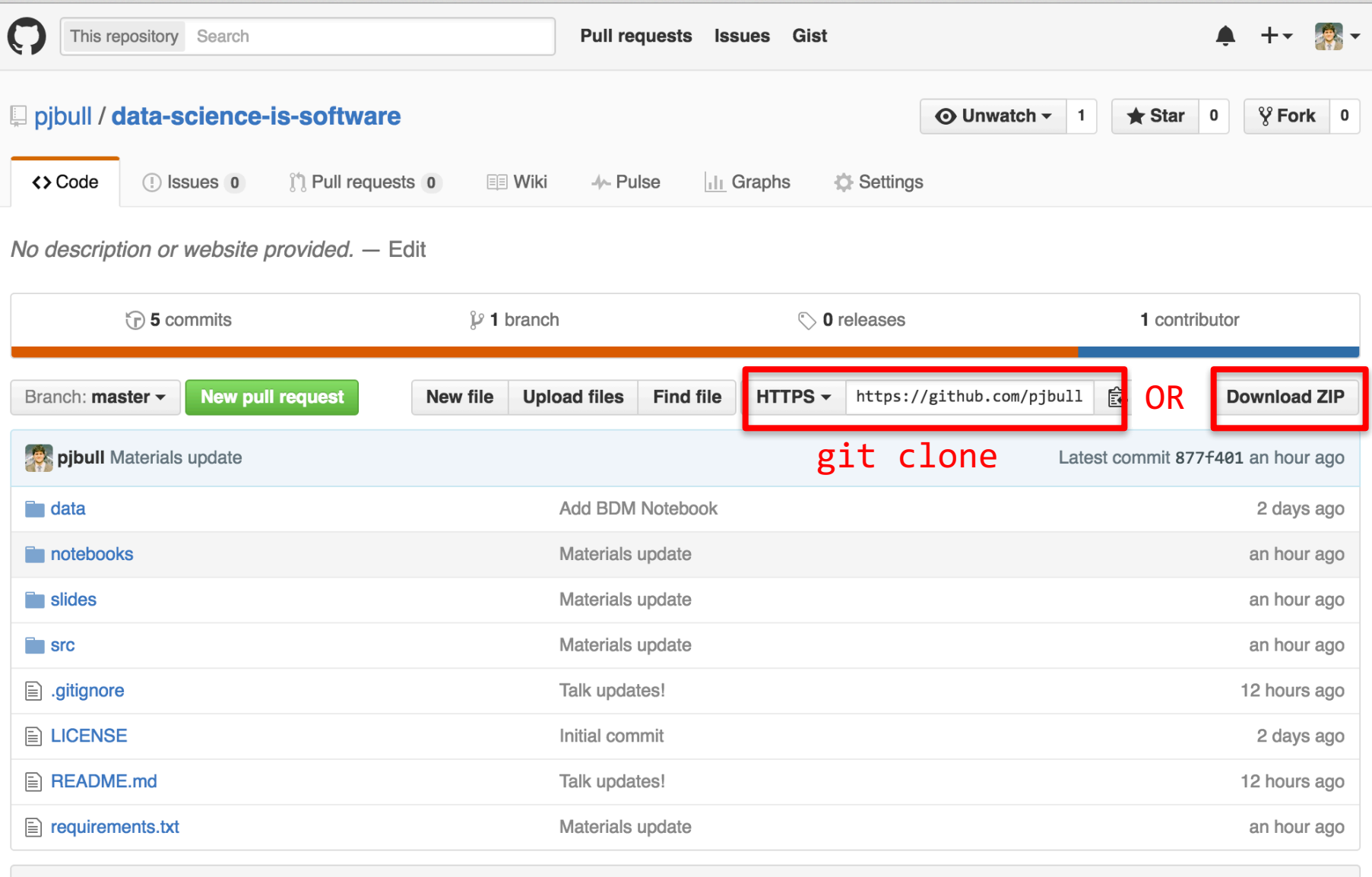
Data Science is Software:

Developer #lifehacks for the Jupyter Data Scientist

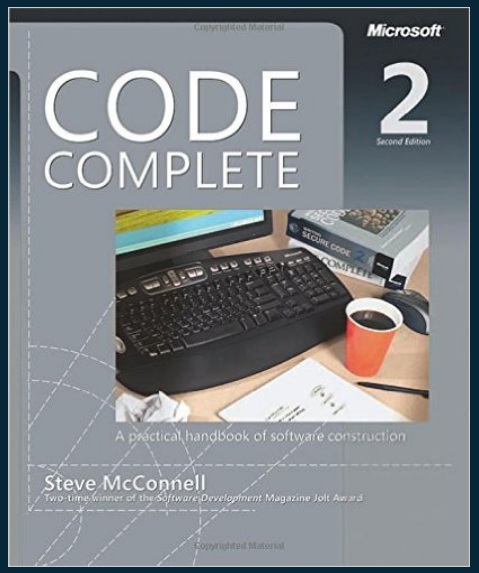
PETER BULL
@drivendataorg / @pjbull

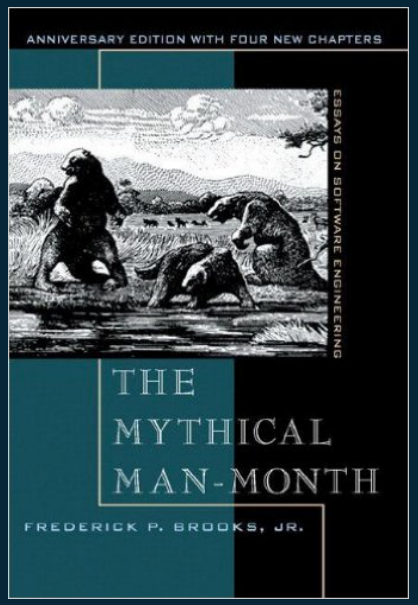
bit.ly/jupyter-lifehacks

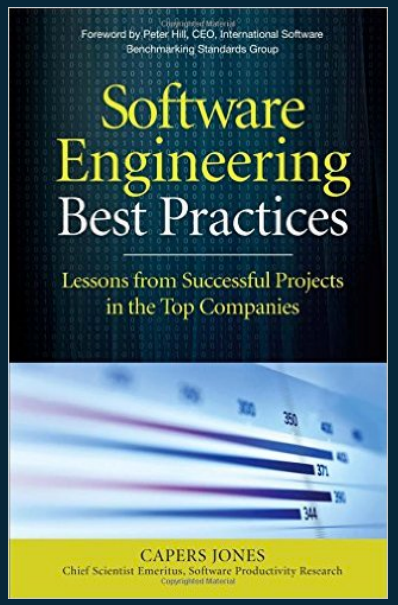


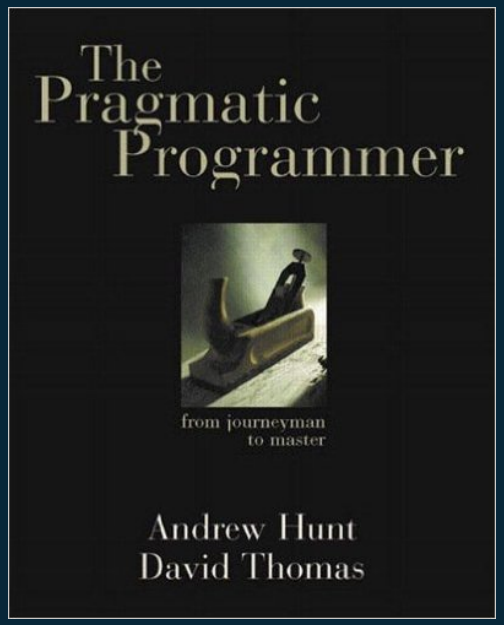
“It's easy to enhance a FORTRAN compiler to compile COBOL as well; it's just a SMOP.”

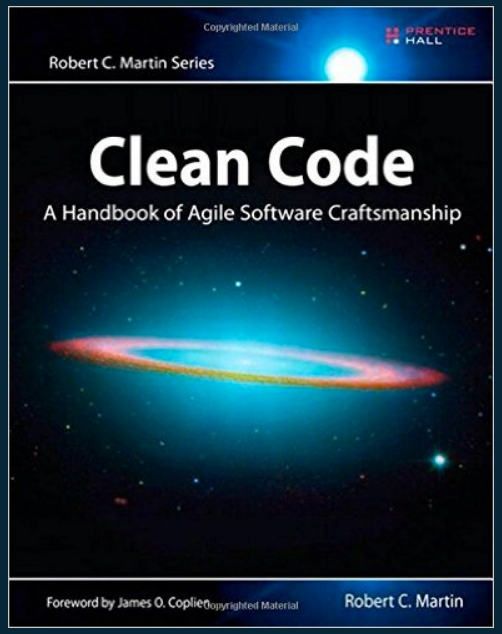
— SMOP entry in The Jargon File (a comprehensive compendium of hacker slang)

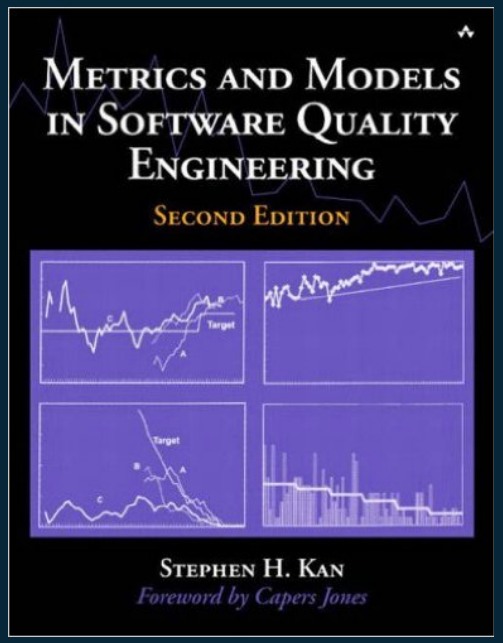

Microsoft
CODE COMPLETE
2
Second Edition
A practical handbook of software construction
Steve McConnell
Two-time winner of the Software Development Magazine Jolt Award


ANNIVERSARY EDITION WITH FOUR NEW CHAPTERS
ESSAYS ON SOFTWARE ENGINEERING
THE MYTHICAL MAN-MONTH
FREDERICK P. BROOKS, JR.


Foreword by Peter Hill, CEO, International Software Benchmarking Standards Group
Software Engineering Best Practices
Lessons from Successful Projects in the Top Companies
CAPERS JONES
Chief Scientist Emeritus, Software Productivity Research


The Pragmatic Programmer
from journeyman to master
Andrew Hunt
David Thomas


Robert C. Martin Series
PRENTICE HALL
Clean Code
A Handbook of Agile Software Craftsmanship
Foreword by James O. Coplien
Robert C. Martin

