Yogesh Simmhan
simmhan@iisc.ac.in

Department of Computational and Data Sciences
Indian Institute of Science, Bangalore

DS256 (3:1)

Scalable Systems for Data Science

Scalable Systems for Data Science

- Instructor: **Yogesh Simmhan** (email) (www)
- TAs: Pranjal, Haseeb, Mayank
- Course number: DS256
- Credits: 3:1
- Semester: Jan 2026
- Lecture: **Tue/Thu 330-5pm**
- Room: CDS 202
- Teams: Teams Link (Join using Teams Code **v0xq5zv**)
- Pre-requisites: *Data Structures, Programming and Algorithm concepts. Programming experience required. Basic knowledge of Machine Learning and Deep Learning.*

About the course

- Fundamental “systems” aspects of designing and using scalable data science platforms.
- Store, manage, pre-process and train ML models over datasets that are large, fast and linked
- Data engineering pipelines required to prepare data before DNN and LLM training
- Scalable machine learning methods using distributed and federated approaches
- Big Data platforms used to scale enterprise data.

About the Course

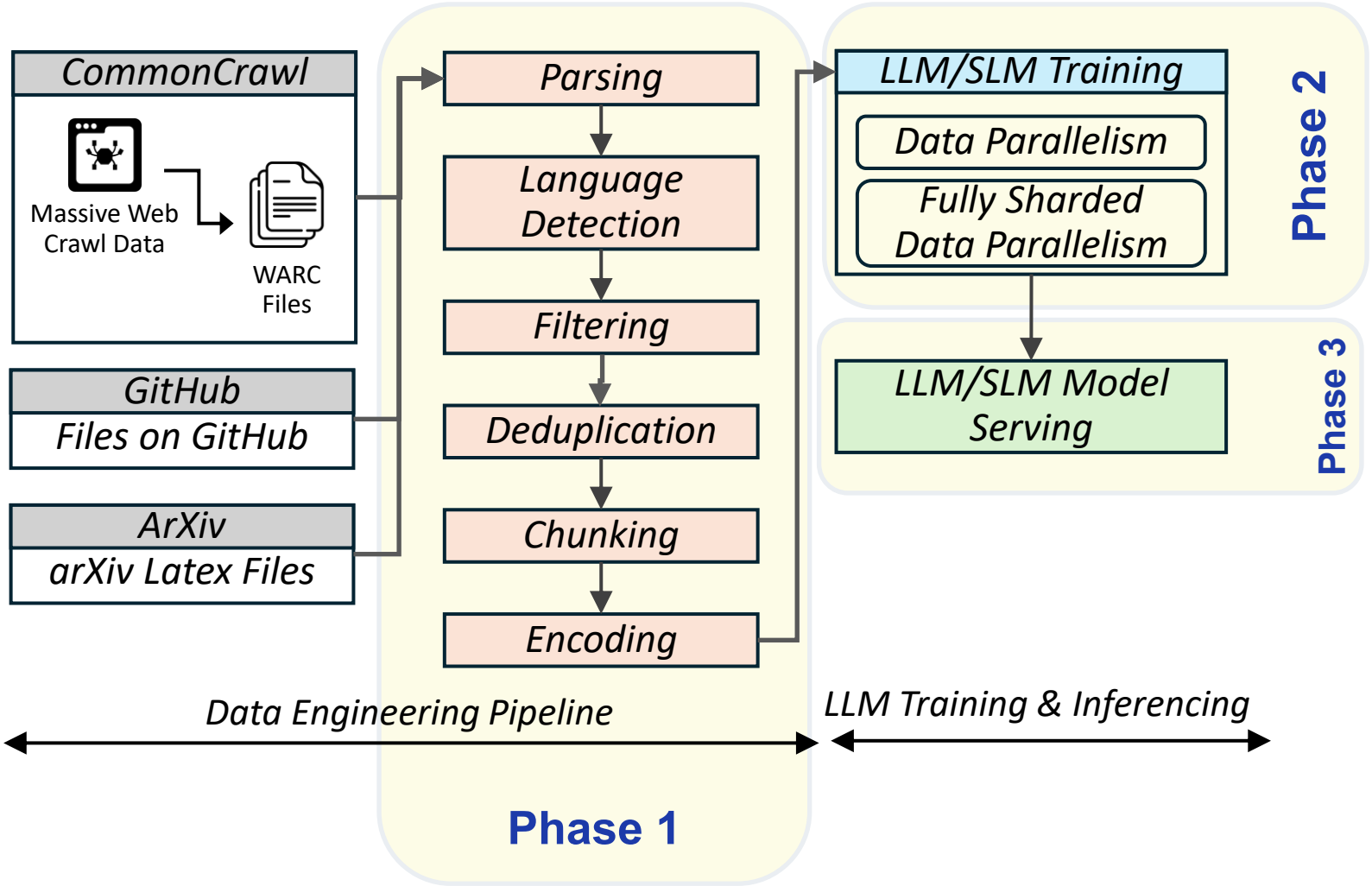
- How do you store and query data at scale, using **distributed file systems** such as *GFS/HDFS and Ceph* and using **cloud/NoSQL databases** such as *HBase and Dynamo*?
- How do you pre-process data at large volumes in preparation for machine learning using **distributed processing systems** on the cloud, such as *Apache Spark*?
- How do you perform scalable training for both classic and deep learning using distributed training patterns and platforms such as parameter server, model/pipeline parallelism, federated learning, *SparkML*, *Pytorch Distributed and DistDGL*? How do we serve LLM and GNN model inferencing at scale on distributed systems?
- How do you **process fast and linked data** for applications such as Internet of Things (IoT) and fintech using platforms such as Kafka, Spark Streaming and Giraph?

New this Semester

- **Architectural design** of data platforms: *Google's GFS/ Apache HDFS, Apache Spark (RDD/DF/ML/Streaming), PyTorch Distributed, Amazon DynamoDB/Apache Cassandra, Apache Kafka, Google's Pregel/Apache Giraph…*
- Examine **Design Pattern for Distributed Systems** used by these to achieve scalability, throughput, reliability…
- Entire class will do a semester-long project on designing a **pre-processing pipeline for training an LLM (SLM)** using *Apache Spark*…
- **Distributed model training** using *Pytorch Distributed* and **serving** the model.

If you like distributed data and computing systems, or are interested in scaling ML Workloads, this course if for you!

Example LLM Pipeline for Course Project



About the Course

- **Module 1 (~4 lectures):** Introduction to Distributed Systems & Big Data Storage
  - Introduction to Big Data. Motivation for scalable data systems. Overview of distributed systems, Cloud computing, strong and weak scaling. Architecture of Google File System/HDFS, Design Patterns of HDFS.
- **Module 2 (~5 lectures/Phase 1):** Processing Large Volumes of Big Data
  - Introduction to Big Data processing systems. Architecture and internals of Apache Spark, DF and SQL. Programming using Spark DF. Design Patterns of Spark.
- **Module 5 (~5 lectures/Phase 2, 3):** Machine Learning at Scale
  - Spark ML, Pytorch Distributed, Data, Model and Pipeline parallelism. Parameter server. Federated Learning. Scalable GNN Training. Training & serving LLMs at scale. Design Patterns of ML training & inferencing.
- **Module 3 (4 lectures):** NoSQL Databases
  - Intro to relational, NoSQL databases. ACID, BASE and CAP Theorem. Architecture of Dynamo/Cassandra, BigTable/Hbase, GraphDB, VectorDB. Overview of ETL/Data lakes.
- **Module 4 (4 lectures):** Fast Data & Linked Data Processing
  - Introduction to streaming and linked data processing. Architecture and programming of distributed streaming systems like Kafka, Spark Streaming. Distributed graph processing systems like Pregel/Giraph.

Pre-requisites

- Familiarity with:
  - Computer systems, operating systems
  - Data structures, algorithms and good programming skills (preferably in Python)
  - Basics of Machine Learning, Deep Learning

Course Material

- Lecture slides
- Paper reading
  - 3-4 per module
  - Must read before class
  - Lectures give overview of papers and supplementary material
- Select textbook chapters
  - **Learning Spark**, 1st and 2nd Editions
  - **Patterns of Distributed Systems**, Umesh Joshi, Martin Fowler, 2023

- *All of these will be part of quizes and final exam*

Evaluation

- **Midterm Quizzes (2 quizzes x 15 points = 30 points)**
  - Quizzes will cover all modules. MCQ and short answer format. In-class, using MS Forms/Moodle.
  - Individual effort. No collaboration/ChatGPT. Honor system.
- **Project Assignments (20+10+5=35 points)**
  - One hands-on programming course project on pre-processing pipeline, training and inferencing for LLMs
  - Team effort (2-4 per team, *TBD*)
- **Paper Reading & Presentation (10 points)**
  - Choose a paper from reading list that will be provided
  - Read in-depth and present in class for 30mins. Peer view and feedback.
  - Individual effort. No collaboration. Honor system.
- **Final exam (25 points)**
  - Individual effort. No collaboration. Honor system. Short and long form answers.

Logistics

- **3:1** Course
- 3 hours of lectures per week, Mon/Wed 330-5pm
  - ~22 instructor lectures, excluding IISc holidays (special class)
  - ~2-3 guest lectures from Industry
- Tutorial ~once a week, 1.5hrs typically Friday, time *TBD* thru poll
- Microsoft Teams for lectures, tutorials, slides
  - Important announcements on Teams Chat, Q&A Chat
- Quizzes likely on Moodle/MS Forms

schedule.xlsx

DS256 Tentative Schedule (Jan 2026 Term)

Module 1: Introduction to Distributed Systems & Big Data Storage (4)			
1	Tue, 13 Jan, 2026		Intro to big data. Contrast Big Data systems. Scalability, weak and strong scaling.
2	Thu, 15 Jan, 2026		Distributed File Systems/HDFS/GFS
T1	Fri, 16 Jan, 2026		Tutorial: Turing setup
3	Tue, 20 Jan, 2026		Distributed File Systems/HDFS/GFS
4	Thu, 22 Jan, 2026		Design Patterns of HDFS
T2	Fri, 23 Jan, 2026		Tutorial: HDFS
Module 2: Processing Large Volumes of Big Data (5)			
5	Tue, 27 Jan, 2026		Big Data Processing with MapReduce and Apache Spark
6	Thu, 29 Jan, 2026		Spark Basics, RDD, transformations, action, Shuffle
T3	Fri, 30 Jan, 2026		***Assignment # 1 Posted***
			Tutorial: Spark Data Frames
7	Tue, 3 Feb, 2026		Spark internals & Spark tuning
8	Thu, 5 Feb, 2026		Spark DataFrames, Spark SQL and Catalyst Optimizer
T4	Fri, 6 Feb, 2026		Tutorial: Introduction to Assignment
9	Tue, 10 Feb, 2026		Design Patterns of Spark
	Thu, 12 Feb, 2026	*ACM Annu*	**Midterm Exam #1**
T5	Fri, 13 Feb, 2026		Tutorial: Introduction to Assignment #1
Module 3: Machine Learning at Scale (5)			
10	Tue, 17 Feb, 2026		ML over Big Data, Spark ML for ML pipelines. Data, Model and Pipeline parallelism. Parameter server.
	Thu, 19 Feb, 2026	*Ugadi Holiday*	
T6	Fri, 20 Feb, 2026		Midterm Exam #1 Review
11	Tue, 24 Feb, 2026		Training and Serving LLMs at scale
12	Thu, 26 Feb, 2026		PyTorch Distributed
T7	Fri, 27 Feb, 2026		***Assignment #1 Due***
			Tutorial: Pytorch Distributed
13	Tue, 3 Mar, 2026		Federated Learning platforms
14	Thu, 5 Mar, 2026		Scalable GNN training
T8	Fri, 6 Mar, 2026		***Assignment #2 Posted***
			Tutorial: Introduction to Assignment #2

