BIOLOGICAL DATA SCIENCE

November 5–November 8, 2014

#biodata14

Anne Carpenter

Broad Institute, @DrAnneCarpenter

Michael Schatz

Cold Spring Harbor Laboratory, @mike_schatz

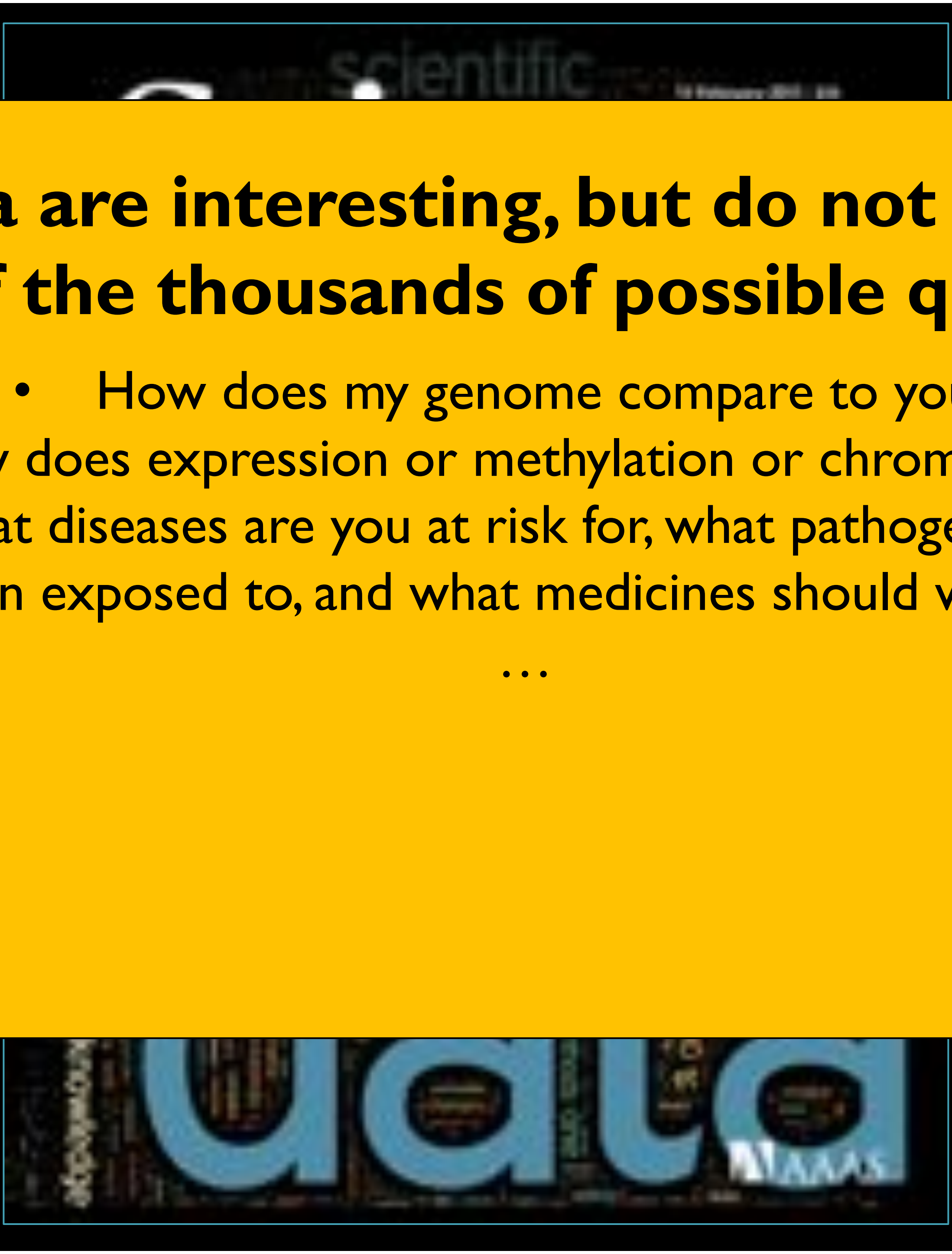
Matt Wood

Amazon Web Services, @mza

CSH Cold Spring Harbor Laboratory
MEETINGS & COURSES

@JasonWilliamsNY

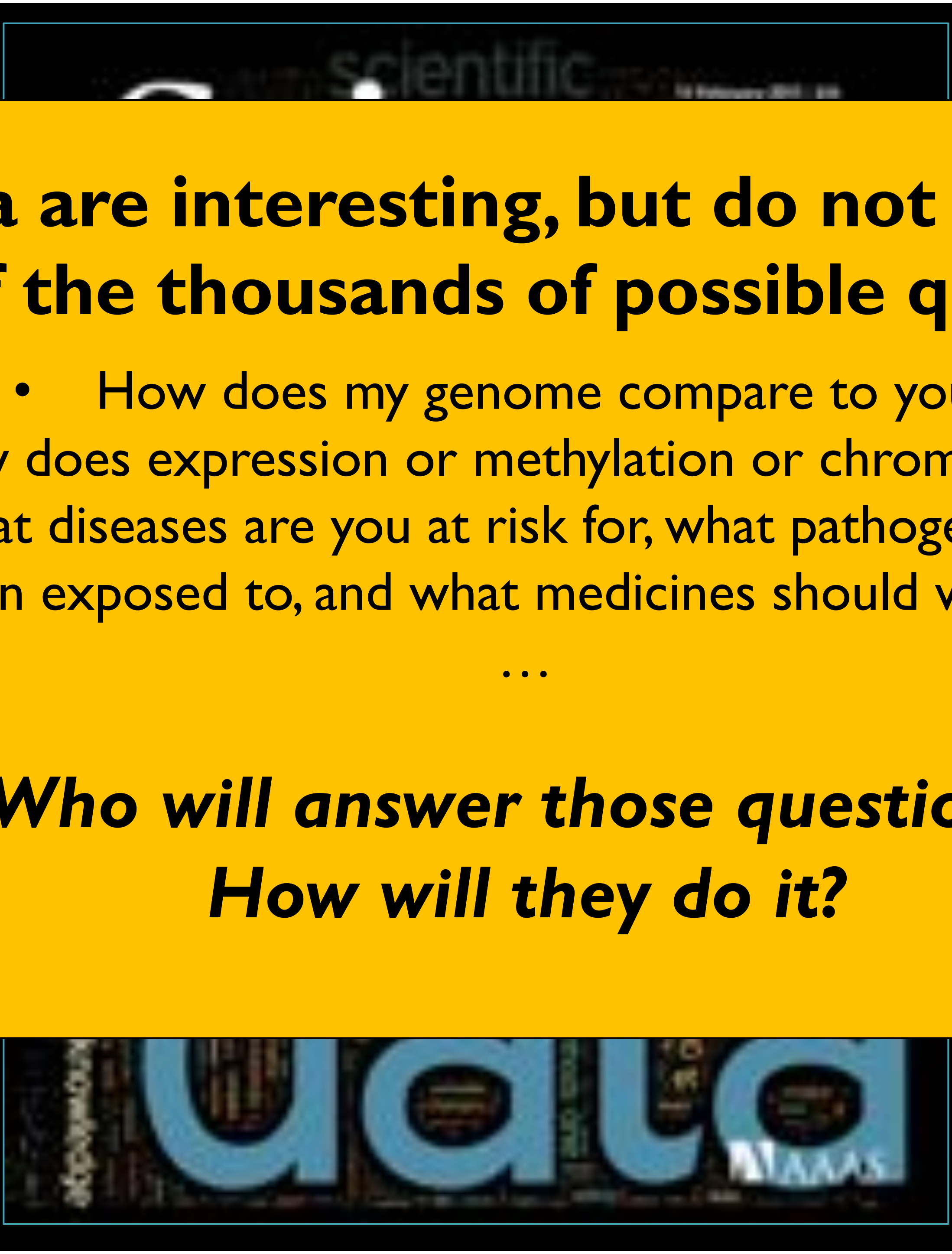
Charla Lambert



Data are interesting, but do not answer any of the thousands of possible questions:

- How does my genome compare to yours?
