

Cold Spring Harbor Laboratory

The View from Nowhere in Computational Infrastructure

Jason Williams

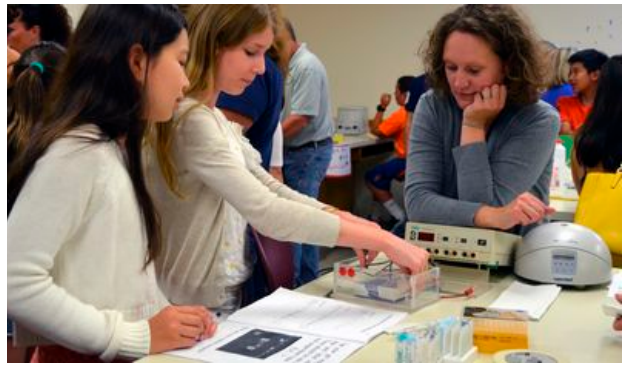
Cold Spring Harbor Laboratory, DNA Learning Center

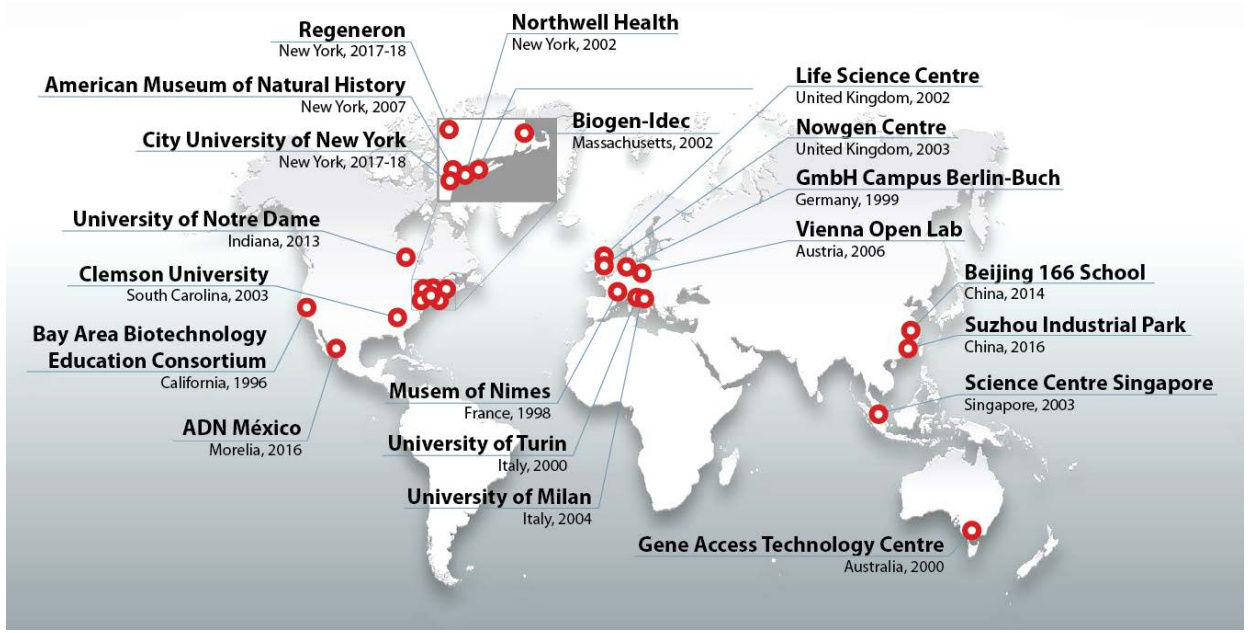
 @JasonWilliamsNY

ELIXIR UK All-hands, University of Dundee, December 2019

CSHL DNA Learning Center





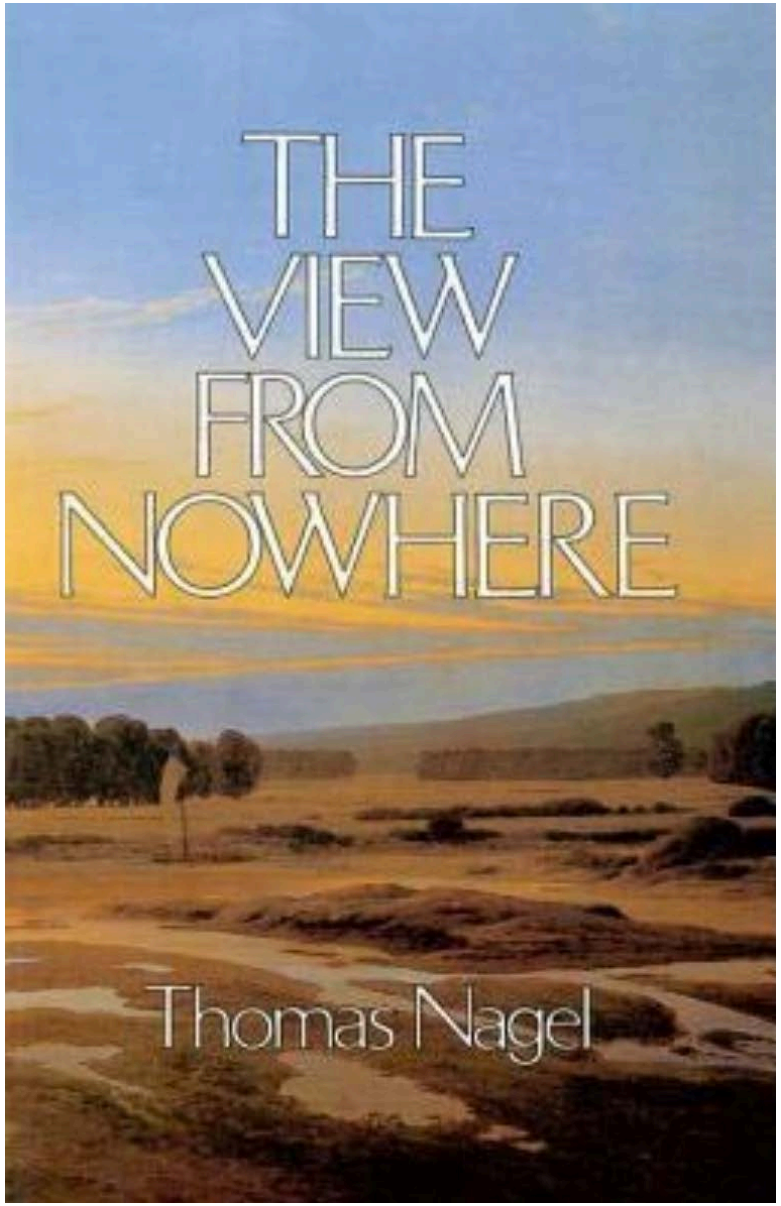


- Hands on education in molecular biology/bioinformatics for secondary students (Grade 6-12); secondary and undergraduate faculty training; websites/multimedia
- 3 Dedicated centers (2 more by 2020) and 16 centers licensed/modeled on DNALC
- Approximately 30K Visitors annually (more than 2 million visitors to-date for all centers)

Warning: observations and conjectures ahead






THE
VIEW
FROM
NOWHERE
Thomas Nagel

Journalism Ethics

The goal of objective and unbiased reporting ("just the facts"), often leaves subjective decisions about the evaluation of competing facts in a news report up to the audience. A journalist who does not help the audience evaluate which set of competing facts are most reasonable has taken the *view from nowhere*.

https://en.wikipedia.org/wiki/View_from_nowhere | Jay Rosen

Infrastructures have responsibility

Infrastructures have responsibility

- Guide users to what is useful
- Protect users from what is harmful/distracting

Infrastructures have responsibility

- Guide users to what is useful
- Protect users from what is harmful/distracting (noise, extraneous information)

Infrastructure (compute, data) that emphasizes availability over usefulness has taken the view from nowhere

What is the Vision of ELIXIR?

