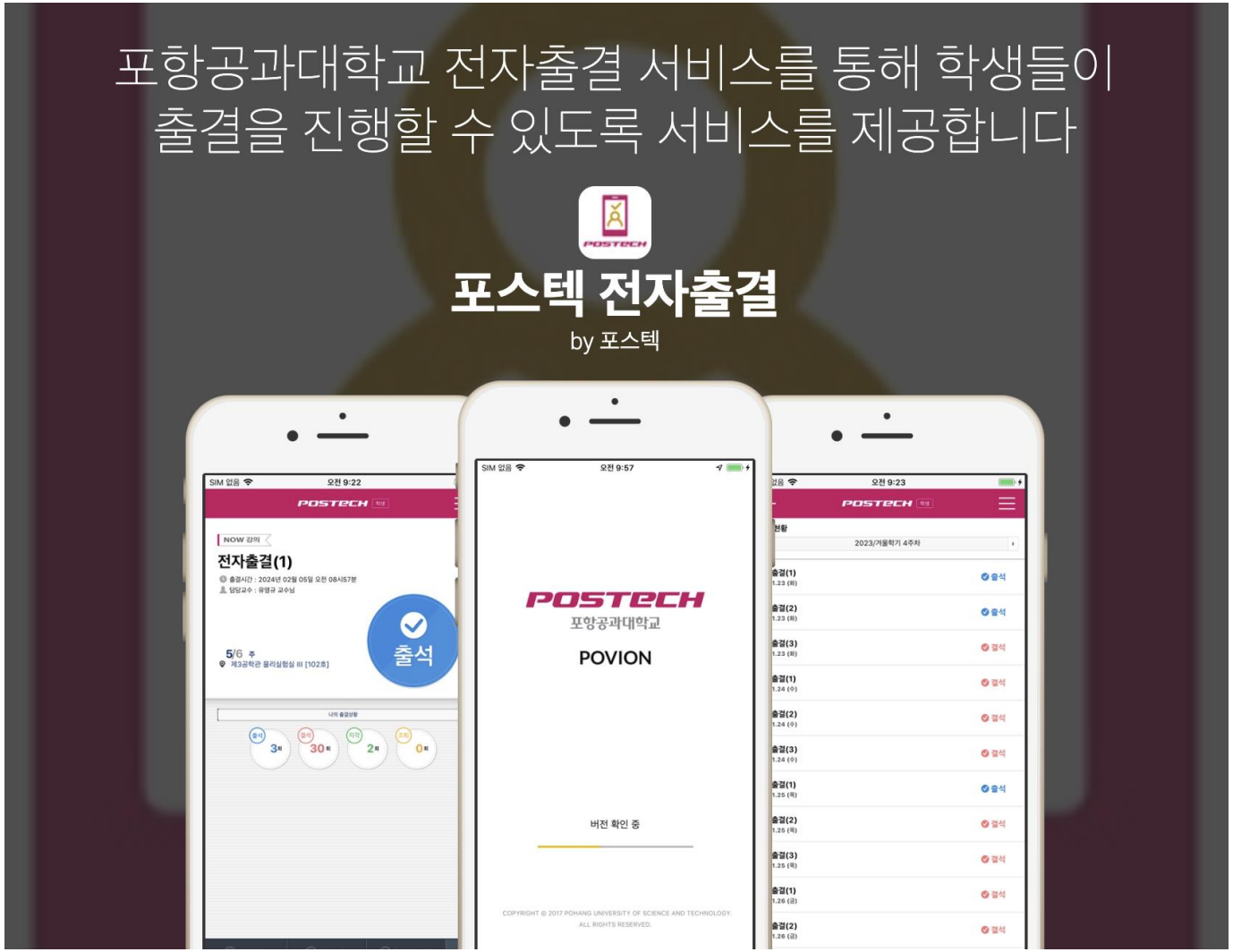
IMEN891K/IDSC501 Data Engineering

2. **Data Collection**

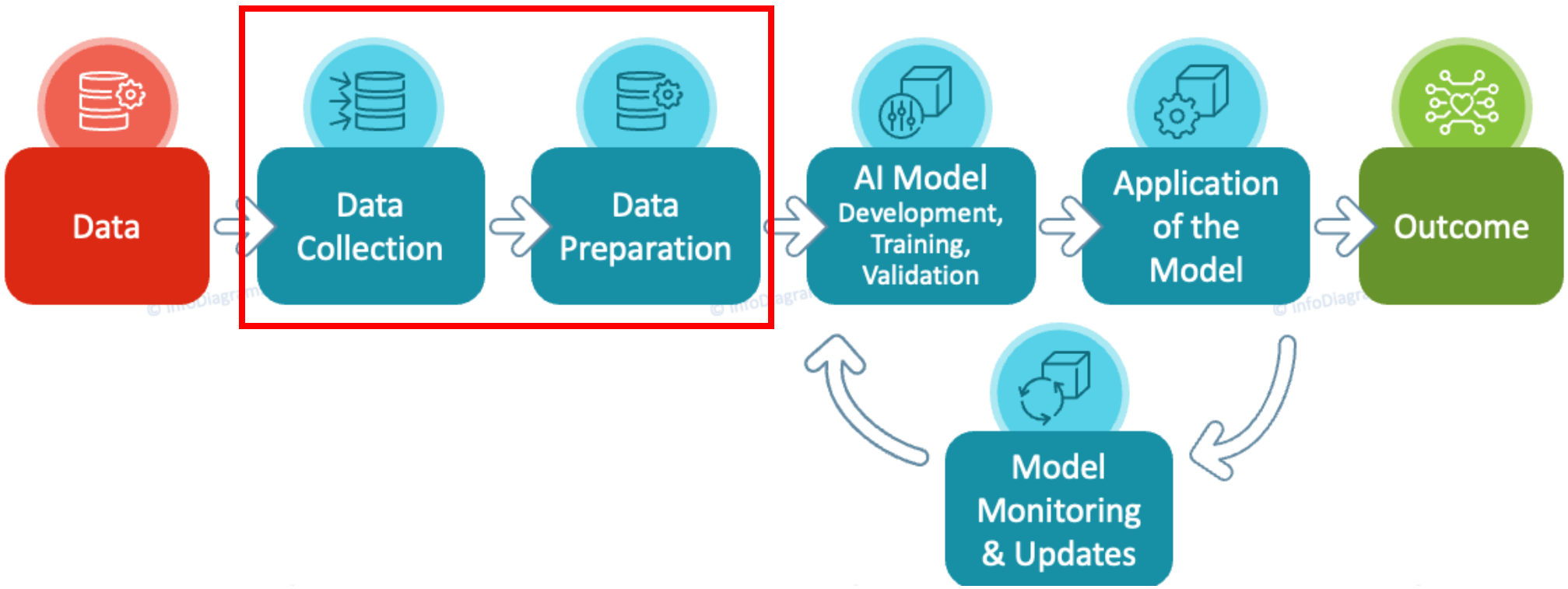
Fall 2025

Instructor: **Sangwoo Mo**

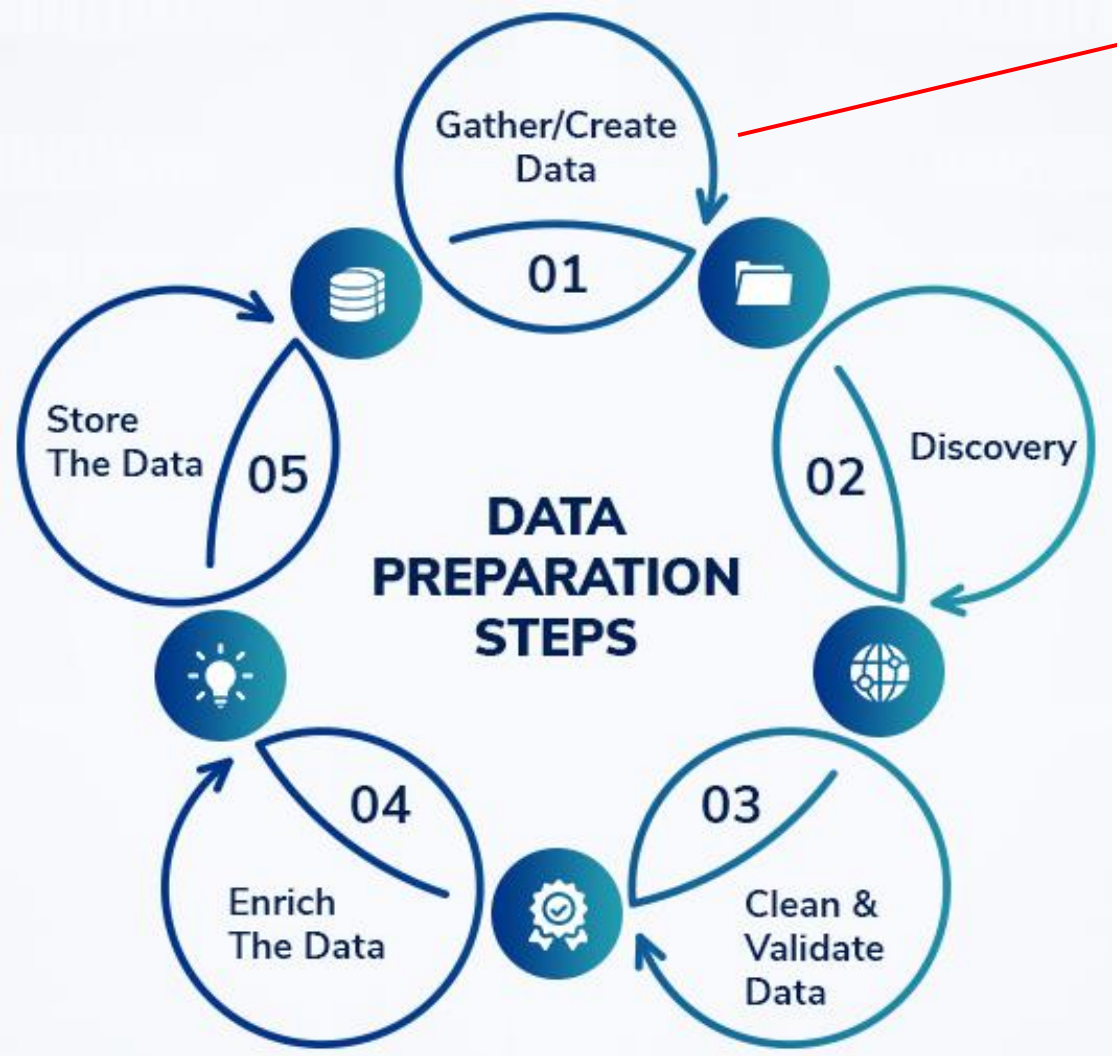
Recap: Attendance



Recap: Course Overview



Recap: Course Overview



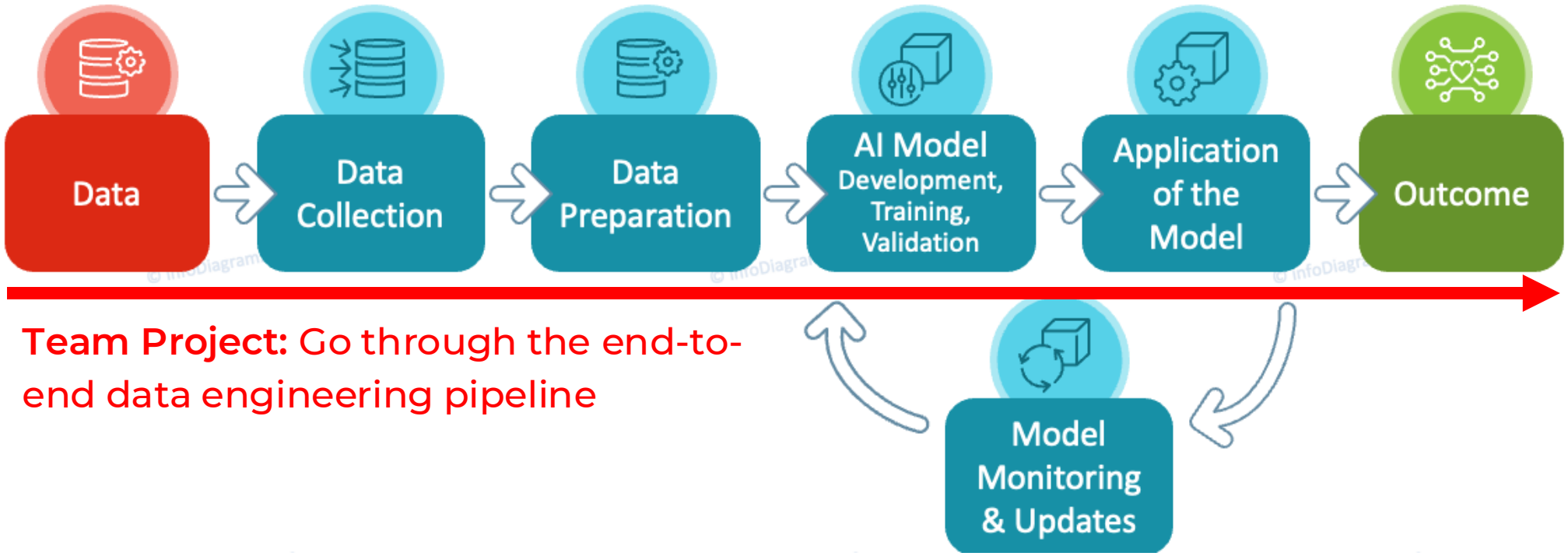
Lecture 2

Recap: Grading Policy

- Attendance (**10%**)
- Midterm Exam (**25%**)
- Final Exam (**25%**)
- Team Project (**40%**)