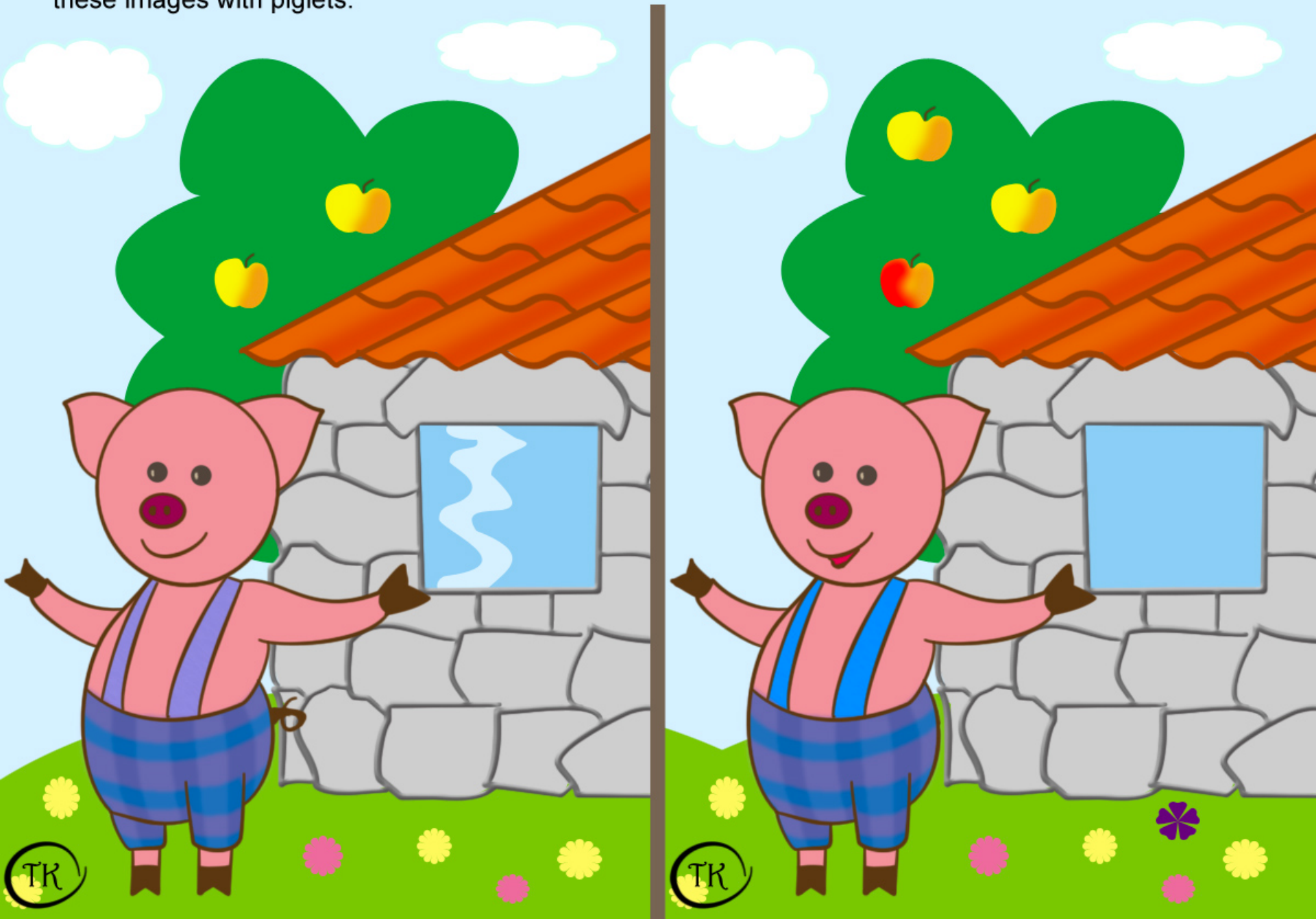

METRICS AND MODELS IN SOFTWARE QUALITY ENGINEERING
SECOND EDITION
STEPHEN H. KAN
Foreword by Capers Jones

1

This is my house

Spot 7 differences between these images with piglets.

http://www.everydayok.com




MRE
MEALS, RE
WITH FLAMELES
PRODUCT OF U.S.A
12
MEALS
THIS CASE CONTAINS TWE
EACH MEAL IS SEALED IN AN AIRTIGHT
• ENTRÉE • SIDE DISH • BREAD • SPREAD
• CONDIMENTS • INSTANT COFFEE • SPO
• STORE AT OR BELOW 72°F FOR 5 YEAR SHELF LIFE
• SEE BOTTOM FOR DATE OF MANUFACTURE
MRE
THREE-COURSE
MEAL, READY TO EAT
WITH FLAMELESS RATION HEATER
MEAL KIT
SUPPLY


