Data Mining

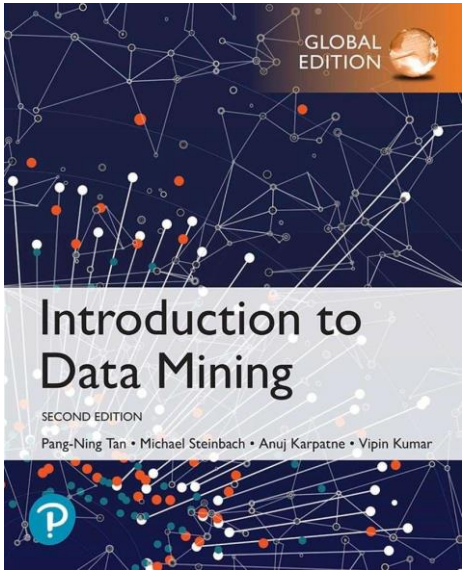
Introduction to Data Mining

Targeted Contents

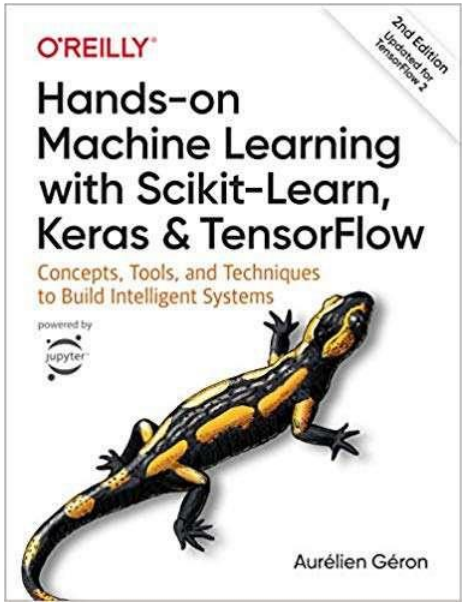
1. Introduction to Data Mining	What is Data Mining? Tasks and Applications The Data Mining Process
2. Cluster Analysis	K-means Clustering, Density-based Clustering, Hierarchical Clustering, Proximity Measures
3. Classification	Nearest Neighbor, Decision Trees and Forests, Rule Learning, Naïve Bayes, 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

Recommended Books

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



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



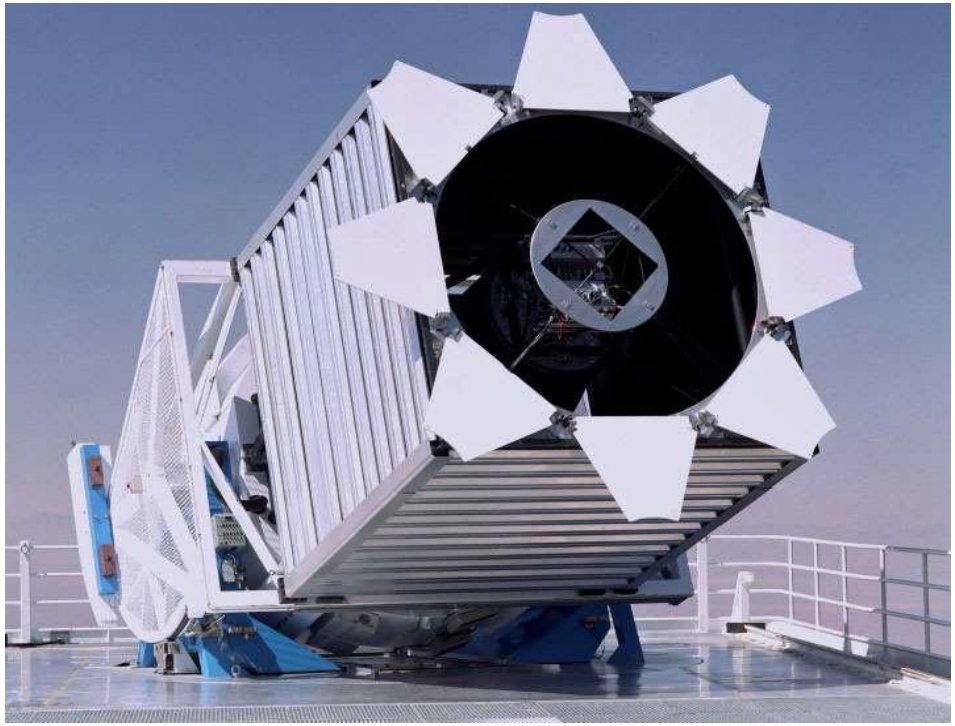
Outline: Introduction to Data Mining

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

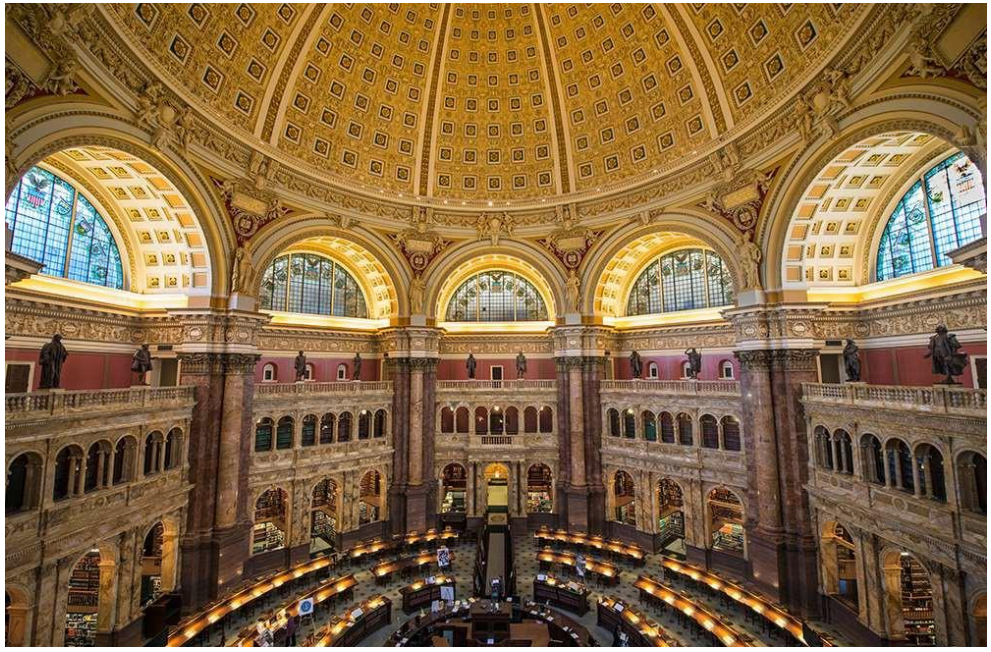
“We are Drowning in Data...”



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...”

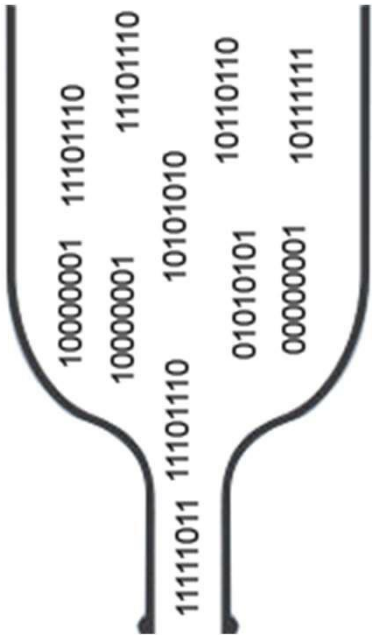
Law enforcement agencies collect unknown amounts of data from various sources

