Introduction to Big Data

2024-2025

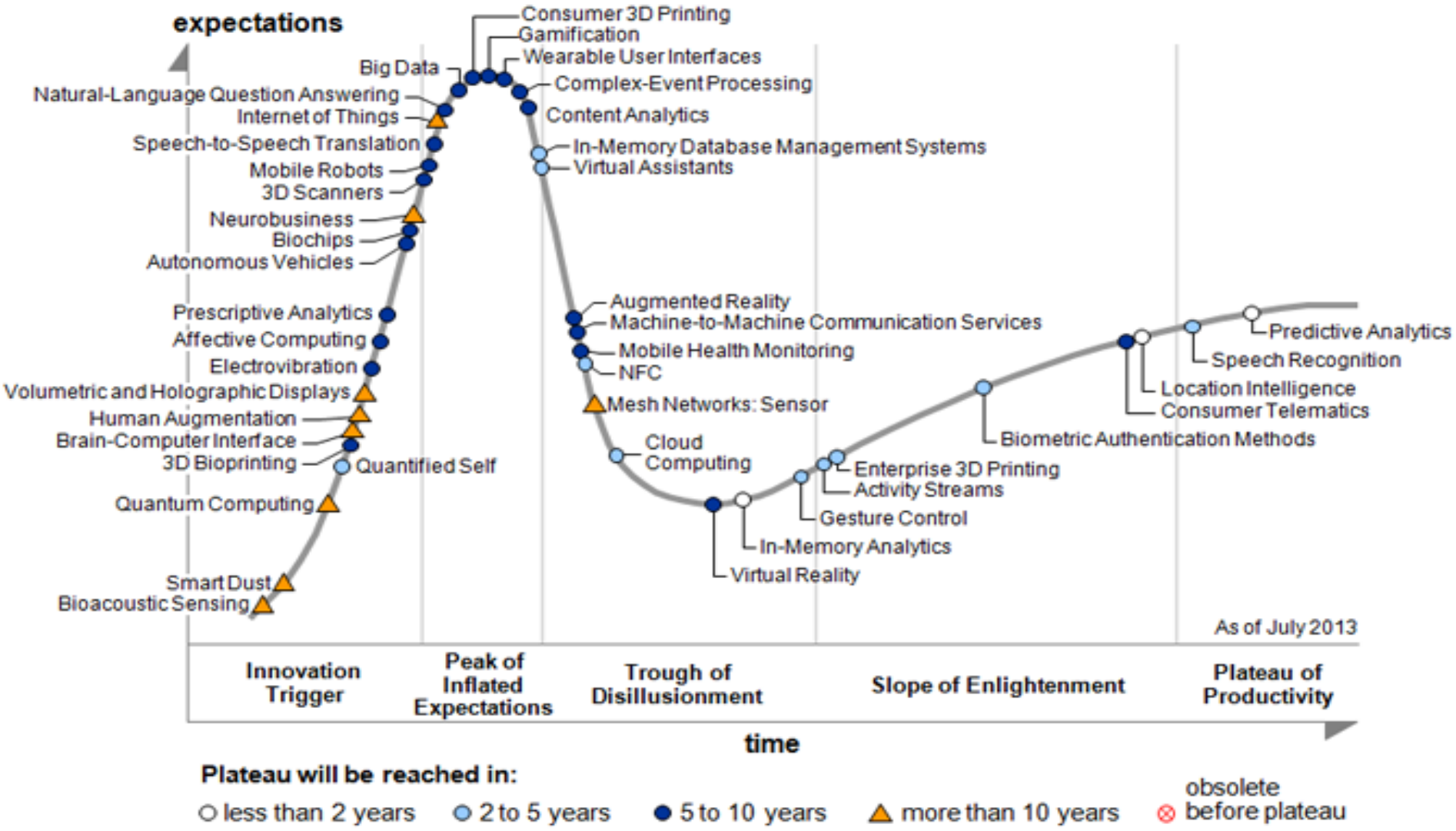
Prof.dr.ing. Florin Pop
florin.pop@cs.pub.ro
Dr.ing. Cătălin Negru
catalin.negru@cs.pub.ro
SL.Dr.ing. Bogdan Mocanu
bogdan_costel.mocanu@upb.ro

