Center of Excellence:
Supercomputing for

Introduction to High Performance Computing, and its impact on life

by: Tassadaq Hussain

Professor Department of Electrical Engineering

Namal University Mianwali

Collaborations:

Barcelona Supercomputing Center, Spain

European Network on High Performance and Embedded Architecture and Compilation

Pakistan Supercomputing Center

- **Speaker Introduction**
- Objectives of this Event
- Mankind Progress
- Age of Big Data
- Artificial Intelligence Past Present and Future Challenges
- Untangling Artificial Intelligence Problems with Supercomputing
- Namal Supercomputing Facility

Introduction

Education:

PhD. Barcelona-Tech
Microsoft Research, Infineon Technologies France, Microsoft Research Cambridge, IBM

Suspenseful record of academic management as Professor and Dean

Enhanced Education Quality by Inculcating Outcome Based Education by Applied and Sustainable Projects

Experience:

19+ year's versatile experience in the area of Computer Architecture, AI, Software Architecture, Big-Data Architecture
Served National and International Academia, Industry and Government

- **Barcelona Science Park Spain**
- **Cambridge Science Park UK**
- **Technopolis Of Sofia-Antipolis, France**

Innovation, Research and Commercialization

Innovation and Research

- **110+ Million Pkr National and Int'l Funding.**
  - Supercomputing and Artificial Intelligence
  - Smart Electric Motor Controllers
  - Biomedical Applications
- **80+ Publications**
- **10 Patents**
- **10 MVPs**
- **5 Int'l Collaborations**

- ## Development & Commercialization

60+ Million of Industrial Investments.
Developed Digital Systems for Industry.
Transform Idea into product.
Innovation and Commercialization for Sustainable economic and industrial development.

- ## Capacity Building:

Conducted more than **50 national and international workshops** and **training on Commercializable research, Writing successful grant proposal, and research and innovation.**

Provides Consultancy and Support for Entrepreneurship, Start-ups, Business Innovation and Technology transfer.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

UVHC
UNIVERSITE DE VALENCIENNES ET DU HAINAUT - CAMBRESIS

NEW ZEALAND COLLEGE OF CHIROPRACTIC
graduating hands, hearts & minds

PAKISTAN SUPERCOMPUTING

EntrepreneurialPark
Research Innovation Commercialization

Foot Analytic

ComputingPark

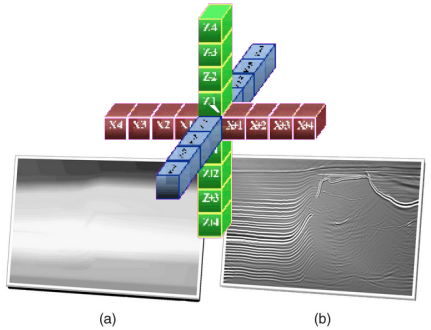
Int'l Projects

- Design Ultra Low Cost Display Camera Interface for Mobile Baseband XGold Chip **(Infineon Techonogies, 200 million single chip)**

Int'l Projects

- Design Ultra Low Cost Display Camera Interface for Mobile Baseband XGold Chip **(Infineon Techonogies, 200 million single chip)**

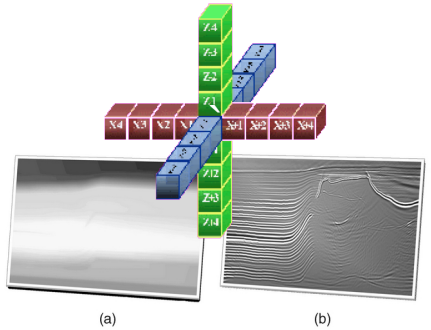
- Implementation of Reverse Time Migration on FPGAs
  (BSC-REPSOL, PLDA Italia, Cambridge Science Park)



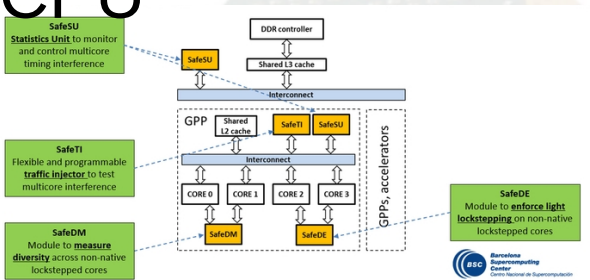
Int'l Projects

- Design Ultra Low Cost Display Camera Interface for Mobile Baseband XGold Chip **(Infineon Techonogies, 200 million single chip)**

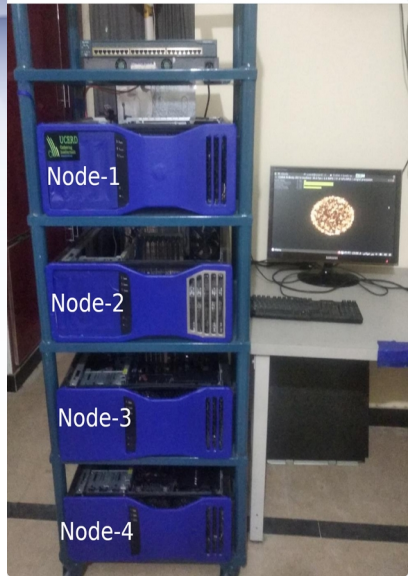
- Implementation of Reverse Time Migration on FPGAs
  (BSC-REPSOL, PLDA Italia, Cambridge Science Park)



- Open source European full-stack ecosystem based on a new RISC-V CPU
  (Barcelona Supercomputing Center)



National

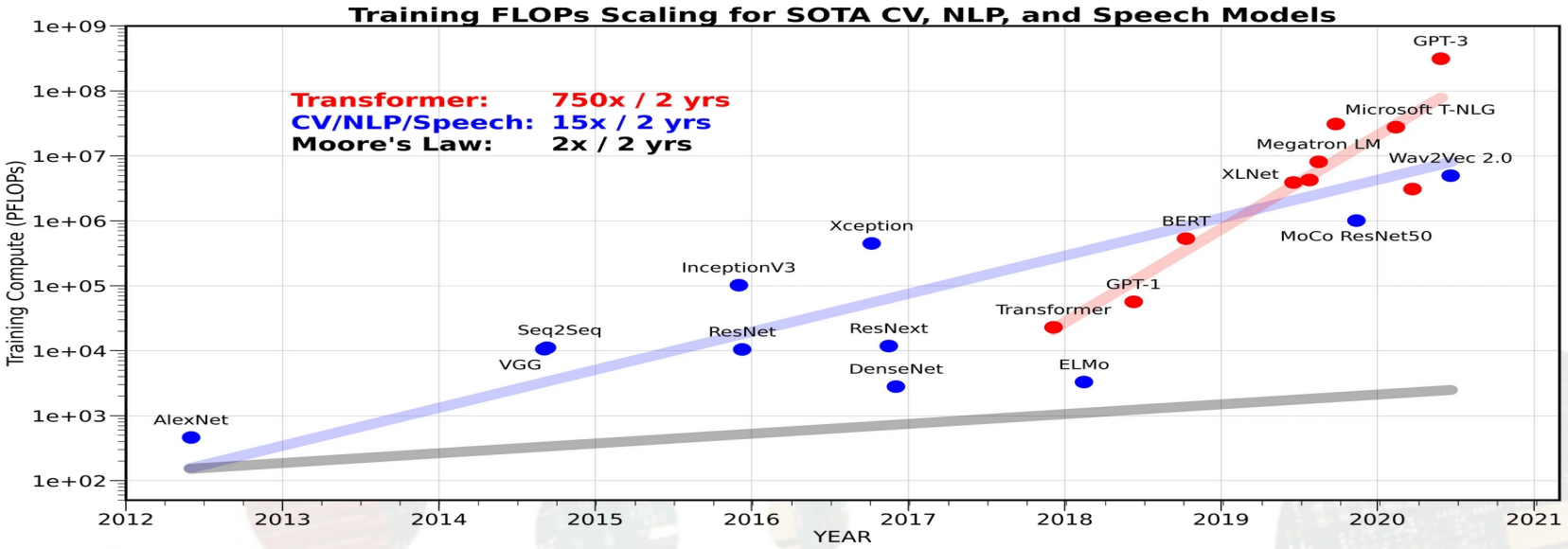


- Supercomputing and AI for Health Sciences
- Pakistan Supercomputing Center
- FPGA Power Supercomputer
- Scalable Heterogeneous Supercomputing System
- Smart Motor Controller
- FootAnalytic
- VR/AR for Rehabilitation
- Live-Stock Breed Identification System

- Speaker Introduction
- **Objectives of this Event**
- Mankind Progress
- Age of Big Data
- Artificial Intelligence Past Present and Future Challenges
- Untangling Artificial Intelligence Problems with Supercomputing
- Namal Supercomputing Facility

Objectives of the School

- World Data Size = 130 Zettabytes, doubling every 18 months.
- The computational demands of AI algorithms are experiencing exponential growth.
- The performance of microelectronics is not increasing in line with Moore's Law.



PAKISTAN SUPERCOMPUTING

Mastery of AI is essential; a lack may result in unforeseen consequences.

Deep and steep

Computing power used in training AI systems

Days spent calculating at one petaflop per second*, log scale

By fundamentals

Language Speech Vision

Games Other

AlphaGo Zero becomes its own teacher of the game Go

AlexNet, image classification with deep convolutional neural networks

3.4-month doubling

Two-year doubling (Moore's Law)

Perceptron, a simple artificial neural network

← First era →

→ Modern era

100 10 1 0.1 0.01 0.001 0.0001 0.00001 0.000001 0.0000001

1960 70 80 90 2000 10 20

Source: OpenAI

*1 petaflop=10^{15} calculations

The Economist

Objectives

Educate , Collaborate and Accelerate

The goal of this school is to foster **interdisciplinary collaboration** and **teamwork across departments** within the **University** through the exploration of high performance computing, big data and artificial intelligence (AI).

by:

Leveraging the collective expertise and resources, challenges and opportunities

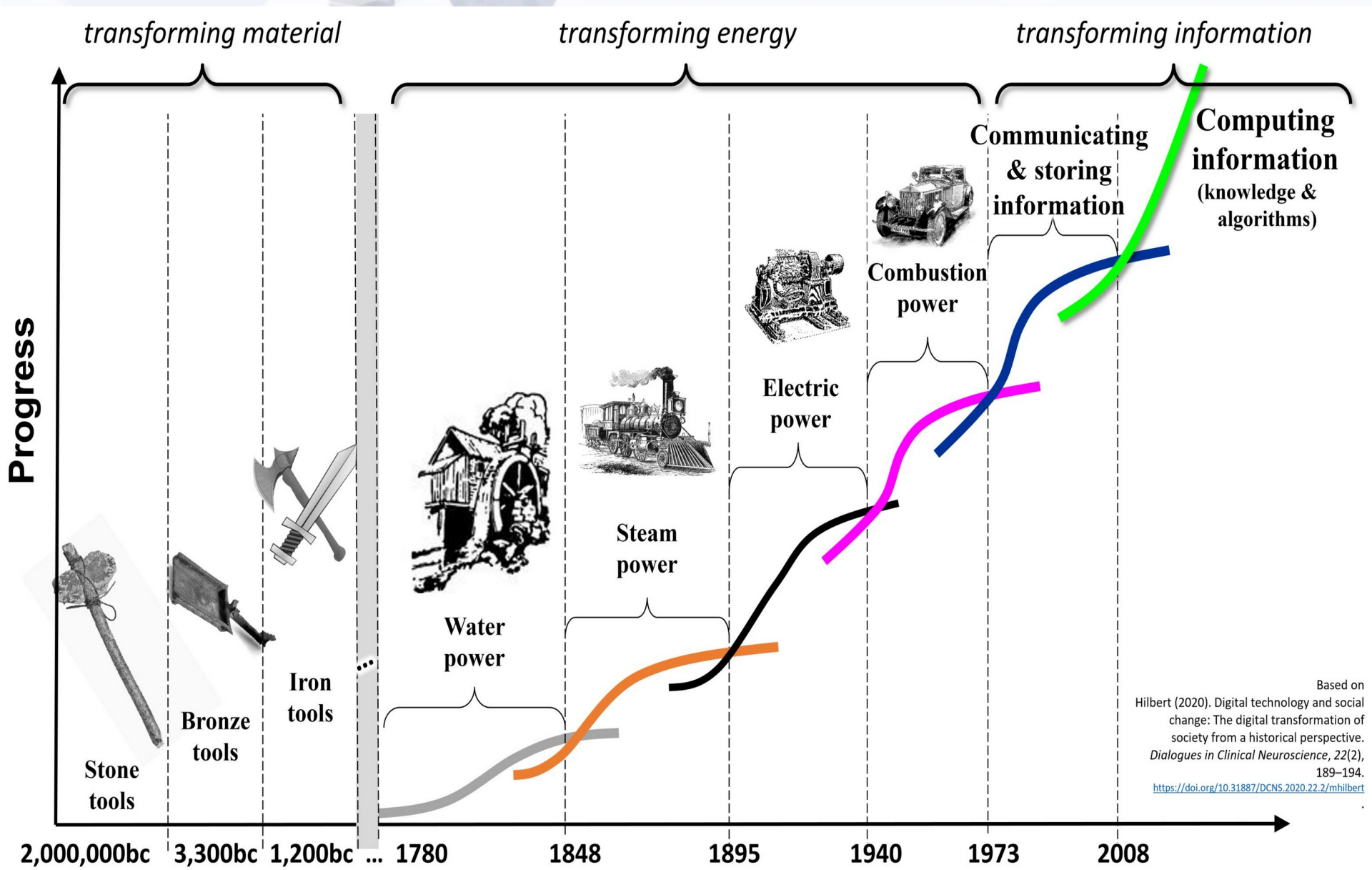
for:

Advancing research, education, and societal impact.

"The future we will "invent" is a choice we make jointly, not something that happens." Jordi

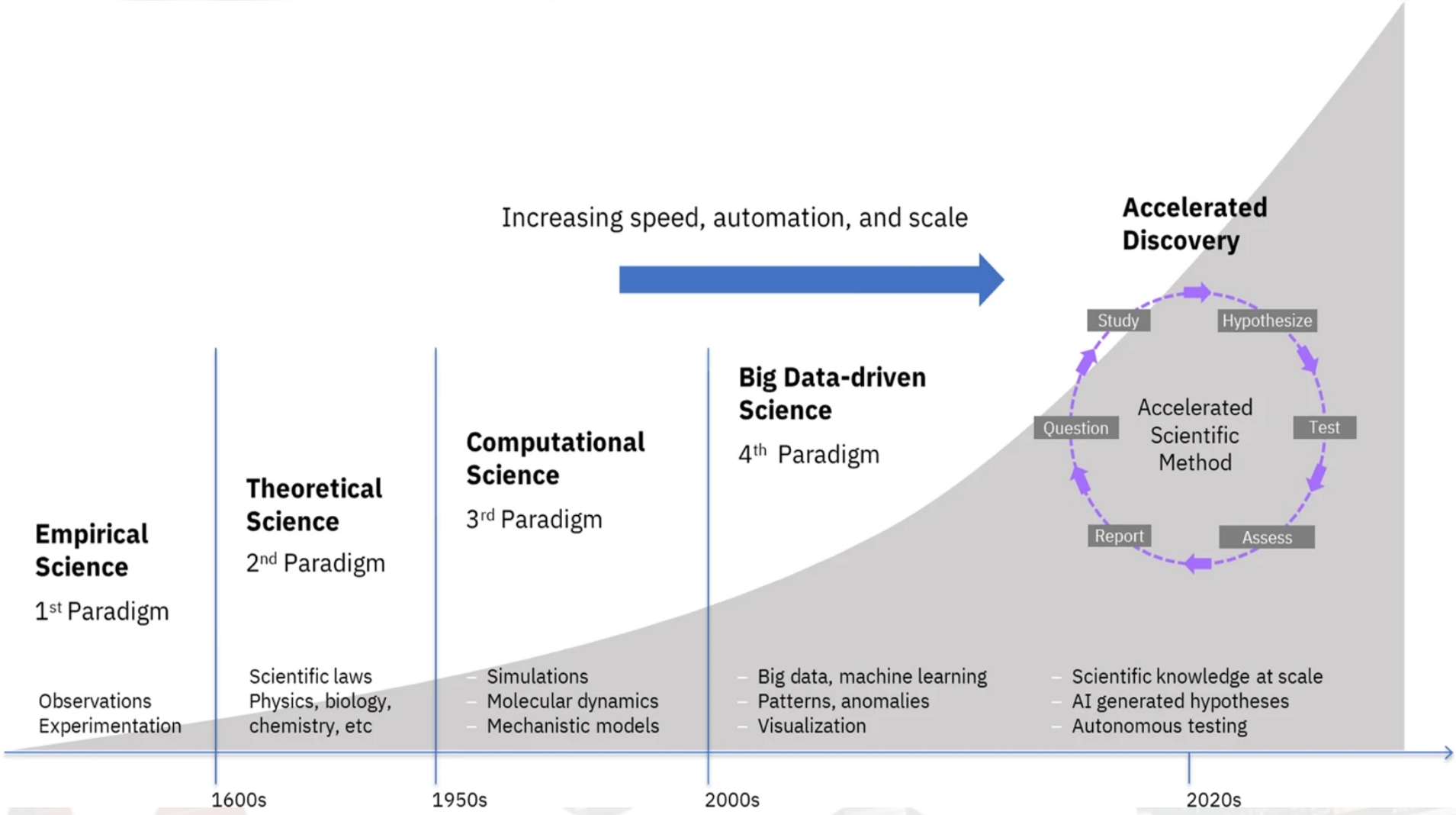
- Speaker Introduction
- Objectives of this Event
- **Mankind Progress**
- Age of Big Data
- Artificial Intelligence Past Present and Future Challenges
- Untangling Artificial Intelligence Problems with Supercomputing
- Namal Supercomputing Facility

Mankind Progress



Based on Hilbert (2020). Digital technology and social change: The digital transformation of society from a historical perspective. *Dialogues in Clinical Neuroscience*, *22*(2), 189–194. https://doi.org/10.31887/DCNS.2020.22.2/mhilbert .

