How We Test Self-Driving Cars

And How We Explain Their Failures

Leilani H. Gilpin
Assistant Professor
Dept. of Computer Science &
Engineering, UC Santa Cruz

Introduction and Disclaimer: My Car Failures

Agenda

<u>Motivate problem: Complex systems are prone to failure</u>

Local sanity checks for vehicle perception

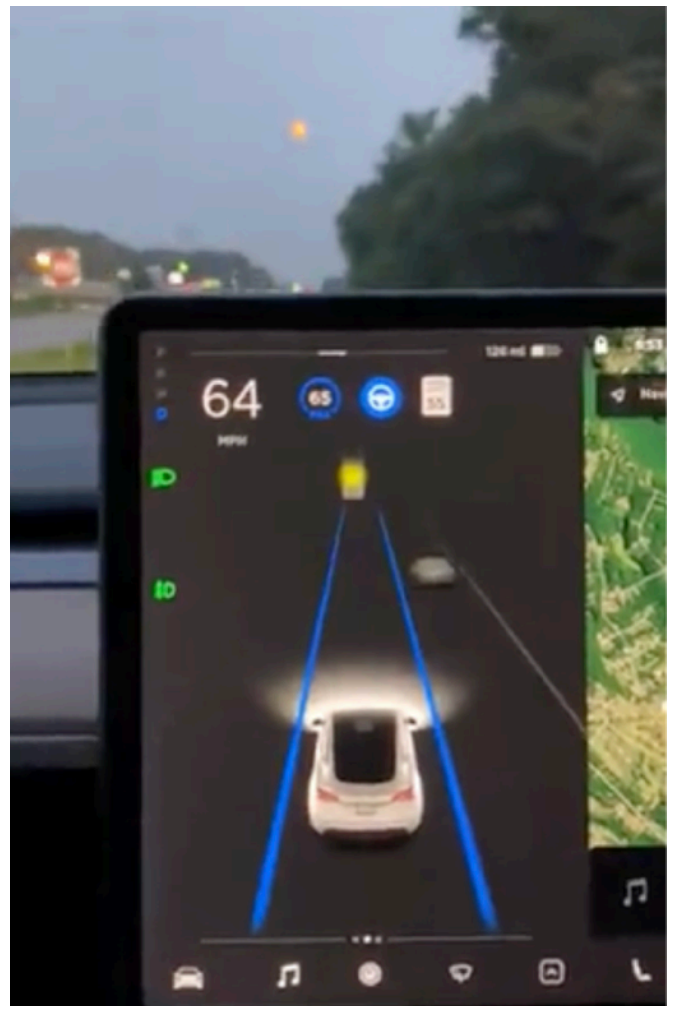
Explanations as an Internal Debugging Language for Complex Systems

Ongoing Work: Testing Autonomous Vehicles by Augmenting Datasets

Question: What are the eXplanatory AI (XAI) methods for testing autonomous vehicles in safety-critical scenarios?

Complex Systems Fail in Complex Ways

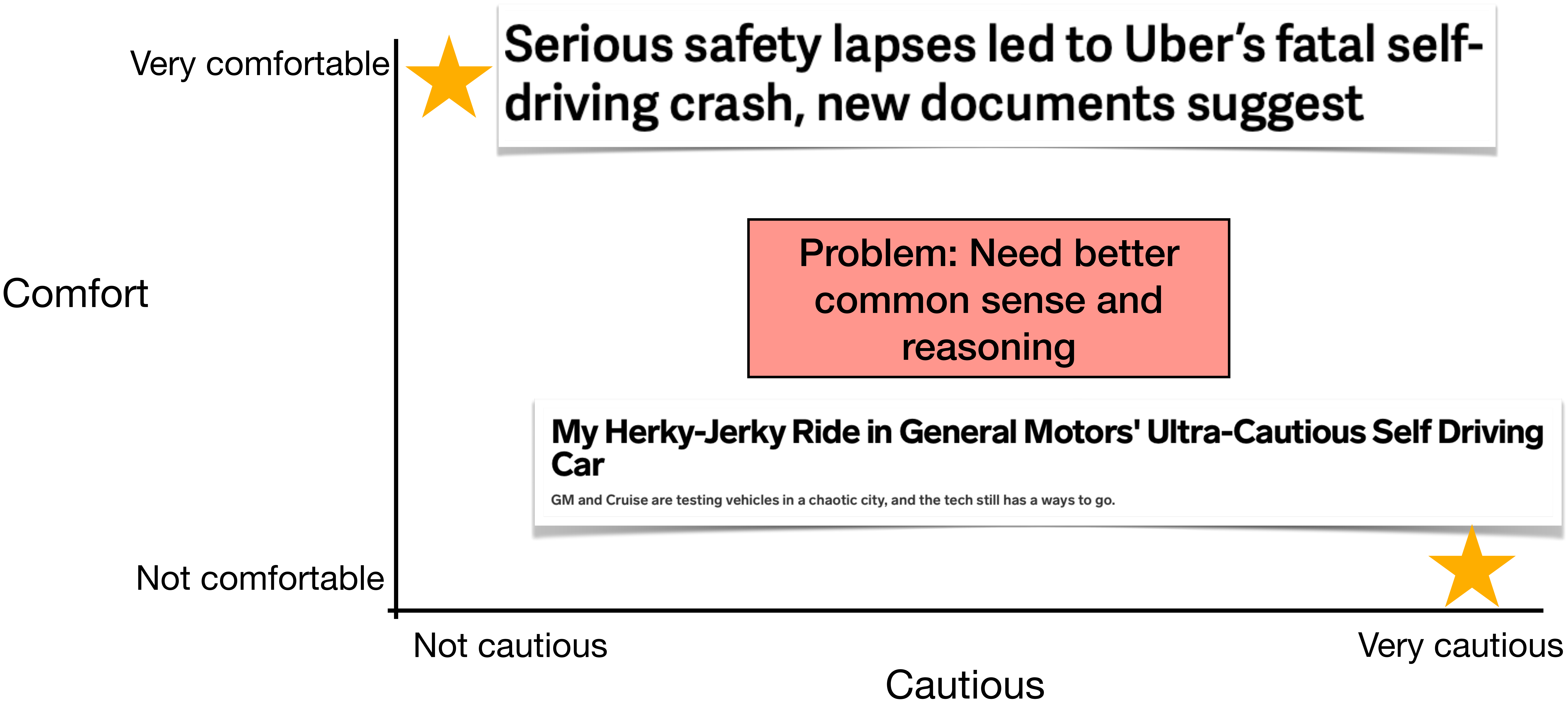
K. Eykholt et al. “Robust Physical-World Attacks on Deep Learning Visual Classification.”



Predictive Inequity in Object Detection

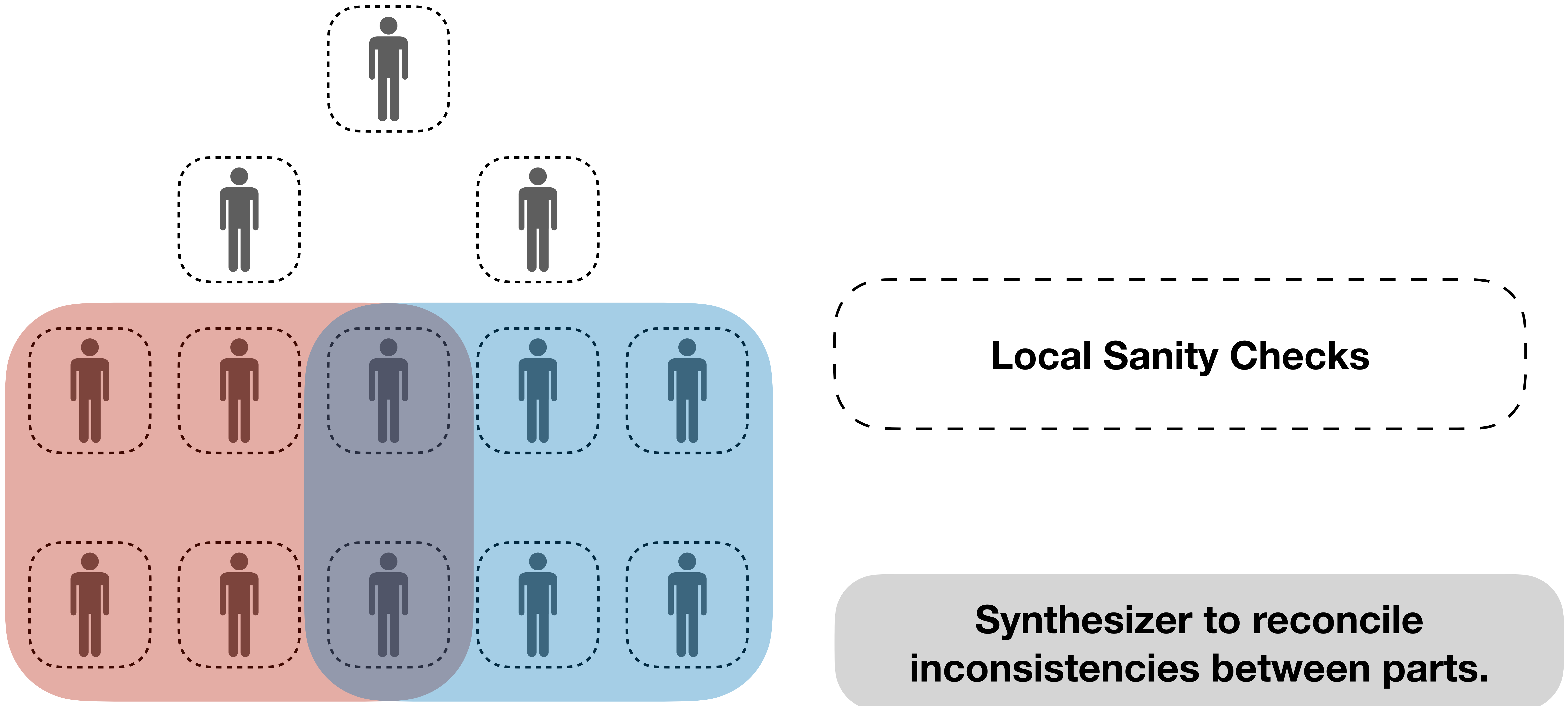
Benjamin Wilson [1] Judy Hoffman [1] Jamie Morgenstern [1]

