Session 1.
The current landscape of Big Data benchmarks

DataBench

Panelists

Arne Berre
Chief Scientist, SINTEF

Axel Ngonga
University of Paderborn, Lead of BDVA TF6 Benchmarking group

Wangling Gao
Assistant professor, Institute of Computing Technology, Chinese Academy of Sciences

Rekha Singhal
Senior scientist and Head of the Computing Systems-Software Research area in TCS

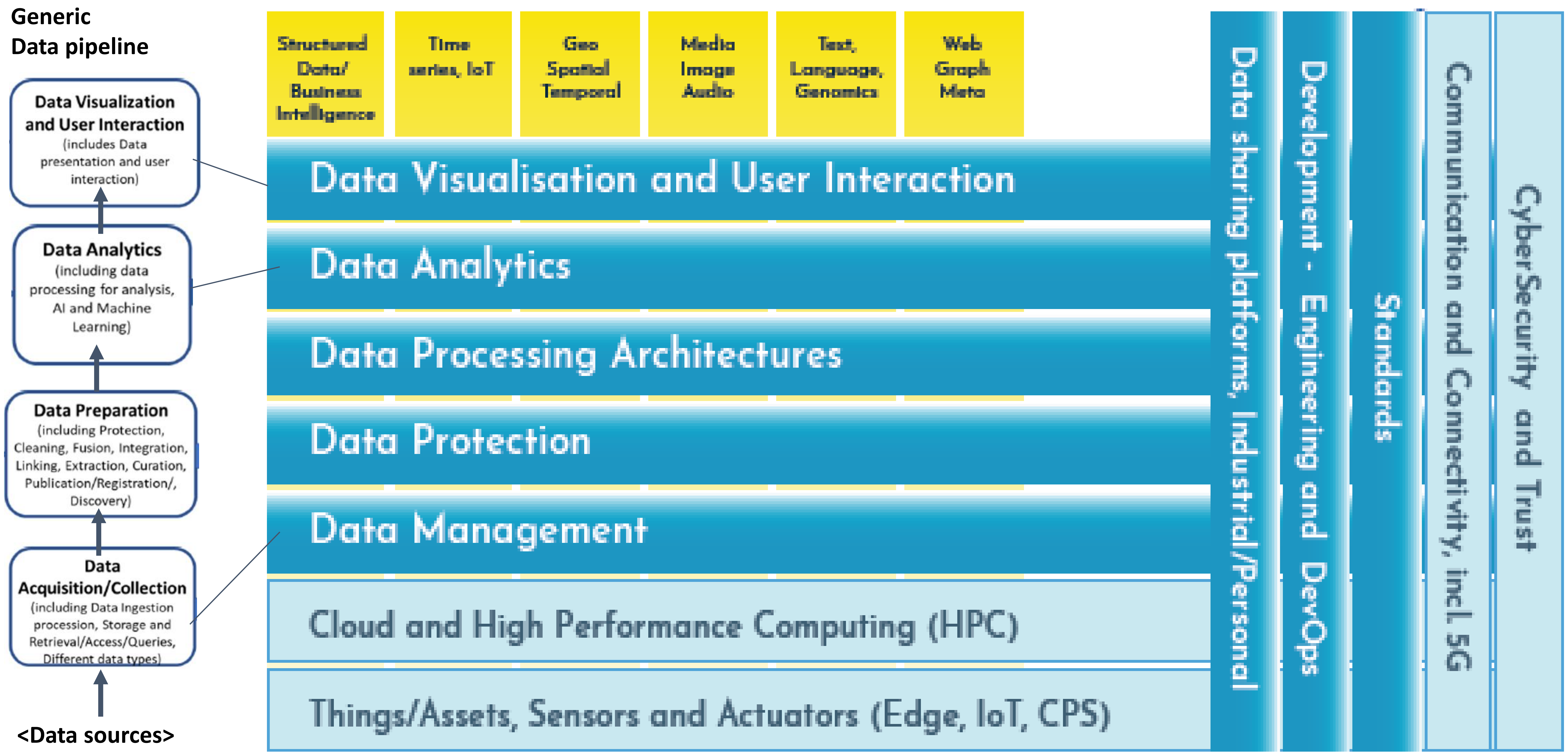
Todor Ivanov
Senior consultant, Lead Consult

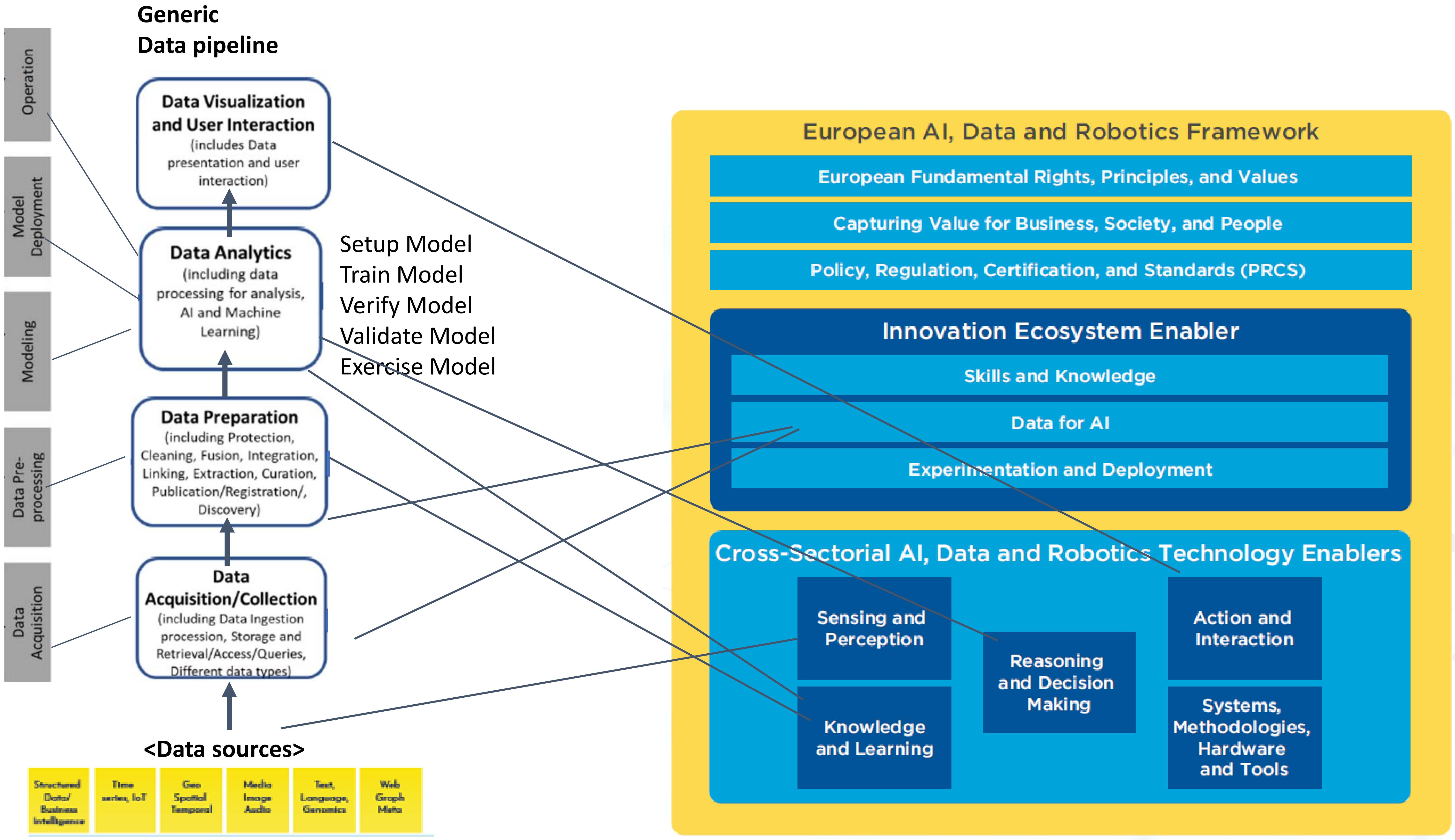
DataBench

10:15-11:15 Session 1.
The current landscape of Big Data and AI benchmarks

- 10:15-10:25: DataBench Framework for Benchmarks, Arne J. Berre, SINTEF
- 10:25-10:40: Benchmarking platforms and AI, Axel Ngonga, BDVA TF6 Benchmark Lead, University of Paderborn
- 10:40-10:55: BenchCouncil Big Data and AI Benchmarks, Wanling Gao, Chinese Academy of Sciences
- 10:55-11:10: MLPerf AI and ABench, Rekha Singhal, Senior Scientist, TCS, India
- 11:10-11:15: Conclusion on Big Data and AI Benchmarks, Todor Ivanov, LeadConsult

BDV – Big Data and Analytics/Machine Learning Reference Model

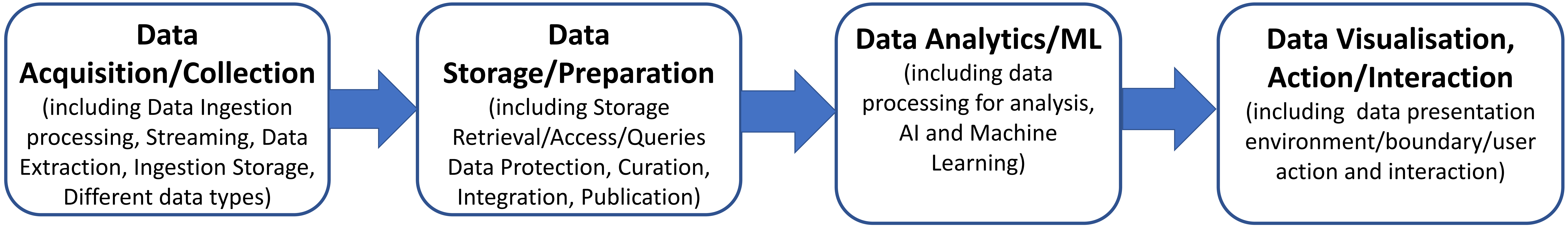




Generic
Data pipeline
Operation
Model Deployment
Modeling
Data Pre-processing
Data Acquisition
Data Visualization and User Interaction (includes Data presentation and user interaction)
Data Analytics (including data processing for analysis, AI and Machine Learning)
Setup Model
Train Model
Verify Model
Validate Model
Exercise Model
Data Preparation (including Protection, Cleaning, Fusion, Integration, Linking, Extraction, Curation, Publication/Registration/, Discovery)
Data Acquisition/Collection (including Data Ingestion procession, Storage and Retrieval/Access/Queries, Different data types)
<Data sources>
Structured Data/ Business Intelligence
Time series, IoT
Geo Spatial Temporal
Media Image Audio
Text, Language, Genomics
Web Graph Meta
European AI, Data and Robotics Framework
European Fundamental Rights, Principles, and Values
Capturing Value for Business, Society, and People
Policy, Regulation, Certification, and Standards (PRCS)
Innovation Ecosystem Enabler
Skills and Knowledge
Data for AI
Experimentation and Deployment
Cross-Sectorial AI, Data and Robotics Technology Enablers
Sensing and Perception
Knowledge and Learning
Reasoning and Decision Making
Action and Interaction
Systems, Methodologies, Hardware and Tools

Top level Generic Big Data and AI Pipeline pattern

(For all Benchmarks and Project pipeline to be related to)



Benchmarks · Knowledge Nuggets · Other Tools · Search

Search

Big Data Analytics

=

Big Opportunities
for EU Companies

Catalogue of technical Benchmarks

Databench has compiled a list of technical benchmarks that can help you with measuring your Big Data system, compare it with others and help in fine-tuning it.

About DataBench Toolbox

Based on existing efforts in big data benchmarking and enabling inclusion of new benchmarks that could arise in the future, the DataBench Toolbox provides a unique environment to search, select and deploy big data benchmarking tools, giving the possibility to generate unified technical metrics and, most importantly, going the extra mile and derive business KPIs for your organization.

User journeys: What type of user are you?

Whether you are more interested in the technical aspects of benchmarking, or your focus lays more on the business aspects we have prepared a set of user-journeys ready to help you while working with this platform.
Just select from the titles below the one that you are more interested in to see a page with advices

Technical	Business	Benchmark provider

<- Back

Benchmark catalogue

Filter..

BigBench V2

The BigBench V2 benchmark addresses some of the limitation of the BigBench (TPCx-BB) benchmark. BigBench V2 separates from TPC-DS with a simple data model. The new data model still has the variety of structured, semi-structured, and unstructured data as the original BigBench data model. The differe...

HiBench

A comprehensive benchmark suite consisting of multiple workloads including both synthetic micro-benchmarks and real-world applications. HiBench features several ready-to-use benchmarks from 4 categories: micro benchmarks, Web search, Machine Learning, and HDFS benchmarks. It is used for both stream...

Yahoo Streaming Benchmark (YSB)

It is an end-to-end pipeline that simulates a real-world advertisement analytics pipeline. Currently implemented in Kafka, Storm, Spark, Flink and Redis. Yahoo reported the following as background of why they developed YSB: "At Yahoo we have adopted >Apache Storm as our stream processing p...

Yahoo! Cloud Serving Benchmark (YCSB)

A benchmark designed to compare emerging cloud serving systems like Cassandra, HBase, MongoDB, Riak and many more, which do not support ACID. It provides a core package of 6 pre-defined workloads A-F, which simulate a cloud OLTP application. Web references https://github.com/brianfrankcoope...

ABench

ABench is as a big data architecture stack benchmark. It aims to evaluate big data system across multiple layers of big data architecture, including cloud services, data storage, batch processing, interactive processing, streaming and machine learning. The benchmark supports re-using of existing be...

AdBench

It combines Ad-Serving, Streaming Analytics on Ad-serving logs, streaming ingestion and updates of various data entities, batch-oriented analytics (e.g. for Billing), Ad-Hoc analytical queries, and Machine learning for Ad targeting. While this benchmark is specific to modern Web or Mobile advertisi...

AIBench

AIBench is an industry standard Internet service AI benchmark suite, designed specifically for modern Internet services with microservice-based architecture. The benchmark spans sixteen AI problem domains from three most widely used Internet service domains: search engine, social network, and e-com...

Hobbit Benchmark

<- Back

Benchmark suite (47) Sandbox/VM (6) Inhouse/On-Premise (40) Cloud (53) Data Quality (6) Data Management (38) Data Visualization (5) Execution time (68) Throughput (40) CPU and Memory (28) Hybrid (28) Graphs or linked data (24) NoSQL (27) Distributed (49) Centralized (11) Volume (23) Execution performance (60) Graph Databases (16) Data visualization and User interaction (10) Data analytics (52) Data processing architectures (48) Data management (46) Data acquisition/Collection (64) Data preparation (71) Data analytics (49) Data Visualization and User interaction (9) Graph (17)

Description

The HOBBIT evaluation platform is a distributed FAIR benchmarking platform for the Linked Data lifecycle. This means that the platform was designed to provide means to: (1) benchmark any step of the Linked Data lifecycle, including generation and acquisition, analytics and processing, storage and curation as well as visualization and services;(2) ensure that benchmarking results can be found, accessed, integrated and reused easily (FAIR principles); (3) benchmark Big Data platforms by being the first distributed benchmarking platform for Linked data.

