amplab

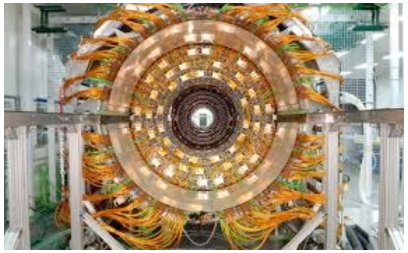
A Berkeley View of Big Data

Ion Stoica
UC Berkeley

BEARS
February 17, 2011

Big Data is Massive…

- Facebook:
  - 130TB/day: user logs
  - 200-400TB/day: 83 million pictures
- Google: > 25 PB/day processed data



- Data generated by LHC: 1 PB/sec

- Total data created in 2010:
  1 ZettaByte (1,000,000 PB)/year
  - ~60% increase every year

…and Grows Bigger and Bigger!

- More and more devices

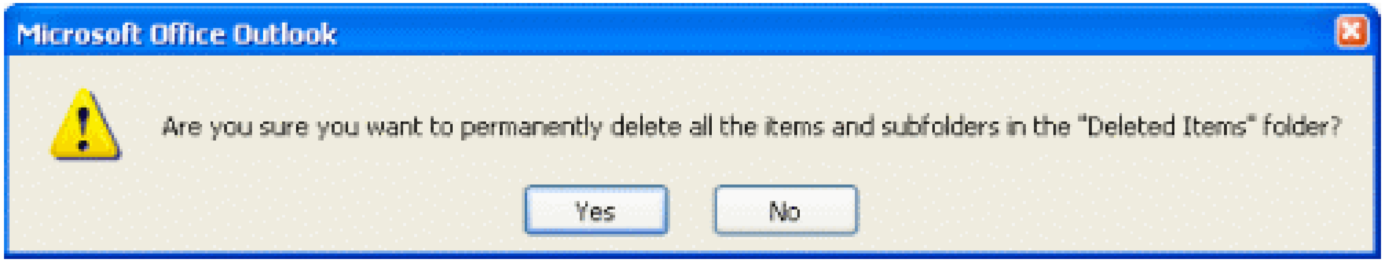
- More and more people

- Cheaper and cheaper storage
  - ~50% increase in GB/$ every year

…and Grows Bigger and Bigger!

- Log everything!
  - Don’t always know what question you’ll need to answer

- Hard to decide what to delete



Microsoft Office Outlook

Are you sure you want to permanently delete all the items and subfolders in the "Deleted Items" folder?

  - Thankless decision: people know only when you are wrong!
  - “Climate Research Unit (CRU) scientists admit they threw away key data used in global warming calculations”

- Stored data grows faster than GB/$

amplab

What is Big Data?

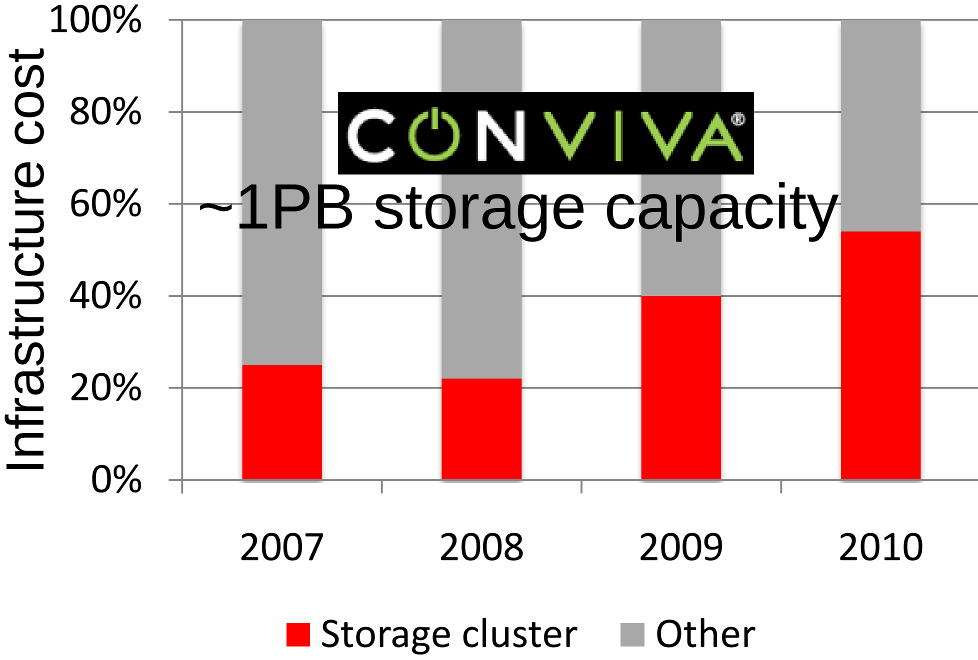
Data that is expensive to manage, and hard to extract value from

- You don’t need to be big to have big data problem!
  - Inadequate tools to analyze data
  - Data management may dominate infrastructure cost

Big Data is not Cheap!

- Storing and managing 1PB data: $500K-$1M/ year
  - Facebook: 200 PB/year

- “Typical” cloud-based service startup (e.g., Conviva)
  - Log storage dominates infrastructure cost



~1PB storage capacity

Hard to Extract Value from Data!

- Data is
  - Diverse, variety of sources
  - Uncurated, no schema, inconsistent semantics, syntax
  - Integration a huge challenge

