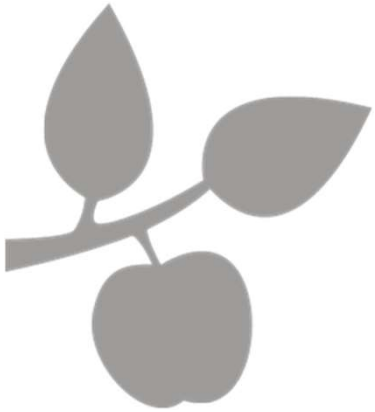


DM583

Data Mining

Introduction

DATA SCIENCE, BIG DATA, EDA, KDD

Contextualization and Motivation

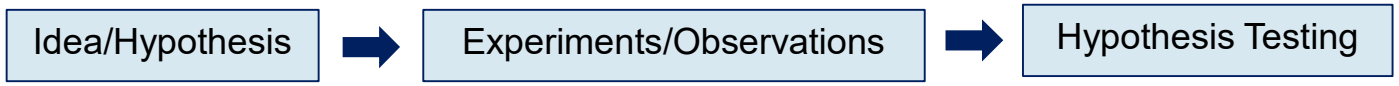
What is Data Science?

- "*Data science… is an interdisciplinary field … to extract knowledge or insights from data in various forms… It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science…*"

Wikipedia Entry on **Data Science**

(April 2017)

Origins of Data Science



- Historically, the **traditional setup** for statistical analysis is such that one has an idea, or hypothesis, and they attempt to validate this *hypothesis* by testing it against observations, which are possibly the outcomes of carefully designed *experiments*
- Clearly, such traditional setup is still very important in many application scenarios

Origins of Data Science

Adapted from:

DM868 DM870
DS804

(Arthur Zimek)

Knowledge Discovery from Data

Data Deluge

- Current Reality:
  Huge amounts of data collected in various forms and domains
- Manual analysis...?

SDU
SYDDANSK UNIVERSITET

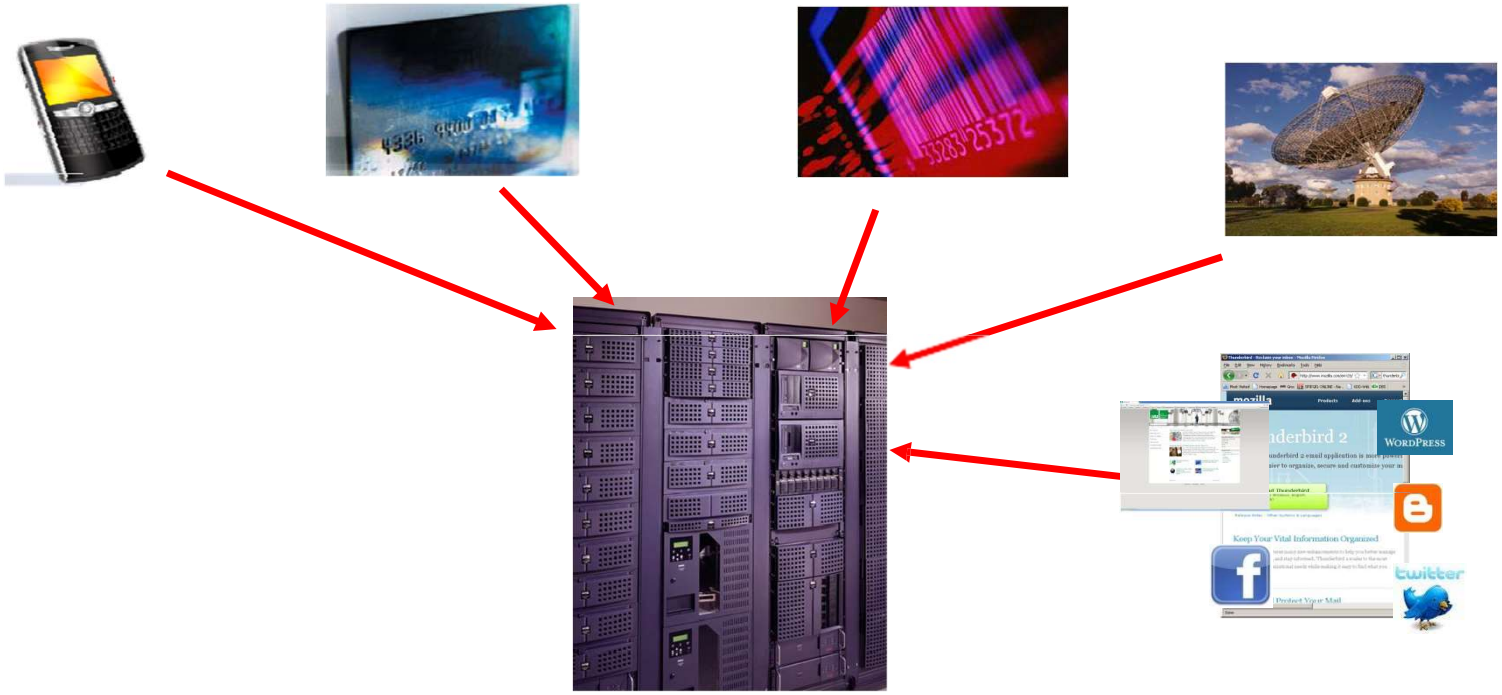
Data Deluge Requires Data Scientists

Adapted from:

DM868 DM870
DS804

(Arthur Zimek)

Knowledge Discovery from Data

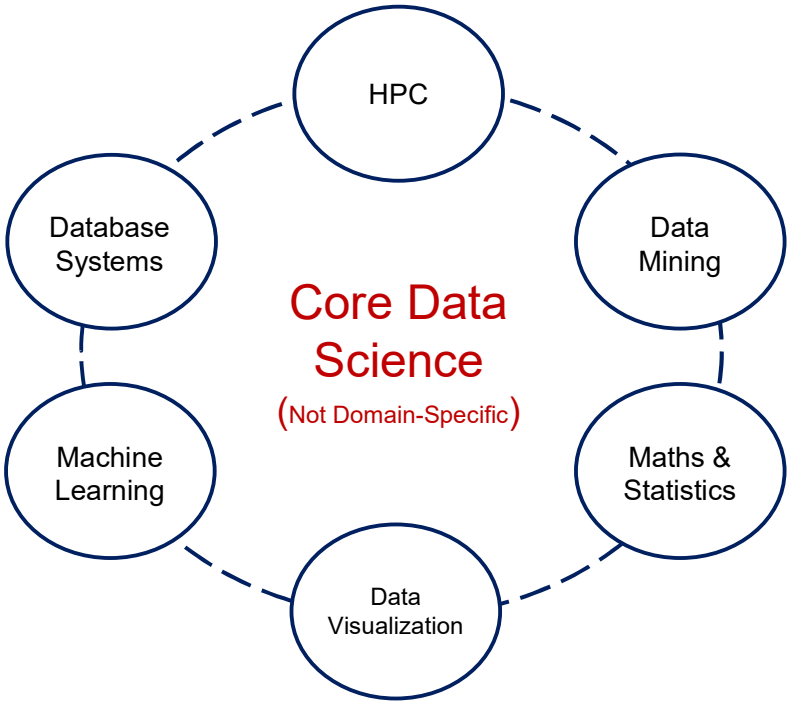


- **Data scientists** can play different key roles: design, prototyping, development, deployment, experiments, analysis, interpretation, communication (domain-experts, end-users, stakeholders), etc.

