

Big Data Meets DNA

How Biological Data Science is improving our health, foods, and energy needs

Michael Schatz

May 22, 2014

Procter and Gamble



DNA: The secret of life



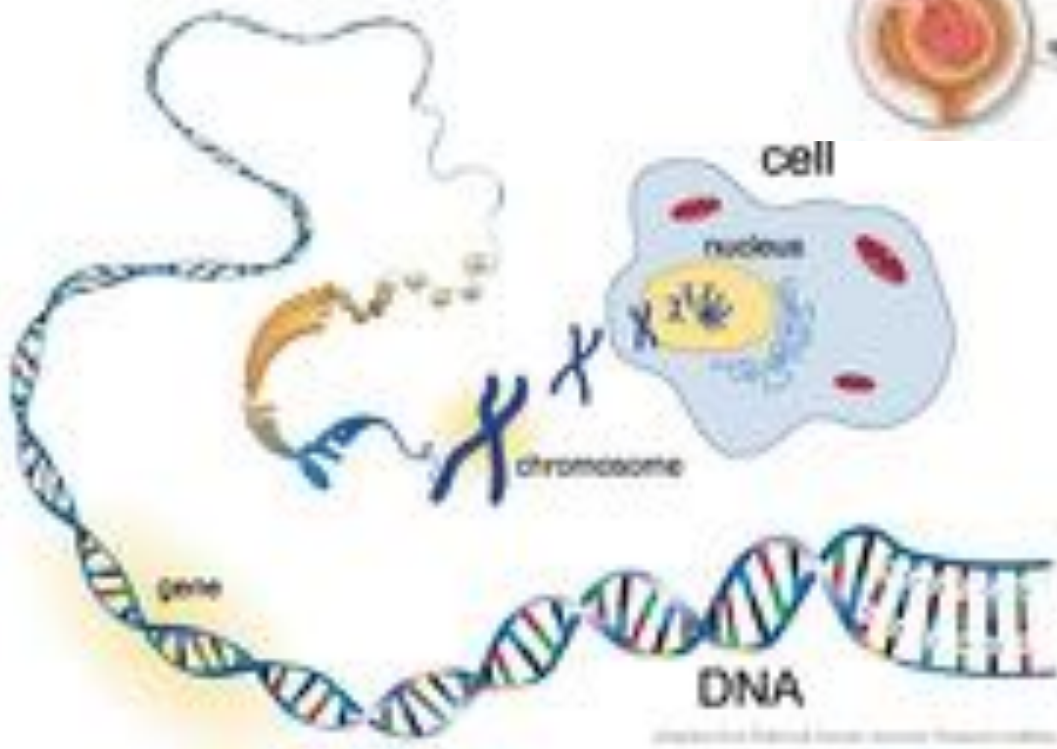
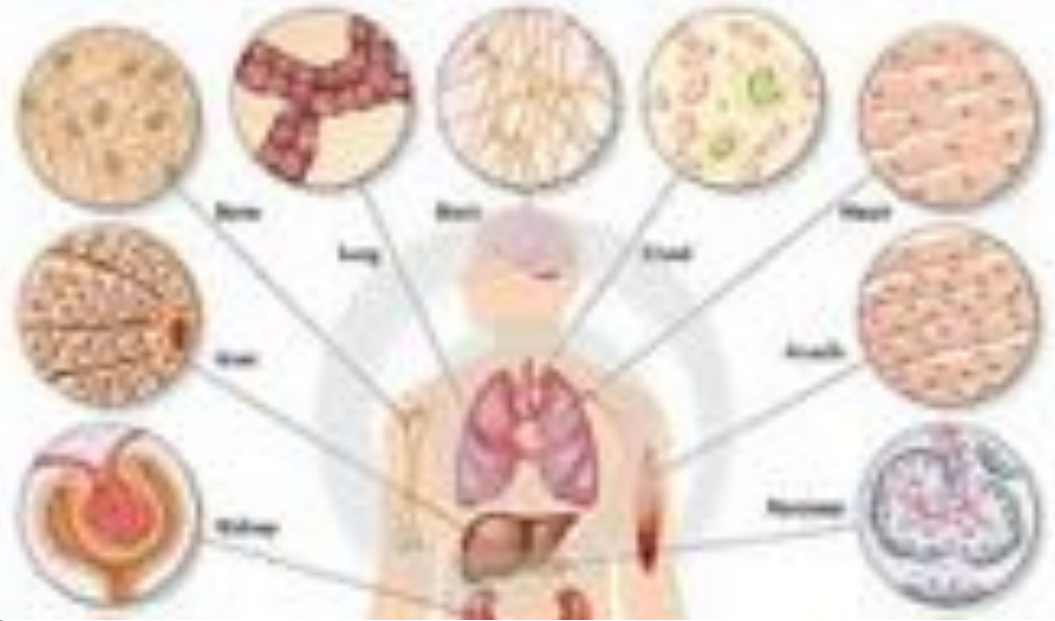
Your DNA, along with your environment and experiences, shapes who you are

- Height
- Hair, eye, skin color
- Broad/narrow, small/large features
- Susceptibility to disease
- Response to drug treatments
- Longevity and cognition

Physical traits tend to be strongly genetic, social characteristics tend to be strongly environmental, and everything else is a combination

Cells & DNA

Each cell of your body contains an exact copy of your 3 billion base pair genome.



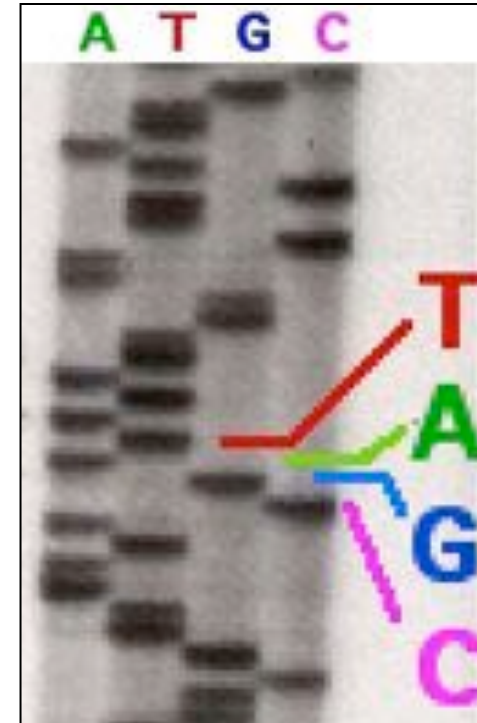
Your specific nucleotide sequence encodes the genetic program for your cells and ultimately your traits

The Origins of DNA Sequencing



Sanger et al. (1977) Nature
1st Complete Organism
Bacteriophage ϕ X174; 5375 bp

Awarded Nobel Prize in 1980



Radioactive Chain Termination
5000bp / week / person

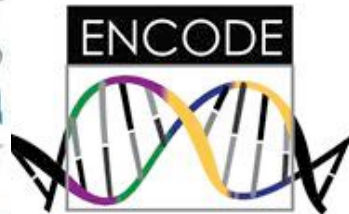
<http://en.wikipedia.org/wiki/File:Sequencing.jpg>
<http://www.answers.com/topic/automated-sequencer>

Milestones in DNA Sequencing



(TIGR/Celera, 1995-2001)

Genomics across the tree of life



Unsolved Questions in Biology

- What is your genome sequence?

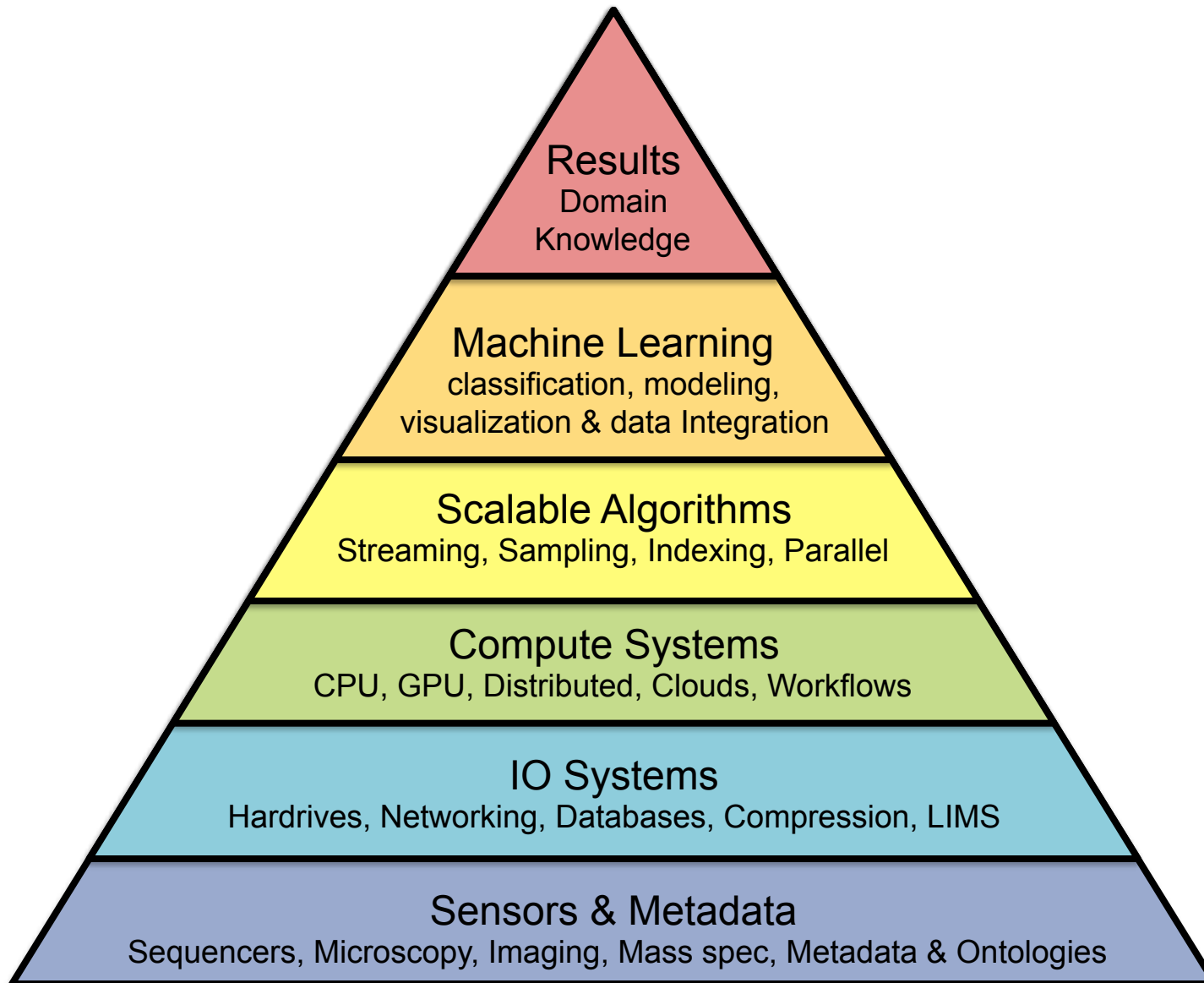
The instruments provide the data, but none of the answers to any of these questions.

