The New DBfication of ML/AI

Arun Kumar

UC San Diego
JACOBS SCHOOL OF ENGINEERING
Computer Science and Engineering

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

DB & AI Seminar, China

June 26, 2022

Golden Age of ML/AI

Golden Age of ML/AI

amazon FACEBOOK Google Microsoft

MAYO CLINIC

Healthcare

AMERICAN FAMILY INSURANCE

Insurance

Walmart

Retail

Sciences

Golden Age of ML/AI

amazon FACEBOOK Google Microsoft

MAYO CLINIC Healthcare

AMERICAN FAMILY INSURANCE® Insurance

Walmart Retail

 Sciences

$ 38 billion in 2019*

→

$ 500 billion! by 2024*

*International Data Corporation

Golden Age of ML/AI

amazon FACEBOOK Google Microsoft

MAYO CLINIC

Healthcare

AMERICAN FAMILY INSURANCE

Insurance

Walmart

Retail

Sciences

$ 38 billion
in 2019*

$ 500 billion!
by 2024*

scikit learn dmlc XGBoost APACHE Spark ML

TensorFlow PyTorch

*International Data Corporation

Golden Age of ML/AI

amazon FACEBOOK Google Microsoft

MAYO CLINIC Healthcare

AMERICAN FAMILY INSURANCE Insurance

Walmart Retail

 Sciences

$ 38 billion
in 2019*

→

$ 500 billion!
by 2024*

scikit learn dmlc XGBoost APACHE Spark ML

TensorFlow PyTorch

amazon Google Microsoft

*International Data Corporation

Golden Age of ML/AI

amazon FACEBOOK Google Microsoft

MAYO CLINIC — Healthcare

AMERICAN FAMILY INSURANCE — Insurance

Walmart — Retail

 Sciences

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

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

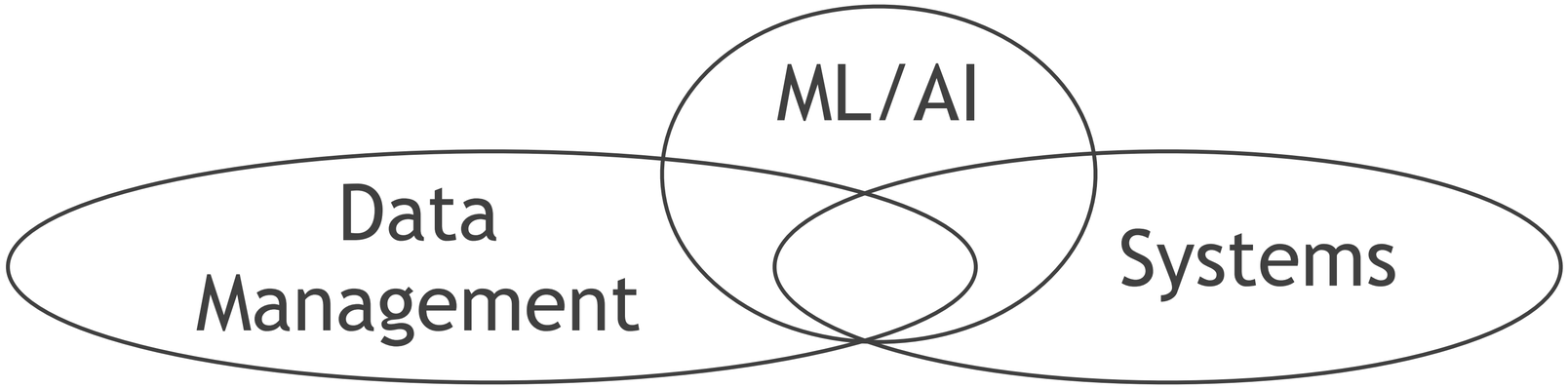
New abstractions, algorithms, and software systems to "***democratize***" ML/AI-based data analytics from a data management/systems standpoint

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

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)



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

Data + ML Systems Implementations

Research Approach: ***Abstract*** key steps + ***Formalize*** computation + ***Automate*** grunt work + ***Optimize*** execution

https://ADALabUCSD.github.io

Outline

- The New DBfication of ML/AI
- Two Examples from My Research
- Accelerating the DBfication of ML/AI

DB-style Practical Concerns in ML

DB-style Practical Concerns in ML

Key concerns in ML:

DB-style Practical Concerns in ML

Key concerns in ML:

Accuracy

DB-style Practical Concerns in ML

Key concerns in ML:

Accuracy

Runtime efficiency (sometimes)

DB-style Practical Concerns in ML

Key concerns in ML:

Accuracy

Runtime efficiency (sometimes)

Additional key *practical* concerns in ML Systems:

DB-style Practical Concerns in ML

Key concerns in ML:

Accuracy

Runtime efficiency (sometimes)

Additional key *practical* concerns in ML Systems:

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

DB-style Practical Concerns in ML

Key concerns in ML:

Accuracy

Runtime efficiency (sometimes)

Additional key *practical* concerns in ML Systems:

Scalability (and efficiency at scale)

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

DB-style Practical Concerns in ML

Key concerns in ML:

Accuracy

Runtime efficiency (sometimes)

Additional key *practical* concerns in ML Systems:

Scalability (and efficiency at scale)

Q: How are the features and models configured?

DB-style Practical Concerns in ML

Key concerns in ML:

Accuracy

Runtime efficiency (sometimes)

Additional key *practical* concerns in ML Systems:

Scalability (and efficiency at scale)

Usability

Q: How are the features and models configured?

DB-style Practical Concerns in ML

Key concerns in ML:

Accuracy

Runtime efficiency (sometimes)

Additional key *practical* concerns in ML Systems:

Scalability (and efficiency at scale)

Usability

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

DB-style Practical 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

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

DB-style Practical 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

Q: How to simplify the implementation of such systems?

DB-style Practical 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 to simplify the implementation of such systems?

DB-style Practical 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

DB-style Practical 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

} *Long-standing concerns in the **DB systems** world!*

DB-style Practical 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

} *Long-standing concerns in the **DB systems** world!*

Can often trade off accuracy a bit to gain on the rest!

The Push to “Platformize” ML/AI

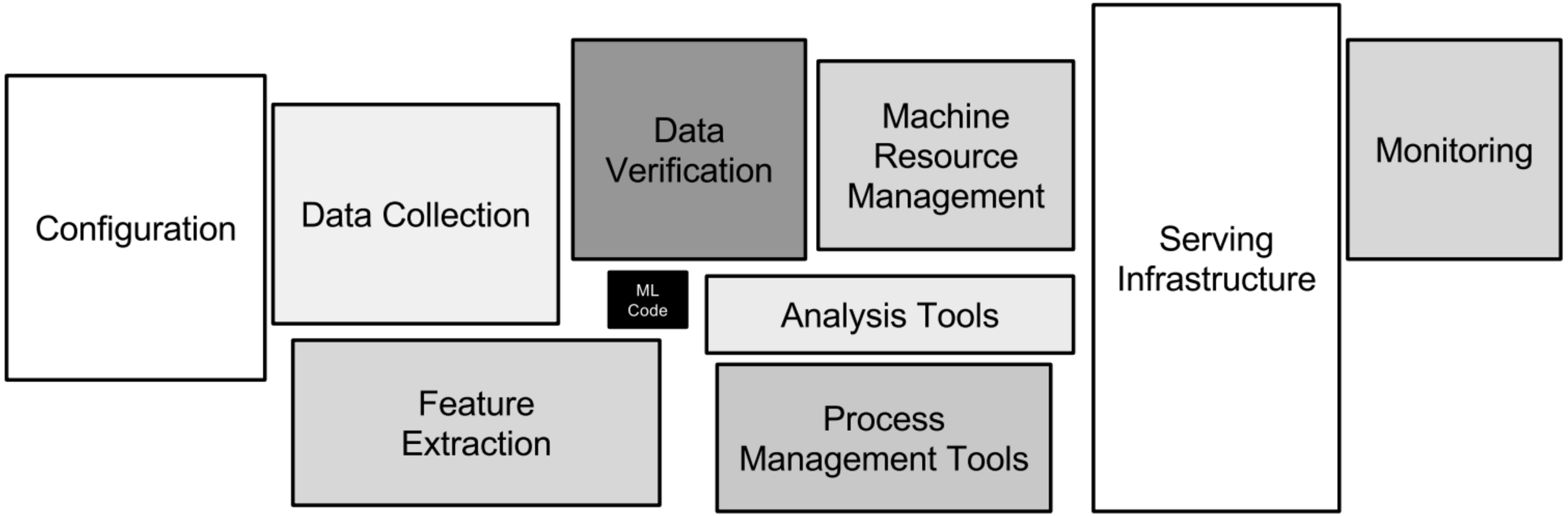


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Google

Hidden Technical Debt in Machine Learning Systems. ***NIPS’15***

A Brief History of “Platformizing” ML

A Brief History of “Platformizing” ML

1980s

S

A Brief History of “Platformizing” ML

Mid
1990s

1980s

S

A Brief History of “Platformizing” ML

Mid
1990s

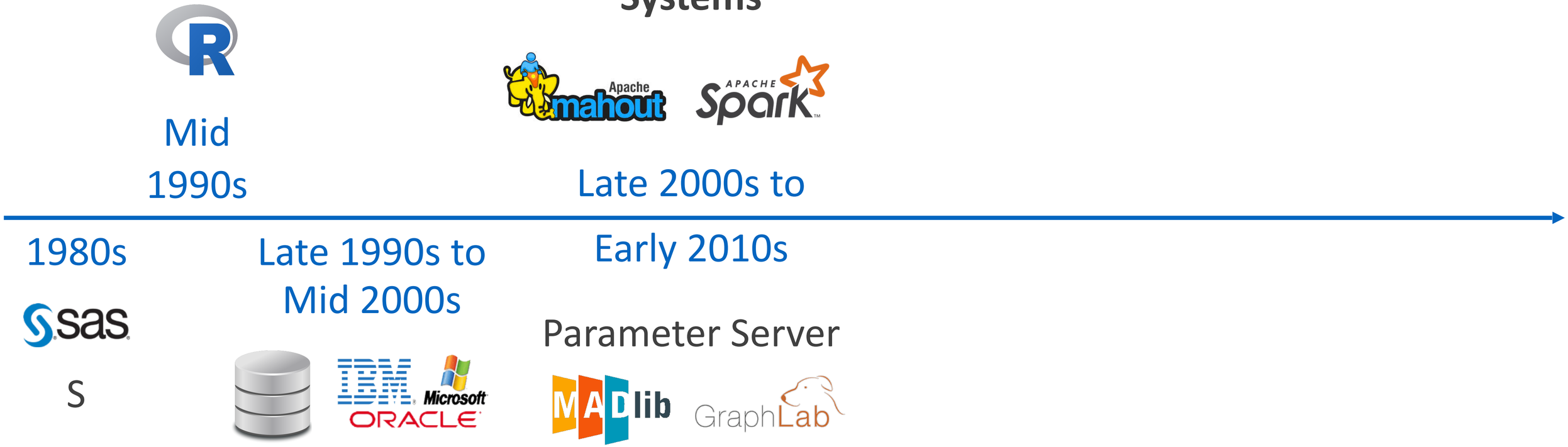
1980s

S

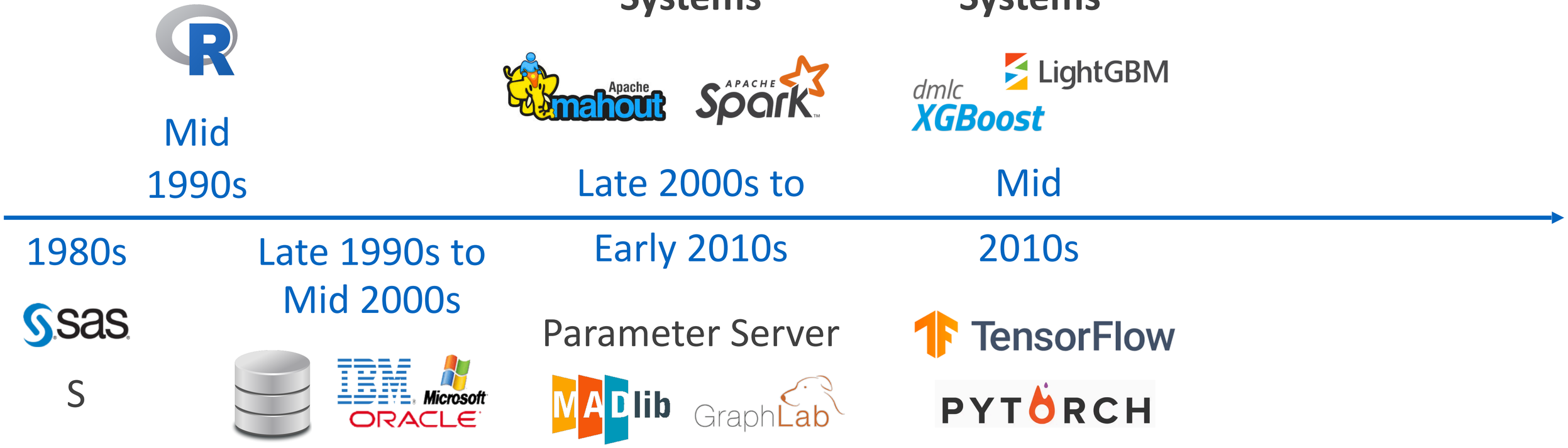
Late 1990s to
Mid 2000s

In-RDBMS
ML Systems

A Brief History of “Platformizing” ML



A Brief History of “Platformizing” ML



A Brief History of “Platformizing” ML

1980s

S

Mid 1990s

Late 1990s to Mid 2000s

In-RDBMS ML Systems

Late 2000s to Early 2010s

ML on Dataflow Systems

Parameter Server

ML System Abstractions

Mid 2010s

Tree Learning Systems

Deep Learning Systems

Late 2010s Onwards

Amazon SageMaker

Cloud ML/AI Services

ML Platforms and Feature Stores

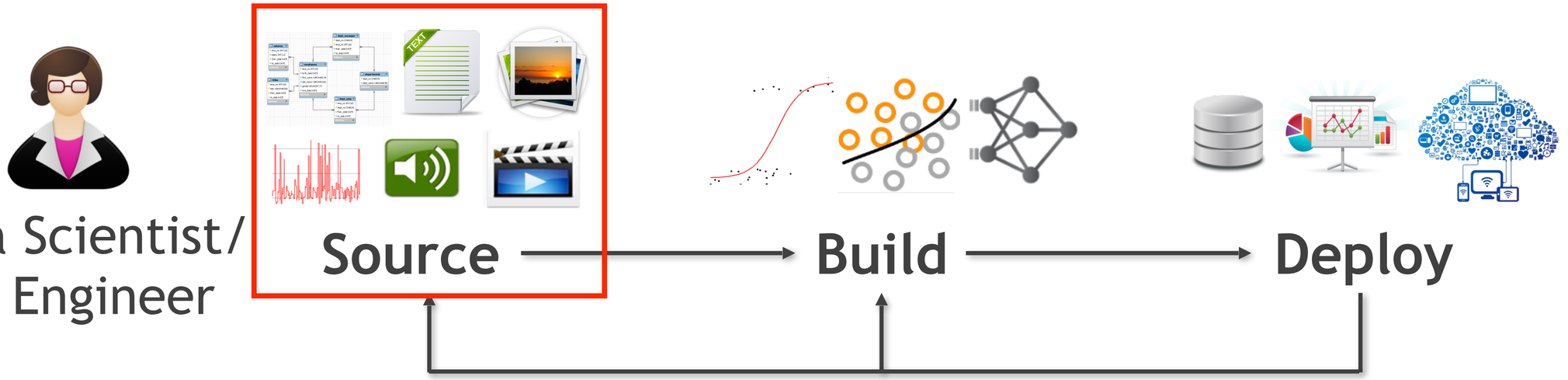
The way I see it, the rise of ML systems/ platforms today resembles the rise of RDBMSs circa early 1980s

The New DBfication of ML/AI

Data Scientist/
ML Engineer

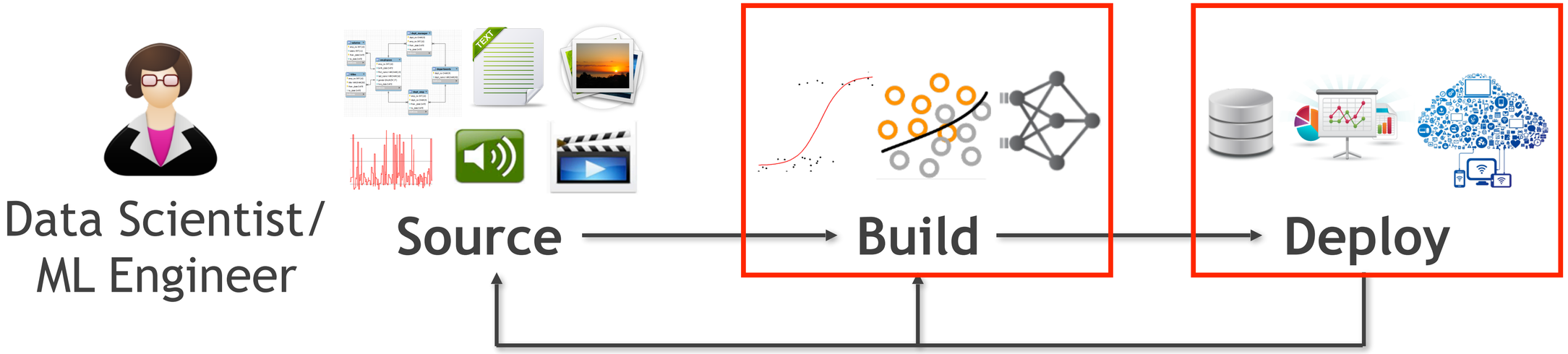
Source → Build → Deploy

The New DBfication of ML/AI



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

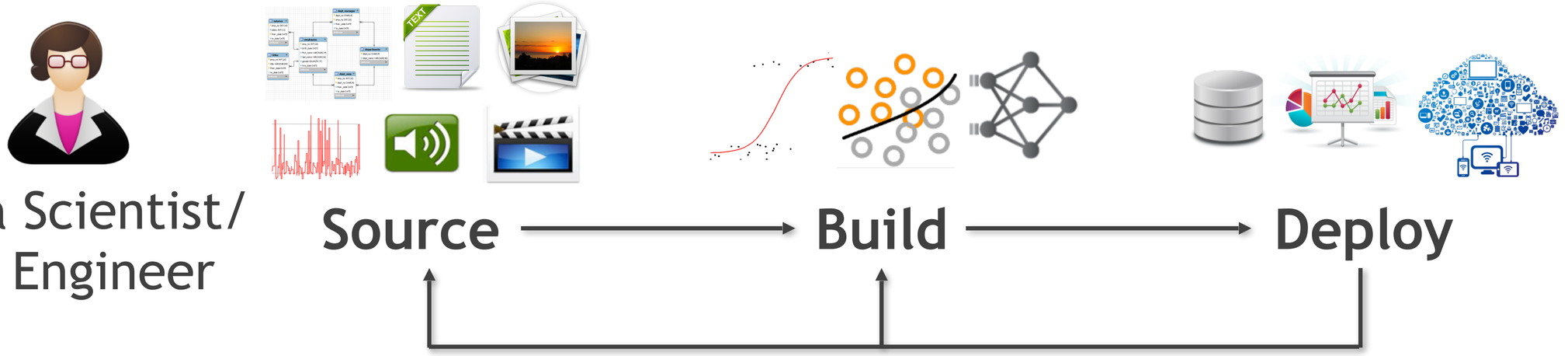
The New DBfication of ML/AI



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
…

The New DBfication of ML/AI



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.

