Reproducible machine learning

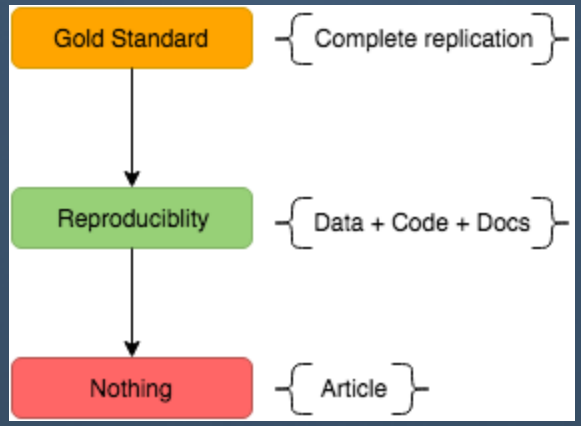
Aneesh Karve, @akarve
Adam Sah, asah@quiltdata.io

Who we are

Triangle of reproducibility

- code
- data
- models

Reproduction vs Replication



[AP]

Why

- trust
- compliance
- auditing
- teams
- banking, healthcare, transportation, insurance

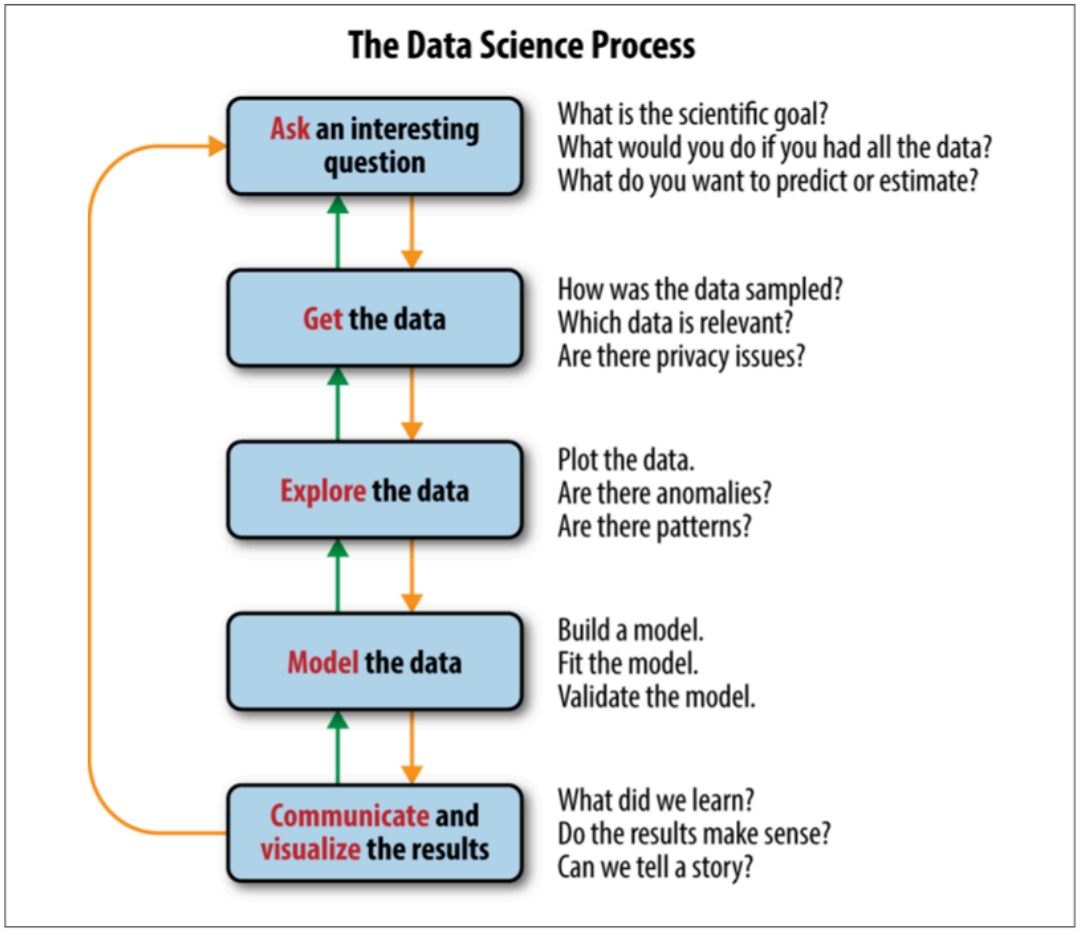


Figure 1-1. One representation of the data science process (courtesy of Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course)

