The next 10 years of quantitative biology

Michael Schatz

March 25, 2014

Keystone Meeting on Big Data in Biology

@mike_schatz / #KSBigData

Unsolved Questions in Biology

- What is your genome sequence?
- How does your genome compare to my genome?

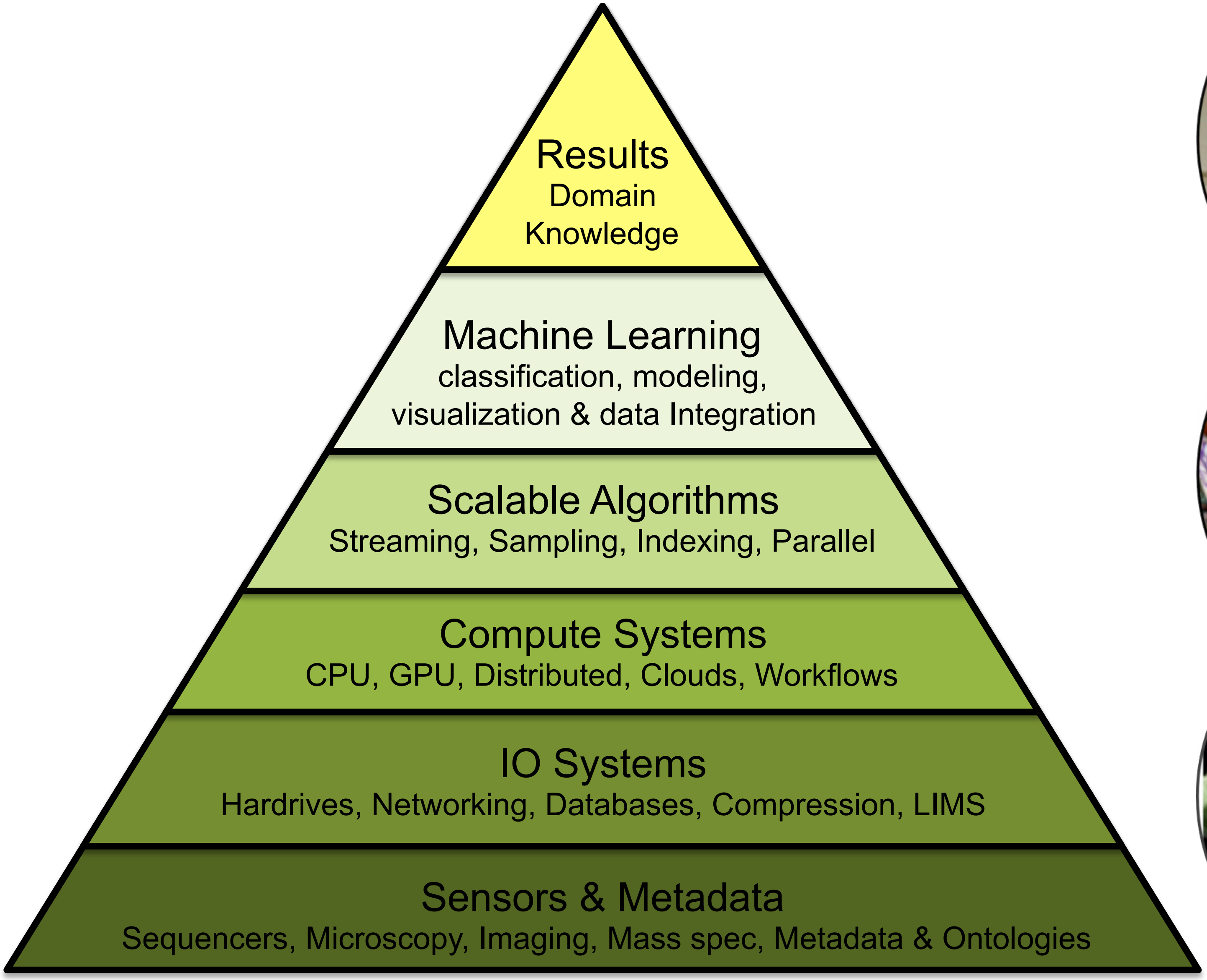
- Where are the genes and how active are they?
- How does gene activity change during development?
- How does splicing change during development?

- How does methylation change during development?
- How does chromatin change during development?
- How does is your genome folded in the cell?
- Where do proteins bind and regulate genes?

- What virus and microbes are living inside you?
- How do your mutations relate to disease?
- What drugs should we give you?

- Plus hundreds and hundreds more

Quantitative Biology Technologies



Results
Domain Knowledge

Machine Learning
classification, modeling, visualization & data Integration

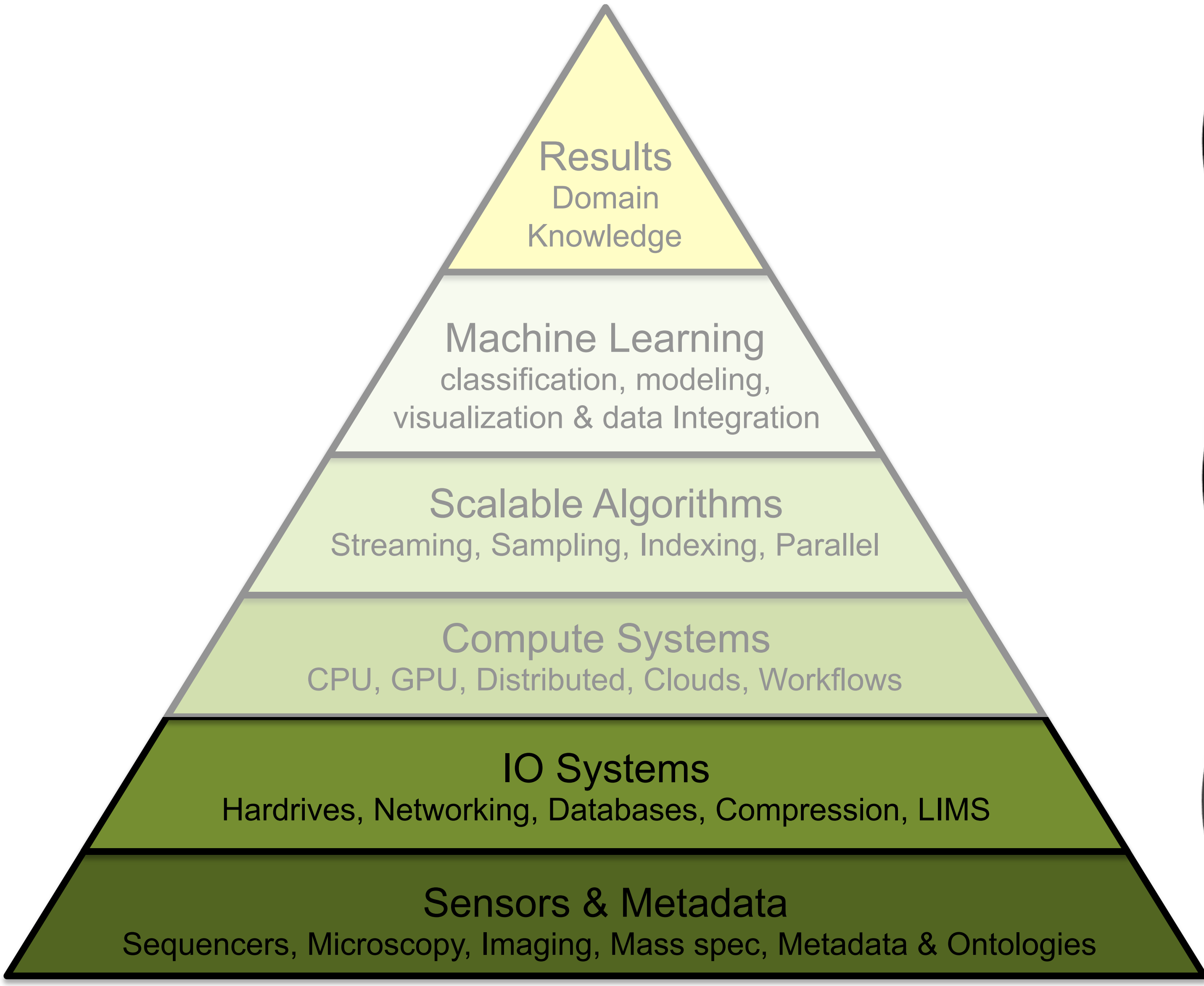
Scalable Algorithms
Streaming, Sampling, Indexing, Parallel

Compute Systems
CPU, GPU, Distributed, Clouds, Workflows

IO Systems
Hardrives, Networking, Databases, Compression, LIMS

Sensors & Metadata
Sequencers, Microscopy, Imaging, Mass spec, Metadata & Ontologies

Quantitative Biology Technologies



Results
Domain Knowledge

Machine Learning
classification, modeling, visualization & data Integration

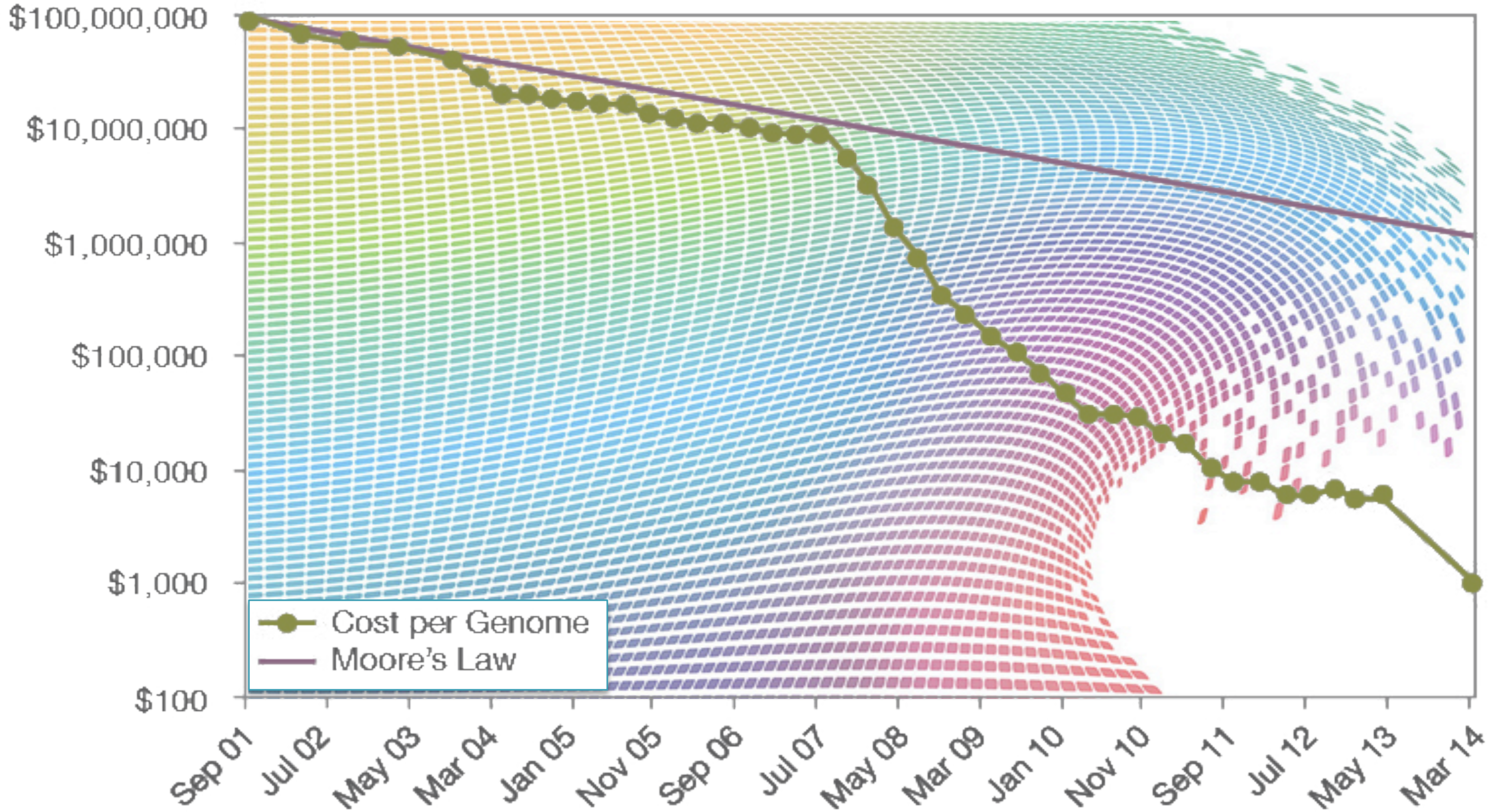
Scalable Algorithms
Streaming, Sampling, Indexing, Parallel

