

Keeping Your Code Organized as You Scale It Up

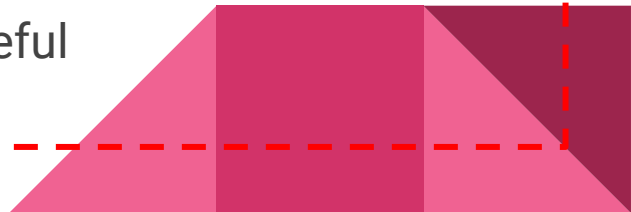
Nicholas A. Del Grosso

Super Python Talks and PyData Munich

February 2018

Size, Complexity, and Organizational Tools

- **10 LOC:** Variable Names become Helpful
- **30 LOC:** Data Collections and Iteration becomes necessary
- **80 LOC:** Functions, please! I can't see the logic on-screen!
- **200 LOC:** This File is a Mess. Split it up!
- **500 LOC:** How do the files relate to each other, again? Glue them together!
- **500 LOC:** This thing breaks whenever we touch it. We need some kind of automated tests to confirm that it will still work!
- **500 LOC:** You know, some of these functions are useful for my other project...

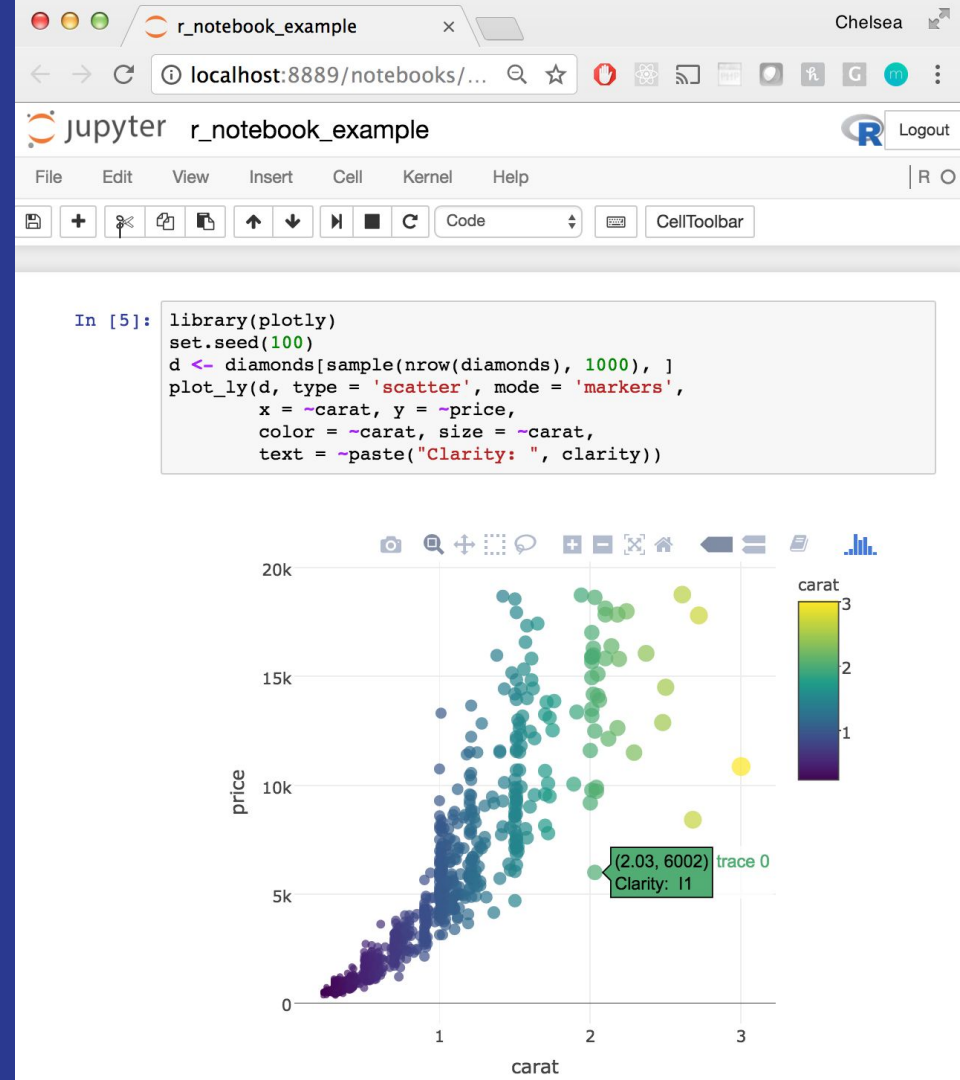
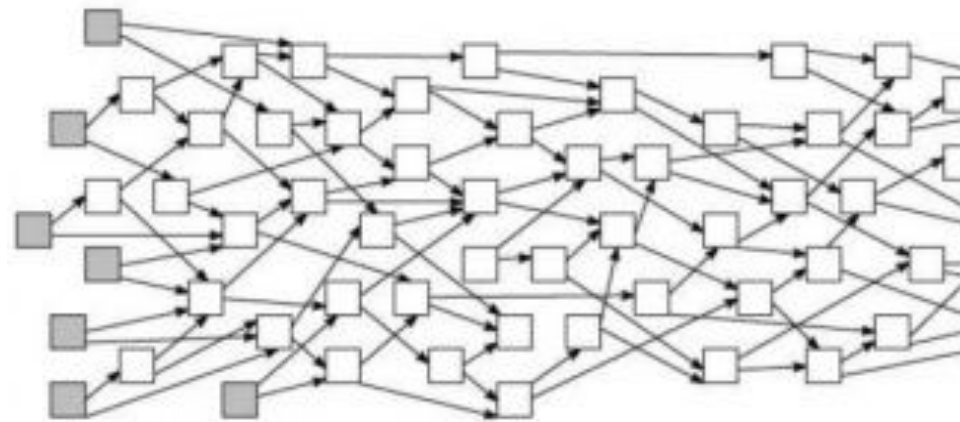


