Big data science Day 1

F. Legger - INFN Torino

https://github.com/leggerf/MLCourse-INFN-2021

Schedule

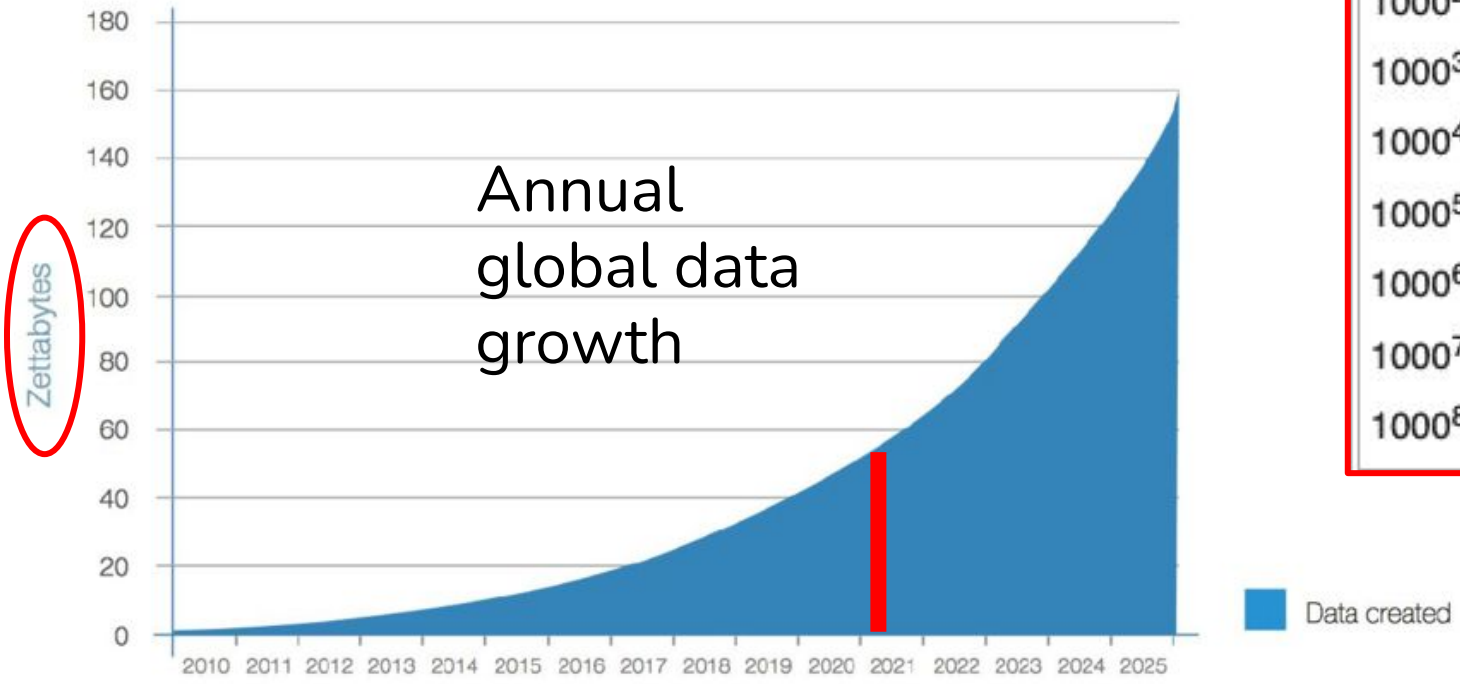
- **Mon 11 - Fr 15:**
  - 10:00 - 12:00 Lectures
  - 12:00 - 16:00 1 h lecture on practical aspects and 1 h hands-on or 2 h hands-on

Today

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

What is big data?

- Data that is too big to be analysed traditionally

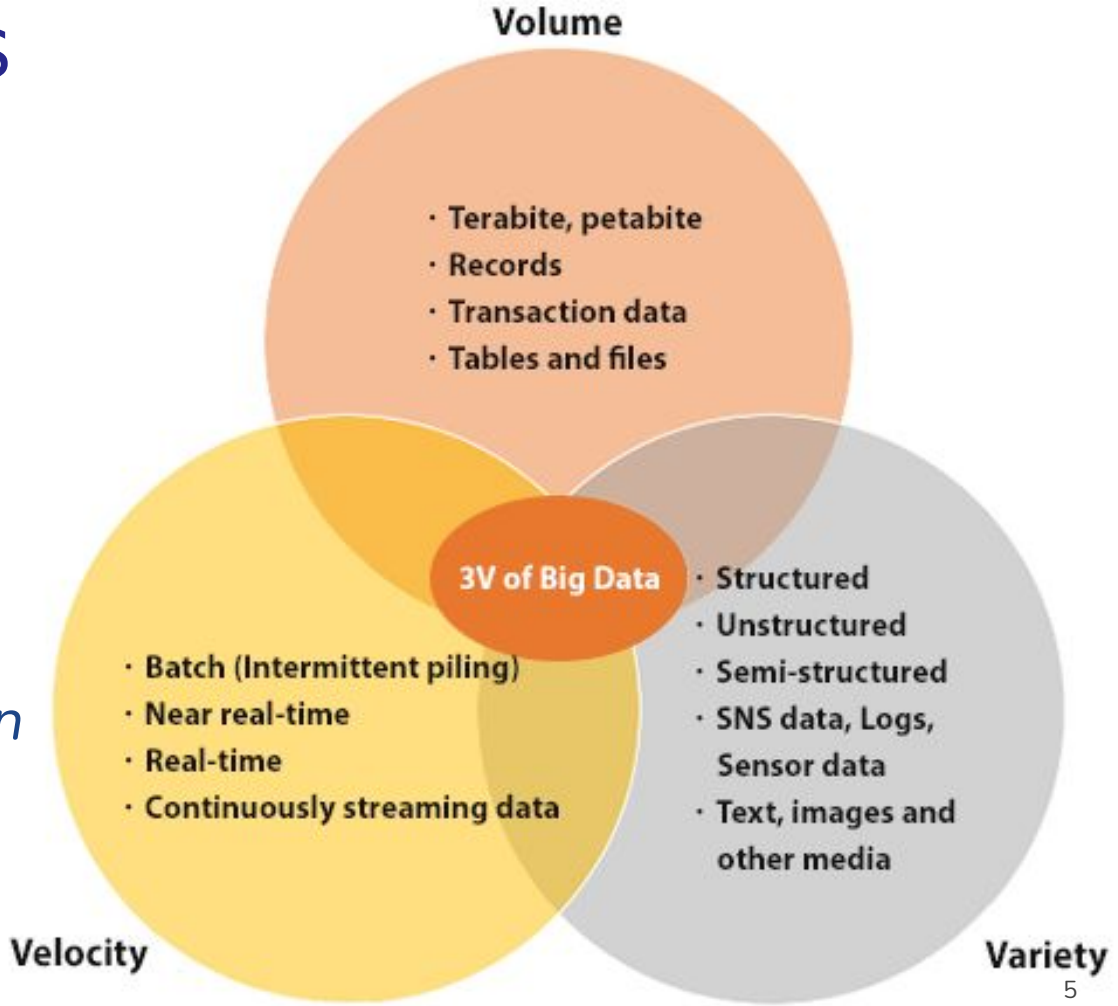


Decimal		
Value	Metric	
1000	kB	kilobyte
1000^2	MB	megabyte
1000^3	GB	gigabyte
1000^4	TB	terabyte
1000^5	PB	petabyte
1000^6	EB	exabyte
1000^7	ZB	zettabyte
1000^8	YB	yottabyte

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)

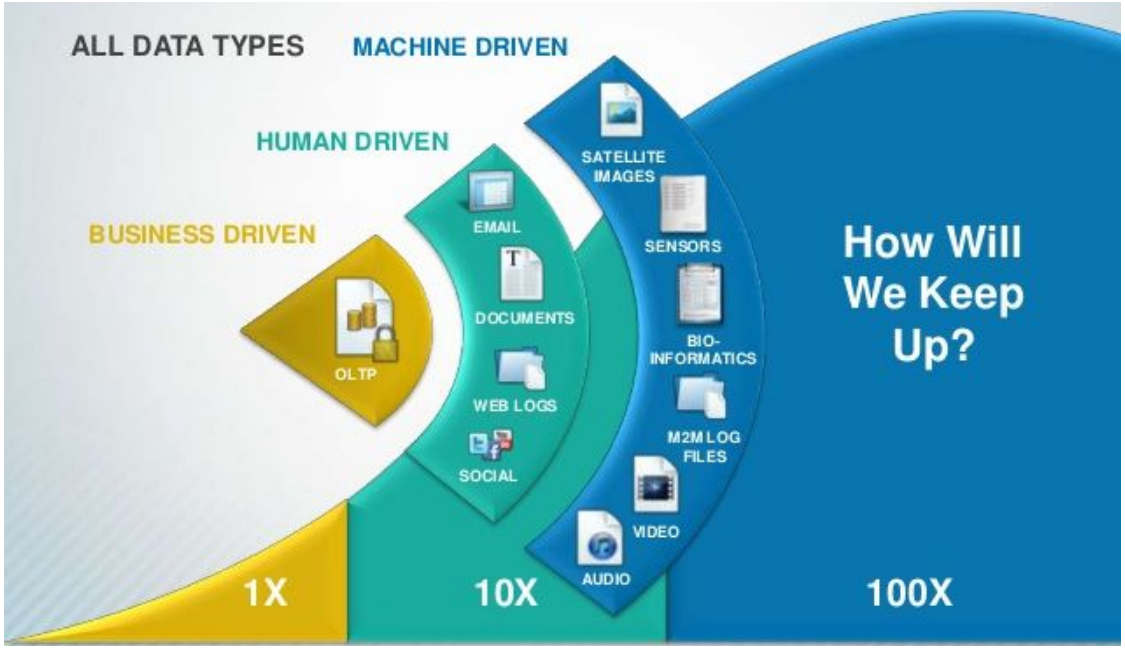


Even more Vs

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

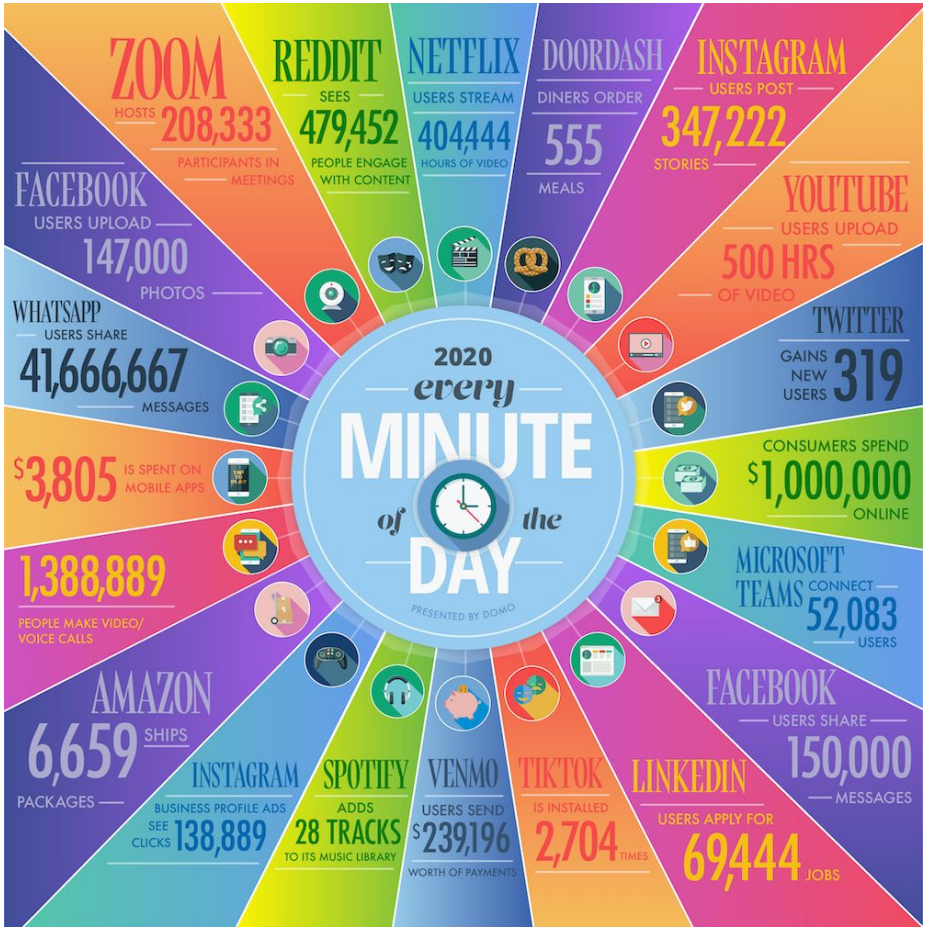
https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx

Big data sources: Business



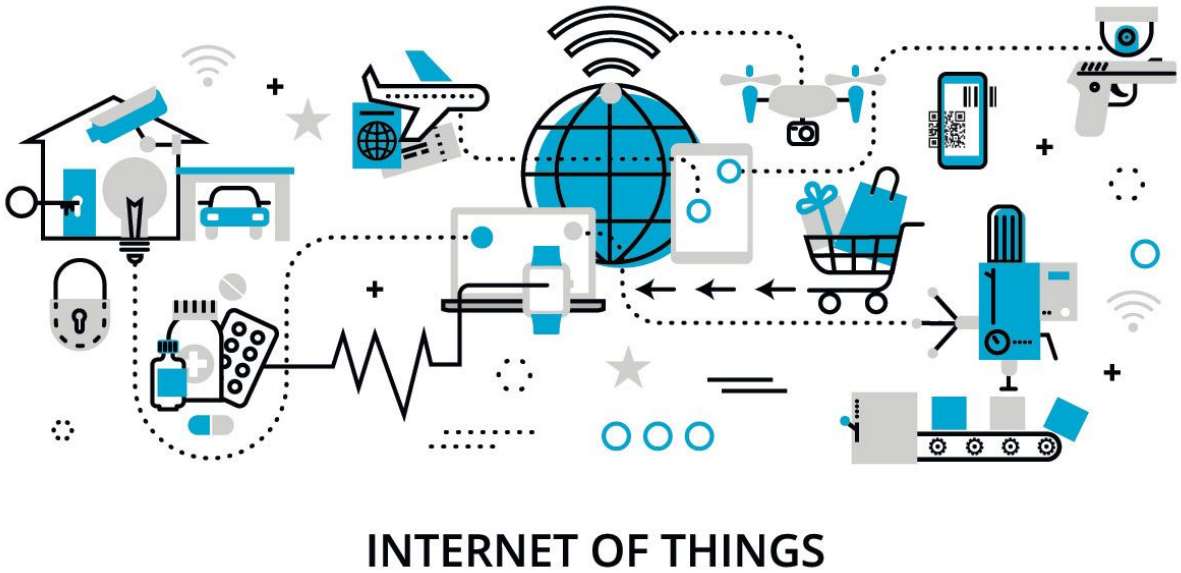
- **Traditional Business Systems**
  - Commercial transactions
  - Banking/stock records
  - E-commerce
  - Credit cards
  - Medical records

Big data sources: Human (x10)



- **Social Networks**
  - Twitter and Facebook
  - Blogs and comments
  - Video conferences, Zoom, Teams
  - Videos: YouTube, Netflix
  - Internet searches
  - Deliveries
  - User-generated maps
  - E-Mail

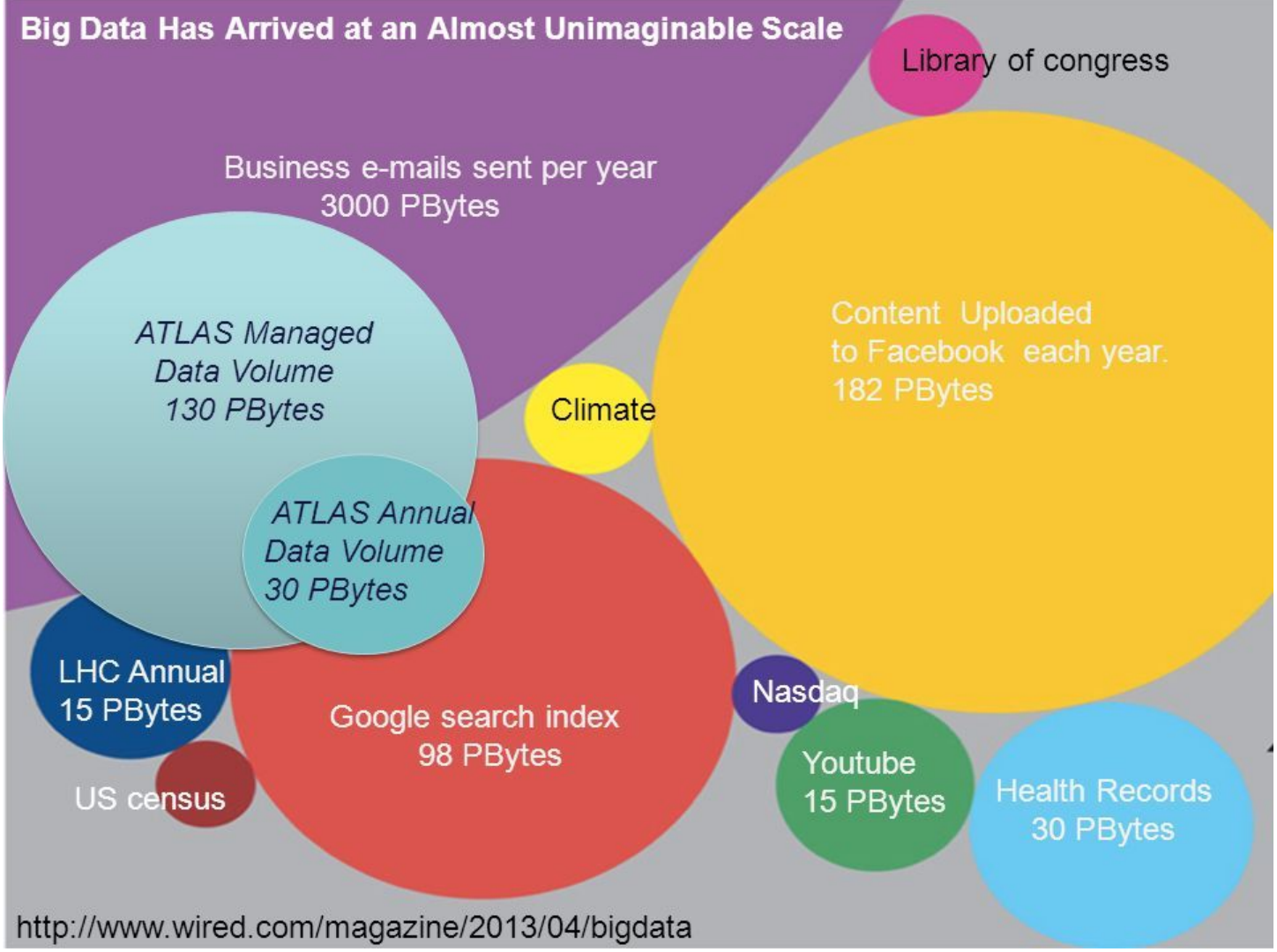
Big data sources: Machine (x100)

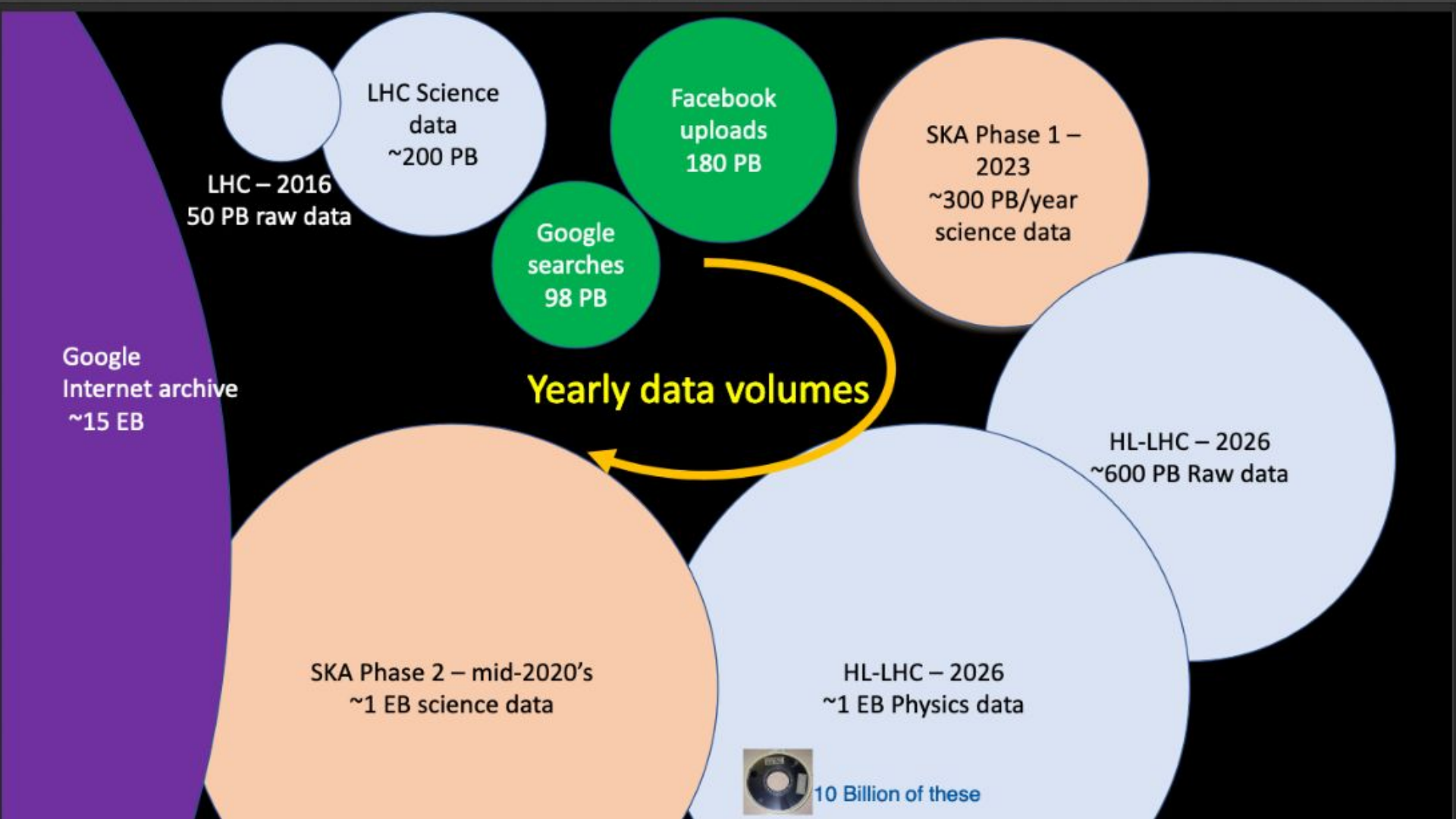


- **Internet of Things (IoT)**
  - Sensors: traffic, weather, mobile phone location, etc.
  - Security, surveillance videos, and images
  - Satellite images
  - Data from computer systems (logs, web logs)

Big data applications

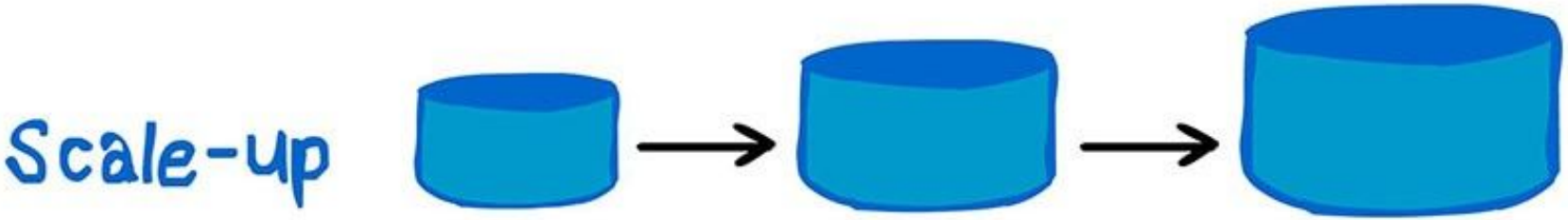
- **Entertainment:** Netflix and Amazon make shows and movie recommendations to their users.
- **Insurance:** predict illness, accidents and price their products accordingly.
- **Driverless Cars**: Google's driverless cars collect about 1 GB/s.
- **Automobile:** Rolls Royce fits hundreds of sensors into its engines and propulsion systems. Real time data are used to schedule maintenance.
- **Government:** analyse patterns and influence election results (Cambridge Analytica Ltd.), people movement during COVID lockdown, ...


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
Content Uploaded
to Facebook each year.
182 PBytes
Climate
ATLAS Annual
Data Volume
30 PBytes
LHC Annual
15 PBytes
Google search index
98 PBytes
Nasdaq
US census
Youtube
15 PBytes
Health Records
30 PBytes
http://www.wired.com/magazine/2013/04/bigdata

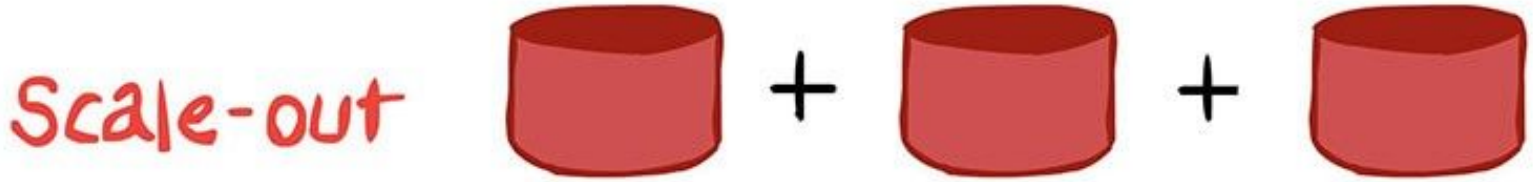

LHC – 2016
50 PB raw data
LHC Science data
~200 PB
Facebook uploads
180 PB
Google searches
98 PB
SKA Phase 1 – 2023
~300 PB/year
science data
Google
Internet archive
~15 EB
Yearly data volumes
HL-LHC – 2026
~600 PB Raw data
SKA Phase 2 – mid-2020's
~1 EB science data
HL-LHC – 2026
~1 EB Physics data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

Scale up vs scale out

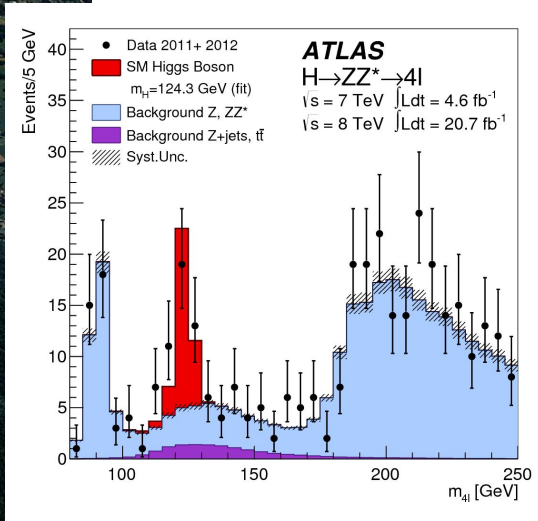
HPC, super computers



Grid, cloud



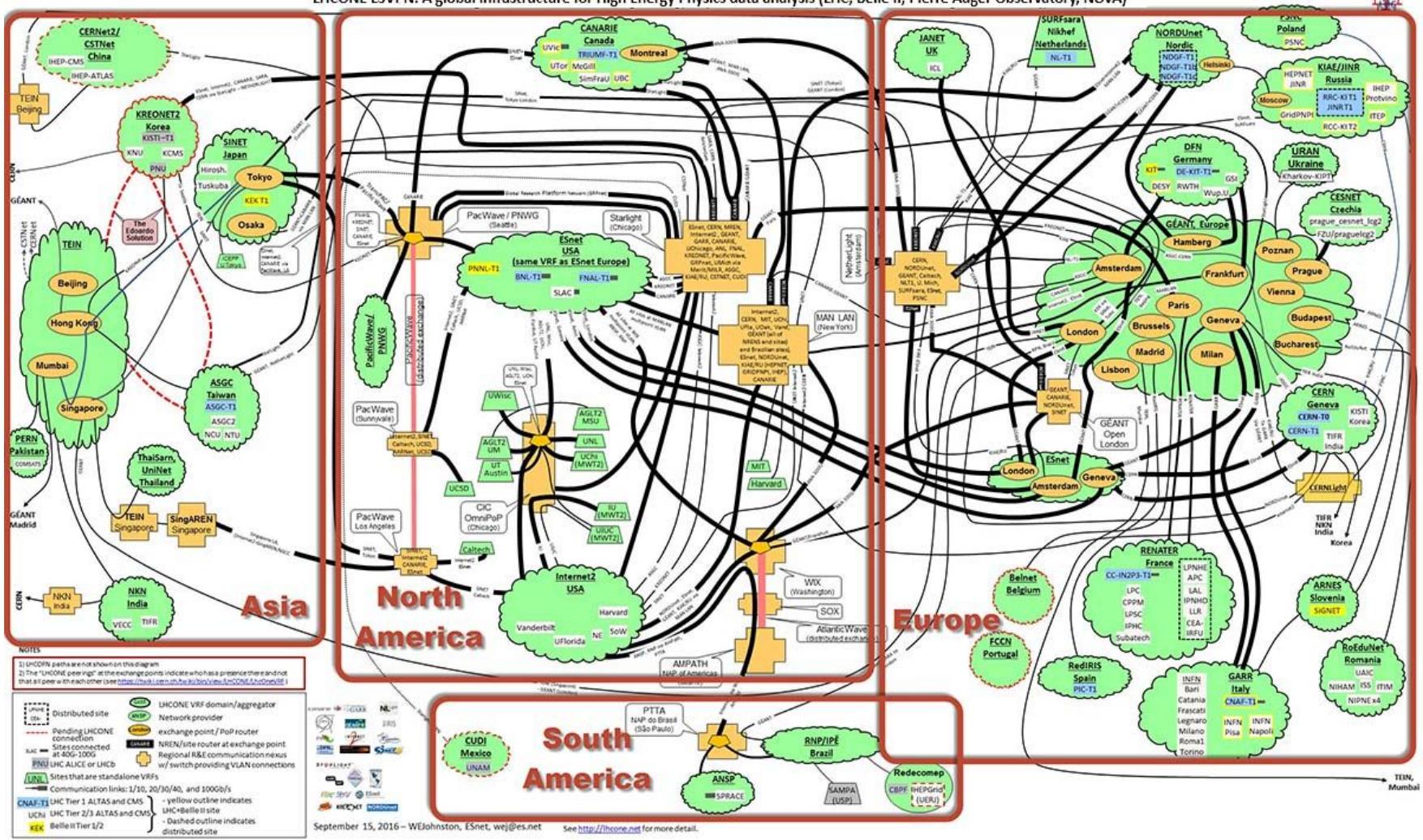
- pp (or Pb-Pb) collisions
- 4 experiments (ATLAS, CMS, LHCb, ALICE)
- Discovery of Higgs boson
- Nobel prize for physics 2013



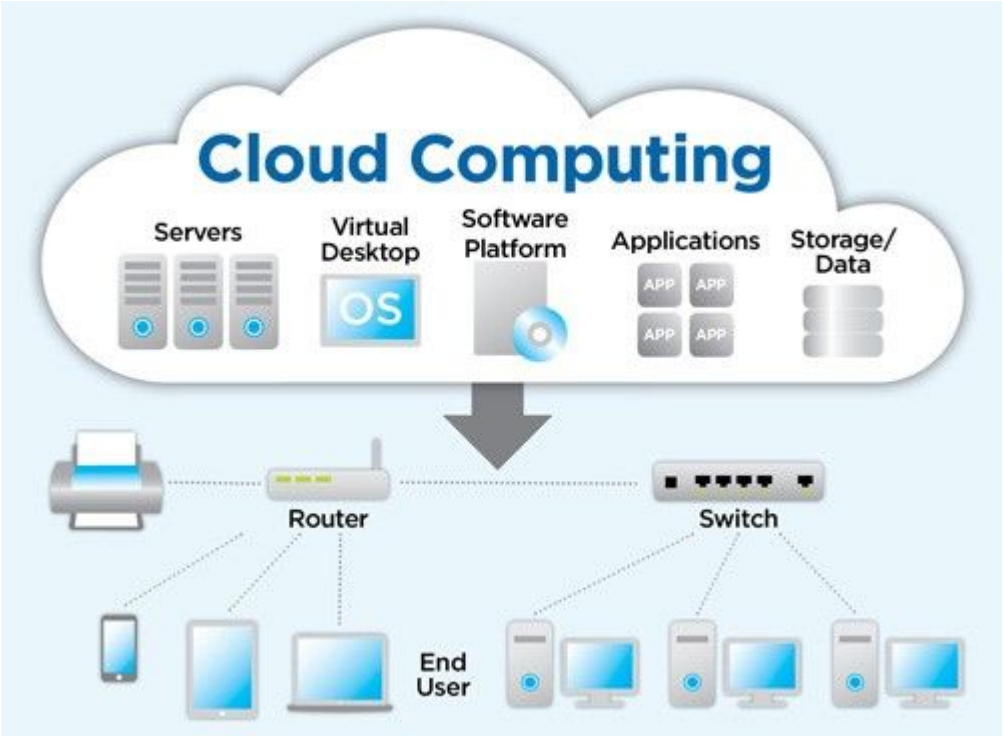
Big data @LHC

- ATLAS/CMS: 100-megapixel digital cameras that take 40 million “pictures” per second == 40 TB raw data/second
- Trigger system (real time): “empty” pictures immediately thrown away: 1GB/second
- Only one in a billion is an Higgs boson!
- Data stored, processed and analysed using the **Worldwide LHC Computing Grid (WLCG)**


11/28/2013 11:44:13 am
11:40 am
11:51 am
Running jobs: 214268
Transfer rate: 42.74 GiB/sec
World
LHC
Computing
Grid
42 countries
170 computing centres
Over 2 million tasks daily
1 million computer cores
1 exabyte of storage
US Dept of State Geographer
© 2013 Google
Image Landsat
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Google earth


LHCONE L3VPN: A global infrastructure for High Energy Physics data analysis (LHC, Belle II, Pierre Auger Observatory, NOvA)
Asia
North America
Europe
South America
CERNet2/ CSTNet China
IHEP-CMS
IHEP-ATLAS
TEIN Beijing
KREONET2 Korea
KISTI-T1
KNU
KCMS
PNU
SINET Japan
Hirosh.
Tokyo
Tsukuba
KEK T1
Osaka
The Edoardo Solution
GÉANT
TEIN
Beijing
Hong Kong
Mumbai
Singapore
ASGC Taiwan
ASGC-T1
ASGC2
NCU
NTU
PERN Pakistan
COMSATS
ThaiSarn, UniNet Thailand
TEIN Singapore
SingAREN Singapore
GÉANT Madrid
CERN
NKN India
VECC
TIFR
CANARIE Canada
UVic
Montreal
UTor
McGill
SimFraU
UBC
PacWave / PNWG (Seattle)
Starlight (Chicago)
ESnet USA (same VRF as ESnet Europe)
PNNL-T1
BNL-T1
FNAL-T1
SLAC
PacificWave/ PNWG
PacificWave (distributed exchange)
PacWave (Sunnyvale)
PacWave Los Angeles
UWISC
AGLT2 MSU
AGLT2 UM
UNL
UCHi (MWT2)
UT Austin
UCSD
CIC OmniPoP (Chicago)
IU (MWT2)
UIUC (MWT2)
Caltech
MIT
Harvard
Internet2 USA
Vanderbilt
Harvard
SoW
UFlorida
NE
WIX (Washington)
SOX
AtlanticWave
AMPATH NAP of Americas
MAN LAN (New York)
NetherLight (Amsterdam)
JANET UK
ICL
SURFsara Nikhef Netherlands
NL-T1
NORDUnet Nordic
NDGF-T1
NDGF-T1d
NDGF-T1c
Helsinki
PSNC Poland
KIAE/JINR Russia
HEPNET JINR
Moscow
GridPNPI
RRC-KI T1
JINR T1
IHEP Protvino
ITEP
RCC-KI T2
DFN Germany
KIT
DE-KIT-T1
DESY
RWTH
GSI
Wup.U.
URAN Ukraine
Kharkov-KIPT
CESNET Czechia
prague_cesnet_lcg2
FZU/praguelcg2
GÉANT, Europe
Hamberg
Amsterdam
Frankfurt
Poznan
Prague
Vienna
Paris
Brussels
Geneva
Budapest
London
Madrid
Milan
Bucharest
Lisbon
GÉANT Open London
ESnet
London
Amsterdam
Geneva
CERN Geneva
CERN-T0
CERN-T1
KISTI Korea
TIFR India
CERNLight
TIFR NKN India
Korea
Belnet Belgium
FCCN Portugal
RENATER France
CC-IN2P3-T1
LPC
CPPM
LPSC
IPHC
Subatech
LPNHE
APC
LAL
IPNHO
LLR
CEA-IRFU
ARNES Slovenia
SIGNET
RoEduNet Romania
NIHAM
ISS
ITIM
NIPNE x4
RediRIS Spain
PIC-T1
GARR Italy
CNAF-T1
INFN Bari
Catania
Frascati
Legnaro
Milano
Roma1
Torino
INFN Pisa
INFN Napoli
CUDI Mexico
UNAM
PTTA NAP do Brasil (São Paulo)
RNP/IPÊ Brazil
ANSP
SPRACE
SAMPA (USP)
Redecomep
CBPF
RHEPGrid (UERJ)
TEIN, Mumbai
NOTES
1) LHCOPN paths are not shown on this diagram
2) The "LHCONE peerings" at the exchange points indicate who has a presence there and not that all peer with each other (see https://twiki.cern.ch/twiki/bin/view/LHCONE/LhcOneVrf)
Distributed site
Pending LHCONE connection
Sites connected at 40G-100G
LHC ALICE or LHCb
Sites that are standalone VRFs
Communication links: 1/10, 20/30/40, and 100Gb/s
LHC Tier 1 ALTAS and CMS
LHC Tier 2/3 ALTAS and CMS
Belle II Tier 1/2
LHCONE VRF domain/aggregator
Network provider
exchange point / PoP router
NREN/site router at exchange point
Regional R&E communication nexus w/ switch providing VLAN connections
- yellow outline indicates LHC+Belle II site
- Dashed outline indicates distributed site
September 15, 2016 – WEJohnston, ESnet, wej@es.net
See http://lhcone.net for more detail.

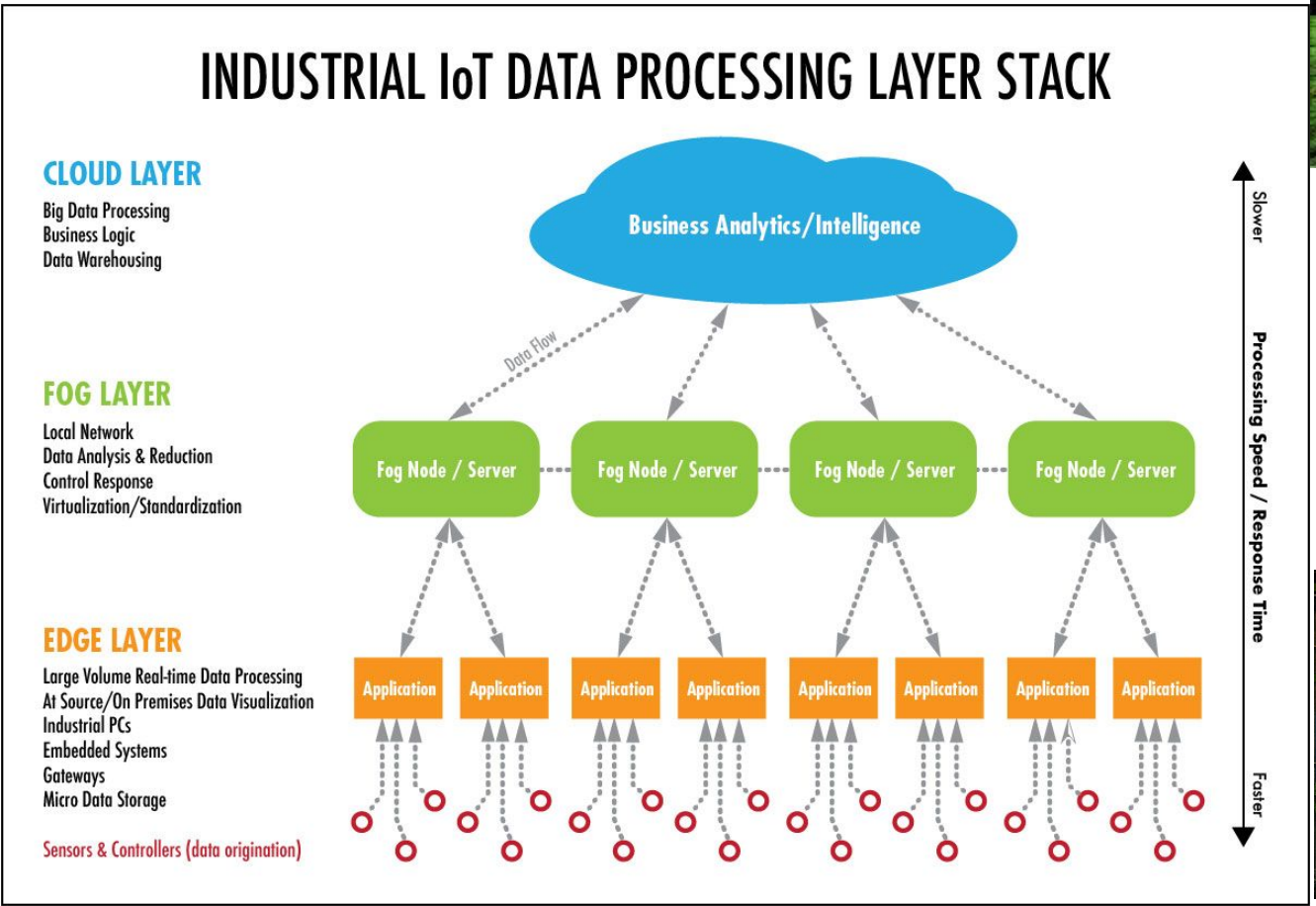
Cloud computing

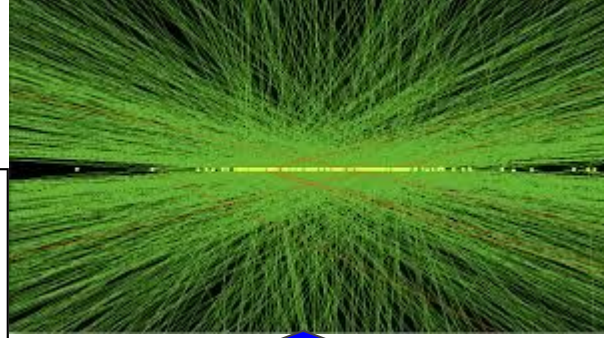


On-demand availability of computer system resources, especially data storage and computing power, without direct active management by the user

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

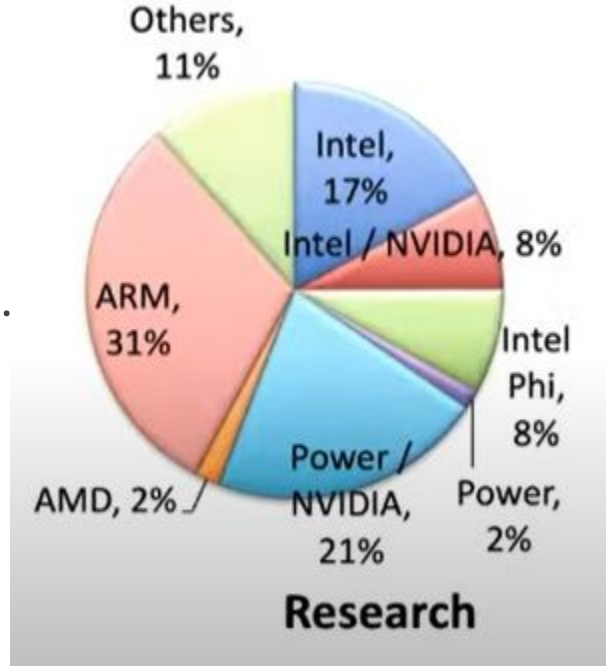
Edge and fog computing





High Performance Computing (HPC)

- Typically involves supercomputers
- Top 500 #1: Fugaku
  - A64FX ARM v8.2-A,
    - Scalable Vector Extension (SVE) instructions and a 512-bit implementation.
    - on-die Tofu-D network BW (~400 Gbps)



https://www.top500.org/

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442,010.0	537,212.0	29,899
2	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	200,794.9	10,096
3	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94,640.0	125,712.0	7,438
4	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,			
5	**Perlmutter** - HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, HPE DOE/SC/LBNL/NERSC United States				

Top European HPCs

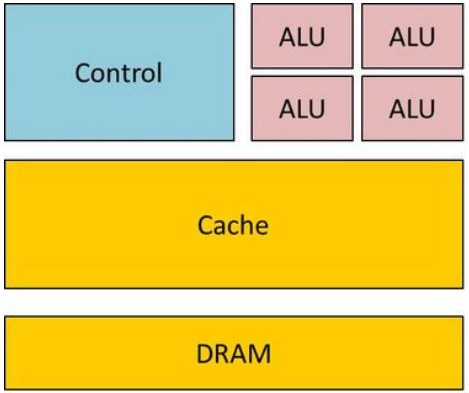
8	**JUWELS Booster Module** - Bull Sequana XH2000 , AMD EPYC 7402 24C 2.8GHz, NVIDIA A100, Mellanox HDR InfiniBand/ParTec ParaStation ClusterSuite, Atos Forschungszentrum Juelich (FZJ) Germany	449,280	44,120.0	70,980.0	1,764
9	**HPC5** - PowerEdge C4140, Xeon Gold 6252 24C 2.1GHz, NVIDIA Tesla V100, Mellanox HDR Infiniband, Dell EMC Eni S.p.A. Italy	669,760	35,450.0	51,720.8	2,252

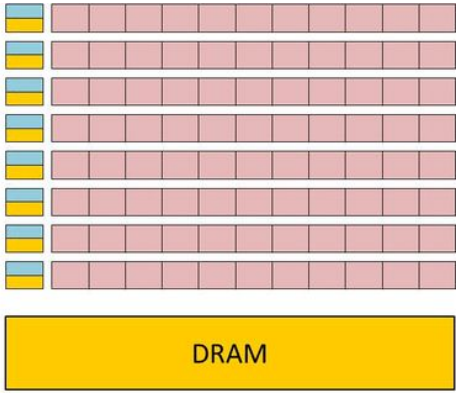
Heterogeneous architectures

CPU vs GPU

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

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





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

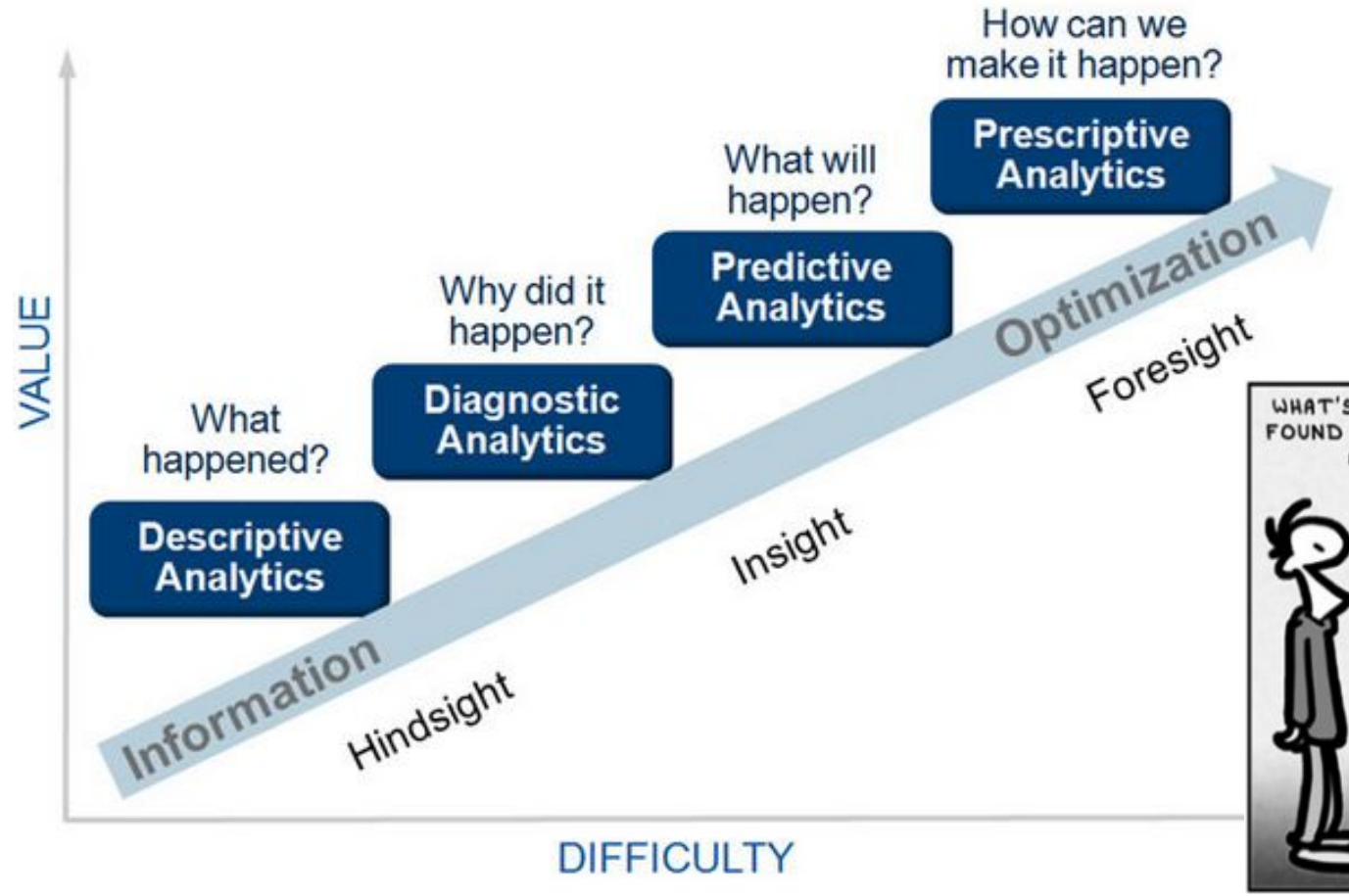
GPU
- Medium-to-large models, datasets
- Image, video processing
- Application on CUDA or OpenCL

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

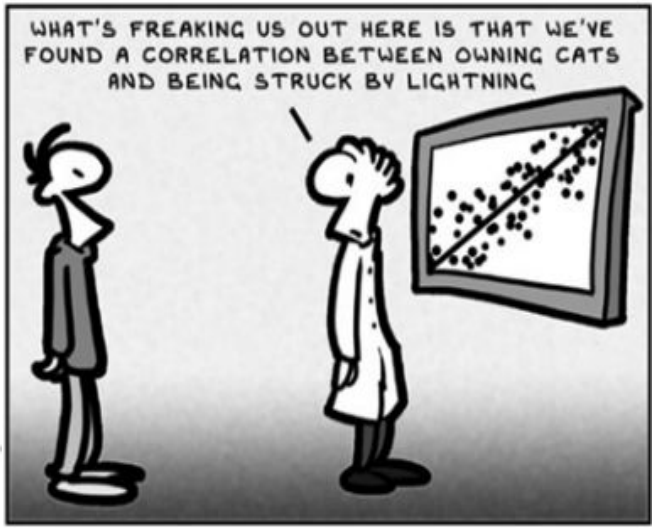
FPGA
- Large datasets, models
- Compute intensive applications
- High performance, high perf./cost ratio

TPU: Tensorflow Processing Unit
FPGA: Field Programmable Gate Array

Analytics


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

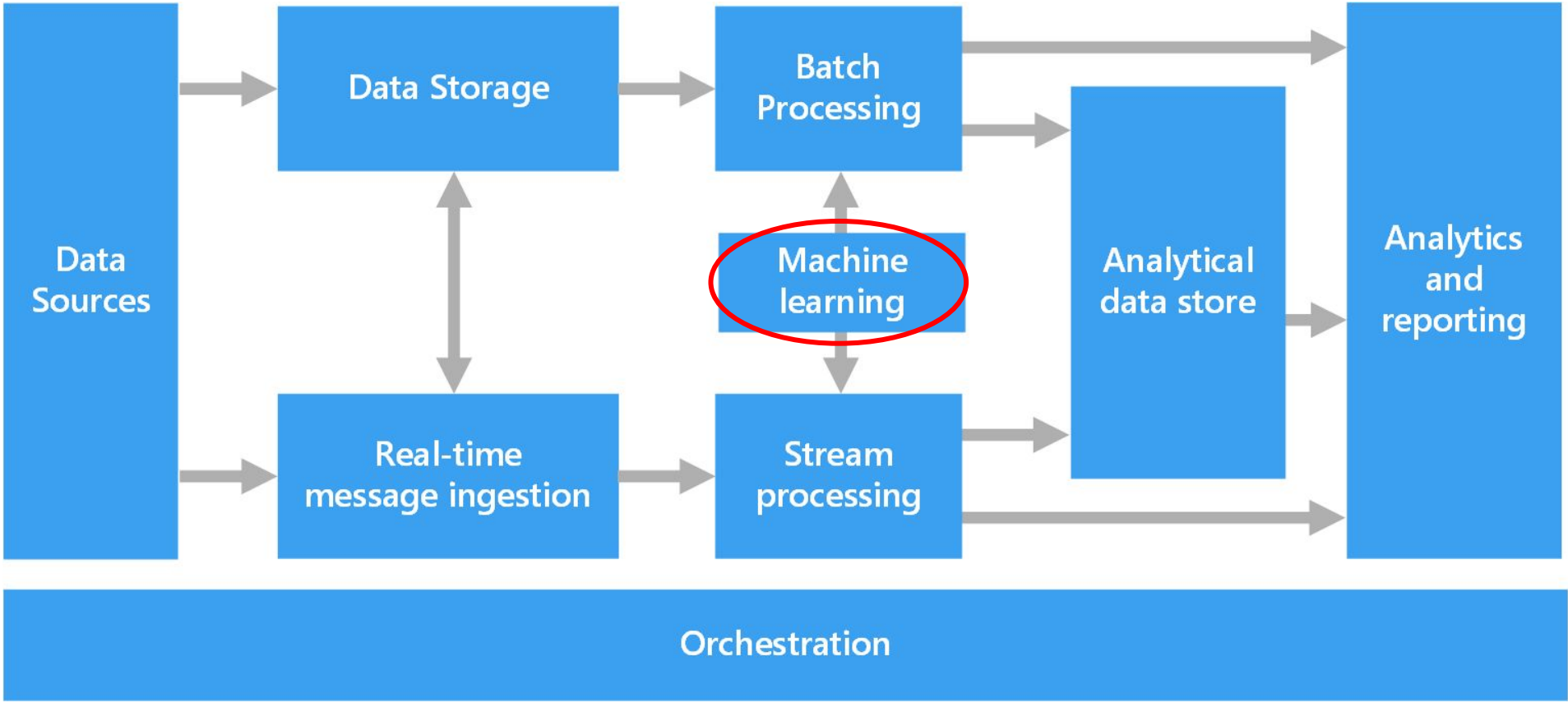

SHOULD WE ALWAYS IGNORE WHAT THE DATA SAYS, OR IS THIS MORE OF A ONE-TIME THING?


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

Analytics applications

- **Churn prediction**
  - Customers switching from one company to another
- **Recommendation system**
  - Netflix reported that **2/3** of the movies watched are recommended
  - Google News stated that recommendations generate **38%** more click-through
  - Amazon claimed that **35%** sales come from recommendations
- **Sentiment analysis**
- **Operational analytics**
  - Automatization
- **Medicine**
  - Remote diagnosis, prevention

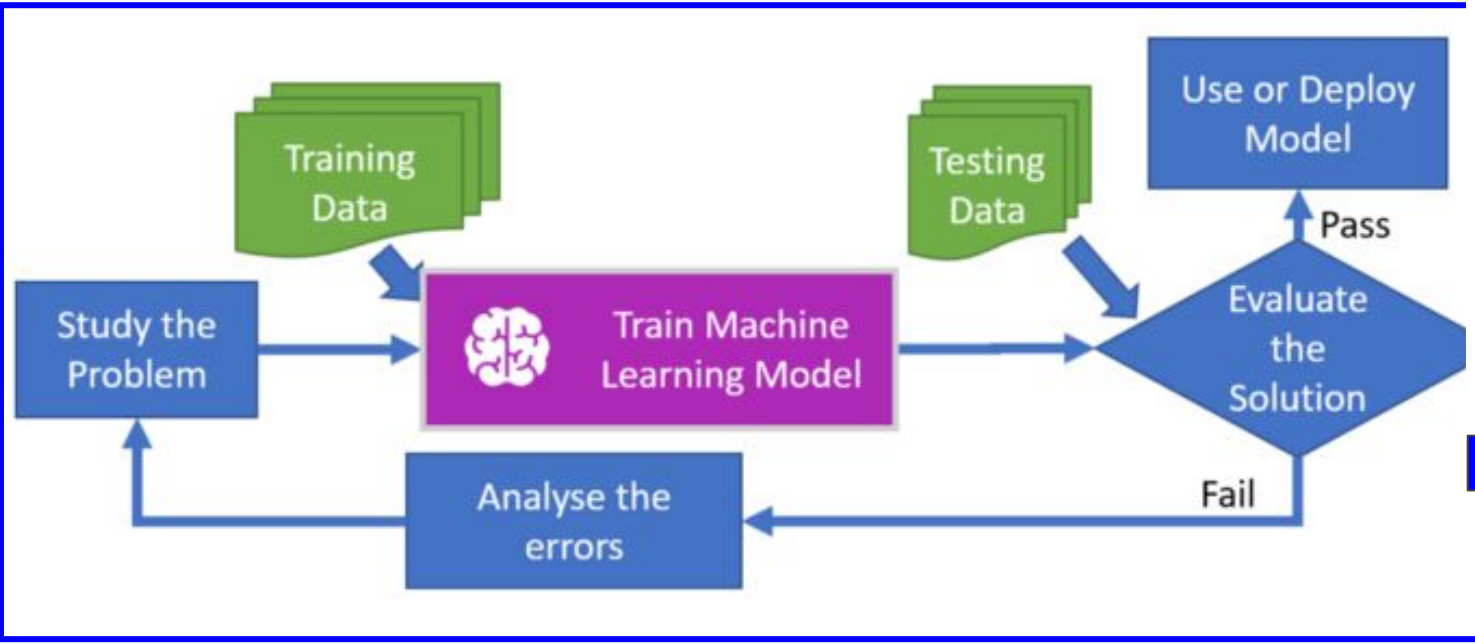
The big data pipeline

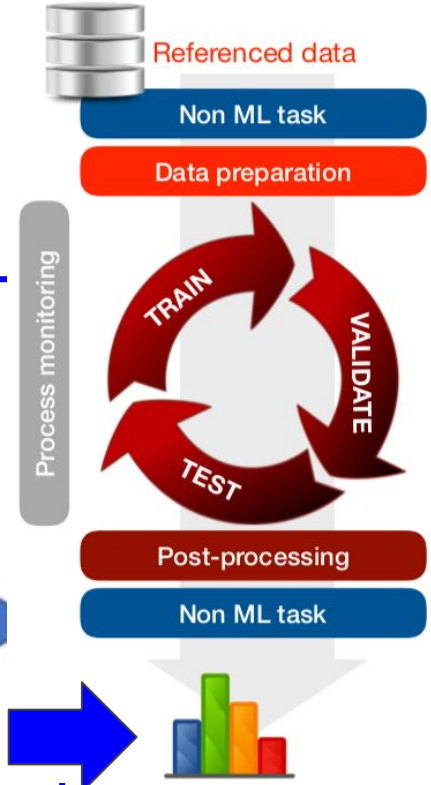


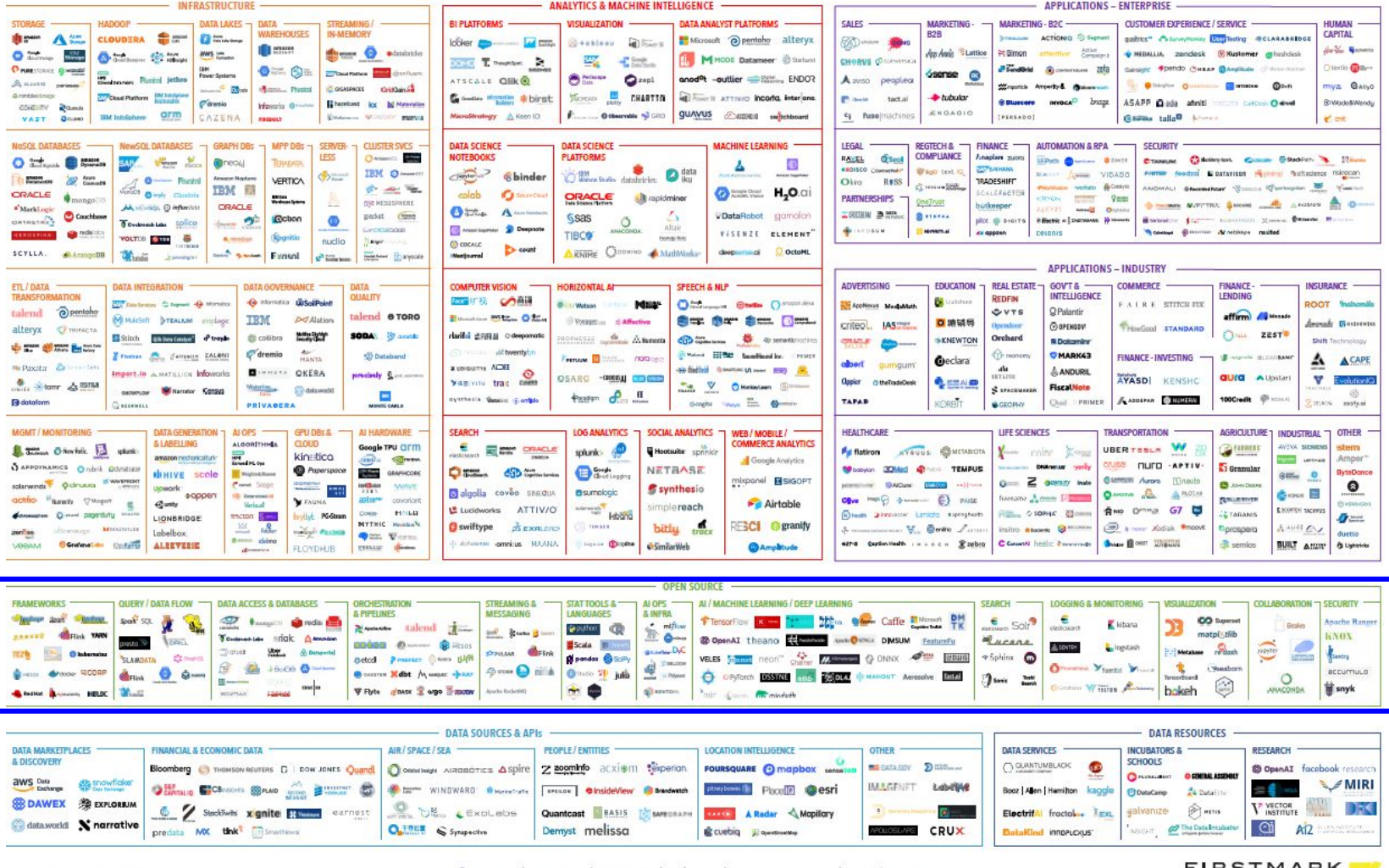
The machine learning pipeline

Input data is typically split into:

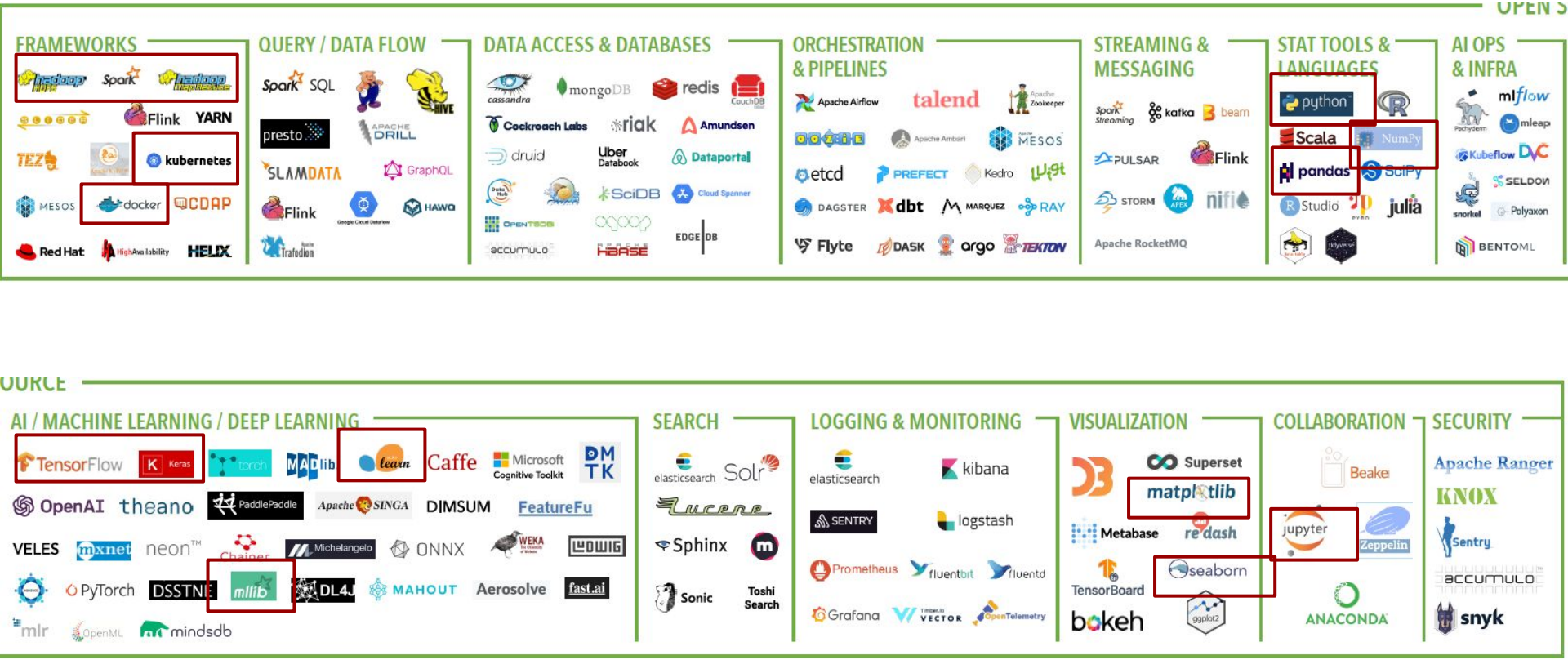
- Training dataset
- Testing dataset

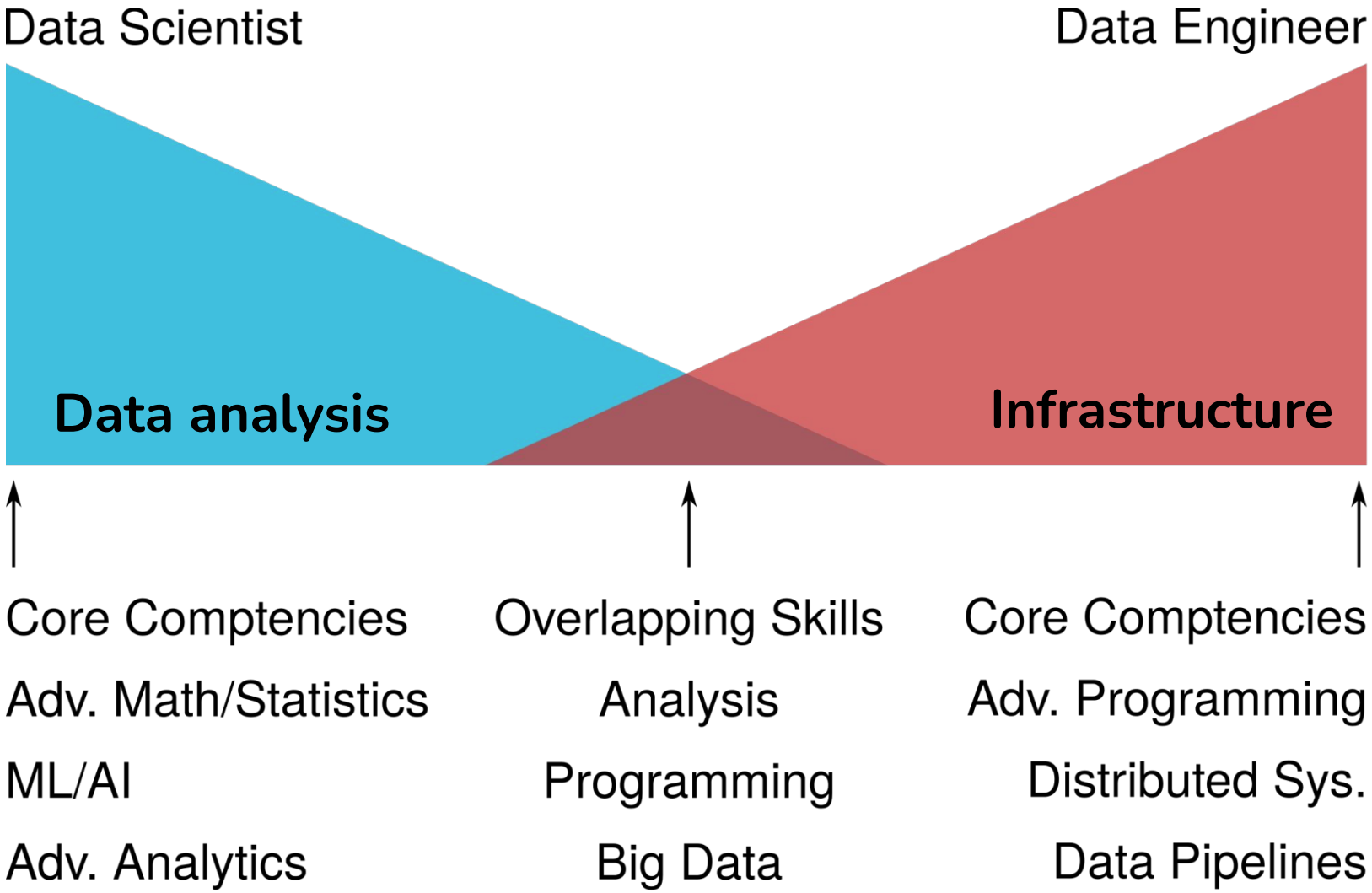


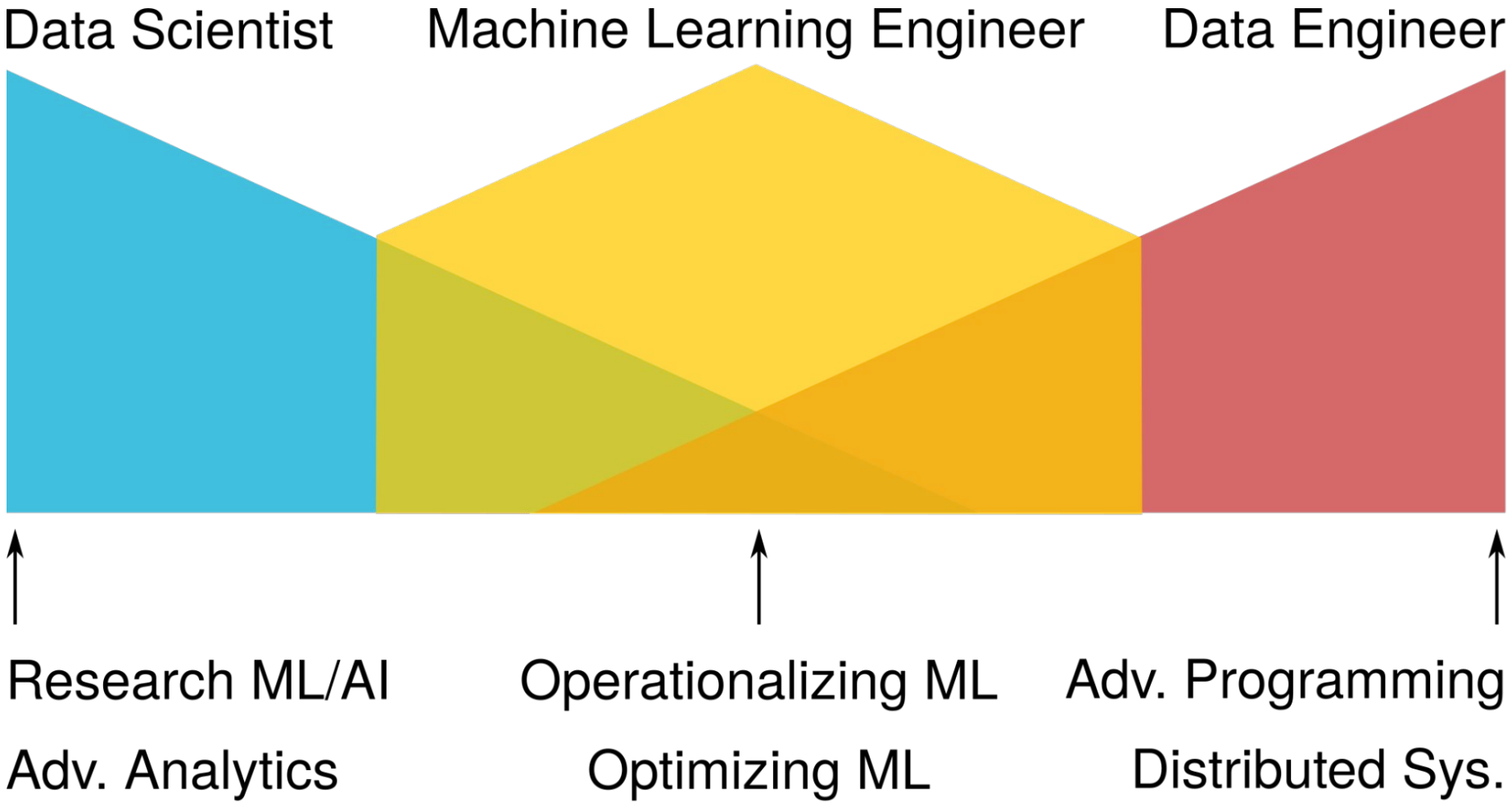



DATA & AI LANDSCAPE 2020
INFRASTRUCTURE
STORAGE
HADOOP
DATA LAKES
DATA WAREHOUSES
STREAMING / IN-MEMORY
NoSQL DATABASES
NewSQL DATABASES
GRAPH DBs
MPP DBs
SERVER-LESS
CLUSTER SVCS
ETL / DATA TRANSFORMATION
DATA INTEGRATION
DATA GOVERNANCE
DATA QUALITY
MGMT / MONITORING
DATA GENERATION & LABELLING
AI OPS
GPU DBs & CLOUD
AI HARDWARE
ANALYTICS & MACHINE INTELLIGENCE
BI PLATFORMS
VISUALIZATION
DATA ANALYST PLATFORMS
DATA SCIENCE NOTEBOOKS
DATA SCIENCE PLATFORMS
MACHINE LEARNING
COMPUTER VISION
HORIZONTAL AI
SPEECH & NLP
SEARCH
LOG ANALYTICS
SOCIAL ANALYTICS
WEB / MOBILE / COMMERCE ANALYTICS
APPLICATIONS – ENTERPRISE
SALES
MARKETING - B2B
MARKETING - B2C
CUSTOMER EXPERIENCE / SERVICE
HUMAN CAPITAL
LEGAL
REGTECH & COMPLIANCE
FINANCE
AUTOMATION & RPA
SECURITY
PARTNERSHIPS
APPLICATIONS – INDUSTRY
ADVERTISING
EDUCATION
REAL ESTATE
GOVT & INTELLIGENCE
COMMERCE
FINANCE - LENDING
INSURANCE
FINANCE - INVESTING
HEALTHCARE
LIFE SCIENCES
TRANSPORTATION
AGRICULTURE
INDUSTRIAL
OTHER
OPEN SOURCE
FRAMEWORKS
QUERY / DATA FLOW
DATA ACCESS & DATABASES
ORCHESTRATION & PIPELINES
STREAMING & MESSAGING
STAT TOOLS & LANGUAGES
AI OPS & INFRA
AI / MACHINE LEARNING / DEEP LEARNING
SEARCH
LOGGING & MONITORING
VISUALIZATION
COLLABORATION
SECURITY
DATA SOURCES & APIs
DATA MARKETPLACES & DISCOVERY
FINANCIAL & ECONOMIC DATA
AIR / SPACE / SEA
PEOPLE / ENTITIES
LOCATION INTELLIGENCE
OTHER
DATA RESOURCES
DATA SERVICES
INCUBATORS & SCHOOLS
RESEARCH

Open Source

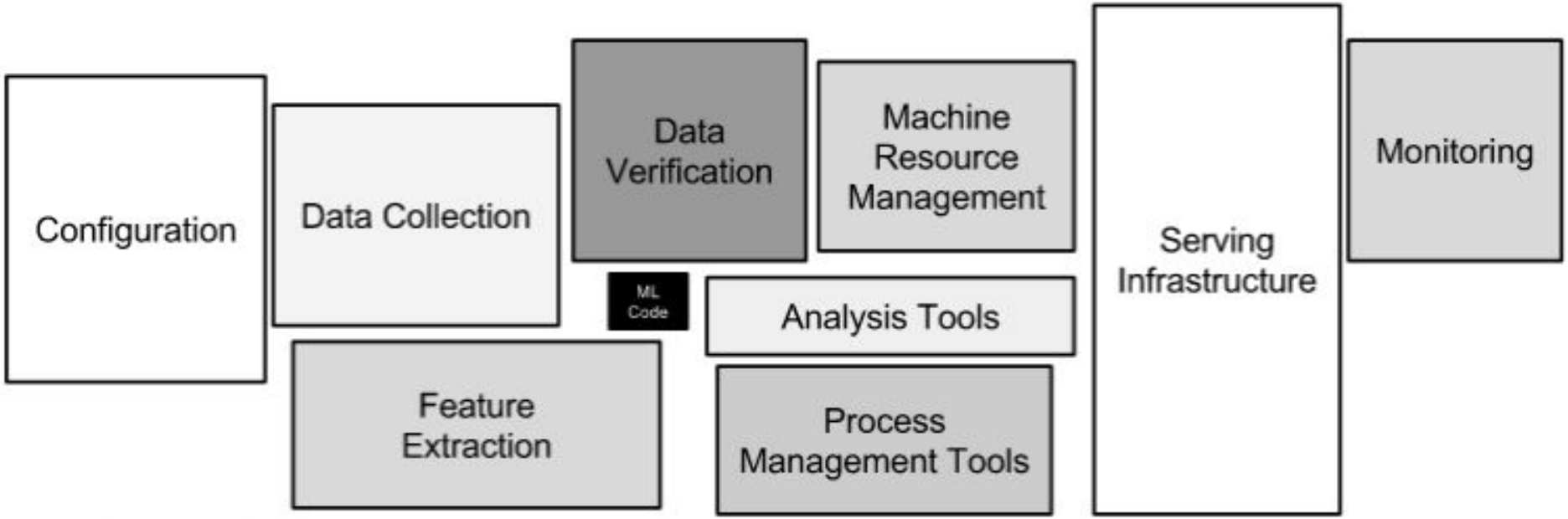

FRAMEWORKS
QUERY / DATA FLOW
DATA ACCESS & DATABASES
ORCHESTRATION & PIPELINES
STREAMING & MESSAGING
STAT TOOLS & LANGUAGES
AI OPS & INFRA
AI / MACHINE LEARNING / DEEP LEARNING
SEARCH
LOGGING & MONITORING
VISUALIZATION
COLLABORATION
SECURITY
Hadoop
Spark
Flink
YARN
kubernetes
docker
MESOS
Red Hat
CDAP
HELIX
Spark SQL
HIVE
presto
APACHE DRILL
SLAMDATA
GraphQL
Flink
HAWQ
Trafodion
cassandra
mongoDB
redis
Cockroach Labs
riak
Amundsen
druid
Uber Databook
Dataportal
SciDB
Cloud Spanner
OpenTSDB
accumulo
HBASE
EDGE DB
Apache Airflow
talend
Apache Zookeeper
OOZIE
Apache Ambari
MESOS
etcd
PREFECT
Kedro
Luigi
DAGSTER
dbt
MARQUEZ
RAY
Flyte
DASK
argo
TEKTON
Spark Streaming
kafka
beam
PULSAR
Flink
STORM
APEX
nifi
Apache RocketMQ
python
R
Scala
NumPy
pandas
SciPy
RStudio
julia
tidyverse
mlflow
Pachyderm
mleap
Kubeflow
DVC
SELDON
snorkel
Polyaxon
BENTOML
TensorFlow
Keras
torch
MADlib
learn
Caffe
Microsoft Cognitive Toolkit
DMTK
OpenAI
theano
PaddlePaddle
Apache SINGA
DIMSUM
FeatureFu
VELES
mxnet
neon
Chainer
Michelangelo
ONNX
WEKA
LOWIG
PyTorch
DSSTNE
mllib
DL4J
MAHOUT
Aerosolve
fast.ai
mlr
OpenML
mindsdb
elasticsearch
Solr
Lucene
Sphinx
Sonic
Toshi Search
elasticsearch
kibana
SENTRY
logstash
Prometheus
fluentbit
fluentd
Grafana
VECTOR
OpenTelemetry
Superset
matplotlib
Metabase
re:dash
jupyter
Zeppelin
TensorBoard
seaborn
bokeh
ggplot2
Beaker
ANACONDA
Apache Ranger
KNOX
Sentry
accumulo
snyk


Data Scientist
Data Engineer
Data analysis
Infrastructure
Core Comptencies
Adv. Math/Statistics
ML/AI
Adv. Analytics
Overlapping Skills
Analysis
Programming
Big Data
Core Comptencies
Adv. Programming
Distributed Sys.
Data Pipelines


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

Typical ML workflow

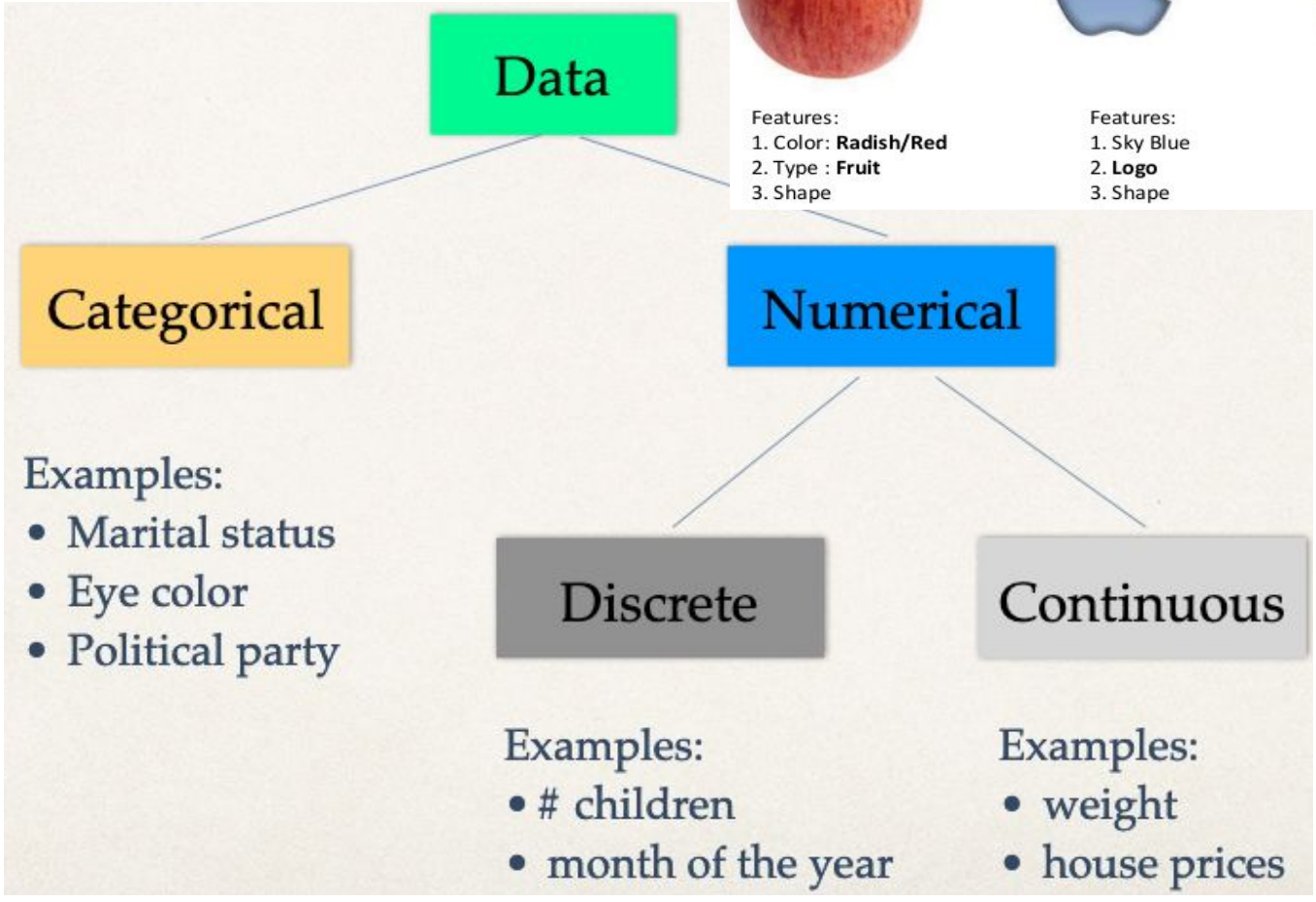
Where is ML?



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

Data types

Typically strings



Features:
1. Color: **Radish/Red**
2. Type : **Fruit**
3. Shape

Features:
1. Sky Blue
2. **Logo**
3. Shape

Features:
1. **Yellow**
2. **Fruit**
3. Shape

Data preprocessing

- Most of the time will be spent in this step
- Data clean-up, data transformation, feature engineering
  - **data transformation**
    - scaling and normalization
    - encoding, aggregation features, log-transformation (to remove outliers)
  - **data visualization, exploration**
  - **data augmentation, bucketing, binning, ...**
  - **dimensionality reduction**
- Your programming skills will be required here: **R, Python,** ...

Data transformation

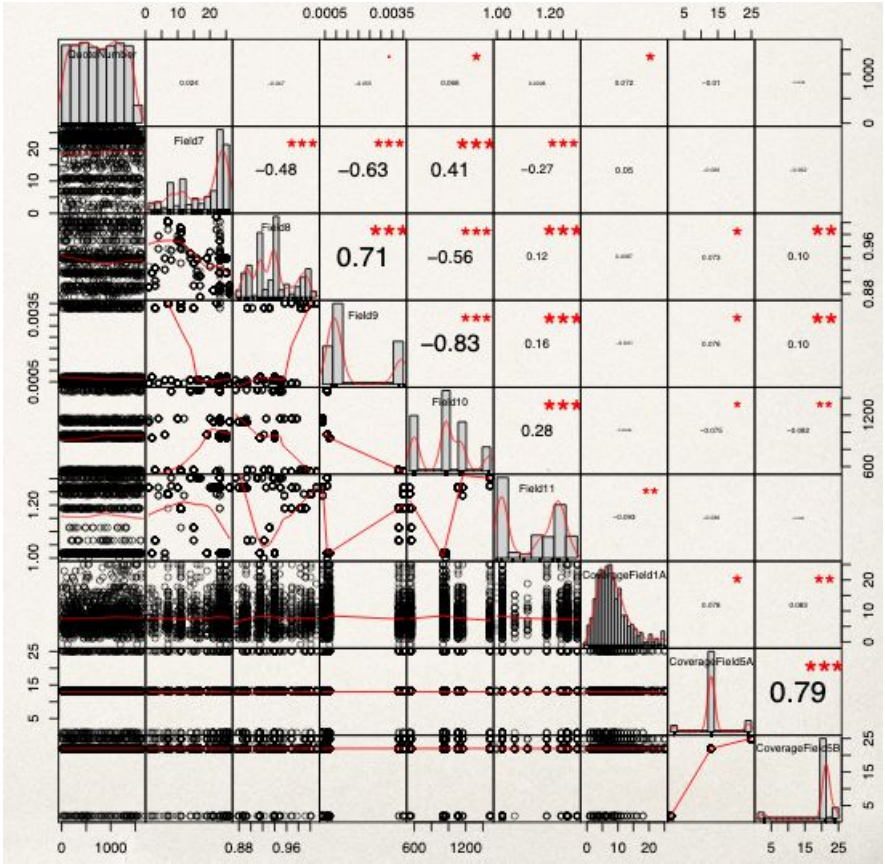
$$x' = \frac{x - \bar{x}}{\sigma}$$

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Data **transformation and *aggregation***: log, sum of values, average, ...
- **Scaling**: a technique to scale data to a given range [0,1] or any other range
- **Normalization/Standardization:** a technique to scale data to mean with zero and and unit-variance
- **Augmentation:** a technique to create additional data based on input sample which slightly differ from it, e.g. image rotation, flip, scale, crop, etc.
- **Bucketing/Binning**: a technique to place similar values into buckets/bins

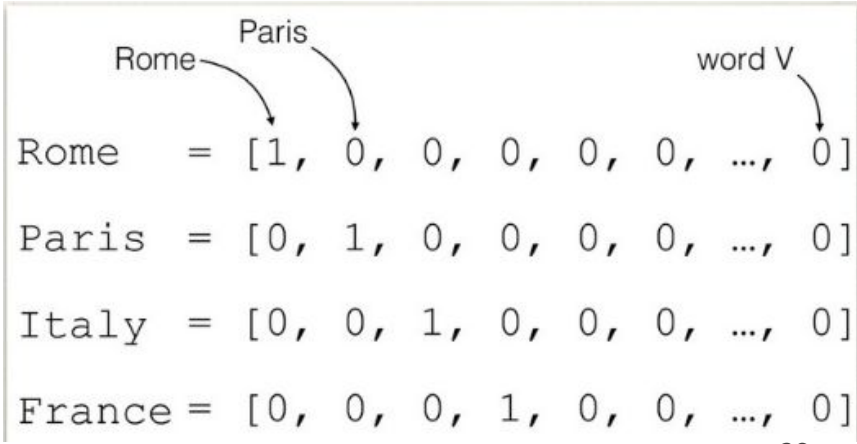
Data visualization

- Graphical representation may reveal important features of the data
  - find correlations, identify range, etc.
- Identify features which may require transformations, e.g. see outliers or skewness (asymmetry in probability distribution) in data
- It helps to identify a strategy how to deal with different features



One-hot encoding

- Technique to handle **categorical** data
- “One-Hot” refers to a state in electrical engineering where all of the bits in a circuit are 0, except a single bit with a value of 1 (“hot”)
- It represents a categorical column as a vector of words
- You need to define the word vector for the full set of data (train + test datasets)
  - Issues with NULL or missing data
    - delete rows with missing data
    - input data for missing values
  - Problematic with high cardinality



https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f

Leave-one-out encoding

- Effective by high cardinality
- Y is what we want to predict
- Encode UserID:
  - Train dataset:
    - Take mean of Y's for all rows with same UserID except the one you want to encode
    - multiply random noise
  - Test dataset
    - No Y, just use frequency of UserID

Mean of [1,1,0]

mean_y* random

Split	UserID	Y	mean_y	random	newID
Train	A1	~~0~~	0.667	1.05	0.70035
Train	A1	1	0.333	0.97	0.32301
Train	A1	1	0.333	0.98	0.32634
Train	A1	0	0.667	1.02	0.68034
Test	A1	-	0.5	1	0.5
Test	A1	-	0.5	1	0.5
Train	A2	0			

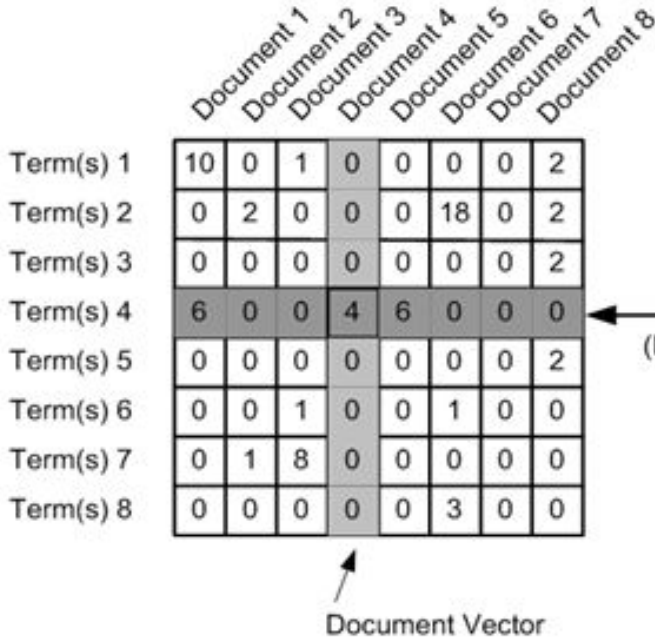
https://www.kaggle.com/c/caterpillar-tube-pricing/discussion/15748

Word embedding

- A way to capture multi-dimensional relationships between categories
  - you define a dimension of word vector up-front
    - it projects categorical variables into another phase space, e.g. days may be sunny or rainy, season or off season, Sunday and Saturday may have similar effect while other days may be treated independently
  - Use neural networks or other ML algorithms to train the model to find the best representation of embedded variables

Frequency based word embedding

- **Count Vector**
  - Corpus C of D documents {d1,d2.....dD} and N unique tokens (words) in C
  - The N tokens will form our dictionary and the size of the Count Vector matrix M will be given by D X N. Each row in the matrix M contains the frequency of tokens in D(i)

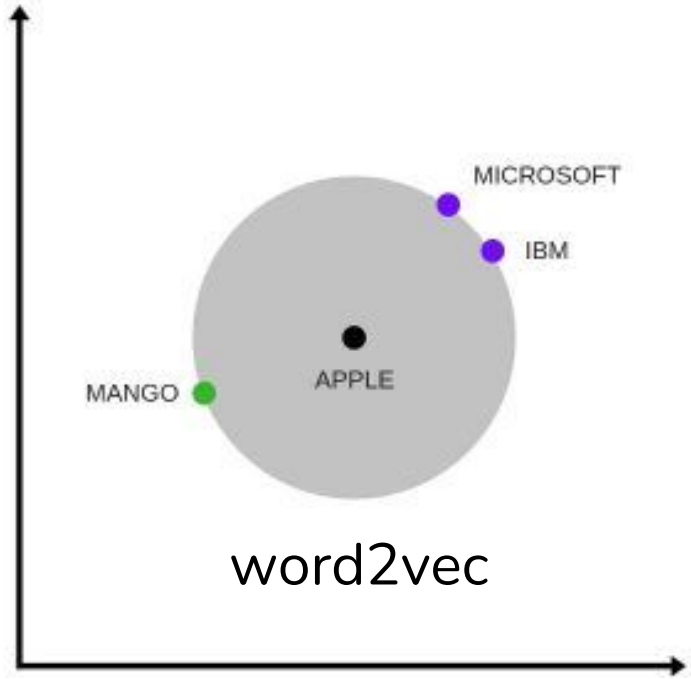


- **TF-IDF Vector**
  - Similar to Count vector, but frequency is calculated with respect to all documents
- **Co-Occurrence Vector**
  - Based on frequency of words appearing together (for example, it is)

Prediction based word embedding

- **Word2vec** based on neural networks
  - **Continuous Bag of words (CBOW)**: predicts the probability of a word given a context
  - **Skip-Gram model:** predicts the context given a word

https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/



word2vec