EMIS/DS 1300: A Practical Introduction to Data Science

Slides by Michael Hahsler
Agenda

What is Data Science?

Who is a Data Scientist?

The Data Science Process

Data Science Tools

Visualization

Ethics, Privacy and Security Issues
Data + Science = Results?
What is Data Science?

“Data science is a concept to unify statistics, data analysis, machine learning and their related methods in order to understand and analyze actual phenomena with data.”

[Hayashi, Chikio "What is Data Science?"]
What is Statistics?

• “Statistics is a branch of mathematics dealing with data collection, organization, analysis, interpretation and presentation.” [Wikipedia]

• Techniques:
  • Design of experiments (sampling)
  • Descriptive statistics
  • Statistical inference (estimation, testing)
What are Analytics and Data Mining?

- Analytics and Data Mining focus on the discovery and the communication of meaningful patterns in data.

- Analytics relies on the simultaneous application of statistics, computer programming and optimization to quantify performance.

- Analytics often favors data visualization to communicate insight.

- Data Mining focuses on predictive models.

[Wikipedia]
What are Artificial Intelligence and Machine Learning?

- **AI** is the study of intelligent agents, devices that perceive its environment and takes actions that maximize its chance of successfully achieving its goals.

- **Machine learning (ML)** is the study of algorithms and statistical models that computer systems use to progressively improve their performance on a specific task. The goal is to make accurate predictions or decisions without being explicitly programmed to perform the task.

[Wikipedia]
Why do companies care about Data

Businesses collect and warehouse lots of data:
- Bank/credit card transactions
- Web data, e-commerce
- Social media
- Internet of things (IOT)

Computers are cheaper and more powerful.
- SaaS/IaaS/PaaS

Competition to provide better services.
- Mass customization and recommendation systems
- Targeted advertising
- Improved logistics
Assignment: Why should you care?

Answer the following questions.

Examples of information that is collected about you in your daily life.

Who do you think collects the information?

Who do you think has access to the information?

How may this information be used/misused?
Some Applications of Data Science

- Uber
- Airbnb
- Netflix
- Amazon
- Logistics
- Banking, loans, insurance
- Pharmaceutical industry
- Healthcare
- Sports
Agenda

- What is Data Science?
- Who is a Data Scientist?
- The Data Science Process
- Data Science Tools
- Visualization
- Ethics, Privacy and Security Issues
Who does all this?  
And who gets the big paycheck?
Good luck finding this person! Probably a team effort!

Source: T. Stadelmann, et al., Applied Data Science in Europe
What Does a Data Scientist Do?

- Identifying data analytics opportunities.
- Find/collect the correct data sets and variables.
- Clean the data and ensure accuracy and completeness.
- Decide on appropriate models and algorithms to mine the data. Identify patterns and trends.
- Interpret the results to data to discover solutions and opportunities.
- Communicate findings to stakeholders using visualization and prototypes.
Agenda

What is Data Science?

Who is a Data Scientist?

The Data Science Process

Data Science Tools

Visualization

Ethics, Privacy and Security Issues
How to do a Data Science project?
CRISP-DM Reference Model

- **CRISP-DM Reference Model**
- Cross Industry Standard Process for Data Mining
- De facto standard for conducting data mining and knowledge discovery projects.
- Defines tasks and outputs.
- Now developed by IBM as the Analytics Solutions Unified Method for Data Mining/Predictive Analytics (ASUM-DM).
- SAS has SEMMA and most consulting companies use their own process.
### Tasks in the CRISP-DM Model

**Business Understanding**
- **Determine Business Objectives**
  - Background
  - Business Objectives
  - Business Success Criteria
- **Assess Situation**
  - Inventory of Resources Requirements, Assumptions, and Constraints
  - Risks and Contingencies
  - Terminology
  - Costs and Benefits

**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**
- **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
  - Dataset
  - Dataset Description

