Hochschule Düsseldorf
University of Applied Sciences

Fachbereich Wirtschaftswissenschaften
Faculty of Business Studies

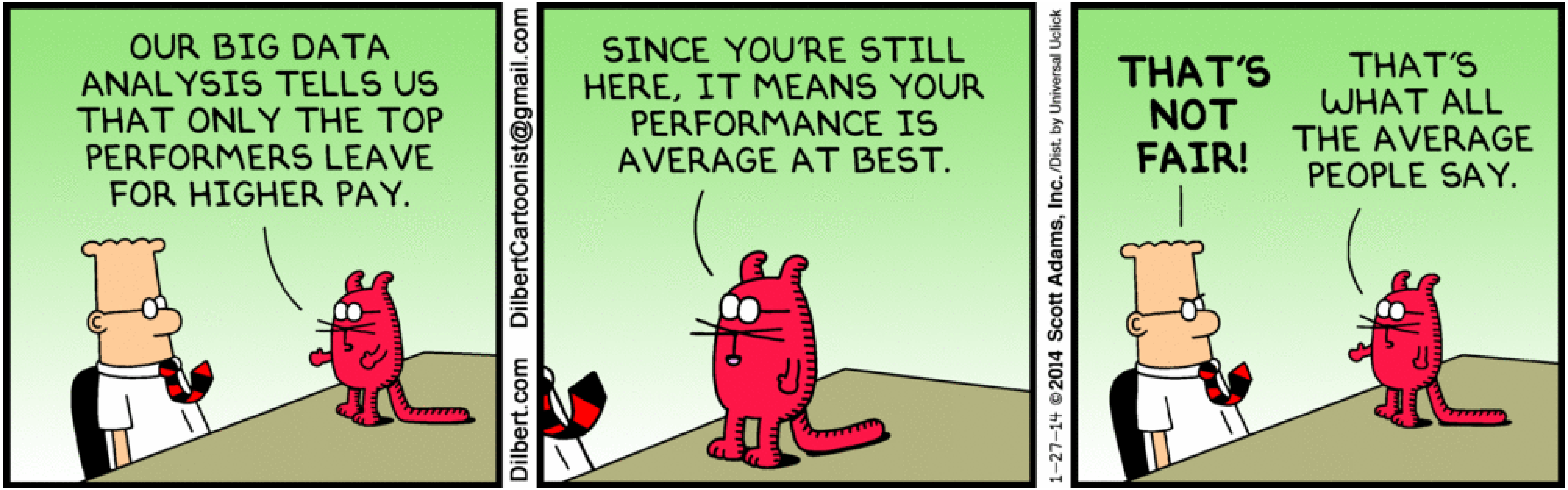
Business Analytics (M.Sc.)
IT in Business Analytics

IT APPLICATIONS IN BUSINESS ANALYTICS

SS2016 / Lecture 02 – CRISP DM

Thomas Zeutschler

Let's get started…



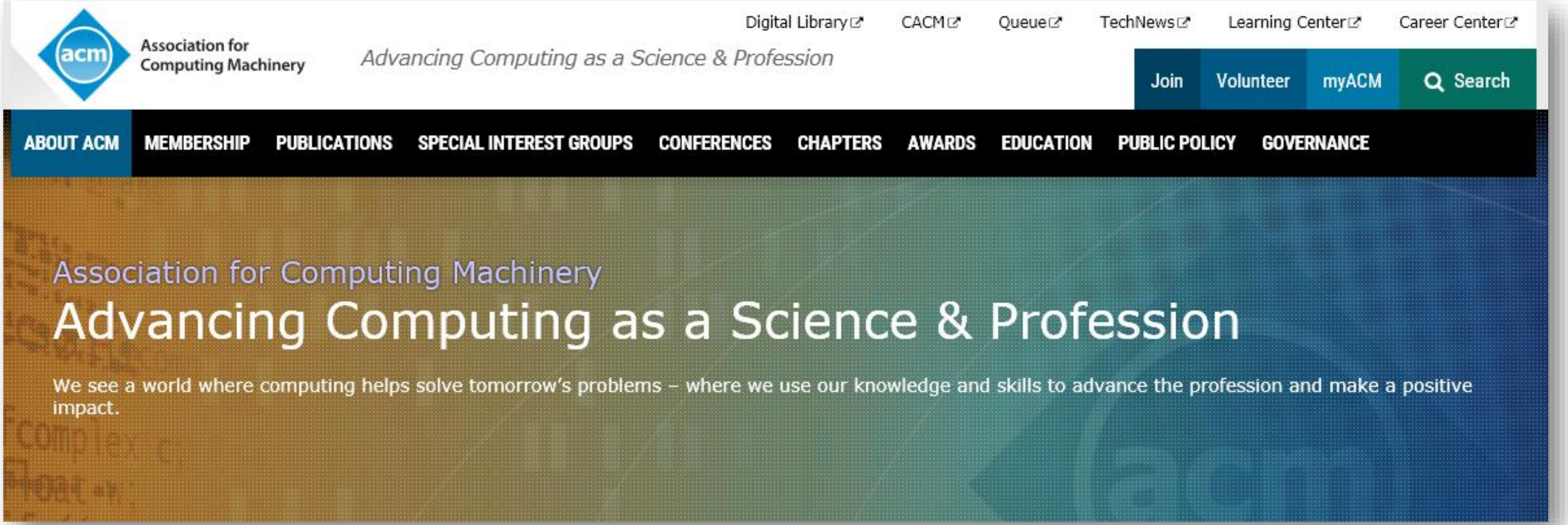
Data Mining

Data Mining

*“**Data Mining** is an **interdisciplinary** subfield of computer science.
It is the **computational process** of **discovering patterns** in large data sets involving methods at the intersection of **artificial intelligence**, **machine learning**, **statistics**, and **database systems**.”*
Source: Wikipedia “Data Mining”

