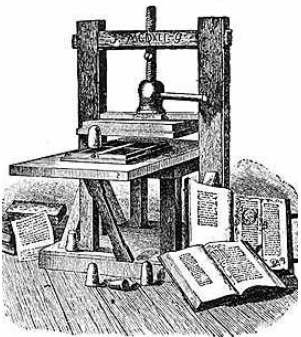


What Bioinformaticians Need to Know About Digital Publishing Beyond the PDF

Philip E. Bourne PhD
pbourne@ucsd.edu



Where My Biased Perspective Comes From..

- Computational biologist – interests in systems pharmacology, evolution, protein structure
- Developer of the RCSB PDB
- Founding Editor in Chief of PLOS Computational Biology
- Got interested in scholarly communication



Scholarly Communication is Being Disrupted – Witness The Story of Meredith

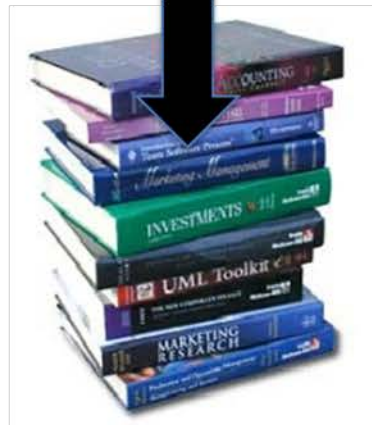
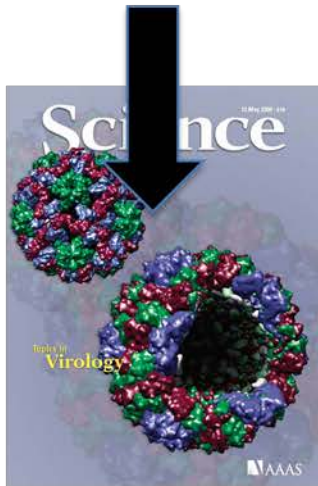
The screenshot shows the FORA.tv website interface. At the top, there is a navigation bar with 'FORA.tv CONFERENCE AND EVENT VIDEO' on the left, and 'Join Now' or 'Log In' buttons next to a search bar labeled 'Search Videos'. To the right of the search bar are social media icons for Twitter and Facebook. Below the navigation bar is a horizontal menu with categories: PAY-PER-VIEW, BUSINESS, ENVIRONMENT, POLITICS, SCIENCE, TECHNOLOGY, and CULTURE. Underneath this is a secondary menu with sub-categories: SPACE, EVOLUTION, PHYSICS, SOCIAL SCIENCES, NATURAL SCIENCES, DNA, PSYCHOLOGY, BIOTECH, MEDICINE, ANTHROPOLOGY, and ASTRONOMY. A prominent banner features a 'WATCH LIVE' button with a timer showing '117 hrs 43 mins 13 secs', a video thumbnail for 'THE U.S. HAS NO DOG IN THE FIGHT IN SYRIA', and the text 'presented by intelligence² DEBATES'. Below the banner, the video title 'Congress Unplugged - Phil Bourne' is displayed, along with a 'WATCH FULL VIDEO' link and information about the conference 'Sage Bionetworks Commons Congress 2012' and partner 'Sage Bionetworks'. The main content area is split into two panels. The left panel contains promotional text for the '3rd Sage Bionetworks Commons Congress April 20-21, 2012' and the Sage Bionetworks logo. The right panel features the SAS logo with the tagline 'THE POWER TO KNOW.' and the word 'ANALYTICS' in large letters. Below this, it says 'Turn what they say into why they stay.' and includes a button that reads 'CLICK FOR PAPER ON BUILDING A MARKETING ANALYTICS FRAMEWORK'.

http://fora.tv/2012/04/20/Congress_Unplugged_Phil_Bourne

The Era of Open Has The *Potential* to Deinstitutionalize (1)



Daniel Hulshizer/Associated Press

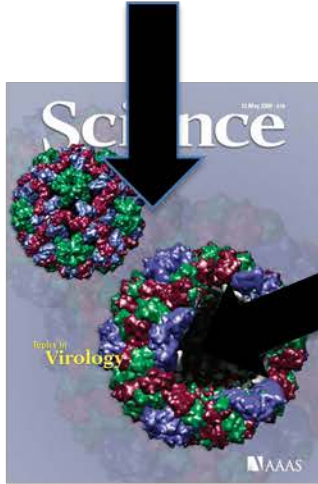


WIKIPEDIA
The Free Encyclopedia

The Era of Open Has The *Potential* to Deinstitutionalize (2)



Daniel Hulshizer/Associated Press



Most Academic Institutions Have Yet
to Realize This

Funding Agencies Could Provide the
Wake Up Call

Publishing is Also Being Deinstitutionalized

- Today:
 - Approx 10,000 publishers
 - Publishing approx 25,000 journals
 - Which publish approx 1.5 million articles per year (almost 1 million of which appear in PubMed)

Witness the 'Open Access Mega Journal'

1. Very very large

- Publishing thousands of articles per year
- and benefiting from economies of scale

