Agenda

Boost your reproducibility with Binder

- 13:00 Registration and introductions
- 13:10 Introduction to the workshop and The Turing Way
- 13:20 Presentation: Why you need a reproducible computing environment and how Binder can help
- 14:30 Coffee break
- 15:00 Code along demo: Zero to Binder, build a Binder resource
- 16:00 Build your own Binder
- 16:30 Feedback, group picture and close

The Alan Turing Institute

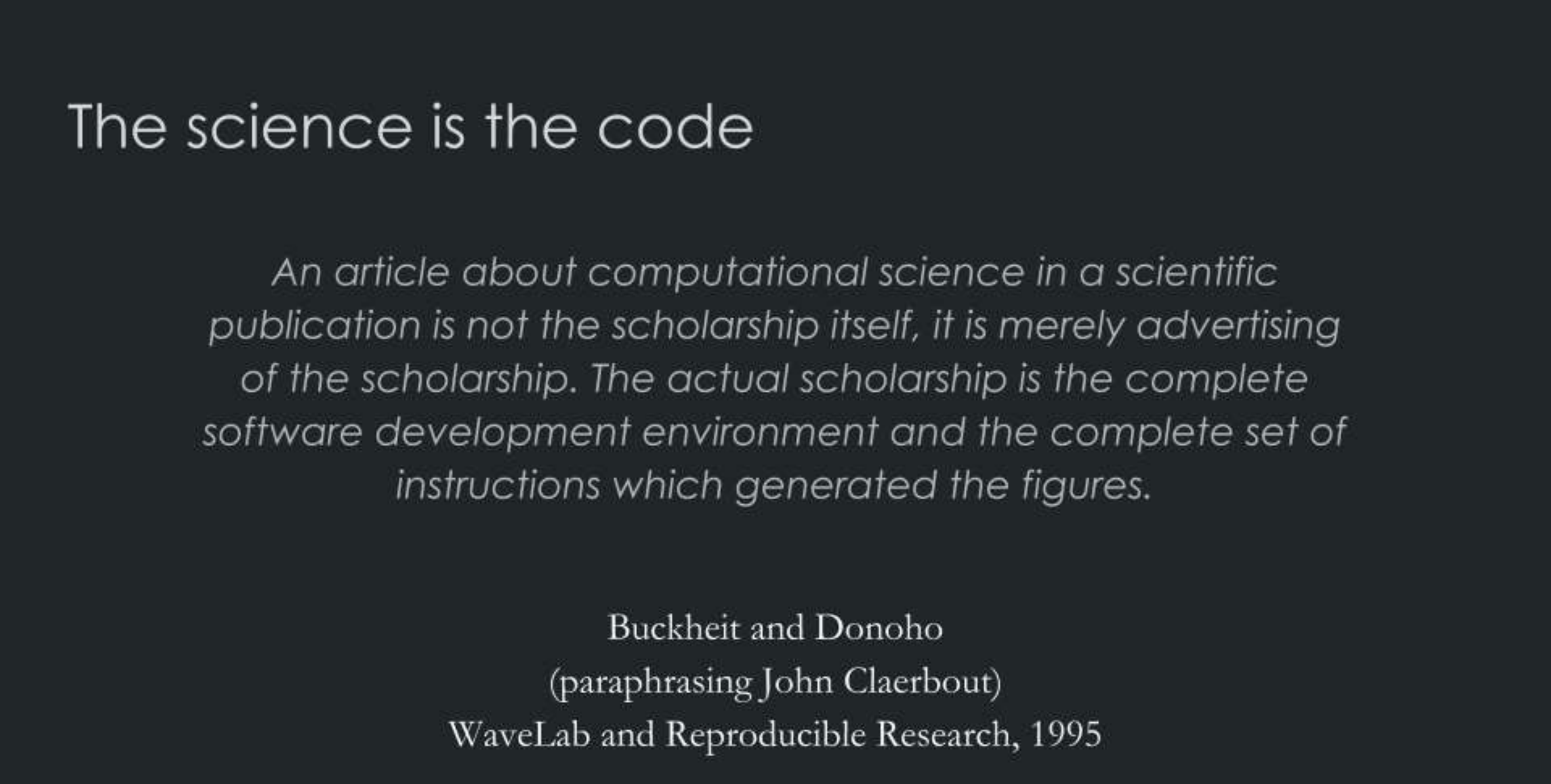
Reproducible Computational Environments

Kirstie Whitaker

Pronouns: she/her

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.2598529



The science is the code

An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.

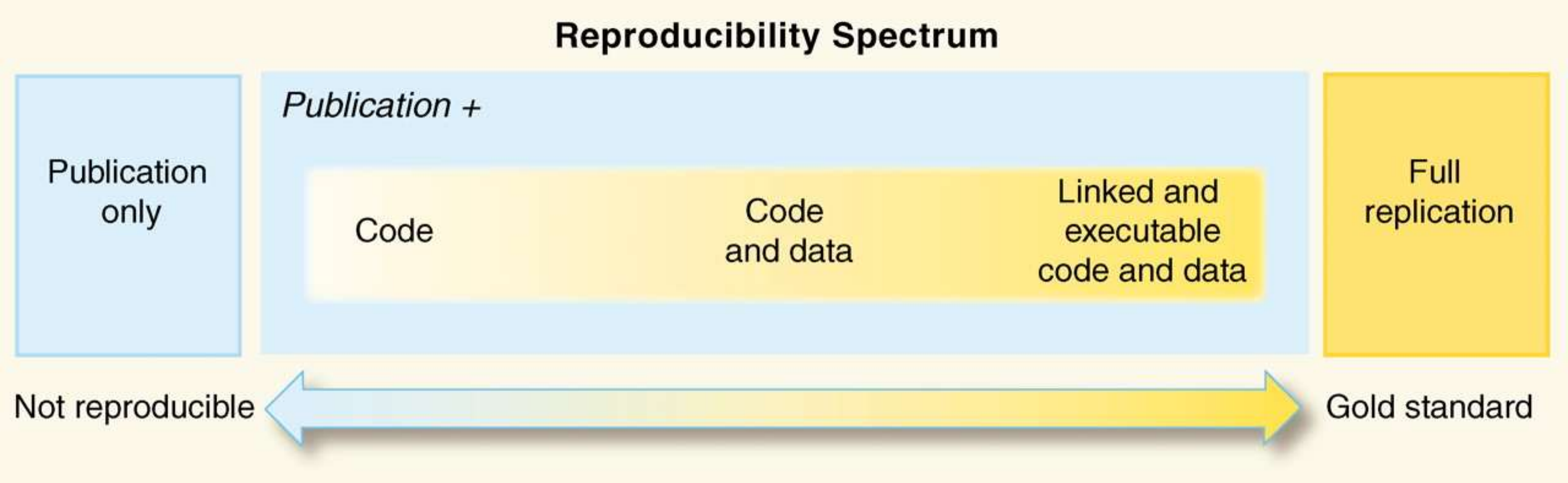
Buckheit and Donoho
(paraphrasing John Claerbout)
WaveLab and Reproducible Research, 1995

Slide courtesy of Chris Holdraf and the Jupyter Team

Upsetting take home message

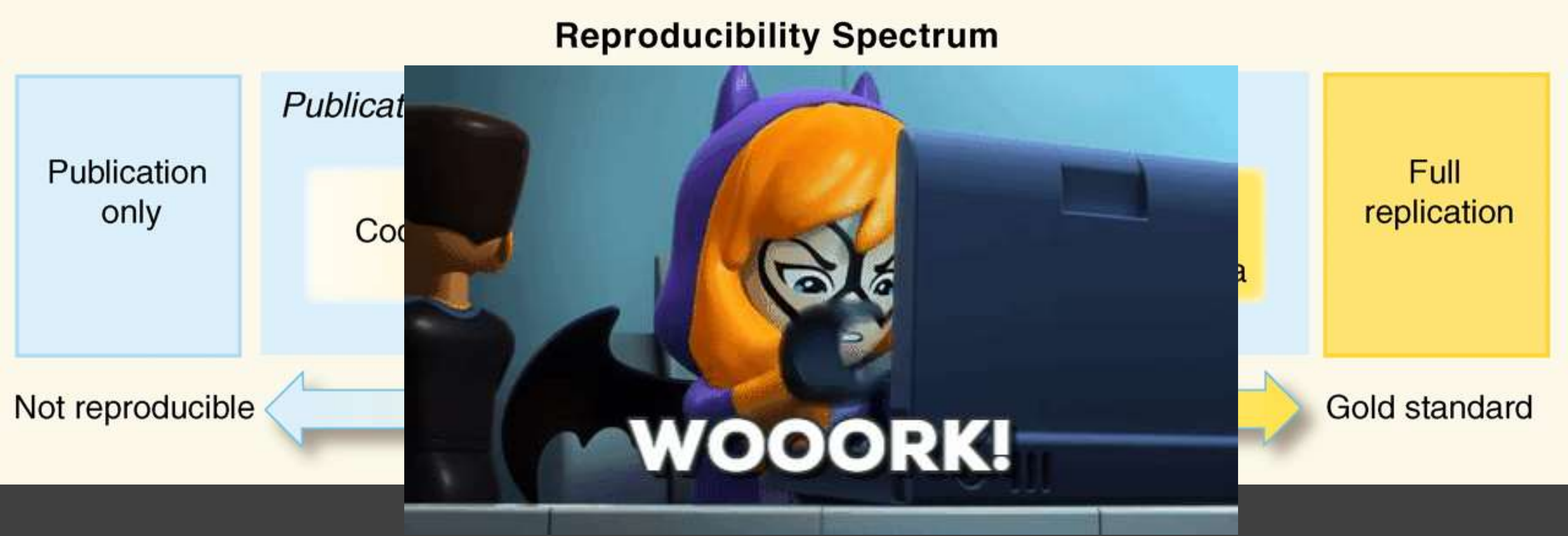
Sharing your code and data isn't enough

You need the computational environment too



Peng, 2011, doi: 10.1126/science.1213847

You need the computational environment too



#TuringWay @kirstie_j @mybinderteam

Peng, 2011, doi: 10.1126/science.1213847

https://doi.org/10.5281/zenodo.2598529

The computational environment includes:

- Hardware (GPU, CPU)
- Operating system (mac, windows, linux)
- Software
  - Language version
  - Package version(s)

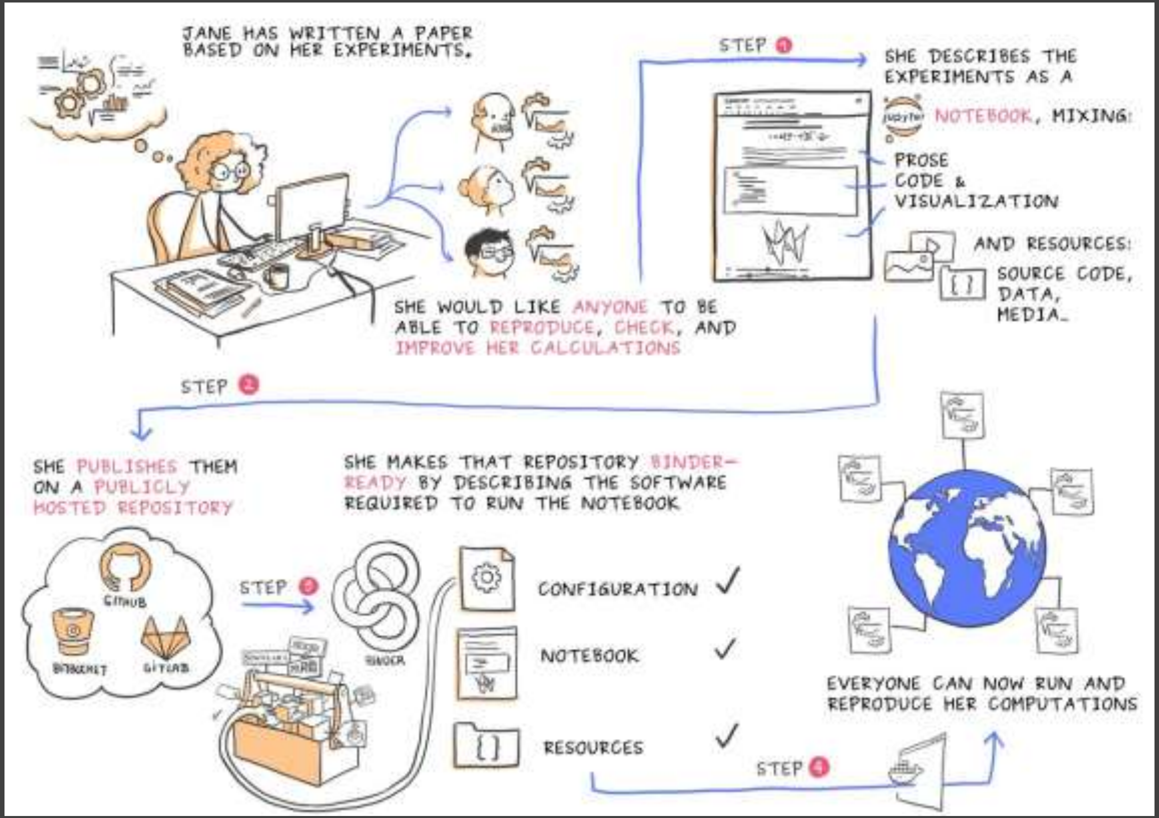
And all the interactions between the layers

#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.2598529

What is Binder?

#TuringWay @kirstie_j @mybinderteam

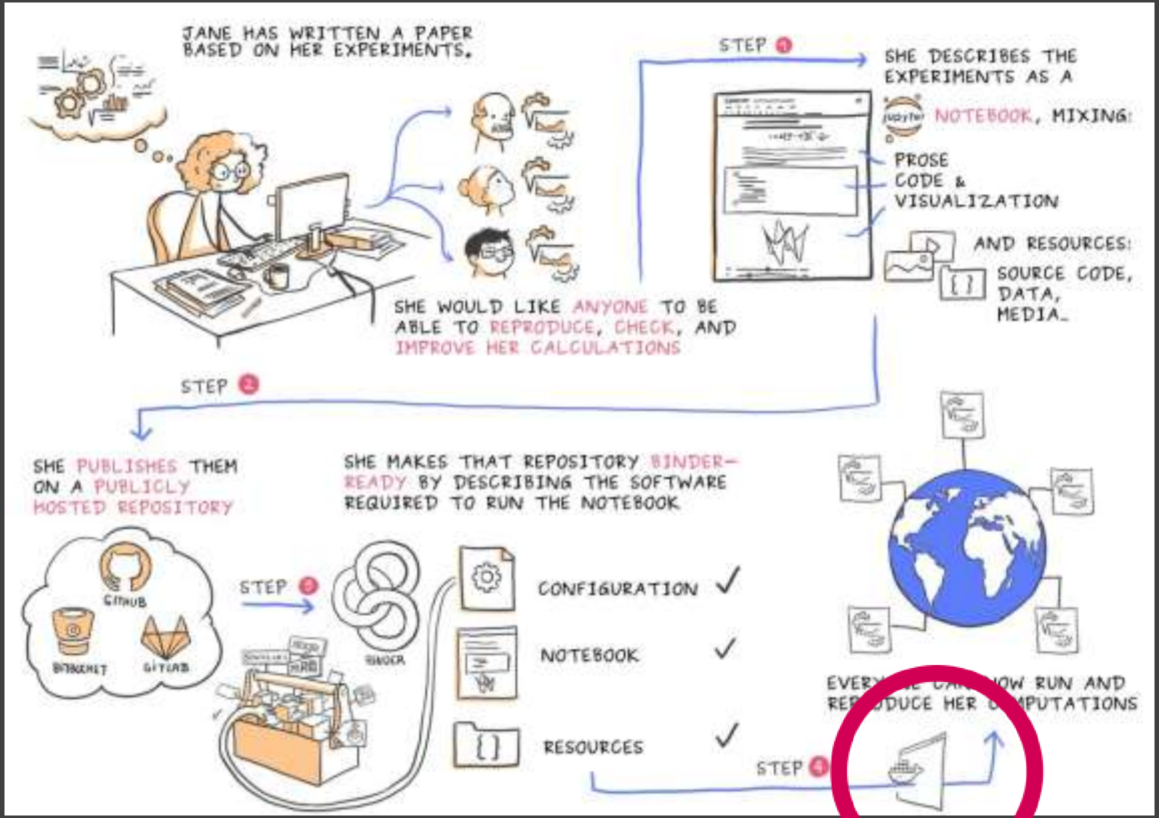
https://doi.org/10.5281/zenodo.2598529


JANE HAS WRITTEN A PAPER BASED ON HER EXPERIMENTS.
SHE WOULD LIKE ANYONE TO BE ABLE TO REPRODUCE, CHECK, AND IMPROVE HER CALCULATIONS
STEP 1
SHE DESCRIBES THE EXPERIMENTS AS A
jupyter NOTEBOOK, MIXING:
PROSE
CODE &
VISUALIZATION
AND RESOURCES:
SOURCE CODE,
DATA,
MEDIA...
STEP 2
SHE PUBLISHES THEM ON A PUBLICLY HOSTED REPOSITORY
GITHUB
BITBUCKET
GITLAB
STEP 3
BINDER
SHE MAKES THAT REPOSITORY BINDER-READY BY DESCRIBING THE SOFTWARE REQUIRED TO RUN THE NOTEBOOK
CONFIGURATION ✓
NOTEBOOK ✓
RESOURCES ✓
STEP 4
EVERYONE CAN NOW RUN AND REPRODUCE HER COMPUTATIONS

Courtesy of Juliette Taka: https://twitter.com/mybinderteam/status/1082556317842264064

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.2598529

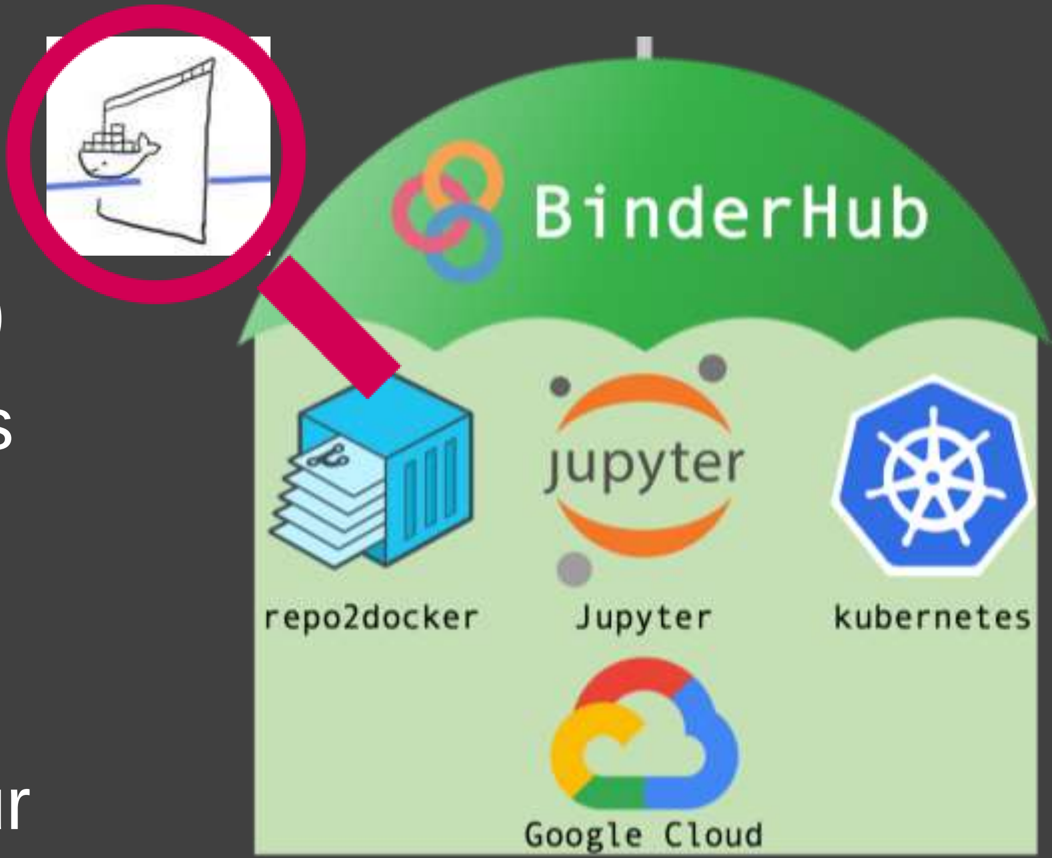

JANE HAS WRITTEN A PAPER BASED ON HER EXPERIMENTS.
SHE WOULD LIKE ANYONE TO BE ABLE TO REPRODUCE, CHECK, AND IMPROVE HER CALCULATIONS
STEP 1
SHE DESCRIBES THE EXPERIMENTS AS A jupyter NOTEBOOK, MIXING:
PROSE
CODE &
VISUALIZATION
AND RESOURCES:
SOURCE CODE,
DATA,
MEDIA...
STEP 2
SHE PUBLISHES THEM ON A PUBLICLY HOSTED REPOSITORY
GITHUB
BITBUCKET
GITLAB
STEP 3
BINDER
SHE MAKES THAT REPOSITORY BINDER-READY BY DESCRIBING THE SOFTWARE REQUIRED TO RUN THE NOTEBOOK
CONFIGURATION ✓
NOTEBOOK ✓
RESOURCES ✓
STEP 4
EVERY... CAN NOW RUN AND REPRODUCE HER COMPUTATIONS

