
Universitat de Barcelona
Data Science & Big Data
A part-time course to train the new generation of data scientists.
www.ub.edu/datascience

Petia Radeva

Oriol Pujol

Francesc Dantí

Jordi Vitrià

Laura Igual

Eloi Puertas

Santi Seguí

Universitat de Barcelona

AMB LA COL·LABORACIÓ DE

BBVA DATA & ANALYTICS

TELEVISIÓ DE CATALUNYA

TÜVRheinland

synergic partners

Barcelona Media

Kernel analytics

ANAXOMICS Systems biology solutions

eDreams

COUNTEREST

Introduction

Universitat de Barcelona's Data Science and Big Data course offers students a program that covers the concepts and tools you will need throughout the entire data science pipeline: asking the right questions; wrangling and cleaning data; generating hypothesis; making inferences; visualizing data; assessing solutions; and building data products.

Schedule

March 17, 2015 - July 14, 2015, every Tuesday 18h-21h and Thursday 18h-21h

Workload

12 hours per week (including lectures and homework)

Location

Edifici Històric de la Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007, Barcelona.

Universitat de Barcelona

AMB LA COL·LABORACIÓ DE

BBVA DATA & ANALYTICS

3 TELEVISIÓ DE CATALUNYA

TÜVRheinland

synergic partners

Barcelona Media

Kernel analytics

ANAXOMICS Systems biology solutions

eDreams

COUNTEREST

Requirements:

The program is specially designed for students with a background in computer science, mathematics, and applied statistics, but other scientific and engineering backgrounds can be considered.

We will require to follow lessons and complete class exercises using personal laptops. You will not be able to complete all your assignments in class if you rely solely on desktop equipment at home.

Information:

Campus Virtual UB (after official enrollment):

campusvirtual2.ub.edu

Now:
datascienceub.sitedrop.com

password: datascienceub2015

Content (I)

Introduction: Understanding “data science” and “big data” concepts. Programming in Python. Use Python tools for exploratory analysis and reproducible research.

Data Gathering, Cleaning & Storage: Work with various data and file formats. Use tools for web data scraping. Write scripts for data. Extract features from unstructured data. Understand and use tools for representing natural language data.

Data Science Basic Toolbox: Use the Python ecosystem for numerical computation and tabular data analysis. Use visualization techniques for data exploration. Use software development and engineering tools.

Computational Statistics for Data Exploration: Use basic statistics descriptors to make initial hypotheses. Use distributions to represent data populations. Estimate parameters from data. Use regression models to compute quantitative responses. Learn about types and assumptions of predictive model’s performance criteria, and principles of validation. Assess the performance of a predictive model by using different methods and metrics.

Content (II)

Bayesian Statistics: Learn about representing states of the world in terms of degrees of belief. Identify prior beliefs about what results are likely for a problem, and then update those according to the data we collect.

Machine Learning: Understand different approaches to machine learning. Use model selection and model evaluation techniques. Recognize appropriate uses of linear models. Recognize appropriate uses of Naïve Bayes and Random Forests. Evaluate the use of ensemble methods. Search for the best set of features. Use dimensionality reduction for data exploration and representation. Define clustering and identify appropriate use cases.

