

---

# 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*



**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](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

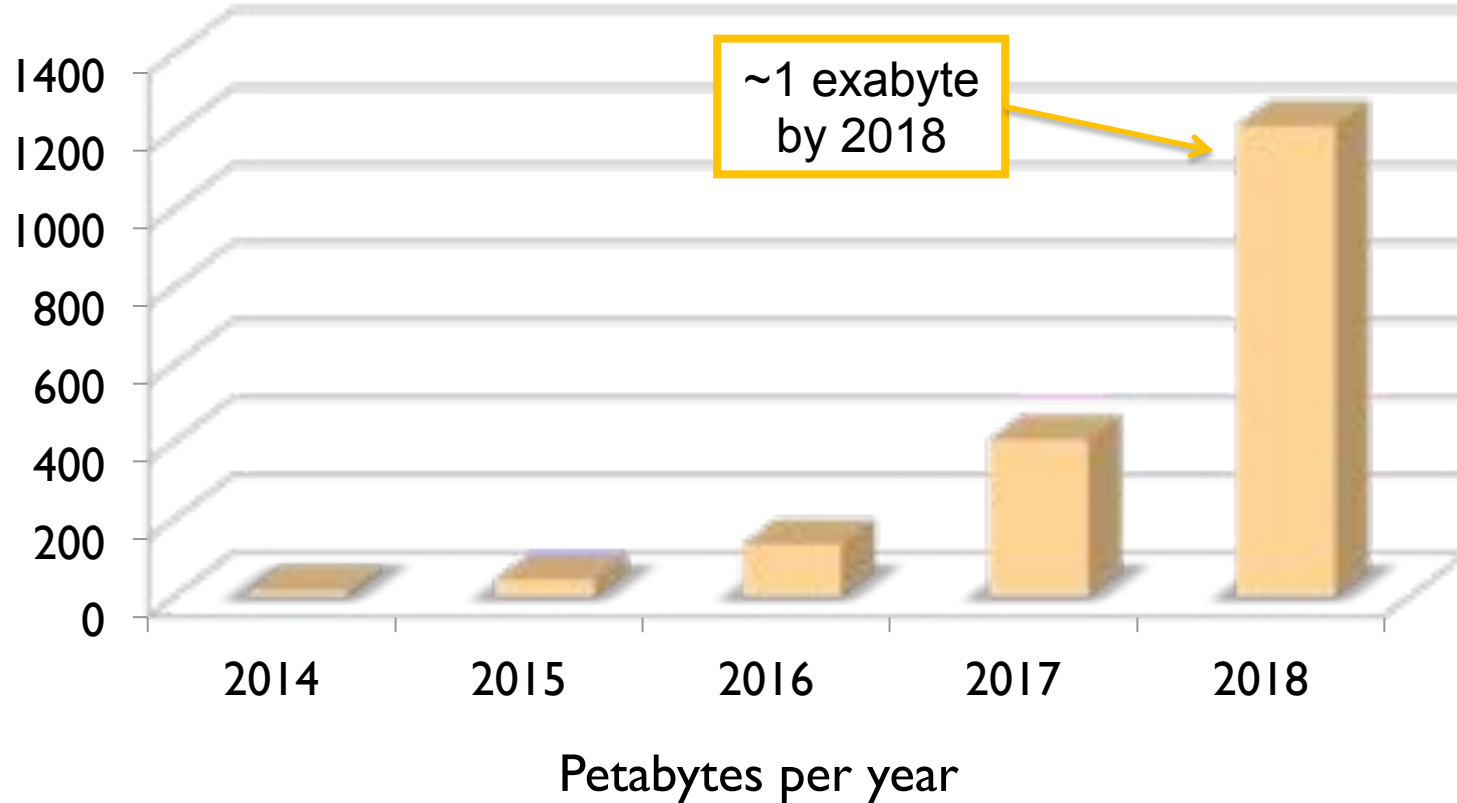


500 2 TB drives  
\$500k



