Big data science
Day 1

F. Legger - INFN Torino

Course plan

- **Mon-Thu: Theory + hands-on**
  - Each day you will receive a jupyter notebook
- **Friday: 2-hour hands-on and final test**
  - Final notebook will be a summary of all hands-on sessions

Today

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

Next

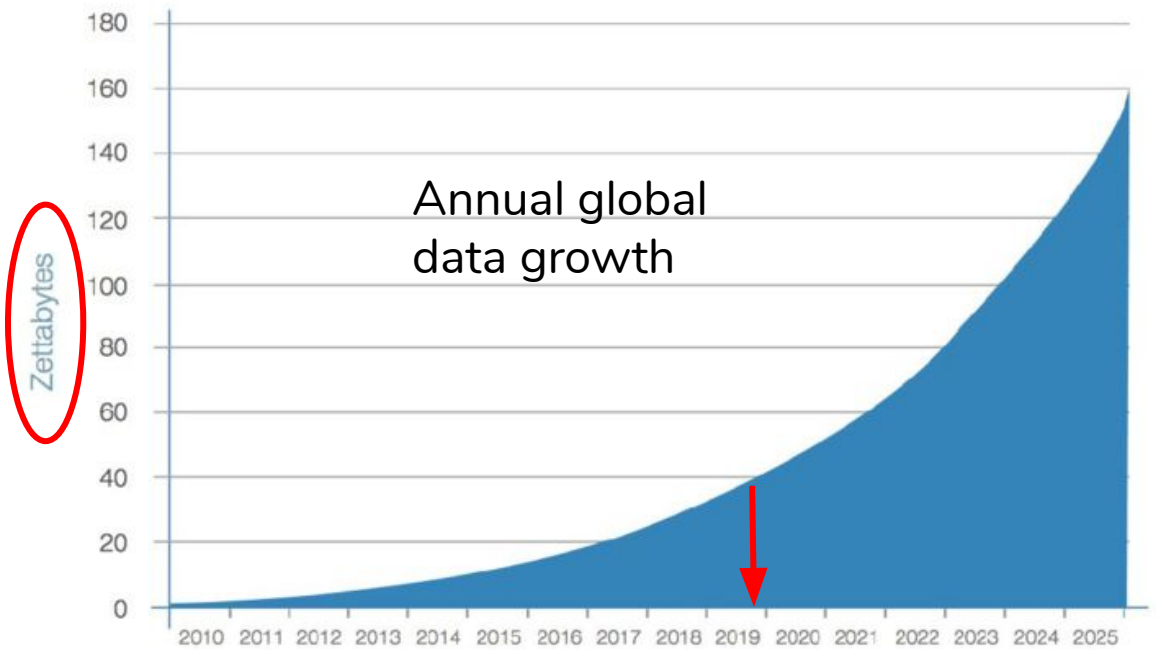
- Machine learning and deep learning
- Parallelisation
- Heterogeneous architectures

Hands-on

Input dataset for ML

What is big data?

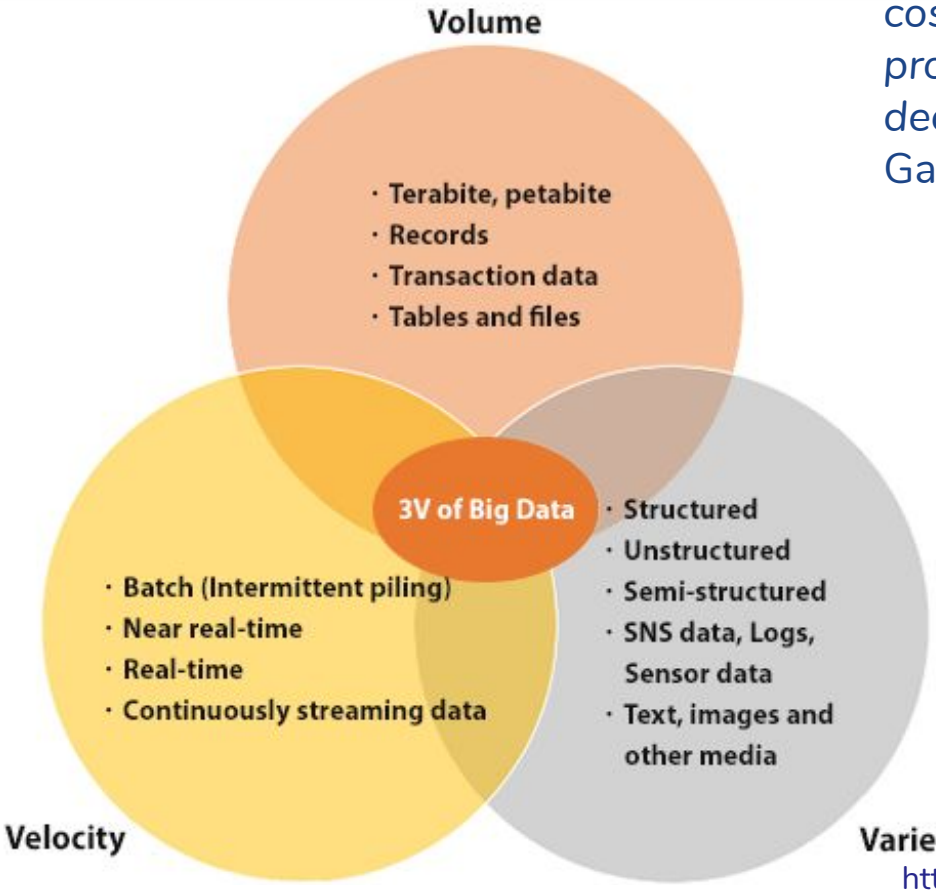
- Data that is too big to be analysed traditionally



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

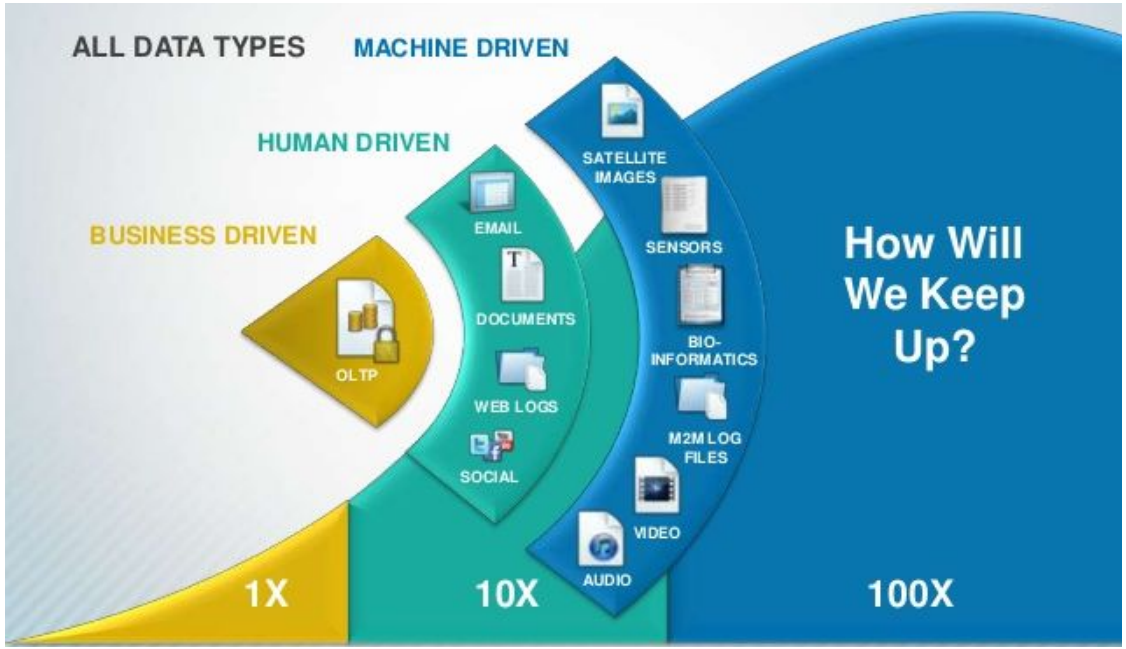


"Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."
Gartner (2012)

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

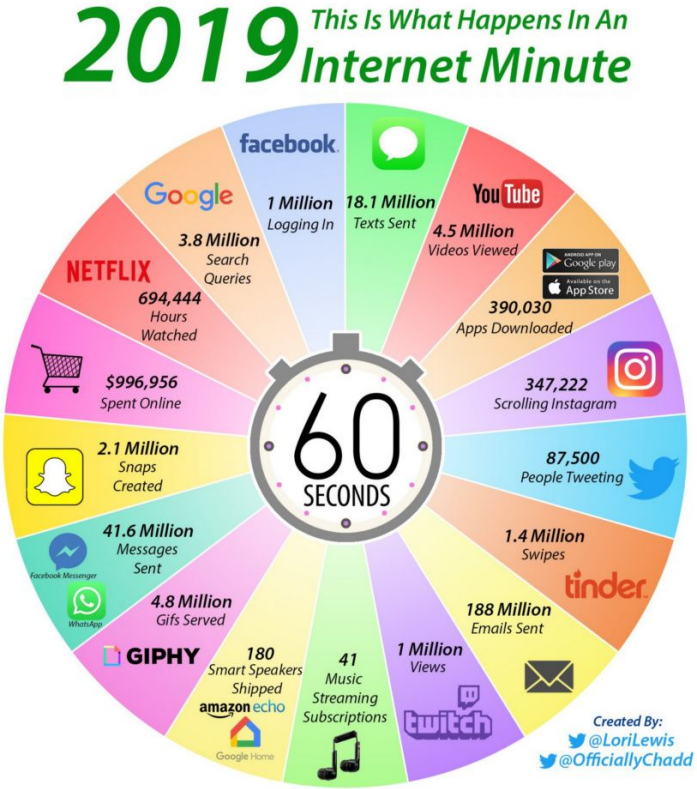
https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx

Big data sources: Business



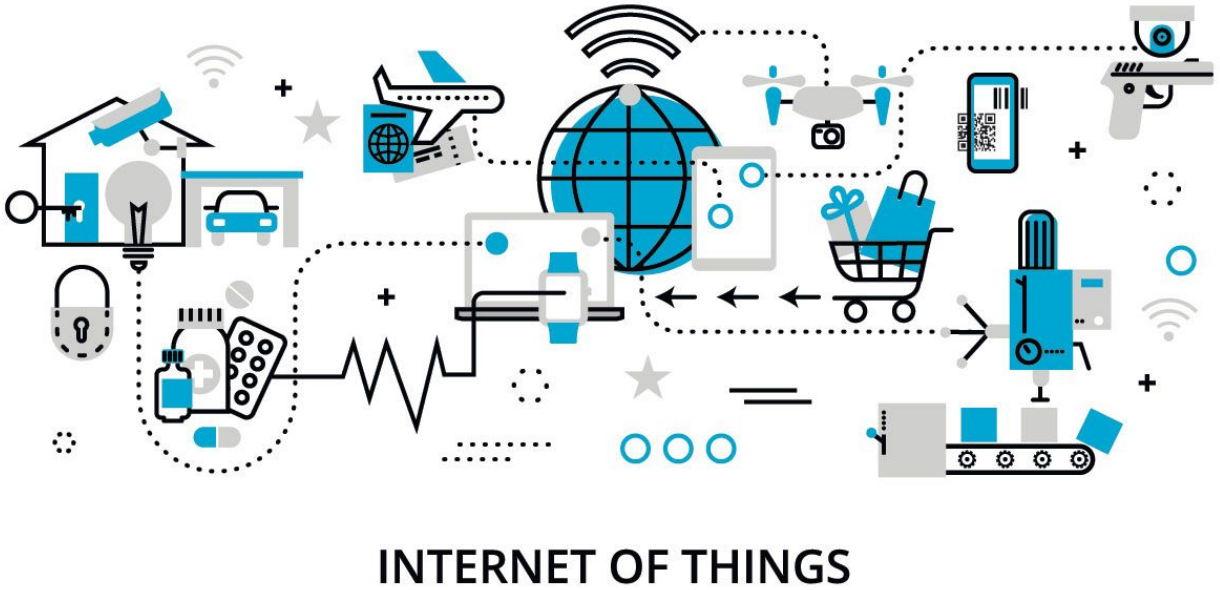
- **Traditional Business Systems**
  - Commercial transactions
  - Banking/stock records
  - E-commerce
  - Credit cards
  - Medical records

Big data sources: Human (x10)



- **Social Networks**
  - Twitter and Facebook
  - Blogs and comments
  - Pictures: Instagram, Flickr, Picasa, etc.
  - Videos: YouTube
  - Internet searches
  - Mobile data content (text messages)
  - User-generated maps
  - E-Mail

Big data sources: machine (x100)

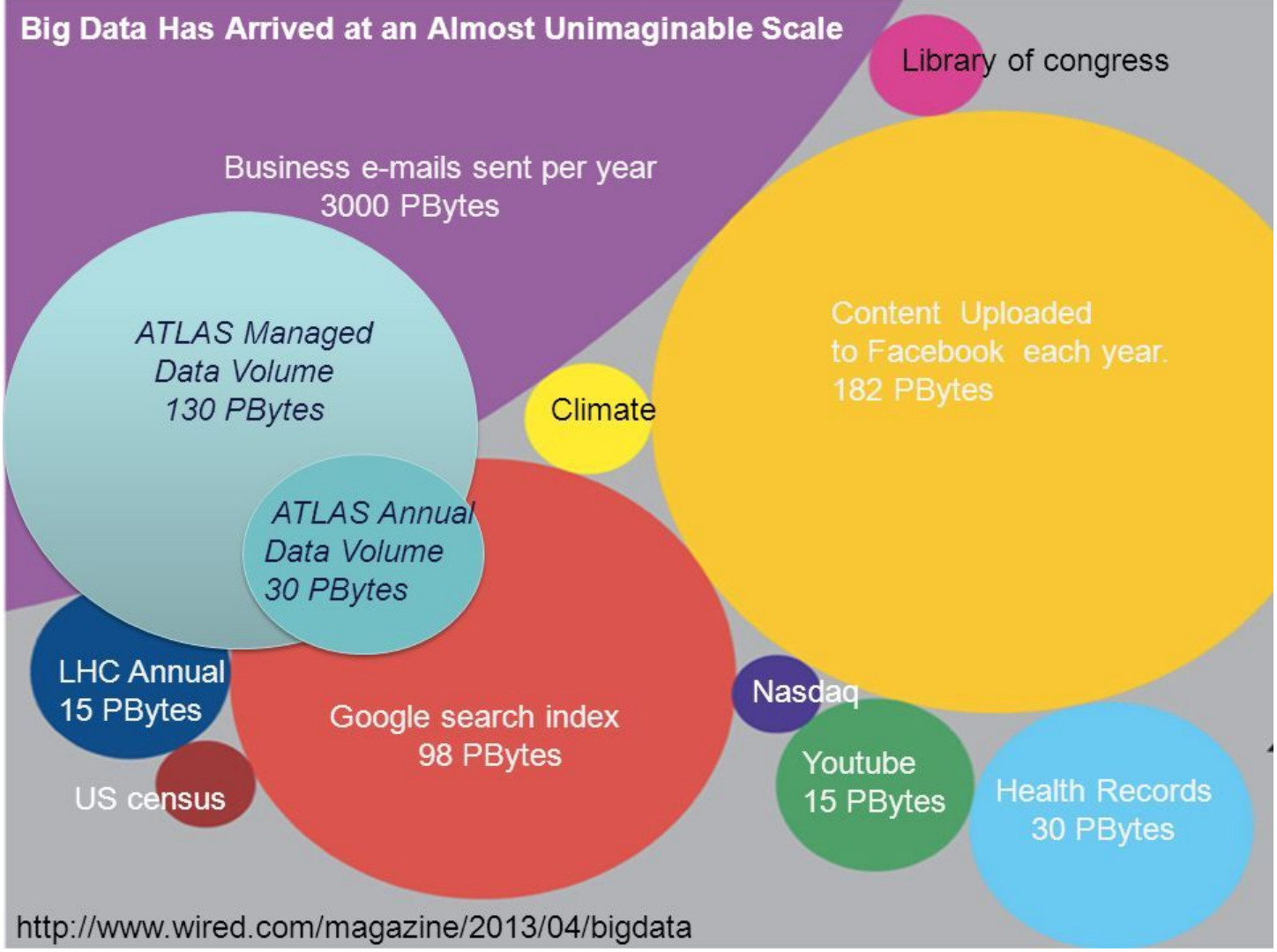


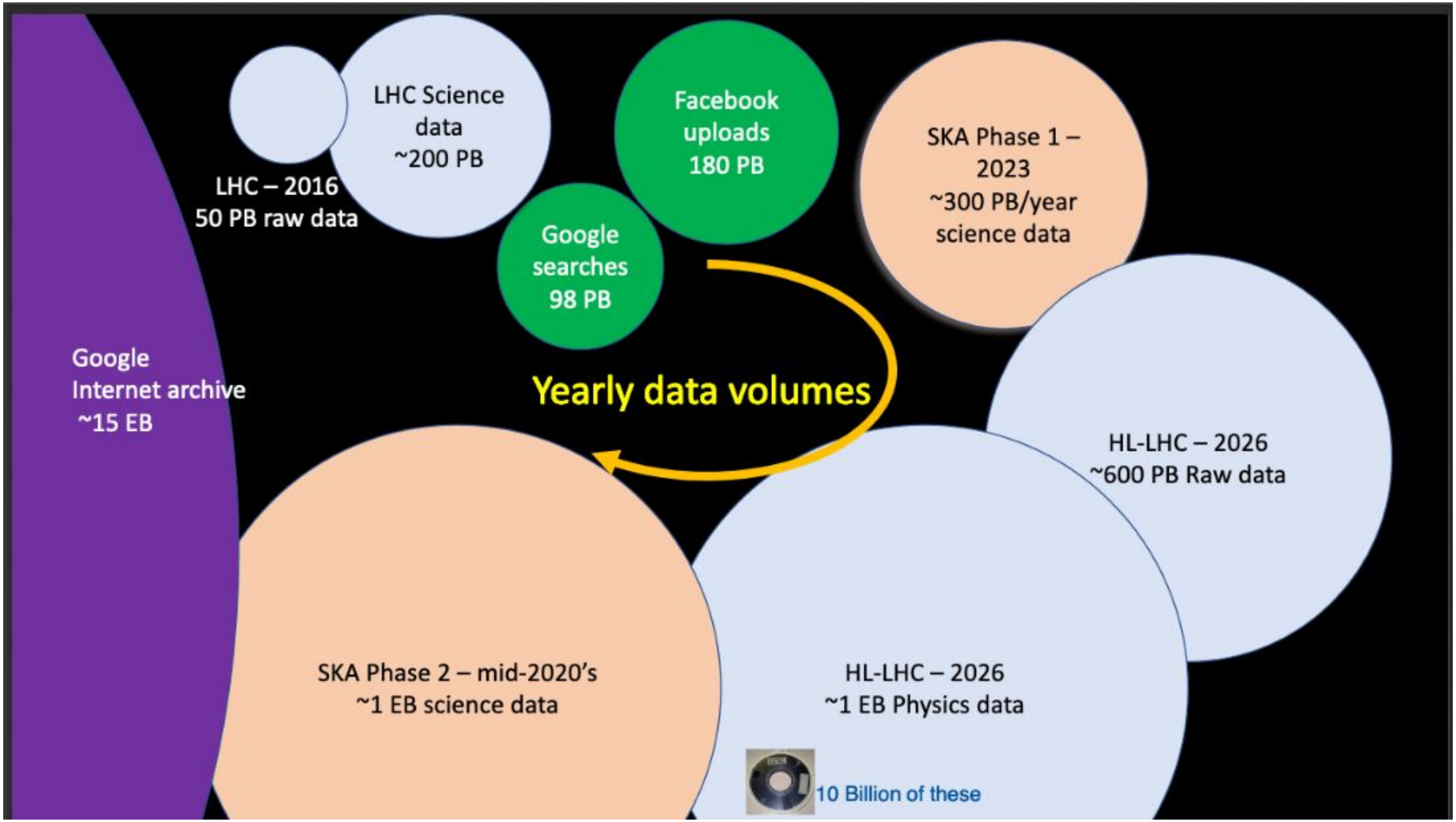
- **Internet of Things (IoT)**
  - Sensors: traffic, weather, mobile phone location, etc.
  - Security, surveillance videos, and images
  - Satellite images
  - Data from computer systems (logs, web logs)

