The Next Frontier in Embodied AI: Autonomous Driving

CUED Guest Lecture – 25 April 2024

Anthony Hu - Researcher

Road death statistics

[WHO Road Safety]

1,350,000

road deaths every year

#1

killer of people aged between 5-29

1 in 2

road deaths are pedestrians, cyclists or motorcyclists

Every

24 seconds

someone is killed on a road

Every year between 20 and 50 million people are non-fatally injured

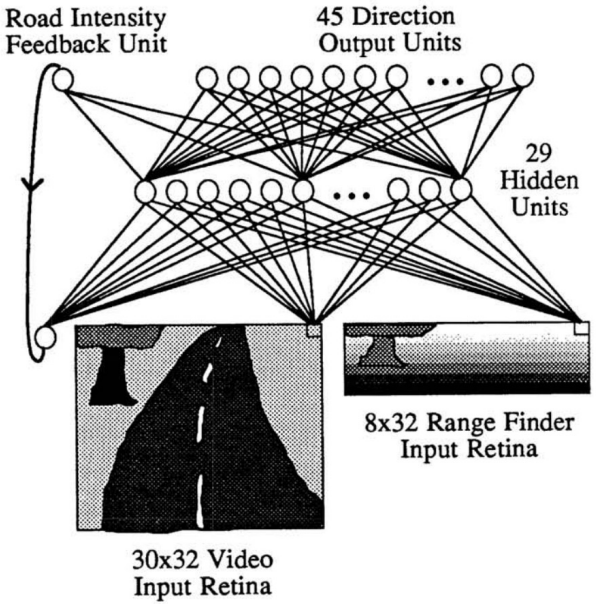
Human error accounts for the vast majority of road accidents

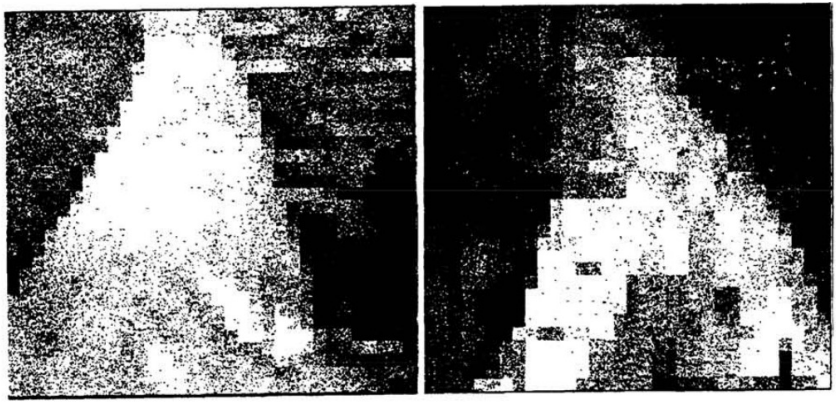
Speeding
Aggressive and reckless driving
Distracted driving
Inebriation
Drowsy driving

[NHTSA 2016]

ALVINN:
AN AUTONOMOUS LAND VEHICLE IN A NEURAL NETWORK

Dean A. Pomerleau
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213



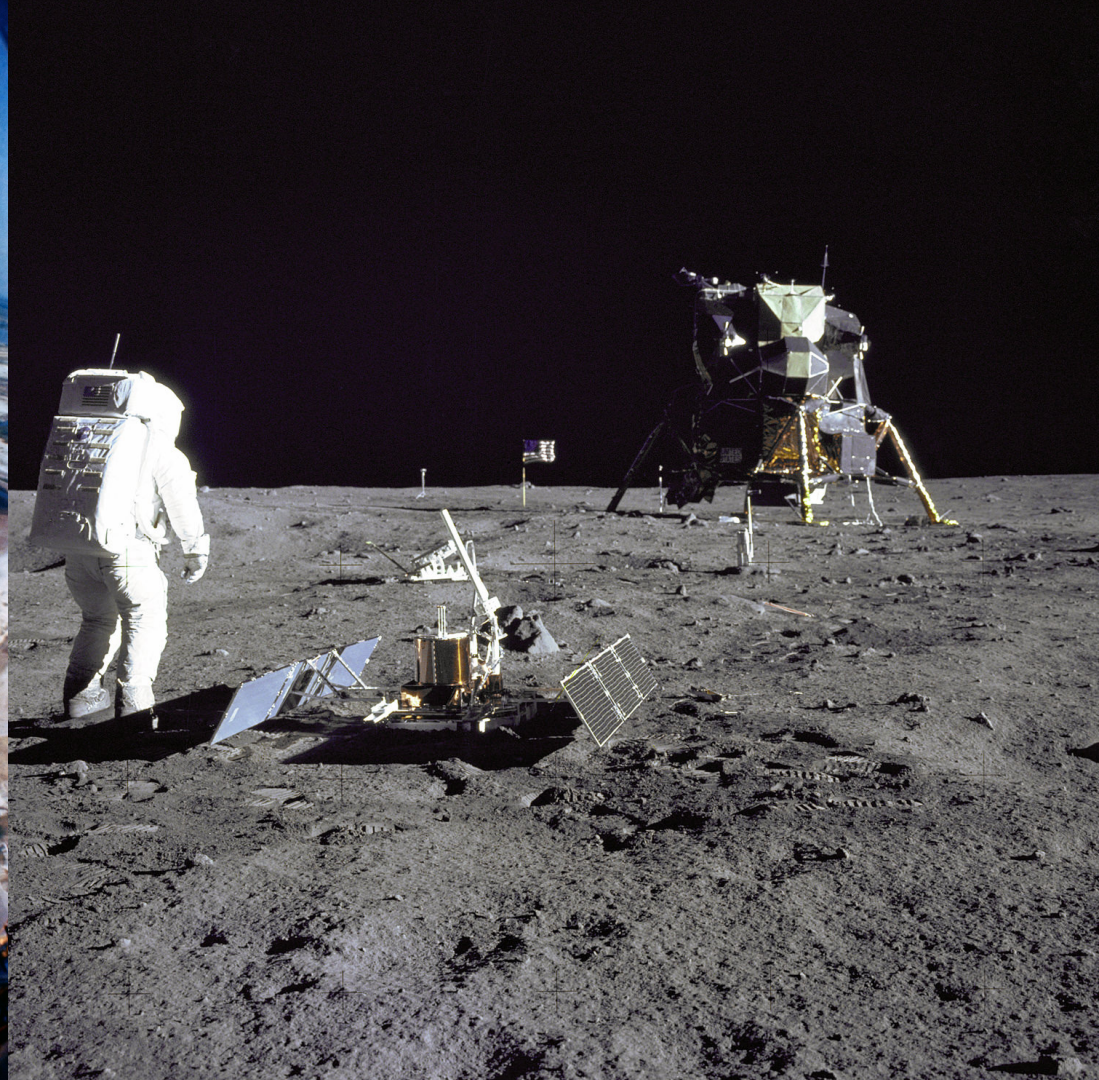


[Pomerleau 1988]

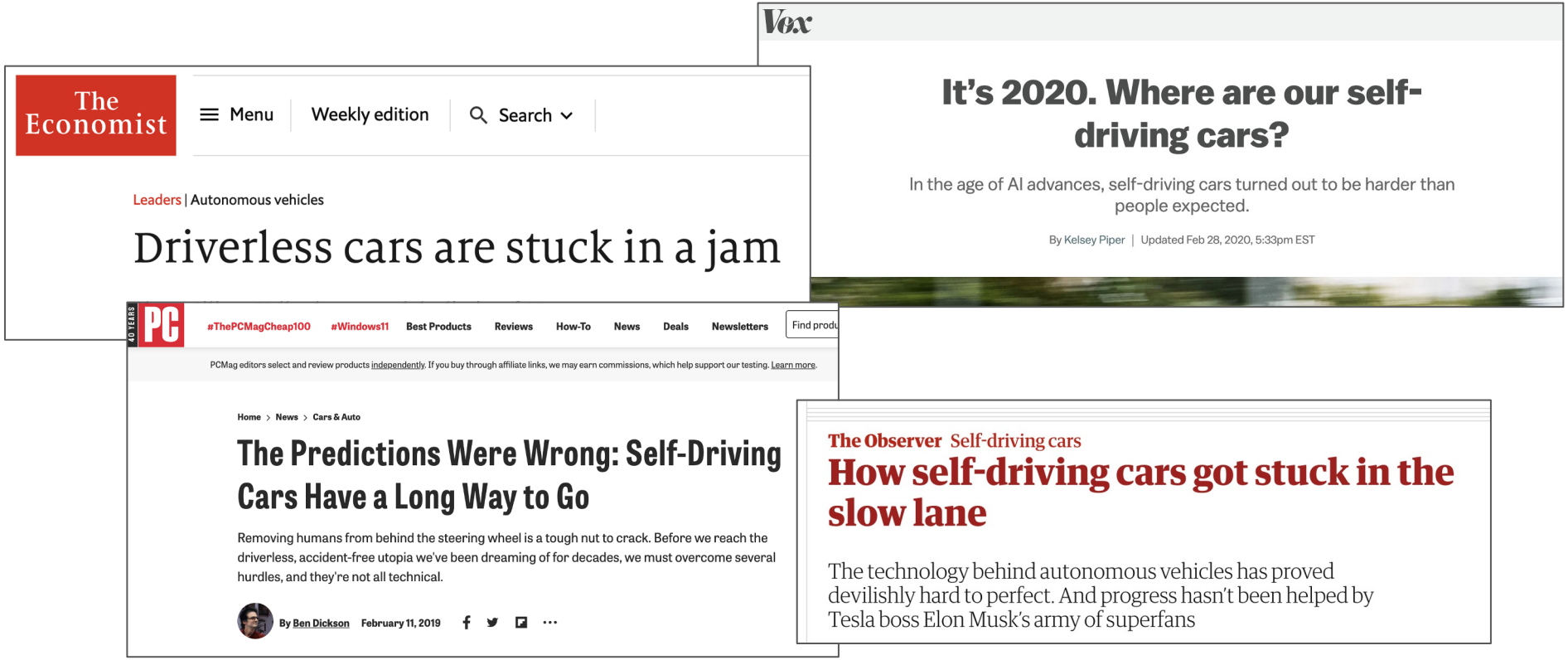
DARPA Grand Challenge 2004-2005

DARPA Urban Challenge 2007

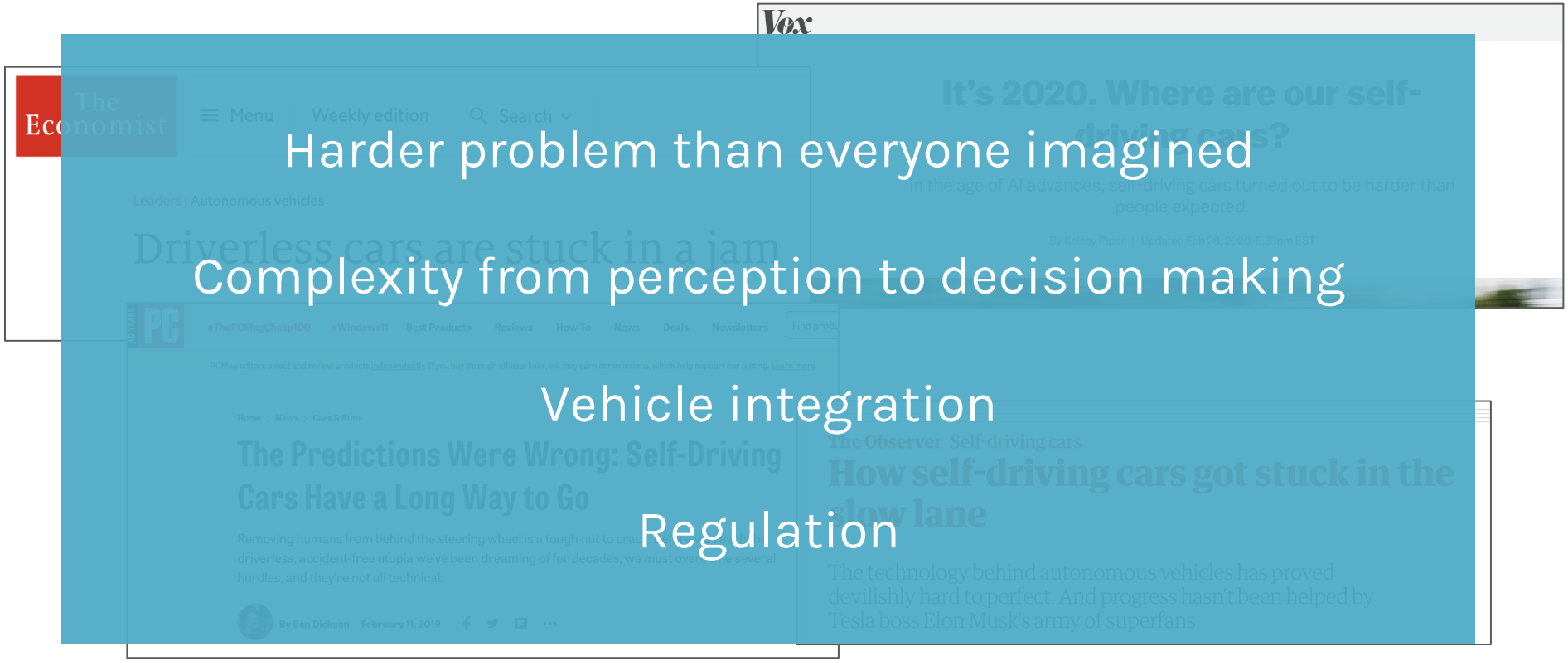

WAYMO
cruise
ZOOX
pony.ai
Waabi
autox
mobileye
WAYVE



Why aren’t self-driving cars here yet?



