NOVA IMS

Information Management School

2

# CRISP-DM PROCESS MODEL

**Data Science for Marketing**

Acreditações e Certificações

Instituto Superior de Estatística e Gestão da Informação
Universidade Nova de Lisboa

# Summary

1. Introduction
2. Global overview
3. Business understanding phase

2.1

# Introduction

CRISP-DM process model

# Why use a standard process model

- Framework to record and replicate projects
- Assists project planning and management
- Encourage best practices and the obtention of better results
- Provides a base for new practitioners:
  - Demonstrates the maturity of Data Mining/Data Science
  - Reduces dependency of ”experts”

# Data Mining/Data Science processes

- KDD: Knowledge Discovery in Databases
- SEMMA: Sampling, Exploring, Modifying, Modelling, and Assessing
- CRISP-DM: CRoss Industry Standard Process for Data Mining

2.2

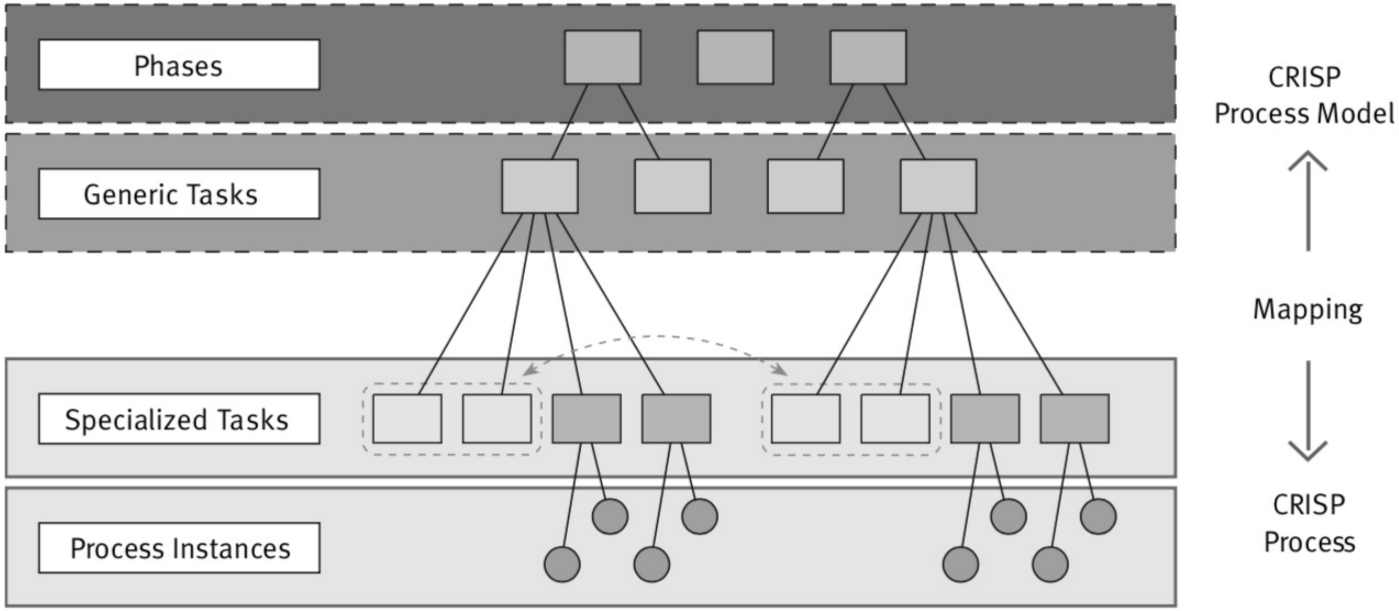
# Global overview

CRISP-DM process model

# CRISP-DM

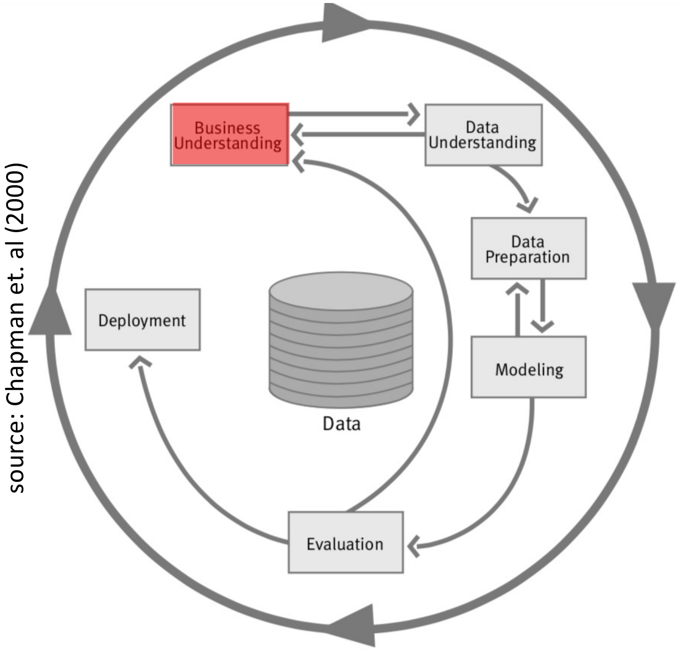
- Applies not only to DM projects, but also to Text Mining, Statistics, and Descriptive and Predictive Analytics
- Used in academy and by DM practitioners
- Non-proprietary
- Tool neutral
- Focus both on the application and the technical perspectives
- Most-often used process model

# Four level breakdown



source: Chapman et. al (2000)

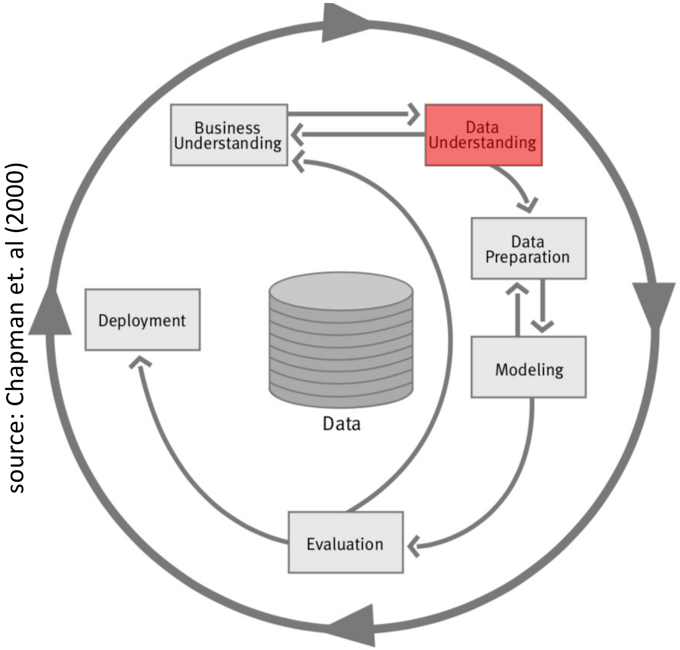
# Phase: Business understanding



source: Chapman et. al (2000)

- Determine business objectives
  - Background
  - Business objectives
  - Business success criteria
- Assess situation
  - Resources
  - Requirements
  - Risks and contingencies
- Determine data mining goals and success criteria
- Produce project plan

