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
- Phone: +49 621 181 2677
- eMail: chris@informatik.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 1.03
- eMail: alex.brinkmann@informatik.uni-mannheim.de

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

Hallo

- **M. Sc. Inf. Sven Hertling**
- Graduate Research Associate
- Research Interests:
  - Semantic Technologies / Semantic Web
  - Linked Data
  - Knowledge Graphs
- Room: B6, 26, B 0.01
- eMail: sven@informatik.uni-mannheim.de

- Will teach one Python exercise group 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

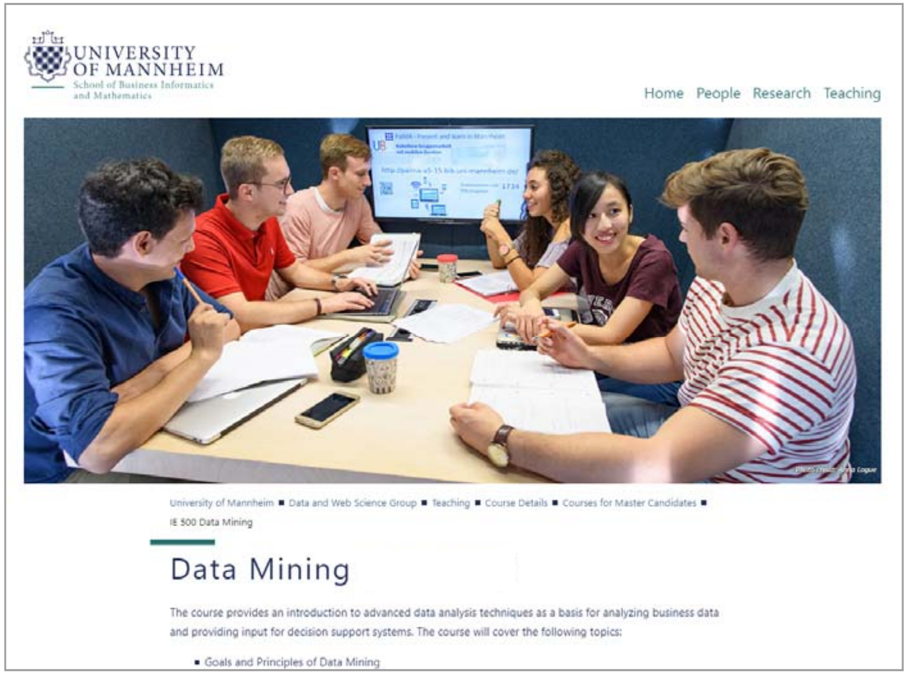
- Will teach one Python exercise group and will supervise student projects.

Course Organisation

- Lecture
  - introduces the principle 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, video lectures, 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, ZOOM
  - Exercise:
    - Thursday, 10.15 - 11.45,
Room A 104 (B6 , Bauteil A) – 16 places
    - Thursday, 12.00 - 13.30,
ZOOM (online) – unrestricted places
    - Thursday, 13.45 - 15.15,
Room A 104 (B6 , Bauteil A) – 16 places



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

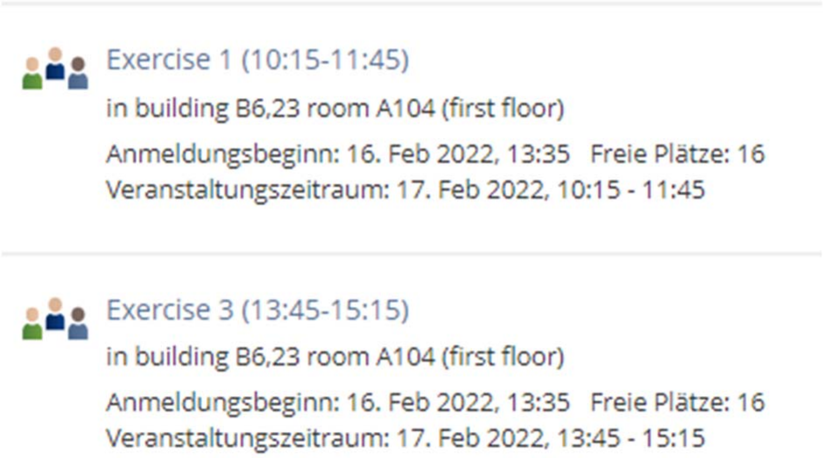
Data Mining

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

Registration for on-site exercises

- Registration for on-site exercises on a weekly basis via ILIAS
- Registration will be opened every Tuesday at 13:35
- Maximum number of slots per exercise: 16
- First Come, First Serve (with waitinglist)
- Please deregister if you change your mind, so others can take your place!
- Online exercise is unrestricted



Exercise 1 (10:15-11:45)
in building B6,23 room A104 (first floor)
Anmeldungsbeginn: 16. Feb 2022, 13:35 Freie Plätze: 16
Veranstaltungszeitraum: 17. Feb 2022, 10:15 - 11:45

Exercise 3 (13:45-15:15)
in building B6,23 room A104 (first floor)
Anmeldungsbeginn: 16. Feb 2022, 13:35 Freie Plätze: 16
Veranstaltungszeitraum: 17. Feb 2022, 13:45 - 15:15

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

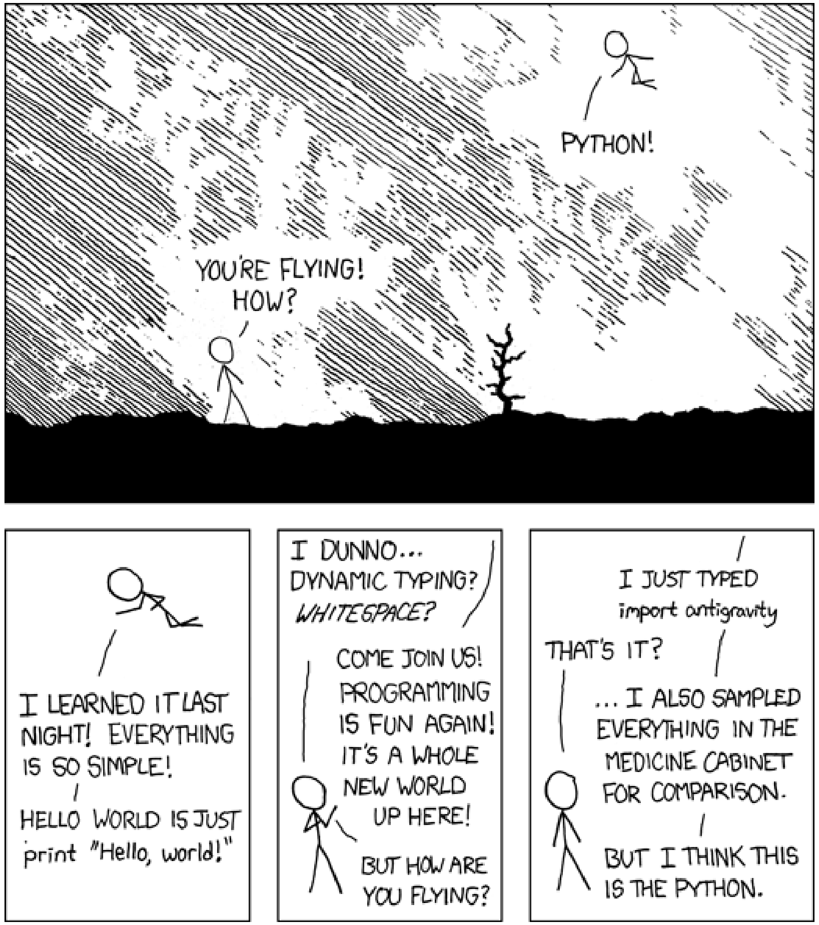
Schedule

Bold = session takes place live via ZOOM

Week	Wednesday	Thursday
16.02.2022	**Lecture: Introduction to Data Mining** **Tutorial: Introduction to Python**	Exercise: Preprocessing/Visualization
23.02.2022	Video Lecture: Cluster Analysis	Exercise: Cluster Analysis
30.02.2022	Video Lecture: Classification 1	Exercise: Classification
02.03.2022	Video Lecture: Classification 2	Exercise: Classification
09.03.2022	Video Lecture: Classification 3 **Question and Answer Session 1**	Exercise: Classification
16.03.2022	Video Lecture: Regression	Exercise: Regression
23.03.2025	Video Lecture: Text Mining	Exercise Text Mining
30.03.2022	Video Lecture: Association Analysis **Question and Answer Session 2**	Exercise Association Analysis
06.04.2022	**Introduction to the Student Projects and Group Formation**	Preparation of Project Outlines
	- Easter Break	
27.04.2022	**Feedback on Project Outlines**	Project Work
04.05.2022	**Feedback on demand**	Project Work
11.05.2022	**Feedback on demand**	Project Work
25.05.2022	**Feedback on demand**	Project Work
01.06.2022	**Presentation of project results**	**Presentation of project results**

Introduction to Python

- Today (16.02) at 15:30-17:00 in ZOOM-Lehre-051 (Online!)
- Topics:
  - Setup of environment (Anaconda, Jupyter Notebooks)
  - Python Infos / Design Goals
  - Basic programming concepts in Python
- Support
  - Help with environment setup
  - Q&A
- Material
  - Tutorial slides available in ILIAS
  - Try to install Anaconda before the tutorial


PYTHON!
YOU'RE FLYING! HOW?
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.

