

Using Big D to Fight the Big C

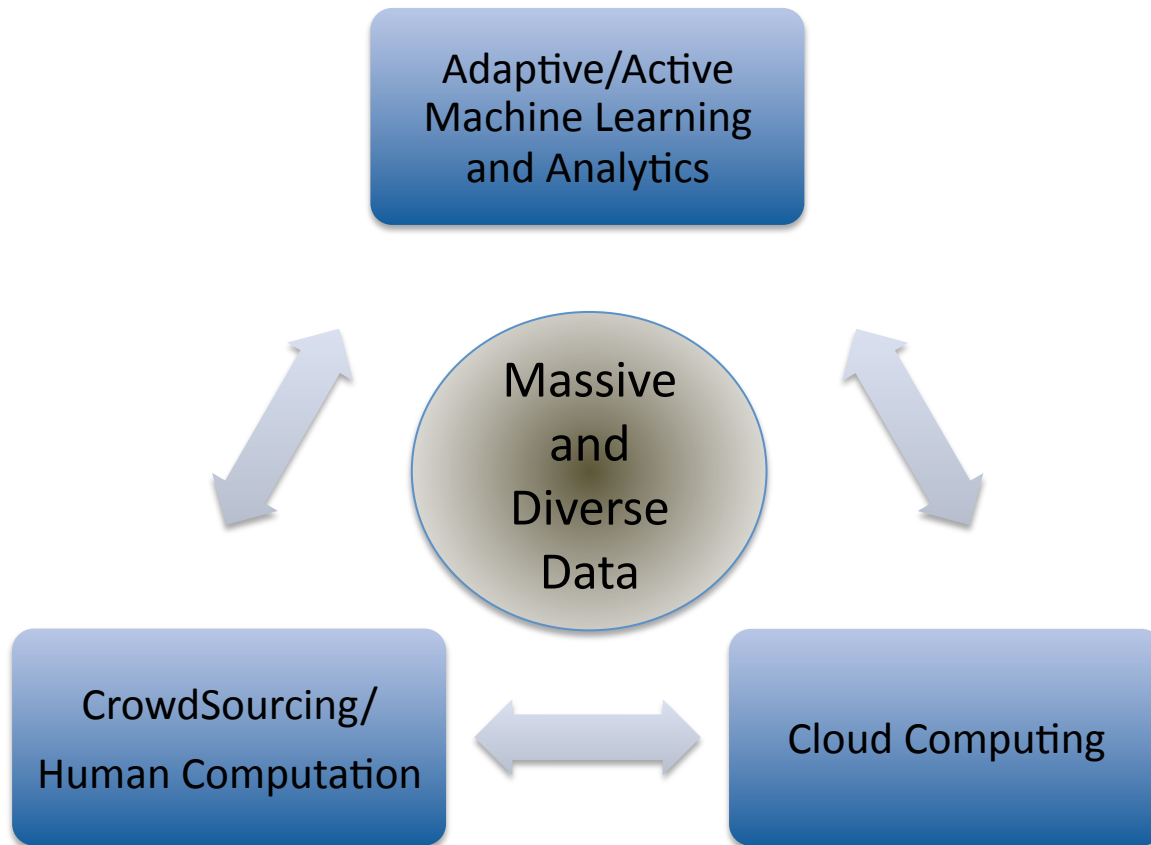
David Patterson
February 14, 2013



Outline

- AMPLab Overview
- How can Computer Scientists Help?
- Genetics 101
- Berkeley's fastest genome aligner: SNAP
- Fighting Cancer in the Future
- A 1M Genome Cancer Warehouse
- Conclusion

AMP Lab: Algorithms, Machines & People



- 2011-2017
- Machine Learning, Databases, Systems, + Networking
- Release Berkeley Data Analysis Stack (BDAS)

AMP Expedition



Office of Science and Technology Policy
Executive Office of the President
New Executive Office Building
Washington, DC 20502

FOR IMMEDIATE RELEASE

March 29, 2012

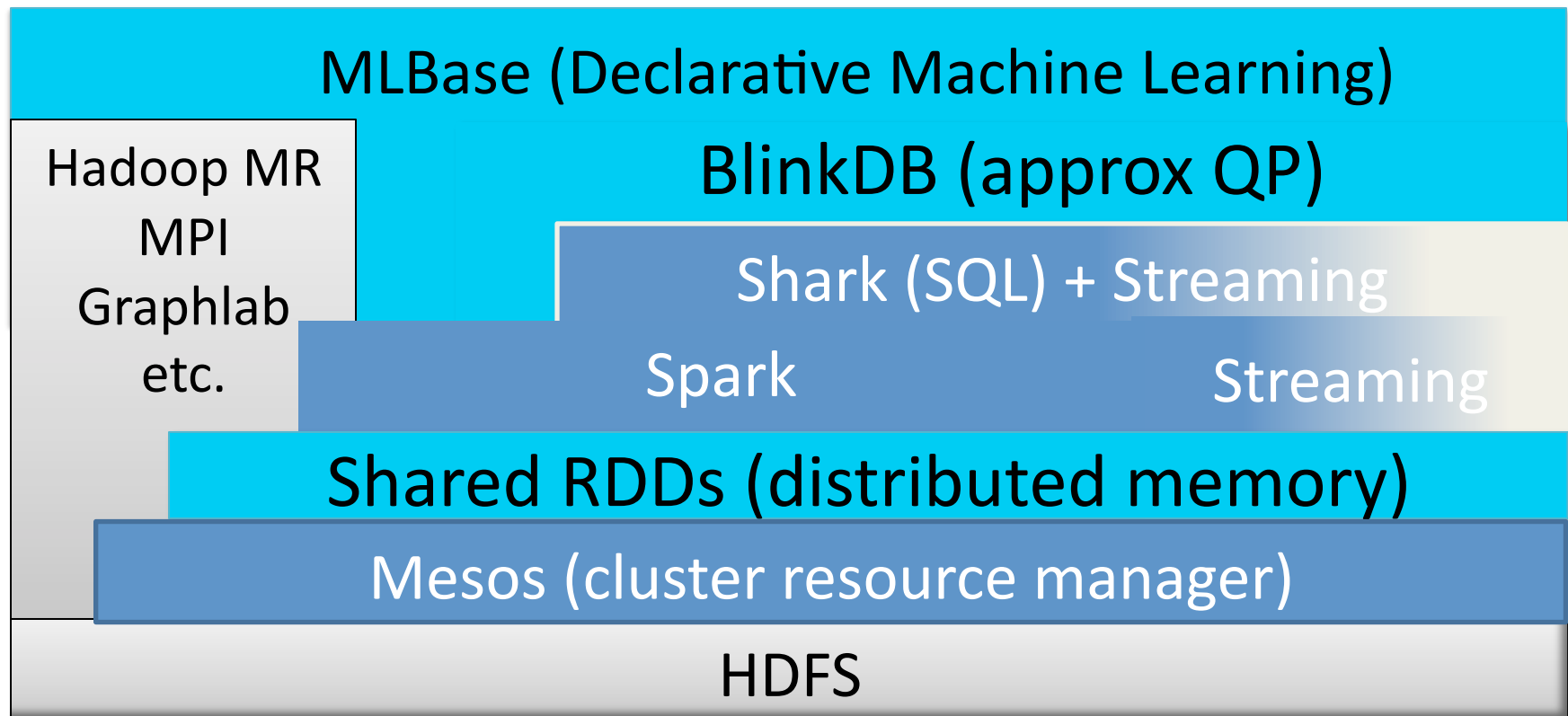
Contact: Rick Weiss 202 456-6037 rweiss@ostp.eop.gov
Lisa-Joy Zgorski 703 292-8311 lisajoy@nsf.gov

