ampcamp
Big Data Bootcamp

AMP Camp Introduction

Michael Franklin

August 21, 2012

amplab
UC BERKELEY

Expeditions in Computing

NSF

The Future is Data-Based

nature

THE BITER BIT
Viral infections for viruses

TROPICAL CYCLONES
The strong get stronger

BLACK HOLE PHYSICS
A new window on the
Galactic Centre

BIG DATA

MICHAEL LEWIS

Moneyball

THE ART OF WINNING AN UNFAIR GAME

The Economist

Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs

The data deluge

AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

McKinsey Global Institute

May 2011

Big data: The next frontier for innovation, competition, and productivity

Thomas H. Davenport • Jeanne G. Harris

Competing on Analytics

The New Science of Winning

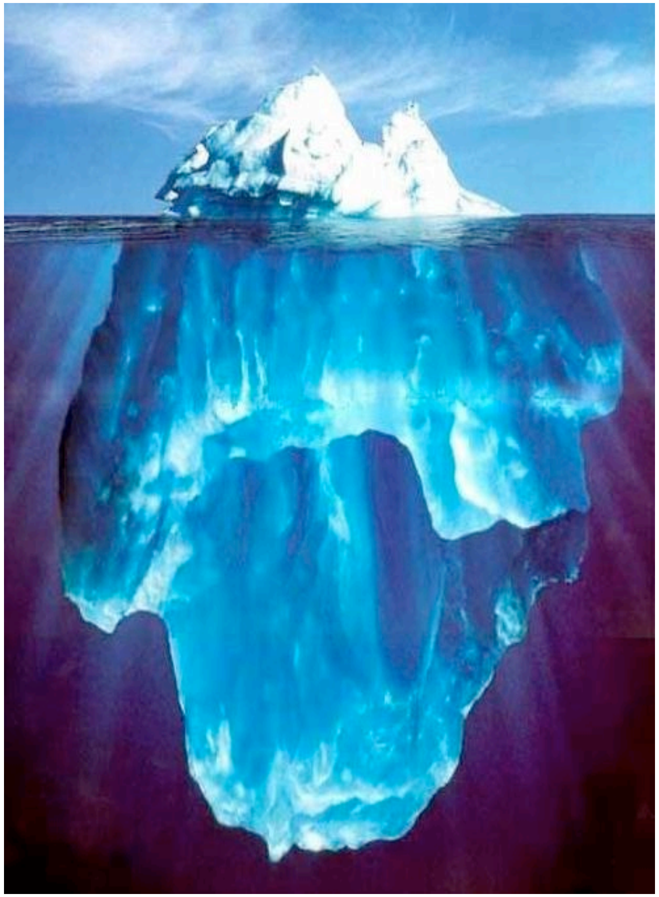
Foreword by Gary Loveman,
CEO of Harrah's Entertainment, Inc.

