# BIOLOGICAL DATA SCIENCE

November 5–November 8, 2014 #biodata14


Anne Carpenter

*Broad Institute, @DrAnneCarpenter*

Michael Schatz

*Cold Spring Harbor Laboratory, @mike_schatz*

Matt Wood

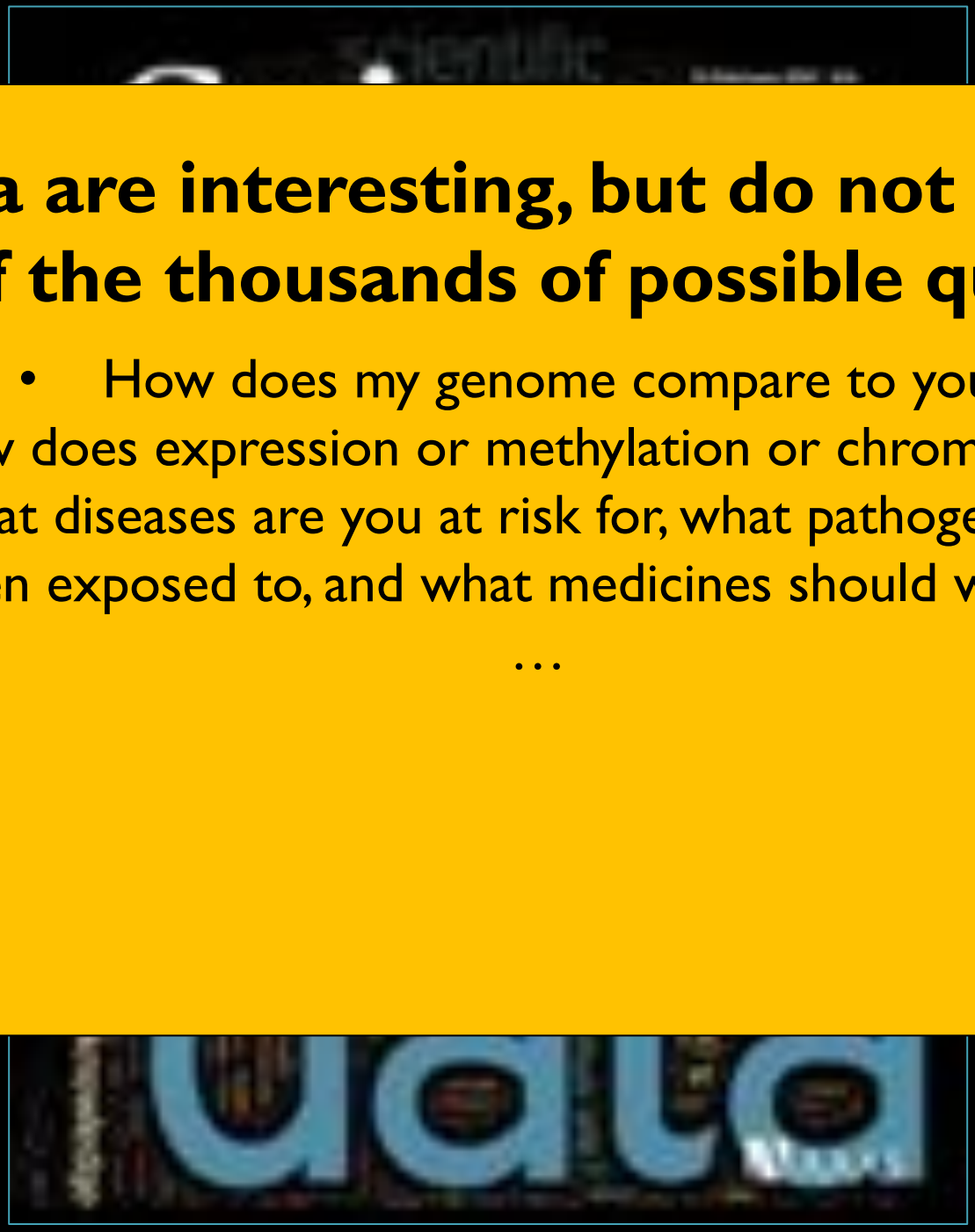
