Using Big D to Fight the Big C

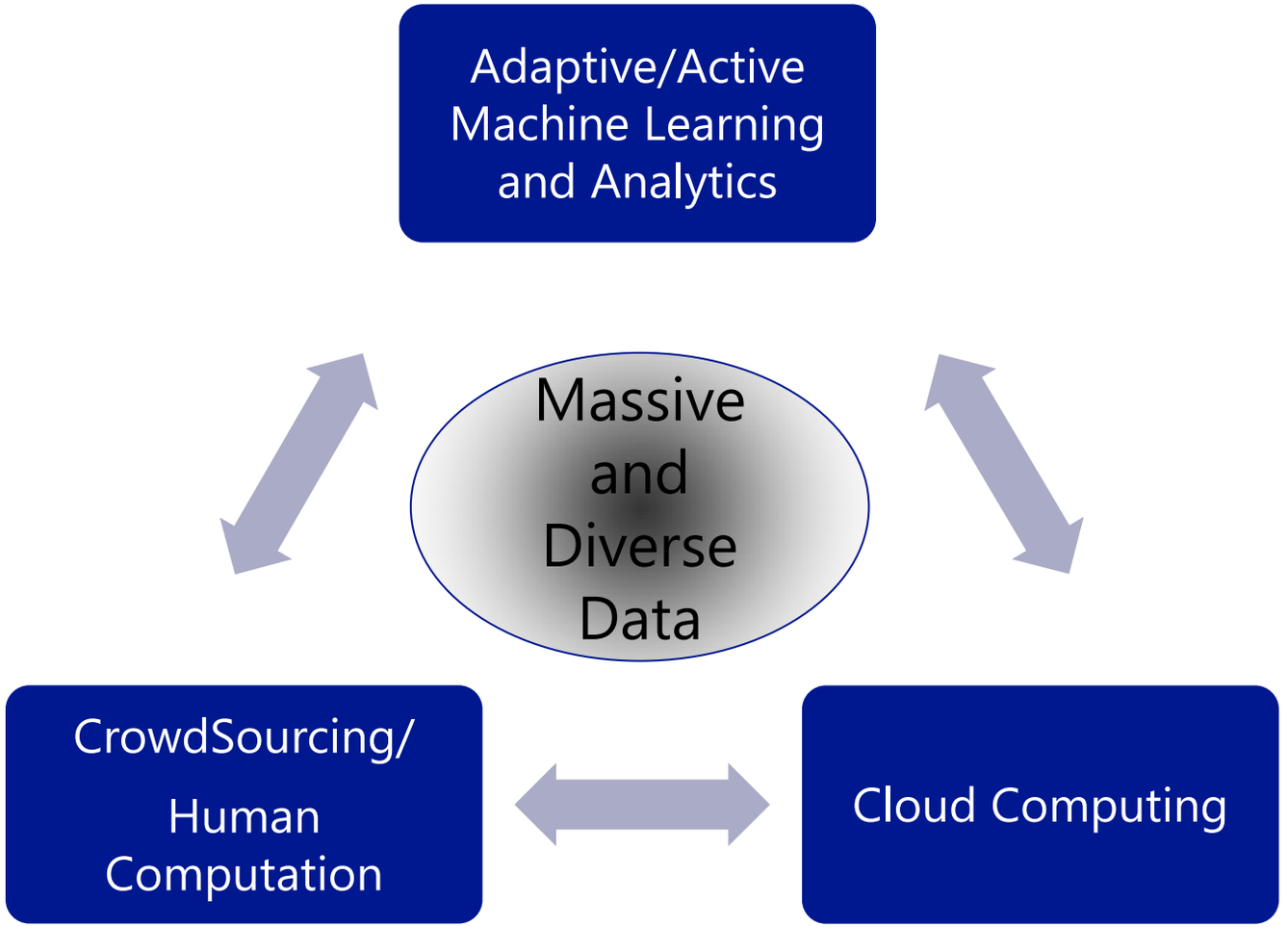
David Patterson

July 16, 2013

Outline

- AMPLab Overview
- How can Computer Scientists Help?
- Berkeley's fastest genome aligner: SNAP
- Fighting Cancer in the Future
- A 1M Genome Cancer Warehouse
- Benchmarks to Accelerate Progress
- 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