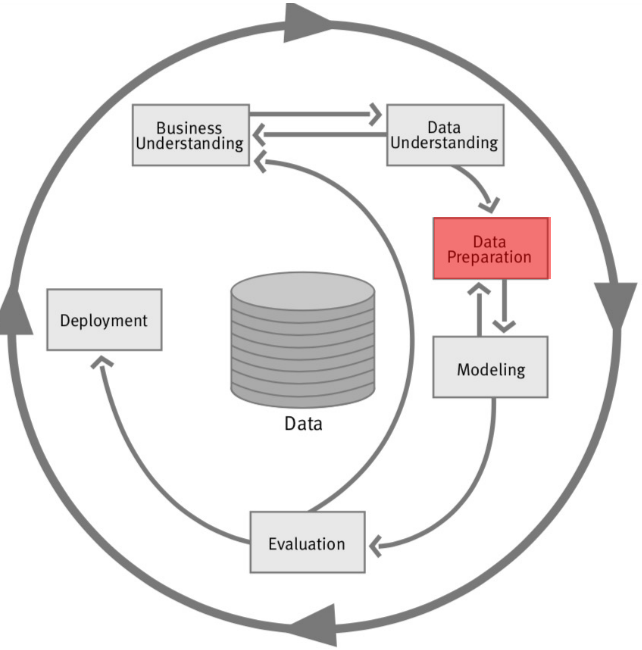
# Phase: Data understanding



source: Chapman et. al (2000)

- Collect initial data
- Describe data
- Explore data
- Verify data quality

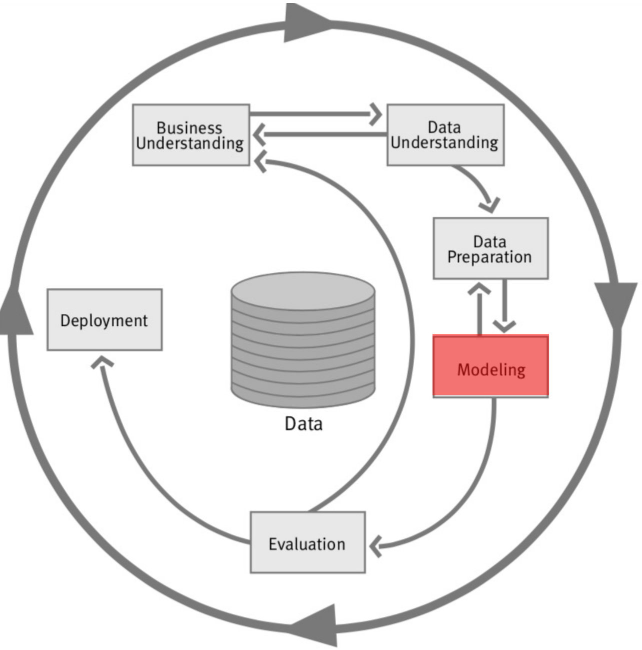
# Phase: Data preparation



source: Chapman et. al (2000)

- Select data
- Clean data
- Construct data
- Integrate/merge data
- Format data

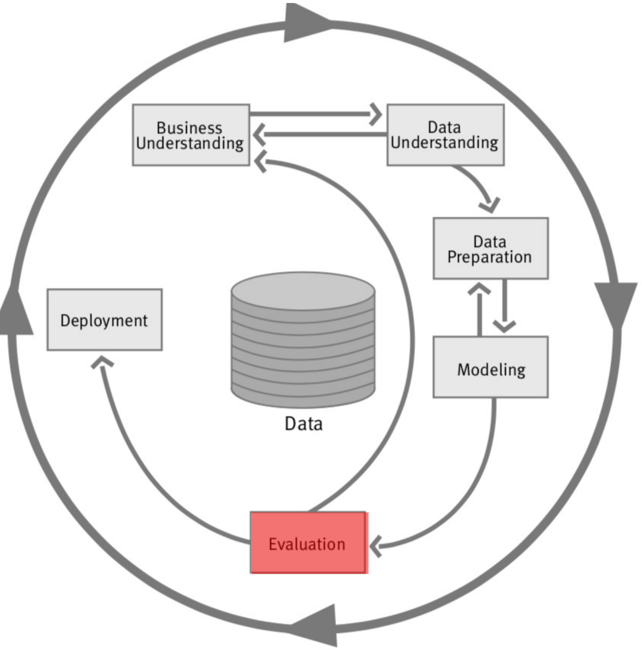
# Phase: Modeling



source: Chapman et. al (2000)

- Select modeling techniques
  - Algorithm selection
  - Modeling assumptions
- Generate test design
- Build model
- Assess model