Compute Systems
CPU, GPU, Distributed, Clouds, Workflows

IO Systems
Hardrives, Networking, Databases, Compression, LIMS

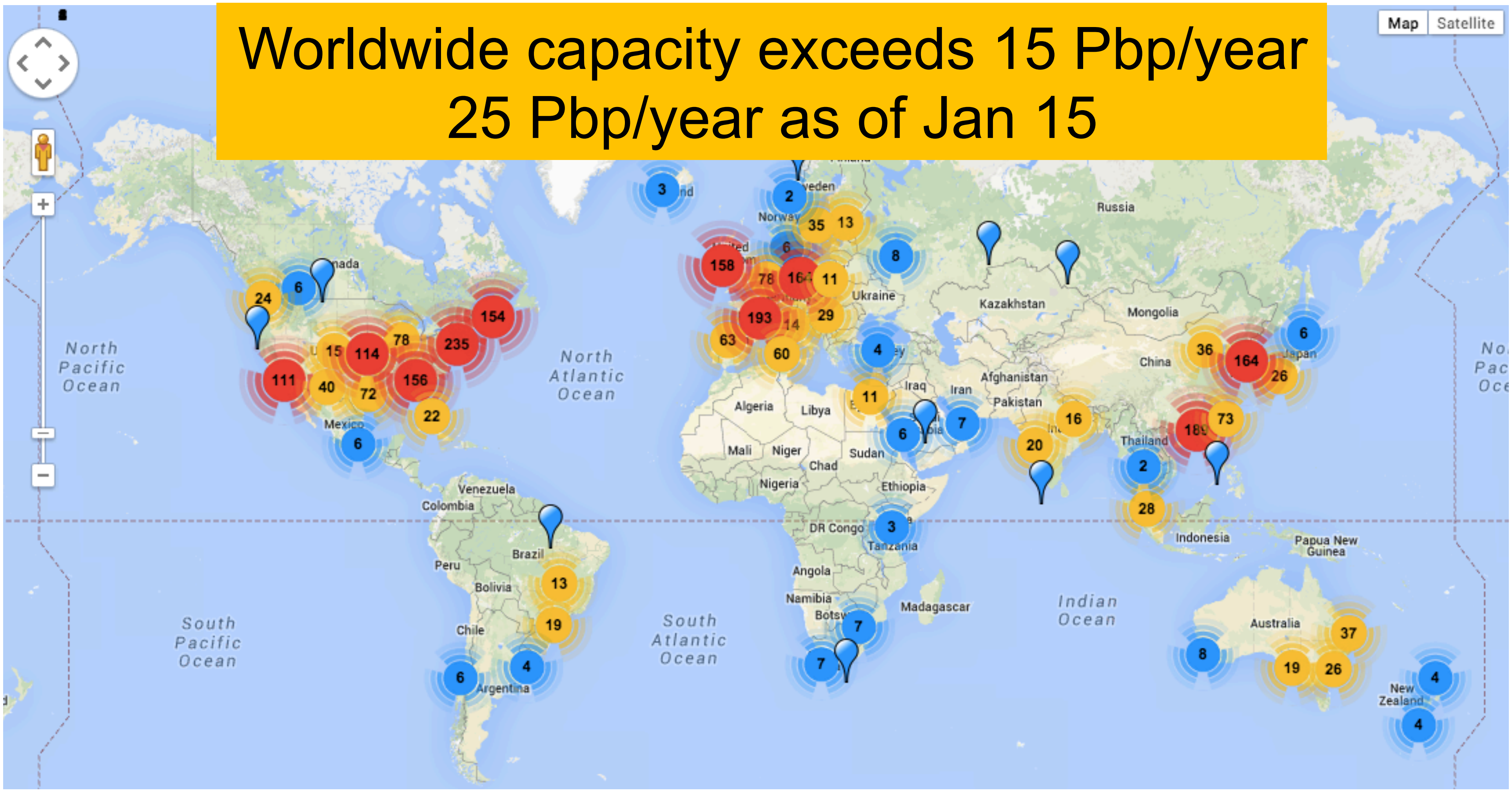
Sensors & Metadata
Sequencers, Microscopy, Imaging, Mass spec, Metadata & Ontologies

Cost per Genome



http://www.genome.gov/sequencingcosts/

Sequencing Centers



Worldwide capacity exceeds 15 Pbp/year
25 Pbp/year as of Jan 15

Next Generation Genomics: World Map of High-throughput Sequencers
http://omicsmaps.com

How much is a petabyte?

Unit	Size
Byte	1
Kilobyte	1,000
Megabyte	1,000,000
Gigabyte	1,000,000,000
Terabyte	1,000,000,000,000
Petabyte	1,000,000,000,000,000

*Technically a kilobyte is 2^{10} and a petabyte is 2^{50}

How much is a petabyte?

100 GB / Genome
4.7GB / DVD
~20 DVDs / Genome

X

10,000 Genomes

=

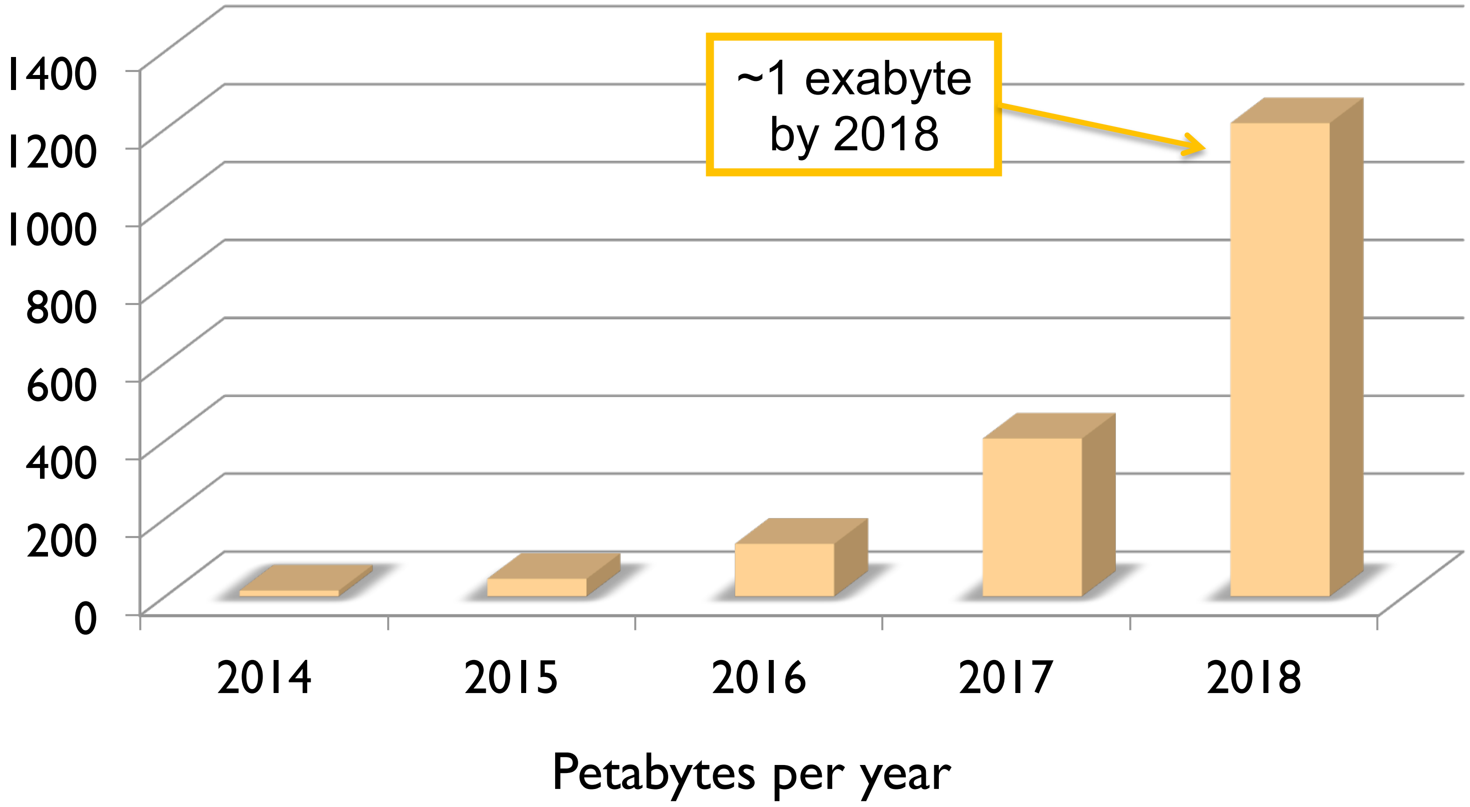
1PB Data
200,000 DVDs

787 feet of DVDs
~1/6 of a mile tall

500 2 TB drives
$500k

DNA Data Tsunami

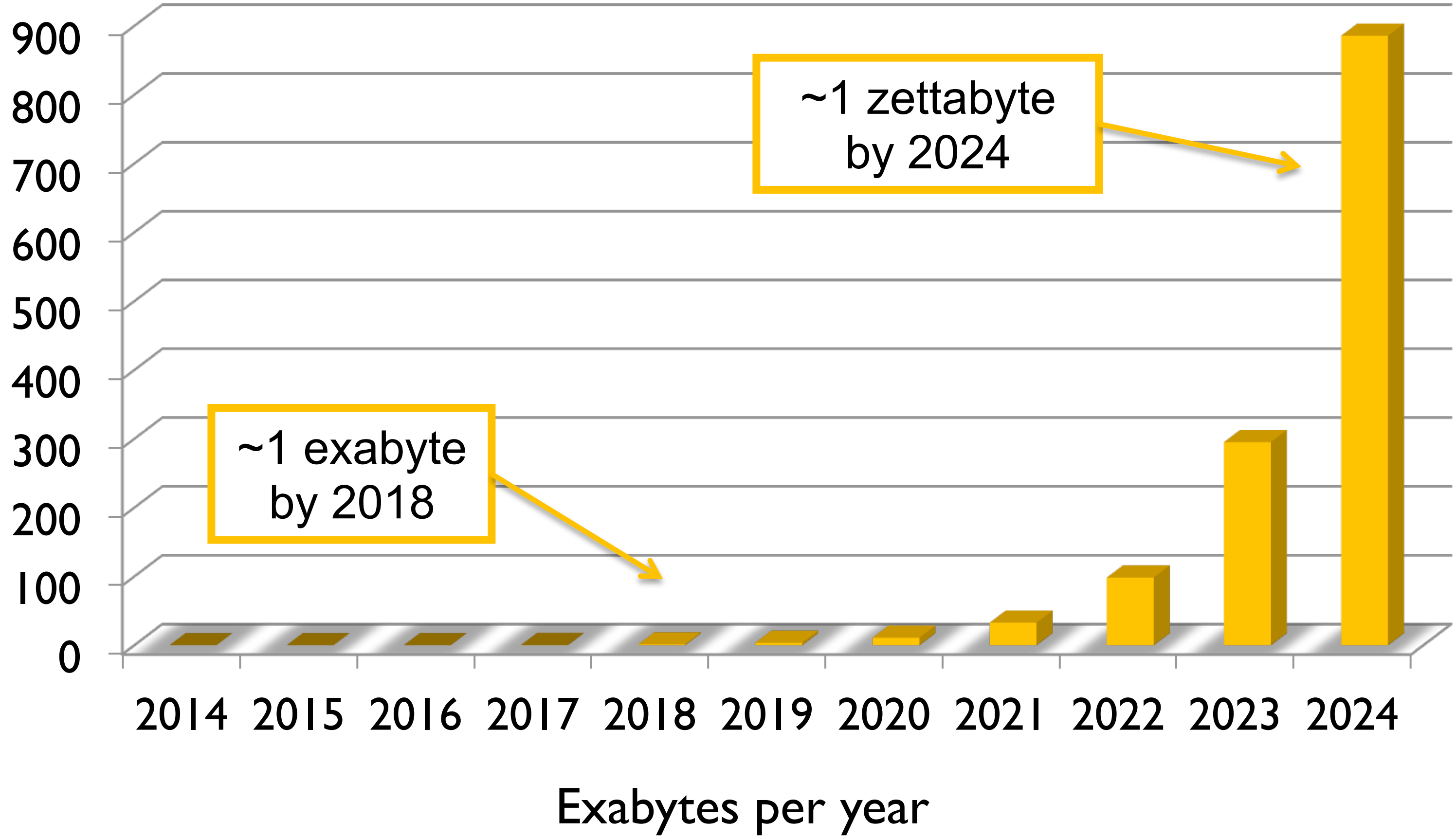
Current world-wide sequencing capacity is growing at ~3x per year!



~1 exabyte by 2018

DNA Data Tsunami

Current world-wide sequencing capacity is growing at ~3x per year!



~1 zettabyte by 2024

~1 exabyte by 2018

How much is a zettabyte?

Unit	Size
Byte	1
Kilobyte	1,000
Megabyte	1,000,000
Gigabyte	1,000,000,000
Terabyte	1,000,000,000,000
Petabyte	1,000,000,000,000,000
Exabyte	1,000,000,000,000,000,000
Zettabyte	1,000,000,000,000,000,000,000

How much is a zettabyte?

