Improving imaging research

Matthew Brett

December 7th 2016

Thesis

We are currently seeing the effects of a misguided attempt to apply 1950's production-line methods to scientific enquiry.

The primary damage is a loss of engagement in the quality of scientific work.

We should address this by conscious attention to scientific culture, training for understanding and an emphasis on correct process.

Our goal is not just reproducibility, but a single-minded commitment on quality.

Plan of the talk

- we probably have an excess of false findings in neuroimaging;
- the academic reward system may be a driver, but we are unlikely to be able to change that soon;
- business as usual is unlikely to correct this;
- two examples of fields that have increased quality by conscious intervention to stimulate engagement and refine process;
- stimulating engagement in neuroimaging;
- attention to process in neuroimaging.

The ubiquity of error

> *The scientific method's central motivation is the ubiquity of error - the awareness that mistakes and self-delusion can creep in absolutely anywhere and that the scientist's effort is primarily expended in recognizing and rooting out error.*

David L. Donoho *et al* (2009) "Reproducible research in computational harmonic analysis" Computing in Science & Engineering 11 p8-18.

Error in neuroimaging

> *I have occasionally asked respected colleagues what percent of published neuroimaging findings they think would replicate, and the answer is generally very depressing. My own guess is* ***way*** *less than 50%.*

Nancy Kanwisher (2013) commenting on Daniel Bor's blog post.

My straw poll

> *Let us say you took a random sample of papers using functional MRI over the last five years. For each study in the sample, you repeated the same experiment. What proportion of your repeat experiments would substantially replicate the main findings of the original paper?*

Answers from people running neuroimaging labs vary from 5% to 50%.

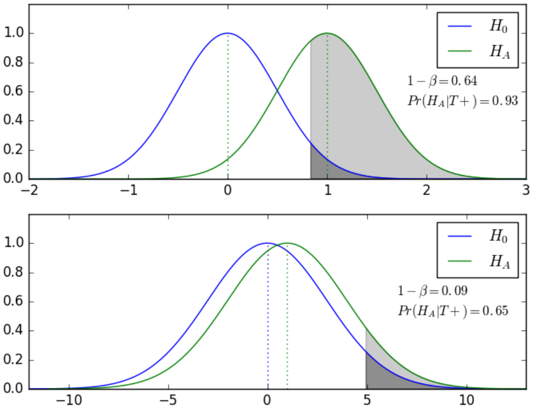
Risks for error

Increased risk of false findings for:

1. small sample size (low power);
2. small effect size (low power);
3. large number of tests (analysis bias);
4. greater flexibility in analysis (analysis bias);
5. greater financial interests (analysis bias);
6. larger numbers of groups studying same effects (publication bias);

John P. A. Ioannidis (2005). "Why most published research findings are false." PLoS medicine 2 (8): e124. See also exposition on Ioannidis 2005.

Low power increases false reports



- low power leads to lower probability that alternative is true, given a significant test statistic;
- effect size and the winner's curse.

Low power is typical for neuroimaging

In neuroimaging studies of brain volume abnormalities:

> *Our results indicated that the median statistical power of these studies was 8% across 461 individual studies contributing to 41 separate meta-analyses, which were drawn from eight articles that were published between 2006 and 2009.*

Katherine S. Button *et al* (2013) “Power failure: why small sample size undermines the reliability of neuroscience”. Nature Reviews Neuroscience 14, 365-376

Low power and false positives

> *2.1.5. Corpus Callosum*
>
> *Corpus callosum is found to be correlated with the ASD. . . . two successive longitudinal RBV studies . . . have found persistent reductions in the total corpus callosum volumes in the autistic subjects compared to the healthy controls.*

Ismail MM *et al* (2016) “Studying Autism Spectrum Disorder with Structural and Diffusion Magnetic Resonance Imaging: A Survey.” Front. Hum. Neurosci.

