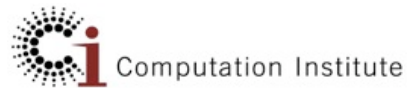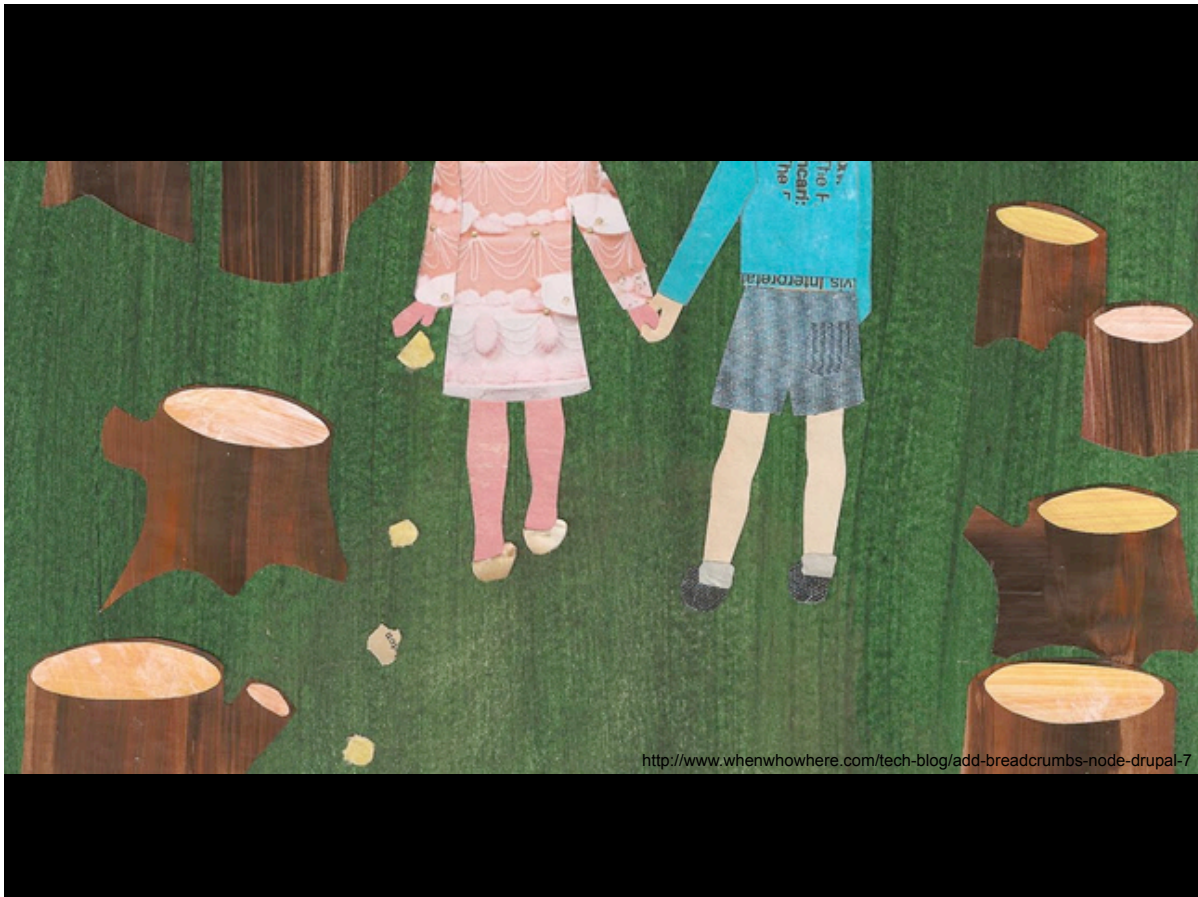# Fairly sharing the costs of Reproducibility:
Precedents and Possibilities