Harvard Business School Press

amplab

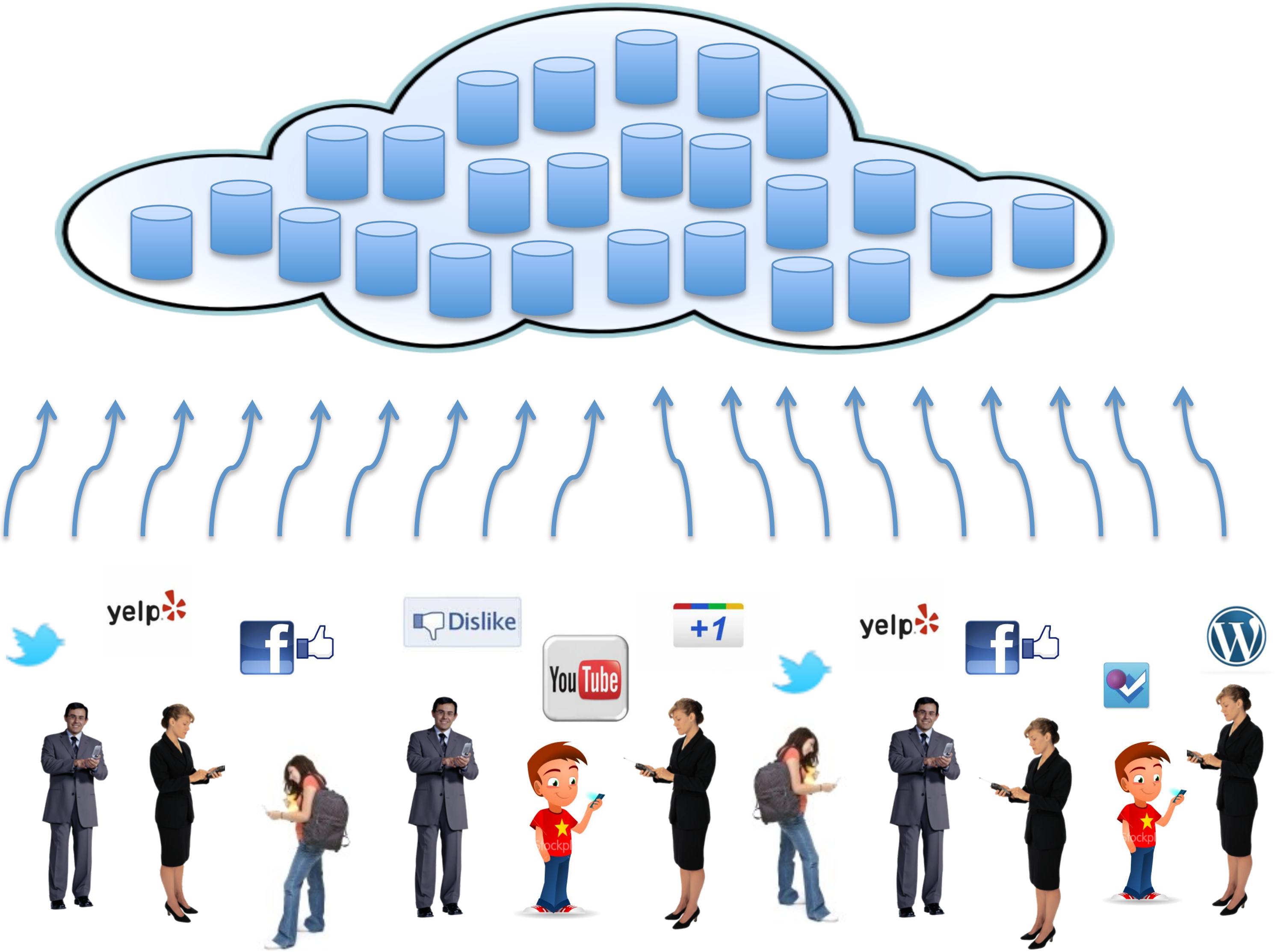
It’s All Happening On-line

- Every:
  - – Click
  - – Ad impression
  - – Wall post, friending, …
  - – Billing event
  - – Fast Forward, pause,…
  - – Server request
  - – Transaction
  - – Network message
  - – Fault
  - – …