Deadlines

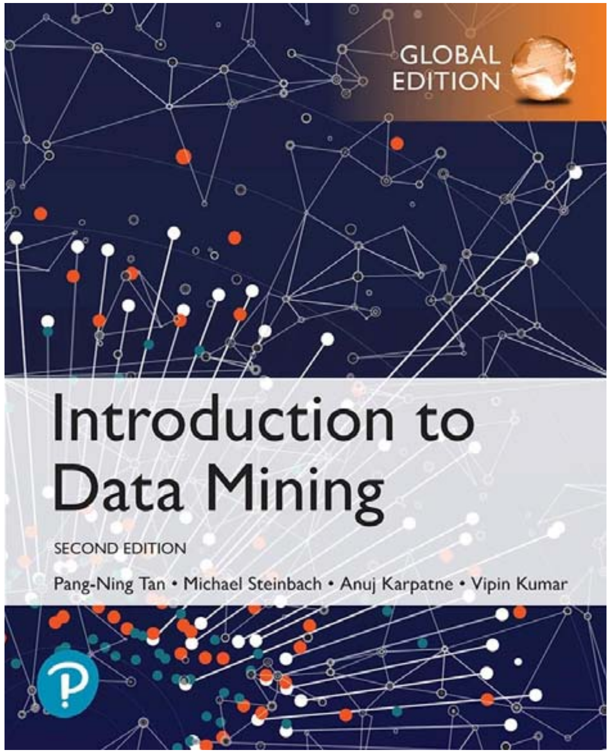
- Submission of project proposal
  - Friday, April 22nd, 23:59
- Submission of final project report
  - Sunday, May 29th, 23:59
- Project presentations
  - Wednesday June 1st, Thursday, June 2nd
  - everyone has to attend the presentations

Final Exam

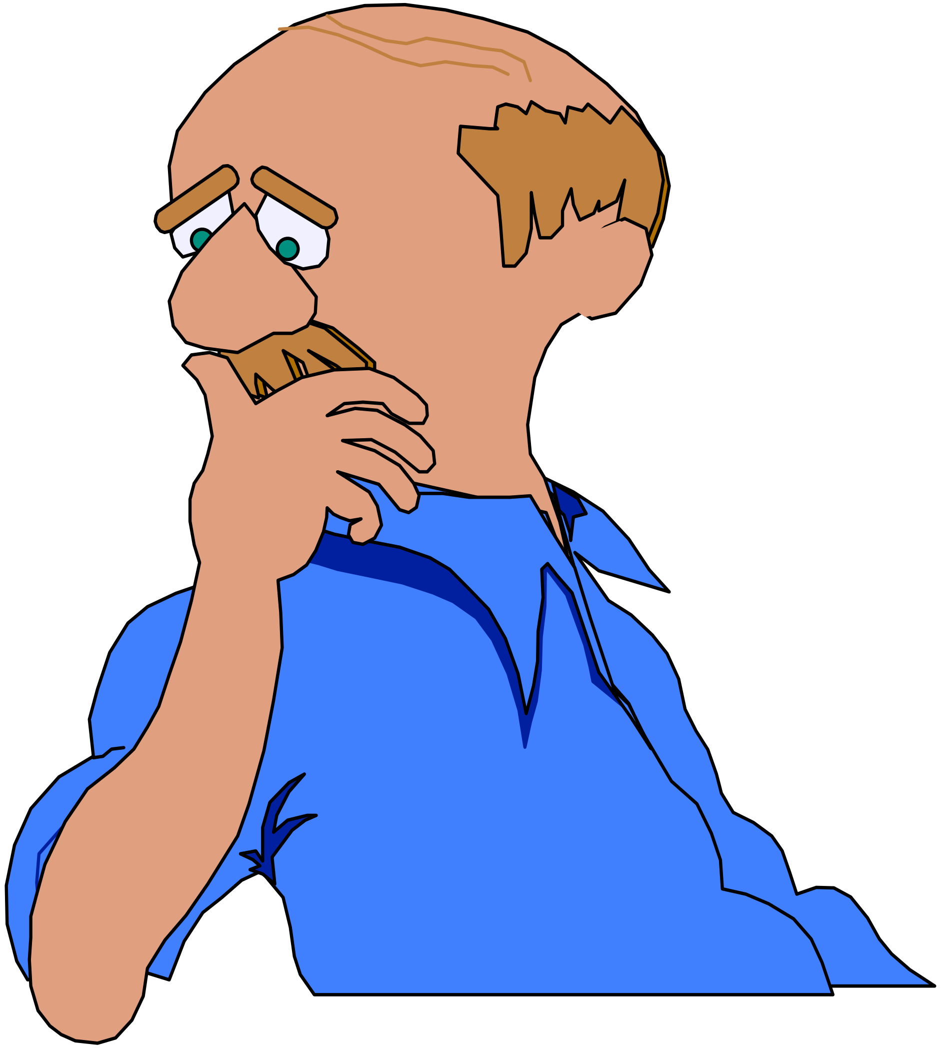
- Date and Time: tbd
- Room: tba
- Duration: 60 minutes
- Structure: 6 open questions that
  - Goal is to check whether you have understood the lecture content
    - we try to cover all major chapters of the lecture: clustering, classification, regression, association analysis, text mining
  - Require 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.


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

Questions?