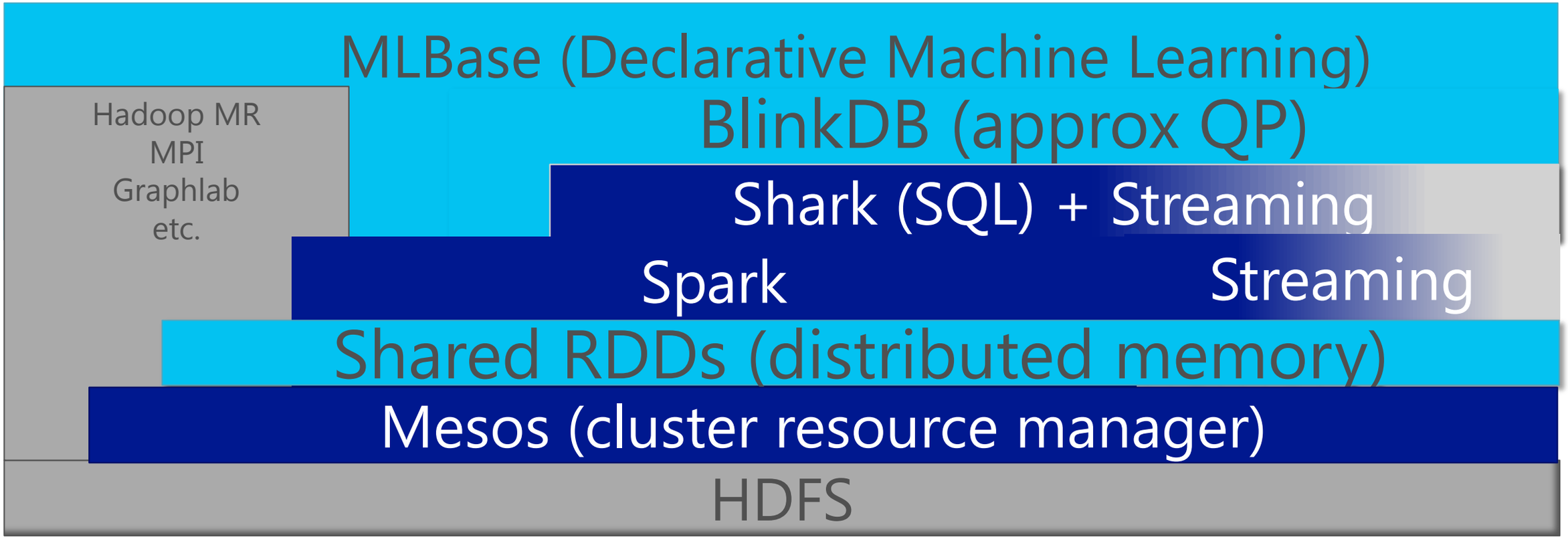
Contact: Rick Weiss 202 456-6037 rweiss@ostp.eop.gov
Lisa-Joy Zgorski 703 292-8311 lisajoy@nsf.gov

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:

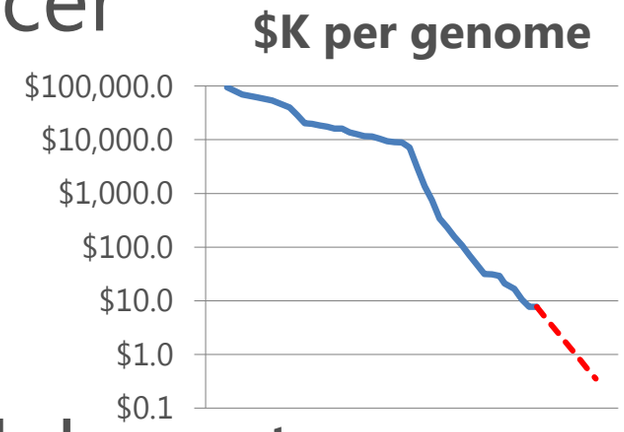
- Encouraging research universities to develop interdisciplinary graduate programs to prepare the next generation of data scientists and engineers;
- Funding a $10 million Expeditions in Computing project based at the University of California, Berkeley, that will integrate three powerful approaches for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

Berkeley Data Analytics System



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



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
3. Import benchmarking culture to accelerate progress

