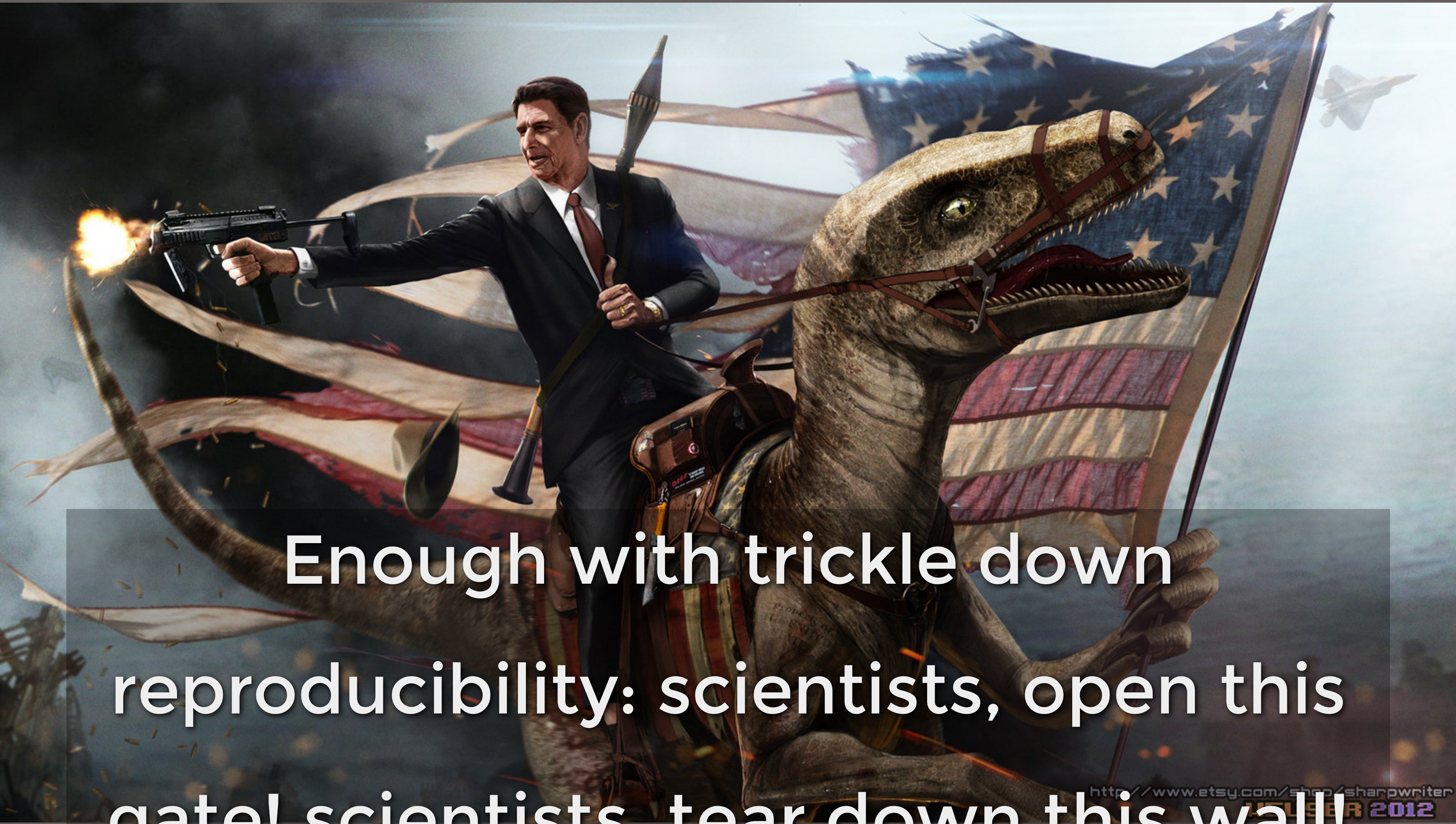


Enough with trickle down reproducibility: scientists, open this gate! scientists, tear down this wall!

How does reproducible research actually work in practice?

