Opportunities for Data-Management Research in the Era of Horizontal AI/ML

Panelists: Theo Rekatsinas (UW Madison)
Sudeepa Roy (Duke Univ.)
Manasi Vartak (Verta.AI)
Ce Zhang (ETH Zurich)

Moderator: Alkis Polyzotis (Google Research)

Starting points

ML is blooming as a field

- Rapid innovation and impact in research and industry
- Growing base of researchers and practitioners
- It’s now harder to get a NeurIPS registration than a ticket to Hamilton :-)

Starting points

ML is blooming as a field

- Rapid innovation and impact in research and industry
- Growing base of researchers and practitioners
- It’s now harder to get a NeurIPS registration than a ticket to Hamilton :-)

There is a strong link between ML and data management

- Data is the fuel for ML ⇒ Data management in the context of ML
- ML training/serving is a data flow ⇒ Optimizations from DB systems
- ML can crack hard problems ⇒ ML-driven DB system optimizations

Starting points

ML is blooming as a field

- Rapid innovation and impact in research and industry
- Growing base of researchers and practitioners
- It's now harder to get a NeurIPS registration than a ticket to Hamilton :-)

There is a strong link between ML and data management

- Data is crucial for ML ⇒ Data management in the context of ML
- ML training/serving is a data flow ⇒ Optimizations from DB systems
- ML can crack hard problems ⇒ ML-driven DB system optimizations

Good news for everyone in this room!

ML is becoming horizontal

ML is becoming horizontal

ML applies to more domains of increasing diversity

- Medical diagnosis, farming, chip design, transportation, astronomy, ...

ML is becoming horizontal

ML applies to more domains of increasing diversity

- Medical diagnosis, farming, chip design, transportation, astronomy, ...

Integration of ML in the stack is becoming wider and deeper

- Servers vs phones, machine-learned modules, hardware innovations...

ML is becoming horizontal

ML applies to more domains of increasing diversity

- Medical diagnosis, farming, chip design, transportation, astronomy, ...

Integration of ML in the stack is becoming wider and deeper

- Servers vs phones, machine-learned modules, hardware innovations...

More users, of varying skill sets, are relying on ML

- Engineers, analysts, scientists, ...

ML is becoming horizontal

ML applies to more domains of increasing diversity

- Medical diagnosis, farming, chip design, transportation, astronomy, ...

Integration of ML in the stack is becoming wider and deeper

- Servers vs phones, machine-learned modules, hardware innovations...

More users, of varying skill sets, are relying on ML

- Engineers, analysts, scientists, ...

What does this expansion imply for data management? ⇐ This panel!

Panel Structure

Question 1: Research opportunities (or, the good news!)

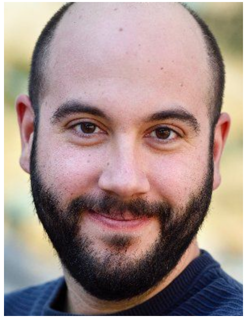
Question 2: How do we publicize our research?

Question 3: How do we train our students?

For each question:

- Panelists make their case (audience: hold your fire!)
- Open discussion (audience participation strongly encouraged)
- Next question

Panelists



Theo Rekatsinas
UW Madison

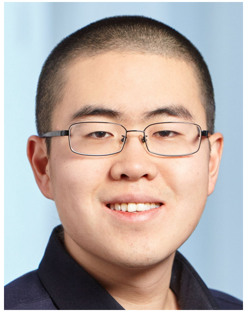
"As a teenager I used to juggle devil sticks. My first set was a gift from a psychiatrist."

Sudeepa Roy
Duke Univ.

"My other current research is on learning new nursery rhymes for my 18 months old daughter."

Manasi Vartak
Verta.AI

"My company's name is not based on my last name, just a need for available domain names ;) and also `ver=true`"



Ce Zhang
ETH Zurich

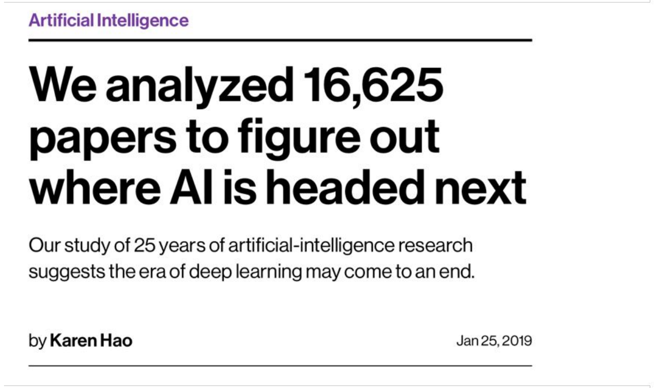
"I am trying to cycle around every single non-trivial lake in Switzerland, and I am almost 40% done."

Research opportunities

Theo

Are we seeing the whole picture?

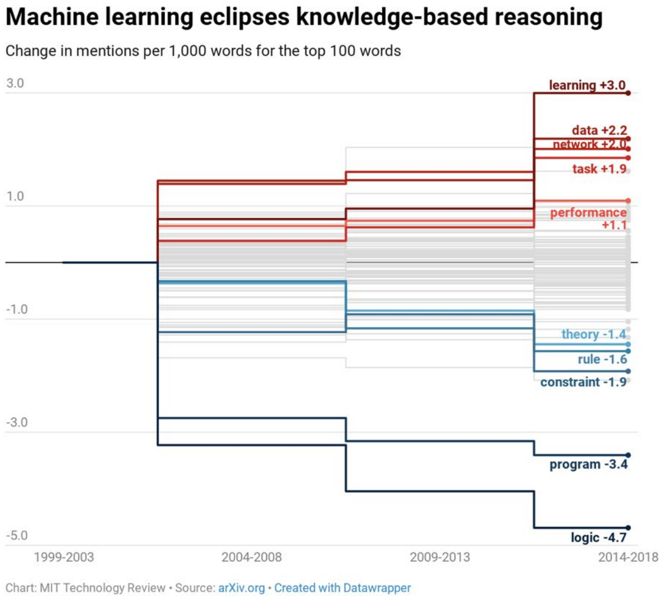
Let’s see where AI is headed next

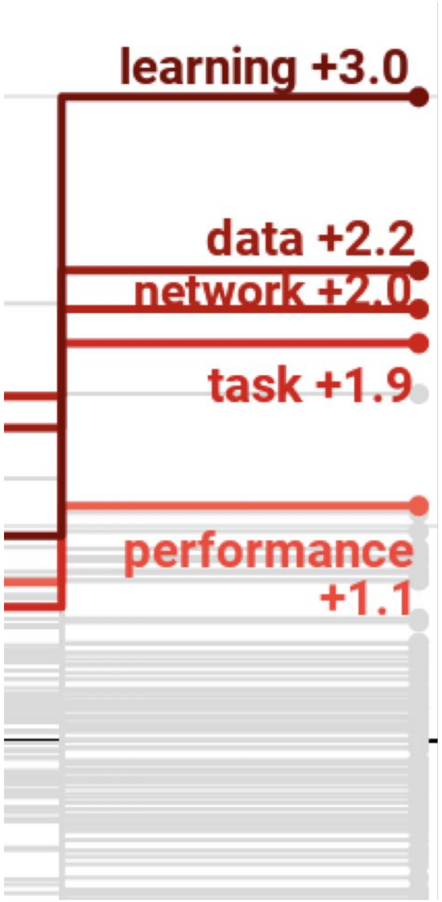

Artificial Intelligence

We analyzed 16,625 papers to figure out where AI is headed next

Our study of 25 years of artificial-intelligence research suggests the era of deep learning may come to an end.