Big data applications

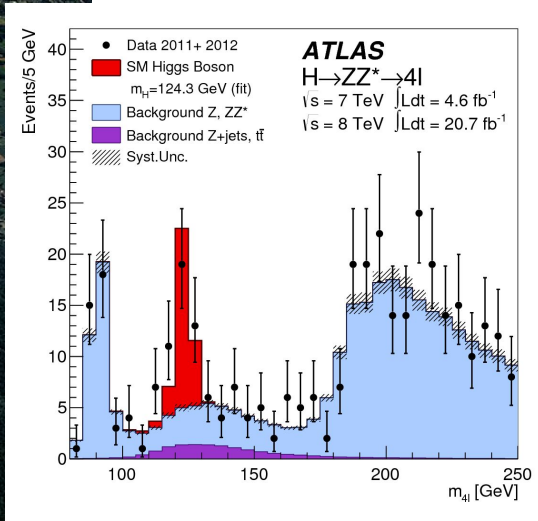
- **Entertainment:** Netflix and Amazon use Big Data to make shows and movie recommendations to their users.
- **Insurance:** Uses Big data to predict illness, accidents and price their products accordingly.
- **Driver-less Cars**: Google's driver-less cars collect about one gigabyte of data per second.
- **Automobile:** Rolls Royce has embraced Big Data by fitting hundreds of sensors into its engines and propulsion systems, which record every tiny detail about their operation. The changes in data in real-time are reported to engineers who will decide the best course of action such as scheduling maintenance or dispatching engineering teams should the problem require it.
- **Government:** A very interesting use of Big Data is in the field of politics to analyse patterns and influence election results (Cambridge Analytica Ltd.)

Size of big data (in 2013!!!)




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

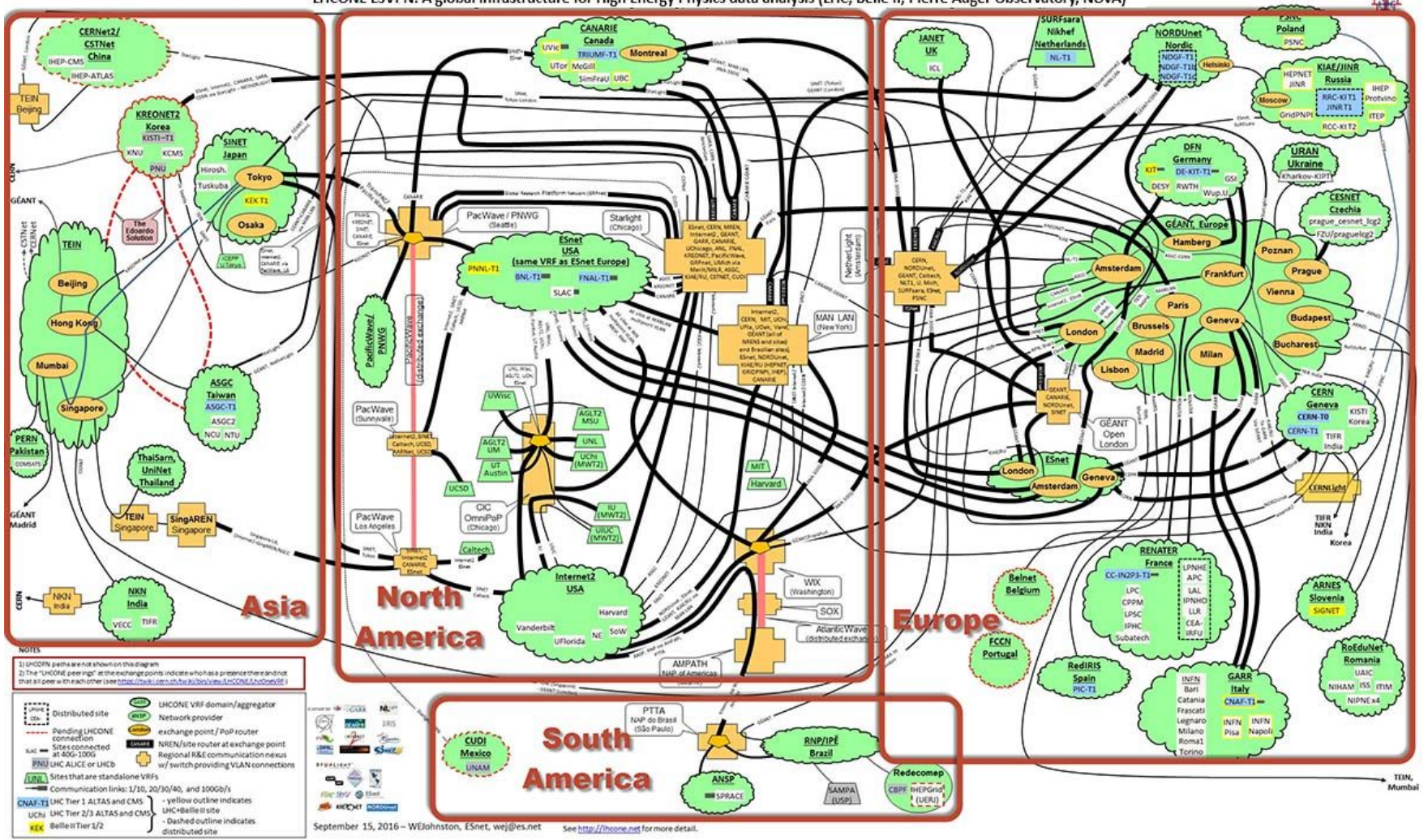
- pp (or Pb-Pb) collisions
- 4 experiments (ATLAS, CMS, LHCb, ALICE)
- Discovery of Higgs boson
- Nobel prize for physics 2013



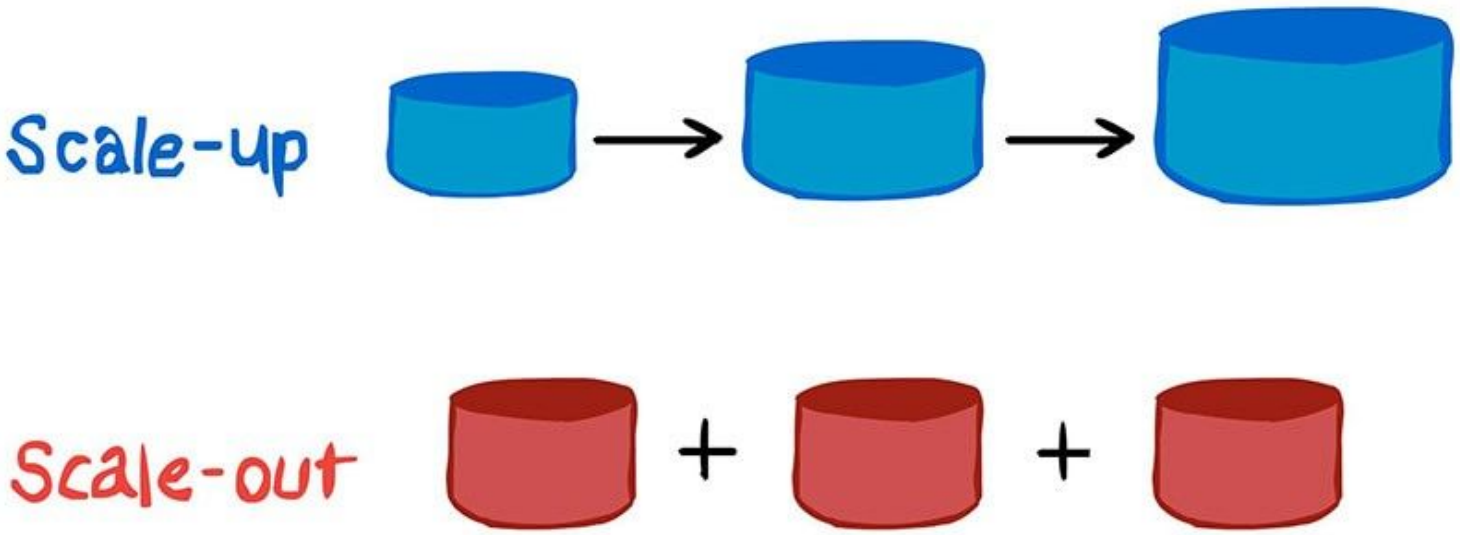
Big data @LHC

- 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)**, consisting of 200 computing sites located in more than 40 countries and holding over 600 PB (600,000TB) in more than 1 billion files.