			Module 4: NoSQL Databases (4)
15	Tue, 10 Mar, 2026		Consistency models and CAP theorem/BASE
16	Thu, 12 Mar, 2026		Amazon Dynamo/Cassandra distributed key-value store
T9	Fri, 13 Mar, 2026		Tutorial: LLM Model Service, e.g., vLLM?
17	Tue, 17 Mar, 2026		Overview of HBase/Big Table, Graph Databases, Vector Databases. Data Warehousing, Data Lakes, ETL, Cloud NoSQL
18	Thu, 19 Mar, 2026		Design Patterns of HBase, Dynamo
T10	Fri, 20 Mar, 2026		***Assignment #2 Due***
	Tue, 24 Mar, 2026	*SEACEN T*	**Midterm Exam #2**
			Module 5: Processing Fast Data & Linked Data (4)
19	Thu, 26 Mar, 2026		
T11	Fri, 27 Mar, 2026		***Assignment #3 Posted***
			Tutorial: Introduction to Assignment #3
	Tue, 31 Mar, 2026	*Mahavir Jayanti Holiday*	
20	Thu, 2 Apr, 2026		
	Fri, 3 Apr, 2026	*Good Friday Holiday*	
21	Tue, 7 Apr, 2026		
22	Thu, 9 Apr, 2026		
T12	Fri, 10 Apr, 2026		***Assignment #3 Due***
			Module 6: Guest Lectures (2)
23	Tue, 14 Apr, 2026		M365 Lecture
24	Thu, 16 Apr, 2026		NPCI/IBM Lecture
25	Fri, 17 Apr, 2026		**Paper Presentations**
	21-30/Apr		**Final Exams**

In-class/Off-class Participation

- Expect the lectures to be interactive

- Use Teams for asynchronous Q&A
  - Encourage active collaboration. Anyone can respond to questions (unless it's a direct solution to an assignment)
  - TA/Yogesh will answer or confirm answer

IISc POLICY FOR ACADEMIC INTEGRITY

- Acknowledge and cite use of others' material
- Acknowledges all contributors to a piece of work
- All work submitted is his or her own in a course
- Produce academic work without the aid of impermissible materials or collaboration.
- Obtains all results by ethical means and reports them accurately
- Neither facilitates dishonesty by others nor obstructs their academic progress.
- Use of Generative AI in completing any of the assessments is not permitted.
- Violation will result in **penalties**.
- **Do NOT actively collaborate on quizzes, assignments, final exam, etc.**

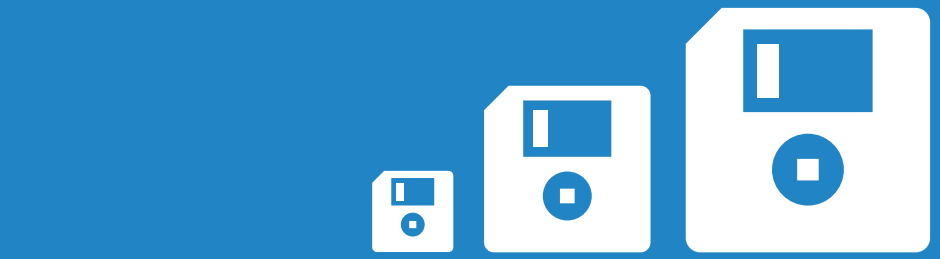
https://www.iisc.ac.in/about/student-corner/academic-integrity/

Module 1

Introduction to Big Data & Distributed Storage

Mandatory Reading before Next Class

1. ***Scalable problems and memory-bounded speedup***, *Sun and Ni, JPDC, 1993*
2. ***The Google File System***, *Sanjay Ghemawat Howard Gobioff Shun-Tak Leung, ACM SOSP, 2003*



Big Data Concepts

What, Where, Why?

What is Big Data?

Image credits: http://www.seekbig.in/1128-tnpsc-economics-questions/

The term ***is*** fuzzy … *Handle with care!*

Wordle of "Thought Leaders'" definition of Big Data, © Jennifer Dutcher, 2014
https://datascience.berkeley.edu/what-is-big-data/

Data Generation View

"Big data refers to the approach to data of 'collect now, sort out later'…The low cost of storage and better methods of analysis mean that you generally don't need to have a specific purpose for the data in mind before you collect it."

Rohan Deuskar, CEO and Co-Founder, Stylitics

Wordle of "Thought Leaders'" definition of Big Data, © Jennifer Dutcher, 2014
https://datascience.berkeley.edu/what-is-big-data/

Image Credits: https://community.uservoice.com/wp-content/uploads/benefits-of-effective-questions-800x448-300x168.jpg

Data Systems View

"Big data is when your business wants to use data to solve a problem, answer a question, produce a product, etc., but the standard, simple methods break down on the size of the data set, causing time, effort, creativity, and money to be spent crafting a solution to the problem that leverages the data without simply sampling or tossing out records."

John Foreman, Chief Data Scientist, MailChimp

Wordle of "Thought Leaders'" definition of Big Data, © Jennifer Dutcher, 2014
https://datascience.berkeley.edu/what-is-big-data/

Image Credits: http://heae.tk/question/

Data Analysis View

"While the use of the term is quite nebulous … I've understood "big data" to be about analysis for data that's really messy or where you don't know the right questions or queries to make — analysis that can help you find patterns, anomalies, or new structures amidst otherwise chaotic or complex data points."

Philip Ashlock, Chief Architect, Data.gov

Wordle of "Thought Leaders'" definition of Big Data, © Jennifer Dutcher, 2014
https://datascience.berkeley.edu/what-is-big-data/

Image Credits:http://www.clipartpanda.com/categories/question-and-answer-clipart

So…What is Big Data?

Data whose characteristics exceeds the capabilities of conventional *algorithms*, ***systems*** *and techniques* to derive useful value.

https://www.oreilly.com/ideas/what-is-big-data

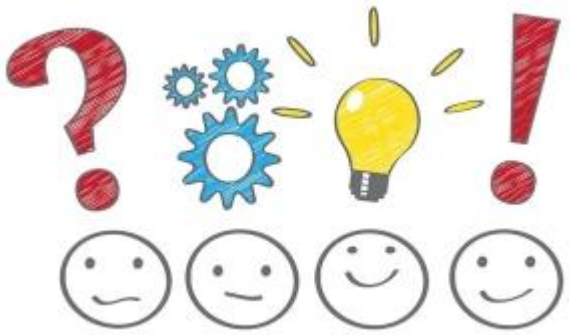
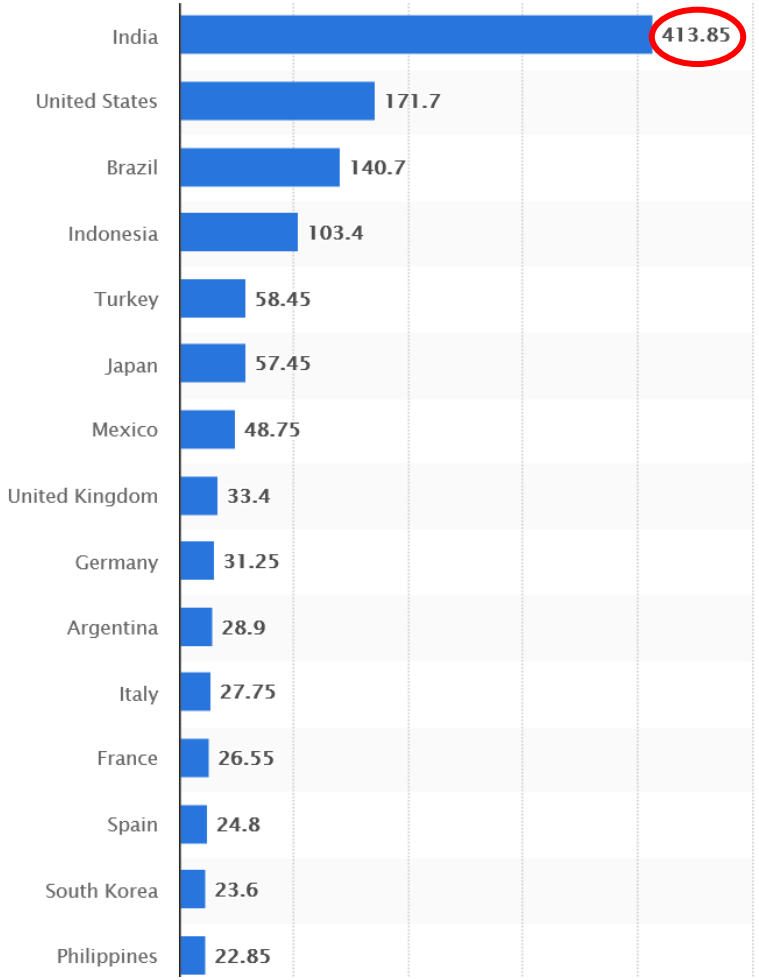
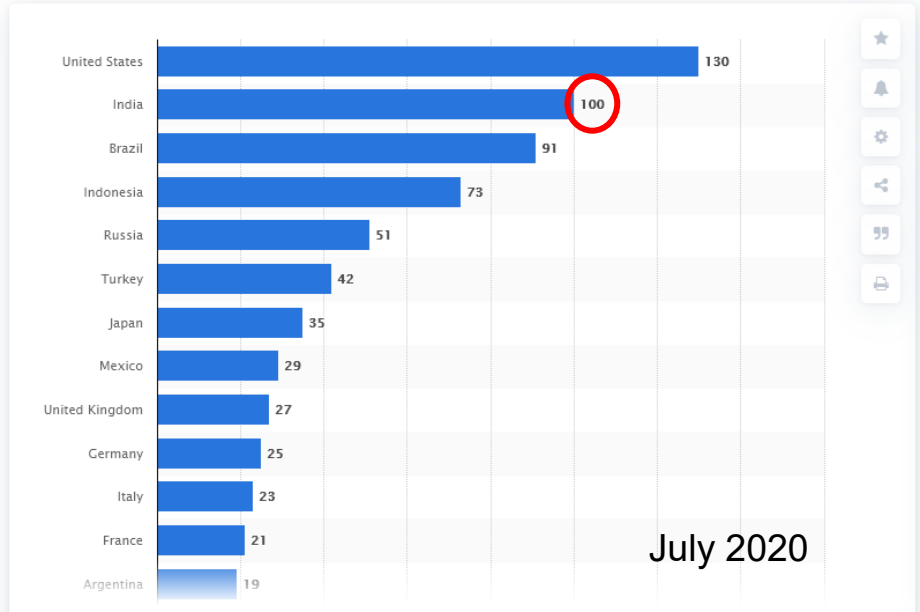


Image Credits: https://community.uservoice.com/wp-content/uploads/benefits-of-effective-questions-800x448-300x168.jpg

So, where does Big Data come from?

