



Using the Open Science Data Cloud for Research in Data Intensive Science

Robert Grossman
University of Chicago
Open Cloud Consortium

June 16, 2014
OSDC PIRE Workshop

Outline

1. Complexity of disease genomic data
2. Big data and the disruption it causes
3. Computing over big data
4. Science clouds
5. Data analysis at scale
6. Genomic clouds
7. How might we organize?



A challenge for students



Cloud Services for the Scientific Community

The OSDC provides petabyte-scale cloud resources that let you easily analyze, manage, and share data.

Get Started Now

OSDC Console Login

Featured on the OSDC

Project Matsu is a collaboration with NASA to process *Earth Observing 1 (EO-1)* satellite imagery to detect fires and floods and provide relevant information to first responders. The data is freely available from the OSDC to interested users.

How can I get involved?

Apply

Fill out a short application for an OSDC resource allocation. Allocations start at 16 dedicated cores and 1TB of storage, but scale depending on the project needs and level of organizational partnership.

Partner

Partner with us and add your own racks to the OSDC (we will manage them for you). Organizations can also join the Open Cloud Consortium (OCC) which is made up of working groups, including the OSDC.

Develop

All of the software developed as part of the OSDC is open source and hosted on GitHub. You can directly help the scientific cloud computing community by contributing to the open source OSDC software stack.

Contact Us

Questions? Comments? Suggestions? Contact us at info@opencloudconsortium.org.

The Open Science Data Cloud hosts over 1PB of scientific data, 10,000 cores and free accounts for everyone in this room.



Cloud Services for the Scientific Community

The OSDC provides petabyte-scale cloud resources that let you easily analyze, manage, and share data.

[Get Started Now](#)

[OSDC Console Login](#)

Featured on the OSDC

Project Matsu NASA

Project Matsu is a collaboration with NASA to process Earth Observing 1 (EO-1) satellite imagery to detect fires and floods and provide relevant information to first responders. The data is freely available from the OSDC to interested users.

How can I get involved?

Apply

Fill out a short application for an OSDC resource allocation. Allocations start at 16 dedicated cores and 1TB of storage, but scale depending on the project needs and level of organizational partnership.

Partner

Partner with us and add your own racks to the OSDC (we will manage them for you). Organizations can also join the Open Cloud Consortium (OCC) which is made up of working groups, including the OSDC.

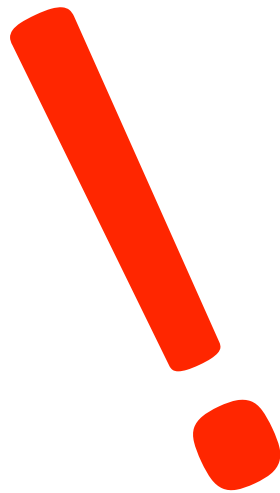
Develop

All of the software developed as part of the OSDC is open source and hosted on GitHub. You can directly help the scientific cloud computing community by contributing to the open source OSDC software stack.

Contact Us

Questions? Comments? Suggestions? Contact us at info@opencloudconsortium.org.

1. Make a discovery using the data in the Open Science Data Cloud.
2. Write an OSDC application so that others can easily make similar discoveries.



An opportunity for
research scientists

Cyber Condo Model

- Research institutions today have access to high performance networks – 10G & 100G.
- They couldn't afford access to these networks from commercial providers.
- Over a decade ago, they got together to buy and light fiber.
- This changed how we do scientific research.
- Cyber condos interoperate with commercial ISPs



Science Cloud Condos

- We're building a Science Cloud condo.
- To provide: i) a sustainable home for large commons of research data; and ii) an infrastructure to compute over it.
- "Tier 1" Science Clouds need to establish peering relationships.
- And to interoperate with CSPs
- The Open Cloud Consortium (OCC) is participating and looking for other science cloud condos to peer with.



OSDC Organization

- Robert Grossman

Director, Open Science Data Cloud

- Maria Patterson

Scientific Lead for Open Science Data Cloud

- Allison Heath

Scientific Lead for Bionimbus Protected Data Cloud and Genomics Data Commons

- Heidi Alvarez

Lead, OSDC NSF PIRE Project

Part 1

What is Big Data. What is Data Science and Why Do We Care?

Degrees of the length of Array 0°-100°	Estimates in lbs.	Centiles		Excess of Observed over Normal
		Observed deviates from 1207 lbs.	Normal p.e = 37	
5	1074	- 133	- 90	+ 43
10	1109	- 98	- 70	+ 28
15	1126	- 81	- 57	+ 24
20	1148	- 59	- 46	+ 13
<i>q</i> ₁ 25	1162	- 45	- 37	+ 8
30	1174	- 33	- 29	+ 4
35	1181	- 26	- 21	+ 5
40	1188	- 19	- 14	+ 5
45	1197	- 10	- 7	+ 3
<i>m</i> 50	1207	0	0	0
55	1214	+ 7	+ 7	0
60	1219	+ 12	+ 14	- 2
65	1225	+ 18	+ 21	- 3
70	1230	+ 23	+ 29	- 6
<i>q</i> ₃ 75	1236	+ 29	+ 37	- 8
80	1243	+ 36	+ 46	- 10
85	1254	+ 47	+ 57	- 10
90	1267	+ 52	+ 70	- 18
95	1293	+ 86	+ 90	- 4

*q*₁, *q*₃, the first and third quartiles, stand at 25° and 75° respectively.
m, the median or middlemost value, stands at 50°.
 The dressed weight proved to be 1198 lbs.

Source: Distribution of the estimates of the dressed weight of a particular living ox, made by 787 different persons, Francis Galton, Nature (1907), No. 1949, Vol. 75, 450-451.

What is Big Data?

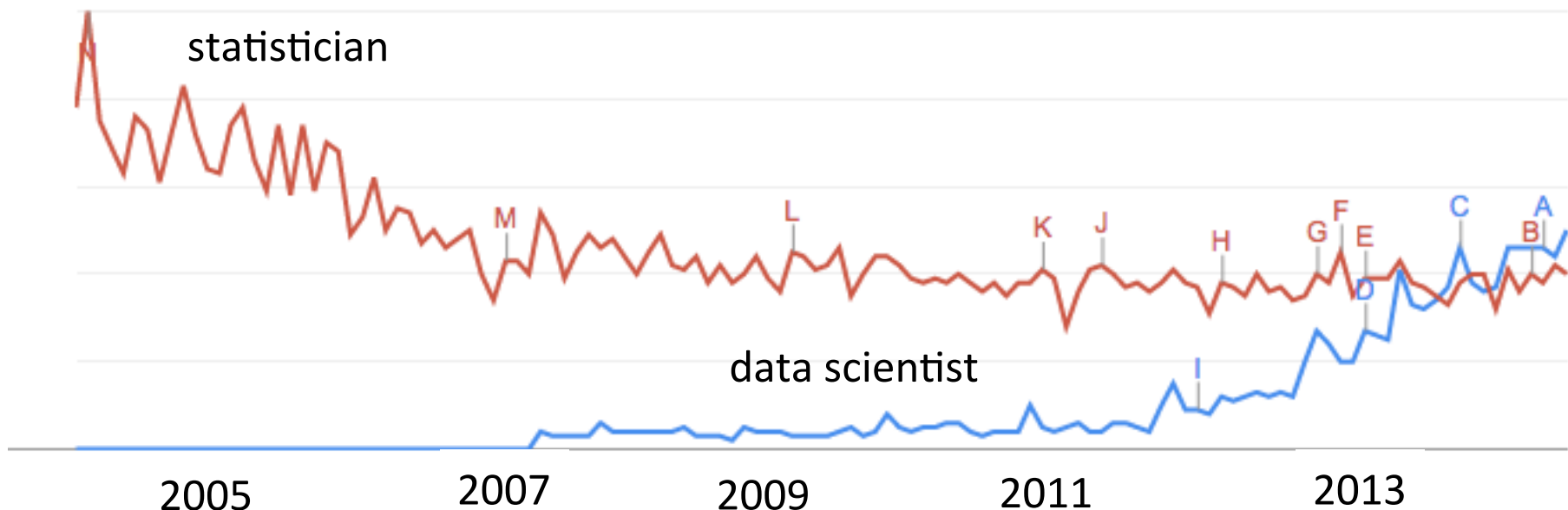
A marketing term introduced by O'Reilly:

Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it.

Edd Dumbill, What is Big Data?, strata.oreilly.com, January 11, 2012.

Data Scientist as a Job Category

- Starting in 2008, Jeff Hammerbacher (at Facebook) and D.J. Patil (at LinkedIn) used the term data scientist as a job title for those who “use both data and science to create



Source: Google Trends (www.google.com/trends)

Data Science as a Scientific Discipline

Bold new partnership launches to harness potential of data scientists and big data

GORDON AND BETTY
MOORE
FOUNDATION



November 12, 2013

New York University, the University of California, Berkeley and the University of Washington launch a 5-year, \$37.8 million cross-institutional effort with support from the Gordon and Betty Moore Foundation and Alfred P. Sloan Foundation

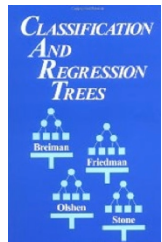
Washington, D.C. – A new multi-million dollar collaboration will enable university researchers to harness the full potential of the data-rich world that characterizes all fields of science and discovery. This ambitious partnership, which includes New York University, the University of

California, Berkeley and the University of Washington, will spur collaborations within and across the three campuses and other partners pursuing similar data-intensive science goals.