Office Hours: **Thursdays 11:00-12:00 pm, Room 319**

Recap: Team Project



Team Project: Go through the end-to-end data engineering pipeline

Recap: Team Project

- **Proposal** (mid-semester):
  - Ideation, data collection & preprocessing, DB management

- **Final** (end-semester):
  - Develop a product that addresses a real-world problem using a large-scale DB (with potential LLM integration)

Recap: Team Project

Evaluation Criteria

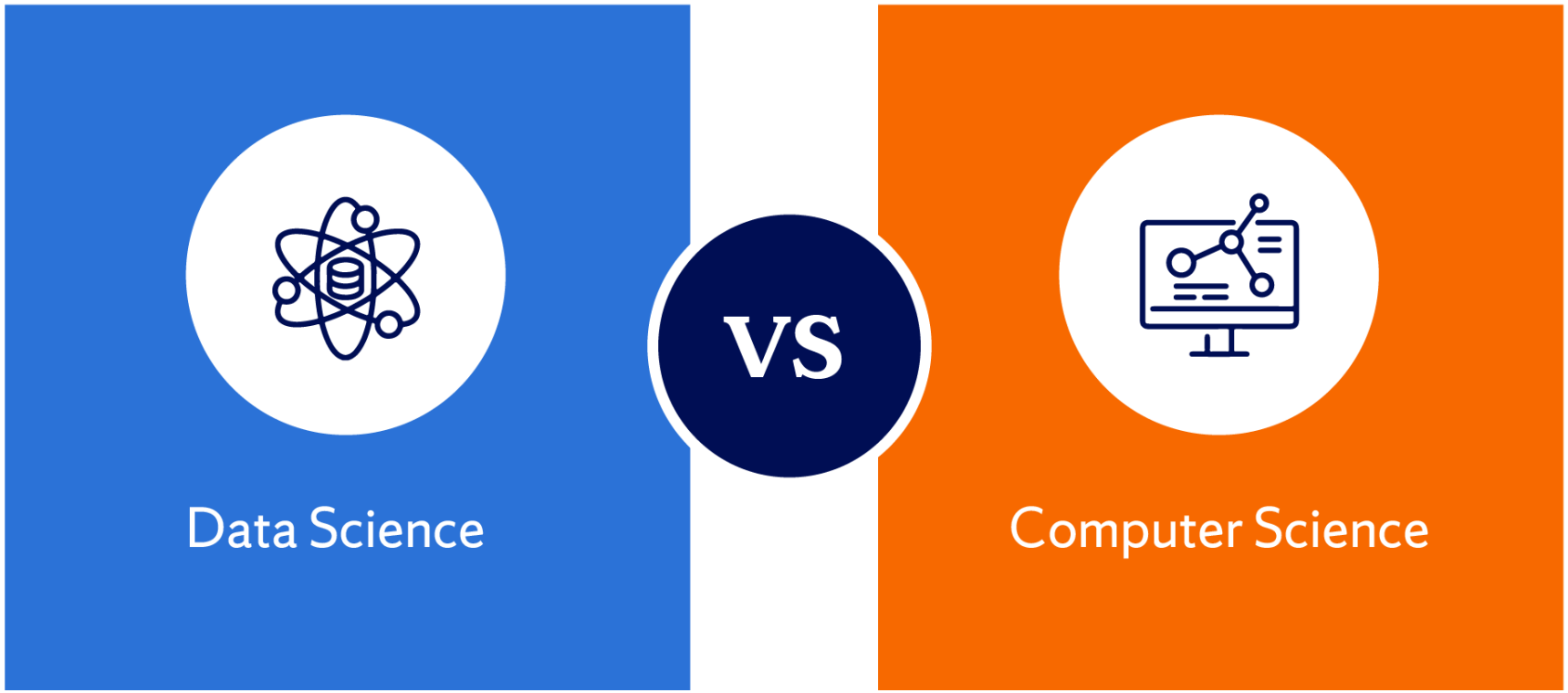
- **Novelty:** originality and creativity of the idea
- **Impact:** relevance and potential usefulness of the solution
- **Technology:** effective use of the techniques and tools covered in class
- **Presentation:** clarity and professionalism of the report and presentation