Process of Data Science



From <https://slidemodel.com/>

But, the Data Exploration
Process Isn't So Linear.



Products of Interactive Workflow Over-Reliance

- “The Single All-Powerful Script/Notebook”
 - One single, massive file that does everything.
- “The Comment Switch”
 - Using comments to change variables’ values
 - Using comments to speed up analysis by skipping steps.
 - Using comments to change script outputs.
- “Upside-Down Code”
 - Relying on developer time, not page layout, to determine where code is written.



Today's Goal: Review some Core Tools and Strategies for Managing Data Science Projects

- Version Control (Git)
- Data Workflow / Pipeline Manager (Bash Scripts, PyDoit)
- Notebook Report Export Tools: (Jupyter NbConvert)

Live Coding: Multi-File Project Refactoring Strategies

Data: Our Meetup Events, using the [Meetup API](#)

'https://api.meetup.com/PyData-Munchen/events?&status=past'





Section 1:

Building a Safe Learning Environment with Git

The Most Important Computer Skills let us be Human.

Ctrl-S
(Strg-S)
(Strg-V)

Ctrl-Z
(Strg-Z)

Ctrl-Shift-Z
(Strg-Shift-Z)

Ctrl-C
(Strg-C)

Ctrl-V

...but they are usually temporary.

Version Control:

Permanent

Ctrl-S, Ctrl-Z, Ctrl-C for Folders

Commit: Save

Checkout: Undo, Redo

Clone, Push/Pull: Copy, Paste



Demo



Takeaway

Git is a Version Control System.

Use it to ***Save***, ***Backup***, and ***Undo*** your code.

