SELF-EXPLANATION AND SELF-DRIVING

no explanation
at all

communication to
non-expert

explanation to
human expert

Leilani H. Gilpin
MIT

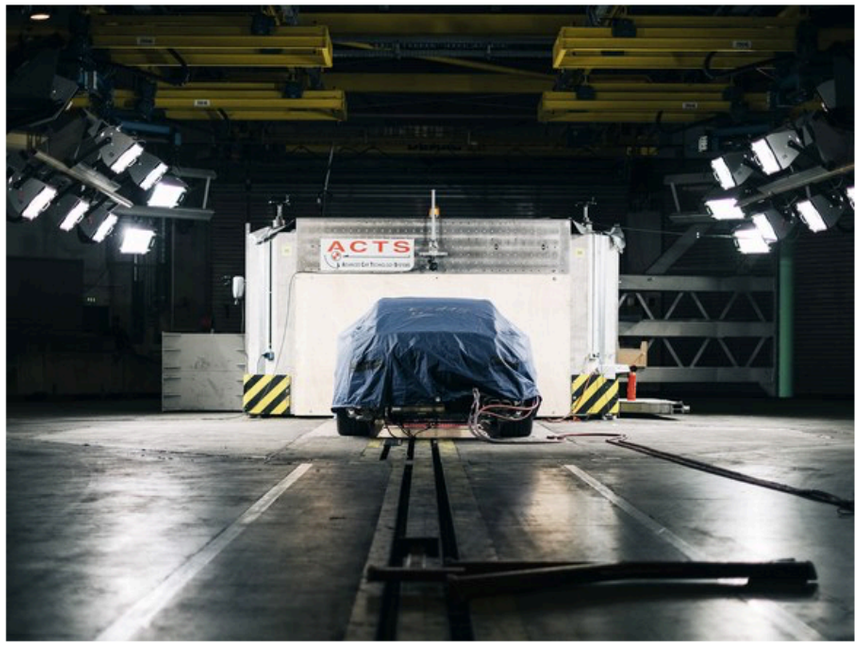
WHO'S AT FAULT?

WHO'S AT FAULT?

Victim of self-driving Uber accident could be to blame, expert says

USA TODAY NETWORK Ryan Randazzo, The Arizona Republic Published 4:20 p.m. ET March 23, 2018

THE UBER CRASH WON'T BE THE LAST SHOCKING SELF-DRIVING DEATH



Tesla said autopilot was activated during a fatal Model X crash last week in California.

WHAT WENT WRONG?