Outline

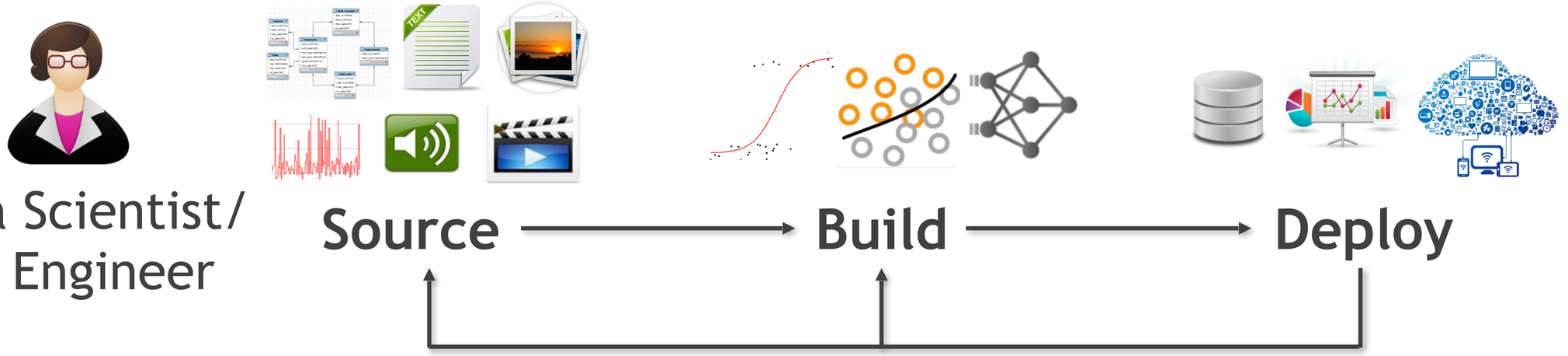
- The New DBfication of ML/AI
- Two Examples from My Research
- Accelerating the DBfication of ML/AI

The New DBfication of ML/AI

Data Scientist/
ML Engineer

Source → Build → Deploy

The New DBfication of ML/AI

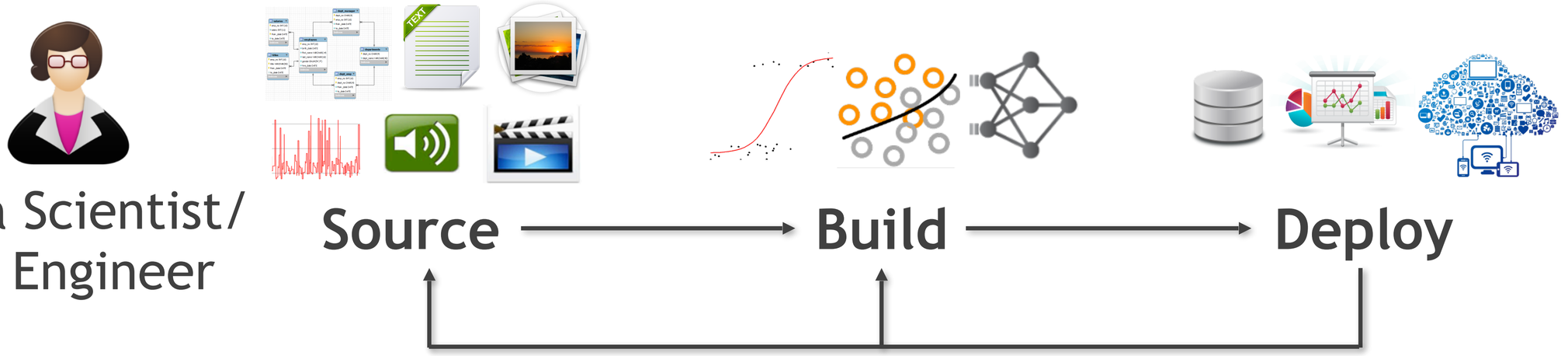


Metadata Management for ML
Data Prep/Cleaning for ML
Multimodal ML Query Models
Data Search, Labeling, etc.
…

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

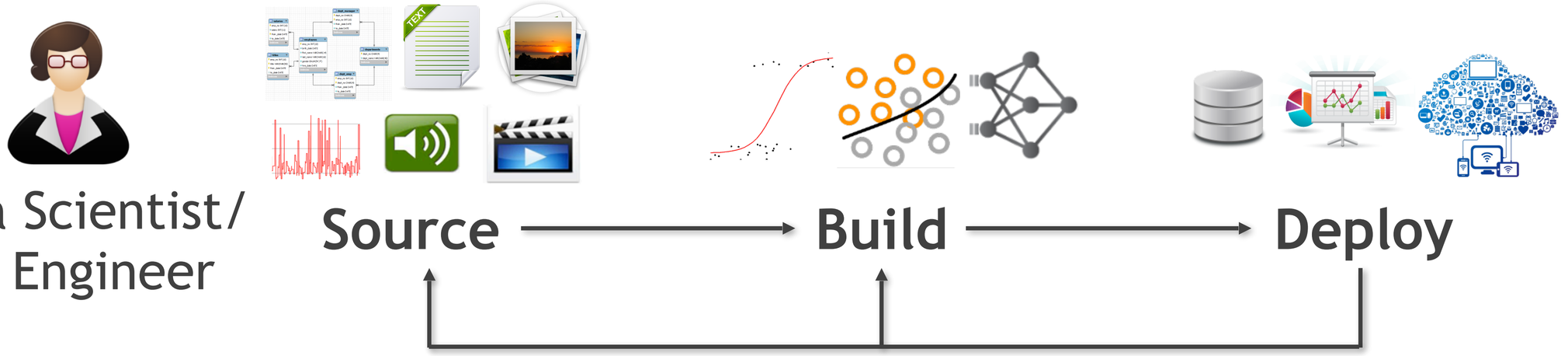


Metadata Management for ML
Data Prep/Cleaning for ML
Multimodal ML Query Models
Data Search, Labeling, etc.
…

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

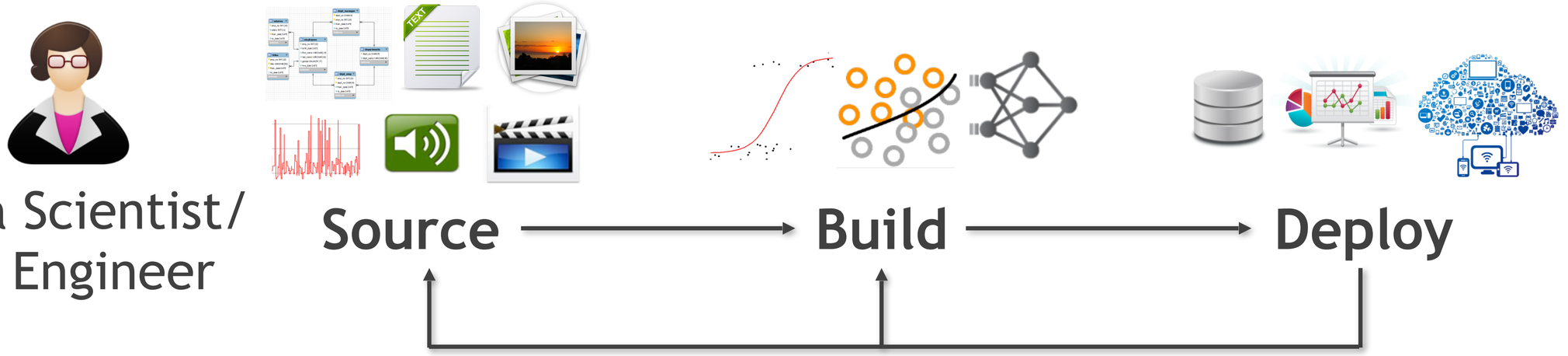


Metadata Management for ML
Data Prep/Cleaning for ML
Multimodal ML Query Models
Data Search, Labeling, etc.
…

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



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.

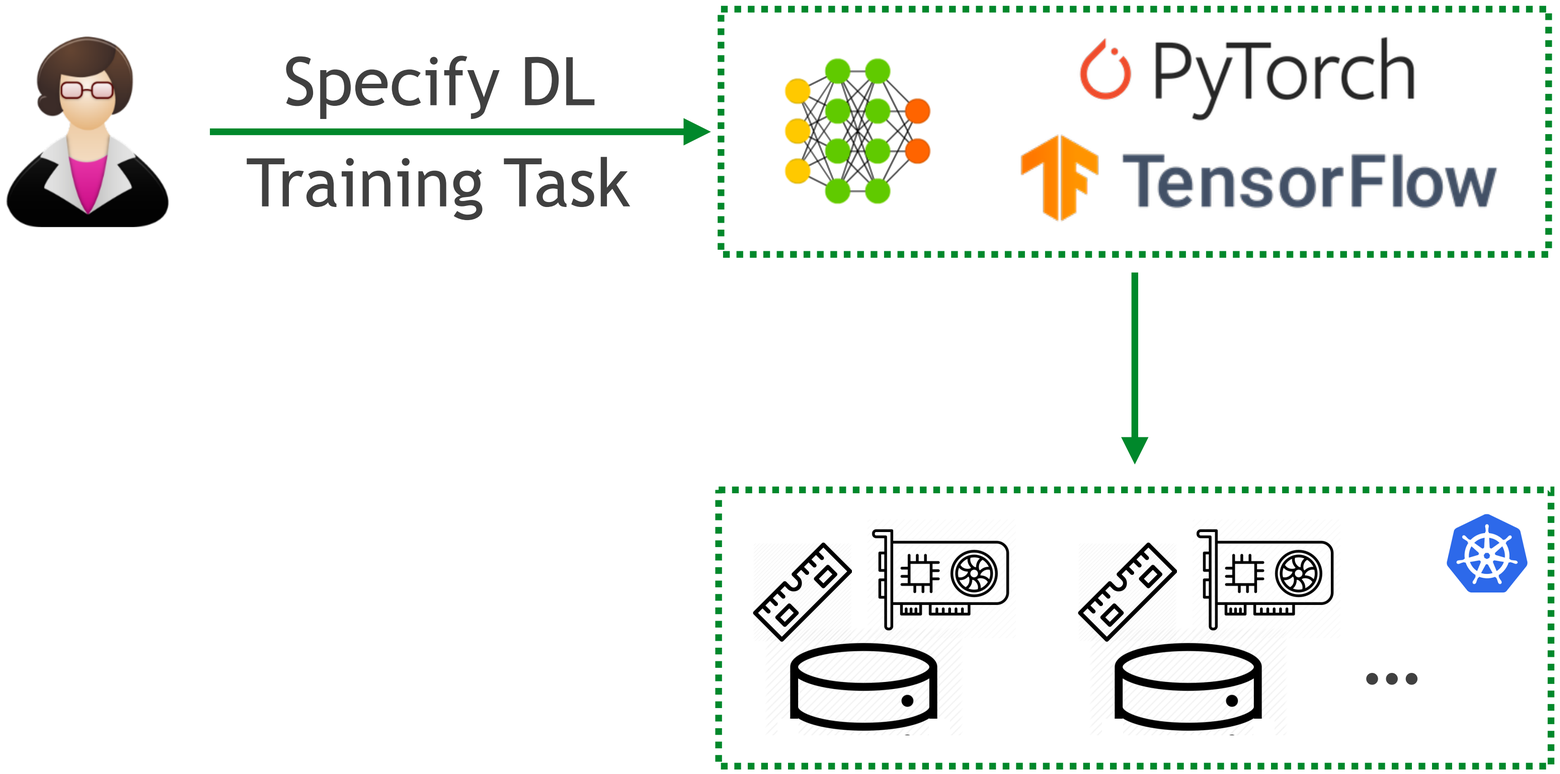
Outline

- The New DBfication of ML/AI
- Two Examples from My Research
  - Example 1: Scalable DL Systems
  - Example 2: Auto Data Prep for ML
- Accelerating the DBfication of ML/AI

Outline

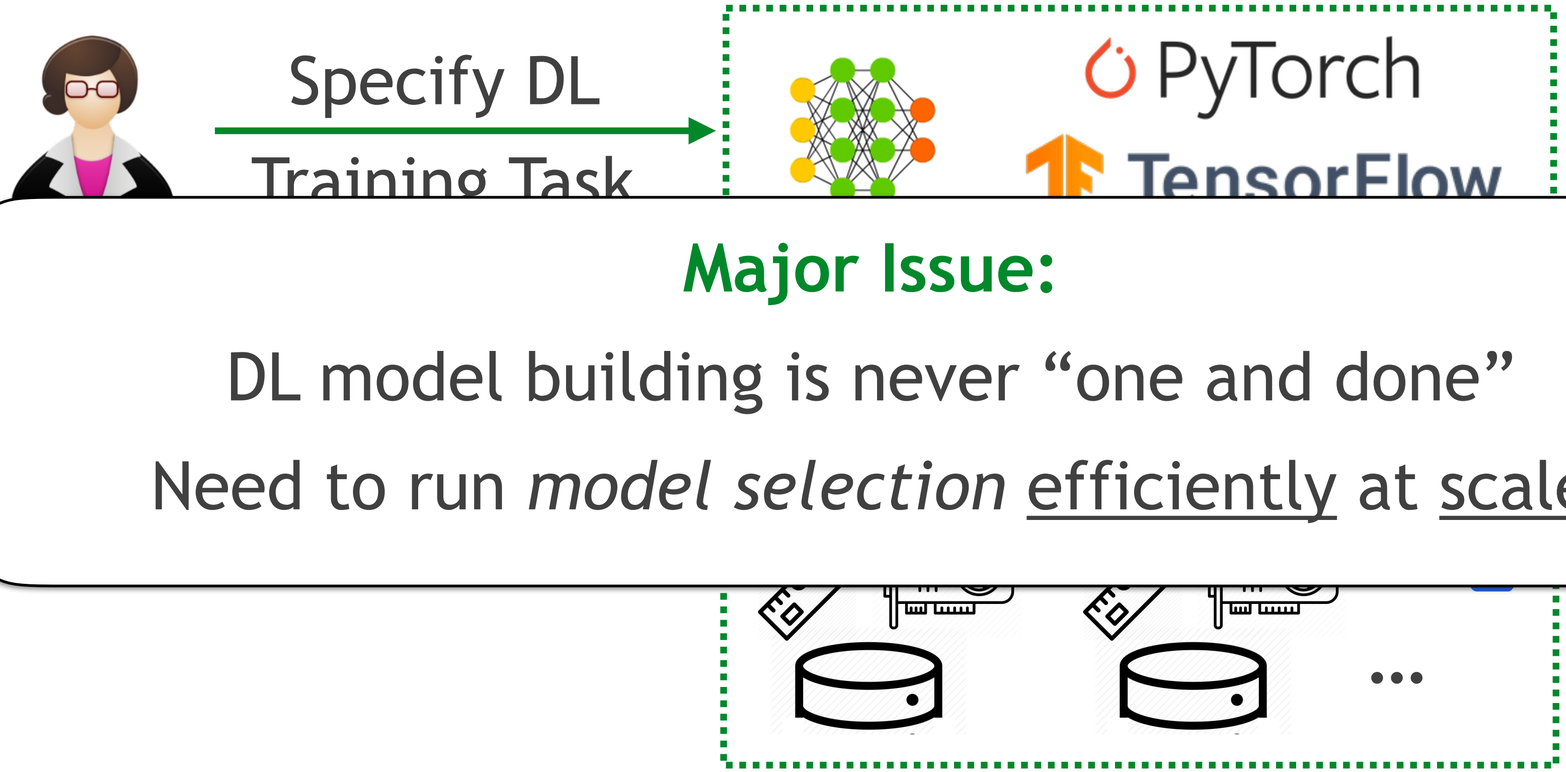
- The New DBfication of ML/AI
- Two Examples from My Research
  - Example 1: Scalable DL Systems
  - Example 2: Auto Data Prep for ML
- Accelerating the DBfication of ML/AI

Example 1: DL Systems at Scale



Cluster: GPU+CPU, memory, disks

Example 1: DL Systems at Scale



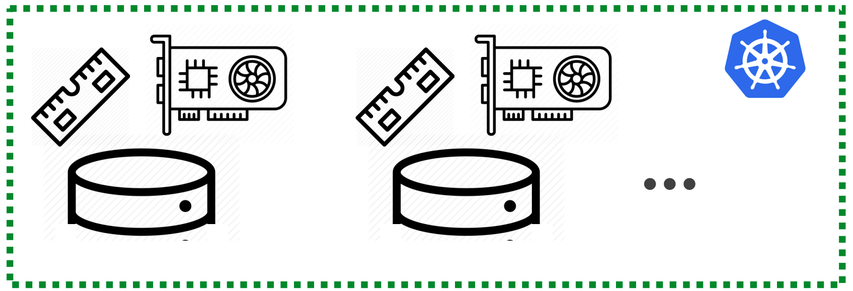
Major Issue:

DL model building is never “one and done”

Need to run *model selection* efficiently at scale

Cluster: GPU+CPU, memory, disks

Project Cerebro



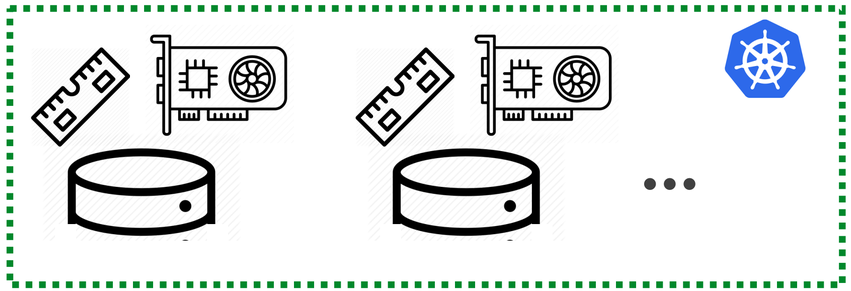
Plain/Kubernetes cluster

Cerebro: A Layered Data Platform for Scalable Deep Learning. ***CIDR'21***

Project Cerebro

Specify DL Model

Selection Holistically



Plain/Kubernetes cluster

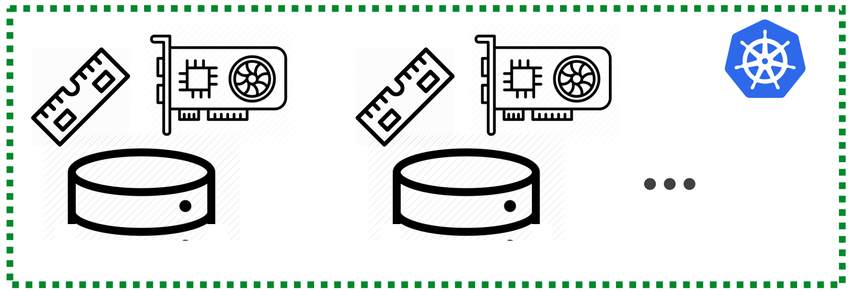
Cerebro: A Layered Data Platform for Scalable Deep Learning. ***CIDR'21***

Project Cerebro

Specify DL Model

Selection Holistically

Magic of Multi-Query Optimization for DL



Plain/Kubernetes cluster

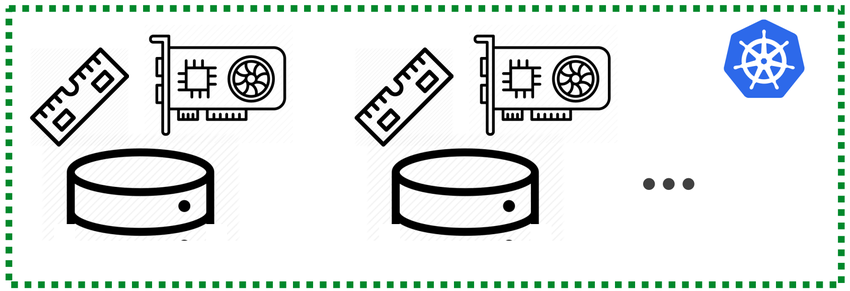
Cerebro: A Layered Data Platform for Scalable Deep Learning. ***CIDR'21***

Project Cerebro

Specify DL Model

Selection Holistically

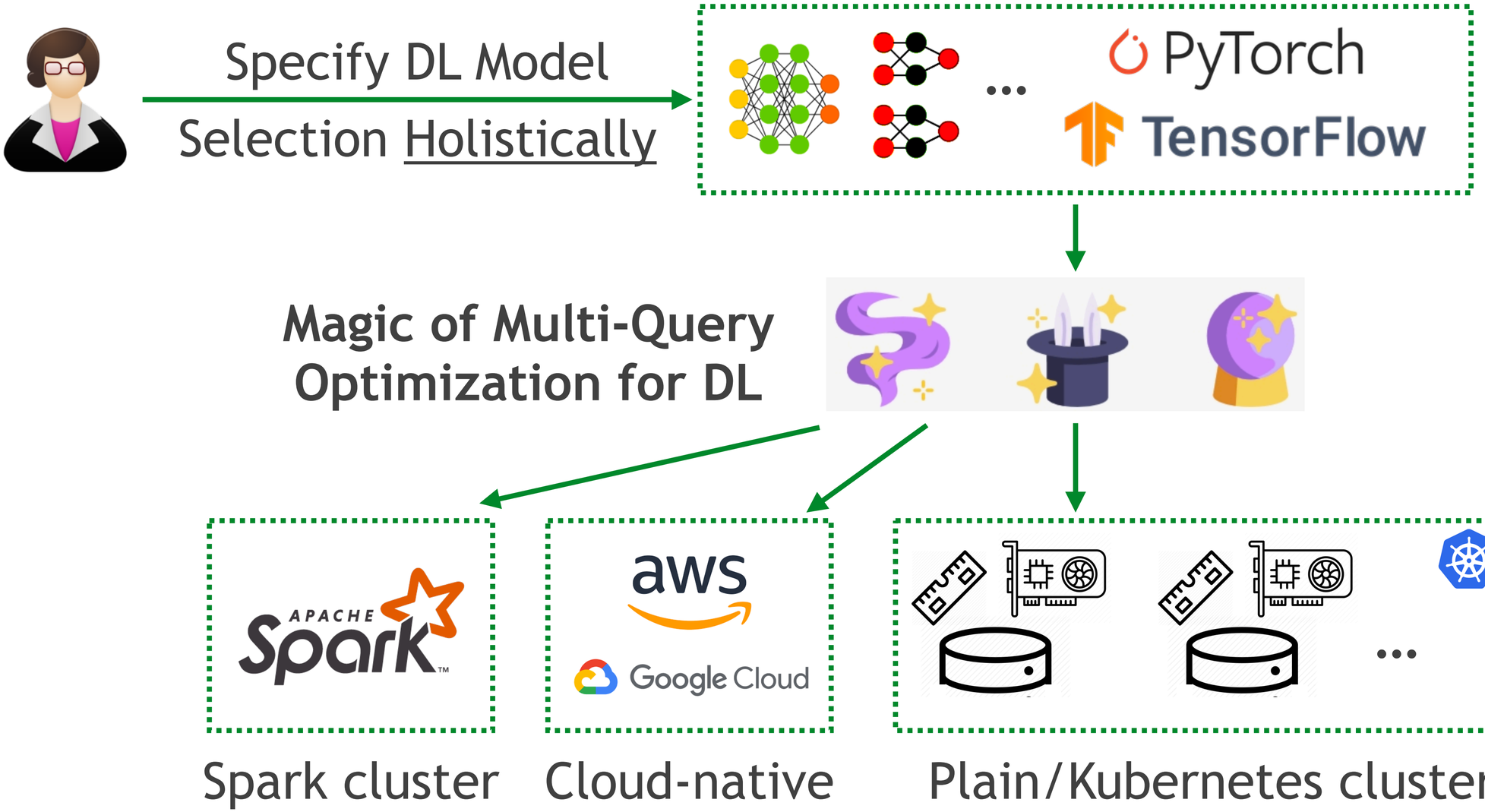
Magic of Multi-Query Optimization for DL