Courtesy of Juliette Taka: https://twitter.com/mybinderteam/status/1082556317842264064

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.2598529

- Coordinate cloud computing resources with Kubernetes (k8s)
- Make it easy for users to access with a JupyterHub
- Set up the environment from your GitHub repository



https://binderhub.readthedocs.io
#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.2598529

Sarah Gibson

“It took me a while to feel like I knew enough to contribute to Binder. But the team are always so excited to have my input. Its really motivating to be part of such a welcoming community.”

https://www.turing.ac.uk/people/researchers/sarah-gibson

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.2598529

- Check analysis on my phone
- Share the responsibility with busy PIs
- Requires version control, capturing environment and new build for each change

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.2598529

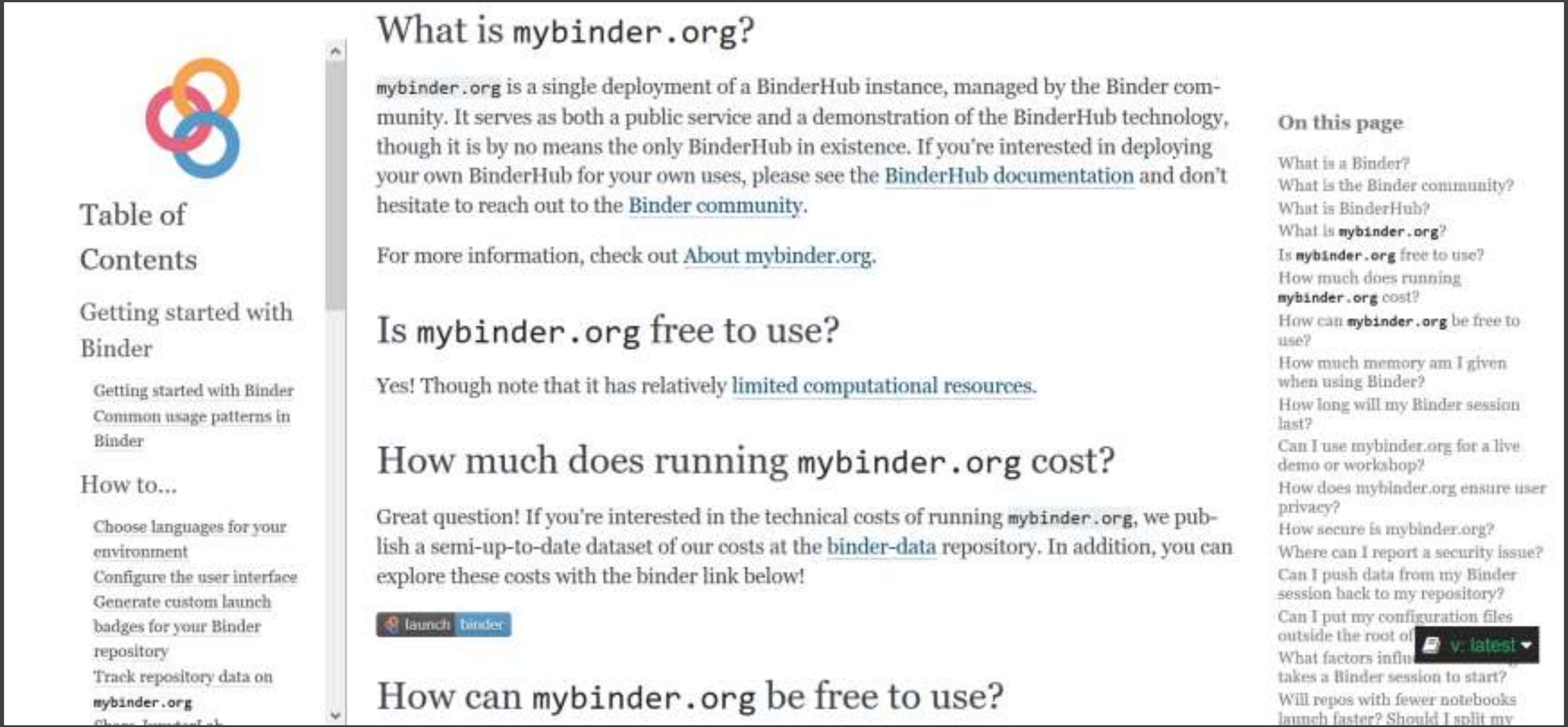


Table of Contents

Getting started with Binder

How to...

What is `mybinder.org`?

`mybinder.org` is a single deployment of a BinderHub instance, managed by the Binder community. It serves as both a public service and a demonstration of the BinderHub technology, though it is by no means the only BinderHub in existence. If you're interested in deploying your own BinderHub for your own uses, please see the BinderHub documentation and don't hesitate to reach out to the Binder community.

For more information, check out About mybinder.org.

Is `mybinder.org` free to use?

Yes! Though note that it has relatively limited computational resources.

How much does running `mybinder.org` cost?

Great question! If you're interested in the technical costs of running `mybinder.org`, we publish a semi-up-to-date dataset of our costs at the binder-data repository. In addition, you can explore these costs with the binder link below!

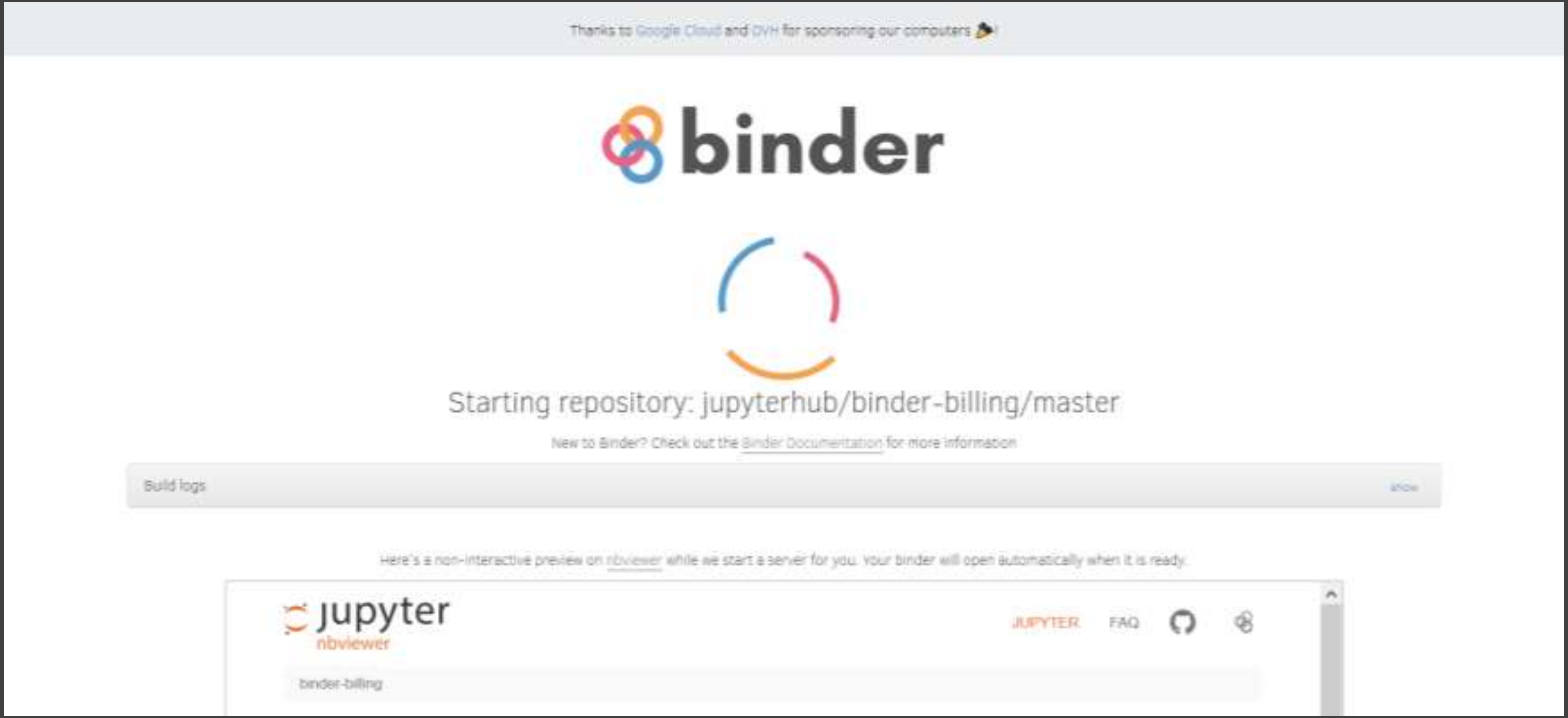
How can `mybinder.org` be free to use?

On this page

https://mybinder.readthedocs.io/en/latest/faq.html#how-much-does-running-mybinder-org-cost

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.2598529



Thanks to Google Cloud and OVH for sponsoring our computers 🎉!

Starting repository: jupyterhub/binder-billing/master

New to Binder? Check out the Binder Documentation for more information.

Build logs show

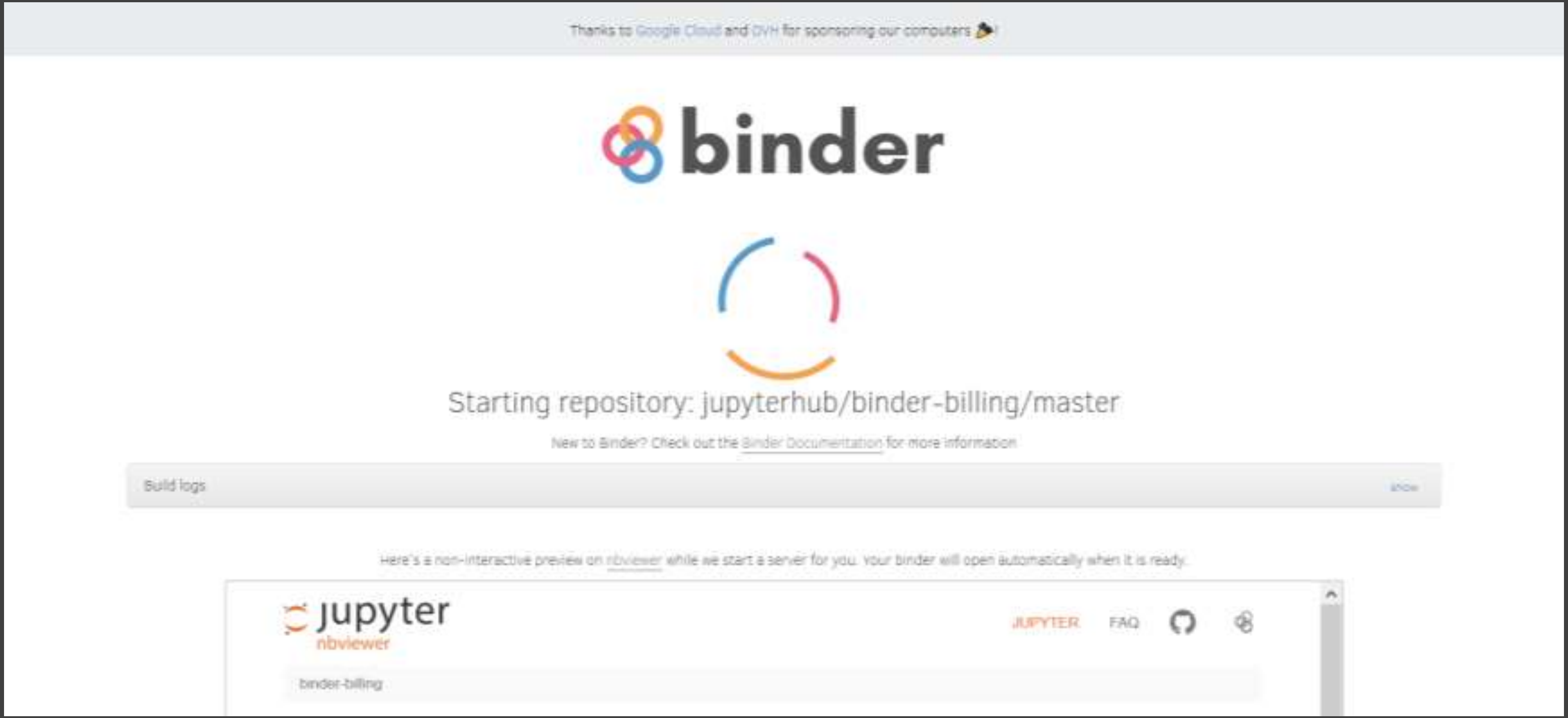
Here's a non-interactive preview on nbviewer while we start a server for you. Your binder will open automatically when it is ready.

binder-billing

https://mybinder.readthedocs.io/en/latest/faq.html#how-much-does-running-mybinder-org-cost

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.2598529



Thanks to Google Cloud and OVH for sponsoring our computers 🎉!

Starting repository: jupyterhub/binder-billing/master

New to Binder? Check out the Binder Documentation for more information.

Build logs

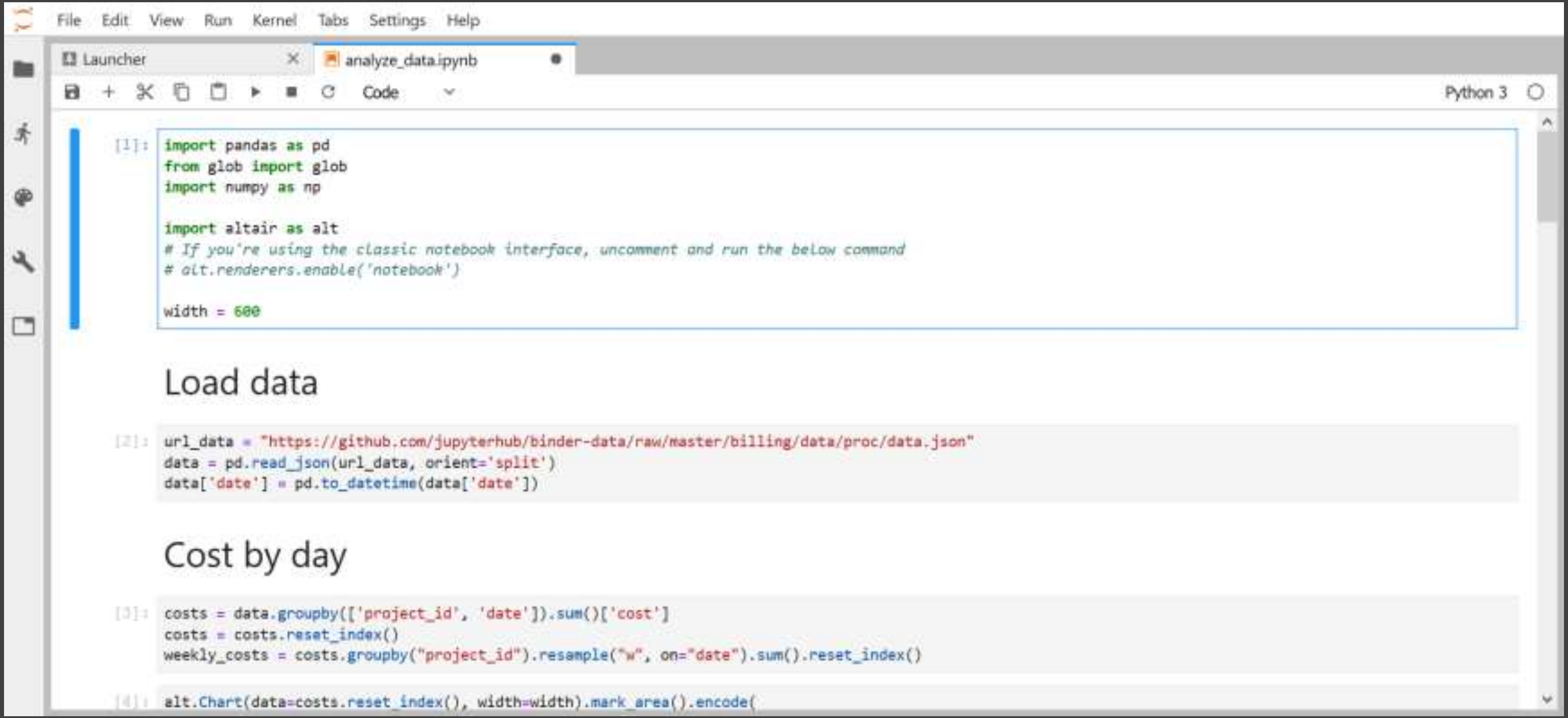
show

Here's a non-interactive preview on nbviewer while we start a server for you. Your binder will open automatically when it is ready.

https://mybinder.readthedocs.io/en/latest/faq.html#how-much-does-running-mybinder-org-cost

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.2598529



```
import pandas as pd
from glob import glob
import numpy as np

import altair as alt
# If you're using the classic notebook interface, uncomment and run the below command
# alt.renderers.enable('notebook')

width = 600
```

Load data

```
url_data = "https://github.com/jupyterhub/binder-data/raw/master/billing/data/proc/data.json"
data = pd.read_json(url_data, orient='split')
data['date'] = pd.to_datetime(data['date'])
```

Cost by day

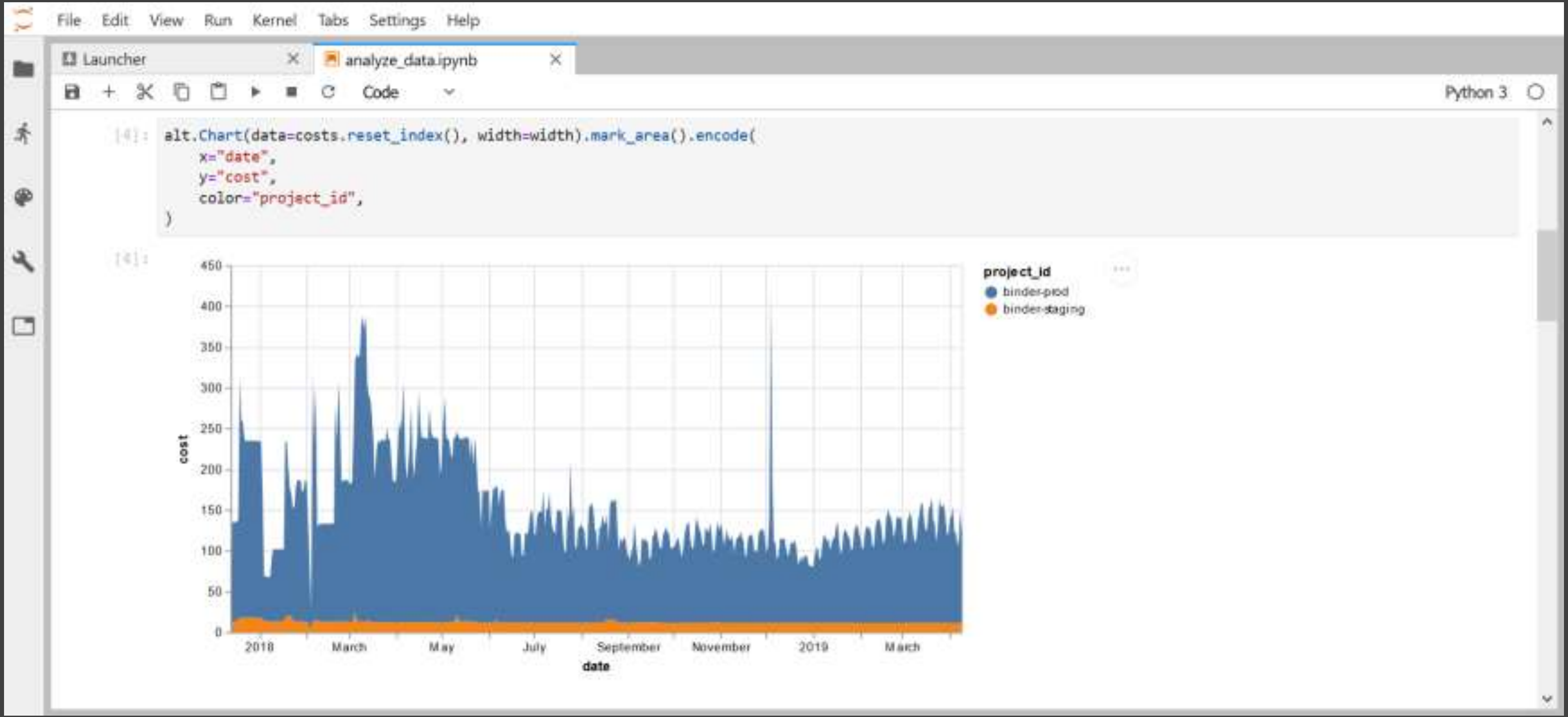
```
costs = data.groupby(['project_id', 'date']).sum()['cost']
costs = costs.reset_index()
weekly_costs = costs.groupby("project_id").resample("w", on="date").sum().reset_index()
```

```
alt.Chart(data=costs.reset_index(), width=width).mark_area().encode(
```

https://mybinder.readthedocs.io/en/latest/faq.html#how-much-does-running-mybinder-org-cost