What software and systems will?

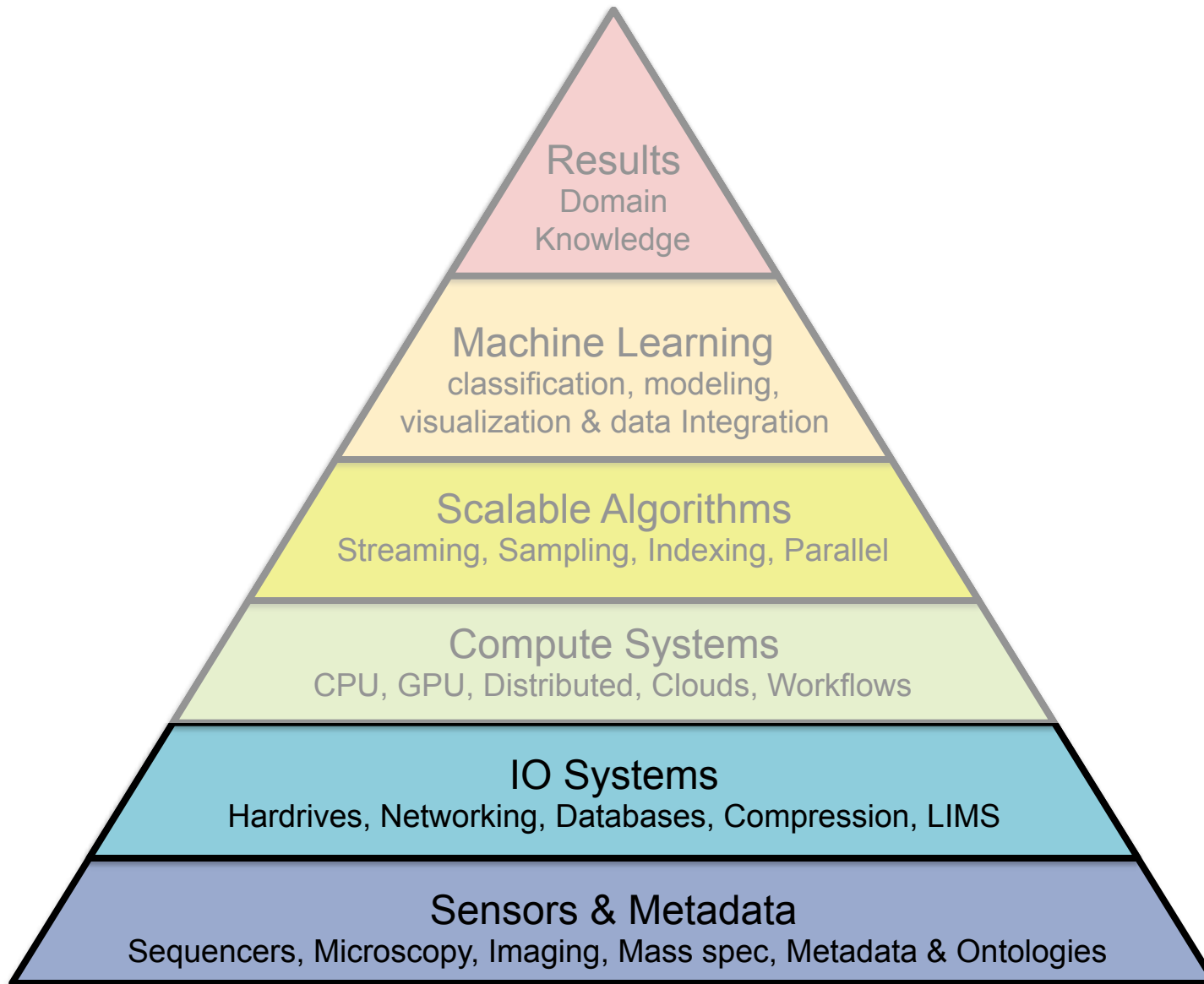
- How do your mutations relate to disease?
- What drugs and treatments should we give you?
- ***Plus hundreds and hundreds more***