Plain/Kubernetes cluster

Cerebro: A Layered Data Platform for Scalable Deep Learning. ***CIDR'21***

Project Cerebro



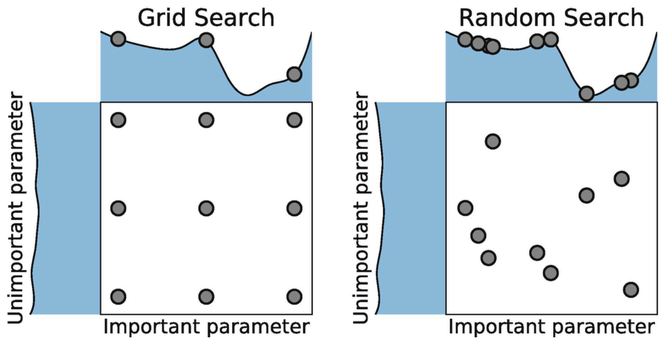
Cerebro: A Layered Data Platform for Scalable Deep Learning. ***CIDR'21***

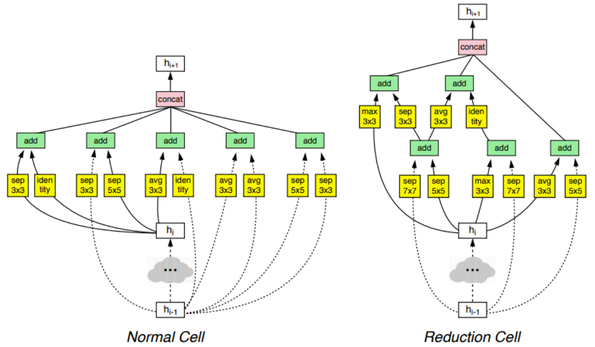
Key Example: MQO for DL Model Selection

Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

Key Example: MQO for DL Model Selection

Workload: Hyper-parameter tuning and neural architecture exploration yield numerous (even 100s) of DL training configs



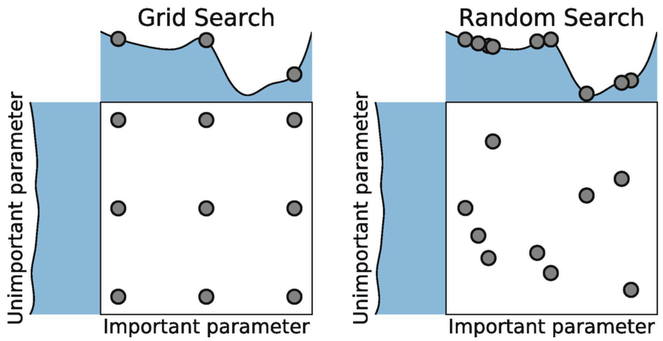


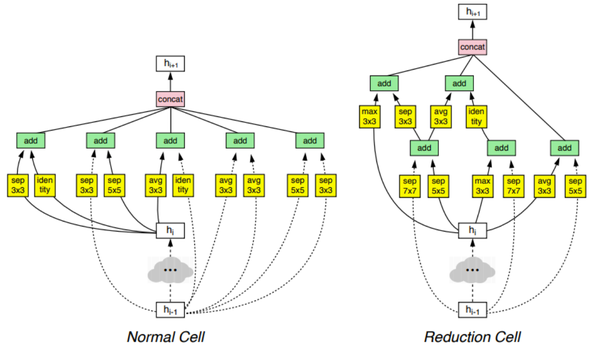
HyperBand, AutoKeras, ASHA, PBT, HyperOpt, NAS, etc.

Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

Key Example: MQO for DL Model Selection

Workload: Hyper-parameter tuning and neural architecture exploration yield numerous (even 100s) of DL training configs





HyperBand, AutoKeras, ASHA, PBT, HyperOpt, NAS, etc.

Multi-Query: Run multiple configs on large dataset in one go

Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

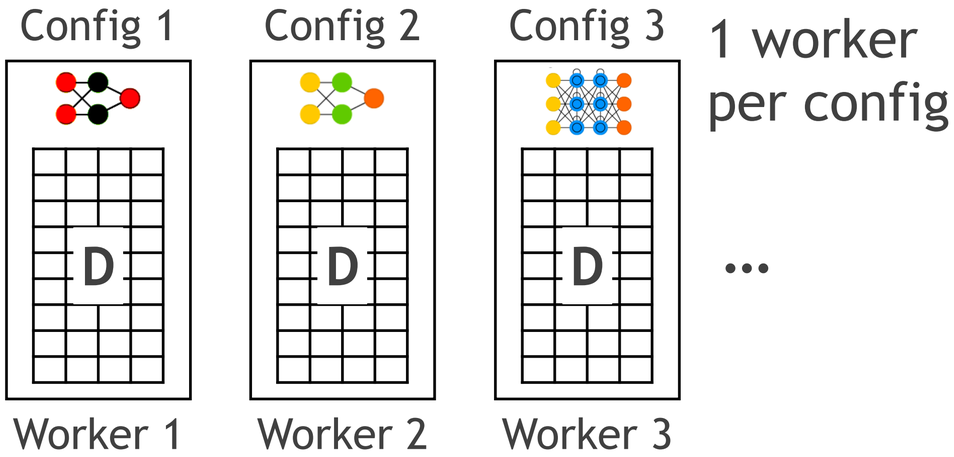
Positioning Cerebro's Technique vs Prior Art

We devise a novel execution strategy called **Model Hopper Parallelism (MOP)**

Positioning Cerebro's Technique vs Prior Art

We devise a novel execution strategy called **Model Hopper Parallelism (MOP)**

Task Parallelism:



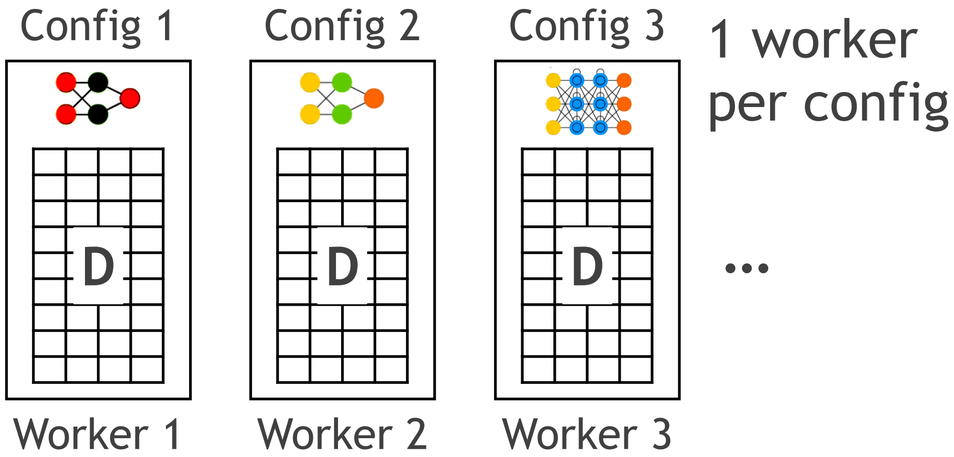
1 worker per config

...

Positioning Cerebro's Technique vs Prior Art

We devise a novel execution strategy called **Model Hopper Parallelism (MOP)**

Task Parallelism:

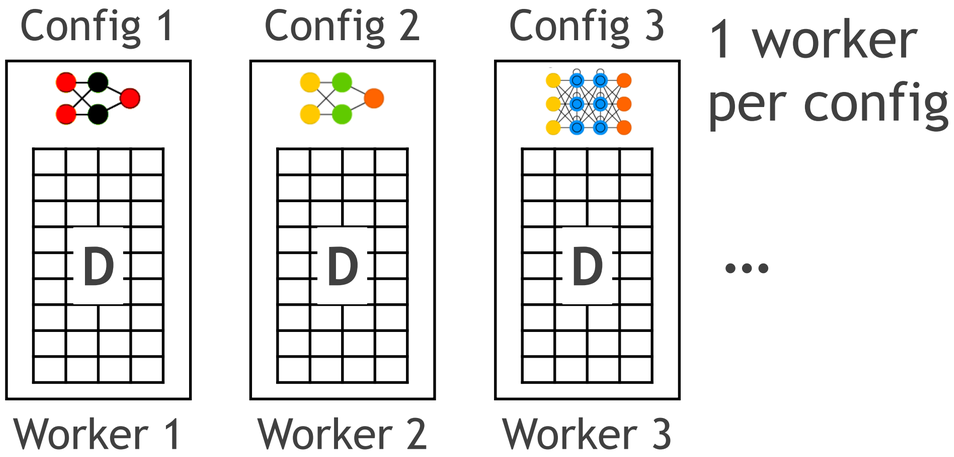


+ High throughput model selection
+ Best accuracy from Sequential SGD
— Low data scalability; wastes space (copy) or network (remote read)

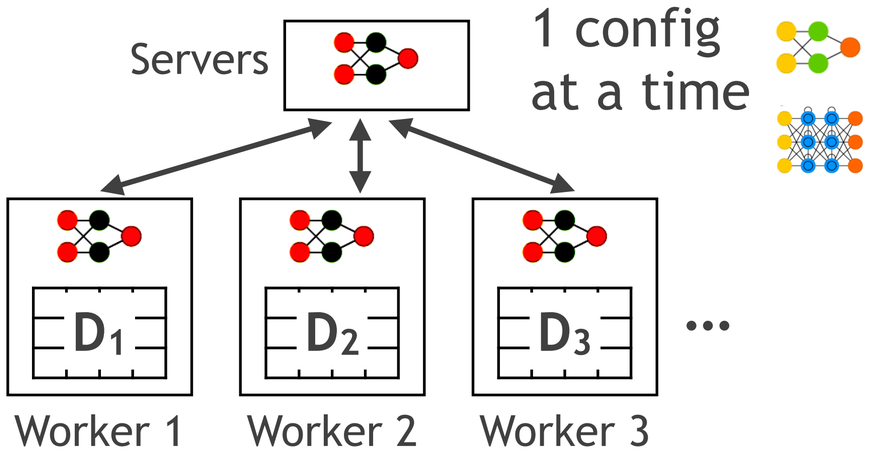
Positioning Cerebro's Technique vs Prior Art

We devise a novel execution strategy called **Model Hopper Parallelism (MOP)**

Task Parallelism:



Data Parallelism:

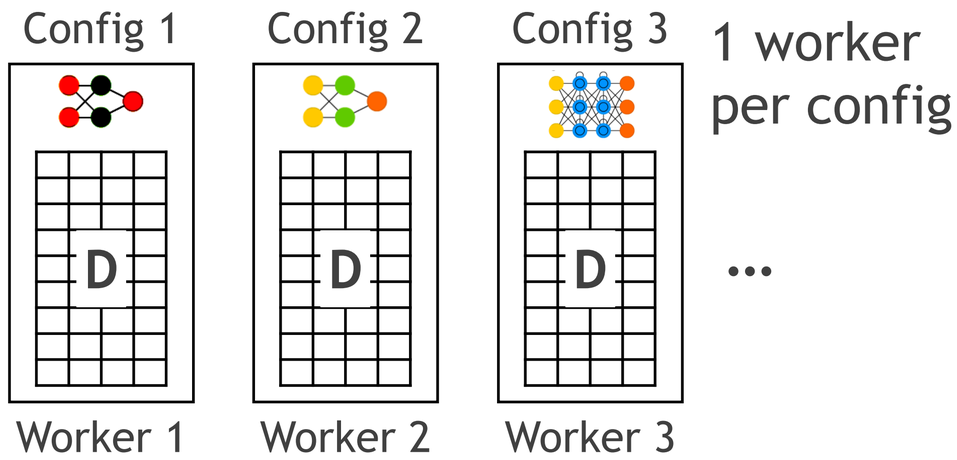


+ High throughput model selection
+ Best accuracy from Sequential SGD
— Low data scalability; wastes space (copy) or network (remote read)

Positioning Cerebro's Technique vs Prior Art

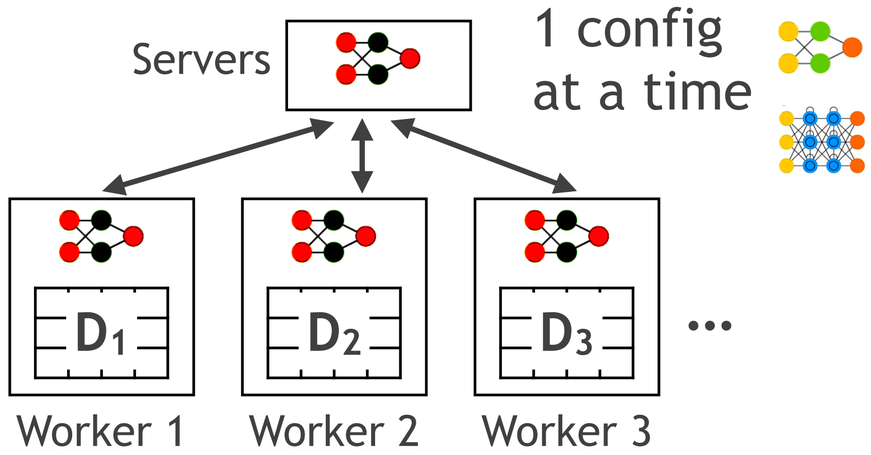
We devise a novel execution strategy called **Model Hopper Parallelism (MOP)**

Task Parallelism:



+ High throughput model selection
+ Best accuracy from Sequential SGD
— Low data scalability; wastes space (copy) or network (remote read)

Data Parallelism:

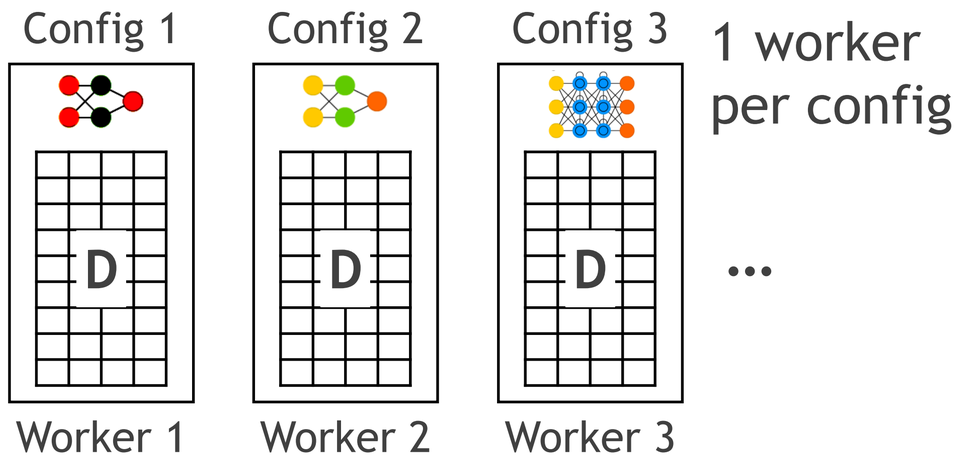


+ High data scalability via sharding
— Shard-level SGD does not converge; mini-batch level is too slow
— Very low throughput overall

Positioning Cerebro's Technique vs Prior Art

We devise a novel execution strategy called **Model Hopper Parallelism (MOP)**

Task Parallelism:



1 worker per config

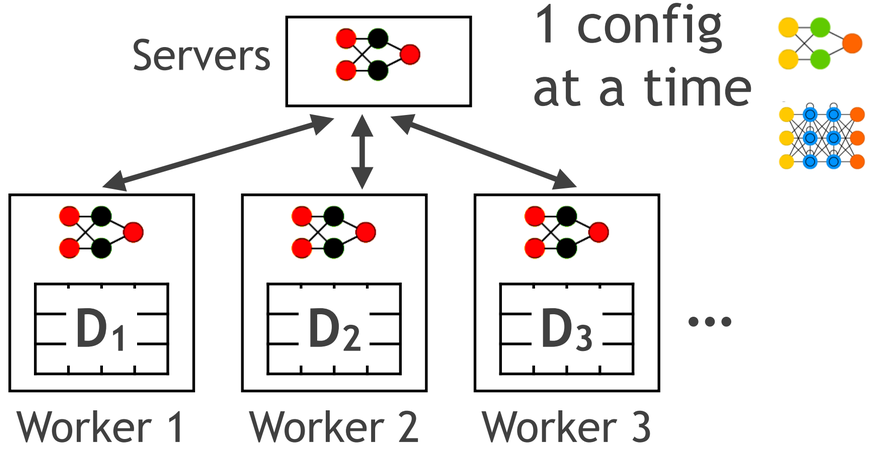
...

+ High throughput model selection
+ Best accuracy from Sequential SGD
— Low data scalability; wastes space (copy) or network (remote read)

MOP:

New hybrid of data and task parallelism
Best of both worlds!

Data Parallelism:



...

+ High data scalability via sharding
— Shard-level SGD does not converge; mini-batch level is too slow
— Very low throughput overall

Model Hopper Parallelism (MOP)

Insight from Optimization Theory:

SGD is robust to *data ordering randomness*

Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

Model Hopper Parallelism (MOP)

Insight from Optimization Theory:

SGD is robust to *data ordering randomness*

Shuffle and partition dataset
Run *n* DNNs on *n* workers

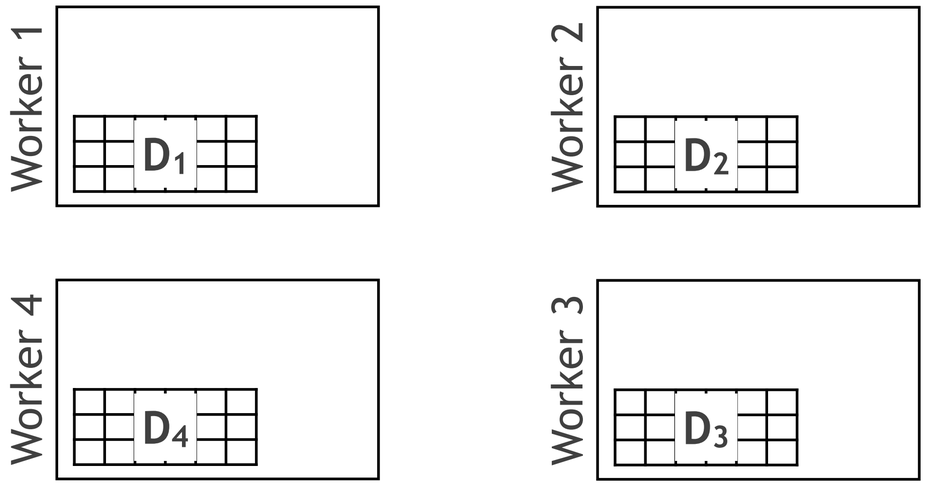
Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

Model Hopper Parallelism (MOP)

Insight from Optimization Theory:

SGD is robust to *data ordering randomness*

Shuffle and partition dataset
Run *n* DNNs on *n* workers



Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

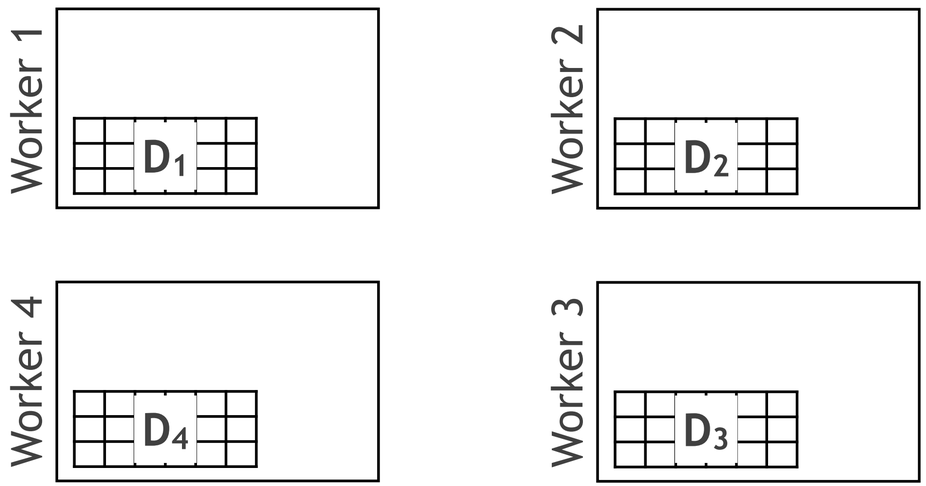
Model Hopper Parallelism (MOP)

Insight from Optimization Theory:

