

Data Science is Software: Developer #lifehacks for the Jupyter Data Scientist

PETER BULL
[@drivendataorg](https://drivendata.org) / [@pjbull](https://twitter.com/pjbull)

bit.ly/jupyter-lifehacks

Screenshot of a GitHub repository page for `pjbull / data-science-is-software`.

The repository has 5 commits, 1 branch, 0 releases, and 1 contributor.

Key UI elements highlighted with red boxes:

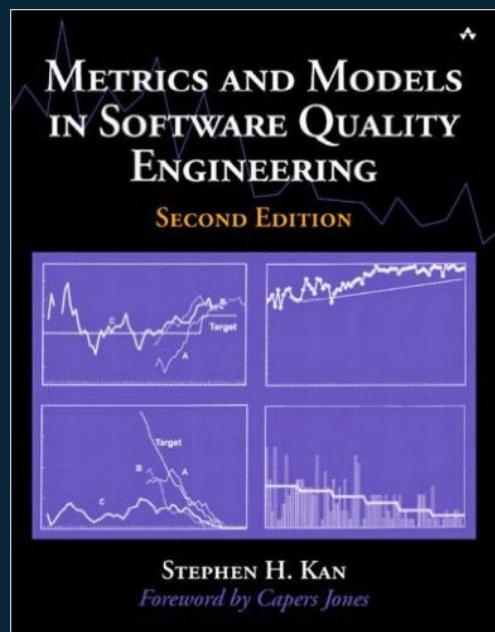
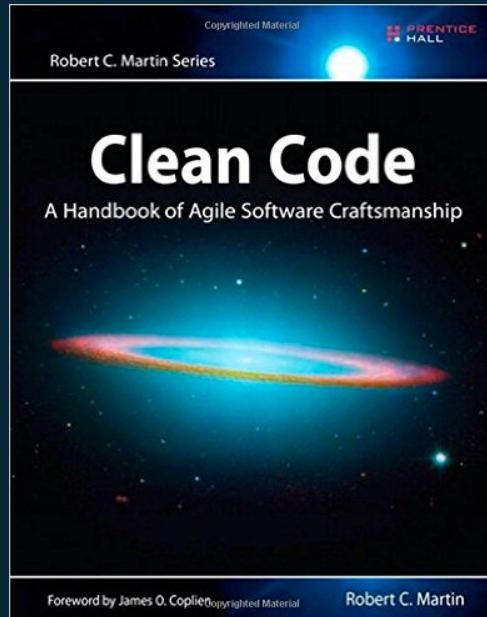
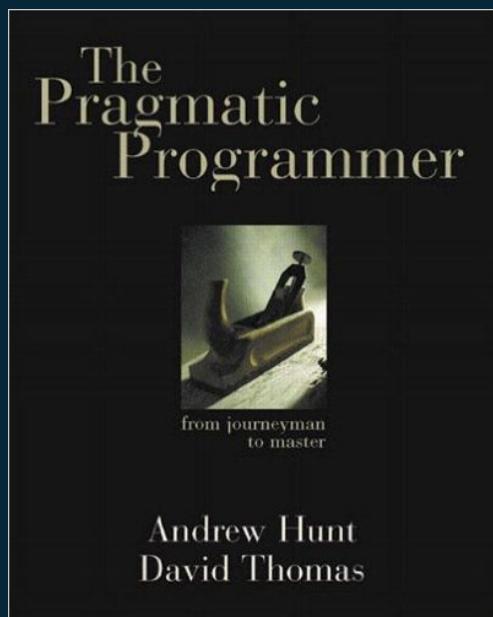
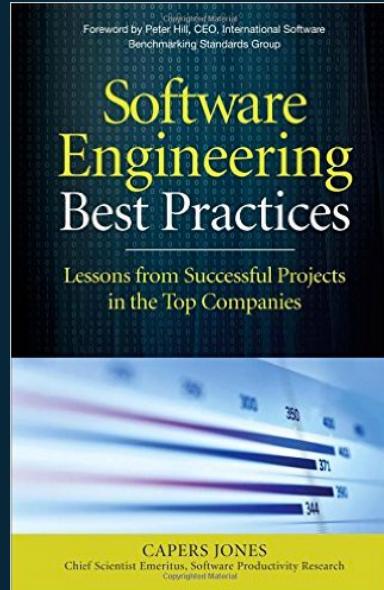
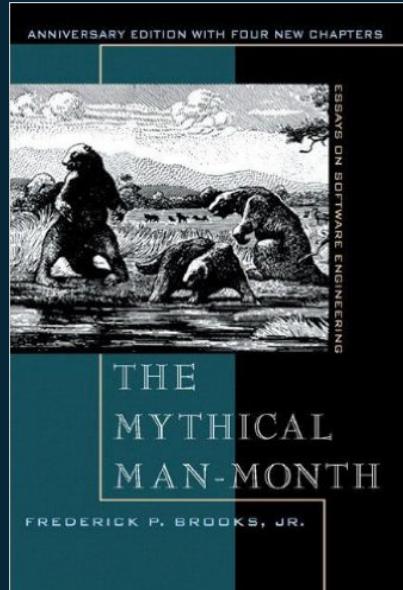
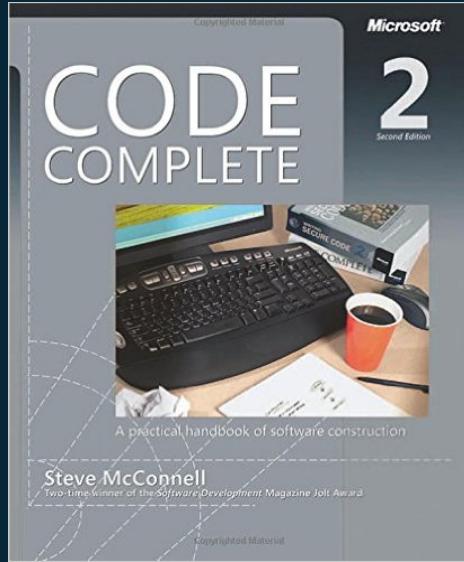
- HTTPS** dropdown and URL (`https://github.com/pjbull`)
- OR** button
- Download ZIP** button