SCPD • SSA • ABD

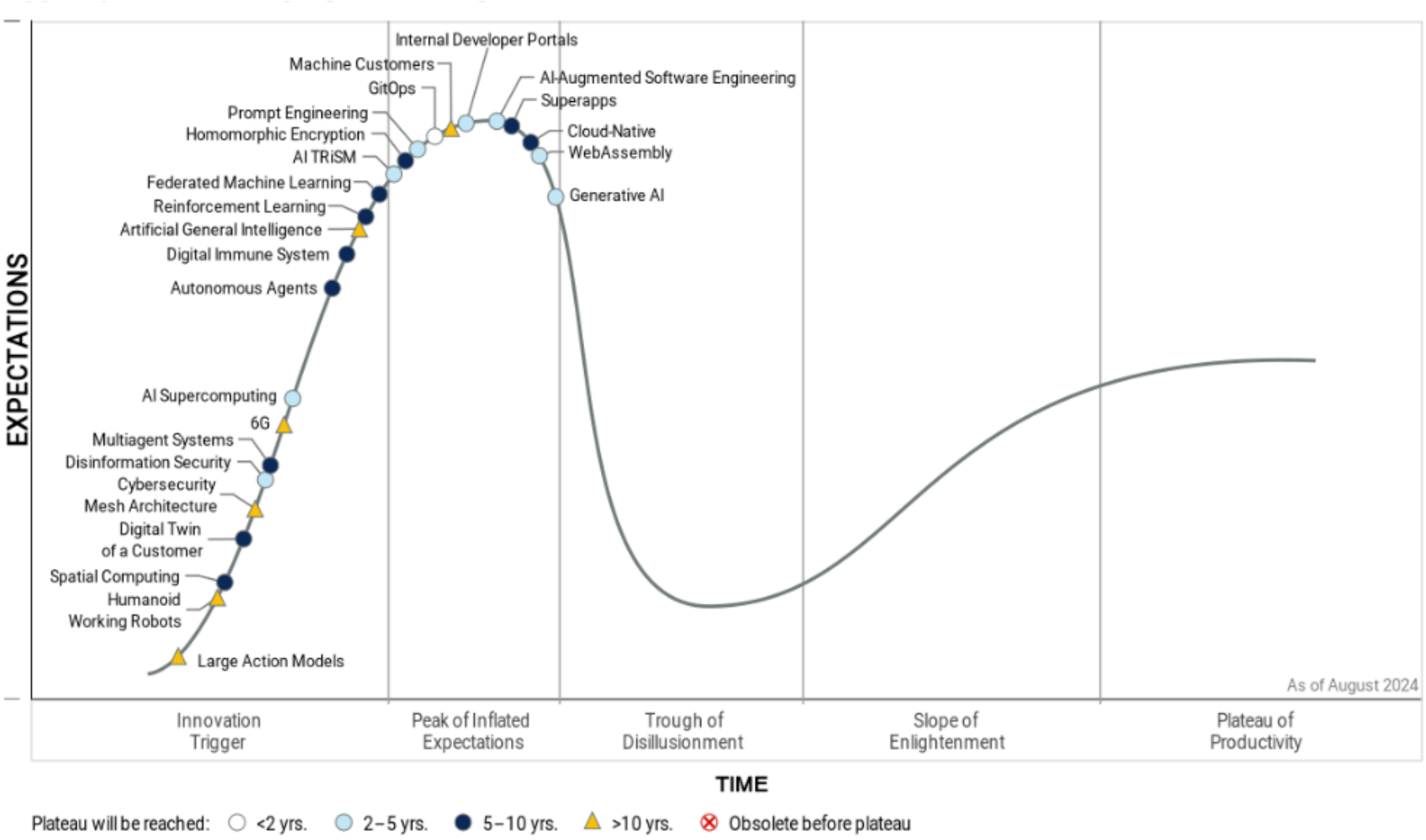
The Human Face of Big Data

- Big Data will impact every part of your life: https://www.youtube.com/watch?v=0Q3sRSUYmys
- What Exactly Is Big Data? https://www.forbes.com/video/4857597029001