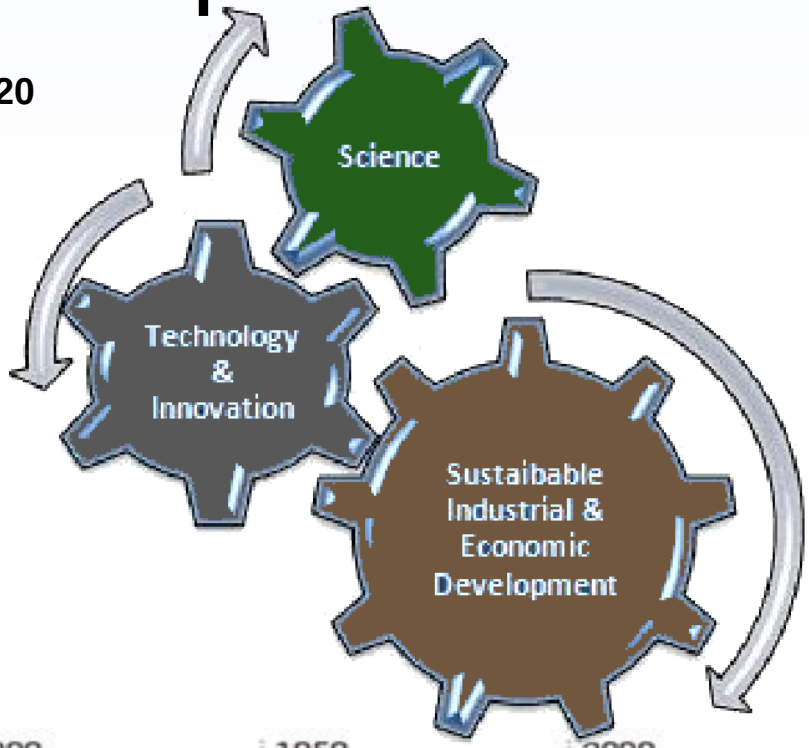
From Age of Empirical Science to Data-Science

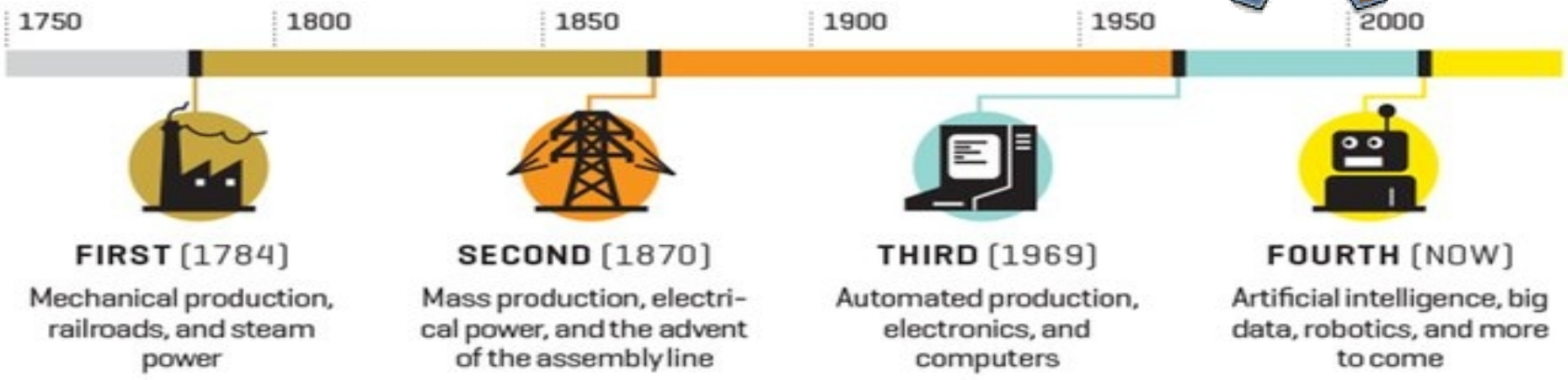


https://www.nature.com/articles/s41524-022-00765-z/figures/1

Industrial Revolutions and Sustainable Developments

Top 10 Biggest Companies By Market Size From 2010 - 2020





Life Science

Earth Science

Social Science

Science

175 ZByte @2025

80%
Data-Sciences

Data

100 ExaFLOPS
@2020

87.04 B$

234.6 B$ @2025

AI

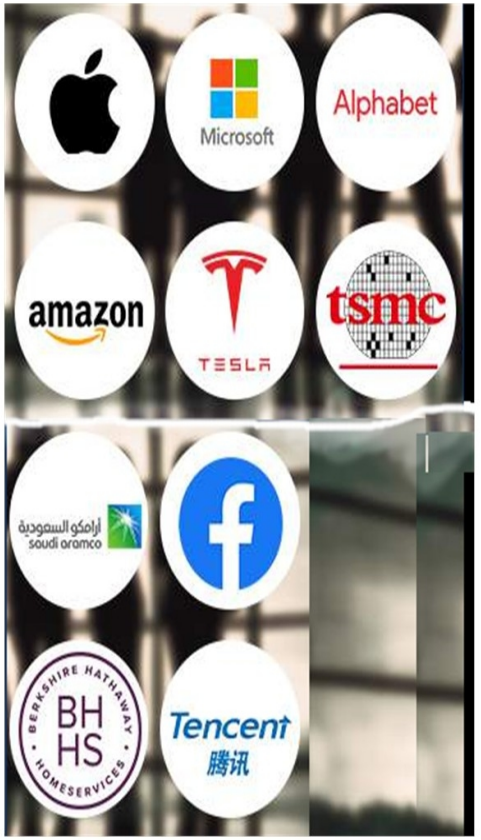
Top500 List
8 PetaFLOPS
@2022

uProcessor
100 B$ @2020

30% Cell Phone
20% Embedded
App
50 Servers, PCs etc.

Computing

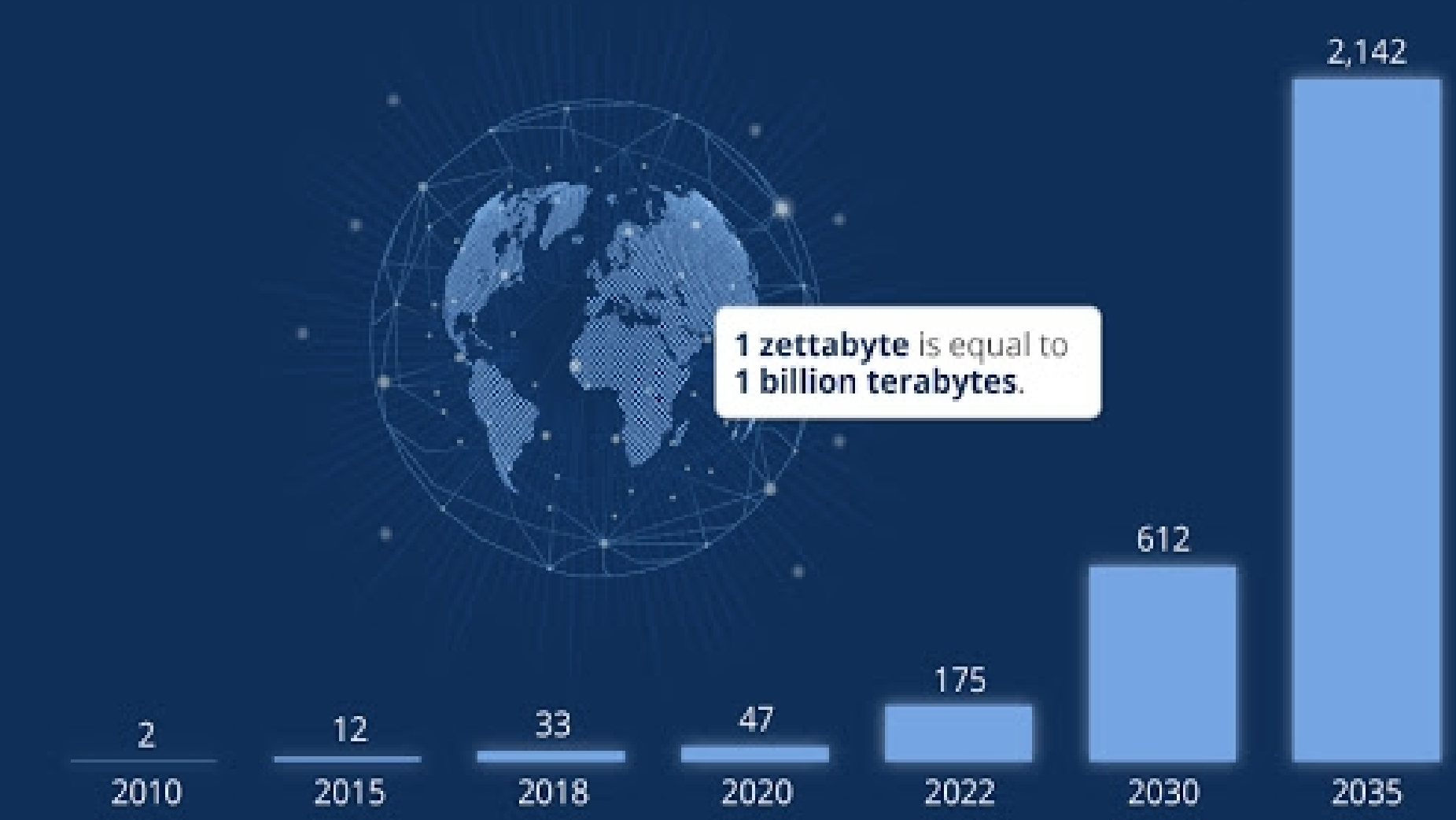
Digital Industrial Age
5.5 Trillion $ Revenue@2021



- Speaker Introduction
- Objectives of this Event
- Mankind Progress
- **Age of Big Data**
- Artificial Intelligence Past Present and Future Challenges
- Untangling Artificial Intelligence Problems with Supercomputing
- Namal Supercomputing Facility

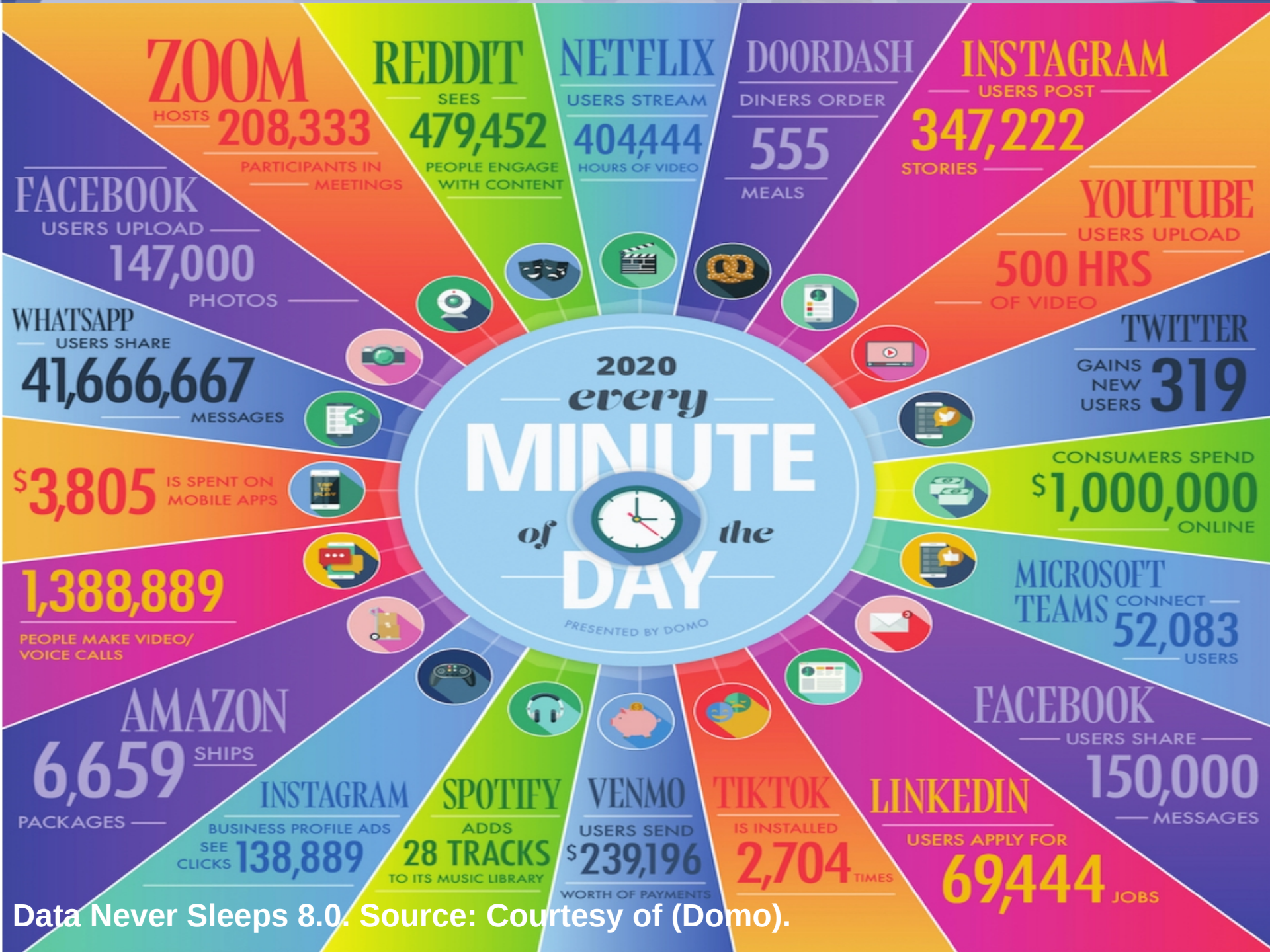
Global Data Creation is About to Explode

Actual and forecast amount of data created worldwide 2010-2035 (in zettabytes)