“The core endeavor in data mining is to extract knowledge from data; this knowledge is captured in a human-understandable structure.”
Source: Data Mining Curriculum, ACM, 2006

Data Mining is about Computing



http://www.acm.org

Data Mining – Steps, Challenges and Issues

Data Mining Curriculum: A Proposal (Version 1.0)

Intensive Working Group of ACM SIGKDD Curriculum Committee:
Soumen Chakrabarti, Martin Ester, Usama Fayyad, Johannes Gehrke,
Jiawei Han, Shinichi Morishita, Gregory Piatetsky-Shapiro, Wei Wang

April 30, 2006

1 Introduction

Recent tremendous technical advances in processing power, storage capacity, and inter-connectivity of computer technology is creating unprecedented quantities of digital data. *Data mining*, the science of extracting useful knowledge from such huge data repositories, has emerged as a young and interdisciplinary field in computer science. Data mining techniques have been widely applied to problems in industry, science, en-

http://www.kdd.org/exploration_files/CURMay06.pdf

Data Mining – Steps, Challenges and Issues

1. Database and Data Management Issues

- Where does the data reside? How is it to be accessed?
- What forms of sampling are needed? are possible? are appropriate?
- What are the implications of the database or data warehouse structure and constraints on data movement and data preparation?

Data Mining – Steps, Challenges and Issues

2. Data Preprocessing

- What are the required data transformations before a chosen algorithm or class of algorithms can be applied to the data?
- What are effective methods for reducing the dimensionality of the data so the algorithms can work efficiently?
- How are missing data items to be modelled?
- What transformations properly encode a priori knowledge of the problem?

Data Mining – Steps, Challenges and Issues

3. Choice of Model and Statistical Inference Considerations

- What are the appropriate choices to ensure proper statistical inference*?
- What are valid approximations?
- What are the implications of the inference methods on the expected results?
- How is the resulting structure to be evaluated and validated?

*__Statistical Inference__ is the process of deducing properties of an underlying distribution by analysis of data

Data Mining – Steps, Challenges and Issues

4. Interestingness Metrics

- What makes the derived structure interesting or useful?
- How do the goals of the particular data mining activity influence the choice of algorithms or techniques to be used?

Data Mining – Steps, Challenges and Issues

5. Algorithmic Complexity Considerations

- What choice of algorithms based on the size and dimensionality of data?
- What about computational resource constraints?
- Requirements on accuracy of resulting models?
- What are the scalability considerations and how should they be addressed?

Data Mining – Steps, Challenges and Issues

6. Post-processing of Discovered Structure

- How are the results to be used?
- What are the requirements for use at prediction time?
- What are the transformation requirements at model application time?
- How are changes in the data or underlying distributions to be managed?

Data Mining – Steps, Challenges and Issues

7. Visualization and Understandability

- What are the constraints on the discovered structure from the perspective of understandability by humans?
- What are effective visualization techniques for the resulting structure?
- How can data be effectively visualized in the context of or with the aid of the discovered structures?

8. Maintenance, Updates, and Model Life Cycle Considerations

- When are models to be changed or updated?
- How must the models change as the utility metrics in the application domain change?
- How are the resulting predictions or discovered structure integrated with application domain metrics and constraints?

CRISP DM

The Data Mining Process

CRoss-InduStry Process for Data Mining

- A **methodology** covering the typical phases of an analytical project, the tasks involved with each phase, and an explanation of the relationships between these tasks.
- A **process model**, as CRISP-DM provides an overview of the data mining life cycle.

