Big data management

Intro & course logistics

Course staff

Zoi

Maria

Alexandru

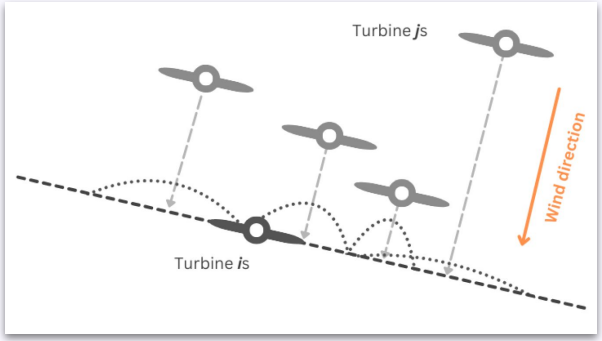
Richard

- Scalable algorithms for data intensive applications
- Applications related to sustainability
- Timeseries forecasting, graph neural networks

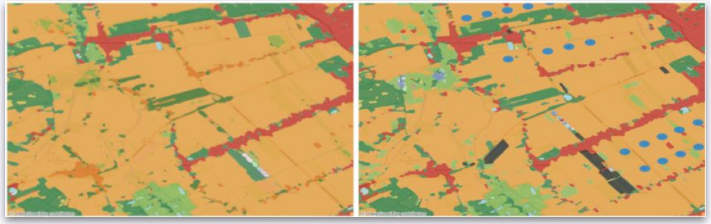
Maria Astefanoaei
Assistant professor

https://mariaast.github.io/
msia@itu.dk

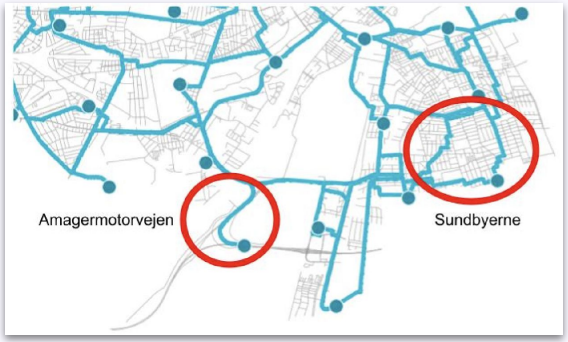
Projects supervision



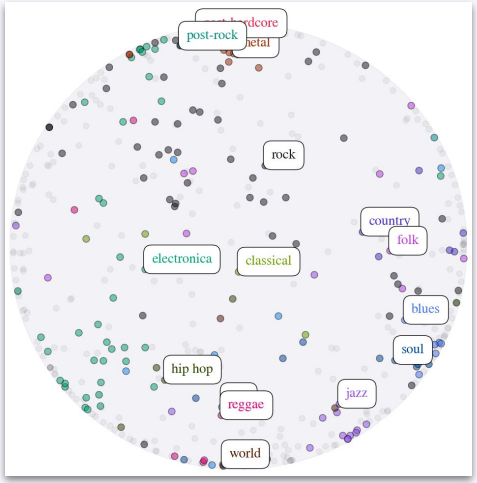
graph neural networks for predicting wind energy



land use and land cover changes



cycling network expansion



interactive music genre exploration

- Data management for/with machine learning
- Unifying data systems (cross-platform data processing)
- Machine learning and knowledge graphs

Zoi Kaoudi

Associate Professor

http://itu.dk/~zoka/

zoka@itu.dk

Research Group at ITU:

Main site: https://dasya.itu.dk/

General info on courses and projects: https://dasya.itu.dk/for-students/

Proposals for projects: https://dasya.itu.dk/for-students/proposals/

People at a glance: https://dasya.itu.dk/people/

AI making headlines

Microsoft, OpenAI plan $100 billion data-center project, media report says

By **Reuters**

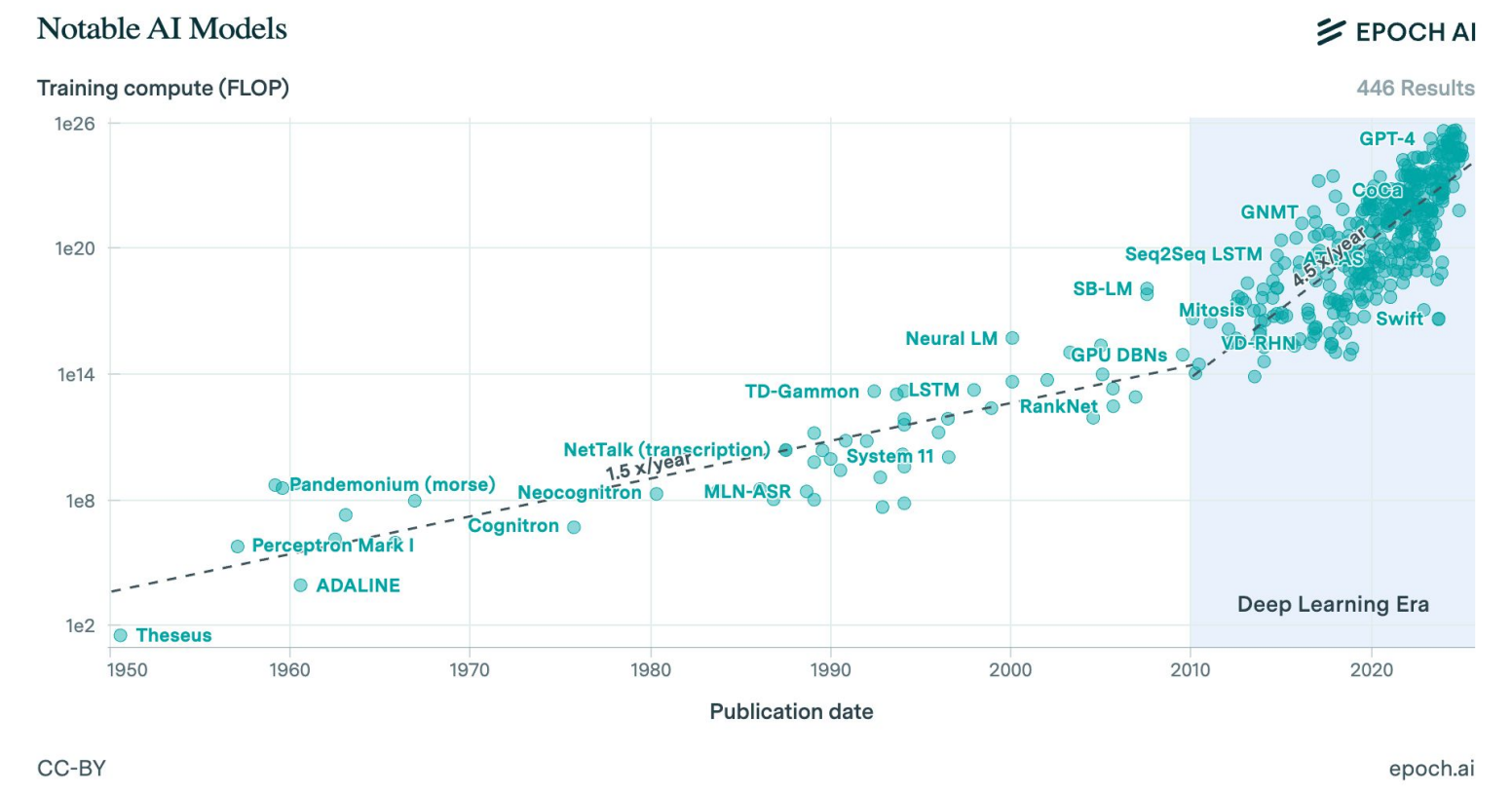
March 29, 2024 10:14 PM GMT+1 · Updated 10 months ago

Google to Build Seven Nuclear Reactors to Power Artificial Intelligence Systems

HEADLINE OCT 16, 2024

The announcement comes just a month after Microsoft said it would fund the reopening of the Three Mile Island nuclear plant, the site of the worst nuclear disaster in U.S. history.

An ever-growing amount of floating point operations



The amount of compute used to train notable AI models has grown about 4-5x/year between 2010 and May 2024. Much of the growth comes from increased spending, but not only.

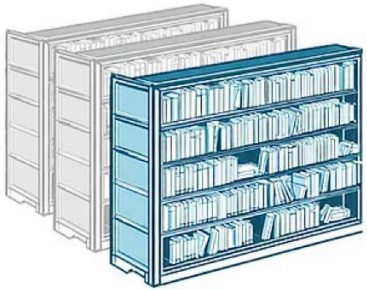
GPT4 Model Estimates

Training Size

of Book shelves for 13T tokens

650 kms

Long line of Library Shelves



100000 tokens per Book
100 Books per shelf
2 Shelves per meter

Compute Size

Compute time for 2.15 e25 FLOPs

7 million years

On mid-size Laptop (100GFLOPs)

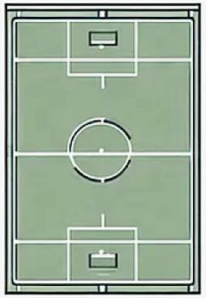
100GLOPs per second

Model Size

Size of Excel Sheet for 1.8T params

30,000

Football Fields sized Excel Sheet



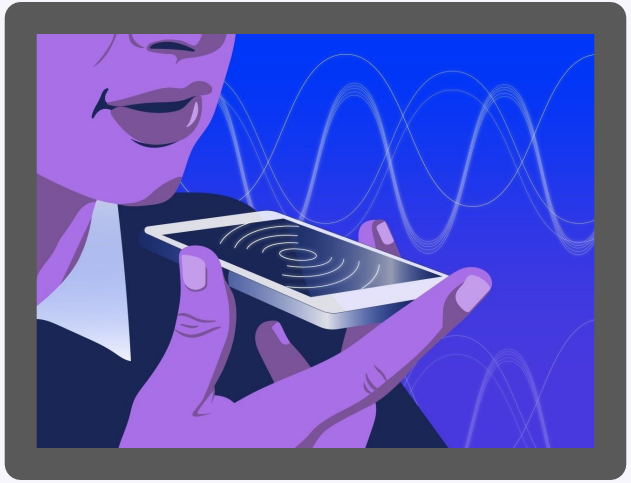
1x1 cm per Excel cell
100 x 60 meters Field Size

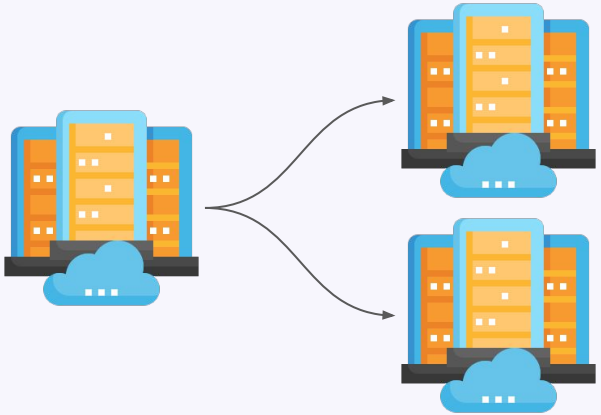
Source: https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked

Technology has come a long way

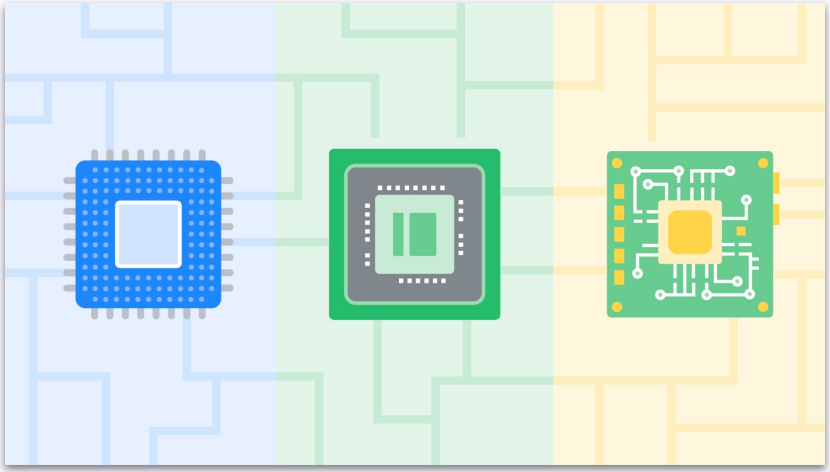
Around 2013 at  :

If Android users (a bit over 1 billion at the time) began to use Google's *voice search* for just **three minutes** a day it would require Google to **double its number of data centers**.





