
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?

...

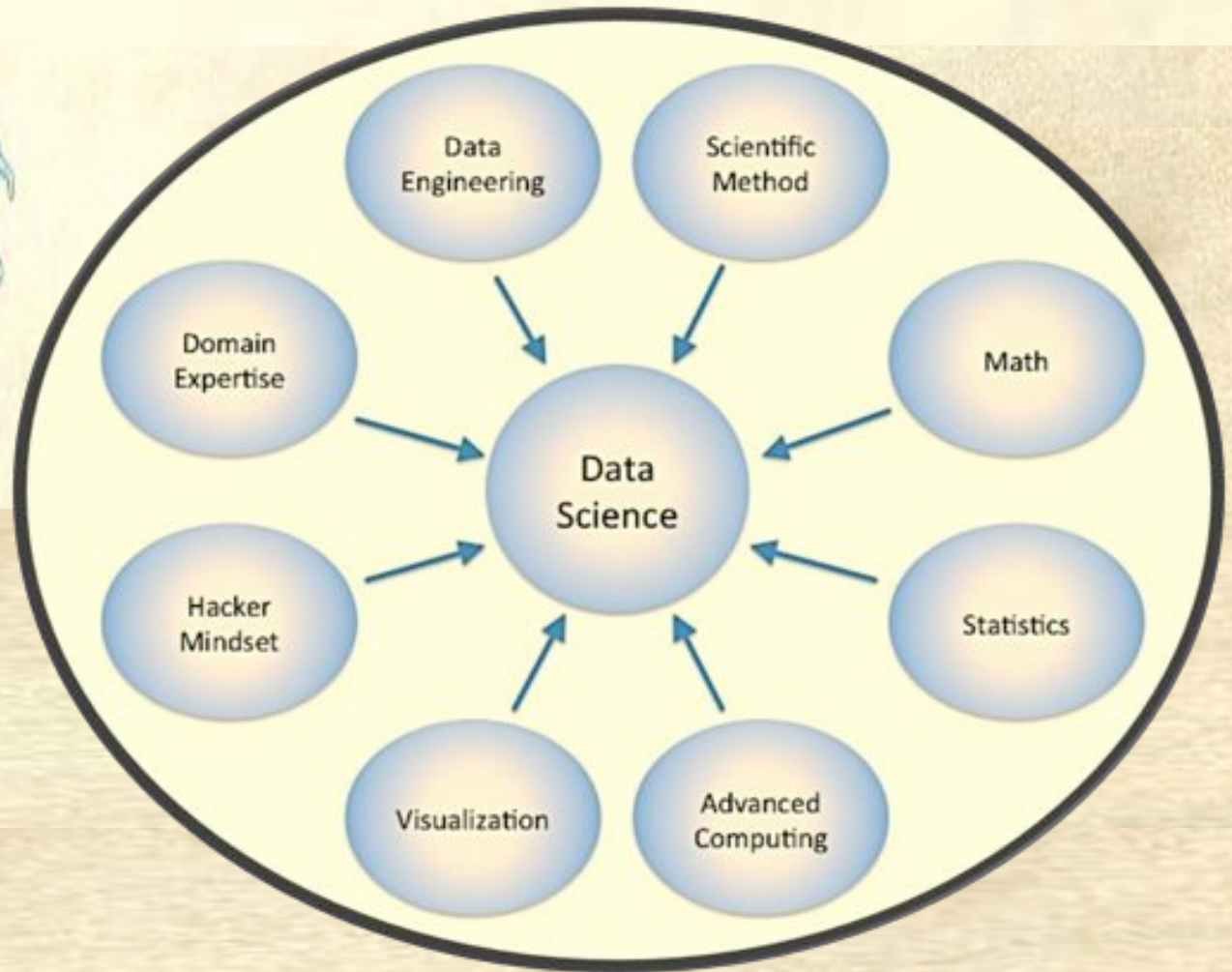
Data are interesting, but do not answer any of the thousands of possible questions:

- 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



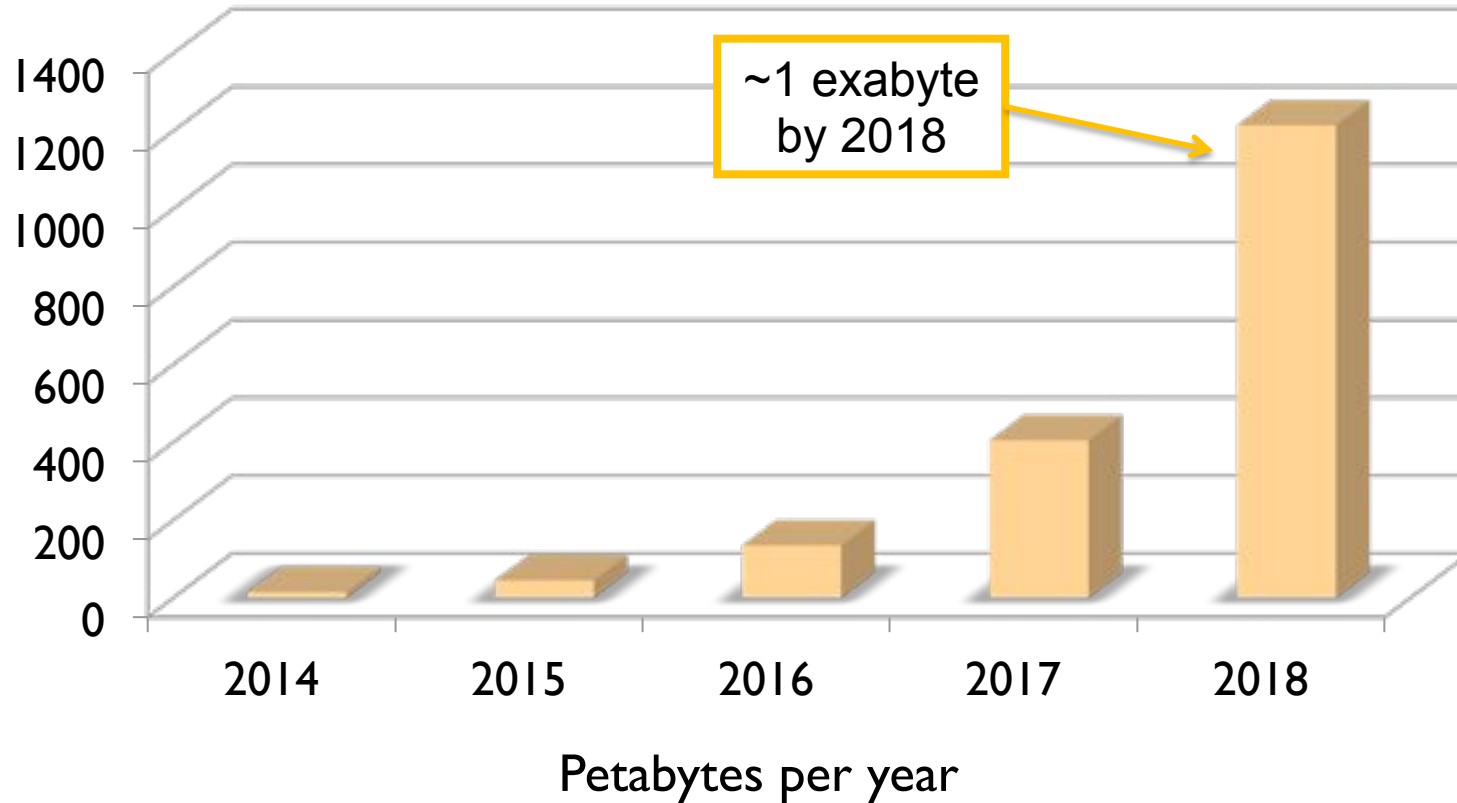
787 feet of DVDs
~1/6 of a mile tall



500 2 TB drives
\$500k

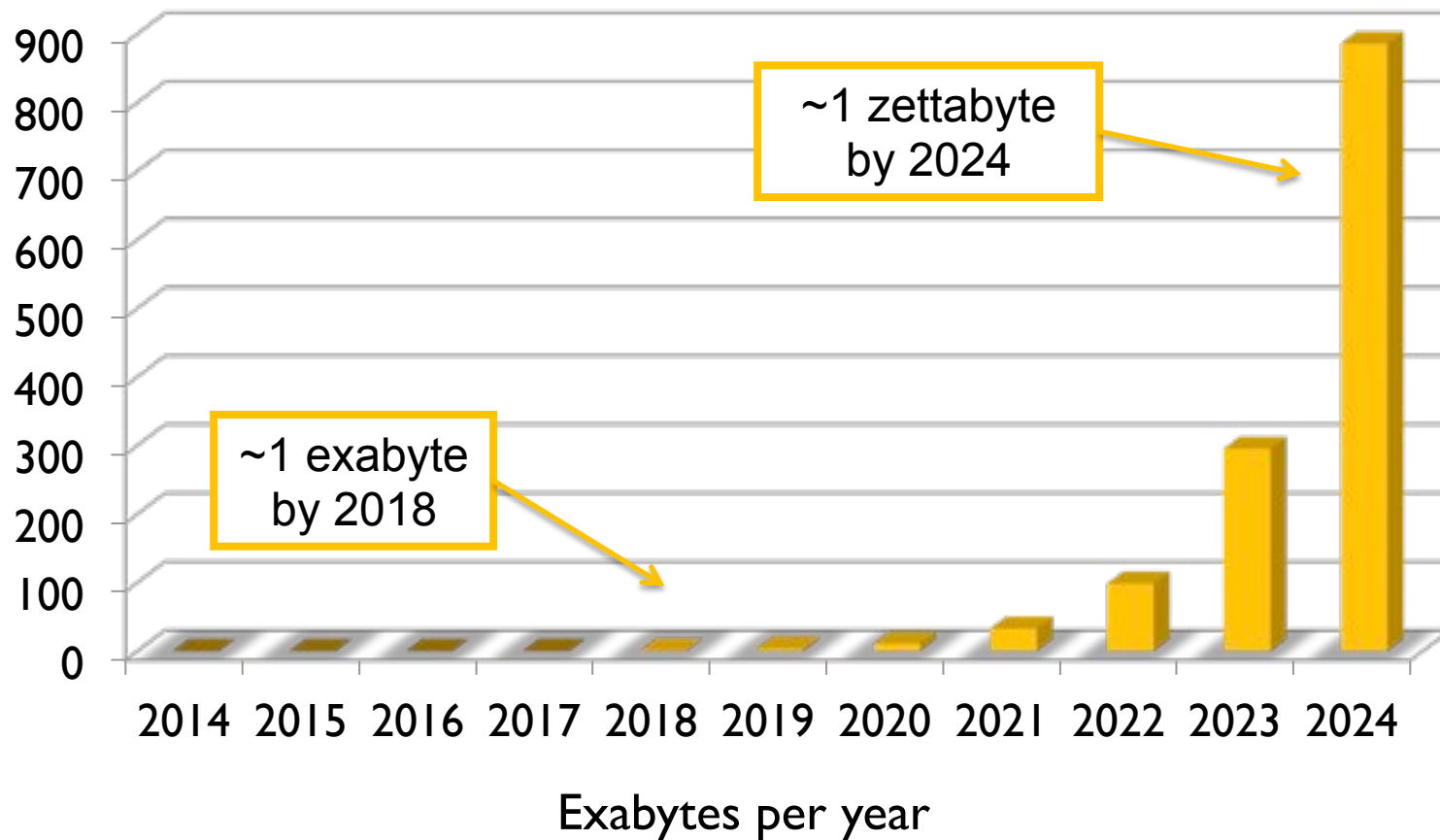