Karthik Ram

BIDS
BERKELEY INSTITUTE
FOR DATA SCIENCE

ROpenSci

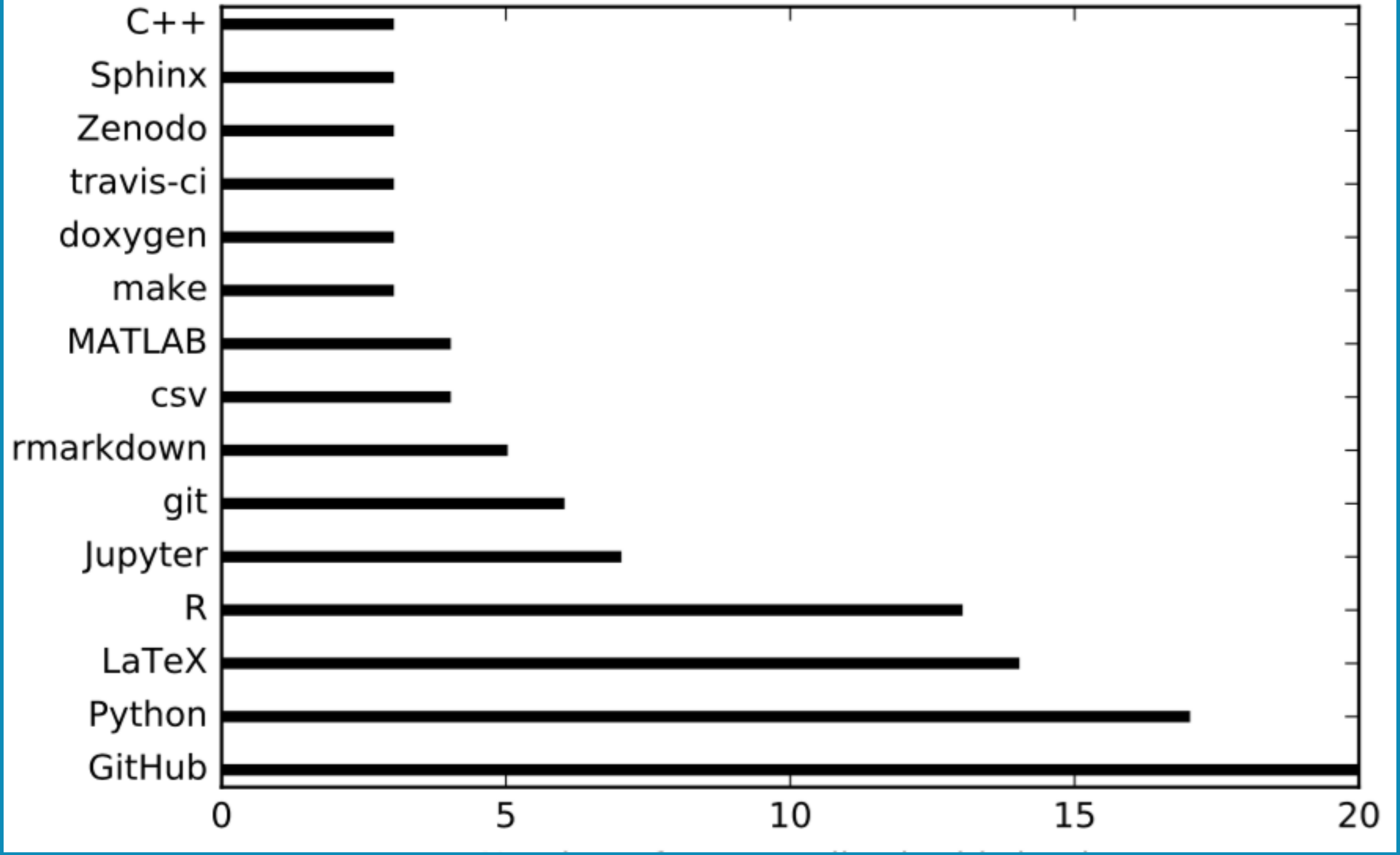
What exactly is reproducibility anyway?

4 kinds of reproducibility

Computational reproducibility and transparency	*Computational correctness*
Scientific reproducibility & transparency	*Statistical reproducibility*

Millman et al , 2016

What tools are scientists using?


C++
Sphinx
Zenodo
travis-ci
doxygen
make
MATLAB
csv
rmarkdown
git
Jupyter
R
LaTeX
Python
GitHub
0
5
10
15
20

Huff 2016

What are some obstacles around making research reproducible?