Why aren’t self-driving cars here yet?



WAYVE

Outline

1. Technical challenges

2. How to build a self-driving car

3. Research highlights

WAYVE

1. Technical challenges

What are the technical challenges?

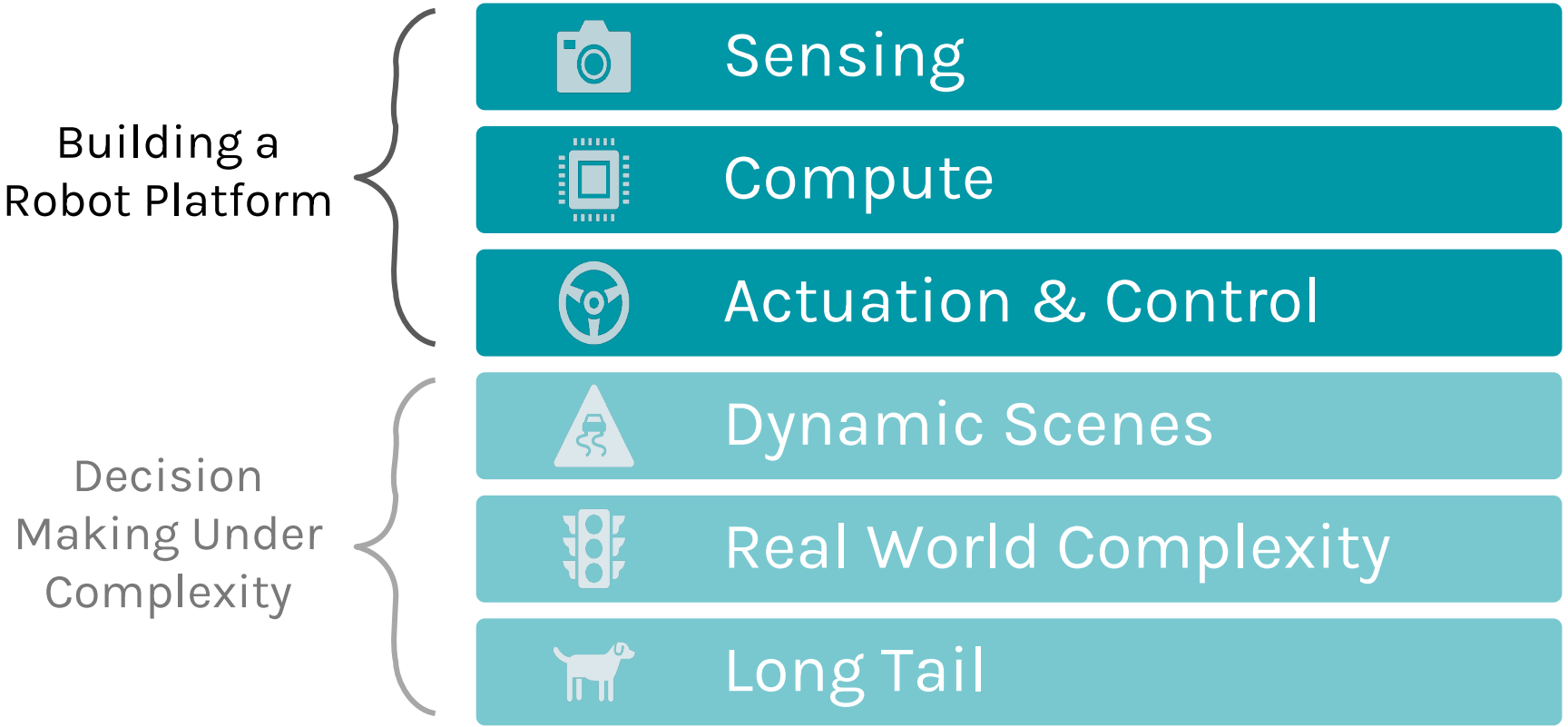
Building a Robot Platform

- Sensing
- Compute
- Actuation & Control

Decision Making Under Complexity

- Dynamic Scenes
- Real World Complexity
- Long Tail

What are the technical challenges?



Sensing

The dominant sensor modalities used in robotics are:

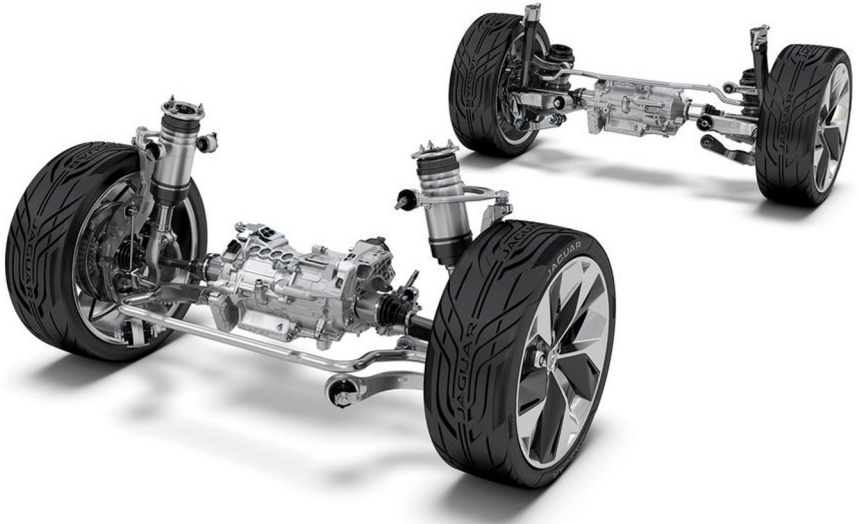
Proprioceptive (internal state)

- Actuators (i.e. motor speed, position)
- Inertial Measurement Unit (IMU)

Exteroceptive (external state)

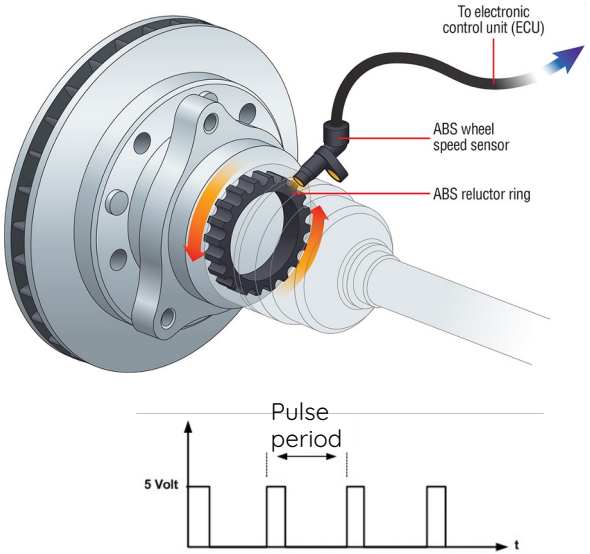
- Global Navigation Satellite System (GNSS)
- Cameras
- RADAR
- LiDAR

Sensing: Actuators

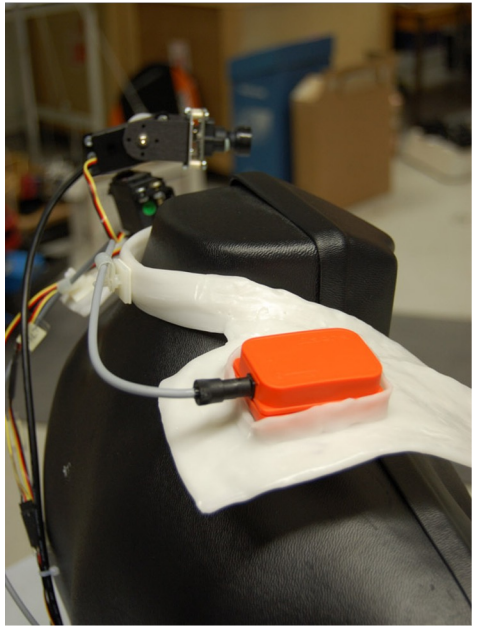


Wheel speed

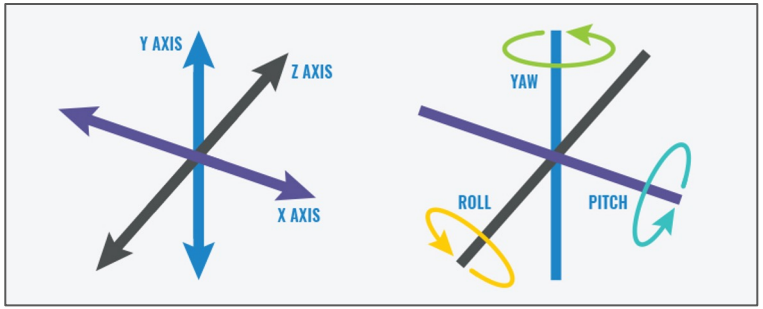
Measure the motion of teeth past a sensor



Sensing: Inertial Measurement



- Acceleration sensing (3D)
- Angular velocity sensing (3D)



IMUs are extremely useful, but suffer from drift over time

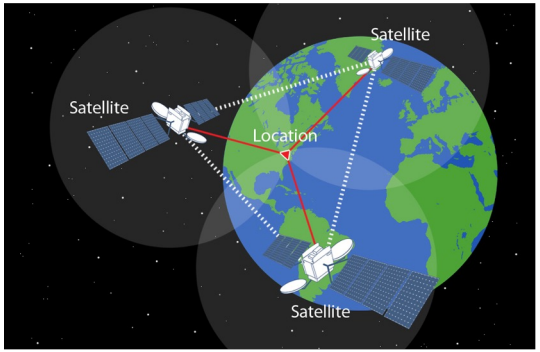
Sensing: Global Navigation Satellite System

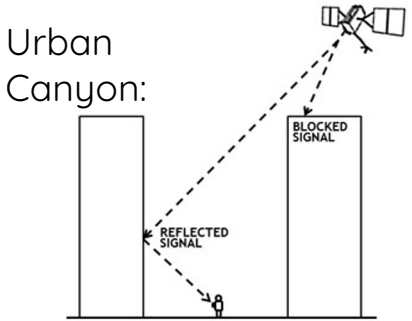
Pros:

- Global 2.5D positioning: [x, y, θ]

Cons:

- ~1-10m accuracy
- Consumer-grade limited to ~5Hz
- Urban canyons hugely degrade GNSS performance: multipath effects + blocked signal





Sensing: Cameras

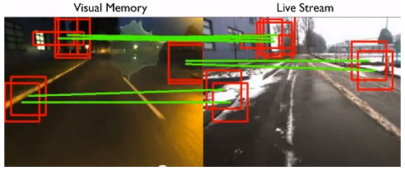
Typical camera:

~1-8MP, ~8-14 bit colour (Red, Green, Blue), 20-200Hz

Tradeoff: frame rate vs resolution

Monocular cameras

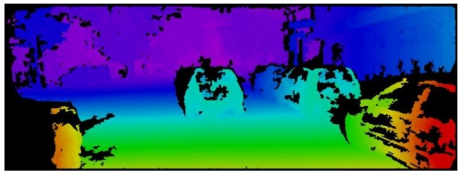
Optical flow, visual odometry, and localisation