2. Open Access

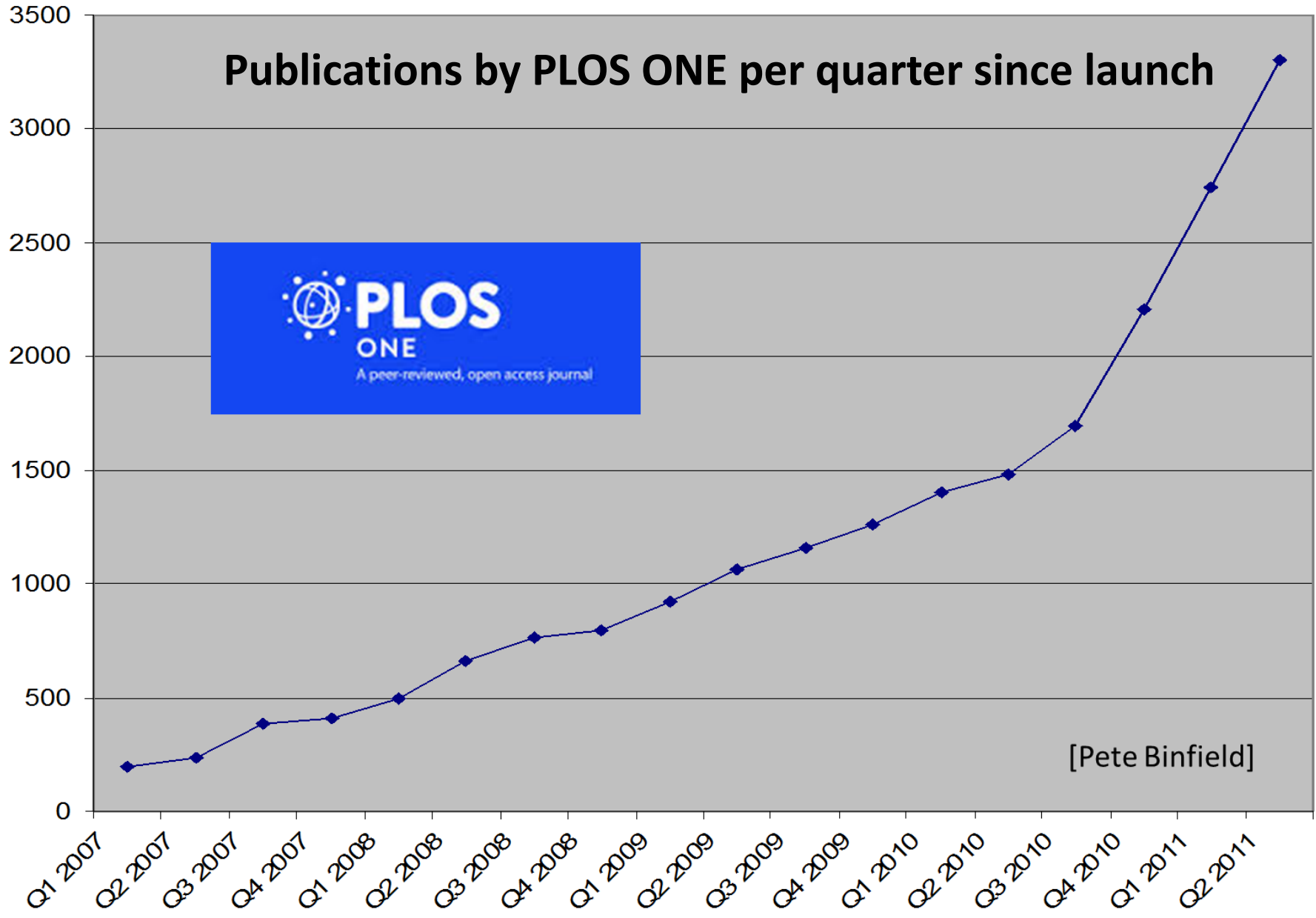
- Because no one will pay a subscription fee for a journal that large (and growing that fast)
- and using an OA Business Model where each article pays for its own costs

3. (Preferably) without any 'artificial' constraints on its ability to grow

- For example, a desire to only publish 'high impact; papers

[Pete Binfield]

Publications by PLOS ONE per quarter since launch



[Pete Binfield]

“Open Access Mega Journals”

– One Name, Two Flavours

- ‘Clones’ of PLoS ONE (not selective)
 - SAGE Open
 - BMJ Open
 - Scientific Reports (Nature)
 - AIP Advances (Am Inst Physics)
 - G3 (Genetics Soc of America)
 - Biology Open (Company of Biologists)
- ‘Pseudo-Clones’ of PLoS ONE (probably selective)
 - Physical Review X (Am Physical Society)
 - Open Biology (Royal Society)
 - Cell Reports (Elsevier, Cell Press)

[Pete Binfield]

This Still Places the Research Article as
the Central Focus of the Academic
Enterprise...

Maybe the Article is Only One View

Paper as Portal

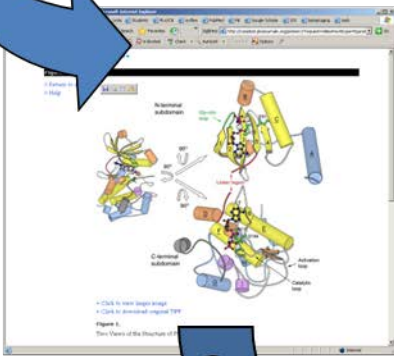
0. Full text of PLoS papers stored in a database



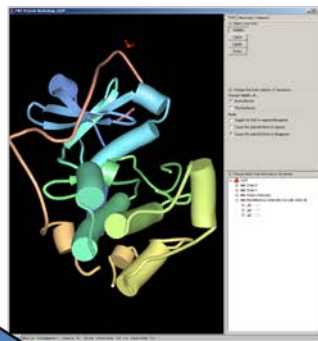
4. The composite view has links to pertinent blocks of literature text and back to the PDB



1. A link brings up figures from the paper



3. A composite view of journal and database content results



2. Clicking the paper figure retrieves data from the PDB which is analyzed

1. User clicks on thumbnail
2. Metadata and a webservice call provide a renderable image that can be annotated
3. Selecting a features provides a database/literature mashup
4. That leads to new papers

PLoS Comp. Biol. 2005 1(3) e34



Given This Disruption It is Worth Thinking About... (1)



