The CRISP-DM User Guide

Brussels SIG Meeting

Pete Chapman

NCR Systems Engineering Copenhagen

email: Pete.Chapman@Copenhagen.NCR.com

Agenda

- CRISP-DM Objectives and Benefits
- CRISP-DM Deliverables
- CRISP-DM Methodology, Phases and Tasks
- CRISP-DM User Guide
- Possible CRISP-DM Futures

Objectives and Benefits of CRISP-DM

- 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)

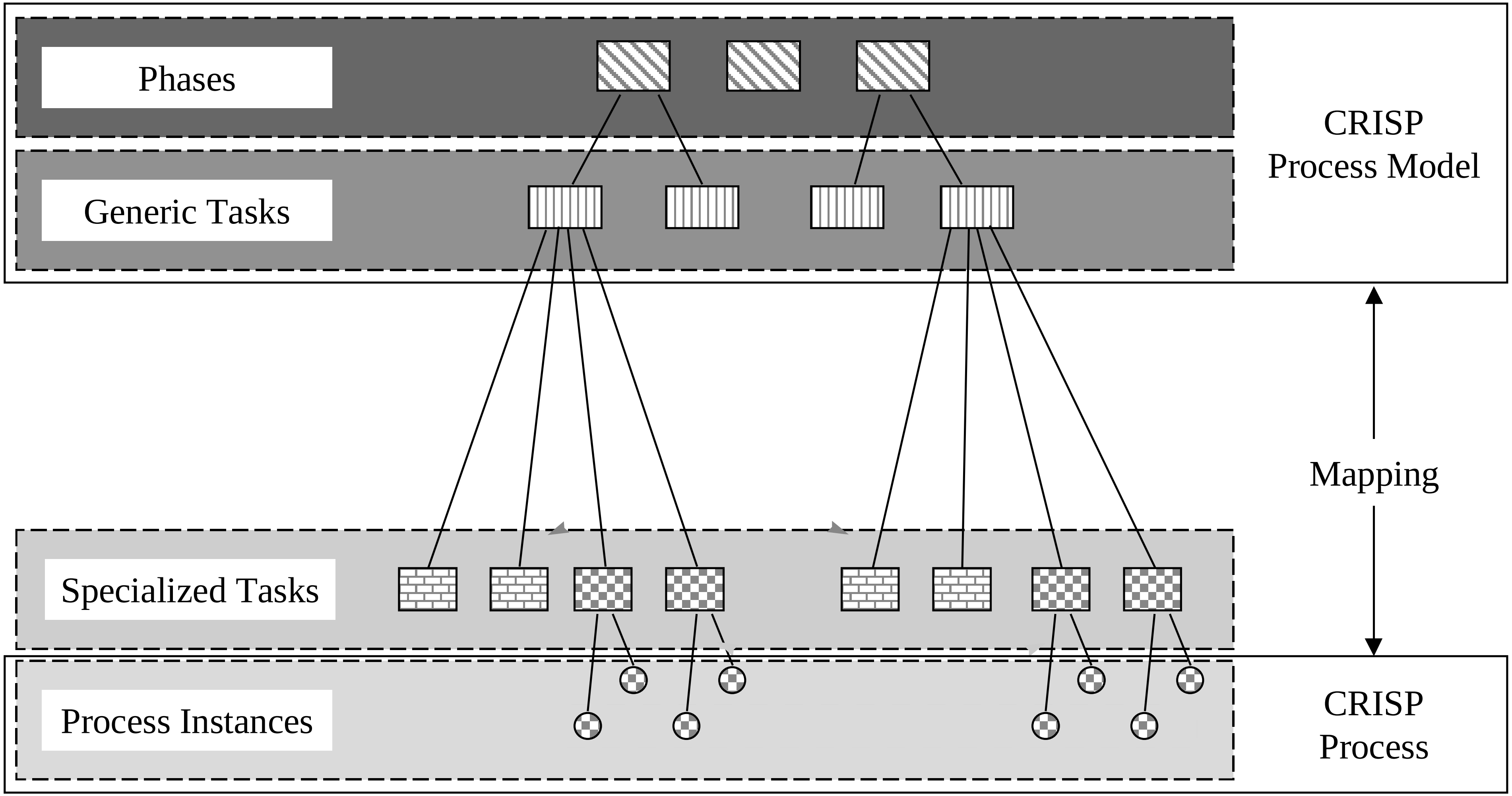
- tool and technique independent
- tool supportable