- 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

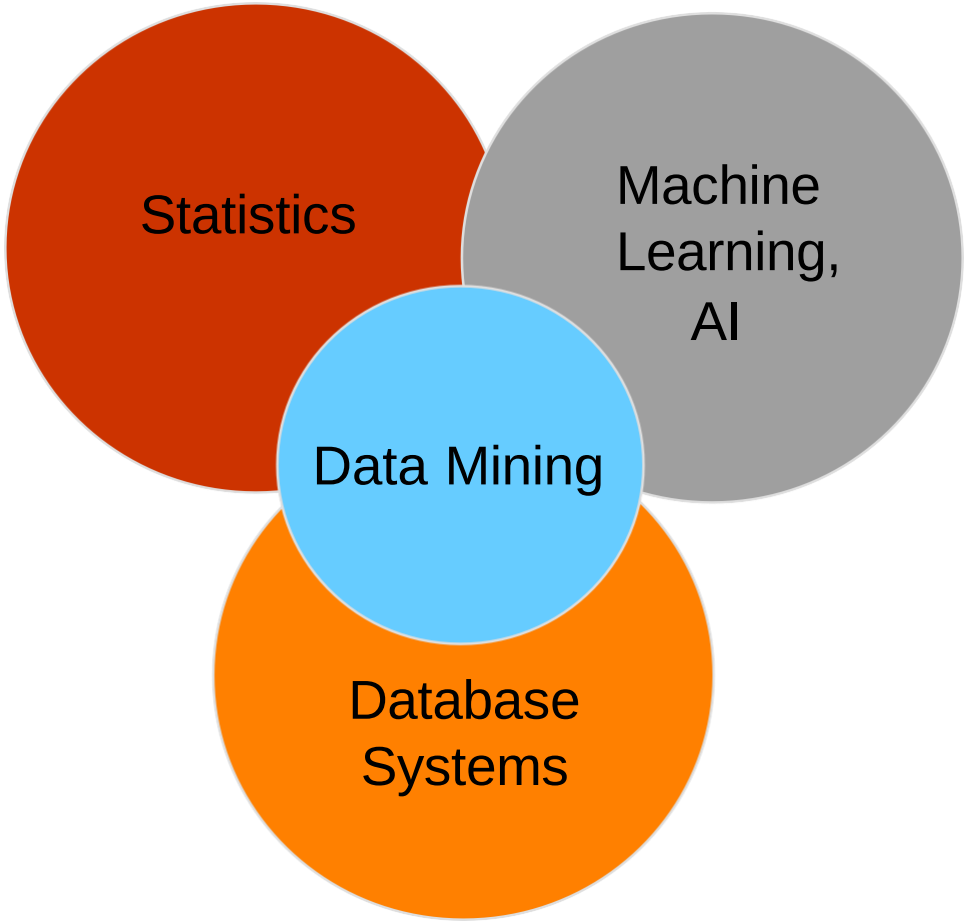
Exploration & analysis,
of large quantities of data
in order to discover
meaningful patterns.

Non-trivial extraction of
– **implicit,**
– **previously unknown, and**
– **potentially useful**
information from data.

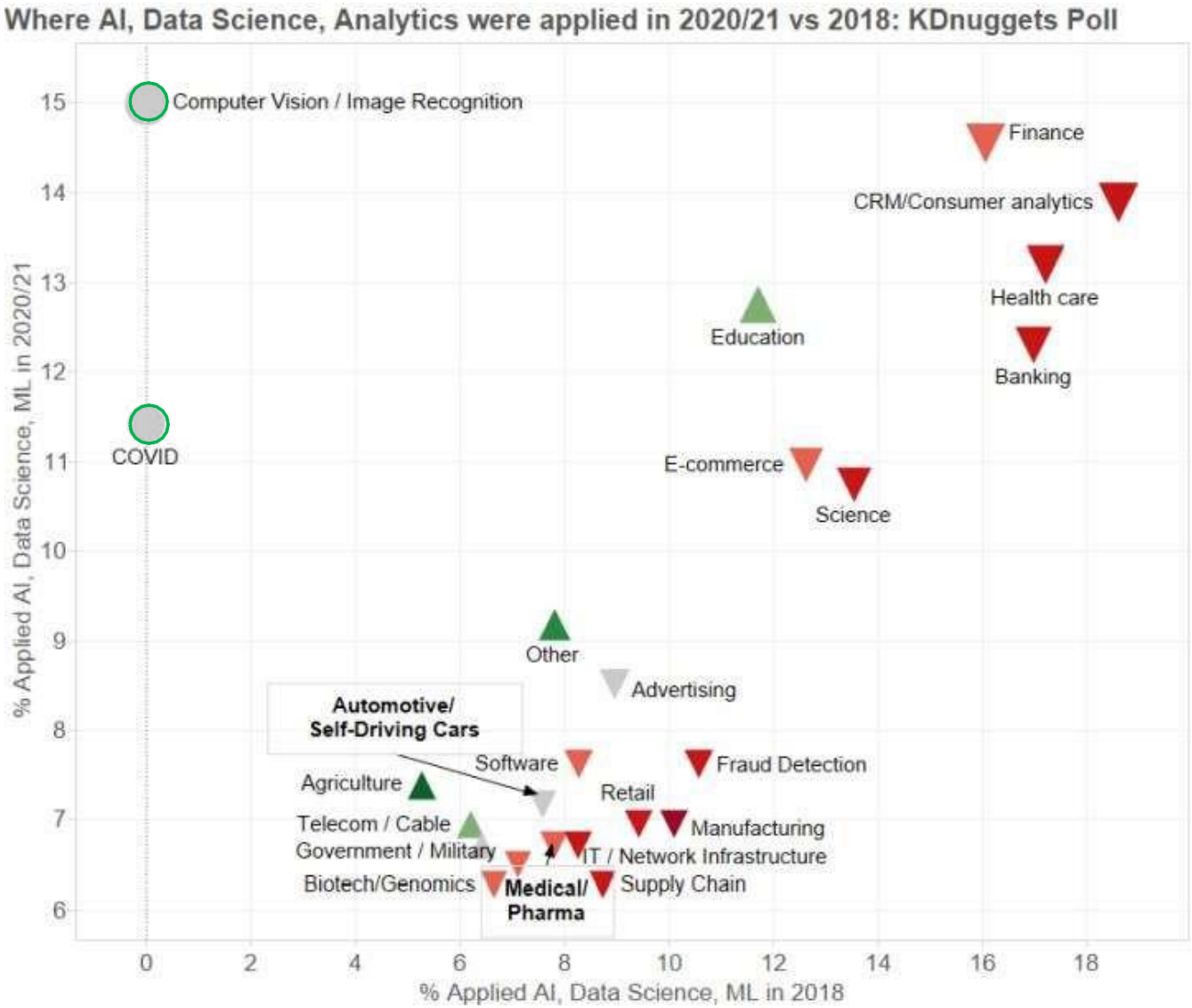
– Data Mining methods
  1. detect interesting patterns in large quantities of data
  2. **support** human decision making by providing such patterns
  3. **predict** the outcome of a future observation based on the patterns

Origins of Data Mining

- Data Mining combines ideas from statistics, machine learning, artificial intelligence, and database systems
- Tries to overcome 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, 447 (2021) and 435 (2018) participants
https://www.kdnuggets.com/2021/06/poll-where-analytics-data-science-ml-applied.html

2. Tasks and Applications

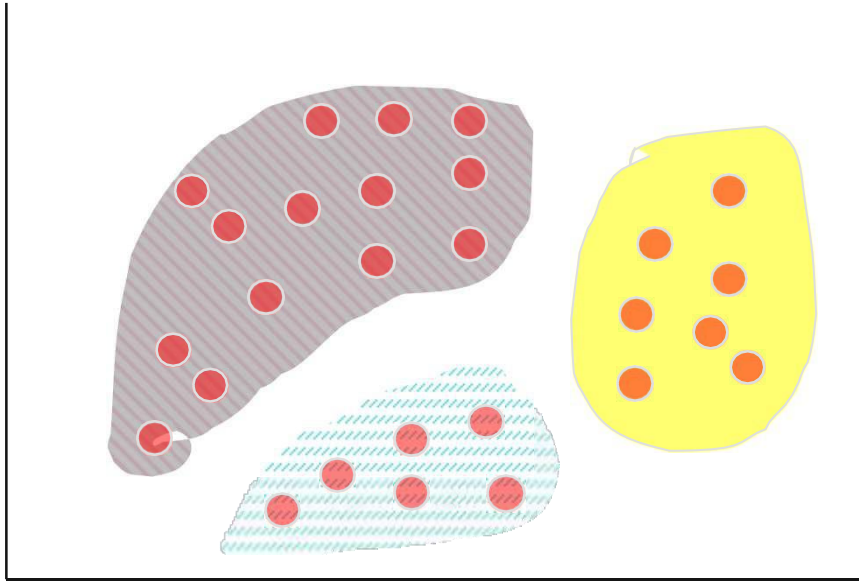
- Descriptive Tasks
  - Goal: Find patterns in the data.
  - Example: W*hich products are often bought together?*
- Predictive Tasks
  - Goal: Predict unknown values of a variable
    - given observations (e.g., from the past)
  - Example: W*ill a person click a online advertisement?*
    - given her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

