Advanced Topics in Data Analytics

-

The Data Analytics Process

Andreas Rauber

Department of Software Technology and
Interactive Systems

Vienna University of Technology
rauber@ifs.tuwien.ac.at
http://www.ifs.tuwien.ac.at/~andi

Outline

- Data Analytics Process
- Self Organizing Map
- Ethics, Privacy, Reproducibility, Explainability

Outline

- Data Analytics Process
  - How to do Data Mining?
  - Types of machine learning
  - Attribute types
  - Data preprocessing: coding, scaling
  - Summary

What is Big Data?

Data as “the new oil”…

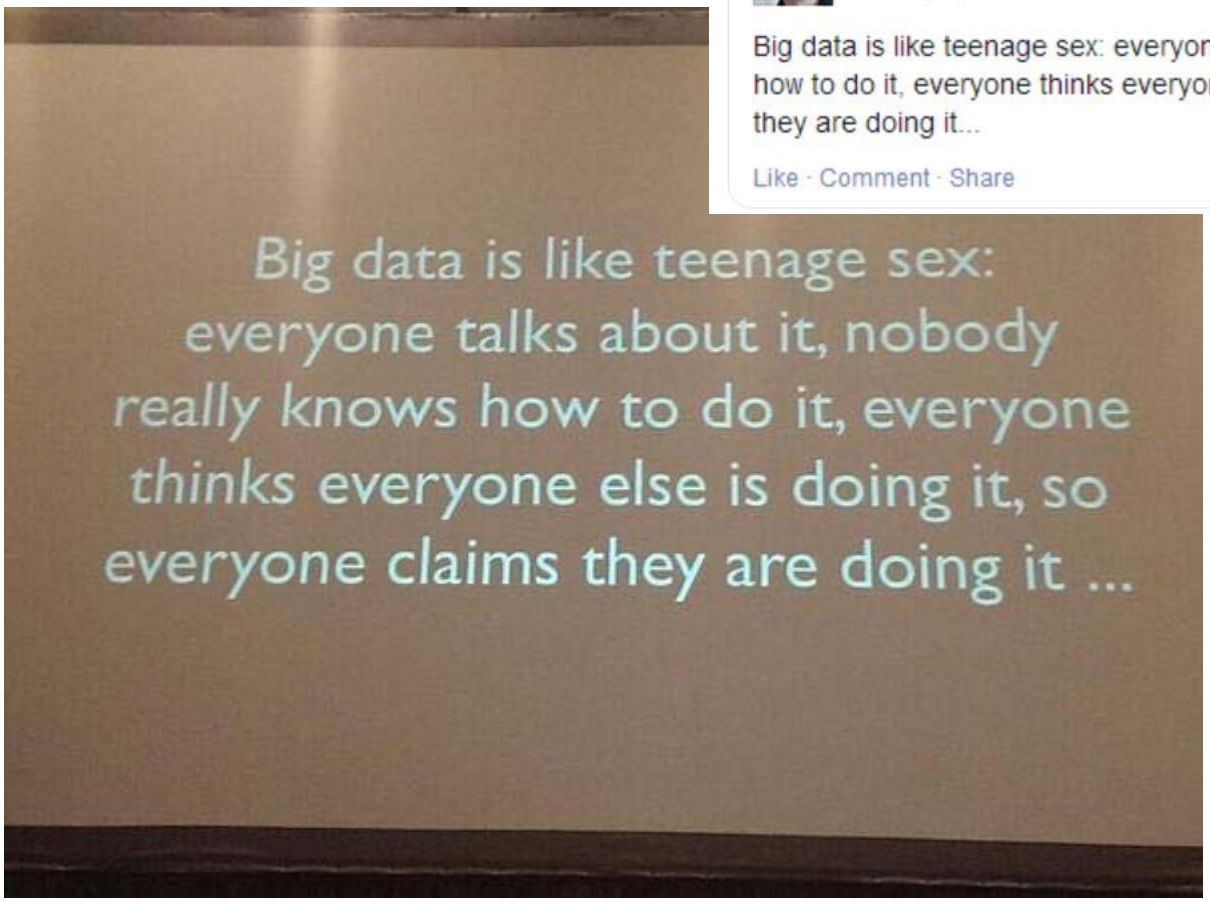
… or “the new water”…

… or “the new light”….

The Economist, May 6 2017
https://www.economist.com/printedition/2017-05-06

What is “Big Data”?

- What is “Big Data”?



2013

Dan Ariely, Professor of Psychology & Behavioral Economics at Duke University

Data Science – The Sexiest Job

Harvard Business Review

ANALYTICS

Data Scientist: The Sexiest Job of the 21st Century

Data Science – The Sexiest Job

Data Driven Business

- Hype or not hype…
- Data is at the core of almost every domain today
- Value of data (including historical) is increasingly being recognized
  - Data warehousing, business analytics, ...
- Basis: being able to integrate, use and re-use data

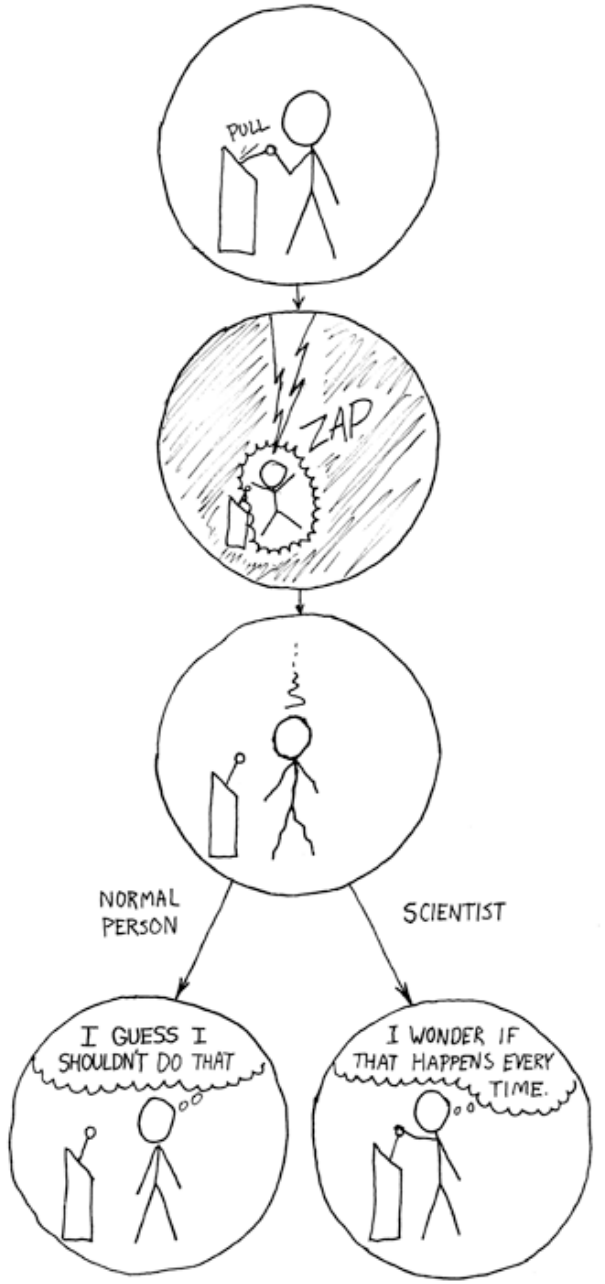
How to do Data Mining?

- Data Mining / Data Analysis / Knowledge Discovery
- How do we do it?
  - Download/extract the data
  - Send it through some machine learning tool
  - Summarize the results
- Done?
- Art vs. Science (but not magic!!)
- A more formal process
  - Fayyad's KDD process
  - CRISP-DM
  - ASUM-DM

Reproducibility

- Reproducibility is core to the scientific method
  (Data *Science* vs. Alchemy vs. Art)
- Focus not on misconduct –
  but on complexity and the will to produce good work

- Should be easy
  - Get the code, compile, run, …
  - Why is it difficult?



https://xkcd.com/242/

Challenges in Reproducibility

- Carmen M. Reinhart and Kenneth S. Rogoff: *Growth in a Time of Debt.* American Economic Review: Papers and proceedings 100:573-578, May 2010.
- Study on relationship btw. debt and economic growth
  - Tipping point at 90% of government debt
  - Published after the Greek crisis
  - Analysis supporting budget cuts
  - Stimulus vs austerity
  - Strong political influence

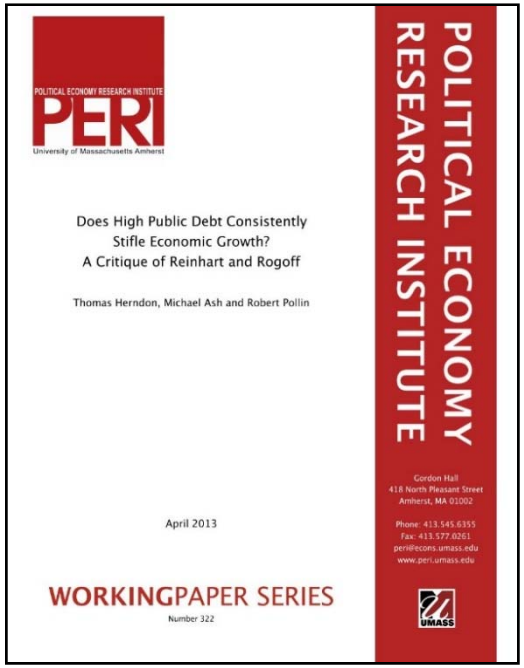
Growth in a Time of Debt

By Carmen M. Reinhart and Kenneth S. Rogoff*

https://scholar.harvard.edu/files/rogoff/files/growth_in_time_debt_aer.pdf

Reproducibility and politics

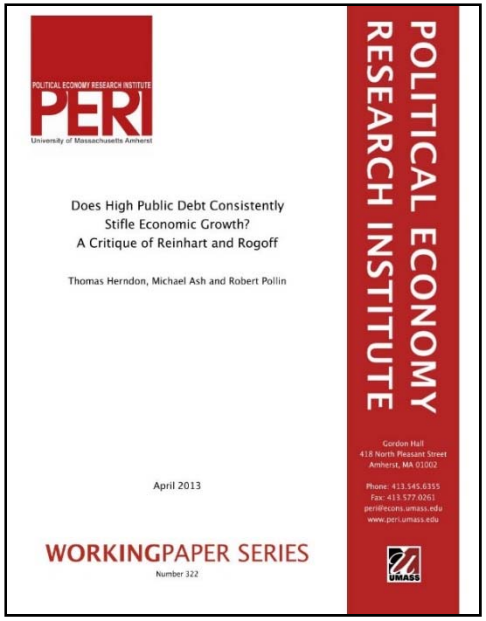
- Carmen M. Reinhart and Kenneth S. Rogoff: *Growth in a Time of Debt.* American Economic Review: Papers and proceedings 100:573-578, May **2010**.
- Others could not reproduce the result: Thomas Herndon, Michael Ash, Robert Pollin:
  Does High Public Debt Consistently Stifle Economic Growth?
  A Critique of Reinhart and Rogoff
  UMASS Working Paper Series 322, April **2013**



https://www.peri.umass.edu/fileadmin/pdf/working_papers/working_papers_301-350/WP322.pdf

Reproducibility and politics

- Carmen M. Reinhart and Kenneth S. Rogoff (2010) *vs.* Thomas Herndon, Michael Ash, Robert Pollin (2013)
- Original spreadsheet investigated
  - Some data excluded on purpose
  - Questionable statistical procedures
  - **Excel error**
    - Accidentally missed 5 rows of data!
    - Average Annual Growth changed from -0.1 to 2.2 after correction
- Lead to prominent coverage on importance of transparency, reproducibility



https://www.newyorker.com/news/john-cassidy/the-reinhart-and-rogoff-controversy-a-summing-up
https://www.nytimes.com/2013/04/19/opinion/krugman-the-excel-depression.html

Reproducibility & Verifiability

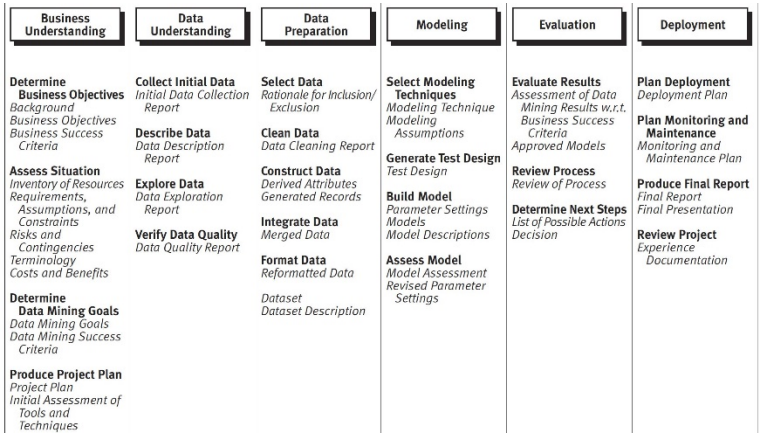
- Currently, data “science” (CS?) often resembles alchemy… (or wizardry?)

Pieter Bruegel the Elder: De Alchemist (Source: British Museum, London)
http://www.britishmuseum.org/research/collection_online/collection_object_details/collection_image_gallery.aspx?assetId=62085001&objectId=1335345&partId=1

Reproducibility & Verifiability

- From Alchemy to chemistry – from wizard to data analyst:
  - Structured processes
  - Documentation
  - Traceability, reproducibility
- To ensure trust, efficiency, correctness
- A more formal process to remove *some* of the “arts” aspects

Pieter Bruegel the Elder: De Alchemist
(Source: British Museum, London)



CRISP-DM Process Model

BI Process Models

Data analytics process models

- Fayyad's KDD process
- SEMMA
- CRISP-DM
- ASUM-DM

- Reference models
  - Decide and adapt process to organizational needs!
  - Balance structure – flexibility!

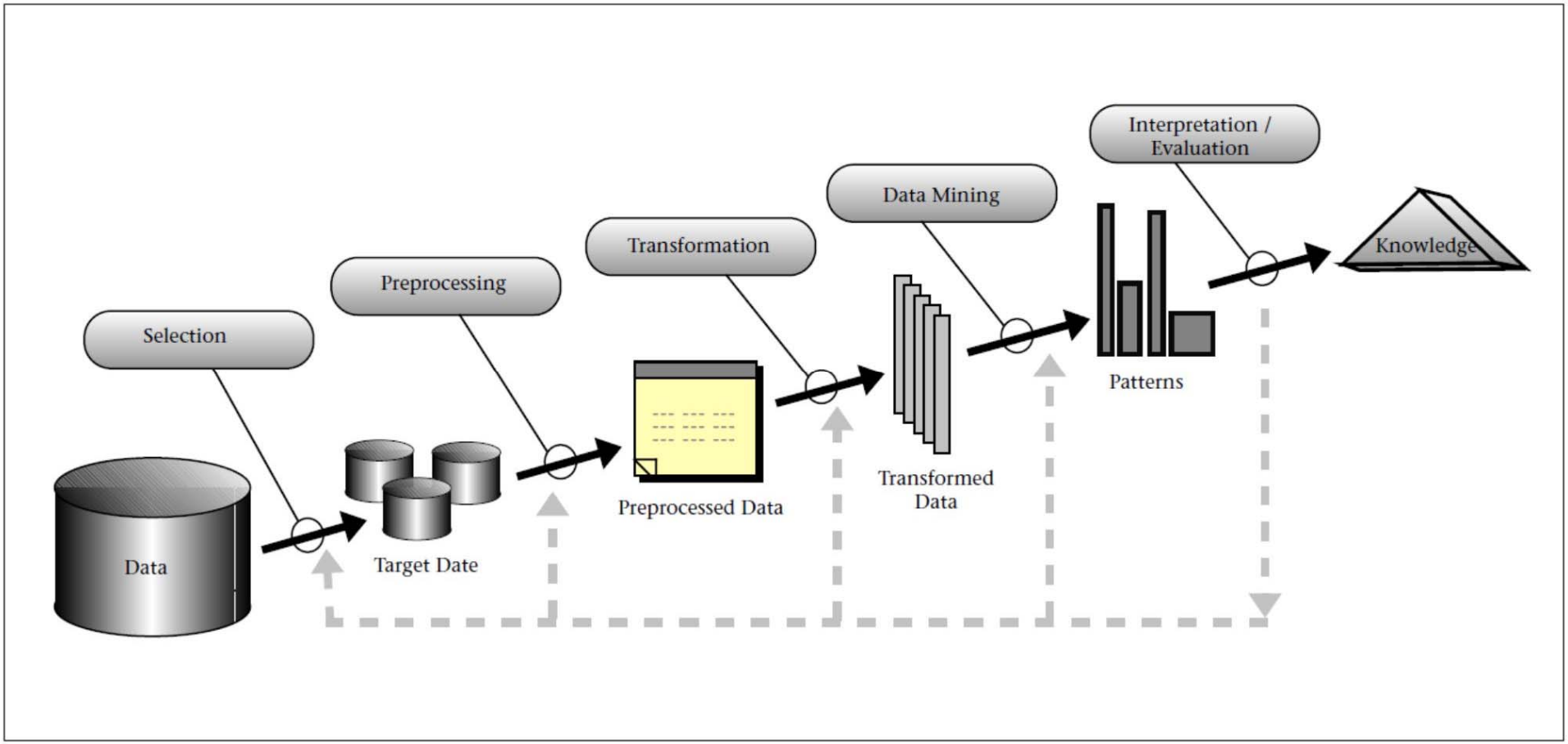
How to do Data Mining

Fayyad's KDD Process

- Usama Fayyad, Gregory Platetsky-Shapiro, Padhraic Smyth: From Data Mining to Knowledge discovery in Databases. AI Magazine 17(3):37-54, 1996
- "…mapping low-level data into other forms that might be more abstract or more useful."
- "Data Warehousing helps set the stage for KDD in two important ways: Data cleaning and Data Access"
- "Data Mining is a step in the KDD process that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data."

How to do Data Mining

Fayyad's 5-step KDD process



BI Process Models

Data analytics process models

- Fayyad's KDD process
- SEMMA
- CRISP-DM
- ASUM-DM

- Reference models
  - Decide and adapt process to organizational needs!
  - Balance structure – flexibility!