Autonomous Vehicle Solutions are at Two Extremes



Architecture Inspired by Human Organizations

Communication and Sanity Checks

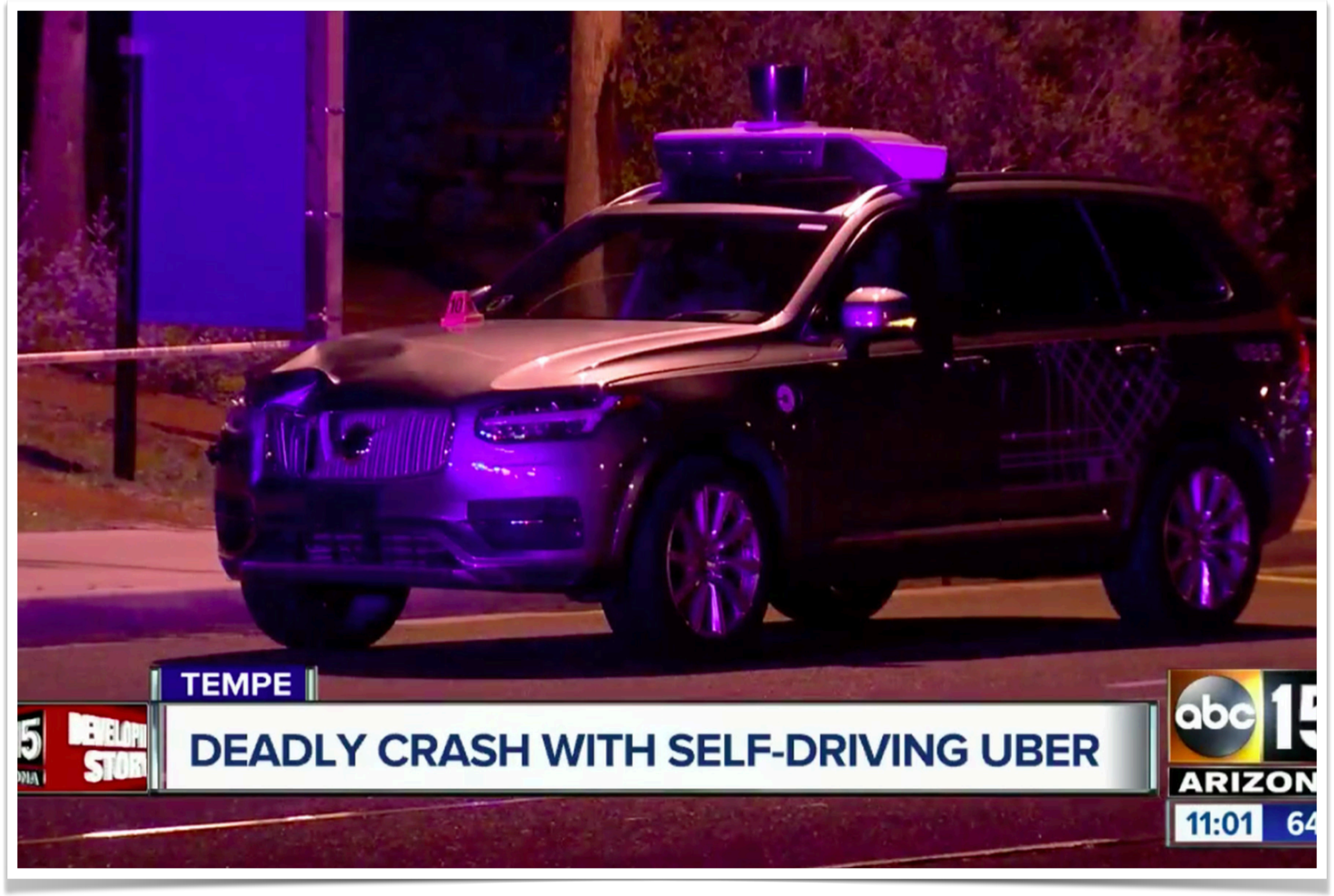


1. Hierarchy of overlapping committees.
2. Continuous interaction and communication.
3. When failure occurs, a story can be made, combining the members’ observations.

An Architecture to Mitigate Common Problems

Synthesizer to reconcile inconsistencies between parts.

Local Sanity Checks




future tense

The Trollable Self-Driving Car

Reconcile conflicting reasons.

Justify new examples.

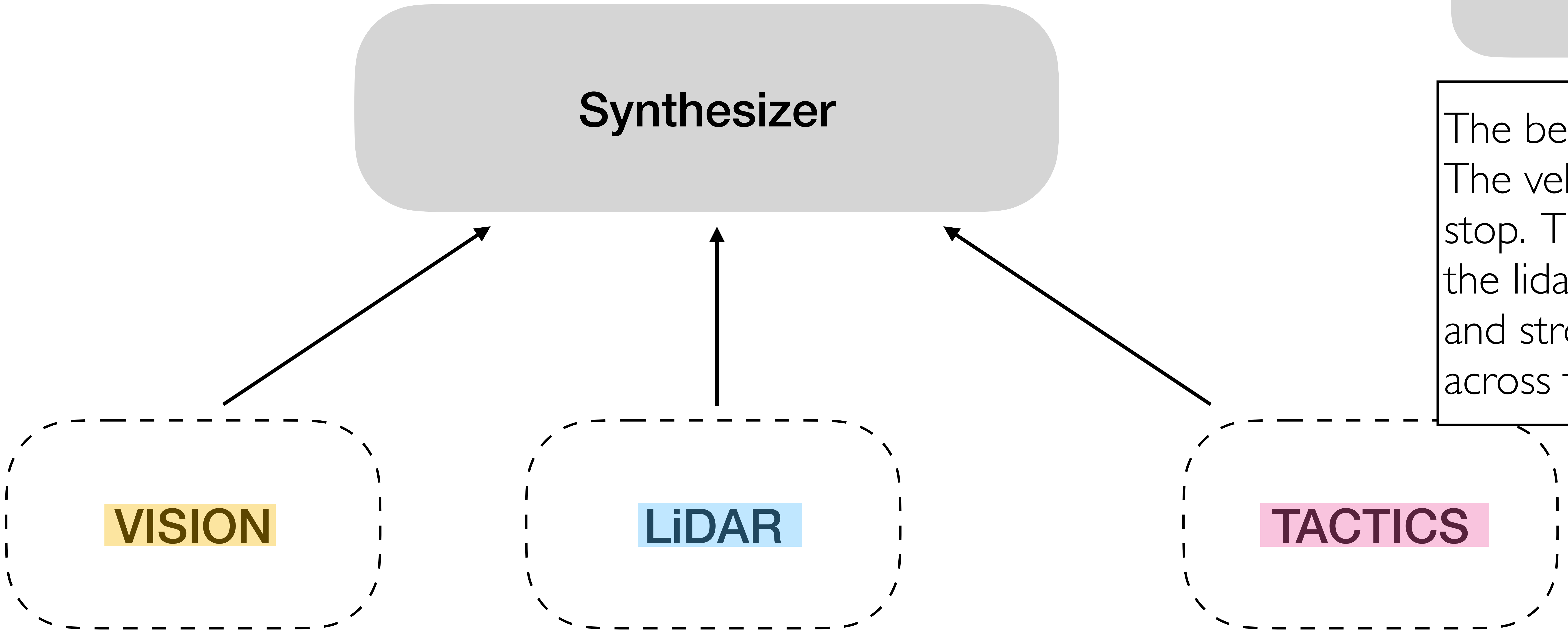
An Existing Problem

The Uber Accident

Solution: Internal Communication

Anomaly Detection through Explanations

Synthesizer to reconcile inconsistencies between monitor outputs.



The best option is to veer and slow down. The vehicle is traveling too fast to suddenly stop. The vision system is inconsistent, but the lidar system has provided a reasonable and strong claim to avoid the object moving across the street.

Agenda

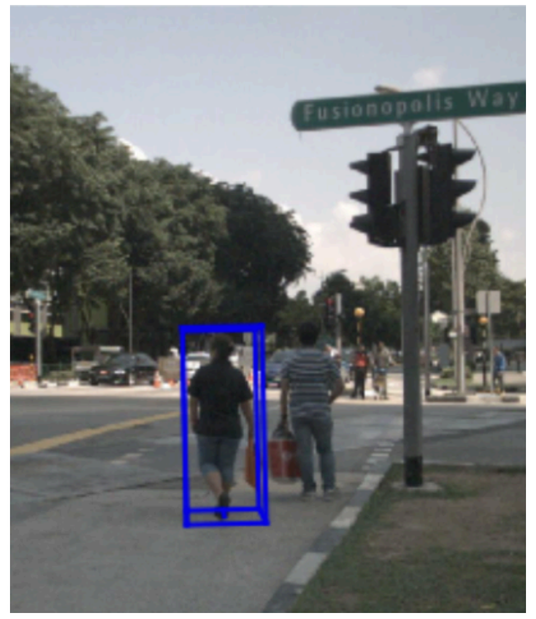
Motivate problem: Complex systems are prone to failure

<u>Local sanity checks for vehicle perception</u>

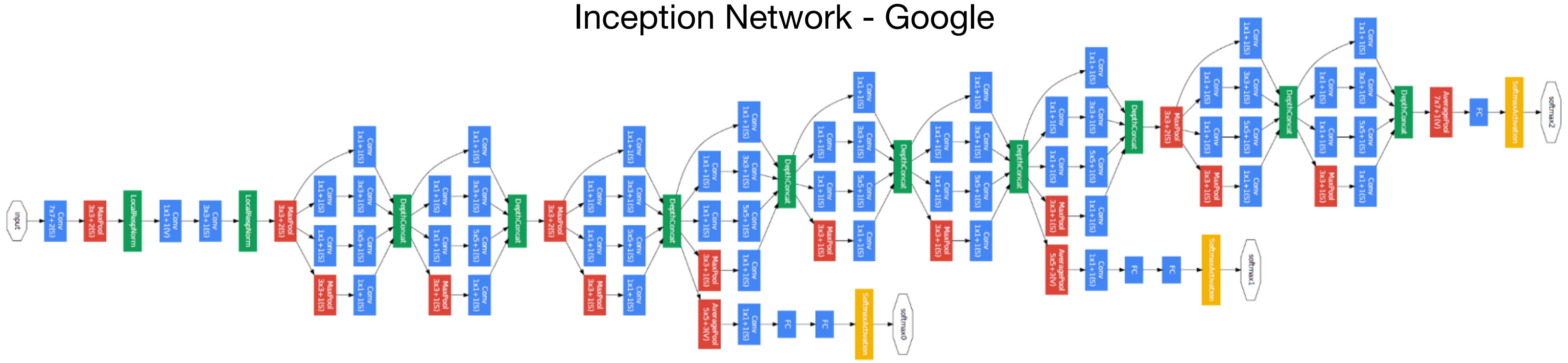
Explanations as an internal debugging language

