Democratizing Reproducibility

Juliana Freire
Computer Science & Engineering
Visualization and Data Analysis (ViDA)
Center for Data Science (CDS)
Center for Urban Science and Progress (CUSP)

Joint work with Fernando Chirigati, Remi Rampin,
Dennis Shasha, Cláudio Silva, Vicky Steeves, VisTrails Team

Why Reproducibility?

- Reproducibility is the cornerstone of science
- *"If I have seen further it is by standing on the shoulders of giants."*

  Isaac Newton
- Science is incremental and self-correcting
  - If we can't trust (or reproduce) previous results, we have to start over from scratch
- Reproducibility for *selfish reasons*: Increased impact, visibility [Vandewalle et al. 2009] and research quality [Begley and Ellis 2012]
  - Ability to stand on our own shoulders!

Why Reproducibility?

The New York Times

Science

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

ENVIRONMENT | SPACE & COSMOS

Nobel Laureate Retracts Two Papers Unrelated to Her Prize

By KENNETH CHANG
Published: September 23, 2010

Linda B. Buck, who shared the 2004 Nobel Prize
Medicine for deciphering the workings of the sens
retracted two scientific papers after she and her c
unable to repeat the findings.

The Economist

World politics | Business & finance

Unreliable research

Trouble at the lab

Scientists like to think of science as self-correctin

Oct 19th 2013 | From the print edition

NewScientist **Opinion**

search New Scientist | Go

Home | News | In-Depth Articles | Opinion | CultureLab | Galleries | Topic Guides | Last Word | Subscribe | Dating

SPACE | TECH | ENVIRONMENT | HEALTH | LIFE | PHYSICS&MATH | SCIENCE IN SOCIETY

Test drive our new beta website >

Home | Opinion | Health | Opinion

Is medical science built on shaky foundations?

- 17 September 2012 by Elizabeth Iorns
- Magazine issue 2882. Subscribe and save
- For similar stories, visit the Comment and Analysis Topic Guide

More than half of biomedical findings cannot be reproduced – we urgently need a way to ensure that discoveries are properly checked

REPRODUCIBILITY is the cornerstone of science. What we hold as definitive scientific fact has been tested over and over again. Even when a fact has been tested in this way, it may still be superseded by new knowledge. Newtonian mechanics became a special case of Einstein's general relativity; molecular biology's mantra "one gene, one protein" became a special case of DNA transcription and translation.

One goal of scientific publication is to share results in enough detail to allow other research teams to reproduce them and build on them. However, many recent reports have raised the alarm that a shocking amount of the published literature in fields ranging from cancer biology to psychology is not

Like 1.4k | Tweet 338 | +1 73

Share 303

NYU TANDON SCHOOL OF ENGINEERING

Why Reproducibility?

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2 × 2 table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let *R* be the ratio of the number of "true relationships" to "no relationships" among those tested in the field. *R*

The Essay section contains opinion pieces on topics of broad interest to a general medical audience.

COMM

AVIAN INFLUENZA Shift expertise to track mutations where they emerge p.534

EARTH SYSTEMS Past climates give valuable clues to future warming p.537

S. GSCHMEISSNER/SPL

Many Ian

CORRESPONDENCE

LINK TO ORIGINAL ARTICLE

Believe it or not: how much can we rely on published data on potential drug targets?

Florian Prinz, Thomas Schlange and Khusru Asadullah

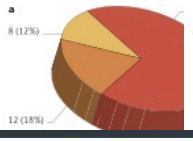
A recent report by Arrowsmith noted that the success rates for new development projects in Phase II trials have fallen from 28% to 18% in recent years, with insufficient efficacy being the most frequent reason for failure (Phase II failures: 2008–2010. *Nature Rev. Drug Discov.* **10**, 328–329 (2011))[1]. This indicates the limitations of the predictivity of disease models and also that the validity of the targe investigated is frequently questionabl is a crucial issue to address if success clinical trials are to be improved.