- No easy way to get answers that are
  - High-quality
  - Timely

- Challenge: maximize value from data by getting best possible answers

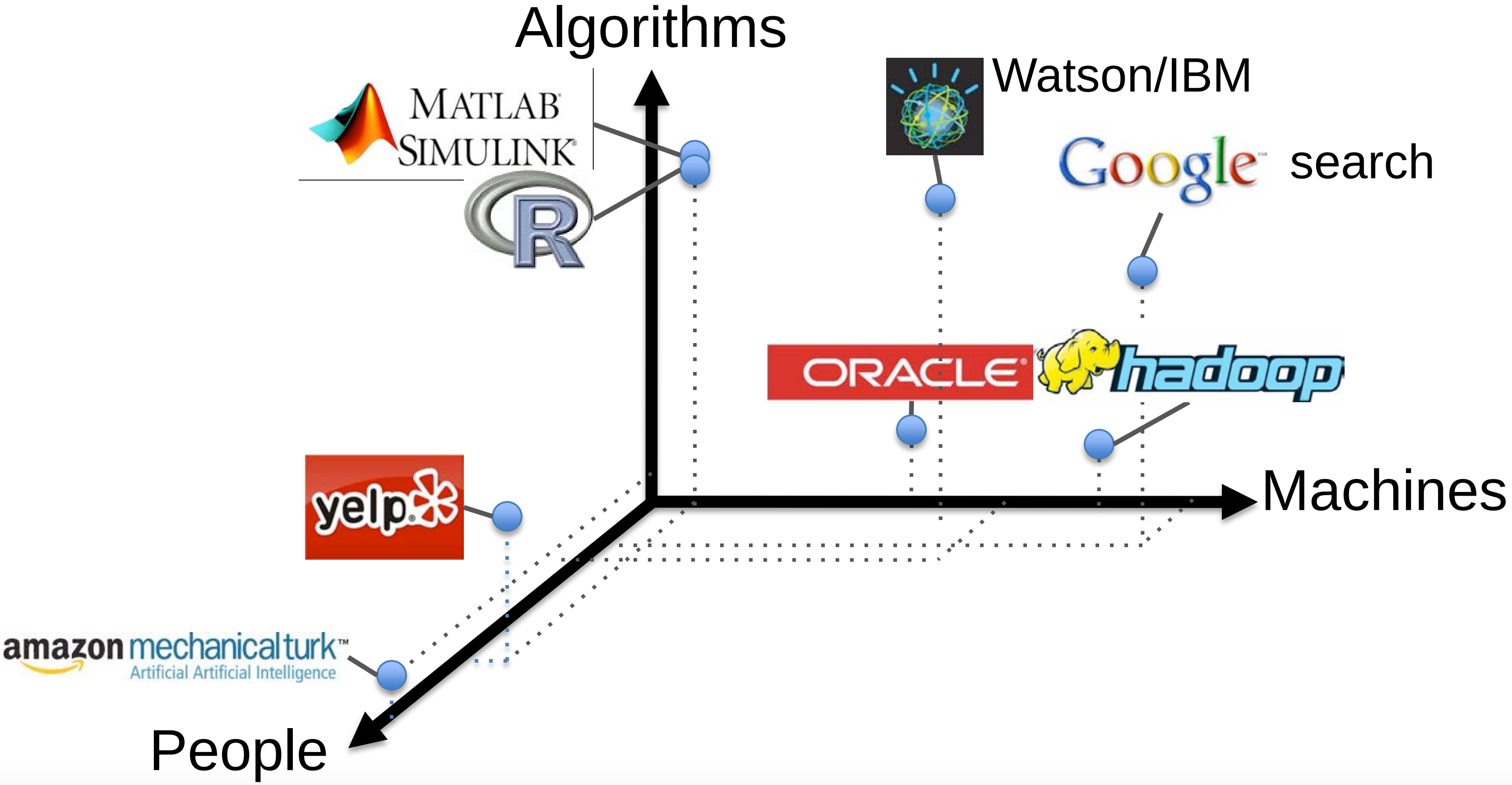
Requires Multifaceted Approach

- Three dimensions to improve data analysis
  - Improving scale, efficiency, and quality of algorithms (**Algorithms)**
  - Scaling up datacenters (**Machines)**
  - Leverage human activity and intelligence **(People)**

- Need to adaptively and flexibly combine all three dimensions

Algorithms, Machines, People

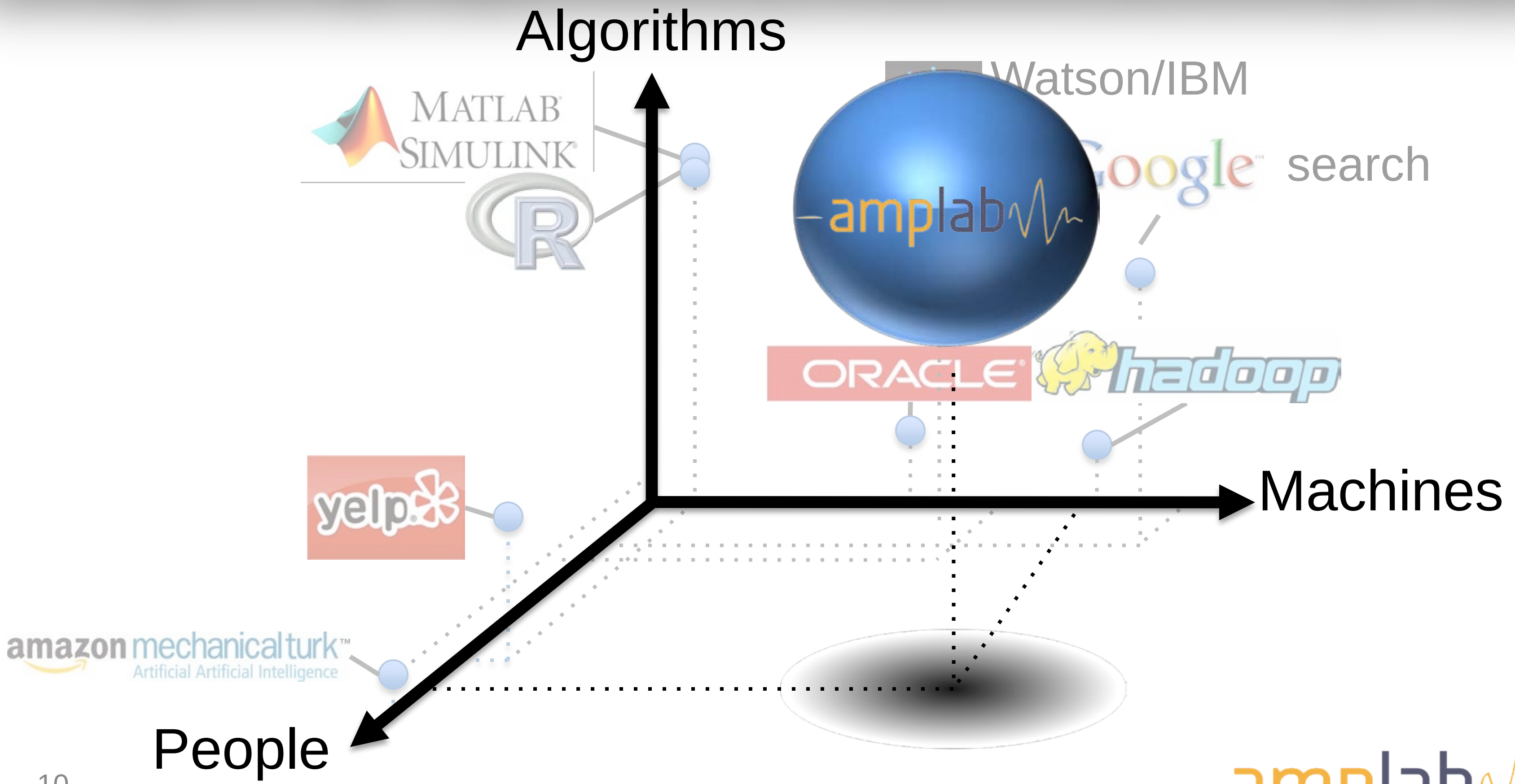
- Today's apps: fixed point in solution space



Need techniques to dynamically pick best operating point

The AMP Lab

Make sense of data at scale by tightly integrating algorithms, machines, and people



AMP Faculty and Sponsors

- Faculty
  - Alex Bayen (mobile sensing platforms)
  - Armando Fox (systems)
  - Michael Franklin (databases): Director
  - Michael Jordan (machine learning): Co-director
  - Anthony Joseph (security & privacy)
  - Randy Katz (systems)
  - David Patterson (systems)
  - Ion Stoica (systems): Co-director
  - Scott Shenker (networking)
- Sponsors: Google™ SAP®