BI Process Models

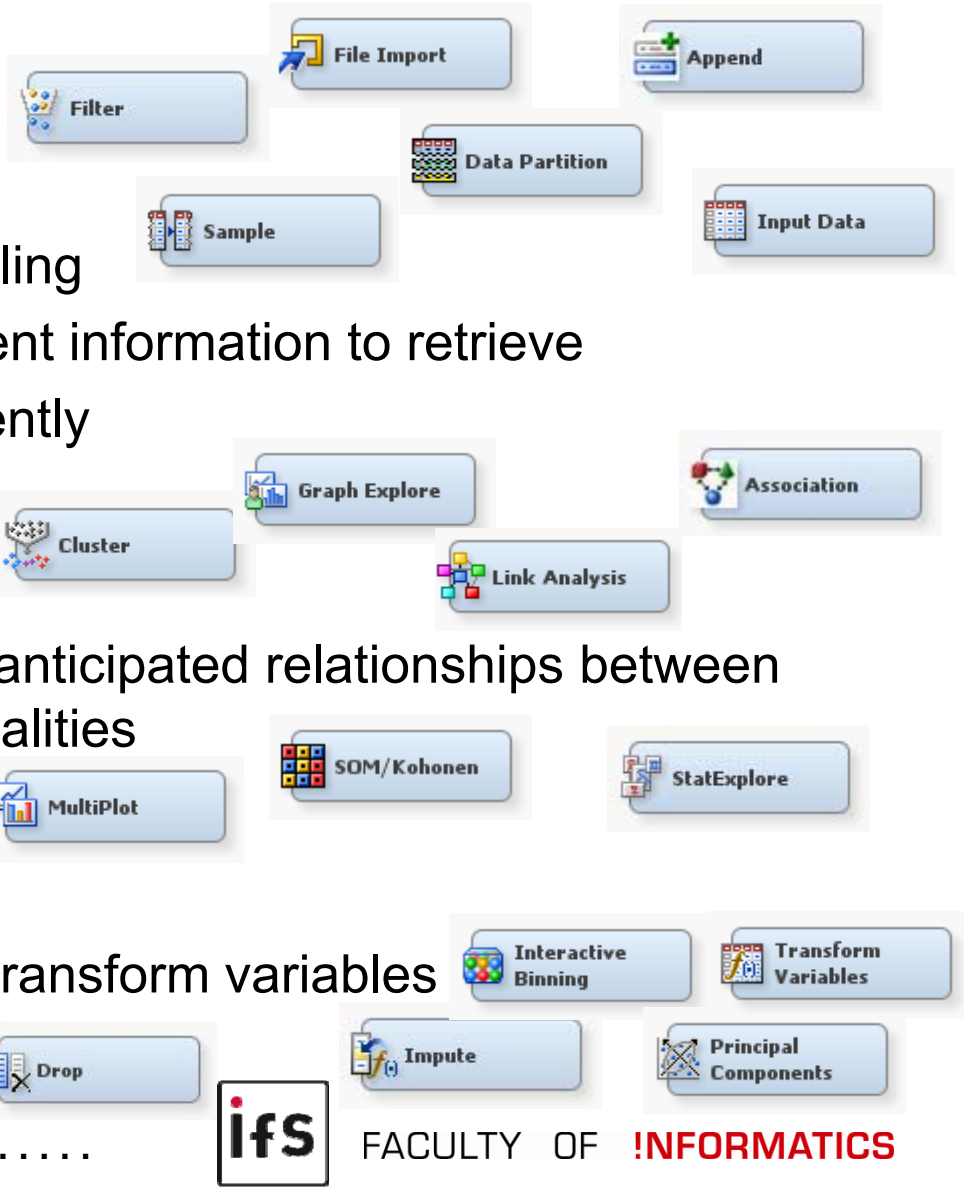
SEMMA

- *Sample*, *Explore*, *Modify*, *Model*, and *Assess*
- Model by SAS Institute
- Focused on SAS Enterprise Miner, but still generic
- Data Mining Using SAS(R) Enterprise Miner(TM): A Case Study Approach, Third Edition
  - http://support.sas.com/documentation/cdl/en/emcs/66392/HTML/default/viewer.htm
  - http://support.sas.com/documentation/cdl/en/emcs/66392/HTML/default/viewer.htm#n0pejm83csbja4n1xueveo2uoujy.htm
- Similar to Fayad: focus (only) on the core data mining process

BI Process Models

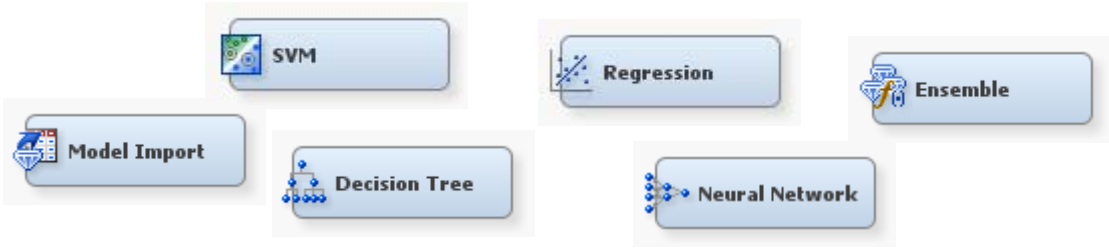
SEMMA

- **Sample**
  - selecting the data set for modeling
  - large enough to contain sufficient information to retrieve
  - small enough to be used efficiently
- **Explore**
  - understanding of the data
  - discovering anticipated and unanticipated relationships between the variables, and also abnormalities
  - includes data visualization
- **Modify**
  - methods to select, create and transform variables
  - preparation for data modeling

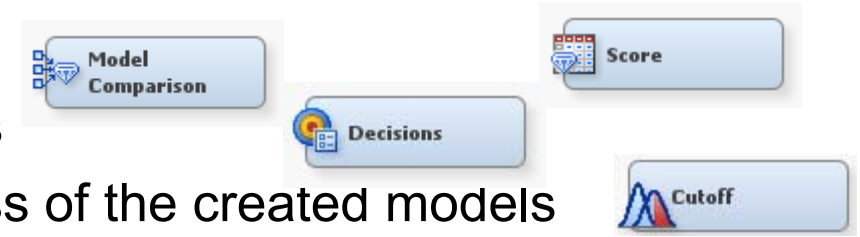


BI Process Models

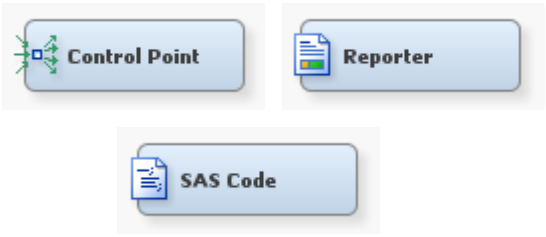
SEMMA



- **Model**
  - apply various modeling (data mining) techniques to create models that possibly provide the desired outcome
- **Assess**
  - evaluation of the modeling results
  - Verify the reliability and usefulness of the created models



- **Utility nodes**
  - For control flow, reporting, programming, …



BI Process Models

Data analytics process models

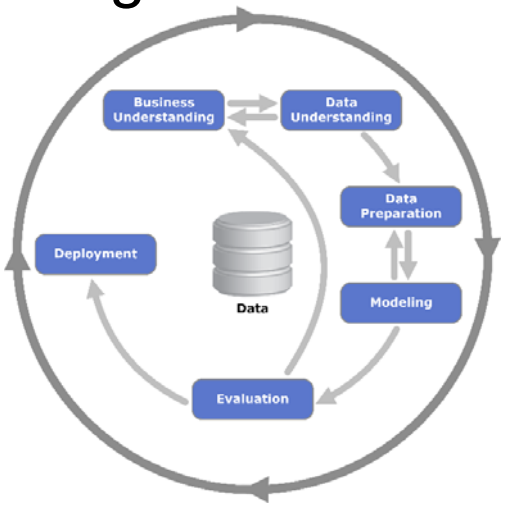
- Fayyad's KDD process
- SEMMA
- CRISP-DM
- ASUM-DM

- Reference models
  - Decide and adapt process to organizational needs!
  - Balance structure – flexibility!

How to do Data Mining

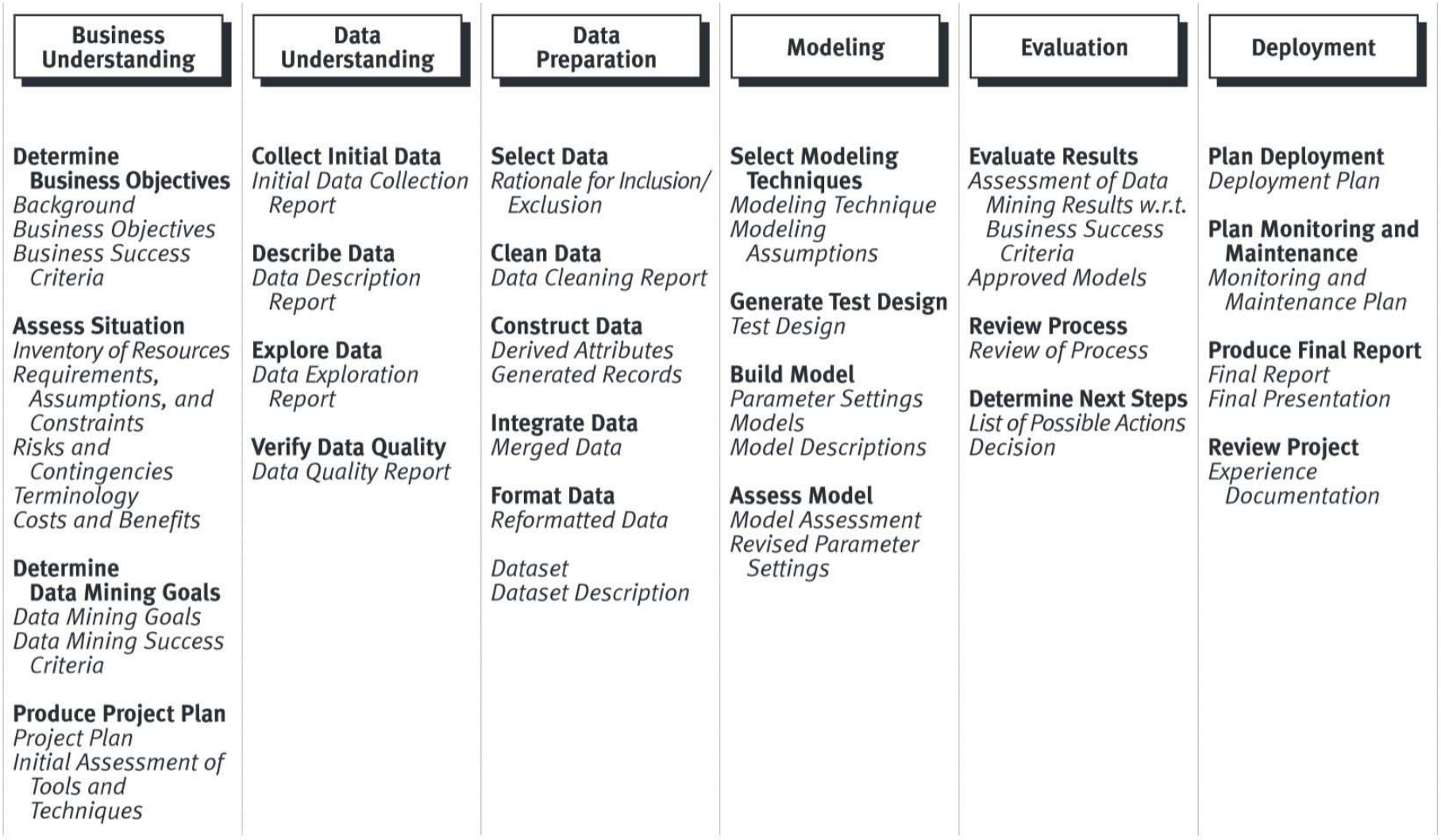
CRISP-DM

- Cross-Industry Standard Process for Data Mining
- Initiated by 3 industry members in 1996
  - Daimler-Chrysler, SPSS (then ISL), NCR
  - published in 1999
  - Over 200 members joined
  - Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS), Rüdiger Wirth (DaimlerChrysler): CRISP-DM Step-by-Step Data Mining Guide, 76pp, 1999.



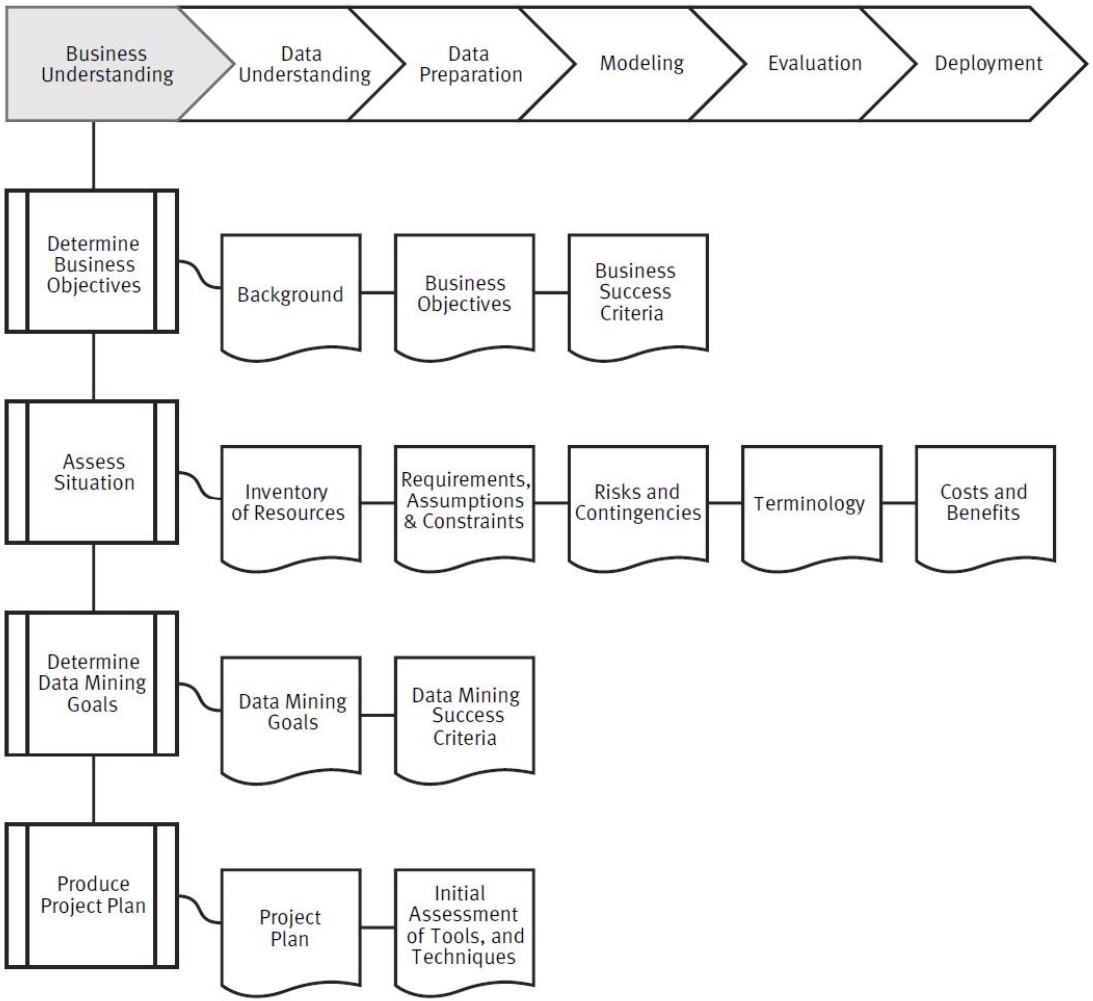
How to do Data Mining

CRISP-DM



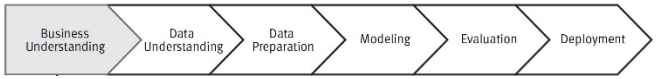
- Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

CRISP-DM – Phase 1: Business Understanding



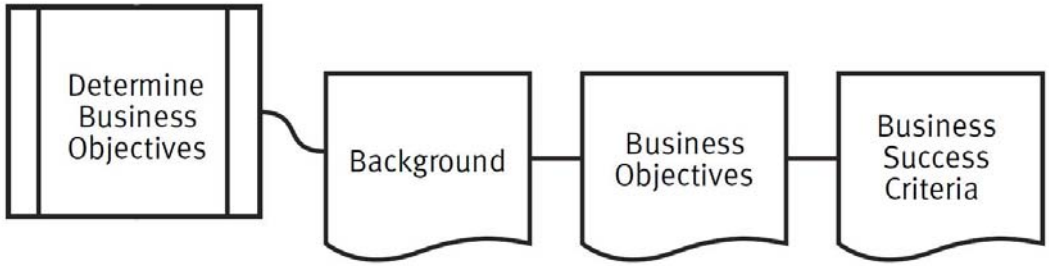
- 4 Tasks
- 12 Outputs

CRISP-DM



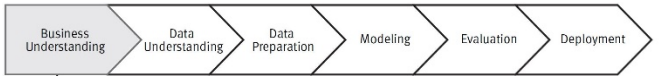
1 Business Understanding

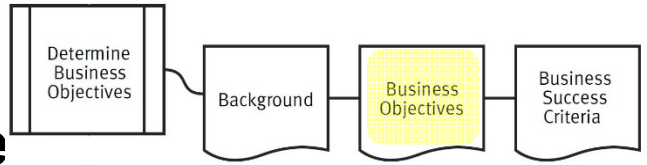
Task 1.1 Determine Business Objectives:



- What does the customer want to accomplish?
- Competing objectives and constraints
- Uncover factors that can influence the final outcome
- Risk: investing a great deal of effort producing the correct answers to the wrong questions.
- 3 Outputs:
  - **1.1.1 Background**
  - **1.1.2 Business objectives**
  - **1.1.3 Business success criteria**

CRISP-DM



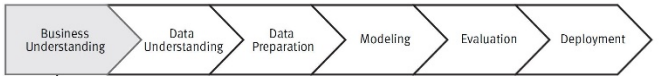


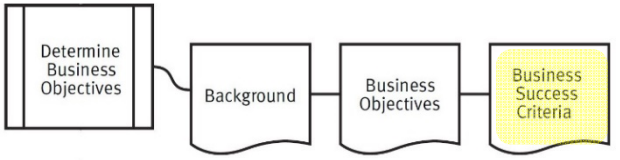
1 Business understanding

Task 1.1 Determine Business Objective

- **Output 1.1.2 Business objectives**
  - Customer's primary objective, from a business perspective
  - Identify related business
- Activities
  - Informally describe the problem to be solved
  - Specify all business questions as precisely as possible
  - Specify expected benefits in business terms
  - Beware of setting unattainable goals—make them as realistic as possible

CRISP-DM



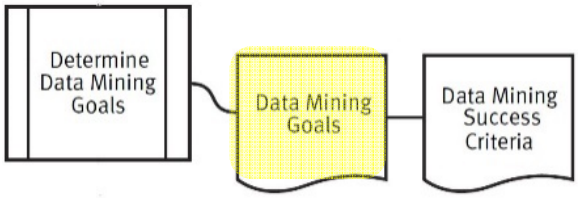


1 Business understanding

Task 1.1 Determine business objectives

- **Output 1.1.3 Business success criteria**
  - Describe criteria for a successful or useful outcome
    - specific and readily measurable
    - general and subjective
      (indicate who will make the subjective judgment!)
- Activities
  - Specify business success criteria
  - Identify who assesses the success criteria
  - Each of the success criteria should relate to at least one of the specified business objectives

CRISP-DM



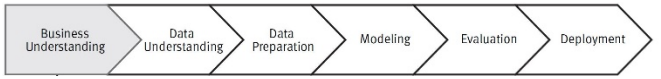
1 Business understanding

Task 1.3 Determine data mining goals

Output 1.3.1: Data mining goals

- Goal:
  - Describe outputs that enable achievement of the business objectives
- Activities
  - Translate the business questions to data mining goals (e.g., marketing campaign requires segmentation of customers; the level/size of the segments should be specified).
  - Specify DM problem type (e.g., classification, regression, clustering)
  - Note: double-check correct match between business and DM goals!
    - E.g. "*Improve quality of products*" -> *"given process monitoring data predict quality of resulting product"? "given process data, at what time can I detect a deviation from the ideal process? Recommend corrections?" …*

CRISP-DM



1 Business understanding

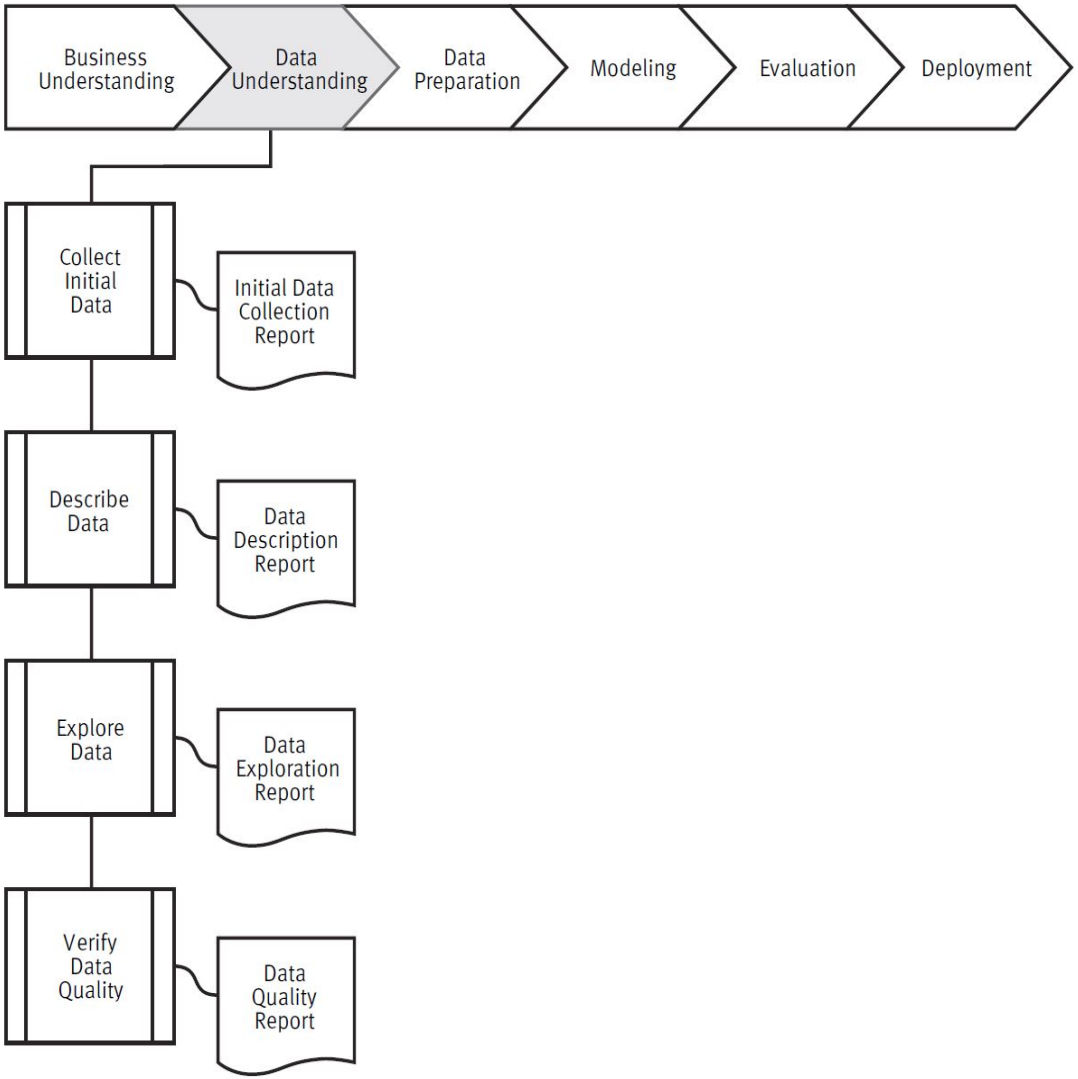
Task 1.3 Determine data mining goals

Output 1.3.2 Data mining success criteria

- Goals:
  - Define criteria for a successful outcome in technical terms
  - For subjective criteria identify persons making the judgment
- Activities
  - Specify criteria for model assessment (e.g., model accuracy, performance, robustness, complexity)
  - Define benchmarks for evaluation criteria
  - Specify criteria which address subjective assessment criteria (e.g., model explain ability)
  - Consider deployment aspects
- Note: data mining success criteria are different than the business success criteria!