Candidate drug targets in indu derived from various sources, inclu house target identification campai licensing and public sourcing, in pa based on reports published in the litera presented at conferences. During the of projects from an academic to a c setting, the focus changes from 'inte

to 'feasible/marketable', and the financial costs of pursuing a full-blown drug discovery and development programme for a particular target could ultimately be hundreds of millions of Euros. Even in the earlier stages, investments in activities such as high-throughput screening programmes are substantial, and thus the validity of published data on potential targets

results that are published are hard to reproduce. However, there is an imbalance between this apparently widespread impression and its public recognition (for example, see REFS 2,5), and the surprisingly few scientific publications dealing with this topic. Indeed, to our knowledge, so far there has been no published in-depth, systematic analysis that compares reproduced results with published results for wet-lab experiments related to target identification and validation.

Early research in the pharmaceutical industry, with a dedicated budget and scientists who mainly work on target validation to increase the confidence in a project, provides a unique opportunity to generate a broad data set on the reproducibility of published data. To substantiate our incidental observations that published reports are frequently not reproducible with



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For A

News & Comment › News › 2015 › June › Article

NATURE | NEWS

Irreproducible biology research costs put at $28 billion per year

THE CHRONICLE OF HIGHER EDUCATION research in the United States.

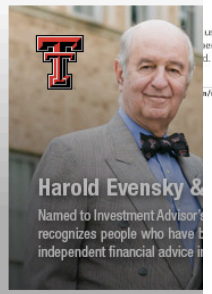
June 26, 2015

Home | News | Global | Opinion & Ideas | Facts & Figures | Blogs | Advice | Forums | Jobs | Search The Chr

Research

March 16, 2015

Amid a Sea of False Findings, the NIH Tries Reform



Harold Evensky &

Named to Investment Advisor

recognizes people who have

independent financial advice

Get Academ

NYU TANDON SCHOOL OF ENGINEERING

Why Not Reproducibility?

- Intellectual property concerns: patents/commercial prospects
- It is too hard and too time consuming

*"authors have complained that the process **requires too much work for the benefit derived**"*

Bonnet et al., SIGMOD Record 2011

***"Insufficient time** is the main reason why scientists do not make their data and experiment available and reproducible."*

Carol Tenopir, Beyond the PDF 2 Conference

***"77%** claim that they do not have **time to document and clean up the code**."*

Victoria Stodden, Survey of the Machine Learning Community – NIPS 2010

***"**It would require **huge amount of effort** to make our code work with the latest versions of these tools."*

Collberg et al., Repeatability and Benefaction in Computer Systems Research, University of Arizona TR 14-04

NYU TANDON SCHOOL OF ENGINEERING

Methods and Tools

Managing Rapidly-Evolving Scientific Workflows

Juliana Freire, Cláudio T. Silva, Steven P. Callahan,
Emanuele Santos, Carlos E. Scheidegger, and Huy T. Vo

University of Utah

Abstract. We give an overview of VisTrails, a system that provides an infrastructure for systematically capturing detailed provenance and streamlining the data exploration process. A key feature that sets VisTrails apart from previous visualization and scientific workflow systems is a novel action-based mechanism that uniformly captures provenance for data products and workflows used to generate these products. This mechanism not only ensures reproducibility [IPAW 2006] plifies data exploration by allowing scientists the space of workflows and parameter settings for an exploration task.

Querying and Re-Using Workflows with VisTrails

Carlos E. Scheidegger Huy T. Vo David Koop Juliana Freire Cláudio T. Silva
SCI Institute and School of Computing – University of Utah
{cscheid, hvo, dakoop, juliana, csilva}@cs.utah.edu

ABSTRACT

We show how workflow systems can be augmented to leverage provenance information to enhance usability. In particular, we will demonstrate new mechanisms and intuitive user interfaces designed to allow users to query workflows by example and to refine workflows by analogies. These techniques are implemented in VisTrails, an open-source provenance-enabled scientific workflow system that can be combined with a wide range of tools, libraries, and visualization systems. We will show different scenarios where these techniques can be used to simplify the notoriously hard tasks of creating and refining workflows.