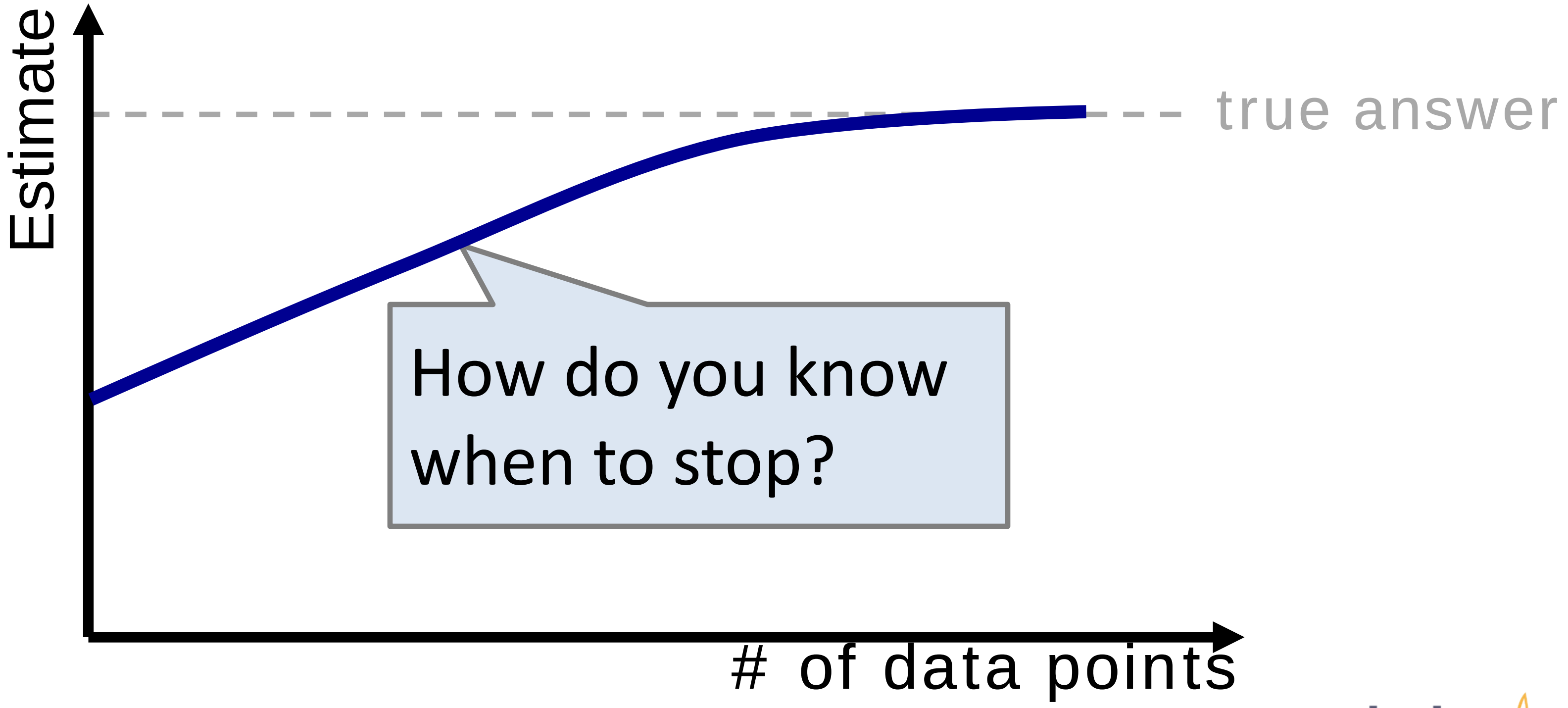
amazon.com web services · ebay™ · HUAWEI · IBM® · intel® · Microsoft®

NetApp · NEC · vmware® · cloudera

amplab

Algorithms

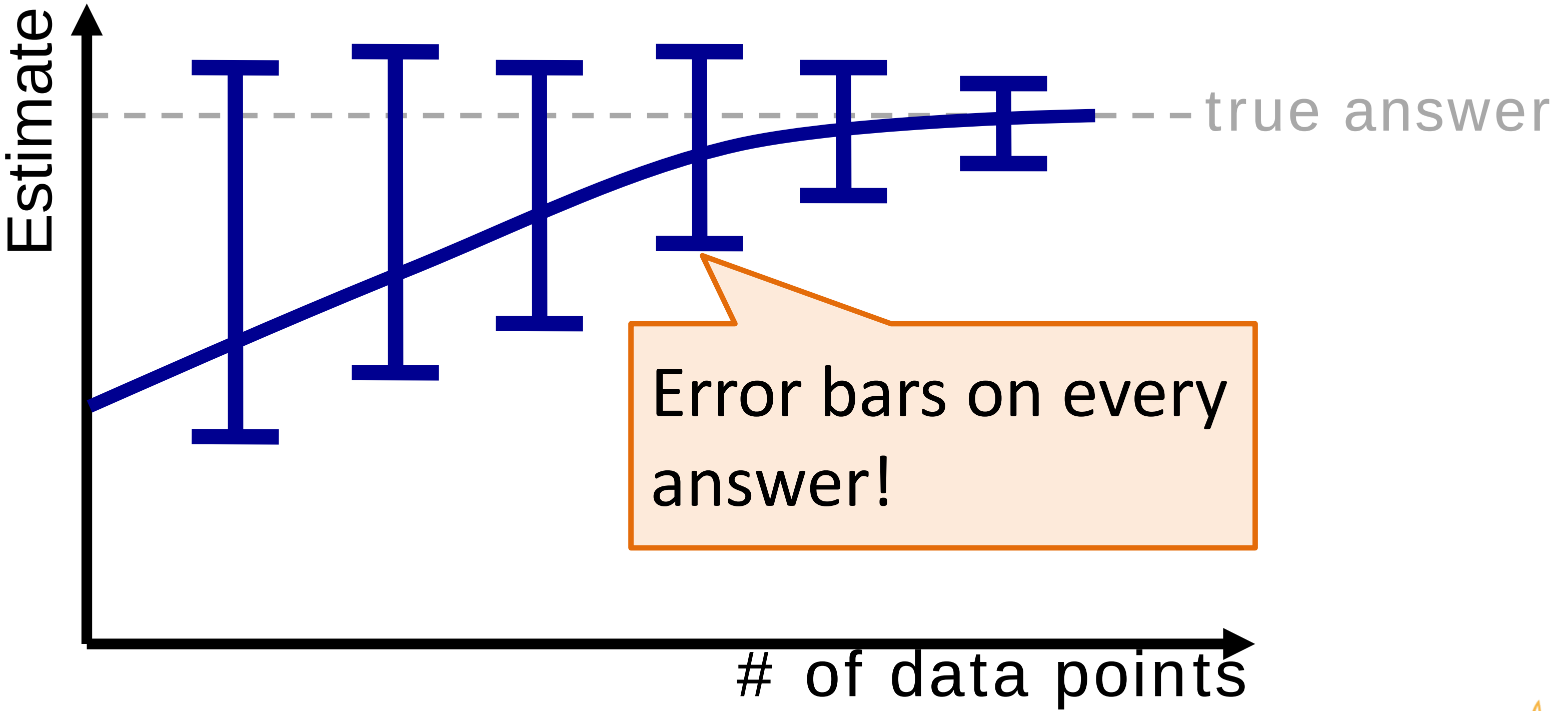
- State-of-art Machine Learning (ML) algorithms do not scale
  - – Prohibitive to process all data points



How do you know when to stop?