- Generates Streams of Data that can be Analyzed

User Generated Content

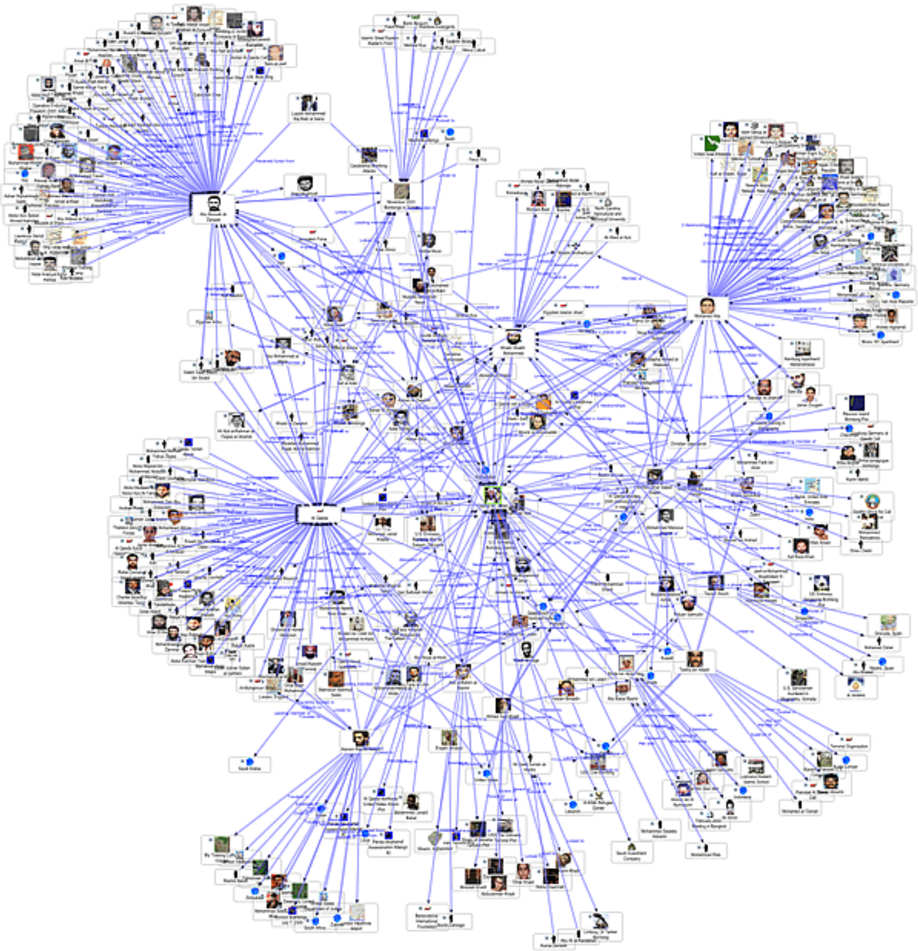


Credit: Mike Carey, UCI

Graph Data

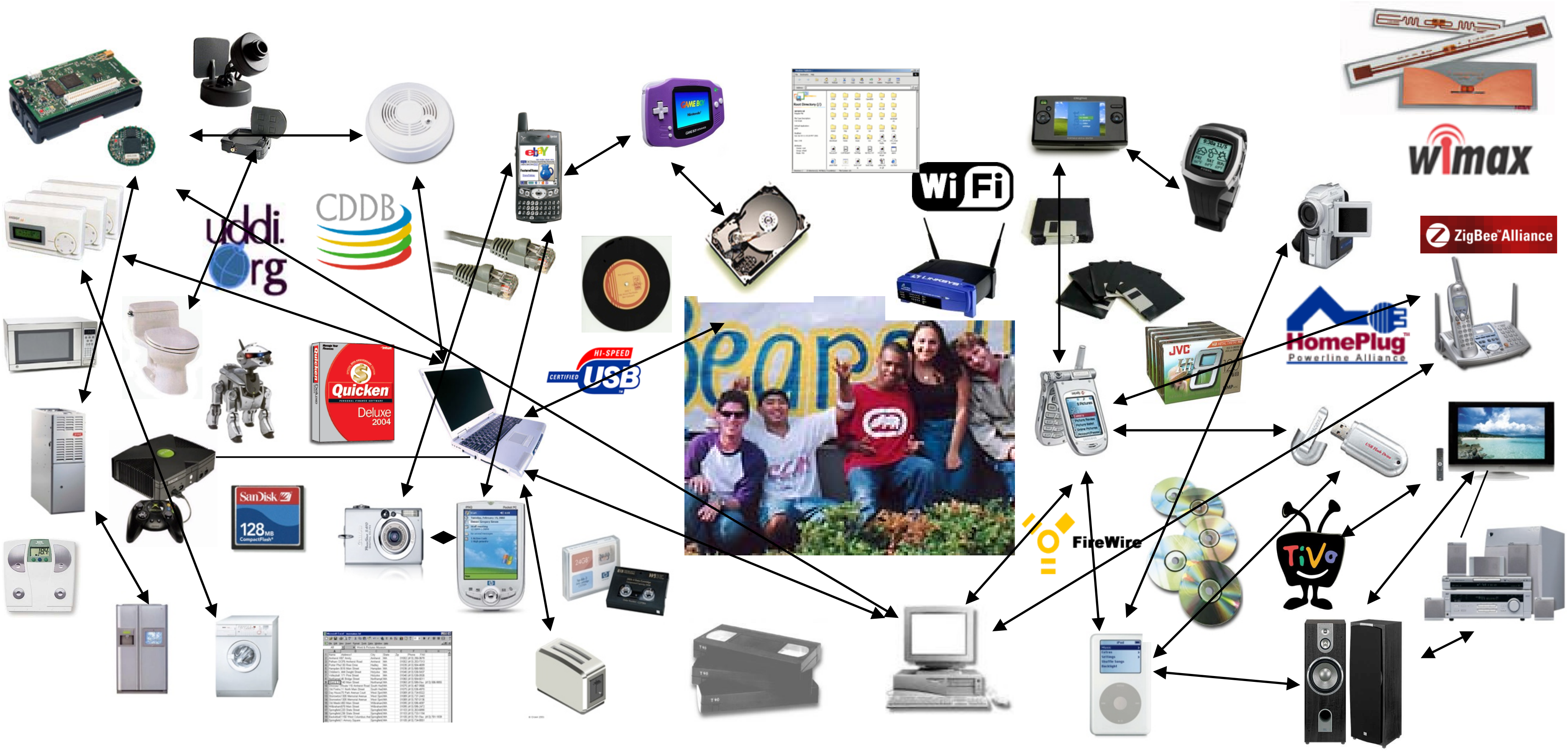
Lots of interesting data has a graph structure:

- Social networks
- Communication networks
- Computer Networks
- Road networks
- Citations
- Collaborations/Relationships
- …

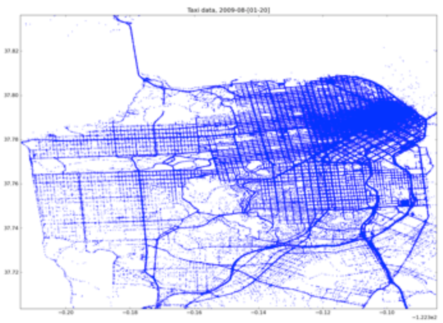


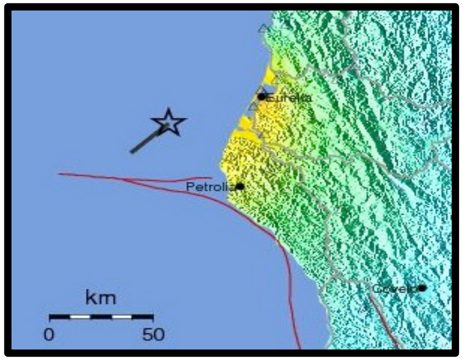
Some of these graphs can get quite large (e.g., Facebook's user graph)

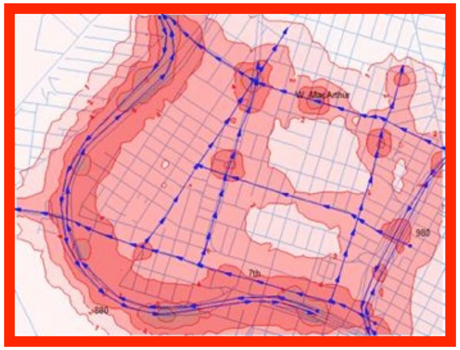
M2M - Internet of things



Fusion: e.g., NextGen Maps







Crowdsourcing + physical modeling + sensing + data assimilation

to produce:

Earthquake
Strong Shaking
in 5 seconds
Get under desk

From Alex Bayen, UCB

What can you do with the data

- Reporting
  - Post Hoc
  - Real time
- Monitoring (fine-grained)
- Exploration
- Finding Patterns
- Root Cause Analysis
- Closed-loop Control
- Model construction
- Prediction
- …

Big Data Explained

Size

(Volume, Velocity)

+

Complexity
(Variety)

=

Answers that don’t meet
quality, **time** and **cost** requirements.