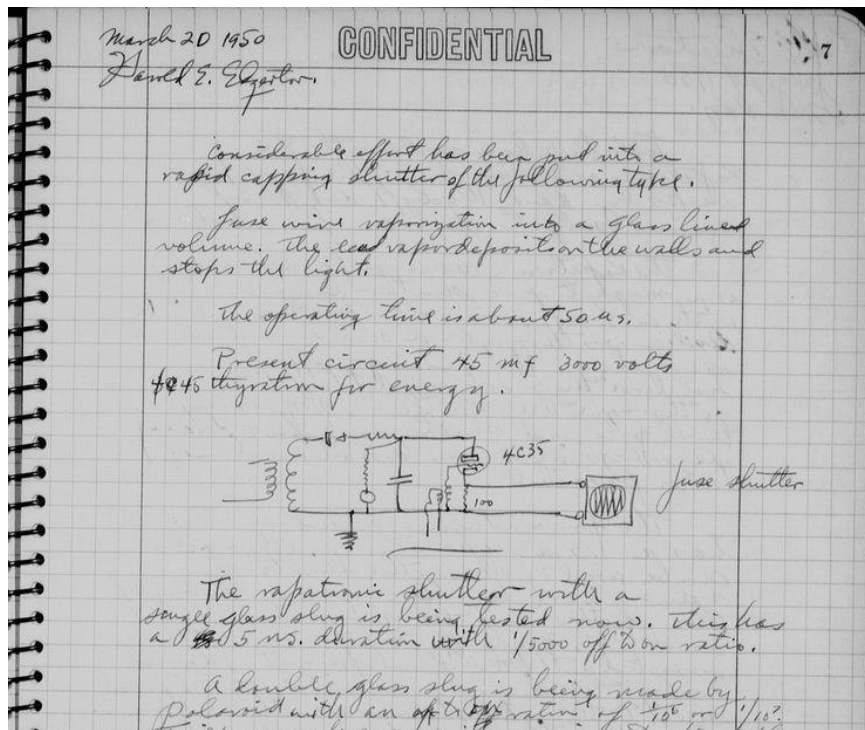




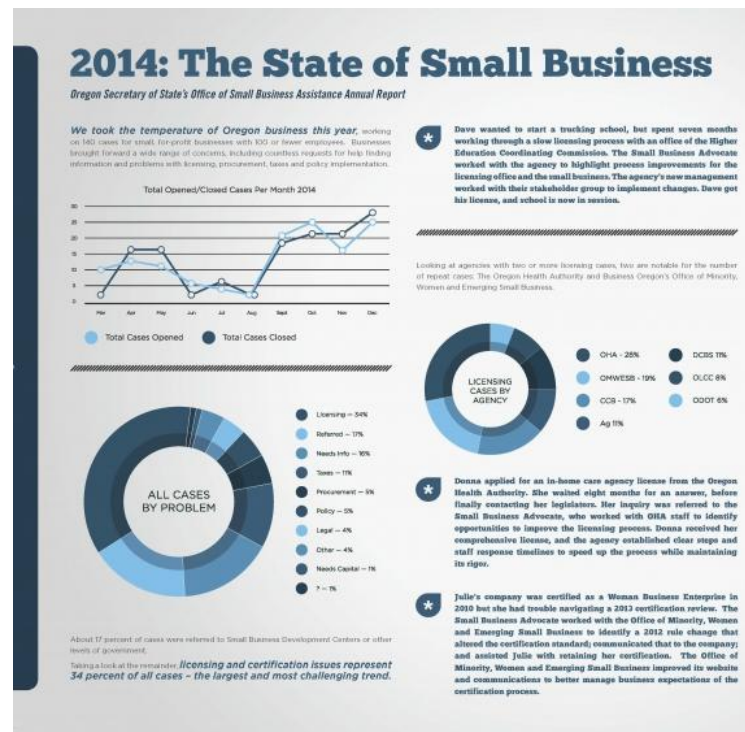
Section 2: Exporting Your Notebook to a Read-Only Format