Gartner Tempers The Expectations Of Big Data

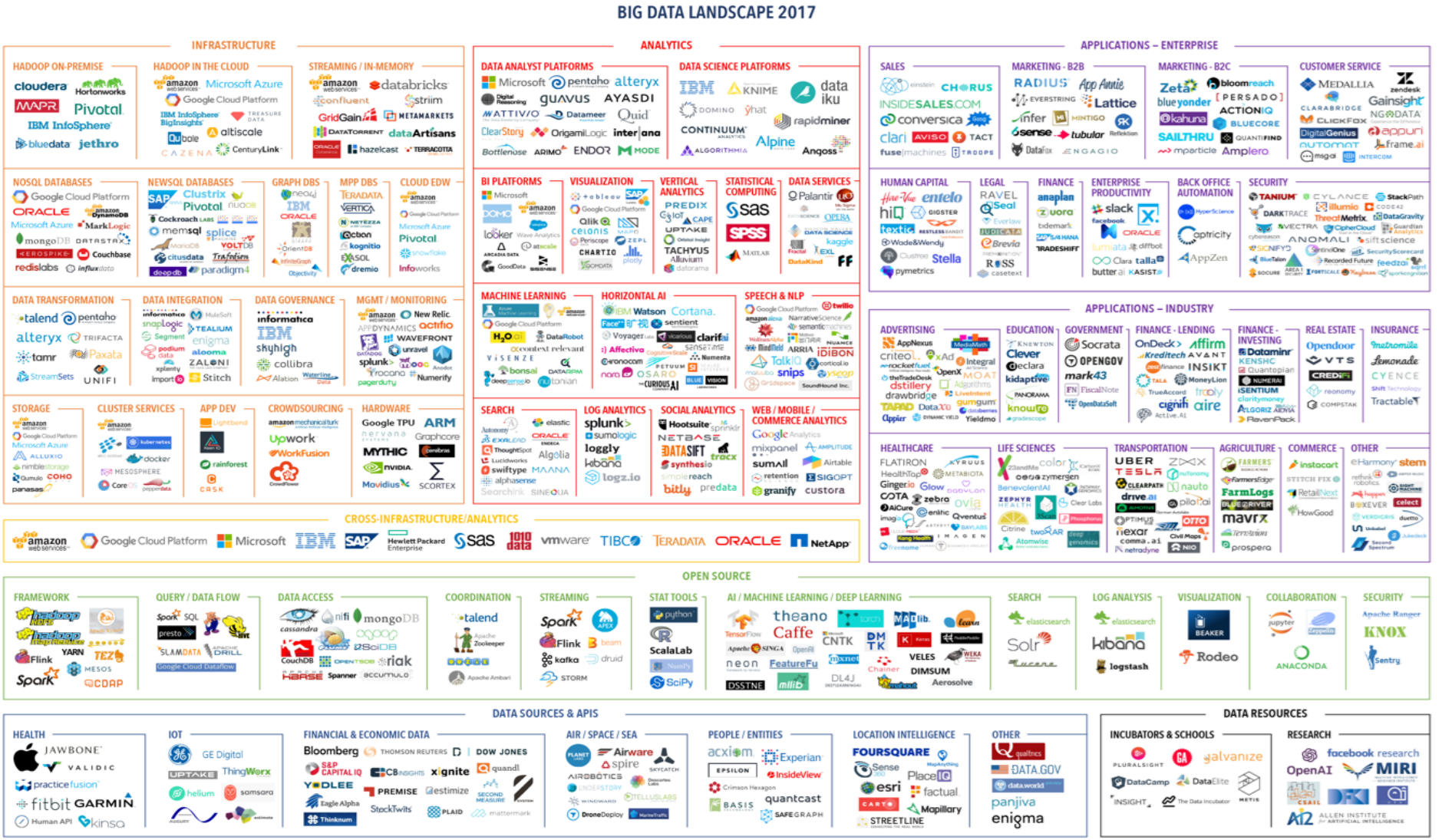


Gartner Tempers The Expectations Of Emerging Technologies



Gartner

The 2017 Big Data Landscape

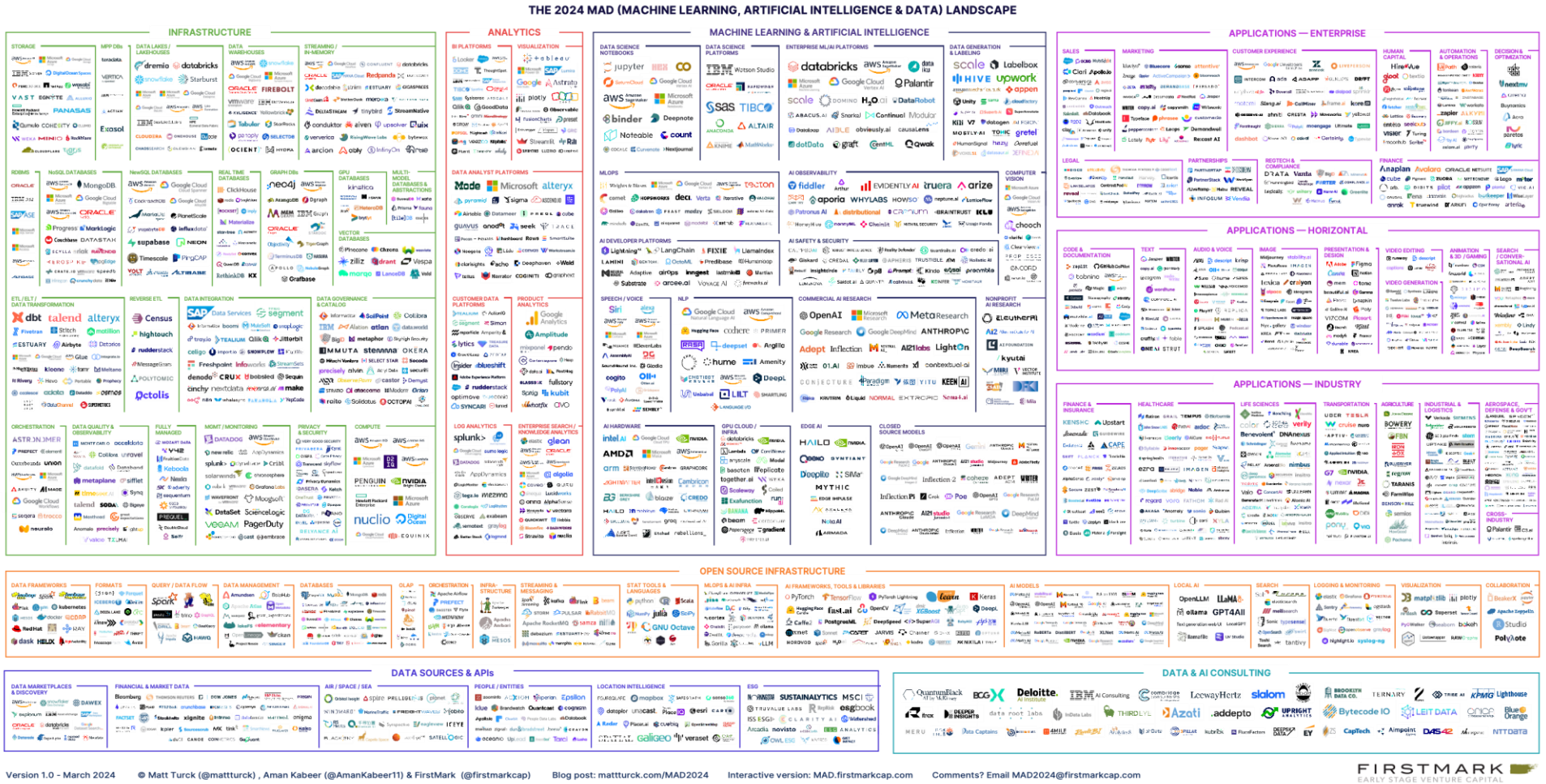


V2 – Last updated 5/3/2017

mattturck.com/bigdata2017

FIRSTMARK EARLY STAGE VENTURE CAPITAL

The 2024 MAD Landscape



Version 1.0 - March 2024 Blog post: mattturck.com/MAD2024 Interactive version: MAD.firstmarkcap.com Comments? Email MAD2024@firstmarkcap.com

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

AI Landscape in 2024: Training Costs, Open Source, and Running Out of Data

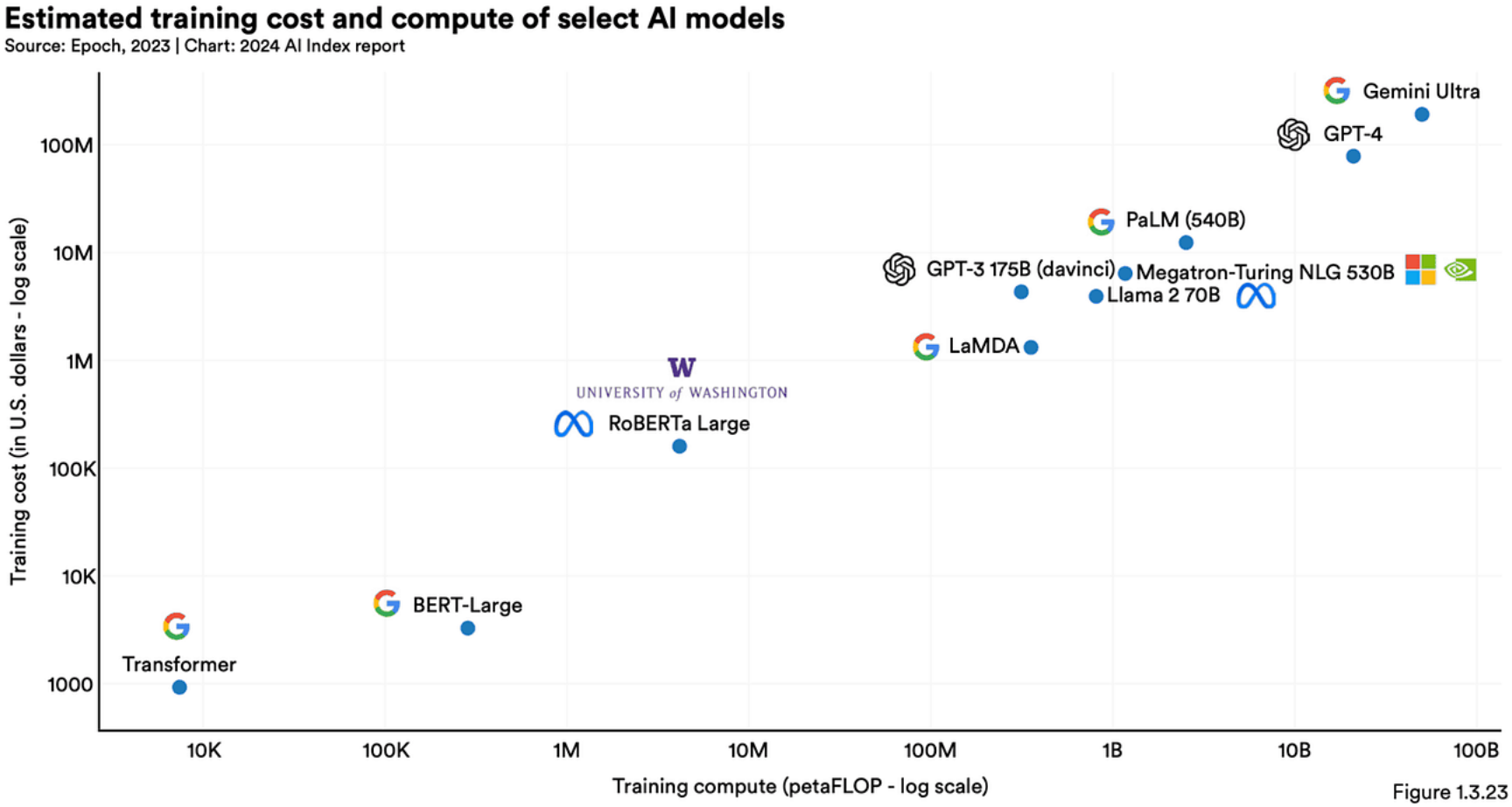


Figure 1.3.23

Understanding Big Data

Value	Symbol	Name
1024	KB	Kilobyte
1024^2	MB	Megabyte
1024^3	GB	Gigabyte
1024^4	TB	Terabyte
1024^5	PB	Petabyte
1024^6	EB	Exabyte
1024^7	ZB	Zettabyte
1024^8	YB	Yottabyte

WHAT HAPPENS EVERY MINUTE

via Internet Live Stats

6,123 TB
TRAFFIC PRODUCED BY USERS

84,000
INSTAGRAM PHOTOS UPLOADED

5,200,000
GOOGLE SEARCHES

305,000
SKYPE CALLS

185,000,000
E-MAILS SENT

BIG DATA GLOBAL MARKET SIZE FORECAST

by 2025 **$250* billion**
According to MarketsandMarkets, Adroit Market Research

by 2021 **$70* billion**
According to Wikibon, Frost & Sullivan, MarketsandMarkets, Statista