1 Leveling up skills

Biggest bottleneck to adoption of reproducible research practices was related to diversity of skills

More homogeneity in tool familiarity = better reproducibility

2 Dependencies, build systems, and packaging

Scientific software often built on numerous dependencies

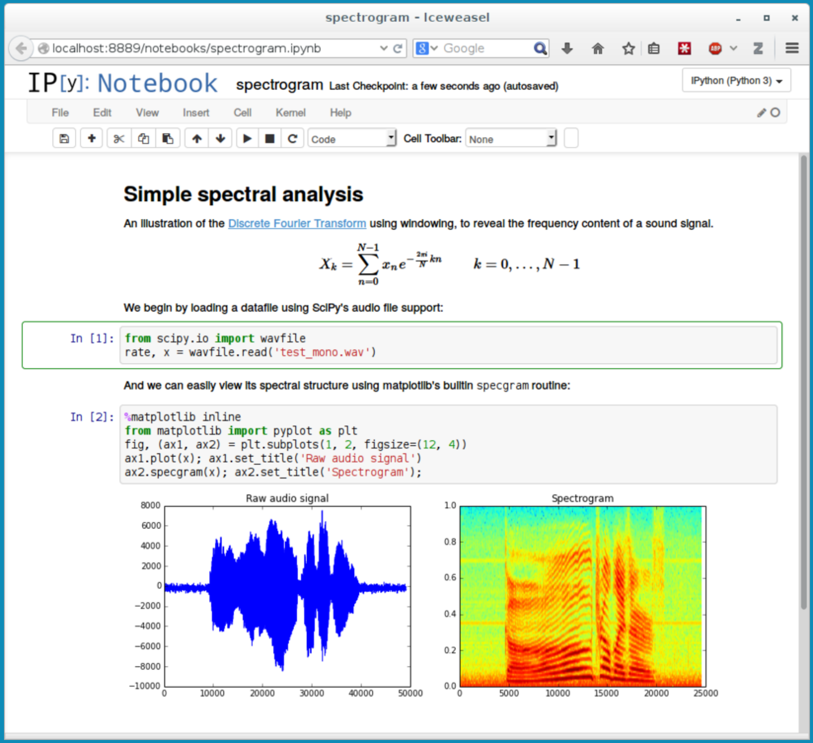
Improved build systems
for software, data &
workflows

3 Testing

Code that went beyond simple script reported testing systematically

However, many scientists were discouraged by the perceived effort of unit testing

4 Publishing



Simple spectral analysis

An illustration of the Discrete Fourier Transform using windowing, to reveal the frequency content of a sound signal.

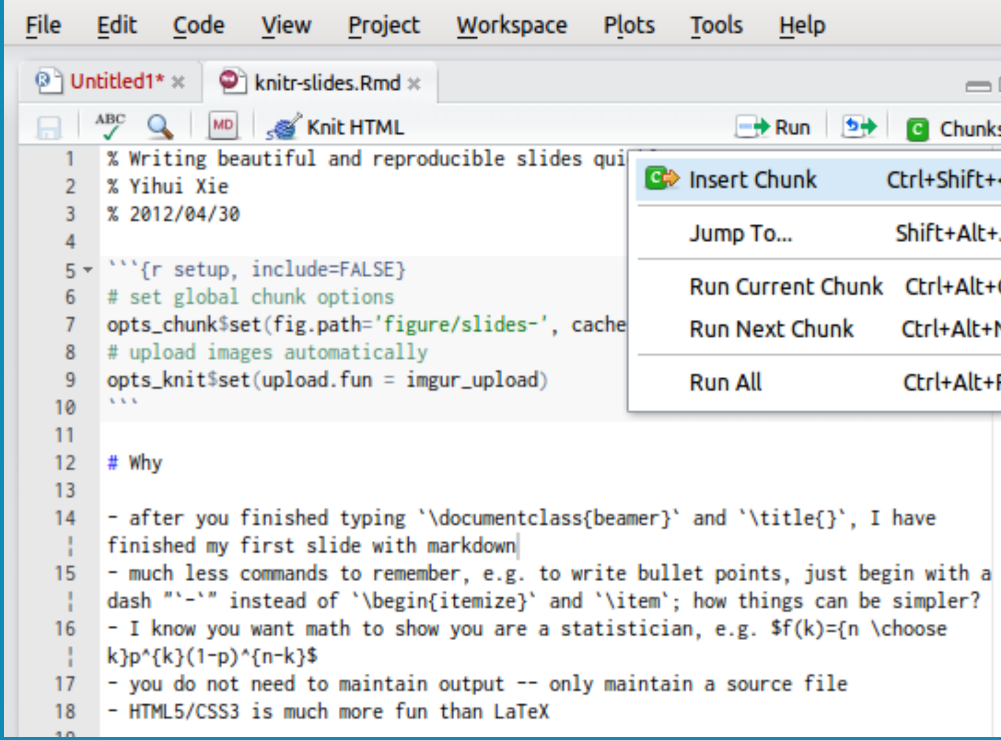
$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \qquad k = 0, \ldots, N-1$$