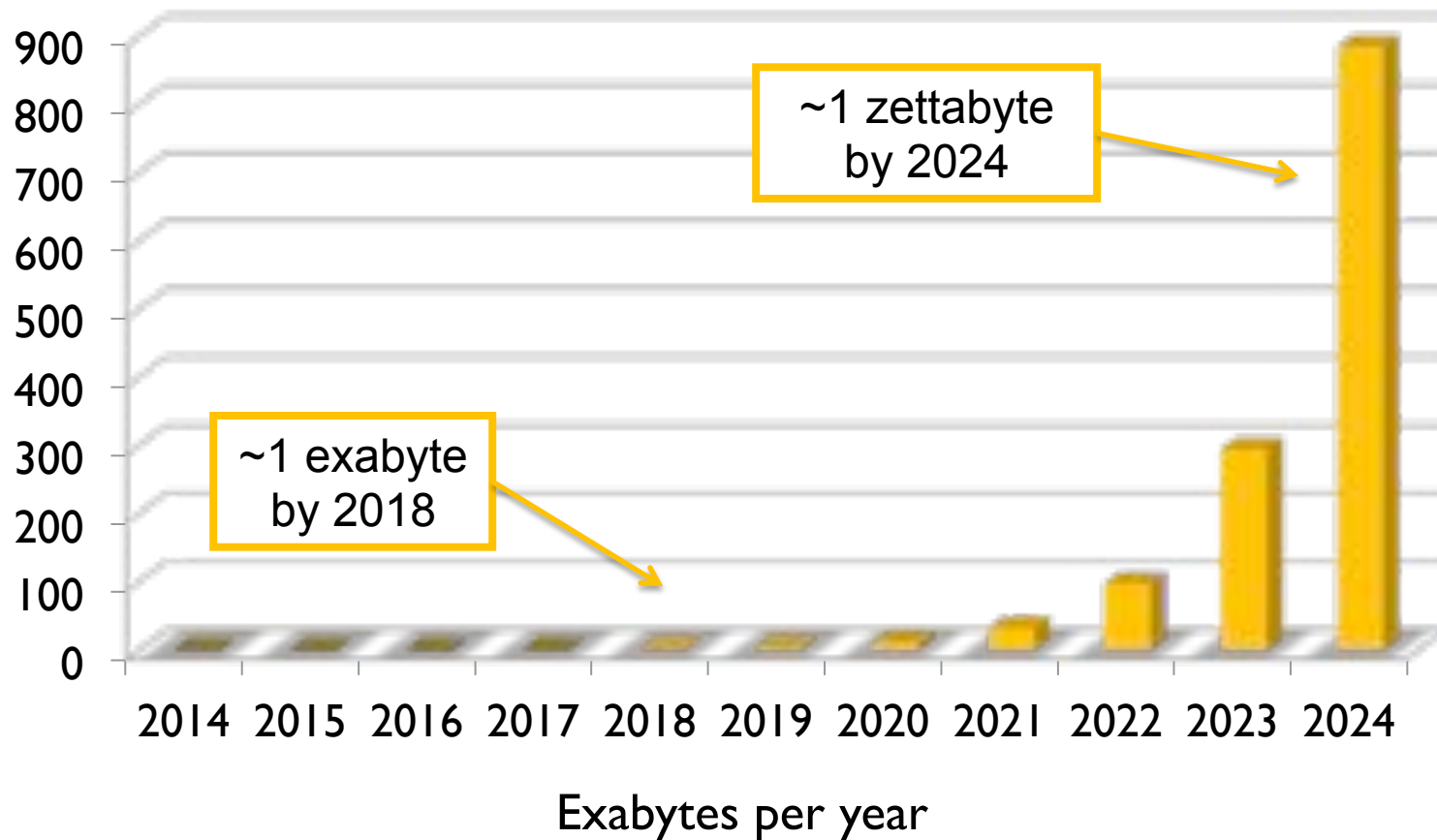
# DNA Data Tsunami

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



# 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?



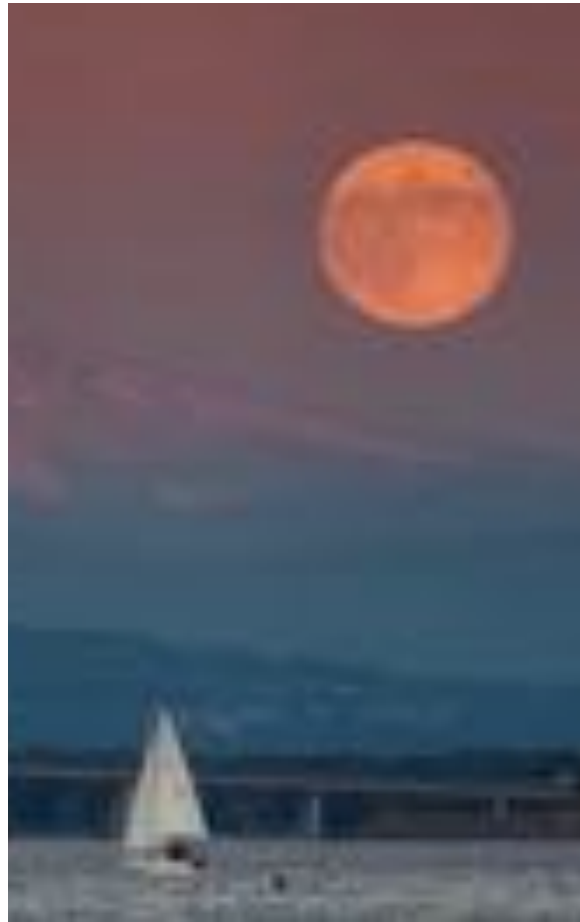
100 GB / Genome  
4.7GB / DVD  
~20 DVDs / Genome

