Bringing research to life

Innovation to facilitate data sharing in the life sciences and biomedicine

NAOMI PENFOLD
Innovation Officer

Slides available at

https://github.com/npscience/csvconfv3-presentation

doi:10.6084/m9.figshare.4964999

About eLife

@eLifeInnovation

innovation@elifesciences.org

eLife is a non-profit organisation inspired by research funders and led by scientists


eLIFE
Helping scientists accelerate discovery by
operating a platform for research communication
that encourages and recognises the most
responsible behaviours in science

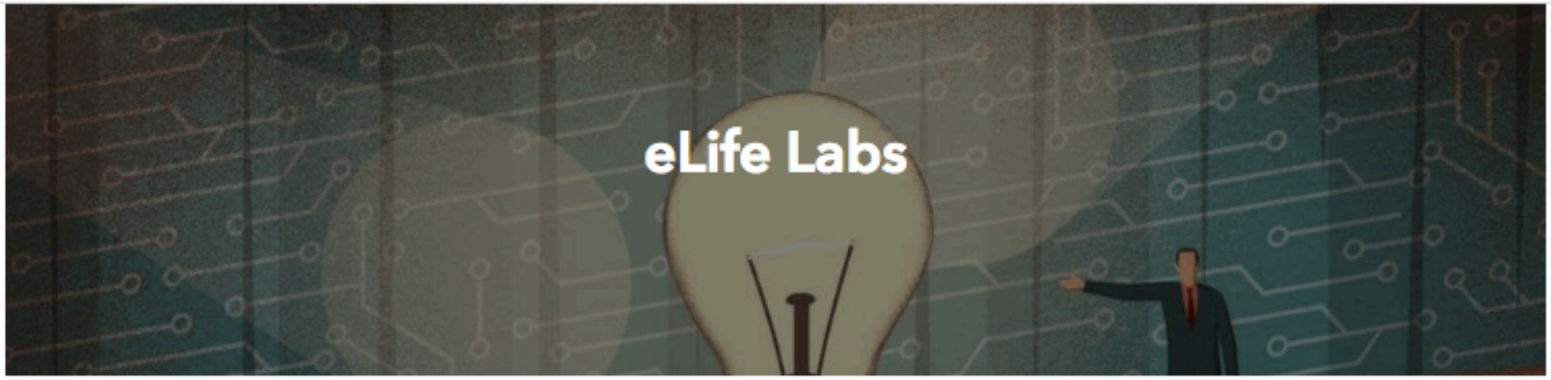
eLife Innovation Initiative

We invest in open source technologies, tools and processes that improve the way cutting-edge research is discovered, shared, consumed and evaluated

eLIFE

ABOUT LABS COMMUNITY SUBMIT MY RESEARCH

MENU HOME MAGAZINE

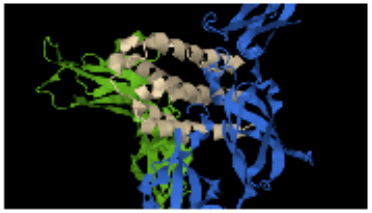


Exploring open-source solutions at the intersection of research and technology. Learn more about innovation at eLife, or follow us on Twitter.

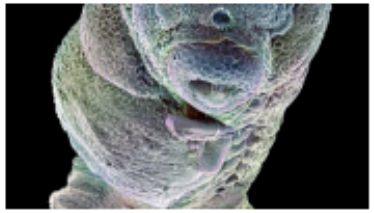
Latest

Composing reproducible manuscripts using R Markdown

Hack Cambridge Recurse entries: eXplore, Knowledge Direct, SciChat

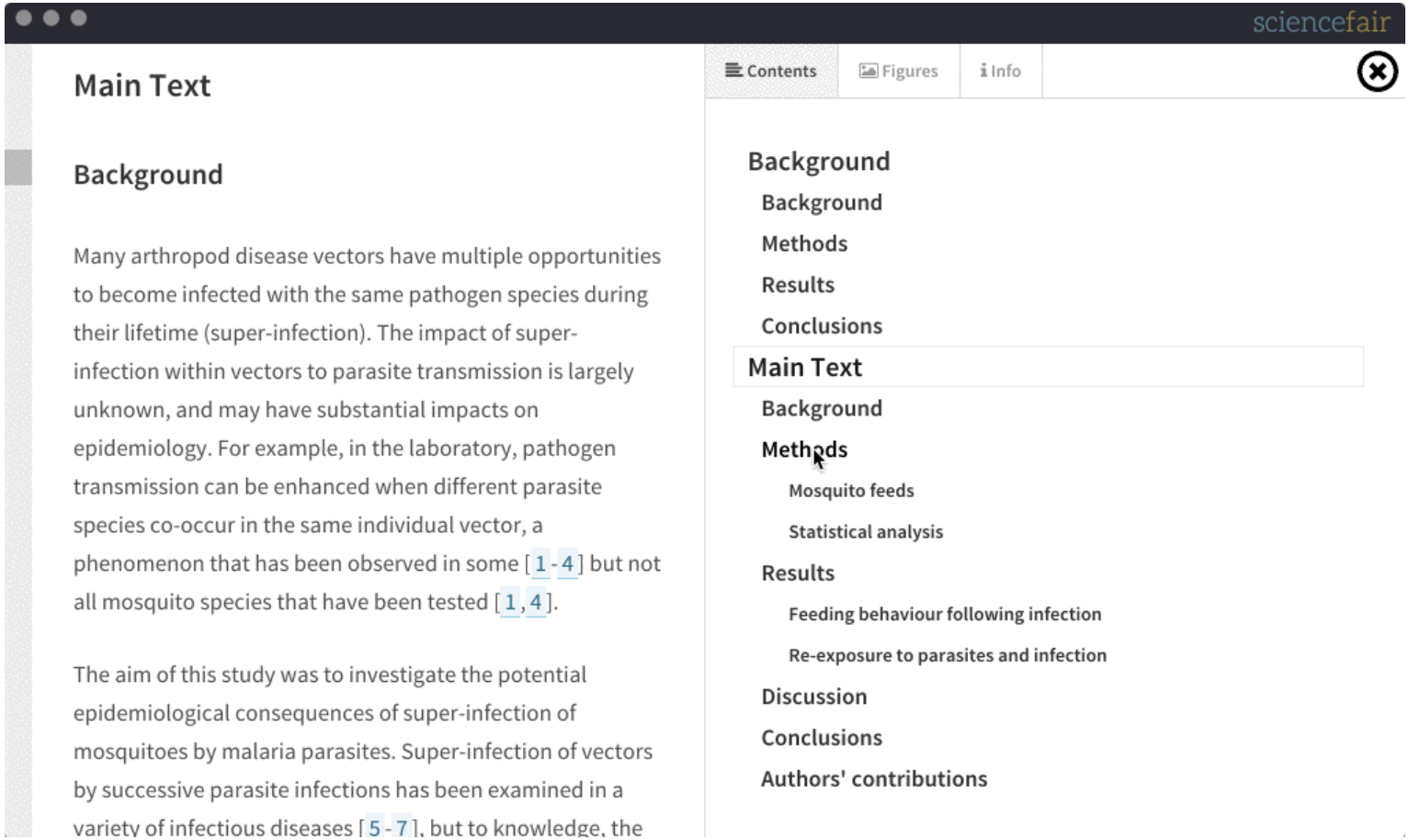


Proteopedia for sharing macromolecule concepts online



The International Image Interoperability Framework (IIIF) for science publishers

ScienceFair



https://github.com/codeforscience/sciencefair

Rik Smith-Unna
@blahah404

Follow

This evening, after ~1year of work, I've finally cut a beta release for ScienceFair. Seeking early adopter feedback.
github.com/codeforscience…
10:20 PM - 3 May 2017

codeforscience/sciencefair

sciencefair - :microscope: :book: a desktop science library that users control

github.com

13 19

Ideas?

Email Naomi at innovation@elifesciences.org

Why share data+ ?

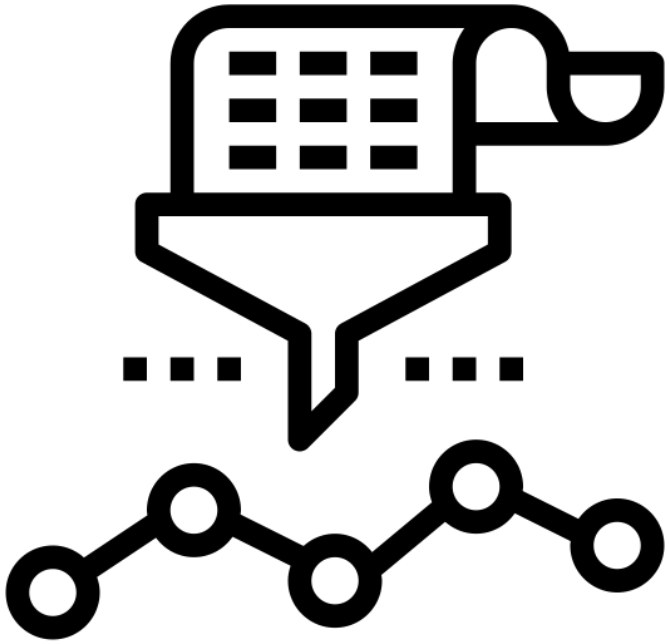
@eLifeInnovation

innovation@elifesciences.org

"If I have seen further,
it is by **standing**
on the shoulders of giants"

Isaac Newton

“Show me the evidence”



Created by Becris
from Noun Project

"True science **thrives best** in glass houses, where **everyone can look in**"

