Introduction to Data Science
Summary of this lesson

„We are drowning in information, but starving for knowledge”
-John Naisbett

What is knowledge?

*This lesson refers to chapters 1 and 2 of the GIDS book
Content of this lesson

- What is Data Science?
- The Data Science Process
- Data Science: An Example
What is Data Science?
**Data**

- refer to single instances (single objects, people, events, points in time, etc.)
- describe individual properties
- are often available in large amounts (databases, archives)
- are often easy to collect or to obtain (e.g., scanner cashiers in supermarkets, Internet)
- do not allow us to make predictions or forecasts

**Knowledge**

- refers to *classes* of instances (*sets* of objects, people, events, points in time, etc.)
- describes general patterns, structures, laws, principles, etc.
- consists of as few statements as possible
- is often difficult and time consuming to find or to obtain (e.g., natural laws, education)
- allows us to make predictions and forecasts
Criteria to assess knowledge

- **correctness** (probability, success in tests)
- **generality** (domain and conditions of validity)
- **usefulness** (relevance, predictive power)
- **comprehensibility** (simplicity, clarity, parsimony)
- **novelty** (previously unknown, unexpected)
What is Data Science?

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.

Knowledge discovery in databases (KDD) is the process of (semi-)automatic extraction of knowledge from databases which is valid, previously unknown, and potentially useful.
Some Clarity about Words

- \textit{(semi)-automatic}: no manual analysis, though some user interaction required
- \textit{valid}: in the statistical sense
- \textit{previously unknown}: not explicit, no “common sense knowledge“
- \textit{potentially useful}: for a given application
- \textit{structured data}: numbers
- \textit{unstructured data}: everything else (images, texts, networks, chem. compounds, …)
Valid?

99.98%
Valid?
customer age ∈ [18, 150]
(in 9,999 of 10,000 cases)
Previously Unknown?

A => B (in 100% of all cases)
Previously Unknown?

Pregnant => Female
Useful?

A => B

(with $s = 0.81\%$ and $c = 21.3\%$)
Useful?

Beer => Diapers
(with s =0.81% and c = 21.3%)
Valid, Interesting, and Useful?

Books A and B => Book C

(with $s = 0.81\%$ and $c = 21.3\%$)
The Data Science Process
The Data Science Process

- **SEMMA**
  - Sample, Explore, Modify, Model, Assess

- **CRISP-DM**
  - Cross Industry Standard Process for Data Mining

- **KDD**
  - Knowledge Discovery in Databases

<table>
<thead>
<tr>
<th>Project Understanding</th>
<th>Data Understanding</th>
</tr>
</thead>
<tbody>
<tr>
<td>What exactly is the problem, the expected benefit? How would a solution look like? What is known about the domain?</td>
<td>What data do we have available? Is the data relevant to the problem? Is it valid? Does it reflect our expectations? Is the data quality, quantity, recency sufficient?</td>
</tr>
<tr>
<td>Does data suit problem?</td>
<td>Which data should we concentrate on? How is the data best transformed for modeling? How may we increase the data quality?</td>
</tr>
<tr>
<td>Data Preparation</td>
<td>Modeling</td>
</tr>
<tr>
<td>Technical quality improvable? Likely</td>
<td>What kind of model architecture suits the problem best? What is the best technique/method to get the model? How good does the model perform technically?</td>
</tr>
<tr>
<td>Evaluation</td>
<td>Deployment</td>
</tr>
<tr>
<td>Business objective achieved?</td>
<td>How good is the model in terms of project requirements? What have we learned from the project?</td>
</tr>
<tr>
<td>Close project</td>
<td>How is the model best deployed? How do we know that the model is still valid?</td>
</tr>
</tbody>
</table>
It always starts with some data …
The Data Science Life Cycle

- **Creation**
  - Blend & Transform
  - Model & Visualize

- **Production**
  - Validate & Deploy
  - Consume & Interact
  - Optimize & Capture
  - Monitor & Update