We begin by loading a datafile using SciPy's audio file support:

```
In [1]: from scipy.io import wavfile
        rate, x = wavfile.read('test_mono.wav')
```

And we can easily view its spectral structure using matplotlib's builtin `specgram` routine:

```
In [2]: %matplotlib inline
        from matplotlib import pyplot as plt
        fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 4))
        ax1.plot(x); ax1.set_title('Raw audio signal')
        ax2.specgram(x); ax2.set_title('Spectrogram');
```



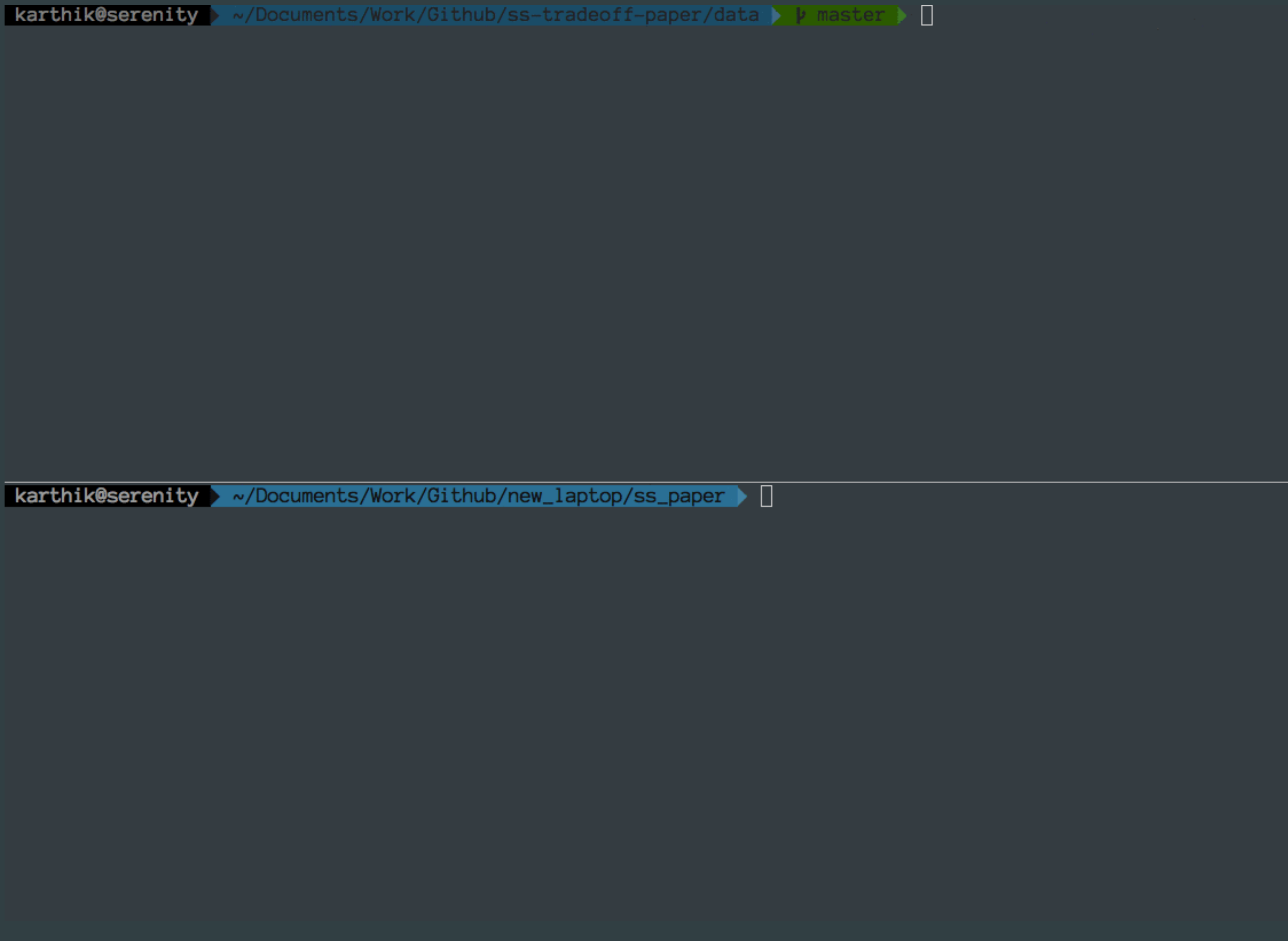
~~~
% Writing beautiful and reproducible slides qui
% Yihui Xie
% 2012/04/30

