Agenda

Boost your reproducibility with Binder

- 13:30 Registration and introductions
- 13:40 Introduction to the workshop and The Turing Way
- 13:50 Presentation: Why you need a reproducible computing environment and how Binder can help
- 15:00 Coffee break
- 15:30 Code along demo: Zero to Binder, build a Binder resource
- 16:30 Build your own Binder
- 16:50 Feedback, group picture and close

The Alan Turing Institute

The Turing Way

Boost your reproducibility with Binder workshop

Kirstie Whitaker

Pronouns: she/her

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.3632909

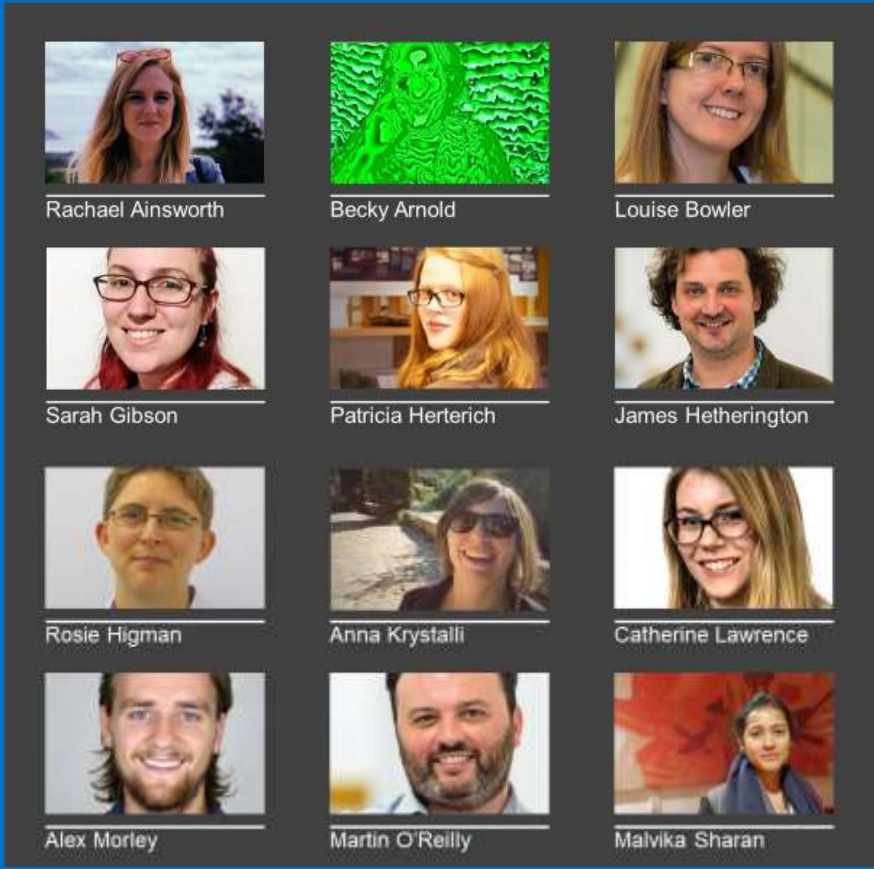
Agenda

Boost your reproducibility with Binder

- 13:30 Registration and introductions
- 13:40 Introduction to the workshop and The Turing Way
- 13:50 Presentation: Why you need a reproducible computing environment and how Binder can help
- 15:00 Coffee break
- 15:30 Code along demo: Zero to Binder, build a Binder resource
- 16:30 Build your own Binder
- 16:50 Feedback, group picture and close

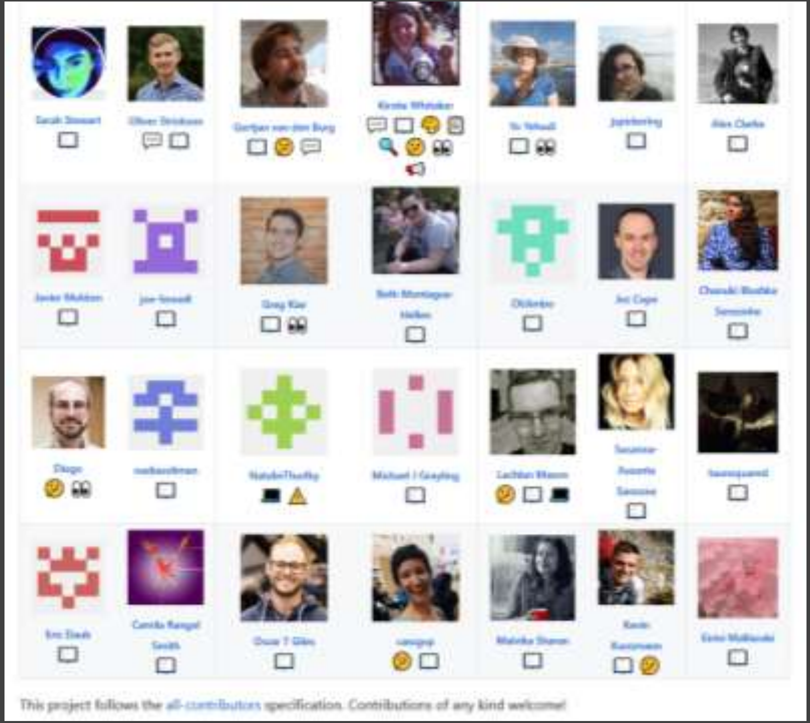
The Turing Way is:

- a book
- a community
- a global collaboration
- a whole tonne of work



#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.3632909

Thank you to all our contributors



This project follows the all-contributors specification. Contributions of any kind welcome!

https://github.com/alan-turing-institute/the-turing-way#contributors
https://allcontributors.org/docs/en/emoji-key

#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.3632909

The Turing Institute

https://www.turing.ac.uk/news/enigma-machine-goes-display-alan-turing-institute

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.3632909

https://www.bbc.co.uk/programmes/p0704h04

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.3632909

https://bletchleypark.org.uk

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.3632909

University network

UNIVERSITY OF CAMBRIDGE

THE UNIVERSITY of EDINBURGH

UNIVERSITY OF OXFORD

UCL

WARWICK
THE UNIVERSITY OF WARWICK

University of BRISTOL

UNIVERSITY OF BIRMINGHAM

UNIVERSITY OF EXETER

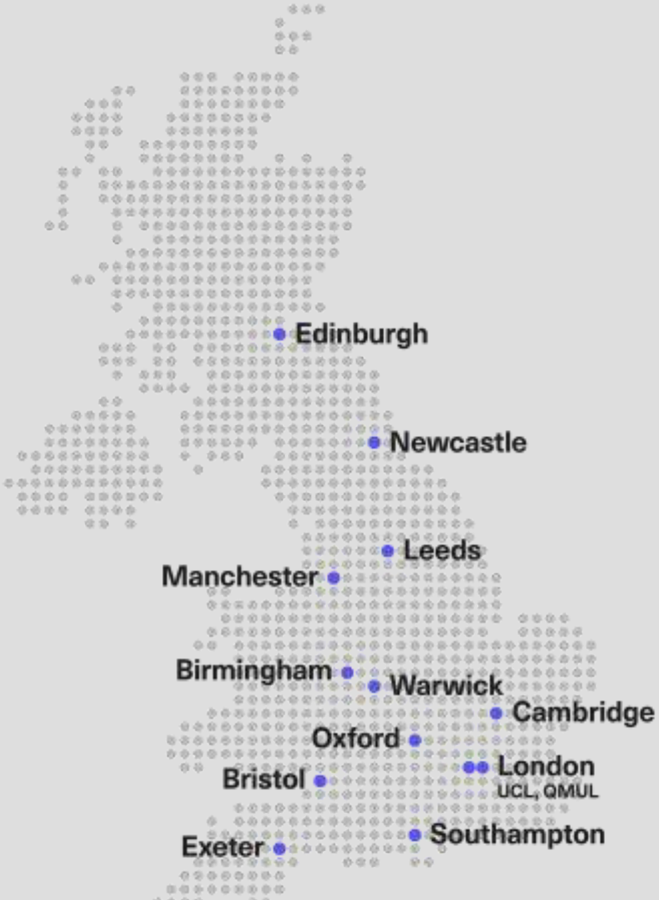
UNIVERSITY OF LEEDS

MANCHESTER 1824
The University of Manchester

Queen Mary
University of London

Newcastle University

UNIVERSITY OF Southampton



#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.3632909

The Institute's partners and collaborators

Lloyd's Register Foundation

intel

GCHQ

dstl

Ministry of Defence

HSBC

uclh

accenture

NATS

British Heart Foundation

Microsoft

TOYOTA mobility FOUNDATION

Office for National Statistics

GREATER LONDON AUTHORITY

CRAY

nationalgrid

SIEMENS

LLOYDS BANKING GROUP

Cystic Fibrosis Trust

ROLLS ROYCE

Scottish Enterprise

Nuffield Foundation

SAMSUNG

BRITISH AIRWAYS

Challenges

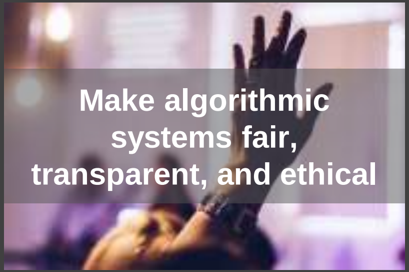
Advance data science and artificial intelligence to…

Revolutionise healthcare

Deliver safer, smarter engineering

Manage security in an insecure world

Shine a light on our economy

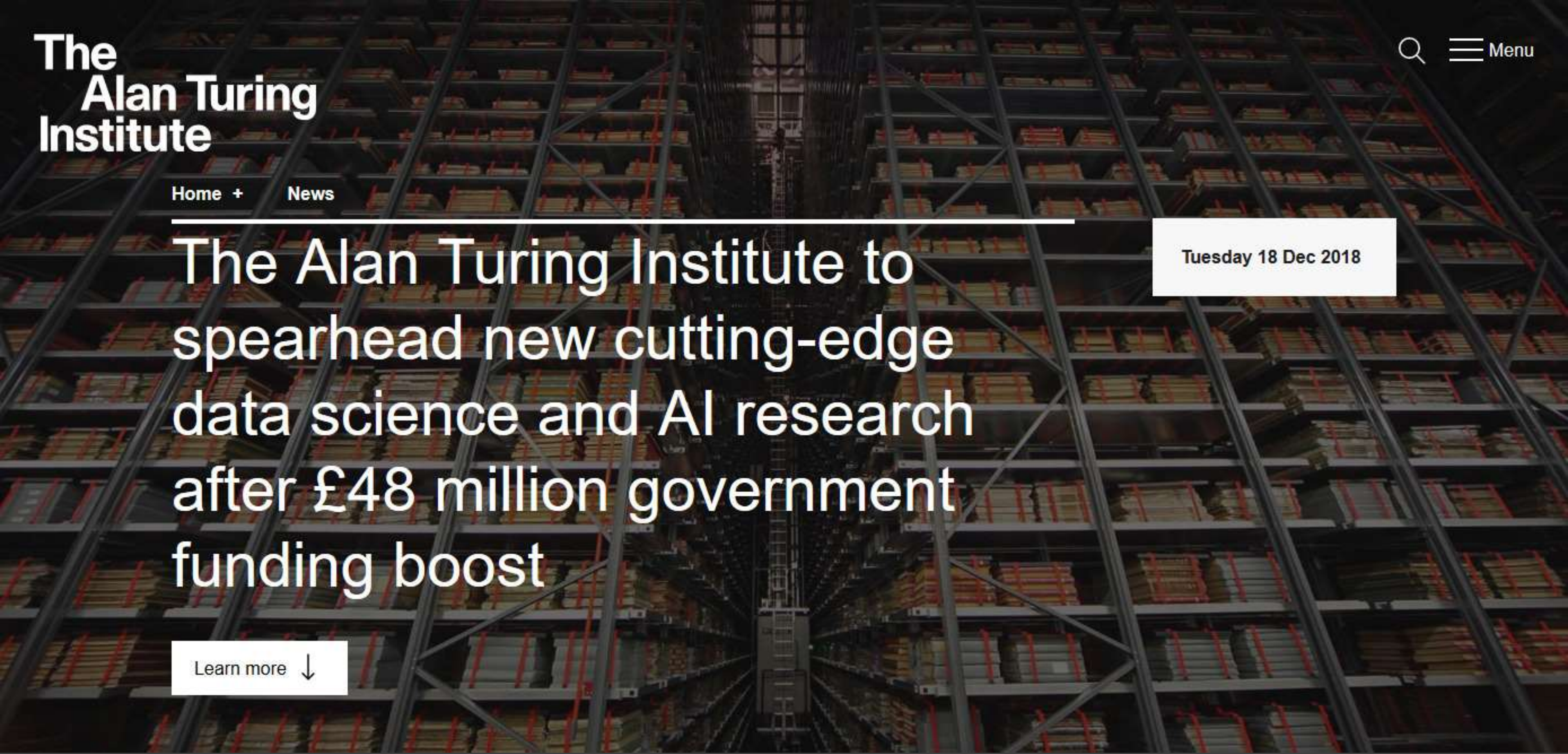


Make algorithmic systems fair, transparent, and ethical

Design computers for the next generation of algorithms

Supercharge research in science and humanities

Foster government innovation



The Alan Turing Institute to spearhead new cutting-edge data science and AI research after £48 million government funding boost

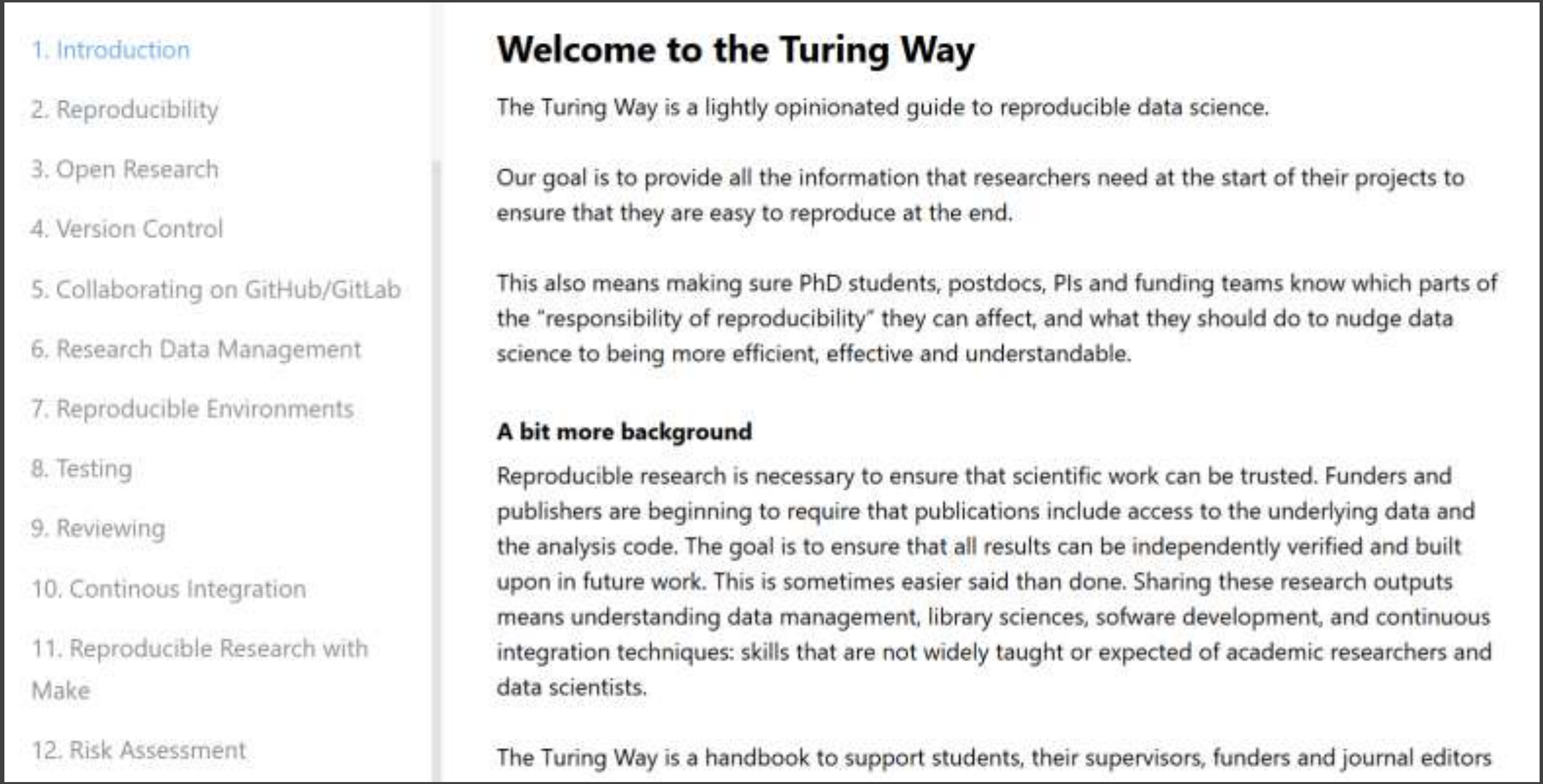
Tuesday 18 Dec 2018

https://www.turing.ac.uk/news/alan-turing-institute-spearhead-new-cutting-edge-data-science-and-artificial-intelligence

#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.3632909

The Turing Way

#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.3632909


1. Introduction
2. Reproducibility
3. Open Research
4. Version Control
5. Collaborating on GitHub/GitLab
6. Research Data Management
7. Reproducible Environments
8. Testing
9. Reviewing
10. Continous Integration
11. Reproducible Research with Make
12. Risk Assessment

Welcome to the Turing Way

The Turing Way is a lightly opinionated guide to reproducible data science.

Our goal is to provide all the information that researchers need at the start of their projects to ensure that they are easy to reproduce at the end.

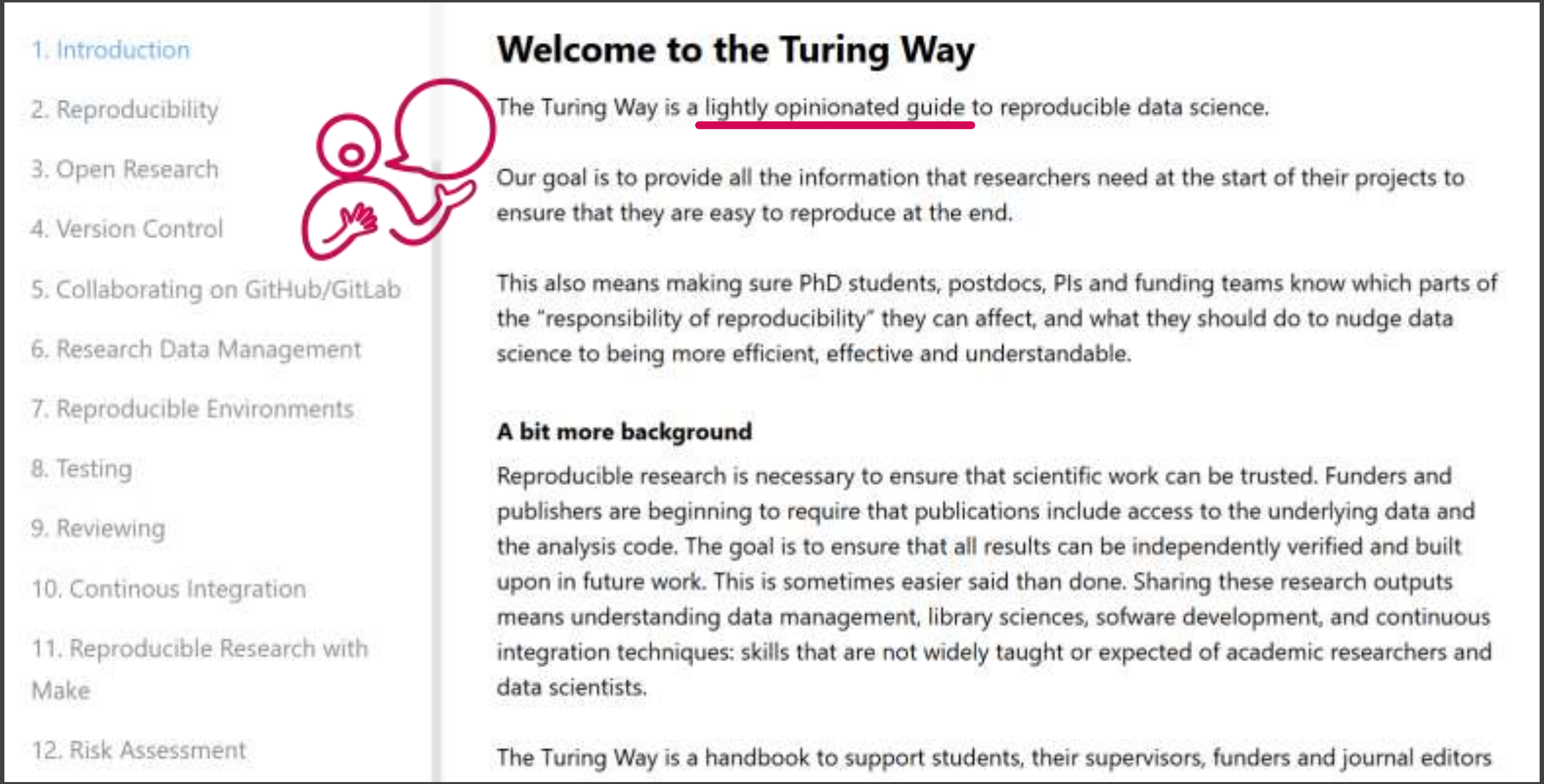
This also means making sure PhD students, postdocs, PIs and funding teams know which parts of the "responsibility of reproducibility" they can affect, and what they should do to nudge data science to being more efficient, effective and understandable.