** - average assumption*

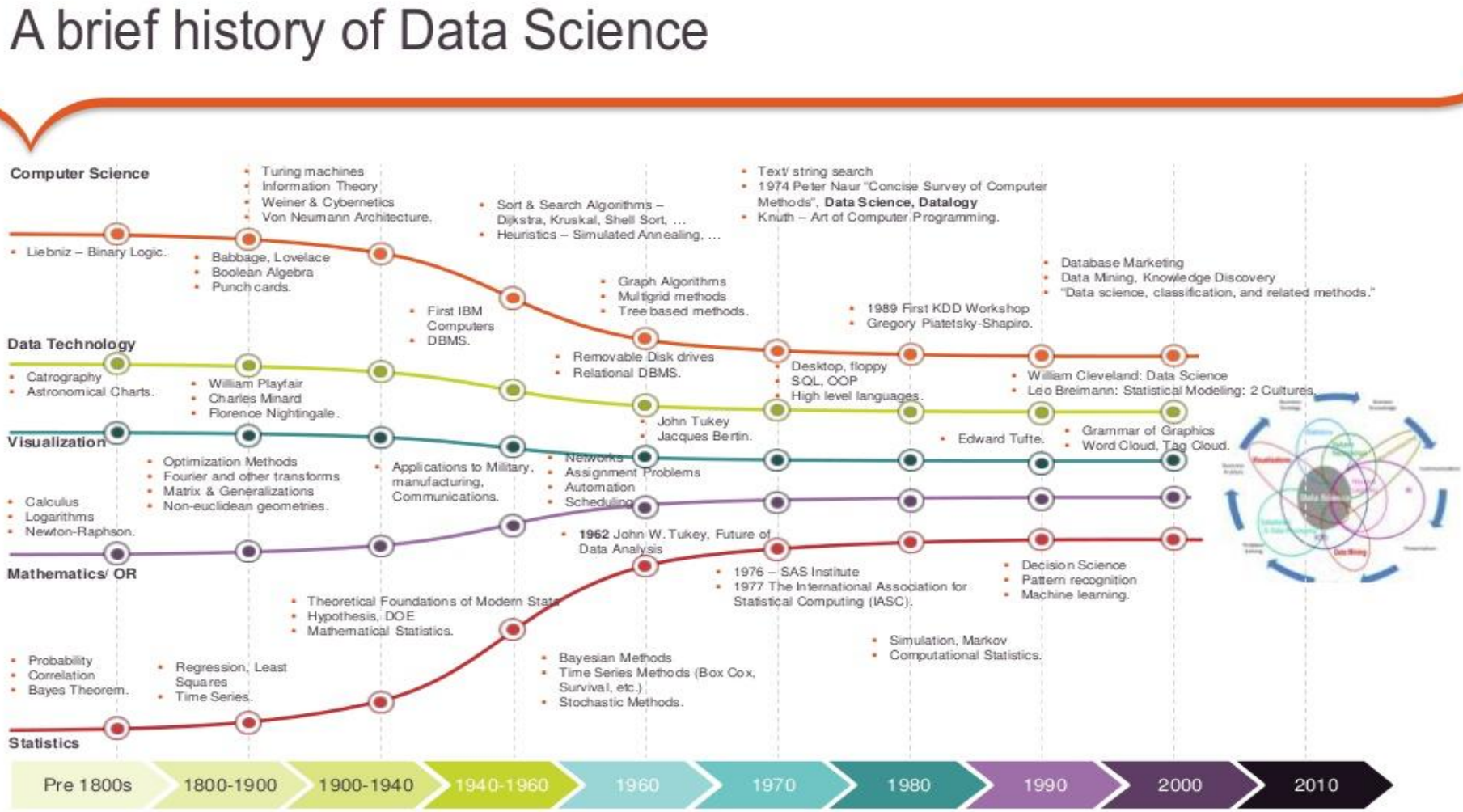
2.5 quintillion bytes of data each day

Dataset whose volume, velocity, variety and complexity are beyond the ability of commonly used tools to capture, process, store, manage and analyze them can be termed as ***BIG DATA****.*

Complex situations?

- **Data distribution**
  - The large data set is split into chunks or smaller blocks and distributed over N number of nodes or machines.
  - Distributed File System.
- **Parallel processing**
  - The distributed data gets the power of N number of servers and machines in which data is residing and works in parallel for the processing and analysis
  - MapReduce (Google), Hadoop, Spark, Flink, etc.
- **Fault tolerance**
  - We keep the replica of a single block (or chunk) of data more than once.
- **Commodity hardware**
  - We don’t need specialized hardware with special RAID as Data container.
- **Flexibility and Scalability**
  - add more and more of rack space into the cluster as the demand for space increases.