Max Perutz

Rufus Pollock
@rufuspollock

Follow

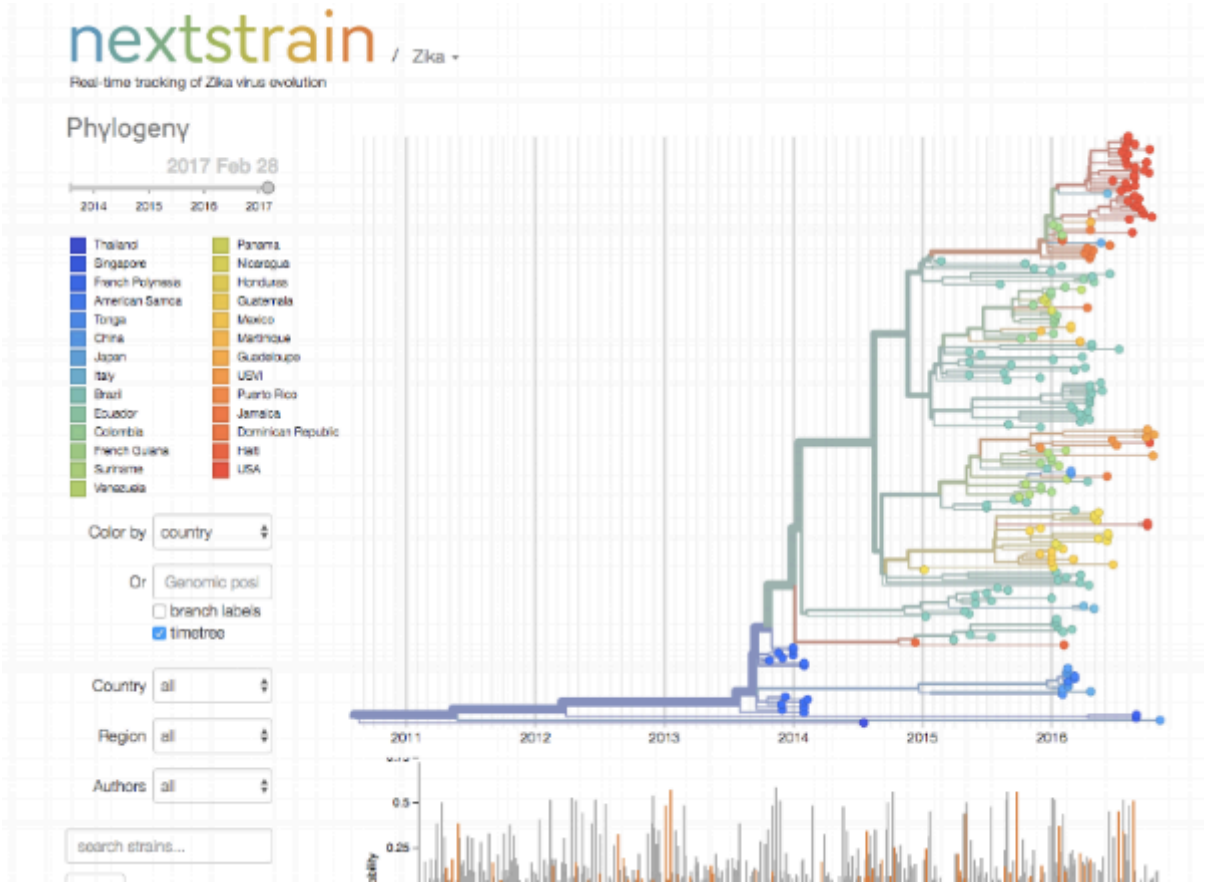
The best thing to do with your data will be thought of by someone else

Thank you @brianglanz for sharing#OpenData #OpenKnowledge

2:43 PM - 20 Apr 2017

52 64

Sharing data accelerates discovery and innovation



Sharing data allows community to spot problems…

Gene name errors are widespread in the scientific literature

Mark Ziemann, Yotam Eren and Assam El-Osta

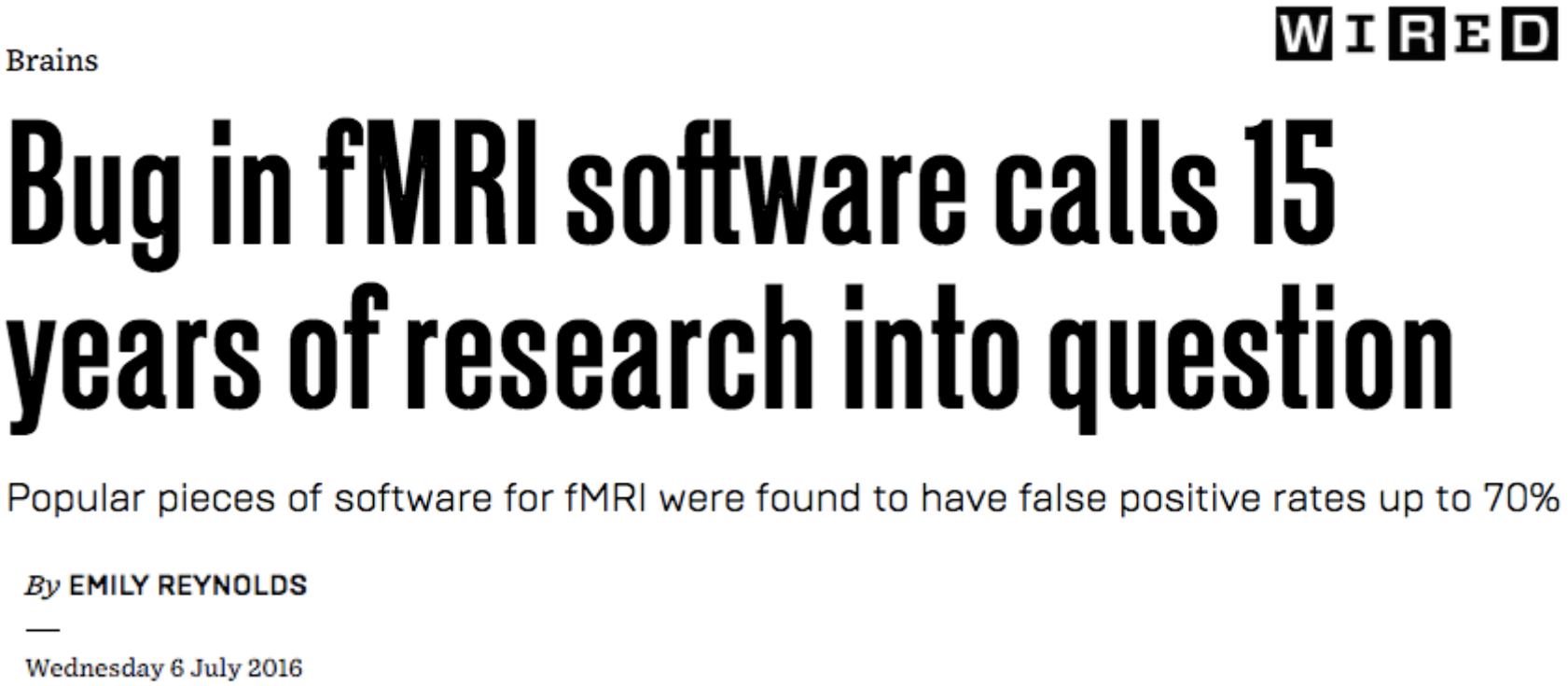
Genome Biology 2016 17:177 | DOI: 10.1186/s13059-016-1044-7 | © The Author(s). 2016

Published: 23 August 2016

- 'SEPT2' (Septin 2) → '2-Sep'
- '2310009E13' (RIKEN identifier) → '2.31E+13'

1 in 5

…and adjust with advancements



Brains

WIRED

Bug in fMRI software calls 15 years of research into question

Popular pieces of software for fMRI were found to have false positive rates up to 70%

By EMILY REYNOLDS

—

Wednesday 6 July 2016

The state of Open Data

@eLifeInnovation

innovation@elifesciences.org

Open… and FAIR?

F Findable

A Accessible

I Interoperable

R Re-usable

https://www.force11.org/group/fairgroup/fairprinciples

Recent surveys

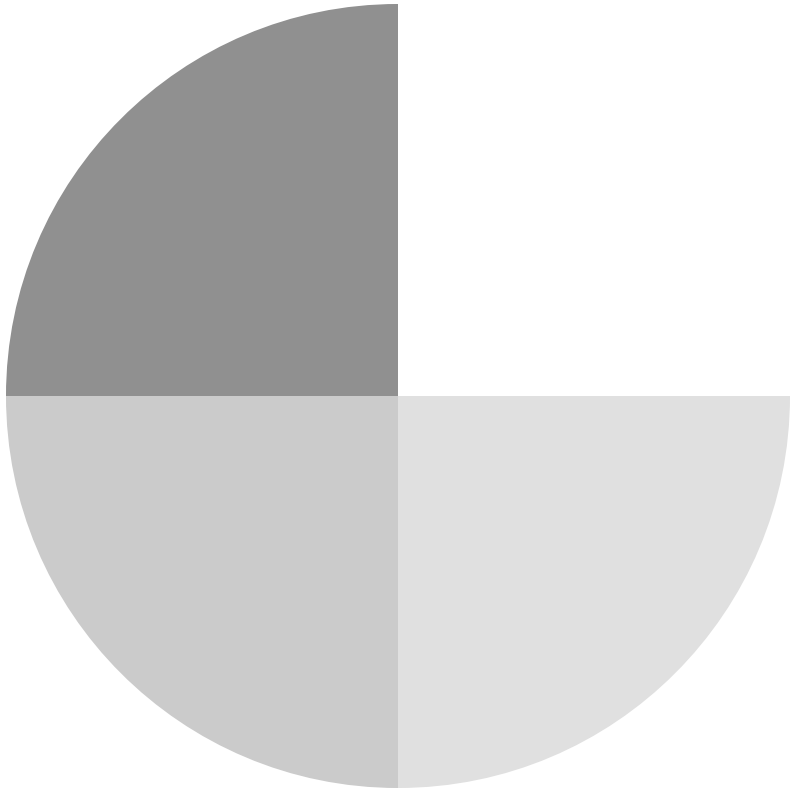
Van den Eynden, Veerle et al. (2016) Towards Open Research: Practices, experiences, barriers and opportunities. Wellcome Trust. https://dx.doi.org/10.6084/m9.figshare.4055448