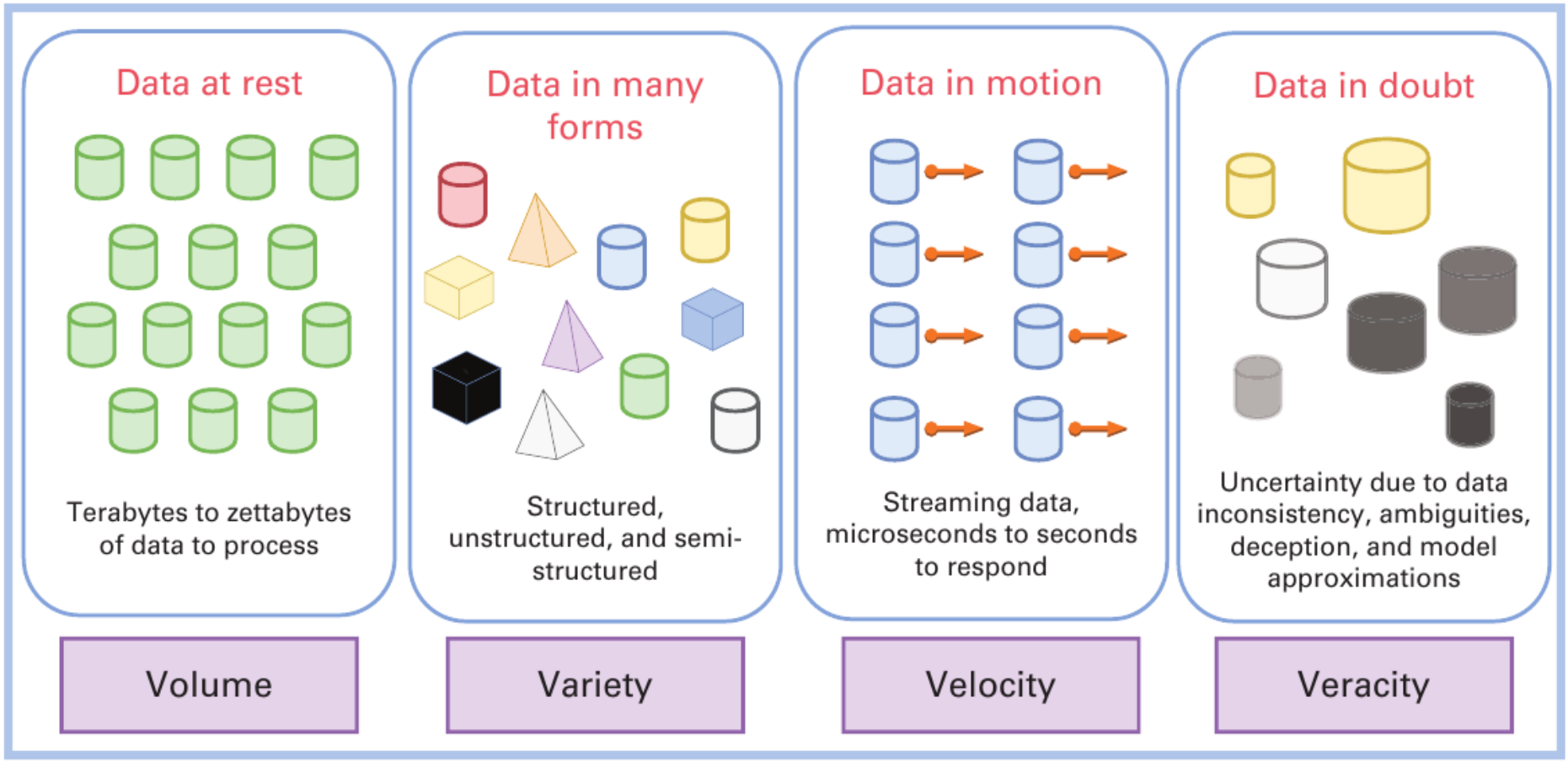


57.76 US$

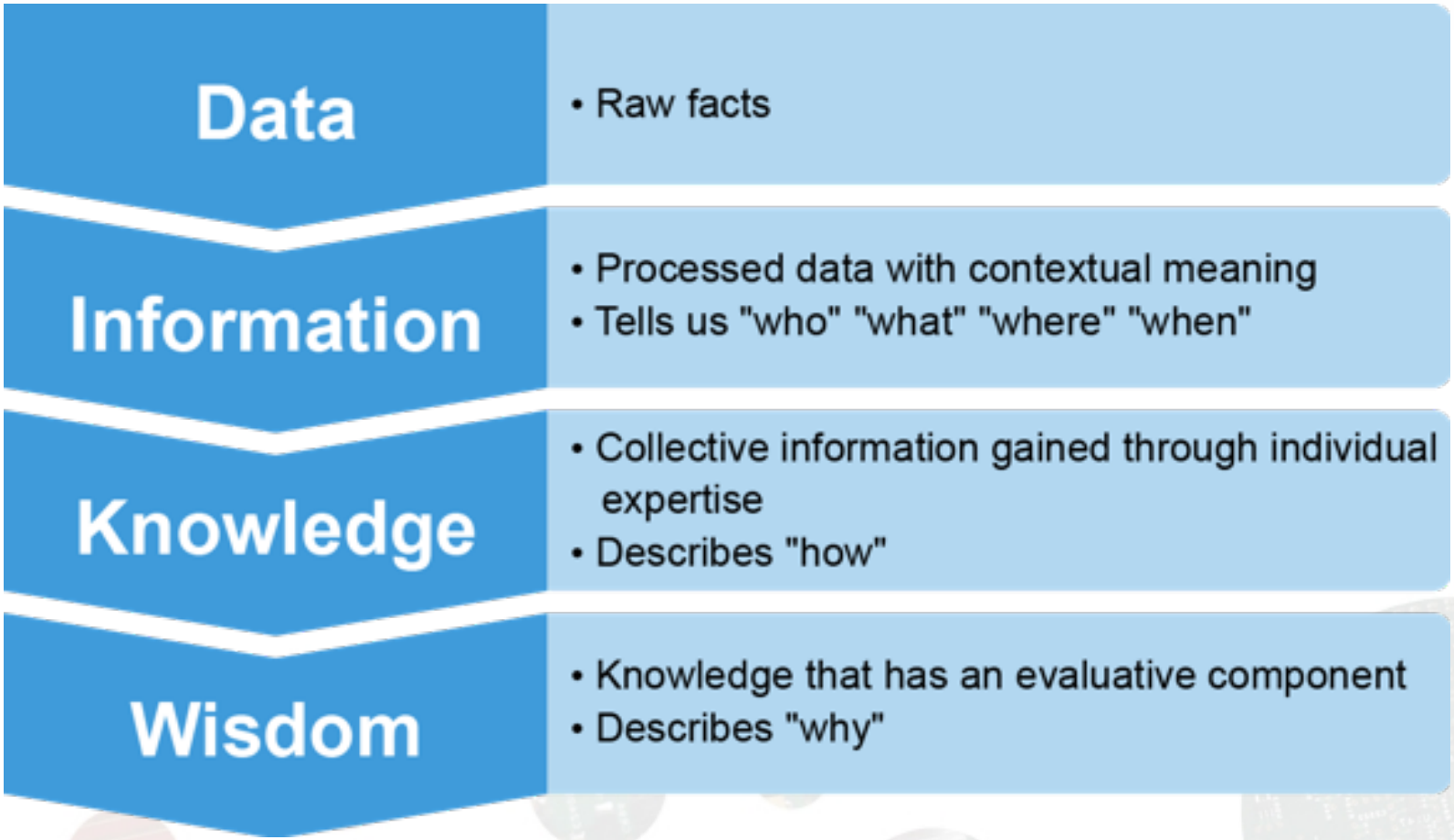
statista



Data Never Sleeps 8.0. Source: Courtesy of (Domo).


Data at rest
Terabytes to zettabytes of data to process
Data in many forms
Structured, unstructured, and semi-structured
Data in motion
Streaming data, microseconds to seconds to respond
Data in doubt
Uncertainty due to data inconsistency, ambiguities, deception, and model approximations
Volume
Variety
Velocity
Veracity

The Data-Information-Knowledge-Wisdom



https://www.researchgate.net/publication/332400827_How_can_nutrition_models_increase_the_production_efficiency_of_sheep_and_goat_operations/figures?lo=1

- Speaker Introduction
- Objectives of this Event
- Mankind Progress
- Age of Big Data
- **Artificial Intelligence Past Present and Future Challenges**
- Untangling Artificial Intelligence Problems with Supercomputing
- Namal Supercomputing Facility

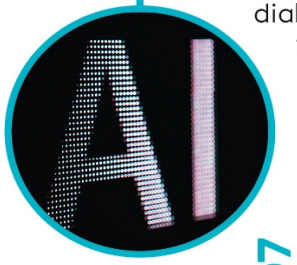
1642 — First mechanical calculating machine built by French mathematician and inventor Blaise Pascal.

1837 — First design for a programmable machine, by Charles Babbage and Ada Lovelace.

1943 — Foundations of neural networks established by Warren McCulloch and Walter Pitts, drawing parallels between the brain and computing machines.

1950 — Alan Turing introduces a test—the Turing test—as a way of testing a machine's intelligence.

1955 — 'Artificial intelligence' is coined during a conference devoted to the topic.



1965 — ELIZA, a natural language program, is created. ELIZA handles dialogue on any topic; similar in concept to today's chatbots.

1980s — Edward Feigenbaum creates expert systems which emulate decisions of human experts.

1997 — Computer program Deep Blue beats world chess champion Garry Kasparov.

2002 — iRobot launches Roomba, an autonomous vacuum cleaner that avoids obstacles.

2009 — Google builds the first self-driving car to handle urban conditions.

2011 — IBM's Watson defeats champions of US game show Jeopardy!

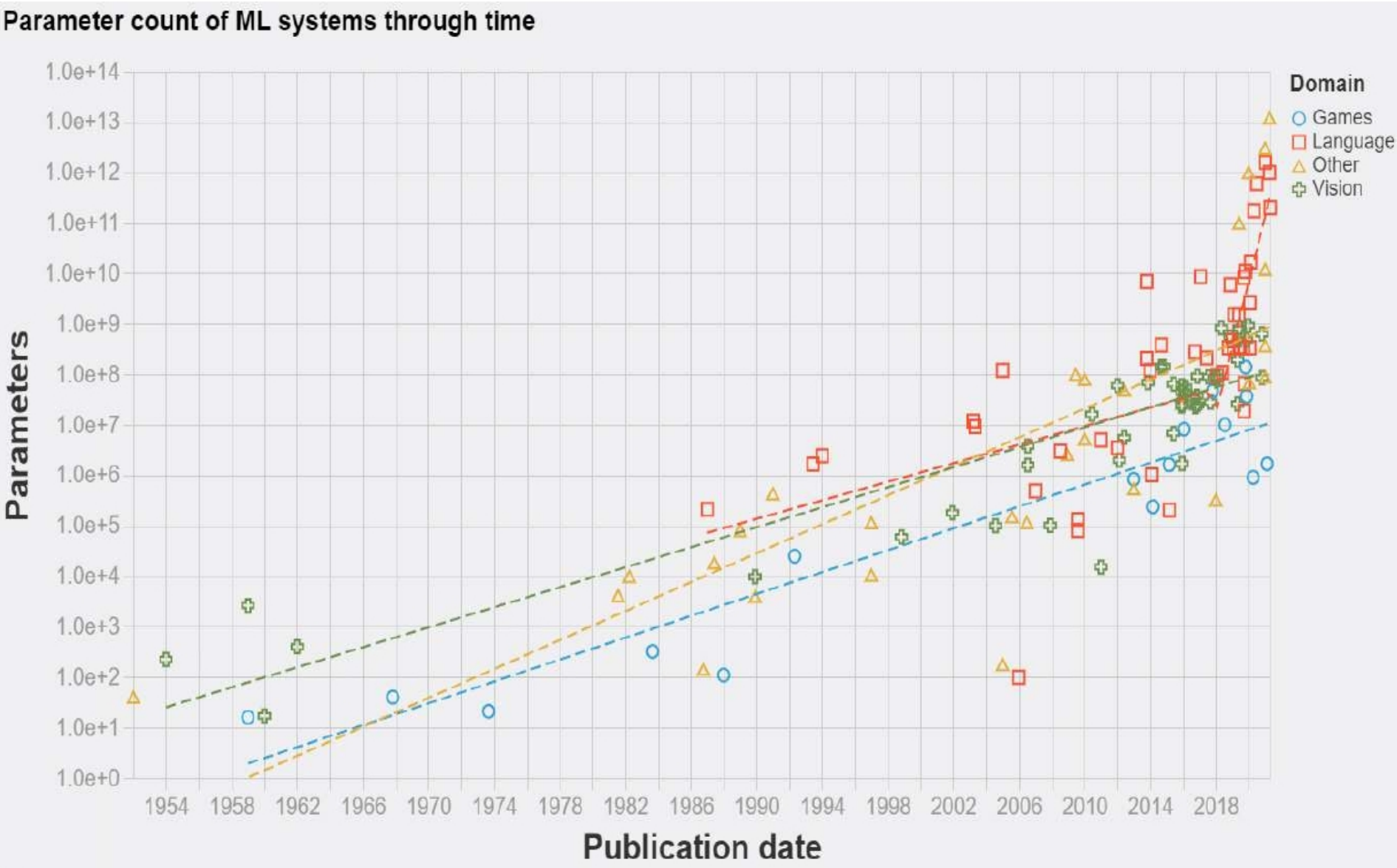
2011–2014 — Personal assistants like Siri, Google Now, Cortana use speech recognition to answer questions and perform simple tasks.

2014 — Ian Goodfellow comes up with Generative Adversarial Networks (GAN).

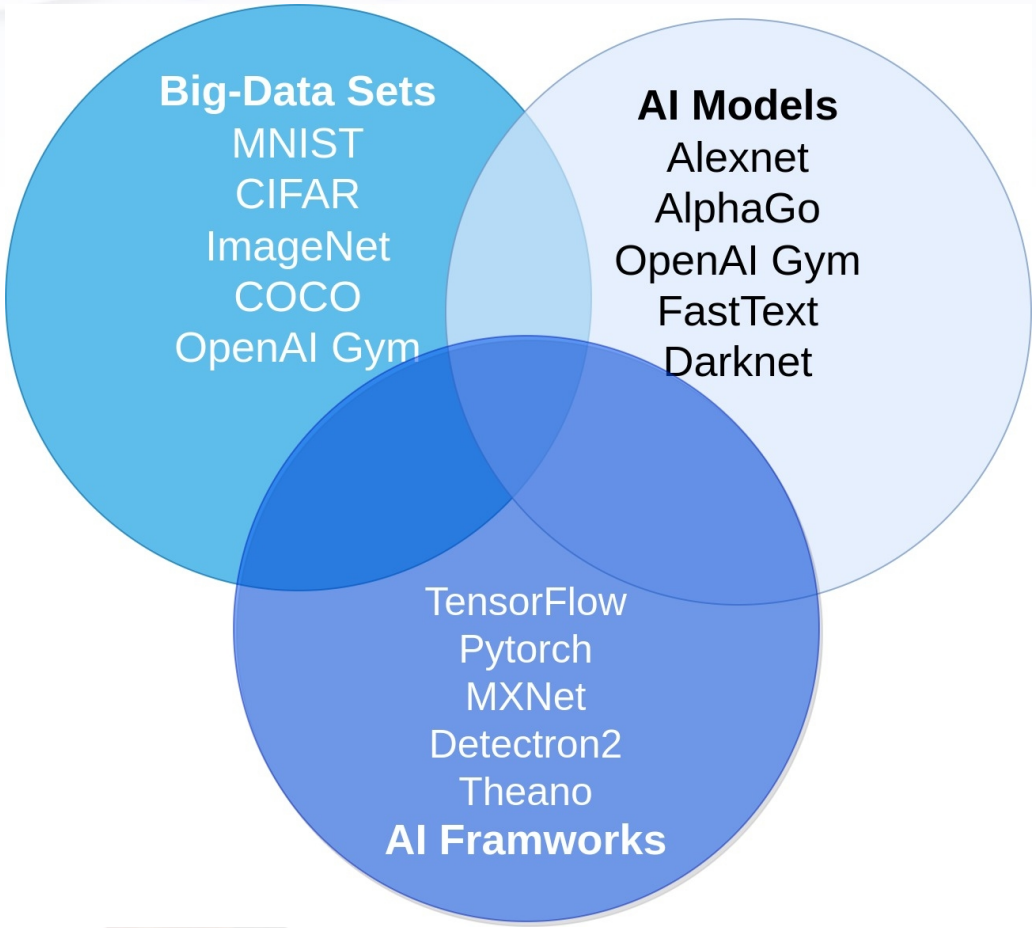
2016 — AlphaGo beats professional Go player Lee Sedol 4-1.

2018 — Most universities have courses in Artificial Intelligence.