Data Scientist

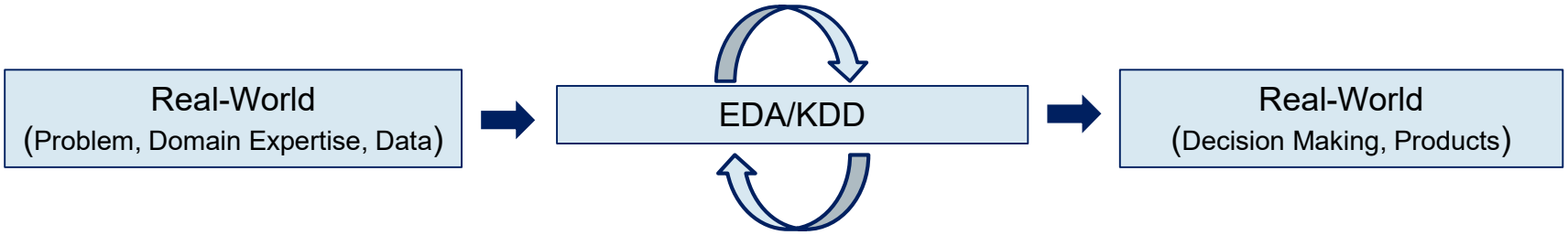
- Data Scientists have skills in a broad spectrum of areas, but very importantly, they shall be able to communicate, understand, interact, and collaborate with other experts in those areas
- Data Science is **interdisciplinary**
  - and real-world data analysis is rarely carried out by a single person

Core Data Science Pillars



Data Science, EDA and KDD

- The ultimate goal of data science is to make sense of data
- Data science is broader, but closely related to the concepts of **Knowledge Discovery from Databases** (**KDD**) and **Exploratory Data Analysis** (**EDA**)
  - These concepts precede the rise of Data Science as a discipline
  - They are focused on the central core of the data analysis process

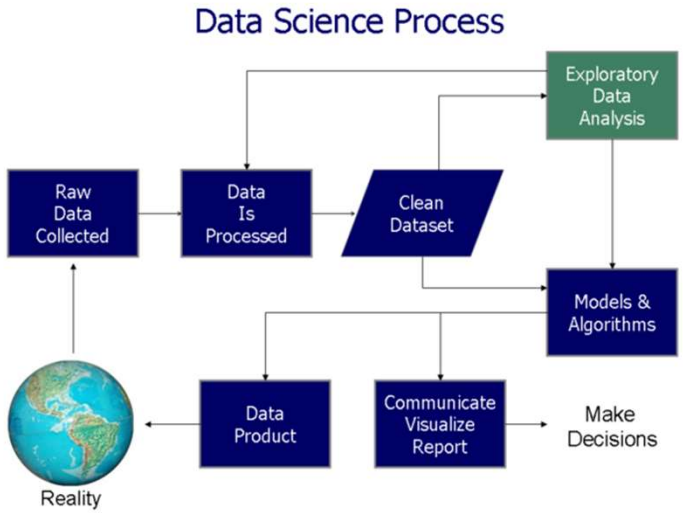


Data Science, EDA and KDD

- Data Science is broader in the sense that it involves the whole process from an initial problem or idea to the final solution (decision-making) or product
  - This can include pre- and post-interactions with domain experts, end-users & stakeholders
  - This can also require sophisticated data engineering technology, such as end-to-end pipelines with database integration, HPC/Cloud infrastructure, etc.

Data Science, EDA and KDD

- EDA and KDD focus on the inner part of this cycle, which involves certain aspects of data processing, interactive exploration, modelling and analysis:
  - How to extract useful insights, patterns, and ultimately knowledge from data



Data Science, EDA and KDD

- **Exploratory Data Analysis** (**EDA**):

  "*In statistics, **exploratory data analysis** (**EDA**) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not… (EDA) was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments…*"

Wikipedia Entry on **Exploratory Data Analysis** (April 2017)

Data Science, EDA and KDD

- **Knowledge Discovery from Databases (KDD):**
  - It is the process of identifying valid, novel, useful, and understandable patterns in data
  - It is a more recent concept related to EDA, yet broader in scope
    - Traditional EDA focuses mainly on visualization and descriptive statistics
    - KDD also encompass modern **data mining** and **machine learning** algorithms and models

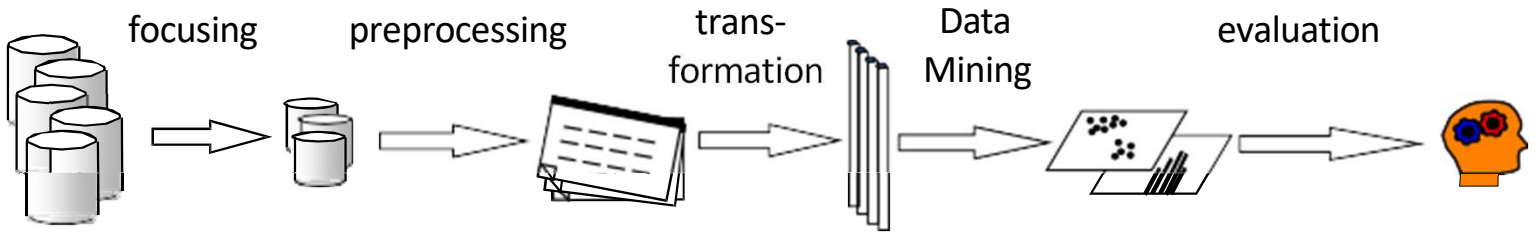
KDD (Original Definition)

"KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." [Fayyad et al., 1996]

- *data*: set of facts (e.g., entries in a database)
- *pattern*: expression in some language to describe a data subset (e.g., mathematical model)
- *process*: can involve several steps or iterations
- *nontrivial*: more complex than search, inference, simple aggregations
- *valid*: applicable to new data with a certain degree of reliability
- *novel*: for the system / user
- *potentially useful*: beneficial for user of application
- *ultimately understandable*: if not immediately then given some post processing

