Open Science, Open Software, and Reproducible Code

a marriage of FOSS and Science

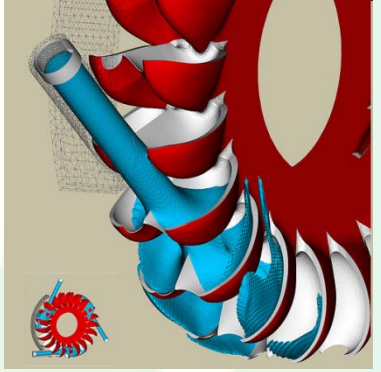
Bill Hoffman CTO Founder Kitware Inc, “the CMake guy”, Barefoot runner

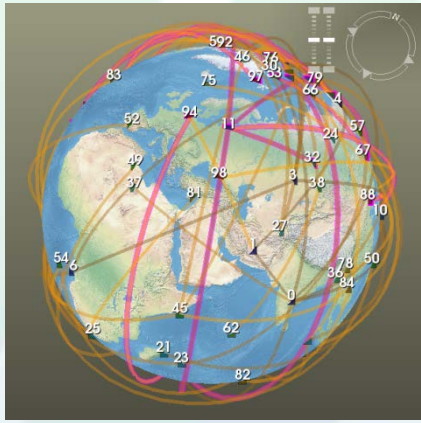
FOSDEM 2013

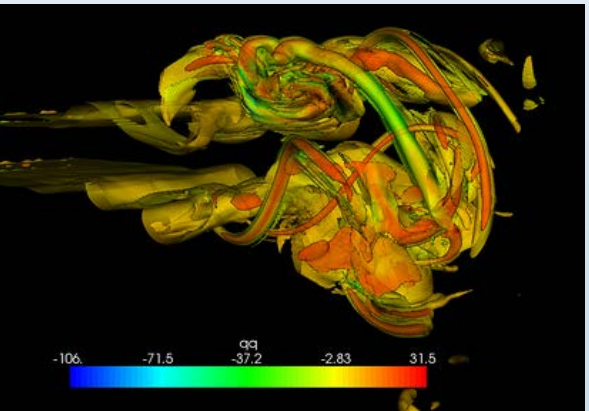
Kitware, Inc.

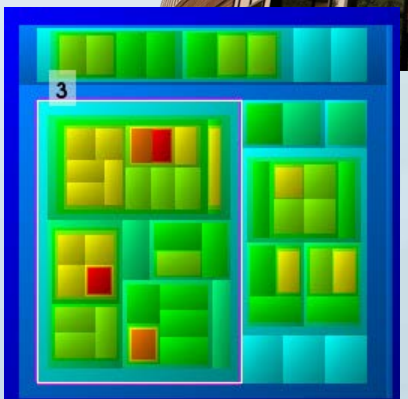
Open Source
Scientific Computing Software

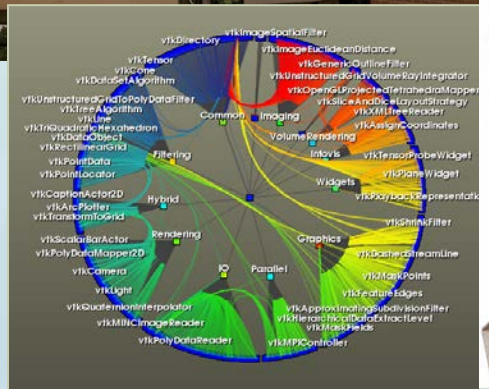
Software Services

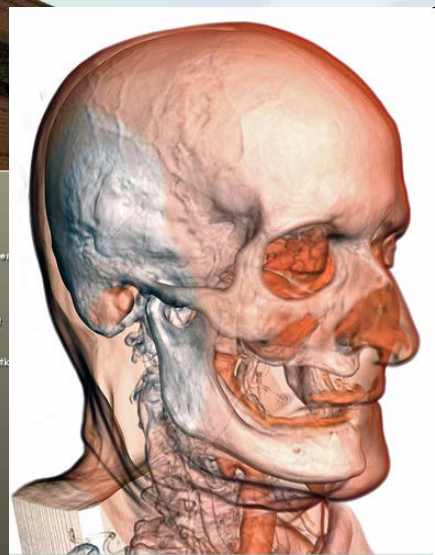


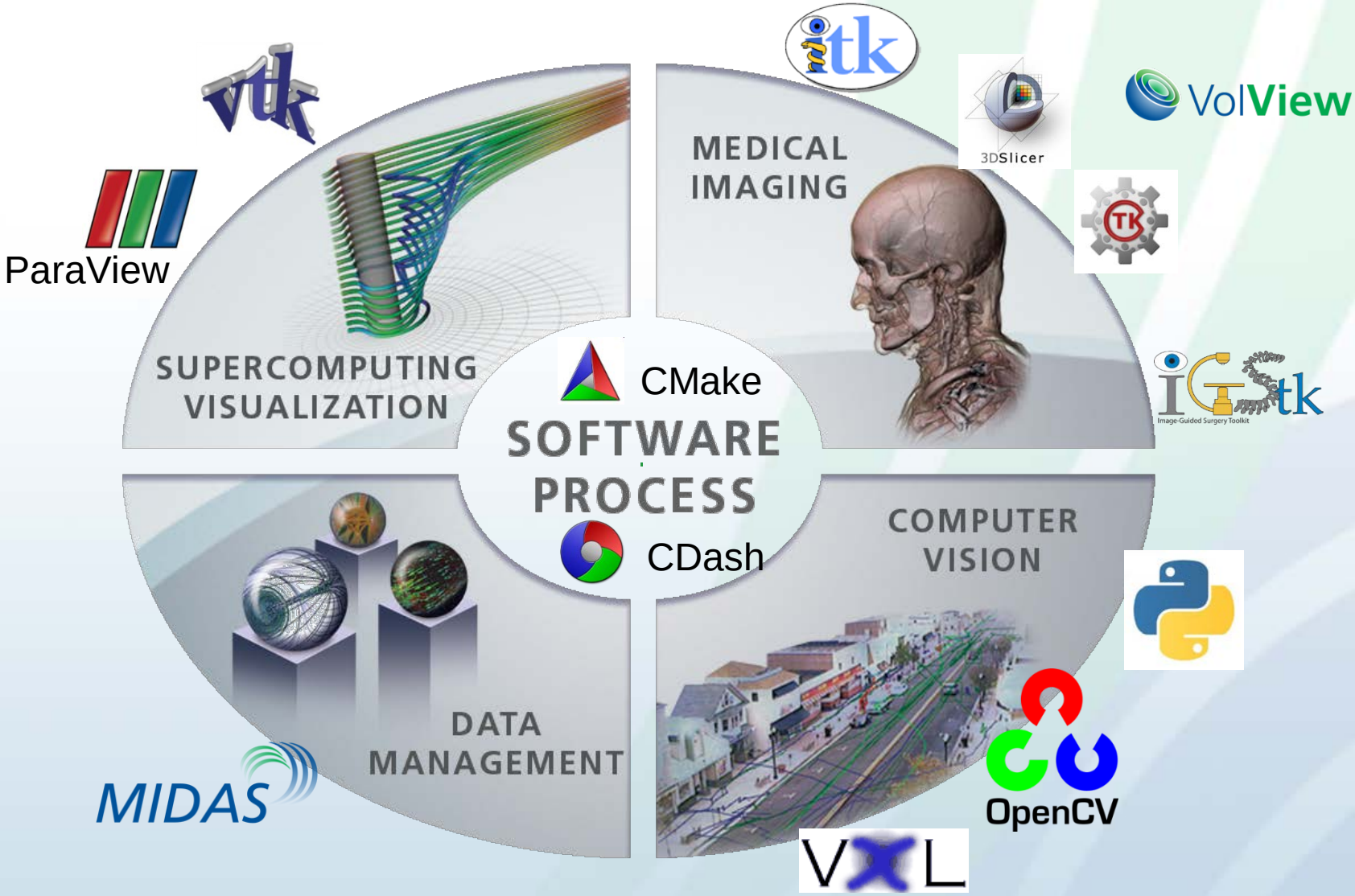




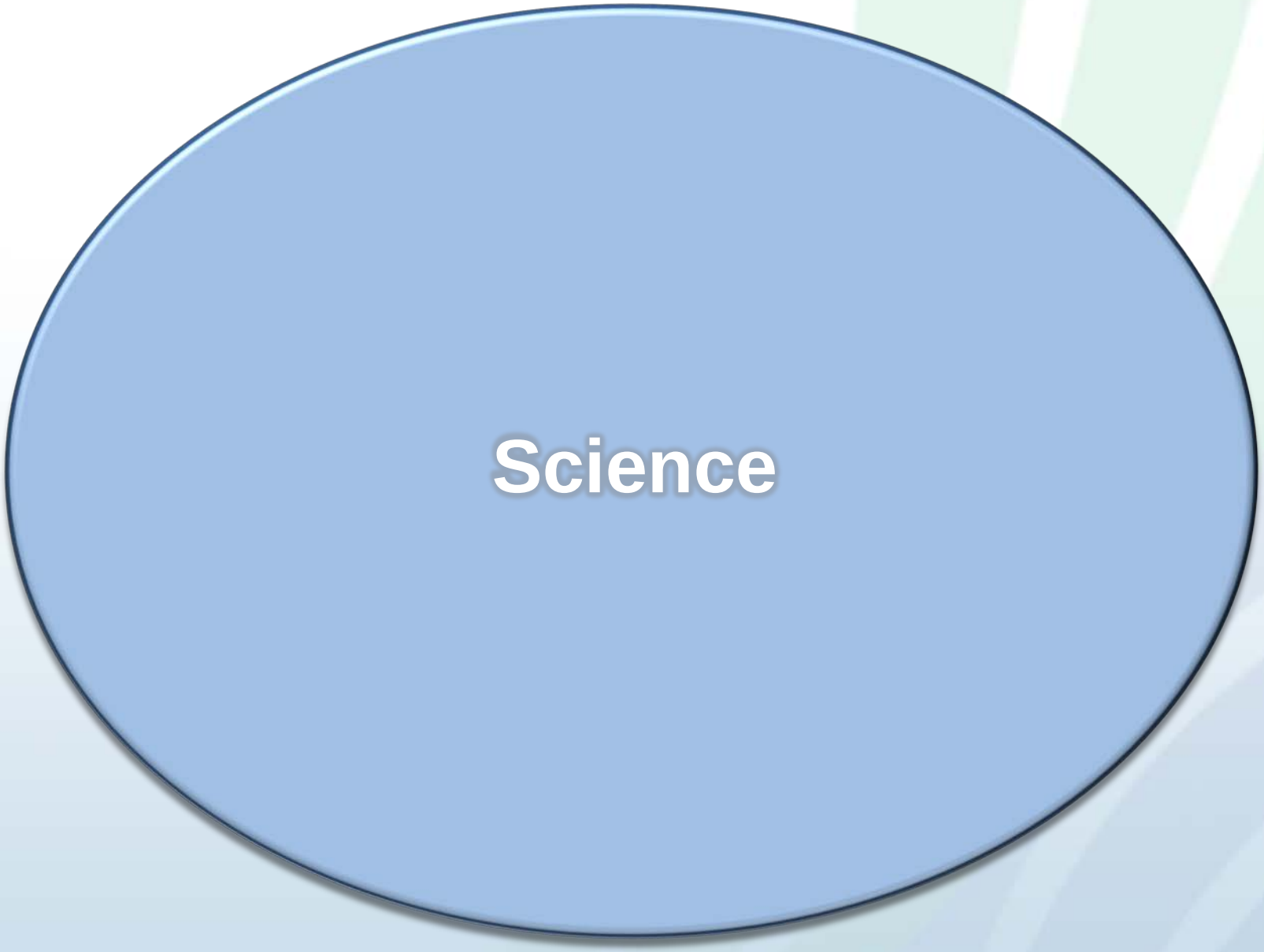




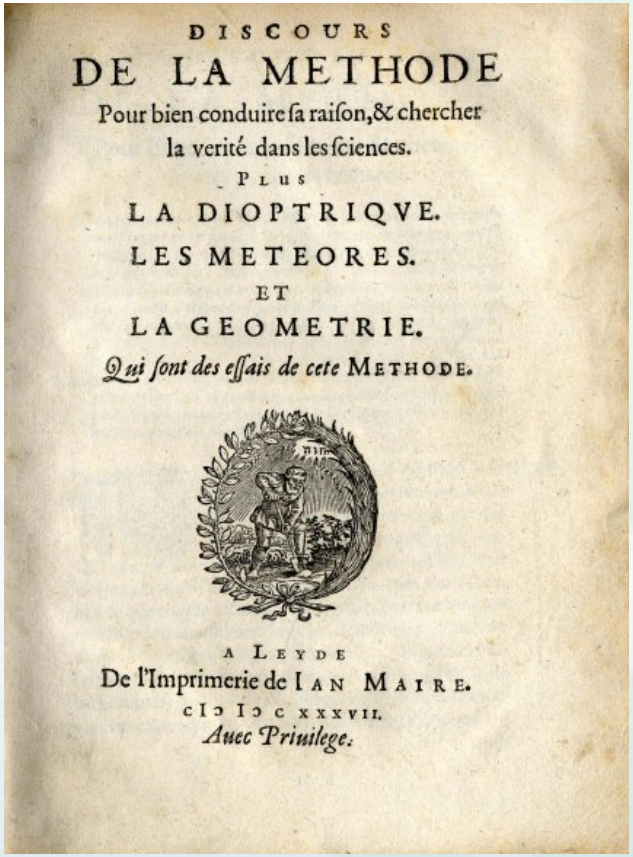



vtk
ParaView
SUPERCOMPUTING VISUALIZATION
itk
3DSlicer
VolView
MEDICAL IMAGING
IGSTK
Image-Guided Surgery Toolkit
CMake
SOFTWARE PROCESS
CDash
DATA MANAGEMENT
MIDAS
COMPUTER VISION
OpenCV
VXL

Kitware



Kitware



DISCOURS
DE LA METHODE
Pour bien conduire ſa raiſon,& chercher
la verité dans les ſciences.
PLUS
LA DIOPTRIQVE.
LES METEORES.
ET
LA GEOMETRIE.
Qui ſont des eſſais de cete METHODE.