100 GB / Genome
4.7GB / DVD
~20 DVDs / Genome

X

10,000,000,000 Genomes

=

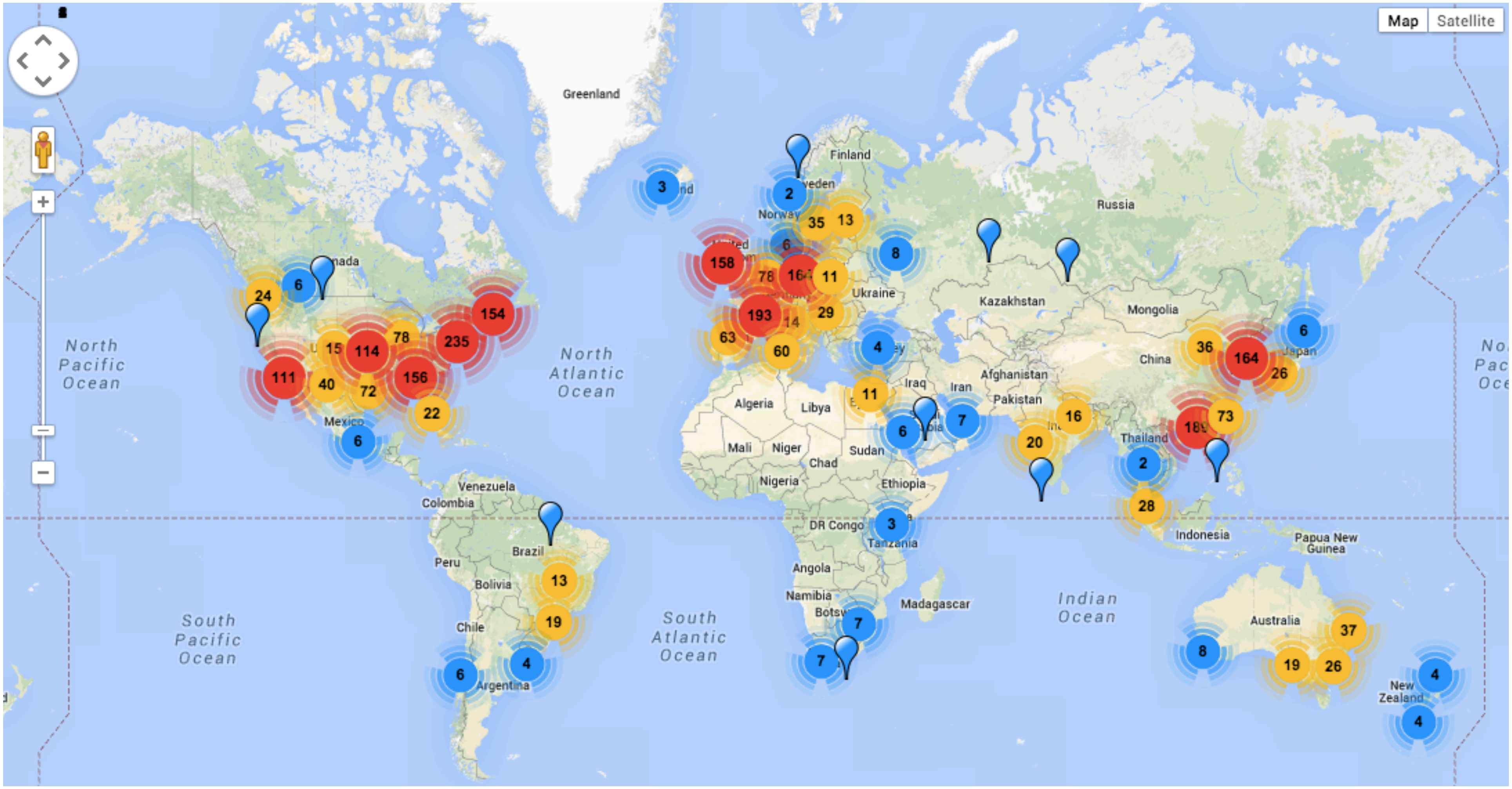
1ZB Data
200,000,000,000 DVDs

150,000 miles of DVDs
~ ½ distance to moon

YouTube

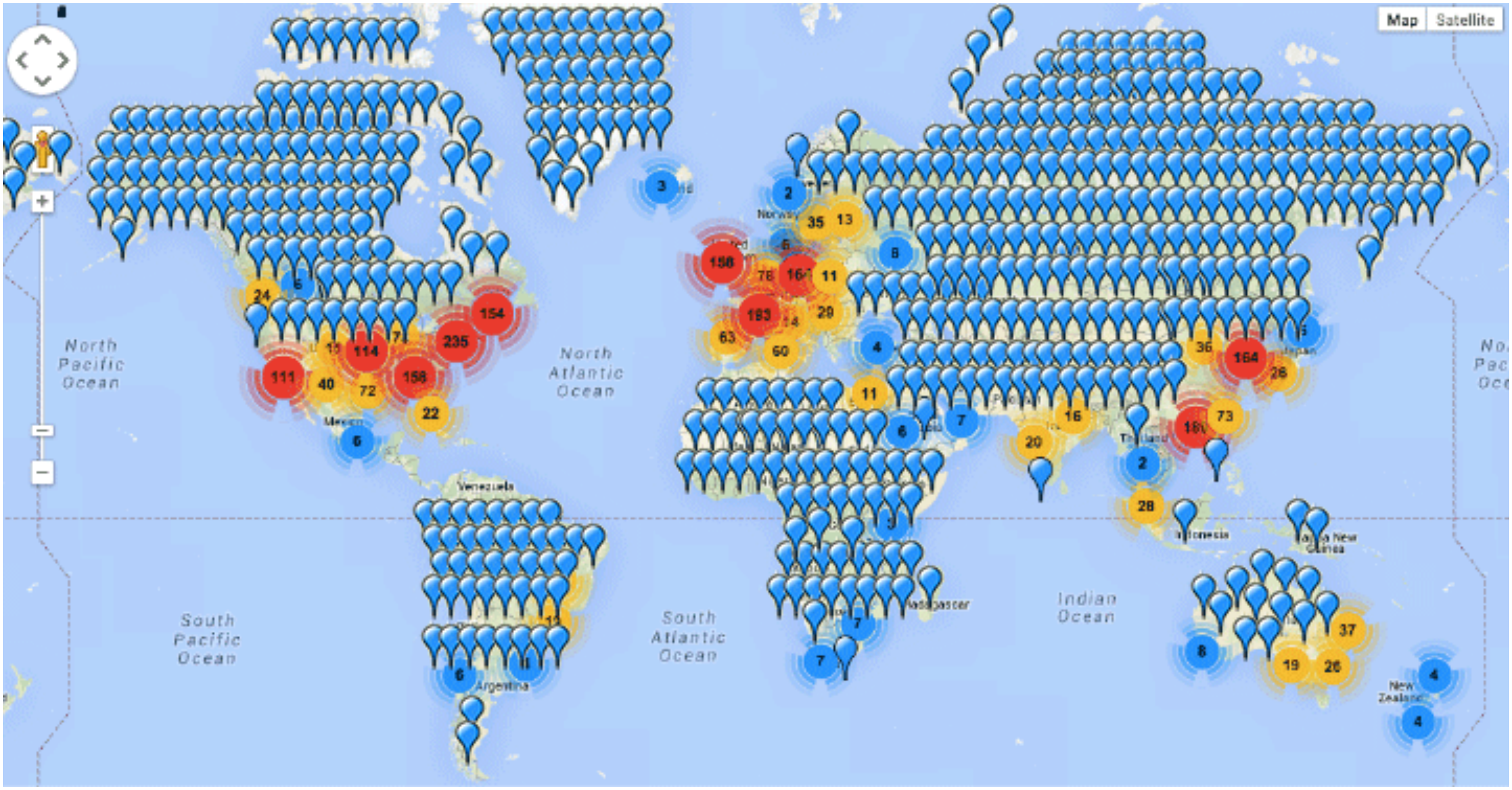
Both currently ~100Pb
But growing exponentially

Sequencing Centers



Next Generation Genomics: World Map of High-throughput Sequencers
http://omicsmaps.com

Sequencing Centers



Next Generation Genomics: World Map of High-throughput Sequencers
http://omicsmaps.com

Biological Sensor Network

(@ewanbirney)

(@latimes)

The rise of a digital immune system
Schatz, MC, Phillippy, AM (2012) GigaScience 1:4

Data Production & Collection

Expect massive growth to sequencing and other biological sensor data over the next 10 years

- Exascale biology is certain, zettascale on the horizon
- Compression helps, but need to aggressively throw out data
- Requires careful consideration of the "preciousness" of the sample

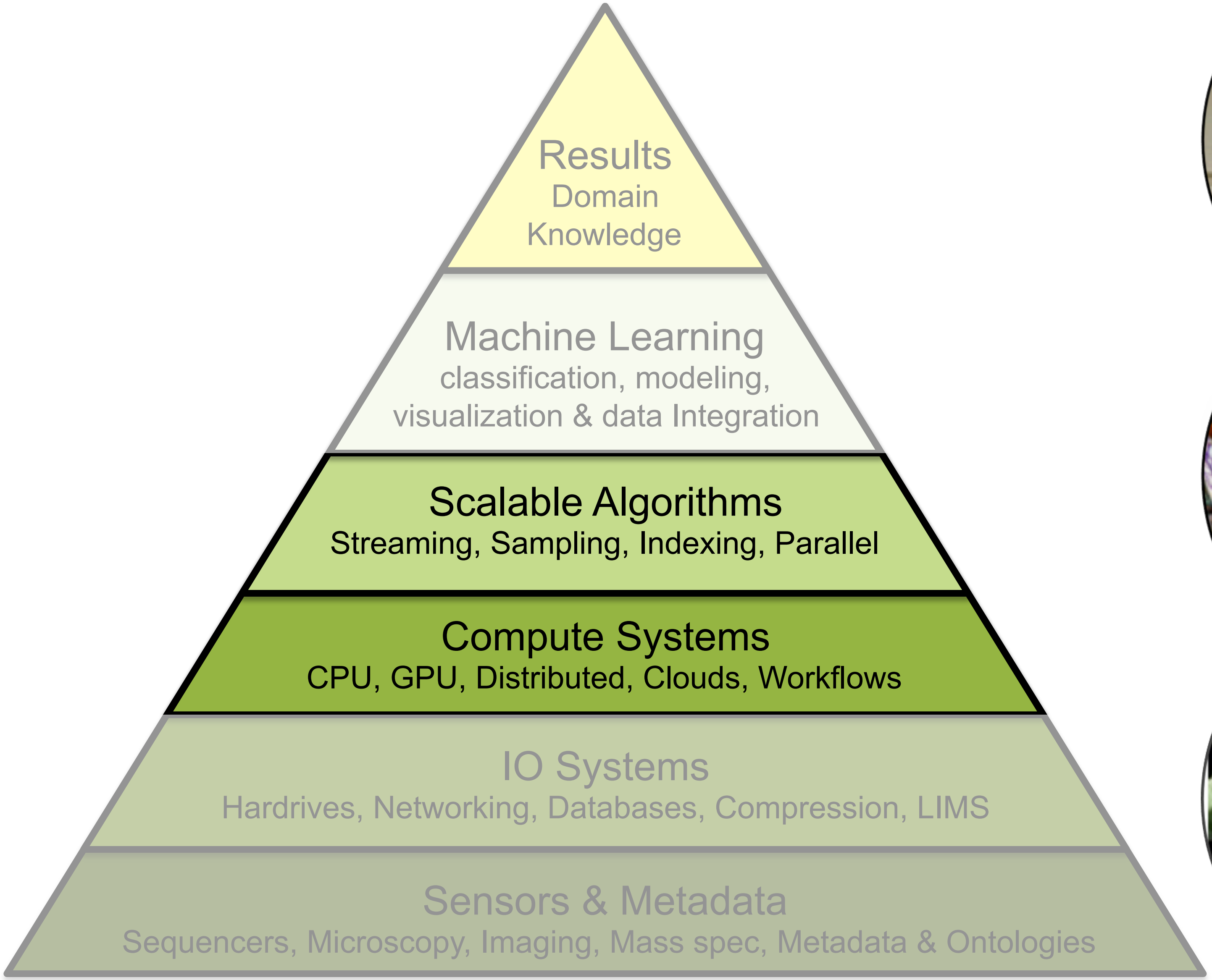
Major data producers concentrated in hospitals, universities, agricultural companies, research institutes

- Major efforts in human health and disease, agriculture, bioenergy

But also widely distributed mobile sensors

- Schools, offices, sports arenas, transportations centers, farms & food distribution centers
- Monitoring and surveillance, as ubiquitous as weather stations
- The rise of a digital immune system?