Gordon Kindlmann

glk@uchicago.edu
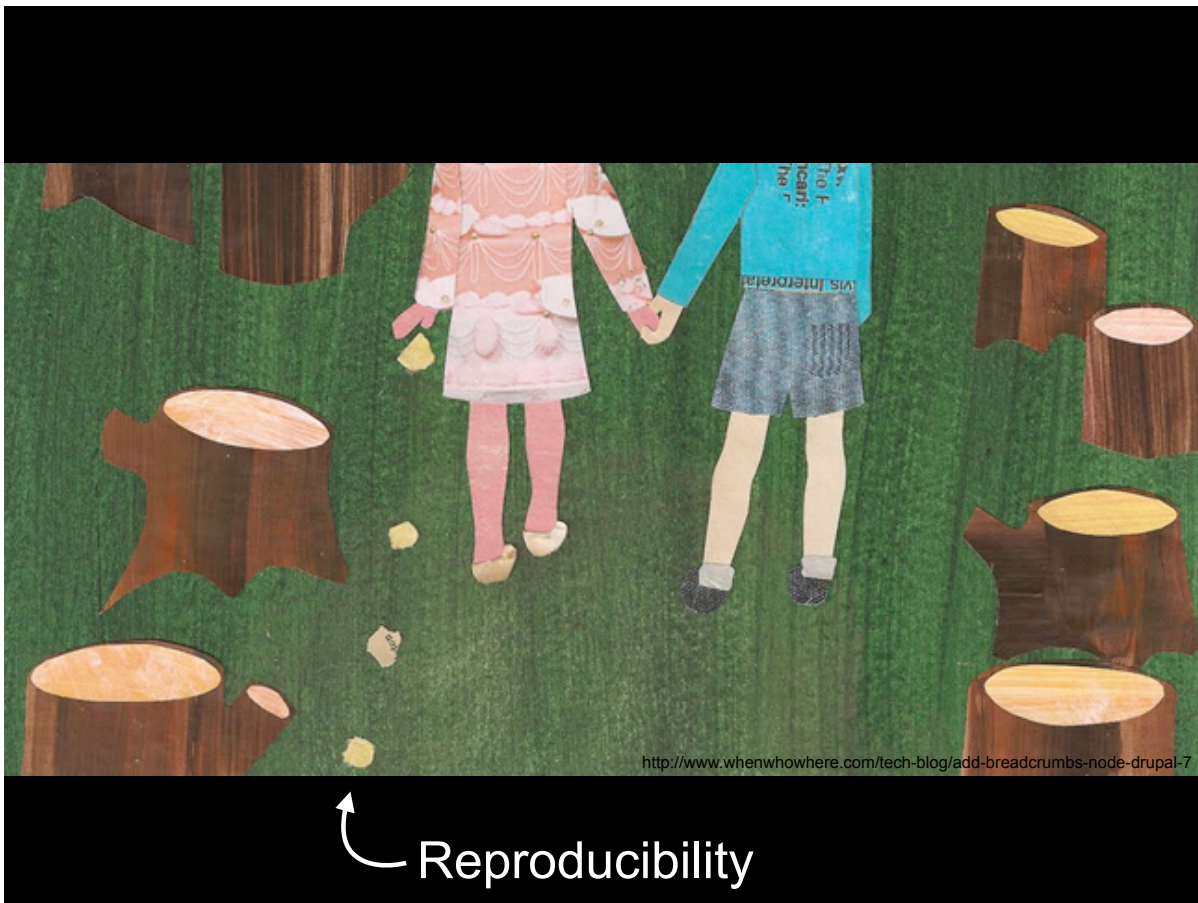
THE UNIVERSITY OF **CHICAGO**
COMPUTER SCIENCE

Computation Institute

http://www.whenwhowhere.com/tech-blog/add-breadcrumbs-node-drupal-7

Reproducibility

EuroRV³ topics

Reproducibility

Verification

Validation

} Discussed by Torsten yesterday, and Mike today

in Visualization

Growing ways to facilitate reproducibility is an active and vital research area.

How will Visualization choose to build on this?