Treadway, Jon; Hahnel, Mark; Leonelli, Sabina; Penny, Dan; Groenewegen, David; Miyairi, Nobuko; Hayashi, Kazuhiro; O'Donnell, Daniel; Science, Digital; Hook, Daniel (2016): The State of Open Data Report. figshare. https://doi.org/10.6084/m9.figshare.4036398.v1

73% of biologists aware

50% make data available

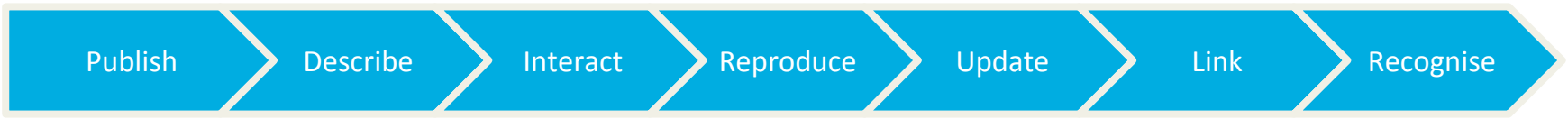
25% make full dataset available



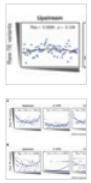
Improving data+ sharing

@eLifeInnovation

innovation@elifesciences.org


Publish
Describe
Interact
Reproduce
Update
Link
Recognise

Data sharing policies



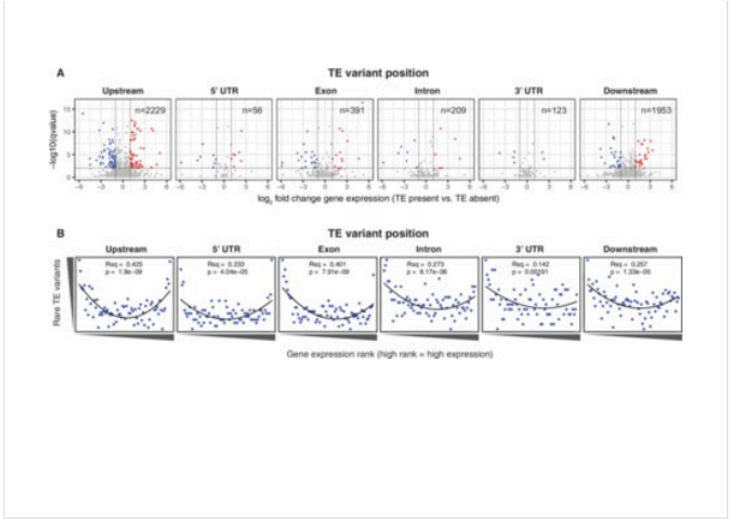


Figure 4. Download figure | Open in new tab

Differential transcript abundance associated with TE variant presence/absence.

(**A**) Transcript abundance differences for genes associated with TE insertion variants at different positions, indicated in the plot titles. Genes with significantly different transcript abundance in accessions with a TE insertion compared to accessions without a TE insertion are colored blue (lower transcript abundance in accessions containing TE insertion) or red (higher transcript abundance in accessions containing TE insertion). Vertical lines indicate ±2 fold change in FPKM. Horizontal line indicates the 1% false discovery rate. (**B**) Relationship between rare TE variant counts and gene expression rank. Cumulative number of rare TE variants in equal-sized bins for gene expression ranks, from the lowest-ranked accession (left) to the highest-ranked accession (right). Lines indicate the fit of a quadratic model.

DOI: http://dx.doi.org/10.7554/eLife.20777.025

Figure 4—source data 1.

Differentially expressed genes associated with TE presence/absence.

List of genes differentially expressed dependent on the presence/absence of nearby TE variants.

DOI: http://dx.doi.org/10.7554/eLife.20777.026

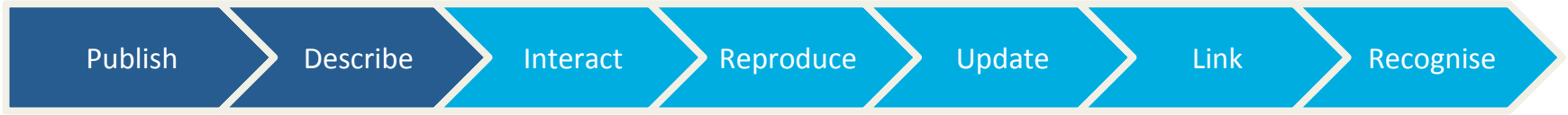
Download source data [figure-4—source-data-1.media-6.tsv]

Publish | Describe | Interact | Reproduce | Update | Link | Recognise

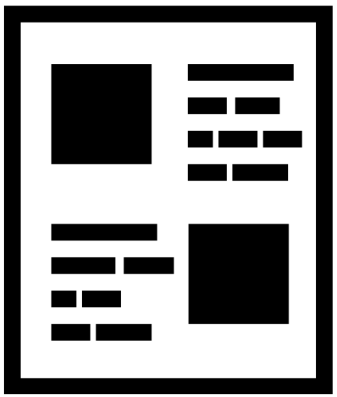
Data publications

(GIGA)n SCIENCE

SCIENTIFIC DATA

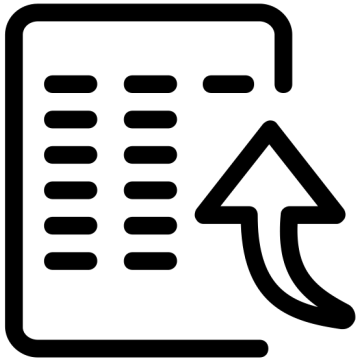


Can we bring the data closer to the narrative?



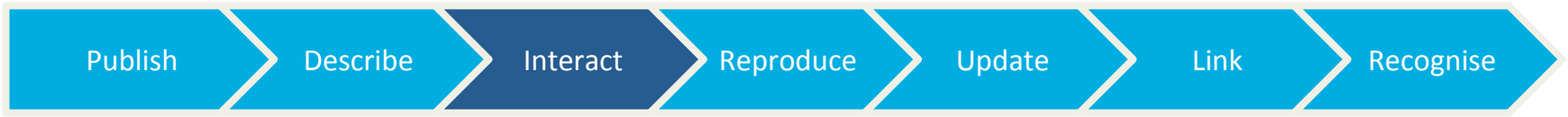
Created by Dmitry Podluzny
from Noun Project