Web references

https://project-hobbit.eu/

Date of last description update

2018

Originating group

Hobbit H2020 Project

Time – first version, last version

2016-2018

Type/Domain

Linked Data Benchmark Platform

Workload

Real world application workloads

Mapping between the ***Pipeline Steps*** and the ***Benchmark Ecosystem*** → matrix available in the **DataBench ToolBox**

Year	Benchmarks	PI-DV	PI-DA	PI-DS	PI-DI
1999	TPC-H			X	X
2002	TPC-DS v1			X	X
2004	Linear Road				
2007	Hadoop Workload Examples			X	X
2007	GridMix			X	X
2008	PigMix			X	X
2008	MRBench			X	X
2009	CALDA			X	X
2010	HiBench		X	X	X
2010	Liquid				
2010	YCSB			X	X
2011	SWIM				X
2012	CloudRank-D		X	X	X
2012	PUMA Benchmark Suite			X	X
2012	CloudeSuite			X	
2012	MRBS		X	X	X
2013	AMP Lab Big Data Benchmark			X	X
2013	BigBench		X	X	X
2013	BigDataBench		X	X	X
2013	LinkBench			X	X
2013	BigFrame			X	X
2013	PRIMEBALL			X	X
2014	Semantic Publishing Benchmark (SPB		X	X	X
2014	Social Network Benchmark			X	X
2014	ALOJA				
2014	gmark			X	X
2014	Convnet		X		X
2014	WatDiv				X
2014	StreamBench			X	X
2014	TPCx-HS			X	X

Year	Benchmarks	PI-DV	PI-DA	PI-DS	PI-DI
2015	SparkBench		X	X	X
2015	TPCx-V			X	X
2015	IoTAbench			X	X
2015	BigFUN			X	X
2015	TPC-DS v2			X	X
2015	TPCx-BB		X	X	X
2015	CityBench			X	X
2015	Graphalytics			X	X
2015	Yahoo Streaming Benchmark (YSB)				X
2015	ShenZhen Transportation System (SZTS)				
2016	DeepBench		X		X
2016	DeepMark		X		X
2016	TensorFlow Benchmarks		X		X
2016	Fathom		X		X
2016	AdBench		X	X	X
2016	RIoTBench		X	X	X
2016	Hobbit Benchmark	X	X	X	X
2017	TPCx-HS v2			X	X
2017	BigBench V2		X	X	X
2017	Sanzu		X	X	
2017	AIM Benchmark		X	X	X
2017	GARDENIA		X	X	X
2017	Penn machine learning benchmark (PMLB)	X	X		
2017	OpenML benchmark suites		X		
2017	Benchip		X		
2017	Deep Learning Benchmarking Suite (DLBS)		X		
2017	TPCx-IoT			X	X
2017	Senska			X	X
2017	DAWNBench		X		X
2017	BlockBench				X
2017	IDEBench	X			X
2017	ABench		X	X	X
2018	Stream WatDiv				X
2018	TERMinator Suite			X	
2018	HERMIT		X		X
2018	MLBench Services		X		X
2018	MLBench Distributed	X	X		X
2018	MLPerf		X		
2018	Training Benchmark for DNNs (TBD)		X		
2018	PolyBench				X
2018	NNBench-X	X	X		X
2019	GDPRbench			X	X
2019	BenchIoT			X	
2019	IoTBench				
2019	Visual Road	X	X	X	X
2019	AdaBench		X	X	
2019	MiDBench			X	X
2019	CBench-Dynamo			X	X
2019	Edge AIBench				
2019	AIBench	X	X	X	X
2019	HPC AI500		X	X	X
2019	SparkAIBench		X	X	X
2019	AI Matrix		X	X	X

10:15-11:15 Session 1.
The current landscape of Big Data and AI benchmarks

10:25-10:40: Benchmarking platforms and AI, Axel Ngonga, BDVA TF6 Benchmark Lead, University of Paderborn

10:40-10:55: BenchCouncil Big Data and AI Benchmarks, Wanling Gao, Chinese Academy of Sciences

10:55-11:10: MLPerf AI and ABench, Rekha Singhal, Senior Scientist, TCS, India

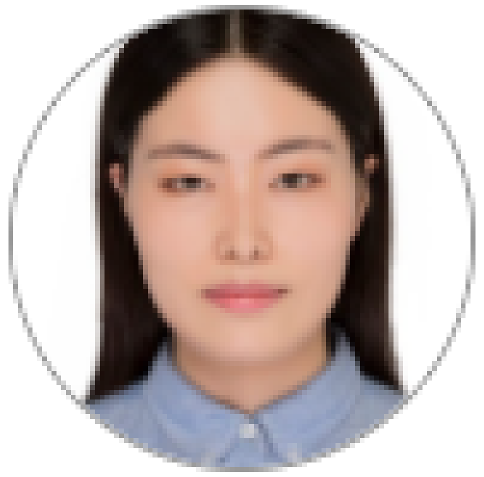
11:10-11:15: Conclusion on Big Data and AI Benchmarks, Todor Ivanov, LeadConsult

Axel Ngonga

University of Paderborn, Lead of BDVA TF6 Benchmarking group

Evaluation schemes for Big data and AI Performance of high Business impact (DataBench project sponsored session)

Axel's research focuses on methods to improve the life cycle of knowledge graphs with a strong focus a machine learning techniques. He has participated in/led a number of German and European H2020 projects (including HOBBIT, SAKE, GEISER, RAKI and DAIKIRI) in which he developed techniques for the extraction, integration and fusion of knowledge graphs at scale. Axel currently leads the KnowGraphs Training Network, in which early-stage researchers address some of the core challenges in the representation, extraction, management and use of knowledge graphs. Axel also leads the development of popular benchmarking frameworks such as GERBIL, IGUANA and HOBBIT. He is full professor of Data Science at Paderborn University, where he also leads the activities on Digital Humanities



Wanling Gao

Assistant Professor at Institute of Computing Technology, Chinese Academy of Sciences

Evaluation schemes for Big data and AI Performance of high Business impact (DataBench project sponsored session)

Wanling Gao received the B.S. degree from the Huazhong University of Science and Technology, in 2012, and the Ph.D. degree in computer science and engineering from the University of Chinese Academy of Sciences, in 2019. She has been an Assistant Professor in computer science with the Institute of Computing Technology, Chinese Academy of Sciences, and with the University of Chinese Academy of Sciences, since 2019. Her research interests include big data and AI benchmarking and systems. She is in particular involved with BenchCouncil and related benchmarks like AIBench, HPC AI500, Edge AIBench, AIoTBench and others

Rekha Singhal

Senior Scientist, Head Computing Systems at Tata Consultancy Services

Evaluation schemes for Big data and AI Performance of high Business impact (DataBench project sponsored session)

She focuses on accelerating development and deployment of enterprise applications in data-driven programming environment. Her research interests are heterogeneous architectures for accelerating ML pipelines, high-performance data analytics systems, big data performance analysis, query optimization, storage area networks, and distributed systems. She is associated with Spec RG Big data and ML Perf. She has filed patents and 13 granted in international territories. She also has several publications in international and national conferences, workshops, and journals. She has been awarded with ACM Senior member. She had led the project on Disaster Recovery appliance which was runner up for NASSCOm award. She has received her M.Tech. and Ph.D. in Computer Science from IIT, Delhi, and has been a visiting researcher at Stanford University, United States

Todor Ivanov

Senior Consultant at Lead Consult

Evaluation schemes for Big data and AI Performance of high Business impact (DataBench project sponsored session)

Dr. Todor Ivnaov is an expert in the design, implementation and benchmarking of distributed big data systems and data-intensive applications. Prior to that, he has worked as a senior researcher in multiple projects in the field of databases and big data benchmarking as well as a senior software engineer developing Flight Information Display Systems (FIDS) for different international airports.

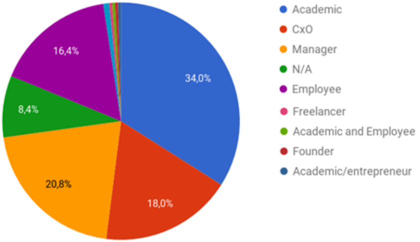
Benchmarking Platforms and AI
The Example of HOBBIT

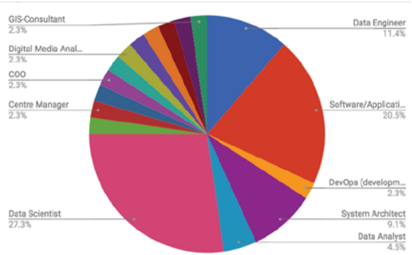
Axel-Cyrille Ngonga Ngomo

DICE Research Group
InfAI

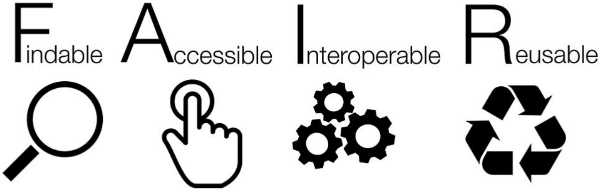
EBDVF 2020

- Two surveys in industry and academia (97 CxOs, 87 academics)
- Core of the survey: Requirements to benchmarking platforms (AI, KG)
- Result: Hobbit, a holistic benchmarking platform

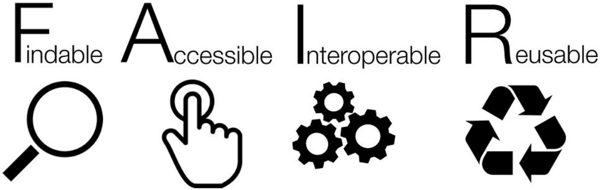




1. FAIR environment (Linked Data principles)

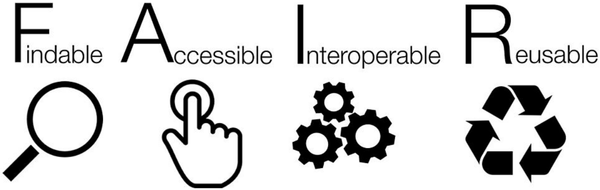


1. FAIR environment (Linked Data principles)
2. Fair conditions (separation of benchmark and systems)

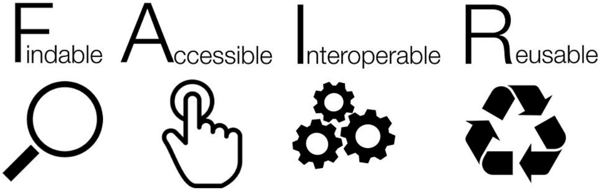


Requirements

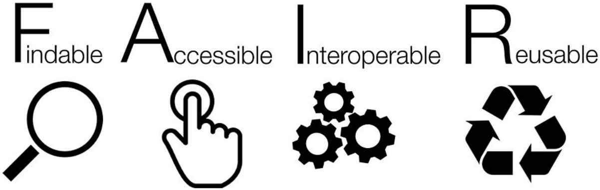
1. FAIR environment (Linked Data principles)
2. Fair conditions (separation of benchmark and systems)
3. Scalable (distributed execution and benchmarking)



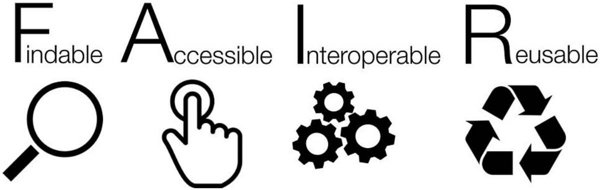
1. FAIR environment (Linked Data principles)
2. Fair conditions (separation of benchmark and systems)
3. Scalable (distributed execution and benchmarking)
4. Open (open-source, extensible)

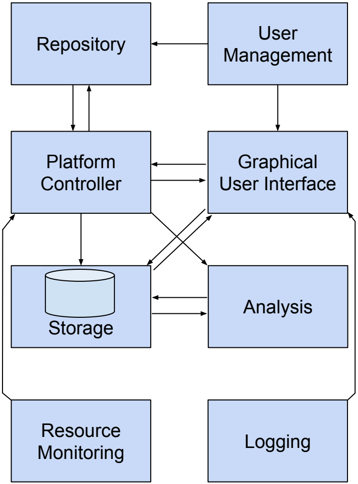


1. FAIR environment (Linked Data principles)
2. Fair conditions (separation of benchmark and systems)
3. Scalable (distributed execution and benchmarking)
4. Open (open-source, extensible)
5. Safe and secure (encryption,)

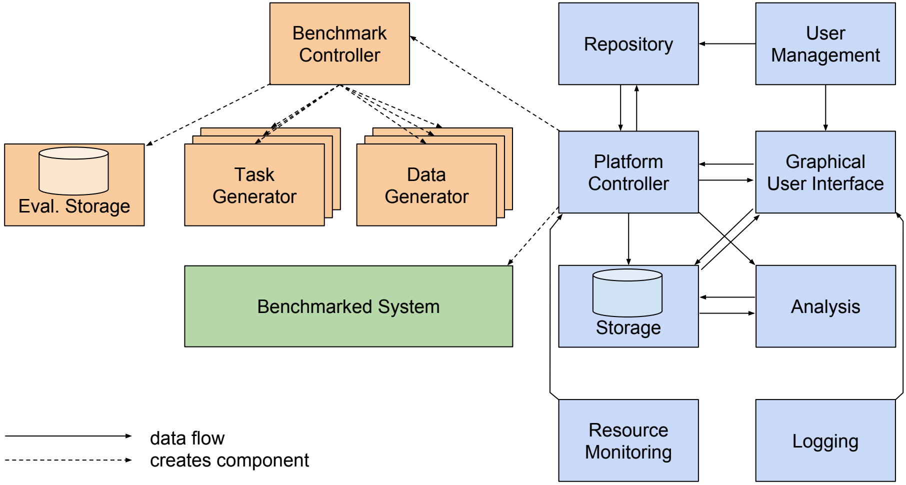


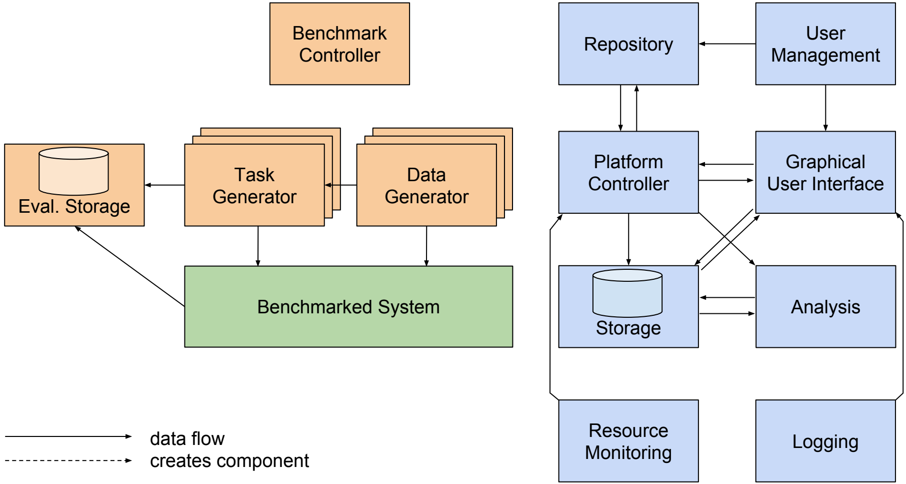
1. FAIR environment (Linked Data principles)
2. Fair conditions (separation of benchmark and systems)
3. Scalable (distributed execution and benchmarking)
4. Open (open-source, extensible)
5. Safe and secure (encryption,)
6. Portable (e.g., AWS)