The repository contains the following files and their descriptions:

File	Description	Last Commit
<code>data</code>	Add BDM Notebook	2 days ago
<code>notebooks</code>	Materials update	an hour ago
<code>slides</code>	Materials update	an hour ago
<code>src</code>	Materials update	an hour ago
<code>.gitignore</code>	Talk updates!	12 hours ago
<code>LICENSE</code>	Initial commit	2 days ago
<code>README.md</code>	Talk updates!	12 hours ago
<code>requirements.txt</code>	Materials update	an hour ago

“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)



1

This is my house

Spot 7 differences between
these images with piglets.

<http://www.everydayok.com>







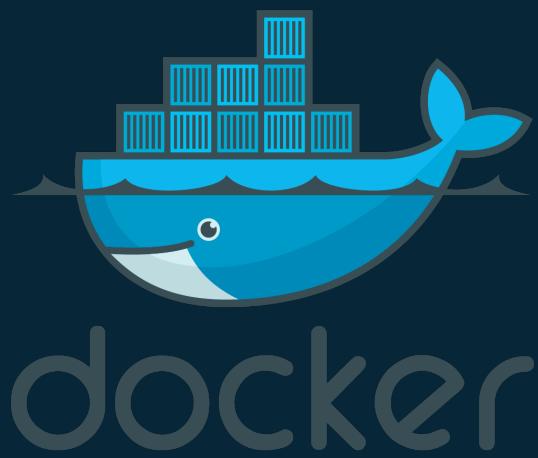
WATERMARK

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

```
.  
|   ├── Gemfile  
|   ├── Gemfile.lock  
|   ├── Guardfile  
|   ├── LICENSE  
|   ├── README.md  
|   ├── README.nitrous.md  
|   └── Rakefile  
|       ├── app  
|       |   ├── assets  
|       |   ├── controllers  
|       |   ├── helpers  
|       |   ├── mailers  
|       |   ├── models  
|       |   └── views  
|       ├── bin  
|       |   ├── bundle  
|       |   ├── rails  
|       |   └── rake  
|       ├── config  
|       |   ├── application.rb  
|       |   ├── boot.rb  
|       |   ├── cucumber.yml  
|       |   ├── database.yml.example  
|       |   ├── environment.rb  
|       |   ├── environments  
|       |   ├── initializers  
|       |   ├── locales  
|       |   └── routes.rb  
|       ├── config.ru  
|       └── db  
|           ├── migrate  
|           ├── schema.rb  
|           └── seeds.rb  
|       ├── features  
|       |   ├── signing_in.feature  
|       |   ├── step_definitions  
|       |   └── support  
|       ├── lib  
|       |   ├── assets  
|       |   └── tasks  
|       ├── log  
|       ├── public  
|       |   ├── 404.html  
|       |   ├── 422.html  
|       |   ├── 500.html  
|       |   ├── assets  
|       |   ├── favicon.ico  
|       |   └── robots.txt  
|       ├── script  
|       |   └── cucumber  
|       ├── spec  
|       |   ├── controllers  
|       |   ├── factories.rb  
|       |   ├── helpers  
|       |   ├── models  
|       |   ├── requests  
|       |   ├── spec_helper.rb  
|       |   └── support  
|       └── vendor  
|           └── assets
```