Computationally Intensive Statistics

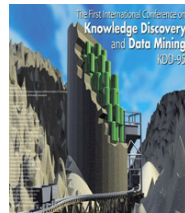


Direct marketing

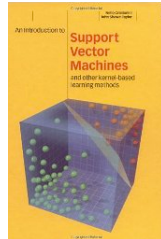


1984

Data Mining & KDD



POS



1993

Predictive Analytics



Internet

PageRank



2004

Big Data / Data Science

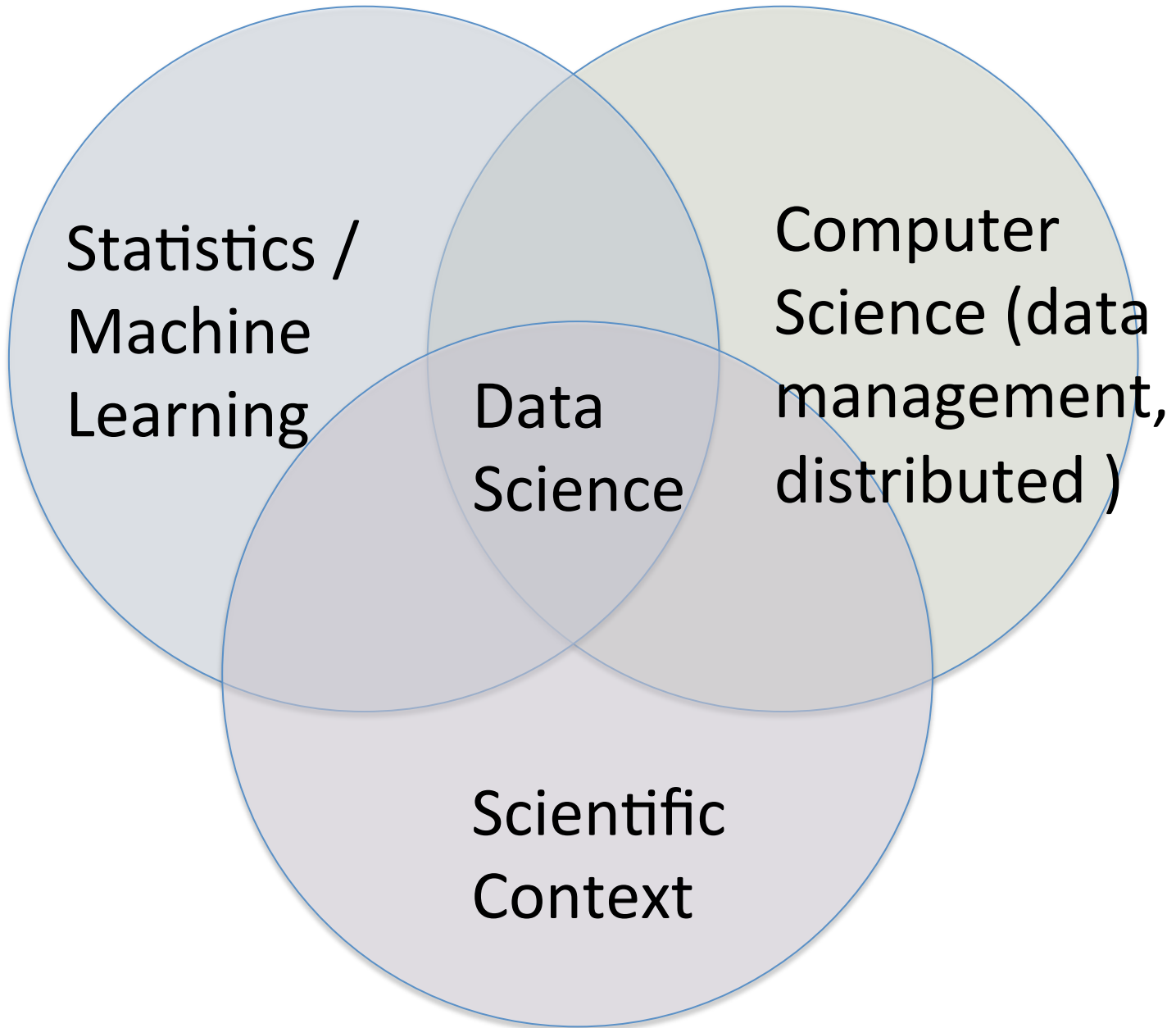


Devices

Spanner TX algorithm



2011

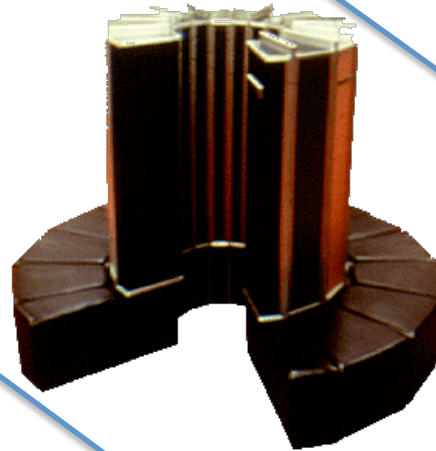


2004
10x-100x



data
science

1976
10x-100x



simulation
science

1670
250x



experimental
science

1609
30x

Data Science as an Academic Discipline

- Computer science emerged in the 1960's & 70s'
 - Computer science emerged out of math, engineering and management
 - There was skepticism at the beginning whether computer science was a separate discipline.
- Today there is an debate whether data science is a separate discipline or whether it will eventually be integrated into astronomy, biology, physics, etc.

Do We Need a New Discipline?

- Statistics has traditionally analyzed analyzes scalar and vector data (and times series and spatial fields built from them).
- Data Science analyzes a wider variety of data, including heterogeneous and unstructured data, media, etc.
- Dhar: Data science is the study of the generalizable extraction of knowledge from data.
- Wladawsky-Berger: Statistics explains, while data science extracts actionable knowledge from data.
- Note that this is essentially the same way that “data mining” was distinguished from statistics during the period 1995-2005.

Part 2

Data Center Scale Computing (aka “Cloud Computing”)



Source: Interior of one of Google's Data Center, www.google.com/about/datacenters/

What instrument do we use to make big data discoveries in cancer genomics and big data biology?



How do we build a “datascope?”



Self Service

Scale

Forgot cloud computing. Focus on data centers & the software they run.

Software stack that scales to a data center

Use of automation



MORGAN & CLAYPOOL PUBLISHERS

The Datacenter as a Computer

*An Introduction to the Design
of Warehouse-Scale Machines*

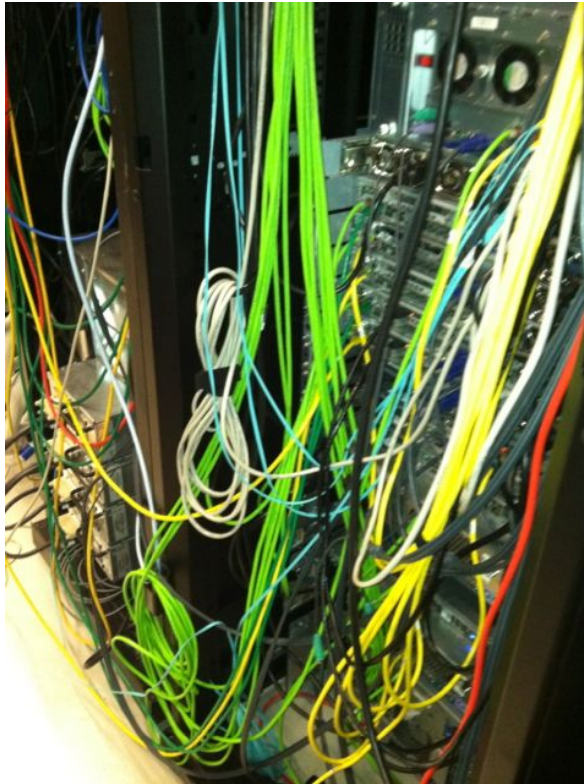
Second Edition

Luiz André Barroso
Jimmy Clidaras
Urs Hölzle

**SYNTHESIS LECTURES ON
COMPUTER ARCHITECTURE**

Mark D. Hill, *Series Editor*

Source: Luiz André Barroso, Jimmy Clidaras and Urs Hölzle, The Datacenter as a Computer, Morgan & Claypool Publishers, Second Edition, 2013, www.morganclaypool.com/doi/pdf/10.2200/S00516ED2V01Y201306CAC024

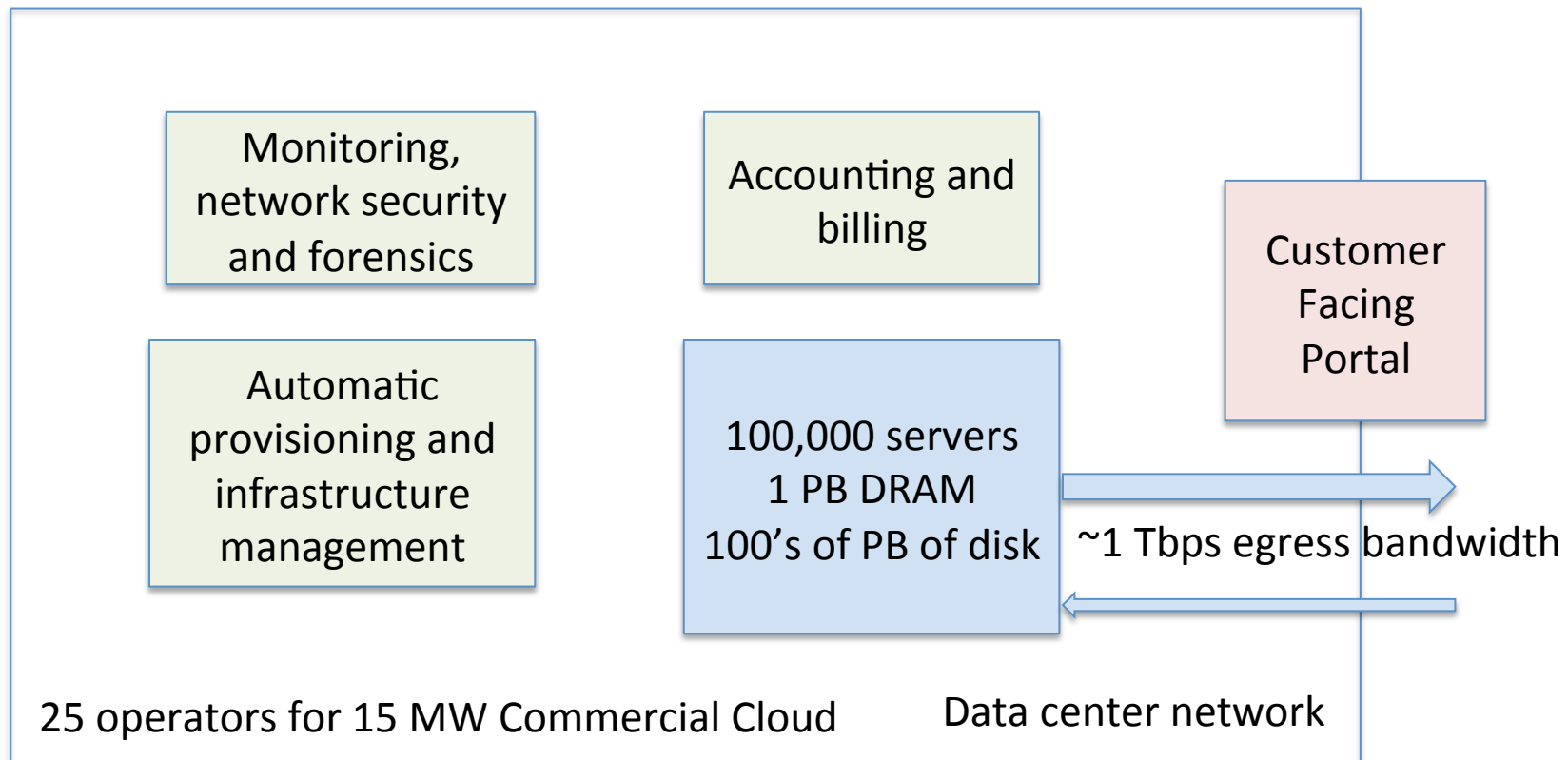


This is not data center
scale computing...



Setting up and operating large scale efficient, secure and compliant racks of computing infrastructure is out of reach for most labs, but essential for the community.