AMP-Microsoft-Intel Genome Team

UC Students/ Post-Docs

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

External

- Bill Bolosky (MS/MSR)
- Christopher Hartl (Broad)
- 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

Expertise

- Computational Biology/Medicine
- Machine Learning
- Systems

Collaborators

- David Haussler, UCSC
- Gaddy Getz, The Broad
- Mark Depristo, The Broad

Lack of SW Engineering by Scientists

- 2008 survey
  - Most scientists are self-taught in programming
  - Only ⅓ think formal training in SW Eng is important
  - < ½ 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

First Result: SNAP

- Scalable Nucleotide Alignment Program (SNAP)
- Came from question at AMP retreat
- Hash table approach vs. Burroghs-Wheeler Algorithm (BWA)
  - Longer seeds
  - Overlapping seeds
  - O(nd) vs. O(n^2) edit distance [Ukkonen]
- Even better as read lengths increase
  - 100 BP 2012 to 400 BP 2014

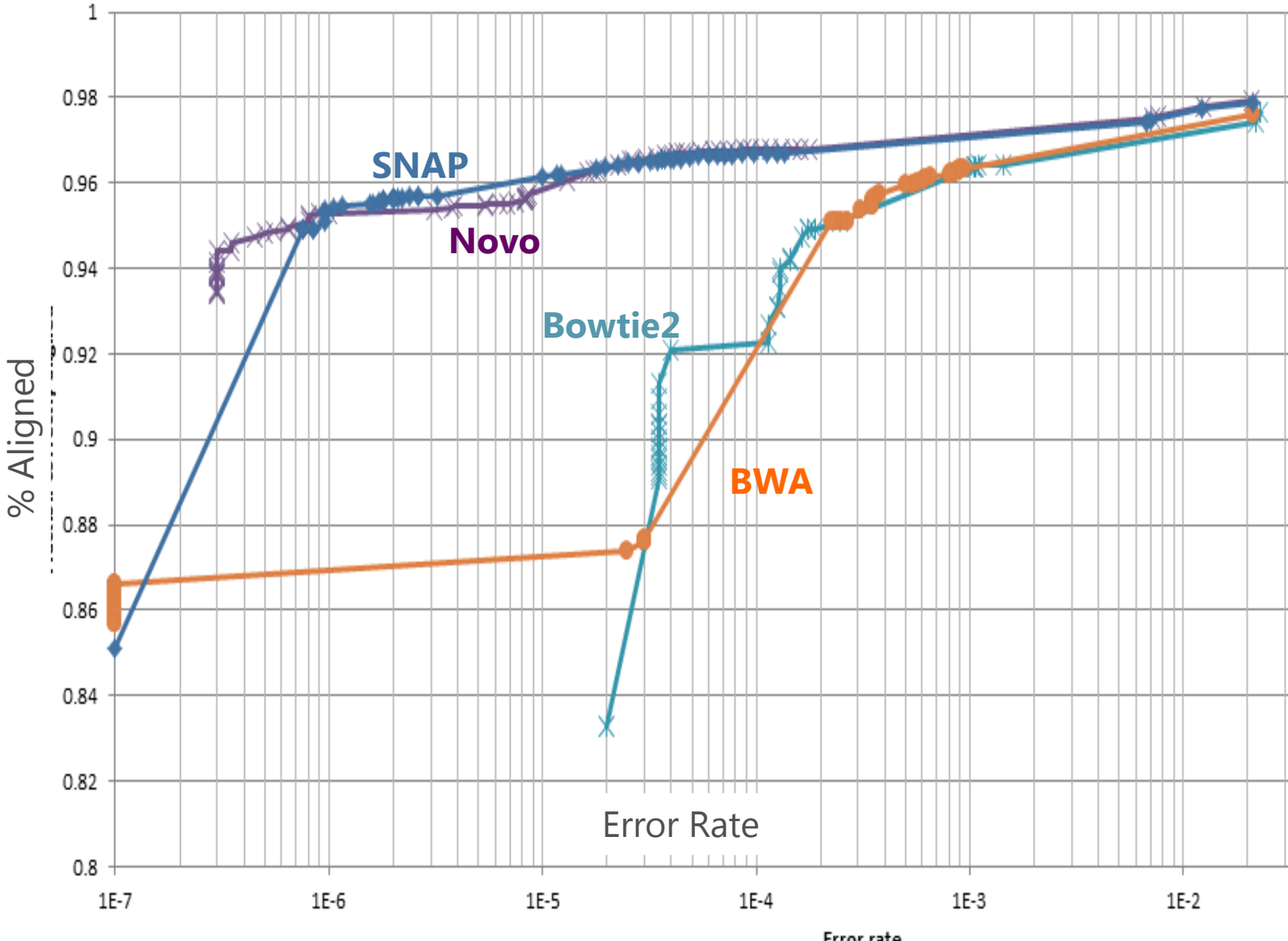
http://snap.cs.berkeley.edu/