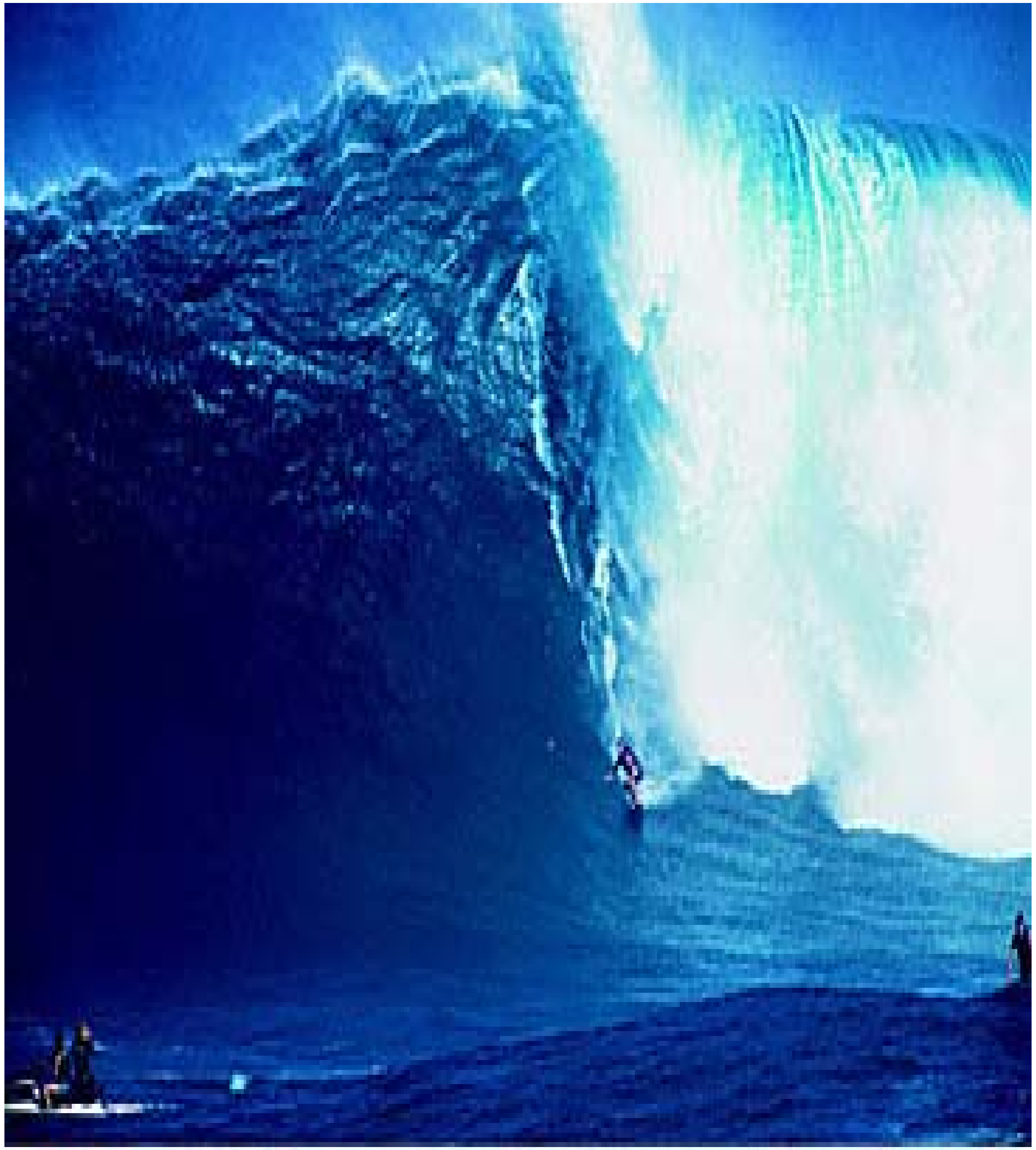


Outline: Introduction to Data Mining

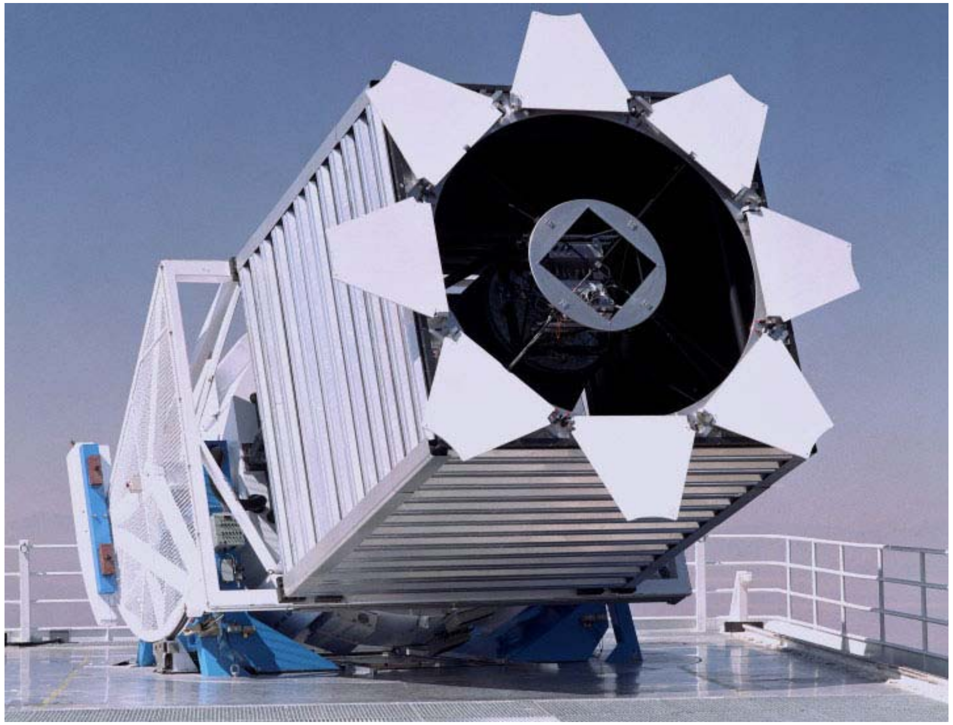
1. What is Data Mining?
2. Tasks and Applications
3. The Data Mining Process
4. Data Mining Software

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...”

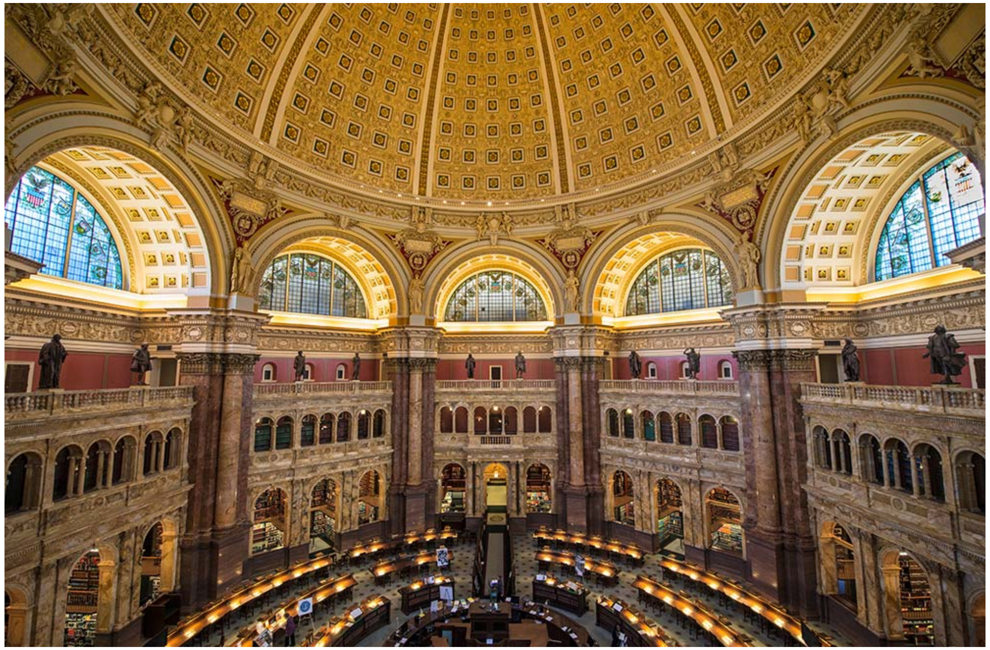


Sloan Digital Sky Survey
≈ 200 GB/day
≈ 73 TB/year

Predict

- Type of sky object: Star or galaxy?

“We are Drowning in Data...”



US Library of Congress

≈ 235 TB archived

≈ 40 Wikipedias

arXiv Preprint Server

> 2 million papers

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

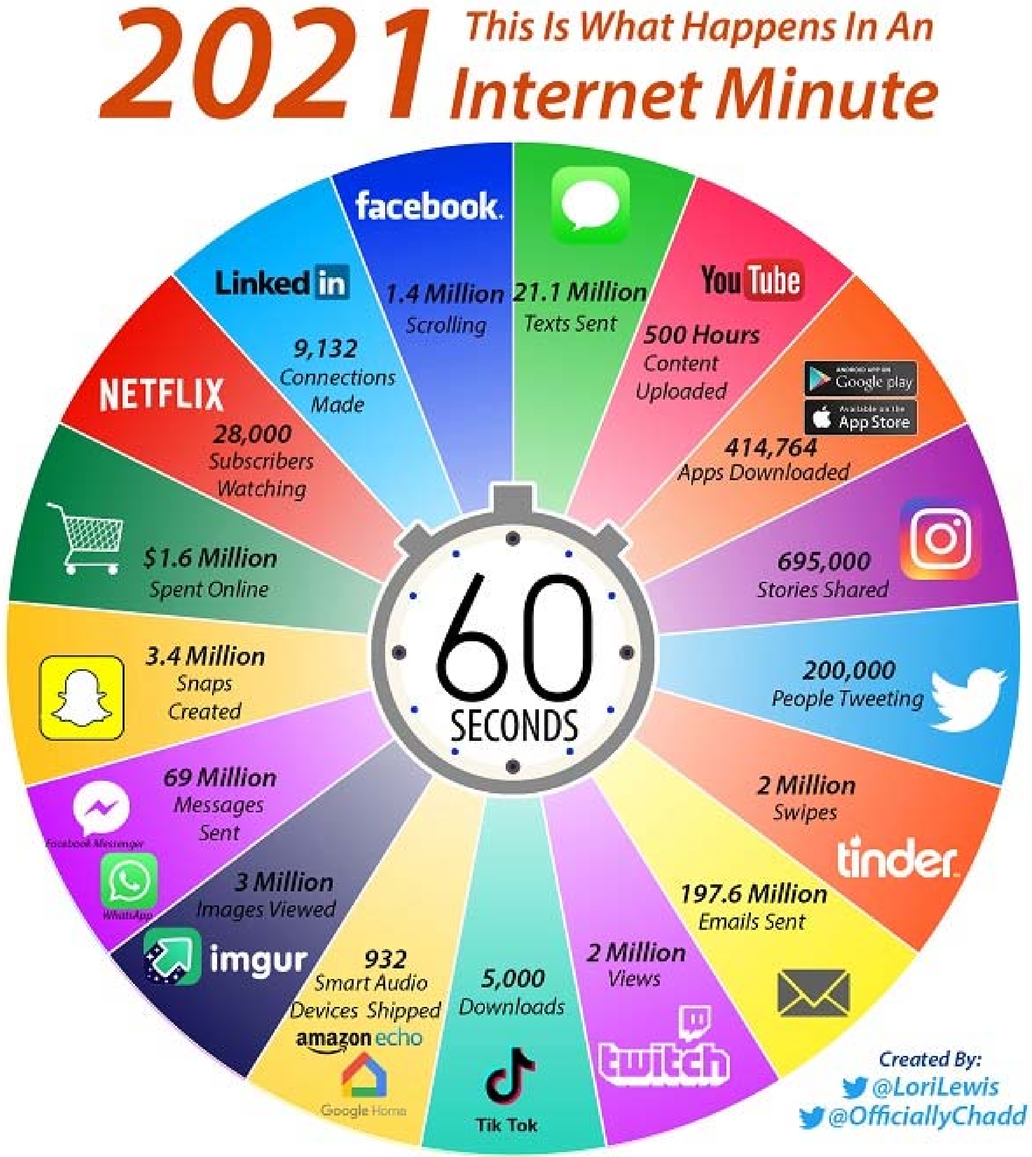
- 4 Petabyte of new data generated every day
- over 300 Petabyte in Facebook‘s data warehouse

