CS294:
AI for Systems and Systems for AI
Logistics, Overview, Trends

Joey Gonzalez and Ion Stoica

January 23, 2019

Course Information

Course website:

- https://ucbrise.github.io/cs294-ai-sys-sp19/
  - It is on Github so you can contribute content!

Please subscribe to piazza:

- http://piazza.com/berkeley/spring2019/cs294159/home

Schedule and reading list are subject to change

Tentative Lecture Format (not today!)

First 1/3 of each lecture presented by faculty

Second 2/3 covers papers presented by students

- Two presentations of 15min each, followed by 20min discussion

Reading assignments

All students are required to read all papers

All students are required to submit “reviews” for each paper:

- Answer short questions on google form: identify key insights, strengths and weaknesses
- We will send out how to signup for papers later today

Grading Policy

40% Class Participation

- Answer questions, join discussion, and present papers

10% Initial Project Proposal Presentation

- Presented in class on 3/11

25% Project Poster Session

- TBA

25% Final Project Reports

- TBA

What do you expect from this class?

Main goal

Identify and solve big/impactful problems at the intersection between AI and Systems

How?

Study:

- Major AI developments through systems' lens
- System-related problems that might leverage AI techniques

Enable System and AI students to collaborate on projects

Play to Berkeley's strengths

  - World-class in both AI and Systems
  - Long tradition of collaboration between AI and Systems

A Short Story…

RAD Lab

EECS Berkeley

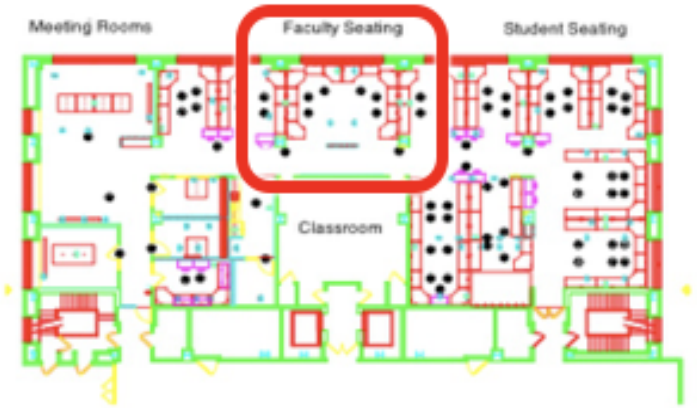
ABOUT PEOPLE PUBLICATIONS SPONSORS MEDIA CENTER RESEARCH LOGIN

2006-2011

SEARCH

- Get Involved
- Come Visit
- Contact

Featured Posts



About

Although large-scale Internet services such as eBay and Google Maps have revolutionized the Web, today it takes a large organization with tremendous resources to turn a prototype or idea into a robust distributed service that can be relied on by millions.

Our vision is to enable one person to invent and run the next revolutionary IT service, operationally expressing a new business idea as a multi-million-user service over the course of a long weekend. By doing so we hope to enable an Internet "Fortune 1 million".

The image part with relationship ID rId3 was not found in the file.

RAD Lab

EECS Berkeley UNIVERSITY OF CALIFORNIA

ABOUT PEOPLE PUBLICATIONS SPONSORS MEDIA CENTER RESEARCH LOGIN

2006-2011

SEARCH

- Get Involved
- Come Visit
- Contact

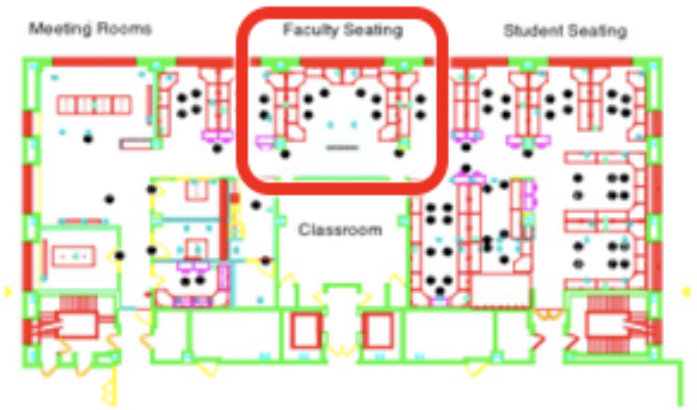
About

Although large-scale Internet services such as eBay and Google Maps have revolutionized the Web, today it takes a large organization with tremendous resources to turn a prototype or idea into a robust distributed service that can be relied on by millions.