by **Karen Hao** Jan 25, 2019

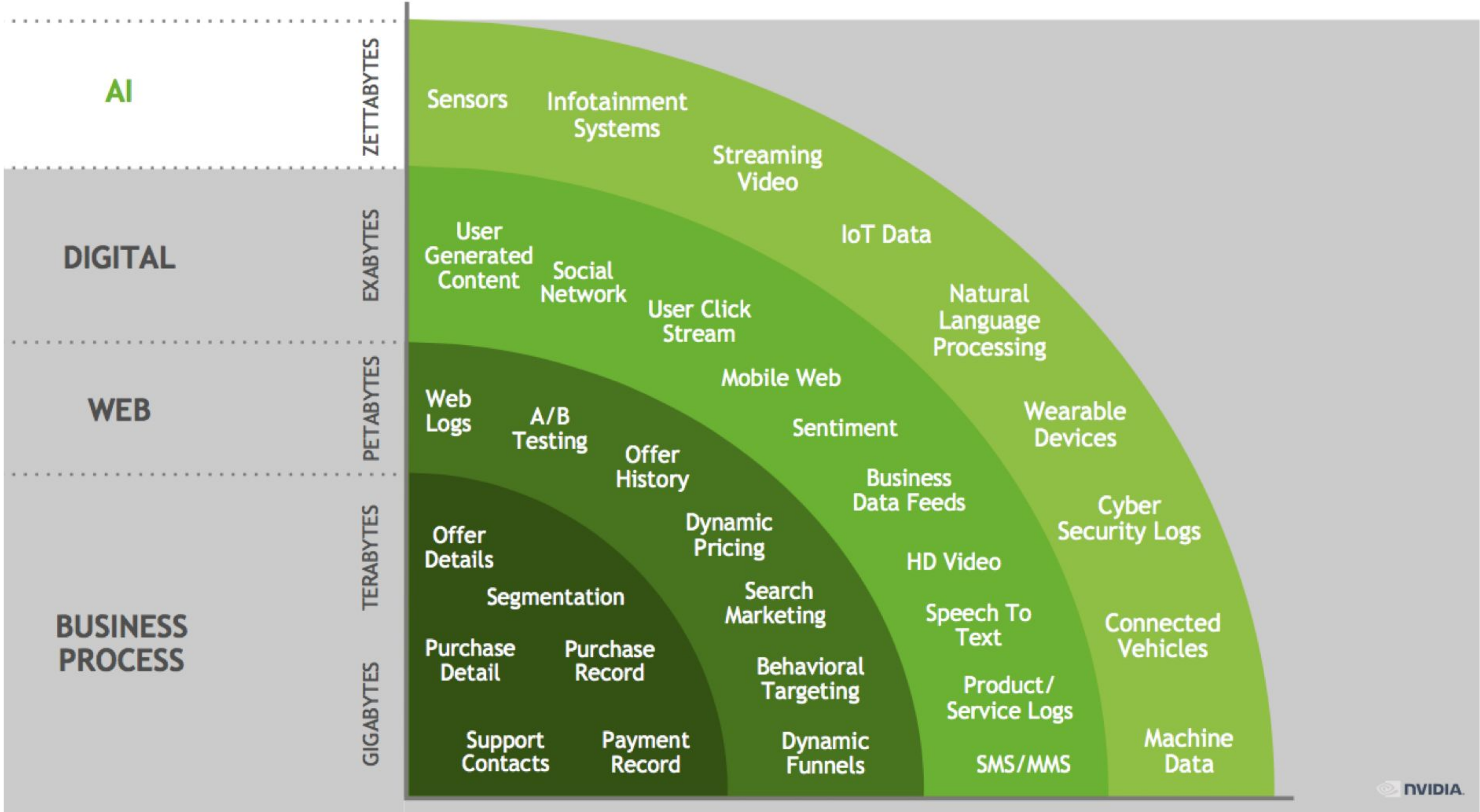




‘data’, ‘task’, and ‘performance’ are terms familiar to this community!

“What is THE most exciting challenge for AI (and Data Management)?”

Exploding data combined with shrinking time to act



AI
DIGITAL
WEB
BUSINESS PROCESS
ZETTABYTES
EXABYTES
PETABYTES
TERABYTES
GIGABYTES
Sensors
Infotainment Systems
Streaming Video
IoT Data
Natural Language Processing
Wearable Devices
Cyber Security Logs
Connected Vehicles
Machine Data
User Generated Content
Social Network
User Click Stream
Mobile Web
Sentiment
Business Data Feeds
HD Video
Speech To Text
Product/ Service Logs
SMS/MMS
Web Logs
A/B Testing
Offer History
Dynamic Pricing
Search Marketing
Behavioral Targeting
Dynamic Funnels
Offer Details
Segmentation
Purchase Detail
Purchase Record
Support Contacts
Payment Record
NVIDIA

The Achilles' Heel of Modern AI

- Data discovery: Explore data collections
- Data preparation: More than data cleaning (standardize, sample, augment/enhance)
- Data labeling: The necessary human cost

The Achilles' Heel of Modern AI

Many modern data management systems are being developed to address aspects of this issue:

HoloClean: Automated data enrichment

Snorkel: A System for Fast Training Data Creation

Google's TFX: TensorFlow Data Validation

Amazon's SageMaker

Amazon's Deequ: Data Quality Validation for ML Pipel

Opinion: Research in this area goes beyond data management

Example (from HoloClean):

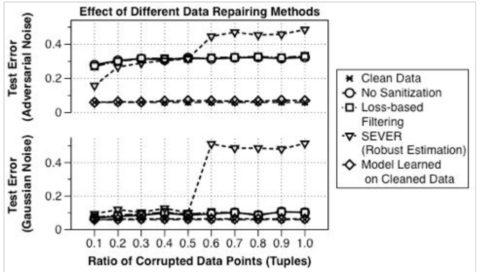
We started from data cleaning (a data management problem)

our intuition helped us solve an open problem in statistical learning theory

And now we are building new systems and theory for robust machine learning

HoloClean: Holistic Data Repairs with Probabilistic Inference

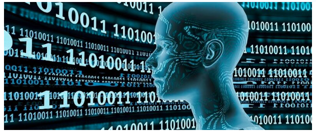
Approximate Inference in Structured Instances with Noisy Categorical Observations



Opinion: Data management is key to the success of AI.

Sudeepa

DM + ML/AI research opportunities



DM-4-ML

- Learning index, schema,
query optimization, access patterns
- Cardinality estimation
- Approximate Query Processing
- Regret-bounded query processing
- ....

We will talk about these anyway! :-)

ML-4-DM

- Systems for ML
- Faster inference
- Pushing ML through a query plan
- Curation and optimization of ML pipeline
- Automated training data generation
- Hardware for ML
- Distributed ML
- Linear algebra based analytics
- ....