OBAMA ADMINISTRATION UNVEILS “BIG DATA” INITIATIVE: ANNOUNCES \$200 MILLION IN NEW R&D INVESTMENTS

National Science Foundation: In addition to funding the Big Data solicitation, and keeping with its focus on basic research, NSF is implementing a comprehensive, long-term strategy that includes new methods to derive knowledge from data; infrastructure to manage, curate, and serve data to communities; and new approaches to education and workforce development. Specifically, NSF is:

- Encouraging research universities to develop interdisciplinary graduate programs to prepare the next generation of data scientists and engineers;
- Funding a \$10 million Expeditions in Computing project based at the University of California, Berkeley, that will integrate three powerful approaches for turning data into information - machine learning, cloud computing, and crowd sourcing;

Berkeley Data Analytics System



 3rd party  AMPLab (released)  AMPLab (in progress)

What is Spark?



- Fast, MapReduce-like engine
 - In-memory storage for very fast iterative queries
 - General execution graphs
 - Up to 100x faster than Hadoop (2-10x even on-disk data)
- Language-integrated API in Scala, Java, + Python
- Compatible with Hadoop's storage APIs
 - Can access HDFS, HBase, S3, SequenceFiles, etc
- Matei Zahari will talk about Spark in PhD Session

Where CS can Help with War on Cancer

1. Create easy-to-use, fast, accurate, reliable genetic analysis software pipelines
2. Create massive, cheap, easy-to-use, privacy-protecting repository for cancer treatments showing tumor genomes over time, therapies, and outcomes

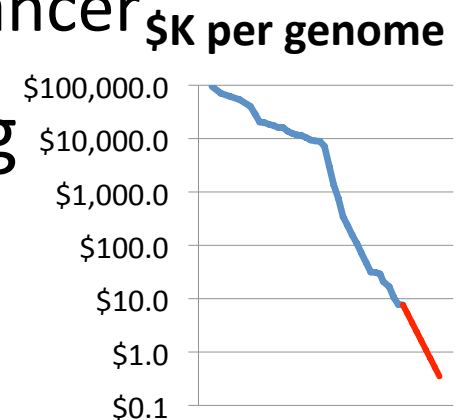
Cancer: Good and Bad News

- Bad news: Cancer is pervasive: 1/3 ♀, 1/2 ♂
- Good news: Cancer is a genetic disease
 - Accidental DNA cell copy flaws + carcinogen-based mutations lead to cancer

- Good news: Sequencing Price Falling

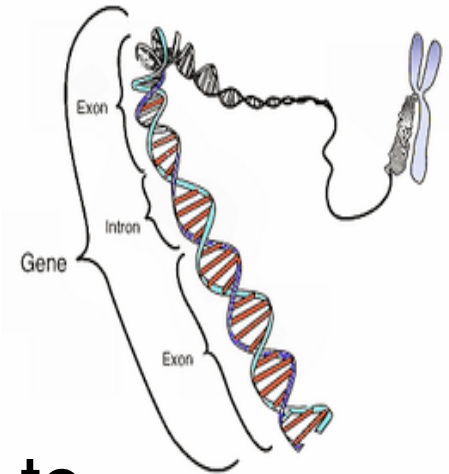
- Bad news:

- DNA processing SW built by scientists
- DNA Data Processing costs > DNA Wet lab costs
- No repository of tumor DNA over time + treatments + patient outcomes to enable personalized medicine