- How does expression or methylation or chromatin change?
- What diseases are you at risk for, what pathogens have you been exposed to, and what medicines should we give you?

…



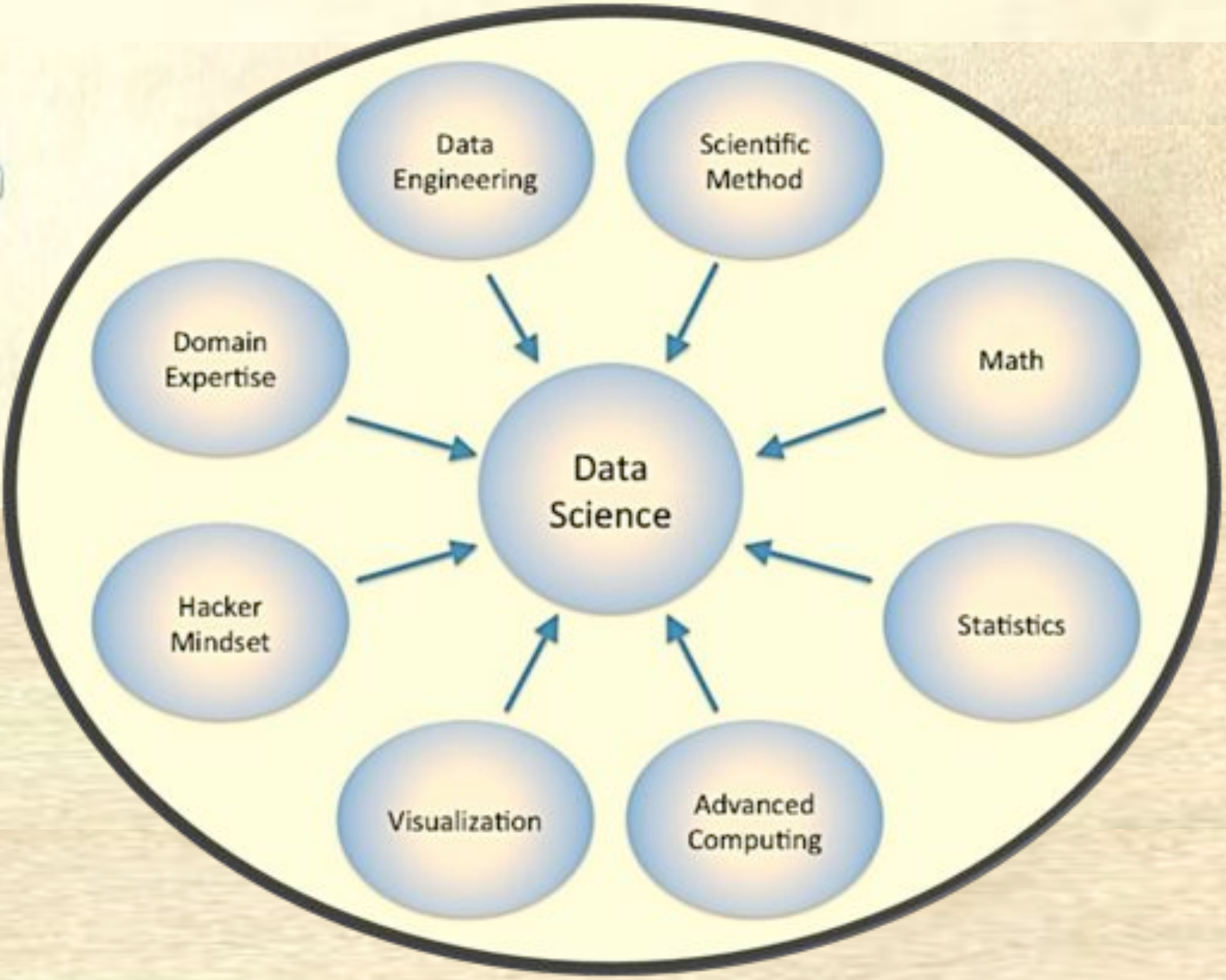
Data are interesting, but do not answer any of the thousands of possible questions:

- How does my genome compare to yours?
- How does expression or methylation or chromatin change?
- What diseases are you at risk for, what pathogens have you been exposed to, and what medicines should we give you?