DL Relentless growth in model size



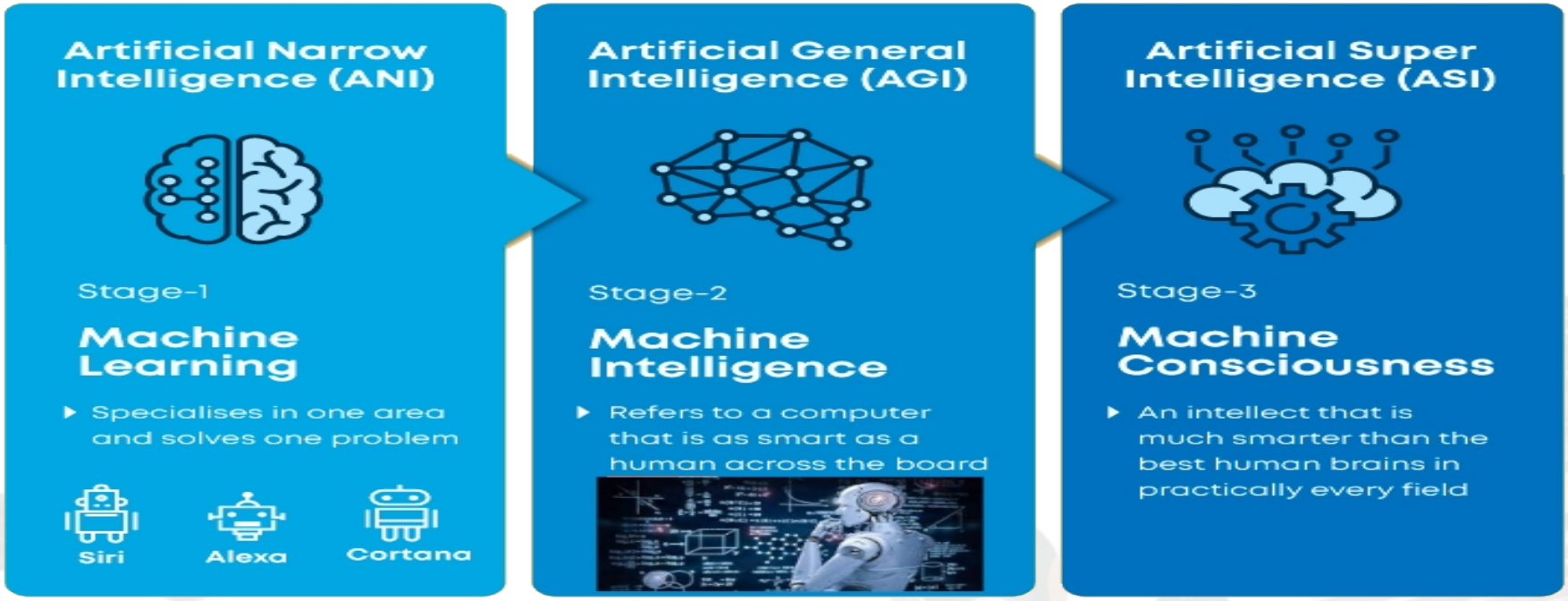
Open-Source



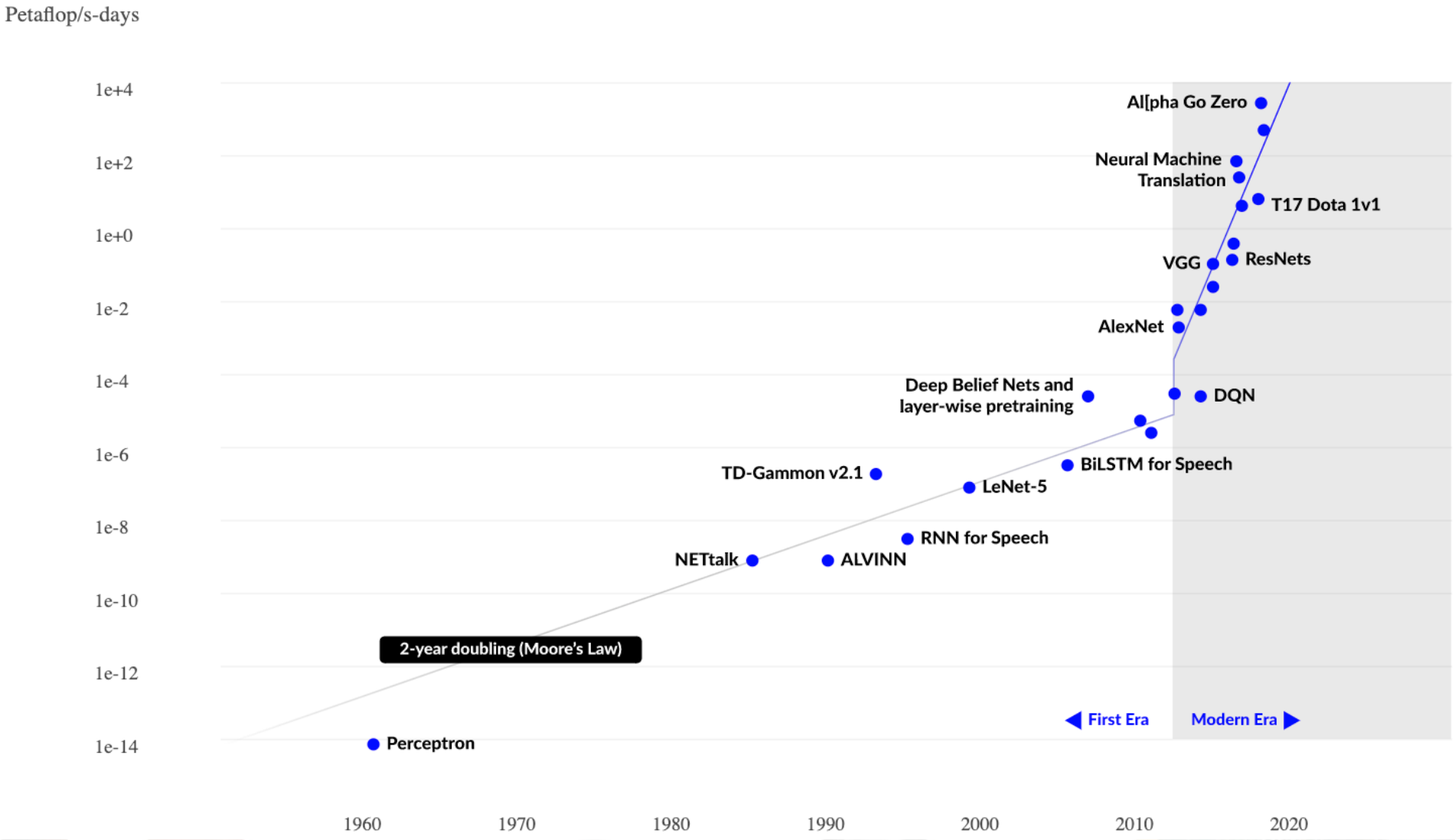
Big datasets, and open-source DL framework, play an important role to create “big” algorithms.

Intelligent Algorithms

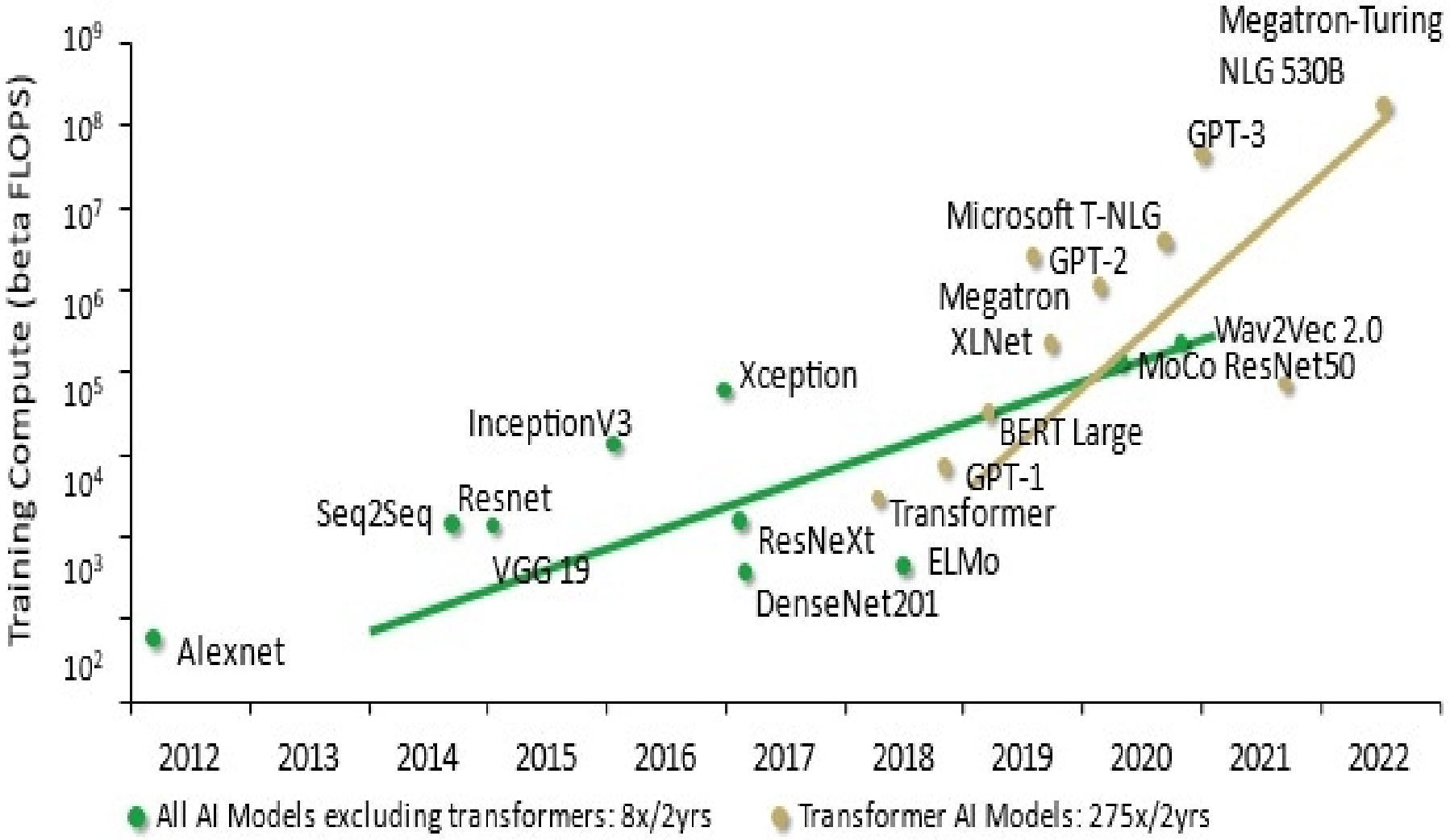
- **Performance**
  - **Execution Time**
  - **Accuracy** “The accuracy of the model is inherently tied to the quality, diversity, and representativeness of the data used for training and evaluation."
  - **Scalability** “Methods that scale with computation are the future of Artificial Intelligence” — Rich Sutton,



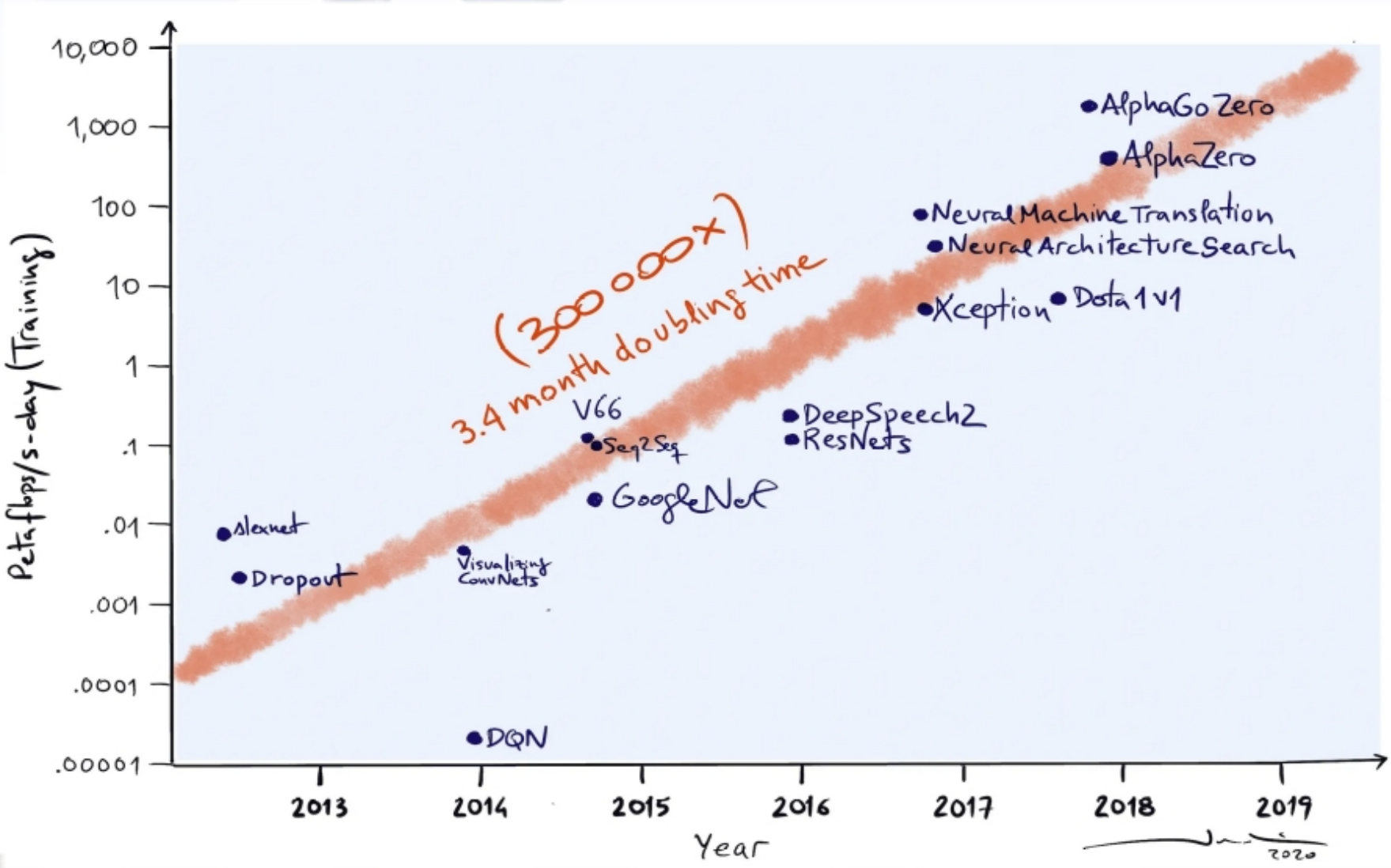
Computation Demand



The total amount of compute, in petaflop/s-days,[2] used to train selected results that are relatively well known, used a lot of compute for their time, and gave enough information to estimate the compute used.

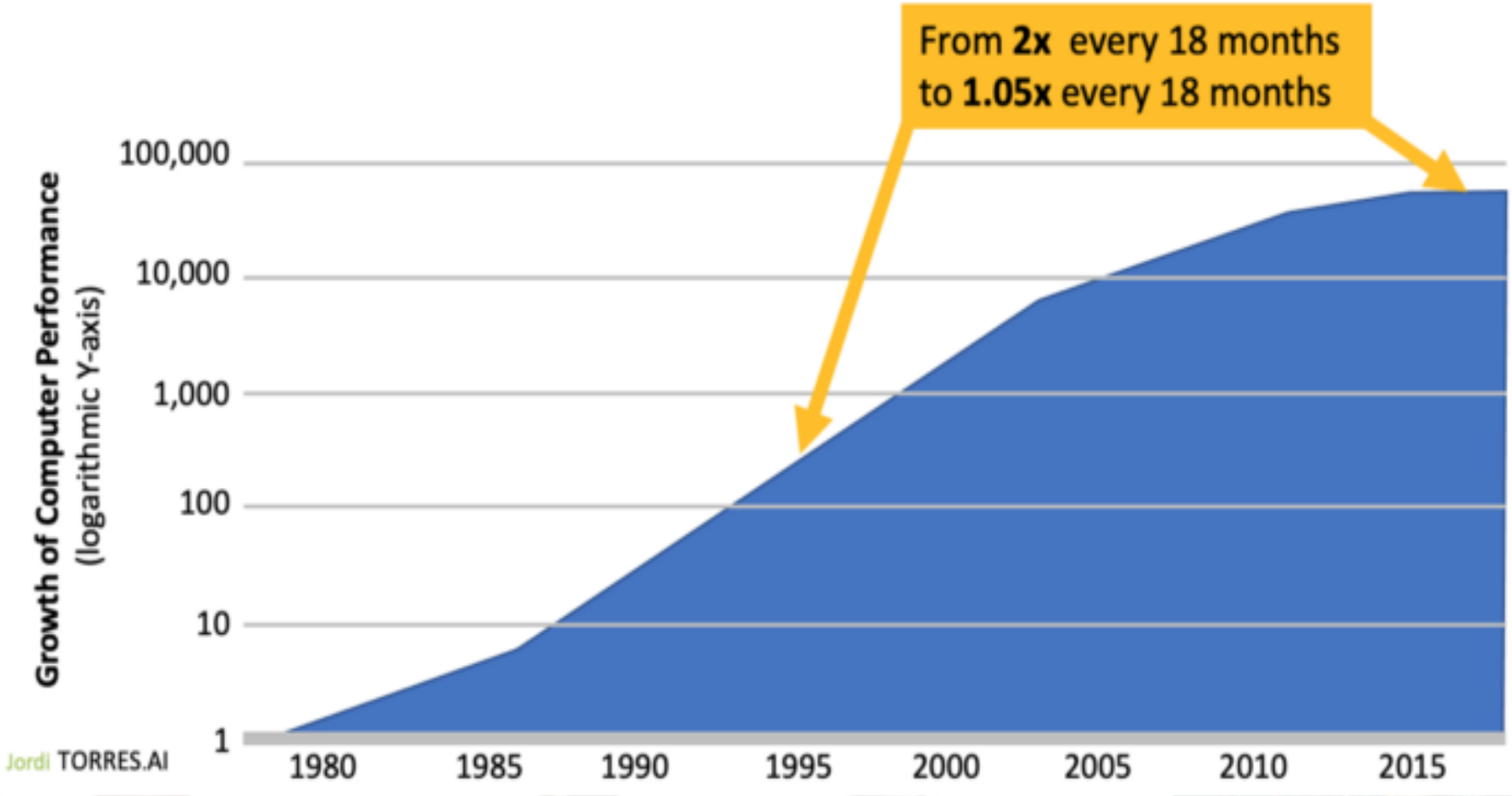

Training Compute (beta FLOPS)
10^9
10^8
10^7
10^6
10^5
10^4
10^3
10^2
Megatron-Turing NLG 530B
GPT-3
Microsoft T-NLG
GPT-2
Megatron
XLNet
Wav2Vec 2.0
MoCo ResNet50
Xception
InceptionV3
BERT Large
GPT-1
Transformer
Seq2Seq
Resnet
ResNeXt
ELMo
VGG 19
DenseNet201
Alexnet
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
All AI Models excluding transformers: 8x/2yrs
Transformer AI Models: 275x/2yrs