Genetics 101

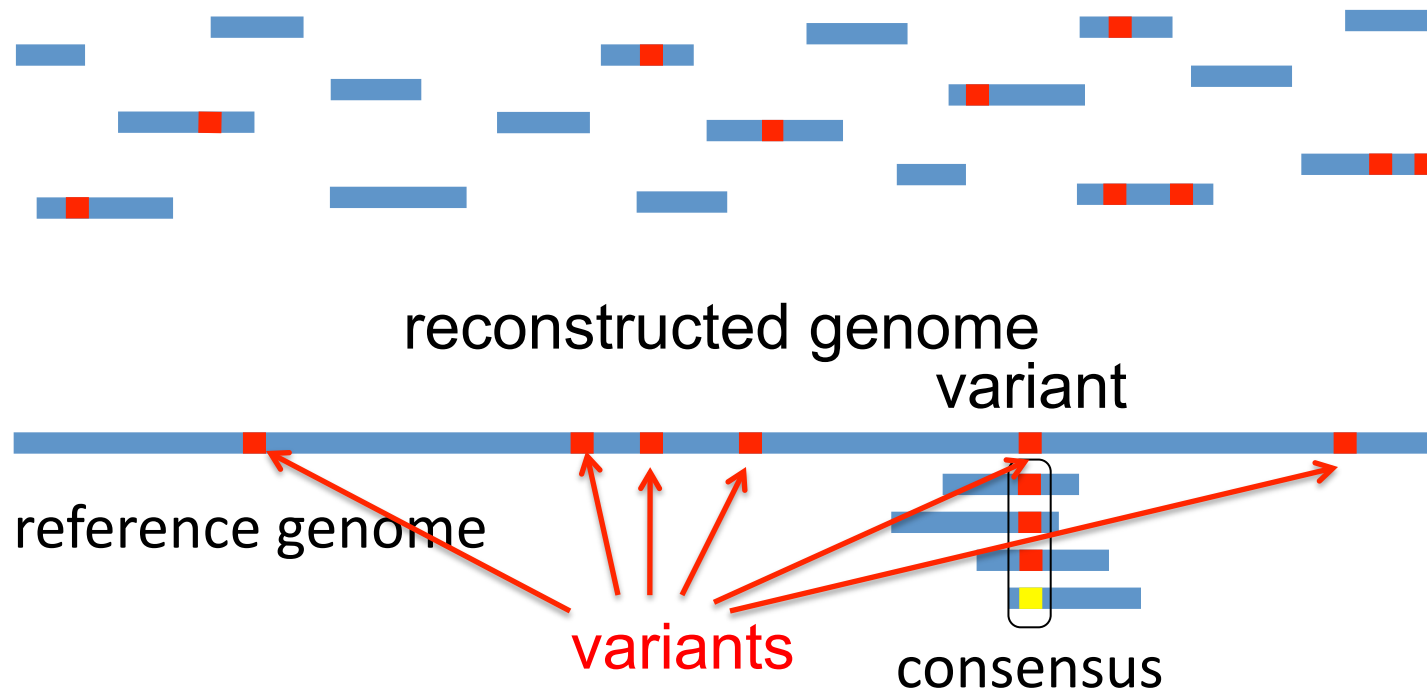
- Double stranded DNA: ½ Mom, ½ Dad
- *Base pairs*: links between 2 bases
 - Guanine-Cytosine, Adenine-Thymine
 - Human DNA = 3.2B base pairs
- *Gene*: unit of heredity that corresponds to stretches of DNA
 - Humans have ~25,000 genes, avg ~25,000 bp / gene
 - Metabolic role of gene to produce *enzymes*, which controls a protein
- *Pathway*: chemical reactions in a cell that maintain organism, controlled in part by genes