Ongoing Work: Testing Autonomous Vehicles by Augmenting Datasets

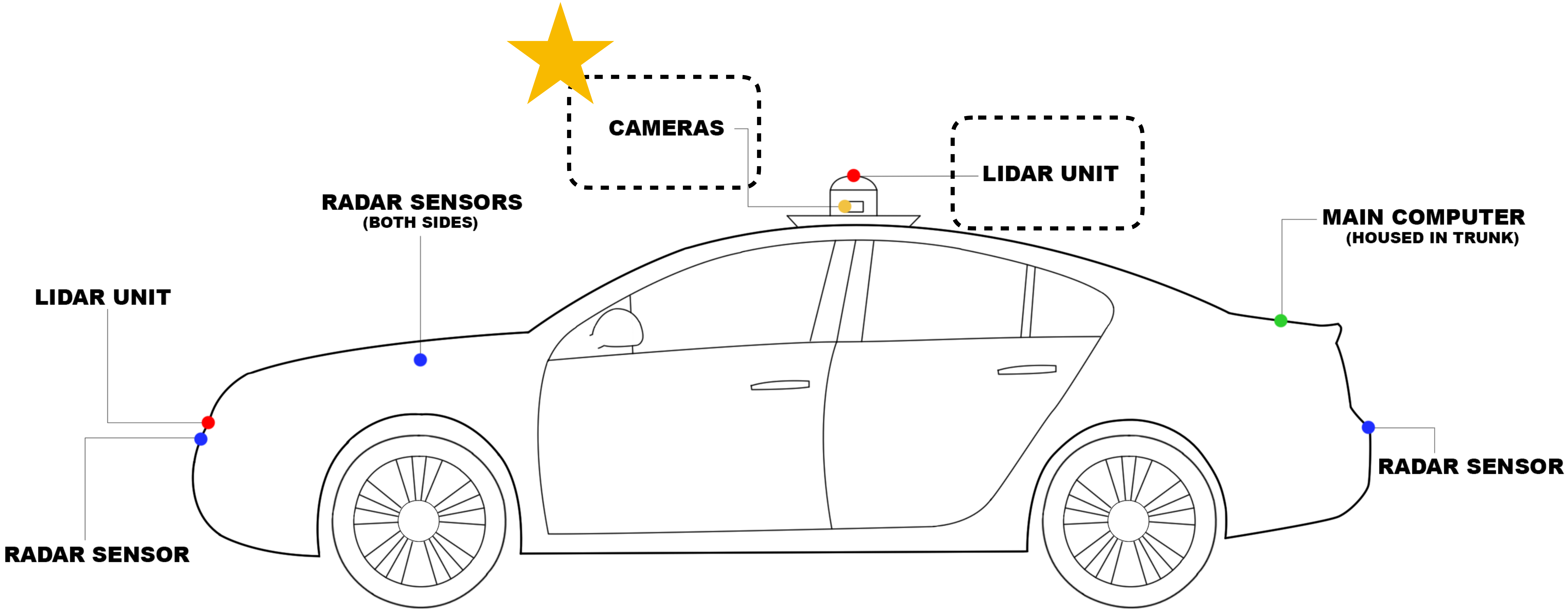
A Neural Network Labels Camera Data



Inception Network - Google

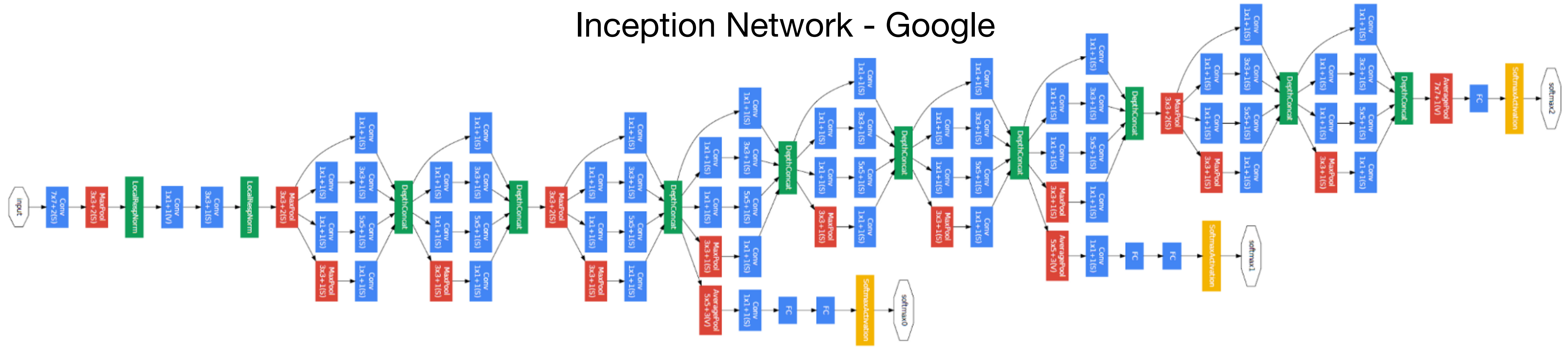


Label
e.g. pedestrian



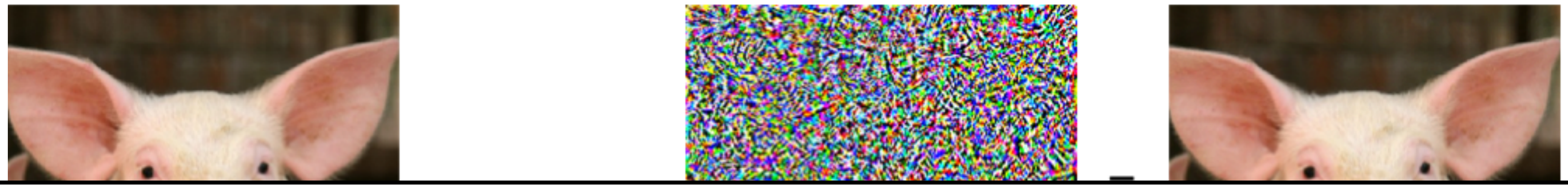
Problem: Neural Networks are Brittle

Inception Network - Google



Label
e.g. pedestrian

"pig" "airliner"



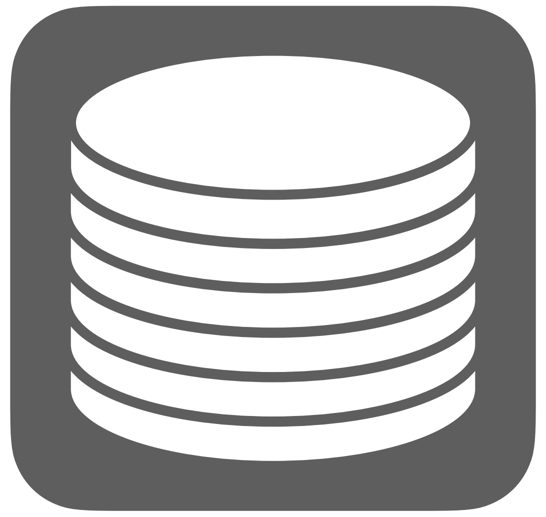
For self-driving, and other mission-critical, safety-critical applications, these mistakes have CONSEQUENCES.