…

Who will answer those questions?
How will they do it?

Who is a Data Scientist?



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

Biological Data

1 Illumina X-Ten sequences a genome every 30 minutes
~100k whole human genomes sequenced
Worldwide capacity exceeds 25 Pbp/year

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

=

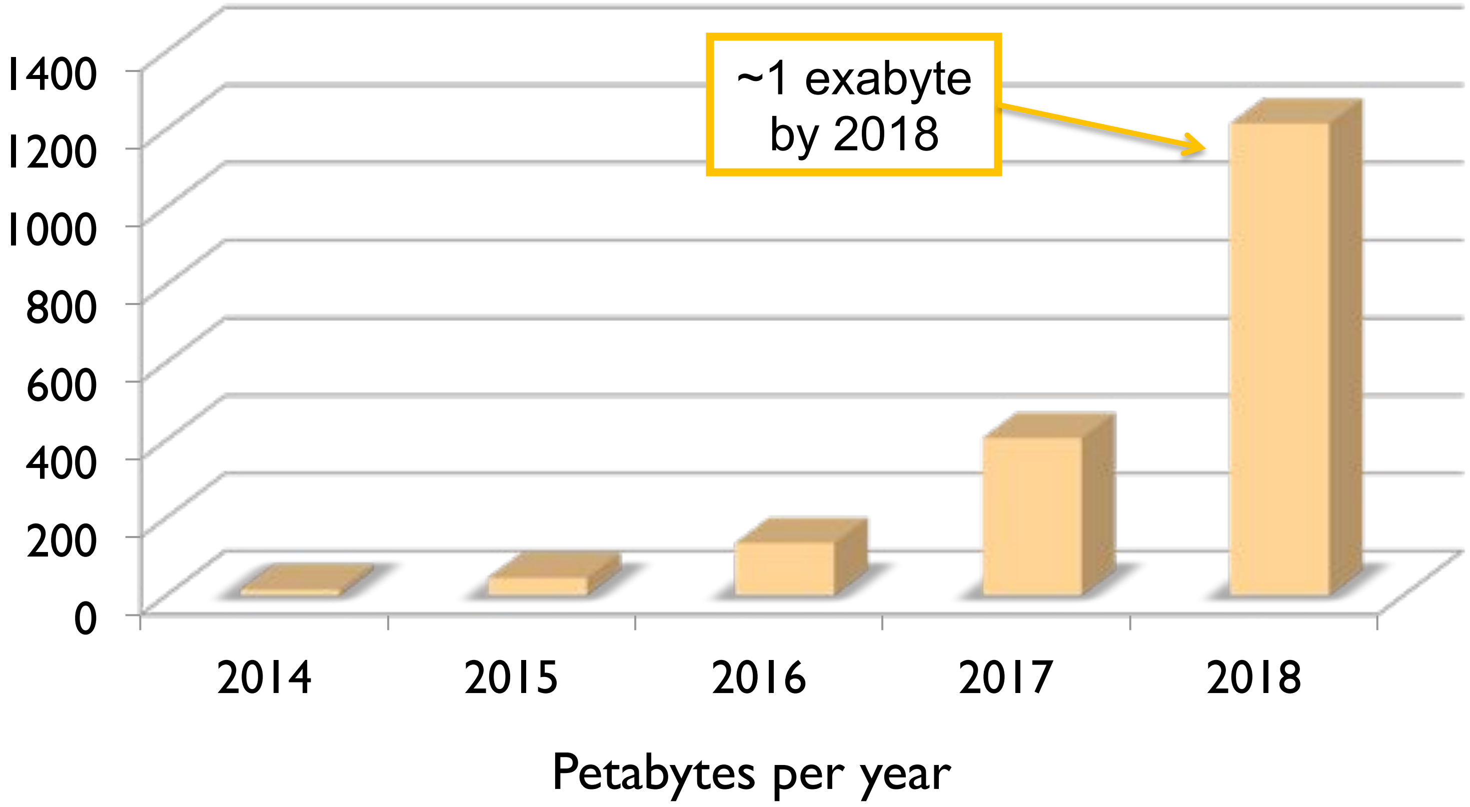
1PB Data
200,000 DVDs

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

500 2 TB drives
$500k

DNA Data Tsunami

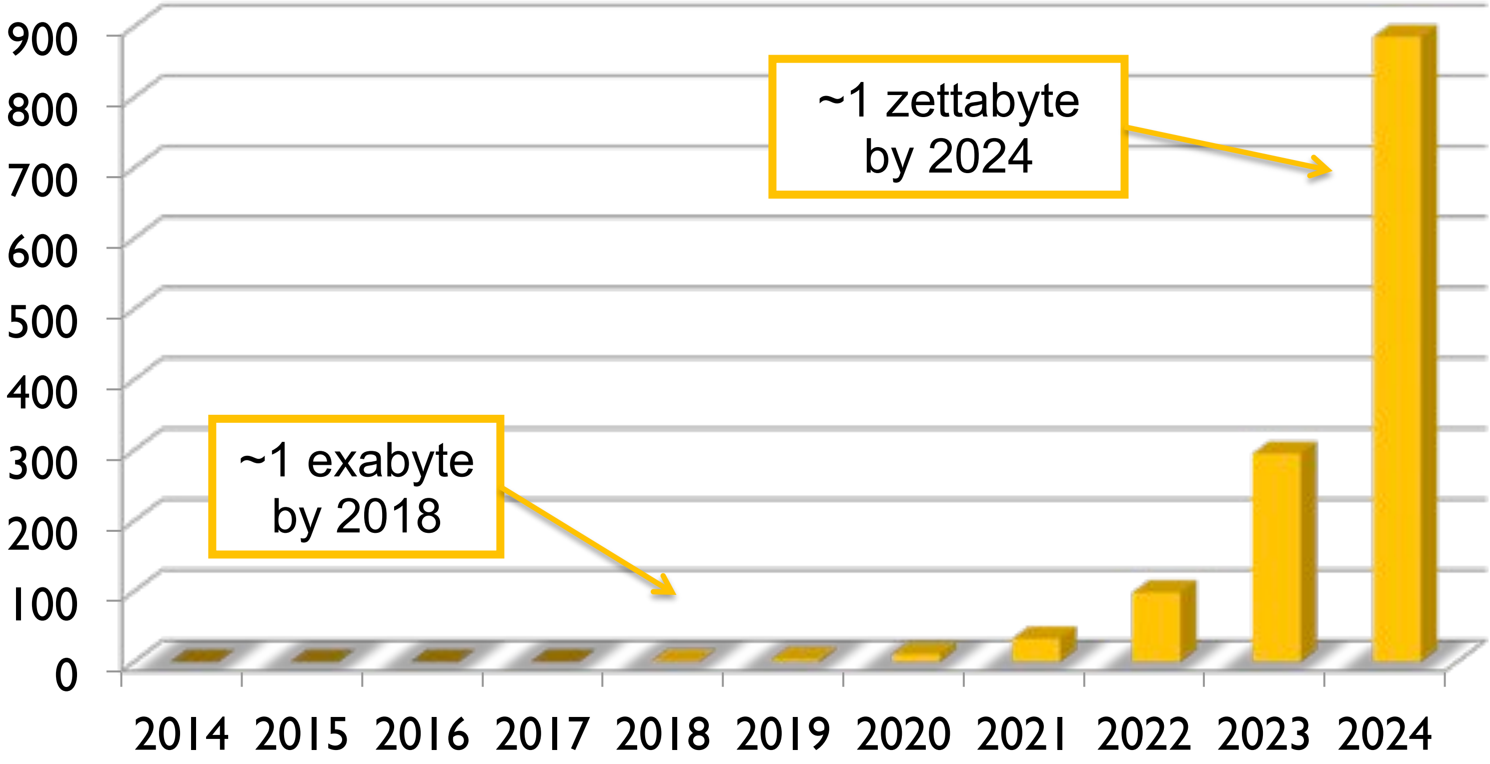
Current world-wide sequencing capacity is growing at ~3x per year!