Quantitative Biology Technologies



Quantitative Biology Technologies

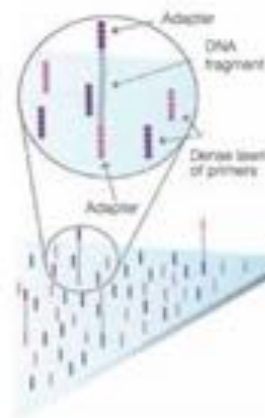


Massively Parallel Sequencing

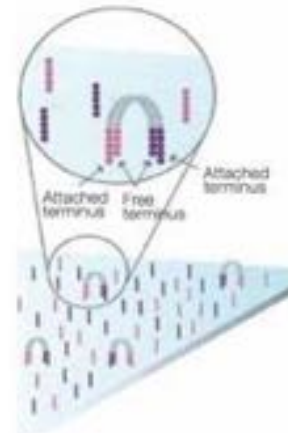


Illumina HiSeq 2000
Sequencing by Synthesis

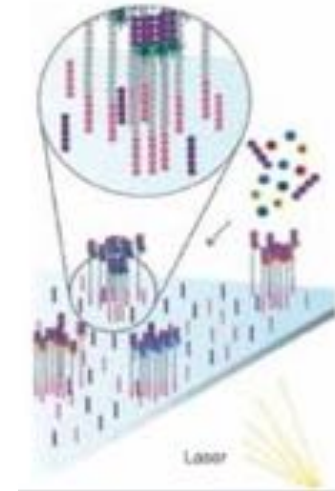
>60Gbp / day



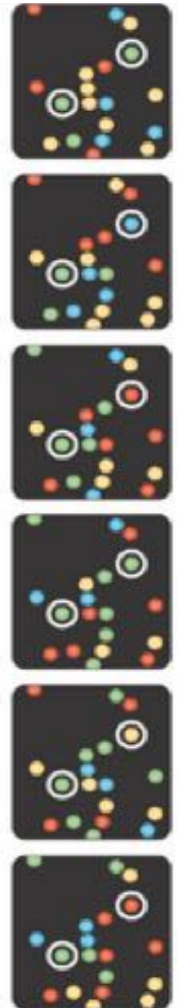
1. Attach



2. Amplify

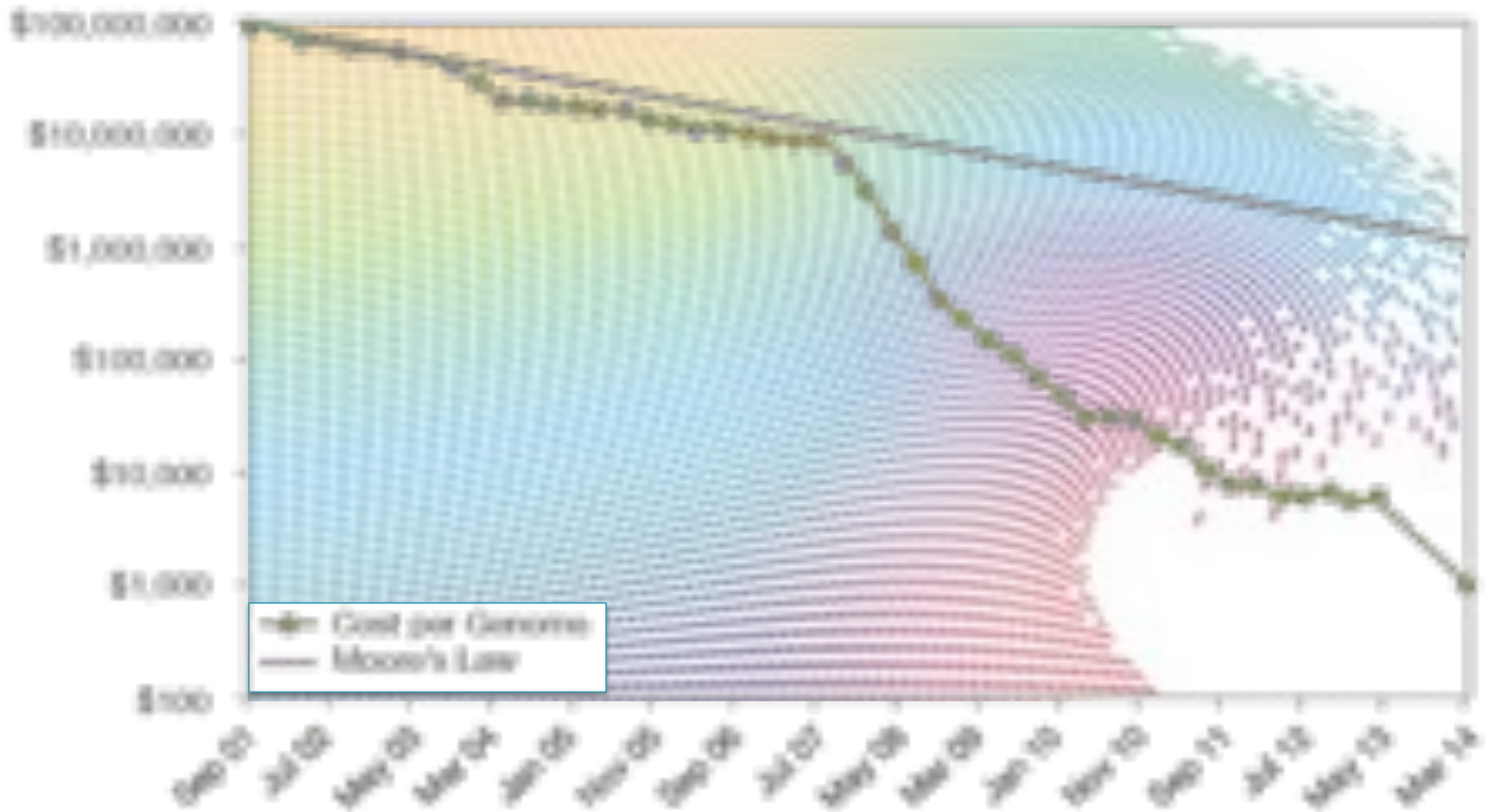


3. Image



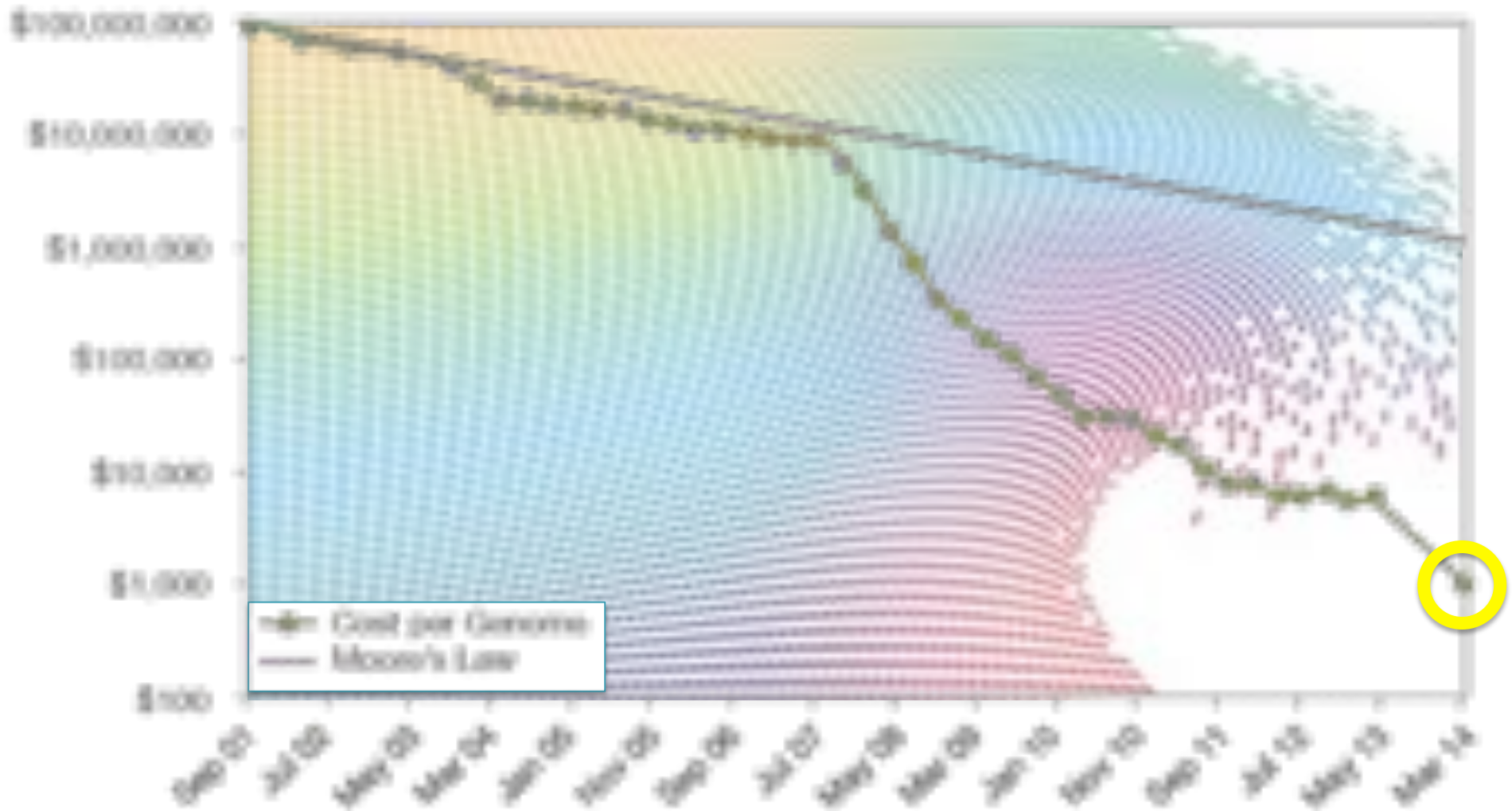
Metzker (2010) Nature Reviews Genetics 11:31-46
<http://www.youtube.com/watch?v=I99aKKHcxC4>

Cost per Genome



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

Cost per Genome



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

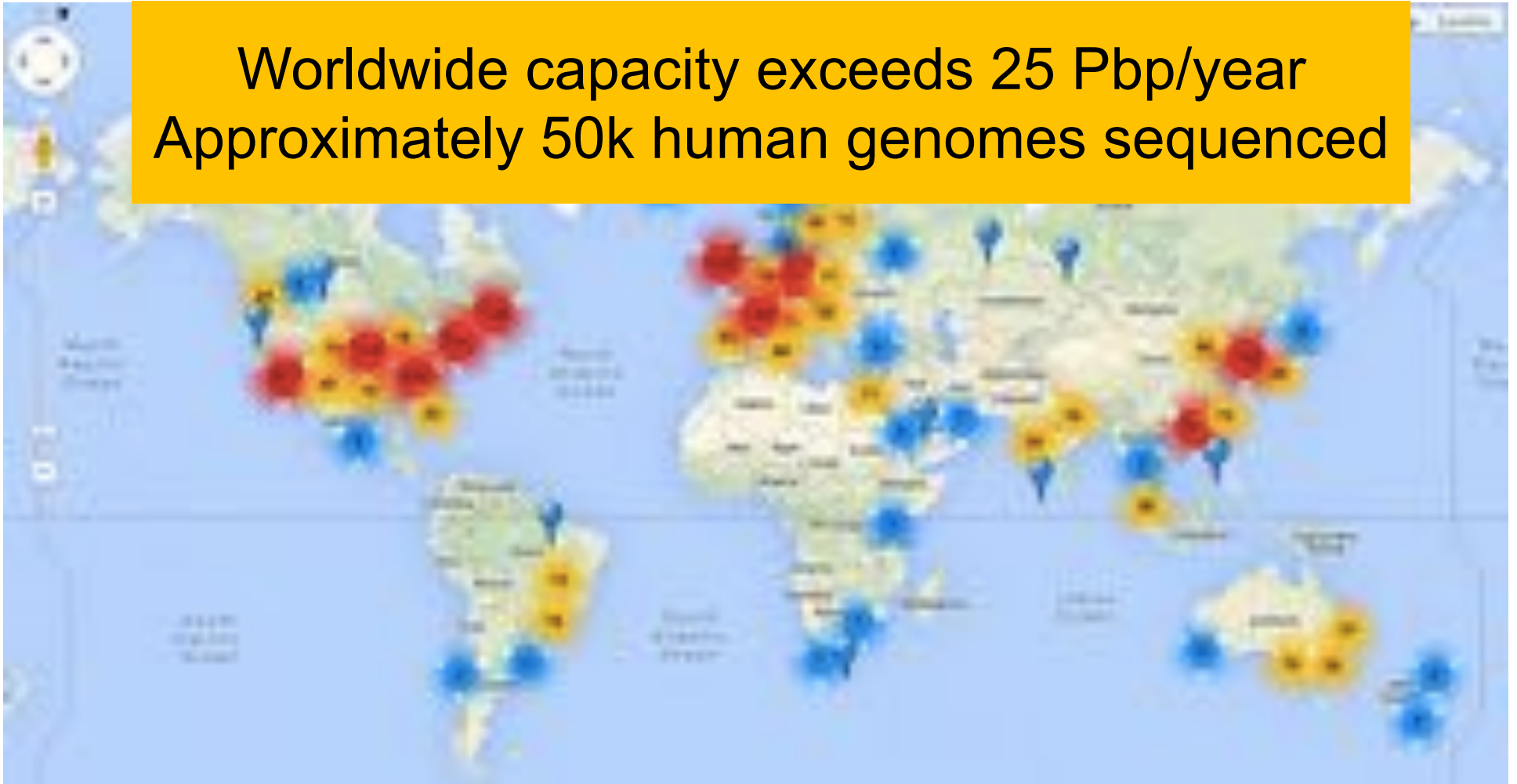
HiSeq X Ten



320 genomes per week / 18,000 genomes per year
\$1000 per genome / ~\$10 M per instrument