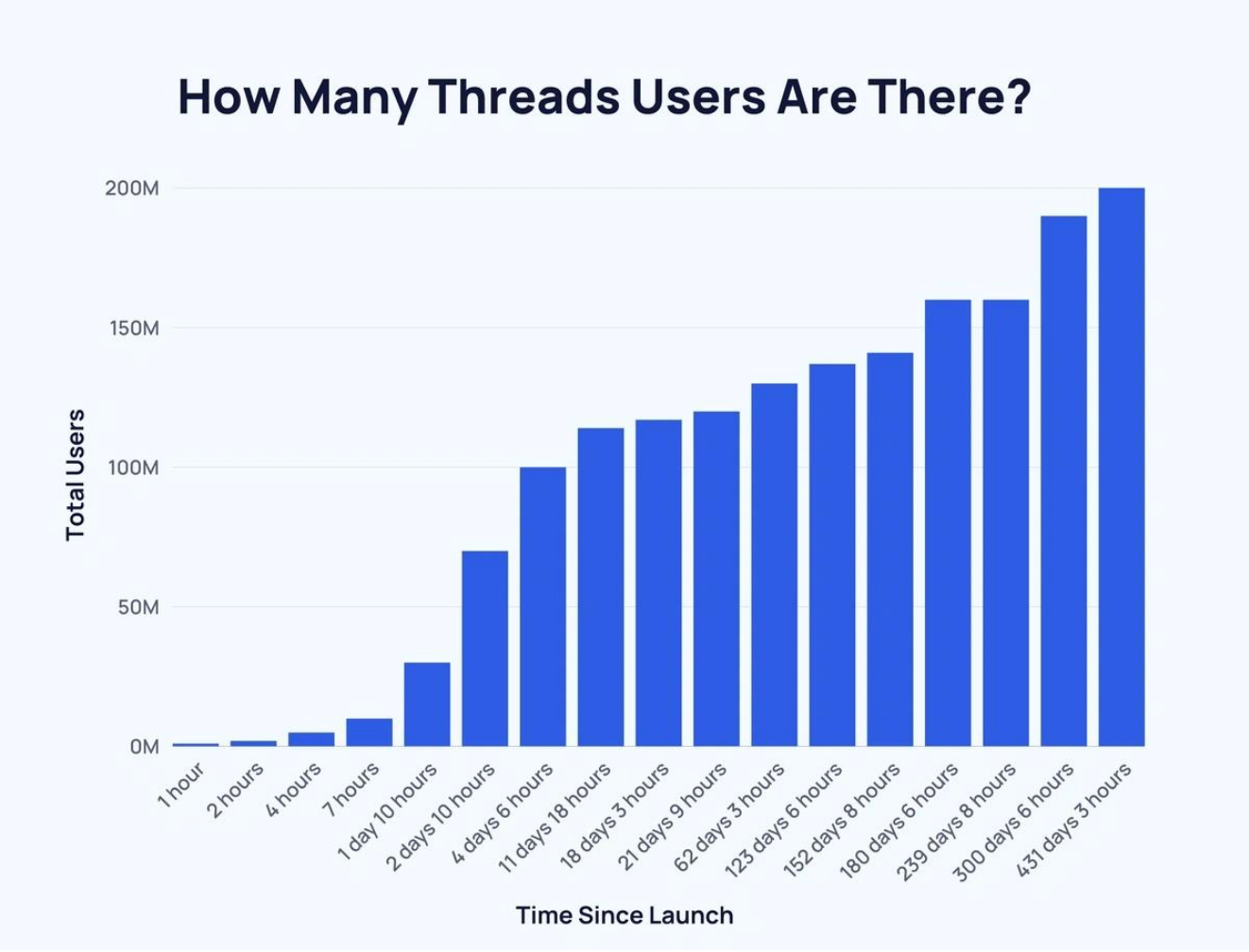
Instagram Users across the World (Feb 2025)





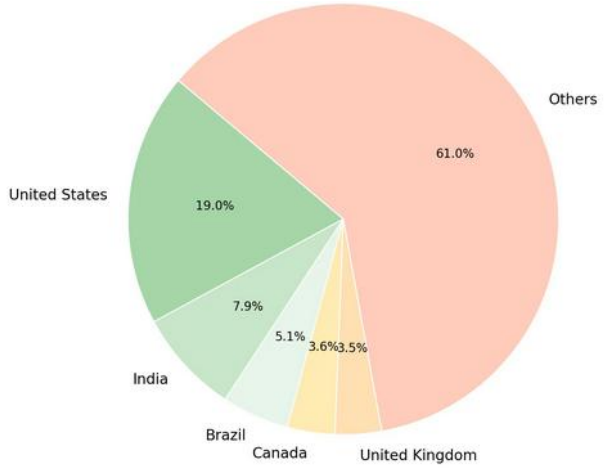
https://www.statista.com/statistics/578364/countries-with-most-instagram-users/

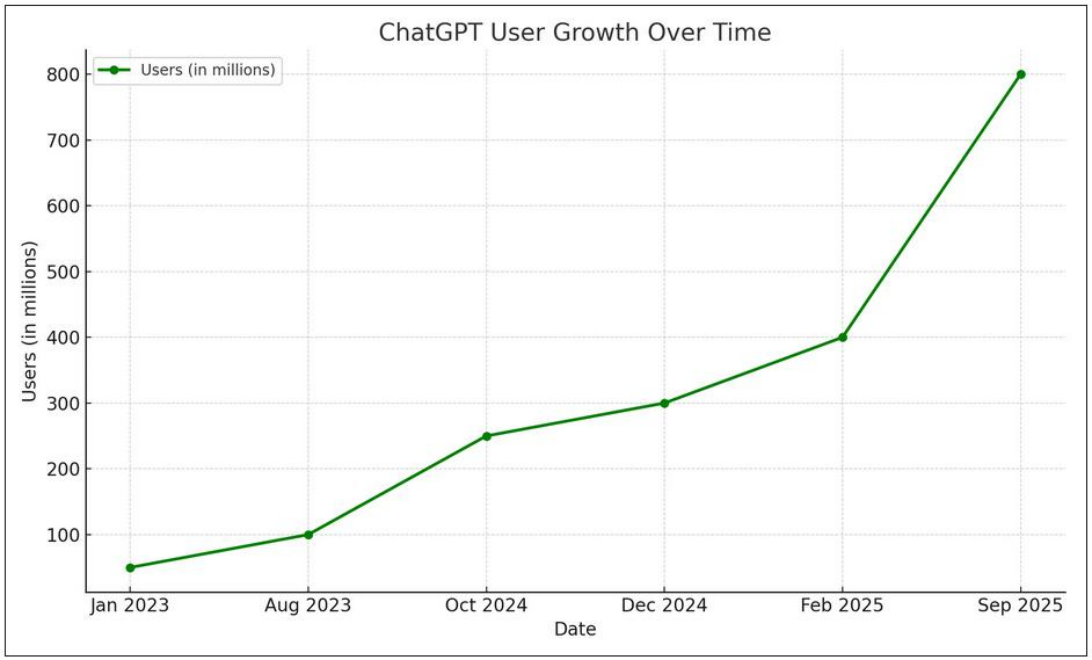
Rapid Ramp-up of Threads sign-ups worldwide since Jul 5, 2023 launch

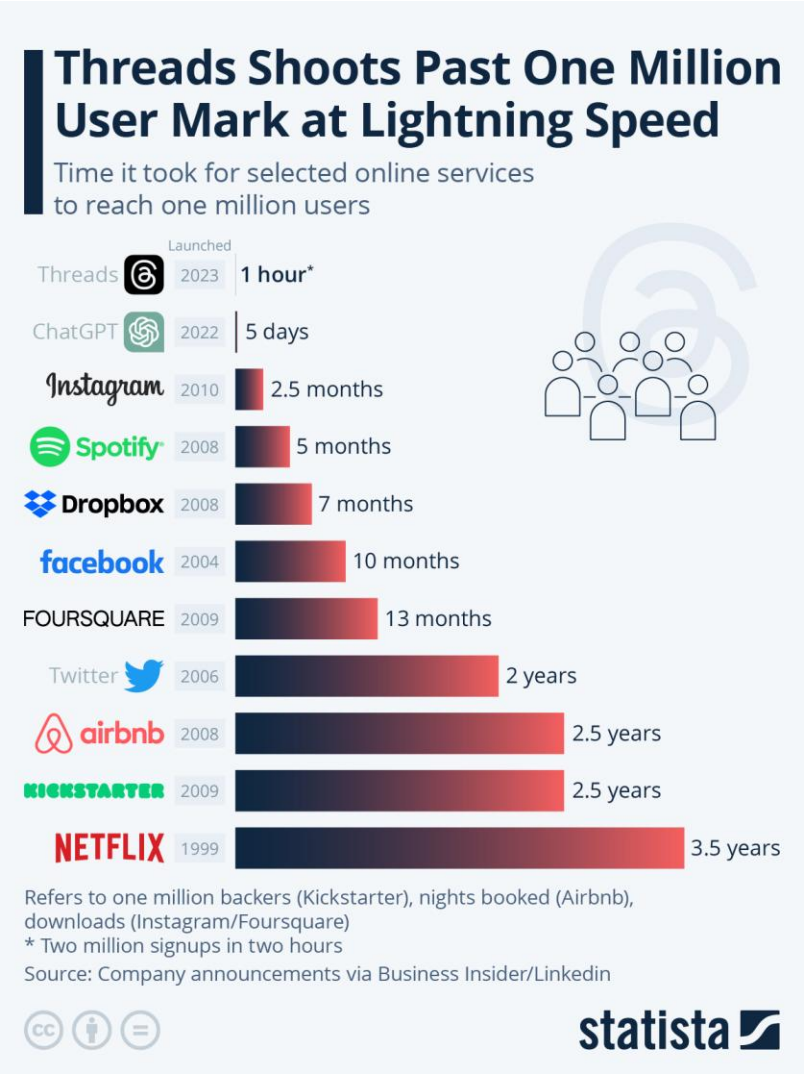


https://explodingtopics.com/blog/threads-users

OpenAI/ChatGPT Users

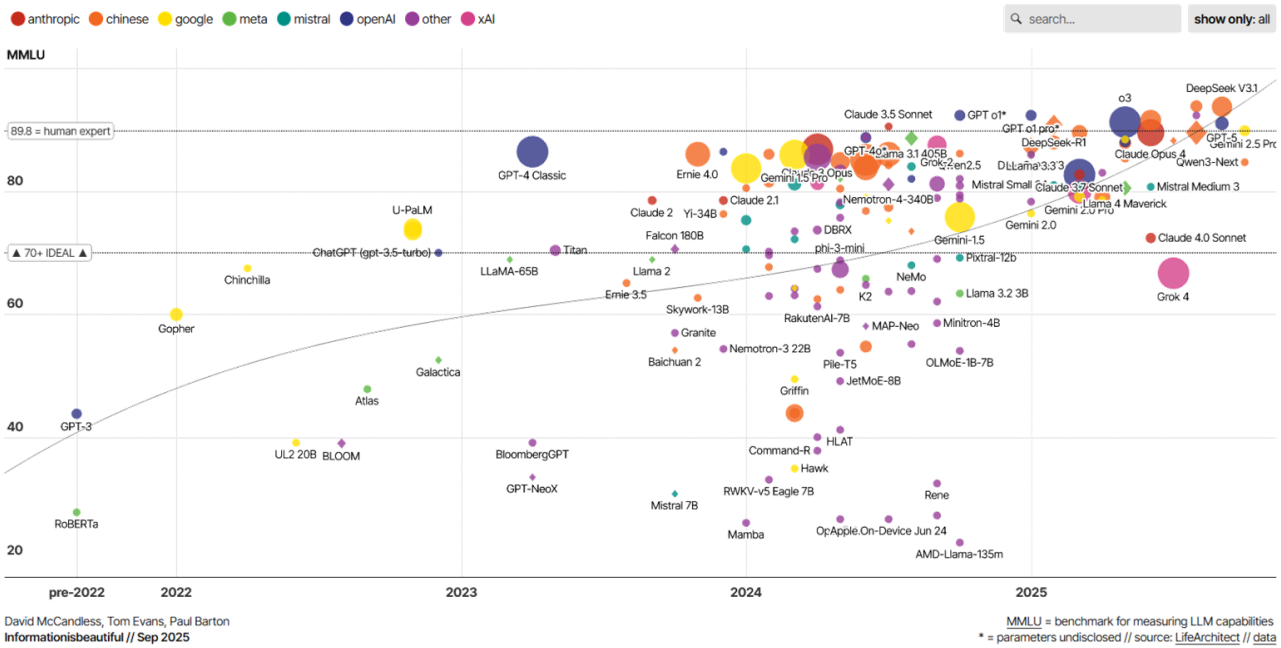


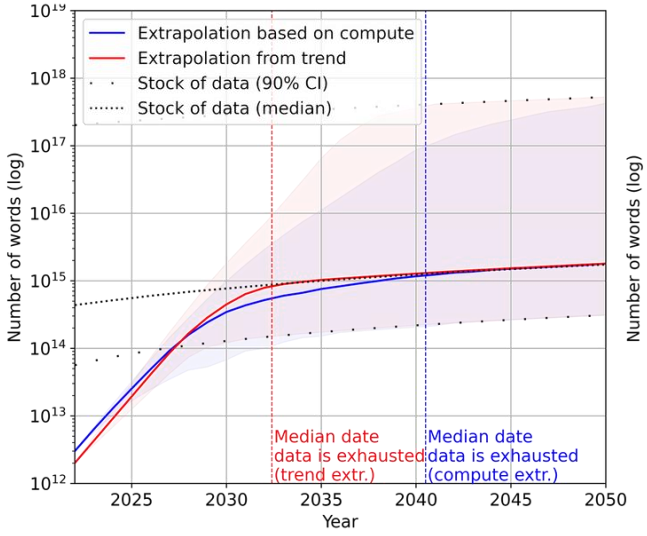




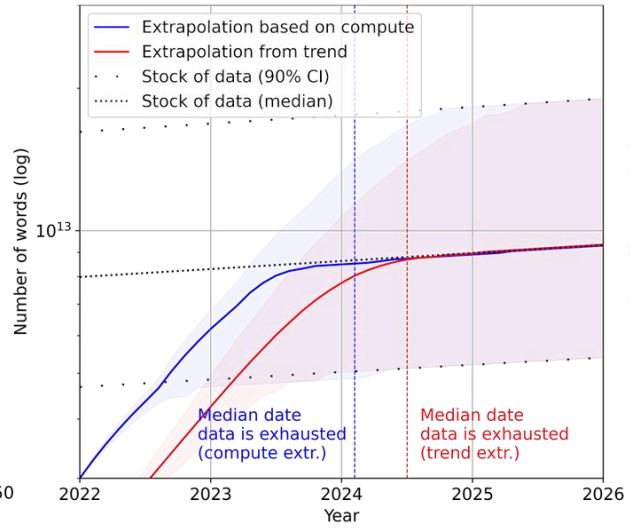
https://nerdynav.com/chatgpt-statistics/
https://www.statista.com/chart/29174/time-to-one-million-users/