Algorithms

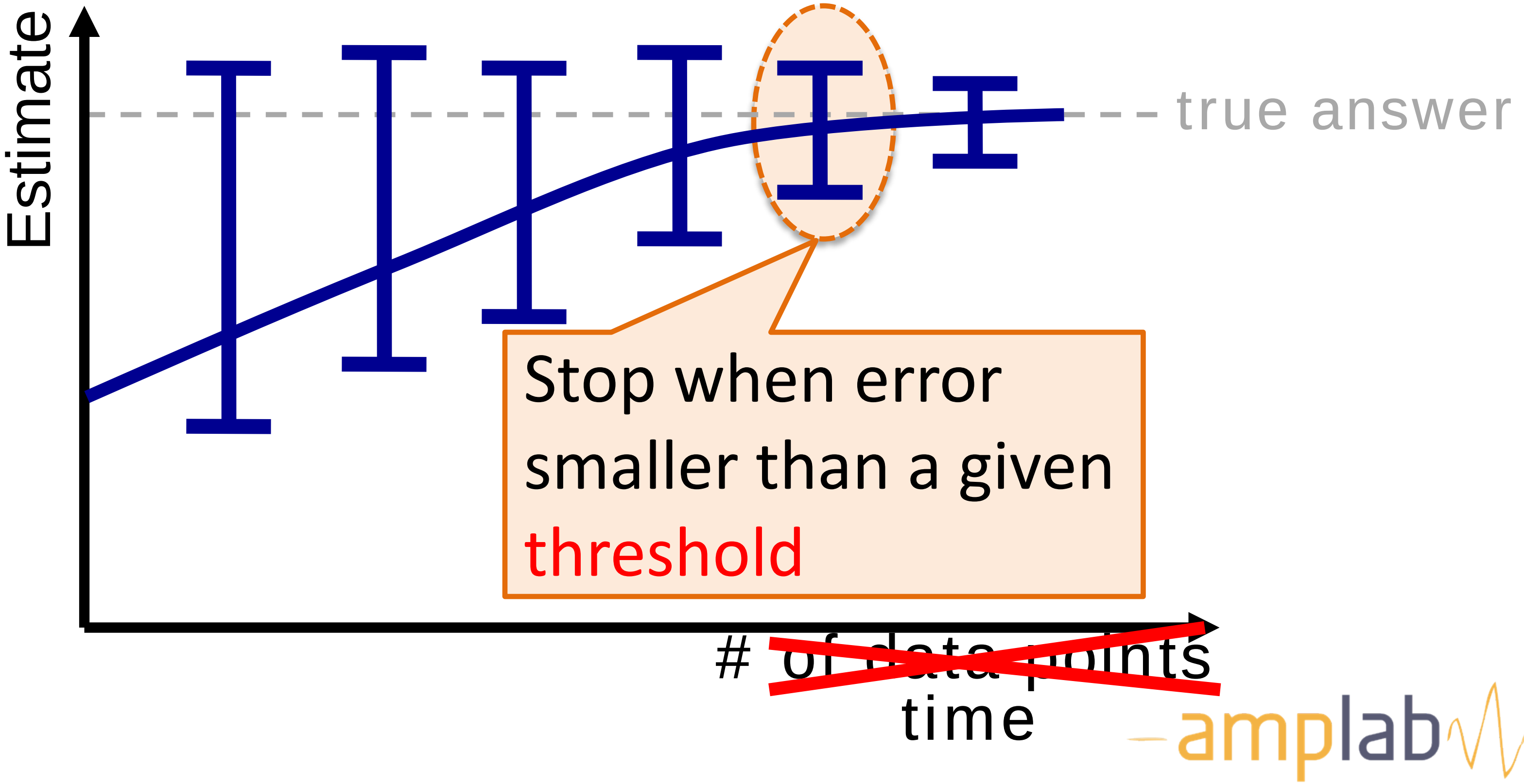
- Given any problem, data and a budget
  - – Immediate results with continuous improvement
  - – Calibrate answer: provide error bars



Error bars on every answer!

Algorithms

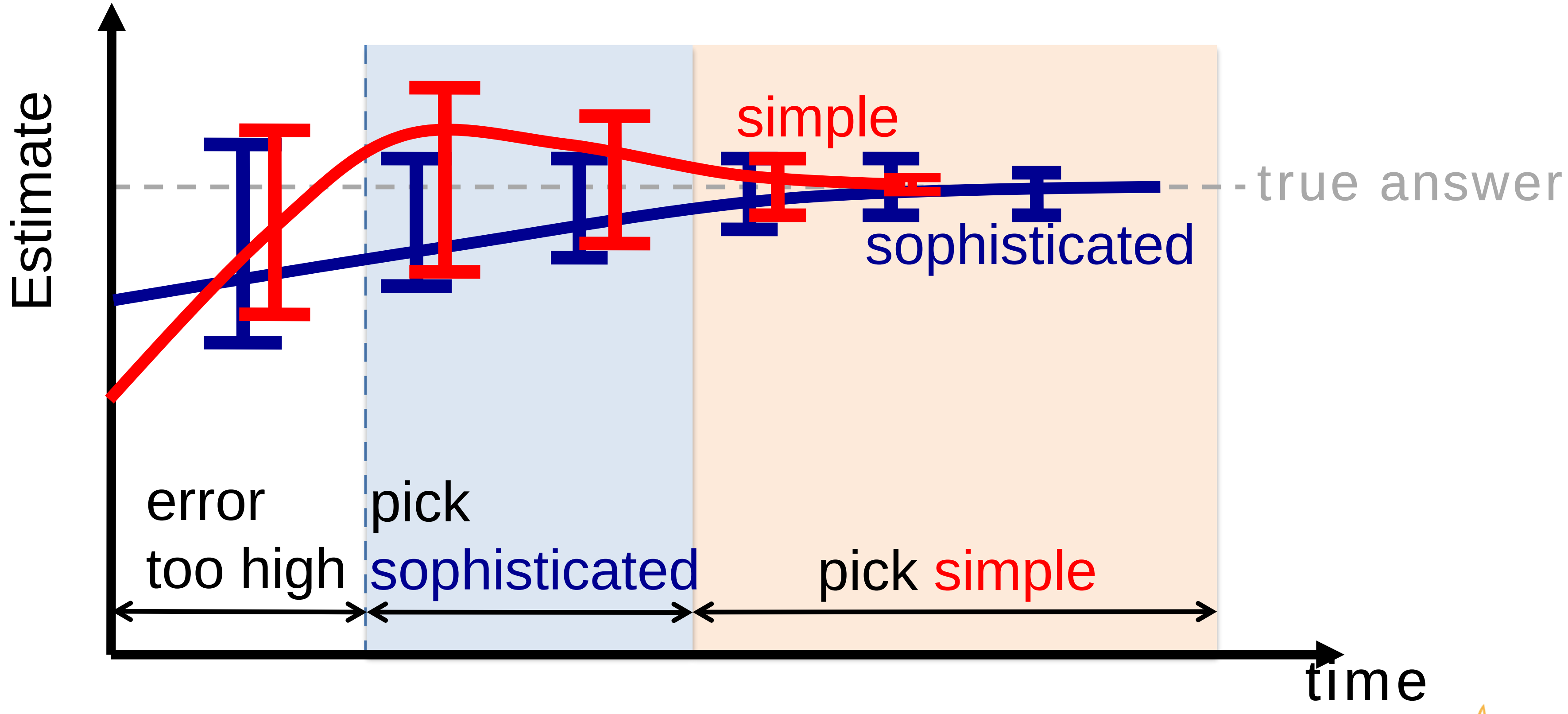
- Given any problem, data and a time budget
  - Immediate results with continuous improvement
  - Calibrate answer: provide error bars



Stop when error smaller than a given threshold

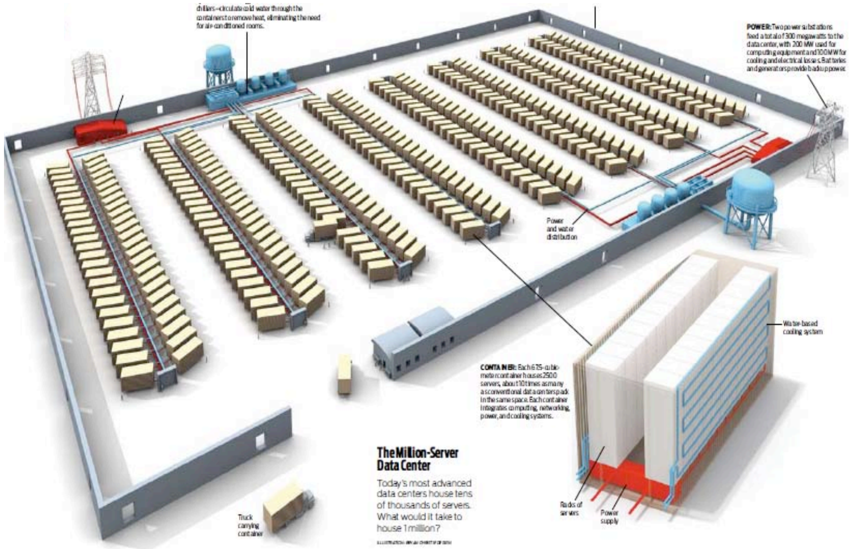
Algorithms

- Given any problem, data and a time budget
  - – Automatically pick the best algorithm