A LEYDE
De l'Imprimerie de IAN MAIRE.
cIↃ IↃ c xxxvii.
Auec Priuilege.

Discourse on the (Scientific) Method,
Descartes 1637

DOUBTING EVERYTHING, and only believe in those things that are evidently true (REPRODUCIBLE)

Kitware

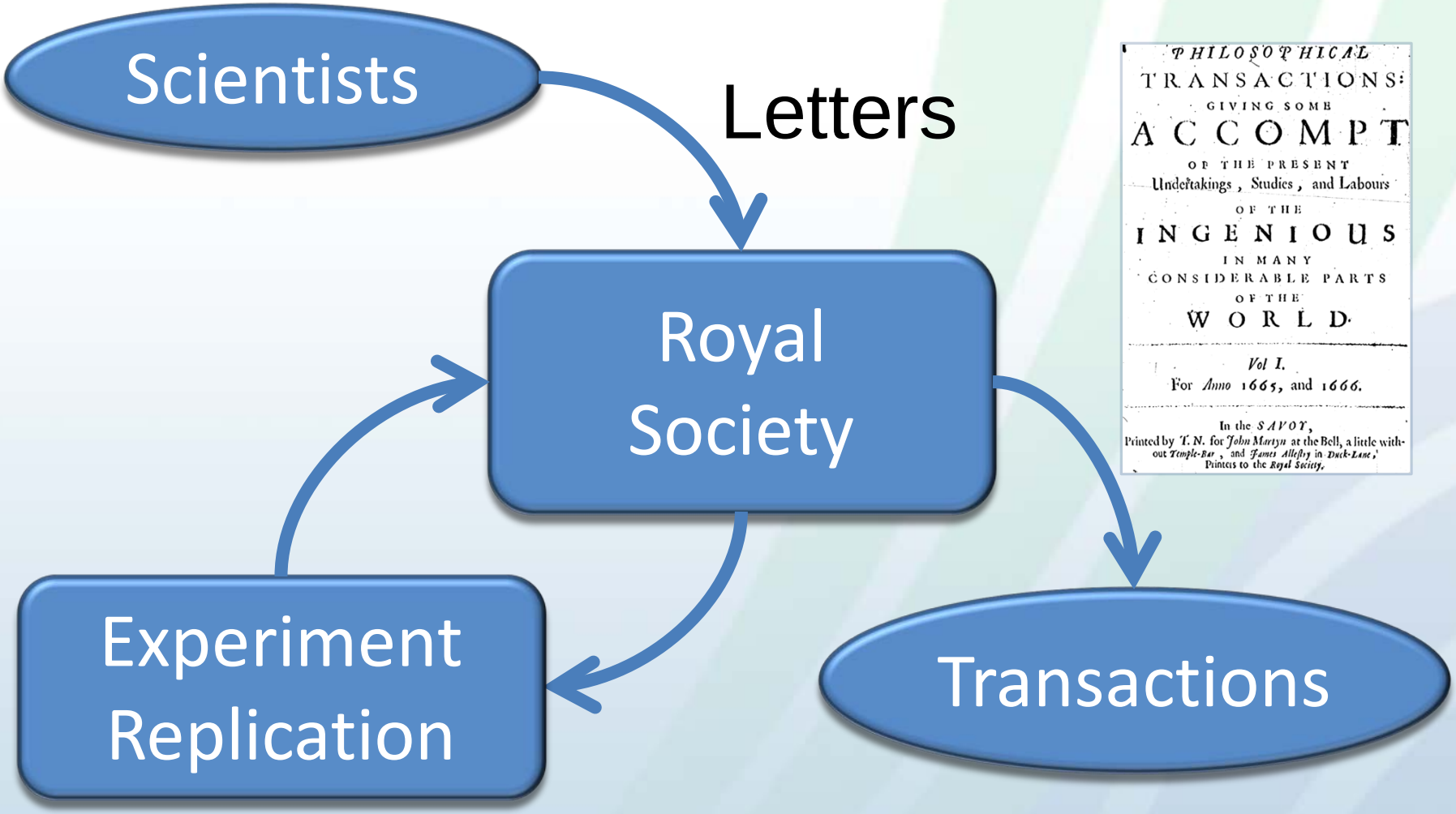
If it's not reproducible, it's not Science

Nullius in Verba

"take nobody's word for it"
Royal Society 1640

Kitware

Scientific Publishing Origins

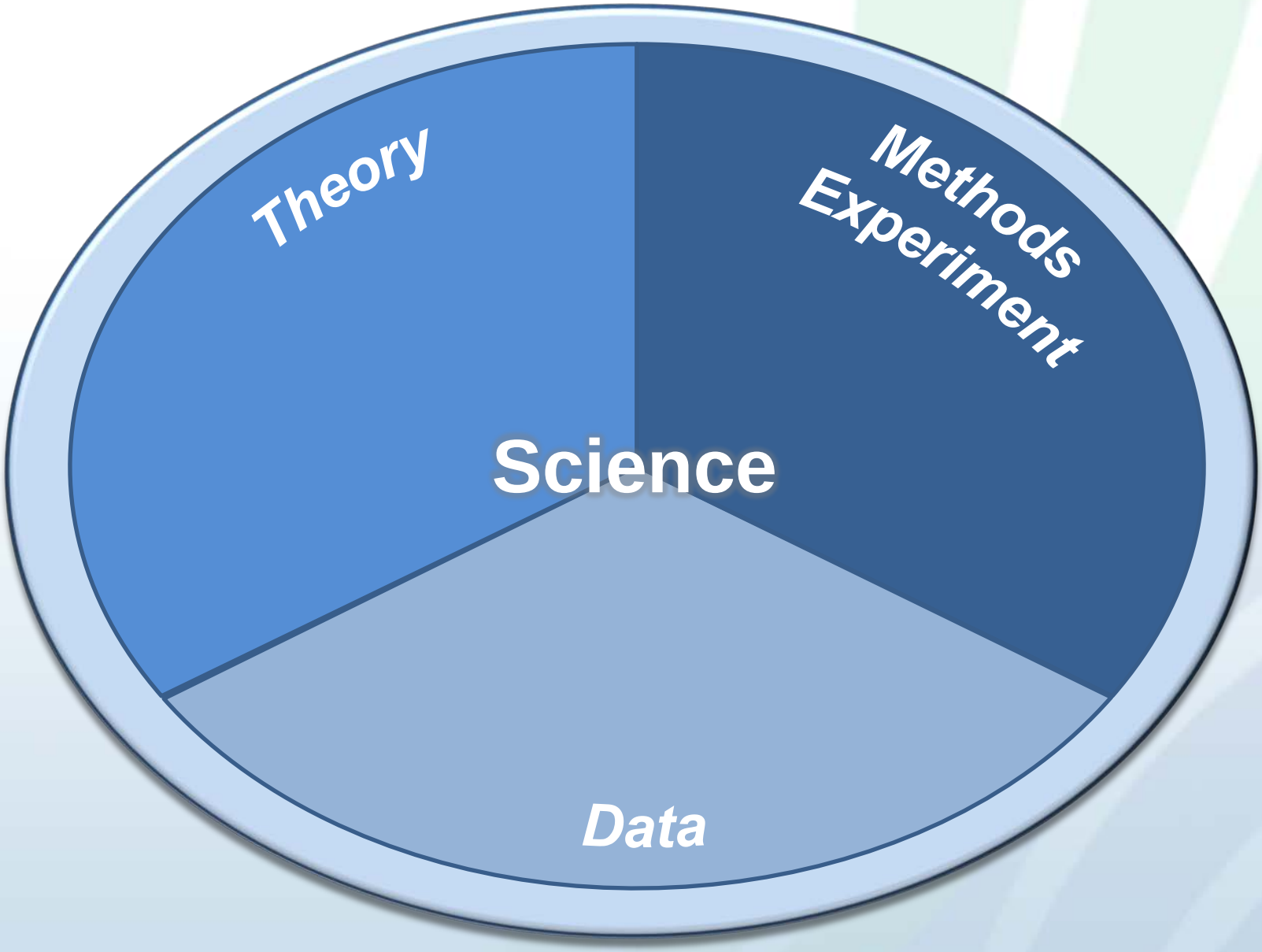


PHILOSOPHICAL TRANSACTIONS: GIVING SOME ACCOMPT OF THE PRESENT Undertakings, Studies, and Labours OF THE INGENIOUS IN MANY CONSIDERABLE PARTS OF THE WORLD.

Vol I. For Anno 1665, and 1666.

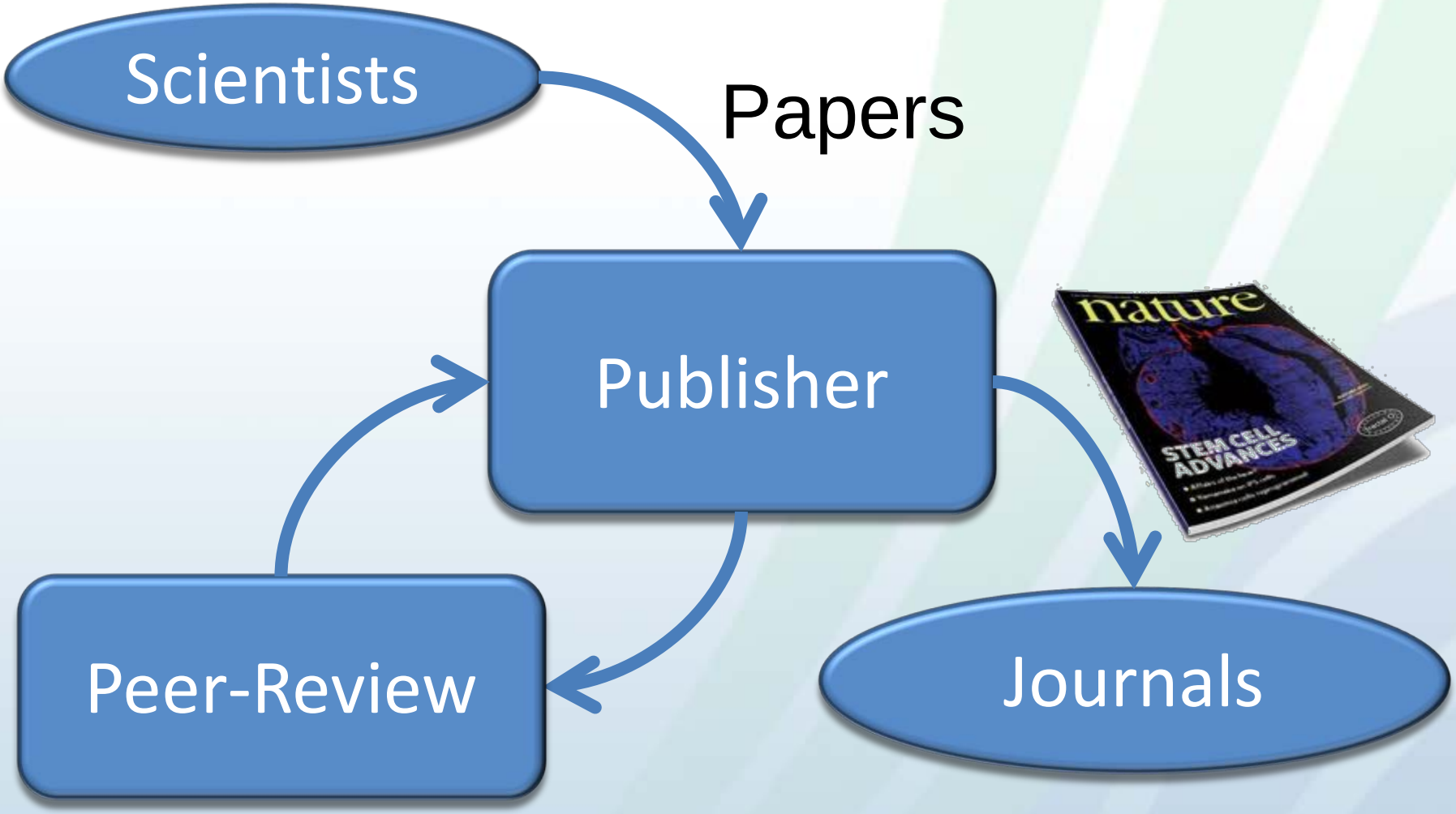
In the SAVOY, Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane, Printers to the Royal Society.