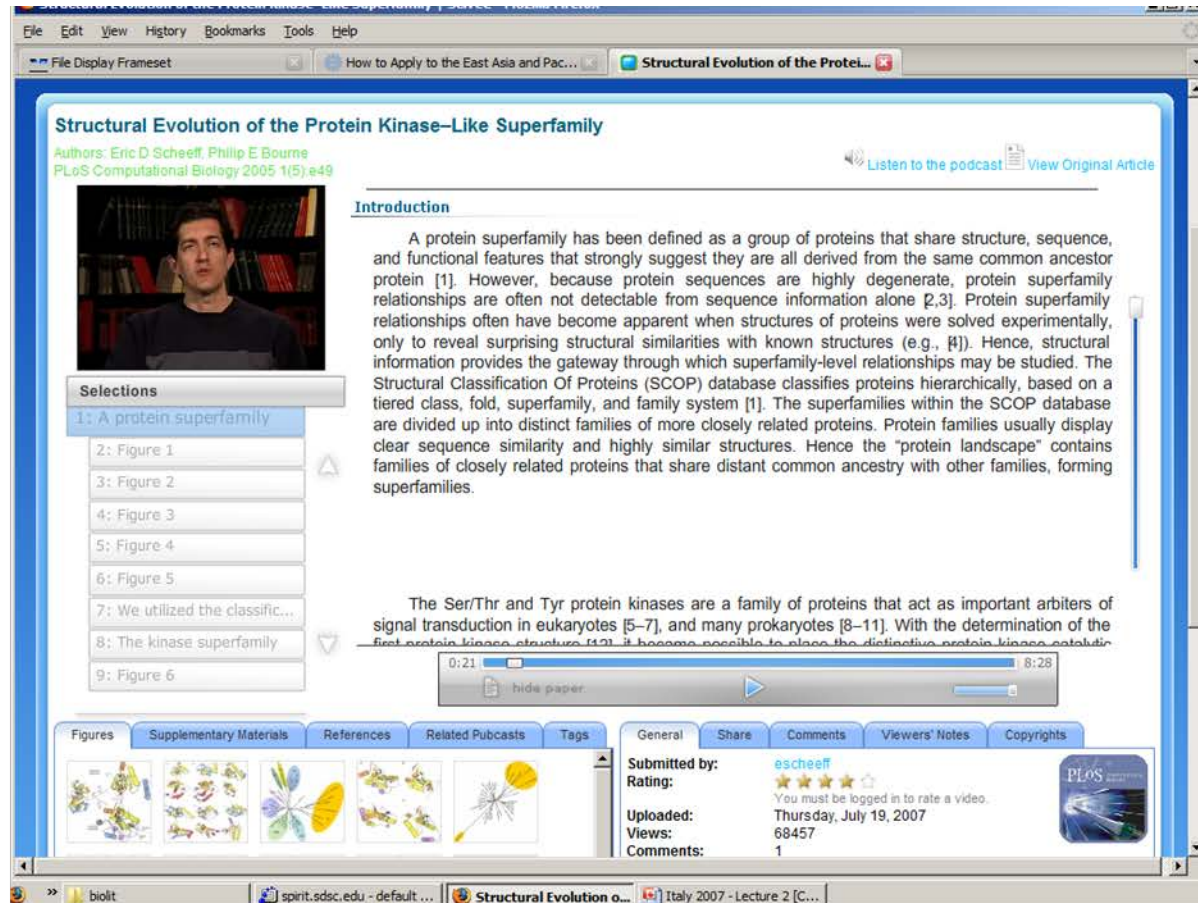
- A paper as only *one* form of knowledge discovery
- The use of interaction and rich media from which to learn and actually *do* science
- Reproducibility
- Reward structures
- Better management of the *research lifecycle*

P.E. Bourne 2005 In the Future will a Biological Database Really be Different from a Biological Journal? *PLoS Comp. Biol.* 1(3) e34

MOOCs As Research



Pubcast – Video Integrated with the Full Text of the Paper



Structural Evolution of the Protein Kinase-Like Superfamily
Authors: Eric D Scheeff, Philip E Bourne
PLoS Computational Biology 2005 1(5): e49

[Listen to the podcast](#) [View Original Article](#)

Introduction

A protein superfamily has been defined as a group of proteins that share structure, sequence, and functional features that strongly suggest they are all derived from the same common ancestor protein [1]. However, because protein sequences are highly degenerate, protein superfamily relationships are often not detectable from sequence information alone [2,3]. Protein superfamily relationships often have become apparent when structures of proteins were solved experimentally, only to reveal surprising structural similarities with known structures (e.g., [4]). Hence, structural information provides the gateway through which superfamily-level relationships may be studied. The Structural Classification Of Proteins (SCOP) database classifies proteins hierarchically, based on a tiered class, fold, superfamily, and family system [1]. The superfamilies within the SCOP database are divided up into distinct families of more closely related proteins. Protein families usually display clear sequence similarity and highly similar structures. Hence the "protein landscape" contains families of closely related proteins that share distant common ancestry with other families, forming superfamilies.

The Ser/Thr and Tyr protein kinases are a family of proteins that act as important arbiters of signal transduction in eukaryotes [5–7], and many prokaryotes [8–11]. With the determination of the first protein kinase structure [12], it became possible to place the distinctive protein kinase catalytic

0:21 8:28
hide paper

Submitted by: [escheeff](#)
Rating: ★★★★★
You must be logged in to rate a video.
Uploaded: Thursday, July 19, 2007
Views: 68457
Comments: 1

Figures Supplementary Materials References Related Pubcasts Tags



Making science visible

<http://www.scivee.tv>

<https://www.coursera.org/course/drugdiscovery> Spontaneous Groups Formed from All Over the World

coursera | Global Partners Courses Partners About ▾ | Sign In Sign Up

UC San Diego

Drug Discovery, Development & Commercialization


Williams S. Ettouati, Pharm.D. and Joseph D. Ma

Students will learn the process of drug discovery and development through specific examples of case studies to better understand the issues facing the challenges of delivering a new drug on the market. At the completion of this course you will be able to have a better understanding of how a small or large molecule becomes a pharmaceutical drug.

Workload: 3-4 hours/week

Taught In: English

Subtitles Available In: English



Sessions:

Apr 19th 2013 (9 weeks long) [Enroll](#)

Future sessions [Add to Watchlist](#)

67 [Tweet](#) 49 [+1](#) 516 [Like](#)



Given This Disruption It is Worth Thinking About... (2)



- A paper as only *one* form of knowledge discovery
- The use of interaction and rich media from which to learn and actually *do* science
- **Reproducibility**
- Reward structures
- Better management of the *research lifecycle*

P.E. Bourne 2005 In the Future will a Biological Database Really be Different from a Biological Journal? *PLoS Comp. Biol.* 1(3) e34

Attitudes are Changing

datasets
data collections
algorithms
configurations
tools and apps
codes
workflows
scripts
code libraries
services,
system software
infrastructure,
compilers
hardware

Carole Goble]

“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**, [the complete data] and the complete set of instructions which generated the figures.”