SNAP vs. Other Aligners

- Hours/genome

SNAP	**2.5**
Novo	**33.5**
Bowtie2	**7.5**
BWA	**28.0**

- Added RNA aligner (open source win)
- Added sorted BAM output



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

OHSU DNA Pipeline: GATK vs SNAP

Whole Genome Sample: G15512.HCC1954.5 with >100x coverage(~3.9B reads)

Introduced Thread-level Parallelism

SNAP+ does in 10 hours vs. 93 hours

Step	# of Threads	Runtime (hours)
Read Alignment	16	16
Sampe	**1**	**59**
Import	1	19
Sort + Index	1	25
MarkDuplicates + Index	1	21
UnifiedGenotyper*	16	12
RealignerTargetCreator	16	1.5
IndelRealigner* + Index	**1**	**32.5**
BaseRecalibrator*	**1**	**96**
PrintReads* + Index + Flagstat	**1**	**44.5**
TOTAL (hours)		332

16	16
16	**12**
1	19
1	25
1	21
1	5.5
16	1.5
64	**16**
64	**15**
64	**26**
	169

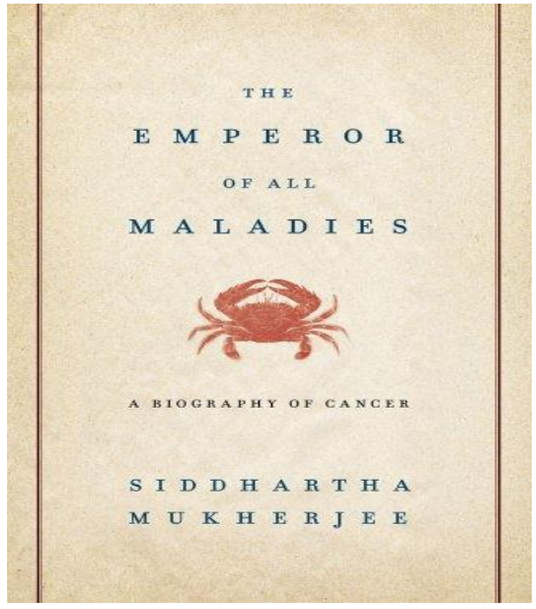
Introduced Cluster-level Parallelism

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
3. Import benchmarking culture to accelerate progress