Object detection, tracking, segmentation

Stereo Cameras

Depth sensing from a pair of images



Sensing: Radar

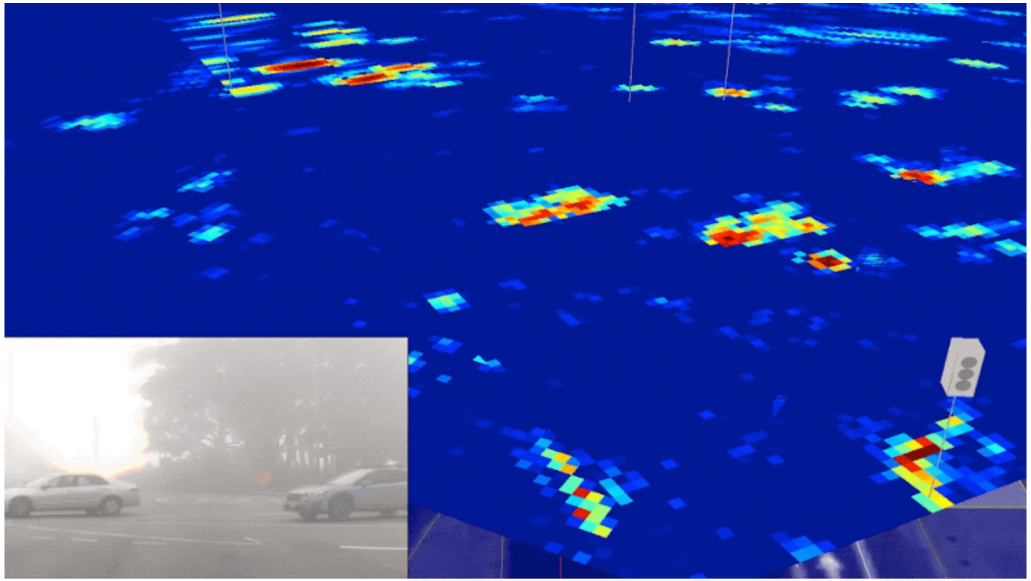
Pros:

Metric depth sensing robust to weather, lighting conditions, ~200m+ range, can see through objects

Cons:

Noisy, multipath effects, not 3D

E.g., Continental ARS441



E.g., Waymo imaging radar visualisation

Sensing: Lidar

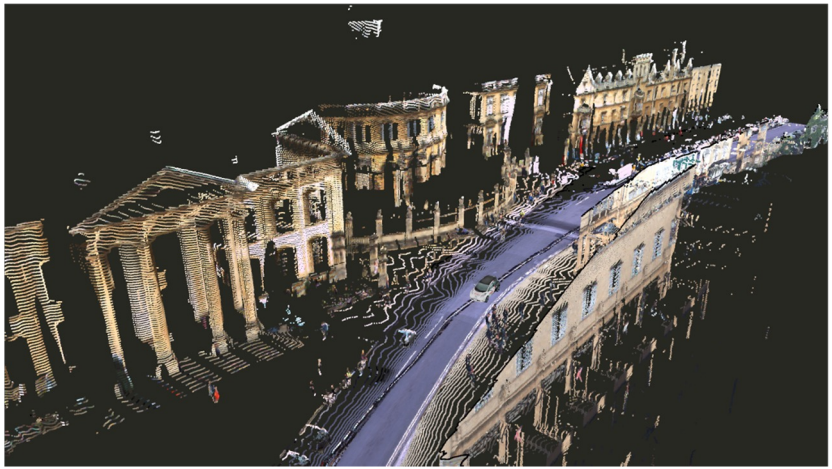
Pros:

- Depth sensing robust to lighting conditions, very accurate / low noise
- 100-300m+ range pointcloud: [x, y, z]
- 0.3-10M points/second at 5-20Hz

Cons:

- Degraded by rain, snow
- Sparse signal at distance
- Expensive (though improving)
- Poor longevity (though improving)

E.g., Velodyne 3D lidar



Compute

More compute -> better predictions

Real-time requirements (e.g. latency)

Space, power, thermal limited

Effect on EV battery life

Actuation & Control

Steering

Brake

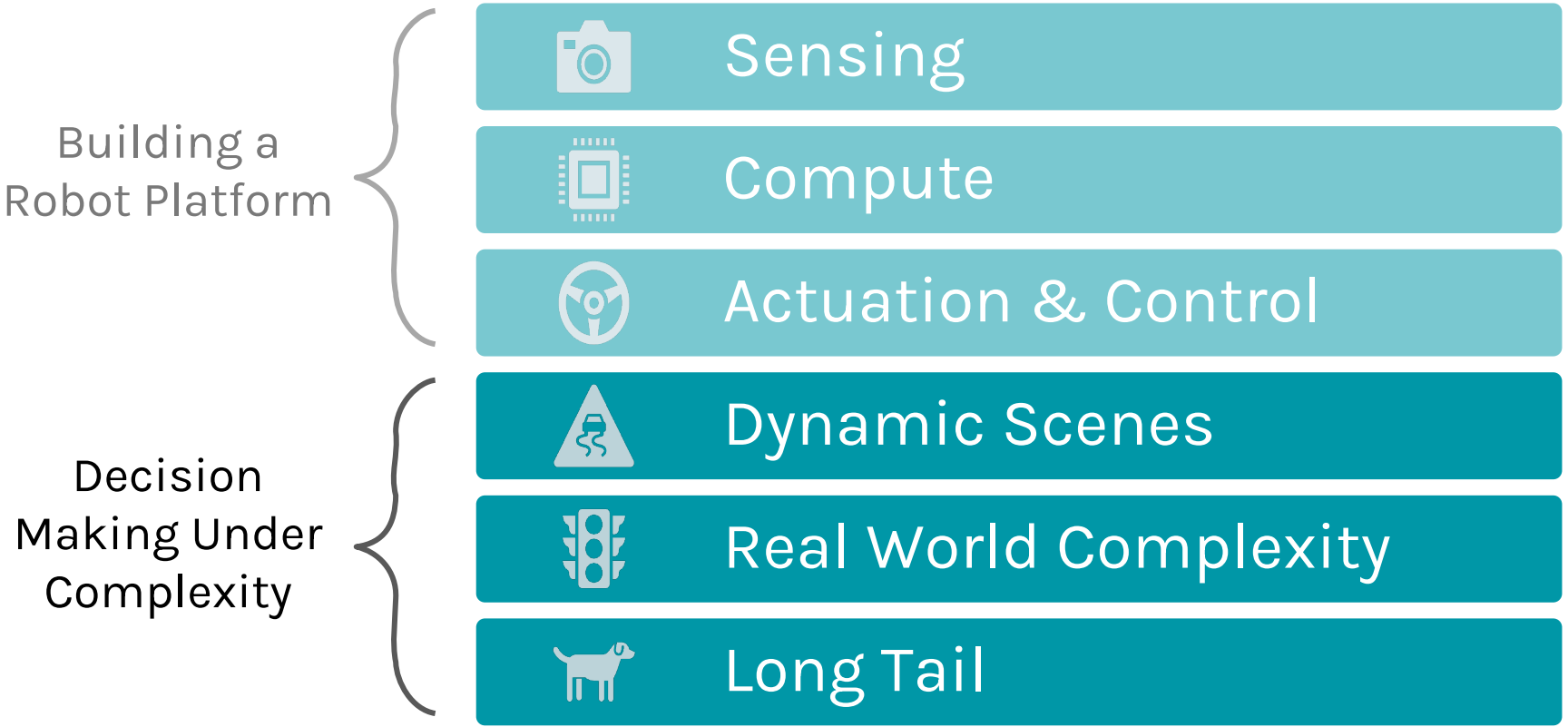
Accelerator

Indicators

Horn? Hazard Lights?

Doors? Locks?

What are the technical challenges?



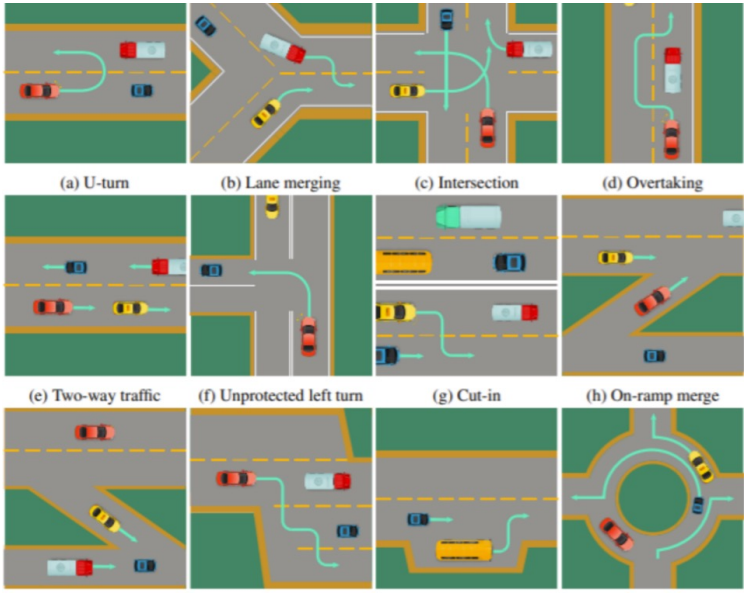
Dynamic Scenes

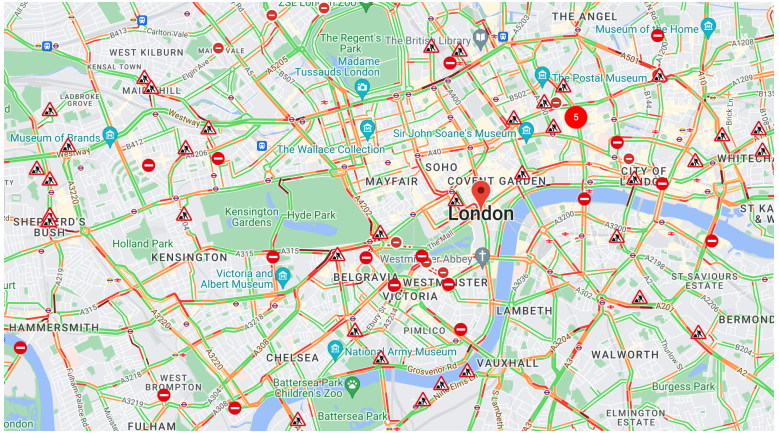
The world is dynamic

Humans are often highly unpredictable

Even the static environment evolves

(e.g. >100 different roadworks per day in Liverpool UK)