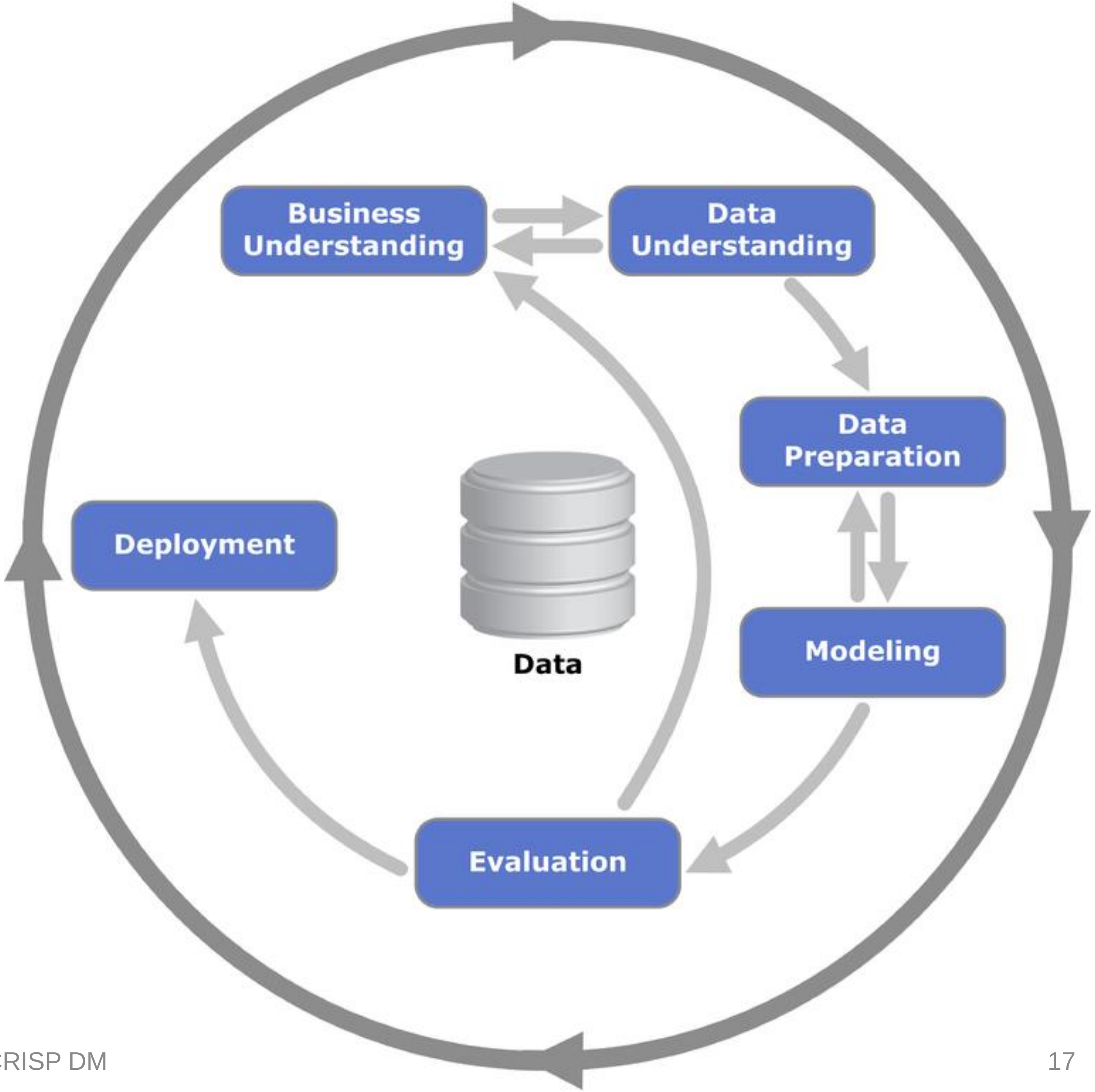
CRISP-DM was conceived in 1996 and first published in 1999 by SPSS, NCR and Mercedes and is reported as the **leading methodology for data mining/predictive analytics projects**.

IBM has released a new implementation method for Data Mining/Predictive Analytics projects in 2015 called **Analytics Solutions Unified Method for Data Mining & Predictive Analytics** (ASUM-DM) which is a refined and extended CRISP-DM. *But it's a little bit too complex start with…*

Introduction

„The process of knowledge discovery in data mining has to be reproducible and reliable.

Especially for people who have no background in data science.“



CRISP DM

CRoss-InduStry Process for Data Mining

- A **methodology** covering the typical phases of an analytical project, the tasks involved with each phase, and an explanation of the relationships between these tasks.
- A **process model**, as CRISP-DM provides an overview of the data mining life cycle.

CRISP-DM was conceived in 1996 and first published in 1999 by SPSS, NCR and Mercedes and is reported as the **leading methodology for data mining/predictive analytics projects**.

IBM has released a new implementation method for Data Mining/Predictive Analytics projects in 2015 called **Analytics Solutions Unified Method for Data Mining & Predictive Analytics** (ASUM-DM) which is a refined and extended CRISP-DM. *But it's a little bit too complex start with…*

