# Big data science Day 1

F. Legger - INFN Torino

https://github.com/Course-bigDataAndML/MLCourse-2425

- So long, and thanks for all the images!
  - *Taken freely from the web*
  - *Credits go to original creators*

# Schedule

- **Nov 24-28, 14:00-16:00, Aula Castagnoli**
  - Lectures + hands-on sessions
- The **exam** will be a practical test (exercises on a Jupyter notebook) and can be turned in till **Jan 6th**

# Objectives

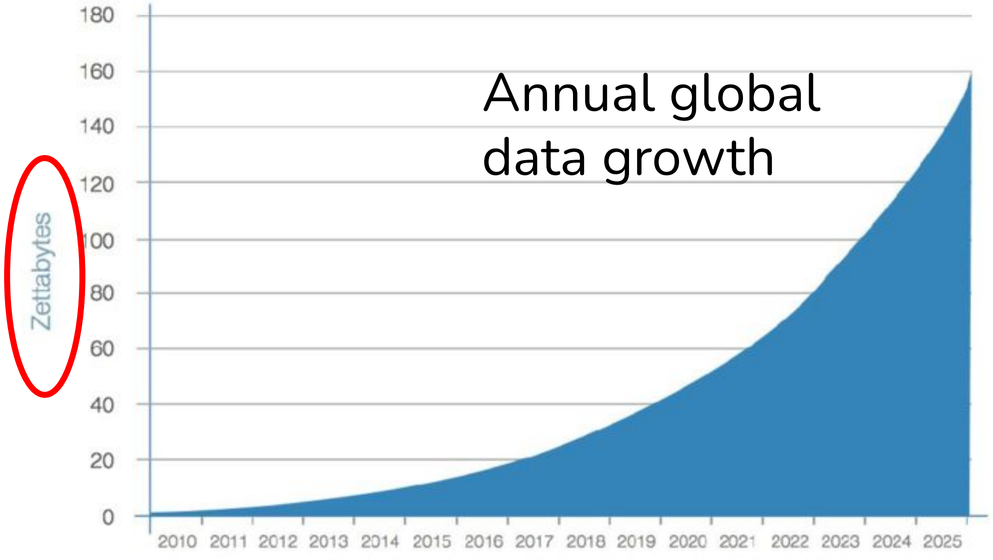
- **Machine Learning 101**
  - We will only introduce transformers
- **There is no ML without distributed computing**
  - We will start with big Data concepts

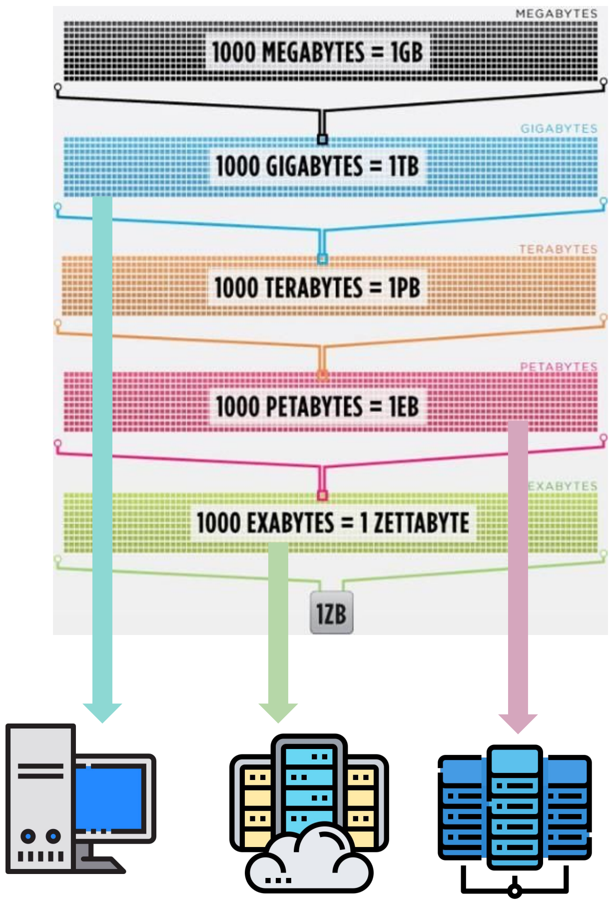
# Today

- **Introduction to big data**
  - Definition, applications, sources
- **The big data pipeline**
  - Infrastructure, technologies
- **Analytics**
  - Data mining, data structures, data processing

# What is big data?

- Data that is too big to be analysed traditionally

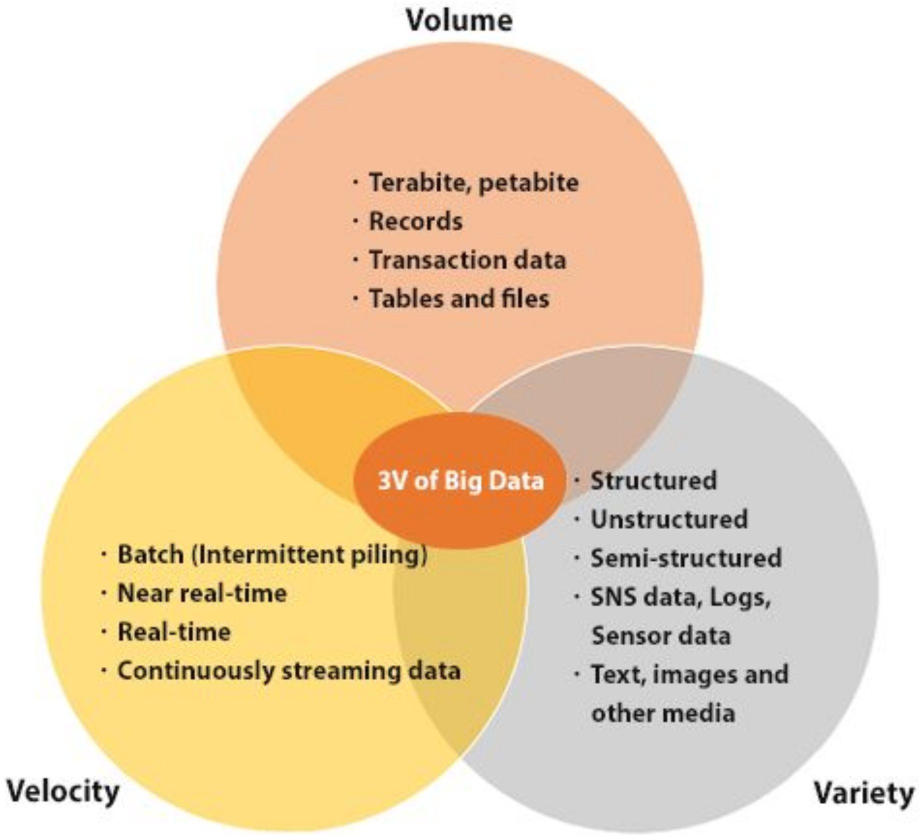



https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

# It's all about Vs

*"Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."* Gartner (2012)



# Even more Vs

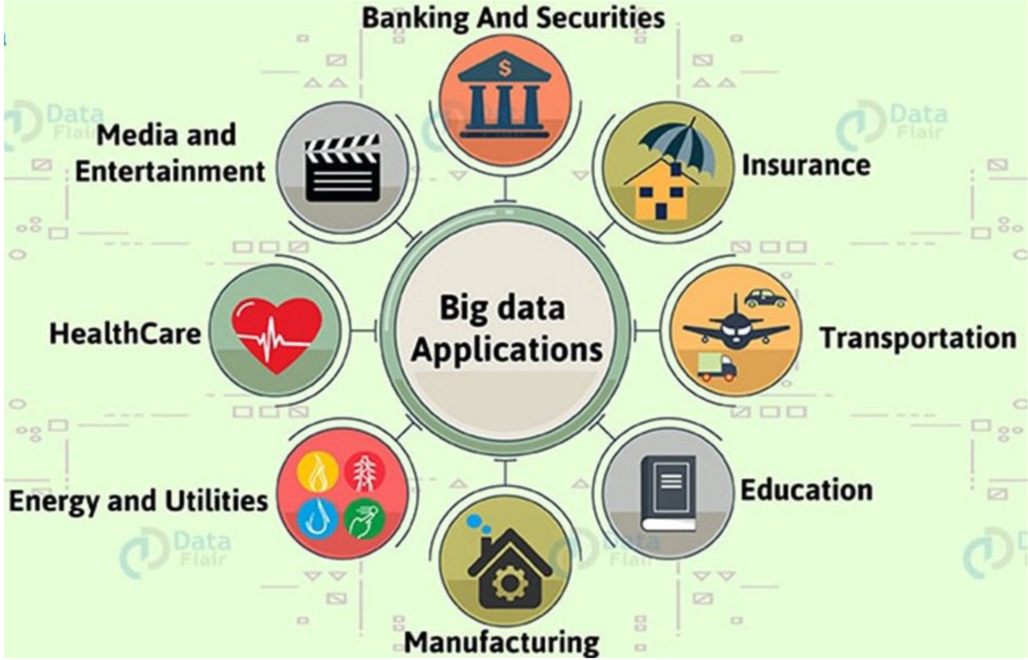
- Veracity
- Variability
- Visualization
- Validity
- Vulnerability
- Volatility
- Value

Complex and heterogeneous

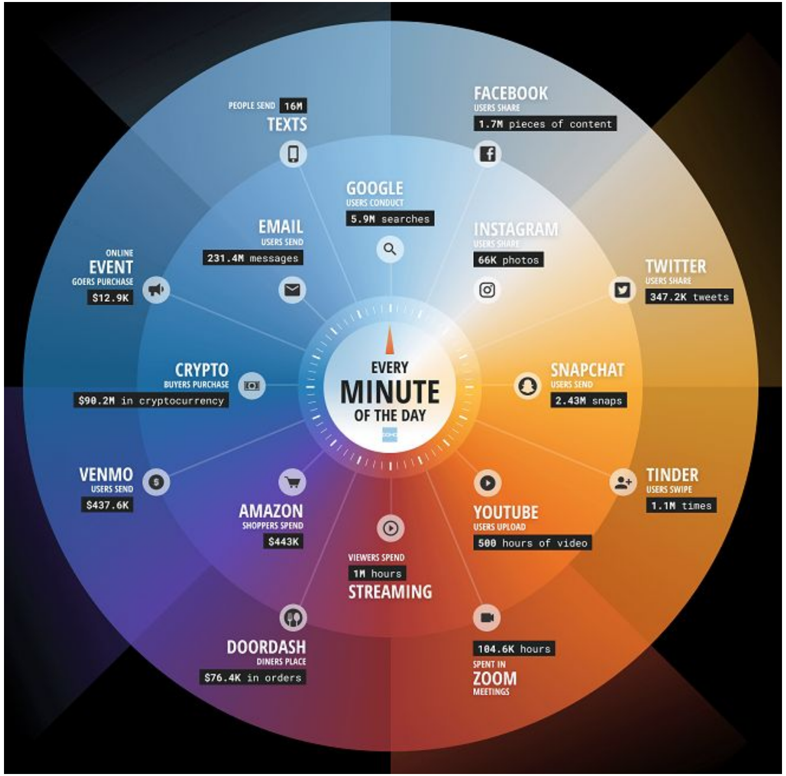
https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx

# Big data sources: Business (x1)

- **Traditional Business Systems**
  - Commercial transactions
  - Banking/stock records
  - E-commerce
  - Credit cards
  - Medical records



# Big data sources: Human (x10)




- **Social Networks**
  - X, Facebook, TikTok
  - Blogs and comments
- **Video conferences**
  - Zoom, Teams
- **Streaming:**
  - YouTube, Netflix
- Internet searches
- User-generated maps
- E-Mail

# Big data sources: Machine (x100)

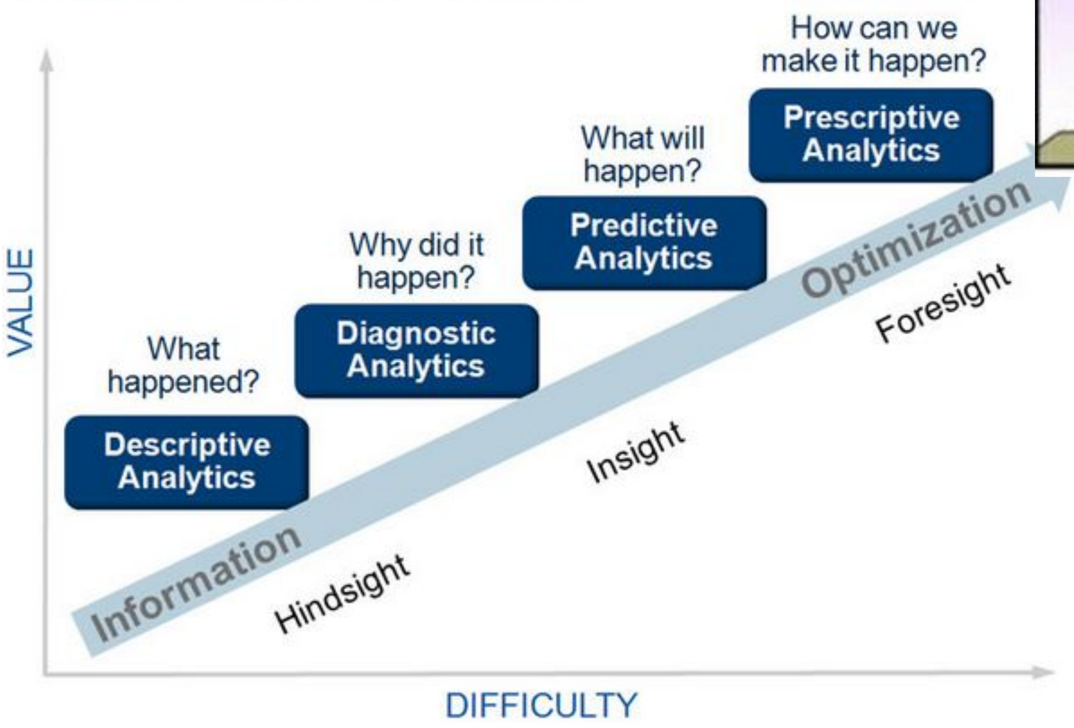
**Internet of Things (IoT)**

- Sensors: traffic, weather, locations
- Security, surveillance videos, images
- Satellite images
- Data from computer systems (logs, web)

INTERNET OF THINGS

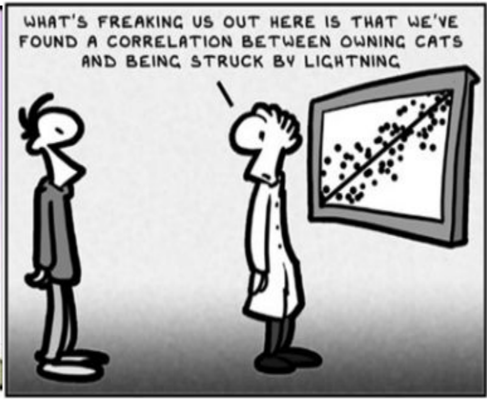
**And AI?** Current estimations: 50-90% AI-generated content

# Analytics

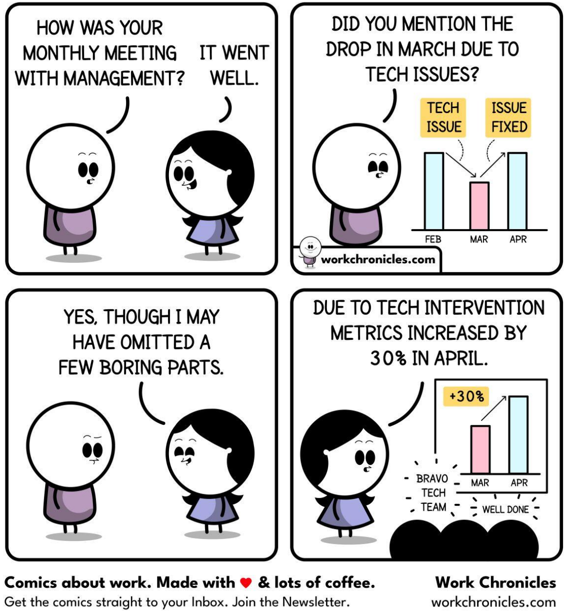


VALUE
DIFFICULTY
Information
Insight
Optimization
Hindsight
Foresight
What happened?
Descriptive Analytics
Why did it happen?
Diagnostic Analytics
What will happen?
Predictive Analytics
How can we make it happen?
Prescriptive Analytics


SHOULD WE ALWAYS IGNORE WHAT THE DATA SAYS, OR IS THIS MORE OF A ONE-TIME THING?

WHAT'S FREAKING US OUT HERE IS THAT WE'VE FOUND A CORRELATION BETWEEN OWNING CATS AND BEING STRUCK BY LIGHTNING

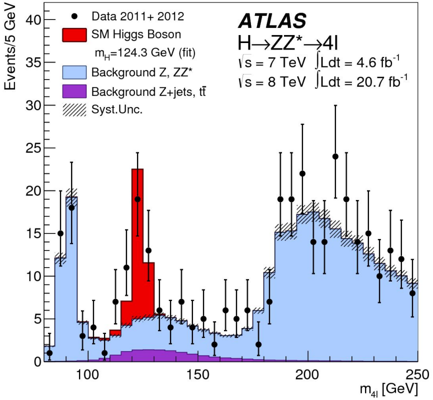



HOW WAS YOUR MONTHLY MEETING WITH MANAGEMENT?
IT WENT WELL.
DID YOU MENTION THE DROP IN MARCH DUE TO TECH ISSUES?
TECH ISSUE
ISSUE FIXED
FEB
MAR
APR
workchronicles.com
YES, THOUGH I MAY HAVE OMITTED A FEW BORING PARTS.
DUE TO TECH INTERVENTION METRICS INCREASED BY 30% IN APRIL.
+30%
MAR
APR
BRAVO TECH TEAM
WELL DONE
Comics about work. Made with ♥ & lots of coffee.
Get the comics straight to your Inbox. Join the Newsletter.
Work Chronicles
workchronicles.com


LHCb
CERN Prévessin
SUISSE
FRANCE
CMS
LHC 27 km


Once upon A Time...


- pp (or Pb-Pb) collisions
- 4 experiments (ATLAS, CMS, LHCb, ALICE)
- Discovery of Higgs boson
- Nobel prize for physics 2013



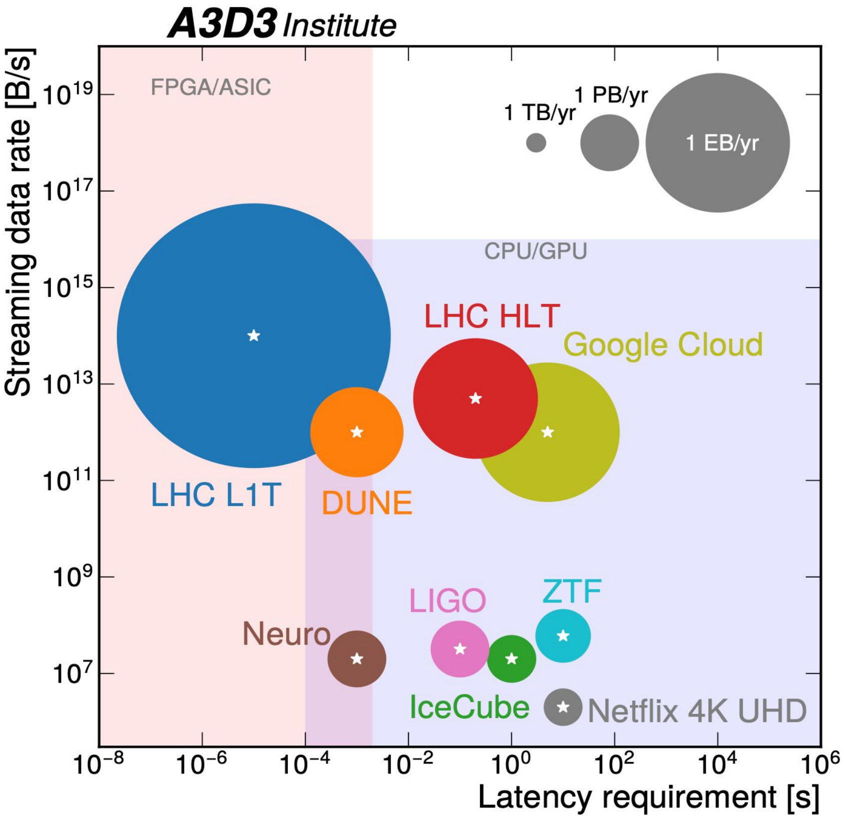
# Big data @LHC

- ATLAS/CMS: 100-megapixel digital cameras that take 40 million "pictures" per second
    - collisions/s = $40 \times 10^6 \times 20 = 0.8 \times 10^9$