The Classic KDD Process Model

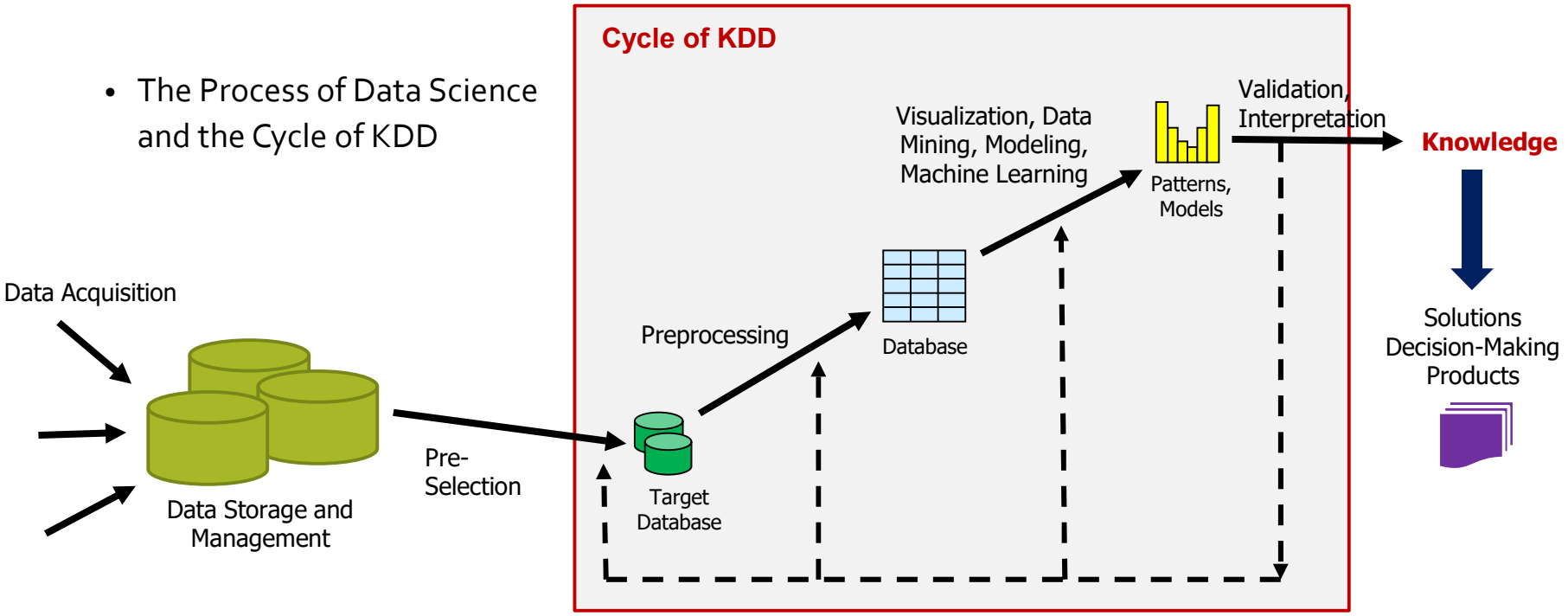
KDD process model (cf. Fayyad et al. [1996])



- *focusing*
  - get the data
  - organize data (file/database)
  - select relevant data
- *preprocessing*
  - integrate heterogeneous data
  - check for completeness
  - check for consistency
- *transformation*
  - convert / modify attributes
  - infer new attributes
  - select relevant attributes
- *data mining*
  - generate patterns or models
- *evaluation*
  - assess "interestingness" for the user
  - validate models statistically

Data Science, EDA and KDD

- The Process of Data Science and the Cycle of KDD



Big Data Era and Challenges

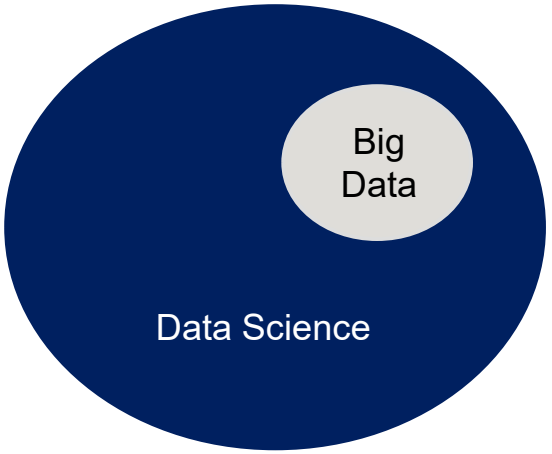
"We are drowning in information, but we are starved for knowledge"

John Naisbitt, "Megatrends", 1982

- Generally speaking, **Big Data** refers broadly to problems in data science that are complex enough to impose a level of difficulty to traditional methods or technologies that requires special approaches to deal with
  - obviously, a blurred definition that adjusts itself dynamically with context over time
- Typical Challenges:
  - **Volume** (too big or too small...), **velocity, variety, dimensionality, scarcity of labels**

Big Data Era and Challenges

- **Data Science is not only about Big Data…**
  - **…but Big Data Demands Data Scientists !**



DATA MINING

Contextualization and Motivation

Data Analysis in a Broad Sense

Illustration by David Parkins (detail).

Source: Silberzahn and Uhlmann [2015].

- Data Analysis: learning from data, finding patterns in data, understanding databases

Data Mining (DM)

Adapted from:

DM868 DM870
DS804

(Arthur Zimek)

Data Science

Illustration by David Parkins (detail).

Source: Silberzahn and Uhlmann [2015].

- Data Analysis: learning from data, finding patterns in data, understanding databases
- Data Mining: computational *methods* for data analysis

Data Mining (DM)

Adapted from:

DM868 DM870
DS804

(Arthur Zimek)

Data Science

Illustration by David Parkins

Source: Silberzahn and Uhlmann [2015].

- Data Analysis: learning from data, finding patterns in data, understanding databases
- Data Mining: computational *methods* for data analysis
- Different methods may deliver different pictures of the data...