Jupyter: A Lab Notebook and Reporting Tool

Lab Notebooks: Messy Permanent Records



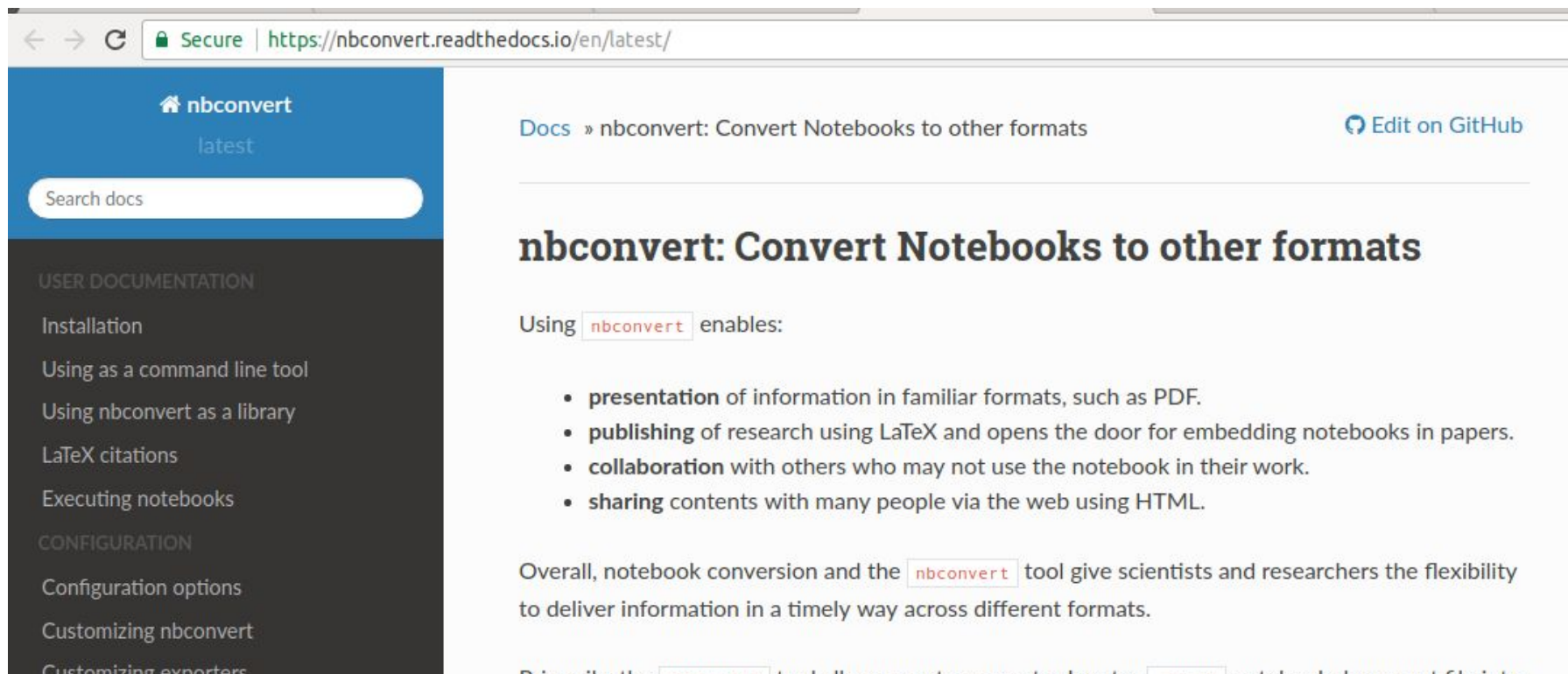
Reports: Pretty Self-Documenting Figures



```
jupyter nbconvert --to html *.ipynb
```


```
jupyter nbconvert --to html *.ipynb --FilesWriter.build_directory=html
```

```
jupyter nbconvert --to notebook --execute --inplace *.ipynb
```



The screenshot shows a web browser displaying the nbconvert documentation page. The browser's address bar shows the URL `https://nbconvert.readthedocs.io/en/latest/`. The page has a blue header with the nbconvert logo and the word "latest". A search bar is present in the header. On the left, there is a dark sidebar with a list of navigation links under "USER DOCUMENTATION" and "CONFIGURATION". The main content area has a title "nbconvert: Convert Notebooks to other formats" and a link to "Edit on GitHub". Below the title, it says "Using nbconvert enables:" followed by a bulleted list of features: presentation, publishing, collaboration, and sharing. At the bottom, a paragraph states that notebook conversion and the nbconvert tool give scientists and researchers the flexibility to deliver information in a timely way across different formats.