```{r setup, include=FALSE}
# set global chunk options
opts_chunk$set(fig.path='figure/slides-', cache
# upload images automatically
opts_knit$set(upload.fun = imgur_upload)
```


# Why

- after you finished typing `\documentclass{beamer}` and `\title{}`, I have finished my first slide with markdown
- much less commands to remember, e.g. to write bullet points, just begin with a dash "`-`" instead of `\begin{itemize}` and `\item`; how things can be simpler?
- I know you want math to show you are a statistician, e.g. $f(k)={n \choose k}p^{k}(1-p)^{n-k}$
- you do not need to maintain output -- only maintain a source file
- HTML5/CSS3 is much more fun than LaTeX
~~~

There is still a need for publication formats that allow for effortless collaboration.

5 Data sharing & versioning

Versioning data is hard, as is finding reliable places to archive them

zenodo



```
karthik@serenity  ~/Documents/Work/Github/ss-tradeoff-paper/data   master 
```

```
karthik@serenity  ~/Documents/Work/Github/new_laptop/ss_paper 
```

ROpenSci

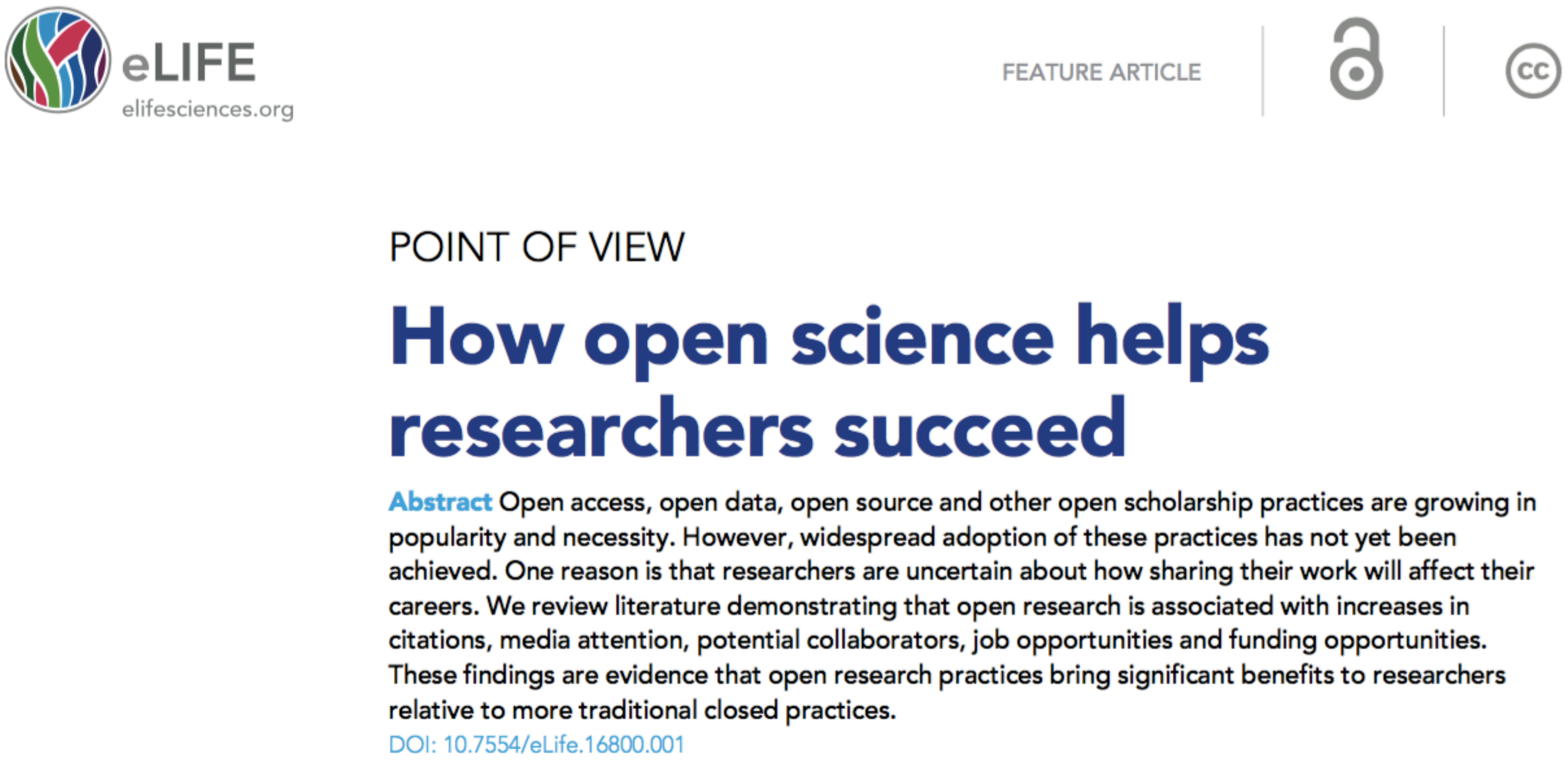