The goal of ELIXIR is to coordinate these resources so that they form a single infrastructure.

What we think we value often differs from what we actually value

“Training and consulting get short shrift”

ECAR
EDUCAUSE CENTER FOR APPLIED RESEARCH

Table 7-3. Investment Preferences

Technology Element	Percentage Responding
Expanding infrastructure (networks, computation, and data storage)	46.0%
Providing research-related IT support (e.g., training and consulting)	20.8%
Offering applications and tools (e.g., visualization, display, statistical analysis)	17.9%
Enhancing operations (e.g., security procedures, licensing support, facility management)	15.3%

IT Engagement in Research: A Baseline Study
EDUCAUSE Center For Applied Research, 2006 –
Study of 328 Universities/Institutions in US and Canada

“Training and consulting get short shrift”

ECAR
EDUCAUSE Center for Applied Research

Table 7-3. Investment Preferences

Technology Element	Percentage Responding
Expanding infrastructure (networks, computation, and data storage)	46.0%
Providing research-related IT support (e.g., training and consulting)	20.8%
Offering applications and tools (e.g., visualization, display, statistical analysis)	17.9%
Enhancing operations (e.g., security procedures, licensing support, facility management)	15.3%

“Even if there were fewer real-world constraints, infrastructure would still rule. The more surprising number is the second-place showing of training and consulting, consistently at the bottom of the pile in actual funding for both the past and the next three years … training and consulting get the short shrift when bumped against the realities of running an IT operation

IT Engagement in Research: A Baseline Study
EDUCAUSE Center For Applied Research, 2006 –
Study of 328 Universities/Institutions in US and Canada

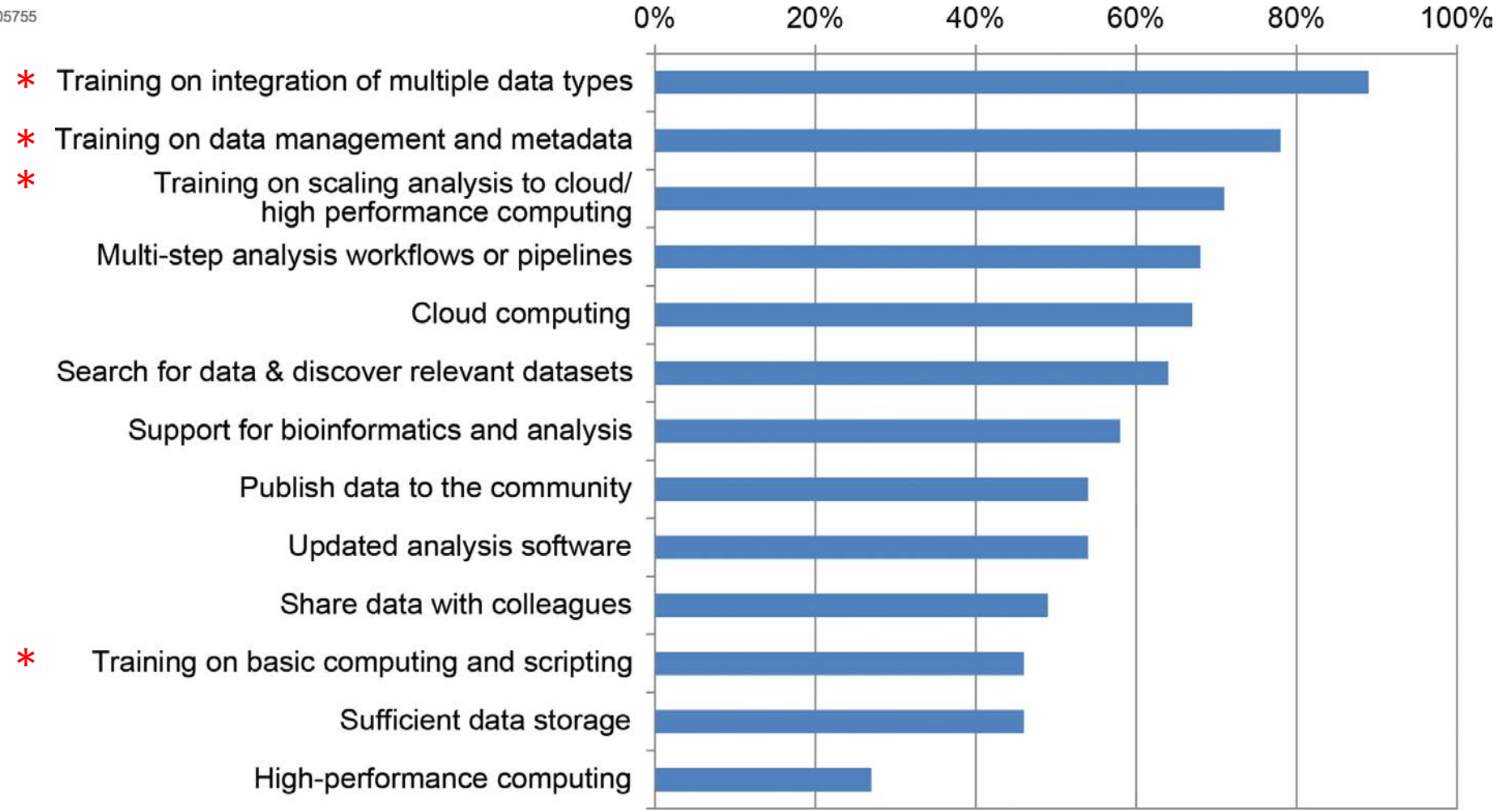
PLOS | COMPUTATIONAL BIOLOGY

Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators

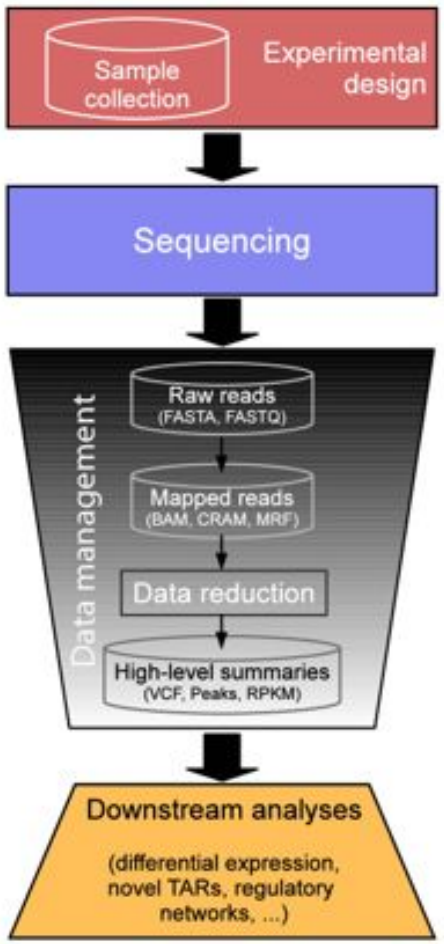
Lindsay Barone, Jason Williams, David Micklos

