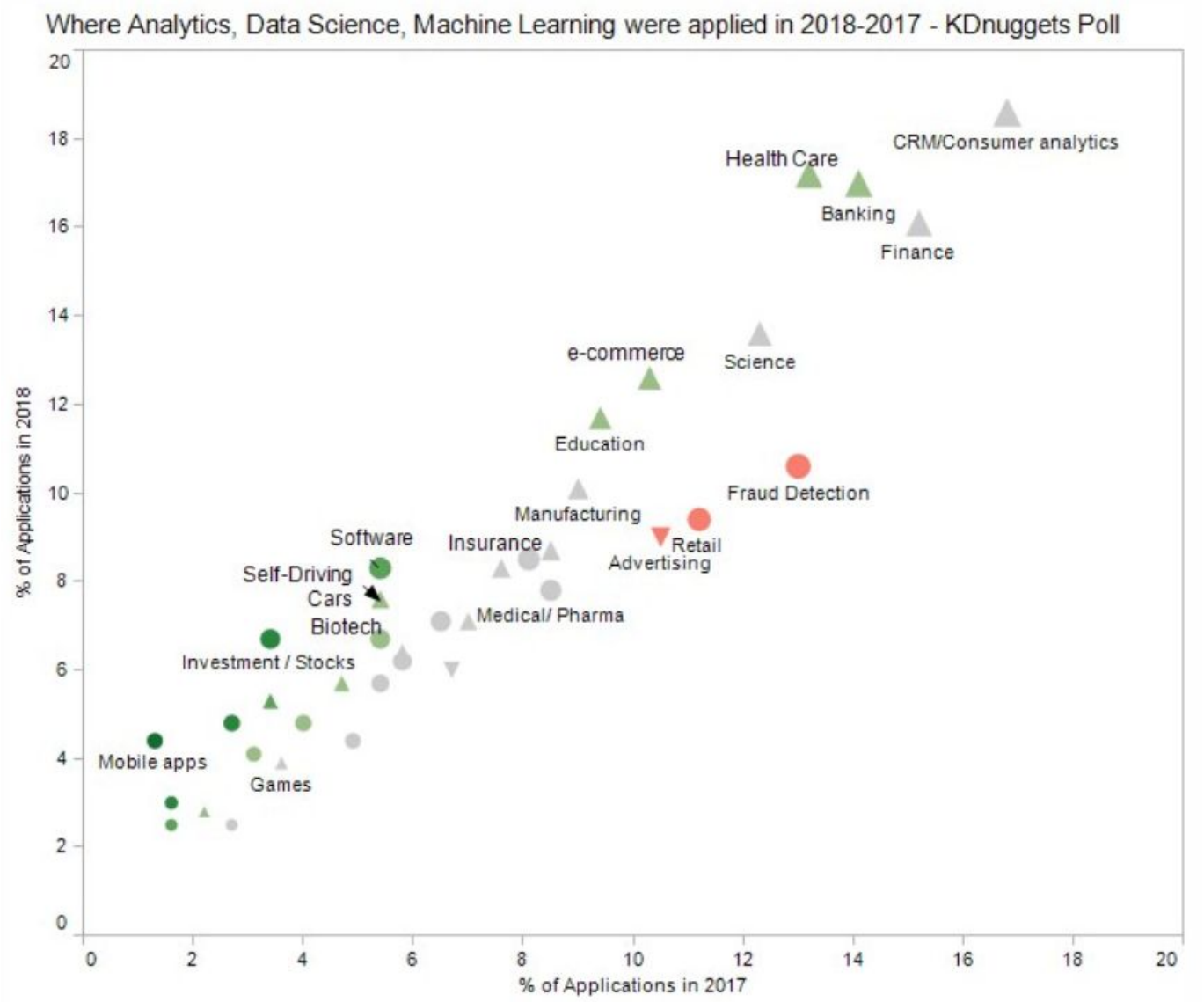
- Data Mining methods
  1. detect interesting patterns in large quantities of data
  2. **support** human decision making by providing such patterns
  3. **predict** the outcome of a future observation based on the patterns

Origins of Data Mining

- Data Mining combines ideas from statistics, machine learning, artificial intelligence, and database systems
- Tries to overcome short-comings 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, 435 and 446 participants
https://www.kdnuggets.com/2019/03/poll-analytics-data-science-ml-applied-2018.html

2. Tasks and Applications

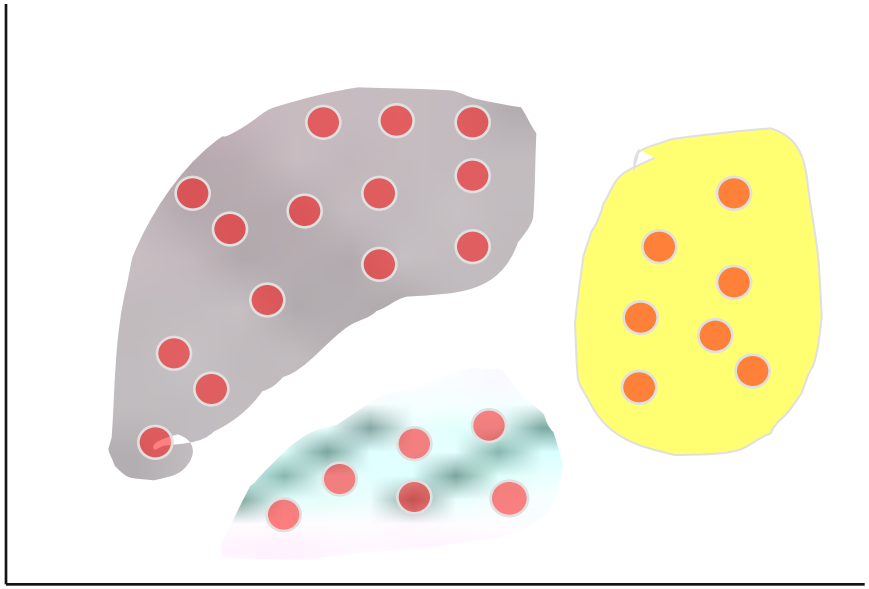
- Descriptive Tasks
  - Goal: Find patterns in the data.
  - Example: W*hich products are often bought together?*
- Predictive Tasks
  - Goal: Predict unknown values of a variable
    - given observations (e.g., from the past)
  - Example: W*ill a person click a online advertisement?*
    - given her 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