DNA Data Tsunami

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
~ 1/2 distance to moon



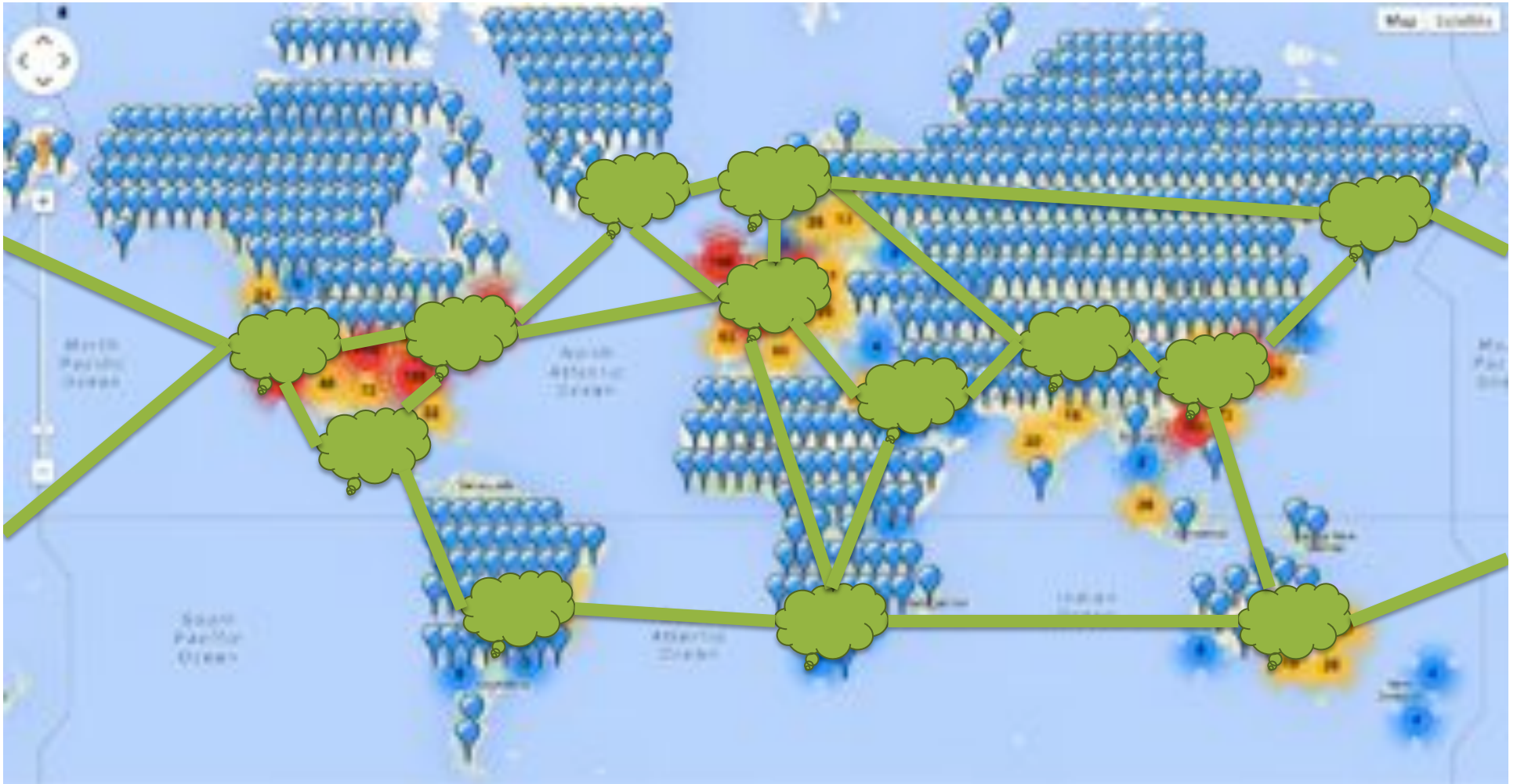
Both currently ~100Pb
And growing exponentially

