Data Science Lab - 1

Academic year 2019-2020

Lecturer Falco J. Bargagli Stoffi

IMT School for Advanced Studies Lucca & KU Leuven

General Overview

- The aims of this course are:

  1. introduce you to main concepts of *Data Science*
  2. to introduce you to `R` language fundamentals and basic syntax

- As, by now, you will be probably sick and tired of *frontal lectures* this course is thought as a fully applied class

Outline of the Classes

1. Introduction to Data Science + Basic R
2. Exploratory Data Analysis in R
3. Data Modeling in R
4. Predictive Analysis in R
5. Causal Machine Learning in R

What is Data Science?

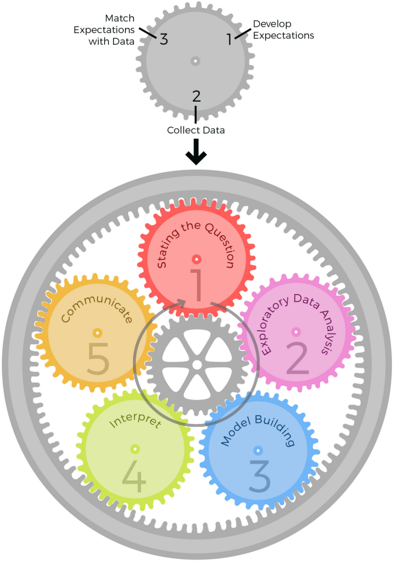
- Xiao-Li Meng, director of the HDSR, introduces the concept by stating what DS is not (Meng, 2019):
  1. DS is not just machine learning or statistics (Lo et al., 2019)
  2. DS is not just about predictions (Sanders, 2019)
  3. DS is not all about data analysis (Wing, 2019)
  4. DS does not sit merely within STEM (Leonelli, 2019)
  5. DS is not a single discipline

What is Data Science?

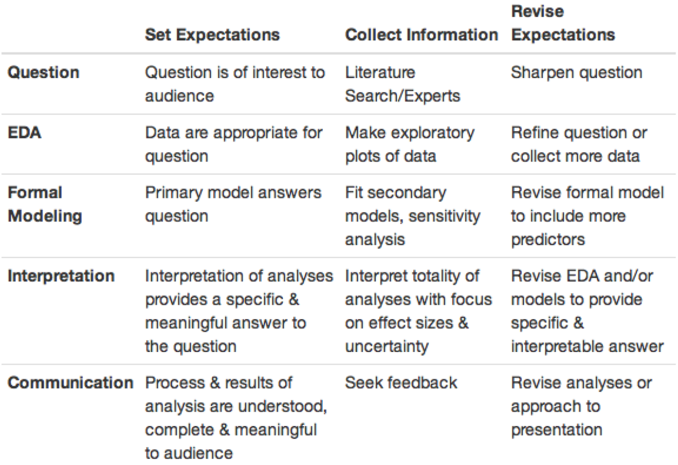
Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.

- The key components are:
  1. Integrated and multi-disciplinary approach
  2. Scientific method
  3. Knowledge extraction from various data sources

Epicycles of Analysis (Peng and Matsui, 2016)



Epicycles of Analysis (Peng and Matsui, 2016)



	Set Expectations	Collect Information	Revise Expectations
Question	Question is of interest to audience	Literature Search/Experts	Sharpen question
EDA	Data are appropriate for question	Make exploratory plots of data	Refine question or collect more data
Formal Modeling	Primary model answers question	Fit secondary models, sensitivity analysis	Revise formal model to include more predictors
Interpretation	Interpretation of analyses provides a specific & meaningful answer to the question	Interpret totality of analyses with focus on effect sizes & uncertainty	Revise EDA and/or models to provide specific & interpretable answer
Communication	Process & results of analysis are understood, complete & meaningful to audience	Seek feedback	Revise analyses or approach to presentation

Research Question

- 6 types of RQs:

1. Descriptive
2. Explanatory
3. Inferential
4. Predictive
5. Causal
6. Mechanistic

A Good Research Question

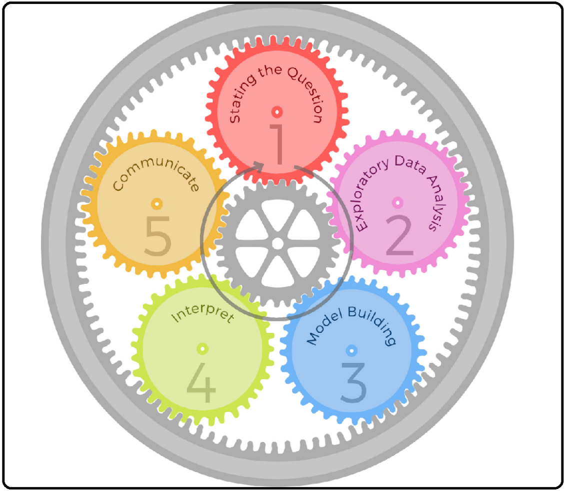
- A good RQ has the following characteristics:

  1. It should be of interest to the scientific community
  2. It should also not already been answered
  3. It should stem from a plausible framework
  4. You should be able to answer to it
  5. It should be specific
  6. It should be not based on a large set of assumptions (Occam's razor)

- Always discuss your research ideas with your colleagues

From the RQ to Data Analysis

Programmming



Why R for DS?

1. Optimized for DS (`tidyverse`)
2. Open source software
3. Robust statistical software
4. Most advanced statistical learning packages
5. Advanced plotting libraries (`ggplot2`)
6. Integrated with other programming languages (`reticulate`)
7. Optimal choice for reproducible research (`markdown`)
8. Widely used in industry

Bibliography

Leonelli, S. (2019). Data Governance is Key to Interpretation: Reconceptualizing Data in Data Science. *Harvard Data Science R.*

Lo, A. W., Siah, K. W., Wong, C. H. 2019. Machine Learning with Statistical Imputation for Predicting Drug Approvals. *Harvard Data Science Review*.

Meng, X.L., 2019. Data Science: An Artificial Ecosystem. *Harvard Data Science Review*.

Peng, R.D. and Matsui, E., 2015. *The Art of Data Science. A Guide for Anyone Who Works with Data.* Skybrude Consulting, LLC.

Sanders, N. (2019). A Balanced Perspective on Prediction and Inference for Data Science in Industry. *Harvard Data Science Review*.

Wing, J. M. (2019). The Data Life Cycle. *Harvard Data Science Review*.

Wickham, H. and Grolemund, G., 2016. *R for data science: import, tidy, transform, visualize, and model data.* O'Reilly Media, Inc..