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.2598529



```
alt.Chart(data=costs.reset_index(), width=width).mark_area().encode(
    x="date",
    y="cost",
    color="project_id",
)
```

https://mybinder.readthedocs.io/en/latest/faq.html#how-much-does-running-mybinder-org-cost

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.2598529

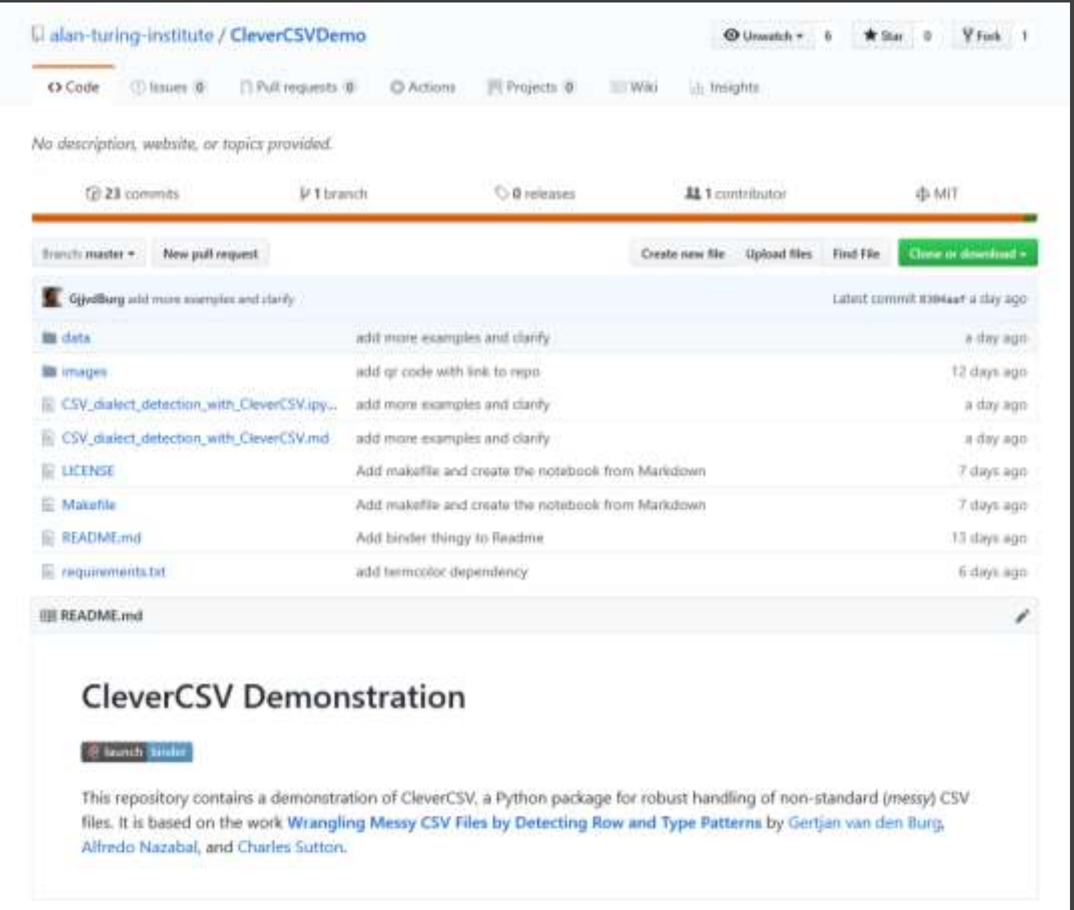
Gertjan van den Burg

“The fun part of data science is the modelling. Being able to read in information from a csv file should not be the hardest part.”

https://gertjanvandenburg.com

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.2598529



No description, website, or topics provided.

GjjvdBurg add more examples and clarify — Latest commit [illegible] a day ago

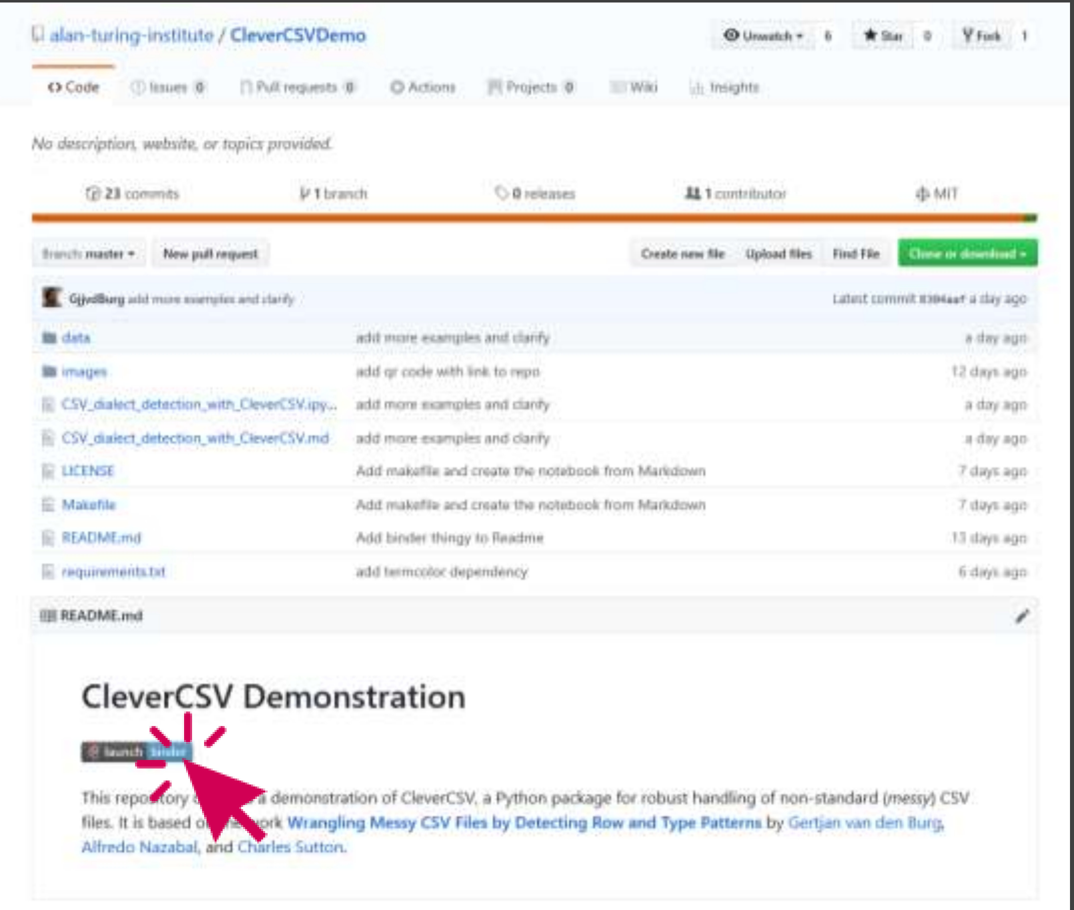
data	add more examples and clarify	a day ago
images	add qr code with link to repo	12 days ago
CSV_dialect_detection_with_CleverCSV.ipy...	add more examples and clarify	a day ago
CSV_dialect_detection_with_CleverCSV.md	add more examples and clarify	a day ago
LICENSE	Add makefile and create the notebook from Markdown	7 days ago
Makefile	Add makefile and create the notebook from Markdown	7 days ago
README.md	Add binder thingy to Readme	13 days ago
requirements.txt	add termcolor dependency	6 days ago

README.md

CleverCSV Demonstration

This repository contains a demonstration of CleverCSV, a Python package for robust handling of non-standard (*messy*) CSV files. It is based on the work **Wrangling Messy CSV Files by Detecting Row and Type Patterns** by Gertjan van den Burg, Alfredo Nazabal, and Charles Sutton.

– https://github.com/alan-turing-institute/CleverCSVDemo



No description, website, or topics provided.

GjjvdBurg add more examples and clarify — Latest commit [illegible] a day ago

data	add more examples and clarify	a day ago
images	add qr code with link to repo	12 days ago
CSV_dialect_detection_with_CleverCSV.ipy...	add more examples and clarify	a day ago
CSV_dialect_detection_with_CleverCSV.md	add more examples and clarify	a day ago
LICENSE	Add makefile and create the notebook from Markdown	7 days ago
Makefile	Add makefile and create the notebook from Markdown	7 days ago
README.md	Add binder thingy to Readme	13 days ago
requirements.txt	add termcolor dependency	6 days ago

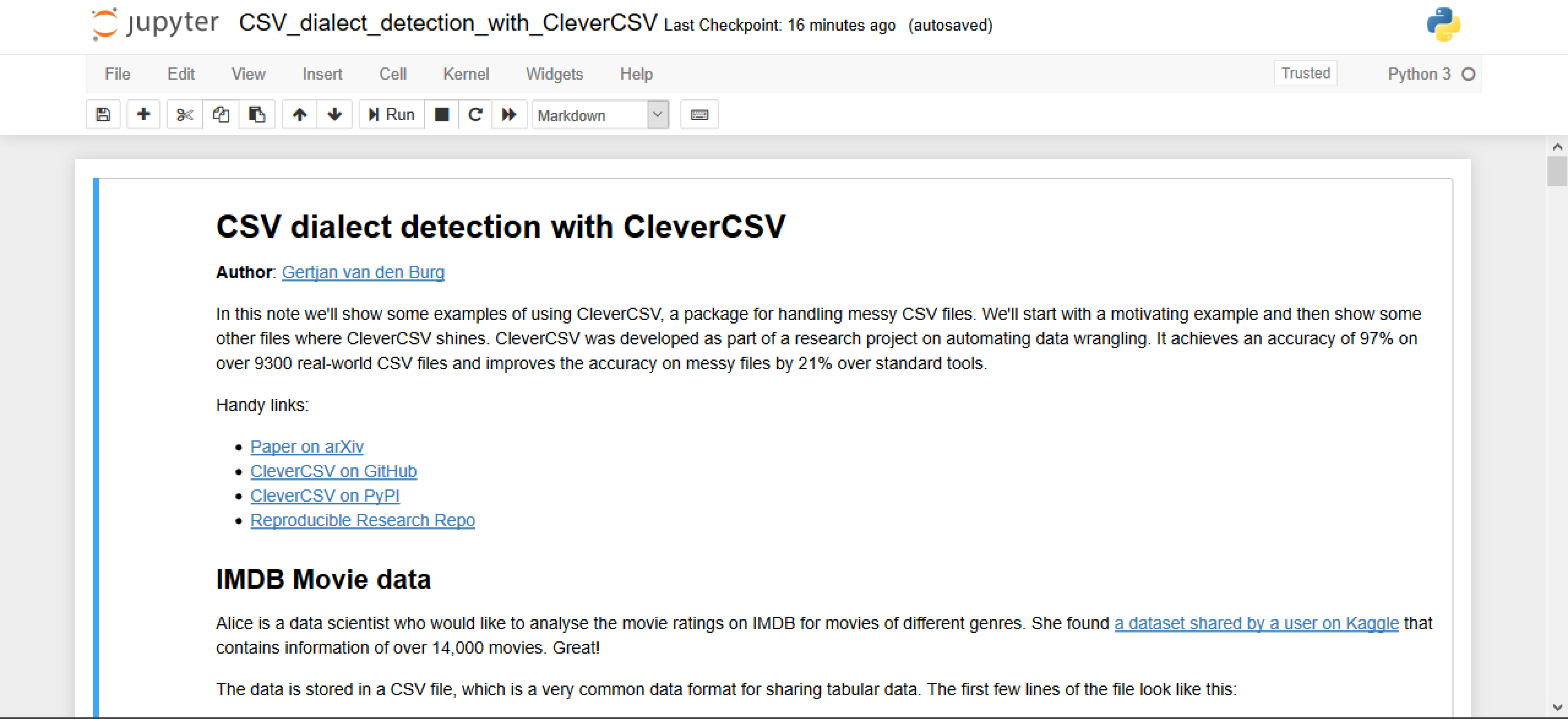
README.md

CleverCSV Demonstration

This repository [illegible] a demonstration of CleverCSV, a Python package for robust handling of non-standard (*messy*) CSV files. It is based [illegible] work **Wrangling Messy CSV Files by Detecting Row and Type Patterns** by Gertjan van den Burg, Alfredo Nazabal, and Charles Sutton.

- https://github.com/alan-turing-institute/CleverCSVDemo
- "Wrangling Messy CSV Files by Detecting Row and Type Patterns" arXiv:1811.11242

#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.2598529


jupyter CSV_dialect_detection_with_CleverCSV Last Checkpoint: 16 minutes ago (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Markdown

CSV dialect detection with CleverCSV

Author: Gertjan van den Burg

In this note we'll show some examples of using CleverCSV, a package for handling messy CSV files. We'll start with a motivating example and then show some other files where CleverCSV shines. CleverCSV was developed as part of a research project on automating data wrangling. It achieves an accuracy of 97% on over 9300 real-world CSV files and improves the accuracy on messy files by 21% over standard tools.

Handy links:

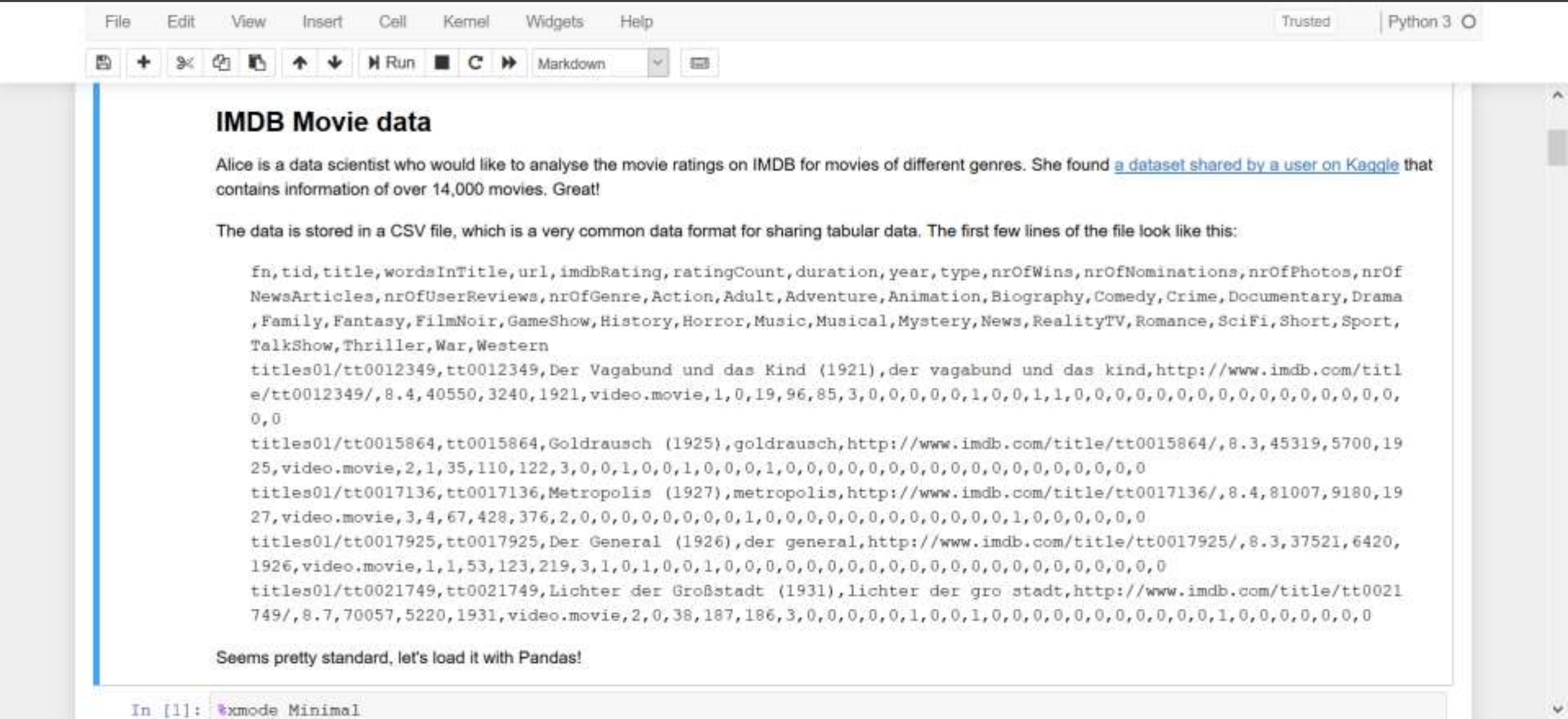
- Paper on arXiv
- CleverCSV on GitHub
- CleverCSV on PyPI
- Reproducible Research Repo

IMDB Movie data

Alice is a data scientist who would like to analyse the movie ratings on IMDB for movies of different genres. She found a dataset shared by a user on Kaggle that contains information of over 14,000 movies. Great!

The data is stored in a CSV file, which is a very common data format for sharing tabular data. The first few lines of the file look like this:

https://github.com/alan-turing-institute/CleverCSVDemo
#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.2598529



IMDB Movie data

Alice is a data scientist who would like to analyse the movie ratings on IMDB for movies of different genres. She found a dataset shared by a user on Kaggle that contains information of over 14,000 movies. Great!

The data is stored in a CSV file, which is a very common data format for sharing tabular data. The first few lines of the file look like this:

```
fn,tid,title,wordsInTitle,url,imdbRating,ratingCount,duration,year,type,nrOfWins,nrOfNominations,nrOfPhotos,nrOf
NewsArticles,nrOfUserReviews,nrOfGenre,Action,Adult,Adventure,Animation,Biography,Comedy,Crime,Documentary,Drama
,Family,Fantasy,FilmNoir,GameShow,History,Horror,Music,Musical,Mystery,News,RealityTV,Romance,SciFi,Short,Sport,
TalkShow,Thriller,War,Western
titles01/tt0012349,tt0012349,Der Vagabund und das Kind (1921),der vagabund und das kind,http://www.imdb.com/titl
e/tt0012349/,8.4,40550,3240,1921,video.movie,1,0,19,96,85,3,0,0,0,0,0,1,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0
titles01/tt0015864,tt0015864,Goldrausch (1925),goldrausch,http://www.imdb.com/title/tt0015864/,8.3,45319,5700,19
25,video.movie,2,1,35,110,122,3,0,0,1,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
titles01/tt0017136,tt0017136,Metropolis (1927),metropolis,http://www.imdb.com/title/tt0017136/,8.4,81007,9180,19
27,video.movie,3,4,67,428,376,2,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0
titles01/tt0017925,tt0017925,Der General (1926),der general,http://www.imdb.com/title/tt0017925/,8.3,37521,6420,
1926,video.movie,1,1,53,123,219,3,1,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
titles01/tt0021749,tt0021749,Lichter der Großstadt (1931),lichter der gro stadt,http://www.imdb.com/title/tt0021
749/,8.7,70057,5220,1931,video.movie,2,0,38,187,186,3,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0
```

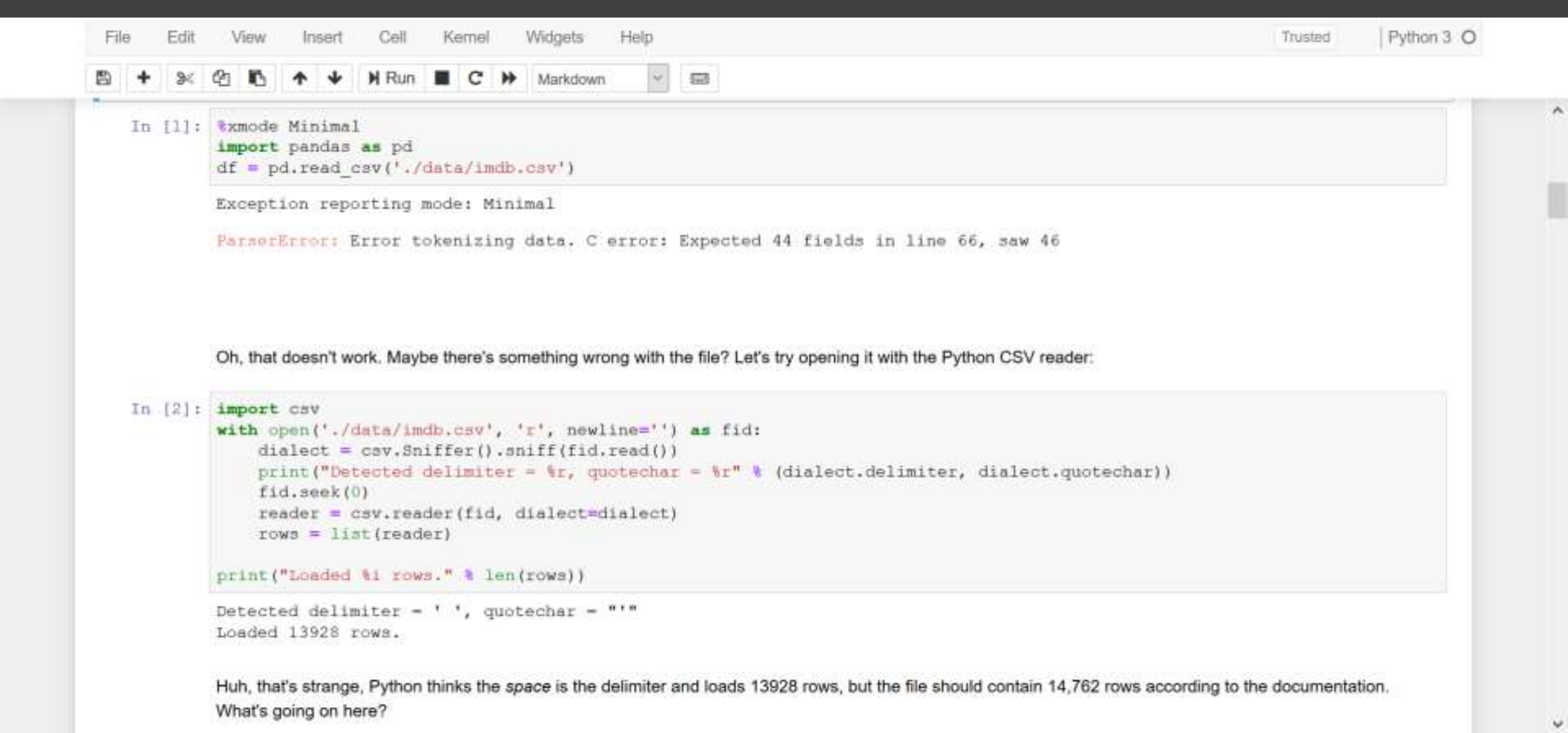
Seems pretty standard, let's load it with Pandas!

```
In [1]: %xmode Minimal
```

https://github.com/alan-turing-institute/CleverCSVDemo

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.2598529



In [1]:
```
%xmode Minimal
import pandas as pd
df = pd.read_csv('./data/imdb.csv')
```

```
Exception reporting mode: Minimal

ParserError: Error tokenizing data. C error: Expected 44 fields in line 66, saw 46
```