More Data → Better Answers?

More Rows: Algorithmic complexity kicks in

More Columns: Exponentially more hypotheses

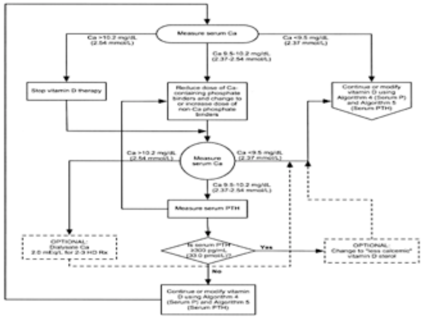
Another formulation of the problem:

- Given an inferential goal and a fixed computational budget, provide a guarantee that the quality of inference will increase monotonically as data accrue (without bound)
- In other words:

Data should be a resource, not a load

Due to Mike Jordan, UCB

The Vision: Algorithms, Machines, People


Measure serum Ca
Ca >10.2 mg/dL (2.54 mmol/L)
Ca 9.5-10.2 mg/dL (2.37-2.54 mmol/L)
Ca <9.5 mg/dL (2.37 mmol/L)
Stop vitamin D therapy
Reduce dose of Ca-containing phosphate binders and change to or increase dose of non-Ca phosphate binders
Continue or modify vitamin D using Algorithm 4 (Serum P) and Algorithm 5 (Serum PTH)
Measure serum Ca
Ca >10.2 mg/dL (2.54 mmol/L)
Ca <9.5 mg/dL (2.37 mmol/L)
Ca 9.5-10.2 mg/dL (2.37-2.54 mmol/L)
Measure serum PTH
OPTIONAL: Dialysate Ca 2.0 mEq/L for 2-3 HD Rx
Is serum PTH ≥300 pg/mL (33.0 pmol/L)?
Yes
OPTIONAL: Change to "less calcemic" vitamin D sterol
No
Continue or modify vitamin D using Algorithm 4 (Serum P) and Algorithm 5 (Serum PTH)

Adaptive/Active Machine Learning and Analytics

Massive and Diverse Data

CrowdSourcing/ Human Computation

Cloud Computing

Why AMP Now?

- Even new "Big Data" stacks respect traditional intellectual borders
  - Need Machine Learning/Systems/Database Co-Design
  - Requires Machine Learning/Systems/Database Cohabitation and Collaboration

- Opportunity to rethink fundamental design points for time-cost-quality:
  - Low Latency
  - Variable Consistency
  - Cloud-based Elastic Resources

- Need to consider role of people *throughout the entire* analytics lifecycle

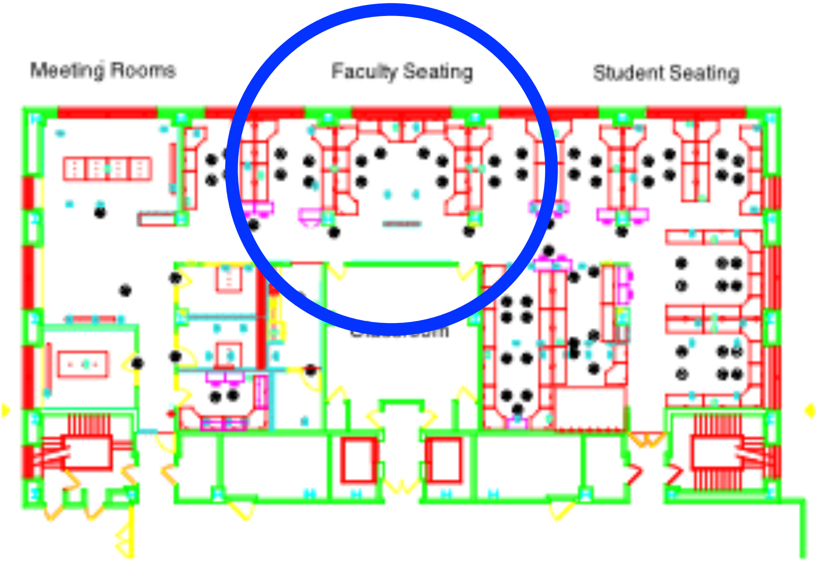
AMPLab Facts

An integration of Faculty Interests (*Directors):

Alex Bayen (Mobile Sensing)	Anthony Joseph (Sec./ Privacy)
Ken Goldberg (Crowdsourcing)	Randy Katz (Systems)
*Michael Franklin (Databases)	Dave Patterson (Systems)
Armando Fox (Systems)	*Ion Stoica (Systems)
*Mike Jordan (Machine Learning)	Scott Shenker (Networking)

~60 World-leading students, post-docs & visitors

Organized for Collaboration:



AMP Facts (continued)

- Started February 2011; 5 (+1) Yr Duration
- Strong industry relationships & support