Our vision is to enable one person to invent and run the next revolutionary IT service, operationally expressing a new business idea as a multi-million-user service over the course of a long weekend. By doing so we hope to enable an Internet "Fortune 1 million".

Featured Posts

The image part with relationship ID rId3 was not found in the file.



RAD Lab

EECS Berkeley

ABOUT PEOPLE PUBLICATIONS SPONSORS MEDIA CENTER RESEARCH LOGIN

2006-2011

SEARCH

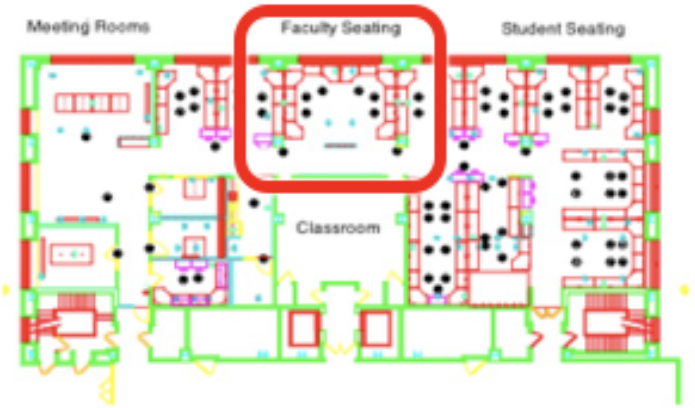
- Get Involved
- Come Visit
- Contact

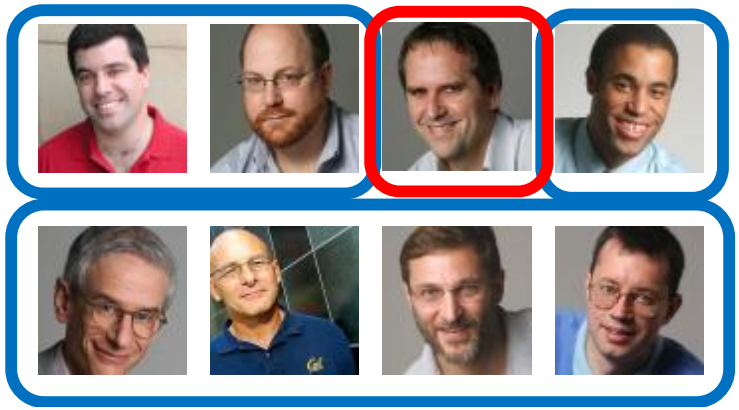
About

Although large-scale Internet services such as eBay and Google Maps have revolutionized the Web, today it takes a large organization with tremendous resources to turn a prototype or idea into a robust distributed service that can be relied on by millions.

Our vision is to enable one person to invent and run the next revolutionary IT service, operationally expressing a new business idea as a multi-million-user service over the course of a long weekend. By doing so we hope to enable an Internet "Fortune 1 million".

Featured Posts





Systems,
Networking
Databases

2009

Lester Mackey
(postdoc w/ Michael Jordan)

