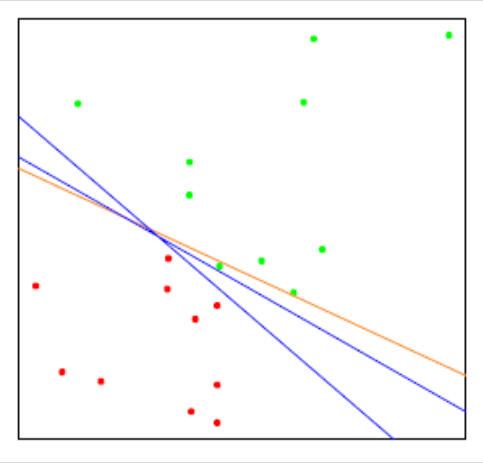


Big Data Analytics

Basic concepts of analyzing very large amounts of data

Dr. – Ing. Morris Riedel
Adjunct Associated Professor
School of Engineering and Natural Sciences, University of Iceland
Research Group Leader, Juelich Supercomputing Centre, Germany

LECTURE BDA2

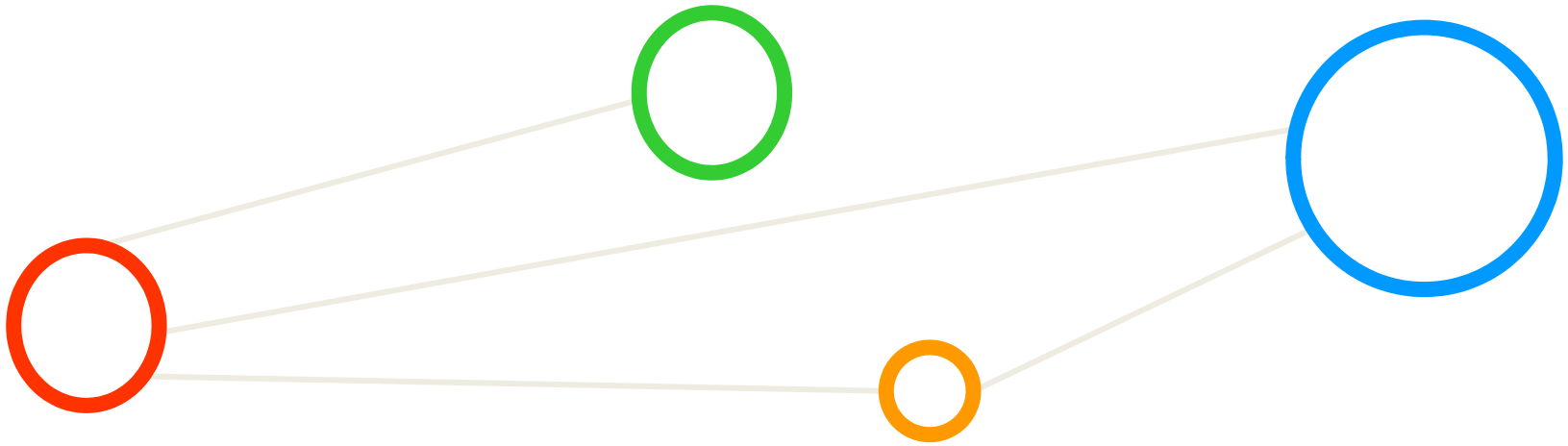
Data Mining & Analysis Process Models

2014-01-01 (v2)
Online Material

UNIVERSITY OF ICELAND
SCHOOL OF ENGINEERING AND NATURAL SCIENCES
FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

JÜLICH
FORSCHUNGSZENTRUM

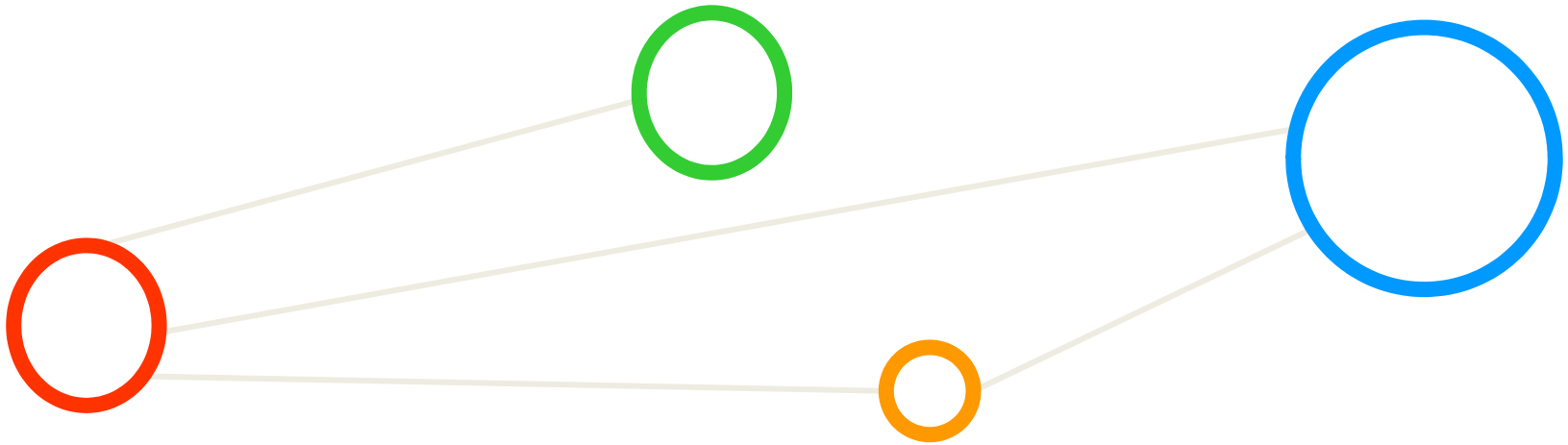
Outline



Outline

- CRISP-DM Process Model
  - Cross-Industry Standard Process for Data Mining (CRISP-DM)
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment

CRISP-DM Process Model



Overview of Six Known Data Mining Phases

- **A typical data mining project includes six known phases: (1) Problem Understanding; (2) Data Understanding; (3) Data Preparation; (4) Modeling; (5) Evaluation; (6) Deployment**

- Six phases for typical data mining project
  - Cross-Industry Standard Process for Data Mining (CRISP-DM)

[1] CRISP-DM Model

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

[3] CRISP-DM Wikipedia

1 – Business Understanding

- **The Business Understanding phase consists of four distinct tasks: (A) Determine Business Objectives; (B) Situation Assessment; (C) Determine Data Mining Goal; (D) Produce Project Plan**

- Task A – Determine Business Objectives
  - Background, Business Objectives, Business Success Criteria
- Task B – Situation Assessment
  - Inventory of Resources, Requirements, Assumptions, and Contraints
  - Risks and Contingencies, Terminology, Costs & Benefits
- Task C – Determine Data Mining Goal
  - Data Mining Goals and Success Criteria
- Task D – Produce Project Plan
  - Project Plan, Initial Assessment of Tools & Techniques