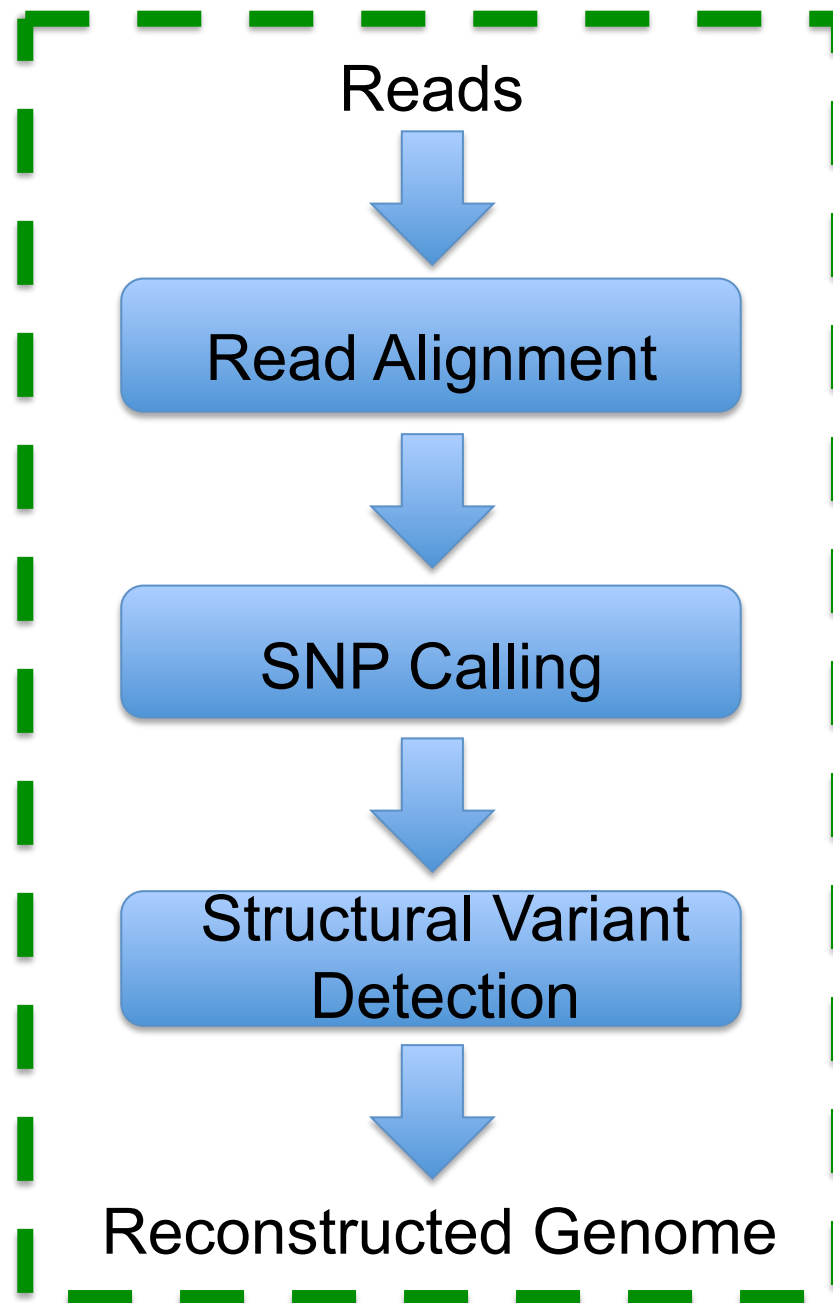
Cancer 101

- Normal cells have built-in limits to cell division/growth
- Cancer cells are immortal and mutate
- Cancer tumors are heterogeneous colonies of cancer cells
- Later spread throughout body (metastasize)
- Some pathways are associated with cancer cell growth / survival
- Taxonomy: location in body vs. genetics
 - 1000s of subtypes of cancer?
- 1.6M new cases per year in US

Genetic Sequencing Overview



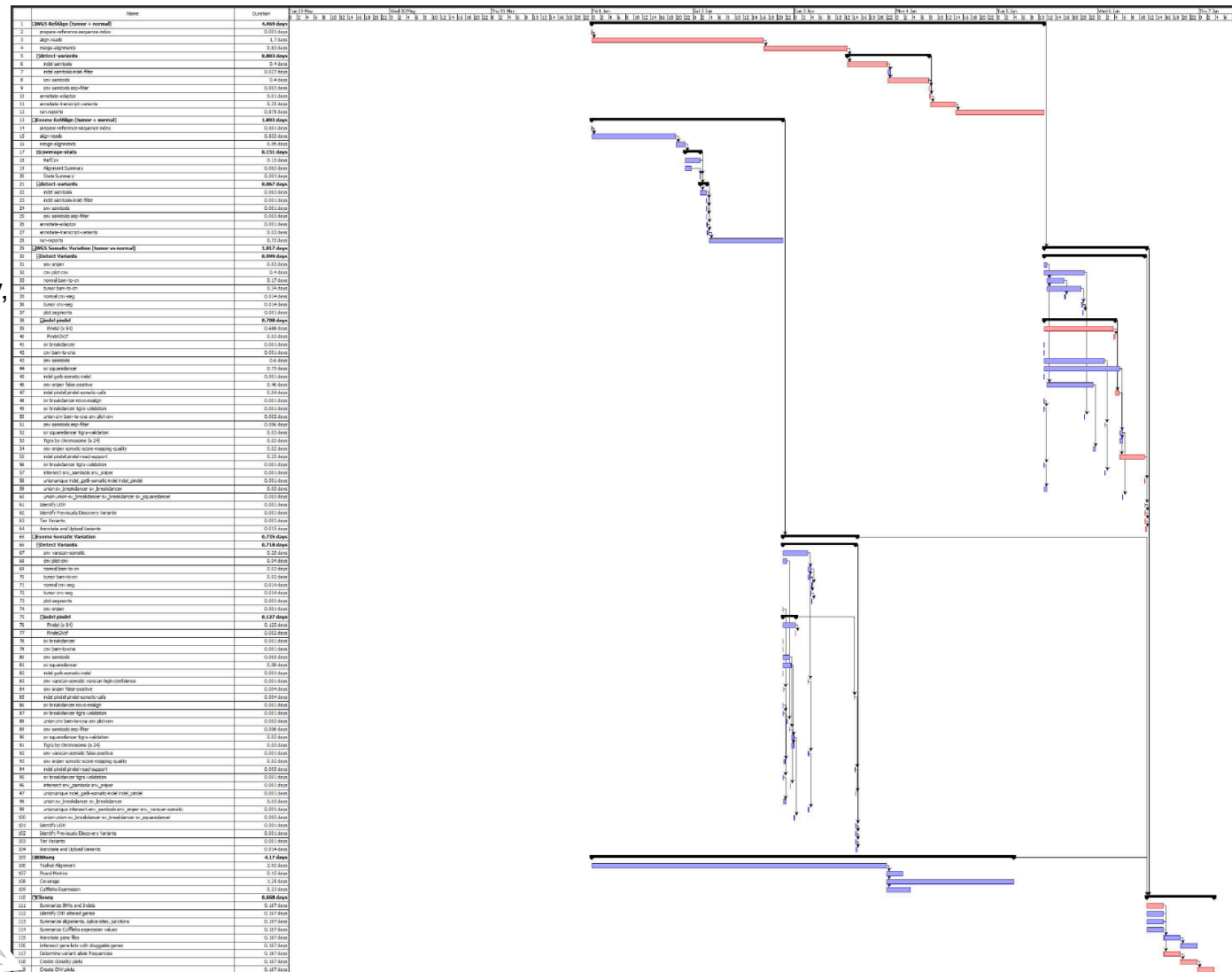
Genetic
Data
Processing
Software
Pipeline



“The reality of implementing such a pipeline and optimizing underlying tasks is complex”

Malachi Griffith,
Washington University,
August 19, 2012
“Cancer genome and
transcriptome
sequencing – analysis
challenges
and bottlenecks”

119th
step



Lack of SW Engineering by Scientists

- 2008 survey
 - Most scientists are self-taught in programming
 - Only $\frac{1}{3}$ think formal training in SW Eng is important
 - $< \frac{1}{2}$ have a good understanding of SW testing
- For example, bug in SW supplied by another research lab forced UCSD Scripps Prof to retract 5 papers
 - *Science, Journal of Molecular Biology, and Proceedings of the National Academy of Sciences*

“Computational science: ...Error...why scientific programming does not compute,” by Zeeya Merali, 13 October 2010, *Nature* 467, 775-777