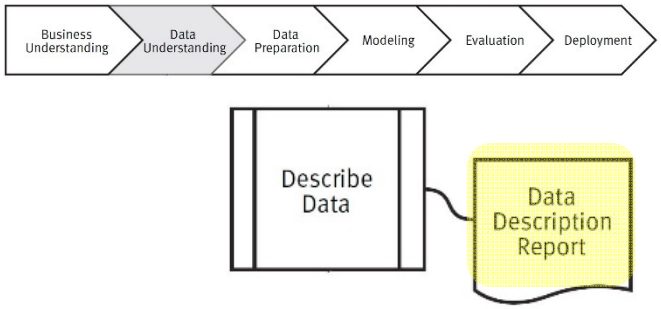
CRISP-DM

CRISP-DM – Phase 2: Data Understanding



- 4 Tasks
- 4 Outputs

CRISP-DM



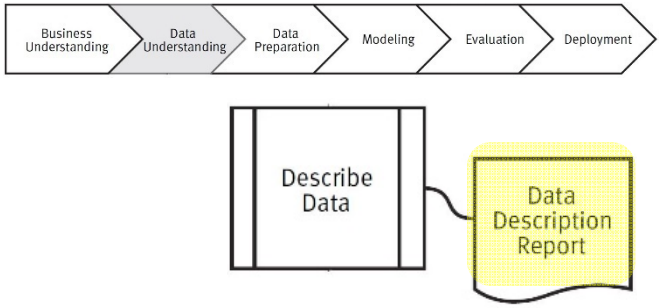
2 Data Understanding

Task 2.2 Describe data

Output 2.2.1 Data description report

- Activities
  - Attribute types and values checking
  - Volumetric analysis of data
    - Identify data and method of capture
    - Perform basic statistical analyses
    - Report tables and their relations
    - Check data volume, number of multiples, complexity
    - Check specifically for free text entries

CRISP-DM



2 Data Understanding

Task 2.2 Describe data

Output 2.2.1 Data description report

- Activities
  - Attribute types and values checking
  - Volumetric analysis of data
    - Identify data and method of capture
    - Perform basic statistical analyses
    - Report tables and their relations
    - Check data volume, number of multiples, complexity
    - Check specifically for free text entries

Excursion: Attribute Types

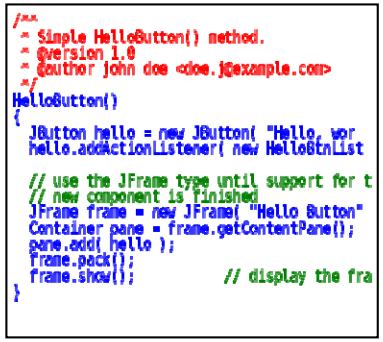
Definitions

- **Concepts**: things that can be learned
  - E.g. list of topics for texts, spam/non-spam for email, groups of similar animals, sub-groups in a social network, correlation between smoking and lung cancer, …
- **Instance**: example of a concept, data point
  - E.g. individual text documents; animals; social network nodes; individual persons
- **Attribute**: measurement/description of an instance
  - E.g. text described by BOW using tfidf; animals described by characteristics such as #legs, fur/feathers, food; social network nodes represented by their connections to other nodes; people described by smoking habits and degree of cancer

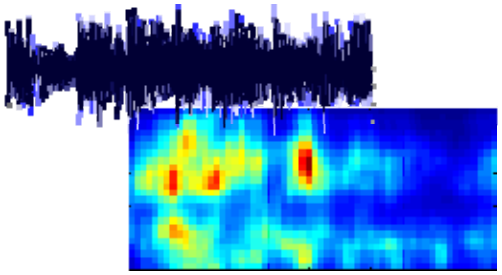
Excursion: Attribute Types

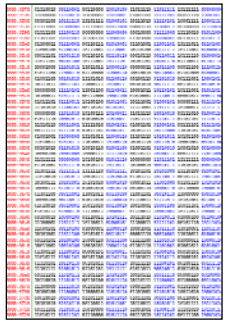
- Instances are described by attributes
- Which types of attributes exist for different types of data?

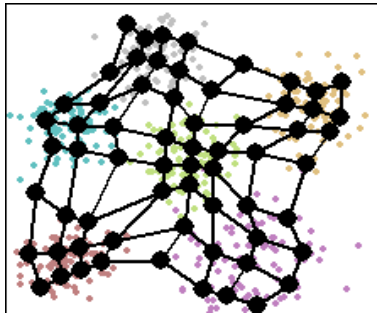

EVALUATION OF FEATURE EXTRACTORS AND PSYCHO-ACOUSTIC TRANSFORMATIONS FOR MUSIC GENRE CLASSIFICATION

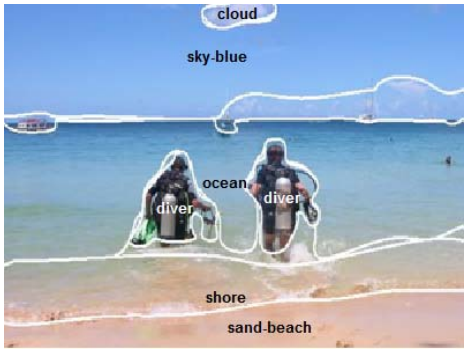

```
/**
 * Simple HelloButton() method.
 * @version 1.0
 * @author john doe <doe.j@example.com>
 */
HelloButton()
{
  JButton hello = new JButton( "Hello, wor
  hello.addActionListener( new HelloBtnList

  // use the JFrame type until support for t
  // new component is finished
  JFrame frame = new JFrame( "Hello Button"
  Container pane = frame.getContentPane();
  pane.add( hello );
  frame.pack();
  frame.show();              // display the fra
}
```









Excursion: Attribute Types

- Instances are described by attributes
- Which types of attributes exist for different types of data?
- **4 main types**
  - **Nominal**
  - **Ordinal**
  - **Interval**
  - **Ratio**
- Different representations
  - Flat file
  - Complex structures -> may need to be flattened
- Influences choice of ML algorithms

Excursion: Attribute Types

1. Nominal

- Latin for „name“
- Distinct labels from a defined vocabulary
- Classification: class labels are nominal values
  - Music: genres (jazz, pop, rock, …)
  - Text: spam/non-spam; sports, politics, weather; report, interview
- Attributes can be nominal too
  - Persons: eye color, hair color, city of birth
  - Nominal attributes can be numeric (e.g. Zip-code, numeric encodings of categories)
- Math: only **equality**! (don‘t subtract ZIP-codes!)

Excursion: Attribute Types

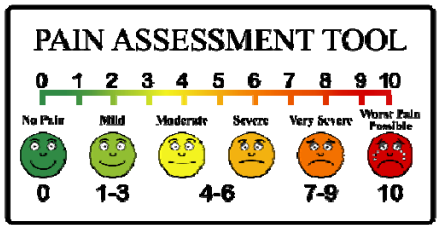
2. Ordinal (aka categorical)

- Impose an order on discrete categories
- But: **no distance** defined!
- Distinct labels from a defined vocabulary, numeric or strings
  - Temperature: cold < cool < mild < hot < very hot
  - Grades: A > B > C > D > E > F; 1 > 2 > 3 > 4 > 5
- Math: ordering: larger, smaller, equal
  no additions / subtractions!

Excursion: Attribute Types

3. Interval

- Ordered elements with fixed distance in-between
- Discrete or continuous values
- Distinct labels from a defined vocabulary, numeric or strings
  - Time: year -> can calculate the difference between 2011 and 2018
  - Levels of pain

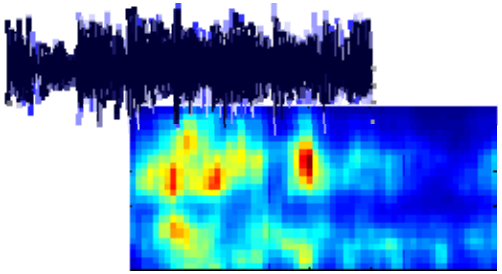


- Math:
  - ordering: larger, smaller, equal
  - Difference / distance -> subtraction
  - no additions! (the Year 2011 + the Year 2018 do not make sense) (note 3 years Bachelor + 2 years Master + 3 years PhD do make sense – that‘s a different type of attribute!)

Excursion: Attribute Types

4. Ratio

- Continuous values, zero-point defined
- Usually represented as real numbers
  - Textmining: BOW using tfxidf
  - Images: color histograms
  - Audio: features extracted from frequency spectrum
  - Sensor mesurements



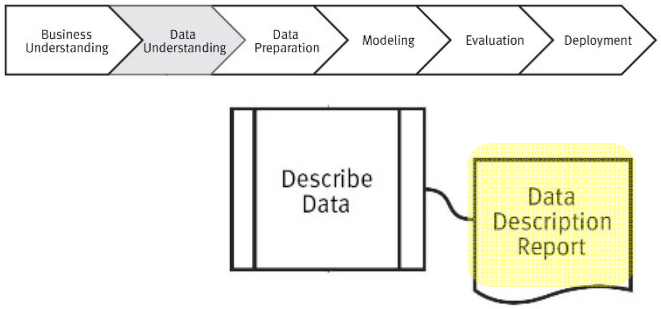
- Cannot be used as class labels! (-> binning or regression)

- Math: all operations allowed

Excursion: Attribute Types

- Understand attributes and attribute types
  - from a data mining perspective
  - from a business perspective!
- Explicitly determine
  - Attribute type
  - Meaning (special values)
  - Encoding
  - Value ranges

CRISP-DM



2 Data Understanding

Task 2.2 Describe data

Output 2.2.1 Data description report

- Activities
  - Attribute types and values checking
  - Volumetric analysis of data
    - Identify data and method of capture
    - Perform basic statistical analyses
    - Report tables and their relations
    - Check data volume, number of multiples, complexity
    - Check specifically for free text entries

Explore Statistical Properties

- "For each attribute compute the basic statistics"
  - Average
  - Min/max values
  - Variance, standard deviation, mode, skewness, …
  - Histogram: encoding issues (0, 99, -1, 1.1.1900, …)
  - Correlation between attributes

- Beyond this: **look at the data!**
  - Scroll through (subsample of) the data
  - Statistics don't tell you everything!
  - Use visualizations!

Explore Statistical Properties

- Anscombe's Quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Explore Statistical Properties

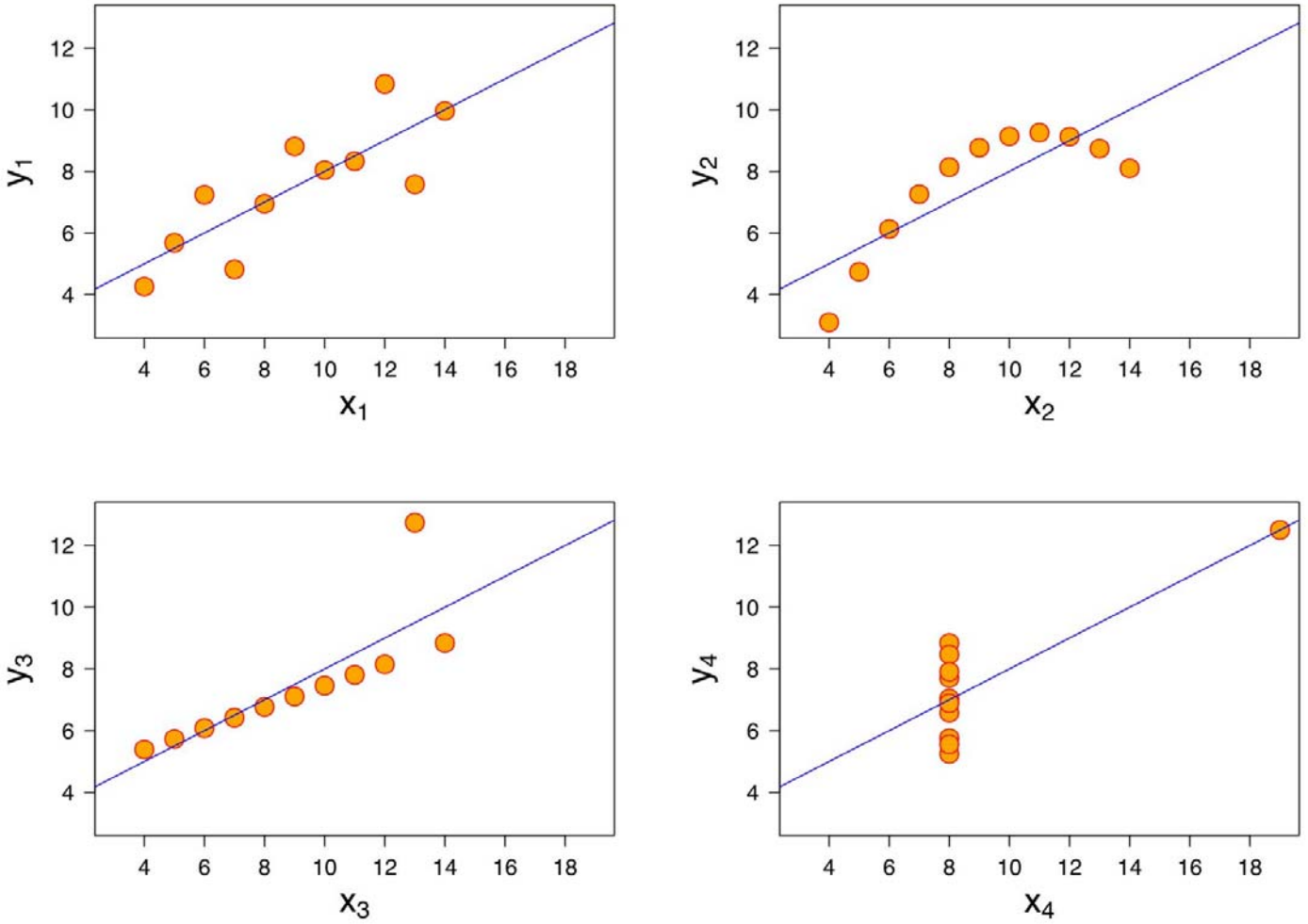
- Anscombe's Quartet

Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x*	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y*	4.125	plus/minus 0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression	0.67	to 2 decimal places

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Explore Statistical Properties

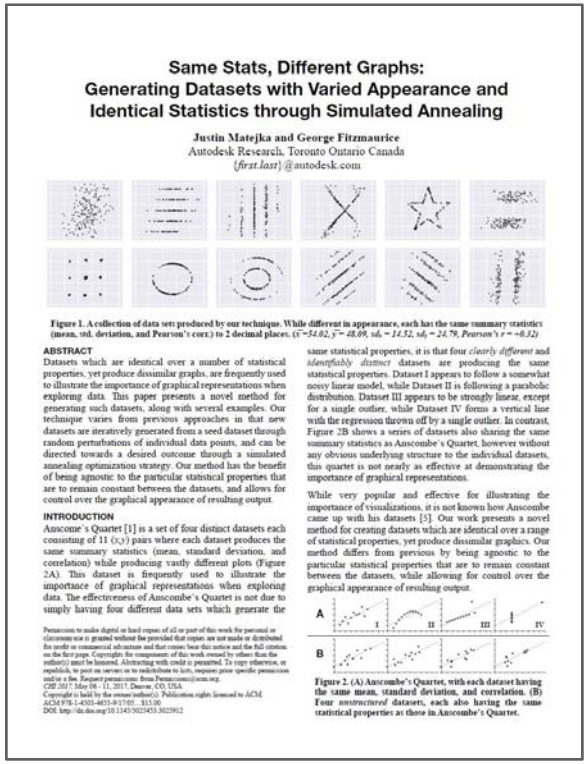
- Anscombe's Quartet



https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Explore Statistical Properties

- **Justin Matejka and George Fitzmaurice:** Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing.
  Proceedings of *CHI 2017*,
  May 06 - 11, 2017, Denver, CO, USA.
  DOI: 10.1145/3025453.3025912



Same Stats, Different Graphs:
Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing

Justin Matejka and George Fitzmaurice
Autodesk Research, Toronto Ontario Canada
{first.last}@autodesk.com

Figure 1. A collection of data sets produced by our technique. While different in appearance, each has the same summary statistics (mean, std. deviation, and Pearson's corr.) to 2 decimal places. ($\bar{x}$=54.02, $\bar{y}$= 48.09, sd_x = 14.52, sd_y = 24.79, Pearson's r = -0.32)

ABSTRACT
Datasets which are identical over a number of statistical properties, yet produce dissimilar graphs, are frequently used to illustrate the importance of graphical representations when exploring data. This paper presents a novel method for generating such datasets, along with several examples. Our technique varies from previous approaches in that new datasets are iteratively generated from a seed dataset through random perturbations of individual data points, and can be directed towards a desired outcome through a simulated annealing optimization strategy. Our method has the benefit of being agnostic to the particular statistical properties that are to remain constant between the datasets, and allows for control over the graphical appearance of resulting output.

INTRODUCTION
Anscombe's Quartet [1] is a set of four distinct datasets each consisting of 11 (x,y) pairs where each dataset produces the same summary statistics (mean, standard deviation, and correlation) while producing vastly different plots (Figure 2A). This dataset is frequently used to illustrate the importance of graphical representations when exploring data. The effectiveness of Anscombe's Quartet is not due to simply having four different data sets which generate the same statistical properties, it is that four *clearly different* and *identifiably distinct* datasets are producing the same statistical properties. Dataset I appears to follow a somewhat noisy linear model, while Dataset II is following a parabolic distribution. Dataset III appears to be strongly linear, except for a single outlier, while Dataset IV forms a vertical line with the regression thrown off by a single outlier. In contrast, Figure 2B shows a series of datasets also sharing the same summary statistics as Anscombe's Quartet, however without any obvious underlying structure to the individual datasets, this quartet is not nearly as effective at demonstrating the importance of graphical representations.

While very popular and effective for illustrating the importance of visualizations, it is not known how Anscombe came up with his datasets [5]. Our work presents a novel method for creating datasets which are identical over a range of statistical properties, yet produce dissimilar graphics. Our method differs from previous by being agnostic to the particular statistical properties that are to remain constant between the datasets, while allowing for control over the graphical appearance of resulting output.