Sequencing Centers

Worldwide capacity exceeds 25 Pbp/year
Approximately 50k human genomes sequenced



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



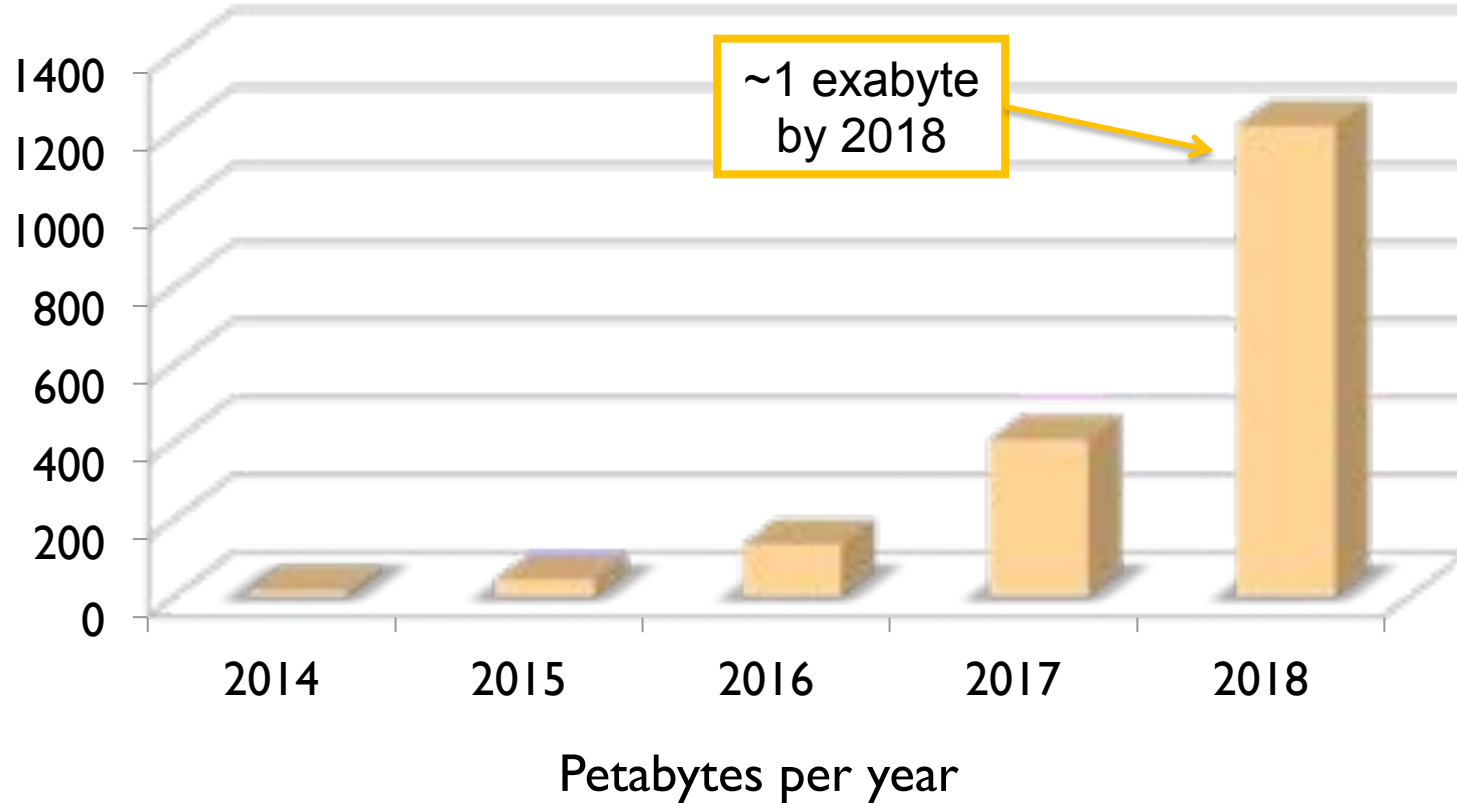
787 feet of DVDs
~1/6 of a mile tall



500 2 TB drives
\$500k

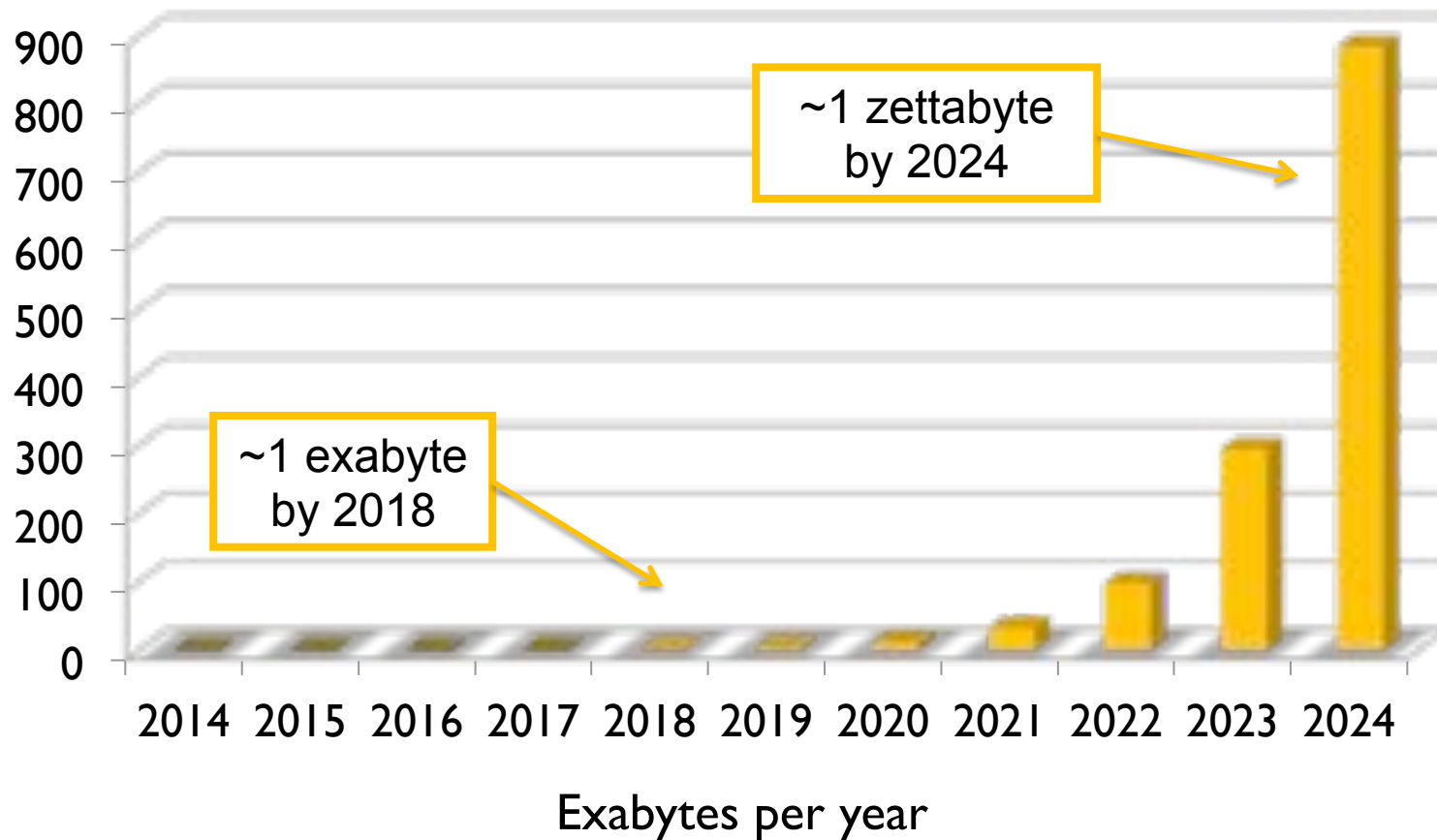
DNA Data Tsunami

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



DNA Data Tsunami

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



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
~ 1/2 distance to moon



Both currently ~100Pb
And growing exponentially

Sequencing Centers 2014



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

Sequencing Centers 2024



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

Biological Sensor Network



Oxford Nanopore



DC Metro via the LA Times

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

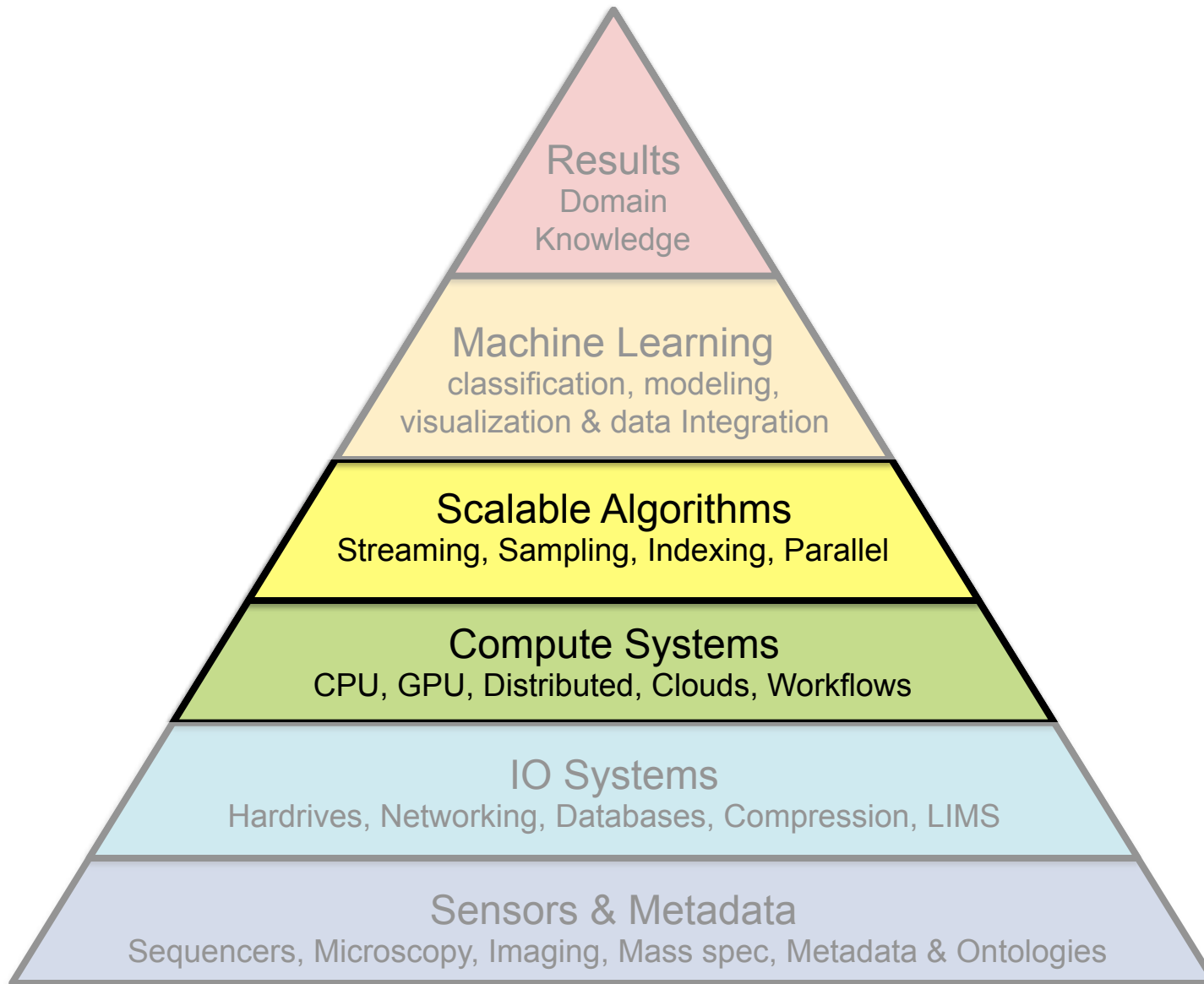
- Major efforts in human health and disease, agriculture, bioenergy

