Cancer Genomics and Computing UC Berkeley CS 294-75

David Patterson (UCB) and Taylor Sittler (UCSF)

August 28, 2011

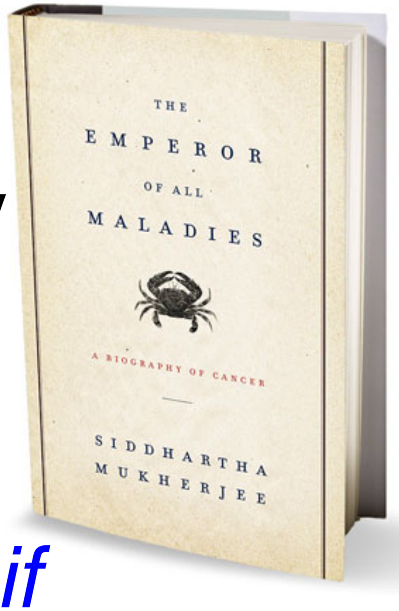
Outline

- Cancer Challenge
- Big Data Technology Opportunity/ Obligation
- Big Data / AMP Lab
- Course Overview
- Genetics 101, by Dr. Andy Poggio

Big Data, Genomics, and Cancer

- Cancer: a perversion of a normal cell
  - Limitless growth, evolves, then spreads: immortal disease
- Cancer is a genetic disease
- Accidental DNA cell copy flaws + carcinogen-caused mutations lead to cancer
- Turns on growth accelerator (*oncogenes*) or turns off tumor brakes (*anti-oncogenes*)

Big Data, Genomics, and Cancer