Who created this data product and when? When was it modified and by whom? What was the process used to create the data product? Were two data products derived from the same raw data? Not only is the process time-consuming, but also error-prone.

Workflow systems have therefore grown in popularity within the scientific community (see e.g., [3, 8, 9]). Not only do they support the automation of repetitive tasks, but they can al
of det
for th [ACM SIGMOD 2008]

Wh

Querying and Creating Visualizations by Analogy

Carlos E. Scheidegger, Huy T. Vo, David Koop, Juliana Freire, *Member, IEEE*, and Cláudio T. Silva,

Abstract— While there have been advances in visualization systems, particularly in multi-view visualizations and visual ex
process of building visualizations remains a major bottleneck in data exploration. We show that provenance metadata col
the creation of pipelines can be reused to suggest similar content in related visualizations and guide semi-automated
introduce the idea of query-by-example in the context of an ensemble of visualizations, and the use of analogies as first-cla
in a system to guide scalable interactions. W hniques in VisTrails, a publi
open-source system.

[IEEE TVCG 2007]

Index Terms—visualization systems, query-by

The ALPS project release 2.0: open source software for strongly correlated systems

B Bauer[1], L D Carr[2], H G Evertz[3], A Feiguin[4], J Freire[5], S Fuchs[6], L Gamper[1], J Gukelberger[1], E Gull[7], S Guertler[8]
Show full author list

Published 4 May 2011 • IOP Publishing Ltd
Journal of Statistical Mechanics: Theory and Experiment, Volum

[JSTAT 2011]

https://github.com/VisTrails

https://github.com/ViDA-NYU/reprozip

http://www.crowdlabs.org/

DOI: 10.1111/j.1467-8659.2010.01830.x

COMPUTER GRAPHICS *forum*
Volume 30 (2011), number 1 pp. 75–84

Using VisTrails and Provenance for Teaching Scientific Visualization

[CGF 2011]

Cláudio T. Silva, Erik Anderson, Emanuele Santos

NYU TANDON SCHOOL OF ENGINEERING

Methods and Tools

Managing Rapidly-Evolving Scientific Workflows

Juliana Freire, Cláudio T. Silva, Stev
Emanuele Santos, Carlos E. Scheidegge

University of Utah

Abstract. We give an overview of VisTrails,
an infrastructure for systematically capturing
streamlining the data exploration process. A k
Trails apart from previous visualization and sci
is a novel action-based mechanism that uniform
for data products and workflows used to gener
mechanism not only ensures reproducibility
plifies data exploration by allowing scientists
the space of workflows and parameter settings

[IPAW 2006]

Querying and Re-Using Workflows with VisTrails

Carlos E. Scheidegger Huy T. Vo David Koop Juliana Freire Cláudio T. Silva

School of Computing – University of Utah
{dakoop, juliana, csilva}@cs.utah.edu

lever-
In par-
tuitive
rkflows
These
-source
can be
ualiza-
tion systems. We will show different scenarios where these techniques can be used to simplify the notoriously hard tasks of creating and refining workflows.

Who created this data product and when? When was it modified and by whom? What was the process used to create the data product? Were two data products derived from the same raw data? Not only is the process time-consuming, but also error-prone.

Workflow systems have therefore grown in popularity within the scientific community (see e.g., [3, 8, 9]). Not only do they support the automation of repetitive tasks, but they can al
of det
for th

[ACM SIGMOD 2008]

50+ papers
2 best-paper awards
3 PhD dissertations
Open-source systems

Querying and Creating Visualizations by Analogy

Carlos E. Scheidegger, Huy T. Vo, David Koop, Juliana Freire, *Member, IEEE*, and Cláudio T. Silva,

Abstract— While there have been advances in visualization systems, particularly in multi-view visualizations and visual ex
process of building visualizations remains a major bottleneck in data exploration. We show that provenance metadata col
the creation of pipelines can be reused to suggest similar content in related visualizations and guide semi-automated
introduce the idea of query-by-example in the context of an ensemble of visualizations, and the use of analogies as first-cla
in a system to guide scalable interactions. W hniques in VisTrails, a publi
open-source system.