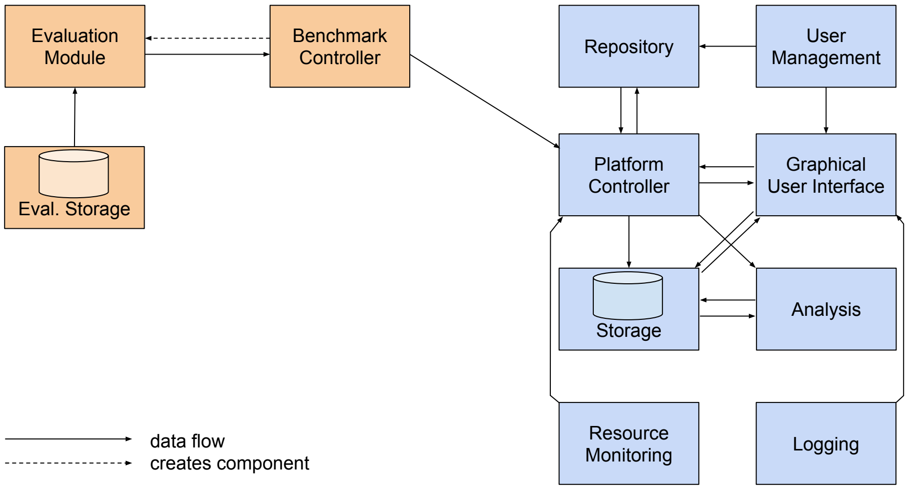




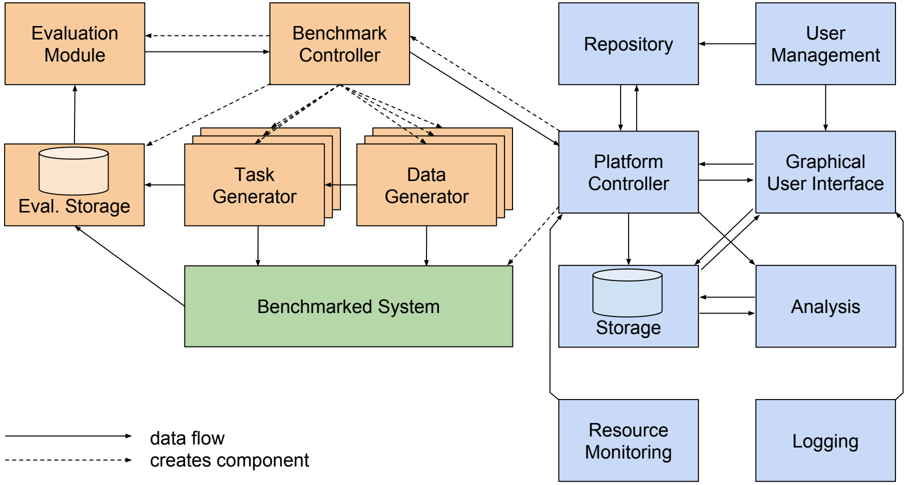
data flow
creates component


Benchmark Controller
Repository
User Management
Eval. Storage
Task Generator
Data Generator
Platform Controller
Graphical User Interface
Benchmarked System
Storage
Analysis
Resource Monitoring
Logging
data flow
creates component


Benchmark Controller
Eval. Storage
Task Generator
Data Generator
Benchmarked System
Repository
User Management
Platform Controller
Graphical User Interface
Storage
Analysis
Resource Monitoring
Logging
data flow
creates component

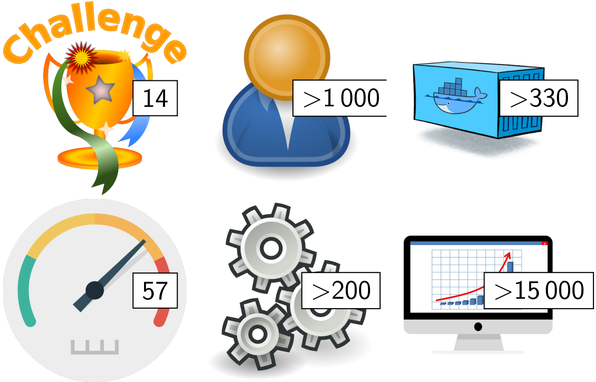


Evaluation Module
Benchmark Controller
Repository
User Management
Eval. Storage
Platform Controller
Graphical User Interface
Storage
Analysis
Resource Monitoring
Logging
data flow
creates component


Evaluation Module
Benchmark Controller
Repository
User Management
Eval. Storage
Task Generator
Data Generator
Platform Controller
Graphical User Interface
Benchmarked System
Storage
Analysis
Resource Monitoring
Logging
data flow
creates component

Platform

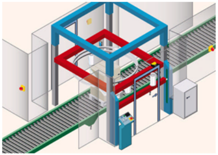
Current status



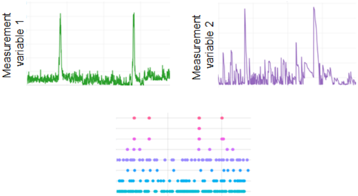
https://github.com/hobbit-project

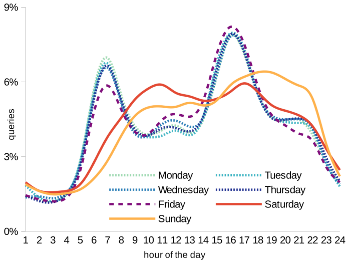
Data Generators

Industry 4.0, Transport

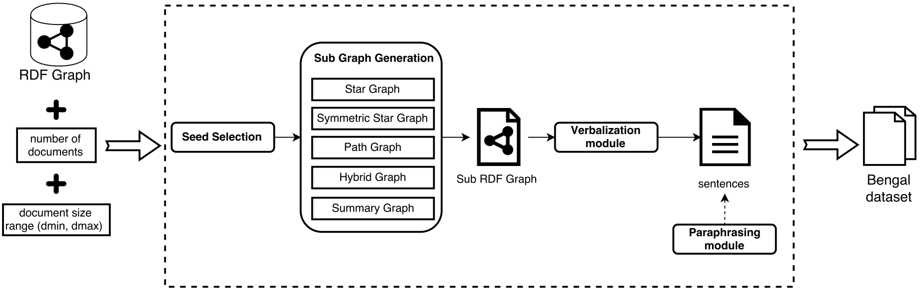


http://www.barr-thorp.com/wp-content/uploads/2012/09/Machine-Safeguarding-Image1.bmp





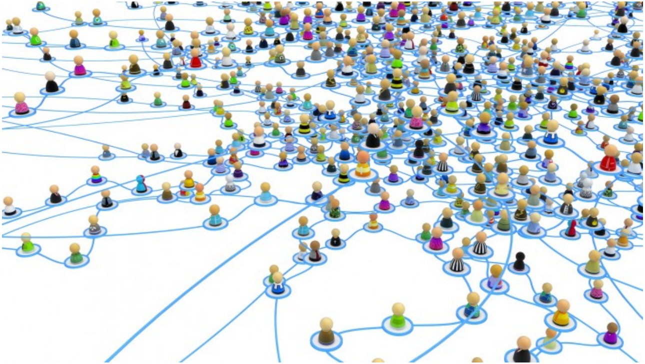
- 5 data generators based on real data
- Various domains (transport, industrial machinery, cars, social networks, ...)



- Dataset: 450×10^6 real tweets
- Classification: Entity recognition, disambiguation and relation extraction

Benchmarks

Storage



- Dataset: Extension of LDBC's SNB (scaling factor up to 30)
- Queries: Analytics and retrieval

- Dataset: 10^6 points, real marine traffic data from Big Data Ocean)
- Classification: Destination ports (name)
- Regression: Arrival time (timestamp)

Benchmarks

Visualisation and Interaction


```
SELECT DISTINCT ?uri WHERE {
    ?uri dbo:influencedBy dbr:J.R.R._Tolkien.
    ?uri dbo:award dbr:Hugo_Award
}
```

Automatic

What is person whose influenced by is J. R. R. Tolkien, and award is the Hugo Award.

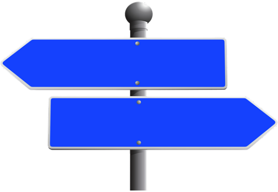
Manual

Name someone influenced by J. R. R. Tolkien, who won the Hugo Award?

- Dataset: 800 + 5000 question (manually edited + automatically generated) (scaling factor up to 30)
- Queries: question answering, keywords

Hobbit
Next Steps

1. KnowGraphs ITN (2019 – 2023)
   - 7 universities, 8 companies
   - AI applications on knowledge graphs
   - https://knowgraphs.eu/
2. Port to Kubernetes
3. Additional benchmarks
   - Embeddings
   - Machine translation
   - Natural language generation
   - Question answering
   - ...



BenchCouncil AI and Big Data Benchmarks

Wanling Gao

European Big Data Value Forum (EBDVF)

2020.11.4

Acknowledgement

- Thanks for the invitation of Dr. Arne J. Berre and Dr. Todor Ivanov

- Thanks for the forum organization of DataBench Toolbox group

BenchCouncil

- **International Open Benchmark Council (BenchCouncil)**
  - http://www.benchcouncil.org
  - a non-profit international organization
    - Aiming to promote the standardization, benchmarking, evaluation, incubation, and promotion of Chip, AI, Big Data, Block Chain, and other emerging techniques.

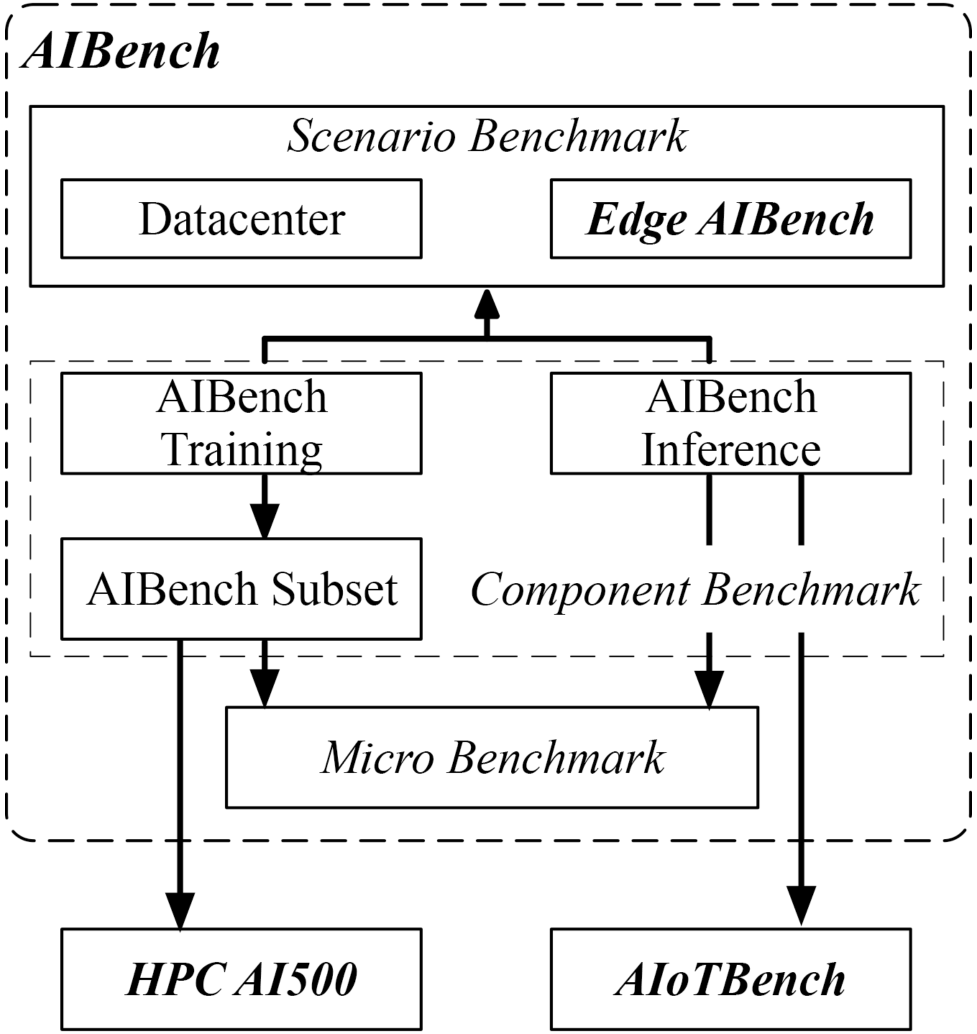
- Fundamental responsibilities
  - incubates benchmark projects and hosts the BenchCouncil benchmark projects
  - encourages benchmark-based quantitative approaches to tackle multi-disciplinary challenges.

AI Benchmarking Targets

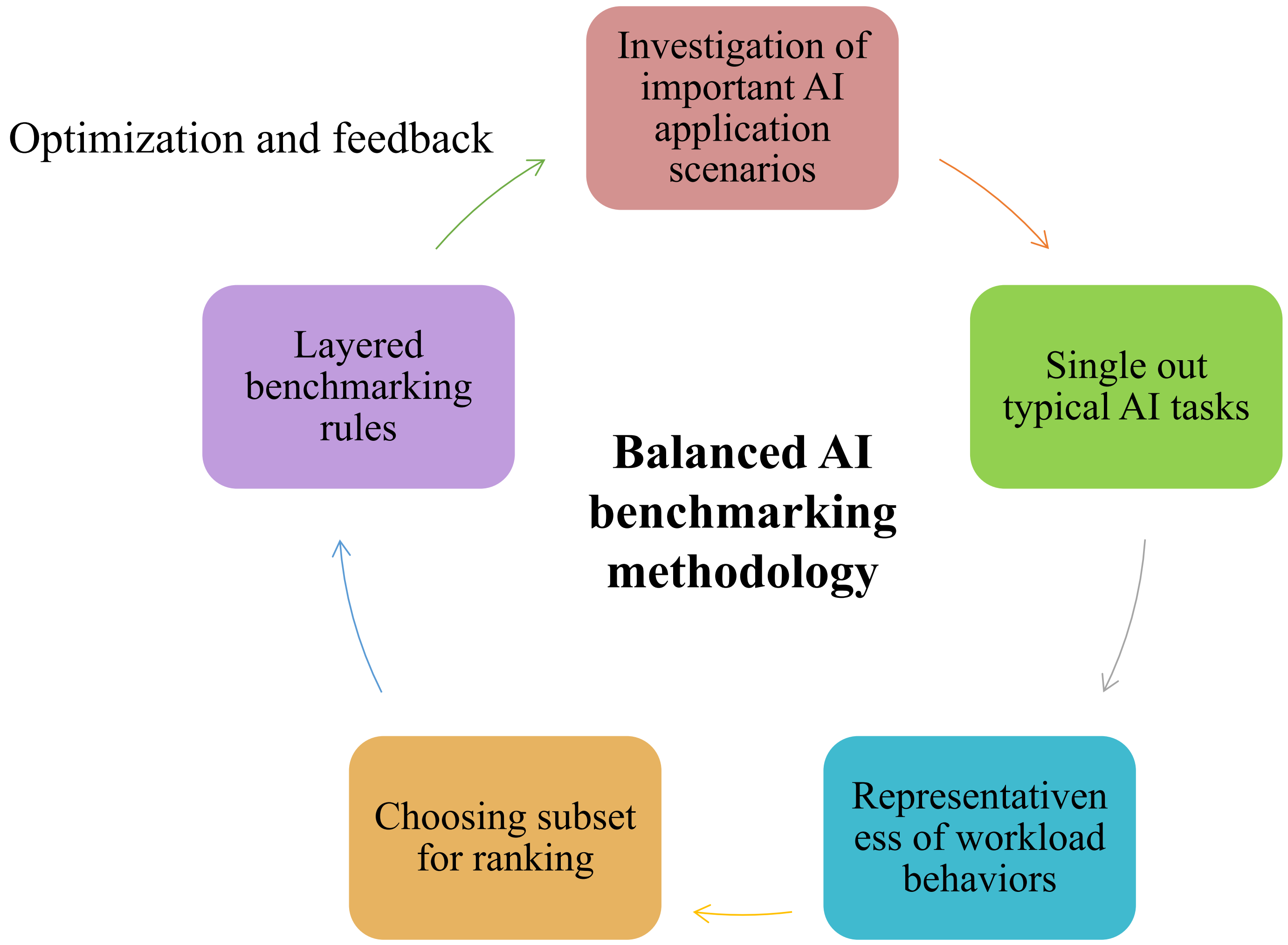
- Diverse application domains
  - Datacenter、HPC、 AIoT、Edge
- Balanced methodology that considers different benchmarking requirements
  - *Scenario benchmarks*
    - **The first AI benchmark** that provides real-world scenario modelling
      - E.g., the complete use cases of autonomous driving scenario in edge computing
    - For overall system evaluation
  - *Component benchmarks*
    - Comprehensive workload behaviors
      - Algorithm/System/Micro-architectural Characteristics
    - Providing component subset for ranking
      - Fairness, affordability, representativeness
  - *Micro benchmarks*
    - Hotspot functions and code optimizations