Commercial Cloud Service Provider (CSP) 15 MW Data Center





www.openstack.org

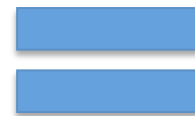


OPEN
Compute Project

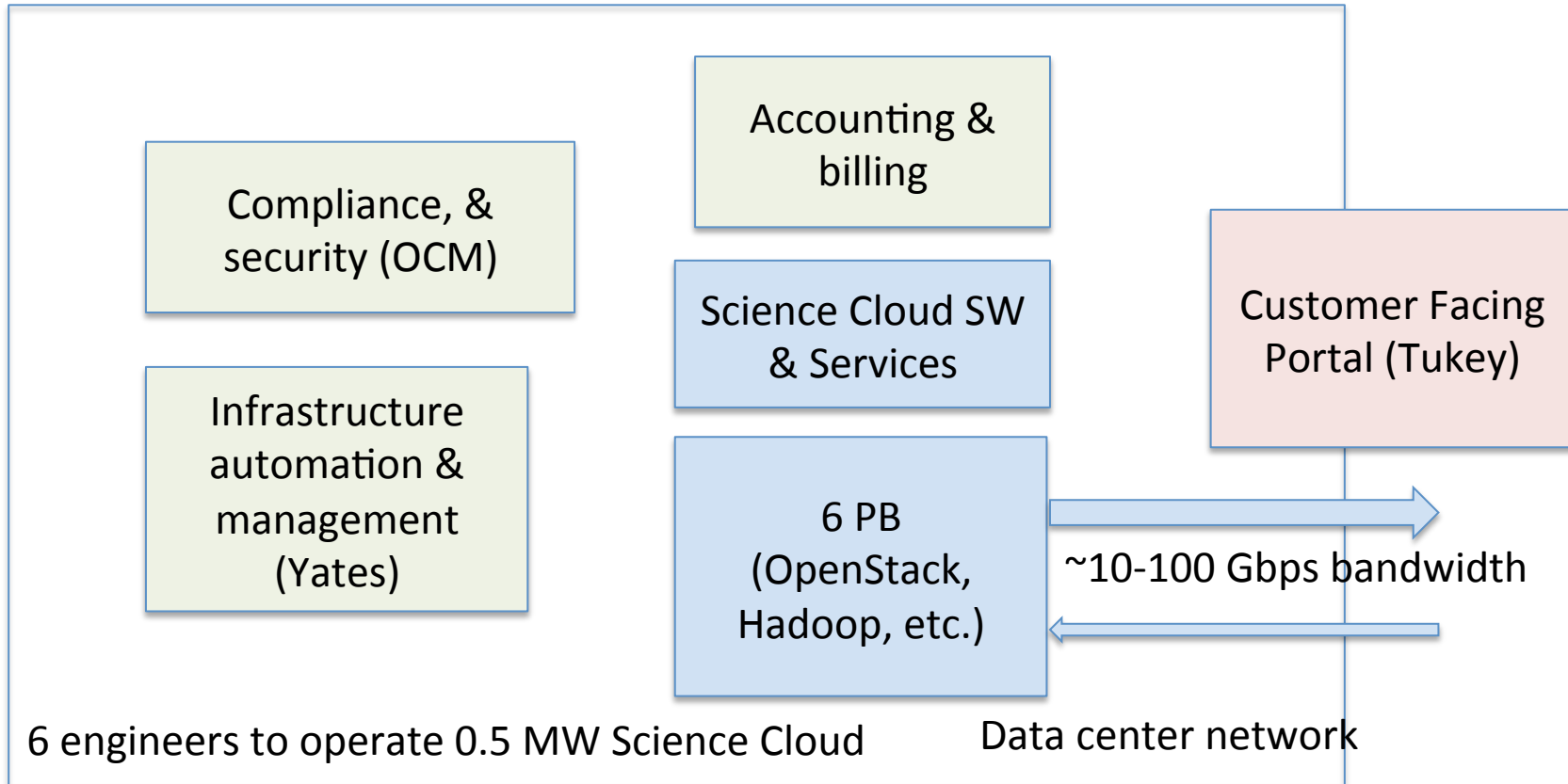
opencompute.org



hadoop.apache.org



Open Science Data Cloud (OSDC)



Why not (only) use Amazon Web Services (AWS)?

Commercial Cloud Service Providers (CSP)

- Scale / capacity
- Simplicity of a credit card
- Wide variety of offerings.

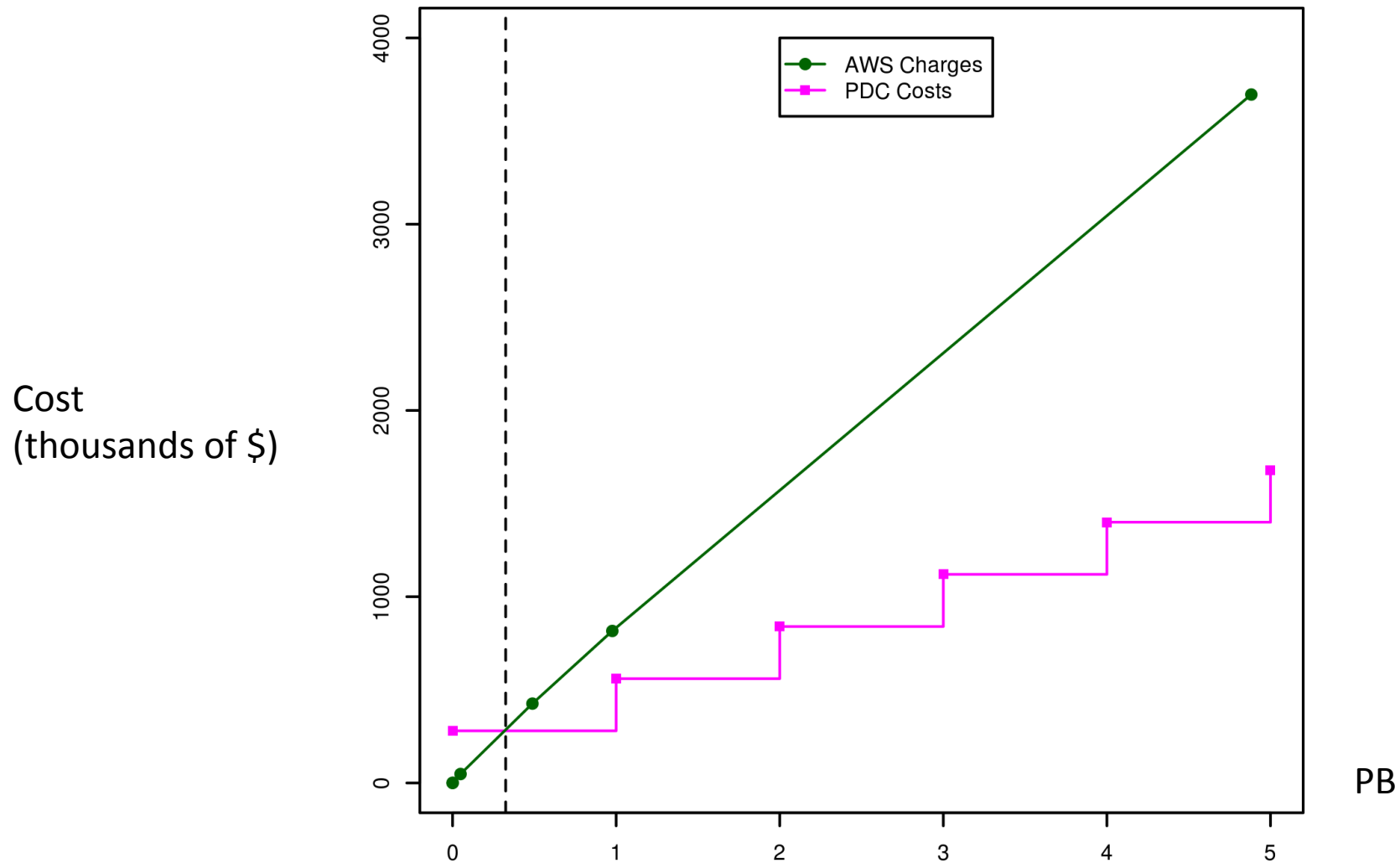
vs.

Community science and biomedical clouds

- Lower cost (at medium scale)
- We can build specialized infrastructure for science.
- We can build specialized infrastructure for security & compliance.
- The data is too important to trust exclusively with a commercial provider.

It is still essential to interoperate with CSP whenever possible by compliance and security policies.

Cost of a medium private/community cloud vs large public cloud.



Source: Allison P. Heath, Matthew Greenway, Raymond Powell, Jonathan Spring, Rafael Suarez, David Hanley, Chai Bandlamudi, Megan McNerney, Kevin White and Robert L Grossman, Bionimbus: A Cloud for Managing, Analyzing and Sharing Large Genomics Datasets, Journal of the American Medical Informatics Association, 2014.

Reliability over Commodity Computing Components Is Difficult as is Data Locality



- Hadoop enables reliable computation over thousands of low costs, unreliable computing nodes.
- Hadoop efficiently computes over the data instead of moving the data.
- The programming model of Hadoop (MapReduce) is in practice more efficient in terms of software development than the programming traditionally used by high performance computing (message passing) (but usually does not fully utilize the underlying hardware)

Latency is Difficult

DOI:10.1145/2408776.2408784

Software techniques that tolerate latency variability are vital to building responsive large-scale Web services.

BY JEFFREY DEAN AND LUIZ ANDRÉ BARROSO

The Tail at Scale

as overall use increases. Temporary high-latency episodes (unimportant in moderate-size systems) may come to dominate overall service performance at large scale. Just as fault-tolerant computing aims to create a reliable whole out of less-reliable parts, large online services need to create a predictably responsive whole out of less-predictable parts; we refer to such systems as “latency tail-tolerant,” or simply “tail-tolerant.” Here, we outline some common causes for high-latency episodes in large online services and describe techniques that reduce their severity or mitigate their effect on whole-system performance. In many cases, tail-tolerant techniques can take advantage of resources already deployed to achieve fault-tolerance, resulting in low additional overhead. We explore how these techniques allow system utilization to be driven higher without lengthening the latency tail, thus avoiding wasteful overprovisioning.