65+ TOOLS	R, C++, NODE	LARGE CONTRIBUTOR COMMUNITY

1. **Data retrieval** (APIs, data s torage services, journals)
2. **Data visualization** (e.g. plot.ly)
3. **Data sharing** (figshare, Zenodo, d at)
4. **Reproducibility**

6 Time and incentives

"time and efforts spent on creating reproducible research are not very well rewarded"

Ram & Marwick, 2016



eLIFE elifesciences.org

FEATURE ARTICLE

POINT OF VIEW

How open science helps researchers succeed

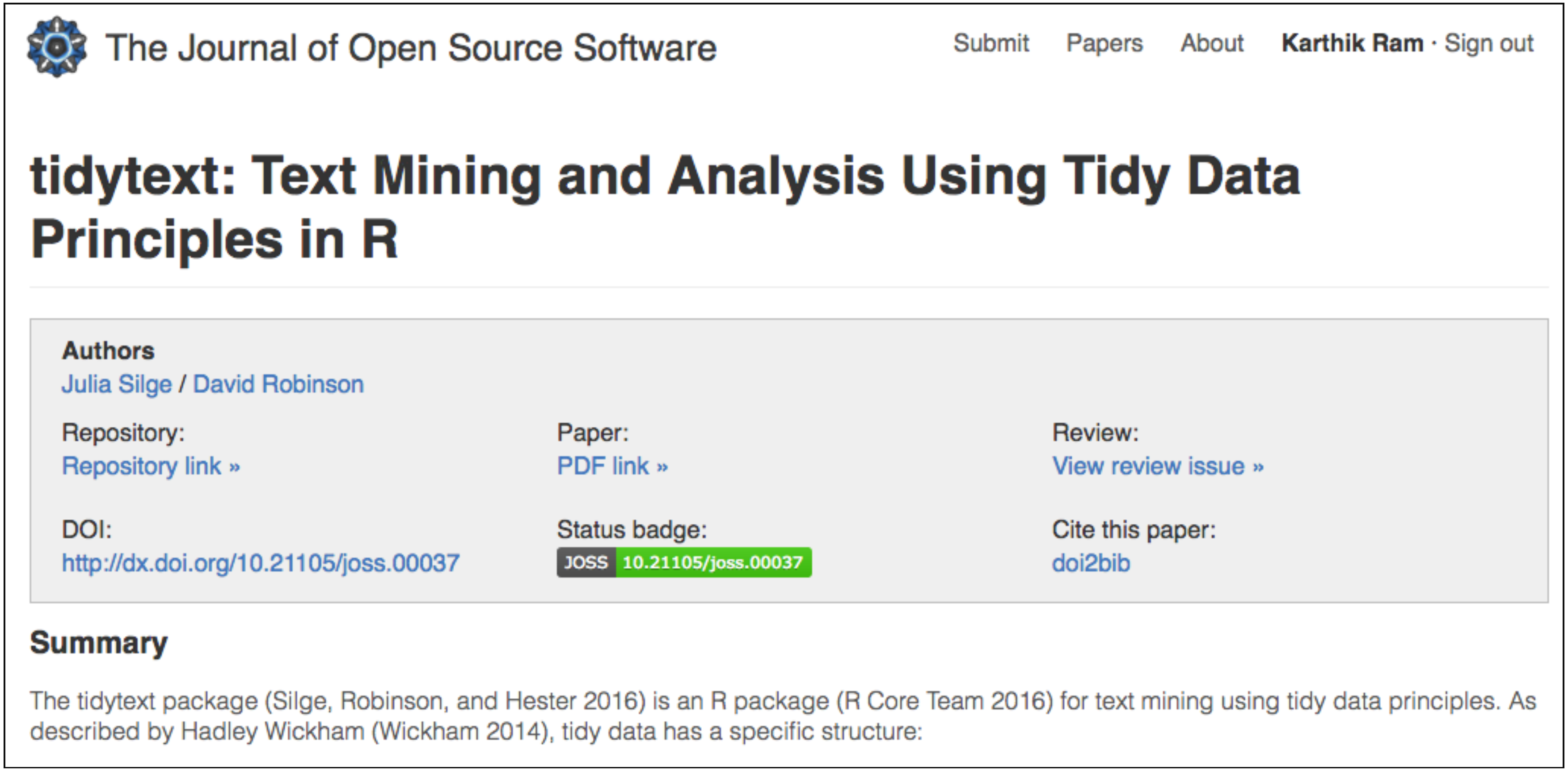
Abstract Open access, open data, open source and other open scholarship practices are growing in popularity and necessity. However, widespread adoption of these practices has not yet been achieved. One reason is that researchers are uncertain about how sharing their work will affect their careers. We review literature demonstrating that open research is associated with increases in citations, media attention, potential collaborators, job opportunities and funding opportunities. These findings are evidence that open research practices bring significant benefits to researchers relative to more traditional closed practices.