CRISP DM – Current Industry Standard

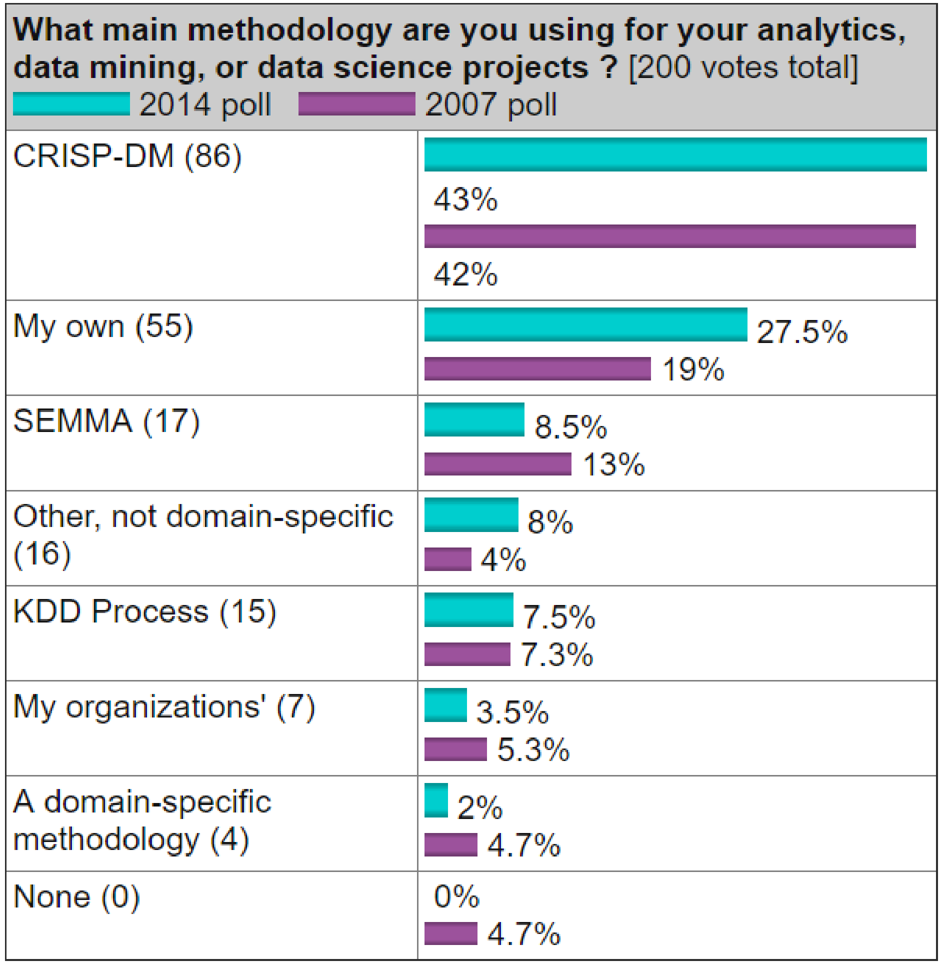
Other approaches:

KDD

- „**K**nowledge **D**iscovery in **D**atabases“ developed by Usama Fayyad (Microsoft Research, 1996) describes methods and technologies to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data.

SEMMA

- SEMMA is an acronym that stands for **S**ample, **E**xplore, **M**odify, **M**odel and **A**ssess. It is a list of sequential steps developed by SAS Institute in 2009.
- **Criticism:** SEMMA mainly focuses on the modeling tasks of data mining projects, leaving the business aspects out. Focussed on the usage of SAS products.



Source: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

CRISP DM – Objectives and Benefits

- Ensure quality of knowledge discovery project results
- Reduce skills required for knowledge discovery
- Reduce costs and time

- General purpose (i.e., stable across varying applications)
- Robust (i.e., insensitive to changes in the environment)

- Tool and technique independent
- Tool supportable

- Support documentation of projects
- Capture experience for reuse
- Support knowledge transfer and training

CRISP DM – Phases and Tasks

Business Understanding

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.

Data Understanding

Collect Initial Data
Initial Data Collection Report.

Describe Data
Data Description Report.

Explore Data
Data Exploration Report.

Verify Data Quality
Data Quality Report.

Data Preparation

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.

Modelling

Select Modelling Technique
Modelling Technique.
Modelling Assumptions.

Generate Test Design
Test Design.

Build Model
Parameter Settings
Models.
Model Description.

Assess Model
Model Assessment.
Revised Parameter Settings.

Evaluation

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.

Deployment

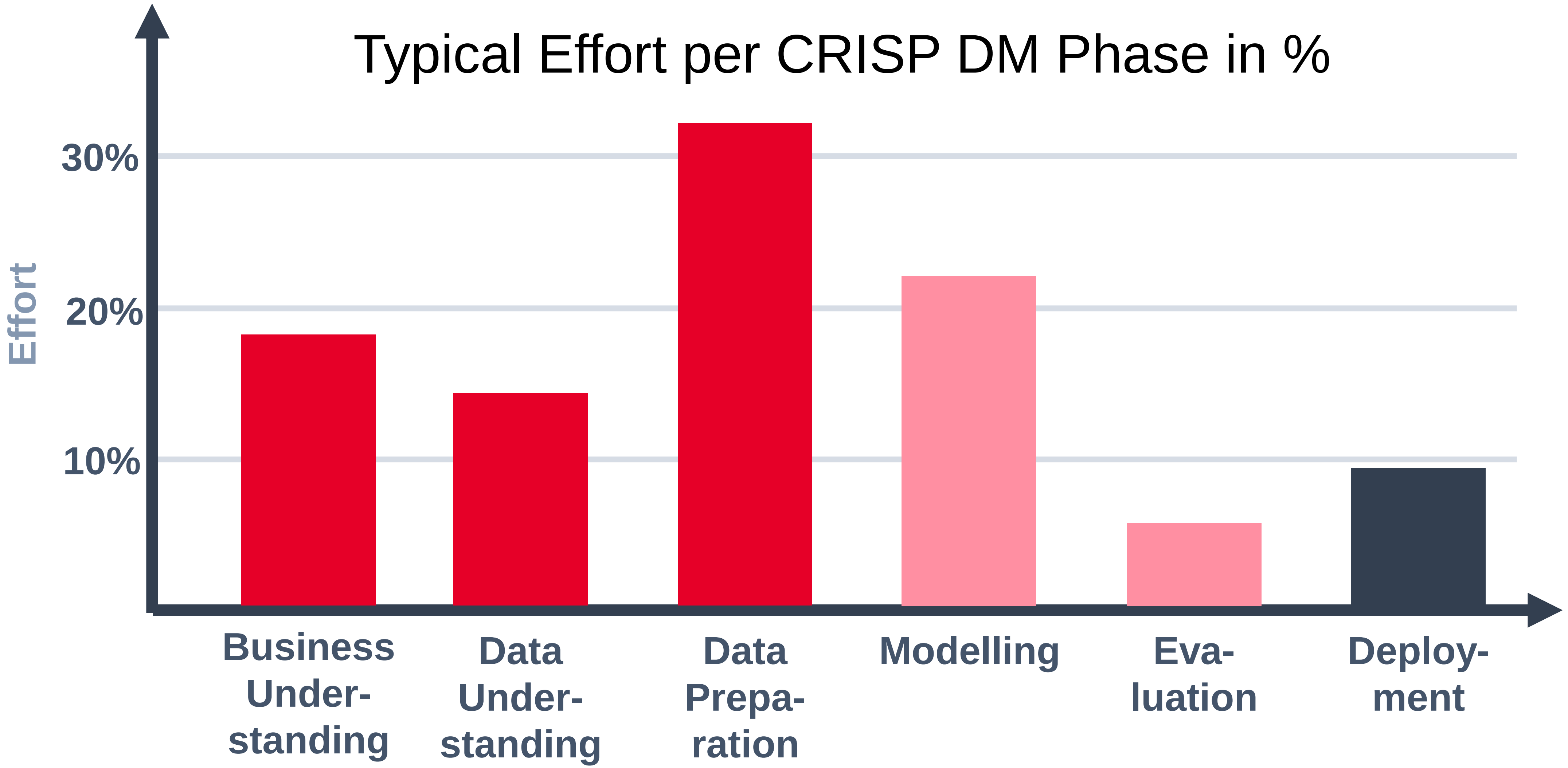
Plan Deployment
Deployment Plan.