K. Eykholt et al. "Robust Physical-World Attacks on Deep Learning Visual Classification."

Monitor Opaque Subsystems for Reasonableness

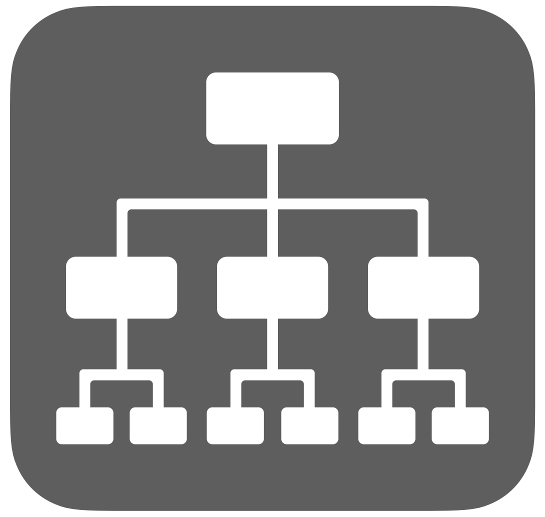
Label
e.g. pedestrian

Opaque
Mechanism



Commonsense
Knowledge Base

+



Flexible
Representation

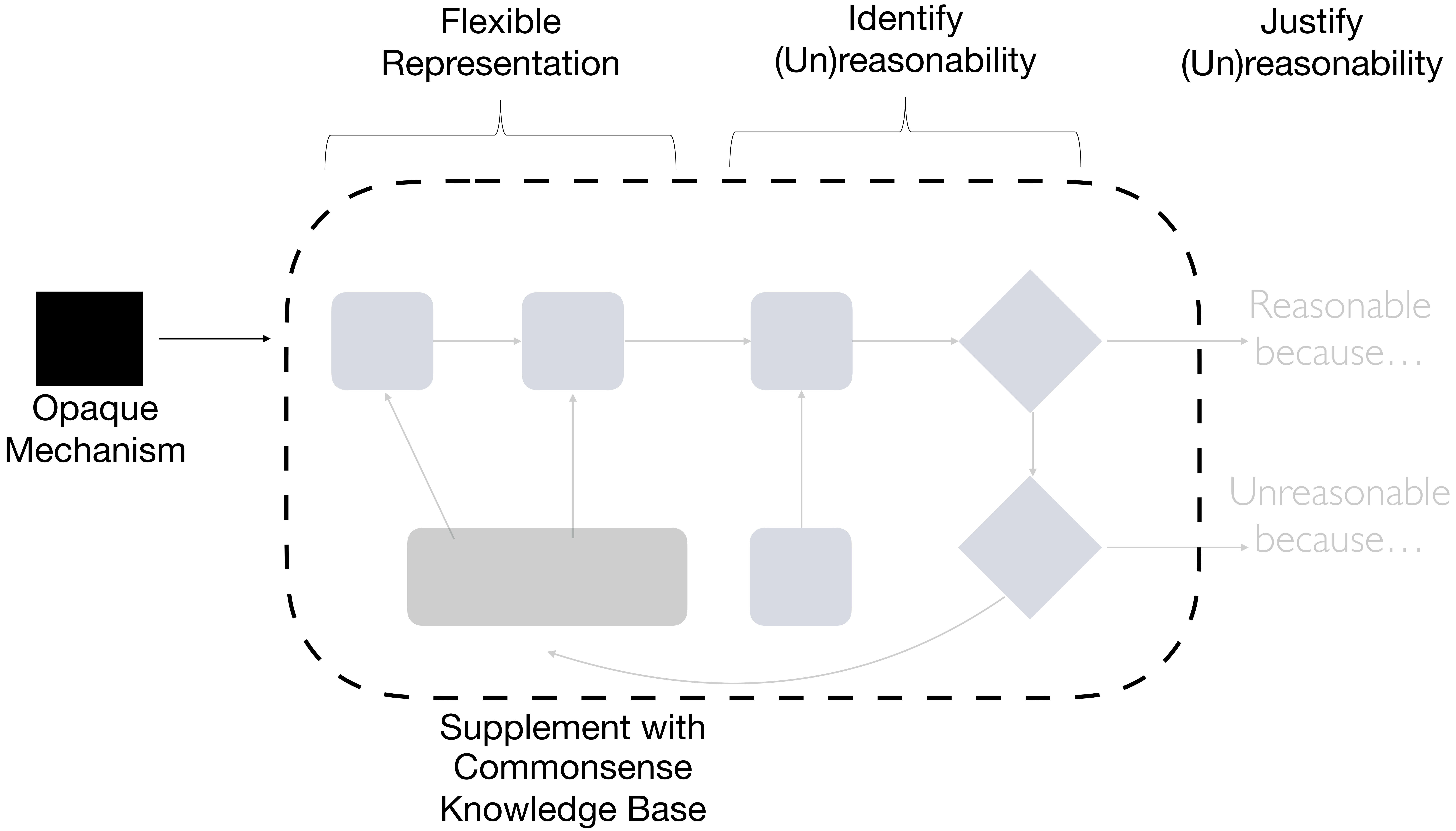
+

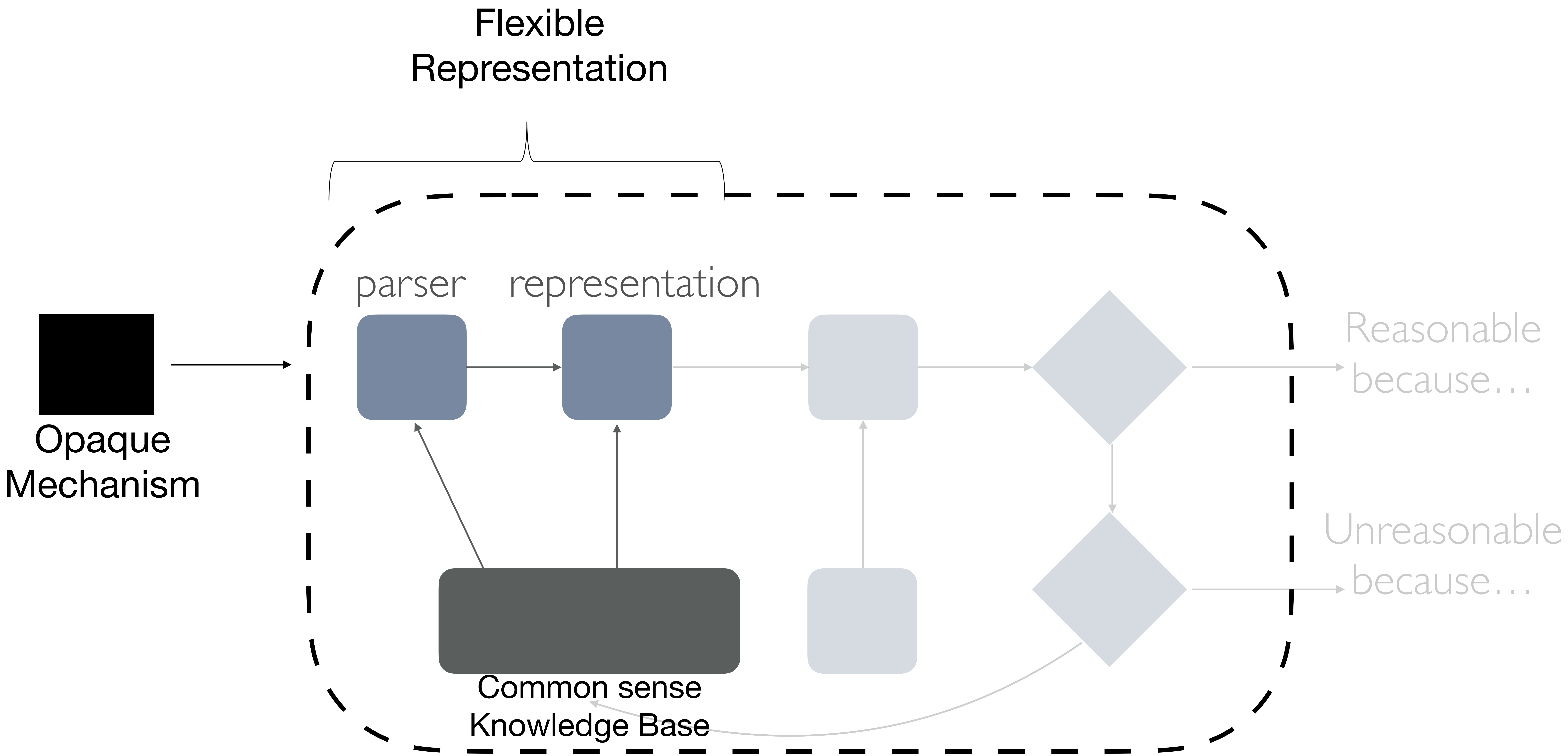
Identify
(Un)reasonability

+

Justify
(Un)reasonability

1. Judgement of reasonableness
2. Justification of reasonableness


Flexible Representation
Identify (Un)reasonability
Justify (Un)reasonability
Opaque Mechanism
Reasonable because…
Unreasonable because…
Supplement with Commonsense Knowledge Base


Flexible
Representation
parser
representation
Reasonable
because…
Unreasonable
because…
Opaque
Mechanism
Common sense
Knowledge Base

Primitive Representations

Encode Understanding

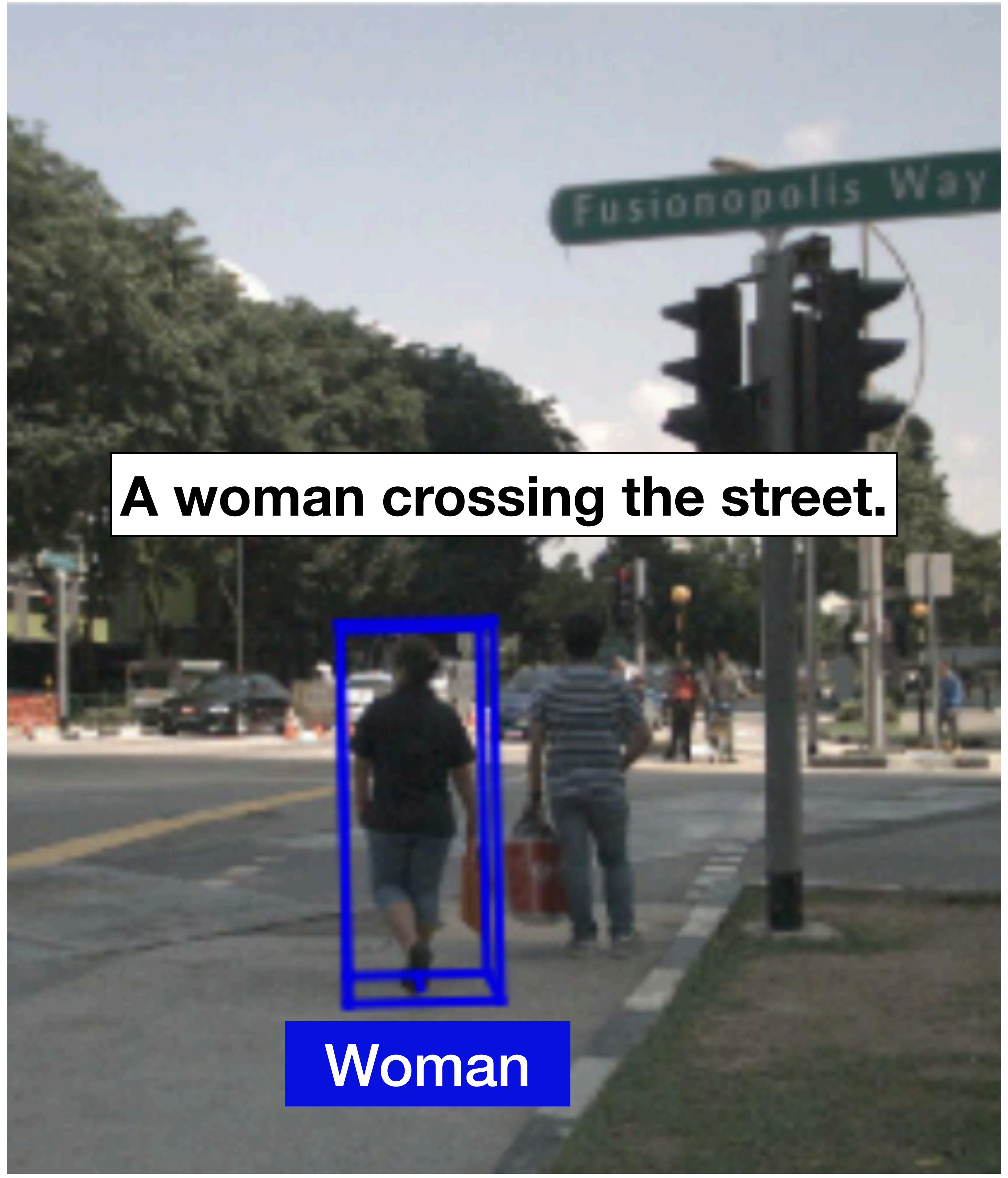
Conceptual Dependency Theory (CD), Schank 1975

11 primitives to account for *most* actions:

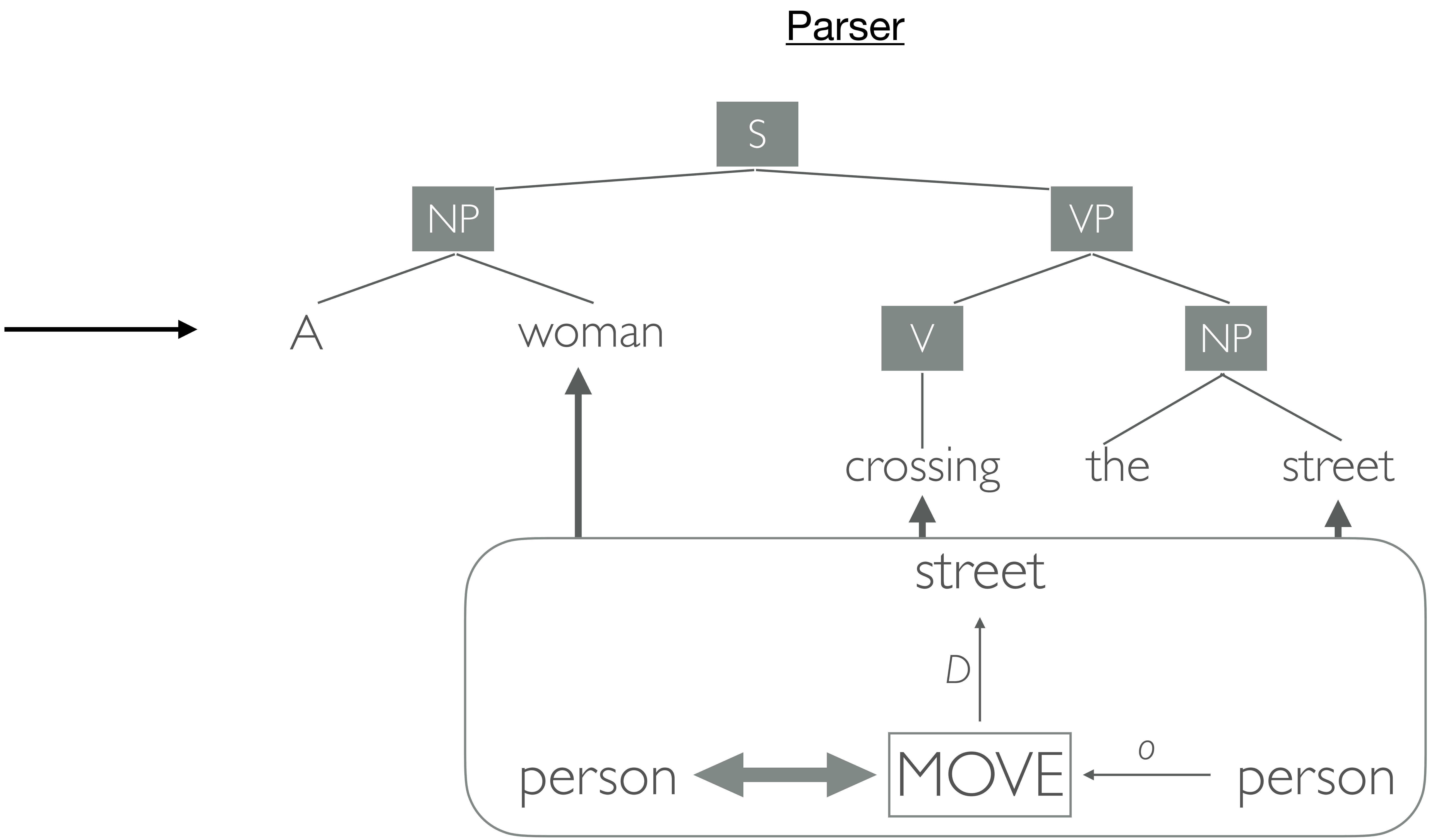
ATRANS
ATTEND
INGEST
EXPEL
GRASP
MBUILD
MTRANS
MOVE
PROPEL
PTRANS
SPEAK

5 for physical actions
Extended to vehicle primitives

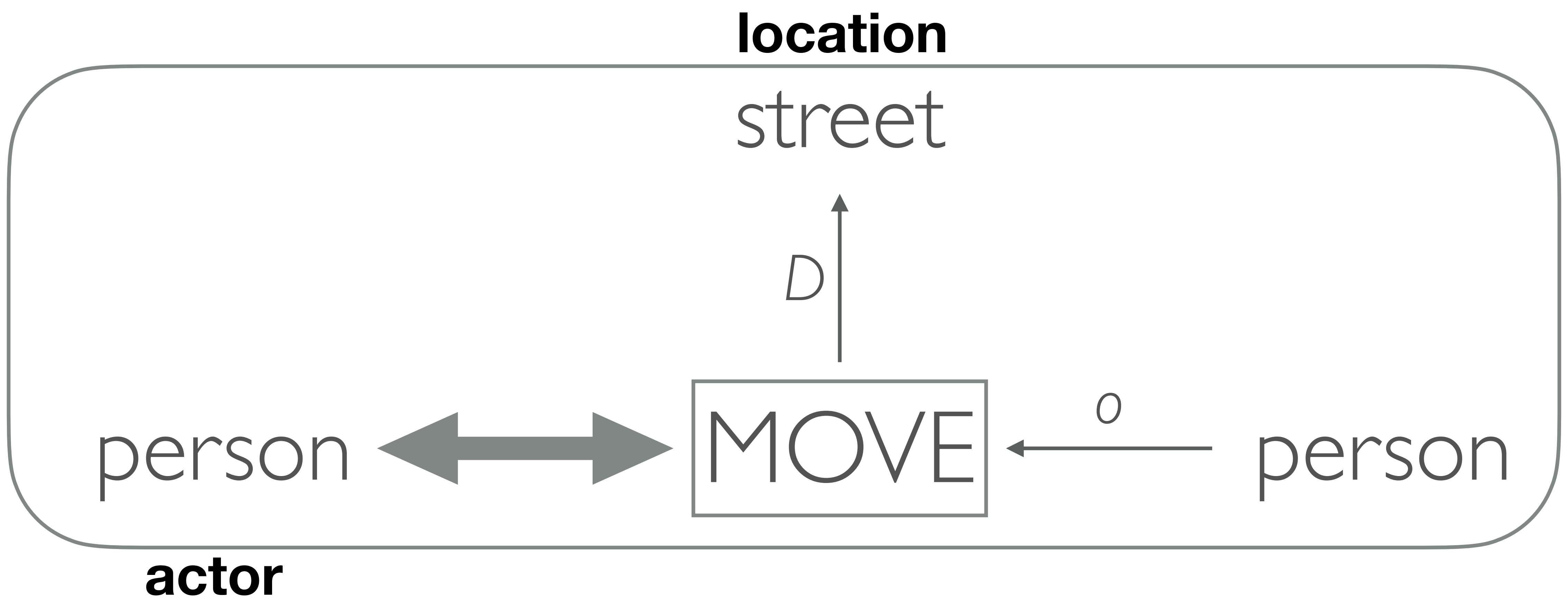
Parse Natural Language into Representation



