Starting Data Science with Kaggle

Learning, Community, Career, Fun

Gerrit Gruben

September 9, 2016

Kaggle Berlin

Table of contents

Our Meetup group

Meetup group

First and foremost this group is about kaggling.

Secondarily, topics relating to kaggle and its contests are of interest, this includes (among others) *machine learning*, *applied mathematics*, *data analytics tooling*, and *career in data science*.

History

- Originally by Ezzeri Esa and more of *a tutorial group*
- Sister group: Advanced Machine Learning by Marcel Ackermann, see `https://www.meetup.com/de-DE/Advanced-Machine-Learning-Study-Group`
- Since last year Gerrit organizes the group. More *hackathon oriented*.

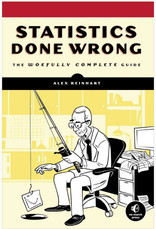
Insights

Lessons from one year of community building:

- Do not make a community dependent on a single interest group
- Keep audience updated, *bias for communication*
- Others are more helpful than expected
- Be *receptive* to community contributions
- RSVP discipline is low, probably hardest problem to deal with

Ethics

- Politeness is inexpensive and should be used in abundance
- Listen and understand other's opinions, discuss about *evidence*
- Proactively work for a *proper* use of statistics.



Looking for you

Organizer

- Open and friendly attitude
- Either long-term kaggler or academic
- Willing to thoroughly check handed-in talks

Presenter

Give a talk about *own kaggle experience* or a data science topic in general.

Navigating through Data Science

Data Scientist

Data scientist *somewhat vague*, mostly one of:

- A classical *data or BI analyst*
- *CEO whisperer* with super powers in computing sciences, mathematics, and business knowledge.
- Concession to a top performer among *software engineers* (or getting some of them at all)

? Knows machine learning, big data, or some other black magic

Data Scientist

Data scientist *somewhat vague*, mostly one of:

- A classical *data or BI analyst*
- *CEO whisperer* with super powers in computing sciences, mathematics, and business knowledge.
- Concession to a top performer among *software engineers* (or getting some of them at all)

? Knows machine learning, big data, or some other black magic rather: *data engineer*, will converge to canonical CS knowledge

Bottom up (or Forward Selection?)

$$\text{machine learning, statistics, programming} \subseteq hardskills(DS)$$

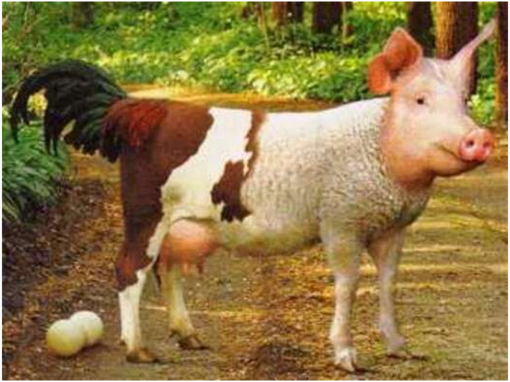
$$\text{presentation, communication} \subseteq softskills(DS)$$

Eierlegende Wollmilchsau

For non-native Germans: What is a *eierlegende Wollmilchsau*?

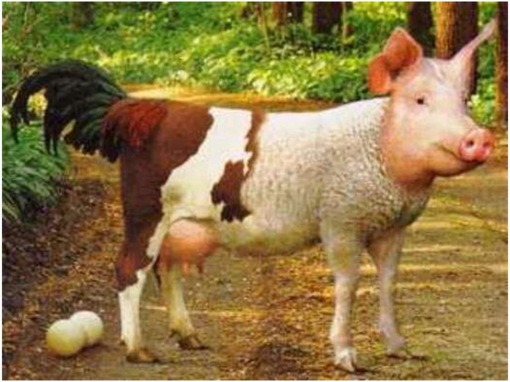
Eierlegende Wollmilchsau

For non-native Germans: What is a *eierlegende Wollmilchsau*?



