Big Data
Why it matters

Patrice KOEHL
Department of Computer Science
Genome Center
UC Davis

The three I’s of Big Data

Big Data is:

- Ill-defined (what is it?)
- Immediate (we need to do something about it now)
- Intimidating (what if we don’t)

(loosely adapted from Forbes)

<table>
<thead>
<tr>
<th>Byte</th>
<th>Kilobyte</th>
<th>Megabyte</th>
<th>Gigabyte</th>
<th>Terabyte</th>
<th>Petabyte</th>
<th>Exabyte</th>
<th>Zettabyte</th>
<th>Yottabyte</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1000 bytes</td>
<td>1000 KB</td>
<td>1000 MB</td>
<td>1000 GB</td>
<td>1000 TB</td>
<td>1000 PB</td>
<td>1000 ZB</td>
<td>1000 YB</td>
</tr>
</tbody>
</table>
Big Data: Volume

Big Data: Volume

Big Data: Volume, Velocity
Big Data: Volume, Velocity, Variety

- Numbers
- Text
- Images
- Sound

Big Data: Challenges

- Volume and Velocity
- Variety
  - Structured, Unstructured....
  - Images, Sound, Numbers, Tables,...
- Security
- Reliability, Integrity, Validity

**Large N:**
“Any dataset that is collected by a scientist whose data collection skills are far superior to her analysis skills”

**Computing issues:**
- Data transfer
- Scalability of algorithms
- Memory limitations
- Distributed computing
Big Data: Challenges

Visualization issues:
The black screen problem

Big Data: Challenges and Opportunities

- Fourth Paradigm: data driven science
  - Basic Data ➔ Knowledge ➔ Societal Benefit

- Holistic approaches to major research efforts
- New paradigms in computing
- Digital Humanities

Big Data Dreams: Genomics

Table 2: Comparison of prices of large government projects circa 1996 with their projected useful life-span.

<table>
<thead>
<tr>
<th>Proposed project</th>
<th>Projected cost ($ billion)</th>
<th>Target completion date</th>
<th>Estimated life-span (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space Station Freedom</td>
<td>6.0</td>
<td>1996</td>
<td>30</td>
</tr>
<tr>
<td>Earth Observing System</td>
<td>12.0</td>
<td>2000</td>
<td>10</td>
</tr>
<tr>
<td>Superconducting Super Collider</td>
<td>11.0</td>
<td>1999</td>
<td>30</td>
</tr>
<tr>
<td>Human Genome Project</td>
<td>3.0</td>
<td>2003</td>
<td>Perpetual</td>
</tr>
<tr>
<td>Habitable Space Telescope</td>
<td>5.5</td>
<td>1994</td>
<td>15 to 20</td>
</tr>
</tbody>
</table>
Big Data Dreams: Genomics

Genomics: Sequencing costs

Genomics: Game changing technologies

**Illumina HiSeq 2000**

- Capable of 600 Gb per run -> 1,000+
- 55 Gb/day
- 6 billion paired-end reads
- < $4,000 per human/plant genome
- < $200 per transcriptome
- Multiplex 384 pathogen isolates/lane
- ≈ $10 (+ $50 library construction)/isolate

Challenges: Library preparation & data analysis

Gary Schroth (Illumina): “A single lab with one HiSeq is able to generate more sequences than was in GenBank in 2009, every four days”
Massively parallel DNA sequencing
- 2 Illumina Genome Analyzers
- 1 Illumina HiSeq 2000, 2 MiSeq
- 1 Roche 454 Junior
- 1 Pacific Biosystems RS

GoldenGate SNP genotyping
- iScan, BeadArray & BeadExpress

Cancer Genomics: Molecular Diagnostics

Genomics: actual costs

“A single lab with one HiSeq is able to generate more sequences than was in GenBank in 2009, every four days.”

Gary Schroth (Illumina)
“A single lab with one HiSeq is able to generate more sequences than was in GenBank in 2009, every four days.”

Gary Schroth (Illumina)

Assembling 22GB conifer genome:

- 16 billion pair reads (100 bases)
- 10 days for error correction
- 6 days for assembling "super-reads"
- 60 days to build contigs/scaffold
- 8 days to fill in gaps


Social Consequences of Commodity Sequencing

- The danger of misuse
  predict sensitivities to various industrial or environmental agents
- The impact of information that is likely to be incomplete
  an indication of a 25 percent increase in the risk of cancer?
- Reversal of knowledge paradigm
- Are the "products" of the Human Genome Project to be patented and commercialized?
  Myriad genetics and BRCA1/2
- How to educate about genetic research and its implications?