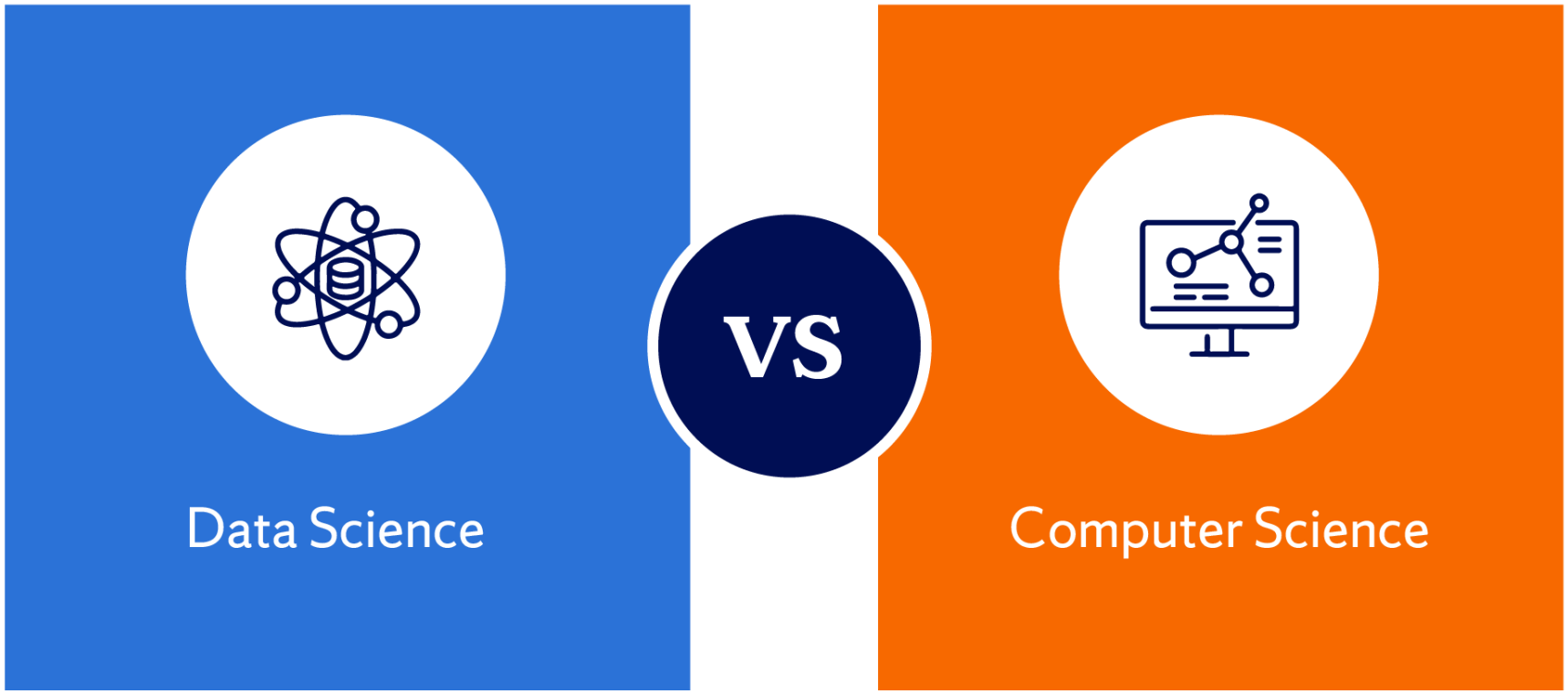
Team Size: **3~5**

Recap: Team Project

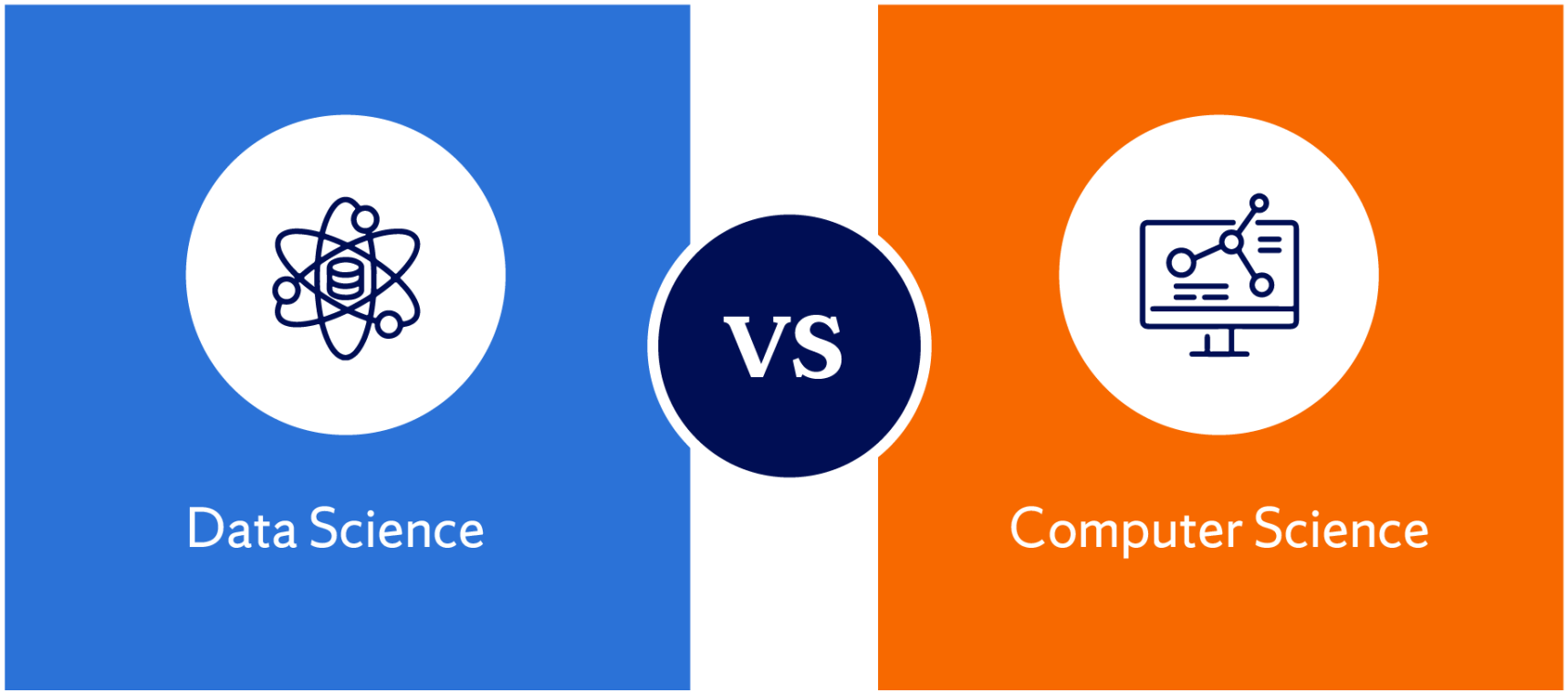
Build your team! **(Deadline: Sep 18)**




VS
Data Science
Computer Science



Technology itself



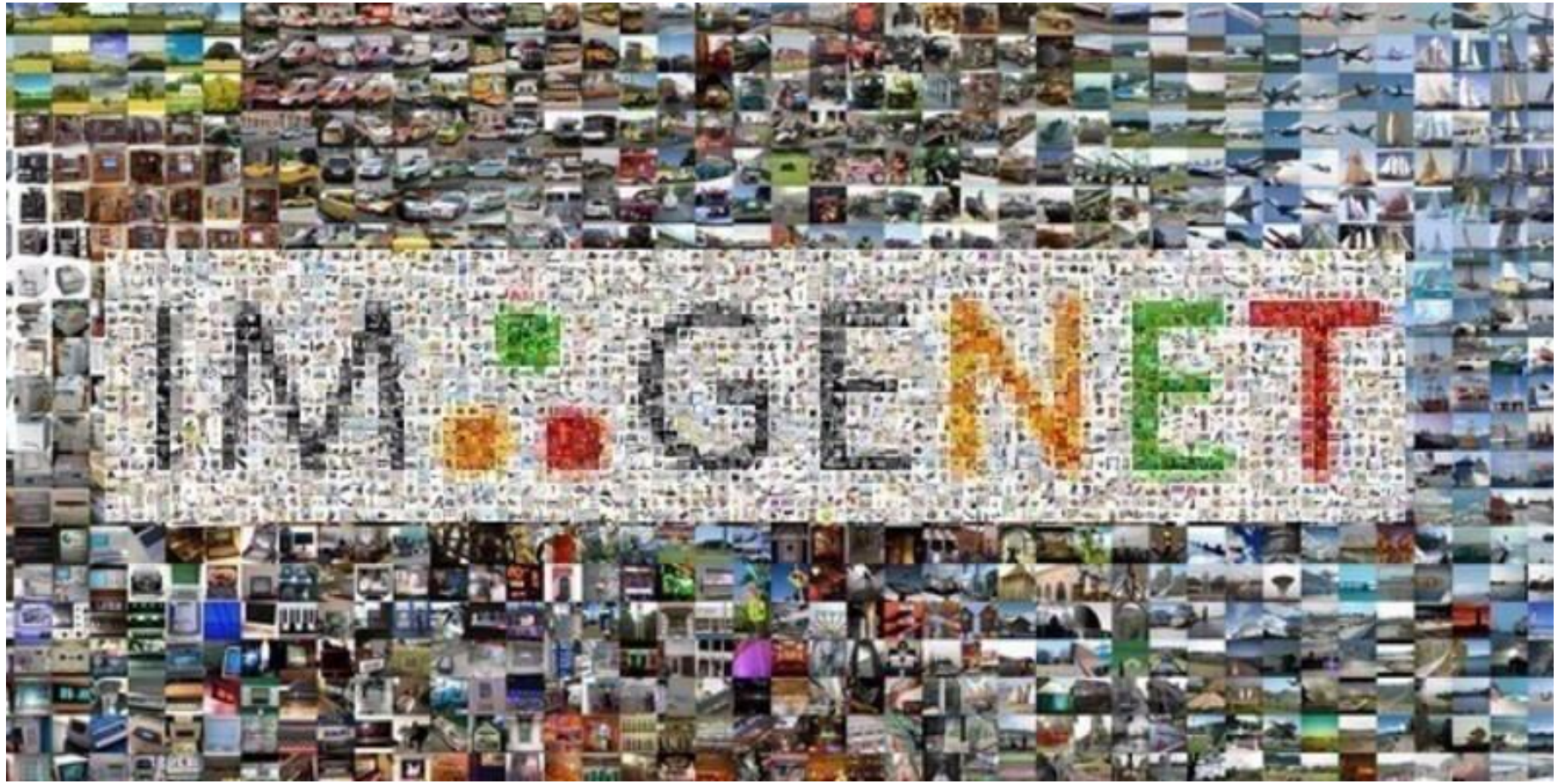
Solve **real** problems

Technology itself

Data Collection



Data Collection



No need to worry. Just use **benchmark** datasets

Data Collection

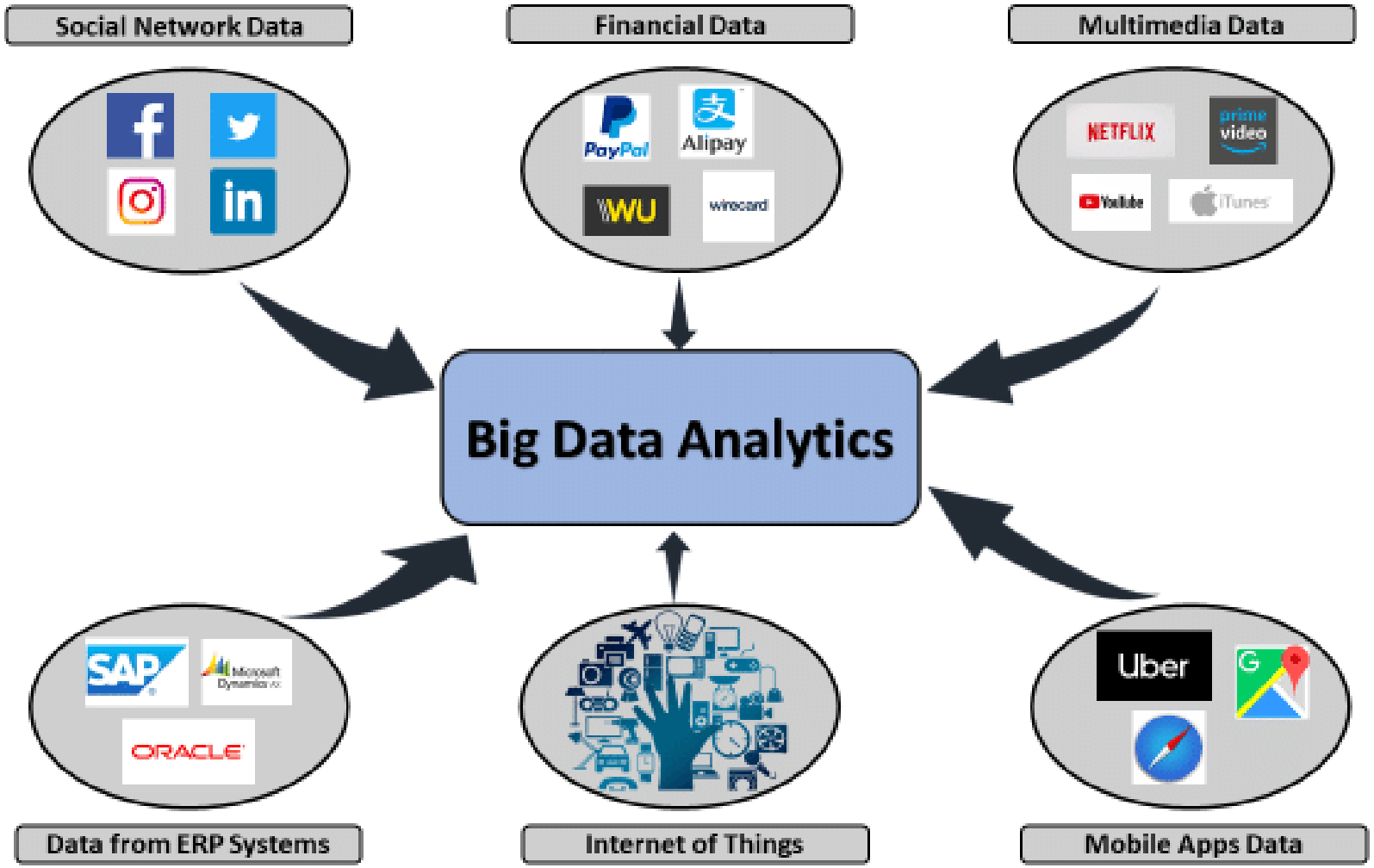


...or **crawl** data from Google

Data Collection

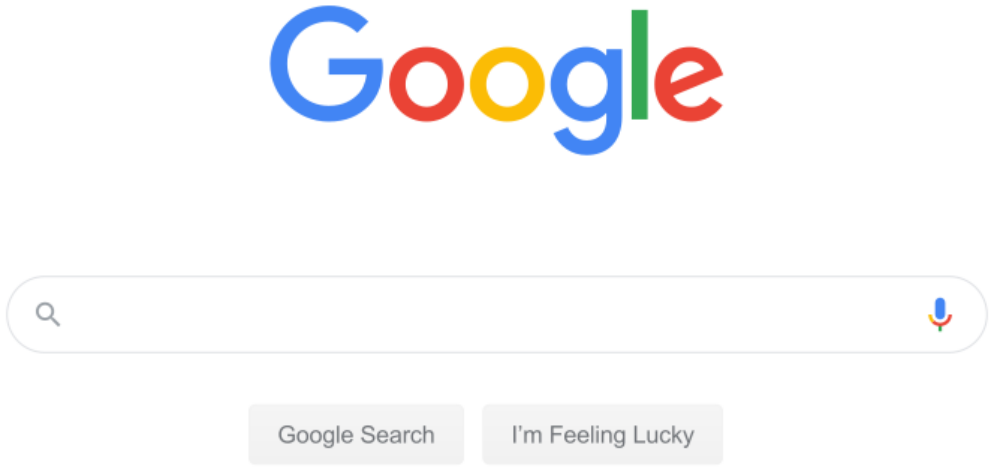


Data Collection

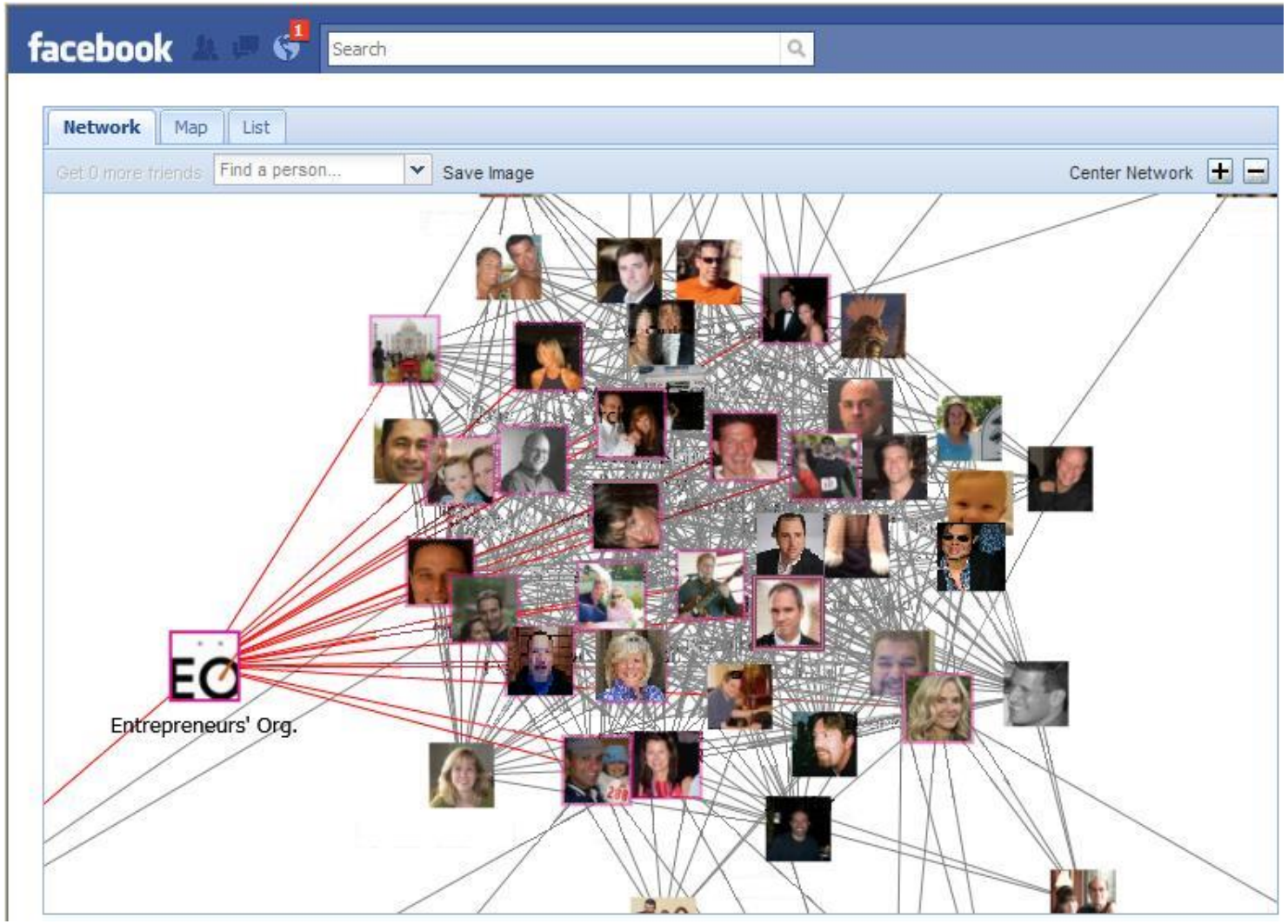


Data Collection is the Core Competence

Data Collection is the Core Competence



Data Collection is the Core Competence



Data Collection is the Core Competence

amazon.in Hello Select your address All Hello, Sign in Account & Lists Returns & Orders 0 Cart

All Best Sellers Mobiles Customer Service Today's Deals Fashion Electronics Prime New Releases Home & Kitchen Amazon Pay Computers Shopping made easy | Download the app

Your Amazon.in Your Browsing History Recommended For You Improve Your Recommendations Your Profile Learn more

Top picks for you



boAt Airdopes 141 True Wireless Earbuds with 42H Playtime,...
4.0 58,038
₹1,499.00
Get it by **Saturday, April 23**



Redmi 9A Sport (Coral Green, 2GB RAM, 32GB Storage) | 2GHz...
4.0 159,691
₹6,999.00
Get it by **Saturday, April 23**



OnePlus Nord CE 2 5G (Bahamas Blue, 8GB RAM, 128GB Storage)
4.5 12,639
₹24,999.00
Get it by **Saturday, April 23**



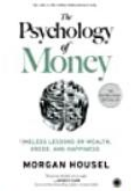
boAt Airdopes 121v2 True Wireless Earbuds with Upto 14...
4.0 116,391
₹1,299.00
Get it by **Saturday, April 23**



APLUS Super Saver Wheat Flour (Maida) 500 g
4.5 489
₹30.00



Mattel Uno Playing Card Game
4.5 32,216
₹104.00
Get it by **Saturday, April 23**



The Psychology of Money
4.5 31,996
₹321.00
Get it by **Saturday, April 23**



Tata Sampann Unpolished Toor Dal/ Arhar Dal, 1kg
4.5 12,726
₹135.00 - ₹189.00



Maggi 2-Minute Noodles Masala, 70g (Pack of 12)
4.5 47,101
₹136.00 - ₹330.00



Fortune Premium Kachi Ghani Pure Mustard Oil, 1tr PET Bottle
4.5 13,858
₹205.00 - ₹214.00



Happilo 100% Natural Premium California Almonds 1kg Value Pack Pouch | High Quality...
4.0 19,151