**Modeling**
- **Select Modeling Techniques**
  - Modeling Technique
  - Assumptions
- **Generate Test Design**
  - Test Design
- **Build Model**
  - Parameter Settings
  - Models
  - Model Descriptions
- **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

---

*Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model*
The Data Science Process

Source: The Data Science Process, Springboard
https://www.kdnuggets.com/2016/03/data-science-process.html
Tools

2018 Magic Quadrant for Data Science and Machine Learning Platforms

Gartner®

As of January 2018 © Gartner, Inc
Tools - Popularity

Tools - Types

• **Data**: Relational databases *(SQLite)*, NoSQL databases
• **Spreadsheet**: Excel, Google Sheets
• **Visualization**: Tableau, Microsoft Power BI, SAS jmp

• **Data Science Platforms**
  • Simple graphical user interface
  • Process oriented
  • Programming oriented
Tools
Simple GUI

• **Weka**: Waikato Environment for Knowledge Analysis (Java API)

• **Rattle**: GUI for Data Mining using R
Tools - Process oriented

- SAS Enterprise Miner
- IBM SPSS Modeler
- RapidMiner
- Knime
- Orange
Tools - Programming oriented

- Python
  - Scikit-learn, pandas
  - IPython, notebooks

- R
  - Rattle for beginners
  - RStudio IDE, markdown, shiny
  - Microsoft Open R

→ Both have similar capabilities. Slightly different focus:
  - R: Statistical computing and visualization
  - Python: Machine learning and big data
Data Warehouse
Agenda

- What is Data Science?
- Who is a Data Scientist?
- The Data Science Process
- Data Science Tools
- Visualization
- Ethics, Privacy and Security Issues
Data Visualization

Infoviz is a field of its own.

Eat **fruits** when they are in season!!!
Agenda

What is Data Science?

Who is a Data Scientist?

The Data Science Process

Data Science Tools

Visualization

Ethics, Privacy and Security Issues
Legal, Privacy and Security Issues

Questions:
- Are we allowed to **collect** the data?
- Are we allowed to **use** the data?
- Is it ethical to use and **act** on the data?
- Is **privacy** preserved in the process?

**Problem:** Internet is global, but legislation is local!
EU law on data protection and privacy for all individuals within the European Union (EU) and the European Economic Area (EEA)
Implementation: 25 May 2018

Personal data may not be processed unless there is at least one legal basis to do so. Lawful purposes are:

- Consent by the individual (Opt-in)
- Legal obligations of the data controller
- Protect the vital interests of a data subject or another individual
- To perform a task in the public interest or in official authority
- For the legitimate interests of a data controller

- Applies to US companies doing business in the EU.
- California passed a similar bill called The California Consumer Privacy Act of 2018.
Effective: May 25, 2018

With over 12,500 members from around the globe, Institute for Operations Research and the Management Sciences (INFORMS) is the leading international association for professionals in operations research and analytics.
Privacy

BERLIN — Angry Birds, the top-selling paid mobile app for the iPhone in the United States and Europe, has been downloaded more than a billion times by devoted game players around the world, who often spend hours slingling squawking fowl at groups of egg-stealing pigs.

When Jason Hong, an associate professor at the Human-Computer Interaction Institute at Carnegie Mellon University, surveyed 40 users, all but two were unaware that the game was storing their locations so that they could later be the targets of ads....
Pokémon Go’s constant location tracking and camera access required for gameplay, paired with its skyrocketing popularity, could provide data like no app before it.

“Theyir privacy policy is vague,” Hong said. “I’d say deliberately vague, because of the lack of clarity on the business model.”

The agreement says Pokémon Go collects data about its users as a “business asset.” This includes data used to personally identify players such as email addresses and other information pulled from Google and Facebook accounts players use to sign up for the game.

If Niantic is ever sold, the agreement states, all that data can go to another company.
How hackers stole millions of credit card records from Target

How did the cyberattack on Target, which resulted in the theft of millions of records, take place?


Banking, Social Security info of more than 1.4 million people exposed in hack involving Washington state auditor