AI Computational Requirements

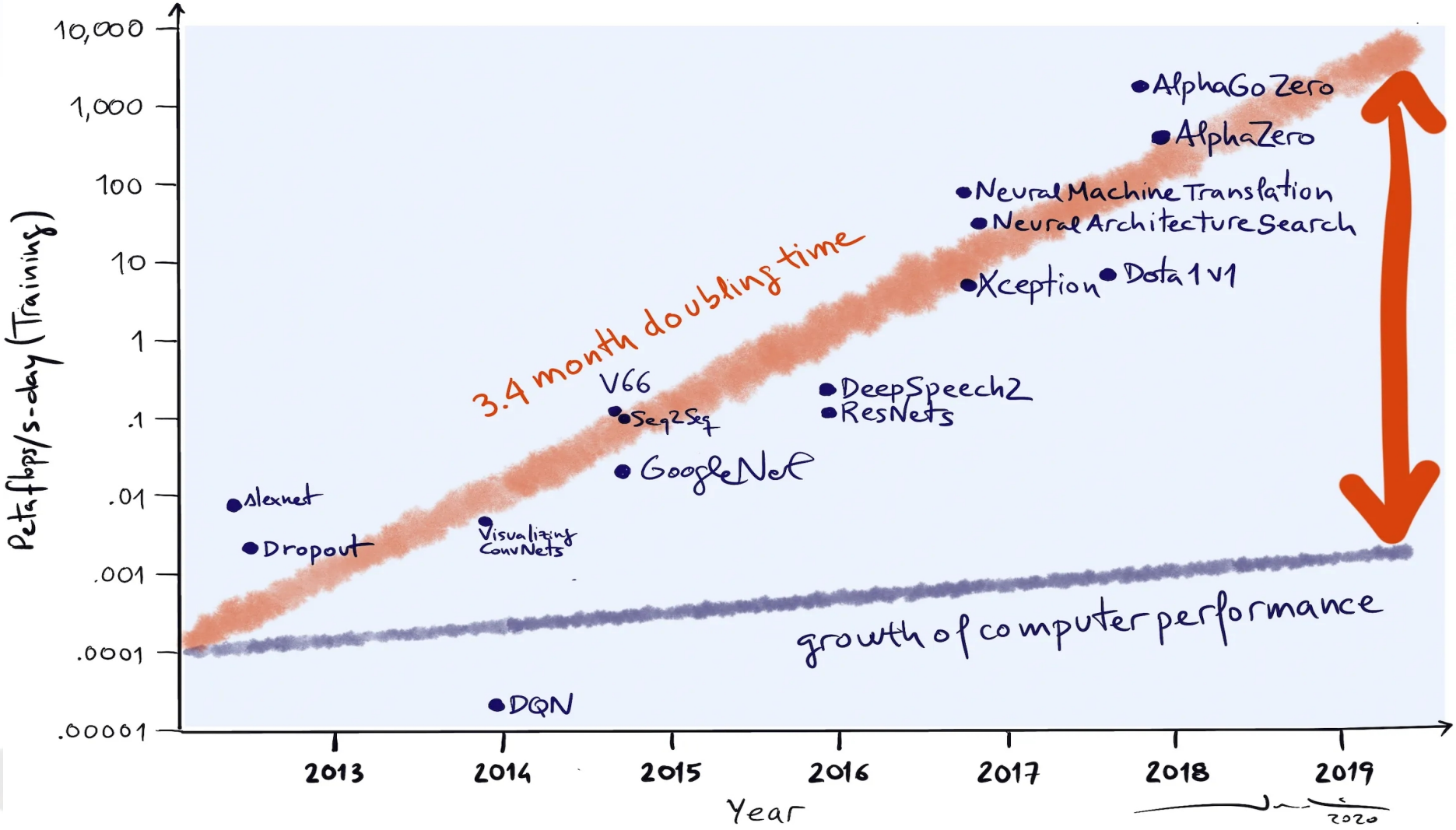


https://towardsdatascience.com/artificial-intelligence-is-a-supercomputing-problem-4b0edbc2888d

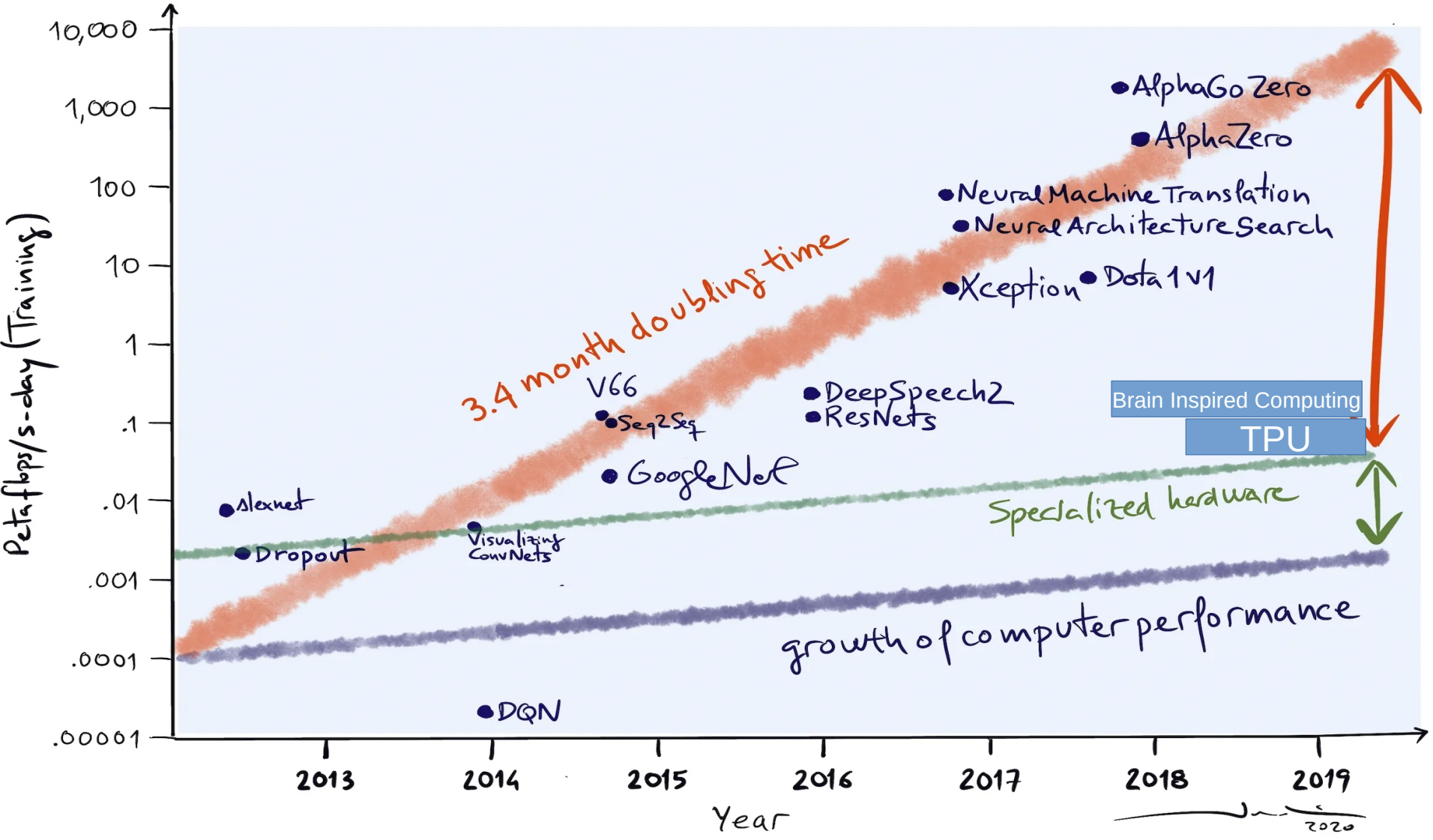
Processor Chip: Moore's law



AI and Computing Performance Gap



AI and Specialized Accelerators Performance Gap

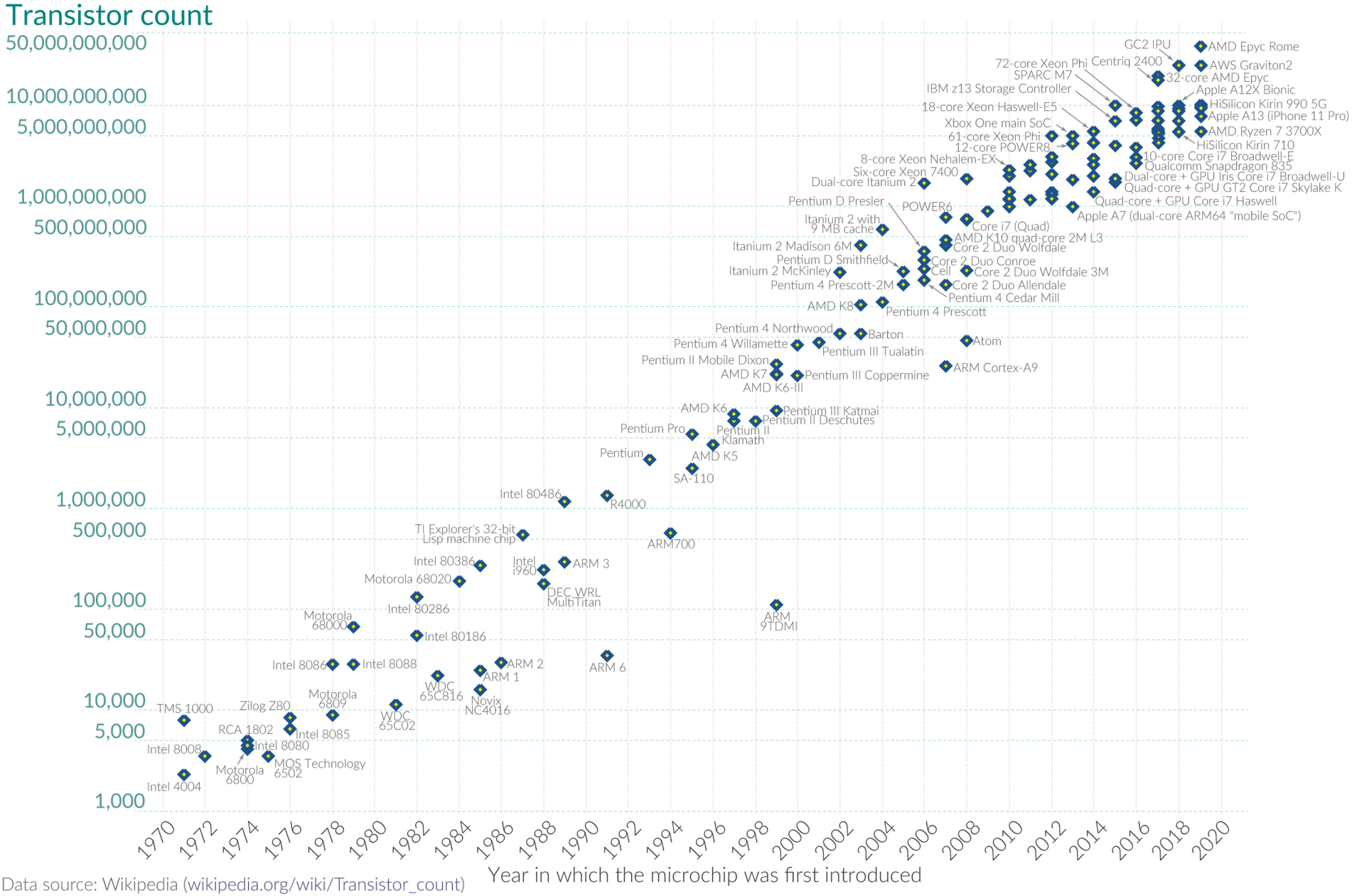


- Speaker Introduction
- Objectives of this Event
- Mankind Progress
- Age of Big Data
- Artificial Intelligence Past Present and Future Challenges
- **Untangling Artificial Intelligence Problems with Supercomputing**
- Namal Supercomputing Facility

Moore's Law: The number of transistors on microchips doubles every two years

Our World in Data

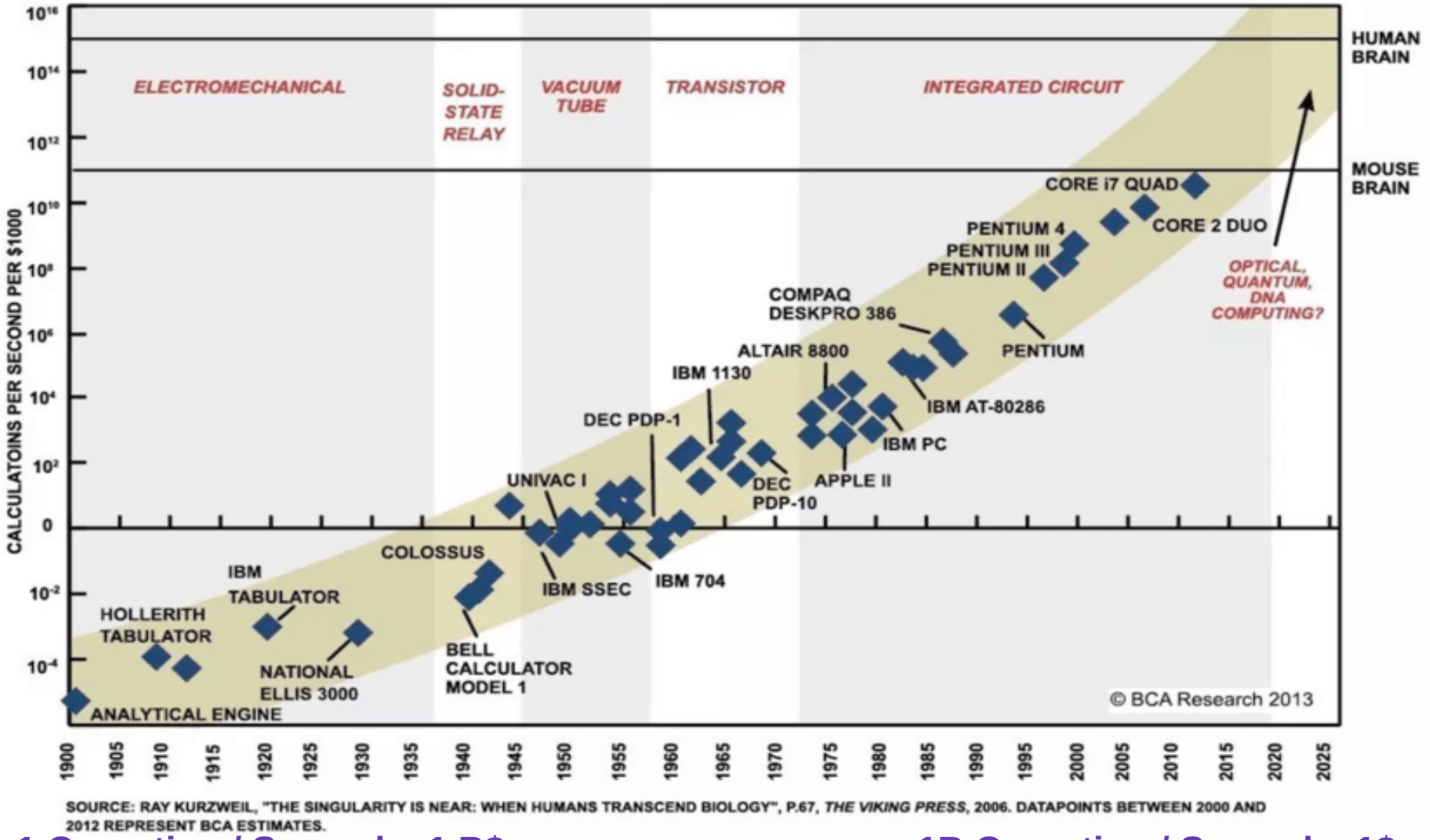
Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.



Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)

OurWorldinData.org – Research and data to make progress against the world's largest problems.