- support documentation of projects
- capture experience for reuse
- support knowledge transfer and training

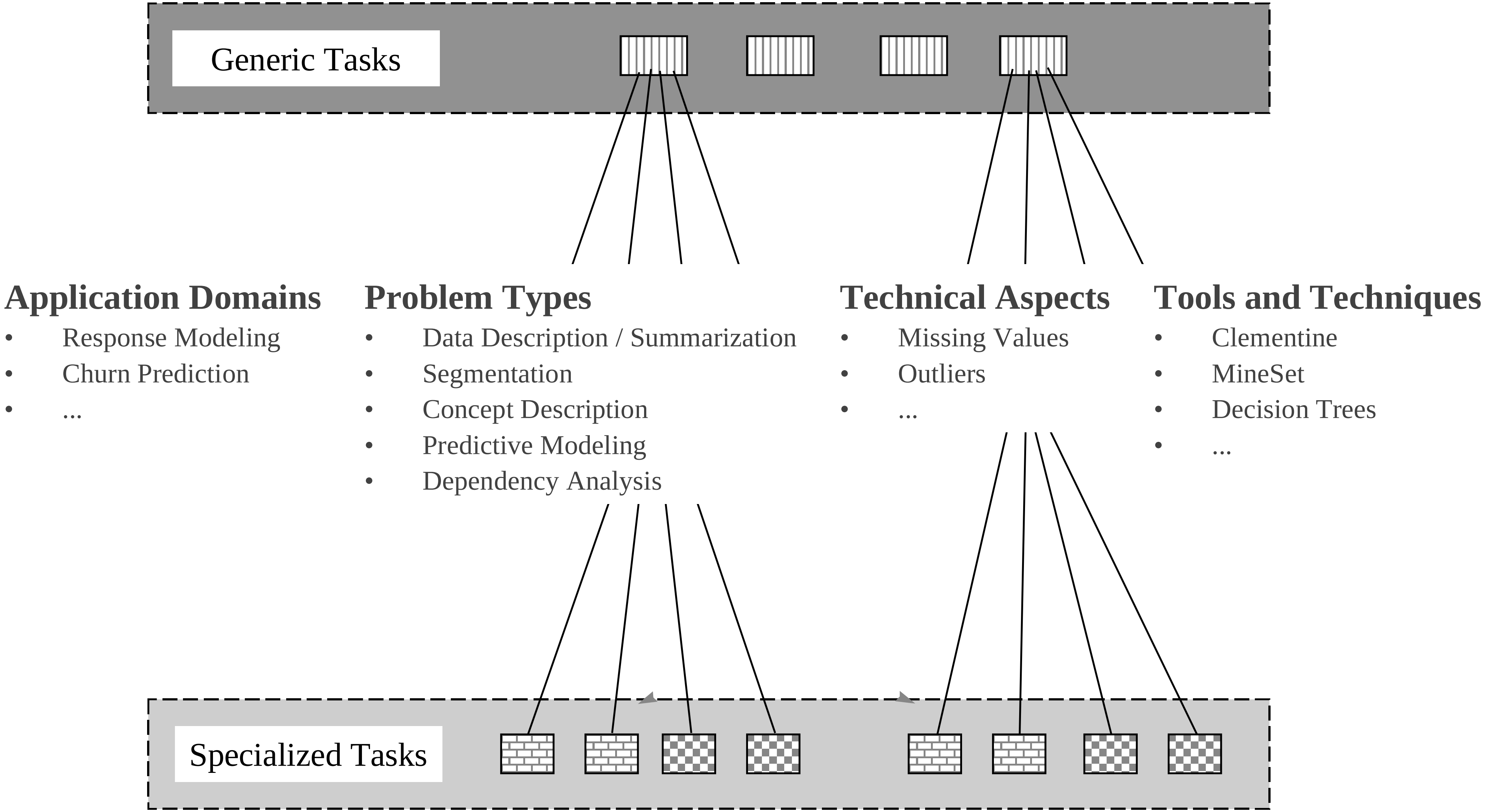
CRISP-DM Deliverables

- Process Model
  - Methodology
  - Reference Model
  - User Guide
  - Output (Deliverable/Templates)
- Tool Support
  - Tool Support Definitions
  - Stream Library
- Experimentation
  - Experimentation Reports
  - CRISP-DM SIG User Feedback

CRISP-DM Methodology



Data Mining Contexts



Application Domains

- Response Modeling
- Churn Prediction
- ...