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



Sequencing Centers 2024



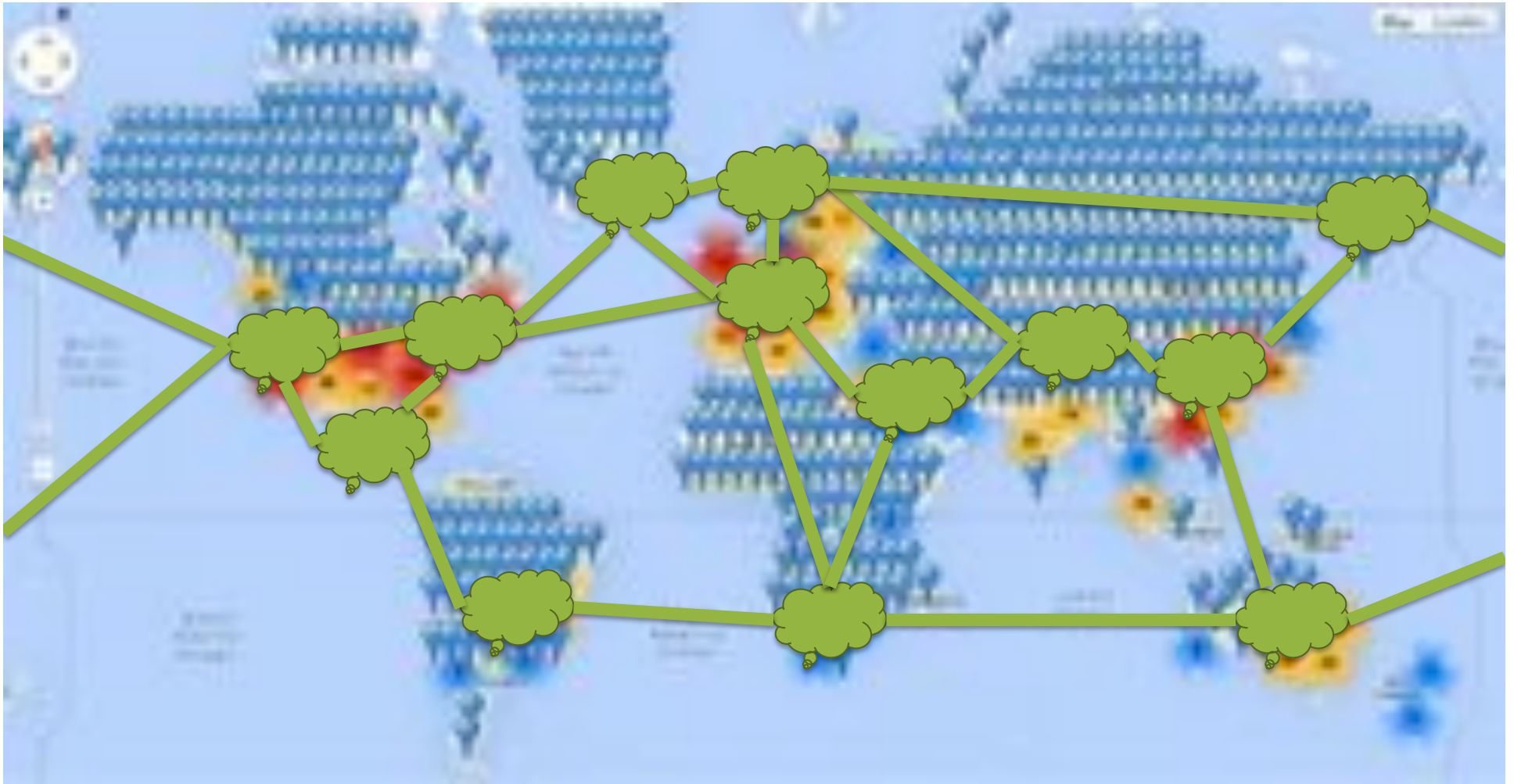
Informatics Centers 2024



The DNA Data Deluge

Schatz, MC and Langmead, B (2013) *IEEE Spectrum*. July, 2013

Informatics Centers 2014



The DNA Data Deluge

Schatz, MC and Langmead, B (2013) *IEEE Spectrum*. July, 2013

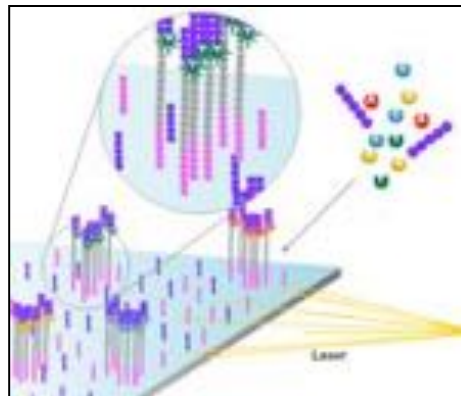
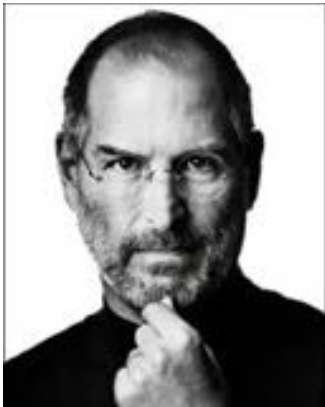
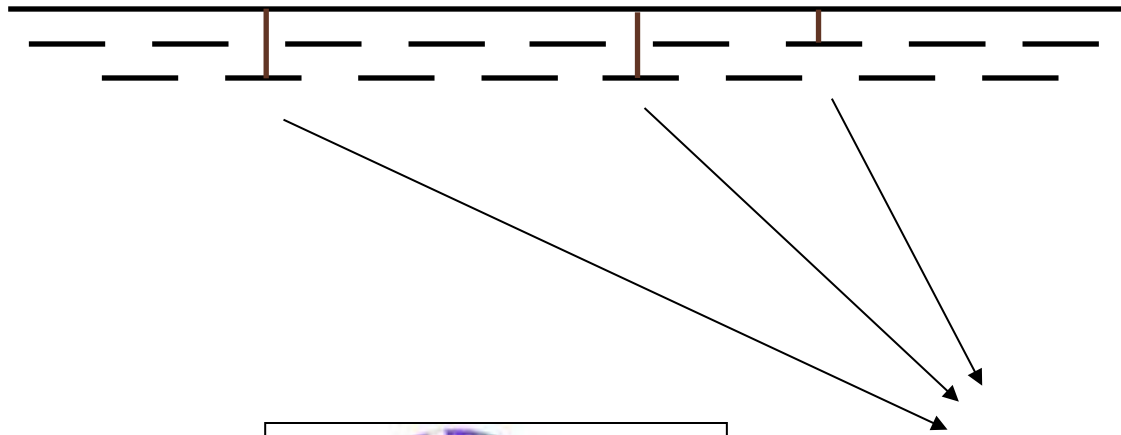
DOE Systems Biology Knowledgebase



<http://kbase.us>: Predictive Biology in Microbes, Plants, and Meta-communities

Personal Genomics

How does your genome compare to the reference?

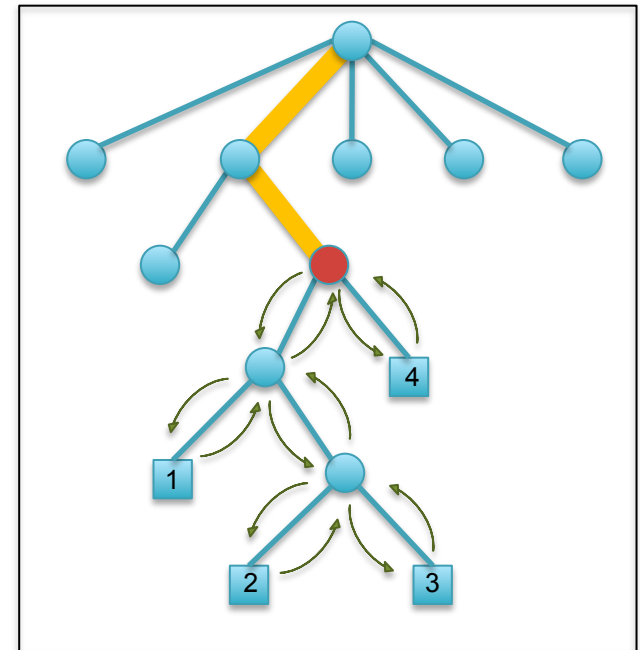


Heart Disease	—	—	—
Cancer	—	—	—
Creates magical technology	—	—	—
	—	—	—

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.