Kitware



Kitware

Evolution



Kitware

Career Pressures

"Publish or Perish" or what they taught me in Graduate School

Author

Kitware

Science is becoming computation

- "Softwa
  mathem
  languag
  Seidel f

SUVAT equations

In elementary physics the above

$$v = u + at \quad [1]$$

$$s = ut + \frac{1}{2}at^2 \quad [2]$$

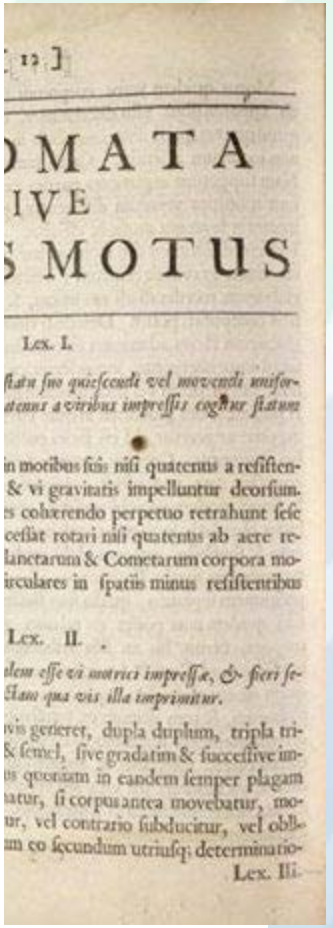
$$s = \frac{1}{2}(u + v)t \quad [3]$$

$$v^2 = u^2 + 2as \quad [4]$$

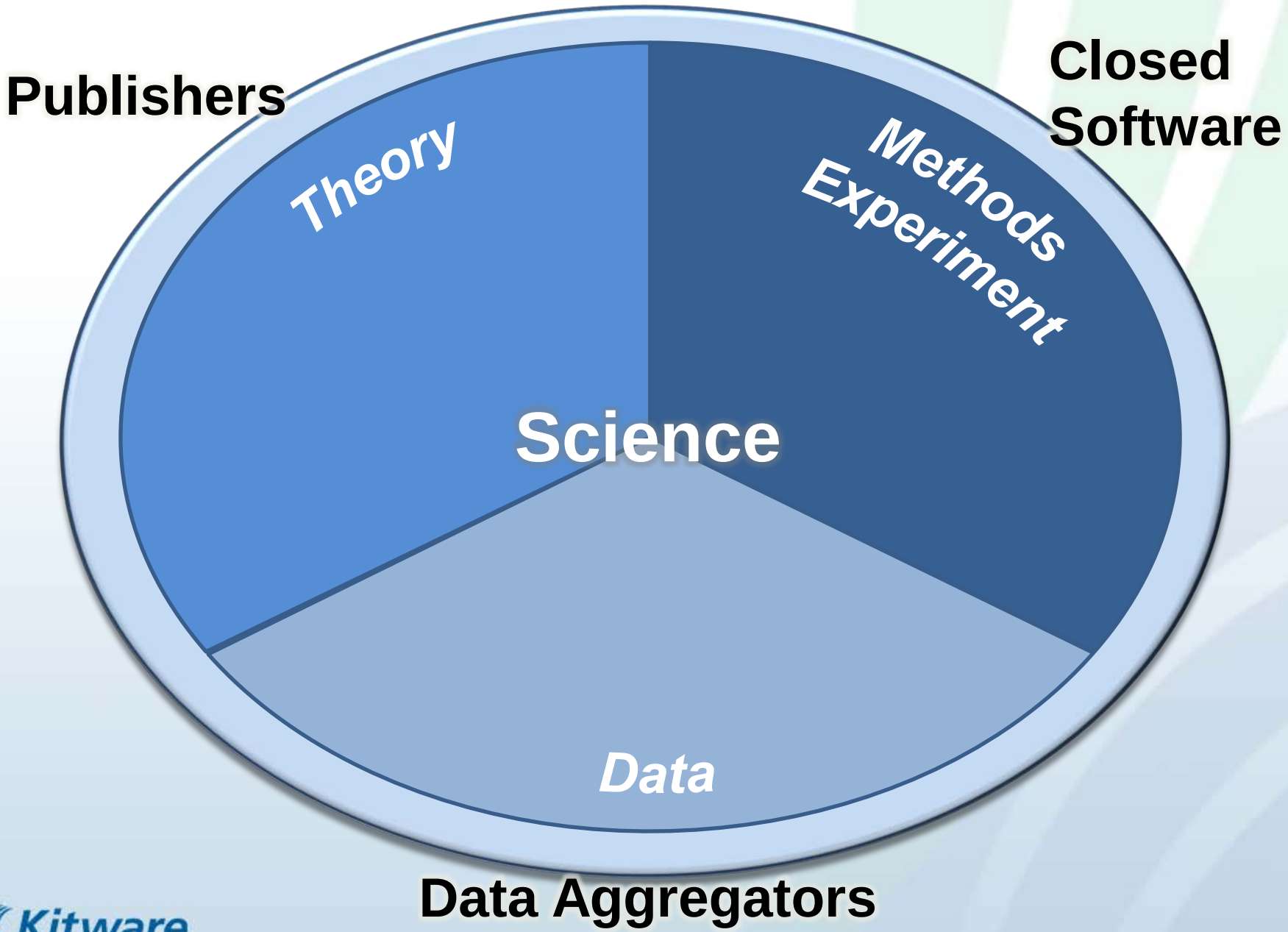
$$s = vt - \frac{1}{2}at^2 \quad [5]$$

where u has replaced v_0, s repla
displacement), u = initial velocity

```
switch (m_symmetry[i]) {
case S:
  m_moIndices[i] = indexMO++;
  m_cIndices.push_back(static_cast<unsigned int>(m_gtoCN.size()));
  // Normalization of the S-type orbitals (normalization used in JMol)
  // (8 * alpha^3 / pi^3)^0.25 * exp(-alpha * r^2)
  for(unsigned j = m_gtoIndices[i]; j < m_gtoIndices[i+1]; ++j) {
    m_gtoCN.push_back(m_gtoC[j] * pow(m_gtoA[j], 0.75) * 0.71270547);
  }
  break;
case P:
  m_moIndices[i] = indexMO;
  indexMO += 3;
  m_cIndices.push_back(static_cast<unsigned int>(m_gtoCN.size()));
  // Normalization of the P-type orbitals (normalization used in JMol)
  // (128 alpha^5 / pi^3)^0.25 * [x|y|z]exp(-alpha * r^2)
  for(unsigned j = m_gtoIndices[i]; j < m_gtoIndices[i+1]; ++j) {
    m_gtoCN.push_back(m_gtoC[j] * pow(m_gtoA[j], 1.25) * 1.425410941);
    m_gtoCN.push_back(m_gtoCN.back());
    m_gtoCN.push_back(m_gtoCN.back());
  }
  break;
case D:
  // Cartesian - 6 d components
  // Order in xx, yy, zz, xy, xz, yz
  m_moIndices[i] = indexMO;
  indexMO += 6;
  m_cIndices.push_back(static_cast<unsigned int>(m_gtoCN.size()));
  // Normalization of the P-type orbitals (normalization used in JMol)
  // xx|yy|zz: (2048 alpha^7/9pi^3)^0.25 [xx|yy|zz]exp(-alpha r^2)
  // xy|xz|yz: (2048 alpha^7/pi^3)^0.25 [xy|xz|yz]exp(-alpha r^2)
  for(unsigned j = m_gtoIndices[i]; j < m_gtoIndices[i+1]; ++j) {
    m_gtoCN.push_back(m_gtoC[j] * pow(m_gtoA[j], 1.75) * 1.645922781);
    m_gtoCN.push_back(m_gtoCN.back());
    m_gtoCN.push_back(m_gtoCN.back());
```



Kitware


Publishers
Closed Software
Theory
Methods Experiment
Science
Data
Data Aggregators

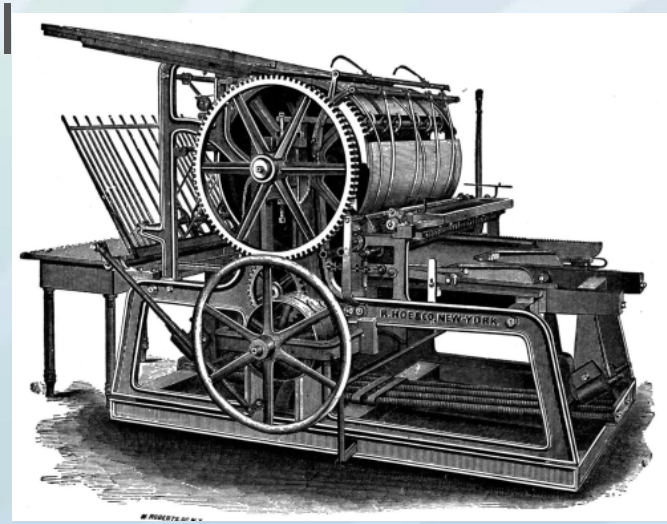
Kitware

Publishing in the Modern Age?

- Time to post a PDF file on the Web
  - Typically 1 hour, ~0 marginal cost

vs

- Time to publish a paper in a journal
  - Typically 2 years
- Cost to publish a paper in a journal
  - About 500€ / paper
- Cost to read the same paper
  - About 30€ / paper



Kitware

Failure of Reproducibility

- ***Nature*** *(March 2012)*
  - Glenn Begley, former head of cancer research at pharma giant Amgen
  - Lee M. Ellis, cancer researcher at the University of Texas

Found that more than 90% of papers published in science journals describing "landmark" breakthroughs in preclinical cancer research, are not reproducible, and are thus just plain wrong.

Example Reproducibility Challenge: White Matter Tracts in Medical Imaging (DTI Imaging at *MICCAI 2011)*

- 8 international teams participated
- 3D visualization and standardized comparison of different tractography
- All used the same diffusion MRI dataset

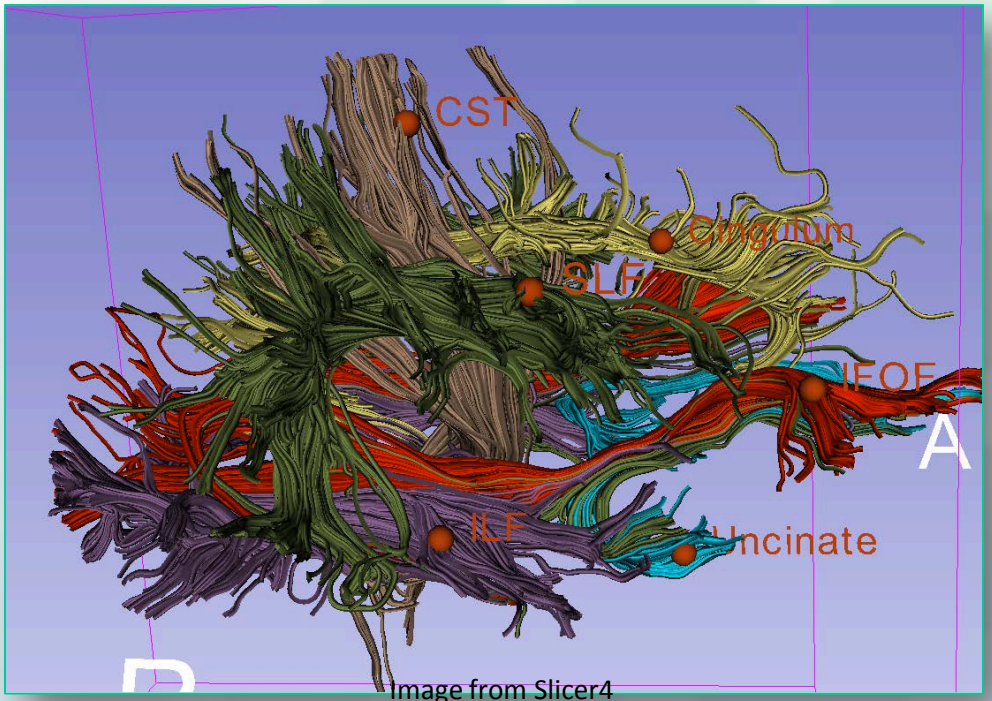
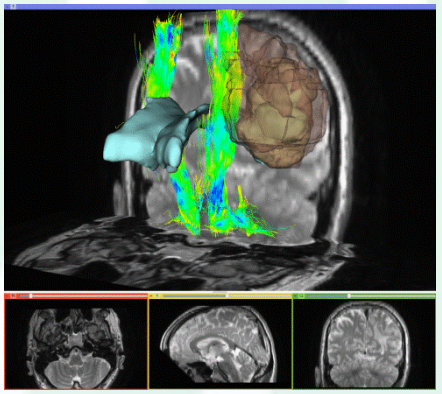