Advancements in hardware



Central Processing Units (CPUs) are designed as the jack-of-all-trades general-purpose “brains” for a computer.

Graphical processing units (GPUs), at the time, were specialized chips designed to work in tandem with a CPU to accelerate complex tasks in graphics, video rendering, and simulations.

Tensor processing units (TPUs) were built for AI; they are chips designed for a specific purpose: running matrix and vector based mathematics that’s needed for building and running AI models.

Ripple effects

Market Summary > NVIDIA Corp

142,32 USD

+136.06 (2,173.48%) ↑ past 5 years

24 Jan, 14.41 GMT-5 • Disclaimer

+ Follow

1D | 5D | 1M | 6M | YTD | 1Y | 5Y | Max

Open	148,37	Mkt cap	3,49T	52-wk high	153,13
High	148,97	P/E ratio	56,09	52-wk low	59,94
Low	141,94	Div yield	0,028 %		

Data centers, backbone of the digital economy, face water scarcity and climate risk

AUGUST 30, 2022 · 6:07 AM ET

By Michael Copley

link

Data centers have become integral to a global economy that's powered by digital information. However, many of the facilities depend on water to keep from overheating. That is further straining water resources in places like California, where Lake Oroville is almost dry due to severe drought that's being fueled by climate change.

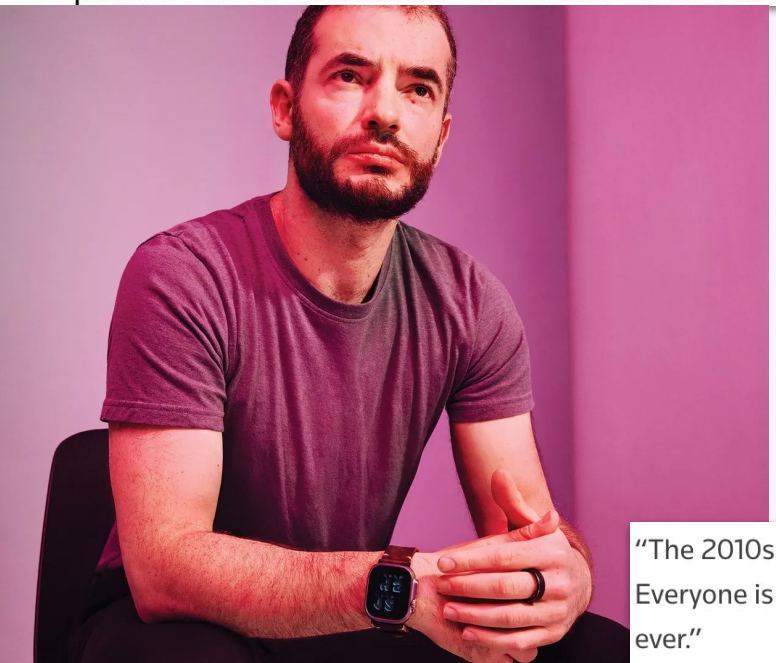
Justin Sullivan/Getty Images

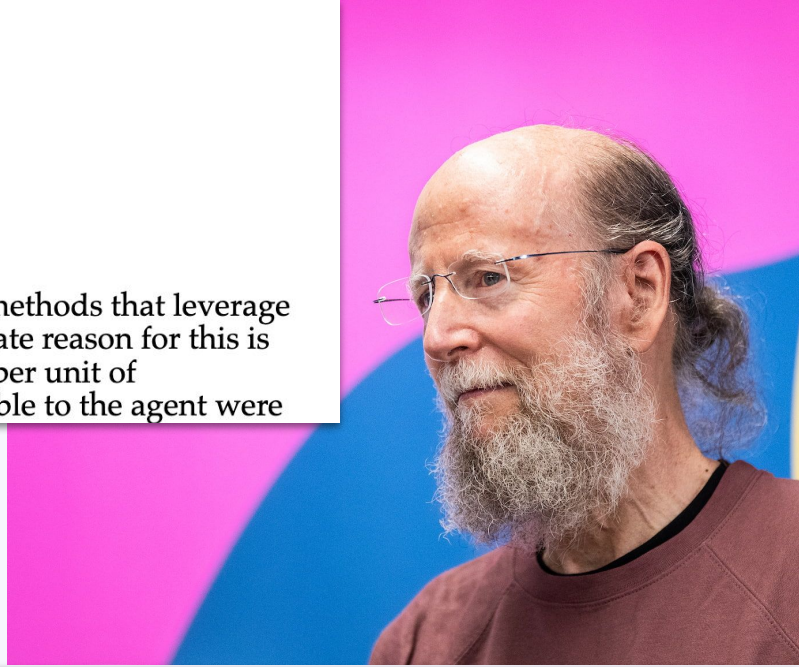
The Bitter Lesson

Rich Sutton

March 13, 2019

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were





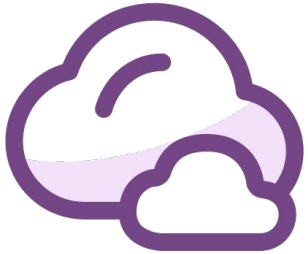
OpenAI and others seek new path to smarter AI as current methods hit limitations

By **Krystal Hu** and **Anna Tong**

November 15, 2024 10:11 AM GMT+1 · Updated 2 months ago

"The 2010s were the age of scaling, now we're back in the age of wonder and discovery once again. Everyone is looking for the next thing," Sutskever said. "Scaling the right thing matters more now than ever."

Big data



What are some challenges related to big data?

where do we store it?

what are the storage/compute costs?

what does it contain?

how can we analyse it?

how do we store it?

Big data

how to represent it?

what are the hidden costs?

is it representative?

do we need it?

is it safe? is it private?

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 Glossary -

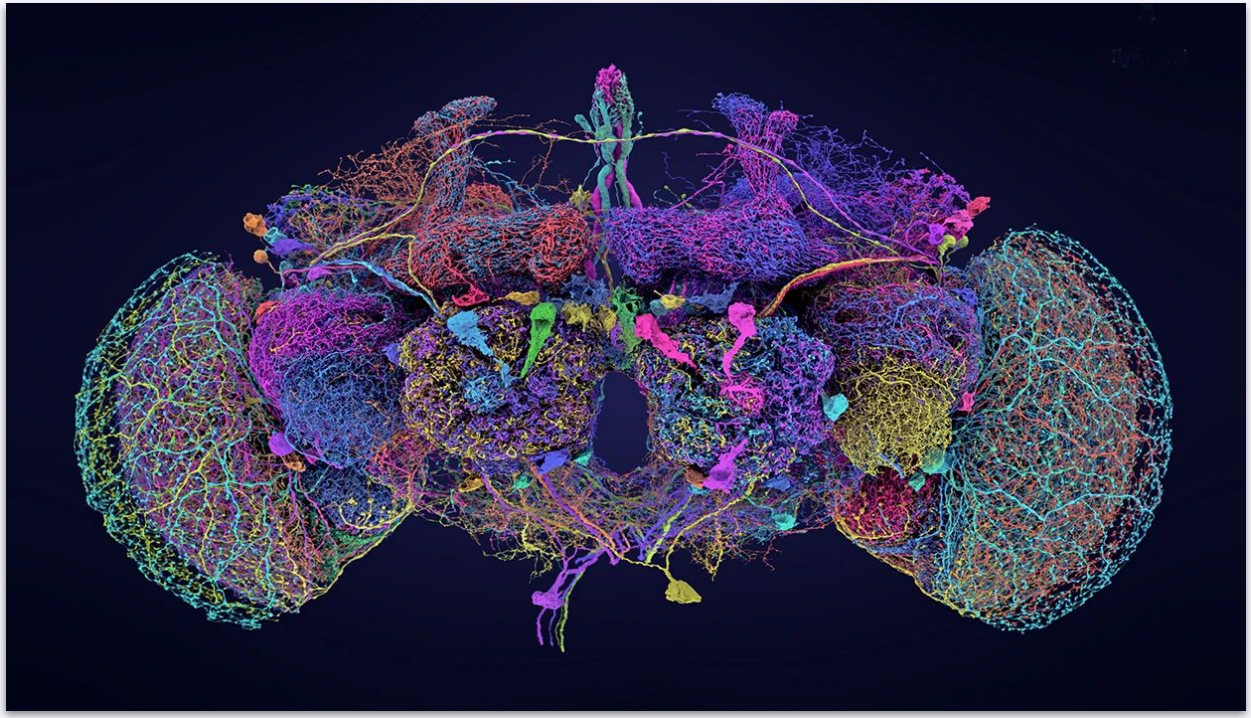
3 V's of big data

High volume data

Abbreviation	Unit	Value	Size (in bytes)	
b	bit	0 or 1	1/8	
B	byte	8 bytes	1	
KB	kilobyte	1024 bytes	~1000	
MB	megabyte	1024 KB	~1,000,000	*A small novel*
GB	gigabyte	1024 MB	~1,000,000,000	*0.5 hours of video*
TB	terabyte	1024 GB	~1,000,000,000,000	*all the X-rays in a large hospital*
PB	petabyte	1024 TB	~1,000,000,000,000,000	*2.5 years of non-stop movies*
EB	exabyte	1024 PB	~1,000,000,000,000,000,000	*A video call of 237,823 years*
ZB	zettabyte	1024 EB	~1,000,000,000,000,000,000,000	*Digital data in the world in 2012*
YB	yottabyte	1024 ZB	~1,000,000,000,000,000,000,000,000	*Atoms in 7,000 human bodies*

Mapping the brain of a fly

2024: the first complete map of any complex brain (neuron-by-neuron and synapse-by-synapse)

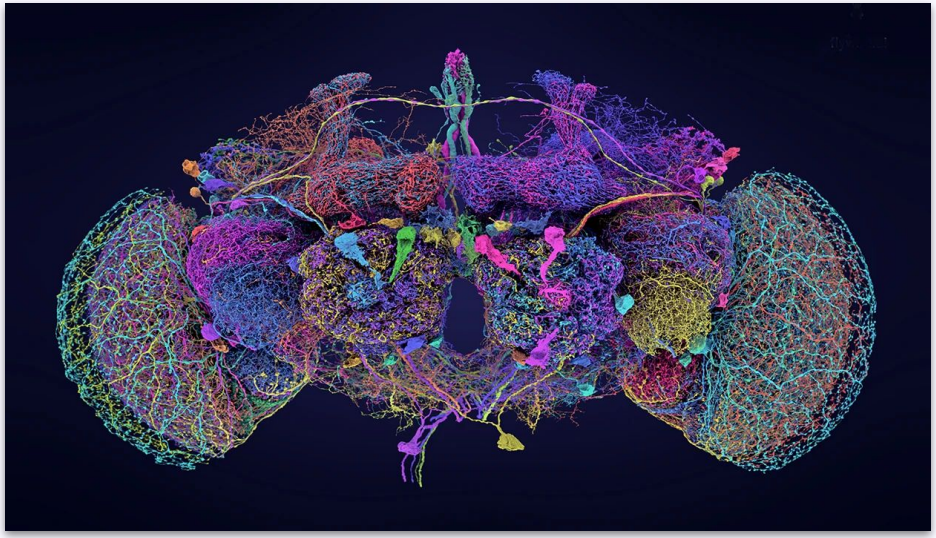


Mapping the brain of a fly

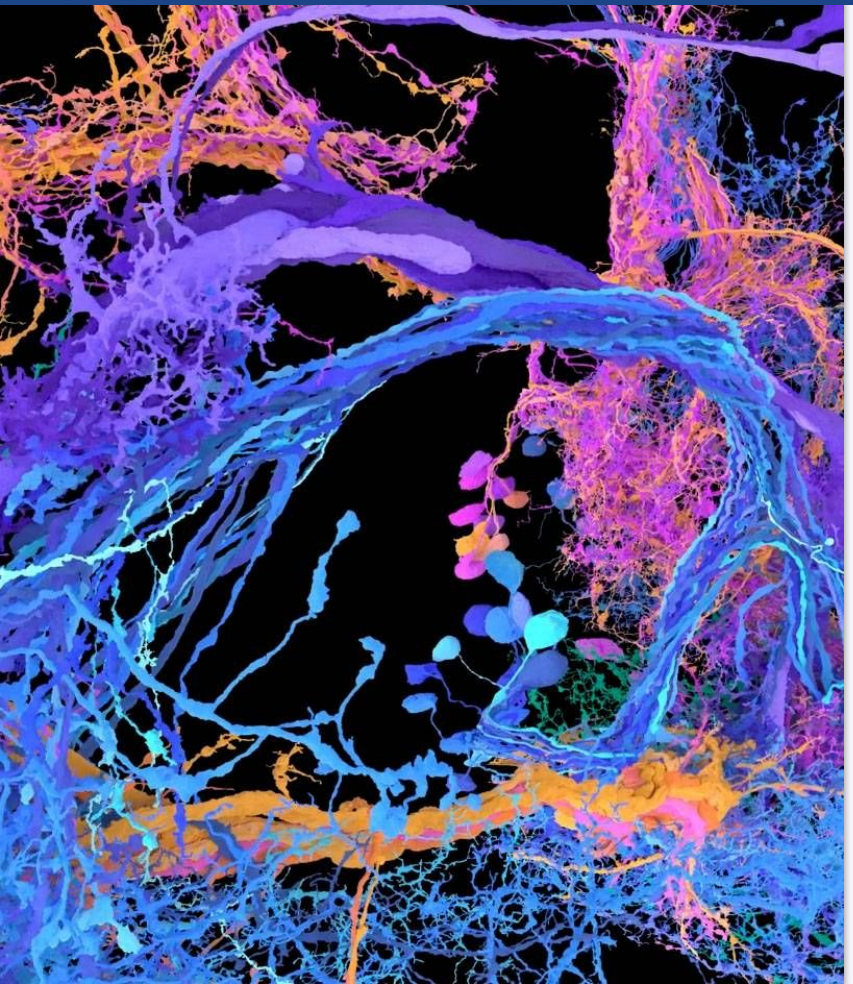
21 million images (electron microscopy)

+ AI model that recognized the cross-sections of neurons in each picture and stacked them into labeled 3-D shapes of the cells.