Founding Sponsors:

amazon.com web services   Google   SAP

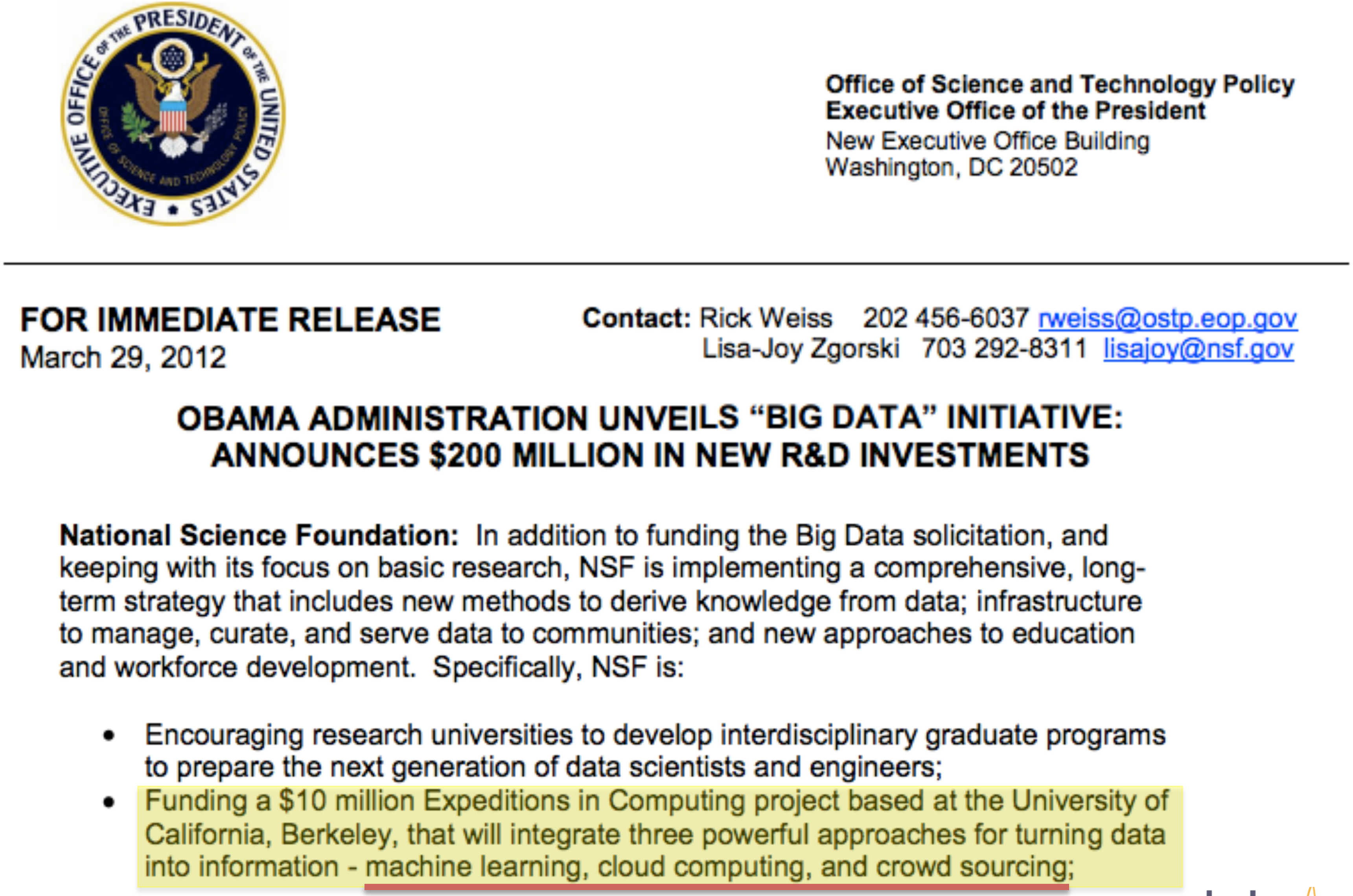
Sponsors and Affiliates:

BLUE GOJI   CISCO   cloudera   ERICSSON   GE imagination at work   hp   HUAWEI

intel   Microsoft   NetApp   ORACLE   Quanta Computer   splunk>   vmware

- NSF Expedition and Darpa XData
- All software released as BSD Open Source

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:

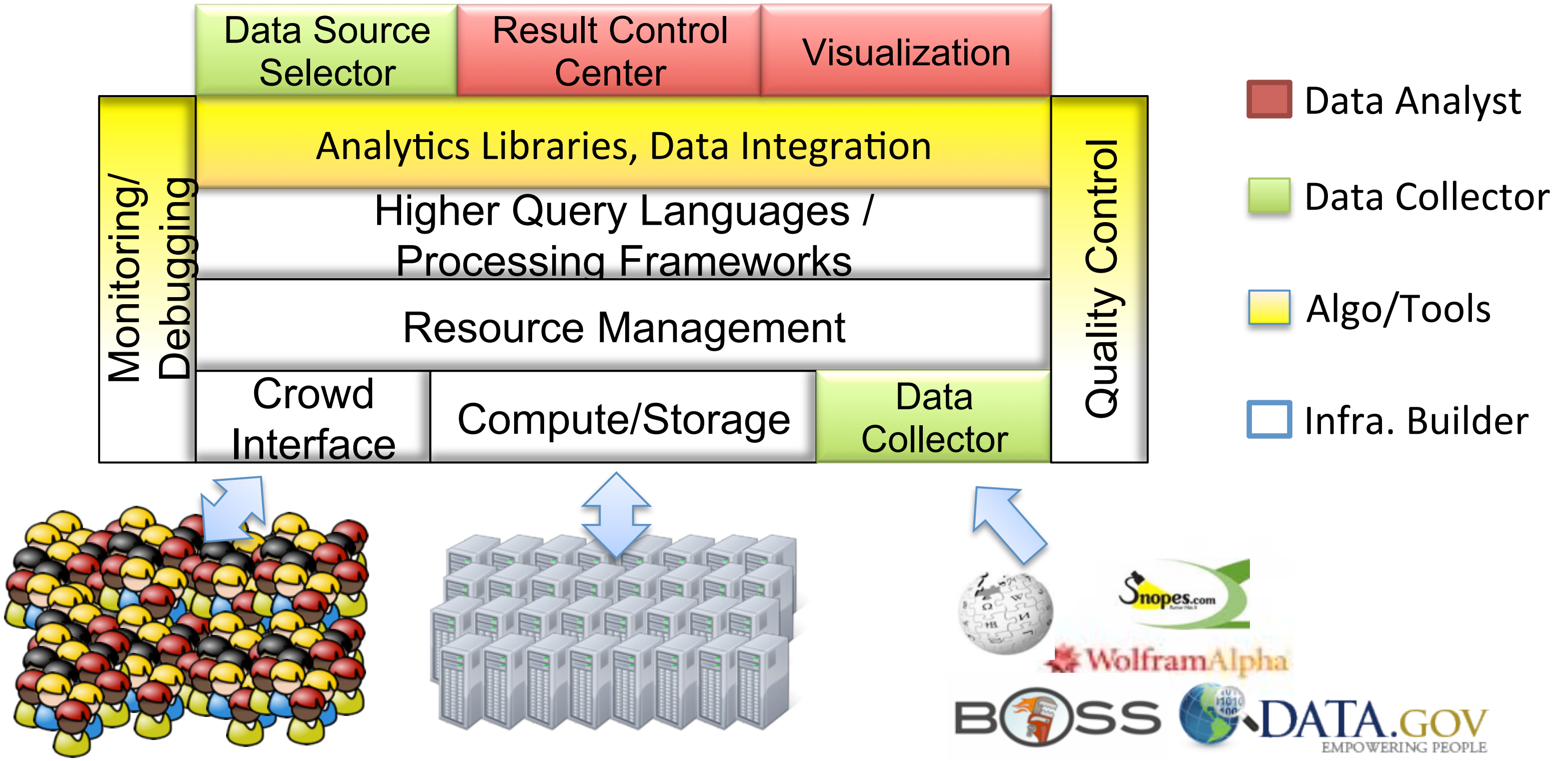
- Encouraging research universities to develop interdisciplinary graduate programs to prepare the next generation of data scientists and engineers;
- Funding a $10 million Expeditions in Computing project based at the University of California, Berkeley, that will integrate three powerful approaches for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

How we work with Industry

- Industry relationships are one of our “unfair advantages”
- Key relationships, insights, problems, guidance, funding, data
- Twice-yearly, in-depth research retreats
  - 20+ Companies and Labs
  - AMP Camp is result of sponsor feedback at previous retreat
- Internships and Collaborations
- Open source and technology transfer => research results and tools will be widely available

ampcamp
Big Data Bootcamp

BDAS: Berkeley Data Analysis System



Data Source Selector

Result Control Center

Visualization

Analytics Libraries, Data Integration

Higher Query Languages / Processing Frameworks

Resource Management

Crowd Interface

Compute/Storage

Data Collector

Monitoring/ Debugging

Quality Control

🟥 Data Analyst

🟩 Data Collector

🟨 Algo/Tools

⬜ Infra. Builder

A new open source software stack to:

Effectively manage cluster resources

Efficiently extract value out of big data

Continuously optimize Cost, Time, and Answer Quality

Application Partners

Participatory Sensing

Mobile Millenium (Alex Bayen)

Collective Discovery

Opinion Space (Ken Goldberg)

Urban Planning and Simulation

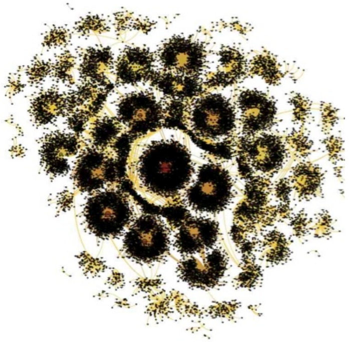
UrbanSim (Paul Waddell)

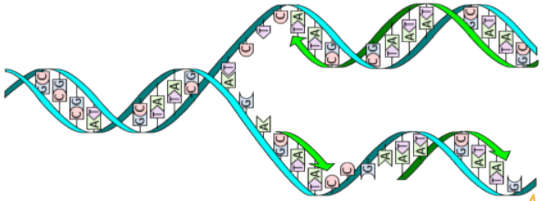
Cancer Genomics/Personalized Medicine

X-Prize(Taylor Sittler, UCSF)

Internet Security

VAST (Vern Paxson)