← → ↻ Secure | <https://nbconvert.readthedocs.io/en/latest/>

 nbconvert
latest

Search docs

USER DOCUMENTATION

- Installation
- Using as a command line tool
- Using nbconvert as a library
- LaTeX citations
- Executing notebooks

CONFIGURATION

- Configuration options
- Customizing nbconvert
- Customizing exporters

Docs » nbconvert: Convert Notebooks to other formats [Edit on GitHub](#)

nbconvert: Convert Notebooks to other formats

Using `nbconvert` enables:

- **presentation** of information in familiar formats, such as PDF.
- **publishing** of research using LaTeX and opens the door for embedding notebooks in papers.
- **collaboration** with others who may not use the notebook in their work.
- **sharing** contents with many people via the web using HTML.

Overall, notebook conversion and the `nbconvert` tool give scientists and researchers the flexibility to deliver information in a timely way across different formats.

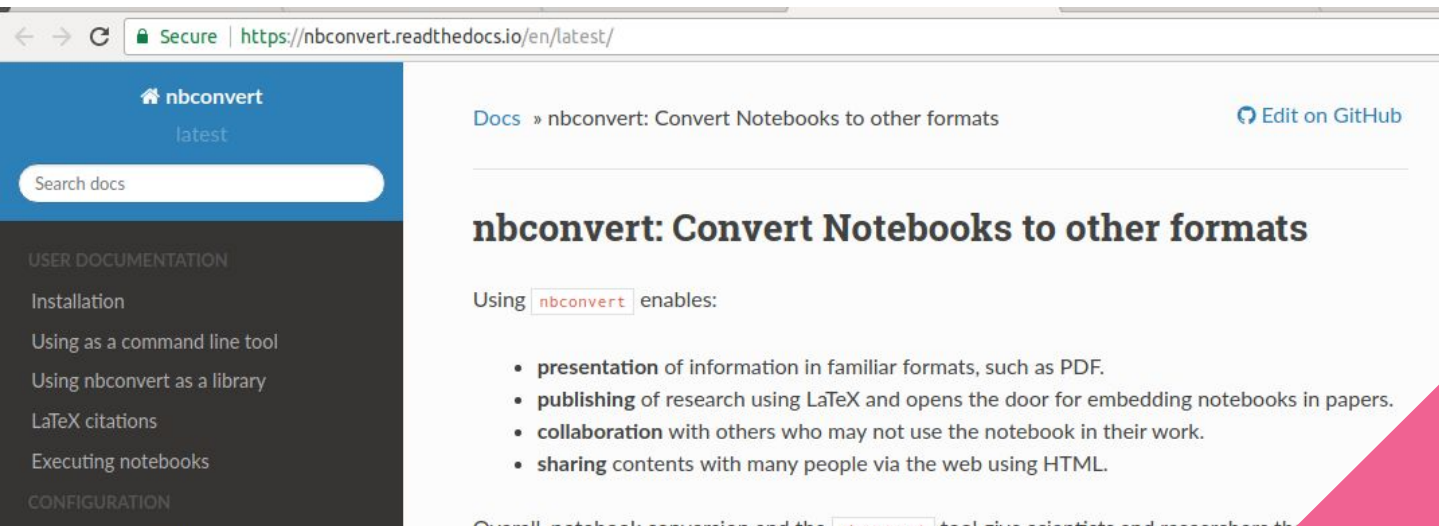
Print this page

Demo: Making Reports Directory

```
jupyter nbconvert --to html *.ipynb
```

```
jupyter nbconvert --to html *.ipynb --FilesWriter.build_directory=html
```

```
jupyter nbconvert --to notebook --execute --inplace *.ipynb
```