The Tail at Scale, Jeffrey Dean, Luiz André Barroso Communications of the ACM, Volume 56 Number 2, Pages 74-80

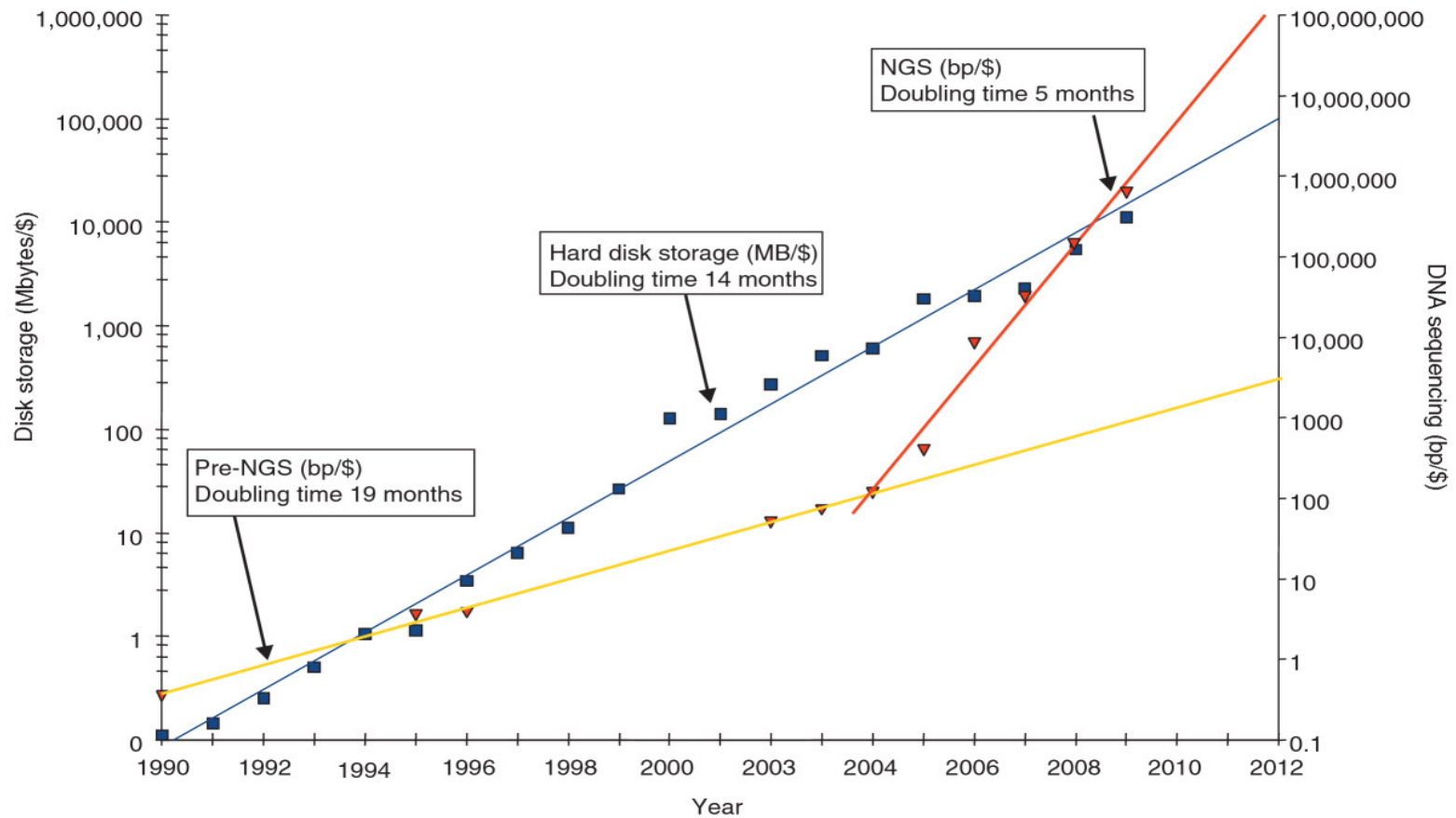
The Rack Test



Call an algorithm and computing infrastructure is “**big-data scalable**” if adding a rack of data (and corresponding processors) does not increase the time required to complete the computation but enables the computation to run on a rack more of data. Most of our community’s algorithms today fail this test.

Part 3

Science Clouds



Source: Lincoln Stein

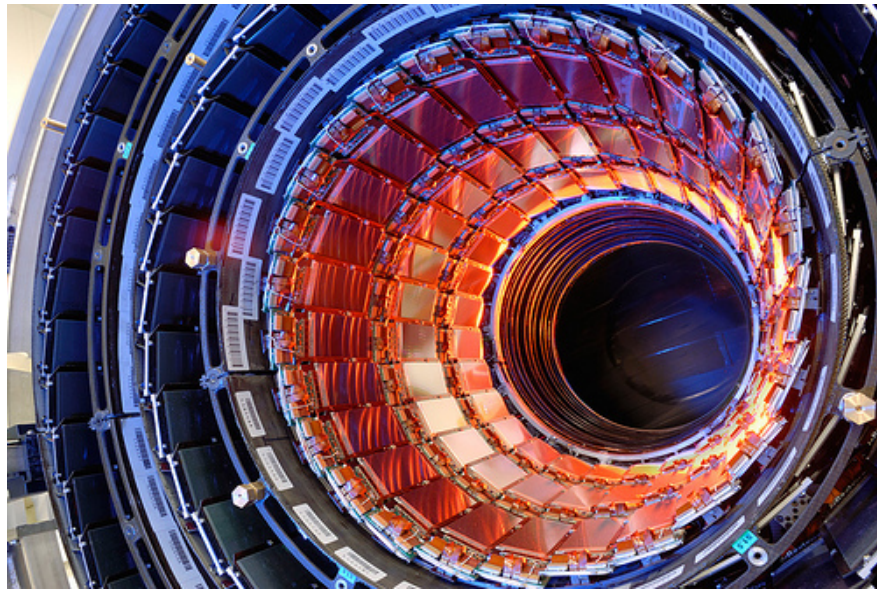
Big Data Sciences

Discipline	Duration	Size	# Devices
HEP - LHC	10 years	15 PB/year*	One
Astronomy - LSST	10 years	12 PB/year**	One
Genomics - NGS	2-4 years	0.4 TB/genome	1000's

*At full capacity, the Large Hadron Collider (LHC), the world's largest particle accelerator, is expected to produce more than 15 million Gigabytes of data each year. ... This ambitious project connects and combines the IT power of more than 140 computer centres in 33 countries. Source: http://press.web.cern.ch/public/en/Spotlight/SpotlightGrid_081008-en.html

**As it carries out its 10-year survey, LSST will produce over 15 terabytes of raw astronomical data each night (30 terabytes processed), resulting in a database catalog of 22 petabytes and an image archive of 100 petabytes. Source: <http://www.lsst.org/News/enews/teragrid-1004.html>

Common computing, storage & transport infrastructure



VS



- Ten years and \$10B
- No business value
- Big data culture
- Little compliance

- Five years and \$1M
- Business value
- Culture of small data
- Compliance

Source: A picture of Cern's Large Hadron Collider (LHC). The LHC took about a decade to construct, and cost about \$4.75 billion.
Source of picture: Conrad Melvin, Creative Commons BY-SA 2.0, www.flickr.com/photos/58220828@N07/5350788732

Science vs Commercial Clouds

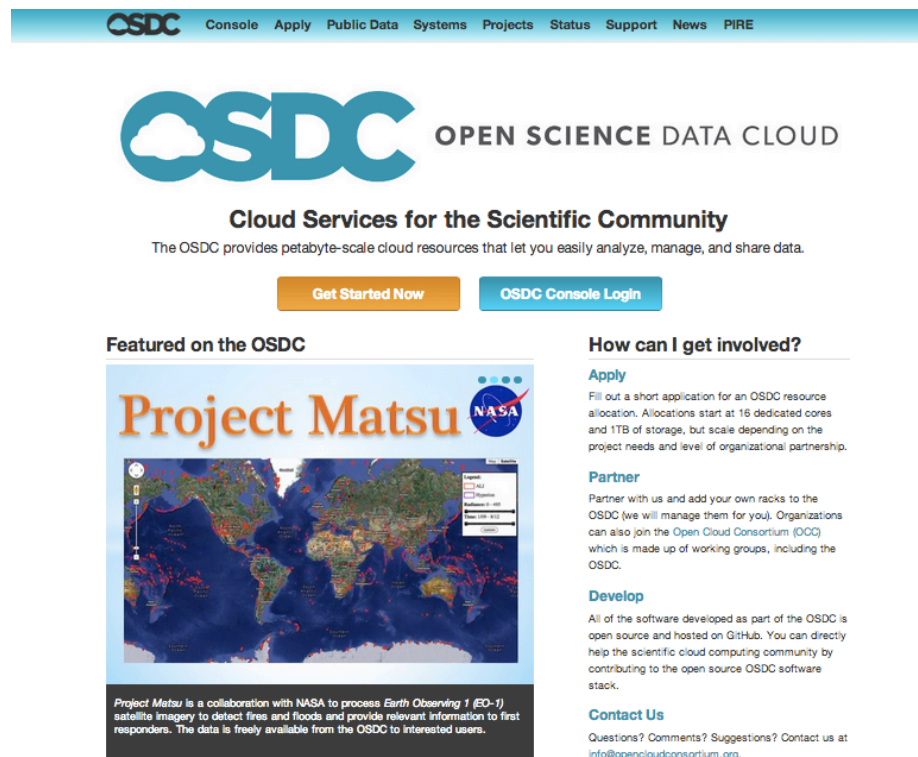
	Science CSP	Commercial CSP
Perspective	Democratize access to data. Integrate data to make discoveries. Long term archive.	As long as you pay the bill; as long as we keep the our business model.
Data	Data intensive computing	Internet style scale out
Flows	Large data flows in and out	Lots of small web flows
Lock in	Data and compute portability essential	Lock in is good

A Key Question

- Is scientific computing at the scale of a data center important enough for the research community to do or do we *only* outsource to commercial cloud service providers (certainly we will interoperate with commercial cloud service providers)?

Part 4


Open Science Data Cloud



The screenshot shows the OSDC website homepage. At the top is a navigation bar with the OSDC logo and links for Console, Apply, Public Data, Systems, Projects, Status, Support, News, and PIRE. Below the navigation bar is the OSDC logo and the text "OPEN SCIENCE DATA CLOUD". Underneath is the tagline "Cloud Services for the Scientific Community" and a sub-headline "The OSDC provides petabyte-scale cloud resources that let you easily analyze, manage, and share data." Two buttons are present: "Get Started Now" and "OSDC Console Login".

Featured on the OSDC

Project Matsu



The map shows a global view with a legend for "Project Matsu" and a NASA logo. The map displays satellite imagery with red and yellow markers indicating fire and flood detection.

Project Matsu is a collaboration with NASA to process Earth Observing 1 (EO-1) satellite imagery to detect fires and floods and provide relevant information to first responders. The data is freely available from the OSDC to interested users.

How can I get involved?

Apply

Fill out a short application for an OSDC resource allocation. Allocations start at 16 dedicated cores and 1TB of storage, but scale depending on the project needs and level of organizational partnership.

Partner

Partner with us and add your own racks to the OSDC (we will manage them for you). Organizations can also join the Open Cloud Consortium (OCC) which is made up of working groups, including the OSDC.

Develop

All of the software developed as part of the OSDC is open source and hosted on GitHub. You can directly help the scientific cloud computing community by contributing to the open source OSDC software stack.

Contact Us

Questions? Comments? Suggestions? Contact us at info@opencloudconsortium.org.