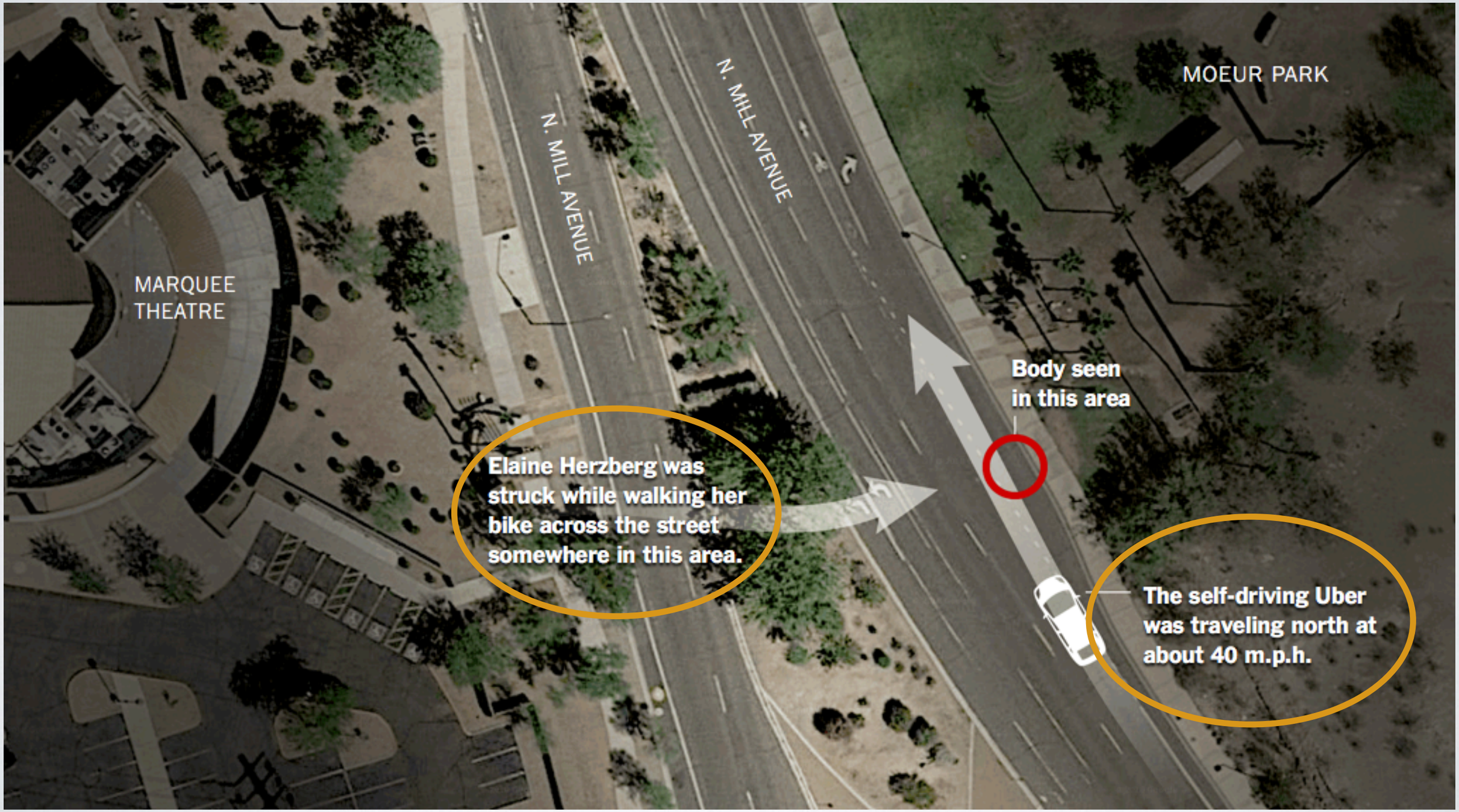


photo courtesy of the New York Times

WHAT WENT WRONG?

- Who's at fault?
  - Human (safety driver) error
  - Pedestrian error
  - Vehicle error

ABC-15, via Associated Press

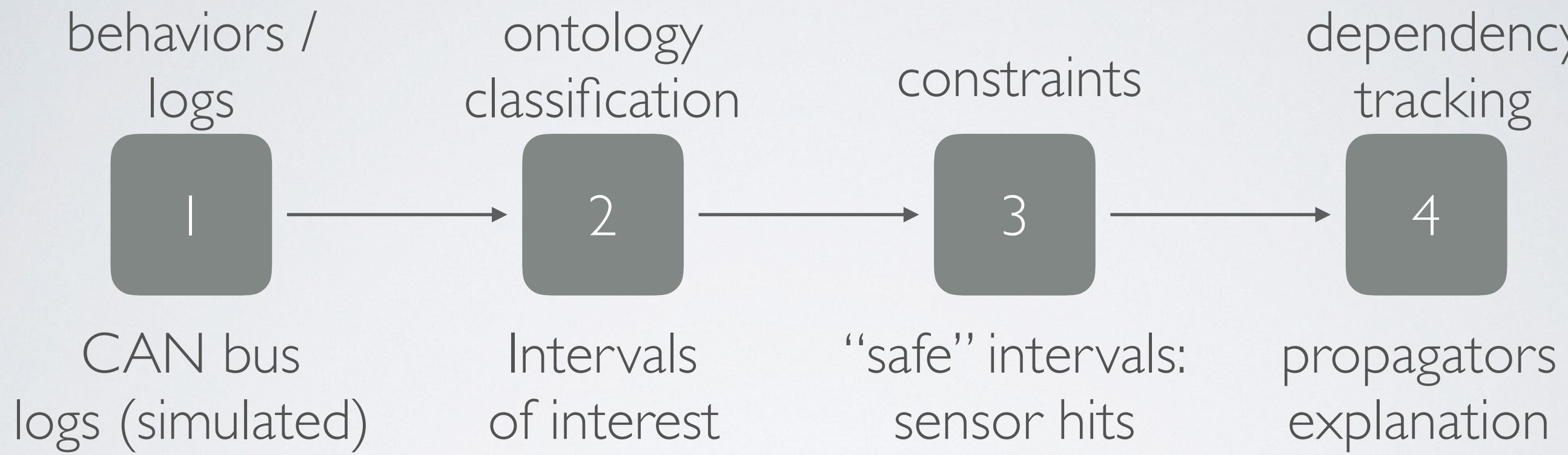
WHAT WENT WRONG?