X

10,000,000,000 Genomes

=

1ZB Data  
200,000,000,000 DVDs



150,000 miles of DVDs  
~ 1/2 distance to moon



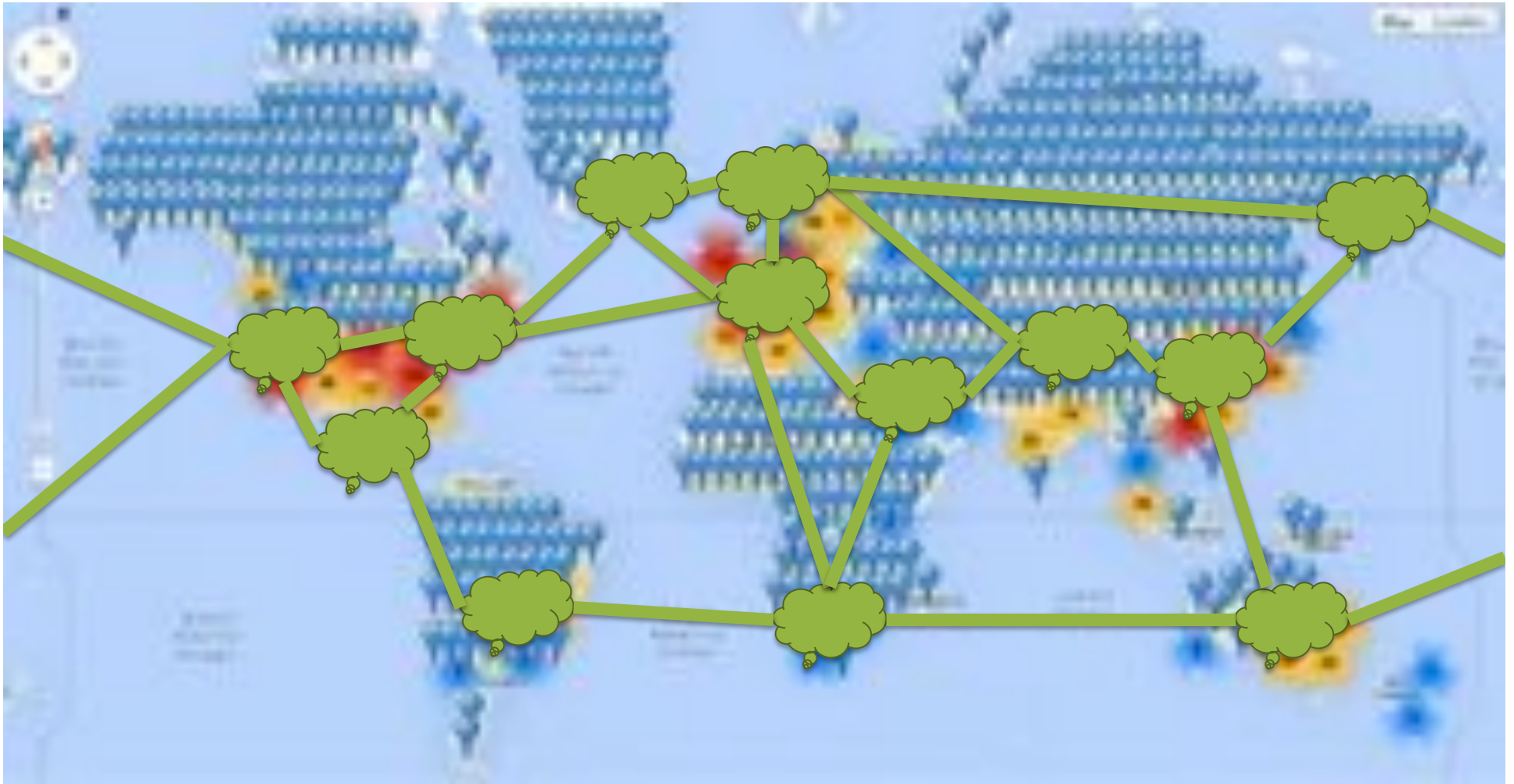
Both currently ~100Pb  
And growing exponentially