NETFLIX Prize

anonymized movie rating dataset

best recommendation algorithm

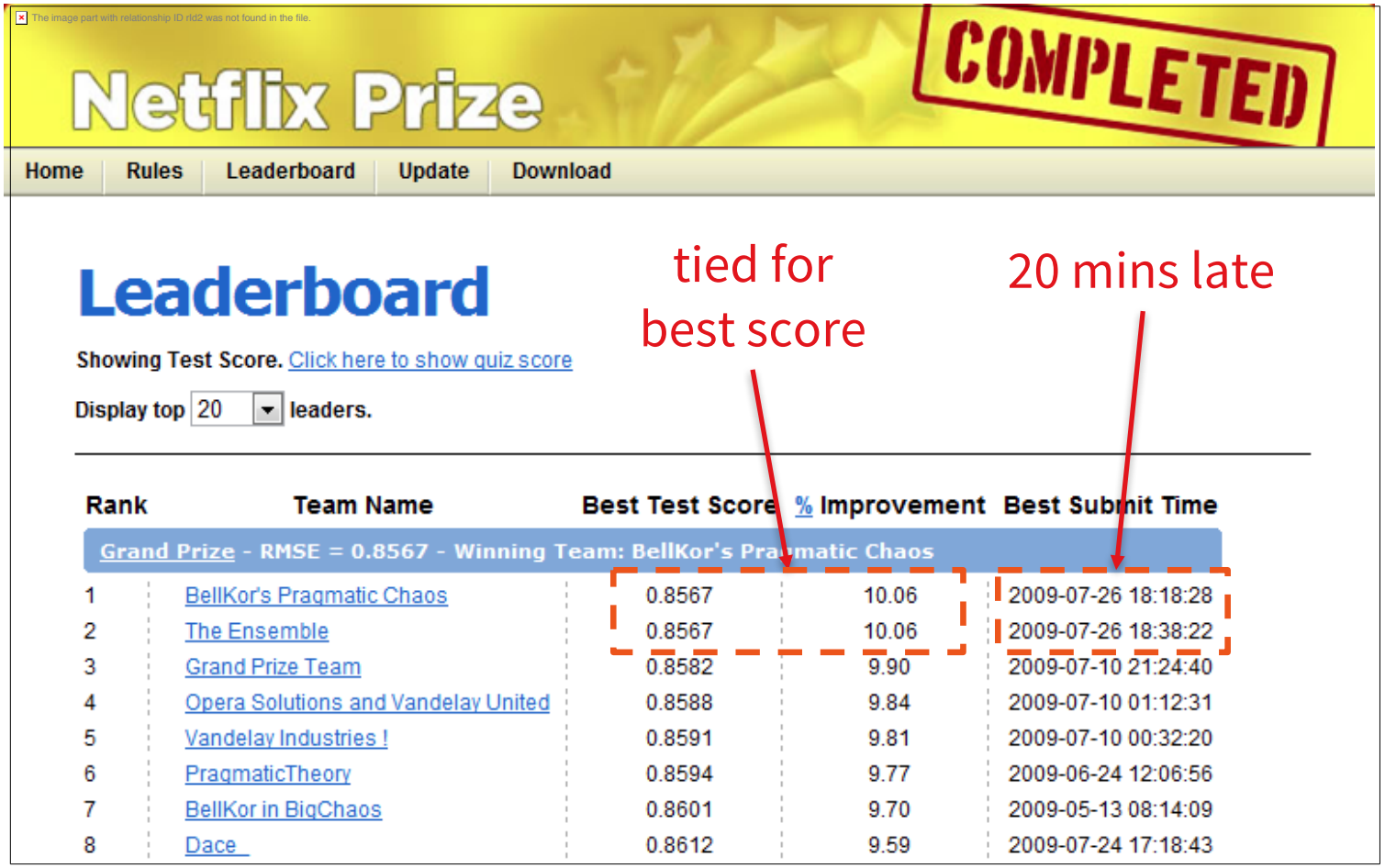
$1m


hadoop

Lester Mackey

Matei Zaharia


APACHE
Spark
TM


The image part with relationship ID rId2 was not found in the file.
Netflix Prize
COMPLETED
Home
Rules
Leaderboard
Update
Download
Leaderboard
tied for best score
20 mins late
Showing Test Score. Click here to show quiz score
Display top 20 leaders.
Rank
Team Name
Best Test Score
% Improvement
Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos
1 BellKor's Pragmatic Chaos 0.8567 10.06 2009-07-26 18:18:28
2 The Ensemble 0.8567 10.06 2009-07-26 18:38:22
3 Grand Prize Team 0.8582 9.90 2009-07-10 21:24:40
4 Opera Solutions and Vandelay United 0.8588 9.84 2009-07-10 01:12:31
5 Vandelay Industries ! 0.8591 9.81 2009-07-10 00:32:20
6 PragmaticTheory 0.8594 9.77 2009-06-24 12:06:56
7 BellKor in BigChaos 0.8601 9.70 2009-05-13 08:14:09
8 Dace 0.8612 9.59 2009-07-24 17:18:43