~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

Exabytes 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

=

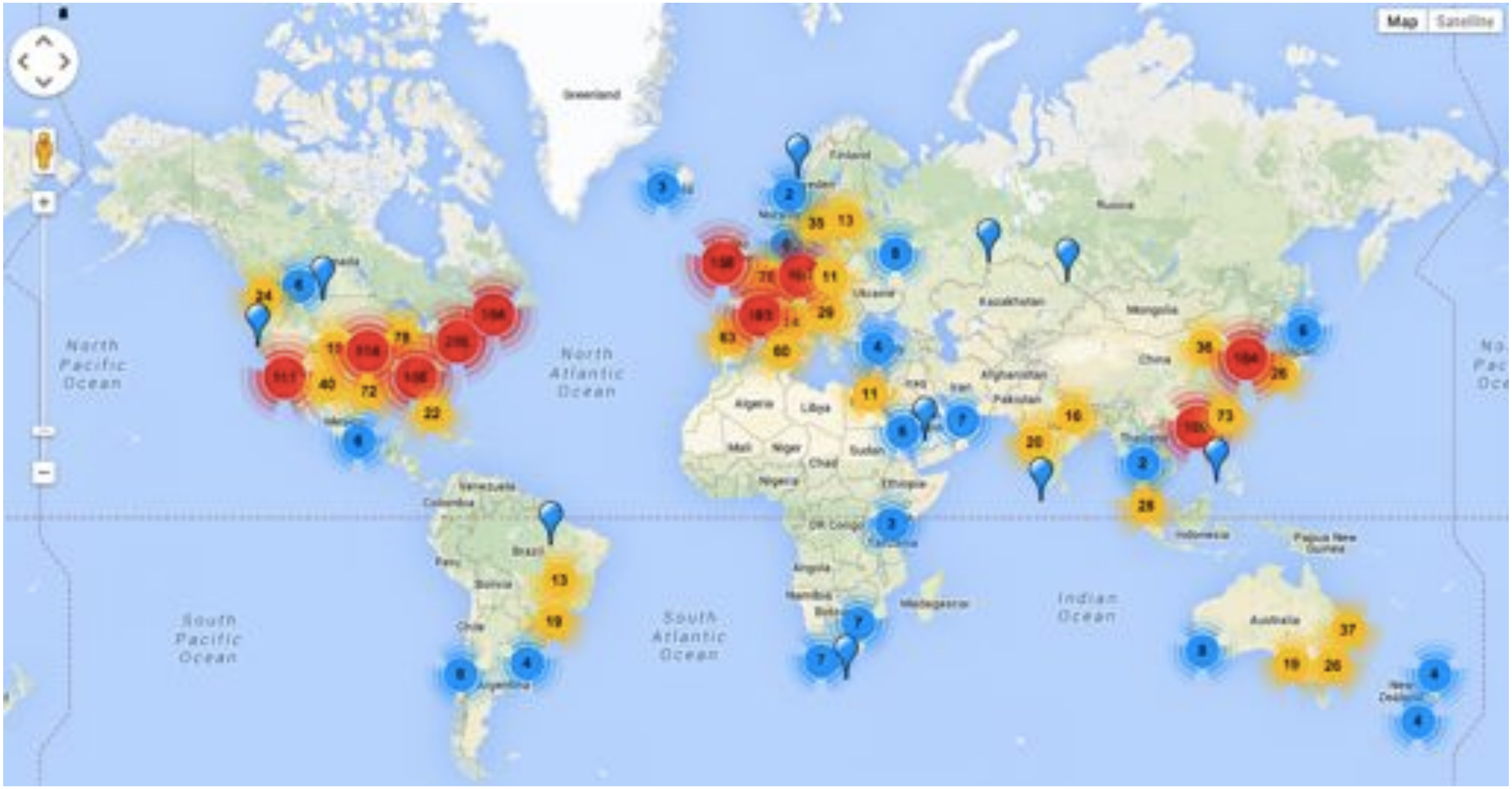
1ZB Data
200,000,000,000 DVDs

150,000 miles of DVDs
~ ½ distance to moon

YouTube

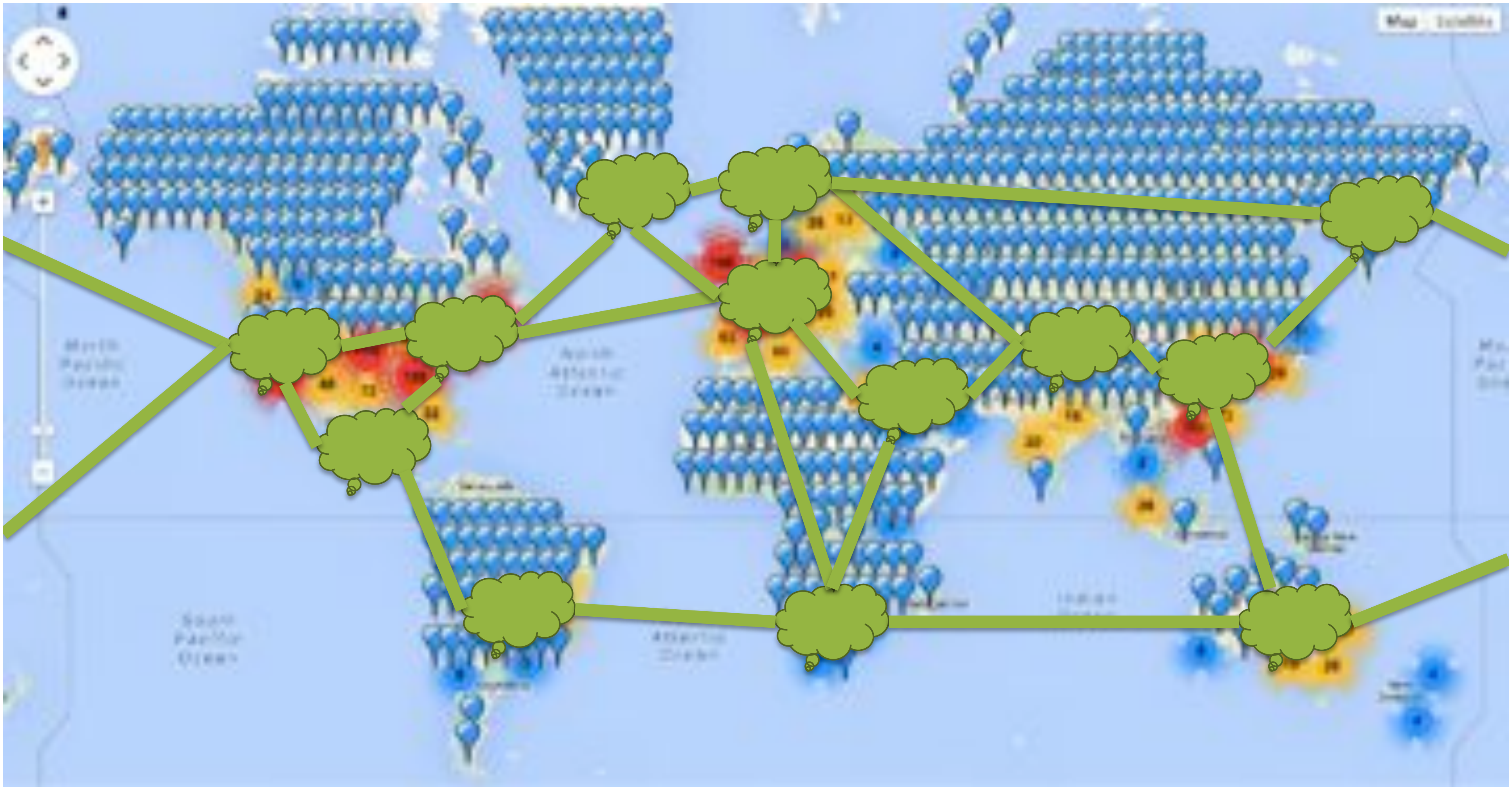
Both currently ~100Pb
And growing exponentially