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...”

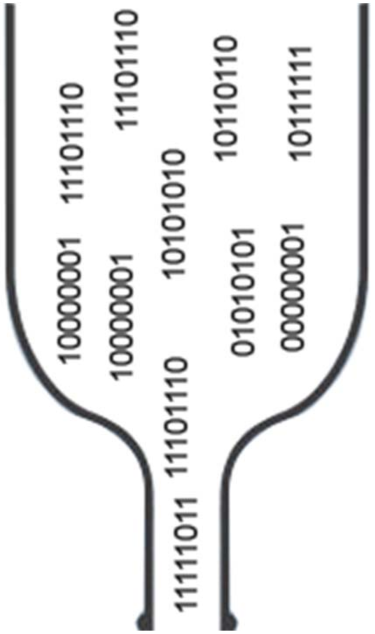
Law enforcement agencies collect unknown amounts of data from various sources

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

Predict

- Terrorist or not?
- Trustworthiness

“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

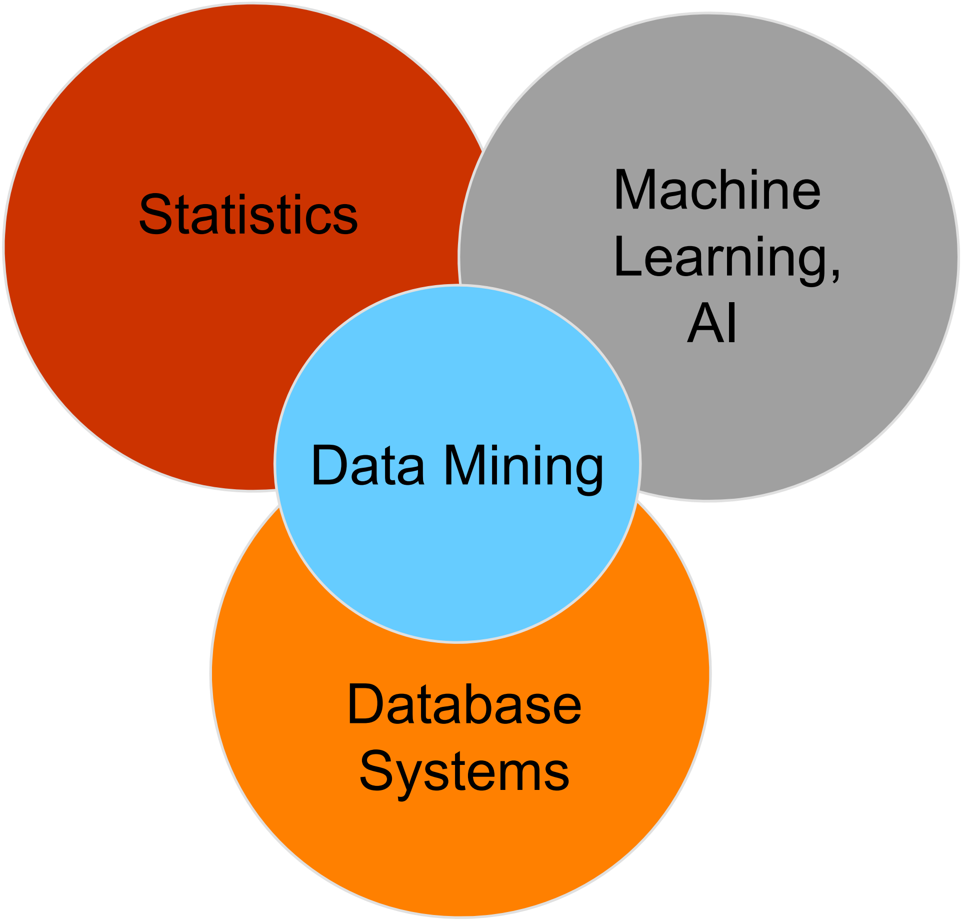
- **implicit,**
- **previously unknown, and**
- **potentially useful**

information from data.

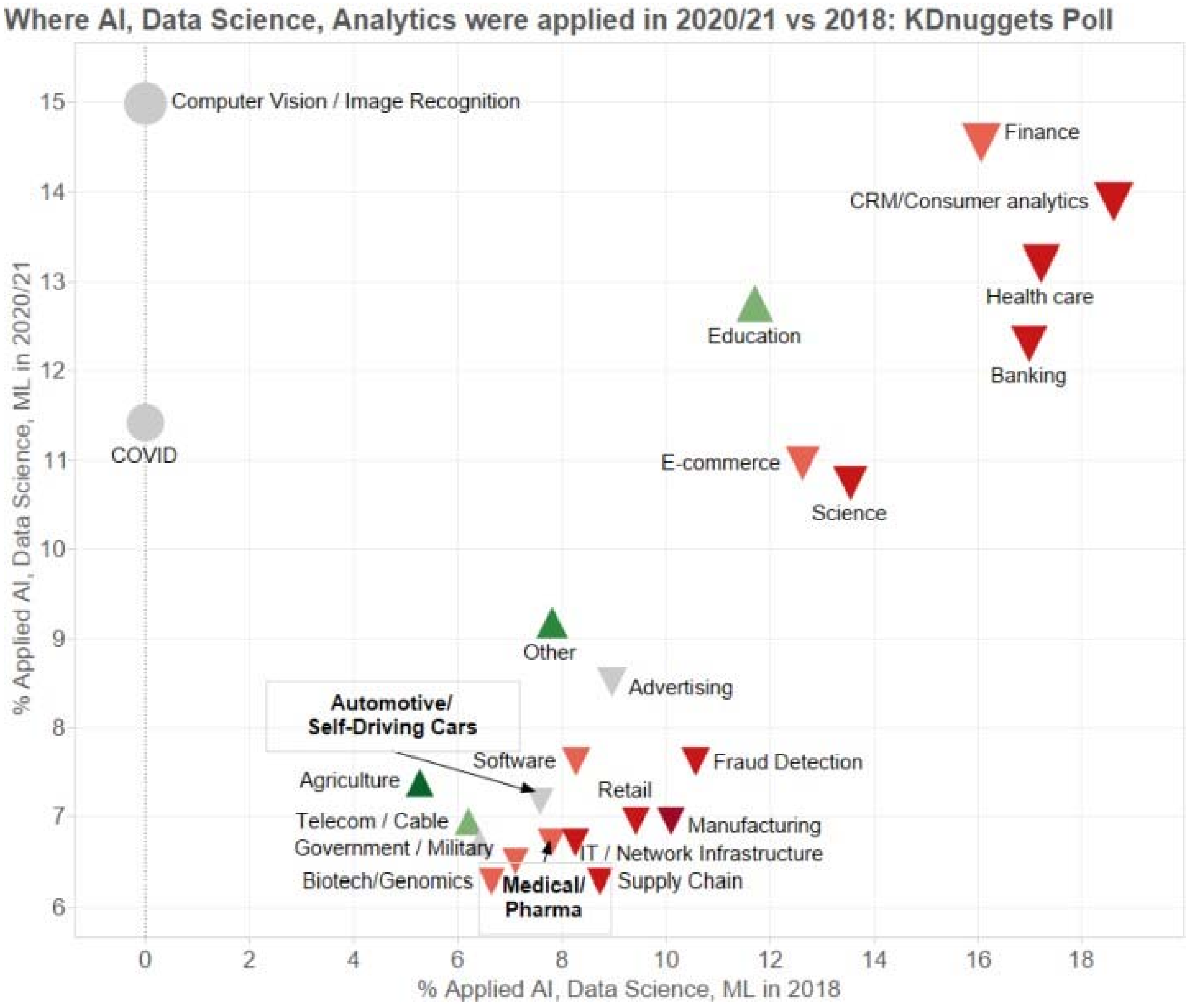
- Data Mining methods
  1. detect interesting patterns in large quantities of data
  2. **support** human decision making by providing such patterns
  3. **predict** the outcome of a future observation based on the patterns