# Sequencing Centers 2014



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

# Informatics Centers 2014

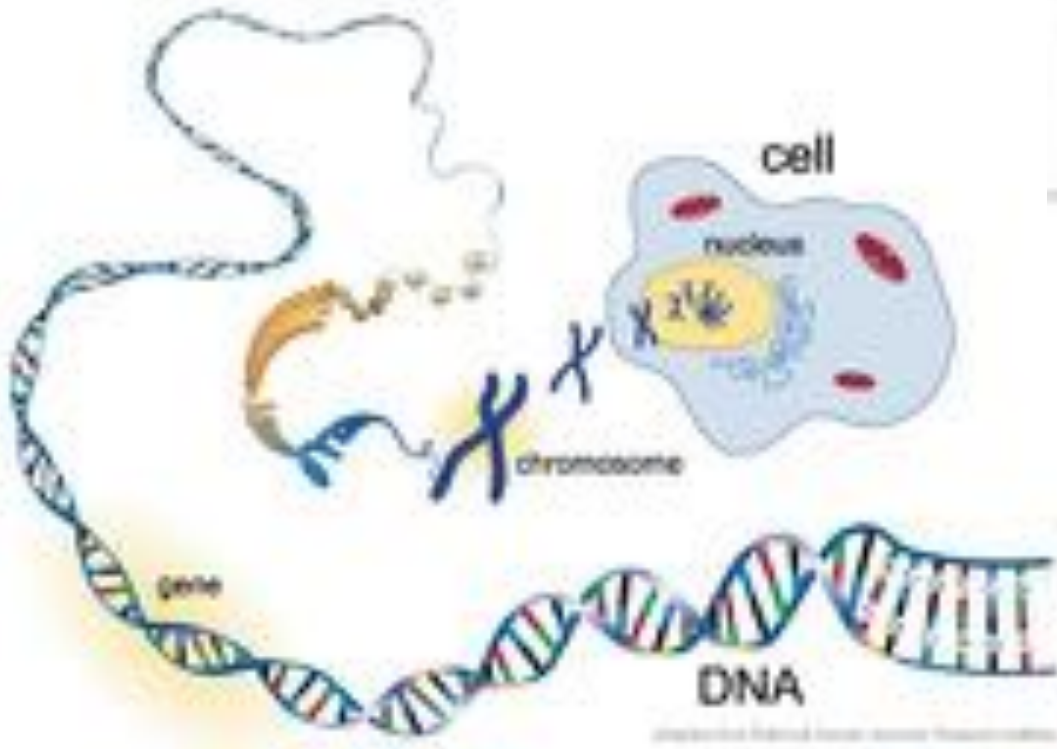


## ***The DNA Data Deluge***

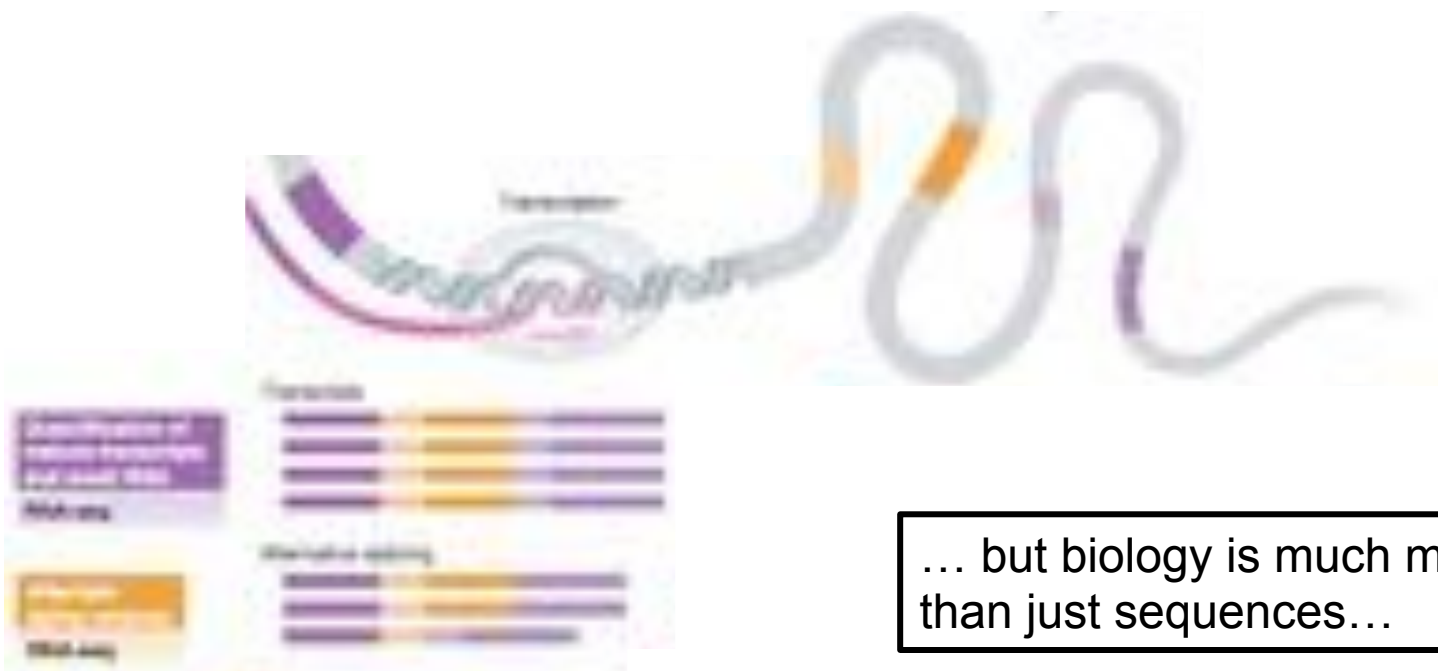
Schatz, MC and Langmead, B (2013) *IEEE Spectrum*. July, 2013

# Biological Data

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



... but biology is much more than just genomes...