Data Mining Tasks

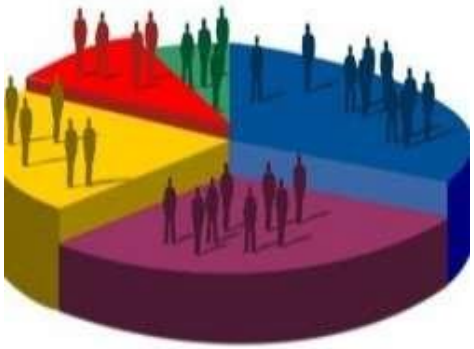
1. Cluster Analysis [Descriptive]
2. Classification [Predictive]
3. Regression [Predictive]
4. Association Analysis [Descriptive]

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



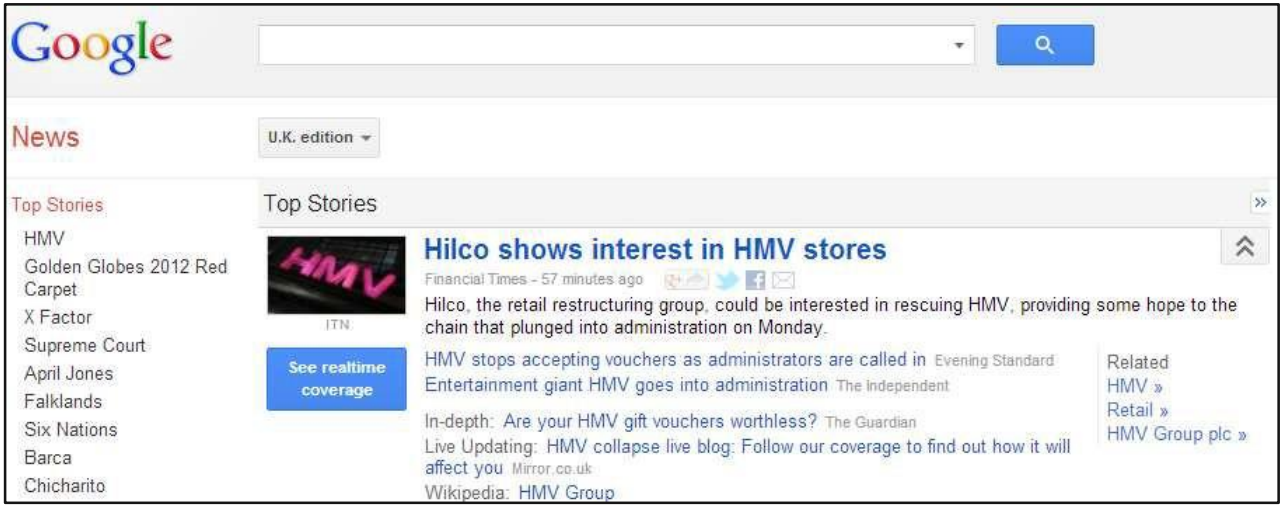
Cluster Analysis: Application 1



- 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

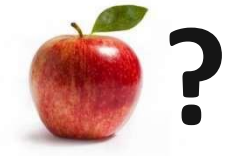
Cluster Analysis: Application 2

- 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



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"

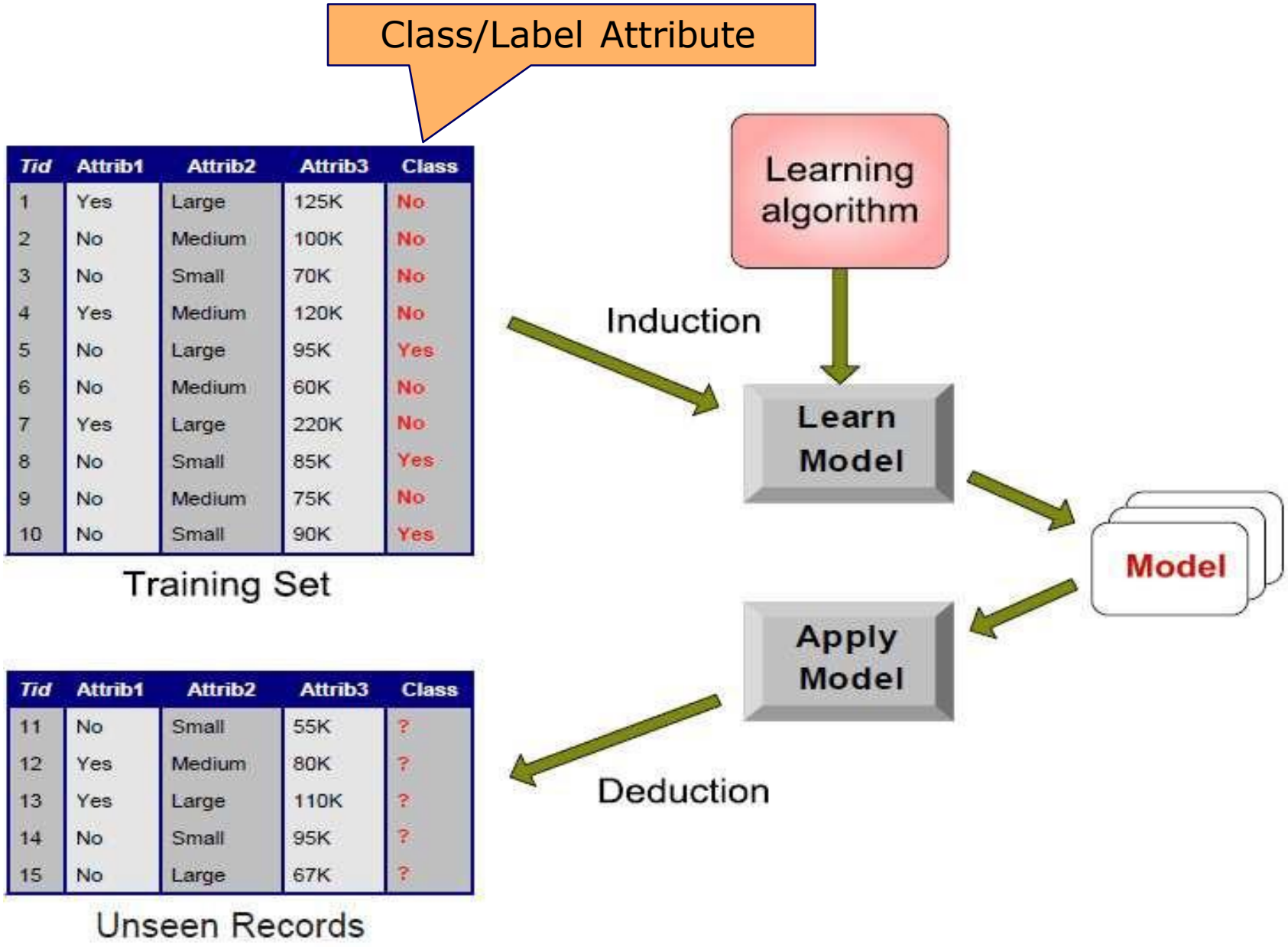
"not a tree"

"not a tree"

"not a tree"

