DSC 102
Systems for Scalable Analytics

Winter 2021

Arun Kumar
2009: Bachelors in CSE from IIT Madras, India

2009–16: MS and PhD in CS from UW-Madison
PhD thesis area: Data systems for ML workloads

2016-: Asst. Prof. at UC San Diego CSE
2019-: + Asst. Prof. at UC San Diego HDSI

Summers: 110F!
Winters: –40F!

Ahh! :)
My Current Research

New abstractions, algorithms, and software systems to “democratize” ML-based data analytics from a data management/systems standpoint

Democratization = System Efficiency + Human Efficiency
(Lower resource costs) (Higher productivity)

Practical and scalable data systems for ML analytics
Inspired by relational database systems principles
Exploit insights from learning theory and optimization theory
My Current Research

Data Scientist/ML Engineer

Source → Build → Deploy

ML/AI + Data Systems Infrastructure

Research Approach:

Abstract key steps + Formalize computation + Automate grunt work + Optimize execution

https://adalabucsd.github.io/
What is this course about? Why take it?
1. Netflix’s “spot-on” recommendations

**Stranger Things**

- **NETFLIX ORIGINAL**
- **Rating:** 95% Match
- **Year:** 2017
- **Seasons:** 2
- **Format:** 4K Ultra HD
- **Audio:** 5.1

When a young boy vanishes, a small town uncovers a mystery involving secret experiments, terrifying supernatural forces and one strange little girl.

*Winona Ryder, David Harbour, Matthew Modine*

*TV Shows, TV Sci-Fi & Fantasy, Teen TV Shows*
How does Netflix know that?
Large datasets + Machine learning!

Log all user behavior (views, clicks, pauses, searches, etc.)
Recommender systems apply ML to TBs of data from all users and movies to deliver a tailored experience
2. Structured data with search results

Pradeep Khosla - UC San Diego Office of the Chancellor - University ...
chancellor.ucsd.edu/chancellor-khosla
Pradeep K. Khosla became UC San Diego's eighth Chancellor on August 1, 2012. As UC San Diego's chief executive officer, he leads a campus with more than ...

Pradeep K. Khosla - UC San Diego Office of the Chancellor
chancellor.ucsd.edu/chancellor-khosla/khosla-biography
Chancellor, University of California San Diego. Pradeep K. Khosla, an internationally renowned electrical and computer engineer, is the eighth Chancellor of the ...

Pradeep Khosla - Wikipedia
https://en.wikipedia.org/wiki/Pradeep_Khosla
Pradeep K. Khosla is an academic computer scientist and university administrator. He is the current chancellor of the University of California, San Diego. He was ...

Pradeep Khosla | LinkedIn
https://www.linkedin.com/in/pradeepkhosla
Greater San Diego Area - Chancellor, UC San Diego - Avigilon
View Pradeep Khosla's professional profile on LinkedIn. LinkedIn is the world's largest business network, helping professionals like Pradeep Khosla discover ...

Robotics Institute: Pradeep Khosla
www.ri.cmu.edu/people

Pradeep Khosla
Chancellor of the University of California, San Diego

Pradeep K. Khosla is an academic computer scientist and university administrator. He is the current chancellor of the University of California, San Diego. Wikipedia

Born: March 13, 1957 (age 60 years), Mumbai, India
Spouse: Thespine Kavoulakis
Education: Indian Institute of Technology Kharagpur, Carnegie Mellon University
Residence: Audrey Geisel University House, La Jolla, CA
How does Google know that?
Large datasets + Machine learning!

Knowledge Base Construction (KBC) process extracts tabular/relational data from large amounts of text data
3. AlphaGo defeats human champion!
How did AlphaGo achieve that?
Breakthrough powered by deep learning!

Architecture of AlphaGo

Deep CNNs to visually process board status in plays

Innumerable “enterprise” applications
FUTURE FARMS
small and smart

SURVEY DRONES
Aerial drones survey the fields, mapping weeds, yield and soil variation. This enables precise application of crops, mapping spread of pests, weed blackgrass could increasing Wheat yield by 3.5%.

FARMING DATA
The farm generates vast quantities of rich and varied data. This is stored in the cloud. Data can be used to optimize crop yields, inputs and machinery. Your farm will now be able to carry out farm-mapping saving on average £3,500 per farm per year.

TEXTING COWS
Sensors attached to livestock allowing monitoring of animal health and wellbeing. They can send alerts to alert farmers when a cow goes into labour or develops infection increasing herd survival and increasing milk yield by 10%.