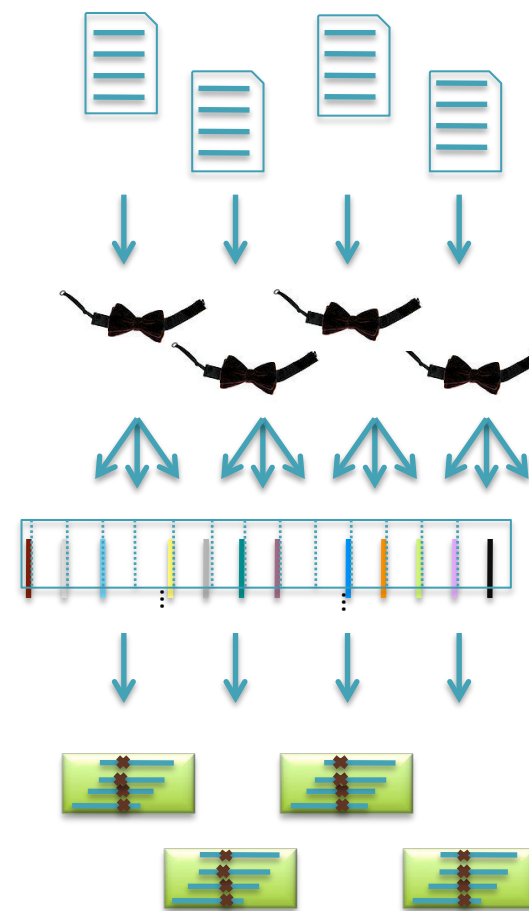


Crossbow

<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; Today's costs <\$30

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

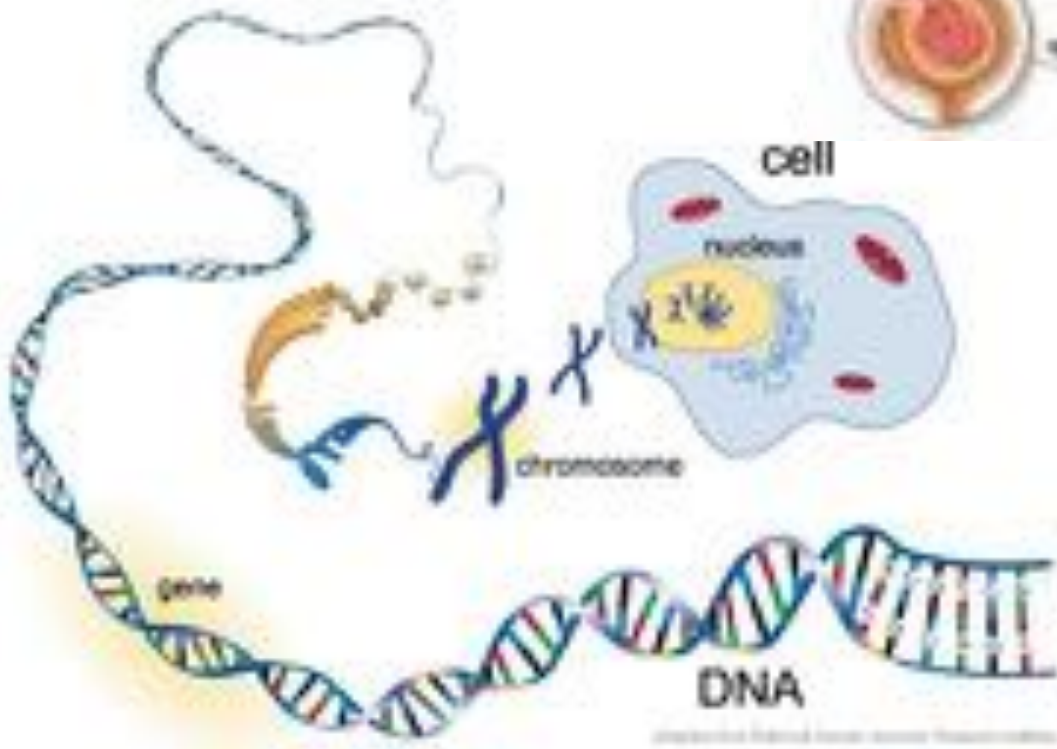
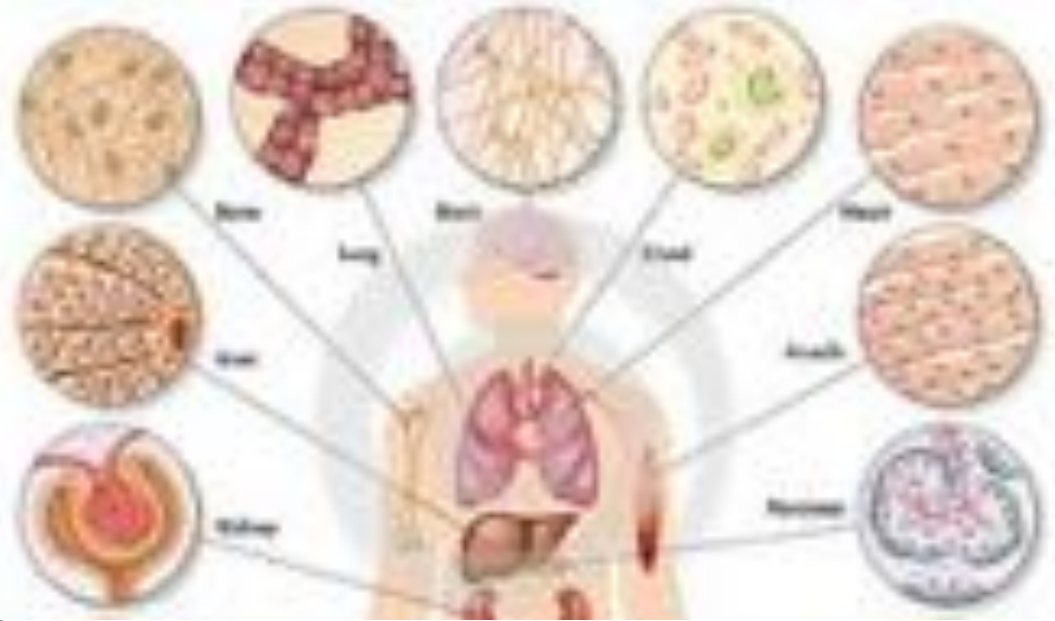


Searching for SNPs with Cloud Computing.

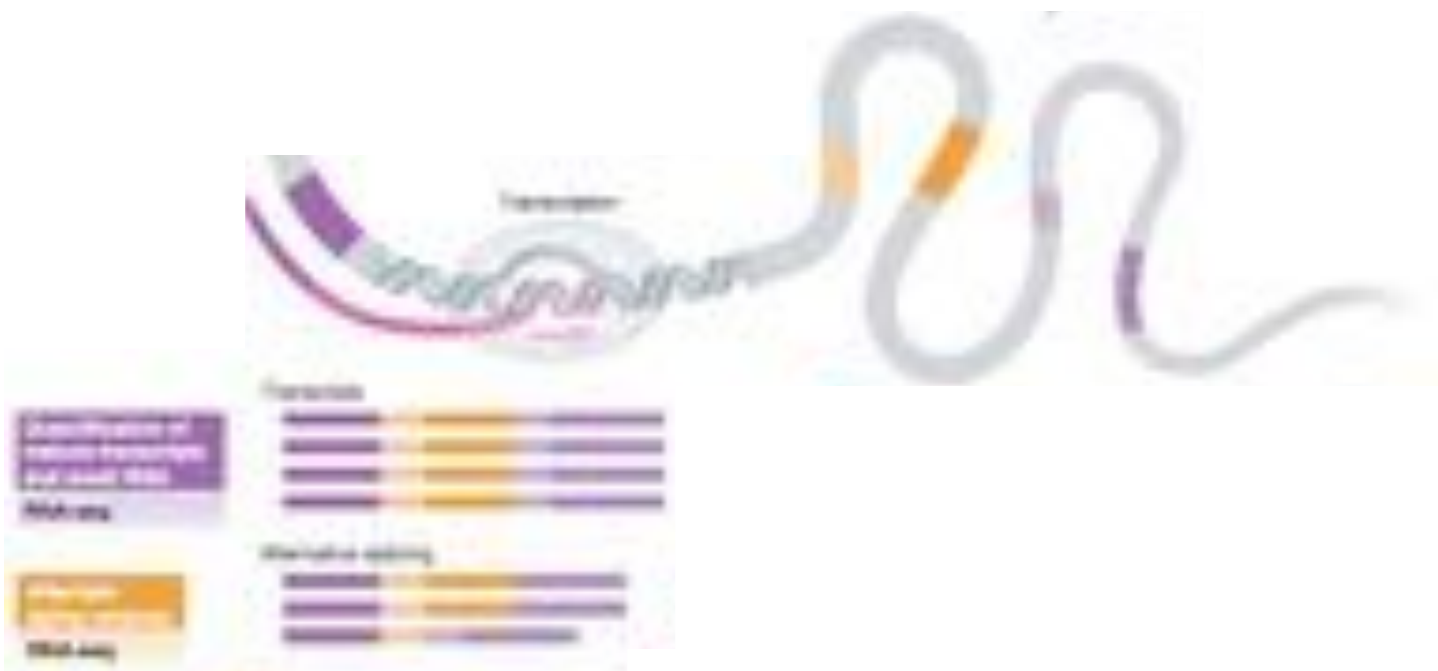
Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL (2009) *Genome Biology*. **10**:R134

Cells & DNA

Each cell of your body contains an exact copy of your 3 billion base pair genome.