- "Indeed, as the fraction of those affected by cancer creeps inexorably in some nations form one in four to one in three to on in *two*, cancer will, indeed, become the new normal–an inevitability. The question will not be *if* we will encounter this immortal disease in our lives, but *when*."
- ¼ US deaths, 7M/year worldwide
- ⅓ US women will get cancer
- ½ US men will get cancer

Five Steps from Genes to Diagnosis

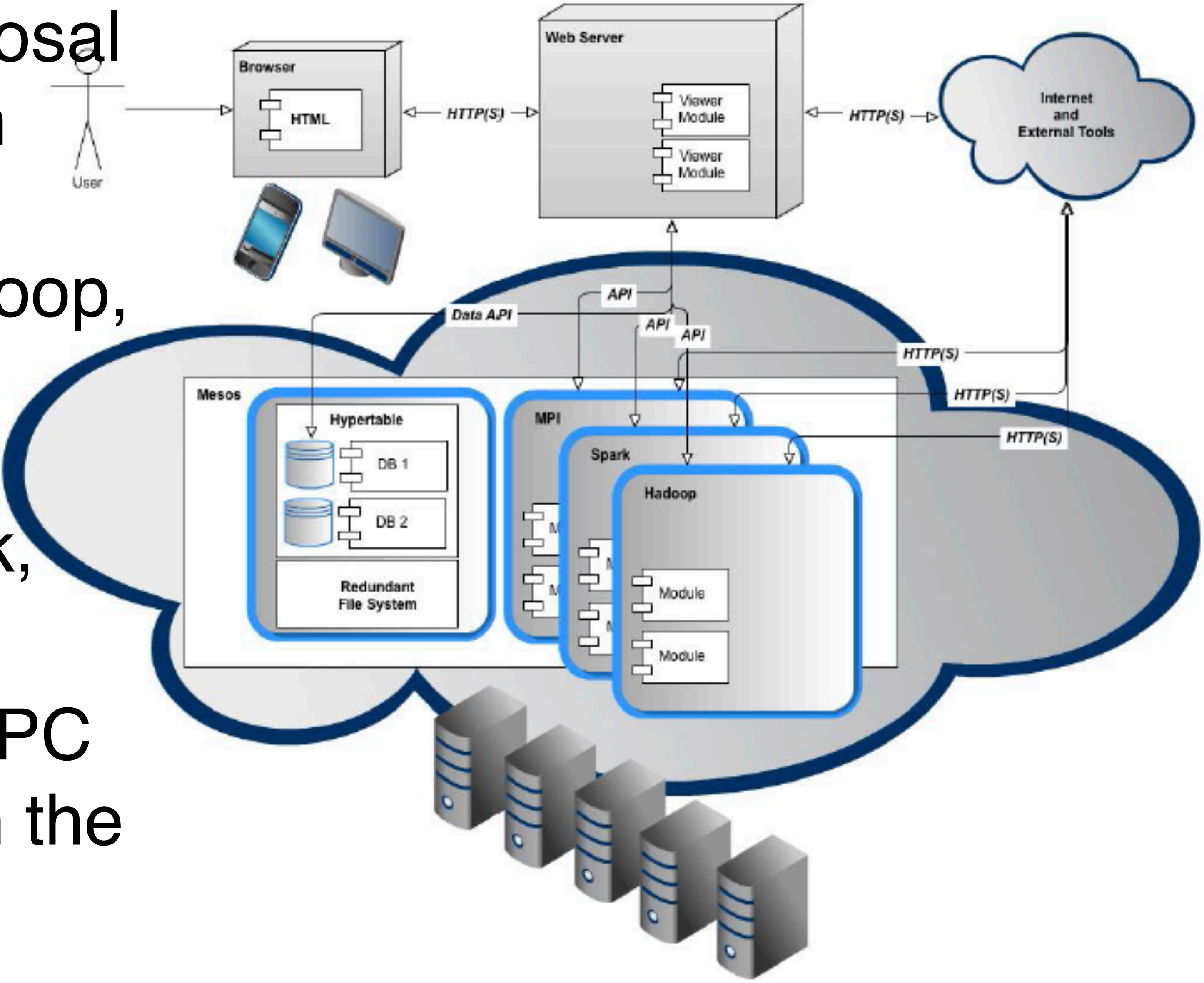
1. Sequencing machine that identifies many 300 base pair segments from a cancer tumor
2. Create full genome sequence of the cancer tumor from many segments
3. Verify correctness of this sequence
4. Insights from comparing many sequences to cancer tumor genome
5. Diagnose and suggest of therapeutic targets for cure or non-progression based on tumor genome comparisons and patient records