Machines

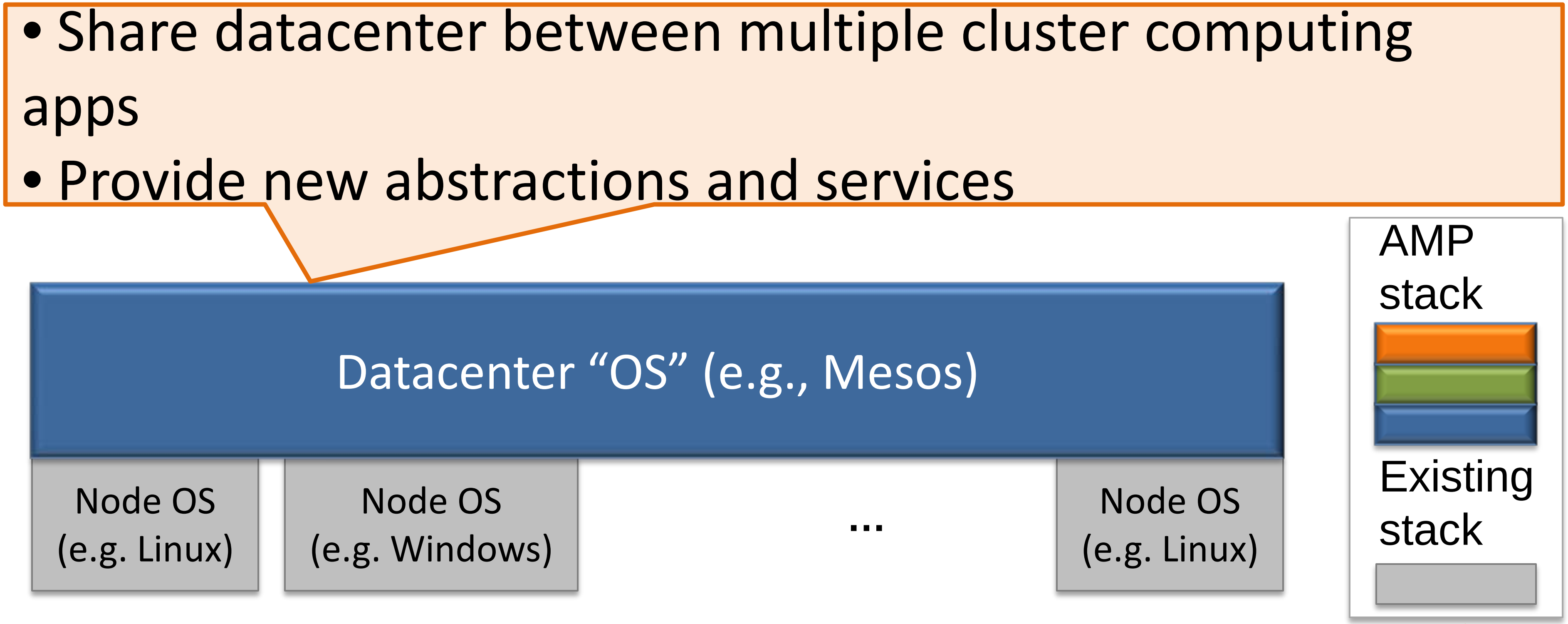
- “The datacenter as a computer” still in its infancy
  - Special purpose clusters, e.g., Hadoop cluster
  - Highly variable performance
  - Hard to program
  - Hard to debug


chillers—circulate cold water through the containers to remove heat, eliminating the need for air-conditioned rooms.
POWER: Two power substations feed a total of 300 megawatts to the data center, with 200 MW used for computing equipment and 100 MW for cooling and electrical losses. Batteries and generators provide backup power.
Power and water distribution
Water-based cooling system
CONTAINER: Each 675-cubic-meter container houses 2500 servers, about 10 times as many as conventional data centers pack in the same space. Each container integrates computing, networking, power, and cooling systems.
The Million-Server Data Center
Today's most advanced data centers house tens of thousands of servers. What would it take to house 1 million?
Truck carrying container
Racks of servers
Power supply

=?

Machines

- Make datacenter a real computer!



- Share datacenter between multiple cluster computing apps
- Provide new abstractions and services

Datacenter "OS" (e.g., Mesos)

Node OS (e.g. Linux)

Node OS (e.g. Windows)

…

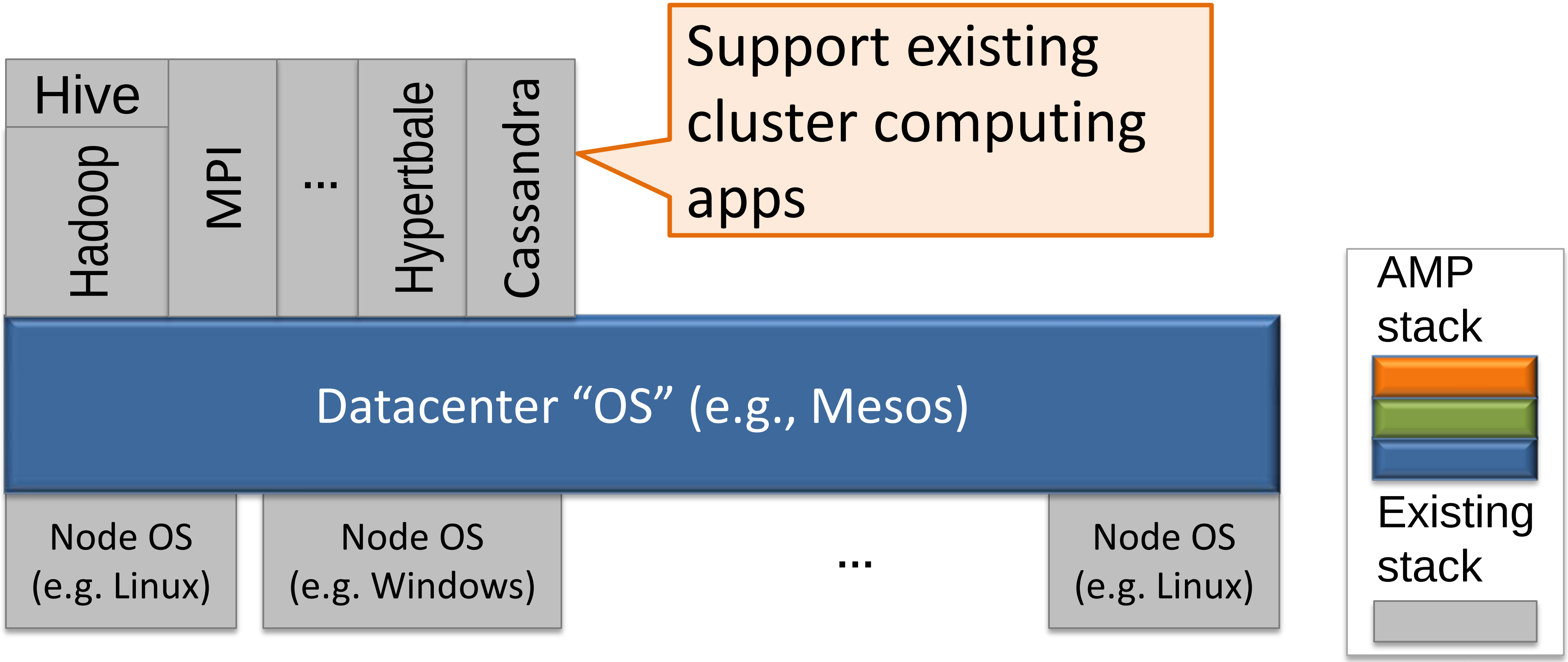
Node OS (e.g. Linux)

AMP stack

Existing stack

Machines

- Make datacenter a real computer!