My thoughts on research opportunities

1. Based on my research experience

2. From ML researchers’ experience

My thoughts on research opportunities

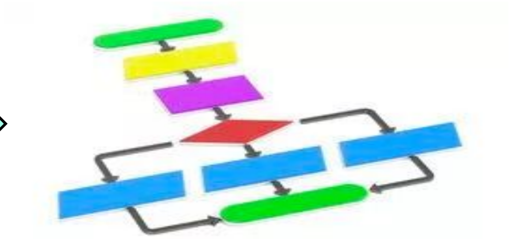
1. Based on my research experience

Relatively recent but interesting research using ML/AI
e.g., “Using regression to explain outliers” or “Learning to sample”

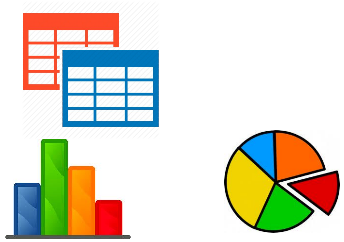
Interpretability/Explanations and Causality

Interpretability and Explanations

Input Data
D



Algorithm or Query
Q



Output(s)
Q[D]

How do we interpret and understand the output?

"Why do I see this output?"
"Why do I see an outlier?"
"Why is one value higher than the other?"
"Why is input-A classified as Type-B?"
"Why is sales in Jan predicted to be higher?"

Why Interpretability?

Transparency

Accountability

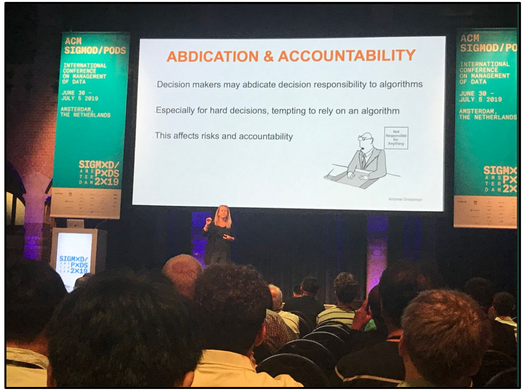
Ethics

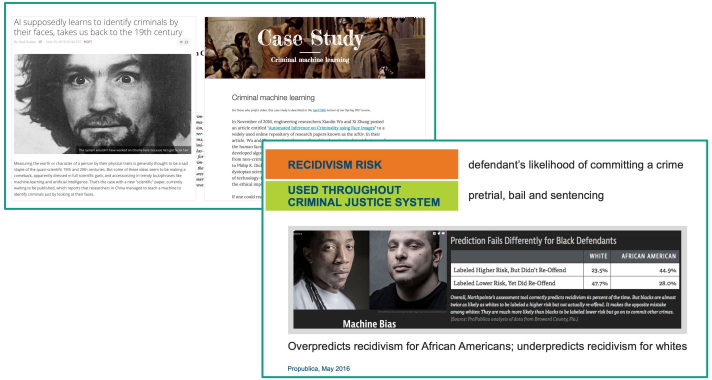
Actions

Fairness

Debugging

Maintainability





SIGMOD’19 Keynote by Lise Getoor on “**Responsible Data Science**”

SIGMOD’19 Panel on “**Data Ethics**”

Courtesy: Lise Getoor and SIGMOD’19 twitter account

How do we interpret and understand the output?

"Why do I see this output?"
"Why do I see an outlier?"
"Why is one value higher than the other?"
"Why is input-A classified as Type-B?"
"Why is sales in Jan predicted to be higher?"

Tracking "provenance" may not be enough

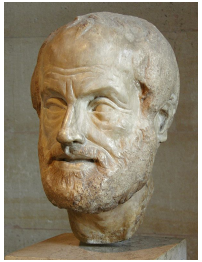
What are the main factors resulting in this prediction/classification/outlier?

How do we explain them to an analyst, decision maker, or scientist who does not hold an advanced degree in CS?

Ideally, "Why" = Find the "Cause"

Causes!

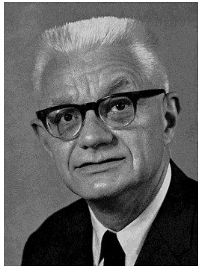
What are the ~~main factors~~ resulting in this prediction/classification/outlier?



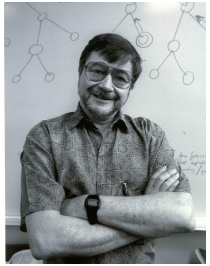
Aristotle
(384-322 BC)
Metaphysics

David Hume
(1738)
A Treatise of Human Nature

Karl Pearson
(1911)
The Grammar of Science



Carl Gustav Hempel
(1965)
Aspects of Scientific Explanation and Other Essays



Judea Pearl
Causality
Graphical Models

Beyond interpretability:
Causality has broader applications in sound "prescriptive" data analysis!

Helping decide whether or not a data-driven decision is wise

Correlation is not causation!

How much

- *“Does smoking cause lung cancer?”*
- *“Does drug A cure disease B?”*
- *“Does increasing tax on cigarettes reduce lung problems?”*
- *“Does a reduction in interests encourage people to buy houses?”*
- *“Does an increased icecream sale increase crime rate?”*

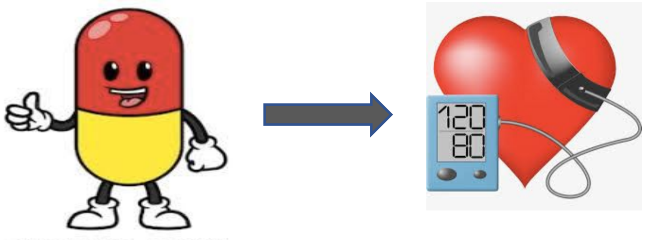
We cannot increase tax on icecream sales to stop crime!

* Both increase during summer

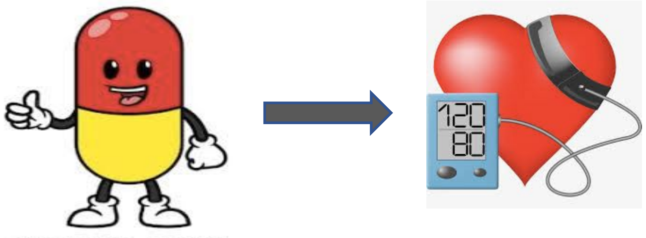
Going only by prediction or learning models for data-driven decisions,
the effect can be disastrous

Need to measure causality

Controlled experiment



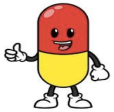
Controlled experiment



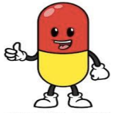
Compute average
and take difference

Randomization is crucial to estimate causal effect without bias

At random



Drug (treatment)



Placebo (control)

What if we cannot do randomized controlled experiments?