Large Language Models & Training Data Size





Low-quality language data



High-quality language data

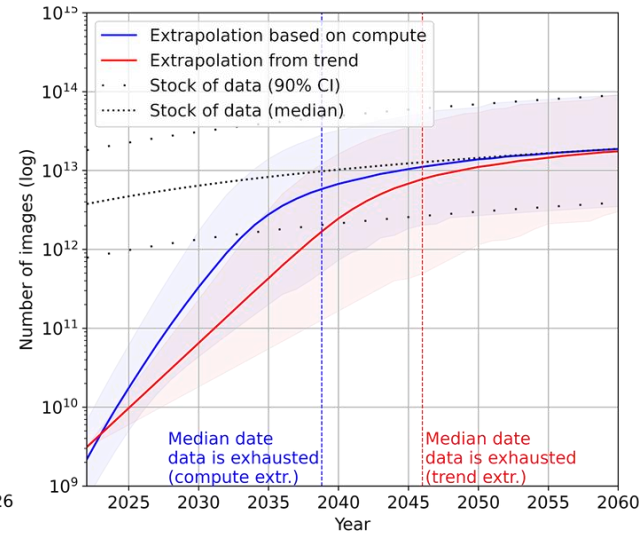


Image data

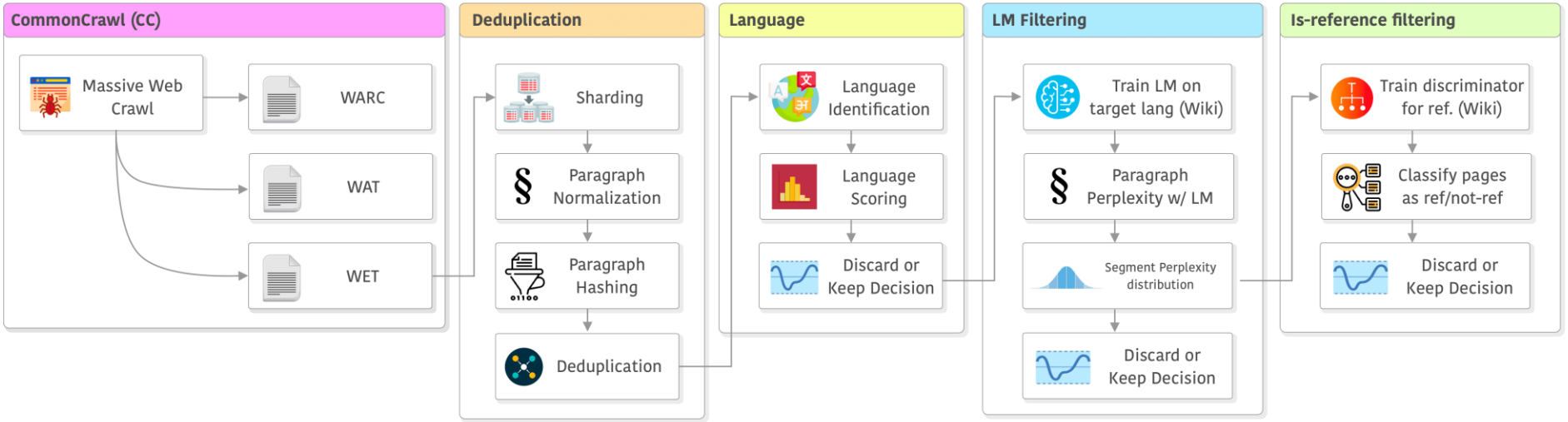
https://www.lesswrong.com/posts/asqDCb9XzXnLjSfgL/trends-in-training-dataset-sizes#Yearly_growth1
https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt/

LLaMA Training

LLaMA-65B training on **1.4T tokens** took ~**21 days** on **2048 NVIDIA A100 GPUs (80GB)**. Throughput of ~**380 tokens/sec/GPU**.

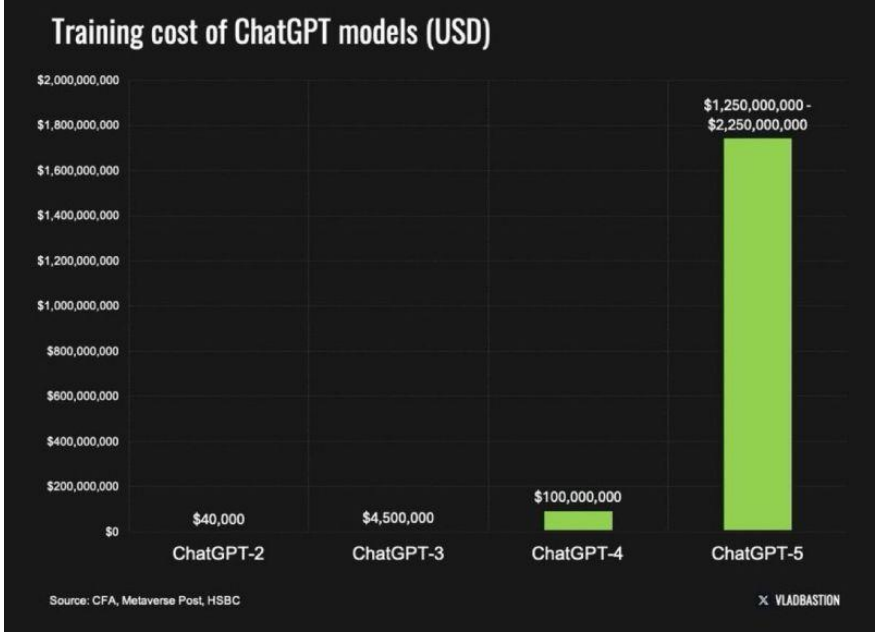
Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

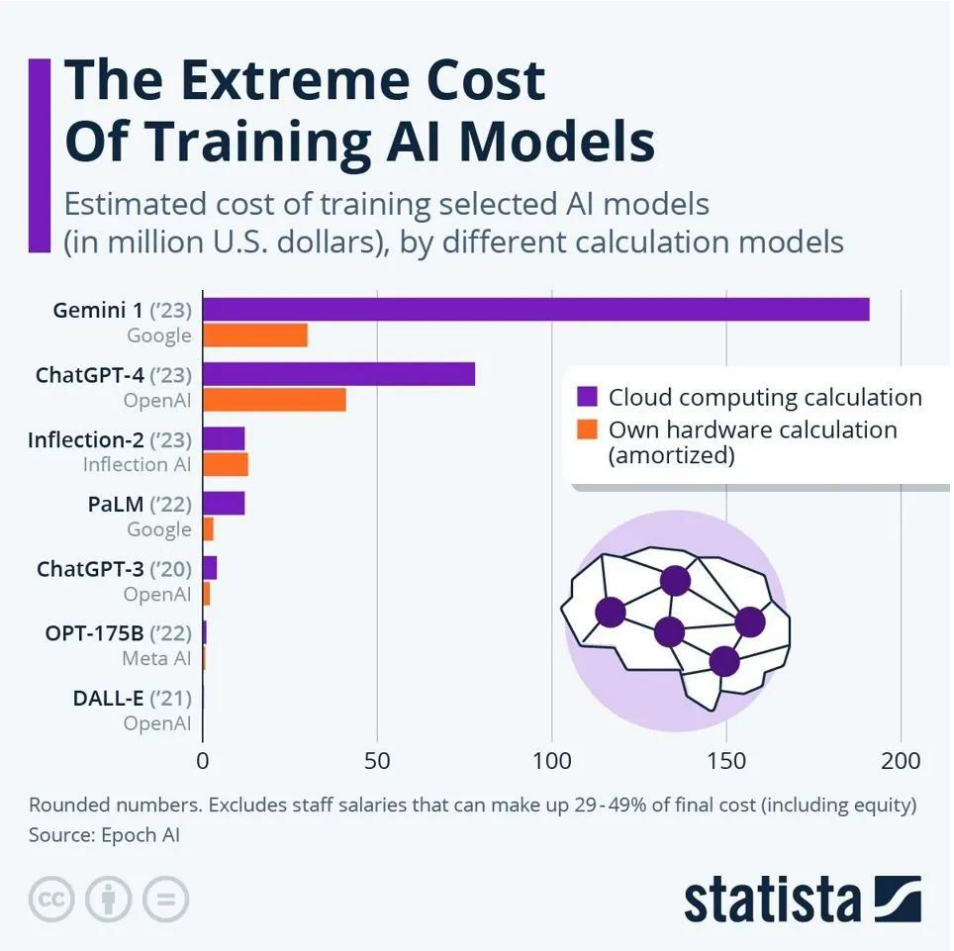
Table 1: **Pre-training data.** Data mixtures used for pre-training, for each subset we list the sampling proportion, number of epochs performed on the subset when training on 1.4T tokens, and disk size. The pre-training runs on 1T tokens have the same sampling proportion.



https://blog.christianperone.com/2023/06/appreciating-llms-data-pipelines/
https://arxiv.org/abs/2302.13971v1

Training is Costly!
And so is inferencing!





OpenAI training and inference costs could reach $7bn for 2024, AI startup set to lose $5bn - report

Details leak about its Microsoft Azure compute cluster

July 24, 2024 By: Sebastian Moss Comment

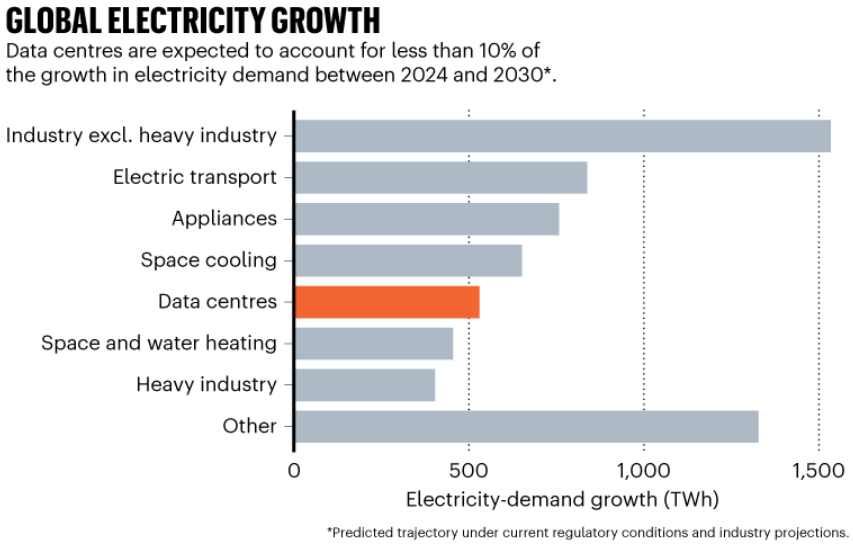
OpenAI is set to spend billions of dollars on training and inference this year, and may be forced to raise more money to cover growing losses.