A bit more background

Reproducible research is necessary to ensure that scientific work can be trusted. Funders and publishers are beginning to require that publications include access to the underlying data and the analysis code. The goal is to ensure that all results can be independently verified and built upon in future work. This is sometimes easier said than done. Sharing these research outputs means understanding data management, library sciences, sofware development, and continuous integration techniques: skills that are not widely taught or expected of academic researchers and data scientists.

The Turing Way is a handbook to support students, their supervisors, funders and journal editors

https://the-turing-way.netlify.com/introduction/introduction
#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.3632909



Welcome to the Turing Way

The Turing Way is a lightly opinionated guide to reproducible data science.

Our goal is to provide all the information that researchers need at the start of their projects to ensure that they are easy to reproduce at the end.

This also means making sure PhD students, postdocs, PIs and funding teams know which parts of the "responsibility of reproducibility" they can affect, and what they should do to nudge data science to being more efficient, effective and understandable.

A bit more background

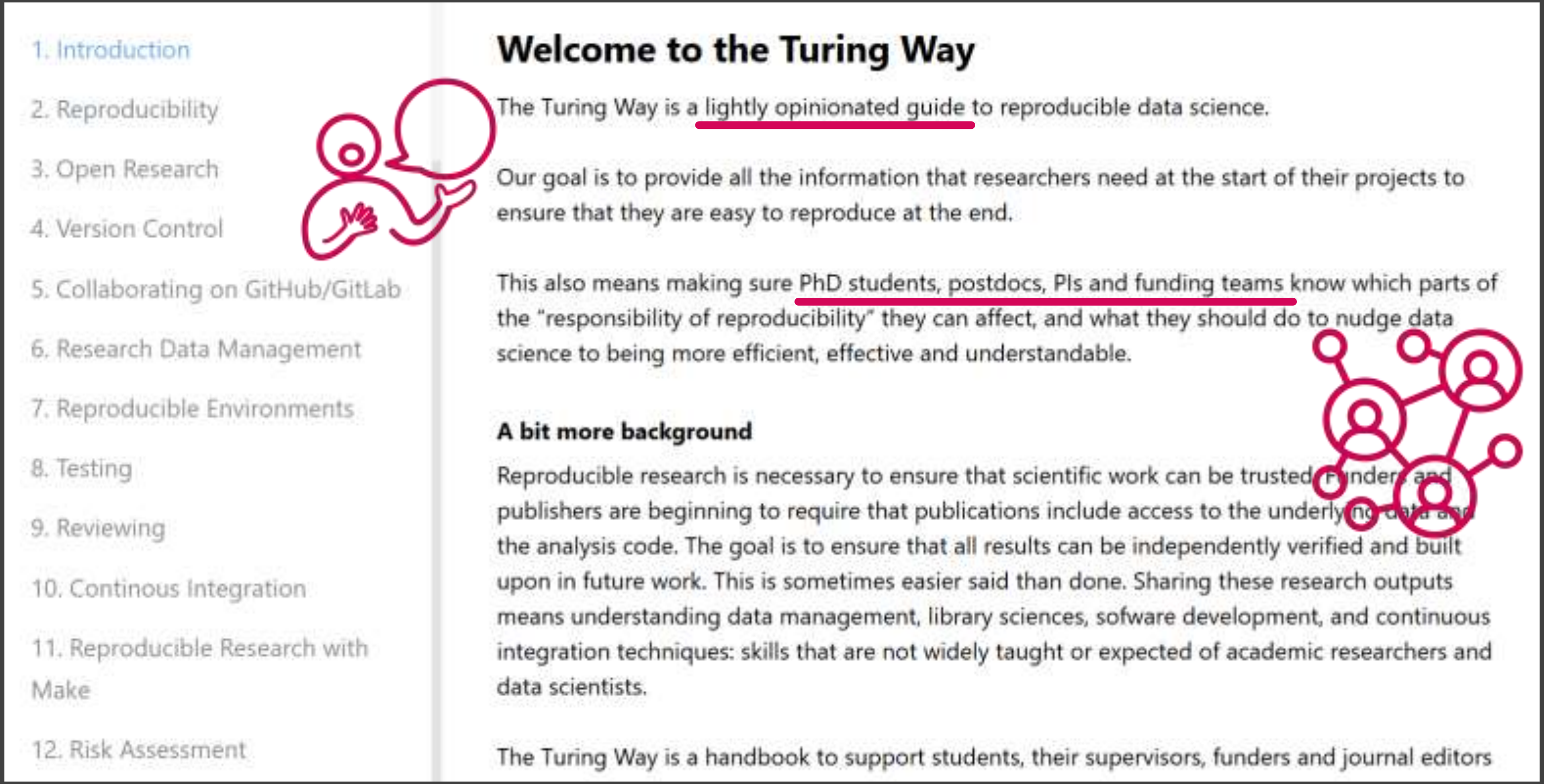
Reproducible research is necessary to ensure that scientific work can be trusted. Funders and publishers are beginning to require that publications include access to the underlying data and the analysis code. The goal is to ensure that all results can be independently verified and built upon in future work. This is sometimes easier said than done. Sharing these research outputs means understanding data management, library sciences, sofware development, and continuous integration techniques: skills that are not widely taught or expected of academic researchers and data scientists.

The Turing Way is a handbook to support students, their supervisors, funders and journal editors

https://the-turing-way.netlify.com/introduction/introduction

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.3632909



Welcome to the Turing Way

The Turing Way is a lightly opinionated guide to reproducible data science.

Our goal is to provide all the information that researchers need at the start of their projects to ensure that they are easy to reproduce at the end.

This also means making sure PhD students, postdocs, PIs and funding teams know which parts of the "responsibility of reproducibility" they can affect, and what they should do to nudge data science to being more efficient, effective and understandable.

A bit more background

Reproducible research is necessary to ensure that scientific work can be trusted. Funders and publishers are beginning to require that publications include access to the underlying data and the analysis code. The goal is to ensure that all results can be independently verified and built upon in future work. This is sometimes easier said than done. Sharing these research outputs means understanding data management, library sciences, sofware development, and continuous integration techniques: skills that are not widely taught or expected of academic researchers and data scientists.

The Turing Way is a handbook to support students, their supervisors, funders and journal editors

https://the-turing-way.netlify.com/introduction/introduction

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.3632909

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

https://the-turing-way.netlify.com/reproducibility/03/definitions.html
#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.3632909

Barriers to reproducible research

Is not considered for promotion

Held to higher standards than others

Publication bias towards novel findings

Requires additional skills

Plead the 5th

Support additional users

Takes time

https://doi.org/10.6084/m9.figshare.5537101

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.3632909

Barriers to reproducible research

Is not considered for promotion

Held to higher standards than others

Publication bias towards novel findings

Requires additional skills

Plead the 5th

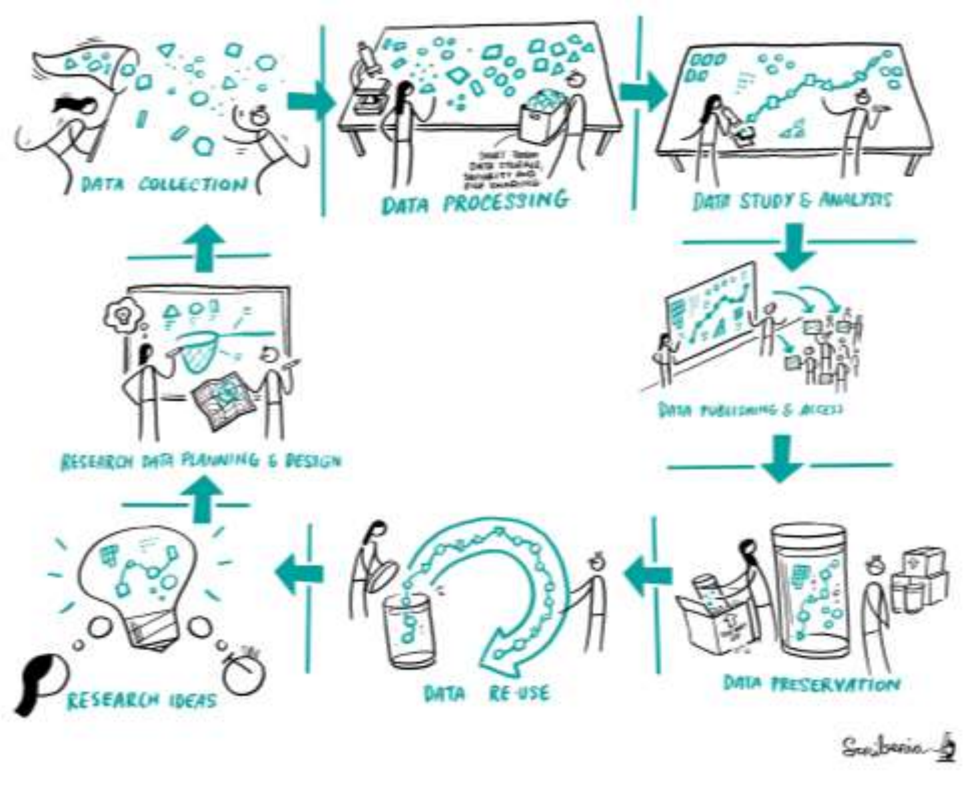
Support additional users

Takes time

https://doi.org/10.6084/m9.figshare.5537101
#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.3632909

To be fully reproducible we have to cover all the steps of the research cycle

And that is super overwhelming…but we’re here to help



PERSISTANT
FINDABLE
MEANINGFUL INTERACTION
ACCESSIBLE
FAIR
PRINCIPLES
INTEROPERABLE
FULL DISCLOSURE
REUSABLE

#TuringWay @kirstie_j @mybinderteam

Wilkinson et al., Sci. Data, 2016. doi: 10.1038/sdata.2016.18

https://doi.org/10.5281/zenodo.3632909

Testing for research

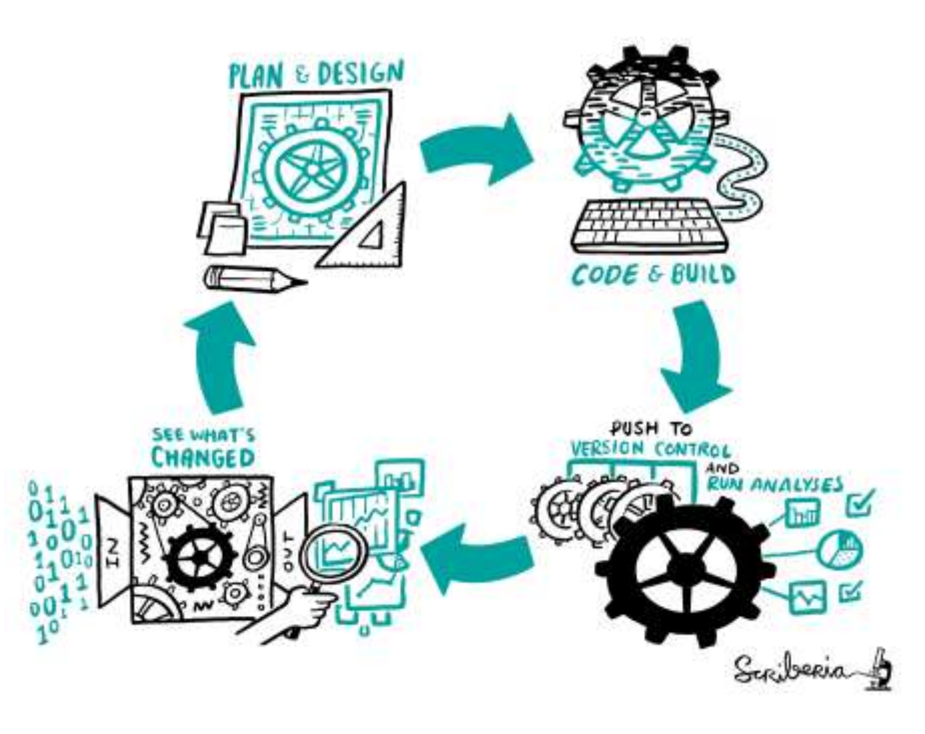
```
Assert.AreEqual(
   GetTimeOfDay(),
      “Morning” )
```

```
Assert.AreEqual(
    GetTimeOfDay(),
    "Morning" );
```

https://www.toptal.com/qa/how-to-write-testable-code-and-why-it-matters
https://the-turing-way.netlify.com/testing/testing.html

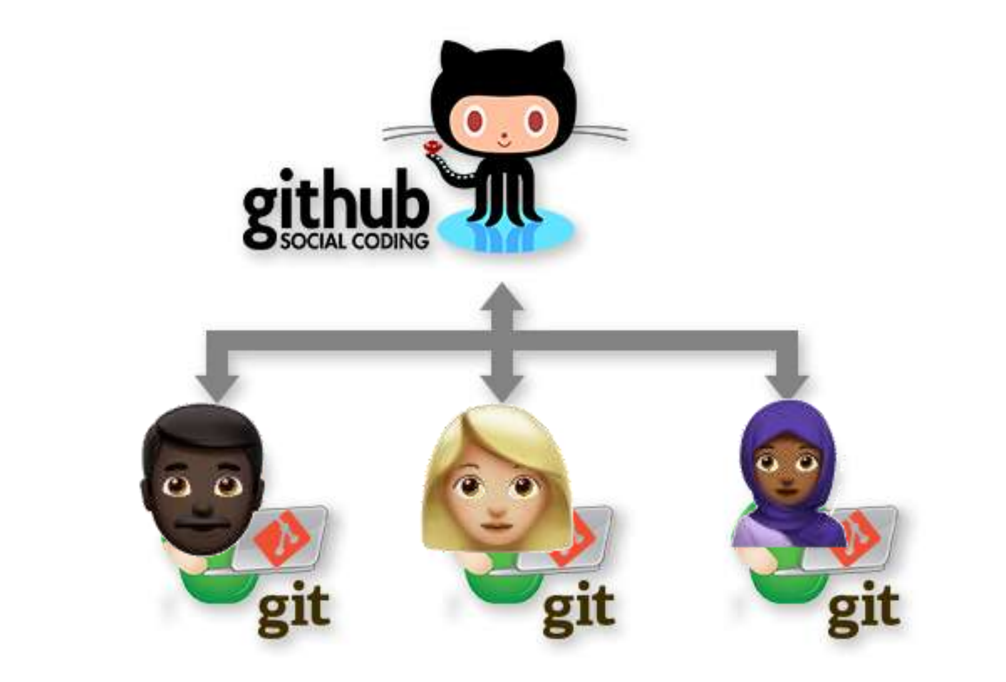
Continuous integration

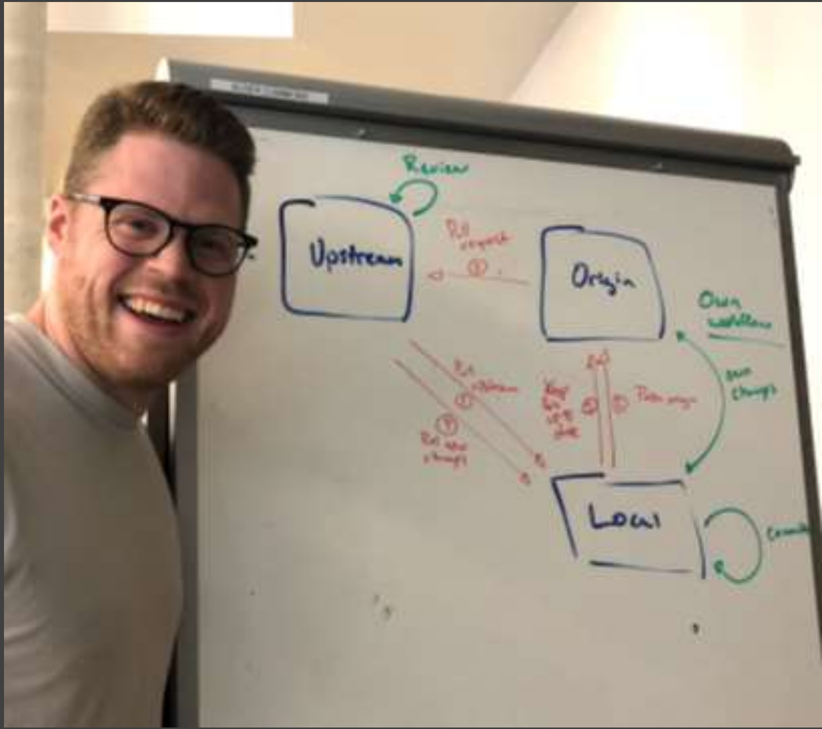
- Plan and design your experiment
- Write down the steps in code
- Push to version control and run the analyses
- Test to see what's changed



https://the-turing-way.netlify.com/continuous_integration/continuous_integration.html

#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.3632909



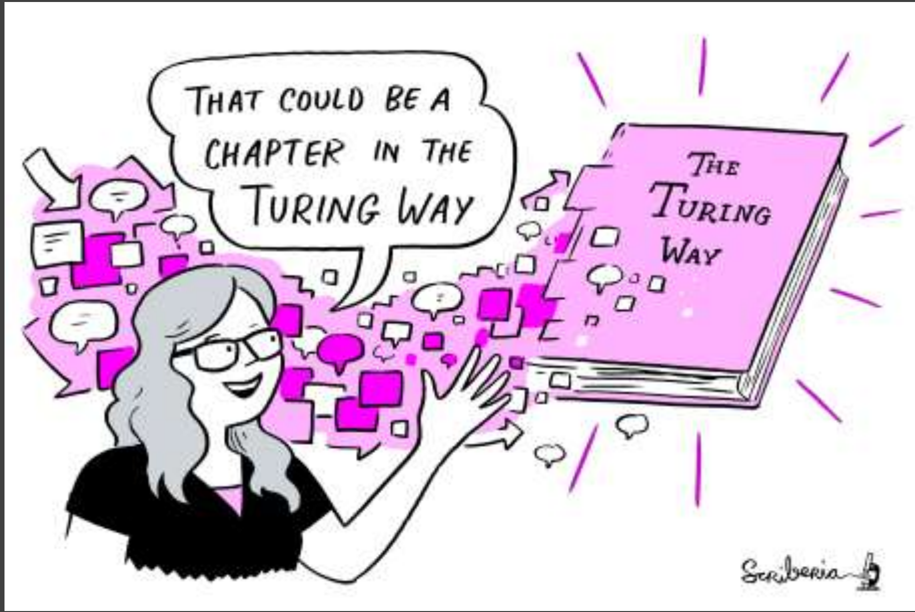


https://the-turing-way.netlify.com/collaborating_github/collaborating_github.html
https://the-turing-way.netlify.com/version_control/version_control.html
https://neurohackademy.org

#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.3632909

Extension in 2020

- Expand scope to all data science practices
  - Reproducibility
  - Scoping and designing a data science project
  - Ethics
  - Communication and visualisation
  - Collaborative working



https://github.com/alan-turing-institute/the-turing-way/blob/master/project_management/tps-funding-application-20190429.md

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.3632909

A global collaboration

Patricia Herterich

“What really sets The Turing Way apart is HOW we’re writing the book. The focus on community, the commitment to transparency and working open right from the beginning is an exciting (and terrifying) new way of working.”

https://rd-alliance.org/users/patricia-herterich

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.3632909

Open Leadership Principles

Understanding
You make the work accessible and clear

Sharing
You make the work easy to adapt, reproduce, and spread

Participation & Inclusion
You build shared ownership and agency to make the work inviting and sustainable for all.

Read more
https://mozilla.github.io/olm-whitepaper

#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.3632909

Agenda

Boost your reproducibility with Binder

- 13:30 Registration and introductions
- 13:40 Introduction to the workshop and The Turing Way
- 13:50 Presentation: Why you need a reproducible computing environment and how Binder can help
- 15:00 Coffee break
- 15:30 Code along demo: Zero to Binder, build a Binder resource
- 16:30 Build your own Binder
- 16:50 Feedback, group picture and close

Goals for the workshop

- Understand how your computational environment impacts reproducibility
- Learn what Binder is and how it can help make your research reproducible
- Build your own Binder!

#TuringWay @kirstie_j @mybinderteam

https://doi.org/10.5281/zenodo.3632909

Our Code of Conduct

"The Turing Way team are dedicated to providing a welcoming and supportive environment for all people…we do not tolerate behaviour that is disrespectful to our community members or that excludes, intimidates, or causes discomfort to others."

https://github.com/alan-turing-institute/the-turing-way/blob/master/CODE_OF_CONDUCT.md

#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.3632909

– Be respectful of different viewpoints and experiences.

– Use welcoming and inclusive language.

– Do not harass people.

– Respect the privacy and safety of others.

Please do not take pictures of anyone without their permission.

– Be considerate of others’ participation.

– Don’t be a bystander.

https://github.com/alan-turing-institute/the-turing-way/blob/master/CODE_OF_CONDUCT.md

- – Be respectful of different viewpoints and experiences.
- – Use welcoming and inclusive language.
- – Do not harass people.
- – Respect the privacy and safety of others.