Corpus callosum and autism with large sample size

> *Our meta-analysis suggested a group difference in CC size; however, the studies were heavily underpowered (20% power to detect Cohen's d 5 .3). In contrast, we did not observe significant differences in the Autism Brain Imaging Data Exchange cohort, despite having achieved 99% power.*

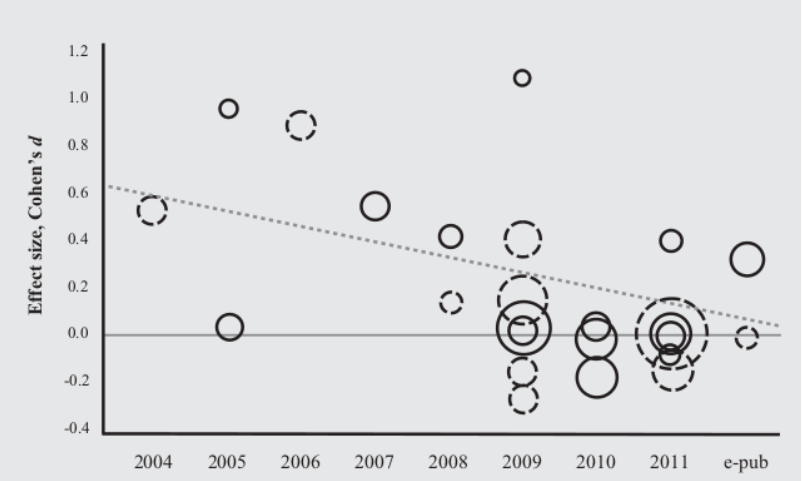
Aline Lefebvre *et al* (2016) "Neuroanatomical Diversity of Corpus Callosum and Brain Volume in Autism: Meta-analysis, Analysis of the Autism Brain Imaging Data Exchange Project, and Simulation" Biological Psychiatry 78:126–134

Genetic markers and hippocampal volume

> *. . . previously identified polymorphisms associated with hippocampal volume showed little association in our meta-analysis (BDNF, TOMM40, CLU, PICALM, ZNF804A, COMT, DISC1, NRG1, DTNBP1). . .*

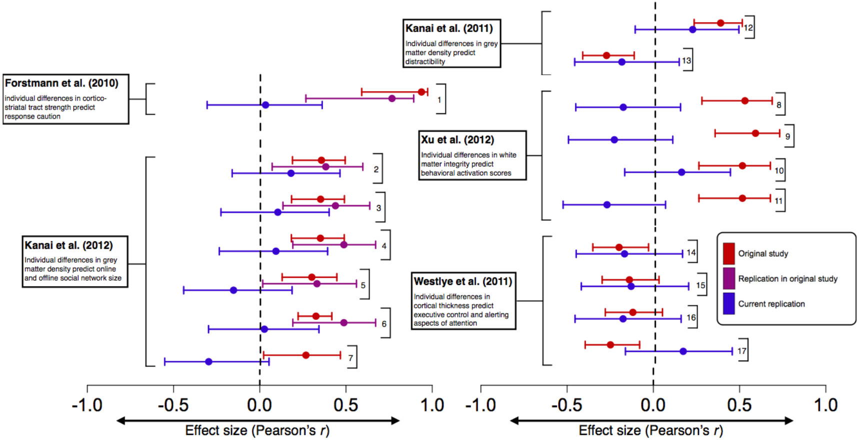
Jason Stein *et al* (2012) “Identification of common variants associated with human hippocampal and intracranial volumes”. Nat Genet. 44(5): 552–561.

Publication bias and effect size



Molendijk *et al* (2012) "A systematic review and meta-analysis on the association between BDNF val(66)met and hippocampal volume–a genuine effect or a winners curse?" Am J Med Genet B Neuropsychiatr Genet 159B(6):731-40

Replication of anatomy-behavior correlations



Wouter Boekel *et al* (2013). "A purely confirmatory replication study of structural brain-behavior correlations". J. Neurosci 12, 4745–65