OPEN CLOUD CONSORTIUM



OPEN CLOUD CONSORTIUM

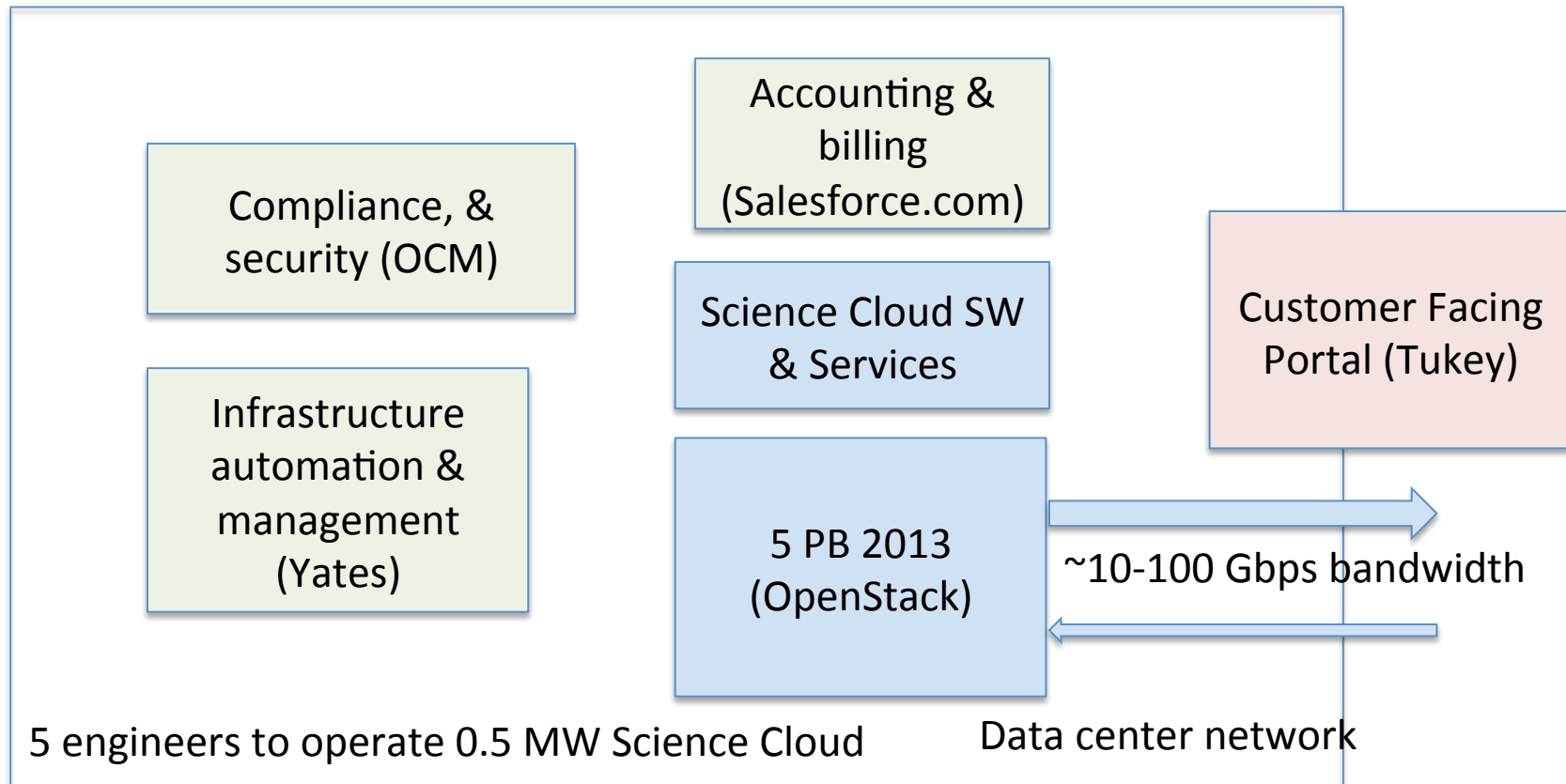
- U.S based not-for-profit corporation that develops and operates cloud computing infrastructure for scientific projects around the world.
- Manages cloud computing infrastructure to support scientific research: Open Science Data Cloud.
- Manages cloud computing testbeds: Open Cloud Testbed.
- Manages cloud computing infrastructure to support medical and health care research: Biomedical Commons Cloud

www.opencloudconsortium.org

OCC Members & Partners

- Companies: Cisco, Yahoo!, Intel, ...
- Universities: University of Chicago, Northwestern Univ., Johns Hopkins, Calit2, ORNL, University of Illinois at Chicago, ...
- Federal agencies and labs: NASA
- International Partners: Univ. Amsterdam, Univ. Edinburgh, AIST (Japan), ...
- Partners: National Lambda Rail

2014 Open Science Data Cloud (IaaS)



- Virtual Machine (VM) containing common applications & pipelines
- Tukey (OSDC portal & middleware v0.3)
- Yates (infrastructure automation and management v0.1)



Third party open source software

+

Tukey

Yates



OPEN CLOUD CONSORTIUM

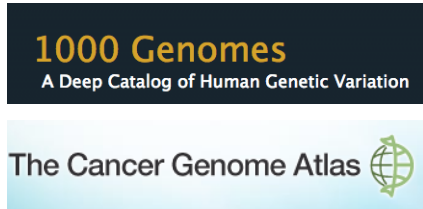
+

Open source software developed by the OCC and open standards



Data center

+



Data with permissions

+



Authorization of users access to data

+



OpenFISMA

Policies, procedures, controls, etc.

+



Governance, legal agreements

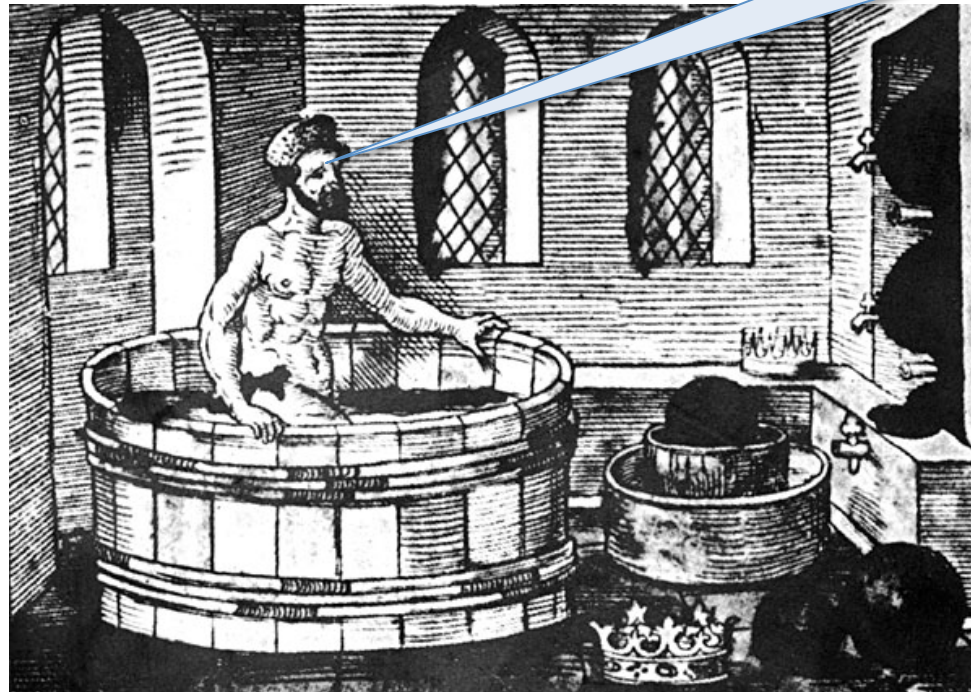
+



Sustainability model

Discoveries

Team: you
and your
colleagues



+ correlation
algorithms

+ Instrument:
3000 cores /
5 PB OSDC
science cloud

+

Data: 1 PB of OSDC
data across several
disciplines

Public Data Sets

Repository for public data sets of scientific interest, hosted on the OSDC.

The data sets below can be downloaded over the internet or high performance networks such as Internet2, as well as computed over directly on the OSDC. Currently, the OSDC hosts about 600 TB of data and the plan is to steadily increase this to the petabyte level. If you have suggestions about data that should be included, please let us know at info@opencloudconsortium.org.

1000 Genomes Project

Human sequence data from populations around the world with the goal of cataloging human genetic variation.

Total Size: 376.9TB

Categories: [genomics](#), [biology](#)

Last Modified: June 4, 2013, 3:30 p.m. UTC

City of Chicago Public Datasets

Public data made available by the City of Chicago

Total Size: 9.7GB

Categories: [social science](#)

Last Modified: Oct. 25, 2012, 2:03 p.m. UTC

Complete Genomics Public Data

Whole human genome sequence data sets provided by Complete Genomics, containing 69 standard, non-diseased samples as well as two matched tumor and normal sample pairs.

Total Size: 47.1TB

Categories: [genomics](#), [biology](#)

Last Modified: June 4, 2013, 3:30 p.m. UTC

Earth Observing-1 Mission

Data gathered by the Advanced Land Imager (ALI) Hyperspectral Imager (Hyperion) instruments on NASA's Earth Observing-1 Mission (EO-1) satellite.

Total Size: 49TB

Categories: [earth science](#), [satellite imagery](#)

Last Modified: April 24, 2013, 6:56 p.m. UTC

Enron Emails

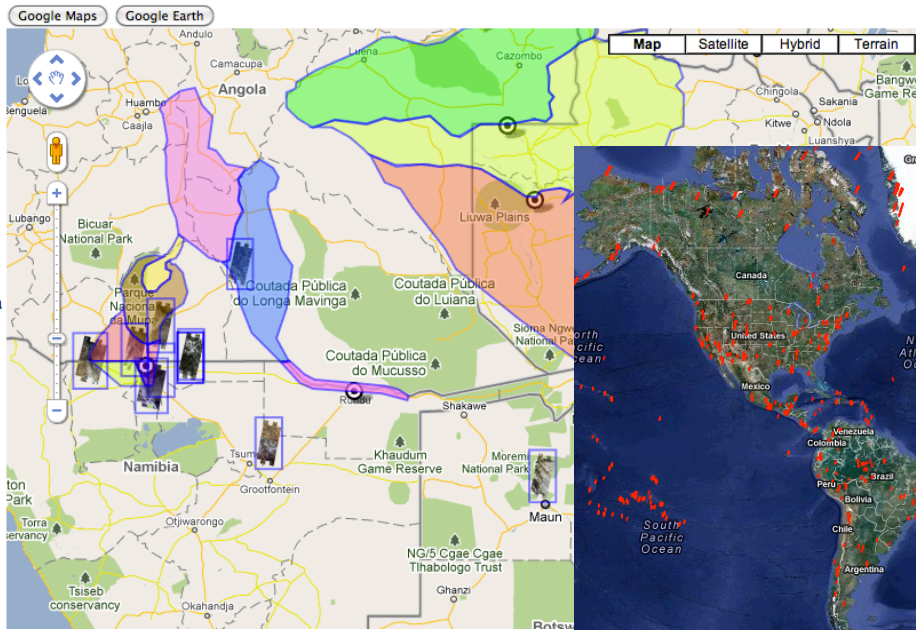
OSDC Public Data Sets

- Over 800 TB of open access data in the OSDC
- Earth sciences data
- Biological sciences data
- Social sciences data
- Digital humanities