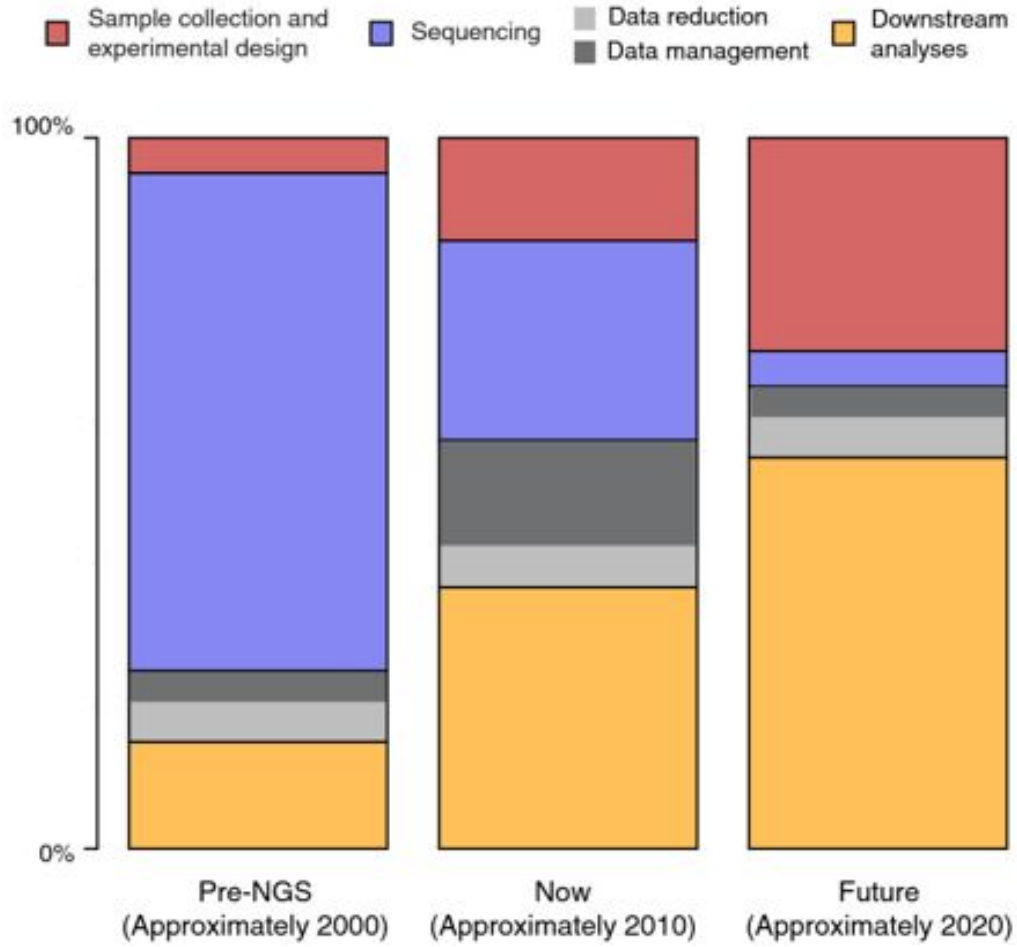
Published: October 19, 2017 • https://doi.org/10.1371/journal.pcbi.1005755

"Does your institution meet this need?"
('no' responses)



Information (and cost) flows



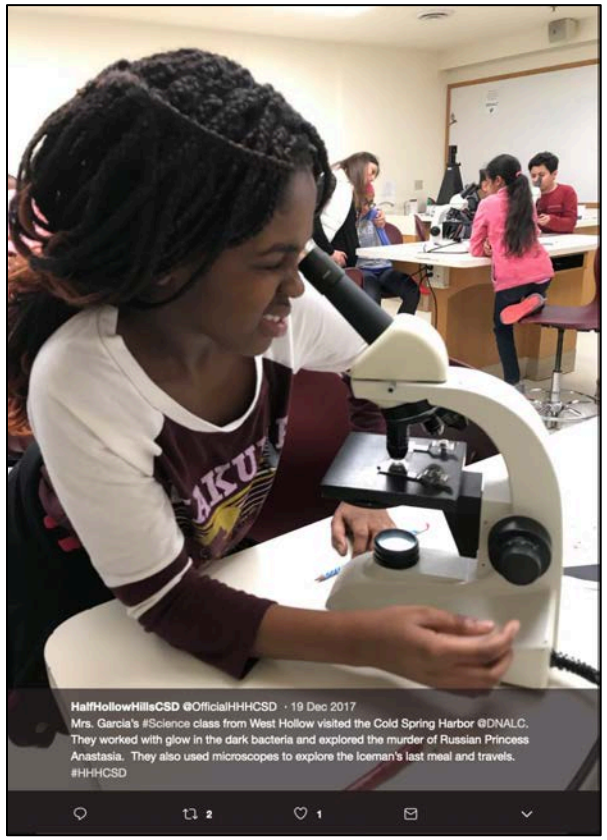


The real cost of sequencing: higher than you think!
Sboner, A., Mu, X.J., Greenbaum, D. *et al.*
Genome Biol **12,** 125 (2011) doi:10.1186/gb-2011-12-8-125

Are we delivering infrastructure is that is useful or merely usable?