*Amazon Web Services, @mza*



Cold Spring Harbor Laboratory
MEETINGS & COURSES

@JasonWilliamsNY

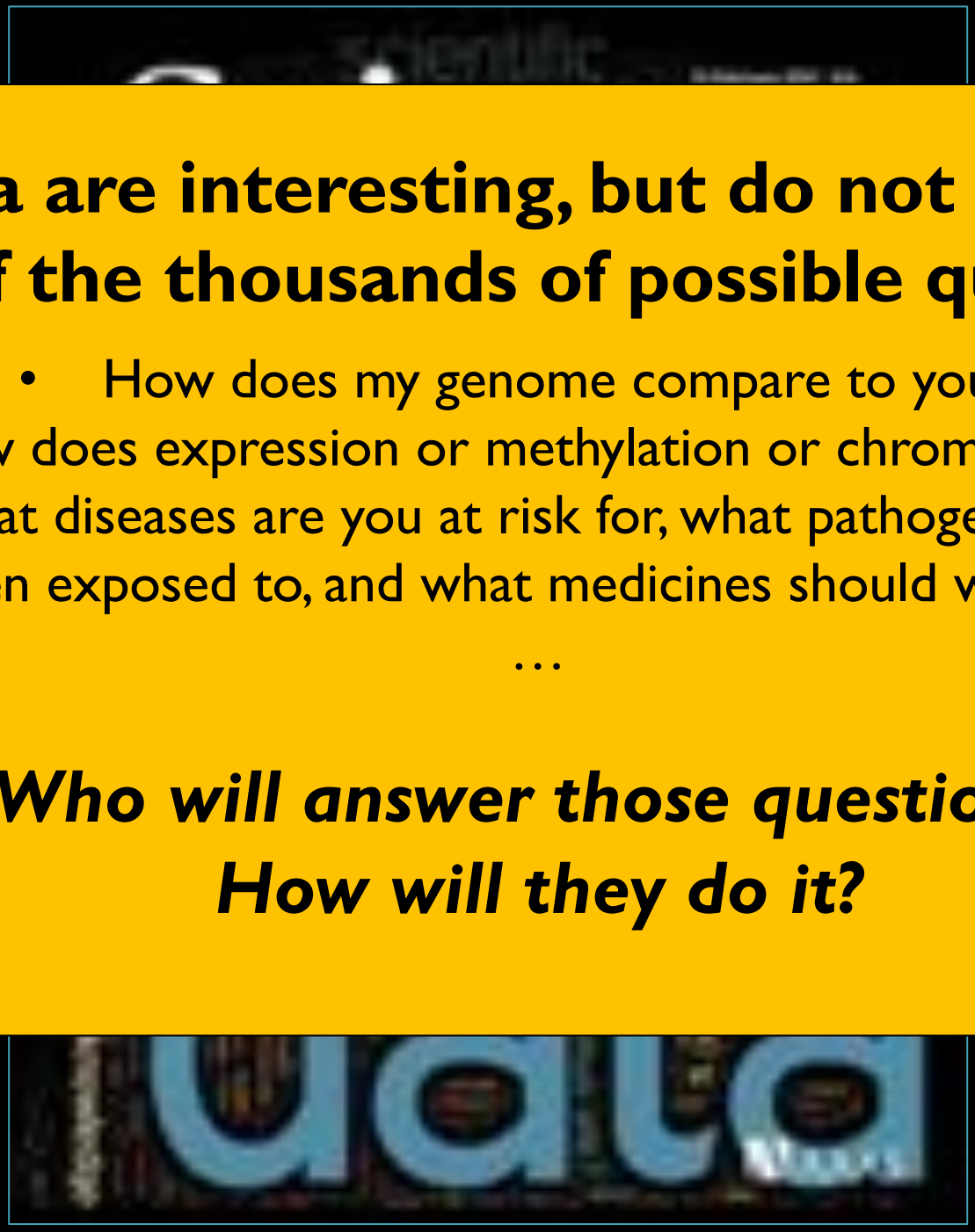
Charla Lambert




**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?

…



