

# Building a Data Science Environment: A View from the Trenches

Juliana Freire  
New York University



# Some History

---

- Moore and Sloan foundations invited 15 universities to submit a proposal to create “data science environments”
- NYU, UC Berkeley and U of Washington were selected
  - 5-year, \$37.8 million cross-institutional collaboration



ALFRED P. SLOAN  
FOUNDATION



Berkeley  
UNIVERSITY OF CALIFORNIA



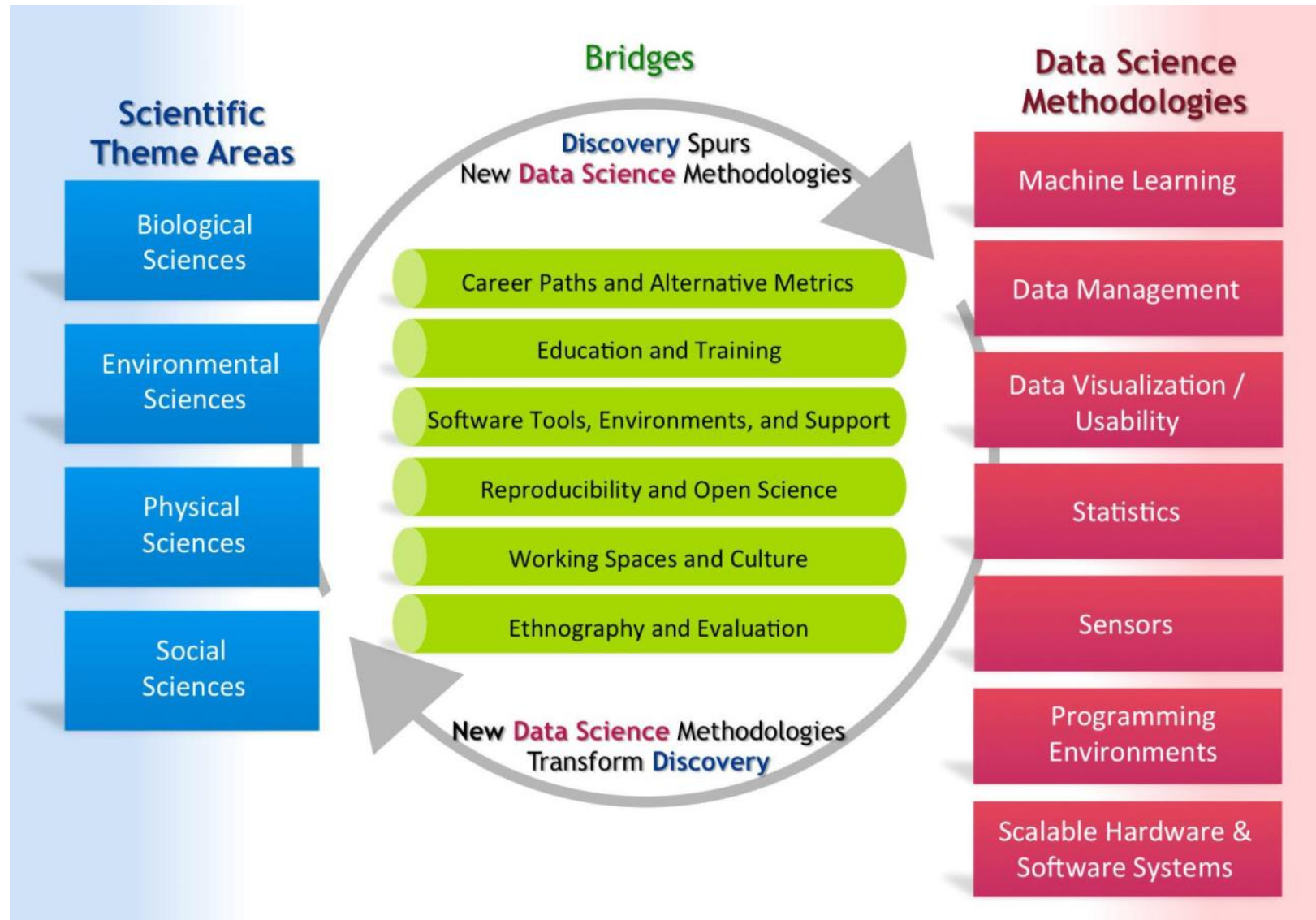


# Goals

---

- Transform the process of discovery and the institutional environments in which discovery takes place
  - Go beyond a small number of narrow successes, identifying and tackling a variety of impediments to the broad and sustainable adoption of data-intensive discovery
- Create a *virtuous cycle* in which advances in data science methodologies transform the process of discovery and drive new discoveries, while at the same time, the needs of scientific discovery stimulate the creation of new data science methodologies

# Challenges



# Approach

---

- Develop **sustained interactions and collaborations** between researchers from different disciplines to move science forward;
- Establish long-term **career paths** to retain scientists whose research is multi-disciplinary and unconventional; and
- Build an **ecosystem of analytical tools and research practices** that is sustainable, reusable, extensible, learnable, easy to translate across research areas and enables researchers to spend more time focusing on their science.