Biologist's Culture Shock

I went to school for…



I need to know this for my research…

```
import urllib2
eutils = 'http://www.ncbi.nlm.nih.gov/entrez/eutils/'
efetch = 'efetch.fcgi?'
s = eutils + efetch

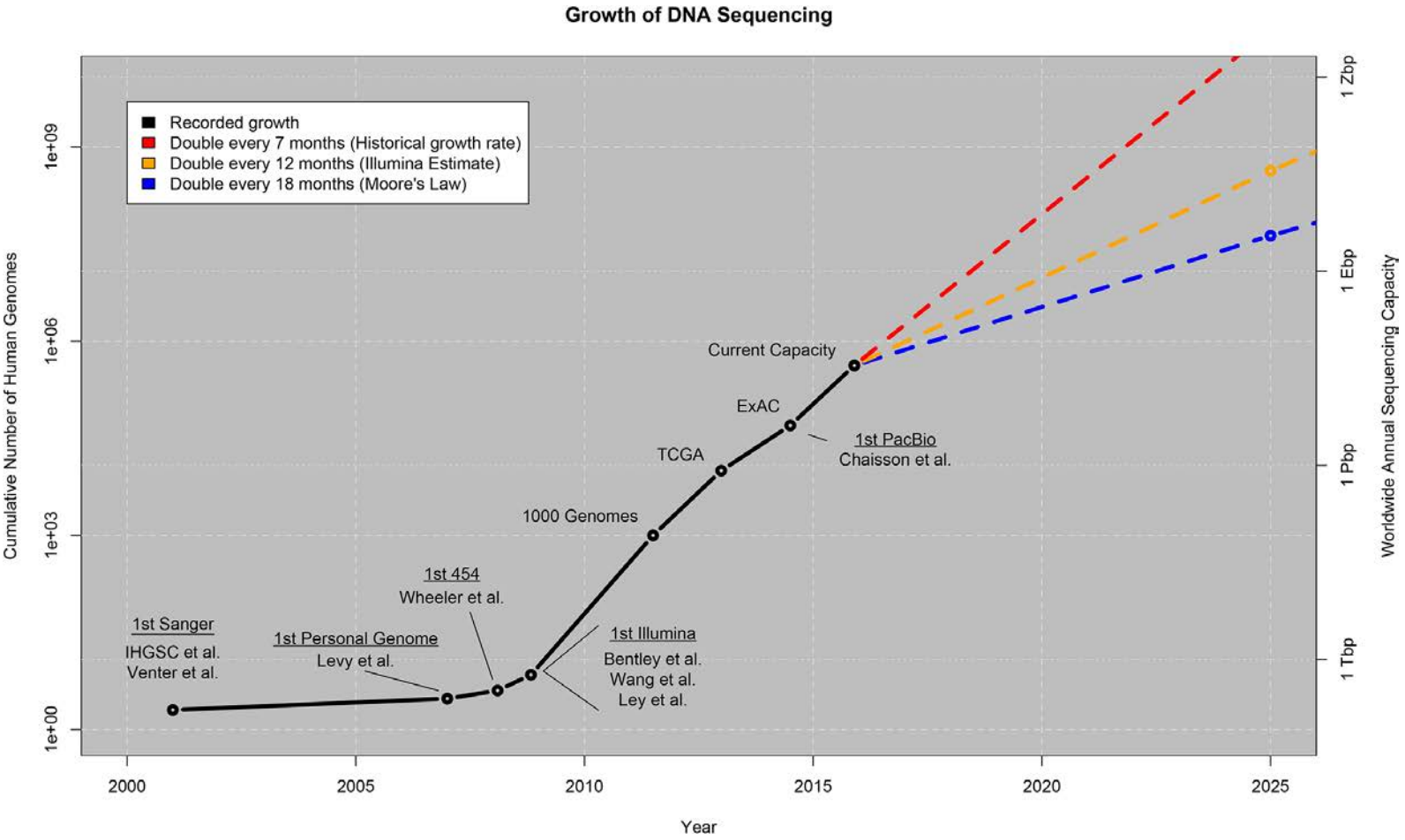
targets = ['J04243','M60064']
idString = 'id=' + ','.join(targets)
s += idString + '&db=nucleotide&rettype=fasta'
fileObject = urllib2.urlopen(s)
data = fileObject.read().strip()

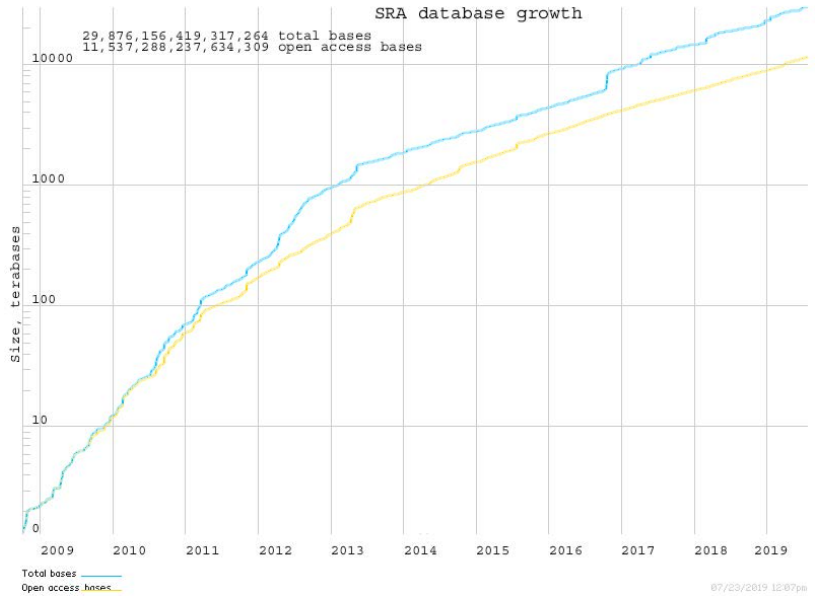
entries = data.split('\n\n')
title, sequence = entries[0].split('\n', 1)
print title.split(' ',1)[0]

# prints:
# >gi|154102|gb|J04243.1|STYHEMAPRF
```

Why do data and computation present special challenges?

There is so much data





https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/

Big Data: Astronomical or Genomical?
Zachary D. Stephens Skylar Y. Lee Faraz Faghri Roy H. Campbell
Chengxiang Zhai Miles J. Efron Ravishankar Iyer Michael C. Schatz Saurabh Sinha Gene E. Robinson

The transformation is recent



https://www.timetoast.com/timelines/telescope-7f500ed9-828b-4dff-9ec6-53322e051786

First telescope 1608

(Lippershey)



~400 Years



Kepler telescope 2009

The transformation is recent

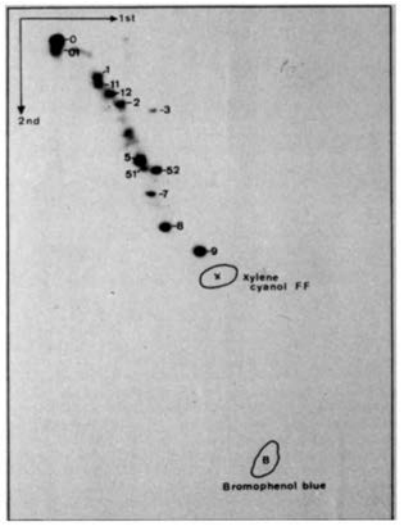
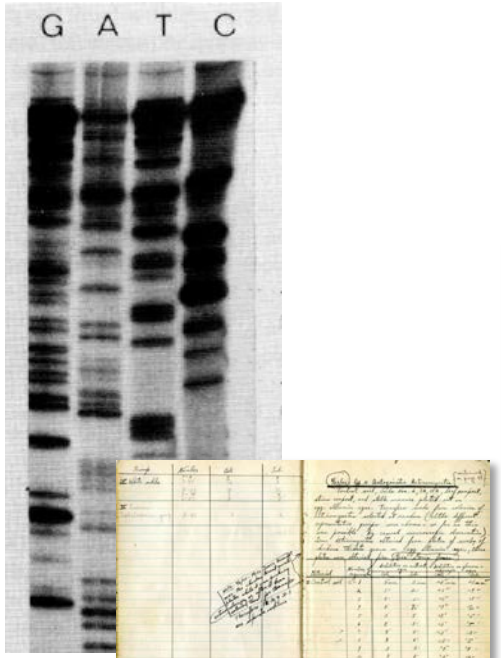


Fig. 1 Two-dimensional separation of band β4. The RNA of band β4 was extracted and subjected to two-dimensional gel electrophoresis. The arrows indicate the directions of the run: the first dimension was on an 8% gel in 0.025 M citric acid+6 M urea; subsequent separation was on a 16% gel in 0.04 M Tris-acetate, pH 8. The positions of the dye markers xylene cyanol FF (X) and bromophenol blue (B) are marked on the figure. Spots *b*1, *b*2 and *b*5 correspond to different regions of the coat protein cistron (Table 2). Spot *b*12 of this particular separation is a rarely observed fragment starting around the GGCGGA sequence (coding for amino-acids 13–14) and going till arrow 14 in Fig. 2.