  Please do not take pictures of anyone without their permission.
- – Be considerate of others’ participation.
- – Don’t be a bystander.
- – **Anita**, **Felix** and **Jeremy** are here to help 👋

https://github.com/alan-turing-institute/the-turing-way/
blob/master/CODE_OF_CONDUCT.md

#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.3632909

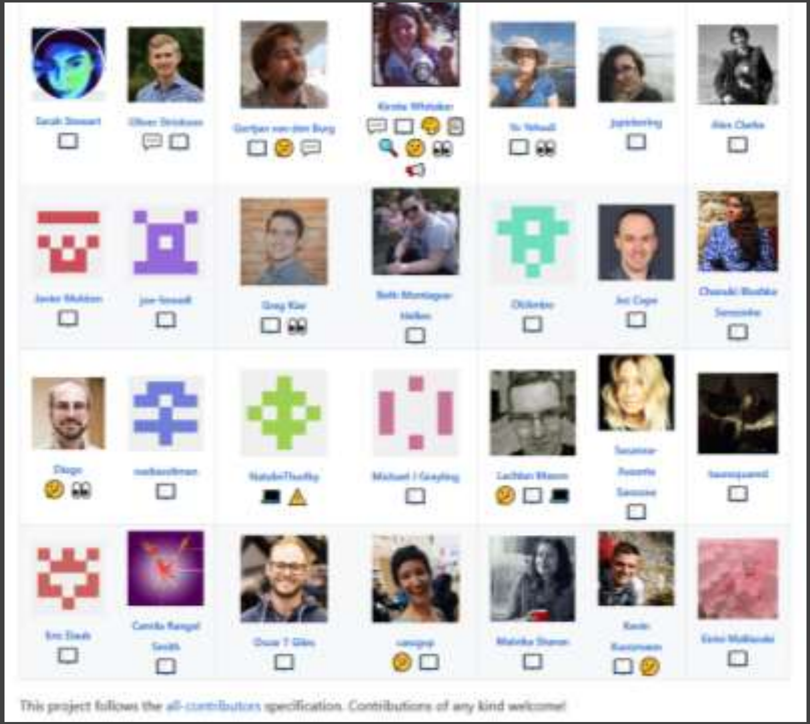
Thank you to current (& future) contributors

Rachael Ainsworth, Tarek Allam, Tania Allard, Becky Arnold, Louise Bowler, Stephan Druskat, Stephen Eglen

Oliver Forrest, Jason Gates, Sarah Gibson, Richard Gilham, Tim Head, Patricia Herterich, Rosie Higman

Ian Hinder, Hao Hoang, Dan Hobley, Chris Holdgraf, Will Hulme, Anna Krystalli, Clare Liggins

Robin Long, Alexander Morley, Martin O'Reilly, Rosa Rosekoff, James Robinson, Ali Seyhun Saral, Andrew Stewart



Sarah Stewart, Oliver Strickson, Gertjan van den Burg, Kirstie Whitaker, Yo Yehudi, jsjesbering, Alex Clarke

Javier Moldon, joe-tessadl, Greg Kiar, Beth Montague-Hellen, Olcturbo, Jez Cope, Chanuki Illushka Seresinhe

Diego, nadiasoliman, NatalieThurlby, Michael J Grayling, Lachlan Mason, Susanna-Assunta Sansone, basnsquared

Eric Daub, Camila Rangel Smith, Oscar T Giles, sangrep, Malvika Sharan, Kevin Kunzmann, Enko Mallianski

This project follows the all-contributors specification. Contributions of any kind welcome!

https://github.com/alan-turing-institute/the-turing-way#contributors
https://allcontributors.org/docs/en/emoji-key

#TuringWay @kirstie_j @mybinderteam
https://doi.org/10.5281/zenodo.3632909

Thank you

The Alan Turing Institute

moz://a

UNIVERSITY OF CAMBRIDGE

- Book: https://the-turing-way.netlify.com
- Newsletter: https://tinyletter.com/TuringWay
- GitHub: https://github.com/alan-turing-institute/the-turing-way
- Chat: https://gitter.im/alan-turing-institute/the-turing-way
- This work was supported by The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1, particularly the "Tools, Practices and Systems" theme within that grant, and by The Alan Turing Institute under the EPSRC grant EP/N510129/1
- Unsplash photos by Adolfo Felix, James Pond, Jose Alejandro Cuffia, Kinson Leung, Mateo Vrbnjak, Mimi thian, Omar Albeik, Perry Grone, Toa Heftiba, Tomasz Frankows, Wilmer Martinez
- Noun Project icons by Aybige, Luis Prado, Edward Boatman, Becris, Rose Alice Design, Hyemm.work
- Original artwork by Scriberia: https://doi.org/10.5281/zenodo.3332807

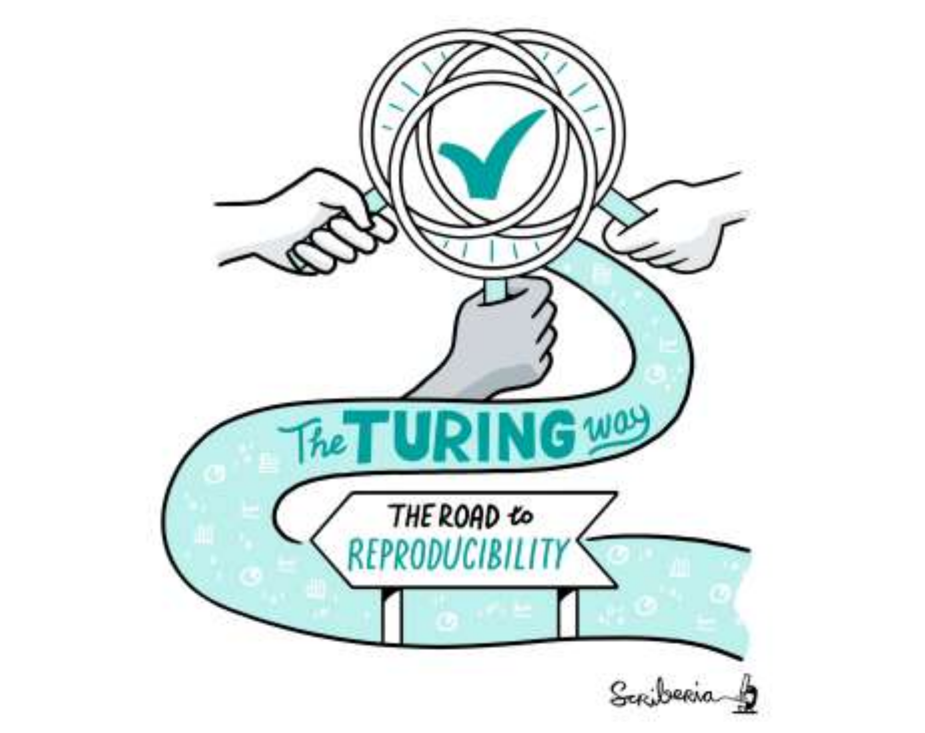
Agenda

Boost your reproducibility with Binder

- 13:30 Registration and introductions
- 13:40 Introduction to the workshop and The Turing Way
- 13:50 Presentation: Why you need a reproducible computing environment and how Binder can help
- 15:00 Coffee break
- 15:30 Code along demo: Zero to Binder, build a Binder resource
- 16:30 Build your own Binder
- 16:50 Feedback, group picture and close

A note on the name

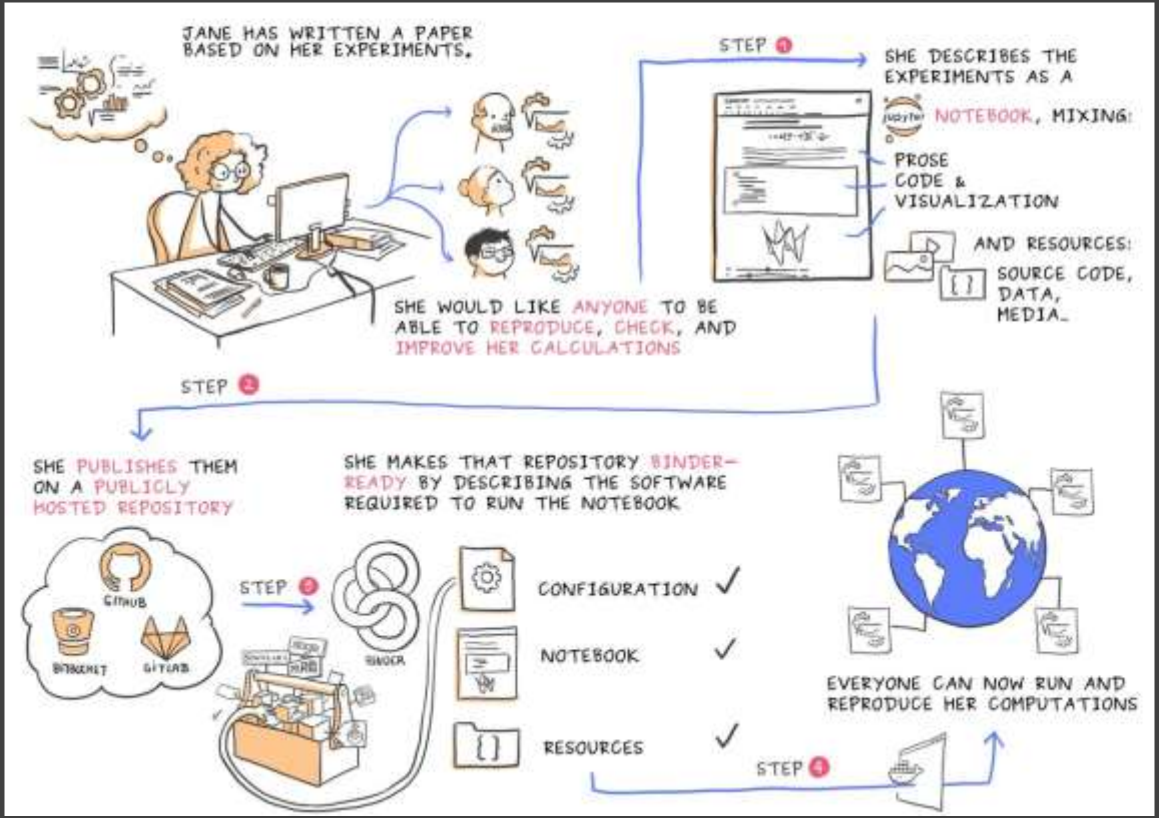
- I never thought the name would be approved!
- This is not a Turing project (although it has great support from the Institute)
- We are creating guidance together, the way is a journey not a set of rules



https://www.turing.ac.uk/research/impact-stories/changing-culture-data-science

#OpenMRB2020 #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3615258

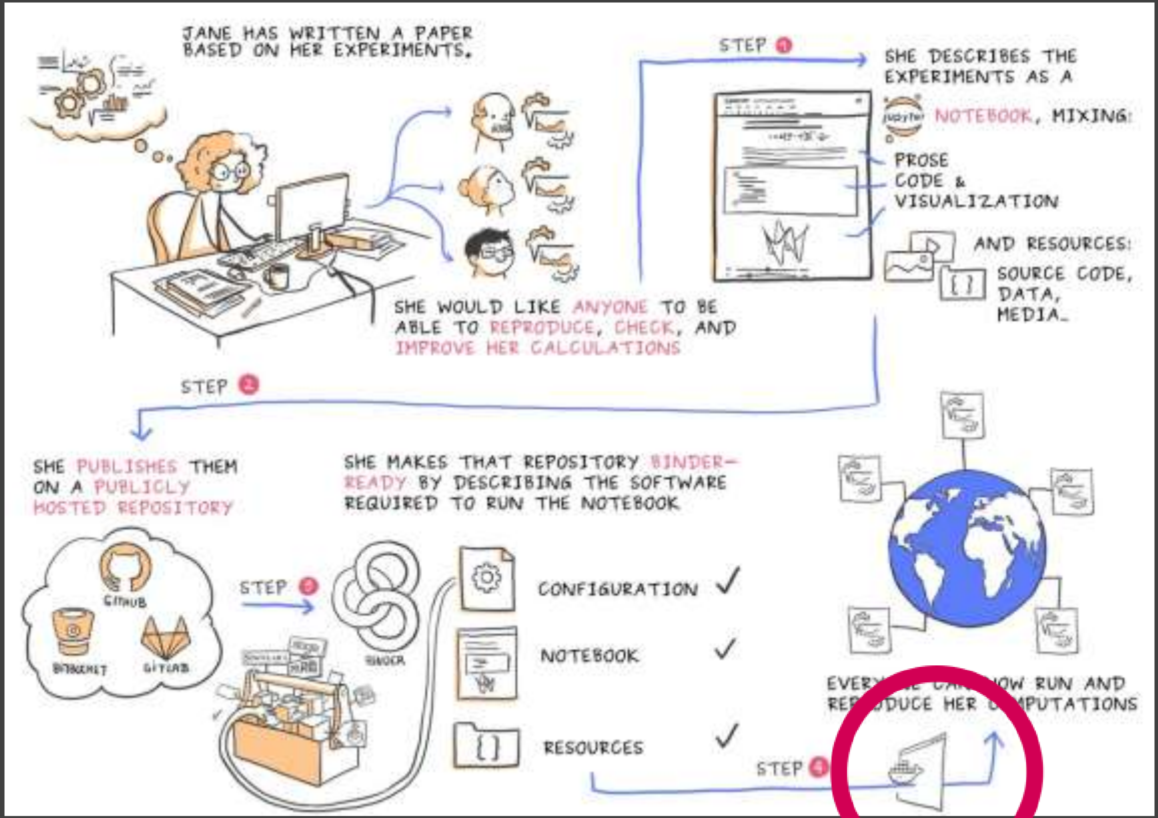
Turing Way & Binder


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

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3632909

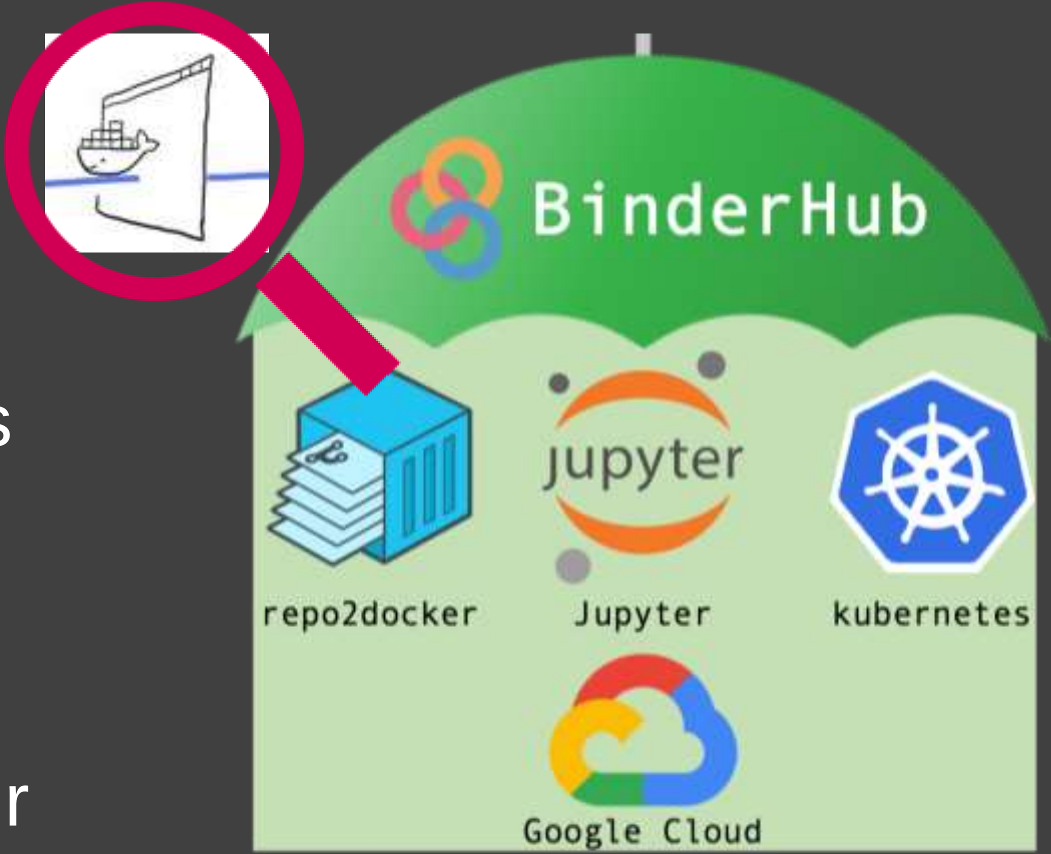

JANE HAS WRITTEN A PAPER BASED ON HER EXPERIMENTS.
SHE WOULD LIKE ANYONE TO BE ABLE TO REPRODUCE, CHECK, AND IMPROVE HER CALCULATIONS
STEP 1
SHE DESCRIBES THE EXPERIMENTS AS A
jupyter NOTEBOOK, MIXING:
PROSE
CODE &
VISUALIZATION
AND RESOURCES:
SOURCE CODE,
DATA,
MEDIA...
{ }
STEP 2
SHE PUBLISHES THEM 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

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3632909

- 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

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3632909

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.”

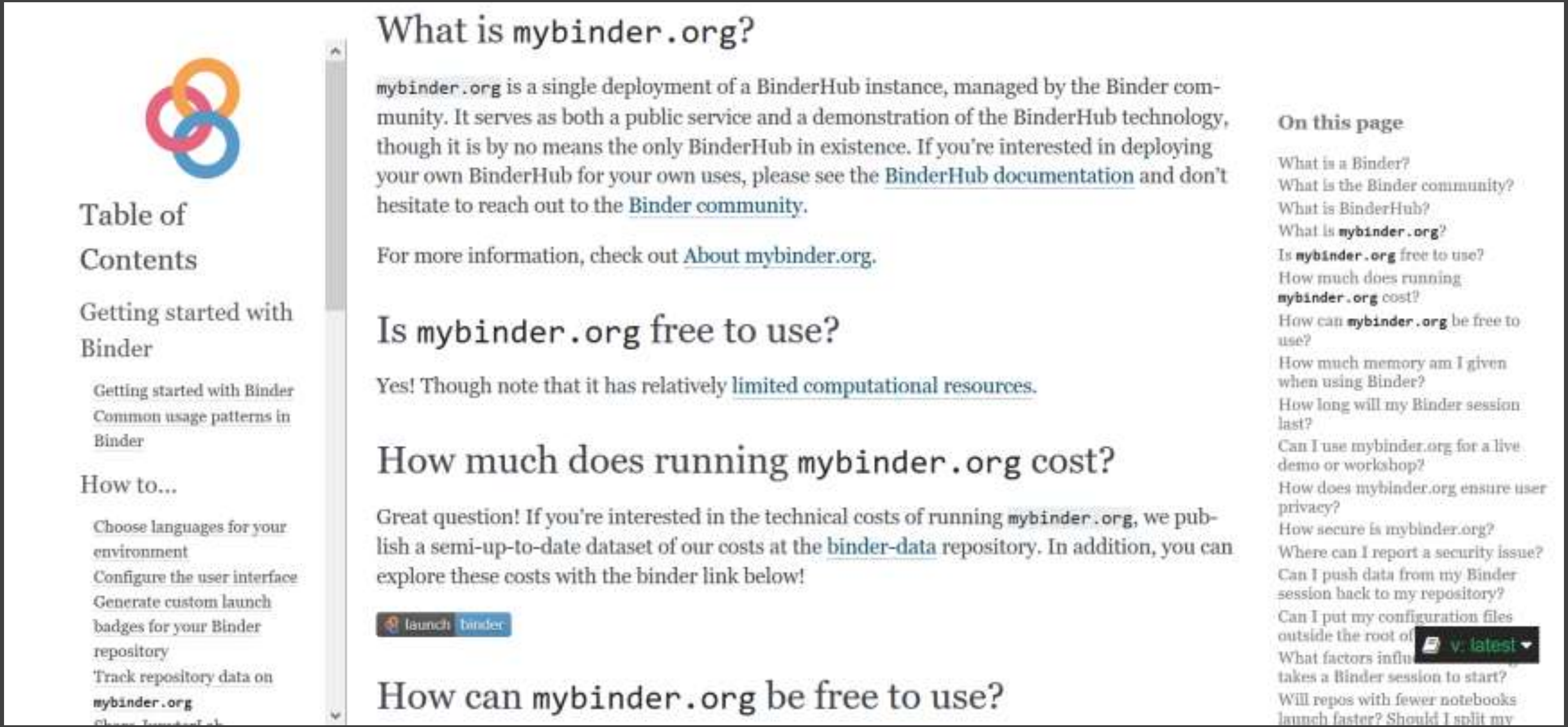
https://www.turing.ac.uk/people/researchers/sarah-gibson

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3632909

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

https://doi.org/10.5281/zenodo.3632909



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?