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

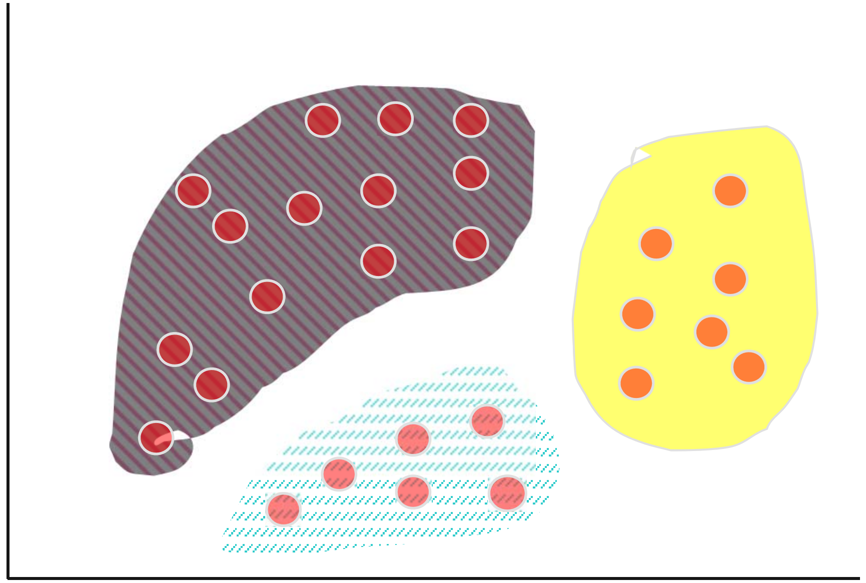
- Machine Learning Terminology
  - descriptive = unsupervised
  - predictive = supervised

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

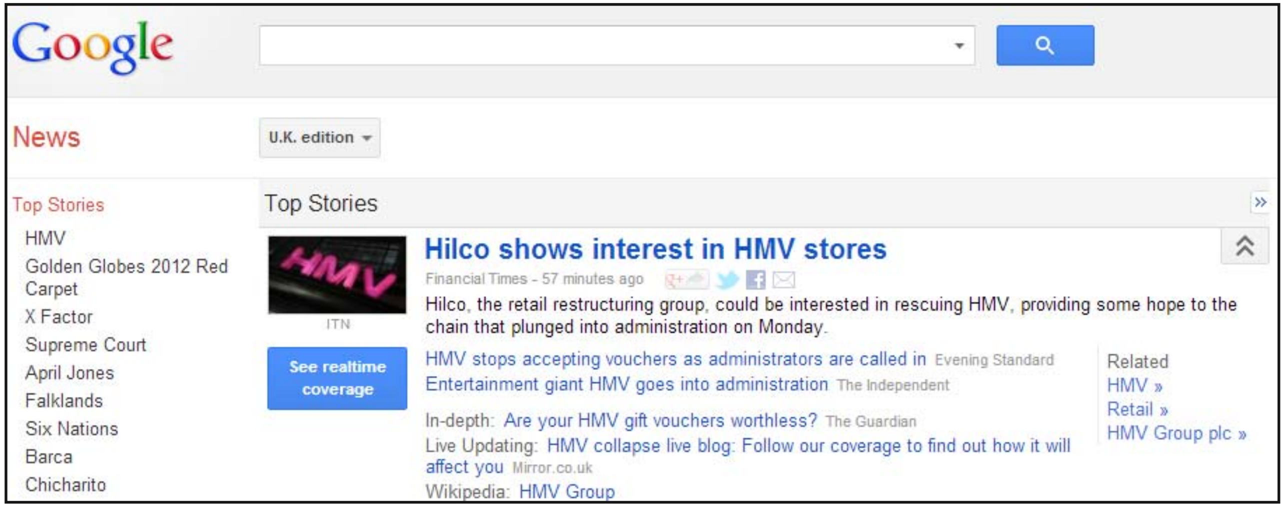
Cluster Analysis: Application 2

- 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

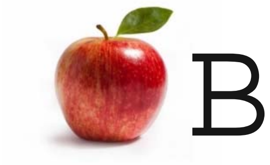
In-depth: Are your HMV gift vouchers worthless? The Guardian

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"

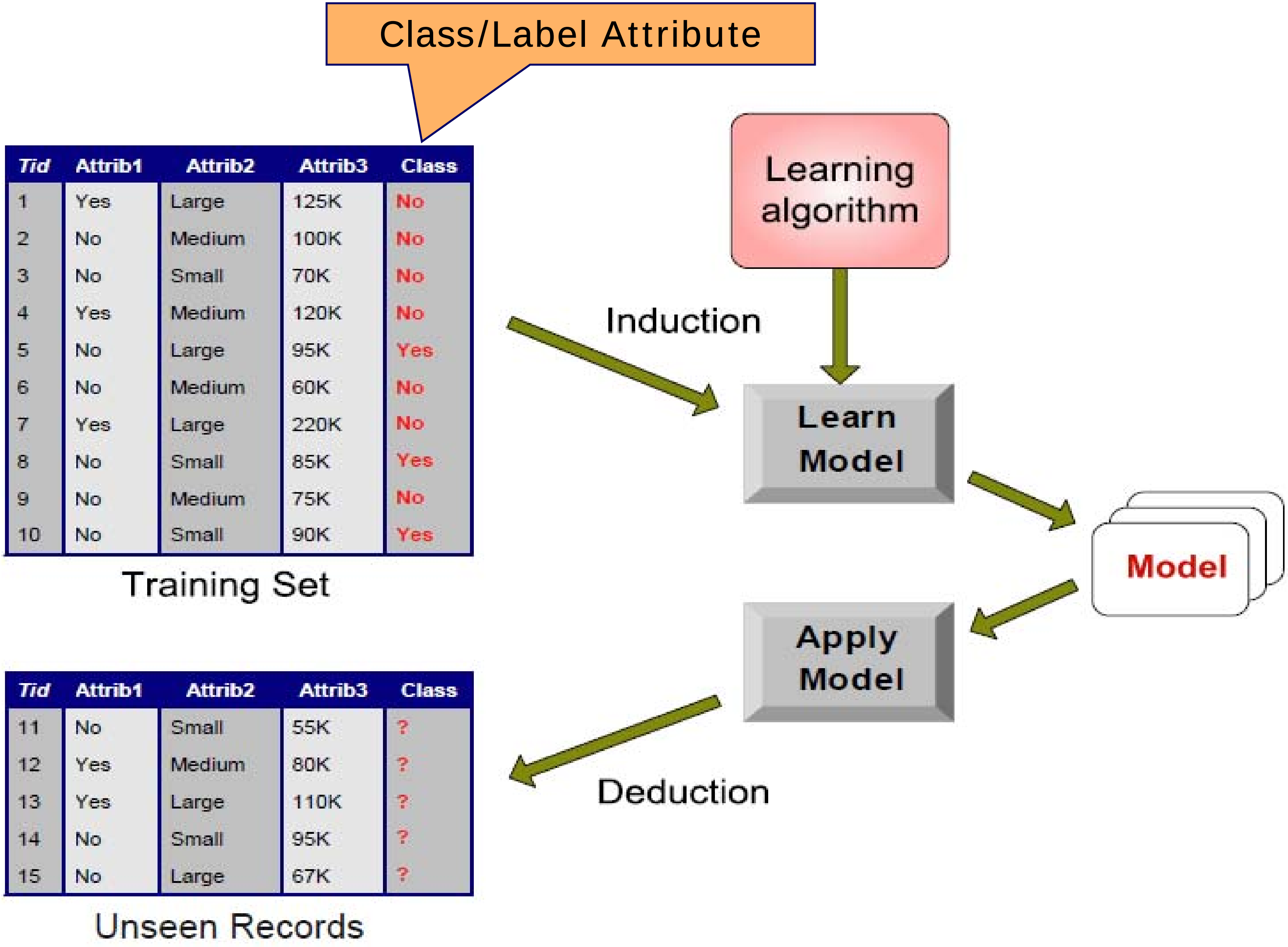
"not a tree"