Data Mining and Statistical Learning

- DM overlaps with the related areas of *statistical learning* and *machine learning*
- **Statistical Learning** (SL) refers broadly to the process of learning — *by using some sort of statistically sound technique* — a model that helps understand a data sample and possibly the underlying real-world phenomenon behind that sample
  - Statistical learning is the term most used by statisticians when referring to data mining
- However, there are certain DM techniques that do not fully fit the description of SL either because no model of the data is learnt, or because the process used for learning is more computational and/or heuristic than statistical in nature
  - One may argue, e.g., that certain association rule mining methods are not doing SL
    - Instead, they can be viewed as a computationally clever procedure to efficiently compute the frequency of many possible patterns to select the most interesting ones

Data Mining and Machine Learning

- **Machine Learning** (ML) is another area that highly overlaps with both DM and SL, but is closer to computing science than to statistics
- Traditionally, ML is seen as a process whereby computers automatically learn how to solve tasks from previous experience. A computer is expected to learn from examples, improving its performance on the given task
- Unlike SL, ML methods are *not necessarily* supported on statistical theory
- Unlike DM, ML methods are *not necessarily* intended for data analysis/KDD
  - E.g., self-driving cars, self-navigating robots, and other AI tasks
- Certain Tasks are both ML and DM tasks (e.g. regression, classification):
  - ML perspective generally focuses on prediction power/effectiveness only
  - DM perspective in these tasks can be different or broader…

SDU
SYDDANSK UNIVERSITET

Data Mining/ML: Patterns and Models

Adapted from:

DM868 DM870
DS804

(Arthur Zimek)

Data Mining Methods

- ***Predictive Methods***:
  A predictive model should be learnt from known data in a way suitable to make predictions on unknown, future data
- ***Descriptive Methods***:
  Provide insight into the data and help understand their major patterns, underlying structure(s), properties, etc.
- Sometimes a data mining method can be both, descriptive and predictive
- ***Prescriptive Tasks***:
  DM model and or predictions can be used for strategic decision making and action planning (typically related to optimization problems, e.g., in businesses)

Data Mining/ML: Training Strategies

- **Supervised Learning** (e.g., classification, regression, anomaly detection):
  - A dependent or output variable, Y, is modelled as a function of a collection of independent or input variable(s), X, so-called predictor(s)
  - For instance, credit risk as a function of customers' banking history and/or personal details
  - Past observations from both (X, Y) serve as a "teacher" for a model to be trained from data, and the training of such a model is referred to as supervised learning

Supervised Learning Task (Regression)

- **Regression**:
  - The dependent variable Y is real-valued (numeric, continuous)
  - It is presumed that it can be described as a function f of the predictor(s) X, i.e.

$$Y = f(X) + \varepsilon$$

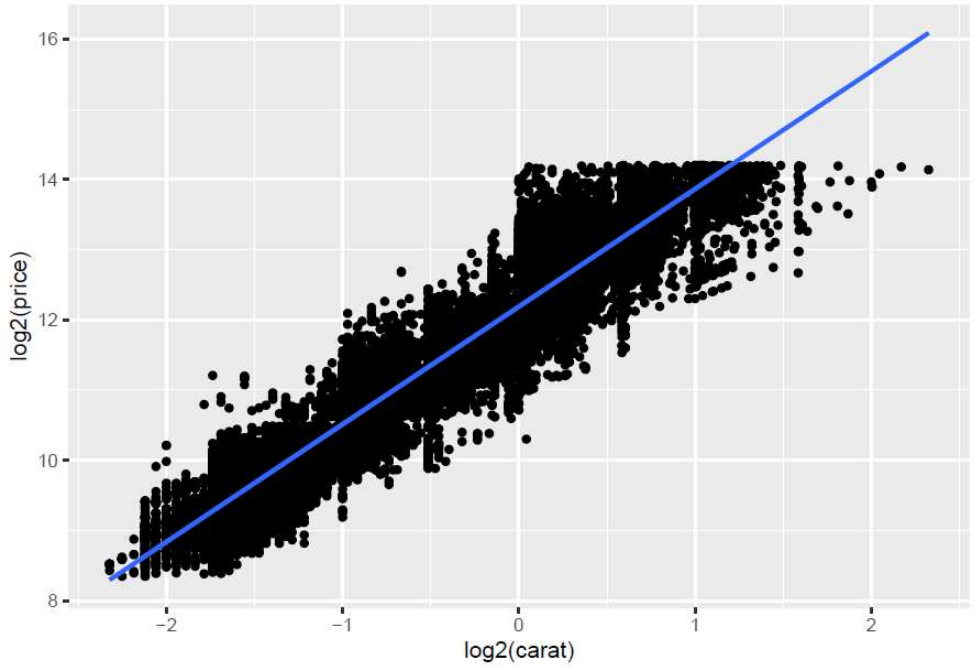
    Except for a component ε that depends on unobserved variables (obs. error, noise)

  - Main goal is to learn a model $\hat{f}$ for the unknown (possibly non-linear) mapping f from a data set containing a representative collection of observations (X, Y)

  - **ML Perspective**: typically, only or mostly focused on *prediction* (e.g. *black-box* models)
  - **DM (and SL) *Potentially* Extended Perspective**: e.g., which predictors most influence the output? how do such predictors influence the output? what do $\hat{f}$ tell us about the domain?

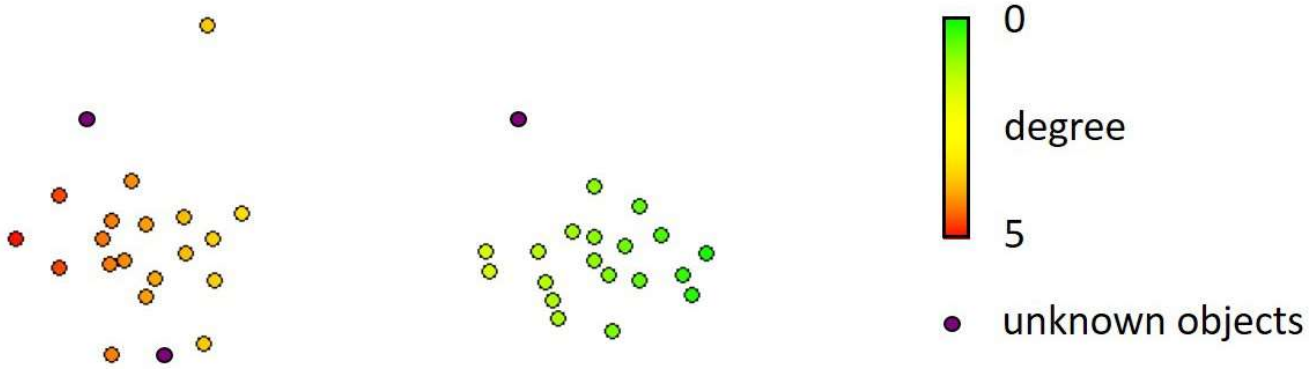
Supervised Learning Task (Regression)

- **(Simple Linear) Regression Example:** $Y = f(X) + \varepsilon \rightarrow \hat{Y} = \hat{f}(X) = b_0 + b_1 X_1$



Regression Example: `diamonds` dataset

Regression (Another Example)