[2] CRISP-DM User Guide

2 – Data Understanding

- **The Data Understanding phase consists of four distinct tasks:**
  (A) Collect Initial Data; (B) Describe Data; (C) Explore Data; (D) Verify Data Quality

- Task A – Collect Initial Data
  - Initial Data Collection Report
- Task B – Describe Data
  - Data Description Report
- Task C – Explore Data
  - Data Exploration Report
- Task D – Verify Data Quality
  - Data Quality Report

[2] CRISP-DM User Guide

3 – Data Preparation

- **The Data Preparation phase consists of six distinct tasks: (A) Data Set; (B) Select Data; (C) Clean Data; (D) Construct Data; (E) Integrate Data; (F) Format Data**

- Task A – Data Set
  - Data set description
- Task B – Select Data
  - Rationale for inclusion / exclusion
- Task C – Clean Data
  - Data cleaning report
- Task D – Construct Data
  - Derived attributes, generated records
- Task E – Integrate Data
  - Merged data
- Task F – Format Data
  - Reformatted data

[2] CRISP-DM User Guide

4 – Modeling

- **The Data Preparation phase consists of four distinct tasks: (A) Select Modeling Technique; (B) Generate Test Design; (C) Build Model; (D) Assess Model;**

- Task A – Select Modeling Technique
  - Modeling assumption, modeling technique
- Task B – Generate Test Design
  - Test design
- Task C – Build Model
  - Parameter settings, models, model description
- Task D – Assess Model
  - Model assessment, revised parameter settings

[2] CRISP-DM User Guide

5 – Evaluation

- **The Data Preparation phase consists of three distinct tasks: (A) Evaluate Results; (B) Review Process; (C) Determine Next Steps**

- Task A – Evaluate Results
  - Assessment of data mining results w.r.t. business success criteria
  - List approved models
- Task B – Review Process
  - Review of Process
- Task C – Determine Next Steps
  - List of possible actions, decision