"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

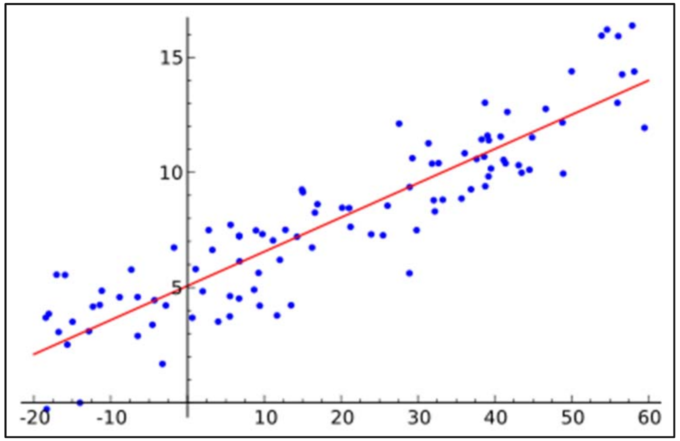
Classification: Application 2

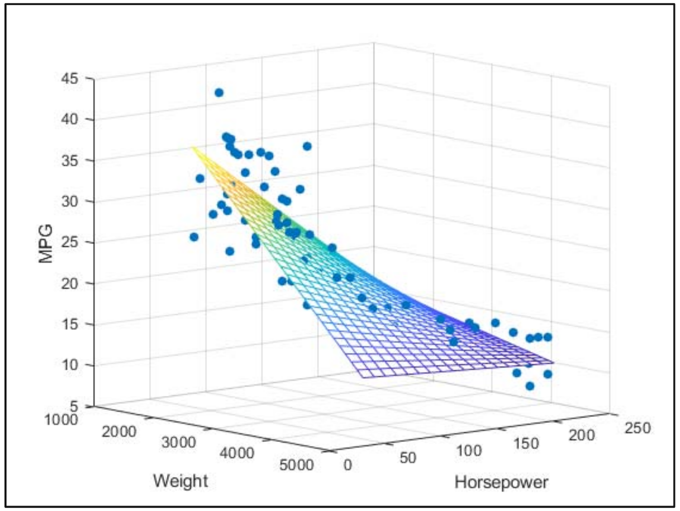
- Application area: Direct Marketing

- 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 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}