Real World Complexity

Real World Complexity


30

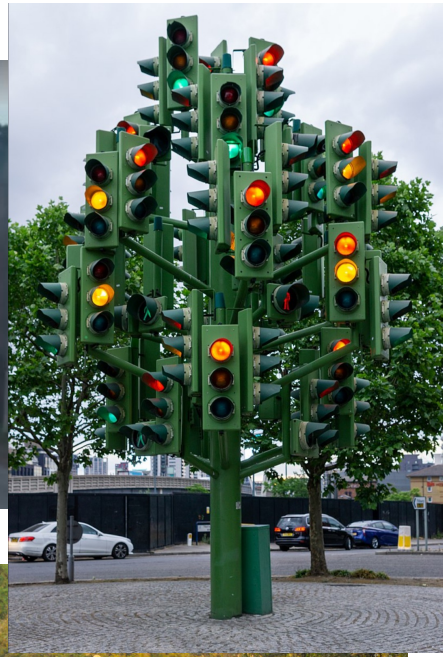

Pilgrim Payne


KW19 VJO


20

Long Tail


SAVOY THEATRE
Fiddler on the Roof
SAVOY




BUS
STOP
© Ross Parry


CPL
CONVOI EXCEPTIONNEL


@vipchaudhry

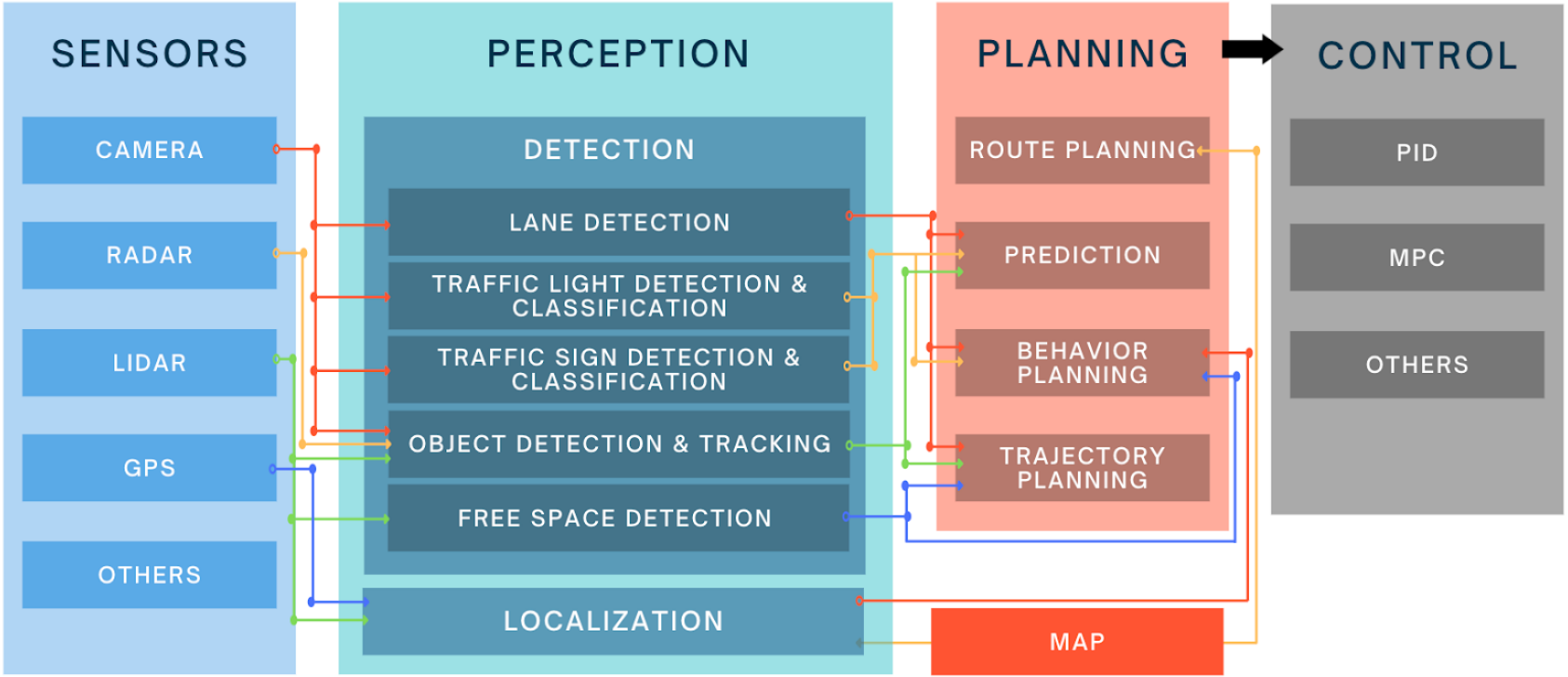
WAYVE

2. How to build your self-driving car

WAYVE

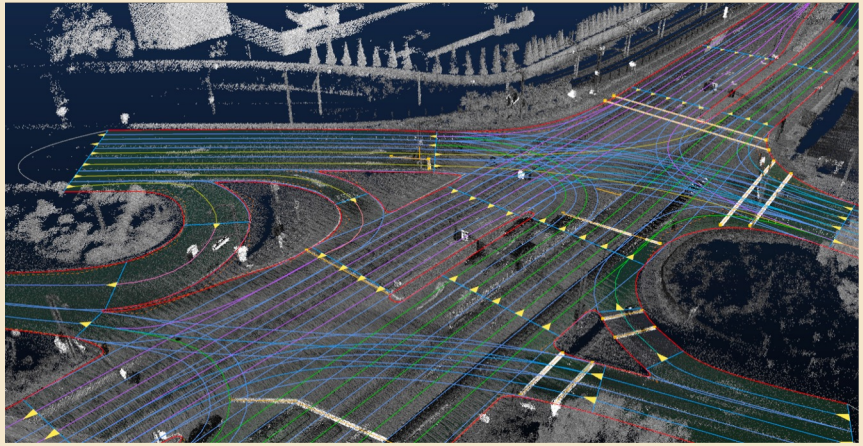
The “Traditional” Approach

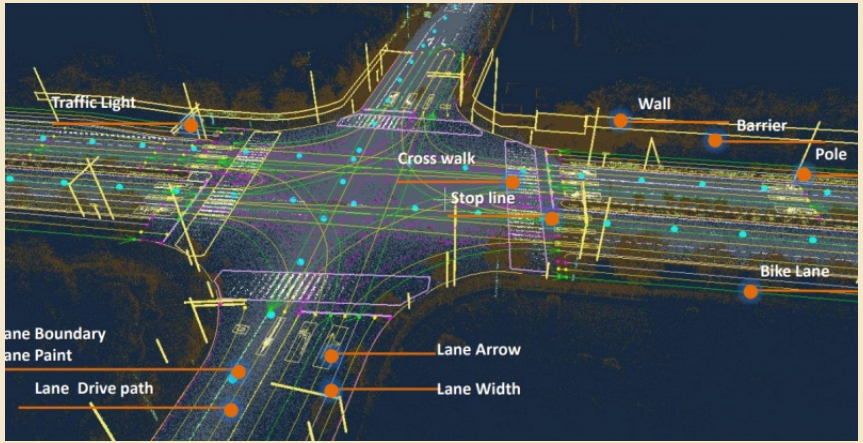
AV 1.0



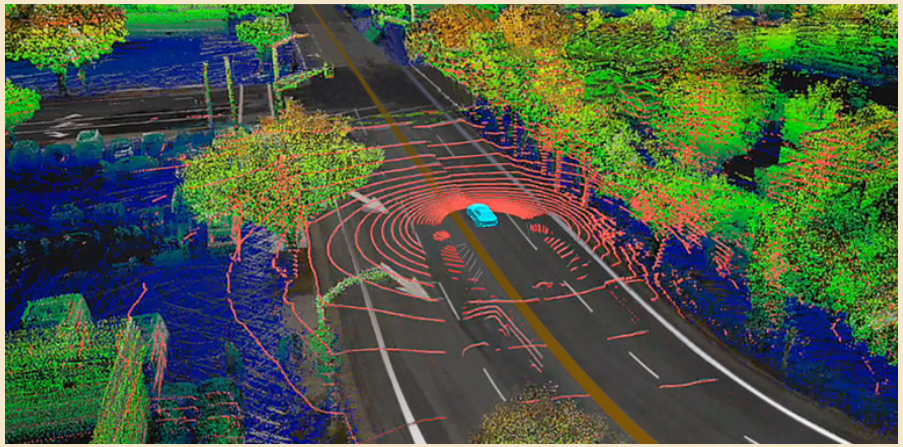
Driving Intelligence System

AV 1.0

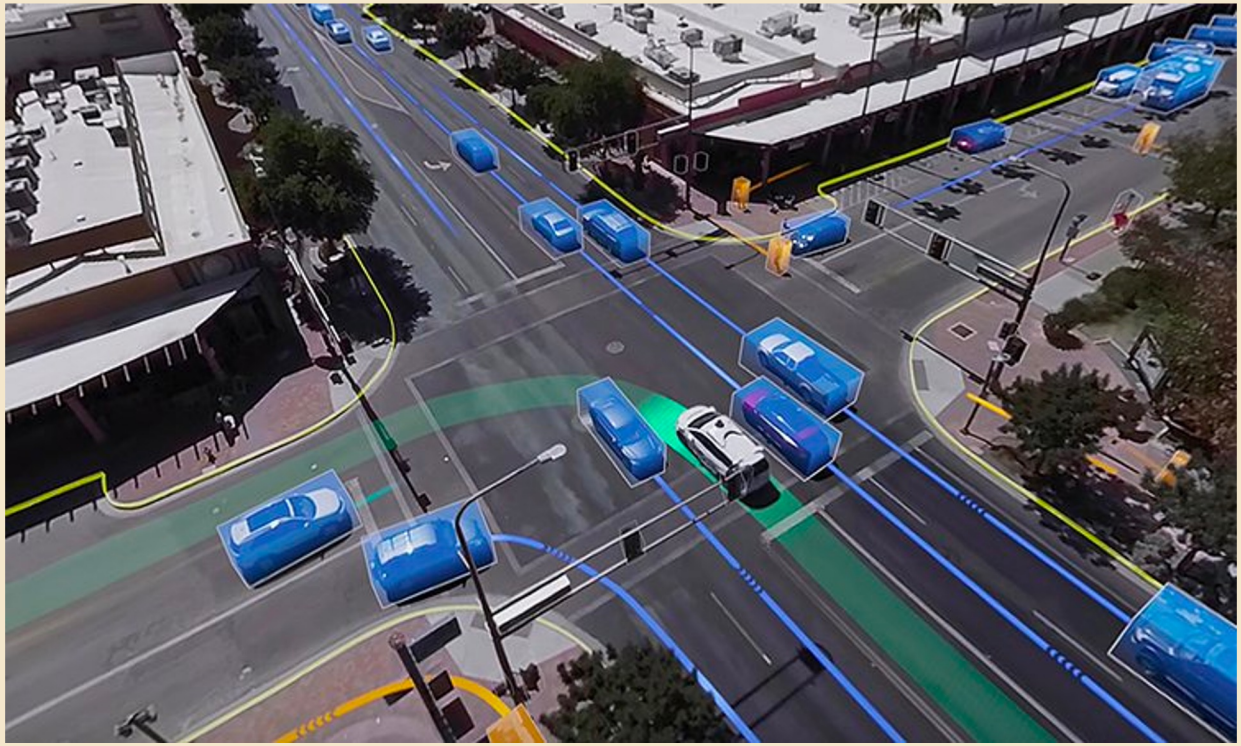




Example HD Maps

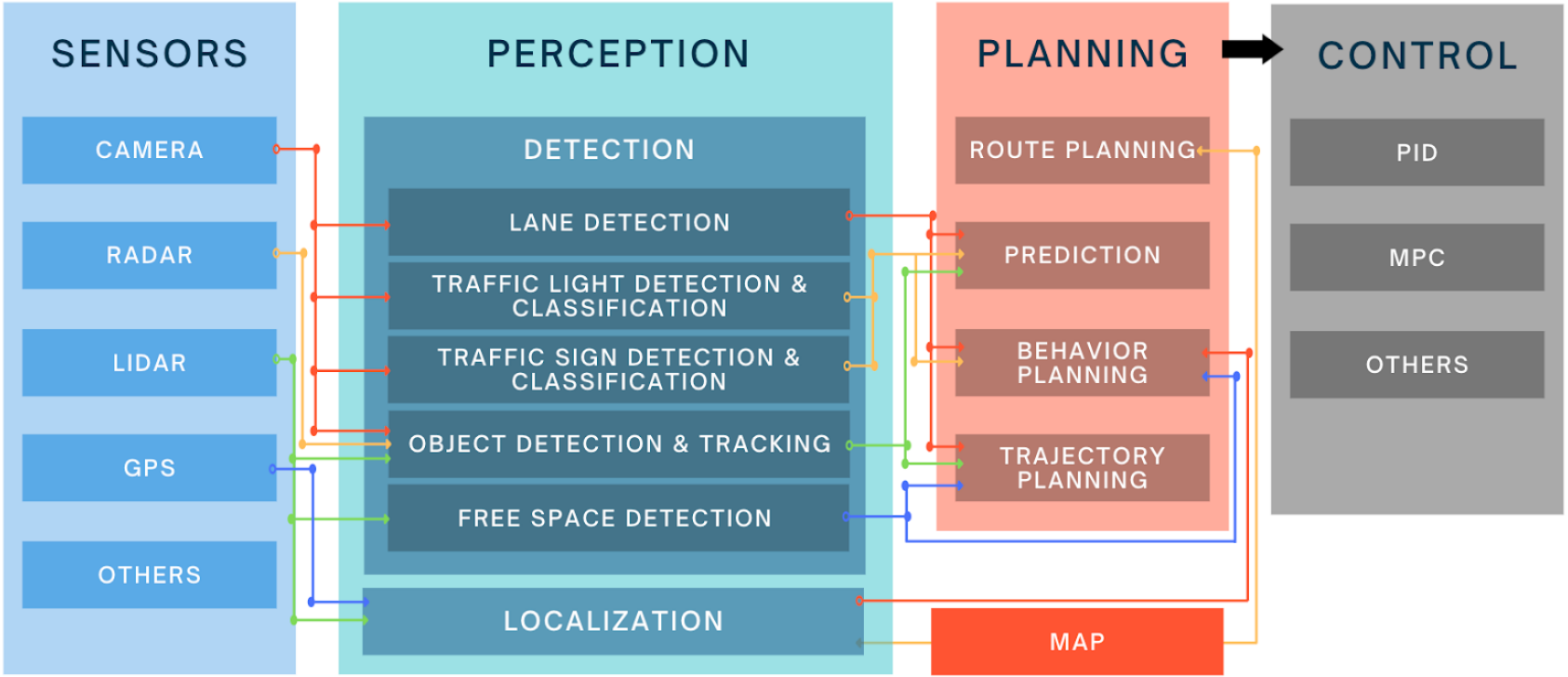


Localizing Lidar against an HD map



Example Perception Outputs, Predictions, and Driving Plan

Image: Waymo



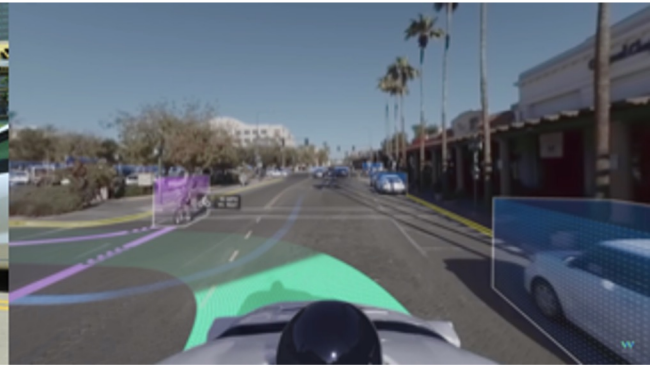
Heavy sensor stack

Many hand-designed interfaces between learned perception and planning modules.
High-definition mapping and precise localization typically required.