Quantitative Biology Technologies



Results
Domain Knowledge

Machine Learning
classification, modeling, visualization & data Integration

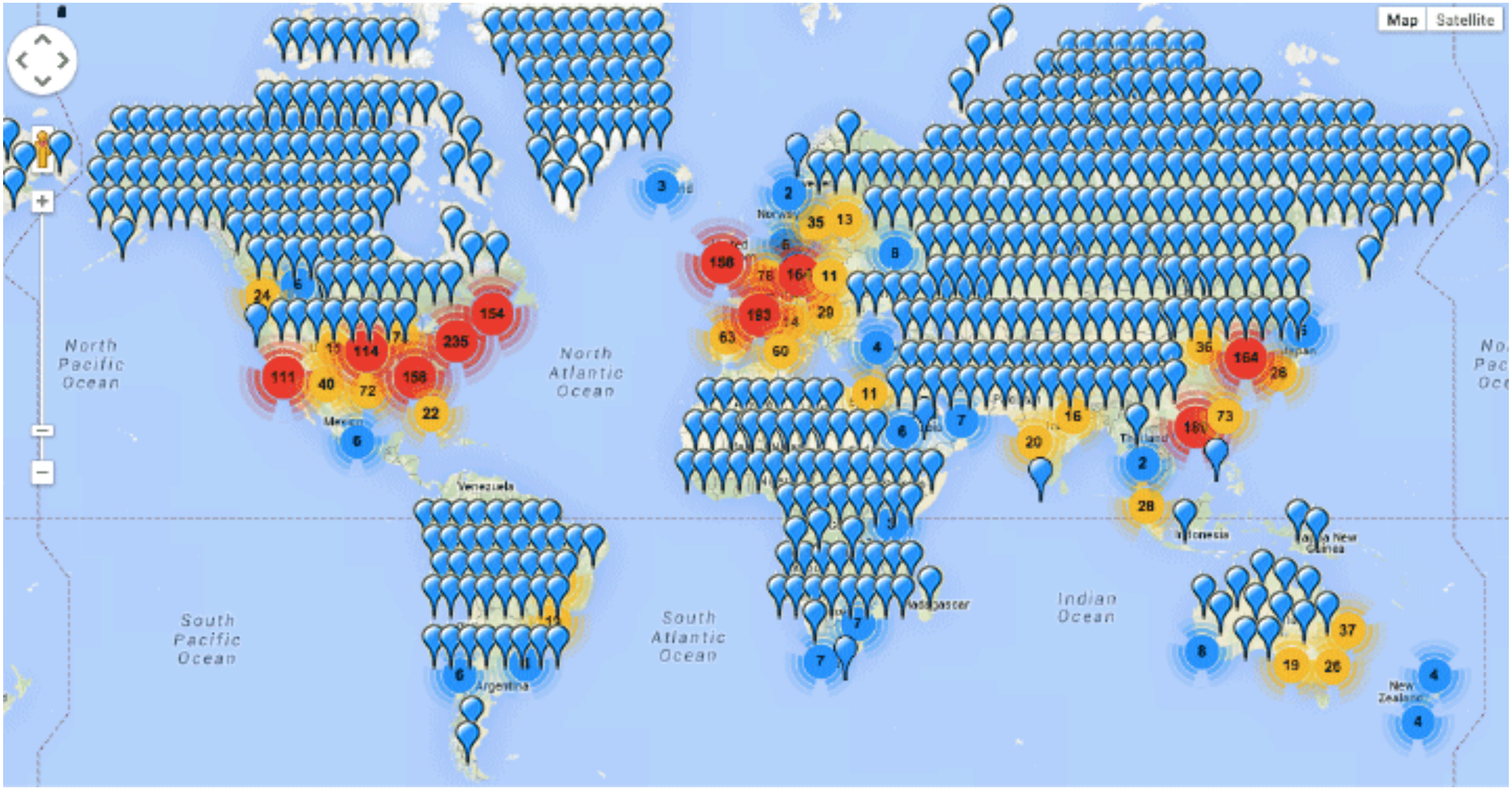
Scalable Algorithms
Streaming, Sampling, Indexing, Parallel

Compute Systems
CPU, GPU, Distributed, Clouds, Workflows

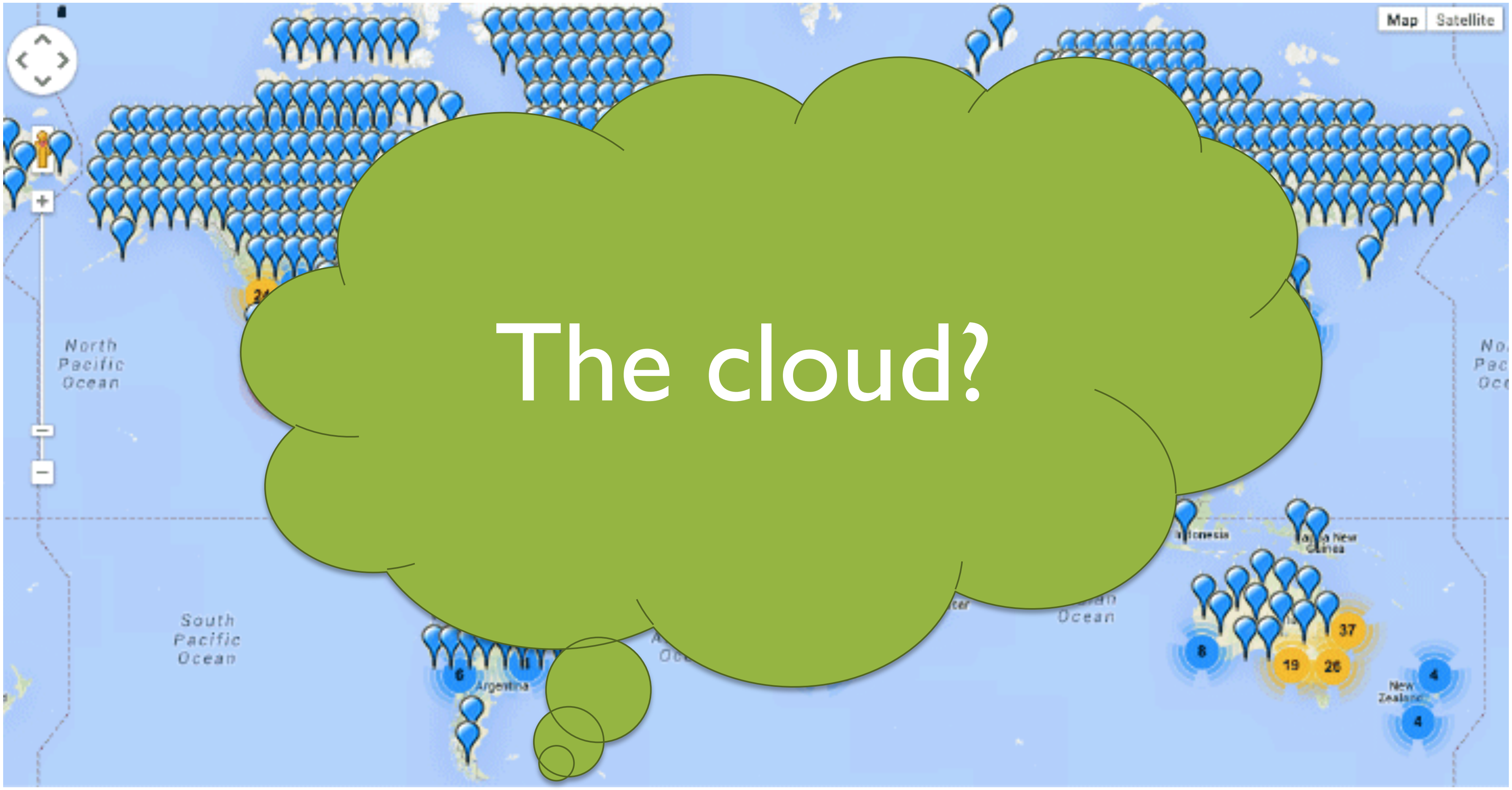
IO Systems
Hardrives, Networking, Databases, Compression, LIMS

Sensors & Metadata
Sequencers, Microscopy, Imaging, Mass spec, Metadata & Ontologies

Sequencing Centers



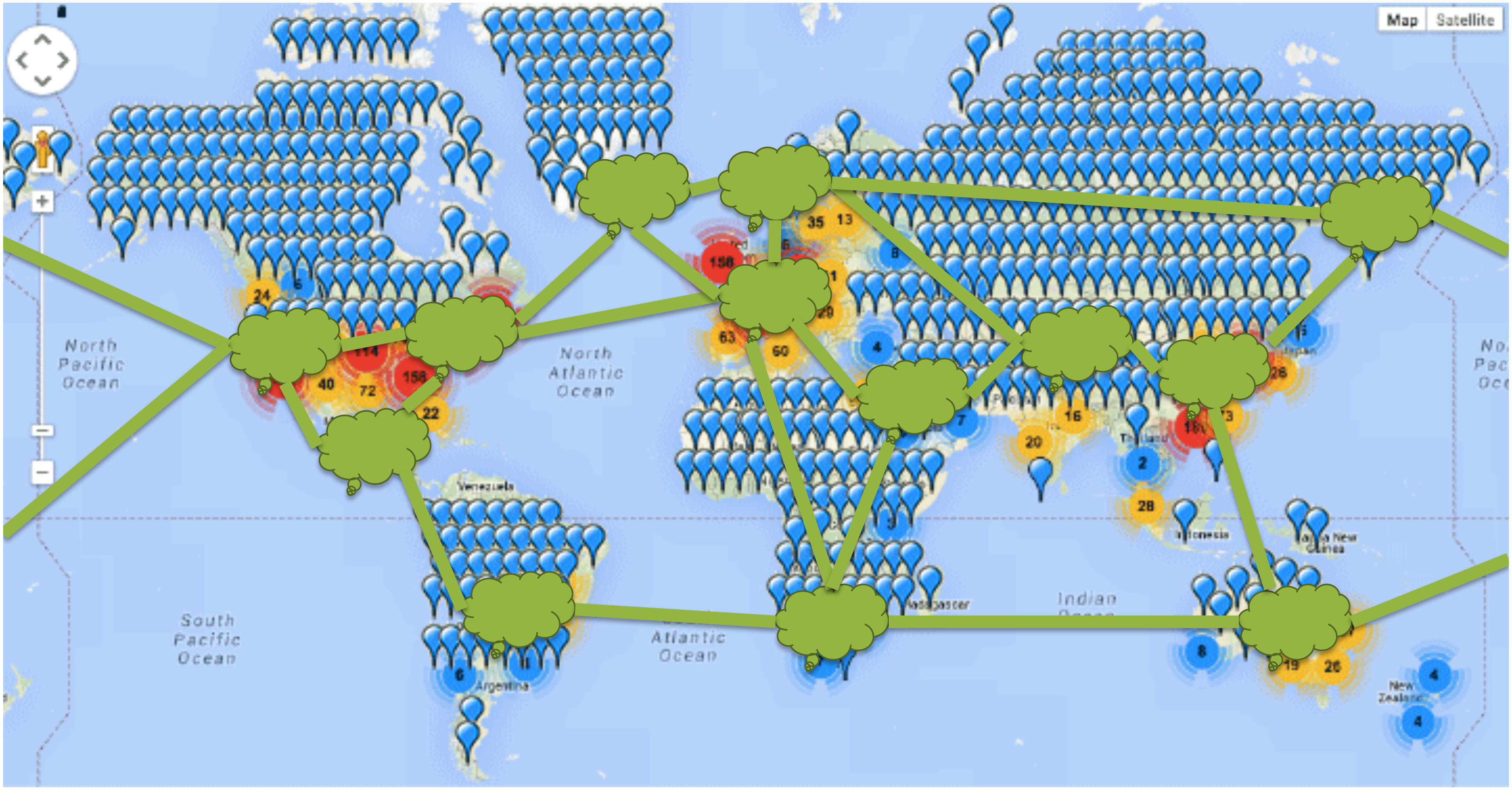
Informatics Centers



The cloud?

The DNA Data Deluge
Schatz, MC and Langmead, B (2013) *IEEE Spectrum*. July, 2013

Informatics Centers



The DNA Data Deluge
Schatz, MC and Langmead, B (2013) *IEEE Spectrum*. July, 2013

Parallel Algorithm Spectrum

Embarrassingly Parallel

Cluster Computing

Each item is Independent

Loosely Coupled

MapReduce

Independent-Sync-Independent

Tightly Coupled

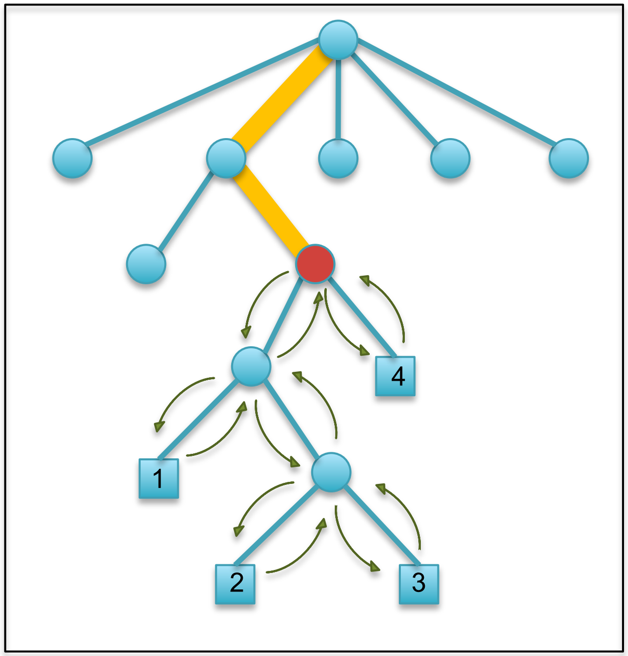
Graphs & MD simulations

Constant Sync

MUMmerGPU

http://mummergpu.sourceforge.net

- Index reference using a suffix tree
  - Each suffix represented by path from root
  - Reorder tree along space filling curve
- Map many reads simultaneously on GPU
  - Find matches by walking the tree
  - Find coordinates with depth first search
- Performance on nVidia GTX 8800
  - Match kernel was ~10x faster than CPU
  - Search kernel was ~4x faster than CPU
  - End-to-end runtime ~4x faster than CPU



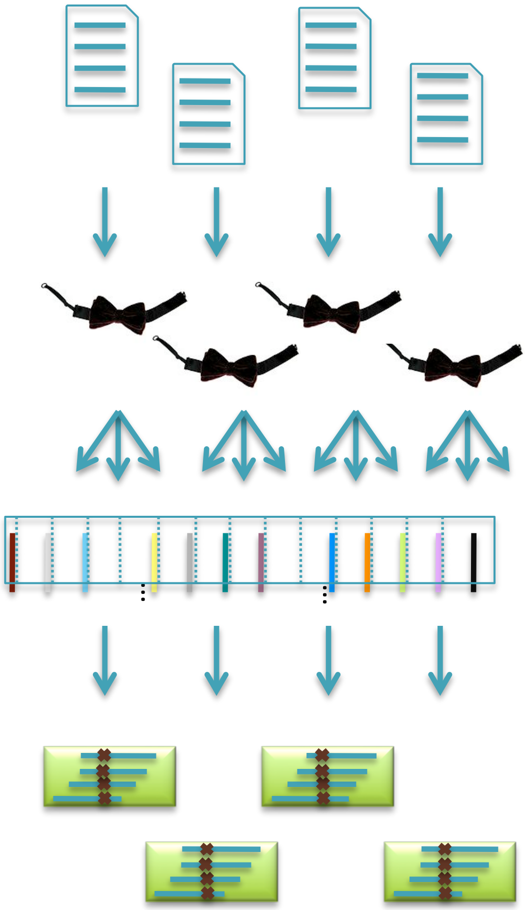
- Cores are only part of the solution.
- Need storage, fast IO
- Locality is king

High-throughput sequence alignment using Graphics Processing Units.
Schatz, MC, Trapnell, C, Delcher, AL, Varshney, A. (2007) BMC Bioinformatics 8:474.