David Donoho, “Wavelab and Reproducible Research,” 1995

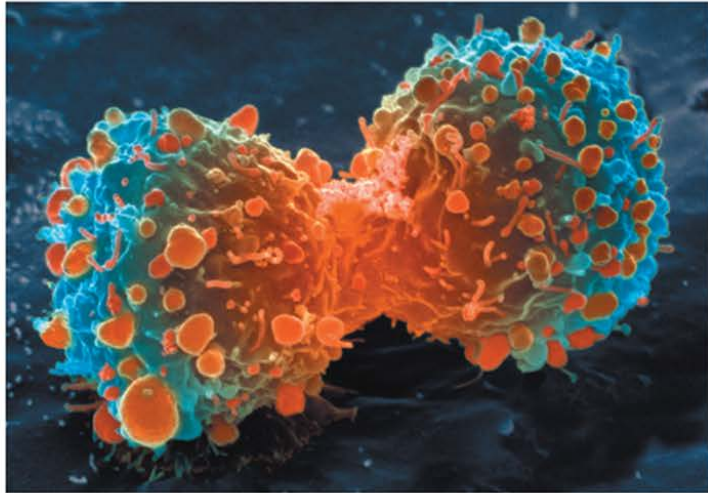
COMMENT

AVIAN INFLUENZA Shift expertise to track mutations where they emerge **p.524**

EARTH SYSTEMS Past climates give valuable clues to future warming **p.527**

HISTORY OF SCIENCE Descartes' lost letter tracked using Google **p.540**

OBITUARY Wylie Vale and an elusive stress hormone **p.542**



Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability

to identify and test drugs in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will

be approved. Investigators must reassess their approach to translating discovery research into greater clinical success and impact.

Many factors are responsible for the high failure rate, notwithstanding the inherently difficult nature of this disease. Certainly, the limitations of preclinical testing

47/53 “landmark” publications could not be replicated
Begley, Ellis Nature, 483, 2012]

Must try harder

Too many sloppy mistakes are creeping into scientific papers. Lab heads must look more rigorously at the data — and at themselves.

Error prone

Biologists must realize the pitfalls of work on massive amounts of data.

If a job is worth doing, it is worth doing twice

Researchers and funding agencies need to put a premium on ensuring that results are reproducible, argues Jonathan F. Russell.

The case for open computer programs

Six red flags for suspect work

C. Glenn Begley explains how to recognize the preclinical papers in which the data won't stand up.

Know when your numbers are significant
[Carole Goble]

computational analyses almost irreproducible. [Supplementary information S2 \(reference list\)](#) lists 50 papers randomly selected from 378 manuscripts published in 2011 that use the Burrows-Wheeler Aligner¹⁵ for mapping Illumina reads. Most papers (31) provide neither a version nor the parameters used, and neither do they provide the exact version of the genomic reference sequence. From the remaining 19 publications, only four studies provide settings, eight studies list the version, and only seven studies list all necessary details. More than half of the studies (26 out of 50) do not provide access to the primary data sets. In two cases, authors provided links to their own websites, where data were deposited; however, in both cases, links were broken.

Nekrutenko & Taylor, Next-generation sequencing data interpretation: enhancing, reproducibility and accessibility, Nature Genetics 13 (2012)

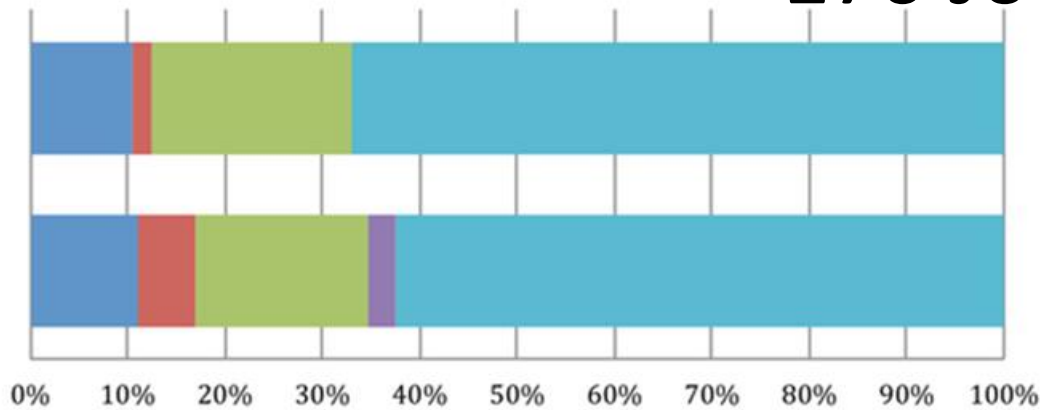
59% of papers in the 50 highest-IF journals comply with (often weak) data sharing rules.

Alsheikh-Ali et al Public Availability of Published Research Data in High-Impact Journals. PLoS ONE 6(9) 2011

[Carole Goble]

170 Journals, 2011 - 2012

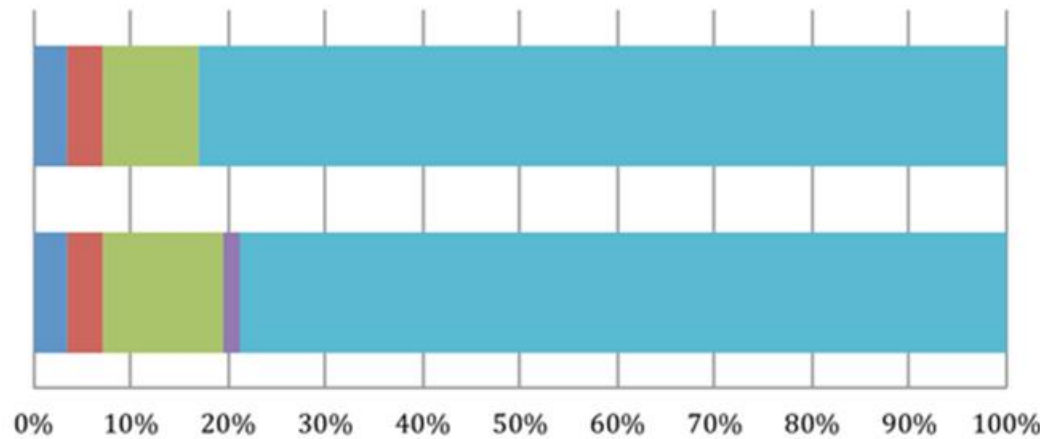
Data Sharing Policy



- 1 Required as condition of publication
- 2 Required but may not affect decisions
- 3 Explicitly encouraged
- 4 Implied
- 5 No mention