AV 1.0

Challenges of AV 1.0

- Expensive to engineer
- Hard to integrate
- Challenging to scale



HD Maps

(brittle / slow to build / expensive to maintain)

LiDAR Sensors

(expensive / short lifespan / hard to integrate)

Hand-Designed Rules/Pipeline/Interfaces

(rigid / brittle / clunky)

WAYVE

The Modern Approach

Deep learning has already achieved superhuman performance in domains that are comparably complex to autonomous driving (but more accessible and structured)

IMAGE RECOGNITION

ImageNet considered a solved problem in 2017

NATURAL LANGUAGE & VISION

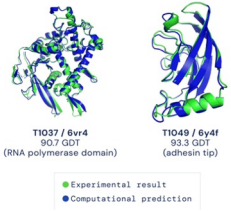
DALL-E 2 / Imagen creating images from text (OpenAI / Google)

GAMES

AlphaStar winning Starcraft (DeepMind)

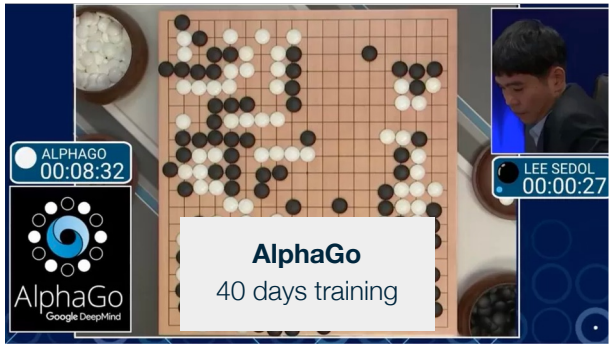
MuZero learning Go, Chess, Shogi, and Atari (DeepMind)

BIOCHEMISTRY



AlphaFold solving protein folding (DeepMind)

Scale of data + compute is driving AI breakthrough after breakthrough



AlphaGo
40 days training

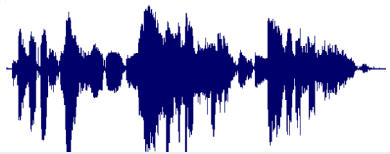
CLIP
400M image, text pairs

AlphaStar
44 days training (=200 years of experience)

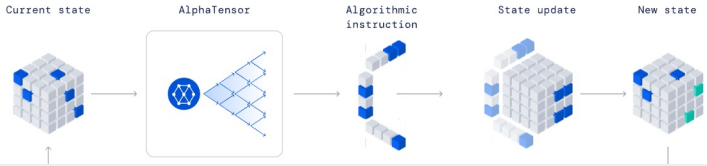
StableDiffusion
2.3B images

```
Q: What is your favorite animal?
A: My favorite animal is a dog.

Q: Why?
A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a bad mood?
A: Two reasons that a dog might be in a bad mood are if it is hungry or if it is hot.
```

GPT-3
175B parameters



Whisper
700,000 hours (80 years!) of audio



AlphaTensor
600,000 iterations batch size 2048 across 64 TPUs


WAYVE
Autonomous driving is an embodied intelligence problem
JAGUAR
I-PACE
LF19 MSK
WAYVE
Admiral

SENSORS

CAMERA

RADAR

GPS

END-TO-END AI

SAFETY & VERIFICATION COMPUTER

CONTROL

VERIFIED MOTION PLAN

DRIVE BY WIRE

Lightweight and flexible sensor stack

Single end-to-end deep neural network

No HD mapping or localization

Trained using self-supervised learning and expert human demonstration

AV 2.0

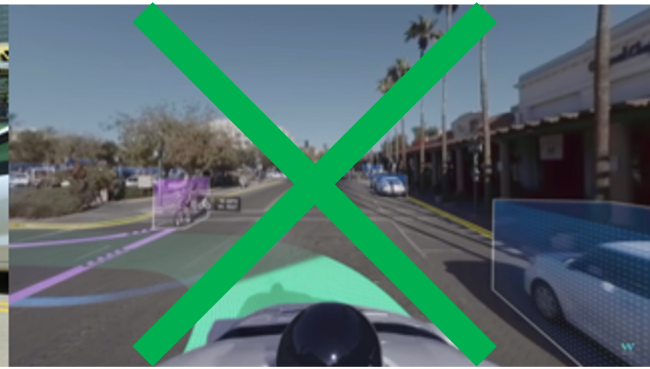
WAYVE

HD Maps

(brittle / slow to build / expensive to maintain)

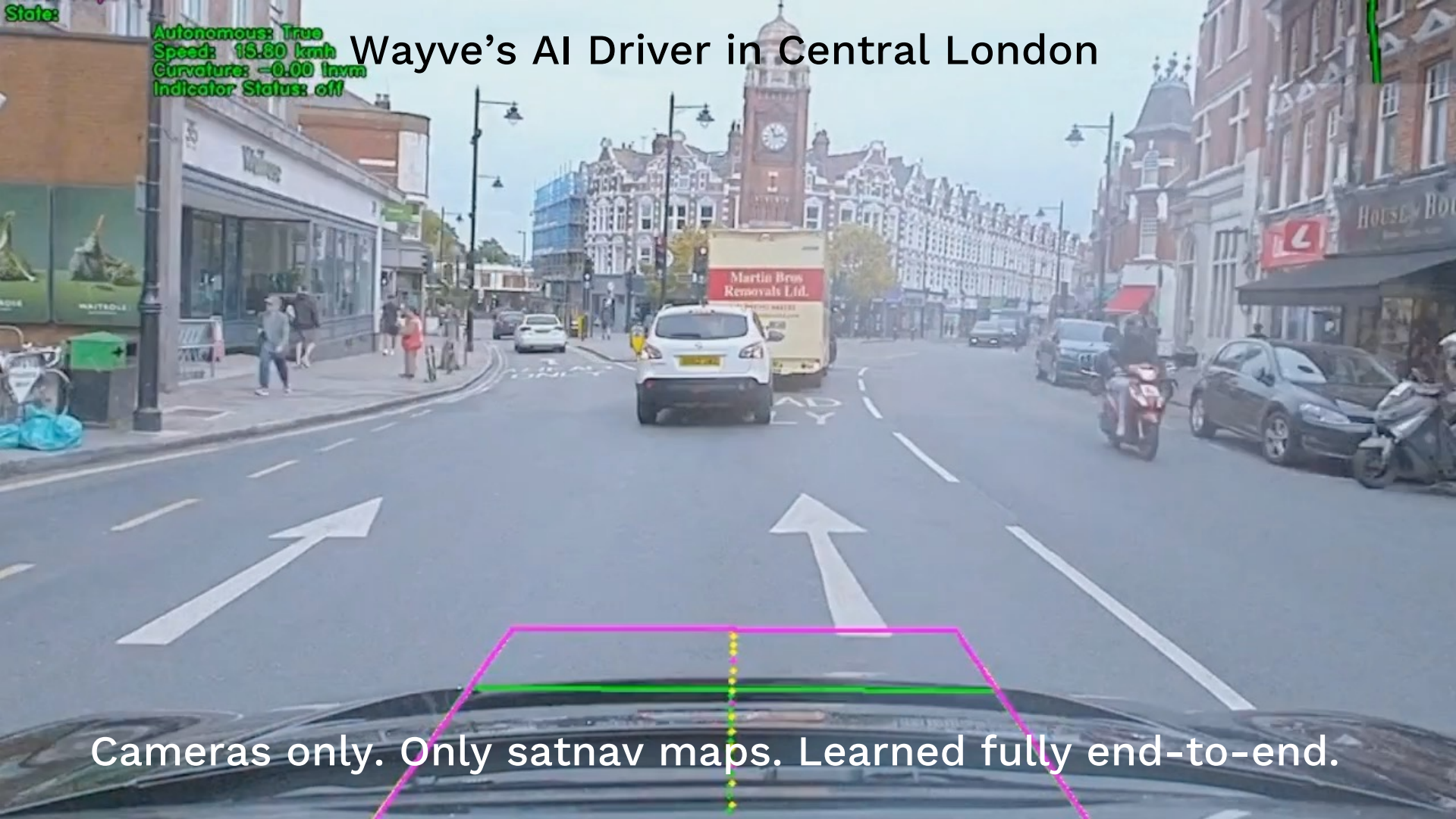
LiDAR Sensors

(expensive / short lifespan / hard to integrate)



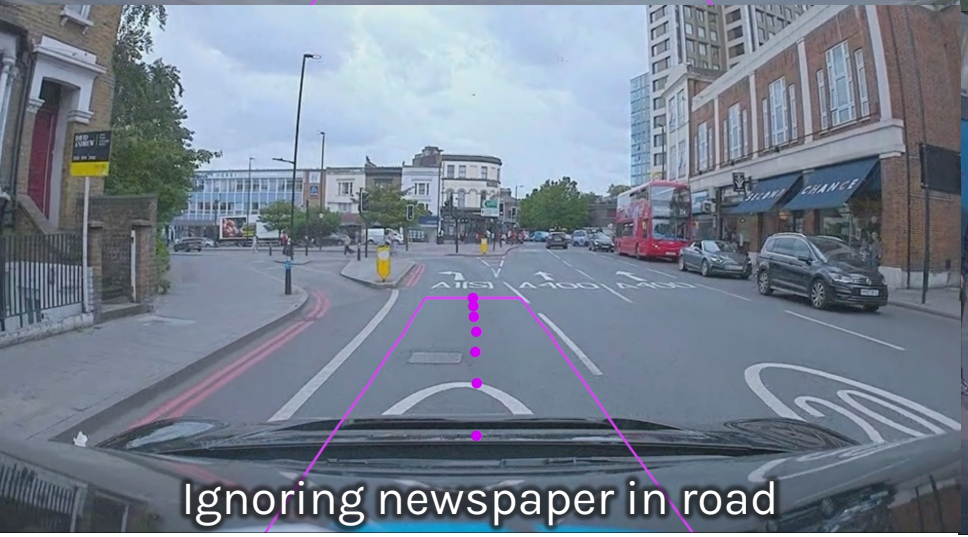
Hand-Designed Rules/Pipeline/Interfaces

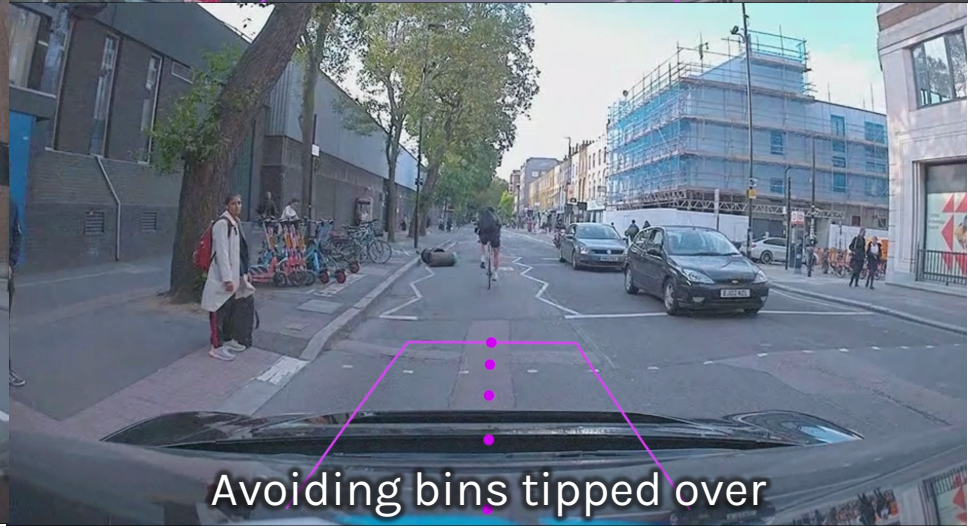
(rigid / brittle / clunky)


State:
Autonomous: True
Speed: 15.80 kmh
Curvature: -0.00 1nvm
Indicator Status: off
Wayve’s AI Driver in Central London
Martin Bros
Removals Ltd.
Cameras only. Only satnav maps. Learned fully end-to-end.


Complex intersections


Negotiation in narrow road


Ignoring newspaper in road


Avoiding bins tipped over


Autonomous: True
Speed: 11.54 mph
WAYVE

WAYVE

3. Research highlights

World Models

Accumulate knowledge, common-sense

“Understand the rules of the world before understanding the rules of the road.”