Sequencing Centers 2014



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

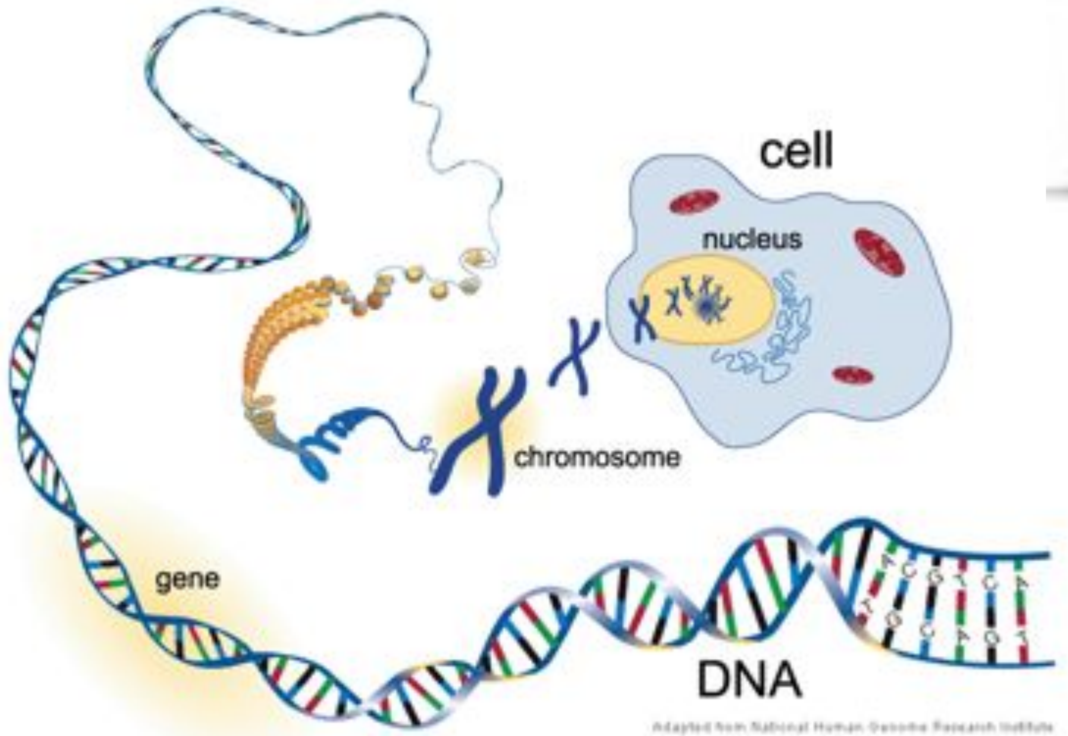
Informatics Centers 2014



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

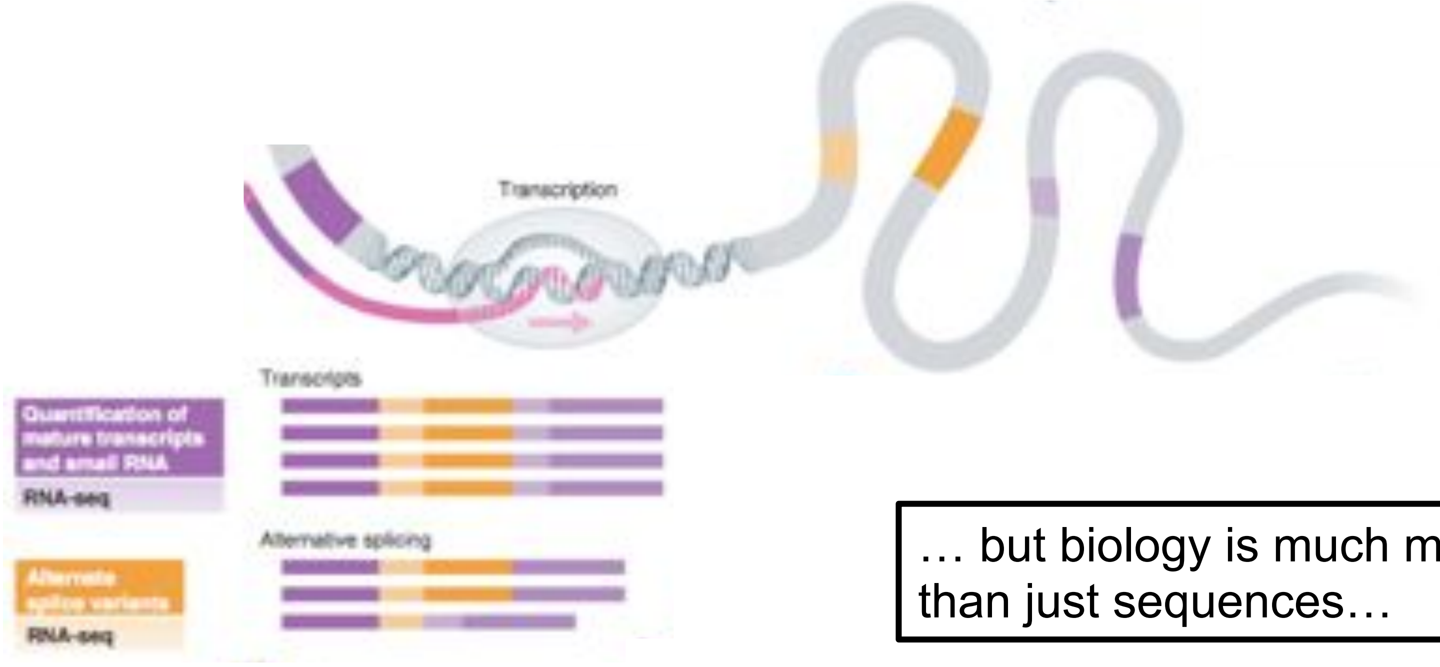
Biological Data

Much of the capacity is used to sequence genomes (or exomes) of individuals…



Adapted from National Human Genome Research Institute

… but biology is much more than just genomes…


Transcription
Transcripts
Quantification of mature transcripts and small RNA
RNA-seq
Alternative splicing
Alternate splice variants
RNA-seq

Soon et al., Molecular Systems Biology, 2013

… but biology is much more than just sequences…

Genomic

Other ‘omic

Imaging

Phenotypic

Exposure

Clinical

Complexity

Courtesy of NHGRI

Phil Bourne, Associate Director of Data Science for NIH
http://www.slideshare.net/pebourne/wiki-mania080914

Biological Data Science

ARTICLE

doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

ARTICLE

OPEN

doi:10.1038/nature12634

Mutational landscape and significance across 12 major cancer types

ARTICLE

doi:10.1038/nature11532

A map of rice genome variation reveals the origin of cultivated rice

LETTER

doi:10.1038/nature09990

A *cis*-regulatory map of the *Drosophila* genome

ARTICLE

doi:10.1038/nature13908

The contribution of *de novo* coding mutations to autism spectrum disorder

ARTICLE

doi:10.1038/nature11209

A framework for human microbiome research

The Human Microbiome Project Consortium*

© 1994 Oxford University Press Nucleic Acids Research, 1994, Vol. 22, No. 22 4673–4680

CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice

Julie D.Thompson, Desmond G.Higgins+ and Toby J.Gibson*

European Molecular Biology Laboratory, Postfach 102209, Meyerhofstrasse 1, D-69012 Heidelberg, Germany