$10^9$ (LHC events/s) x 1 MB/event = 1 PB / second

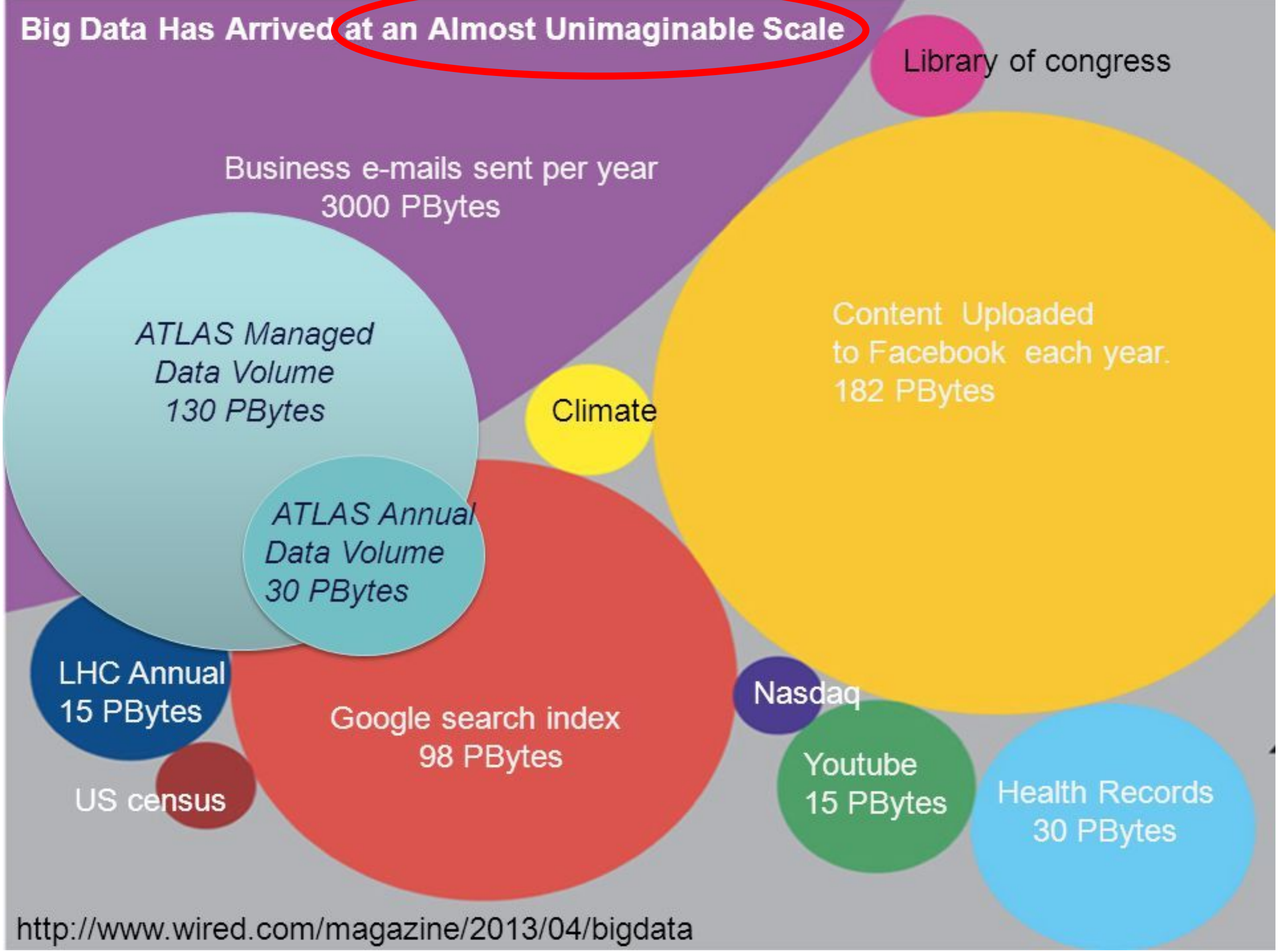
- 
    - We cannot store this!
    - Only one in a billion is an Higgs boson!

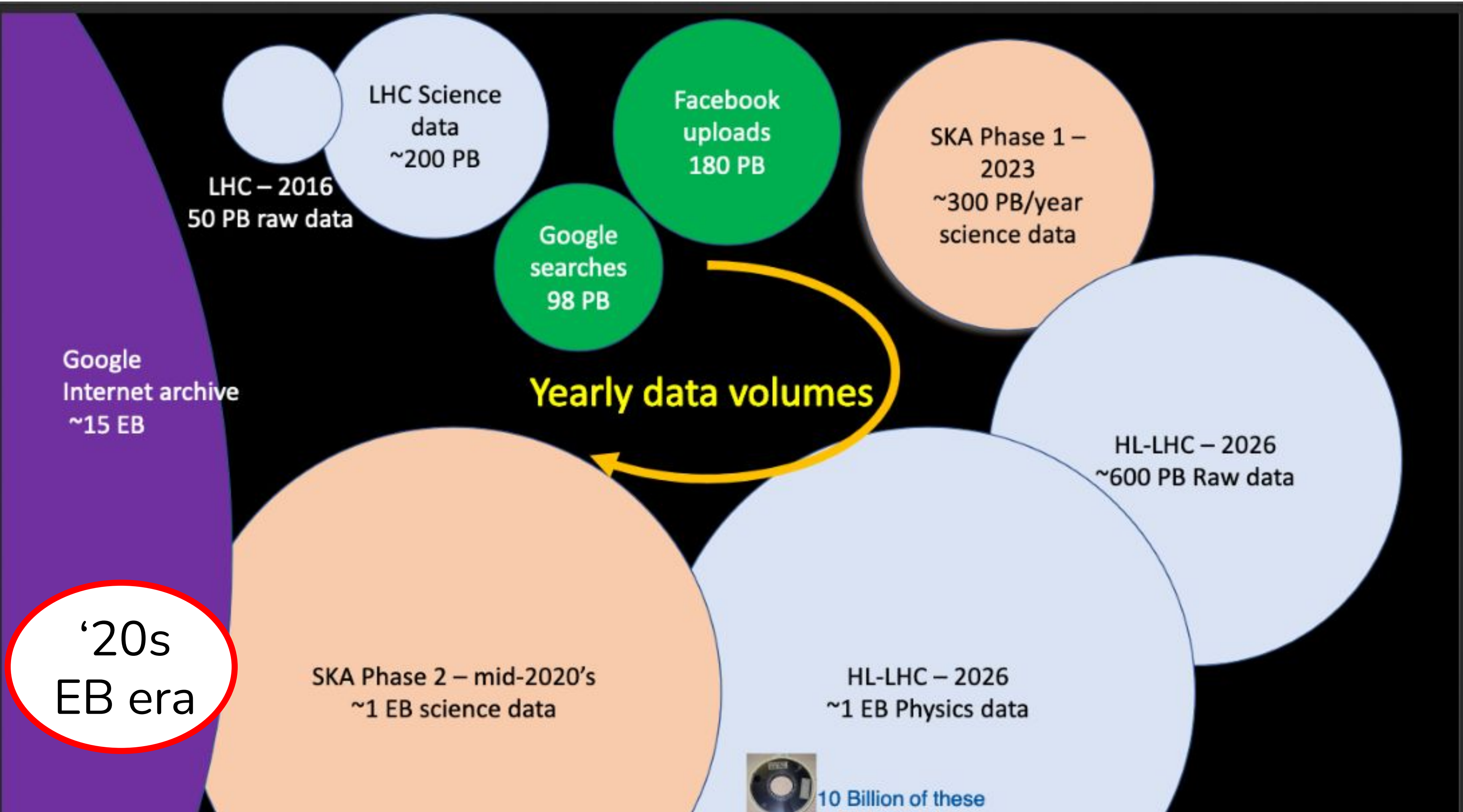
- Trigger system (real time): “empty” pictures immediately thrown away -> 1GB/second
- Data stored, processed and analysed using a distributed computing infrastructure, known as the **Worldwide LHC Computing Grid (WLCG)**




$10^3$ (events/s) x $10^7$ (s/year) x 1 MB/event = 10 PB / year

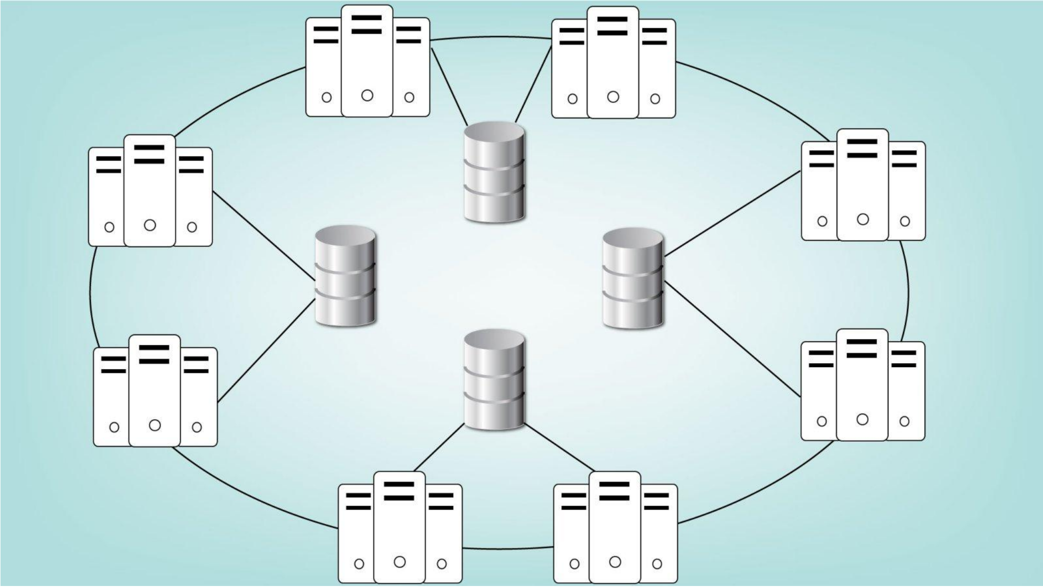
2013,
PB era


Big Data Has Arrived at an Almost Unimaginable Scale
Library of congress
Business e-mails sent per year
3000 PBytes
ATLAS Managed
Data Volume
130 PBytes
Climate
Content Uploaded
to Facebook each year.
182 PBytes
ATLAS Annual
Data Volume
30 PBytes
LHC Annual
15 PBytes
Google search index
98 PBytes
Nasdaq
Youtube
15 PBytes
Health Records
30 PBytes
US census
http://www.wired.com/magazine/2013/04/bigdata

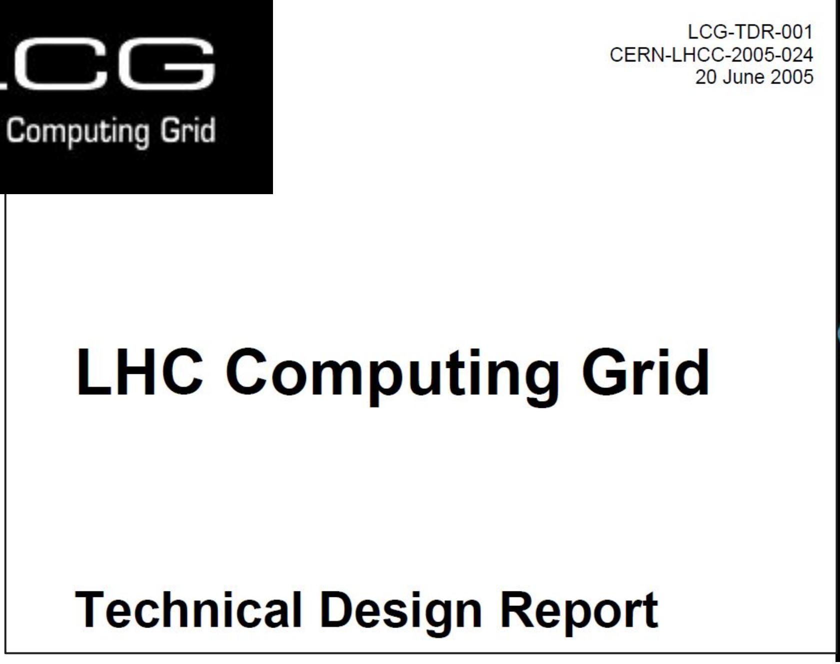

LHC – 2016
50 PB raw data
LHC Science data
~200 PB
Facebook uploads
180 PB
Google searches
98 PB
SKA Phase 1 – 2023
~300 PB/year science data
Google
Internet archive
~15 EB
Yearly data volumes
HL-LHC – 2026
~600 PB Raw data
'20s
EB era
SKA Phase 2 – mid-2020's
~1 EB science data
HL-LHC – 2026
~1 EB Physics data
10 Billion of these