- 200 sites
- >40 countries
- 750000 cores
- 2 million jobs/day
- 600 PB storage
- 10-100 GB links


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
Tuskuba
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
GÉANT Madrid
TEIN Singapore
SingAREN Singapore
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
UFlorida
NE
SoW
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-T1c
NDGF-T1d
Helsinki
PSNC Poland
KIAE/JINR Russia
HEPNET JINR
RRC-KI T1
JINR T1
IHEP Protvino
ITEP
Moscow
GridPNPI
RCC-KI T2
URAN Ukraine
Kharkov-KIPT
DFN Germany
KIT
DE-KIT-T1
DESY
RWTH
GSI
Wup.U.
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
RENATER France
CC-IN2P3-T1
LPNHE
APC
LAL
IPNHO
LLR
CEA-IRFU
LPC
CPPM
LPSC
IPHC
Subatech
Belnet Belgium
FCCN Portugal
RediRIS Spain
PIC-T1
ARNES Slovenia
SIGNET
RoEduNet Romania
UAIC
NIHAM
ISS
ITIM
NIPNE x4
GARR Italy
CNAF-T1
INFN Bari
Catania
Frascati
Legnaro
Milano
Roma1
Torino
INFN Pisa
INFN Napoli
TEIN, Mumbai
CUDI Mexico
UNAM
PTTA NAP do Brasil (São Paulo)
RNP/IPÊ Brazil
ANSP
SPRACE
SAMPA (USP)
Redecomep
CBPF
RHEPGrid (UERJ)
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.

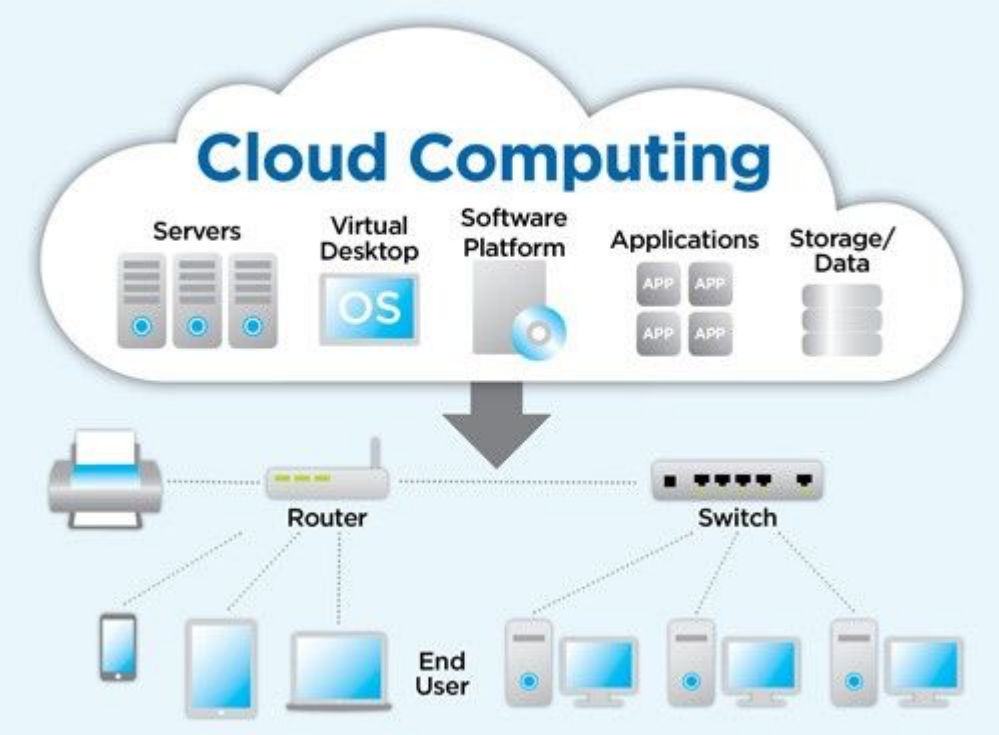
Scale up vs scale out



HPC, super computers

Grid, cloud

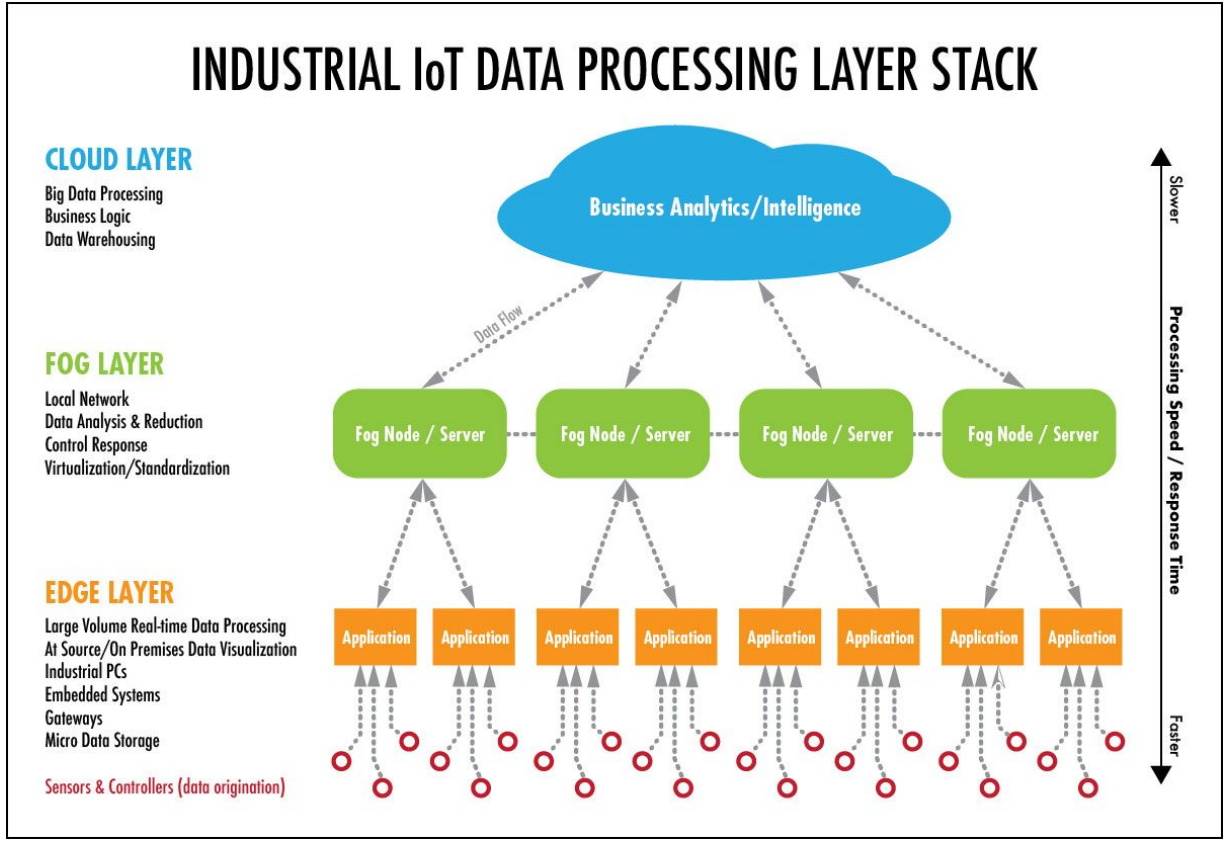
Cloud computing

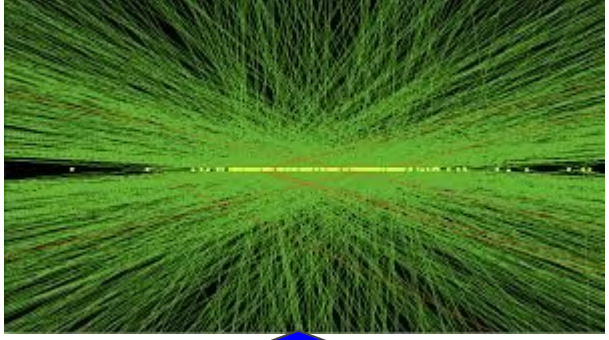


On-demand availability of computer system resources, especially data storage and computing power, without direct active management by the user

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

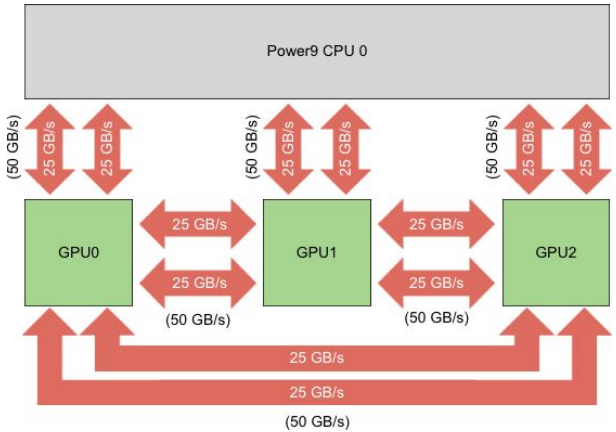
Edge and fog computing





High Performance Computing (HPC)

- Typically involves supercomputers
- Summit (2018), #1 HPC at Oakridge
  - hybrid architecture
  - 4608 nodes connected in a non-blocking fat-tree using a dual-rail Mellanox EDR InfiniBand interconnect.
  - each node contains
    - multiple IBM POWER9 CPUs
    - multiple NVIDIA Volta GPUs
    - Connection through NVIDIA's high-speed NVLink.
    - half a terabyte of coherent memory (high bandwidth memory + DDR4) addressable by all CPUs and GPUs
    - 800GB of non-volatile RAM that can be used as a burst buffer or as extended memory

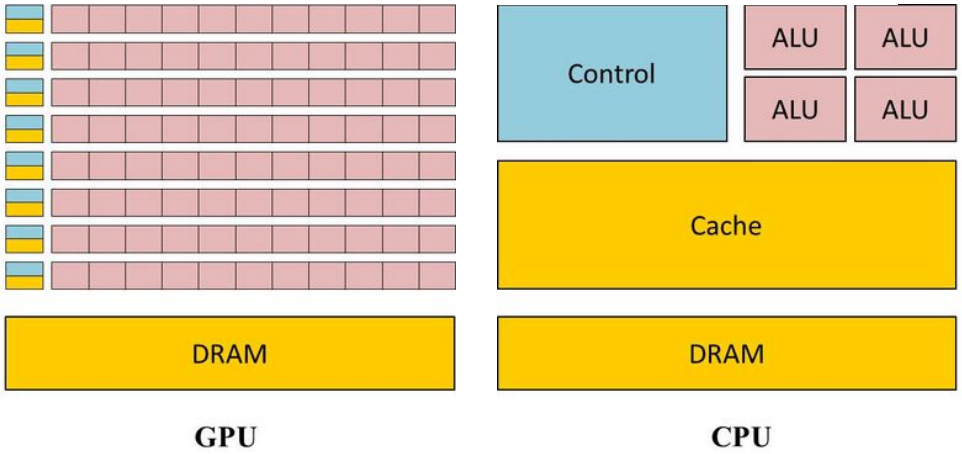


https://www.top500.org/

Heterogeneous architectures

GPU vs CPU

GPU	CPU
• hundreds of simpler cores	• few very complex cores
• thousand of concurrent hardware threads	• single-thread performance optimization
• maximize floating-point throughput	• transistor space dedicated to complex ILP
• most die surface for integer and fp units	• few die surface for integer and fp units