Glucon-D Glucon D Instant Energy Health Drink Nimbu Pani - 1kg...
4.5 4,335
₹309.00



boAt Bassheads 100 in Ear Wired Earphones with Mic(Black)
4.0 317,716
₹399.00



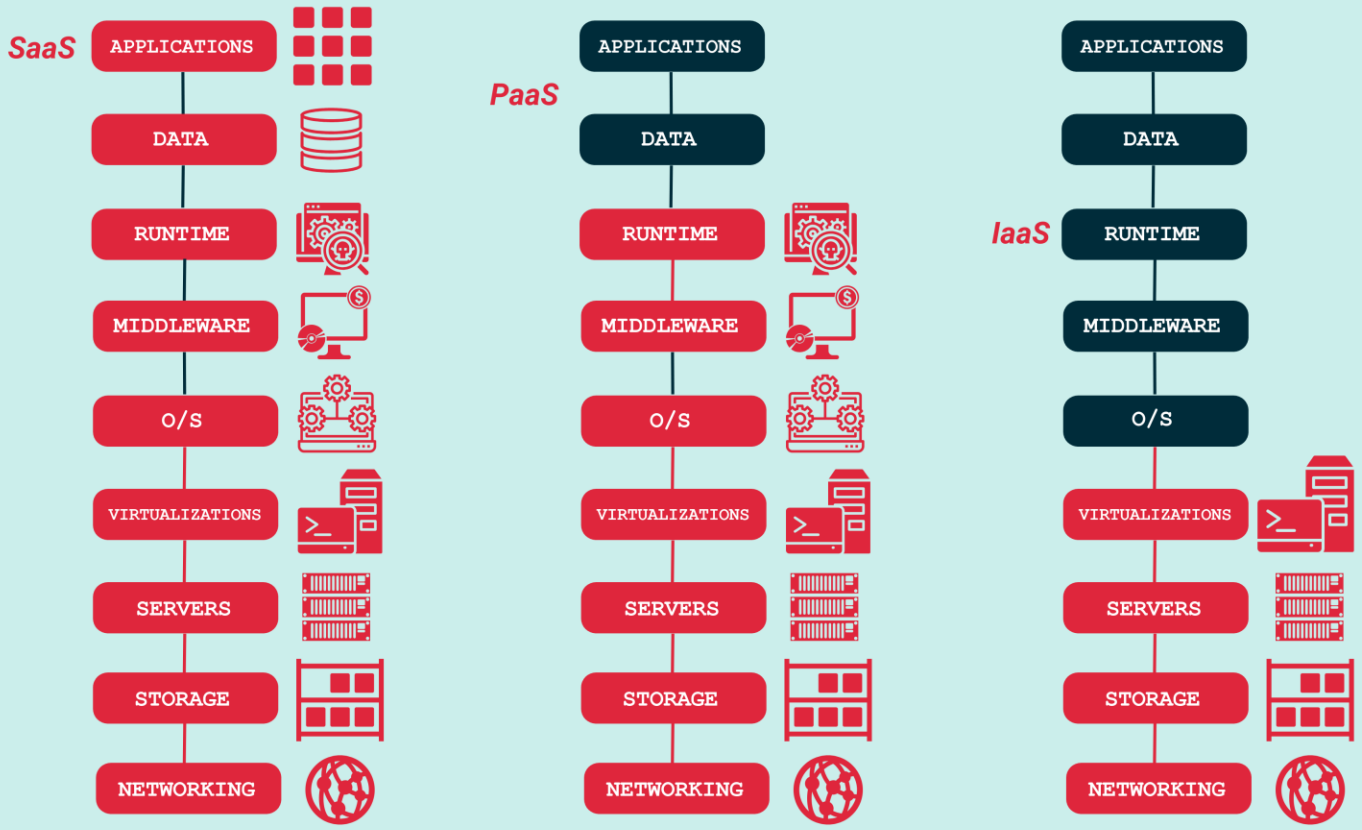
Apple 20W USB-C Power Adapter (for iPhone, iPad & AirPods)
4.5 38,383
₹1,899.00

Data Collection is the Moat

Data Collection is the Moat

AI Business Models

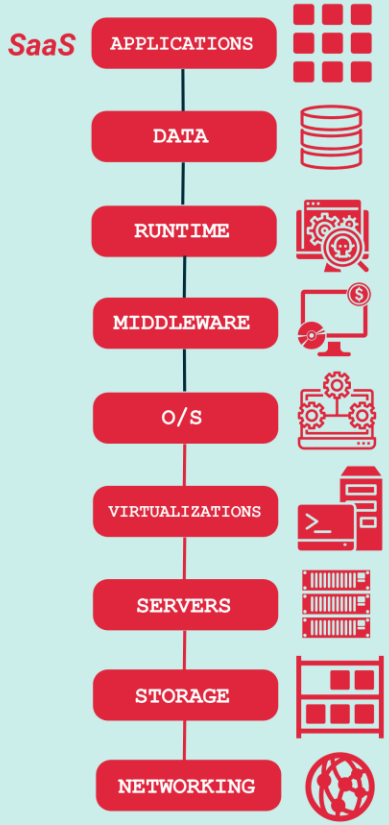
AI business models have exploded, as companies like DeepMind and OpenAI have created large-scale generative models, which are propelling an entire industry of AI software developed on top of those models.



Data Collection is the Moat

AI Business Models

AI business models have exploded, as companies like DeepMind and OpenAI have created large-scale generative models, which are propelling an entire industry of AI software developed on top of those models.

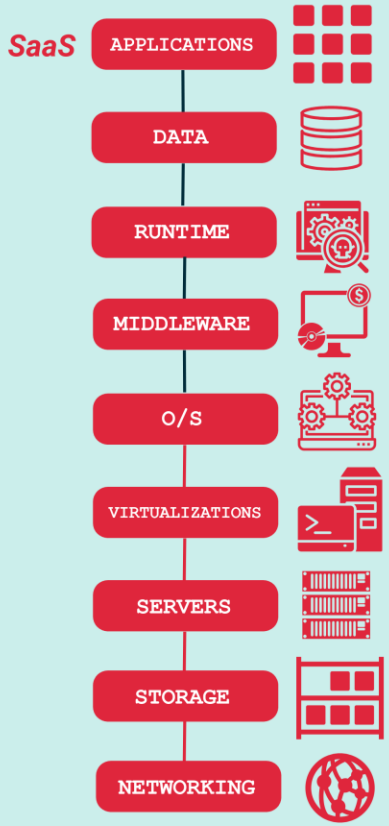


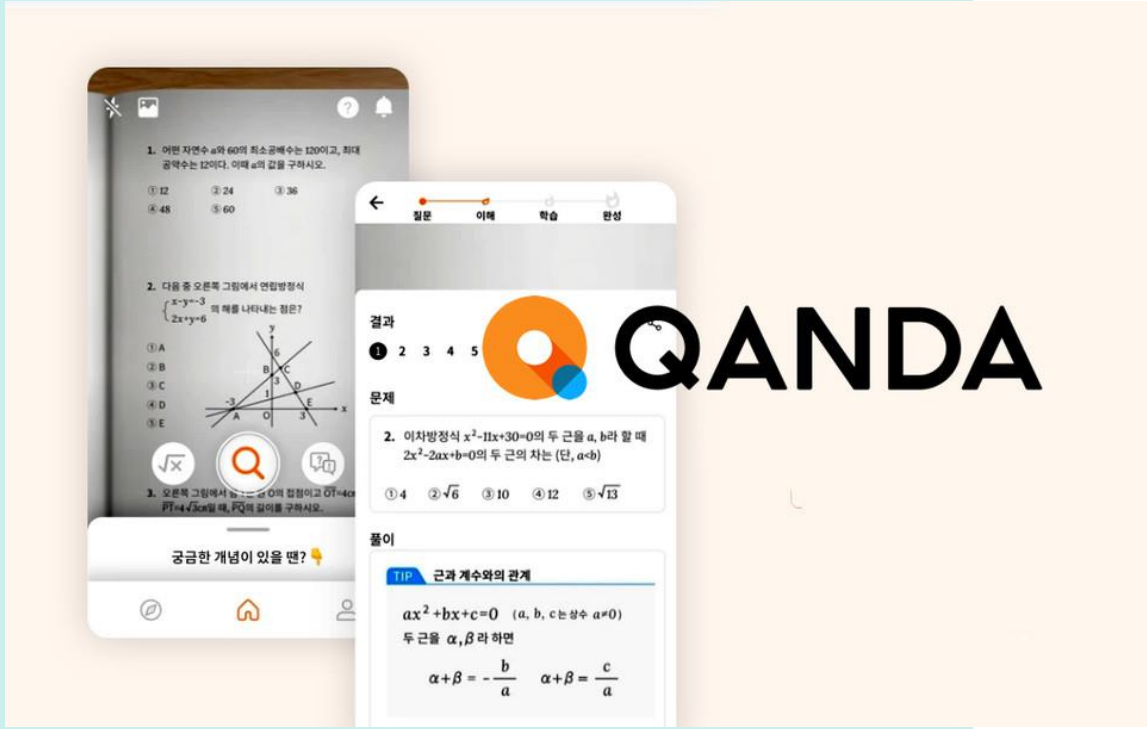


Data Collection is the Moat

AI Business Models

AI business models have exploded, as companies like DeepMind and OpenAI have created large-scale generative models, which are propelling an entire industry of AI software developed on top of those models.