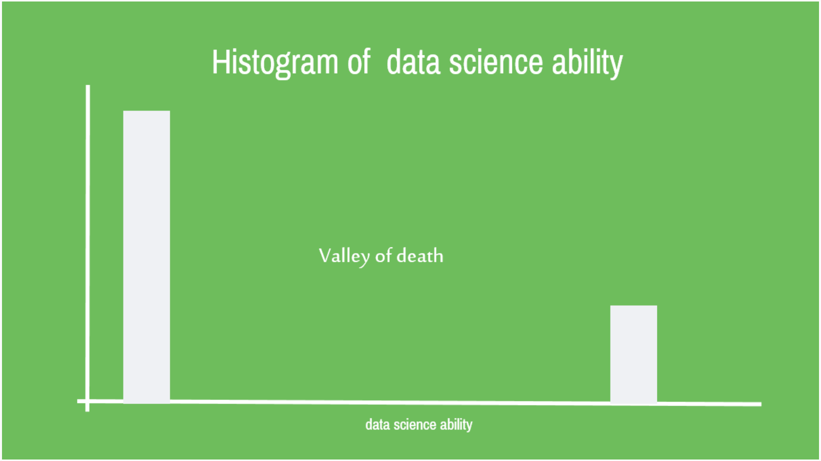
Eierlegende Wollmilchsau

For non-native Germans: What is a *eierlegende Wollmilchsau*?



Fred, Data Scientist

Dichotomy



Taken from "A full ML pipeline in Scikit-learn vs Scala/Spark" by Jose Quesda and David Anderson

Why Kaggle?

$$S = L + MV \times RV$$

*S*uccess, *L*uck, *M*arket *V*alue, *R*eal *V*alue

Why Kaggle?

$$S = L + MV \times RV$$

*S*uccess, *L*uck, *M*arket *V*alue, *R*eal *V*alue

Btw. this is *dating advice* from Quora

Why Kaggle? - Altruism

Galaxy Zoo - The Galaxy Challenge	Classify the morphologies of distant galaxies in our Universe 2 years ago · Research	326 teams $16,000
Flavours of Physics: Finding $\tau \rightarrow \mu\mu\mu$	Identify a rare decay phenomenon A year ago · Featured	673 teams 806 kernels $15,000
March Machine Learning Mania 2015	Predict the 2015 NCAA Basketball Tournament A year ago · Featured	341 teams $15,000
Click-Through Rate Prediction	Predict whether a mobile ad will be clicked 2 years ago · Featured	1,604 teams $15,000
March Machine Learning Mania	Tip off college basketball by predicting the 2014 NCAA Tournament 2 years ago · Featured	248 teams $15,000
Higgs Boson Machine Learning Challenge	Use the ATLAS experiment to identify the Higgs boson 2 years ago · Featured	1,785 teams $13,000
Practice Fusion Analyze This! 2012 - Prediction Challenge	Start digging into electronic health records and submit your ideas for the most promising, impactful or interesting predictive modeling competitions 4 years ago · Prospect	0 teams $10,500
Right Whale Recognition	Identify endangered right whales in aerial photographs 8 months ago · Research	364 teams $10,000
Truly Native?	Predict which web pages served by StumbleUpon are sponsored A year ago · Featured	274 teams $10,000
Grasp-and-Lift EEG Detection	Identify hand motions from EEG recordings A year ago · Research	379 teams 627 kernels $10,000

