

The challenge of reproducible research in the computer age

Production is not the application of tools to materials.
It is the application of logic to work.
—Peter Drucker, *The practice of management* (1954)

K. Jarrod Millman
Helen Wills Neuroscience Institute
University of California, Berkeley
Applied Mathematics Perspectives 2011
Reproducible Research: Tools and Strategies for Scientific Computing

Culture matters



Build quality into the process



The radical novelty of computing

The concept of radical novelties is of contemporary significance because, while we are ill-prepared to cope with them, science and technology have now shown themselves expert at inflicting them upon us.

— Edsger Dijkstra, *The Cruelty of Really Teaching Computer Science* (1988)

Better, faster, cheaper

- Are we doing a good (enough) job? How would we know?
- How long does it take to go from the idea as presented in (say) lab meeting to the paper being submitted?
- What proportion of measured data makes it to publication?
- Are we duplicating work that other people have done already?
- Are we doing work for other people because they don't know how to do it?
- Are there tasks that can be automated?

“truth will sooner come out of error than from confusion.”

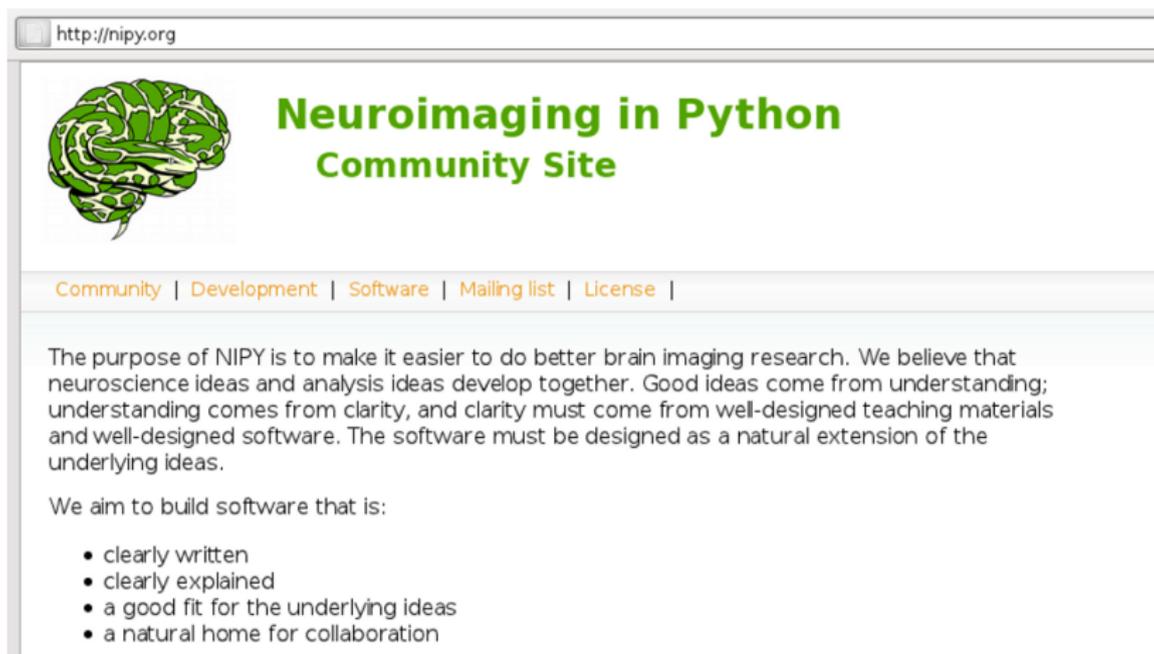
...so when a man tries all kinds of experiments without method or order, this is mere groping in the dark; but when he proceeds with some direction and order in his experiments, it is as if he were led by the hand...
— Francis Bacon, *Novum Organum* (1620)

Neuroimaging



Deep magic begins here...