SGD is robust to *data ordering randomness*

Shuffle and partition dataset
Run *n* DNNs on *n* workers

Epoch 1.1 starts in parallel



Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

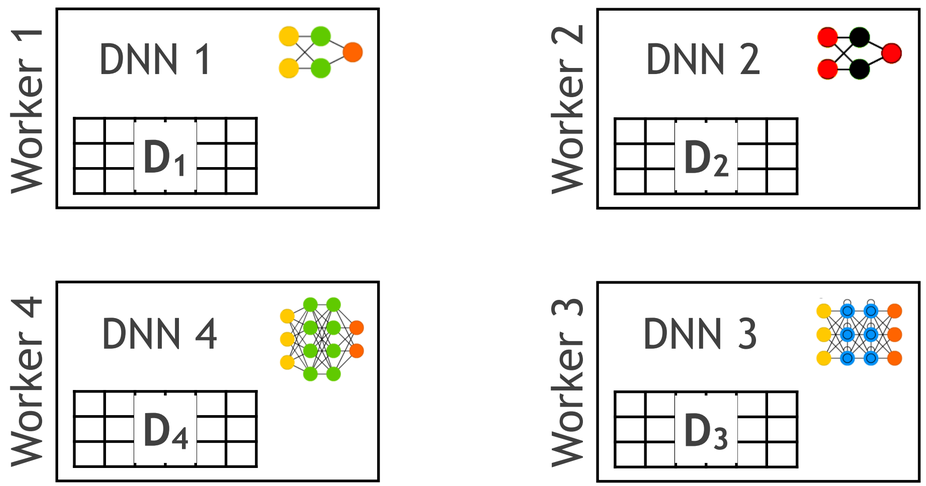
Model Hopper Parallelism (MOP)

Insight from Optimization Theory:

SGD is robust to *data ordering randomness*

Shuffle and partition dataset
Run *n* DNNs on *n* workers

Epoch 1.1 starts in parallel



Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

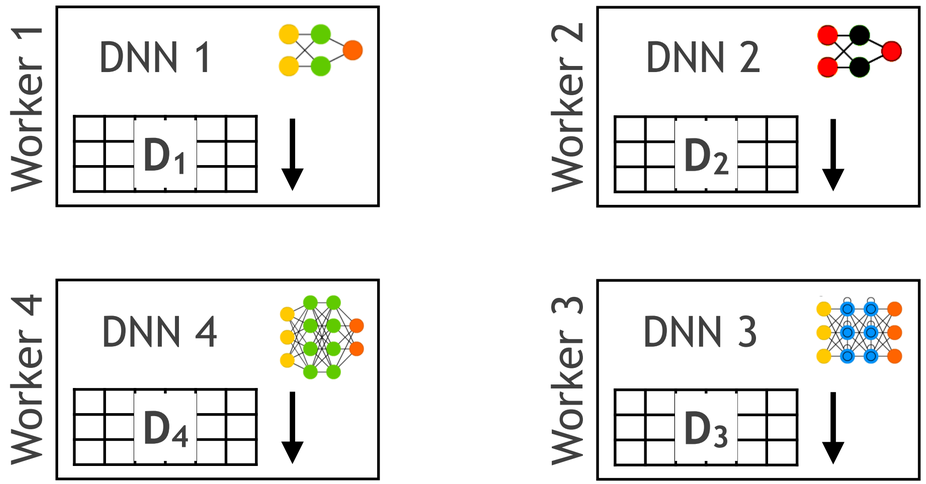
Model Hopper Parallelism (MOP)

Insight from Optimization Theory:

SGD is robust to *data ordering randomness*

Shuffle and partition dataset
Run *n* DNNs on *n* workers

Epoch 1.1 starts in parallel



Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

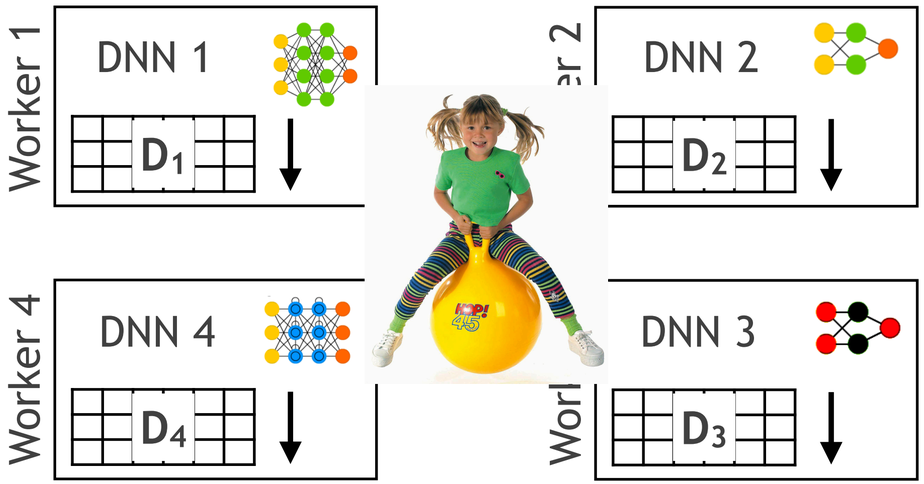
Model Hopper Parallelism (MOP)

Insight from Optimization Theory:

SGD is robust to *data ordering randomness*

Shuffle and partition dataset
Run *n* DNNs on *n* workers

Epoch 1.1 starts in parallel



Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

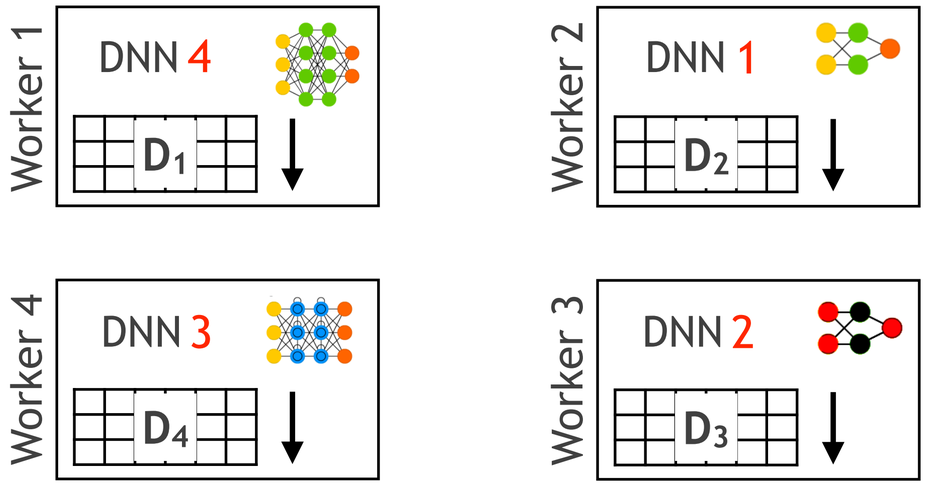
Model Hopper Parallelism (MOP)

Insight from Optimization Theory:

SGD is robust to *data ordering randomness*

Shuffle and partition dataset
Run *n* DNNs on *n* workers

Epoch 1.1 starts in parallel



Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

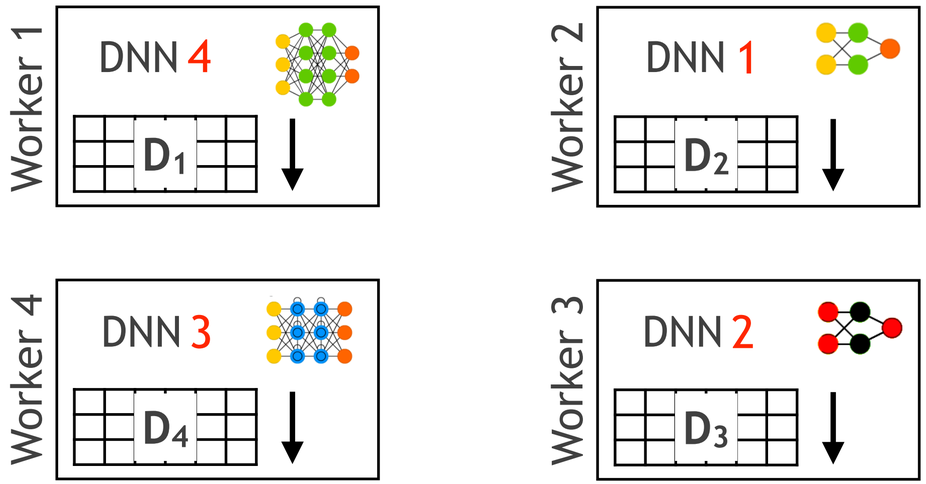
Model Hopper Parallelism (MOP)

Insight from Optimization Theory:

SGD is robust to *data ordering randomness*

Shuffle and partition dataset
Run *n* DNNs on *n* workers

Epoch 1.2 starts in parallel



Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

Model Hopper Parallelism (MOP)

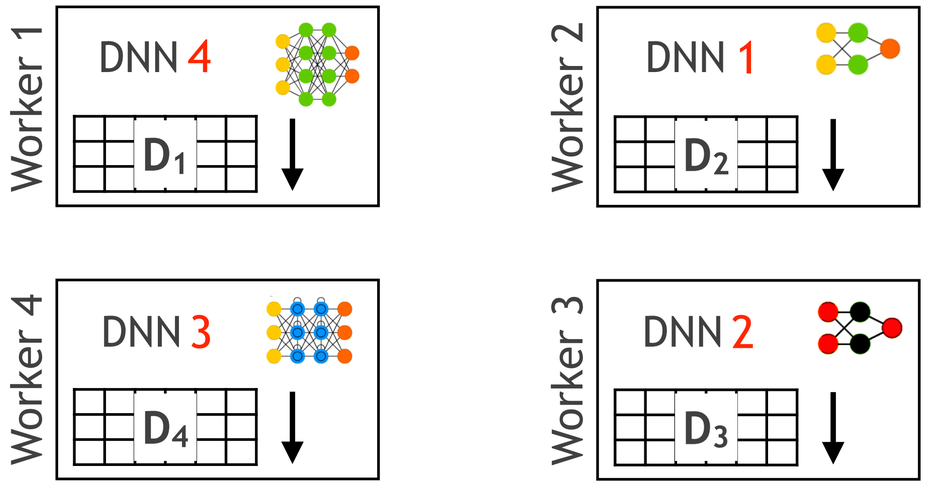
Insight from Optimization Theory:

SGD is robust to *data ordering randomness*

Shuffle and partition dataset
Run *n* DNNs on *n* workers

Each model keeps "hopping" until it sees all of D

Epoch 1.2 starts in parallel



Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

Model Hopper Parallelism (MOP)

Insight from Optimization Theory:

SGD is robust to *data ordering randomness*

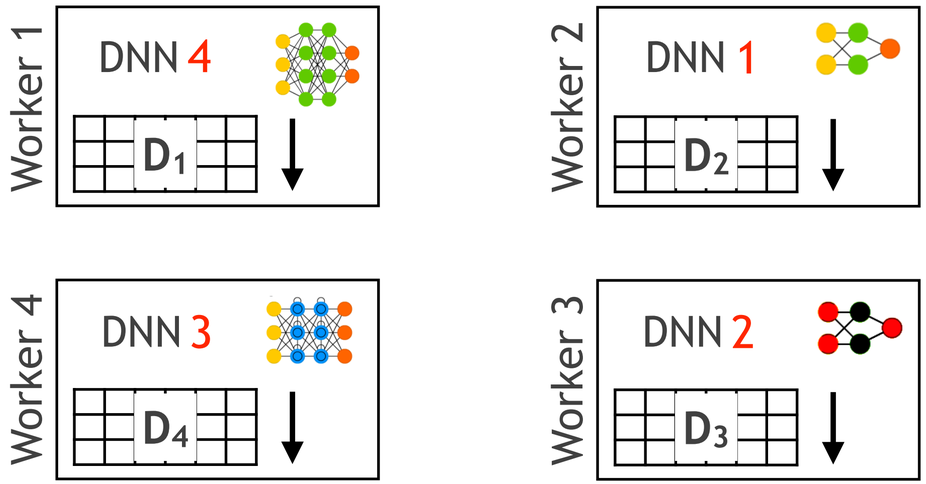
Shuffle and partition dataset
Run *n* DNNs on *n* workers

Each model keeps "hopping" until it sees all of D

Strong theoretical guarantees:

1. *Equivalence* to sequential SGD
2. Hits lower bound on comm. cost

Epoch 1.2 starts in parallel



Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

Model Hopper Parallelism (MOP)

Insight from Optimization Theory:

SGD is robust to *data ordering randomness*

Technical Challenges (See paper for details):

Resource-aware scheduling of evolving model configs' hops

Support data replication, fault tolerance, and elasticity

Non-disruptive integration with existing DL systems

1. *Equivalence* to sequential SGD
2. Hits lower bound on comm. cost

Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

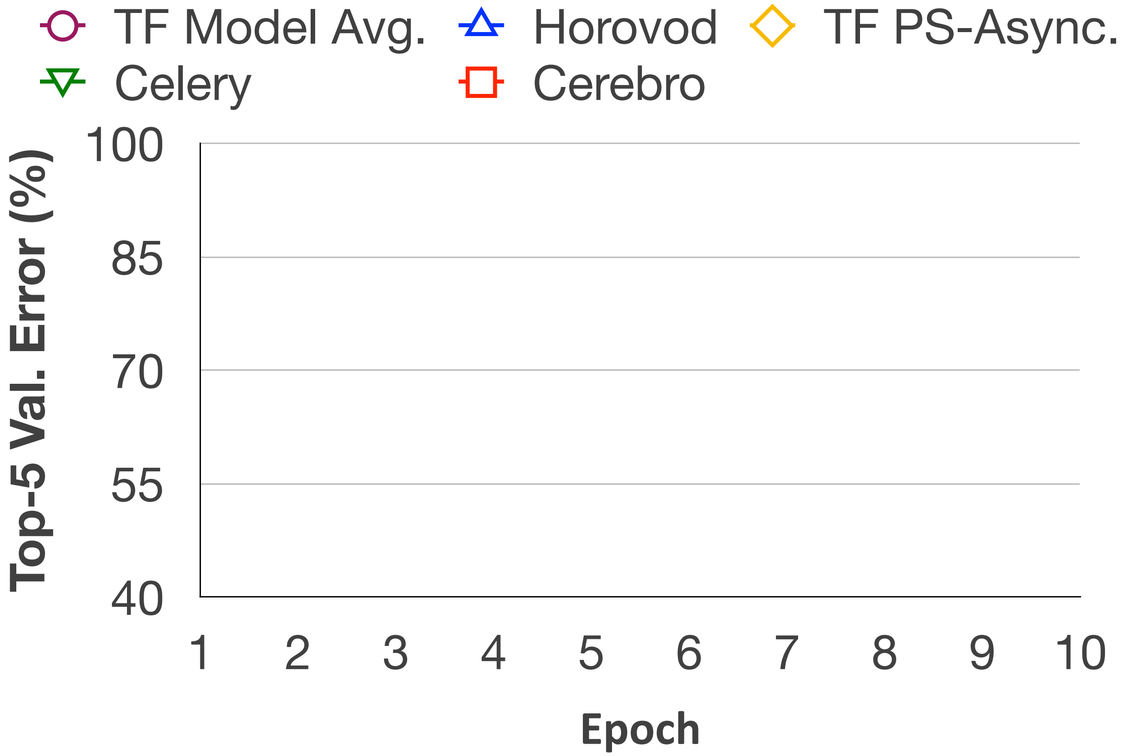
Experimental Evaluation

Setup: ImageNet; 16 CNN configurations; TensorFlow; 8 GPU nodes

Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

Experimental Evaluation

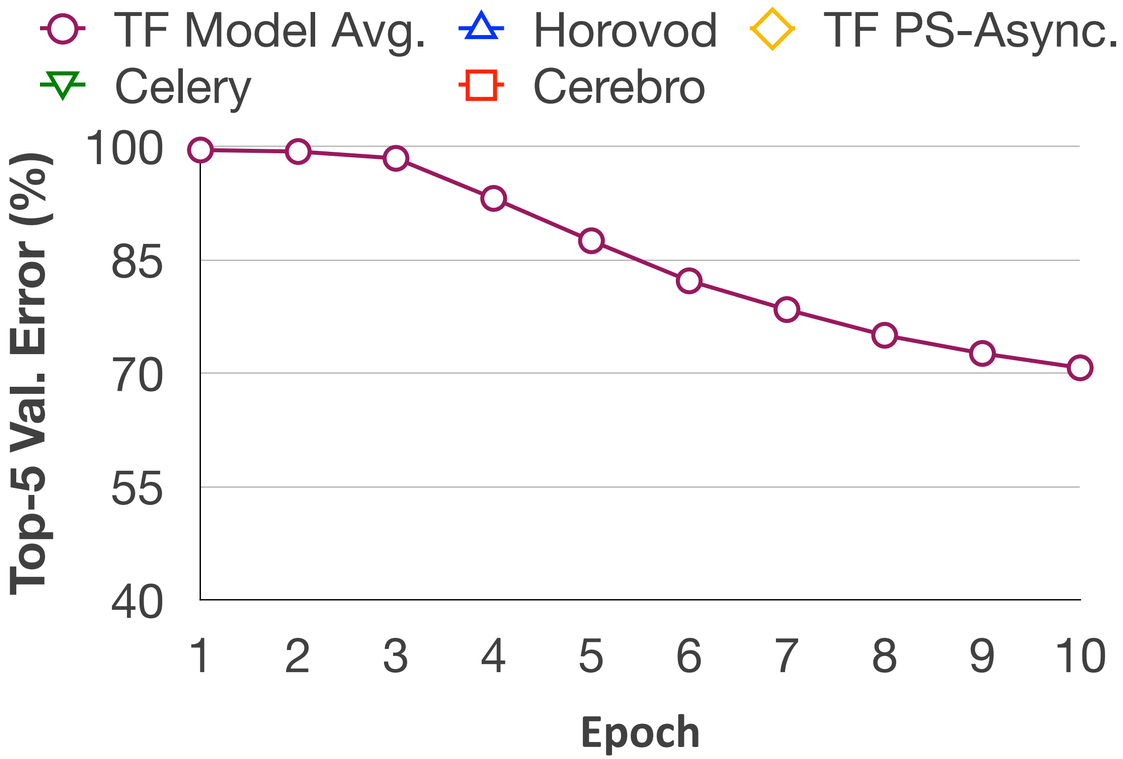
Setup: ImageNet; 16 CNN configurations; TensorFlow; 8 GPU nodes



Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

Experimental Evaluation

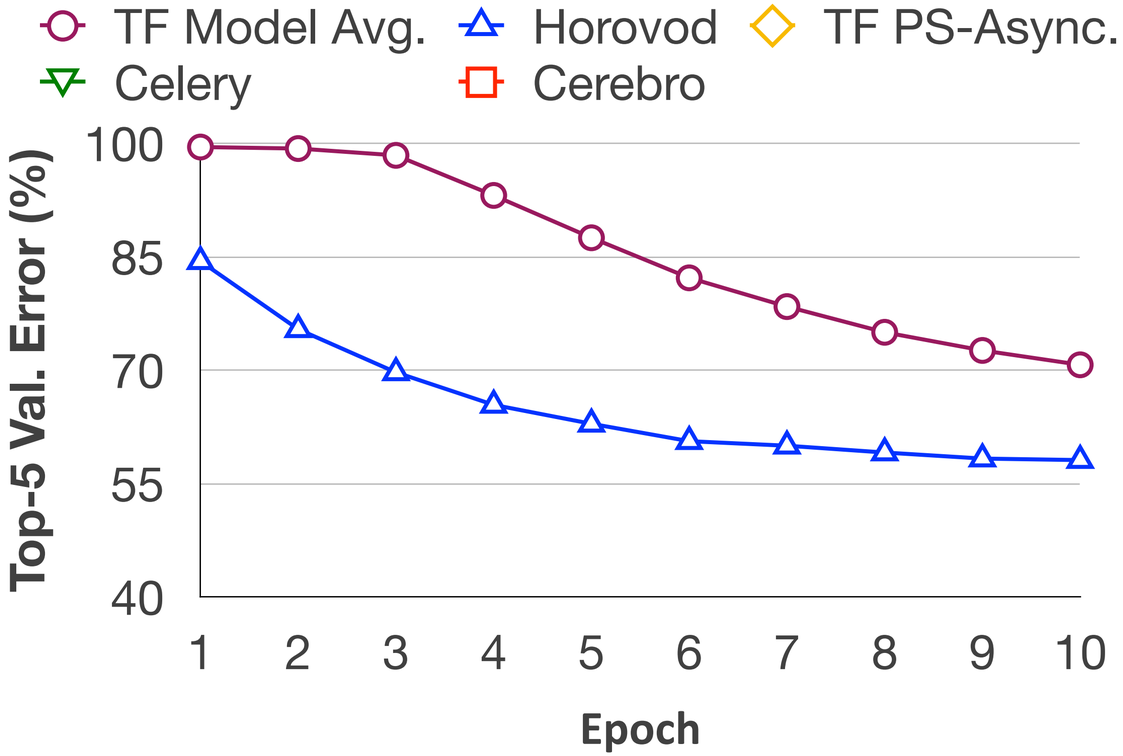
Setup: ImageNet; 16 CNN configurations; TensorFlow; 8 GPU nodes



Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

Experimental Evaluation

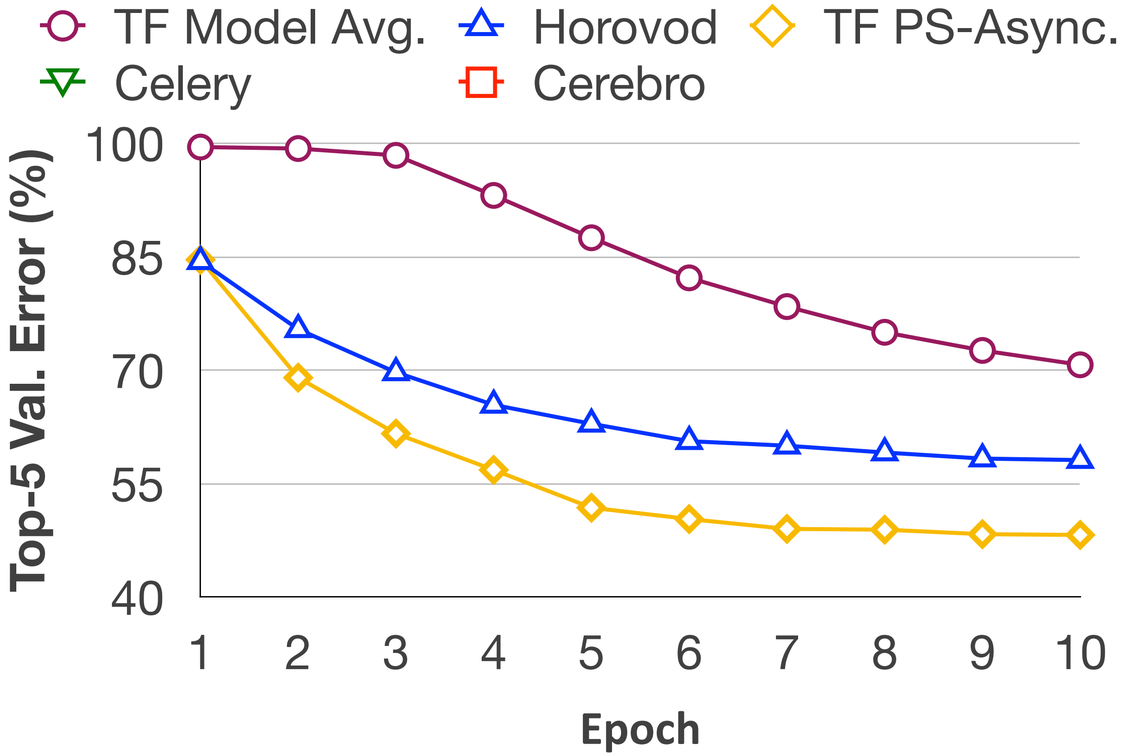
Setup: ImageNet; 16 CNN configurations; TensorFlow; 8 GPU nodes



Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

Experimental Evaluation

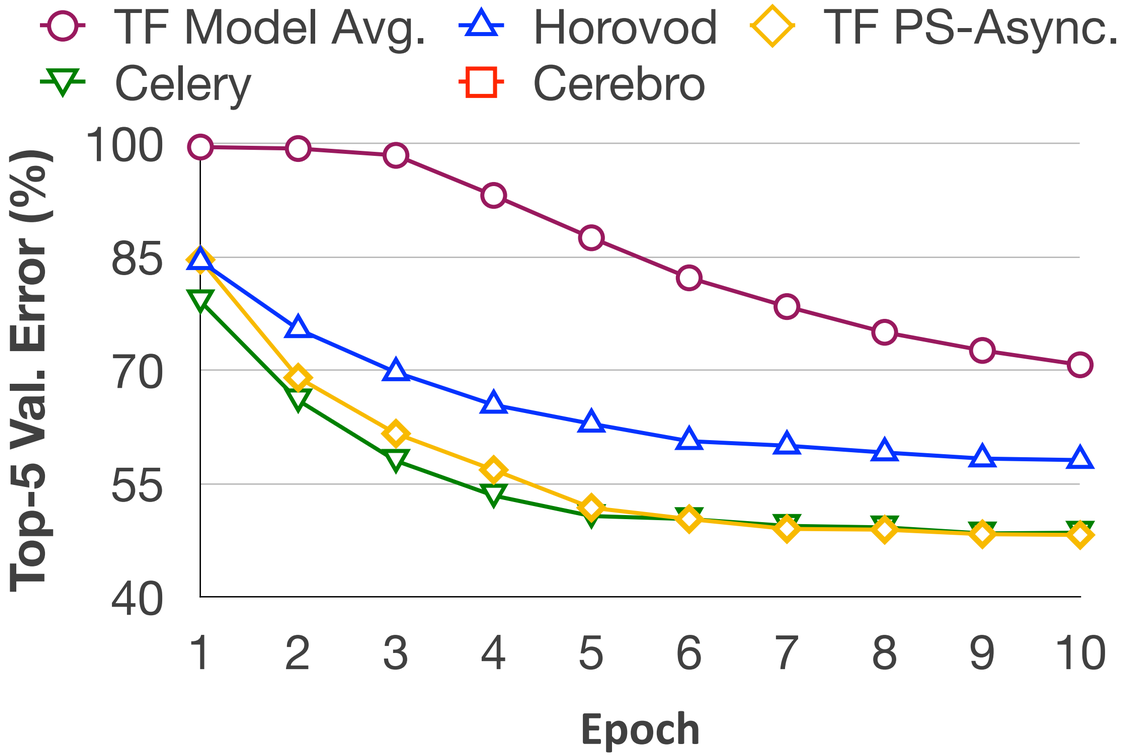
Setup: ImageNet; 16 CNN configurations; TensorFlow; 8 GPU nodes



Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

Experimental Evaluation

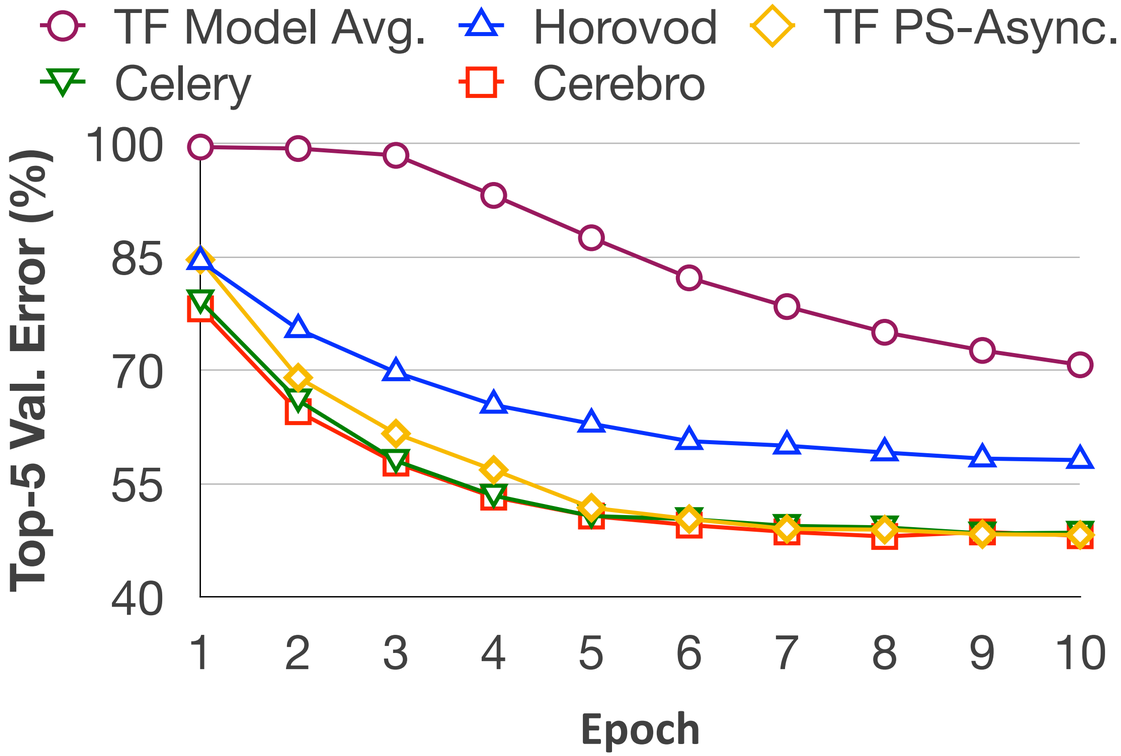
Setup: ImageNet; 16 CNN configurations; TensorFlow; 8 GPU nodes



Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

Experimental Evaluation

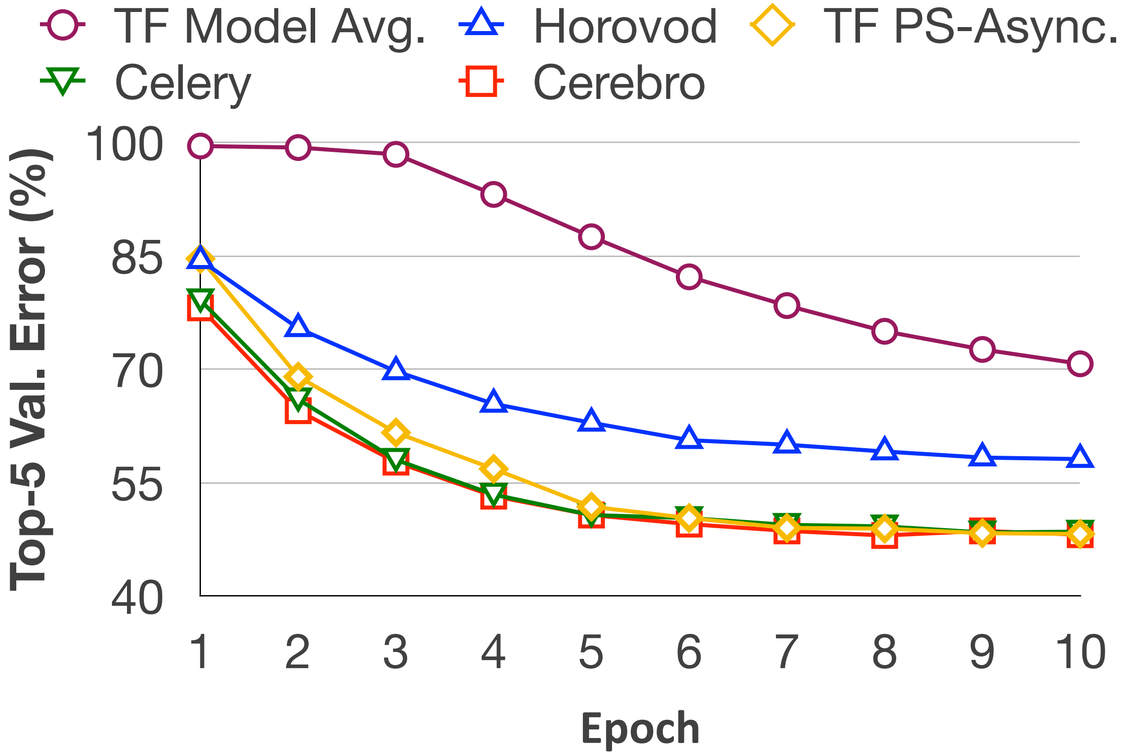
Setup: ImageNet; 16 CNN configurations; TensorFlow; 8 GPU nodes



Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

Experimental Evaluation

Setup: ImageNet; 16 CNN configurations; TensorFlow; 8 GPU nodes

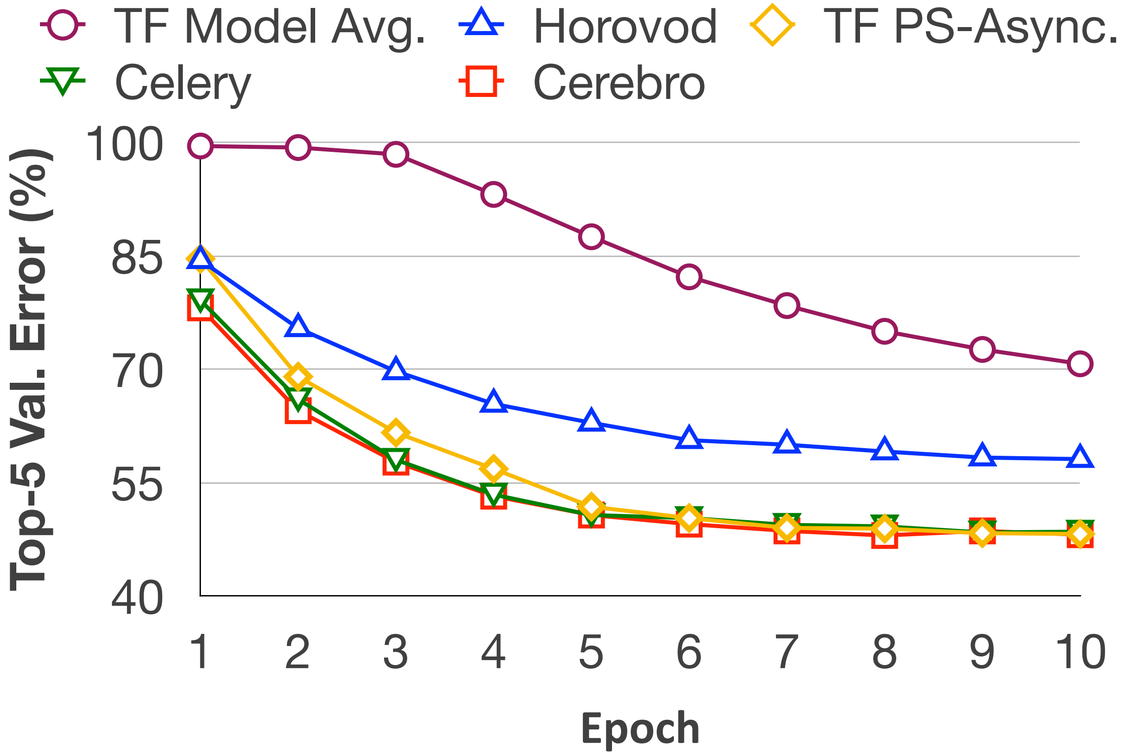


Method	Runtime (hrs)	Memory
TF PS-Async.	190.0 (!)	200 GB
Horovod	54.2	200 GB
TF Model Averaging	19.7	200 GB
Celery (Task-Parallel)	17.2-19.0*	1600 GB (!)
MOP/Cerebro	17.7	200 GB

Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

Experimental Evaluation

Setup: ImageNet; 16 CNN configurations; TensorFlow; 8 GPU nodes

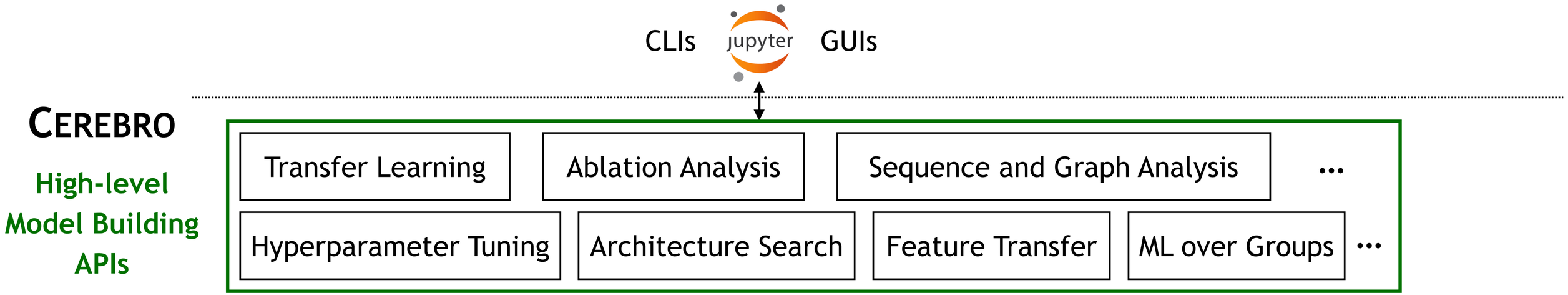


Method	Runtime (hrs)	Memory
TF PS-Async.	190.0 (!)	200 GB
Horovod	54.2	200 GB
TF Model Averaging	19.7	200 GB
Celery (Task-Parallel)	17.2-19.0*	1600 GB (!)
MOP/Cerebro	17.7	200 GB

Cerebro offers overall best combination of resource efficiency and accuracy

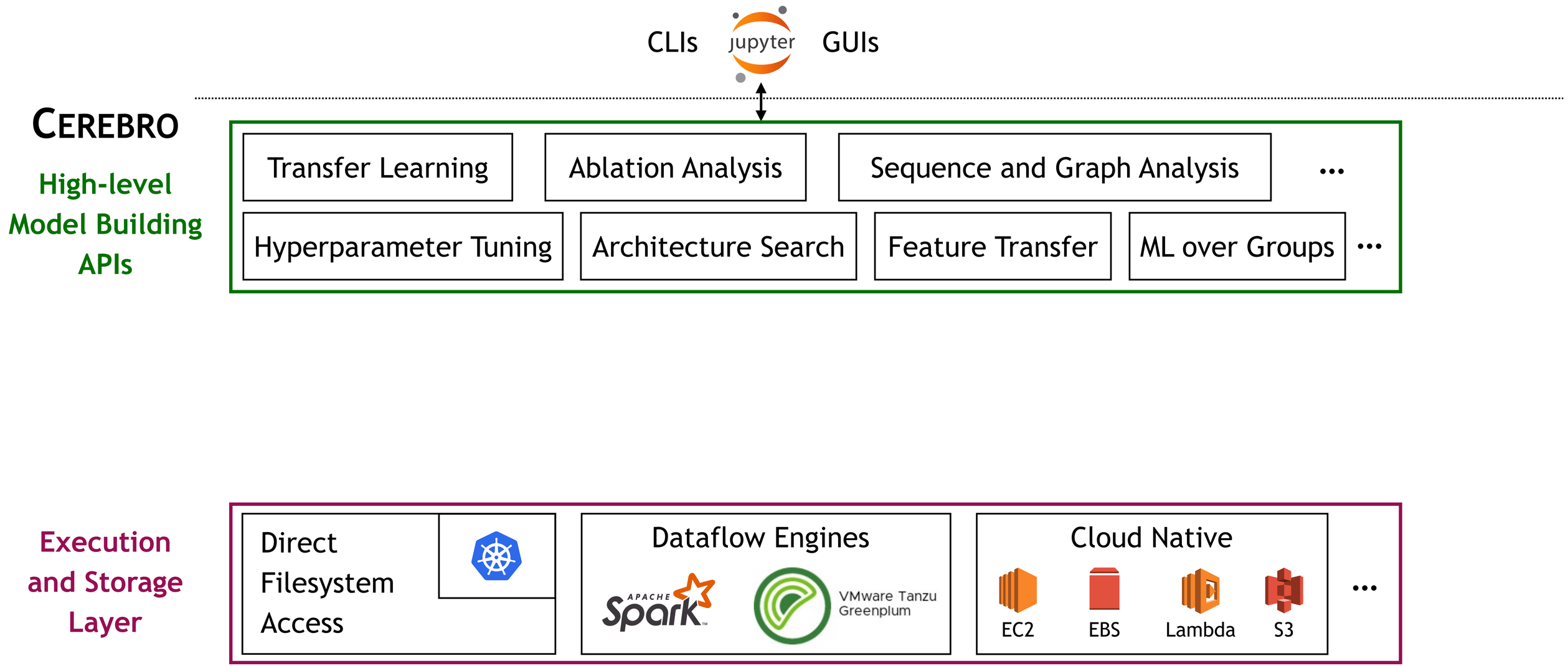
Cerebro: Efficient and Reproducible Model Selection on Deep Learning Systems. ***SIGMOD DEEM'19***
Cerebro: A Data System for Optimized Deep Learning Model Selection. ***VLDB'20***

Full Vision of the Cerebro Platform



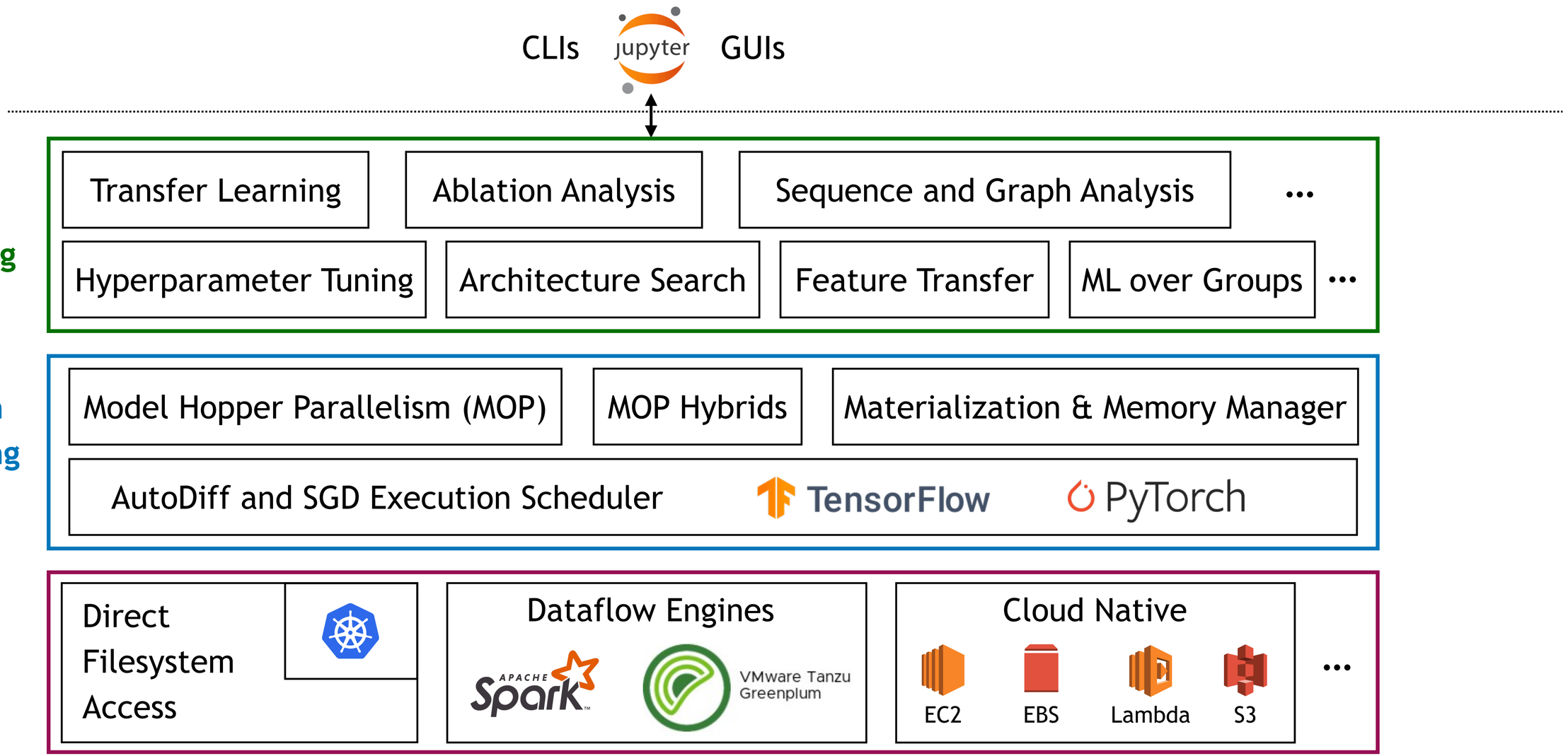
Cerebro: A Layered Data Platform for Scalable Deep Learning. ***CIDR'21***