jupyter

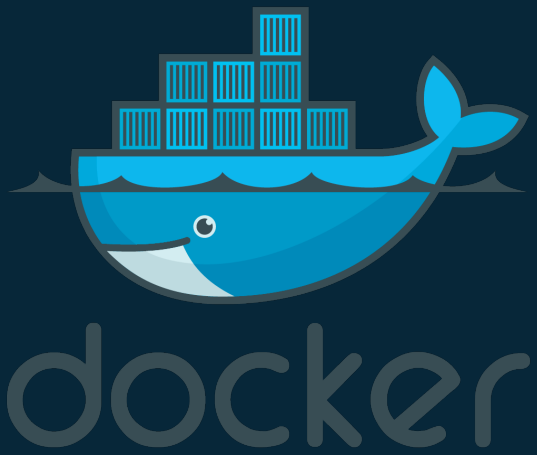
WATERMARK

VIRTUALENV

VIRTUALENVWRAPPER

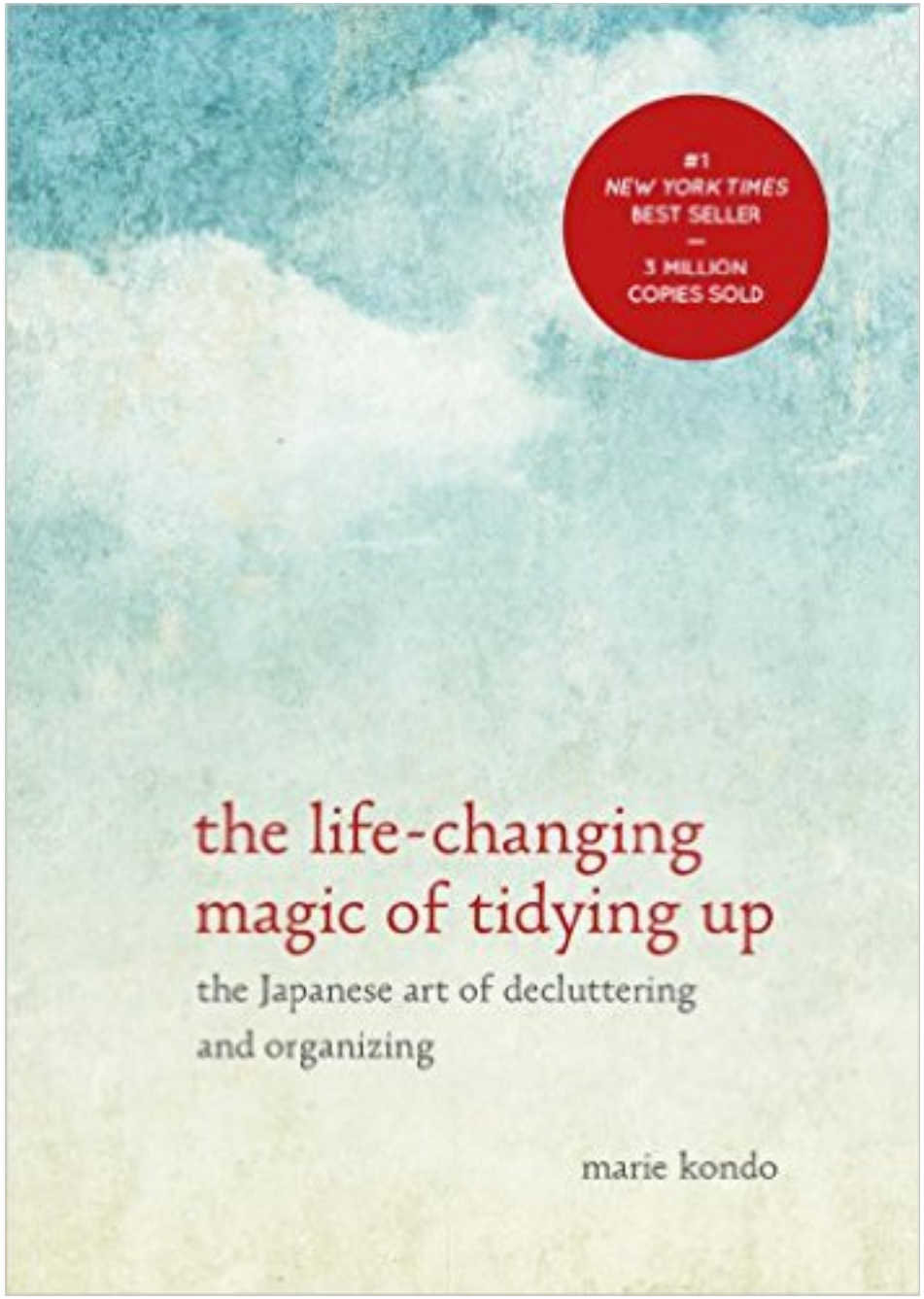
PIP REQUIREMENTS.TXT

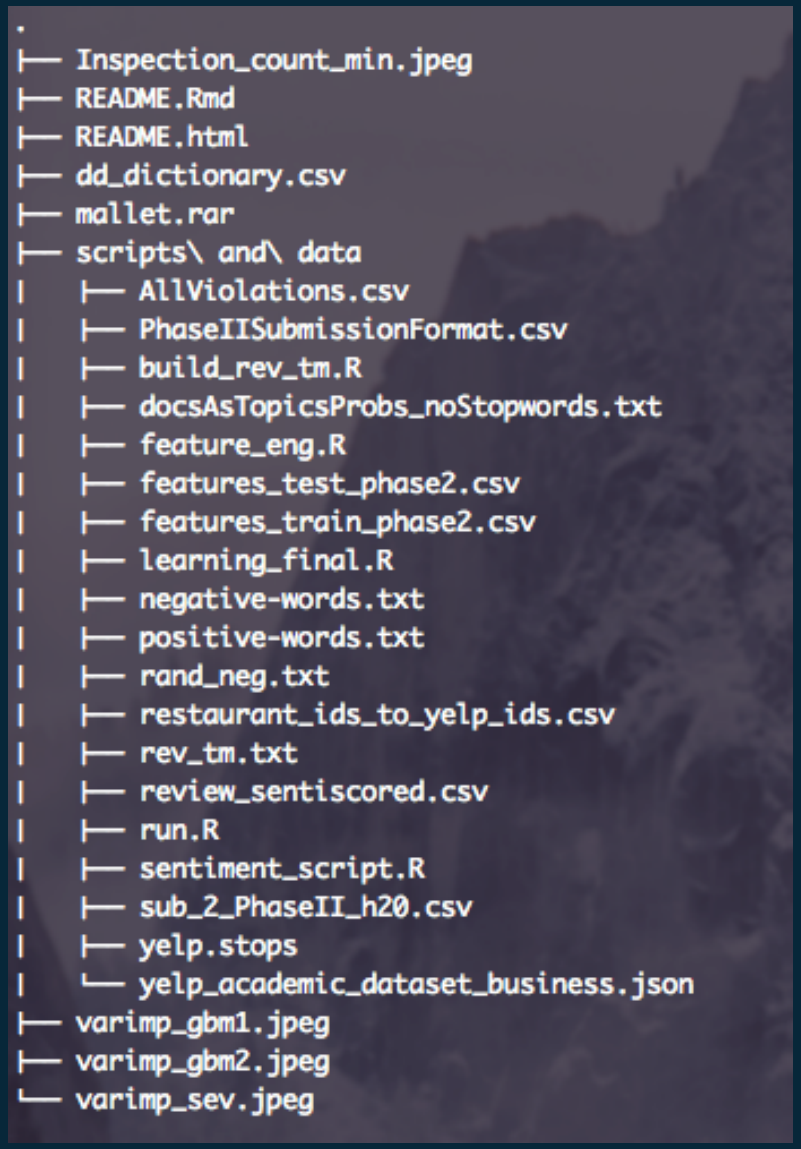
Other options (for more complex environments)