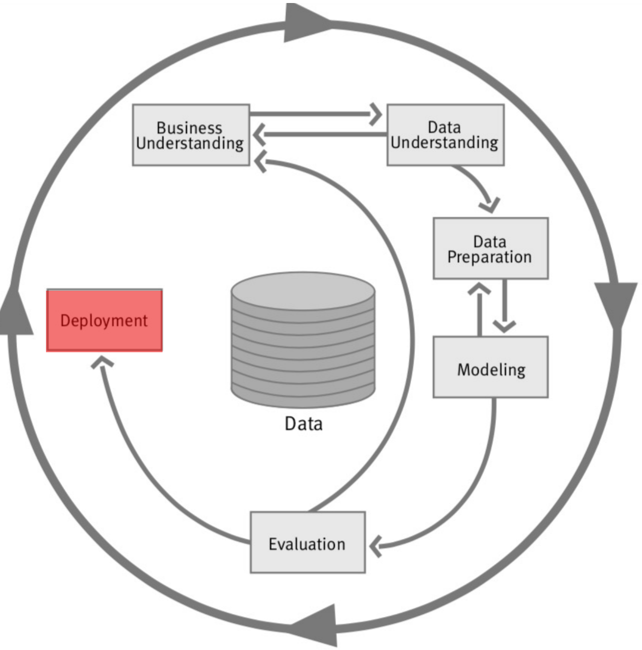
# Phase: Evaluation



source: Chapman et. al (2000)

- Evaluate results
  - Assess data mining results vs business success criteria
  - Approve model
- Review process
- Determine next steps
  - Production or not?
  - Additional requirements?

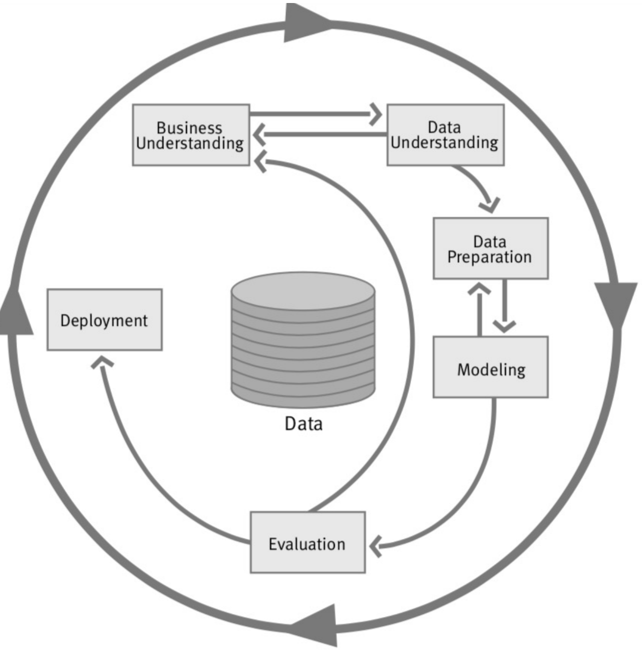
# Phase: Deployment



source: Chapman et. al (2000)

- Plan deployment
  - Strategy to deploy the model, including integration in business processes
- Plan monitoring and maintenance
  - Performance assessment
  - Models' update
- Produce final report
- Review project

# Cyclical nature



source: Chapman et. al (2000)

- DM does not end once a project is deployed!!!
- Lessons learned during the project development and from the deployed project can trigger more-focused business questions