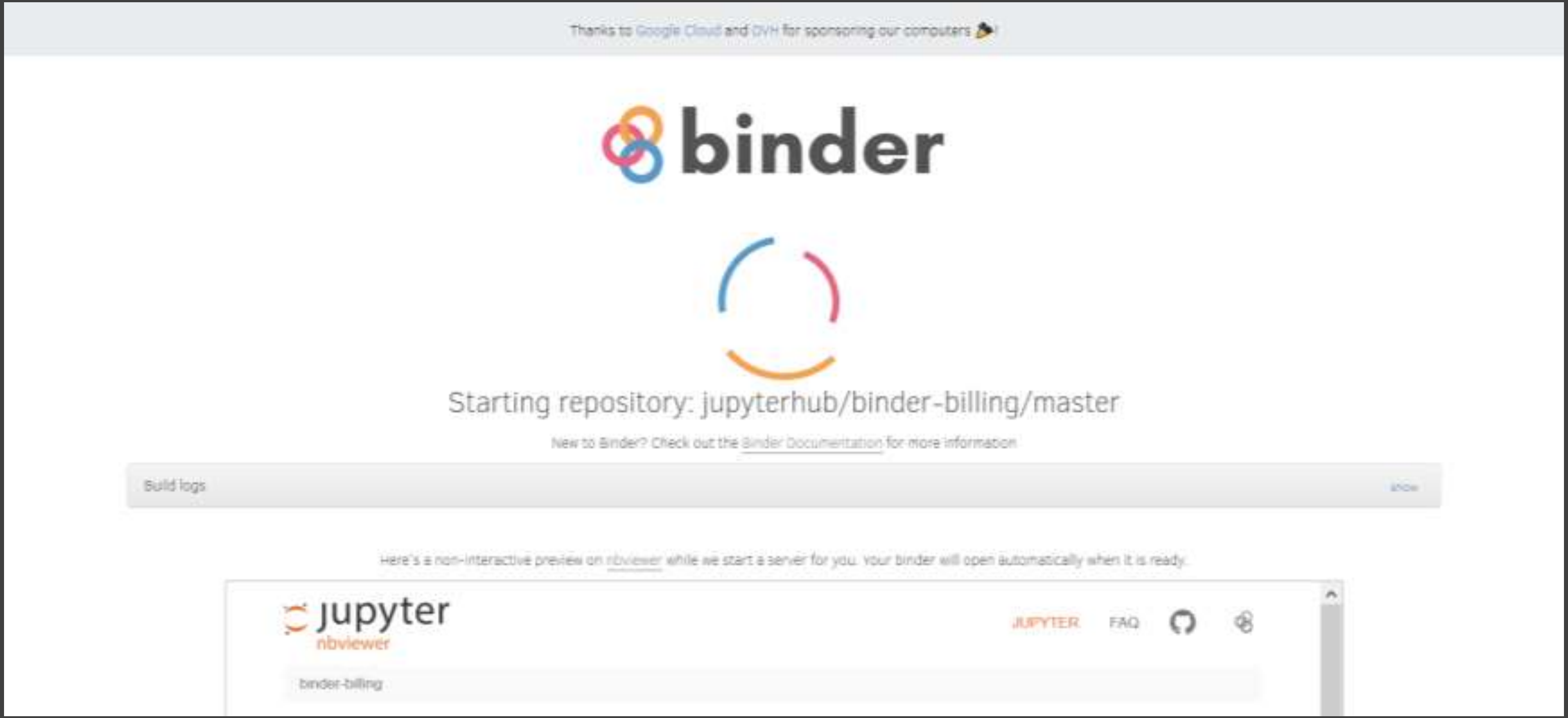
Great question! If you're interested in the technical costs of running `mybinder.org`, we publish a semi-up-to-date dataset of our costs at the binder-data repository. In addition, you can explore these costs with the binder link below!

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

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

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3632909



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

binder

Starting repository: jupyterhub/binder-billing/master

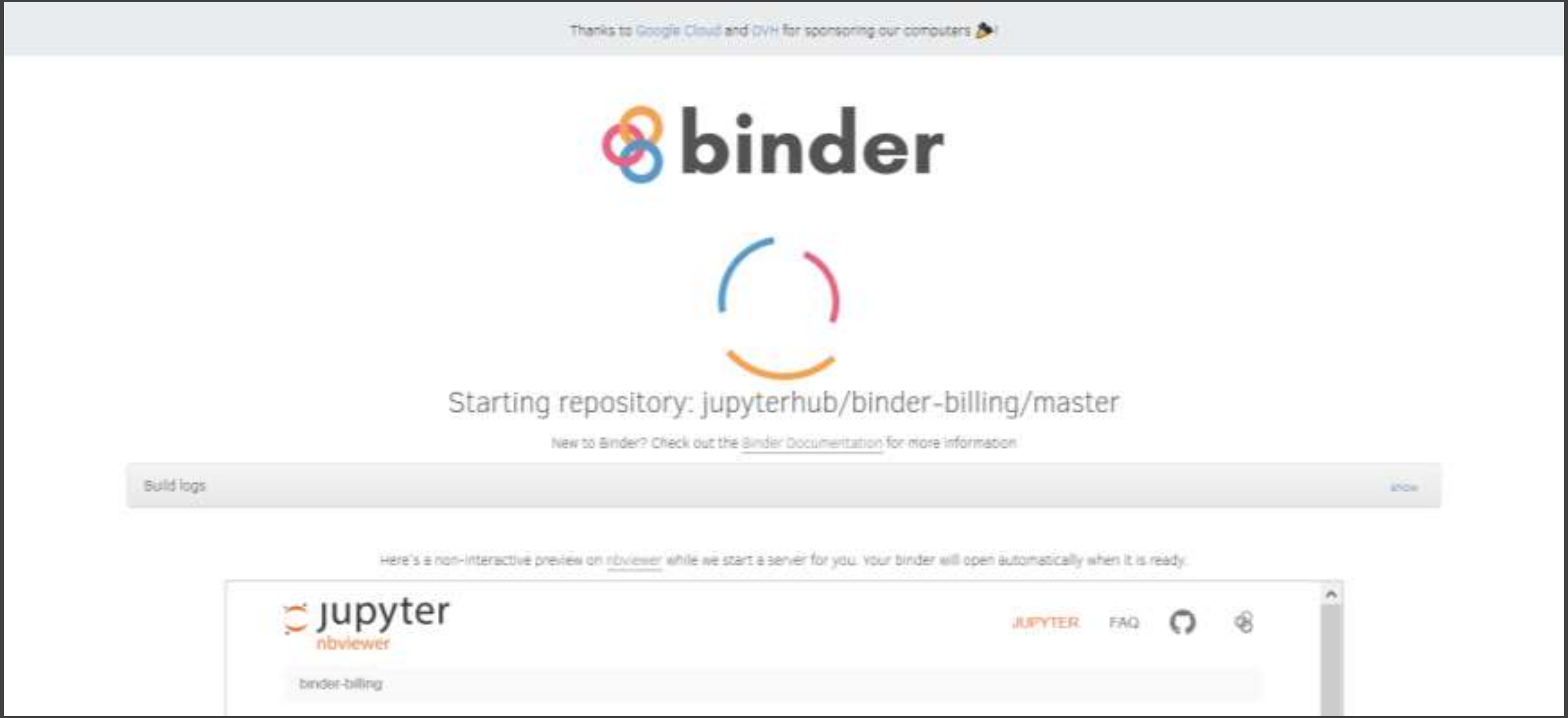
New to Binder? Check out the Binder Documentation for more information.

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



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

Starting repository: jupyterhub/binder-billing/master

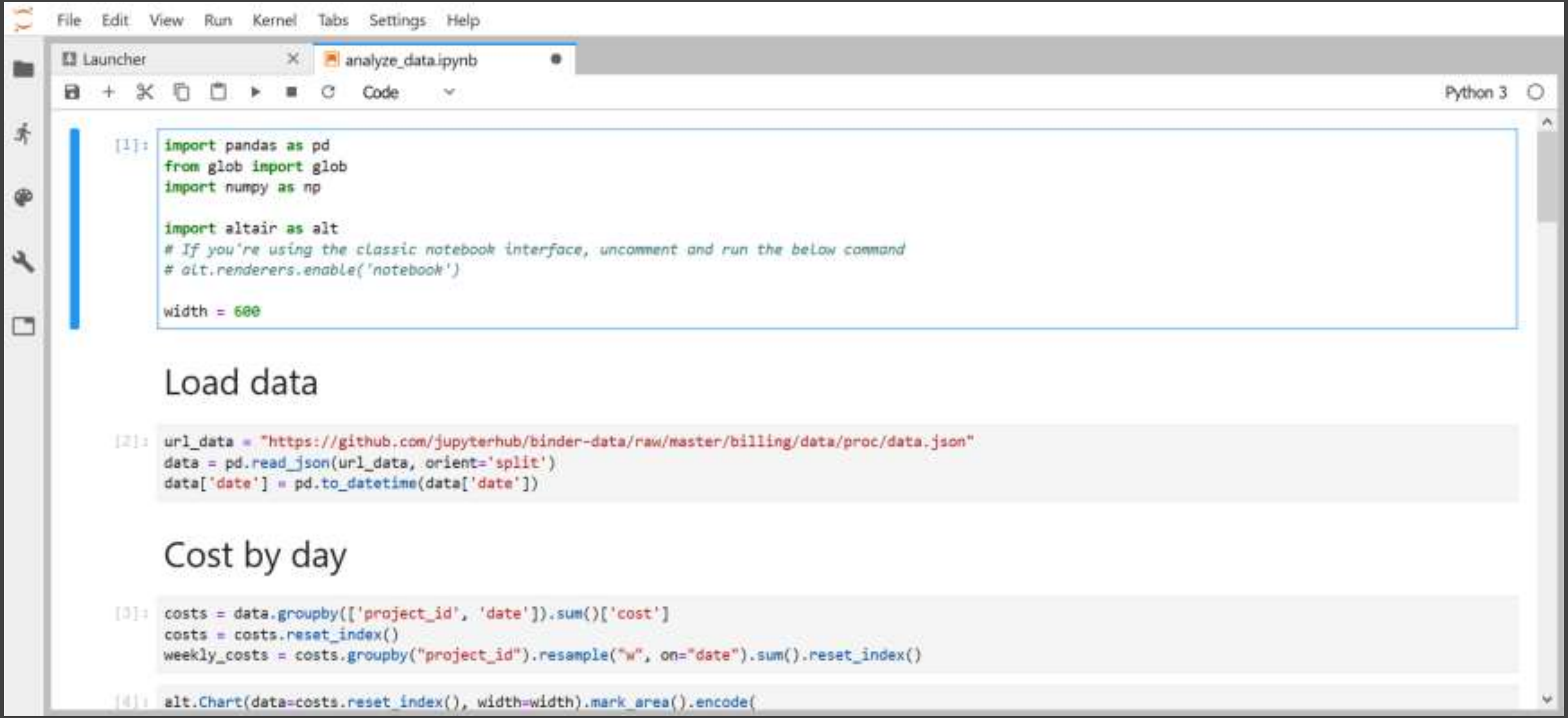
New to Binder? Check out the Binder Documentation for more information

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

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3632909



```
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

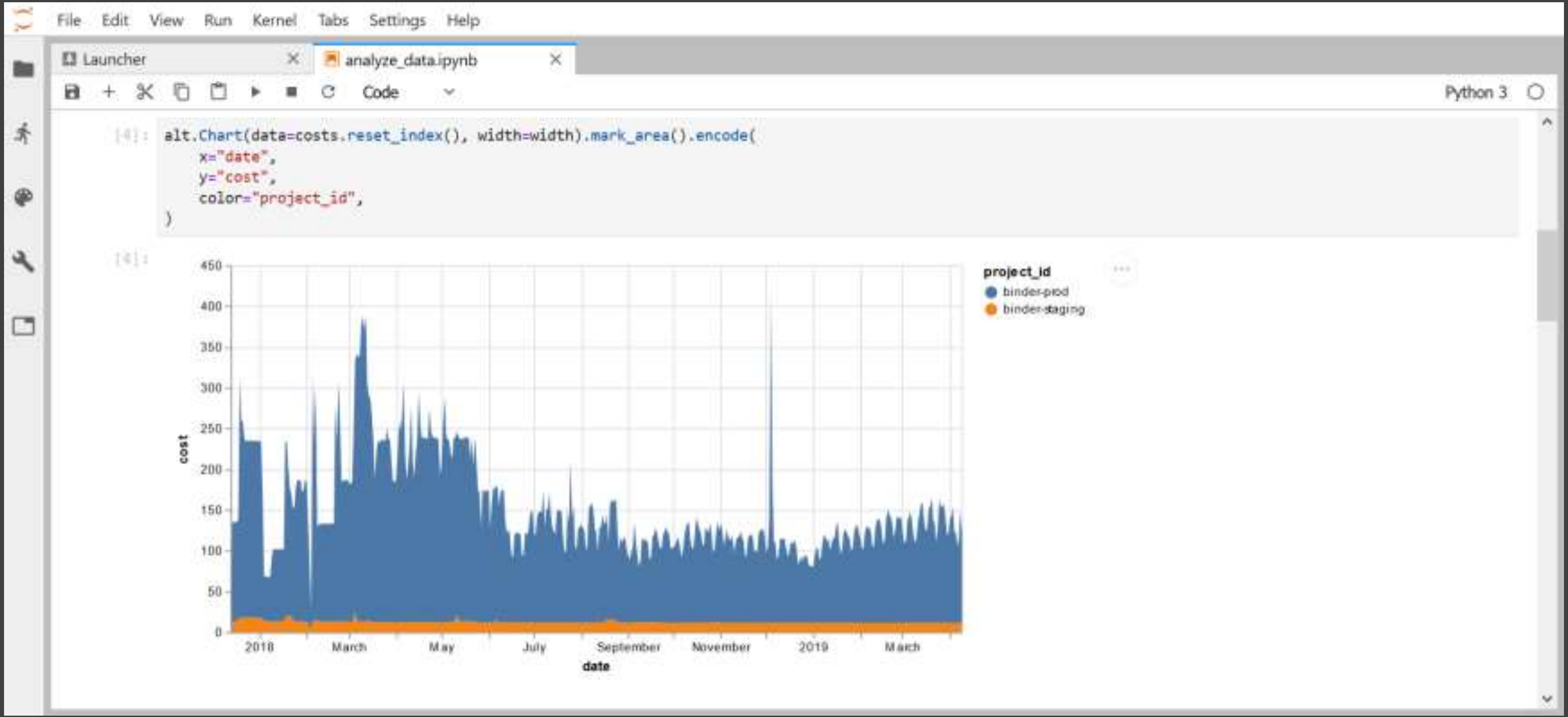
```
costs = data.groupby(['project_id', 'date']).sum()['cost']
costs = costs.reset_index()
weekly_costs = costs.groupby("project_id").resample("w", on="date").sum().reset_index()
```

```
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

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3632909

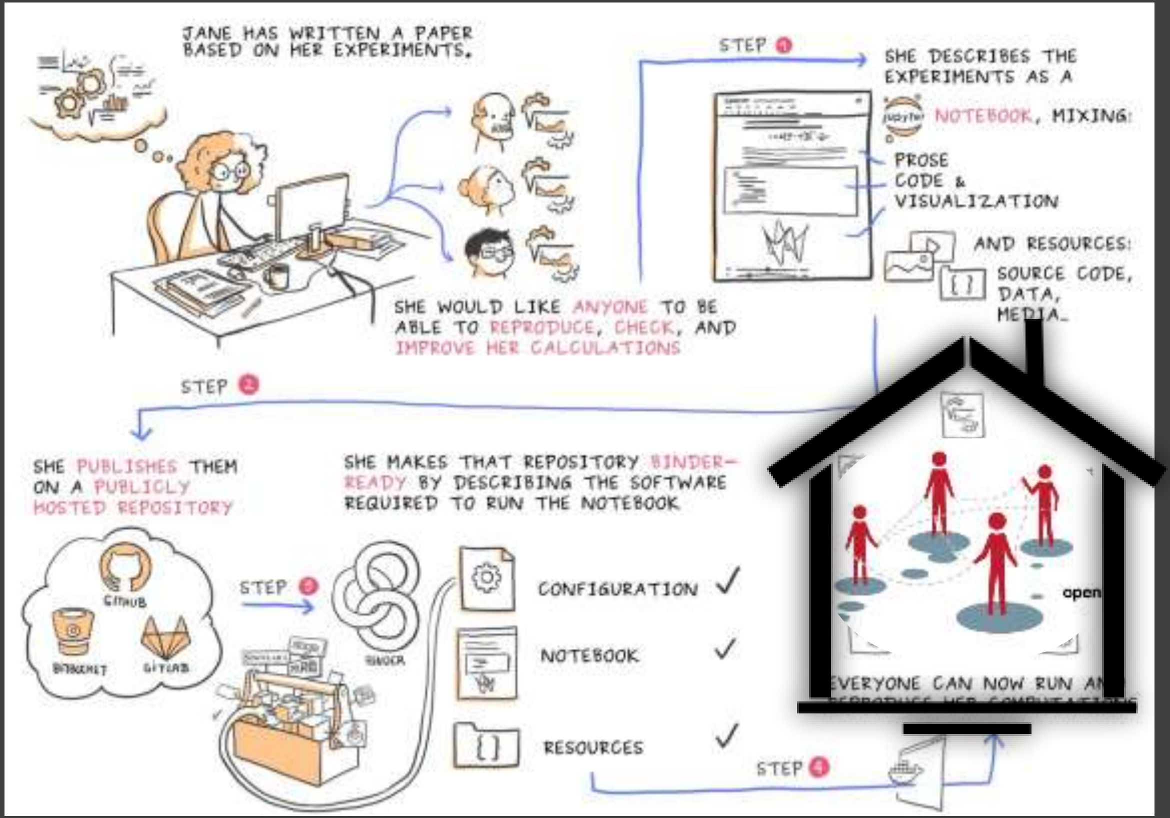


```
alt.Chart(data=costs.reset_index(), width=width).mark_area().encode(
    x="date",
    y="cost",
    color="project_id",
)
```

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

#PyDataLDN #TuringWay @kirstie_j

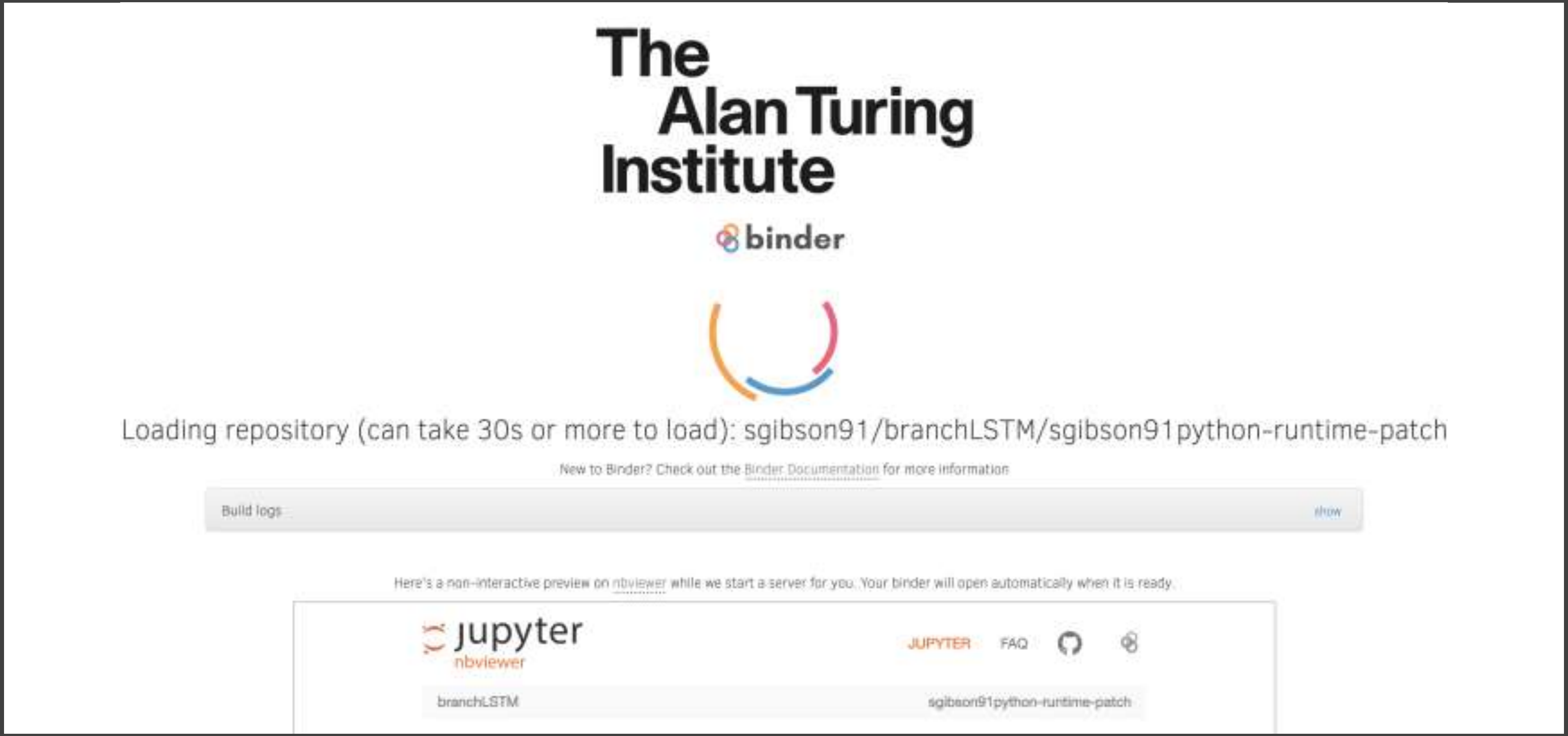
https://doi.org/10.5281/zenodo.3632909


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
SHE MAKES THAT REPOSITORY BINDER-READY BY DESCRIBING THE SOFTWARE REQUIRED TO RUN THE NOTEBOOK
STEP 3
BINDER
CONFIGURATION ✓
NOTEBOOK ✓
RESOURCES ✓
STEP 4
open
EVERYONE CAN NOW RUN AN
REPRODUCE HER COMPUTATIONS