Why Kaggle? - Altruism

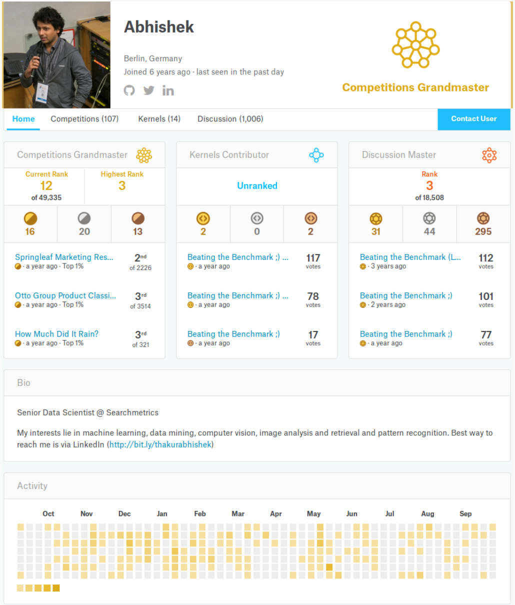
Competition	Teams / Kernels	Prize
Galaxy Zoo - The Galaxy Challenge Classify the morphologies of distant galaxies in our Universe 2 years ago · Research	326 teams	$16,000
Flavours of Physics: Finding τ → μμμ Identify a rare decay phenomenon A year ago · Featured	673 teams 806 kernels	$15,000
March Machine Learning Mania 2015 Predict the 2015 NCAA Basketball Tournament A year ago · Featured	341 teams	$15,000
Click-Through Rate Prediction Predict whether a mobile ad will be clicked 2 years ago · Featured	1,604 teams	$15,000
March Machine Learning Mania Tip off college basketball by predicting the 2014 NCAA Tournament 2 years ago · Featured	248 teams	$15,000
Higgs Boson Machine Learning Challenge Use the ATLAS experiment to identify the Higgs boson 2 years ago · Featured	1,785 teams	$13,000
Practice Fusion Analyze This! 2012 - Prediction Challenge Start digging into electronic health records and submit your ideas for the most promising, impactful or interesting predictive modeling competitions 4 years ago · Prospect	0 teams	$10,500
Right Whale Recognition Identify endangered right whales in aerial photographs 8 months ago · Research	364 teams	$10,000
Truly Native? Predict which web pages served by StumbleUpon are sponsored A year ago · Featured	274 teams	$10,000
Grasp-and-Lift EEG Detection Identify hand motions from EEG recordings A year ago · Research	379 teams 627 kernels	$10,000

Honorable mention: DSSG `http://dssg-berlin.org/`

Why Kaggle? - Domain variety

"The best thing about being a statistician is that you get to play in everyone's backyard." — *John Turkey*

Why Kaggle? - Visibility



Why Kaggle? - Learn from the best

Exploring Survival on the Titanic

by Megan Risdal · last run 5 months ago · R notebook · 40759 vie...
using data from Titanic: Machine Learning from Disaster

231
voters

Report | Code | Output (2) | Comments (83) | Log | Versions (5) | Forks (232) | Fork Script

Report

Exploring the Titanic Dataset

Megan L. Risdal

6 March 2016

- 1 Introduction
  - 1.1 Load and check data
- 2 Feature Engineering
  - 2.1 What's in a name?
  - 2.2 Do families sink or swim together?
  - 2.3 Treat a few more variables …
- 3 Missingness
  - 3.1 Sensible value imputation
  - 3.2 Predictive imputation
  - 3.3 Feature Engineering: Round 2
- 4 Prediction
  - 4.1 Split into training & test sets
  - 4.2 Building the model
  - 4.3 Variable importance
  - 4.4 Prediction!
- 5 Conclusion

Summary

Kaggling will benefit *you* in these terms:

- Teaches applied machine learning techniques not found in textbook
- Create a Data Science project portfolio
- Get to learn several domains
- Help *mankind*
- Learn best practices from experts working on the same problem

Summary

Kaggling in this group will benefit *you* in these terms:

- Teaches applied machine learning techniques not found in textbook
- Create a Data Science project portfolio
- Get to learn several domains
- Help *mankind*
- Learn best practices from experts working on the same problem
- Improve your presentation skills
- Make friends and team mates

vs. competitive programming

Kaggling is sometimes put in the same basket as competitive programming, though:

- Diminishing returns much earlier in competitive programming
- Kaggle projects are more open
- Crowd structurally different
- *Knowledge gained by kaggling is more applicable to real life*