Work was done by the *FlyWire Consortium* - an unlikely collaboration among gamers, professional tracers, and neuroscientists who fixed over three million mistakes by hand.

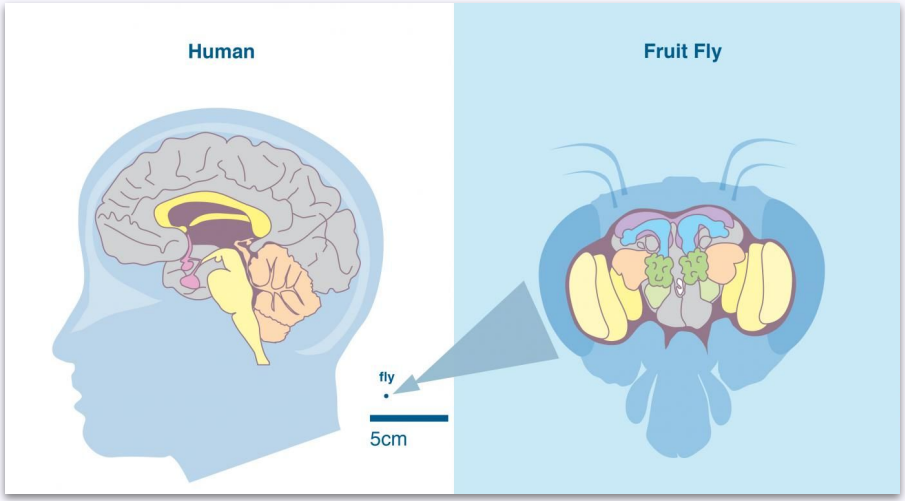


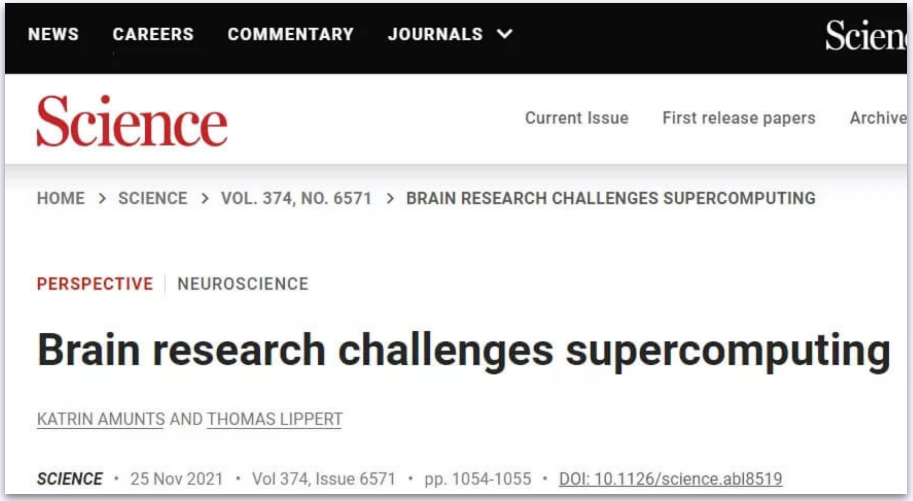
Towards an understanding of the human brain



- Fruit flies share 60% of human DNA, including genes for learning, Down syndrome and jet lag.
- 3 in 4 human genetic diseases have a parallel in fruit flies.
- Fruit flies age like we do, can get drunk, can be kept awake with coffee, and serenade their romantic interests.
- NASA sent fruit flies into space in 1947 — and they returned alive, paving the way for all astronauts.

Towards understanding the human brain



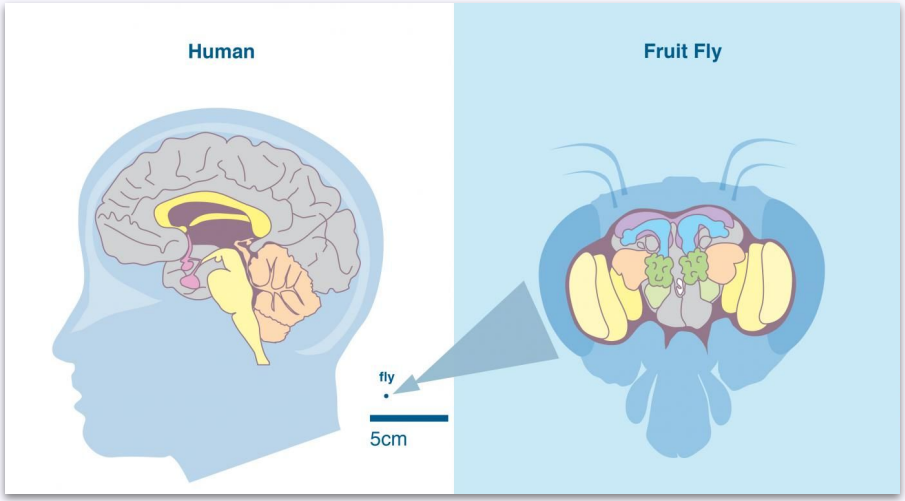


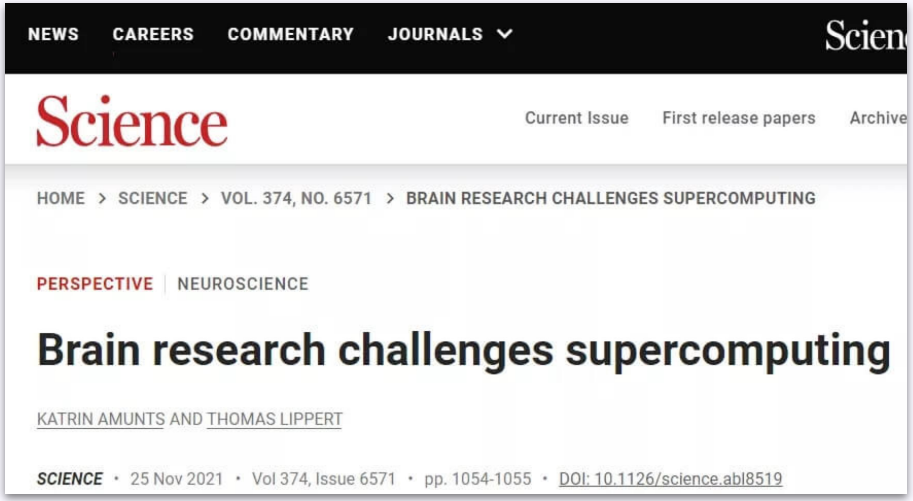
Fruit fly brain: 21 million images - more than 100 TB of data.

Electron microscopy of a whole brain would amount to more than one Exabyte of data.

TB	terabyte	1024 GB	~1,000,000,000,000	*all the X-rays in a large hospital*
PB	petabyte	1024 TB	~1,000,000,000,000,000	*2.5 years of non-stop movies*
EB	exabyte	1024 PB	~1,000,000,000,000,000,000	*A video call of 237,823 years*
ZB	zettabyte	1024 EB	~1,000,000,000,000,000,000,000	*Digital data in the world in 2012*
YB	yottabyte	1024 ZB	~1,000,000,000,000,000,000,000,000	*Atoms in 7,000 human bodies*

Towards understanding the human brain




NEWS CAREERS COMMENTARY JOURNALS

Science

Current Issue First release papers Archive

HOME > SCIENCE > VOL. 374, NO. 6571 > BRAIN RESEARCH CHALLENGES SUPERCOMPUTING

PERSPECTIVE | NEUROSCIENCE

Brain research challenges supercomputing

KATRIN AMUNTS AND THOMAS LIPPERT

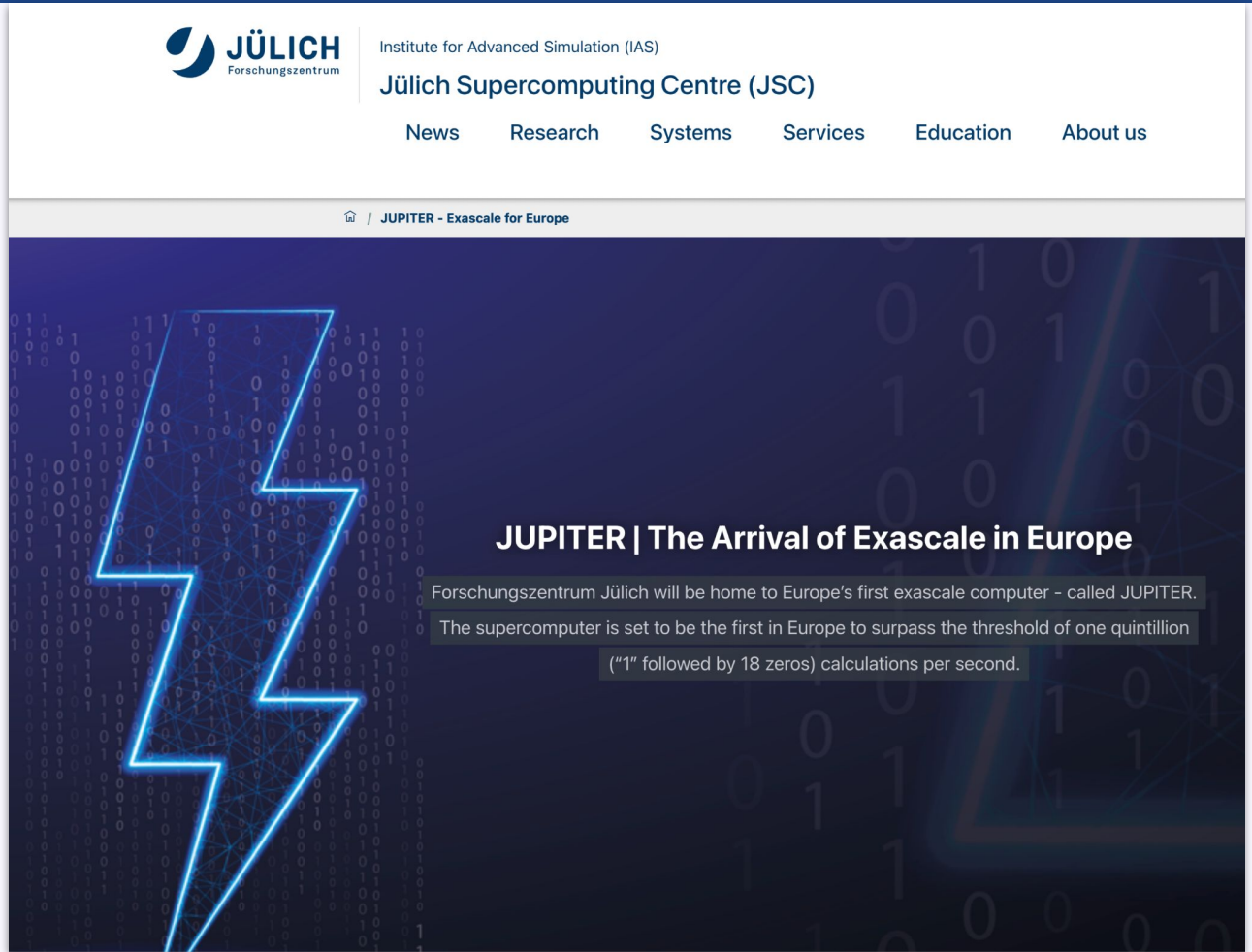
SCIENCE • 25 Nov 2021 • Vol 374, Issue 6571 • pp. 1054-1055 • DOI: 10.1126/science.abl8519

Fruit fly brain: 21 million images - more than 100 TB of data.