Outline

Motivation

Space of reproducibility

Selected samples

Looking forward

Outline

Motivation

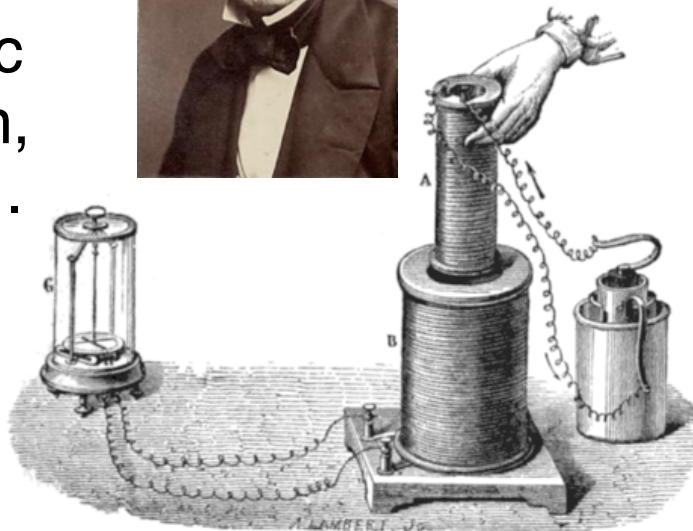**Space of reproducibility**

Selected samples

Looking forward

Historic Scientific Reproducibility

Michael Faraday (1791-1867)

Described electromagnetic fields, induction, diamagnetism...

Clearly written notebooks enabled reproduction

Current Scientific Reproducibility

Proceedings of the National Academy of Sciences (PNAS)

(x) Materials and Data Availability. To allow others to replicate and build on work published in PNAS, authors must make materials, data, and associated protocols available to readers. Authors

http://www.pnas.org/site/misc/iforc.pdf

Science

"... To address the growing complexity of data and analyses, *Science* is extending our data access requirement ... **to include computer codes involved in the creation or analysis of data**. ..."

http://www.sciencemag.org/content/331/6018/649.short

---

http://www.nature.com/news/2010/101013/full/467753a.html

nature
International weekly journal of science

| nature news home | news archive | specials | opinion | features | news blog | nature journal |

comments on this story

Published online 13 October 2010 | *Nature* **467**, 753 (2010) | doi:10.1038/467753a

Column: World View

Publish your computer code: it is good enough

Freely provided working code — whatever its quality — improves programming and enables others to engage with your research, says Nick Barnes.

Nick Barnes

Stories by subject

. Lab life

This article elsewhere

Blogs linking to this article

Add to Digg

Add to Facebook

Add to Newsvine

Add to Del.icio.us

Add to Twitter

Naturejobs

Laboratory Tec
The Nestlé Institute

Scientist for NI
Max Planck Institut

More science j

Post a job for

Resources

PDF Forma

Send to a F

Reprints &

RSS Feeds

I am a professional software engineer and I want to share a trade secret with scientists: most professional computer software isn't very good. The code inside your laptop, television, phone or car is often badly documented, inconsistent and poorly tested.

Why does this matter to science? Because to turn raw data into published research papers often requires a little programming, which means that most scientists write software. And you scientists generally think the code you write is poor. It doesn't contain good comments, have sensible variable names or proper indentation. It

Reproducibility and Impact

V Stodden et al. "How Journals are Adopting Open Data and Code Policies"

1st Global Thematic IASC Conference on the Knowledge Commons: Governing Pooled Knowledge Resources, Belgium, Sept 2012

http://www.stanford.edu/~vcs/papers/IASC2012-STODDEN-Sept122012.pdf

- Journal publication offers a *quid pro quo*: you give us (publisher) reproducibility info, we give you professional visibility

- Correlation (0.55; p=0.0017) between **Impact Factor** and having open data/code requirement for publication

Outline

Motivation

Space of reproducibility

Selected samples

Looking forward

Who pays for reproducibility?

Reviewers

Conference

Research

Community

Publisher

Institution
& support staff

Colleagues
& Group

Author(s)

For How Long

Millennia

Until publisher goes
out of business

Until I don't
get tenure

Now

For what kind of paper

By Whom

Reviewers

Conference

Research Community

Publisher

Institution
& support staff

Colleagues
& Group

Author(s)

Real Datasets
Test Datasets
Source Code
Executables
Provenance
Documentation
Software Portability

What

For Whom

General Public

For How Long

Millennia

Until publisher goes out o...

Conjecture:
Miscommunication about location or extent in this space underlies our disagreements about reproducibility

get tenure

Now

For what kind of paper

Real Datasets
Test Datasets

Executables
Provenance
Documentation

General Public

What

For Whom

Software Portability

Whom

Costs we already pay

Learning to code

LaTeX .tex, .bib, .cls

Presenting in English

Airfare, registration

Outline

Motivation

Space of reproducibility

Selected samples

Looking forward

Who pays for reproducibility?