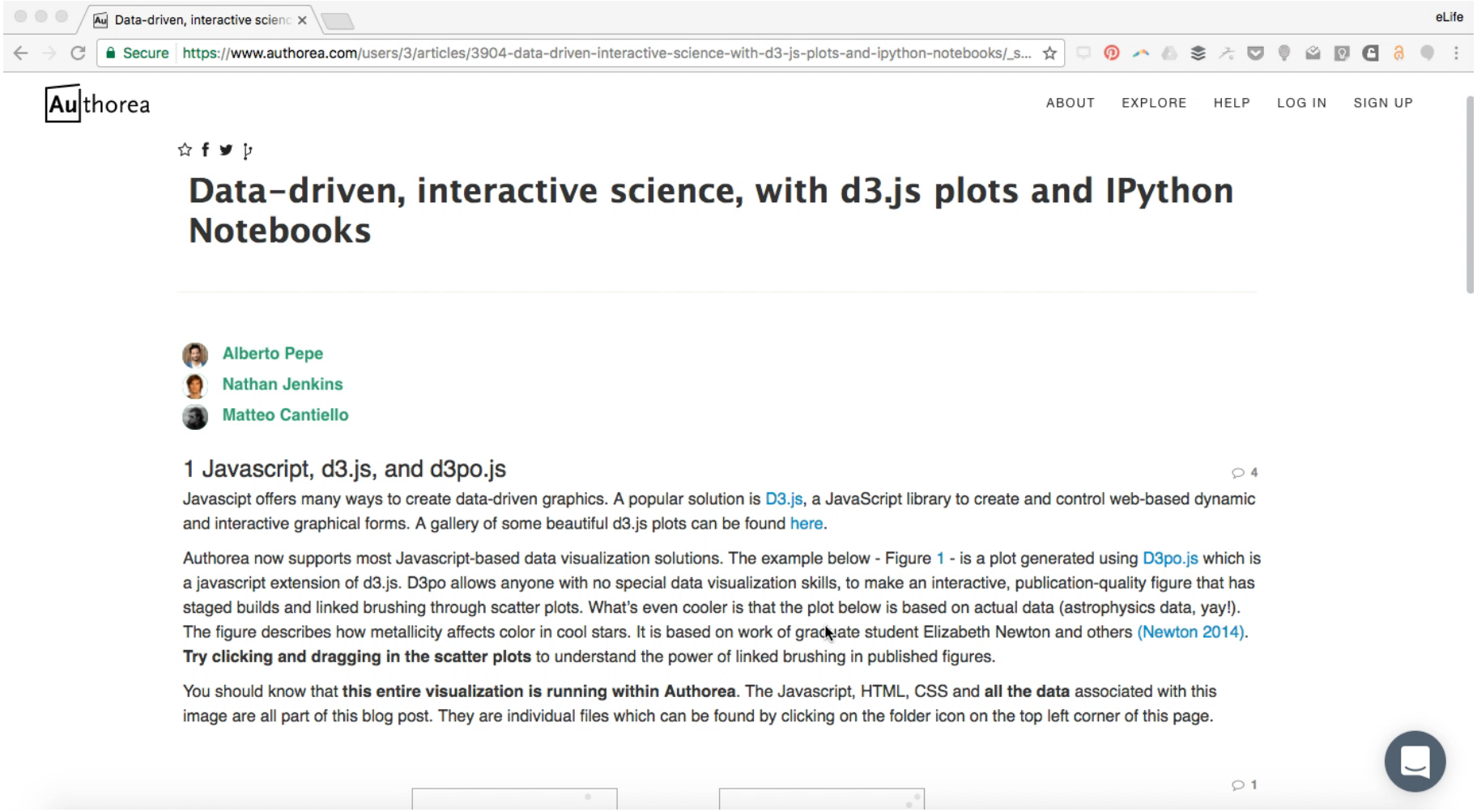
+



Created by IconfactoryTeam
from Noun Project

The interactive figure

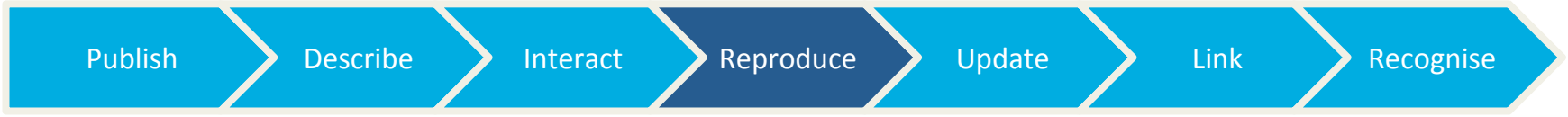


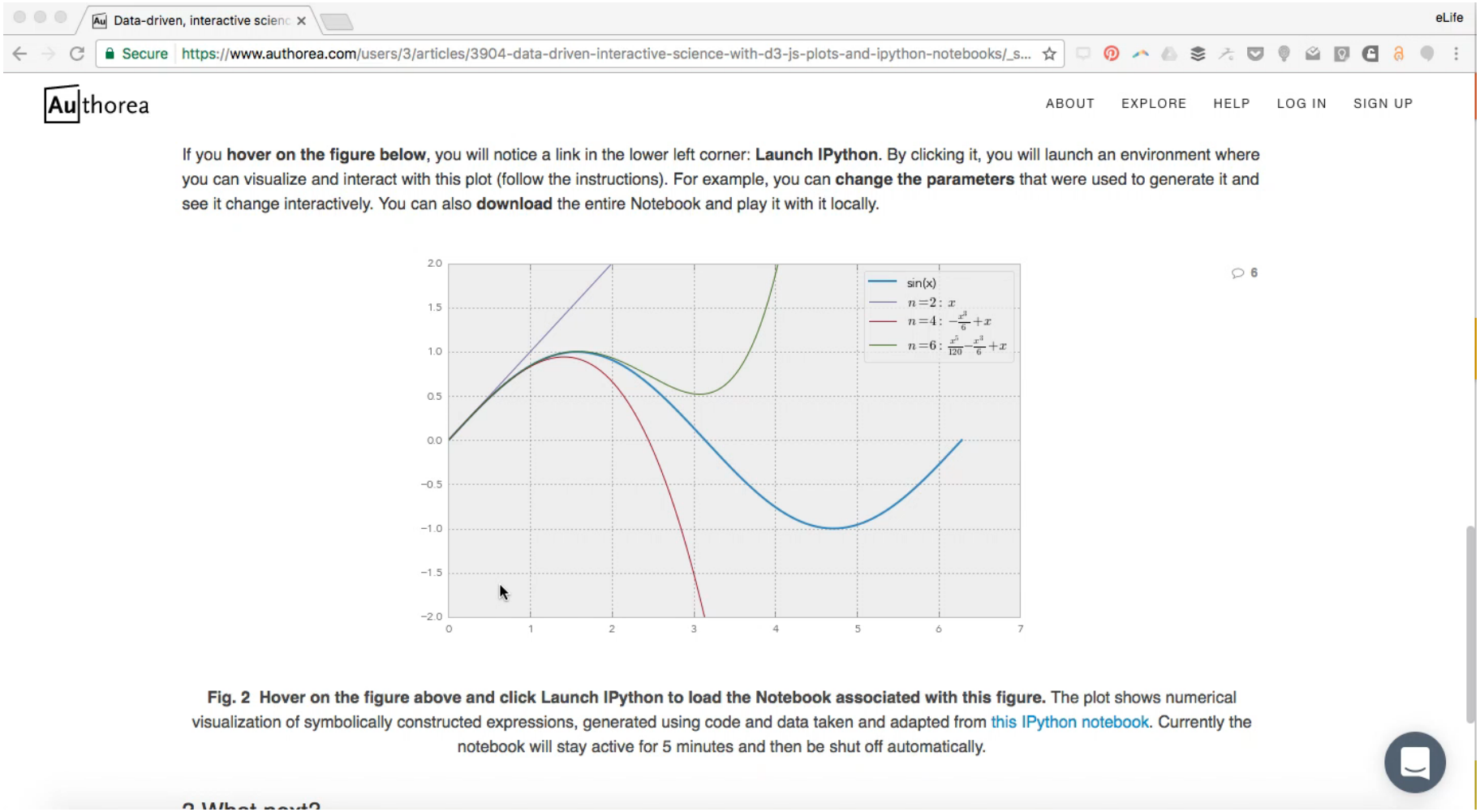


Some steps removed and sequences shortened

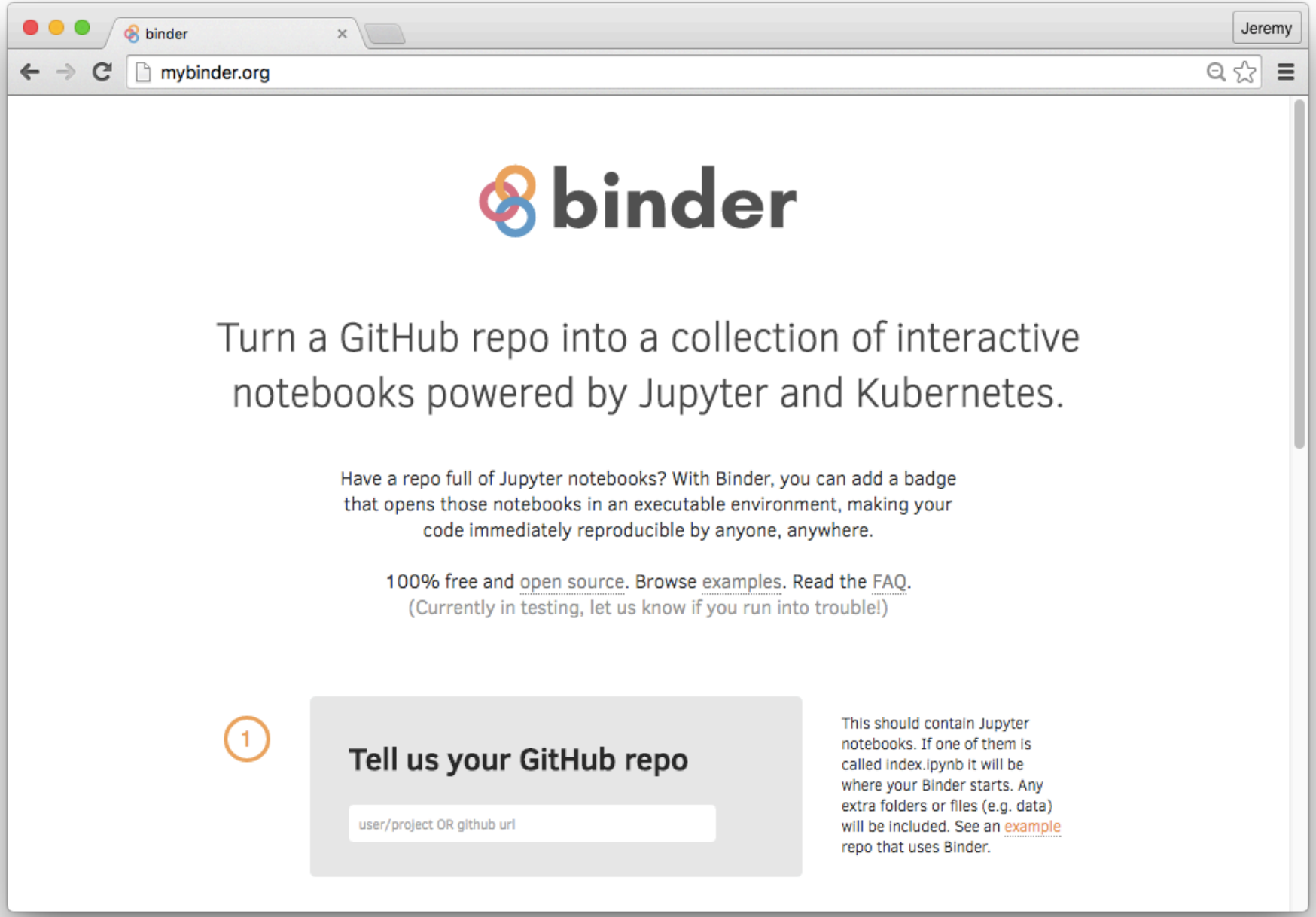
The executable figure

Some steps removed and sequences shortened

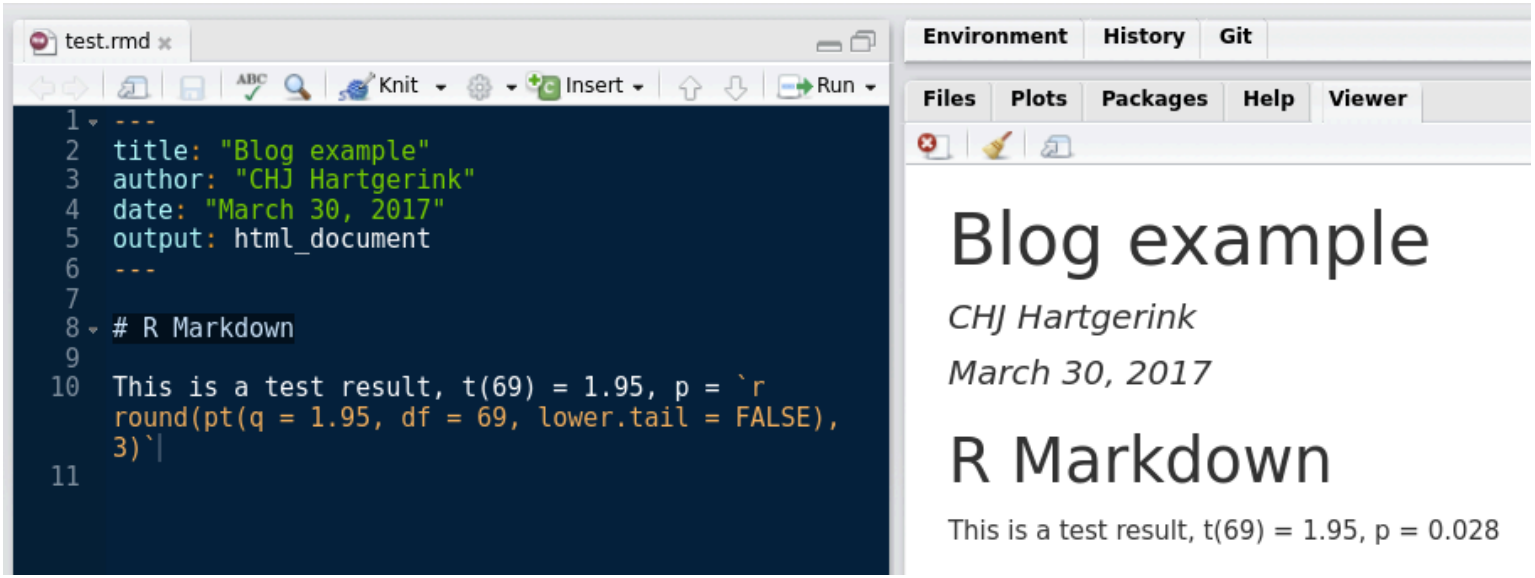




Some steps removed and sequences shortened


binder
Jeremy
mybinder.org
binder
Turn a GitHub repo into a collection of interactive notebooks powered by Jupyter and Kubernetes.
Have a repo full of Jupyter notebooks? With Binder, you can add a badge that opens those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.
100% free and open source. Browse examples. Read the FAQ.