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

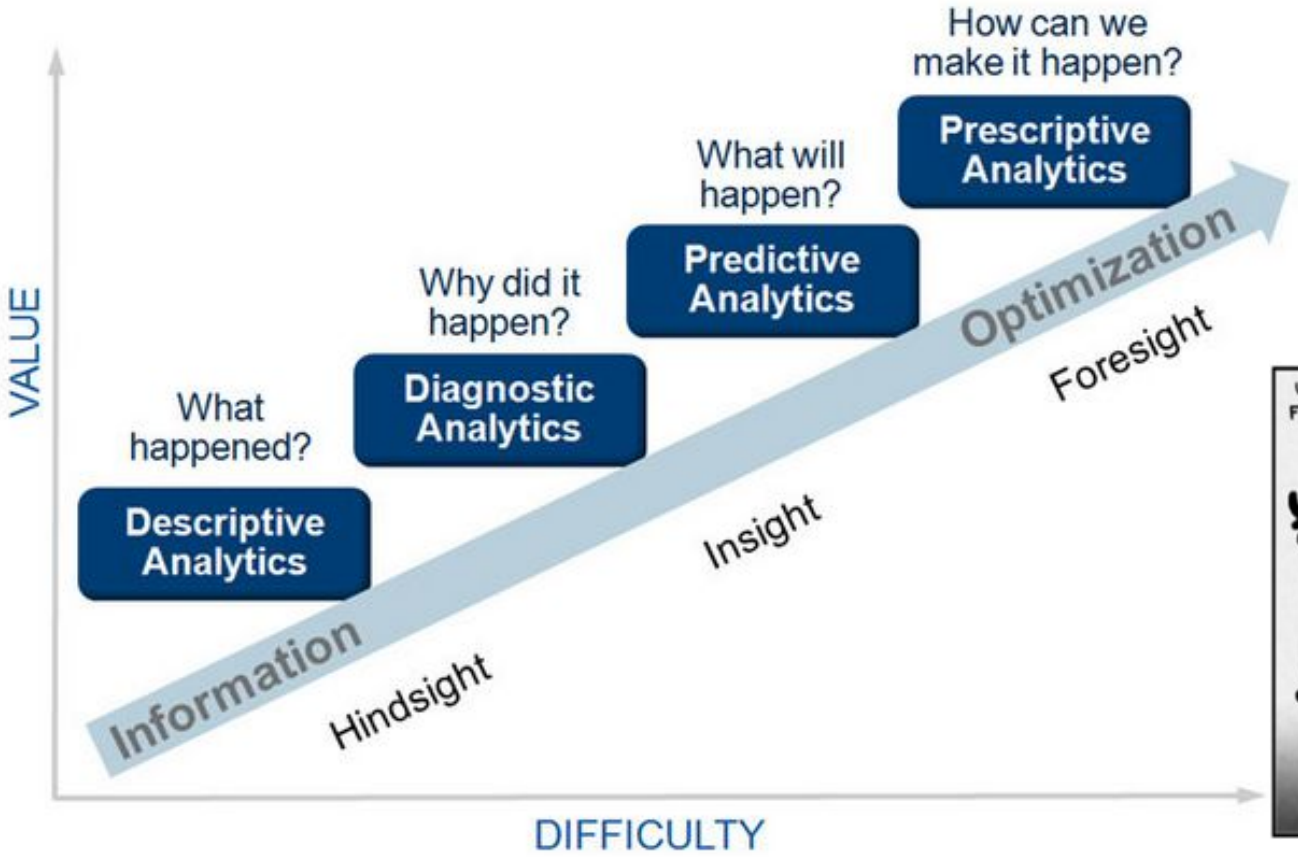
GPU
- Medium-to-large models, datasets
- Image, video processing
- Application on CUDA or OpenCL

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

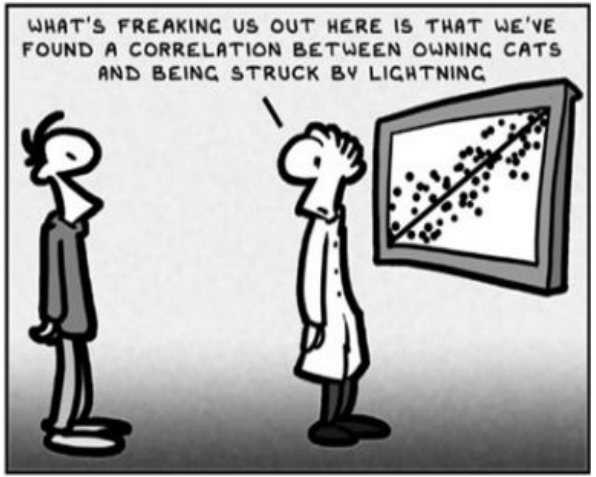
FPGA
- Large datasets, models
- Compute intensive applications
- High performance, high perf./cost ratio

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

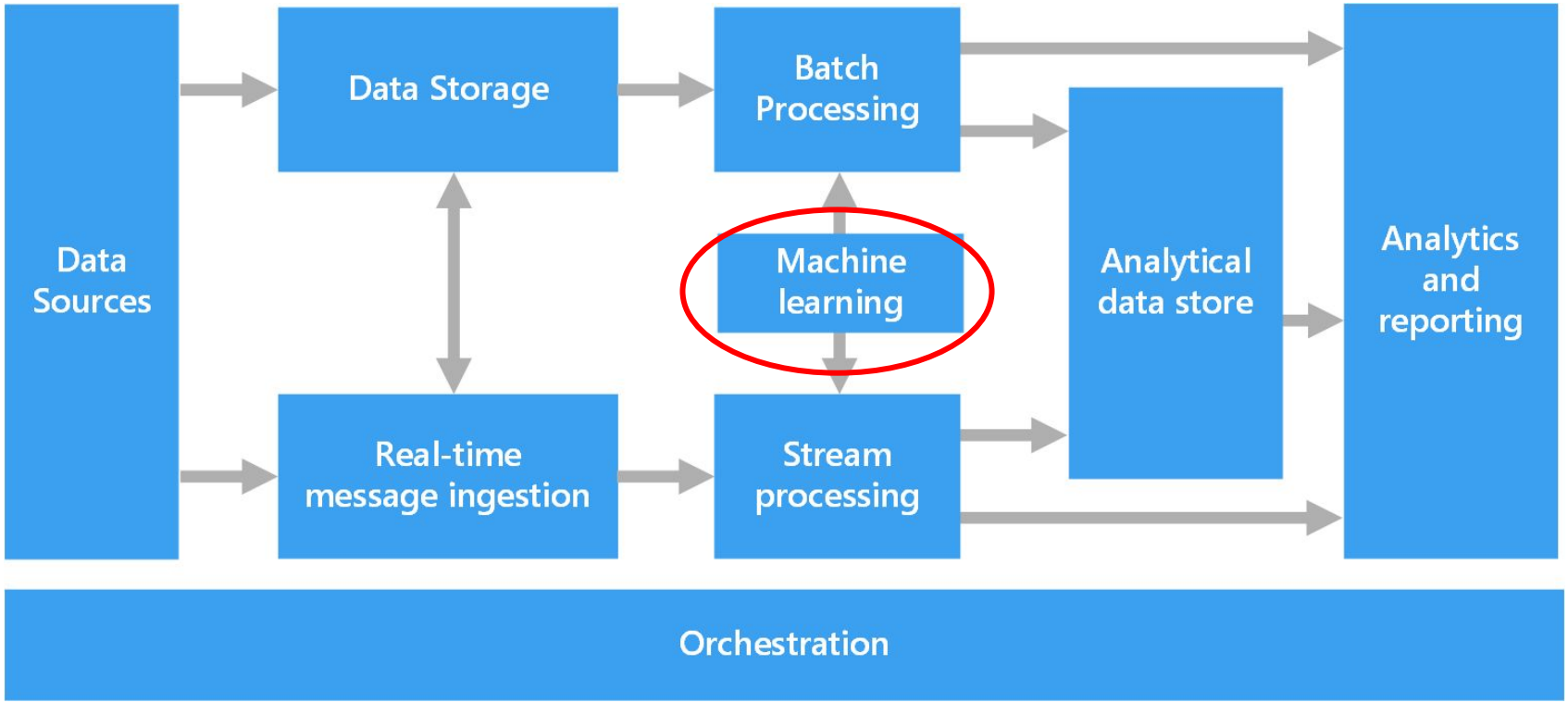

SHOULD WE ALWAYS IGNORE WHAT THE DATA SAYS, OR IS THIS MORE OF A ONE-TIME THING?