Crossbow

http://bowtie-bio.sourceforge.net/crossbow

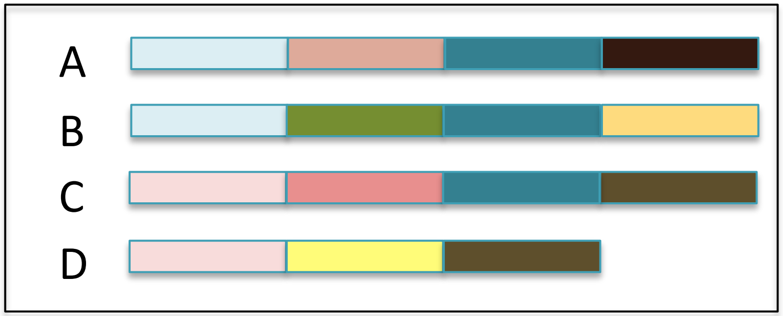
- Align billions of reads and find SNPs
  - Reuse software components: Hadoop Streaming
  - Mapping with Bowtie, SNP calling with SOAPsnp
- 4 hour end-to-end runtime including upload
  - Costs $85; Todays costs <$30



- Very compelling example of cloud computing in genomics
- Transfer takes time, but totally depends on institution
- Need more applications!

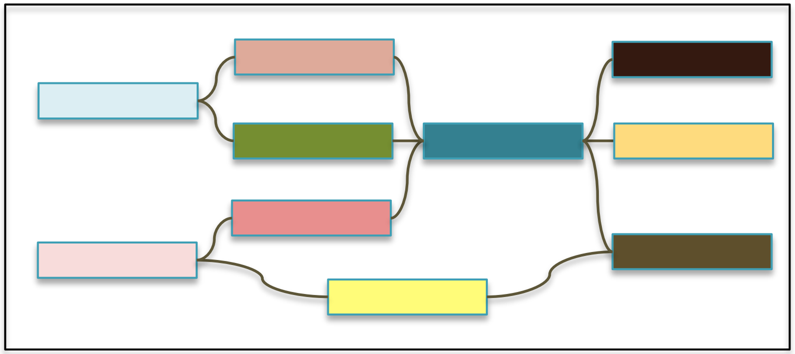
Searching for SNPs with Cloud Computing.
Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL (2009) *Genome Biology.* **10**:R134

Pan-Genome Alignment & Assembly



Time to start considering problems for which N complete genomes is the input to study the "pan-genome"

- Available today for many microbial species, near future for higher eukaryotes



Pan-genome colored de Bruijn graph

- Encodes all the sequence relationships between the genomes
- How well conserved is a given sequence?
- What are the pan-genome network properties?

Rapid pan genome analysis with augmented suffix trees
Marcus, S, Schatz, MC (2014) *In preparation*

Compute & Algorithmic Challenges

Expect to see many dozens of major informatics centers that consolidate regional / topical information

- Clouds for Cancer, Autism, Heart Disease, etc
- Plus many smaller warehouses down to individuals
- Move the code to the data

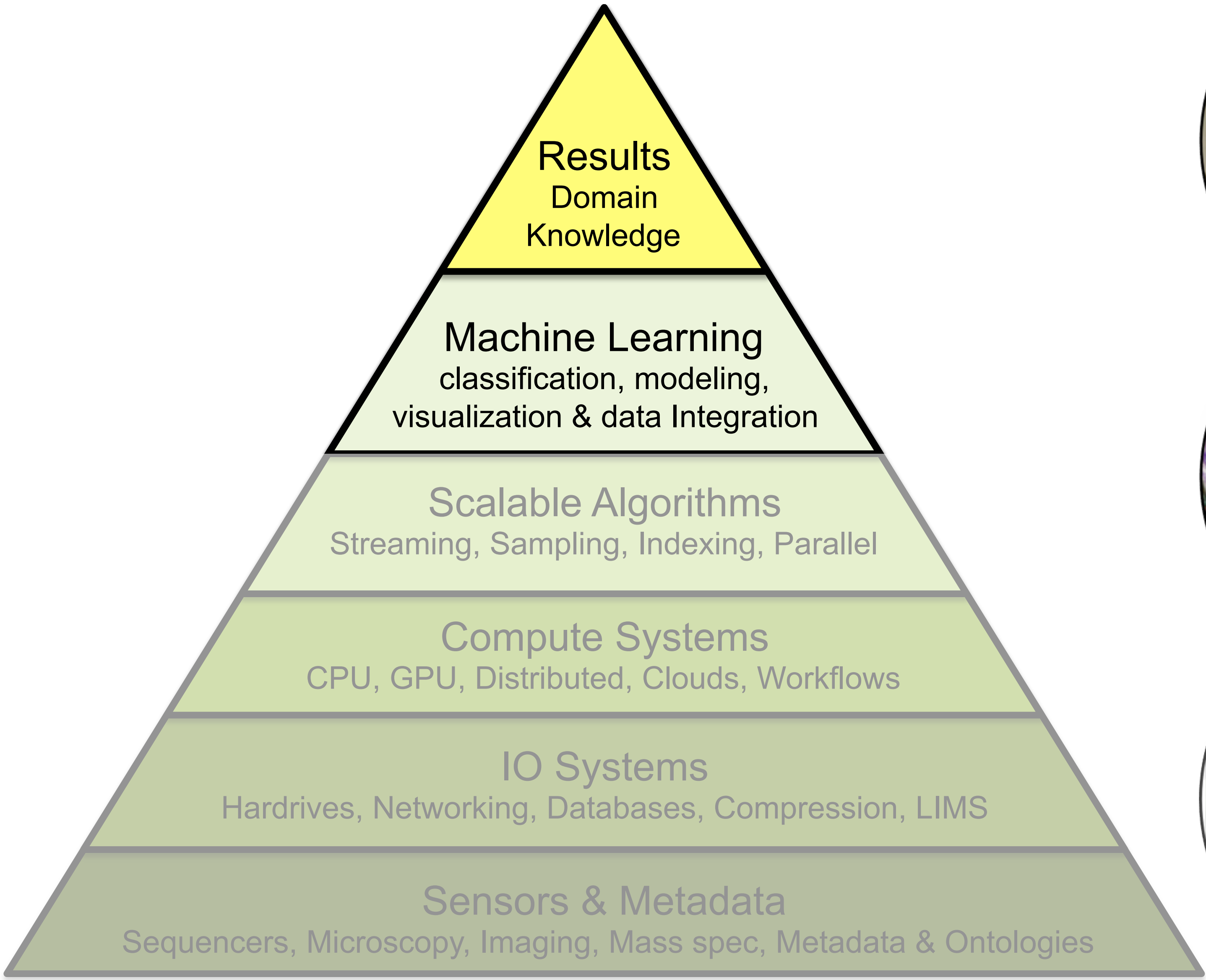
Parallel hardware and algorithms are required

- Expect to see >1000 cores in a single computer
- Compute & IO needs to be considered together
- Rewriting efficient parallel software is complex and expensive

Applications will shift from individuals to populations

- Read mapping & assembly fade out
- Population analysis and time series analysis fade in
- Need for network analysis, probabilistic techniques

Quantitative Biology Technologies



Results
Domain Knowledge

Machine Learning
classification, modeling, visualization & data Integration

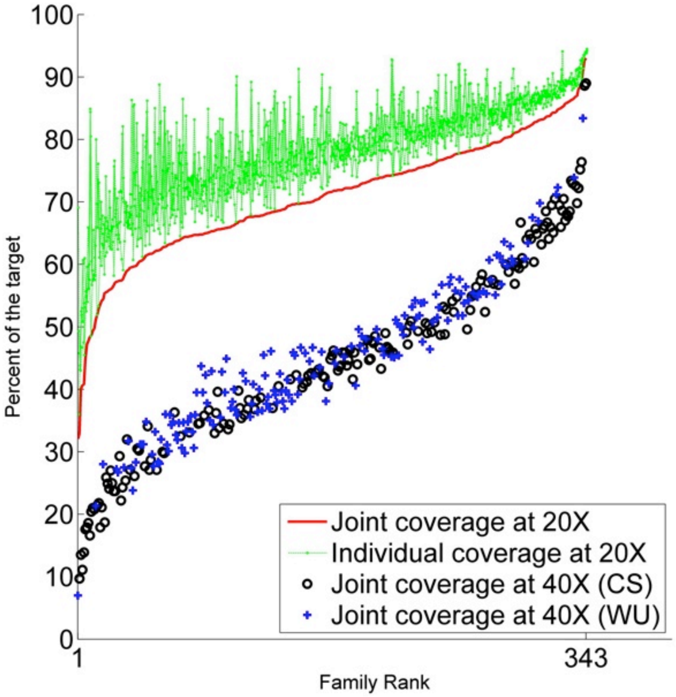
Scalable Algorithms
Streaming, Sampling, Indexing, Parallel

Compute Systems
CPU, GPU, Distributed, Clouds, Workflows

IO Systems
Hardrives, Networking, Databases, Compression, LIMS

Sensors & Metadata
Sequencers, Microscopy, Imaging, Mass spec, Metadata & Ontologies

Exome sequencing of the SSC



Last year saw 3 reports of >593 families from the Simons Simplex Collection

- Parents plus one child with autism and one non-autistic sibling
- All attempted to find "gene killing mutations" specific to the autistic children to find genes associated with the disease
- Iossifov (343) and O'Roak (50) used GATK, Sanders (200) didn't attempt to identify indels

De novo gene disruptions in children on the autism spectrum
Iossifov *et al.* (2012) *Neuron*. 74:2 285-299

De novo mutations revealed by whole-exome sequencing are strongly associated with autism
Sanders *et al.* (2012) *Nature*. 485, 237–241.

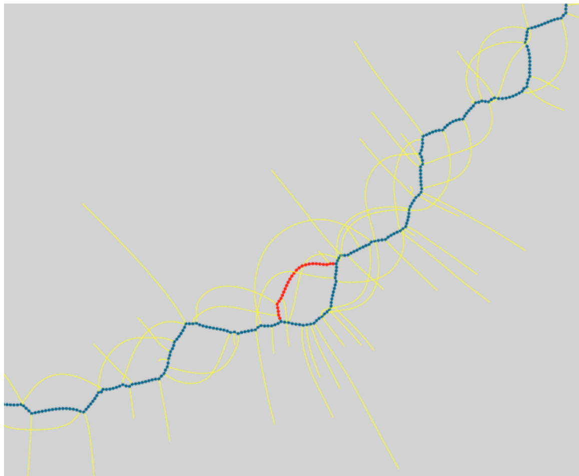
Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations
O'Roak *et al.* (2012) *Nature*. 485, 246–250.

Scalpel: Haplotype Microassembly

DNA sequence **micro-assembly** pipeline for accurate detection and validation of *de novo* mutations (SNPs, indels) within exome-capture data.