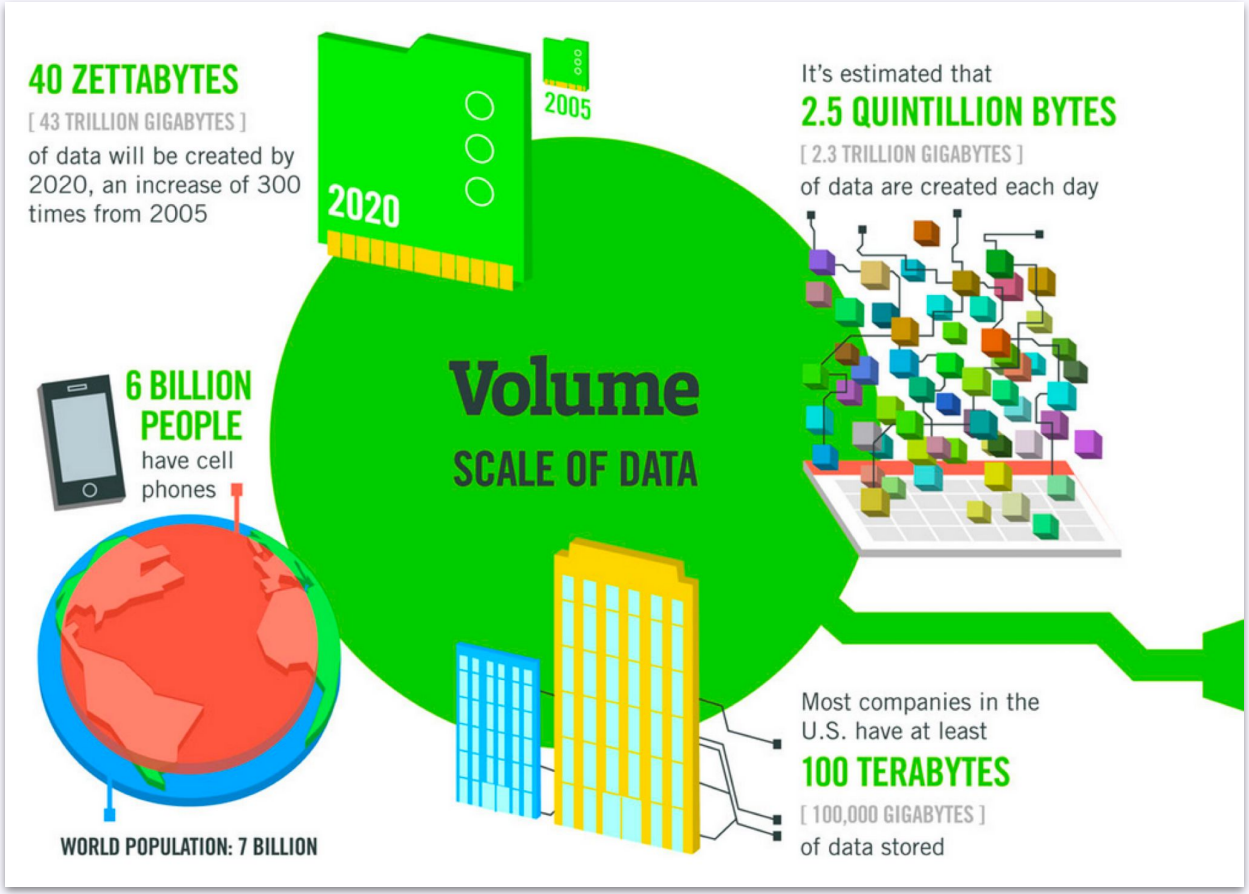
Electron microscopy of a whole brain would amount to more than one Exabyte of data.

How do we store such massive amounts of data?

Exascale computation

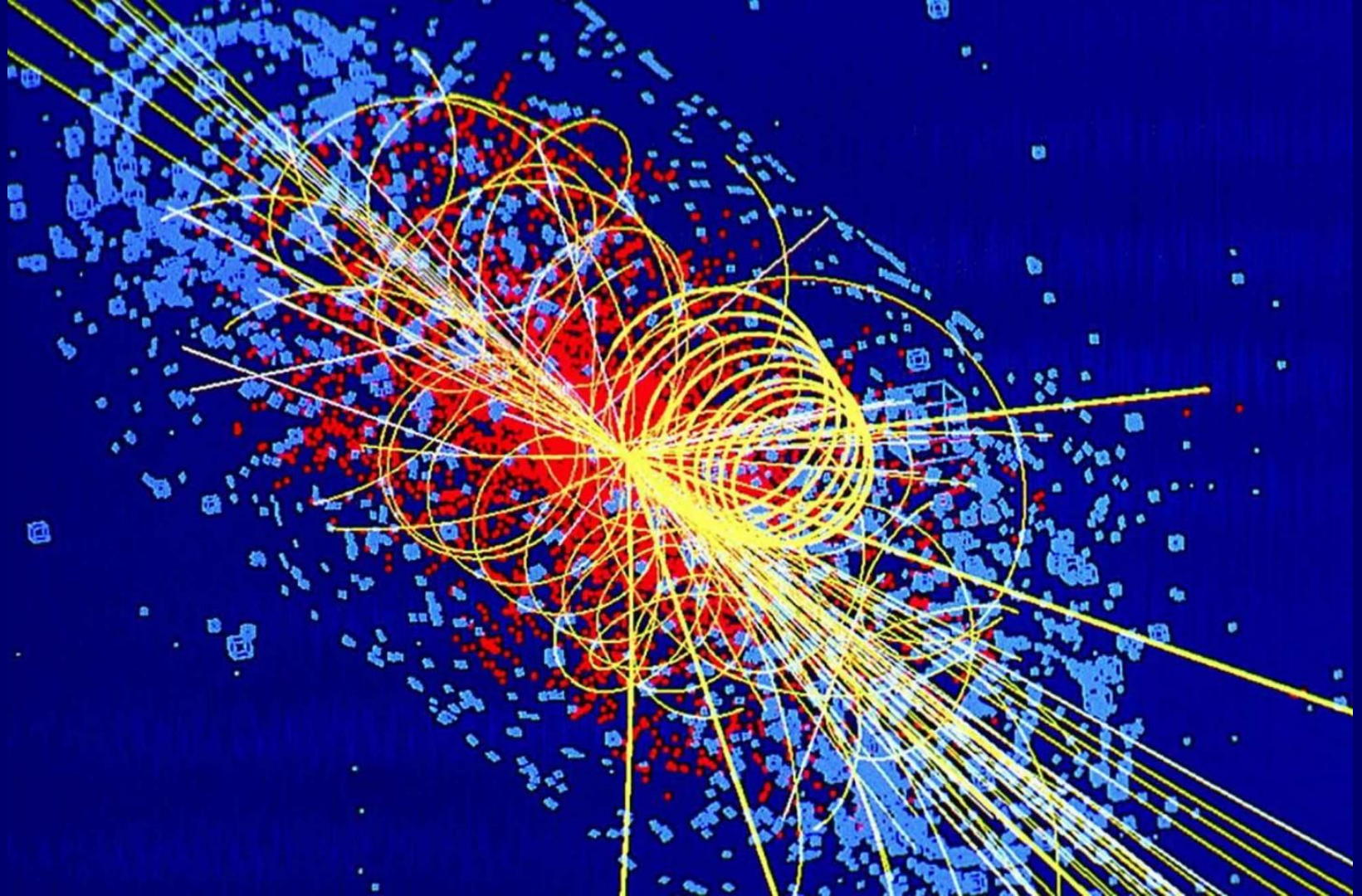


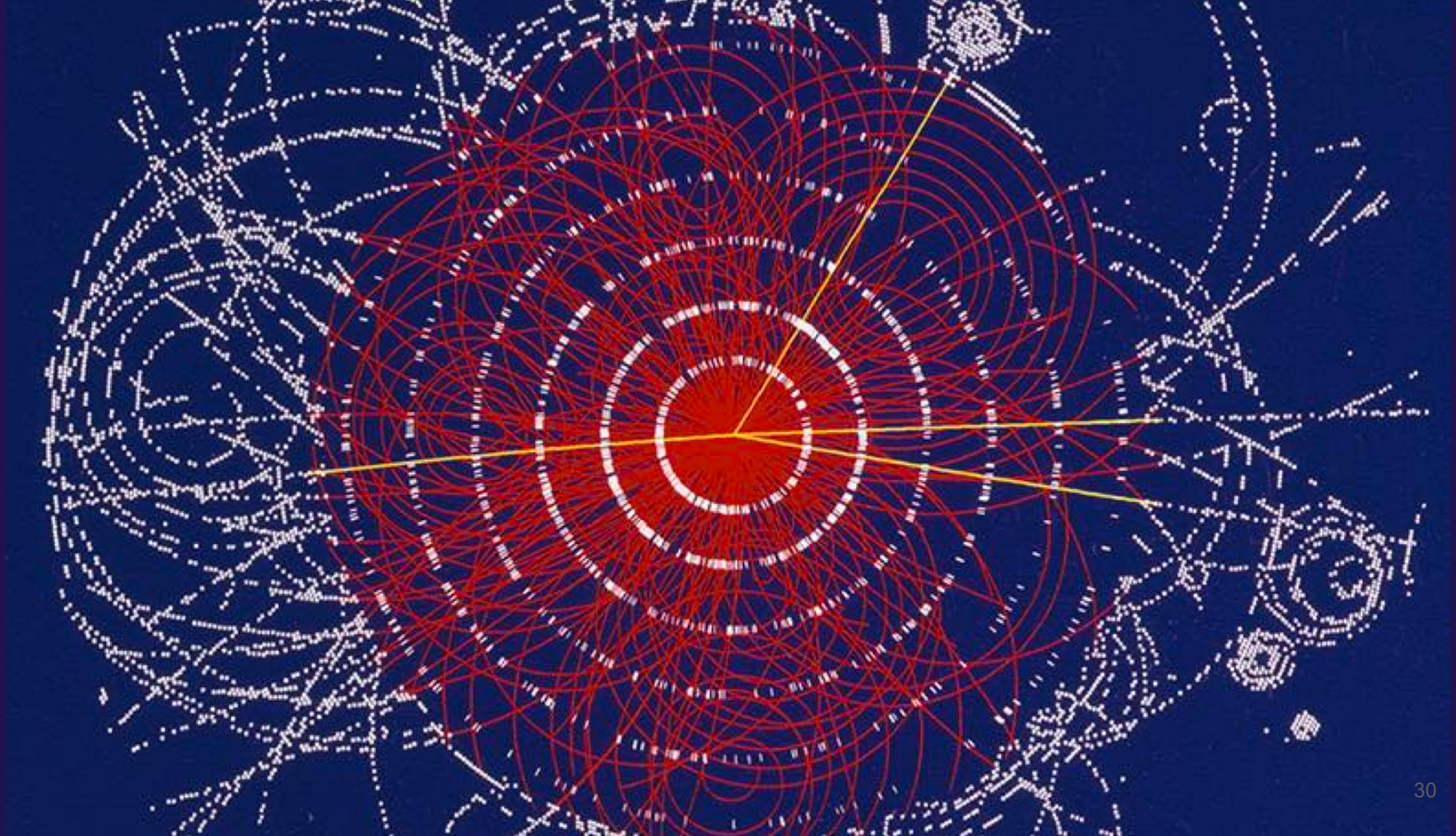
Other examples of high volume data?