# Distributed Computing

- distributed storage
- distributed compute nodes
- connected over fast network
- some way to distribute computing workloads according to available data/computing power

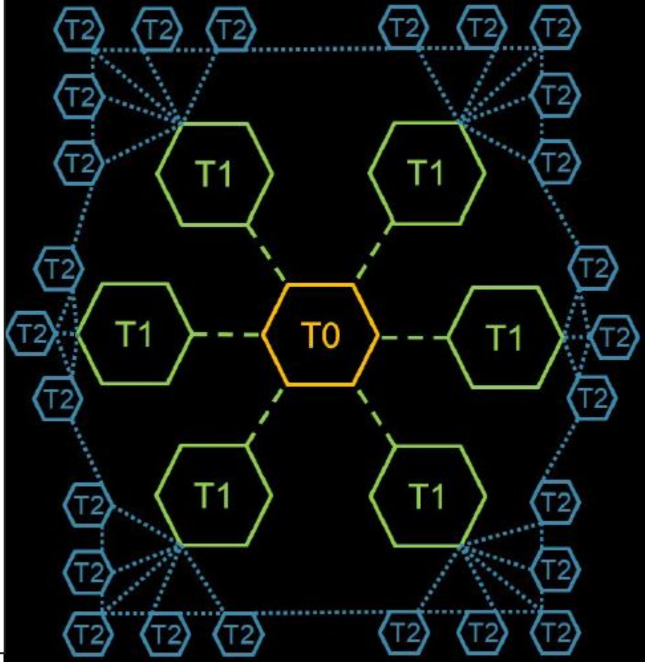


WLCG
Worldwide LHC Computing Grid


LCG-TDR-001
CERN-LHCC-2005-024
20 June 2005

**LHC Computing Grid**

**Technical Design Report**




Work started in early 2000's

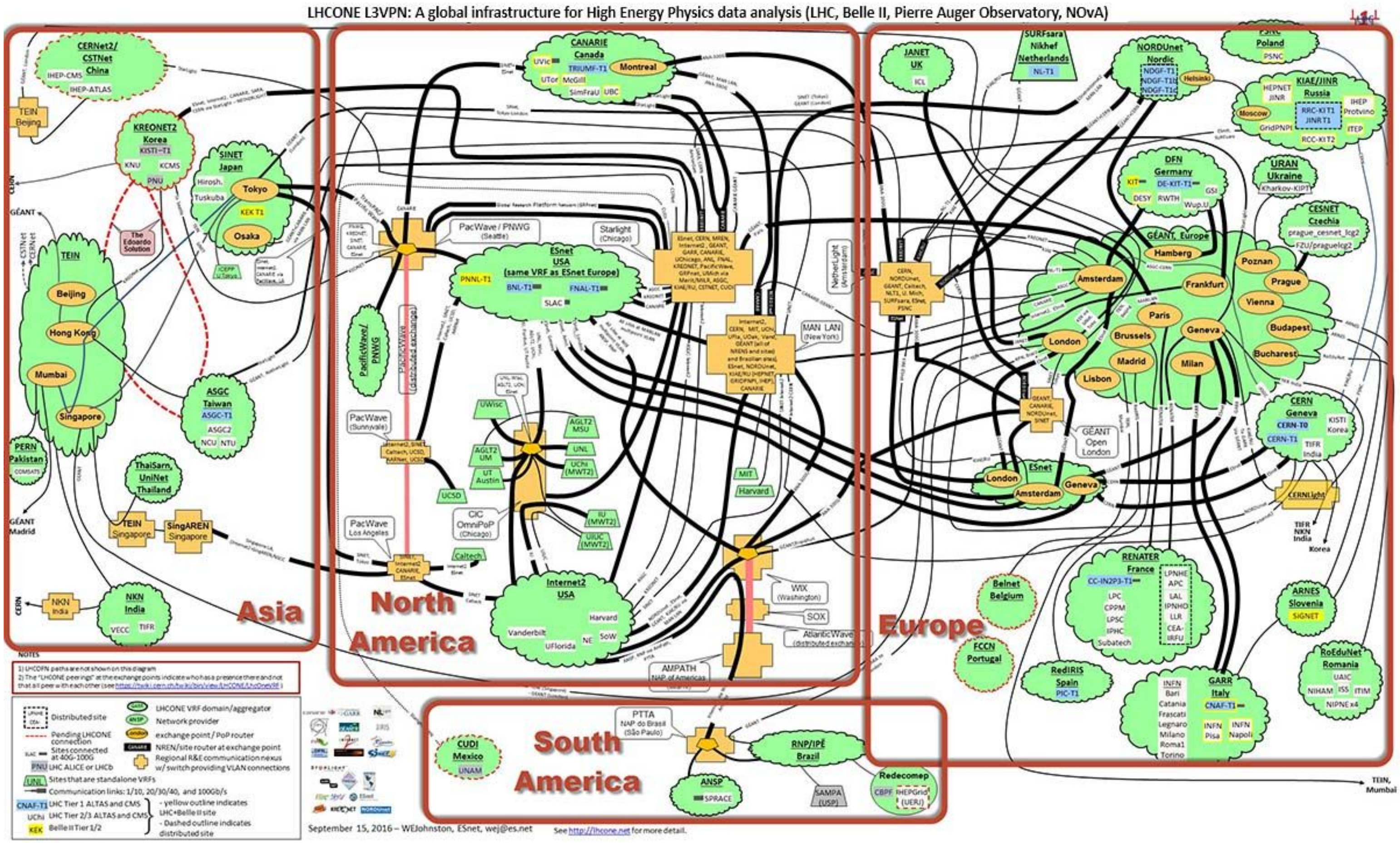
- Tiered architecture:
  - **Tier 0 - Tier 1**
  - **Tier 2 - Tier 31**

- **Workload Management**
- **Data Management**
- **Sites and facilities**

# **W**orld
# **L**HC
# **C**omputing
# **G**rid

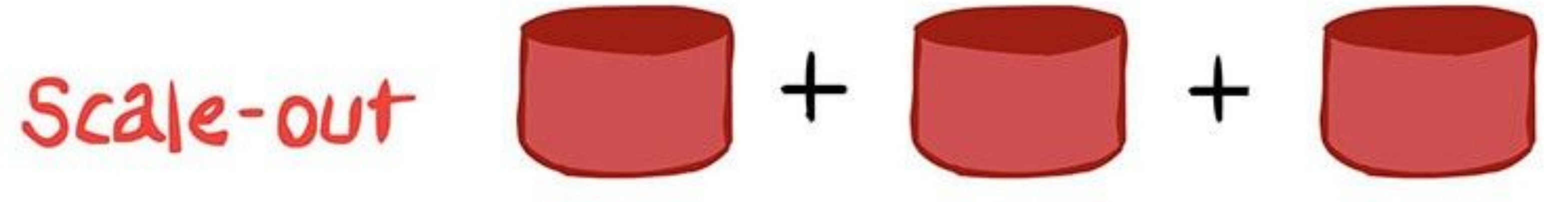

- 42 countries
- 170 computing centres
- Over 2 million tasks daily
- 1 million computer cores
- 2 exabytes of storage: 1.2 Tape + 0.8 Disk

- **LHCOPN** >1 Tbps from Tier 0 to 14 Tier 1s
- **LHCOne** overlay of 10-100 Gbps networks to connect Tier 1,2,3

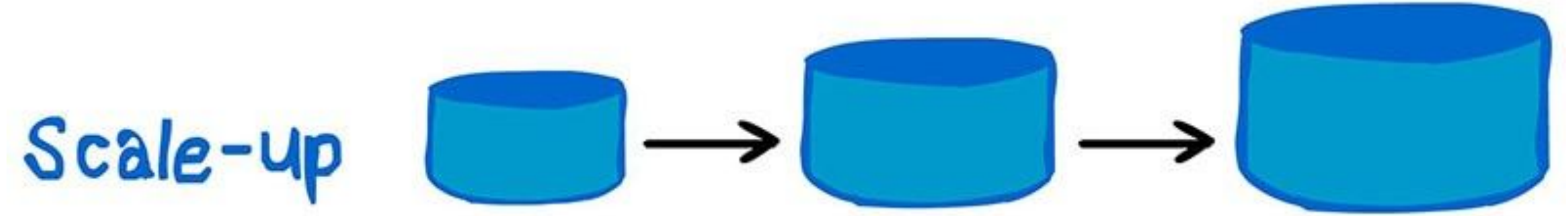


# Scale up vs scale out

Grid, cloud

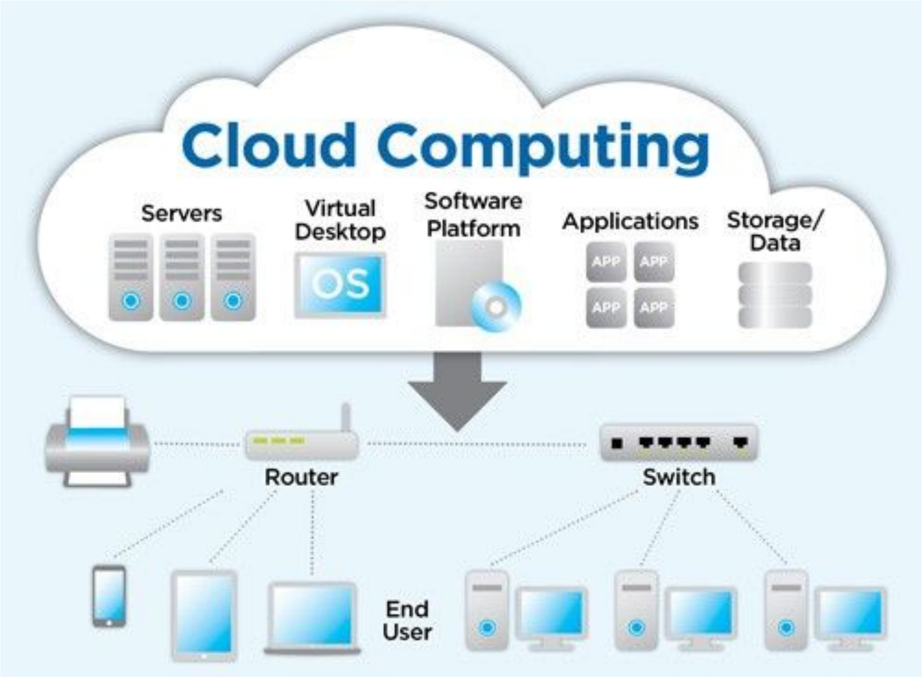



HPC, super computers



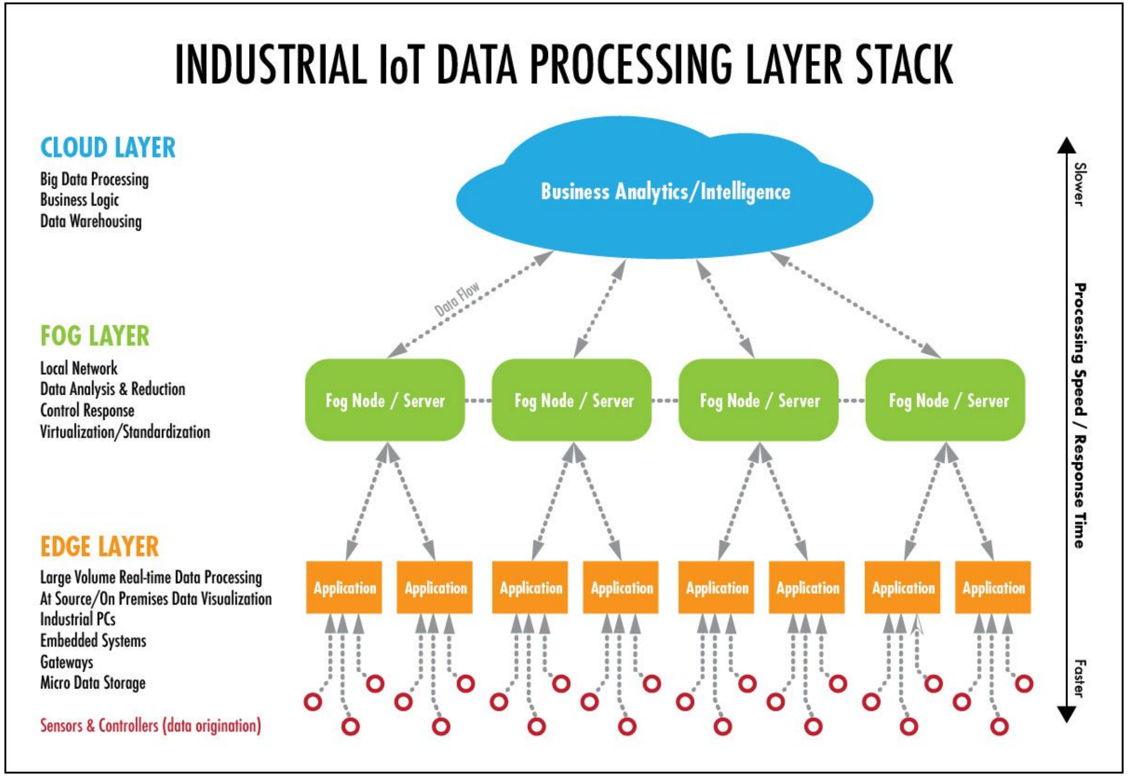
# Cloud computing

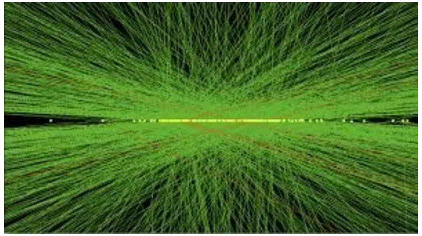
*On-demand availability of resources, especially data storage and computing power, without direct active management by the user*




1. Resources Pooling
2. On-Demand Self-Service
3. Easy Maintenance
4. Large Network Access
5. Availability
6. Automatic System
7. Economical
8. Security
9. Pay as you go (commercial)
10. Measured Service

# Edge and fog computing



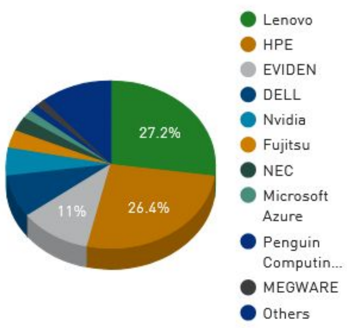
# High Performance Computing (HPC)

- Top 500, 3 exascale machines:
  - **El Capitan**, #1 Exascale machine (Nov, 2024)
  - Argonne **Aurora** becomes the second exascale machine (May, 2024)
  - Oak Ridge’s **Frontier** first to break the Exaflop Ceiling (May, 2022)

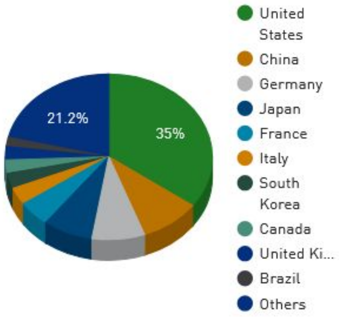
Top European HPC

1. **El Capitan** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE
2. **Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE
3. **Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel
4. **JUPITER Booster** - BullSequana XH3000, GH Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, RedHat Enterprise Linux, EVIDEN
5. **Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure
6. **HPC6** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, RHEL 8.9, HPE
7. **Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu
8. **Alps** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Cray OS, HPE
9. **LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE
10. **Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, EVIDEN

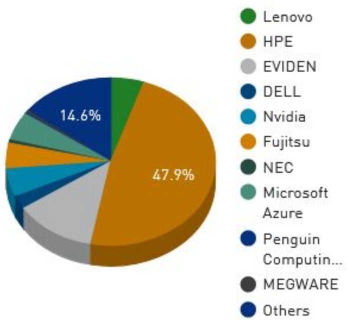
**Vendors System Share**



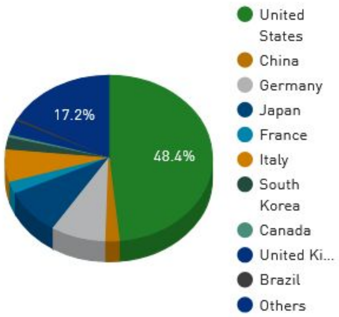

**Countries System Share**




**Vendors Performance Share**




**Countries Performance Share**




https://www.top500.org/

# Graphical Processing Units (GPU)

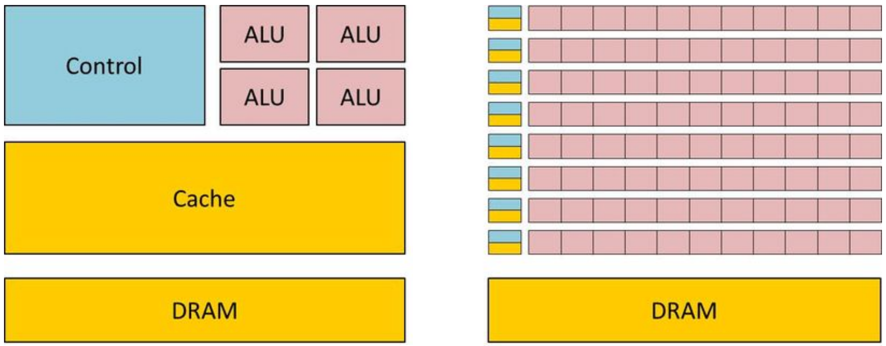
## CPU vs GPU

**CPU**

- few very complex cores
- single-thread performance optimization
- transistor space dedicated to complex ILP
- few die surface for integer and fp units

**GPU**

- hundreds of simpler cores
- thousand of concurrent hardware threads
- maximize floating-point throughput
- most die surface for integer and fp units




- Performances
  - Excels at matrix and vector operations
  - Gaming, rendering… and ML
- Power consumption (more FLOPS/Watt)

# Heterogeneous architectures

**CPU**
- Small models
- Small datasets
- Useful for design space exploration

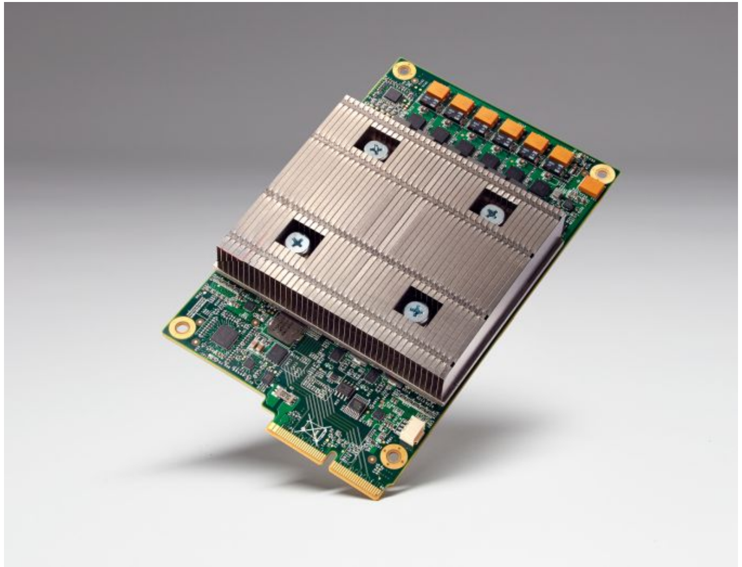
**GPU**
- Medium-to-large models, datasets
- Image, video processing
- Application on CUDA or OpenCL

**TPU**
- Matrix computations
- Dense vector processing
- No custom TensorFlow operations

**FPGA**
- Large datasets, models
- Compute intensive applications
- High performance, high perf./cost ratio

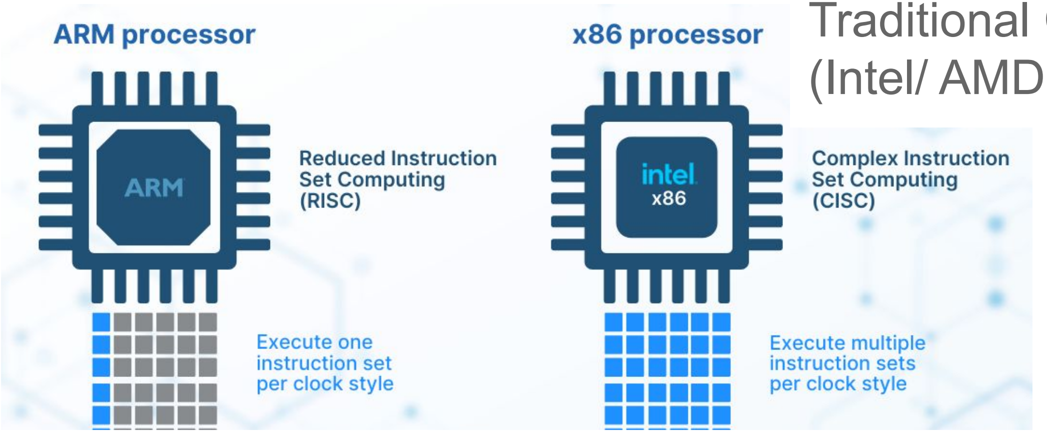
TPU: Tensor(flow) Processing Unit

FPGA: Field Programmable Gate Array

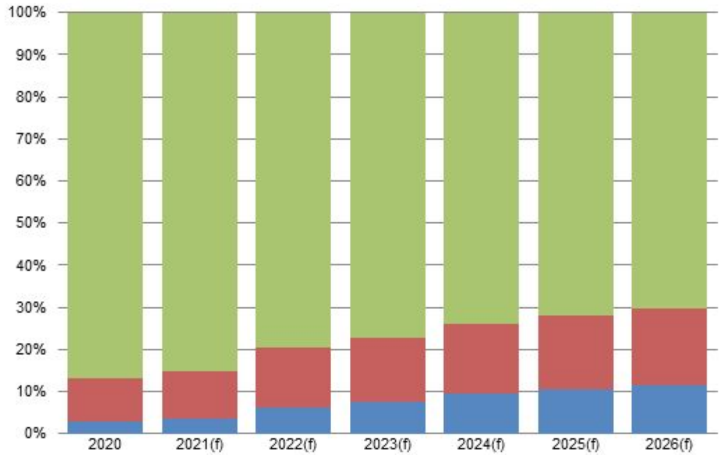


# ARM (Advanced RISC Machine)

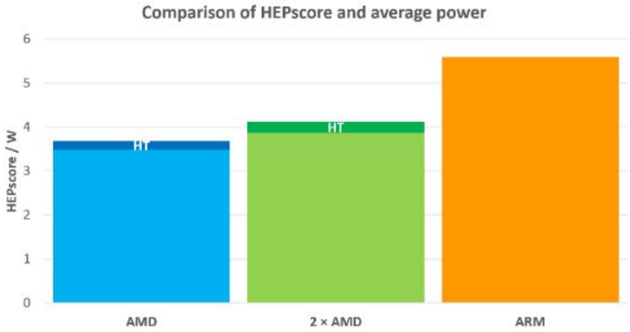
Traditional CPU
(Intel/ AMD)




Market share

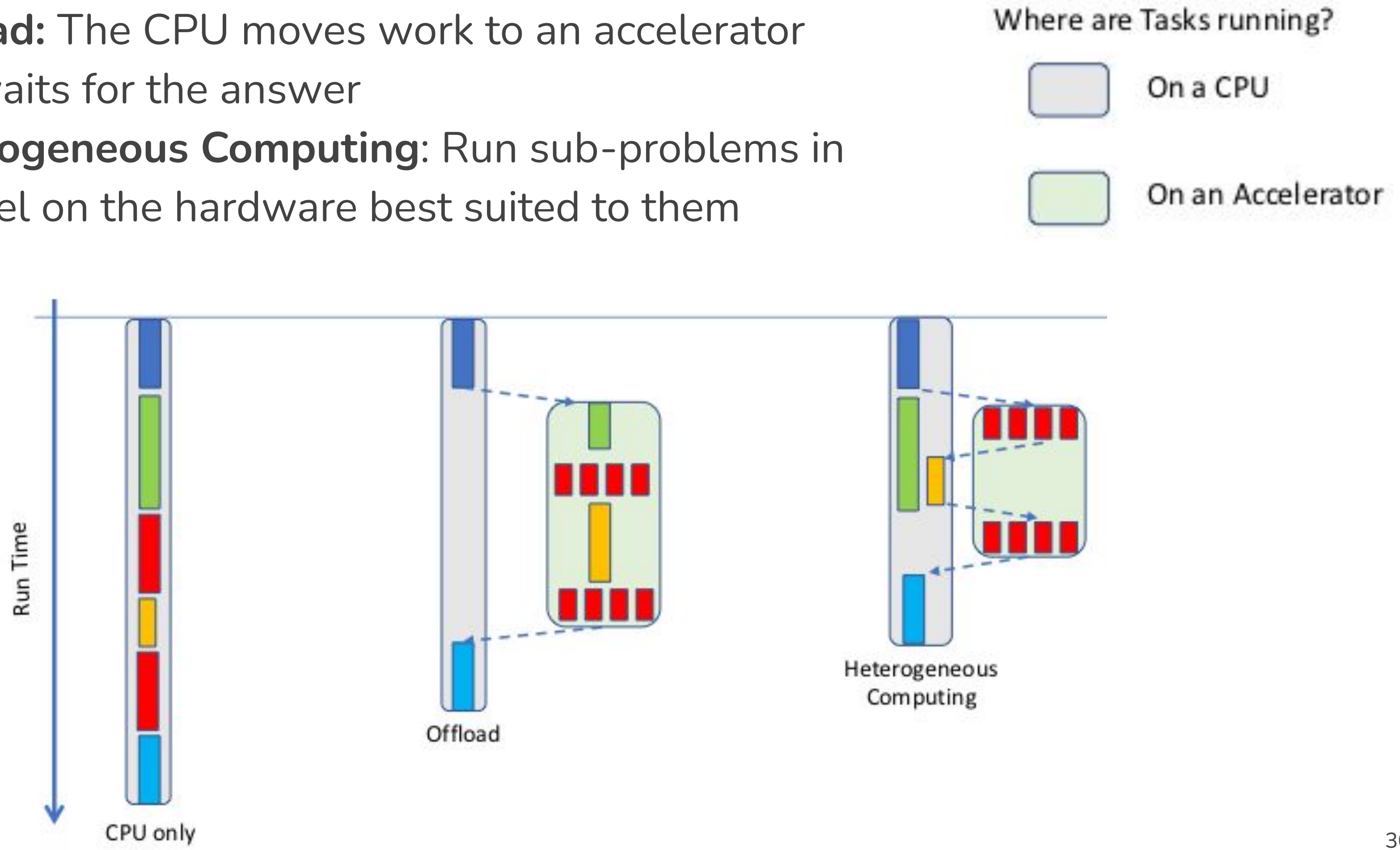



| | 2020 | 2021(f) | 2022(f) | 2023(f) | 2024(f) | 2025(f) | 2026(f) |
|---|---|---|---|---|---|---|---|
| Intel | 86.9% | 85.3% | 79.5% | 77.4% | 74.1% | 72.0% | 70.3% |
| AMD | 10.1% | 11.2% | 14.3% | 15.2% | 16.3% | 17.4% | 18.2% |
| Arm and others | 3.0% | 3.5% | 6.2% | 7.4% | 9.6% | 10.6% | 11.5% |

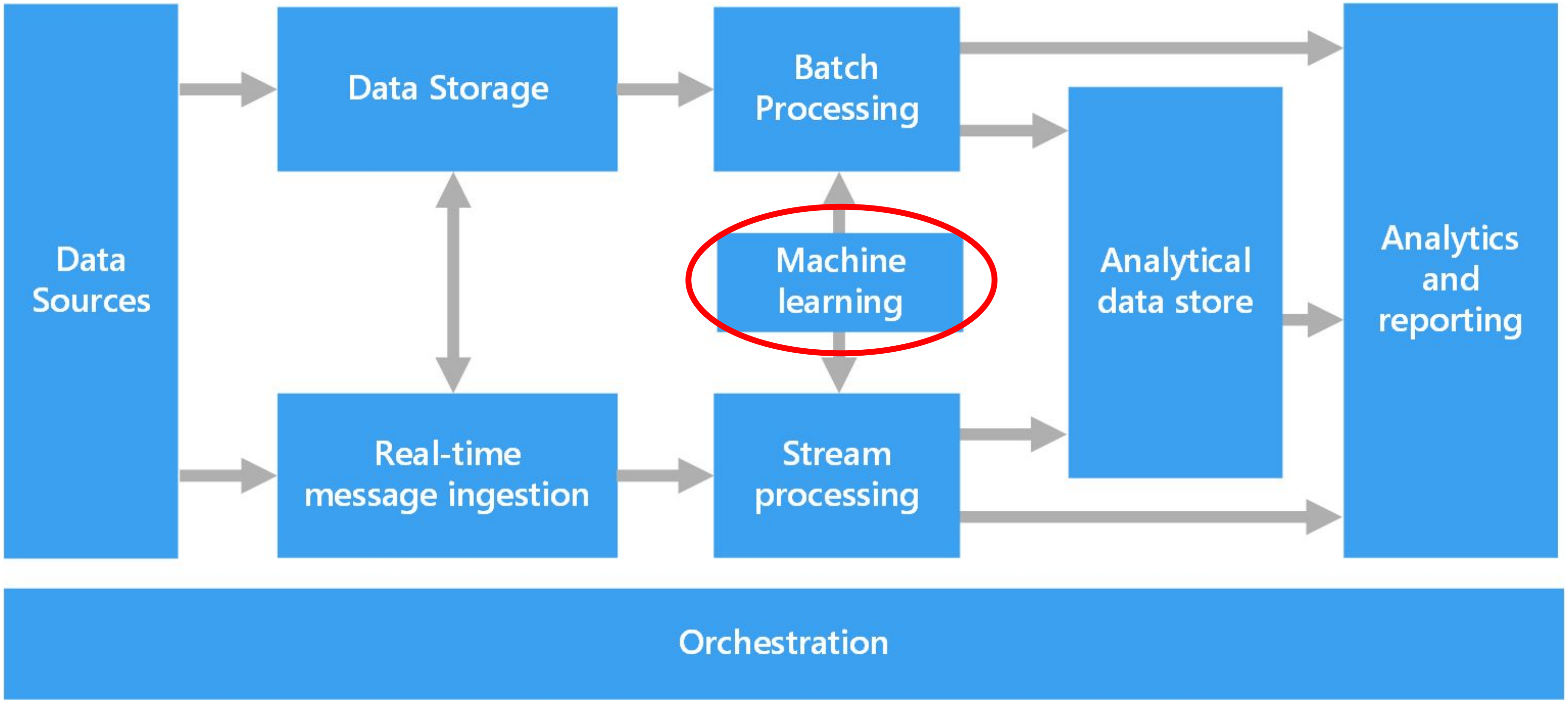



HEPScore
benchmark

- **Offload:** The CPU moves work to an accelerator and waits for the answer
- **Heterogeneous Computing**: Run sub-problems in parallel on the hardware best suited to them

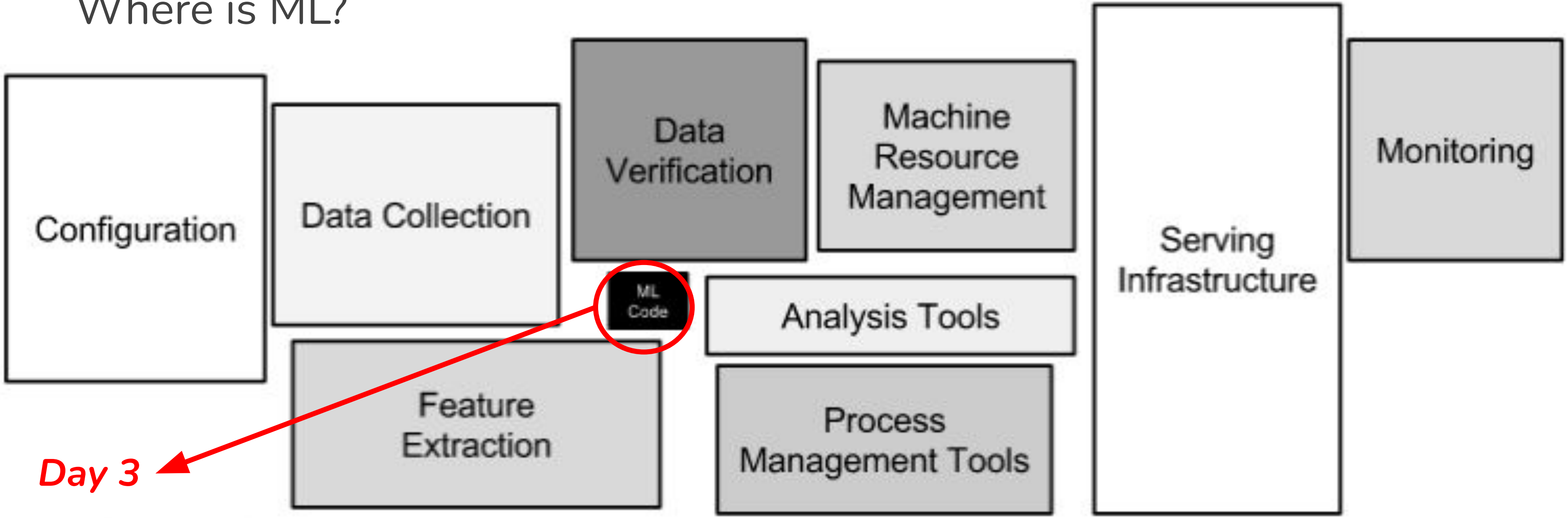


# The big data pipeline

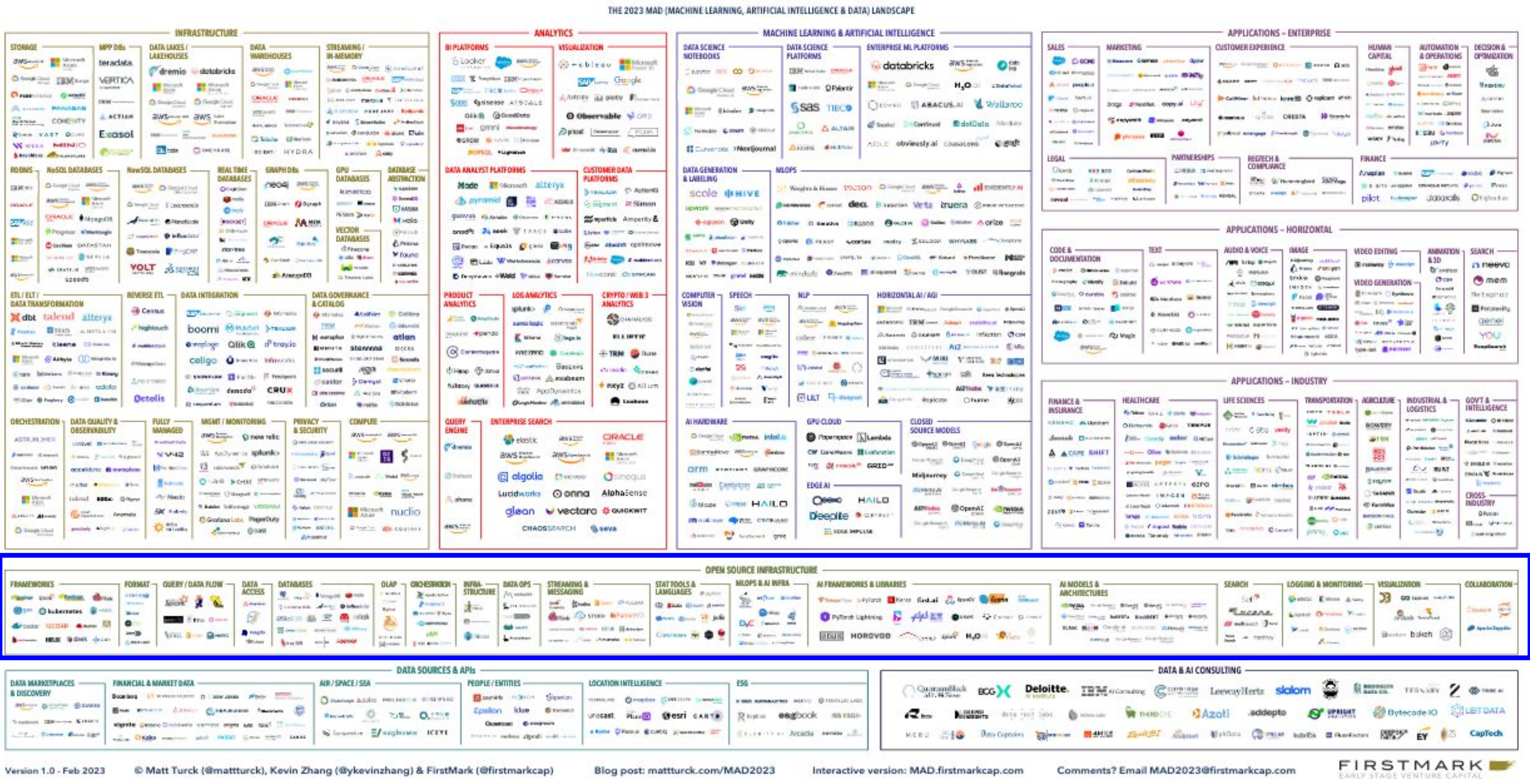


# Typical ML workflow

Where is ML?

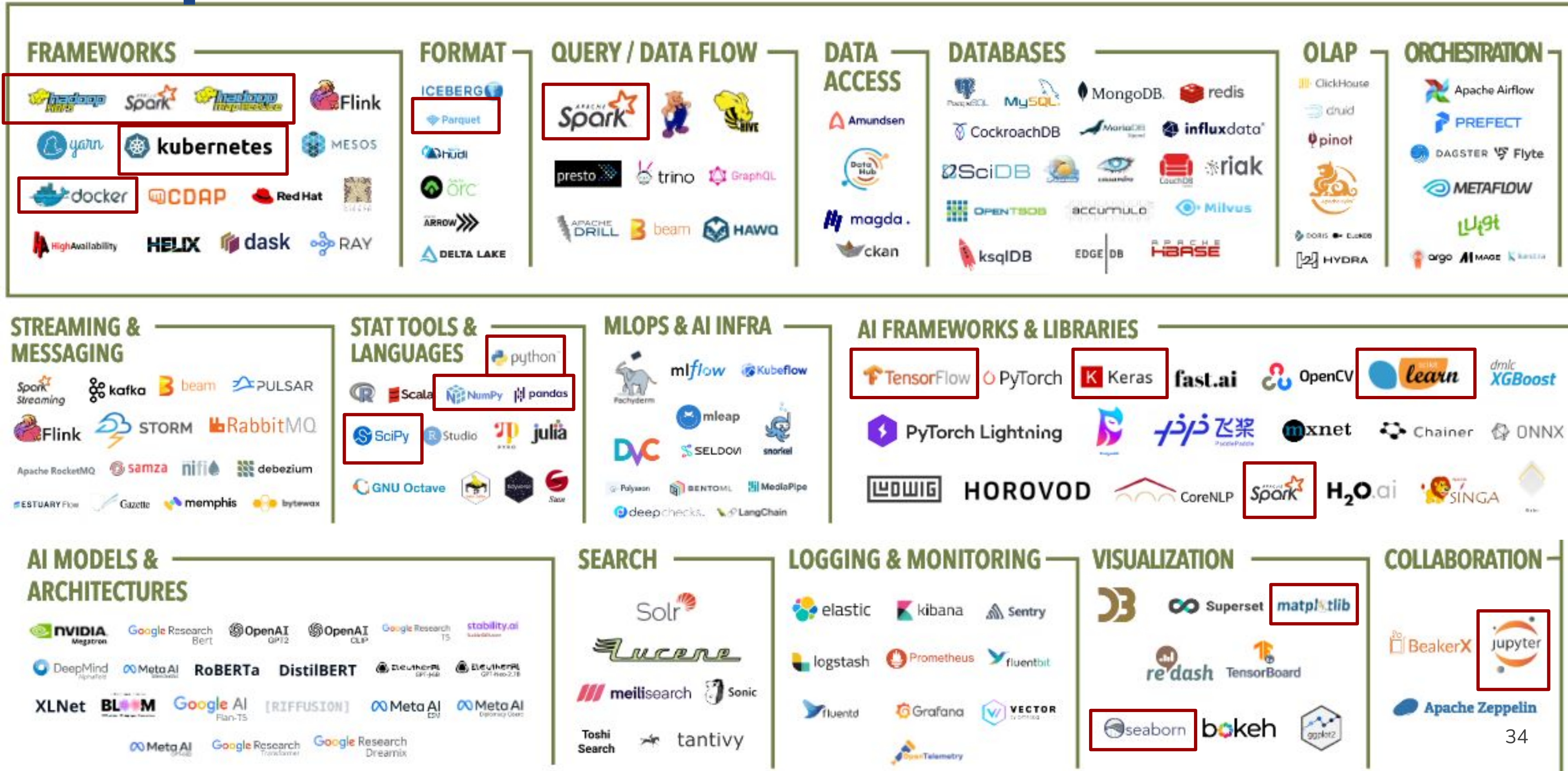



https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

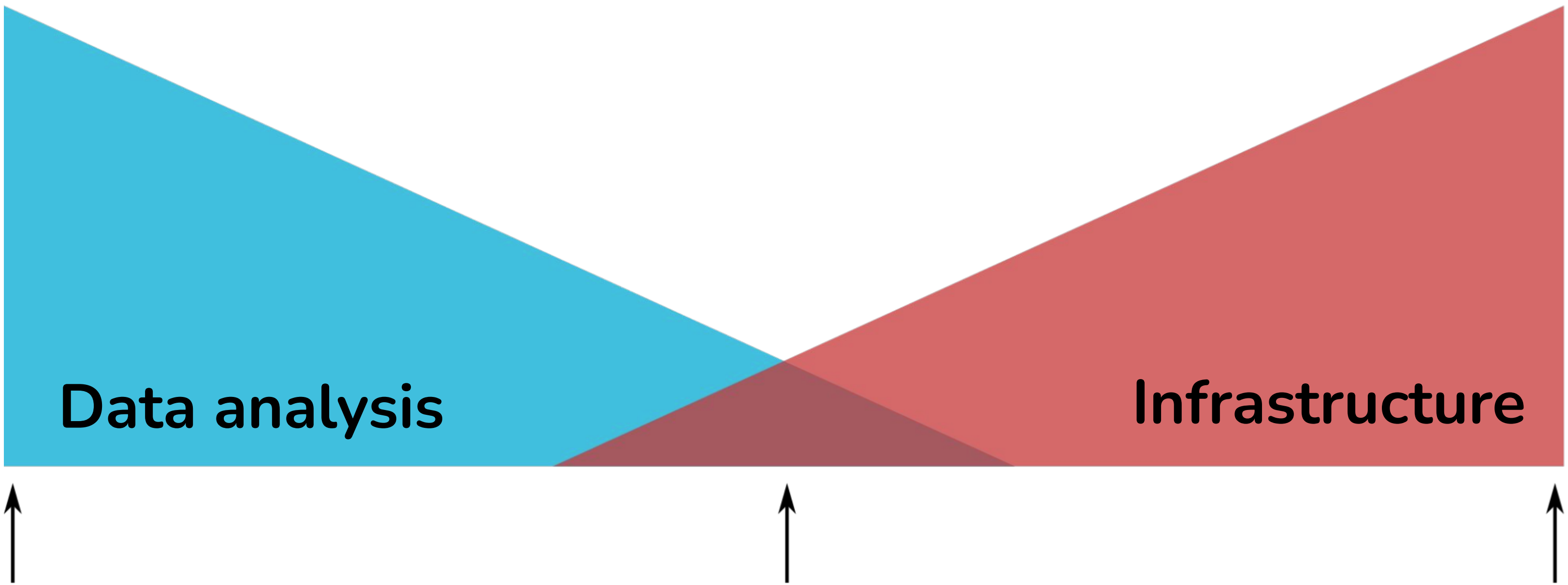



https://mattturck.com/landscape/mad2023.pdf

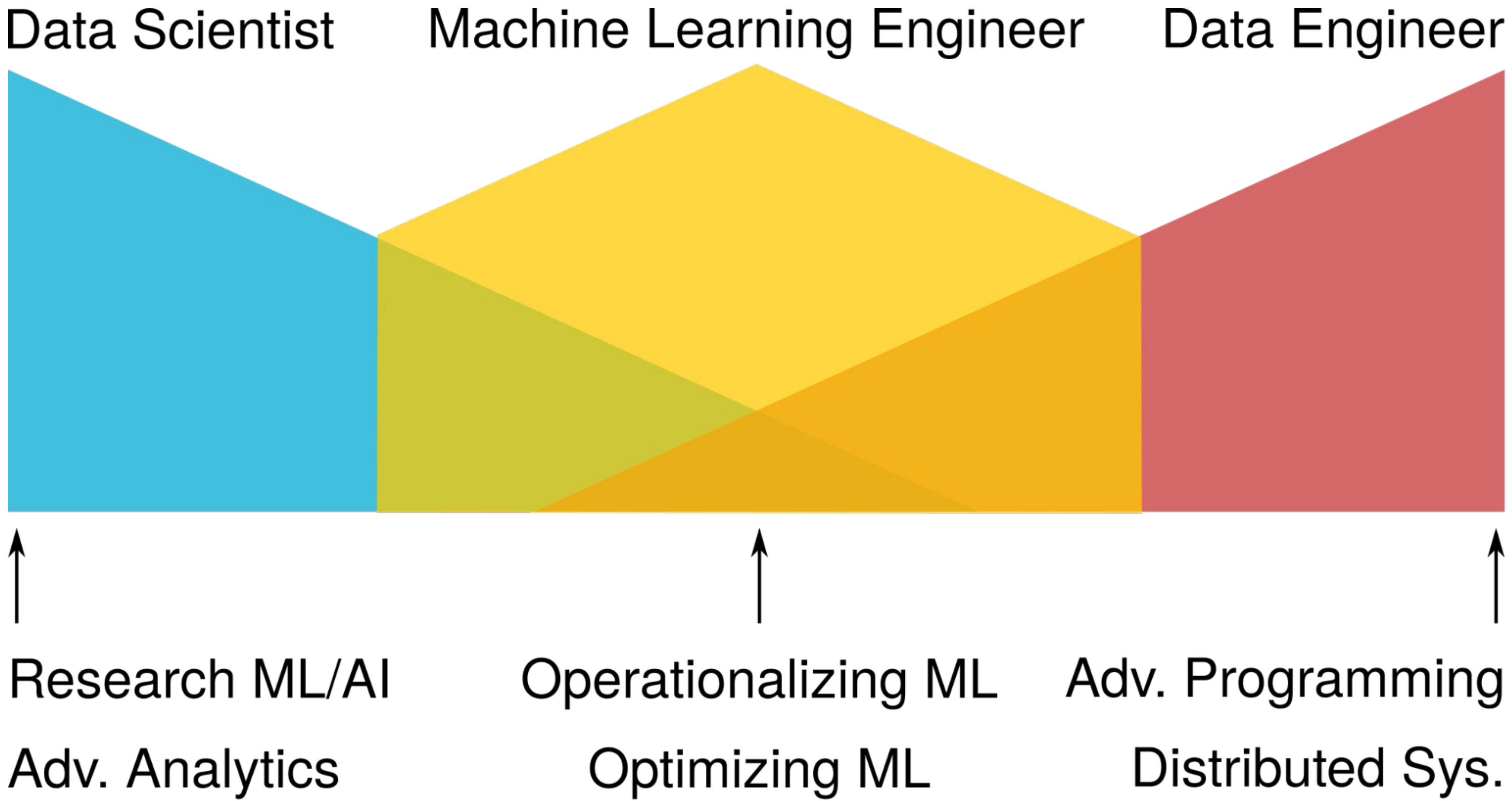
# Open Source


FRAMEWORKS
Spark
Flink
yarn
kubernetes
MESOS
docker
CDAP
Red Hat
HighAvailability
HELIX
dask
RAY
FORMAT
ICEBERG
Parquet
Hudi
orc
ARROW
DELTA LAKE
QUERY / DATA FLOW
Spark
HIVE
presto
trino
GraphQL
DRILL
beam
HAWQ
DATA ACCESS
Amundsen
Data Hub
magda.
ckan
DATABASES
PostgreSQL
MySQL
MongoDB
redis
CockroachDB
MariaDB
influxdata
SciDB
CouchDB
riak
OPENTSDB
accumulo
Milvus
ksqlDB
EDGE DB
HBASE
OLAP
ClickHouse
druid
pinot
HYDRA
ORCHESTRATION
Apache Airflow
PREFECT
DAGSTER
Flyte
METAFLOW
Luigi
argo
MAGE
STREAMING & MESSAGING
Spark Streaming
kafka
beam
PULSAR
Flink
STORM
RabbitMQ
Apache RocketMQ
samza
nifi
debezium
ESTUARY Flow
Gazette
memphis
bytewax
STAT TOOLS & LANGUAGES
python
Scala
NumPy
pandas
SciPy
RStudio
julia
GNU Octave
MLOPS & AI INFRA
Pachyderm
mlflow
Kubeflow
mleap
DVC
SELDON
snorkel
BENTOML
MediaPipe
deepchecks
LangChain
AI FRAMEWORKS & LIBRARIES
TensorFlow
PyTorch
Keras
fast.ai
OpenCV
learn
XGBoost
PyTorch Lightning
飞桨
mxnet
Chainer
ONNX
LUDWIG
HOROVOD
CoreNLP
Spark
H2O.ai
SINGA
AI MODELS & ARCHITECTURES
NVIDIA Megatron
Google Research Bert
OpenAI GPT2
OpenAI CLIP
Google Research T5
stability.ai
DeepMind
Meta AI
RoBERTa
DistilBERT
EleutherAI
XLNet
BLOOM
Google AI Flan-T5
[RIFFUSION]
Meta AI
Google Research
Google Research Dreamix
SEARCH
Solr
Lucene
meilisearch
Sonic
Toshi Search
tantivy
LOGGING & MONITORING
elastic
kibana
Sentry
logstash
Prometheus
fluentbit
fluentd
Grafana
VECTOR
OpenTelemetry
VISUALIZATION
D3
Superset
matplotlib
redash
TensorBoard
seaborn
bokeh
ggplot2
COLLABORATION
BeakerX
jupyter
Apache Zeppelin

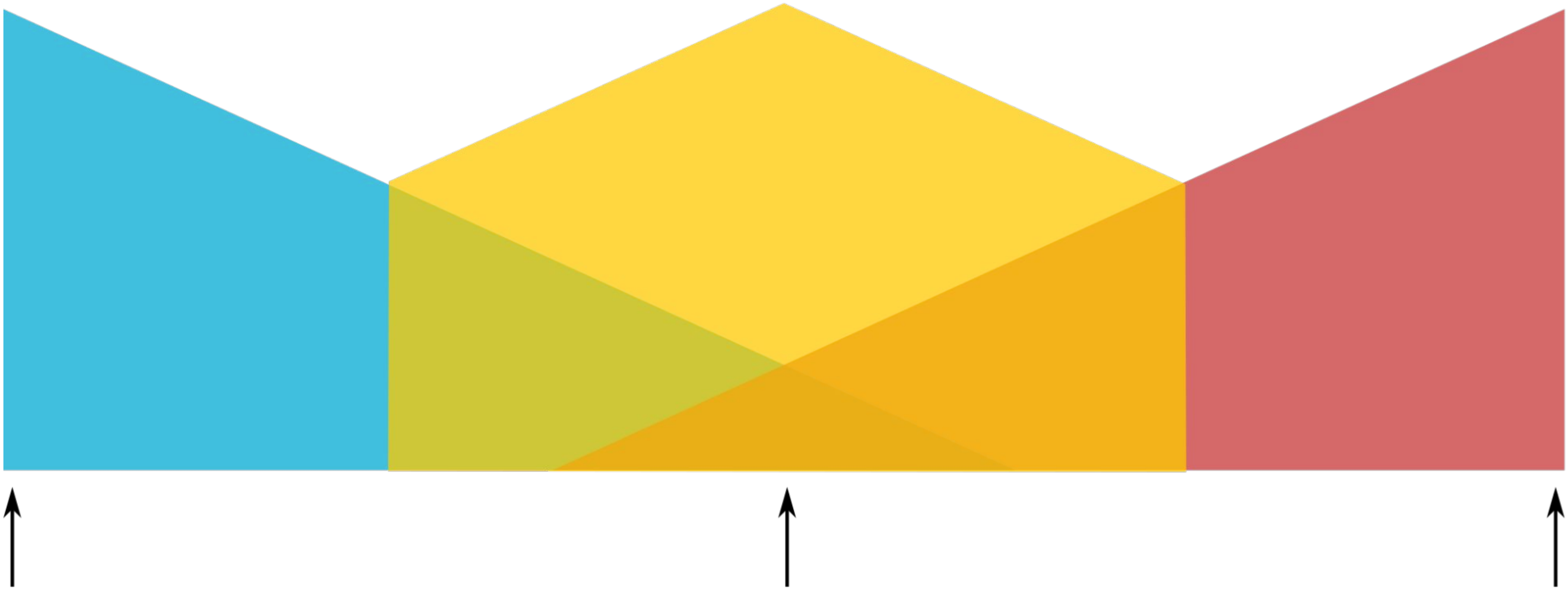
| Data Scientist | | Data Engineer |
|---|---|---|




| | | |
|---|---|---|
| Core Comptencies | Overlapping Skills | Core Comptencies |
| Adv. Math/Statistics | Analysis | Adv. Programming |
| ML/AI | Programming | Distributed Sys. |
| Adv. Analytics | Big Data | Data Pipelines |


Data Scientist
Machine Learning Engineer
Data Engineer
Research ML/AI
Adv. Analytics
Operationalizing ML
Optimizing ML
Adv. Programming
Distributed Sys.

Data Scientist | Machine Learning Engineer | Data Engineer



Research ML/AI | Operationalizing ML | Adv. Programming

Adv. Analytics | Optimizing ML | Distributed Sys.