Full Vision of the Cerebro Platform



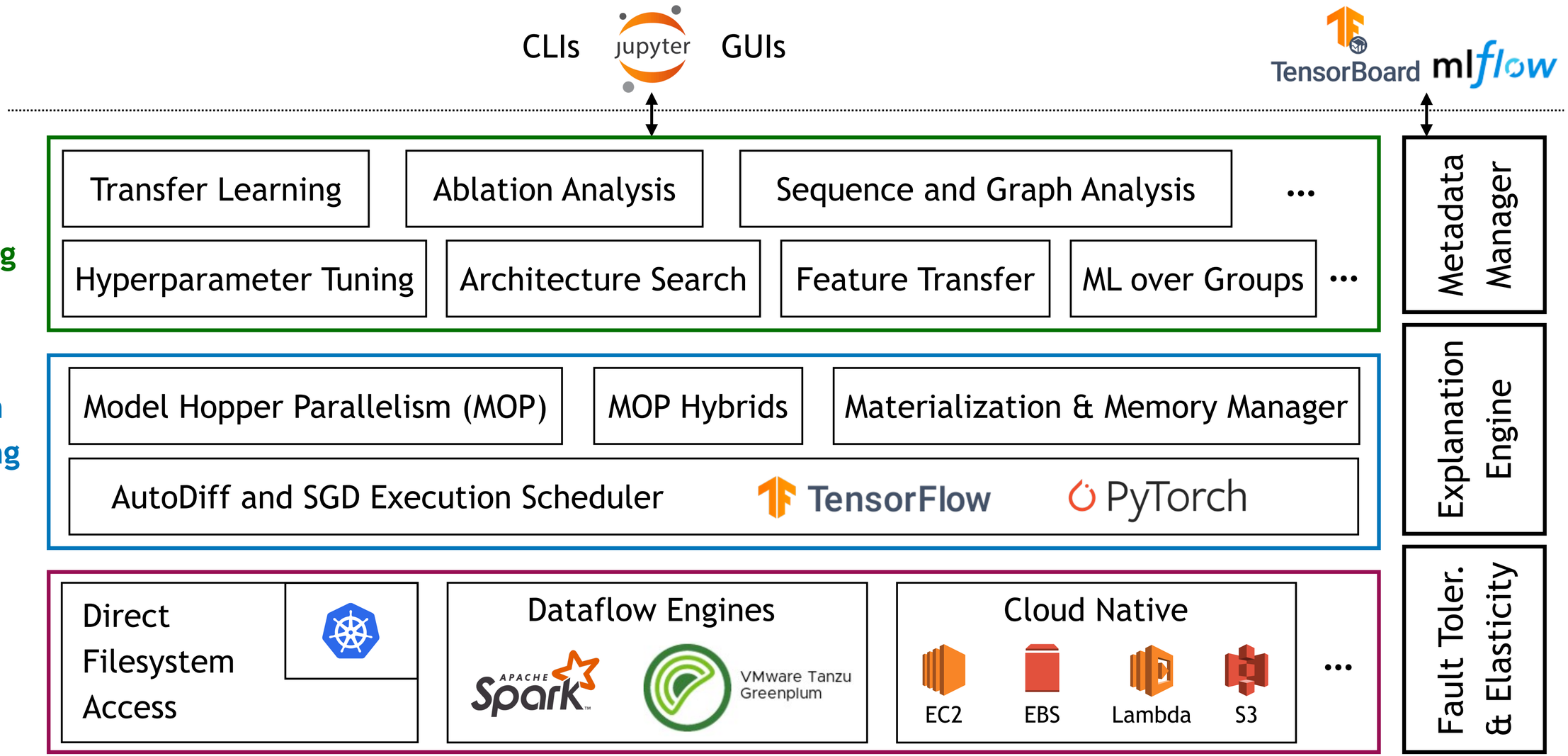
Cerebro: A Layered Data Platform for Scalable Deep Learning. ***CIDR'21***

Full Vision of the Cerebro Platform



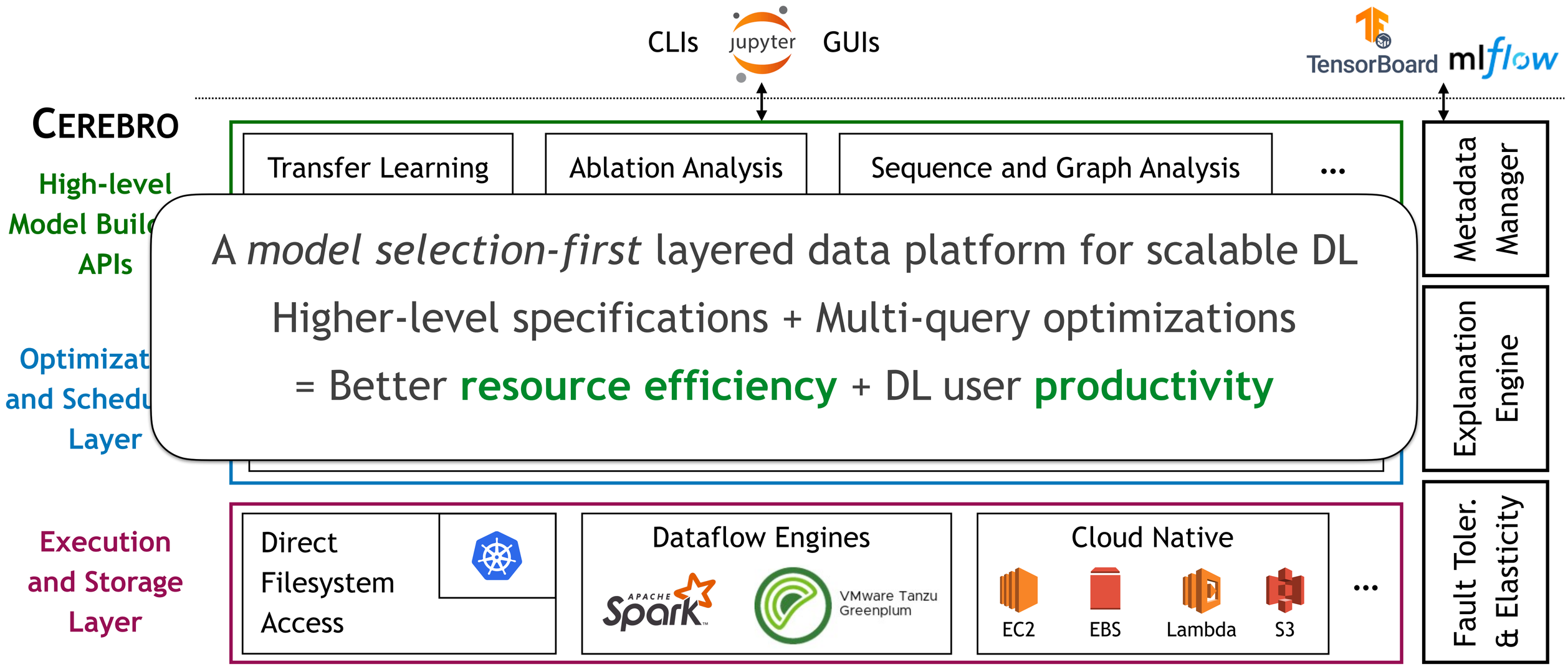
Cerebro: A Layered Data Platform for Scalable Deep Learning. ***CIDR'21***

Full Vision of the Cerebro Platform



Cerebro: A Layered Data Platform for Scalable Deep Learning. ***CIDR'21***

Full Vision of the Cerebro Platform



Cerebro: A Layered Data Platform for Scalable Deep Learning. ***CIDR'21***

Cerebro: Early Impact and Trajectory

Used on TBs by UCSD Public Health scientists; Physics & Materials Eng. next

Shipped with Apache MADlib by Pivotal; VMware collaboration

Cerebro+Spark released; all open sourced under Apache License v2.0

Coming soon: More workloads + optimizations + domain science use cases

Cerebro: A Layered Data Platform for Scalable Deep Learning. ***CIDR'21***

Cerebro: Early Impact and Trajectory

Used on TBs by UCSD Public Health scientists; Physics & Materials Eng. next

Shipped with Apache MADlib by Pivotal; VMware collaboration

Cerebro+Spark released; all open sourced under Apache License v2.0

Coming soon: More workloads + optimizations + domain science use cases

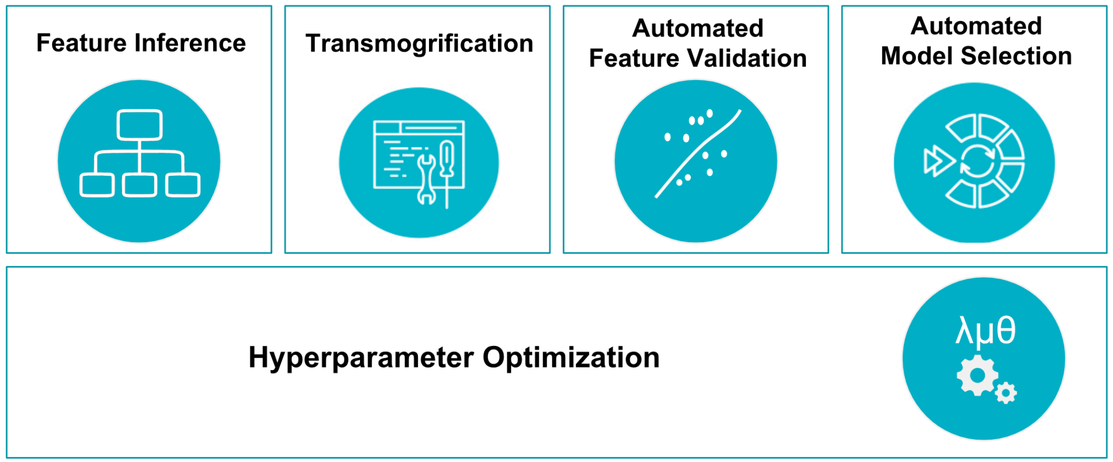
Cerebro: A Layered Data Platform for Scalable Deep Learning. ***CIDR'21***

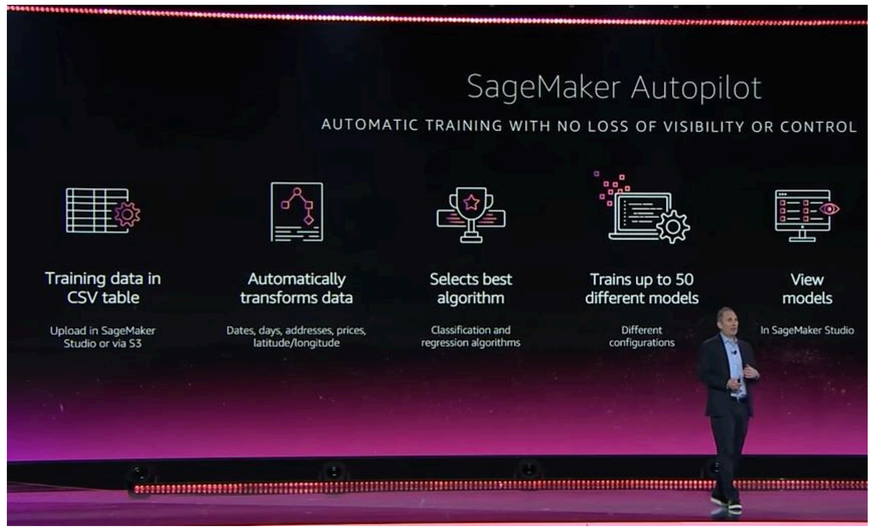
Outline

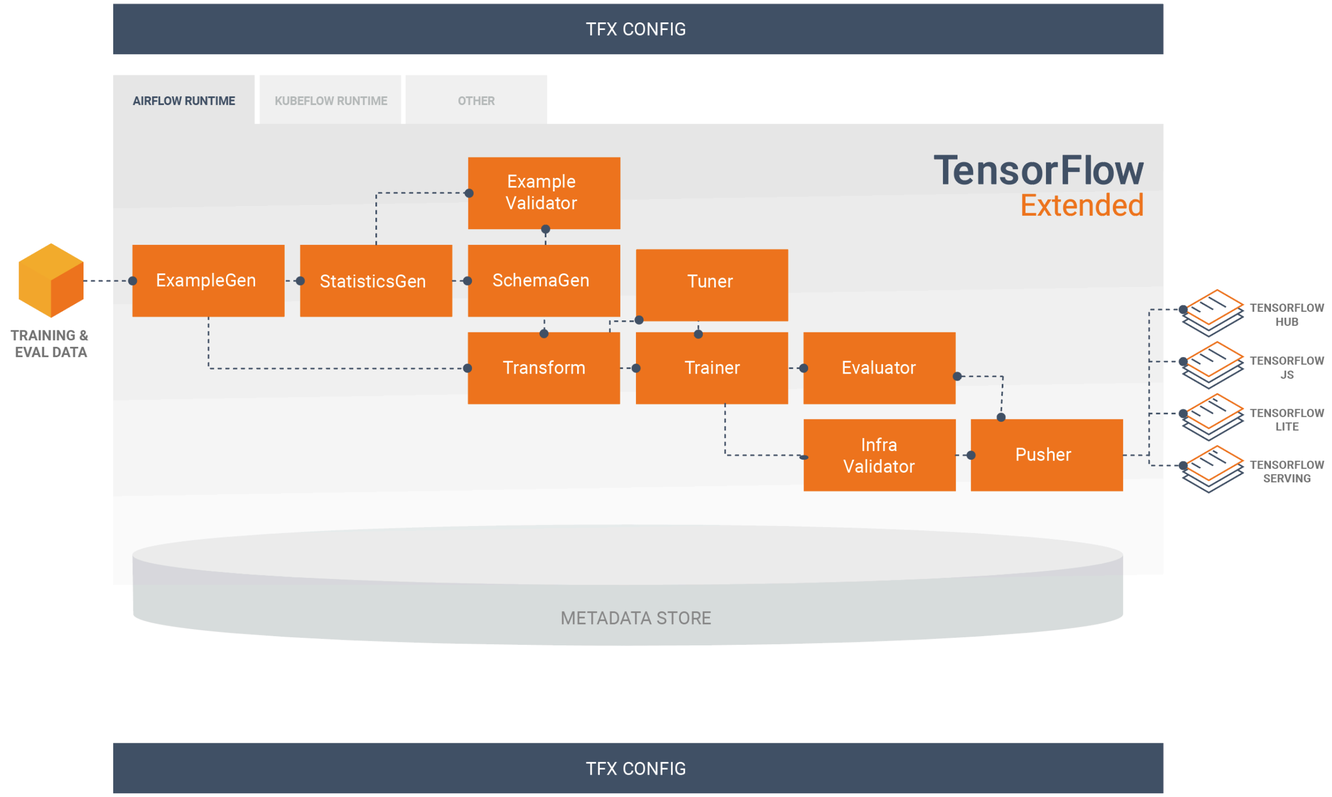
- The New DBfication of ML/AI
- Two Examples from My Research
  - Example 1: Scalable DL Systems
  - Example 2: Auto Data Prep for ML
- Accelerating the DBfication of ML/AI