(Currently in testing, let us know if you run into trouble!)
1
Tell us your GitHub repo
user/project OR github url
This should contain Jupyter notebooks. If one of them is called index.ipynb it will be where your Binder starts. Any extra folders or files (e.g. data) will be included. See an example repo that uses Binder.

The reproducible document



Publish > Describe > Interact > Reproduce > Update > Link > Recognise

"It took me a couple of hours to…

REPRODUCE EXACTLY the analysis presented in the manuscript...

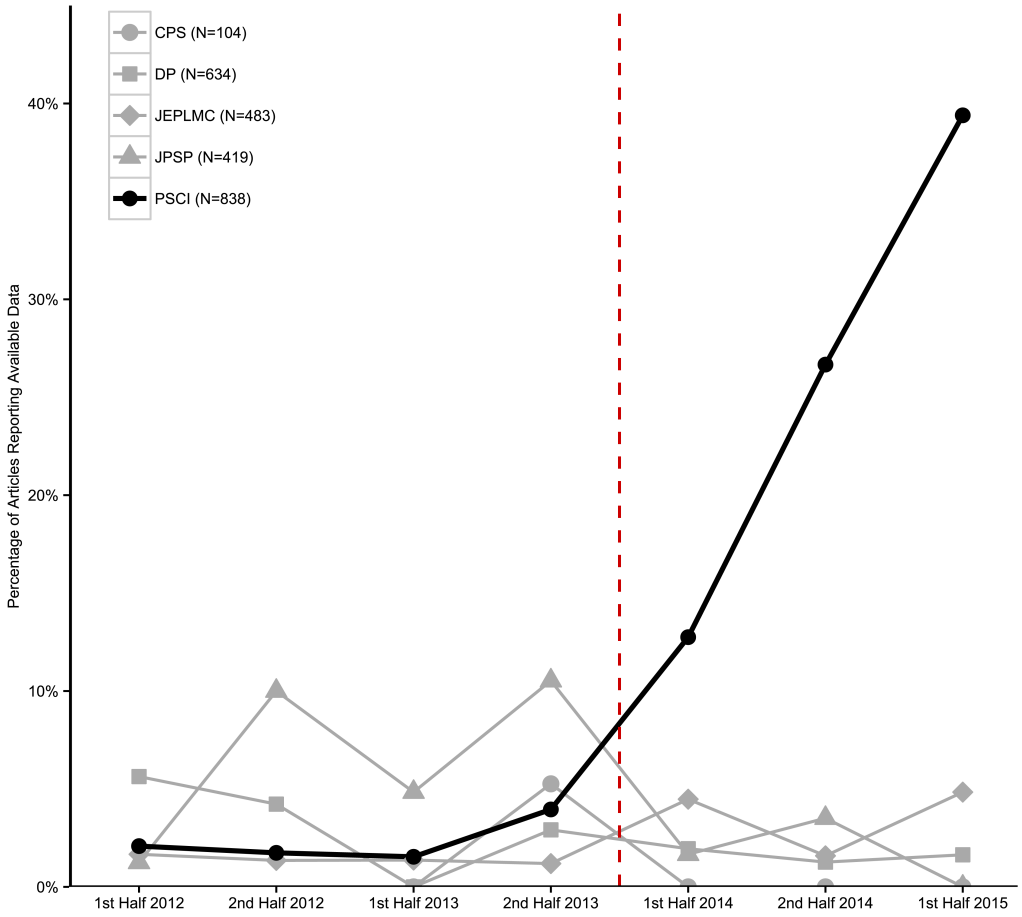
With few more hours, **I was able to modify the authors' code**

to change a linear scale for a log scale for their Fig. 4."

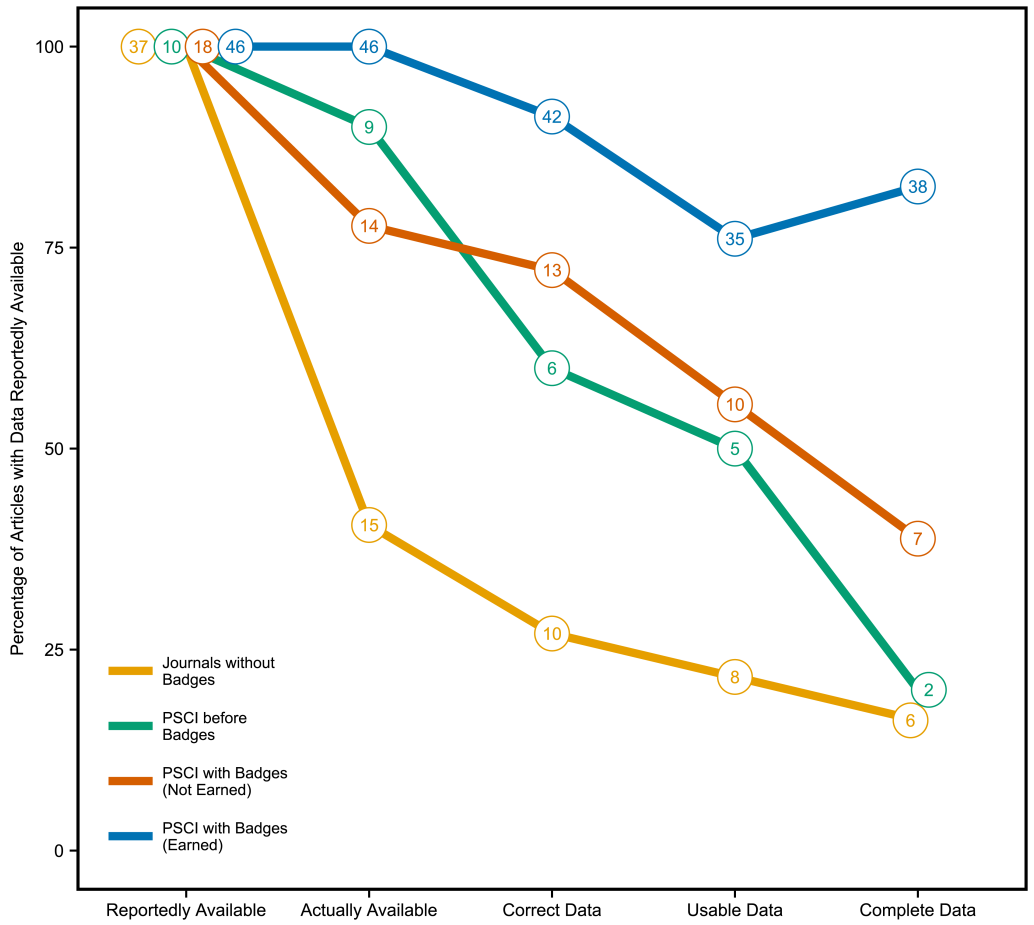
Christophe Pouzat, reviewer
GigaScience blog: http://gigasciencejournal.com/blog/qa-on-dynamic-documents