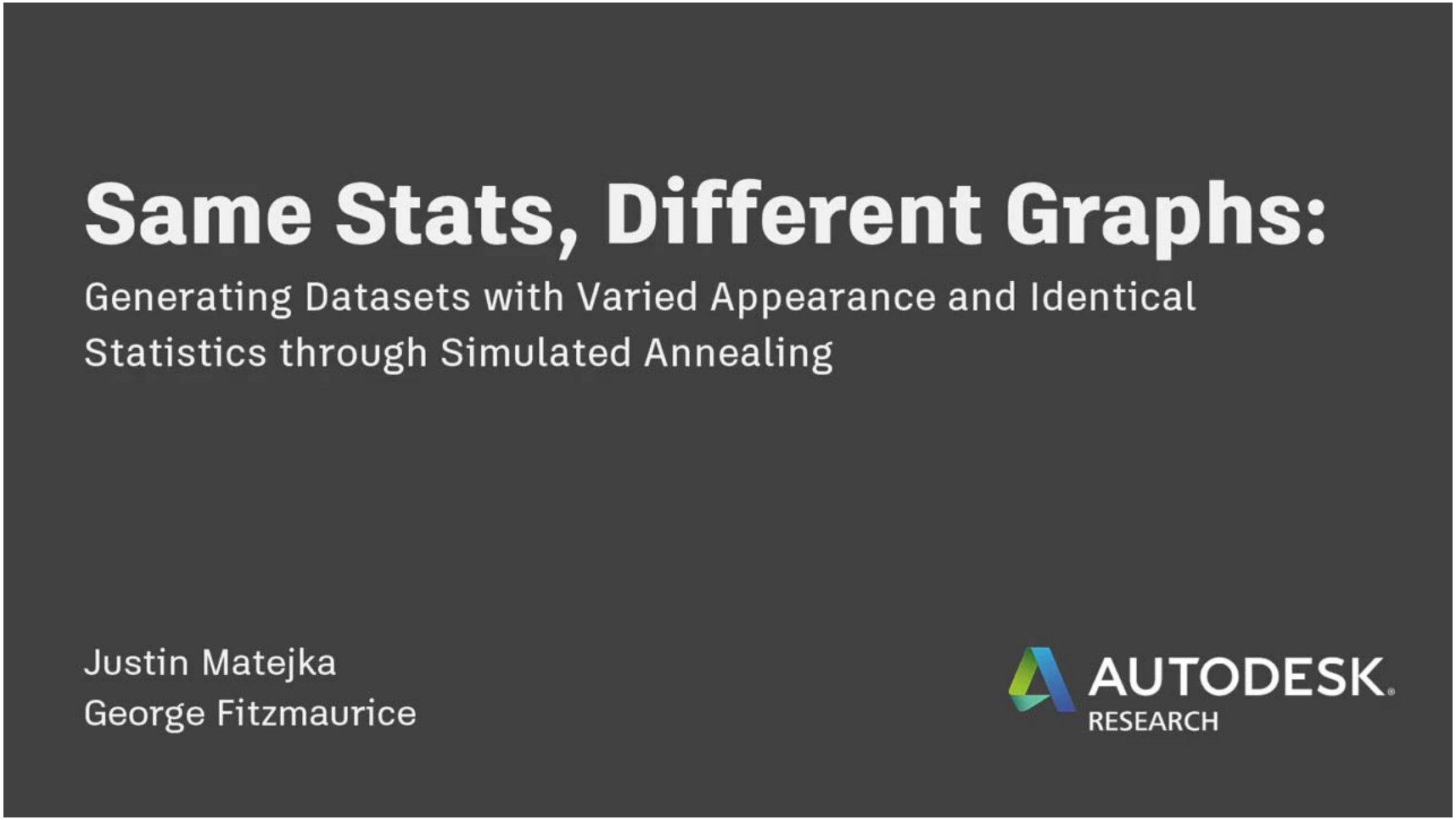
CHI 2017, May 06 - 11, 2017, Denver, CO, USA

ACM 978-1-4503-4655-9/17/05…$15.00
DOI: http://dx.doi.org/10.1145/3025453.3025912

Figure 2. (A) Anscombe's Quartet, with each dataset having the same mean, standard deviation, and correlation. (B) Four *unstructured* datasets, each also having the same statistical properties as those in Anscombe's Quartet.

Explore Statistical Properties

- **Justin Matejka and George Fitzmaurice:** CHI'17



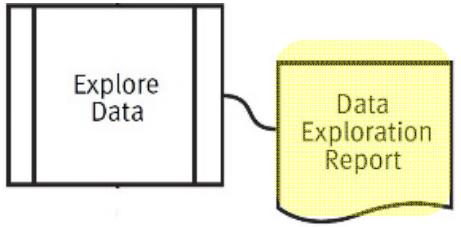
https://www.youtube.com/watch?v=DbJyPELmhJc

Explore Statistical Properties

- "For each attribute compute the basic statistics"
  - Average
  - Min/max values
  - Variance, standard deviation, mode, skewness, …
  - Histogram: encoding issues (0, 99, -1, 1.1.1900, …)
  - Correlation between attributes

- Beyond this: look at the data!
  - Scroll through (subsample of) the data
  - Statistics don't tell you everything!
  - **Cross-check semantics and attribute values!**
  - Next step (2.3 Explore Data) focuses also on visual exploration

CRISP-DM



2 Data Understanding

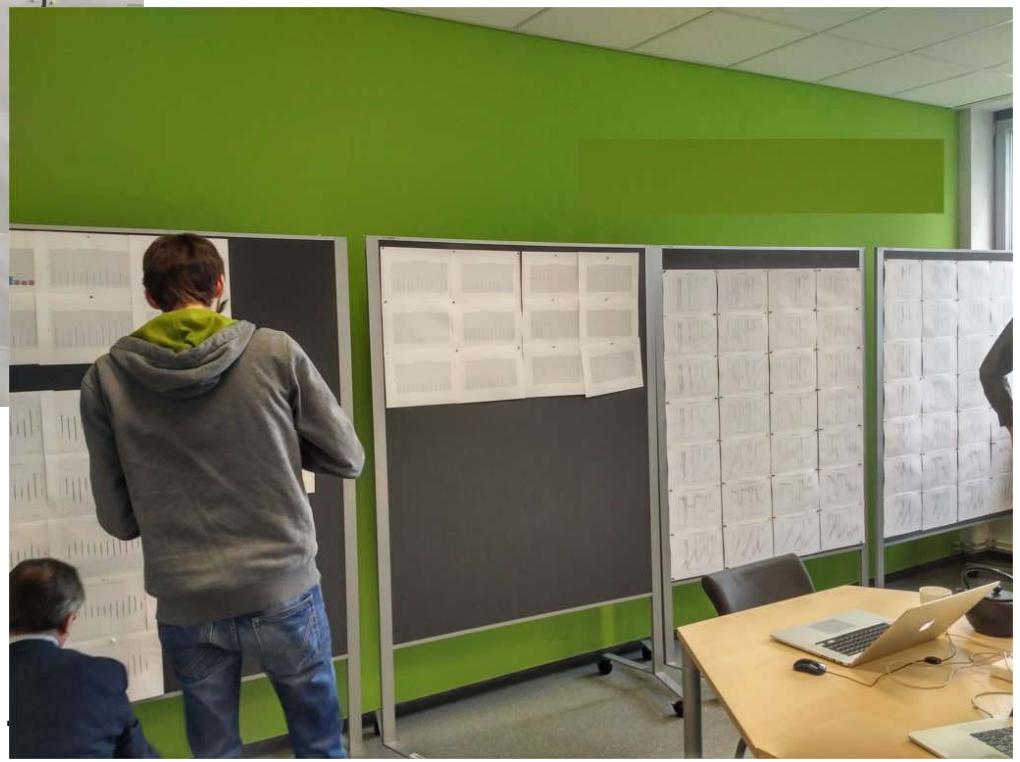
Task 2.3 Explore data:

Output 2.3.1 Data exploration report

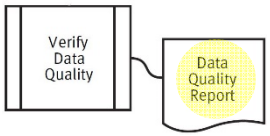
- Activities
  - Data exploration
    - Analyze (**visualize!**) properties of interesting attributes in detail (e.g., basic statistics, interesting sub-populations)
    - Identify characteristics of sub-populations
  - Form suppositions for future analysis
    - Form hypotheses and identify actions
    - Transform the hypothesis into a data mining goal, if possible
    - Clarify data mining goals or make them more precise ("blind" search may be useful as well)
    - Perform basic analysis to verify the hypotheses

Explore Statistical Properties

- **Example: predictive maintenance**



CRISP-DM



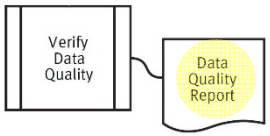
2 Data Understanding

Task 2.4 Verify data quality:

Output 2.4.1 Data quality report

- Activities
  - Review attributes
    - Identify special values and catalog their meaning
    - Check coverage (e.g., are all possible values represented?)
    - Verify that the meanings of attributes and contained values fit
    - Identify missing attributes and blank fields
    - **Establish the meaning of missing data ! Why is it missing?**
    - Check for attributes with different values that have similar meanings (e.g., low fat vs. diet used in different places/times)
    - Check spelling and format of values (e.g., same value but sometimes beginning with lower-case sometimes with upper-case letter)

CRISP-DM



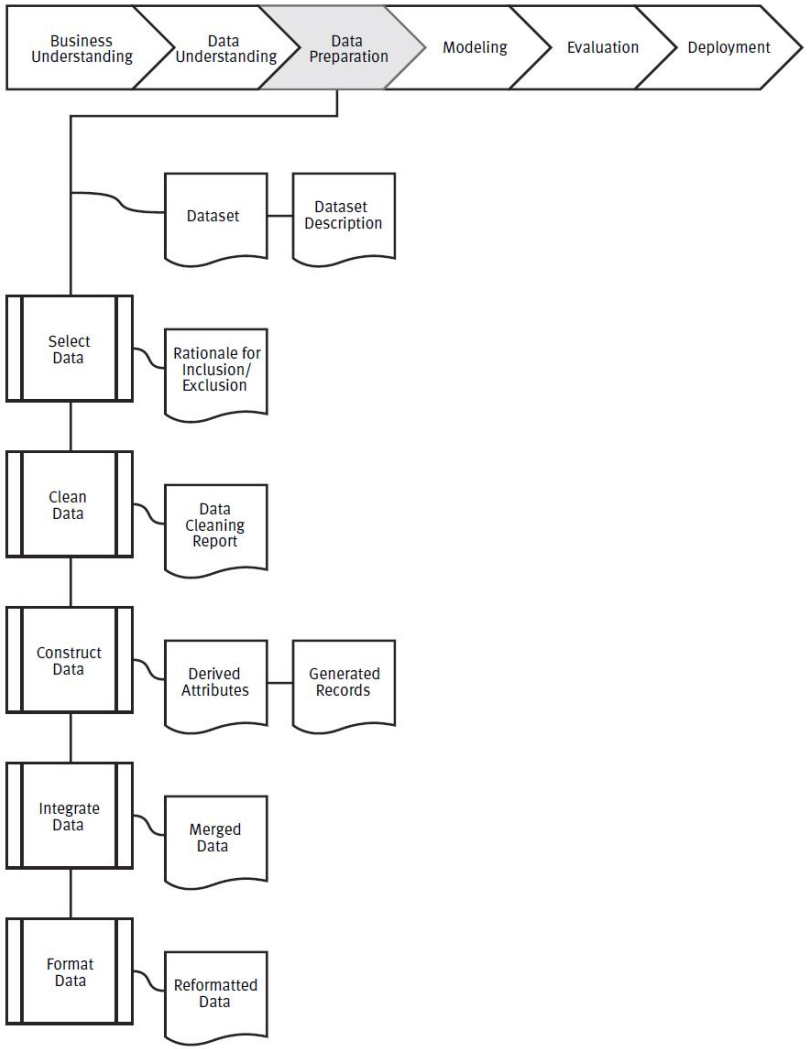
2 Data Understanding

Task 2.4 Verify data quality

Output 2.4.1 Data quality report

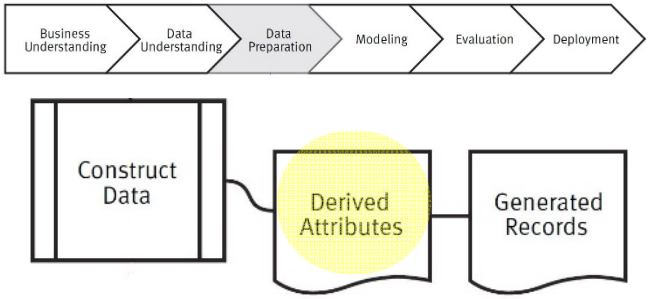
- Activities (cont.)
  - Review attributes (cont.)
    - Check for deviations, decide whether it is “noise” or may indicate an interesting phenomenon
    - Check for plausibility of values (e.g., all fields having the same or nearly the same values)
    - Review any attributes that give answers that conflict with common sense (e.g., teenagers with high income levels – *unless dataset contains YouTube influencers / Start-up millionaires*)
    - If flat files, check delimiter used and consistency within attributes
    - If flat files, check the number of fields in each record to see if they coincide

CRISP-DM – Phase 3: Data Preparation



- 5 Tasks
- 8 Outputs

CRISP-DM



3 Data Preparation

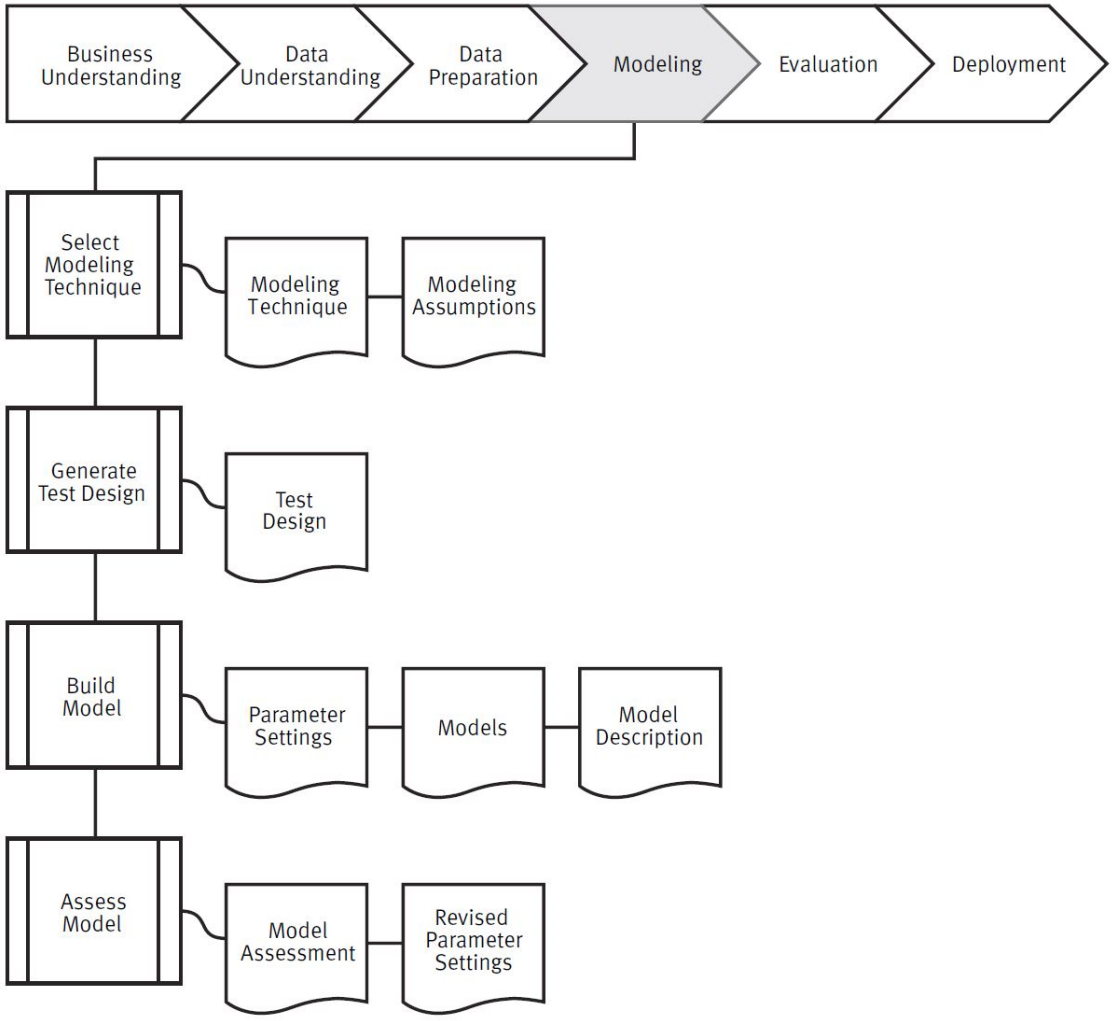
Task 3.3 Construct data

Output 3.3.1 Derived attributes

- Activities
  - Transform to different attribute types (Binning, 1-to-n coding, …)
  - Decide if any attribute should be normalized (e.g., k-means clustering algorithm with age and income)
  - How can missing attributes be constructed or imputed? Decide type of construction (e.g., aggregate, average, induction)
  - Add new attributes to the accessed data
  - Consider adding new information on the relevant importance of attributes by adding new attributes (e.g. weighted normalization)