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

Classification: Workflow



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

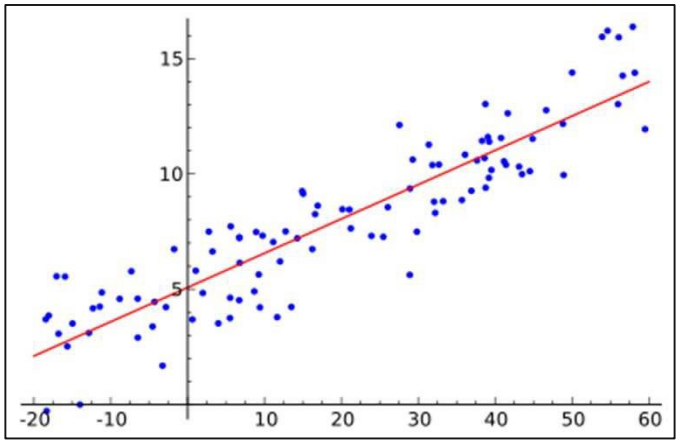
Classification: Application 2

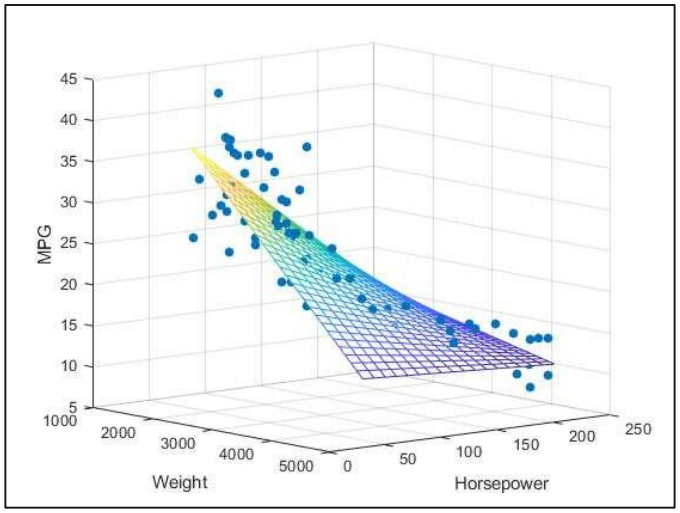
- Application area: Direct Marketing
- Goal: Reduce cost of a mailing campaign by targeting only the set of consumers that likely to buy a new product

- 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	Juice, Bread
3	Juice, Coke, Diaper, Milk
4	Juice, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Frequent Itemsets

- {Diaper, Milk, Juice}
- {Milk, Coke}

Association Rules

- {Diaper, Milk} -> {Juice}
- {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