A screenshot of a web browser displaying the nbconvert website. The browser's address bar shows the URL 'https://nbconvert.readthedocs.io/en/latest/'. The website has a blue header with the 'nbconvert' logo and a search bar. A dark sidebar on the left contains a 'USER DOCUMENTATION' menu with links to 'Installation', 'Using as a command line tool', 'Using nbconvert as a library', 'LaTeX citations', and 'Executing notebooks'. The main content area has a breadcrumb trail 'Docs » nbconvert: Convert Notebooks to other formats' and an 'Edit on GitHub' link. The title 'nbconvert: Convert Notebooks to other formats' is prominently displayed. Below it, the text 'Using nbconvert enables:' is followed by a bulleted list of features: presentation, publishing, collaboration, and sharing. The bottom of the page shows the start of a paragraph about overall notebook conversion.

Secure | <https://nbconvert.readthedocs.io/en/latest/>

nbconvert
latest

Search docs

USER DOCUMENTATION

- Installation
- Using as a command line tool
- Using nbconvert as a library
- LaTeX citations
- Executing notebooks

CONFIGURATION

Docs » nbconvert: Convert Notebooks to other formats [Edit on GitHub](#)

nbconvert: Convert Notebooks to other formats

Using `nbconvert` enables:

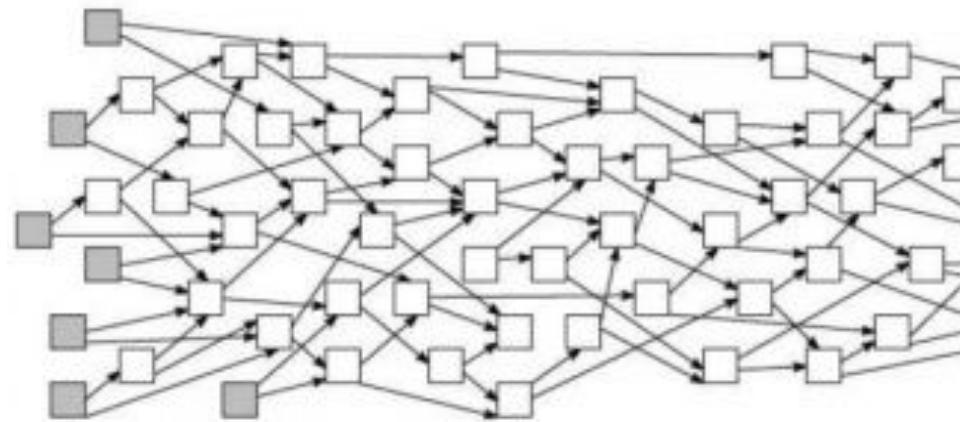
- **presentation** of information in familiar formats, such as PDF.
- **publishing** of research using LaTeX and opens the door for embedding notebooks in papers.
- **collaboration** with others who may not use the notebook in their work.
- **sharing** contents with many people via the web using HTML.

Overall, notebook conversion and the `nbconvert` tool give scientists and researchers the



Section 3: Gluing Your Scripts Together

Workflow Managers
Explicitly Describe the Order
and Dependencies of a
Project.