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan.

Produce Final Report
Final Report.
Final Presentation.

Review Project
Experience Documentation.

CRISP DM – Objectives and Benefits



CRISP DM – 1 Business Understanding

1.1 Determine Business Objectives

Background.
Business Objectives.
Business Success Criteria.

1.2 Assess Situation

Inventory of Resources, Requirements, Assumptions and Constraints.
Risks and Contingencies Terminology.
Costs and Benefits.

1.3 Determine Data Mining Goals

Data Mining Goals.
Data Mining Success Criteria.

1.4 Produce Project Plan

Project Plan.
Initial Assessment of Tools and Techniques.

CRISP DM – 2 Data Understanding

2.1 Collect Initial Data

Initial Data Collection Report.

2.2 Describe Data

Data Description Report.

2. 3 Explore Data

Data Exploration Report.

2.4 Verify Data Quality

Data Quality Report.

CRISP DM – 3 Data Preparation

3.1 Select Data
Rationale for Inclusion / Exclusion.

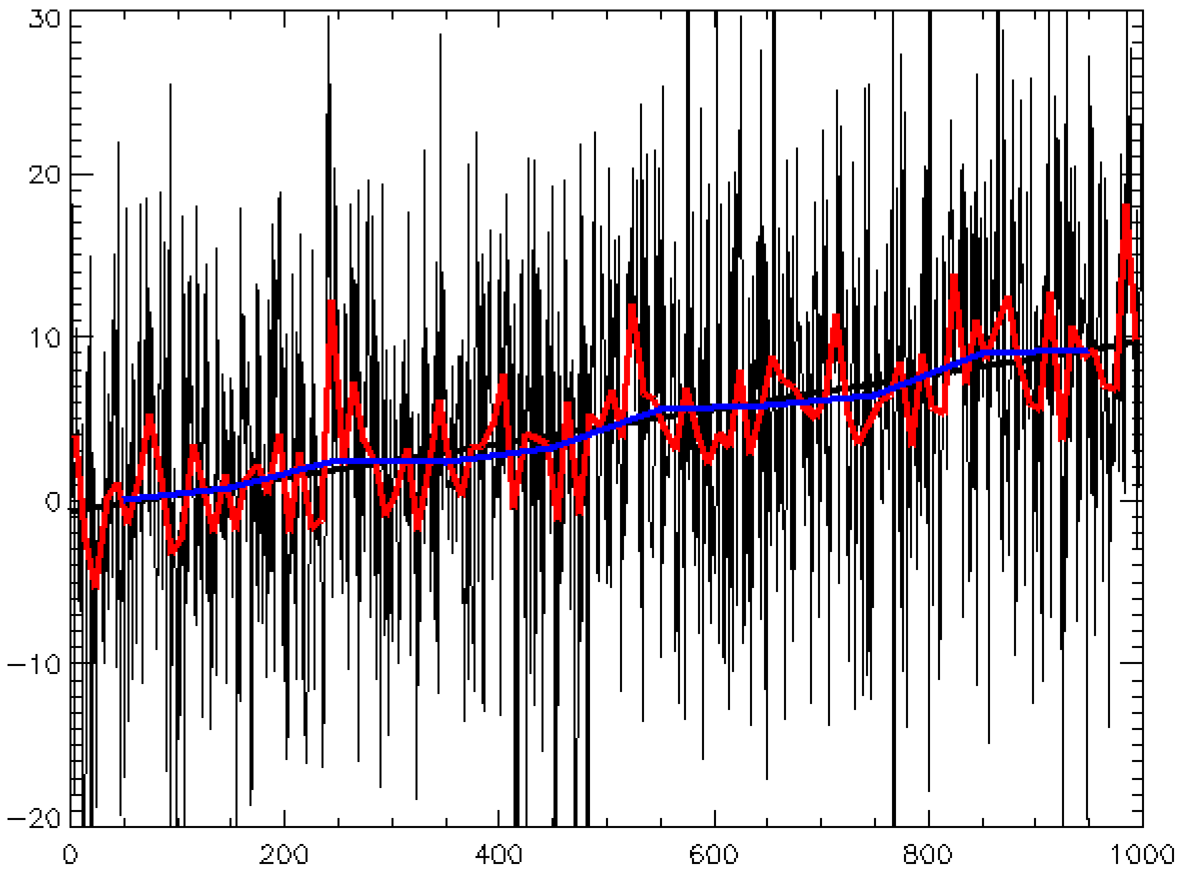
3.2 Clean Data
Data Cleaning Report.