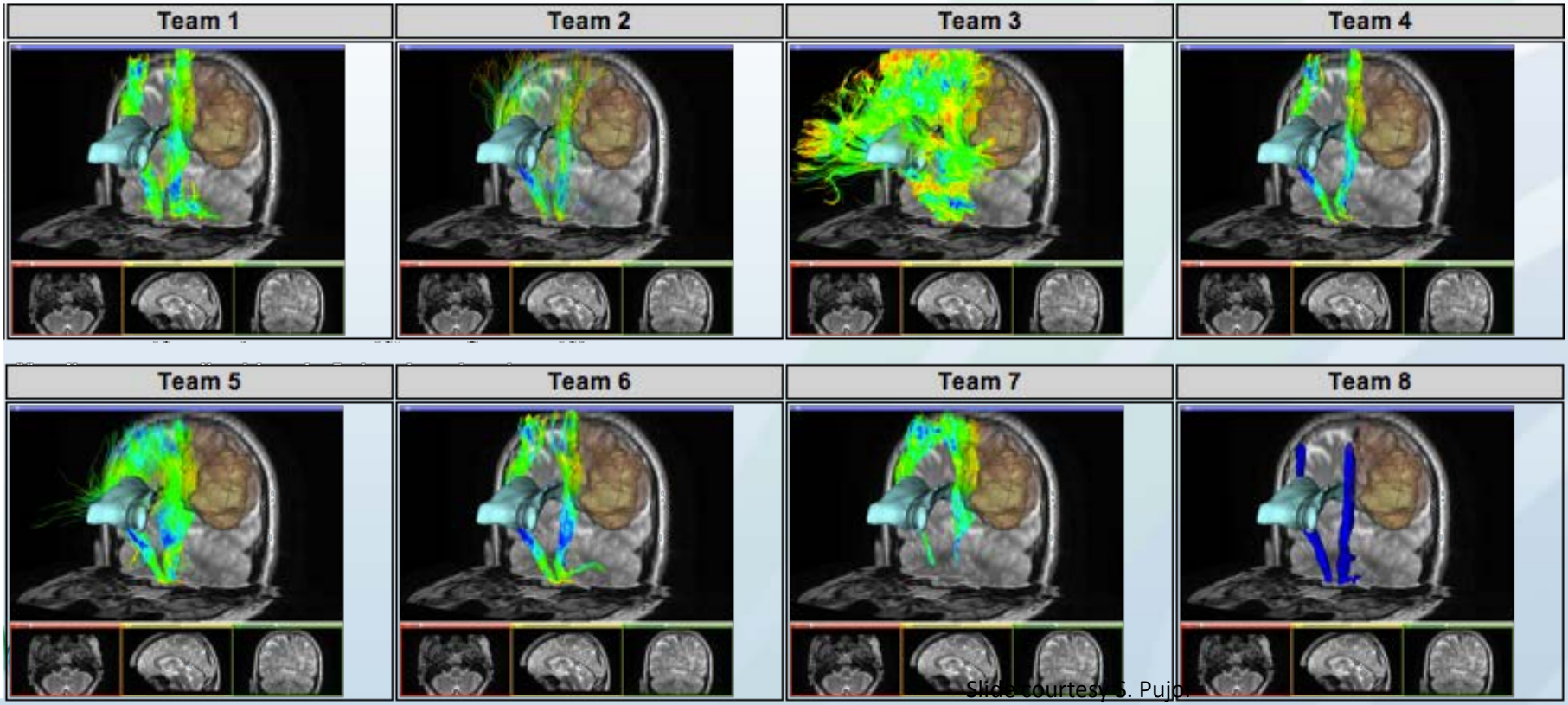
Image from Slicer4

Kitware

MICCAI Workshop Results

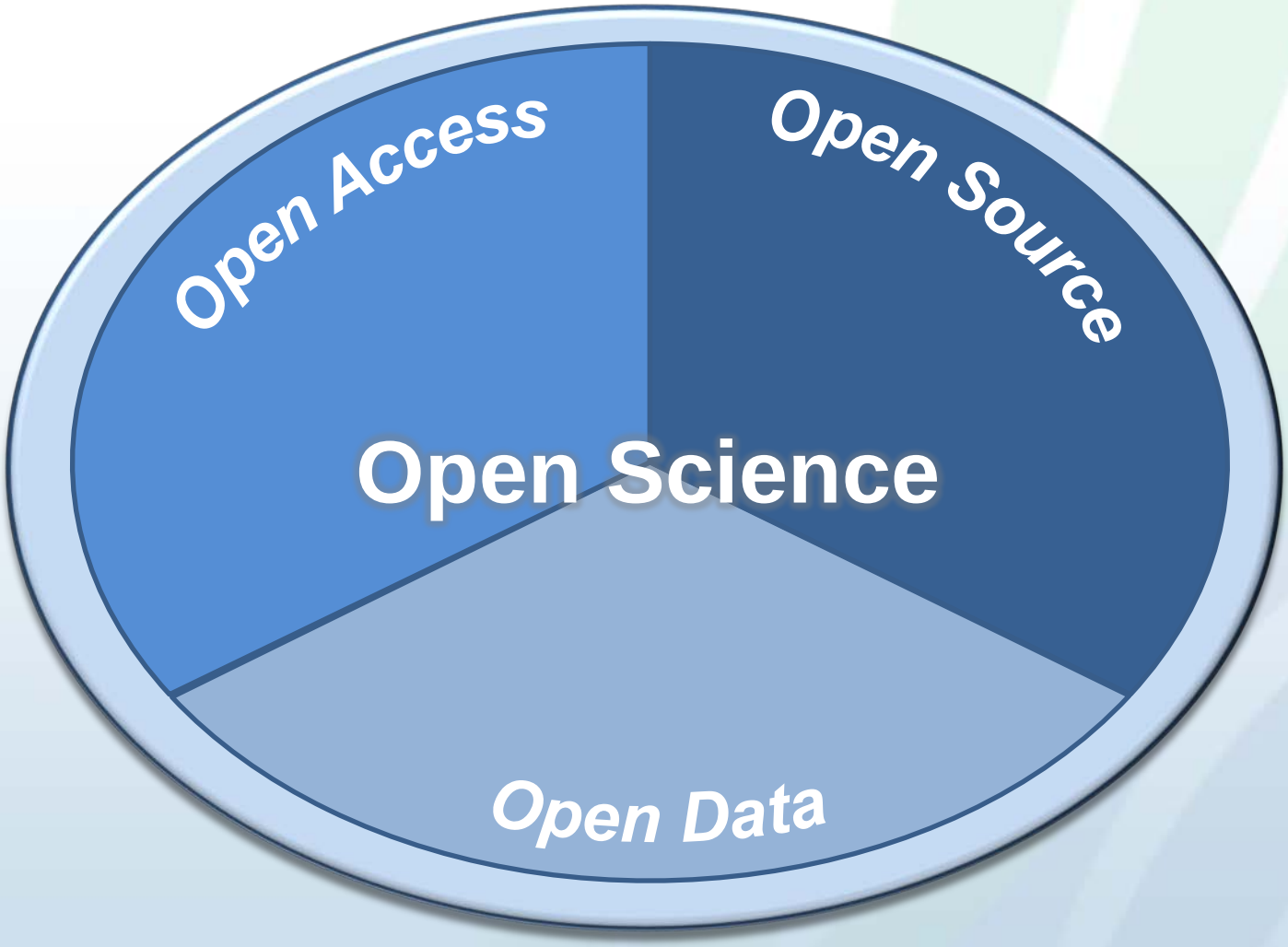
- Large **inter-algorithm** variability in finding the CST (cortico-spinal tract)
- How to compare?





Slide courtesy S. Pujol

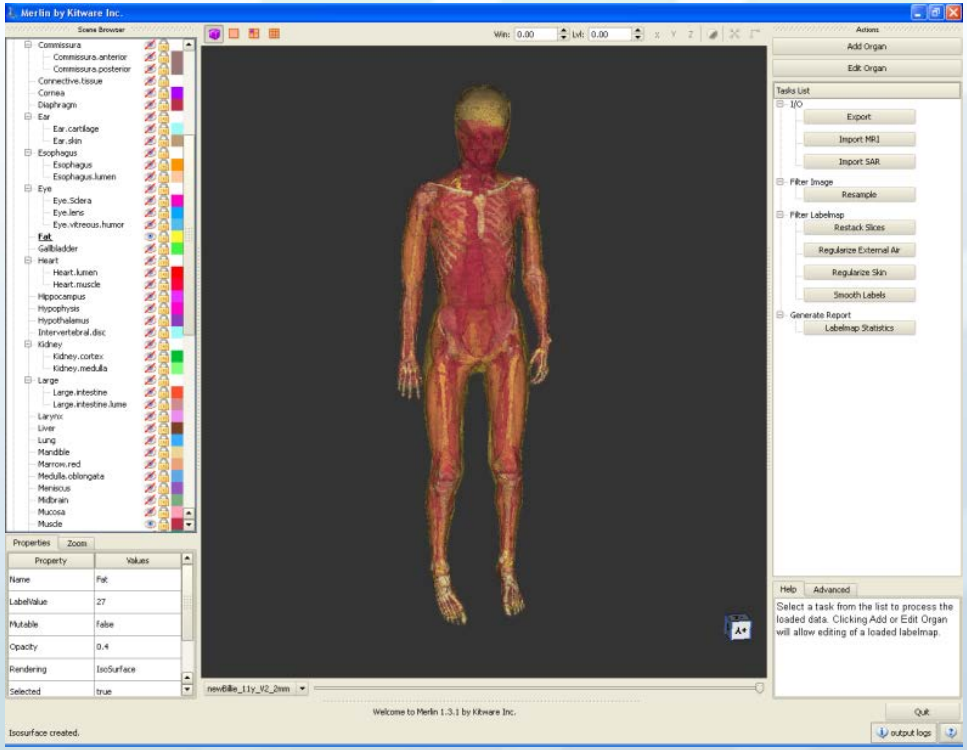
There is a better way



Kitware

CMake history in open science

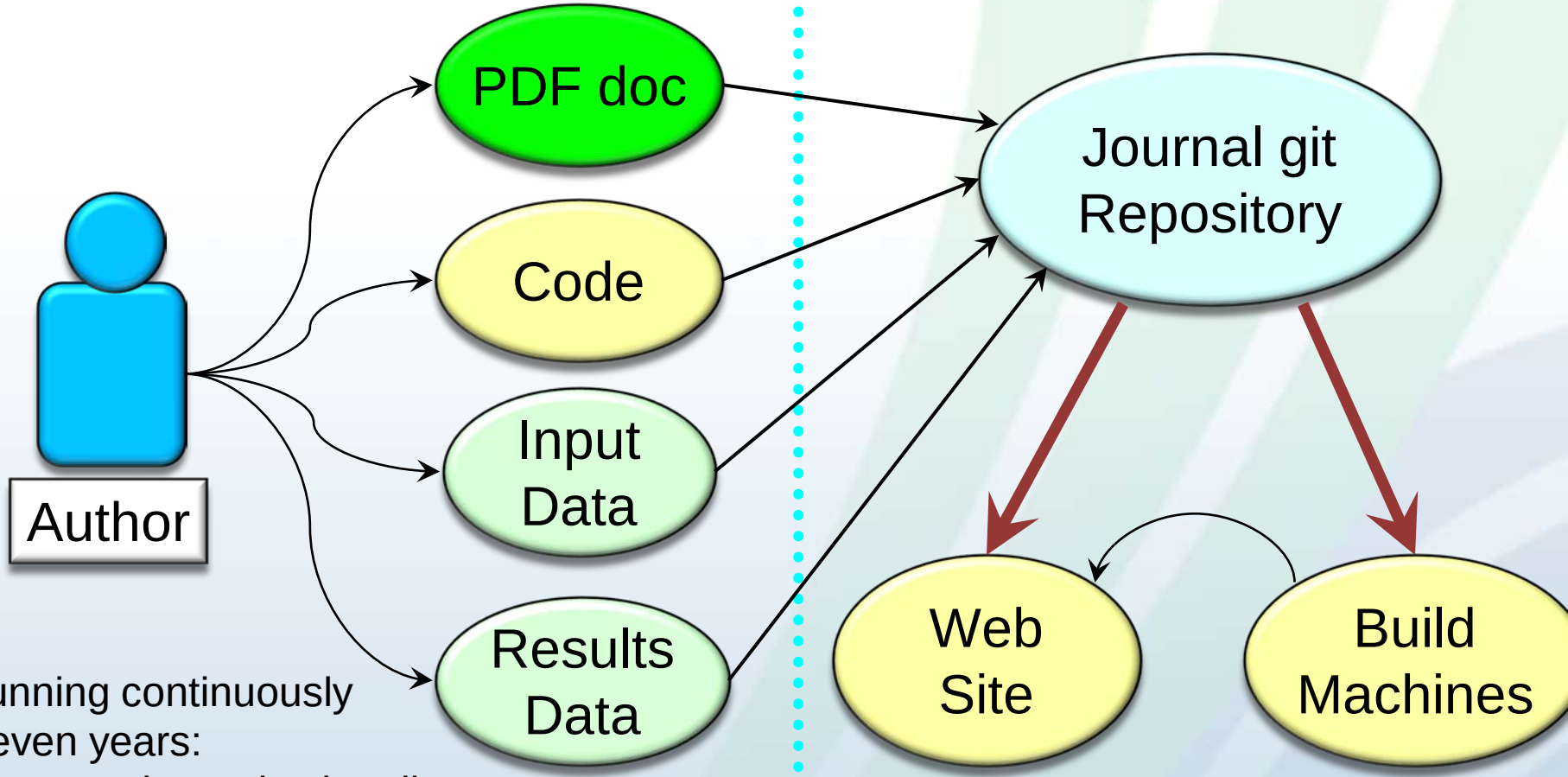
- US NIH Visible Human Project
  - First Data, CT/MR/Slice
  - Second Code (ITK)
- Happy to hear CMake in many of the presentations at FOSDEM



