CS 744: SUMMARY

Shivaram Venkataraman Fall 2020

Quick Poll on Papers! https://forms.gle/xuTEEQjd9B5m5uMn8

ADMINISTRIVIA

- Midterm 2 on Thursday!
- Final Project presentations next week! Signup?
- Final report due Dec 17th
- AEFIS Course feedback form

Big Data Landscape 2016 (Version 3.0) Infrastructure Analytics **Applications** Analyst Data Science Human Hadoop Hadoop in Analytics Sales & Marketing Customer Service Spark **Cluster Services** Visualization Legal On-Premise the Cloud Platforms Platforms Platforms RADIUS' Gainsight' Capital MEDALLIA amazon 🔆 + a b | e a ı context relevant databricks RAVEL **Q** Palantir Microsoft Soloomreach Zeta ğıld CONTINUUM To DataRobot ATTENJITY 🐵 LEVERSTRING LIVERYRE Hortonwork O Google Cloud Platform JUDICATA AYASDI quavus CLARABRIDGE MAPR Pivotal Qlik @ looker blue yonder Lattice ₩ CLICKFOX GridGain 4 **d**ockei IBM InfoSphere Connectifi Datameer Quid enigma MODE plotly Roamb @kahuna √infer SAILTHRU Everlav STELLAService MESOSPHERE textic IBM InfoSphere SISENSE TOOMDATA **TACHYON** Core OS pepperdata. Digital Reasoning IG&DATA ... Preact Bottlenose. persado AVISO Ósense ODMINO Sense Brevia entelo 🐧 altiscale 🔲 bole bluedata jethro Stack IQ ORBITAL INSIGHT inter ana ŷhat ▲ ALGORITHMIA CHARTIO QUANTIFIND ACTIONIC **DigitalGenius** hi□ PREM®NITION fuse|machines #NGAGIO appuri Wiseio Social **NoSQL Databases** NewSQL Databases BI Platforms Statistical Log Analytics Ad Optimization Security Vertical AI Computing Analytics Power BI wamazon amazon O Google Cloud Platform Clustrix Pivotal splunk> Hootsuite □ Y L ∧ N □ E Applications AppNexus MediaMath **S**sas **ORACLE** paradigm4 sumologic CounterTack cybereason Microsoft Azure MarkLogic NETB^SE criteo. Threat Metrix. **DATASIFT** kıbana **splice ≉** birst SPSS OpenX ∞rocketfuel mongoDB DATASTAX AREA 1 GoodData SentinelOne tracx bitly Recorded Future Guardian CLOUD PHYSICS Integral () the Trade Desk Clara citusdata Couchbase platfora ♠ MATLAB **synthesio** | Adgorithms dstillery deepdb Trafodion Cockroach LABS SequoiaDB redislabs @ influxdata loggly **KASIST** (1) atscale **TFORT**SCALE *sift science TAPAD Data \mathfrak{X} u **Oppi**er $M \bigcirc A$ SICNIFY: θedzaί & SICNIFY: Graph MPP Data Speech & NLP Horizontal AI Cloud EDW Real-Time Machine Learning Databases Databases Transformation Integration IBM Watson amazon Publisher Govt / Regulation Finance amazon informatica sentient alteryx **⊿ffirm iiiLending**Club @neo4j Google Coud Tools T METAMARKETS H₂O.0 NUANCE W Socrata **VERTIC**↑ vicarious Microsoft Azure talend Outbrain OnDeck> ...Kreditech striim! MuleSoft Dato NETEZZA Pivotal (7) OPENGOV TRIFACTA ngro 🕞 🕰 Numenta Tab99la zest finance LendÚp 💔 Kabbage snapLogic confluen **Oction** tamr 🙀 🎤 quantcast FN FiscalNote tidemark. 📆 INSIKT **BedrockData** Descartes clarifai OrientDB DATATORRENT StreamSets Chartbeat 🗷 UOra' 👼 Dataminr' 矝 Lenddo EXASOL Odremio dataArtisans PREDPOL Infoworks △ Alation xplenty IDIBON (%) yieldbot 🖒 mark43 **ISENTIUM** Management Security Storage App Dev Crowd-Search Data Services For Business Web / Mobile Quantopian sentient Yieldmo / Commerce / Monitoring TANIUM" sourcing Analysts UO OPERA apigee X illumio Google Analytics New Relic. 35 EXALEAD **Origami**Logio Life Sciences Industries Education/ CODE42 mixpanel Aicrosoft Azu Lucidworks amazon Octifio PATHWAY GENOMICS CASK Keen K **KEXL** Learning OP@WER eHarmony ClearStory DataGravity panasas/ RJMetrics DLUECO Counsul elastic <a>ThoughtSpot ■ RetailNext KNEWTON splunk> X Recombine CipherCloud' nimblesto **-** Typesafe AMPLITUDE 69 granify CrowdFlower DATA SCIENCE CIRRO **S**VECTR∧ M∧∧N∧ 💋 swiftype KYRUUS FLATIRON STITCH FIX 0 Clever соно sum All Airtable oe⊘ezymergen HealthTap® retention custora TACHYUS Seeo FarmLogs Trocana DRIVEN Qumulo import (io @eclara Cross-Infrastructure/Analytics PANORAMA HowGood celect C MACHINE Ginger.io * transcriptic Glow @enlitic DAiCure 🗘 🎎 statmuse B@XEVER mazon Google 🖁 Microsoft 🏗 🐼 SAS 🕍 🕼 🗥 🚾 VERTICA VMWATE'T IBC 🗸 TERADATA ORACLE 👖 NetApp **Open Source** Stat Tools Framework Query / Data Flow Data Access Coordination Real-Time Machine Learning 1 Search Security HEASE mongoDP Apache SINGA MADlib. 🍣 storm Spark Apache Range talend oggang cassandra YARN A MESOS ScalaLab Apache Zookeeper Caffe | CNTK TensorFlo Visualization CouchDB **riak | PPENTS Flink Solr SLAMDATA APACHE Spark VELES WEKA DIMSUM Jupyter DL4J Flink QCDAP Apache Ambari ** TACHYON - druid Data Sources & APIs Incubators & Schools Location / People / Entities Health IOT Financial & Economic Data Air / Space / Sea Other Bloomberg D | DOW JONES PLURALSIGHT IAWBONE GARMIN **△** spire acxiem Experian Epsilon InsideView ThingWorx Y-DLEE PREMISE CAPITAL IQ DataCamp INSIGHT GARMIN. COURSEUGIE WINDWARD 🏥 STREETLINE 🍓 esri practice fusion : fitbit helium samsara A DataElite panjiva Withings VALIDIC netatmo Crimson Hexagon CARTODB factual. quandl xignite CBINSIGHTS ★ The Data Incubator ★ METIS ★ M Human API SteckTwits @estimize PLAID Airware Tone Deploy CIRCULATE placemeter BASIS Sense ■ ĐATA.GOV kinsa Last Updated 3/23/2016 © Matt Turck (@mattturck), Jim Hao (@ijmrhao), & FirstMark Capital (@firstmarkcap) FIRSTMARK