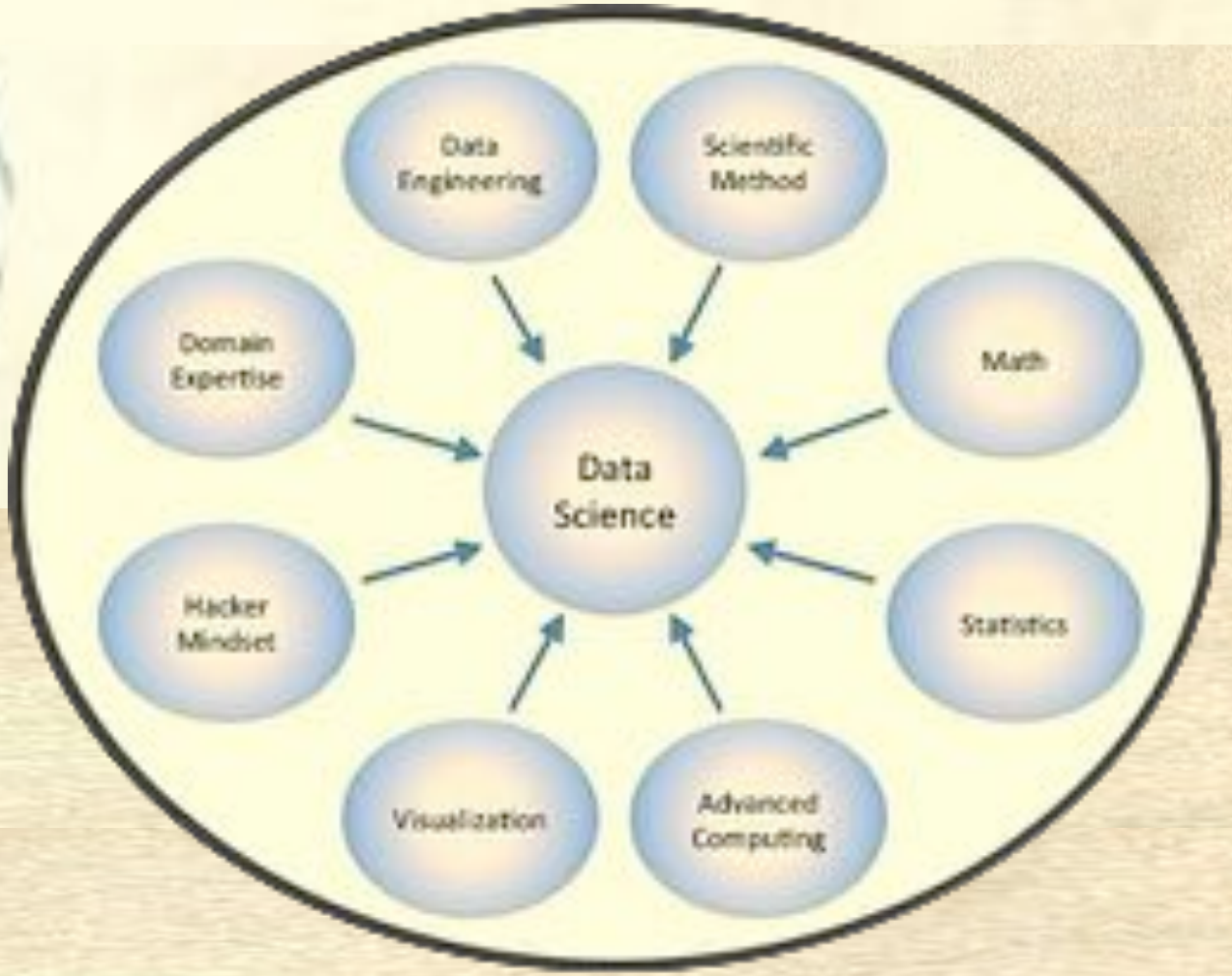

**Data are interesting, but do not answer any of the thousands of possible questions:**

- 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

# 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

=