BenchCouncil AI Benchmark Suites

- AIBench
  - Benchmarking datacenter AI systems and chips
- HPC AI500
  - Benchmarking HPC AI Systems
- Edge AIBench
  - Benchmarking Edge Computing
- AIoTBench
  - Benchmarking Mobile and Embedded device Intelligence



Benchmarking Methodology



Investigation of Internet Service Applications

Internet Service	Core Scenario	Involved AI Tasks
Search Engine	Content-based image retrieval (e.g., face, scene)	Object detection; Classification; Spatial transformer; Face embedding; 3D face recognition
	Advertising and recommendation	Recommendation
	Maps search and translation	3D object reconstruction; Text-to-Text translation; Speech recognition; Neural architecture search
	Data annotation and caption (e.g., text, image)	Text summarization; Image-to-Text
	Search result ranking	Learning-to-rank
	Image resolution enhancement	Image generation; Image-to-Image
	Data storage space and transfer optimization	Image compression; Video prediction
Social Network	Friend or community recommendation	Recommendation; Face embedding; 3D face recognition;
	Vertical search (e.g., image, people)	Classification; Spatial transformer; Object detection;
	Language translation	Text-to-Text translation; Neural architecture search
	Automated data annotation and caption	Text summarization; Image-to-Text; Speech recognition
	Anomaly detection (e.g., spam image detection)	Classification
	Image resolution enhancement	Image generation; Image-to-Image
	Photogrammetry (3D scanning)	3D object reconstruction
	Data storage space and transfer optimization	Image compression; Video prediction
	News feed ranking	Learning-to-rank
E-commerce	Product searching	Classification; Spatial transformer; Object detection
	Product recommendation and advertising	Recommendation
	Language and dialogue translation	Text-to-Text translation; Speech recognition; Neural architecture search
	Automated data annotation and caption	Text summarization; Image-to-Text
	Virtual reality (e.g., virtual fitting)	3D object reconstruction; Image generation; Image-to-Image
	Data storage space and transfer optimization	Image compression; Video prediction
	Product ranking	Learning to rank
	Facial authentication and payment	Face embedding; 3D face recognition;

Typical AI Tasks

- Cover mainstream neural network models（CNN, ResNet, LSTM, GRU, Attention, etc）
  - Text processing (5)
    - Text-to-Text Translation, Text Summarization, Learning-to-Rank, Recommendation, Neural Architecture Search
  - Image processing (8)
    - Image Classification, Image Generation, Image-to-Text, Image-to-Image Translation, Face Embedding, Object Detection, Image Compression, Spatial Transformer
  - Audio processing (1)
    - Speech Recognition
  - Video processing(1)
    - Video Prediction
  - 3D data processing (2)
    - 3D Face Recognition, 3D Object Reconstruction

Bench Council

Workloads of AIBench Training

No.	Component Benchmark	Algorithm	Data Set	Target Quality
TrC1	Image Classification	ResNet50 [38]	ImageNet	74.9% (accuracy)
TrC2	Image Generation	WassersteinGAN [18]	LSUN	N/A
TrC3	Text-to-Text translation	Transformer [64]	WMT English-German	55% (accuracy)
TrC4	Image-to-Text	Neural Image Caption Model [66]	Microsoft COCO	4.2 (perplexity)
TrC5	Image-to-Image Translation	CycleGAN [73]	Cityscapes	N/A
TrC6	Speech Recognition	DeepSpeech2 [17]	Librispeech	23.5% (WER)
TrC7	Face Embedding	Facenet [60]	VGGFace2, LFW	90% (accuracy)
TrC8	3D Face Recognition	3D face models [65]	77,715 samples from 253 face IDs	94.64% (accuracy)
TrC9	Object Detection	Faster R-CNN [56]	VOC2007	76% (mAP)
TrC10	Recommendation	Neural collaborative filtering [39]	MovieLens	63.5% (HR@10)
TrC11	Video Prediction	Motion-Focused predictive models [33]	Robot pushing data set	72 (MSE)
TrC12	Image Compression	Recurrent neural network [63]	ImageNet	0.99 (MS-SSIM)
TrC13	3D Object Reconstruction	Convolutional encoder-decoder network [68]	ShapeNet Data set	45.83% (IU)
TrC14	Text Summarization	Sequence-to-sequence model [51]	Gigaword data set	41 (Rouge-L)
TrC15	Spatial Transformer	Spatial transformer networks [42]	MNIST	99% (accuracy)
TrC16	Learning-to-Rank	Ranking distillation [62]	Gowalla	14% (accuracy)
TrC17	Neural Architecture Search	Efficient neural architecture search [55]	PTB [50]	100 (perplexity)

Workload Characterization

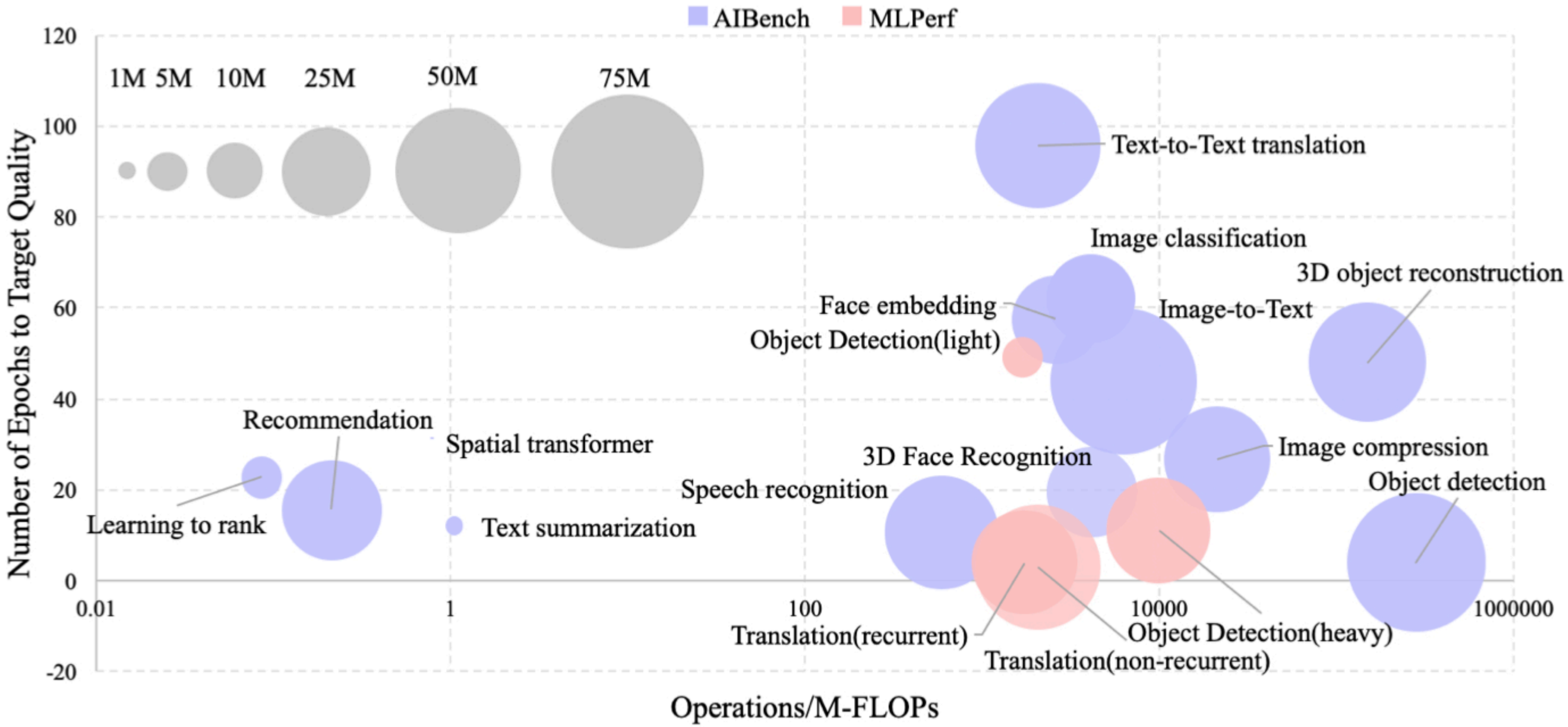
- Micro-architecture level
  - FLOPs computation、memory access pattern、computation pattern、I/O pattern
- System level
  - Throughput、run-to-run variation、and convergence characteristics
- Algorithm level
  - Model architecture, model complexity (parameters)

AIBench Training vs. MLPerf Training

- Concurrent work
- AIBench has wider coverage

		AIBench Training v1.0	MLPerf Training V0.5
Methodology		Balanced methodology considering conflicting requirements	According to commercial and research relevance
Algorithm		Seventeen tasks and models	Five tasks and seven models
Dataset		Text, image, 3D, audio, and video data	Text and image data
Model behavior	Computation	0.09 to 282830 MFLOPs	0.21 to 24500 MFLOPs
	Complexity	0.03 to 68.4 million parameters	5.2 to 49.53 million parameters
	Convergence	6 to 96 epochs	3 to 49 epochs
System behavior		30 hot functions	9 hot functions
Micro-architecture behavior	Achieved occupancy	0.14 to 0.61	0.28 to 0.54
	IPC efficiency	0.25 to 0.77	0.39 to 0.74
	Gld efficiency	0.28 to 0.94	0.52 to 0.85
	Gst efficiency	0.27 to 0.98	0.75 to 0.98
	DRAM utilization	0.12 to 0.61	0.52 to 0.61

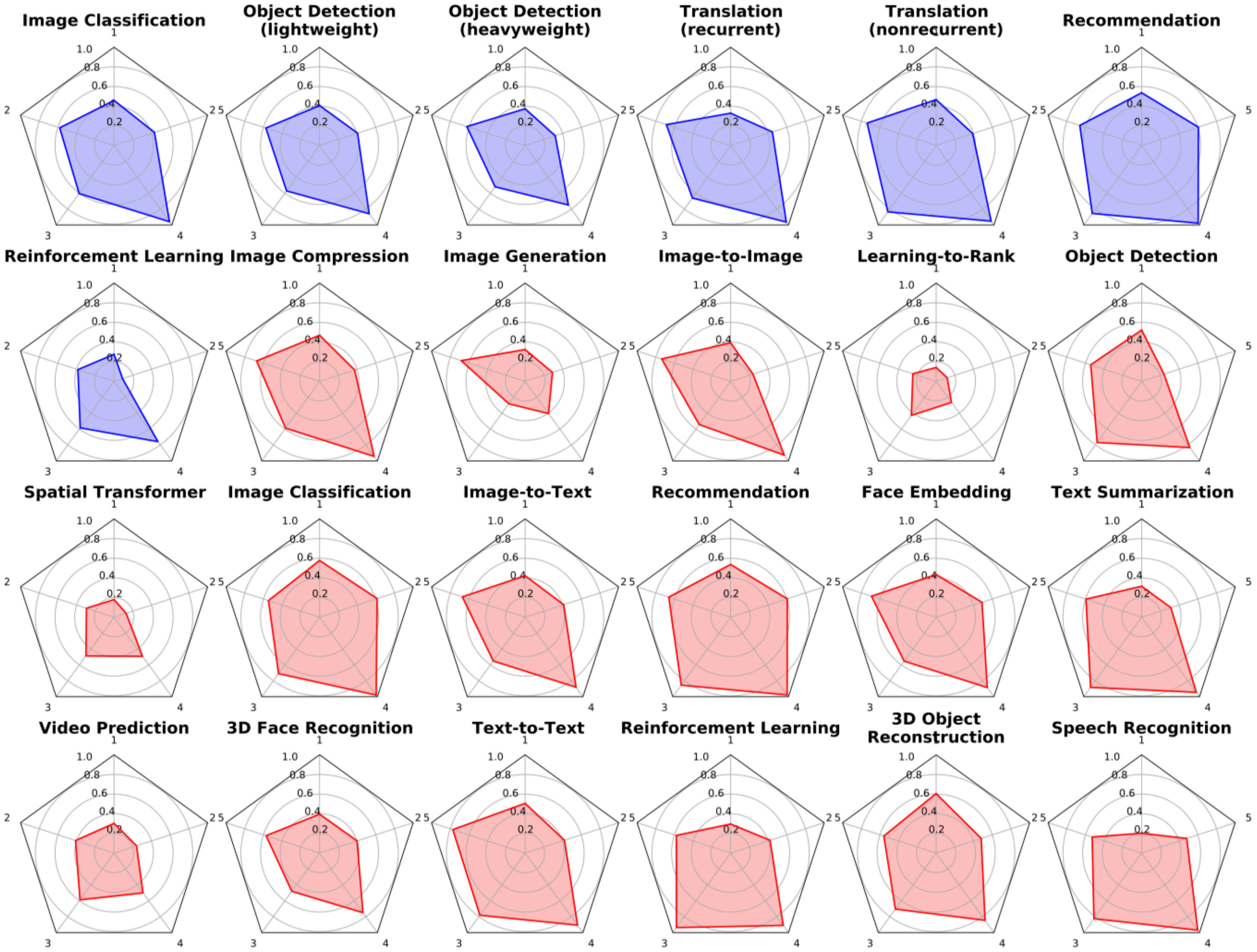
Comparisons of AIBench against MLPerf



The Comparisons of AIBench against MLPerf from the Perspectives of Model Complexity, Computational Cost, and Convergent Rate.

Bench Council

Micro-architectural Comparison



1: achieved occupancy

2: ipc efficiency

3: gld efficiency

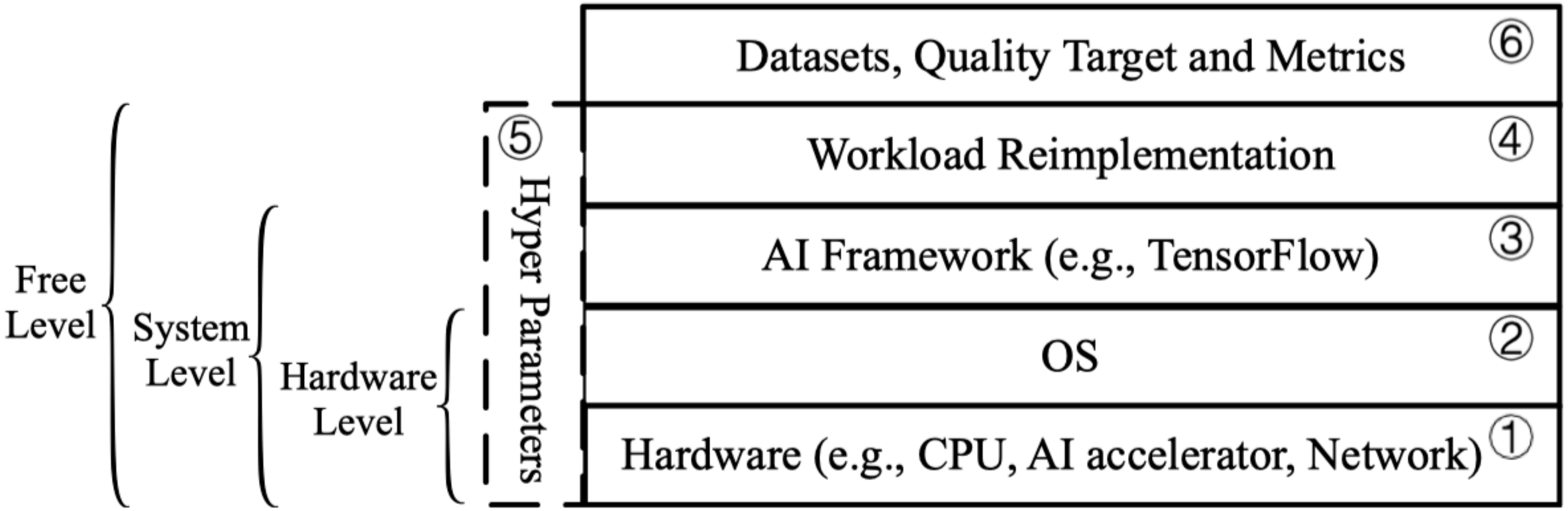
4: gst efficiency

5: dram utilization

BenchCouncil

Layered benchmarking rules

- Hardware Level
  - allows the modifications of hardware, OS, and hyper-parameter layers, with the other layers unchanged
- Software Level
  - allows the modifications of hardware, OS, AI framework, and hyper-parameter layers, while the others are fixed.
- Free Level
  - allows the modifications of all layers except for the datasets and metrics layer.