A brief history of Data Science



Big Data Timeline v2.0

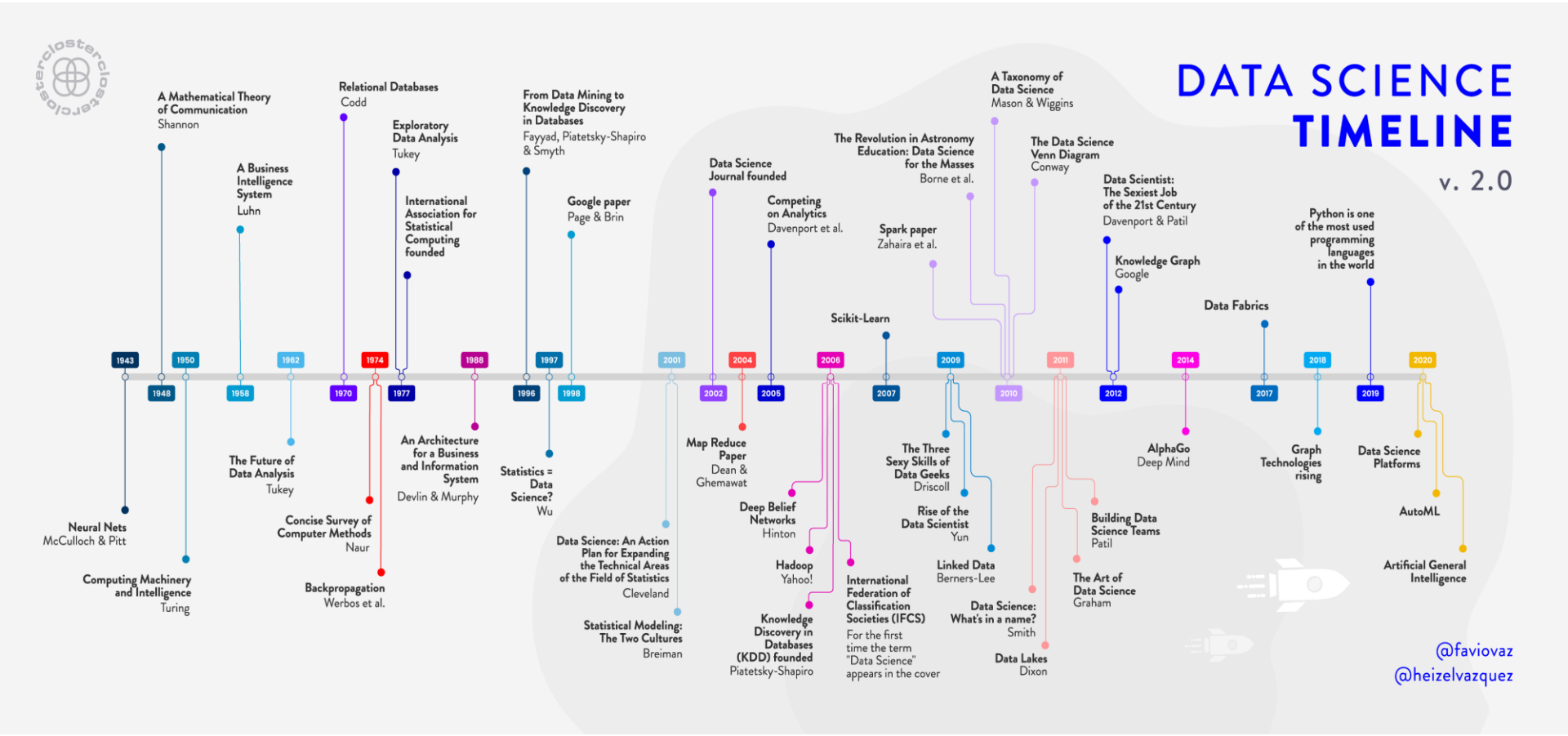


Illustration by Héizel Vázquez

Big Data Research Challenges

- **Heterogeneity and incompleteness**
  - machine analysis algorithms expect homogeneous data;
  - even after data cleaning and error correction, some incompleteness and some errors in data are likely to remain.
- **Scale**
  - managing large and rapidly increasing volumes of data;
  - clock speeds have largely stalled and processors are being built with increasing numbers of cores;
  - parallelism across nodes in a cluster;
  - move towards cloud computing;
  - large clusters requires new ways of determining how to run and execute data processing jobs.
- **Timeliness**
  - acquisition rate challenge and a timeliness challenge;
  - many situations in which the result of the analysis is required immediately;
  - ex: a full analysis of a user's purchase history is not likely to be feasible in real-time.
- **Privacy**
  - important to rethink security for information sharing in Big Data use cases;
  - we do not understand what it means to share data.
- **Human collaboration**
  - analytics for Big Data will be designed to have a human in the loop;
  - crowd-sourcing;
  - issues of uncertainty and error => participatory-sensing;
  - the extra challenge here is the inherent uncertainty of the data collection devices.

Major Benefits of Big Data

- **Answer more questions, more completely** (powerful big data business intelligence platform)
- **Faster and better decision making**
- **Become confident in your accurate data** (solving the problem of inaccurate, incomplete view of the data)
- **Empower a new generation of employees** (data scientists)
- **Cost reduction and revenue generation**

Big Remark

We must to support and we need to encourage

fundamental research towards addressing all scientific and technical challenges

if we want to achieve **the promised benefits of Big Data**.

Introduction in Big Data. Distributed platforms

Brief introduction

Datacenters, Cloud computing, HPC, Fog and Edge computing; Mobile computing; Vanet (V2X). Distributed Applications (practical use cases): Smart Cities; eHealth; eGovernment; Scientific Applications

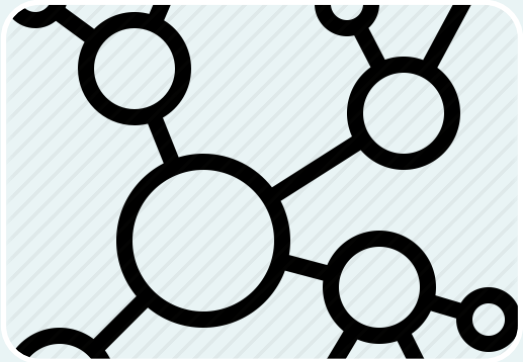
Datacenters

- Processing Big Data, the Google Way

Datacenters components

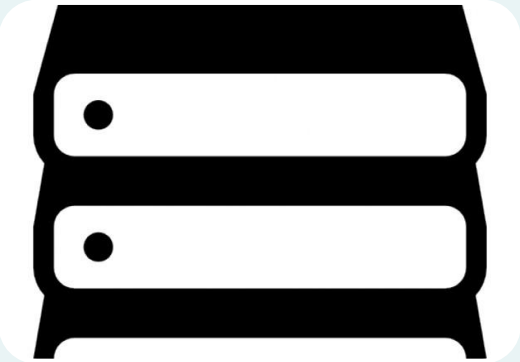
Computing Infrastructure

- Processing resources
- Memory resources



Network Infrastructure

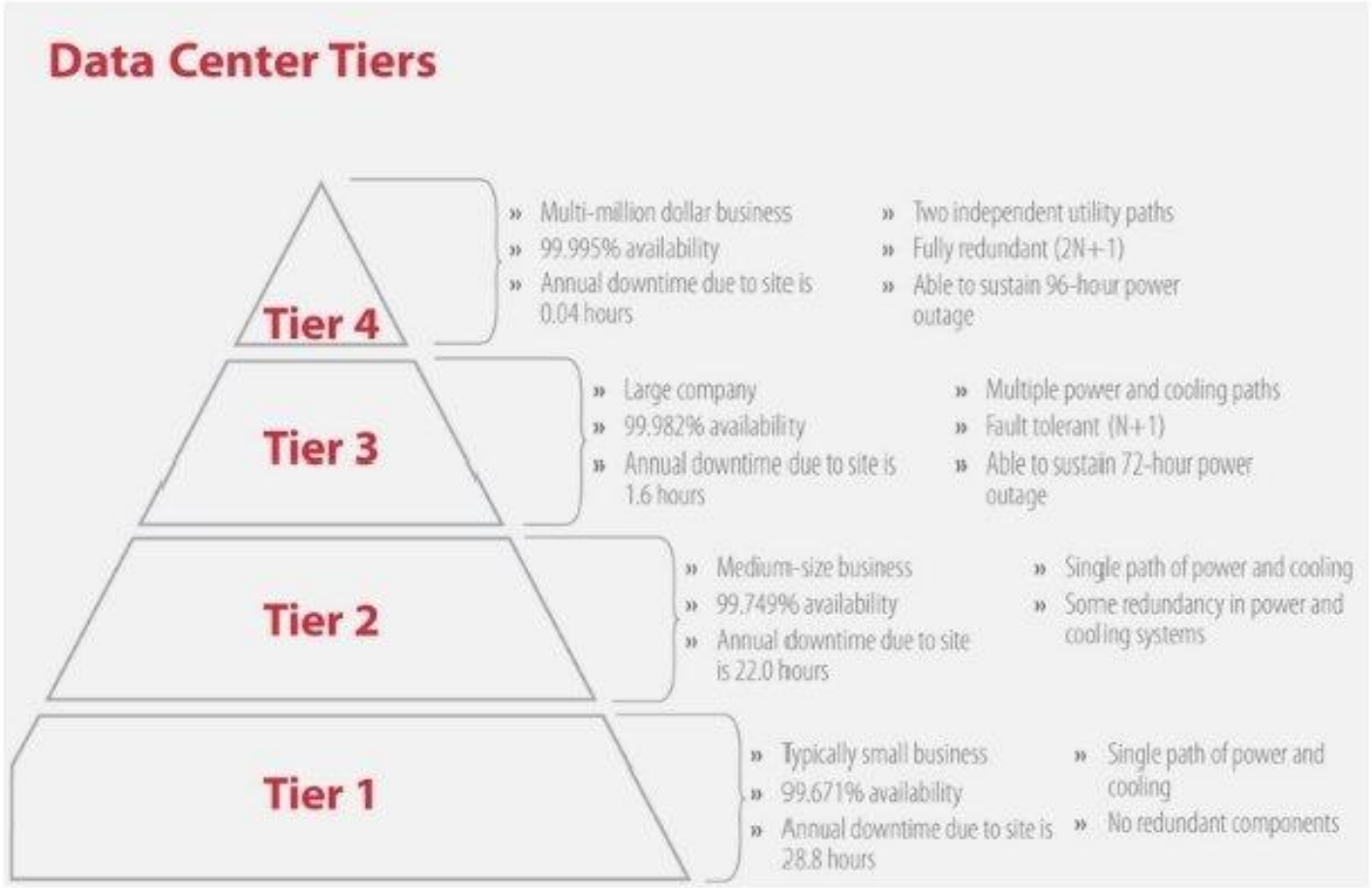
- Connected servers (physical and virtualized)
- Data centre services
- External connectivity