~40 Years



First RNA Holly sequence, 1965
Sanger Sequence 1977

Solexa Genome Analyzer 2006

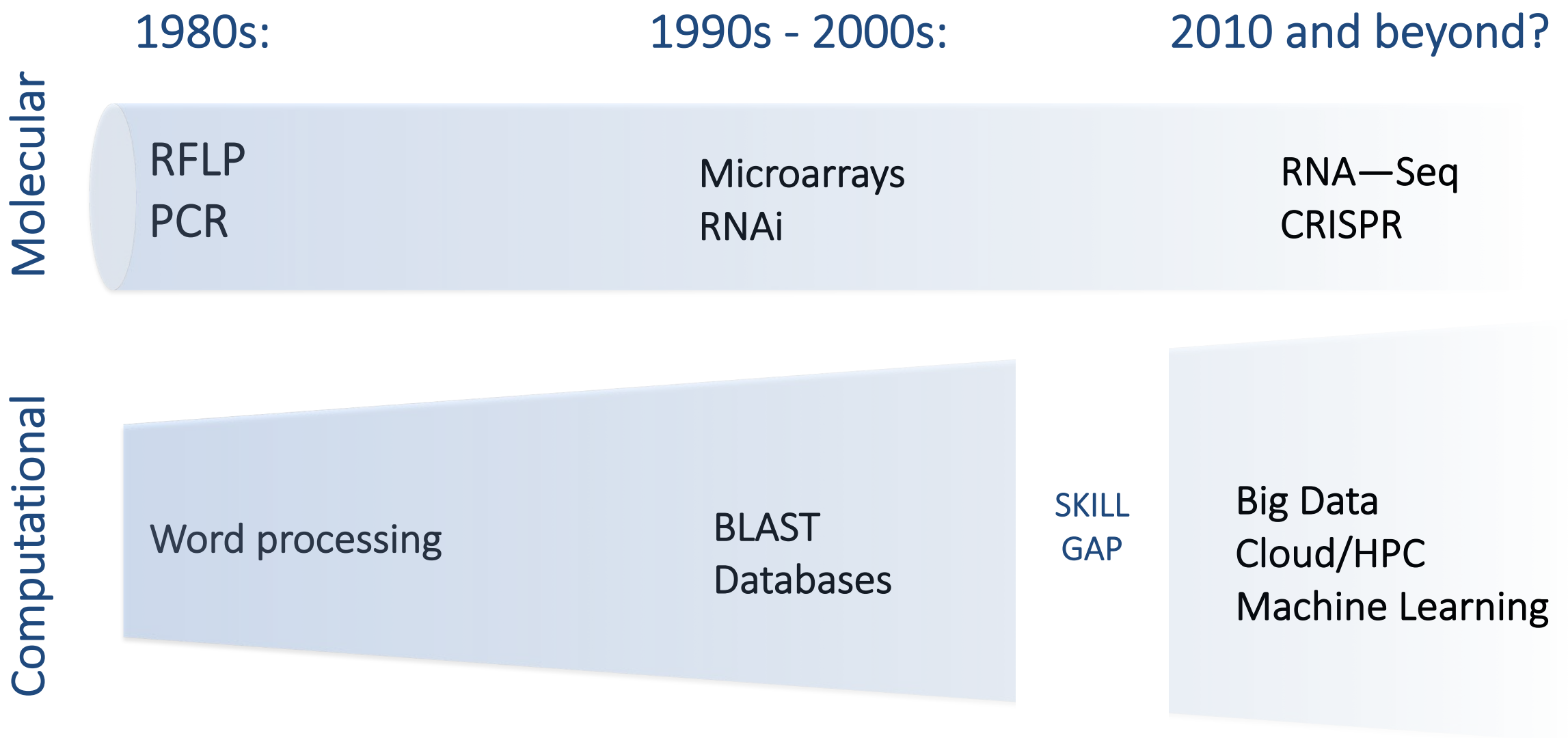
https://www.nature.com/articles/237082a0.pdf
http://dnaftb.org/23/animation.html

New faculty aren't integrating bioinformatics

Decade of Highest Degree	Formal Bioinformatics Training (%)	Faculty Integrating Bioinformatics (%)
1980-1989	8.4	35.4
1990-1999	11.3	41.9
2000-2009	35.1	41.7
2010-2016	48.3	25.2

Williams JJ, Drew JC, Galindo-Gonzalez S, Robic S, Dinsdale E, et al. (2019) **Barriers to integration of bioinformatics into undergraduate life sciences education: A national study of US life sciences faculty uncover significant barriers to integrating bioinformatics into undergraduate instruction**. PLOS ONE 14(11): e0224288. https://doi.org/10.1371/journal.pone.0224288

What skills are in the biology toolbox?



Good infrastructure scales people and compute

"The computer is incredibly fast, accurate, and stupid. Man is unbelievably slow, inaccurate, and brilliant. The marriage of the two is a force beyond calculation." *– Leo Cherne*

How do we tune into the slow, inaccurate brilliance of persons?

CyVerse vision



Transforming science through data-driven discovery

More than 70K users, PBs of data, and hundreds of publications, courses, and discoveries

CyVerse evolution



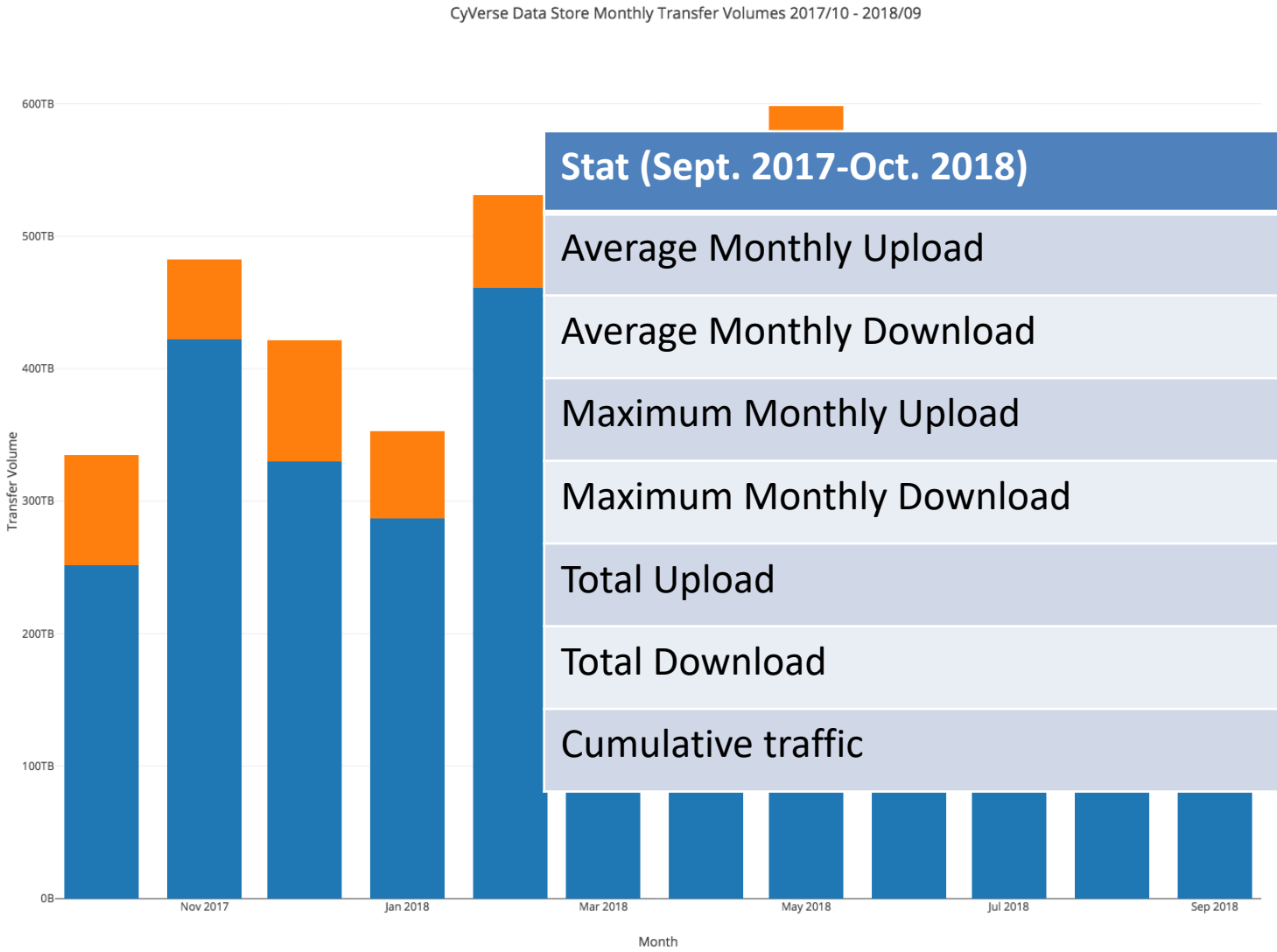
iPlant 2008
Empowering a New Plant Biology