AIBench Subset for Ranking

- A minimum set to represent the maximum workload characteristics
  - Algorithm/AI model: FLOPs、parameter Size、epochs to accuracy
  - System/Micro-architectural： system occupancy、IPC、load、store、dram utilization
- Widely accepted evaluation metrics
  - E.g., GAN-based model has no widely accepted metrics, which may incurs inconsistence for different users
- Repeatability with low run-to-run variation

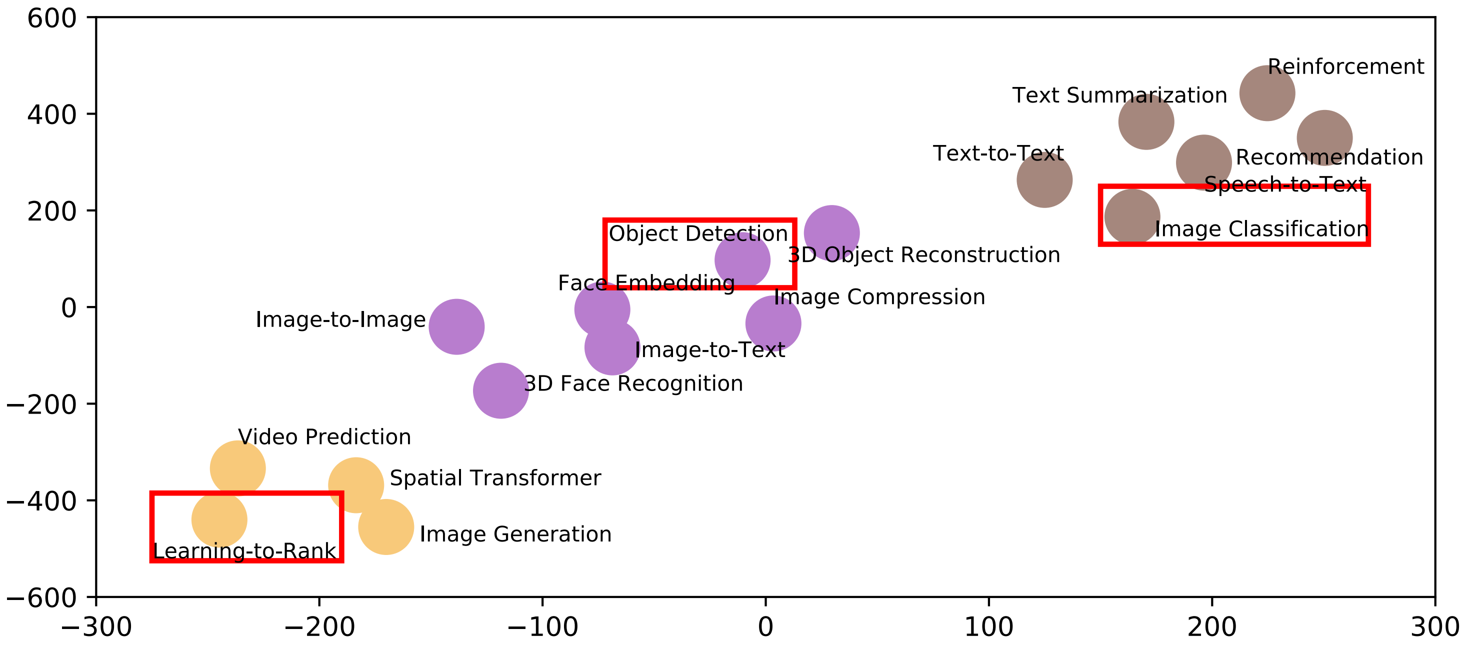
Bench Council

Run-to-run Variation

No.	Component Benchmark	Time Per Epoch (second)	Total Time (hour)	Variation	Repeat Times
TrC1	Image Classification	4440	76.25	1.12%	5
TrC2	Image Generation	3935.75	N/A	N/A	N/A
TrC3	Text-to-Text translation	64.83	1.72	9.38%	6
TrC4	Image-to-Text	845.02	10.21	23.53%	5
TrC5	Image-to-Image	251.67	N/A	N/A	N/A
TrC6	Speech Recognition	14326.86	42.78	12.08%	4
TrC7	Face Embedding	214.73	3.43	5.73%	8
TrC8	3D Face Recognition	36.99	12.02	38.46%	4
TrC9	Object Detection	1859.96	2.06	0	10
TrC10	Recommendation	36.72	0.16	9.95%	5
TrC11	Video Prediction	24.99	2.11	11.83%	4
TrC12	Image Compression	763.44	5.67	22.49%	4
TrC13	3D Object Reconstruction	28.41	0.38	16.07%	4
TrC14	Text Summarization	1923.33	6.41	24.72%	5
TrC15	Spatial Transformer	6.38	0.06	7.29%	4
TrC16	Learning-to-Rank	60.1	0.14	1.90%	4
TrC17	Neural Architecture Search	932.79	7.47	6.15%	6

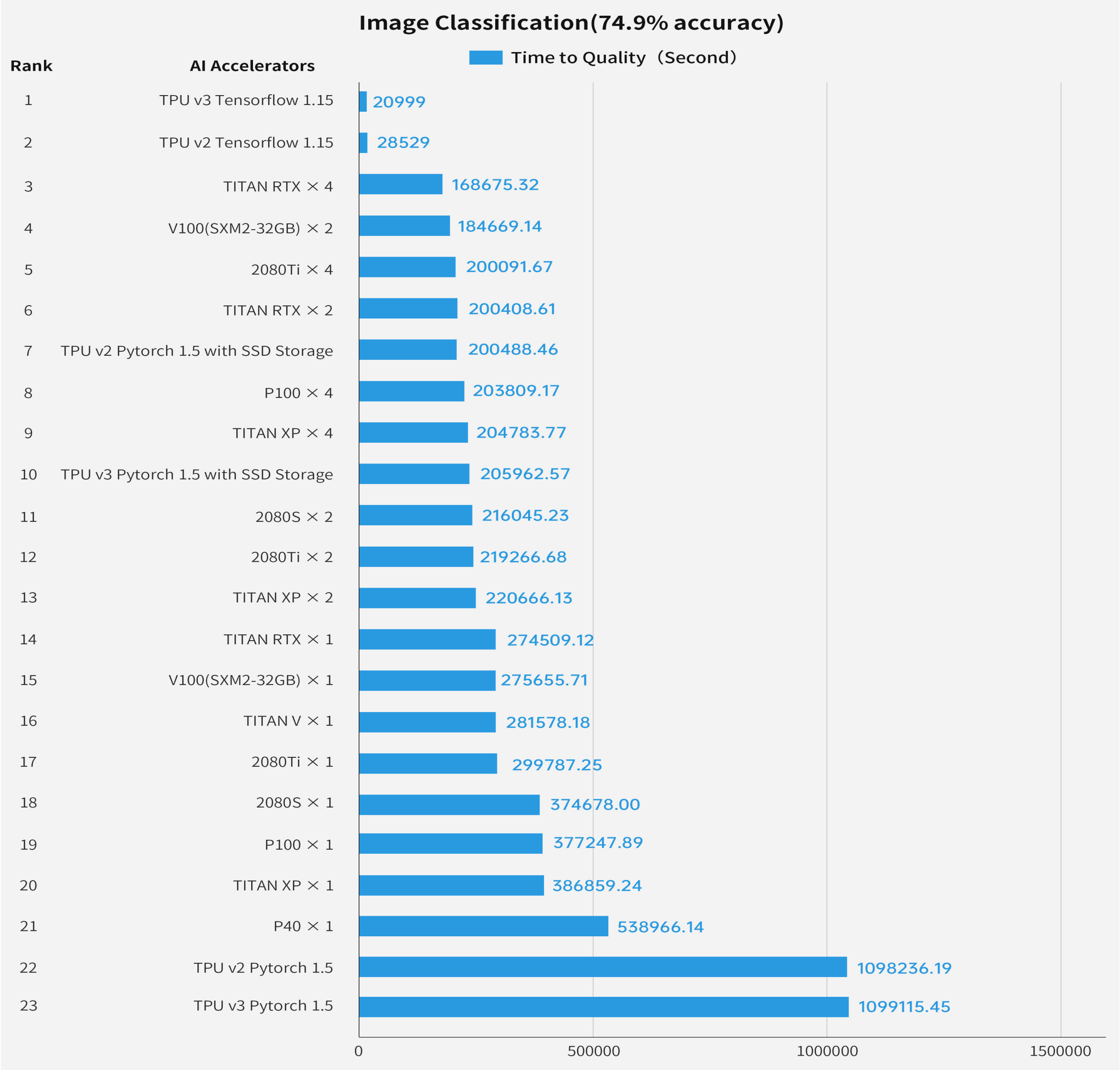
AIBench Subset

- K-Means Clustering（system occupancy、IPC、load、store、dram utilization）



- To achieve repeatability, affordability, representativeness
  - Image classification, Object detection, Learning-to-Rank

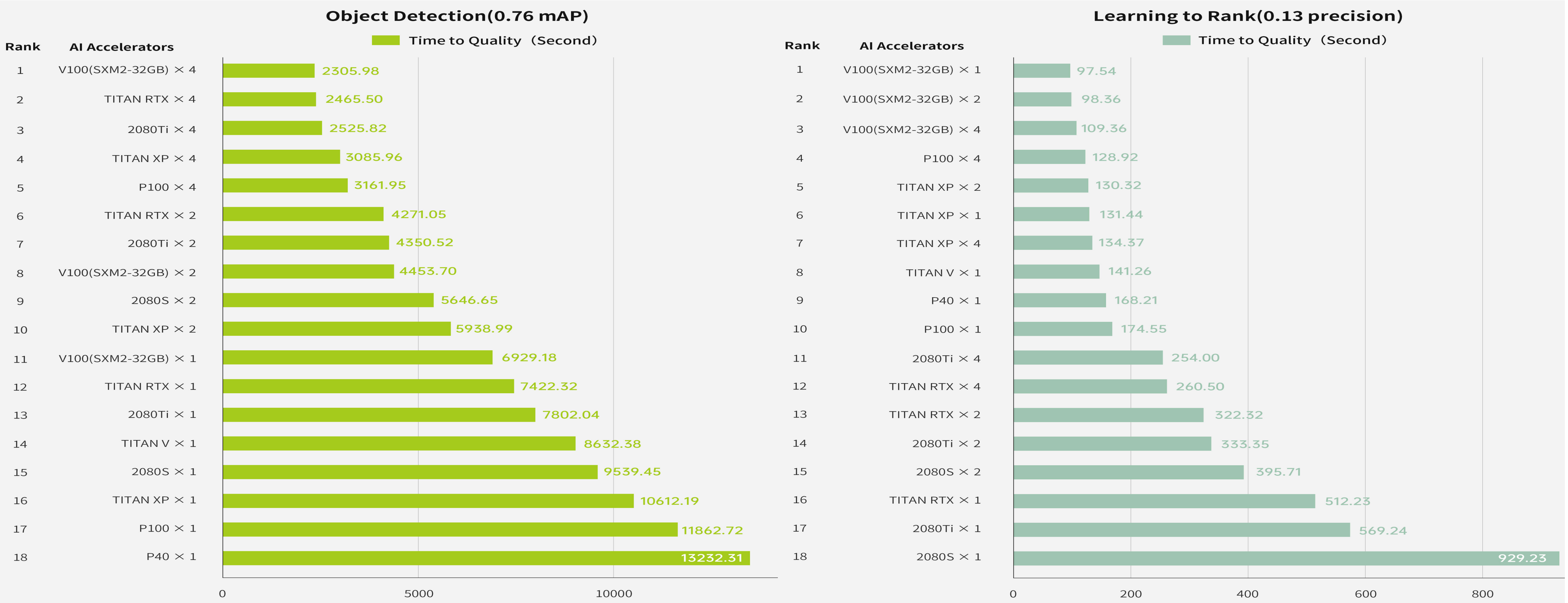
Intelligent Chips Ranking – Image Classification



BenchCouncil

Intelligent Chips Ranking

■ Object detection & Learning-to-Rank



Bench Council

HPC AI500 For Benchmarking HPC AI Systems

- Metrics
  - Representativeness, Affordability
  - Repeatability
  - Computation
  - Tasks, Models, Datasets
  - Scalability

- AIBench Subset nearly satisfies the above all metrics

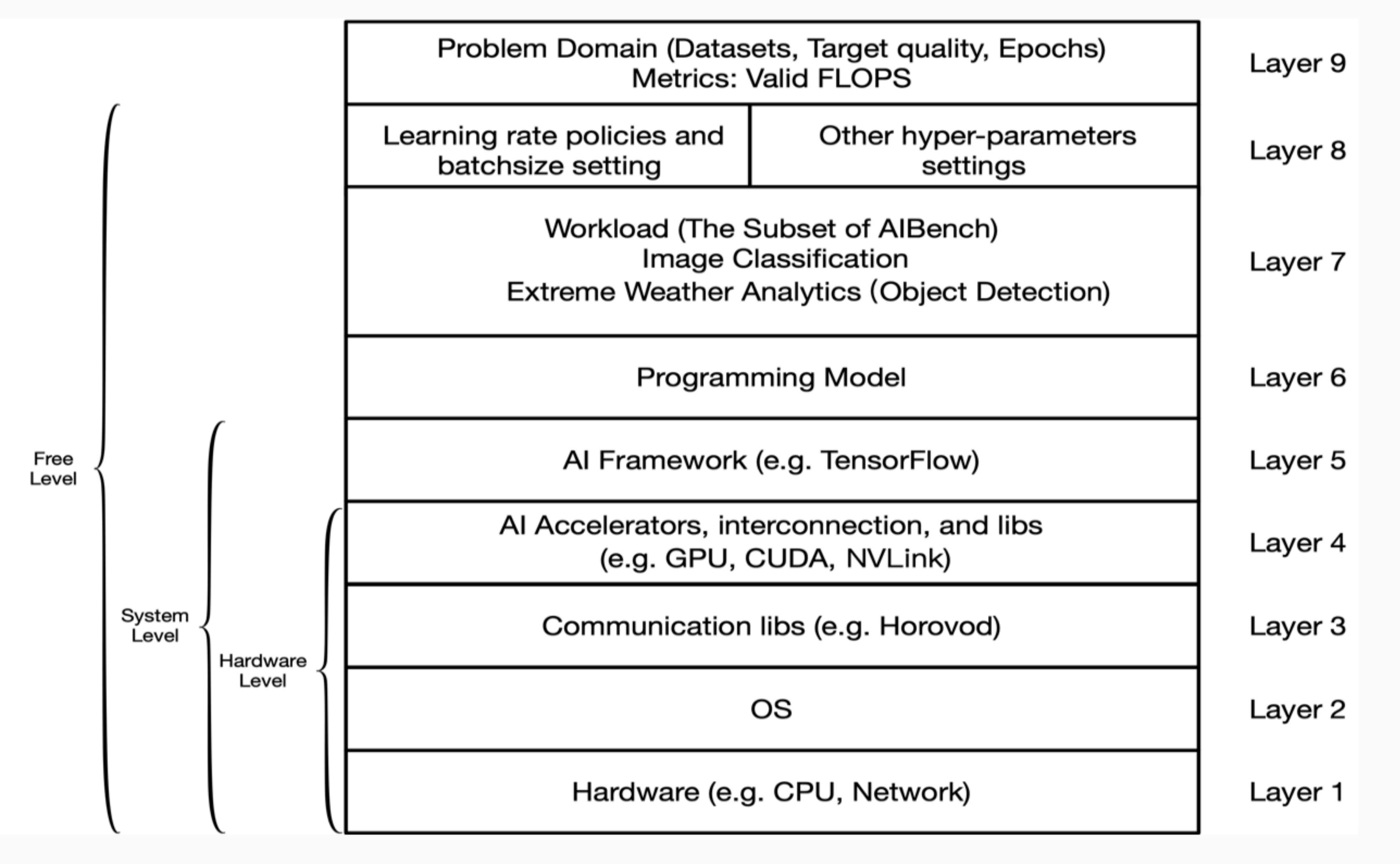
Computation

- The computation complexity of AIBench subset

Workload	FLOPs (Single batch)
Image Classification	23 G
Object Detection	691 G
Learning to Rank	0.08 M