Storage Infrastructure

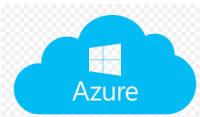
- Storage systems used for data storage

Data centre tier infrastructure

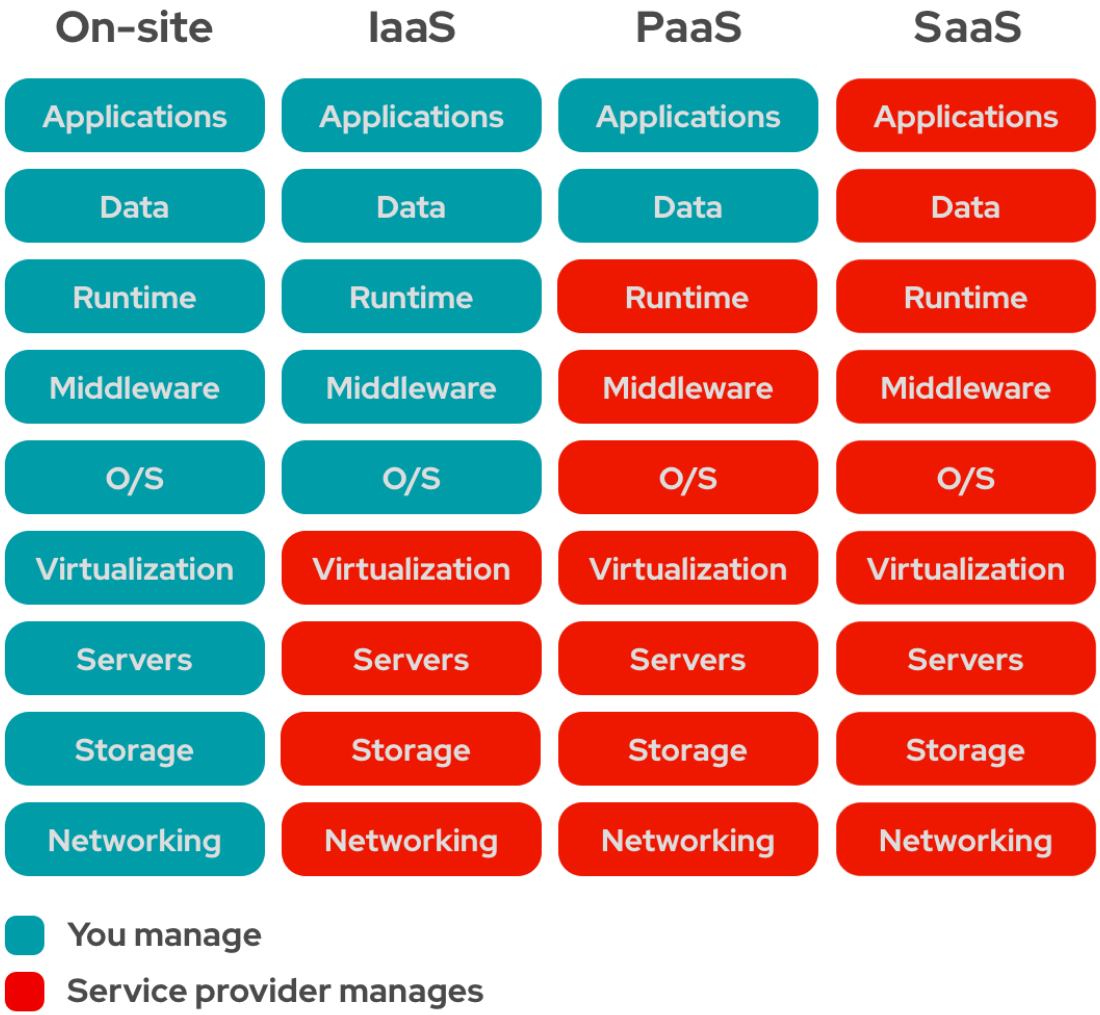


Cloud Computing

- Definition by IBM:
  - **Cloud computing, often referred to as simply "the cloud," is the delivery of on-demand computing resources—everything from applications to data centers—over the internet on a pay-for-use basis.**
- Cloud Computing Services:
  - **Software as a Service (SaaS):**
    - applications are supplied by the cloud provider;
    - applications are accessible from various client devices (e.g. web browser, API);
    - Examples: GMAIL, Dropbox, Office 365, Amazon Web Services
  - **Platform as a Service (PaaS)**
    - deploy applications onto the Cloud infrastructure
    - programming languages, tools and libraries are supported by the provider;
    - control over the deployed applications
  - **Infrastructure as a Service (IaaS)**
    - provision with fundamental computing resources: VM's (CPU, storage, network);
    - resources are distributed and support dynamic scaling;
    - deploy and run arbitrary software, meaning middleware, and operating systems;
    - control over operating systems, storage, and deployed applications
    - User does not manage or control the underlying Cloud infrastructure.