Credit

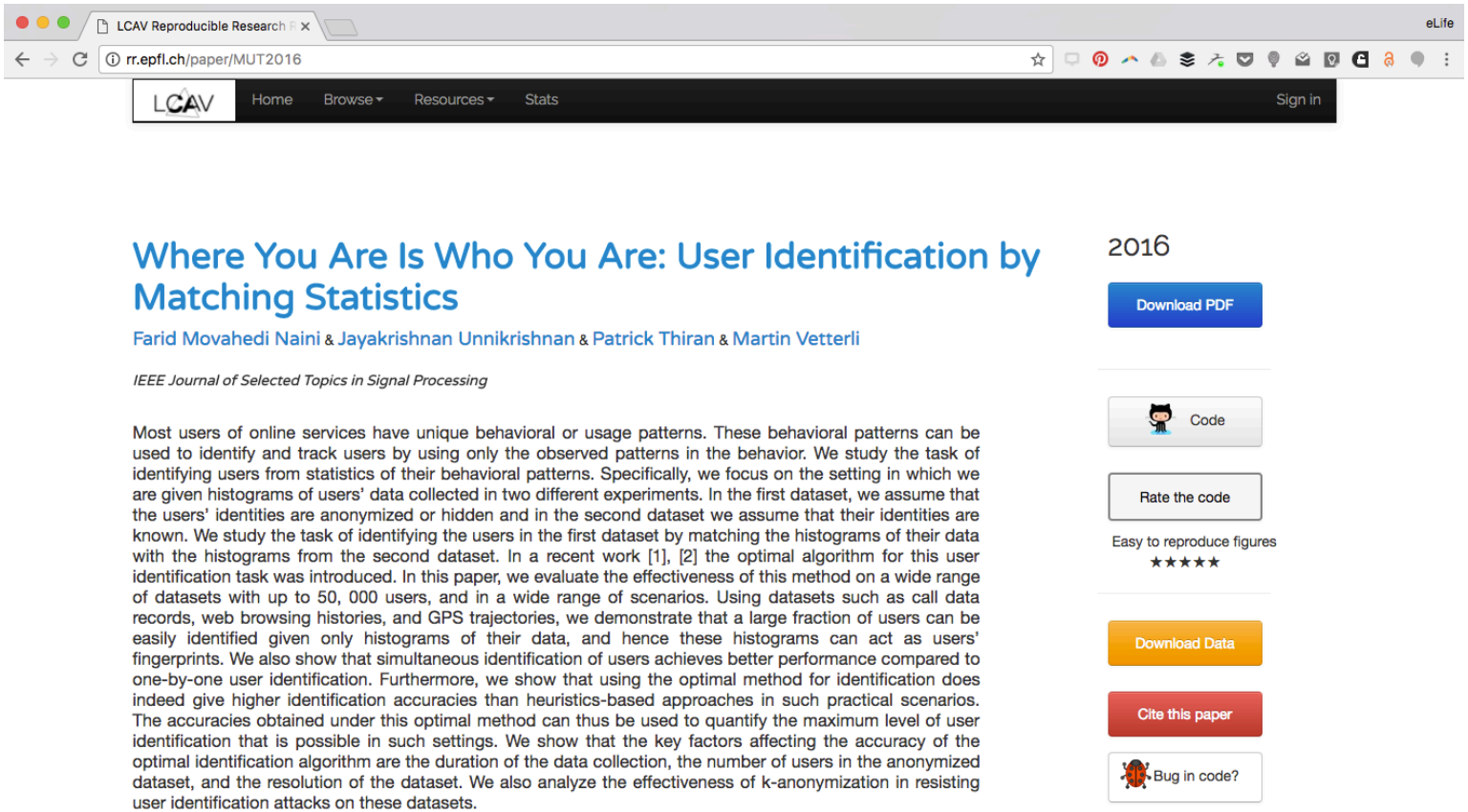
Publish | Describe | Interact | Reproduce | Update | Recognise | Link



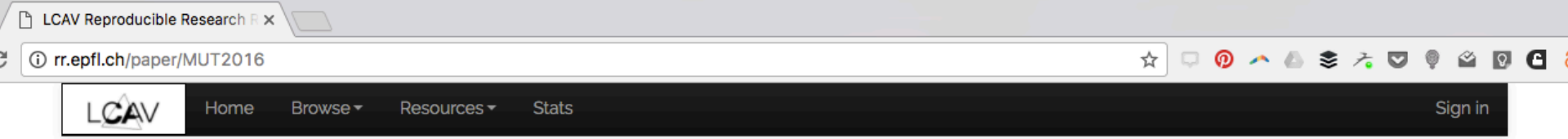
PLOS Biology, 2016; https://doi.org/10.1371/journal.pbio.1002456.g002



PLOS Biology, 2016; https://doi.org/10.1371/journal.pbio.1002456.g002



Publish | Describe | Interact | Reproduce | Update | Recognise | Link



Where You Are Is Who You Are: User Identification by Matching Statistics

Farid Movahedi Naini & Jayakrishnan Unnikrishnan & Patrick Thiran & Martin Vetterli

IEEE Journal of Selected Topics in Signal Processing

Most users of online services have unique behavioral or usage patterns. These behavioral patterns can be used to identify and track users by using only the observed patterns in the behavior. We study the task of identifying users from statistics of their behavioral patterns. Specifically, we focus on the setting in which we are given histograms of users' data collected in two different experiments. In the first dataset, we assume that the users' identities are anonymized or hidden and in the second dataset we assume that their identities are known. We study the task of identifying the users in the first dataset by matching the histograms of their data with the histograms from the second dataset. In a recent work [1], [2] the optimal algorithm for this user identification task was introduced. In this paper, we evaluate the effectiveness of this method on a wide range of datasets with up to 50, 000 users, and in a wide range of scenarios. Using datasets such as call data records, web browsing histories, and GPS trajectories, we demonstrate that a large fraction of users can be easily identified given only histograms of their data, and hence these histograms can act as users' fingerprints. We also show that simultaneous identification of users achieves better performance compared to one-by-one user identification. Furthermore, we show that using the optimal method for identification does indeed give higher identification accuracies than heuristics-based approaches in such practical scenarios. The accuracies obtained under this optimal method can thus be used to quantify the maximum level of user identification that is possible in such settings. We show that the key factors affecting the accuracy of the optimal identification algorithm are the duration of the data collection, the number of users in the anonymized dataset, and the resolution of the dataset. We also analyze the effectiveness of k-anonymization in resisting user identification attacks on these datasets.