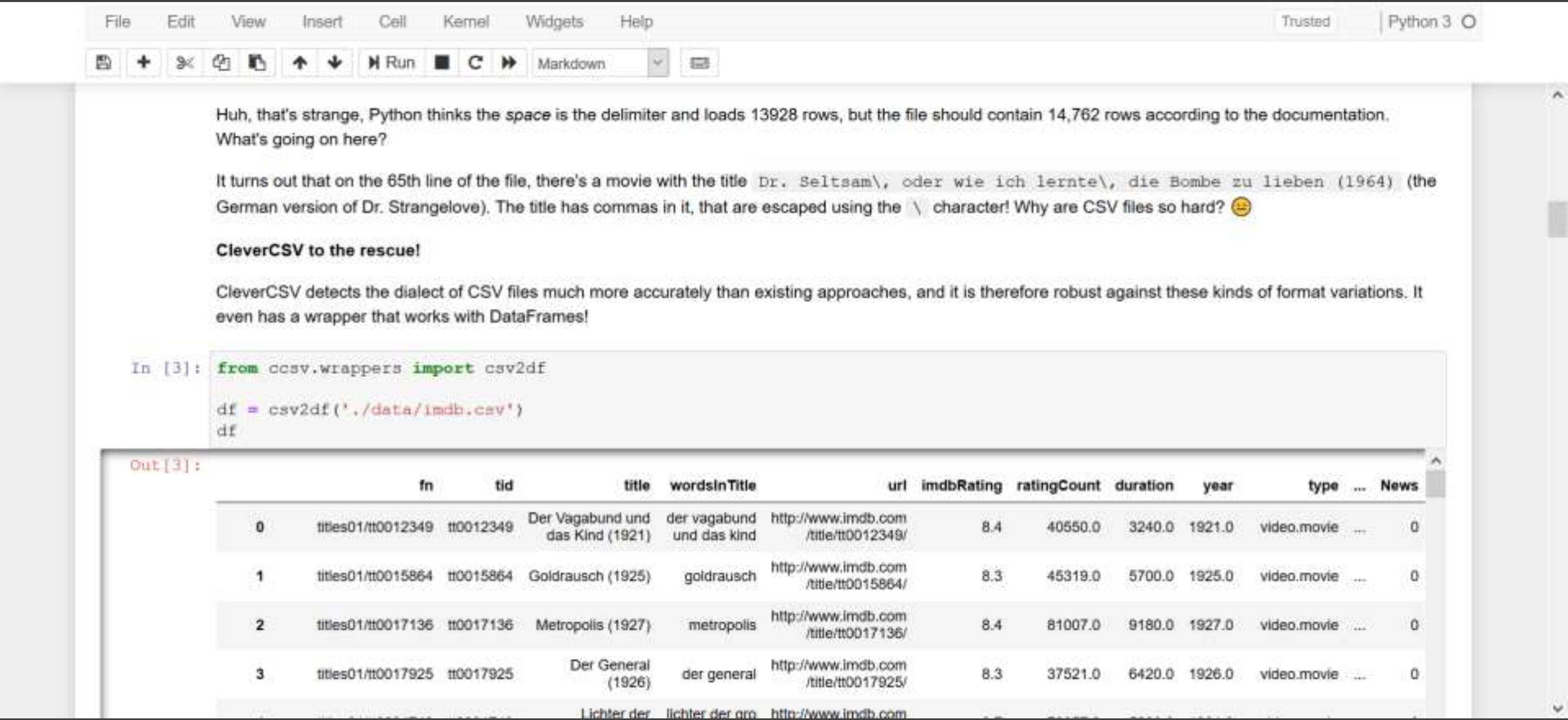
Oh, that doesn't work. Maybe there's something wrong with the file? Let's try opening it with the Python CSV reader:

In [2]:
```
import csv
with open('./data/imdb.csv', 'r', newline='') as fid:
    dialect = csv.Sniffer().sniff(fid.read())
    print("Detected delimiter = %r, quotechar = %r" % (dialect.delimiter, dialect.quotechar))
    fid.seek(0)
    reader = csv.reader(fid, dialect=dialect)
    rows = list(reader)

print("Loaded %i rows." % len(rows))
```

```
Detected delimiter = ' ', quotechar = "'"
Loaded 13928 rows.
```

Huh, that's strange, Python thinks the *space* is the delimiter and loads 13928 rows, but the file should contain 14,762 rows according to the documentation. What's going on here?

https://github.com/alan-turing-institute/CleverCSVDemo
#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.2598529



Huh, that's strange, Python thinks the *space* is the delimiter and loads 13928 rows, but the file should contain 14,762 rows according to the documentation. What's going on here?

It turns out that on the 65th line of the file, there's a movie with the title `Dr. Seltsam\, oder wie ich lernte\, die Bombe zu lieben (1964)` (the German version of Dr. Strangelove). The title has commas in it, that are escaped using the `\` character! Why are CSV files so hard? 😑

CleverCSV to the rescue!

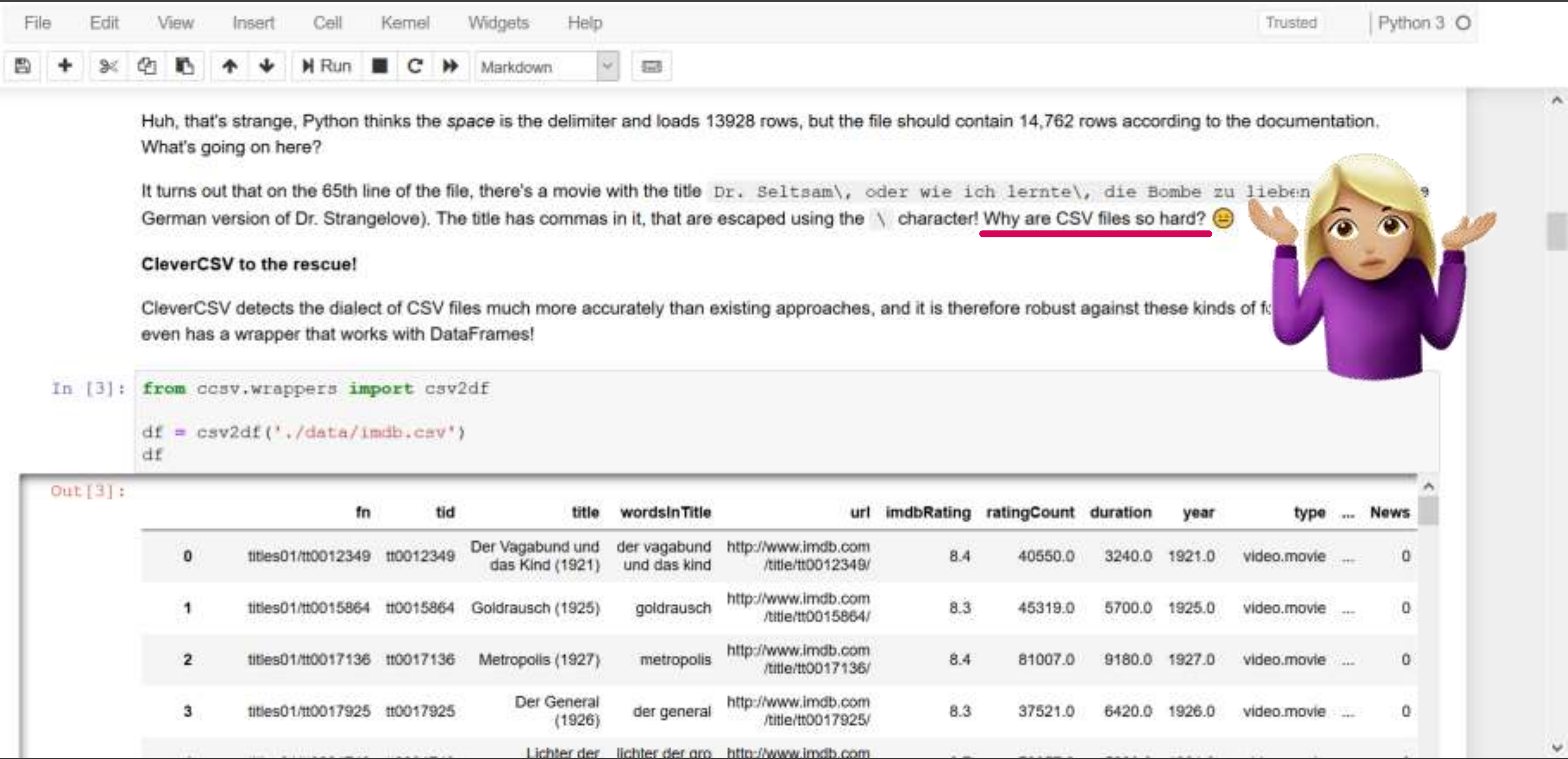
CleverCSV detects the dialect of CSV files much more accurately than existing approaches, and it is therefore robust against these kinds of format variations. It even has a wrapper that works with DataFrames!

```
In [3]: from ccsv.wrappers import csv2df

        df = csv2df('./data/imdb.csv')
        df
```

Out[3]:

	fn	tid	title	wordsInTitle	url	imdbRating	ratingCount	duration	year	type	...	News
0	titles01/tt0012349	tt0012349	Der Vagabund und das Kind (1921)	der vagabund und das kind	http://www.imdb.com /title/tt0012349/	8.4	40550.0	3240.0	1921.0	video.movie	...	0
1	titles01/tt0015864	tt0015864	Goldrausch (1925)	goldrausch	http://www.imdb.com /title/tt0015864/	8.3	45319.0	5700.0	1925.0	video.movie	...	0
2	titles01/tt0017136	tt0017136	Metropolis (1927)	metropolis	http://www.imdb.com /title/tt0017136/	8.4	81007.0	9180.0	1927.0	video.movie	...	0
3	titles01/tt0017925	tt0017925	Der General (1926)	der general	http://www.imdb.com /title/tt0017925/	8.3	37521.0	6420.0	1926.0	video.movie	...	0
			Lichter der	lichter der gro	http://www.imdb.com							

https://github.com/alan-turing-institute/CleverCSVDemo
#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.2598529



Huh, that's strange, Python thinks the *space* is the delimiter and loads 13928 rows, but the file should contain 14,762 rows according to the documentation. What's going on here?

It turns out that on the 65th line of the file, there's a movie with the title `Dr. Seltsam\, oder wie ich lernte\, die Bombe zu lieben` German version of Dr. Strangelove). The title has commas in it, that are escaped using the `\` character! Why are CSV files so hard? 😑

CleverCSV to the rescue!

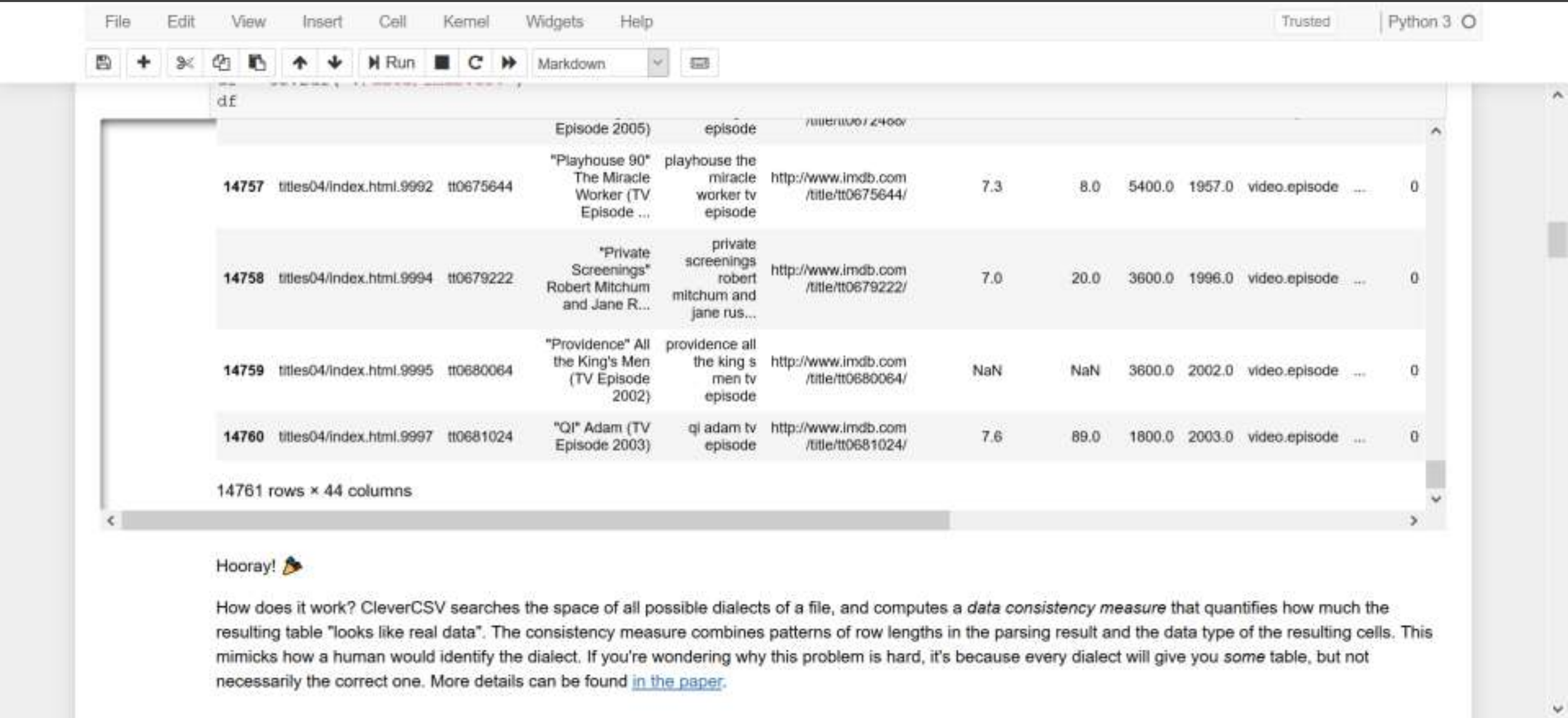
CleverCSV detects the dialect of CSV files much more accurately than existing approaches, and it is therefore robust against these kinds of f even has a wrapper that works with DataFrames!

```
In [3]: from ccsv.wrappers import csv2df

        df = csv2df('./data/imdb.csv')
        df
```

Out[3]:

	fn	tid	title	wordsInTitle	url	imdbRating	ratingCount	duration	year	type	...	News
0	titles01/tt0012349	tt0012349	Der Vagabund und das Kind (1921)	der vagabund und das kind	http://www.imdb.com/title/tt0012349/	8.4	40550.0	3240.0	1921.0	video.movie	...	0
1	titles01/tt0015864	tt0015864	Goldrausch (1925)	goldrausch	http://www.imdb.com/title/tt0015864/	8.3	45319.0	5700.0	1925.0	video.movie	...	0
2	titles01/tt0017136	tt0017136	Metropolis (1927)	metropolis	http://www.imdb.com/title/tt0017136/	8.4	81007.0	9180.0	1927.0	video.movie	...	0
3	titles01/tt0017925	tt0017925	Der General (1926)	der general	http://www.imdb.com/title/tt0017925/	8.3	37521.0	6420.0	1926.0	video.movie	...	0
			Lichter der	lichter der gro	http://www.imdb.com							

https://github.com/alan-turing-institute/CleverCSVDemo
#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.2598529



```
df
```

			Episode 2005)	episode	/title/tt0672466/							
14757	titles04/index.html.9992	tt0675644	"Playhouse 90" The Miracle Worker (TV Episode ...	playhouse the miracle worker tv episode	http://www.imdb.com /title/tt0675644/	7.3	8.0	5400.0	1957.0	video.episode	...	0
14758	titles04/index.html.9994	tt0679222	"Private Screenings" Robert Mitchum and Jane R...	private screenings robert mitchum and jane rus...	http://www.imdb.com /title/tt0679222/	7.0	20.0	3600.0	1996.0	video.episode	...	0
14759	titles04/index.html.9995	tt0680064	"Providence" All the King's Men (TV Episode 2002)	providence all the king s men tv episode	http://www.imdb.com /title/tt0680064/	NaN	NaN	3600.0	2002.0	video.episode	...	0
14760	titles04/index.html.9997	tt0681024	"QI" Adam (TV Episode 2003)	qi adam tv episode	http://www.imdb.com /title/tt0681024/	7.6	89.0	1800.0	2003.0	video.episode	...	0

14761 rows × 44 columns

Hooray! 🎉

How does it work? CleverCSV searches the space of all possible dialects of a file, and computes a *data consistency measure* that quantifies how much the resulting table "looks like real data". The consistency measure combines patterns of row lengths in the parsing result and the data type of the resulting cells. This mimicks how a human would identify the dialect. If you're wondering why this problem is hard, it's because every dialect will give you *some* table, but not necessarily the correct one. More details can be found in the paper.

https://github.com/alan-turing-institute/CleverCSVDemo
#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.2598529