Motivations

Equip our driving models with a strong understanding of the world:

- Semantics
- 3D geometry
- Motion
- Interaction

To drive well, you need to be able to predict the future... and the different possibilities the future holds

A world model is a generative model that predicts what happens next conditioned on some context: $P(s_{t+1}|s_t, c_t)$

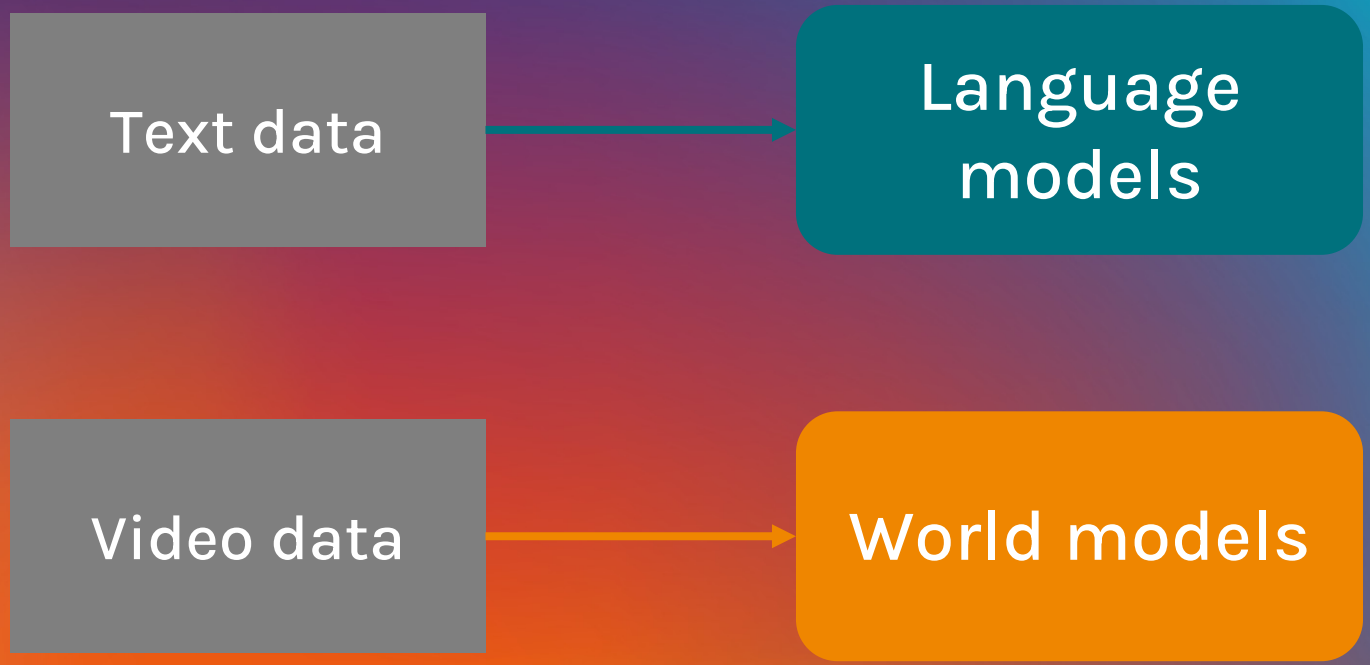
Applications

Representation learning	Learned Simulator	Model-based reinforcement learning	Search-based planning

... and more

How do we train a world model?

Self-supervised training

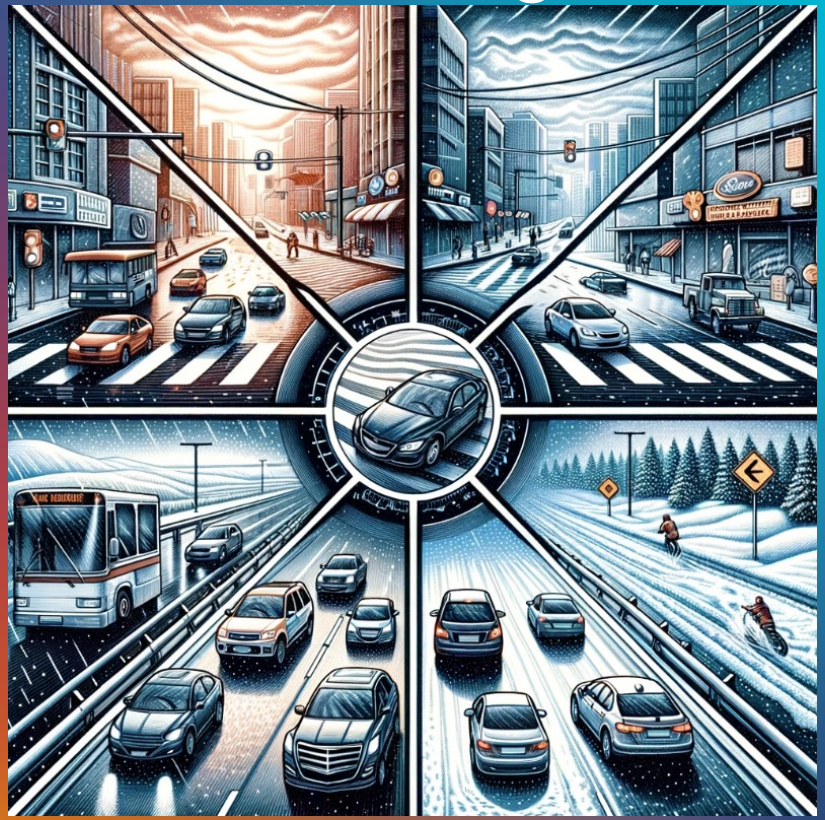


World models can learn from any data

- Imperfect driving

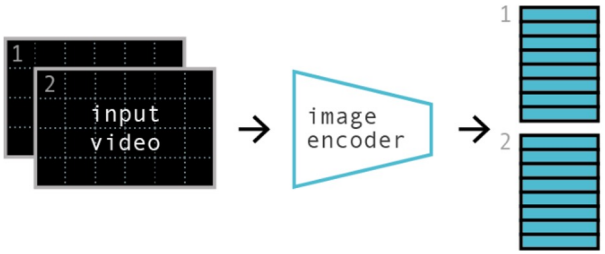
- Videos not even related to driving

-> Contribute to enrich the model's knowledge about the world.




WAYVE
GAIA-1
Generative AI for Autonomy

GAIA-1 Architecture

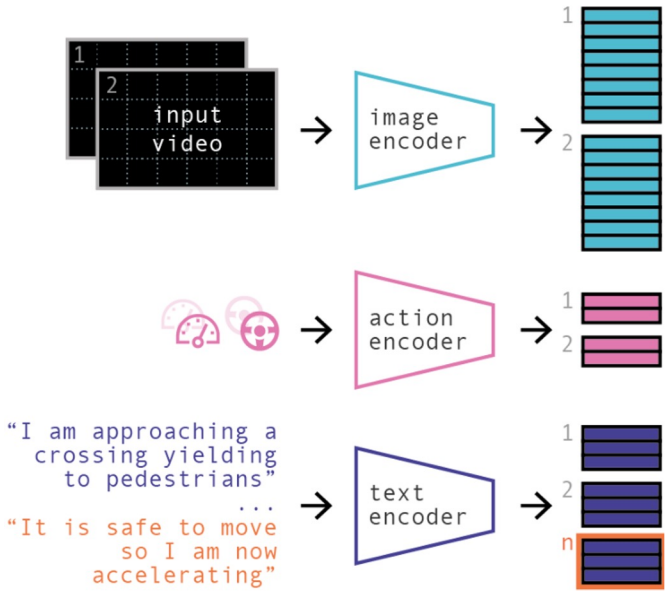


Objectives of the image tokenizer:

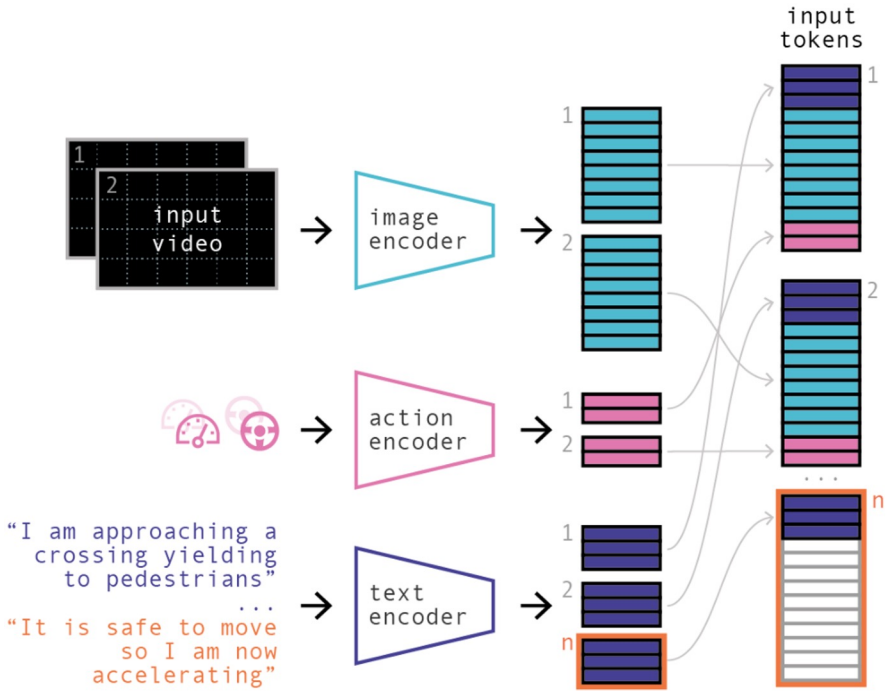
- **Compress the information** from raw pixels to make the sequence modelling problem tractable.

- Guide the compression towards **meaningful representations** instead of high frequency signals.