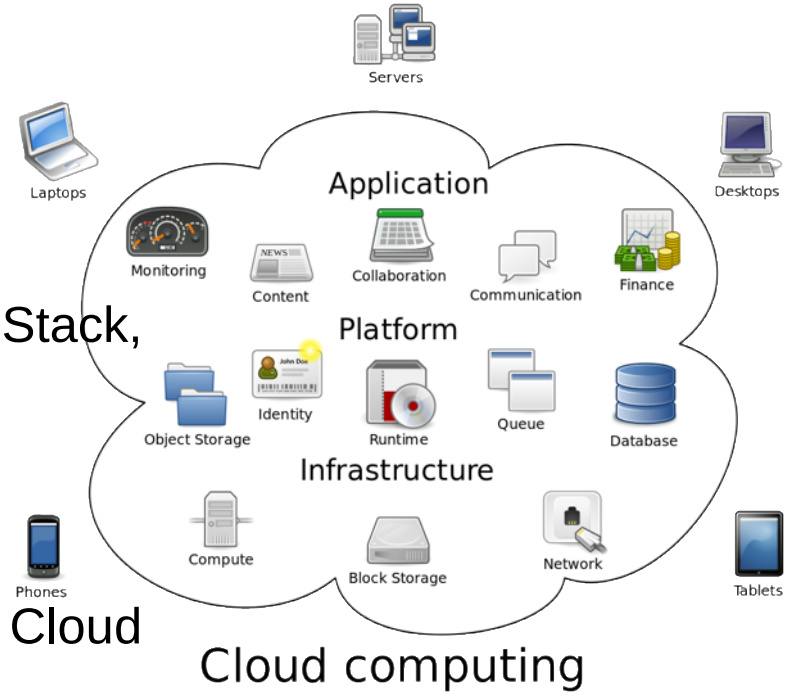


Cloud computing



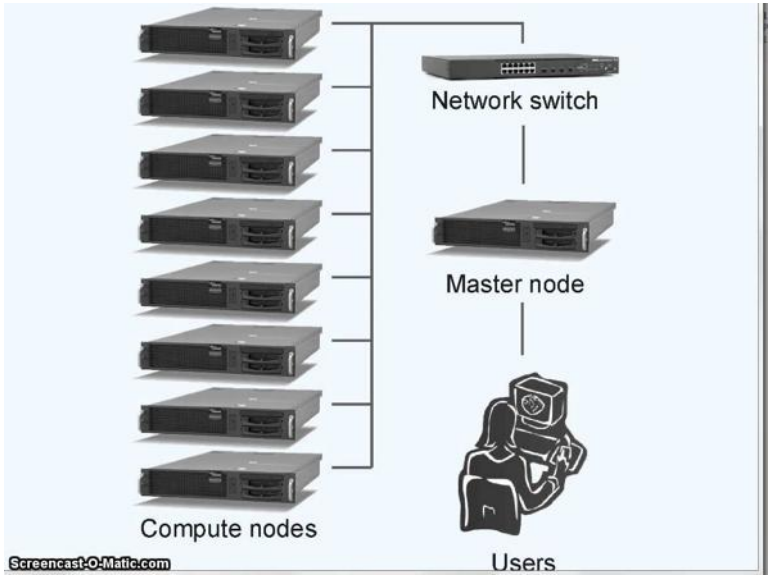
Cloud Computing - Deployment models

- Private
  - Single-tenant architecture
  - On-premises hardware
  - Direct control of infrastruscure
  - Ex: IBM, Red Had, Microsoft, OpenStack, ownCloud
- Public
  - Multi-tenant architecture
  - Pay-as-you-go pricing model
  - Ex: AWS, Microsoft Azure, Google Cloud Platform
- Hybrid
  - Cloud bursting capabilities
  - Benefit of both public and private environments
  - Ex: combination of both public and private providers



HPC

- *High Performance Computing most generally refers to the practice of* ***aggregating computing power in a way that delivers much higher performance than one could get out of a typical desktop computer or workstation*** *in order to solve large problems in science, engineering, or business.*