[CB]

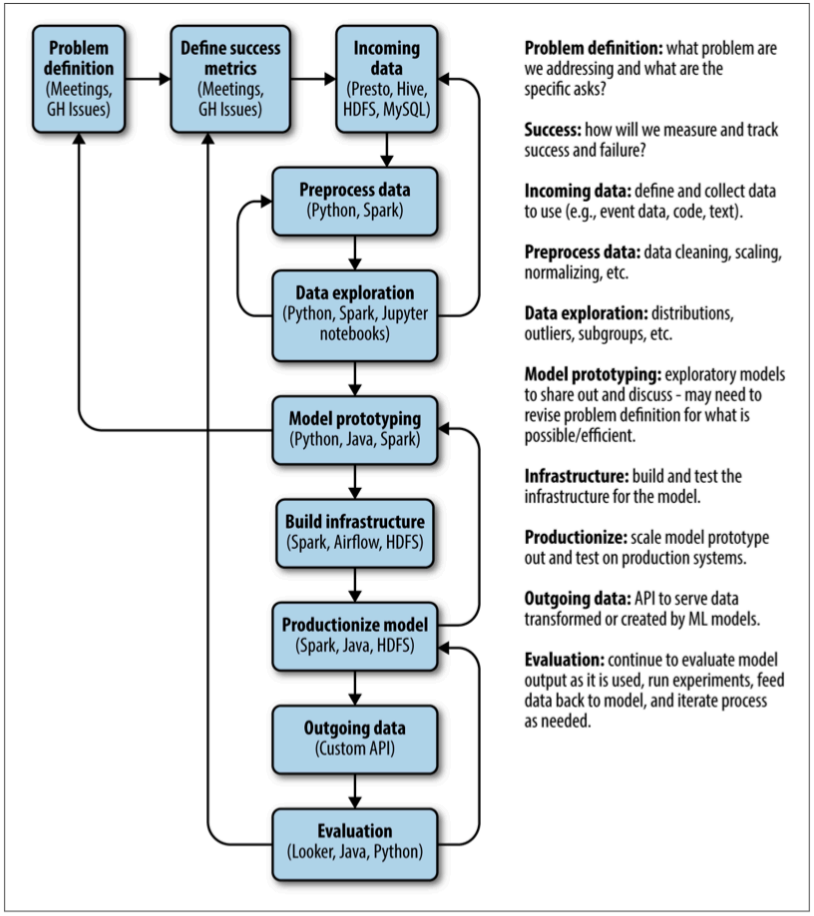


Figure 1-2. The data science workflow of GitHub's machine learning team

[CB]

Out of scope

- Deploying models to production
- Big data (Spark, Hadoop)
- Docker
- But many of the tools discussed still apply

Tools for reproducibility

- code = git + conda + jupyter
- data = pandas + quilt
- models = quilt + { sklearn | TensorFlow | ...}

Package managers

Now you've got two problems

pip or conda?

- pip - python packages in any env
- conda - any package in conda envs
- plz stahp: sudo pip install

[JV]

conda

- packages + environments
- run from the command line
- no packages in root
- all packages in same env
- all packages in same environment (use "which")
- `conda create -n MY_ENV python=3.6`
- `pip freeze`

Turnkey solutions FTW

- Paperspace
- FloydHub
- Domino

Data versioning

Don't put data in git

- Slows down repo
- Does not serialize
- Does not handle large data (incl. LFS)
- Conflates code and data changes
- pip install quilt

Demo

Quilt + TensorFlow

- Virtual File System
- Minimal code changes to use Quilt
- Alpha stage

Quilt + Keras/TF

- Reproducibility = specific data version, verified hashes
- Load input data from Quilt
- Load model checkpoints from Quilt (vs local files)
- Save and share model checkpoints

```
# map local directories/files to Quilt packages
import quilt.vfs
quilt.vfs.setup("<PKGNAME>:version", mappings={...})

# existing code - unchanged
train_x, train_y, trail_l = get_data_set()
test_x, test_y, test_l = get_data_set("test")
```

Demo

Source code, slides

https://github.com/quiltdata/reproducible-ml

Further learning

- A tour of artificial intelligence and its limitations
- [CB] Development Workflows for Data Scientists, Claire Byrne
- [AP] Reproducible Research in Computational Sciences, Arnu Pretorius
- [JV] Installing Python Packages from a Jupyter Notebook, Jake VanderPlas
- [TM] Machine Learning, Tom Mitchell