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
Volume
SCALE OF DATA
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Human Brain Project

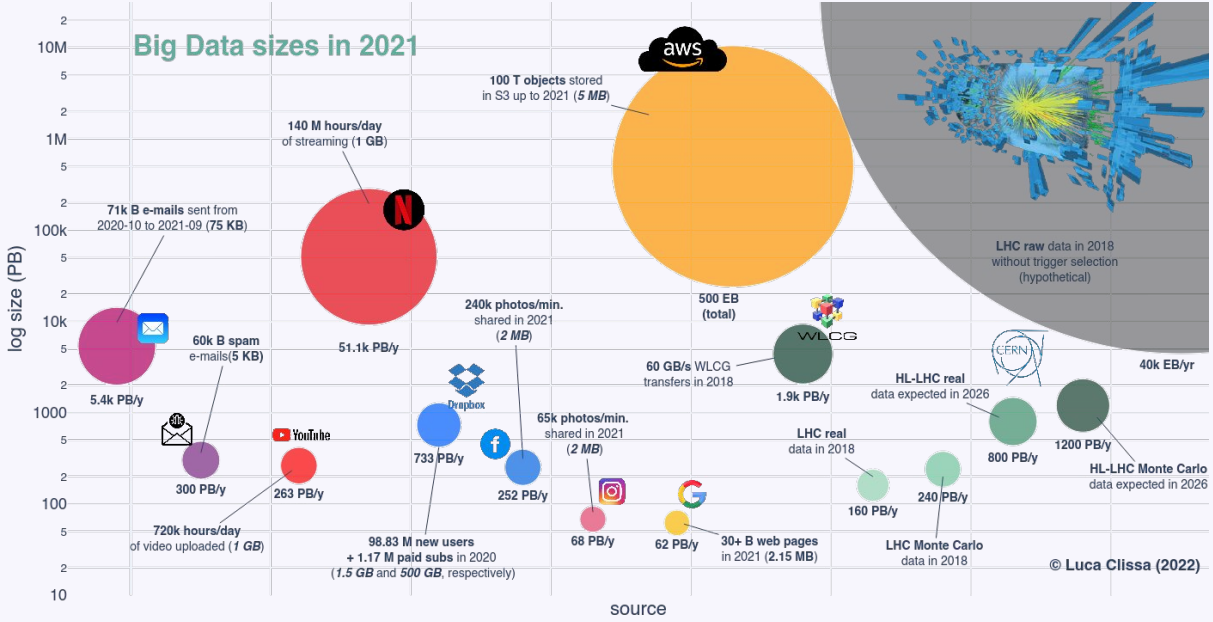
www.humanbrainproject.eu





High velocity data

[Pushing the limits of real-time ML: Nanosecond inference for Physics Discovery at the Large Hadron Collider]



Work presented at NeurIPS (Conference on Neural Information Processing Systems), the largest annual ML and AI research conference.

LHC computing grid

Model: grid computing

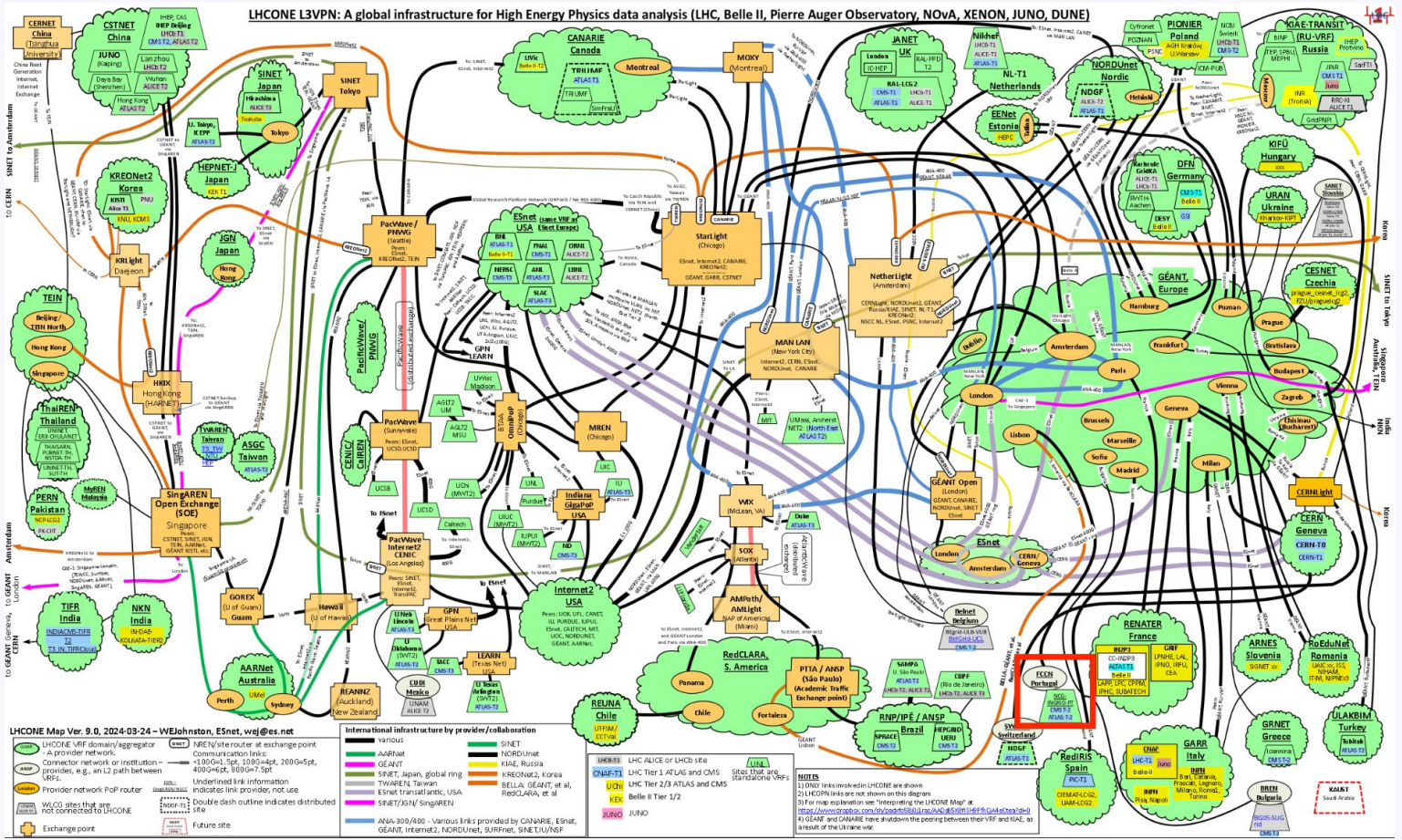
started in 2003

170 computing centers across

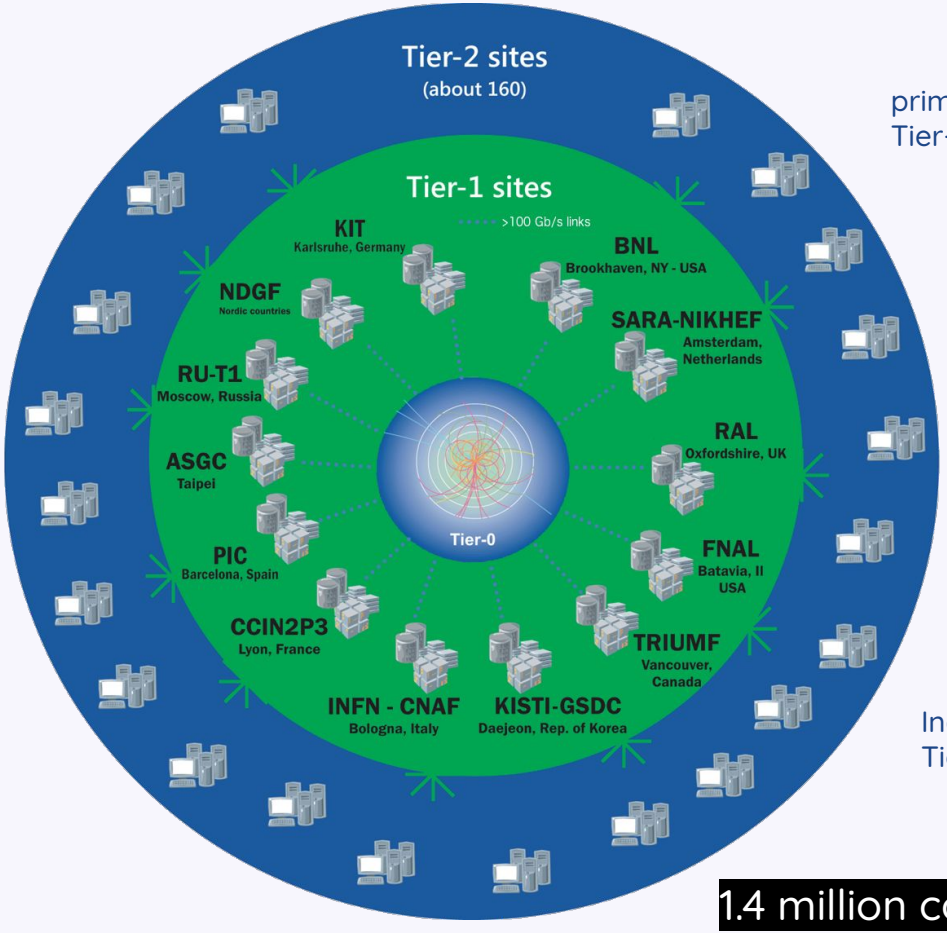
42 countries

> 10 000 physicists

LHC computing grid



LHC computing grid



primary backup will be recorded on tape at CERN, the Tier-0 centre of LCG

processed data will be distributed to a series of Tier-1 centres

- large computer centres with sufficient storage capacity and with round-the-clock support for the Grid

Tier-1 centres will make data available to Tier-2 centres

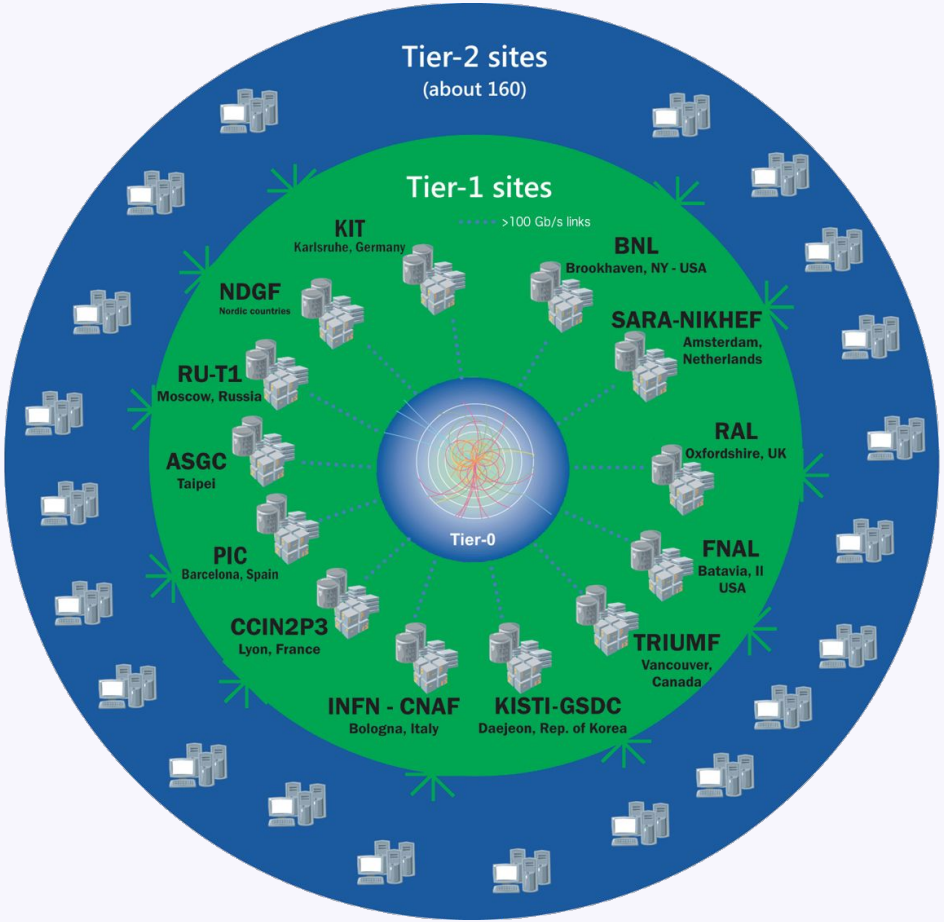
- each consists of one or several collaborating computing facilities

Individual scientists will access these facilities through Tier-3 computing resources

- can consist of local clusters in a University Department or even individual PCs

1.4 million computer cores, 1.5 exabytes of storage

LHC computing grid



What is a computing grid?