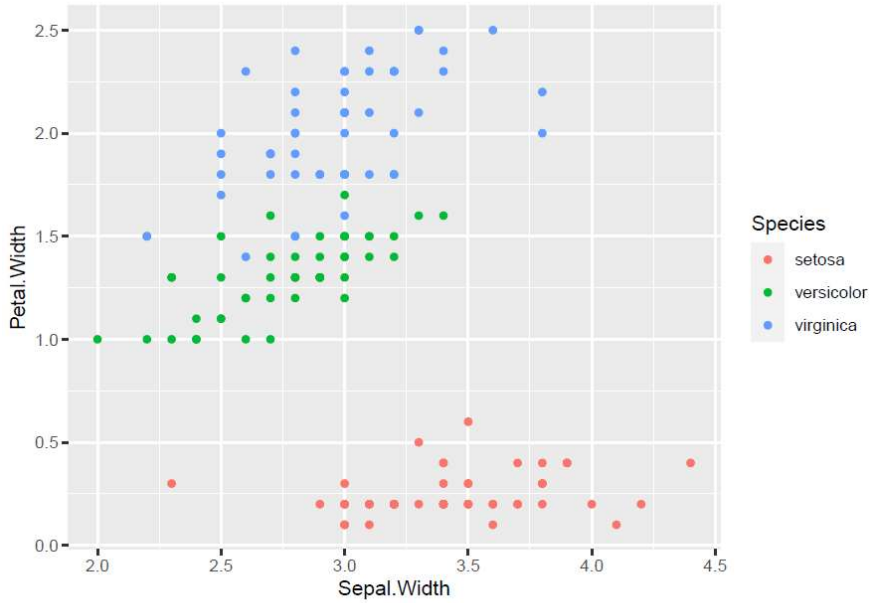


Supervised Learning Task (Classification)

- **Classification**:
  - The dependent variable *Y* is categorical
    - It takes on a finite set of nominal values (*class labels*)
  - It is also presumed that it can be described as a function *f* of the predictor(s) *X*, except for a component ε that depends on unobserved variables (observation error, noise)
  - Main goal is to learn a model $\hat{f}$ for the unknown mapping *f* from a data set containing a representative collection of observations (*X*, *Y*)
  - **ML Perspective**: typically, only or mostly focused on *prediction* (e.g. *black-box* models)
  - **DM (and SL) *Potentially* Extended Perspective**: e.g., which predictors most influence the output? how do such predictors influence the output? what do $\hat{f}$ tell us about the domain?

Supervised Learning Task (Classification)

- **Classification Example:** $Y = f(X) + \varepsilon \rightarrow \hat{Y} = \hat{f}(X)$

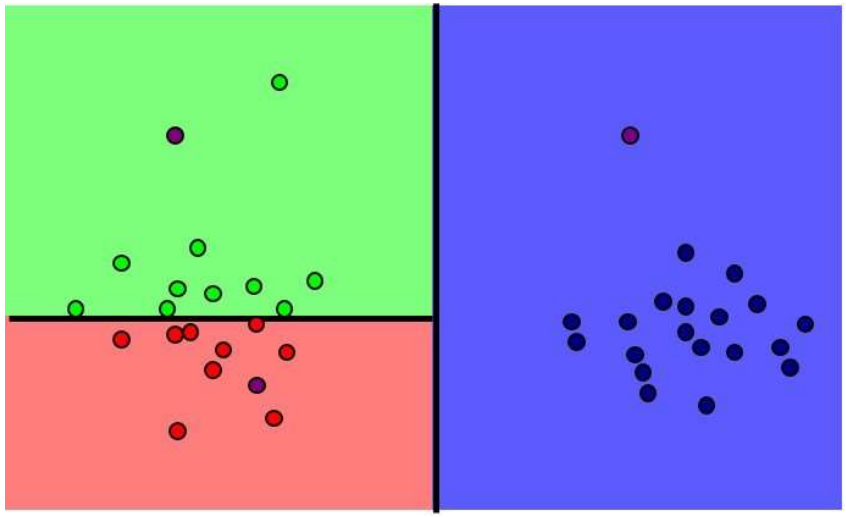


Classification Example: `iris` dataset

```
head(iris, 10)
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1           5.1         3.5          1.4         0.2  setosa
## 2           4.9         3.0          1.4         0.2  setosa
## 3           4.7         3.2          1.3         0.2  setosa
## 4           4.6         3.1          1.5         0.2  setosa
## 5           5.0         3.6          1.4         0.2  setosa
## 6           5.4         3.9          1.7         0.4  setosa
## 7           4.6         3.4          1.4         0.3  setosa
## 8           5.0         3.4          1.5         0.2  setosa
## 9           4.4         2.9          1.4         0.2  setosa
## 10          4.9         3.1          1.5         0.1  setosa
```

Classification (Another Example)



- screws
- nails
- staples

} training-data

- unknown objects

Data Mining/ML: Training Strategies

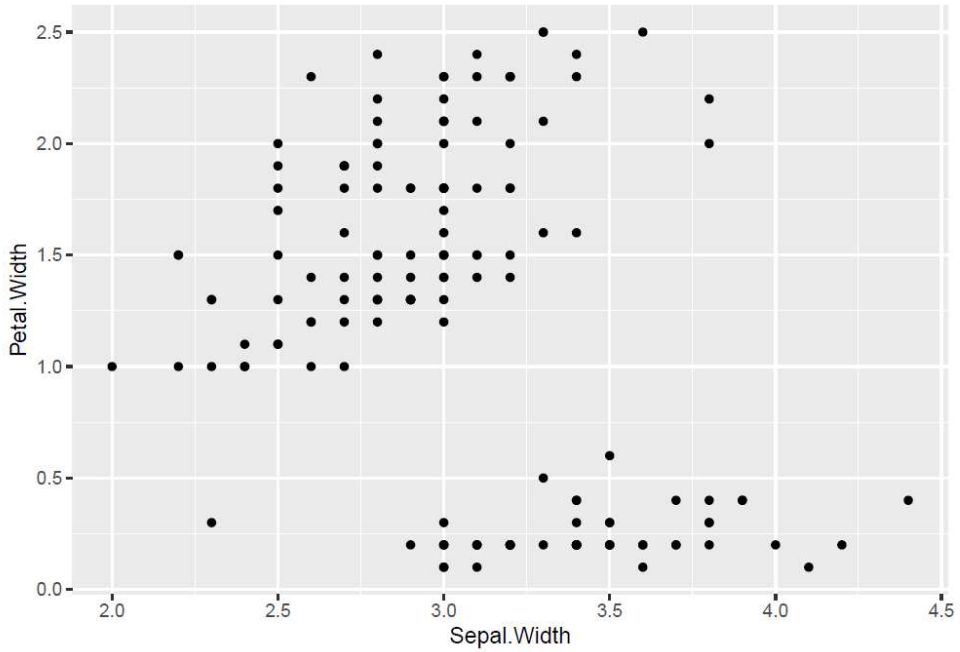
- **Unsupervised Learning** (e.g., clustering, outlier detection, association rules):
  - There is no dependent/output variable, only the input variables X
  - Patterns are learnt from X only
    - without explicit teacher/guidance on desired outputs
  - e.g., groups of similar customers based on their purchase behavior and personal details
- Emphasis on descriptive tasks