Due to ethical, time, or cost constraints

- *"Does smoking cause lung cancer?"*
- *"Does growing up in a poor neighborhood make a child earn less as an adult?"*

Fortunately, we can do
"Observational Causal Studies"
Under certain assumptions

Donald Rubin
Harvard Statistics
Potential Outcome
Framework for Causality

Observational Causal Study (+ DM)

Find "units" (e.g. patients) who look similar (called "matching")

- E.g., of same age, gender, height, ethnicity, …
- "Confounding covariates"

SQL Group-By

Many tools are available

But for small, simple data

With large data, SQL wins by a margin!

```
SELECT Age, Race, Gender, State, Education,
    ((SUM(T*Y)/SUM(T)) – (SUM(1-T)*Y)/(COUNT(*)-SUM(T))) AS ATE
FROM  Population
GROUP BY Age, Race, Gender, State, Education
HAVING SUM(T)>= 1 AND SUM(T) <= COUNT(*) - 1
```

4 Lines of SQL ⇒ Our two collaborative projects on causality and ML/AI!

DM-4-ML/AI



Lise Getoor **Babak Salimi** **Dan Suciu**

UCSC UW

- Causal analysis on large complex data
- Causal discovery
- Automatic assessment of key assumptions


Cynthia Rudin **Alexander Volfovsky**

Duke CS Duke Statistics

NIH

- Fast matching methods for large data using DM and ML techniques
- with applications in health data

e.g., Stopping flu-spread in college dorms

(with UNC Global Health)

New insights in data analysis or DM problems

ML-4-DM

SIGMOD’19 best paper by
Salimi et al. on fairness by causality!

My thoughts on research opportunities

ML-4-DM

DM-4-ML/AI

2. From ML researchers' experience

Do they face any data related problems?
Which problems they would like to solve?

Sometimes running batch scripts work for large data!

Some challenges faced in ML: 1/2

- Real-time systems and easy data flow and tensor flows
  - e.g., real-time neural network with frequent updates
- Infrastructure to work with Electronic Health Record and Medical Data
  - Privacy, updates, dataflow
- Efficient pre-processing in NLP
  - e.g., Find word-tuples appearing frequently and prune by some measures
- Image databases and image retrieval
  - Use the high level image structure (scene, objects, people, their spatial relation) , and find images whose structure satisfies some property?

Some challenges faced in ML: 2/2

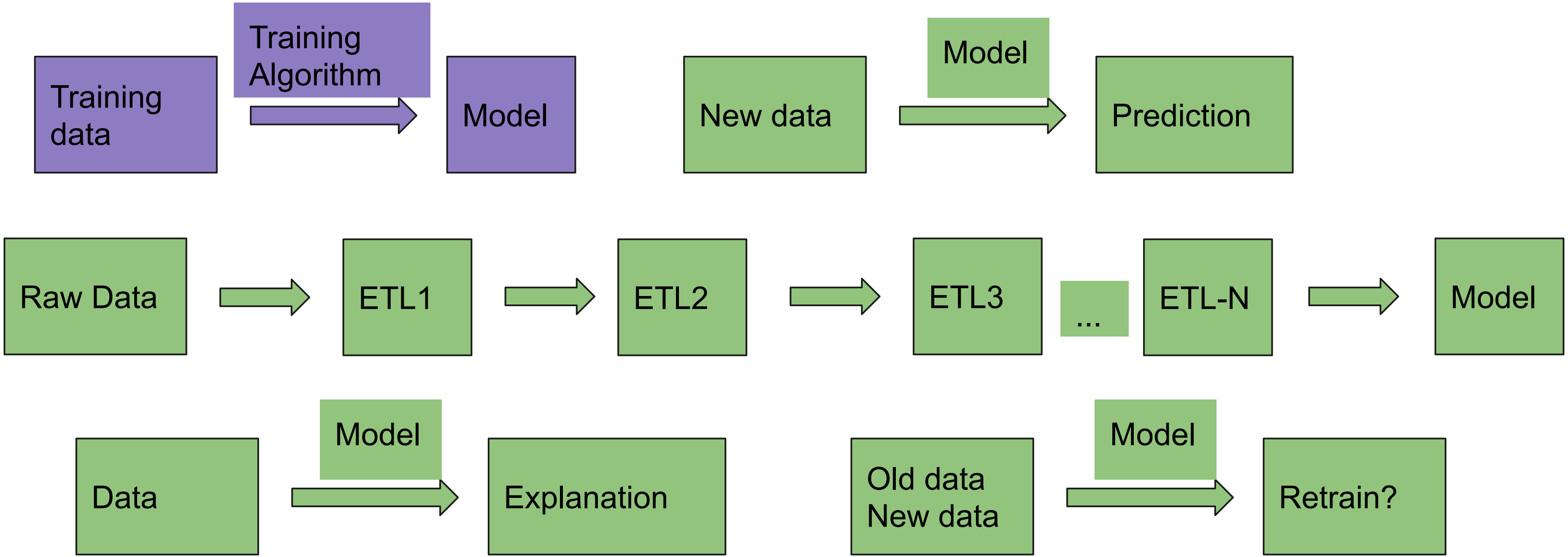
- Storing large data in computational genomics
  - Genome has 3 billion DNA-bases so genome-wide predictions are hard to store
  - Can be compressed well, but does compression work with ML method?

- Storing and analyzing 1600 hours of video data
  - extract gestures, conversations, etc. and model the behavior of the individuals there

Some problems may be worth looking also from DM viewpoint.
Collaboration and co-advising students would help.

Manasi

ML & AI is a Data Game



But We Are NOT Where the Workloads Are

Problem 1: Better abstractions for ETL for ML

??

Problem 1: Better abstractions for ETL for ML

Problem 2: Data Versioning, Discovery, Lineage

Principles of dataset versioning: exploring the recreation/storage tradeoff

Full Text: PDF Get this Article

Authors:		
	Souvik Bhattacherjee	University of Maryland, College Park
	Amit Chavan	University of Maryland, College Park
	Silu Huang	University of Illinois, Urbana-Champaign
	Amol Deshpande	University of Maryland, College Park
	Aditya Parameswaran	University of Illinois, Urbana-Champaign

Collaborative data analytics with DataHub

Full Text: PDF Get this Article

Authors:		
	Anant Bhardwaj	MIT
	Amol Deshpande	U. Maryland (UMD)
	Aaron J. Elmore	U. Chicago
	David Karger	MIT
	Sam Madden	MIT
	Aditya Parameswaran	U. Illinois (UIUC)
	manyam	MIT
		Columbia

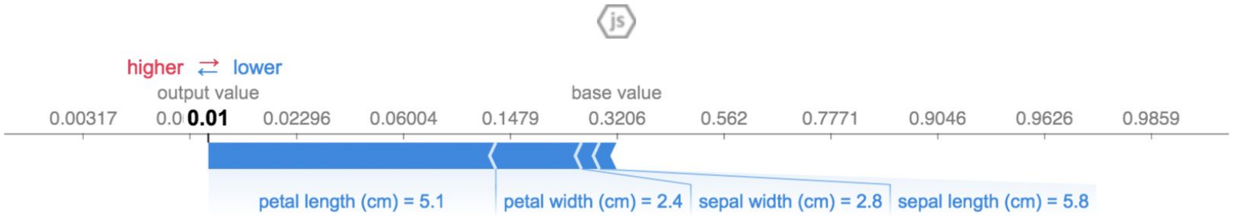
Aurum: A Data Discovery System

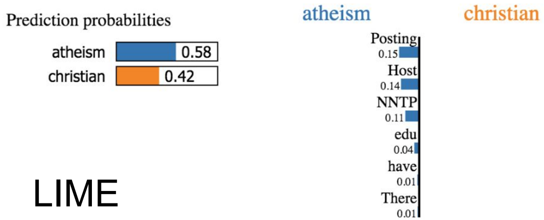
Raul Castro Fernandez, Ziawasch Abedjan[#], Famien Koko, Gina Yuan, Sam Madden, Michael Stonebraker