Kitware

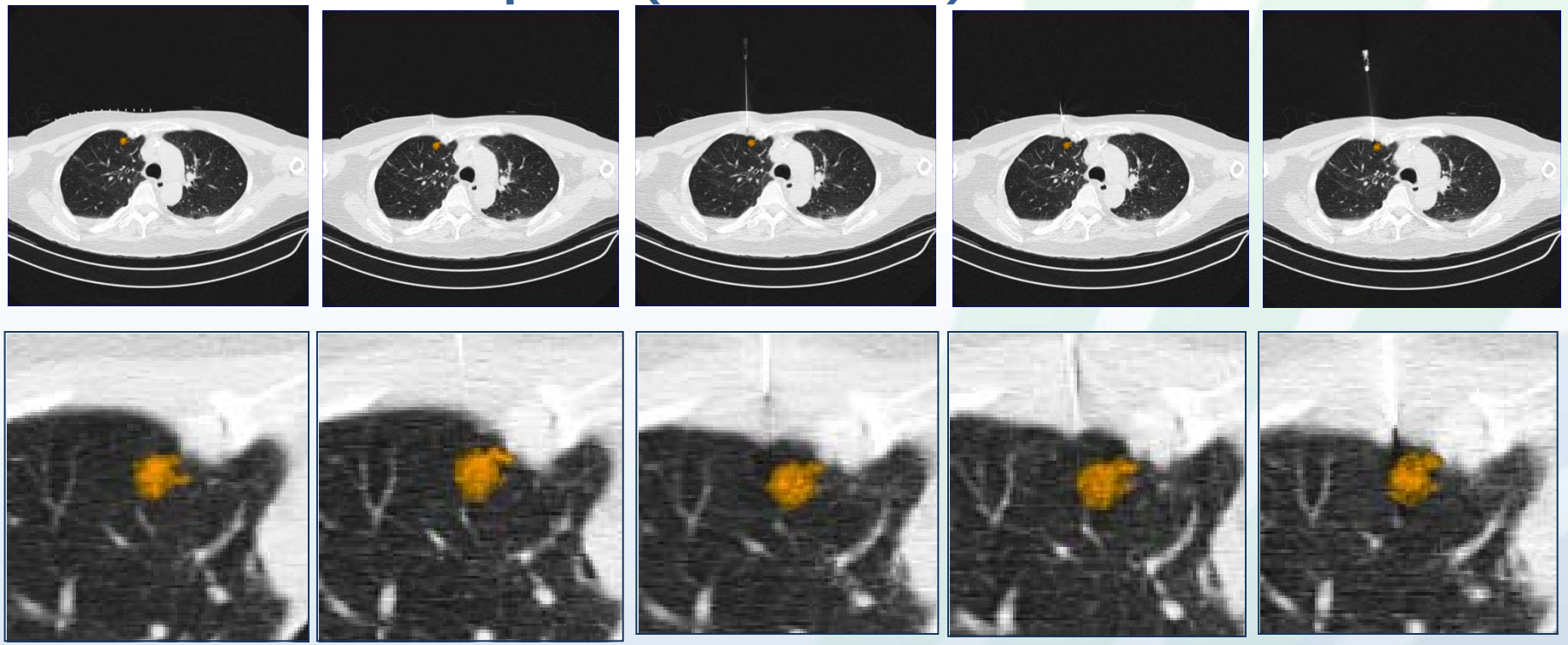
Reproducibility in action

The Insight Journal (since 2005): Submission & Automatic (Code) Review



Running continuously seven years:
3,571 registered subscribers
536 published articles
802 reviews

Lung Cancer Lesion Sizing
LSTK Example (NL0026)



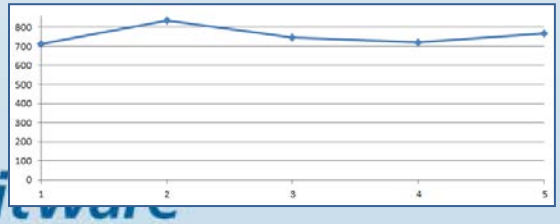
Series 1:
713 mm^3

Series 2:
836 mm^3

Series 3:
745 mm^3

Series 4:
722 mm^3

Series 5:
768 mm^3

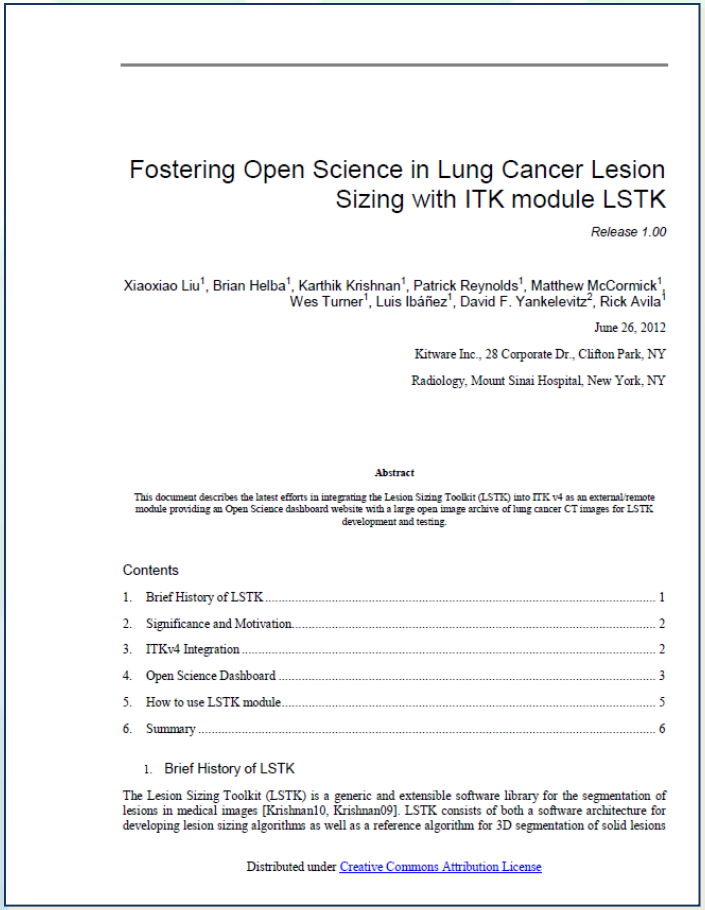


Mean
756.8 mm^3

Standard Deviation
49.2 mm^3

Kitware

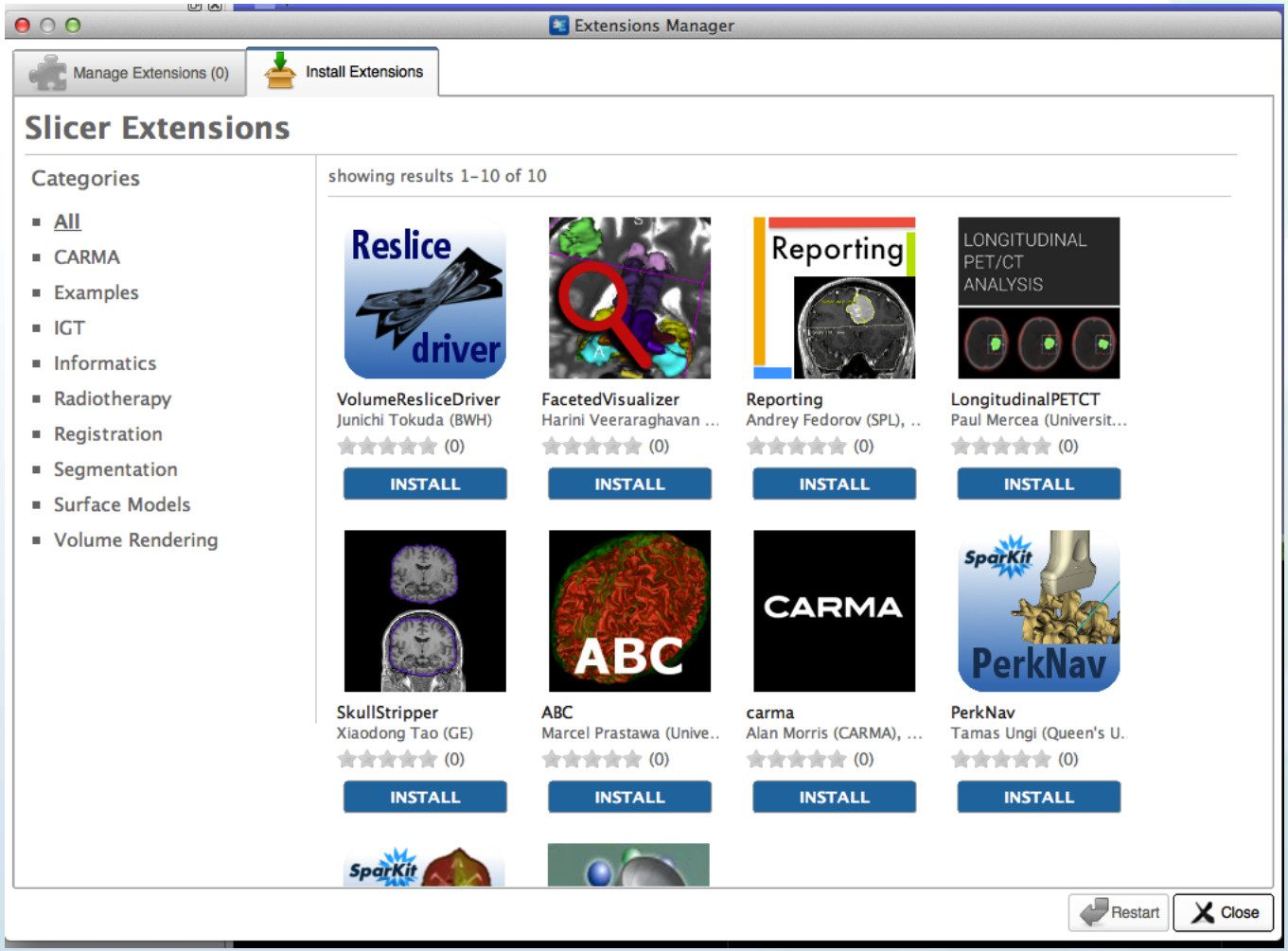
Open Access Publication on LSTK



http://www.insight-journal.org/browse/publication/869

Kitware

Slicer Extension Catalog



- Follows the “App Store” paradigm
- Extensions built nightly dashboards or contributed by users
- Manage revisions and dependencies
- Multiple CLI, Loadable, Python modules per extension