Features

1. Combine mapping and assembly
2. Exhaustive search of haplotypes
3. De novo mutations

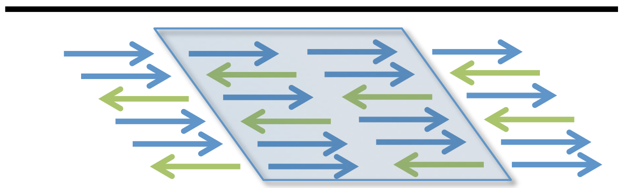


NRXN1 *de novo* SNP
(auSSC12501 chr2:50724605)

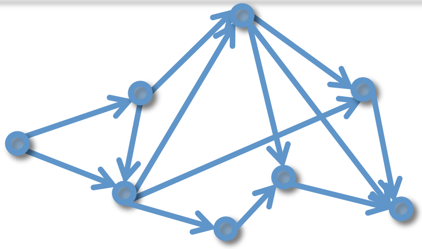
Accurate detection of de novo and transmitted INDELs within exome-capture data using micro-assembly
Narzisi, G, O'Rawe, J, Iossifov, I, Lee, Y, Wang, Z, Wu, Y, Lyon, G, Wigler, M, Schatz, MC (2014) *Under review.*

Scalpel Pipeline

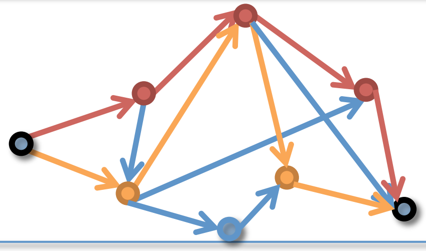
Extract reads mapping within the exon including (1) well-mapped reads, (2) soft-clipped reads, and (3) anchored pairs



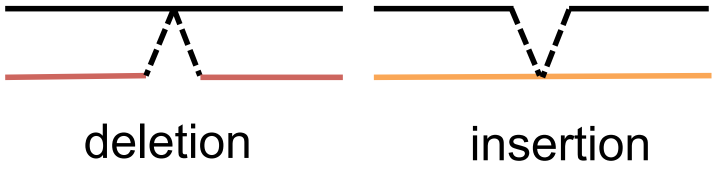
Decompose reads into overlapping *k*-mers and construct de Bruijn graph from the reads



Find end-to-end haplotype paths spanning the region



Align assembled sequences to reference to detect mutations



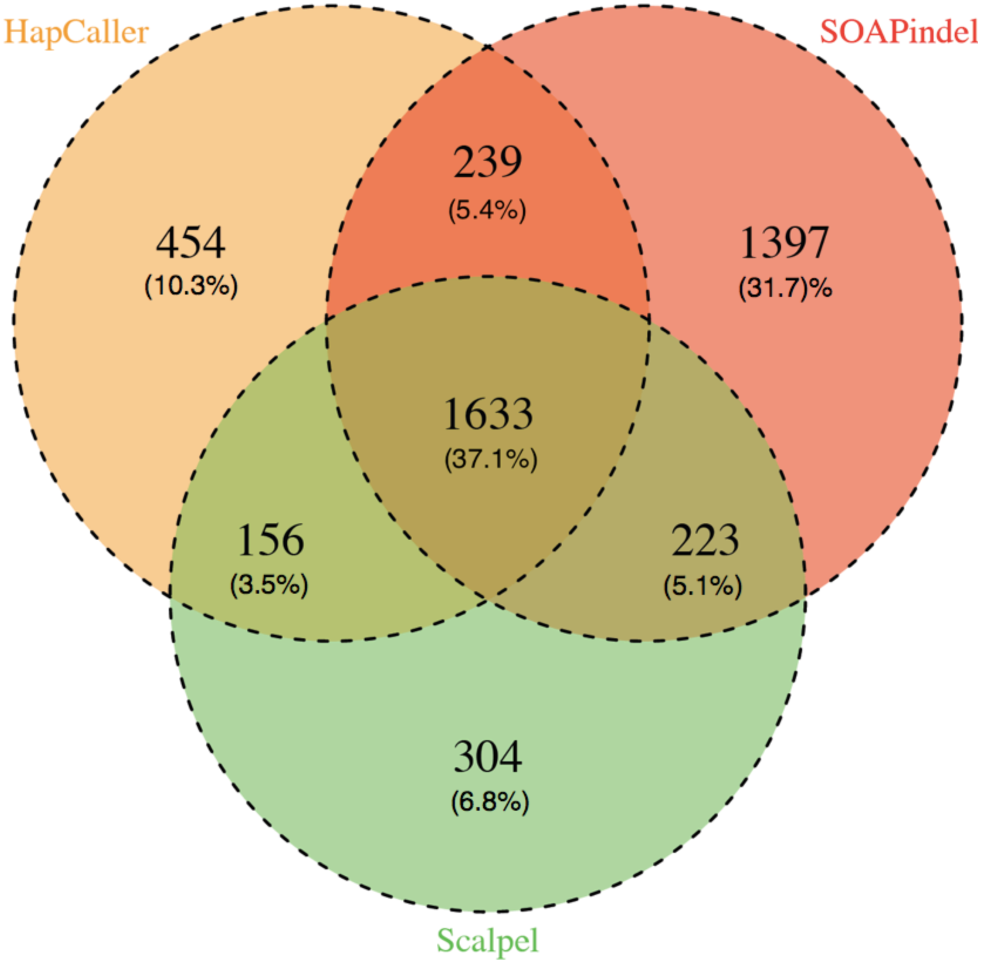
Experimental Analysis & Validation

Selected one deep coverage exome for deep analysis

- Individual was diagnosed with ADHD
- 80% of the target at >20x coverage
- Evaluated with Scalpel, SOAPindel, and GATK Haplotype Caller

1000 indels selected for validation

- 200 Scalpel
- 200 GATK Haplotype Caller
- 200 SOAPindel
- 200 within the intersection
- 200 long indels (>30bp)



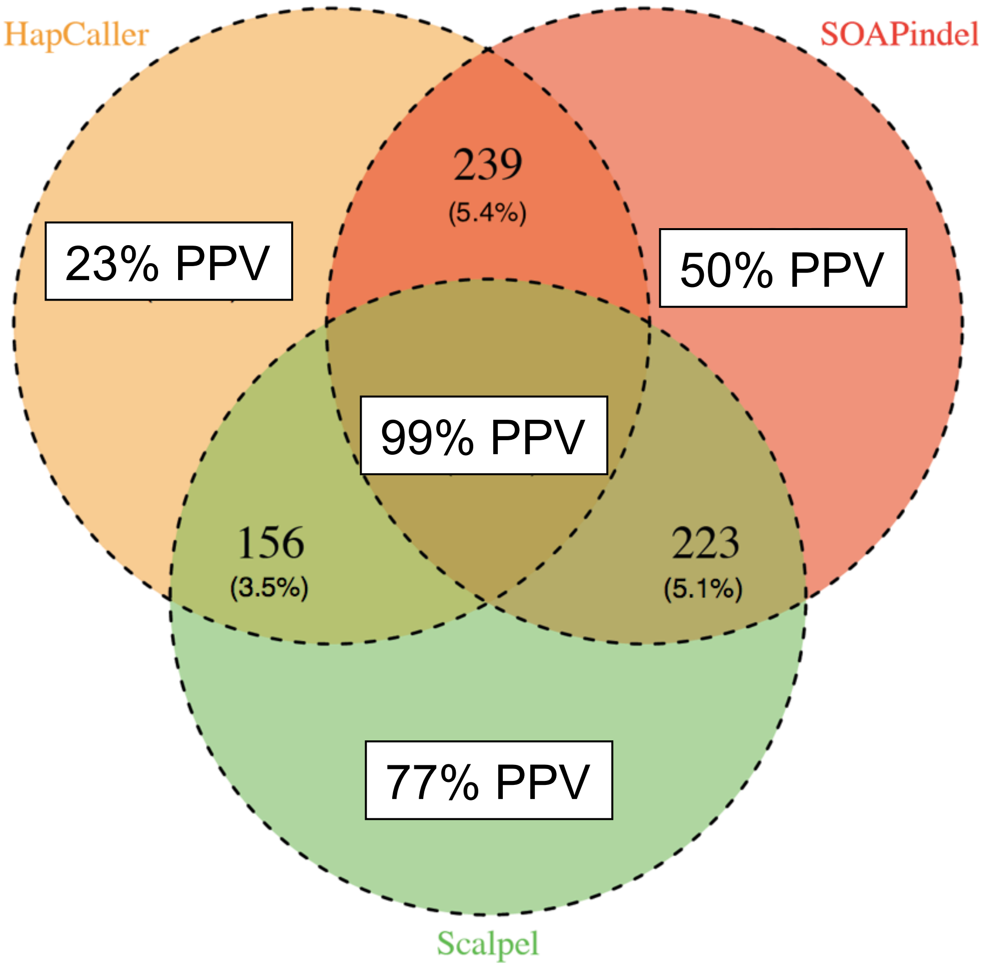
Experimental Analysis & Validation

Selected one deep coverage exome for deep analysis

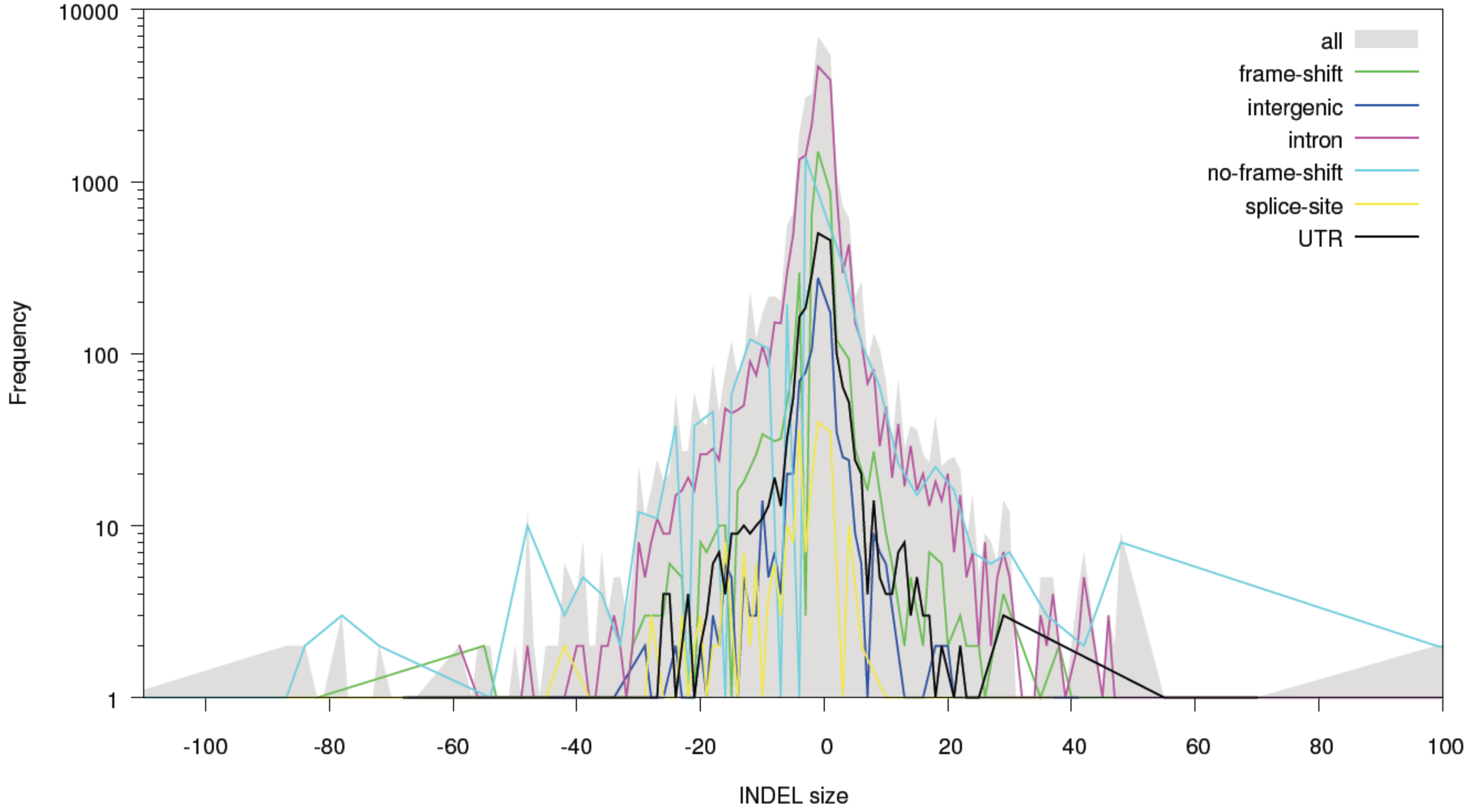
- Individual was diagnosed with ADHD (See Gholson for details)
- 80% of the target at >20x coverage
- Evaluated with Scalpel, SOAPindel, and GATK Haplotype Caller

1000 indels selected for validation

- 200 Scalpel
- 200 GATK Haplotype Caller
- 200 SOAPindel
- 200 within the intersection
- 200 long indels (>30bp)



Revised Analysis of the SSC

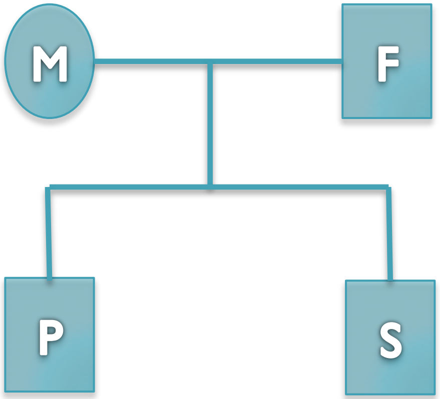


Constructed database of >1M transmitted and de novo indels
Many new gene candidates identified, population analysis underway

De novo mutation discovery and validation

Concept: Identify mutations not present in parents.