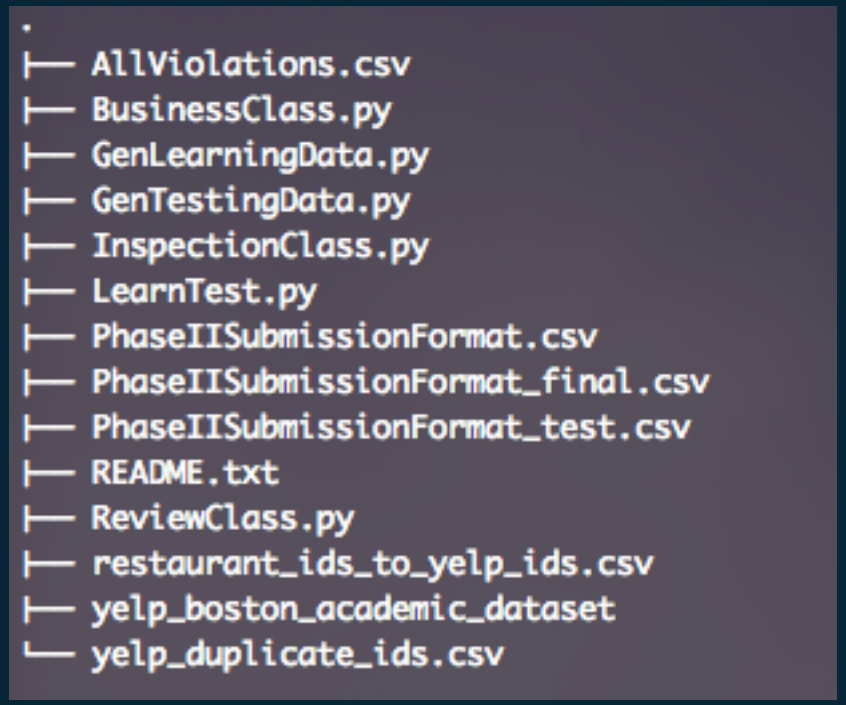


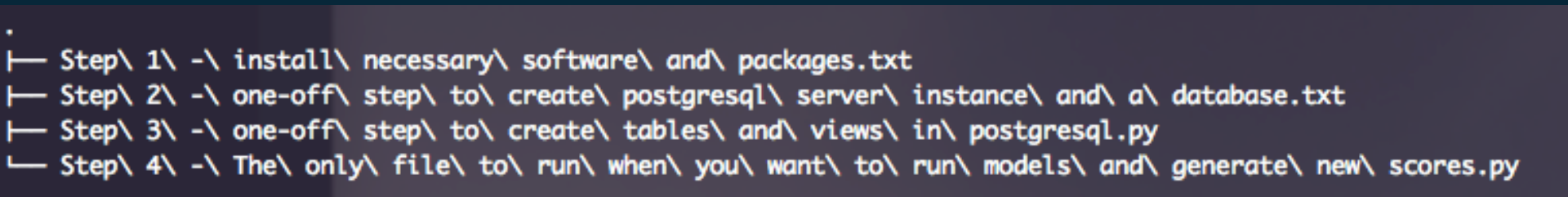
2

The Life-Changing Magic of Tidying Up

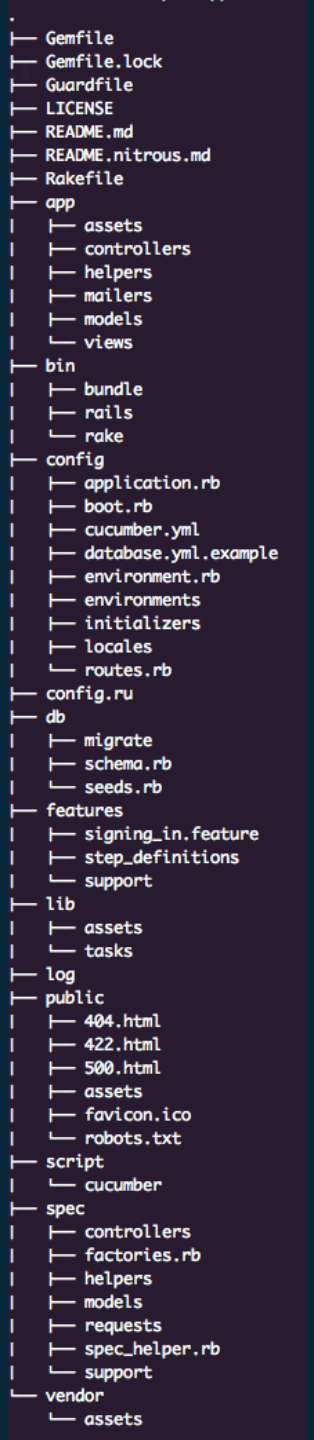

#1
NEW YORK TIMES
BEST SELLER
—
3 MILLION
COPIES SOLD
the life-changing
magic of tidying up
the Japanese art of decluttering
and organizing
marie kondo


.
├── Inspection_count_min.jpeg
├── README.Rmd
├── README.html
├── dd_dictionary.csv
├── mallet.rar
├── scripts\ and\ data
| ├── AllViolations.csv
| ├── PhaseIISubmissionFormat.csv
| ├── build_rev_tm.R
| ├── docsAsTopicsProbs_noStopwords.txt
| ├── feature_eng.R
| ├── features_test_phase2.csv
| ├── features_train_phase2.csv
| ├── learning_final.R
| ├── negative-words.txt
| ├── positive-words.txt
| ├── rand_neg.txt
| ├── restaurant_ids_to_yelp_ids.csv
| ├── rev_tm.txt
| ├── review_sentiscored.csv
| ├── run.R
| ├── sentiment_script.R
| ├── sub_2_PhaseII_h20.csv
| ├── yelp.stops
| └── yelp_academic_dataset_business.json
├── varimp_gbm1.jpeg
├── varimp_gbm2.jpeg
└── varimp_sev.jpeg

