Uniwersytet Ekonomiczny we Wrocławiu

IT in Finance

Methodology of Data Mining

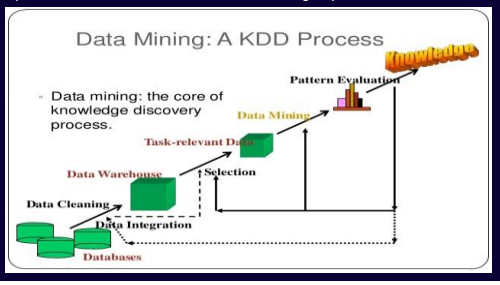
Jerzy KORCZAK
email: jerzy.korczak@ue.wroc.pl
http://www.korczak-leliwa.pl

1

What is Data Mining?

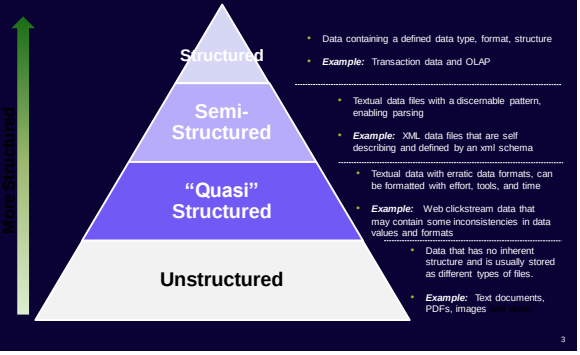
Many definitions

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns



2

Data Structures



- Data containing a defined data type, format, structure
- ***Example:*** Transaction data and OLAP

- Textual data files with a discernable pattern, enabling parsing
- ***Example:*** XML data files that are self describing and defined by an xml schema

- Textual data with erratic data formats, can be formatted with effort, tools, and time
- ***Example:*** Web clickstream data that may contain some inconsistencies in data values and formats

- Data that has no inherent structure and is usually stored as different types of files.
- ***Example:*** Text documents, PDFs, images

3

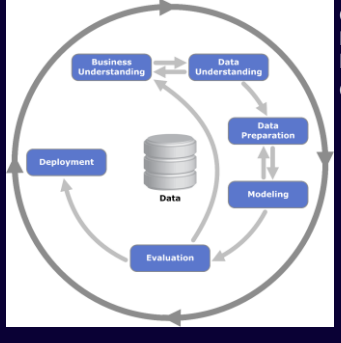
Business Requirements

Current Business Problems Provide Opportunities for Organizations to Become More Analytical & Data Driven

	Driver	Examples
1	Desire to optimize business operations	Sales, pricing, profitability, efficiency
2	Desire to identify business risk	Customer churn, fraud, default
3	Predict new business opportunities	Upsell, cross-sell, best new customer prospects
4	Comply with laws or regulatory requirements	Anti-Money Laundering, Fair Lending, Basel II

4

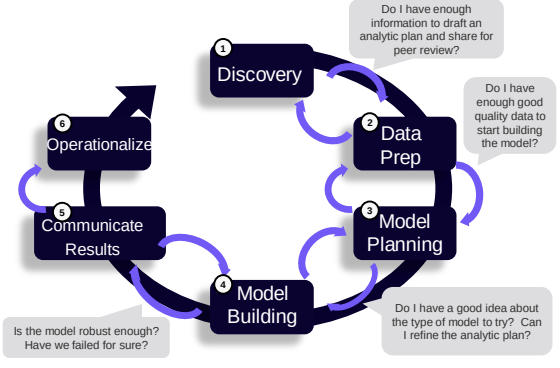
Methodology



Cross Industry Standard Process for Data Mining known by its acronym **CRISP-DM** *[ESPRIT, 1996].*