2016

Download PDF

Code

Rate the code

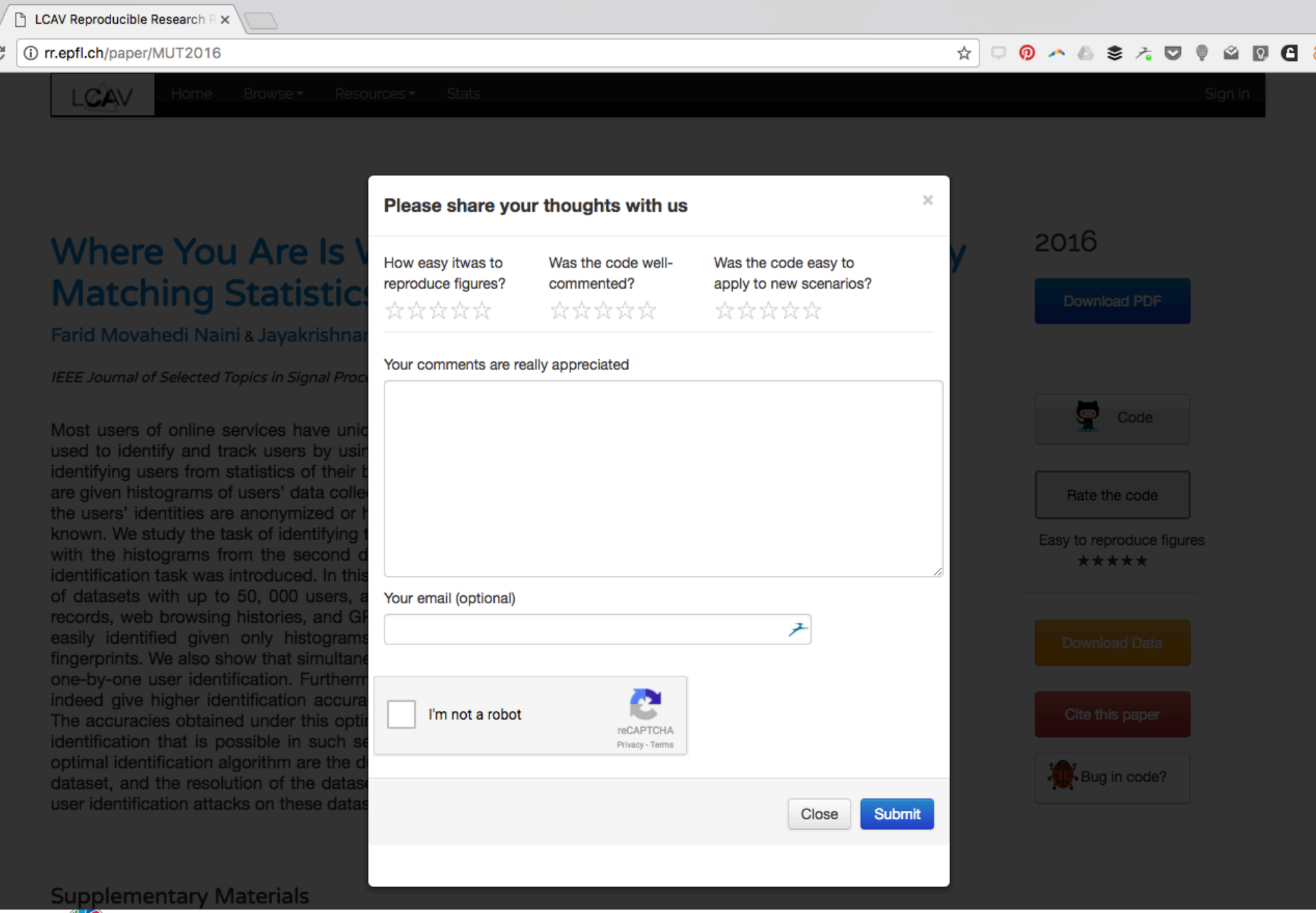
Easy to reproduce figures

★★★★★

Download Data

Cite this paper

Bug in code?


LCAV Reproducible Research
rr.epfl.ch/paper/MUT2016
LCAV
Home
Browse
Resources
Stats
Sign in
Where You Are Is
Matching Statistic
Farid Movahedi Naini & Jayakrishnan
IEEE Journal of Selected Topics in Signal Proc
Most users of online services have unic
used to identify and track users by usin
identifying users from statistics of their b
are given histograms of users' data colle
the users' identities are anonymized or h
known. We study the task of identifying t
with the histograms from the second d
identification task was introduced. In this
of datasets with up to 50, 000 users, a
records, web browsing histories, and GP
easily identified given only histograms
fingerprints. We also show that simultane
one-by-one user identification. Furtherm
indeed give higher identification accura
The accuracies obtained under this opti
identification that is possible in such se
optimal identification algorithm are the d
dataset, and the resolution of the datas
user identification attacks on these datas
Supplementary Materials
Please share your thoughts with us
How easy itwas to reproduce figures?
Was the code well-commented?
Was the code easy to apply to new scenarios?
Your comments are really appreciated
Your email (optional)
I'm not a robot
reCAPTCHA
Privacy - Terms
Close
Submit
2016
Download PDF
Code
Rate the code
Easy to reproduce figures
Download Data
Cite this paper
Bug in code?

easily identified given only histograms of their data, and hence these histograms can act as users' fingerprints. We also show that simultaneous identification of users achieves better performance compared to one-by-one user identification. Furthermore, we show that using the optimal method for identification does indeed give higher identification accuracies than heuristics-based approaches in such practical scenarios. The accuracies obtained under this optimal method can thus be used to quantify the maximum level of user identification that is possible in such settings. We show that the key factors affecting the accuracy of the optimal identification algorithm are the duration of the data collection, the number of users in the anonymized dataset, and the resolution of the dataset. We also analyze the effectiveness of k-anonymization in resisting user identification attacks on these datasets.

Download Data

Cite this paper

Bug in code?

Supplementary Materials

How to run the code

First, download the code and data files, uncompress them and copy/paste the content of the data files into the `generatingResultsFigures/` folder. Then, to generate the figures in the paper, simply open **MATLAB** and navigate to the `generatingResultsFigures/` folder (in the code files) and follow the instructions below.

Note that since only the WBH and the GL datasets are publicly available (in here and here, respectively), the codes generate only the figures related to these two datasets.

Figure 8 and 9

In MATLAB, type

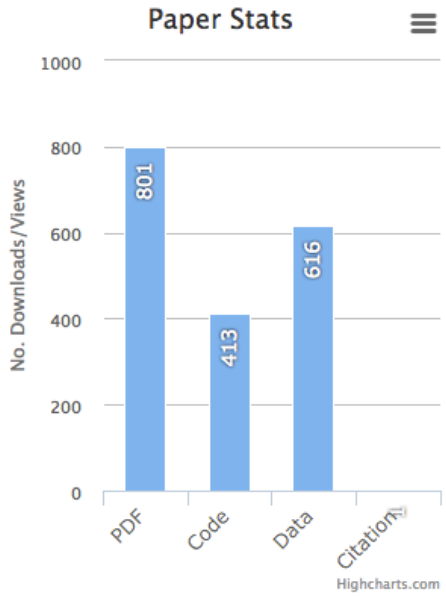
```
>> WBH_results1.m
```

and

```
>> WBH_results2.m
```

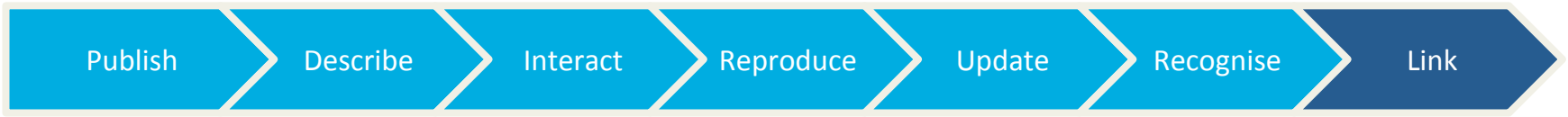
to reproduce Figures 8 and 9, respectively.

Figure 10 and 11



Connecting projects

- Linked open data
- Human curation



ReFigure

RE figure

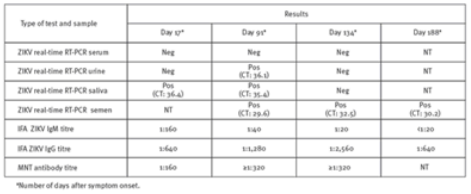
<basic navigation menu>

Figures hyperlinked to original paper

Longest post onset detection of Zika RNA in semen in multiple case studies

Report/Flag

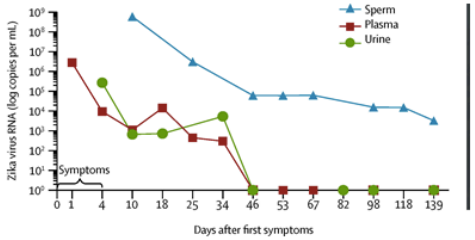
Christa Osuna



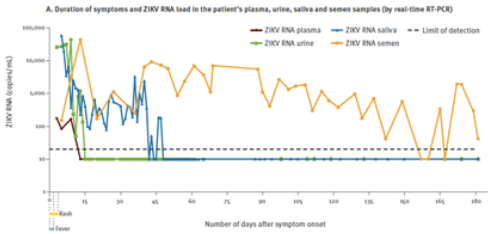
Type of test and sample	Results			
	Day 17*	Day 91*	Day 134*	Day 188*
ZIKV real-time RT-PCR serum	Neg	Neg	Neg	NT
ZIKV real-time RT-PCR urine	Neg	Pos (CT: 36.1)	Neg	NT
ZIKV real-time RT-PCR saliva	Pos (CT: 36.4)	Pos (CT: 35.4)	Neg	NT
ZIKV real-time RT-PCR semen	NT	Pos (CT: 29.4)	Pos (CT: 32.5)	Pos (CT: 30.2)
IFA ZIKV IgM titre	1:160	1:40	1:20	<1:20
IFA ZIKV IgG titre	1:640	1:1,280	1:2,560	1:640
MNT antibody titre	1:160	≥1:320	≥1:320	NT

*Number of days after symptom onset.

Laboratory findings related to Zika virus infection in a traveller returning from Haiti to Italy, Feb-July 2016
Nicostri et al
Eurosurveillance



Zika virus in semen and spermatozoa
Mansuy et al
Lancet Infectious Diseases



Clinical and laboratory findings in a patient returning from Haiti to Italy, Jan 2016
Barzon et al
Euro Surveillance

<scroll bar>

http://refigure.org/index.html

Where next?