DOI: 10.7554/eLife.16800.001

10.7554/eLife.16800.001

Journal of Open Source Software

JOSS 10.21105/joss.00037



tidytext: Text Mining and Analysis Using Tidy Data Principles in R

Authors
Julia Silge / David Robinson

Repository:
Repository link »

Paper:
PDF link »

Review:
View review issue »

DOI:
http://dx.doi.org/10.21105/joss.00037

Status badge:
JOSS 10.21105/joss.00037

Cite this paper:
doi2bib

Summary

The tidytext package (Silge, Robinson, and Hester 2016) is an R package (R Core Team 2016) for text mining using tidy data principles. As described by Hadley Wickham (Wickham 2014), tidy data has a specific structure:

joss.theoj.org

Will reproducibility always be this hard?

Practices you can adopt now

Version your code

Automate everywhere

Open your **data**

Document your processes

Test everything

Avoid excessive dependencies

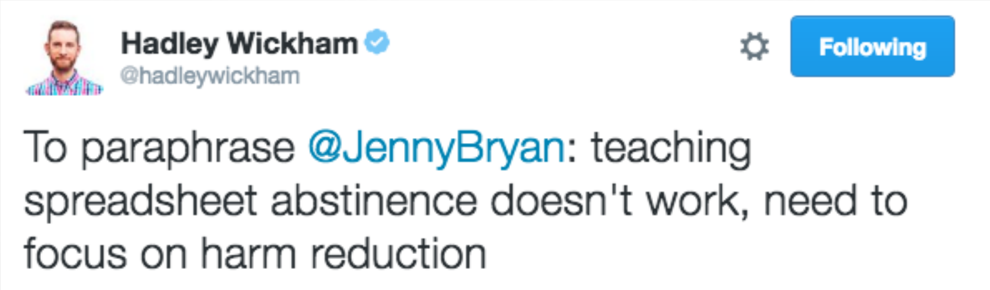
DOIs everywhere

Avoid spreadsheets *

Workflow and provenance

frameworks are hard to adopt

*



Hadley Wickham
@hadleywickham

To paraphrase @JennyBryan: teaching spreadsheet abstinence doesn't work, need to focus on harm reduction