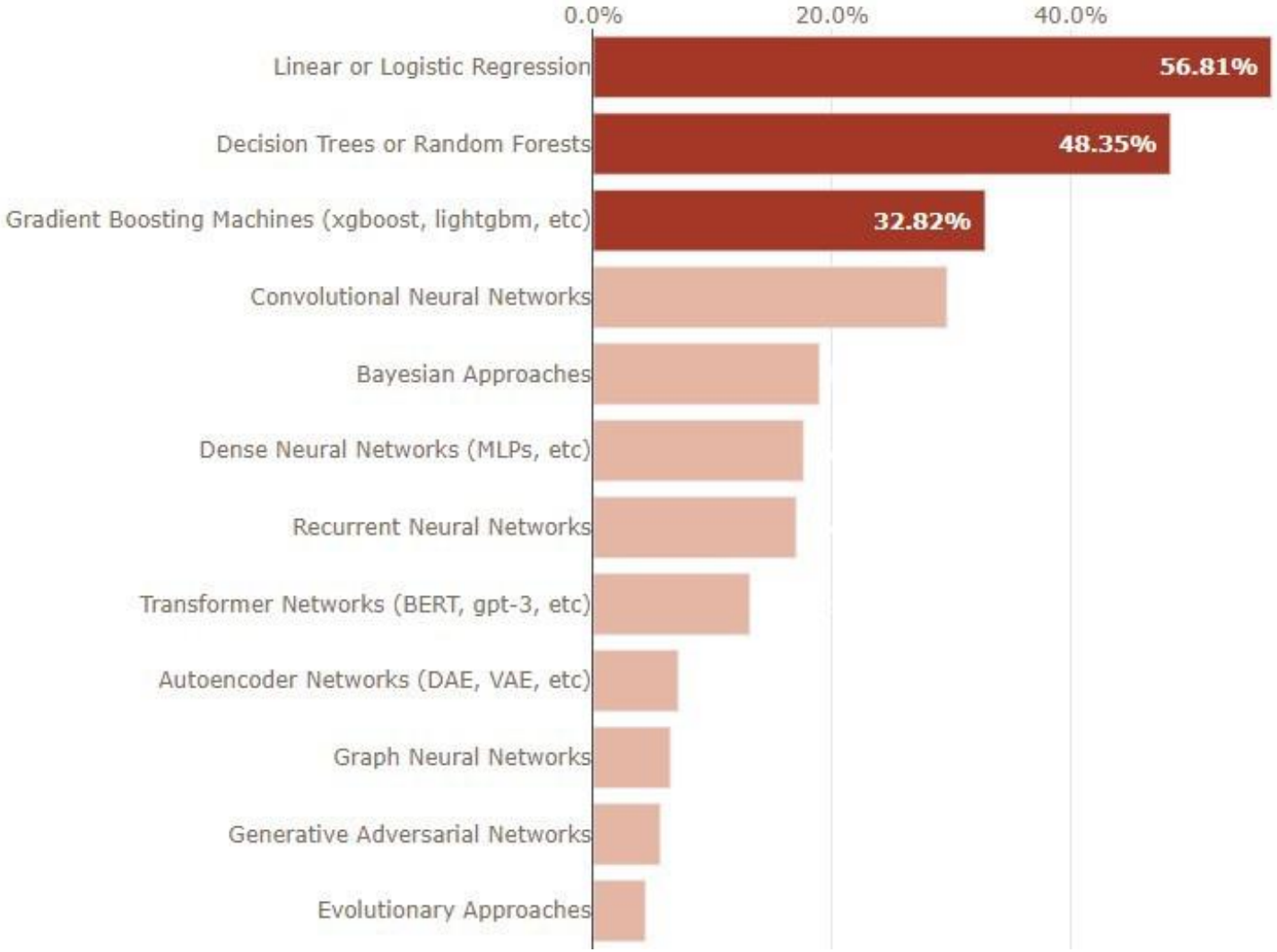
- Application area: Sales Promotion

Association Rule Discovery: Application 2

- Application area:
  Inventory Management

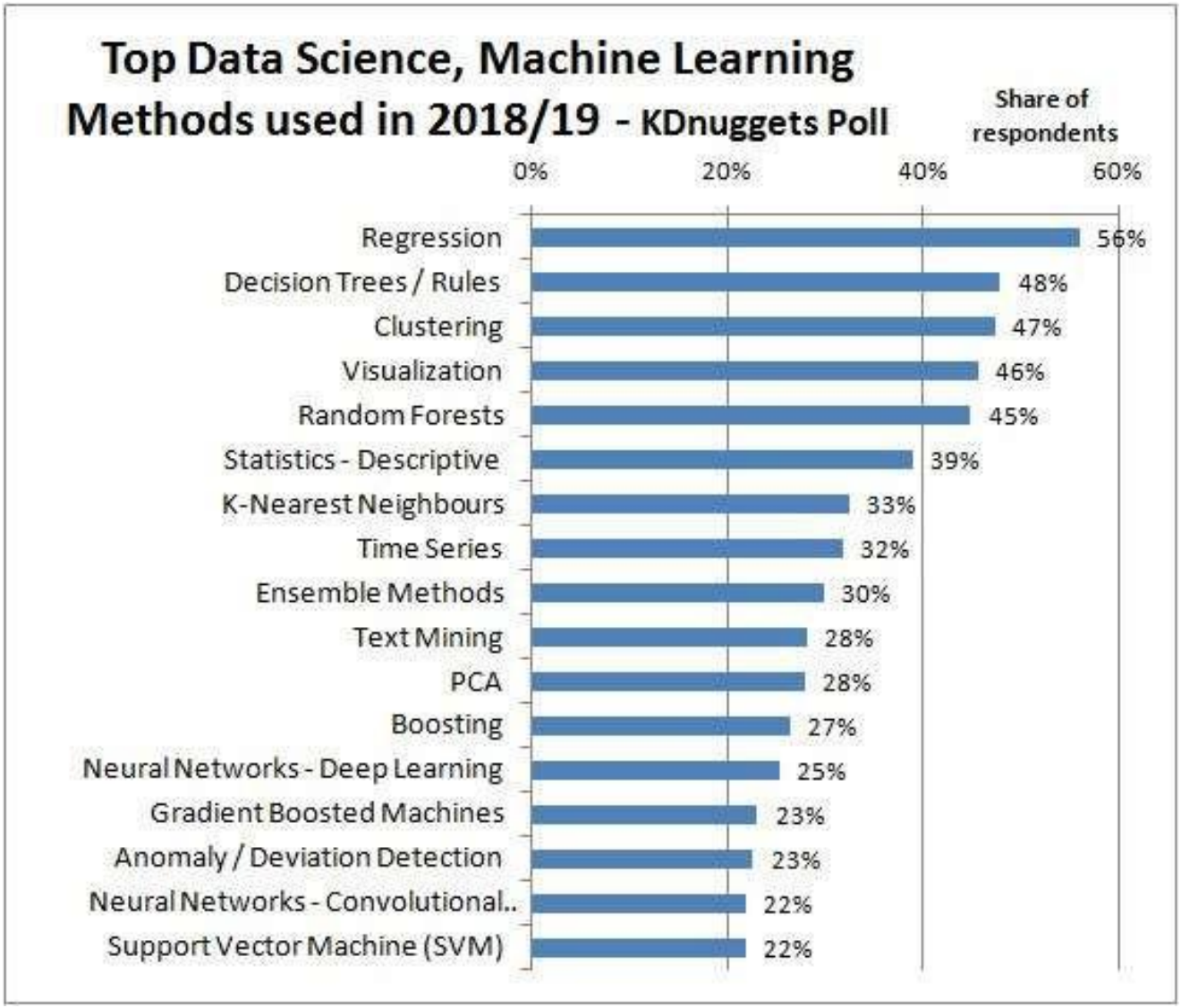
- Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households
- Approach: Process the data on tools and parts required in previous repairs at different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

Which Methods are Used in Practice?



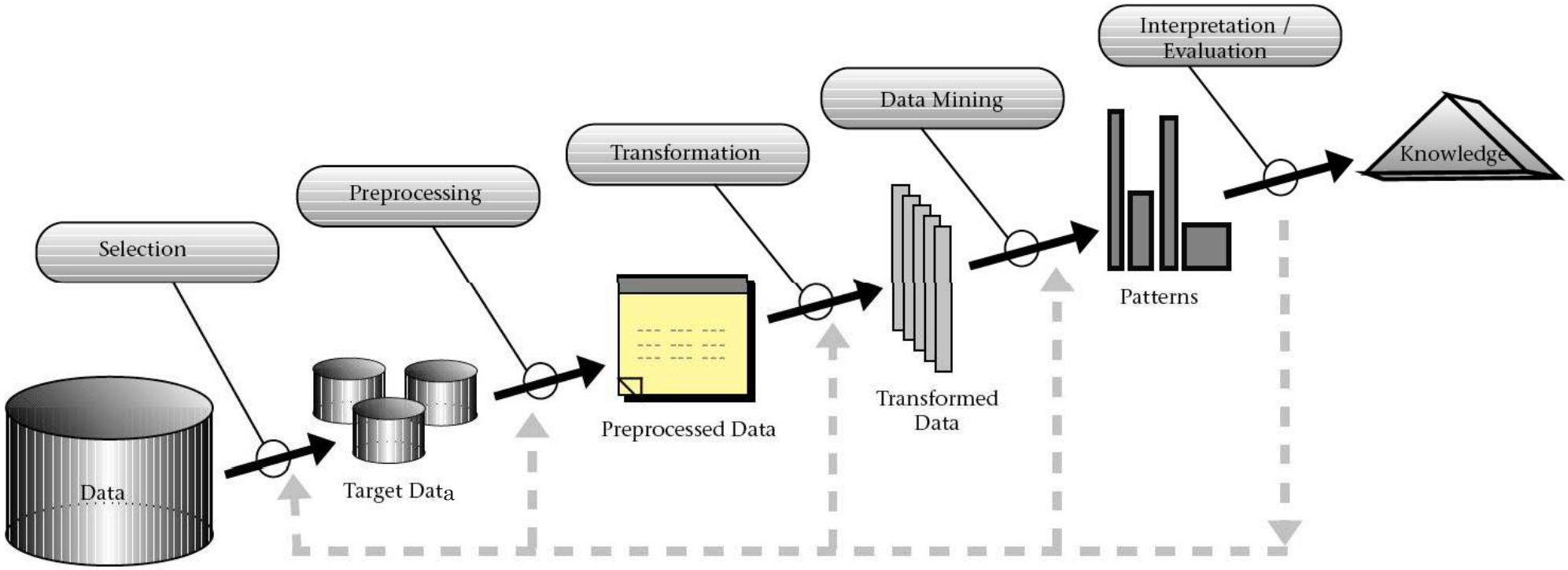
Source: Kaggle online poll 2022, 23,997 respondents,
https://www.kaggle.com/code/eraikako/data-science-and-mlops-landscape-in-industry

Which Methods are Used in Practice?



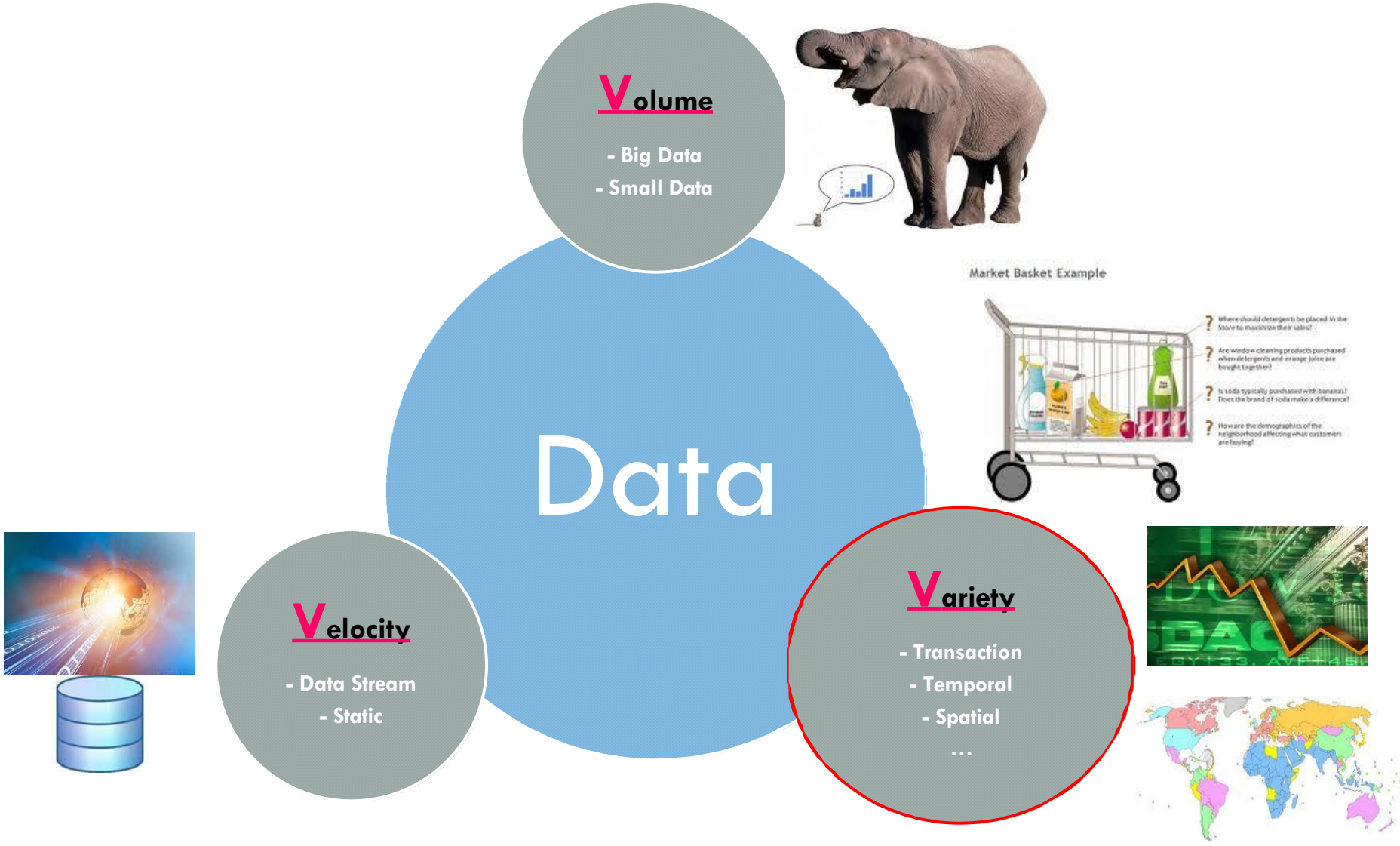
Source: KDnuggets online poll, 833 votes, question: methods used last year for real-world app?
https://www.kdnuggets.com/2019/04/top-data-science-machine-learning-methods-2018-2019.html