@eLifeInnovation

innovation@elifesciences.org

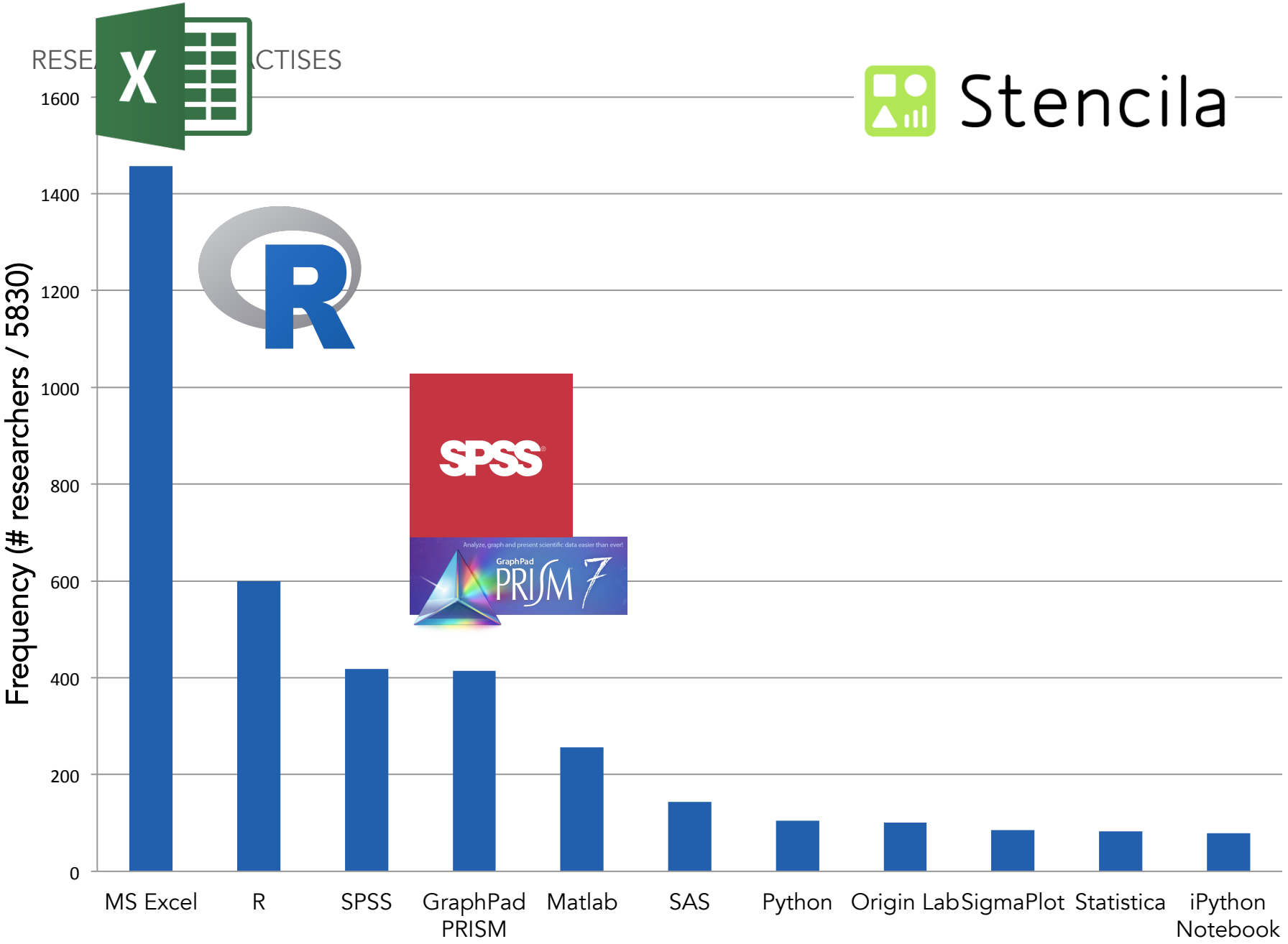
Publisher constraints

- Static version of peer-reviewed article
- Persistence
- Accuracy
- Who stores? Who hosts? Who computes?

…and researcher practises

RESE CTISES

Stencila



Outstanding issues

- Improve quality of data sharing
- Incentives

"No time or funding" to learn new tools, document data well

"No benefit"

Opportunities

- Facilitating data sharing **via the journal** means:
  - Persistence
  - Leverages incentive system
  - Streamline the process from initial discovery to replication
  - Engages community

- Doing it **open source** means:
  - Progress can be shared
  - Stimulates further innovation
  - Engages community

Your thoughts?

Naomi Penfold, Innovation Officer

Tweet: @eLifeInnovation (or @npscience)

Email: innovation@elifesciences.org

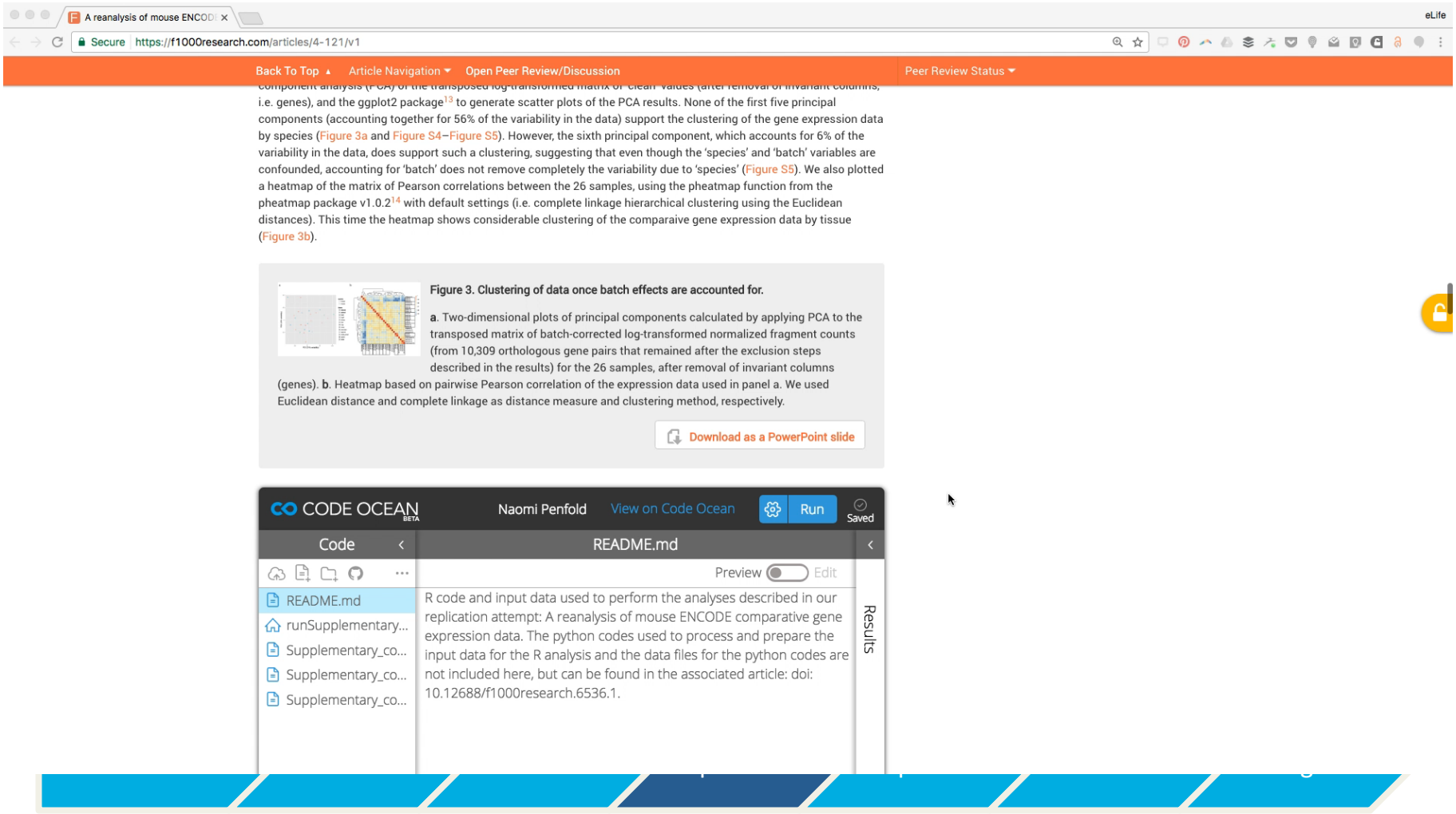
Slides: github.com/npscience/csvconfv3-presentation

"The **impact** we cherish is **discovery** in science"

Randy Schekman, eLife Editor-in-Chief

The reproducible project

F1000: introducing the CodeOcean widget: https://blog.f1000.com/2017/04/20/reanaly-seas-making-reproducibility-easier-with-code-ocean-widgets/


A reanalysis of mouse ENCOD
eLife
Secure https://f1000research.com/articles/4-121/v1
Back To Top
Article Navigation
Open Peer Review/Discussion
Peer Review Status
component analysis (PCA) of the transposed log-transformed matrix of clean values (after removal of invariant columns, i.e. genes), and the ggplot2 package[13] to generate scatter plots of the PCA results. None of the first five principal components (accounting together for 56% of the variability in the data) support the clustering of the gene expression data by species (Figure 3a and Figure S4–Figure S5). However, the sixth principal component, which accounts for 6% of the variability in the data, does support such a clustering, suggesting that even though the 'species' and 'batch' variables are confounded, accounting for 'batch' does not remove completely the variability due to 'species' (Figure S5). We also plotted a heatmap of the matrix of Pearson correlations between the 26 samples, using the pheatmap function from the pheatmap package v1.0.2[14] with default settings (i.e. complete linkage hierarchical clustering using the Euclidean distances). This time the heatmap shows considerable clustering of the comparaive gene expression data by tissue (Figure 3b).
Figure 3. Clustering of data once batch effects are accounted for.
a. Two-dimensional plots of principal components calculated by applying PCA to the transposed matrix of batch-corrected log-transformed normalized fragment counts (from 10,309 orthologous gene pairs that remained after the exclusion steps described in the results) for the 26 samples, after removal of invariant columns (genes). b. Heatmap based on pairwise Pearson correlation of the expression data used in panel a. We used Euclidean distance and complete linkage as distance measure and clustering method, respectively.
Download as a PowerPoint slide
CODE OCEAN
Naomi Penfold
View on Code Ocean
Run
Saved
Code
README.md
Preview
Edit
README.md
runSupplementary...
Supplementary_co...
Supplementary_co...
Supplementary_co...
R code and input data used to perform the analyses described in our replication attempt: A reanalysis of mouse ENCODE comparative gene expression data. The python codes used to process and prepare the input data for the R analysis and the data files for the python codes are not included here, but can be found in the associated article: doi: 10.12688/f1000research.6536.1.
Results