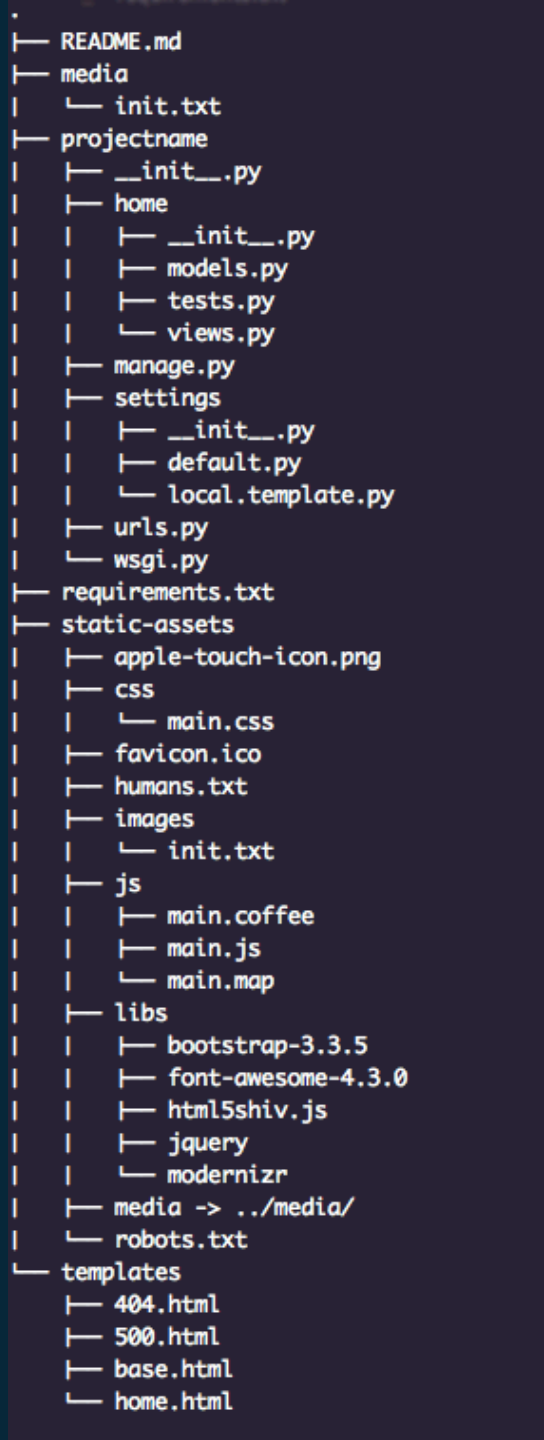

.
├── AllViolations.csv
├── BusinessClass.py
├── GenLearningData.py
├── GenTestingData.py
├── InspectionClass.py
├── LearnTest.py
├── PhaseIISubmissionFormat.csv
├── PhaseIISubmissionFormat_final.csv
├── PhaseIISubmissionFormat_test.csv
├── README.txt
├── ReviewClass.py
├── restaurant_ids_to_yelp_ids.csv
├── yelp_boston_academic_dataset
└── yelp_duplicate_ids.csv

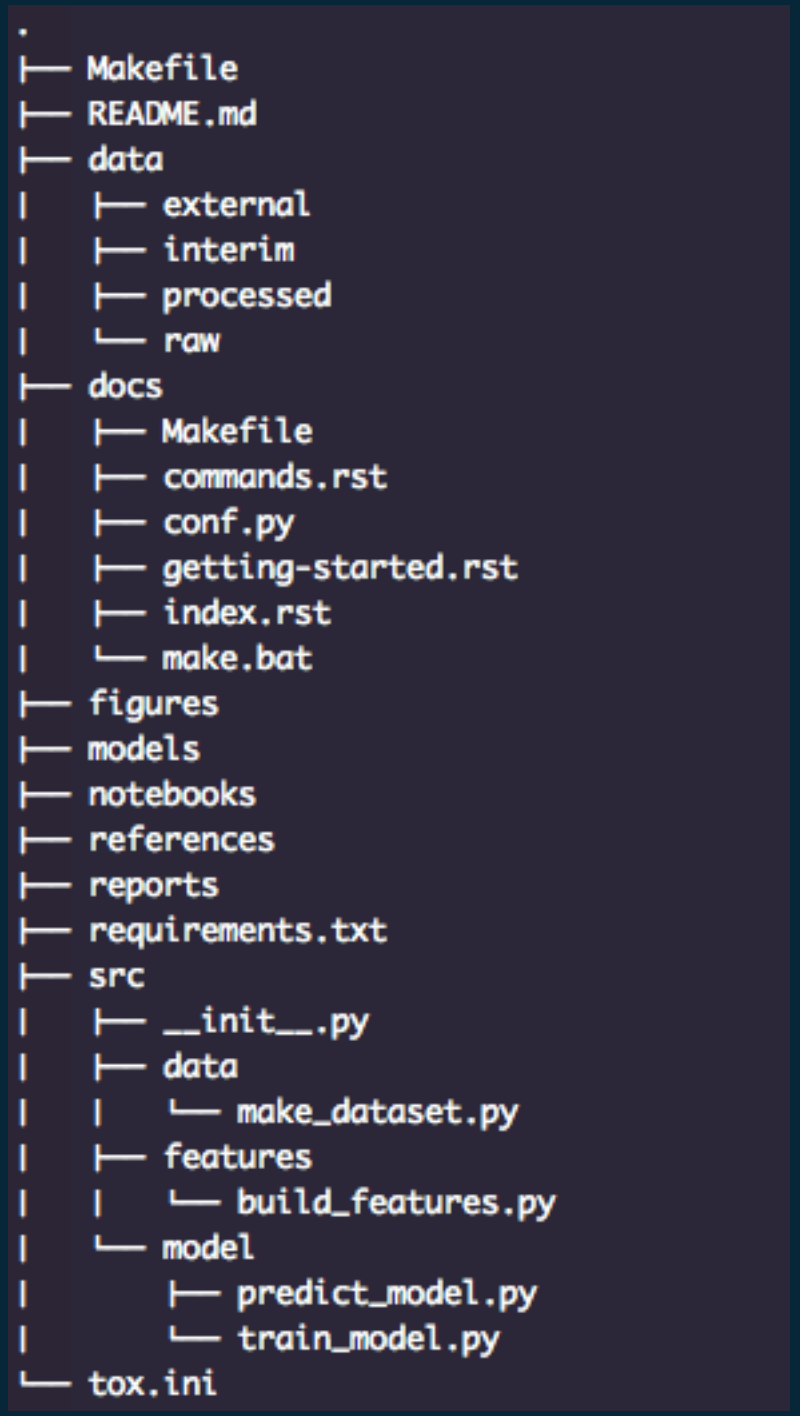

.
├── Step\ 1\ -\ install\ necessary\ software\ and\ packages.txt
├── Step\ 2\ -\ one-off\ step\ to\ create\ postgresql\ server\ instance\ and\ a\ database.txt
├── Step\ 3\ -\ one-off\ step\ to\ create\ tables\ and\ views\ in\ postgresql.py
└── Step\ 4\ -\ The\ only\ file\ to\ run\ when\ you\ want\ to\ run\ models\ and\ generate\ new\ scores.py