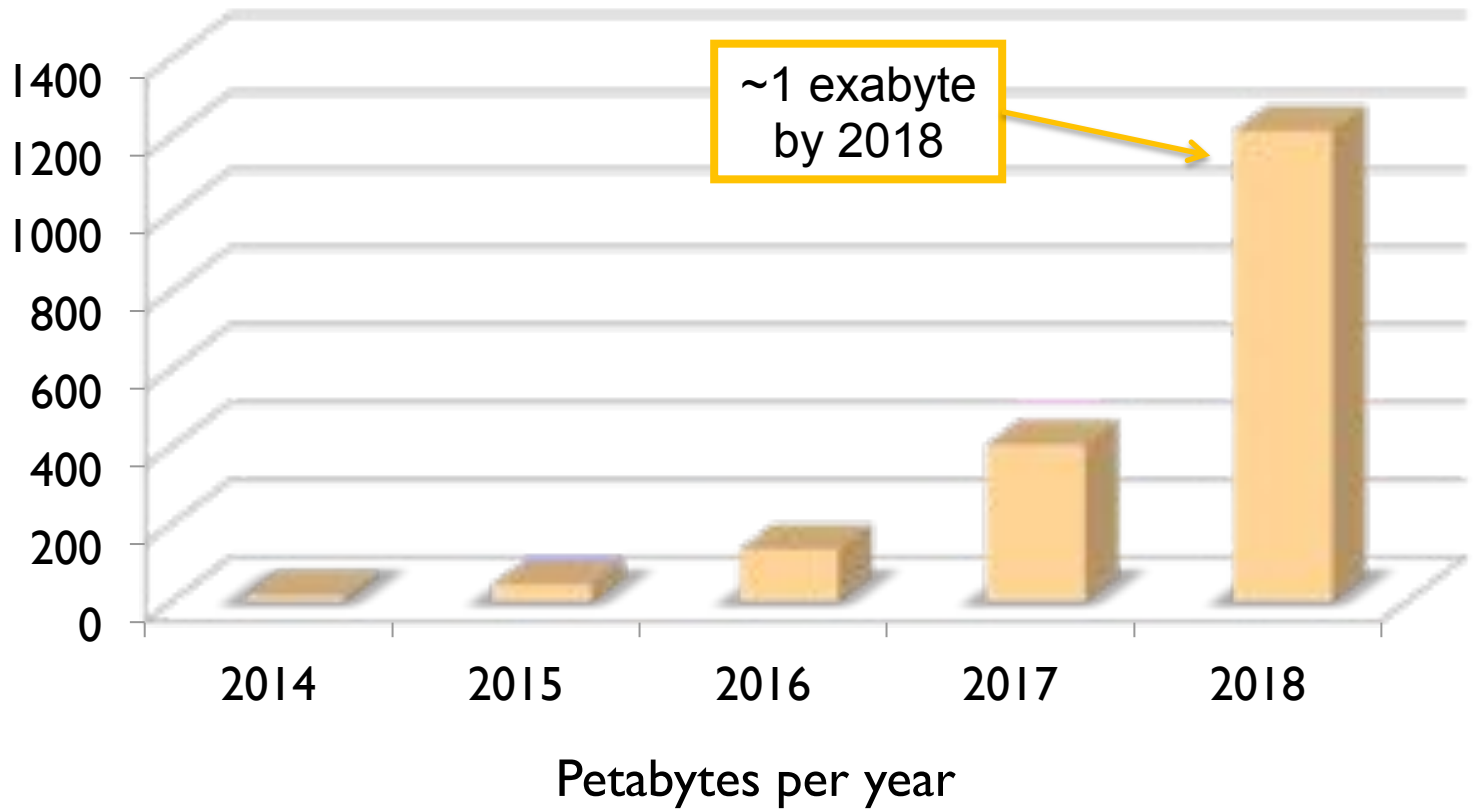
1PB Data
200,000 DVDs

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

500 2 TB drives
$500k

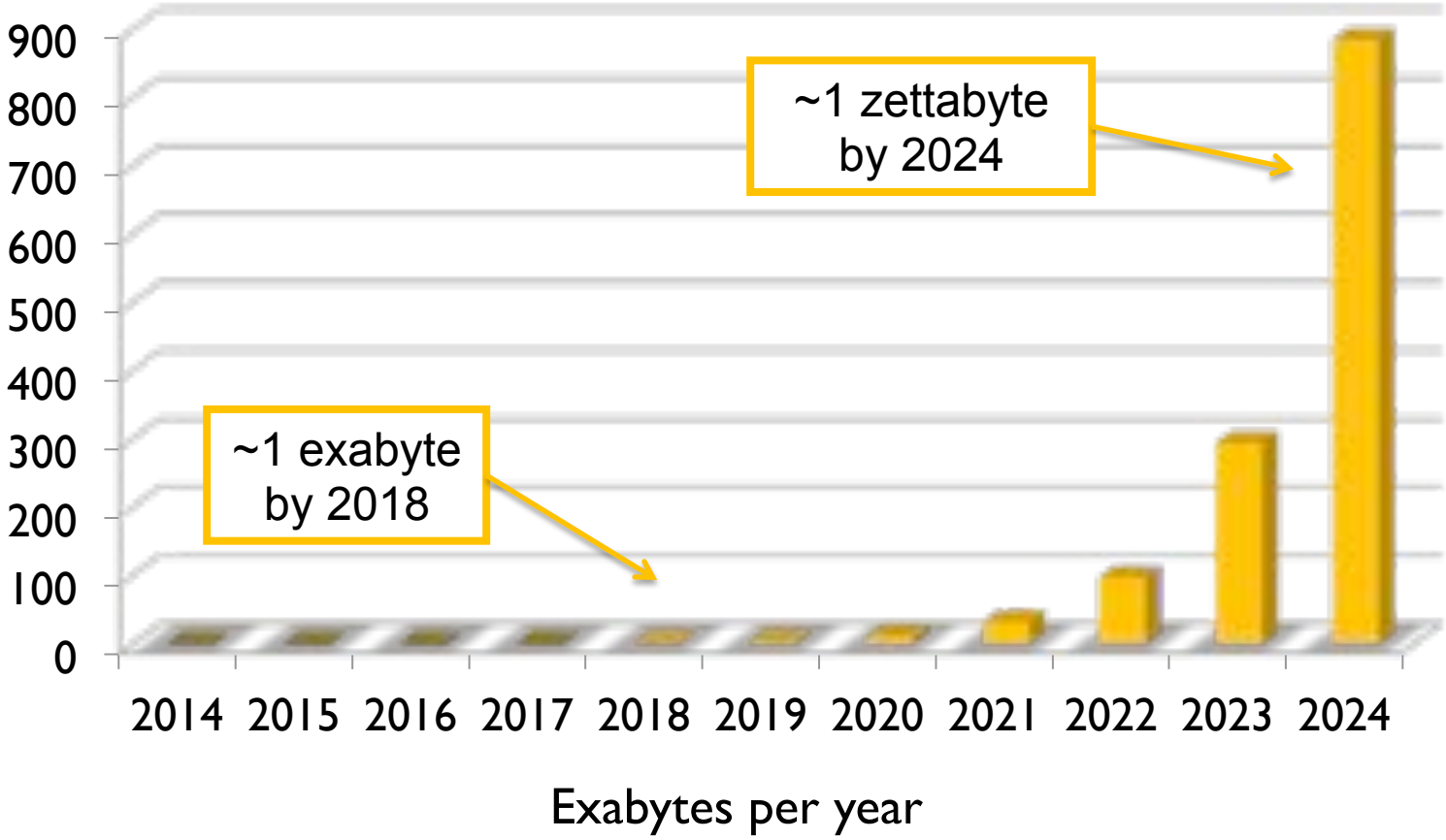