The Information reports that, as of March, the company was set to spend nearly $4 billion this year on using Microsoft's servers to run inference workloads for ChatGPT.

OpenAI recently generated $283m in total revenue per month, which could mean full-year sales of between $3.5bn and $4.5bn.

https://www.forbes.com/sites/katharinabuchholz/2024/08/23/the-extreme-cost-of-training-ai-models/
https://www.linkedin.com/posts/jeremyprasetyo_leaked-chatgpt-5-will-cost-up-to-25-billion-activity-7201898490275254273-8BQq/
https://www.datacenterdynamics.com/en/news/openai-training-and-inference-costs-could-reach-7bn-for-2024-ai-startup-set-to-lose-5bn-report/

Power Saturation, Sustainability Challenges



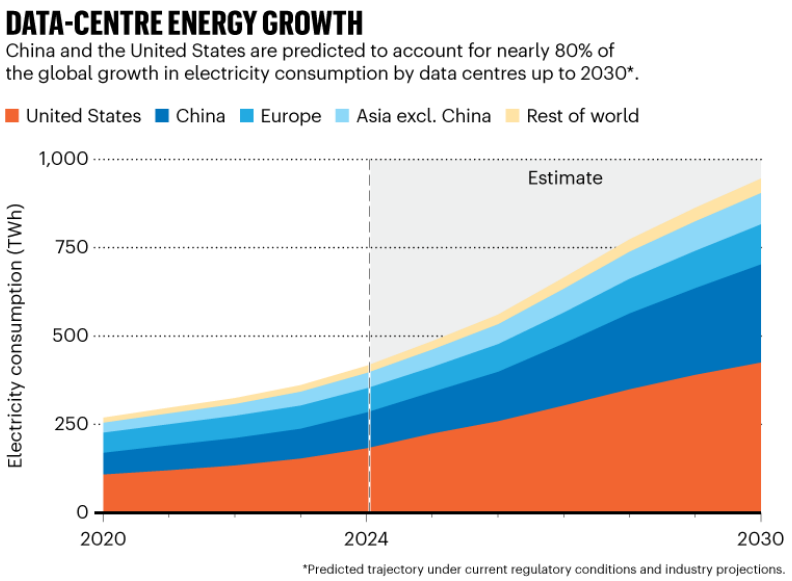
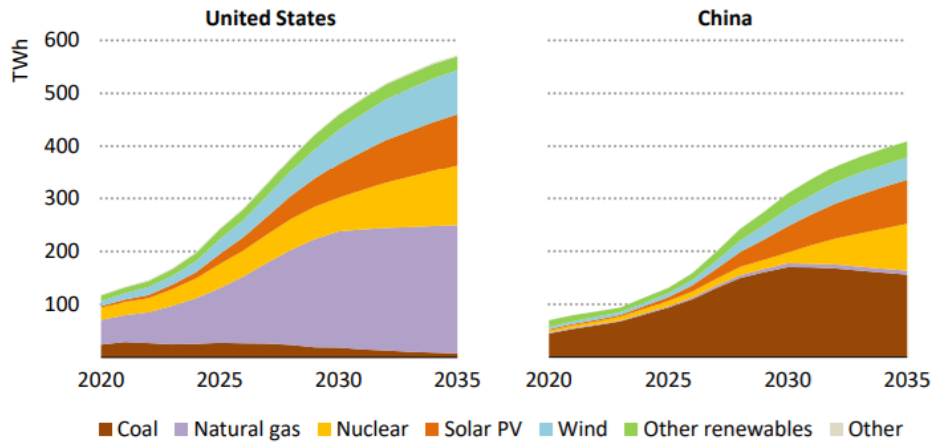
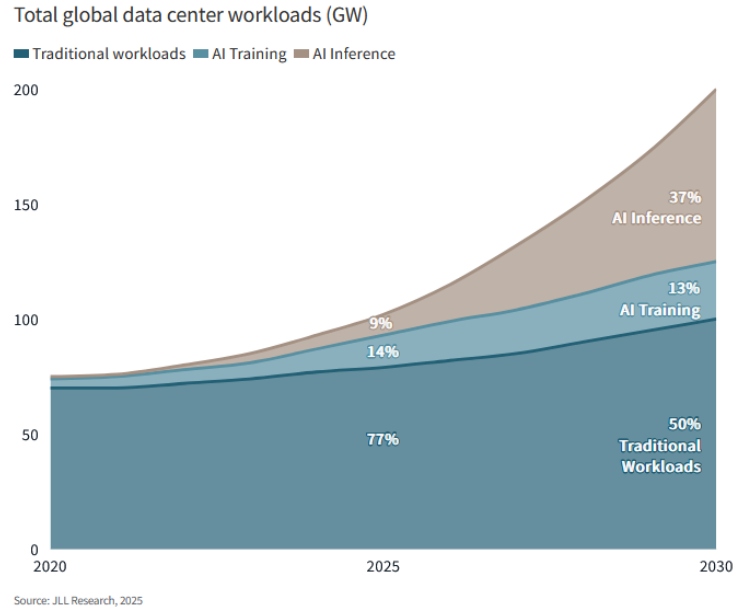


Figure 2.21 ▷ Electricity generation for data centres in the United States and China in the Base Case, 2020-2035

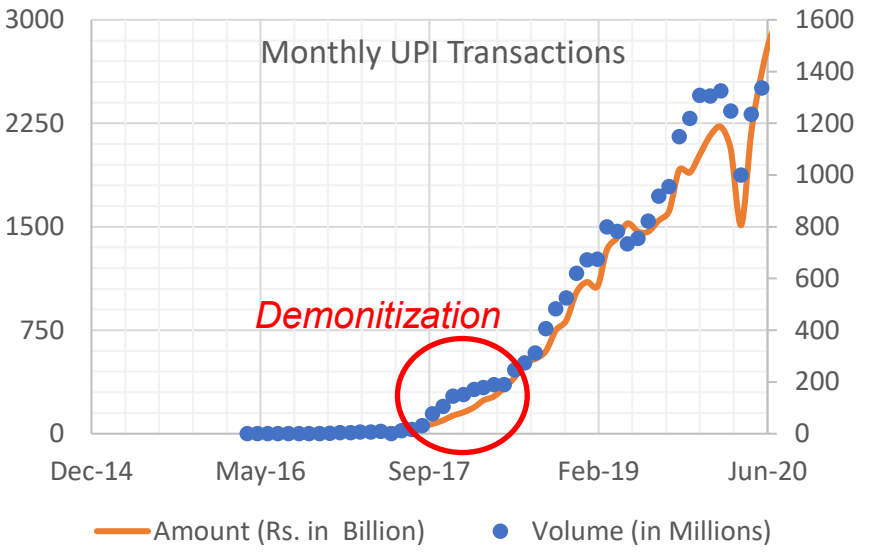


India as a country uses 1622 TWh

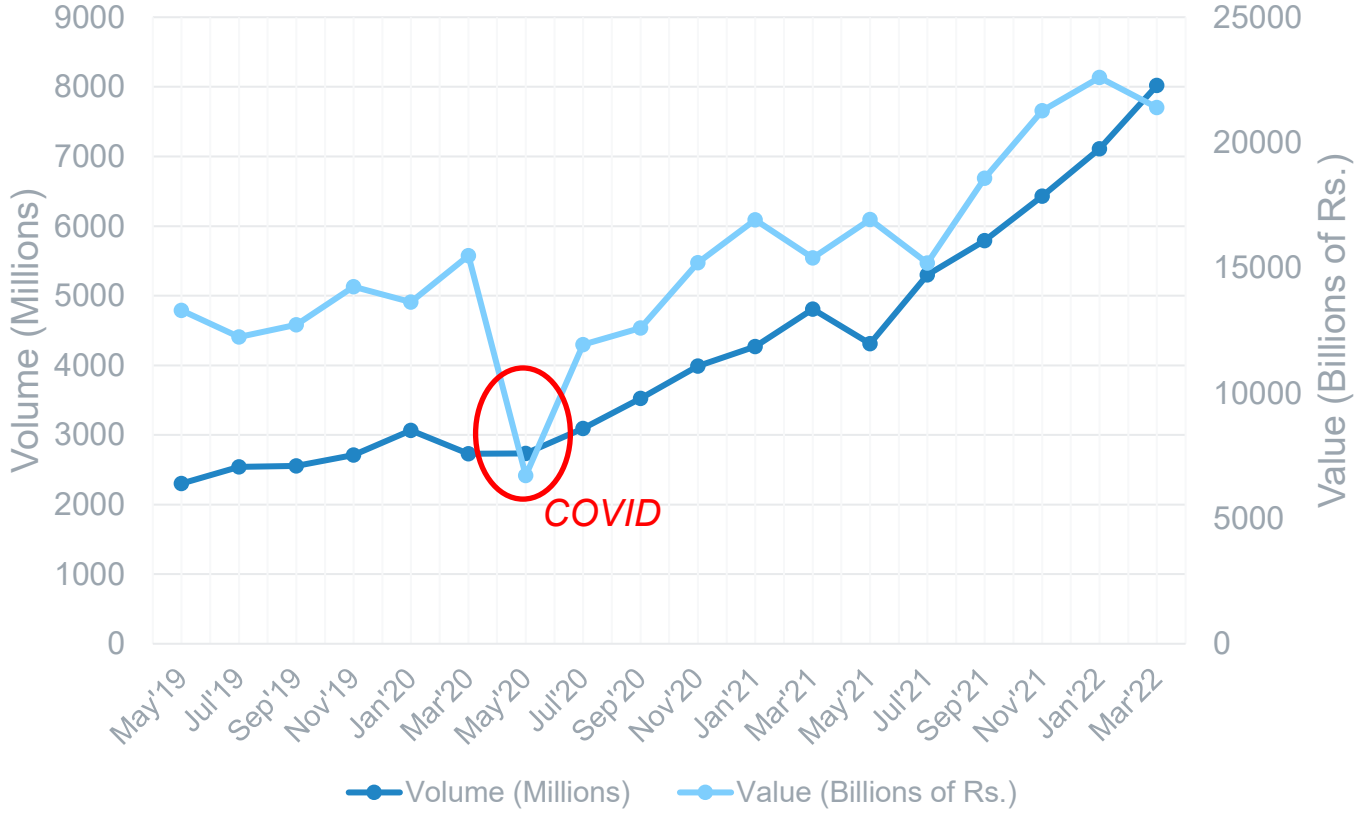


https://www.nature.com/articles/d41586-025-01113-z
https://www.jll.com/en-us/insights/market-outlook/data-center-outlook
https://www.fierce-network.com/cloud/will-new-nuclear-arrive-us-time-save-data-centers

FinTech: UPI Transactions



Monthly NCPI Retail Payment Transactions



https://www.npci.org.in/statistics

NPCI FinTech Case Study

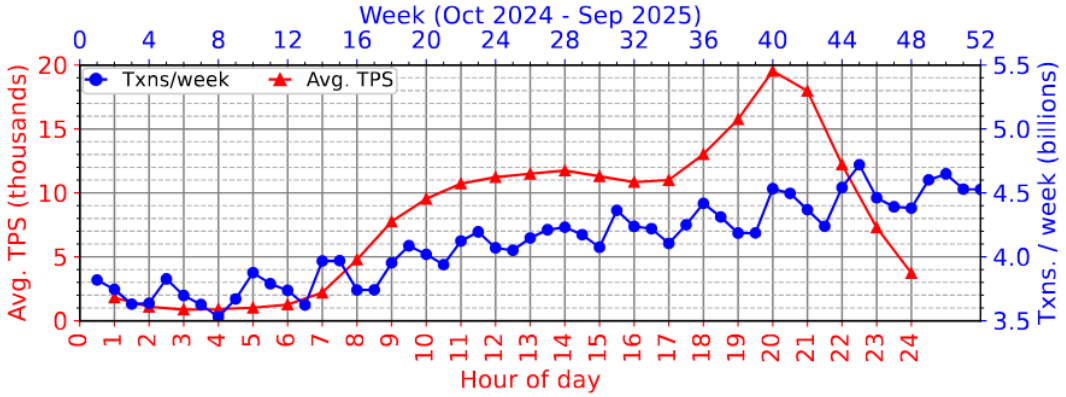
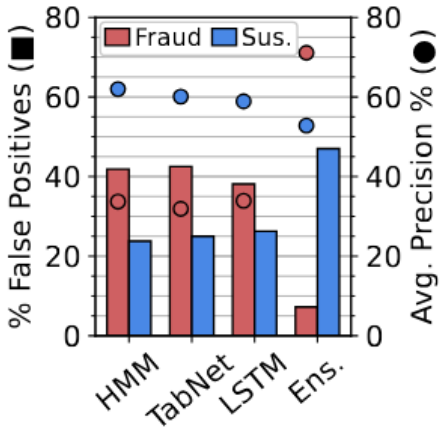


Fig. 1: TPS, averaged hourly for one day in Q1 of 2025 *(red, bottom X, left Y)*, and Weekly # of transactions in Q4 2024–Q3 2025 [7] *(blue, top X, inner right Y)* for UPI.