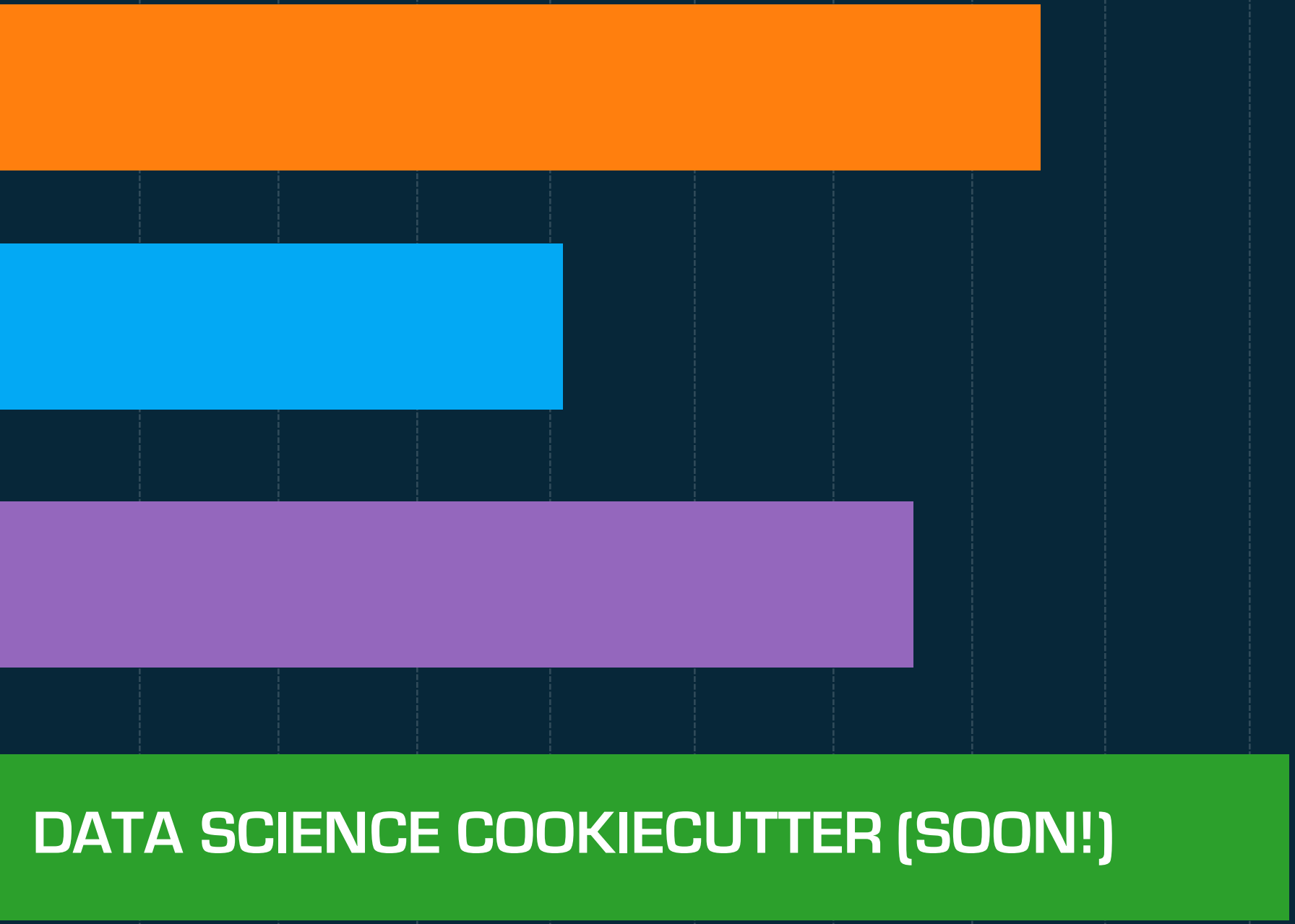
Ruby on Rails Application



Django Application




.
├── Makefile
├── README.md
├── data
│   ├── external
│   ├── interim
│   ├── processed
│   └── raw
├── docs
│   ├── Makefile
│   ├── commands.rst
│   ├── conf.py
│   ├── getting-started.rst
│   ├── index.rst
│   └── make.bat
├── figures
├── models
├── notebooks
├── references
├── reports
├── requirements.txt
├── src
│   ├── __init__.py
│   ├── data
│   │   └── make_dataset.py
│   ├── features
│   │   └── build_features.py
│   └── model
│   ├── predict_model.py
│   └── train_model.py
└── tox.ini


DATA SCIENCE COOKIECUTTER (SOON!)