**Do breakthrough science, and enable  
breakthrough science**

<http://msdse.org>

People:  
The most important component



# Career Paths for Data Scientists

---

- Need professional data scientists with
  - Deep knowledge of data science methodology, e.g., data management, machine learning, visualization
  - Curiosity beyond their home disciplines and are energized by the opportunity to work across disciplines
- No career paths at universities beyond soft-money positions
- Intervention: Created prestigious positions that are academic alternatives to faculty pathways
- Attracted great talent!
- Challenges:
  - Identifying and attracting data scientists with a university salary
  - Balancing independence/autonomy and organizational obligations
  - Offering job security

# Different Positions

---

- *Research Engineers*: design, build, and deploy software as well as support data science across the University
- *Data Science Fellows*: independent scientists are expected to run interdisciplinary research programs
- *Post-doctoral researchers*: associated with existing research group – contribute data science expertise to the group's research program
- *Junior Data Scientists*: MS-level students assigned to work on incubator projects and advised by Research Engineers and faculty
- *Outreach staff*: inform broad NYU community about data science and best practices and help build new collaborations with the NYU MSDSE

# Interdisciplinary Faculty

---

- The NYU Center for Data Science has hired 7 interdisciplinary faculty members strong both in advancing data science methodology and in putting it to work in some field
- Developed a protocol for joint hires between CDS and domain departments
  - Address challenges related to different process and culture in different departments

# Many Benefits

---

- Critical mass in Data Science covering many different methodological areas and scientific domains
  - Helps in recruiting new personnel!
- Positively contributed both to the research and educational objectives of the University
- Enabled many new (funded) projects – lots of bridges with different NYU units
- Interaction with other researchers and students
  - Advising projects
  - Teaching courses and giving tutorials
  - Teaching best practices for open software development
  - Sharing experience in solving (real) data science problems



# Ongoing Work

---

- Collaborating with UW and UC Berkeley to institute data-science-specific criteria for promotion, tenure and merit increases
  - Explicit credit for open-source code and data releases
  - Technology transfer
- How to adapt current university rules and regulations regarding ladder-rank faculty members, adjuncts and lecturers, research scientists, and research staff to create rewarding career paths for data scientists?

# Software to Enable Data-Driven Science



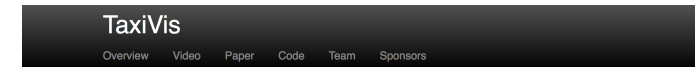
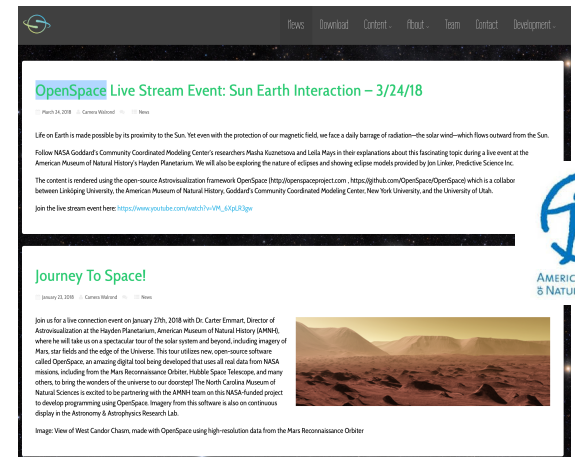
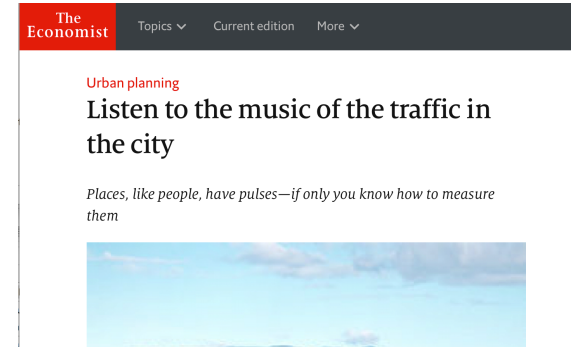
# Software Development

---

- Enable development, hardening, dissemination, and use of software tools and infrastructure that support data-driven research
  - Deliver high-impact, usable and generalizable software for science
  - Promote open-source software practices
  - Create, strengthen and deepen connections between tool developers and investigators in data-driven domains
- Challenges
  - Domain scientists are not equipped to develop and deliver the advanced software they require
  - Computer scientists have little incentive to harden, sustain, share, and integrate their techniques into a reusable software infrastructure
- Intervention: infrastructure for matchmaking and *personnel* dedicated to software development