MIT <raulcf, fakoko, gyuan, madden, stonebraker>@csail.mit.edu [#] *TU Berlin* abedjan@tu-berlin.de

Problem 3: Data-Driven Model Explanations

SHAP





LIME

MISTIQUE: A System to Store and Query Model Intermediates for Model Diagnosis

Full Text: PDF Get this Article

Authors:
- Manasi Vartak — Massachusetts Institute of Technology, Cambridge, MA, USA
- Joana M. F. da Trindade — Massachusetts Institute of Technology, Cambridge, MA, USA
- Samuel Madden — Massachusetts Institute of Technology, Cambridge, MA, USA
- Matei Zaharia — Stanford University, Palo Alto, CA, USA

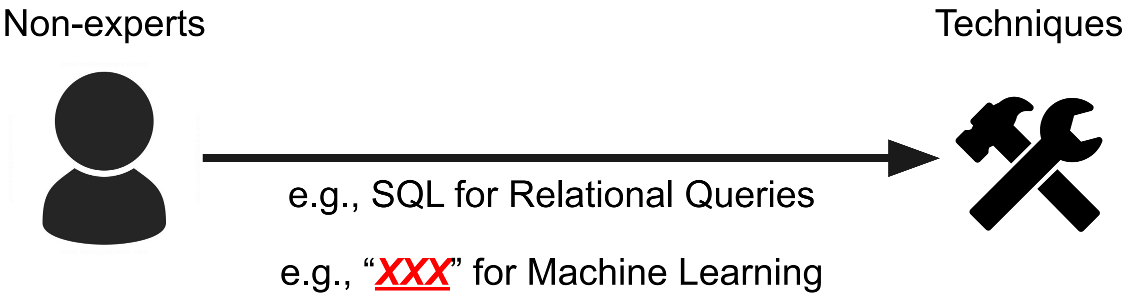
2018 Article

Bibliometrics
- Citation Count: 2
- Downloads (cumulative): 283

Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, Rory Sayres

Ce



How does the next generation Machine Learning platform look like for non-expert users to unleash the full potential of ML?

Usability of learning systems -- we are excited about this because I believe there are no other community more suitable than us to answer this question -- ***ML is just another way of analyzing the data, whatever we did to make SQL awesome and accessible, we need to redo it for ML.***

Let me share with you three research opportunities we realized over time (two are “embarrassingly obvious”).

SPEED! SPEED! SPEED!

- *Once upon a time…*

EC2 Instance: g2.8xlarge

- 4x GRID K520
- ~ TFLOPS

- *Today…*

EC2 Instance: p3.16xlarge

- 8x V100
- ~ PFLOPS

Training ResNet-50 on ImageNet in 5h = $120

https://www.nvidia.com/en-us/data-center/tesla-v100/

SPEED! SPEED! SPEED!

- *Once upon a time…*

EC2 Instance: g2.8xlarge
- 4x GRID K520
- ~ TFLOPS

- *Today…*

EC2 Instance: p3.16xlarge
- 8x V100
- ~ PFLOPS

Training ResNet-50 on ImageNet in 5h = $120

- Speed is still a huge, huge problem (many models on *mid*-size dataset still takes weeks with a cluster of GPUs)

We should continue to play a role here, especially when distributed learning systems are becoming more sophisticated and require more tuning, just like a relational DB.

https://www.nvidia.com/en-us/data-center/tesla-v100/

Speed is necessary but not sufficient

1 Biologist + 8 V100 $\neq$ 1 ML Model

AUTOMATION! AUTOMATION! AUTOMATION!

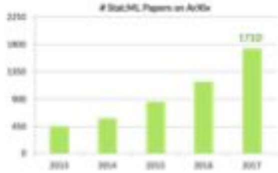
- *Even when training is fast, users are overwhelmed by choices.*

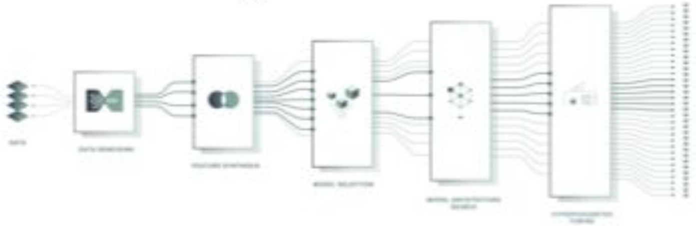
AUTOMATION! AUTOMATION! AUTOMATION!

- *Even when training is fast, users are overwhelmed by choices.*

AlexNet, ResNet, GoogLeNet, DenseNet...

ResNet-18
ResNet-34
ResNet-50
...
Other hyper-parameters



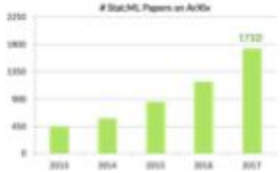


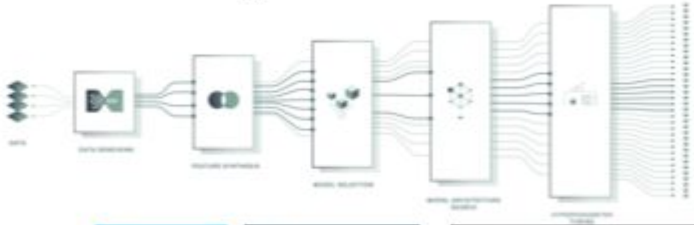
AUTOMATION! AUTOMATION! AUTOMATION!

- *Even when training is fast, users are overwhelmed by choices.*

AlexNet, ResNet, GoogLeNet, DenseNet...

ResNet-18
ResNet-34
ResNet-50
...
Other hyper-parameters





DataRobot

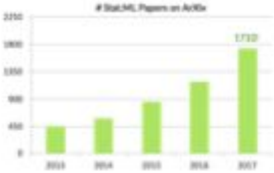
"Shameless self-advertisement" →

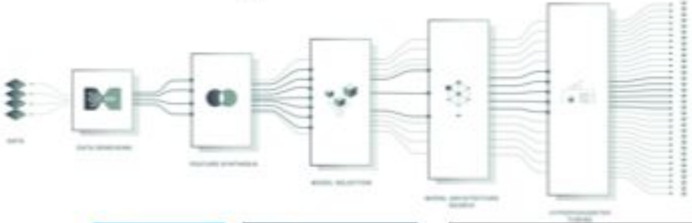
AUTOMATION! AUTOMATION! AUTOMATION!

- *Even when training is fast, users are overwhelmed by choices.*

AlexNet, ResNet, GoogLeNet, DenseNet...

ResNet-18
ResNet-34
ResNet-50
...
Other hyper-parameters





- Automation is still a huge, huge problem (search space, search alg., data/computation sharing, etc.)