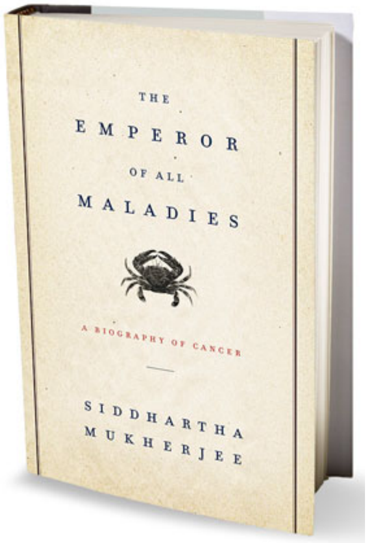


amplab

Application Requirements

	Mobile Millennium	Opinion Space	Tumor Genomics	Urban Planning	Internet Security
Large Data Volumes	●	◓	●	●	●
Data Integration	◓	◓	●	●	
Crowdsourcing	●	●	◓	◓	
Computationally Intensive	●		●	●	●
Real Time Analysis	●	●			●
Sensor/Physical Data	●		●	◓	◓
Text/ Unstructured		●	●	●	●

Big Data, Societal-Scale App?



- Cancer Tumor Genomics
- Vision: Personalized Therapy
  - "…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."

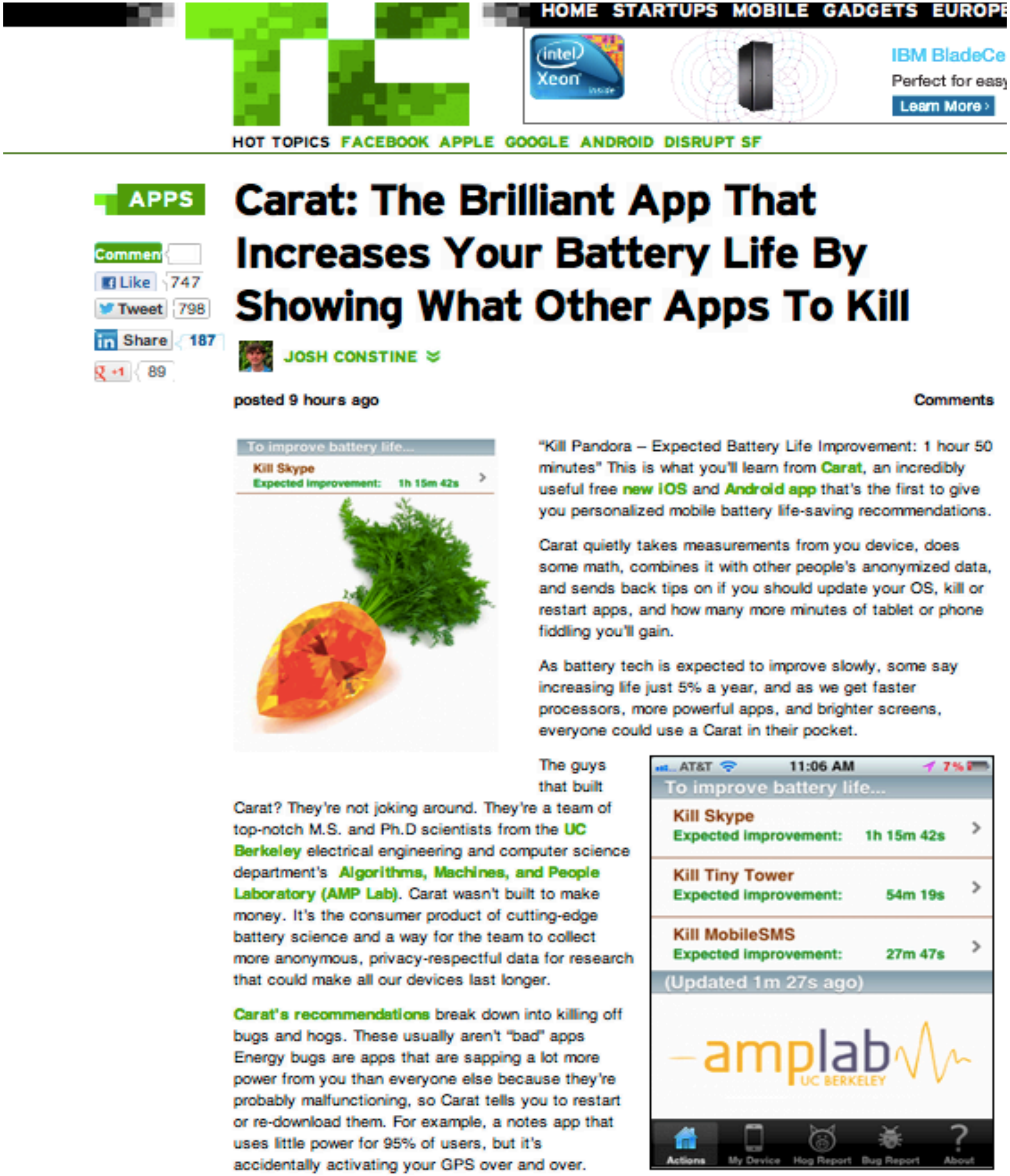
Director, The Cancer Genome Atlas (TCGA), *Time Magazine,* 6/13/11

- Sequencing costs ↓ (150X) → Big Data ↑
- Opportunity: UCSF cancer researchers + UCSC cancer genetic database + AMP Lab
  - TCGA: 5 PB = 20 cancers x 1000 genomes

"DNA Sequencing Caught in Deluge of Data," Andrew Pollack, *New York Times*, 11/30/11

Opportunity or Obligation?