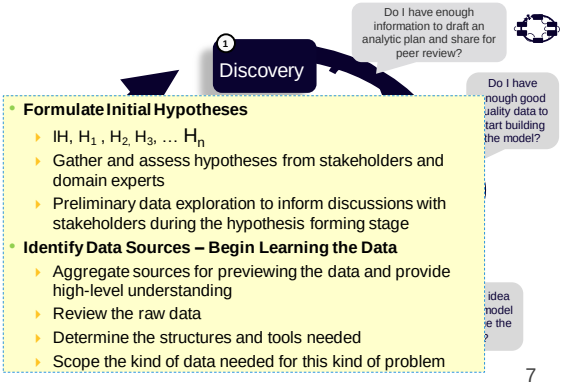
5

Data Analytics – Development

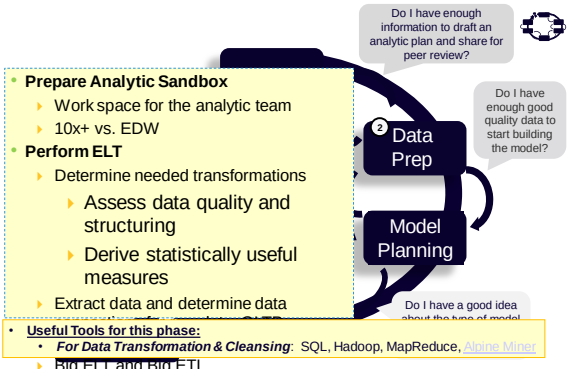


6

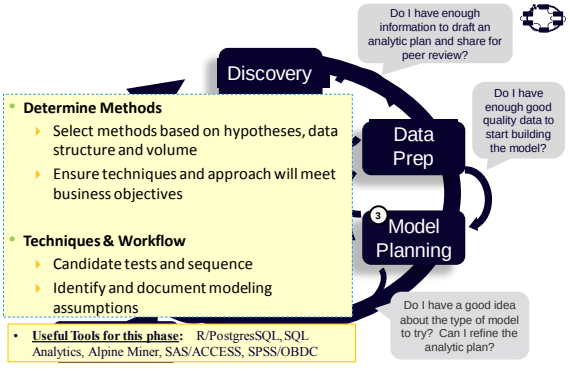
Data Analytics – Development (cont.)



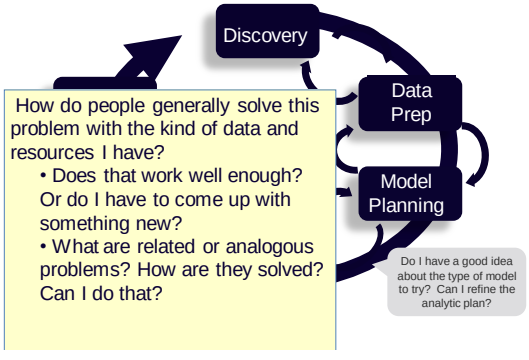
Data Analytics – Development (cont.)



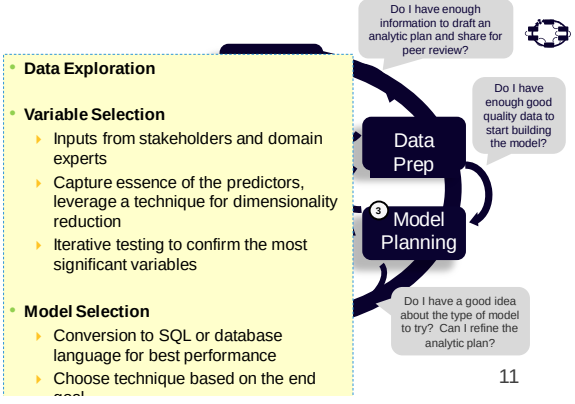
Data Analytics – Development (cont.)



Data Analytics – Development (cont.)



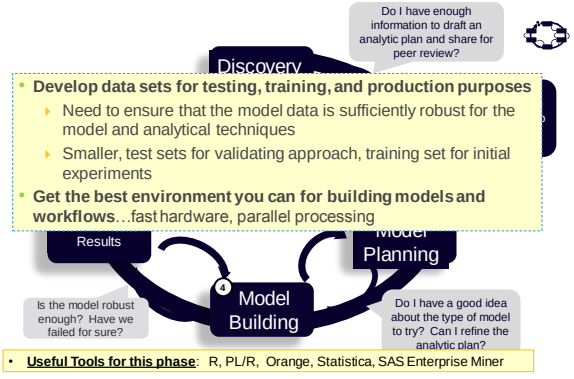
Data Analytics – Development (cont.)