NETFLIX
2009
DATE 09.21.09
PAY TO THE ORDER OF BellKor's Pragmatic Chaos
$ 1,000,000 00
AMOUNT ONE MILLION
00/100
FOR The Netflix Prize
Reed Hastings

Meanwhile…

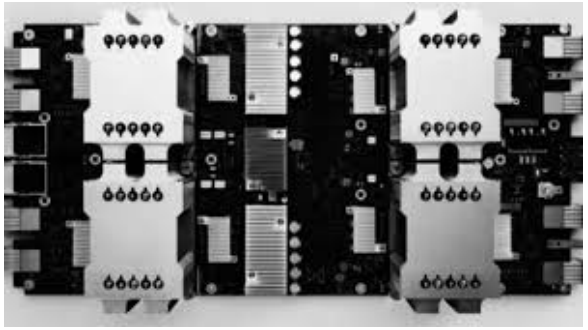
Joey morphing from an ML to a system student ;-)

GraphLab

Today

Jeff Dean's favorite talk about how

- AI can get rid of all hacks and heuristics in systems
- And how hardware & software can revolutionize AI



TensorFlow

Proliferation of venues…

Fri Dec 7th 08:00 AM -- 06:30 PM @ Room 510 ABCD
MLSys: Workshop on Systems for ML and Open Source Software
Aparna Lakshmiratan · Sarah Bird · Siddhartha Sen · Joseph Gonzalez · Daniel Crankshaw

Workshop

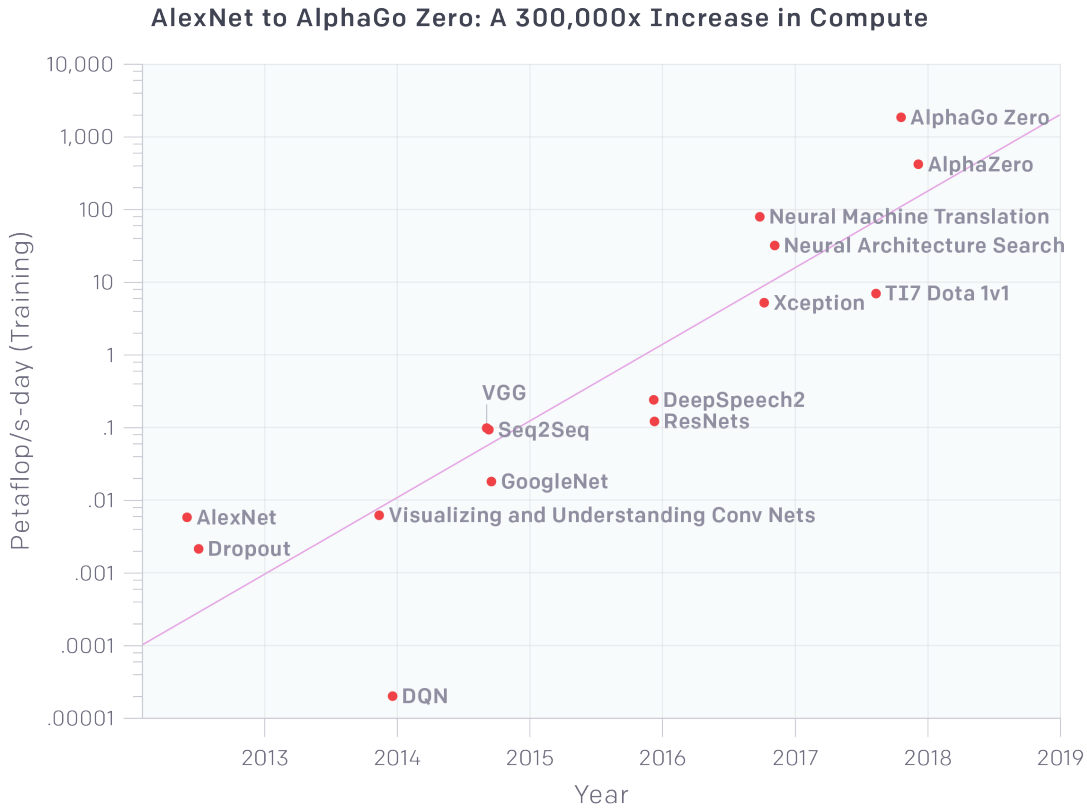
Sat Dec 8th 08:00 AM -- 06:30 PM @ Room 510 AC
Machine Learning for Systems
Anna Goldie · Azalia Mirhoseini · Jonathan Raiman · Kevin Swersky · Milad Hashemi