- Specialization
- Lack of patience
- Lack of understanding
- Confusion, frustration, and helplessness



The screenshot shows a web browser window with the address bar containing "http://nipy.org". The main content area features a green brain icon on the left and the title "Neuroimaging in Python Community Site" in green text. Below the title is a navigation menu with links for "Community", "Development", "Software", "Mailing list", and "License". The main text block explains the purpose of NIPY and lists four goals: clearly written, clearly explained, a good fit for the underlying ideas, and a natural home for collaboration.

http://nipy.org



Neuroimaging in Python Community Site

[Community](#) | [Development](#) | [Software](#) | [Mailing list](#) | [License](#) |

The purpose of NIPY is to make it easier to do better brain imaging research. We believe that neuroscience ideas and analysis ideas develop together. Good ideas come from understanding; understanding comes from clarity, and clarity must come from well-designed teaching materials and well-designed software. The software must be designed as a natural extension of the underlying ideas.

We aim to build software that is:

- clearly written
- clearly explained
- a good fit for the underlying ideas
- a natural home for collaboration

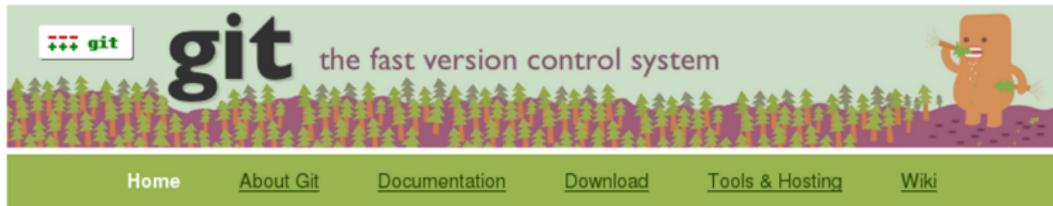
The process

- How many mistakes do you make?
- What do they cost?
- Could you have made mistakes you don't know about?

Data & code sharing

- Could you send someone else in the lab an email with all the information they need to rerun your analysis?
- How long would it take to write that email?

Git for everything



Git is...

Git is a **free & open source, distributed version control system** designed to handle everything from small to very large projects with speed and efficiency.

Every Git clone is a full-fledged repository with complete history and full revision tracking capabilities, not dependent on network access or a central server.

Branching and merging are fast and easy to do.

Git is used for version control of files, much like tools such as [Mercurial](#), [Bazaar](#), [Subversion](#), [CVS](#), [Perforce](#), and [Team Foundation Server](#).

Projects using Git

- [Git](#)
- [Linux Kernel](#)
- [Perl](#)
- [Eclipse](#)
- [Gnome](#)
- [KDE](#)
- [Qt](#)
- [Ruby on Rails](#)
- [Android](#)
- [PostgreSQL](#)
- [Debian](#)
- [X.org](#)

Download Git

The latest stable Git release is

v1.7.6

[release notes](#) (2011-06-26)



[Windows](#)



[Mac OSX](#)



[Source](#)

[Other Download Options](#)
[Git Source Repository](#)

Python

The screenshot shows the IEEE Xplore Digital Library interface. At the top, the browser address bar displays the URL: `http://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=5725228`. Below the address bar, the IEEE Xplore logo is visible, along with a search bar and navigation links for "Advanced Search", "Preferences", "Search Tips", and "Search within contents".

The main navigation bar includes "BROWSE", "MY SETTINGS", "CART", "SIGN OUT", and "About IEEE Xplore". The breadcrumb trail reads: "Browse > Journals > Computing in Science & Enginee ... Volume 13 Issue 2".

On the left side, there is a "QUICK SEARCH" section with input fields for "Volume:", "Issue:", and "Start Page:", and a "GO" button. Below it is a "TITLE HISTORY" section showing "(1994 - 1998) Computational Science & Engineering, IEEE".

The main content area features the journal title "computing in SCIENCE & ENGINEERING" in large red and black text. Below the title is a small image of a tablet displaying a blue waveform. To the right of the image, there is an "Early Access:" section with a "VIEW ARTICLES" button. Below this, there are dropdown menus for "Year:" (set to 2011) and "Volume:" (set to Volume: 13 Issue: 2), followed by a "VIEW CONTENTS" button.