Analysis flexibility

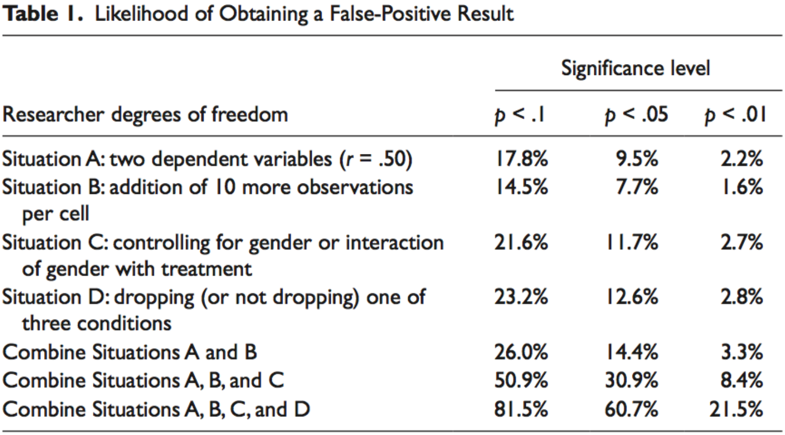


Table 1. Likelihood of Obtaining a False-Positive Result

	Significance level		
Researcher degrees of freedom	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

Joseph P. Simmons *et al* (2011) “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant” Psychological Science 22(11) 1359–1366.

Analysis flexibility is characteristic of imaging

> *Ten analysis steps for which multiple strategies appear in the literature were identified, and two to four strategies were enumerated for each step. Considering all possible combinations of these strategies yielded 6,912 unique analysis pipelines.*

Joshua Carp (2012) “On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments” Front. Neurosci.

This all looks a lot like

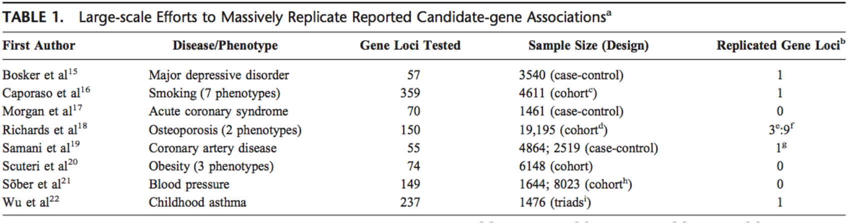


TABLE 1. Large-scale Efforts to Massively Replicate Reported Candidate-gene Associations[a]

First Author	Disease/Phenotype	Gene Loci Tested	Sample Size (Design)	Replicated Gene Loci[b]
Bosker et al[15]	Major depressive disorder	57	3540 (case-control)	1
Caporaso et al[16]	Smoking (7 phenotypes)	359	4611 (cohort[c])	1
Morgan et al[17]	Acute coronary syndrome	70	1461 (case-control)	0
Richards et al[18]	Osteoporosis (2 phenotypes)	150	19,195 (cohort[d])	3[e]:9[f]
Samani et al[19]	Coronary artery disease	55	4864; 2519 (case-control)	1[g]
Scuteri et al[20]	Obesity (3 phenotypes)	74	6148 (cohort)	0
Sõber et al[21]	Blood pressure	149	1644; 8023 (cohort[h])	0
Wu et al[22]	Childhood asthma	237	1476 (triads[i])	1

- few selected candidate risk factors;
- small sample size;
- “substantial” reporting bias

Ioannidis *et al* (2011) “The False-positive to False-negative Ratio in Epidemiologic Studies” Epidemiology 22(4) p450-6

A lack of concern

> *Computing results are now being presented in a very loose, "breezy" way—in journal articles, in conferences, and in books. All too often one simply takes computations at face value. This is spectacularly against the evidence of my own experience. I would much rather that at talks and in referee reports, the possibility of such error were seriously examined.*

David L. Donoho (2010). "An invitation to reproducible computational research" Biostatistics 11(3) p385-8

Data sharing might lead to refutation

> *A second concern held by some is that a new class of research person will emerge — people who had nothing to do with the design and execution of the study but use another group's data for their own ends, possibly stealing from the research productivity planned by the data gatherers, or even use the data to try to disprove what the original investigators had posited.*

Dan L. Longo, Jeffrey M. Drazen, editorial (2016) "Data Sharing"
N Engl J Med 374:276-277

Failed replications are uninteresting

> *Recent hand-wringing over failed replications in social psychology is largely pointless, because unsuccessful experiments have no meaningful scientific value.*
> *Because experiments can be undermined by a vast number of practical mistakes, the likeliest explanation for any failed replication will always be that the replicator bungled something along the way . . .*
> *Whether they mean to or not, authors and editors of failed replications are publicly impugning the scientific integrity of their colleagues.*

