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
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?

<table>
<thead>
<tr>
<th>Unit</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Byte</td>
<td>1</td>
</tr>
<tr>
<td>Kilobyte</td>
<td>1,000</td>
</tr>
<tr>
<td>Megabyte</td>
<td>1,000,000</td>
</tr>
<tr>
<td>Gigabyte</td>
<td>1,000,000,000</td>
</tr>
<tr>
<td>Terabyte</td>
<td>1,000,000,000,000</td>
</tr>
<tr>
<td>Petabyte</td>
<td>1,000,000,000,000,000,000</td>
</tr>
</tbody>
</table>

*Technically a kilobyte is $2^{10}$ and a petabyte is $2^{50}$*
How much is a petabyte?

100 GB / Genome
4.7 GB / DVD
~20 DVDs / Genome
\[ \times \]
10,000 Genomes

= 1 PB 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!

![Graph showing DNA Data Tsunami growth from 2014 to 2018 with an estimate of ~1 exabyte by 2018.](image-url)
DNA Data Tsunami

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

![Graph showing DNA data growth per year from 2014 to 2024. By 2018, ~1 exabyte, and by 2024, ~1 zettabyte.](image)
How much is a zettabyte?

<table>
<thead>
<tr>
<th>Unit</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Byte</td>
<td>1</td>
</tr>
<tr>
<td>Kilobyte</td>
<td>1,000</td>
</tr>
<tr>
<td>Megabyte</td>
<td>1,000,000</td>
</tr>
<tr>
<td>Gigabyte</td>
<td>1,000,000,000</td>
</tr>
<tr>
<td>Terabyte</td>
<td>1,000,000,000,000,000</td>
</tr>
<tr>
<td>Petabyte</td>
<td>1,000,000,000,000,000,000,000</td>
</tr>
<tr>
<td>Exabyte</td>
<td>1,000,000,000,000,000,000,000,000</td>
</tr>
<tr>
<td>Zettabyte</td>
<td>1,000,000,000,000,000,000,000,000,000</td>
</tr>
</tbody>
</table>
How much is a zettabyte?

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

X

10,000,000,000 Genomes

= 

1 ZB Data
200,000,000,000 DVDs

150,000 miles of DVDs
~ ½ distance to moon

Both currently ~100Pb
And growing exponentially
Sequencing Centers 2014

Next Generation Genomics: World Map of High-throughput Sequencers
http://omicsmaps.com
The DNA Data Deluge
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
Identifying Personal Genomes by Surname Inference

Sharing sequencing data sets without identifiers, we report that surnames can be reused repeatedly on the Internet. These publicly accessible data sets can be used to manipulate the identity of the relatives of those seeking to use them for identification for U.S. states. This further heightens the probability of identity theft by generating a false identity for a real individual.

Predicting Social Security numbers from public data

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 obtained a correlation of 0.5 between individuals’ SSNs and their birth data and found that for younger cohorts the correlation allows statistical inference of private SSNs. The inference is 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 should raise privacy concerns at the SBA.

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

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

- IO Systems
  - Hardrives, Networking, Databases, Compression, LIMS

- Compute Systems
  - CPU, GPU, Distributed, Clouds, Workflows

- Algorithmics
  - Streaming, Sampling, Indexing, Parallel

- Machine Learning
  - classification, modeling, visualization & data Integration

- Results
  - Domain Knowledge
<table>
<thead>
<tr>
<th>Day</th>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wednesday</td>
<td>7:30 pm</td>
<td>Introduction</td>
</tr>
<tr>
<td></td>
<td>8:00 pm</td>
<td>Keynote Speaker</td>
</tr>
<tr>
<td>Thursday</td>
<td>9:00 am</td>
<td>1. Data and Data Mining I</td>
</tr>
<tr>
<td>Thursday</td>
<td>1:30 pm</td>
<td>2. Data and Data Mining II</td>
</tr>
<tr>
<td>Thursday</td>
<td>3:00 pm</td>
<td>3. Poster Session I</td>
</tr>
<tr>
<td>Thursday</td>
<td>4:30 pm</td>
<td>Wine and Cheese Party</td>
</tr>
<tr>
<td>Thursday</td>
<td>7:30 pm</td>
<td>4. Compute Infrastructure</td>
</tr>
<tr>
<td>Friday</td>
<td>9:00 am</td>
<td>5. Algorithmics</td>
</tr>
<tr>
<td>Friday</td>
<td>1:30 pm</td>
<td>6. Biological Software</td>
</tr>
<tr>
<td>Friday</td>
<td>4:30 pm</td>
<td>Master Lecture</td>
</tr>
<tr>
<td>Friday</td>
<td>5:30 pm</td>
<td>7. Poster Session II and Cocktails</td>
</tr>
<tr>
<td>Friday</td>
<td>7:00 pm</td>
<td>Banquet</td>
</tr>
<tr>
<td>Saturday</td>
<td>9:00 am</td>
<td>8. Human Biology</td>
</tr>
</tbody>
</table>
“Homomorphic encryption as a tool to preserve privacy in genomic computation”

Friday @ 4:30pm
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 Peraklis, Ph.D.
Harvard Medical School
Keynote Introduction

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

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
Thank you!
@mike_schatz / #biodata14