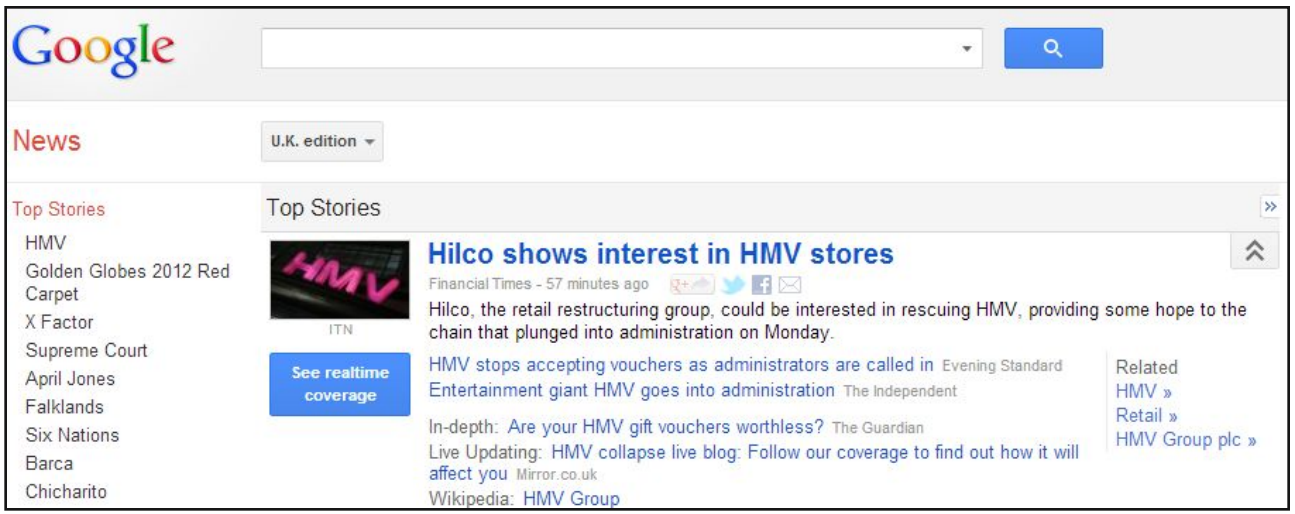


Cluster Analysis: Application 1

- 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



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"

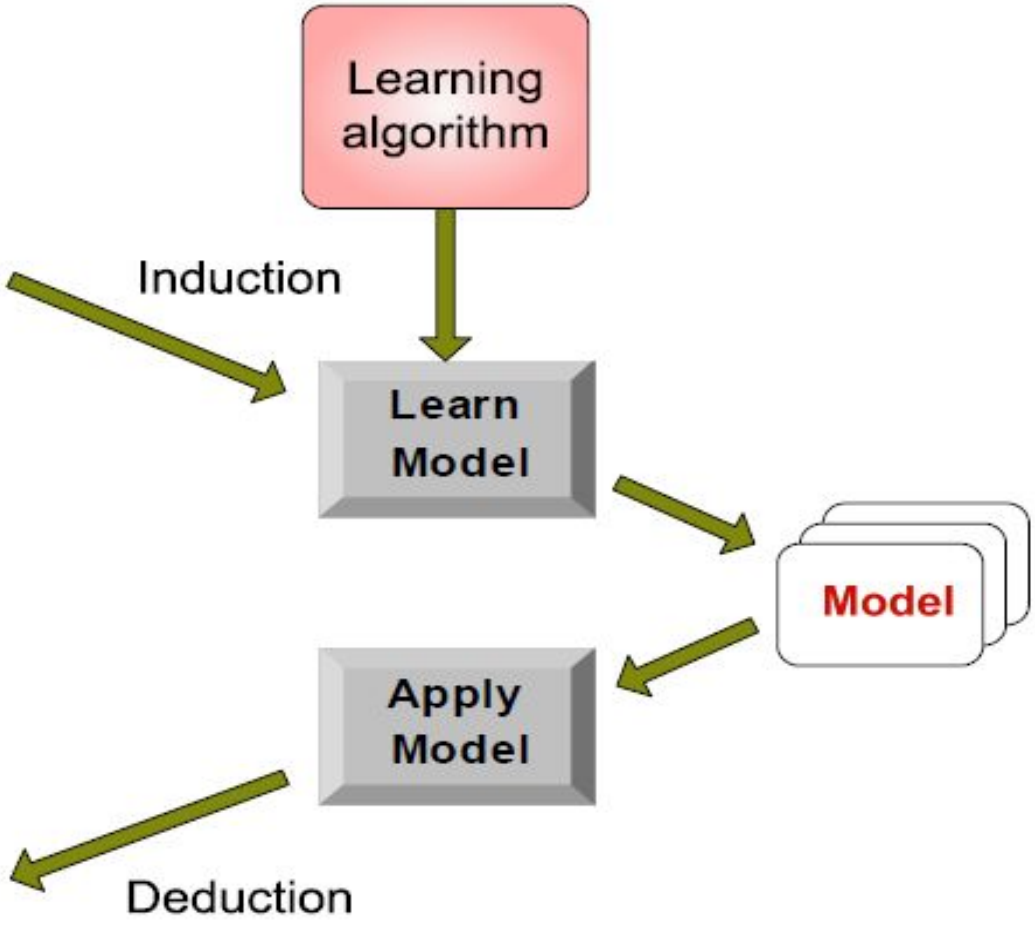
- Learned model: "Trees are big, green plants without wheels."