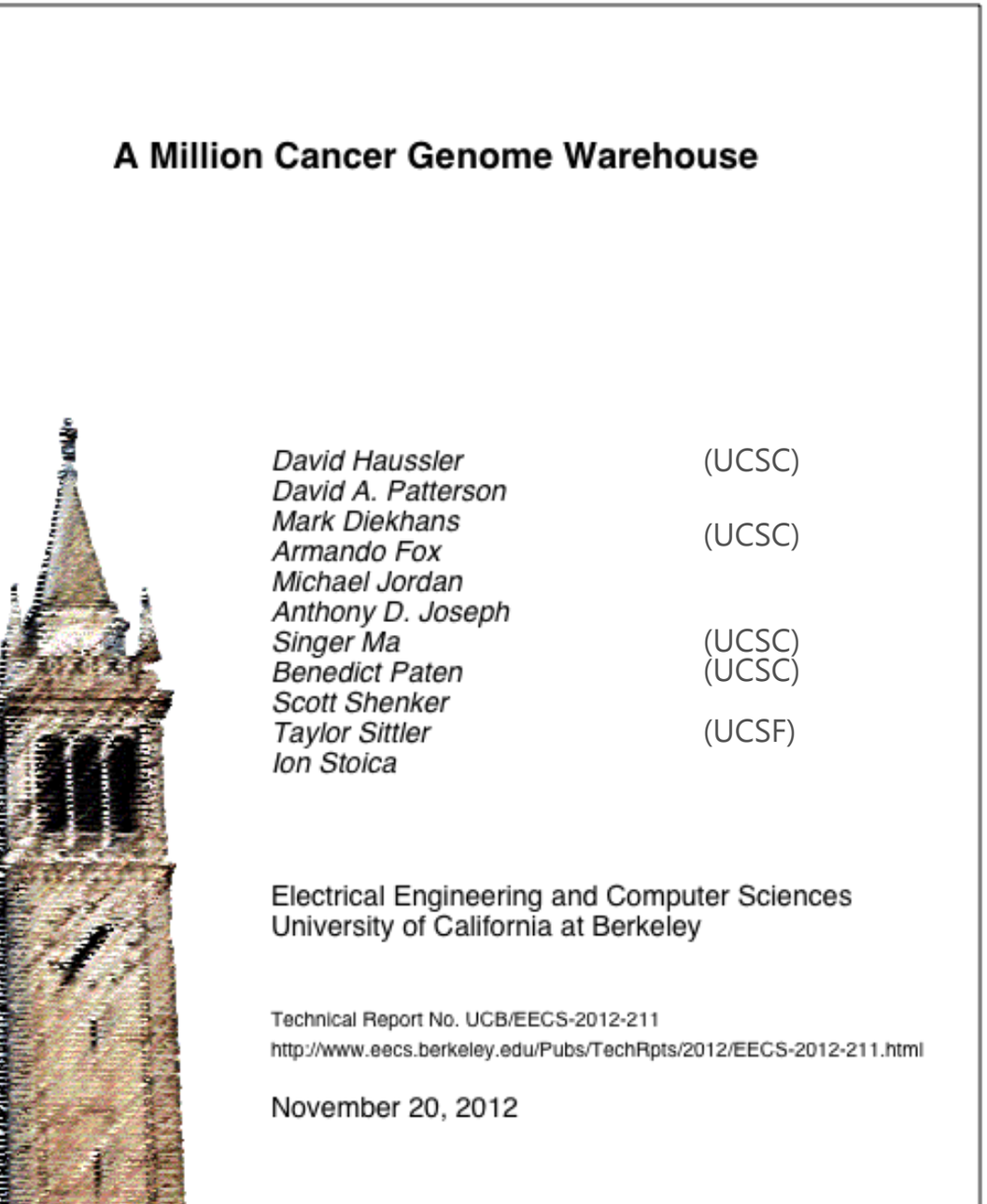
Fighting Cancer in Future

- 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

A Million Cancer Genome Warehouse



A Million Cancer Genome Warehouse

David Haussler (UCSC)
David A. Patterson
Mark Diekhans
Armando Fox (UCSC)
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

- Accelerate pace of cancer research and integration into clinical practice
- Why 1M?
  - Sufficient samples for cancer subtypes to discover meaningful patterns in the data (enough statistical power)
- Participating in effort to form international alliance with similar goals