Reviewers

Conference

Research

Community

Publisher

Institution
& support staff

Colleagues
& Group

Author(s)

Lab pays
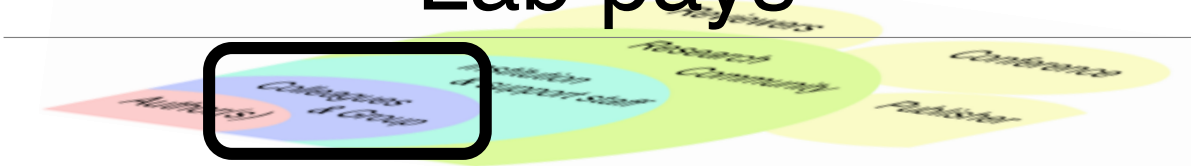
Wavelab http://www-stat.stanford.edu/~wavelab/
http://www-stat.stanford.edu/~wavelab/Wavelab_850/wavelab.pdf

Matlab routines for wavelet analysis

Using & building Wavelab is part of Donoho's lab culture

"An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures."   --David Donoho on WaveLab

Author pays
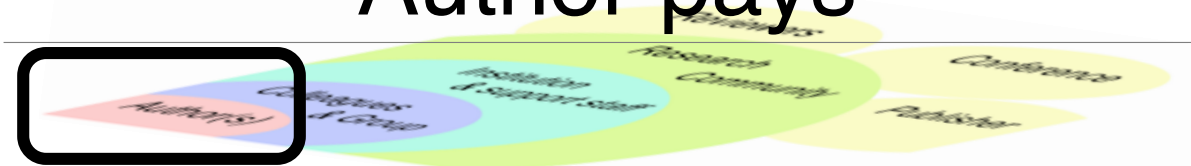
Sweave http://www.stat.uni-muenchen.de/~leisch/Sweave/

Embed data analysis R code in LaTeX

"Literate Programming" -Knuth

PTU https://www.usenix.org/system/files/conference/tapp13/tapp13-final18.pdf

Provenance-to-Use: software, its inputs, and its execution, are packaged into self-contained unit

Author, lab, others?

VisTrails http://www.vistrails.org

Open-source scientific workflow and provenance management system

Does entail a new way of working

Pay-off: **Knowledge Reuse**

User community increases power

Lab/Institution pays

(Network connectivity, web servers)

University of Chicago Research Computing Center http://rcc.uchicago.edu

"We'll help write and implement your NSF Data Management plan"

Research Computing Center   THE UNIVERSITY OF CHICAGO

Resources   Services   Access and Docs   About
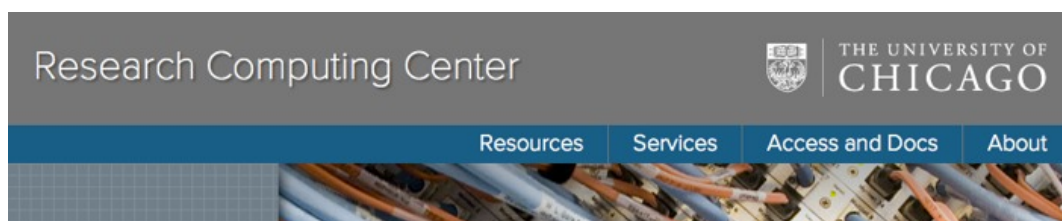
Author, Community pays

ROOT http://root.cern.ch/drupal/
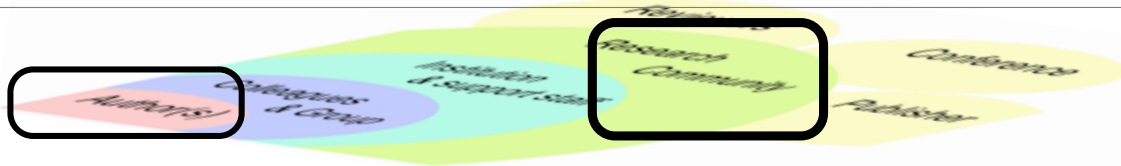
> Software from CERN to handle and analyze large amounts of High Energy Physics (and other areas)

http://www.runmycode.org/

> Authors create companion website to paper; computation behind figures runs in the cloud