Jason Mitchell (2014) "On the emptiness of failed replications" archived blog post

Cell culture contamination

- 1967 - Stanley Gartler: 18 of 18 human cell lines were HeLa;
- 1975-81 - Walter Nelson-Rees: widespread cross-contamination;
- 2007 - Roland Nardone: "Eradication of cross-contaminated cell lines: a call for action."
- 2015 estimate is 20% of cell-lines contaminated;
- 2013 survey - 19% published papers reported cell-line authentication;

Leonard P. Freedman *et al* (2015) "Reproducibility: changing the policies and culture of cell line authentication" Nature Methods 12(6) 493-7.

Cell culture contamination

> *[scientists show] a general and quite remarkable concern for truth. Would-be authors are forever going back to their benches to check small points raised by referees . . . It would be tragic if these civilized habits were to be corrupted by the activities of the self-appointed vigilantes.*

John Maddox, editorial (1981) “Responsibility for trust in research” Nature 289 p211-2.

Citations of false findings

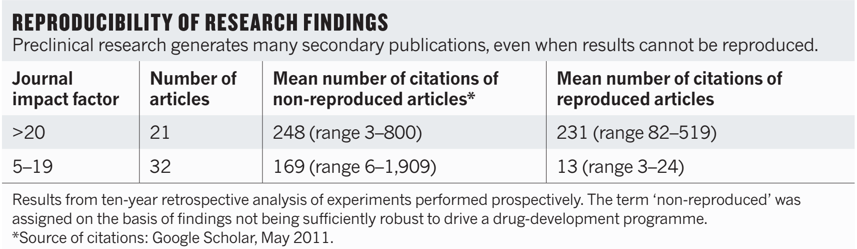
Scientists at Amgen (a drug company) tried to reproduce findings from 53 “landmark” studies.

> *. . . when findings could not be reproduced, an attempt was made to contact the original authors, discuss the discrepant findings, exchange reagents and repeat experiments under the authors’ direction, occasionally even in the laboratory of the original investigator.*

Of 53 studies, only 6 replicated (11%).

Glenn Begley and Lee Ellis (2012) “Raise standards for preclinical cancer research” Nature 483

Citations of false findings



REPRODUCIBILITY OF RESEARCH FINDINGS

Preclinical research generates many secondary publications, even when results cannot be reproduced.

Journal impact factor	Number of articles	Mean number of citations of non-reproduced articles*	Mean number of citations of reproduced articles
>20	21	248 (range 3–800)	231 (range 82–519)
5–19	32	169 (range 6–1,909)	13 (range 3–24)

Results from ten-year retrospective analysis of experiments performed prospectively. The term 'non-reproduced' was assigned on the basis of findings not being sufficiently robust to drive a drug-development programme.

*Source of citations: Google Scholar, May 2011.

Glenn Begley and Lee Ellis (2012) "Raise standards for preclinical cancer research" Nature 483

How did this happen?

> *.. compared to non-rewarded subjects, subjects offered a task-extrinsic incentive choose easier tasks, are less efficient in using the information available to solve novel problems, and tend to be answer oriented and more illogical in their problem-solving strategies. They seem to work harder and produce more activity, but the activity is of lower quality, contains more errors, and is more stereotyped and less creative than the work of comparable nonrewarded subjects working on the same problems"*

J Condry (1977) Journal of Personality and Social Psychology (quoted in "Punished by rewards" by Alfie Kohn).

See also: Edward L. Deci *et al* (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychological bulletin, 125(6), 627.

What can we do?

> *Among all the relevant stakeholders, concerns about the culture of research are often on matters that they think are outside their control or are someone else's responsibility"*

Nuffield Council on Bioethics (2014) “The culture of scientific research in the UK”

What can we do?