Applications

Machine Learning

SQL

Streaming

Graph





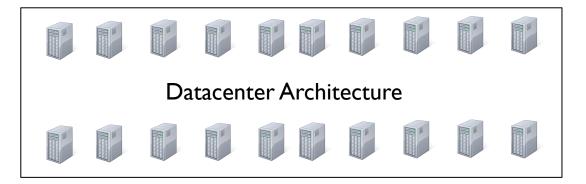




Computational Engines

Scalable Storage Systems

Resource Management





Open Compute Project

OUTLINE

Fairness in ML

Survey results, Discussion

Big data systems: Looking forward

JASON TASHEA OPINION 04.17.17 07:00 AM

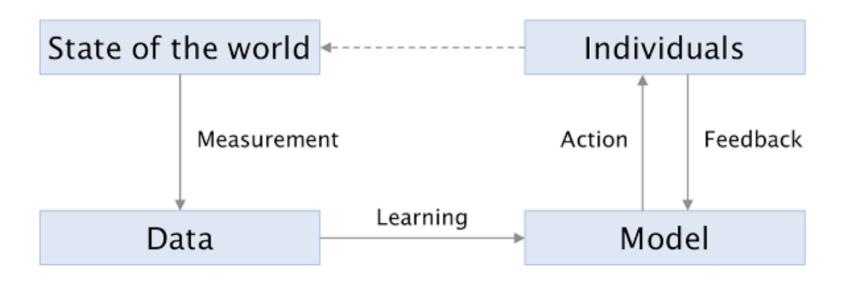
COURTS ARE USING AI TO SENTENCE CRIMINALS. THAT MUST STOP NOW



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.

ML TRAINING LOOP



MEASUREMENT

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

Defining a target variable

"credit-worthiness"

ImageNet class names from WordNet

person

ballplayer, baseball player

groom, bridegroom

scuba diver

http://ludo.mit.edu/~ludo/labeling ui.html

LEARNING

Learning: Data → Models
Calibrates to training data

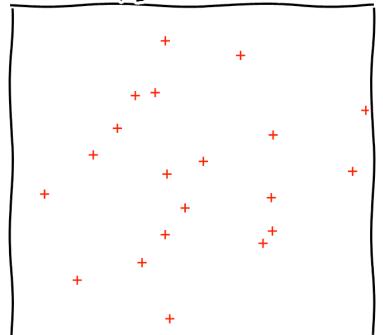
Sample size disparity

¹⁸ Translating from English to Turkish, then back to English injects gender stereotypes.**

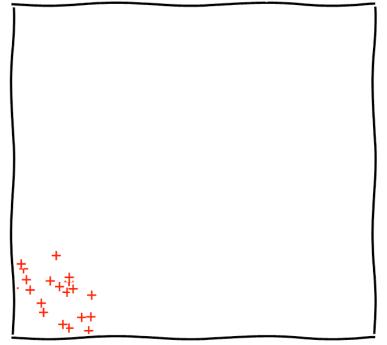


ML ERROR

RANDOM ERRORS



SYSTEMATIC ERRORS



From https://fairmlclass.github.io/

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 → 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

Paper Review

Discussion

Assignment

Project

DISCUSSION

https://forms.gle/KIxsiTUiQqnvfNrY6

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 / Al

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

COURSE SUMMARY

Large scale data analysis has changed the world







COURSE SUMMARY

Applications

Machine Learning SQL Streaming Graph

Your System Here?

Computational Engines

Scalable Storage Systems

Resource Management









kubernetes