- Unavoidable - No way to detect the pedestrian with enough time to swerve out of the way.

- *Possibly* avoidable - Did sensors detect the pedestrian with *enough* time to swerve out of the way?

- Internal errors - Sensors, perception mechanisms, etc. not working as expected?

EX-POST-FACTO EXPLANATION



Coherent Story

STORY-TELLING FOR SAFETY

- For autonomous machines to be safe they need to be able to **explain** themselves

- For autonomous vehicles to be intelligent, they need to **understand** the action and behavior or their underlying parts

VEHICLE STORIES

- Autonomous agents must be able to provide explanations for the following reasons:
  - in order to be **audited**
  - to provide an understandable and coherent story which **justifies** their actions
  - able to be **challenged** in an adversary proceeding
  - if the explanation is inadequate or inappropriate, the agent should either corrected or disabled.

3 MAIN AREAS

- **Explanations**
  - Machinery / software
  - Machine perception
  - Security
    - How can we strengthen vehicle security?
- **Accountability**
  - What are likely [autonomous] vehicle scenarios?
  - How will pedestrians react?
  - How can we use technology to ensure vehicles can provide evidence?

OUR RESEARCH

- Adapted a game simulation to output a "CAN Bus" log
- Edge detection : **When** did the operator apply brakes
- Interval analysis: **How** do intervals relate
- Tell a story of **what** happened
- Begin to tell a **why** story

L.H. Gilpin and B.Z. Yuan. "Getting Up to Speed on Vehicle Intelligence." *The AAAI 2017 Spring Symposium on Science of Intelligence: Computational Principles of Natural and Artificial Intelligence.*

NEED FOR OPEN SOFTWARE

- Availability of code/data to be **evaluated**
- Software available for **accountable** development
  - Simulation
  - Error detection and reasoning

OUR DATA

- Controller Area Network log (CAN Bus)

- Easy to hack
  - simple schema
  - schema: time stamp, CAN bus code, extra information
  - connects to all aspects of a car

- Standard

```
93.79 B3 -24.94 1.15
93.79 120 13 04 50
93.79 244 0.00
93.795 22 0.00
93.795 23 -0.80
93.795 25 0.00
93.795 30 0.00
93.795 B1 81.83 -5.69
93.795 B3 24.24 -56.52
93.795 120 13 04 50
93.795 244 0.00
93.8 22 0.00
93.8 23 0.89
93.8 25 0.00
93.8 30 0.00
93.8 B1 -46.06 -88.97
93.8 B3 21.87 6.62
93.8 120 13 04 50
93.8 244 0.00
93.805 22 0.00
93.805 23 -0.08
93.805 25 0.00
93.805 30 0.00
93.805 B1 -77.20 -5.41
93.805 B3 18.62 -19.38
93.805 120 13 04 50
93.805 244 0.00
93.81 22 0.00
93.81 23 0.21
```

OUR DATA - UP CLOSE

CAN bus code

B1 - front wheels
B3 - rear wheels
120 - drive mode

```
93.795 B1 81.83 81.83
93.795 B3 24.24 24.24
93.795 120 13 04 50
```