Unsupervised Learning Task (Clustering)

- **Clustering**
  - In clustering we are interested in finding groups (clusters) of observations that are somehow more "similar" or "related" to each other than observations in other groups
  - Cluster analysis is a well-established field of statistics at least since the 1940/50s
    - e.g. market segmentation problem
  - It is sometimes referred to as **unsupervised classification**

Unsupervised Learning Task (Clustering)

- **Clustering Example:** 2 major clusters

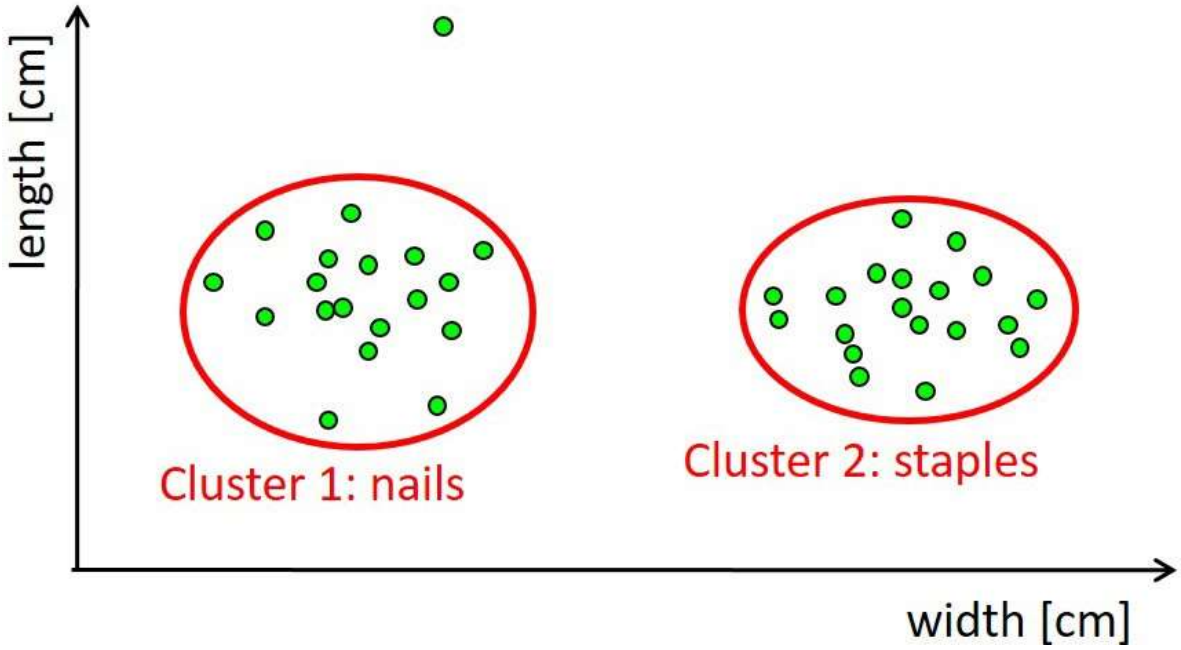


Clustering Example: `iris` dataset

```
head(iris, 10)
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1           5.1         3.5          1.4         0.2  setosa
## 2           4.9         3.0          1.4         0.2  setosa
## 3           4.7         3.2          1.3         0.2  setosa
## 4           4.6         3.1          1.5         0.2  setosa
## 5           5.0         3.6          1.4         0.2  setosa
## 6           5.4         3.9          1.7         0.4  setosa
## 7           4.6         3.4          1.4         0.3  setosa
## 8           5.0         3.4          1.5         0.2  setosa
## 9           4.4         2.9          1.4         0.2  setosa
## 10          4.9         3.1          1.5         0.1  setosa
```

Clustering (Another Example)

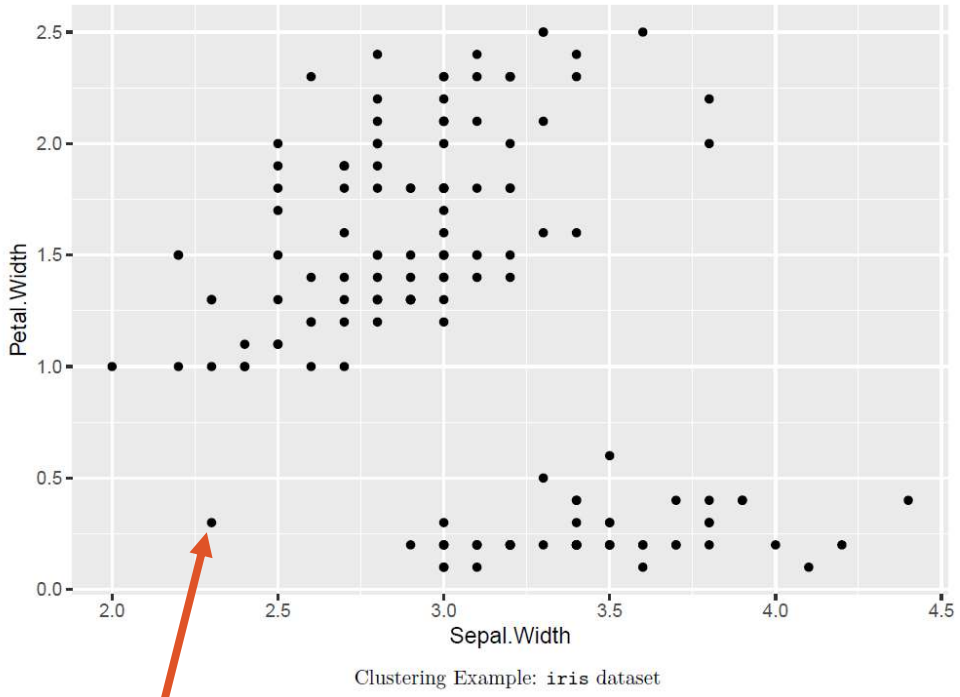


Unsupervised Learning Task (Outliers)

