

Best Practices for Data Science Projects

Hector Corrada Bravo

Center for Bioinformatics and Computational Biology

Libraries

- 1.Connect/access databases
- 2.Data structures for fundamental objects
- 3.Basic operations/algorithms on these structures
- 4.Tools for communication

Reproducibility

- Extremely important aspect of data analysis
 - ‘Starting from the same raw data, can we reproduce your analysis and obtain the same results?’
- Using libraries helps:
 - Since you don’t reimplement everything, reduce programmer error
 - Large user bases serve as ‘watchdog’ for quality and correctness
- Standard practices help:
 - Version control: git
 - Unit testing: RUnit, testthat
 - Share and publish: github

Practical Tips

- Many tasks can be organized in modular manner:
 - Data acquisition
 - Algorithm/tool development
 - Computational analysis
 - Communication of results

Practical Tips

- Many tasks can be organized in modular manner:
 - Data acquisition: get data, put it in usable format (many 'join' operations), clean it up (wrangling)
 - Algorithm/tool development
 - Computational analysis
 - Communication of results

Practical Tips

- Many tasks can be organized in modular manner:
 - Data acquisition: get data, put it in usable format (many 'join' operations), clean it up
 - Algorithm/tool development: if new analysis tools are required
 - Computational analysis
 - Communication of results

Practical Tips

- Many tasks can be organized in modular manner:
 - Data acquisition: get data, put it in usable format (many 'join' operations), clean it up
 - Algorithm/tool development: if new analysis tools are required
 - Computational analysis: use tools to analyze data
 - Communication of results

Practical Tips

- Many tasks can be organized in modular manner:
 - Data acquisition: get data, put it in usable format (many 'join' operations), clean it up
 - Algorithm/tool development: if new analysis tools are required
 - Computational analysis: use tools to analyze data
 - Communication of results: prepare summaries of experimental results, plots, publication, upload processed data to repositories

Practical Tips

- Many tasks can be organized in modular manner:
 - Data acquisition: get data, put it in usable format (many 'join' operations), clean it up
 - Algorithm/tool development: if new analysis tools are required
 - Computational analysis: use tools to analyze data
 - Communication of results: prepare summaries of experimental results, plots, publication, upload processed data to repositories

Usually a single language or tool
does not handle all of these
equally well

Practical Tips

- Many tasks can be organized in modular manner:
 - Data acquisition: get data, put it in usable format (many 'join' operations), clean it up
 - Algorithm/tool development: if new analysis tools are required
 - Computational analysis: use tools to analyze data
 - Communication of results: prepare summaries of experimental results, plots, publication, upload processed data to repositories

Choose the best tool
for the job!

Practical Tips

- Modularity requires organization and careful thought
- In Data Science we wear two hats
 - Algorithm/tool developer
 - **Experimentalist**: we don't get trained to think this way enough!
- It helps two consciously separate these two jobs

Think like an experimentalist

- Plan your experiment
- Gather your raw data
- Gather your tools
- Execute experiment
- Analyze
- Communicate

Think like an experimentalist

- Let this guide your organization. I find structuring my projects like this to be useful:

```
project/  
| data/  
| | processing_scripts  
| | raw/  
| | proc/  
| tools/  
| | src/  
| | bin/  
| exps  
| | pipeline_scripts  
| | results/  
| | analysis_scripts  
| | figures/
```

Think like an experimentalist

- Keep a lab notebook!
- Literate programming tools are making this easier for computational projects
 - http://en.wikipedia.org/wiki/Literate_programming
 - <http://rmarkdown.rstudio.com/>
 - <http://jupyter.org/>

Think like an experimentalist

- Separate experiment from analysis from communication
 - Store results of computations, write separate scripts to analyze results and make plots/tables
- **Aim for reproducibility**
 - There are serious consequences for not being careful
 - Publication retraction
 - Worse: http://videlectures.net/cancerbioinformatics2010_baggerly_irrh/
 - Lots of tools available to help, use them! Be proactive: learn about them on your own!