Data from Nuscenes



Representations with Implicit Rules



A perceived frame is REASONABLE

$$\big((x_1, p_1, y_1), \textbf{isA, REASONABLE}\big) \wedge$$
$$\big((x_2, p_2, y_2), \textbf{isA, REASONABLE}\big) \wedge$$
$$\ldots \wedge$$
$$\big((x_n, p_n, y_n), \textbf{isA, REASONABLE}\big)$$

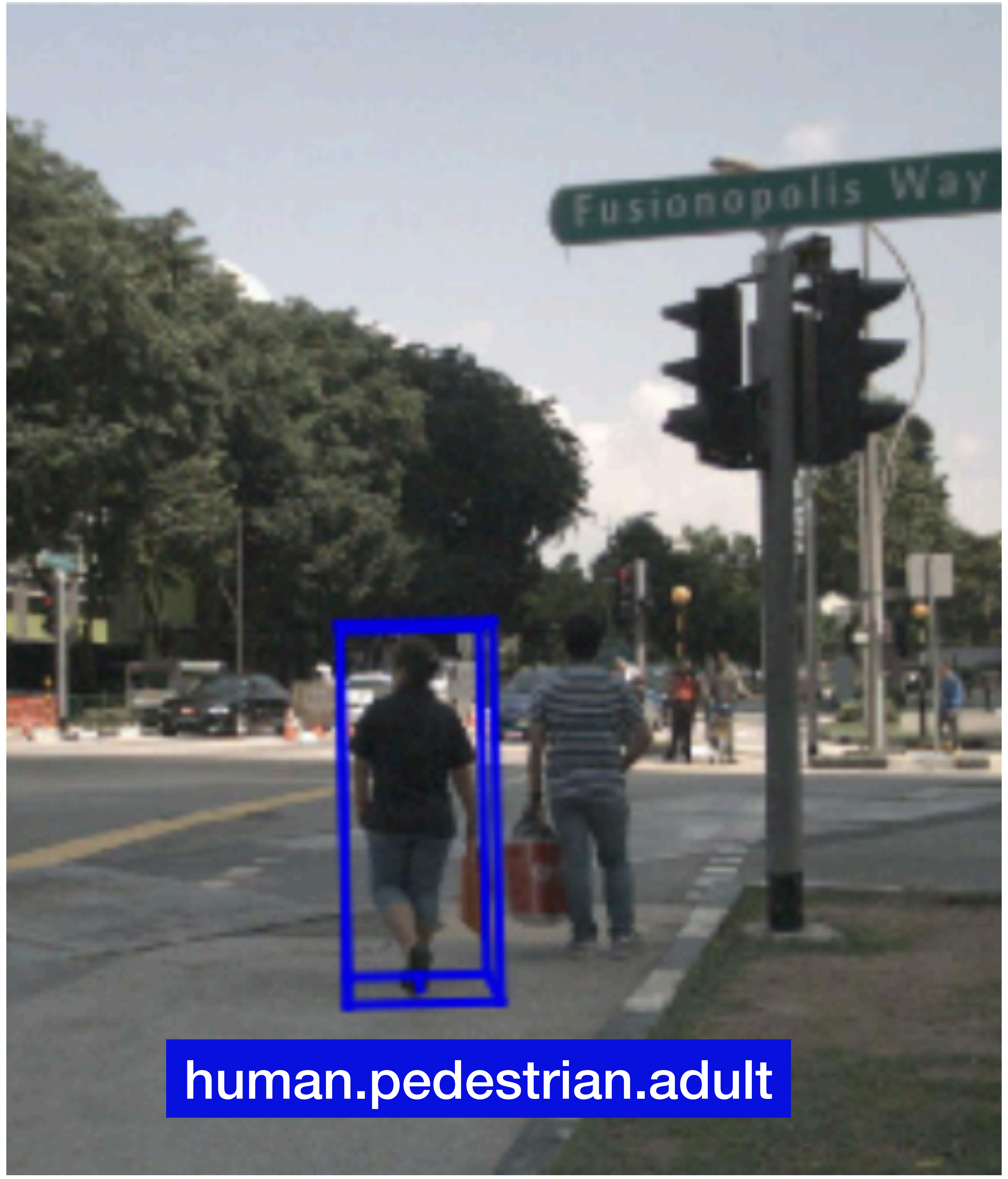
<u>Move Primitive Reasonability</u>

$$(x, hasProperty, animate) \wedge (x, locatedNear, y) \Rightarrow \big((\underset{\textbf{actor}}{x}, MOVE, \underset{\textbf{location}}{y}) \textbf{ isA, REASONABLE}\big)$$

Reasonableness Monitoring on Real Data

NuScenes

{'token': '70aecbe9b64f4722ab3c230391a3beb8',
 'sample_token': 'cd21dbfc3bd749c7b10a5c42562e0c42',
 'instance_token': '6dd2cbf4c24b4caeb625035869bca7b5',
 'visibility_token': '4',
 'attribute_tokens': ['4d8821270b4a47e3a8a300cbec48188e'],
 'translation': [373.214, 1130.48, 1.25],
 'size': [0.621, 0.669, 1.642],
 'rotation': [0.9831098797903927, 0.0, 0.0, -0.18301629506281616],
 'prev': 'a1721876c0944cdd92ebc3c75d55d693',
 'next': '1e8e35d365a441a18dd5503a0ee1c208',
 'num_lidar_pts': 5,
 'num_radar_pts': 0,
 'category_name': 'human.pedestrian.adult'}



Data from NuScenes

Commonsense is Unorganized

ConceptNet

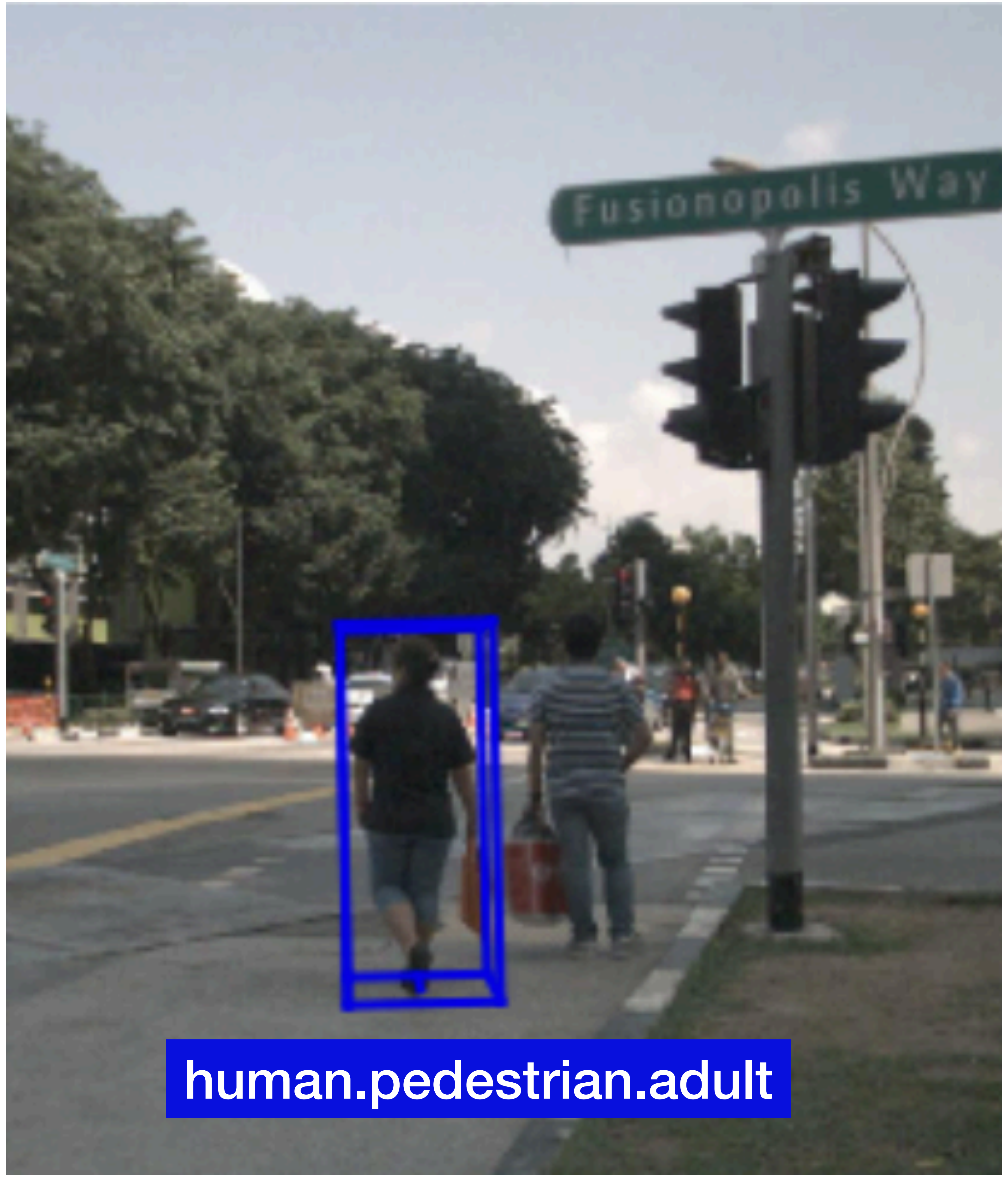
adult is a type of...

- en animal (n, wn)
- en person (n, wn)
- en animal (n)

adult is capable of...