Problem Types

- Data Description / Summarization
- Segmentation
- Concept Description
- Predictive Modeling
- Dependency Analysis

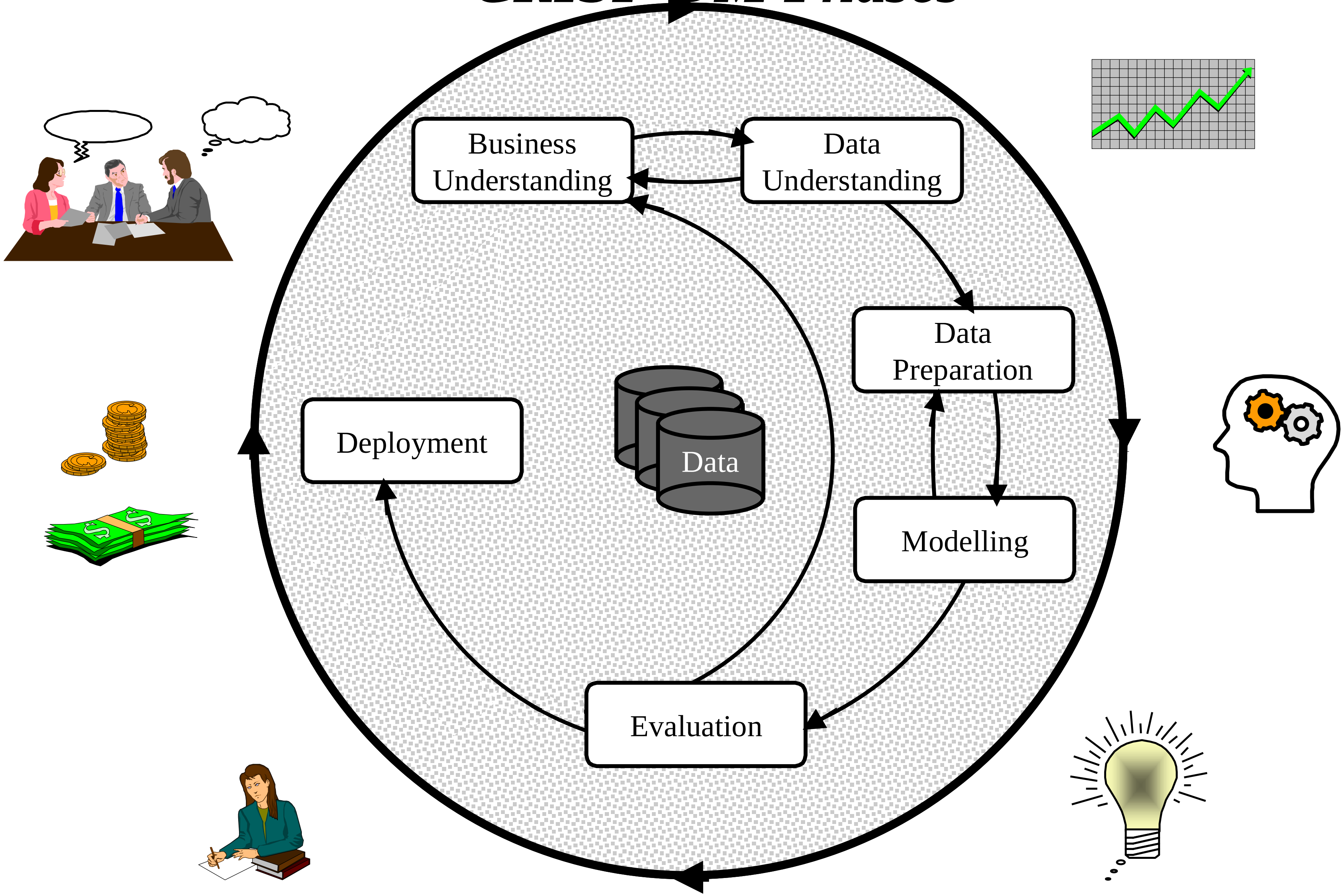
Technical Aspects

- Missing Values
- Outliers
- ...