- **Production Process**
Problem Categories

- **Classification**
  - Predict experiment outcome falling into a finite number of possible results
  - *How credit-worthy is this customer?* Very / Enough / Not enough / Absolutely not
  - *Will this customer respond to our mailing?* Yes / No

- **Regression**
  - Predict numeric values
  - *How will the EUR/USD exchange rate develop?*
  - *What will be the price of this washing machine next week?*

- **Clustering, Segmentation**
  - Group similar cases in order to get overview, detect outliers, or get insights on the data structure
  - *Do my customers separate into different groups?*
  - *How many operating points does the machine have, and what do they look like?*
Problem Categories

- **Association Analysis**
  - Find correlations to better understand the interdependencies of all the attributes
  - Focus in the full record (all the attributes) rather than on a single target variable
  - *Which optional equipment of a car often goes together?*
  - *How do the various qualities in a car influence each other?*

- **Deviation Analysis**
  - Knowing the trend of the data, find subgroups that behave differently
  - *Under which circumstances does the system behave differently?*
  - *Which properties do those customers - who do not follow the crowd - share?*
Data Science: an Example
Example

- Dataset from a hypothetical supermarket chain
  - Customers
  - Products
  - Purchases

- Three tasks
  - Divide customers into different groups according to their purchase behaviour
  - Identify connections between products to implement cross-selling campaigns
  - Helping design a marketing campaign to increase purchases

- Two approaches
  - Naive approach lead by common sense
  - Sound approach using DS techniques
Data Understanding and Pattern Finding: Customer Segmentation

- **Naive Approach**
  - Aggregate purchases to respective customer
  - Join with the customer details
  - No interesting relations highlighted

<table>
<thead>
<tr>
<th>Cluster id</th>
<th>Age</th>
<th>Customer revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>46.5</td>
<td>€ 1,922.07</td>
</tr>
<tr>
<td>2</td>
<td>39.4</td>
<td>€ 11,162.20</td>
</tr>
<tr>
<td>3</td>
<td>39.1</td>
<td>€ 7,279.59</td>
</tr>
<tr>
<td>4</td>
<td>46.3</td>
<td>€ 419.23</td>
</tr>
<tr>
<td>5</td>
<td>39.0</td>
<td>€ 4,459.30</td>
</tr>
</tbody>
</table>
Sound Approach

- Check values for the string attributes (name, employment..)
- Check and add constraints to numeric attributes (e.g. Age between 18-100)
- Look for misleading information (e.g. In the dataset a missing birthdate was by default set to 1970. If not handled properly, this information can lead to errors)
- Use average basket price as estimator for the value of a customer
- Use average number of purchases per month as further estimator
- Apply normalization to average attributes magnitudes
Explanation Finding: Find Product Dependencies

- **Naive Approach**
  - Run Association Rule Mining algorithm with default setting
  - Consider Product ID (differentiating each product)
  - Unintuitive and unuseful result
  - Rules have high confidence but low support values

- **Sound Approach**
  - Consider product categories
  - Rules match with well-known facts
  - Monitor combinations on regular basis

- 'foie gras' (p1231) ← 'champagne Don Huberto' (p2149),
  'truffle oil de Rossini' (p578) [s=1E-5, c=75%]

- 'Tortellini De Cecco 500g' (p3456)
  ← 'De Cecco Sugo Siciliana' (p8764) [s=1E-5, c=60%]

- tomatoes ← capers, pasta [s=0.007, c=32%]
- tomatoes ← apples [s=0.013, c=22%]
Predicting the Future: Forecast customers reactions to coupon mailings

### Naive Approach
- No detailed analysis
- Send coupon with discounts after a certain purchase amount
- Just monitor the results
- Fail: customers only combine shopping trips, no additional revenues
- The data analyst is in the end fired

### Sound Approach
- Discriminate valuable customers => exploit earlier segmentation
- Derive meaningful attributes, e.g. Customers underperforming on specific category, distance
- Build black box classifier model
Thank you