CS 744: SUMMARY

Shivaram Venkataraman
Fall 2020

Quick Poll on Papers! https://forms.gle/xuTEEQj9B5m5uMn8
• Midterm 2 on Thursday!

• Final Project presentations next week! Signup?

• Final report due Dec 17\textsuperscript{th}

• AEFIS Course feedback form
OUTLINE

Fairness in ML

Survey results, Discussion

Big data systems: Looking forward
The UK used a formula to predict students’ scores for canceled exams. Guess who did well.

The formula predicted rich kids would do better than poor kids who’d earned the same grades in class.

By Kelsey Piper | Aug 22, 2020, 7:30am EDT
ML TRAINING LOOP

State of the world → Measurement → Data → Learning → Model → Feedback

Individuals

Action
MEASUREMENT

Why is this hard? E.g., measuring demographics over time

Defining a target variable
“credit-worthiness”
ImageNet class names from WordNet

http://ludo.mit.edu/~ludo/labeling_ui.html
LEARNING

Learning: Data $\rightarrow$ Models
Calibrates to training data

Sample size disparity

18 Translating from English to Turkish, then back to English injects gender stereotypes.
ACTION – FEEDBACK LOOP

ML reveals correlations, but often used as if causation!
Prediction affects outcome
   Traffic congestion

ML Feedback loop
   Search engine sort by pages linked more often
   More user clicks $\rightarrow$ more often linked to
   Feedback loop: Rank more highly
WHAT CAN WE DO

Toy Example of Hiring

Use ML to make predictions based on GPA, interview score
Predict ”job performance” based on that
What could go wrong?

Intervention

Include diversity criterion in objective function
CHALLENGES AND OPPORTUNITIES

Limitations on what we can measure: unbiased measurements infeasible

Data-driven decision-making potential to be more transparent
   Need for explainable ML models

New research shows effective interventions (read rest of the book?)
SURVEY RESULTS
LEARNING OBJECTIVES

At the end of the course you will be able to

• Explain the design and architecture of big data systems
• Compare, contrast and evaluate research papers
• Develop and deploy applications on existing frameworks
• Design, articulate and report new research ideas
DISCUSSION

https://forms.gle/KlxxsiTUqmvfNvY6
What is one application that you have used or worked on that could have similar issues to ones described in the chapter?
What were some of your goals when you started the course? (Think about the first survey.) Reflect on what part of your goals have been achieved and how.
What are some other trends you have noticed across the papers in the class? (e.g., specialization vs unification) Or what are some commonalities across papers/topics?
LOOKING FORWARD
NEXT-GENERATION BIG DATA SYSTEMS?

- Workloads
- Data Processing Systems
- Hardware
TRENDS IN WORKLOADS

New functionalities
  Data science / AI
  Robotics

New data sources
  Bio-medical data
  Video streams
  IoT / edge devices
WHAT CAN SYSTEMS RESEARCH DO?

More than performance?
  Latency, throughput, efficiency
  Ease of use

Some other goals to consider?
  Security, Privacy
  Robustness
  Data bias / ethics
Large scale data analysis has changed the world
COURSE SUMMARY

Applications

- Machine Learning
- SQL
- Streaming
- Graph

Computational Engines

Scalable Storage Systems

Resource Management

Your System Here?