Pipeline goals

Build a faster, more scalable, more accurate pipeline

Apply to both medicine & research → focus on cancer

Interdisciplinary team: UC Berkeley, Intel, Microsoft, UCSF

AMP-Microsoft-Intel Genome Team

UC Students/ Post-Docs

- Ma'ayan Bresler
- Kristal Curtis
- Jesse Liptrap
- Ameet Talwalkar
- Jonathan Terhorst
- Matei Zaharia
- Yuchen Zhang

Expertise

- Computational Biology/Medicine
- Machine Learning
- Systems

External

- Bill Bolosky (MS/MSR)
- Mishali Naik (Intel)
- Paolo Narvaez (Intel)
- Ravi Pandya (MS)
- Abirami Prabhakaran (Intel)
- Taylor Sittler (UCSF)
- Gans Srinivasa (Intel)
- Arun Wiita (UCSF)

UC Faculty

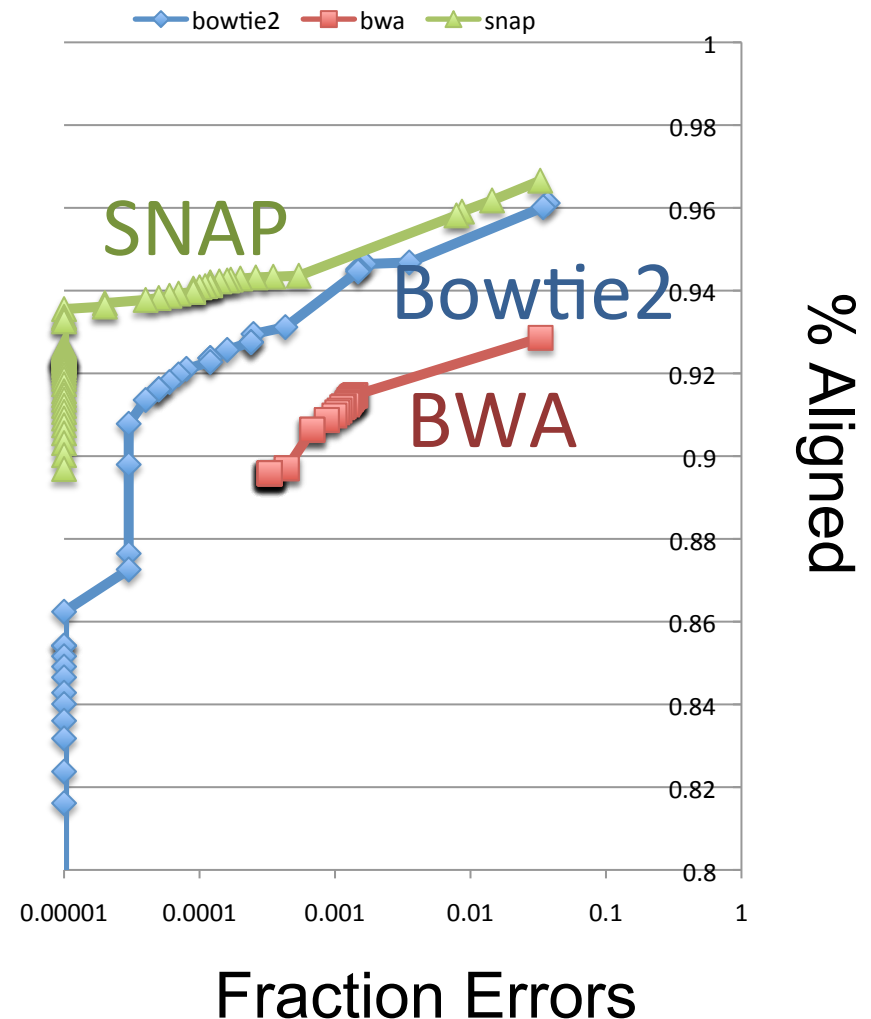
- Michael Jordan
- David Patterson
- Satish Rao
- Scott Shenker
- Yun Song
- Ion Stoica

First Result: SNAP

- Scalable Nucleotide Alignment Program (SNAP)
- Longer seeds, Overlapping seeds, $O(nd)$ vs. $O(n^2)$ edit distance [Ukkonen]