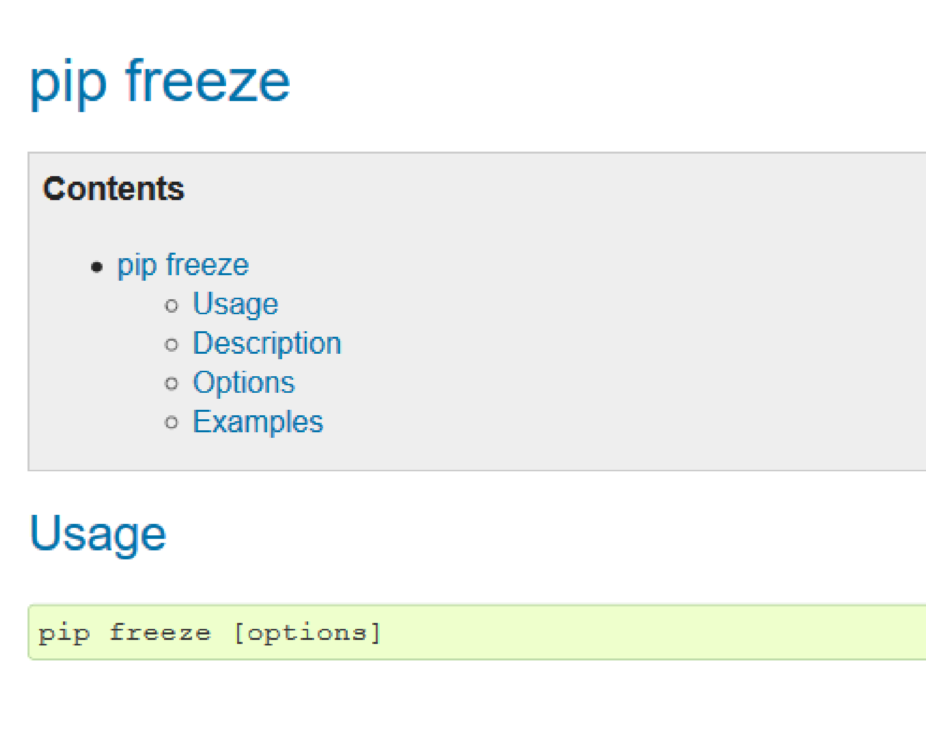
Defining your computational environment

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.2598529

pip freeze

- Python Package Index (PyPI)
- Pip is a recursive acronym that can stand for either "Pip Installs Packages" or "Pip Installs Python"



pip freeze

Contents

- pip freeze
  - Usage
  - Description
  - Options
  - Examples

Usage

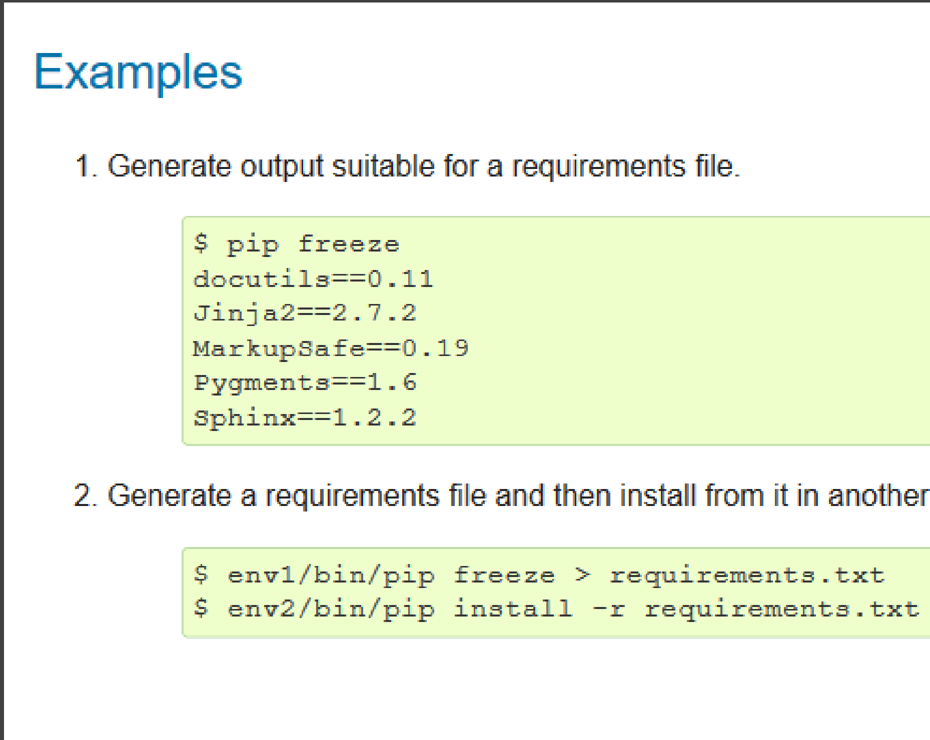
```
pip freeze [options]
```

https://pip.pypa.io/en/stable/reference/pip_freeze

#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.2598529

pip freeze

- `pip freeze` captures the versions of all packages that you’re currently using
- Can print to screen or save in a file called `requirements.txt`



Examples

1. Generate output suitable for a requirements file.

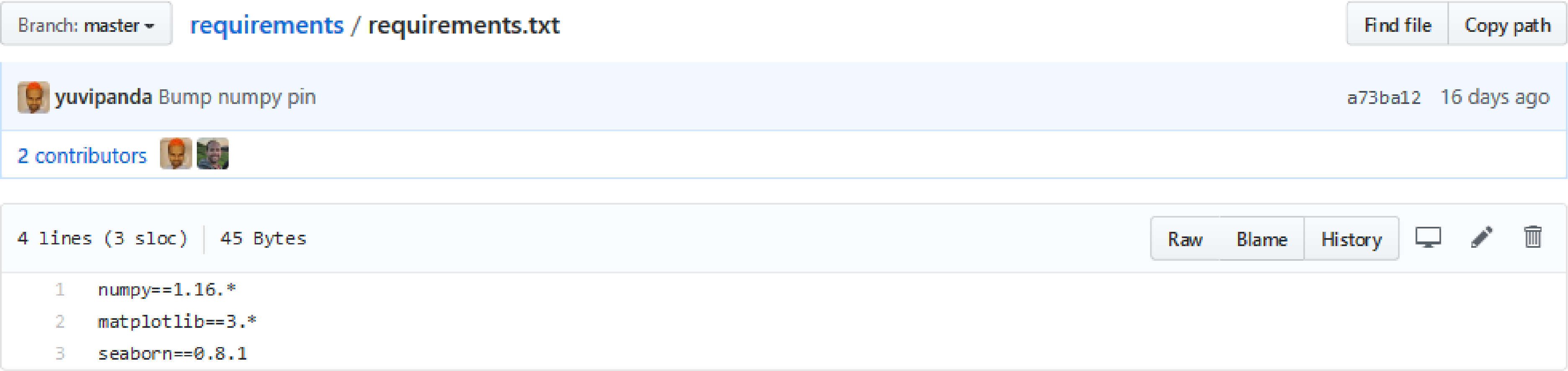
```
$ pip freeze
docutils==0.11
Jinja2==2.7.2
MarkupSafe==0.19
Pygments==1.6
Sphinx==1.2.2
```

2. Generate a requirements file and then install from it in another

```
$ env1/bin/pip freeze > requirements.txt
$ env2/bin/pip install -r requirements.txt
```

https://pip.pypa.io/en/stable/reference/pip_freeze

Binder example: requirements.txt



```
numpy==1.16.*
matplotlib==3.*
seaborn==0.8.1
```

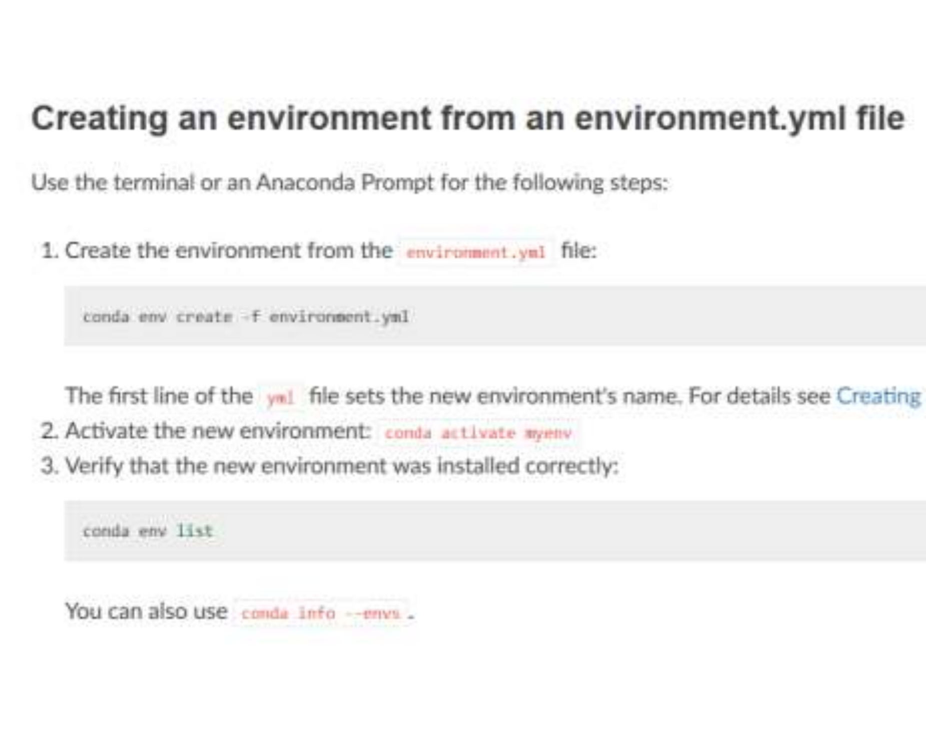
#TuringWay @kirstie_j @mybinderteam

https://github.com/binder-examples/requirements

https://doi.org/10.5281/zenodo.2598529

Conda env create

- Package manager for multiple languages
- Information about installed software saved in file called `environment.yml`



Creating an environment from an environment.yml file

Use the terminal or an Anaconda Prompt for the following steps:

1. Create the environment from the `environment.yml` file:

```
conda env create -f environment.yml
```

   The first line of the `yml` file sets the new environment's name. For details see Creating

2. Activate the new environment: `conda activate myenv`
3. Verify that the new environment was installed correctly:

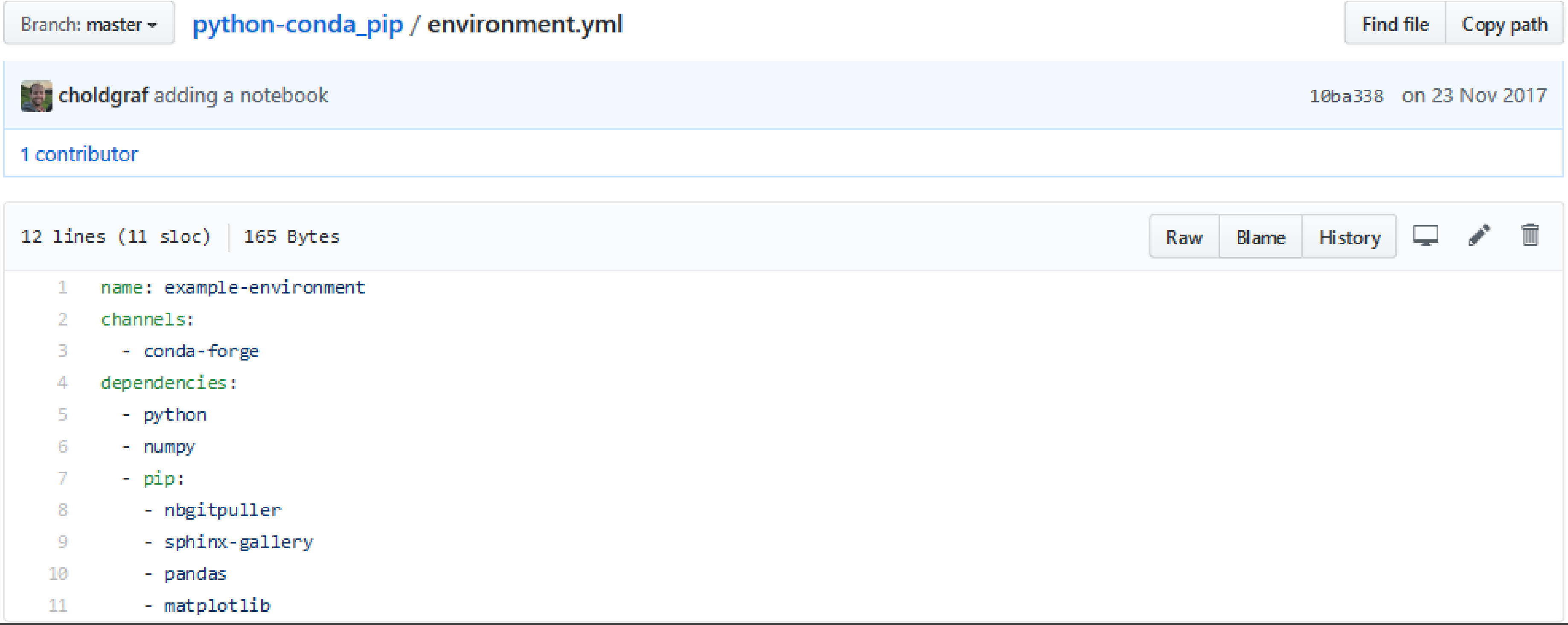
```
conda env list
```

   You can also use `conda info --envs`.

https://docs.conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html
https://the-turing-way.netlify.com/reproducible_environments/02/package-management.html

#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.2598529

Binder example: environment.yml



python-conda_pip / environment.yml

choldgraf adding a notebook 10ba338 on 23 Nov 2017

1 contributor

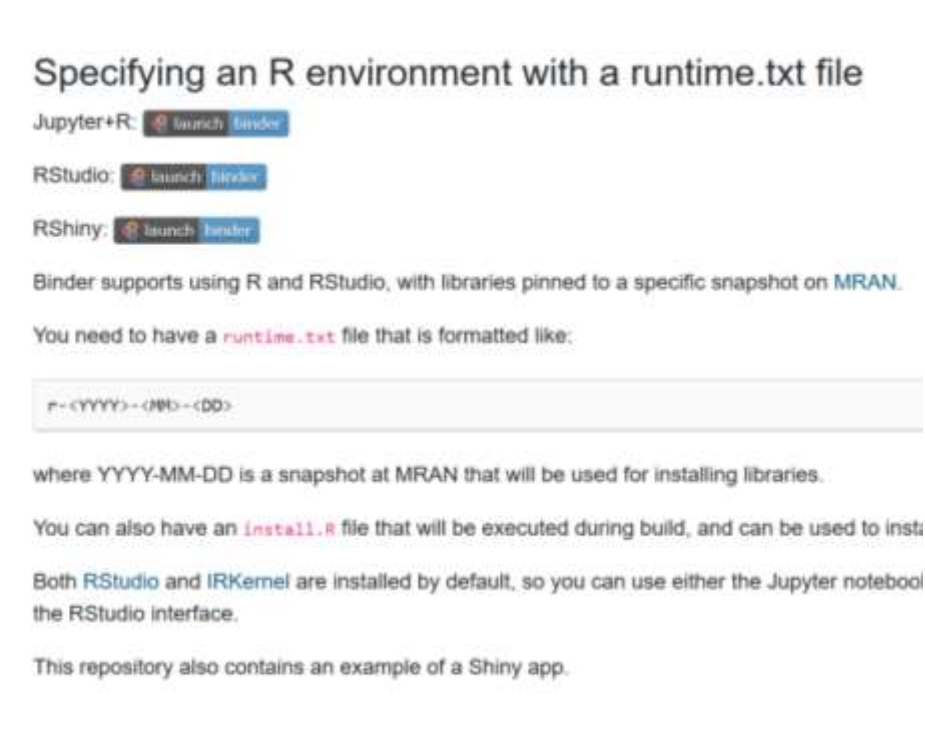
12 lines (11 sloc) | 165 Bytes

```yaml
name: example-environment
channels:
  - conda-forge
dependencies:
  - python
  - numpy
  - pip:
    - nbgitpuller
    - sphinx-gallery
    - pandas
    - matplotlib
```

https://github.com/binder-examples/python-conda_pip

install.R

- Binder interface includes Rstudio
- `install.R` contains required packages
- `runtime.txt` sets the version on MRAN



Specifying an R environment with a runtime.txt file

Jupyter+R: launch binder

RStudio: launch binder

RShiny: launch binder

Binder supports using R and RStudio, with libraries pinned to a specific snapshot on MRAN.

You need to have a `runtime.txt` file that is formatted like:

```
r-<YYYY>-<MM>-<DD>
```

where YYYY-MM-DD is a snapshot at MRAN that will be used for installing libraries.

You can also have an `install.R` file that will be executed during build, and can be used to insta

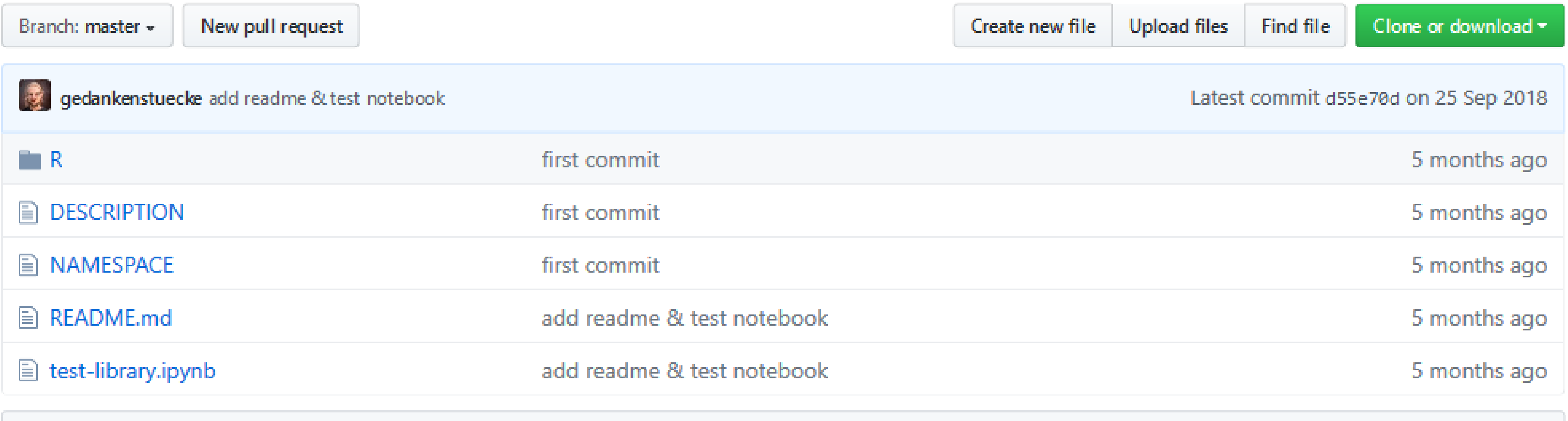
Both RStudio and IRKernel are installed by default, so you can use either the Jupyter noteboo the RStudio interface.

This repository also contains an example of a Shiny app.

https://mybinder.readthedocs.io/en/latest/sample_repos.html#specifying-an-r-environment-with-a-runtime-txt-file

#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.2598529

Binder example: install.R



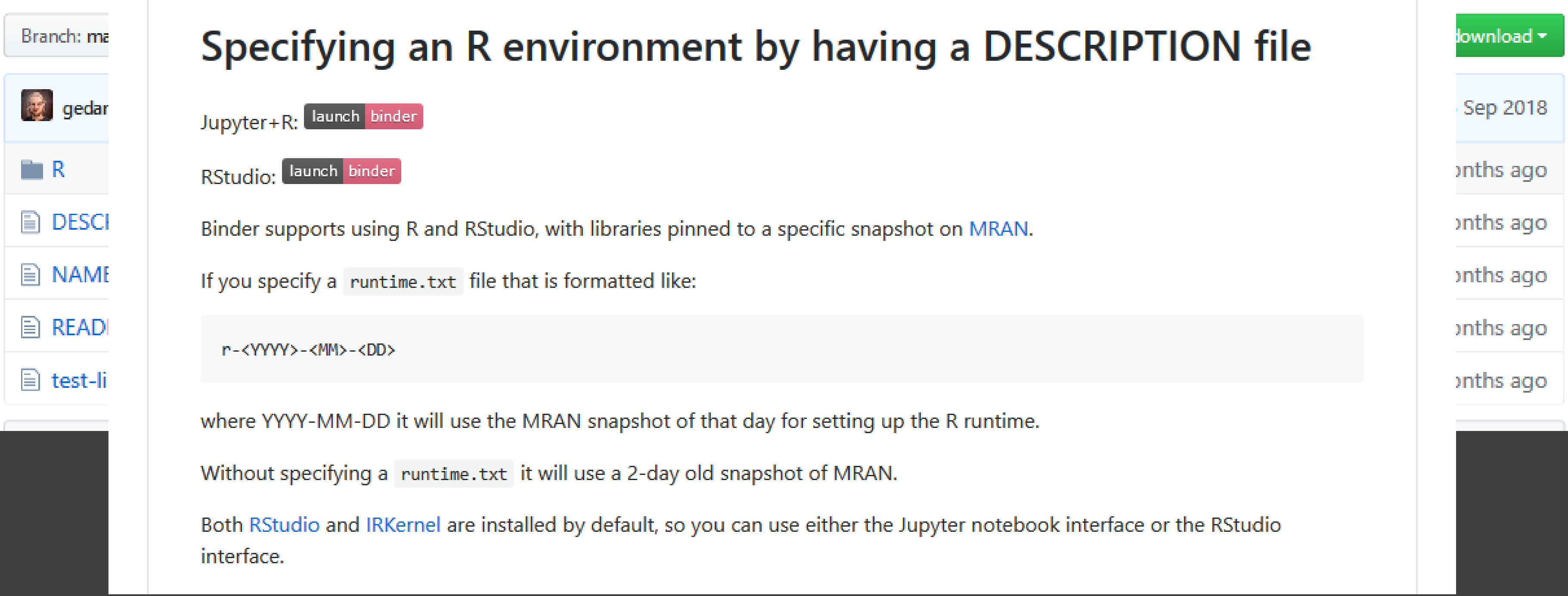
gedankenstuecke add readme & test notebook Latest commit d55e70d on 25 Sep 2018

R	first commit	5 months ago
DESCRIPTION	first commit	5 months ago
NAMESPACE	first commit	5 months ago
README.md	add readme & test notebook	5 months ago
test-library.ipynb	add readme & test notebook	5 months ago

https://github.com/binder-examples/binder-r-description

#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.2598529

Binder example: install.R



Specifying an R environment by having a DESCRIPTION file

Jupyter+R: launch binder

RStudio: launch binder

Binder supports using R and RStudio, with libraries pinned to a specific snapshot on MRAN.

If you specify a `runtime.txt` file that is formatted like:

```
r-<YYYY>-<MM>-<DD>
```

where YYYY-MM-DD it will use the MRAN snapshot of that day for setting up the R runtime.

Without specifying a `runtime.txt` it will use a 2-day old snapshot of MRAN.