(a) False +ve & Precision%

- 200M in daily active unique accounts
- 185B transactions annually
- 100+ ML features
- Ingest rate of over 20,000 Transactions per Second (TPS)
- 500ms latency budget within the data center

NPCI FinTech Case Study

Operation	Scalable Framework used	Data size	Number of entities	Data size
Transaction Tuples to Temporal Graph (Multiple data joins and map operations)	PySpark	1TB -> 350 GB	500 million users, 5 billion trans	4.5TB
Graph Ft extract	WICM	350 GB -> 4 TB	500 million users * 91 days * 12 features	4 TB
Time series extracation	PySpark	4 TB -> 200 GB	500 million users * 120 features	~5TB

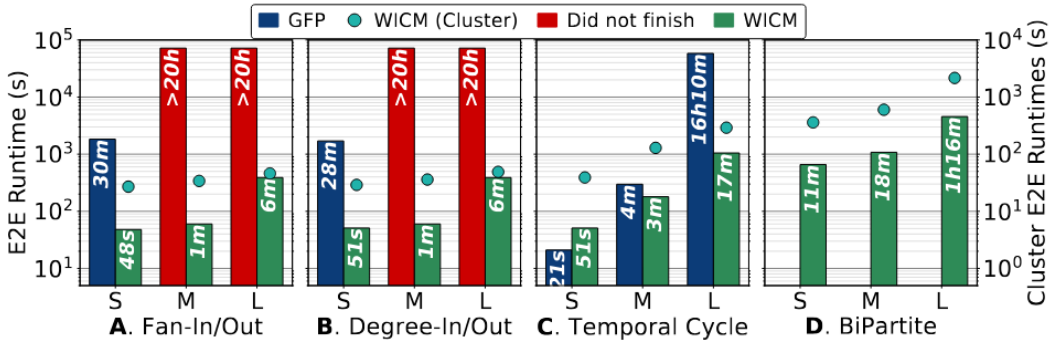


Fig. 2: GFP vs WICM E2E Runtime Comparisons

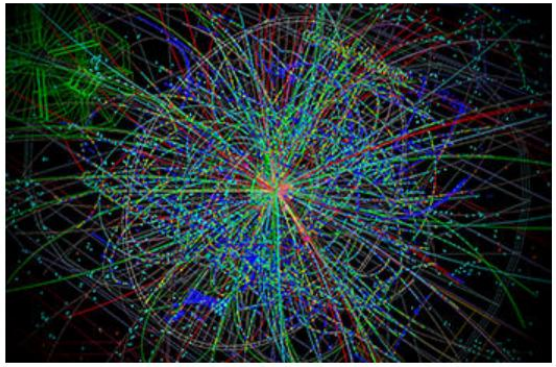
Towards Scalable Mining of Temporal GraphMotifs over Large-Scale Transaction Networks, *HiPC SRS* 2026

Big Data and Science

AI's early proving ground: the hunt for new particles

Particle physicists began fiddling with artificial intelligence (AI) in the late 1980s, just as the term "neural network" captured the public's imagination. Their field lends itself to AI and machine-learning algorithms because nearly every experiment centers on finding subtle spatial patterns in the countless, similar readouts of complex particle detectors—just the sort of thing at which AI excels. "It took us several years to convince people that this is not just some magic, hocus-pocus, black box stuff," says Boaz Klima, of Fermi National Accelerator Laboratory (Fermilab) in Batavia, Illinois, one of the first physicists to embrace the techniques. Now, AI techniques number among physicists' standard tools.

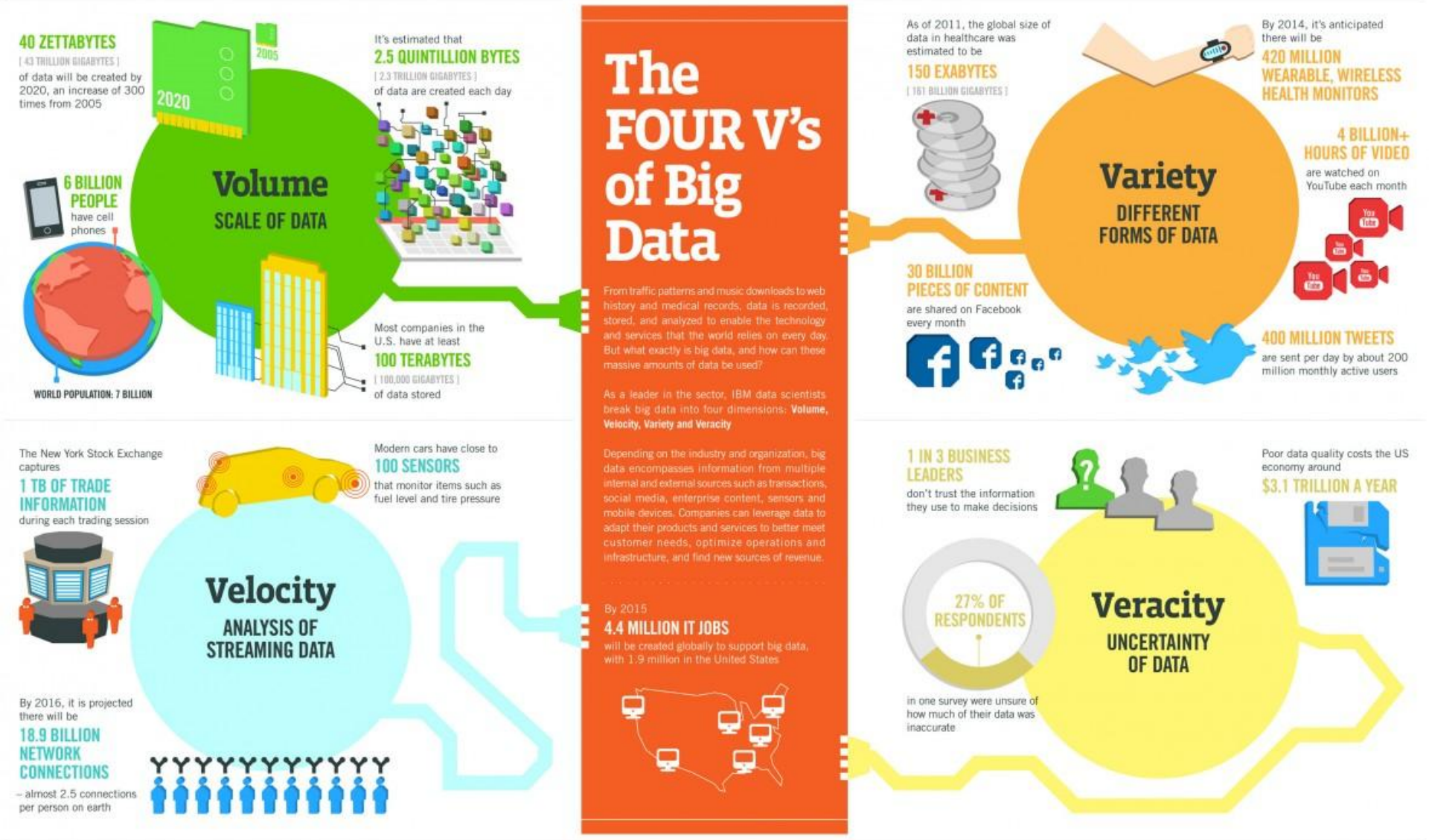
Particle physicists strive to understand the inner workings of the universe by smashing subatomic particles together with enormous energies to blast out exotic new bits of matter. In 2012, for example, teams working with the world's largest proton collider, the Large Hadron Collider (LHC) in Switzerland, discovered the long-predicted Higgs boson, the fleeting particle that is the linchpin to physicists' explanation of how all other fundamental particles get their mass.



Neural networks search for fingerprints of new particles in the debris of collisions at the LHC. © 2012 CERN, FOR THE BENEFIT OF THE ALICE COLLABORATION

Such exotic particles don't come with labels,

https://www.sciencemag.org/news/2017/07/ai-changing-how-we-do-science-get-glimpse



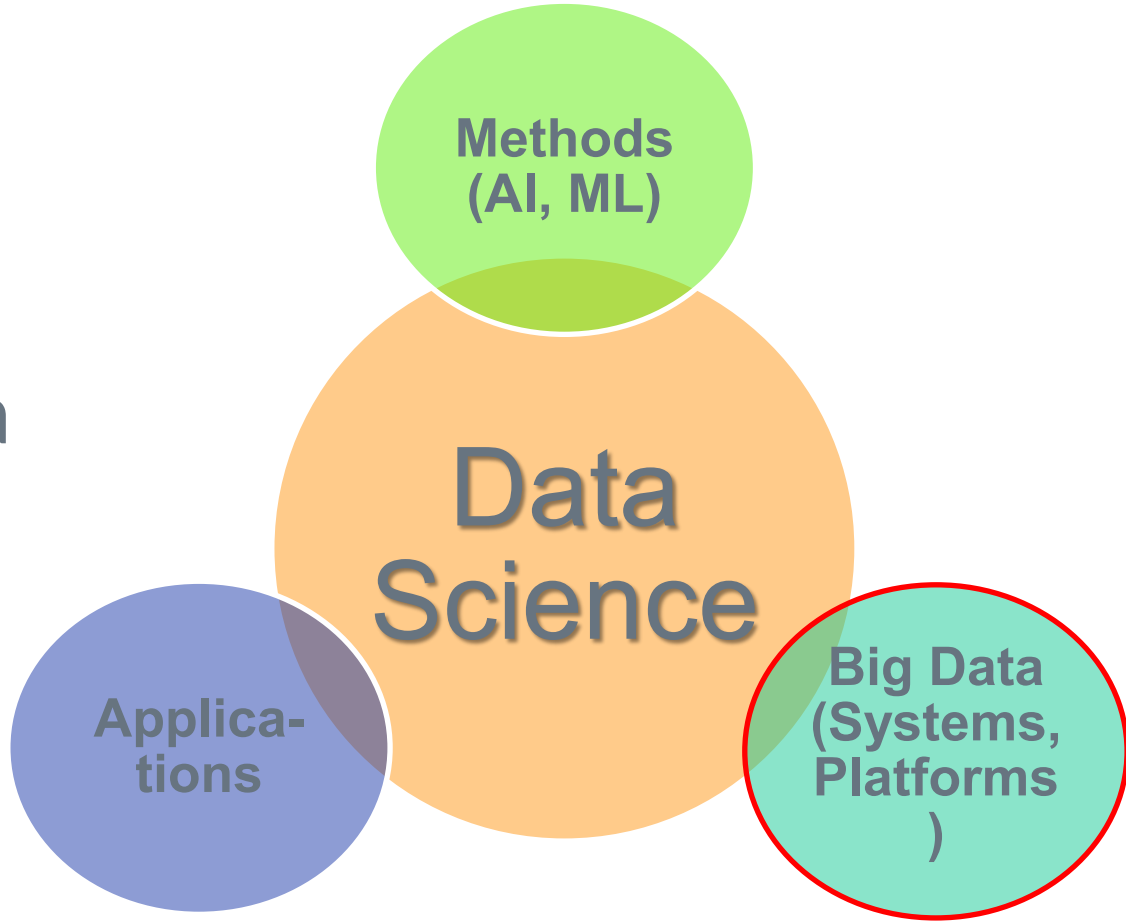
http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Big Data Landscape