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

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3632909



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

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.

branchLSTM

sgibson91python-runtime-patch

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

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3632909

Champion: Elena Kochkina

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

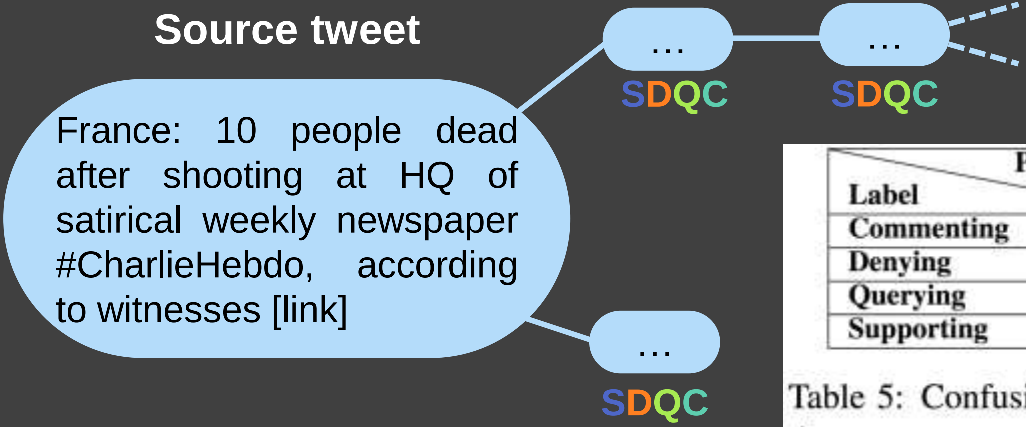
Elena Kochkina, Maria Liakata, Isabelle Augenstein

Support

Deny

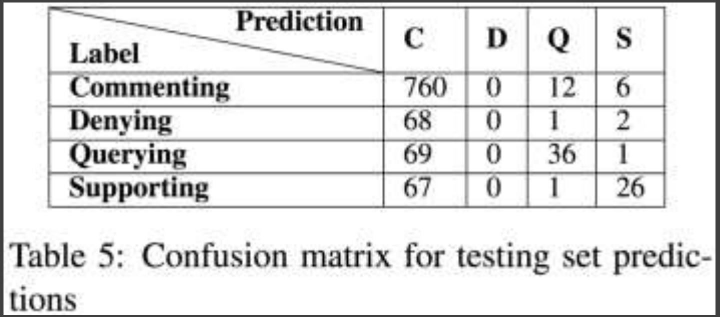
Query

Comment



Source tweet

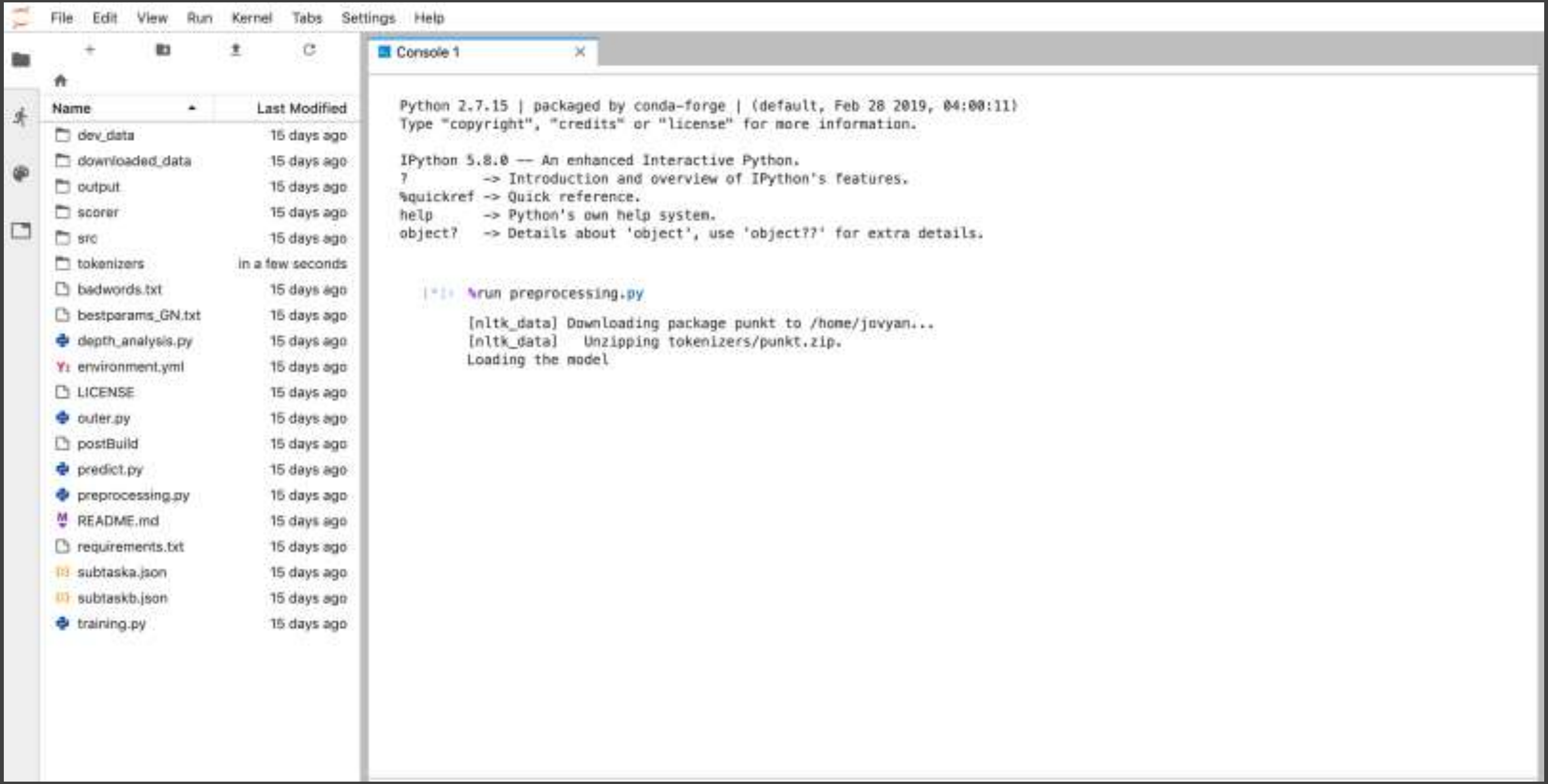
France: 10 people dead after shooting at HQ of satirical weekly newspaper #CharlieHebdo, according to witnesses [link]



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



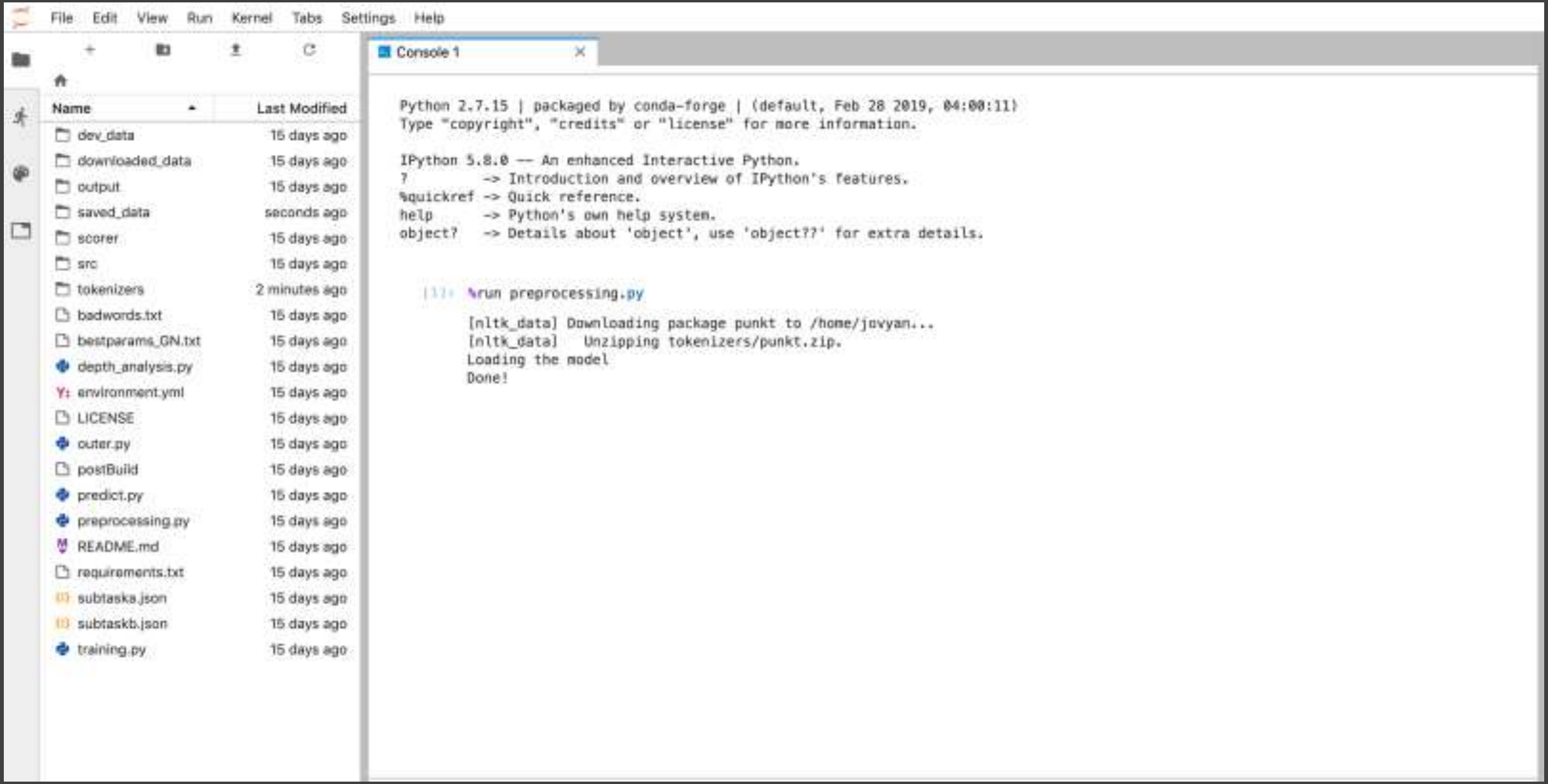
Name	Last Modified
dev_data	15 days ago
downloaded_data	15 days ago
output	15 days ago
scorer	15 days ago
src	15 days ago
tokenizers	in a few seconds
badwords.txt	15 days ago
bestparams_GN.txt	15 days ago
depth_analysis.py	15 days ago
environment.yml	15 days ago
LICENSE	15 days ago
outer.py	15 days ago
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.3632909



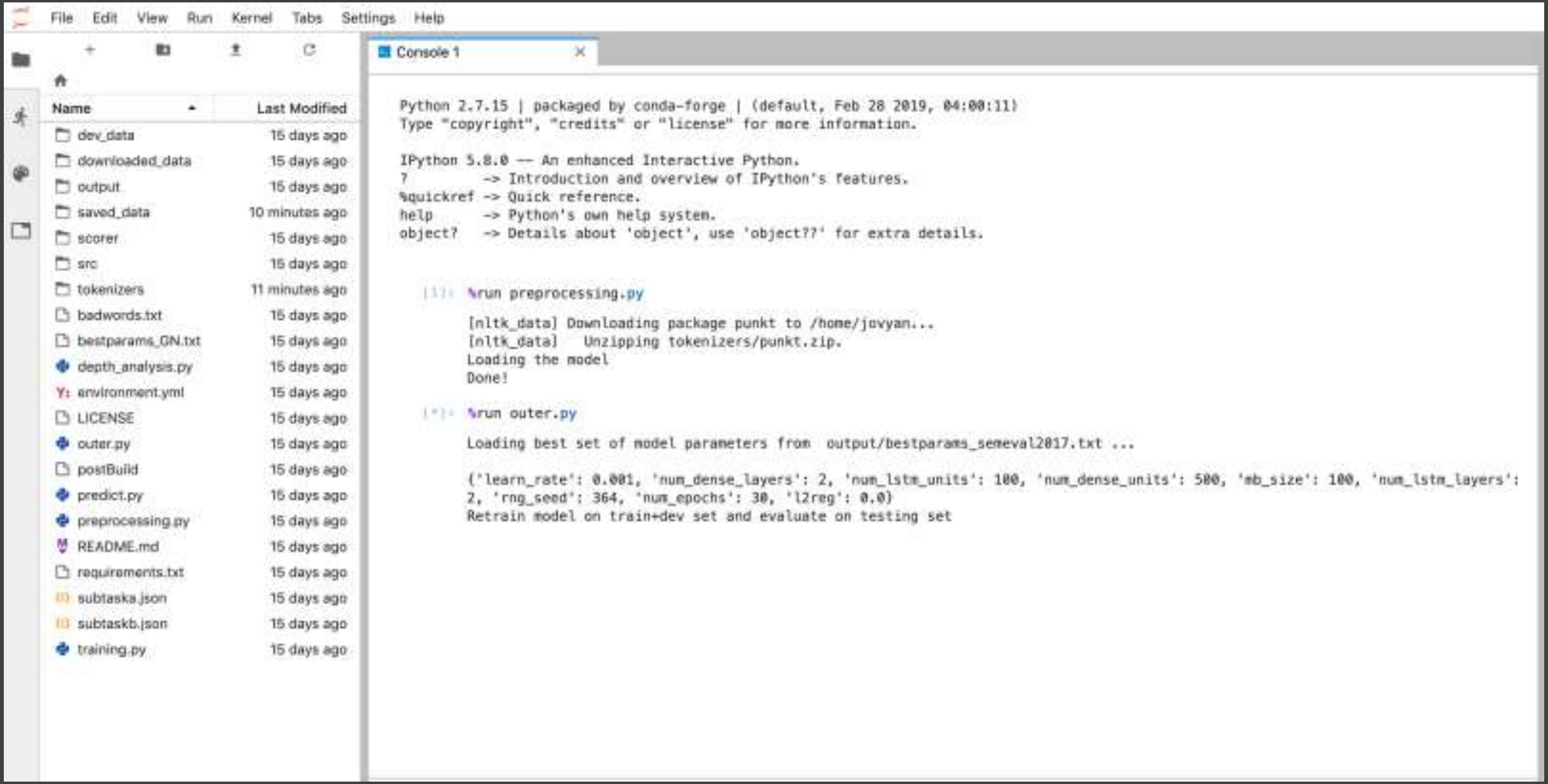
Name	Last Modified
dev_data	15 days ago
downloaded_data	15 days ago
output	15 days ago
saved_data	seconds ago
scorer	15 days ago
src	15 days ago
tokenizers	2 minutes ago
badwords.txt	15 days ago
bestparams_GN.txt	15 days ago
depth_analysis.py	15 days ago
environment.yml	15 days ago
LICENSE	15 days ago
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.3632909



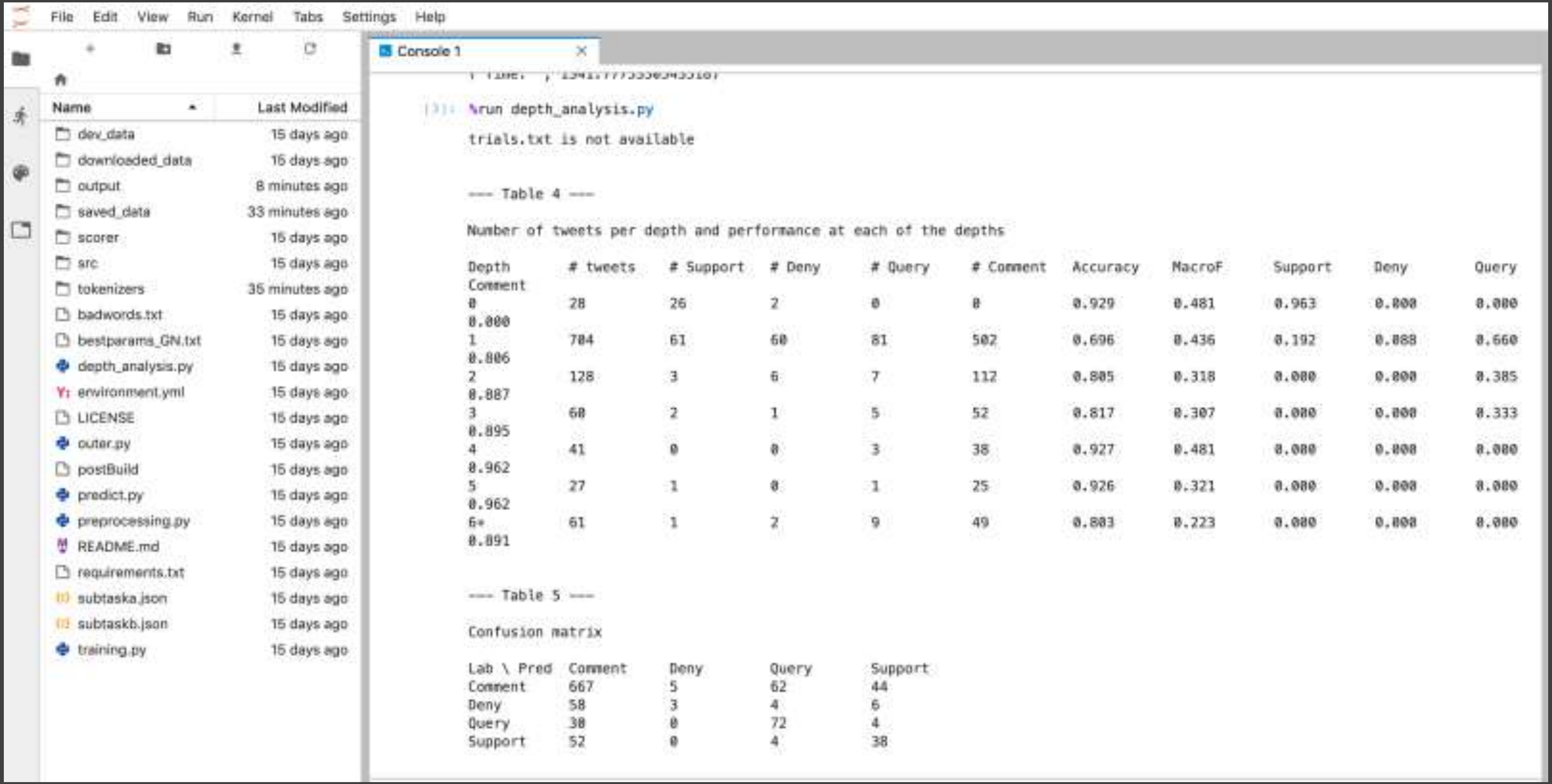
Name	Last Modified
dev_data	15 days ago
downloaded_data	15 days ago
output	15 days ago
saved_data	10 minutes ago
scorer	15 days ago
src	15 days ago
tokenizers	11 minutes ago
badwords.txt	15 days ago
bestparams_GN.txt	15 days ago
depth_analysis.py	15 days ago
environment.yml	15 days ago
LICENSE	15 days 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.3632909