DataRobot

"Shameless self-advertisement" →

ML In Three Days: The Space

Day 1

Goal: To get your first ML model as fast/easy as possible.

Speed

Automation

Too powerful -- users are overwhelmed.

ML In Three Days: The Space

Day 0

Goal: To get your first ML model as fast/easy as possible.

Feasibility Study & Sanity Check

Day 1

Goal: To get your first ML model as fast/easy as possible.

Speed

Automation

Day 2

Goal: To get a sequence of models that gets better and better.

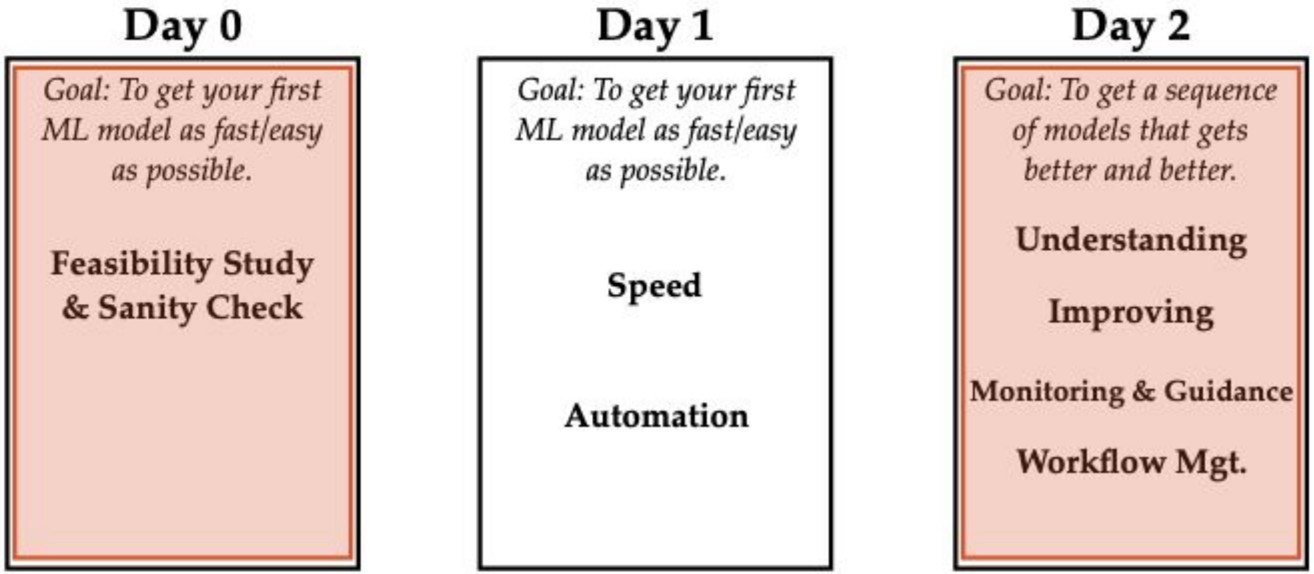
Understanding

Improving

Monitoring & Guidance

Workflow Mgt.

ML In Three Days: The Space



If ML is "Software 2.0", users need "Software Engineering 2.0" -- and deep down, I believe this is our opportunities to lose.

How do we publicize our research?

Theo

The Data Management ambassadors

An increasing number of data management researchers are turning their attention to ICML, NeurIPS, KDD, Systems for Machine Learning Conference.

These people are our ambassadors!

The Data Management ambassadors

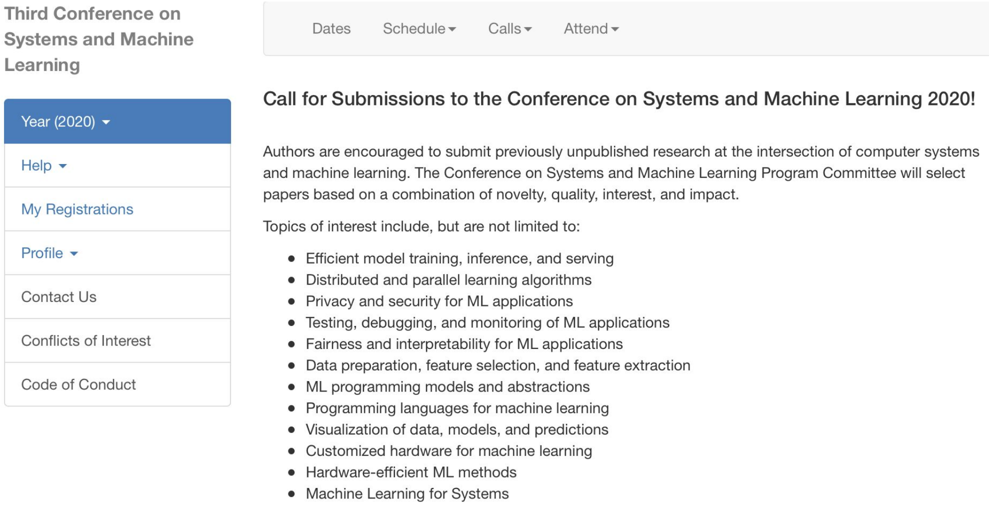
An increasing number of data management researchers are turning their attention to ICML, NeurIPS, KDD, Systems for Machine Learning Conference.

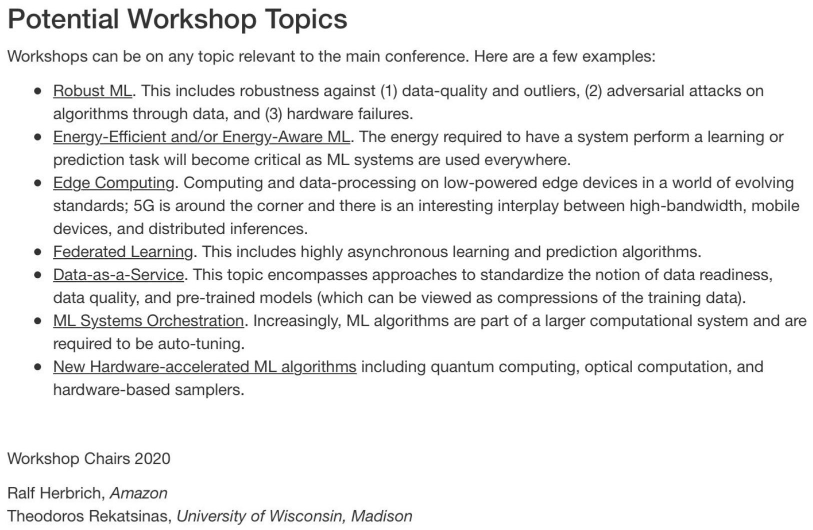
These people are our ambassadors!

Opinion: These works do not focus on what one would call traditional data management problems. This is why other venues can be more attractive.

Why ambassadors matter: They bring (1) visibility and (2) expertise that can help diversify the current agenda of the data management conferences.

Systems and Machine Learning Conference: An example of a diverse agenda





Give the stage to the ambassadors of other fields

KDD2018 HOME PROGRAM ATTENDING CALLS SPONSORS ORGANIZERS KDD CUP CONTACT US

benefit all aspects of society. KDD 2018 is looking to be an amazing year.