Aligner	Reads/ sec	Time (hours)
Bowtie2	14,400	22
BWA	9,000	35
SNAP	180,000	2

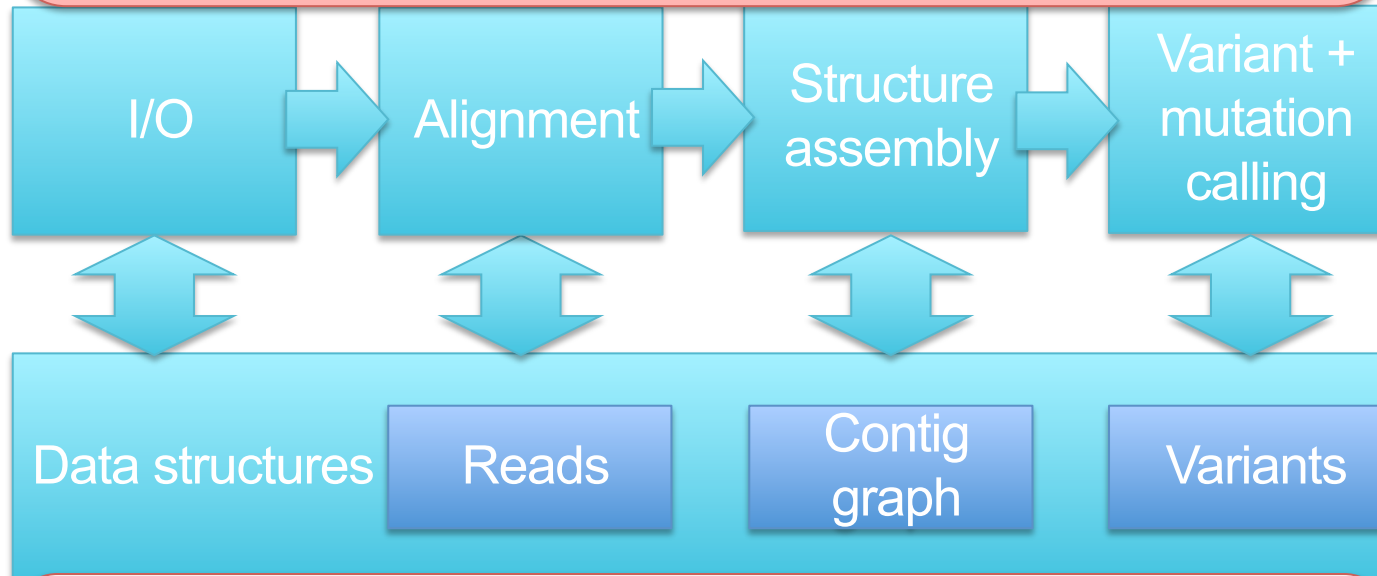
For 16 cores



<http://snap.cs.berkeley.edu/>

SNAP C++ pipeline engine

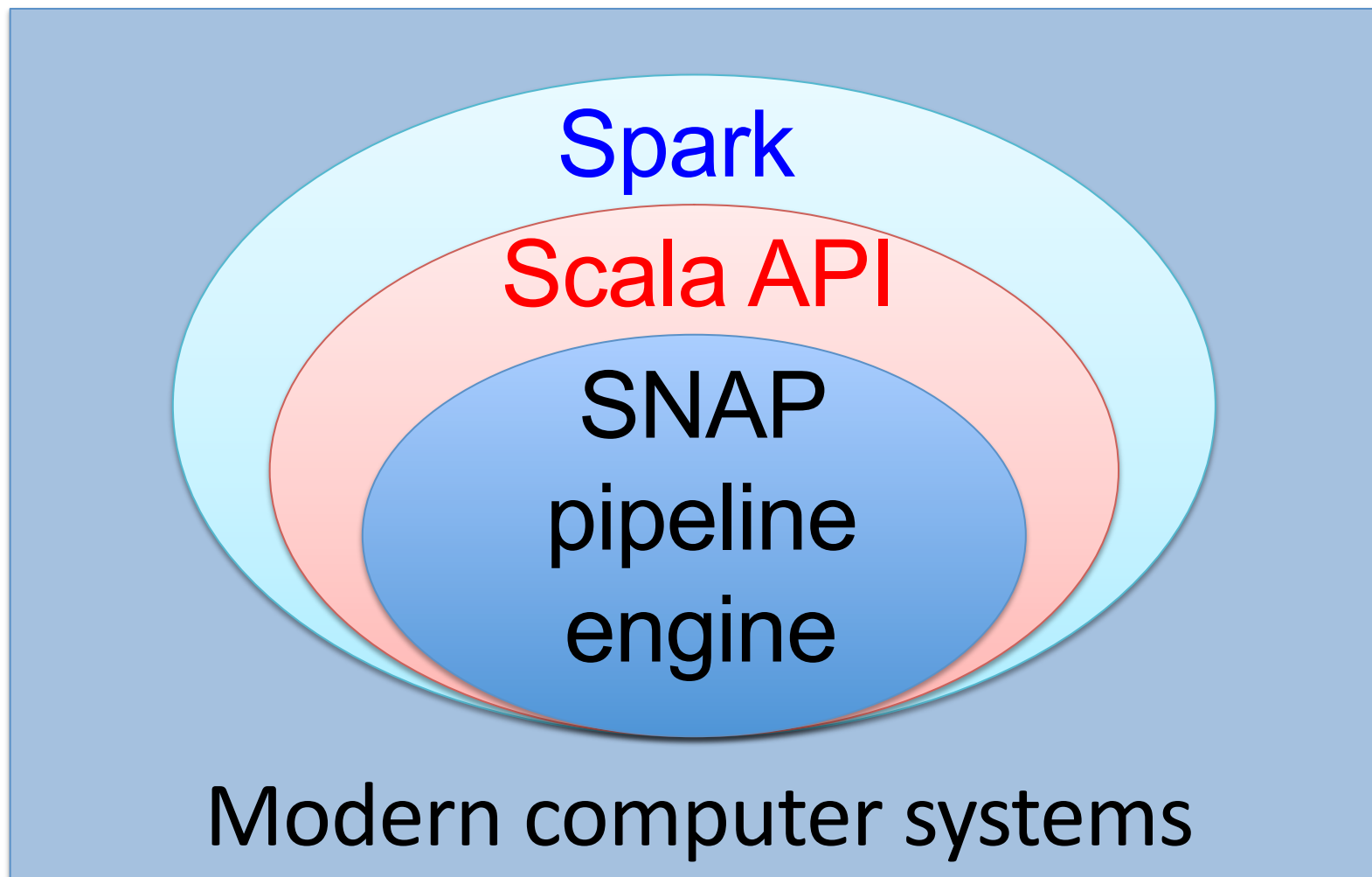
Many-core shared-memory parallelism
Fast multi-threaded I/O
Low-level optimization



Common
algorithms
+ data
structures

End-to-end in-memory pipeline
Efficient memory usage

Pipeline design

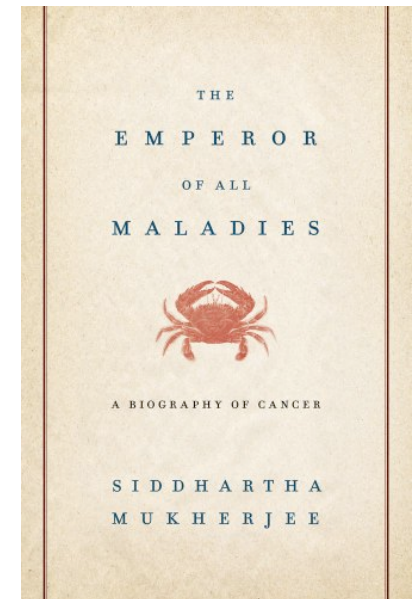


Where CS can Help with War on Cancer

1. Create easy-to-use, fast, accurate, reliable genetic analysis software pipelines
2. Create massive, cheap, easy-to-use, privacy-protecting repository for cancer treatments showing tumor genomes over time, therapies, and outcomes