CRISP-DM

CRISP-DM – Phase 4: Modeling

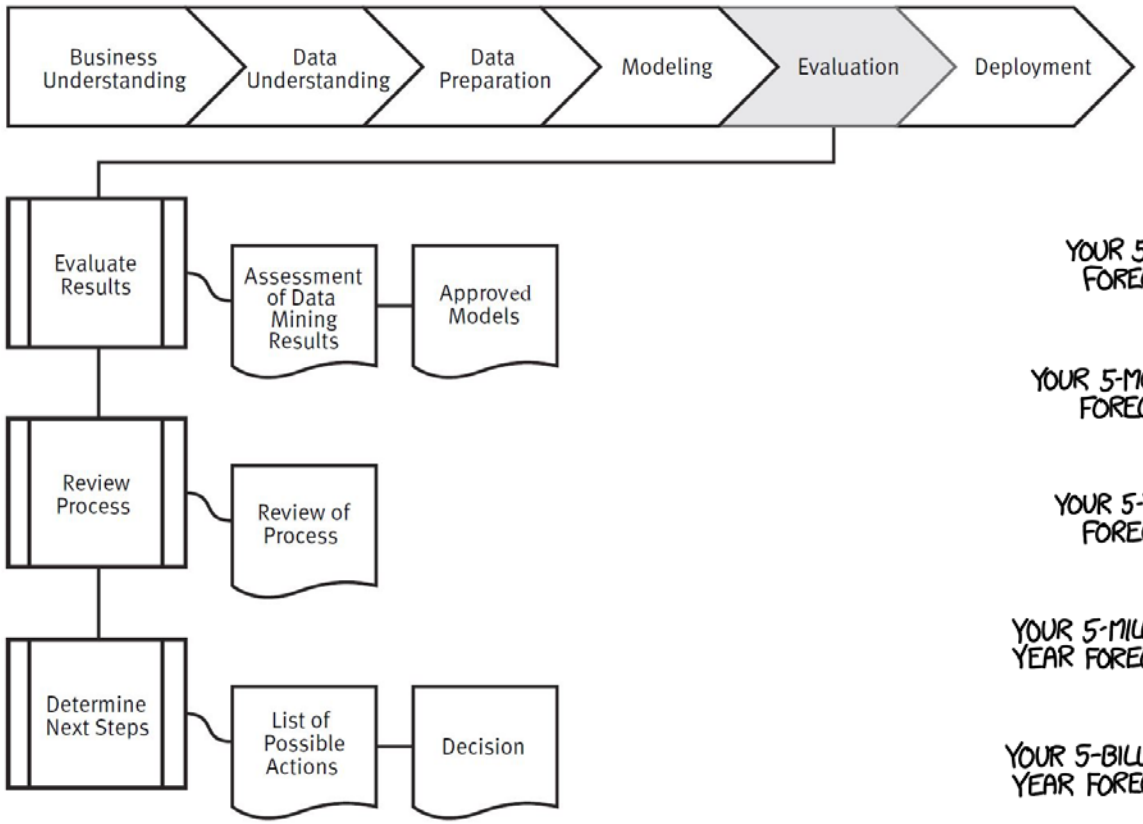


- 4 Tasks
- 8 Outputs

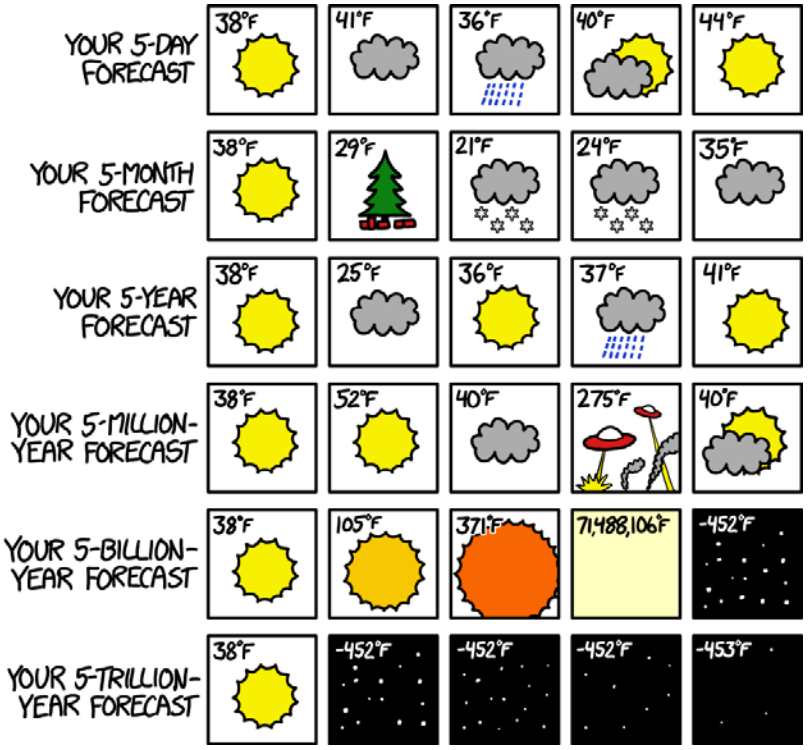
xkcd: https://xkcd.com/1838/

CRISP-DM

CRISP-DM – Phase 5: Evaluation

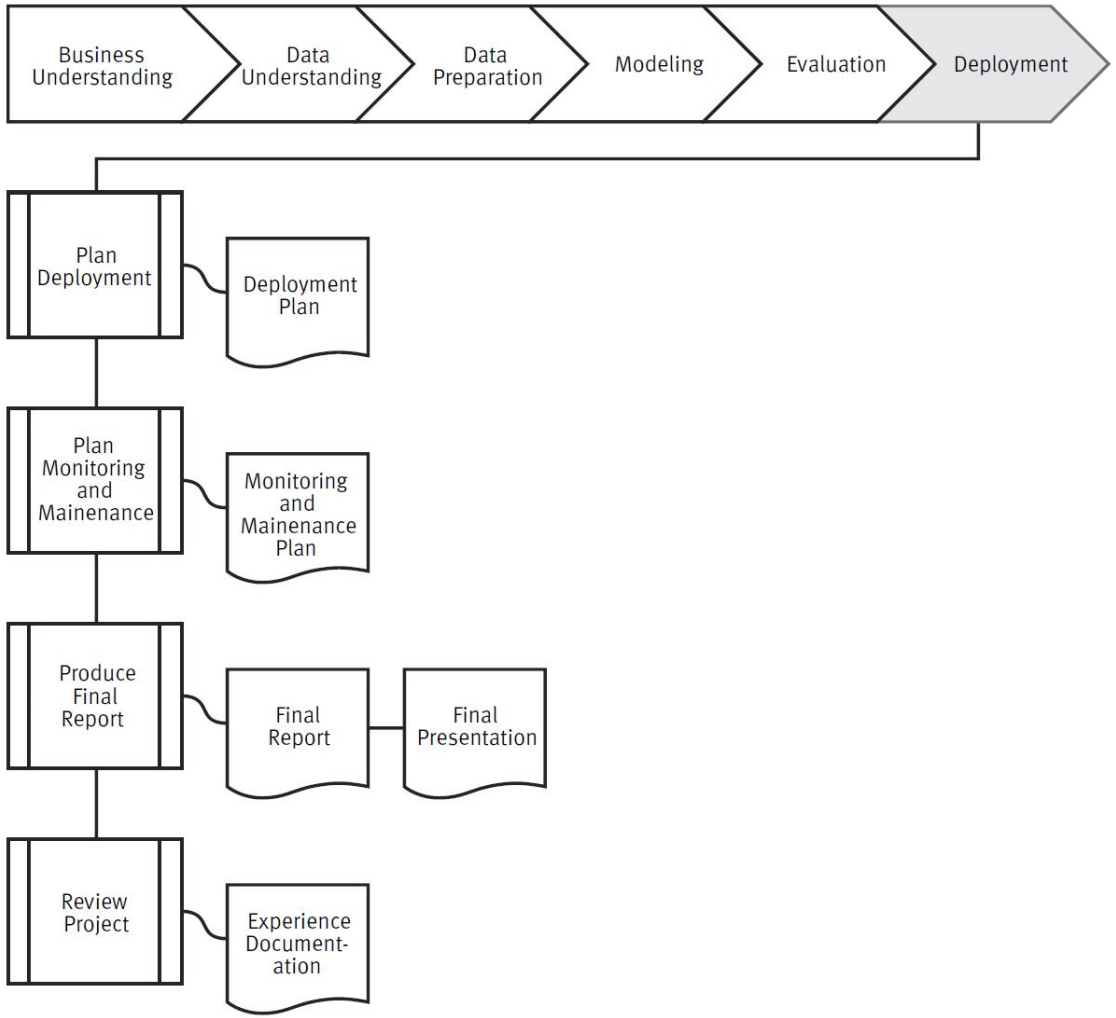


- 3 Tasks
- 5 Outputs



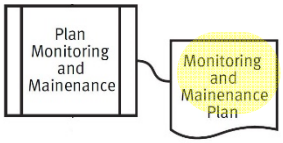
https://xkcd.com/1606/

CRISP-DM – Phase 6: Deployment



- 4 Tasks
- 5 Outputs

CRISP-DM

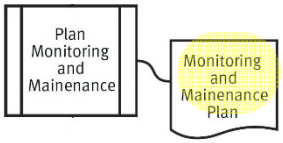


6 Deployment

Task 6.2 Plan monitoring and maintenance

- Goal:
  - Monitoring and maintenance are essential in continuous use
  - Monitoring for data drift, bias, …
  - Needs to be (semi-)automated!
  - Maintenance strategy
  - Avoid unnecessarily long periods of incorrect usage of data mining results
  - **What to do if system "misbehaves"?**
    (Is "pulling the plug" an option?)
- **Output 6.2.1 Monitoring and maintenance plan**
  - Summarize monitoring and maintenance strategy, including necessary steps and how to perform them

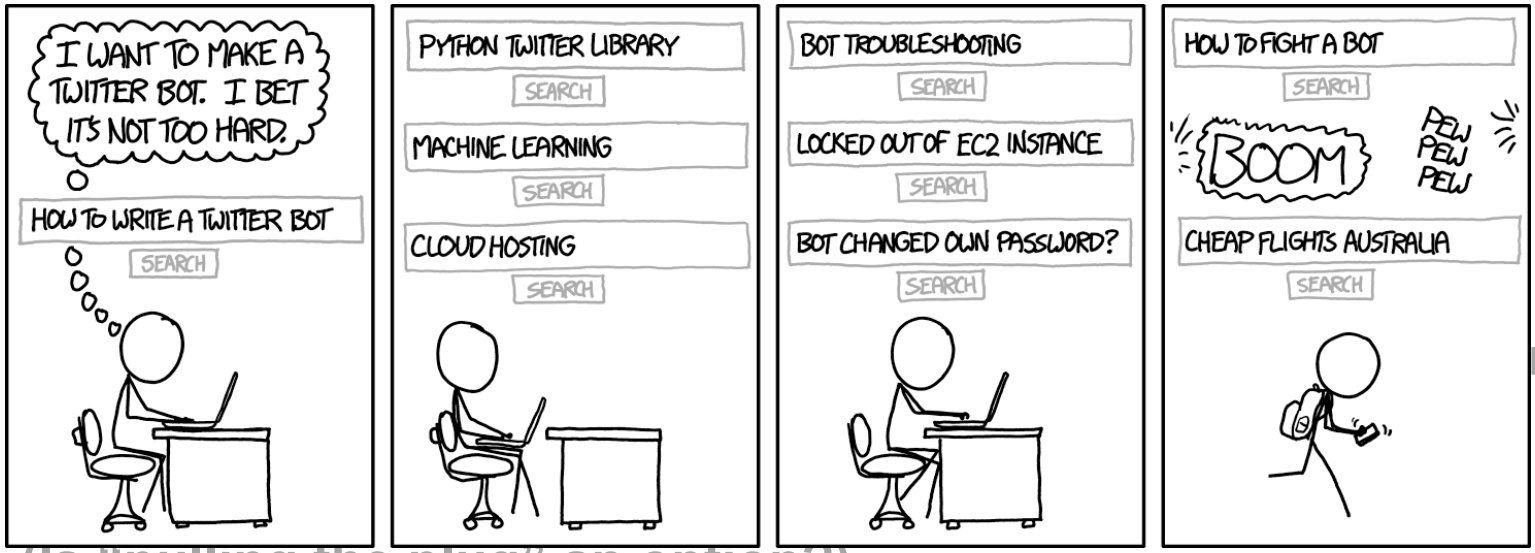
CRISP-DM



6 Deployment

Task 6.2 Plan monitoring and maintenance

- Goal:
  -
  -
  -
  -
  -
  -



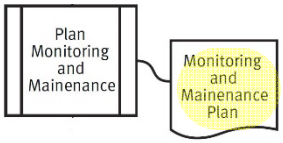
ng

(Is "pulling the plug" an option?)

https://xkcd.com/1646/

- **Output 6.2.1 Monitoring and maintenance plan**
  - Summarize monitoring and maintenance strategy, including necessary steps and how to perform them

CRISP-DM



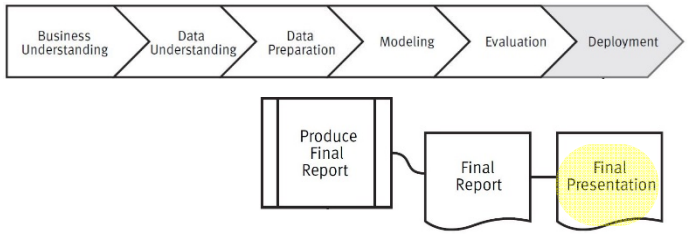
6 Deployment

Task 6.2 Plan monitoring and maintenance

Output 6.2.1 Monitoring and maintenance plan

- Activities
  - Check for dynamic aspects (i.e., what things could change in the environment?)
  - Decide how accuracy/errors/… will be monitored
  - Determine when the result or model should not be used any more
  - Identify criteria (validity, threshold of accuracy, new data, change in the application domain, etc.), and what should happen if the model or result could no longer be used
    (update model, set up new data mining project, etc.).
  - Will business objectives of the use of the model change over time?
  - Develop monitoring and maintenance plan

CRISP-DM



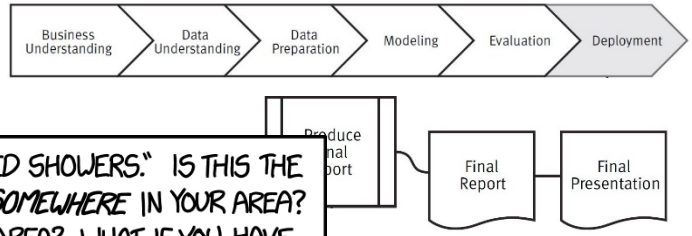
6 Deployment

Task 6.3 Produce final report

Output 6.3.2 Final presentation

- (Final) presentation(s) to summarize the project
- To the management sponsor, key stakeholders, PR, …
- Activities
  - Decide on target group for the presentation
  - Select which items from the final report to be included in presentation
  - Communicate clearly, **addressing the target groups**!
  - Training, assistance in deployment
  - Expectation management: tasks in operation!

CRISP-DM

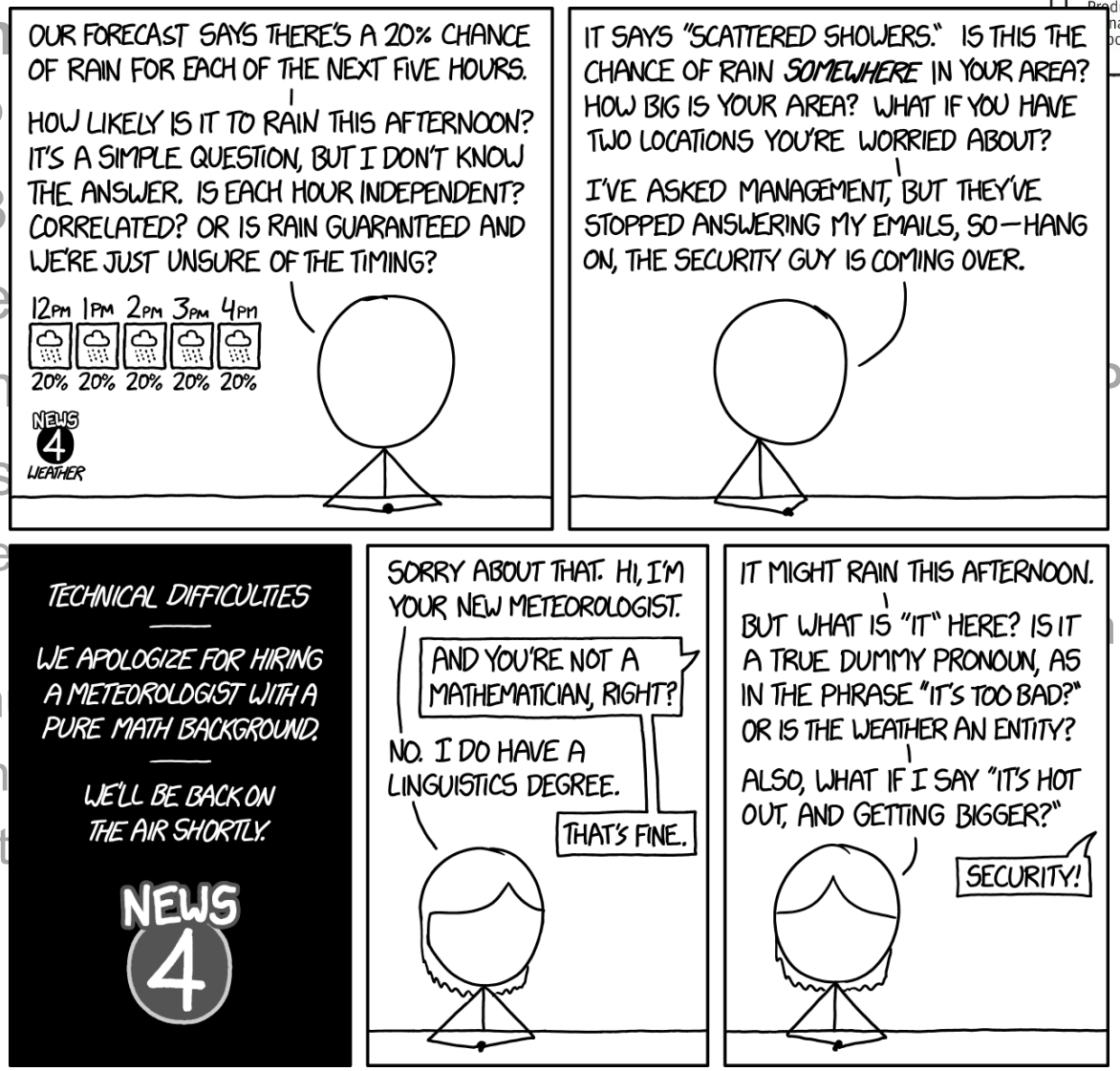


6 Deploym

Task 6.3 P

Output 6.3

- Final pre
- To the m PR, …
- Activities
  - Decide
  - Select presentation
  - Comm
  - Trainin
  - Expect



https://xkcd.com/1985/

CRISP-DM

CRISP-DM

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	**Select Data** *Rationale for Inclusion/ Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data* *Dataset* *Dataset Description*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Descriptions* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

BI Process Models

Data analytics process models

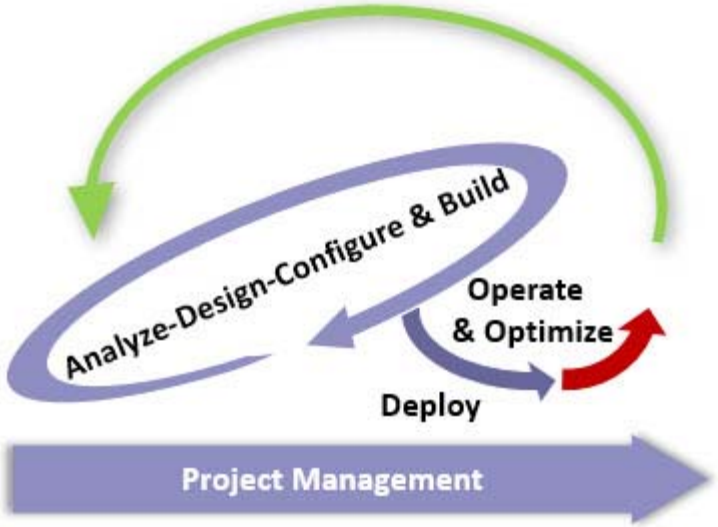
- Fayyad's KDD process
- SEMMA
- CRISP-DM
- ASUM-DM

- Reference models
  - Decide and adapt process to organizational needs!
  - Balance structure – flexibility!