[Carole Goble]

Code Sharing Policy



- 1 Required as condition of publication
- 2 Required but may not affect decisions
- 3 Explicitly encouraged
- 4 Implied
- 5 No mention

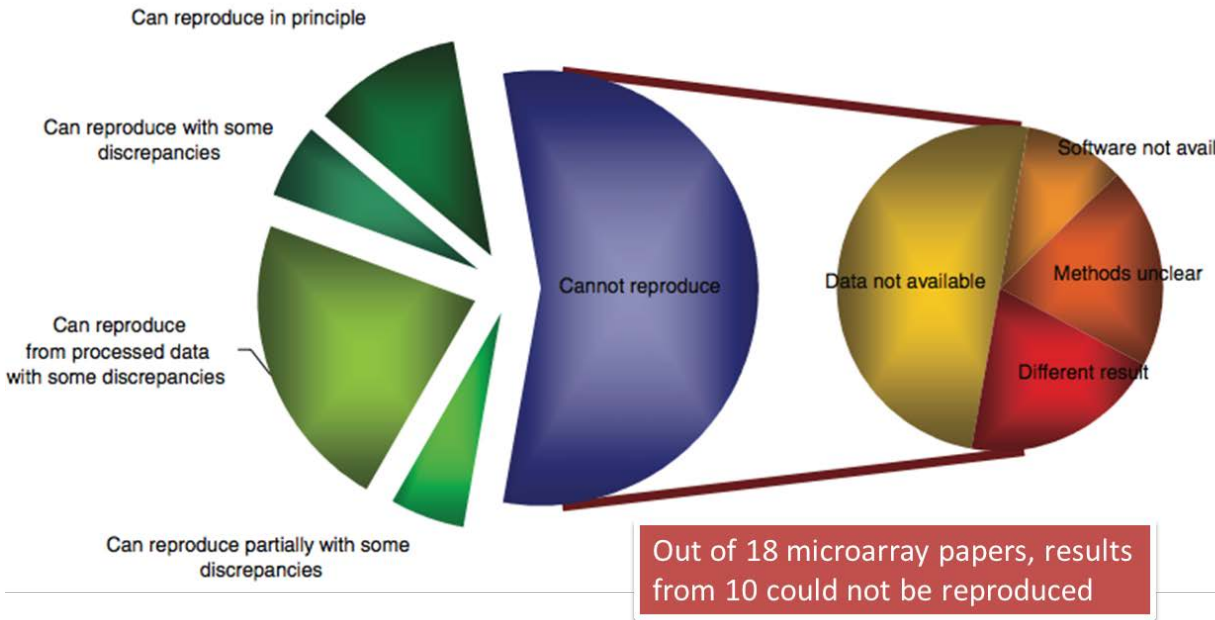
Stodden V, Guo P, Ma Z (2013) Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals. PLoS ONE 8(6): e67111. doi:10.1371/journal.pone.0067111

Flaws Are Becoming More Obvious

More retractions:

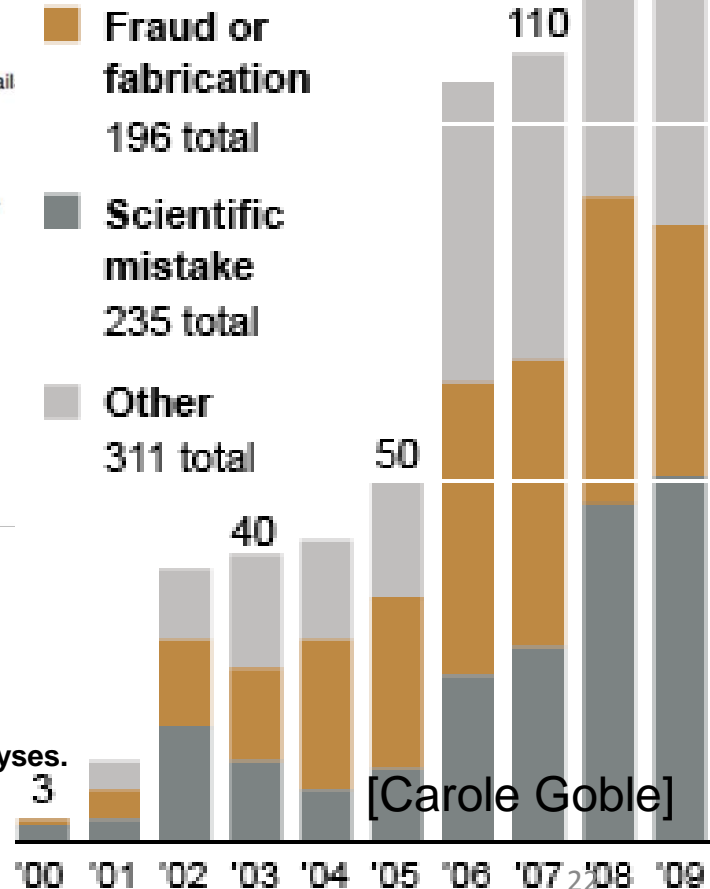
>15X increase in last decade

At current % > by 2045 as many papers published as retracted



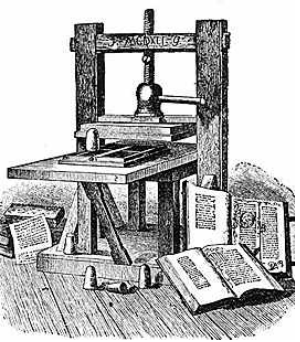
Retractions On the Rise

A study of the PubMed database found that the number of articles retracted from scientific journals increased substantially between 2000 and 2009.