3.3 Construct Data
Derived Attributes.
Generated Records.

3.4 Integrate Data
Merged Data.

3.5 Format Data
Reformatted Data.

3.6 Dataset
Dataset Description.



CRISP DM – 4 Modelling

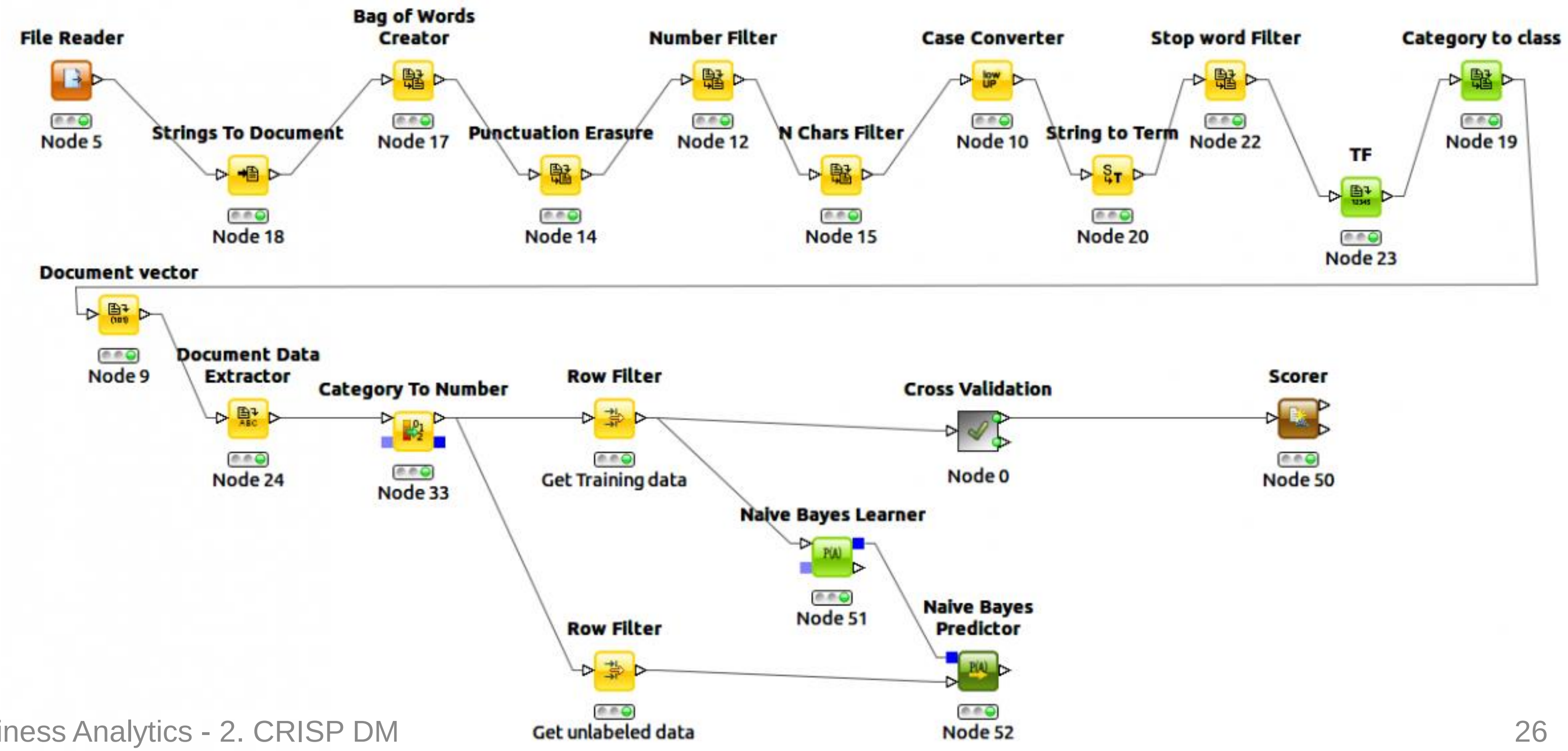
4.1 Select Modelling Technique
Modelling Technique.
Modelling Assumptions.