Global Alliance to Enable Responsible Sharing of Genomic and Clinical Data

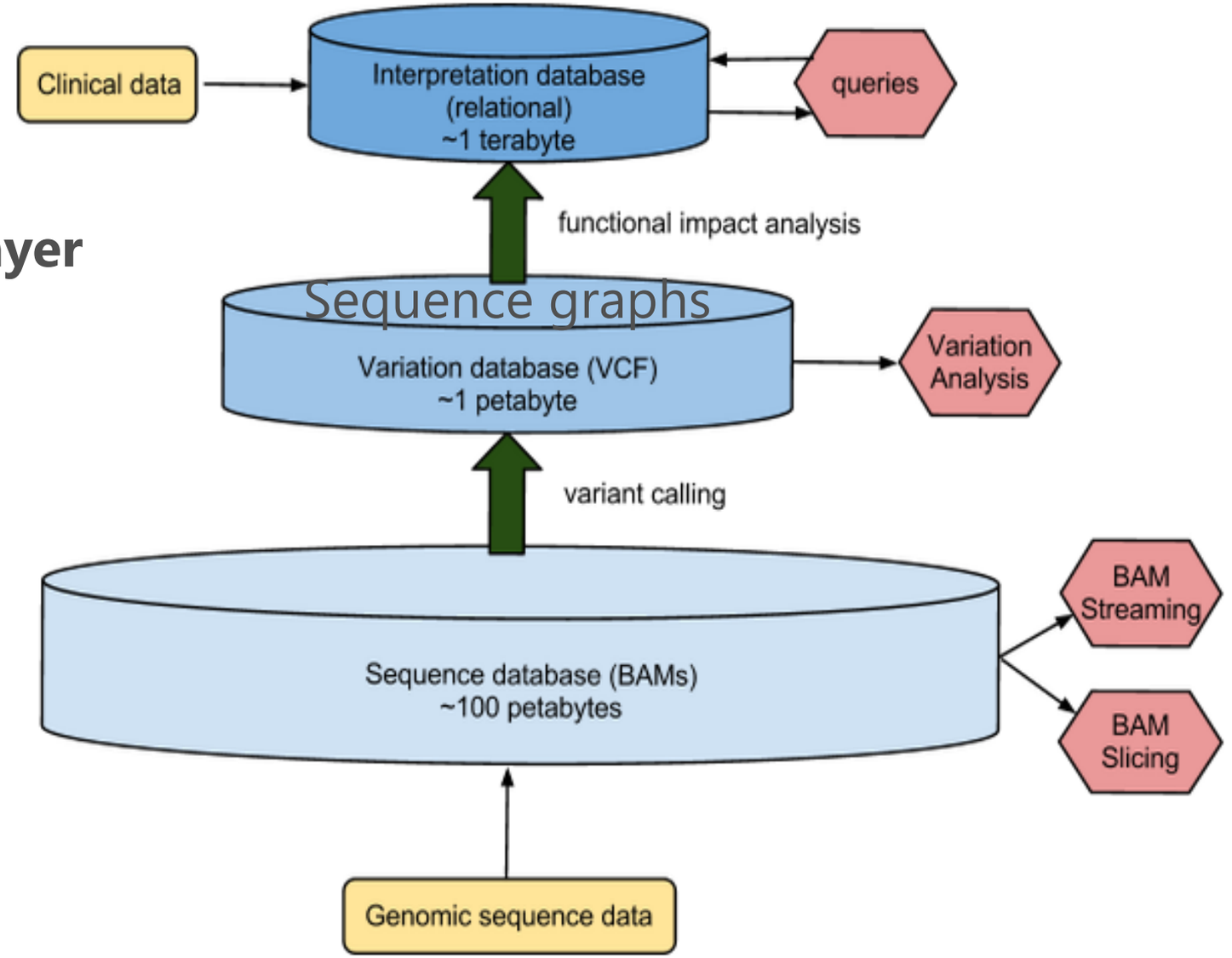
- Founding partners on June 5, 2013: 70+ leading health care, research, and disease advocacy organizations from over 40 countries
- Mission: to enable rapid progress in biomedicine
- Plan:
  - create and maintain the interoperability of technology platform standards for managing and sharing genomic data in clinical samples;
  - develop guidelines and harmonizing procedures for privacy and ethics in the international regulatory context;
  - engage stakeholders across sectors to encourage the responsible and voluntary sharing of data and of methods.

Possible Genome Commons Architecture

Interpretation Layer

Variation Layer

Read Layer



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

- Our estimate is ~ $50/genome/year in 2014 to store and analyze 1M whole genomes (~ 100 petabytes, 2 months of YouTube growth)
  - 25,000 disks and 100,000 processor cores
  - Including operating costs: space, electricity, operators
  - Including 2nd center to protect against disasters
- Note that cancer is the high water mark for global genome commons requirements, requirements for other diseases are smaller, less complex, assuming cancer includes full germline and somatic cell analysis

Dave Patterson, www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-211.html

Different Requirements for 1M Genomes

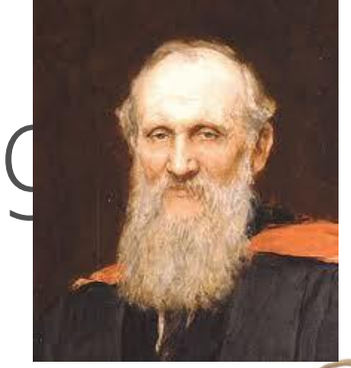
- Different types of data interactions:
  - Support both research and clinical practice
  - Compute within a provided cloud
  - Separately URIed, metadata-tagged parts of a single patient file supporting 3rd party mashups and tools
- Harmonized portable consents, sample donor has fined-grained control of who can access their data parts, trusts the security provided
- APIs, not file formats. 3rd parties must be able to build on it: goal to enable research and clinical analysis, not usurp it
- Benchmarking so all can use system to improve methods, e.g. variant calling

