Machine Learning Systems

(294-162)

Joseph E. Gonzalez
Co-director of the RISE Lab
jegonzal@cs.berkeley.edu

About Me

About Me

- Co-director of the RISE Lab
- Co-founder of Turi Inc.

Research

- Machine Learning
- Distributed Systems
- Computer Vision
- Autonomous Driving
- Secure Learning
- …

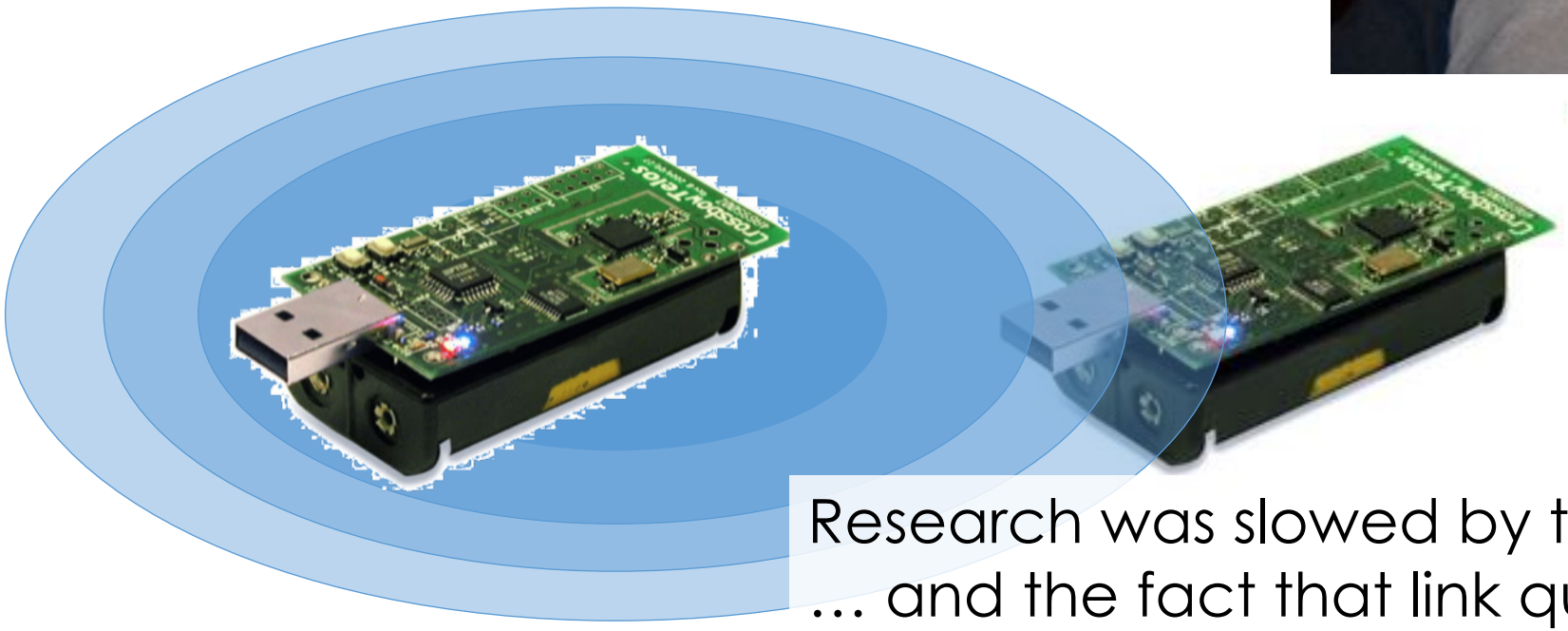
My story …

Machine Learning → Learning Systems

Back in 2007

Back in 2007

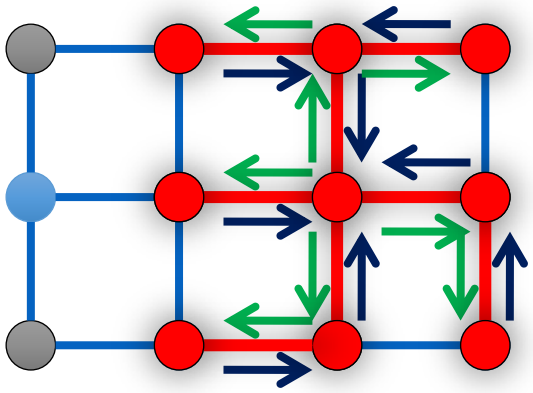
I started studying the use of Gaussian Process (GP) models for wireless link quality estimation

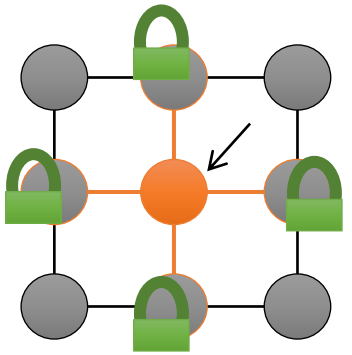


Research was slowed by the speed of training
… and the fact that link quality is difficult to predict