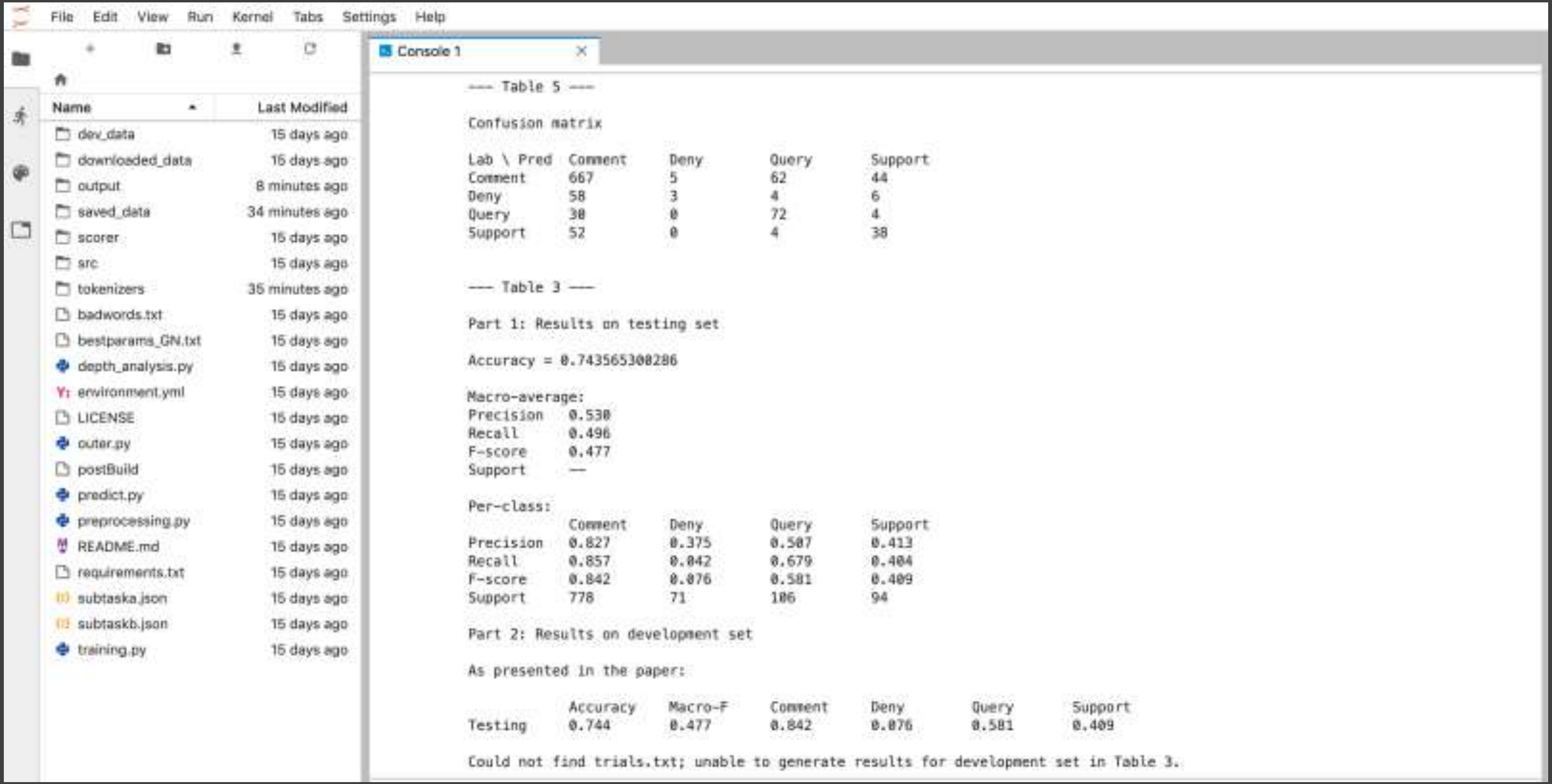
Name	Last Modified
dev_data	15 days ago
downloaded_data	15 days ago
output	8 minutes ago
saved_data	33 minutes ago
scorer	15 days ago
src	15 days ago
tokenizers	35 minutes ago
badwords.txt	15 days ago
bestparams_GN.txt	15 days ago
depth_analysis.py	15 days ago
environment.yml	15 days ago
LICENSE	15 days 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

```
%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.3632909



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.3632909

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.3632909

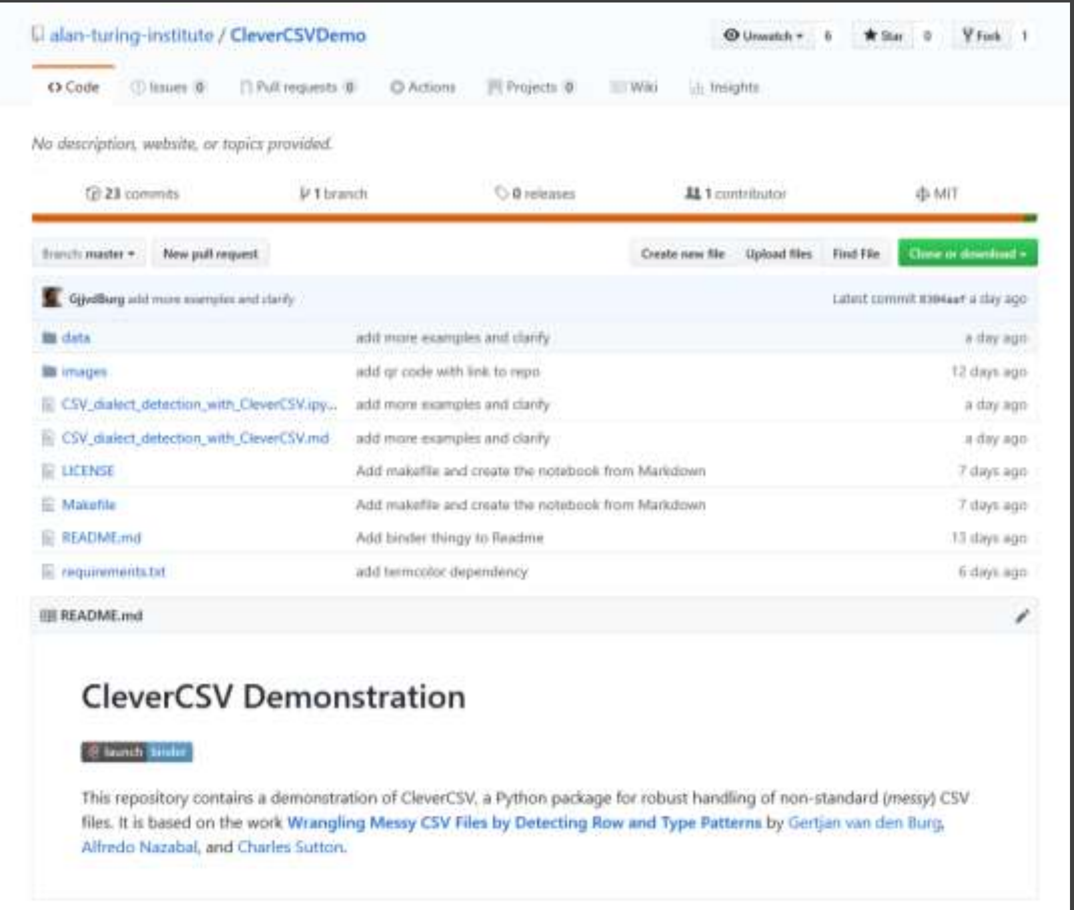
Gertjan van den Burg

“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

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3632909



No description, website, or topics provided.

GjjvdBurg add more examples and clarify — Latest commit 8394aaf 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

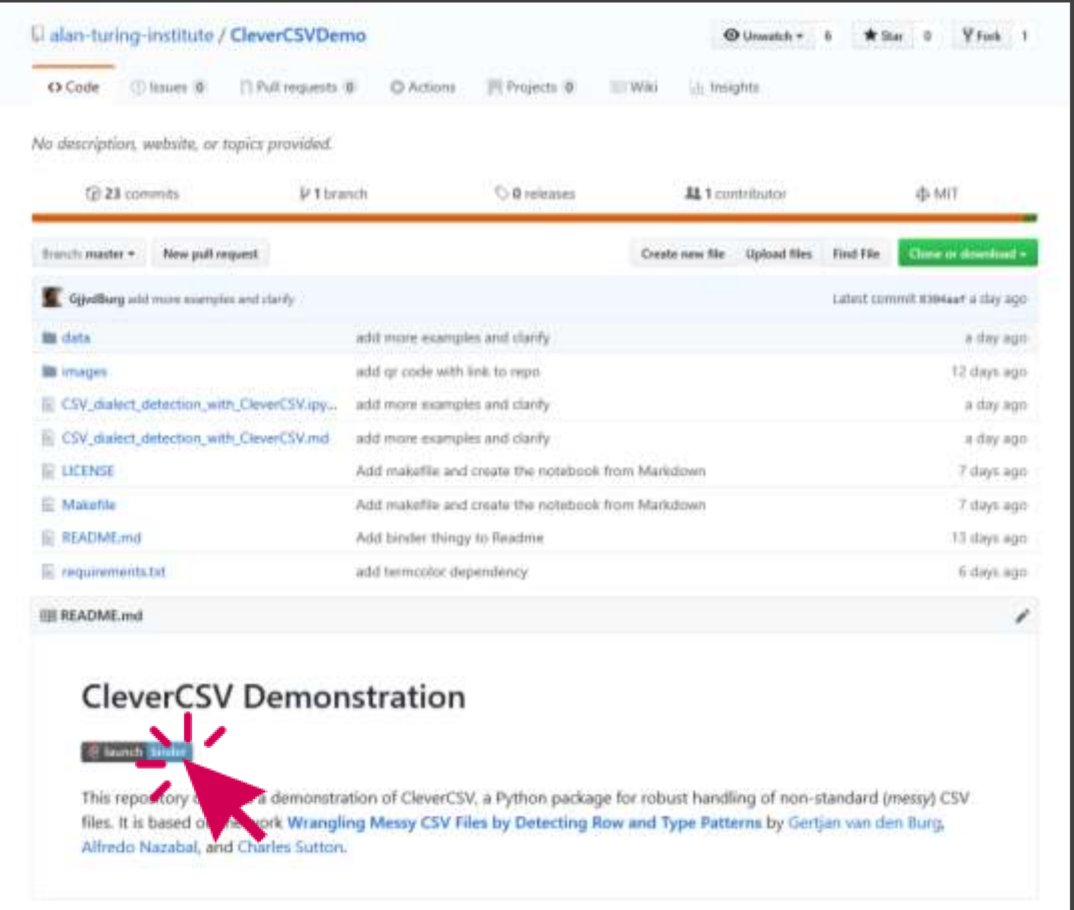
README.md

CleverCSV Demonstration

launch binder

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 8384aaf a day ago

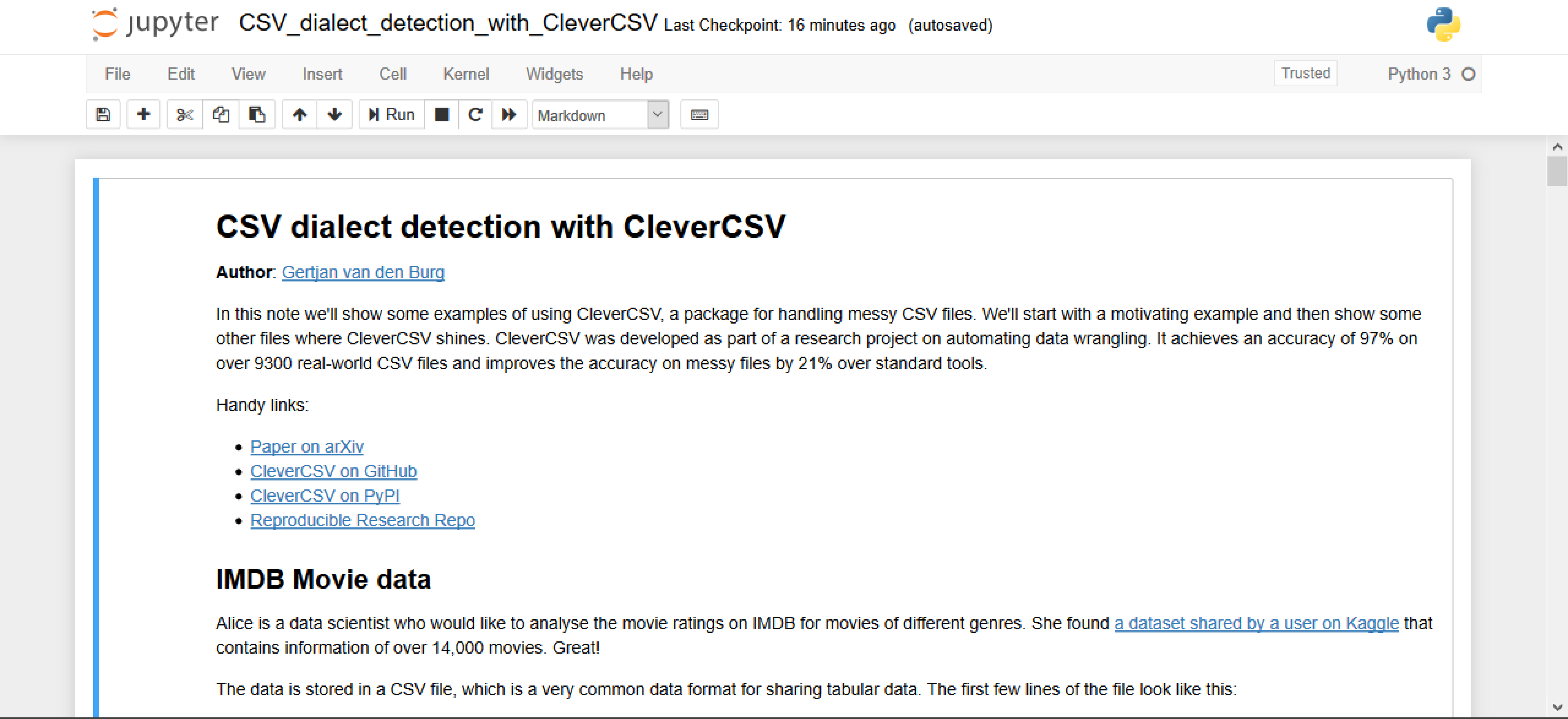
data	add more examples and clarify	a day ago
images	add qr code with link to repo	12 days ago
CSV_dialect_detection_with_CleverCSV.ipy...	add more examples and clarify	a day ago
CSV_dialect_detection_with_CleverCSV.md	add more examples and clarify	a day ago
LICENSE	Add makefile and create the notebook 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

CleverCSV Demonstration

This repository [illegible] a demonstration of CleverCSV, a Python package for robust handling of non-standard (*messy*) CSV files. It is based o[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


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:

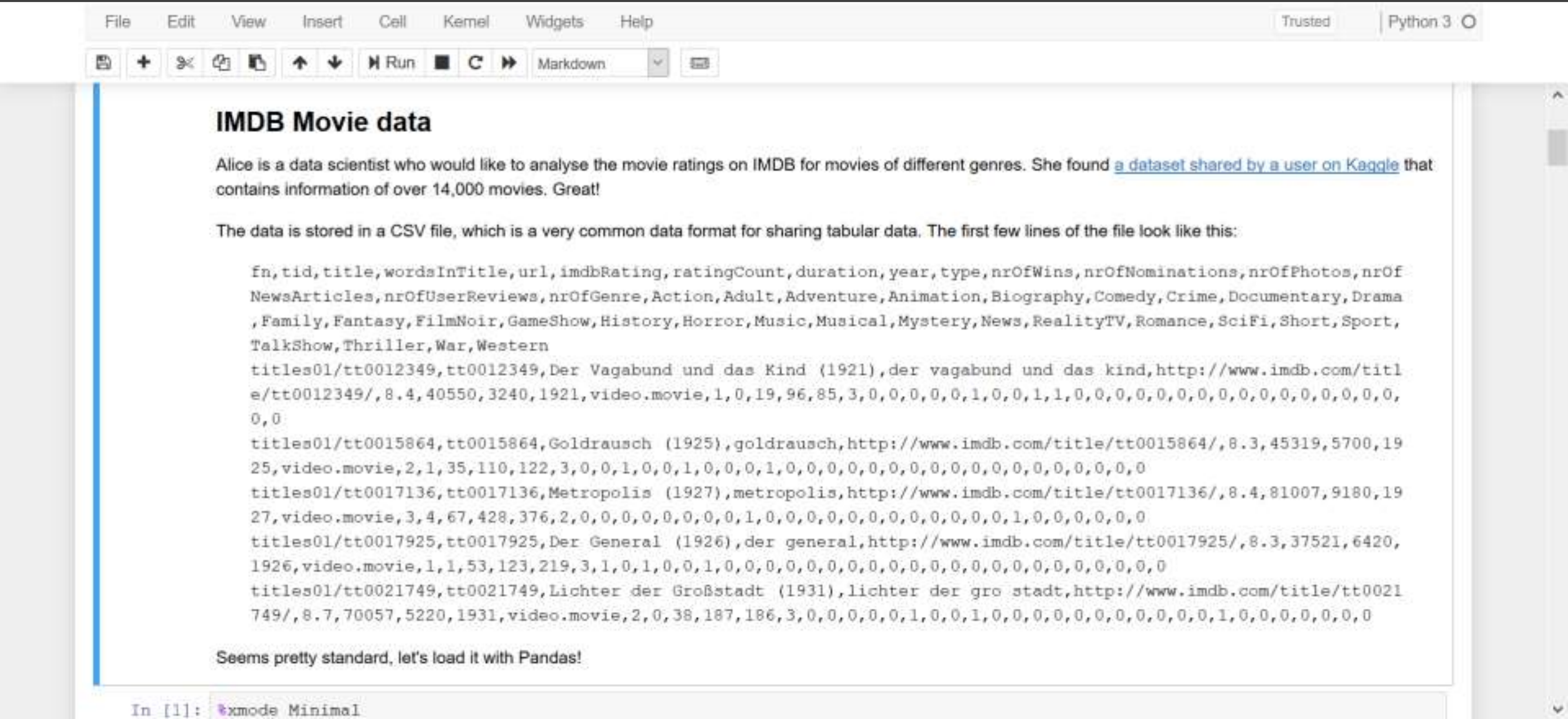
- Paper on arXiv
- CleverCSV on GitHub
- CleverCSV on PyPI
- Reproducible Research Repo

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
#PyDataLDN #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3632909



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,
0,0
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
titles01/tt0017136,tt0017136,Metropolis (1927),metropolis,http://www.imdb.com/title/tt0017136/,8.4,81007,9180,19
27,video.movie,3,4,67,428,376,2,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0
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
```

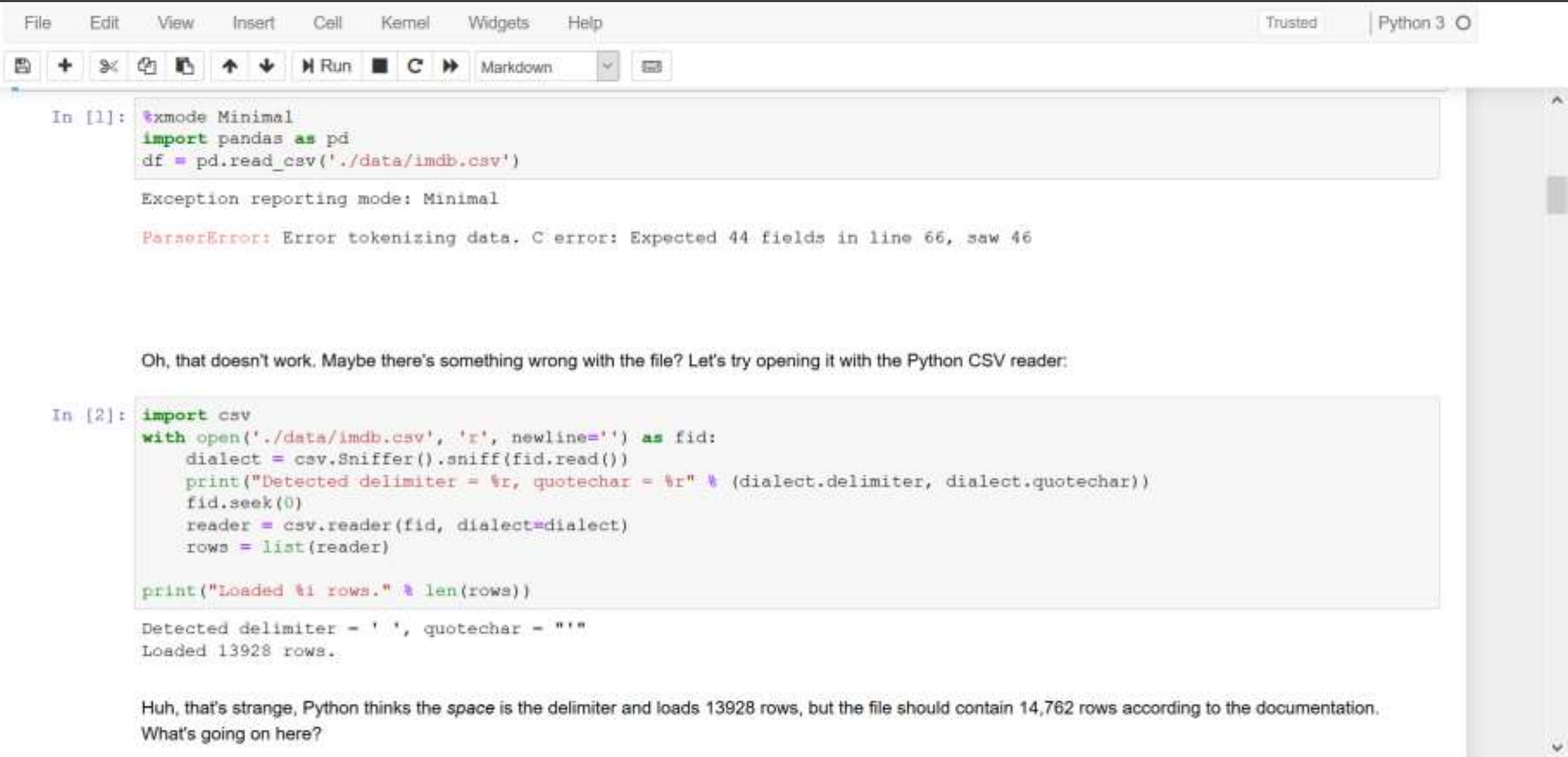
Seems pretty standard, let's load it with Pandas!