Sequencing Centers 2014



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

Informatics Centers 2014

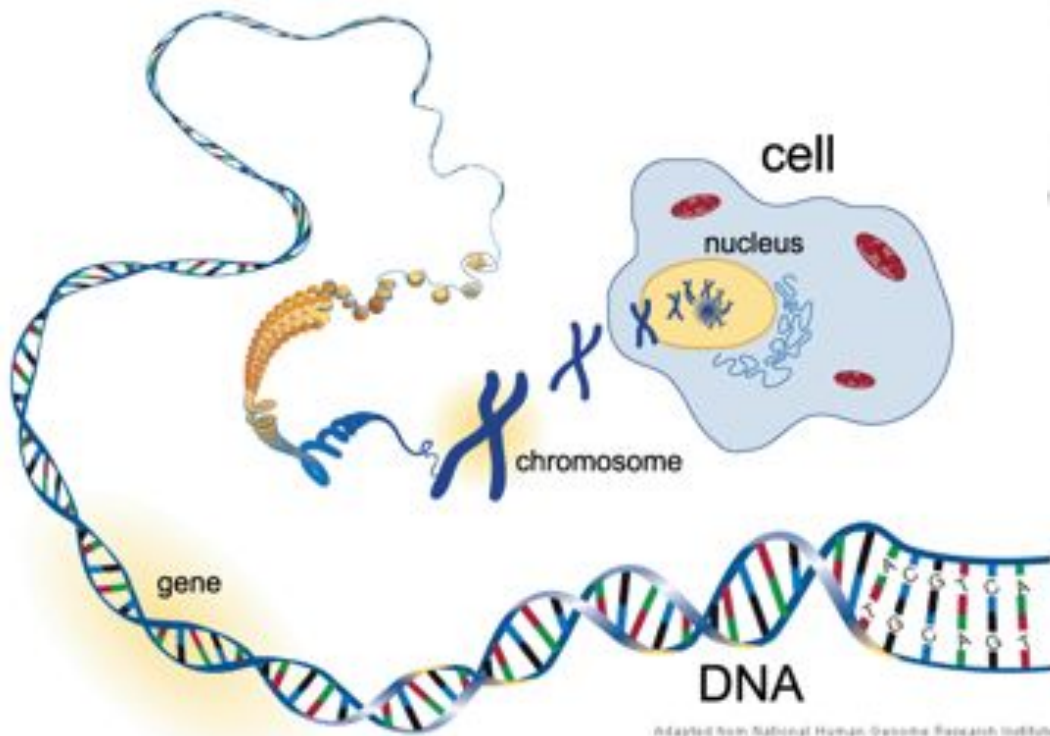


The DNA Data Deluge

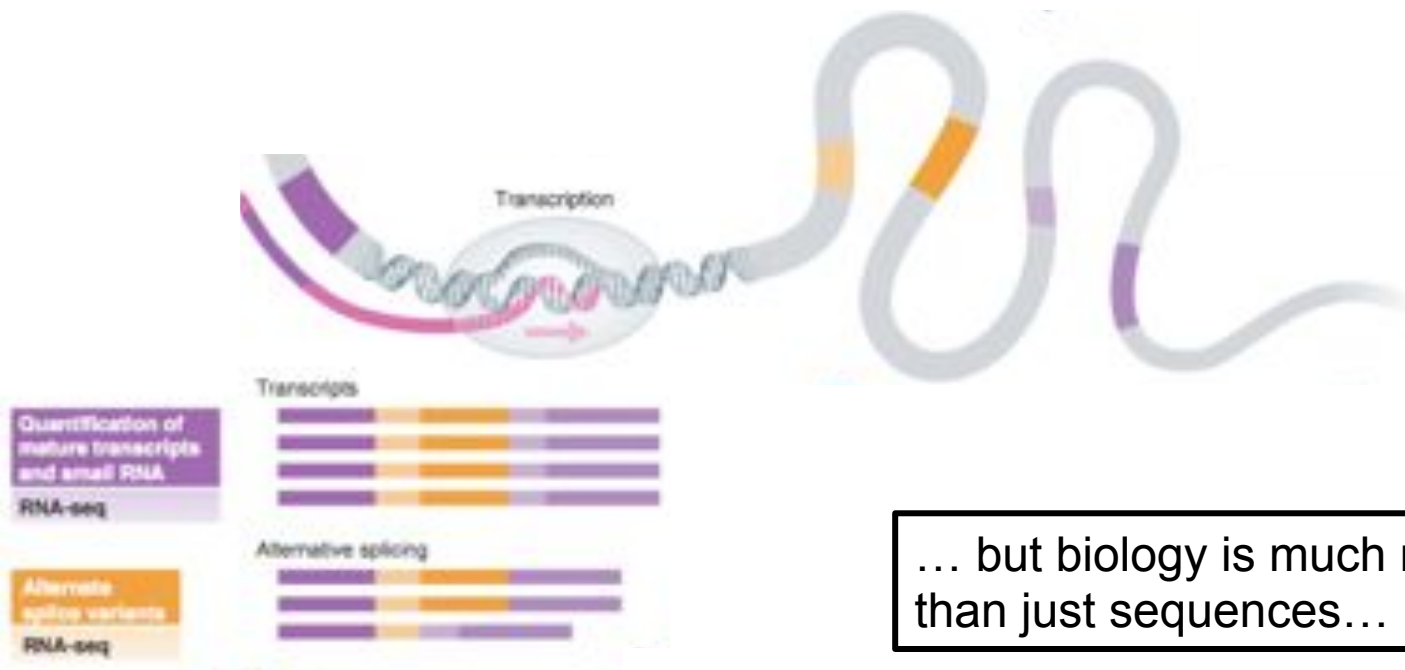
Schatz, MC and Langmead, B (2013) *IEEE Spectrum*. July, 2013

Biological Data

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



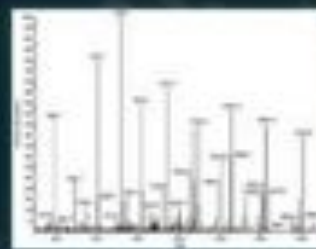
... but biology is much more than just genomes...



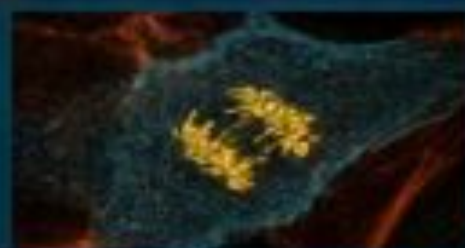
... but biology is much more than just sequences...



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>

Biological Data Science

ARTICLE

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome. In particular, outside of the well-studied protein-coding regions, we

discovered

new

transcription

start

sites

and

found

novel

regulatory

elements

that

control

gene

expression

and

found

novel

transcription

start

sites

and

found

novel

regulatory

elements

that

control

gene

expression

and