1. Sequencing Machine Costs

- Improving faster than Moore's Law
- 2007: $1,000,000
- 2009: $10,000
- 2012-3: $1,000
- 2007 wet lab processing problem => 2012 digital processing problem
- Looks like sequencing machines not the bottleneck in speed *or* cost

2 & 3. Full sequence and verification

- UCSF proposal to use open source SW, Cloud, Hadoop, Hypertable, Berkeley tools (Spark, Mesos)
- >1 year on PC to <1 day in the cloud



4. Compare Sequences

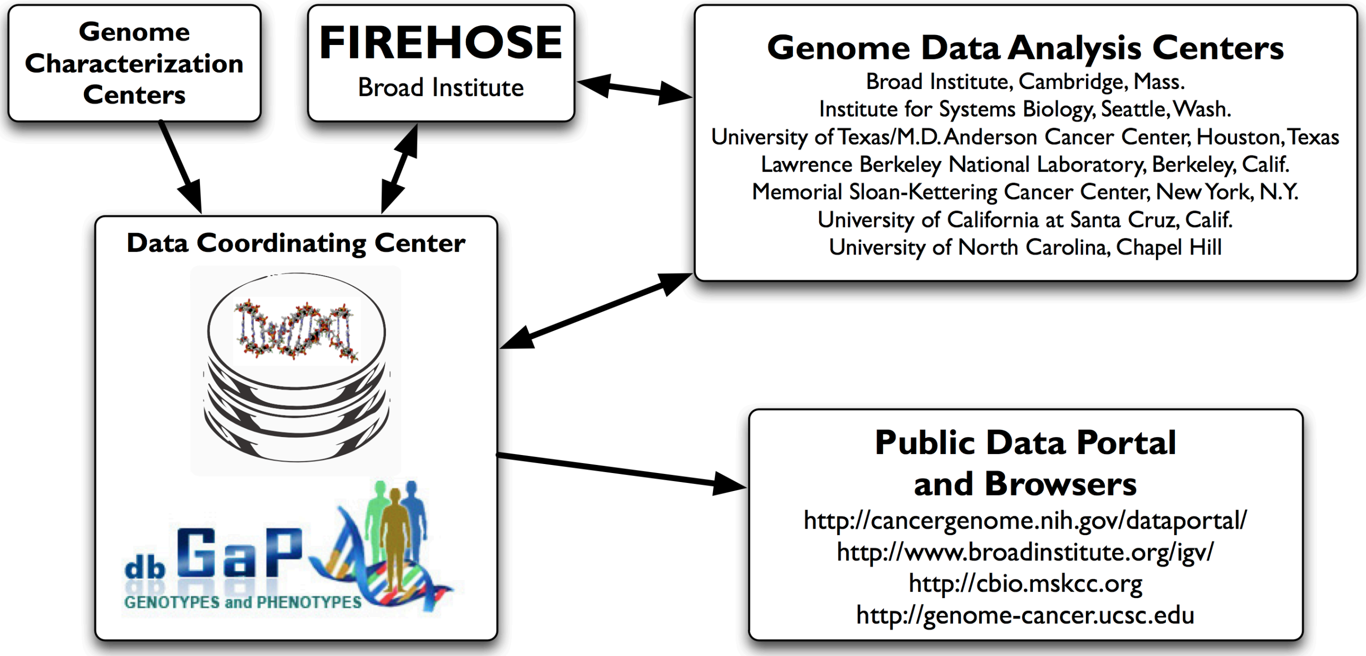
- Sequence Alignment
- Machine Learning + Data Analytics
- Cloud Programming Frameworks

5. Clinical Diagnosis

- Suggest effective therapeutic targets for cure or stabilization
  - Often events are relatively rare mutations, and not identifiable by traditional statistical methods (solutions lie in long tail of rare mutations)
- Use patient records
  + natural language processing?
- Crowd Sourcing?
  - Can turn it into a "foldit" like game?

Big Data Opportunity

- The Cancer Genome Atlas (TCGA)
  - 20 cancer types, 500 tumors each: 5 petabytes
  - Datacenter of David Haussler opens 10/1/11
  - Place Berkeley cluster next to 5 PB of data
  - Novelty: Academic Access yet Important Big Data



Genome Characterization Centers

FIREHOSE
Broad Institute

Genome Data Analysis Centers
Broad Institute, Cambridge, Mass.
Institute for Systems Biology, Seattle, Wash.
University of Texas/M.D. Anderson Cancer Center, Houston, Texas
Lawrence Berkeley National Laboratory, Berkeley, Calif.
Memorial Sloan-Kettering Cancer Center, New York, N.Y.
University of California at Santa Cruz, Calif.
University of North Carolina, Chapel Hill

Data Coordinating Center

Public Data Portal and Browsers
http://cancergenome.nih.gov/dataportal/
http://www.broadinstitute.org/igv/
http://cbio.mskcc.org
http://genome-cancer.ucsc.edu

Slide from David Haussler, UCSC, "Cancer Genomics," AMP retreat, 5/24/11

TCGA Potential Impact?

“We fully expect that 10 years from now, each cancer patient is going to want to get a genomic analysis of their cancer and will expect customized therapy based on that information.”

Brad Ozenberger
TCGA program director
“Cracking Cancer's Code”
Time Magazine
June 2, 2011

Information Technology Obstacle?

“There is a growing gap between the generation of massively parallel sequencing output and the ability to process and analyze the resulting data. New users are left to navigate a bewildering maze of base calling, alignment, assembly and analysis tools with often incomplete documentation and no idea how to compare and validate their outputs. Bridging this gap is essential, or the coveted $1,000 genome will come with a $20,000 analysis price tag.”

John D McPherson
“Next-generation gap,”
Nature Methods,
October 15, 2009

An Opportunity or Obligation?

- Given increasing genomic databases, next breakthroughs in cancer fight more likely to come from computer scientists than from biological scientists
- If it is plausible that we could help millions of cancer patients live longer and better lives, as moral people, don't we have an obligation to try?