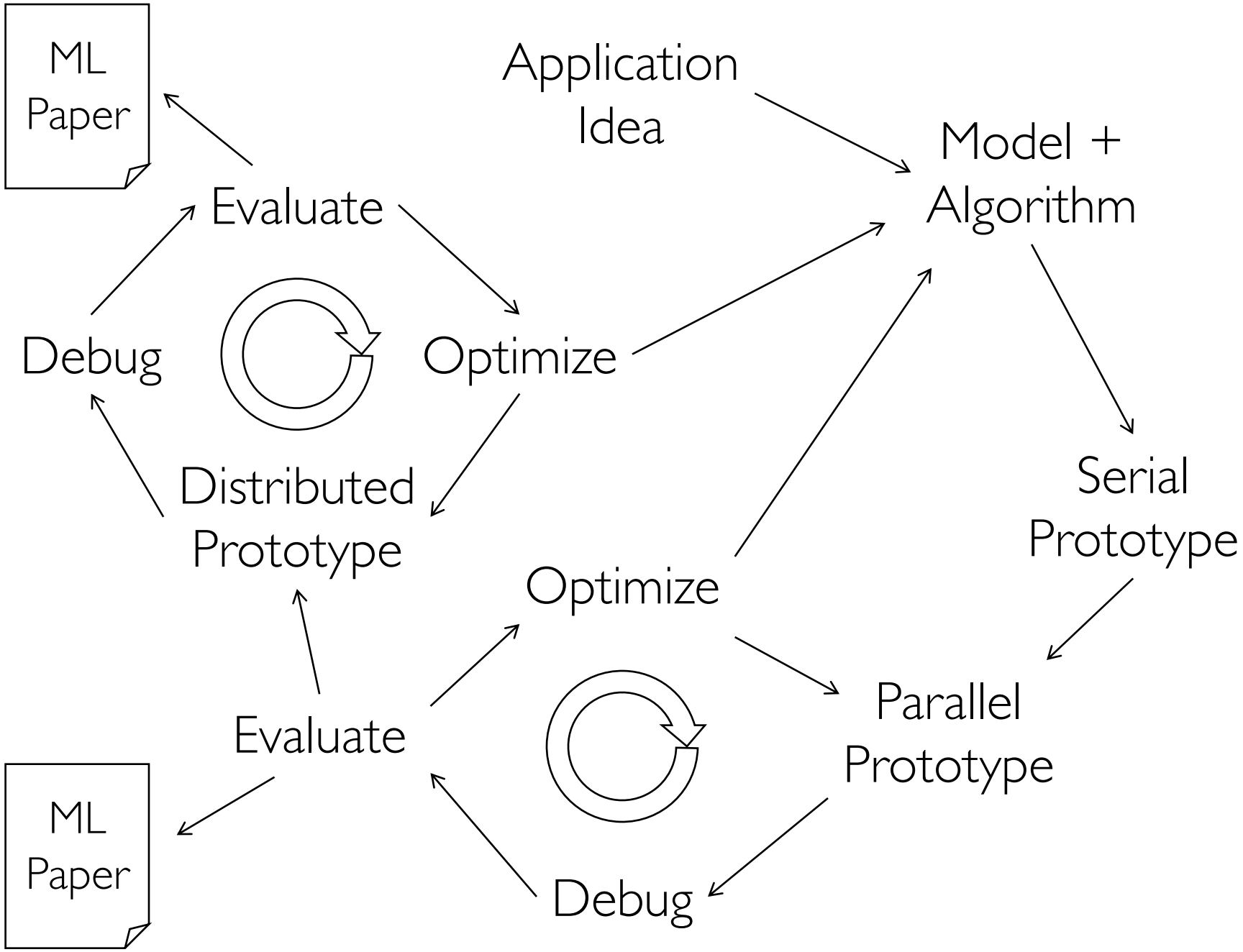
Back in 2008

I started studying parallel inference algorithms and systems





I designed and implemented parallel learning algorithms on top of low-level primitives …


ML
Paper
Application
Idea
Model +
Algorithm
Evaluate
Debug
Optimize
Distributed
Prototype
Serial
Prototype
Optimize
Parallel
Prototype
Evaluate
ML
Paper
Debug

Low-Level Primitives

- Shared Memory Parallelism: *pThreads* & *OpenMP*
- Distributed Communication: *Actors* & *MPI*
  - Yes,… I built an RPC framework and then an actor framework
- Hardware APIs + Languages: *GPU Programming (CUDA)*

Low-level parallel primitives offer
fine grained control over parallel execution.

Advantages of the Low-Level Approach

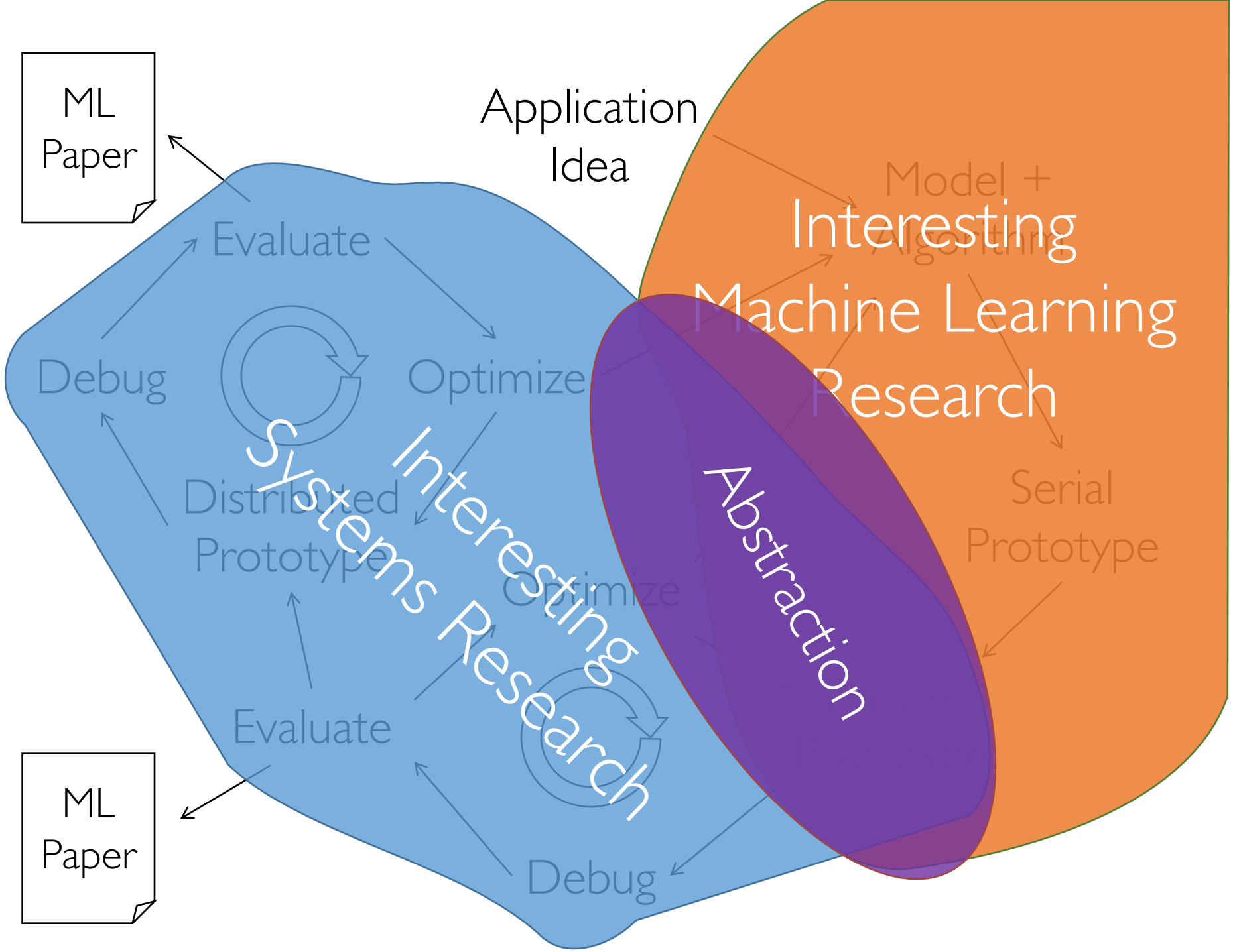
- Extract maximum performance from hardware
- Enable exploration of more complex algorithms
  - Fine grained locking
  - Atomic data-structures
  - Distributed coordination protocols