Index Terms—visualization systems, query-by

[IEEE TVCG 2007]

The ALPS project release 2.0: open source software for strongly correlated systems

B Bauer[1], L D Carr[2], H G Evertz[3], A Feiguin[4], J Freire[5], S Fuchs[6], L Gamper[1], J Gukelberger[1], E Gull[7], S Guertler[8]

Show full author list

Published 4 May 2011 • IOP Publishing Ltd

Journal of Statistical Mechanics: Theory and Experiment, Volum

[JSTAT 2011]

https://github.com/VisTrails

https://github.com/ViDA-NYU/reprozip

http://www.crowdlabs.org/

DOI: 10.1111/j.1467-8659.2010.01830.x

COMPUTER GRAPHICS *forum*
Volume 30 (2011), number 1 pp. 75–84

Using VisTrails and Provenance for Teaching Scientific Visualization

Cláudio T. Silva, Erik Anderson, Emanuele Santos

[CGF 2011]

NYU TANDON SCHOOL OF ENGINEERING

Methods and Tools

Managing Rapidly-Evolving Scientific Workflows

Juliana Freire, Cláudio T. Silva, Stev
Emanuele Santos, Carlos E. Scheidegge

University of Utah

Abstract. We give an overview of VisTrails, an infrastructure for systematically capturing streamlining the data exploration process. A k Trails apart from previous visualization and sci is a novel action-based mechanism that uniforr for data products and workflows used to gener mechanism not only ensures reproducibility plifies data exploration by allowing scientists the space of workflows and parameter settings

[IPA

Querying and Creating Vis

Carlos E. Scheidegger, Huy T. Vo, David Koop, Juliana Freir

Abstract— While there have been advances in visualization systems, particularly in multi-view visualizations and visual ex process of building visualizations remains a major bottleneck in data exploration. We show that provenance metadata col the creation of pipelines can be reused to suggest similar content in related visualizations and guide semi-automated introduce the idea of query-by-example in the context of an ensemble of visualizations, and the use of analogies as first-cla in a system to guide scalable interactions. hniques in VisTrails, a publi open-source system.

Index Terms—visualization systems, query-by

[IEEE TVCG 2007]

DOI: 10.1111/j.1467-8659.2010.01830.x

COMPUTER GRAPHICS *forum*
Volume 30 (2011), number 1 pp. 75–84

Using VisTrails and Provenance for Teaching Scientific Visualization

Cláudio T. Silva, Erik Anderson, Emanuele Santos

[CGF 2011]

50+ papers
2 best-paper awards

MISSION: ACCOMPLISHED

Querying and Re-Using Workflows with VisTrails

Carlos E. Scheidegger Huy T. Vo David Koop Juliana Freire Cláudio T. Silva

chool of Computing – University of Utah
{dakoop, juliana, csilva}@cs.utah.edu

lever-
In par-
ituitive
rkflows
These
-source
can be
sualiza-
s where these
isly hard tasks

Who created this data product and when? When was it modified and by whom? What was the process used to create the data product? Were two data products derived from the same raw data? Not only is the process time-consuming, but also error-prone.

Workflow systems have therefore grown in popularity within the scientific community (see e.g., [3, 8, 9]). Not only do they support the automation of repetitive tasks, but they can al
of det
for th

[ACM SIGMOD 2008]

ct release 2.0: open source software for strongly ems

B Bauer[1], L D Carr[2], H G Evertz[3], A Feiguin[4], J Freire[5], S Fuchs[6], L Gamper[1], J Gukelberger[1], E Gull[7], S Guertler[8]
Show full author list

Published 4 May 2011 • IOP Publishing Ltd
Journal of Statistical Mechanics: Theory and Experiment, Volum

[JSTAT 2011]