Big Data and Pasteur’s Quadrant

Research is inspired by:

		Consideration of use? No	 Yes
Quest for Fundamental Understanding?	Yes	Pure Basic Research (Bohr)	**Use-inspired Basic Research (Pasteur)** **Attack Big Data by Helping Fight Cancer?**
	No		Pure Applied Research (Edison)

From *Pasteur’s Quadrant: Basic Science and Technological Innovation*, Donald E. Stokes, 1997
Slide from “Engineering Education and the Challenges of the 21st Century,” Charles Vest, 9/22/09

Outline

- Cancer Challenge
- Big Data Technology Opportunity/ Obligation
- Big Data / AMP Lab
- Course Overview
- Genetics 101, by Dr. Andy Poggio

Big Data is ..

- Massive
  - Facebook: 200-400TB/day: 83 million pictures
  - Google: > 25 PB/day processed
- Growing
  - More devices (cell phones), More people (3rd world), Bigger disks (2TB/$100)
- Dirty
  - Diverse, No Schema, Uncurated, Inconsistent Syntax and Semantics

Many biologists don't trust any data that they didn't produce themselves. An important distinction is that biologists typically do trust tools they didn't create themselves, despite a similar potential for inaccuracy. They tend to acknowledge data bugs but not software bugs.

Current Biologist Tools

- These algorithms implemented as tools are critical
- Tools come from diverse individuals or small, ad hoc groups in the computational biology community
  - Not writing in modern ways or modern languages; Perl is popular
- Obvious improvements needed in interoperability and parallelism

"The current state of genomics data is turning biologists into Perl programmers."

Gene Myers,
Celera Human Genome Project

“Big Data”: Working Definition

When the normal application of current technology doesn’t enable users to obtain **timely** and **cost-effective** answers of sufficient **quality** to data-driven questions

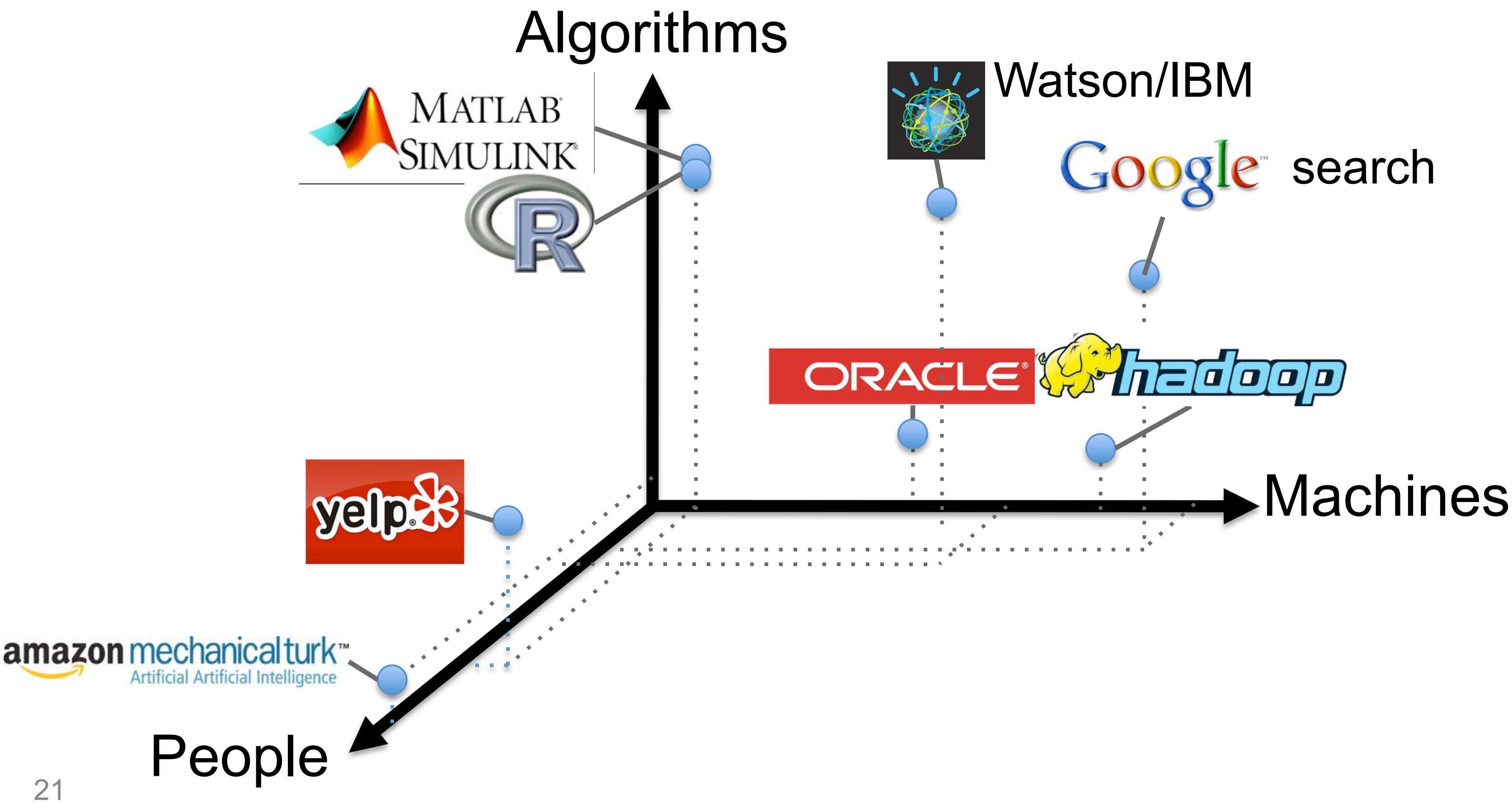
Challenge: Use algorithms and people to extract value from Big Data while decreasing the cost of maintaining it