Your specific nucleotide sequence encodes the genetic program for your cells and ultimately your traits



Soon et al., Molecular Systems Biology, 2013

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

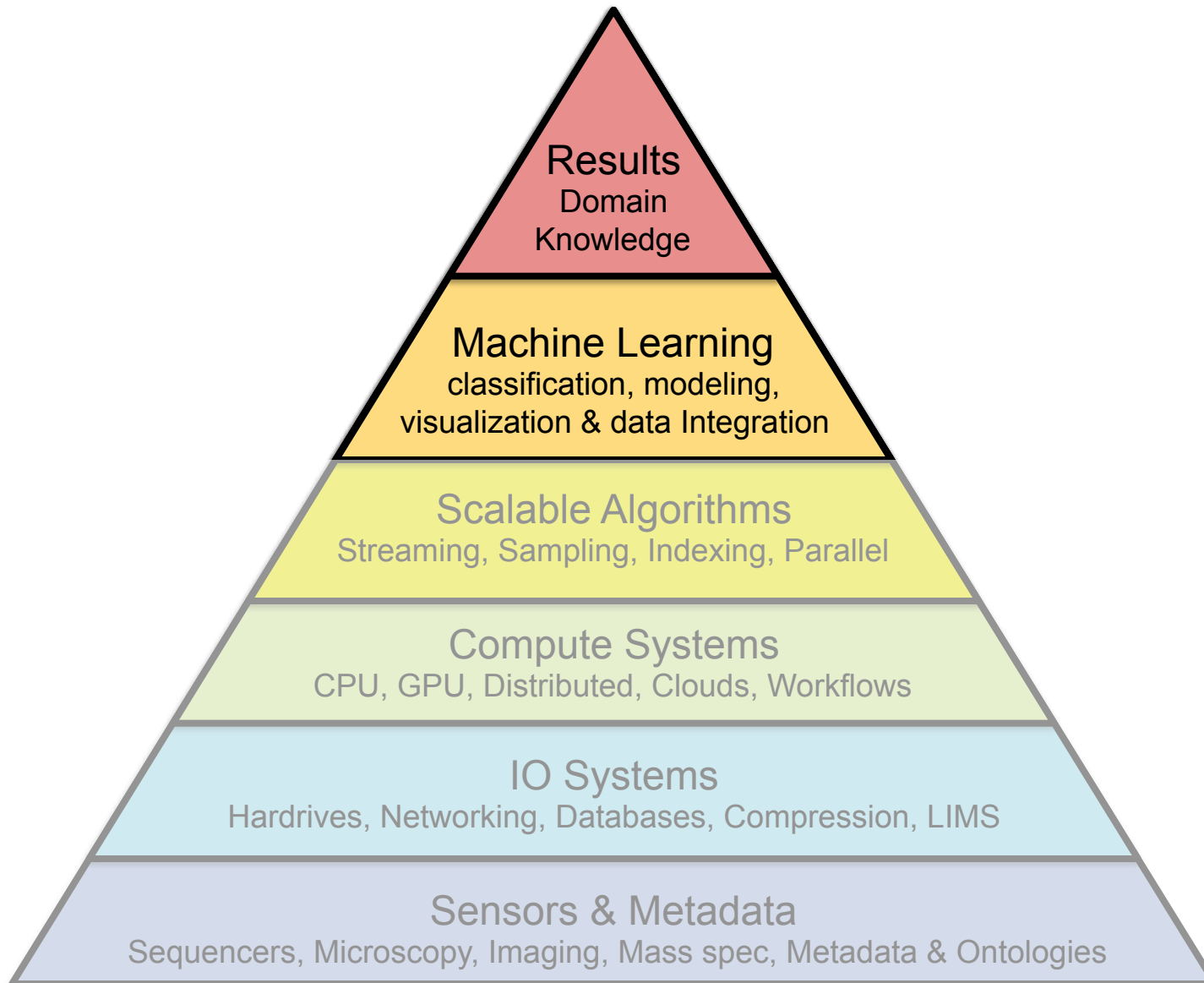
- Expect to see >1000 cores in a single computer
- Compute & IO needs to be considered together
- Rewriting efficient parallel software is complex and expensive

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



Genetic Basis of Autism Spectrum Disorders



Complex disorders of brain development

- Characterized by difficulties in social interaction, verbal and nonverbal communication and repetitive behaviors.
- Have their roots in very early brain development, and the most obvious signs of autism and symptoms of autism tend to emerge between 2 and 3 years of age.

U.S. CDC identify around 1 in 68 American children as on the autism spectrum

- Ten-fold increase in prevalence in 40 years, only partly explained by improved diagnosis and awareness.
- Studies also show that autism is four to five times more common among boys than girls.
- Specific causes remain elusive

What is Autism?

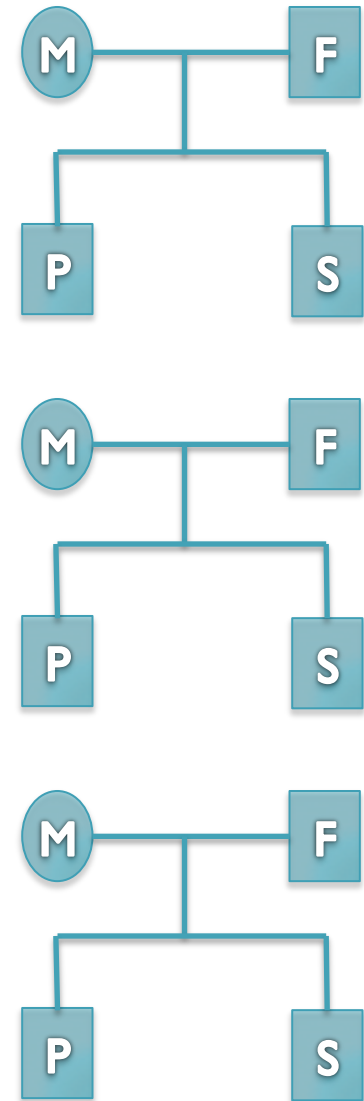
<http://www.autismspeaks.org/what-autism>

Searching for the genetic risk factors

Search Strategy

- Thousands of families identified from a dozen hospitals around the United States
- Large scale genome sequencing of “simplex” families: mother, father, affected child, unaffected sibling
- Unaffected siblings provide a natural control for environmental factors

Are there any genetic variants present in affected children, that are not in their parents or unaffected siblings?



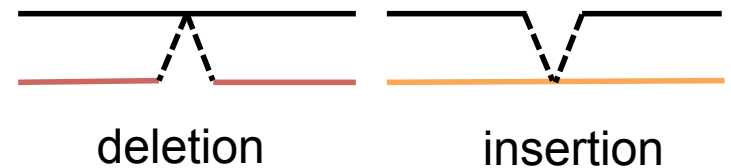
Scalpel: Haplotype Microassembly

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



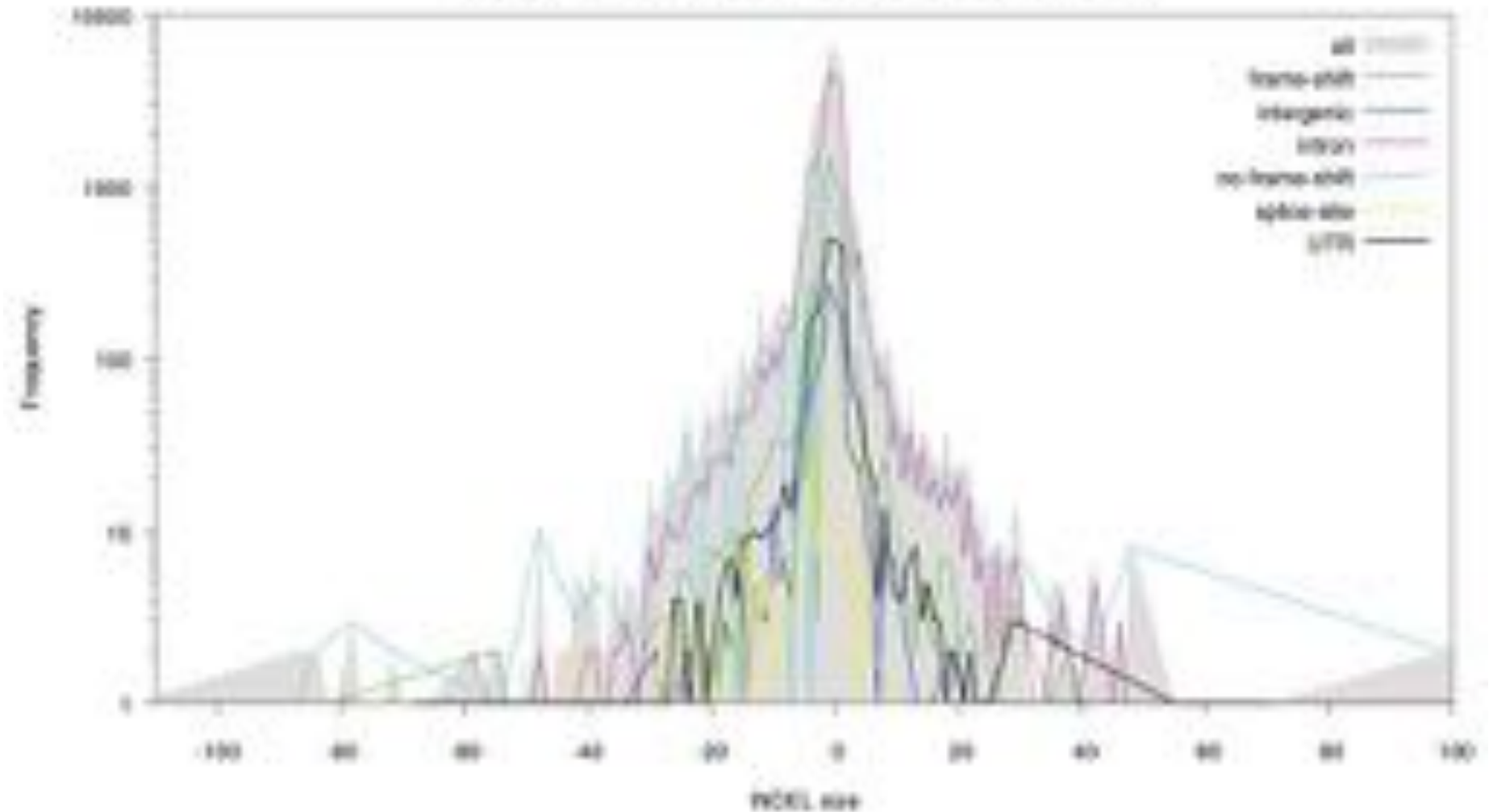
Features

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



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) *In press*

Population Analysis of the SSC

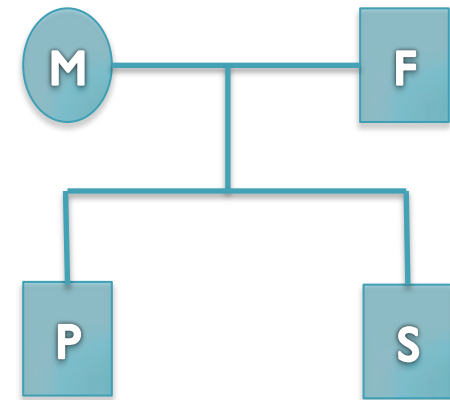


Constructed database of >1M transmitted and de novo indels

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: . . . TCAAATCCTTTTAAATAAGAAGAGCTGACA . . .