Both RStudio and IRKernel are installed by default, so you can use either the Jupyter notebook interface or the RStudio interface.

https://github.com/binder-examples/binder-r-description

Becky Arnold

“There are a lot of things you need to know before you can jump into continuous integration.

Version control is a prerequisite for pretty much everything.”

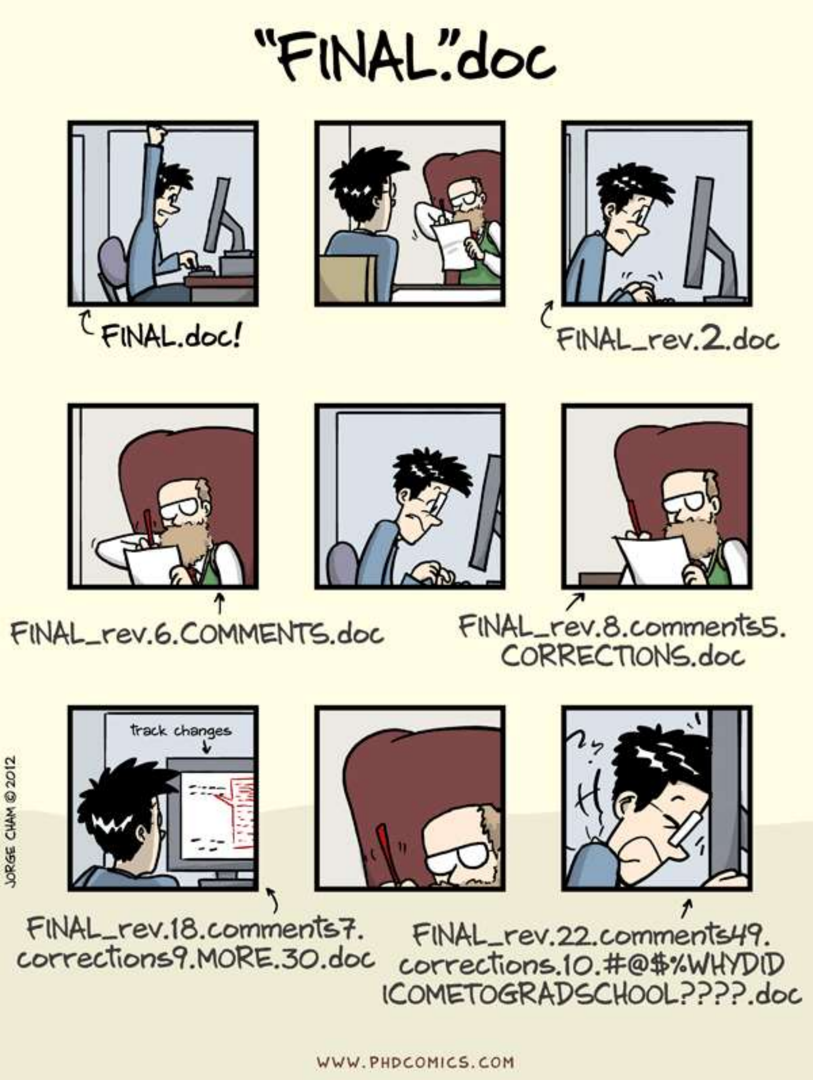
https://software.ac.uk/about/fellows/becky-arnold

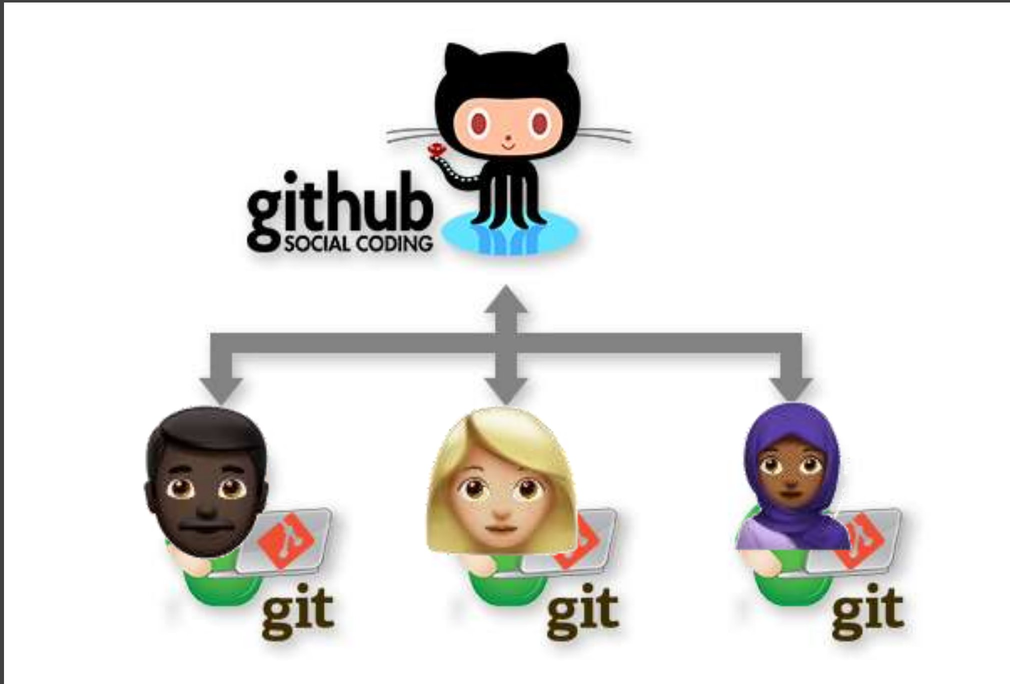
#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.2598529

Distributed and remote version control

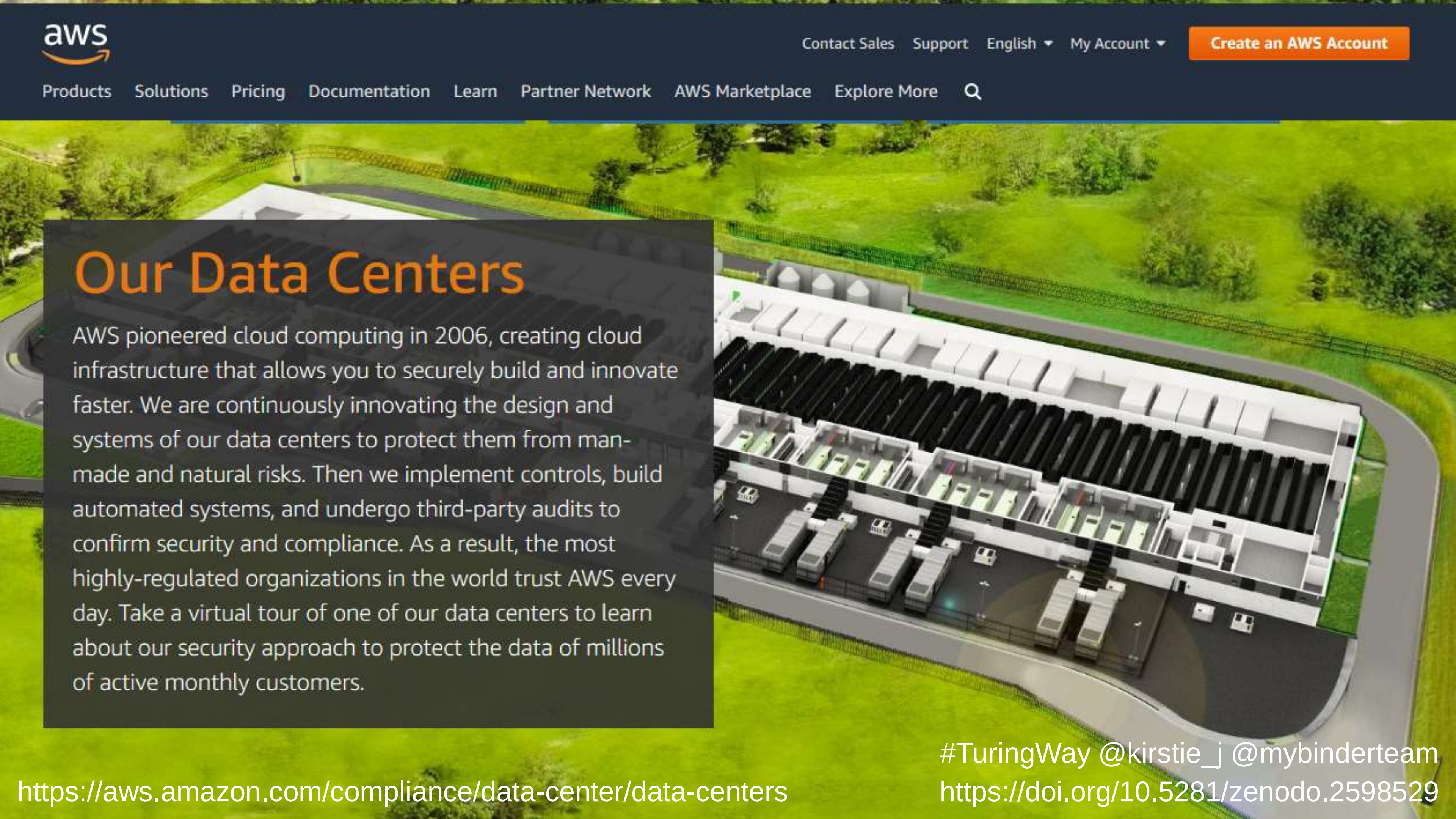
#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.2598529





http://phdcomics.com/comics/archive.php?comicid=1531
#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.2598529

The cloud is just someone else’s computer



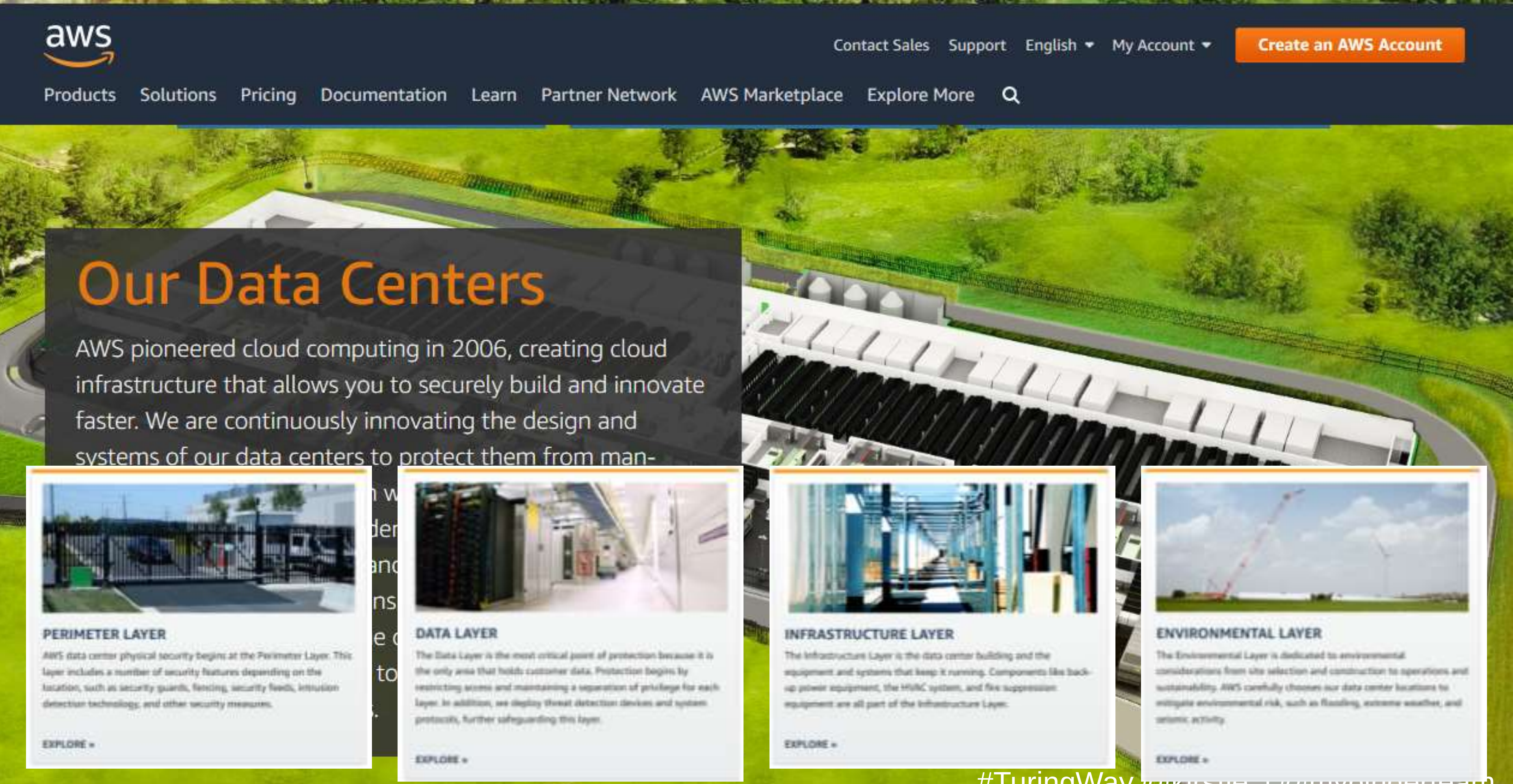
Our Data Centers

AWS pioneered cloud computing in 2006, creating cloud infrastructure that allows you to securely build and innovate faster. We are continuously innovating the design and systems of our data centers to protect them from man-made and natural risks. Then we implement controls, build automated systems, and undergo third-party audits to confirm security and compliance. As a result, the most highly-regulated organizations in the world trust AWS every day. Take a virtual tour of one of our data centers to learn about our security approach to protect the data of millions of active monthly customers.

#TuringWay @kirstie_j @mybinderteam

https://aws.amazon.com/compliance/data-center/data-centers

https://doi.org/10.5281/zenodo.2598529



Our Data Centers

AWS pioneered cloud computing in 2006, creating cloud infrastructure that allows you to securely build and innovate faster. We are continuously innovating the design and systems of our data centers to protect them from man-

PERIMETER LAYER

AWS data center physical security begins at the Perimeter Layer. This layer includes a number of security features depending on the location, such as security guards, fencing, security feeds, intrusion detection technology, and other security measures.

EXPLORE »

DATA LAYER

The Data Layer is the most critical point of protection because it is the only area that holds customer data. Protection begins by restricting access and maintaining a separation of privilege for each layer. In addition, we deploy threat detection devices and system protocols, further safeguarding this layer.

EXPLORE »

INFRASTRUCTURE LAYER

The Infrastructure Layer is the data center building and the equipment and systems that keep it running. Components like back-up power equipment, the HVAC system, and fire suppression equipment are all part of the Infrastructure Layer.

EXPLORE »

ENVIRONMENTAL LAYER

The Environmental Layer is dedicated to environmental considerations from site selection and construction to operations and sustainability. AWS carefully chooses our data center locations to mitigate environmental risk, such as flooding, extreme weather, and seismic activity.

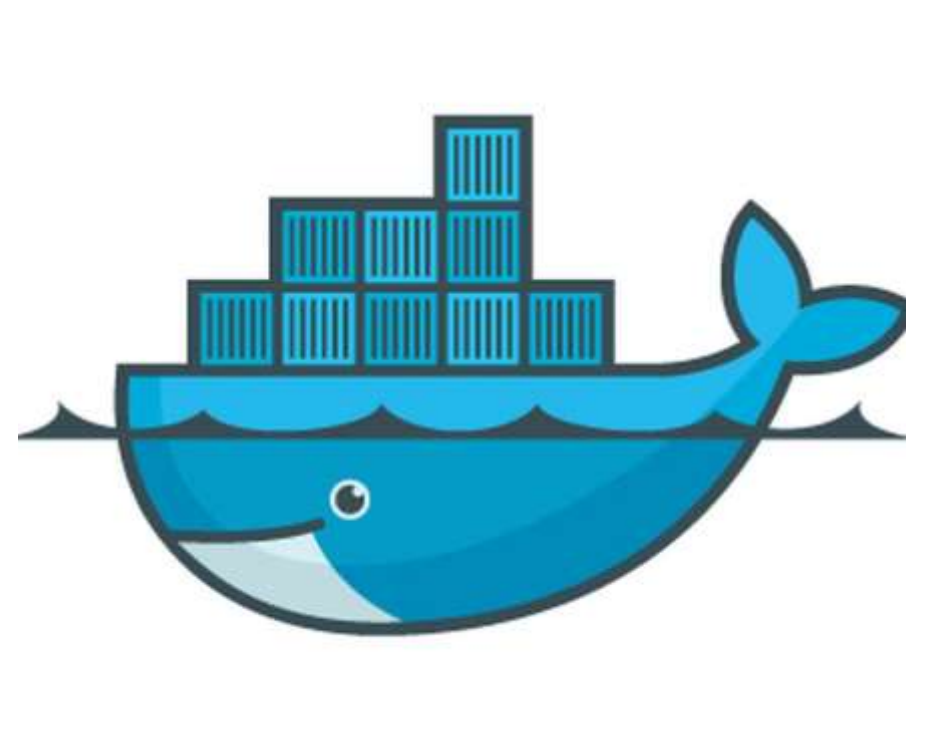
EXPLORE »

https://aws.amazon.com/compliance/data-center/data-centers

Containers

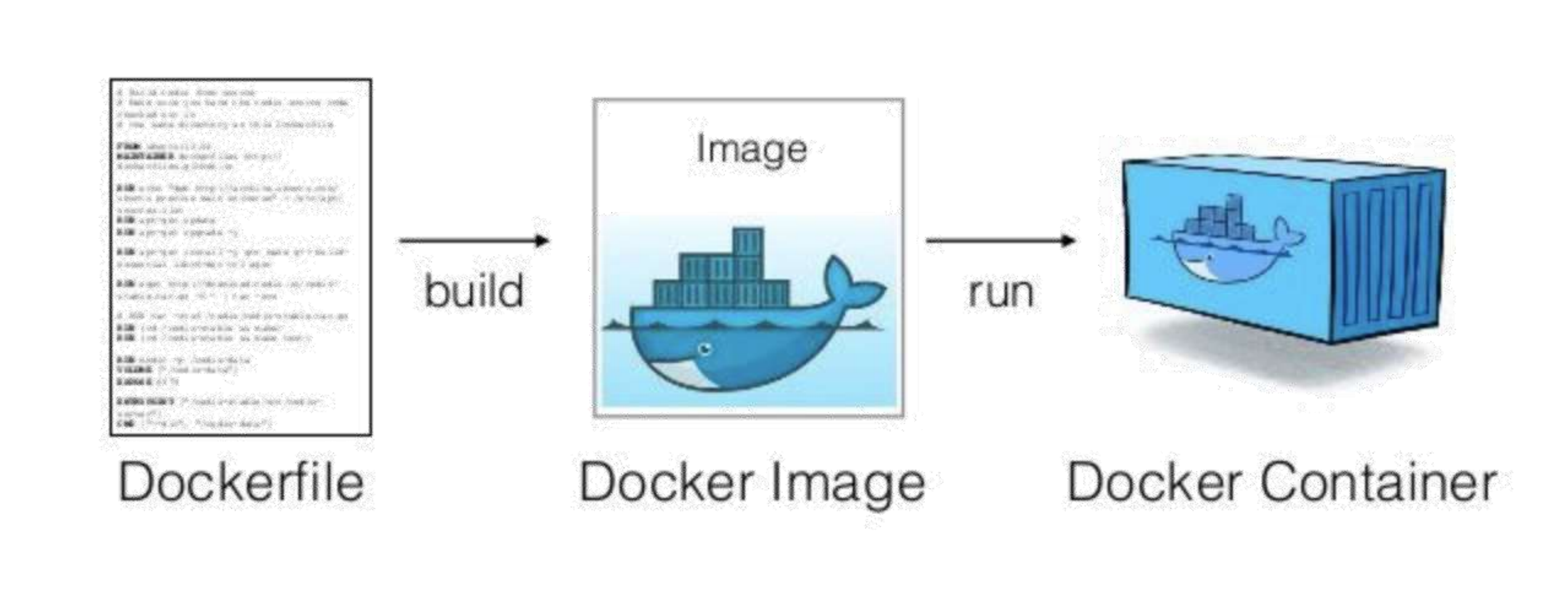
Docker

– A container that bundles all the infrastructure and software together.