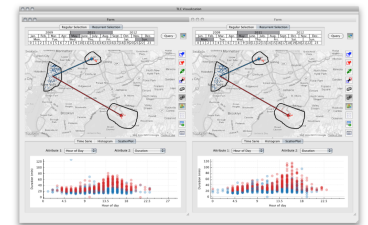
# Software Incubator

- Led by research engineers and faculty members
- Staffed by graduate students
- Some projects:
  - The Standard Cortical Observer (Medical School)
  - QuantEcon (Economics)
  - ADAGE (Physics)
  - Glacial Movement (Applied Math)
  - Community discovery in urban environments (Social Sciences)
  - Open Space (Astronomy, Museum of Natural History)
  - TaxiVis (NYC TLC and DoT)



## Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips

As increasing volumes of urban data are captured and become available, new opportunities arise for data-driven analysis that can lead to improvements in the lives of citizens through evidence-based decision making and policies. In this project, we focus on a particularly important urban data set: taxi trips. Taxis are valuable sensors and information associated with taxi trips can provide unprecedented insight into many different aspects of city life, from economic activity and human behavior to mobility patterns. But analyzing these data presents many challenges. The data are complex, containing geographical and temporal components in addition to multiple variables associated with each trip. Consequently, it is hard to specify exploratory queries and to perform comparative analyses (e.g., compare different regions over time). This problem is compounded due to the size of the data—there are on average 500,000 taxi trips each day in NYC. We propose a new model that allows users to visually query taxi trips. Besides standard analytics queries, the model supports origin-destination



Using TaxiVis to compare taxi trips from Lower Manhattan to JFK and LGA airports in May 2011.

# Impact

---

- Wider adoption of open-source tools developed by the group
- Several inter-disciplinary collaborations
- Educating researchers about software development practices and open-source policies/licenses
- Educating graduate students!

# Space



# Dedicated Space

---

- Programmatic: house people and hold activities
- Intellectual: space for cross-disciplinary collaboration
- Political: “neutral turf” – welcoming to all
- Cultivate the water cooler effect – serendipitous exchange of information that leads to collaboration and innovation
- Intervention:
  - Dedicated space for the Center for Data Science – not connected to any department or school
  - Mixed space adaptable to a range of activities
    - Quiet: Traditional offices, hoteling
    - Reconfigurable rooms of different sizes for meetings and seminars
    - (Vibrant) open space
    - Lounges/kitchens with espresso machine

# CDS Space

---

## **Before CDS Move (726 Broadway):**

63% Faculty did not visit CDS in a week

30% Satisfied access to light

75% Dissatisfied with amount of space

## **After CDS Move (60 5<sup>th</sup> Ave):**

86% Faculty visited CDS at least once/week

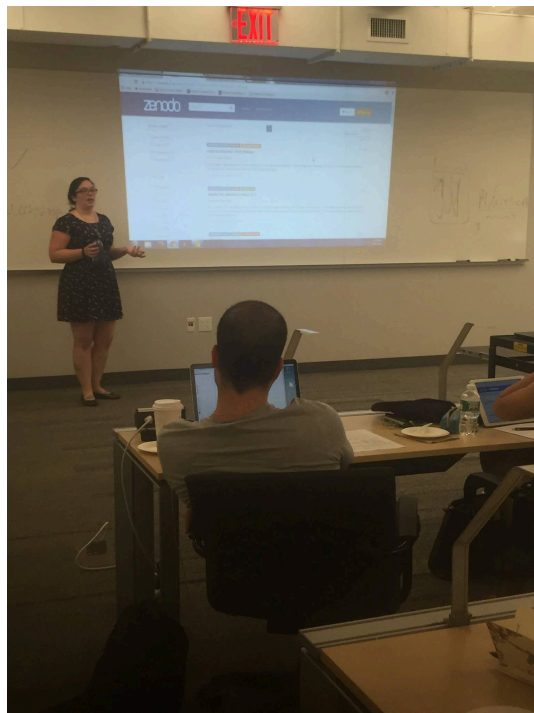
80% Satisfied access to light

92% Satisfied amount of space

More cohesive sense of community: Over 55% of CDS and MSDSE constituencies – students, faculty, staff, research scientists – come to the new CDS space four or more days per week.