3. The Data Mining Process

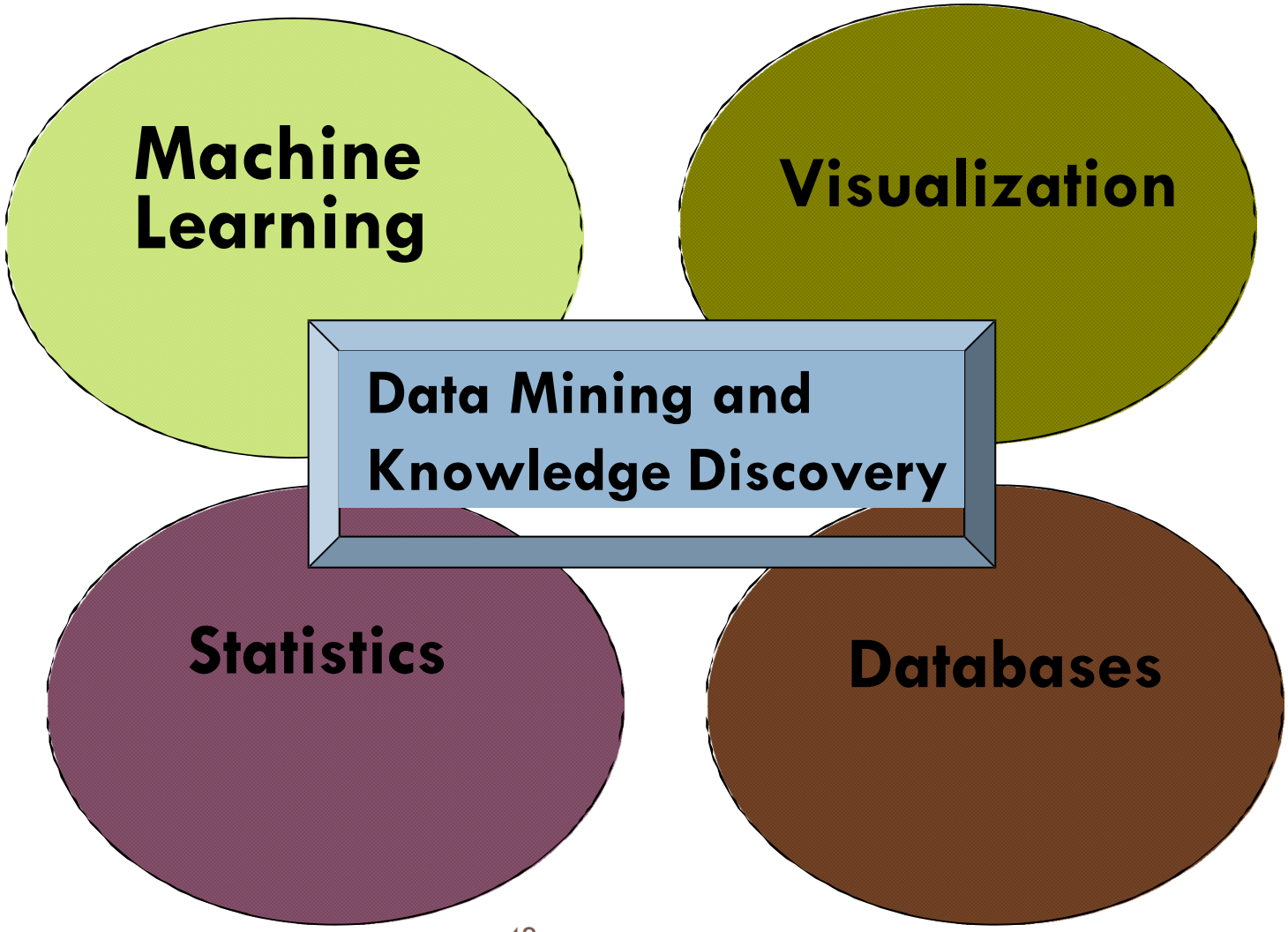


Source: Fayyad et al. (1996)