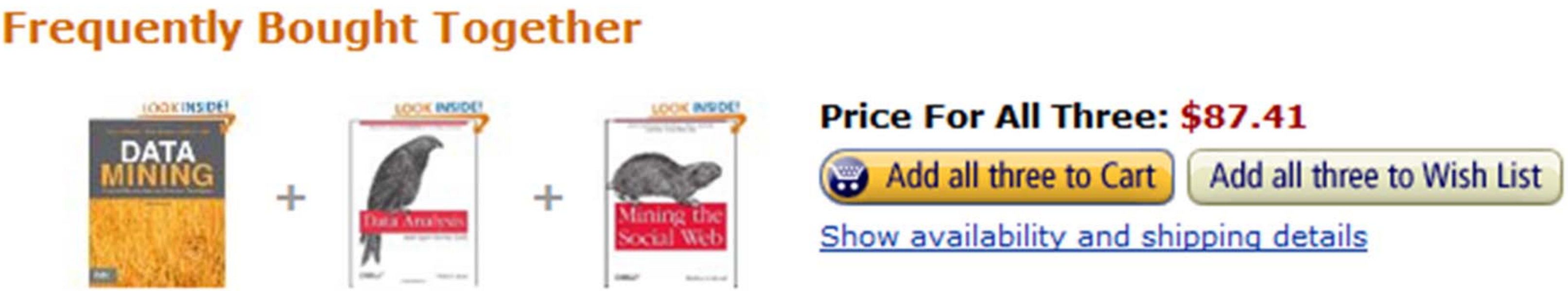
{Milk} --> {Coke}

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

amazon.com



Frequently Bought Together

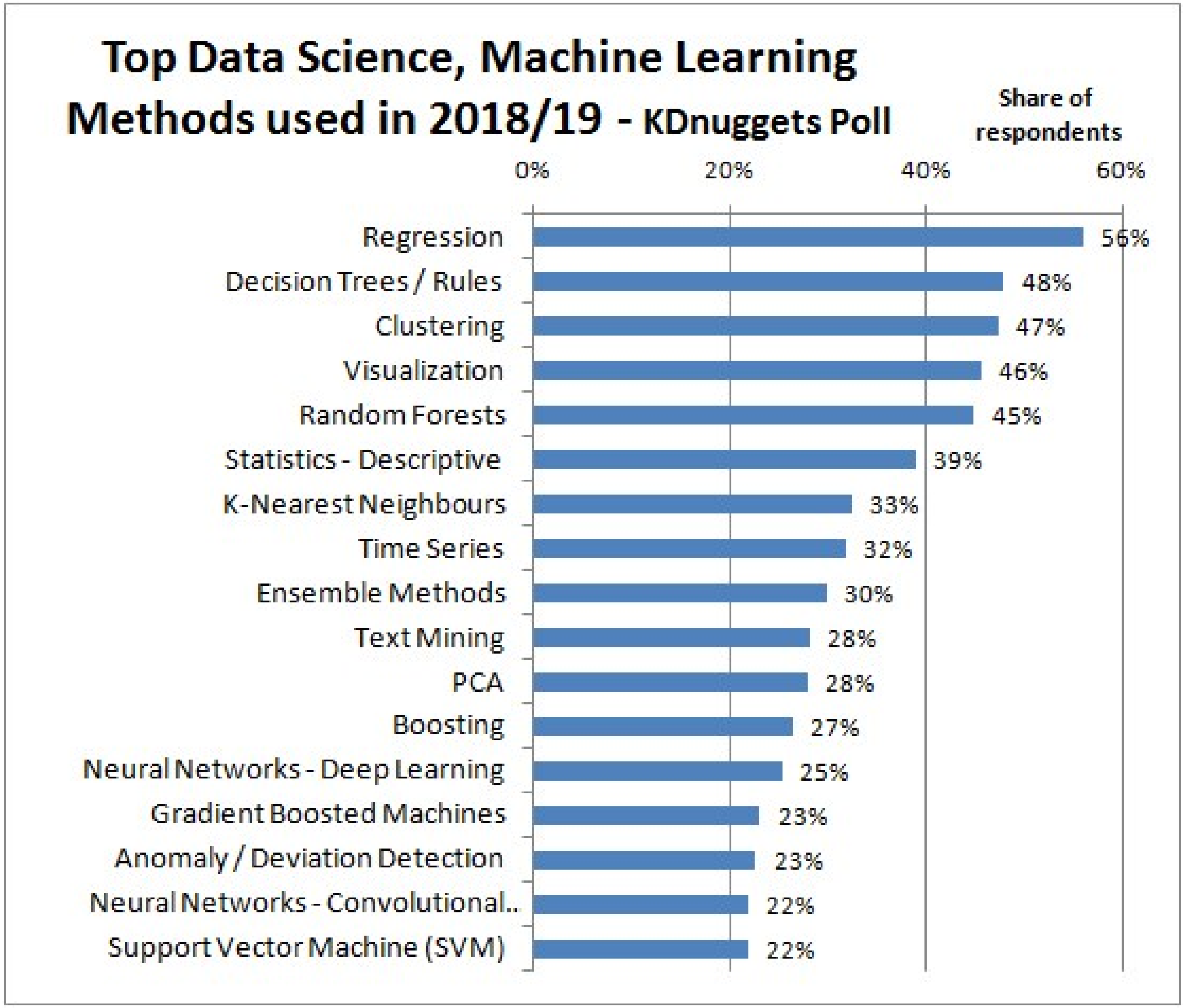
Price For All Three: $87.41

Association Rule Discovery: Application 2

- Application area:
  Inventory Management

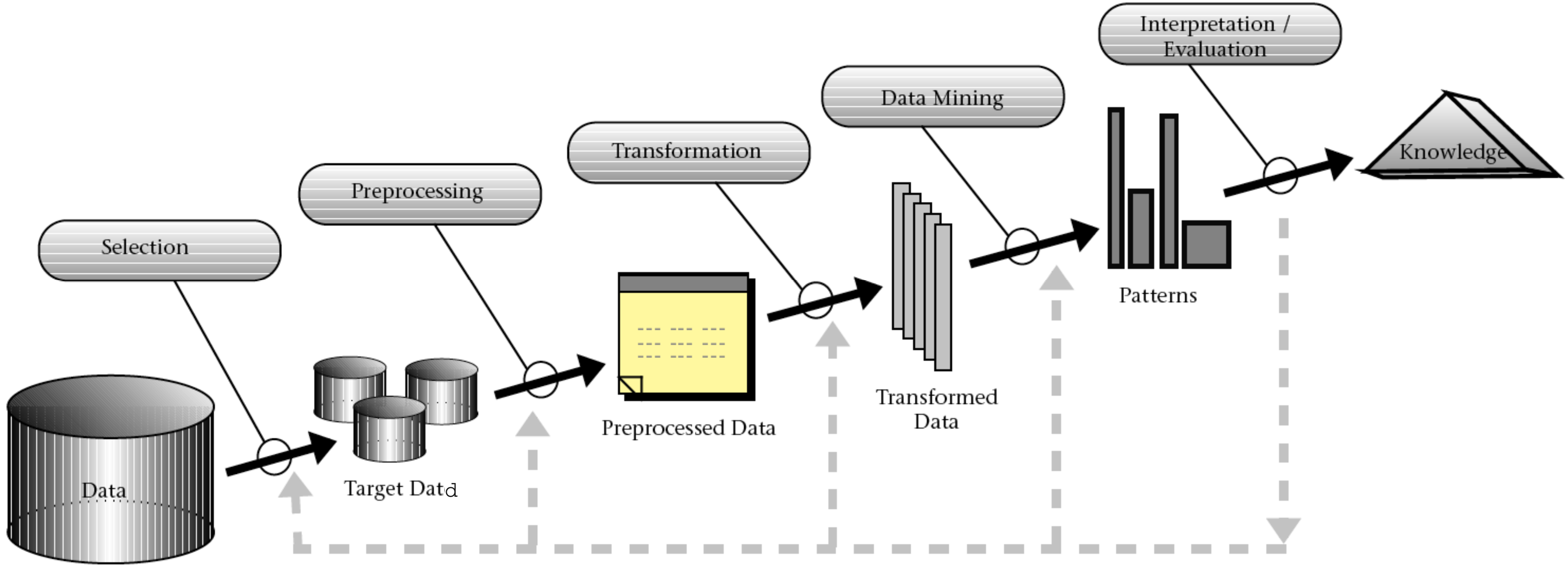
- Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households
- Approach: Process the data on tools and parts required in previous repairs at different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

Which Methods are Used in Practice?



Source: 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

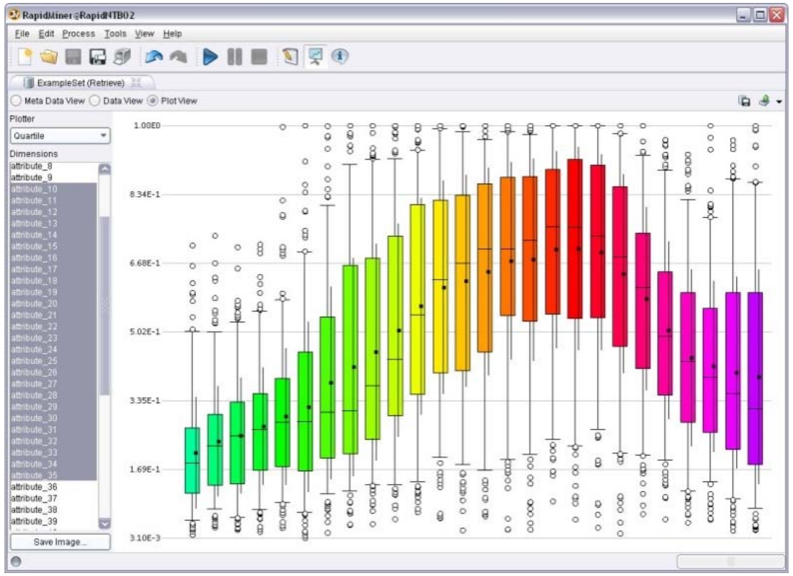
3. The Data Mining Process

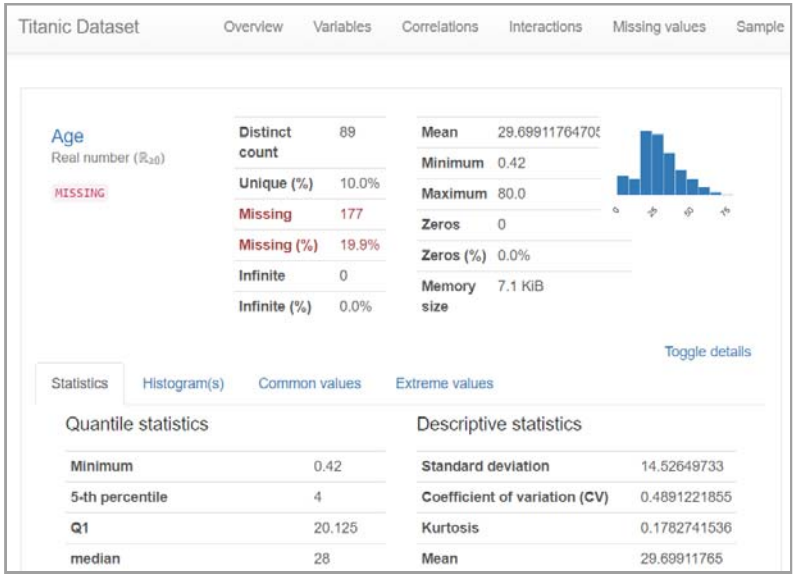


Source: Fayyad et al. (1996)

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
Unique (%)	10.0%
Missing	177
Missing (%)	19.9%
Infinite	0
Infinite (%)	0.0%

Mean	29.6991176470
Minimum	0.42
Maximum	80.0
Zeros	0
Zeros (%)	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