Classification: Workflow

Class/Label Attribute

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.
  - Label past transactions as fraud or fair transactions
    This forms the class attribute
  1. Learn a model for the class attribute from the transactions
  - Use this model to detect fraud by observing credit card transactions on an account

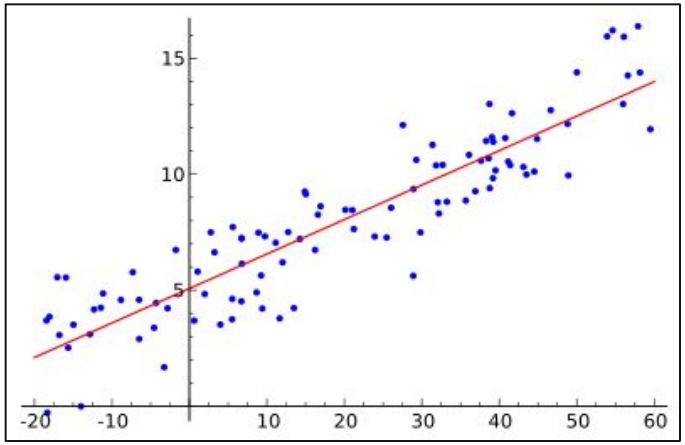
Classification: Application 2

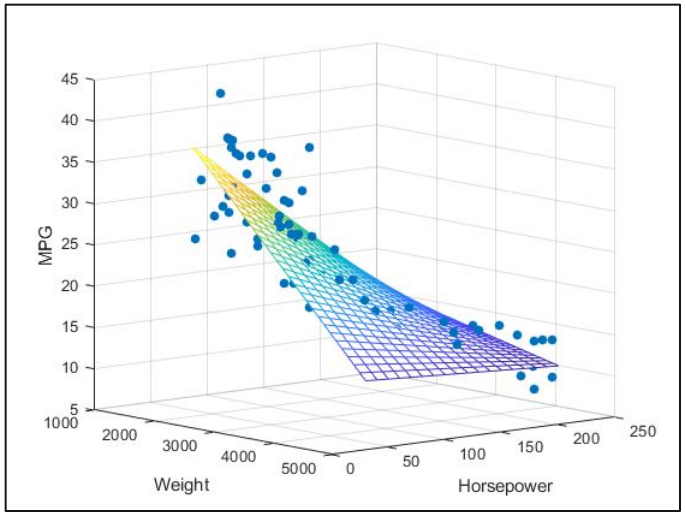
- Application area: Direct Marketing
- Goal: Reduce cost of a mailing campaign by targeting only the set of consumers that likely to buy a new product
- Approach:

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
   - Collect various demographic, lifestyle, and company-interaction related information about the customers
     - age, profession, location, income, marriage status, visits, logins, etc.
   - Use this information to learn a classification model
1. 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 sales amounts of new product based on advertising expenditure
  - Predicting the price of a house or car
  - 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