Image Classification and Object Detection suit for the requirements of HPC AI benchmarking

Layered benchmarking rules of HPC AI500



Metrics for HPC AI500

- Considering both computing complexity and model quality: VFLOPS

$$VFLOPS = FLOPS \times \left(\frac{achieved\_quality}{target\_quality}\right)^n$$

achieved_quality represents the actual model quality achieved in the evaluation ;

target_quality represents the state-of-the-art model quality that has been predefined in HPC AI500 benchmark ;

The value of n is apositive integer, which is used to define the sensitivity to the model quality.。

HPC AI500 Introduction

- Workload
  - Extreme Weather Analysis， EWA in short
    - To identify various extreme weather patterns (e.g. tropical depression), which is essentially object detection
  - Image Classification
    - A fundamental task in AI research
- Datasets
  - The extreme weather dataset: 16 channels, 768*1052, 2 TB
    - The first AI benchmark for HPC that uses the real-world scientific dataset
  - ImageNet 2012: 3 channels, 256*256, 136 GB
- Models
  - Faster-RCNN
  - ResNet-50 V1.5

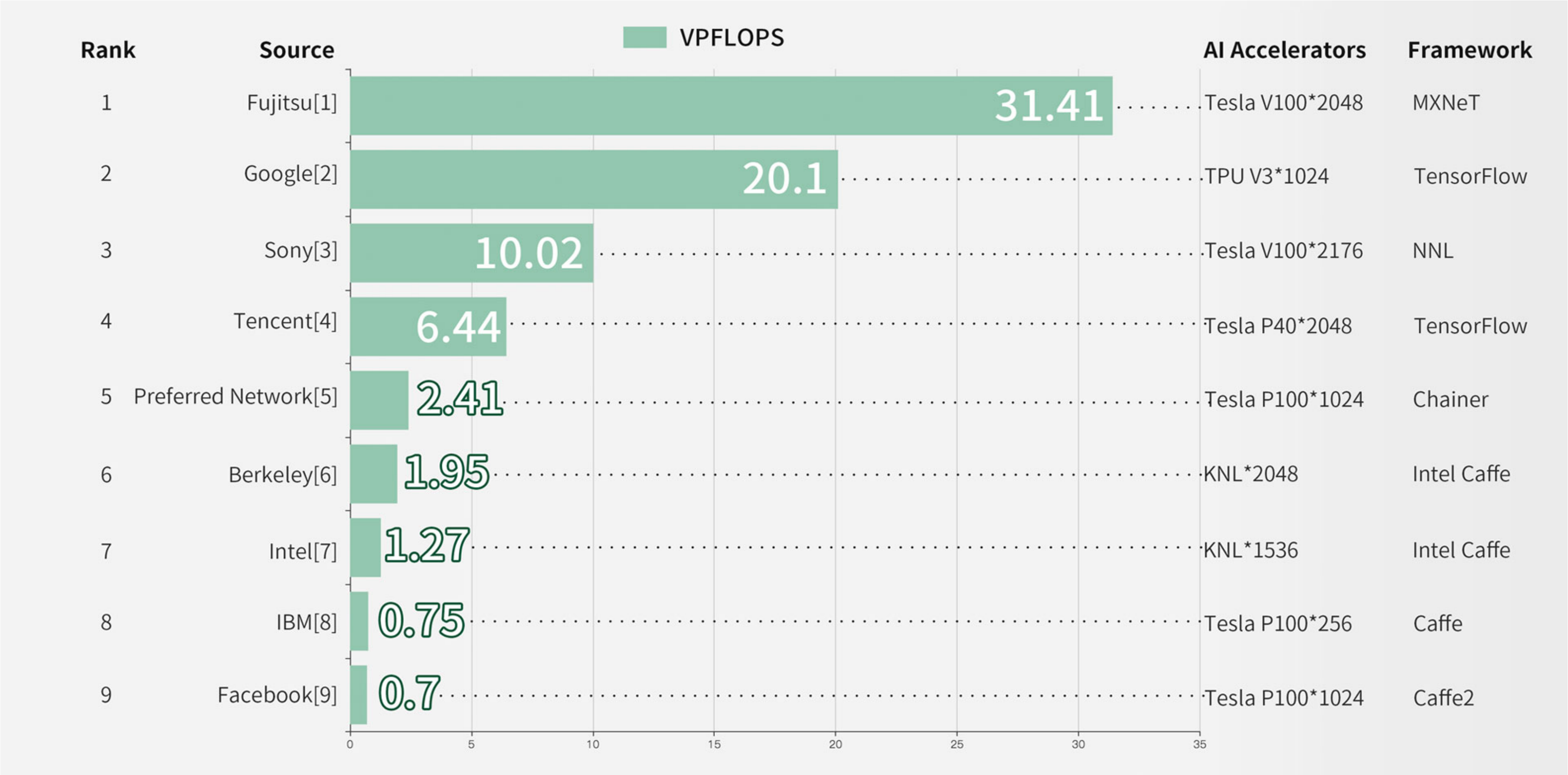
HPC AI500 Workloads

Problem domain	Model	Dataset	Target quality	Epochs	Communication	Library	AI framework
EWA	Faster-RCNN	The extreme weather dataset	mAP@[IoU=0.5]=0.35	50	MPI、NCCL2	CUDA, cuDNN, NCCL	TensorFlow
Image Classification	ResNet50 V1.5	ImageNet 2012	TOP1 Accuracy= 0.763	90			

BenchCouncil

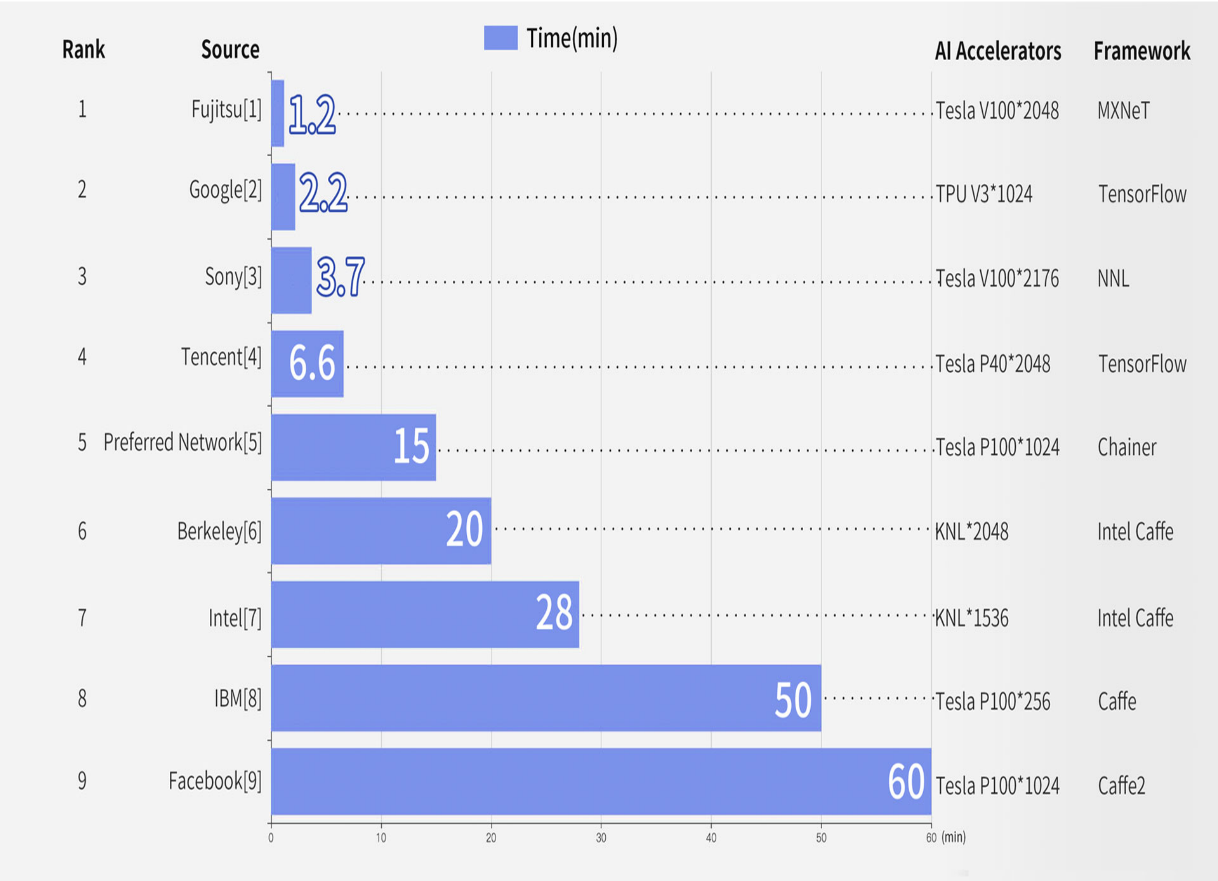
The First HPC AI Ranking

HPC AI500, Image Classification, Free level

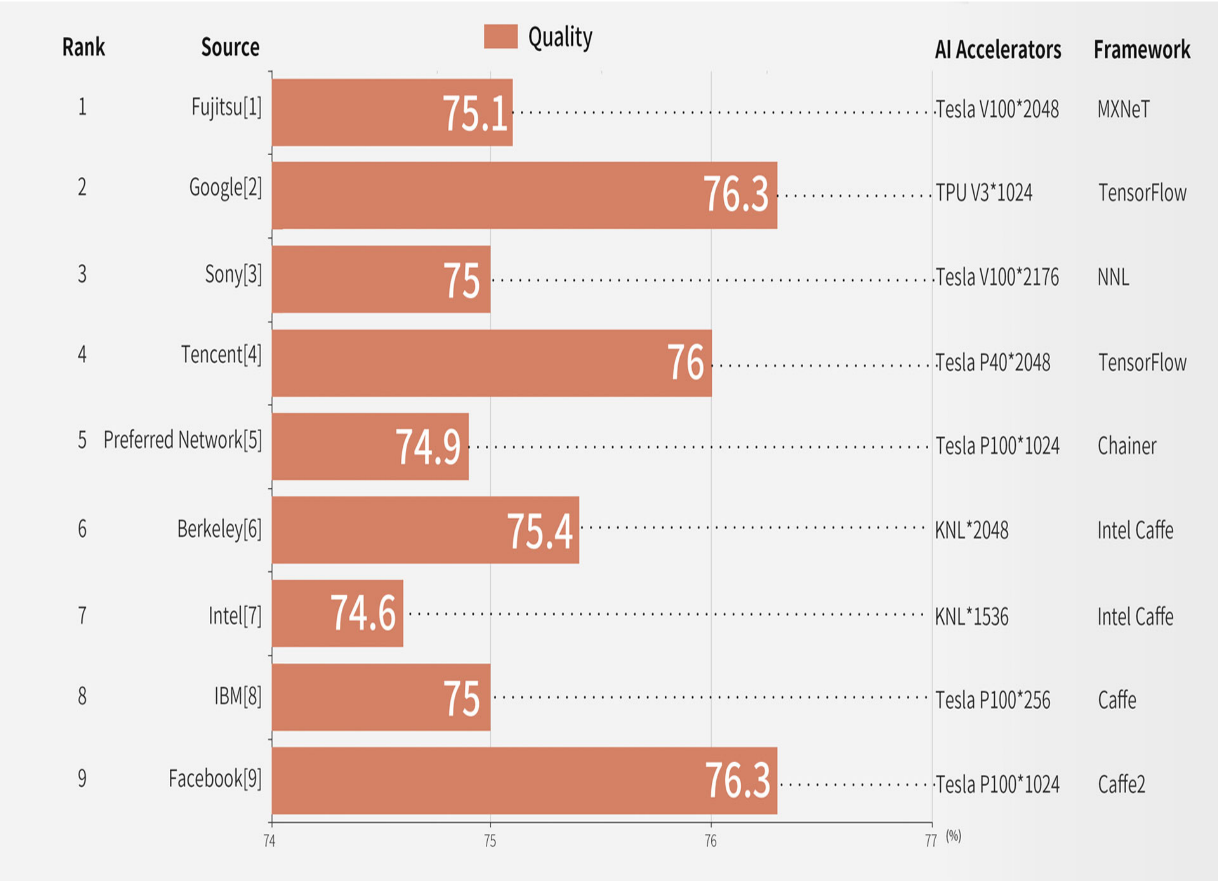


The First HPC AI Ranking

Training time



Model Quality



Conclusion

- BenchCouncil AI Benchmarks
  - https://www.benchcouncil.org/aibenchmark.html
  - *AIBench: An Industry Standard AI Benchmark Suite from Internet Services*
    - *https://arxiv.org/abs/2004.14690*
  - *HPC AI500: The Methodology, Tools, Roofline Performance Models, and Metrics for Benchmarking HPC AI Systems*
    - https://www.benchcouncil.org/file/HPC_AI500TR.pdf
  - Ranking
    - https://www.benchcouncil.org/ranking.html