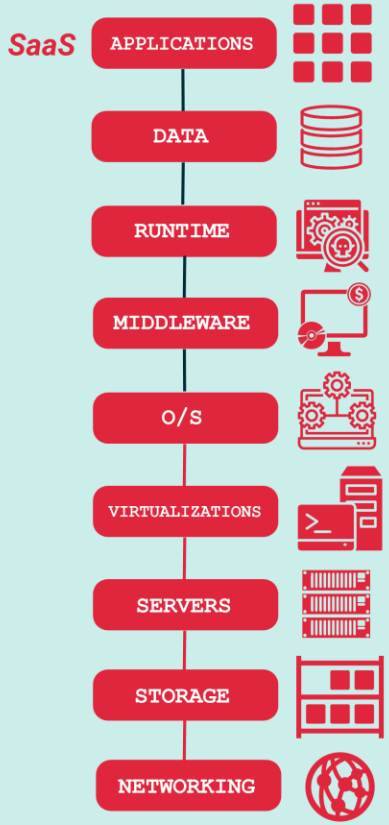


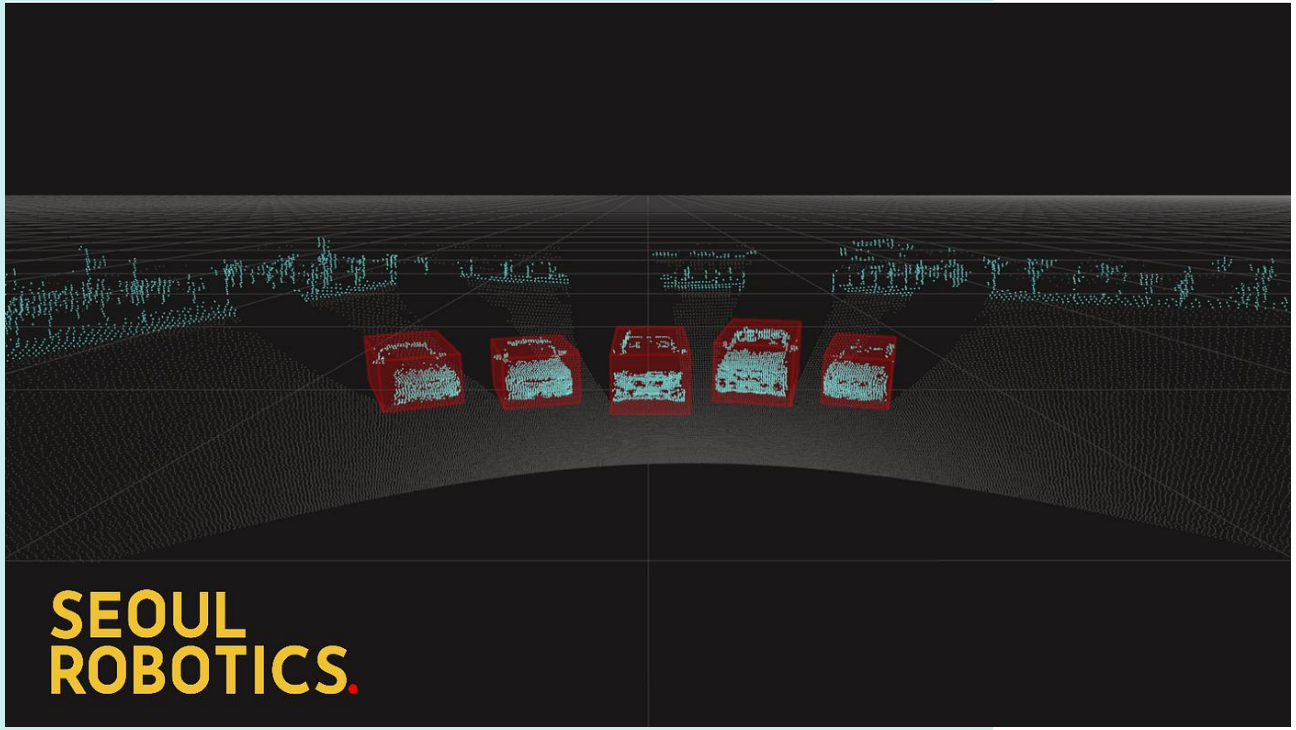


Data Collection is the Moat

AI Business Models

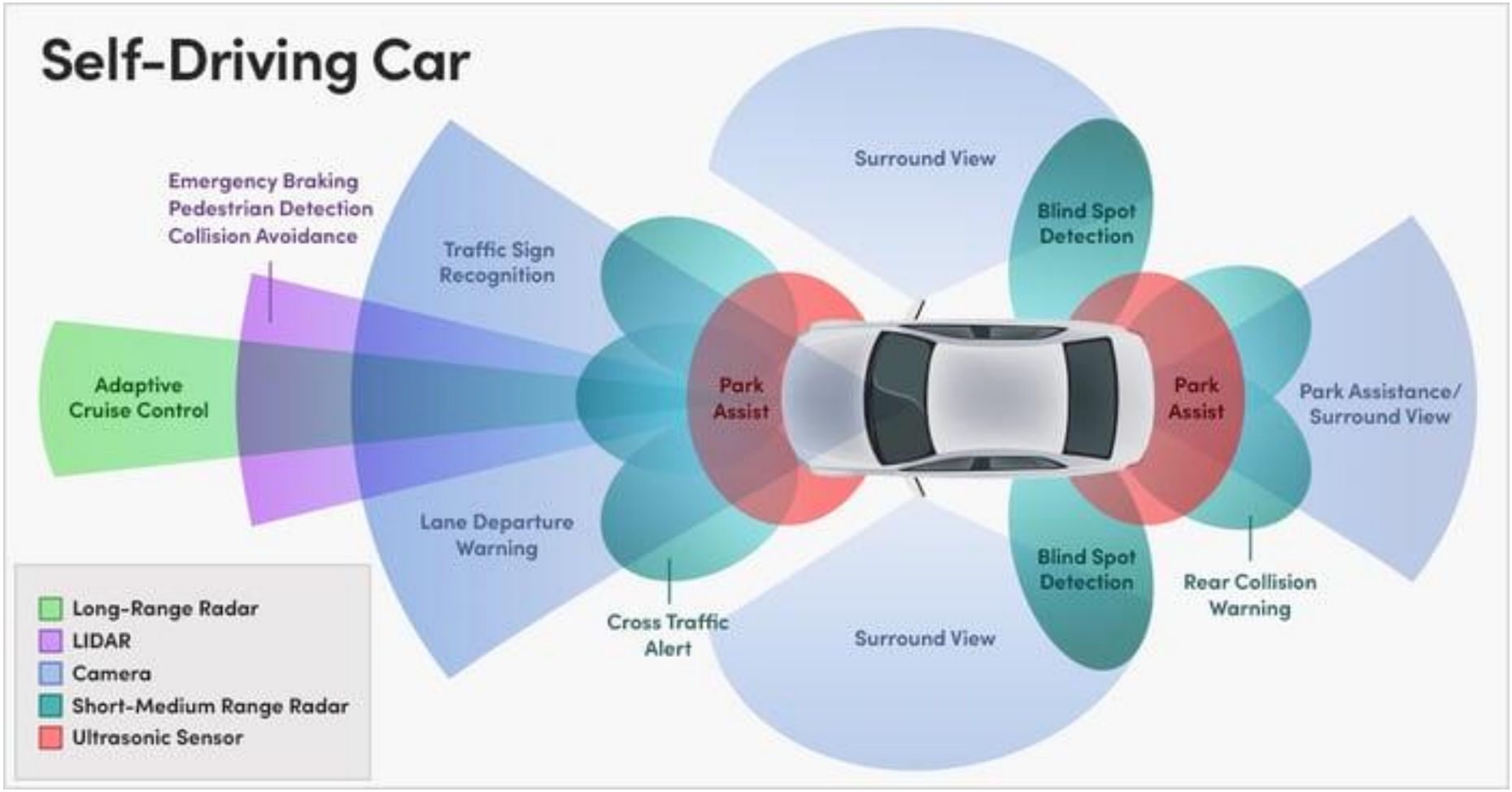
AI business models have exploded, as companies like DeepMind and OpenAI have created large-scale generative models, which are propelling an entire industry of AI software developed on top of those models.



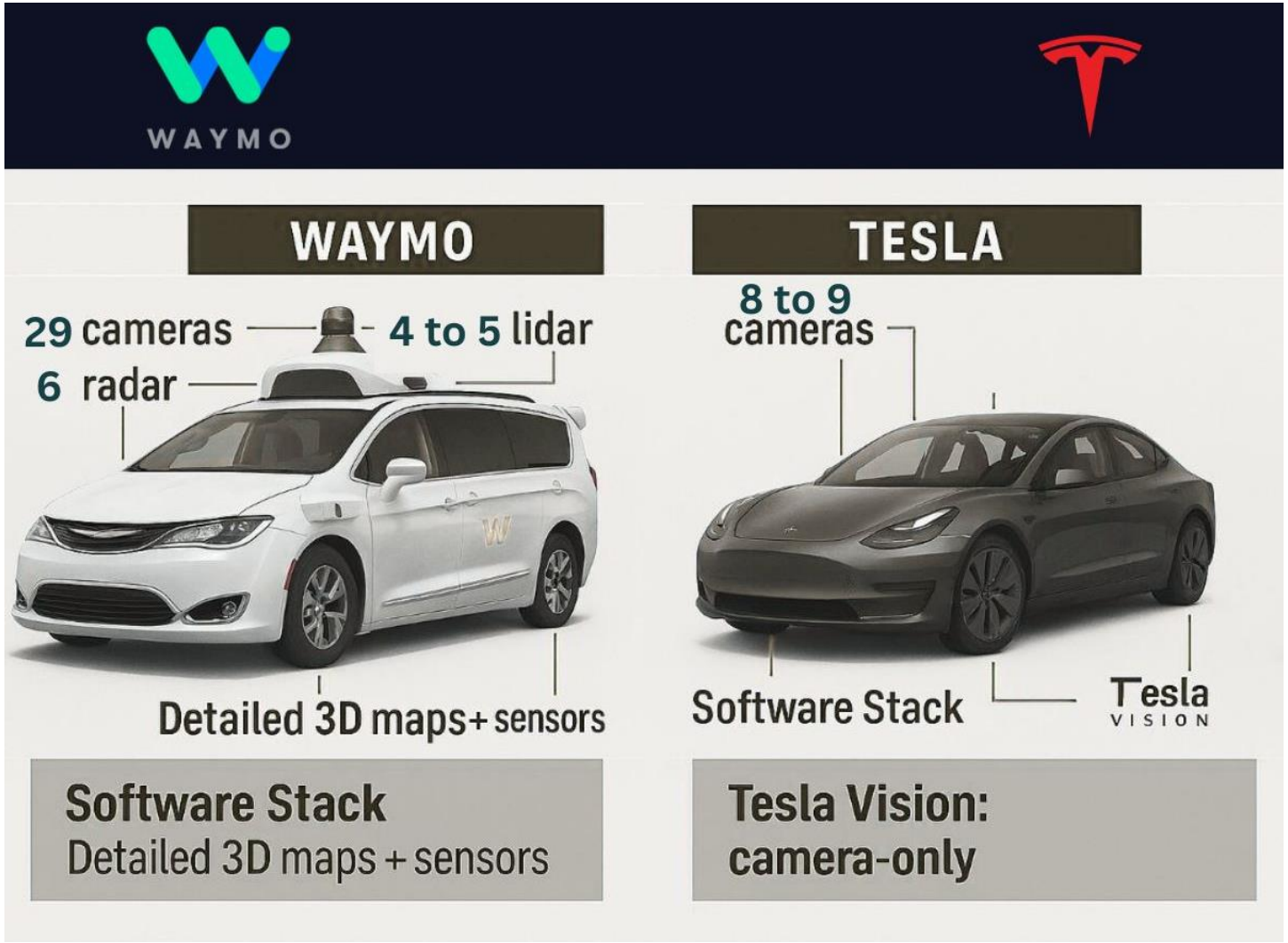


Data Collection is the Strategy

Data Collection is the Strategy



Data Collection is the Strategy

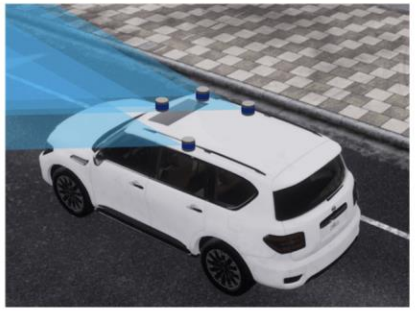
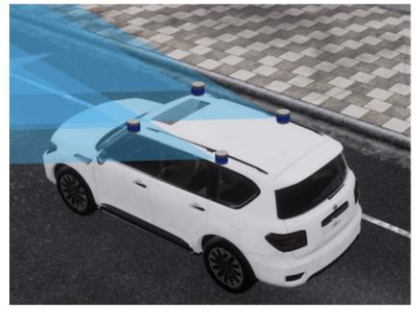
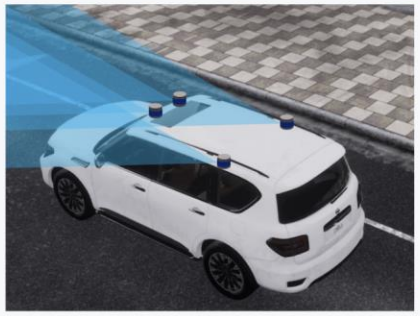
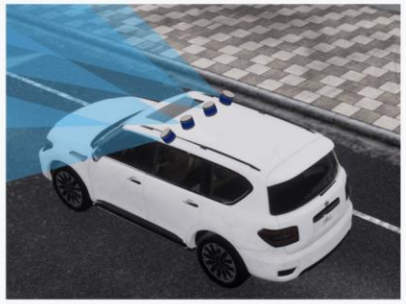
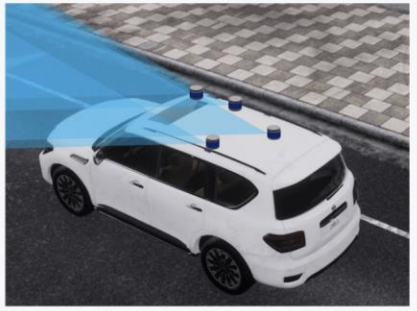


Data Collection is the Strategy

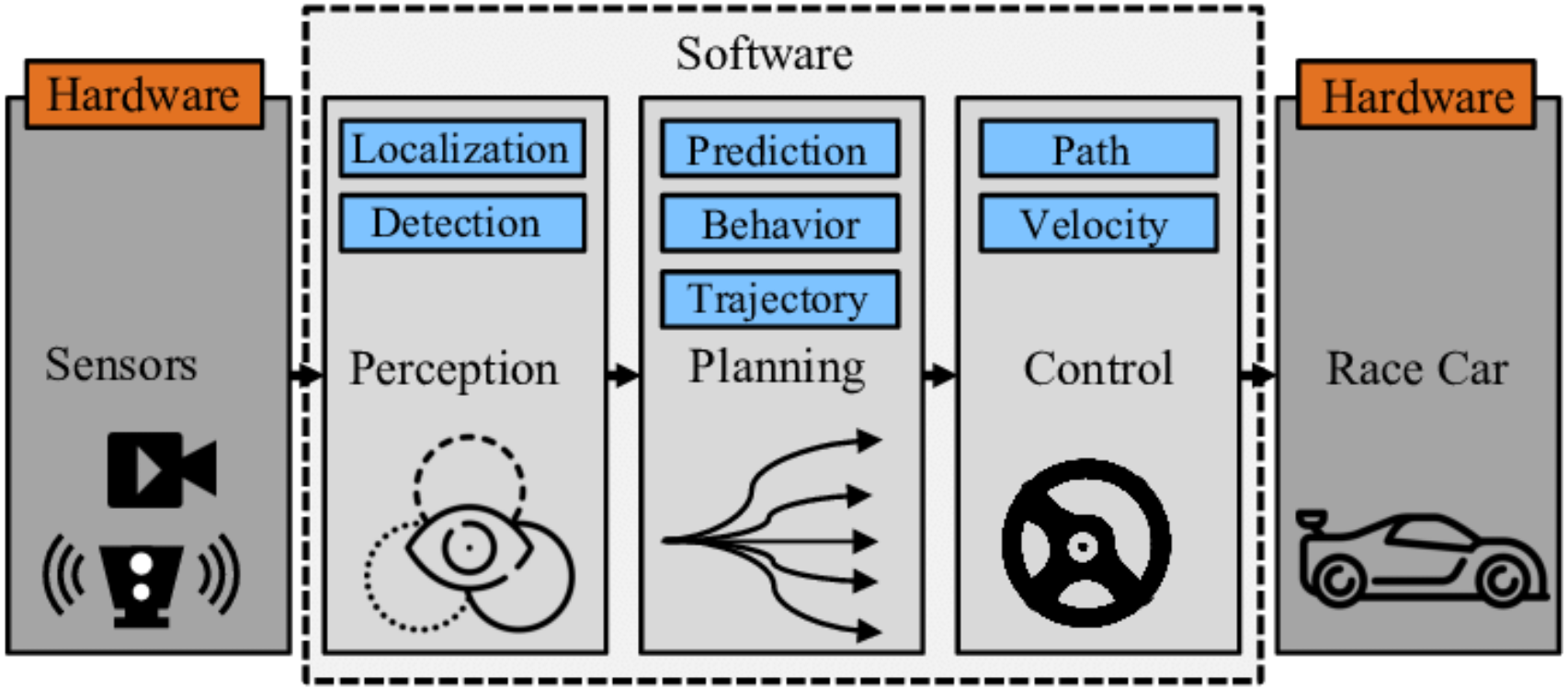
			
Waymo	Motional	Cruise	Pony.ai
			
Zoox	Toyota	Momenta	Ford

Data Collection is the Strategy

🚙 Sensor Placement

			
Center	Line	Pyramid	Square
			
Trapezoid	Line-Roll	Ours (Det)	Ours (Seg)

Data Collection is the Strategy



Data Collection Strategy is

the Art of Data Business!

Data Collection Strategy

1. **What data** to collect?
2. **Where** to collect from?
3. How to maximize **efficiency & quality**?
4. Can it create a virtuous **feedback loop**?

Data Collection Strategy is

the Art of Data Business!

Data Collection Strategy is

the Art of Data Business!



