

The Data Science workflow

Gianluca Campanella

The Data Science workflow

1. Define the **research question**
2. **Get** the data
3. **Explore** the data
 - (Re)format, clean, merge, stratify...
 - Identify trends and outliers
4. **Model** the data
 - Select and build model(s)
 - Evaluate and refine model(s)
5. **Summarise** the results
 - Condense findings into recommendations
 - Describe assumptions and limitations
 - Identify follow-up research questions

Which takes longer?

Time allocation

In decreasing order...

1. Defining the problem
2. Obtaining the data
3. Cleaning and exploring the data
4. Managing expectations
5. Summarising the results
6. Learning new things
7. Modelling

The 'PR problem' of Data Science

Inevitably the data are...

- Not quite what you need to solve your problem
- Too limited, too large, too inaccurate, too expensive to obtain...

But (eventually) you...

- End up with a 'nice' dataset
- Apply some models

...and it **looks** incredibly easy from the outside!

Define the research question

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

Questions to ask

- Which metric(s) need to be improved?
- Which are possible actions to solve the problem?
- Which information is necessary and sufficient?
- What is the benefit of solving the problem?

Get the data

- **Ideal vs available** ('opportunistic' usage)
- Limitations

Questions to ask

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

Explore the data

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

Questions to ask

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

Model the data

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

Questions to ask

- Is there an outcome?
- What is an appropriate model for the data?
- How can we 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
- Interpret results sceptically

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?

Summarise the results

- **Storytelling** and **visual aids** to interpretation
- Assumptions and limitations

Questions to ask

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

The Data Science workflow is
non-linear and **iterative**