Workshop

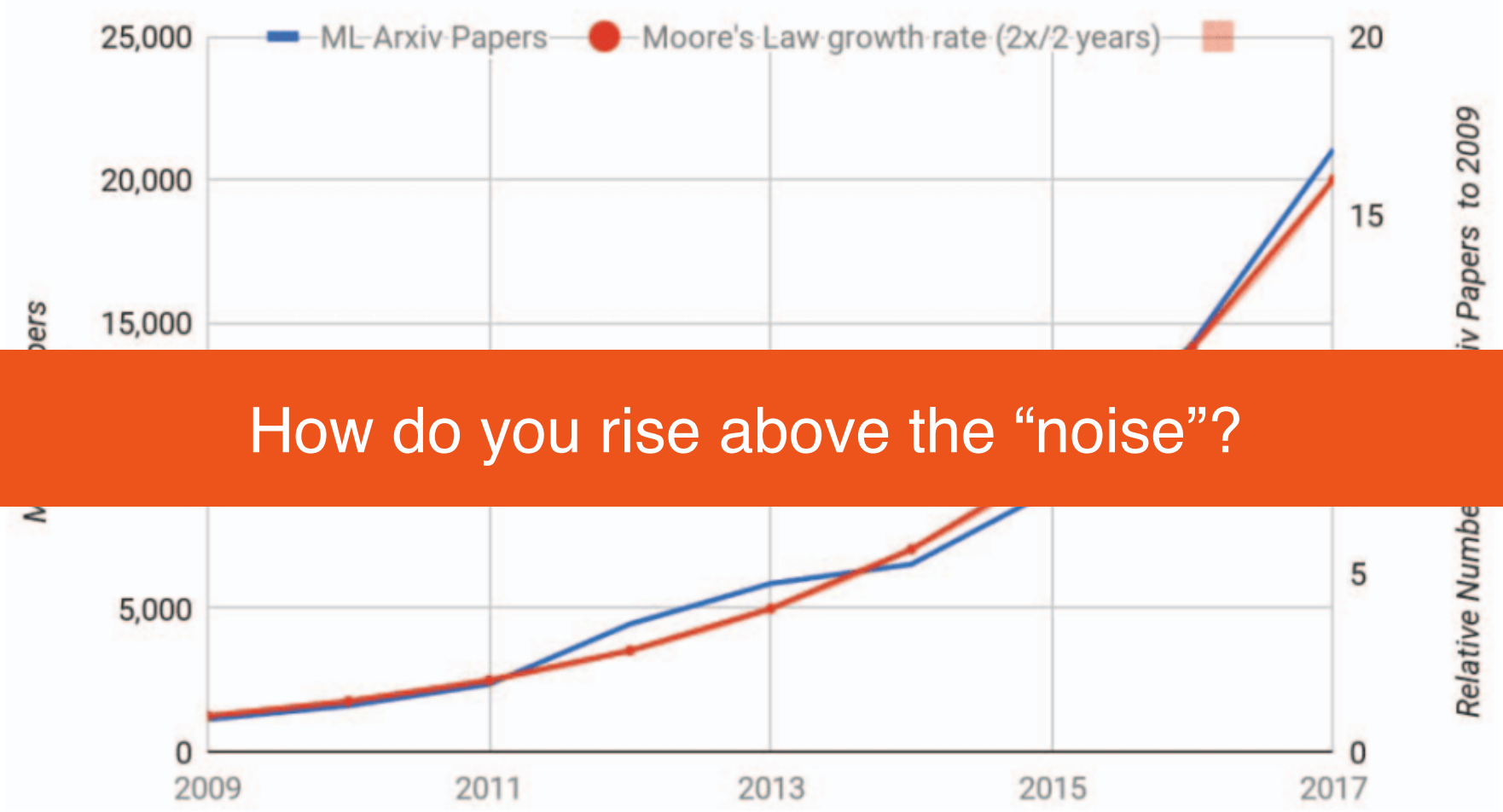
SysML Conference

March 31 – April 2, 2019
Stanford, CA

The need for systems



AI and Compute (https://blog.openai.com/ai-and-compute/)



How do you rise above the “noise”?