The New York Times

1. Ioannidis et al., 2009. Repeatability of published microarray gene expression analyses. *Nature Genetics* 41: 14
2. Science publishing: The trouble with retractions
<http://www.nature.com/news/2011/111005/full/478026a.html>
3. Bjorn Brembs: Open Access and the looming crisis in science <https://theconversation.com/open-access-and-the-looming-crisis-in-science-14950>



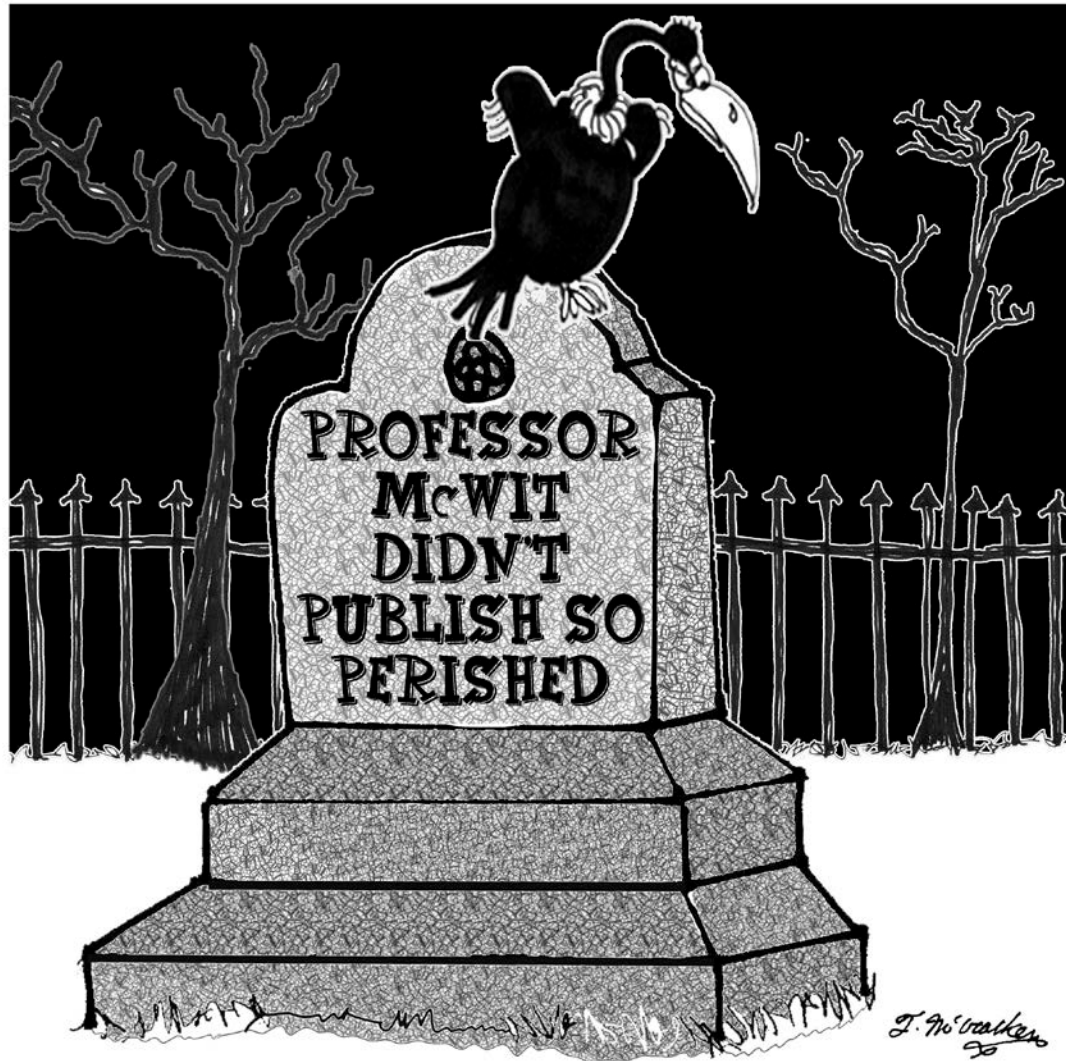
Given This Disruption It is Worth Thinking About... (3)



- A paper as only *one* form of knowledge discovery
- The use of interaction and rich media from which to learn and actually *do* science
- Reproducibility
- **Reward structures**
- Better management of the *research lifecycle*

P.E. Bourne 2005 In the Future will a Biological Database Really be Different from a Biological Journal? *PLoS Comp. Biol.* 1(3) e34

Unfortunately the Metrics of Success Remain...



[Carole Goble]

This makes no sense when you ask yourself the question:
What is more valuable a dataset used and cited by 100 scientists or a paper you wrote that only you cite?

Case in point...

Title / Author	Cited by	Year
<input type="checkbox"/> The protein data bank HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, IN ... Nucleic acids research 28 (1), 235-242	16287	2000

What can you do today to change the situation?



**WHAT YOU DO
TODAY
CAN IMPROVE
ALL YOUR
TOMORROWS.**

What Can You Do?

- Support emergent community commons/portals
- Be involved in the support and development of metadata standards
- Contribute to workflow development etc. to drive an open research lifecycle
- Educate your mentors on the importance of open science and scholarly communication
- Write software thinking of an App model

Portals

INSTITUTE FOR TRANSLATIONAL SYSTEMS BIOLOGY

3D VIRTUAL CELL

HOME ABOUT CATALOG NEWSROOM MULTIMEDIA EVENTS CONTACT

WHAT IS 3DVC?

Advancing the field of cell biology through open, community-based science and technology innovation, shared infrastructure, and new, sustainable business models.

The 3D Virtual Cell (3DVC) is a collaborative project funded by the NSF to create a software platform with the ultimate goal of gathering a community of scientists together to create the first 3D Virtual Cell. Currently under development, the institute strives to address the cyber-infrastructure challenges related to data accessibility, software development, software reuse and sustainability. We are challenged with creating new algorithm development and analysis tools that can operate on an increasingly large, diverse, complex and widely distributed body of digital biological data. Broad objectives are to create the 3D virtual cell to speed biological discovery and lead to educational tools for K-12 and beyond.