Knowledge Discovery in Data: Challenges

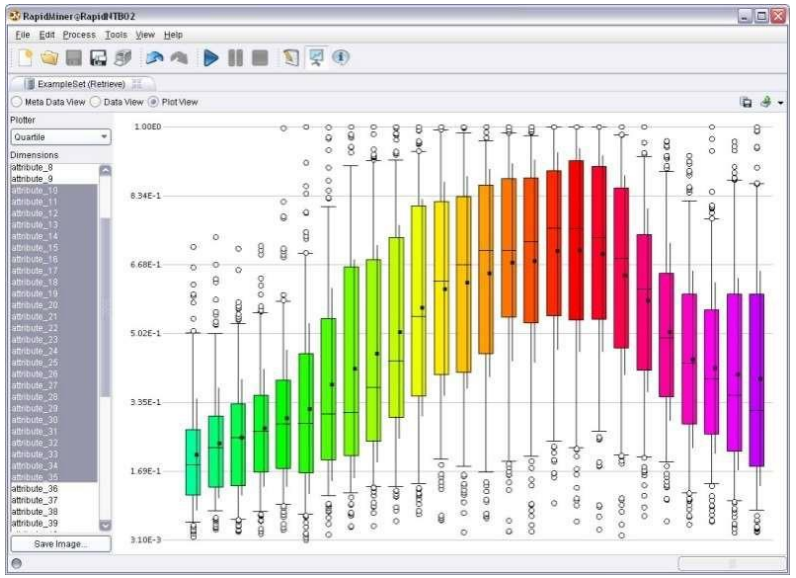


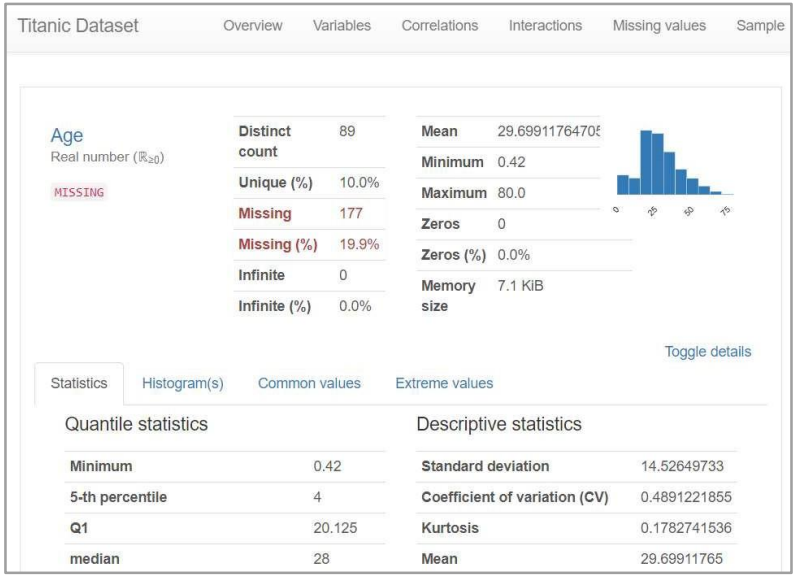
Related Field



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





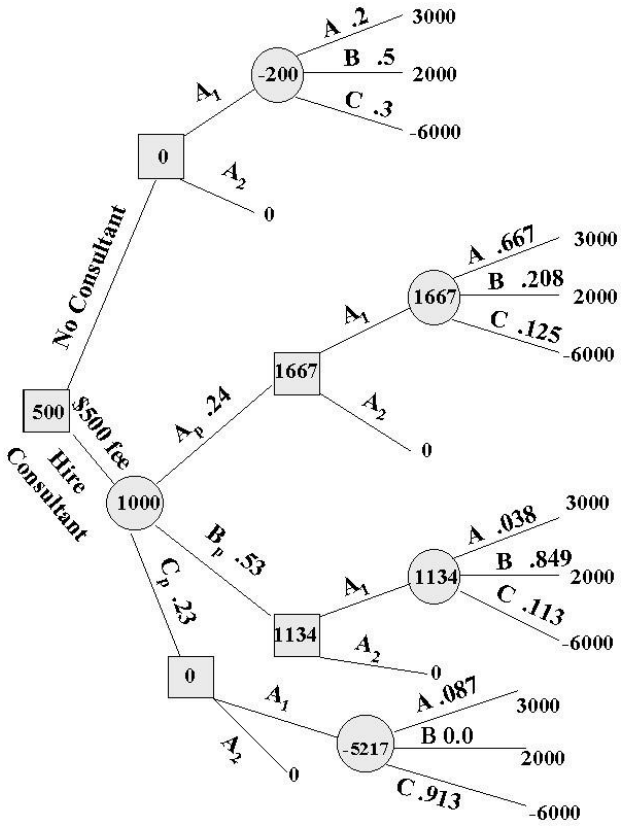
3.2 Preprocessing and Transformation

- 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

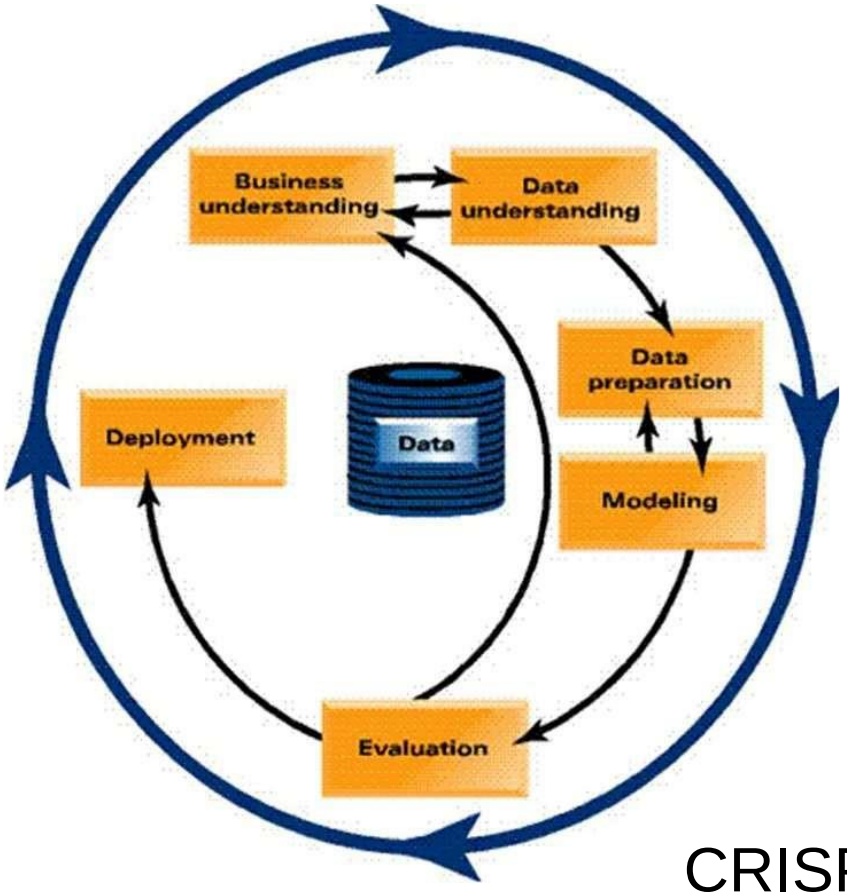
- Input: Preprocessed Data
- Output: Model / Patterns

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



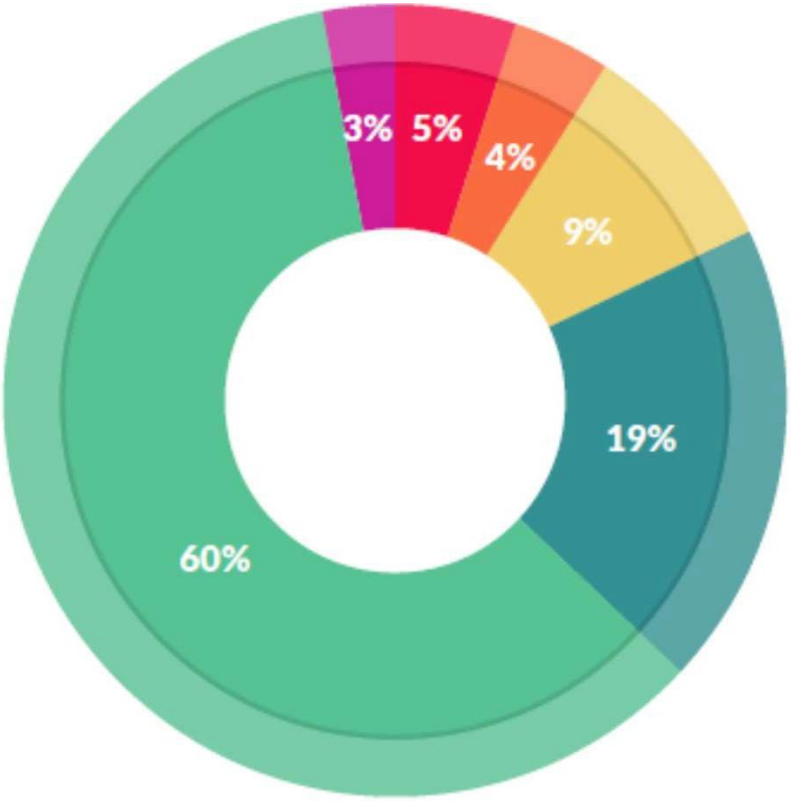
3.4 Deployment

- 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?