“A New Golden Age in Computer Architecture: Empowering the Machine-Learning Revolution”, https://ieeexplore.ieee.org/document/8259424

Work on great problems!

Disclaimer

My (biased) opinion

Based on my experience

Not the only path to success

You want “proof”?

How did the last AI revolution started?

Providing state-of-the-art results on OCR

Gradient-Based Learning Applied to Document Recognition

16,000+ citations

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner

You want “proof”?

How did the last AI revolution started?

Providing state-of-the-art results on image recognition

ImageNet Classification with Deep Convolutional Neural Networks

34,500+ citations

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Another line of argument…

All of you are good problem solvers

- Good grades
- Good scores
- …

As such, the biggest differentiator is the problem you are working on

What is a good problem to work on?

Impact: someone cares about solving the problem

- The more people care, the better

Priority: you are among the first to work on the problem

- Increases the chances of being first to solve this problem
- Caveat: This is just “nice to have”. If you believe it’s an important problem and you are passionate bout it, go for it!

Edge: You have an expertise (edge) to solve the problem

Incremental: can be partitioned into sub-problems

- Help you make rapid progress, iterate, better understand the problem

Some problems I worked on

Problem	***Impact*** ***(problem formulation)***	***Priority***	***Edge***	***Incremental***	***Comments***
Chord	Need for decentralized (no liability), efficient P2P system	Among several groups working on it	Experts in distributed systems and algorithms	Followed by many refinements	At some point, most cited paper **13,600+ citations**

Some problems I worked on

Problem	***Impact*** ***(problem formulation)***	***Priority***	***Edge***	***Incremental***	***Comments***
Chord	Need for decentralized (no liability), efficient P2P system	Among several groups working on it	Experts in distributed systems and algorithms	Followed by many refinements	At some point, most cited paper **13,600+ citations**
Spark	Need for fast, general-purpose big data analytics	Among first to see ML use cases (AMPLab) and interactive queries (industry; Conviva, FB)	Already experts in big data (in academia); started working on this 3 years earlier	Core Spark (3,000 LoC); Shark, Streaming, ML libs	High impact in industry **3,940+ citations**

Some problems I worked on

Problem	***Impact*** ***(problem formulation)***	***Priority***	***Edge***	***Incremental***	***Comments***
Chord	Need for decentralized (no liability), efficient P2P system	Among several groups working on it	Experts in distributed systems and algorithms	Followed by many refinements	At some point, most cited paper **13,600+ citations**
Spark	Need for fast, general-purpose big data analytics	Among first to see ML use cases (AMPLab) and interactive queries (industry; Conviva, FB)	Already experts in big data (in academia); started working on this 3 years earlier	Core Spark (3,000 LoC); Shark, Streaming, ML libs	High impact in industry **3,940+ citations**
Dominant Resource Fairness (DRF)	Need for better utilization for multi resource types workloads	Among first to see the problem (Mesos)	Experts in scheduling algorithms	Followed by other papers in different areas	Intellectual impact (microeconomy); implemented in a few systems **800+ citations**

Impact (1/2)

Do things that have not been done before

- "It always seems impossible until is done" – Nelson Mandela
- Examples: zero-knowledge proof, beating chess/go world champion, outperform experts in medical diagnosis, …

Do things much better than before (e.g., 10x)

- Examples: GPUs, Distributed Hash Tables (routing efficiency), in-memory databases, CNNs, …

Impact (2/2)

Theory

- Improve space/time complexity of wide used algo (e.g., SGD)
- Extend existing techniques to new domains/environments, e.g., fair scheduler from network to CPU, distributed graph algorithms
- Prove new desirable properties for popular techniques (increase practitioners confidence in using them), e.g., show an allocation algo is strategy proof, show an algo is guaranteed to converge

Systems

- Support new workloads impractical before, e.g., big data processing
- Significantly improve efficiency/speed/cost, e.g., GPUs, dist algos
- “Democratize” technology, e.g., cloud computing, serverless, TensorFlow

Priority (1/2)

You experience problem first hand

- One of the main reasons we are building systems/artifacts
  1. Build system solving a “need”
  2. Have people use it
  3. Understand things people want to do with your system, you haven’t thought about
  4. Now you have a problem, you’ve “seen” first!