Fighting Cancer in Future

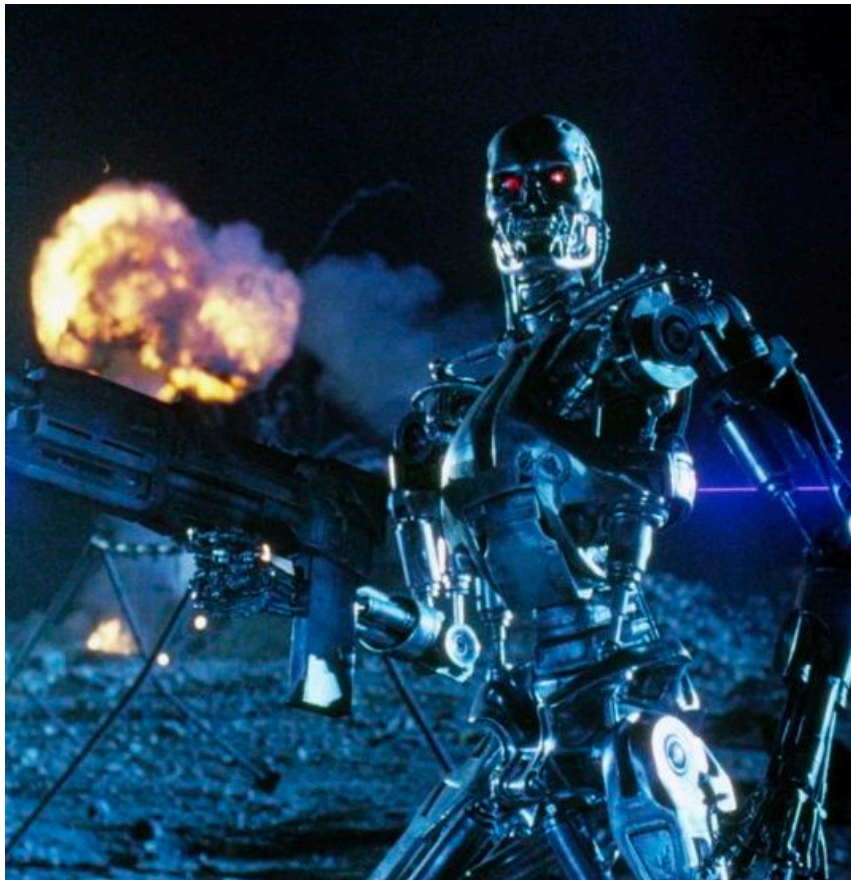
- Patient arrives at oncologist with entire sequence of cancer's genome
 - Mutations organized into key pathways
- Software identifies key pathways contributing to growth of cancer
- Therapies target these pathways after tumor removed
- Patients starts with 1st drug cocktail, switch to 2nd when cancer mutates, switch to 3rd when mutates again ...
 - Take some medicine for rest of life?
- 2050????



*Emperor of
All Maladies,
page 464*

Feel Like Character in SciFi Movie


- Terminator 2 to prevent SkyNet from being invented: Sarah Connors



A Million Cancer Genome Warehouse

A Million Cancer Genome Warehouse

- Why 1M?
- Sufficient number of samples for cancer subtypes to discover meaningful patterns in the data (enough statistical power)



David Haussler (UCSC)
David A. Patterson
Mark Diekhans (UCSC)
Armando Fox
Michael Jordan
Anthony D. Joseph
Singer Ma (UCSC)
Benedict Paten (UCSC)
Scott Shenker
Taylor Sittler (UCSF)
Ion Stoica