My implementation is better than your implementation.

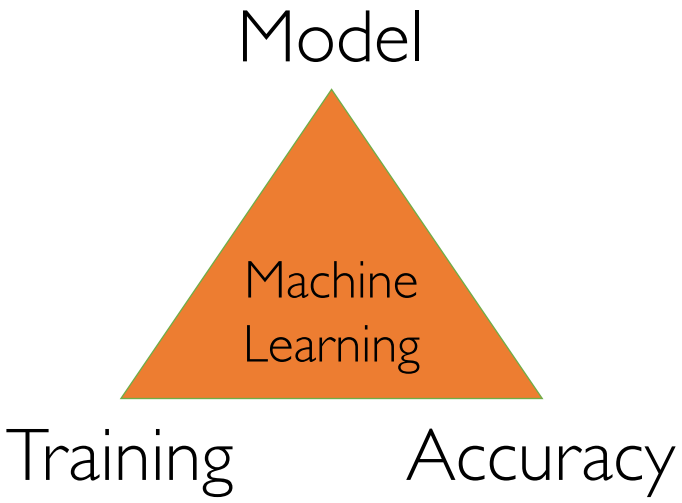
Limitations of the Low-Level Approach

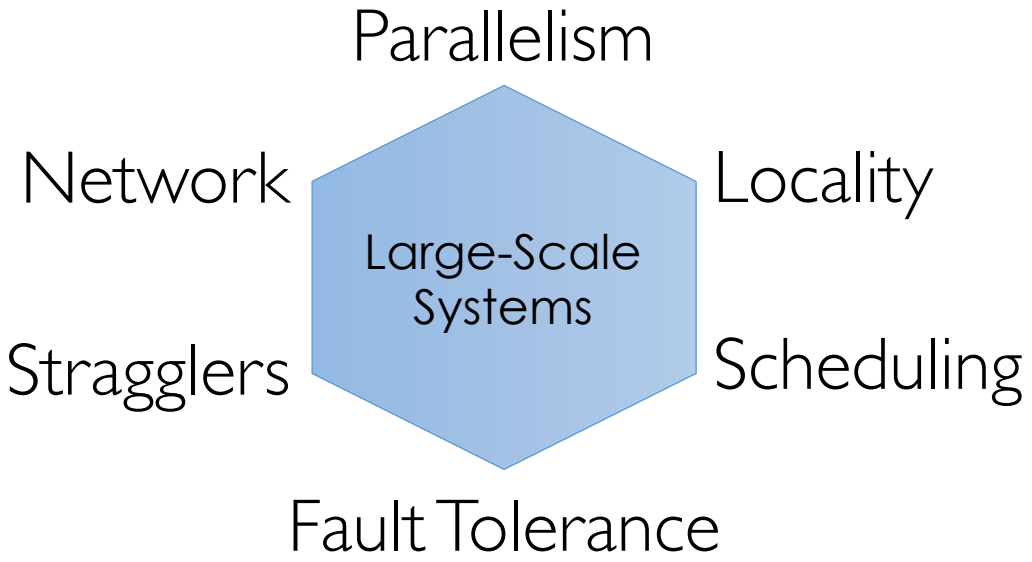
- Repeatedly address the same system challenges
- Algorithm conflates learning and system logic
- Difficult to debug and extend
- Typically does not address issues at scale: hardware failure, stragglers, …

Months of tuning and engineering
for one problem.