4.2 Generate Test Design
Test Design.

4.3 Build Model
Parameter Settings Models.
Model Description.

4.4 Assess Model
Model Assessment.
Revised Parameter Settings.

	Data Mining Context			
Dimension	Application Domain	Data Mining Problem Type	Technical Aspect	Tool and Technique
Examples	Response Modeling	Description and Summarization	Missing Values	Clementine
	Churn Prediction	Segmentation	Outliers	MineSet
	...	Concept Description	...	Decision Tree
		Classification		...
		Prediction		
		Dependency Analysis		



CRISP DM – 5 Evaluation

5.1 Evaluate Results

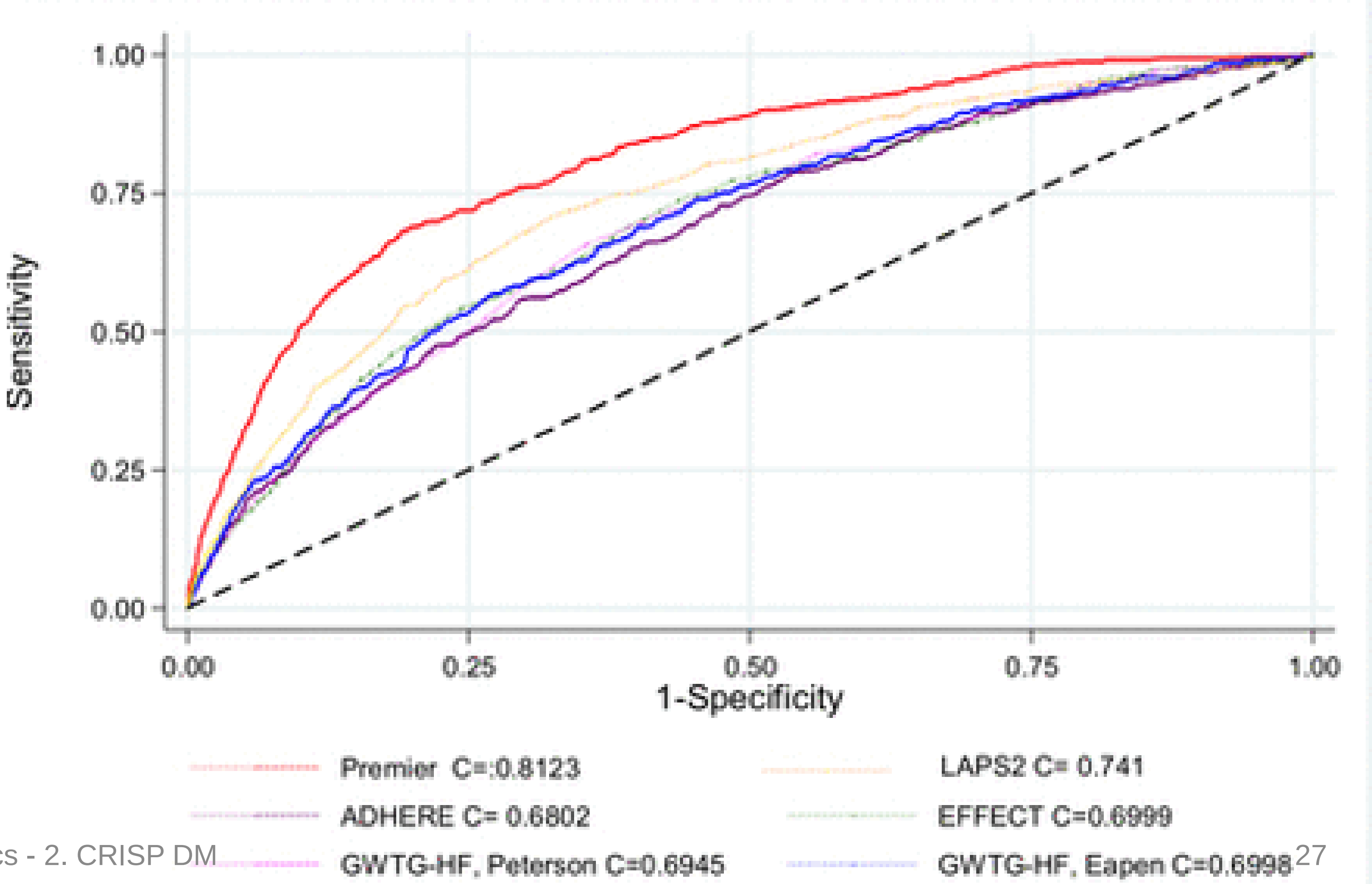
Assessment of Data.
Mining Results with respect to Business Success Criteria.
Approved Models.

5.2 Review Process

Review of Process.

5.3 Determine Next Steps

List of Possible Actions.
Decision.