... 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>



# Privacy & Security

## Identifying Personal Genomes by Surname Inference

Melissa Gymrek,<sup>1,2,3,4</sup> Amy L. McGuire,<sup>5</sup> David Faloutsos,<sup>6</sup> Evan Haber,<sup>7,8</sup> Paul Fieguth,<sup>9</sup> and George V. Panagiotou<sup>1,2,3,4</sup>

Sharing sequencing data sets without identifiers, we report that surnames can be recovered from the Y chromosome (Y-DNA) and we show that a combination of 4 surnames can be used to triangulate the identity of the males on free, publicly accessible Internet identification for U.S. males. We further show with high probability the identities of males

**S**urnames are generally inherited in a human population, resulting in their segregation with Y-chromosomal haplotypes (Y-DNA). Based on this observation, multiple genealogy websites offer services to locate and partition relatives by generating a list of

<sup>1</sup>Wellcome Institute for Human Genetics, 100 Brookings Drive, Cambridge MA 02142, USA; <sup>2</sup>Harvard-MIT Center for Technology, 320 Stovessandt Hall, Cambridge, MA 02139, USA; <sup>3</sup>Harvard School of Public Health, 665 Huntington Ave, Boston, MA 02115, USA; <sup>4</sup>Harvard University, 77 Avenue Louis Pasteur, Cambridge, MA 02138, USA; <sup>5</sup>Department of Statistics, MIT, Cambridge, MA 02139, USA; <sup>6</sup>Department of Computer Science, MIT, Cambridge, MA 02139, USA; <sup>7</sup>Department of Computer Science, MIT, Cambridge, MA 02139, USA; <sup>8</sup>Department of Computer Science, MIT, Cambridge, MA 02139, USA; <sup>9</sup>Department of Computer Science, MIT, Cambridge, MA 02139, USA.

The authors' contributions should be addressed to g.panagiotou@mit.edu.

By combining other pieces of demographic information, such as sex and place of birth, they fully exposed the identity of their biological fathers. Lander et al. (17) were the first to speculate that this technique could expose the full identity of individuals in anonymous datasets (October 1977)

## Predicting Social Security numbers from public data

Alexandru Argente<sup>1</sup> 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 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 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 availability 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

number (SN). The USA agency provides information about the process through which ANs, CNs, and DNs are issued (1). ANs are currently assigned based on the zip code of the mailing address provided in the SSN application form (SSNEXPLAIN) (2). Low-population states and certain U.S. possessions are allowed 1 AN each, whereas other states are allocated sets of ANs (for instance, an individual applying from a zip code within New York state may be assigned any of 10 possible first 3 SN digits). Within each SSA area, CNs are assigned in a strictly but nonconsecutive order between 00 and 99 (SSNEXPLAIN) (3). SSNs for sets of ANs assigned to different states and the assigned CNs are publicly available (see www.socialsecurity.gov/numbers).

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.