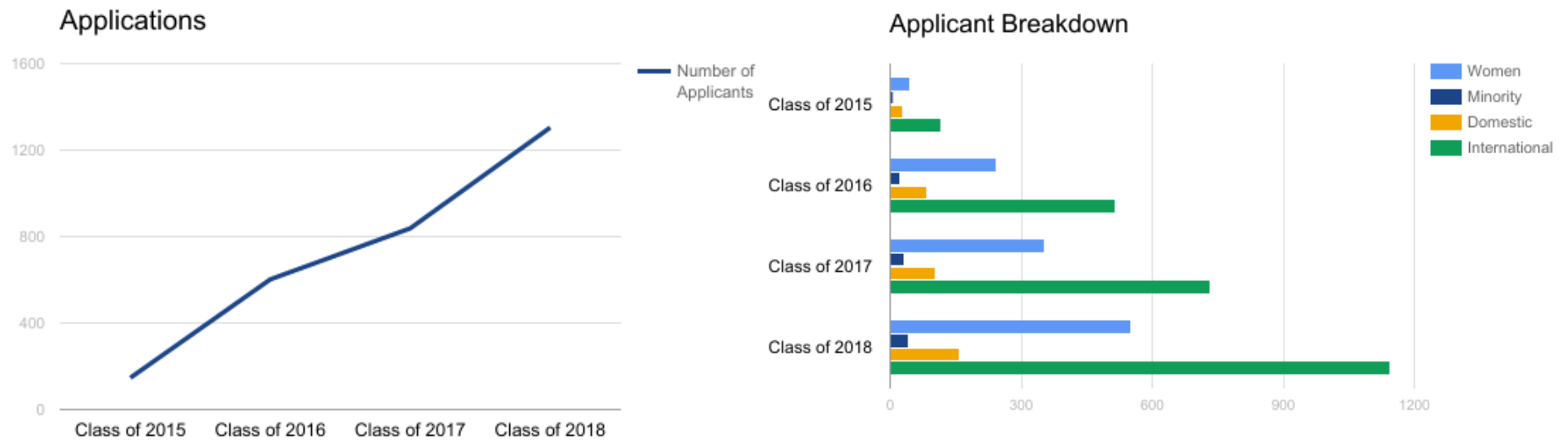


# Teaching and Training



# Formal Education

## MS in Data Science



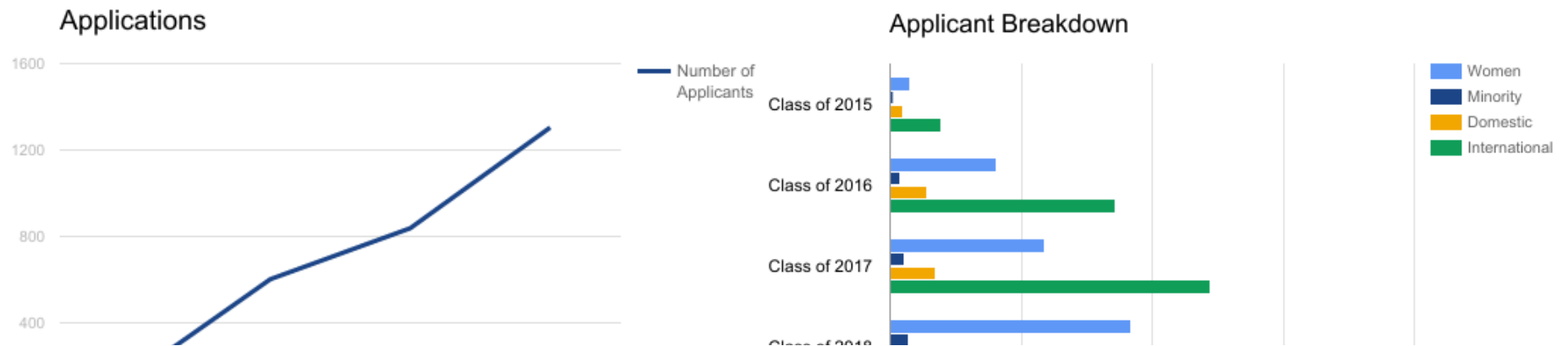
**2016-2017: 80 Matriculated Students**



PhD program started in 2017

# Formal Education

## MS in Data Science



Flexible curriculum

Tracks: Big data, NLP, Biology, Physics...