# DNA Data Tsunami

*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
~ ½ distance to moon

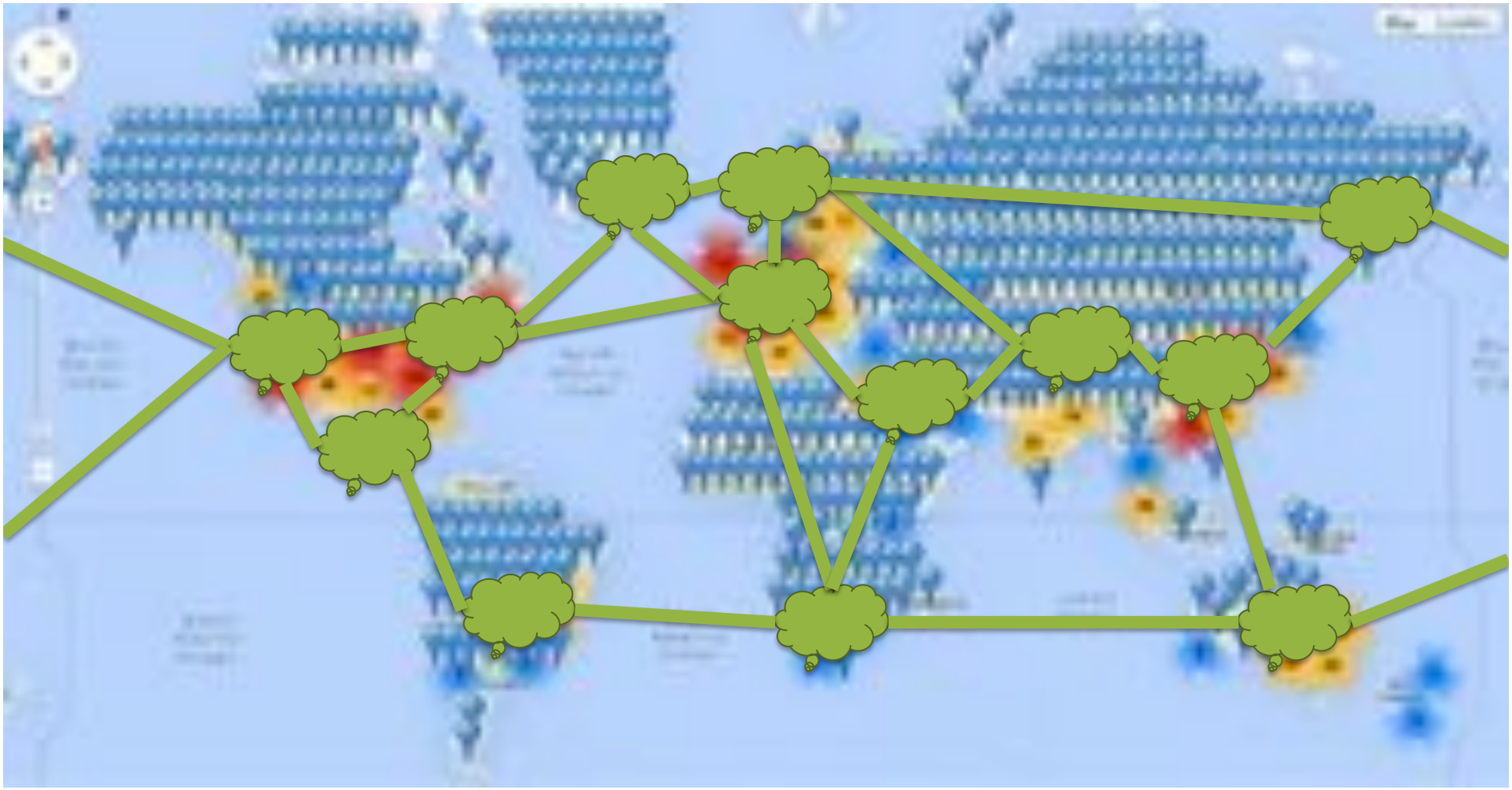
YouTube

Both currently ~100Pb
And growing exponentially

# Sequencing Centers 2014

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