Largely determined by

(implicit) domain knowledge

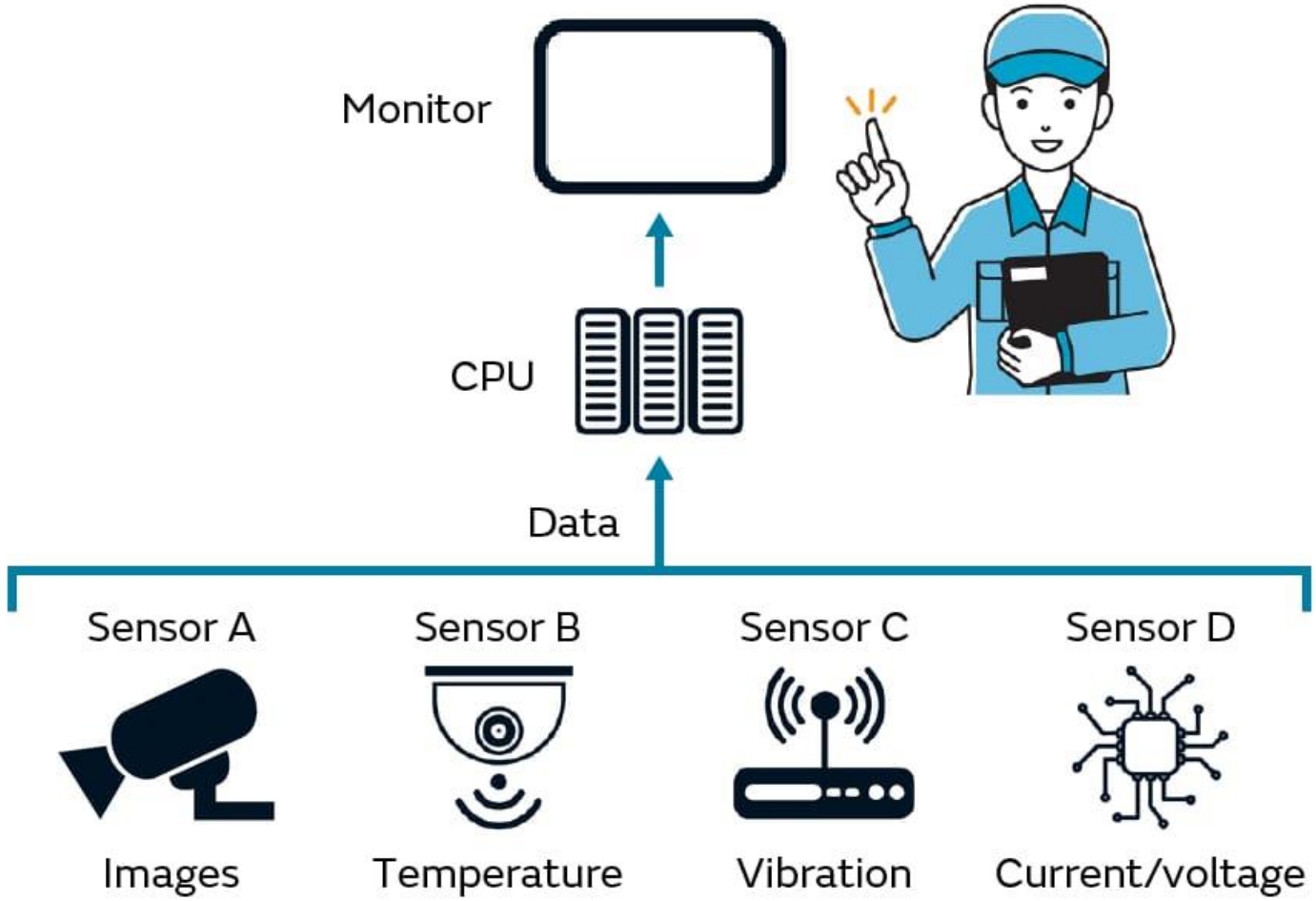
Data Collection in... Ergonomics



Data Collection in... Logistics



Data Collection in… Manufacturing



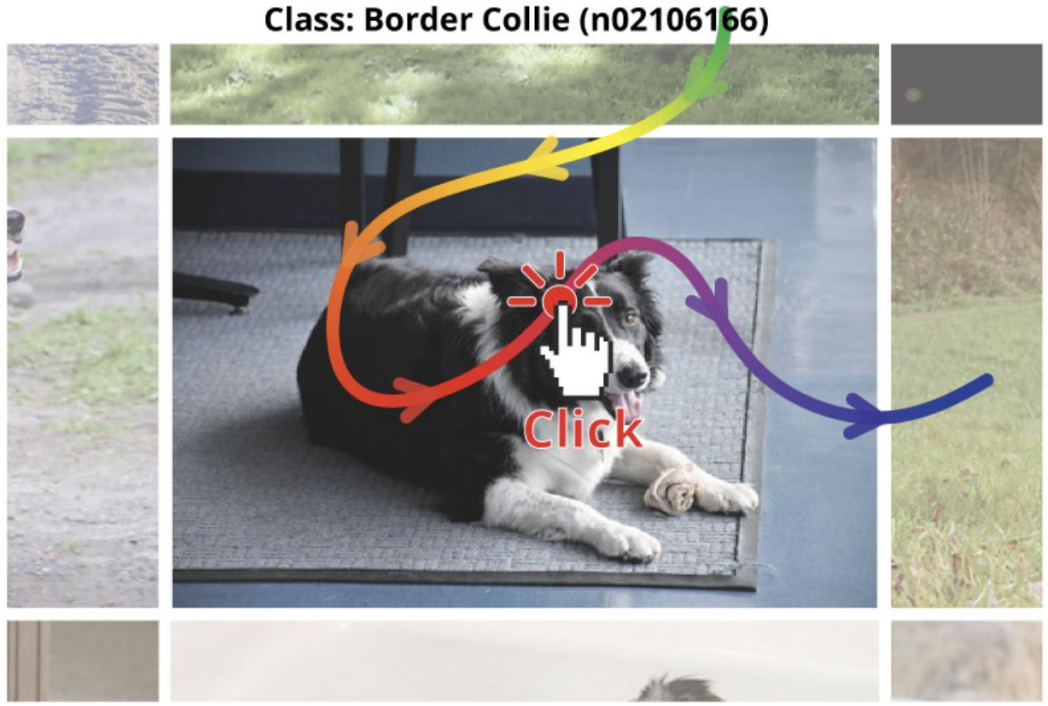
Data: Online vs. Offline



Asset & Log

Sensor

Everything Online can be Logged



Selected?
True

Time	16769	16832	16855	16881
Mouse X	.575	.498	.502	.770
Mouse Y	.991	.716	.684	.619
Click?	False	False	True	False

Neglected Free Lunch – Learning Image Classifiers Using Annotation Byproducts. ICCV, 2023.

Everything Online can be Logged

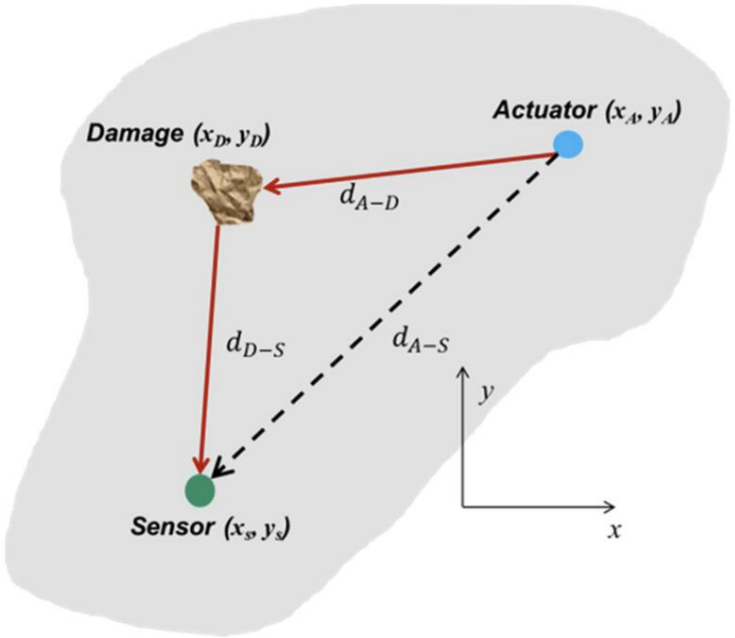


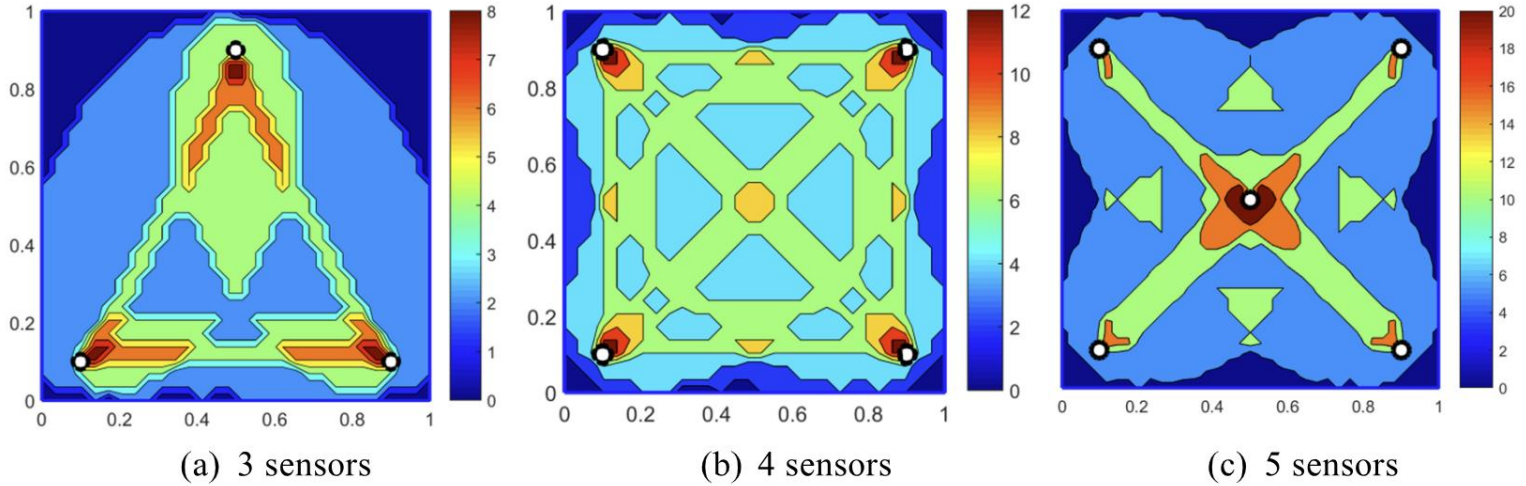
https://www.realeye.io/features/online-webcam-eyetracking

Offline Sensors are Trickier

Optimal Sensor Placement

Mathematical Optimization



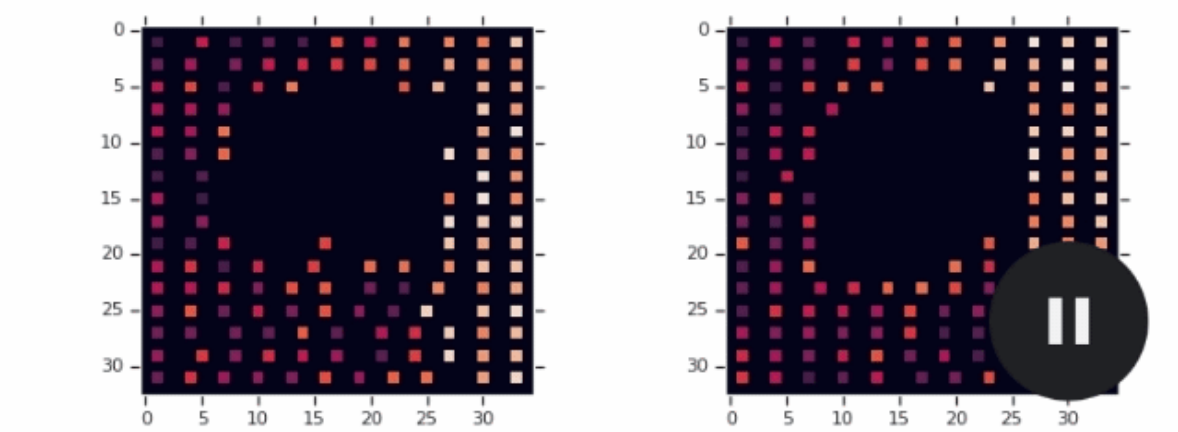


(a) 3 sensors (b) 4 sensors (c) 5 sensors

Optimal sensor placement for maximum area coverage (MAC) for damage localization in composite structures. Smart Materials and Structures, 2016.