Cost Vs Performance: Electromechanical to ICs

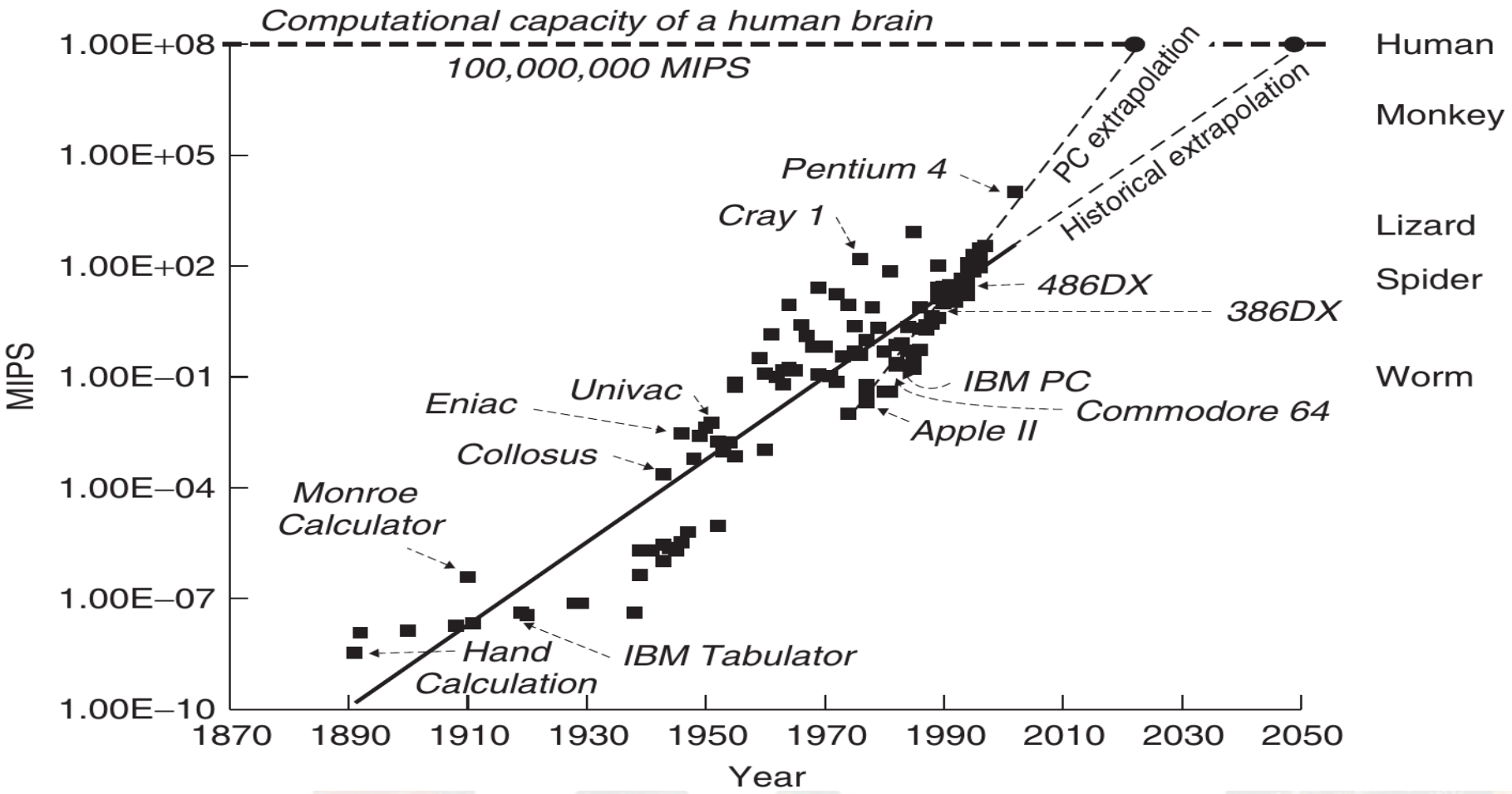


SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, *THE VIKING PRESS*, 2006. DATAPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.

1 Operation / Second = 1 B$

1B Operation / Second < 1$

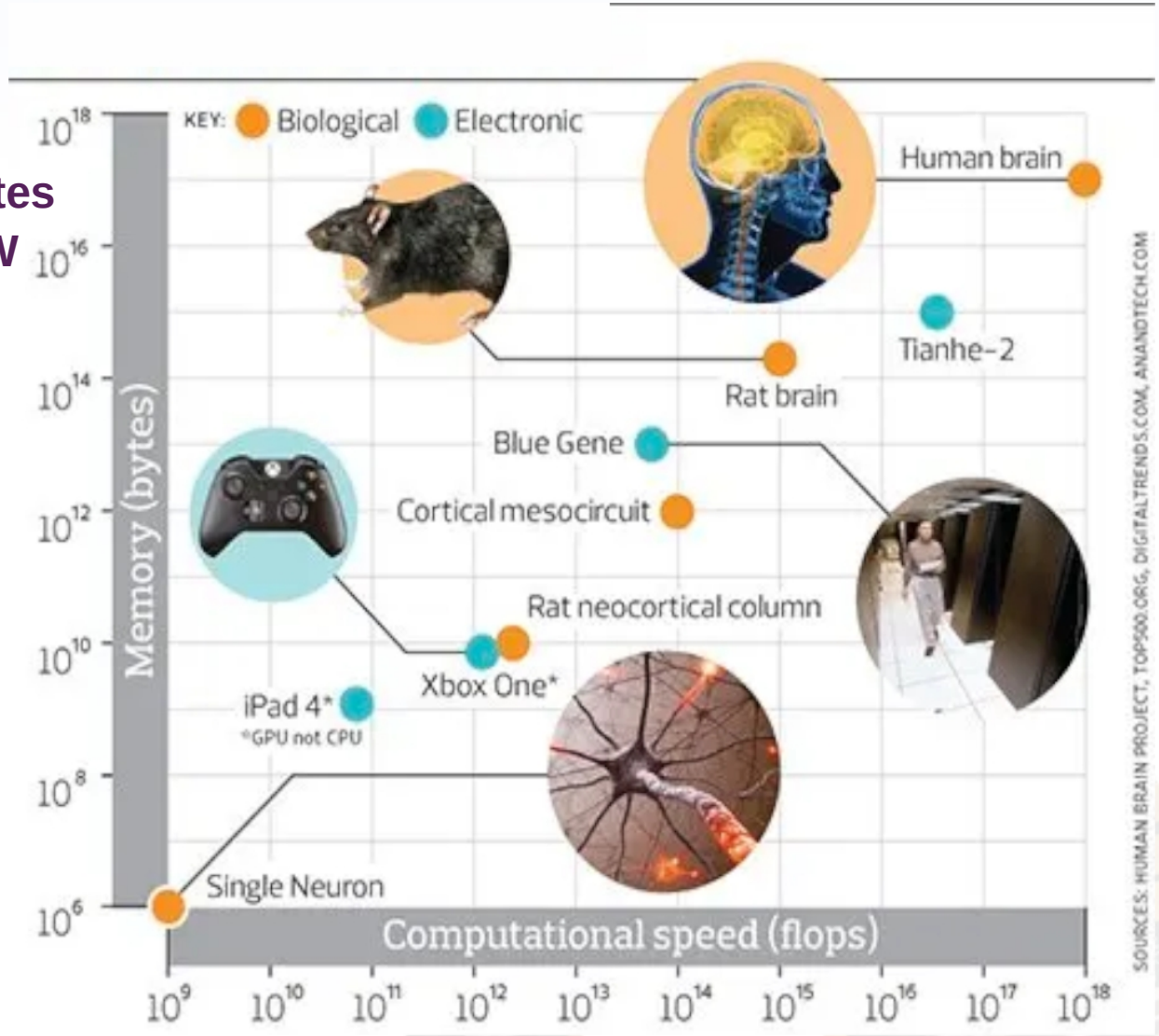
Computational Capability ?



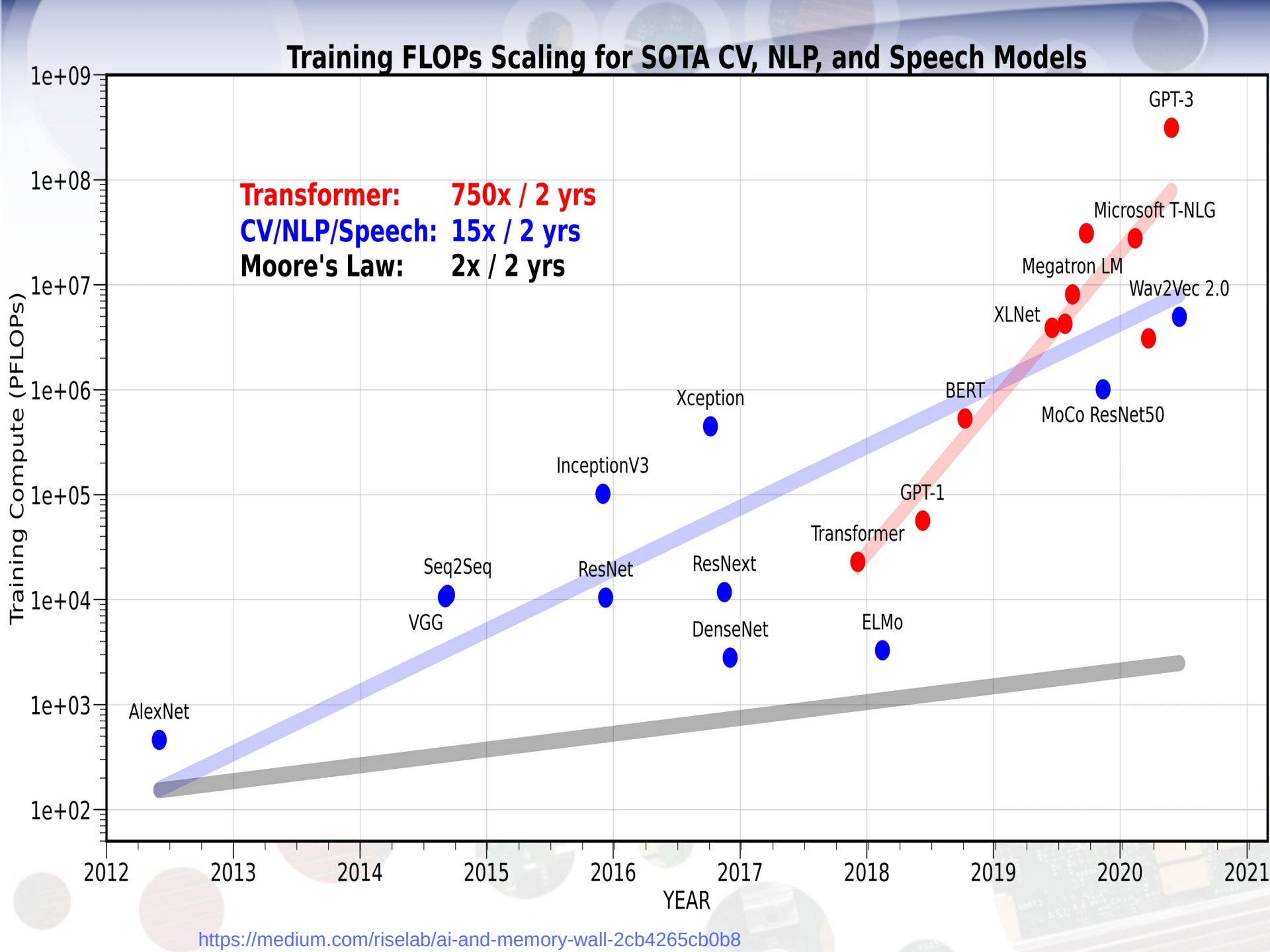
It is estimated that sometime between the years **2025 and 2050**, a **personal computers** will exceed the calculation power of a human brain.

Compute Vs Intellectual Capability

- **Human Brain:**
- **1 Exa FLOPS**
- **Memory 100 Peta bytes**
- **100 Tera OPS @ 10 W**

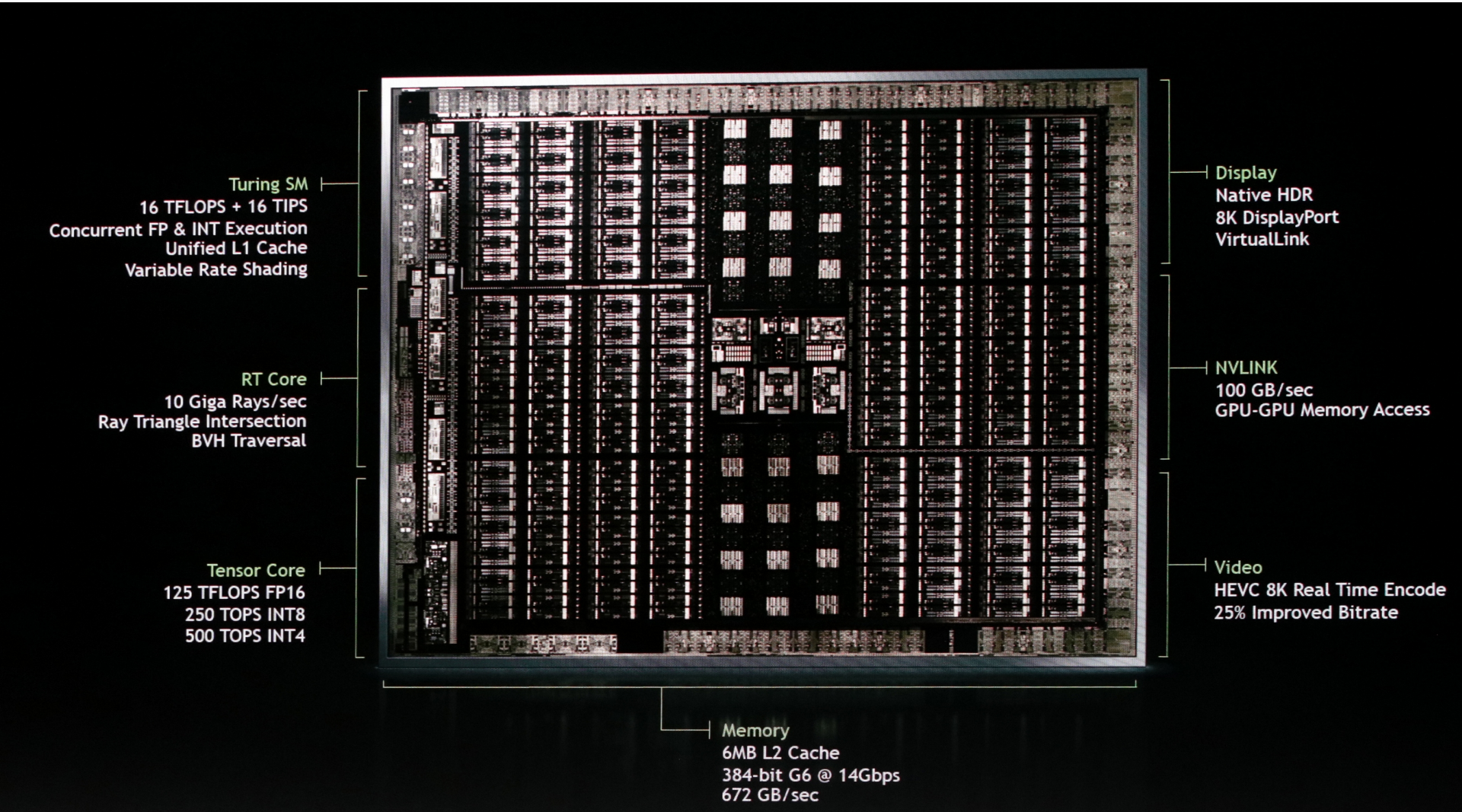


https://www.theguardian.com

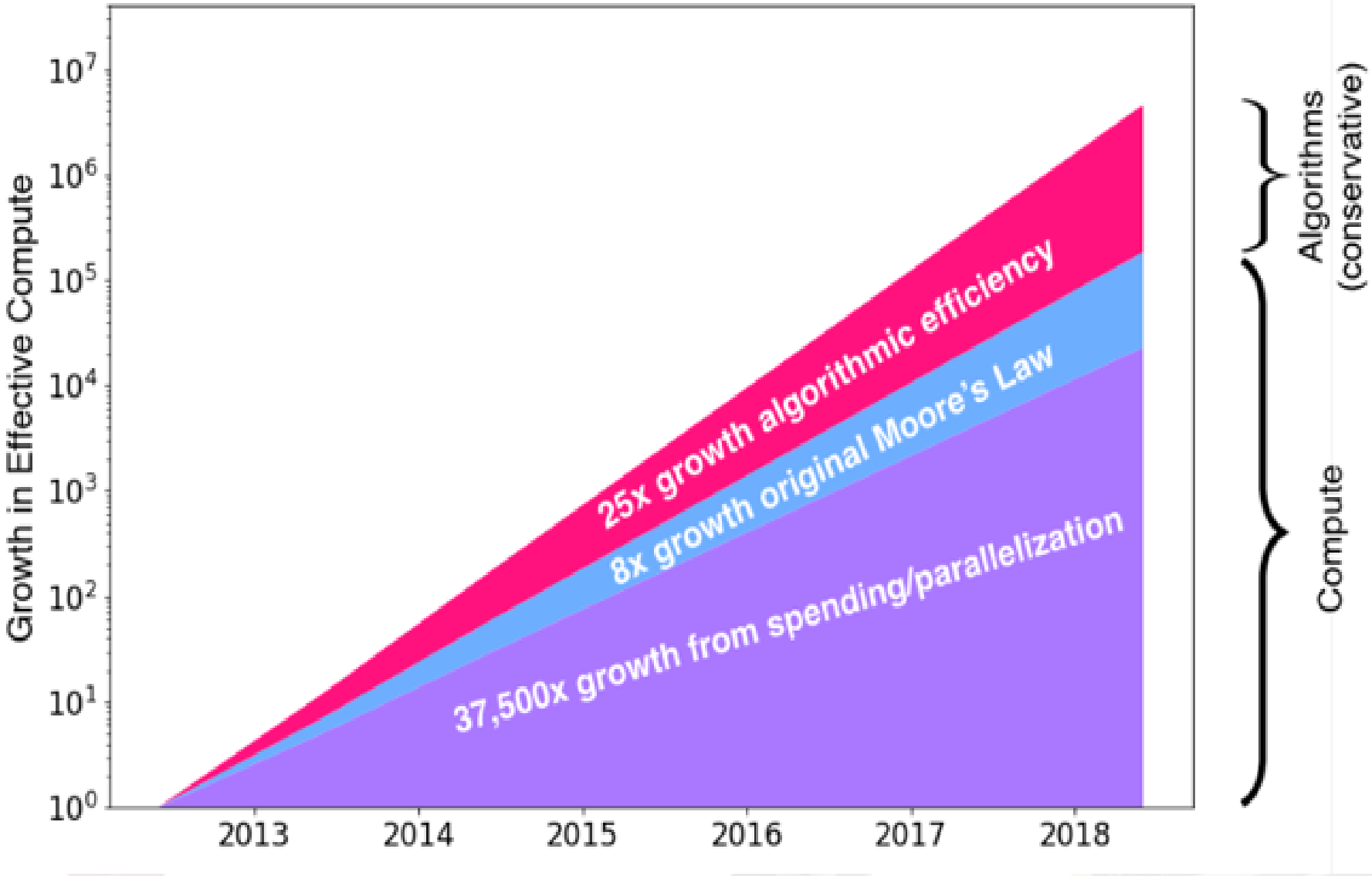


https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8

NVIDIA Next-Gen Turing GPU Architecture:



Challenges: AI Efficiency Growth



https://arxiv.org/pdf/2005.04305.pdf

ChatGPT: Use Case

- The cost to train 175 billion parameter language model is nearly $12 million dollars (200x the price of GPT-2)

OpenAI GPT-3 paper.