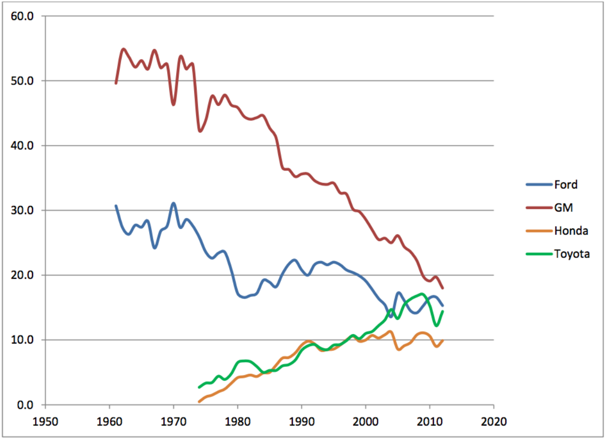
Two fields where conscious intervention was effective in increasing output quality:

- car manufacture.
- software projects;

Note emphasis on:

- engagement;
- process.

Toyota and General Motors



Source: Wards Automotive.

Susan Helper, Rebecca Henderson (2014) "Management Practices, Relational Contracts and the Decline of General Motors". Harvard Business School Working Paper 14-062

Toyota and General Motors

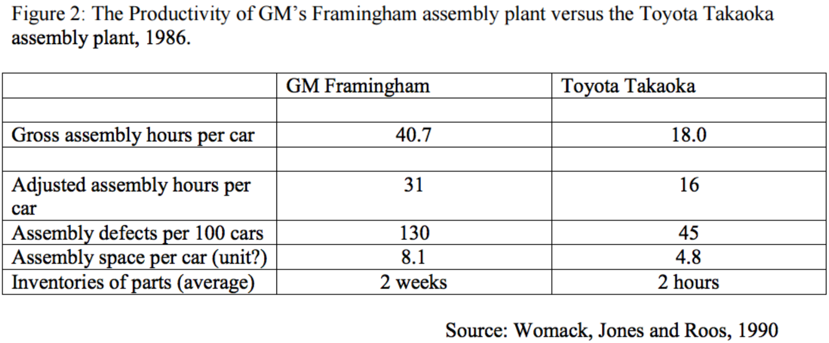


Figure 2: The Productivity of GM's Framingham assembly plant versus the Toyota Takaoka assembly plant, 1986.

	GM Framingham	Toyota Takaoka
Gross assembly hours per car	40.7	18.0
Adjusted assembly hours per car	31	16
Assembly defects per 100 cars	130	45
Assembly space per car (unit?)	8.1	4.8
Inventories of parts (average)	2 weeks	2 hours

Source: Womack, Jones and Roos, 1990

Susan Helper, Rebecca Henderson (2014)

Culture at General Motors

> *General Motors was a kind of throw it over the wall organization. Each department, we were very compartmentalized, and you design that vehicle, and you'd throw it over the wall to the manufacturing guys.*

Ernie Schaefer, GM manager, interviewed in "NUMMI"; This American Life episode 403 (2010).

Culture at Toyota

14 principles in four sections:

1. Long-term Philosophy;
2. The Right Process Will Produce the Right Results;
3. Add Value to the Organization by Developing Your People;
4. Continuously Solving Root Problems Drives Organizational Learning

The Toyota Way - process

2. The Right Process Will Produce the Right Results;
   - 2.5 Build a culture of stopping to fix problems, to get quality right the first time. Quality takes precedence (Jidoka).
   - 2.6 Standardized tasks and processes are the foundation for continuous improvement and employee empowerment.
   - 2.7 Use visual control so no problems are hidden.

The Toyota Way - developing people

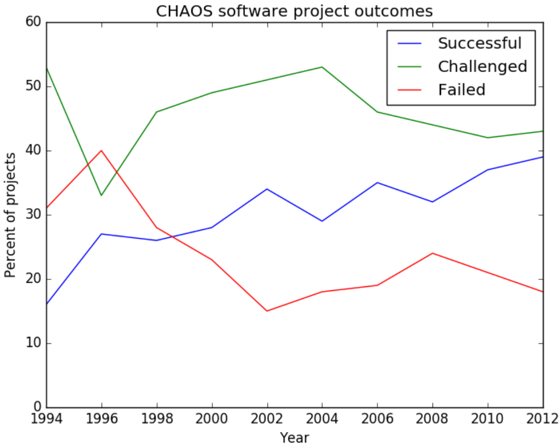
3. Add Value to the Organization by Developing Your People;
    - 3.9 Grow leaders who thoroughly understand the work, live the philosophy, and teach it to others.
    - 3.10 Develop exceptional people and teams who follow your company's philosophy.