found

novel

transcription

start

sites

and

found

novel

regulatory

elements

that

control

gene

expression

and

found

novel

transcription

start

sites

and

found

novel

regulatory

elements

that

control

gene

expression

and

found

LETTER

A cis-regulatory map of the *Drosophila* genome

Nicolas Nègre¹, Christopher D. Brown¹, Li-Jia Ma¹, Christopher Aaron Britton¹, Steven W. Miller¹, Ulrich Wagner¹, Peter Kheradpour¹, Matthew L. Eaton¹, Paul Lorkovic¹, Rachel Sealfon¹, Zhongqi Li¹, Haruhiko Ishii¹, Rebecca F. Stocker¹, Jia Chen¹, Lindsay Huang¹, Chao Chang^{1,2,3,4}, Richard P. Auburn¹, Melissa B. Davis¹, Marc Dommans¹, Parantim K. Saha¹, Carolyn A. Morrison¹, Jennifer Zerbe¹, Sarah Stucky¹, Alec Vickrey¹, Nicholas A. Bell¹, A. Jason Grundler¹, David Hamby¹, David M. Mackinnon¹, Marius Mueyer¹, Koen Verhaeghe¹, Hugo Belien¹, Robert Urban¹, Mark Gerstein^{1,5}, Steven Russell¹, Robert L. Grossman^{1,6}, Bing Ren^{1,7}, James W. Posakony¹, Manolis Kellis¹ & Kevin P. White¹