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

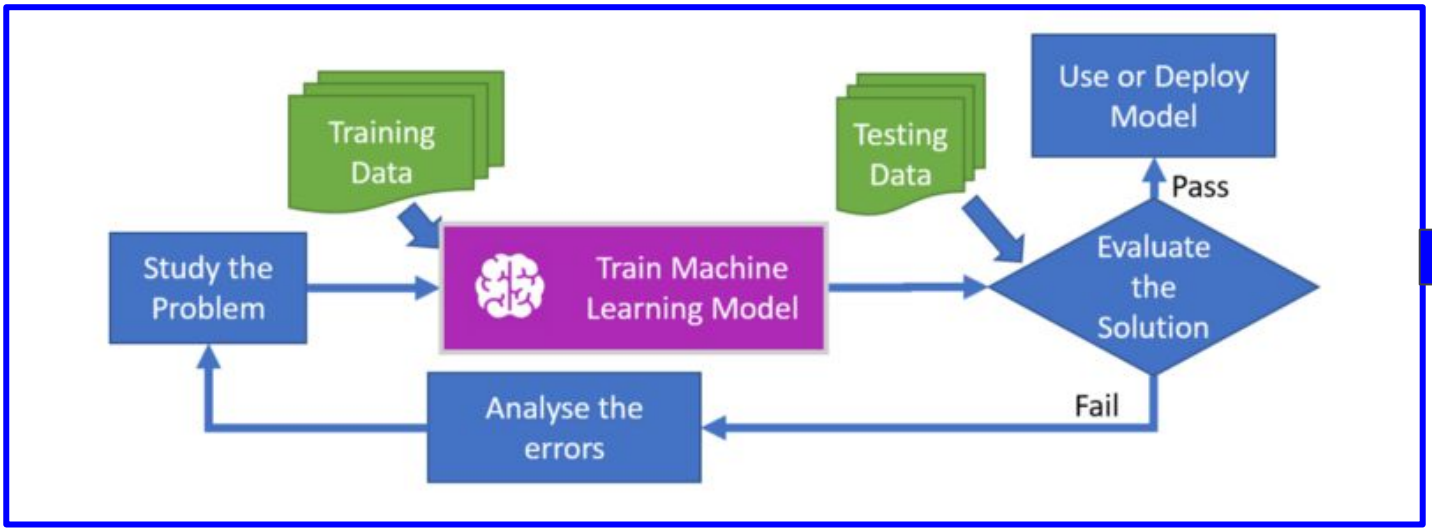
Analytics applications

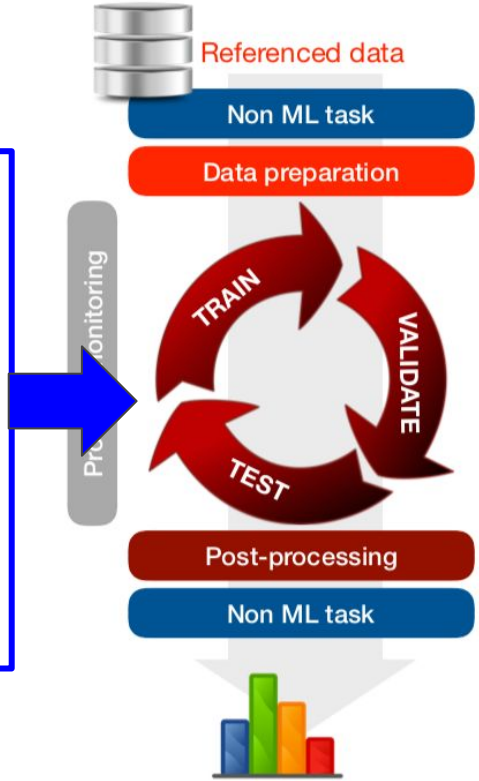
- Segmentation
  - Credit score, health insurance
- Churn prediction
  - Customers switching from one company to another
- Recommendation system
  - Netflix reported that 2/3 of the movies watched are recommended
  - Google News stated that recommendations generate 38% more click-through
  - Amazon claimed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