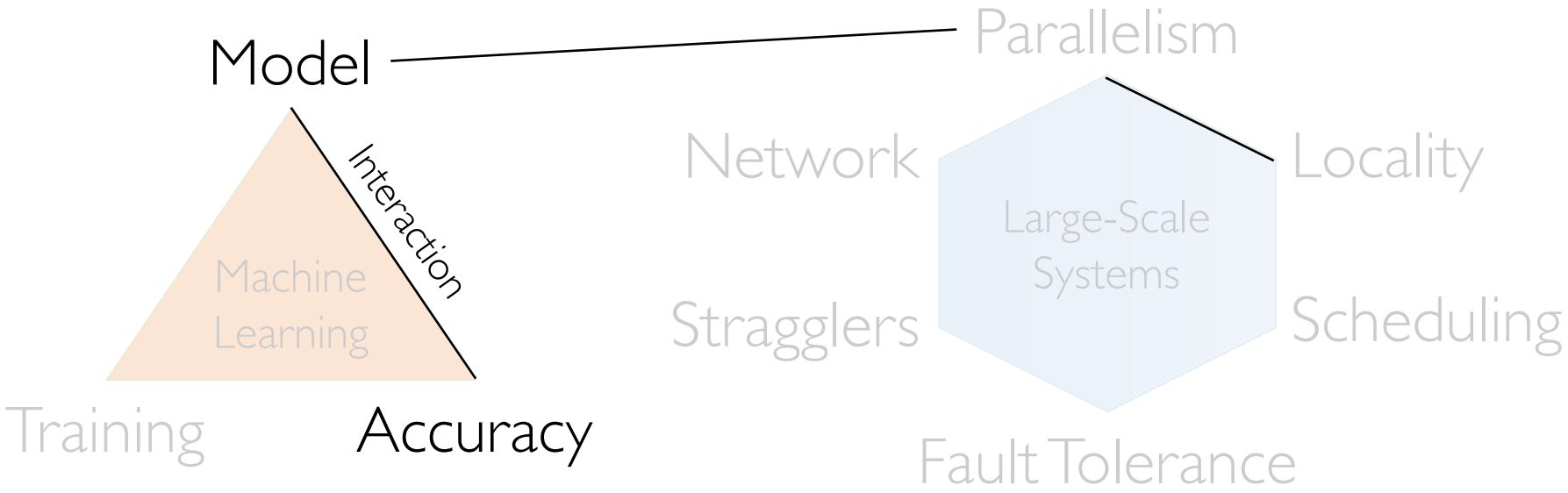

ML
Paper
Application
Idea
Model +
Algorithm
Interesting
Machine Learning
Research
Evaluate
Debug
Optimize
Distributed
Prototype
Interesting
Systems Research
Abstraction
Serial
Prototype
Optimize
Evaluate
ML
Paper
Debug