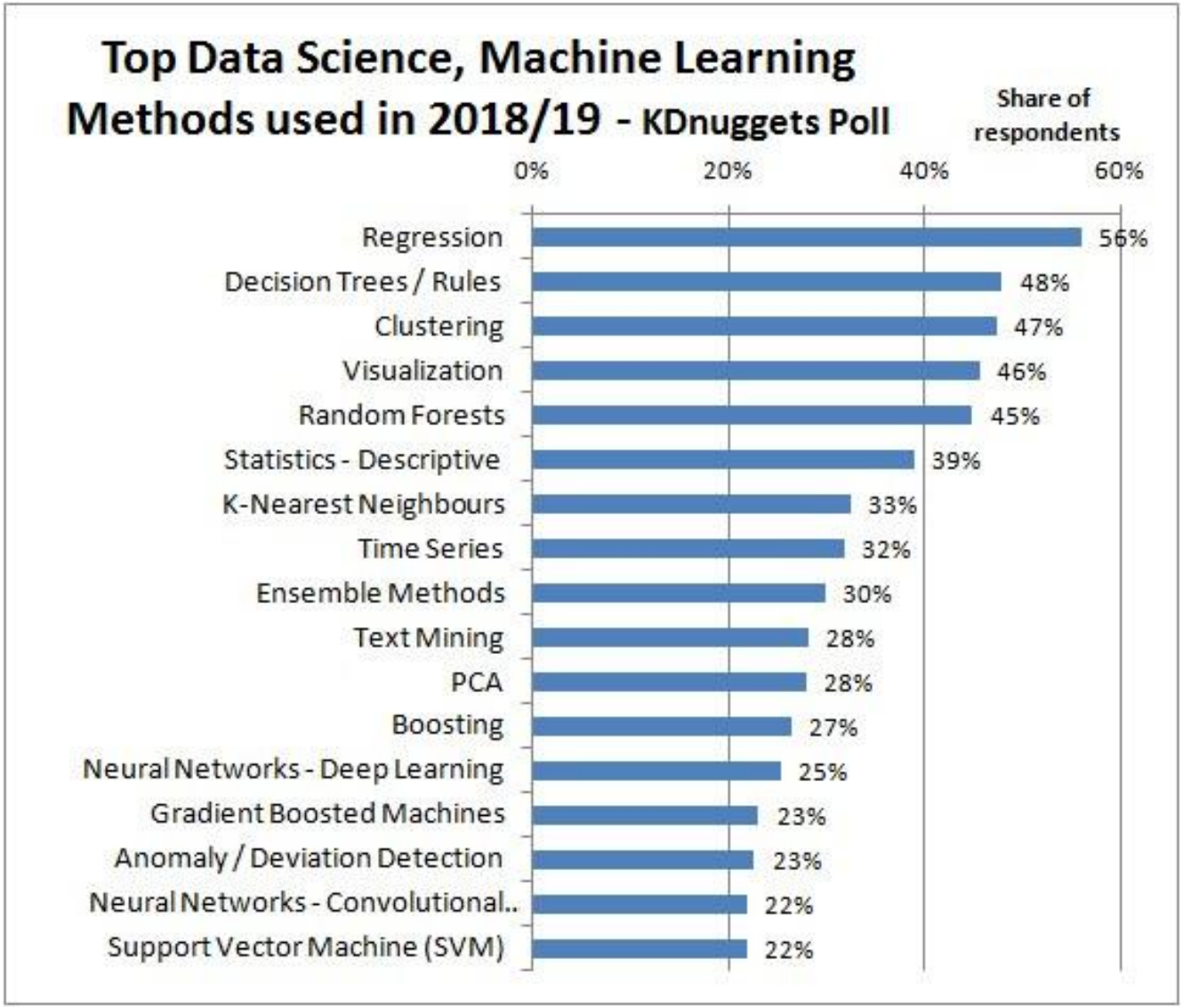
- Application area: Sales Promotion

Association Rule Discovery: Application 2

- Application area:
  Inventory Management

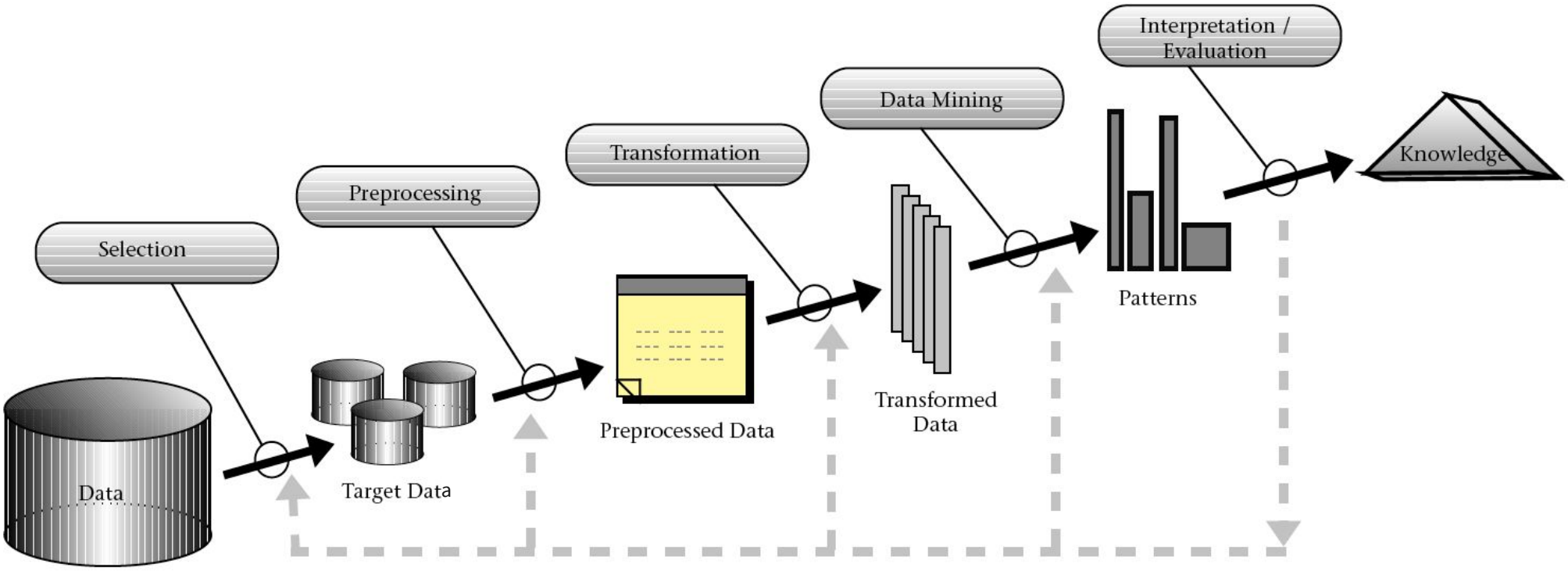
- Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households
- Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns

Which Methods are Used in Practice?