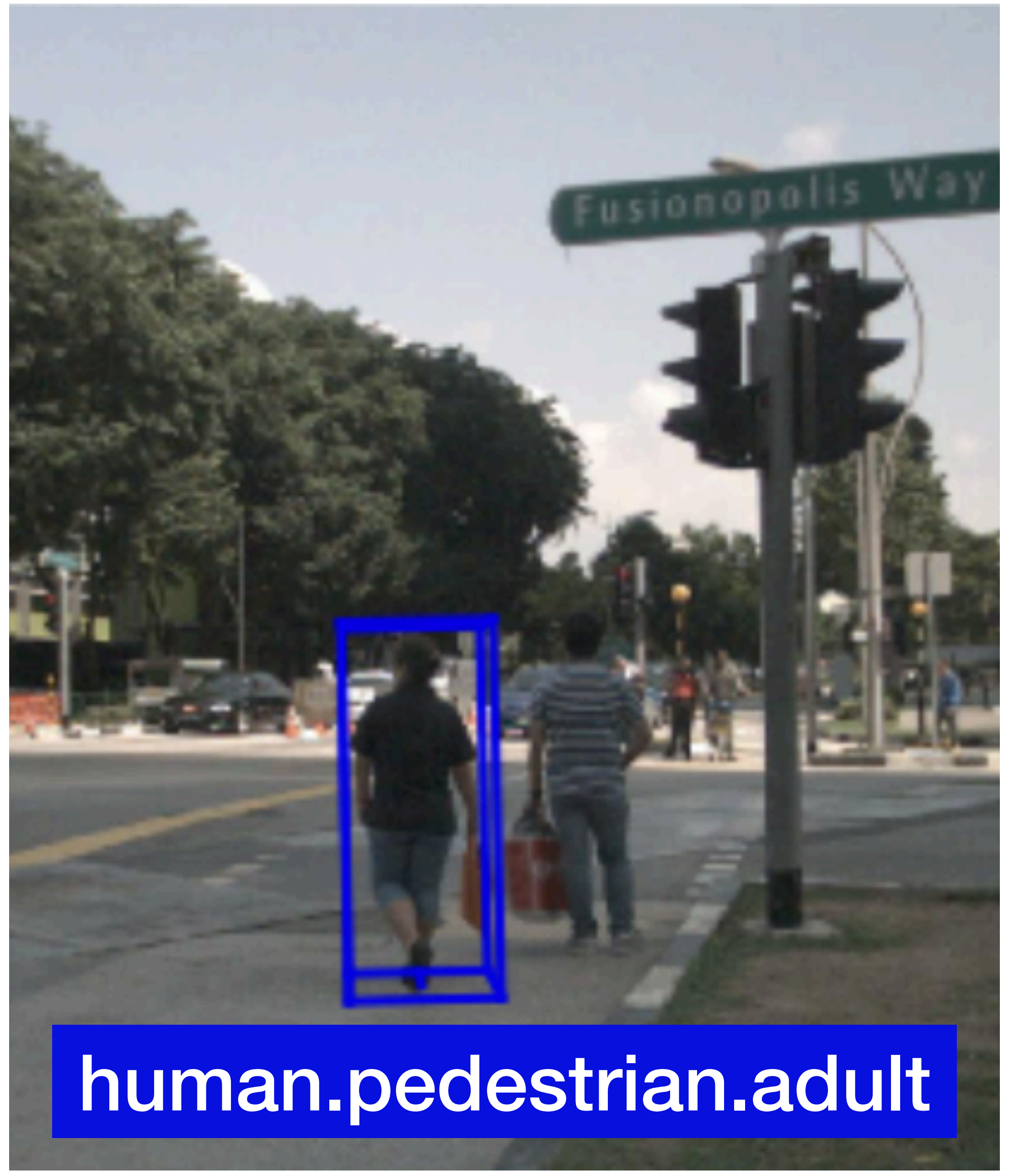
- en help a child
- en dress herself
- en sign a contract
- en drink beer
- en work
- en act like a child
- en dress himself
- en drive a car
- en drive a train
- en explain the rules to a child

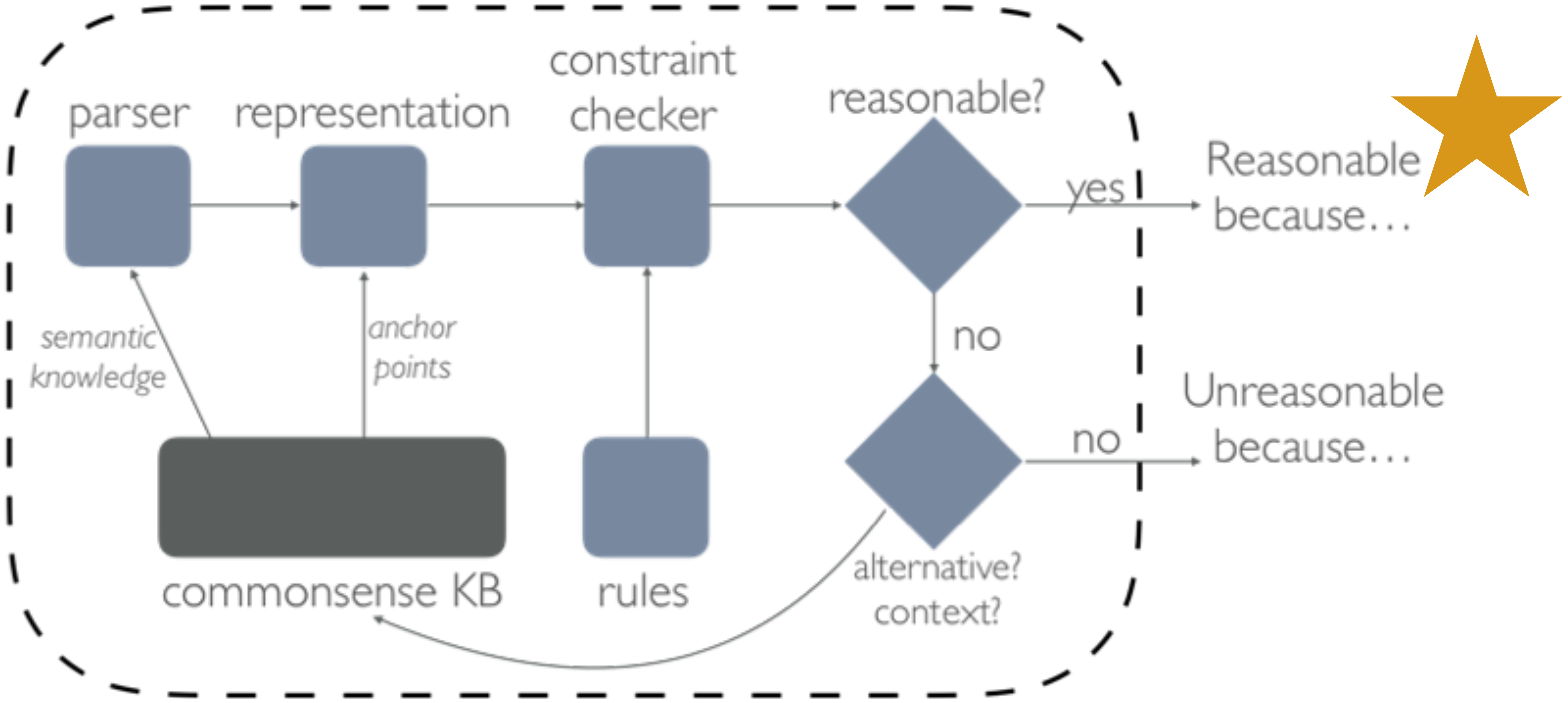
```
('adult, 'typeOf, 'animal)
('adult, 'isA, 'bigger than a child')
…
```



Data from NuScenes

Monitor Outputs a Judgement and Justification





This perception is reasonable. An adult is typically a large person. They are usually located walking on the street. Its approximate dimensions of [0.621, 0.669, 1.642] is approximately the correct size in meters.

Agenda

Motivate problem: Complex systems are prone to failure

Local sanity checks for vehicle perception

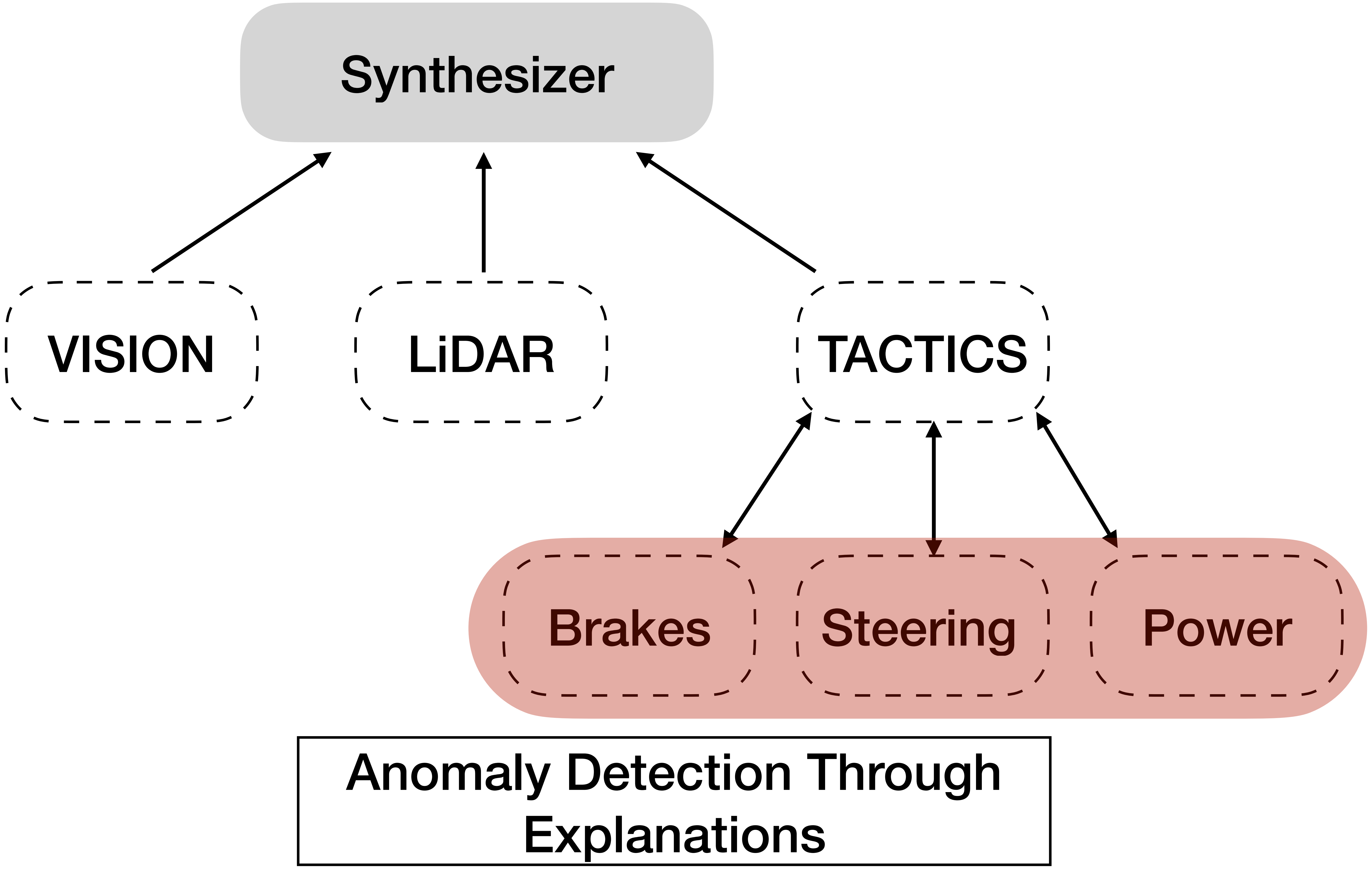
<u>Explanations as an internal debugging language</u>