Kitware

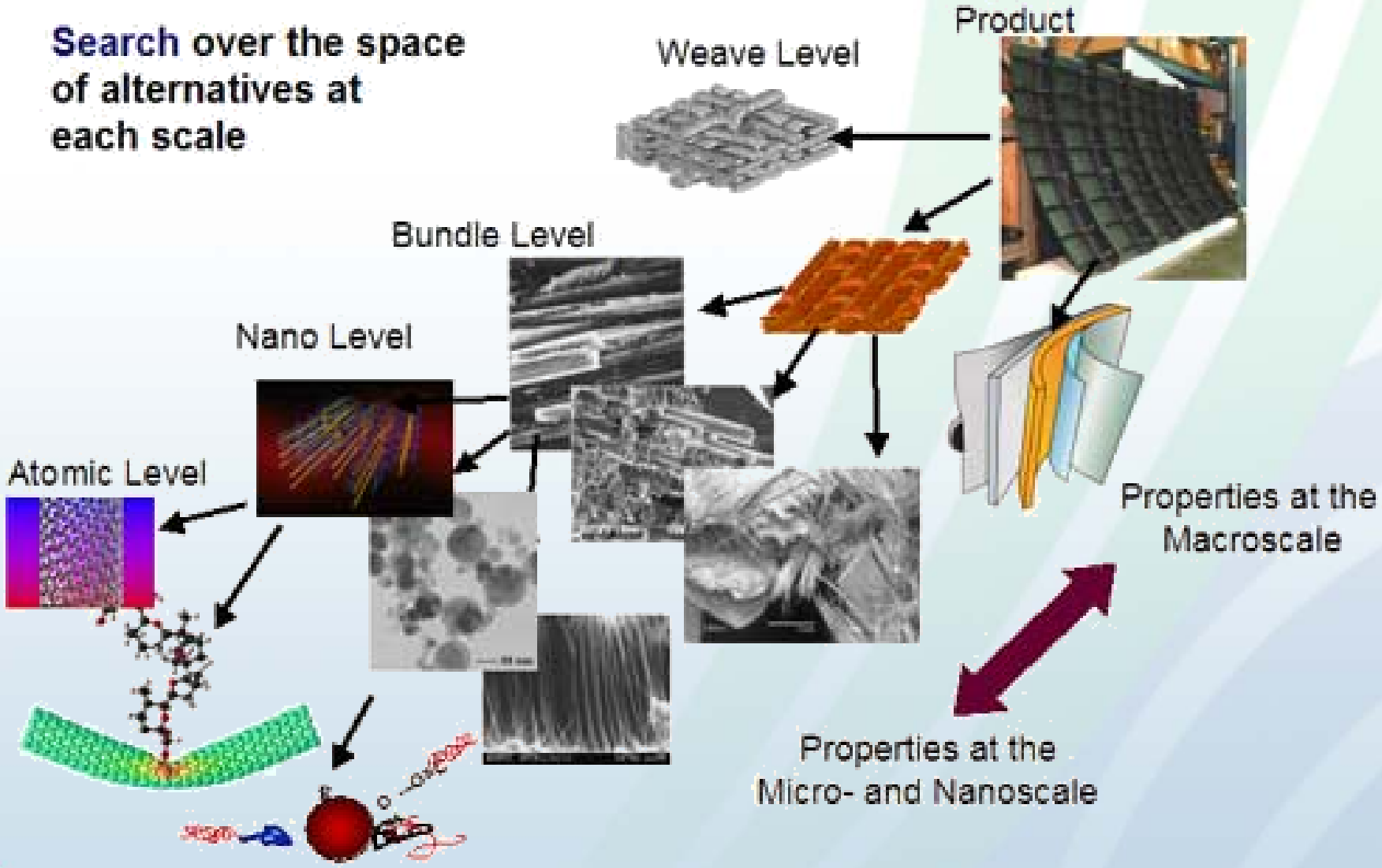
RunMyCode


Register | Sign In
runmycode
Search here ...
Search
Home
First visit?
Our offering
Submit your code
Search by themes
Advanced search
Help/FAQ
Our partners
The team
Contact us
The concept
As simple as 1,2,3
1. A researcher has an idea.
2. The researcher writes a paper based on this idea.
3. Using RunMyCode, the researcher creates a companion website associated with this paper. The companion website allows people to implement the methodology presented in the paper.
Learn more >>
1 Idea
2 Paper
3 Companion website
About
Concept
Purpose
Create your own companion website >>
RunMyCode goes global
Companion websites
Most Popular
Latest

Science is not done by one person and problems are getting bigger

Multiscale Design

Search over the space of alternatives at each scale

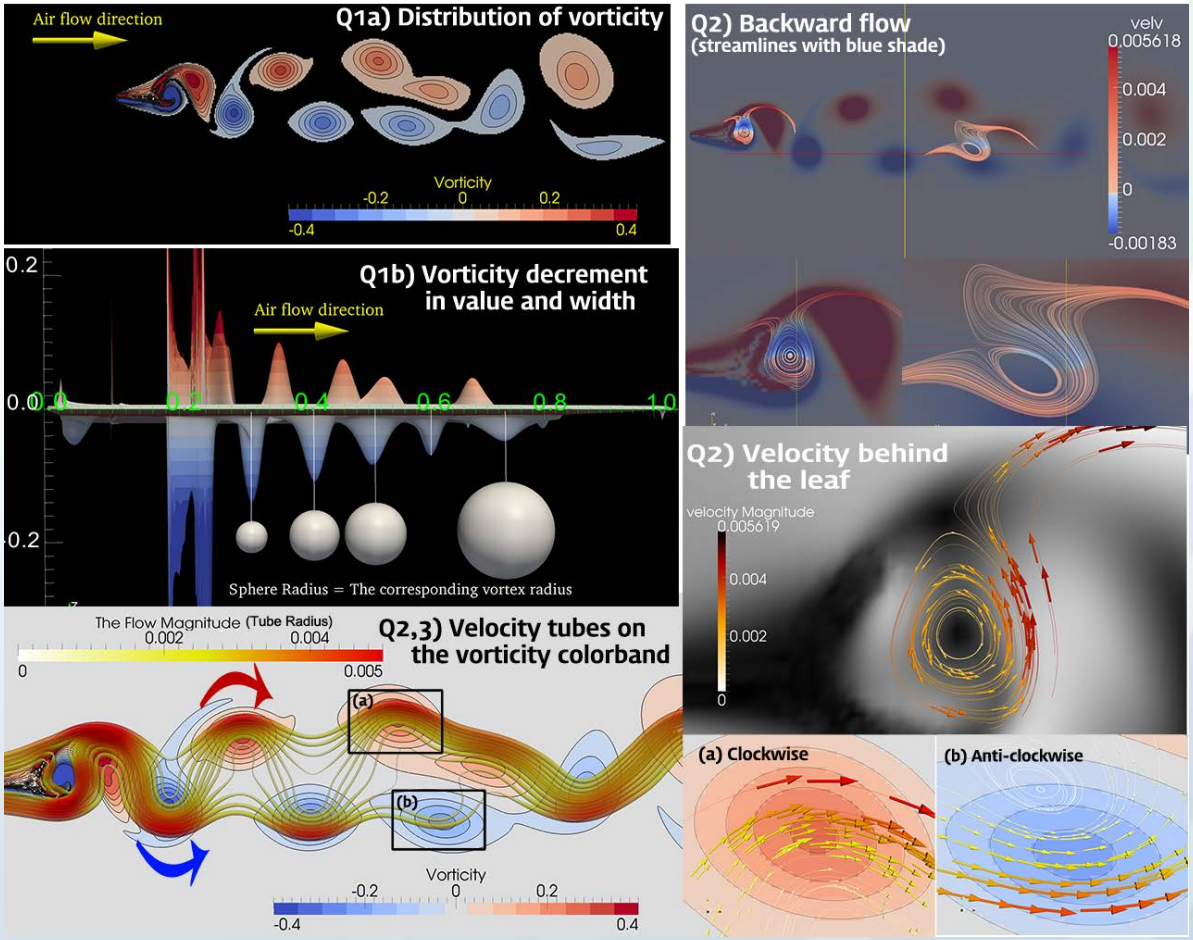


Kitware

Courtesy SCOREC RPI

Multi-Disciplinary

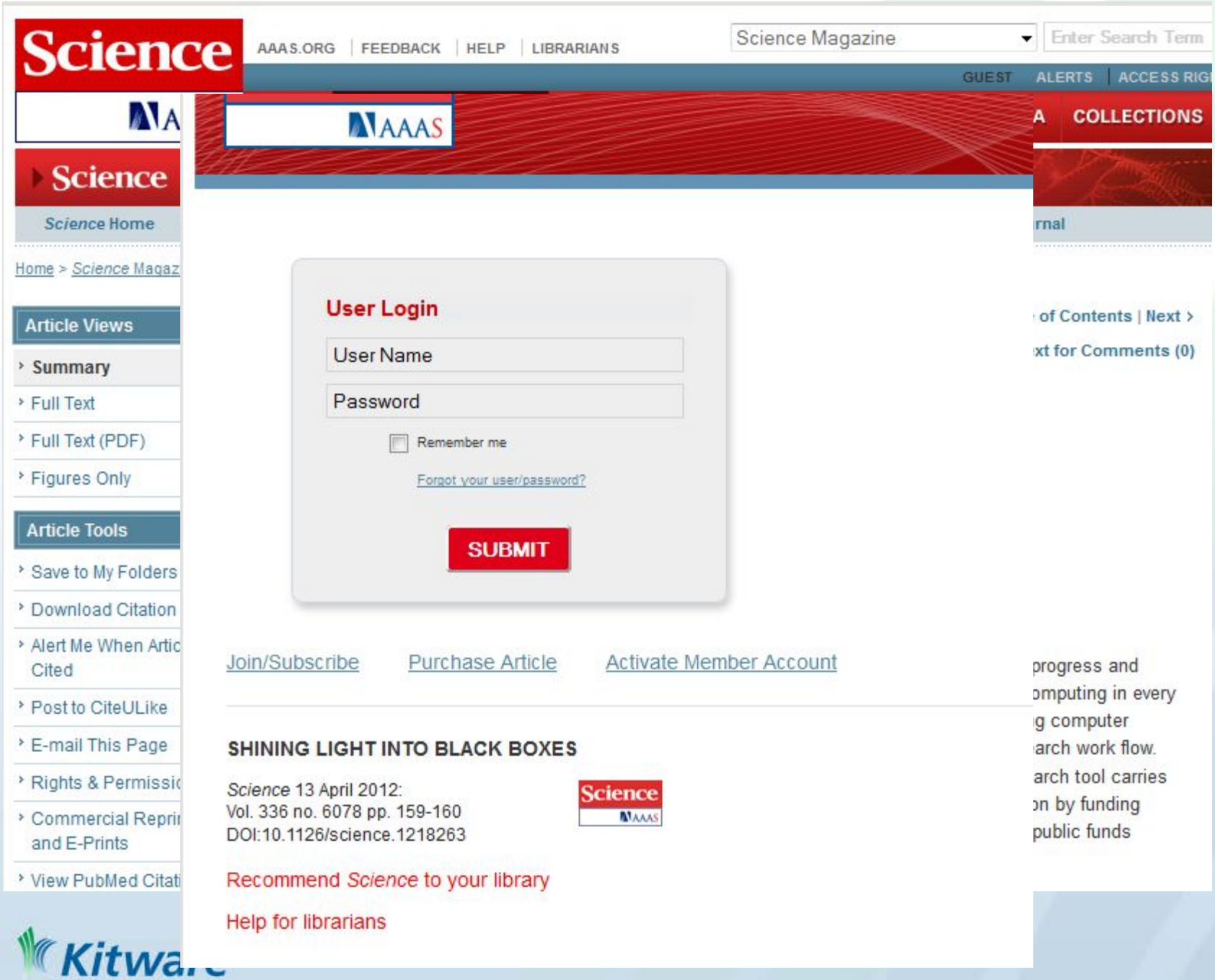
- Analysis
- Simulation
- Optimization



ParaView, Joo Hwi Lee and Namdi Brandon, UNC Visualization Class

Kitware

Signs and calls for change


Science
AAAS.ORG | FEEDBACK | HELP | LIBRARIANS
Science Magazine
Enter Search Term
GUEST ALERTS | ACCESS RIG
AAAS
COLLECTIONS
Science
Science Home
rnal
Home > Science Magaz
Article Views
Summary
Full Text
Full Text (PDF)
Figures Only
Article Tools
Save to My Folders
Download Citation
Alert Me When Artic Cited
Post to CiteULike
E-mail This Page
Rights & Permissic
Commercial Reprir and E-Prints
View PubMed Citati
User Login
User Name
Password
Remember me
Forgot your user/password?
SUBMIT
of Contents | Next >
xt for Comments (0)
Join/Subscribe
Purchase Article
Activate Member Account
progress and
omputing in every
g computer
arch work flow.
arch tool carries
on by funding
public funds
SHINING LIGHT INTO BLACK BOXES
Science 13 April 2012:
Vol. 336 no. 6078 pp. 159-160
DOI:10.1126/science.1218263
Science
AAAS
Recommend Science to your library
Help for librarians

PLOS | Open for Discovery

HOME ABOUT PUBLICATIONS PUBLISH SUPPOR

Open Access

The Case for Open Access (OA)

Open Access stands for unrestricted access and unrestricted reuse. Here's why that matters.

Most publishers own the rights to the articles in their journals. Anyone who wants to read the articles must pay to access them. Anyone who wants to use the articles in any way must obtain permission from the publisher and is often required to pay an additional fee.

Although many researchers can access the journals they need via their institution and think that their access is free, in reality it is not. The institution has often been involved in lengthy negotiations around the price of their site license, and re-use of this content is limited.

Paying for access to content makes sense in the world of print publishing, where providing content to each new reader requires the production of an additional copy, but online it makes much less sense to charge for content when it is possible to provide access to all readers anywhere in the world.