Optimal Sensor Placement

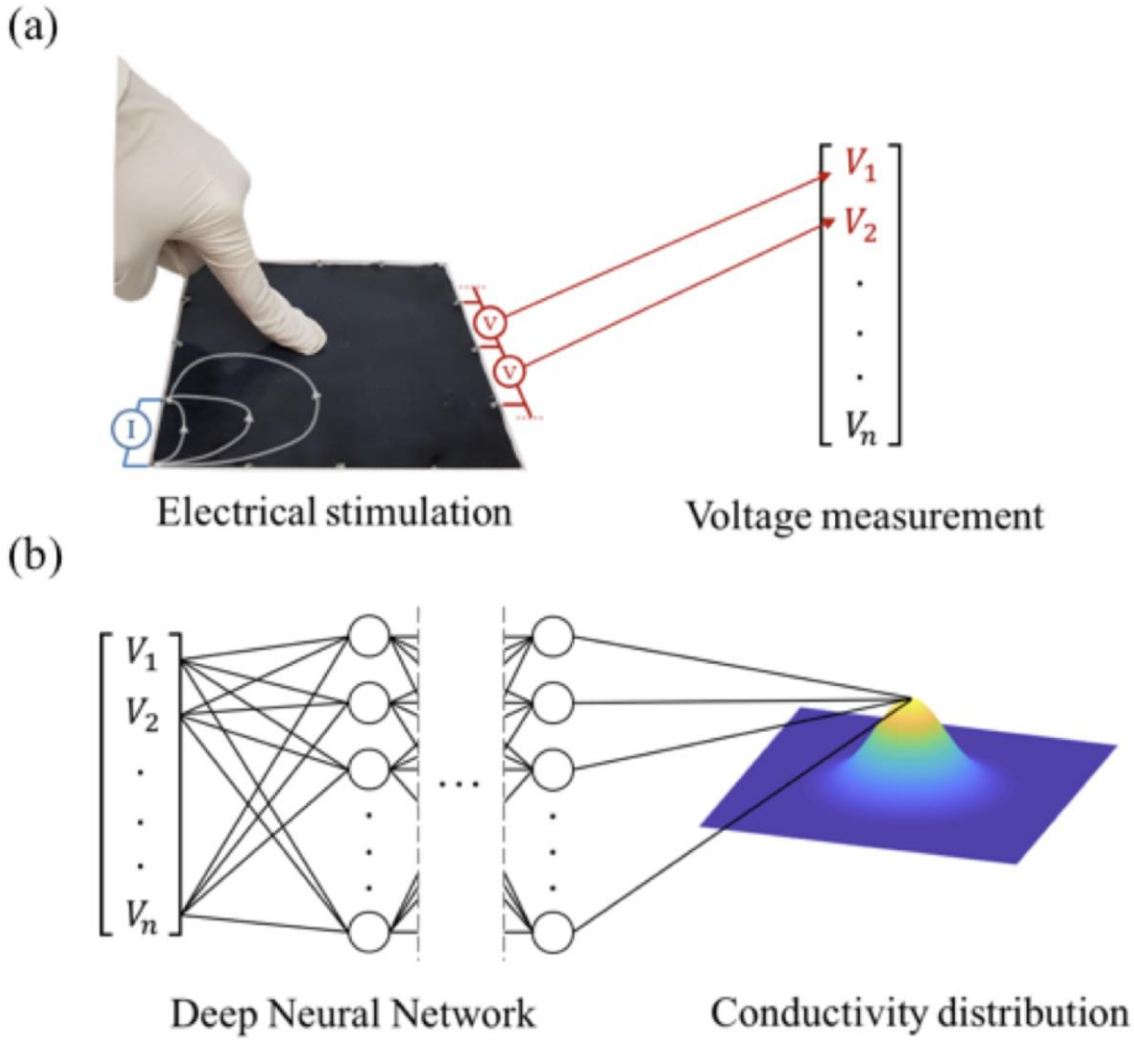
Reinforcement Learning



Left: Animation showing AlphaChip placing the open-source, Ariane RISC-V CPU, with no prior experience. Right: Animation showing AlphaChip placing the same block after having practiced on 20 TPU-related designs.

A graph placement methodology for fast chip design. Nature, 2021.

Measurement → Physical Quantity



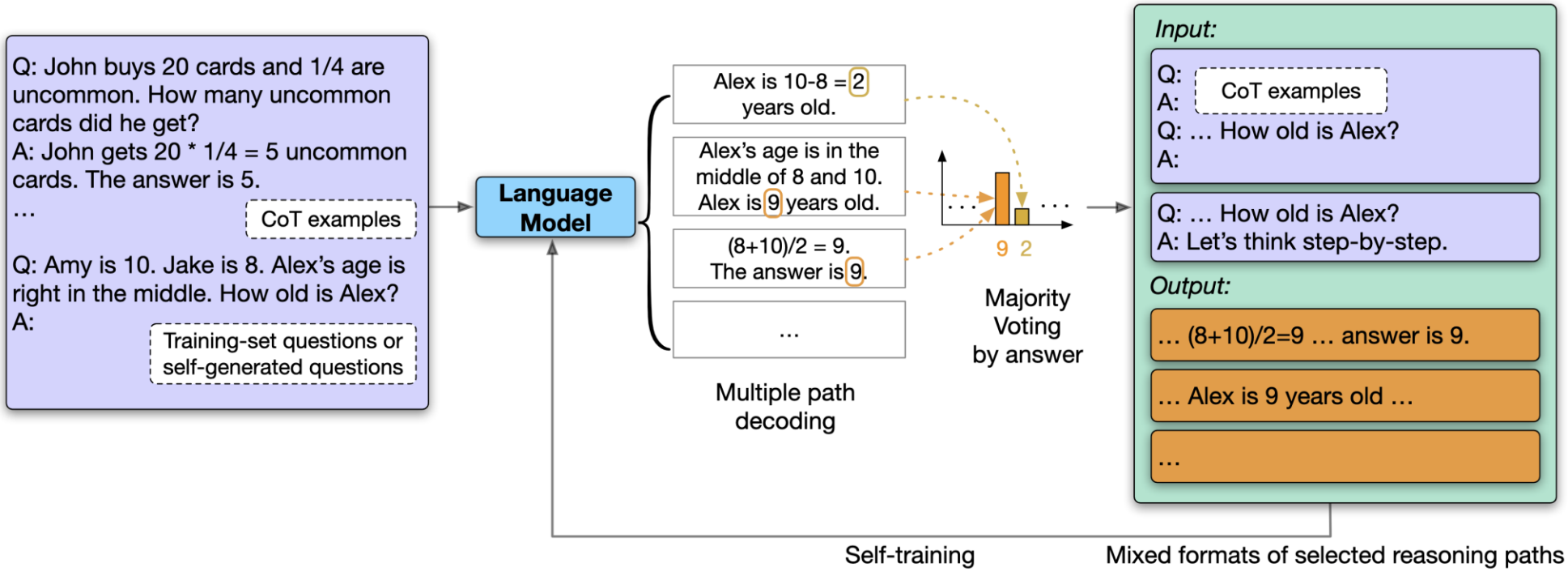
Deep Neural Network Approach in Electrical Impedance Tomography-based Real-time Soft Tactile Sensor. IROS, 2019.

Recent Trend:

AI-Generated Data

GenAI as a Data Augmentation Tool

LLMs are bootstrapped by LLM-generated text



Large Language Models Can Self-Improve. EMNLP, 2023.

GenAI as a Data Augmentation Tool

(V)LLMs refine image captions



Original: Cafe Van goch" stock image

Ours: The first image shows the exterior of a cafe with a yellow and green facade. The cafe has a large glass window with red frames, and the sign above it reads 'Cafe Van Gogh'. There are several hanging lights with a warm glow, and the tables outside are set with white tablecloths, plates, and glasses. The chairs are black with a woven seat. The floor is wooden, and there is a small potted plant...



Original: Moyie BC railway

Ours: A yellow train with red and blue accents is traveling along a railway track that runs parallel to a body of water. The train is surrounded by a landscape with autumnal colors, featuring trees with yellow and orange leaves. The railway track is bordered by a grassy area with some bushes and a dirt mound on the left side. The water is calm, reflecting the train and the trees.



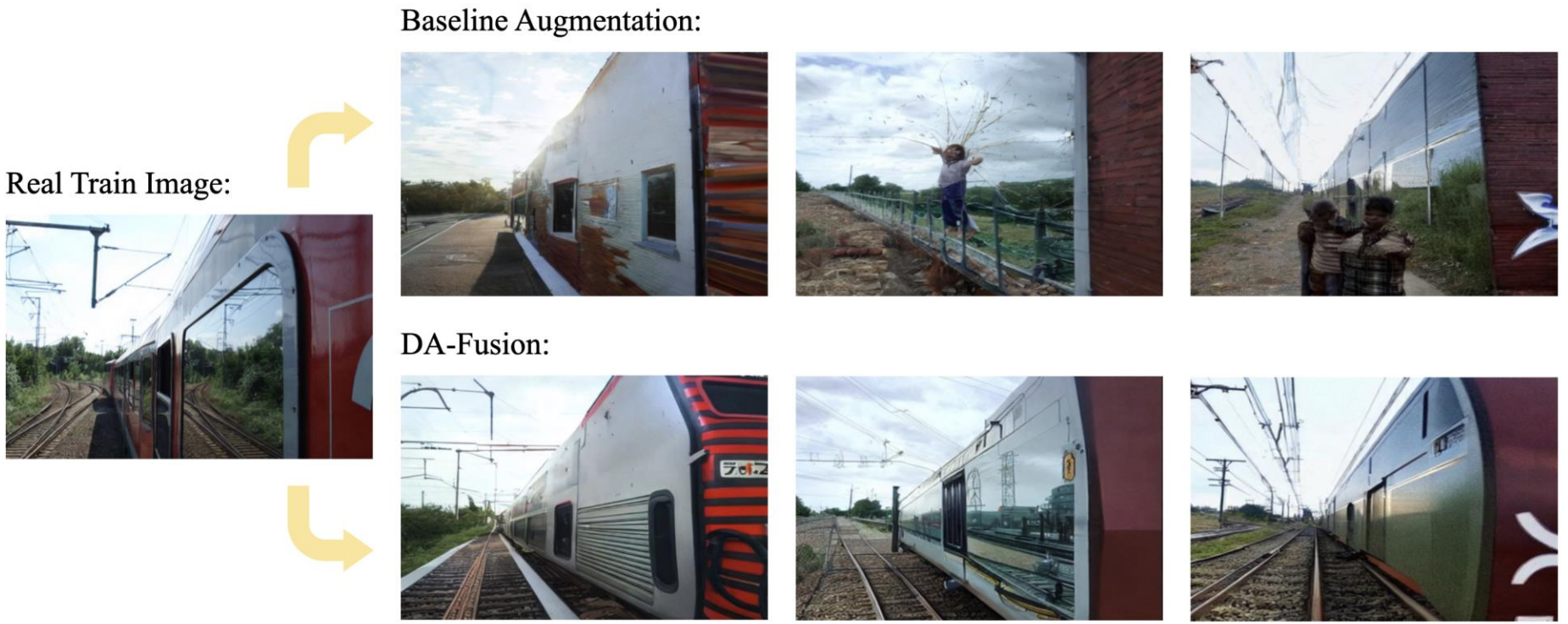
Original: Deluxe Twin Room

Ours: A modern hotel room with a large bed covered in white and red bedding, a nightstand with a lamp, and a chair with a small table in front of it. The room has a large window with purple curtains, a white ceiling with a light fixture, and a dark wall with a patterned wallpaper. There are two colorful rugs on the floor, and the room is well-lit with natural light coming through the window.

What If We Recaption Billions of Web Images with LLaMA-3? ICML, 2025.

GenAI as a Data Augmentation Tool

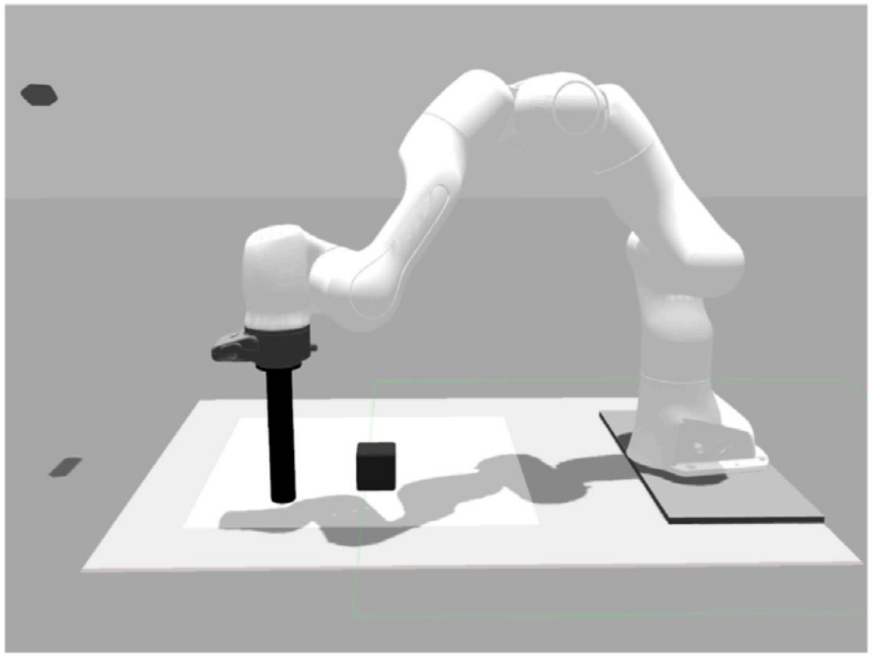
Synthesize images, especially for rare cases



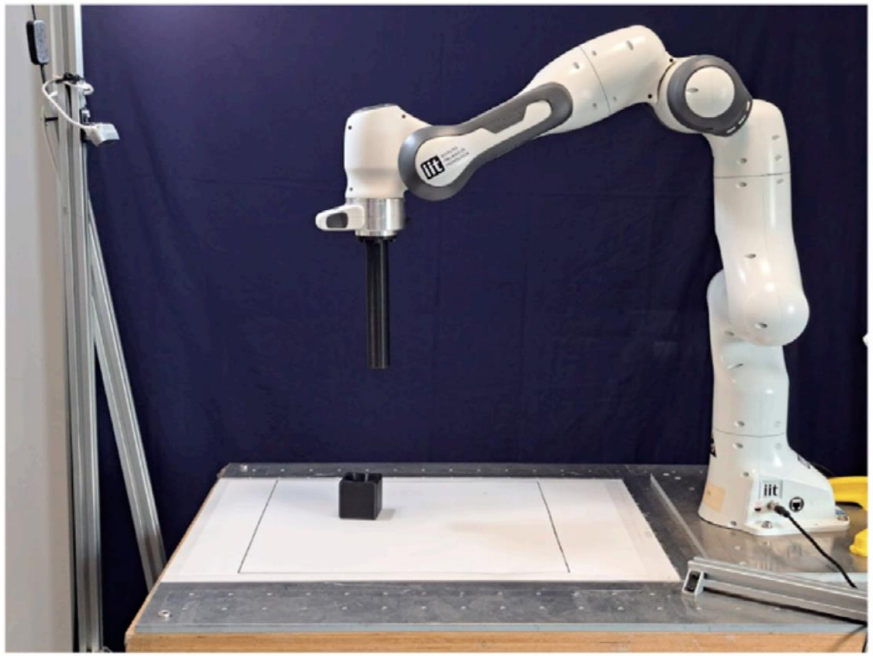
Effective Data Augmentation With Diffusion Models. ICLR, 2024.