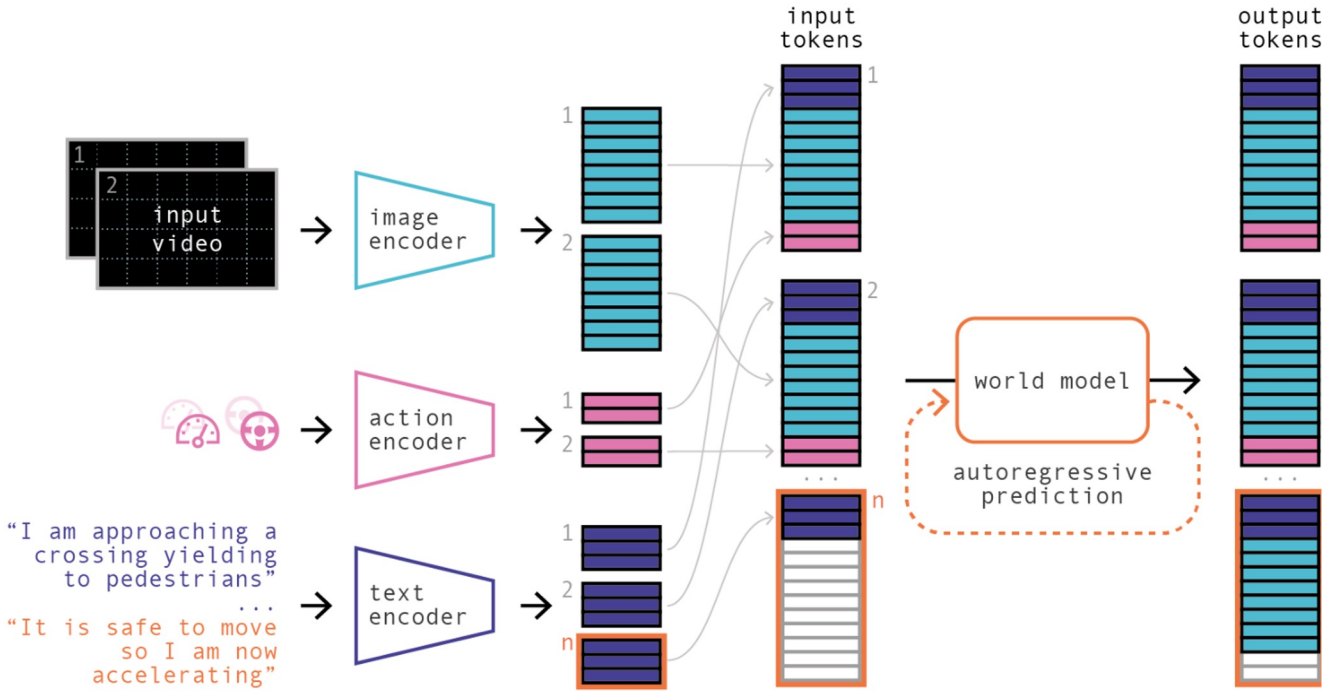
GAIA-1 Architecture



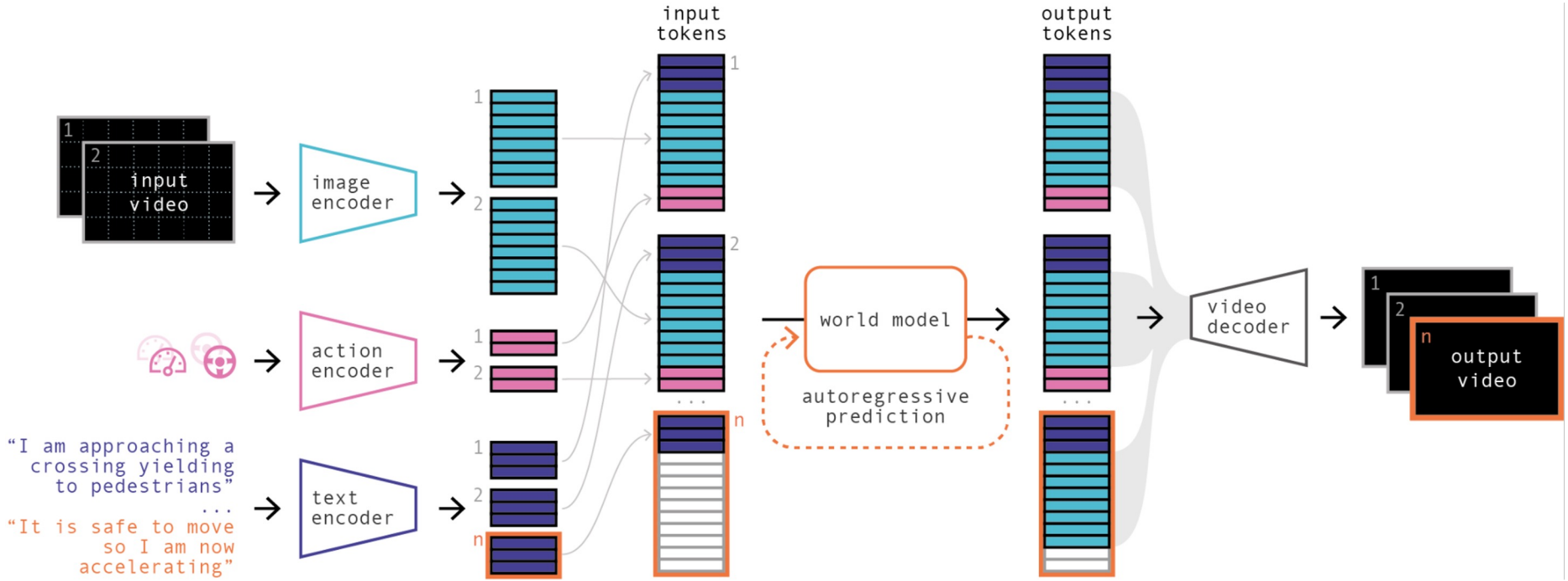
GAIA-1 Architecture



GAIA-1 Architecture



GAIA-1 Architecture

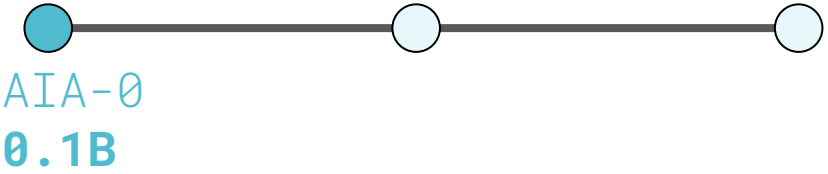


WAYVE

Scaling GAIA-1

WAYVE

GAIA-0 (November 2022)



GAIA-0 (Nov. 2022)

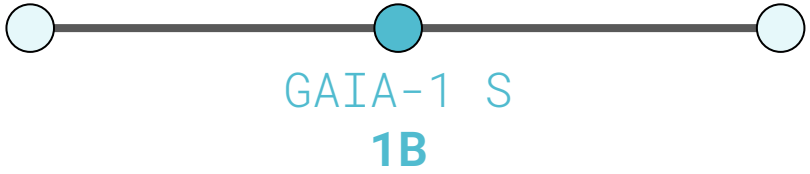
- 0.1B parameters
- Trained for 20 GPU days
- On 20B training tokens

Forcing a left drift, then a right drift

Forcing the model to stop, then accelerate

WAYVE

GAIA-1 S (June 2023)



GAIA-1 S (June 2023)

- 1B parameters **(10x more parameters)**
- Trained for 120 GPU days **(6x more compute)**
- On 70B training tokens **(4x more data)**

Multiple futures

Action-conditioned rollouts

Context video

Steer left

Go straight

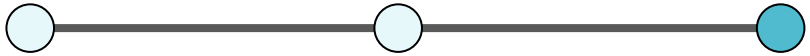
Steer right

Out-of-distribution examples

Out-of-distribution examples

WAYVE

GAIA-1 (Today)



GAIA-1 (Today)

- 10B parameters **(10x more parameters)**
- Trained for 1500 GPU days **(12x more compute)**
- On 300B training tokens **(4x more data)**

Comparison

GAIA-0
0.1B

GAIA-1 S
1B

GAIA-1
10B

GAIA-0 (0.1B parameters)

GAIA-1 S (1B parameters)

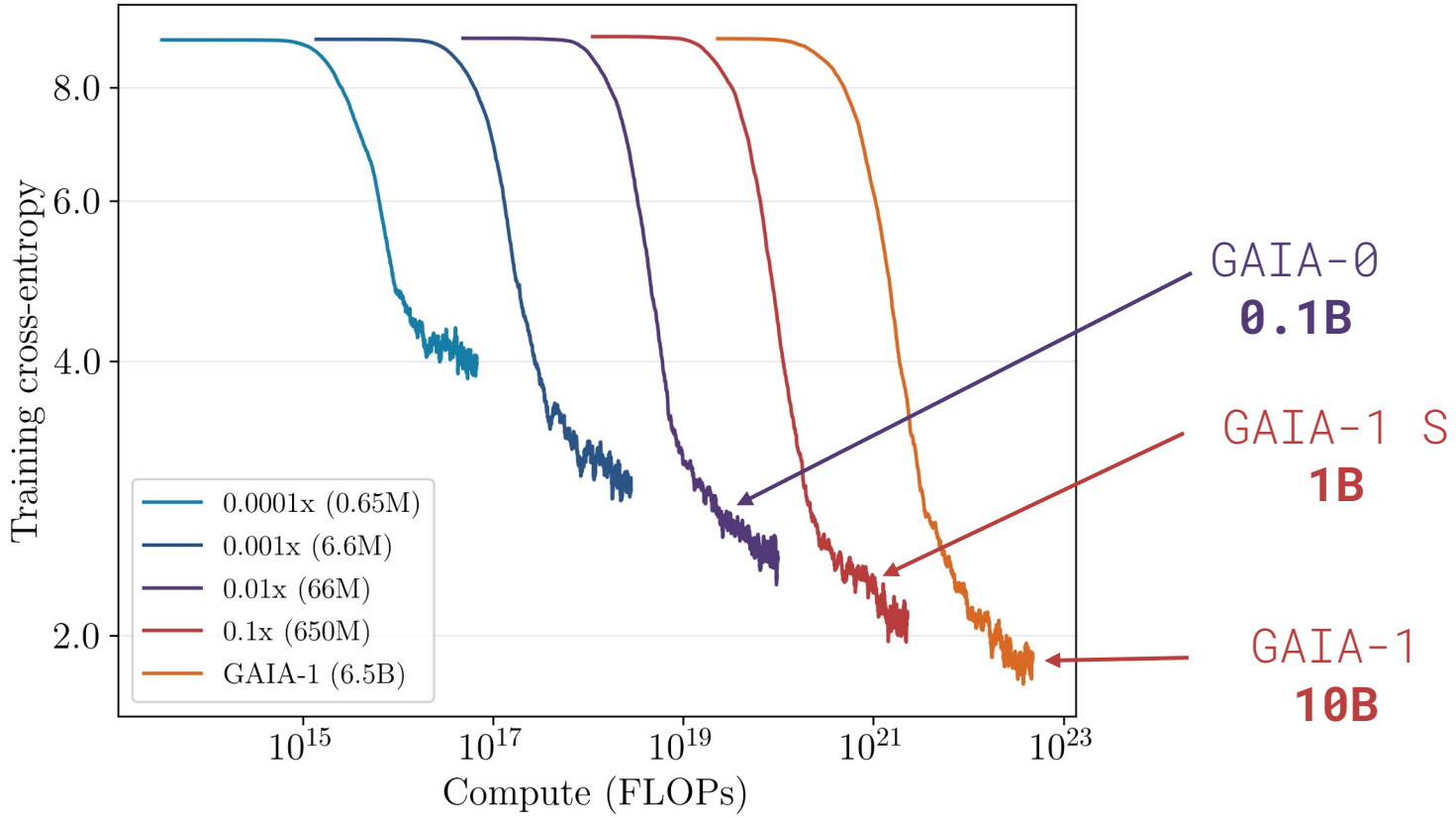
GAIA-1 (10B parameters)