Thank you！

Bench
Council

Framework for Benchmarking ML/DL Workloads

Rekha Singhal

Senior Scientist & Head Computing Systems, Tata Consultancy Services

Rekha.Singhal@tcs.com

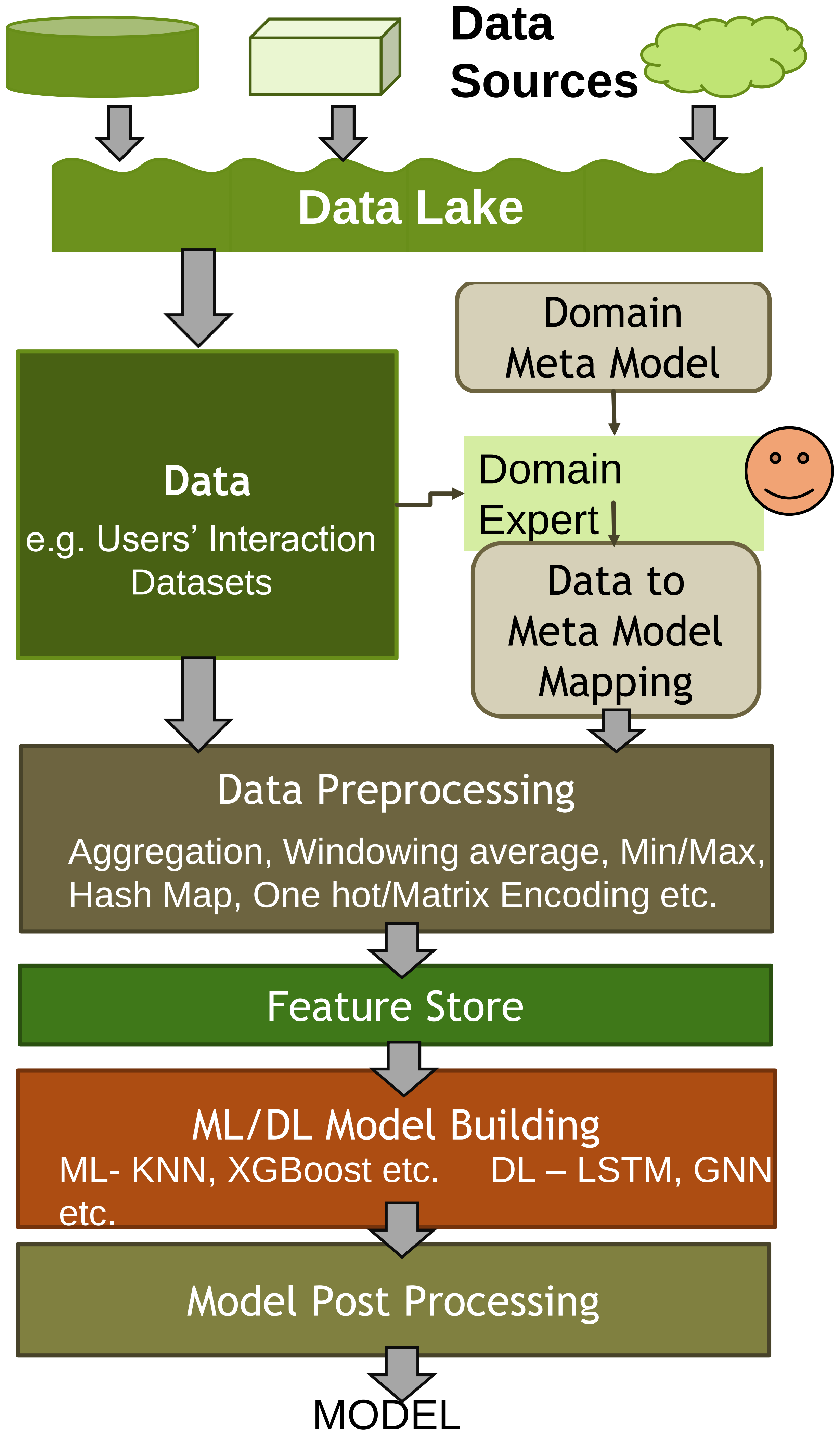
TATA CONSULTANCY SERVICES

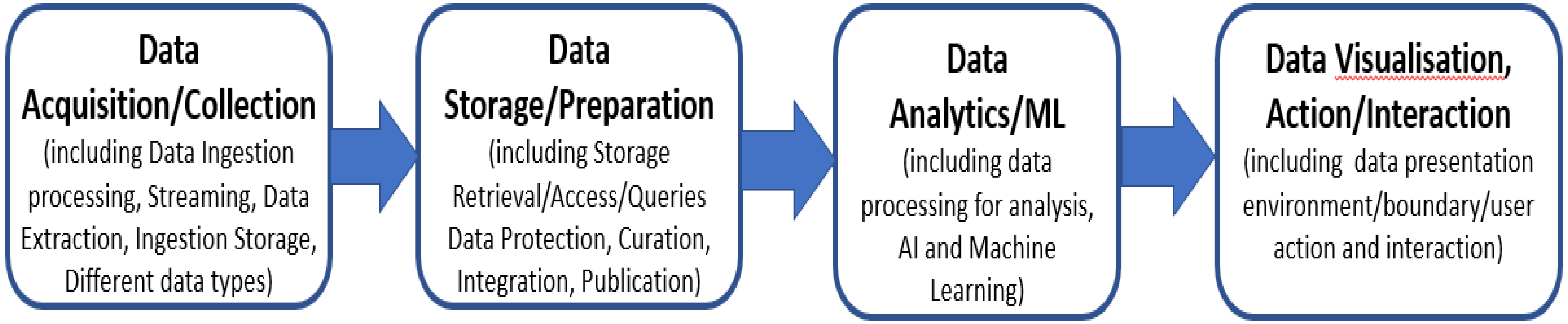
ML/DL Pipeline

New challenges for ML/DL workloads

- Heterogeneous hardware- CPU, GPU., TPU, ASIC
- Many frameworks- Pytorch, Tensorflow
- Heterogenous big data technology in the Pipeline
- Many kinds of NN models - RNN, CNN, LSTM, RL…..

iPrescribe: Recommendation Benchmark





SparkScala, SparkPython, Ignite

MongoDB, Ignite

Python(Training) Tornado (Inference)

Ref: Fast Online 'Next Best Offers' using Deep Learning, ACM COMAD 2019

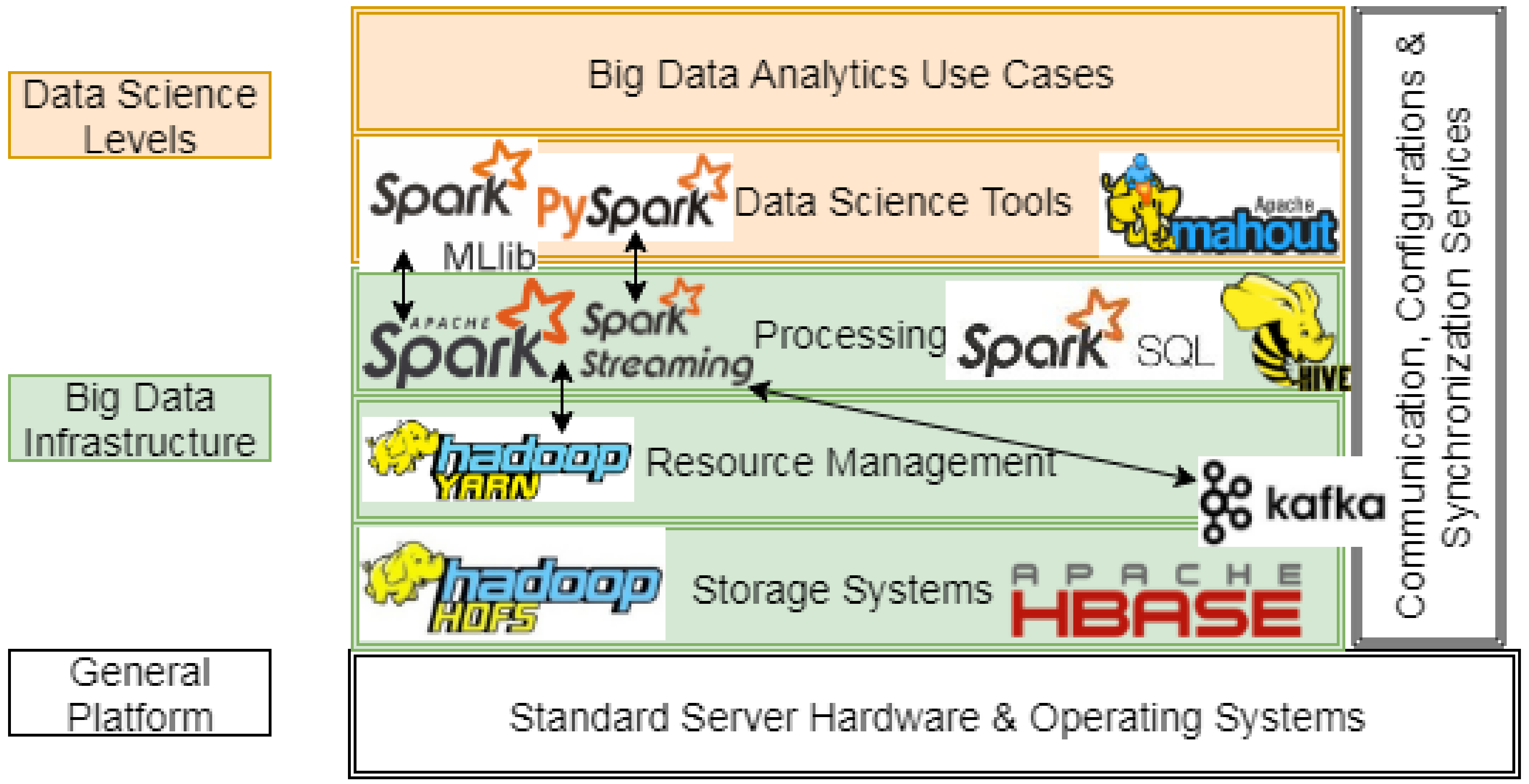
ABench

ABench: Big Data Architecture Stack Benchmark *
Todor Ivanov and Rekha Singhal

* Published in Proceedings of the 9th ACM/SPEC International Conference on Performance Engineering (ICPE 2018), April 9-13, Berlin, Germany, 2018

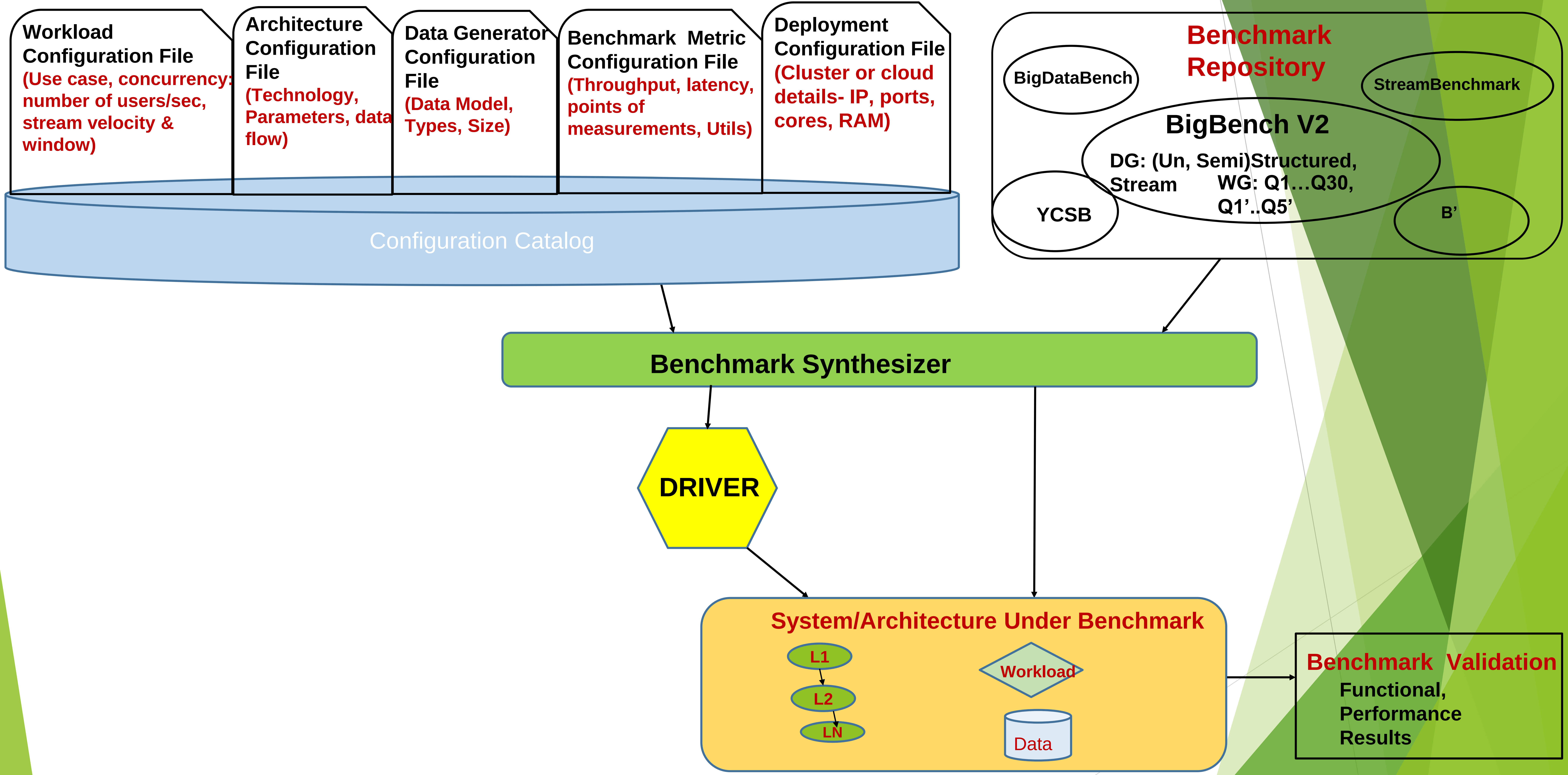
Motivation

- Growing number of new ***Big Data technologies*** and ***connectors*** in the Big Data Stacks
  - Challenges for Solution Architects, Data Engineers, Data Scientist, Developers, etc.