BLOGS

- 10.29.13 **microRNAs** targeting seeds of destruction
- 10.28.13 **test post**
- 10.28.13 **Stem cells: the future, the future, die Zukunft, il futuro, masa przyszlosci, et futuro**
- 10.28.13 **Changing the currency of science to solve our greatest challenges**
- 10.28.13 **Science and storytelling: great conversations at Brave New Worlds event**

TWITTER

If you want to participate in better discussions, please use the hashtag #force11 and your tweet will show up here.

Force 11.org

3D Virtual Cell

3D VIRTUAL CELL

The 3DVC will do for cell biology what the Large Hadron Collider (LHC) does for particle physics, but through a virtual rather than physical resource. It will bring together collaborations around a shared infrastructure to advance the field through efficient, groundbreaking science and technology, the results of which will be broadly disseminated to an audience ranging from K-12 to professionals. The 3DVC is committed to open science, yet strives for sustainability through new business models that leverages that open science.

MULTIMEDIA

Philip Bourne, PI of the 3D Virtual Cell Project

[JOIN OUR COMMUNITY](#)

<http://www.3dvcell.org/>

FORCE11
The Future of Research Communications and e-Scholarship

CONNECT [social media icons]

search this site

Home About Community Members Groups Resources Publications News + Events Contact

Join now!

CREATE ACCOUNT or login below:

pbourne@uctsd.edu

LOG IN

Forgot your password?

We currently have 483 active members.

INVITE OTHERS

DISCUSSIONS

GET INVOLVED

TOOL + RESOURCES

CALENDAR

GROUPS

PROMOTE FORCE11

FORCE 11 HIGHLIGHTS

WHAT IS FORCE 11?

FORCE11 is a community working together in support of the goal of advancing scholarly communication. The website has been set up for this growing community to gather and disseminate information. We are a neutral open information community and do not endorse or seek to block any relevant work.

Tool and Resource Catalog LIVE

GREAT NEWS: The Force 11 Catalog is live!!! Finally a place on the web where our members can add in one place all tools and resources relating to e-scholarship.

Force 11 Forms Non-Profit Organization

Force11 announces the conclusion of its incorporation process, as a non-profit organization, and the selection of its first slate of Offices and Directors: Executive Director: Mariann Martone; President: Philip Bourne; Vice President: Daniel O'Donnell; Secretary: Anita de Waard; Treasurer: Tim Clark, Treasurer; and Directors: Paul Groth, Ivan Herman, Eduard Hovy, Cameron Neylon

NEW FEATURE...Start a Community Working Group or Project

Force 11 has added space on the website for working groups to collaborate on projects. If you wish to start a working group fill out an application here or contact admin@force11.org for setup assistance.

DATA CITATION WORKING GROUP - 25+ organizations working together

A Force 11 working group representing over 25 organizations working together to produce a consolidated set of data citation principles in order to encourage broad adoption of a consistent policy for data citation across disciplines and venues

FORCE11 has a Mendeley Group!

We've created a group in Mendeley for papers relevant to the future of research communication and e-scholarship. The group is public, so feel free to add your own papers.

<http://www.mendeley.com/groups/2475471/force11/>

COMMUNITY NEWS

- 10.27.13: What will future publications be like? + READ MORE
- 10.26.13: FORCE 11 SESSION AT MOZFEEST 2013 + READ MORE
- 10.25.13: Force 11 Tool and Resource Catalog + READ MORE

TWITTER

Ingeborg @Mrs_Ingborg Storm in the Netherlands! #force11 #Autumn

nobleundshawn @nobleundshawn Bit windy here #force11. It should be into the North Sea in 2 hours from now, 28 Oct 0620. Trees down, transport disrupted, power out.

Compose new Tweet...

FEATURED

What will future publications be like?

POST DATE: SUNDAY, OCTOBER 27, 2013 - 07:36

<http://www.force11.org/>

What Can You Do?

- Support emergent community commons/portals
- Be involved in the support and development of metadata standards
- Contribute to workflow development etc. to drive an open research lifecycle
- Educate your mentors on the importance of open science and scholarly communication
- Write software thinking of an App model

We Need Innovative Contributions to the Research Lifecycle (1)

Authoring
Tools

Lab
Notebooks

Data
Capture

Software
Repositories

Analysis
Tools

Visualization

Scholarly
Communication

IDEAS – HYPOTHESES – EXPERIMENTS – DATA - ANALYSIS - COMPREHENSION - DISSEMINATION

Discipline-
Based Metadata
Standards

Git-like
Resources
By Discipline

Community Portals

Data Journals

Commercial &
Public Tools

New Reward
Systems

Training

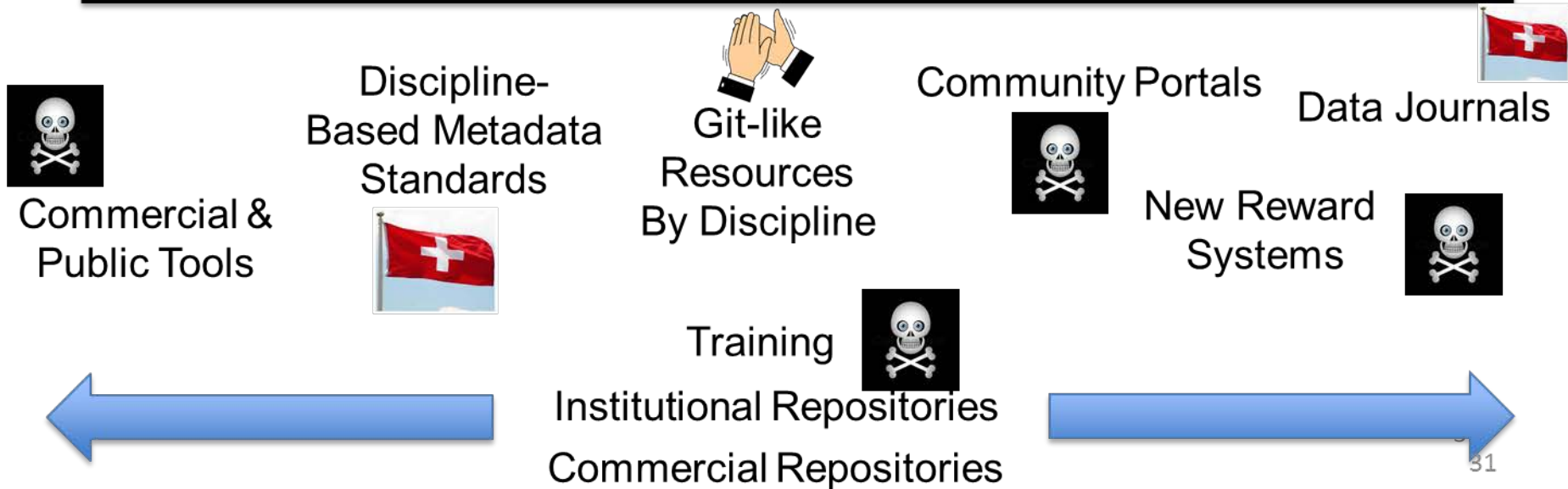
Institutional Repositories
Commercial Repositories