2.3

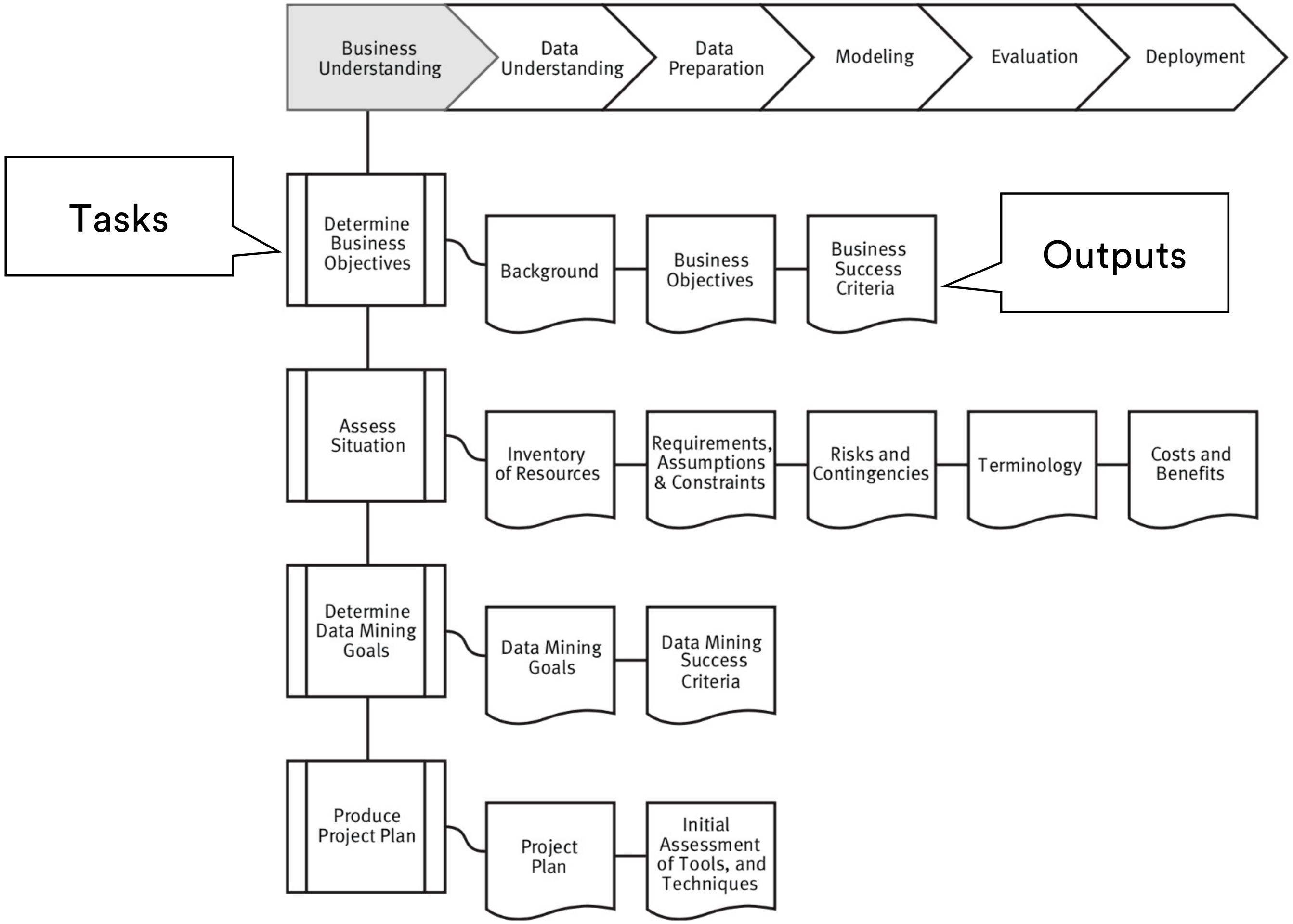
# Business understanding phase

CRISP-DM process model

*“A possible consequence of* ***neglecting this step*** *is to expend a great deal of effort* ***producing the right answers to the wrong questions****”*

Abbott (2015)

# Business understanding



# Determine business objectives
# Background

## ORGANIZATION

- Identify key persons
- Identify the internal sponsor and main expert
- Define steering committee
- Identify affected business units

## PROBLEM AREA

- Identify the problem area (e.g., marketing)
- Describe the problem in general terms
- Identify target groups (e.g., users or managers)
- Identify users’ needs and expectations

## CURRENT SOLUTION

- Identify and describe current solution used to address the problem (if any)
- Describe the pros and cons of the current solution (if any)

## Determine business objectives

# Business objectives

- Informally describe the **problem to be solved** (e.g., increase customers loyalty to increase sales)
- Specify all business questions as precisely as possible
- Specify any other business requirements (e.g., vouchers cannot exceed 25% of the benefits)
- Specify **expected benefits in business terms** (e.g., identify customers visiting patterns and try to increase the number of visits) – should be AS REALISTIC AS POSSIBLE!