### Introduction

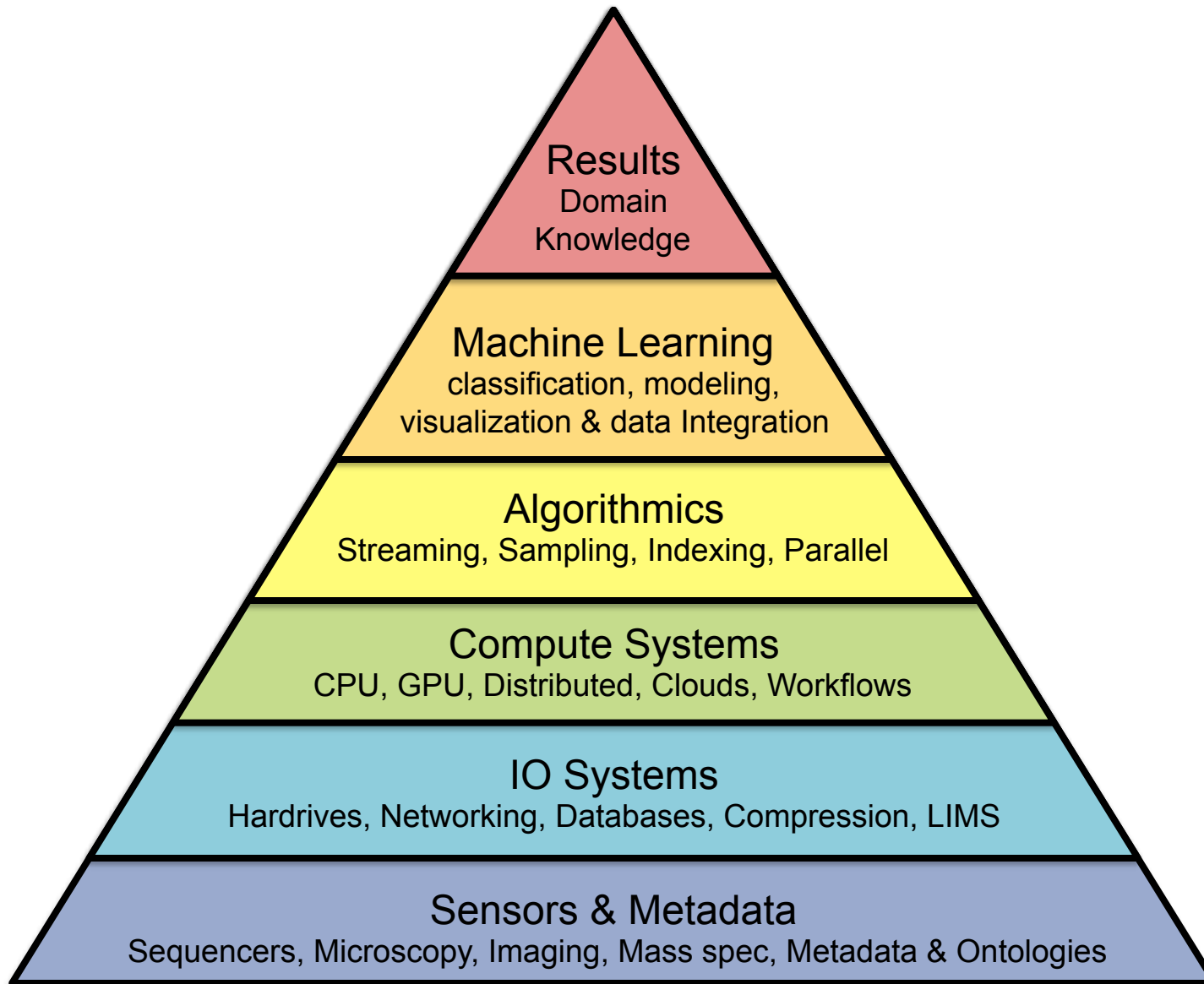
Information about an individual's place and date of birth can be exploited to predict his or her Social Security number (SSN) using only publicly available information. We observed a correlation between individuals' SSNs and their birth data and found that for younger cohorts the correlation allows statistical 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 availability 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

# 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



# BIOLOGICAL DATA SCIENCE



Wednesday	7:30 pm	Introduction
	8:00 pm	<i>Keynote Speaker</i>
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	<i>Wine and Cheese Party</i>
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 AAI**

**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