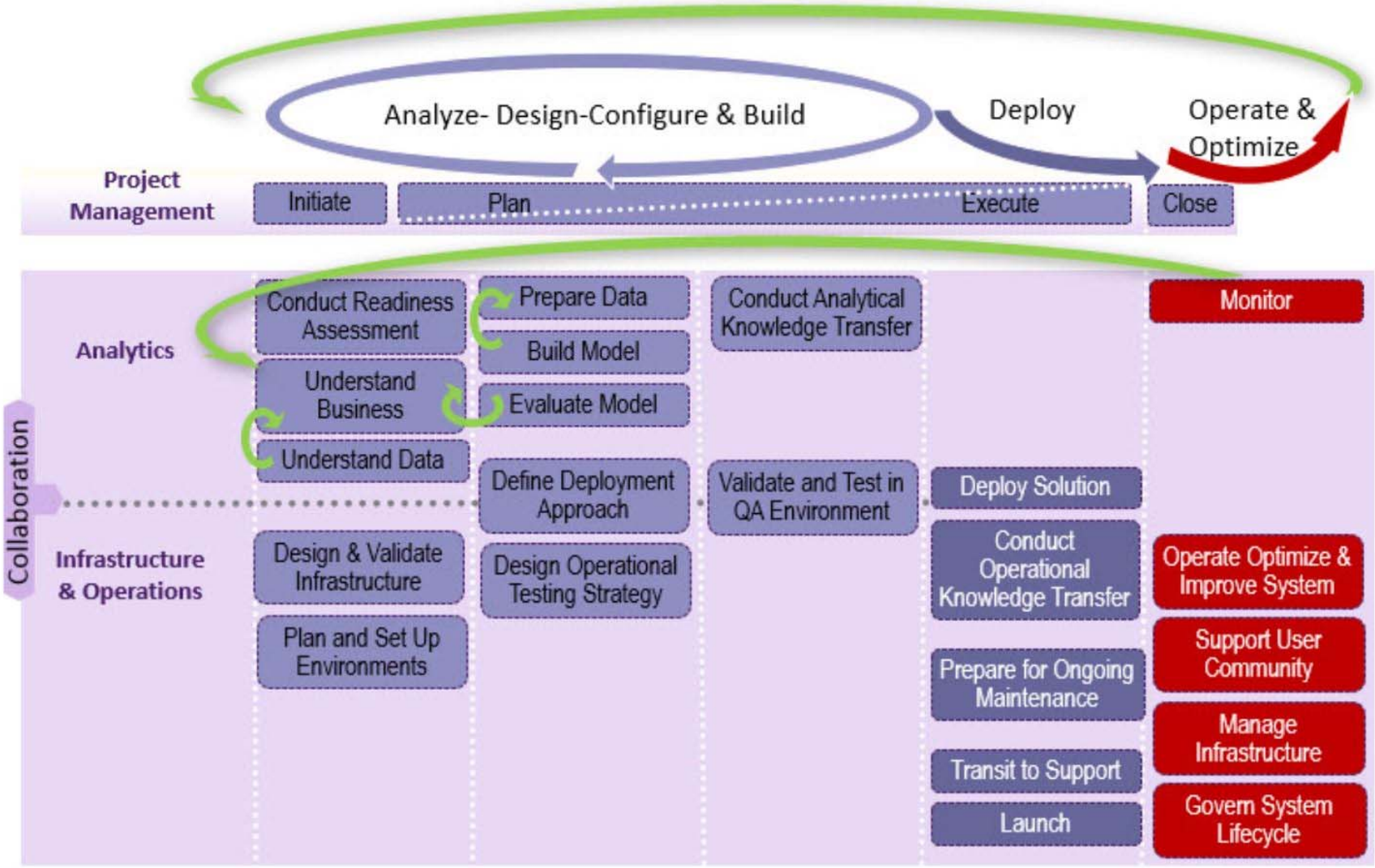
How to do Data Mining

ASUM-DM

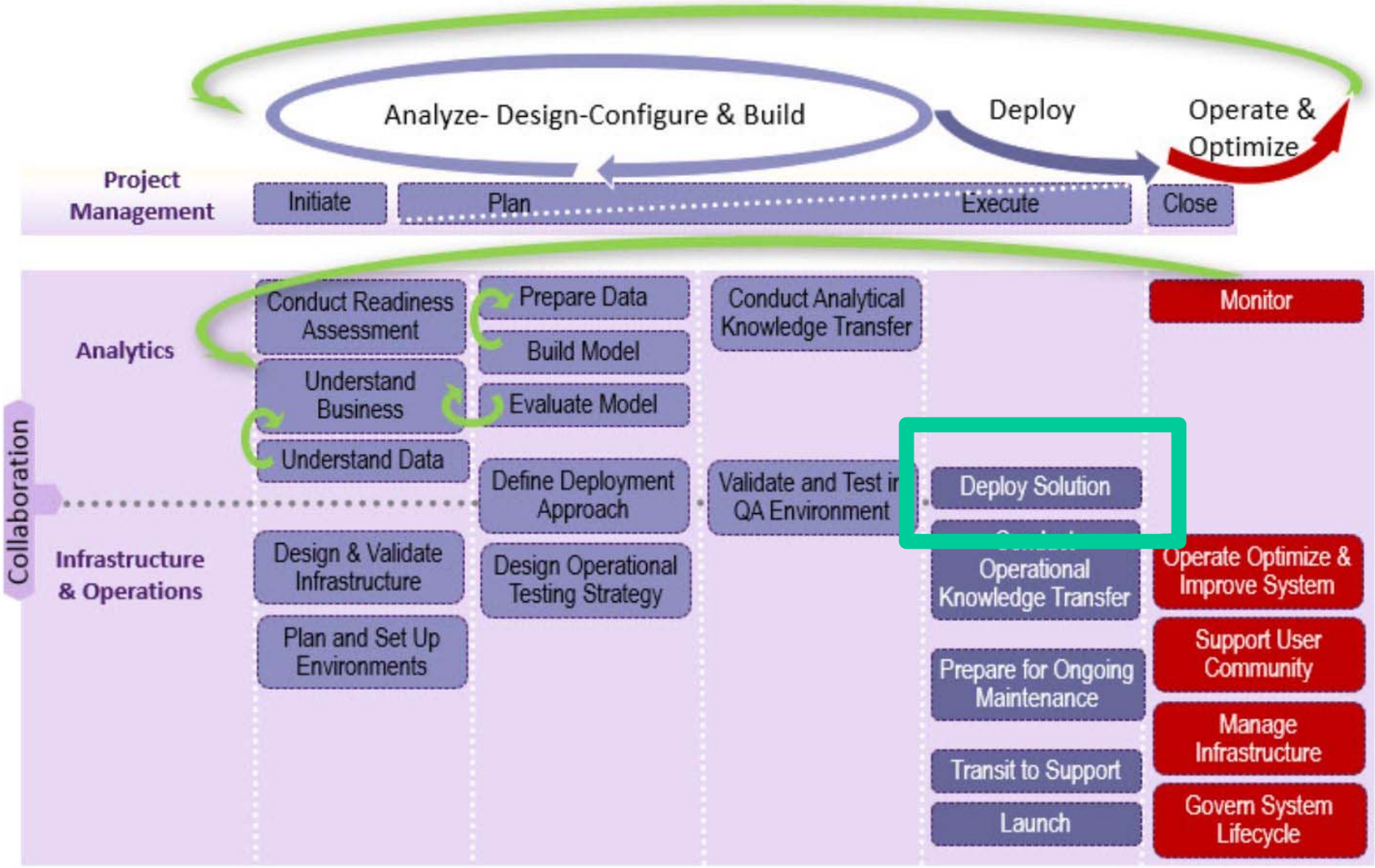
- Analytics Solutions Unified Method for Data Mining/Predictive Analytics
- Distributed after registration as EXE-file (!), installing a set of html pages
- Extension of CRISP-DM by IBM
  - Infrastructure / operations aspects
  - Project management
  - Deployment



BI Process Models

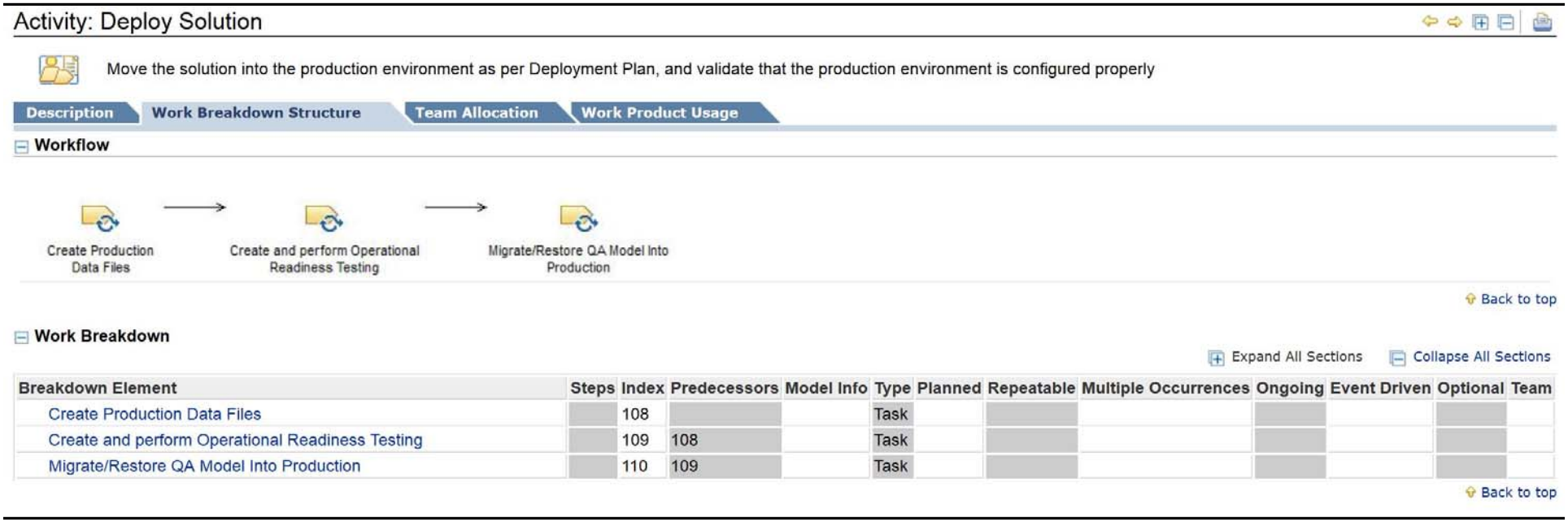


ASUM-DM


Analyze- Design-Configure & Build
Deploy
Operate & Optimize
Project Management
Initiate
Plan
Execute
Close
Analytics
Conduct Readiness Assessment
Prepare Data
Build Model
Understand Business
Evaluate Model
Understand Data
Conduct Analytical Knowledge Transfer
Monitor
Collaboration
Define Deployment Approach
Validate and Test in QA Environment
Deploy Solution
Infrastructure & Operations
Design & Validate Infrastructure
Design Operational Testing Strategy
Operational Knowledge Transfer
Operate Optimize & Improve System
Plan and Set Up Environments
Support User Community
Prepare for Ongoing Maintenance
Manage Infrastructure
Transit to Support
Govern System Lifecycle
Launch

ASUM-DM

Activity: *Deploy*: 3 Tasks



Activity: Deploy Solution

Move the solution into the production environment as per Deployment Plan, and validate that the production environment is configured properly

Description | Work Breakdown Structure | Team Allocation | Work Product Usage

Workflow

Create Production Data Files → Create and perform Operational Readiness Testing → Migrate/Restore QA Model Into Production

Back to top

Work Breakdown

Expand All Sections | Collapse All Sections

Breakdown Element	Steps	Index	Predecessors	Model Info	Type	Planned	Repeatable	Multiple Occurrences	Ongoing	Event Driven	Optional	Team
Create Production Data Files		108			Task							
Create and perform Operational Readiness Testing		109	108		Task							
Migrate/Restore QA Model Into Production		110	109		Task							

Back to top

ASUM-DM

Activity: *Deploy* – Task 1

Task: Create Production Data Files

Load all the data needed for the operation of the solution in production

Purpose

Load all the data needed for the operation of the solution in production

Back to top

Relationships

Roles	Primary Performer: • Client Key System Users • Data Miner/Data Scientist • Enterprise Architect	Additional Performers:
Process Usage	• ASUM-DM > Deploy > Deploy Solution > Create Production Data Files	

Back to top

ASUM-DM

Task: Create and perform Operational Readiness Testing

Create and execute the test to ensure that the production environment is ready to receive and handle the built solution

Purpose

- Create the test plan and verify that the solution is ready for use in production.
- Perform fixes and perform regression testing, rolling back if necessary
- Fine tune system as necessary (This might be required to accommodate any differences between the QA and production environments).

Back to top

Relationships

Roles	Primary Performer: • Enterprise Architect	Additional Performers: • Client Database Administrator • Client Network Administrator • Client Security Administrator • Client Tool Administrator
Process Usage	• ASUM-DM > Deploy > Deploy Solution > Create and perform Operational Readiness Testing	

Back to top

ASUM-DM

Activity: *Deploy* – Tasks 3

Task: Migrate/Restore QA Model Into Production

Migrate/Restore QA Model Into Production

Purpose

Migrate/Restore QA Model Into Production

Back to top

Relationships

Roles	Primary Performer: • Client Database Administrator • Client Security Administrator • Enterprise Architect	Additional Performers:
Process Usage	• ASUM-DM > Deploy > Deploy Solution > Migrate/Restore QA Model Into Production	

Back to top

ASUM-DM

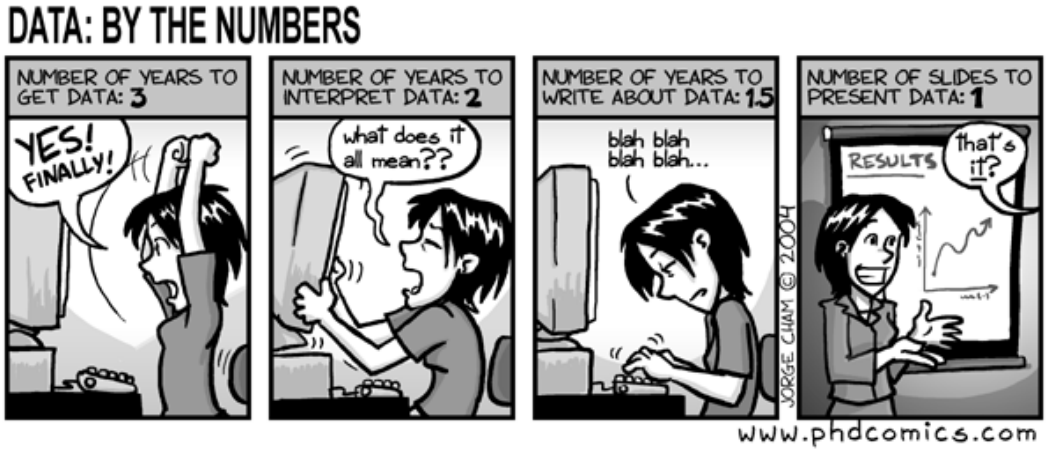
Deploy – Team Breakdown

Breakdown Element	Model Info	Team	Type
Client Database Administrator			Role
Migrate/Restore QA Model Into Production	Performs as Owner		Task
Create and perform Operational Readiness Testing	Performs as Additional		Task
Client Key System Users			Role
Create Production Data Files	Performs as Owner		Task
Client Network Administrator			Role
Create and perform Operational Readiness Testing	Performs as Additional		Task
Client Security Administrator			Role
Migrate/Restore QA Model Into Production	Performs as Owner		Task
Create and perform Operational Readiness Testing	Performs as Additional		Task
Client Tool Administrator			Role
Create and perform Operational Readiness Testing	Performs as Additional		Task
Data Miner/Data Scientist			Role
Create Production Data Files	Performs as Owner		Task
Enterprise Architect			Role
Create and perform Operational Readiness Testing	Performs as Owner		Task

BI Process Models

Summary

- A number of process models
- Focus expanding increasingly beyond core data mining
- …to business and data understanding
- … …to deployment and monitoring
- Most of the time spent in first two phases!



https://phdcomics.com/comics/archive.php?comicid=462, 31.5.2004

How to do Data Mining

- Data Analysis / Business Intelligence are both art and science
- Important to understand the goals (c.f. later when we talk about evaluation)
- Important to document process, be able to trace results, to analyse process → repeatability and verifiability (note: that's why simple approaches are so prominent)
- Important to know what you are doing (i.e. which tools you are using, how they behave, how to interpret results,…)

Outline

- How to do Data Mining
- Types of machine learning
- Data preprocessing: coding, scaling
- Summary

Definitions

- **AI**: Deals with intelligence of machines
  - *system that perceives its environment & takes actions which maximize its chances of success*
  - *science and engineering of making intelligent machines*

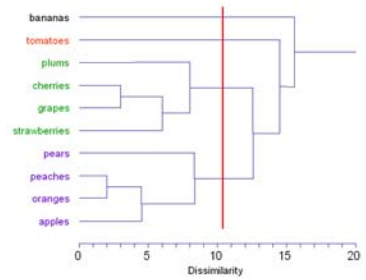
- Areas/problems of AI
  - Deduction, reasoning, problem solving
  - Knowledge representation (reasoning)
  - Planning / scheduling
  - Natural language processing
  - ***Machine learning***
  - …

Definitions

- **Concepts**: things that can be learned
  - E.g. list of topics for texts, spam/non-spam for email, groups of similar animals, sub-groups in a social network, correlation between smoking and lung cancer, …
- **Instance**: example of a concept, data point
  - E.g. individual text documents; animals; social network nodes; individual persons
- **Attribute**: measurement/description of an instance
  - E.g. text described by BOW using tfidf;
    animals described by characteristics such as #legs, fur/feathers, food;
    social network nodes represented by their connections to other nodes;
    people described by smoking habits and degree of cancer

Supervised vs. Unsupervised Learning

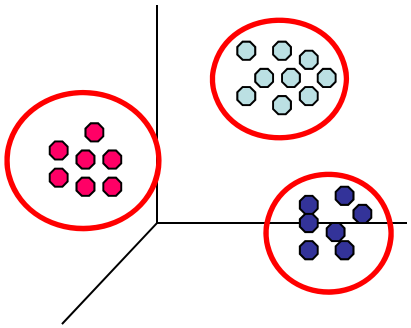
- **Unsupervised learning**
  - Data ***not*** **labelled**
    No information on which and how many classes or other structures
  - Goal:
    - find structures (e.g. clustering)
    - association rules learning
  - Most often associated with ***Data Mining***

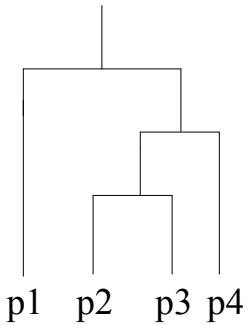


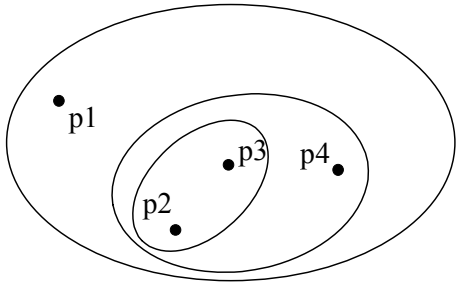
- **Supervised learning**
  - Data **labelled** with actual output variable
  - Regression and Classification
  - Goal: correctly label unknown data
  - Sometimes equivalently used with "**machine learning**"
    (but: ML is both unsupervised and supervised learning)

Unsupervised Learning

- Unsupervised learning
  - Data not ***labelled***
    No information on which and how many classes or other structures
- Clustering: find groups of data that belong together
  - K-means, tree-based algorithms
- Associations: find relationships between attributes in data
  - Association rule mining: find rules that show the relationship between certain attributes
  - Evaluated wrt. support and coverage of the rules

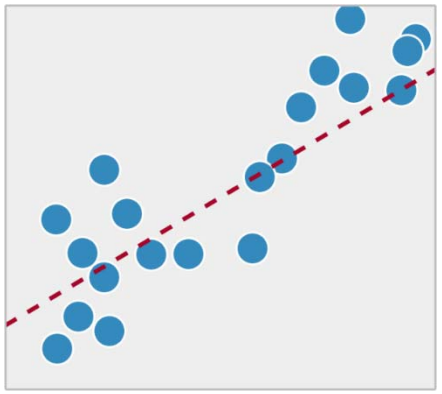






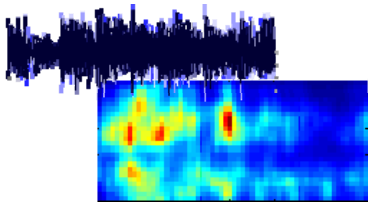
Supervised Learning: Regression

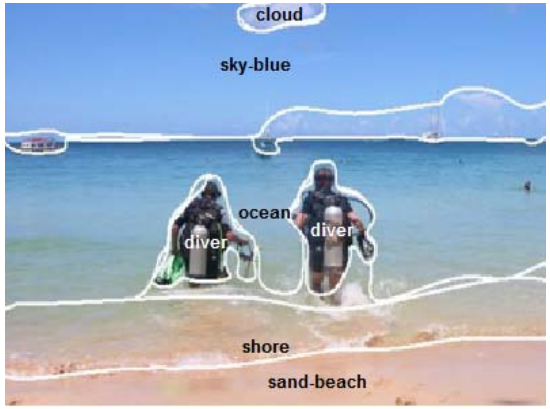
- **Regression** tries to predict a ***continuous*** variable
  - e.g. the temperature, depending on overcast, wind, humidity…,
  - Statistics, e.g. linear regression



Supervised Learning: Classification

- **Classification**: **discrete** output variable (pre-defined set of values) referred to as "**class**"





Supervised Learning: Classification

- **Classification**: **discrete** output variable (pre-defined set of values)
- Widely used in text mining
  - email SPAM filtering
  - Document routing, document classification
  - Document genre recognition
  - Image: classification of hand-written letters for OCR; automatic labelling of images
  - Music: classification of music into genres
  - Medicine: classification of whether a person has an illness, based on secondary features

Supervised Learning: Classification

- **Example:** Data set describing characteristics of humans
  - Gender
  - Age
  - Smoker yes/no
  - Eye colour

- Want to predict whether a person will get lung cancer
- Available: some data labelled
- (note: this example is not entirely correct from a medical perspective!!)

Supervised Learning: Classification

Attributes / Variables / Features

Output Variable

Training Samples

gender	age	smoker	eye colour	lung cancer
male	19	yes	green	No
female	44	yes	grey	yes
male	49	yes	blue	yes
male	12	no	brown	no
female	37	no	brown	no
female	60	no	brown	yes
male	44	no	blue	no
female	27	yes	brown	no
female	51	yes	green	yes
female	81	yes	grey	no
male	22	yes	brown	no
male	29	no	blue	no

Ground Truth / Gold standard

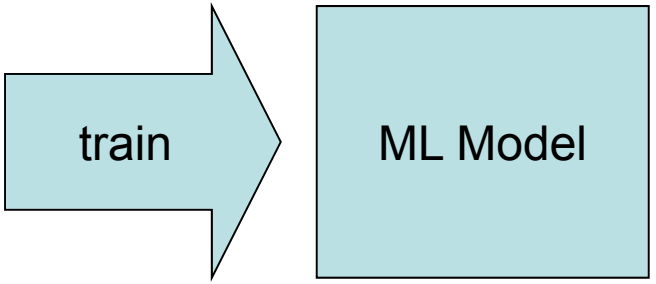
male	77	yes	grey	?
male	19	yes	green	?
female	44	no	grey	?