PhD program started in 2017

# Training

---

- Reach out to broad NYU community
- Data Carpentry (<http://www.datacarpentry.org>)
- Software Carpentry (<https://software-carpentry.org>)
- Reproducibility and data management tutorials (collaboration with NYU Libraries)
- Teaching through the incubator
  - Domain scientists: “teach a person to fish”
  - Graduate students learn to solve real problems and to work in an interdisciplinary environment

# Community Building and Outreach Activities



# Bringing People Together

---

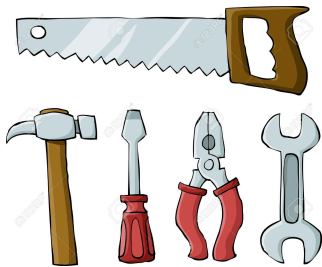
- Seminars and events: general and domain specific
- Data science lunches: informal gathering of NYU Data Science affiliated persons to discuss a broad range of data science related topics
- NYU-wide Data Science Showcases: inform the members of the broad NYU community about data science research and development, while at the same time, create new opportunities for collaboration
- AstroHackWeek: space-exploring informal educational event
- Text as Data (<https://cds.nyu.edu/text-data-speaker-series>)
- Math and Data (<http://mad.cds.nyu.edu>)
- ...

# Seed Grants

---

- Awards open to all NYU faculty, aiming to bring together data scientists and domain scientists to foster collaborations and generate new ideas (<https://cds.nyu.edu/nyu-data-science-seed-grant>)
- Project members become affiliated members of the MSDSE

# Reproducibility and Open Science



Tools and  
Infrastructure



Outreach and  
education



Incentives



# Tool Development

- ReproZip

<https://vida-nyu.github.io/reprozip/>

ReproZip

About

News

Documentation

GitHub Project

Examples

PyPI Packages

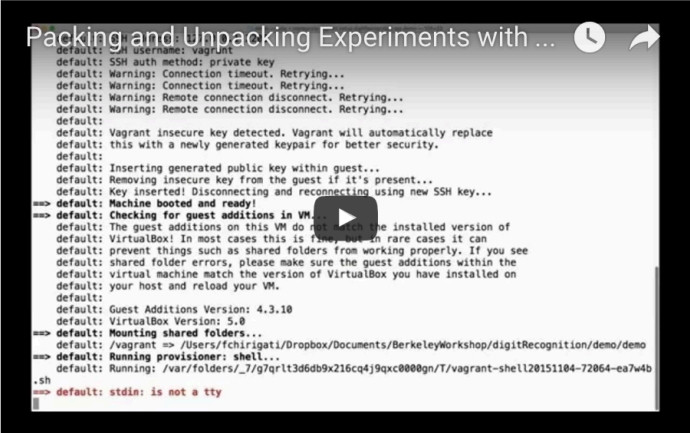
Installers

## Examples


Various examples of ReproZip packages, including instructions on how to reproduce them, are available in the [reprozip-examples](#) GitHub repository.

The following video shows how to use ReproZip to make your experiment reproducible. The example used in the video is based on [this blog post from B. Hanzra](#) on digit recognition using OpenCV and scikit-learn, and is available for [download here](#).


Packing and Unpacking Experiments with ...




## Tweets about reprozip



**ChrisAldrich** @ChrisAldrich  
#DtMH2016 @merbroussard: Reprozip was developed by NYU Center for Data and we're using it to save the code, data, and environment.  
14 Oct



**Juliana Freire** @jfreirenet  
ReproZip being used at NIST @nyuvida @nyupoly @MooreFound @SloanFoundation twitter.com/newgovrepos/st...  
09 Oct



**Vicky Steeves** @VickySteeves  
Hey all!!! We've added a graphical interface to #ReproZip!! Now you don't even have to use the command line to reproduce something  
twitter.com/pypi\_updates2/...

[Embed](#)

[View on Twitter](#)

<https://vida-nyu.github.io/reprozip/>

# ReproZip: Uses

- Recommended tool for the SIGMOD Reproducibility Evaluation and Information Systems Journal

The screenshot shows a web browser window with the URL [www.sciencedirect.com/science/article/pii/S0306437915301113](http://www.sciencedirect.com/science/article/pii/S0306437915301113). The page displays the article title "Reproducible experiments on dynamic resource allocation in cloud data centers" by Andreas Wolke, Martin Bichler, Fernando Chirigati, and Victoria Steeves. The article is from the journal "Information Systems", Volume 52, August–September 2015, Pages 83-95. The page includes a "Highlights" section with four bullet points, a "ReproZip" download link, and a list of attached data files. The "reprozip.rpz" file is highlighted with a red circle.

Download PDF Export Search ScienceDirect Advanced search

## Reproducible experiments on dynamic resource allocation in cloud data centers

Andreas Wolke<sup>a</sup>, Martin Bichler<sup>a</sup>, Fernando Chirigati<sup>b, 1</sup>, Victoria Steeves<sup>b, 1</sup>

[Show more](#)

<http://dx.doi.org/10.1016/j.is.2015.12.004> [Get rights and content](#)

**Refers To** Andreas Wolke, Boldbaatar Tsend-Ayush, Carl Pfeiffer, Martin Bichler  
**More than bin packing: Dynamic resource allocation strategies in cloud data centers**  
*Information Systems, Volume 52, August–September 2015, Pages 83-95*  
[Purchase PDF - \\$35.95](#)

**Highlights**

- Simulation and experimentation framework that allows an extensive evaluation of VM allocation strategies.
- Supports initial VM allocation controllers and dynamic controllers with VM migrations and random VM arrivals/departures.
- More than 200 time series data sets describing workload in enterprise data centers.
- The software framework allows for the design of new experiments and the replication of those published in Wolke et al. [1].