Father: . . . TCAAATCCTTTTAAATAAGAAGAGCTGACA . . .

Mother: . . . TCAAATCCTTTTAAATAAGAAGAGCTGACA . . .

Sibling: . . . TCAAATCCTTTTAAATAAGAAGAGCTGACA . . .

Proband(1): . . . TCAAATCCTTTTAAATAAGAAGAGCTGACA . . .

Proband(2): . . . TCAAATCCTTTTAAAT****AAGAGCTGACA . . .

4bp heterozygous deletion at chr15:9352406 | 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 FMR1
 - Related to neuron development and synaptic plasticity
 - Also strong overlap with chromatin remodelers

The potential for big data?

PLoS ONE | Vol 5(2) February 2010 | doi:10.1371/journal.pone.0011811

LETTERS

Detecting influenza epidemics using query data

Jeremy Ginsberg¹, Matthew J. Heuley², and David Foray³

Abstract

Influenza epidemics are a leading cause of death and illness worldwide. Early detection of disease response, can reduce the impact of influenza. One way to improve health seeking behavior is through search engines, which are submitted by millions of users around the world each day. Here we present a method of analyzing large numbers of Google search queries to track influenza like illness (ILI) that a random physician visit in a particular region is related to an influenza epidemic. This is equivalent to the percentage of ILI-related physician visits. A direct relationship was found with the probability that a random physician visit in a particular region is related to an influenza epidemic.

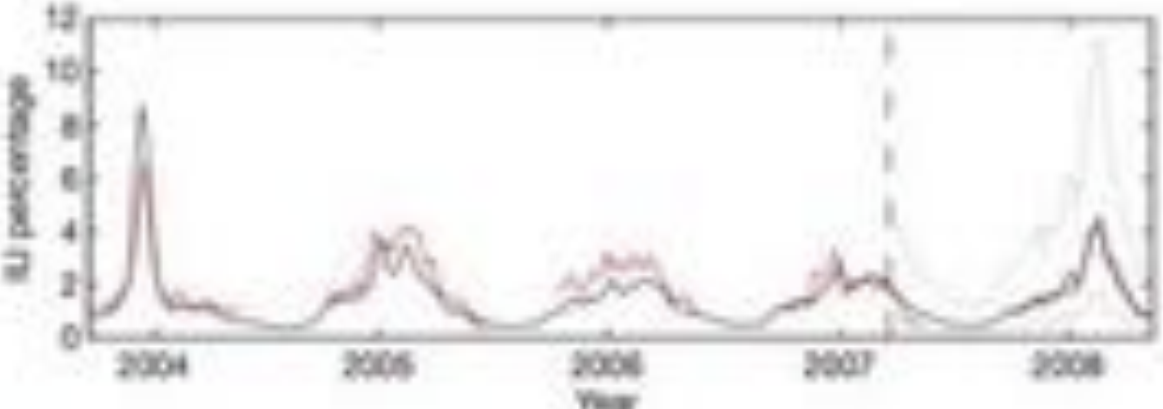


Figure 2 | A comparison of model estimates for the mid-Atlantic region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.83 was obtained over 128 points from this region to which the model was fit, whereas a correlation of 0.96 was obtained over 42 validation points. Dotted lines indicate 95% prediction intervals. The region comprises New York, New Jersey and Pennsylvania.

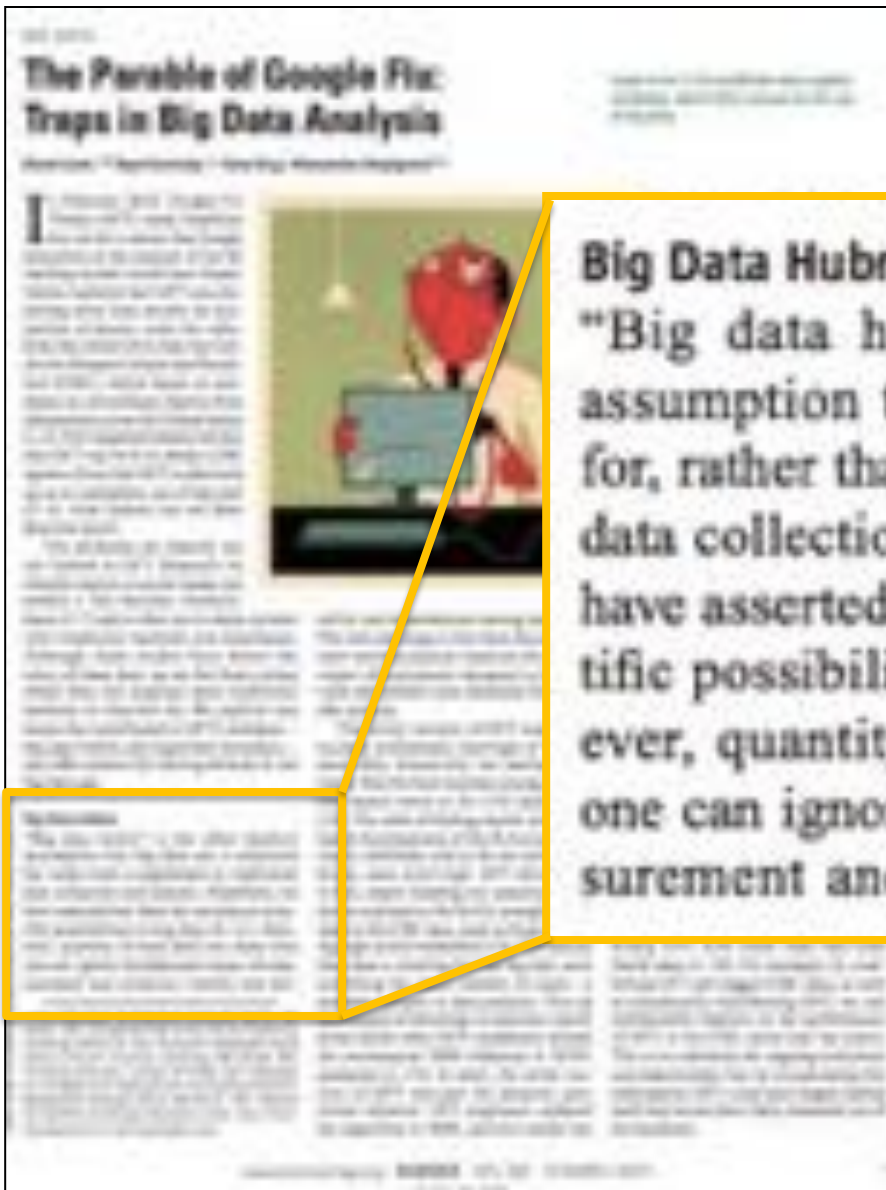
Introduction

Influenza epidemics are a leading cause of death and illness worldwide. Early detection of disease response, can reduce the impact of influenza. One way to improve health seeking behavior is through search engines, which are submitted by millions of users around the world each day. Here we present a method of analyzing large numbers of Google search queries to track influenza like illness (ILI) that a random physician visit in a particular region is related to an influenza epidemic. This is equivalent to the percentage of ILI-related physician visits. A direct relationship was found with the probability that a random physician visit in a particular region is related to an influenza epidemic.

Discussion

The results of this study suggest that search engines can be used to track influenza like illness (ILI) in a particular region. This is equivalent to the percentage of ILI-related physician visits. A direct relationship was found with the probability that a random physician visit in a particular region is related to an influenza epidemic.

The fallacy of big data?



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

Alexandry Acquaito¹ and Ralph Green

Carnegie Mellon University, Pittsburgh, PA, USA

Communicated by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, USA, 5, 2009 (received for review January 16, 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 date 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 personal information from multiple sources, such as data brokers or public records, or working sites. Our results highlight the complex interactions among sources in modern information economies and the risks associated with information revelation in

PNAS

In a modern information economy, sensitive personal data is often disseminated through multiple channels. Such is the case in the United States. Coupled with data brokers, individual records (2), they have a wide reach. Information services (2), becoming one of the most used means of identity theft. The Administration (SSA), which issues them, has a long history of providing (3), maintaining with their public records (4). After extensive work on the part of the SSA, attempts to strengthen their security and compliance (5). How has already left the SSA. The American

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.

the authors

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



Biological Data Sciences

Cold Spring Harbor Laboratory, Nov 5 - 8, 2014

Michael Schatz, Anne Carpenter, Matt Wood



10:30 Algorithms & Cloud Computing
1:00 Advances in Sequencing technology
2:00 Big Data in Biology
3:00 Microbial and Metagenomics