ChatGPT: Use Case

- The cost to train 175 billion parameter language model is nearly $12 million dollars (200x the price of GPT-2)
- Google presents a model for Multilingual translation quality with 600 Billion parameters. The training takes 22 Years on 1TPU.

OpenAI GPT-3 paper.

ChatGPT: Use Case

- The cost to train 175 billion parameter language model is nearly $12 million dollars (200x the price of GPT-2)
- Google presents a model for Multilingual translation quality with 600 Billion parameters. The training takes 22 Years on 1TPU.
- While distributing the training over 2048 TPUs, achived results in 4 days.

OpenAI GPT-3 paper.

Can the bottlenecks be alleviated with new designs by taking advantage of HPC technologies?

Can RDMA-enabled high-performance interconnects benefit big data processing and deep learning?

Can HPC clusters with high-performance storage systems (e.g., SSD, parallel file systems) benefit big data and deep learning applications?

How much performance benefits can be achieved through enhanced designs?

What are the major bottlenecks in current big data processing and deep learning middleware (e.g., Hadoop, Spark)?

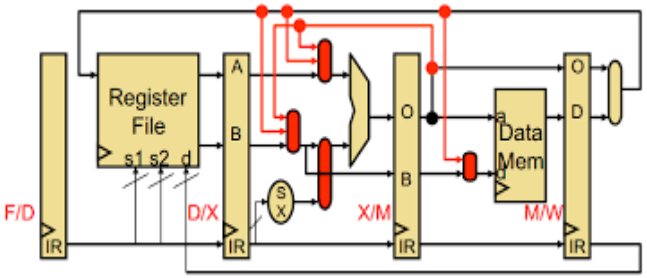
How to design benchmarks for evaluating the performance of big data and deep learning middleware on HPC clusters?

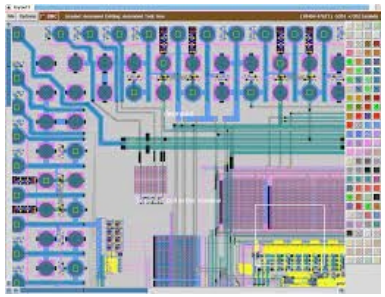
Bring HPC, big data processing, and deep learning into a "convergent trajectory"!

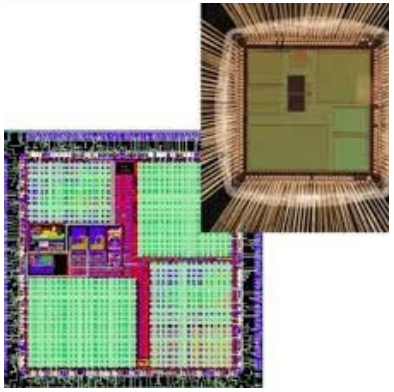
- Speaker Introduction
- Objectives of this Event
- Mankind Progress
- Age of Big Data
- Artificial Intelligence Past Present and Future Challenges
- Untangling Artificial Intelligence Problems with Supercomputing
- **Namal Supercomputing Facility**

Microelectronics Solutions for AI Compute Capability

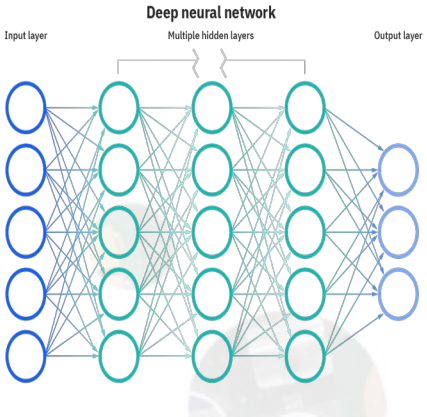
- **OpenSource Full-Stack Ecosystem for RISC-V Processor System**





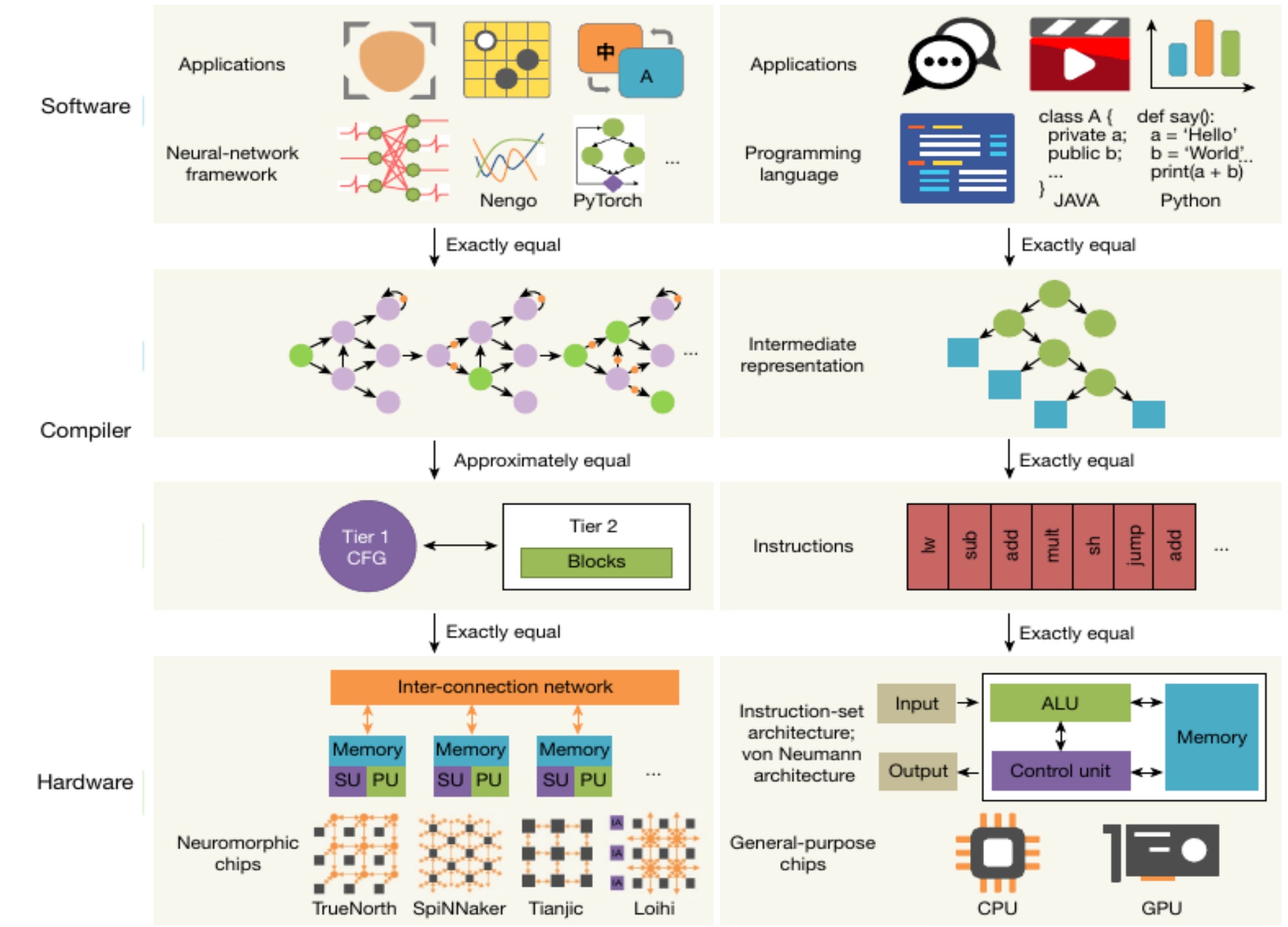


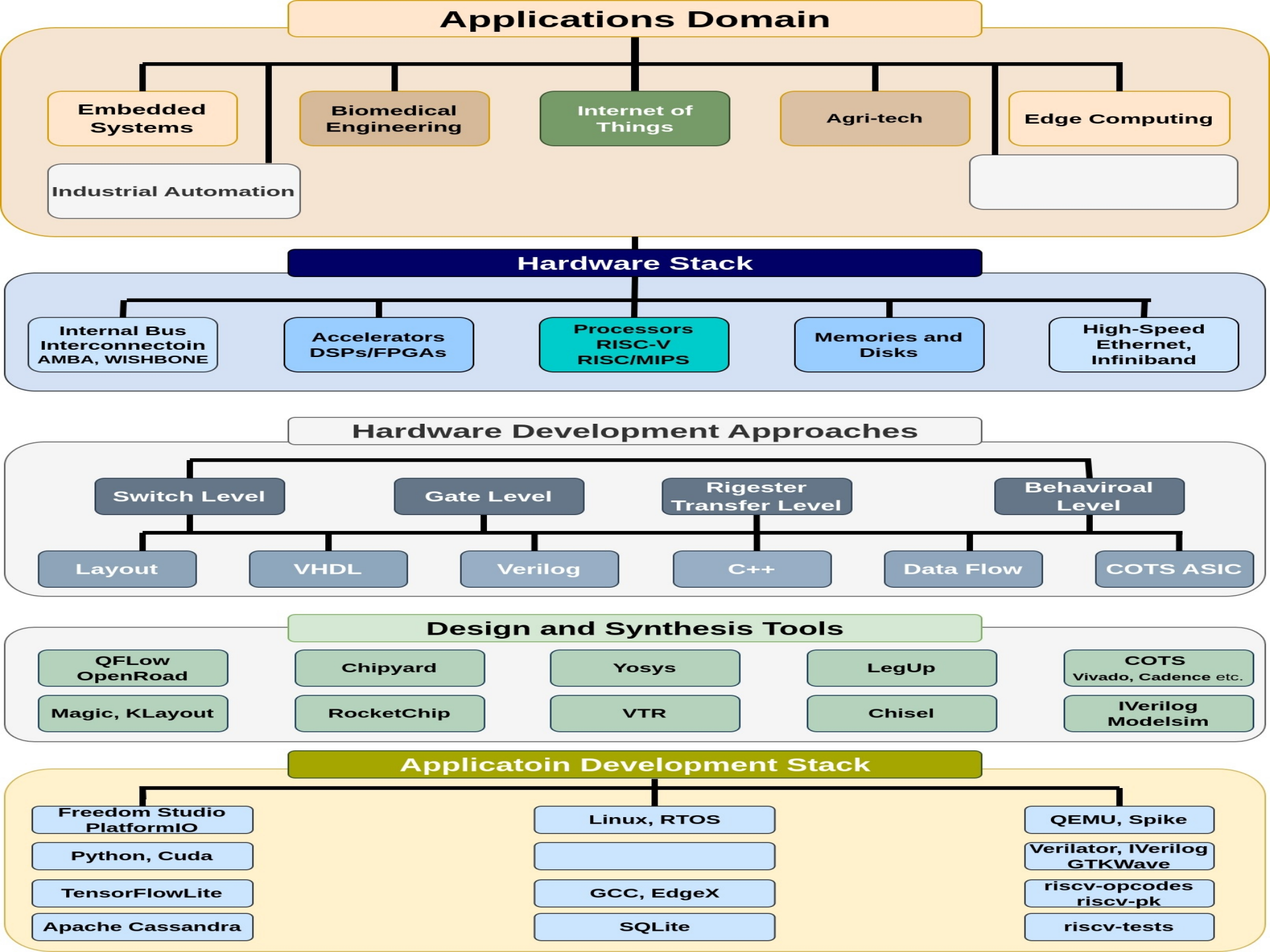
- **Supercomputing for AI and BigData Applications**



OpenSource Full-Stack Ecosystem for RISC-V Processor Architecture

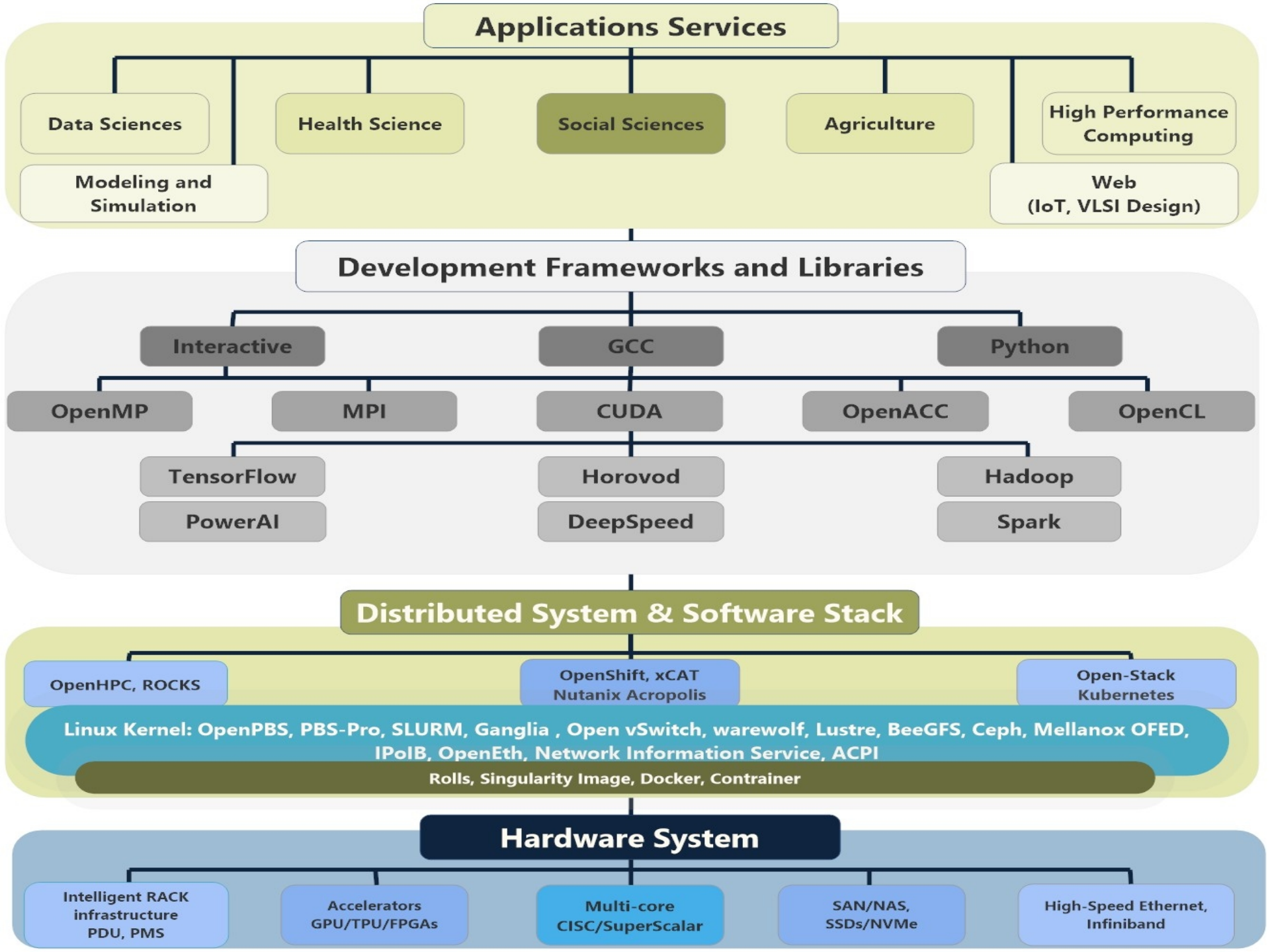
- **Hardware Architecture**
  - Low Performance and Low Cost Digital System
  - Uni/Multi Core System on a Chip
- **Single Board Computer**
  - Hardware Software Co-Design
  - High Performance Computing
- **Intelligent and Real-time Applications**
  - Industrial Automation
  - Machine Learning


Software
Applications
Neural-network framework
Nengo
PyTorch
...
Applications
Programming language
class A {
private a;
public b;
...
}
JAVA
def say():
a = 'Hello'
b = 'World'
print(a + b)
Python
Exactly equal
Exactly equal
Compiler
...
Intermediate representation
Approximately equal
Exactly equal
Tier 1 CFG
Tier 2
Blocks
Instructions
lw
sub
add
mult
sh
jump
add
...
Exactly equal
Exactly equal
Hardware
Inter-connection network
Memory
SU
PU
Memory
SU
PU
Memory
SU
PU
...
Neuromorphic chips
TrueNorth
SpiNNaker
Tianjic
Loihi
Instruction-set architecture; von Neumann architecture
Input
Output
ALU
Control unit
Memory
General-purpose chips
CPU
GPU