Design Complexity

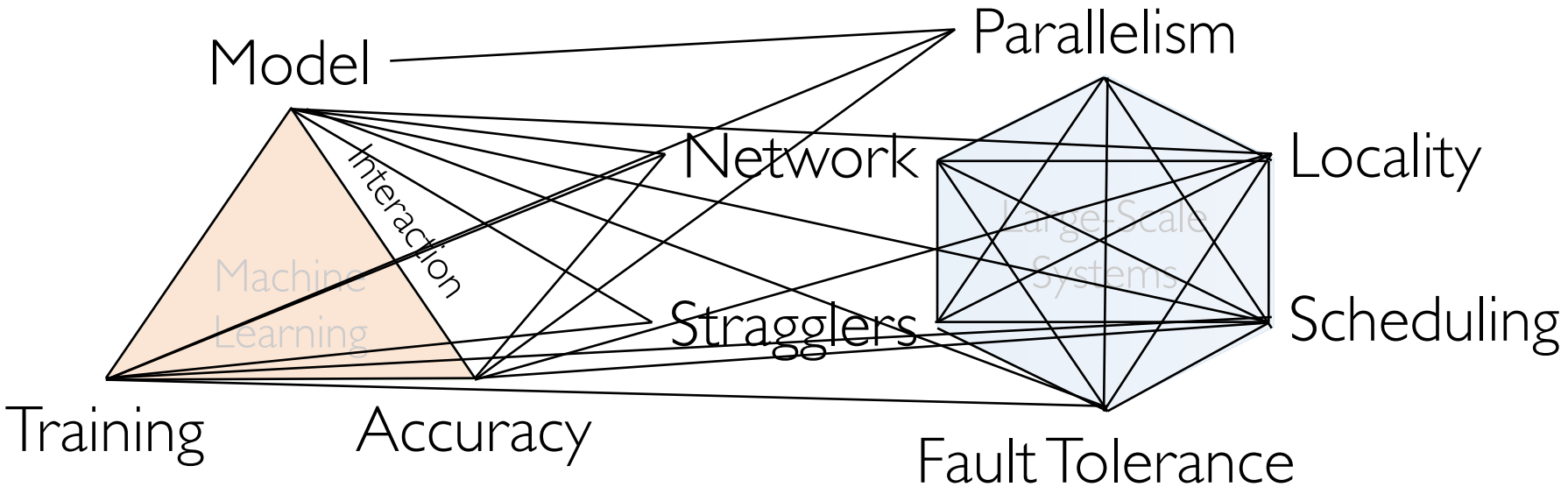




Design Complexity

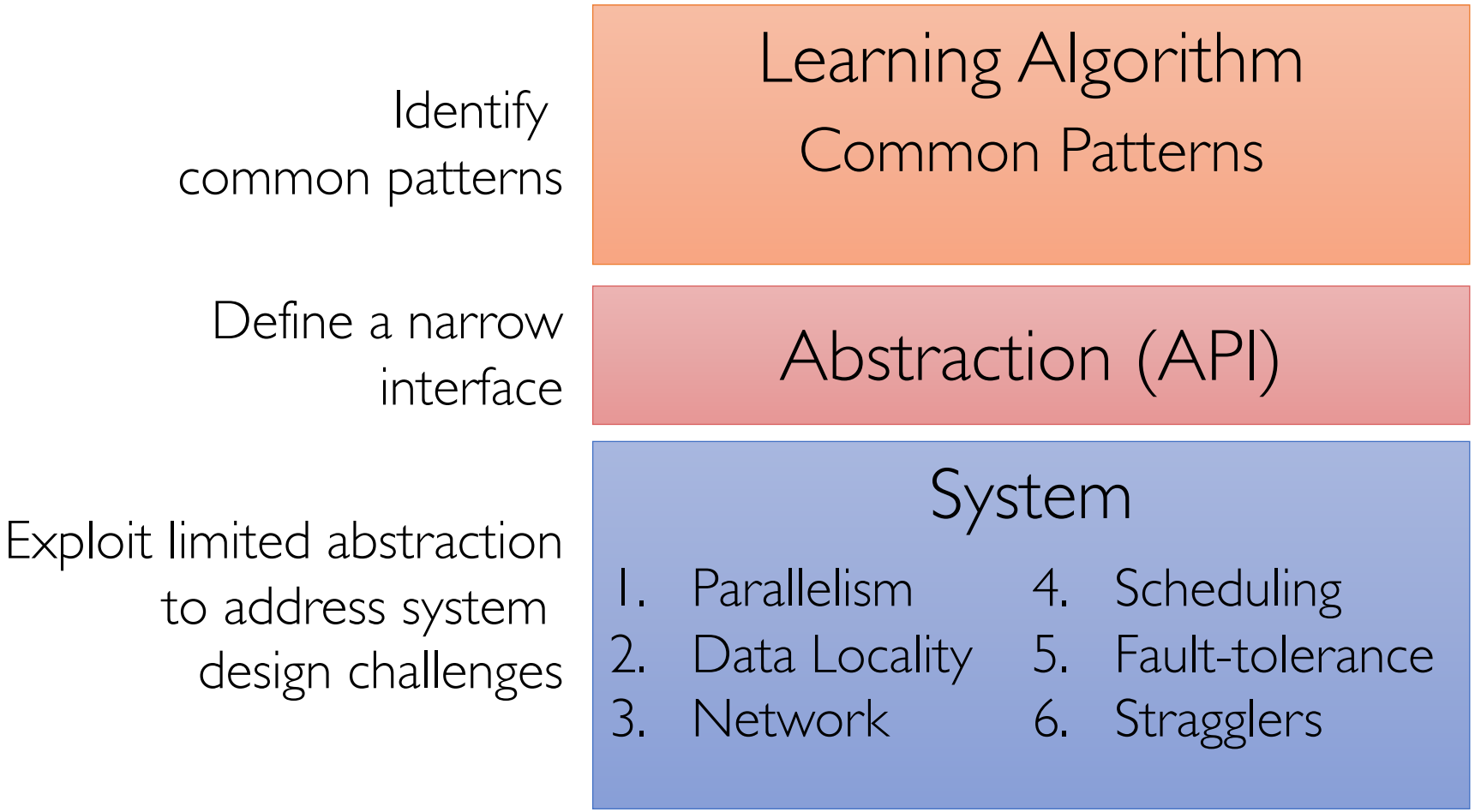


Design Complexity



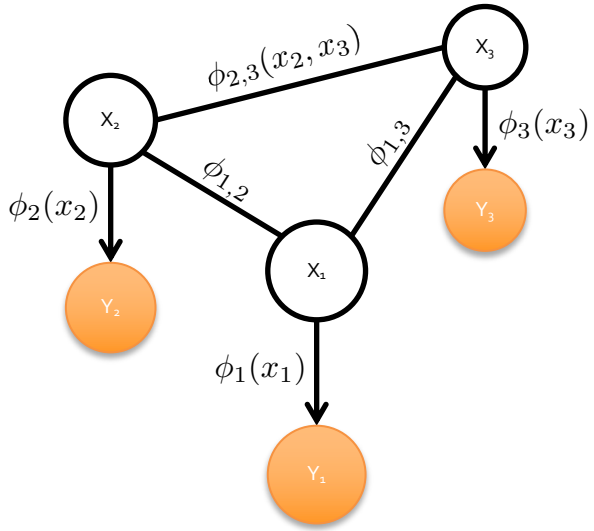
Learning systems combine the complexities
of machine learning with system design

Managing Complexity Through Abstraction

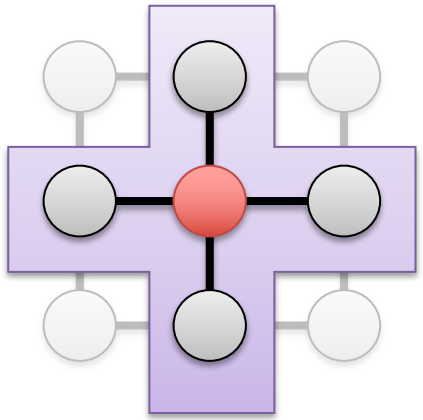


PhD in Machine Learning from CMU 2013

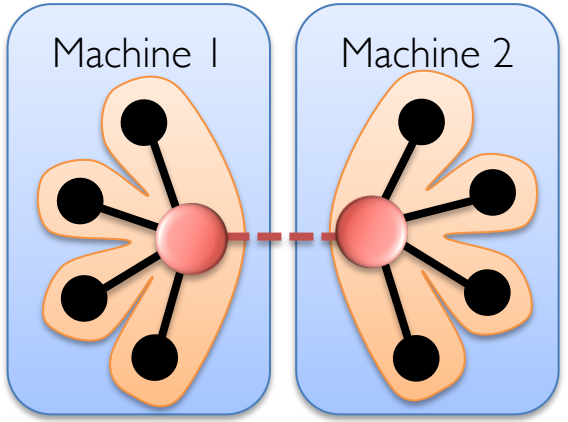
Machine Learning | Abstractions | Scalable Systems