Keynote Speakers

David Hand

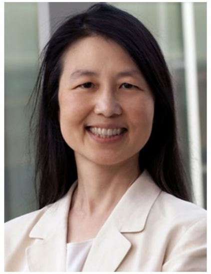
SENIOR RESEARCH INVESTIGATOR EMERITUS PROFESSOR OF MATHEMATICS, IMPERIAL COLLEGE

Alvin E. Roth

NOBEL MEMORIAL PRIZE IN ECONOMICS
PROFESSOR OF ECONOMICS, STANFORD UNIVERSITY

Yee Whye Teh

PROFESSOR,DEPARTMENT OF STATISTICS, UNIVERSITY OF OXFORD
RESEARCH SCIENTIST, DEEPMIND



Jeannette M. Wing

AVANESSIANS DIRECTOR OF THE DATA SCIENCES INSTITUTE
COLUMBIA UNIVERSITY

Opinion: More keynote talks by people outside our area! KDD is a great example!

Give the stage to the ambassadors of other fields

Opinion: Accept original works that address problems in non-traditional data management/database areas (e.g., systems for scaling ML workloads).

But… we need to be careful to accept papers that would only be accepted at top-tier conferences. VLDB and SIGMOD are precious and should not become 2nd-tier ML conferences.

We need external expertise to ensure the above. Let’s bring in experts to help!

Sudeepa

What can we do as a community?

Discuss

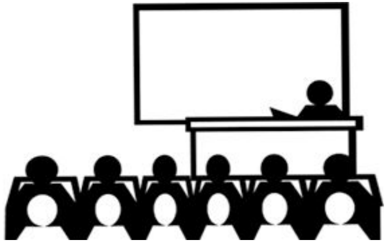
Research

NII SHONAN MEETING

Workshops
cost/overhead?

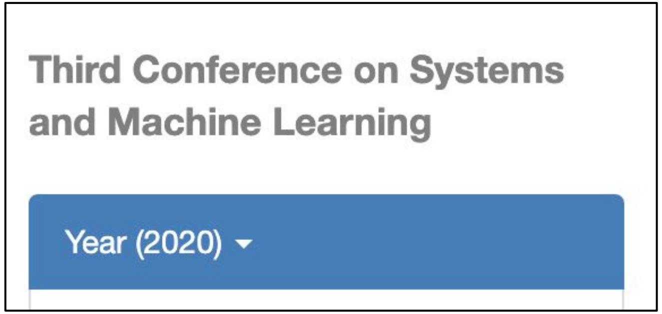
SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

Collaborate



More keynote from ML/AI
in major DM conferences

Publication venues?



Something similar for non-systems / theory / application-based research combining ML/AI and DM?

Publication in NeurIPS, ICML, AAAI, IJCAI
Review process, acceptance of DM ideas?

Give more DM-related talks in ML conferences and workshops?

Manasi

Conferences != Publicity

Why is Tensorflow so famous?

- It solves a real problem
- It’s good software
- Google pushed hard to publicize it

Democratization ⇒ Non-researchers can appreciate and use

Solve problems based on current use cases

snorkel

Helix

MLDB

The ML Data Prep Zoo: Towards Semi-Automatic Data Preparation for ML

Vraj Shah, Arun Kumar
University of California, San Diego
{vps002,arunkk}@eng.ucsd.edu

NoScope: 1000x Faster Deep Learning Queries over Video

by Daniel Kang, John Emmons, Firas Abuzaid, Peter Bailis, and Matei Zaharia

awslabs / **deequ**

<> Code | Issues 33 | Pull requests 0 | Projects 0 | Security | Insights

Deequ is a library built on top of Apache Spark for defining "unit tests for data", which measure data quality in large datasets.

dataquality spark unit-testing scala

Blogs, Twitter, Talks & Reusable Code

Medium

Big Tech Cos,
Meetups,
Demos

Beware the pitfalls of Open-Source

2 reasons:

- I: Reproducibility or selling point of paper
- II: Actually want people to adopt it

If II:

- Need significant support, software engineering resources
- Meetups, outreach
- *If you aren't able to do this, don't open-source*

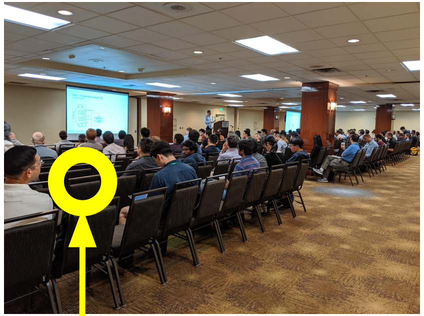
Ce

Today, VLDB/SIGMOD is not on many people's radar for ML Systems -- People think about VLDB/SIGMOD when they want to read about DB, not ML Sys.

All of my students in their 1st year were surprised that we send our best ML System work every year to VLDB/SIGMOD instead of NIPS/ICML.

We need to establish VLDB/SIGMOD as the top venue for most, if not all ML System topics.

Yesterday

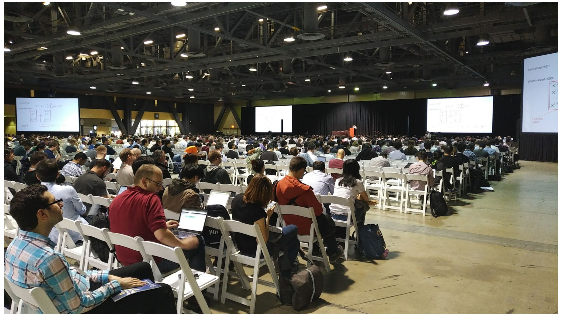


(~120 People)

SysML 2019

(~500 Registrations)

NIPS 2017



We should publicize VLDB/SIGMOD such that many of these people come to our ML sessions looking for the best ML system work.

“But do we have the expertise to assess ML Sys Papers?”

Program Committee

Dan Alistarh, Gustavo Alonso, Anima Anandkumar, David Andersen, Peter Bailis, Sarah Bird, Joseph Bradley, John Canny, Nicholas Carlini, Bryan Catanzaro, Eric Chung, William Dally, Christopher De Sa, Inderjit Dhillon, Alex Dimakis, Pradeep Dubey, Kayvon Fatahalian, Lise Getoor, Phillip Gibbons, Garth Gibson, Joseph Gonzalez, Justin Gottschlich, Song Han, Kim Hazelwood, Cho-Jui Hsieh, Furong Huang, Martin Jaggi, Prateek Jain, Kevin Jamieson, Yangqing Jia, Gauri Joshi, Rania Khalaf, Jason Knight, Jakub Konecný, Tim Kraska, Arun Kumar, Anastasios Kyrillidis, Aparna Lakshmiratan, Jing Li, Brendan McMahan, Erik Meijer, Ioannis Mitliagkas, Rajat Monga, Dimitris Papailiopoulos, Gennady Pekhimenko, Alex Ratner, Theodoros Rekatsinas, Afshin Rostamizadeh, Hanie Sedghi, Siddhartha Sen, Evan Sparks, Ion Stoica, Vivienne Sze, Ameet Talwalkar, Madeleine Udell, Jaoquin Vanschoren, Shivaram Venkataraman, Markus Weimer, Andrew Wilson, Ce Zhang

We do have expertise to assess ML system papers!