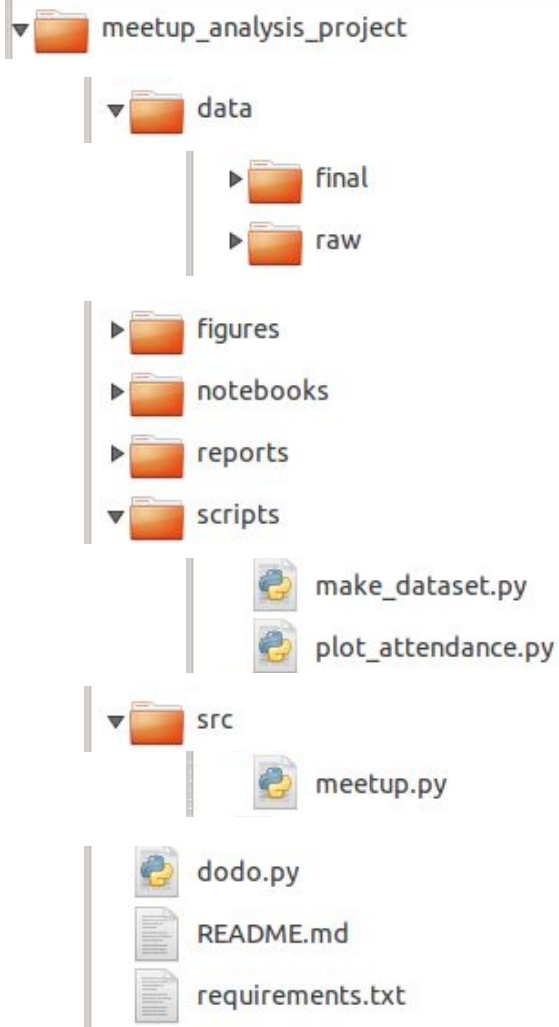
Make (software)

From Wikipedia, the free encyclopedia

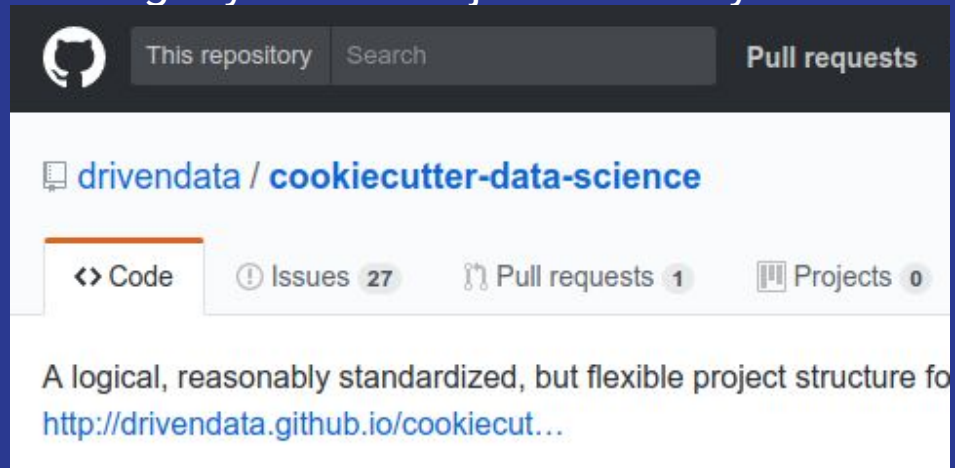
DoIt Automation Tool

Wtgit





In the process, we'll build up a slightly-better Project Directory



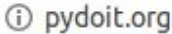
The resulting directory structure











The directory structure of your new project looks like this:

```
├── LICENSE
├── Makefile
├── README.md
├── data
│   ├── external
│   ├── interim
│   ├── processed
│   └── raw
```

<- Makefile with commands like `make data` or `make tra`
<- The top-level README for developers using this proje
<- Data from third party sources.
<- Intermediate data that has been transformed.
<- The final, canonical data sets for modeling.
<- The original, immutable data dump.

PyDoIt is Easy to Learn, Lightweight, and Flexible.

 pydoit.org



DoIt Automation Tool

[Home](#) | [Documentation](#) | [Code](#) | [Issues](#) | [Download](#) | [StackOverflow](#) | [Forum](#) | [Twitter](#)

***doit* is a task management & automation tool**

***doit* comes from the idea of bringing the power of build-tools to execute any kind of task**

doit is a modern open-source build-tool written in python designed to be simple to use and flexible to deal with complex work-flows. It is specially suitable for building and managing custom work-flows where there is no out-of-the-box solution available.

Table Of Contents

- [introduction](#)
- [What people are saying about *doit*](#)
- [Project Details](#)
- [Status](#)

Quick search

Live Coding

Project Refactoring:
Big Notebook to Big Project



Thank you!

Project Files Available at