Dave Patterson, www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-211.html

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
3. Import benchmarking culture to accelerate progress

State of Variant Benchmarking

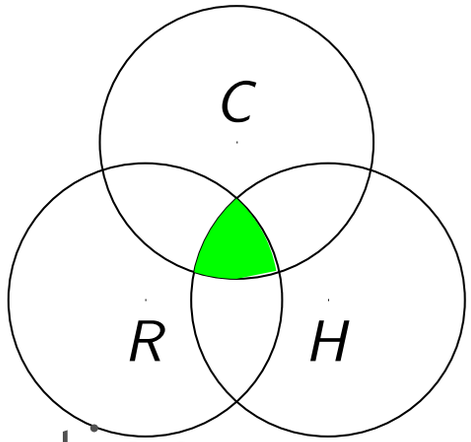


If you cannot measure it,
you cannot improve it.
- Lord Kelvin

- No fully sequenced, error-free, human genomes to measure success
  - We don't have the technology
- Limited agreement on evaluation metrics
- No agreement on common data sets
- Papers rely on (own) simulated data
- Evaluation based on consensus
  - If programs A, B, C, … all call it a variant, then must be correct

Ideal Benchmark

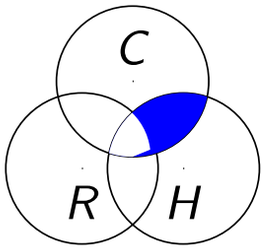
- A "benchmarking dataset" consists of
  1. Reference (for alignment of reads)
  2. Sample: Short reads input (high-coverage)
  3. Validation data (to compare predictions against) with known error bars
- Three desired properties:
  - Real (non-synthetic) reads (R)
  - Comprehensive over genome (C)
  - Human (H)
- Currently no dataset with all 3 properties



Practical Benchmark

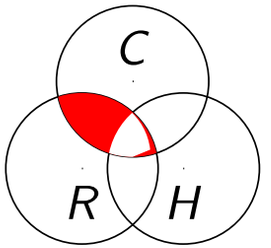
Synthetic:

1. Reference: Human Reference Genome
2. Sample: Simulated reads from simNGS [Massingham, 2012])
3. Validation: Simulated genome with variation from Venter (using TVsim, in-house simulator); no errors!



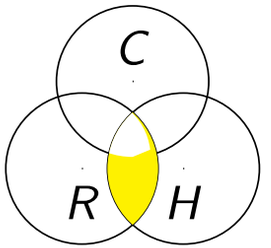
Mouse:

1. Reference: Derived from a (inbred) mouse strain
2. Sample: Real reads from actual mouse reference (another strain)
3. Validation: Mouse reference itself provides validation data



Sampled Human:

1. Reference: Human Reference
2. Sample: Short reads from Illumina and 1000Genomes
3. Validation: SNPs from HapMap, SVs from HGSVP ("Mullikin fosmids")



SMASH Results: SNPs

	Precision		Recall	
	GATK	mpileup	GATK	mpileup
Synthetic	98.3±0.0	96.8±0.0	91.3±0.00	96.9±0.00
Mouse	98.6±0.3	98.4±0.3	89.7±0.20	87.6±0.20
Sampled Human	*pending*	*pending*	98.4±0.04	80.9±0.04

	$/Genome		Hours/Genome	
	GATK	mpileup	GATK	mpileup
Synthetic	$67	$5	28	2
Mouse	$72	$17	30	7
Sampled Human	$192	$22	80	9

Time/cost on AWS EC2 (cc2.8xlarge, 16 cores, 60GB RAM)

Conclusion: Societal-Scale Big Data App

- Genetic sequencing costs 1,000,000X less
  - $1000 per genome soon?
- 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
  - Introduce benchmark culture to accelerate progress
- Accelerate Personalized Cancer Therapy from ~2050 to ~20??

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