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

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

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3632909



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

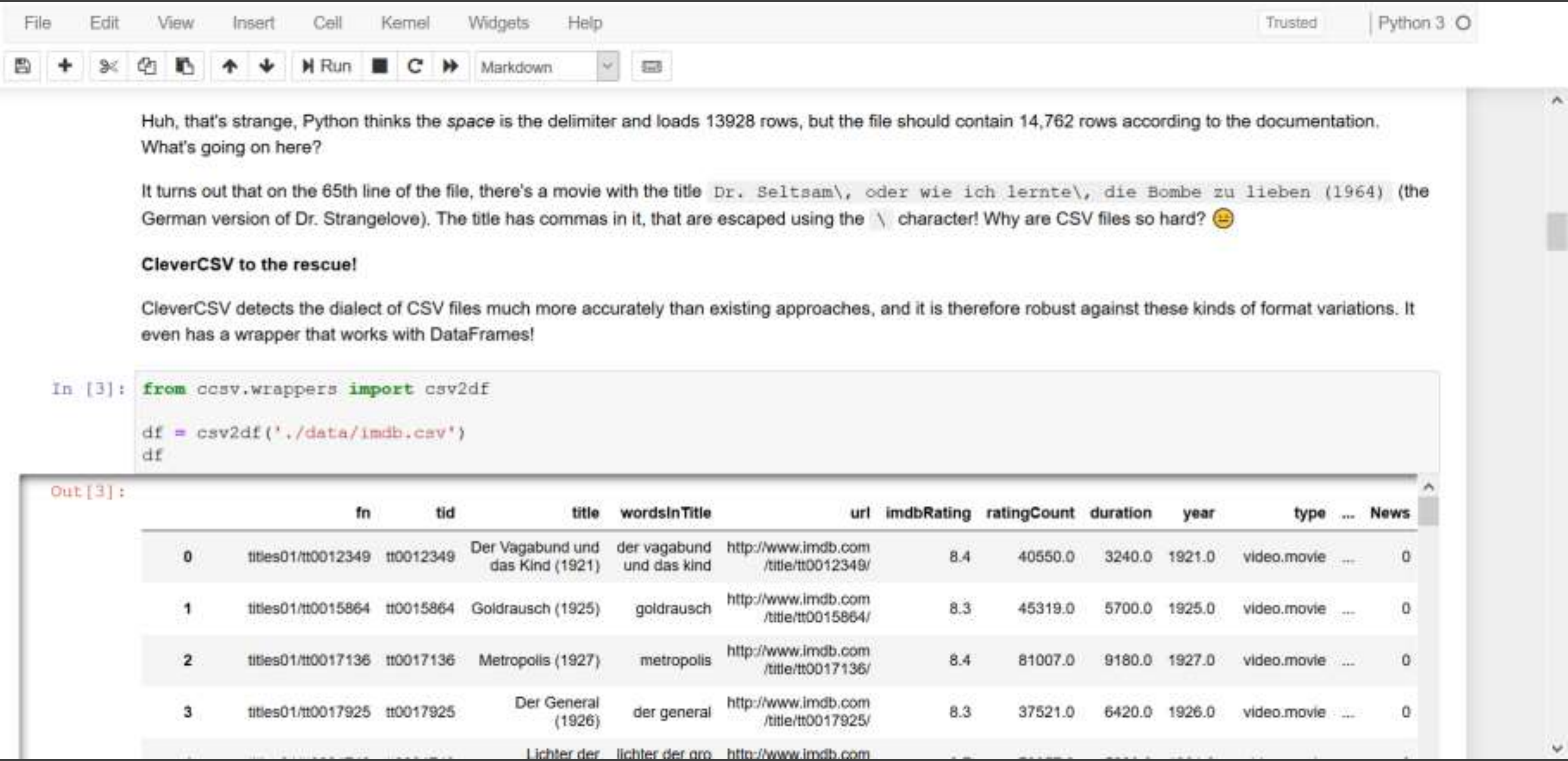
```
Exception reporting mode: Minimal

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?



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!

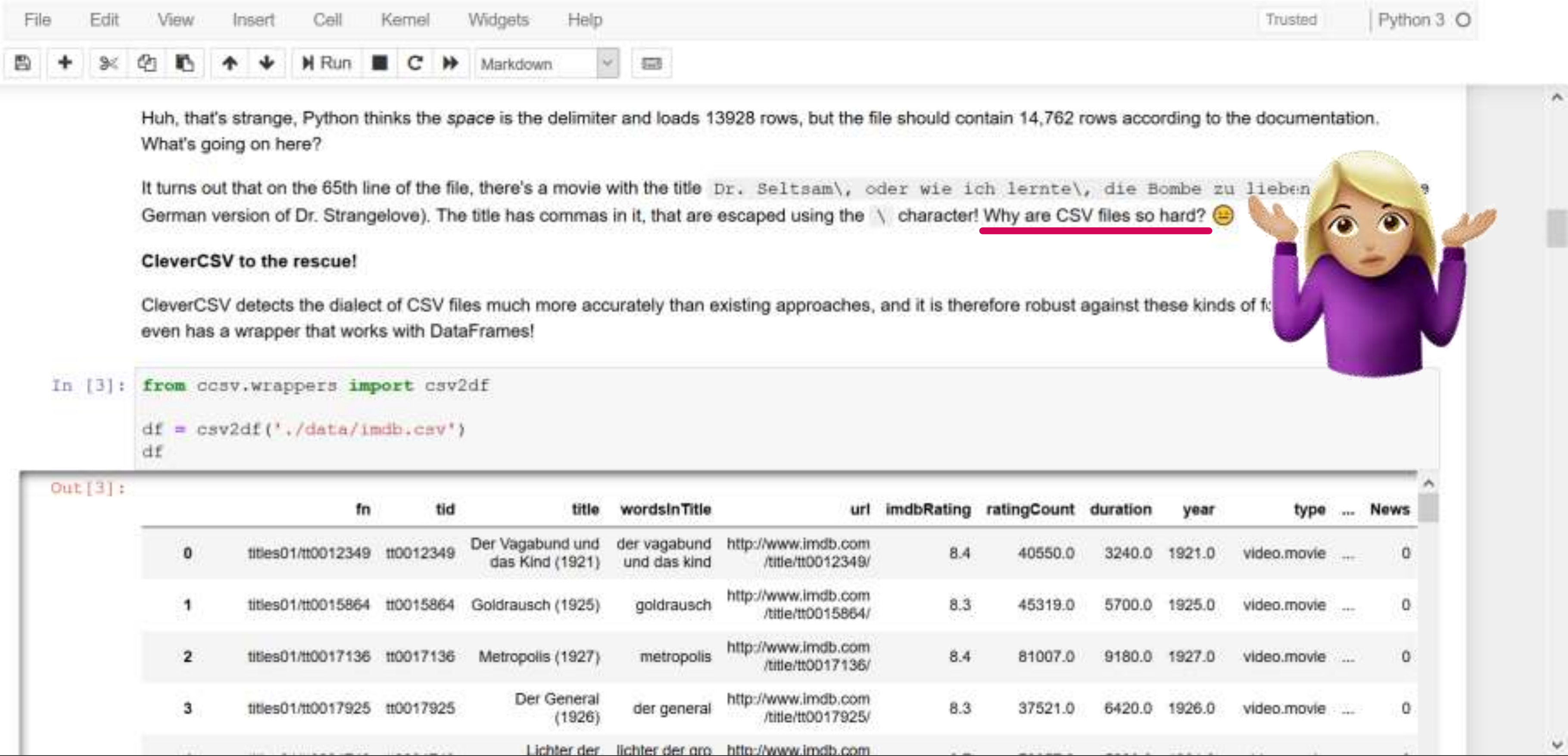
In [3]:

```
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
#PyDataLDN #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3632909



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!

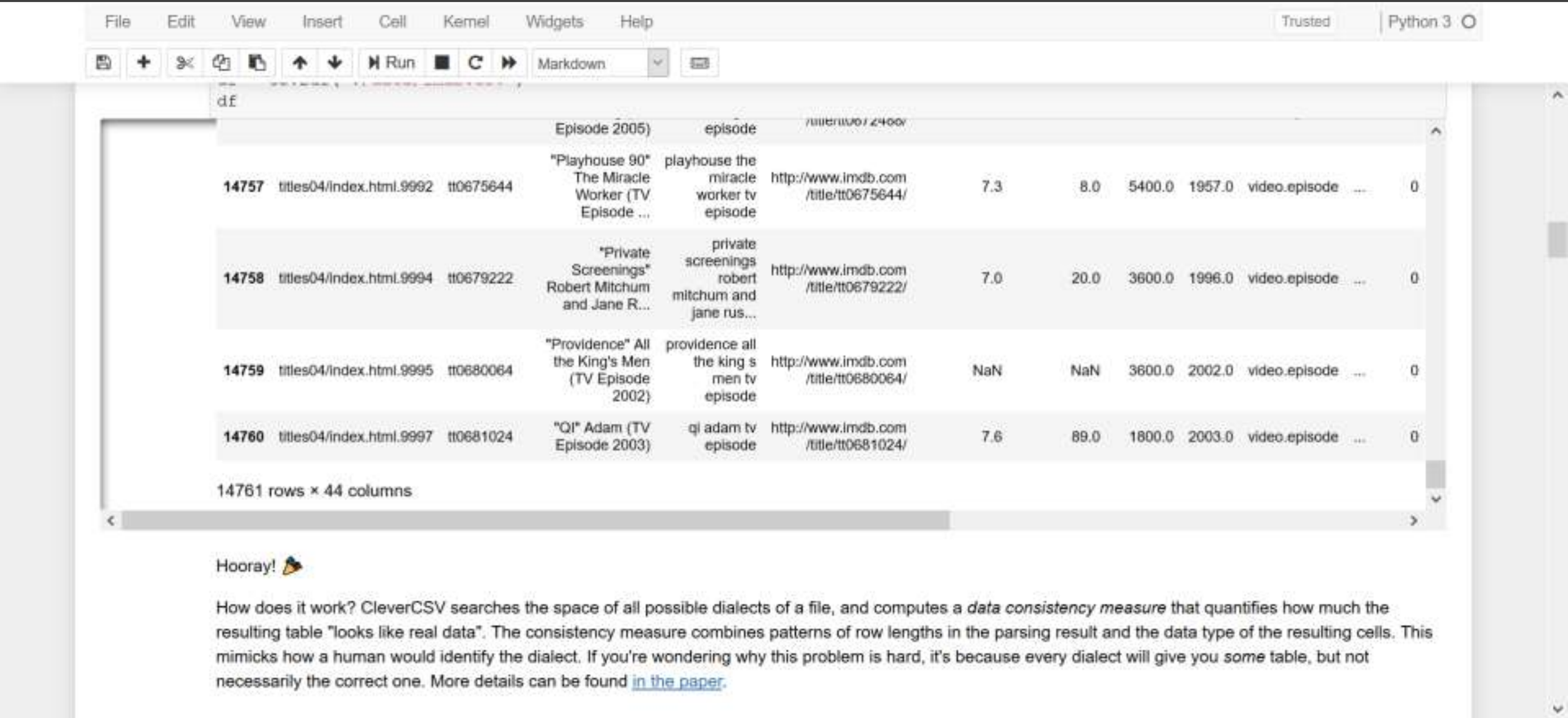
CleverCSV detects the dialect of CSV files much more accurately than existing approaches, and it is therefore robust against these kinds of f even has a wrapper that works with DataFrames!

```
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
#PyDataLDN #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3632909