A project template

Goal

Provide a technical environment to do Data Science in:

- Isolation: Project environment should not interact with other parts of the system if not necessary
- Reproducibility: Results should be reproducible by others or on other devices
- Structure: Provide a easy to understand structure to reduce *context switch costs*
- Low barrier: Avoid throwing documentation at people
- No boundaries: Make the template itself extensible and use open, freely available tech (*Open Source*)

Goal

Provide a technical environment to do Data Science in:

- Isolation: Project environment should not interact with other parts of the system if not necessary
- Reproducibility: Results should be reproducible by others or on other devices
- Structure: Provide a easy to understand structure to reduce *context switch costs*
- Low barrier: Avoid throwing documentation at people
- No boundaries: Make the template itself extensible and use open, freely available tech (*Open Source*)

complexity $\rightarrow$ min!

Goal

Provide a technical environment to do Data Science in:

- Isolation: Project environment should not interact with other parts of the system if not necessary
- Reproducibility: Results should be reproducible by others or on other devices
- Structure: Provide a easy to understand structure to reduce *context switch costs*
- Low barrier: Avoid throwing documentation at people
- No boundaries: Make the template itself extensible and use open, freely available tech (*Open Source*)

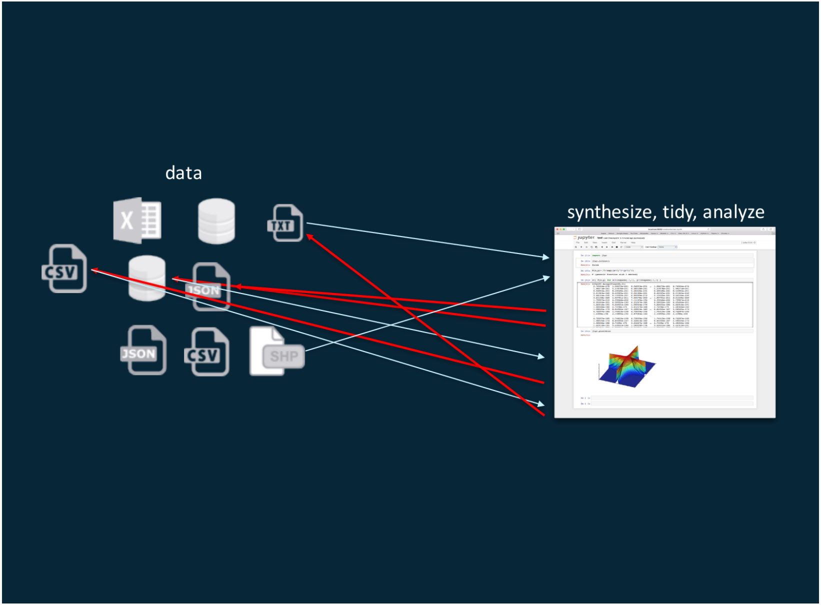
complexity $\rightarrow$ min!

We use Python...