Received July 12, 1994; Revised and Accepted September 23, 1994

ABSTRACT

NONPARAMETRIC ESTIMATION FROM INCOMPLETE OBSERVATIONS*

E. L. Kaplan

University of California Radiation Laboratory

and

Paul Meier

University of Chicago

MapReduce: Simplified Data Processing on Large Clusters

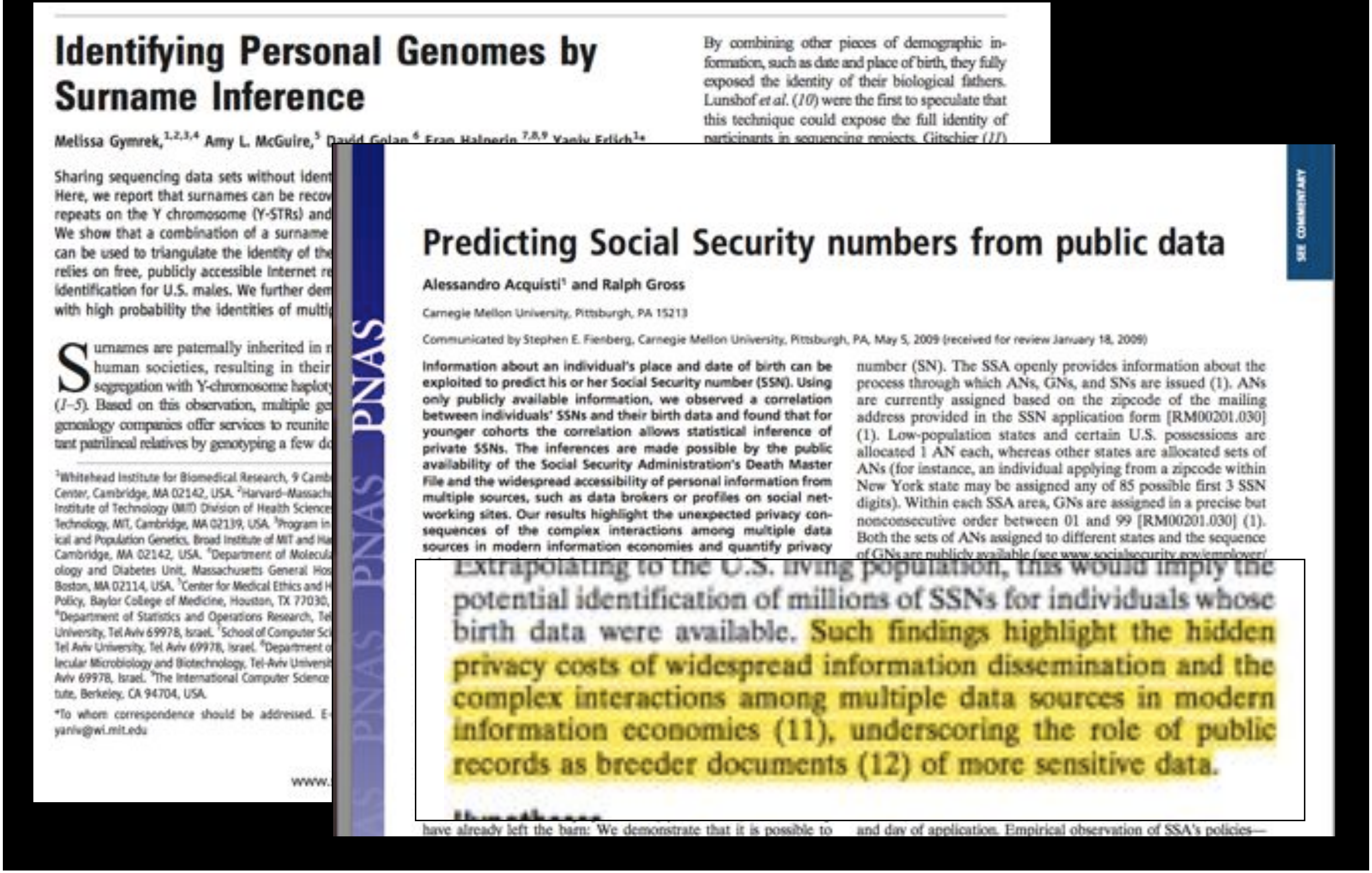
Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.

Abstract

Privacy & Security



Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

Sharing sequencing data sets without ident
Here, we report that surnames can be recov
repeats on the Y chromosome (Y-STRs) and
We show that a combination of a surname
can be used to triangulate the identity of the
relies on free, publicly accessible Internet re
identification for U.S. males. We further dem
with high probability the identities of multi

Surnames are paternally inherited in m
human societies, resulting in their
segregation with Y-chromosome haploty
(1–5). Based on this observation, multiple ge
genealogy companies offer services to reunite
tant patrilineal relatives by genotyping a few do

By combining other pieces of demographic information, such as date and place of birth, they fully exposed the identity of their biological fathers. Lunshof *et al.* (10) were the first to speculate that this technique could expose the full identity of participants in sequencing projects. Gitschier (11)

[1]Whitehead Institute for Biomedical Research, 9 Camb
Center, Cambridge, MA 02142, USA. [2]Harvard–Massach
Institute of Technology (MIT) Division of Health Science
Technology, MIT, Cambridge, MA 02139, USA. [3]Program in
ical and Population Genetics, Broad Institute of MIT and Ha
Cambridge, MA 02142, USA. [4]Department of Molecula
ology and Diabetes Unit, Massachusetts General Hos
Boston, MA 02114, USA. [5]Center for Medical Ethics and H
Policy, Baylor College of Medicine, Houston, TX 77030,
[6]Department of Statistics and Operations Research, Tel
University, Tel Aviv 69978, Israel. [7]School of Computer Sci
Tel Aviv University, Tel Aviv 69978, Israel. [8]Department o
lecular Microbiology and Biotechnology, Tel-Aviv Universit
Aviv 69978, Israel. [9]The International Computer Science
tute, Berkeley, CA 94704, USA.

*To whom correspondence should be addressed. E-
yaniv@wi.mit.edu

www.