- Missing benchmarks for ***each technology***, ***connector*** or a ***combination of them***

- Consequence: ***Increasing complexity in the Big Data Architecture Stacks***


Workload Configuration File
(Use case, concurrency: number of users/sec, stream velocity & window)
Architecture Configuration File
(Technology, Parameters, data flow)
Data Generator Configuration File
(Data Model, Types, Size)
Benchmark Metric Configuration File
(Throughput, latency, points of measurements, Utils)
Deployment Configuration File
(Cluster or cloud details- IP, ports, cores, RAM)
Configuration Catalog
Benchmark Repository
BigDataBench
StreamBenchmark
BigBench V2
DG: (Un, Semi)Structured, Stream
WG: Q1…Q30, Q1'..Q5'
YCSB
B'
Benchmark Synthesizer
DRIVER
System/Architecture Under Benchmark
L1
L2
LN
Workload
Data
Benchmark Validation
Functional, Performance Results

MLPerf

https://mlperf.org/

MLPerf Inference Benchmark - Vijay J. Reddi et al. http://arxiv.org/abs/1911.02549
MLPerf Training Benchmark - Peter Mattson et al. http://arxiv.org/abs/1910.01500

MLPerf Goals

- Accelerate progress in ML via fair and useful measurement
- Serve both the commercial and research communities
- Enable fair comparison of competing systems yet encourage innovation to improve the state-of-the-art of ML
- Enforce replicability to ensure reliable results
- Keep benchmarking effort affordable (so all can play)

MLPerf

Benchmarks Considered for MLPerf

Area	*Vision*	*Language*	*Audio*	*Commerce*	*Action / RL*	*Other*
Problem	Image Classification Object Detection / Segmentation Face ID HealthCare (Radiology) Video Detection Self-Driving	Translation Language Model Word Embedding	Speech Recognition Text-to-Speech Question Answering Keyword Spotting Language Modeling Chatbots Speaker ID Graph embeddings Content ID	Rating Recommendations Sentiment Analysis Next-action Healthcare (EHR) Fraud detection Anomaly detection Time series prediction Large scale regression	Games Go Robotics Health Care Bioinformatics	GANs 3D point clouds Word embeddings
Datasets	ImageNet COCO	WMT English-German	LibriSpeech SQuAD LM-Benchmark	MovieLens-20M Amazon IMDB	Atari Go Chess Grasping	
Models	ResNet-50 TF Object Detection Detectron	Transformer OpenNMT	Deep Speech 2 SQuAD Explorer	Neural Collaborative Filtering CNNs	DQN PPO	
Metrics	COCO mAP Prediction accuracy	BLEU	WER Perplexity	Prediction accuracy	Prediction accuracy Win/Loss	

MLPerf metric: **Training time** to reach quality target + cost **or** power

- Quality target is *specific for each benchmark* and *close to state-of-the-art*
    - Updated w/ each release to keep up with the state-of-the-art
- Time includes preprocessing, validation over median of 5 runs
- Available: reference implementations that achieve quality target

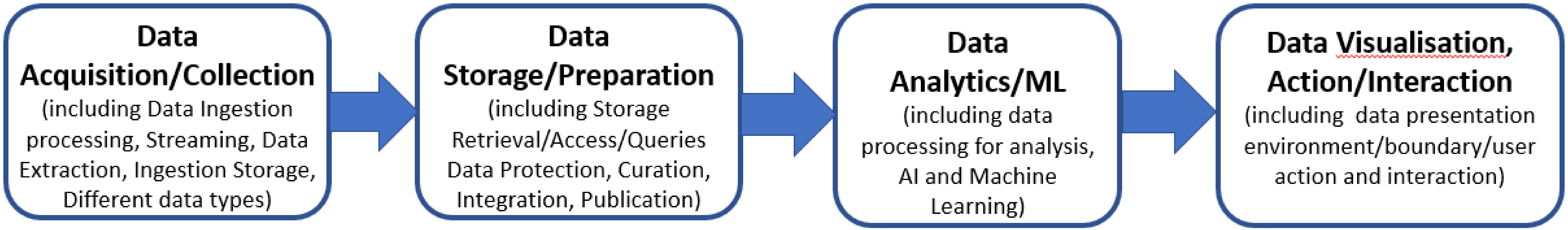
In addition, *either*:

- *Cost* of public cloud resources *(no spot/preemptible instances)*
- *Power utilization* for on-premise hardware

Important for benchmark to capture
both performance and quality

MLPerf

Conclusions



Recommendation – ML/DL	Recommendation – ML/DL	Recommendation – ML/DL
Supply Chain- RL	Supply Chain- RL	Supply Chain- RL (distributed training)
Recommendation –RL	Recommendation –RL	Recommendation –RL (distributed training)

Thank You for your attention!

Any Questions ?

Conclusion on Big Data and AI Benchmarks

Todor Ivanov (LeadConsult)

DataBench

DataBench Pipeline Methodology

Data Acquisition/Collection
(including Data Ingestion processing, Streaming, Data Extraction, Ingestion Storage, Different data types)

Data Storage/Preparation
(including Storage Retrieval/Access/Queries Data Protection, Curation, Integration, Publication)

Data Analytics/ML
(including data processing for analysis, AI and Machine Learning)

Data Visualisation, Action/Interaction
(including data presentation environment/boundary/user action and interaction)

BDVA Reference Model

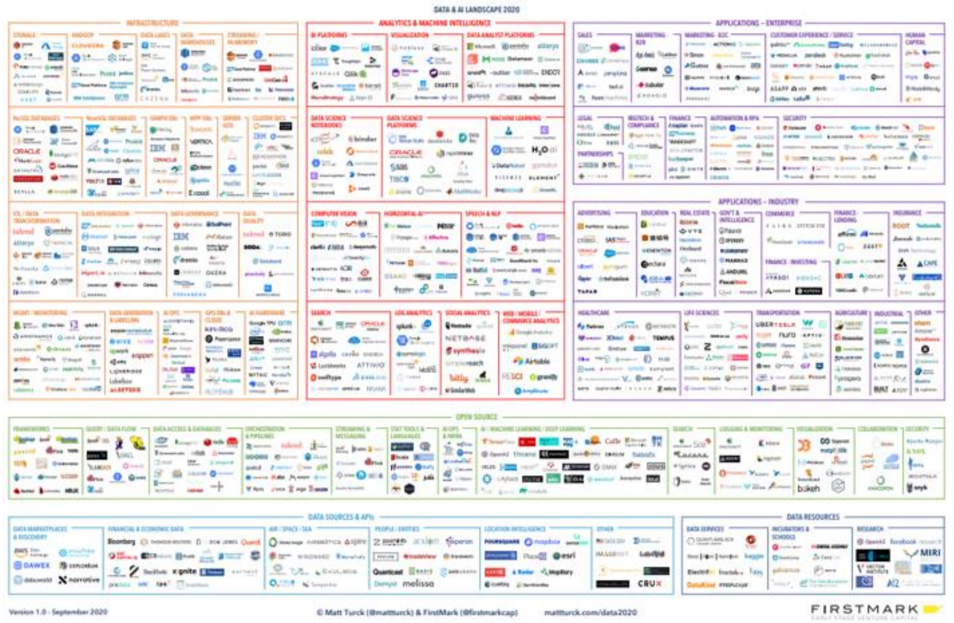
- Streaming/ Realtime Processing
- Interactive Processing
- Batch Processing

- Data Privacy/Security
- Data Governance/Mgmt
- Data Storage

- Industrial Analytics (Descriptive, Diagnostic, Predictive, Prescriptive)
- Machine Learning, AI, Data Science

- Visual Analytics

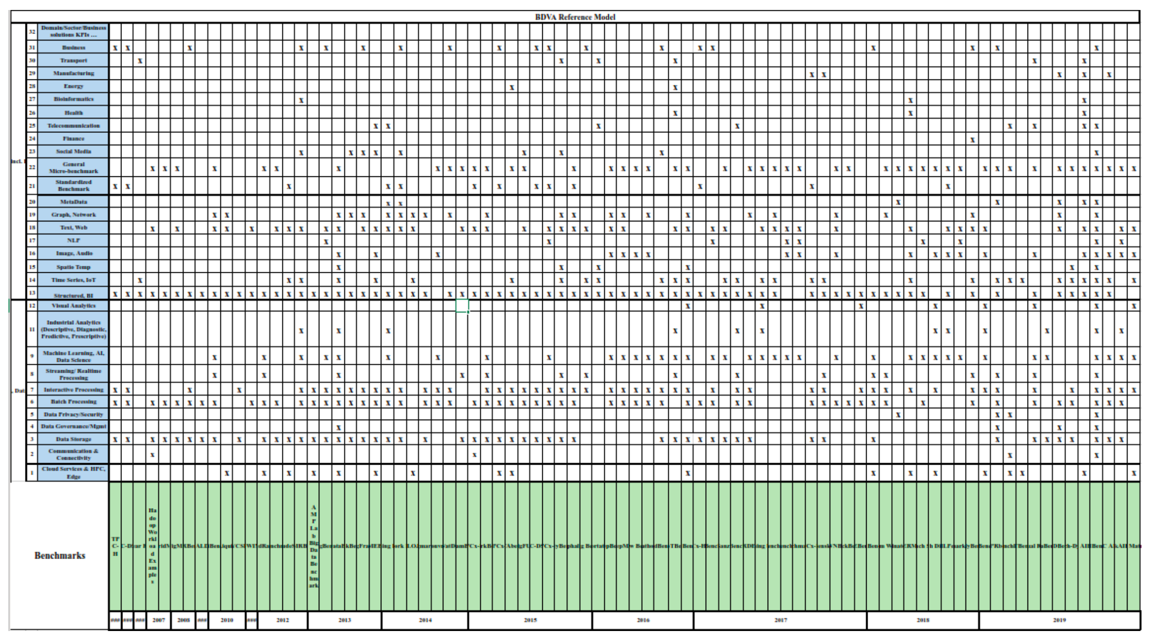
Big Data and AI Landscape



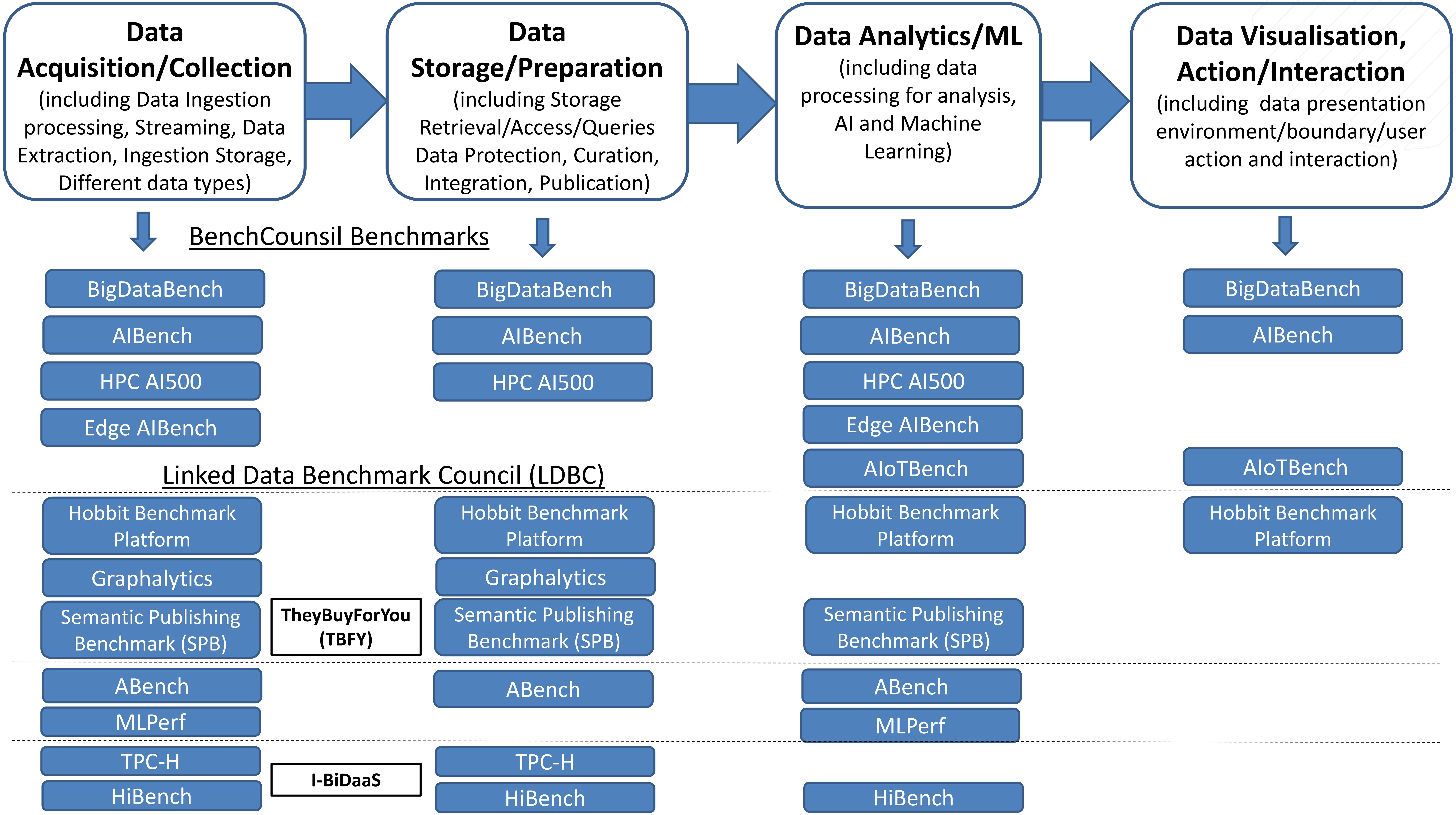
ICT Big Data PPPICT-13-14 Projects

- **I-BiDaaS**
- **TheyBuyForYou (TBFY)**
- **Track&Know**
- **DataBio**
- **DeepHealth**

Big Data and AI Benchmark Ecosystem

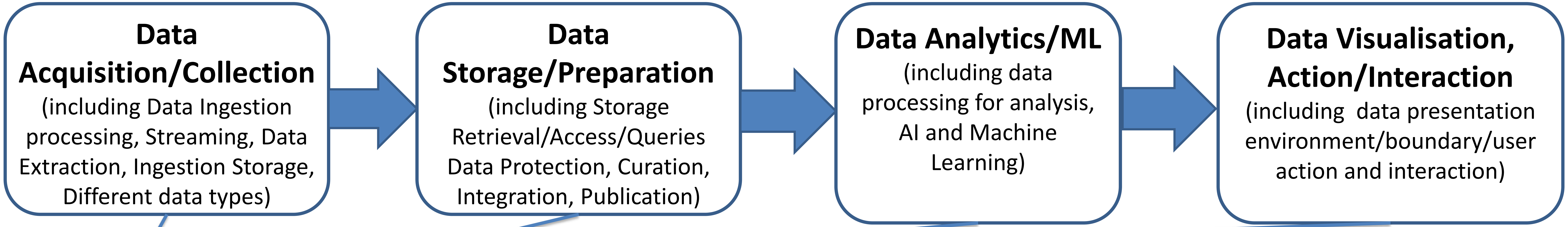


DataBench Pipeline Methodology



Mapping between the ***DataBench Pipeline Steps*** and the ***Benchmark Ecosystem***
→ matrix available in the **DataBench ToolBox**

DataBench Pipeline Methodology



Benchmarks	Year	PI-DV	PI-DA	PI-DS	PI-DI
SparkBench	2015		X	X	X
TPCx-V	2015			X	X
IoTAbench	2015			X	X
BigFUN	2015			X	X
TPC-DS v2	2015			X	X
TPCx-BB	2015		X	X	X
CityBench	2015			X	X
Graphalytics	2015			X	X
Yahoo Streaming Benchmark (YSB)	2015				X
ShenZhen Transportation System (SZTS)	2015				
DeepBench	2016		X		X
DeepMark	2016		X		X
TensorFlow Benchmarks	2016		X		X
Fathom	2016		X		X
AdBench	2016		X	X	X
RIoTBench	2016		X	X	X
Hobbit Benchmark	2016	X	X	X	X
TPCx-HS v2	2016			X	X
BigBench V2	2017		X	X	X
Sanzu	2017		X	X	
AIM Benchmark	2017		X	X	X
GARDENIA	2017		X	X	X
Penn machine learning benchmark (PMLB)	2017	X	X		
OpenML benchmark suites	2017		X		
Benchip	2017		X		
Deep Learning Benchmarking Suite (DLBS)	2017		X		
TPCx-IoT	2017			X	X
Senska	2017			X	X
DAWNBench	2017		X		X
BlockBench	2017				X
IDEBench	2017	X			X
ABench	2017		X	X	X
Stream WatDiv	2018				X
TERMinator Suite	2018			X	
HERMIT	2018		X		X
MLBench Services	2018		X		X
MLBench Distributed	2018	X	X		X
MLPerf	2018		X		
Training Benchmark for DNNs (TBD)	2018		X		
PolyBench	2018				X
NNBench-X	2018	X	X		X
GDPRbench	2019			X	X
BenchIoT	2019			X	
IoTBench	2019				
Visual Road	2019	X	X	X	X
AdaBench	2019		X	X	
MiDBench	2019			X	X
CBench-Dynamo	2019			X	X
Edge AIBench	2019				
AIBench	2019	X	X	X	X
HPC AI500	2019		X	X	X
SparkAIBench	2019		X	X	X
AI Matrix	2019		X	X	X

Contacts

www.databench.eu

info@databench.eu

@DataBench_eu

DataBench

DataBench Project

DataBench

DataBench Project

EUROPEAN BIG DATA VALUE FORUM

BERLIN + VIRTUAL
3-5 NOVEMBER 2020

DataBench

This project has received funding from the European Horizon 2020 Programme for research, technological development and demonstration under grant agreement n° 780966