Graphical Model Inference



Vertex Program



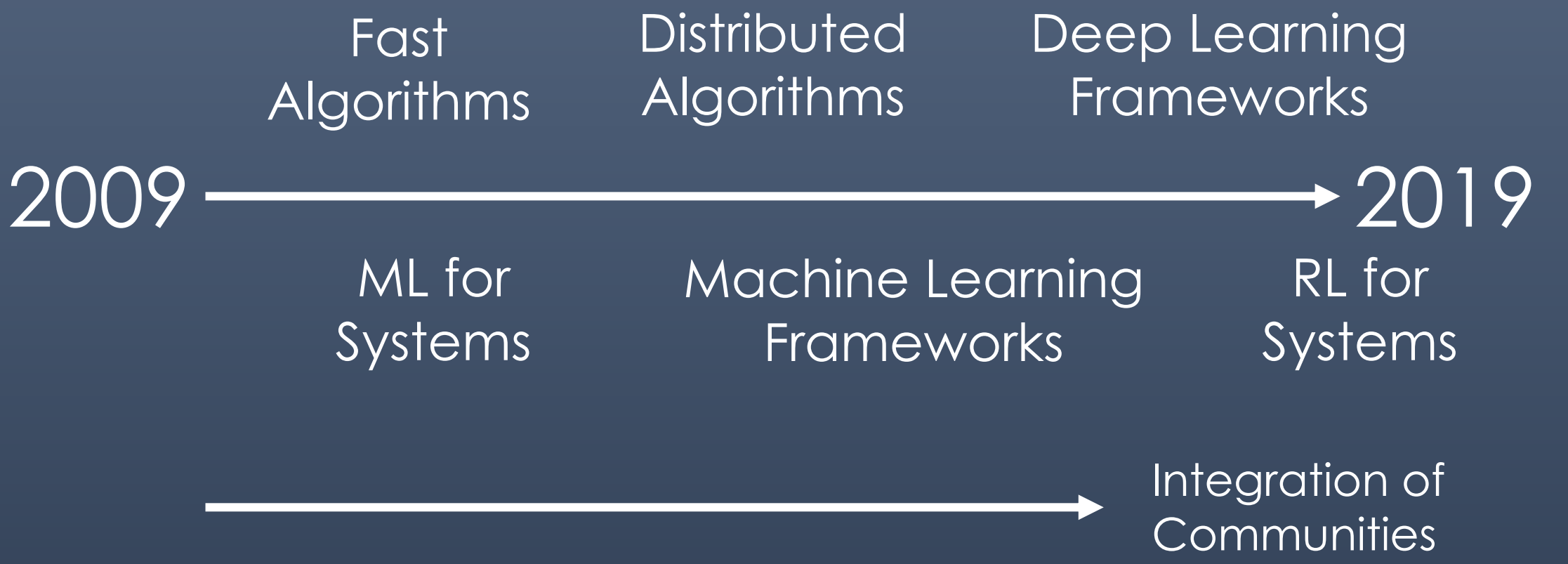
GraphLab/GraphX System

Looking Back on AI Systems

How did the field evolve during the time I was doing my PhD


ML Systems Workshop
ICML'16
OVERVIEW
ACCEPTED PAPERS
CALL FOR PAPERS
IMPORTANT DATES
INVITED SPEAKERS
ORGANIZERS
SCHEDULE
SITEMAP
Overview
The ML Systems Workshop will be held as part of ICML in NYC on June 24, 2016, 8:30am
Location
Microsoft Technology Center
(11 Times Square, 8th avenue between 41st and 42nd streets)
Room Central Park on 6th Floor
Entrance at the intersection of 8th avenue and 41st street
Phone: 212-245-2100
Map/Directions
2 blocks from the Marriot Marquis Hotel, (between 41st and 42nd streets on 8th Avenue)
Overview
This workshop is a follow up to the ICML audience of the well attended Learning Systems workshop at NIPS 2015 and the Software Engineering for Machine Learning workshop at NIPS 2013.
A new area is emerging at the intersection of machine learning (ML) and systems design. This birth is driven by the explosive growth of diverse applications of ML in production, the continued growth in data volume, and the complexity of large-scale learning systems. The goal of this workshop is to bring together experts working at the crossroads of ML, system design and software engineering to explore the challenges faced when building practical large-scale machine learning systems. In particular, we aim to elicit new connections among these diverse fields, identify tools, best practices and design principles. The workshop will cover ML and AI platforms and algorithm toolkits (Caffe, Tensor Flow, Torch etc), as well as dive into Machine learning focused developments in distributed learning platforms, programming languages, data structures and general purpose GPU programming.

Machine learning community has had an evolving focus on AI Systems



Systems and the Third Wave of AI

Waves of AI Research

- **1950 to 1974:** *Birth of AI*
  - 1950 Alan Turing publishes the *Imitation Game*
  - 1951 Marvin Minsky builds first neural network machine (SNARC)
  - DARPA starts to pour money into AI Research (limited oversight)
  - ***Incredible optimism***

"Within ten years a digital computer will be the world's chess champion"
-- Herbert A. Simon and Allen Newell (1958)

"In from three to eight years we will have a machine with the general intelligence of an average human being"
– Marvin Minksy (1970)

Waves of AI Research

1950 to 1974: *Birth of AI*

1974 to 1980: *First AI Winter*

- Technology was unable to meet the high expectations
  - **Insufficient computer power**
  - Over emphasis on combinatorial search
  - Insufficient knowledge (data)
- 1969 Book "Perceptrons" by Minsky and Papert
  - Showed that certain basic functions (e.g., parity) cannot be computed using local connections
  - Strong argument against connectionist approaches to AI (neural networks)
- Government funding dries
  - Not meeting objectives
  - DARPA focuses more on immediate impact

Waves of AI Research

- **1950 to 1974:** *Birth of AI*
- **1974 to 1980:** *First AI Winter*
- **1980 to 1987:** *Second Wave of AI*
  - Expert systems start to solve real-world problems
    - CMU developed XCON (AI for Systems) for DEC → saves $40M annually
    - Combines "knowledge" (expert rules "data") with logic
  - Japanese government invests $850M in AI Research
    - Other governments respond by also investing heavily
  - Emergence of the **AI hardware** and **software industry** to

Waves of AI Research

1950 to 1974: *Birth of AI*

1974 to 1980: *First AI Winter*

1980 to 1987: *Second Wave of AI*

1987 to 1993: *Second AI Winter*

- Expert systems too brittle to maintain
- Collapse of the specialized **AI hardware market**
  - Commodity hardware was improving to rapidly
  - Ultimately over 300 AI companies would vanish from the market
- Massive government funding cuts

Waves of AI Research

- **1950 to 1974:** *Birth of AI*
- **1974 to 1980:** *First AI Winter*
- **1980 to 1987:** *Second Wave of AI*
- **1987 to 1993:** *Second AI Winter*
- **1993 to 2011:** *AI Goes Stealth Mode (aka Machine Learning)*
  - Hardware becomes fast enough, and AI techniques start to work
    - Deep Blue beats Garry Kasparov (1997)
    - OCR, Speech Recognition, Google Search, …
  - Confluence of ideas and techniques: optimization, statistics, probability theory, and information theory