Example 2: Auto Data Prep for ML

TransmogrifAI

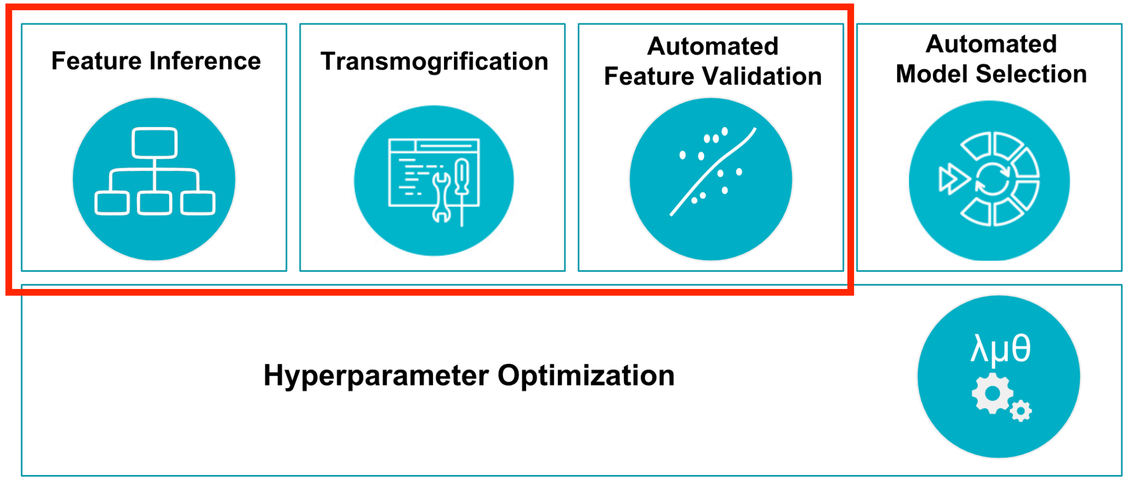


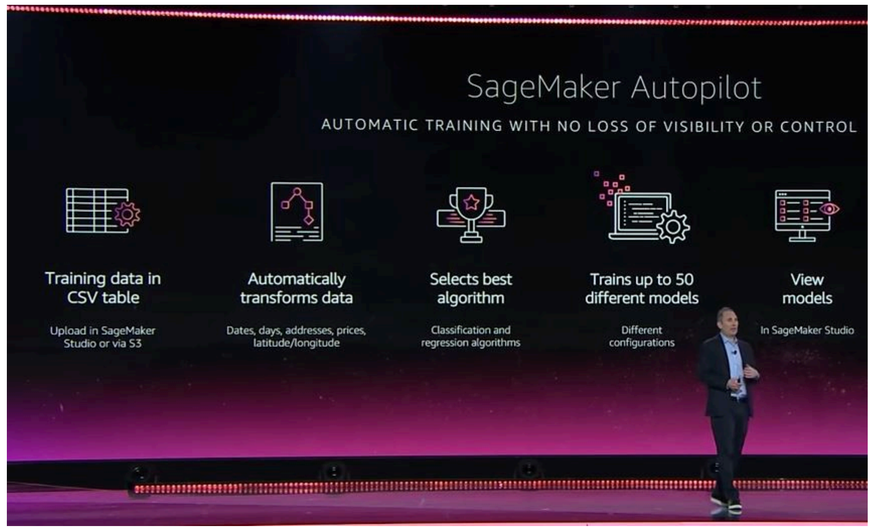


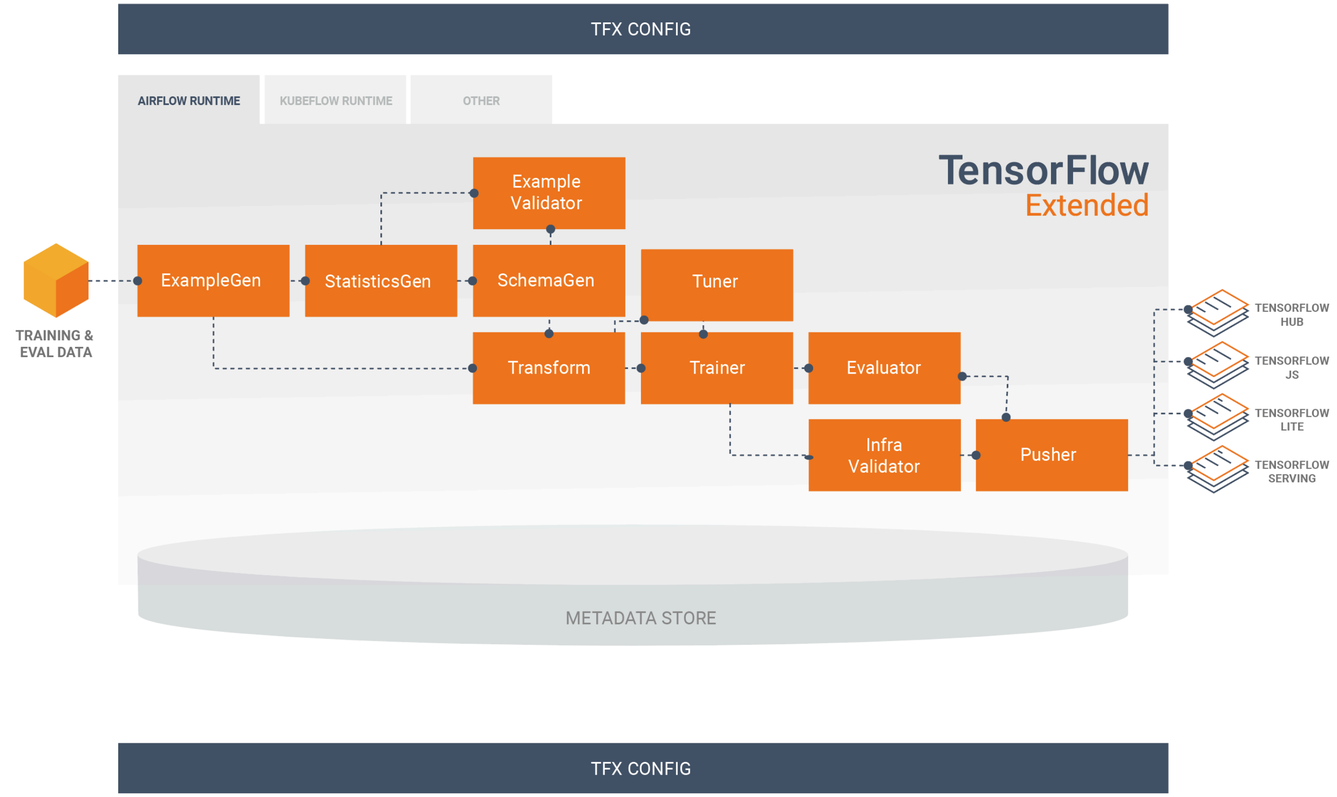


Example 2: Auto Data Prep for ML

TransmogrifAI

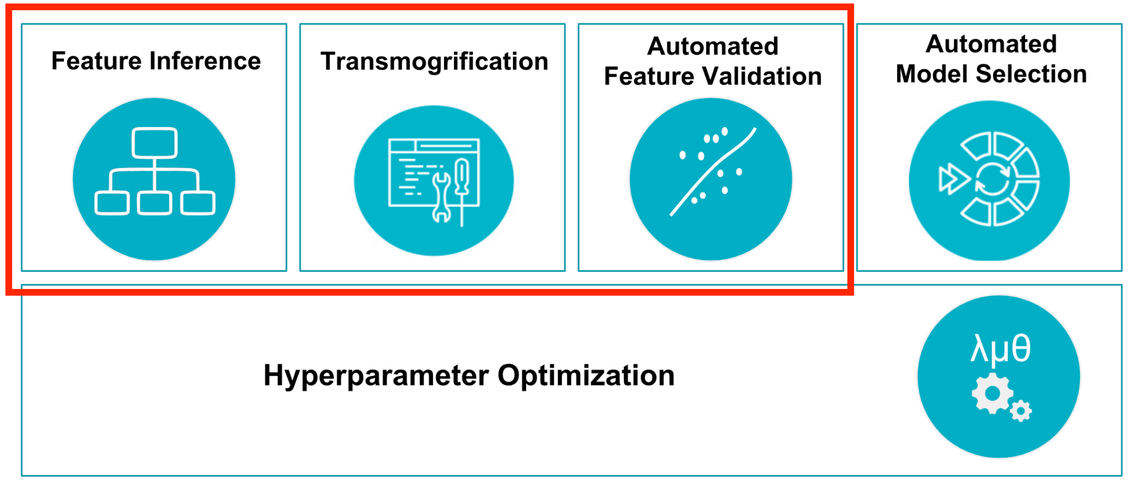


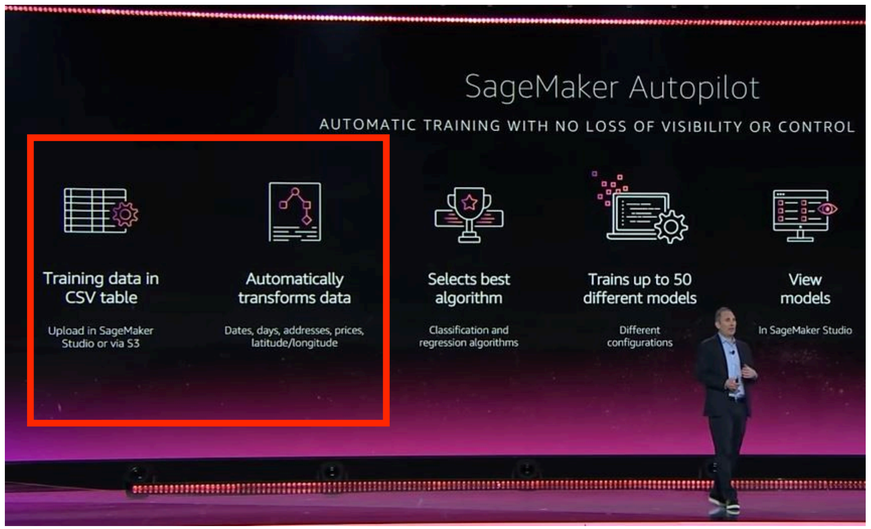


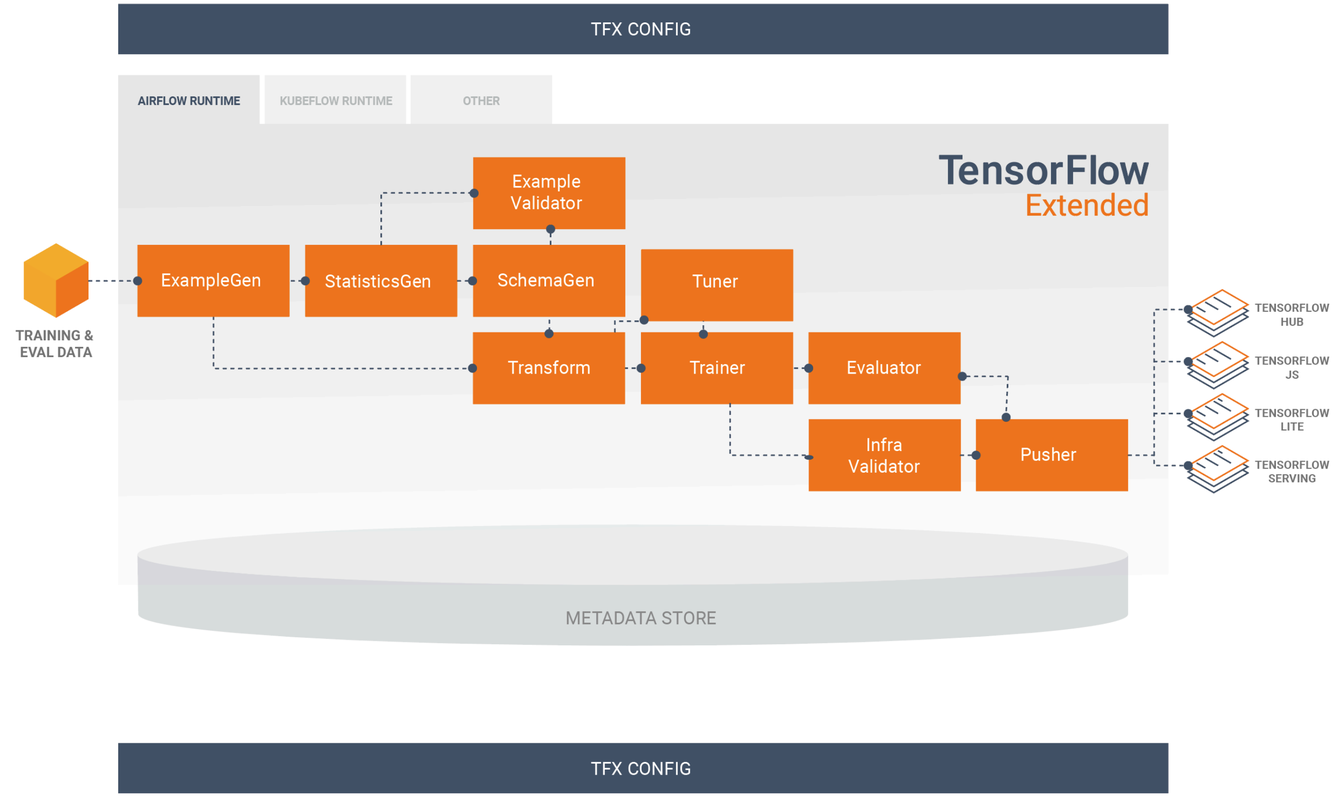


Example 2: Auto Data Prep for ML

TransmogrifAI

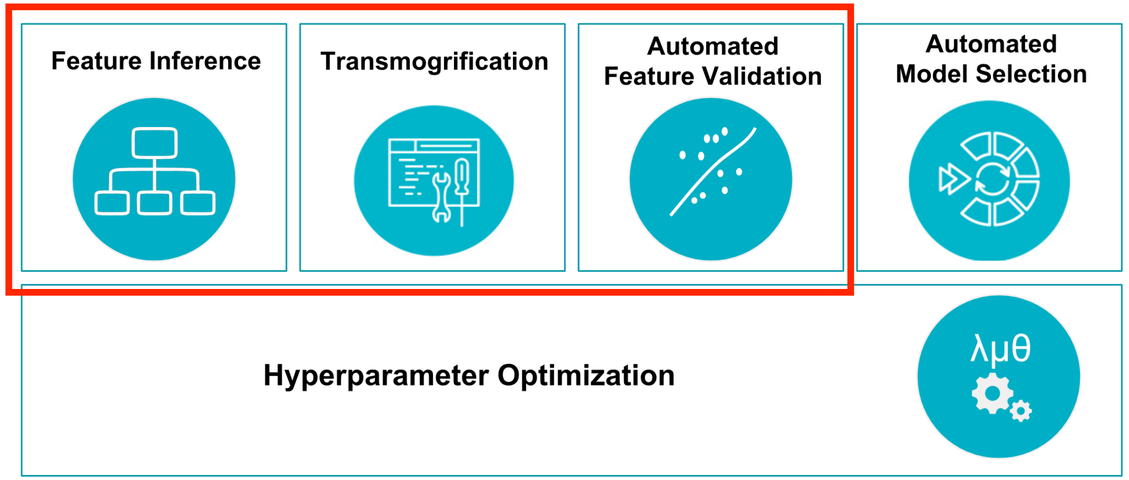


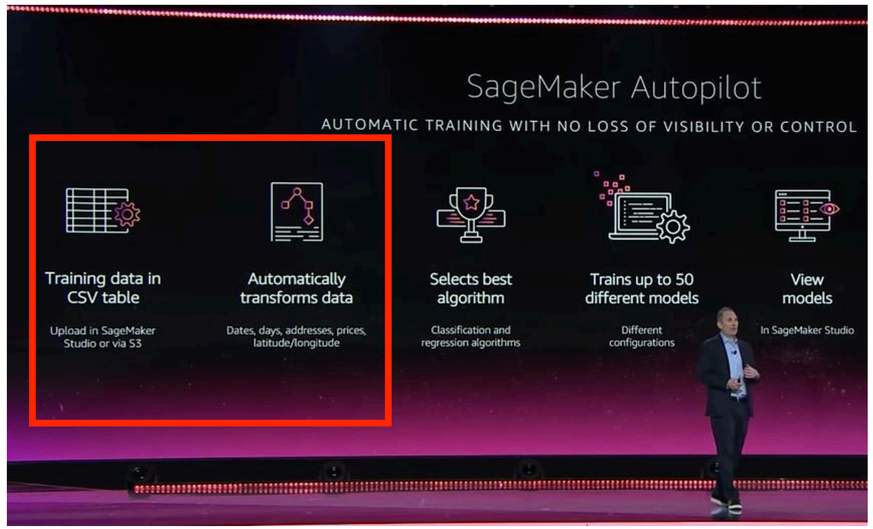


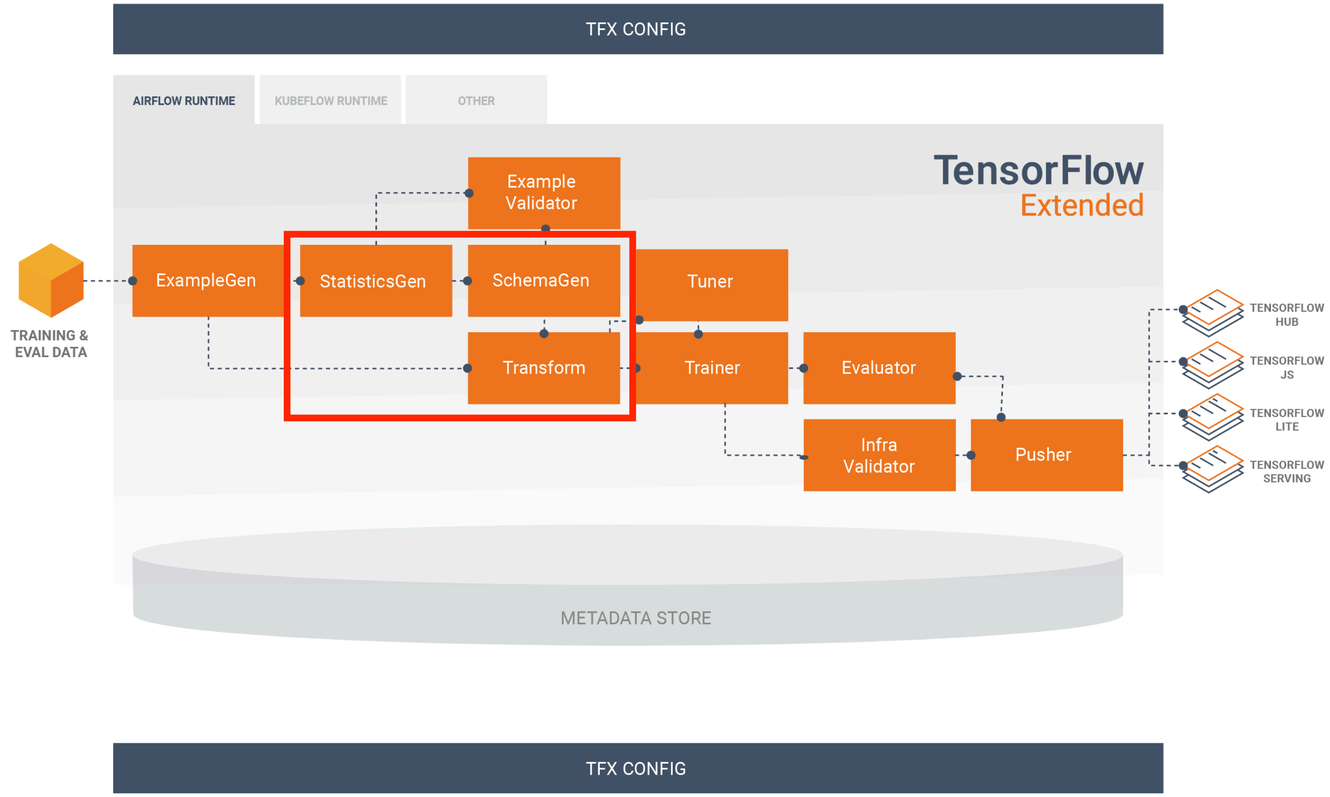


Example 2: Auto Data Prep for ML

TransmogrifAI

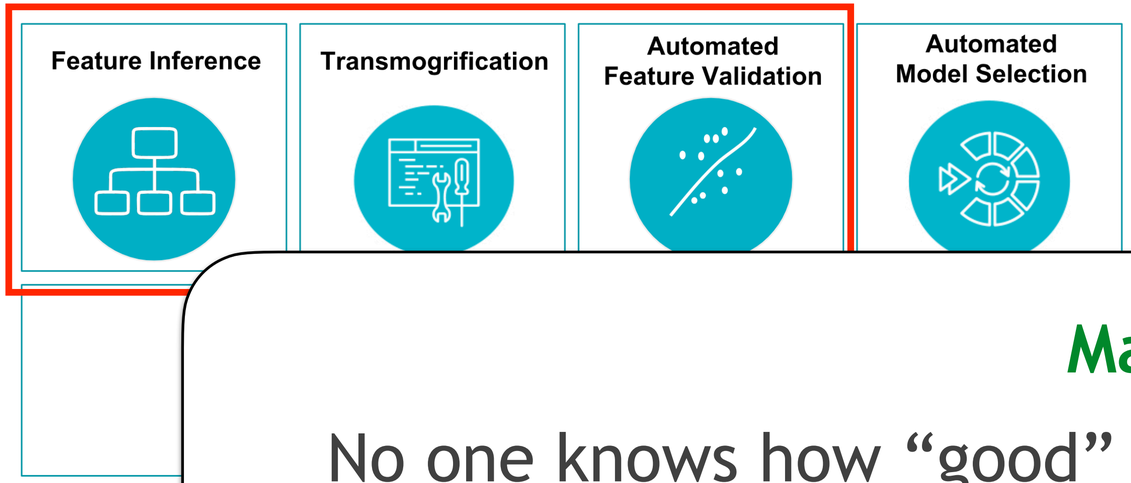


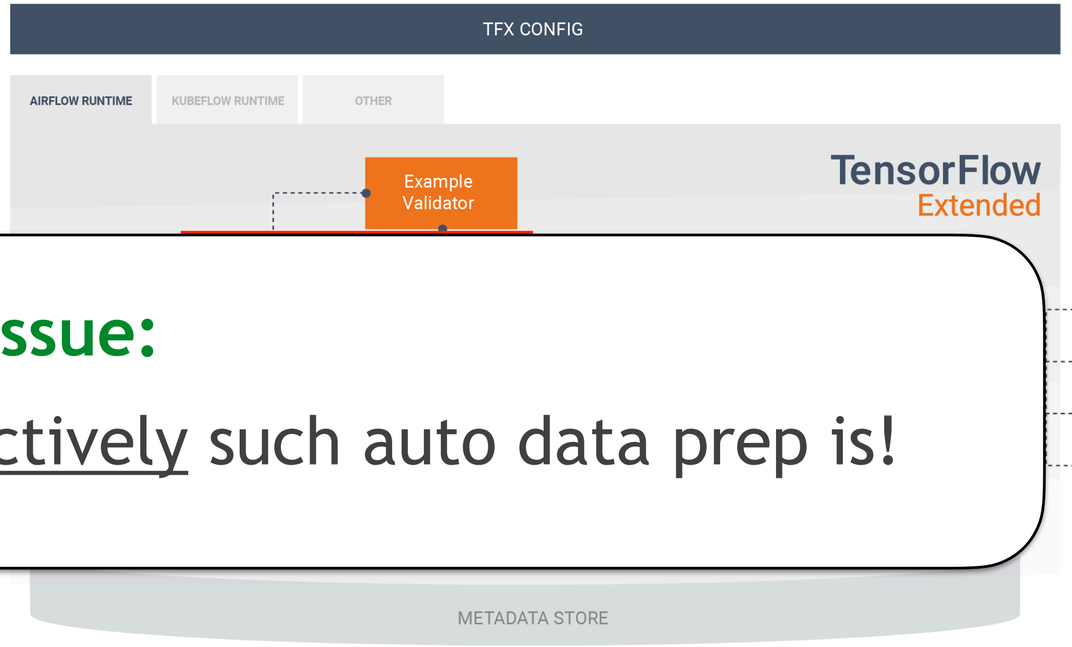




Example 2: Auto Data Prep for ML

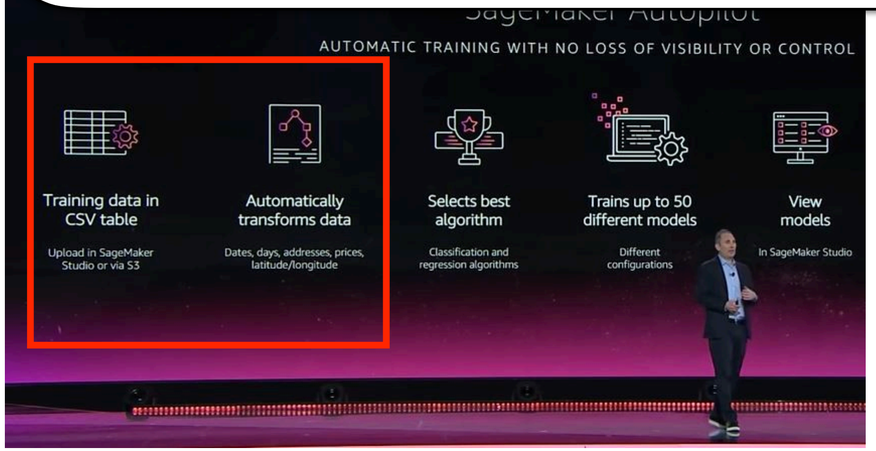
TransmogrifAI





Major Issue:

No one knows <u>how “good” objectively</u> such auto data prep is!



