UNIVERSITY OF MANNHEIM

Data Mining I

Introduction and Course Organisation

Hallo

- **Prof. Dr. Christian Bizer**
- Professor for Information Systems V
- Research Interests:
  - Data and Web Mining
  - Web Data Integration
  - Data Web Technologies
- Room: B6 - B1.15
- eMail: chris@informatik.uni-mannheim.de
- Consultation: Wednesday, 13:30-14:30

- I will teach the lecture introducing the principle methods of data mining.

Hallo

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

- Anna will teach Exercise 1 (RapidMiner) and will supervise the student projects.

Hallo

- **Oliver Lehmberg**
- Graduate Research Associate
- Research Interests:
  - Data and Web Mining
  - Network Analysis
  - Web Data Integration
- Room: B6, 26, C 1.04
- eMail: oli@informatik.uni-mannheim.de

- Oliver will teach Exercise 2 and 3 (Python) and will supervise the student projects.

Outline of Today‘s Lecture

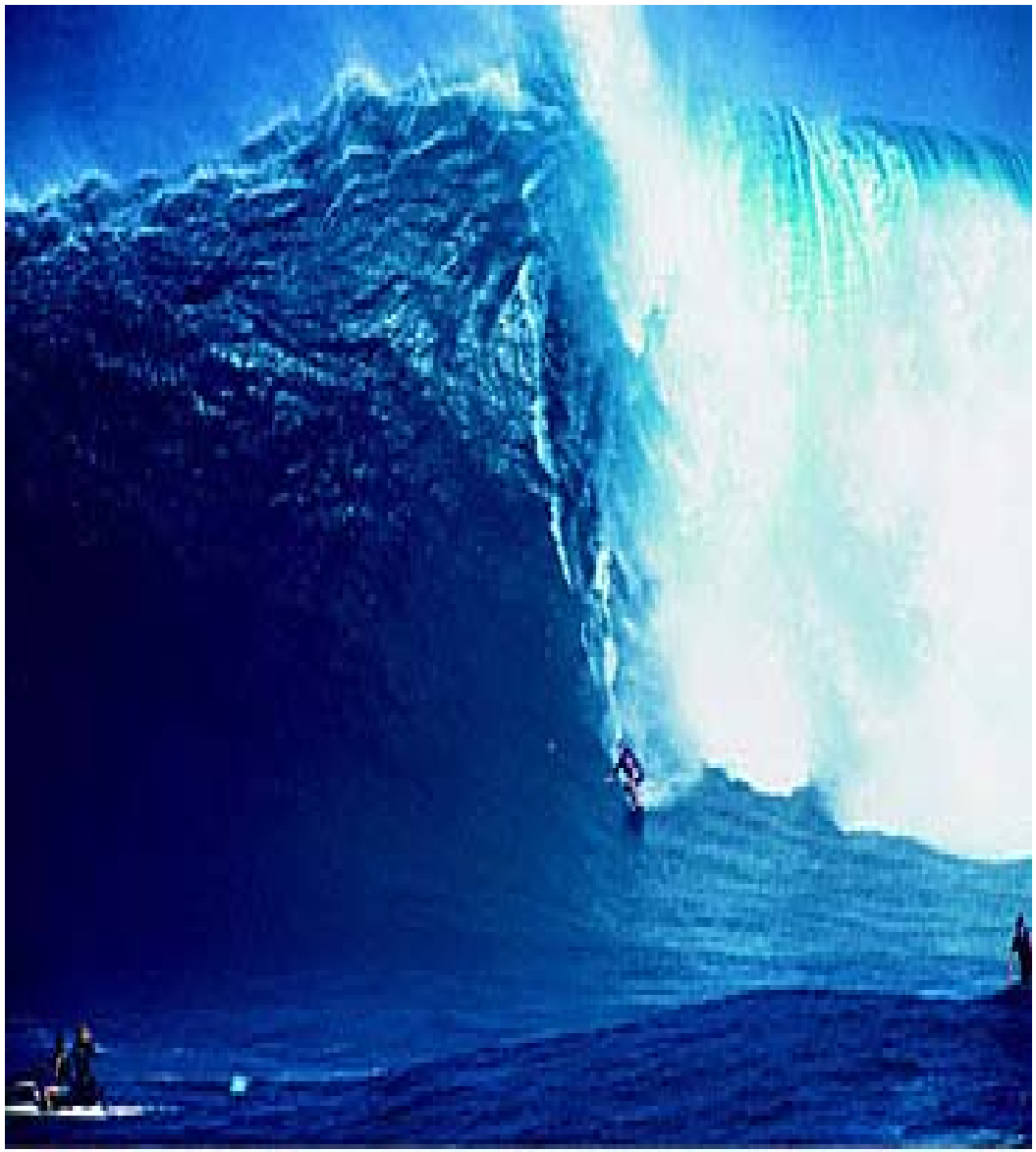
1. Introduction to Data Mining (60 minutes)
    1. What is Data Mining?
    2. Methods and Applications
    3. The Data Mining Process
2. Course Organisation (30 minutes)

1. Introduction to Data Mining

The Data Deluge

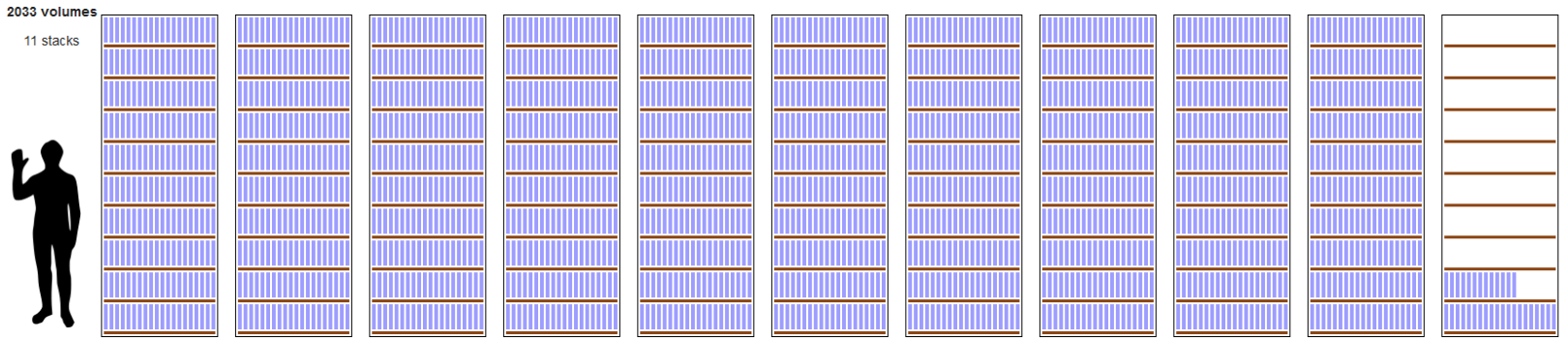
More and more data is generated:

- Transaction data from e-commerce, banking
- Scientific data from astronomy, physics, biology
- Social network sites
- The public Web, twitter, the blogosphere
- Sensor data from machines
- ERP application logs



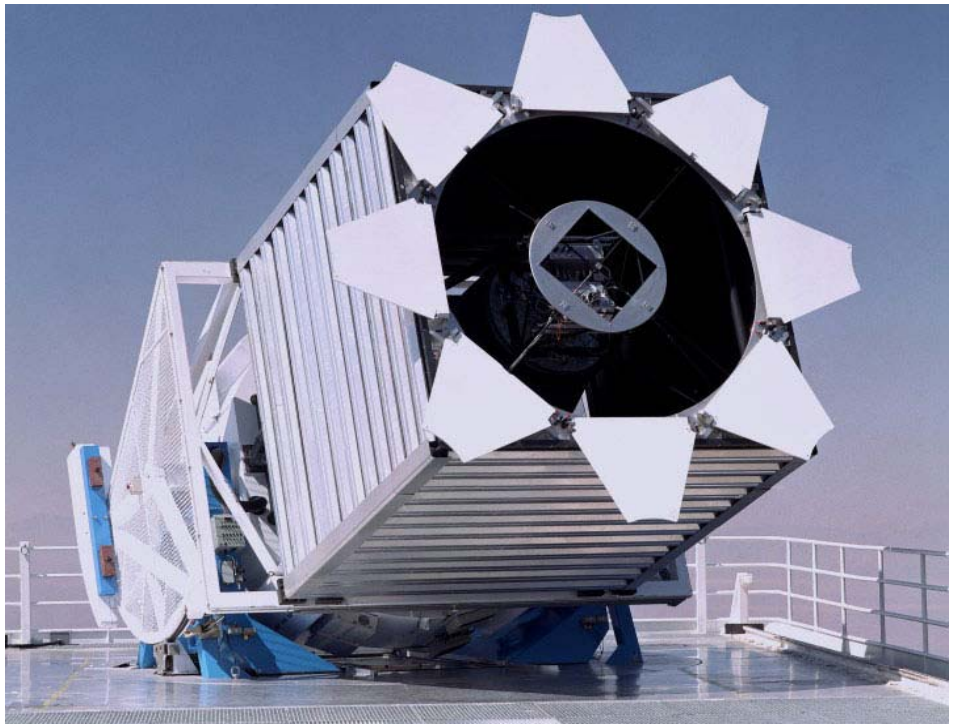
“We are Drowning in Data...”

Wikipedia = Reference Size
≈ 5.9 TB of data



Source: The following slides are taken from Aidan Hogan's course on “Massive Data Processing”

“We are Drowning in Data...”



Sloan Digital Sky Survey
≈ 200 GB/day
≈ 73 TB/year
≈ 12 Wikipedias/year

Analyze
•Type of sky object:
Star or galaxy?

“We are Drowning in Data...”

US Library of Congress
≈ 235 TB archived
≈ 40 Wikipedias

Analyze
•Topic distributions
•Citation networks
•Historic trends*.

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

“We are Drowning in Data...”

Facebook

≈ 12 TB/day added
(*as of Mar. 2010*)
≈ 2 Wikipedias/day
≈ 782 Wikipedias/year

Analyze

•Current interests and behavior of over one billion people.

“We are Drowning in Data...”

Google

≈ 20 PB/day processed *(Jan. 2010)*

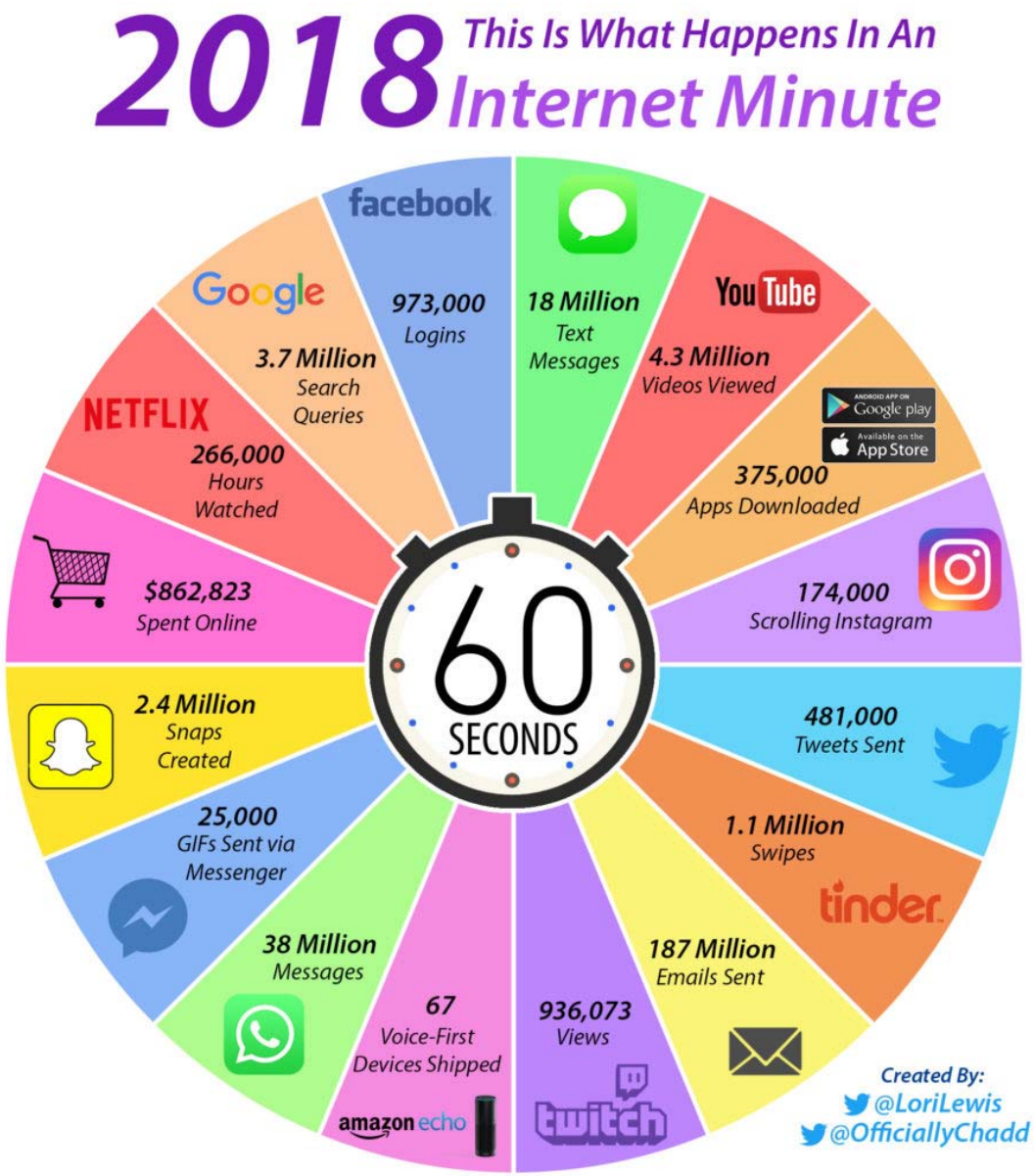
≈ 3,389 Wikipedias/day

≈ 7,300,000 Wikipedias/year

Analyze

•Browsing behavior and interests of users.

“We are Drowning in Data...”



Analyze

- Current behavior and interests of mankind.

“We are Drowning in Data...”

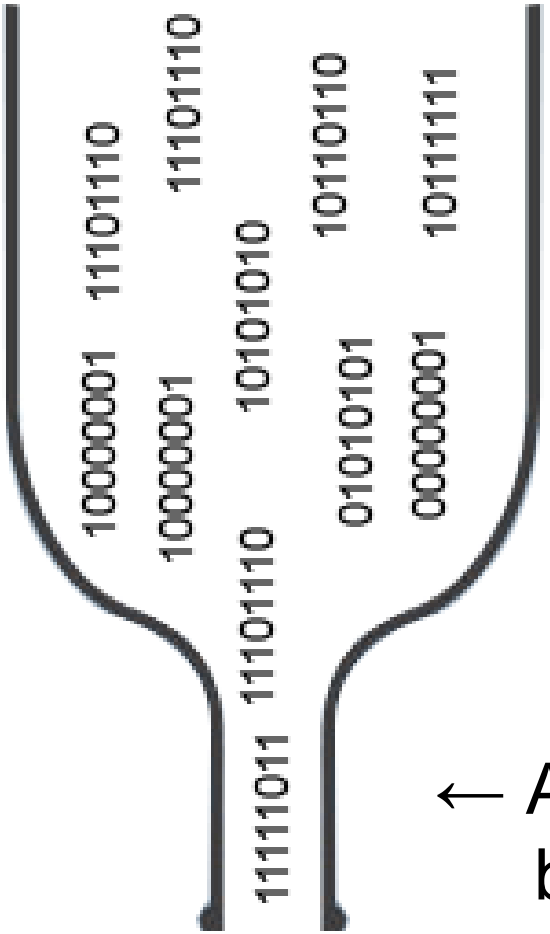
NSA

Unknown amount of communication data from all over the world.

Analyze

- Identify suspects and terrorists.

“We are Drowning in Data...”



← Amount of data that is produced.

← Amount of data that can be looked at by humans.

“...but starving for knowledge!”

The valuable knowledge is “hidden” in the raw data.

Data Mining methods are needed in many cases to

- make sense of data.
- take business decisions based on data.

We are interested in the **patterns** not the data itself.

1.1 What is Data Mining?

– Definitions

Non-trivial extraction of

– **implicit,**

– **previously unknown and**

– **potentially useful**

information from data.

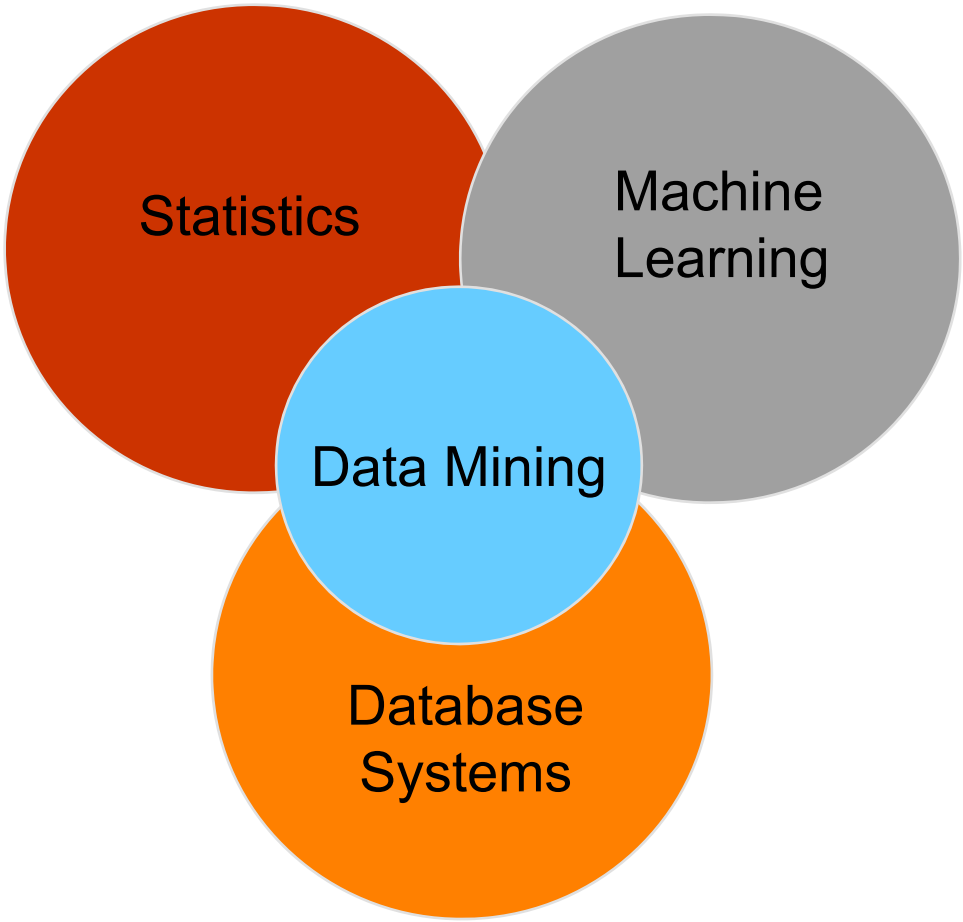
Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns.

– What is needed? Methods that

1. detect patterns and regularities in data
2. support business decisions based on data patterns

Origins of Data Mining

- Draws ideas from machine learning, statistics, and database systems.
- Traditional techniques may be unsuitable due to
  - large amount of data
  - high dimensionality of data
  - heterogeneous and distributed nature of data



Data Mining Application Fields

- Business
  - Customer relationship management, marketing, fraud detection, manufacturing, telecom, health care, …
- Science
  - Data mining helps scientists to formulate hypotheses.
  - Astronomy, physics, drug discovery, social sciences, …
- Web and Social Media
  - Advertising, search engine optimization, spam detection, web site optimization, sentiment analysis, …
- Government
  - Surveillance, crime detection, finding tax cheaters, …

Big Data and the Cloud

- Today, everybody can mine large amounts of data at low costs in the cloud.
- Technical realization
  - massive parallelization using hundreds or thousands of machines
  - using tools like Spark, TensorFlow, Mahout
- Open Data
  - Hundreds of portals offer thousands of data sets
  - https://data.wu.ac.at/portalwatch/
  - https://toolbox.google.com/datasetsearch
- Conference
  - O'Reilly STRATA Conference
  - http://strataconf.com/public/content/home

The Hottest Skills That Got People Hired in 2016

LinkedIn analyzed the skills of members who started new jobs or received interest from recruiters in 2016.

Source: https://business.linkedin.com/talent-solutions/blog/trends-and-research/2016/the-top-10-skills-you-will-be-hiring-for-in-2017

1.2 Methods and Applications

Data Mining Methods

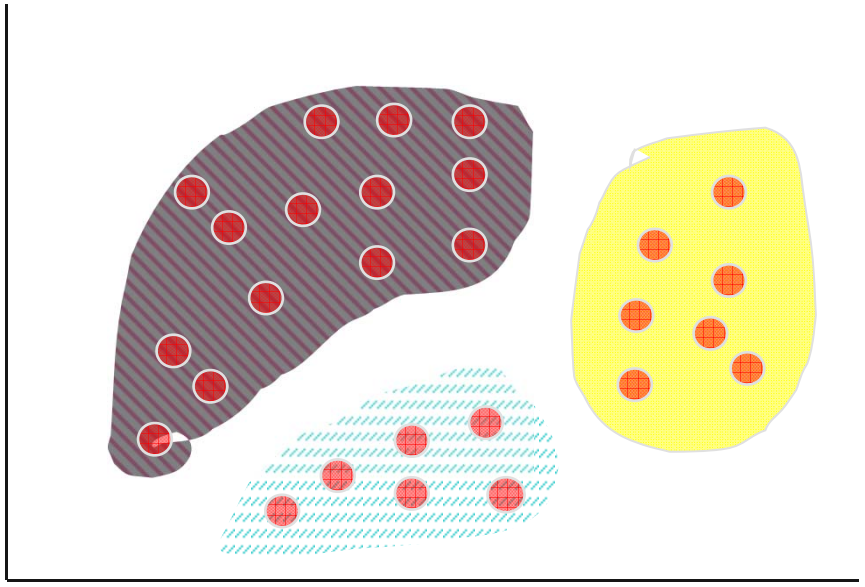
- Descriptive Methods
  - Goal: Find patterns in the data.
  - Example: W*hich products are often bought together?*
- Predictive Methods
  - Goal: Predict unknown values of a variable
    - given observations (e.g., from the past)
  - Example: W*ill a person click a online advertisement?*
    - given her browsing history
- Machine Learning Terminology
  - descriptive = unsupervised
  - predictive = supervised

Data Mining Tasks

1. Clustering [Descriptive]
2. Classification [Predictive]
3. Regression [Predictive]
4. Association Rule Discovery [Descriptive]
5. Time Series Prediction [Predictive, Data Mining II]
6. Sequential Pattern Discovery [Descriptive, Data Mining II]
7. Anomaly Detection [Descriptive, Data Mining II]

1.2.1 Clustering: Definition

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - data points in one cluster are more similar to one another
  - data points in separate clusters are less similar to one another
- Similarity Measures
  - Euclidean distance if attributes are continuous
  - Other problem-specific similarity measures
- Goals
  - Intracluster distances are minimized
  - Intercluster distances are maximized
- Result
  - A descriptive grouping of data points

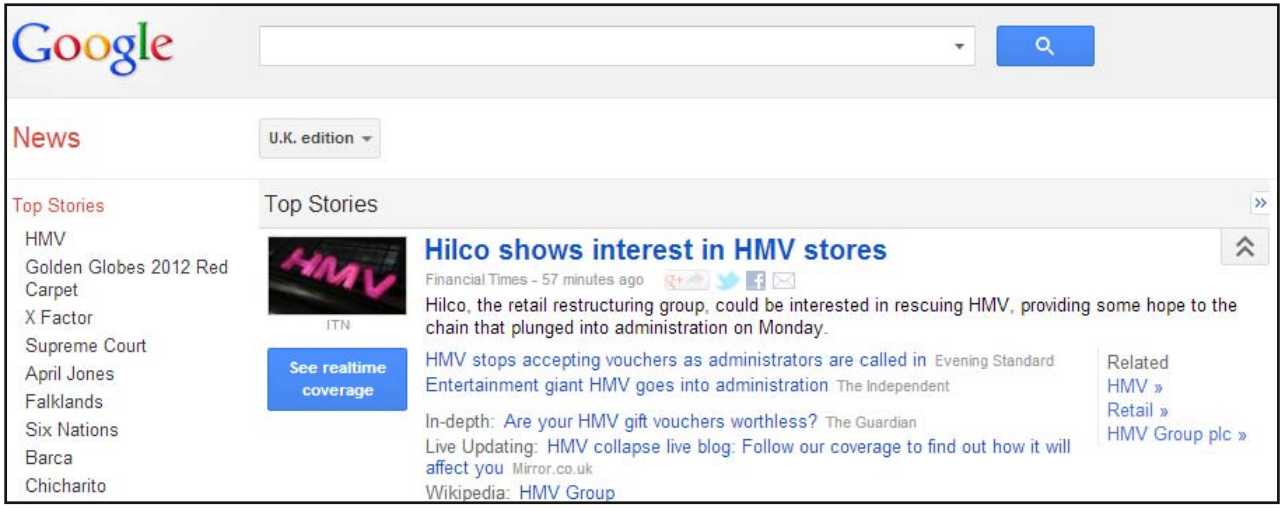


Clustering: Application 1

- Application area: Market segmentation
- Goal: Divide a market into distinct subsets of customers
  - where any subset 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 of customers in same cluster vs. those from different clusters

Clustering: 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



1.2.2 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 of the attributes is the *class (label)* that should be predicted.

– Find a *model* for the class attribute as a function of the values of other attributes.

Classification: Example

- Training set:

"tree"

"tree"

"tree"

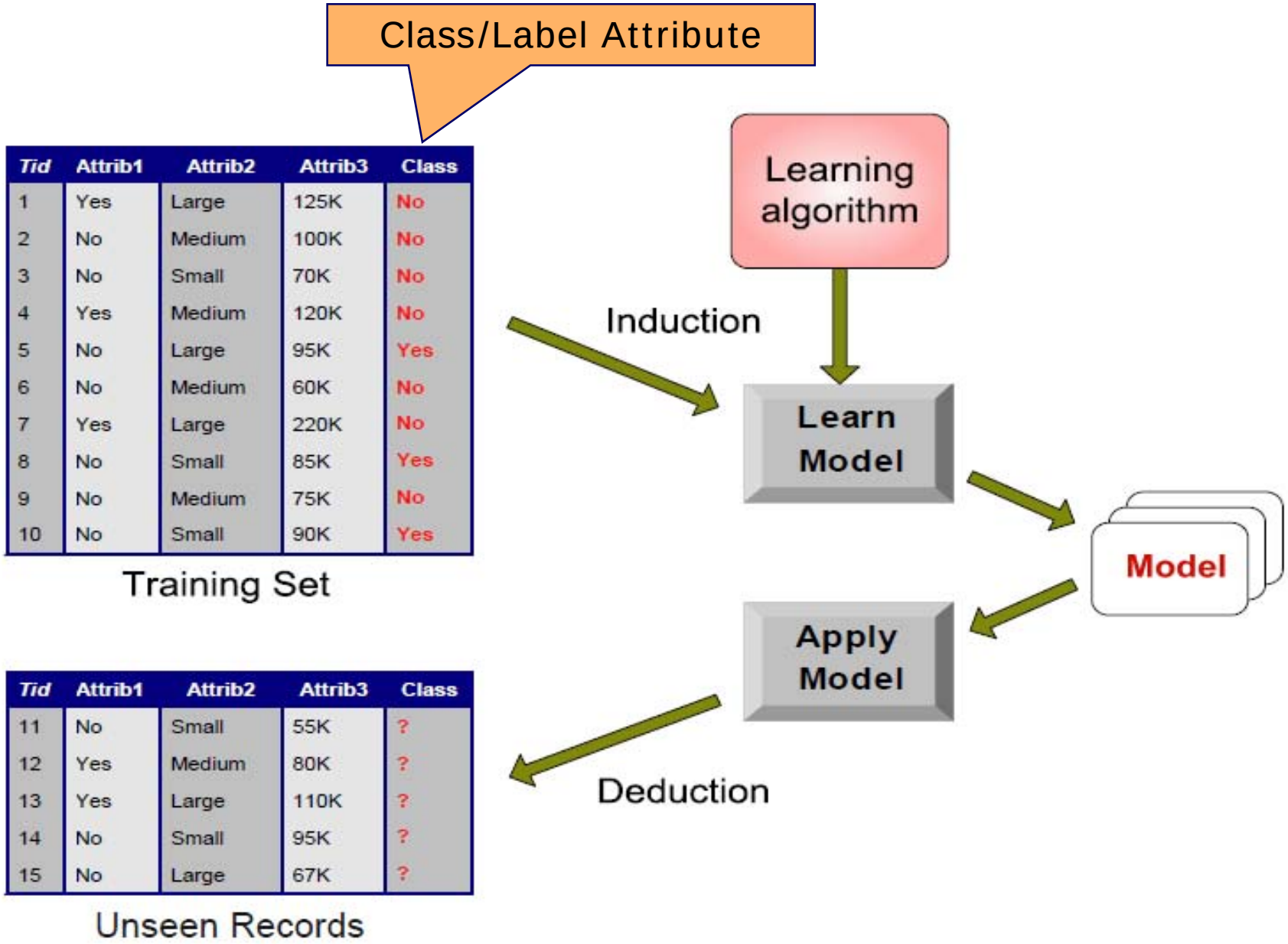
"not a tree"

"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

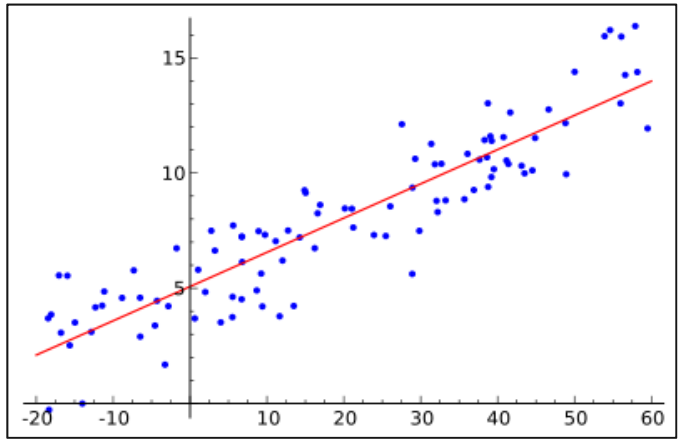
- Application area: Fraud Detection
- Goal: Predict fraudulent cases in credit card transactions.
- Approach:
  1. Use credit card transactions and information about account-holders as attributes.
     - When and where does a customer buy? What does he buy?
     - How often he pays on time? etc.
  2. 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.

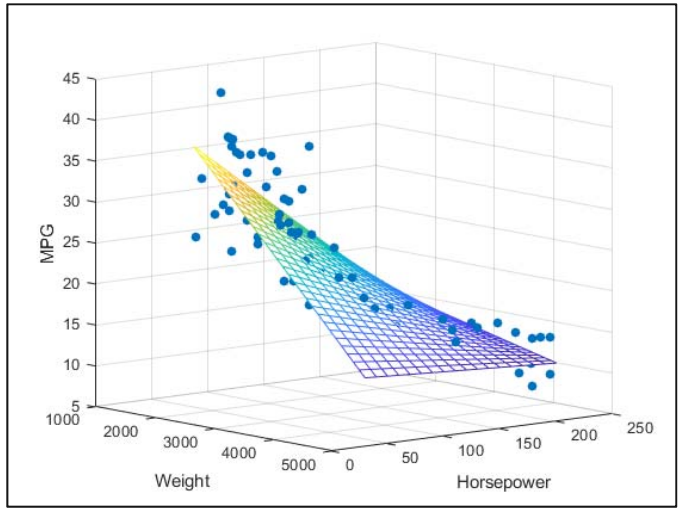
Classification: Application 2

- Application area: Direct Marketing
- Goal: Reduce cost of mailing by targeting the set of consumers likely to buy a new cell-phone product.
- Approach:
  1. Use the data for a similar product introduced before.
     - 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 all such customers.
     - Age, profession, location, income, marriage status, etc.
  3. Use this information as input attributes to learn a classification model.

1.2.3 Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics and neural network field.
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Predicting the realizable price of a house or car.





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

1.2.4 Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection
- produce dependency 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

Rules Discovered
{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: Applications 2

- Application area: Advertising
- Real example:
  - Target (American grocery store)
  - Analyzes customer buying behavior
  - Sends personalized advertisements and coupons
- Famous case in the USA:
  - Teenage girl gets advertisement for baby products
  - ... and her father is mad

Articles about this case

- http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html
- http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/
- http://mashable.com/2014/04/26/big-data-pregnancy/

Association Rule Discovery: Application 3

- 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.

1.2.5 Sequential Pattern Discovery: Definition

- Given a sequence of events (or sets of events)
- Find typical temporal patterns:
  - 1. (A,B) 2. (C) 3. (D,E)
  - 1. (A) 2. (B,C) 3. (D)
  - 1. (C) 2. (A,D) 3. (B)
- Typical pattern: Event C usually takes place before event D.

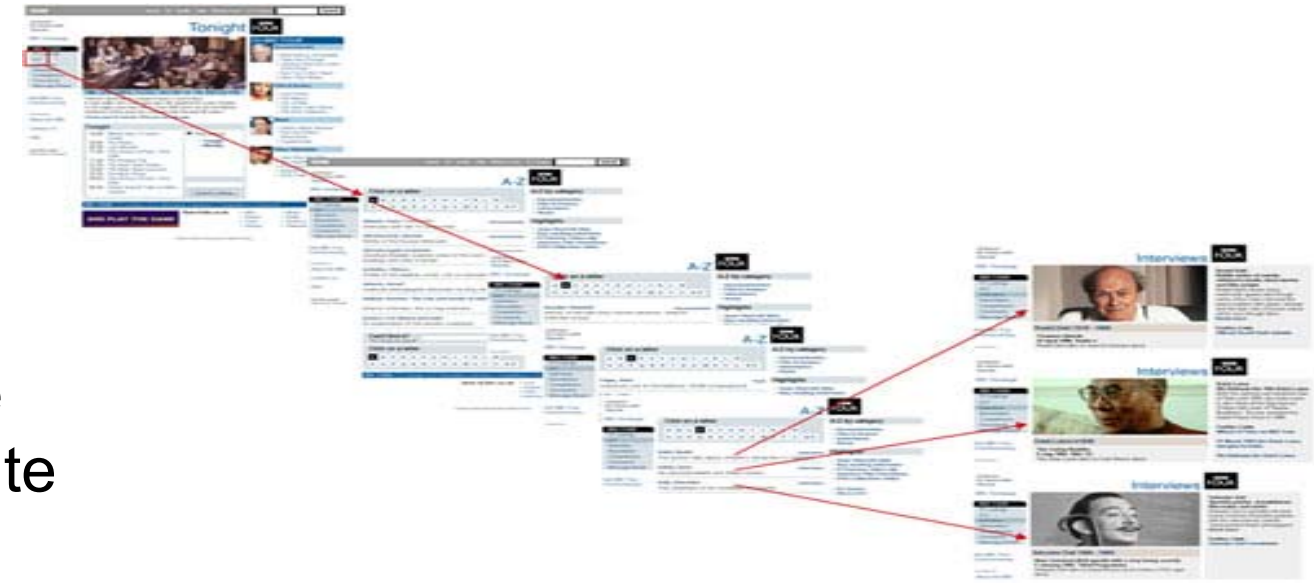
Sequential Pattern Mining: Applications 1

- Application area: Marketing
- Recurring customers
  - Book store example: (Twilight) (New Moon) → (Eclipse)

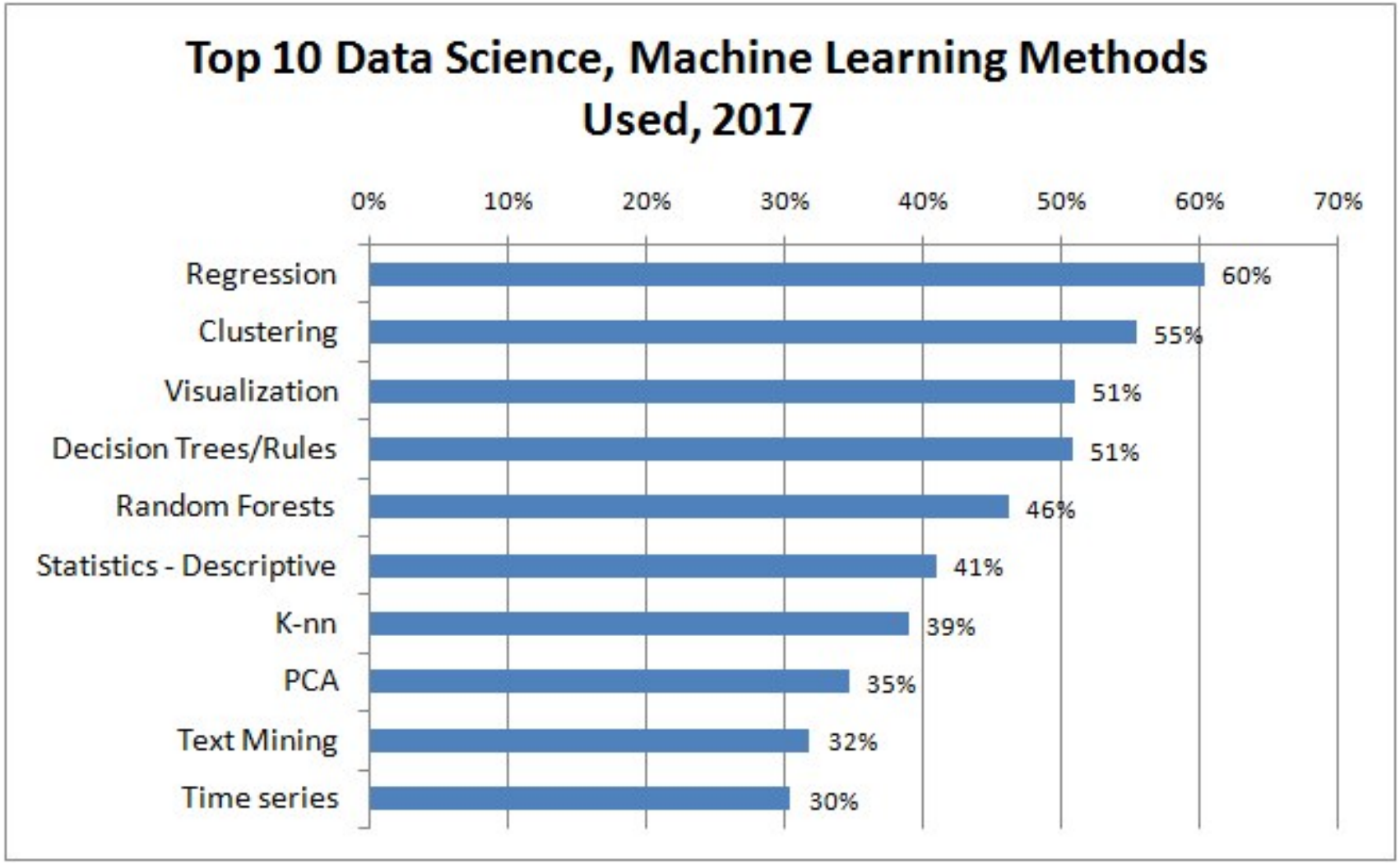
- Sequential patterns allow more fine grained suggestions than frequent pattern mining without sequence information
- Example:
  - *mobile phone* → *charger* vs. *charger* → *mobile phone*
    - are indistinguishable by frequent pattern mining
  - customers will buy a charger after a mobile phone
    - but not the other way around!

Sequential Pattern Mining: Applications 2

- Application area: Web usage mining
- Input
  - Web server logs
- Patterns
  - typical sequences of pages visited
- Goal: Improve structure and navigation of website

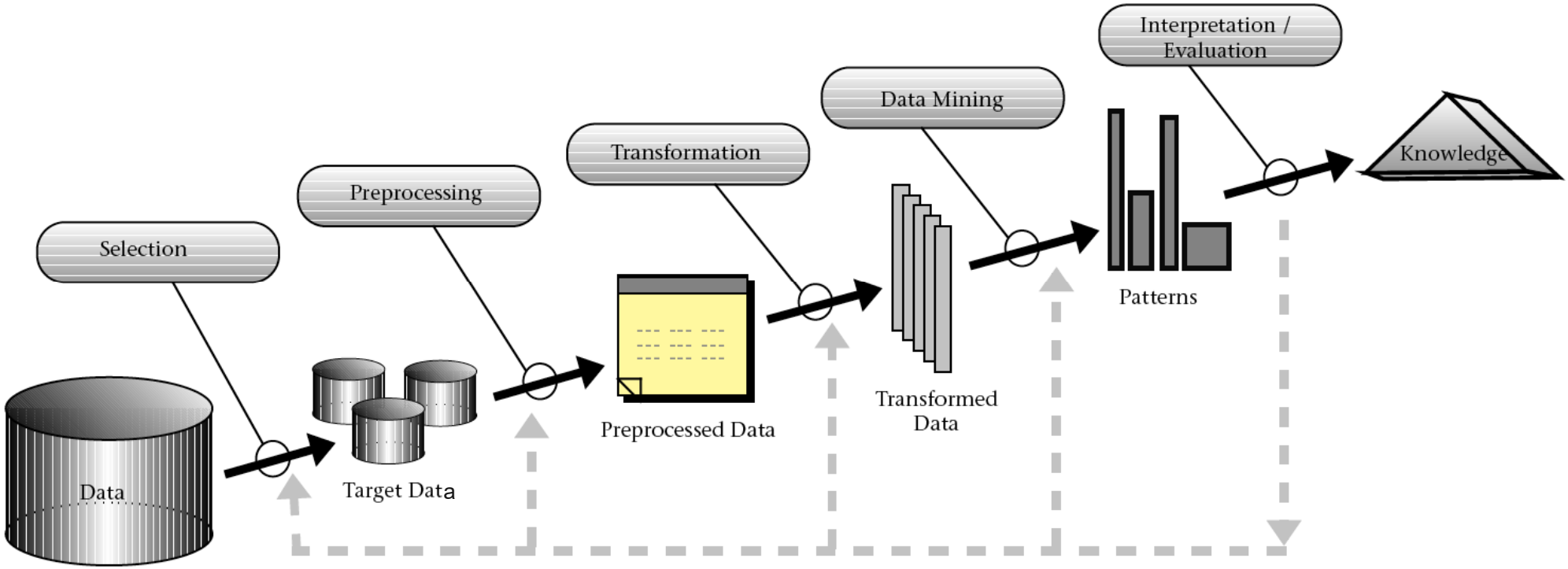


Which Methods are Used in Practice?



Source: KDnuggets online poll, 732 votes, question: methods used last year for real-world app?
https://www.kdnuggets.com/2017/12/top-data-science-machine-learning-methods.html

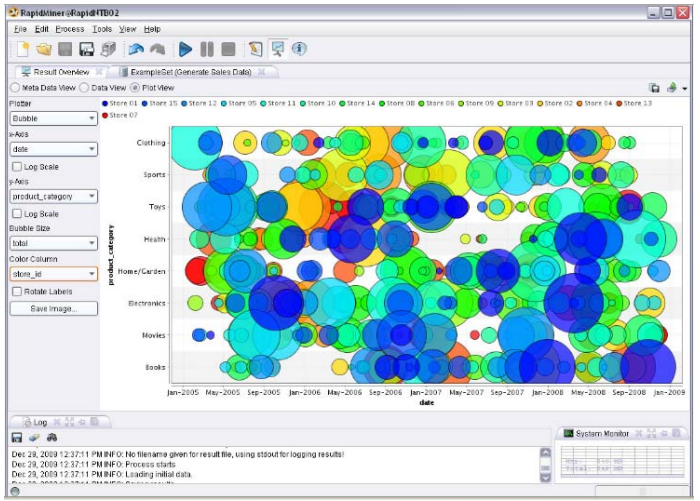
1.3. The Data Mining Process

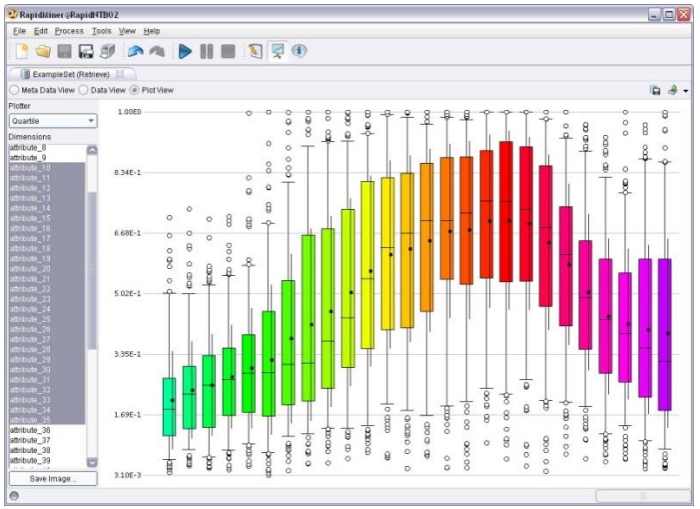


Source: Fayyad et al. (1996)

1.3.1 Selection and Exploration

- Selection
  - What data is available?
  - What do I know about the provenance of this data?
  - What do I know about the quality of the data?
- Exploration
  - Get an initial understanding of the data
  - Calculate basic summarization statistics
  - Visualize the data
  - Identify data problems such as outliers, missing values, duplicate records





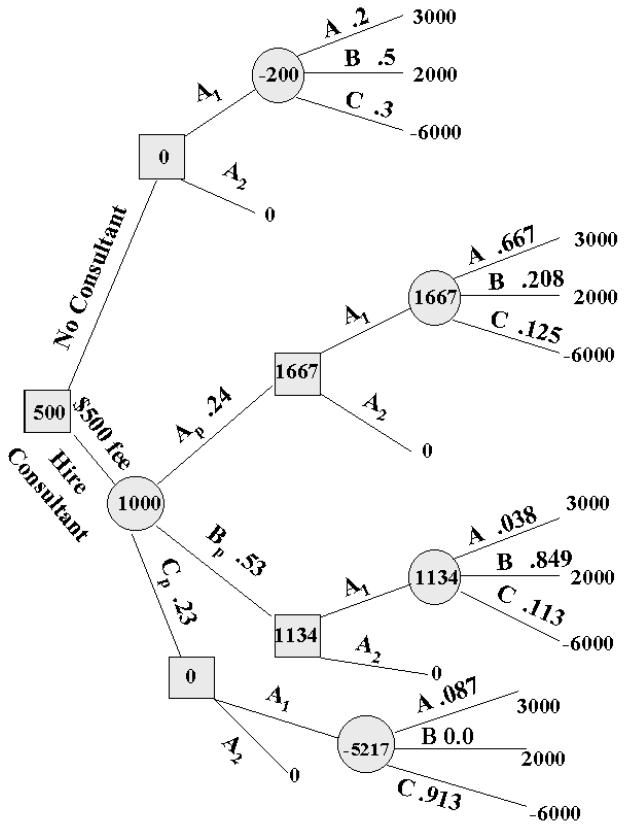
1.3.2 Preprocessing and Transformation

- Transform data into a representation that is suitable for the chosen data mining methods
  - number of dimensions
  - scales of attributes (nominal, ordinal, numeric)
  - amount of data (determines hardware requirements)
- Methods
  - integrate data from multiple sources
  - aggregation, sampling
  - dimensionality reduction / feature subset selection
  - attribute transformation / text to term vector / embeddings
  - discretization and binarization
- 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!

1.3.3 Data Mining

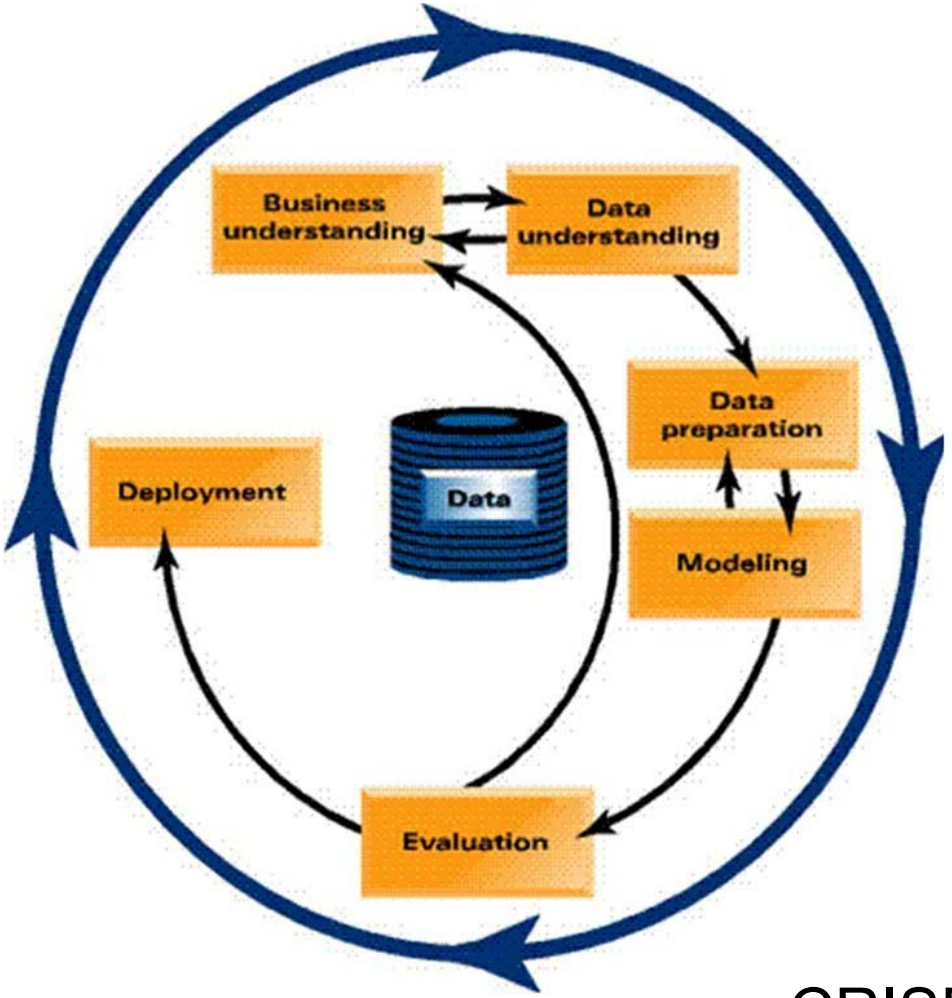
- Input: Preprocessed Data
- Output: Model / Patterns

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
   - Combine/ensemble multiple methods



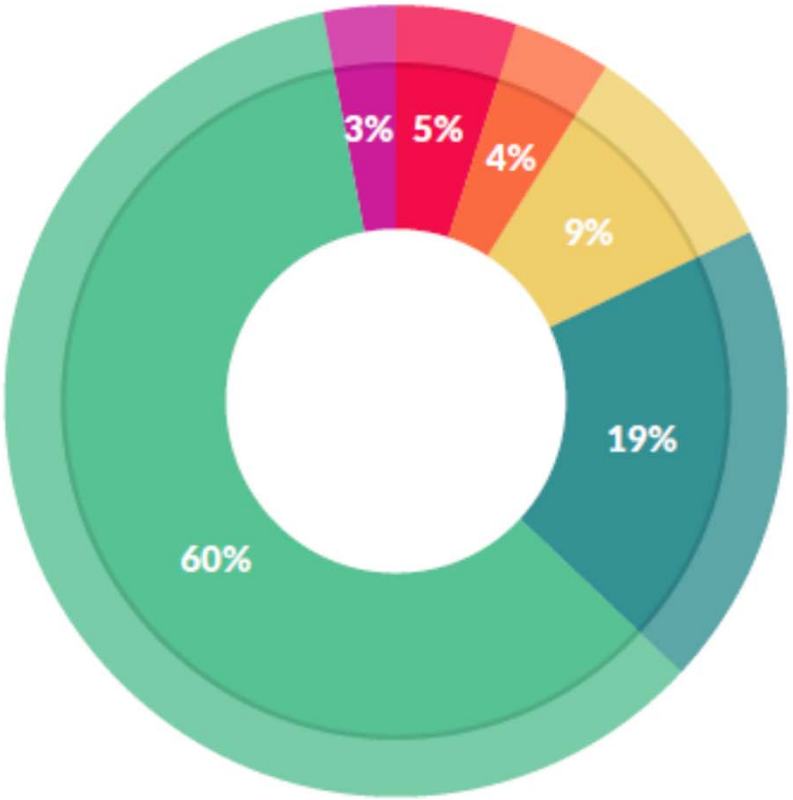
1.3.4 Deployment

– Use model in the business context.



CRISP-DM Process Model

How Do Data Scientists Spend Their Days?



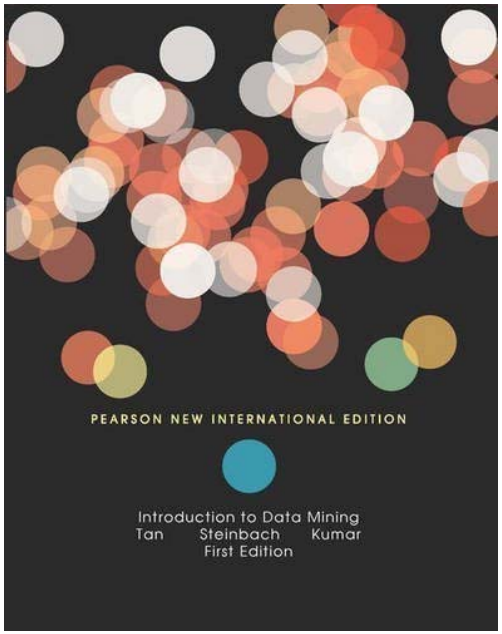
What data scientists spend the most time doing

- *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

Literature Reference for this Chapter

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



Chapter 1: Introduction

Chapter 2: Data

2. 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
- Three alternative Exercise Groups
  - students experiment with data sets 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 summary about their project and present the project results
- Grading
  - 60% written exam, 30% project report, 10% 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
    Room A5, B144
  - Exercise:
    - Thursday, 10.15 - 11.45
      Room B6, A104 (Anna, RapidMiner)
    - Thursday, 12.00 - 13.30,
      Room B6, A104 (Oliver, Python)
    - Thursday, 13.45 - 15.15,
      Room B6, A104 (Oliver, Python)

Lecture Contents

1. Introduction to Data Mining	What is Data Mining? Methods and Applications The Data Mining Process
2. Clustering	K-means Clustering, Density-based Clustering, Hierarchical Clustering, Proximity Measures
3. Classification	Nearest Neighbor, Decision Trees, Model Evaluation, Rule Learning, Naïve Bayes, Neural Networks, Support Vector Machines
4. Regression	Linear Regression, Nearest Neighbor Regression Regression Trees, Time Series
5. Association Analysis	Frequent Item Set Generation, Rule Generation Interestingness Measures
6. Text Mining	Preprocessing Text, Feature Generation, Feature Selection, RapidMiner Text Extension
7. Introduction to Student Projects	Requirements and Organization Overview of proposed data sets and tasks

Schedule

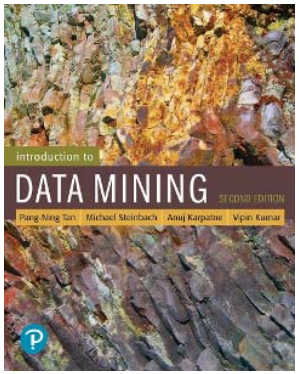
Week	Wednesday	Thursday
13.02.2019	Introduction to Data Mining	Introduction to RapidMiner/Python
20.02.2019	Lecture Clustering	Exercise Clustering
27.02.2019	Lecture Classification 1	Exercise Classification
06.03.2019	Lecture Classification 2	Exercise Classification
13.03.2019	Lecture Classification 3	Exercise Classification
20.03.2019	Lecture Regression	Exercise Regression
27.03.2019	Lecture Text Mining	Exercise Text Mining
03.04.2019	Introduction to Student Projects and Group Formation	Preparation of Project Outlines
10.04.2019	Lecture Association Analysis	Exercise Association Analysis
	- Easter Break -	
01.05.2019	- Holiday -	Feedback on demand
08.05.2019	Project Work	Feedback on demand
15.05.2019	Project Work	Feedback on demand
22.05.2019	Project Work	Submission of project results
29.05.2019	Presentation of project results	- Holiday -
03.06.2019	Final Exam	

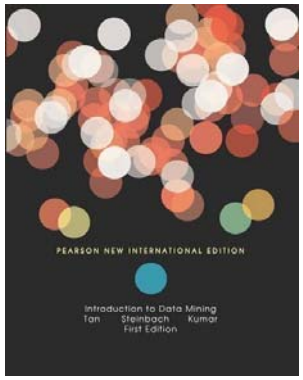
Deadlines

- Submission of project proposal
  - Sunday, April 7th, 23:59
- Submission of final project report
  - Sunday, May 26th, 23:59
- Project presentations
  - Thursday, May 30th
  - everyone has to attend the presentations
- Final written exam
  - Monday, June 3rd

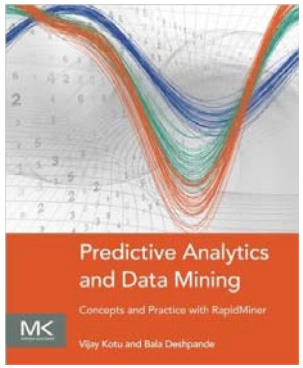
Literature

1. Pang-Ning Tan, Michael Steinbach, Vipin Kumar: **Introduction to Data Mining. 2nd Edition,** Pearson.
   - main reference book for the course!
   - we provide scans of important chapters via ILIAS

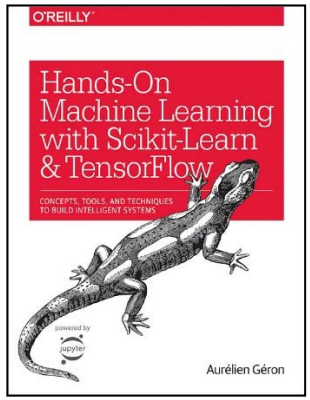




2. Vijay Kotu, Bala Deshpande: **Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner.** Morgan Kaufmann.
   - covers some theory, but also many practical aspects using RapidMiner



3. Aurélien Géron: **Hands-on Machine Learning with Scikit-Learn**. O'Reilly.
   - explains how to apply the covered methods using Python

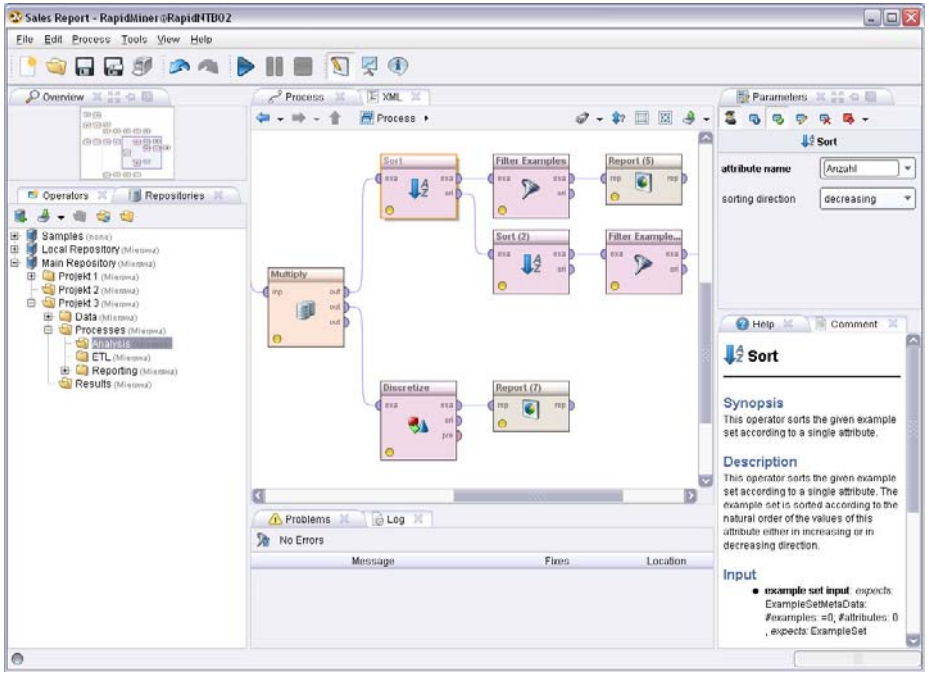


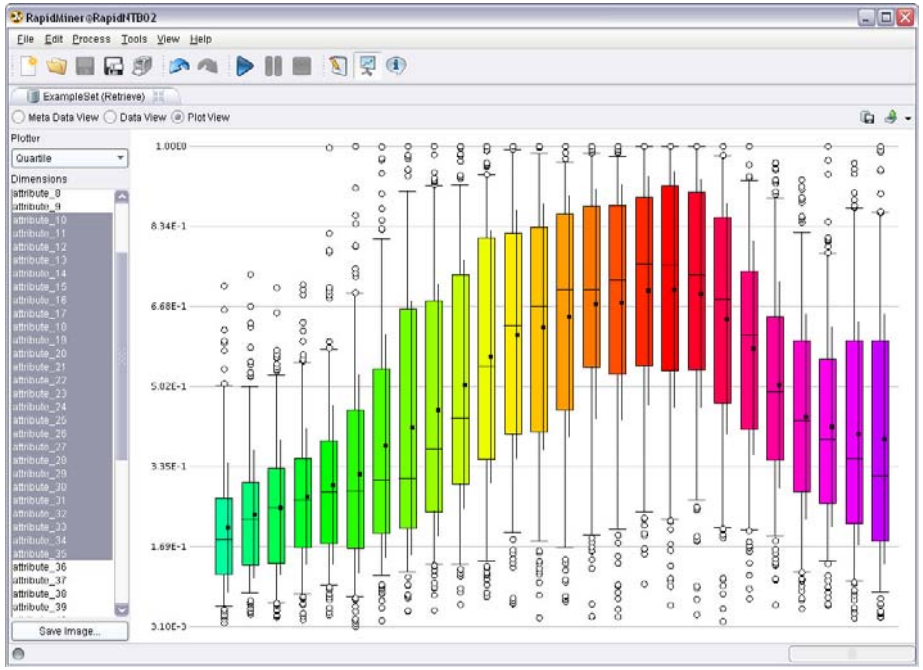
4. Website: **KDnuggets**
   - Overview of tools, online courses, events
   - http://www.kdnuggets.com/

Software

rapidminer

- Powerful data mining suite
- We are using Version 9.0 in the exercise

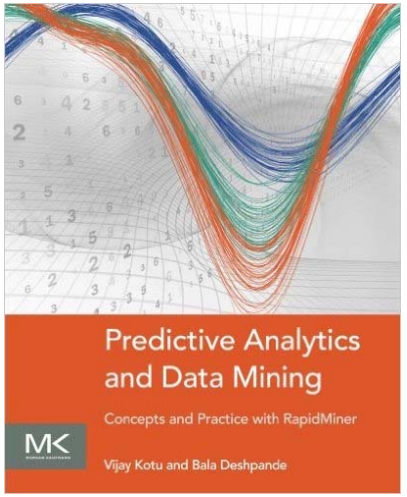


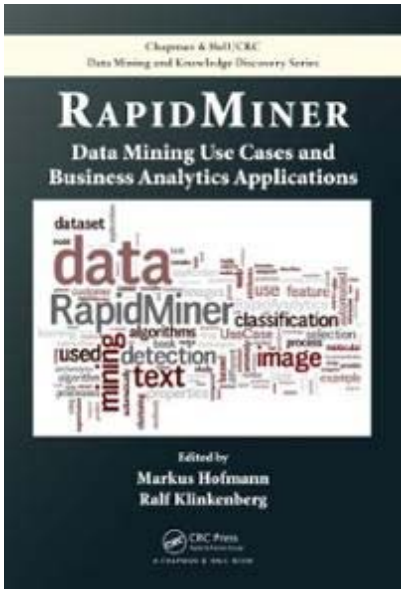


Gartner 2018 Magic Quadrant for Advanced Analytics Platforms

Literature – Rapidminer

1. **Rapidminer – Documentation**
   - http://docs.rapidminer.com
   - https://academy.rapidminer.com/catalog
   - https://www.youtube.com/user/RapidIVideos
2. Vijay Kotu, Bala Deshpande: **Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner.** Morgan Kaufmann.
   - 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





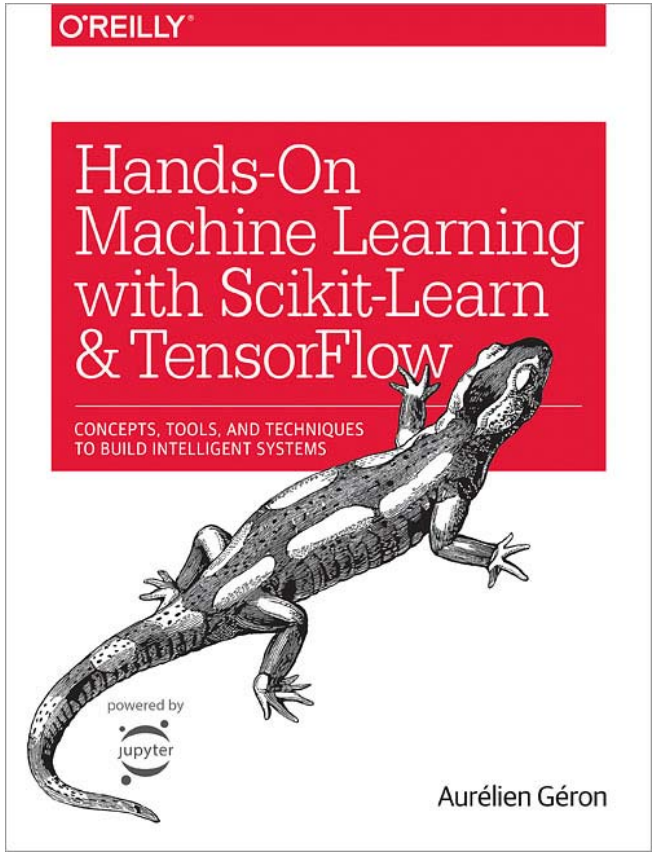
Software

- We use the Anaconda Python distribution, which includes:
  - Pandas, Numpy, Matplotlib
  - Scikit-learn
- Exercises are provided as Jupyter notebooks
- You will need to code!

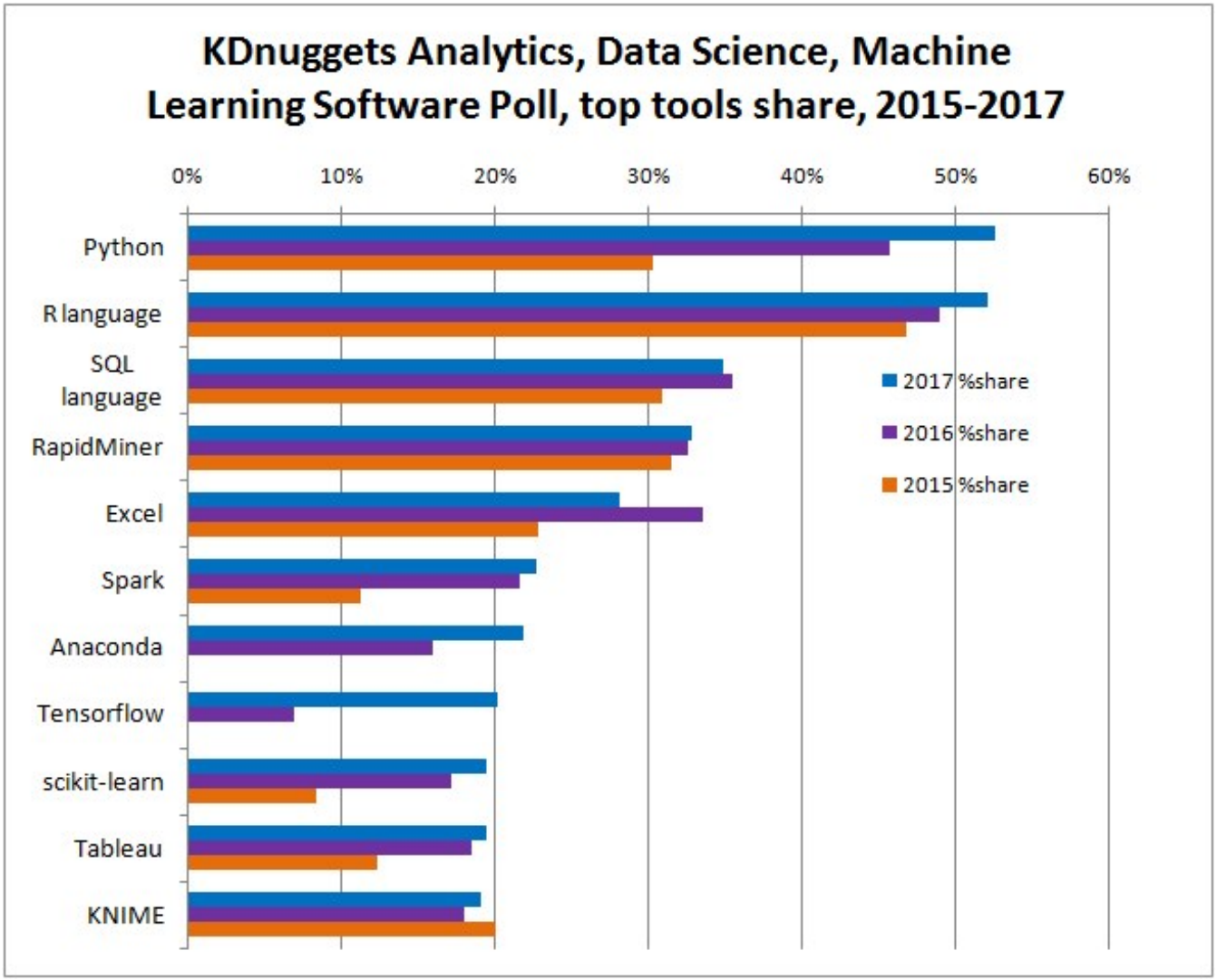

```
cv = StratifiedKFold(n_splits=10, shuffle=True, random_state=42)
naive_bayes = GaussianNB()
target_prediction = cross_val_predict(naive_bayes,
    credit_data.values, credit_target, cv=cv)
cm = confusion_matrix(credit_target, target_prediction)
cost = cm[0][1] * 100 + cm[1][0] * 1
acc = accuracy_score(credit_target, target_prediction)
print("Naive Bayes with accuracy of {} and cost {}".format(acc, cost))
print(confusion_matrix_report(credit_target, target_prediction))
```

Literature – Python

- **Python tutorial**: https://docs.python.org/3/tutorial/
- Great book: **Hands-on Machine Learning with Scikit-Learn** by Aurélien Géron (freely available online via the University library)



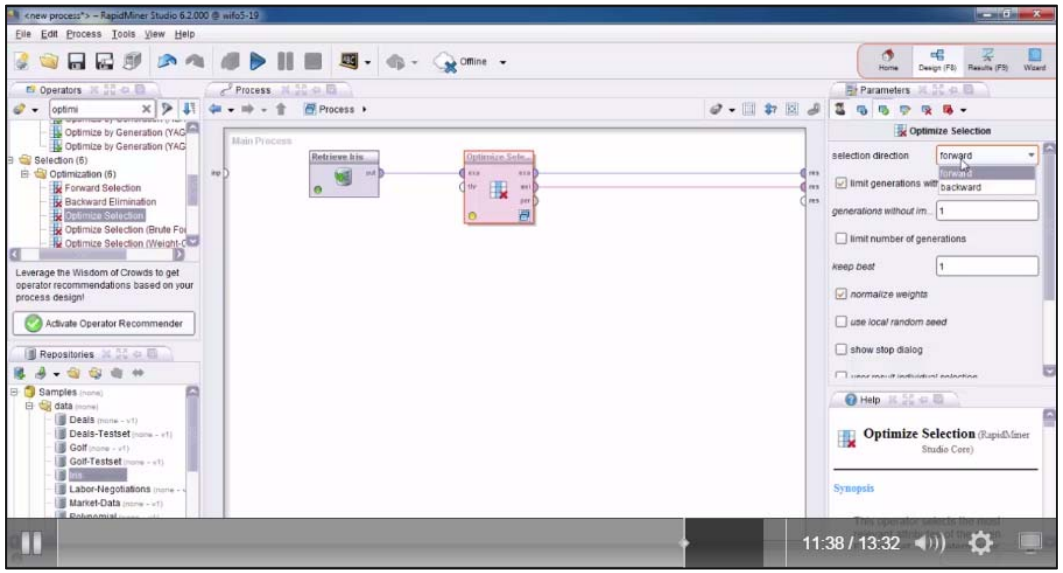
Usage of Python versus R versus RapidMiner



Source: KDnuggets online poll, 2900 votes
https://www.kdnuggets.com/2017/05/poll-analytics-data-science-machine-learning-software-leaders.html

Lecture Videos and Screencasts

1. Video recordings of all lectures from FSS 2015
2. Step-by-step introduction to relevant RapidMiner features
3. Step-by-step solutions and discussion of the exercises



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

Advertisement: Career Fair

20.03.2019 – 16:00-18:00 Uhr

MINT-MARKTPLATZ

Fakultät für Wirtschaftsinformatik & Wirtschaftsmathematik
B6, 30-32, Bauteil E-F (Neubau) im 1.OG

PRAKTIKA
FESTANSTELLUNGEN
VORTRÄGE

Accenture	EXA Deutschland GmbH	Porsche	Xenium
BCG Gamma	Inter-Versicherung	Roche Diagnostics	zeb
Commerzbank	KPMG	SAP AG	
d-fine GmbH	mayato	Scheer Group	
Essity GmbH	MSW & Partner	Stoccard	

Questions?