Challenge: Sequencing errors in the child or low coverage in parents lead to false positive de novos



```
Reference:  ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Father:     ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Mother:     ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Sibling:    ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Proband(1): ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Proband(2): ...TCAAATCCTTTTAAT****AAGAGCTGACA...
```

4bp heterozygous deletion at chr15:93524061 CHD2

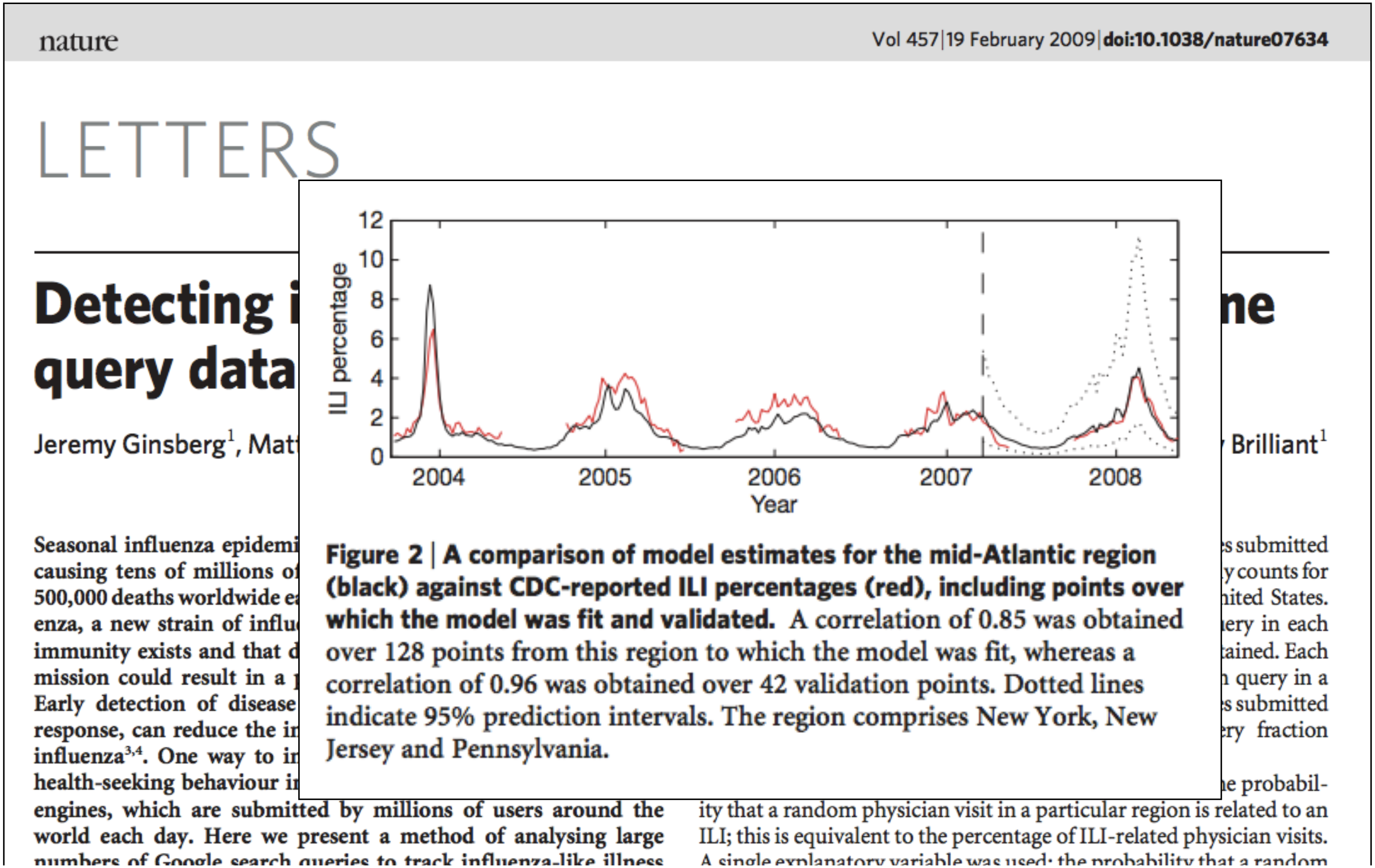
De novo Genetics of Autism

- In 593 family quads so far, we see significant enrichment in de novo ***likely gene killers*** in the autistic kids
  - Overall rate basically 1:1
  - 2:1 enrichment in nonsense mutations
  - 2:1 enrichment in frameshift indels
  - 4:1 enrichment in splice-site mutations
  - Most de novo originate in the paternal line in an age-dependent manner (56:18 of the mutations that we could determine)

- Observe strong overlap with the 842 genes known to be associated with fragile X protein FMPR
  - Related to neuron development and synaptic plasticity
  - Also strong overlap with chromatin remodelers

Accurate detection of de novo and transmitted INDELs within exome-capture data using micro-assembly
Narzisi, G, O'Rawe, J, Iossifov, I, Lee, Y, Wang, Z, Wu, Y, Lyon, G, Wigler, M, Schatz, MC (2014) Under review.

The potential for big data?



nature

Vol 457|19 February 2009|doi:10.1038/nature07634

LETTERS

Detecting i ne
query data

Jeremy Ginsberg[1], Matt Brilliant[1]

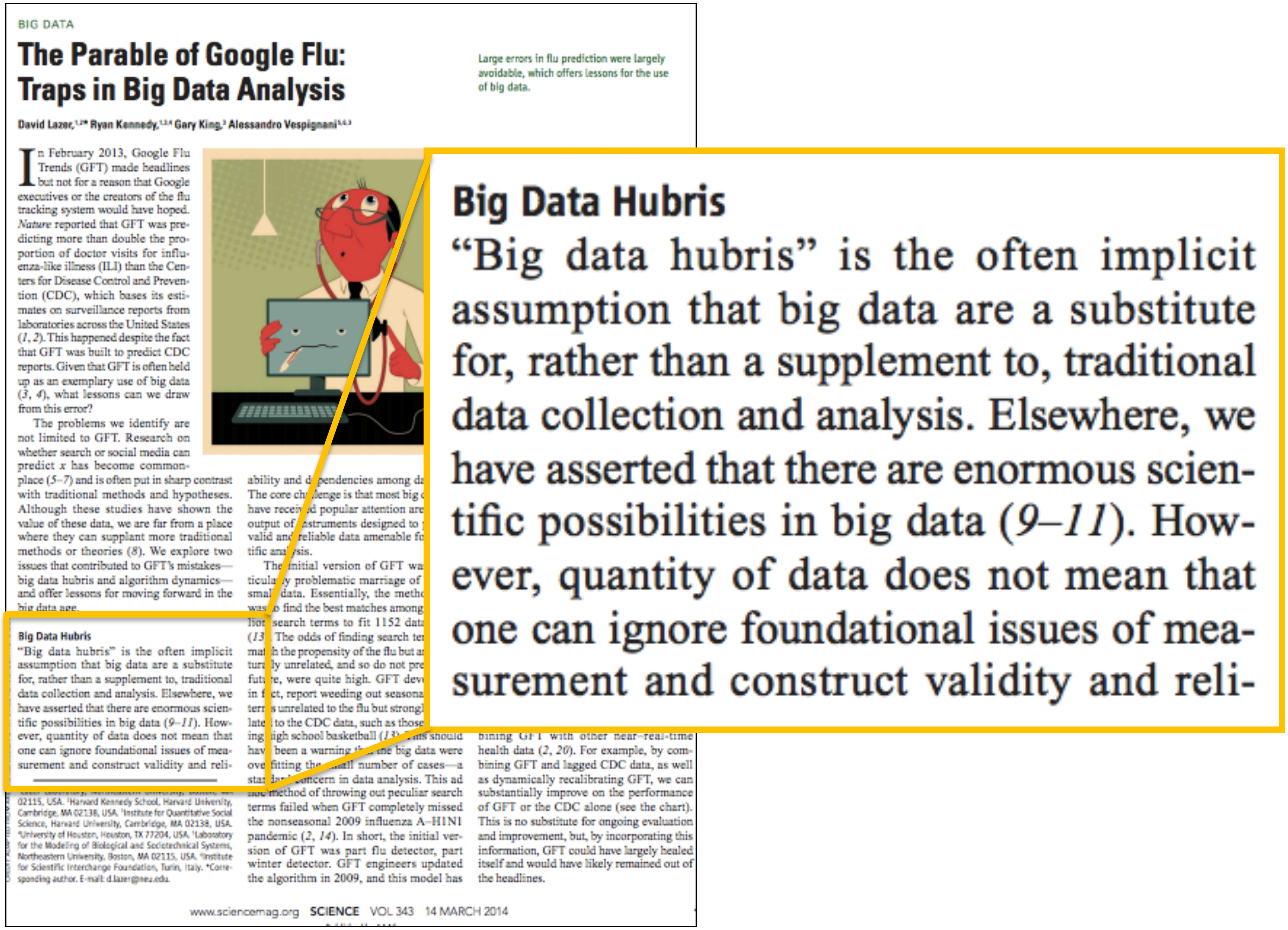
Figure 2 | A comparison of model estimates for the mid-Atlantic region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, whereas a correlation of 0.96 was obtained over 42 validation points. Dotted lines indicate 95% prediction intervals. The region comprises New York, New Jersey and Pennsylvania.

Seasonal influenza epidemi causing tens of millions of 500,000 deaths worldwide ea enza, a new strain of influ immunity exists and that d mission could result in a p Early detection of disease response, can reduce the in influenza[3,4]. One way to in health-seeking behaviour in engines, which are submitted by millions of users around the world each day. Here we present a method of analysing large numbers of Google search queries to track influenza-like illness

s submitted y counts for nited States. ery in each tained. Each n query in a s submitted ery fraction

he probability that a random physician visit in a particular region is related to an ILI; this is equivalent to the percentage of ILI-related physician visits. A single explanatory variable was used: the probability that a random

The fallacy of big data?



BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

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

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 predict *x* has become commonplace (*5*–*7*) and is often put in sharp contrast with traditional methods and hypotheses. Although these studies have shown the value of these data, we are far from a place where they can supplant more traditional methods or theories (*8*). We explore two issues that contributed to GFT's mistakes—big data hubris and algorithm dynamics—and offer lessons for moving forward in the big data age.

Big Data Hubris

"Big data hubris" is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis. Elsewhere, we have asserted that there are enormous scientific possibilities in big data (*9*–*11*). However, quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reli-

02115, USA. [2]Harvard Kennedy School, Harvard University, Cambridge, MA 02138, USA. [3]Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA. [4]University of Houston, Houston, TX 77204, USA. [5]Laboratory for the Modeling of Biological and Sociotechnical Systems, Northeastern University, Boston, MA 02115, USA. [6]Institute for Scientific Interchange Foundation, Turin, Italy. *Corresponding author. E-mail: d.lazer@neu.edu.

ability and d pendencies among da
The core ch lenge is that most big d
have recei d popular attention are
output of astruments designed to p
valid and reliable data amenable fo
tific analysis.

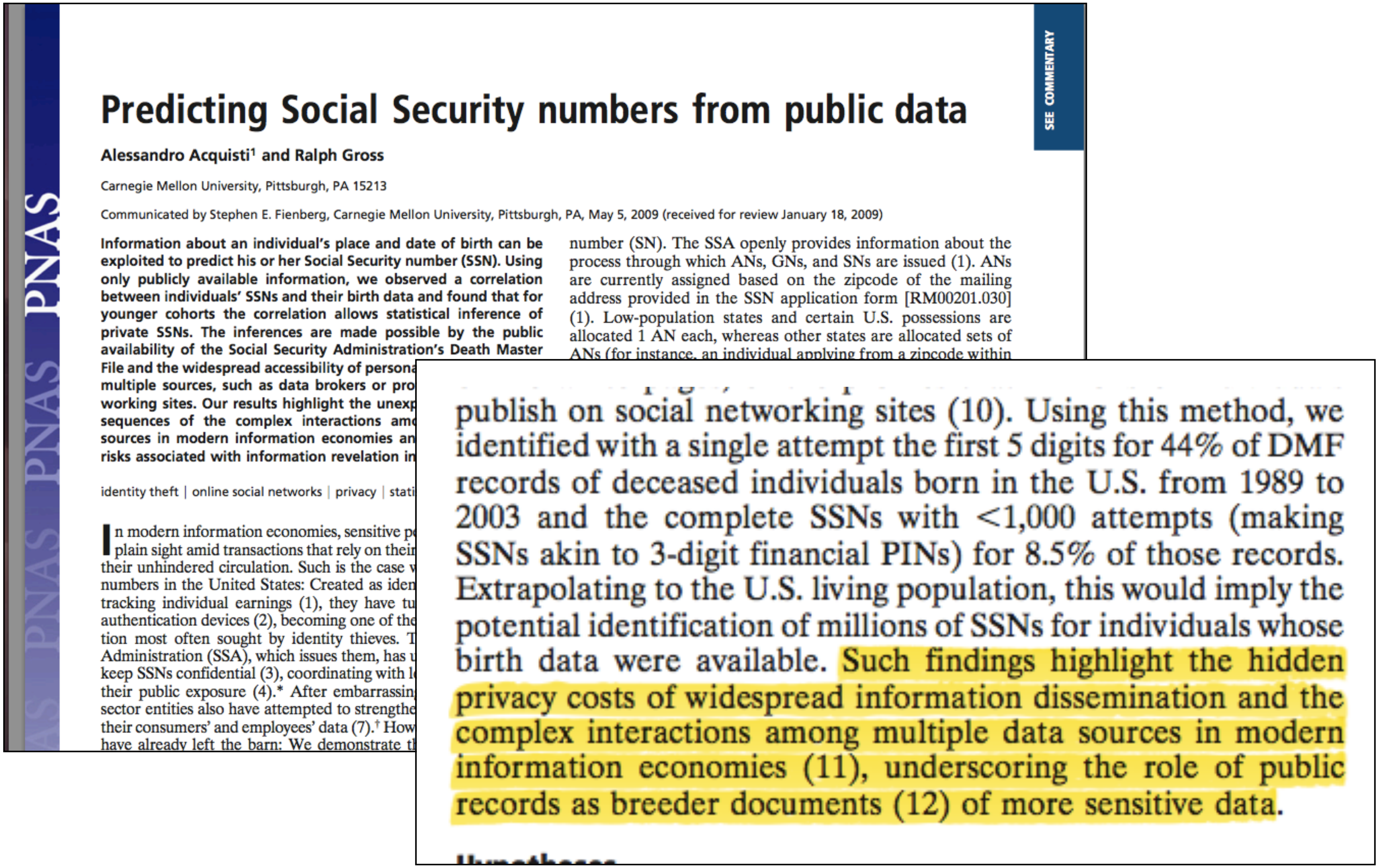
The initial version of GFT wa
ticula ly problematic marriage of
smal data. Essentially, the metho
was o find the best matches among
lion search terms to fit 1152 data
(*13* The odds of finding search ter
mat h the propensity of the flu but ar
tura ly unrelated, and so do not pre
futu e, were quite high. GFT deve
in f ct, report weeding out seasona
terr s unrelated to the flu but strongl
late to the CDC data, such as those
ing igh school basketball (*13*). This should
hav been a warning that the big data were
ove fitting the small number of cases—a
standard concern in data analysis. This ad
hoc method of throwing out peculiar search terms failed when GFT completely missed the nonseasonal 2009 influenza A–H1N1 pandemic (*2*, *14*). In short, the initial version of GFT was part flu detector, part winter detector. GFT engineers updated the algorithm in 2009, and this model has