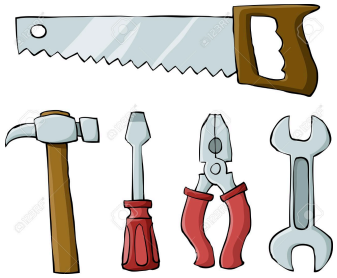
https://github.com/VisTrails

https://github.com/ViDA-NYU/reprozip

http://www.crowdlabs.org/

NYU TANDON SCHOOL OF ENGINEERING

Democratizing Reproducibility



Tools

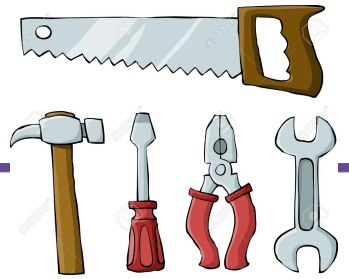
Outreach and education

Incentives

- Make reproducibility really easy → more people will do it!
- Develop and disseminate best practices
- Create and institutionalize an incentive structure

NYU TANDON SCHOOL OF ENGINEERING

Building Tools



- Academic software built by PhD students is a good start
- Tools must be robust, easy to install and use
  - A lot of work that does not lead to publications

ReproZip: Using Provenance to Support Computational Reproducibility

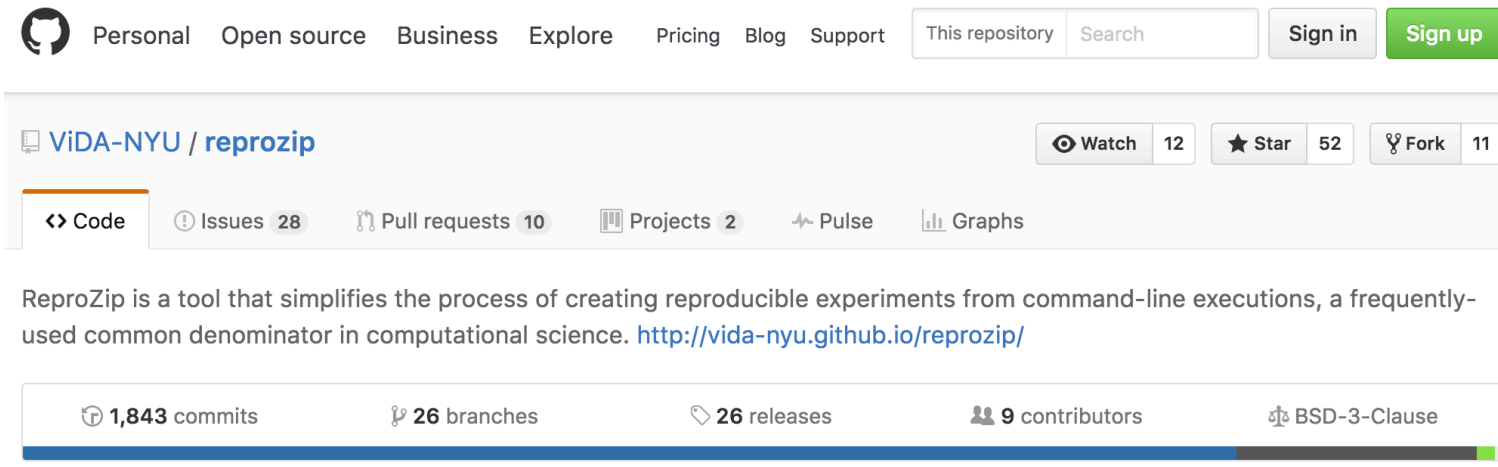
Fernando Chirigati
Polytechnic Institute of NYU
fchirigati@nyu.edu

Dennis Shasha
New York University
shasha@courant.nyu.edu

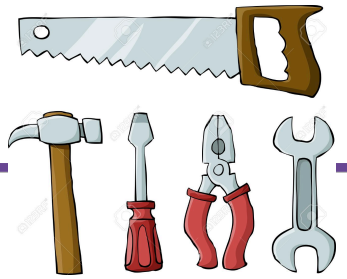
Juliana Freire
Polytechnic Institute of NYU
julic