## Determine business objectives
# Business success criteria

- Specify **business success criteria:**
  - Measurable (e.g., increase the visits by 5% per quarter); or
  - Subjective (e.g., give useful insights into frequent visitors)
- Identify who assesses the success criteria

## Assess Situation

# Inventory of resources (1/3)

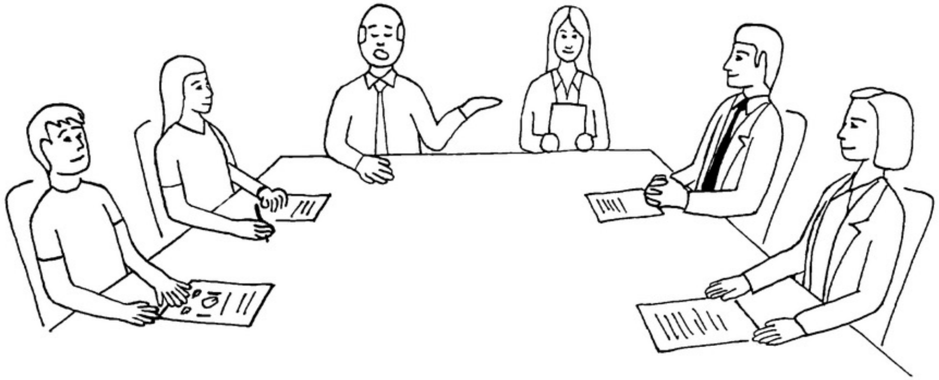
Hardware

- Identify required hardware
- Establish the hardware availability

# Assess Situation
# Inventory of resources (2/3)

Personal

- Identify project sponsor (if not the main sponsor)
- Identify systems, databases, and other technical administrators
- Identify DM experts, statisticians, and other analysts



"IT'S GREAT TO HAVE BUSINESS AND IT IN THE ROOM. EVERYONE PLEASE MEET SHEILA. SHEILA WILL BE YOUR TRANSLATOR TODAY."