Project SortingHat

Benchmarks for ML Data Prep =

The ML Data Prep Zoo: Towards Semi-Automatic Data Preparation for ML. ***SIGMOD DEEM'19***
Towards Benchmarking Feature Type Inference for AutoML Platforms. ***SIGMOD'21***

Project SortingHat

Benchmarks for ML Data Prep = **Formalized task definition**

What are the inputs?
What are the outputs?

The ML Data Prep Zoo: Towards Semi-Automatic Data Preparation for ML. ***SIGMOD DEEM'19***
Towards Benchmarking Feature Type Inference for AutoML Platforms. ***SIGMOD'21***

Project SortingHat

Benchmarks for ML Data Prep = **Formalized task definition** + **Labeled dataset(s)**

What are the inputs?
What are the outputs?

(Input,Output) pairs to assess accuracy

The ML Data Prep Zoo: Towards Semi-Automatic Data Preparation for ML. ***SIGMOD DEEM'19***
Towards Benchmarking Feature Type Inference for AutoML Platforms. ***SIGMOD'21***

Project SortingHat

Benchmarks for ML Data Prep	=	**Formalized task definition**	+	**Labeled dataset(s)**	+	**Downstream predictions**
		What are the inputs? What are the outputs?		(Input,Output) pairs to assess accuracy		Does it really help?

The ML Data Prep Zoo: Towards Semi-Automatic Data Preparation for ML. ***SIGMOD DEEM'19***
Towards Benchmarking Feature Type Inference for AutoML Platforms. ***SIGMOD'21***

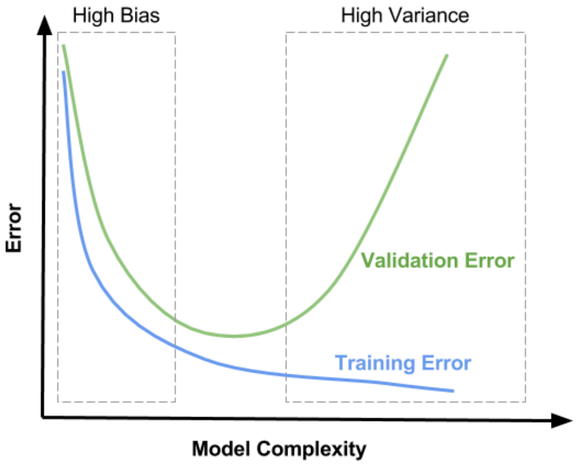
Project SortingHat

Benchmarks for ML Data Prep = **Formalized task definition** + **Labeled dataset(s)** + **Downstream predictions**

Formalized task definition	Labeled dataset(s)	Downstream predictions
What are the inputs? What are the outputs?	(Input,Output) pairs to assess accuracy	Does it really help?

How exactly does data prep affect ML accuracy?

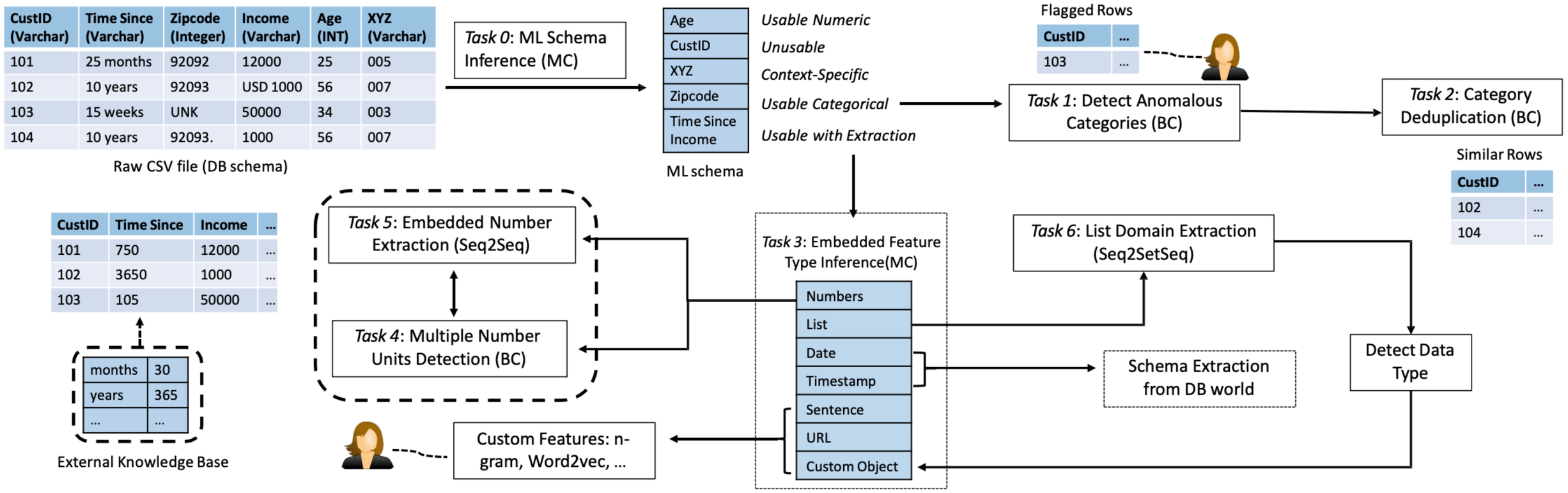
Are “failsafes” possible for errors in auto data prep?



The ML Data Prep Zoo: Towards Semi-Automatic Data Preparation for ML. ***SIGMOD DEEM’19***
Towards Benchmarking Feature Type Inference for AutoML Platforms. ***SIGMOD’21***

Project SortingHat

Current major tasks covered in our **ML Data Prep Zoo** benchmarks:



The ML Data Prep Zoo: Towards Semi-Automatic Data Preparation for ML. ***SIGMOD DEEM'19***
Towards Benchmarking Feature Type Inference for AutoML Platforms. ***SIGMOD'21***

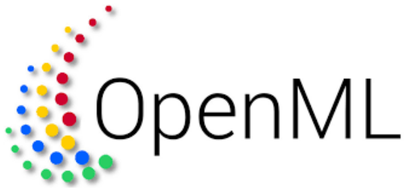
SortingHat: Early Impact and Trajectory

Integrated with TFDV in TFX in collaboration with Google

Being tested on Google's internal AutoML benchmarks

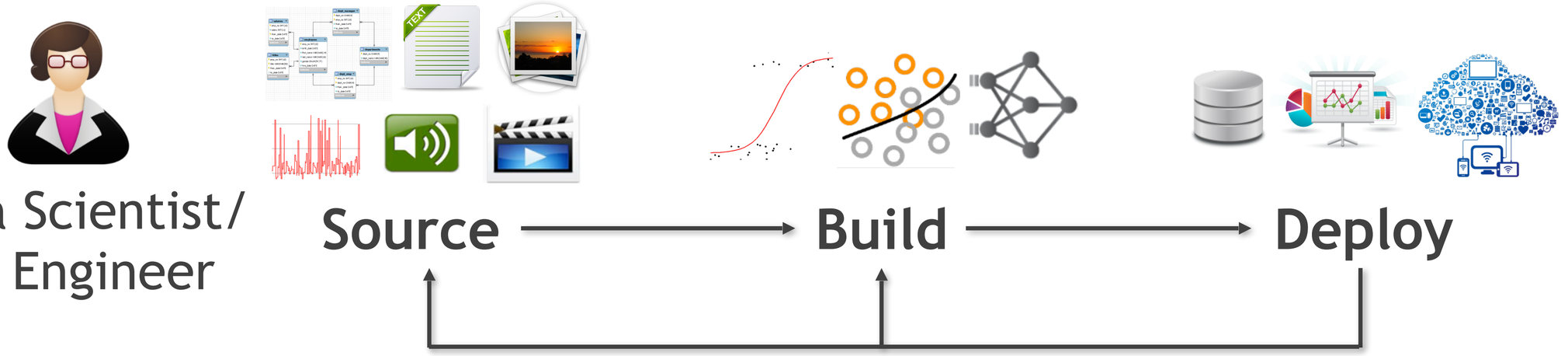
Our model used in NSF-funded "AI Maker" model + dataset search tool

Coming soon: AWS DBMS and ML/AI products; OpenML; catalog search



The ML Data Prep Zoo: Towards Semi-Automatic Data Preparation for ML. ***SIGMOD DEEM'19***
Towards Benchmarking Feature Type Inference for AutoML Platforms. ***SIGMOD'21***

The New DBfication of ML/AI



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.

Leaving these problems open => Huge waste of time/ effort/money/energy/etc. by users of ML/AI!

Leaving these problems open => Huge waste of time/ effort/money/energy/etc. by users of ML/AI!

It is high time for DB, ML/AI, and more areas to join forces to accelerate the DBfication of ML/AI!

Outline

- The New DBfication of ML/AI
- Two Examples from My Research
- Accelerating the DBfication of ML/AI

Accelerating the DBfication of ML/AI

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

Accelerating the DBfication of ML/AI

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.

CSE 234/291: Data Systems for Machine Learning. UC San Diego. 2020.

Data Management in Machine Learning Systems. Morgan & Claypool Publishers. 2019.

Accelerating the DBfication of ML/AI

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

Data Management in Machine Learning: Challenges, Techniques, and Systems. SIGMOD 2017.

Data Management Challenges in Production Machine Learning. SIGMOD 2017.

Data Integration and Machine Learning: A Natural Synergy. VLDB 2018.

Data Collection and Quality Challenges for Deep Learning. VLDB 2020.

Accelerating the DBfication of ML/AI

3. Check out recent "DB for ML" tutorials and papers.

Data Management in Machine Learning: Challenges, Techniques, and Systems. SIGMOD 2017.

Data Management Challenges in Production Machine Learning. SIGMOD 2017.

Data Integration and Machine Learning: A Natural Synergy. VLDB 2018.

Data Collection and Quality Challenges for Deep Learning. VLDB 2020.

4. Check out SIGMOD DEEM and HILDA Workshops. Check out MLSys.

Accelerating the DBfication of ML/AI

5. Check out topical panel discussions on “DB for ML” stuff.

ACM SIGMOD Blog

ML/AI SYSTEMS AND APPLICATIONS: IS THE SIGMOD/VLDB COMMUNITY LOSING RELEVANCE?

Arun Kumar

AUGUST 21, 2018

Databases , Machine Learning

Photos from the workshop. L to R: (1) The DEEM audience. (2) The Panelists: Matei, Joaquin, Jens, Joey, and Manasi (Martin not pictured). (3) Advocatus Diaboli.

SIGMOD DEEM’18

ACM SIGMOD Blog

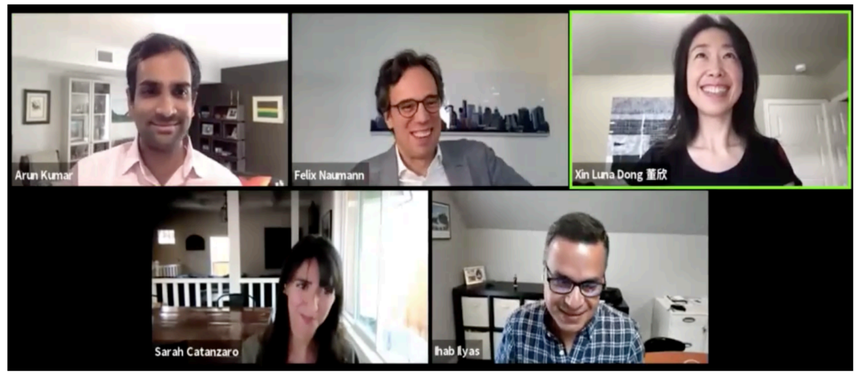
AUTOMATION OF DATA PREP, ML, AND DATA SCIENCE: NEW CURE OR SNAKE OIL?

Arun Kumar

NOVEMBER 10, 2021

AutoML , Data Preparation , Machine Learning

For almost 30 years, the DB / data management community has intensively studied the vexing pains of data integration, cleaning, and transformation. This research has largely been in the contexts of RDBMSs, SQL-oriented business intelligence (BI), and



SIGMOD’21

Accelerating the DBfication of ML/AI

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.

Domain sciences, enterprises, Web companies, cloud vendors, policy, etc.

Create open source artifacts, both software and datasets

Attend/speak at industry venues: Spark/Data+AI Summit, FOSDEM, etc.

…

New Publication Avenues

ACM SIGMOD Blog

Alon Halevy, Arun Kumar and Nesime Tatbul

FEBRUARY 10, 2020

SCALABLE DATA SCIENCE: A NEW RESEARCH TRACK CATEGORY AT PVLDB VOL 14 / VLDB 2021

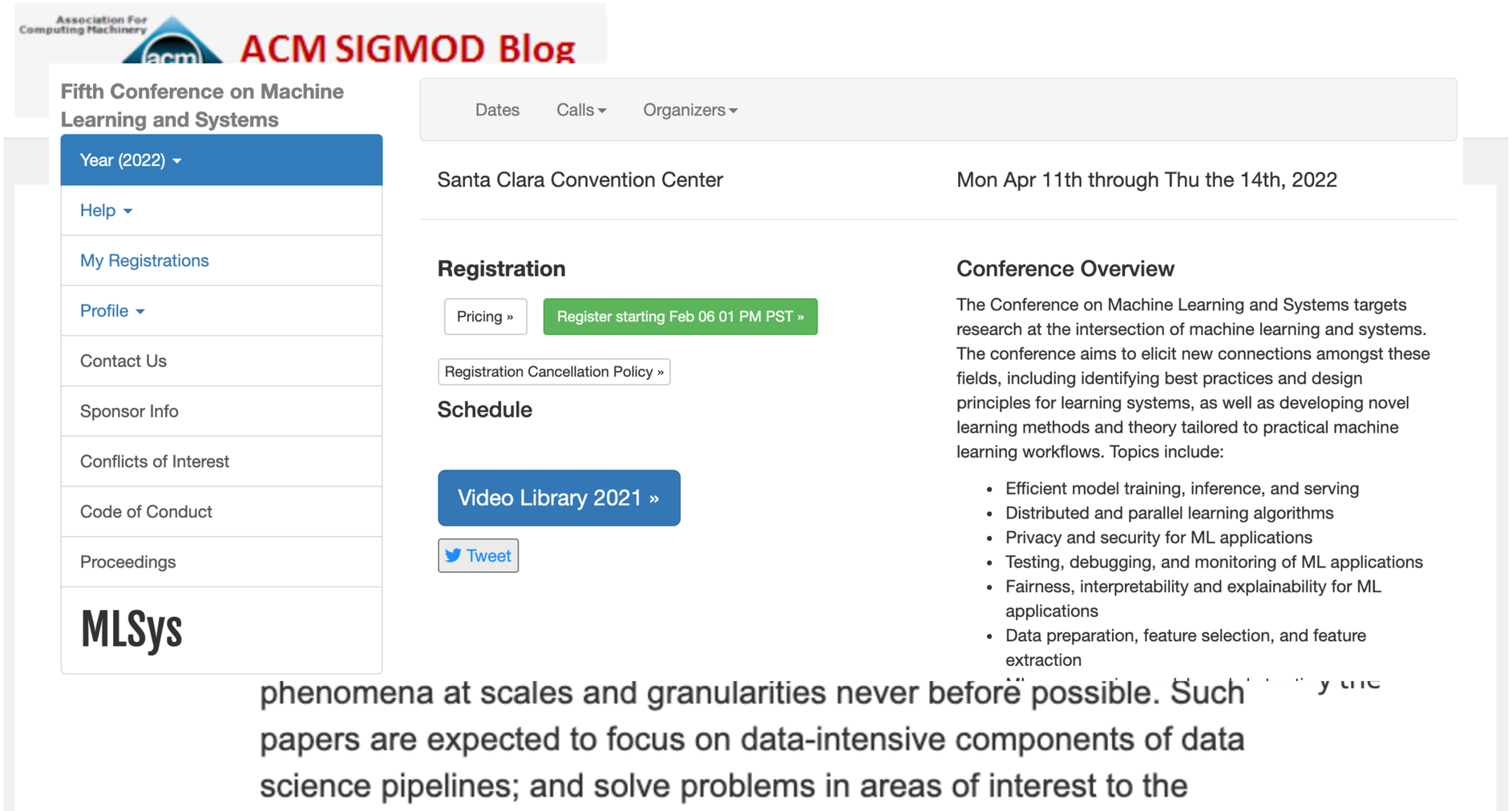
Uncategorized

This post introduces and explains the newly created category of “Scalable Data Science” within the Research Track of PVLDB. This category comes into effect for volume 14, i.e., submissions starting April 1, 2020, which will be evaluated by the Review Board of PVLDB vol 14 for presentation at VLDB 2021.

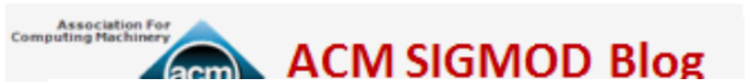
The Growth of Data Science

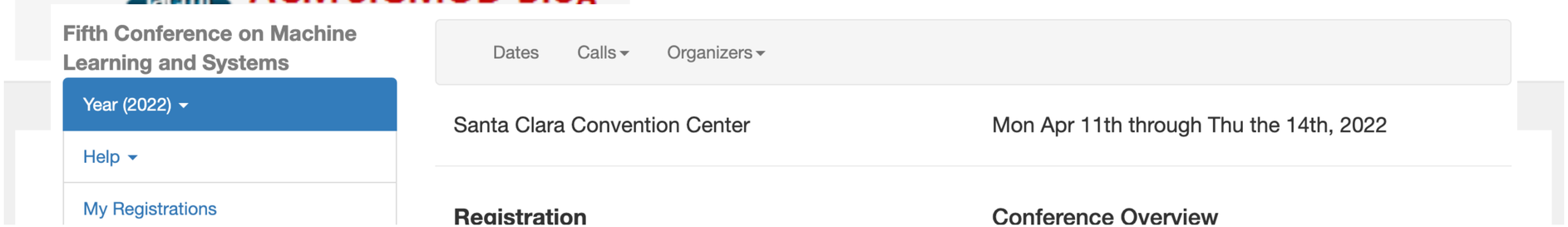
New Publication Avenues

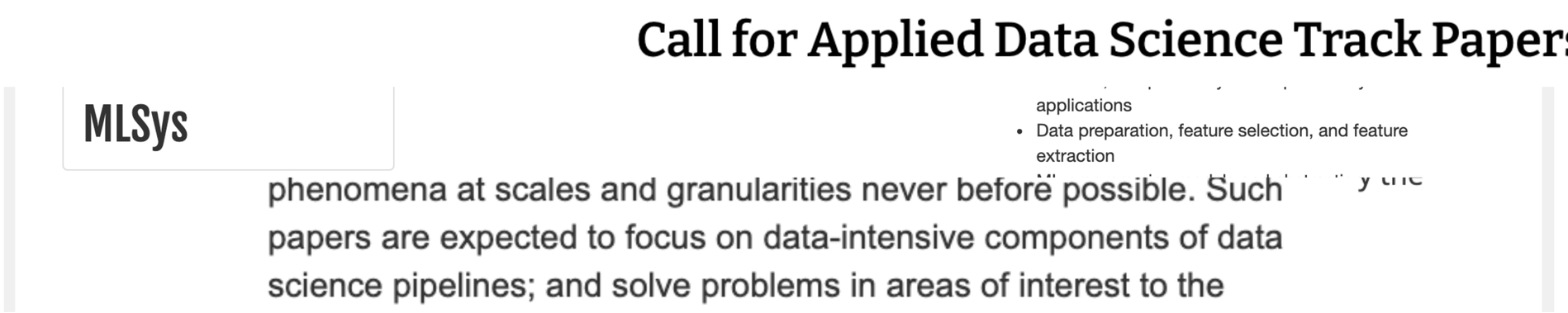
New Publication Avenues



New Publication Avenues





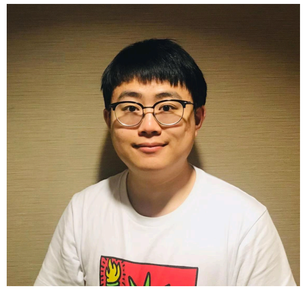


I hope the DB community steps up to the grand challenge of DBfication of ML/AI.

I hope the DB community steps up to the grand challenge of DBfication of ML/AI.

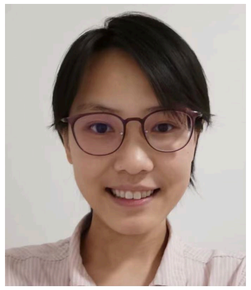
Partner with folks in ML/AI, systems, HCI etc; work with ML/AI users in domain sciences and industry.

My Terrific Advisees



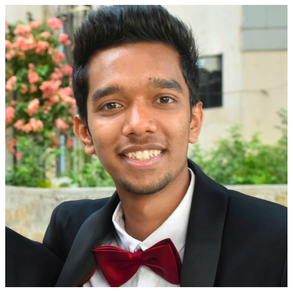
Yuhao Zhang
PhD & MS

Kabir Nagrecha
PhD & BS



Xiuwen Zheng
PhD & MS

Supun Nakandala
PhD'22 -> Databricks



Pradyumna Sridhara
MS

Vignesh Nanda Kumar
MS

Kyle Luoma
PhD

Vraj Shah
PhD'22 -> IBM
Research Almaden

https://adalabucsd.github.io

https://ADALabUCSD.github.io

arunkk@eng.ucsd.edu

github.com/ADALabUCSD

@TweetAtAKK

ACKS: