# **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





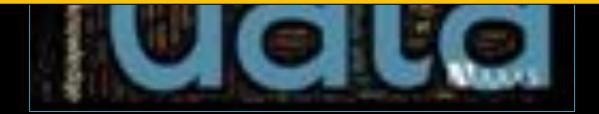


### @JasonWilliamsNY

### Charla Lambert

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

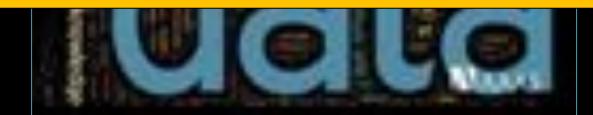
- How does my genome compare to yours?
- How does expression or methylation or chromatin change?
- What diseases are you at risk for, what pathogens have you been exposed to, and what medicines should we give you?

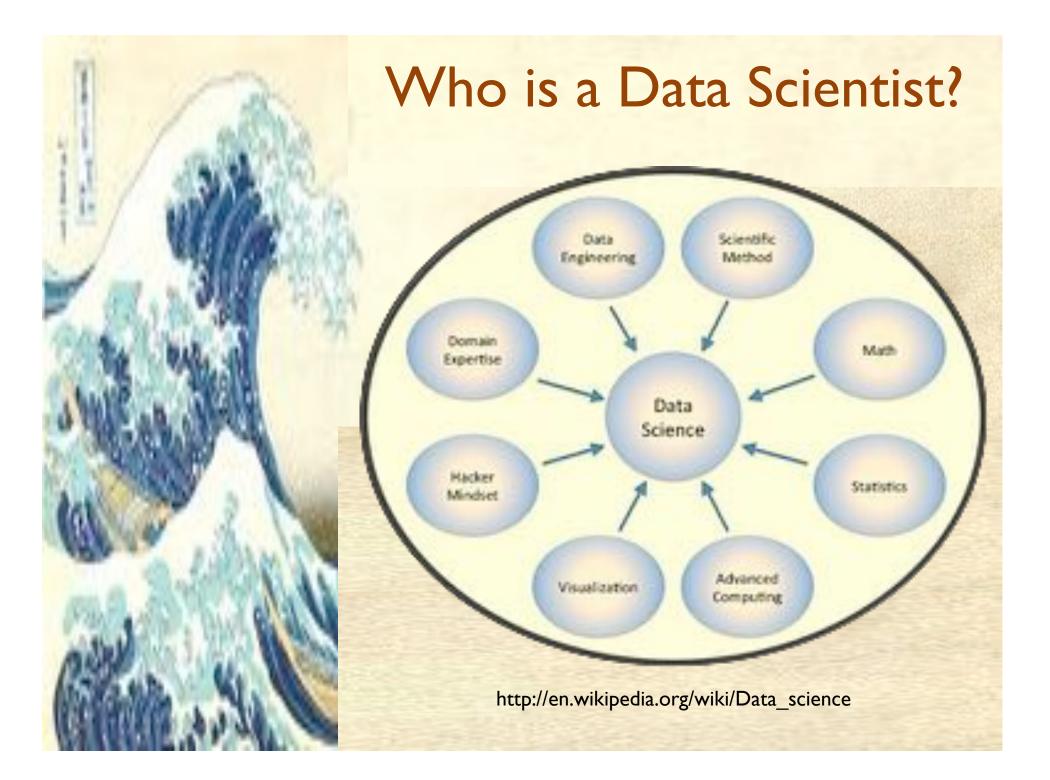


### Data are interesting, but do not answer any of the thousands of possible questions:

- 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?





# **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	
Kilobyte	1,000
Megabyte	1,000,000
Gigabyte	1,000,000,000
Terabyte	I,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

Х

10,000 Genomes

=

1PB Data 200,000 DVDs



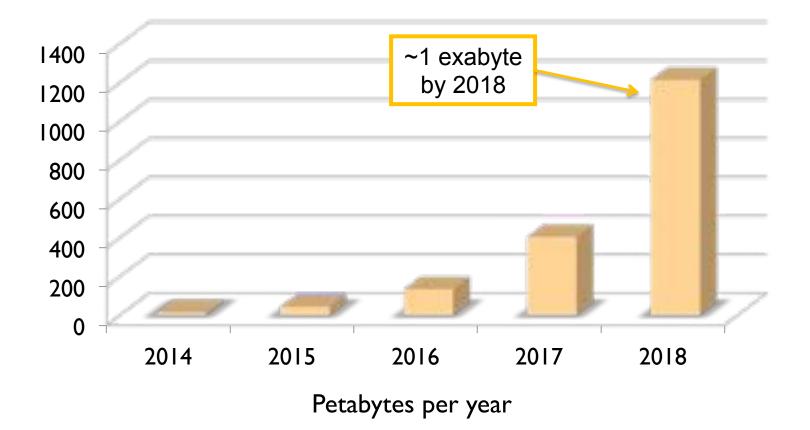
787 feet of DVDs ~1/6 of a mile tall



500 2 TB drives \$500k

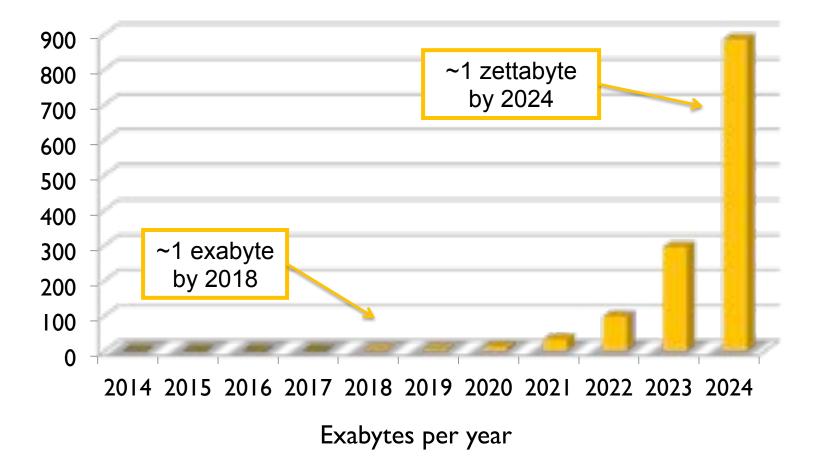
## DNA Data Tsunami

Current world-wide sequencing capacity is growing at  $\sim 3x$  per year!



## DNA Data Tsunami

Current world-wide sequencing capacity is growing at  $\sim 3x$  per year!



# How much is a zettabyte?

Unit	Size
Byte	
Kilobyte	I,000
Megabyte	1,000,000
Gigabyte	I,000,000,000
Terabyte	I,000,000,000,000
Petabyte	1,000,000,000,000,000
Exabyte	I,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

Х

10,000,000,000 Genomes



1ZB Data 200,000,000,000 DVDs







150,000 miles of DVDs  $\sim \frac{1}{2}$  distance to moon

Both currently ~100Pb And growing exponentially

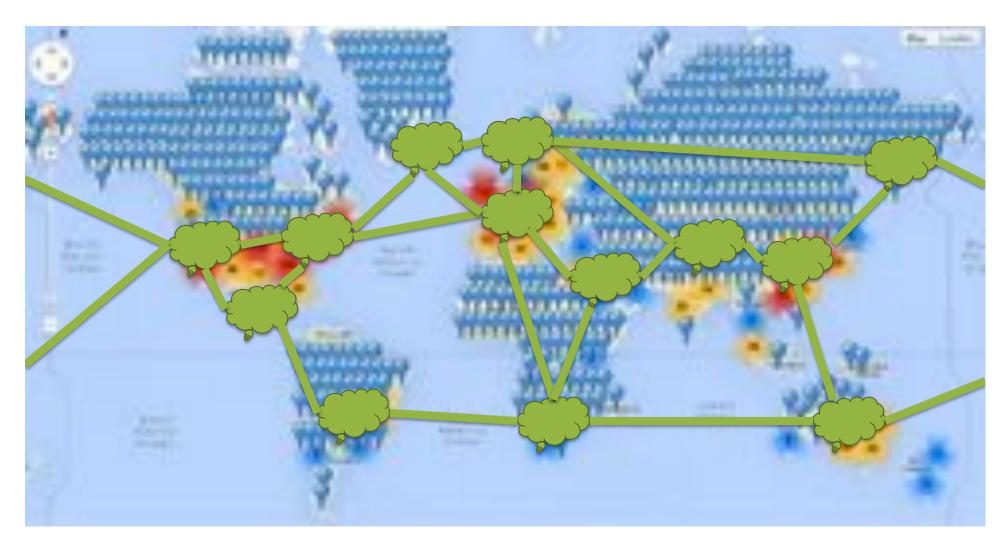
# Sequencing Centers 2014



### Next Generation Genomics: World Map of High-throughput Sequencers

http://omicsmaps.com

# Informatics Centers 2014

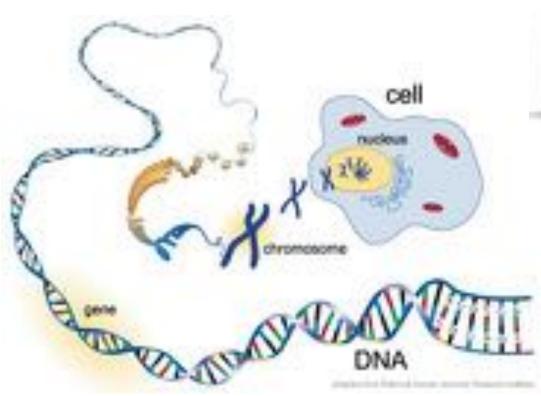


### The DNA Data Deluge

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

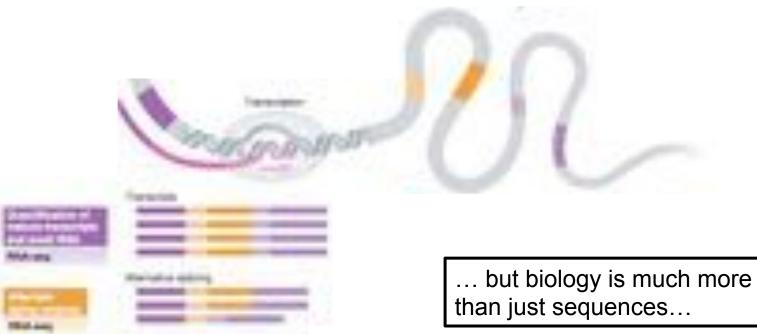
# **Biological Data**

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





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



Soon et al., Molecular Systems Biology, 2013



http://www.slideshare.net/pebourne/wiki-mania080914

### **Biological Data Science**

### ARTICLE

#### An integrated encyclopedia of DNA elements in the human genome

The DALLINE Associations?

tong its generative at the set of the set of

Annual States of the local states of the

All and a second second

these suggests

for pression rule reserve 1,000 general

Annual and a second state

the second second

perma as parter

and the second s

- Spinkenger

and it. See

instantights &

1110

ATTENNER.

### LETTER

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

Sandar G, Bourt Y, Jak Ku, Chennaker James Kinne R. Killer Y. Mu Backer J, Spern Angel Status , Nethol Applied Streng V, Statuber MJ, Schen Marker MM, Sandar S, Sandar M, Sandar M, Sandar M, Sandar MM, Sandar MM, Sandar S, Sandar S, Sandar M, Sandar M, Nathan J, Hu Shellar MM, Sandar S, Sandar S, Sandar S, Sandar M, Nathan J, Hu Shellar MM, Sandar S, Sandar S, Sandar S, Sandar M, Sandar J, Malan S, Sandar MM, Sandar S, Janna & Manlands, "Anima Manageth," Anne "Animati," Singi Ashari, Sahari Kimi," Anja Sano, Manua, Commun." Sing Int.", Society & Designer," Strength Mills, & Long, 7, Wiley.

E 104 Oakel Develop Pres

feature and Asserts, APA, No. 12, No. 12, and 1 and

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

ALC: NOT THE OWNER.

Jule D Thompson, Desmand G/Hggire\* and Tuby J.Otteon\* Surgroup Molecular Bulkgy Lateratory, Postach 10209, Meyertotatees 1, 0-60112 resolutions. Germany

Penalmel 20/y 12, 1984; Revised and Asserted Department 28, 1994

#### ABSTRACT

<text><text><section-header><text><text><text><text><text>

#### The relative processing watching for addressed

inter Same Int. Name Costing (Sci. 14), 18

ARTICLE

#### Mutational landscape and significance across 12 major cancer types

trans a second to be a second to the second se

or finite bits 7000 to an house againing of avoid active bits to react or landow

### ARTICLE

respectively 1.1

#170.

11111111111111111

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

#### INTERACTORY APPRACED FROM ITCOMPLETE OBSLETATIONP

R.L. Bates Terenty of Collecte Automa Solution -Are time

Talains of Thomas

In Statement, we first fallen og, och stelle Bette He offentilteter of De State of concentration of the second of Astronom Conference on Astrowhere any approximate of the second of the second scalars of the s Writes upp of the U is independent of the polarised from time, by previous these sequences measure over of or writes. The proper test measures, in the transport of the sequences of the transport of the polarization, if of the second sequences is the transport of the transport of the origin of the transport of the transport of the transport of the origin of the transport to the transport of the

for the second s

Febra antiquing that are devoted are the oriental attinuing tofact. are also produces, has able for results of factors usingly include by proving 1 and subsend weight (2.6) withinking which weights that there are to architectual to the the limits of observation interesting factors are in a subsend, to they the individual structures interesting factors are associated for share the individual structures. The in both sould all age has then (, the editor's of F(r) is all near seture to fix and fractional estimate, cannot, its observat projection

#### ----

1.1	and the	
~ 0	() Reparatalises	
	LF PLANNARY COMMENT	
	14 Theorem of the 18 and 10 colonest	
	Batalan Colorador Colorador	
	Par Puntur Linge Entropy	
	1.1 TMINER and refr-Mark	
- 1	<ol> <li>How and withow of \$55.</li> </ol>	
	18 THEOR BRIDE CONTRACTOR OF THE OWNER	
	Pair Stativest Notable Sections in the PS-	
	<ol> <li>Allertative is the FI, antitudo</li> </ol>	-
	<ol> <li>Dependence of deploy and terms</li> </ol>	

### ARTICLE

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

Barden Tanayi, Yan Daniya, Hanghao Wali, A. Xuan Mund, Y., Aman Mund, Yuan Ziao, Yuan Xiao, Yuan Xiao, Jiao Yuan, J. Shanayi, Y. Yuan Xiao, Yuan Yuan, Y. Yuan, Yuan, Y. Shanayi, Yanayi, Y. Shanayi, Yanayi, Yuan Yuan, Yu

### ARTICLE

101111111111111111

23

2012105025\*5222122552255255

THURSES:

-----

#### A framework for human microbiome research

#### MapRobuce: Simplified Data Processing on Large Clusters

Artisty Dean and Bargay Chamanas

pillingergin over anger fragmatic over

#### Create Inc.

#### Abstract

Highster is a preparate mole and as your and implementation for processing and generating large data and. Easts specify a map benchm that processes a implement pair or processes a set of intermediate large value pairs and a maker function that surgets all intermediant values associated with the same intermediant key. Many real world tasks are supersafile in this model, as itsner in the paints

Programs termined in this historical style are continued ofly pendidumbana encount in a large chain. If you readily machines. The tax rises comes takin uses of the deals of participants up a new scheduling the pro-gram's members arrows and insufficies, landing the this falses, and samping in regard new matter commutation. The alters preparation without any reporters with parallel and charitened system is on-To calling the component of a large photological scenary. Our implementation of Manholatic rank on a larger ment of community mathins and a legity sin

a traugi Malifeliter uniquiative process same terdepend of data or demonster of starburs. Programmary field the symmetry to new hereined of Highbolics are prior here been implemented and spreach of our deve and Mightainet printing canonid or Graph's states manip also.

#### 1 Detroited into

Over he put his part, he addres not many plan at Groupit have implemented hatelyink of special partyres compilation that protect large measure of new dest, such as control decompose, with separat large, est, in compute contrast balak of derived date, such as invested taliani, revinas tapmeniation of the people enteriner of well-devaluants, extended of the tasking of pages concluding to first, for us of next-frequent parties in a

#### To oppose in \$1000 (see

grow-day pit. Must such compositions are one ely singleficientel Borecost, de Tope des 5 consig-lege and des competences have a fix detributed action hardholt at presents of machines it with a bold in a resentativ assess of loss. The sense of loss to per-elector for companyion, clienthers the loss, and taulity federate onlights to observe the original length origin. taken with here advention of younghin ranks in deal with

As a maximum to this association and inequality are designing the gloss on to option for might compute these one want thing in performing that the output inand of parabelliaries, fast originant, dos the Delibertes and load behaving in a West's Oar distantion is to spinel to the weat with index principles present is 1-pp and many other limitized languages. We realized the term of our comparation territoric applying a language station to such highest "harms?" in our spherics could be compare a set of companies largencies pairs, and then applying a tablest spirature on all the values that charact He was het, in other is combine the detend day of proprietly. Our use of a functional count with our specified and and subset operations affinis as is parel. later large representations leady and to us to constitut to be phine; technism to bell him

The second committeeness of this work are a complex and prevailed minihes the station memory people and intribution of large scale compositions, exciting with an applementation of the specificse that achieves high parts interest inter channels of unstanding (N).

Bucility 2 describes the back programming social and providenced countries. Notice 2 descripts as imple-mentation of the Wagheshad interface latitud control. or closer basel comprise entropeent. Reside 4 de-techna anteral telescone of the proposition would that its have thered used at . Rushing 7 that performance tenergenerists of our implainabilities his a variate of spile. Sectors 6 coglisers due are of MagRenhar within Enopy methoding one experiments in using it as the factor

### Privacy & Security

### Identifying Personal Genomes by Surname Inference

Belliou Genreb, 173,14 Ares L. Bella/re," Annie Enter," Store Holmonie 74.4 Runte Bellehie.

Buring seguencing data sets attituti ident tions, we regard that survames can be receiv repeats on the 1 divancements (2) Othil and 84 shee that a contribution of a surtaine can be used to Trianguide the Identity of the reflex or Teo, publicly accessible internet re identification for 0.5, mass, the further data with high probability the identities of walts.

Structure and paratrally inherital in a former materies, resulting in their copyrights with Y-theoreman topics (7-5) Based on His descention, realized pr paratrally exerption offer services to marite any paratral relative by generating a live in

"Withdraud Autom for Branchul Henrych, Frühle Genter, Dariteritase Ref. 2014, 2014, "Introd-Henrych Roburg, BH, Genteritas, BH 2014, 2014, "Introd-Henrych Roburg, BH, Genterita, BH 2014, 2014, "Integer in Genteritage, BH 2014, 2014, "Organization Genteral Henlands, Hei Elitato, Unit, "Separation Genteral Hen-Roburg, BH 2014, 2014, "Genterfor Bethard Dire and H Roburg, BH 2014, 2014, "Londo- for Bethard Dire and H Roburg, BH 2014, 2014, "Londo- for Bethard Dire and H Roburg, BH 2014, 2014, "Londo- for Bethard Dire and H Roburg, BH 2014, 2014, "Londo- for Bethard Dire and H Roburg, BH 2014, 2014, "Londo- for Bethard Dire and H Roburg, BH 2014, 2014, "Londo- for Bethard Dire and H Roburg, BH 2014, 2014, "Londo- for Bethard Dire and H Roburg, BH 2014, 2014, "Londo- for Bethard Dire and H Roburg, BH 2014, 2014, Scill, Scill, "State of Tarrest Tab. Bethard, Change, The International Langutor Tairrest Tab. Bethard, Chang, Chan, 2014,

"Is abort schementerer should be pathward. It participal mit alls

tainet and

### Predicting Social Security numbers from public data

By condening other pieces of denoiphilits information, such as date and place of http:, they fully reposed the identity of their basing/oil failters.

London and (10) non-the first to speculate that this technique smill mersus the 8.2 utents of

participants in accounting projects. Cittachine 527

Alexandro Acquist? and Reigh Grows

Earnige Meter University, Product p. AA 15210

Extension/Land Rev Extension E. Parmany, Carmage Matters University, Physics 97, 454, 454 5, 2009 (resolved for review lasticary, H), 2009

information aloud an individual's place and date of lartit can be exploited to preside his or her Contal Genurity number (50%) Using only publicly available information, we assumed a contactuation between individually 50% and their birth data and found that he promps colorests the secondation alread states and found that he promps (50%). The information alread states and found that he provide (50%). The information are reade presiden by the public mailability of the field becarity indeconstruction's fourth Waster tits and the nodespressi accessibility of personal information from multiple secures, such as data broken or profiles or social ratentities of the contactor resolutions are profiles of social ratentities of the contactor resolutions around allows and sequences of the contactor resolutions around alloys the sequences of the contactor resolutions and social private stores of the contactor resolutions and social private the stores of the contactor resolutions and social private the stores of the contactor resolutions and social private the stores of the contactor resolutions and social private social stores of the contactor resolutions and social private the tasebar DNO, The KBA specify provides information almost the process through which A/Na, GNa, and DNa are insued (1). A/No ant currently sengred based on the specific of the training address pervalues in the MIP application form DEMERGOLUM! (1) Low-population white and outstate U.S. presentence or admented I A/N each, whereas other mater admented are of A/Na (for instance, an individual applying from a signable which A/Na (for instance, an individual applying from a specific for Size York more may be using out of 15 pointies for 3 500 Eight). Written each MIA with, CPH are assigned in a process bit neurosciencement order between (ii) and 90 (KMIRER), 2001 (1). Buth the sets of A/N assigned to different sizes and the sequence in terms and the applying from the sequence of the sequence.

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.

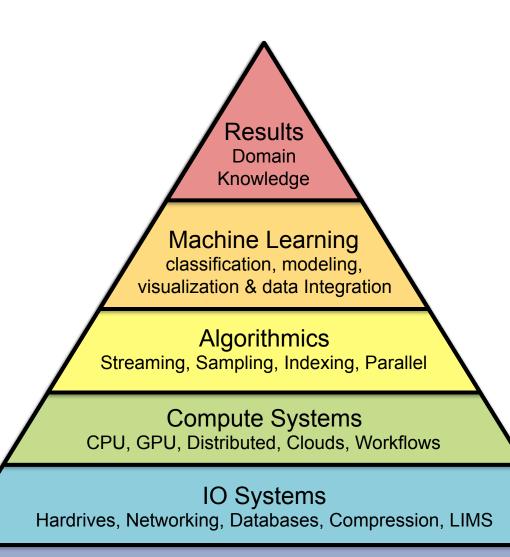
I han a the same

# 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



Sensors & Metadata Sequencers, Microscopy, Imaging, Mass spec, Metadata & Ontologies



# **BIOLOGICAL DATA SCIENCE**



Wednesday	7:30 pm 8:00 pm	Introduction 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 Diclogical Software
Friday	4:30 pm	Master Locture
Friday	5:30 pm	7 Poster Session II and Cocktails
Friday	7:00 pm	Benquet
Saturday	9:00 am	# Human Biology

## Master Lecture



"Homomorphic encryption as a tool to preserve privacy in genomic computation"

Friday @ 4:30pm

Kristin Lauter, Ph.D. Microsoft Research

# Schedule Change



### Saturday Morning: Human Biology

Mark Gerstein will present first in the session

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

**Eric Perakslis, Ph.D.** Harvard Medical School

# **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 / #biodata 14