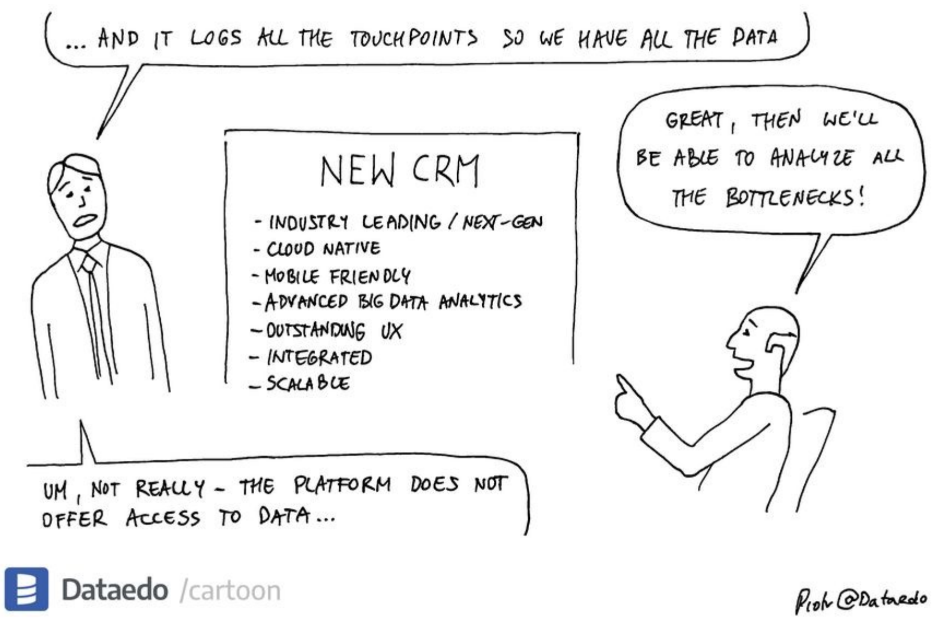
Dataedo /cartoon

Piotr@Dataedo

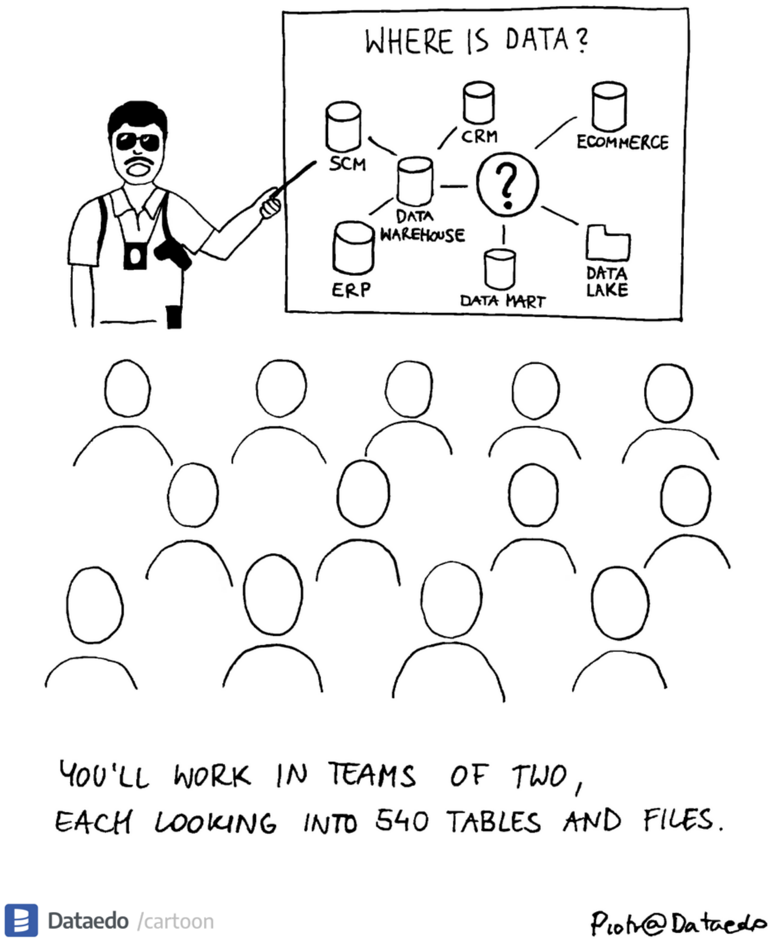
# Assess Situation
# Inventory of resources (3/3)

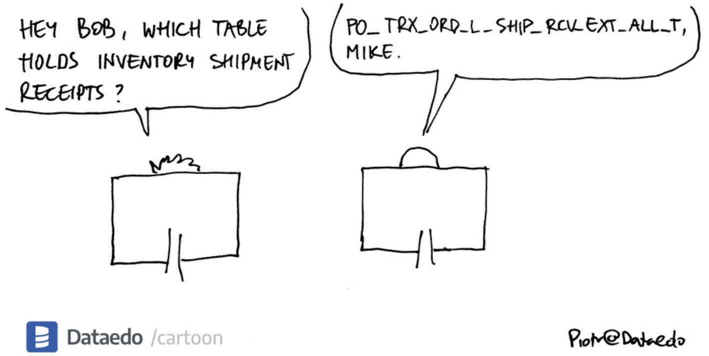
Personal

- Identify knowledge and types of knowledge sources
- Check available tools and techniques
- Identity data and types of data sources (e.g., online, experts, docs., etc.)