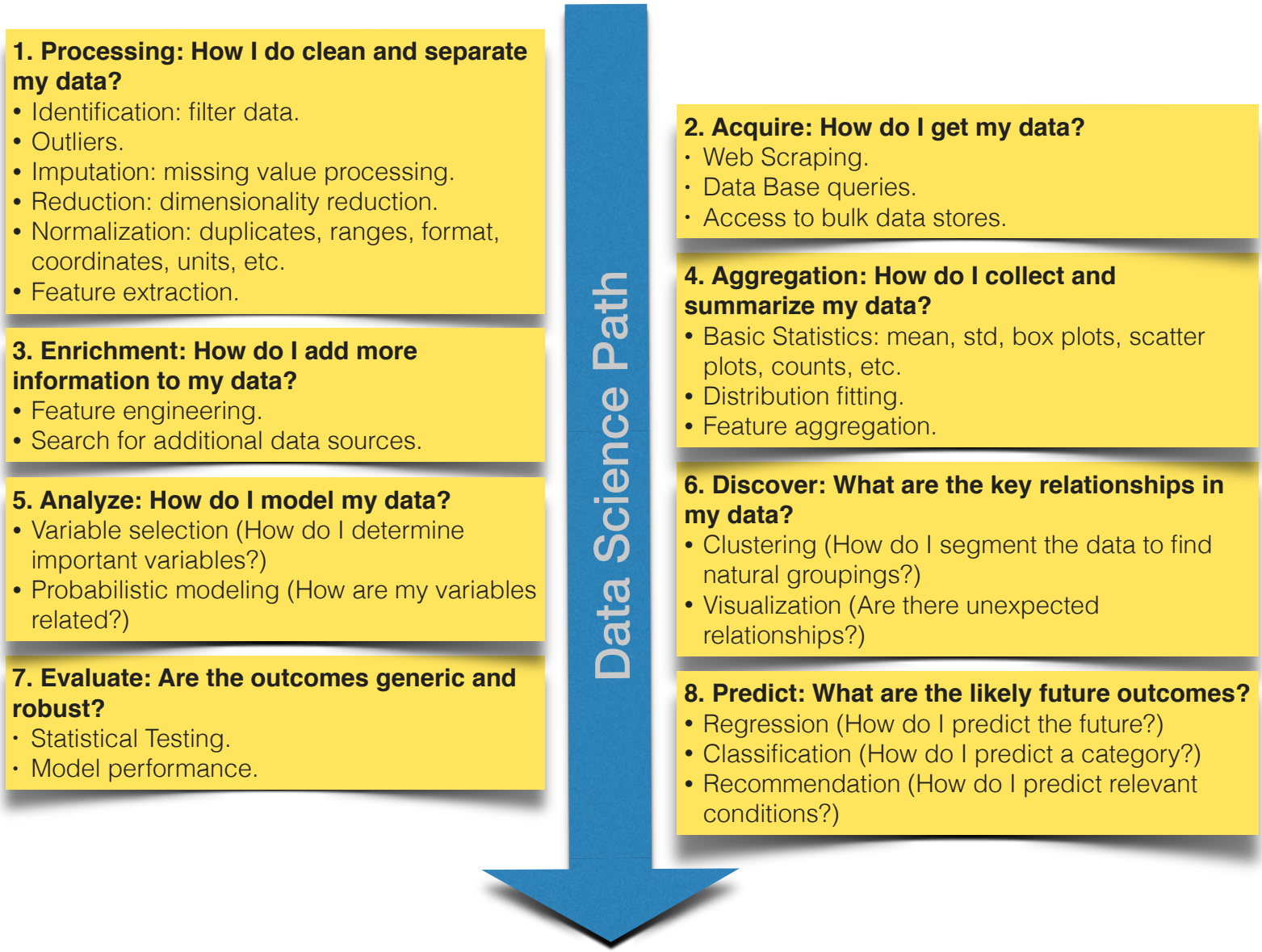
Recommenders: Identify user-based and item-based collaborative filtering techniques. Given a scenario, develop a collaborative filtering implementation.

Graph analysis: Explore and visualize networks. Analyze network data.

Big Data Analysis: Leverage multicore architectures. Run a distributed machine learning algorithm.

Capstone Project: An important part of the course is developing a solution to a real problem, including data collection, trying alternative solutions, describing statistical methods, and getting insights from data.

Raw Data



1. Processing: How I do clean and separate my data?
- Identification: filter data.
- Outliers.
- Imputation: missing value processing.
- Reduction: dimensionality reduction.
- Normalization: duplicates, ranges, format, coordinates, units, etc.
- Feature extraction.

2. Acquire: How do I get my data?
- Web Scraping.
- Data Base queries.
- Access to bulk data stores.

3. Enrichment: How do I add more information to my data?
- Feature engineering.
- Search for additional data sources.

4. Aggregation: How do I collect and summarize my data?
- Basic Statistics: mean, std, box plots, scatter plots, counts, etc.
- Distribution fitting.
- Feature aggregation.

5. Analyze: How do I model my data?
- Variable selection (How do I determine important variables?)
- Probabilistic modeling (How are my variables related?)

6. Discover: What are the key relationships in my data?
- Clustering (How do I segment the data to find natural groupings?)
- Visualization (Are there unexpected relationships?)

7. Evaluate: Are the outcomes generic and robust?
- Statistical Testing.
- Model performance.

8. Predict: What are the likely future outcomes?
- Regression (How do I predict the future?)
- Classification (How do I predict a category?)
- Recommendation (How do I predict relevant conditions?)

Insights

Calendar & Masterclasses

17/03/2015: Introducció – Jordi Vitrià

24/03/2015: Data Science Toolbox – Eloi Puertas

26/03/2015: Data gathering and storage – Oriol Pujol

7/04/2015: Software Carpentry – Eloi Puertas

9/04/2015: Recommenders – Santi Seguí

14/04/2015: Computational Statistics – Petia Radeva

16/04/2015:Statistical Estimation – Jordi Vitrià

21/04/2015: Regression – Laura Igual

28/04/2015: Bayesian Estimation – Jordi Vitrià

30/04/2015: Master Class – BBVA (Dani Villatoro)

5/05/2015: Machine Learning I – Oriol Pujol

7/05/2015: Machine Learning II – Oriol Pujol

12/05/2015: Master Class – TV3 (Daniel Giribet)

14/05/2015:Machine Learning III – Petia Radeva

19/05/2015: Big Data I – Francesc Dantí

21/05/2015: Big Data II – Francesc Dantí

26/05/2015: Master Class – BSC (Jordi Nin) + Presentació projectes

28/05/2015: Graph Analysis -Laura Igual

2/06/2015: Visualization – Santi Seguí

4/06/2015: Master Class – Kernel Analytics (Pau Agulló)

Activities

twitter: @datascienceub | @databeersbcn

linkedin: datascienceub group

Evaluation

Capstone Project: An important part of the course is the IPython process notebook. This notebook details your steps in developing a solution to a real problem, including how you collected the data, alternative solutions you tried, describing statistical methods you used, and the insights you got.