Tools and Techniques

- Clementine
- MineSet
- Decision Trees
- ...

CRISP-DM Phases



Phases and Tasks

Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

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

Determine Data Mining Goal
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Asessment 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

Data Set
Data Set Description

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

Modeling

Select Modeling Technique
Modeling Technique
Modeling 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

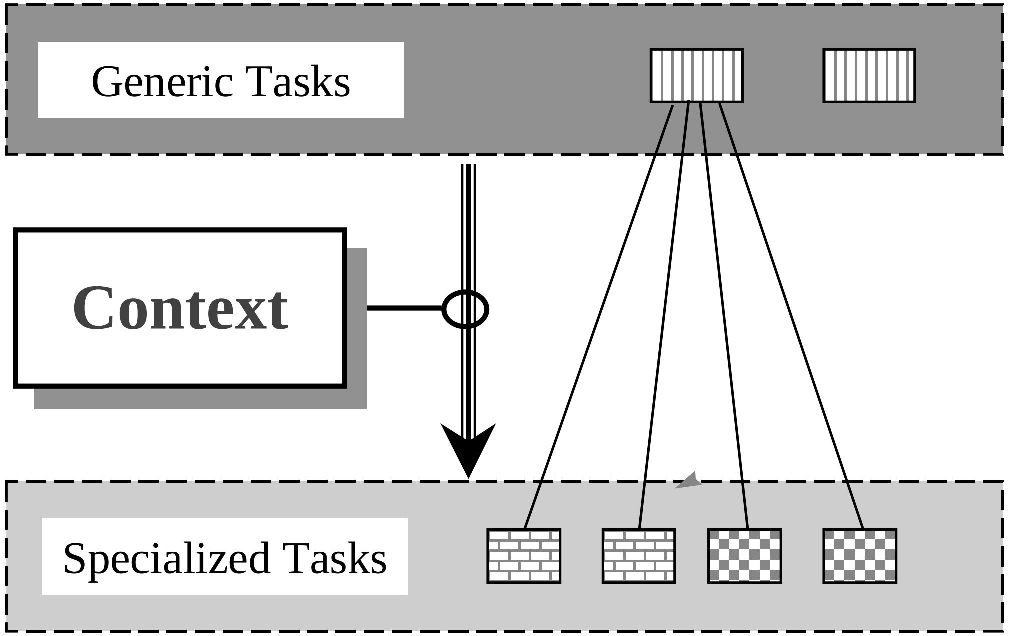
Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

Introduction to the User Guide

Reference Model

What To Do?