Unlabelled samples

Supervised Learning: Classification

gender	age	smoker	eye colour
male	19	yes	green
female	44	yes	grey
male	49	yes	blue
male	12	no	brown
female	37	no	brown
female	60	no	brown
male	44	no	blue
female	27	yes	brown
female	51	yes	green
female	81	yes	grey
male	22	yes	brown
male	29	no	blue

lung cancer
No
yes
yes
no
no
yes
no
no
yes
no
no
no



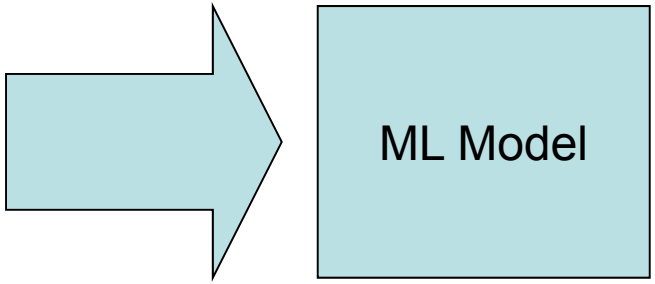
male	77	yes	grey
male	19	yes	green
female	44	no	grey

?
?
?

Supervised Learning: Classification

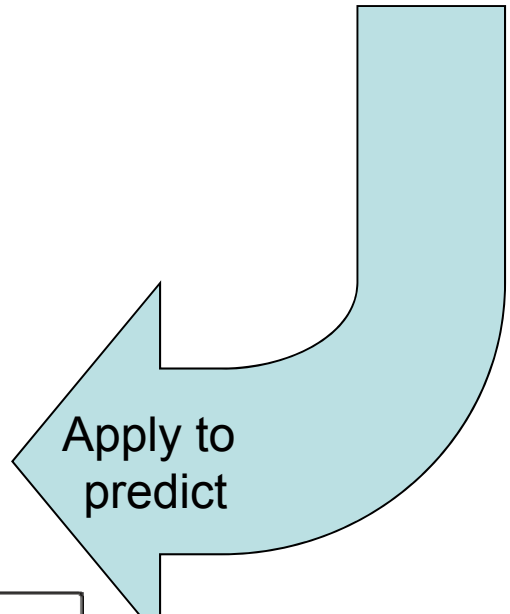
gender	age	smoker	eye colour
male	19	yes	green
female	44	yes	grey
male	49	yes	blue
male	12	no	brown
female	37	no	brown
female	60	no	brown
male	44	no	blue
female	27	yes	brown
female	51	yes	green
female	81	yes	grey
male	22	yes	brown
male	29	no	blue

lung cancer
No
yes
yes
no
no
yes
no
no
yes
no
no
no



male	77	yes	grey
male	19	yes	green
female	44	no	grey

yes
no
yes



Other Learning Models

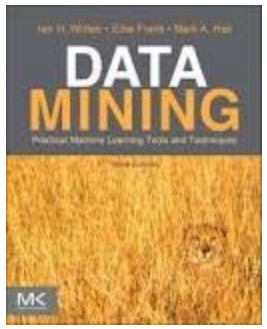
- Semi-supervised learning
  - Using unlabelled data together with labelled data for training
- Positive unary learning (PU)
  - Binary class setting (e.g. normal – (many) non-normal states)
  - Only labelled examples of one class provided
  - Unlabelled set of target class plus others
- Reinforcement Learning
  - Not explicitly presenting input/output pair
  - Rather reward/penalise agent for actions
- Zero-shot Learning
  - Learning a class for which there is no training data
  - Tries to identify intermediary concepts
  - Learns on concept-based description

Literature

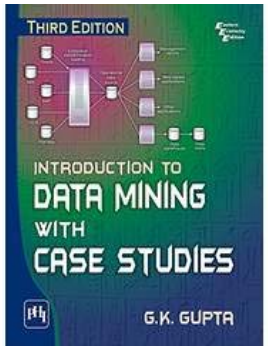
Massive number of books of varying quality

Recommended ones include:

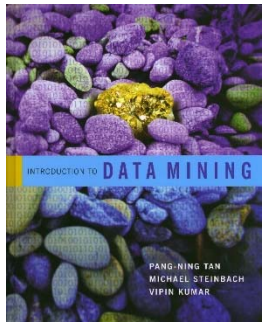
- I. Witten, E. Frank: Data Mining: Practical Machine Learning tools and applications
  - Companion book/software: WEKA
  - Slides: http://www.cs.waikato.ac.nz/ml/weka/book.html



- Gupta: Data Mining with Case Studies
  3rd ed., 2014
  http://www.csse.monash.edu.au/~gopal/Teaching/Datamining/index.html



- Pang-Ning Tan, Michael Steinbach, Vipin Kumar: Introduction to Data Mining.
  http://www-users.cs.umn.edu/~kumar/dmbook/index.php



Outline

- How to do Data Mining
- Types of machine learning
- Data preprocessing: coding, scaling
- Summary

Outline

- How to do Data Mining
- Types of machine learning
- Attribute types
- Data Pre-processing: coding, scaling, missing values
- Summary

Pre-processing

- Task Analysis
- Data Analysis and Cleansing
- Encoding
- Missing Values
- Scaling

- Simple – and tricky: the key to succss!
- Frequently most time-consuming task

Pre-processing

- Vital step for machine learning (supervised and unsupervised)

- ML algorithm will always give you a model
- Quality of that model depends highly on the quality of the input data
- “Garbage in” -> “Garbage out”

- One major goal of data preparation: Eliminate “wrong influence” of variables

Preprocessing

- Defining goal (sometimes not trivial!)
- Selecting which data to use
  - internal data
  - external data! (lots available? which is most useful?)
- Deciding about suitable algorithms
- Transforming and pre-processing:
  - Understanding the data: value ranges, sparsity, missing values, dependency analysis / correlations, …
  - cleansing, missing value handling, transcoding, normalization, outlier detection, …
  - critical success factor
  - garbage-in, garbage-out

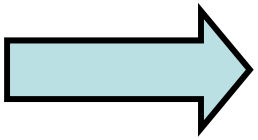
Preprocessing: Coding

- Nominal/ordinal data
- E.g. eye color „gray", "blue", "green„, brown“
- Some ML algorithms can only handle numeric variables (e.g. distance-based algorithms in vector space)
- Solution:
  - Nominal: 1-to-N coding
    (will create attributes with dependencies)
  - Ordinal:
    - 1-to-N coding: looses ordering
    - Transforming to interval/ratio quantity

Preprocessing: Coding

- 1-to-N coding
- Introduces dependencies, increases dimensionality, sparsity

colour
brown
blue
green
grey
brown
green
blue



green	*blue*	*brown*	*grey*
0	0	1	0
0	1	0	0
1	0	0	0
0	0	0	1
0	0	1	0
1	0	0	0
0	1	0	0

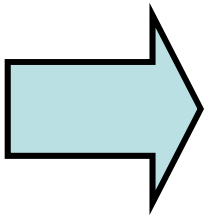
Preprocessing: Coding

- Animal data set
  - Describes animal by some characteristics
  - Instances: cow, horse, duck, eagle, …
- Variables
  - Size
    - tiny, small, medium, large
  - Number of legs
    - 2, 4, 6, 8

	size	*Legs*
bird	small	2
cat	small	4
spider	tiny	8
dog	med	4
cow	large	4
bee	tiny	6
monkey	med	2

Preprocessing: Coding

	size	*Legs*
bird	small	2
cat	small	4
spider	tiny	8
dog	med	4
cow	large	4
bee	tiny	6
monkey	med	2



tiny	*small*	*med*	*large*	*legs*
0	1	0	0	2
0	1	0	0	0
1	0	0	0	8
0	0	1	0	4
0	0	0	1	4
1	0	0	0	6
0	0	1	0	2

Preprocessing: Coding

- Any other pre-processing needed?

	size	*Legs*
bird	small	2
cat	small	4
spider	tiny	8
dog	med	4
cow	large	4
bee	tiny	6
monkey	med	2

Preprocessing: Coding

- Any other pre-processing needed?
- Variable “legs”
  - If considered categorical: defined order ➔ ordinal data
  - Can compute similarity: 2 closer to 4 than to 6
  - Numerical value, can compute distance
  - *Does the number of legs denote similarity?*

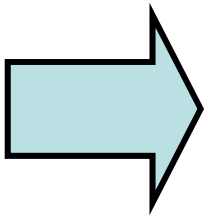
	size	*Legs*
bird	small	2
cat	small	4
spider	tiny	8
dog	med	4
cow	large	4
bee	tiny	6
monkey	med	2

Preprocessing: Coding

- *Does number of legs denote similarity?*
  i.e.,
  - is an animal with 2 legs more similar to one with 4, or with 6?
  - is one with 4 equally similar to the one with 2 and 6?
    - Dog to monkey vs. dog to spider
  - One with 6 equally similar to one with 4 and 8?
    - Bee to spider vs. bee to cow

Preprocessing: Coding

	size	*Legs*
bird	small	2
cat	small	4
spider	tiny	8
dog	med	2
cow	large	4
bee	tiny	6
monkey	med	2

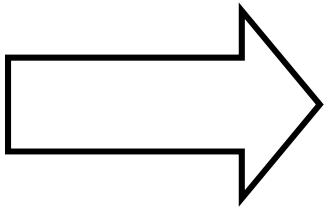


2 legs	*4 legs*	*6 legs*	*8 legs*
1	0	0	0
0	1	0	0
0	0	0	1
1	0	0	0
0	1	0	0
0	0	1	0
1	0	0	0

Preprocessing: Coding

- Transforming to interval quantities
- Careful selection of relative values

Feature
Win
Draw
Win
Loose
Loose
Win
Draw

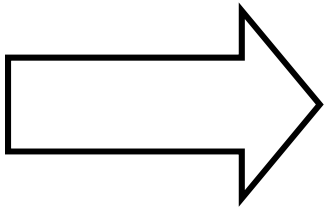


Feature
2
1
2
0
0
2
1

Preprocessing: Coding

- Transforming to interval quantities
- Careful selection of relative values

Feature
Win
Draw
Win
Loose
Loose
Win
Draw



Feature
2
1
2
0
0
2
1

Feature
3
1
3
0
0
3
1

Preprocessing: Distances

- **Calculating distances:**
  What to do with categorical data?

Preprocessing: Distances

- **Calculating distances:**
  What to do with categorical data?

- 1-n coding – then apply any distance function

- Definition of custom distance functions

Preprocessing: Distances

- Definition of custom distance functions
  - Adapt hamming distance

    Hamming distance between two strings of equal length is the number of positions at which the corresponding symbols are different

    ➔ count number of different nominal values

  - Define distance for each attribute, aggregate e.g. via sum

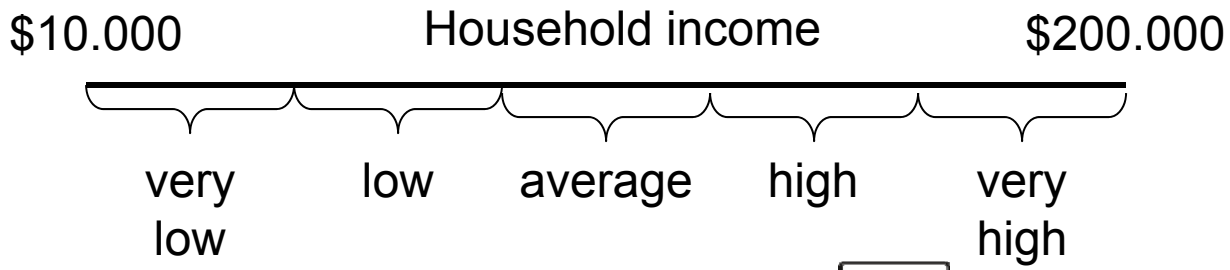
    e.g. implicit transformation to interval quantities

Preprocessing: Coding

Classification with continuous output attribute?

- Classification requires categorical output (continuous output = regression)
- Classification methods can be applied by **binning** (aka „bucketing“) continuous output (loss of prediction accuracy)
- Solution: sub-division into discrete bins (ordinal data)

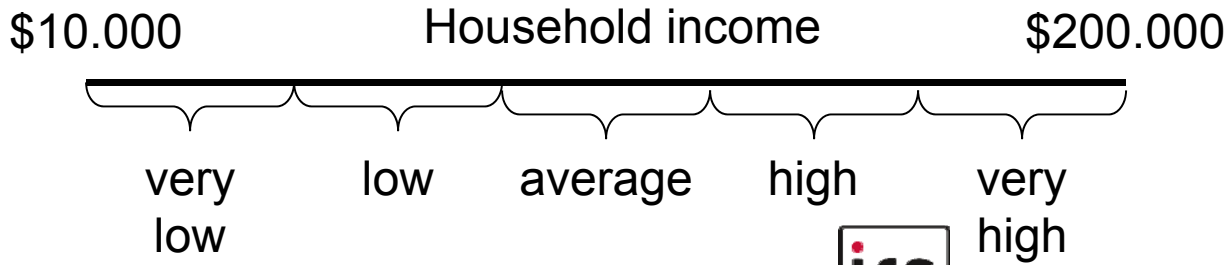
- How can we do this?



Preprocessing: Coding

Binning (aka „Bucketing“) – How?

- What is the optimal binning method?
  - evenly spaced
  - „natural“ boundaries (e.g. age groups)
  - density based
- When to use which? How do you decide?
- How many bins should we use?
  - More bins: finer granularity,
  - But also: more difficult to learn (more classes) fewer datapoints per bin for training



Preprocessing: Missing Values

Dealing with missing values

- How are they encoded?
- Why are they missing?
  - not considered important for purpose at data collection
  - not available
  - errors when reading data
  - systematic errors?!
- How to deal with them?

Preprocessing: Missing Values

Dealing with missing values

- How to deal with them?
  - Delete instance
  - Ignore in calculation
  - Data Imputation: substitute value
    - Mean value of the attribute (computed from other samples)
    - Random selection of value from another (similar?) sample
    - Regression – using other attributes to predict
    - Clustering – values of cluster centroid
    - Nearest Neighbour – value of closest sample
- Careful when making decisions!

(cf. eg. David Howell: Treatment of Missing Data, http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html)

Preprocessing: Missing Values

Scaling data

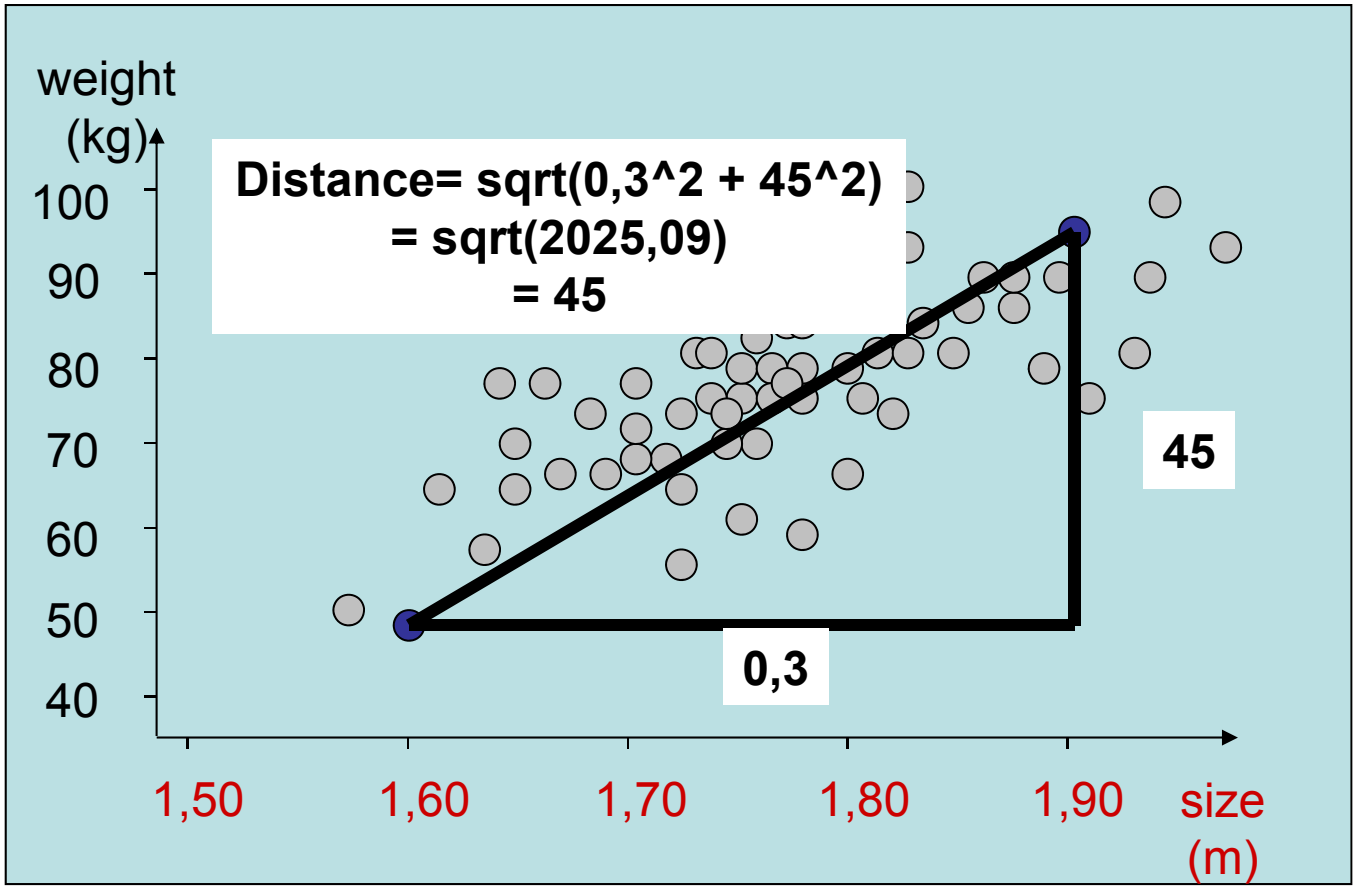
- Different variables may exhibit vastly different value ranges
  - E.g. a length variable measured in cm, inch, or meters
  - Different types of measurements: length, speed, temperature, …
  - Different types of measuring devices capturing different value ranges
  - …
- Why is this a (potential) problem?