- **Outlier Detection Example:**

```
head(iris, 10)
```

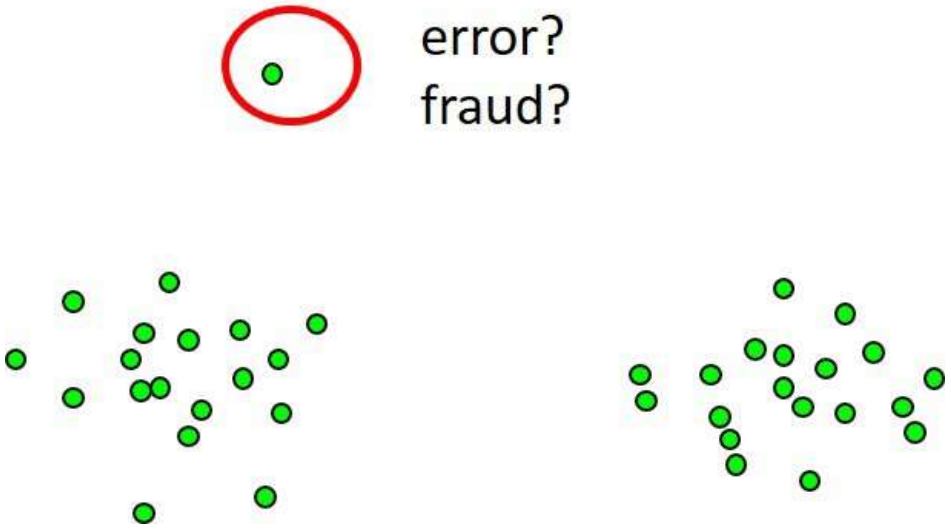
```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1           5.1         3.5          1.4         0.2  setosa
## 2           4.9         3.0          1.4         0.2  setosa
## 3           4.7         3.2          1.3         0.2  setosa
## 4           4.6         3.1          1.5         0.2  setosa
## 5           5.0         3.6          1.4         0.2  setosa
## 6           5.4         3.9          1.7         0.4  setosa
## 7           4.6         3.4          1.4         0.3  setosa
## 8           5.0         3.4          1.5         0.2  setosa
## 9           4.4         2.9          1.4         0.2  setosa
## 10          4.9         3.1          1.5         0.1  setosa
```



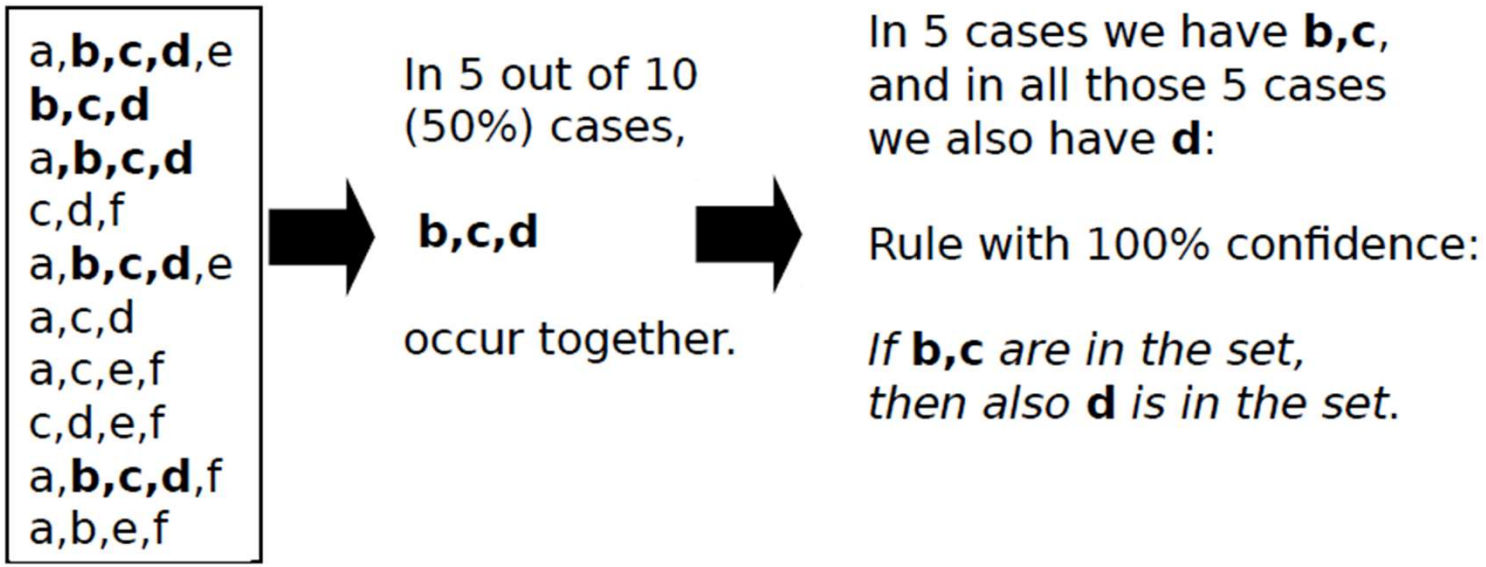
Clustering Example: `iris` dataset

abnormality?

Outlier Detection (Another Example)



Unsupervised Learning Task (Association Rules)

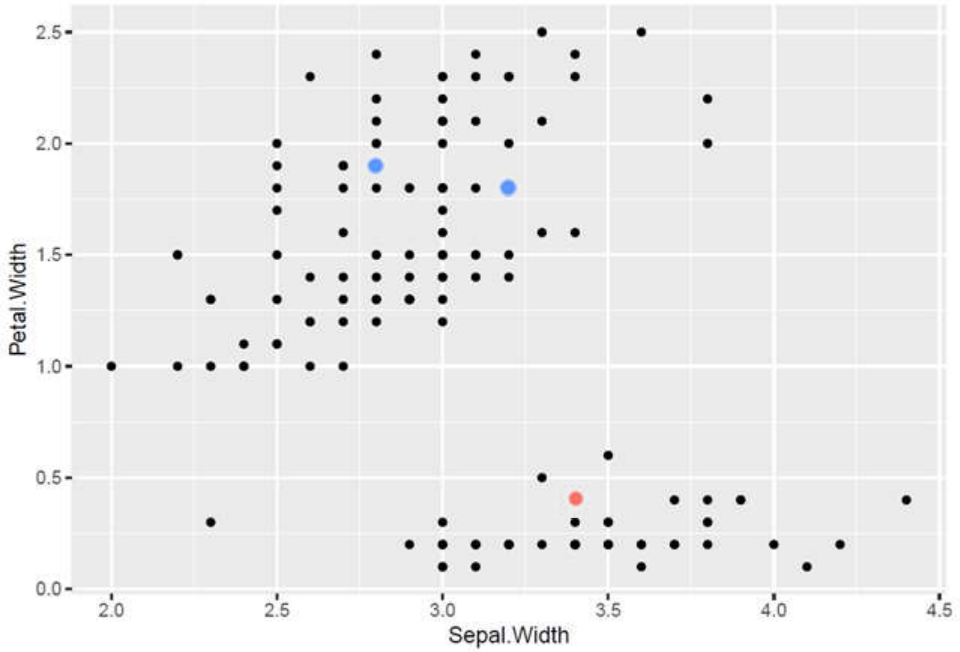


Data Mining/ML: Training Strategies

- **Semi-Supervised Learning** (e.g., classification, anomaly detection, clustering):
  - A response variable (or another form of external/supervised information for training) is only partially available
  - For instance:
    - Only class labels for a very small fraction of the data is available
      - not suitable for the conventional supervised training of a classifier
    - Only labels of one class (e.g., "normal" class), are available
      - again, not suitable for conventional supervised learning
    - Constraints limiting or discouraging certain clustering results