The Toyota Way - solving root problems

4. Continuously Solving Root Problems Drives Organizational Learning

   4.12 Go and see for yourself to thoroughly understand the situation (Genchi Genbutsu).

Software quality



From the Standish CHAOS report 1994-2012.

Code as personal property

> *In the early years of programming, a program was regarded as the private property of the programmer. One would no more think of reading a colleague's program unbidden than of picking up a love letter and reading it. This is essentially what a program was, a love letter from the programmer to the hardware, full of the intimate details known only to partners in an affair. Consequently, programs became larded with the pet names and verbal shorthand so popular with lovers who live in the blissful abstraction that assumes that theirs is the only existence in the universe. Such programs are unintelligible to those outside the partnership.*

Attributed to Michael Marcotty, quoted in Steve McConnell (2004) "Code Complete, second edition" p 842. Microsoft Press.

Developer responsibility

> *.. in my office is a big poster that says “Nothing at Facebook is someone else’s problem”, and the remarkable thing about Facebook as an engineering organization is the degree to which 7000 people all actually agree on that . . . the transition back in the dark ages wasn’t from a healthy relationship [with quality assurance teams], it was from this very dysfunctional, aresponsibile attitude, throw it over the wall, long long cycles, vague feedback. Going from that, to programmers accepting responsibility for the quality of their work, that was a huge step forward.*

Kent Beck (2014) discussing test-first development.

Modern process is more effective

	Agile	Waterfall
Successful	42%	14%
Challenged	49%	57%
Failed	9%	29%

Standish group (2011) “CHAOS report”, summary

Barriers to engagement

- unfamiliarity with ideas and tools;
- “makes sense epistemology”;
- “garbage in, gospel out”;
- black box software and pipelines;

Removing barriers to engagement

- understanding one level down;
- a landscape of concepts;
- commitment to teaching on math;
- prove everything you reasonably can;
- teaching with code;
- opening the black box.

Opening the black box

> *The tools we use have a profound (and devious!) influence on our thinking habits, and, therefore, on our thinking abilities."*

Edsger W. Dijkstra "How do we tell truths that might hurt?" link

Opening the black box

“What I cannot create, I do not understand”

Found on Richard Feynman’s blackboard after his death.

Math etc curriculum

- Floating point calculations;
- Fourier transform;
- Matrix multiplication and linear algebra;
- Principal Component Analysis;
- Optimization;
- Interpolation;
- Convolution;
- Multiple regression;
- Multiple comparison correction.

Improving process

> *In studies for which findings could be reproduced, authors had paid close attention to controls, reagents, investigator bias and describing the complete data set. For results that could not be reproduced, however, data were not routinely analysed by investigators blinded to the experimental versus control groups. Investigators frequently presented the results of one experiment, such as a single Western-blot analysis. They sometimes said they presented specific experiments that supported their underlying hypothesis, but that were not reflective of the entire data set.*

Begley and Ellis (2012).

Process etc curriculum

- version control;
- automation;
- testing;
- documentation;
- code re-use;
- code review.

Wilson *et al* (2014) “Best practices for scientific computing” PLoS Biology doi

Our teaching

- by analogy with math teaching;
- start and continue with “best practice”;
- Practical neuroimaging class;
- Reproducible computational and statistical data science;
- functional MRI methods class
    - syllabus

Is “practical neuroimaging” practical?

- three 4-unit courses?
- or coding and math integrated into other courses

The end

Thanks to JB Poline, Jarrod Millman, Stefan van der Walt, Paul Ivanov and all the Nipy developers.

The NiPy community

> *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*

Nipy “Mission statement”

Tools are a continuation of teaching

- build your own lightsaber.
- transparent (open source, readable language);
- shared (open development, open governance);
- modular and composable;
- tested;
- eulerangles.py
- nibabel