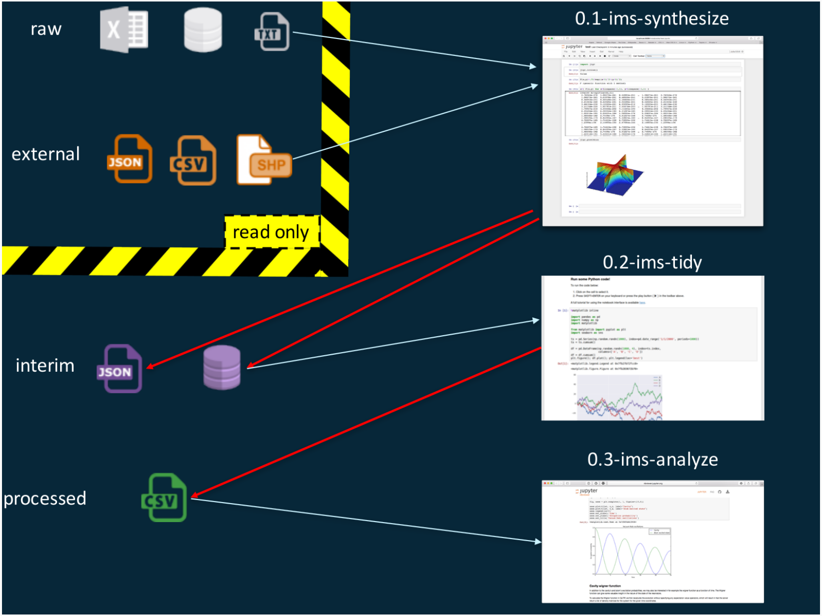
Existing work

- Kaggle scripts uses Docker images for reproducibility
  `http://blog.kaggle.com/2016/02/05/how-to-get-started-with-data-science-in-containers`
- We tried to use a Vagrant based solution in teaching
  `http://www.cs.uni-potsdam.de/~ggruben/vm.html`
- Recent SciPy 2016 talk contains a well-structured project structure and some neat Jupyter tricks
  `http://isaacslavitt.com/2016/07/20/data-science-is-software-talk`
  (next slides are borrowed from it)