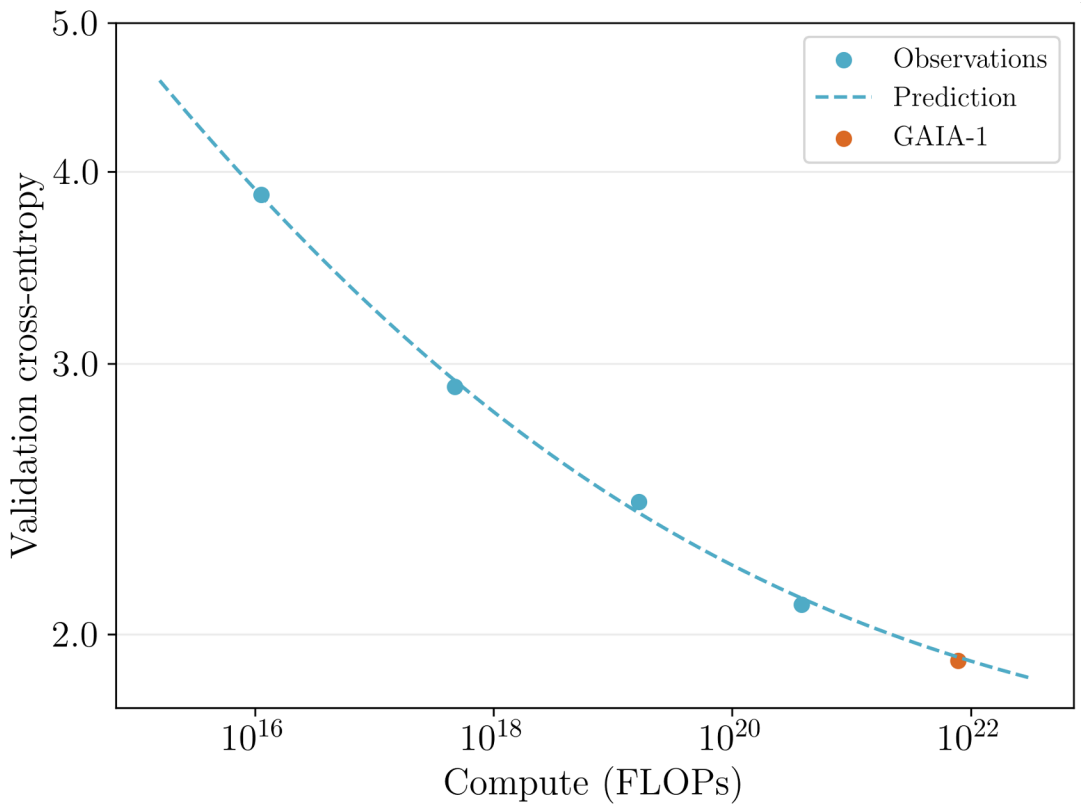

WAYVE

Scaling curves

Scaling curves of models 10,000x to 10x smaller



Final model performance predicted with 20x less compute



WAYVE

Capabilities

Long Video Generation

Generated by GAIA-1

Diverse Futures

Generated by GAIA-1

Action Conditioning

Generated by GAIA-1

Language Conditioned

Generated by GAIA-1

Language Conditioned

Generated by GAIA-1

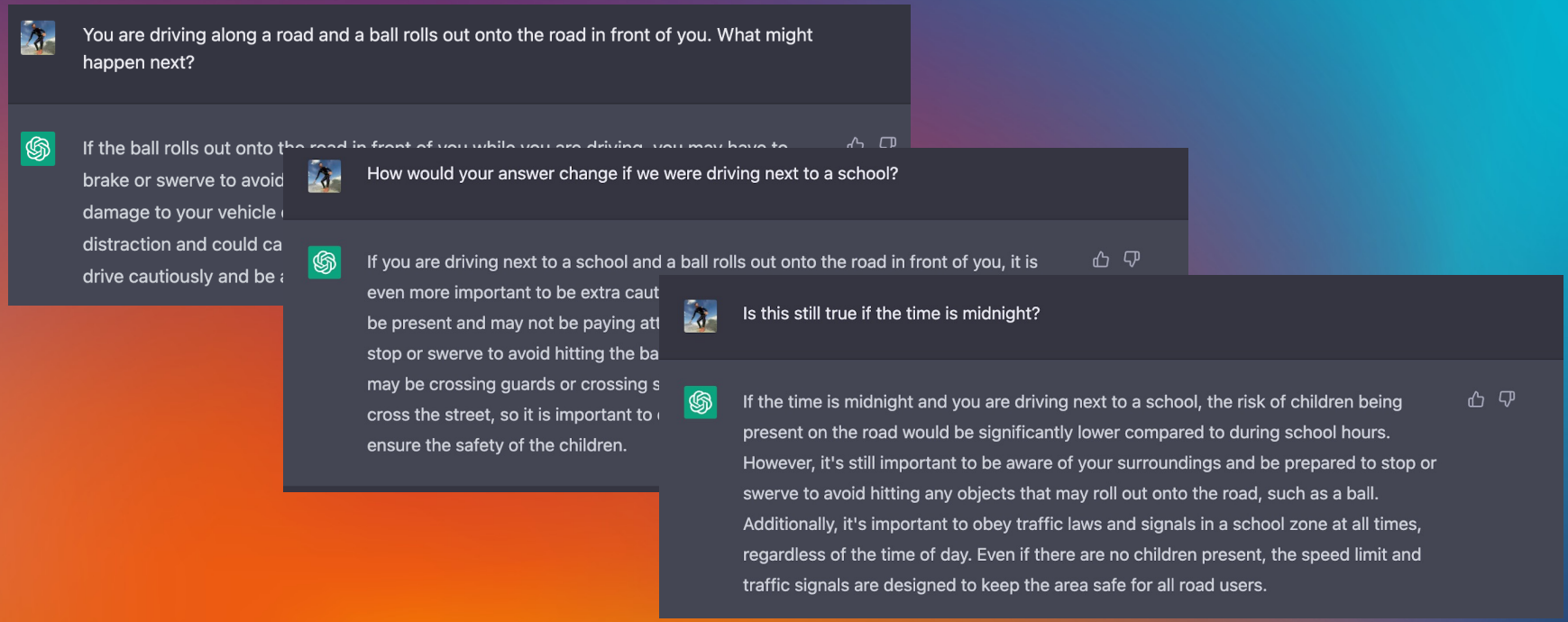
Opportunities ahead

- Integrating GAIA into Wayve's on-road Driving AI
- Scaling up (even more), multi-view generation, multi-sensor
- GAIA as a learned simulator for training and evaluation

Language Meets Driving

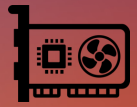
Language meets driving

LLMs encode huge knowledge about driving



Opportunities for language in autonomous driving

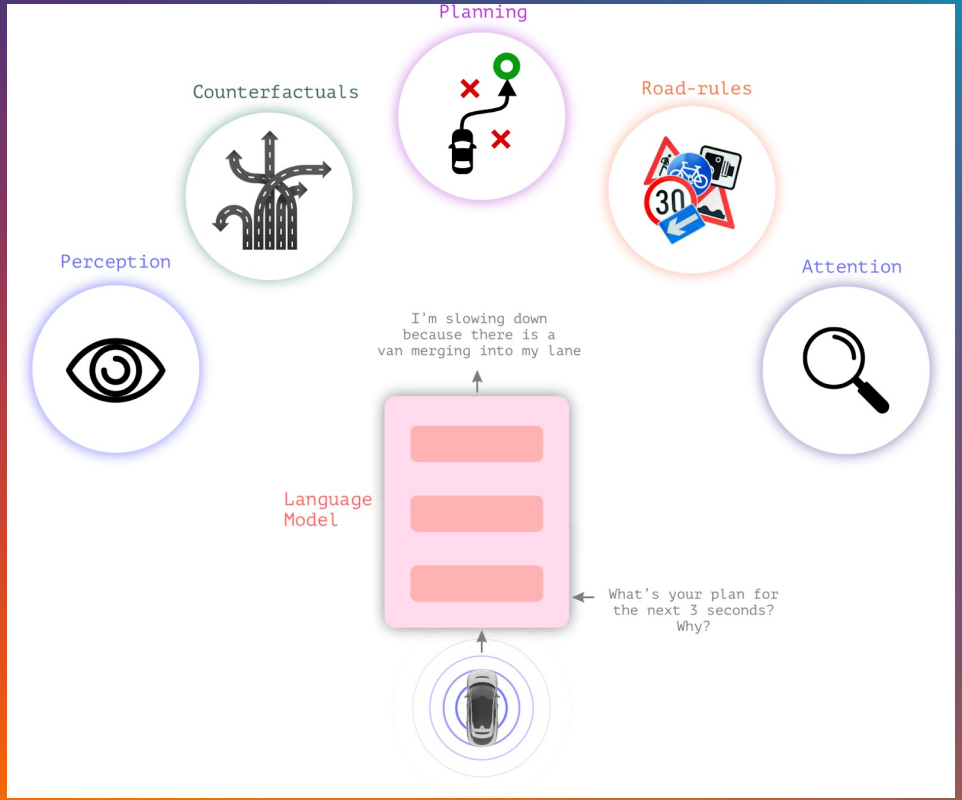
Explainability – user and regulator confidence



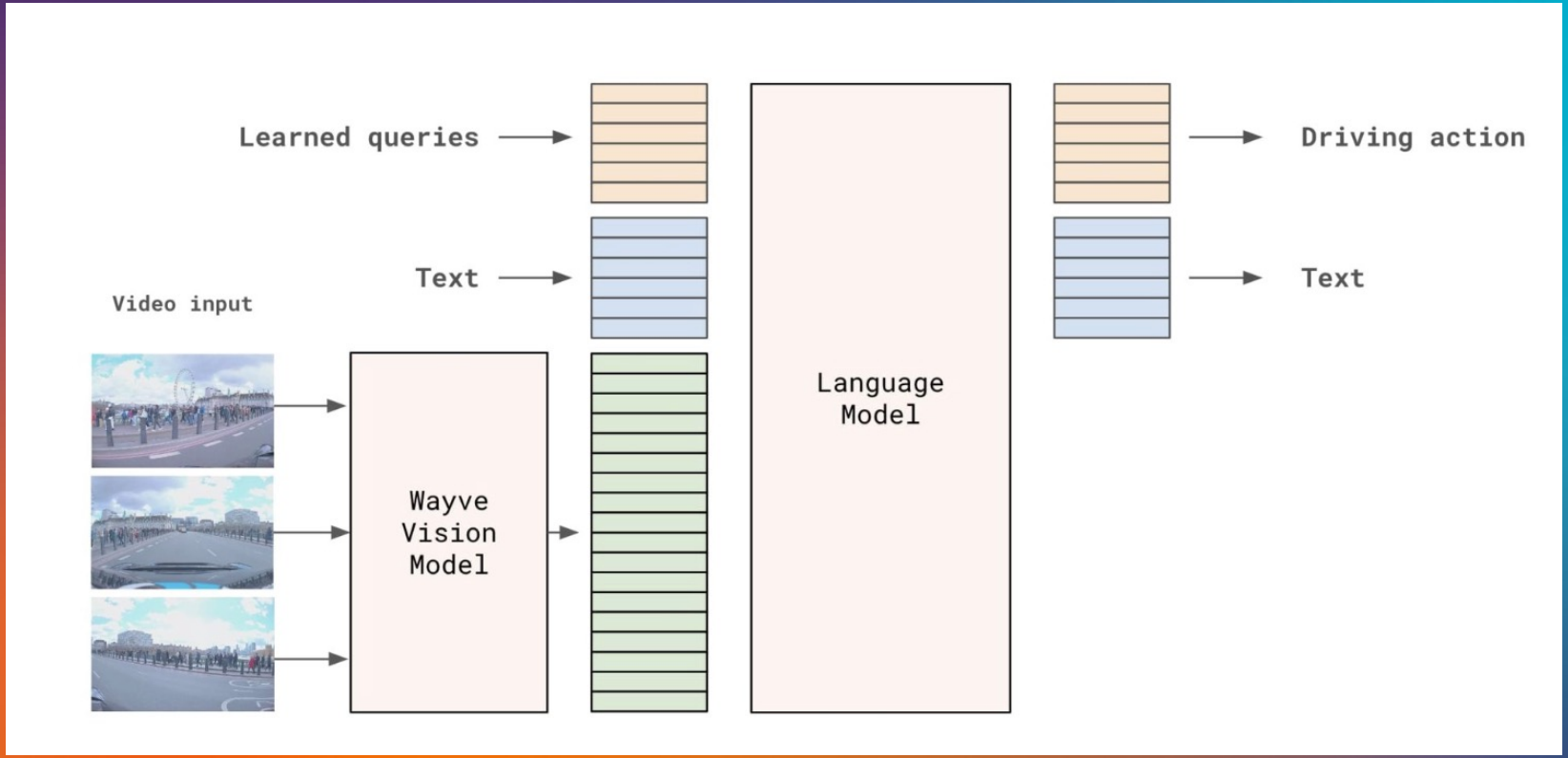
Multimodal training – data efficiency

Common-sense reasoning – solving the long tail

Explainability



LINGO: Natural Language meets Driving



Staying stopped due to the red light.

Future Direction: Multimodal Training

"We must stay stopped until all pedestrians have cleared the zebra crossing.

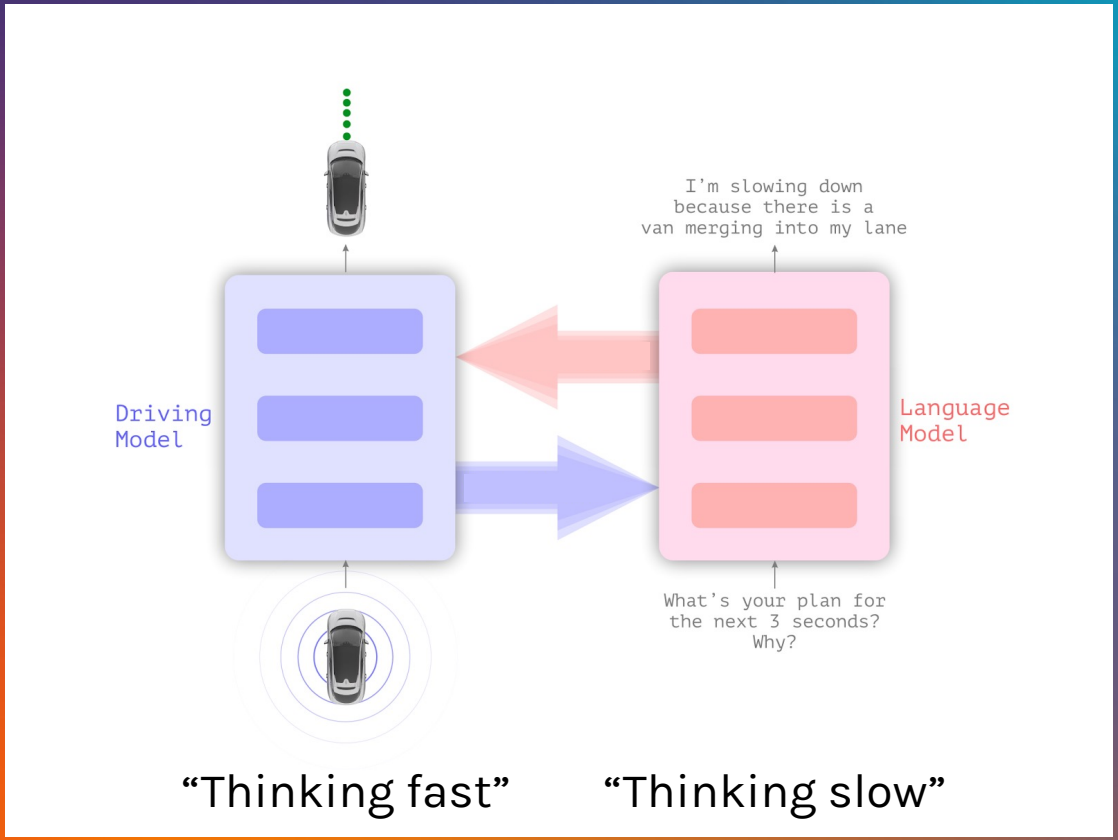
There are several pedestrians ahead we need to pay attention to, even those not in the zebra crossing.

Once they are clear we can proceed to the intersection where we must give way to cross traffic.

There's also a 20mph speed limit sign, so we must adjust our speed accordingly."

Opportunity for enormous data efficiency: a paragraph of text may be worth a thousand videos

Future Direction: Reasoning for Driving



WAYVE

Conclusions

Conclusions

- Autonomous driving is the next major frontier in AI
- End-to-end learned driving offers a scalable solution
- Huge scope for further innovation in data, models, and learning


WAYVE