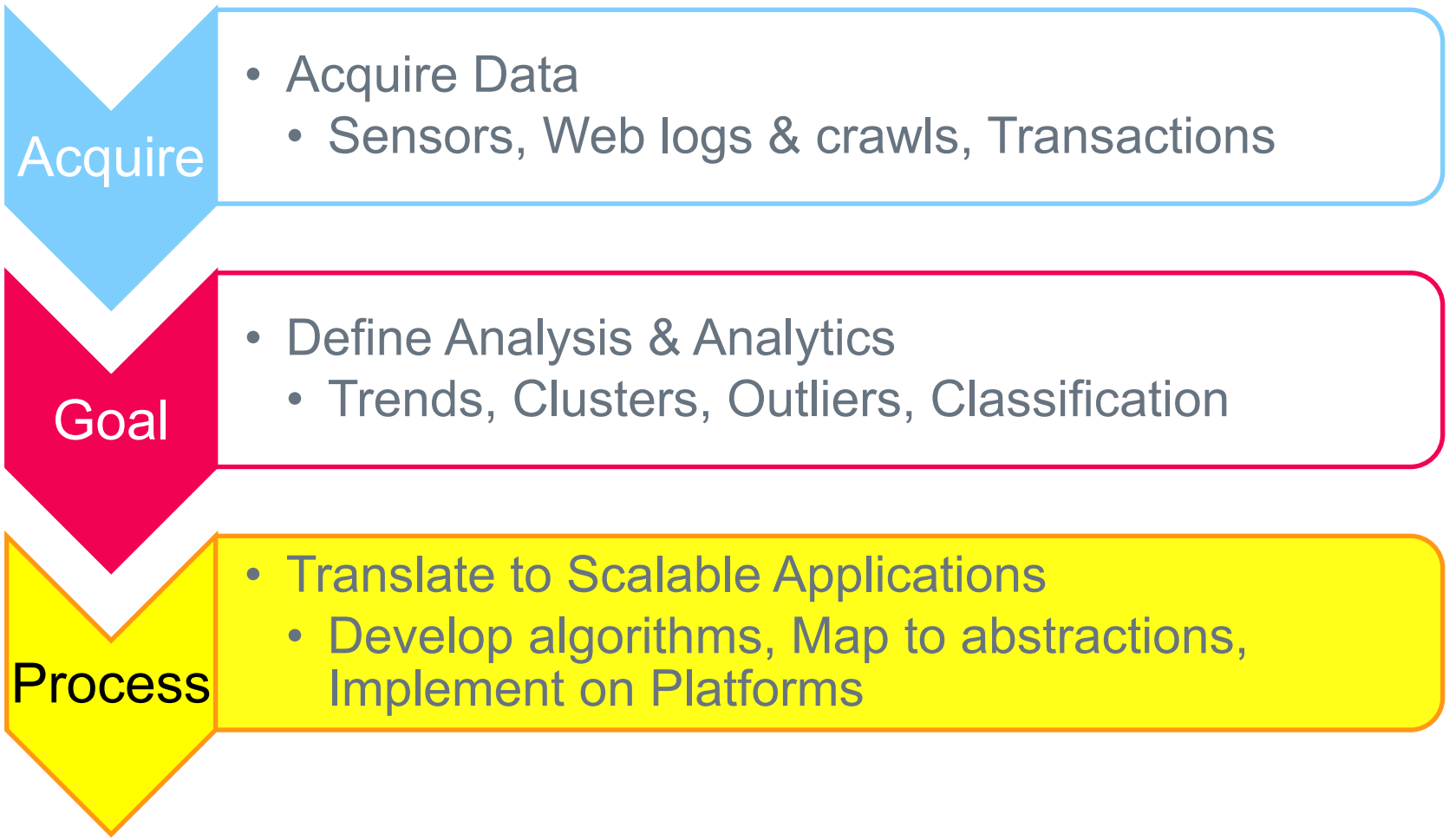
Resilience and Vibrancy: The 2020 Data & AI Landscape, Matt Turck,
https://mattturck.com/data2020/
The 2021 Machine Learning, AI and Data (MAD) Landscape,
https://mattturck.com/data2021/

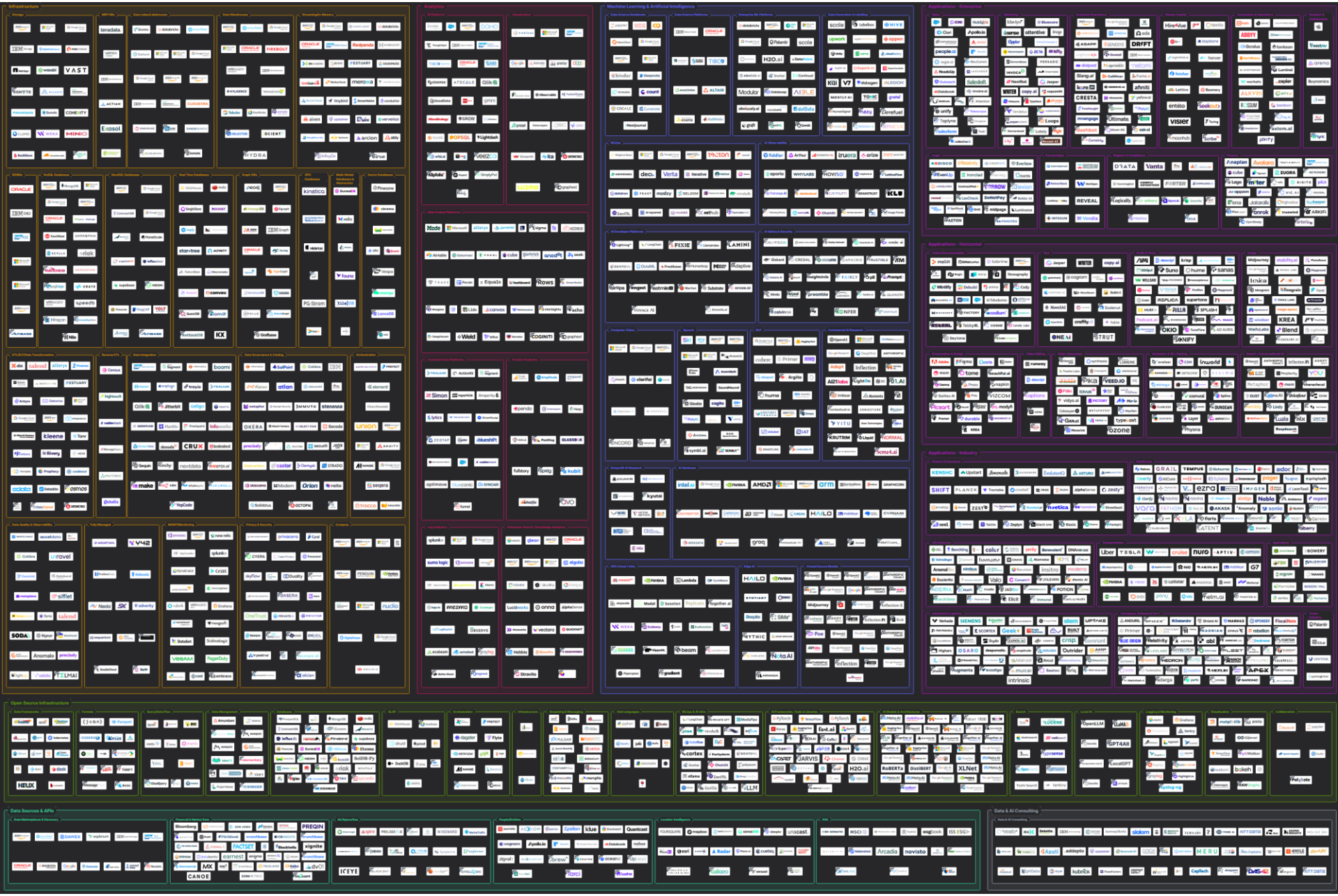
Data Science

Inter-disciplinary domain at the intersection of data analysis methods, Big Data Systems and data-driven applications.

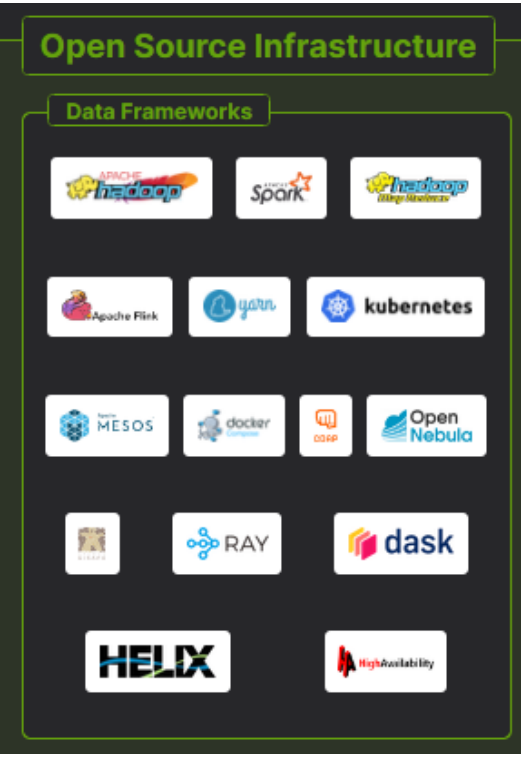


Data Analysis Lifecycle



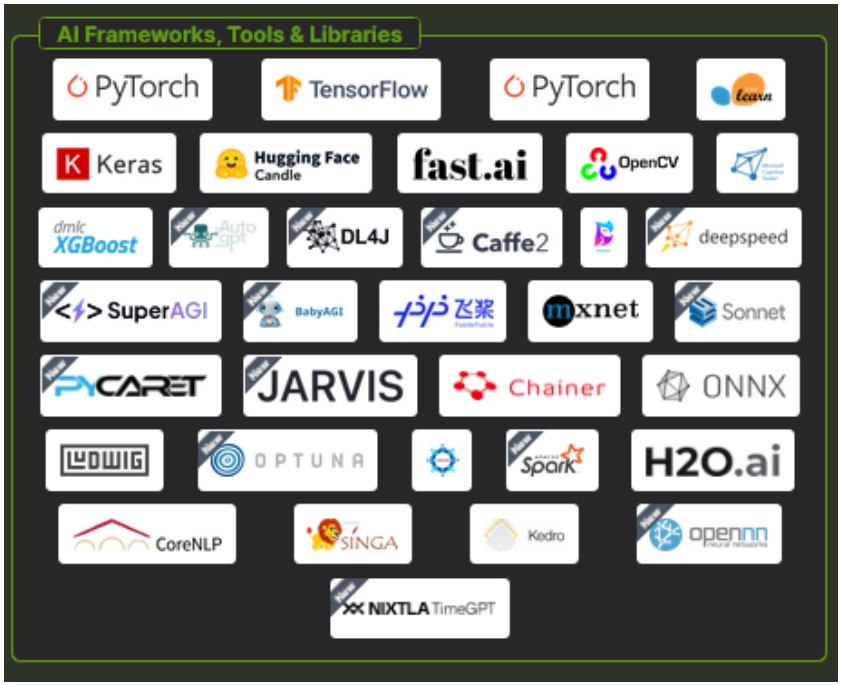


https://mad.firstmark.com/


Open Source Infrastructure
Data Frameworks
Apache Hadoop
Spark
Hadoop
Apache Flink
yarn
kubernetes
MESOS
docker Compose
CORD
Open Nebula
RAY
dask
HELIX
HighAvailability