Data Analytics – Development

The Problem to Solve	The Category of Techniques	Algorithms
I want to group items by similarity. I want to find structure (commonalities) in the data	Clustering	K-means clustering
I want to discover relationships between actions or items	Association Rules	Apriori
I want to determine the relationship between the outcome and the input variables	Regression	Linear Regression Logistic Regression
I want to assign (known) labels to objects	Classification	Naïve Bayes Decision Trees
I want to find the structure in a temporal process I want to forecast the behavior of a temporal process	Time Series Analysis	ACF, PACF, ARIMA
I want to analyze my text data	Text Analysis	Regular expressions, Document representation (Bag of Words), TF-IDF

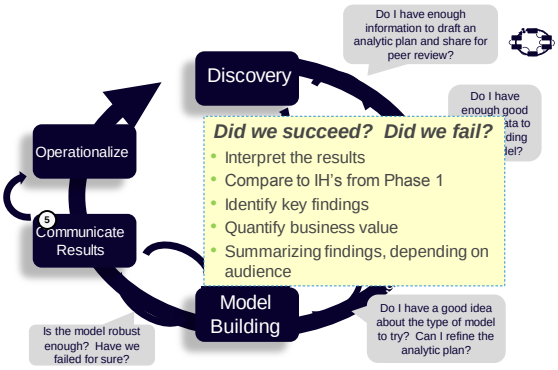
Data Analytics – Development (cont.)



- **Develop data sets for testing, training, and production purposes**
  - Need to ensure that the model data is sufficiently robust for the model and analytical techniques
  - Smaller, test sets for validating approach, training set for initial experiments
- **Get the best environment you can for building models and workflows**…fast hardware, parallel processing

- **Useful Tools for this phase**: R, PL/R, Orange, Statistica, SAS Enterprise Miner

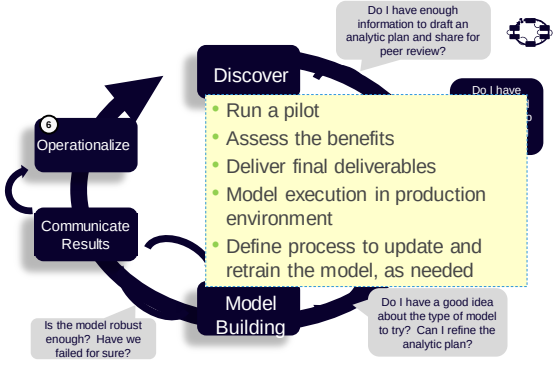
Data Analytics – Development (cont.)



Did we succeed? Did we fail?

- Interpret the results
- Compare to IH's from Phase 1
- Identify key findings
- Quantify business value
- Summarizing findings, depending on audience

Data Analytics – Development (cont.)



- Run a pilot
- Assess the benefits
- Deliver final deliverables
- Model execution in production environment
- Define process to update and retrain the model, as needed

Data Analytics – Core deliverables

Presentation for Project Sponsors

1. "Big picture" takeaways for executive level stakeholders
2. Determine key messages to aid their decision-making process
3. Focus on clean, easy visuals for the presenter to explain and for the viewer to grasp

Presentation for Analysts

1. Business process changes
2. Reporting changes
3. Fellow Data Scientists will want the details and are comfortable with technical graphs (such as ROC curves, density plots, histograms)

Code for technical people

Technical specs of implementing the code

Data Analytics – Key roles

Role	Description
Business User	Someone who benefits from the end results and can consult and advise project team on value of end results and how these will be operationalized
Project Sponsor	Person responsible for the genesis of the project, providing the impetus for the project and core business problem, generally provides the funding and will gauge the degree of value from the final outputs of the working team
Project Manager	Ensure key milestones and objectives are met on time and at expected quality.
Business Intelligence Analyst	Business domain expertise with deep understanding of the data, KPIs, key metrics and business intelligence from a reporting perspective
Data Engineer	Deep technical skills to assist with tuning SQL queries for data management, extraction and support data ingest to analytic sandbox
Database Administrator (DBA)	Database Administrator who provisions and configures database environment to support the analytical needs of the working team
Data Scientist	Provide subject matter expertise for analytical techniques, data modeling, applying valid analytical techniques to given business problems and ensuring overall analytical objectives are met