Support existing cluster computing apps

Hive

Hadoop

MPI

…

Hypertbale

Cassandra

Datacenter “OS” (e.g., Mesos)

Node OS (e.g. Linux)

Node OS (e.g. Windows)

…

Node OS (e.g. Linux)

AMP stack

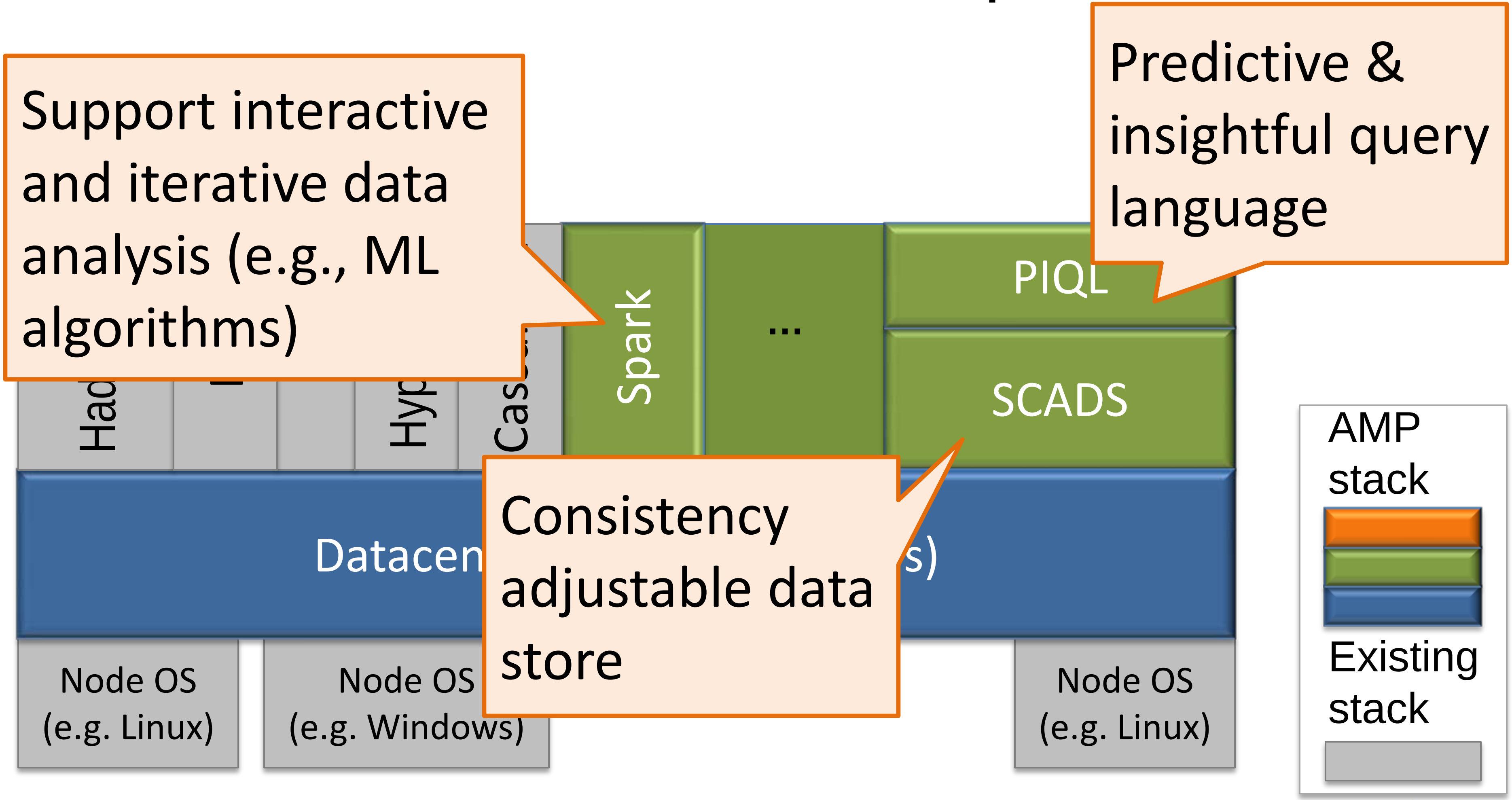
🟧 🟩 🟦

Existing stack

⬜

Machines

- Make datacenter a real computer!



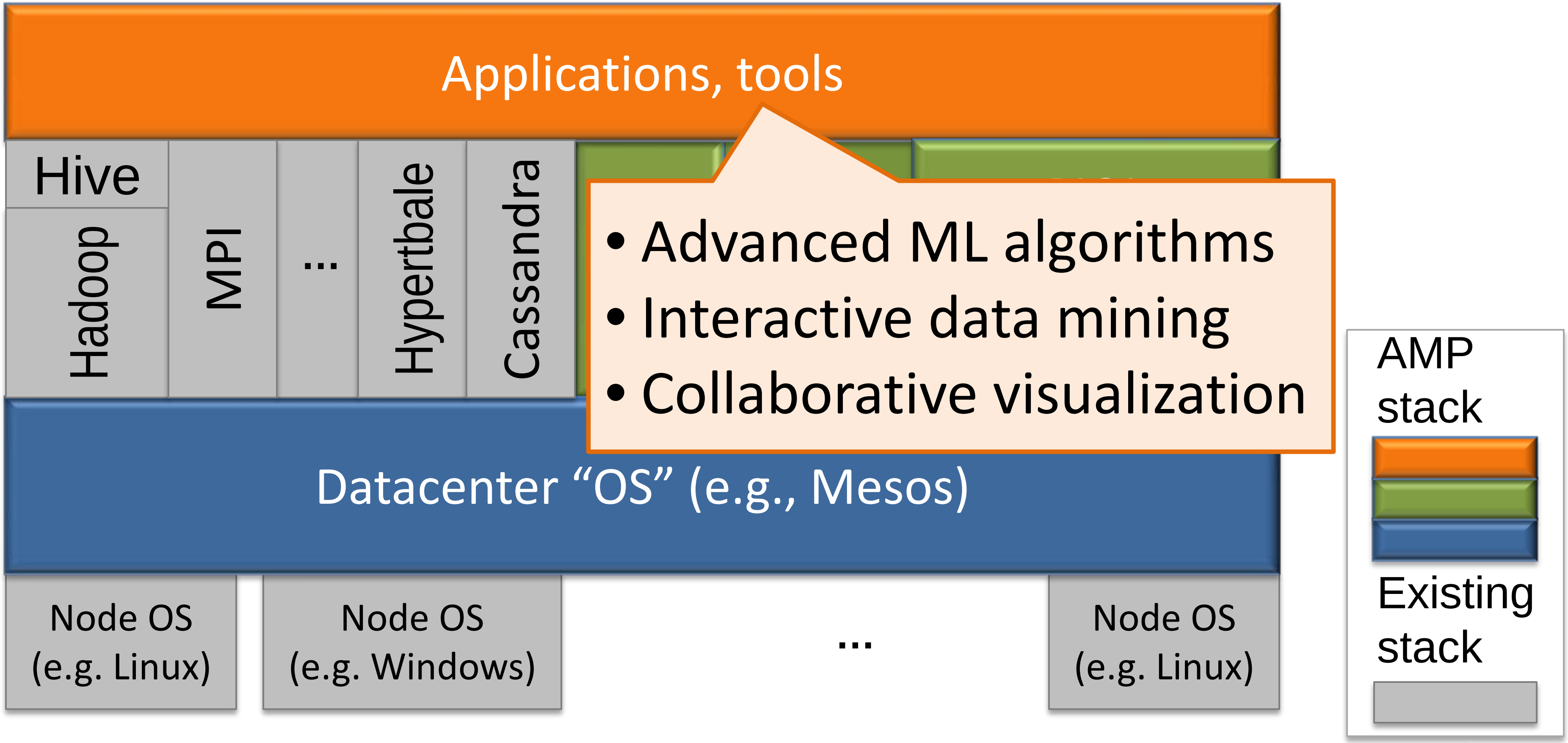
Support interactive and iterative data analysis (e.g., ML algorithms)