Applications Domain
Embedded Systems
Biomedical Engineering
Internet of Things
Agri-tech
Edge Computing
Industrial Automation
Hardware Stack
Internal Bus Interconnectoin AMBA, WISHBONE
Accelerators DSPs/FPGAs
Processors RISC-V RISC/MIPS
Memories and Disks
High-Speed Ethernet, Infiniband
Hardware Development Approaches
Switch Level
Gate Level
Rigester Transfer Level
Behaviroal Level
Layout
VHDL
Verilog
C++
Data Flow
COTS ASIC
Design and Synthesis Tools
QFLow OpenRoad
Chipyard
Yosys
LegUp
COTS Vivado, Cadence etc.
Magic, KLayout
RocketChip
VTR
Chisel
IVerilog Modelsim
Applicatoin Development Stack
Freedom Studio PlatformIO
Linux, RTOS
QEMU, Spike
Python, Cuda
Verilator, IVerilog GTKWave
TensorFlowLite
GCC, EdgeX
riscv-opcodes riscv-pk
Apache Cassandra
SQLite
riscv-tests

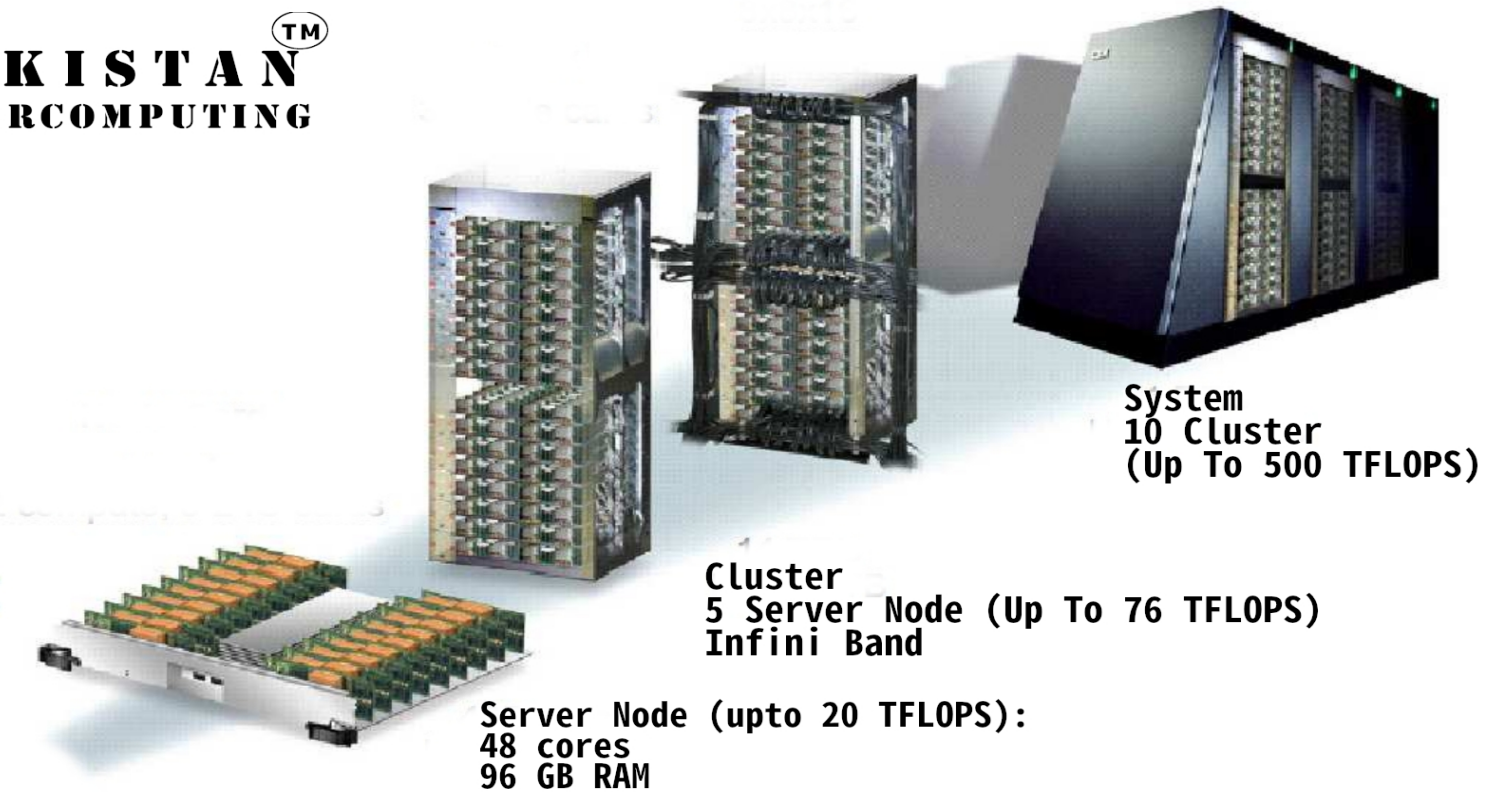
Supercomputing Platform for AI and BigData Applications

- **Bare-Metal and Containerized Cluster Infrastructure:**
  - Distributed Hardware Interfacing, Network Configuration and Distributed Computing Software Deployment
- **Data Center and Cloud Infrastructure:**
  - Storage systems, networking equipment, and software configuration
- **AI Applications for Scientific and Engineering Problems**
  - Distributed AI applications for multi-node bare-metal system
- **HPC Application Parallel Programming**
  - Heterogeneous multi-node parallel processing using parallel programming models


Applications Services
Data Sciences
Health Science
Social Sciences
Agriculture
High Performance Computing
Modeling and Simulation
Web (IoT, VLSI Design)
Development Frameworks and Libraries
Interactive
GCC
Python
OpenMP
MPI
CUDA
OpenACC
OpenCL
TensorFlow
Horovod
Hadoop
PowerAI
DeepSpeed
Spark
Distributed System & Software Stack
OpenHPC, ROCKS
OpenShift, xCAT Nutanix Acropolis
Open-Stack Kubernetes
Linux Kernel: OpenPBS, PBS-Pro, SLURM, Ganglia , Open vSwitch, warewolf, Lustre, BeeGFS, Ceph, Mellanox OFED, IPoIB, OpenEth, Network Information Service, ACPI
Rolls, Singularity Image, Docker, Contrainer
Hardware System
Intelligent RACK infrastructure PDU, PMS
Accelerators GPU/TPU/FPGAs
Multi-core CISC/SuperScalar
SAN/NAS, SSDs/NVMe
High-Speed Ethernet, Infiniband

Developing Supercomputing for AI

PAKISTAN SUPERCOMPUTING™



System
10 Cluster
(Up To 500 TFLOPS)

Cluster
5 Server Node (Up To 76 TFLOPS)
Infini Band

Server Node (upto 20 TFLOPS):
48 cores
96 GB RAM
1 TB Disk
2 GPUs

CentOS Linux

Chip
4 cores
XEON Processor

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

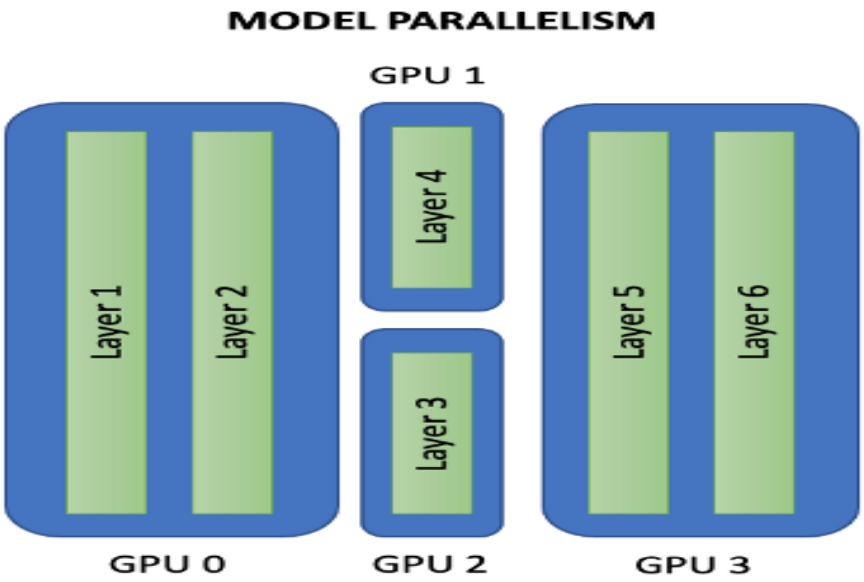
AI Model Parallelism

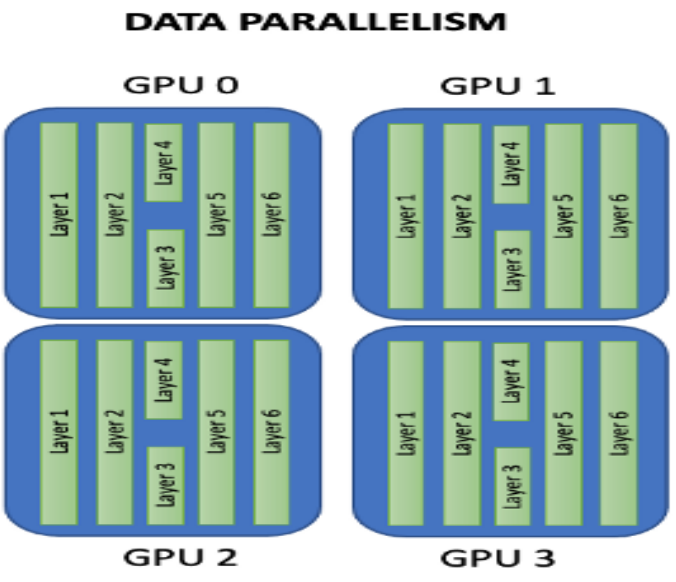
- **Model Parallelism**

  Different layers of the network distributed across different devices

- **Data Parallelism**

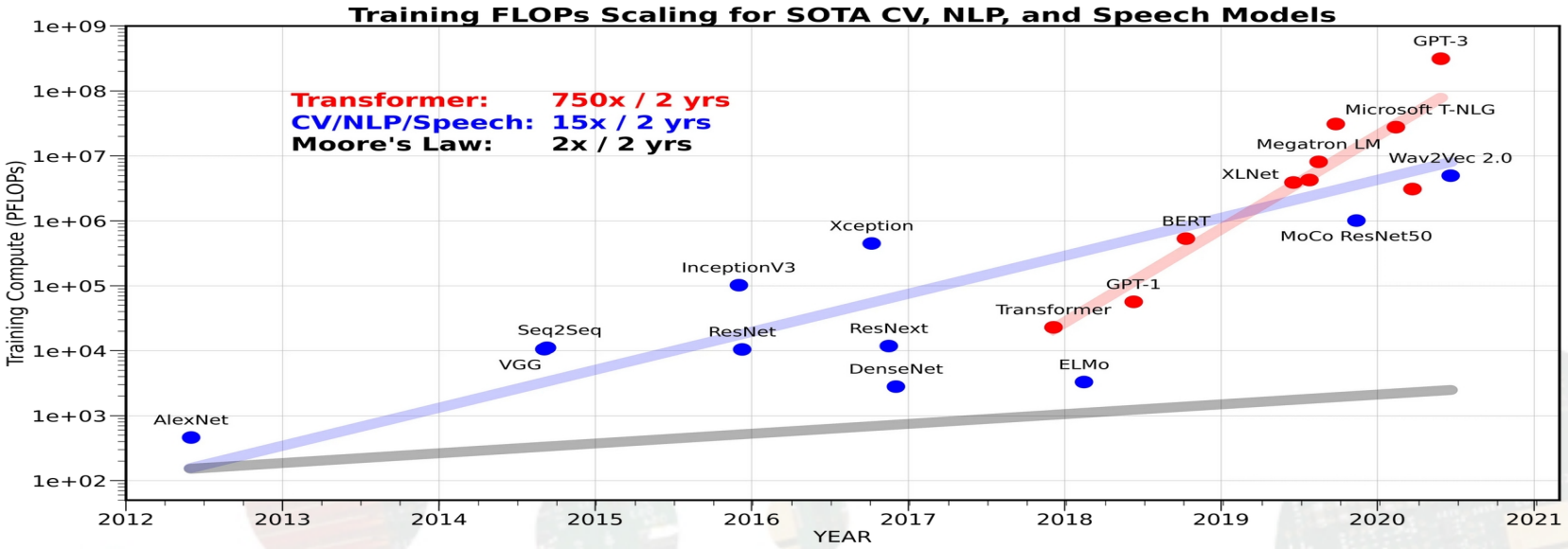
  Same model in every one of the GPUs, each processing a separate piece of the data, a separate portion of the mini-batch.





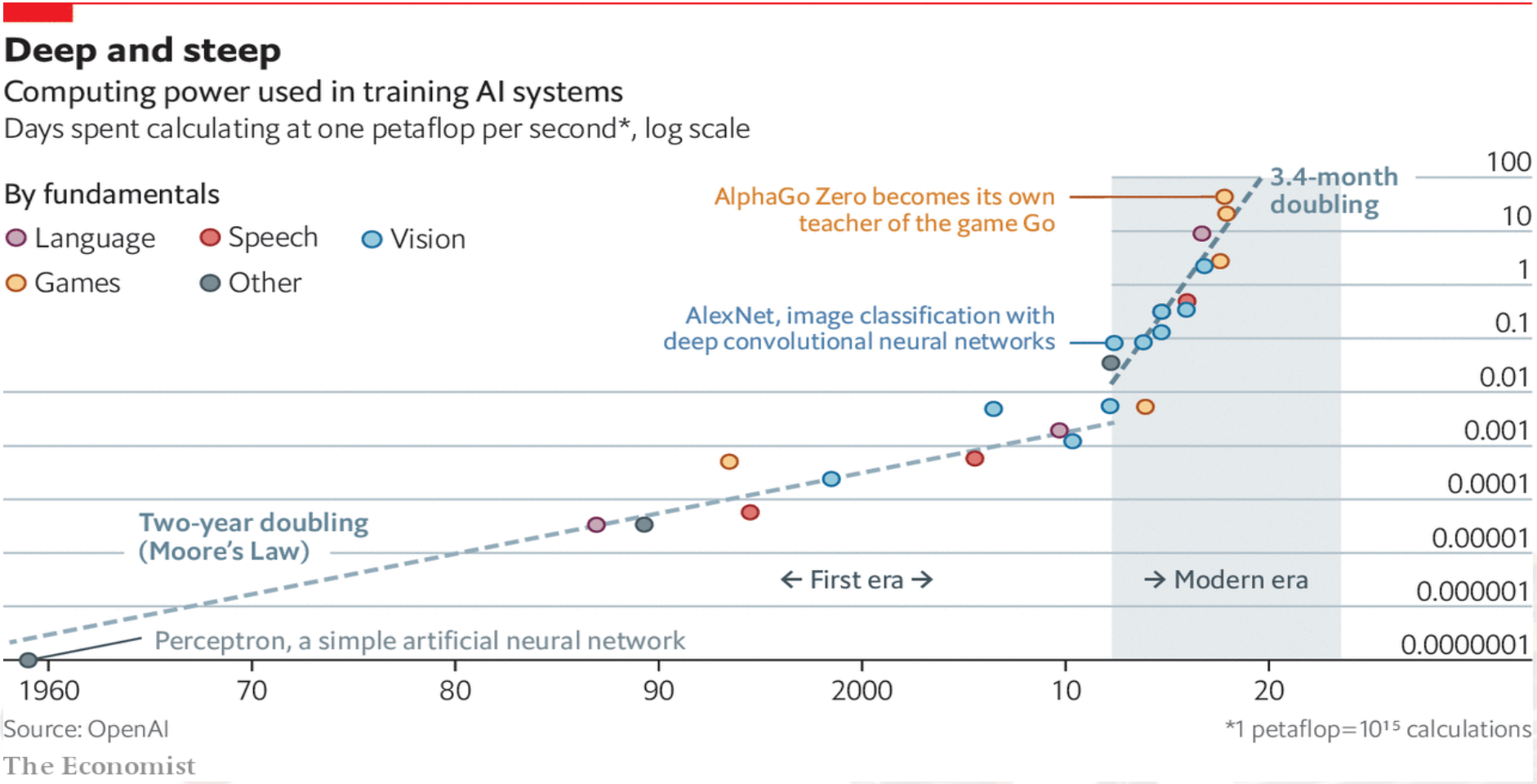
Conclusion

- World Data Size = 200 Zettabytes, doubling every 18 months.
- The computational demands of AI algorithms are experiencing exponential growth.
- The performance of microelectronics is not increasing in line with Moore's Law.



PAKISTAN SUPERCOMPUTING

AI is the need of the day and is definitely penetrating society, like electricity.



Center of Excellence:
Supercomputing for

Supercomputing: Untangling Big Data and Artificial Intelligence Problems

by: Tassadaq Hussain

Professor Department of Electrical Engineering

Namal University Mianwali

Collaborations:

Barcelona Supercomputing Center, Spain

European Network on High Performance and Embedded Architecture and Compilation

Pakistan Supercomputing Center

- Speaker Introduction
- Objectives of this Event
- Mankind Progress
- Age of Big Data
- Artificial Intelligence Past Present and Future Challenges
- Untangling Artificial Intelligence Problems with Supercomputing
- Namal Supercomputing Facility