ARTICLE

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

The sensitivity of the commonly used progressive multiple sequence alignment method has been greatly improved for the alignment of divergent protein sequences. Firstly, individual weights are assigned to each sequence in a partial alignment in order to down-weight near-duplicate sequences and up-weight the most divergent ones. Secondly, amino acid substitution matrices are varied at different alignment stages according to the divergence of the sequences to be aligned. Thirdly, residue-specific gap penalties and locally reduced gap penalties in hydrophilic regions encourage new gaps in potential loop regions rather than regular secondary structure. Fourthly, positions in early alignments where gaps have been opened receive locally reduced gap penalties to encourage the opening up of new gaps at these positions. These modifications are incorporated into a new program, CLUSTAL W which is freely available.

INTRODUCTION

The simultaneous alignment of many nucleotide or amino acid sequences is now an essential tool in molecular biology. Multiple alignment are used to find diagnostic patterns to characterize protein families; to detect or demonstrate homology between new proteins and existing families of proteins; to help predict the secondary and tertiary structure of new sequences; to suggest oligonucleotide primers for PCR; as an essential prelude to molecular evolutionary analysis. The rate of appearance of new sequence data is steadily increasing and the development of efficient and accurate automatic methods for multiple alignment is, therefore, of major importance. The majority of automatic multiple alignments are now carried out using the 'progressive' approach of Feng and Doolittle (1). In this paper, we describe a number of improvements to the progressive multiple alignment method which greatly improve the sensitivity without sacrificing any of the speed and efficiency which makes this approach so

ARTICLE

Mutational landscape and significance across 12 major cancer types

Cyrus Karhadlool¹, Michael D. McCallan¹, Fabio Vandin¹, Kai Ye^{1,2}, Jieping Wang¹, Mingchao Xu¹, Qiyuan Zhang¹, Joshua F. Michalek¹, Matthew A. Wyszynski¹, Mark D. Johnson¹, Christopher A. Miller¹, John S. Wodicka¹, Matthew J. Walter¹, Michael C. Wrenn^{1,3}, Timothy J. Ley^{1,4}, Richard S. Wilson^{1,5}, Benjamin J. Raphael¹ & Li Ding^{1,6,7,8,9}

The Cancer Genome Atlas (TCGA) has used the latest sequencing and analysis methods to identify somatic variants across

ARTICLE

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

Ivan Iossifov¹, Brian J. O'Rourke^{1,2}, Stephan J. Sanders^{1,3}, Michael Rothenberger¹, Niklas Krumm¹, Dan Levy¹, Holly A. Steiner¹, Jeanette Dool¹, Shan Duan^{1,4}, Lutz H. Gelernter¹, Jeffrey D. Mandel¹, Shrikant M. Mane¹, Michael J.omba¹

These exome are we apply which affective disrupting (IG) or novel mutation specific wild variant (IQ), 30 genes are number of per- gers for int ed embryos

These exome are we apply which affective disrupting (IG) or novel mutation specific wild variant (IQ), 30 genes are number of per- gers for int ed embryos

NONPARAMETRIC ESTIMATION FROM INCOMPLETE OBSERVATIONS*

E. L. KAPLAN
University of California Radiation Laboratory
AND
PAUL MEIER
University of Chicago

In life-testing, medical follow-up, and other fields the observation of the times of occurrences of the event of interest (called a death) may be prevented for some of the items of the sample by the previous occurrence of some other event (called a loss). Losses may be either accidental or controlled, the latter resulting from a decision to terminate certain observations. In either case it is usually assumed in this paper that the lifetime (age at death) is independent of the potential loss status; in practice this assumption deserves careful scrutiny. Despite the resulting incompleteness of the data, it is desired to estimate the proportion $P(t)$ of items in the population whose lifetimes would exceed t (in the absence of such losses), without making any assumption about the form of the function $P(t)$. The observation for each item of a suitable initial event, marking the beginning of its lifetime, is presupposed.

For random samples of size N the product-limit (PL) estimate can be defined as follows: List and label the N observed lifetimes (whether to death or loss) in order of increasing magnitude, so that one has $0 \leq t_1 \leq t_2 \leq \dots \leq t_N$. Then $\hat{P}(t) = \prod_{j: t_j \leq t} [N - (j-1)] / [N - (j-1)]$, where r assumes those values for which $t_r \leq t$ and for which t_r measures the time to death. This estimate is the distribution, unrestricted as to form, which maximizes the likelihood of the observations.