[2] CRISP-DM User Guide

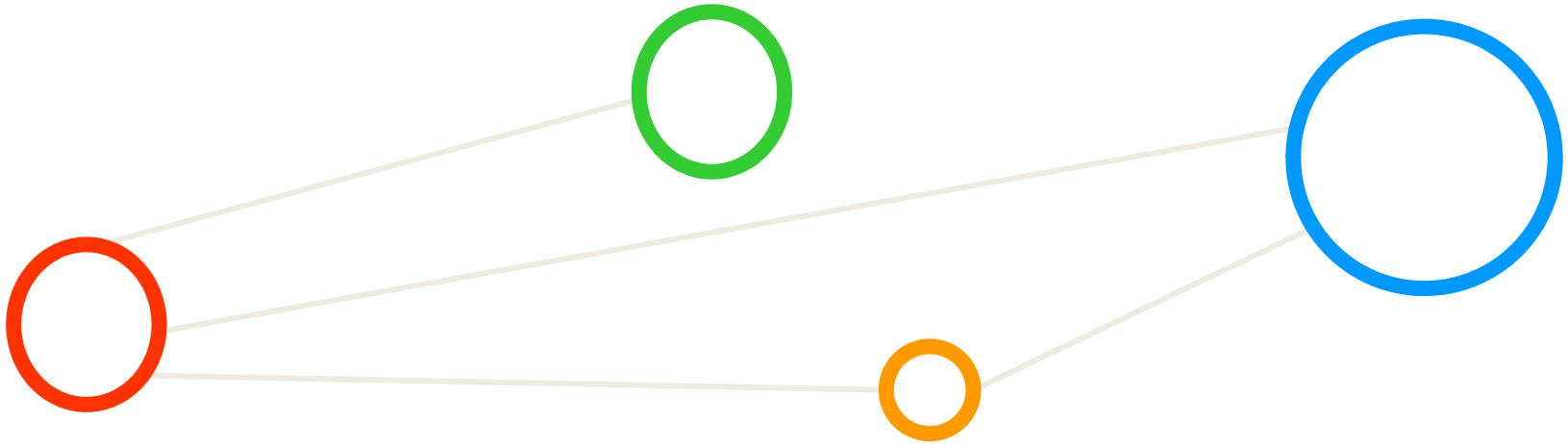
6 – Deployment

- **The Data Preparation phase consists of three distinct tasks: (A) Plan Deployment; (B) Plan Monitoring and Maintenance; (C) Produce Final Report; (D) Review Project**

- Task A – Plan Deployment
  - Establish a deployment plan
- Task B – Plan Monitoring and Maintenance
  - Create a monitoring and maintenance plan
- Task C – Product Final Report
  - Create final report and provide final presentation
- Task D – Review Project
  - Document experience, provide documentation

[2] CRISP-DM User Guide

Lecture Bibliography



Lecture Bibliography

- [1] Shearer C., *The CRISP-DM model: the new blueprint for data mining*, J Data Warehousing, 2000, 5, pages 13—22.
- [2] Pete Chapman, 'CRISP-DM User Guide', 1999, Online: http://lyle.smu.edu/~mhd/8331f03/crisp.pdf
- [3] Wikipedia, 'Cross Industry Standard Process for Data Mining',
  Online: http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining


services
measurement
within
fusion
new
funding
Policy-based
concepts
device
Analysis
forms
cross-disciplinary
resources
Centre
climate
Computer
expertise
Approaches
construction
Fusion
year
like
Computational
Test
E.g
Computing
disciplines
results
Enable
different
Cross-Disciplinary
community
Environment
Modelling
EU
methods
High
big
nodes
PB
Key
scientific
important
Science
centers
storage
basis
Infrastructure
files
model
etc
DKRZ
Use
SMAQ
international
engineering
References
formats
simulations
ENES
technologies
physical
brain
increasing
via
EUDAT
IPCC
Big
Project
Data
processing
Common
TB
Methods
MapReduce
create
computing
Using
European
databases
GB
Supercomputing
Work
images
codes
analysis
stored
Resources
analyze
Large
Summary
data-intensive
Service
Scientific
levels
often
DLCL
Level
external
use
Research
Matters
Concrete
performance
JSC
computational
research
e.g
science
HPC
Juelich
system
datasets
acquisition
large-scale
compute
Landscape
management
System
Climate
modelling
AR5
Hardware
distributed
cores
access
hundreds
Nutshell
Services
Technologies
Earth
Structure
current
network
DLCLs
Understanding
structures
directory
project
General
Germany
Simulation
Processing
Provide
NASA
Energy
systems
day
Health