PLOS Takes a Different Approach

All PLOS content is published under the Creative Commons Attribution License (CC-BY), which was developed to facilitate open access – namely, free immediate access to, and unrestricted reuse of, original works of all types. Under this license, authors agree to make articles legally available for reuse, without permission or fees, for virtually any purpose. Anyone may copy, distribute, or reuse these articles, as long as the author and original source are properly cited. Additionally, the journal platform that PLOS uses to publish research articles is Open Source.

Kitware

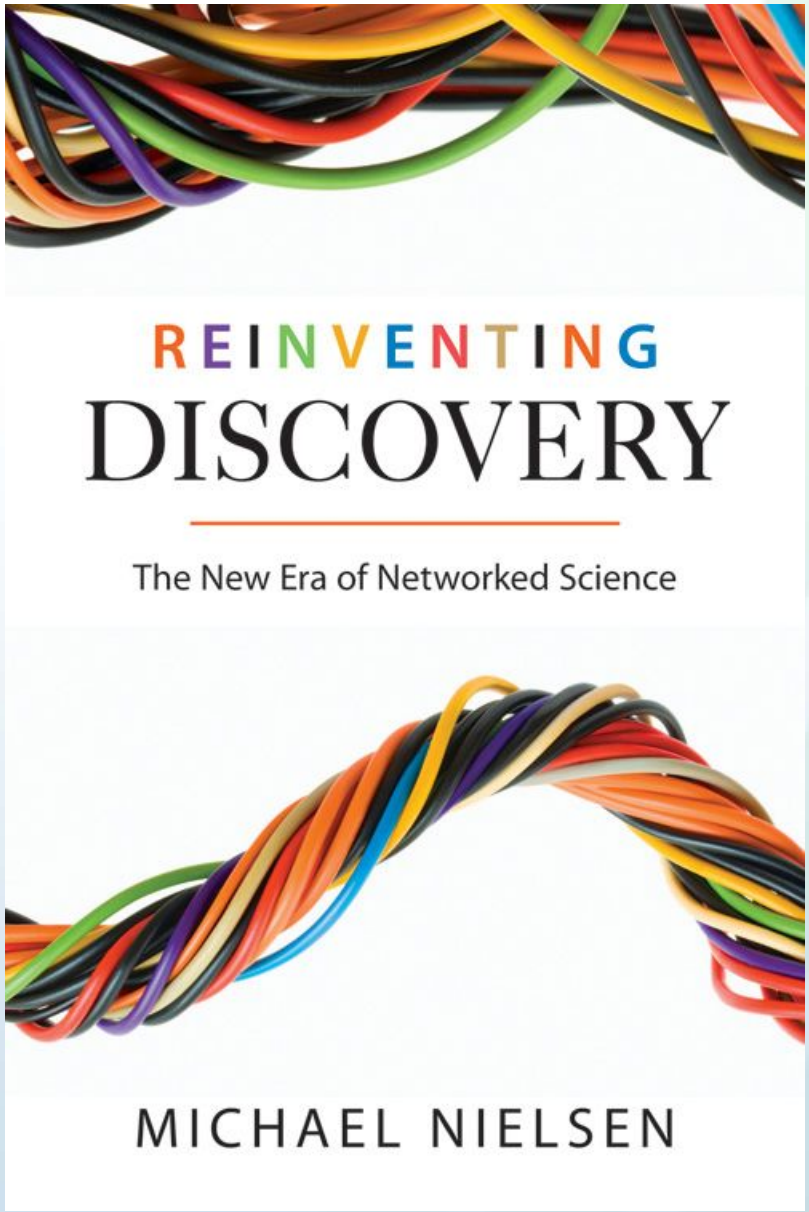
World wide web creator sees open access future for academic publishing

January 29, 2013 by Sunanda Creagh

"I think that the open access activists will win out": world wide web creator, Sir Tim Berners-Lee.

Activists pushing for free, open access to academic papers will eventually defeat publishers who seek to lock scholarly findings behind paywalls, the founder of the world wide web said today.

Kitware


REINVENTING
DISCOVERY
The New Era of Networked Science
MICHAEL NIELSEN

Kitware

Panton Principles

Principles for Open Data in Science

Science is based on building on, reusing and openly criticising the published body of scientific knowledge.

For science to effectively function, and for society to reap the full benefits from scientific endeavours, it is crucial that science data be made **open**.

By open data in science we mean that it is freely available on the public internet permitting any user to download, copy, analyse, re-process, pass them to software or use them for any other purpose without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. **To this end data related to published science should be explicitly placed in the public domain.**

Formally, we recommend adopting and acting on the following principles:

1. Where data or collections of data are published it is critical that they be published with a clear and explicit statement of the wishes and expectations of

Web buttons

Get an **open data web button** for your project!

OPEN DATA OPEN DATA

Related Links

Open Science Working Group - Open Knowledge Foundation

Open Definition - Defining the Open in Open Data and Content

Is It Open Data?

Science Commons - Protocol for Implementing Open Access Data

Open Knowledge Foundation

sciencecodemanifesto.org

Science Code Manifesto

Manifesto Discussion Endorse Resources About

Software is a cornerstone of science. Without software, twenty-first century science would be impossible. Without better software, science cannot progress.

But the culture and institutions of science have not yet adjusted to this reality. We need to reform them to address this challenge, by adopting these five principles:

Code	All source code written specifically to process data for a published paper must be available to the reviewers and readers of the paper.
Copyright	The copyright ownership and license of any released source code must be clearly stated.
Citation	Researchers who use or adapt science source code in their research must credit the code's creators in resulting publications.
Credit	Software contributions must be included in systems of scientific assessment, credit, and recognition.
Curation	Source code must remain available, linked to related materials, for the useful lifetime of the publication.

Kitware

U.S.Department of Health & Human Services

National Institutes of Health Public Access

The Public Access Policy ensures that the public has access to the published results of NIH funded research to help advance science and improve human health.

Home

- 1. Determine Applicability
- 2. Address Copyright
- 3. Submit paper to PMC

Overview

The NIH Public Access Policy ensures that the public has access to the published results of NIH funded research. It requires scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to the digital archive PubMed Central *upon acceptance for publication*. To help advance science and improve human health, the Policy requires that these papers are accessible to the public on PubMed Central no later than 12 months after publication.

RESEARCH COUNCILS UK

Excellence with Impact

- Home
- Research Policy
  - Eligibility for Research Council funding
  - How to apply for research funding

Home > Research Policy > Policy on Open Access

RCUK Policy on Open Access

Free and open access to publicly-funded research offers significant social and economic benefits. The Government, in line with its overarching commitment to transparency and open data, is committed to ensuring that such research should be freely accessible. As major bodies charged with investing public money in research, the Research Councils take very seriously their responsibilities in making the outputs from this research publicly available – not just to other researchers, but also to potential users in business,

ndings (the 'Finch
olished. The report sets
earch Councils have used

RESEARCH in the national interest – enabling the future

Australian Government
Australian Research Council

Applicants

RMS Login | Site Map | Contacts

Search | Search

Home | About ARC | Minister | National Competitive Grants Program | Information for Applicants | Media | General Information | Research Excellence | ARC-Supported Activities

INFORMATION FOR APPLICANTS

- Appeals
- Application closing dates
- ARC Open Access Policy
- Assessor Reports & Rejoinders
- Certification Proforma
- Eligibility Exemption & Ruling
- FOR, RFCD, SEO, ANZSIC Codes
- Funding Agreements
- Funding Outcomes
- Funding Rules
- Information for managers
- Information for researchers
- Instructions to applicants
- International collaboration

You are here: Home › Information for Applicants › ARC Open Access Policy

ARC Open Access Policy

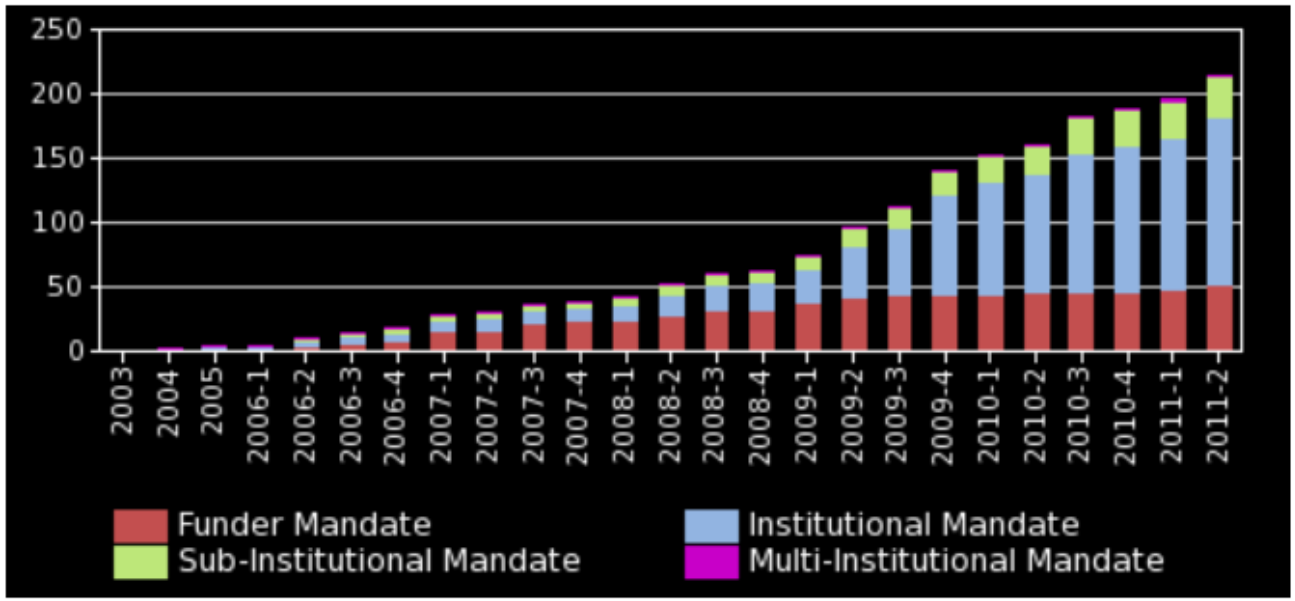
Printer frindly version - PDF Format (450KB) Word Format (125KB)

The ARC has introduced a new open access policy for ARC funded research which takes effect from 1 January 2013. According to this new policy the ARC requires that any publications arising from an ARC supported research project must be deposited into an open access institutional repository within a twelve (12) month period from the date of publication.

The ARC understands that some researchers may not be able to meet the new requirements initially because of current legal or contractual obligations. In these cases, Final Reports must provide reasons why publications derived from a Project, Award, or Fellowship have not been deposited in an open access institutional repository within the twelve month period. The policy will be incorporated into all new Funding Rules and Agreements released after 1 January 2013. It will not be applied retrospectively to pre-existing Funding Rules and Agreements.

http://roarmap.eprints.org/

The Registry of Open Access Mandatory Archiving Policies (ROARMAP [17]) is a searchable international database charting the growth of open-access mandates adopted by universities, research institutions and research funders that require their researchers to provide open access to their peer-reviewed research articles by self-archiving them in an open access repository. To date, mandates have been adopted by over 150 universities and over 50 research funders worldwide (see figure below):



Kitware

Publishing:
Some Economic Repercussions

- **Subscription costs are out of control**
  - **Harvard University:** canceling “too expensive” journal subscriptions due to expense. Asking professors to publish in open access journals.
  - **UK:** Minister of Science David Willetts that all publicly funded research should be published as open access
  - **World Bank** announced that all existing and new publications, reports and documents will be open access by July 2012.
  - **Boycott of Elsevier:**
    - **E.g., In 2011: > $7K for a subscription to *Theoretical Computer Sciences***