iPlant 2013
Cyberinfrastructure for Life Sciences

CyVerse 2016
Transforming Science Through Data-Driven Discovery

CyVerse Mission:
To design, deploy, and expand a national Cyberinfrastructure for Life Sciences research, and to train scientists in its use

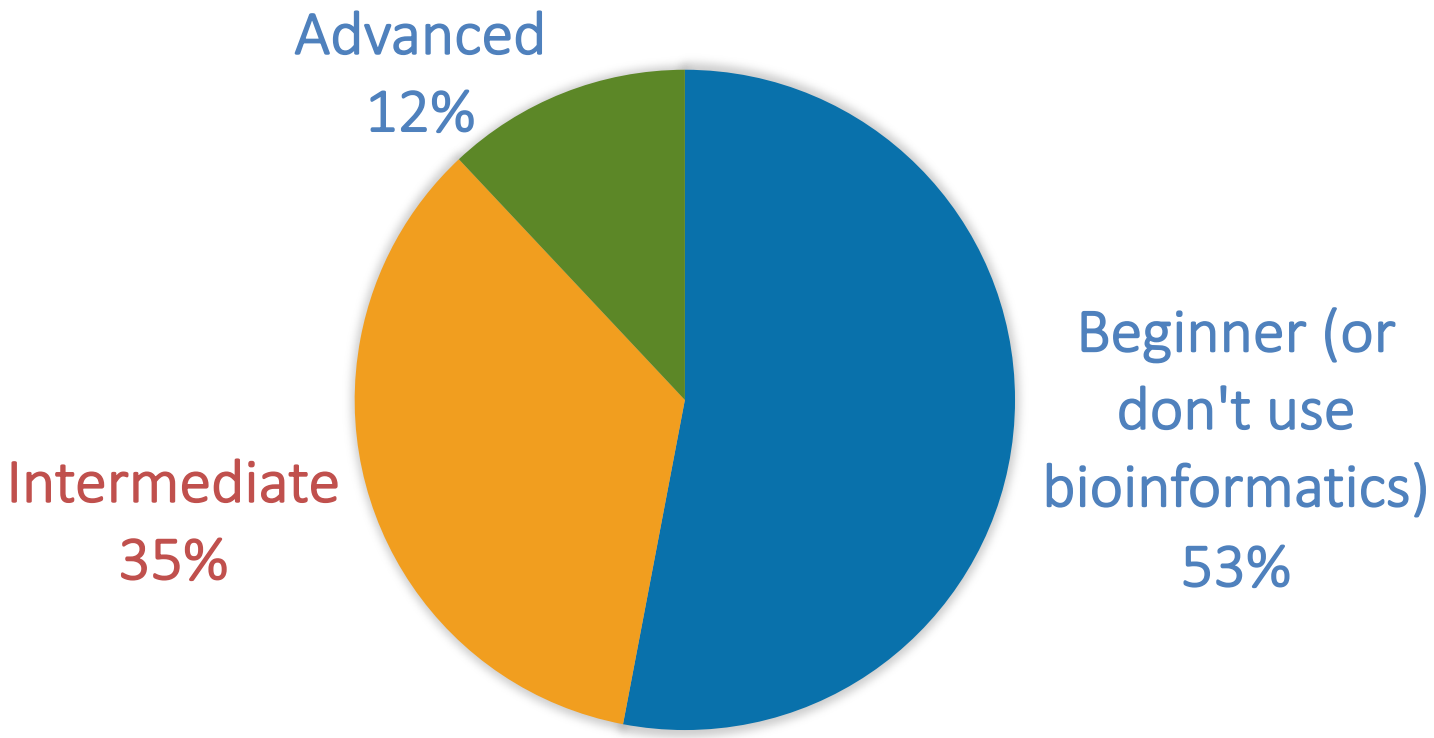
CYVERSE



Stat (Sept. 2017-Oct. 2018)	Size
Average Monthly Upload	72 TB
Average Monthly Download	403 TB
Maximum Monthly Upload	93 TB
Maximum Monthly Download	517 TB
Total Upload	860 TB
Total Download	4834 TB
Cumulative traffic	5.7 PB

CYVERSE

“How do you rate your level of bioinformatics skills?”



Recent NSF program announcement
Data Release, Sharing and Management

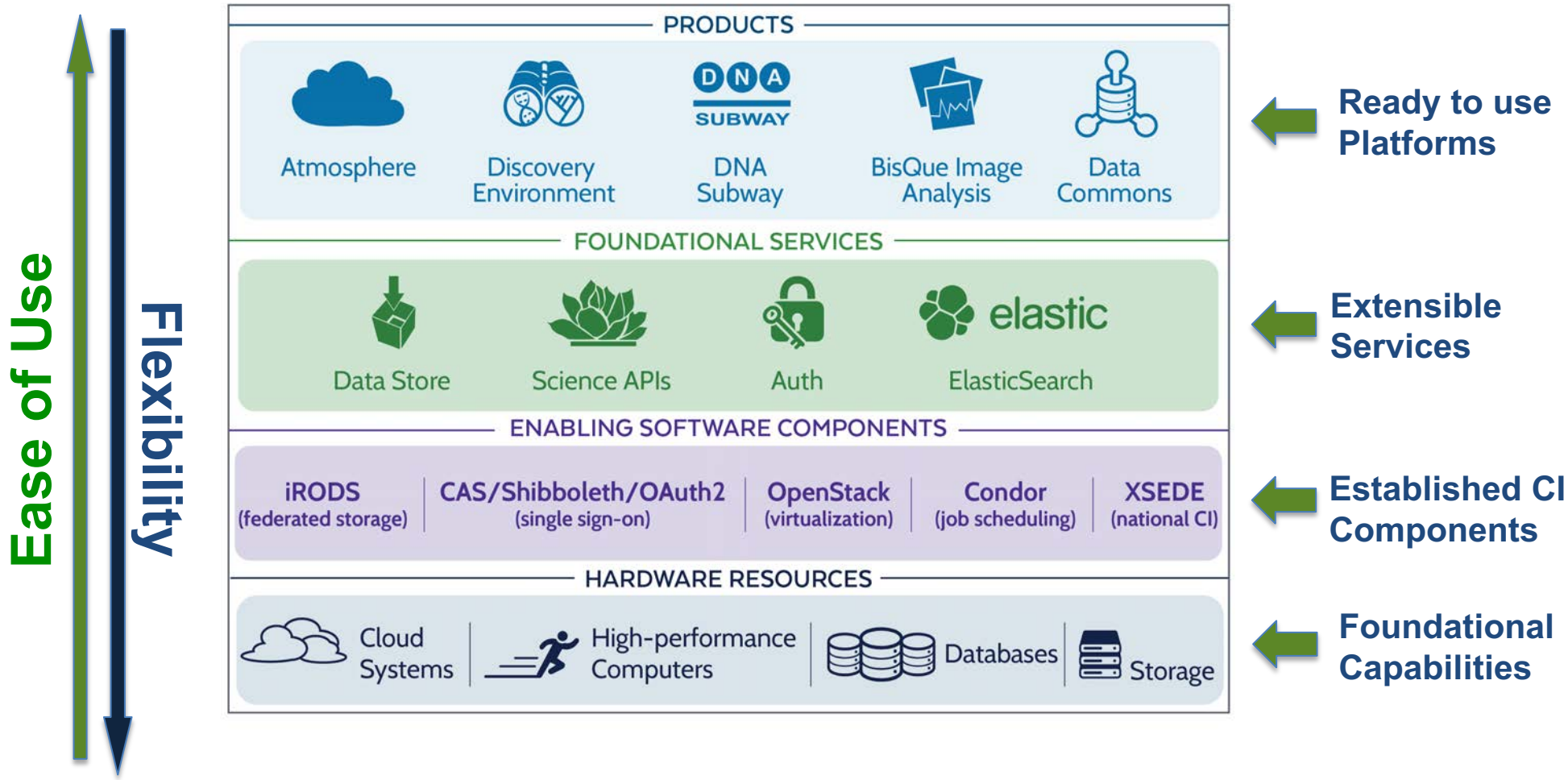
- All software and code must be in a **versioned** code repository (e.g., GitHub/BitBucket) and released immediately.