Other estimates that are discussed are the actuarial estimate (which are also products, but with the number of factors usually reduced by grouping); and reduced-sample (RS) estimates, which require that losses not be accidental, so that the limits of observation (potential loss times) are known even for those items whose deaths are observed. When no losses occur at ages less than t , the estimate of $P(t)$ in all cases reduces to the usual binomial estimate, namely, the observed proportion of survivors.

CONTENTS

| | |
|---|-----|
| 1. Introduction | 458 |
| 1.1 Parametric estimation | 458 |
| 1.2 Nonparametric estimation | 459 |
| 1.3 Examples of the IS and PL estimates | 459 |
| 1.4 Notation | 461 |
| 2. The Product-Limit Estimator | 462 |
| 2.1 Definition and calculation | 462 |
| 2.2 Mean and variance of $\hat{P}(t)$ | 465 |
| 2.3 Mean lifetime | 467 |
| 2.4 The Reduced-Sample Estimator | 468 |
| 3.1 Alternatives to the PL estimate | 469 |
| 3.2 Dependence of deaths and losses | 470 |

* Prepared while the authors were at Bell Telephone Laboratories and Johns Hopkins University respectively. The work was aided by a grant from the Office of Naval Research.

ARTICLE

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

Xuehui Huang¹, Nori Kurauchi¹, Xinghua Wei^{1,2}, Zi-Xuan Wang^{1,2,3}, Abong Wang¹, Qiang Zhao¹, Yan Zhao¹, Kanyun Liu¹, Hengyun Lu¹, Wenjun Li¹, Yunli Guo¹, Yiqi Lu¹, Congcong Zhang¹, Jianlin Fan¹, Qijun Wang¹, Chuanrang Zhu¹, Yao Huang¹, Lei Zhang¹, Yongchun Wang¹, Lei Feng¹, Huiyao Furuzar¹, Takahito Kubo¹, Toshio Miyabayashi¹, Xiangping Yuan¹, Qun Xu¹, Guojun Dong¹, Qilin Zhan¹, Canyang Li¹, Aiao Fujiyama¹, Atsushi Toyoda¹, Tingting Lu¹, Qi Feng¹, Qian Qian¹, Haiyang Li¹ & Bin Han¹

ARTICLE

A framework for human microbiome research

The Human Microbiome Project Consortium*

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat
jeff@google.com, sanjay@google.com
Google, Inc.

Abstract

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, as shown in this paper.

Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system.

Our implementation of MapReduce runs on a large cluster of commodity machines and is highly scalable; a typical MapReduce computation processes many terabytes of data on thousands of machines. Programmers find the system easy to use: hundreds of MapReduce programs have been implemented and executed on Google's clusters every day.

1 Introduction

Over the past five years, the authors and many others at Google have implemented hundreds of special-purpose computations that process large amounts of raw data, such as crawled documents, web request logs, etc., to compute various kinds of derived data, such as inverted indices, various representations of the graph structure of web documents, summaries of the number of pages crawled per host, the set of most frequent queries in a

given day, etc. Most such computations are conceptually straightforward. However, the input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures comprise to obscure the original simple computation with large amounts of complex code to deal with these issues.

As a reaction to this complexity, we designed a new abstraction that allows us to express the simple computations we were trying to perform but hides the messy details of parallelization, fault-tolerance, data distribution and load balancing in a library. Our abstraction is inspired by the map and reduce primitives present in Lisp and many other functional languages. We realized that most of our computations involved applying a map operation to each logical 'record' in our input in order to compute a set of intermediate key/value pairs, and then applying a reduce operation to all the values that shared the same key, in order to combine the derived data appropriately. Our use of a functional model with user-specified map and reduce operations allows us to parallelize large computations easily and to use re-execution as the primary mechanism for fault tolerance.

The major contributions of this work are a simple and powerful interface that enables automatic parallelization and distribution of large-scale computations, combined with an implementation of this interface that achieves high performance on large clusters of commodity PCs. Section 2 describes the basic programming model and gives several examples. Section 3 describes an implementation of the MapReduce interface tailored towards our cluster-based computing environment. Section 4 describes several refinements of the programming model that we have found useful. Section 5 has performance measurements of our implementations for a variety of tasks. Section 6 explores the use of MapReduce within Google including our experiences in using it as the basis

*To whom correspondence should be addressed

*Present address: European Bioinformatics Institute, Hinxton Hall, Hinxton, Cambridge CB10 1RT, UK

Privacy & Security

Identifying Personal Genomes by Surname Inference

Melissa Gymrek,^{1,2,3,4} Amy L. McGuire,⁵ David Golan,⁶ Fran Halperin,^{7,8,9} Yaniv Erlich^{1*}

Sharing sequencing data sets without identifying individuals is a challenge. Here, we report that surnames can be recovered from Y-chromosome repeats on the Y chromosome (Y-STRs) and used to triangulate the identity of the father. We show that a combination of a surname and a Y-STR can be used to identify the father of a child. This method relies on free, publicly accessible Internet resources and does not require re-identification for U.S. males. We further demonstrate that with high probability the identities of multiple

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

¹Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA. ²Harvard–Massachusetts Institute of Technology (MIT) Division of Health Sciences and Technology, MIT, Cambridge, MA 02139, USA. ³Program in Population and Quantitative Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. ⁴Department of Molecular Biology and Diabetes Unit, Massachusetts General Hospital, Boston, MA 02114, USA. ⁵Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, TX 77030, USA. ⁶Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel. ⁷School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel. ⁸Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Tel Aviv 69978, Israel. ⁹The International Computer Science Institute, Berkeley, CA 94704, USA.