Ongoing Work: Testing Autonomous Vehicles by Augmenting Datasets

Reconciling Internal Disagreements

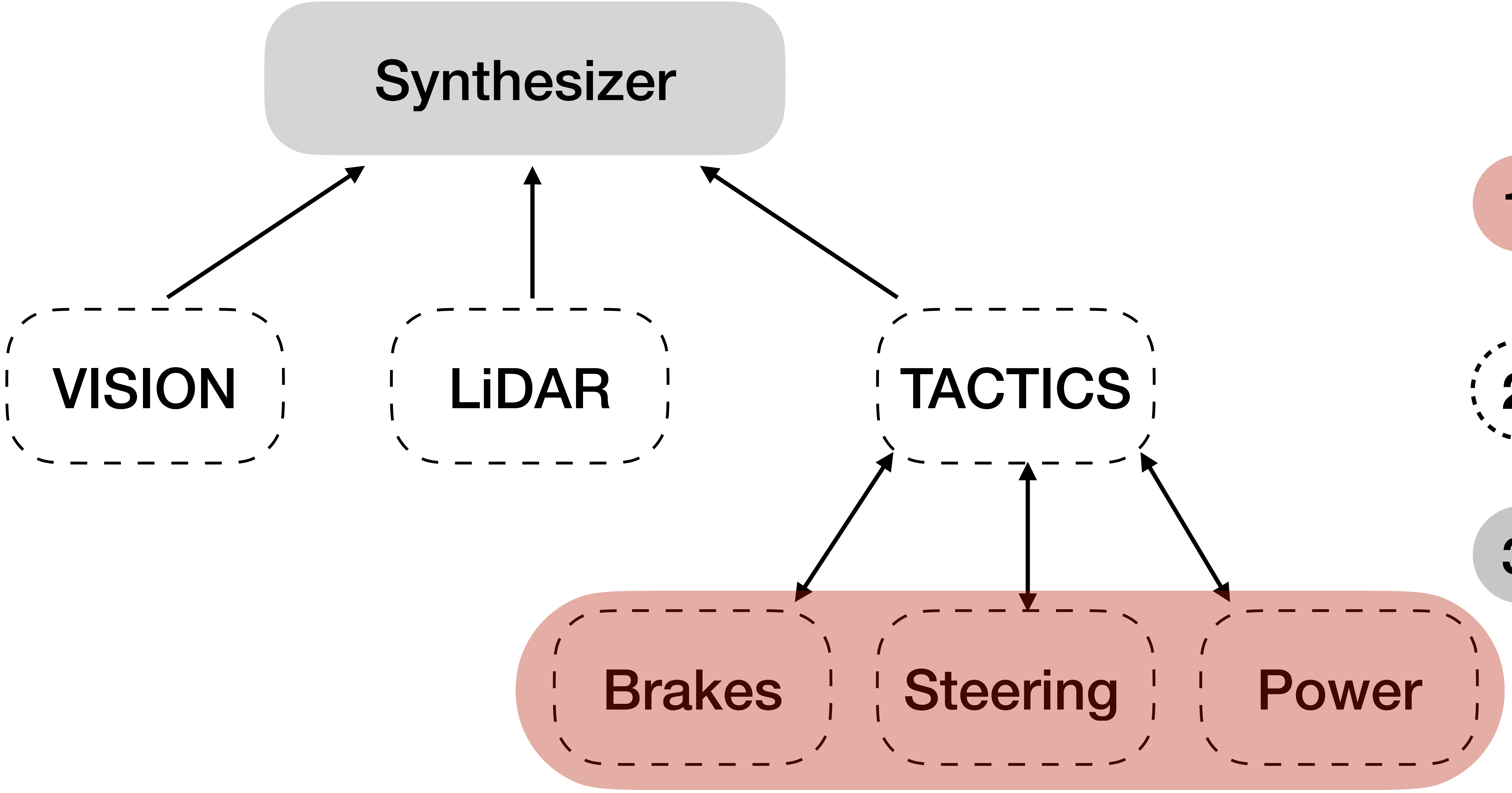
With an Organizational Architecture

- Monitored subsystems combine into a system architecture.
- Explanation synthesizer to deal with *inconsistencies.*
    - Argument tree.
    - Queried for support or counterfactuals.



Anomaly Detection through Explanations

Reasoning in Three Steps



1. Generate Symbolic Qualitative Descriptions for each committee.
2. Input qualitative descriptions into local “reasonableness” monitors.
3. Use a synthesizer to reconcile inconsistencies between monitors.

3. Use a synthesizer to reconcile inconsistencies between monitors.

Synthesizer + Priority Hierarchy → Abstract Goals

- Explanation synthesizer to deal with *inconsistencies.*
  - Argument tree.
  - Queried for support or counterfactuals.

1. Passenger Safety
2. Passenger Perceived Safety
3. Passenger Comfort
4. Efficiency (e.g. Route efficiency)

→ A passenger is safe if:

- The vehicle proceeds at the same speed and direction.
- The vehicle avoids threatening objects.

3. Use a synthesizer to reconcile inconsistencies between monitors.

$$
\begin{aligned}
&(\forall s, t \in STATE, v \in VELOCITY \\
&\quad \big((self, moving, v), \textbf{state}, s\big) \wedge \\
&\quad (t, \textbf{isSuccesorState}, s) \wedge \\
&\quad \big((self, moving, v), \textbf{state}, t\big) \wedge \\
&\quad (\nexists x \in OBJECTS \textbf{ s.t.} \\
&\quad \big((x, isA, threat), \textbf{state}, s\big) \vee \\
&\quad \big((x, isA, threat), \textbf{state}, t\big))) \\
&\quad \Rightarrow \big(\textbf{passenger, hasProperty, safe}\big)
\end{aligned}
$$

A passenger is safe if:

- The vehicle proceeds at the same speed and direction.
- The vehicle avoids threatening objects.

$$
\begin{aligned}
&(\forall s \in STATE, x \in OBJECT, v \in VELOCITY \\
&\quad \big((x, moving, v), \textbf{state}, s\big) \wedge \\
&\quad \big((x, locatedNear, self), \textbf{state}, s\big) \wedge \\
&\quad \big((x, isA, large\_object), \textbf{state}, s\big) \\
&\quad \Leftrightarrow \big((x, isA, threat), \textbf{state}, s\big))
\end{aligned}
$$

3. Use a synthesizer to reconcile inconsistencies between monitors.

```
(monitor, judgement, unreasonable)
(input, isType, labels)
(all_labels, inconsistent, negRel)
(isA, hasProperty, negRel)
…
(all_labels, notProperty, nearMiss)
(all_labels, locatedAt, consistent)
(monitor, recommend, discount)
```

```
(monitor, judgement, reasonable)
(input, isType, sensor)
…
(input_data[4], hasSize, large)
(input_data[4], IsA, large_object)
(input_data[4], moving, True)
(input_data[4], hasProperty, avoid)
…
(monitor, recommend, avoid)
```

!

```
(monitor, judgement, reasonable)
(input, isType, history)
(input_data, moving, True)
(input_data, direction, forward)
(input_data, speed, fast)
(input_data, consistent, True)
(monitor, recommend, proceed)
```

Abstract Goal Tree

```
'passenger is safe',
AND(
  'safe transitions',
  NOT('threatening objects')
```

!

The best option is to veer and slow down. The vehicle is traveling too fast to suddenly stop. The vision system is inconsistent, but the lidar system has provided a reasonable and strong claim to avoid the object moving across the street.

Uber Example in Simulation

L. H. Gilpin, V. Penubarthi and L. Kagal, "Explaining Multimodal Errors in Autonomous Vehicles," *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, 2021, pp. 1-10, doi: 10.1109/DSAA53316.2021.9564178.

Agenda

Motivate problem: Complex systems are prone to failure

Local sanity checks for vehicle perception

Explanations as an internal debugging language

<u>Ongoing Work: Testing Autonomous Vehicles by Augmenting Datasets</u>

Vision: Real World Adversarial Examples

“Realistic” Adversarial examples

L. H. Gilpin, A. Amos-Binks, "Close Syntax but Far Semantics: A Risk Management Problem for Autonomous Vehicles.” *The AAAI Fall Symposium on Cognitive Systems for Anticipatory Thinking.*

Vision: Real World Adversarial Examples

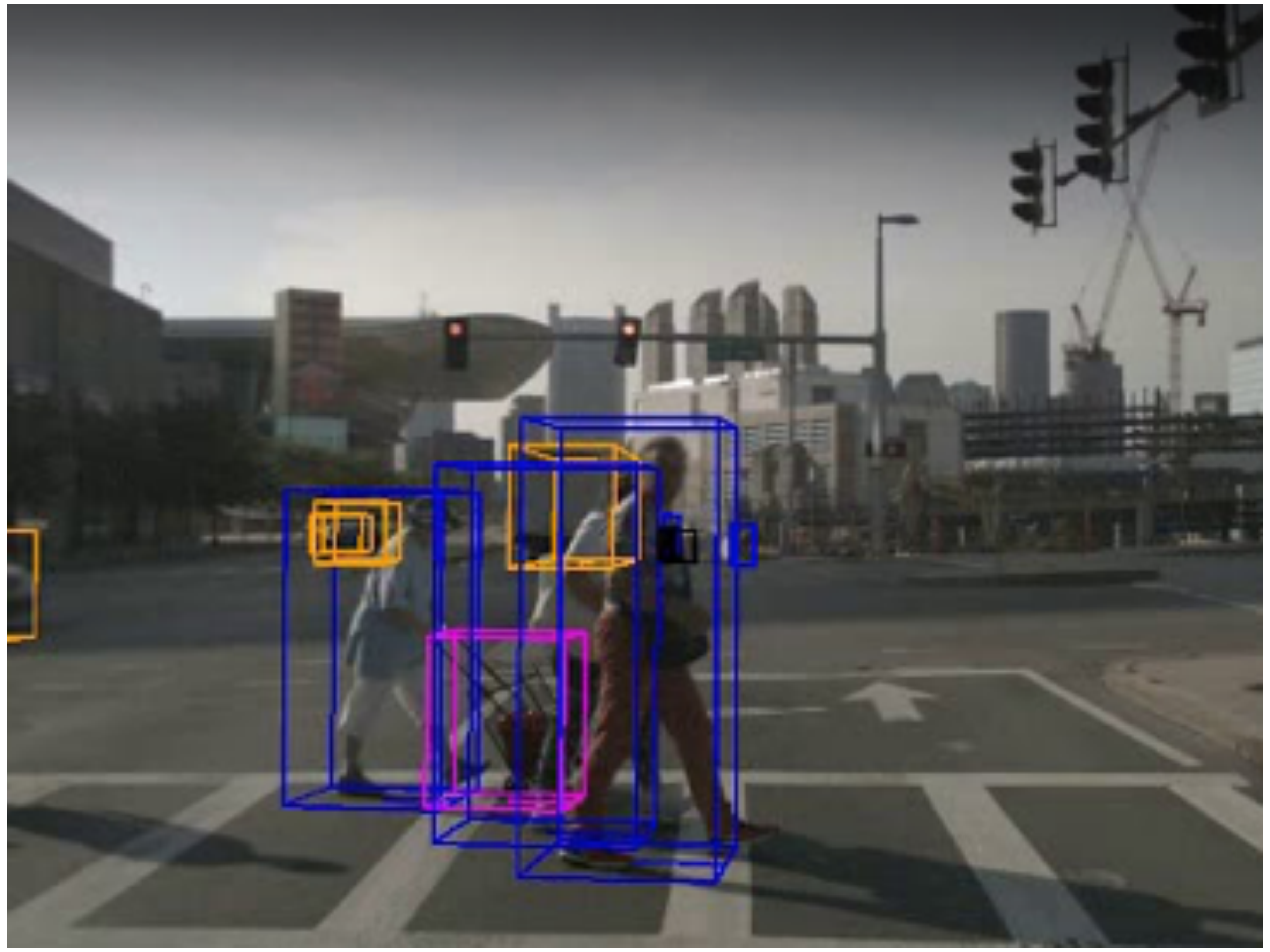
Anticipatory Thinking Layer for Error Detection

“Realistic” Adversarial examples

The traffic lights are on top of the truck. The lights are not illuminated. The lights are moving at the same rate as the truck, therefore this is not a “regular” traffic light for slowing down and stopping at.

Lack of Data and Challenges for AVs

- Existing Challenges
  - Targeted as optimizing a mission or trajectory and not safety.
  - Data is hand-curated.
- Failure data is not available
  - Unethical to get it (cannot just drive into bad situations).
  - Want the data to be realistic (usually difficult in simulation).