Bias, ethics and responsibility

Data Science Lifecycle

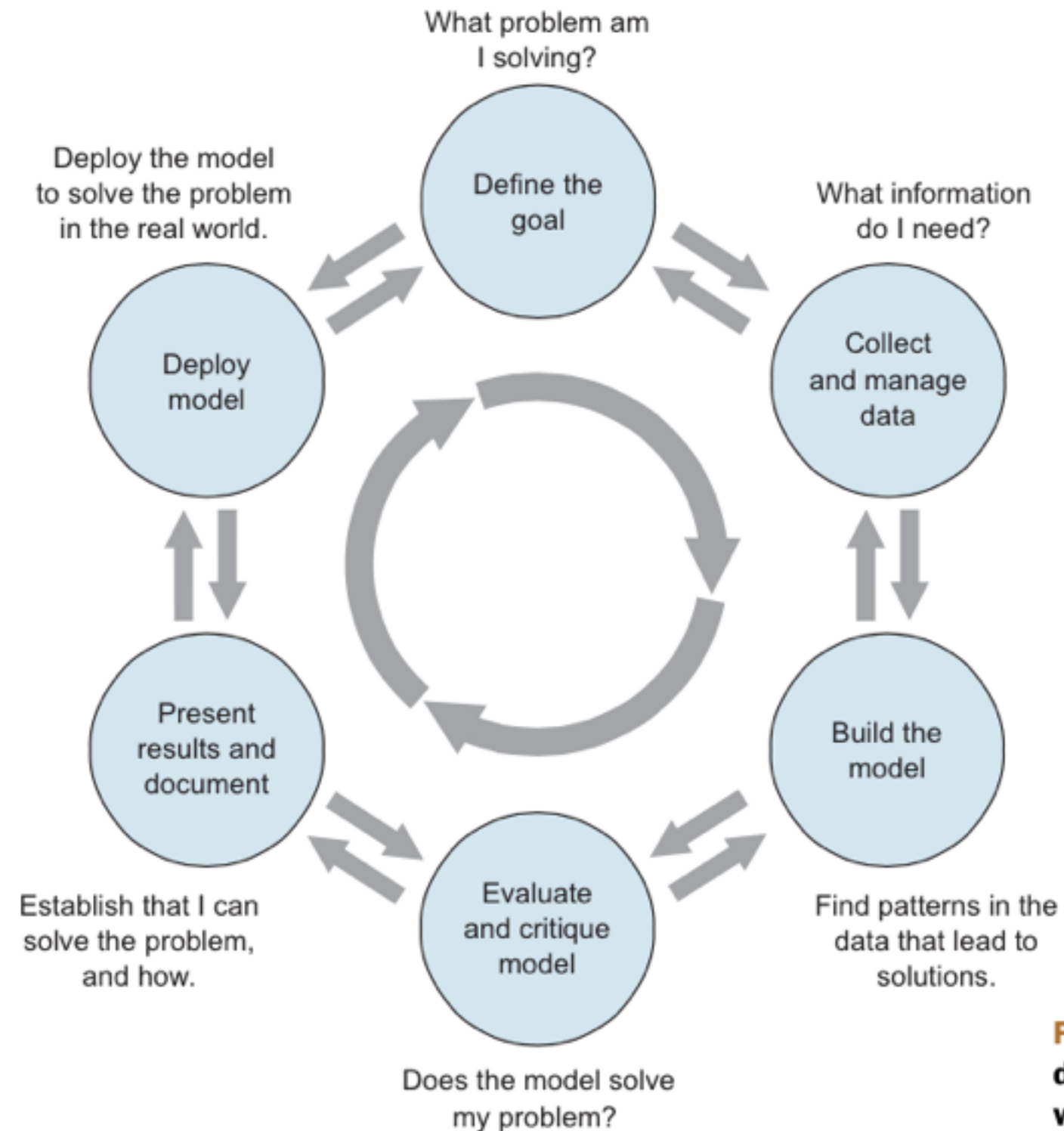


Figure 1.1 The lifecycle of a data science project: loops within loops

Examples

- Genetic testing
 - Genetic tests for heart disorder and race-biased risk (NYTimes)
 - Race-bias in ancestry reports
- Search results / feed optimization
 - Google
 - Facebook

Data collection

- What data should (not) be collected
- Who owns the data
- Whose data can (not) be shared
- What technology for collecting, storing, managing data
- Whose data can (not) be traded
- What data can (not) be merged
- What to do with prejudicial data

[Fung, 2016]

Data Modeling

- Data is biased (known/unknown)
 - Invalid assumptions
 - Confirmation bias
- Publication bias
- Badly handling missing values

[Fung, 2016]

Deployment

- Spurious correlation / over-generalization
- Using “black-box” methods that cannot be explained
- Using heuristics that are not well understood
- Releasing untested code
- Extrapolating
- Not measuring lifecycle performance (concept drift in ML)

[Fung, 2016]

Guiding principles

- Start with clear user need and public benefit
- Use data and tools which have minimum intrusion necessary
- Create robust data science models
- Be alert to public perceptions
- Be as open and accountable as possible
- Keep data secure

[UK cabinet office]

Some references

- Presentation on ethics and data analysis, Kaiser Fung @ Columbia Univ. http://andrewgelman.com/wp-content/uploads/2016/04/fung_ethics_v3.pdf
- O'Neil, Weapons of math destruction. <https://www.amazon.com/Weapons-Math-Destruction-Increases-Inequality/dp/0553418815>
- UK Cabinet Office, Data Science Ethical Framework. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/524298/Data_science_ethics_framework_v1.0_for_publication_1.pdf
- Derman, Modelers' Hippocratic Oath. <http://www.iijsournals.com/doi/pdfplus/10.3905/jod.2012.20.1.035>