**Attached data files:**

- [Results.zip \(63 KB\)](#)
- [github.paper.IS2015-master.zip \(8 MB\)](#)
- [github.workload-master.zip \(222 MB\)](#)
- [Dockerfile \(1 KB\)](#)
- [IS2015.tar.gz \(1.5 GB\)](#)
- [reprozip.rpz \(160 MB\)](#)

[Evaluating REST architectures—Approach, tooling a...](#)  
2016, Journal of Systems and Software [more](#)

[View more articles »](#)

[Citing articles \(0\)](#)

[Related book content](#)

**Open Data with this article**

*Research data on Mendeley Data*

[Reproducible experiments on dynamic resource allocation in cloud data centers](#)  
In Wolke et al. we compare the efficiency of different resourc...

[Feedback](#)

# ReproZip: Uses

---

- Recommended tool for the SIGMOD Reproducibility Evaluation and Information Systems Journal
- Component of CORR: Cloud Infrastructure for storing, disseminating, federating and collaborating on Reproducible Record atoms



# ReproZip: Uses

---

- Recommended tool for the SIGMOD Reproducibility Evaluation and Information Systems Journal
- Component of CORR: Cloud Infrastructure for storing, disseminating, federating and collaborating on Reproducible Record atoms
- Snapshot research projects (Bonneau Lab, NYU)
- Archiving data journalism apps, e.g., <http://stackedup.org>
- And many more: <https://github.com/ViDA-NYU/reprozip-examples>

# Reproducibility Resources

---

- <http://www.reproduciblescience.org>
- UW: education material/tutorials
  - <http://uwescience.github.io/reproducible/git.html>
- ReproMatch

repromatch.engineering.nyu.edu/tools/search/

## ReproMatch

Welcome to ReproMatch!

Are you looking for a reproducibility tool that supports some specific functionality? Give **ReproMatch** a try!

ReproMatch stands for *Reproducibility Match* and it was designed to help you find the tool (or tools) that best matches your reproducibility needs. The tools in the ReproMatch catalog are classified according to different reproducibility tasks, which we organized in a taxonomy. Please see [Reproducibility Tasks](#) for a detailed description of this taxonomy.