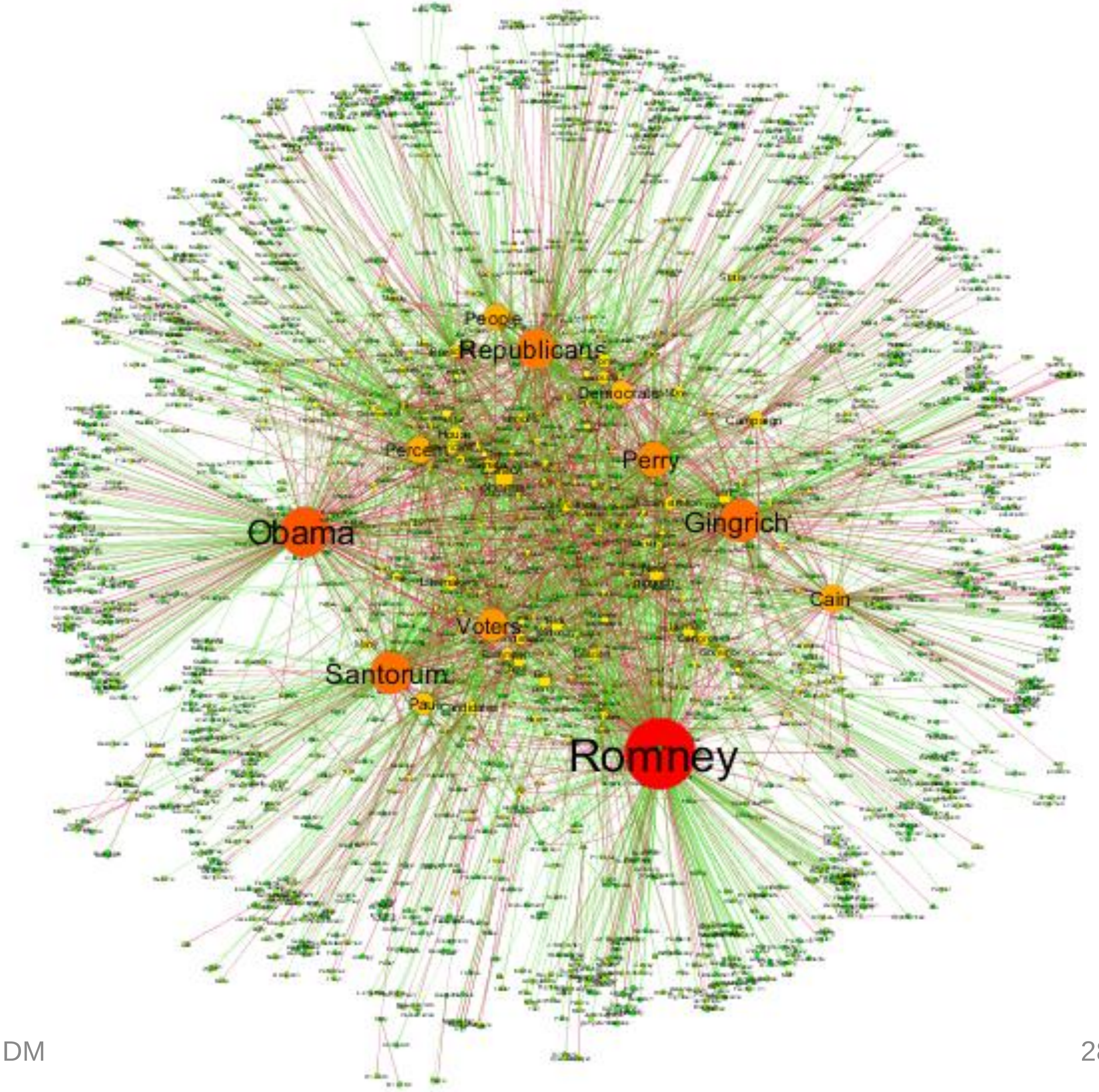
CRISP DM – 6 Deployment

6.1 Plan Deployment
Deployment Plan.

6.2 Plan Monitoring and Maintenance
Monitoring and Maintenance Plan.

6.3 Produce Final Report
Final Report.
Final Presentation.

6.4 Review Project
Experience Documentation.



Lessons Learned

- CRISP DM is a highly adopted and standardized process for data mining projects.
- Ex-ante definition of success criteria is essential for successful projects.
- Data understanding and preparation are typically the most costly and time-consuming (~80%) phases in CRISP DM.
- CRISP DM is an iterative approach. Certain phases are likely to be passed multiple times (modelling and evaluation.

Resources

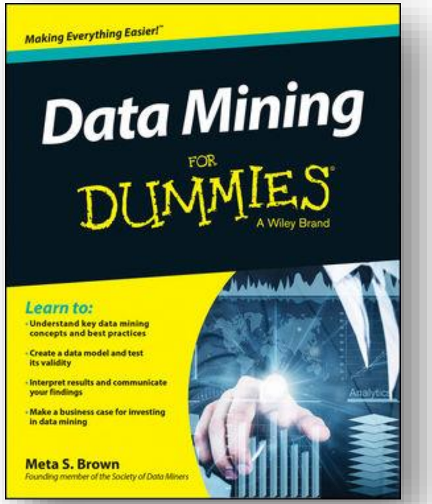
CRISP DM 1.0 Document
https://www.the-modeling-agency.com/crisp-dm.pdf

From Data Mining to Knowledge Discovery in Databases
http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

IBM ASUM DM
https://developer.ibm.com/predictive analytics/2015/10/16/have-you-seen-asum-dm/

Data Mining Curriculum, ACM
http://www.kdd.org/exploration_files/CURMay06.pdf

I do not recommend this, but it‘s great.



Get Prepared (Homework)

- Read the KDD article by Usama Fayyad
  http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

- Read the CRISP DM 1.0 Document
  https://www.the-modeling-agency.com/crisp-dm.pdf

- Read the Data Mining Curriculum
  http://www.kdd.org/exploration_files/CURMay06.pdf

Any Questions?