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

www.

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)

Predicting Social Security numbers from public data

Alessandro Acquisti¹ 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

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/).

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.

have already left the barn: We demonstrate that it is possible to and day of application. Empirical observation of SSA's policies—

SEE COMMENTARY

PNAS

PNAS

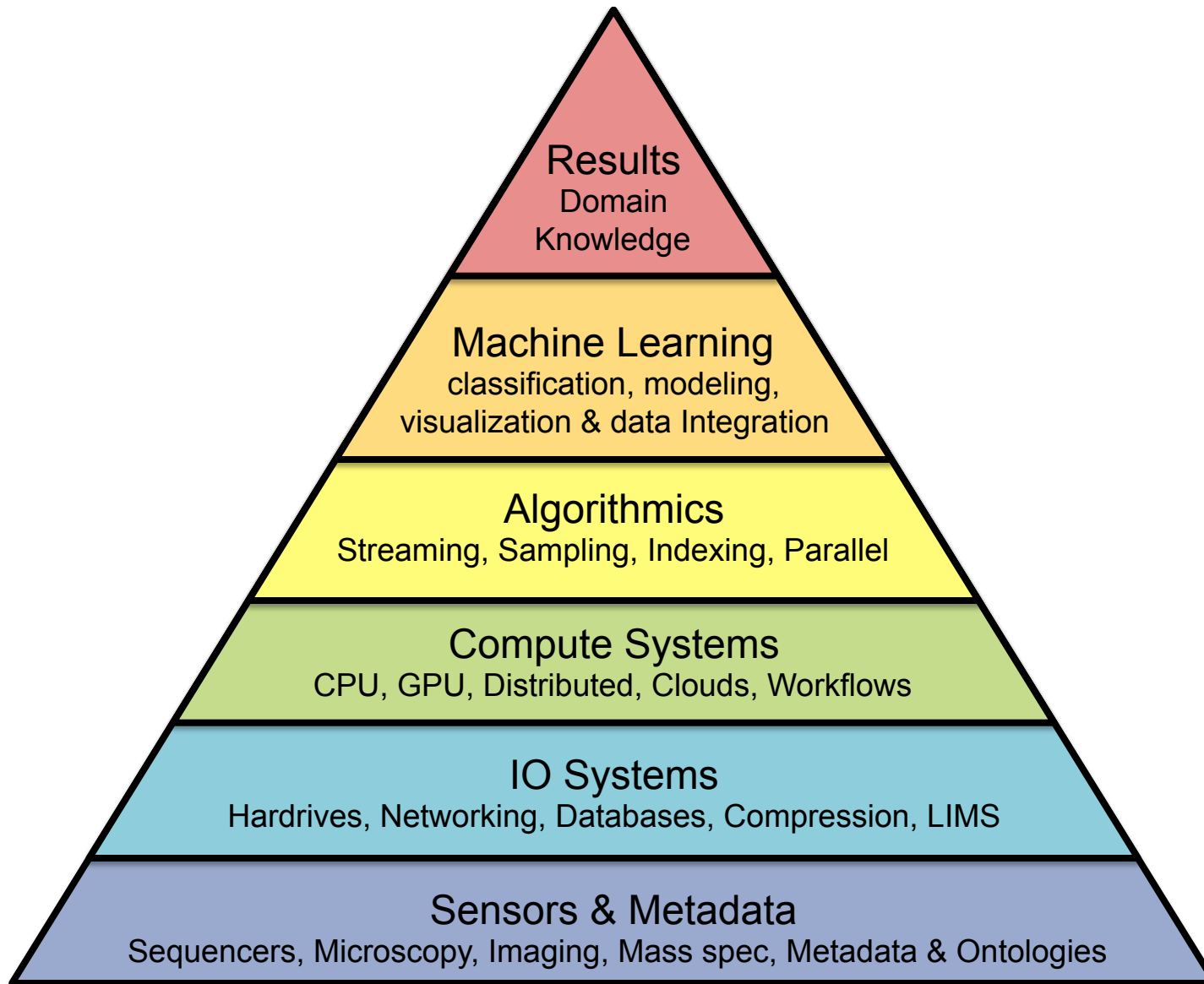
PNAS

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



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