Distributed computing paradigm with:

- Shared heterogeneous computational resources across multiple (geographic disperse) administrative domains loosely coupled over network and controlled centrally (but not managed!)
- Distributed users of Virtual Organizations (VO) with a common access interface
- Goal: Collaboration on complex (compute) projects across different geographic and institutional boundaries to solve a common large-scale problem

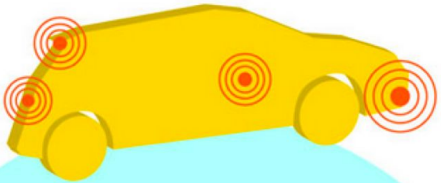
The naming "Grid" is inspired by the power grid

Other examples of high velocity data?

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session



Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

Velocity

ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

High variety data

Data centers on wheels

Computers that power self-driving cars could be a huge driver of global carbon emissions

Study shows that if autonomous vehicles are widely adopted, hardware efficiency will need to advance rapidly to keep computing-related emissions in check.

Adam Zewe | MIT News Office
January 13, 2023

1 billion autonomous vehicles, each driving for **one hour** per day with a computer consuming 840 watts, would consume enough energy to generate about the same amount of emissions as **data centers** currently do.

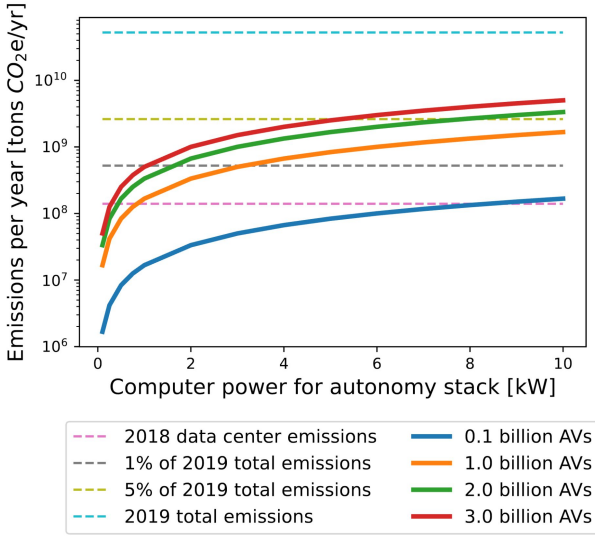


Fig. 1: Emissions from computing onboard AVs driving 1 hr/day. With one billion AVs, an avg. computer power of 0.84 kW yields emissions equal to emissions of all data centers.

From: *Data Centers on Wheels: Emissions from Computing Onboard Autonomous Vehicles*

Data centers on wheels

Computers that power self-driving cars could be a huge driver of global carbon emissions

Study shows that if autonomous vehicles are widely adopted, hardware efficiency will need to advance rapidly to keep computing-related emissions in check.

Adam Zewe | MIT News Office
January 13, 2023

For example, Facebook runs trillions of DNN inferences per day

an AV that drives for an hour per day computing 10 DNN inferences at 60 Hz on each of the inputs of 10 cameras would make 21.6 million inferences per day

one billion AVs would make 21.6 quadrillion inferences per day

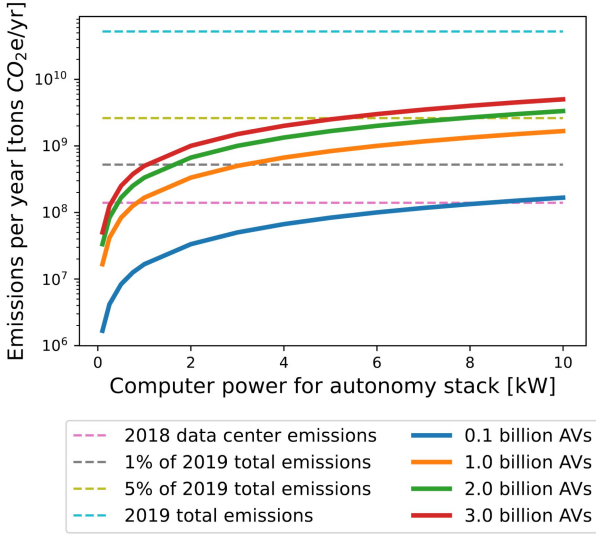
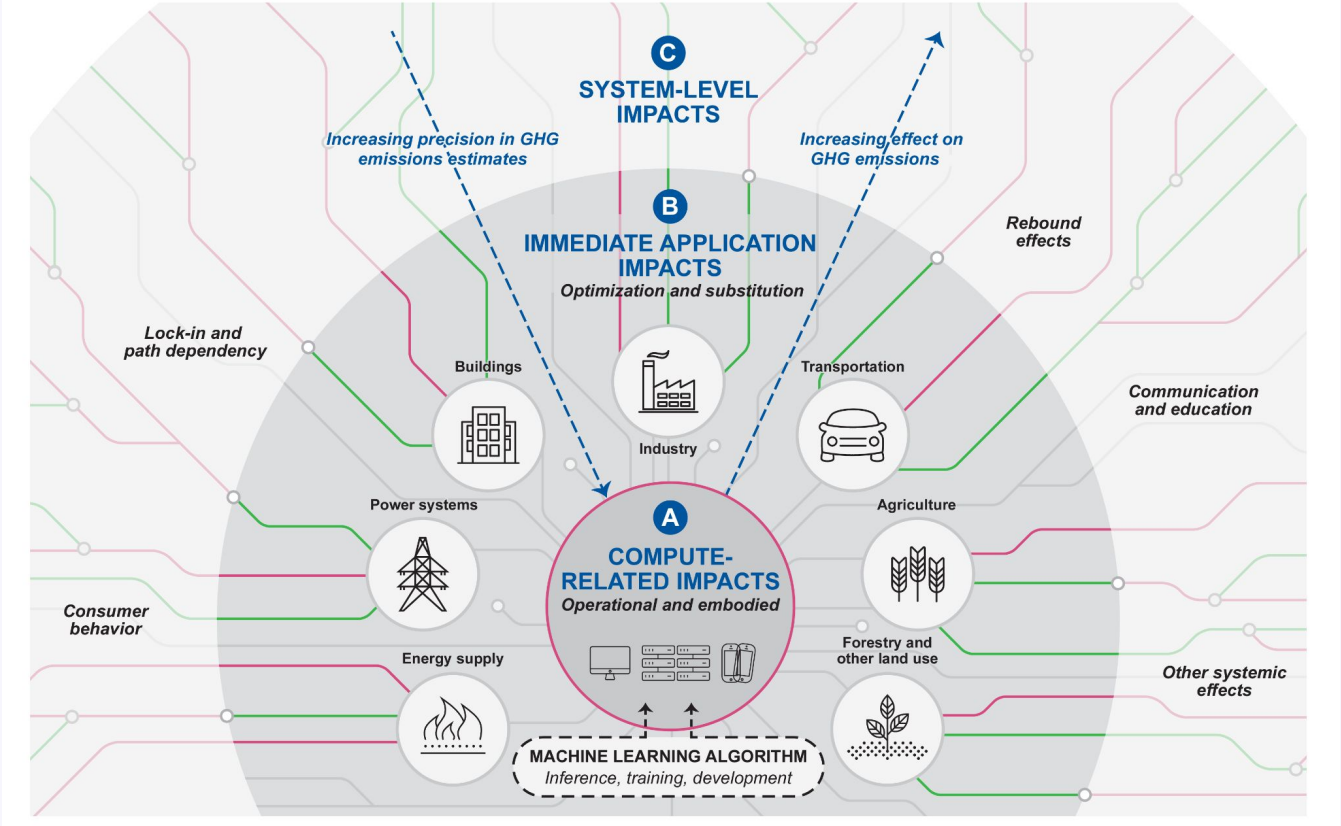


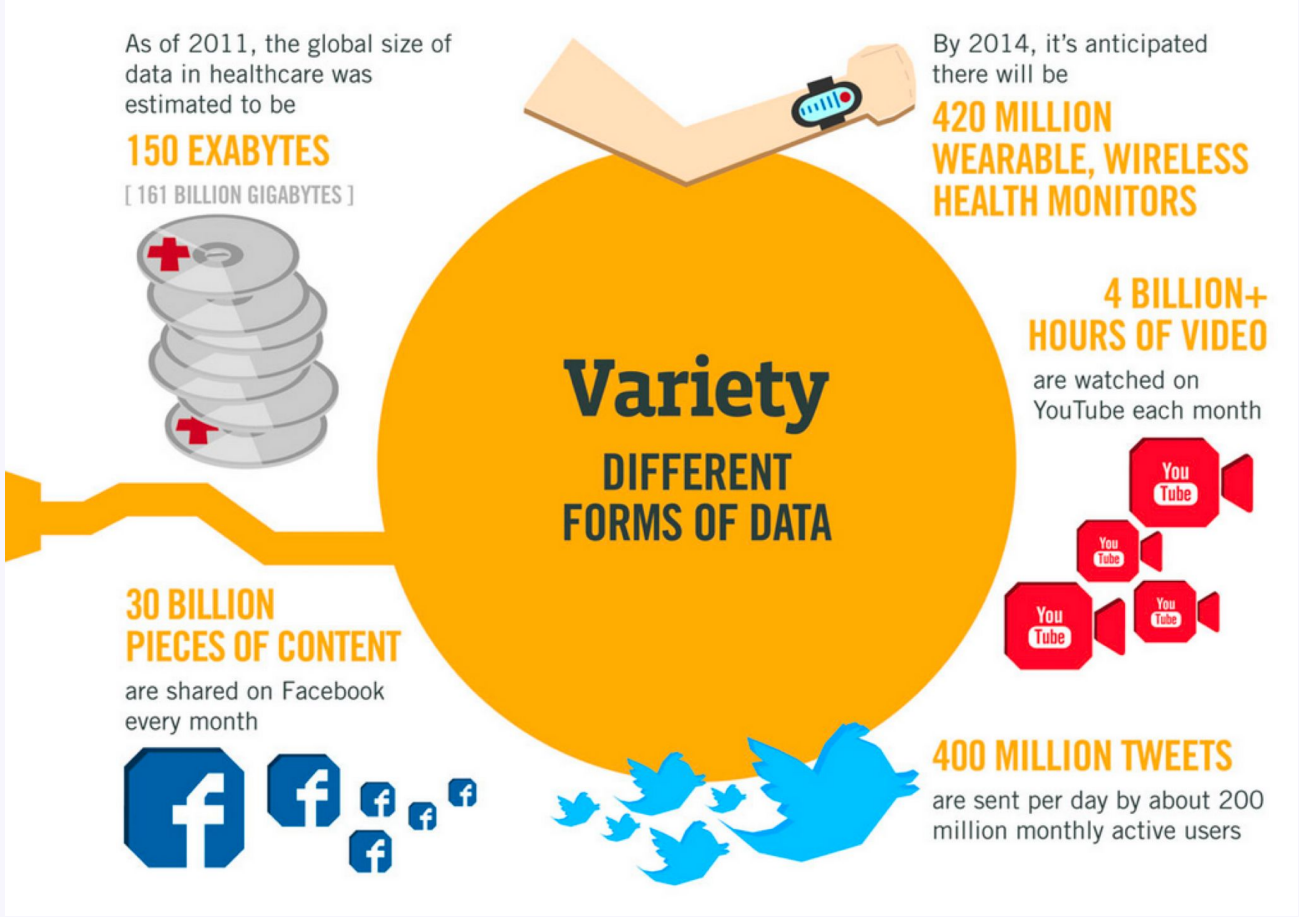
Fig. 1: Emissions from computing onboard AVs driving 1 hr/day. With one billion AVs, an avg. computer power of 0.84 kW yields emissions equal to emissions of all data centers.