Take a bet on a new area and dive in

- Because it’s new area, few people are working on it
- This is one thing Berkeley has been excelling at, e.g.,
  - Complexity theory, approximate algorithms, databases, networking, sensor networks, OSes (UNIX), NOW, Big Data, ML, …

Priority (2/2)

Talk with people from other areas to find new problems, e.g.,

- Spark motivated by Matei's trying to solve Lester Mackey's

Berkeley's labs are set up to encourage you do exactly this!

> *"I notice that if you have the door to your office closed, you get more work done today and tomorrow, and you are more productive than most. But 10 years later somehow you don't know quite know what problems are worth working on; all the hard work you do is sort of tangential in importance. … I can say there is a pretty good correlation between those who work with the doors open and those who ultimately do important things*."
>
> – Richard Hamming, Turing Award Winner
>
> ("You and Your Research" : http://www.cs.virginia.edu/~robins/YouAndYourResearch.html)

Edge

If you do not have expertise to solve the problem, talk with people who do!

- Lester leveraged Matei's help to rank 2nd for Netflix's challenge

Again, labs like RISELab are set up to encourage you do exactly this!

Incrementality

Reach the first meaningful milestone fast

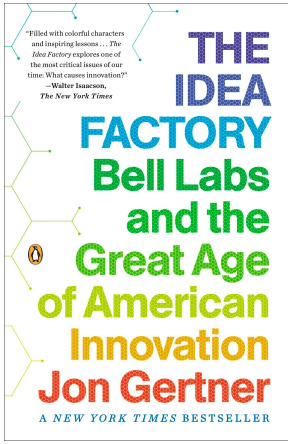
- Critical to get the project off the ground, maintain the excitement of the students and, if a system, start growing community
- Examples: Mesos, Spark provided an useful system within 1 year

Decompose project in a series of milestones providing increasing value

- Examples: Spark core → Shark (SQL) → Streaming → Mllib
- Developing a chess algorithm to first beat an amateur, then a master, then a grand master, and finally the world champion

Holly grail example: Bell Labs achieving its vision of "enabling any two people in the world communicate"

- Enable two people to talk by phone between Boston and NY
- Enable two people to talk by phone across US between NY and SF
- Enable two people to talk by phone across Atlantic between NY and London…. (see "The Idea Factory" book)



Questions to ask when picking a problem

What will happen if I solve this problem? Will anyone care? Will anything change in the way people do things?

- If you hope for a quantitative gain, simplify the problem and see improvement; if improvement not what you hopped for, reconsider

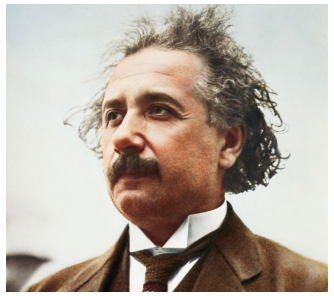
Who else is working on the problem? (Again, big caveat here; if you believe the problem is important an you can solve it go for it!)

Why me? Do I have any edge on solving this problem? Am I expert in some techniques I think it will help me solve it? Did I solve similar problems?

Is there a meaningful milestone I can get to fast?

“The mere formulation of a problem is far more often essential than its solution, which may be merely a matter of mathematical or experimental skill. To raise new questions, new possibilities, to regard old problems from a new angle requires creative imagination and marks real advances in science.”

Albert Einstein (1879 - 1955) Physicist & Nobel Laureate



You want proof?

Why did RL become popular?

Because Atari games, Go

Playing Atari with Deep Reinforcement Learning

1,938 citations

Volodymyr Mnih Koray Kavukcuoglu David Silver Alex Graves Ioannis Antonoglou

Daan Wierstra Martin Riedmiller

nature
International journal of science

Article | Published: 18 October 2017

Mastering the game of Go without human knowledge

1,143 citations

David Silver ✉, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel & Demis Hassabis