Django Application

```
.  
|   ├── README.md  
|   └── media  
|       └── init.txt  
|   └── projectname  
|       ├── __init__.py  
|       └── home  
|           ├── __init__.py  
|           ├── models.py  
|           ├── tests.py  
|           └── views.py  
|       ├── manage.py  
|       ├── settings  
|       |   ├── __init__.py  
|       |   ├── default.py  
|       |   └── local.template.py  
|       ├── urls.py  
|       └── wsgi.py  
|   └── requirements.txt  
└── static-assets  
    ├── apple-touch-icon.png  
    ├── css  
    |   └── main.css  
    ├── favicon.ico  
    ├── humans.txt  
    ├── images  
    |   └── init.txt  
    ├── js  
    |   ├── main.coffee  
    |   ├── main.js  
    |   └── main.map  
    └── libs  
        ├── bootstrap-3.3.5  
        ├── font-awesome-4.3.0  
        ├── html5shiv.js  
        ├── jquery  
        └── modernizr  
    └── media -> ../media/  
    └── robots.txt  
└── templates  
    ├── 404.html  
    ├── 500.html  
    ├── base.html  
    └── home.html
```

```
.  
|   └── 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





AUTORELOAD, PDB, DEBUG

Q, MODULES

ASSERT, UNITTEST

ENGARDE, NUMPY.TESTING



4

Next-level code
inspection

8200

19

Town

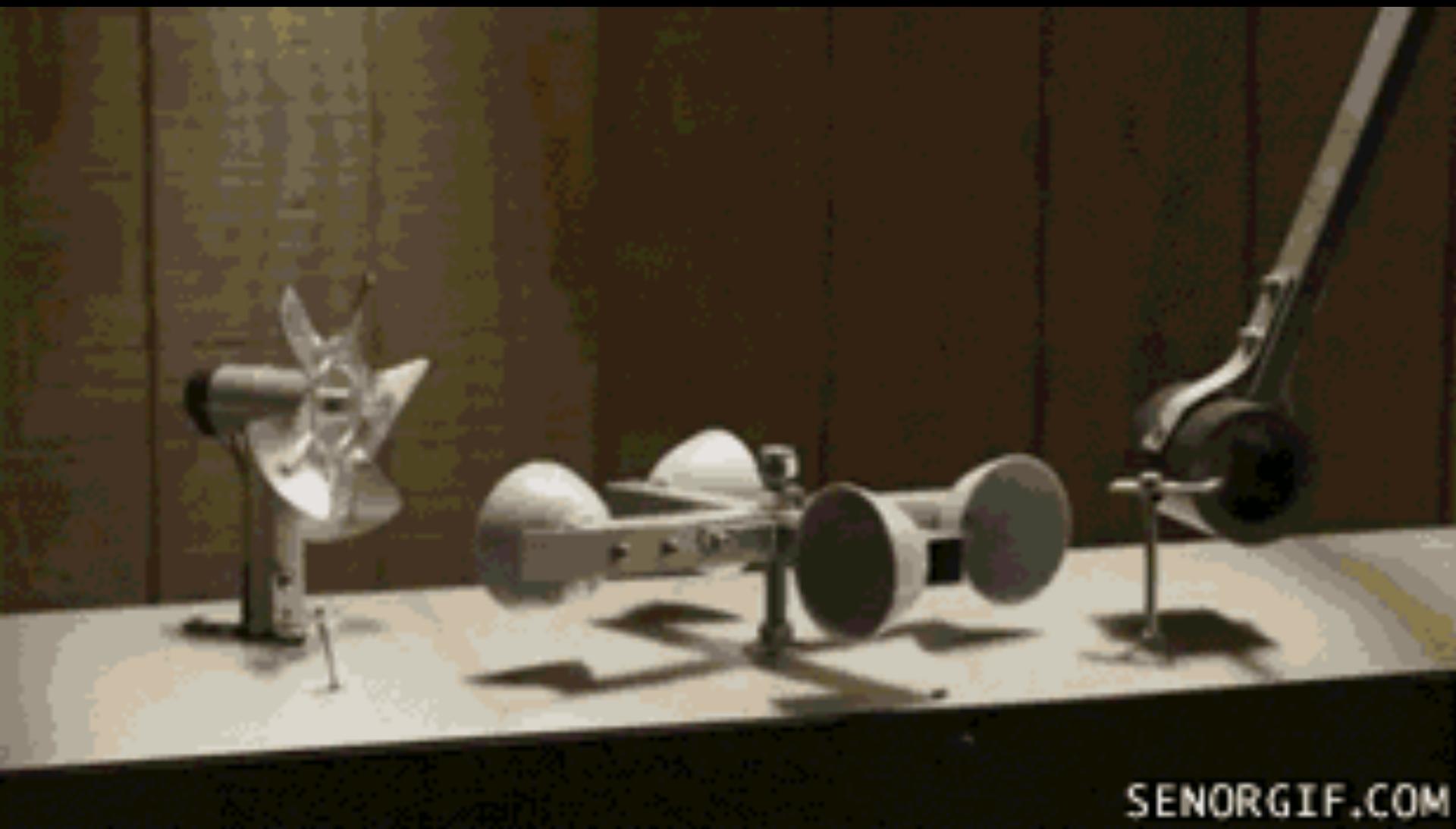


Save

Load







SENORGIF.COM







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](#)

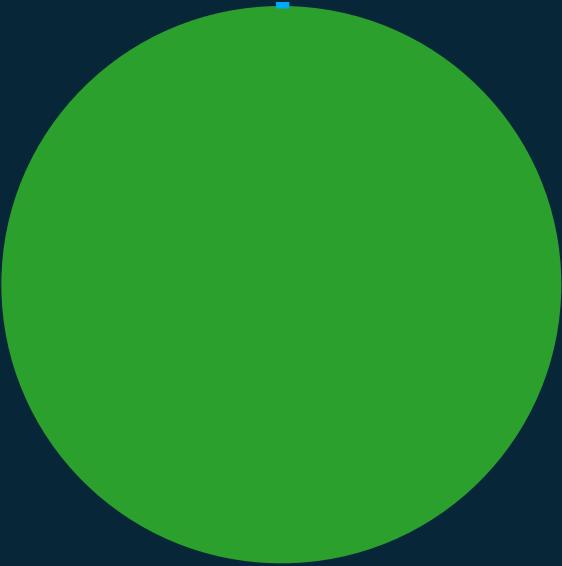
Documentation Generator [Sphinx](#)

Docstrings [Sphinx ReST](#)

Style guide [flake8](#)

Notebook differencing `nbconvert to .py on save`

Configuration isolation `config.py`



Questions?

DRIVENDATA
[@drivendataorg](https://twitter.com/drivendataorg) / [@pjbull](https://twitter.com/pjbull)
peter@drivendata.org