From: *Data Centers on Wheels: Emissions from Computing Onboard Autonomous Vehicles*

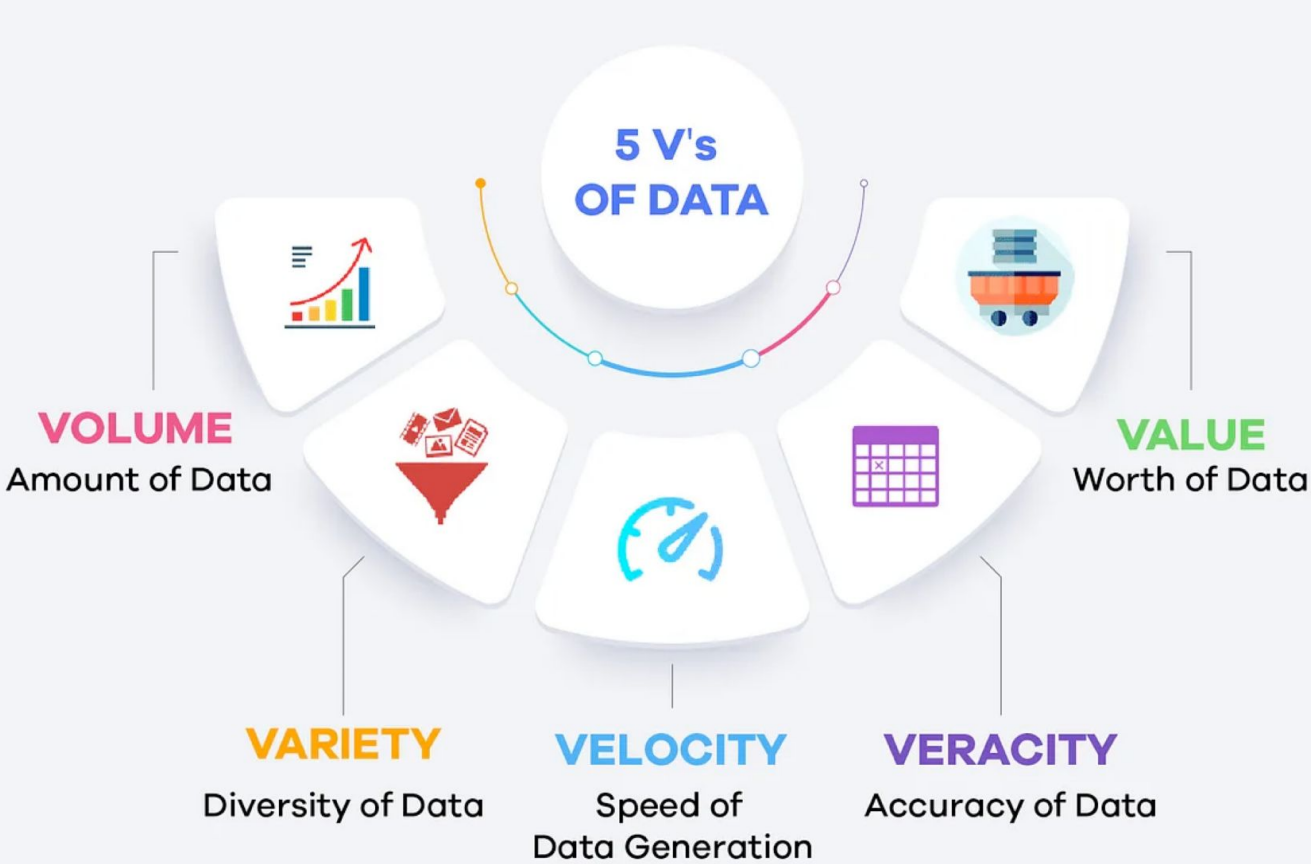
The impact of AI



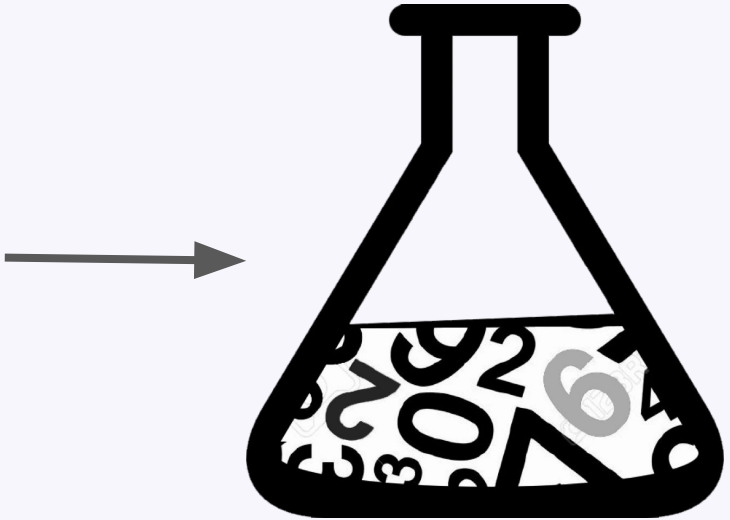
Other examples of high variety data?


As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
Variety
DIFFERENT FORMS OF DATA
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users

More V's of Big Data



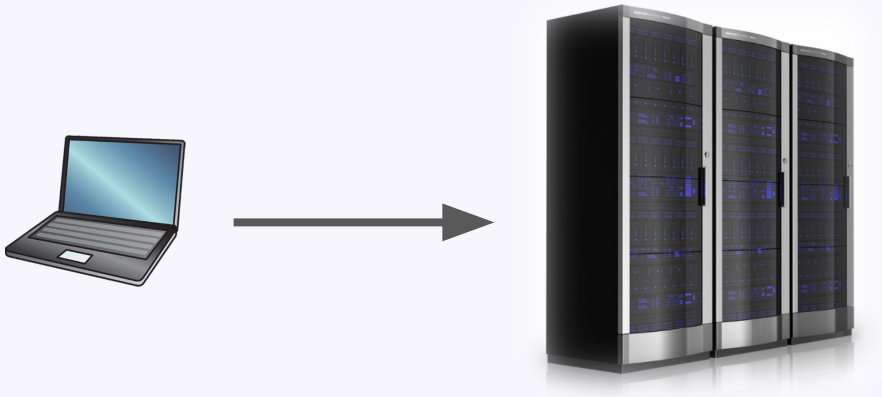
Big data requires data analytics/mining/science



Big data management - course objectives

Address the challenges that emerge during the collection, management, processing, and analytics of large-scale data

- Big data management tools/platforms
- Big data processing and analysis



Course contents

How do we process/manage big data?

How do we implement ML models for big data in practice?

Course contents

How do we process/manage big data?

How do we implement ML models for big data in practice?

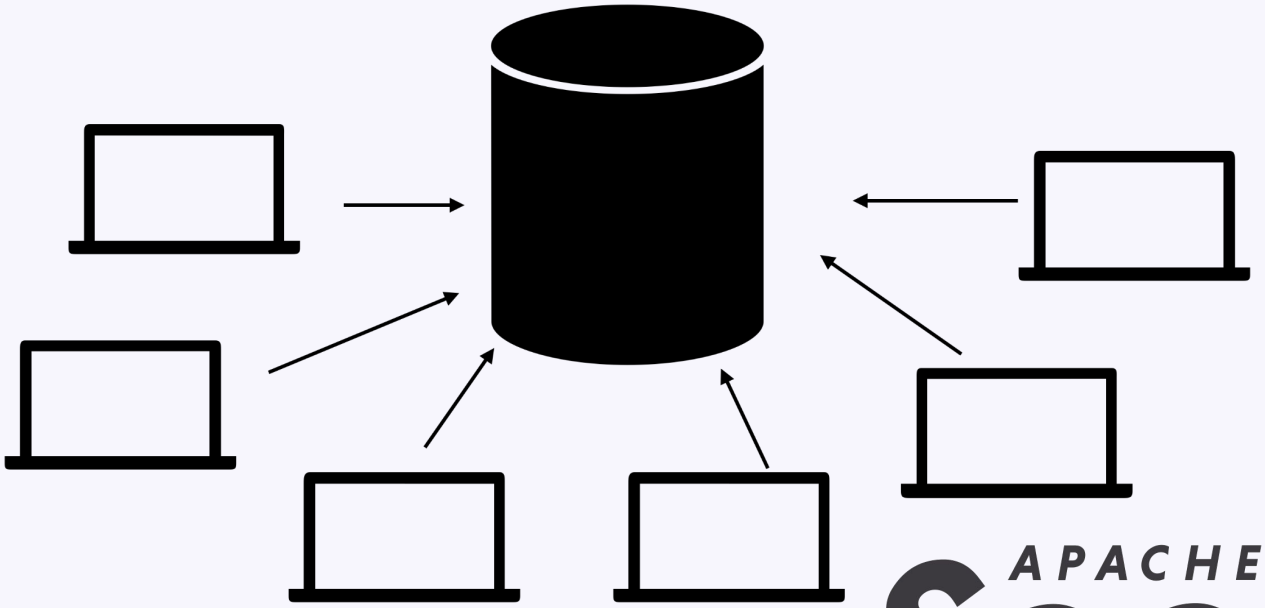
Storage
Querying

How do we process/manage big data?

How do you store data that does not fit in a single machine?

How do we process/manage big data?

How do you store datasets to speed up query answering?



Assignment: query and analyze a dataset of millions of reviews

yelp

The Dataset

6,990,280 reviews

150,346 businesses

200,100 pictures

11 metropolitan areas

APACHE Spark™

Course contents

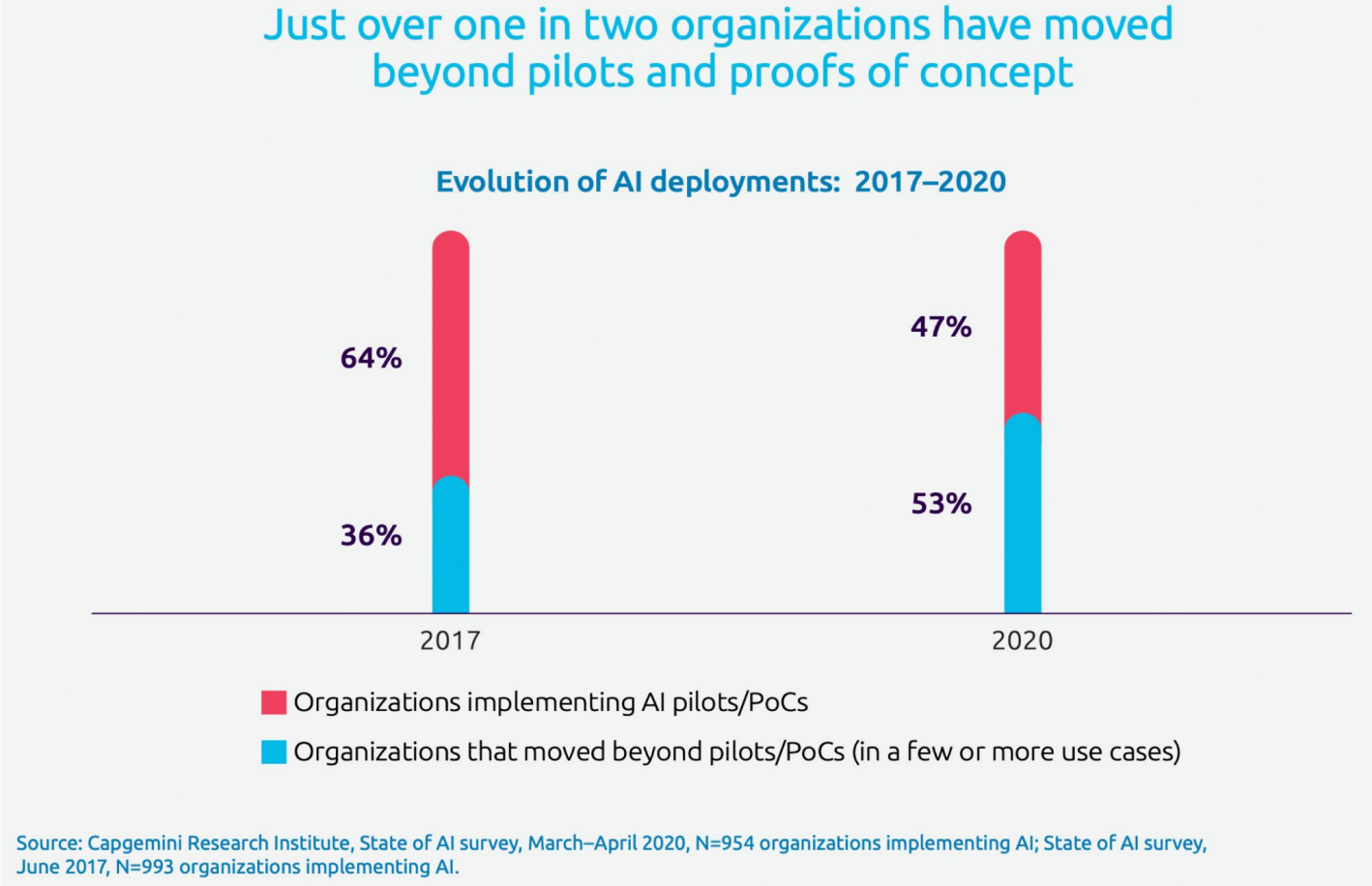
How do we process/manage big data?