- Code should be well **documented** for others to **reuse**. We strongly encourage release of ready-to-use software and code through **integration with computing resources** (e.g., Galaxy, CyVerse)

- In **Virtual Machines** (e.g., AWS, JetStream), and in **Containers** (e.g., Docker/DockerHub).

https://www.nsf.gov/pubs/2018/nsf18579/nsf18579.htm

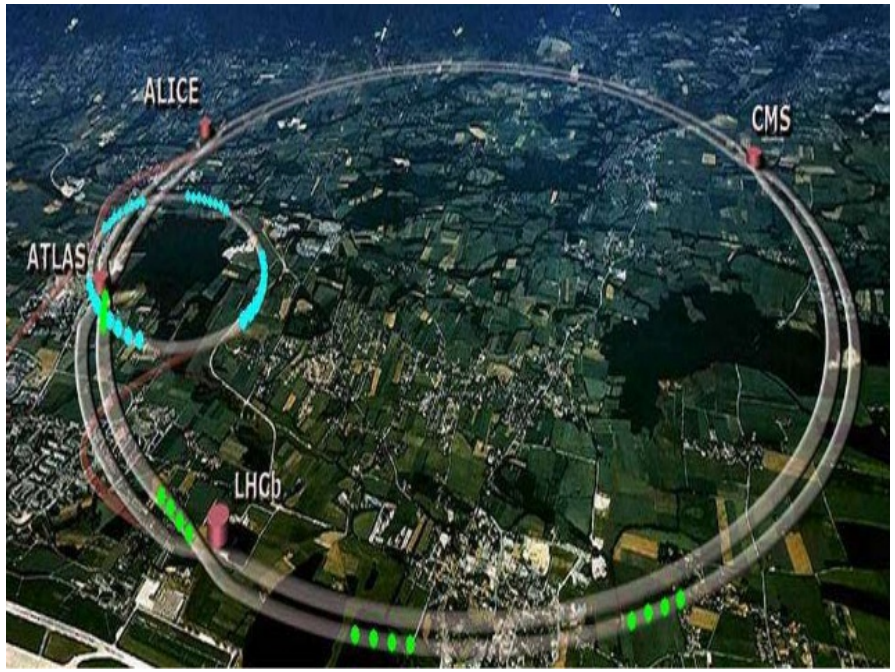
Infrastructure stack



Towards next-gen applications

- Distributed Data Grids
- Container Orchestration
- Distributed Computing
- Integrating with all of the above !
- Knowing how to manage your data and computation flowing through these systems
- No single provider of infrastructure, but a federation

(Physics/Astro) big data, centralized science



Large Hadron Collider
(~1 petabyte per second, ~1petabyte per day processed)

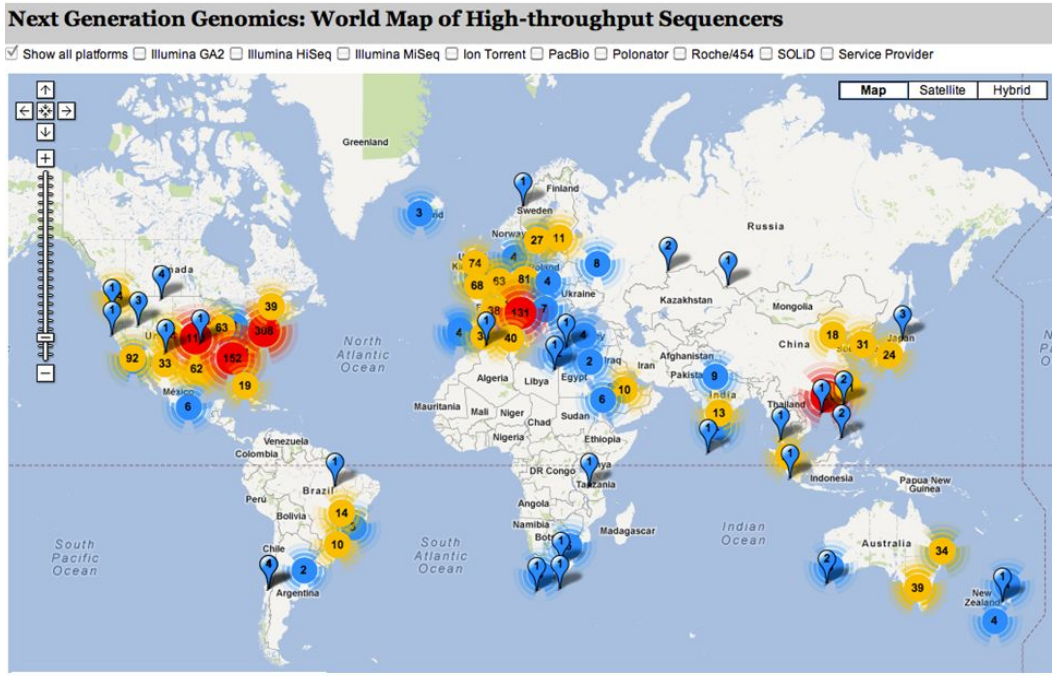


Square Kilometer Array
(~1 Exabyte per day)