PNAS PNAS PNAS

SEE COMMENTARY

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 personal information from multiple sources, such as data brokers or profiles on social networking sites. Our results highlight the unexpected privacy consequences of the complex interactions among multiple data sources in modern information economies and quantify privacy

have already left the barn: We demonstrate that it is possible to

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 New York state may be assigned any of 85 possible first 3 SSN digits). Within each SSA area, GNs are assigned in a precise but nonconsecutive order between 01 and 99 [RM00201.030] (1). Both the sets of ANs assigned to different states and the sequence of GNs are publicly available (see www.socialsecurity.gov/employer/

and day of application. Empirical observation of SSA's policies—

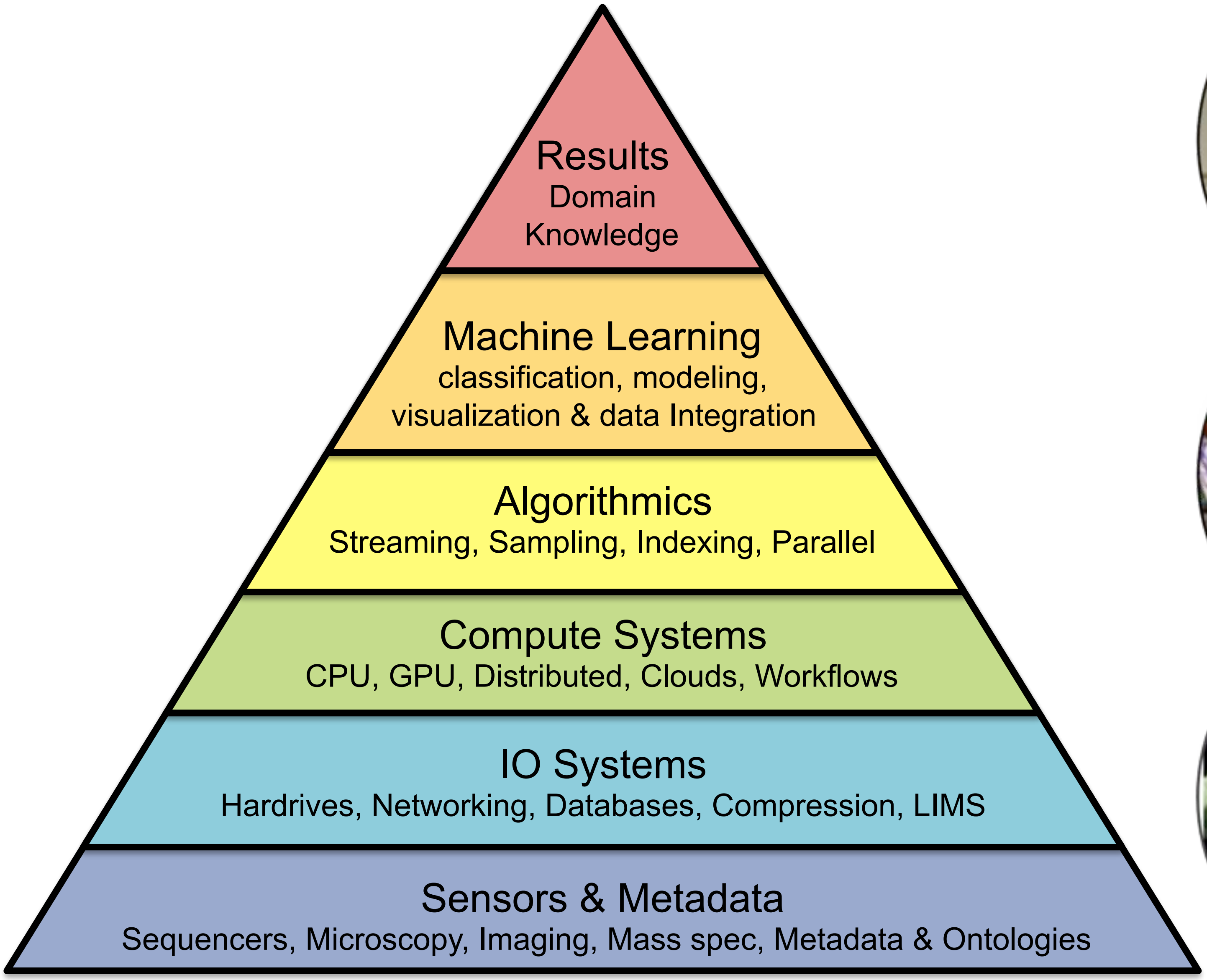
Extrapolating to the U.S. living population, this would imply the potential identification of millions of SSNs for individuals whose birth data were available. Such findings highlight the hidden privacy costs of widespread information dissemination and the complex interactions among multiple data sources in modern information economies (11), underscoring the role of public records as breeder documents (12) of more sensitive data.

Hypotheses

How?

- Integration of multiple data types
- Massively scalable
- Geographically distributed
- Computationally flexible
- Tolerate noise, errors, and artifacts
- Support data exploration and ambiguity
- Reliable, reproducible, and secure

Data Science Technologies



Results
Domain Knowledge

Machine Learning
classification, modeling, visualization & data Integration

Algorithmics
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

BIOLOGICAL DATA SCIENCE

Wednesday	7:30 pm	Introduction
	8:00 pm	**Keynote Speaker**
Thursday	9:00 am	**1** Data and Data Mining I
Thursday	1:30 pm	**2** Data and Data Mining II
Thursday	3:00 pm	**3** Poster Session I
Thursday	4:30 pm	*Wine and Cheese Party*
Thursday	7:30 pm	**4** Compute Infrastructure
Friday	9:00 am	**5** Algorithmics
Friday	1:30 pm	**6** Biological Software
Friday	4:30 pm	**Master Lecture**
Friday	5:30 pm	**7** Poster Session II and Cocktails
Friday	7:00 pm	Banquet
Saturday	9:00 am	**8** Human Biology

Master Lecture

Kristin Lauter, Ph.D.
Microsoft Research

“Homomorphic encryption as a tool to preserve privacy in genomic computation”

Friday @ 4:30pm

Schedule Change

Eric Perakslis, Ph.D.
Harvard Medical School

Saturday Morning: Human Biology

Mark Gerstein will present first in the session

Plan to break for lunch at 11:40am instead of noon

Keynote Introduction

Ph.D. in CS from the Univ. of Colorado at Boulder in 1982

Member of the NAS and the American Academy of Arts and Sciences; Fellow of AAAS and AAAI

Research combines mathematics, computer science, and molecular biology

- Pioneered the use of HMMs and other machine learning techniques for analyzing biological sequences
- Major efforts in the human genome project, and developing the UCSC Genome Browser
- Recently focused on understanding and fighting cancer; sharing of data through the Global Alliance for Genomics and Health

David Haussler, Ph.D.
Distinguished Professor of Biomolecular Engineering at UCSC
Investigator, Howard Hughes Medical Institute
Scientific Director, UC Santa Cruz Genomics Institute

Thank you!

@mike_schatz / #biodata14