We should be ***confident***, and grab the opportunity

60 SysML 2019 Reviewers

11 -- DB/DM -- **18%**

29 -- ML ←

11 -- System

7 -- Architecture

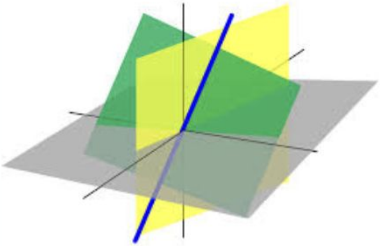
2 -- Other

*If **13** of these reviewers agree to be our external reviewers, we have **40%** of SysML PC.*

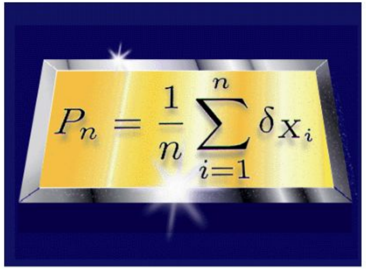
How do we prepare our students?

Theo

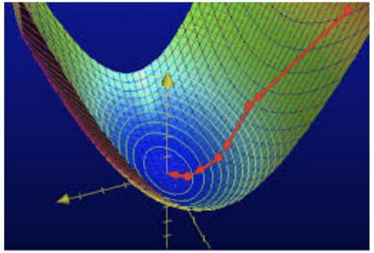
Mathematical Foundations of ML



Linear Algebra

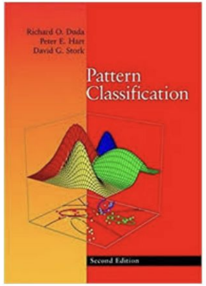


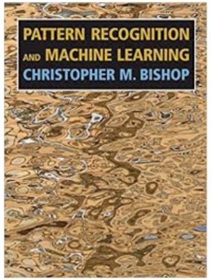
Probability Theory

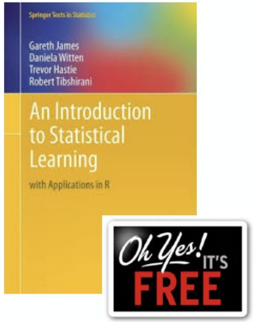


Optimization

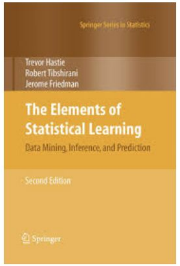
Basic ML methods and mathematical background

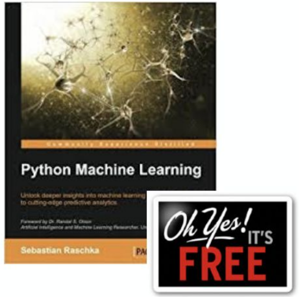




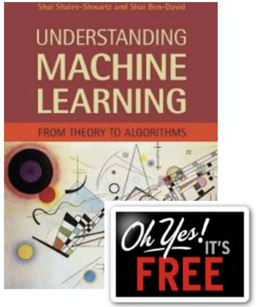


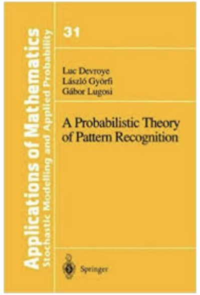
Algorithms and coding





Basic theory (more advanced, 861 level)





And to be a real data management/database researcher, you must take 764!

CS 764, Fall 2017: Topics in Database Management Systems

Coordinates: MWF 9:30-10:45 in **1257 CS (note the change in room)**
Instructor: J. Patel
Office Hours: Wed 10:45-11:45AM or by appointment

Description

This course covers a number of advanced topics in the development of data management systems and the application of such systems in modern applications. The topics discussed include advanced concurrency control and recovery techniques, query processing and optimization strategies, advanced access methods, parallel and distributed data systems, extensible data systems, implications of cloud computing for data platforms, and data analysis on large datasets.

The course material will be drawn from a number of papers in the database literature. We will cover about 2-3 papers per week. All students in this class are expected to read the papers before coming to the lecture.

Sudeepa

What can we do as a community

- A common repository of course material from researchers working on ML + DM?
  - With a common discussion forum? Led by senior students?
  - Challenges:
    - Difficult to sustain if centrally-managed
    - Cost, storage, spam, moderating knowledge flow
- Organize 1-day long bootcamps with SIGMOD/VLDB?
  - Similar to workshops but focused on teaching basics as well as relevant research
  - Similar to tutorials but longer, probably by multiple people

Other ideas

- Take both ML and DM courses

Take advantage of the online courses and material?

- Teach students how to use DMs in data analysis, not just how to build DMs

A module in a DM courses (ML too?)
Or an advanced course on data analysis?

- ML students may not always appreciate the need for DM techniques for modern ML applications. ML/AI courses dealing with large datasets that are too big to store/manage “naively” would be helpful

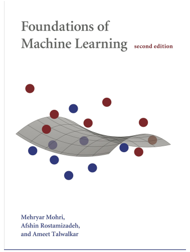
Scalable ML? But ML courses are already popular!
Team up with a colleague in ML?

Manasi

- Move beyond relational data
- Focus on core data processing techniques (ETL, queries, indexing, caching)
- Understand scalability and techniques to tame it
- Need basic understanding of ML (e.g., just like calculus)
- *Be proud that you work with data :)*

Ce

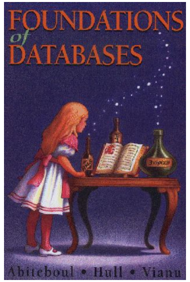
- Given all the excitement around ML, I am not that worried about students not learning ML -- they are smarts, they will learn.
- Sure, we need to provide some guidances to:
  - help them to decouple fundamentals with hypes.
  - make sure they are not only attracted by fancy applications but also the core fundamental theory.

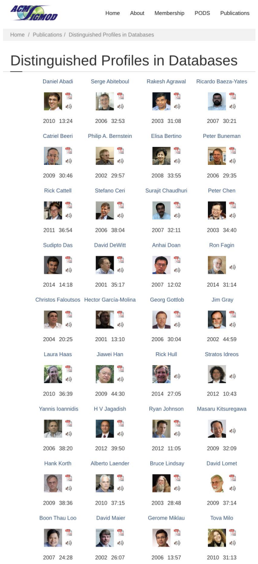


i.e., I am not worried that our students do not know about this book.

***<u>My Bias</u>**: All of my students wanted to do ML instead of DB/DM when they first come to my group -- so I have been "converting" students who want to do ML into DB/DM instead of the other way around.*

- Amid all the excitement around ML, we need to make sure our students learn about DATABASE and DATA MANAGEMENT properly:
  - We need to remind them how ***<u>cool</u>*** DATABASE is.
  - ***<u>History of database research</u>*** -- Not only how things are working today, but also the exploratory process of how we reach where we are today.
  - ***<u>Database Theory</u>*** -- DB goes way beyond systems, it has solid theoretical foundation.





- The DB/DM aspect is what make our student's background unique:
  - We need to make sure they ***<u>realize</u>*** it, ***<u>appreciate</u>*** it, and ***<u>be proud of</u>*** it.