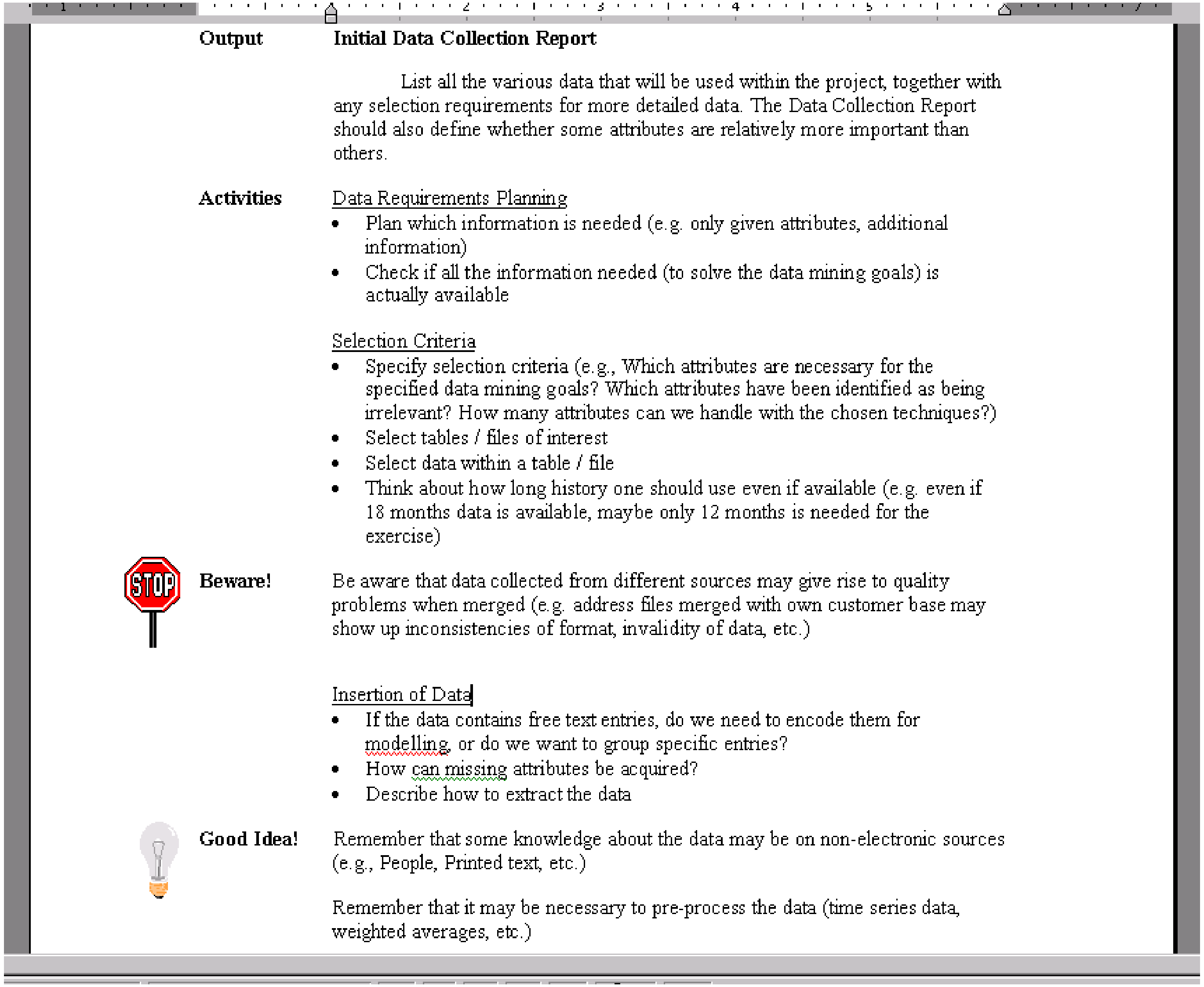


User Guide

How To Do?

- check lists
- questionaires
- tools
- sequences of steps
- decision points
- pitfalls

CRISP-DM User Guide



Output **Initial Data Collection Report**

List all the various data that will be used within the project, together with any selection requirements for more detailed data. The Data Collection Report should also define whether some attributes are relatively more important than others.

Activities Data Requirements Planning

- Plan which information is needed (e.g. only given attributes, additional information)
- Check if all the information needed (to solve the data mining goals) is actually available

Selection Criteria

- Specify selection criteria (e.g., Which attributes are necessary for the specified data mining goals? Which attributes have been identified as being irrelevant? How many attributes can we handle with the chosen techniques?)
- Select tables / files of interest
- Select data within a table / file
- Think about how long history one should use even if available (e.g. even if 18 months data is available, maybe only 12 months is needed for the exercise)

Beware! Be aware that data collected from different sources may give rise to quality problems when merged (e.g. address files merged with own customer base may show up inconsistencies of format, invalidity of data, etc.)

Insertion of Data

- If the data contains free text entries, do we need to encode them for modelling, or do we want to group specific entries?
- How can missing attributes be acquired?
- Describe how to extract the data

Good Idea! Remember that some knowledge about the data may be on non-electronic sources (e.g., People, Printed text, etc.)

Remember that it may be necessary to pre-process the data (time series data, weighted averages, etc.)

How to use the User Guide (i)

- Contents of the User Guide
  - More detailed description of the various tasks using:
    - Activities List
    - Check Lists
    - Good Ideas
    - Warnings!