bining GFT with other near–real-time health data (*2*, *20*). For example, by combining GFT and lagged CDC data, as well as dynamically recalibrating GFT, we can substantially improve on the performance of GFT or the CDC alone (see the chart). This is no substitute for ongoing evaluation and improvement, but, by incorporating this information, GFT could have largely healed itself and would have likely remained out of the headlines.

Big Data Hubris

"Big data hubris" is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis. Elsewhere, we have asserted that there are enormous scientific possibilities in big data (*9*–*11*). However, quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reli-

The risks of big data?



Predicting Social Security numbers from public data

Alessandro Acquisti[1] and Ralph Gross

Carnegie Mellon University, Pittsburgh, PA 15213

Communicated by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, May 5, 2009 (received for review January 18, 2009)

Information about an individual's place and date of birth can be exploited to predict his or her Social Security number (SSN). Using only publicly available information, we observed a correlation between individuals' SSNs and their birth data and found that for younger cohorts the correlation allows statistical inference of private SSNs. The inferences are made possible by the public availability of the Social Security Administration's Death Master File and the widespread accessibility of persona multiple sources, such as data brokers or pro working sites. Our results highlight the unexp sequences of the complex interactions amo sources in modern information economies an risks associated with information revelation in

identity theft | online social networks | privacy | stati

In modern information economies, sensitive pe plain sight amid transactions that rely on their their unhindered circulation. Such is the case w numbers in the United States: Created as iden tracking individual earnings (1), they have tu authentication devices (2), becoming one of the tion most often sought by identity thieves. T Administration (SSA), which issues them, has u keep SSNs confidential (3), coordinating with l their public exposure (4).* After embarrassin sector entities also have attempted to strengthe their consumers' and employees' data (7).† How have already left the barn: We demonstrate th

number (SN). The SSA openly provides information about the process through which ANs, GNs, and SNs are issued (1). ANs are currently assigned based on the zipcode of the mailing address provided in the SSN application form [RM00201.030] (1). Low-population states and certain U.S. possessions are allocated 1 AN each, whereas other states are allocated sets of ANs (for instance, an individual applying from a zipcode within

publish on social networking sites (10). Using this method, we identified with a single attempt the first 5 digits for 44% of DMF records of deceased individuals born in the U.S. from 1989 to 2003 and the complete SSNs with <1,000 attempts (making SSNs akin to 3-digit financial PINs) for 8.5% of those records. Extrapolating to the U.S. living population, this would imply the potential identification of millions of SSNs for individuals whose birth data were available. Such findings highlight the hidden privacy costs of widespread information dissemination and the complex interactions among multiple data sources in modern information economies (11), underscoring the role of public records as breeder documents (12) of more sensitive data.

Learning and Translation

Tremendous power from data aggregation

- Observe the dynamics of biological systems
- Breakthroughs in medicine and biology of profound significance

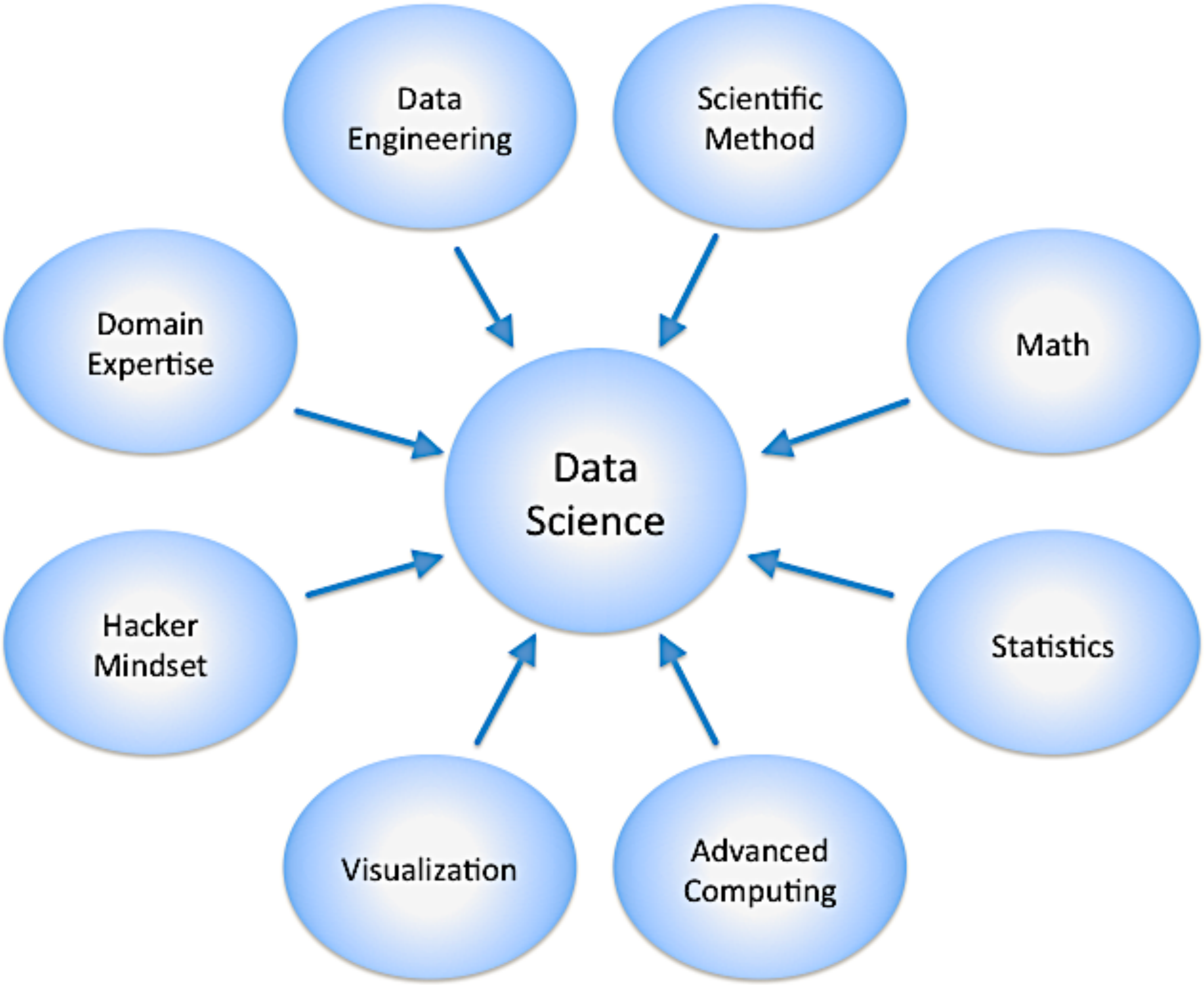
Be mindful of the risks

- The potential for over-fitting grows with the complexity of the data, statistical significance is a statement about the sample size
- Reproducible workflows,APIs are a must
- Caution is prudent for personal data

The foundations of biology will continue to be observation, experimentation, and interpretation

- Technology will continue to push the frontier
- Feedback loop from the results of one project into experimental design for the next

Who is a Data Scientist?



http://en.wikipedia.org/wiki/Data_science

Acknowledgements

Schatz Lab

Giuseppe Narzisi

Shoshana Marcus

James Gurtowski

Srividya Ramakrishnan

Hayan Lee

Rob Aboukhalil

Mitch Bekritsky

Charles Underwood

Tyler Gavin

Alejandro Wences

Greg Vurture

Eric Biggers

Aspyn Palatnick

CSHL

Hannon Lab

Gingeras Lab

Jackson Lab

Iossifov Lab

Levy Lab

Lippman Lab

Lyon Lab

Martienssen Lab

McCombie Lab

Tuveson Lab

Ware Lab

Wigler Lab

IT Department

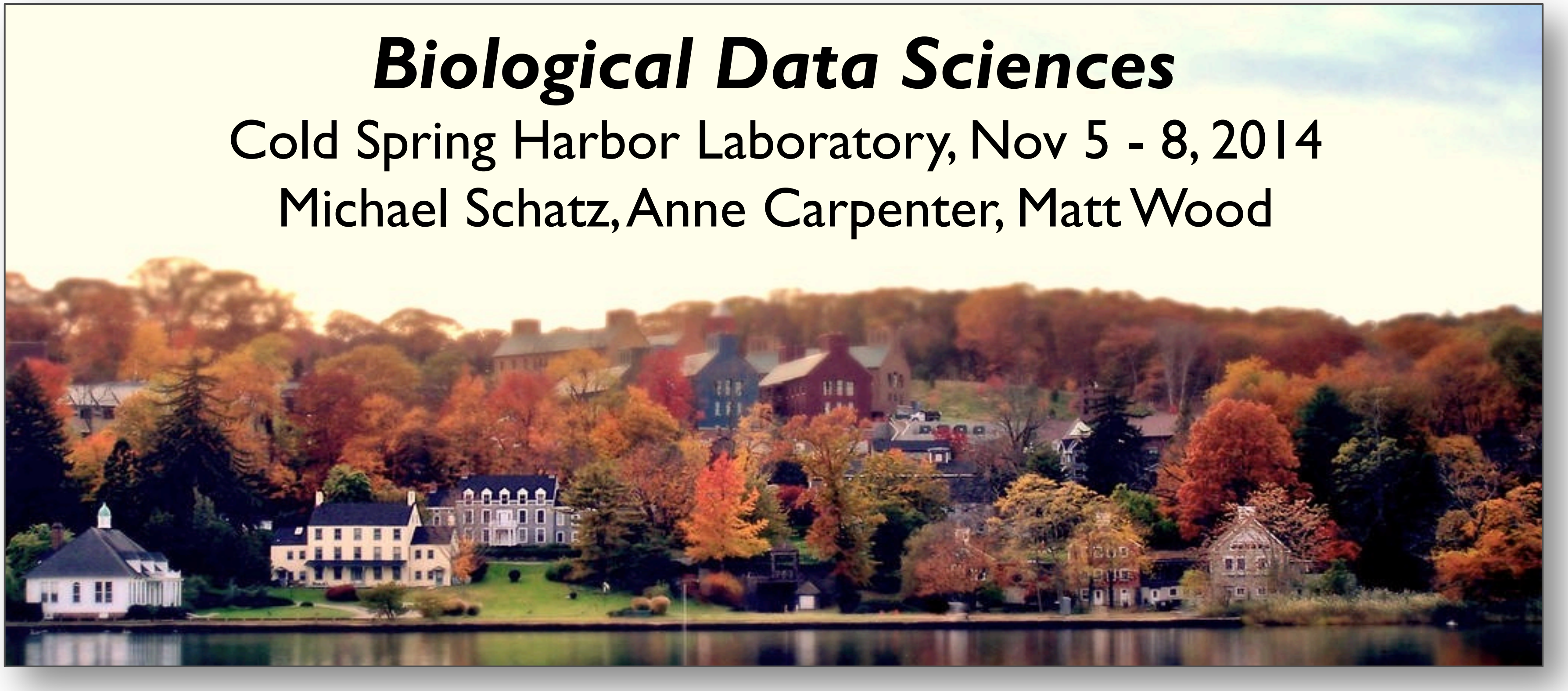
SFARI

SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

National Human Genome Research Institute

U.S. DEPARTMENT OF ENERGY

NSF



Biological Data Sciences

Cold Spring Harbor Laboratory, Nov 5 - 8, 2014

Michael Schatz, Anne Carpenter, Matt Wood

Thank you

http://schatzlab.cshl.edu

@mike_schatz / #KSBigData