Source: 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

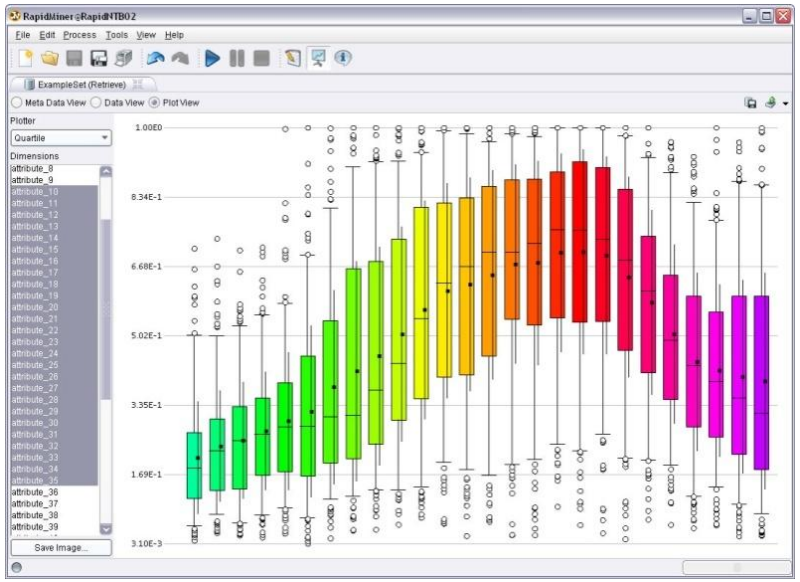
3. The Data Mining Process

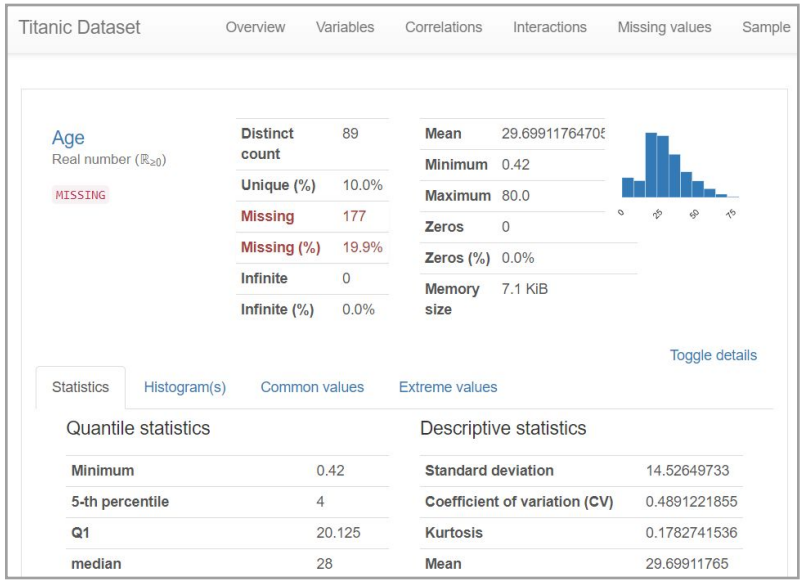


Source: Fayyad et al. (1996)

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





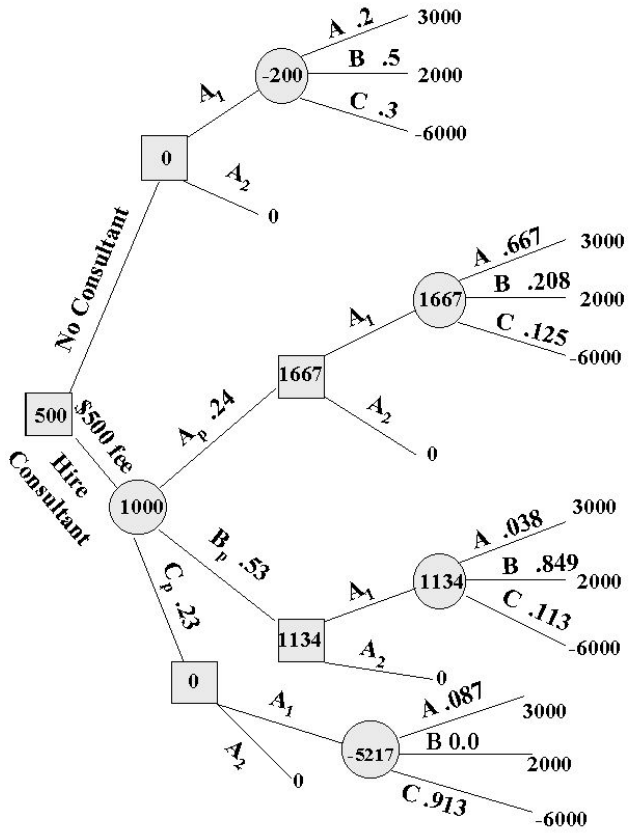
3.2 Preprocessing and Transformation

- 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

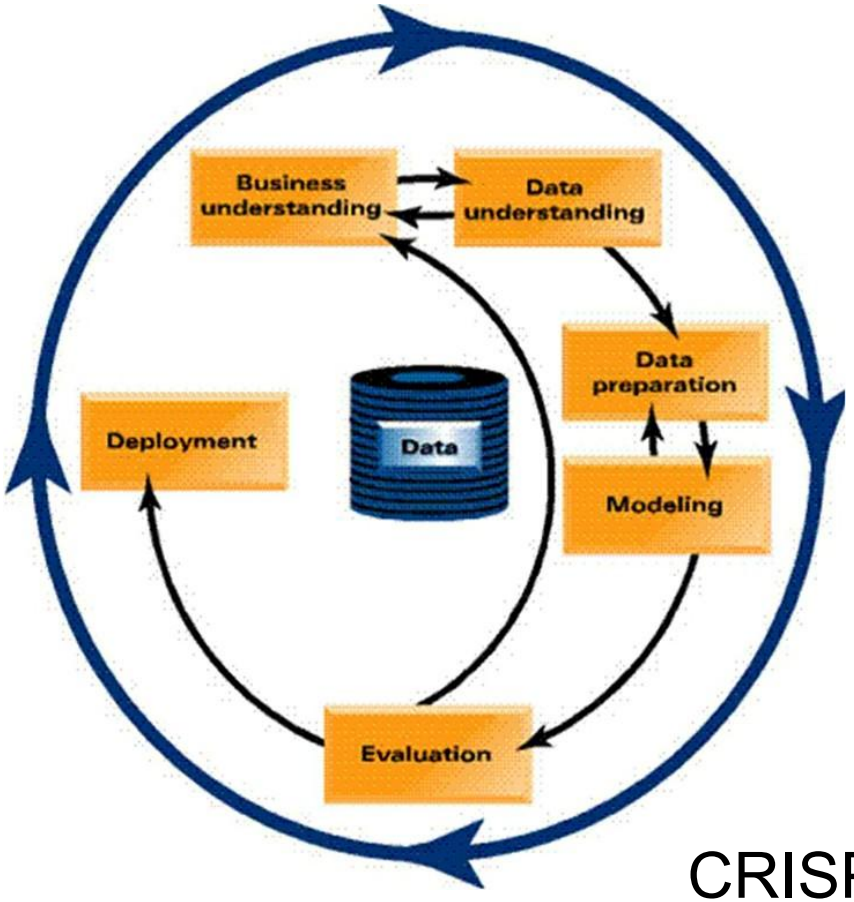
- Input: Preprocessed Data
- Output: Model / Patterns