OSDC Working Groups

Matsu Working Group: Clouds to Support Earth Science

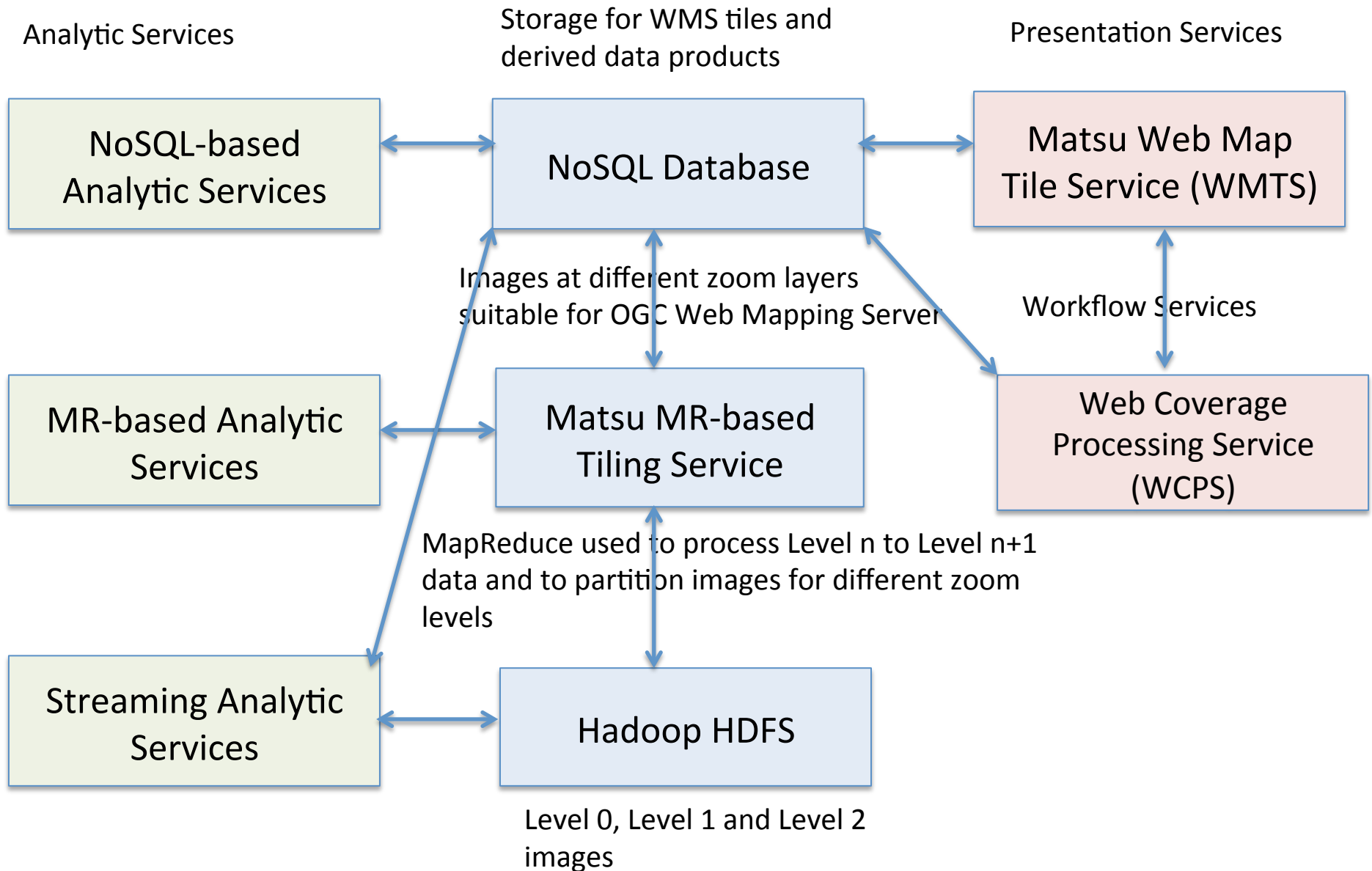
- (Experimental DB)
- Satellite Overlays**
 - EO1 ALI
 - SAR (SRI/Ukraine)
 - ASAR WSM 30-Jan-2010 (SRI/Ukraine)
 - Landsat-5 26-Jan-2010 (SRI/Ukraine)
 - Experimental EO-1
- Flood Classification Product**
- Ground Pics**
 - GeoTagged Images Jan 2010 (SRI/Ukraine)
- Kavango Radarsat Data**
 - March 15, 2011
 - March 17, 2011
 - March 19, 2011
 - March 21, 2011
 - March 22, 2011
 - March 26, 2011
- Cuvelai Radarsat Data**
 - March 18, 2011
 - March 24, 2011 - North
 - March 24, 2011 - South
 - April 1, 2011 - North
 - April 1, 2011 - South
- TRMM Flood Potential**



OPEN CLOUD CONSORTIUM

matsu.opensciencedatacloud.org

Matsu Architecture



Bionimbus Working Group



[OSDC](#) [Console](#) [Apply](#) [Public Data](#) [Status](#) [Projects](#) [Support](#) [News/PIRE](#)

Open Science Data Cloud

Bionimbus

A cloud-based infrastructure for managing, analyzing and sharing genomics datasets.

Bionimbus

Bionimbus is a collaboration between the Institute for Genomics and Systems Biology (IGSB) at the University of Chicago and the Open Science Data Cloud to develop open source technology for managing, analyzing, transporting, and sharing large genomics datasets in a secure and compliant fashion.

Version 2 of Bionimbus is used by the IGSB and their collaborators to manage data from the IGSB's High-Throughput Genome Analysis Core (HGAC).
Version 3 of Bionimbus is beta and uses some of the OSDC's common core services.

Bionimbus Resources

- Bionimbus (v2) [IGSB Private Cloud](#)
- Bionimbus (v3 alpha) [OSDC Community Cloud](#)
- Bionimbus (v3 alpha) [IGSB Private Cloud](#)

Bionimbus Support

- Bionimbus (v3) [tutorials](#)
- Bionimbus [XLDB 12](#) Talk

bionimbus.opensciencedatacloud.org (biological data)

Bionimbus Protected Data Cloud

PDC Console Apply Status

BIONIMBUS PROTECTED DATA CLOUD

Secure cloud services for the scientific community

What is the Bionimbus PDC?

The Bionimbus Protected Data Cloud (PDC) is a collaboration between the Open Science Data Cloud (OSDC) and the IGSB (IGSB,) the Center for Research Informatics (CRI), the Institute for Translational Medicine (ITM), and the University of Chicago Comprehensive Cancer Center (UCCCC).

The PDC allows users authorized by NIH to compute over human genomic data from dbGaP in a secure compliant fashion. Currently, selected datasets from the The Cancer Genome Atlas (TCGA) are available in the PDC.

How can I get involved?

- Apply for an Bionimbus PDC account and use the Bionimbus PDC to manage, analyze and share your data.
- Partner with us and add your own racks to the Bionimbus PDC (we will manage them for you).
- Help us develop the open source Bionimbus PDC software stack.

You can contact us at info@opencloudconsortium.org.

How do I get started?

First, apply for an account. Once your account is approved, you can login to the console and get started. Support questions can be directed to support@opencloudconsortium.org.

Apply for the PDC Now

Login to the PDC Console



OpenFlow-Enabled Hadoop WG

- When running Hadoop some map and reduce jobs take significantly longer than others.
- These are stragglers and can significantly slow down a MapReduce computation.
- Stragglers are common (dirty secret about Hadoop)
- Infoblox and UChicago are leading a OCC Working Group on OpenFlow-enabled Hadoop that will provide additional bandwidth to stragglers.
- We have a testbed for a wide area version of this project.

OSDC PIRE Project



We select OSDC PIRE Fellows (US citizens or permanent residents):

- We give them tutorials and training on big data science.
- We provide them fellowships to work with OSDC international partners.
- We give them preferred access to the OSDC.

Nominate your favorite scientist as an OSDC PIRE Fellow.

www.opensciencedatacloud.org (look for PIRE)

OSDC Software Stack

Tukey



[Console](#) [Apply](#) [Public Data](#) [Status](#) [Projects](#) [Support](#) [News / PIRE](#)

Key Service

[Archival Resource Keys \(ARK\)](#) are provided for data hosted on the OSDC by using the Open Cloud Consortium's Name Assigning Authority Number (NAAN). An ARK provides a persistent identifier for locating data and associated metadata. For example, the ARK for the Earth Observing-1 Mission is [ark:/31807/opd2](#).

The location of this data can be found by using the OSDC key service. This is done by either directly using the URL: <http://www.opensciencedatacloud.org/keyservice/ark:/31807/opd2> or by using the form below.

Locate data from an OCC ARK Key (e.g. [ark:/31807/opd2](#)):

A resource of the  OPEN CLOUD CONSORTIUM and made possible by our [sponsors](#).

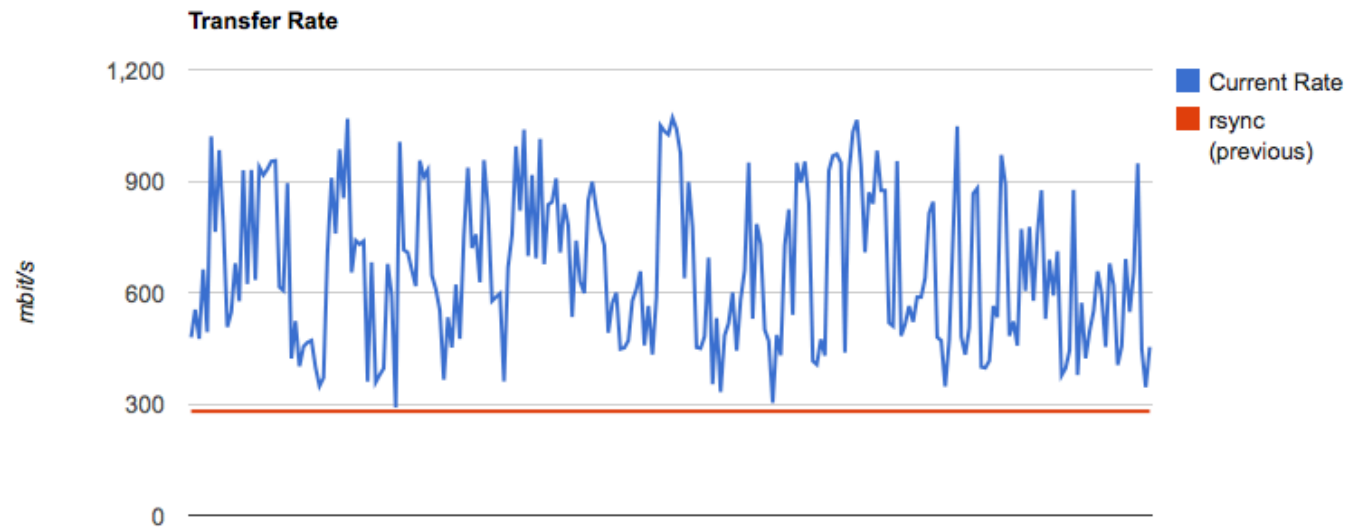
- Tukey (based in part on Horizon) v0.3.0
- We have factored out digital ID service, file sharing, and transport from Bionimbus and Matsu.