- **1950 to 1974:** *Birth of AI*
- **1974 to 1980:** *First AI Winter*
- **1980 to 1987:** *Second Wave of AI*
- **1987 to 1993:** *Second AI Winter*
- **1993 to 2011:** *AI Goes Stealth Mode (aka Machine Learning)*
  - Hardware becomes fast enough, and AI techniques start to work
    - Deep Blue beats Garry Kasparov (1997)
    - OCR, Speech Recognition, Google Search, …
  - Confluence of ideas and techniques: optimization, statistics, probability theory, and information theory

- **1950 to 1974:** *Birth of AI*
- **1974 to 1980:** *First AI Winter*
- **1980 to 1987:** *Second Wave of AI*
- **1987 to 1993:** *Second AI Winter*
- **1993 to 2011:** *AI Goes Stealth Mode (aka Machine Learning)*
- **2011 to 2019:** *Third Wave (AI Goes Deep)*
  - Large quantities of **data** in conjunction with **advances in hardware** and **software** enable the **design** and **training** of **complex models**
  - New applications emerge
    - Autonomous driving, home automation,
  - AI Market frenzy → IDC predicts 77.6B Market for AI-Systems in 2022
  - …

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

New Forces Driving AI Revolution

Data

Benchmarks

Compute

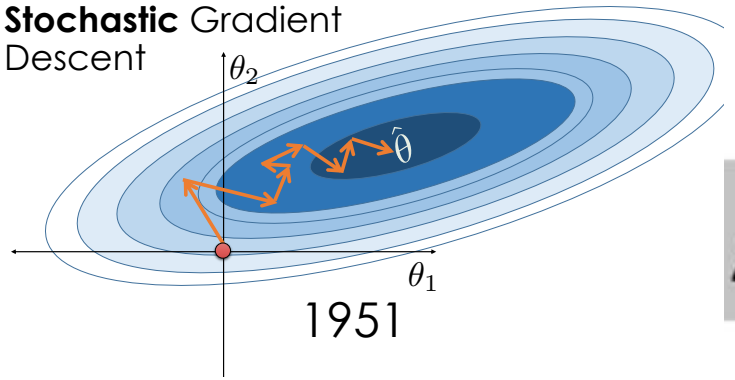
Abstractions

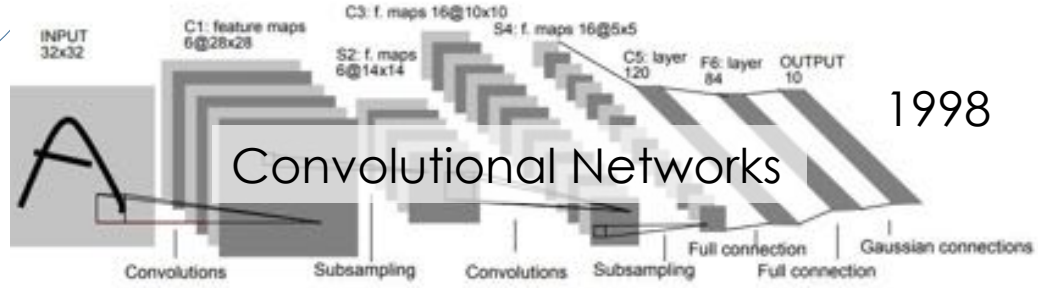
PyTorch

Caffe

TensorFlow

Advances in Algorithms and Models





1998

Convolutional Networks

What is ML/AI Systems Research Today?

What is AI-Systems Research?

- Good AI and Systems research
  - Provides insights to both communities
- Leverages understanding of both domains
  - Examples:
    - Studies tradeoff between statistical and computational efficiency
    - Identify essential abstractions in DNN design
    - Leverages framing of indexes to exploit overfitting
- More than just great software!
- Builds on Big Ideas in AI and Systems Research

Big Ideas in Systems Research

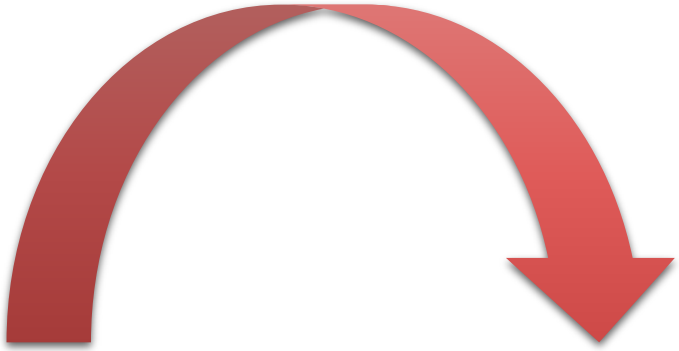
- Problem Framing
  - Identifying the right problem and solution requirements
- Abstraction & Managing Complexity
  - Reducing complex problems into smaller parts
- Tradeoffs
  - Understanding fundamental constraints
- Details: System design and Implementation

Big Ideas in ML Research

- Generalization
  - What is being "learned"?
- Inductive Biases and Representations
  - What assumptions about domain enable efficient learning?
- Efficiency (Data and Computation)
  - How much data and time are needed to learn?
- Details: Objectives/Models/Algorithms

Kinds of AI-Systems Research

Advances in **AI** are being used to address fundamental challenges in **systems**.