The big data pipeline

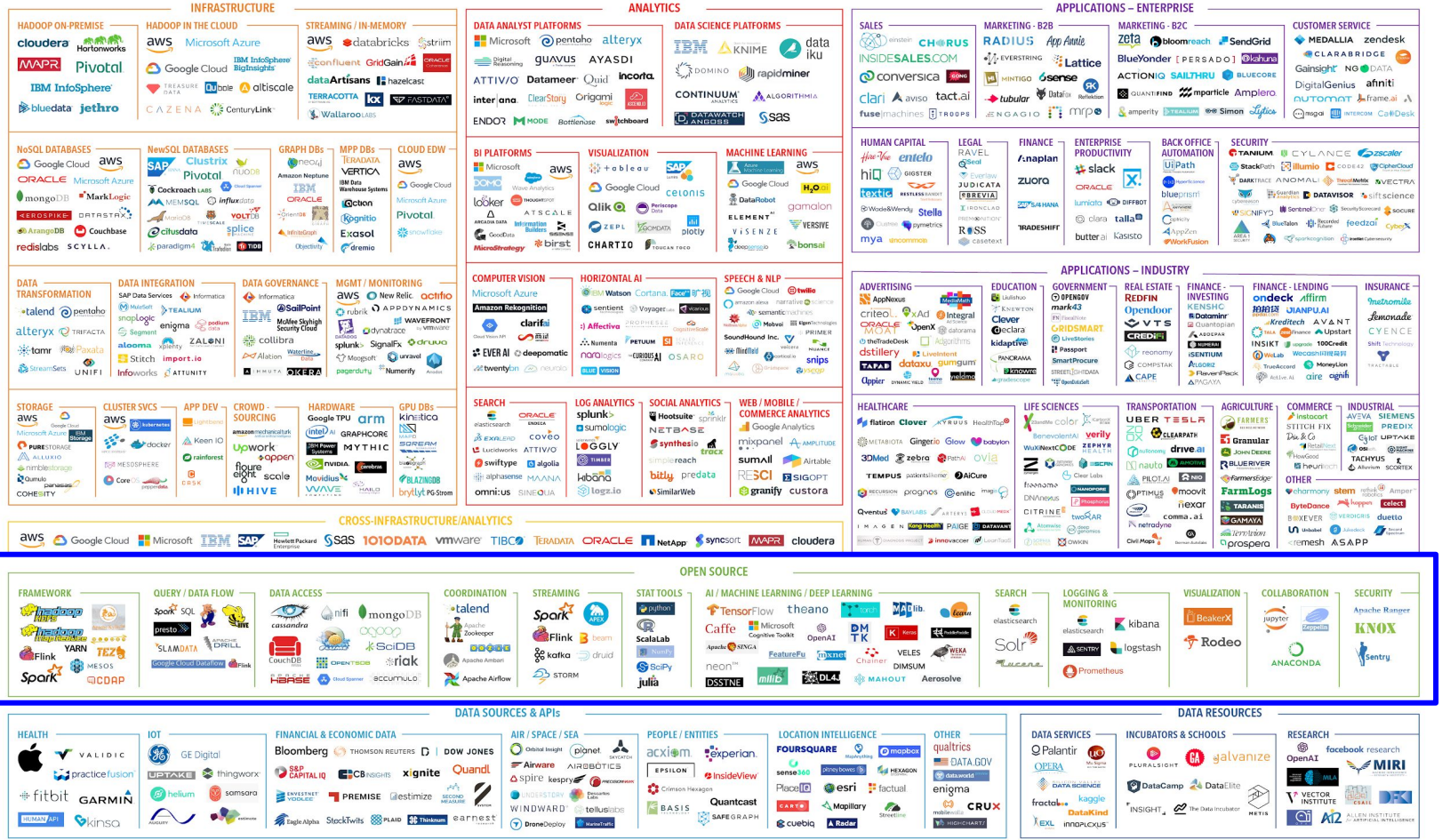


The machine learning pipeline

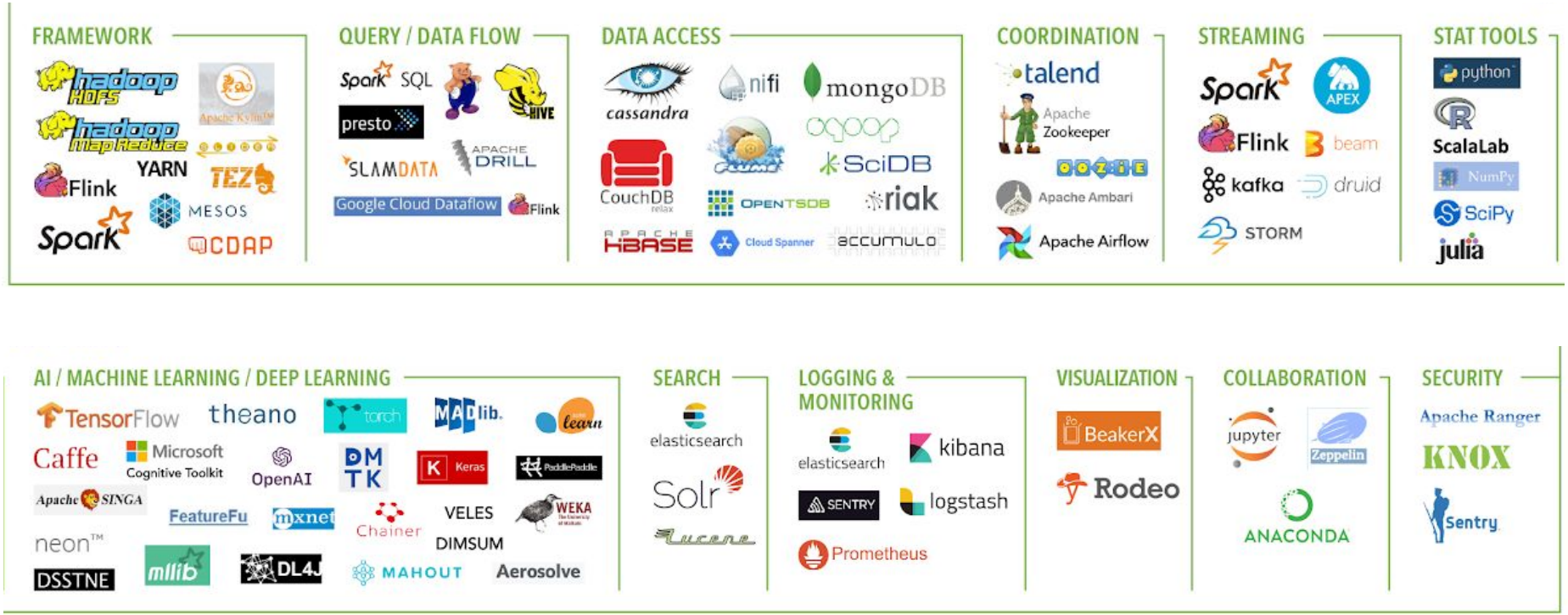


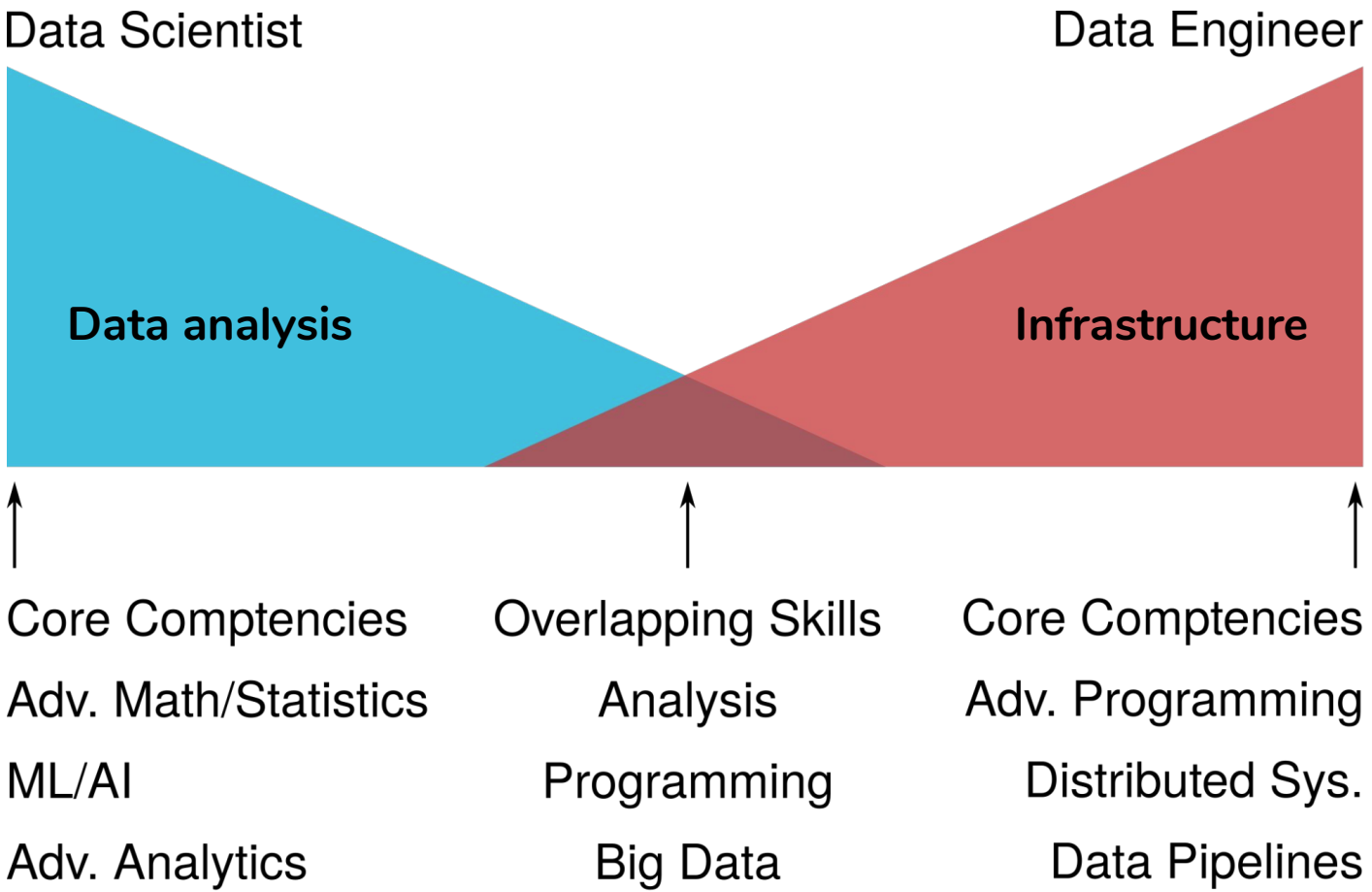


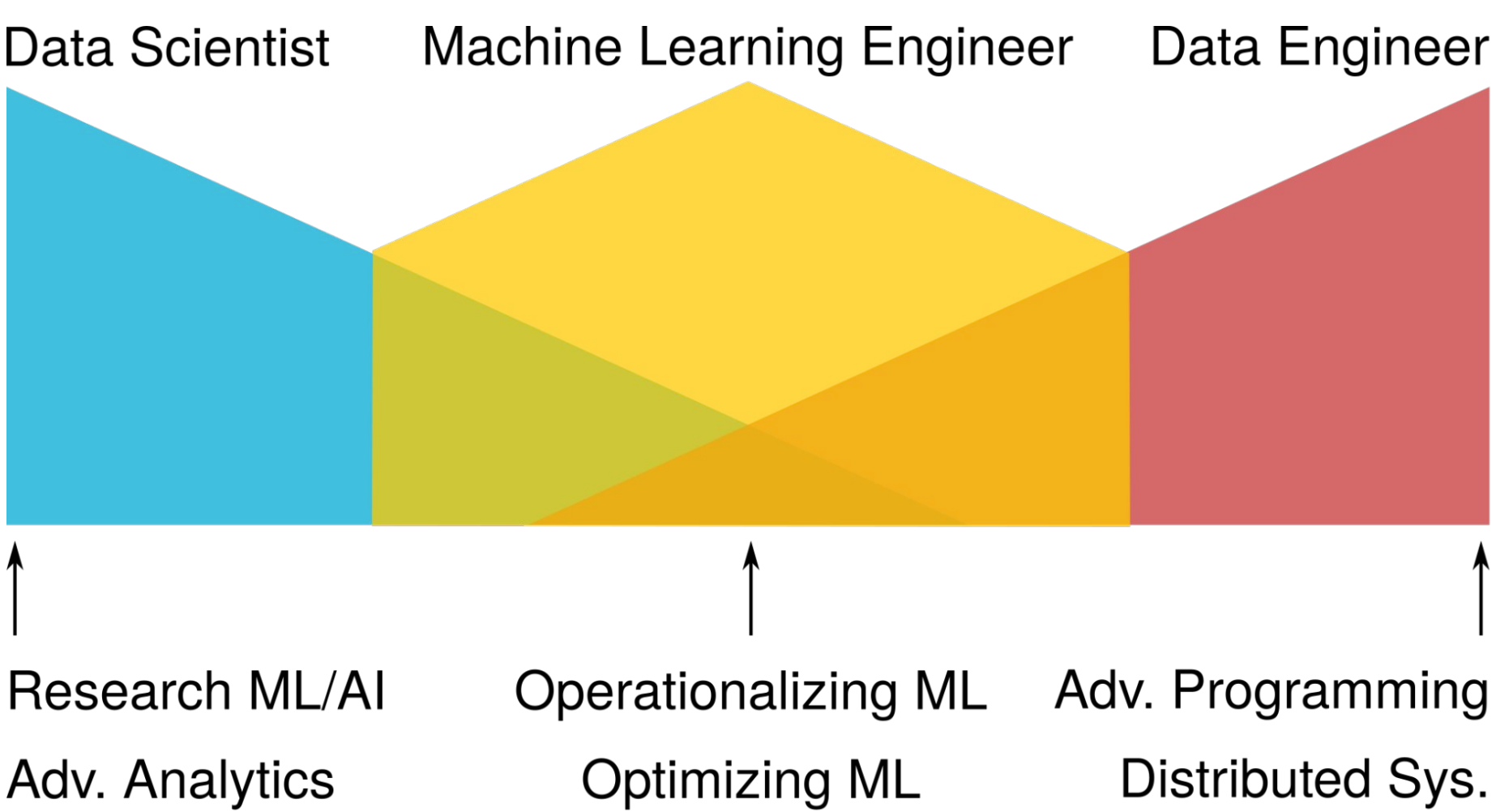
BIG DATA & AI LANDSCAPE 2018



Open Source

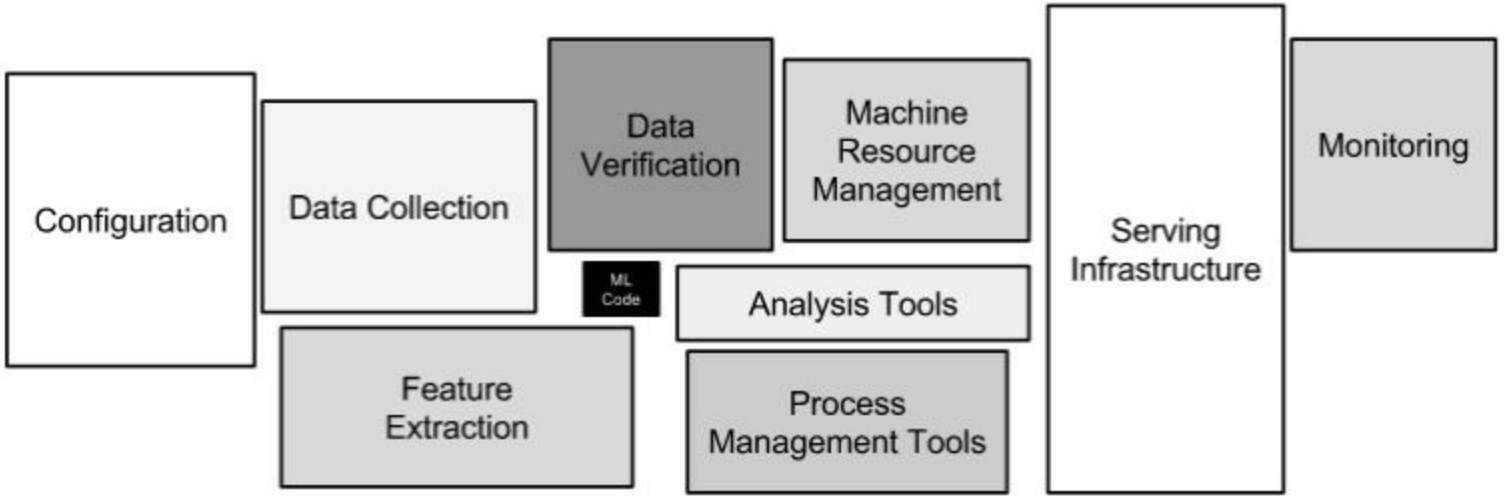

FRAMEWORK
hadoop HDFS
hadoop MapReduce
YARN
TEZ
Flink
MESOS
Spark
CDAP
QUERY / DATA FLOW
Spark SQL
HIVE
presto
SLAMDATA
APACHE DRILL
Google Cloud Dataflow
Flink
DATA ACCESS
cassandra
nifi
mongoDB
SciDB
CouchDB
OPENTSDB
riak
APACHE HBASE
Cloud Spanner
accumulo
COORDINATION
talend
Apache Zookeeper
OOZIE
Apache Ambari
Apache Airflow
STREAMING
Spark
APEX
Flink
beam
kafka
druid
STORM
STAT TOOLS
python
ScalaLab
NumPy
SciPy
julia
AI / MACHINE LEARNING / DEEP LEARNING
TensorFlow
theano
torch
MADlib
learn
Caffe
Microsoft Cognitive Toolkit
OpenAI
DMTK
Keras
PaddlePaddle
Apache SINGA
FeatureFu
mxnet
Chainer
VELES
WEKA
neon™
DIMSUM
DSSTNE
mllib
DL4J
MAHOUT
Aerosolve
SEARCH
elasticsearch
Solr
lucene
LOGGING & MONITORING
elasticsearch
kibana
SENTRY
logstash
Prometheus
VISUALIZATION
BeakerX
Rodeo
COLLABORATION
jupyter
Zeppelin
ANACONDA
SECURITY
Apache Ranger
KNOX
Sentry


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

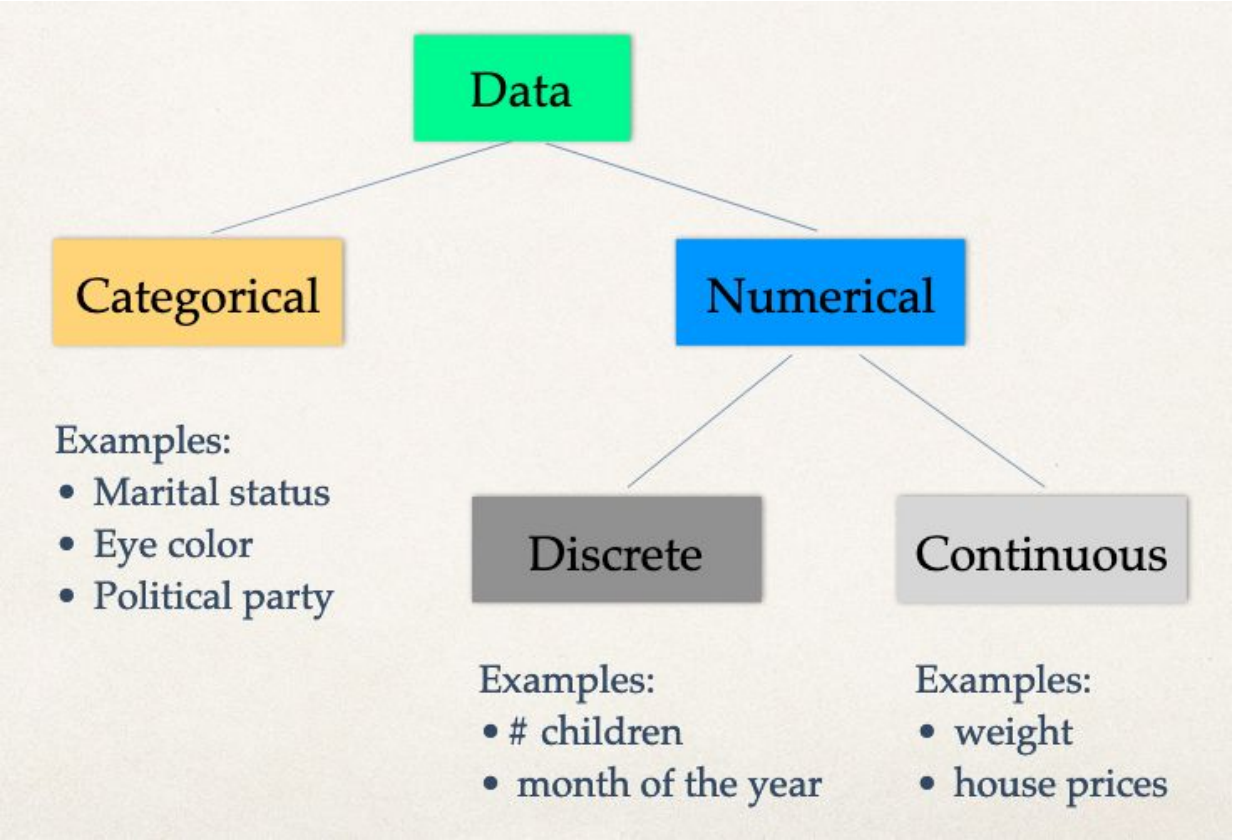
Where is



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

Data types

Typically strings



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, imputing, bucketing, binning, feature interactions
  - dimensionality reduction

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

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

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

- Your programming skills will be required here: R, Python, Databases, etc

Data transformation

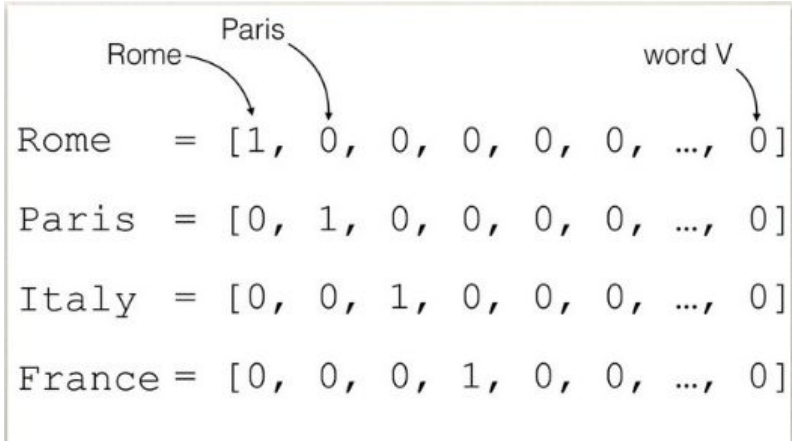