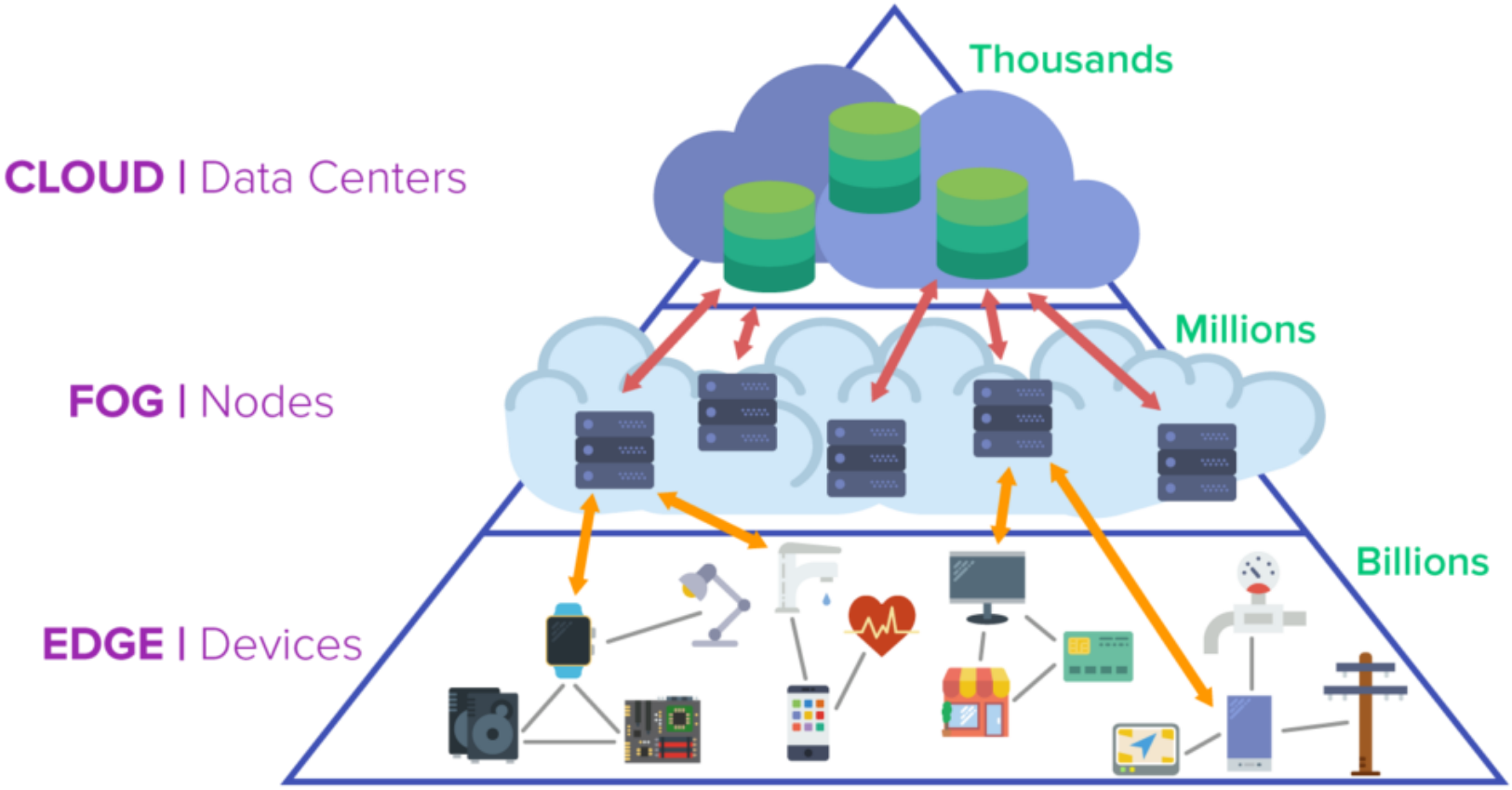


Top 5 World supercomputers

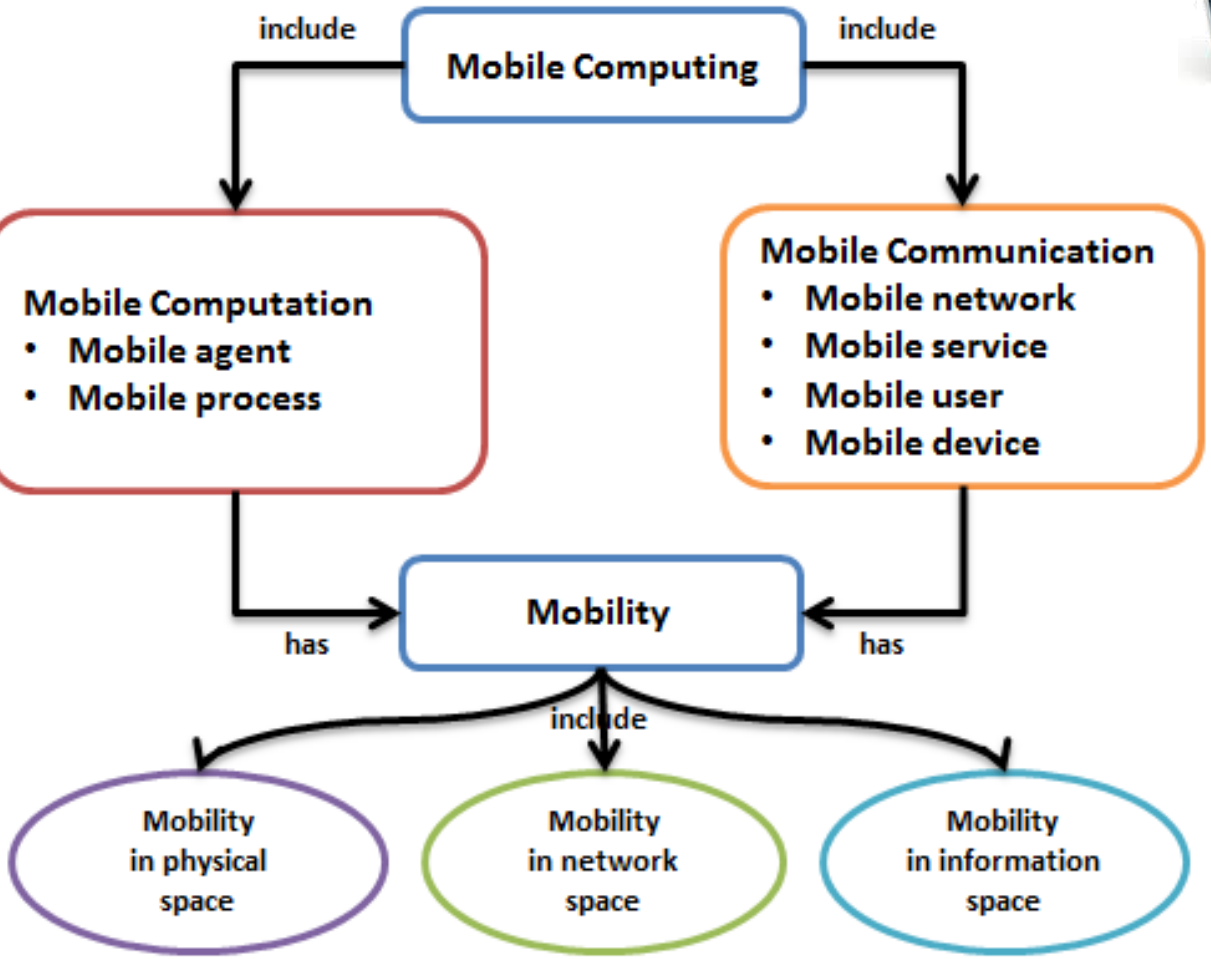
Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,206.00	1,714.81	22,786
2	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698
3	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States	2,073,600	561.20	846.84	
4	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
5	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,752,704	379.70	531.51	7,107

https://top500.org/lists/top500/2024/06/

Fog and Edge Computing

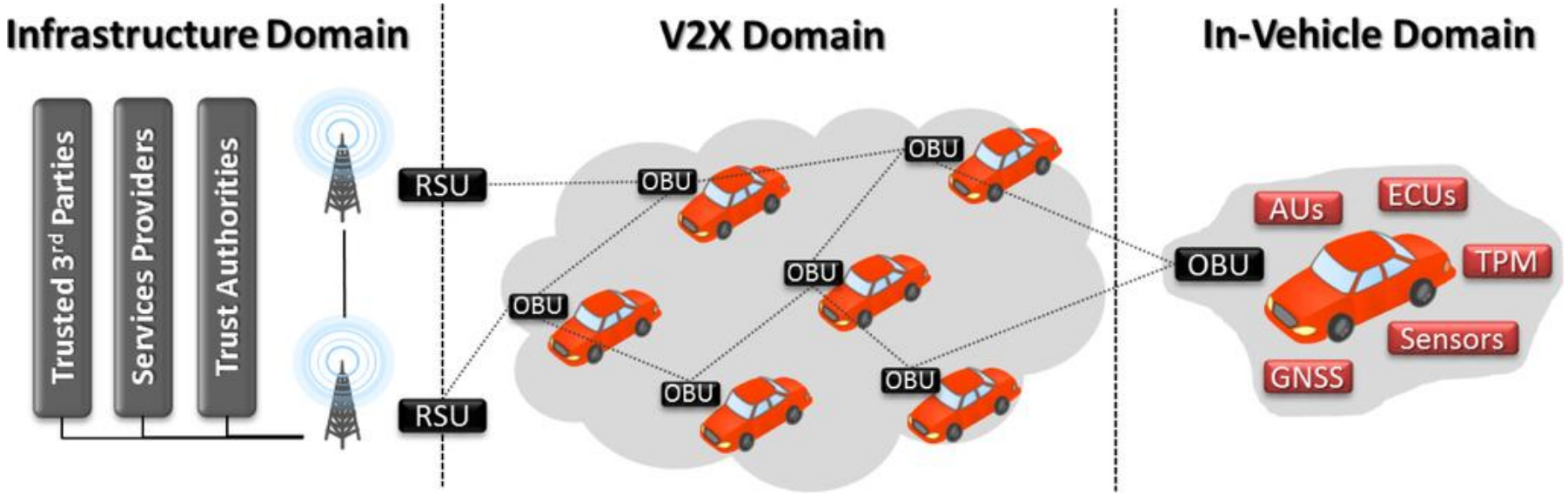


Mobile computing



Vanet (V2X)-Vehicular ad hoc networks

- Vehicle-to-Vehicle (V2V) among vehicles
  - A wireless ad hoc network on the roads, suited for short range vehicular networks;
- Vehicle-to-Infrastructure (V2I) between vehicles and Road-side-Units
  - Long range vehicular networks;
- Vehicle-to-X (V2X) mixed approach
  - Vehicle-to-everything



Practical applications

- Smart Cities
- eHealth
- eGouverment
- Scientific applications
- …