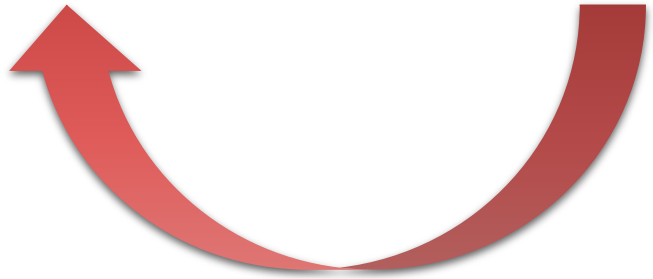


AI + Systems

AI + Systems

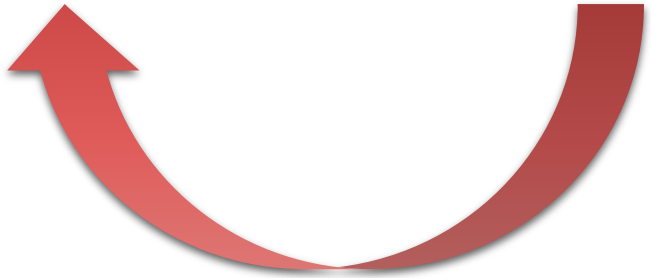
- Reinforcement Learning for
  - Pandas code generation
  - SQL join planning
  - Network packet classification
  - Autoscaling
- Bandit Algorithms for radio link adaptation
- Wireless link quality estimation
- Multi-task learning for straggler mitigation
- VM Selection using Trees ..

AI + Systems



Advances in **systems** are enabling substantial progress in **AI**

AI + Systems



Developing Systems for:

- Autonomous Vehicles
- Reinforcement Learning
- Secure Machine Learning
- Prediction Serving
- Experiment Management

Advancing AI

- Dynamic Neural Nets
- Prediction on Compressed Data
- Distributed Training
- Distributed Auto-ML

Berkeley is the place to be for

AI + Systems

Research

Why?

AI + Systems Research

About You?

Background Requirements

- You have taken **previous courses** in either
  - Systems (for example CS162) → comfortable **building systems** and familiar with **big ideas in system design**
  - Machine Learning (for example CS189) → comfortable **training neural networks** and familiar with **big ideas in ML**

- You are **actively involved in research** in either systems or machine learning (or both)

If this does not describe you, please visit my office hours on **Wednesdays from 4-5 in 773 Soda Hall**.

Goals for the Class

What do you expect from this class?

My Goals for the Class

- Identify and **solve impactful problems** at the intersection of AI and Systems
- Learn about the big ideas and key results in AI systems
- Learn how to read, evaluate, and peer review papers
- Learn how to write papers that get high review scores

How will we achieve these goals?

- **Lectures and Reading:** study developments in AI Systems
  - Identify the key open problems and research opportunities
- **Projects:** collaboration between Systems and AI students
  - Explore open problems and produce top tier publications
- **In Class Reviews:** weekly oral reviews of papers
  - Learn both how to evaluate papers and understand how papers are reviewed

Reasons not to take this class … ☹

If you plan to drop this class, please drop it soon so others can enroll!

- If you want to learn how to **train models** and **use TensorFlow**, **PyTorch**, and **SkLearn**
  - Take: CS C100, CS182, CS189
- If you want learn how to **use big data systems**
  - Take: CS162, CS186
- You are not interested in learning how to **evaluate**, **conduct**, and **communicate** research
  - Why not?
- You are **unable to attend lecture**

Problems

What makes a good problem?

What makes a **good problem?**

- **Impact:** People care about the solution
  - … and progress advances our understanding (**research**)
- **Metrics:** You know when you have succeeded
  - Can you **measure progress** on the solution?
- **Divisible:** The problem can be divided into smaller problems
  - You can identify the first sub-problem.
- **Your Edge:** Why is it a good problem ***for you?***
  - Leverage your strengths and imagine a new path.

Can you Solve a Solved Problem?

- Ideally you want to solve a **new** and **important** problem
- A **new solution** to a solved problem can be impactful if:
  - It supports a **broader set of applications** (users)
  - It **reveals a fundamental trade-off** or
  - Provides a **deeper understanding** of the problem space
  - **10x Better?**
    - Often publishable…
    - Should satisfy one of the three above conditions.

Class Organization

Weekly Topic Organization

I may make changes to topics and format based on class participation.

- Each week we will cover a new topic
  - Big Ideas in ML/Sys
  - ML Life-cycle
  - ML in DBs
  - Model Dev. Frameworks
  - Distributed Training
  - Prediction Serving
  - Autonomous Driving
  - Model Compilation
  - Hardware Acceleration
  - ML → Systems
  - Secure ML
  - Debugging and Interpretability
- There will be **3 required papers** to read each week
- **Monday** I will cover the background for the topic and an overview of the reading for the week
- **Friday** you will participate in an in class "Program Committee Meeting" to discuss the reading

In Class PC Meeting Format (V.0)

For each of the three papers: (30 Minutes per paper)

- **Neutral:** recap of the paper (neutral opinion) [5 Minutes]
- **Advocate:** Strengths of the paper [5 Minutes]
- **Critic:** Weaknesses of the paper [5 Minutes]
- Class will discuss **rebuttal** and **improvements** [10 Minutes]
- Brief in-class vote for acceptance into the AI-Sys prelim

The **Neutral** Presenter will Summarize

- What is the **problem** being solved?
- What was the **solution**? (Summary!)
- What **metrics** did they use to evaluate their solution?
  - What was the **Baselines** of comparison?
- What was the **key insight** or **enabling idea**?
- What are the **claimed technical contributions**?

Advocate and **Critic** Will Discuss

- **Novelty** and **Impact**
  - Are the problem and solution novel and how will the solution affect future research?
- **Technical Qualities**
  - Are the problem **framing** and **assumptions** reasonable
  - Discuss merits of the technical **contributions**
  - Does the **evaluation** support claims and reveal limitations of the proposed approach?
- **Presentation**
  - Discuss the **writing clarity** and **presentation of results**
  - Positioning of **related work**

Simple Grading Policy

20% Class Participation

- **Attend lecture** and **participate** in class discussions
- **Lead 2 to 3 in class** discussions

20% Weekly Reviews

- Submit 3 weekly paper reviews (google form) per week!
- Time consuming but important!

60% Class Projects

- Do great research!

Action Items

- Go to website: https://ucbrise.github.io/cs294-ai-sys-fa19/
- Make sure you are on the course Piazza
  - Needed for announcements

- Signup for **3 discussion** slots as **different roles here**:

https://tinyurl.com/aisysfa19signup