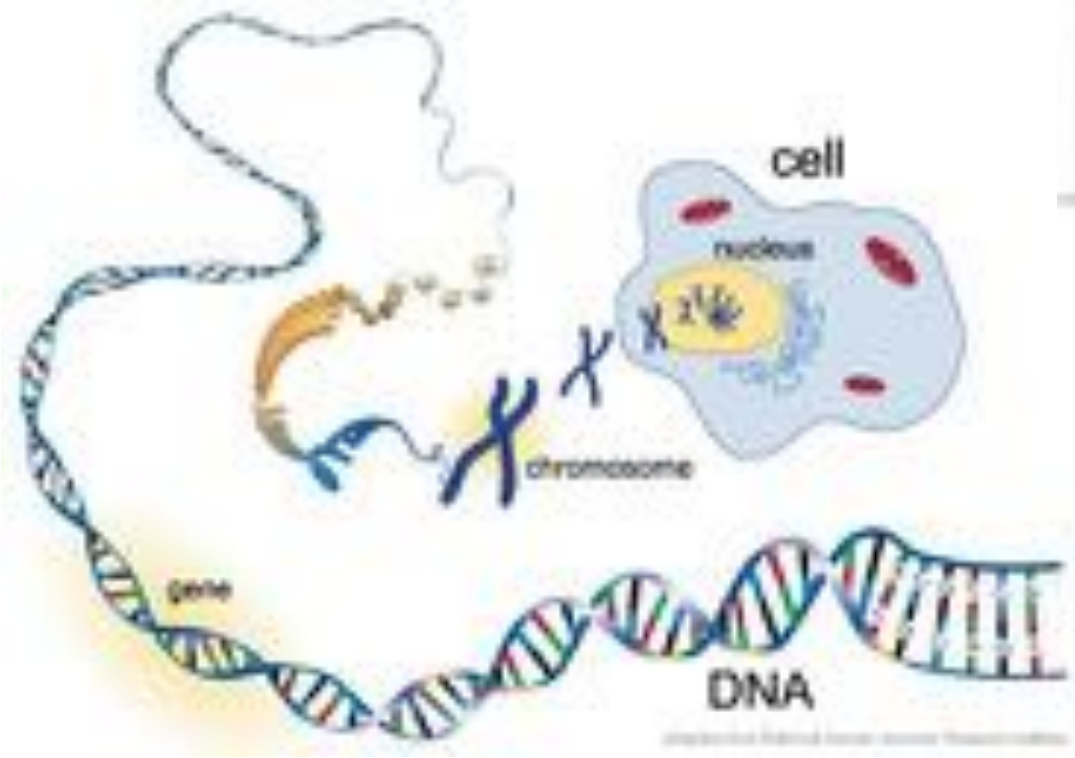
# Informatics Centers 2014



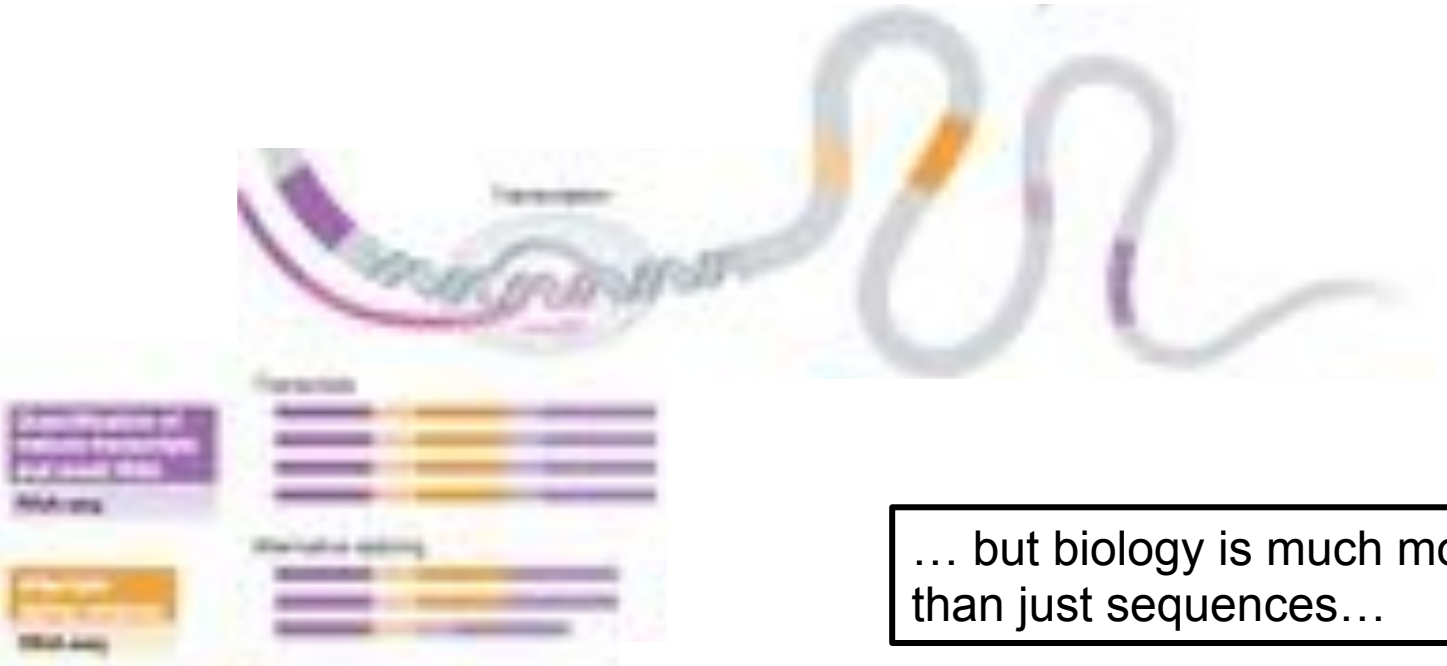
***The DNA Data Deluge***
Schatz, MC and Langmead, B (2013) *IEEE Spectrum*. July, 2013