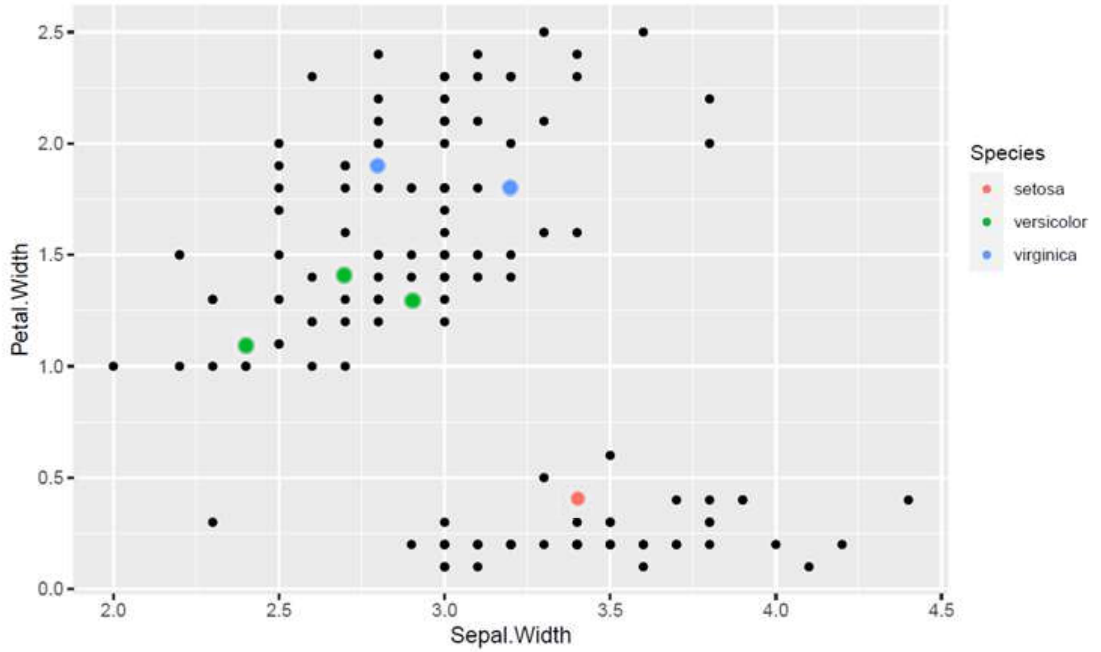
Semi-Supervised Learning Task

- **Semi-Supervised Clustering Example:**



Semi-Supervised Learning Task

- **Semi-Supervised Classification Example:**



Sample of Other, Specialized DM Tasks

- **Recommendation:**
- Goal is to predict whether (or to which extent) an individual will like a certain item, so that the item may possibly be recommended to that individual
- Targeted individuals are typically online users (online shoppers, social media followers, streaming video/music subscribers, etc.)
- Recommended items are typically products (e.g. books, movies, songs, videos, …) or other users (e.g. a match in a dating site or a new contact in a social network)

"*A recommender or recommendation system (sometimes replacing 'system' with a synonym such as platform or engine) is a subclass of information filtering system that seeks to predict the "rating" or "preference" that a user would give to an item.*[1][2]"

[Wikipedia Entry on Recommender Systems, April 2017]

Sample of Other, Specialized DM Tasks

- **Recommendation:**
- Recommendation can sometimes be seen as a very specialized type of *classification* (e.g. like/dislike class) or *regression* (e.g., item rating)
  - However, different from general purpose methods for classification and regression, strategies used in recommendation can be highly specialized
  - For example, a product can be recommended to a user for being similar to other products that the user has previously purchased, or for being highly rated by other users that are similar to the targeted user, or both
- Note that some recommendation strategies can make use of *frequent itemset and association rule mining* techniques, and/or *clustering*

Sample of Other, Specialized DM Tasks

- **Sentiment Analysis:** "*… (sometimes known as opinion mining or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information*" [Wikipedia Entry on Sentiment Analysis, April 2017]

- **Community Detection in Social Networks:** "*In the study of complex networks, a network is said to have community structure if the nodes of the network can be easily grouped into (potentially overlapping) sets of nodes such that each set of nodes is densely connected internally*" [Wikipedia Entry on Community structure, April 2017]

- **Link Analysis/Prediction:** "*In network theory, link analysis is a data-analysis technique used to evaluate relationships (connections) between nodes… Link analysis has been used for investigation of criminal activity (fraud detection, counterterrorism, and intelligence), computer security analysis, search engine optimization, market research, medical research, and art*" [Wikipedia Entry on Link Analysis, April 2017]

References

1. B. S. Baumer, D. T. Kaplan & N. J. Horton. "*Modern Data Science with R*", CRC Press, 2017
2. W. S. Cleveland. "*Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics*", Intl. Statistical Review, V. 69, pp. 21-26, 2001
3. T. H. Davenport & D.J. Patil. "*Data Scientist: The Sexiest Job of the 21st Century*", Harvard Business Review, October 2012
4. Dean, Jeffrey, and Sanjay Ghemawat. "*MapReduce: Simplified Data Processing on Large Clusters*", Communications of the ACM, V. 51, pp. 107-113, 2008.

References

5. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. "*Knowledge Discovery and Data Mining: Towards a Unifying Framework*", In Proc. of the 2nd ACM International Conference on Knowledge Discovery and Data Mining (KDD), Portland, OR, pages 82–88, 1996
6. J. Han, M. Kamber, and J. Pei. "*Data Mining: Concepts and Techniques*", Morgan Kaufmann, 3rd edition, 2011
7. P.-N. Tan, M. Steinbach, and V. Kumar. "*Introduction to Data Mining*", Addison Wesley, 1st Edition, 2006
8. M. J. Zaki and W. Meira JR. "*Data Mining and Analysis: Fundamental Concepts and Algorithms*", Cambridge Univ. Press, 1st Edition, 2014
9. R. Silberzahn and E. L. Uhlmann. "*Many Hands Make Tight Work*", Nature, 526:189–191, 2015. doi: 10.1038/526189a