- Data transformation and *aggregation*: log, sum of values, average, ...
- **Scaling**: a technique to scale data to a given range [0,1] or any other range
- **Normalization/Standardization:** a technique to scale data to mean with zero and and unit-variance
- **Augmentation:** a technique to create 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

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

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

One-hot encoding

- It is a 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 (said to be “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



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're trying to predict
- Encode UserID:
  - Train dataset:
    - Take mean of the values of Y for all rows with same UserID except the one you want to encode
    - Add (multiply) random noise
  - Test dataset
    - No Y
    - Just use frequency of UserID

Mean of [1,1,0]

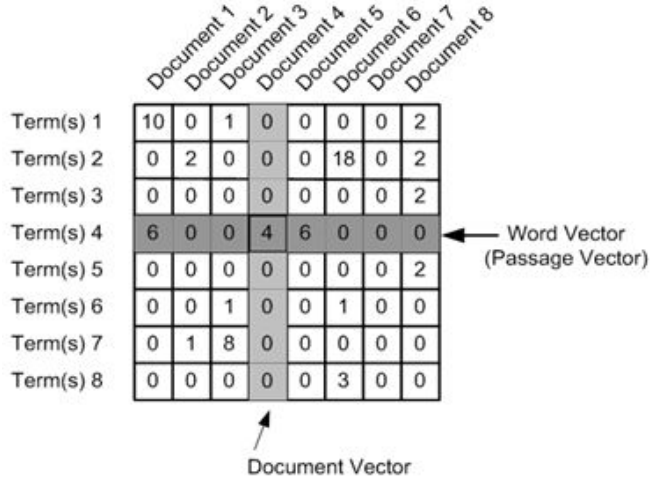
mean_y*random

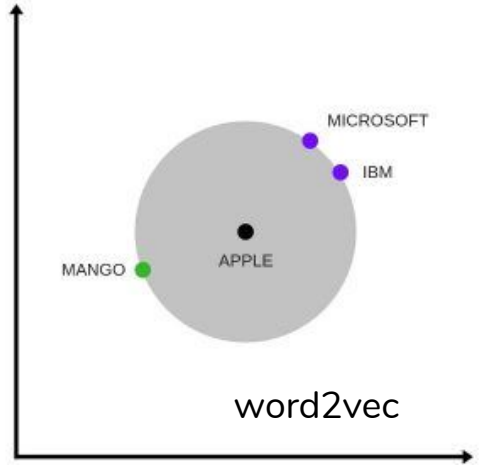
Split	UserID	Y	mean_y	random	newID
Train	A1	~~0~~	0.667	1.05	0.70035
Train	A1	1	0.333	0.97	0.32301
Train	A1	1	0.333	0.98	0.32634
Train	A1	0	0.667	1.02	0.68034
Test	A1	-	0.5	1	0.5
Test	A1	-	0.5	1	0.5
Train	A2	0			

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;
  - e.g. Sunday and Saturday may have similar effect while other days may be treated independently
  - all of these features are hidden from original data representation
  - Use neural networks or other ML algorithms to train the model to find the best representation of embedded variables





Word embedding techniques

- Frequency based 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 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/

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