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



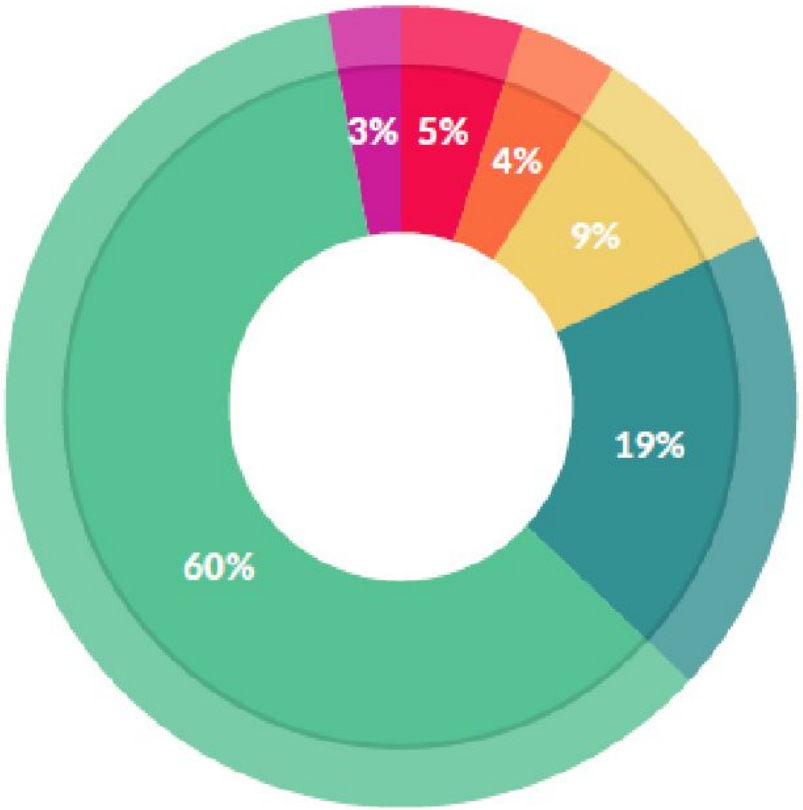
3.4 Deployment

- 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?

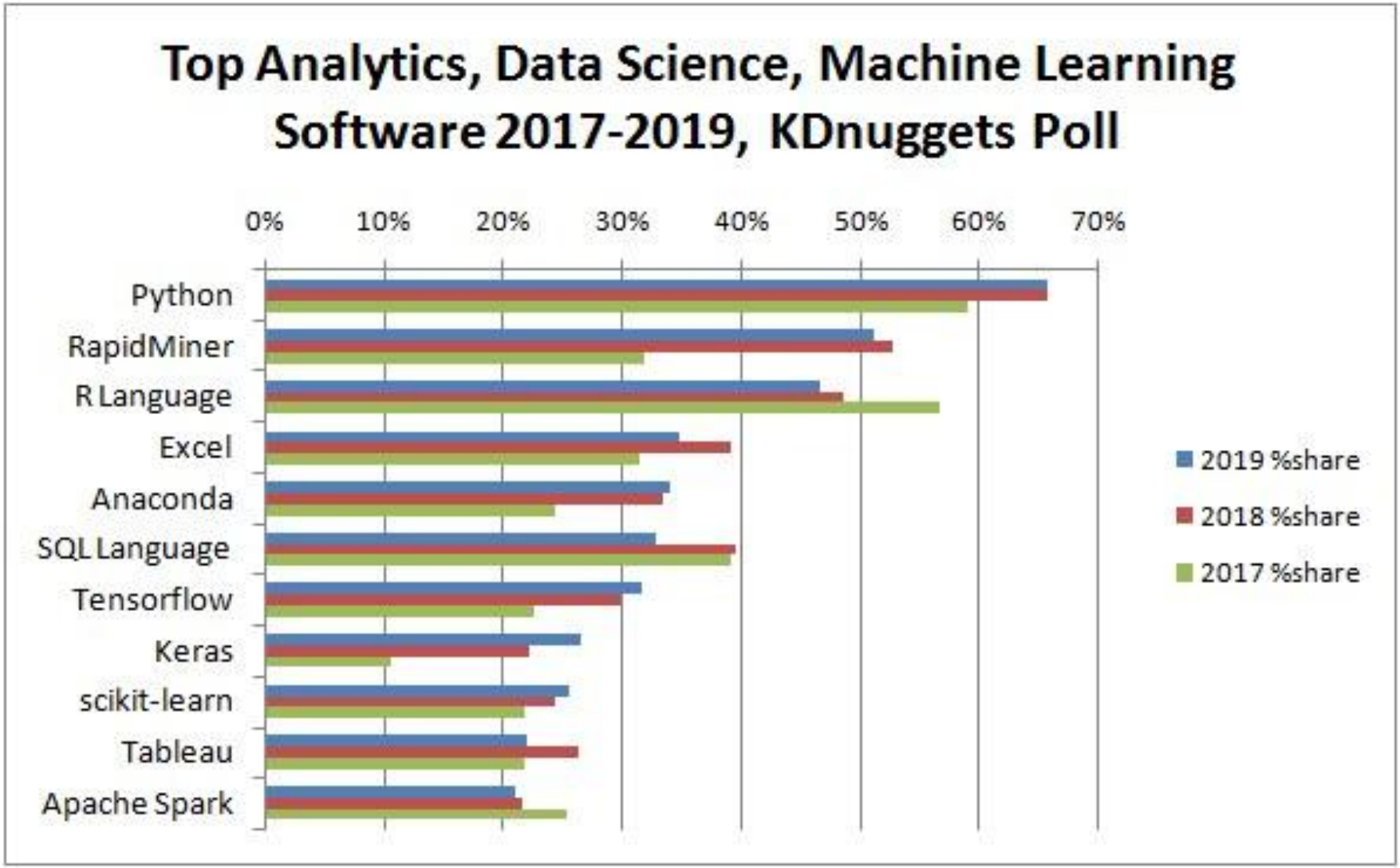


What data scientists spend the most time doing

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

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

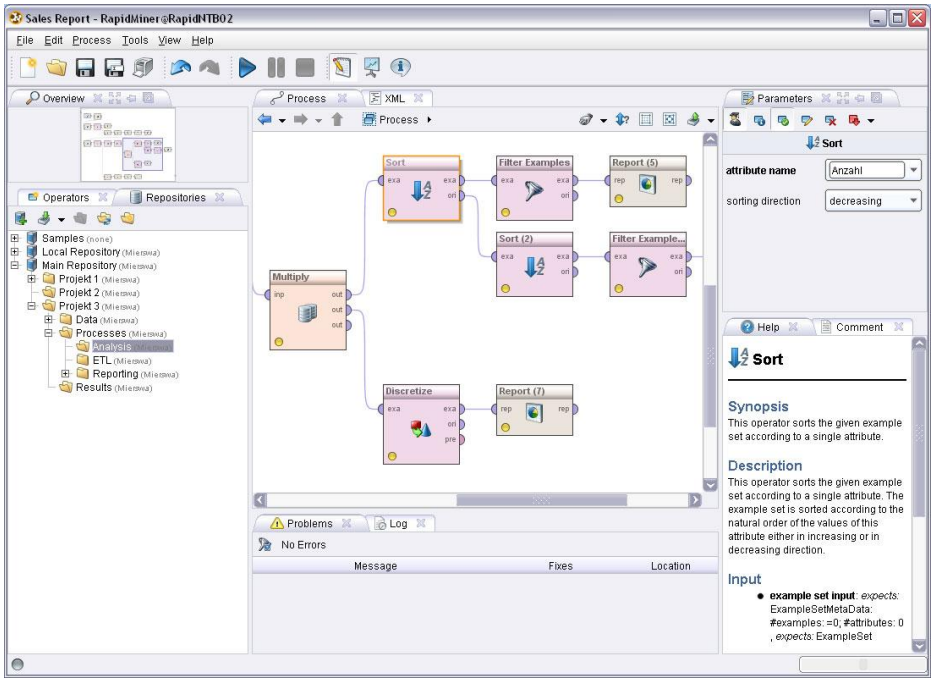
4. Data Mining Software

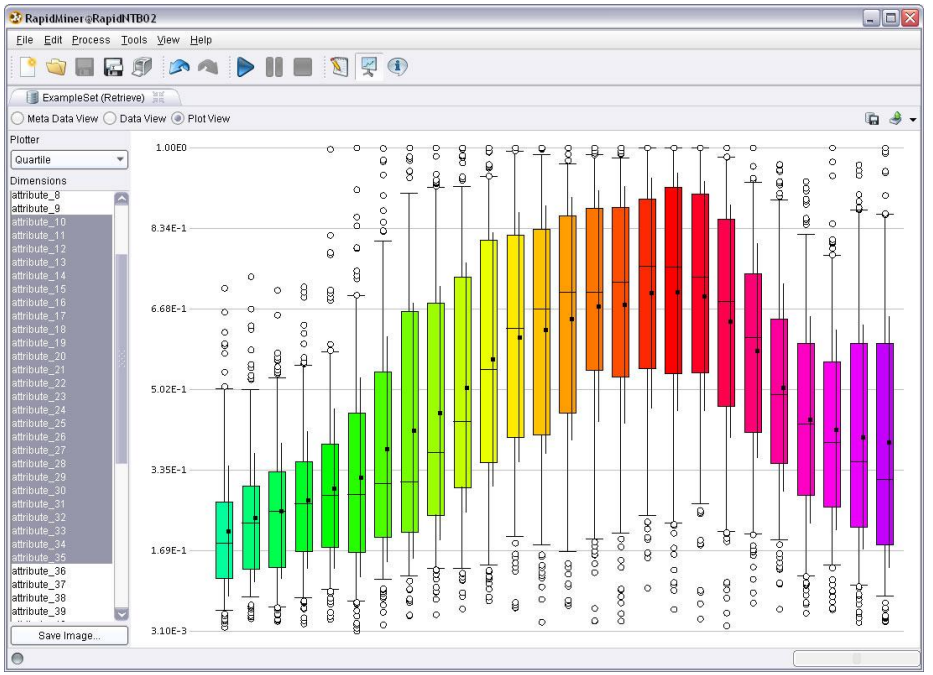


Source: KDnuggets online poll, 1800 votes
https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html

RapidMiner

- Powerful data mining suite
- Visual modelling of data mining pipelines
- Commercial tool, offering educational licenses

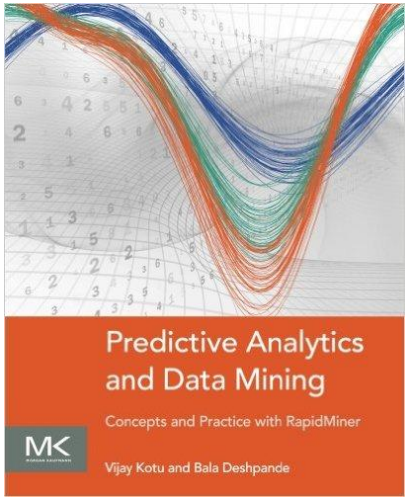


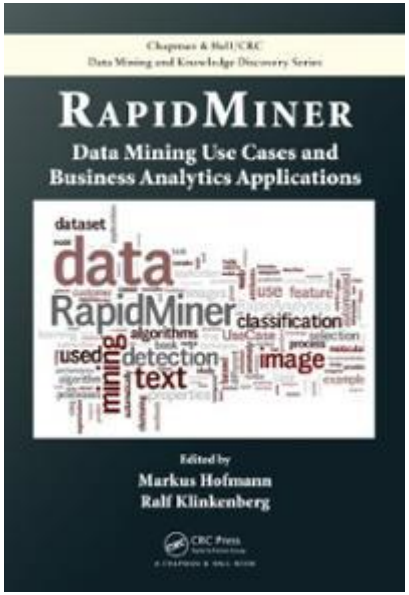


Gartner 2018 Magic Quadrant for Advanced Analytics Platforms

Literature – Rapidminer

1. **Rapidminer – Documentation**
   - http://docs.rapidminer.com
   - https://academy.rapidminer.com/catalog
2. Vijay Kotu, Bala Deshpande: **Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner.** Morgan Kaufmann, 2014.
   - covers theory and practical aspects using RapidMiner
3. Markus Hofmann, Ralf Klinkenberg: **RapidMiner: Data Mining Use Cases and Business Analytics Applications**. Chapman & Hall, 2013.
   - explains along case studies how to use simple and advanced Rapidminer features





Python

We use the Anaconda Python distribution

−includes relevant packages, e.g.

- scikit-learn, pandas
- NumPy, Matplotlib

−includes Jupyter as development environment

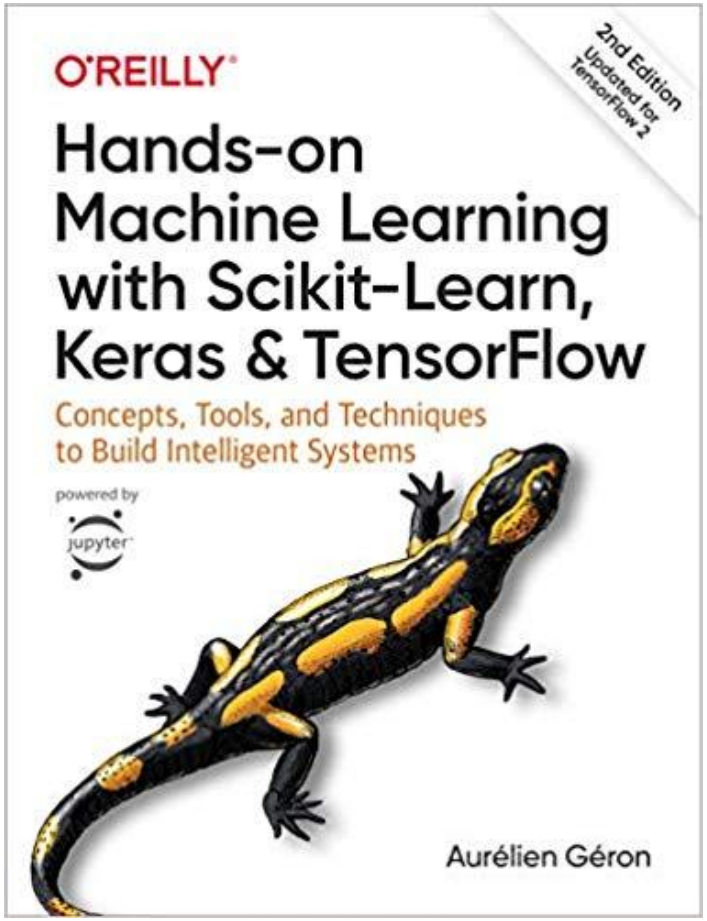
Slide Type Sub-Slide ▾

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import StratifiedKFold
from sklearn.model_selection import GridSearchCV

knn_estimator = KNeighborsClassifier()
parameters = {
    'n_neighbors': range(2, 9),
    'algorithm': ['ball_tree', 'kd_tree', 'brute']
}
stratified_10_fold_cv = StratifiedKFold(n_splits=10, shuffle=True, random_state=42)
grid_search_estimator = GridSearchCV(knn_estimator, parameters, scoring='accuracy',
                                     cv=stratified_10_fold_cv)
grid_search_estimator.fit(iris_data,iris_target)
```

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



Literature for this Chapter

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

Chapter 1: Introduction

Chapter 2: Data