FLEET OF AGRIBOTS
A herd of specialized agribots tend to crops, weeding, fertilizing and harvesting. Robots capable of innovation applications of farming, reduce fertilizer cost by 99.5%.

SMART TRACTORS
50% controlled steering and optimal route planning reduces soil erosion, saving fuel costs by 30%.
“Domain sciences” and healthcare tech are also becoming data+ML intensive
Software systems for data analytics and ML over large and complex datasets are now critical for digital applications in many domains.
Drowning In Big Data - Finding Insight In A Digital Flood

For roughly a decade, there has been a steady stream of information about Big Data. The IDC predicts the industry will expand from $2 billion in 2010 to $43 billion by 2018. What this means is that as companies produce more data, they need to find ways to make sense of it.

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren’t seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, “It was like arriving at a conference reception and realizing you don’t know anyone. So you just stand in the corner sipping your drink—
Data data everywhere,
All the wallets did shrink!
Data data everywhere,
Nor any moment to think?
DSC 102 will get you thinking about the
fundamentals of scalable analytics systems

1. "Systems": What resources does a computer have? How to store and efficiently compute over large data? What is cloud?

2. "Scalability": How to scale and parallelize data-intensive computations?

3. For “Analytics”:
   3.1. Source: Data acquisition & preparation for ML
   3.2. Build: Model selection & deep learning systems
   3.3. Deploying ML models

4. Hands-on experience with scalable analytics tools
The Lifecycle of ML-based Analytics

Data acquisition
Data preparation

Feature Engineering
Training & Inference
Model Selection

Serving
Monitoring

ML/AI + Data Systems Infrastructure

python™, learn, R, TensorFlow, PyTorch, DASK, Spark, AWS
ML Systems

**Q: What is a Machine Learning (ML) System?**

- A data processing system (aka *data system*) for mathematically advanced data analysis operations (inferential or predictive):
  - Statistical analysis; ML, deep learning (DL); data mining (domain-specific applied ML + feature eng.)
  - *High-level APIs* to express ML computations over (large) datasets
  - *Execution engine* to run ML computations efficiently
Categorizing ML Systems

- **Orthogonal Dimensions of Categorization:**
  1. **Scalability:** In-memory libraries vs Scalable ML system (works on larger-than-memory datasets)
  2. **Target Workloads:** General ML library vs Decision tree-oriented vs Deep learning, etc.
  3. **Implementation Reuse:** Layered on top of scalable data system vs Custom from-scratch framework
# Major Existing ML Systems

## General ML libraries:

**In-memory:**
- mlpack
- R
- WEKA
- S.
- SAS
- DASK

**Disk-based files:**
- MADlib
- Apache Spark
- MLlib

**Layered on RDBMS/Spark:**
- Azure Machine Learning
- Amazon SageMaker

**Cloud-native:**
- DataRobot
- H2O.ai

**“AutoML” platforms:**
- Daml
- TensorFlow
- PyTorch

## Decision tree-oriented:
- XGBoost
- LightGBM

## Deep learning-oriented:
Data Systems Concerns in ML

Key concerns in ML:

- Accuracy
- Runtime efficiency (sometimes)

Additional key *practical* concerns in ML Systems:

- Scalability (and efficiency at scale)
- Usability
- Manageability
- Developability

Q: How do “ML Systems” relate to ML?

ML Systems : ML :: Computer Systems : TCS

Long-standing concerns in the DB systems world!

Q: What if the dataset is larger than single-node RAM?

Q: How are the features and models configured?

Q: How does it fit within production systems and workflows?

Q: How to simplify the implementation of such systems?
### Conceptual System Stack Analogy

<table>
<thead>
<tr>
<th>Relational DB Systems</th>
<th>ML Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Theory</strong></td>
<td></td>
</tr>
<tr>
<td>First-Order Logic</td>
<td>Learning Theory</td>
</tr>
<tr>
<td>Complexity Theory</td>
<td>Optimization Theory</td>
</tr>
<tr>
<td><strong>Program Formalism</strong></td>
<td></td>
</tr>
<tr>
<td>Relational Algebra</td>
<td>Matrix Algebra</td>
</tr>
<tr>
<td><strong>Program Specification</strong></td>
<td>SQL</td>
</tr>
<tr>
<td><strong>Program Modification</strong></td>
<td>Query Optimization</td>
</tr>
<tr>
<td><strong>Execution Primitives</strong></td>
<td>Parallel Relational Operator Dataflows</td>
</tr>
<tr>
<td><strong>Hardware</strong></td>
<td>Depends on ML Algorithm</td>
</tr>
<tr>
<td>CPU, GPU, FPGA, NVM, RDMA, etc.</td>
<td></td>
</tr>
</tbody>
</table>
Real-World ML: Pareto Surfaces

Q: Suppose you are given ad click-through prediction models A, B, C, and D with accuracies of 95%, 85%, 90%, and 85%, respectively. Which one will you pick?