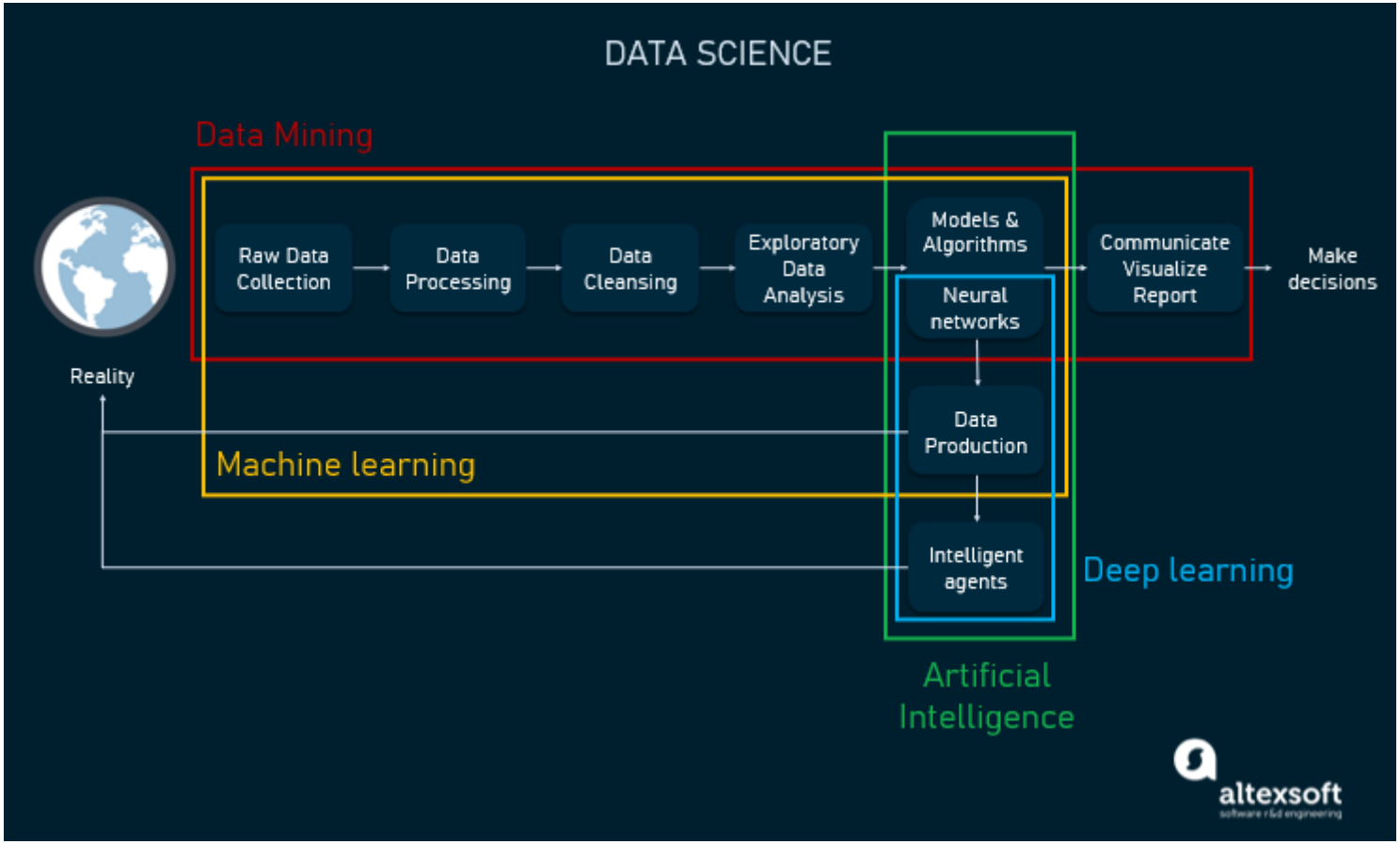
What data scientists spend the most time doing

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

Most frequent yet most overlooked ☺

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

Data Science vs Machine Learning vs AI vs Deep Learning vs Data Mining



https://www.altexsoft.com/blog/data-science-artificial-intelligence-machine-learning-deep-learning-data-mining/

Data Science vs Data Mining

DATA MINING VS DATA SCIENCE

	Data Mining	Data Science
01	More involved with its processes	Broadly focuses on the science of data
02	Primarily used for business purposes	It is essentially implemented for scientific purposes
03	Data mining is a technique that is a part of the KDD process	Data science is a field of study
04	Primarily deals with structured data	It deals with all types of data - structured, unstructured, or semi-structured
05	It is about extracting valuable information from data	It is about collecting, & processing, analyzing & utilizing data in various operations
06	It is a subset of data science as mining activities are in the pipeline of data science	Involves data scraping, cleaning, visualization, stats, etc. Therefore, it is a superset of data mining
07	Its objective is to realize the value of data & make it usable by extracting important Info.	Objective is to build data-dominant products for a venture

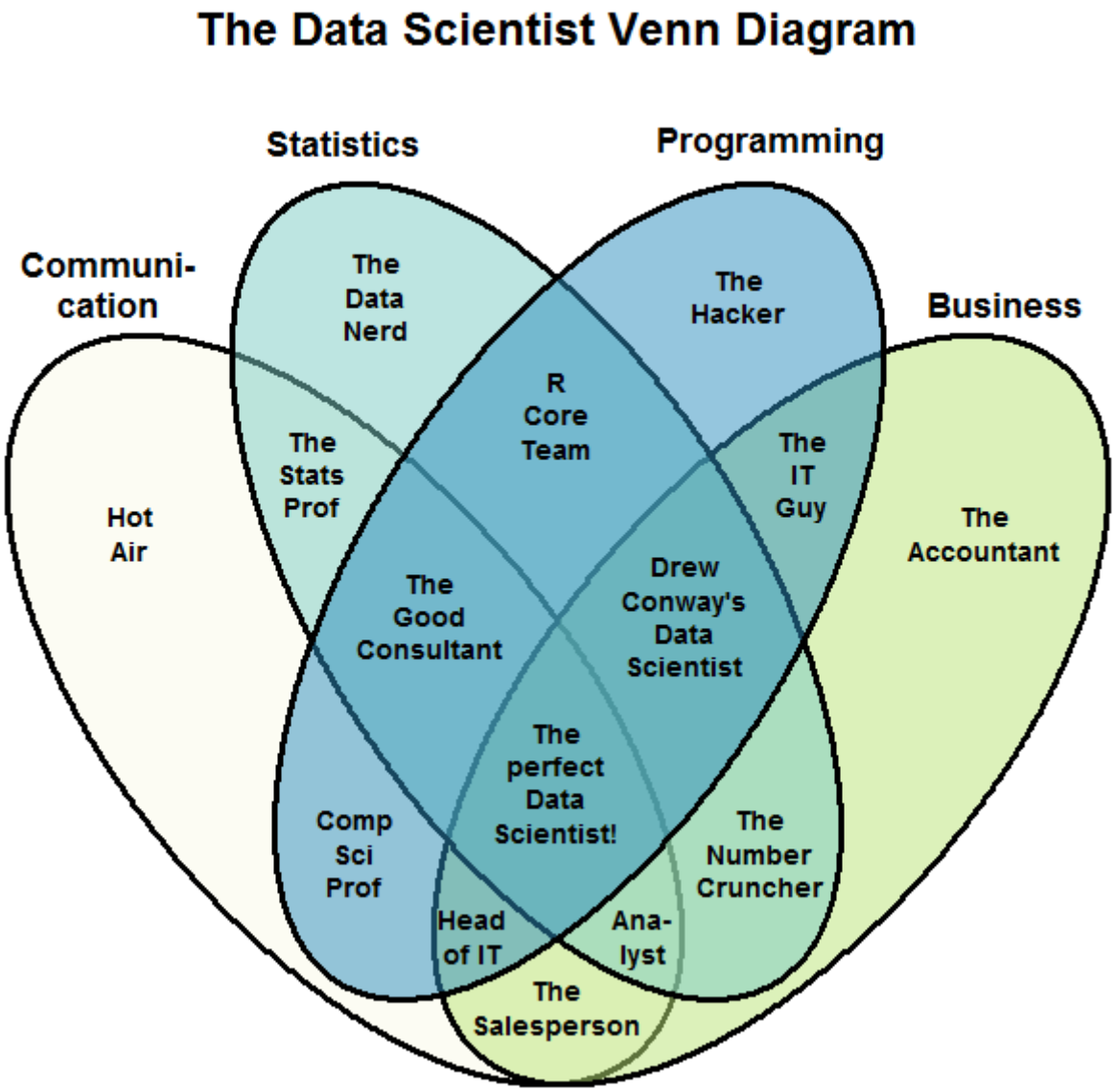
Data Scientist Venn Diagram

Key Insights:

•**Multidisciplinary Field:** Data science requires a blend of different skills, and the most effective data scientists are those who can integrate knowledge from multiple disciplines.

•**Role Specialization:** Different roles in data science may emphasize different skill sets, but the most valuable professionals often have overlapping knowledge areas.

•**Communication:** Effective data scientists must also have good communication skills to convey complex technical information to non-technical stakeholders.



Thank you!

– Are there any questions?

References:

https://www.uni-mannheim.de/dws/

https://www.cs.bu.edu/fac/gkollios/

https://www.iitr.ac.in/media/facspace/patelfec/16Bit/

https://www.kdnuggets.com/

https://hanj.cs.illinois.edu/bk3/bk3_slidesindex.htm