Digital Twins and Physical AI

Digital Twins and Physical AI



Simulated scenario



Real-world scenario

Creating Digital Twins

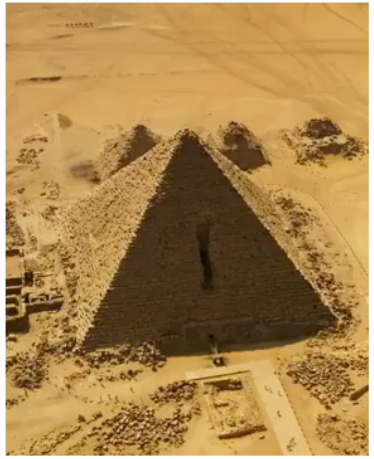
Simulator



Creating Digital Twins

Offline → Online

32 Views



VGGT: Visual Geometry Grounded Transformer. CVPR, 2025.

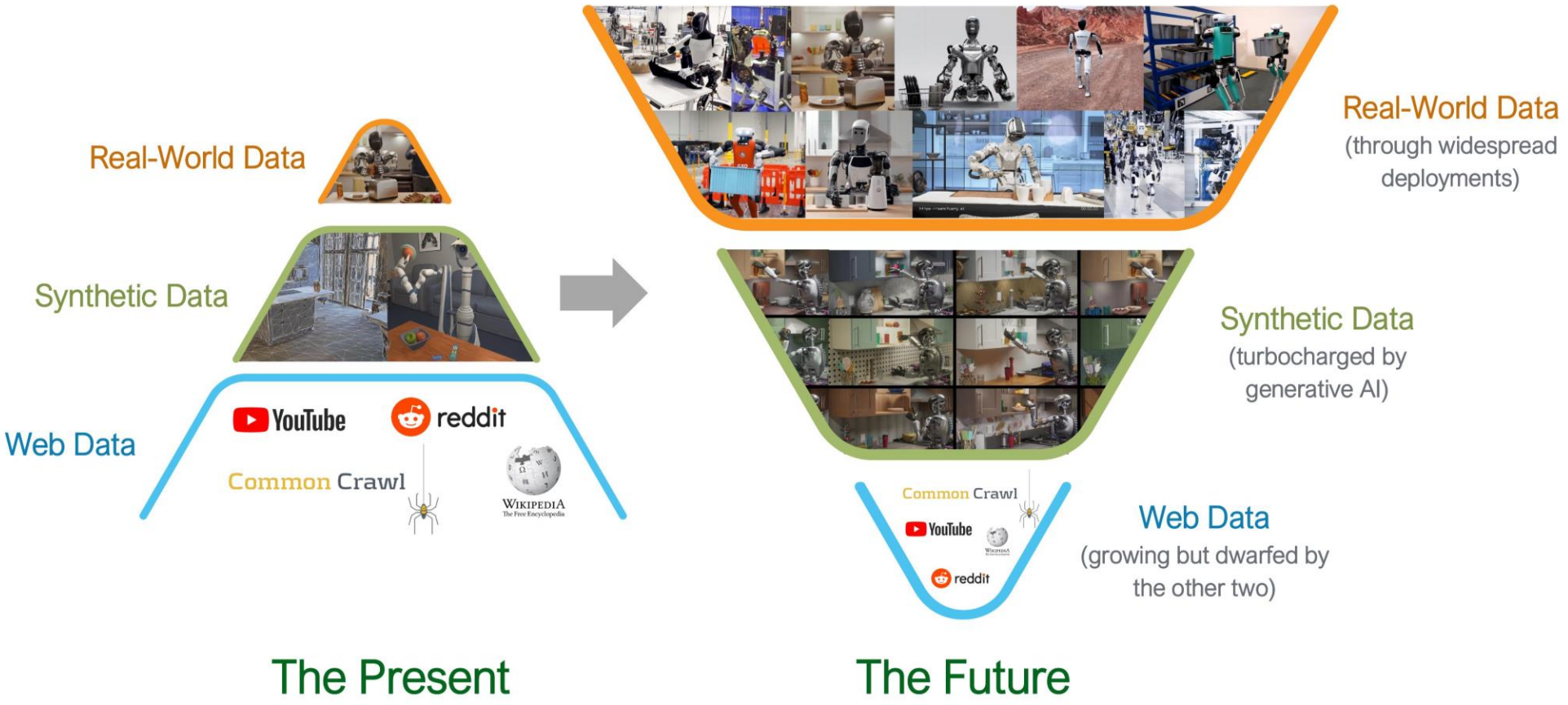
Creating Digital Twins

World Model



https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/

Future: Robots in the Real World



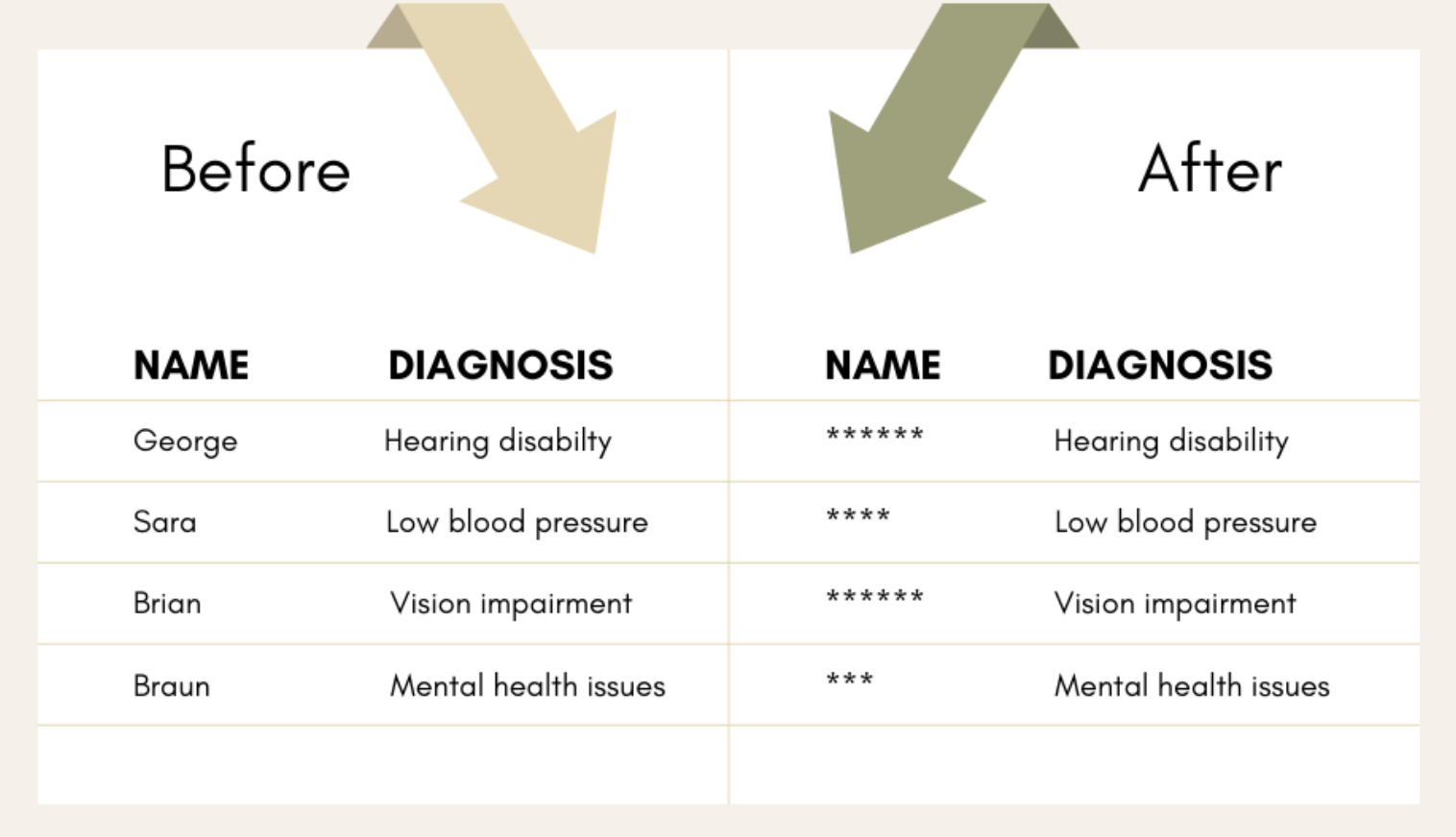
Yuke Zhu, Data Pyramid and Data Flywheel for Robotic Foundation Models

Potential Risks

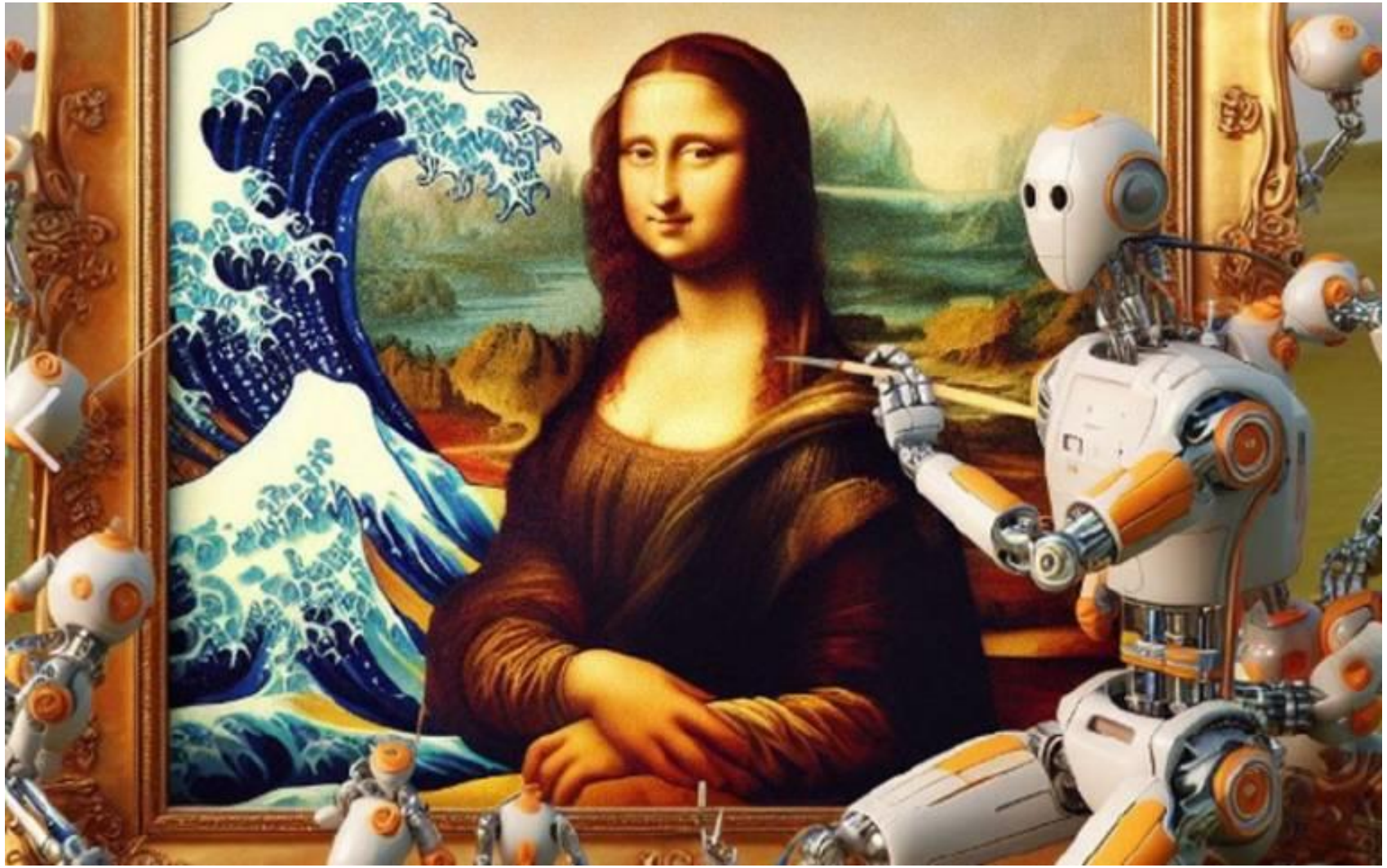
Privacy - GDPR (General Data Protection Regulation)



Privacy - Anonymization



Copyright



Confidentiality



Confidentiality

Tricky Issue: What if GenAI mimics confidential data?

Fairness

Gender Bias

Tendency to perceive one gender as better than the other.



Age Bias

Tendency to judge ability based on the age.



Disability Bias

Tendency to judge individuals based on disability status.



Security


Data Theft
Man-in-the-Middle Attacks
Unauthorized Access
12345678
No Encryption
Malware Spread
Fake Networks

Recap: Team Project

Build your team! **(Deadline: Sep 18)**