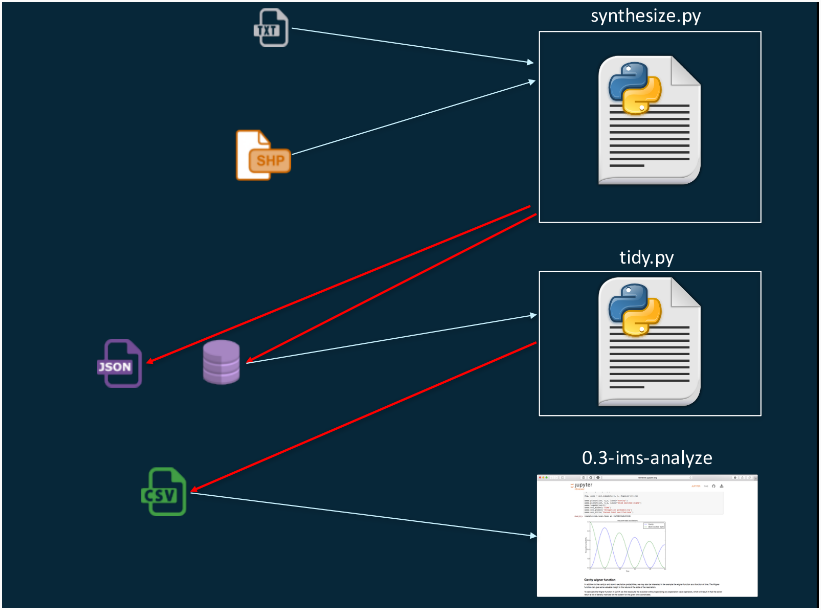
What people do



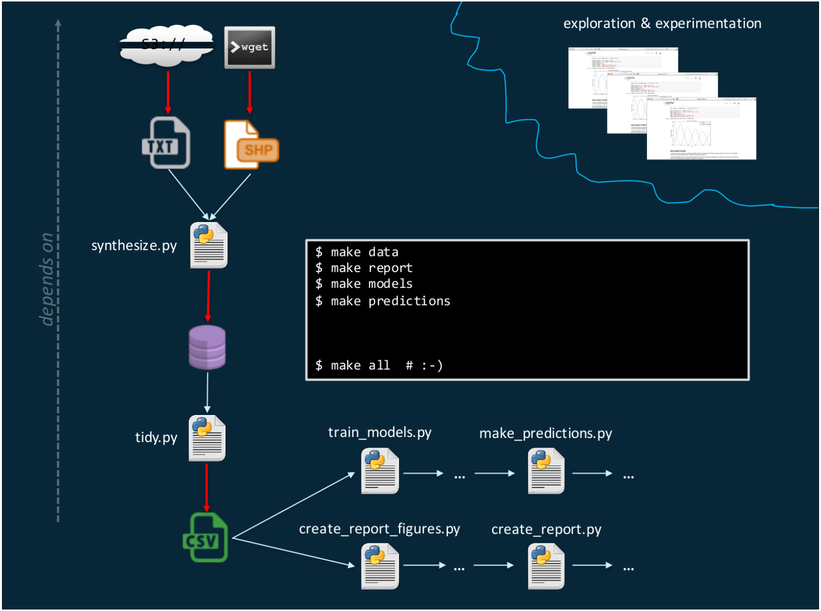
By artifacts



Parts are automatizable



Final


depends on
S3://
wget
TXT
SHP
synthesize.py
tidy.py
CSV
exploration & experimentation
$ make data
$ make report
$ make models
$ make predictions
$ make all # :-)
train_models.py
make_predictions.py
create_report_figures.py
create_report.py
...
...
...
...

Getting started

Setting up a new project from the template

```
$ pip install cookiecutter
$ cookiecutter https://github.com/uberwach/ \
      cookiecutter-kaggle
```

Overview

```
├── data
│   ├── external
│   ├── interim
│   ├── submission
│   └── raw
├── Dockerfile
├── requirements.txt
├── Makefile
├── submissions
├── src
├── notebooks
├── models
├── reports
├── references
└── README.md
```

Data

```
├── data
│   ├── external   <- if you have data from outside
│   ├── interim    <- cleaned, filtered data sets
│   ├── processed  <- final datasets to build predictive
│   └── raw        <- kaggle data files
│
```

Can synchronize with S3 (want to add Dropbox later)

Environment

```
├── Dockerfile
├── requirements.txt
```

Define environment, which packages and libraries are used? *Brings every system on the same page*

Makefile

```
├── Makefile
```

Defines recipes on how artifacts (data files, reports, visualizations).
Can also be used for synchronization, code quality, testing.

Examples: ‘make data’, ‘make data/interim/nn_autoencoder_feats.csv’

Source

```
├── src
├── notebooks
```

SRC is made a Python module (accessible from notebooks). Do versioning with Git.

Models

```
├── models
```

Often benefical to explicitly store models for inspection and later reuse, especially if they take long to train.

Documentation

```
├── reports    <- place for your presentations
├── references <- to store the learning material
and descriptions of data
└── README.md  <- project documentation
(appears on github)
```

Submissions

```
├── submissions
```

Contains the final submissions in the format needed for the contest.

.env

Optionally you can add

```
├── .env
```

That reads *environment variables* that should not be synchronized in public or dependent on your system configuration (AWS authentification keys, Theano flags i.e. GPU)

Demo

Environments

Environment with Anaconda (alternative: virtualenv)

```
$ conda create -n env_name python=3
$ source activate env_name
(env_name) $ ... start to use python like normally ...
# in project path
(env_name) $ pip install -r requirements.txt
# save current dependencies
(env_name) $ pip freeze | requirements.txt
$ conda env list
$ source deactivate
```

Docker 101

Mostly useful if you are not on Linux.

```
$ docker build -t yourproject/tagname .
# wait a while...
# this is based on Kaggle's image (big)
# compatible with Kaggle scripts

# start interactive shell
docker run -i -v $PWD:/tmp/working \
  -w=/tmp/working -t yourproject/tagname \
  /bin/bash
# on windows $PWD -> %cd%
```

Makefile: data

```
data_objs = train_simple_feats.csv test_simple_feats.csv

requirements:
 pip install -q -r requirements.txt

data: requirements $(data_objs)
 echo $(data_objs)

train_simple_feats.csv: requirements data/raw/train.csv
 python src/data/make_dataset.py data/raw/train.csv \
  data/interim/train_simple_feats.csv

test_simple_feats.csv: requirements data/raw/test.csv
 python src/data/make_dataset.py data/raw/test.csv \
  data/interim/test_simple_feats.csv
```