Vis Contest: http://ieeevis.org/year/2012/contest-session/all/all

Reviewers, Conference pays
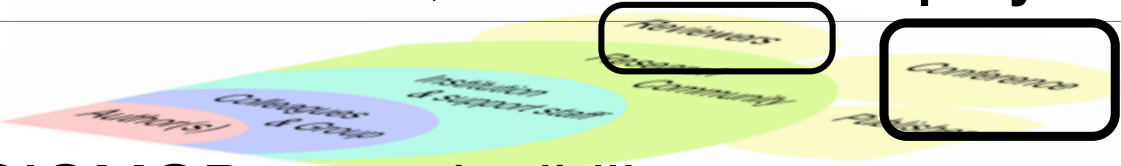
SIGMOD reproducibility http://www.sigmod.org/2013/reproducibility.shtml

Large-scale data management and databases

Papers can earn different levels of reproducibility, as assessed by Reproducibility Committee

"Experimental papers will be most useful when their results have been tested and generalized by objective third parties"

Publisher pays
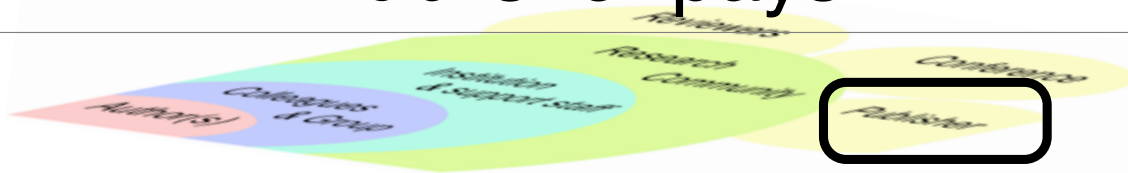
Elsevier **executable paper** challenge

"Publish" from Latin *publicare* "make public," from *publicus*

ACM Trans on Mathematical Software http://toms.acm.org

Procedia Computer Science
Volume 4, 2011, Pages 608–617
Proceedings of the International Conference on Computational Science, ICCS 2011

The Collage Authoring Environment

Piotr Nowakowski[a], Eryk Ciepiela[a], Daniel Harężlak[a], Joanna Kocot[a], Marek Kasztelnik[a], Tomasz Bartyński[a], Jan Meizner[a], Grzegorz Dyk[a], Maciej Malawski[b, c]

[a] ACC CYFRONET AGH, ul. Nawojki 11, 30-950 Kraków, Poland
[b] Institute of Computer Science AGH, al. Mickiewicza 30, 30-059 Kraków, Poland
[c] Center for Research Computing, University of Notre Dame, USA

http://dx.doi.org/10.1016/j.procs.2011.04.064, How to Cite or Link Using DOI
Permissions & Reprints

Abstract

The Collage Authoring Environment is a software infrastructure which enables domain scientists to collaboratively develop and publish their work in the form of executable papers. It corresponds to the

---

Outline

Motivation

Space of reproducibility

Selected samples

Looking forward

How VIS requests reproducibility now

http://ieeevis.org/year/2013/info/call-participation/paper-submission-guidelines

SUPPLEMENTAL MATERIAL AND FORMATTING GUIDELINES
... Submission of code or other supplemental material in order to increase the reproducibility of the work is also encouraged. Those submitting to the conference are urged to make available salient parameter settings of pertinent algorithms and ideally obtain results using open source data. In case specific data sets are employed, we ask that a version of these be made available where possible.

Will happen at the decade time-scale

NSF and NIH have long encouraged sharing

> 2011: NSF Data Management Plan

1980 Bayh-Dole → Tech-Transfer Offices

> Grasped open source licensing within last decade

Journal of the American Statistical Association (JASA)
> (from Victoria Stodden http://www.stanford.edu/~vcs/talks/RRJuly152011-STODDEN.pdf)

> articles using computation: 1996 <50% → 2011 100%

> articles w/ code available: 1996   0%   → 2011 21%

"Visualization + Info Vis" → "Visualization + Info Vis + VAST"
> → "SciVis + InfoVis + VAST + LDAV + BioVis @VisWeek → **VIS**"

World is changing, and moving towards reproducibility.

We are changing. Let's move towards reproducibility.

Lets help each other!

"Evaluation" is expensive

Authors: comparing with results from previous methods

Reviewers: determining if new results are in fact what is claimed

Both easier with reproducibility

Community practices drive this

Lets walk our talk **together**

http://www.whenwhowhere.com/tech-blog/add-breadcrumbs-

Thank you