Yates

- Automation installation of OSDC software stack on rack of computers.
- Based upon Chef
- Version 0.3.0



UDR



- UDT is a high performance network transport protocol (v0.9.4)
- $UDR = rsync + UDT$
- It is easy for an average systems administrator to keep 100's of TB of distributed data synchronized.
- We are using it to distribute c. 1 PB from the OSDC

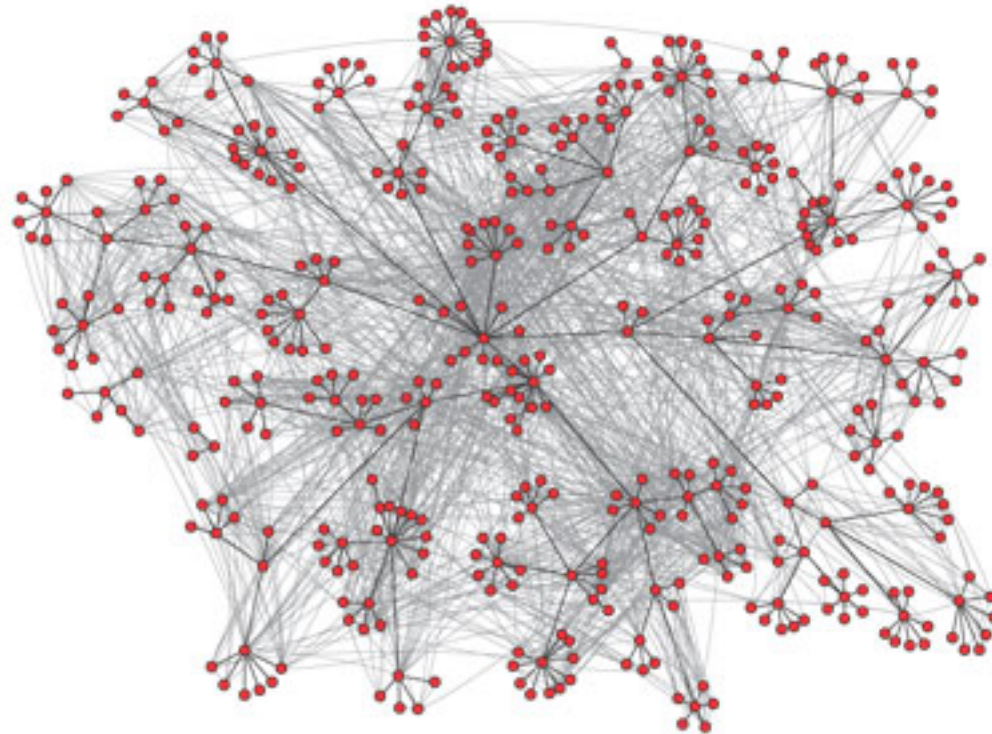
UDT Tuning: Buffers and CPUs

Configuration Change	Observed Transfer Rate	Time to Transfer 1 TB (minutes)
UDT and Linux Defaults	1.6 Gbps	85
Setting buffers sizes to 64 MB	3.3 Gbps	41
Improved CPU on sending side with processor affinity	3.7 Gbps	36
Improved CPU on receiving side with processor affinity	4.6 Gbps	29
Improved CPU on both sides with processor affinity on both sides	6.3 Gbps	21
Turn off CPU frequency scaling and set to max clock speed	6.7 Gbps	20

Open Science Data Cloud Services

- Digital ID services
- Data sharing services
- Data transport services (UDR)
- What other core services are *essential*?
- Of course, working groups and applications always add their own services
- These core services will hopefully make the OSDC attractive as a platform (PaaS) for scientific discovery.

Part 5: Analyzing Data at the Scale of a Data Center



Source: Jon Kleinberg, Cornell University, www.cs.cornell.edu/home/kleinber/networks-book/

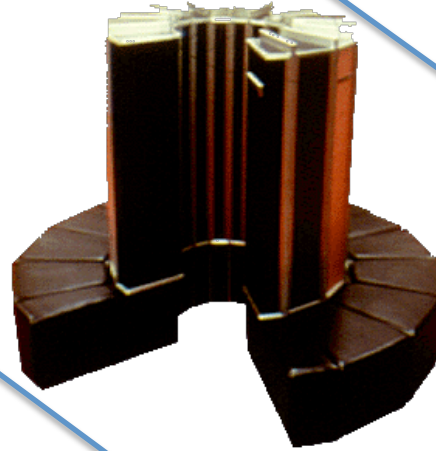


2004
10x-100x



data science
(big data biology,
medicine)

1976
10x-100x



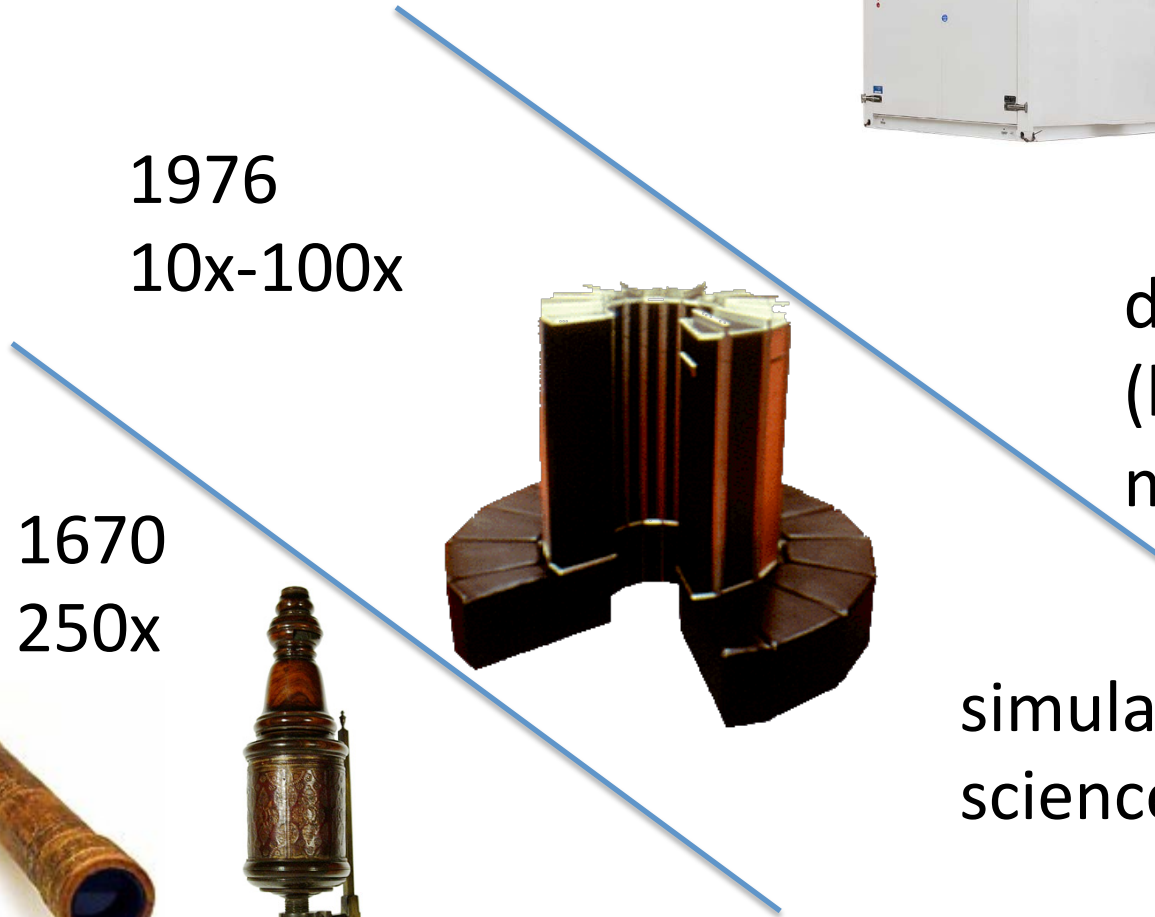
simulation
science

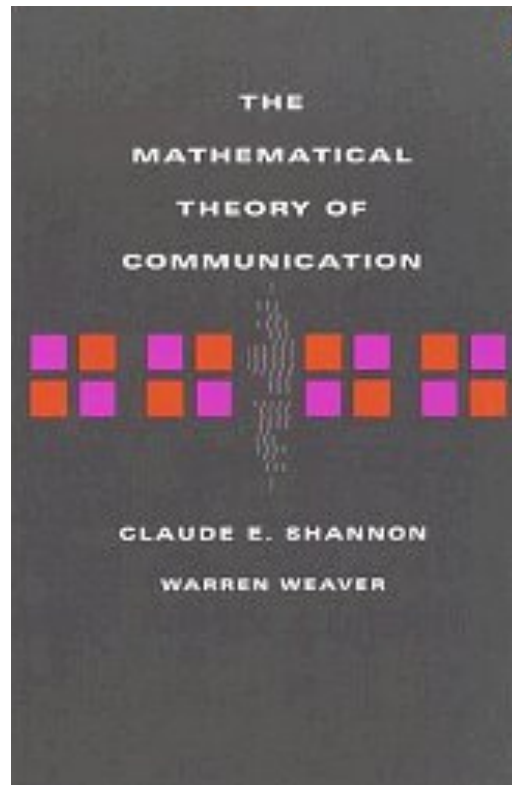
1670
250x



experimental
science

1609
30x





What are the foundations for data science?

Is More Different? Do New Phenomena Emerge at Scale in Biomedical Data?

4 August 1972, Volume 177, Number 4047

SCIENCE

More Is Different

Broken symmetry and the nature of the hierarchical structure of science.

P. W. Anderson

The reductionist hypothesis may still be a topic for controversy among philosophers, but among the great majority of active scientists I think it is accepted without question. The workings of our minds and bodies, and of all the animate or inanimate matter of which we

planation of phenomena in terms of known fundamental laws. As always, distinctions of this kind are not unambiguous, but they are clear in most cases. Solid state physics, plasma physics, and perhaps also biology are extensive. High energy physics and a good part of nuclear physics are intensive. There is always much less intensive research going on than extensive.

less relevance they seem to have to the very real problems of the rest of science, much less to those of society.

The constructionist hypothesis breaks down when confronted with the twin difficulties of scale and complexity. The behavior of large and complex aggregates of elementary particles, it turns out, is not to be understood in terms of a simple extrapolation of the properties of a few particles. Instead, at each level of complexity entirely new properties appear, and the understanding of the new behaviors requires research which I think is as fundamental in its nature as any other. That is, it seems to me that one may array the sciences roughly linearly in a hierarchy, according to the idea: The elementary entities of science X obey the laws of science Y.

X	Y
solid state or many-body physics	elementary particle physics



Characteristic	Colectomy		Gastrectomy	
	Patients (%)	Mortality rate (%)	Patients (%)	Mortality rate (%)
Age > 65	28,243 (58.2)	6.7*	3,482 (54.1)	2.5*
Female gender	26,257 (54.1)	4.8 [†]	2,987 (46.4)	8.4
African American	4,553 (9.4)	5.0	915 (14.2)	8.5
Medicaid	2,843 (5.9)	4.7	659 (10.2)	6.2 [‡]
IHD	7,235 (14.9)	8.6*	942 (14.6)	13.6*
Airway obstruction	1,782 (3.7)	3.7	271 (4.2)	7.0
CHD	4,335 (8.9)	16.8*	613 (9.5)	24.5*
Metastasis	5,953 (12.3)	6.6*	1,099 (17.1)	9.0
PVD	265 (0.6)	18.9*	46 (0.7)	21.7*
COPD	4,004 (8.2)	9.9*	556 (8.6)	16.4*
Diabetes	6,703 (13.8)	6.2*	975 (15.2)	9.0
Dysrhythmia	6,464 (13.3)	14.7*	987 (15.3)	22.9*
All patients	48,582 (100)	4.6	6,434 (100)	8.4

CHD indicates congestive heart disease; COPD, chronic obstructive pulmonary disease; IHD, ischemic heart disease; PVD, peripheral vascular disease.