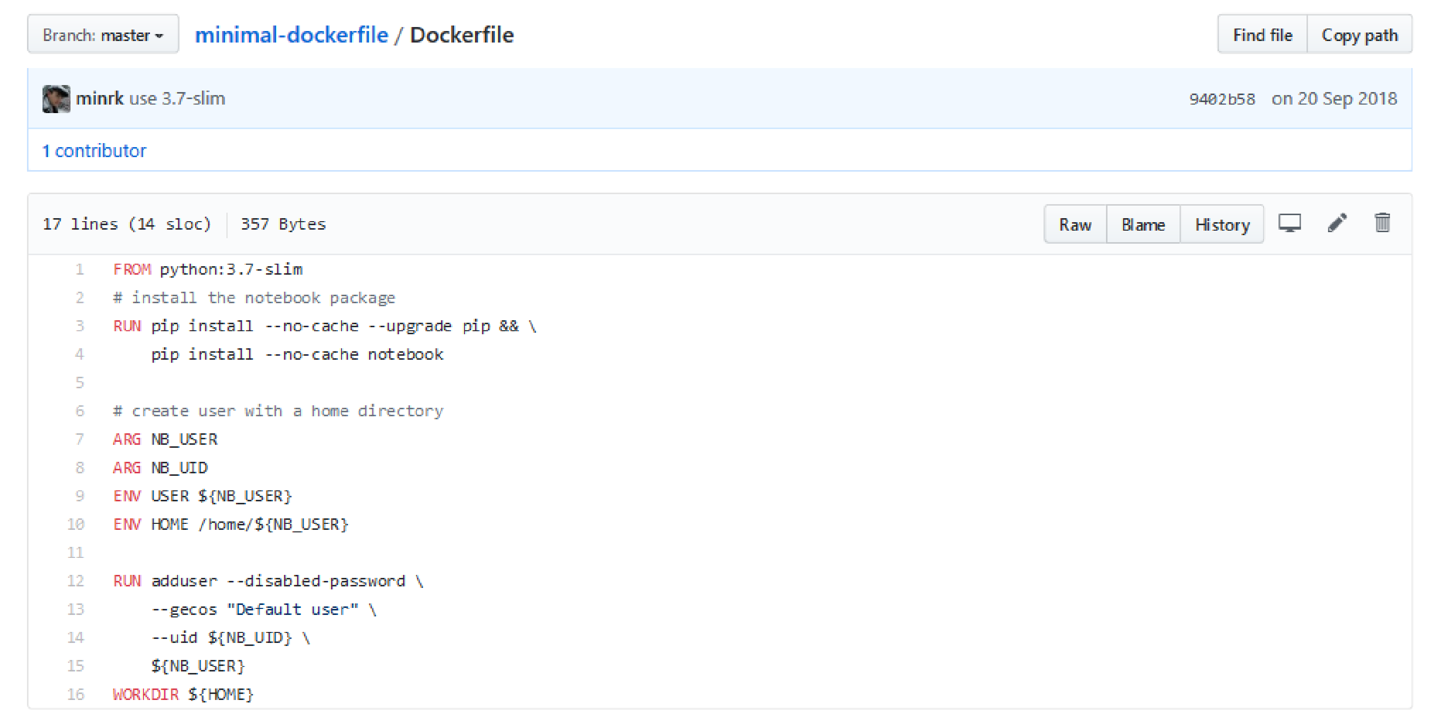


https://mybinder.readthedocs.io/en/latest/sample_repos.html#specifying-an-r-environment-with-a-runtime-txt-file

Human and machine readable files



Binder example: docker



minrk use 3.7-slim 9402b58 on 20 Sep 2018

1 contributor

17 lines (14 sloc) | 357 Bytes

```
FROM python:3.7-slim
# install the notebook package
RUN pip install --no-cache --upgrade pip && \
    pip install --no-cache notebook

# create user with a home directory
ARG NB_USER
ARG NB_UID
ENV USER ${NB_USER}
ENV HOME /home/${NB_USER}

RUN adduser --disabled-password \
    --gecos "Default user" \
    --uid ${NB_UID} \
    ${NB_USER}
WORKDIR ${HOME}
```

https://github.com/binder-examples/binder-r-description

Small group exercise

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.2598529

Small group exercise

Get into groups of 2-3 and explore these examples

- Are there differences between different branches?
- Does that give different results?
- Did you get what you'd expect?

https://github.com/alan-turing-institute/
the-turing-way/blob/master/
workshops/
boost-research-reproducibility-binder/
paired_examples.md

Zero to Binder

Please open these instructions in your browser

https://github.com/alan-turing-institute/

the-turing-way/blob/master/

workshops/

boost-research-reproducibility-binder/

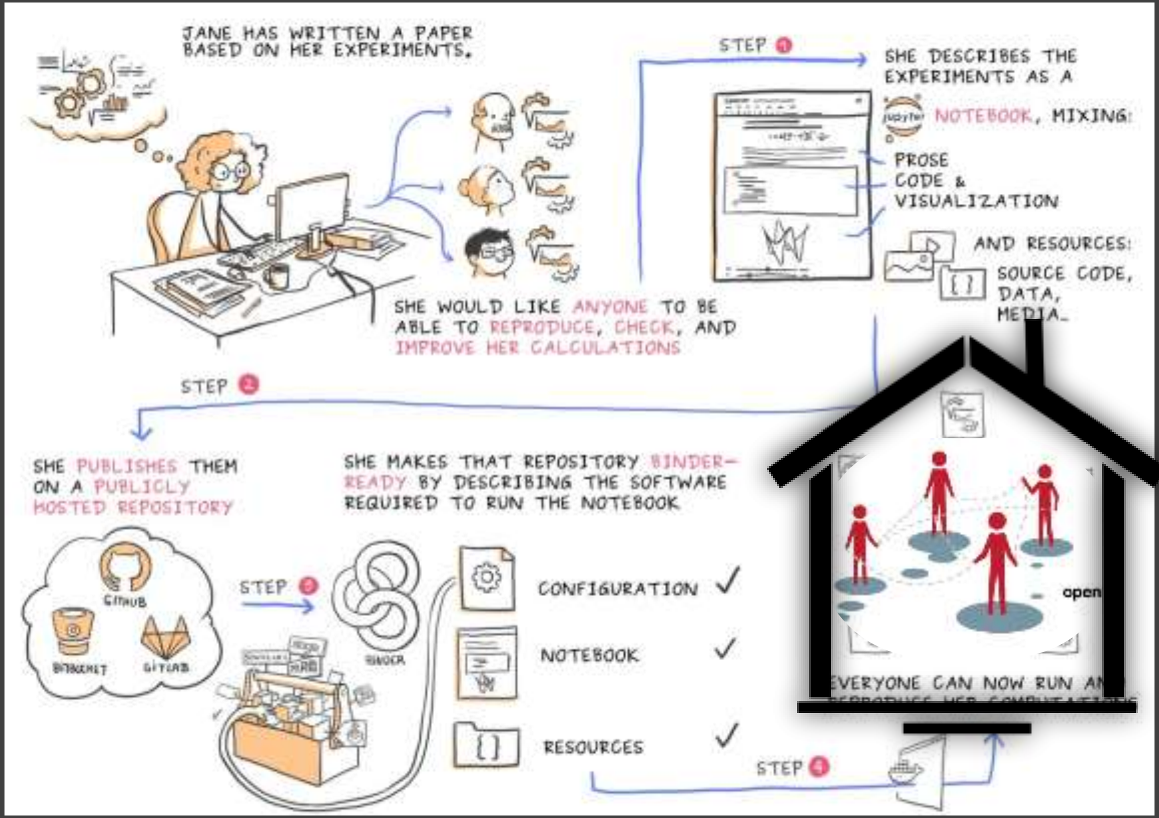
workshop-presentations/

zero-to-binder.md

Agenda

Boost your reproducibility with Binder

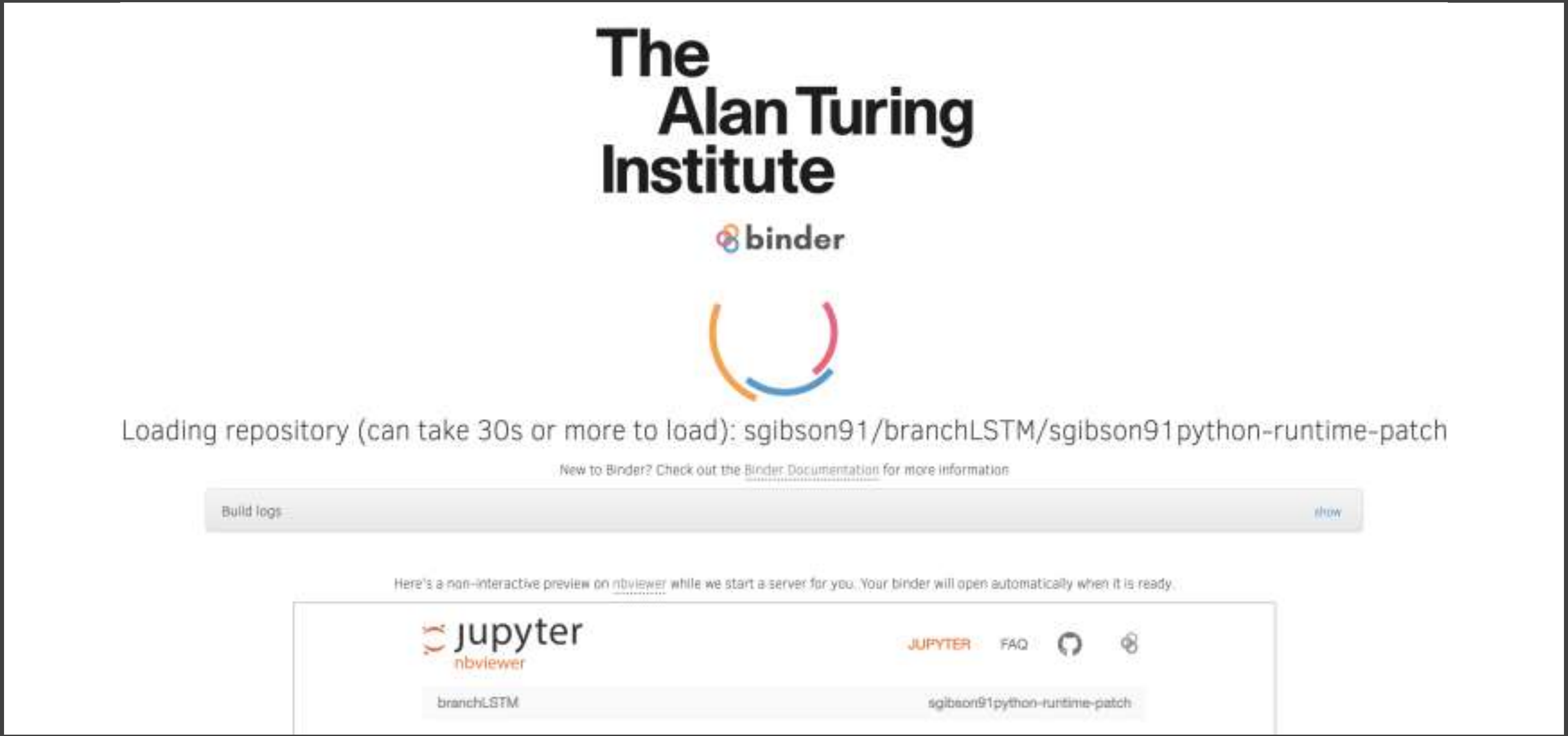
- 13:00 Registration and introductions
- 13:10 Introduction to the workshop and The Turing Way
- 13:20 Presentation: Why you need a reproducible computing environment and how Binder can help
- 14:30 Coffee break
- 15:00 Code along demo: Zero to Binder, build a Binder resource
- 16:00 Build your own Binder
- 16:30 Feedback, group picture and close


JANE HAS WRITTEN A PAPER BASED ON HER EXPERIMENTS.
SHE WOULD LIKE ANYONE TO BE ABLE TO REPRODUCE, CHECK, AND IMPROVE HER CALCULATIONS
STEP 1
SHE DESCRIBES THE EXPERIMENTS AS A
jupyter NOTEBOOK, MIXING:
PROSE
CODE &
VISUALIZATION
AND RESOURCES:
SOURCE CODE,
DATA,
MEDIA...
STEP 2
SHE PUBLISHES THEM ON A PUBLICLY HOSTED REPOSITORY
GITHUB
BITBUCKET
GITLAB
STEP 3
SHE MAKES THAT REPOSITORY BINDER-READY BY DESCRIBING THE SOFTWARE REQUIRED TO RUN THE NOTEBOOK
BINDER
CONFIGURATION ✓
NOTEBOOK ✓
RESOURCES ✓
STEP 4
EVERYONE CAN NOW RUN AN
open

Courtesy of Juliette Taka: https://twitter.com/mybinderteam/status/1082556317842264064

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2598529



The Alan Turing Institute

binder

Loading repository (can take 30s or more to load): sgibson91/branchLSTM/sgibson91python-runtime-patch

New to Binder? Check out the Binder Documentation for more information

Here's a non-interactive preview on nbviewer while we start a server for you. Your binder will open automatically when it is ready.

https://github.com/kochkinaelena/branchLSTM (on Turing Way Hub)

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2598529

Champion: Elena Kochkina

Turing at SemEval-2017 Task 8: Sequential Approach to Rumour Stance Classification with Branch-LSTM

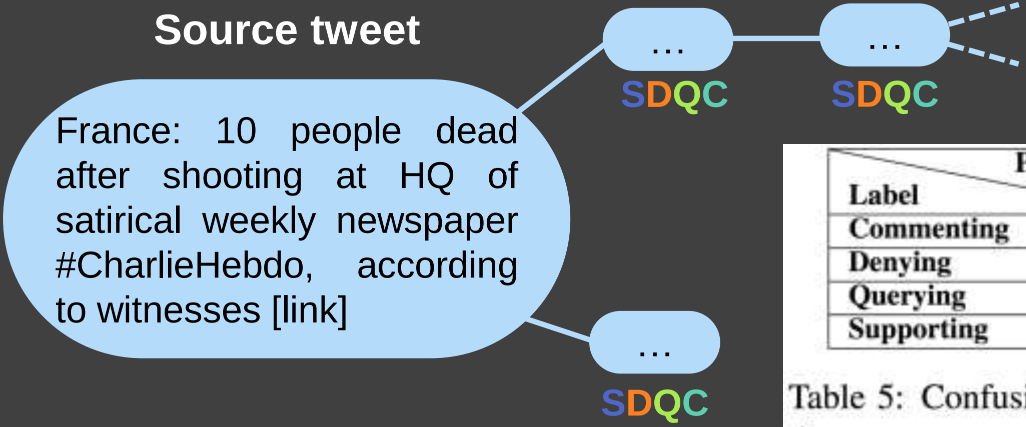
Elena Kochkina, Maria Liakata, Isabelle Augenstein

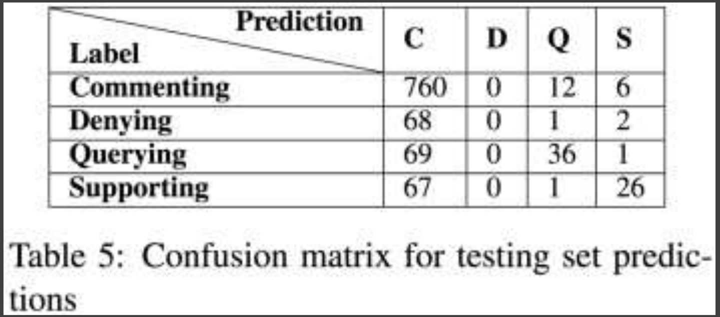
Support

Deny

Query

Comment


Source tweet
France: 10 people dead after shooting at HQ of satirical weekly newspaper #CharlieHebdo, according to witnesses [link]
…
SDQC
…
SDQC
…
SDQC

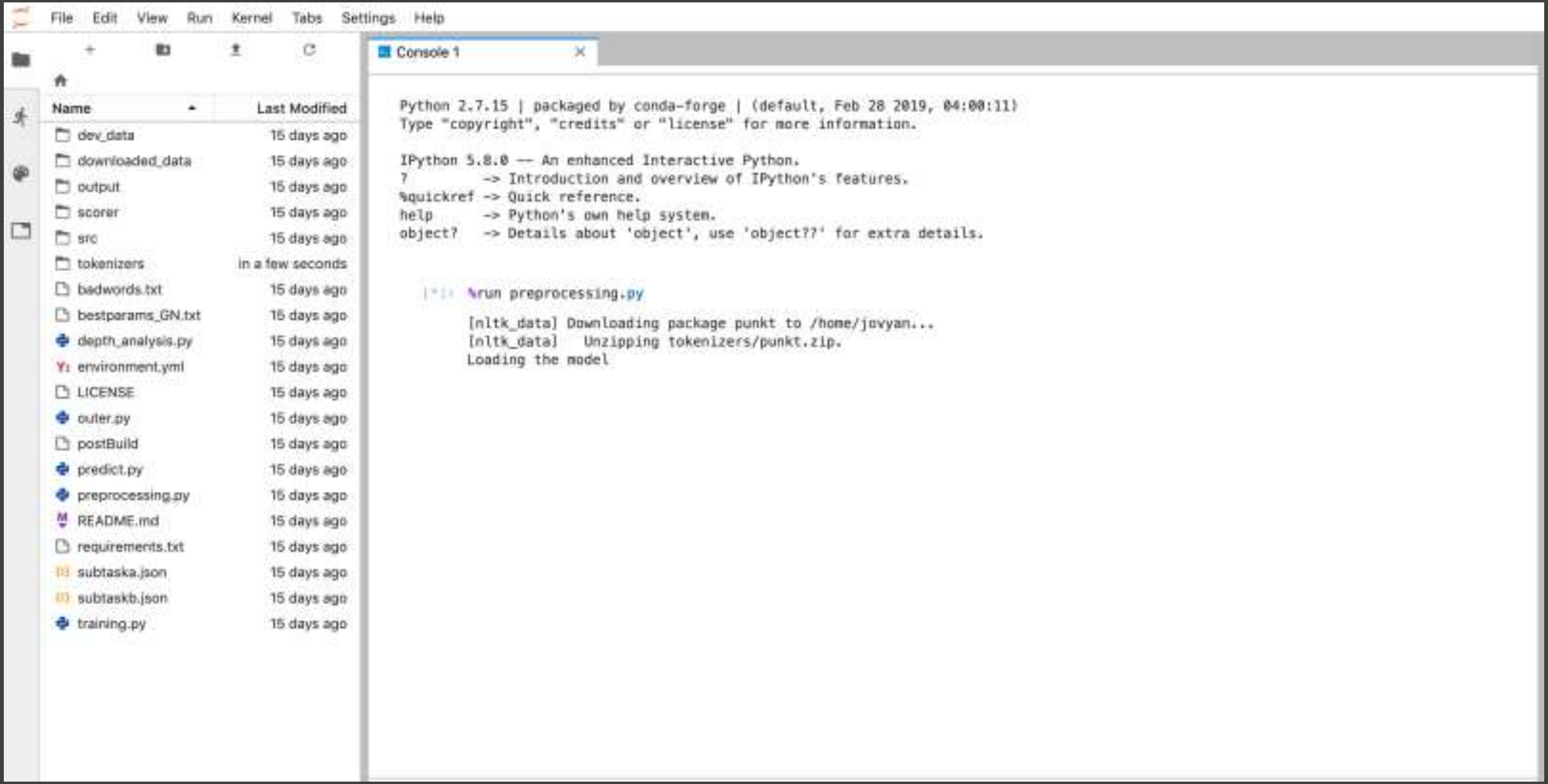


Label \ Prediction	C	D	Q	S
Commenting	760	0	12	6
Denying	68	0	1	2
Querying	69	0	36	1
Supporting	67	0	1	26