Data from NuScenes

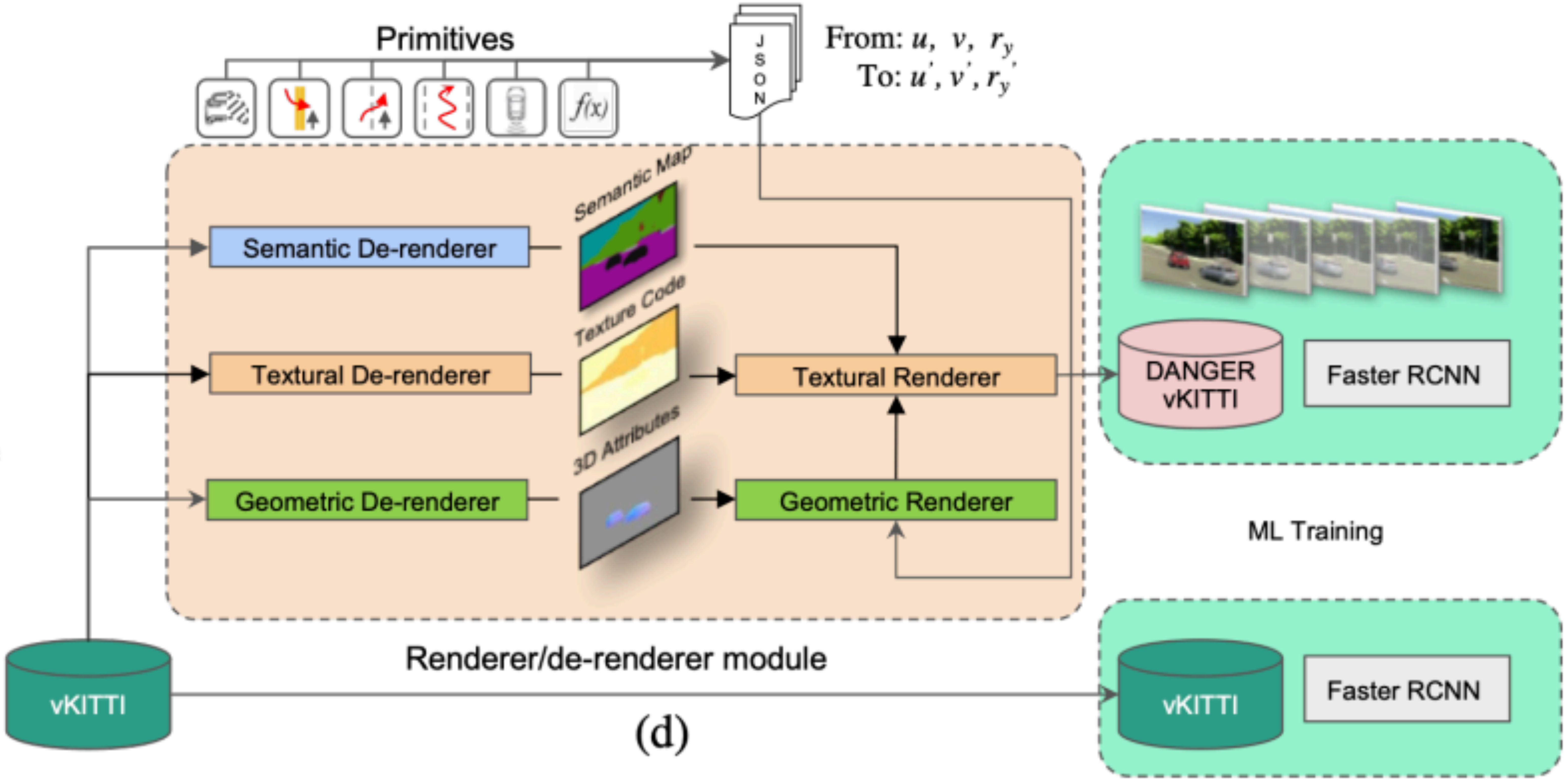
Approach: Content Generation

Anticipatory Thinking Layer for Error Detection

S. Xu, L. Mi and L.H. Gilpin. “A Framework for Generating Dangerous Scenes for Testing Robustness.” Under Review. 2023.

Approach: Content Generation

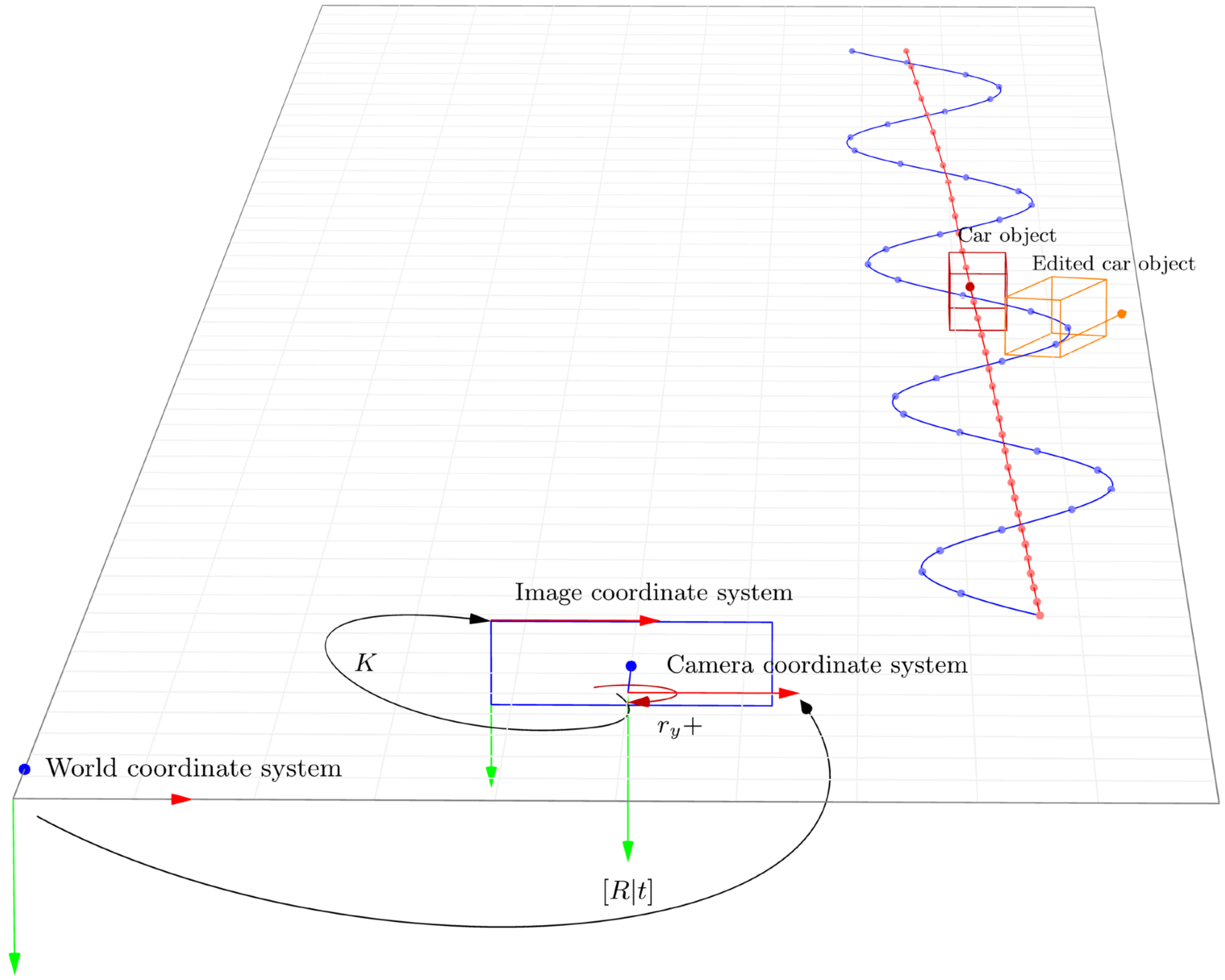
Anticipatory Thinking Laver for Error Detection



S. Xu, L. Mi and L.H. Gilpin. "A Framework for Generating Dangerous Scenes for Testing Robustness." Under Review. 2023.

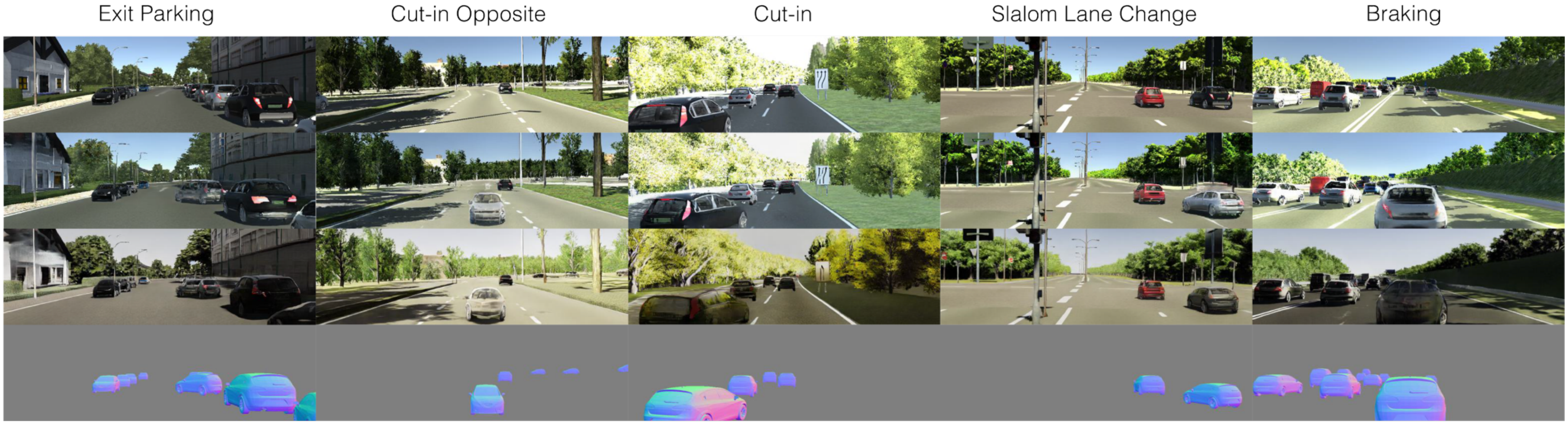
Approach: Content Generation

Anticipatory Thinking Layer for Error Detection



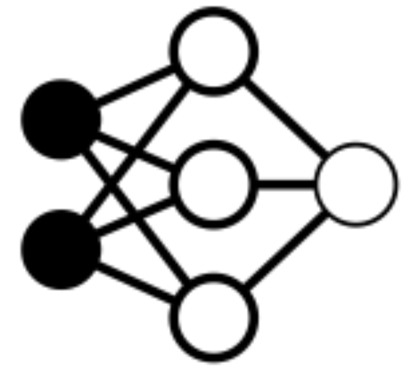
S. Xu, L. Mi and L.H. Gilpin. "A Framework for Generating Dangerous Scenes for Testing Robustness." Under Review. 2023.

Behaviors that are Inherently Explainable

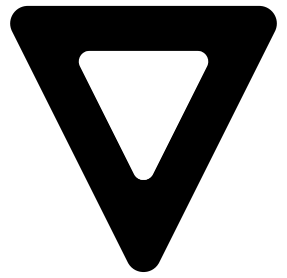


S. Xu, L. Mi and L.H. Gilpin. “A Framework for Generating Dangerous Scenes for Testing Robustness.” Under Review. 2023.

Contributions


Opaque Systems


Autonomous Systems


Error Detection

Motivate problem: Complex systems are prone to failure

Local sanity checks for vehicle perception

Explanations as an internal debugging language

Ongoing Work: Testing Autonomous Vehicles by Augmenting Datasets