Predictive & insightful query language

Consistency adjustable data store

Machines

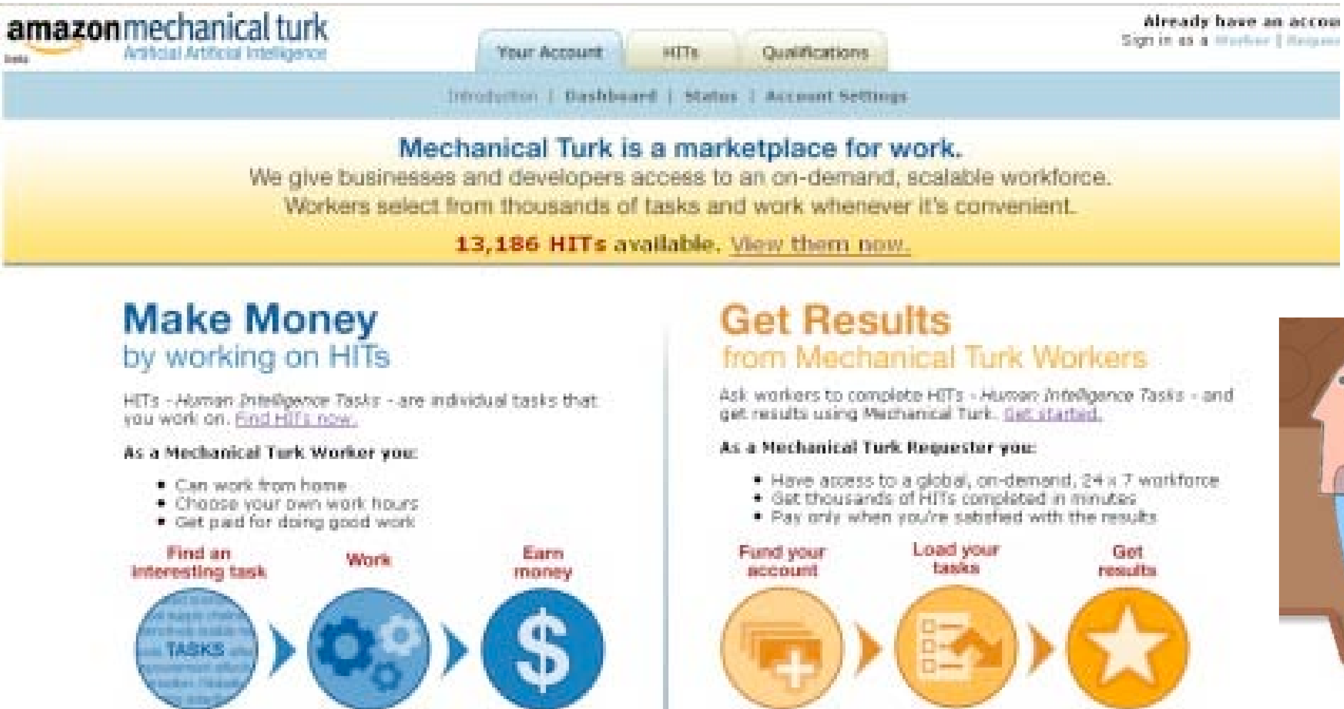
- Make datacenter a real computer!



- Advanced ML algorithms
- Interactive data mining
- Collaborative visualization

People

- Humans can make sense of messy data!



amazon mechanical turk
Artificial Artificial Intelligence

Your Account | HITs | Qualifications

Already have an accou
Sign in as a Worker | Reques

Introduction | Dashboard | Status | Account Settings

Mechanical Turk is a marketplace for work.
We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.

13,186 HITs available. View them now.

Make Money
by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work

Find an interesting task → Work → Earn money

Find HITs Now

or learn more about being a Worker

Get Results
from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. Get started.

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

Fund your account → Load your tasks → Get results

Get Started

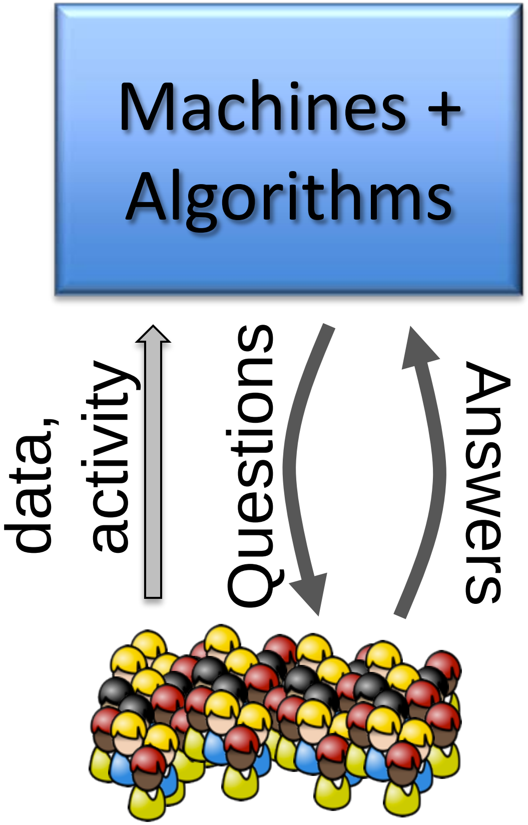
or learn more about being a Requester

FAQ | Contact Us | Careers at Amazon | Developers | Press | Policies
©2005-2008 Amazon.com, Inc. or its Affiliates

Quora

People

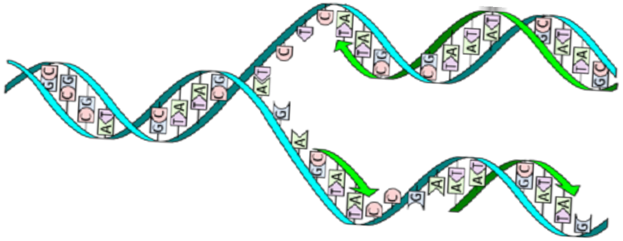
- Make people an integrated part of the system!
  - Leverage human activity
  - Leverage human intelligence (crowdsourcing):
    - Curate and clean dirty data
    - Answer imprecise questions
    - Test and improve algorithms



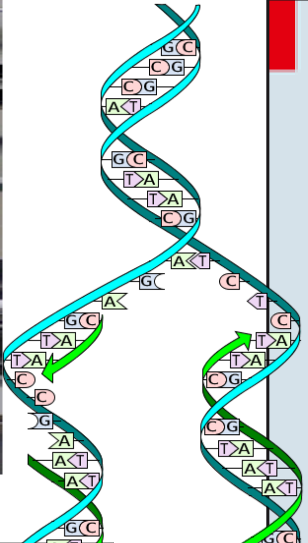
- Challenge
  - Inconsistent answer quality in all dimensions (e.g., type of question, time, cost)