(Biology) big data, decentralized science

http://omicsmaps.com/

Nick Loman and James Hadfield

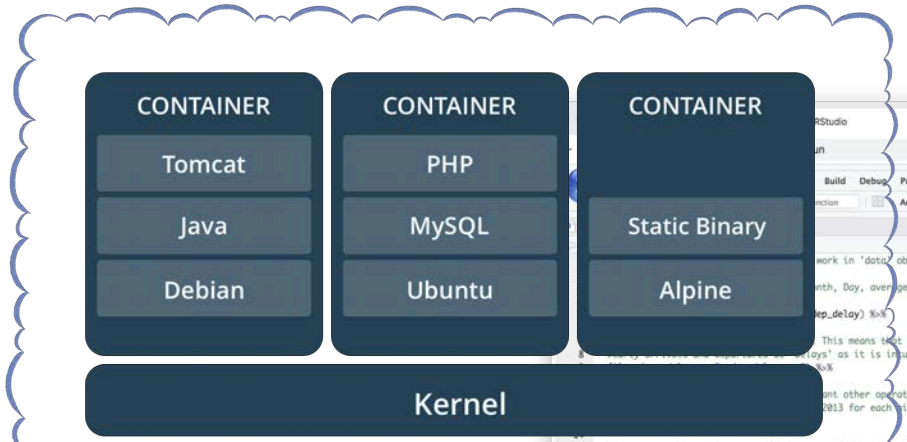


DM Church

Last Updated: 7 May 2012



Modern technology stack enables this transition



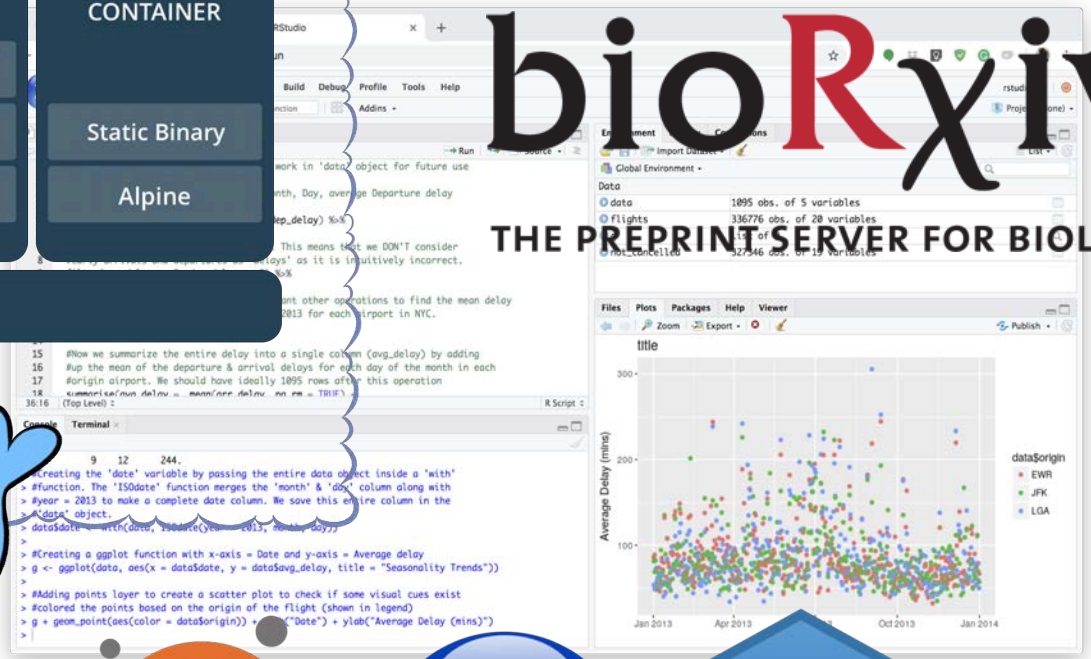
aws

ORCID

GitHub

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY



jupyter

Shiny

Google docs

zenodo

slack

If you can't measure it…

Compute/data infrastructure

- GB/TB/PB of data up/downloads
- # of databases/records
- # of CPUs/GPUs
- Software repositories

Human infrastructure

- # of hackathons?
- # of emails/papers?
- # of under/graduates trained?
- Mental health/happiness?

"But philosophers share the general human weakness for explanations of what is incomprehensible in terms suited for what is familiar and well understood, though entirely different. This has led to the acceptance of implausible accounts of the mental largely because they would permit familiar kinds of reduction"

– T. Nagel

Community is the enabler

What is a community of practice?

"Communities of practice are groups of people who share a concern or a passion for something they do and learn how to do it better as they interact regularly." *- Etienne and Beverly Wenger-Trayner*

Cultural transmission

Cultural transmission is a process of social dissemination in which behaviour patterns, cosmological beliefs, the culture's technological knowledge, etc. are communicated and acquired

Eyssartier, C., Ladio, A.H. & Lozada, M. **Cultural Transmission of Traditional Knowledge in two populations of North-western Patagonia**. *J Ethnobiology Ethnomedicine* **4,** 25 (2008) doi:10.1186/1746-4269-4-25

Cultural transmission

Cultural transmission is a process of social dissemination in which behaviour patterns, cosmological beliefs, the culture's technological knowledge, etc. are communicated and acquired

You should use version control

Eyssartier, C., Ladio, A.H. & Lozada, M. **Cultural Transmission of Traditional Knowledge in two populations of North-western Patagonia**. *J Ethnobiology Ethnomedicine* **4,** 25 (2008) doi:10.1186/1746-4269-4-25

Informal, socio-(cultural) learning



Dictionary style replace multiple items

Asked 8 years, 2 months ago Active 7 months ago Viewed 23k times

mailchimp
Improve ROI with real-time marketing campaign analytics.
Learn More
ADS VIA CARBON

28

18

I have a large data.frame of character data that I want to convert based on what is commonly called a dictionary in other languages.

Currently I am going about it like so:

```
foo <- data.frame(snp1 = c("AA", "AG", "AA", "AA"), snp2 = c("AA", "AT", "AG", "AA"),
foo <- replace(foo, foo == "AA", "0101")
foo <- replace(foo, foo == "AC", "0102")
foo <- replace(foo, foo == "AG", "0103")
```

This works fine, but it is obviously not pretty and seems silly to repeat the `replace` statement each time I want to replace one item in the data.frame.

Is there a better way to do this since I have a dictionary of approximately 25 key/value pairs?

r dataframe bioinformatics

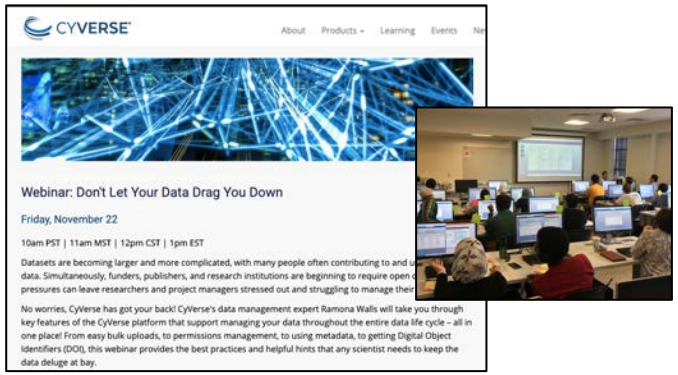
Situating Learning in Communities of Practice

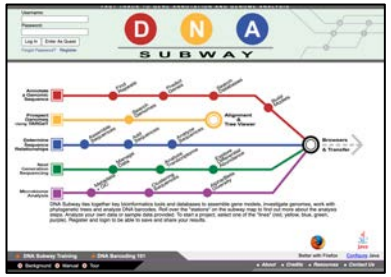
- Jean Lave (1991)

In this chapter, I propose to consider learning not as a process of socially shared cognition that results in the end in the internalization of knowledge by individuals, but as a process of becoming a member of a sustained community of practice. Developing an identity as a member of a community and becoming knowledgeably skillful are part of the same process, with the former motivating, shaping, and giving meaning to the latter, which it subsumes.

CyVerse's path to community-building

Partnering/Basic Training





THE CARPENTRIES

Refactor Infrastructure







Foster CoP





USA **UK**

CYVERSE™

Austria **Australia**

Technology doesn't solve science problems, people do