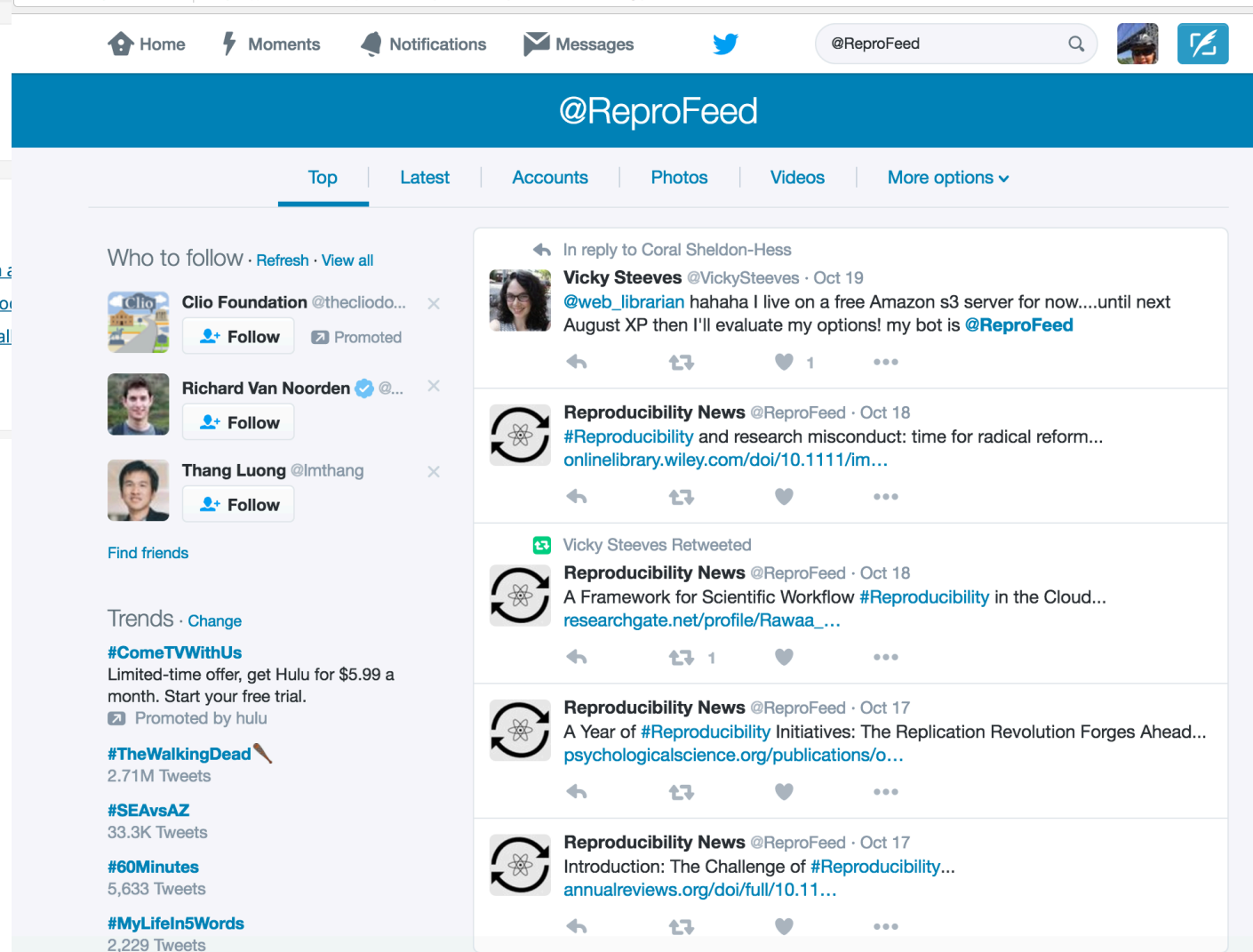
# Reproducibility Resources

<https://reproduciblescience.org/reproducibility-directory/>



## Reproducibility Directory

Twitter, Inc. [US] <https://twitter.com/search?q=%40ReproFeed&src=typd>



Twitter bot

# Training and Outreach

---

- University-wide courses and tutorials on tools (e.g., git, ReproZip)
- Office hours: students, staff, and faculty can ask questions about managing their research data, reproducibility
- Teaching reproducibility to graduate students
  - Reproducible and Collaborative Statistical Data Science (Philip Stark)
  - Applied Statistics: students reproduce the results in a paper and post non-anonymous reviews (Philip Stark)
  - Reproducibility modules in existing courses (e.g., Freire's Big Data course, Shasha's Database course)
  - Citing Code and Data, by Vicky Steeves  
<https://vickysteeves.github.io/DataScience-Citation-Workshop/#/>
  - NYU Data Services: <https://github.com/NYU-DataServices/>

# Training and Outreach

---

- Reproducibility case studies: The Practice of Reproducible Research
  - <https://bids.berkeley.edu/events/practice-reproducible-research-case-studies-and-lessons-data-intensive-sciences>
  - Diverse set of (computational) reproducible experiments from different areas
  - Collaboration between UCB, UW, NYU
  - Goals: serve as a guide for researchers – a collection of "reproducible workflows", lessons learned, tools used, etc.; understand in general what people have been doing towards reproducibility, what's missing, what are the gaps, challenges, and opportunities (based on the case studies and other references)



# Incentives

- Reproducibility badges ACM: incentive for authors

<http://www.acm.org/publications/policies/artifact-review-badging>



## k-Shape: Efficient and Accurate Clustering of Time Series

John Paparrizos  
Columbia University  
[jopa@cs.columbia.edu](mailto:jopa@cs.columbia.edu)



### ABSTRACT

The proliferation and ubiquity of temporal data across many disciplines has generated substantial interest in the analysis and mining of time series. Clustering is one of the most popular data mining methods, not only due to its exploratory power, but also as a preprocessing step or subroutine for other techniques. In this paper, we present *k*-Shape, a novel algorithm for time-series clustering. *k*-Shape relies on a scalable iterative refinement procedure, which creates homogeneous and well-separated clusters. As the clustering procedure

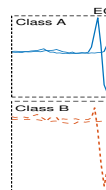


Figure 1:  
for the tw

### k-Shape: Efficient and Accurate Clustering of Time Series

Full Text: PDF Get this Article

Note: **Computationally Reproducible.** The experimental results of this paper were reproduced by a SIGMOD Review Committee and were found to support the central results reported in the paper. Details of the review process are found here: <http://db-reproducibility.seas.harvard.edu/#process>

Authors: [John Paparrizos](#) [Columbia University, New York, USA](#)  
[Luis Gravano](#) [Columbia University, New York, USA](#)