```
df
```

			Episode 2005)	episode	/title/tt0672468/							
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.

Gertjan van den Burg

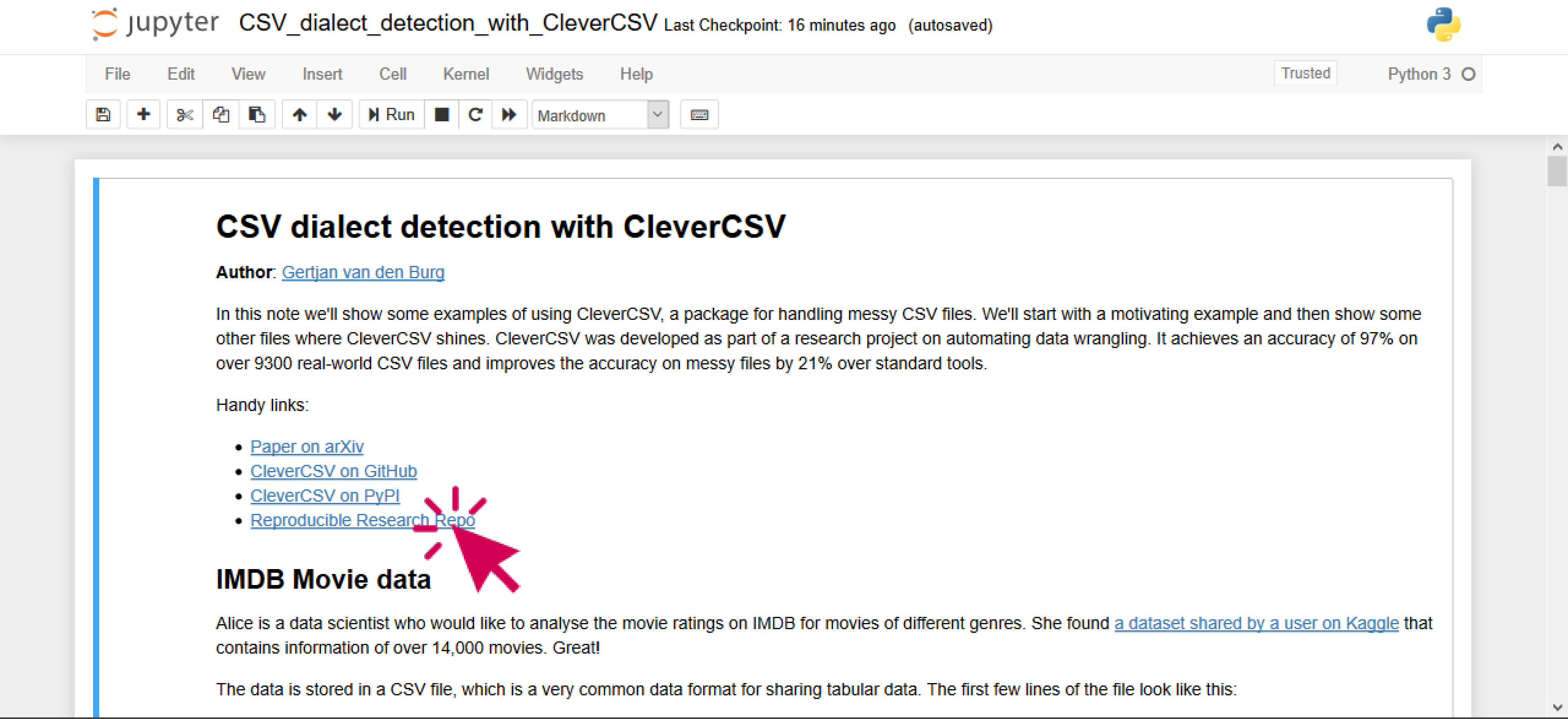
“The fun part of data science is the modelling. Being able to read in information from a csv file should not be the hardest part.

There is no AI. I am the AI.”

https://gertjanvandenburg.com

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3632909


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:

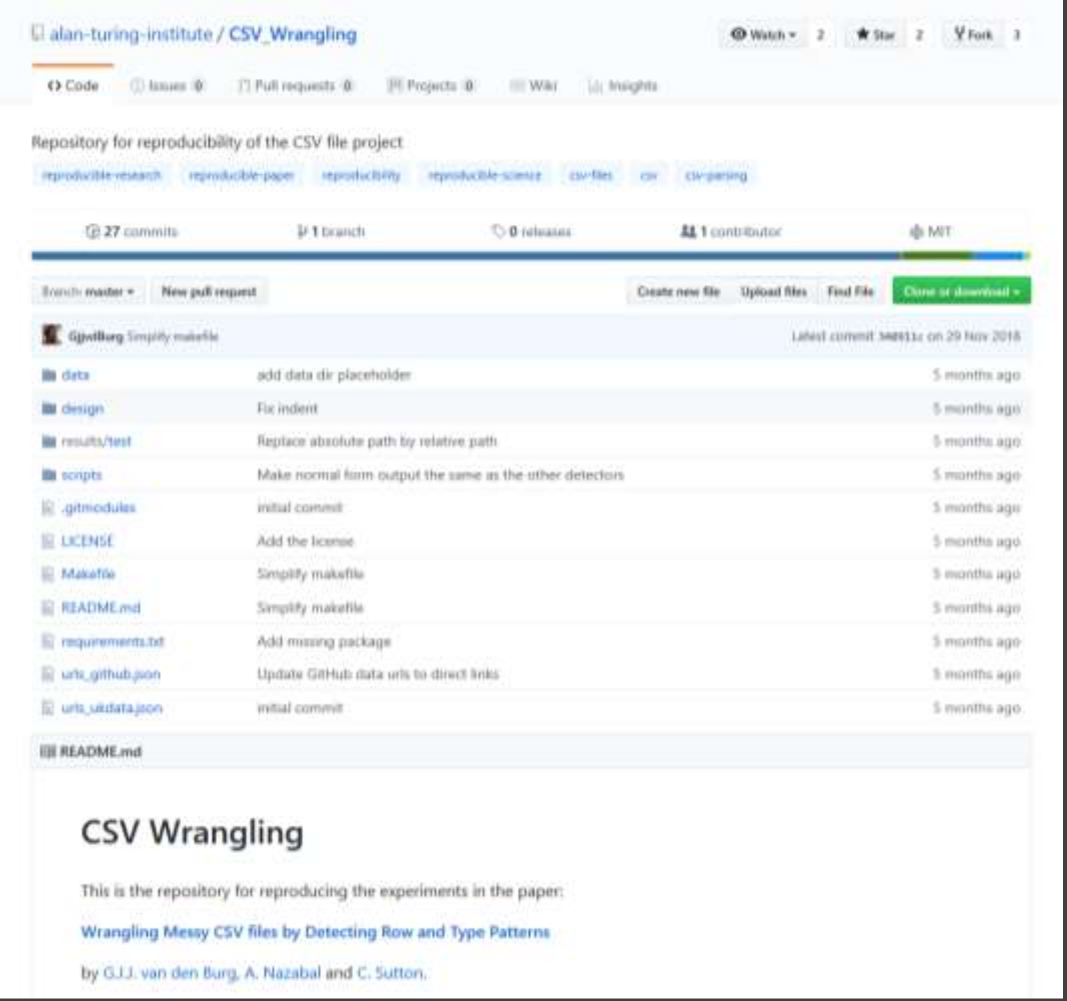
- Paper on arXiv
- CleverCSV on GitHub
- CleverCSV on PyPI
- Reproducible Research Repo

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
#PyDataLDN #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3632909



Repository for reproducibility of the CSV file project

reproducible-research reproducible-paper reproducibility reproducible-science csv-files csv csv-parsing

GjjvdBurg Simplify makefile Latest commit [illegible] on 29 Nov 2018

data	add data dir placeholder	5 months ago
design	Fix indent	5 months ago
results/test	Replace absolute path by relative path	5 months ago
scripts	Make normal form output the same as the other detectors	5 months ago
.gitmodules	initial commit	5 months ago
LICENSE	Add the license	5 months ago
Makefile	Simplify makefile	5 months ago
README.md	Simplify makefile	5 months ago
requirements.txt	Add missing package	5 months ago
urls_github.json	Update GitHub data urls to direct links	5 months ago
urls_ukdata.json	initial commit	5 months ago

README.md

CSV Wrangling

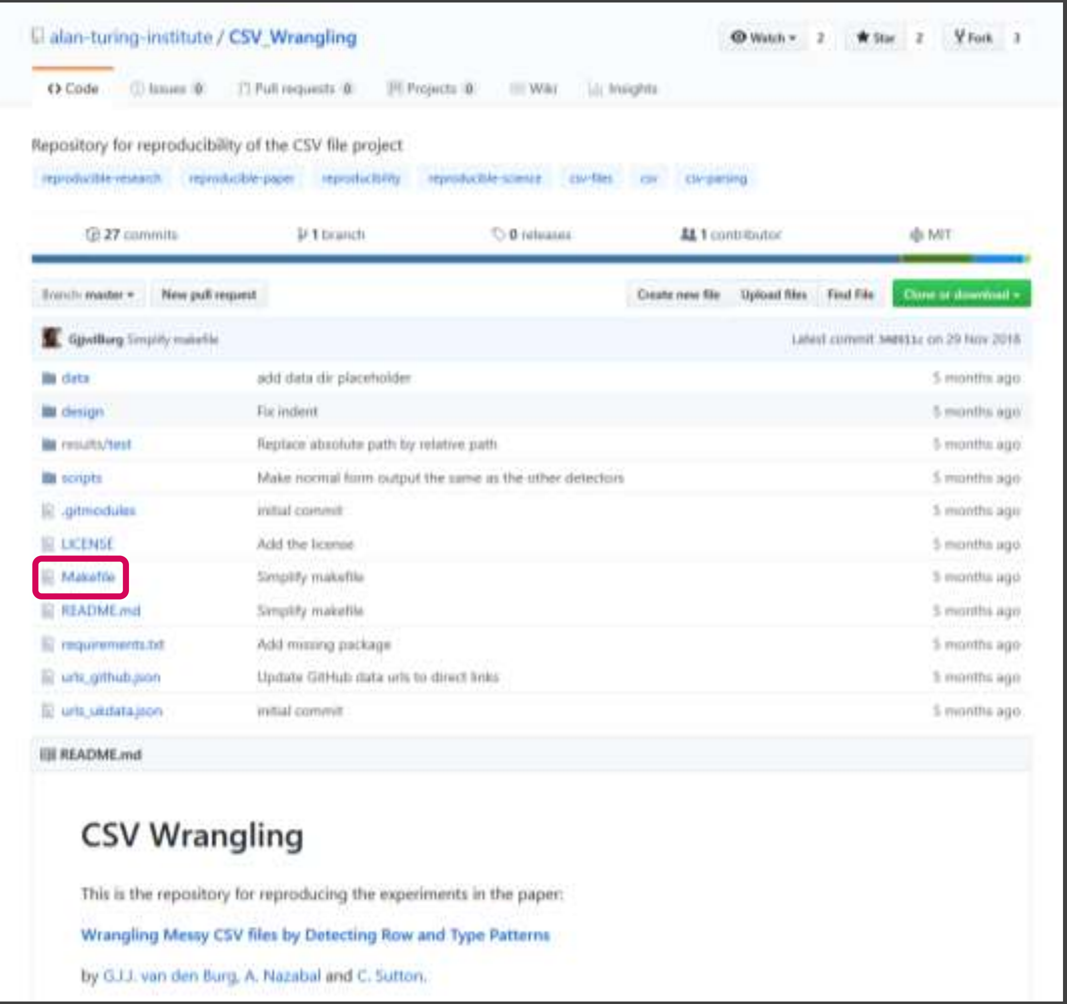
This is the repository for reproducing the experiments in the paper:

Wrangling Messy CSV files by Detecting Row and Type Patterns

by G.J.J. van den Burg, A. Nazabal and C. Sutton.

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

– “Wrangling Messy CSV Files by Detecting Row and Type Patterns”
arXiv:1811.11242



Repository for reproducibility of the CSV file project

GjjvdBurg Simplify makefile — Latest commit 34811c on 29 Nov 2018

data	add data dir placeholder	5 months ago
design	Fix indent	5 months ago
results/test	Replace absolute path by relative path	5 months ago
scripts	Make normal form output the same as the other detectors	5 months ago
.gitmodules	initial commit	5 months ago
LICENSE	Add the license	5 months ago
Makefile	Simplify makefile	5 months ago
README.md	Simplify makefile	5 months ago
requirements.txt	Add missing package	5 months ago
urls_github.json	Update GitHub data urls to direct links	5 months ago
urls_ukdata.json	initial commit	5 months ago

README.md

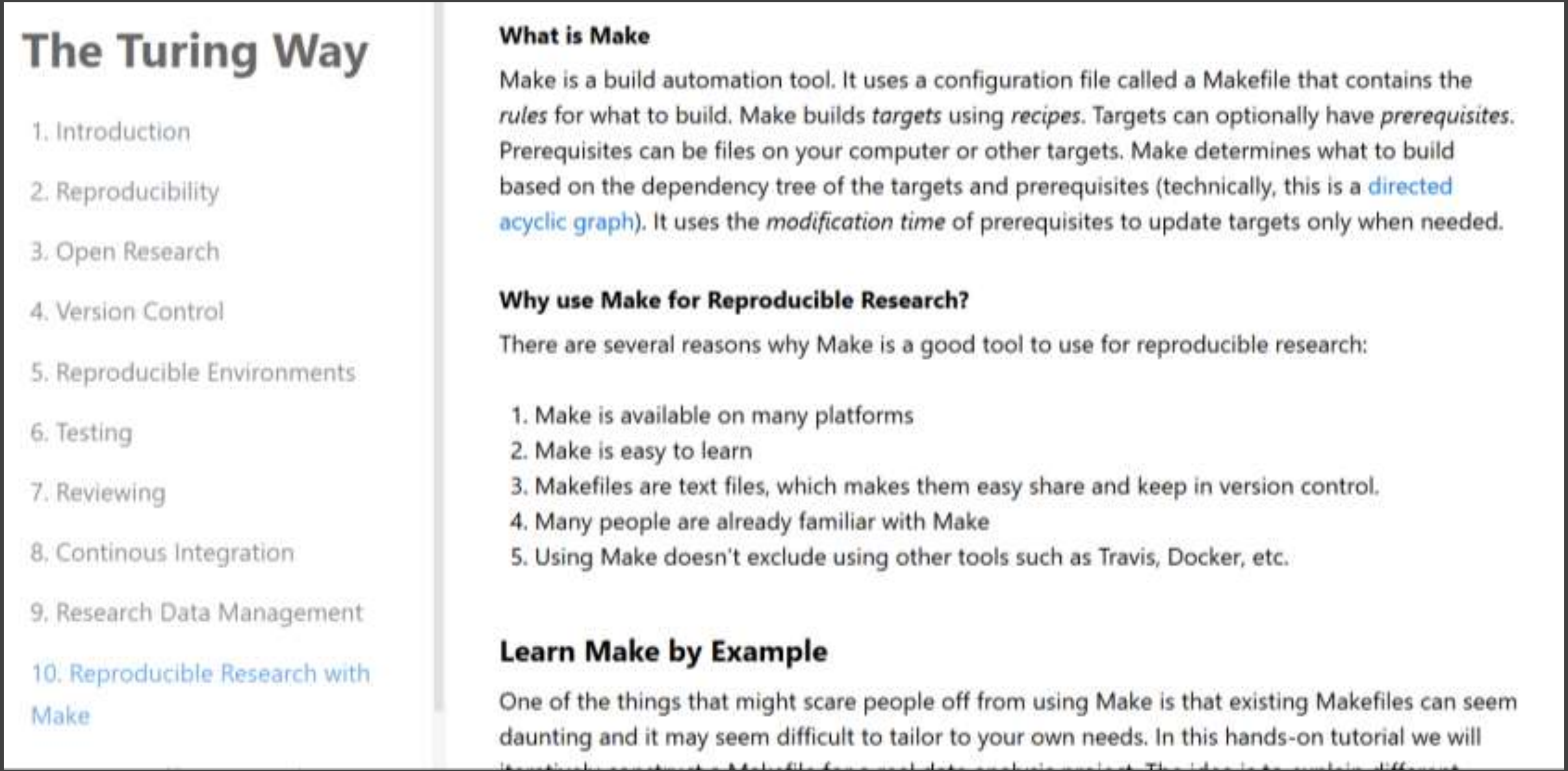
CSV Wrangling

This is the repository for reproducing the experiments in the paper:

Wrangling Messy CSV files by Detecting Row and Type Patterns

by G.J.J. van den Burg, A. Nazabal and C. Sutton.

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



The Turing Way

What is Make

Make is a build automation tool. It uses a configuration file called a Makefile that contains the *rules* for what to build. Make builds *targets* using *recipes*. Targets can optionally have *prerequisites*. Prerequisites can be files on your computer or other targets. Make determines what to build based on the dependency tree of the targets and prerequisites (technically, this is a directed acyclic graph). It uses the *modification time* of prerequisites to update targets only when needed.

Why use Make for Reproducible Research?

There are several reasons why Make is a good tool to use for reproducible research:

1. Make is available on many platforms
2. Make is easy to learn
3. Makefiles are text files, which makes them easy share and keep in version control.
4. Many people are already familiar with Make
5. Using Make doesn't exclude using other tools such as Travis, Docker, etc.

Learn Make by Example

One of the things that might scare people off from using Make is that existing Makefiles can seem daunting and it may seem difficult to tailor to your own needs. In this hands-on tutorial we will

https://the-turing-way.netlify.com/make/make.html

#PyDataLDN #TuringWay @kirstie_j

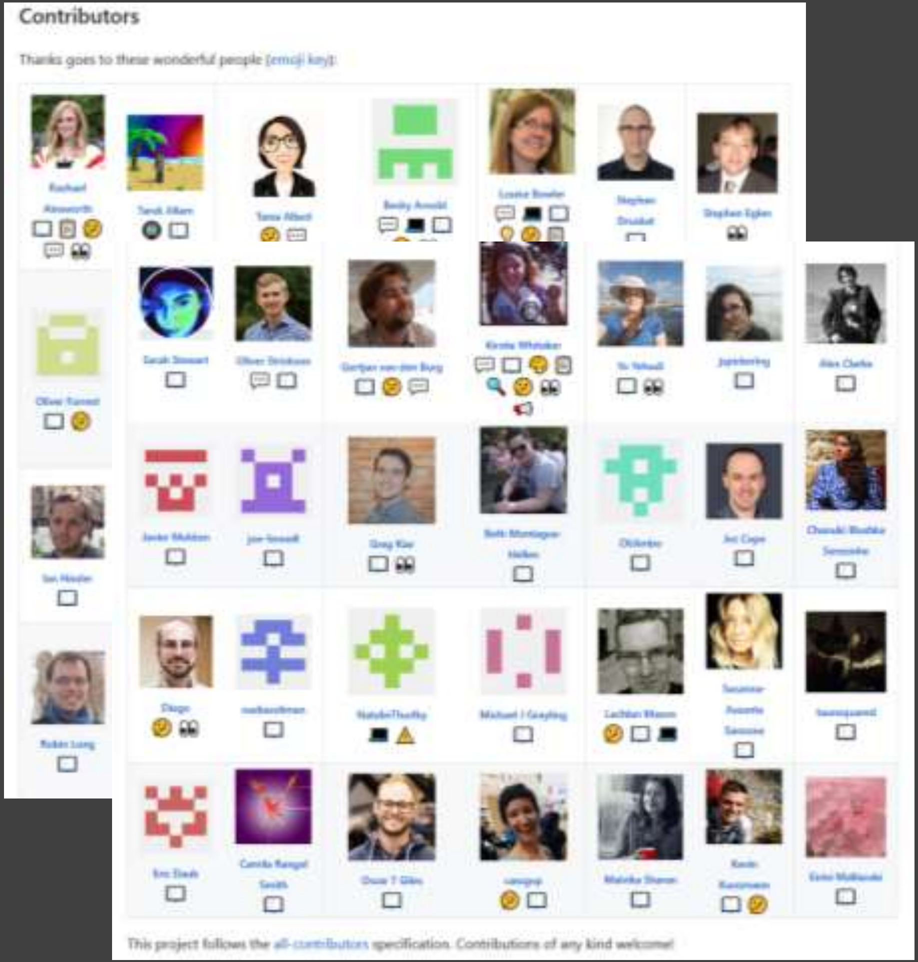
https://doi.org/10.5281/zenodo.3632909

Case studies

- Show that it can be done
- Provide templates and starting points
- Inspire

A global collaboration

#PyDataLDN #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3632909

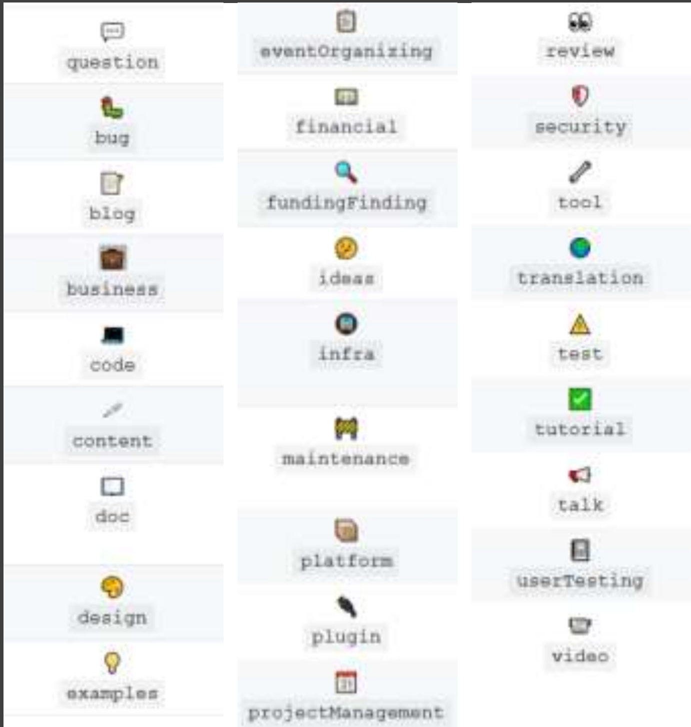


Contributors

Thanks goes to these wonderful people (emoji key):

Rachael Ainsworth
Tarek Allam
Tania Allard
Becky Arnold
Louise Bowler
Stephan Druskat
Stephen Eglen
Oliver Forrest
Sarah Stewart
Oliver Strickson
Gertjan van den Burg
Kirstie Whitaker
Yo Yehudi
Jsjoeberg
Alex Clarke
Ian Hinder
Javier Moldon
joe-lewell
Greg Kiar
Beth Montague-Hellen
Oldenbo
Jez Cope
Chanuki Illushka Seresinhe
Robin Long
Diego
nadiasoliman
NatalieThurlby
Michael J Grayling
Lachlan Mason
Susanna-Assunta Sansone
tauesquared
Eric Daub
Camila Rangel Smith
Oscar T Giles
sangrup
Malvika Sharan
Kevin Kunzmann
Eirini Malliaraki

This project follows the all-contributors specification. Contributions of any kind welcome!



💬 question
🐛 bug
📝 blog
💼 business
💻 code
🖋 content
📖 doc
🎨 design
💡 examples
📋 eventOrganizing
💵 financial
🔍 fundingFinding
🤔 ideas
🚇 infra
🚧 maintenance
📦 platform
🔌 plugin
📆 projectManagement
👀 review
🛡️ security
🔧 tool
🌍 translation
⚠️ test
✅ tutorial
📢 talk
📓 userTesting
📹 video

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

https://github.com/alan-turing-institute/the-turing-way#contributors
https://allcontributors.org/docs/en/emoji-key

#PyDataLDN #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3632909

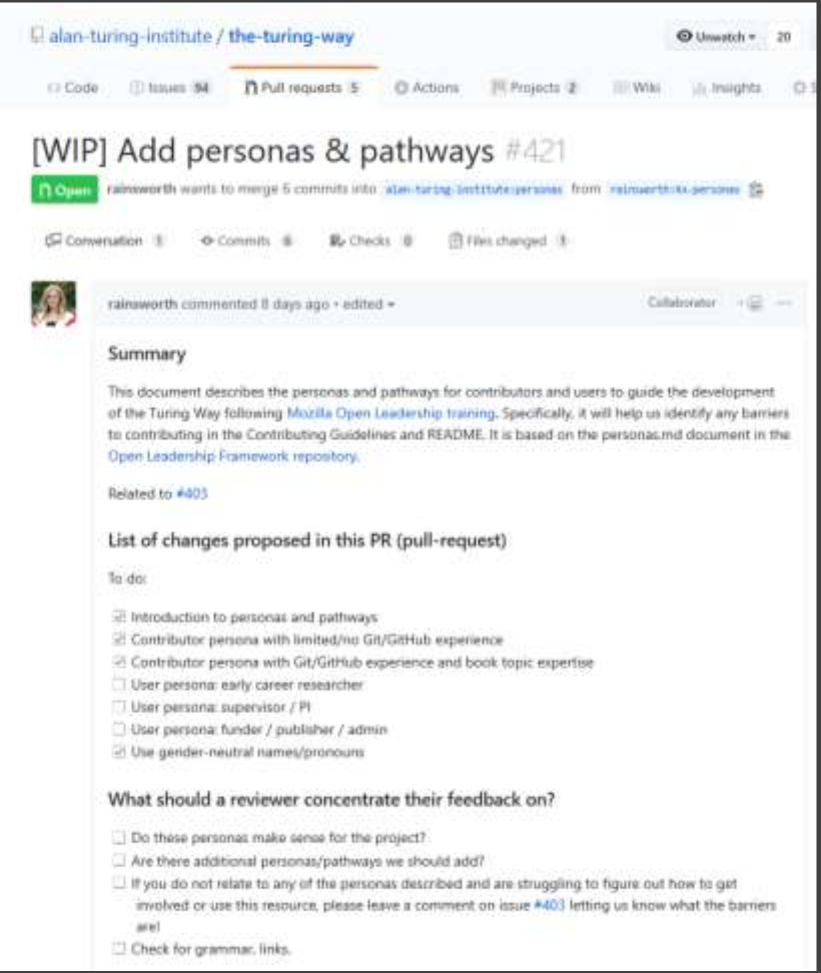
Rachael Ainsworth

“Personas and pathways exercises let me reflect on what people are finding difficult about contributing to The Turing Way. The project can only reach its potential if it is easy for a diverse constellation of contributions.”

https://rainsworth.github.io

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3632909



[WIP] Add personas & pathways #421

rainsworth wants to merge 5 commits into alan-turing-institute:personas from rainsworth:ks-personas

rainsworth commented 8 days ago • edited ▾

Summary

This document describes the personas and pathways for contributors and users to guide the development of the Turing Way following Mozilla Open Leadership training. Specifically, it will help us identify any barriers to contributing in the Contributing Guidelines and README. It is based on the personas.md document in the Open Leadership Framework repository.

Related to #403

List of changes proposed in this PR (pull-request)

To do:

- ☑ Introduction to personas and pathways
- ☑ Contributor persona with limited/no Git/GitHub experience
- ☑ Contributor persona with Git/GitHub experience and book topic expertise
- ☐ User persona: early career researcher
- ☐ User persona: supervisor / PI
- ☐ User persona: funder / publisher / admin
- ☑ Use gender-neutral names/pronouns

What should a reviewer concentrate their feedback on?

- ☐ Do these personas make sense for the project?
- ☐ Are there additional personas/pathways we should add?
- ☐ If you do not relate to any of the personas described and are struggling to figure out how to get involved or use this resource, please leave a comment on issue #403 letting us know what the barriers are!
- ☐ Check for grammar, links.

– **Sam**, who has no GitHub experience

– **Alex**, who has a lot of GitHub experience

– **Amal**, who knows they want to contribute, and does

– **Noor**, who doesn't know they want to contribute, but does

https://github.com/alan-turing-institute/the-turing-way/pull/421
#PyDataLDN #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3632909

WALKING PEOPLE THROUGH THE BASICS
TOOLS
NETWORKING
SHARING BEST PRACTICE
WELCOME!
FULL OF OPPORTUNITIES
HELPING PEOPLE GROW
OUR COMMUNITY
Scriberia

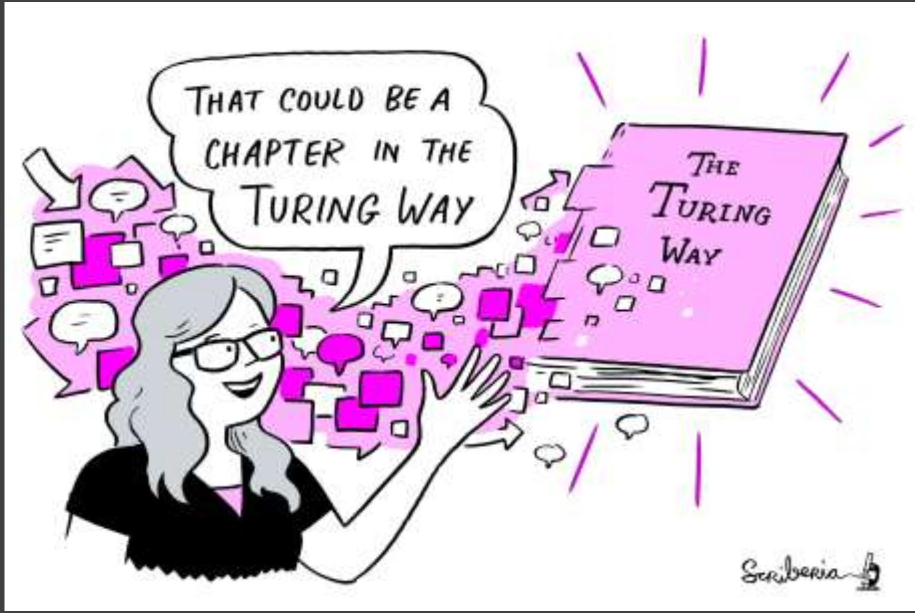
#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3632909

The future

Funding extension

- Expand scope to all data science practices
  - Ethics, model selection, project management, collaborative working
- Full time community manager, contributions from Turing & beyond



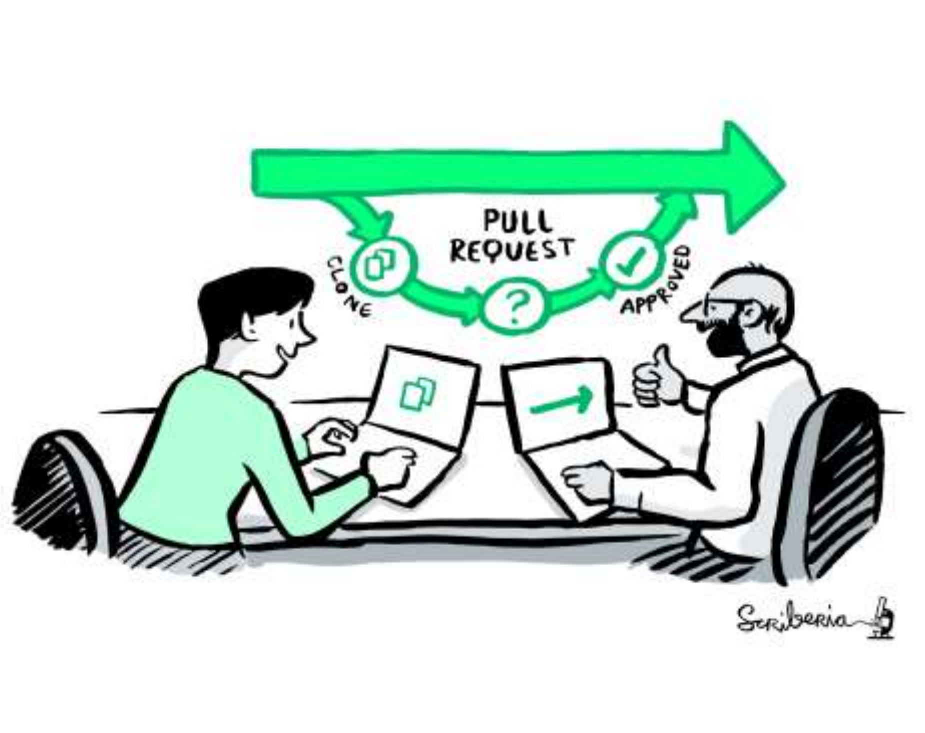
https://github.com/alan-turing-institute/the-turing-way/blob/master/project_management/tps-funding-application-20190429.md

#PyDataLDN #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3632909

Metrics for success

- 20 new chapters
- 100 authors
- 200 contributors
- 1000 mailing list subscribers
- 50 first pull requests
- 20 new contributors to other open source projects



#PyDataLDN #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3632909