Q: What about now?

- Real-world ML users must grapple with multi-dimensional Pareto surfaces: accuracy, monetary cost, training time, scalability, inference latency, tool availability, interpretability, fairness, etc.

- Multi-objective optimization criteria set by application needs / business policies.
Learning Outcomes of this course

- Explain the basic principles of the memory hierarchy, parallelism paradigms, scalable data systems, cloud computing, and containerization.
- Identify the abstract data access patterns of, and opportunities for parallelism and efficiency gains in, data processing and ML algorithms at scale.
- Outline how to use cluster and cloud services, dataflow (“Big Data”) programming with MapReduce and Spark, and deep learning inference with TensorFlow and Keras.
- Apply the above programming skills to create end-to-end pipelines for data preparation, feature engineering, and model selection on large-scale datasets.
- Reason critically about practical tradeoffs between accuracy, efficiency, scalability, usability, and total cost.
What this course is NOT about

❖ NOT a course on databases, relational model, or SQL
  ❖ Take DSC 100 instead (pre-requisite)
❖ NOT a course on internal details of RDBMSs
  ❖ Take CSE 132C instead
❖ NOT a training module for how to use Spark
❖ NOT a course on ML or data mining *algorithmics*; instead, we focus on ML *systems*
Now for the (boring) logistics ...
Prerequisites

❖ **DSC 100** (or equivalent) is necessary
❖ Transitively **DSC 80**; a mainstream ML algorithmics course is necessary
❖ Proficiency in Python programming
❖ For all other cases, email the instructor with proper justification; a waiver can be considered

http://cseweb.ucsd.edu/~arunkk/dsc102_winter21
Components and Grading

❖ 3 Programming Assignments: 35% (7% + 14% + 14%)
  ❖ No late days! Plan your work well ahead.
❖ 4 Quizzes: 20%; time limit: 25min
❖ Midterm Exam: 15%
  ❖ Tue, Feb 9; time limit: 80min + 20min grace time
❖ Final Exam: 25% (cumulative)
  ❖ Thu, Mar 18; time limit: 180min + 30min grace time
❖ Time window for all quizzes/exams: 00:01am—11:59pm PT
❖ Peer Evaluation Activities: 5%
❖ All quizzes/exams/peer submissions will be via Canvas

http://cseweb.ucsd.edu/~arunkk/dsc102_winter21
Hybrid of relative and absolute; grade is better of the two

<table>
<thead>
<tr>
<th>Grade</th>
<th>Relative Bin (Use strictest)</th>
<th>Absolute Cutoff (&gt;=)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A+</td>
<td>Highest 5%</td>
<td>95</td>
</tr>
<tr>
<td>A</td>
<td>Next 10% (5-15)</td>
<td>90</td>
</tr>
<tr>
<td>A-</td>
<td>Next 15% (15-30)</td>
<td>85</td>
</tr>
<tr>
<td>B+</td>
<td>Next 15% (30-45)</td>
<td>80</td>
</tr>
<tr>
<td>B</td>
<td>Next 15% (45-60)</td>
<td>75</td>
</tr>
<tr>
<td>B-</td>
<td>Next 15% (60-75)</td>
<td>70</td>
</tr>
<tr>
<td>C+</td>
<td>Next 5% (75-80)</td>
<td>65</td>
</tr>
<tr>
<td>C</td>
<td>Next 5% (80-85)</td>
<td>60</td>
</tr>
<tr>
<td>C-</td>
<td>Next 5% (85-90)</td>
<td>55</td>
</tr>
<tr>
<td>D</td>
<td>Next 5% (90-95)</td>
<td>50</td>
</tr>
<tr>
<td>F</td>
<td>Lowest 5%</td>
<td>&lt; 50</td>
</tr>
</tbody>
</table>