3

Edit-run-repeat: Stopping the cycle of pain


jupyter

AUTORELOAD, PDB, DEBUG

Q, MODULES

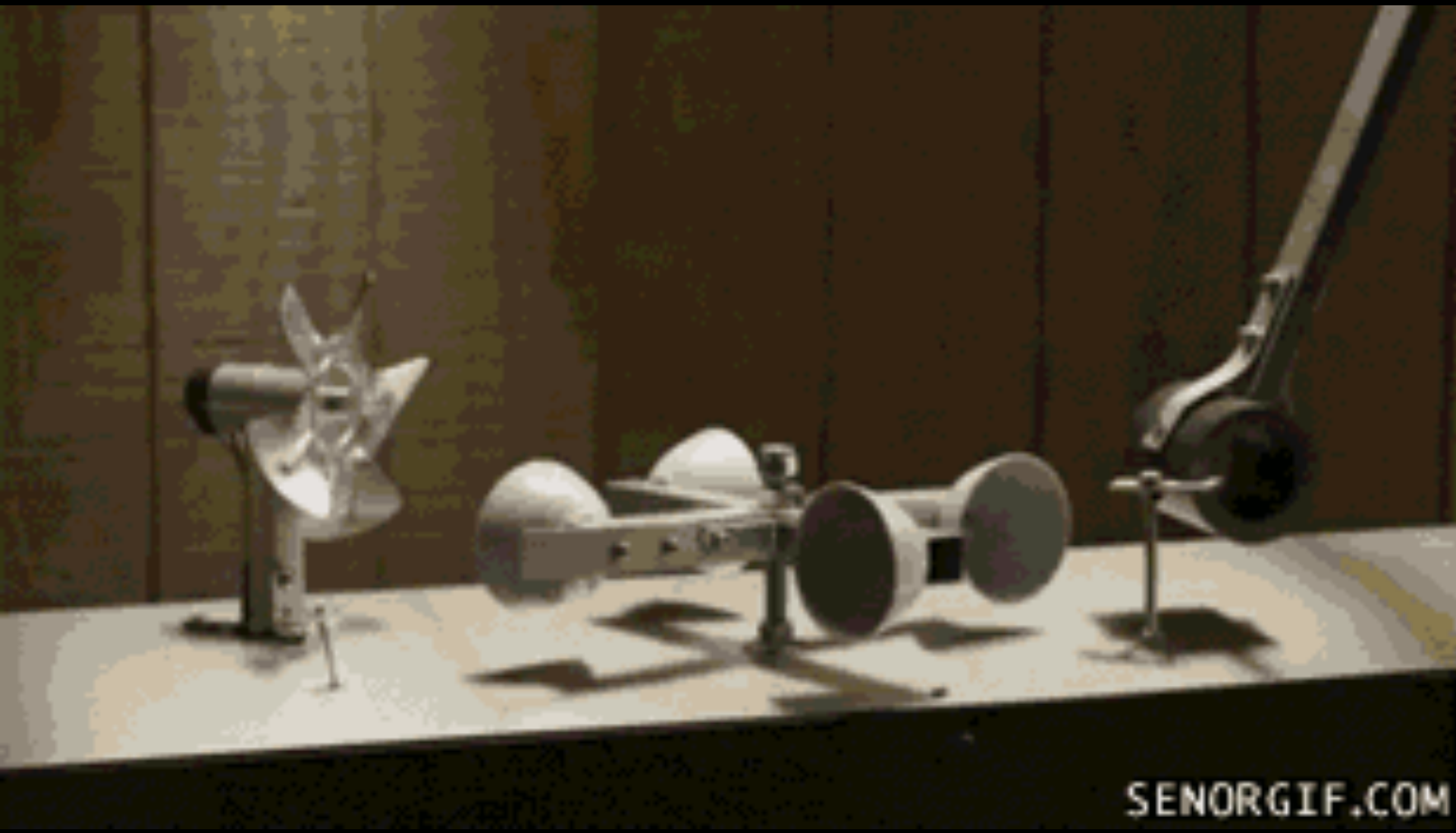
ASSERT, UNITTEST

ENGARDE, NUMPY.TESTING

Next-level code inspection


Town
8200
10
Save
Load


jupyter


SENORGIF.COM


jupyter


jupyter

COVERAGE.PY

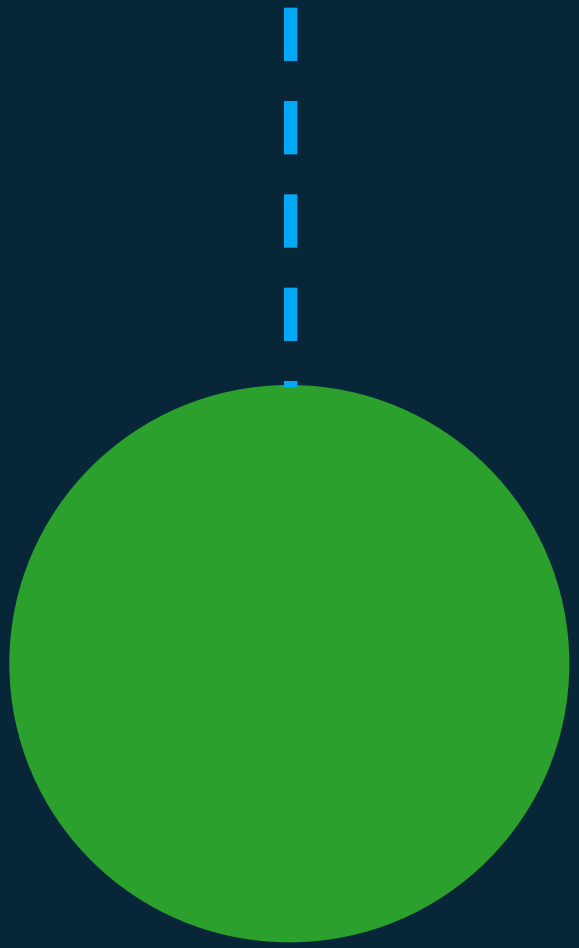
%PRUN

PYCHARM

FLAKE8 LINTING FOR SUBLIME

Not Covered

Version Control	git
Code review	git branch + pull requests
Branching strategy	GitHub Flow
Issue tracking	GitHub Issues + waffle.io
Automating workflows	Make
Documentation Generator	Sphinx
Docstrings	Sphinx ReST
Style guide	flake8
Notebook diffing	nbconvert to .py on save
Configuration isolation	config.py



Questions?

DRIVENDATA
@drivendataorg / @pjbull
peter@drivendata.org