- Provocative Hypothesis: Given fast growing genomic databases, could CS now be a huge help in war on cancer?
- If a *chance* that we 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**



HOME STARTUPS MOBILE GADGETS EUROPE

intel Xeon inside
IBM BladeCe
Perfect for eas
Learn More ›

HOT TOPICS FACEBOOK APPLE GOOGLE ANDROID DISRUPT SF
APPS
Comment
Like 747
Tweet 798
Share 187
+1 89

Carat: The Brilliant App That Increases Your Battery Life By Showing What Other Apps To Kill

JOSH CONSTINE

posted 9 hours ago

Comments

"Kill Pandora – Expected Battery Life Improvement: 1 hour 50 minutes" This is what you'll learn from Carat, an incredibly useful free new iOS and Android app that's the first to give you personalized mobile battery life-saving recommendations.

Carat quietly takes measurements from you device, does some math, combines it with other people's anonymized data, and sends back tips on if you should update your OS, kill or restart apps, and how many more minutes of tablet or phone fiddling you'll gain.

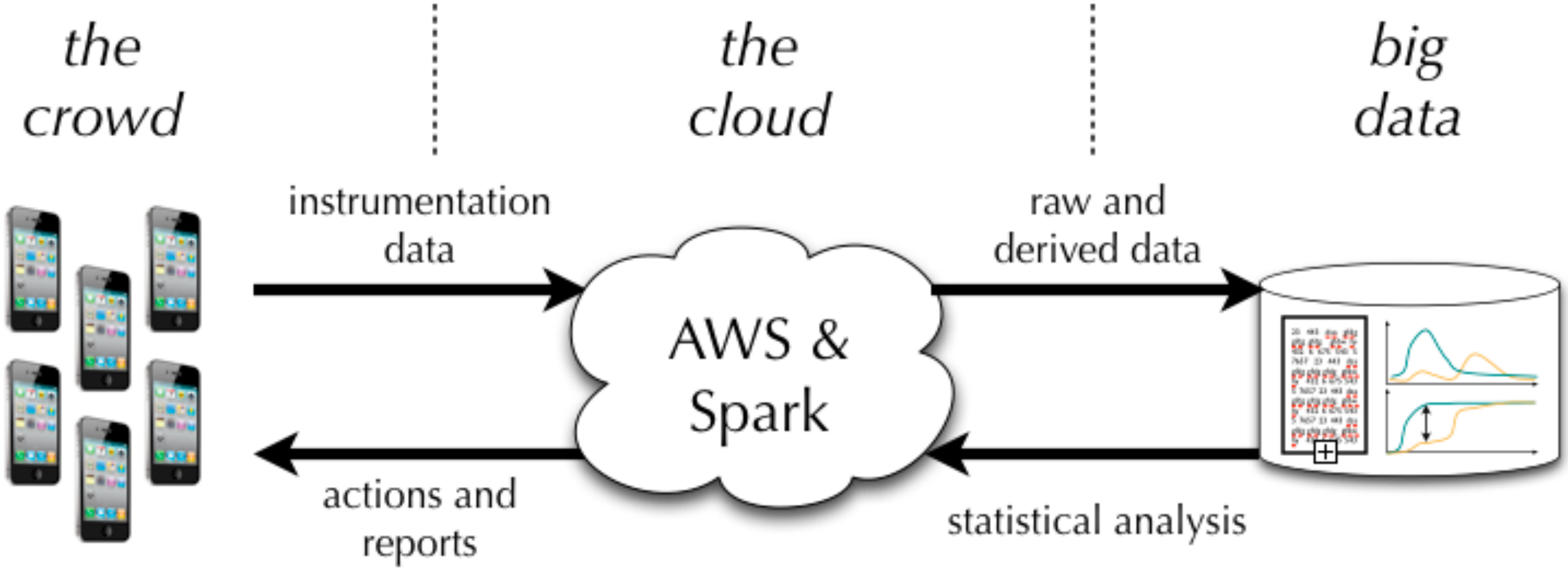
As battery tech is expected to improve slowly, some say increasing life just 5% a year, and as we get faster processors, more powerful apps, and brighter screens, everyone could use a Carat in their pocket.

The guys that built Carat? They're not joking around. They're a team of top-notch M.S. and Ph.D scientists from the UC Berkeley electrical engineering and computer science department's Algorithms, Machines, and People Laboratory (AMP Lab). Carat wasn't built to make money. It's the consumer product of cutting-edge battery science and a way for the team to collect more anonymous, privacy-respectful data for research that could make all our devices last longer.

Carat's recommendations break down into killing off bugs and hogs. These usually aren't "bad" apps Energy bugs are apps that are sapping a lot more power from you than everyone else because they're probably malfunctioning, so Carat tells you to restart or re-download them. For example, a notes app that uses little power for 95% of users, but it's accidentally activating your GPS over and over.

> 300,000 Downloads in 2 Months

Carat: A Quintessential AMP App



Collaborative Detection of Energy Bugs

For More Information

amplab.cs.berkeley.edu

- Papers and Project Pages
- News updates and Blogs

Spark Meet-Up & User Group

Github and Apache Mesos