# Biological Data

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




… but biology is much more than just genomes…



Soon et al., Molecular Systems Biology, 2013

… but biology is much more than just sequences…


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

# Biological Data Science

ARTICLE

An integrated encyclopedia of DNA elements in the human genome

ARTICLE

Mutational landscape and significance across 12 major cancer types

ARTICLE

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

LETTER

A cis-regulatory map of the *Drosophila* genome

ARTICLE

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

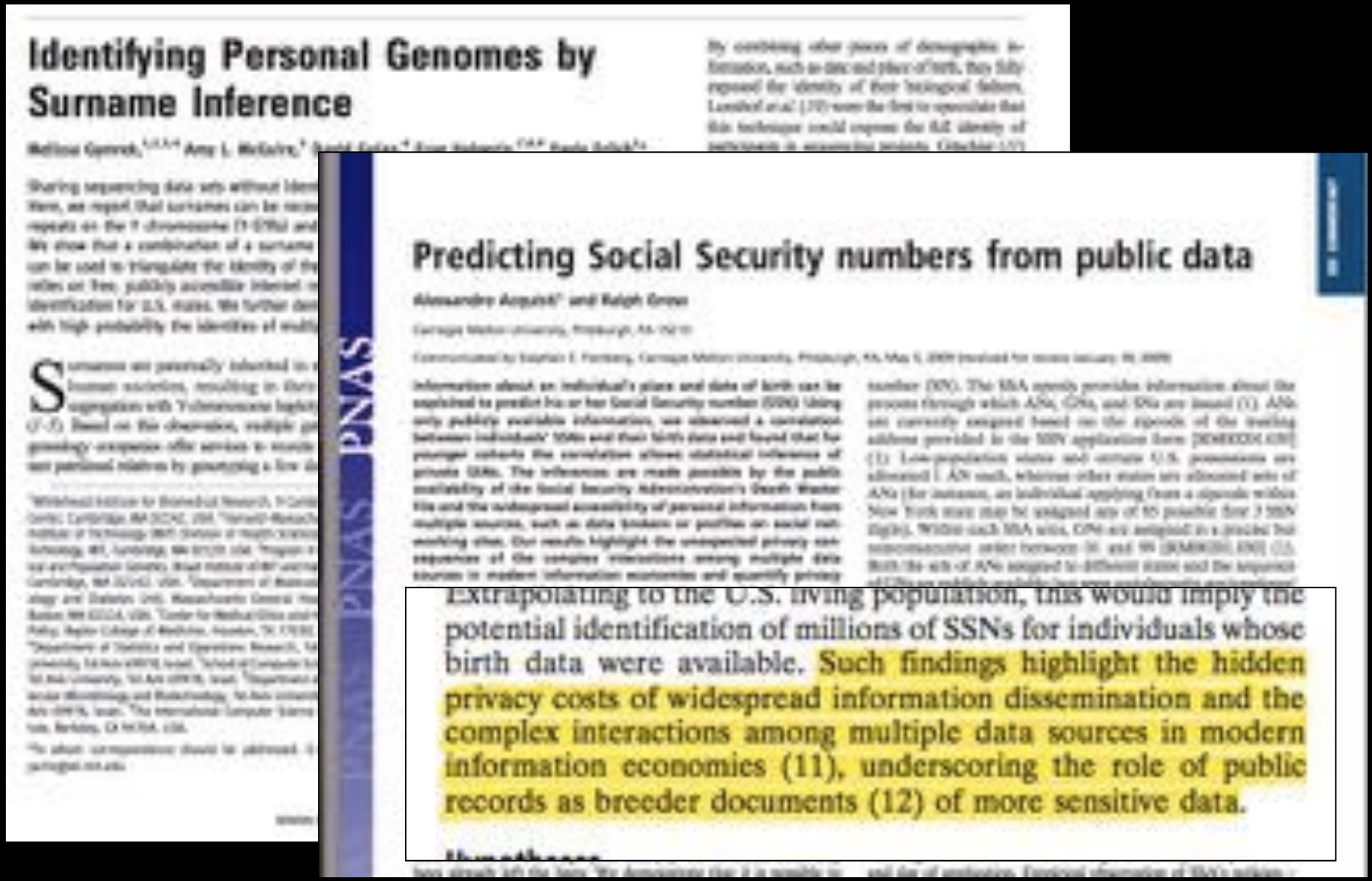
ARTICLE

A framework for human microbiome research

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

MapReduce: Simplified Data Processing on Large Clusters

# Privacy & Security



## Identifying Personal Genomes by Surname Inference

## Predicting Social Security numbers from public data

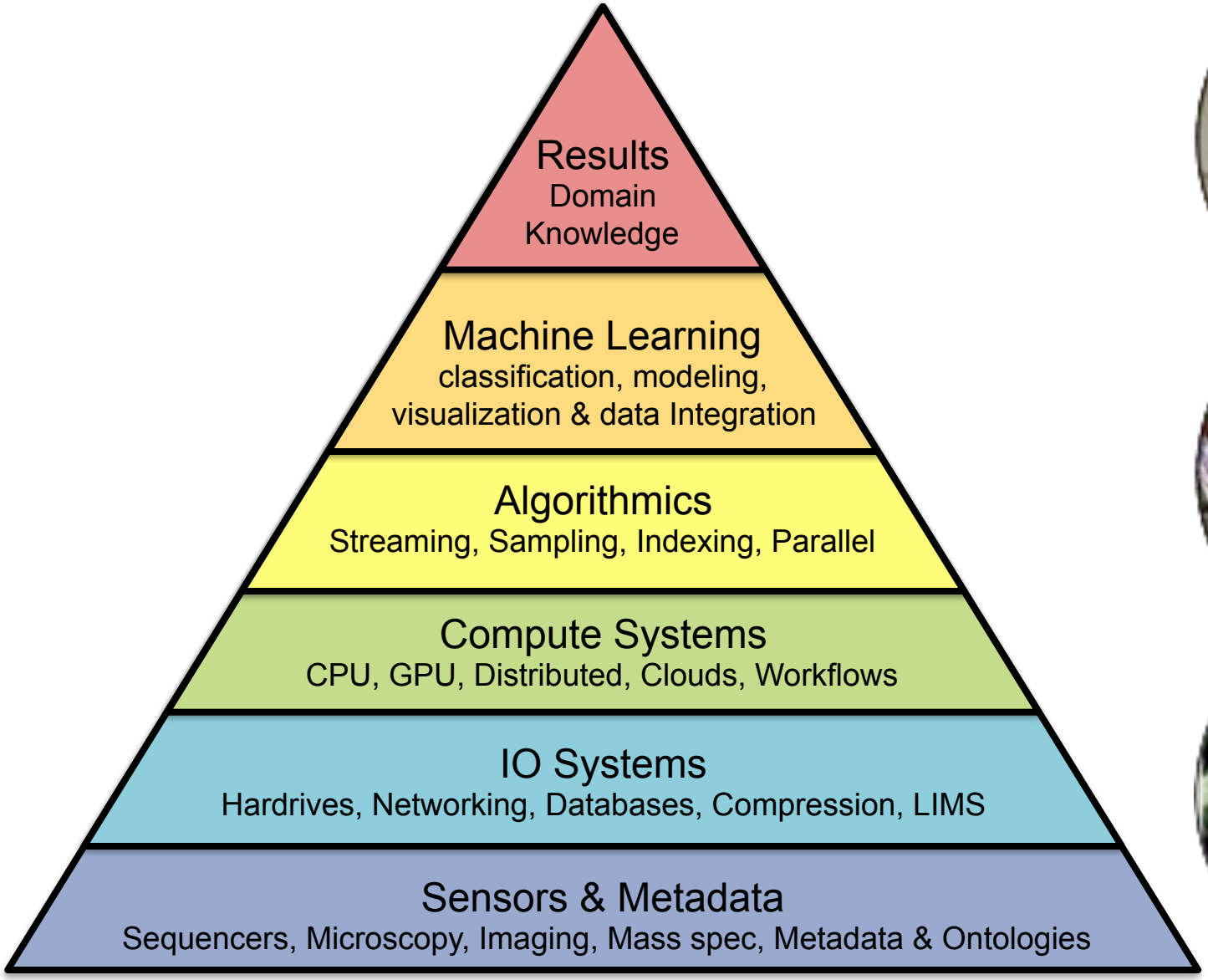
PNAS

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.

# 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 | **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