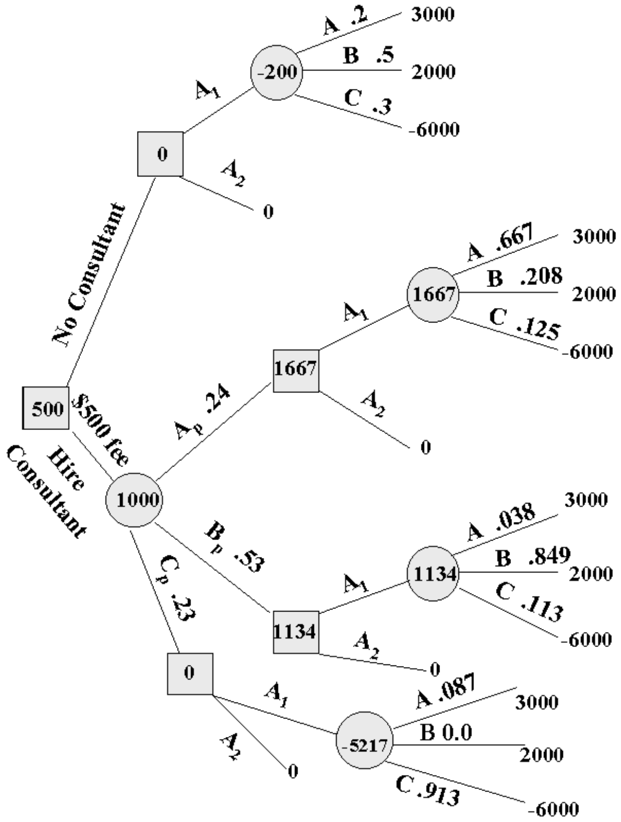
3.2 Preprocessing and Transformation

- 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

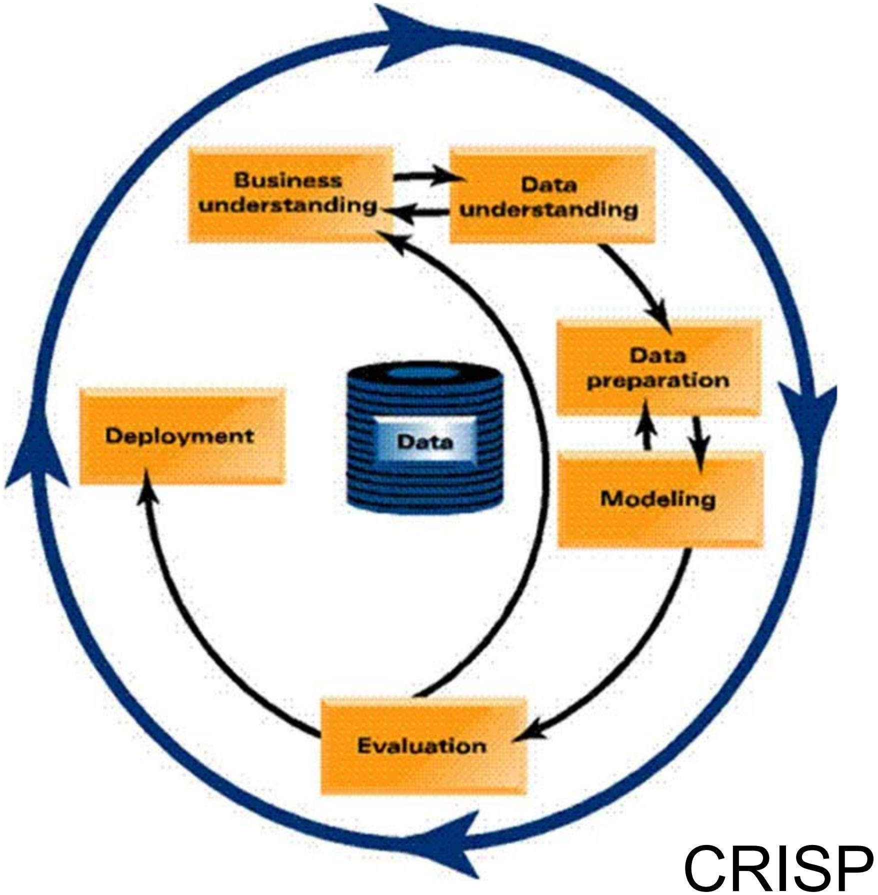
- Input: Preprocessed Data
- Output: Model / Patterns

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



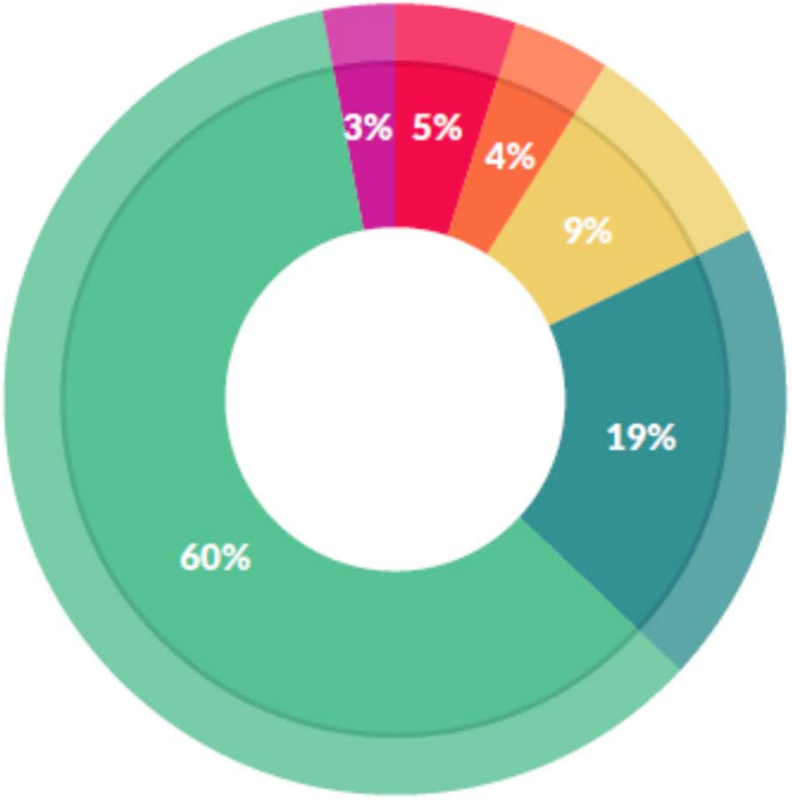
3.4 Deployment

- 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

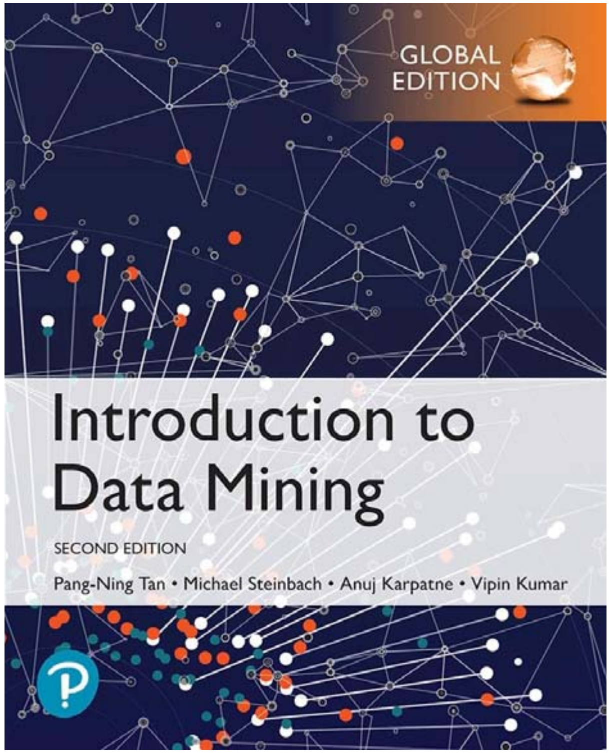
- ● Building training sets: 3%
- ● Cleaning and organizing data: 60%
- ● Collecting data sets; 19%
- ● Mining data for patterns: 9%
- ● Refining algorithms: 4%
- ● Other: 5%

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.

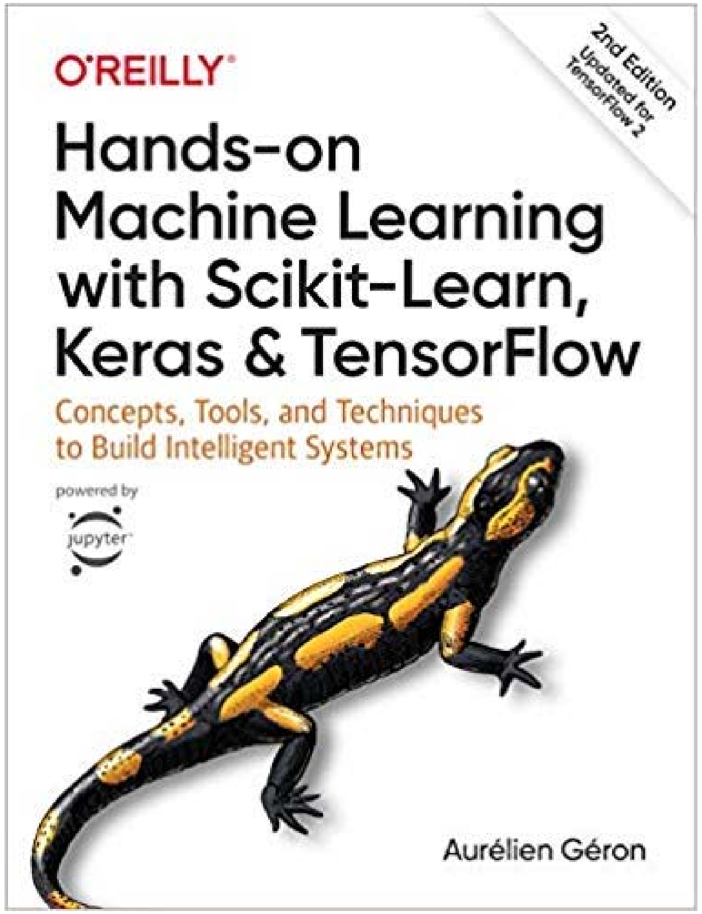

GLOBAL EDITION
Introduction to Data Mining
SECOND EDITION
Pang-Ning Tan • Michael Steinbach • Anuj Karpatne • Vipin Kumar
P

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. attend exercise **Preprocessing/Visualization** on Thursday
  4. watch the video lectures introducing **Cluster Analysis**
  5. attend exercise **Cluster Analysis** next week
  6. …