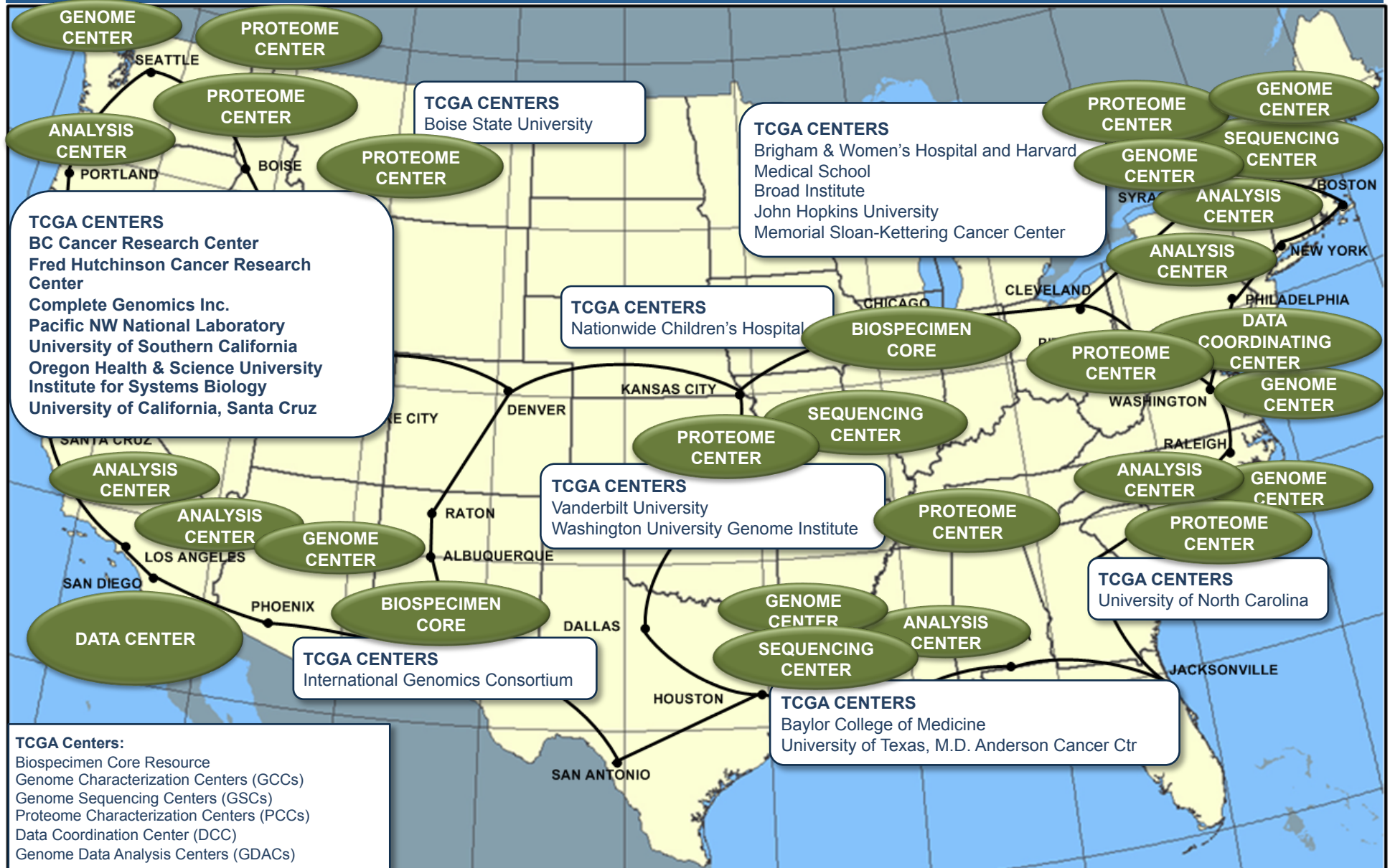
Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2012-211
<http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-211.html>

November 20, 2012

National effort: The Cancer Genome Atlas

10,000 tumors from 20 adult cancers



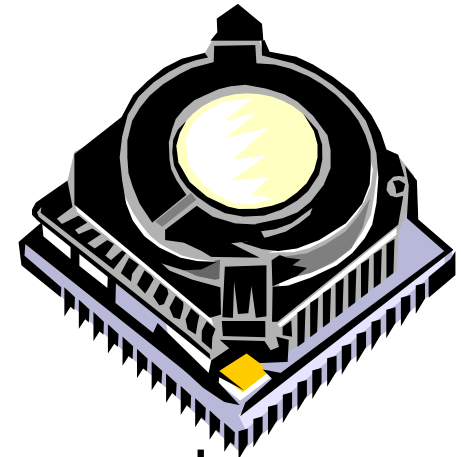
Cancer Genomics Hub (CGHub)

- Built at UCSD supercomputer center to store DNA information in for The Cancer Genome Atlas
- Designed for 50,000 genomes with average of 100 gigabytes per genome: 5 petabytes
- Currently 24,000 files from ~5,500 cases, ~60 gigabytes/case, in total 2 PB of downloads in first 6 months, routine 3 Gbit/s
- Total Cost ~ \$100/year/genome at 50K genomes, i.e. \$5M/year. The technology cost is about 1/2 the total
- Co-location opportunities in same data center for groups who want to compute on the data



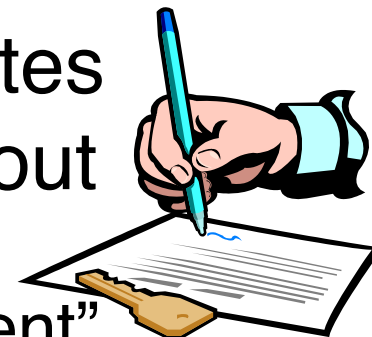
What would it cost to store and analyze 1M Cancer Genomes in 2014?

- Internet network bandwidth
100X cost datacenter networking,
computation should be near data
- Cloud computing demonstrates large
economies of scale of large computer warehouses
- White paper estimates ~ \$25/genome/year in 2014
to store and analyze 100 petabytes reliably
 - 25,000 disks and 100,000 processor cores
 - Including operating costs: space, electricity, operators
 - Including 2nd center to protect against disasters
- Even at \$1000 for wet lab costs, \$25/year is cheap



Non-Technical Challenges for 1MGW

- Patient data: Ethicists v. Disease Advocates
- Cannot aggregate data for research without consent
 - E.g., Wilbanks patient “Portable Legal Consent”
- Can’t use “my” data until we publish
 - CGHub lifts competing publication moratorium after data published or after 18 months, whichever first
- Claiming IP rights on discoveries made from data
 - CGHub data is not encumbered by IP provisions
- Access: all qualified researchers from all institutions
 - CGHub data open to all qualified researcher
- Who: Government? Business? Non-Profit?



Conclusion: Societal-Scale Big Data App

- Genetic sequencing costs 1,000,000X less
 - \$1000 per genome by end of year?
- Cancer: genetic disease that kills 0.6M/yr
- Chance for Computer Scientists to use Big Data technology to help fight Cancer(!)
 - Fast, accurate, easy to use genetics analysis pipeline
 - Fast, cheap, easy to use, privacy protecting repository of cancer genetics, treatments, outcomes
- Accelerate Personalized Cancer Therapy from ~2050 to ~2018?

Using Big D to Fight the Big C: Opportunity or Obligation?

- If a *chance* that Computer Scientists could help millions of cancer patients live longer and better lives, as moral people, *aren't we obligated to try?*

David Patterson,
“Computer Scientists May Have What It Takes to Help Cure Cancer,”
New York Times, 12/5/2011