Table 5: Confusion matrix for testing set predictions

https://github.com/kochkinaelena/branchLSTM
https://doi.org/10.18653/v1/S17-2083

#PyDataLDN #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2598529



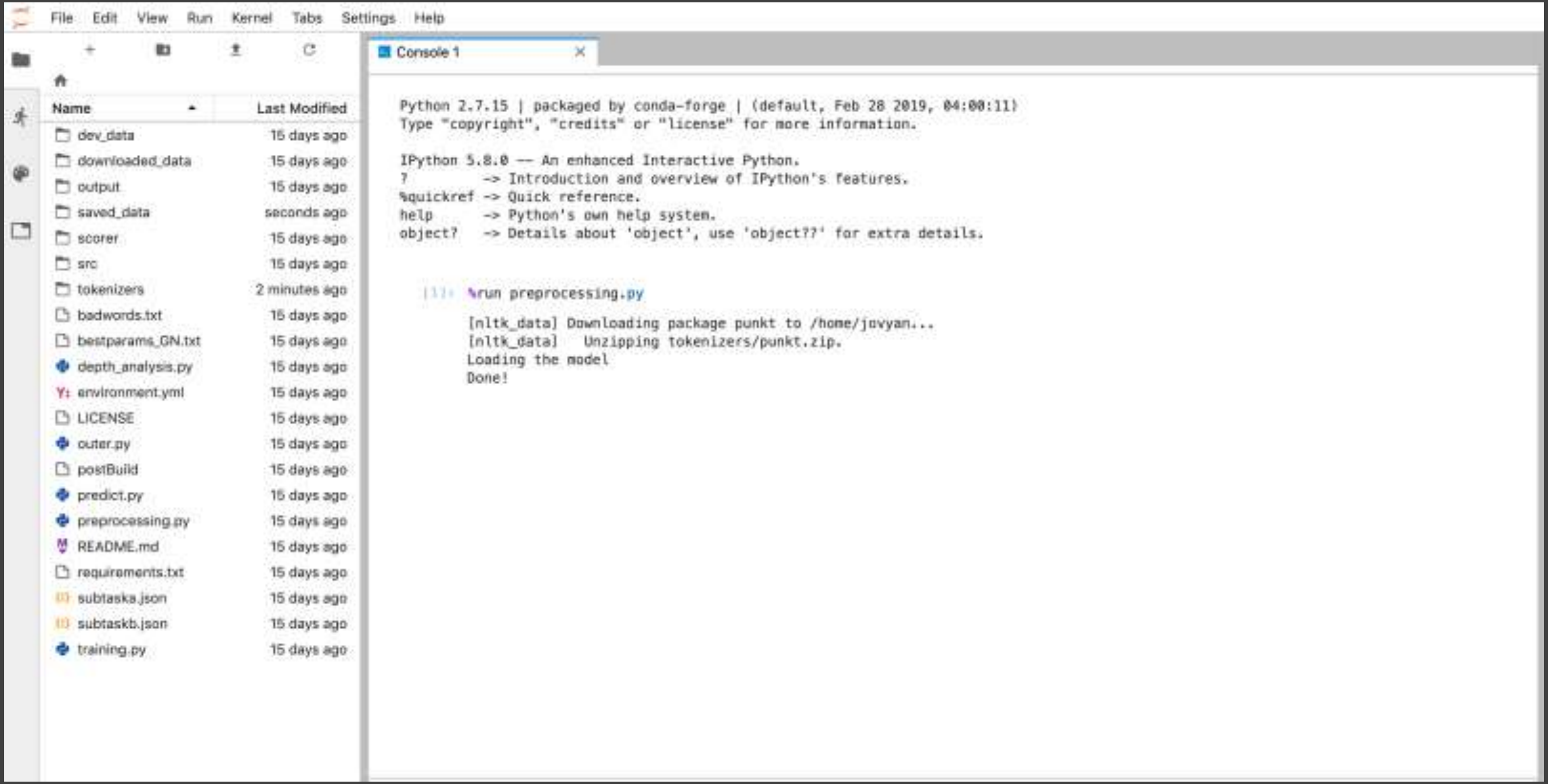
Name	Last Modified
dev_data	15 days ago
downloaded_data	15 days ago
output	15 days ago
scorer	15 days ago
src	15 days ago
tokenizers	in a few seconds
badwords.txt	15 days ago
bestparams_GN.txt	15 days ago
depth_analysis.py	15 days ago
environment.yml	15 days ago
LICENSE	15 days ago
outer.py	15 days ago
postBuild	15 days ago
predict.py	15 days ago
preprocessing.py	15 days ago
README.md	15 days ago
requirements.txt	15 days ago
subtaska.json	15 days ago
subtaskb.json	15 days ago
training.py	15 days ago

```
Python 2.7.15 | packaged by conda-forge | (default, Feb 28 2019, 04:00:11)
Type "copyright", "credits" or "license" for more information.

IPython 5.8.0 -- An enhanced Interactive Python.
?         -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help      -> Python's own help system.
object?   -> Details about 'object', use 'object??' for extra details.


    [*]: %run preprocessing.py

         [nltk_data] Downloading package punkt to /home/jovyan...
         [nltk_data]   Unzipping tokenizers/punkt.zip.
         Loading the model
```

https://github.com/kochkinaelena/branchLSTM (on Turing Way Hub)

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2598529



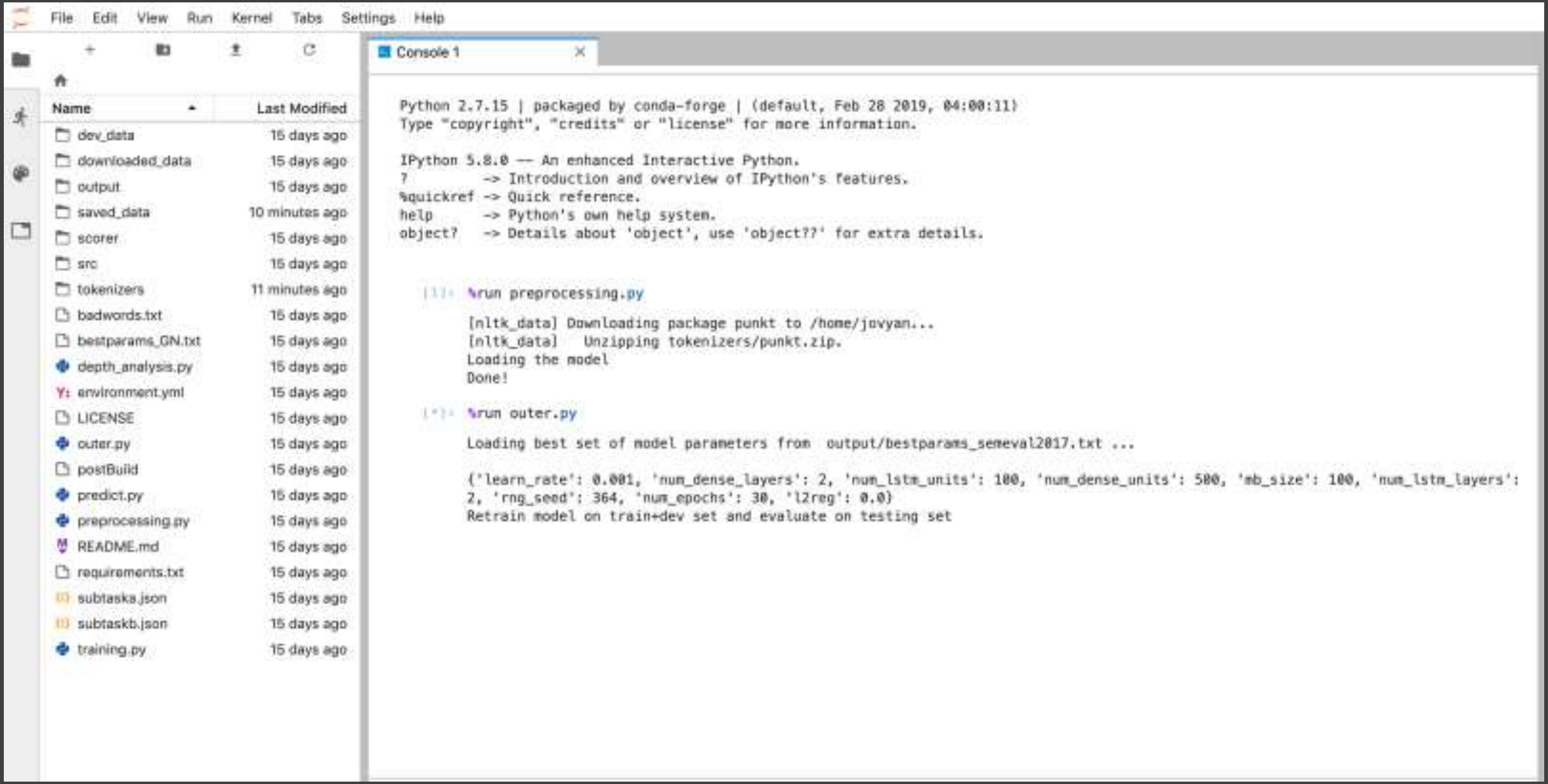
Name	Last Modified
dev_data	15 days ago
downloaded_data	15 days ago
output	15 days ago
saved_data	seconds ago
scorer	15 days ago
src	15 days ago
tokenizers	2 minutes ago
badwords.txt	15 days ago
bestparams_GN.txt	15 days ago
depth_analysis.py	15 days ago
environment.yml	15 days ago
LICENSE	15 days ago
outer.py	15 days ago
postBuild	15 days ago
predict.py	15 days ago
preprocessing.py	15 days ago
README.md	15 days ago
requirements.txt	15 days ago
subtaska.json	15 days ago
subtaskb.json	15 days ago
training.py	15 days ago

```
Python 2.7.15 | packaged by conda-forge | (default, Feb 28 2019, 04:00:11)
Type "copyright", "credits" or "license" for more information.

IPython 5.8.0 -- An enhanced Interactive Python.
?         -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help      -> Python's own help system.
object?   -> Details about 'object', use 'object??' for extra details.


   [1]: %run preprocessing.py

        [nltk_data] Downloading package punkt to /home/jovyan...
        [nltk_data]   Unzipping tokenizers/punkt.zip.
        Loading the model
        Done!
```

https://github.com/kochkinaelena/branchLSTM (on Turing Way Hub)

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2598529



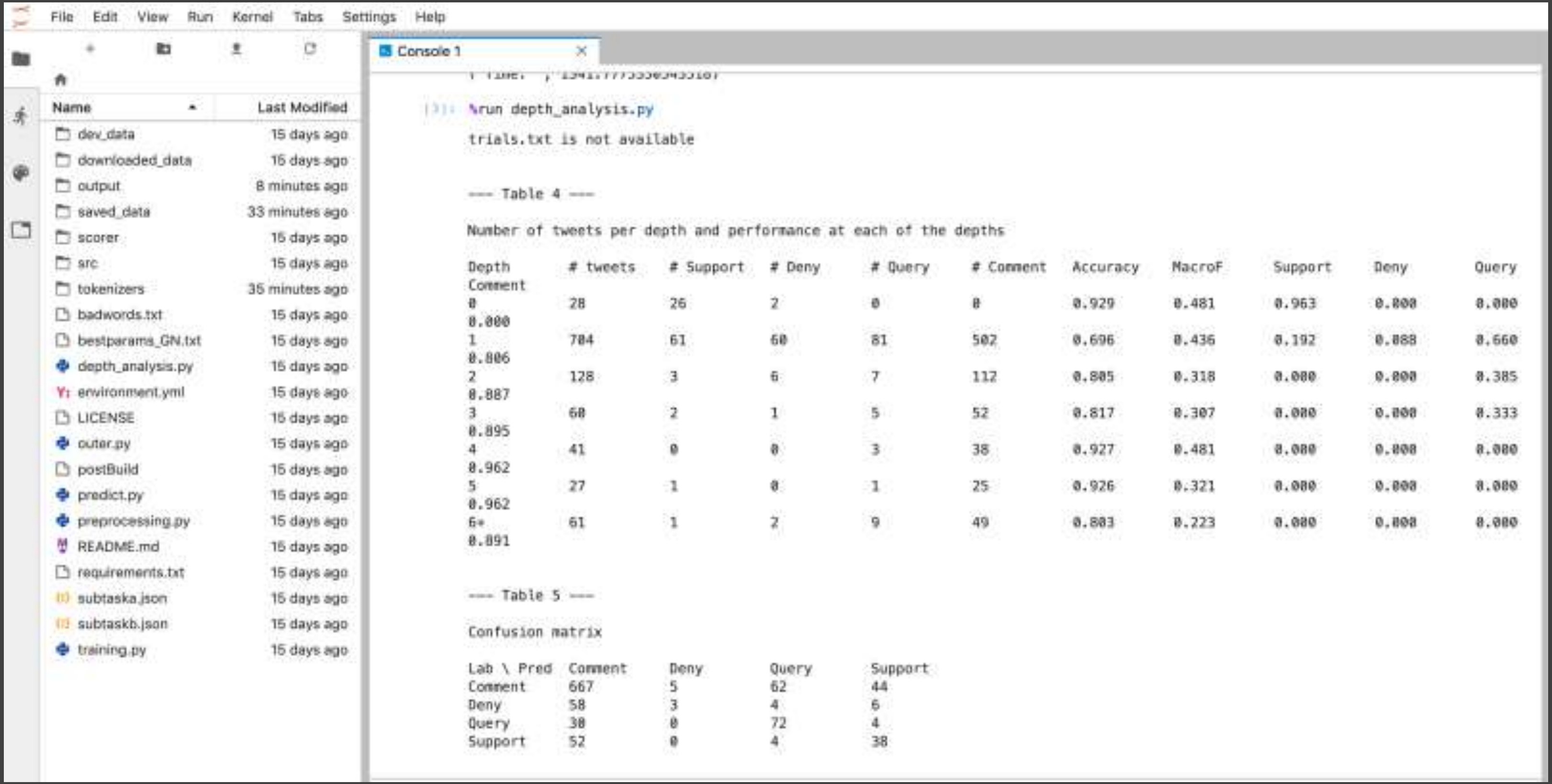
Name	Last Modified
dev_data	15 days ago
downloaded_data	15 days ago
output	15 days ago
saved_data	10 minutes ago
scorer	15 days ago
src	15 days ago
tokenizers	11 minutes ago
badwords.txt	15 days ago
bestparams_GN.txt	15 days ago
depth_analysis.py	15 days ago
environment.yml	15 days ago
LICENSE	15 days ago
outer.py	15 days ago
postBuild	15 days ago
predict.py	15 days ago
preprocessing.py	15 days ago
README.md	15 days ago
requirements.txt	15 days ago
subtaska.json	15 days ago
subtaskb.json	15 days ago
training.py	15 days ago

```
Python 2.7.15 | packaged by conda-forge | (default, Feb 28 2019, 04:00:11)
Type "copyright", "credits" or "license" for more information.

IPython 5.8.0 -- An enhanced Interactive Python.
?          -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help       -> Python's own help system.
object?    -> Details about 'object', use 'object??' for extra details.


  [1]: %run preprocessing.py

       [nltk_data] Downloading package punkt to /home/jovyan...
       [nltk_data]   Unzipping tokenizers/punkt.zip.
       Loading the model
       Done!

  [*]: %run outer.py

       Loading best set of model parameters from  output/bestparams_semeval2017.txt ...

       {'learn_rate': 0.001, 'num_dense_layers': 2, 'num_lstm_units': 100, 'num_dense_units': 500, 'mb_size': 100, 'num_lstm_layers':
       2, 'rng_seed': 364, 'num_epochs': 30, 'l2reg': 0.0}
       Retrain model on train+dev set and evaluate on testing set
```

https://github.com/kochkinaelena/branchLSTM (on Turing Way Hub)

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2598529



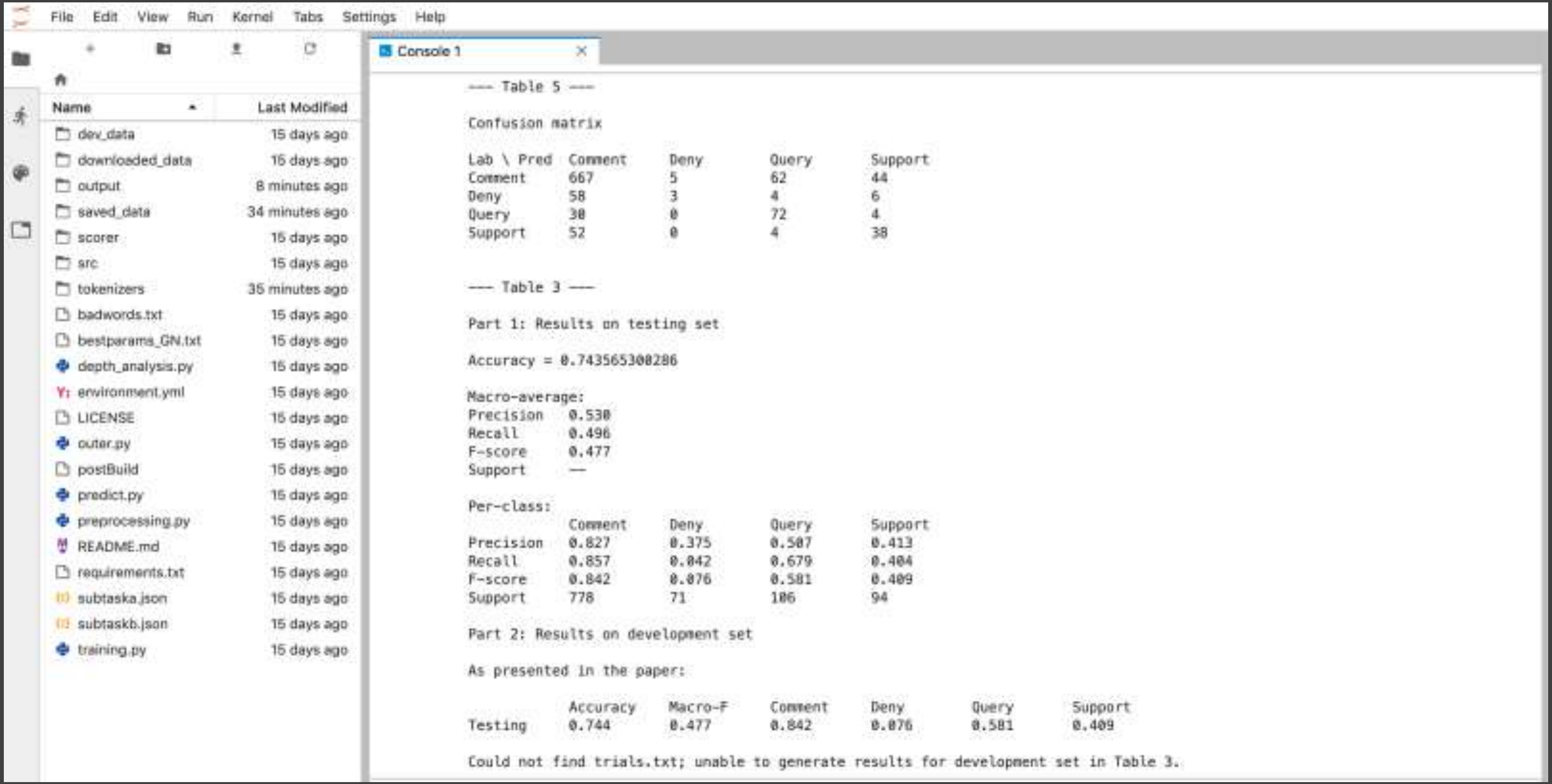
Name	Last Modified
dev_data	15 days ago
downloaded_data	15 days ago
output	8 minutes ago
saved_data	33 minutes ago
scorer	15 days ago
src	15 days ago
tokenizers	35 minutes ago
badwords.txt	15 days ago
bestparams_GN.txt	15 days ago
depth_analysis.py	15 days ago
environment.yml	15 days ago
LICENSE	15 days ago
outer.py	15 days ago
postBuild	15 days ago
predict.py	15 days ago
preprocessing.py	15 days ago
README.md	15 days ago
requirements.txt	15 days ago
subtaska.json	15 days ago
subtaskb.json	15 days ago
training.py	15 days ago

```
[3]: %run depth_analysis.py
     trials.txt is not available


     --- Table 4 ---

     Number of tweets per depth and performance at each of the depths

     Depth       # tweets    # Support   # Deny      # Query     # Comment   Accuracy    MacroF      Support     Deny        Query
     Comment
     0           28          26          2           0           0           0.929       0.481       0.963       0.000       0.000
     0.000
     1           704         61          60          81          502         0.696       0.436       0.192       0.088       0.660
     0.806
     2           128         3           6           7           112         0.805       0.318       0.000       0.000       0.385
     0.887
     3           60          2           1           5           52          0.817       0.307       0.000       0.000       0.333
     0.895
     4           41          0           0           3           38          0.927       0.481       0.000       0.000       0.000
     0.962
     5           27          1           0           1           25          0.926       0.321       0.000       0.000       0.000
     0.962
     6+          61          1           2           9           49          0.803       0.223       0.000       0.000       0.000
     0.891


     --- Table 5 ---

     Confusion matrix

     Lab \ Pred  Comment     Deny        Query       Support
     Comment     667         5           62          44
     Deny        58          3           4           6
     Query       38          0           72          4
     Support     52          0           4           38
```

https://github.com/kochkinaelena/branchLSTM (on Turing Way Hub)

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2598529



Name	Last Modified
dev_data	15 days ago
downloaded_data	15 days ago
output	8 minutes ago
saved_data	34 minutes ago
scorer	15 days ago
src	15 days ago
tokenizers	35 minutes ago
badwords.txt	15 days ago
bestparams_GN.txt	15 days ago
depth_analysis.py	15 days ago
environment.yml	15 days ago
LICENSE	15 days ago
outer.py	15 days ago
postBuild	15 days ago
predict.py	15 days ago
preprocessing.py	15 days ago
README.md	15 days ago
requirements.txt	15 days ago
subtaska.json	15 days ago
subtaskb.json	15 days ago
training.py	15 days ago

```
--- Table 5 ---

Confusion matrix

Lab \ Pred  Comment     Deny        Query       Support
Comment     667         5           62          44
Deny        58          3           4           6
Query       30          0           72          4
Support     52          0           4           38


--- Table 3 ---

Part 1: Results on testing set

Accuracy = 0.743565300286

Macro-average:
Precision   0.530
Recall      0.496
F-score     0.477
Support     --

Per-class:
            Comment     Deny        Query       Support
Precision   0.827       0.375       0.507       0.413
Recall      0.857       0.042       0.679       0.404
F-score     0.842       0.076       0.581       0.409
Support     778         71          106         94

Part 2: Results on development set

As presented in the paper:

            Accuracy    Macro-F     Comment     Deny        Query       Support
Testing     0.744       0.477       0.842       0.076       0.581       0.409

Could not find trials.txt; unable to generate results for development set in Table 3.
```

https://github.com/kochkinaelena/branchLSTM (on Turing Way Hub)

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2598529

Elena Kochkina

“How would I have known that it would be different on a different machine?! I only have access to the university HPC to run deep learning analyses.”

https://warwick.ac.uk/fac/sci/dcs/people/research/mapmbc

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2598529