Threatening access to scientific results

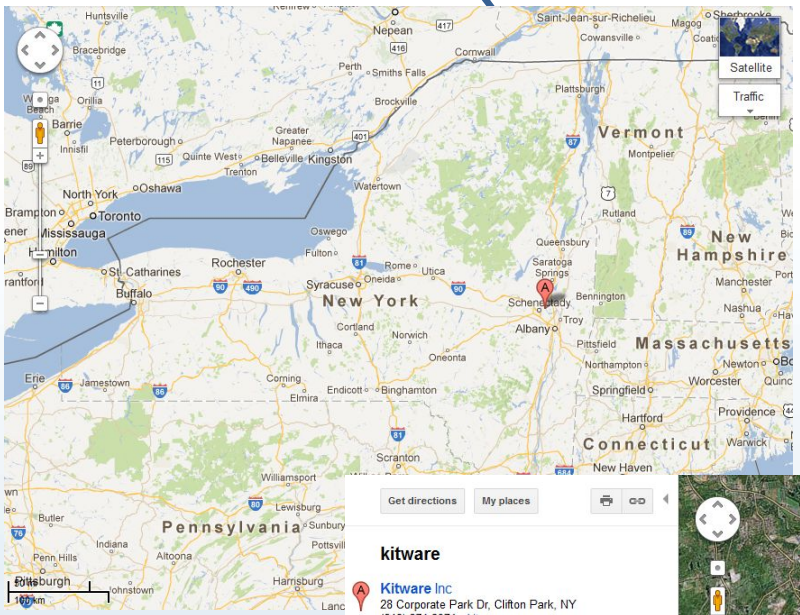
Kitware

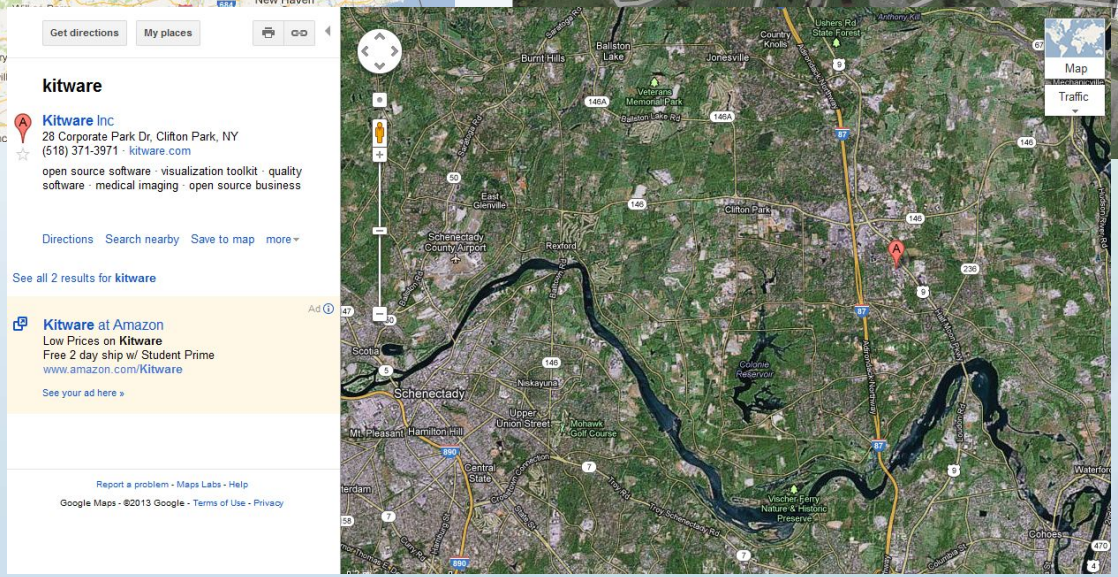
DARPA XDATA

- Current DoD systems and processes for handling and analyzing information cannot be efficiently or effectively scaled to meet this challenge.
- Finally, to enable large scale data processing in a wide range of potential settings, XDATA plans to release open-source software toolkits to enable collaboration among the applied mathematics, computer science and data visualization communities.
- Q48. Please elaborate on your open-source vision. Do you mean public open-source or can it include open APIs, but a proprietary platform with government purpose rights?
- A48. It depends on the proposal. Proprietary platforms with APIs will be considered in exceptional circumstances; however, in order to facilitate transition and use across enterprise platform for the government, unlimited rights and public open source is strongly encouraged.

Science can learn from software devs

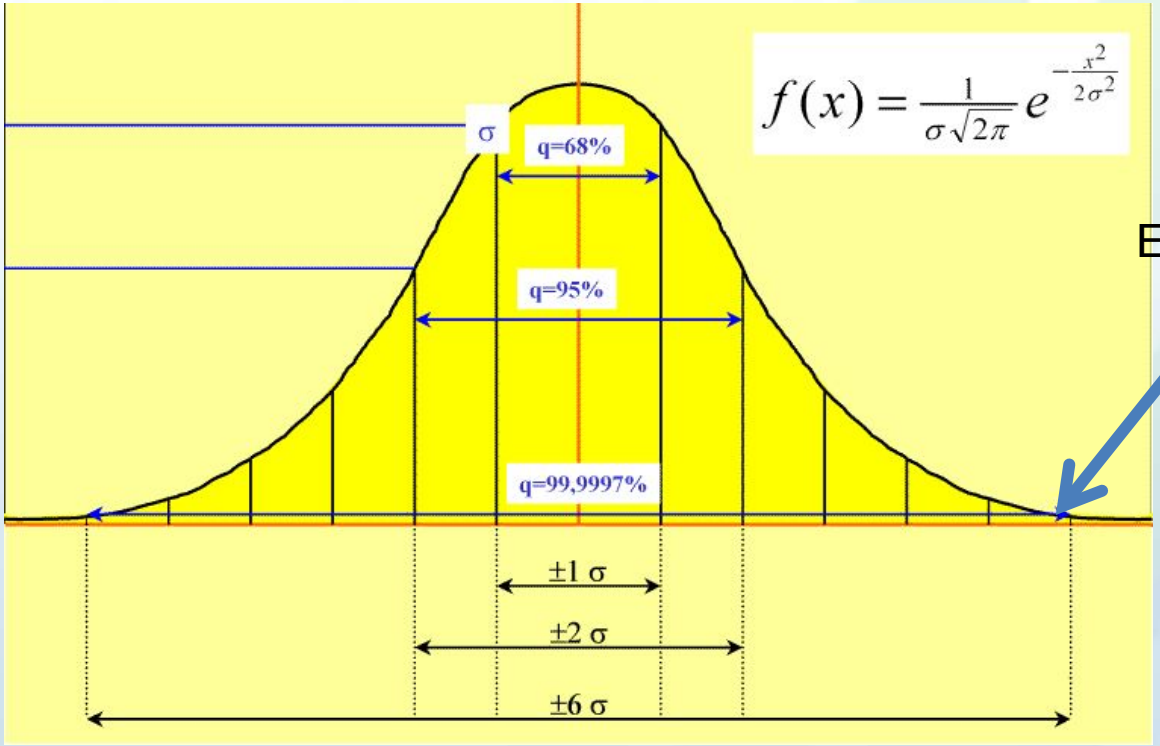
Six Sigma and Quality Research Software (GE Research)





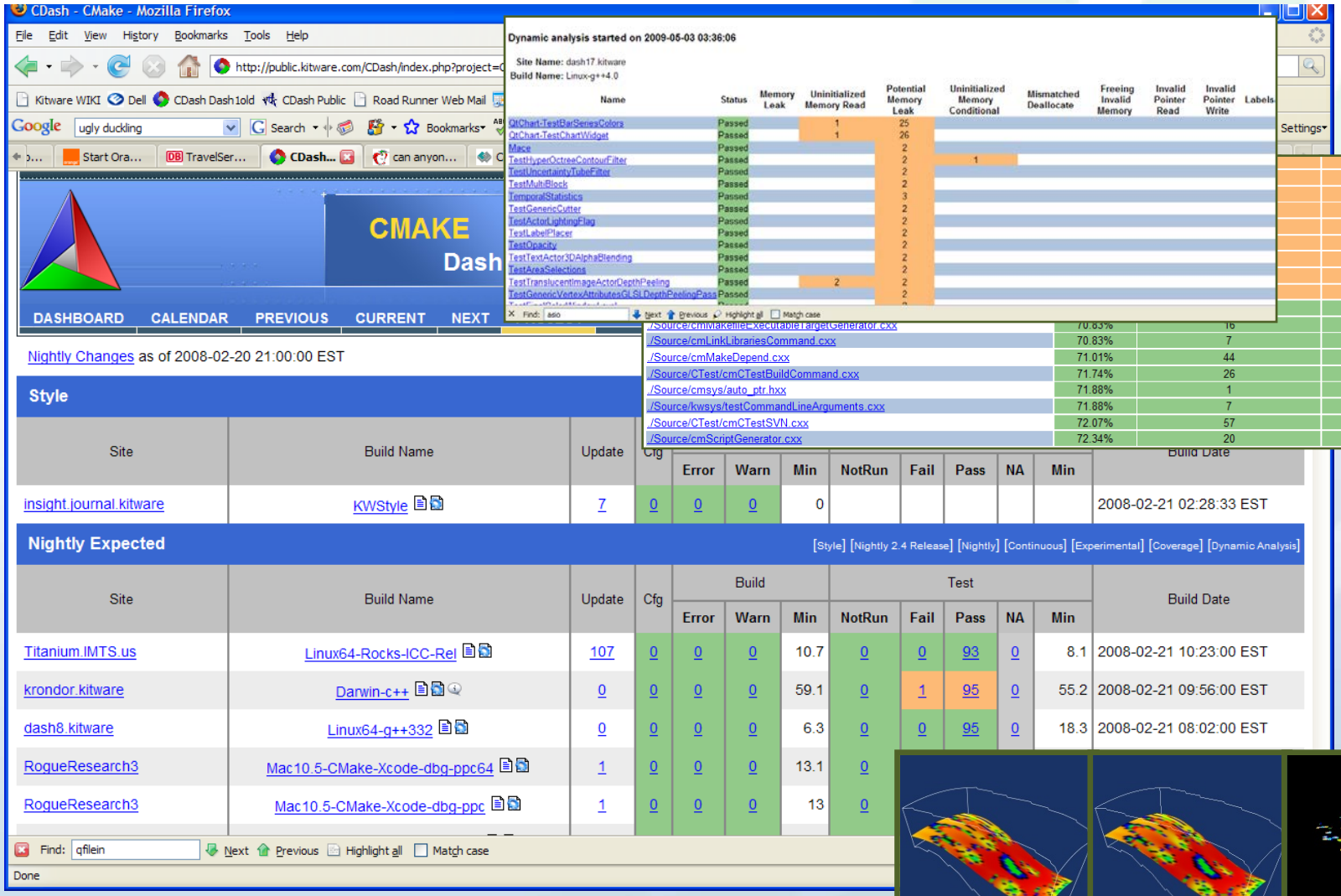
Kitware

Six Sigma and Quality Research Software



Kitware

CDash Dashboard www.cdash.org



Kitware

Software Process – Reproducible Results

Kitware

ExternalData Module - Source

- Tests reference data as if in source tree

```
$ cat CMakeLists.txt
itk_add_test(NAME MyTest COMMAND ... DATA{Baseline/MyTest.png} ...)
```

- File in source tree is a “content link”

```
$ cat Baseline/MyTest.png.md5
081dc468b8b4a18e624757f4a7d0ec2d
```

- Real data in arbitrary content-addressed storage

Road blocks

- The world's colleges now collectively spend at least $10 billion and probably more than $20 billion every year on subscriptions to academic journals and archives like JSTOR.
- Reproducibility is not part of the culture
- No feedback loop, if a student finds a method in a paper failing to work, there is no way to go back to the author
- No money for software infrastructure

From Wikipedia, the free e

For the actor, see Aa

Aaron H. Swartz (Noven
computer programmer, wi

Swartz was involved in the
framework web.py,[3] and
partner after a merger wit
sociology, civic awarenes
Harvard University's Safra
founded the online group
Online Piracy Act, and lat
He also was a contributing

On January 6, 2011, Swa
systematic downloading o
opposed JSTOR's practic
fees it charges for access
limiting public access to a
[10][11]

On January 11, 2013, Sw
apartment where he had h

'Aaron's Law' Proposes Reining in Federal Anti-Hacking Statute

BY KIM ZETTER 02.01.13 5:51 PM

Follow @KimZetter

Aaron Swartz *Photo: Fred Benson / Flickr*

on

FOSS and Science have always had a close relationship

- To this day, the U.S. Army remains one of Red Hat's largest customers by volume
- Open Source

from scientific groups

For the good of all of us: CERN launches open source hardware effort

CERN, the organization behind the Large Hadron Collider experiments, has …

by **Ryan Paul** - July 8 2011, 11:22am EDT

Open source software is used extensively by CERN, the particle physics lab behind the Large Hadron Collider (LHC) experiments. In fact, the organization even maintains its very own Linux distribution —based on Red Hat Enterprise Linux—called Scientific Linux CERN. Inspired by the productivity of Linux development, a group of CERN engineers have decided to bring the advantages of the open source software development model to the world of hardware.

CERN has launched a new community-centric effort called the Open Hardware Repository (OHR) with the aim of encouraging collaborative electronics design. CERN has also developed a new license, called the Open Hardware License (OHL), to govern the distribution of open hardware designs.

LINPACK benchmarks

From Wikipedia, the free encyclopedia

For other uses, see LINPACK (disambiguation).

The **LINPACK Benchmarks** are a measure of a system's floating point computing power. Introduced by Jack Dongarra, they measure how fast a computer solves a dense *n* by *n* system of linear equations $Ax = b$, which is a common task in engineering.

The latest version of these benchmarks is used to build the Top500 list, ranking the world's most powerful supercomputers.[1]

Open Science, Open Software, Reproducible Code

a marriage of FOSS and Science

- Open Data, Open Documentation, Open Code = Reproducibility = Scientific Method

Kitware


Science
Born of truth, service to others
Built on intellectual pursuit
Ruthless in its reach