3 Dimensions to Improve Data Analysis

1. Improve scale, efficiency, and quality of algorithms to increase value from Big Data (**Algorithms**)
2. Use cloud computing to get value from Big Data and enhance datacenter infrastructure to cut costs of Big Data management (**Machines**)
3. Leverage human activity and intelligence to extract value from Big Data cases that are hard for algorithms
(**People**)

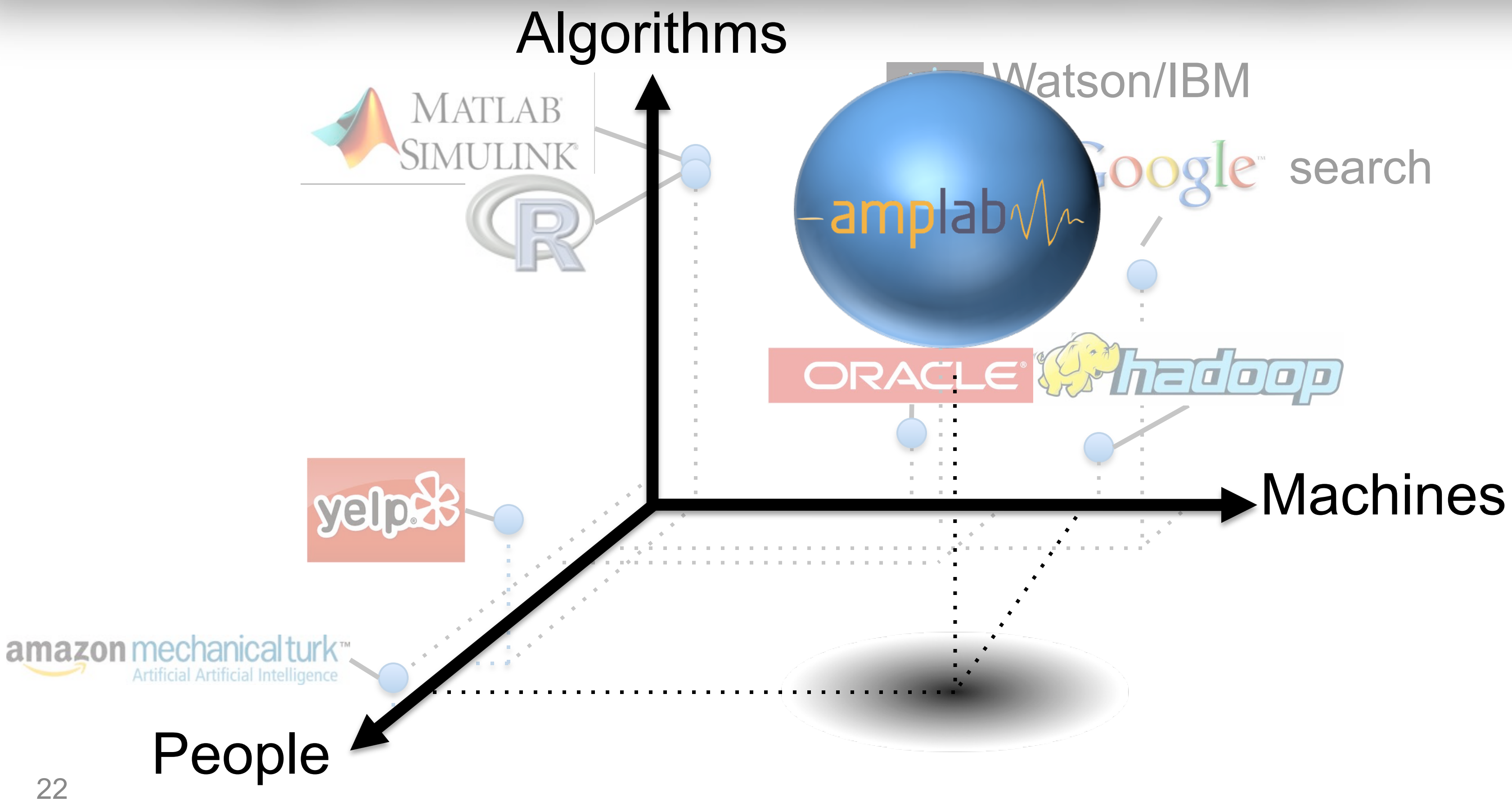
Algorithms, Machines, People

- Today's solutions:

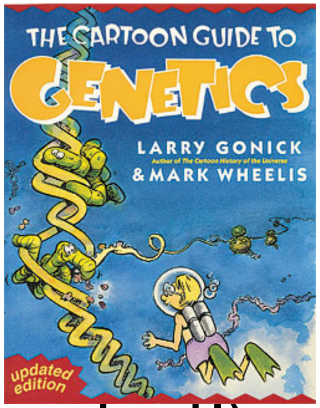


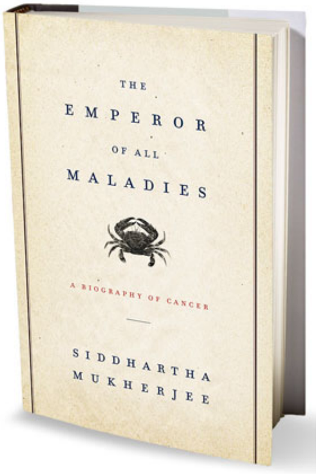
The AMPLab

Making sense at scale by integrating Algorithms, Machines, and People



CS294-75 Course Overview





- Mondays 5:30-7PM, 2 units
- Beginning–read background material, learn jargon
  - – (see who does best on 9/12 jargon test!)
- Middle– Distinguished speakers and read their papers, select project, project choice and project progress
- End– Project Presentations + Reflect on what we learned + Identify low hanging fruit
- See http://tinyurl.com/CS294Cancer

Outline

- Cancer Challenge
- Big Data Technology Opportunity/ Obligation
- Big Data / AMP Lab
- Course Overview
- Genetics 101, by Dr. Andy Poggio

Backup Slides

AMP Faculty and Sponsors

- Started March 2011
- Faculty
  - Alex Bayen (mobile sensing platforms)
  - Armando Fox (systems)
  - Michael Franklin (databases) Director
  - Michael Jordan (machine learning)
  - Anthony Joseph (security & privacy)
  - Randy Katz (systems)
  - David Patterson (systems)
  - Ion Stoica (systems)
  - Scott Shenker (networking)

Google
SAP
amazon.com web services
ERICSSON
HUAWEI
IBM
intel
Microsoft
NetApp
NEC
Quanta Computer
splunk>
Sun
ORACLE
vmware
cloudera