- What is NOT in the User Guide
  - Deliverables/Document Templates (as yet)
  - Description of Techniques and Tools (as yet)
  - Estimates of engagements
  - Quality Indicators

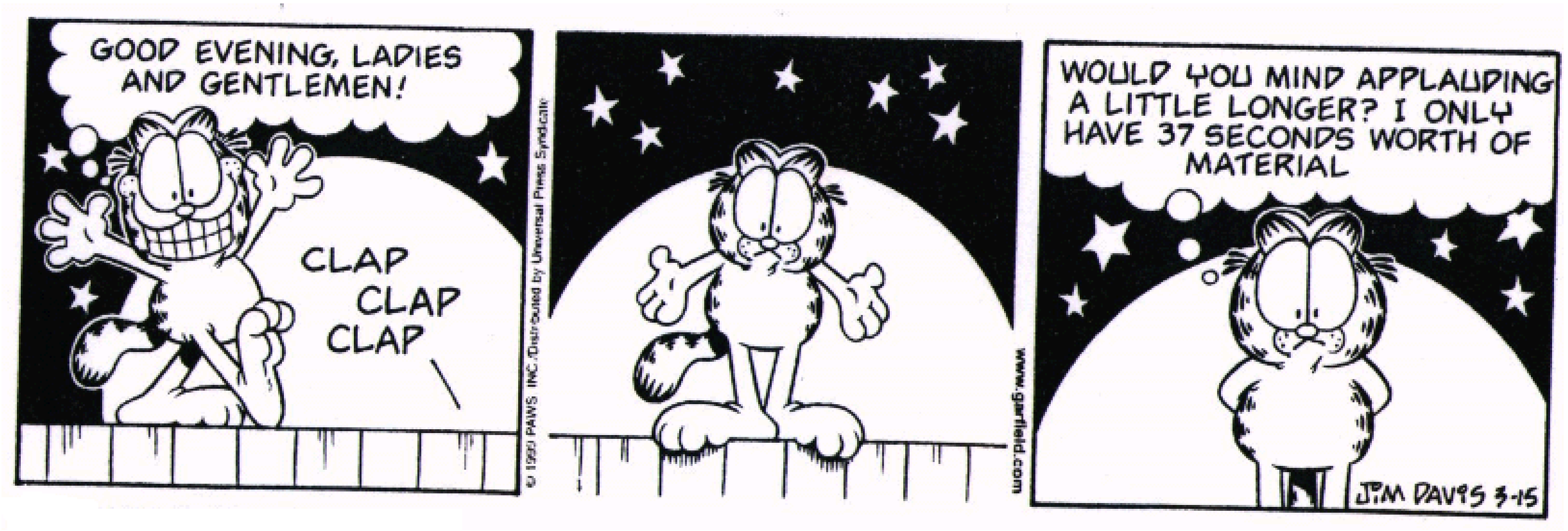
How to use the User Guide (ii)

- Beginning Data Miners
  - What tasks do I need to do?
  - What is the order of the tasks in a Data Mining Engagement?
  - What risks do I run?
  - Are there any “shortcuts” in my tasks?
  - What are the format of the deliverables that I need to resent to management?

- Experienced Data Miners
  - Have I missed any activity?
  - Are there any tasks or activity that I can leave until later?
  - How can I make a Project Plan?
  - How can I document the project for later re-use?

Possible Future CRISP-DM Deliverables

- “CRISP-DM - The Book ”, includes
  - Experiences, feedback from SIG members
  - Reference Model, User Guide updated with experiments
  - Full Deliverables/Document Templates
  - Case Studies
  - Mapping Advice from Generic to Specific Engagements
  - More explicit advice on Tools & Techniques
  - Advice on documentation of engagements, establishment of Data Mining Library,…..


GOOD EVENING, LADIES AND GENTLEMEN!
CLAP CLAP CLAP
© 1990 PAWS INC./Distributed by Universal Press Syndicate
www.garfield.com
WOULD YOU MIND APPLAUDING A LITTLE LONGER? I ONLY HAVE 37 SECONDS WORTH OF MATERIAL
JIM DAVIS 3-15