T. Evgeniou
Professor of Decision Sciences

# Data Analytics

Understand the world. Expand your world.

## Today's plan

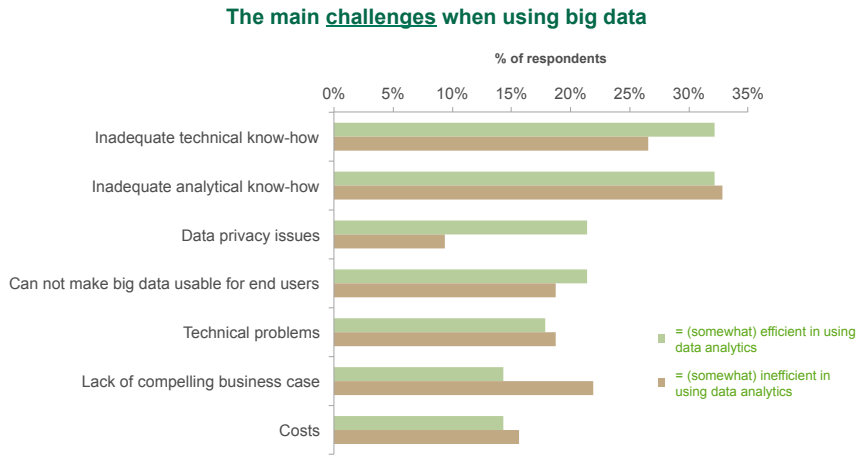
1. Course introduction
2. Data analytics projects
3. Set up

## Course grading

- Group assignment: 40%
- Individual Exercise: 40%
- Class participation: 20%

*How*

## Analyzing data effectively is challenging….




4

## Main Topics

1. Structuring data analytics projects
2. Key Statistical Techniques: Factor Analysis, Cluster Analysis, Classification
3. Some coding

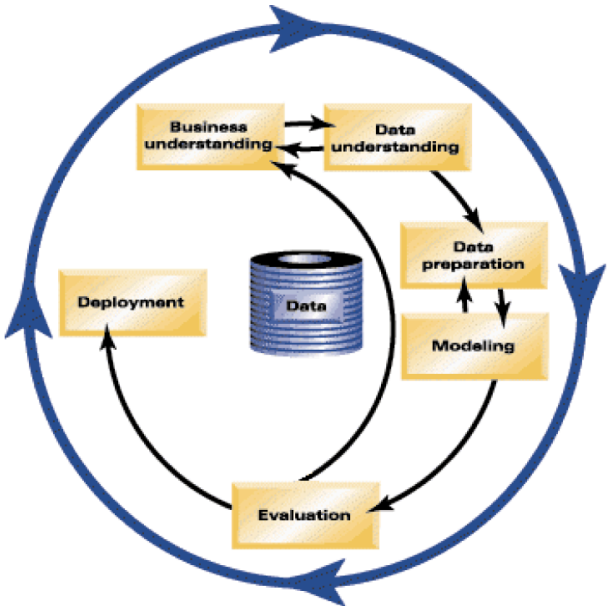
## Main Topics

1. Structuring data analytics projects
2. Key Statistical Techniques: Factor Analysis, Cluster Analysis, Classification
3. Some coding

## Example

**How could you use data for churn management?**

## The CRISP-DM Process



## Step 1 - Business Understanding: Sample Questions

- Describe in detail the problem you want to solve
  - What is defined as "churn"? Customers leaving the next 90 days? 60 days? Ever?
  - Are all customers equally treated?
- Specify expected benefits in business terms
  - What % of churns do we need to predict?
  - Is there a current practice in place? What is a benchmark performance?
  - What is the cost of retention? How does this affect our problem definition?
- Identify key individuals in the organization
  - Who manages churn now and how?
  - Who needs to be involved to activate the solution? Customer support? Call center? Marketing?
- Identify parts of the problem fitting with known tools
  - Is there a regression "hidden" somewhere?
  - Other methods?

**A lot of good statistical analysis is directed at solving the wrong business problem.**

## Step 2 - Data Understanding: Sample Questions

- What data is available?
  - Are there any relevant external data sources?
  - What could possibly affect a customer decision to leave? Is there data about that?
  - Which variables should be used?
- How much history is required?
  - Have there been some major changes in our business/industry recently?
- What is the right level of granularity?
  - Individual or family level? Daily or weekly?
- What data quality issues do we have?
  - Do missing values indicate something?
  - How do we handle non-numeric data?
- Simple hypotheses generation:
  - How do we expect specific variables to affect the solution?

## Step 3 - Data Preparation: Sample Questions

- Merge all data relevant sources
  - Ensure time or any other alignment
- Deal with data quality issues
  - Handle non-numeric data
  - Handle missing values
  - Handle data errors
  - Understand outliers
- Feature engineering:
  - Derive new (simple) features (e.g. "customer on top/bottom 20%)
- Split data in training and testing
  - How will the solution be used in practice? Can we simulate it?

## Step 4 - Modeling: Sample Questions

- Start with simple analyses:
  - Descriptive Statistics and Visualization
- Identify sub-problems fitting with analytics tools:
  - Can we group variables that are highly correlated? (factor analysis)
  - Do we need to develop different solutions for different segments? (clustering)
  - Do we predict binary outcomes (classification)? What is the target variable in that case?
- Estimate and assess model parameters:
  - Are they statistically valid?
  - Do they make sense?

## Step 5 - Evaluation: Sample Questions

- Measure various performance metrics:
  - What is the false positive and false negative rate? What matters most?
  - Lift curve, ROC curve, profit curve, etc
- Rank the candidate models/solutions
- Is there any overfit?
- Are the results easy to explain?
  - Highlight particularly novel or unique findings
- Do the analyses, our judgment, and our business criteria all agree?

## Step 6 - Deployment: Sample Questions

- Who needs to be involved in deployment?
- What is the data pipeline?
  - How are data sources and IT systems integrated?
  - How are data failures handled?
- How to test the solution before full deployment?
  - A/B testing setup
- How do we know our solution/model expired?
  - What metrics do we monitor?

## This is an ITERATIVE process!

- Revisit business objectives.
- Define new objectives.
- Gather and evaluate new data.
- Test new models.
- Test many variations.

***The ability to develop easy to replicate, reproduce, modify, and iterate solutions is key***

## Today's plan

1. Course introduction
2. Data analytics projects
3. Set up

## Setup: Key Software Tools

1. R and Rstudio
   1. .R files for "pure" code
   2. .Rmd files for documents
   3. Shiny for Interactive Documents/Tools
2. Github for collaboration (alternative: dropbox)

## Basic Types of Questions and Tools

1. Market Basket Analysis: which pairs of products are typically sold together? – "On Friday evenings, shoppers who buy diapers also buy beer".
2. Factor Analysis: Finding important dimensions ("factors") that summarize your data, and visualizing your data
3. Clustering: What are the main types of customers we have?
4. Regression Modeling: What are variables that drive a specific outcome?
5. Classification: How can we differentiate between the "high value" and "low value" customers?

## To do for next time

- Read the Boating Case
- Explore the case data
- Class readings
- Make sure you finish the session 1 exercise and you can "compile" the Market Segmentation Process file

INSEAD

The Business School
for the World®