Small data

Medium data

GB

TB

PB



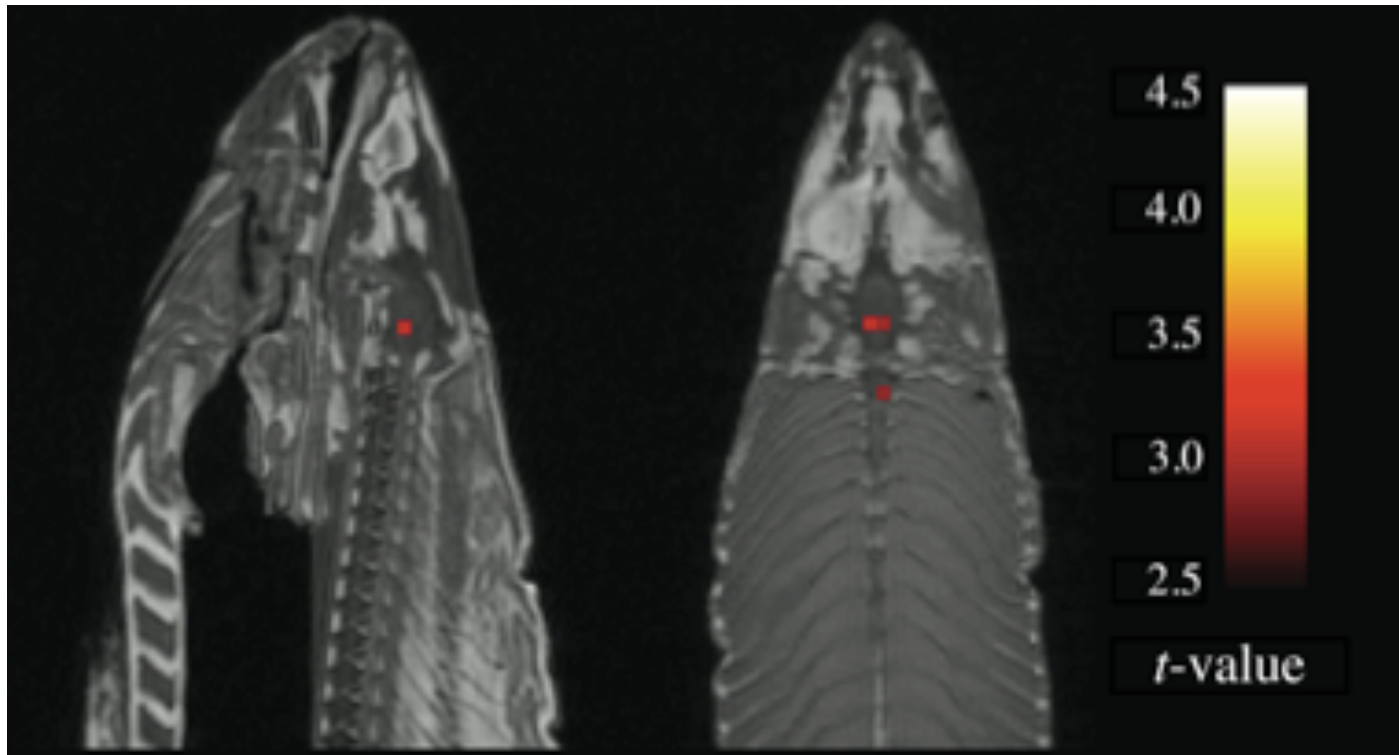
W

KW

MW

Complex models over small data that are highly manual.

Simpler models over large data that are highly automated.



Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed. Out of a search volume of 8064 voxels a total of 16 voxels were significant.

Craig M. Bennett, Abigail A. Baird, Michael B. Miller, and George L. Wolford, Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction, retrieved from <http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf>.

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,^{1,2*} Ryan Kennedy,^{1,3,4} Gary King,³ Alessandro Vespignani^{5,6,3}

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (1, 2). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (3, 4), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can

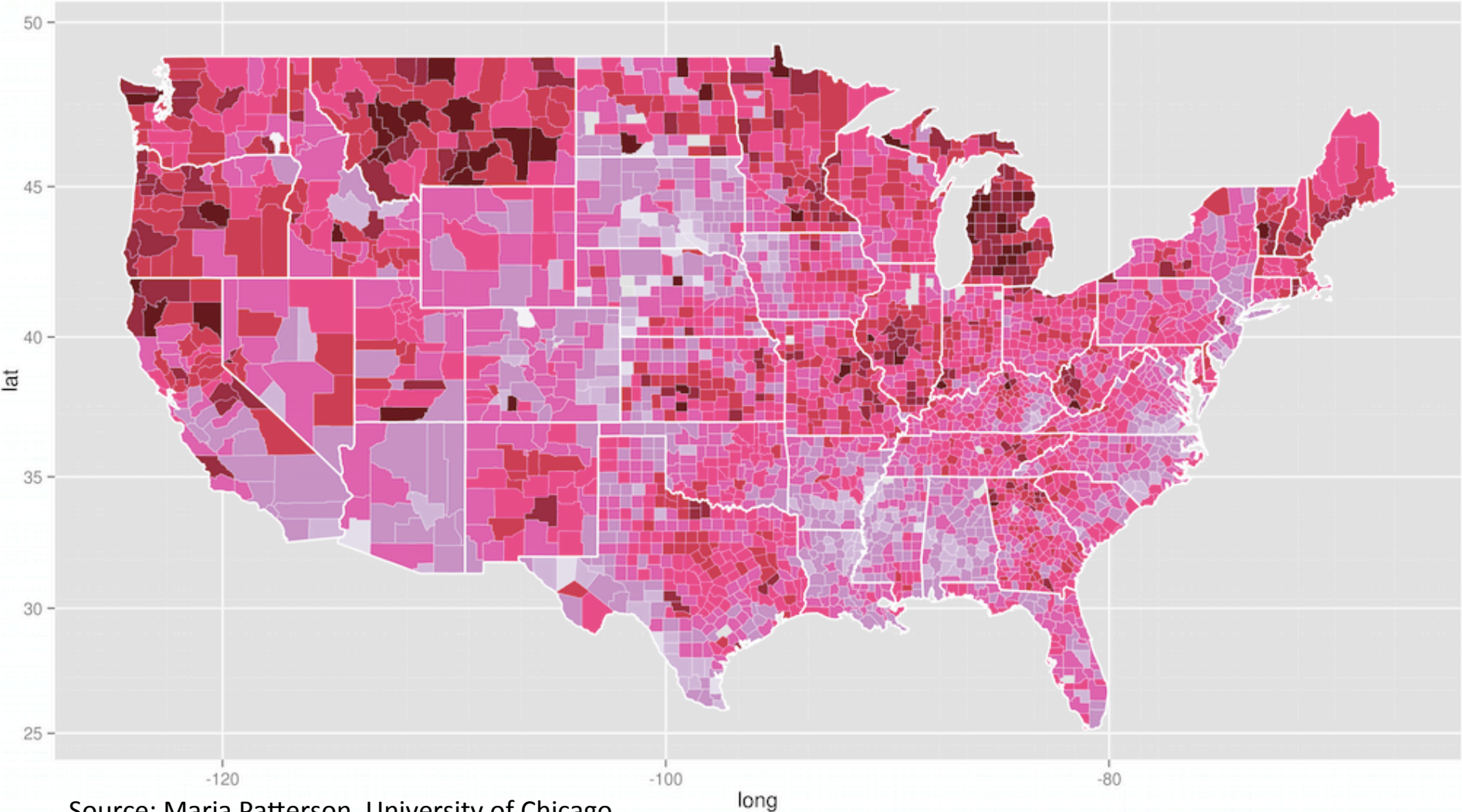


Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

run ever since, with a few changes announced in October 2013 (10, 15).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.

Environmental Factors and Cancer



Source: Maria Patterson, University of Chicago.

Building Models over Big Data

- We know about the “unreasonable effectiveness of ensemble models.” Building ensembles of models over computer clusters works well ...
- ... but, how do machine learning algorithms scale to data center scale science?
- Ensembles of random trees built from templates appear to work better than traditional ensembles of classifiers
- The challenge is often decomposing large heterogeneous datasets into homogeneous components that can be modeled.

Source: Wenxuan Gao, Robert Grossman, Philip Yu, and Yunhong Gu, Why Naive Ensembles Do Not Work in Cloud Computing, Proceedings of the The First Workshop on Large-scale Data Mining: Theory and Applications (LDMTA 2009), 2009.

New Questions

- How would research be impacted if we could analyze *all of the data* each evening?
- How would health care be impacted if we could *analyze of the data* each evening?

Part 6

Key Questions for This Workshop



- Question 1. How can we add partner sites at other locations that extend the OSDC? In particular, how can we *extend* the OSDC to sites around the world? How can the OSDC *interoperate* with other science clouds?
- Question 2. What data can we add to the OSDC to facilitate data intensive cross-disciplinary discoveries?
- Question 3. How can we build a plugin structure so that Tukey can be extended by other users and by other communities?
- Question 4. What tools and applications can we add to the OSDC facilitate data intensive cross-disciplinary discoveries?
- Question 5. How can we better integrate digital IDs and file sharing services into the OSDC?
- Question 6. What are 3-5 grand challenge questions that leverage the OSDC?

Questions



Robert Grossman is a faculty member at the University of Chicago. He is the Chief Research Informatics Officer for the Biological Sciences Division, the Director of the Center for Data Intensive Science (CDIS), a Faculty Member and Senior Fellow at the Computation Institute and the Institute for Genomics and Systems Biology, and a Professor of Medicine in the Section of Genetic Medicine. His research group focuses on big data, data science, biomedical informatics, cloud computing, and related areas.

He is also the Founder and a Partner of Open Data Group, which has been building predictive models over big data for companies for over 12 years.

He recently wrote a book for the general reader that discusses big data (among other topics) called the Structure of Digital Computing: From Mainframes to Big Data, which can be purchased from Amazon.

He blogs occasionally about big data at rgrossman.com.

Major funding and support for the Open Science Data Cloud (OSDC) is provided by the Gordon and Betty Moore Foundation. This funding is used to support the OSDC-Adler, Sullivan and Root facilities.

Additional funding for the OSDC has been provided by the following sponsors:

- The Bionimbus Protected Data Cloud is supported in part by NIH/NCI through NIH/SAIC Contract 13XS021 / HHSN261200800001E.
- The OCC-Y Hadoop Cluster (approximately 1000 cores and 1 PB of storage) was donated by Yahoo! in 2011.
- Cisco provides the OSDC access to the Cisco C-Wave, which connects OSDC data centers with 10 Gbps wide area networks.
- The OSDC is supported by a 5-year (2010-2016) PIRE award (OISE – 1129076) to train scientists to use the OSDC and to further develop the underlying technology.
- OSDC technology for high performance data transport is supported in part by NSF Award 1127316.
- The StarLight Facility in Chicago enables the OSDC to connect to over 30 high performance research networks around the world at 10 Gbps or higher.
- Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, NIH or other funders of this research.

The OSDC is managed by the Open Cloud Consortium, a 501(c)(3) not-for-profit corporation. If you are interested in providing funding or donating equipment or services, please contact us at info@opensciencedatacloud.org.