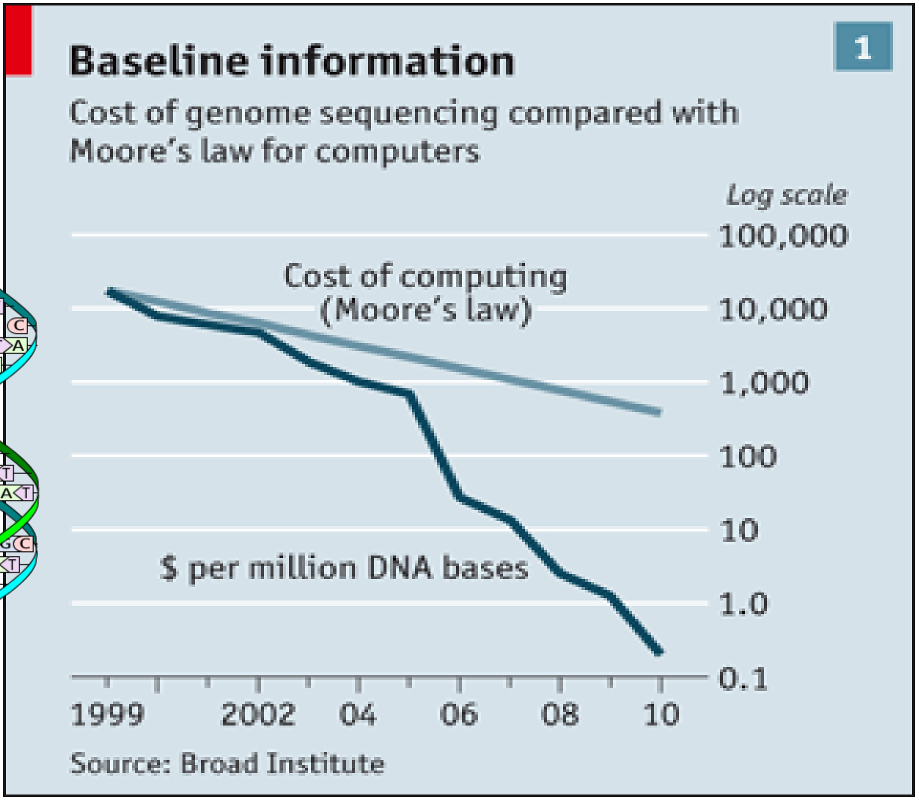
Real Applications

- Mobile Millennium Project
  - *Alex Bayen*, Civil and Environment Engineering, UC Berkeley
- Microsimulation of urban development
  - *Paul Waddell*, College of Environment Design, UC Berkeley
- Crowd based opinion formation
  - *Ken Goldberg*, Industrial Engineering and Operations Research, UC Berkeley
- Personalized Sequencing
  - *Taylor Sittler*, UCSF



Personalized Sequencing





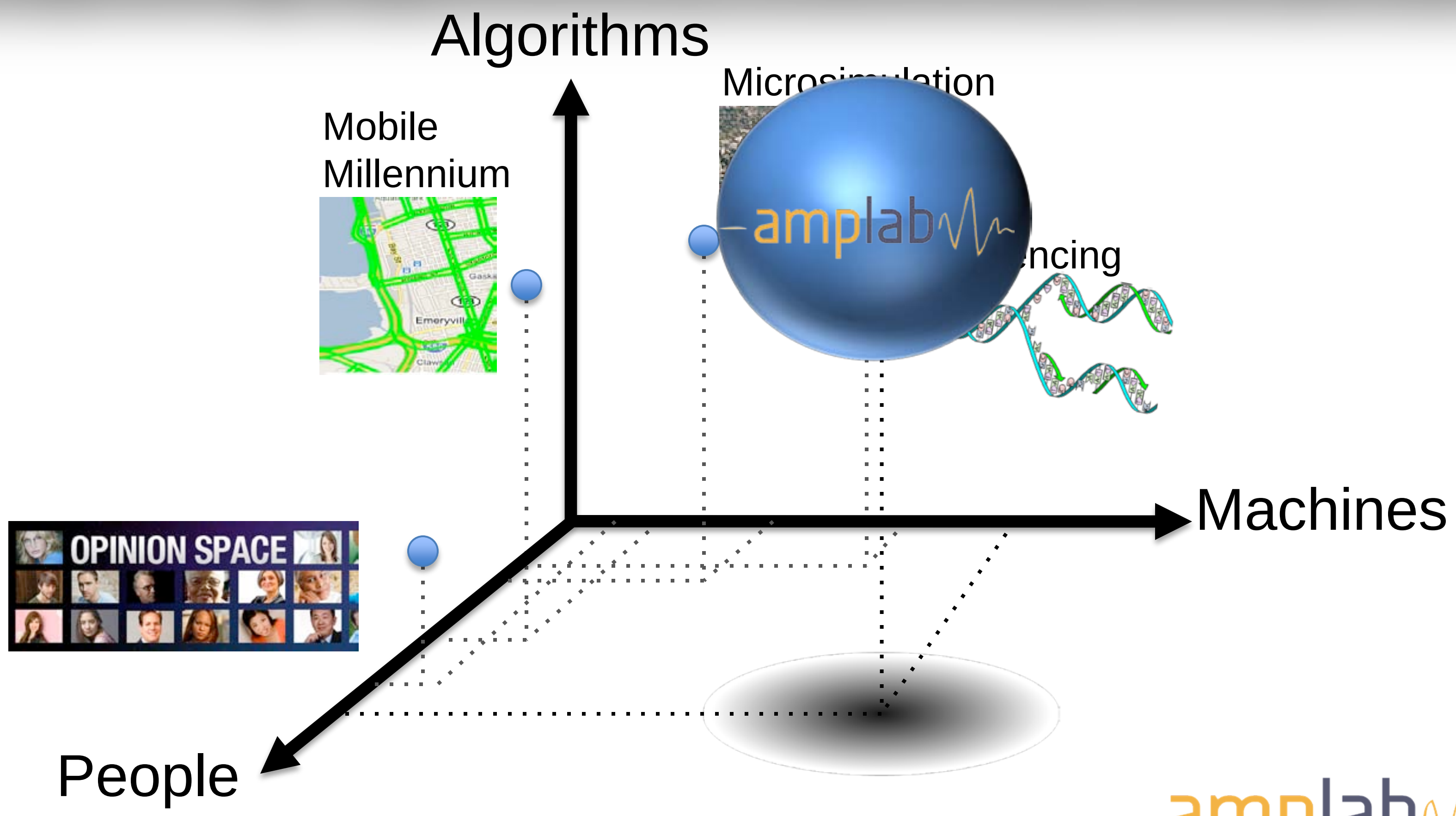
Baseline information

Cost of genome sequencing compared with Moore's law for computers

Source: Broad Institute

The AMP Lab

Make sense of data at scale by tightly integrating algorithms, machines, and people



Big Data in 2020

Almost Certainly:

- Create a new generation of big data scientist
- A real datacenter OS
- ML becoming an engineering discipline
- People deeply integrated in big data analysis pipeline

If We're Lucky:

- System will know what to throw away
- Generate new knowledge that an individual person cannot

Summary

- Goal: Tame Big Data Problem
  - Get results with **right quality** at the right **right time**
- Approach: Holistically integrate **A**lgorithms, **M**achines, and **P**eople
- Huge research issues across many domains