Published in:  
 **Proceeding**  
**SIGMOD '15** Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data  
Pages 1855-1870  
ACM New York, NY, USA ©2015  
[table of contents](#) ISBN: 978-1-4503-2758-9  
doi> [10.1145/2723372.2737793](https://doi.org/10.1145/2723372.2737793)



2015 Article

**Bibliometrics**

- Downloads (6 Weeks): 24
- Downloads (12 Months): 448
- Downloads (cumulative): 682
- Citation Count: 2

# Incentives

- Reproducibility badges ACM: incentive for authors

<http://www.acm.org/publications/policies/artifact-review-badging>

- ACM SIGMOD Most Reproducible Paper award

← → ↻ ⓘ sigmod2017.org/sigmod-call-for-research-papers/

Home

Important Dates

Organization

Events

Call for papers

Participating



**SIGMOD/PODS**  
**2017** *Raleigh, North Carolina*  
*Where Data Learn to Fly*

## Reproducibility

The authors of all accepted SIGMOD 2017 papers will have the option to submit their experiments to the Reproducibility Committee in order to obtain a “Reproducible Label” when the paper appears in the ACM Digital Library. Authors who want to prove the reproducibility of their results will submit data, code and scripts possibly wrapped in a virtual machine. Each paper will be reviewed by one reproducibility reviewer to verify that the experiments and minor variants of the experiments can be reproduced. Should the paper be successfully reproduced, it will be awarded with the “Reproducible Label”. The paper that is the easiest to reproduce successfully will receive the “Best Reproducible Paper Award” which comes with a \$1000 prize. More details and information about tools can be found here: <http://db-reproducibility.seas.harvard.edu>

# Incentives

---

- Reproducibility badges ACM  
<http://www.acm.org/publications/policies/artifact-review-badging>
- ACM SIGMOD Best Reproducible Paper award
- Companion papers implemented by ACM TOMS and Information Systems Journal -- incentive for reviewers
- In progress @NYU:
  - Faculty promotion and evaluation
  - Best reproducible dissertation (or paper?) award

# Collaboration with Industry

<https://cds.nyu.edu/partnership>

# Signals of Success

---

- Great science
  - Successful interdisciplinary research
  - Many best paper awards, cover of Science, Emmy, Sloan
  - Tools adopted by scientists and NYC agencies: SONYC, OpenSpace, ReproZip, TaxiVis
  - More than \$80M in grants awarded to MSDSE faculty and personnel
  - Lots of press (The Economist, NYTime, ...)
- Many new collaborations
- Critical mass and reputation helped us recruit top-notch faculty, researchers and staff
- Joint interdisciplinary faculty hires
  - ML+psychology; ML+linguistics; neural science; political science
- Most fellows and post-docs went on to academic positions



# What our Fellows Say

---

“Being at the MSDSE, and the academic freedom that comes with this, has allowed me to **focus on research** that I’m most interested in, which makes me most happy and productive. I’ve **also learned a lot from the people around me and the larger community, such as ideas on reproducibility and new software tools, which now feature more prominently in my research** than before coming to NYU”

“As a result of being at MSDSE, **my research and the corresponding data/software have been broadly disseminated.** All publications are available on my webpage. I have also engaged journalists to communicate this research to the public, with featured articles in the New York Times, Washington Post, Reuters, CBS News (and radio), NBC news, etc.”

“My own research has been impacted in particular by **meeting experts in applied statistics and machine learning, two fields that I had very little access** to before coming to NYU. ”

# What our Fellows Say

---

“Working here has **refined my research interests toward improving the state of tools and methods in my own field** (music information retrieval). Specifically, I’ve learned a great deal about how to design and produce high-quality scientific software. ”

“My appointment as a Moore-Sloan Data Science Fellow is having a very positive impact on my career. First, **it’s giving me the opportunity to devote time to research outputs that are not traditionally rewarded by hiring and tenure committees, such as the development of open-source research software and tools.** I have also taken this chance to **improve my knowledge of tools for interacting with large datasets,** which has significantly improved the efficiency of my research workflows. For example, with the help of a Data Science Master’s student, I was able to reduce the running time of the code in some of my research from days to just seconds using MapReduce.”

# Lessons Learned

---

- It is hard to create a data science culture at the University: Need to bring together researchers from different fields – making decisions and reaching consensus is challenging
- Need governance, transparent processes, and awareness about these challenges
  - Joint hires is now very easy!
- Space that is conducive to collaboration is essential
- Investment in outreach pays off to *translate* research
- Professional data scientists and research software engineers have been central to our success
- Multi-pronged approach to encouraging best practices in reuse and open science, including tools, incentives, and training is effective



Data science environments can  
transform science!

Obrigada  
благодаря  
Kiitos  
Merci  
धन्यवाद  
Thank you  
Tack  
Danke  
*Ευχαριστω*  
Bedankt