Preprocessing: Scaling

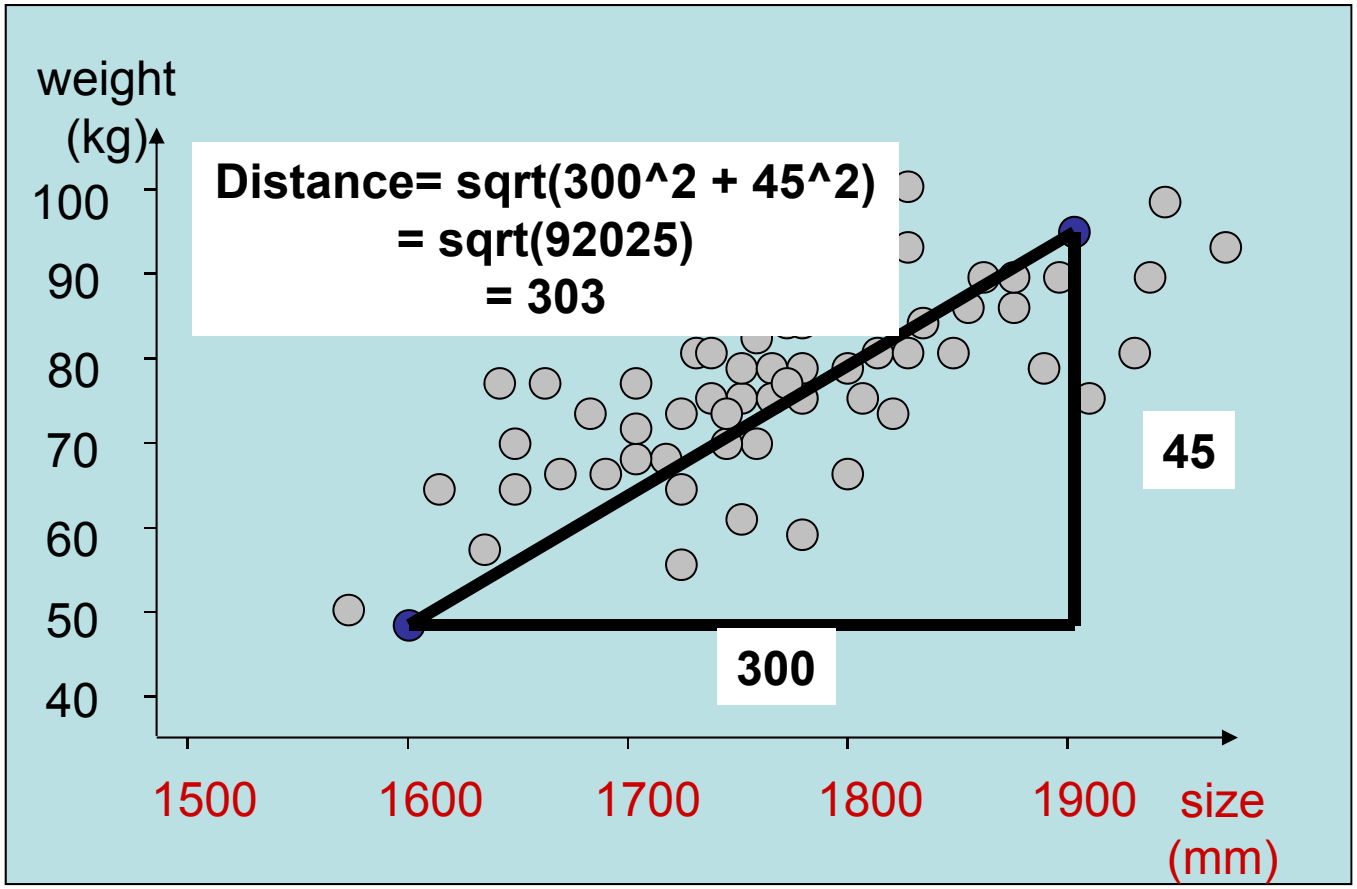
Scaling data

- Some ML algorithms rely on measuring the (numeric) distance between samples
- The value range (scale) should have no impact
- Higher values in one attribute have unproportionally large effect on measure distance
  -> dominate distance metric
  -> might thus dominate learning

Preprocessing: Scaling



Preprocessing: Scaling



Preprocessing: Scaling

- Measuring distance should be independent of measurement unit
- Standardizing attribute values: z-score (zero-mean-unit-variance):
  - subtract mean
  - divide by standard deviation

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

Preprocessing: Scaling

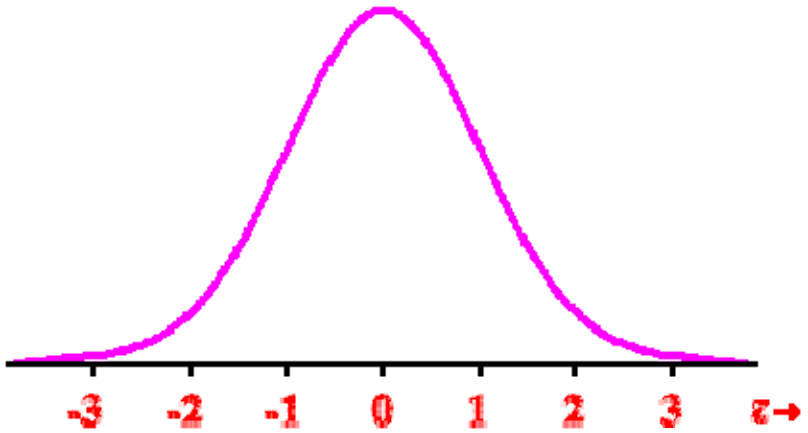
- z-score (zero-mean-unit-variance):
  - What is the value range after applying z-score?

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

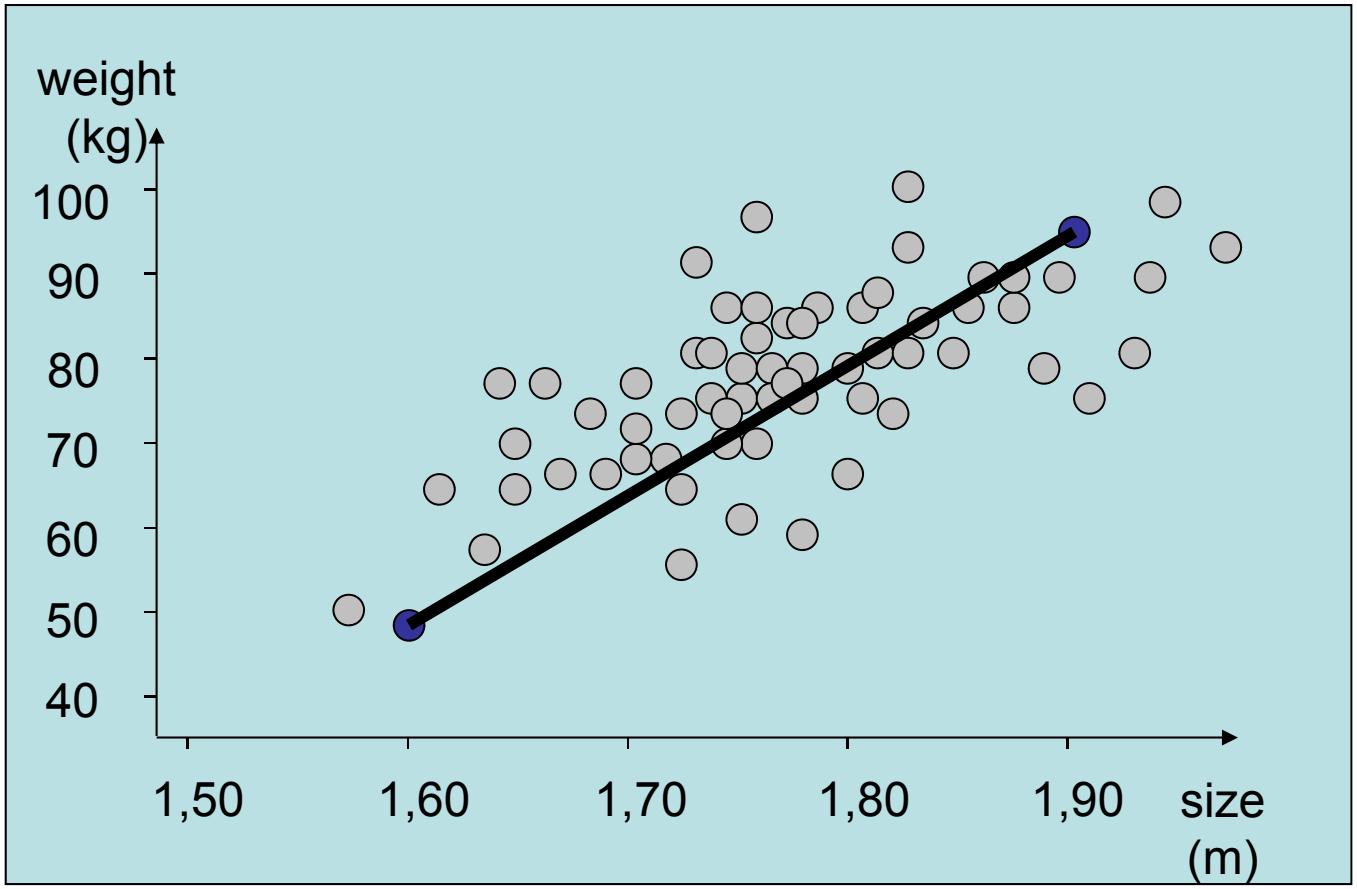
Preprocessing: Scaling

- z-score (zero-mean-unit-variance):
  - What is the value range after applying z-score?

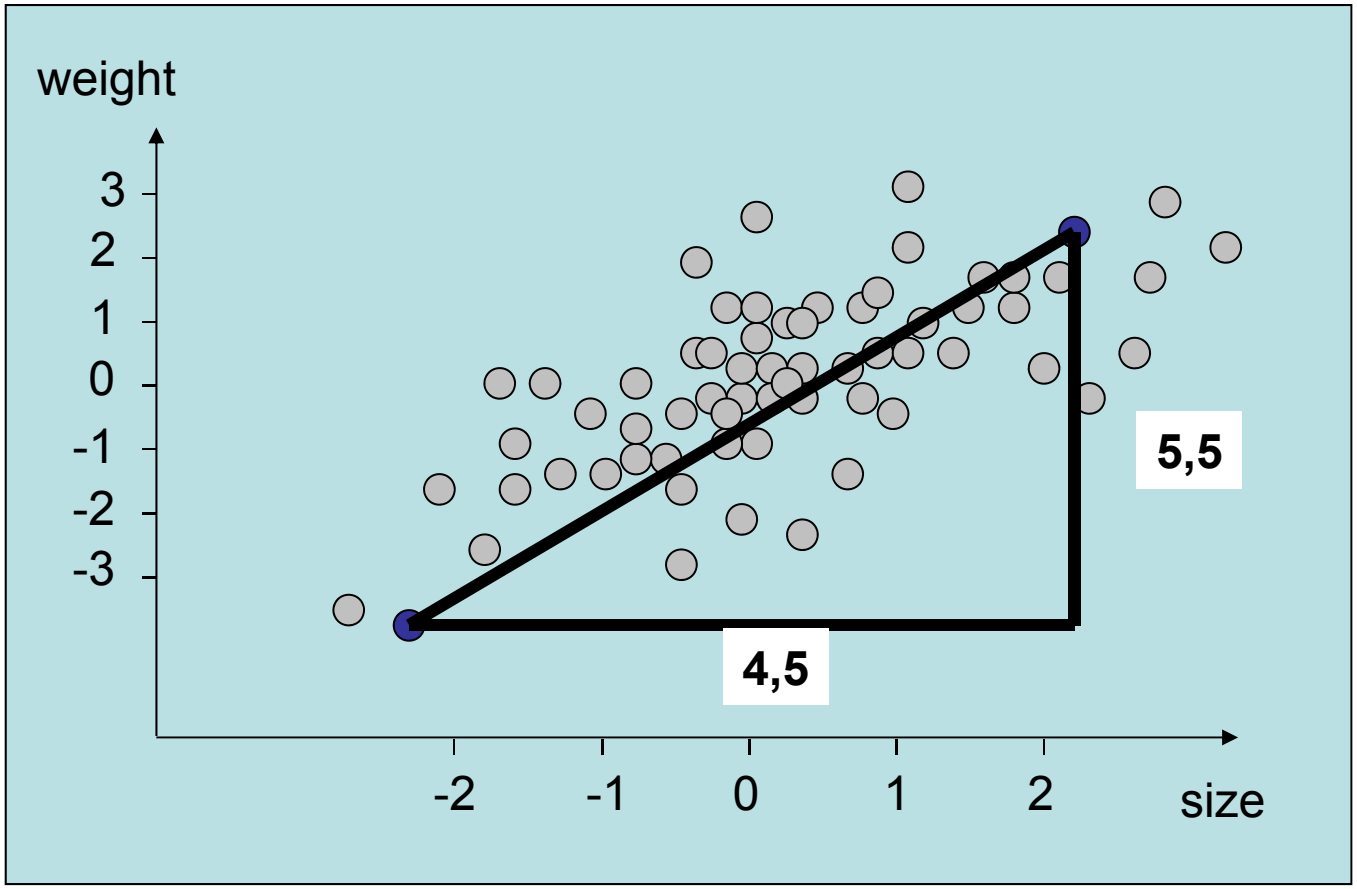
$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$



Preprocessing: Scaling



Preprocessing: Scaling



Preprocessing: Scaling

- Data set now has mean 0, variance 1
- Chebyshev's inequality:
  - 75% of data between -2 and +2
  - 89% of data between -3 and +3
  - 94% of data between -4 and +4
- Other forms of scaling
  - min/max
  - unit length

Preprocessing: Scaling

- Min/Max Scaling
  - Scale all variables to the same (fixed) range
  - Often between 0 and 1
  - Subtract minimum value for each variable
  - Divide by value range of each variable
  - Multiply by new range (if different than 0..1)

$$z_i = \frac{x_i - \min(X)}{\max(X) - \min(X)}$$

Preprocessing: Scaling

Unit Length Scaling

- Example: Features (variables) from text documents (BOW)
  - Each word (term) = one variable
  - Values = count of words in a documents (or tfidf)

	Word 1	Word 2	Word 3	Word 4	Word 5	...	Word n	∑
Doc 1	10	0	0	0	0		0	10
Doc 2	2	0	0	0	2		2	6
Doc 3	1	3	0	0	1		5	10
Doc 4	2	0	0	2	0		2	6
Doc 5	5	2	0	0	1		0	8
...							0	
Doc m	10	4	0	0	2		0	16

Preprocessing: Scaling

Unit Length Scaling

- Text features:
  - Length (size) of the object described influence the values
  - Longer documents -> in general higher values
  - Relative importance of words within a text not higher
  - Two documents with same content but different length -> very different vectors, high distance in-between
  - Normalise data vectors to the same length
  - Divide each attribute by the vector length

$$z_i = \frac{x_i}{\| x \|}$$

Preprocessing: Scaling

- Different forms of scaling
  - When are these useful?
  - When should you NOT use min/max, but rather zero-mean-unit-var?
  - When do they make a difference?
- Do I always need to perform scaling?

Preprocessing: Scaling

- Algorithms relying on distances
  - k-nearest Neighbours
  - …
- Not needed for algorithms that do not use distances, e.g.
  - Naïve Bayes
  - Decision trees
  - …
- Caveat:
  - Many implementations already do this pre-processing implicitly (e.g. WEKA)
  - Check default settings carefully

Preprocessing: Sparsity

- Sparsity: fraction of „zero“ values in vectorial data space
- Challenge in Text Mining!
- Has impact on algorithms being used
- Not all algorithms can deal with sparsity very well
- Some distance measures don‘t work well with sparse data
- Solution: random mapping
  - Producing a random matrix
  - Multiplying each attribute vector with random matrix
- Looses semantics of individual attributes

(Samuel Kaski: Dimensionality reduction by random mapping: fast similarity computation for clustering. Proceedings of The 1998 IEEE International Joint Conference on Neural Networks, 1998. pp. 413–418. doi: 10.1109/IJCNN.1998.682302)

Preprocessing: Sparsity

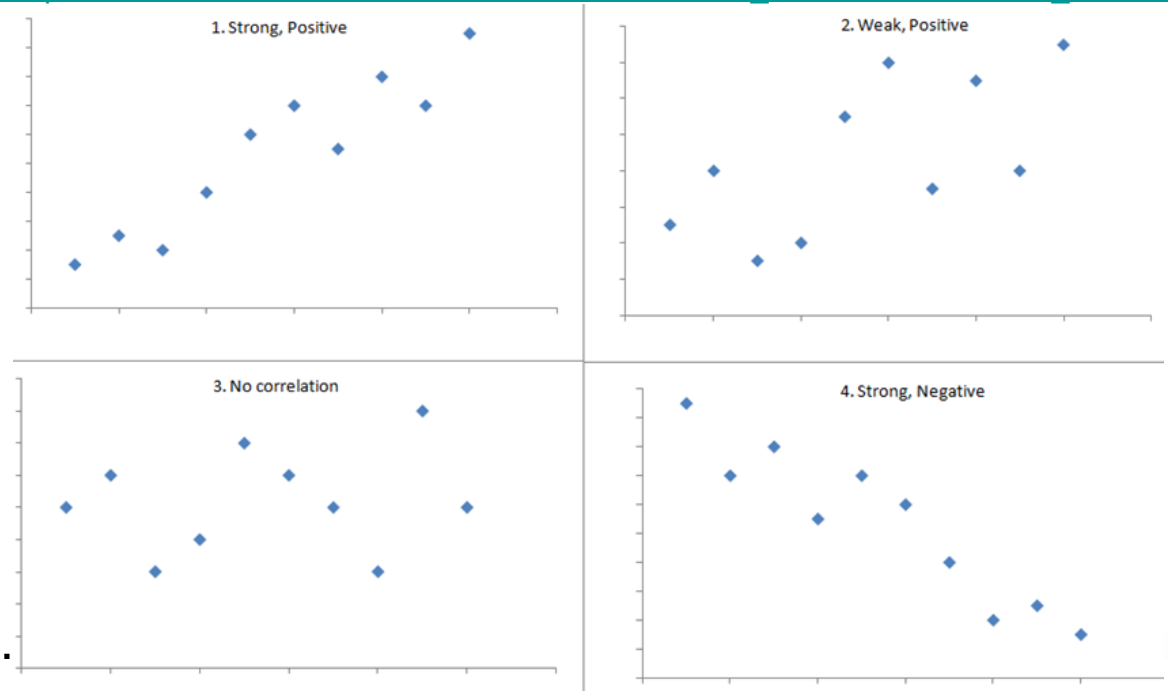
- Random Mapping:
  - Samuel Kaski: Dimensionality reduction by random mapping: fast similarity computation for clustering. Proceedings of the 1998 IEEE International Joint Conference on Neural Networks, 1998. pp. 413–418. doi: 10.1109/IJCNN.1998.682302)
  - Dmitriy Fradkin, David Madigan: Experiments with random projections for machine learning. In: KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, 517-522, 2003.

Preprocessing: Correlations

Correlation analysis

- Data set might contain input variables that directly depend on each other
  - Might have unproportional weight on output prediction
  - Might be better to eliminate such variables
- (pair wise) analysis of correlation

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable5.html



Preprocessing – New Attributes

- Adding new attributes
  - Combine existing attributes:
    area = width x length
  - Derive relative attributes:
    age = current_date – birth_date, windowing over streams,
    days_since_maintenance, number_items_produced_since_restart, …
  - *Group by semantics:*
    day_of_week, working_day, seasons, …
    re-define regions (beyond country/province/zip)
  - Make hierarchical structures explicit
    product loines, geographic, time, …
- Note: only derive attributes that will be available in-operation!
  - Time-to-failure? #re-tweets? %completed? %error_in_batch?
  - Consider time of availability of external data (weather – real-time?)

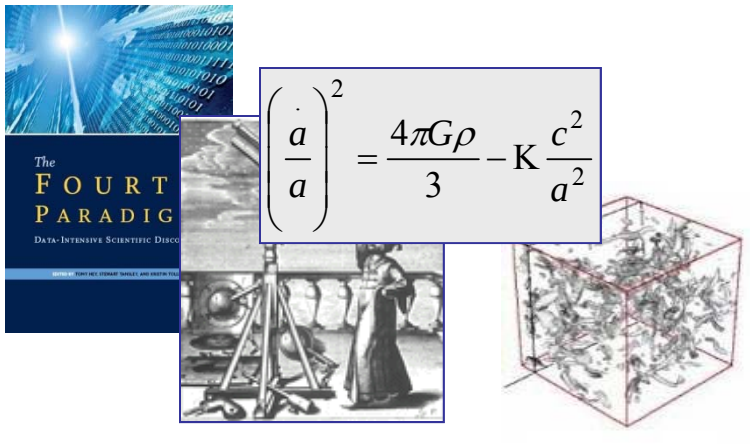
Summary

- Preprocessing is the **most important step** in DM/ML
- Determines performance that can be reached
- Can lead to wrong/spurious patterns being discovered
- Takes a lot of time
- Sometimes: iterative approach:
  pre-process, analyze, find errors, re-pre-process
- Look at data! Understand the data!
- Print graphs of data, analyze value ranges, min/max, histograms, … and document results!
- May involve other DM techniques:
  clustering to understand data
- Document the pre-processing applied!
- **So trivial… and so often forgotten/done wrongly…**

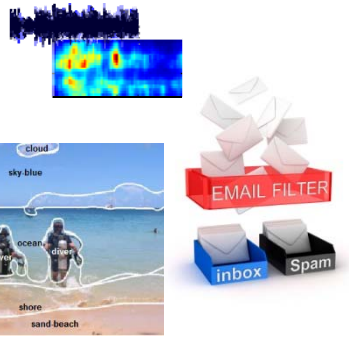
Outline

- How to do Data Mining (and: why?)
- Types of machine learning
- Attribute types
- Data Pre-processing: coding, scaling, missing values
- Summary

Thank you!



$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$



red	blue	green
1	0	0
0	1	0
0	0	1
1	0	0
1	0	0
0	0	1
0	1	0

http://www.ifs.tuwien.ac.at/imp