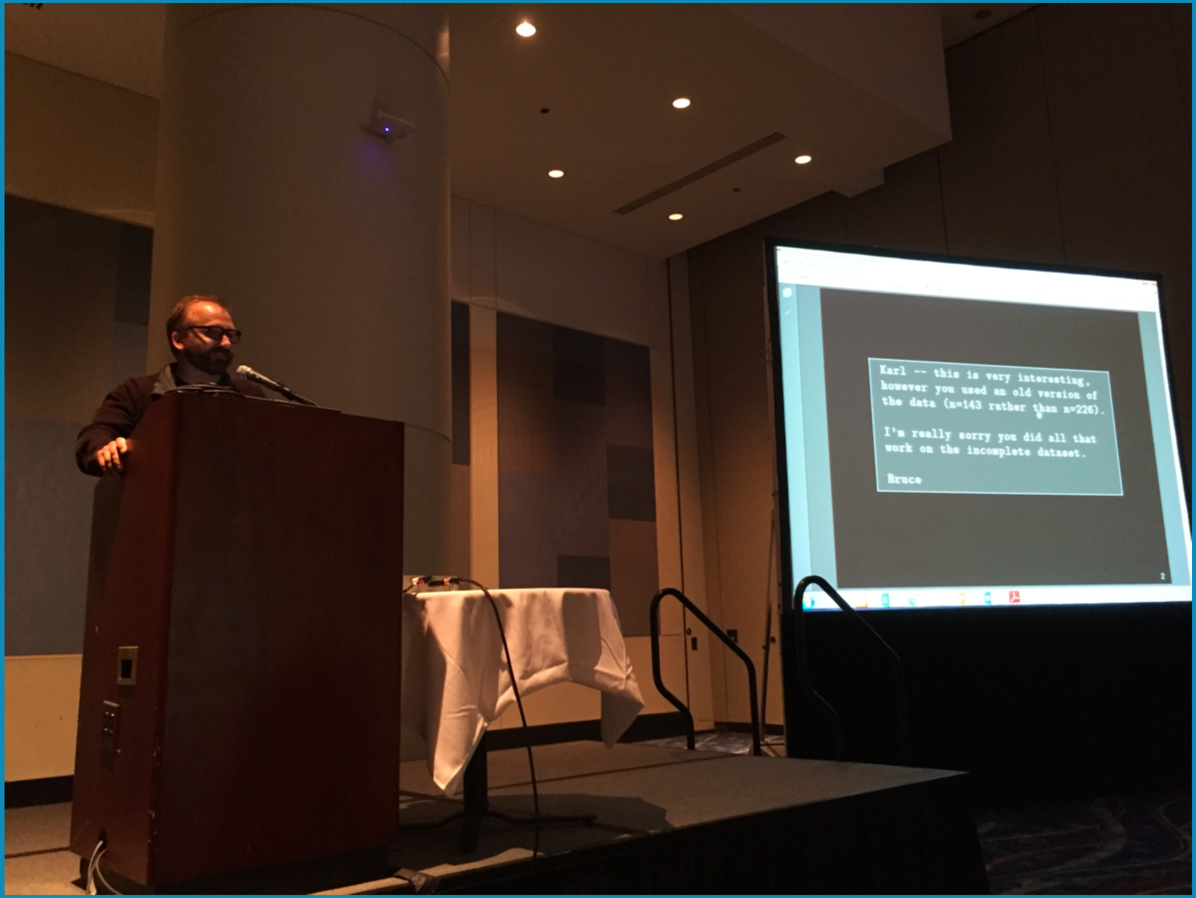
Partial reproducibility is better than nothing

Start small -- provide raw data, post any scripts, and versions of programs you used

Karl Broman

What we need right now is *scientists actually using stuff that already exists*, not *engineers building new stuff that no one will ever use*

C. Titus Brown



Karl -- this is very interesting, however you used an old version of the data (n=143 rather than n=226).

I'm really sorry you did all that work on the incomplete dataset.

Bruce

See previous talk by Karl Broman

kbroman.org/steps2rr/

BIDS
BERKELEY INSTITUTE
FOR DATA SCIENCE

The Practice of Reproducible Research

A collection of case studies to be published in spring 2017

Ben Marwick
Justin Kitzes
Katy Huff

BIDS
BERKELEY INSTITUTE
FOR DATA SCIENCE

Scott Chamberlain
Jeroen Ooms & rOpenSci
contributors

ROpenSci

inundata.org/talks/jsm2016