... AND IT LOGS ALL THE TOUCHPOINTS SO WE HAVE ALL THE DATA
NEW CRM
- INDUSTRY LEADING / NEXT-GEN
- CLOUD NATIVE
- MOBILE FRIENDLY
- ADVANCED BIG DATA ANALYTICS
- OUTSTANDING UX
- INTEGRATED
- SCALABLE
GREAT, THEN WE'LL BE ABLE TO ANALYZE ALL THE BOTTLENECKS!
UM, NOT REALLY - THE PLATFORM DOES NOT OFFER ACCESS TO DATA...
Dataedo /cartoon
Piotr @Dataedo

# Access to data and data knowledge is critical




HEY BOB, WHICH TABLE HOLDS INVENTORY SHIPMENT RECEIPTS?
PO_TRX_ORD_L_SHIP_RCV_EXT_ALL_T, MIKE.
Dataedo /cartoon
Piotr@Dataedo

WE WOULD LIKE TO KNOW WHAT TRX_RCV_INV_AVG_V1 TABLE HOLDS...
Dataedo /cartoon
Piotr@Dataedo

NOVA IMS Information Management School

## Assess Situation

# Requirements, assumptions and constraints

### REQUIREMENTS

- Specify target group
- About comprehensibility, accuracy, deployment, maintenance, etc.
- About security, legal restrictions, privacy, reporting, and project schedule

### ASSUMPTIONS

- Clarify all assumptions (e.g., # of observations)
- List assumptions on data quality, external factors, and costs
- List assumptions about the model explicability or explanation

### CONSTRAINTS

- Check general constraints (e.g., legal, budget, timescale, etc.)
- Check access to data sources (rights and technological issues)
- Check the accessibility of relevant knowledge

## Assess Situation

# Risks and contingencies

### IDENTIFY RISKS

- Identify business risk (e.g., competitor comes up with better results first)
- Identify organization risks (e.g., department requesting project doesn’t have funding)
- Identify financial risks
- Identify technical risks
- Identify risks related to data and data sources

### CONTIGENCY PLANS

- Determine conditions under which each risk may occur
- Develop contingency plans

# Assess Situation
# Terminology

- Check for the existence of a previous glossary
- Talk to domain experts to understand their terminology
- Become familiar with the business terminology

WHO IS A CUSTOMER ?
ANYONE THAT BUYS OUR PRODUCTS
ENTITY WE HAVE A VALID CONTRACT WITH
LEADS IN OUR CRM
ENTRIES IN OUR AR DATABASE
ROWS IN DIM_CUST TABLE
HOW MANY CUSTOMERS DO WE HAVE ?
4,728
1,294
23,763
9,720
12,923
Dataedo /cartoon
Piotr@Dataedo

Assess Situation

# Costs and benefits

- Estimate costs for data collection
- Estimate costs of developing and implementing a solution
- Identify benefits (e.g., improved customer satisfaction, ROI, and increase in revenue)
- Estimate operating costs

Determine DM goals

# Data Mining goals

- Translate the business questions to DM goals (e.g., reduction of visiting frequency)
- Specify DM problem type (e.g., segmentation)

Determine DM goals

# Data Mining success criteria

- Specify criteria for model assessment (e.g., model accuracy, performance and complexity)
- Define benchmarks for evaluation criteria
- Specify criteria which address subjective assessment criteria (e.g., model explicability and insights provided by the model)

## Produce project plan

# Project plan

- Define the initial process and discuss the feasibility with all involved personnel
- Combine all identified goal and selected techniques in a coherent procedure
- Estimate the effort and resources needed to achieve and deploy the solution (e.g., Data understanding: 20-30%, Data preparation: 50-70%, Modeling: 10-20%, Deployment: 5-10%, Other phases: remaining)
- Identify critical steps and major iterations
- Mark decision and review points

Produce project plan

# Initial assessment of tools and techniques

- Create a list of selection criteria for tools and techniques
- Choose potential tools and techniques
- Evaluate appropriateness of techniques
- Review and prioritize applicable techniques according to the evaluation of alternative solutions

# Questions?

**Data Science for Marketing**

© 2021-2022 Nuno António (Rev. 2022-09-13)

Acreditações e Certificações

Instituto Superior de Estatística e Gestão da Informação
Universidade Nova de Lisboa