The New DBfication of ML/AI: Saving Them from Themselves

Arun Kumar
Outline

1. What saving am I doing?
2. Why should we (DB types) care?
3. How can you be a savior too?
Golden Age of ML/AI

amazon  FACEBOOK  Google  Microsoft
Golden Age of ML/AI
Golden Age of ML/AI

$ 38 billion in 2019*

$ 500 billion! by 2024*

*International Data Corporation
Golden Age of ML/AI

Amazon  
Facebook  
Google  
Microsoft

MAYO CLINIC  
AMERICAN FAMILY INSURANCE  
Walmart  
Sciences

Healthcare  
Insurance  
Retail

$ 38 billion in 2019*  
$ 500 billion! by 2024*

TensorFlow  
PyTorch

*XInternational Data Corporation
Golden Age of ML/AI

$38 billion in 2019*

$500 billion! by 2024*

*International Data Corporation
Golden Age of ML/AI

Still, fundamental **efficiency and usability bottlenecks** in the **end-to-end process** of building and deploying ML applications

*International Data Corporation*
My Research

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

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

Democratization = System Efficiency (Reduce costs) + Human Efficiency (Improve productivity)
New abstractions, algorithms, and software systems to “democratize” ML/AI-based data analytics from a data management/systems standpoint.

Democratization = System Efficiency (Reduce costs) + Human Efficiency (Improve productivity)
My Research

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

Democratization = System Efficiency (Reduce costs) + Human Efficiency (Improve productivity)

Practical and scalable data systems for ML/AI analytics
Inspired by relational database systems principles
Exploit insights from learning theory and optimization theory
End-to-End ML Application Lifecycle

Data Scientist/ML Engineer

Source → Build → Deploy

Data + ML Systems Implementations

https://ADALabUCSD.github.io
End-to-End ML Application Lifecycle

Data Scientist/ML Engineer

Source → Build → Deploy

Research Approach:

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

https://ADALabUCSD.github.io
Outline

- What saving am I doing?
- Why should we (DB types) care?
- How can you be a savior too?
The way I see it, ML/AI systems/platforms today resemble RDBMSs circa early ‘80s
The New DBfication of ML/AI

Data Scientist/ML Engineer

Source → Build → Deploy
The New DBfication of ML/AI

Data Scientist/ML Engineer

Source → Build → Deploy

Metadata Management for ML
Data Prep/Cleaning for ML
Multimodal ML Query Models
Data Search, Labeling, etc.

...
The New DBfication of ML/AI

Data Scientist/ML Engineer

Source → Build → Deploy

Metadata Management for ML
Data Prep/Cleaning for ML
Multimodal ML Query Models
Data Search, Labeling, etc.
...

Scalable Data Systems for ML
Query Optimization for ML
Cloud and Streaming Infra.
Provenance and Debugging
...

8
The New DBfication of ML/AI

Data Scientist/ML Engineer

Source \[\rightarrow\] Build \[\rightarrow\] Deploy

Metadata Management for ML
Data Prep/Cleaning for ML
Multimodal ML Query Models
Data Search, Labeling, etc.

Scalable Data Systems for ML
Query Optimization for ML
Cloud and Streaming Infra.
Provenance and Debugging

Benchmark Frameworks and Data
Fairness, Transparency, Privacy, etc.
The New DBfication of ML/AI

Data Scientist/ML Engineer

Source → Build → Deploy

Metadata Management for ML
Data Prep/Cleaning for ML
Multimodal ML Query Models
Data Search, Labeling, etc.

Scalable Data Systems for ML
Query Optimization for ML
Cloud and Streaming Infra.
Provenance and Debugging

Benchmark Frameworks and Data
Fairness, Transparency, Privacy, etc.
The New DBfication of ML/AI

Data Scientist/ML Engineer

Source → Build → Deploy

Metadata Management for ML
Data Prep/Cleaning for ML
Multimodal ML Query Models
Data Search, Labeling, etc.

Scalable Data Systems for ML
Query Optimization for ML
Cloud and Streaming Infra.
Provenance and Debugging

Benchmark Frameworks and Data
Fairness, Transparency, Privacy, etc.
The New DBfication of ML/AI

Data Scientist/ML Engineer

Source \[\rightarrow\] Build \[\rightarrow\] Deploy

Metadata Management for ML
Data Prep/Cleaning for ML
Multimodal ML Query Models
Data Search, Labeling, etc.

Scalable Data Systems for ML
Query Optimization for ML
Cloud and Streaming Infra.
Provenance and Debugging

... Benchmark Frameworks and Data
Fairness, Transparency, Privacy, etc.

...
If we (DB types) do not tackle such DB-style problems, who else will?
If we (DB types) do not tackle such DB-style problems, who else will?
Leaving them open => Huge waste of time/effort/money/energy/etc. by ML/AI types!
Outline

- What saving am I doing?
- Why should we (DB types) care?
- How can you be a savior too?
Becoming a DBesque ML/AI Savior 101

1. Learn the fundamentals of ML/AI algorithms and theory.
   Kinda like learning logic, RA, SQL, etc. for RDBMSs
   Review ML/AI algorithms courses in your institution or online
   3 key books: Hastie et al. (Stat. ML); Mitchell (ML); Courville et al. (DL)
   Need-to-know spectrum: DL for DL sys.; ML theory for accuracy tradeoffs
Becoming a DBesque ML/AI Savior 101

1. Learn the fundamentals of ML/AI algorithms and theory.
   Kinda like learning logic, RA, SQL, etc. for RDBMSs
   Review ML/AI algorithms courses in your institution or online
   3 key books: Hastie et al. (Stat. ML); Mitchell (ML); Courville et al. (DL)
   Need-to-know spectrum: DL for DL sys.; ML theory for accuracy tradeoffs

2. Check out my “DB for ML” grad course and research book? :)
Becoming a DBesque ML/AI Savior 101
Becoming a DBesque ML/AI Savior 101

3. Check out recent “DB for ML” tutorials and papers.
Becoming a DBesque ML/AI Savior 101

3. Check out recent “DB for ML” tutorials and papers.

4. Check out topical panel discussions on “DB for ML” stuff.
*Explosive* mystery panel coming to SIGMOD 2021! ;)

5. Attend SIGMOD DEEM and HILDA Workshops. Check out MLSys.
5. Attend SIGMOD DEEM and HILDA Workshops. Check out MLSys.

6. MOST IMPORTANT: Speak/collaborate with ML/AI users, build REAL stuff, and help transfer research to practice.
   Data scientists, Data analysts, ML engineers, MLOps engineers, etc.
   Create open source artifacts, both software and data
   Enterprises, Web companies, cloud vendors, domain sciences, policy, etc.
   Attend/speak at industry venues: Spark+AI Summit, FOSDEM, etc.
   ...
My Terrific Advisees

Supun Nakandala  
PhD

Tara Mirmira  
PhD

Vraj Shah  
PhD & MS

Xiuwen Zheng  
PhD & MS

Yuhao Zhang  
PhD & MS

Side Li  
MS & BS

Advitya Gemawat  
BS

Kabir Nagrecha  
BS

Shaoqing Yi  
BS

https://ADALabUCSD.github.io
https://ADALabUCSD.github.io
arunkk@eng.ucsd.edu
github.com/ADALabUCSD
@TweetAtAKK

Acks:

NIH, NIDDK, NSF, Hellman Fellows Funds, UC San Diego, AWS, Google, Oracle, VMware