The Data Science workflow

Gianluca Campanella

High-level view

Business goal

↓

Testable hypothesis

↓

Experimentation and modelling

High-level view

Research question

$\downarrow$

Obtain $\longleftrightarrow$ Explore $\longleftrightarrow$ Model

$\downarrow$

Summarise / Operationalise

Time allocation

Which takes longer?

Define the research question

What to do

- Identify the problem and why it should be solved
- Frame it in the context of data collection

What to ask

- Which metrics do I need to improve?
- Which are possible actions to solve the problem?
- What is the benefit of solving the problem?

Obtain the data

What to do

- Measure the gap between ideal and available
- Think about assumptions and limitations

What to ask

- Are there enough data?
- Are they relevant to the research question?
- Can they be trusted?

Explore the data

What to do

- Data dictionary and any other documentation
- Descriptive statistics and visualisations

What to ask

- What kind of simple visualisations can I use?
- Which data types and distributions?
- Are there missing values or outliers?

Model the data

What to do

- Model selection and fitting
- Focus on inference and/or prediction

What to ask

- What is an appropriate model for the data?
- How can I evaluate model performance?
- Can the model be refined?

Modelling misconceptions

Most well-executed Data Science projects don't…

- Use complicated tools
- Fit complicated models

Instead, they do…

- Focus on solving the problem
- Use appropriate — not necessarily big! — data
- Use relatively standard models

The 80—20 rule of modelling

- The first reasonable thing you can do goes 80% of the way
- Everything after that is to get the remaining 20%…
  often at additional cost!

The 80—20 rule of modelling

- The first reasonable thing you can do goes 80% of the way
- Everything after that is to get the remaining 20%…
  often at additional cost!

Is it worth it?

Wait a second…

We are not done yet!

Summarise the findings

What to do

- Storytelling and visual aids to interpretation
- Communicate assumptions and limitations

What to ask

- How can I communicate results effectively?
- What format should I adopt?
- Who are my audience?

Operationalise

What to do

- System integration
- Monitoring and maintenance

What to ask

- What (visual) outputs do I care about?
- How often does the model need retraining?
- Do we need to think about scalability?