We Need Innovative Contributions to the Research Lifecycle (2)



IDEAS – HYPOTHESES – EXPERIMENTS – DATA - ANALYSIS - COMPREHENSION - DISSEMINATION



What Can You Do? (1)

- Support emergent community commons/portals
- Be involved in the support and development of metadata standards
- Contribute to workflow development etc. to drive an open research lifecycle
- Educate your mentors on the importance of open science and scholarly communication
- Write software thinking of an App model

Pressure Your Institutions to Play a Greater Role

- We need institutional data/knowledge sharing plans
- We need *digital universities*
- We need data/information scientists to be better recognized by institutions – its not all about papers – this implies new metrics

A View from the Digital University

Jane scores well in parts of her advanced on-line biology class. Professors who undertake research in the areas where Jane did well are automatically notified of her potential based on a computer analysis of her scores and background interests and Professor Smith interviews her and offers her a research internship for the summer. Over the summer, as she enters details of her experiments related to understanding a widespread neurodegenerative disease in an on-line laboratory notebook, the underlying computer system automatically puts Jane into contact with another student, Jack, in a different department whose notebook reveals he is working on using bacteria for purposes of toxic waste cleanup. Why the connection? It turns out the same gene, which they both reference a number of times in their notes, is linked to two very different disciplines – mental health and the environment. In the analog university they would never have discovered each other, but at the Digital University pooled knowledge can lead to a distinct advantage. The collaboration later results in a patent filing and triggers a notification to a number of biotech companies who might be interested in licensing the technology. A company licenses the technology and hires Jane and Jack to continue working on the project. Professor Smith hires another student using the revenue from the license and this in turn leads to a large federal grant. The students get good jobs, further research is supported and societal benefit arises from the technology. A hypothetical example for why the Digital University makes sense.

Committee on Academic Promotions

- What Counts
 - Money
 - Grants
 - Papers
 - Teaching
 - Service
- What Does Not
 - Sharing data
 - Sharing software
 - Open access
 - Collaboration
 - Patents
 - Startups

Ten Simple Rules for Getting Ahead as a Computational Biologist in Academia
2011 *PLoS Comp Biol* 7(1) e1002001

What Can You Do? (2)

- Support emergent community commons/portals
- Be involved in the support and development of metadata standards
- Contribute to workflow development etc. to drive an open research lifecycle
- Educate your mentors on the importance of open science and scholarly communication
- **Write software thinking of an App model**



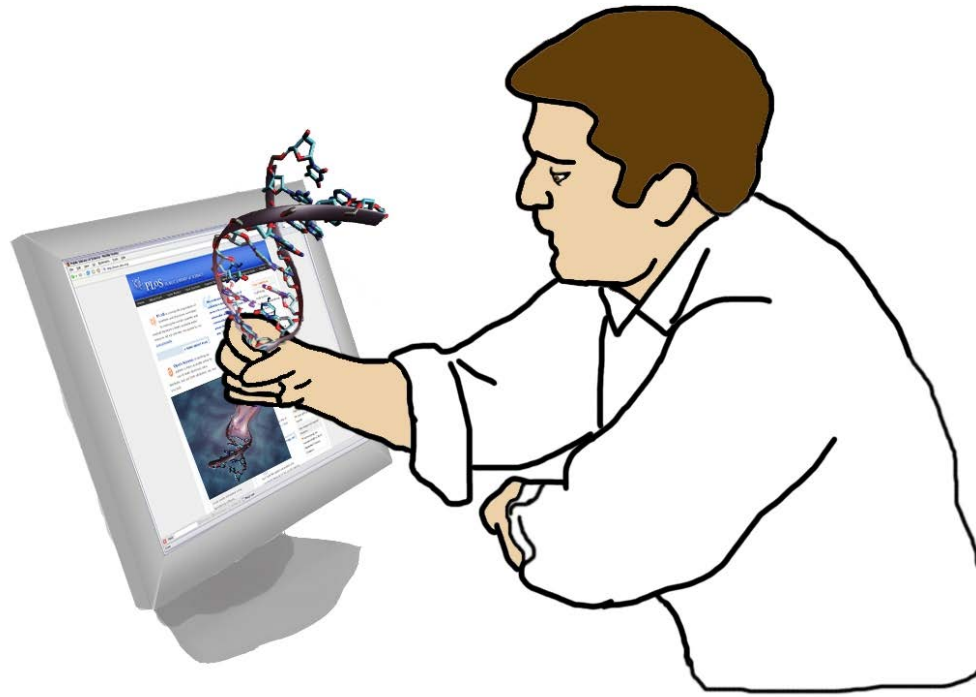
What Do We Need to Do to Get There? An App+ Store?

- The App model
 - Think of it operating on a content base rather than a mobile device
 - Simple and consistent user interface
 - Needs to pass some quality control
 - Has a reward
- The App+ Model
 - Apps interoperate through a generic workflow interface

Summary

- Disruption is occurring
- As bioinformaticians we have the skill set to leverage change and make a difference
- Go for it

pbourne@ucsd.edu



Questions?