How do we implement ML models for big data in practice?

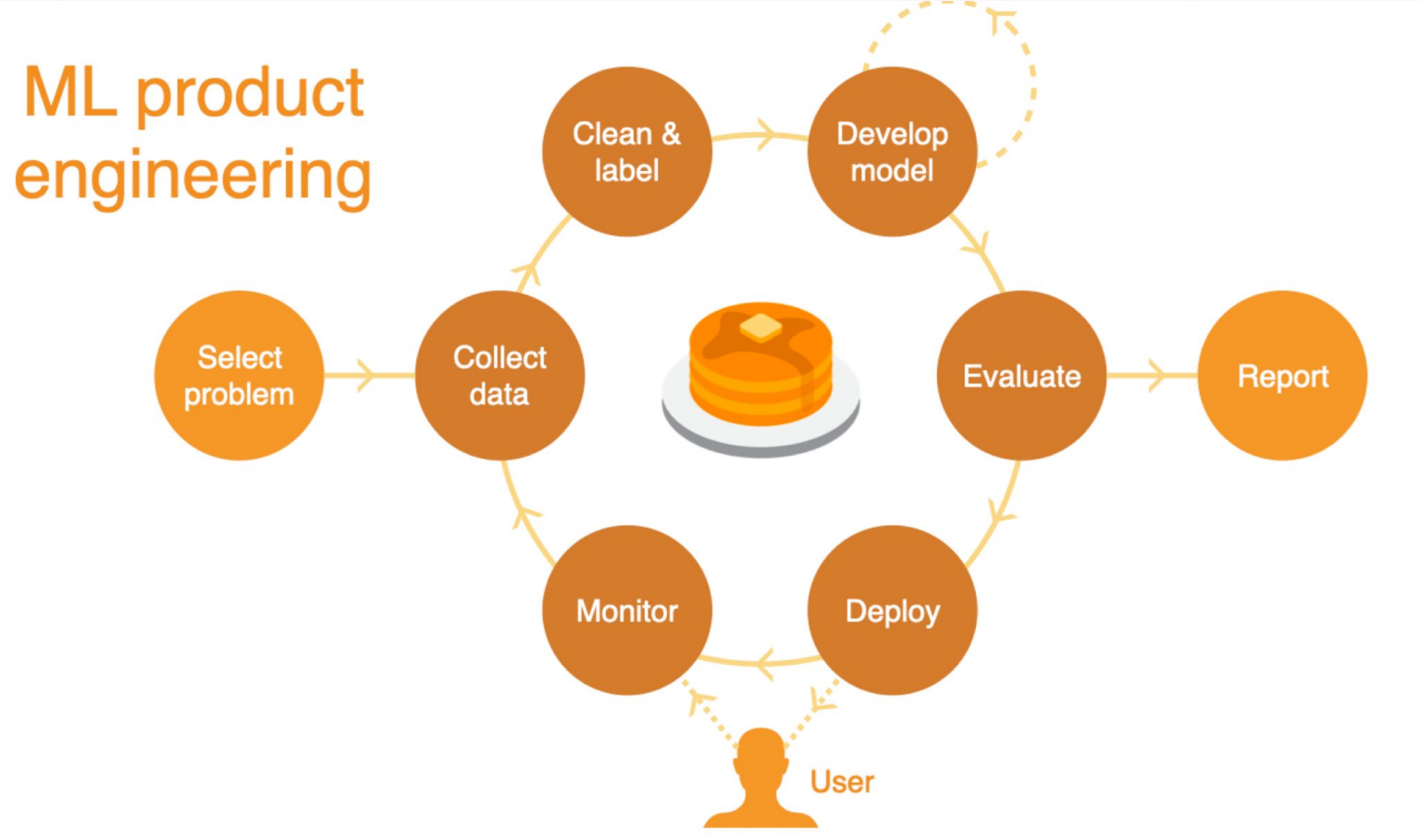
Full ML pipeline
Replication
Impact

How do we implement ML models for big data in practice?

How do we implement ML models for big data in practice?



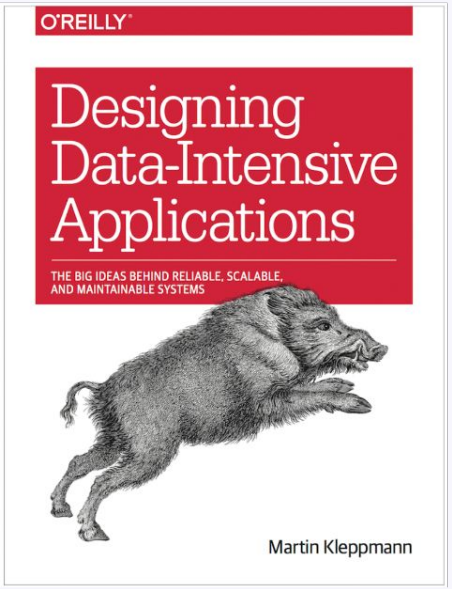
How do we implement ML models for big data in practice?

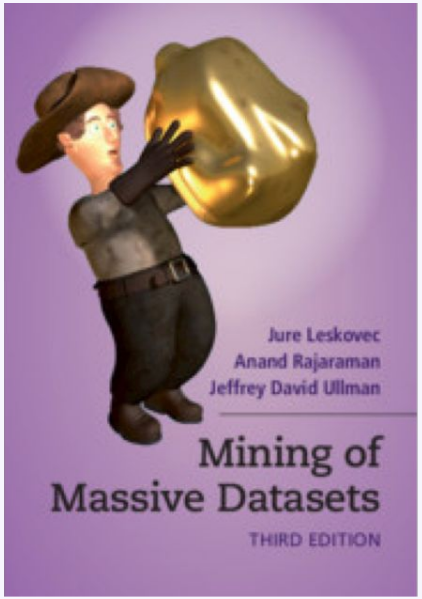


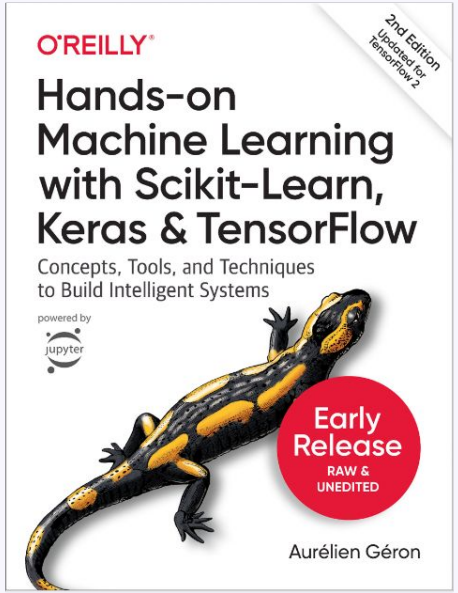
Assignment: implement and deploy a ML pipeline

Tools and technologies

Resources







Papers published in top venues in areas such as:
data management, machine learning, information retrieval

Course schedule

Week	Day (Friday 10.00-12.00)	Lecture	Title	Description	Teacher	Assignments
35	29.08.2025		NO CLASS			
36	05.09.2025	1	Intro+logistics	big data introduction, distributed systems	Maria	
37	12.09.2025	2	Big data storage	replication and partitioning	Zoi	
38	19.09.2025	3	Big data processing 1	HDFS and Mapreduce/hadoop	Zoi	
39	26.09.2025	4	Big data processing 2	Spark	Zoi	
40	03.10.2025	5	Cross-platform data processing	Wayang	Zoi	
41	10.10.2025	6	ML lifecycle I	preprocessing, pipelines	Maria	
42	BREAK					
43	24.10.2025	7	ML lifecycle II	model training	Maria	A1 deadline
44	31.10.2025	8	ML lifecycle III	experiment tracking and deployment	Maria	
45	07.11.2025	9	ITU guest lecture: Virtualization		TBD	
46	14.11.2025	10	ML impact	climate, privacy	Maria	
47	21.11.2025	11	Learning at the edge	Federated learning	Maria	
48	28.11.2025	12	Guest lecture: TBD		TBD	
49	05.12.2025	13	Conclusion/Guest lecture		TBD	A2 deadline
	BREAK					
2	06.01.2026		EXAM on premises			

Course organization

- 2h lectures in Aud 3

Lectures are typically recorded but not the best quality and not guaranteed

- 2h exercise sessions in 4A20-22, 54

Learn tools that will help you with the assignments

Get help setting up

Get to ask questions about the assignments and the exam

- For questions/clarifications: LearnIT forum, message/email, or ask for a meeting

Assignments and exam

2 *non-mandatory but necessary* assignments

- Spark project
- ML pipelines and experiments/deployment project

Non-mandatory: you don’t have to submit the assignments to be admitted in the exam.

BUT

Necessary: without solving the assignments you can’t pass the exam.

Assignments and exam

Exam:

- **75%** will be based on the assignments
- **25%** on material from the lectures

If you submit your assignment on time you get feedback that you can use for your exam later on

Discussion are allowed and encouraged:

- work independently before discussing
- acknowledge sources and collaborators
- submitted work must be your own! Do not plagiarize!

Usage of generative AI tools

1. **Learn, don't copy**: Use GenAI to aid your learning, but never copy-paste any GenAI outputs into your own assessed work. Doing so constitutes academic misconduct.

2. **Ask if uncertain**: Always consult us if you are unclear about the use of GenAI in your assessed work.

3. **Credit use of tools**: Before handing in your assessed work, make sure you acknowledge the use of GenAI, where used.

4. **Respect copyrights**: Never upload copyrighted materials to a GenAI platform without authorization from the copyright owner.

5. **Verify facts**: Always double check GenAI output for factual accuracy, including references and citations.

6. **Diversify sources**: Never rely solely on GenAI; it should supplement, but not replace, traditional sources.

Assignments and exam

Assignments are **not** mandatory

But, working on assignments in the allocated time allows you to:

- Get help from TAs
- Get help from peers
- Get feedback and a grade
- Practice submitting the exam
- Remove a heavy load during the exam period

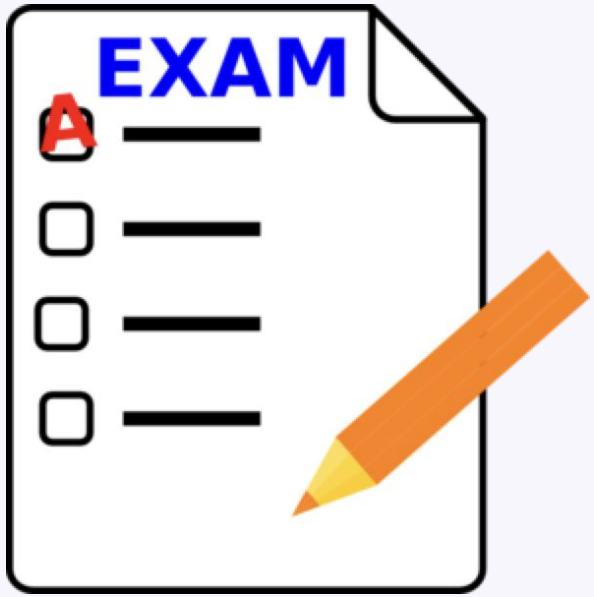
Exam

On-premises written exam: **6th January 9.00-13.00**

You will submit a report through LearnIT:

- 75% based on the assignments
- 25% based on lecture material

Example exam questions posted on LearntIT



Random fraud check

After the exam a random fraud check will take place.

The **day following the submission** we will hold an online meeting with 10% of students selected at random by SAP.

During the check we will ask a few questions about what is written in the report - you should be able to explain:

- your modelling choices
- important functions in your code
- concepts defined in your report and so on

We won't verify if the answers are correct, only if they match what is written in the report.

Expectations

Participation

- Be active, ask questions, share your knowledge

Preparation

- Complete reading/installation tasks
- Reading material is uploaded every week

Responsibility

- Read all instructions
- Manage your time