Example: Score 82 but 33%ile; Rel.: B-; Abs.: B+; so, B+
<table>
<thead>
<tr>
<th>Week</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-3</td>
<td>Introduction; Basics of Computer Org. and Operating Systems</td>
</tr>
<tr>
<td>3</td>
<td>Basics of Cloud Computing</td>
</tr>
<tr>
<td>4-5</td>
<td>Parallel and Scalable Data Processing</td>
</tr>
<tr>
<td>6</td>
<td>Midterm Exam on Tue, Feb 9</td>
</tr>
<tr>
<td>6-7</td>
<td>Dataflow Systems</td>
</tr>
<tr>
<td></td>
<td>ML Data Preparation and Model Selection</td>
</tr>
<tr>
<td></td>
<td>Deep Learning Systems</td>
</tr>
<tr>
<td>9</td>
<td>ML Deployment</td>
</tr>
<tr>
<td>10</td>
<td>Optional: Open Research Questions; Review</td>
</tr>
<tr>
<td>11</td>
<td>Final Exam on Wed, Mar 17</td>
</tr>
</tbody>
</table>

There will likely be 2 industry guest lectures; speakers TBD.
## Tentative Schedule for Prog. Assignments

<table>
<thead>
<tr>
<th>Date</th>
<th>Agenda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 13</td>
<td>PA 0 released</td>
</tr>
<tr>
<td>TBD</td>
<td>Discussion on PA 0 by TA</td>
</tr>
<tr>
<td>Jan 20</td>
<td><strong>PA 0 due</strong></td>
</tr>
<tr>
<td>Jan 21</td>
<td>PA 1 released</td>
</tr>
<tr>
<td>TBD</td>
<td>Discussion on PA 1 by TA</td>
</tr>
<tr>
<td>Feb 15</td>
<td><strong>PA 1 due</strong></td>
</tr>
<tr>
<td>Feb 16</td>
<td>PA 2 released</td>
</tr>
<tr>
<td>TBD</td>
<td>Discussion on PA 2 by TA</td>
</tr>
<tr>
<td>Mar 9</td>
<td><strong>PA 2 due</strong></td>
</tr>
</tbody>
</table>

**No late days!** Plan your work upfront! Try to join the Discussion slot talks by the TAs.
Suggested Textbooks

Aka “CompOrg Book”

Aka “Comet Book”

Aka “Cow Book”

Aka “Spark Book”

Aka “MLSys Book”

(Free PDFs available online; also check out our library)
1. Computer systems are about carefully layering levels of abstraction.

2. Analytics/ML Systems is a recent/emerging area of research.

3. Also, DSC 102 is the first UG course of its kind in the world!
Online-Only Modality Logistics

❖ 2 key tools: Canvas + Zoom
❖ Canvas is the one-stop shop for course announcements, meeting links, quizzes, exams, and submissions.
❖ Lecture recordings will be posted on Canvas Media Gallery.
❖ Use Canvas Discussions for doubts/questions. Help out your peers by answering/discussing.
❖ Zoom is for all live lectures, discussions, and office hours. Do NOT make Zoom links public!

http://cseweb.ucsd.edu/~arunkk/dsc102_winter21
Course Administrivia

❖ Lectures: TueThu 11:00am-12:20pm PT on Zoom
  ❖ Will take live Q&A throughout
  ❖ Videos available for asynchronous viewing too
❖ Instructor: Arun Kumar; arunkk [at] eng.ucsd.edu
❖ Office hours: Thu 1:00-2:00pm PT on Zoom
❖ Canvas Announcements for all Zoom meeting links, logistical announcements, quizzes/exams, and gradebook
❖ TAs: Side Li, Umesh Singla, and Subrato Chakravorty (see webpage for details)

http://cseweb.ucsd.edu/~arunkk/dsc102_winter21
General Dos and Do NOTs

**Do:**
- Follow all announcements on Canvas
- Try to join the synchronous lectures/discussions
- View/review videos asynchronously by yourself
- Participate in discussions in class / on Canvas
- Raise your hand before speaking on Zoom
- Use “DSC102:” as subject prefix for all emails to me/TA

**Do NOT:**
- Record Zoom sessions on your side without permission of instructor and other students
- Harass, intimidate, or intentionally talk over others
- Violate academic integrity on the graded quizzes, exams, PAs, or peer submissions; I am very strict on this matter!
Please submit this Google Form related to the online-only logistics ASAP!

https://forms.gle/r7rMcamNThLjQegh9

http://cseweb.ucsd.edu/~arunkk/dsc102_winter21