“Literate programming”

- Sweave: \LaTeX & R
- Sphinx: reStructuredText & Python

Automate, automate, automate



Programming as a first class citizen

- Read programming articles, books, etc.
- Learn new languages

Agile methodology

- Test driven development
- Pair programming
- Metaprogramming

Programming best practices

http://software-carpentry.org/



Google

Software Carpentry

Helping scientists make better software since 1997

Home About Lectures Blog License Contributing Contact Research

Type text to search here...

About

Since 1997, **Software Carpentry** has taught scientists and engineers the concepts, skills, and tools they need to use and build software more productively. All of the content is freely available under a Creative Commons license, and we are constantly adding and updating lectures, videos, and exercises.

Can Software Carpentry help you? These [comments from former students](#) and our [three-minute pitch](#), can help you decide.

Can you help Software Carpentry? We are an open source/open license project, and there are many ways in which volunteers can [contribute](#). We also have several [funding models](#), so if you would like to help people around the world solve the problems they face today, and prepare them to tackle the larger challenges of parallelism, cloud computing, reproducible research, and global-scale collaboration, please [get in touch](#).

Python

The screenshot shows the SciPy.org website. At the top, there is a navigation bar with the SciPy logo and the text "Sponsored by ENTHOUGHT". Below this is a search bar. On the left side, there is a "Wiki" sidebar with a list of links: Documentation, Mailing Lists, Download, Installing SciPy, Topical Software, Cookbook, Developer Zone, Blogs, Conference, and SciPy (which is highlighted). Below the sidebar is a "Page" section with "Immutable Page", "Info", "Attachments", and a "More Actions:" dropdown menu. The main content area features the "SciPy" title and five circular icons with labels: "Download" (a blue circle with a white 'S' and a green arrow pointing down), "Getting Started" (a yellow circle with a white 'S' and a green arrow pointing up), "Documentation" (a blue circle with a white 'S' and a stack of books), "Report Bugs" (a blue circle with a white 'S' and a red bug), and "Read the Blog" (an orange square with a white RSS symbol). Below these icons is a section titled "Scientific Tools for Python" with a paragraph of text describing SciPy as open-source software for mathematics, science, and engineering, and mentioning its dependencies on NumPy and its use in various operating systems.

http://33bits.org

33 Bits of Entropy

The End of Anonymous Data and what to do about it

HOME

ABOUT 33 BITS

SITEMAP

ARVIND NARAYANAN

Go!

About 33 Bits

This is a blog about my research on privacy and anonymity. The title refers to the fact that there are only 6.6 billion people in the world, so you only need 33 bits (more precisely, 32.6 bits) of information about a person to determine who they are.

This fact has two related consequences. First, a lot of traditional thinking about anonymous data relied on the fact that you can hide in a crowd that's too big to search through. That notion completely breaks down given today's computing power: as long as the bad guy has enough information about his target, he can simply examine every possible entry in the database and select the best match.

The second consequence is that 33 bits is not really a lot. If your hometown has 100,000 people, then knowing your hometown gives me 16 bits of entropy about you, and only 17 bits remain. But the real danger is that information about a person's *behavior*, which was traditionally not considered personally identifying, can be used to cause serious privacy breaches in a variety of different contexts.

Open Research Computation



Now accepting
submissions

Editor-in-Chief
Cameron Neylon (UK)

Open Research Computation is an open access journal that publishes articles describing the development, capacities, and uses of software for researchers in any field. The journal also encourages submissions that review or describe developments relating to software based research tools. All software source code published in *Open Research Computation* is made available under an Open Source Initiative compliant license.

Submit your manuscript and benefit from:

- High visibility for articles through unrestricted online access
- No limits on article length, additional files, colour figures or movies
- Immediate open access publication on acceptance
- Expert peer review

www.openresearchcomputation.com


BioMed Central
The Open Access Publisher