Streaming & Messaging
Spark Streaming
kafka
Apache Flink
beam
Apache Storm
PULSAR
RabbitMQ
Apache RocketMQ
samza
nifi
debezium
ESTUARY Flow
ACTIVE MQ
mosquitto
memphis
bytewax
Gazette

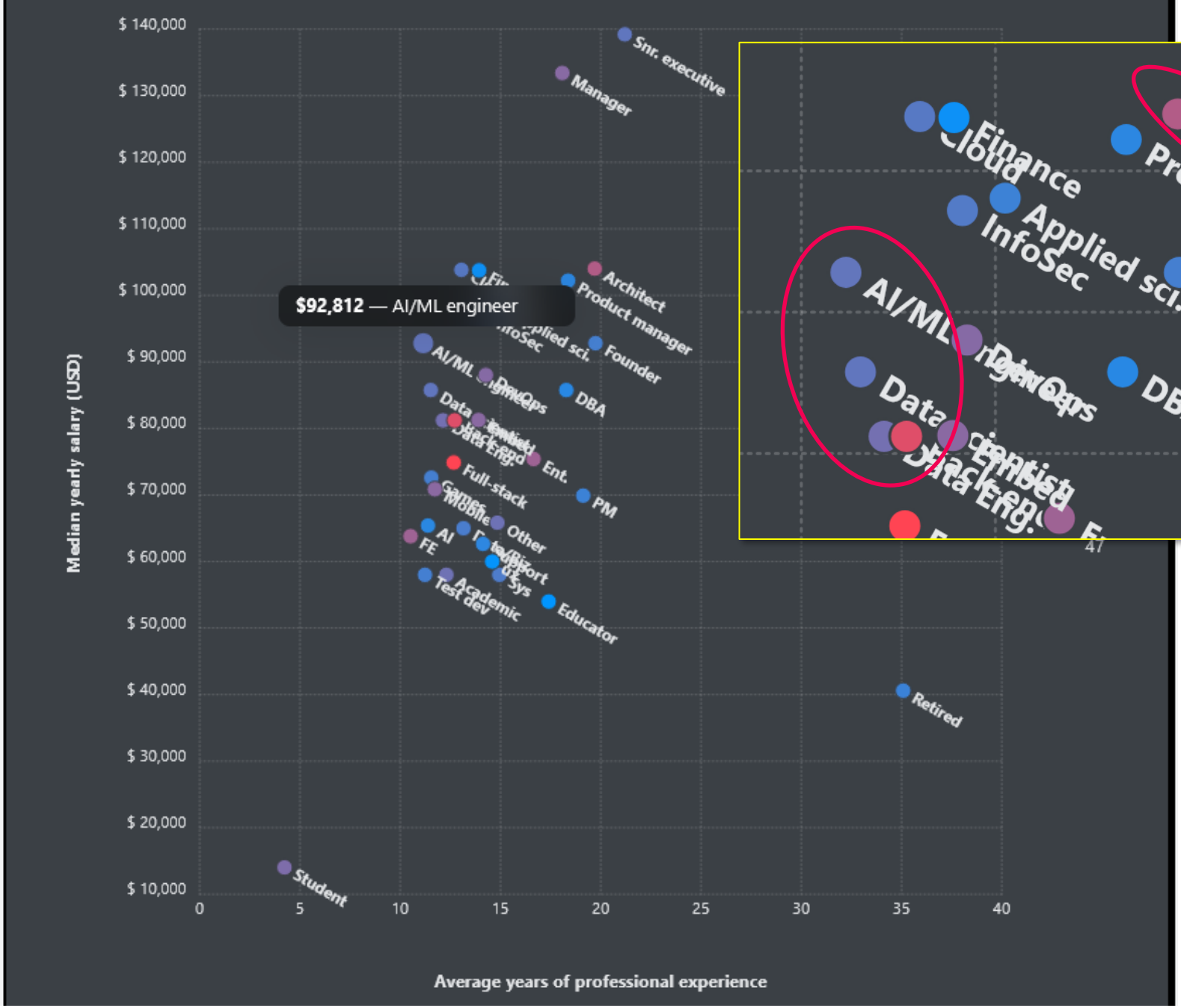

Databases
PostgreSQL
MySQL
MongoDB.
redis
CockroachDB
SQLite
CASSANDRA
MariaDB
influxdb
HBASE
Firebird
supabase
Timescale
SurrealDB
Milvus
Chroma
weaviate
NEON
EDGE DB
RocksDB
SciDB-Py
OpenTSDB
CouchDB
LEVELDB
riak
accumulo
Ignite
FoundationDB
TiKV
spicedb
OCEANBASE


AI Frameworks, Tools & Libraries
PyTorch
TensorFlow
PyTorch
learn
Keras
Hugging Face Candle
fast.ai
OpenCV
dmlc XGBoost
Auto gpt
DL4J
Caffe2
deepspeed
SuperAGI
BabyAGI
飞桨
mxnet
Sonnet
PyCARET
JARVIS
Chainer
ONNX
LUDWIG
OPTUNA
Spark
H2O.ai
CoreNLP
SINGA
Kedro
opennn
NIXTLA TimeGPT

Top Paying Technologies ☺

Topics you are learning in this module are the top-3 paying Databases and Frameworks globally!

https://survey.stackoverflow.co/2022/#top-paying-technologies-other-frameworks-and-libraries



https://survey.stackoverflow.co/2025/work#salary

SageServe: Optimizing LLM Serving on Cloud Data Centers with Forecast Aware Auto-Scaling

SHASHWAT JAISWAL*†, University of Illinois Urbana-Champaign, USA
KUNAL JAIN*‡, Georgia Institute of Technology, USA
YOGESH SIMMHAN, Indian Institute of Science, India
ANJALY PARAYIL, ANKUR MALLICK, RUJIA WANG, RENEE ST. AMANT, CHETAN BANSAL, VICTOR RUHLE, ANOOP KULKARNI, STEVE KOFSKY, SARAVAN RAJMOHAN, Microsoft, India, UK and USA

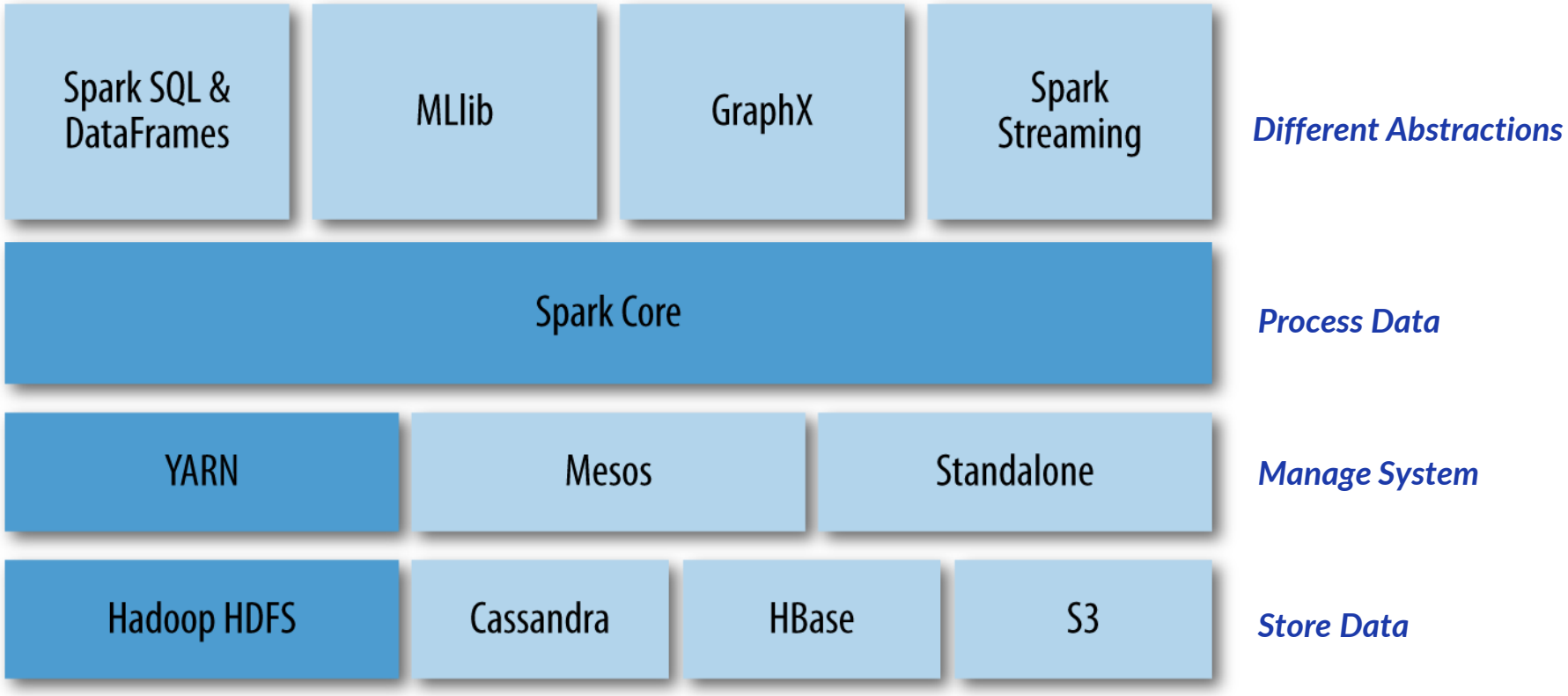
Global cloud service providers handle inference workloads for Large Language Models (LLMs) that span latency-sensitive (e.g., chatbots) and insensitive (e.g., report writing) tasks, resulting in diverse and often conflicting Service Level Agreement (SLA) requirements. Managing such mixed workloads is challenging due to the complexity of the inference serving stack, which encompasses multiple models, GPU hardware,

realistic simulations on 10 million production requests across three regions and four open-source models. We achieve up to 25% savings in GPU-hours compared to the current baseline deployment and reduce GPU-hour wastage due to inefficient auto-scaling by 80%, resulting in a potential monthly cost savings of up to $2.5 million, while maintaining tail latency and meeting SLAs. The workload traces, our simulator harness and the

public studies of Internet-scale LLM workloads. We use these insights to propose SageServe, a comprehensive LLM serving framework that dynamically adapts to workload demands using multi-timescale control knobs. It combines short-term request routing to data centers with long-term scaling of GPU VMs and model placement with higher lead times, and co-optimizes the routing and resource allocation problem using a traffic forecast model and an Integer Linear Programming (ILP) solution. We evaluate SageServe through real runs and realistic simulations on 10 million production requests across three regions and four open-source models. We achieve up to 25% savings in GPU-hours compared to the current baseline deployment and reduce GPU-hour wastage due to inefficient auto-scaling by 80%, resulting in a potential monthly cost savings of up to $2.5 million, while maintaining tail latency and meeting SLAs. The workload traces, our simulator harness and the are available at https://github.com/shashwatj07/SageServe.

ACM SIGMETRICS 2026

Big Data Platform Stack, *Spark Flavor*



https://www.oreilly.com/library/view/data-analytics-with/9781491913734/ch04.html

Additional Reading

- A Survey of Big Data Research, H Fang, et al., *IEEE Network*, September/October 2015, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4617656/
- Beyond the hype: Big data concepts, methods, and analytics, A. Gandomi and M. Haider, *International Journal of Information Management*, Volume 35, Issue 2, 2015, https://doi.org/10.1016/j.ijinfomgt.2014.10.007
- Uncertainty in big data analytics: survey, opportunities, and challenges. R.H. Hariri, et al. *J Big Data* **6**, 44 (2019). https://doi.org/10.1186/s40537-019-0206-3