[TAPP 2013]

Need professional software developers



Building Tools



www.sciencedirect.com/science/article/pii/S0306437915301113

Download PDF | Export | Search ScienceDirect | Advanced search

Reproducible experiments on dynamic resource allocation in cloud data centers

Andreas Wolke[a], Martin Bichler[a], Fernando Chirigati[b, 1], Victoria Steeves[b, 1]

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].

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 cloud data centers
In Wolke et al. we compare the efficiency of diff

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 (1.3 GB)
- reprozip.rpz (160 MB)

Feedback

New Gov Repos @newgovrepos Following

usnistgov/corr-reprozip is a new @github repo by @usnistgov at ift.tt/2drXeEL: An integrated version of the reprozip code to CoRR.

usnistgov/corr-reprozip
corr-reprozip - An integrated version of the reprozip code to CoRR.
github.com

Personal Open source Business Explore Pricing Blog Support This repository Search Sign in Sign up

ViDA-NYU / reprozip

Watch 12 Star 52 Fork 11

Code Issues 28 Pull requests 10 Projects 2 Pulse Graphs

ReproZip is a tool that simplifies the process of creating reproducible experiments from command-line executions, a frequently-used common denominator in computational science. http://vida-nyu.github.io/reprozip/

1,843 commits 26 branches 26 releases 9 contributors BSD-3-Clause

Outreach and Education

- 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 best practices to graduate students
  - 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/
- Resources
  - http://repromatch.engineering.nyu.edu
  - http://reproduciblescience.org

Outreach and Education

- University-wide courses and tutorials on tools (e.g., git, ReproZip)
- Office ho stions about managin
- Teaching
  - Reprodu s Big Data course,
  - Citing C https://v kshop/#/
  - NYU Data Services: https://github.com/NYU-DataServices/
- Resources
  - http://repromatch.engineering.nyu.edu
  - http://reproduciblescience.org

Need dedicated outreach staff

Many benefits beyond education:

- Bridges built across NYU units and with other universities
- New use cases for our tools
- Wider adoption

NYU TANDON SCHOOL OF ENGINEERING

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

g

ACM DL DIGITAL LIBRARY

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 homoge-

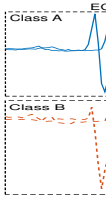


Figure 1:
for the tw

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

2015 Article

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

NYU TANDON SCHOOL OF ENGINEERING

Incentives

- Reproducibility badges ACM: incentive for authors

  http://www.acm.org/publications/policies/artifact-review-badging
- ACM SIGMOD Best 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

UP

Incentives

- Reproducibility badges ACM: incentive for authors

  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

Incentives

- Reproducibility badges ACM: incentive for authors

  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:
  - RPT and faculty evaluation
  - Reproducible dissertation award

Lessons Learned

- Need a multi-pronged approach to democratize reproducibility: tools, training, incentives
- Collaboration across MSDSE partners has been instrumental to our success
- To build high-quality tools, we need research engineers
- Investment in outreach and education have a high payoff
- Institutional/cultural changes take time, but they happen
- Reproducibility is just the beginning, better science will follow!

Acknowledgements

- This work is partially supported by the National Science Foundation grants CNS 1405927, III 1139832, IIS 1050422, IIS 0905385, IIS 0844572, IIS 0746500, CNS 0751152; an IBM Faculty Award; Moore Foundation; Sloan Foundation

GORDON AND BETTY MOORE FOUNDATION

ALFRED P. SLOAN FOUNDATION

National Science Foundation
WHERE DISCOVERIES BEGIN

IBM

NYU TANDON SCHOOL OF ENGINEERING

Obrigada

Gracias

благодаря

Kiitos

Merci

धन्यवाद

Thank you

Tack

Danke

Ευχαριστω

Bedankt