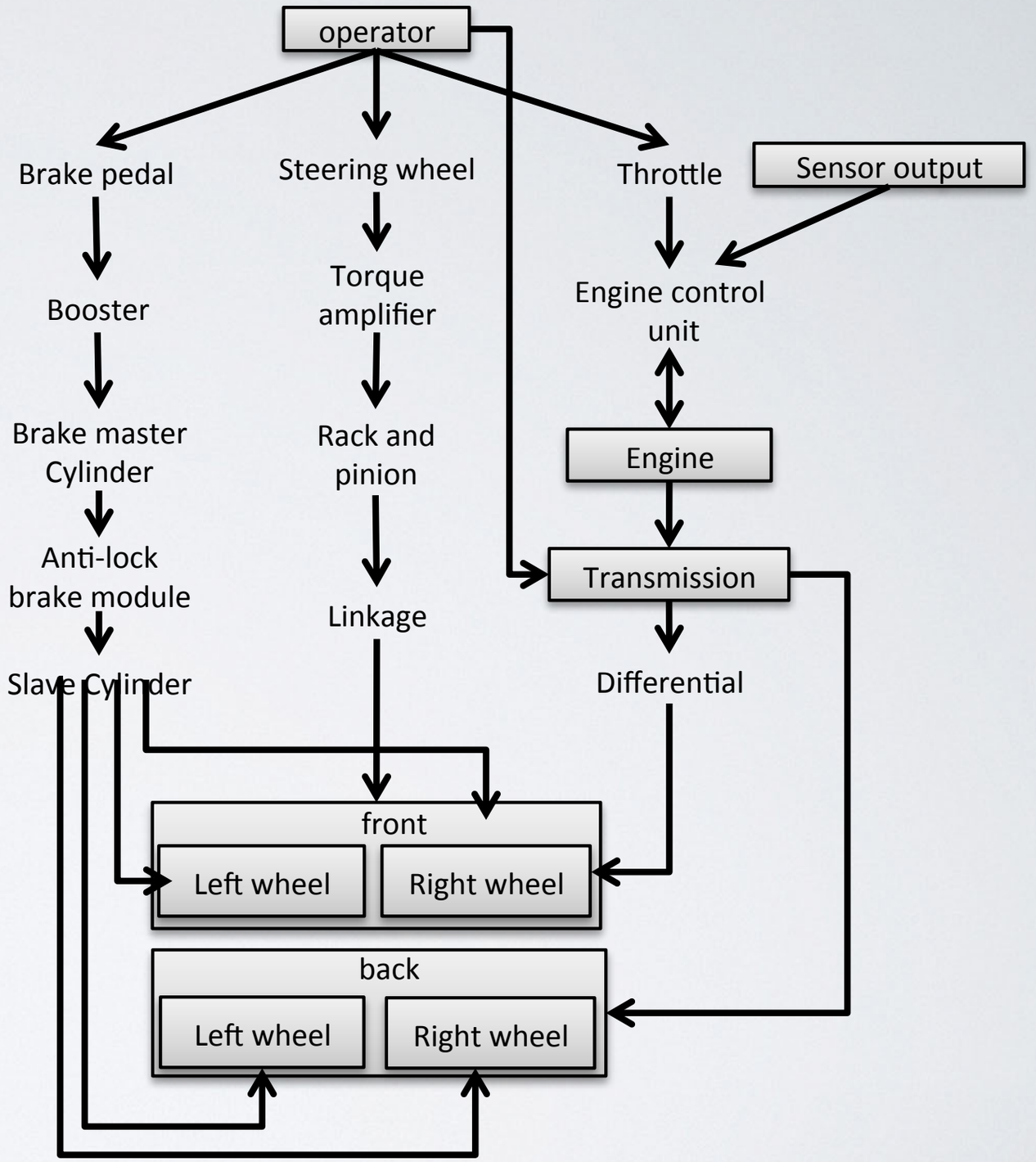
time stamp
in seconds

paramet

right, left wheel rotation
(in km/hr)
13, 50 - Drive
04 - powered

MODELING

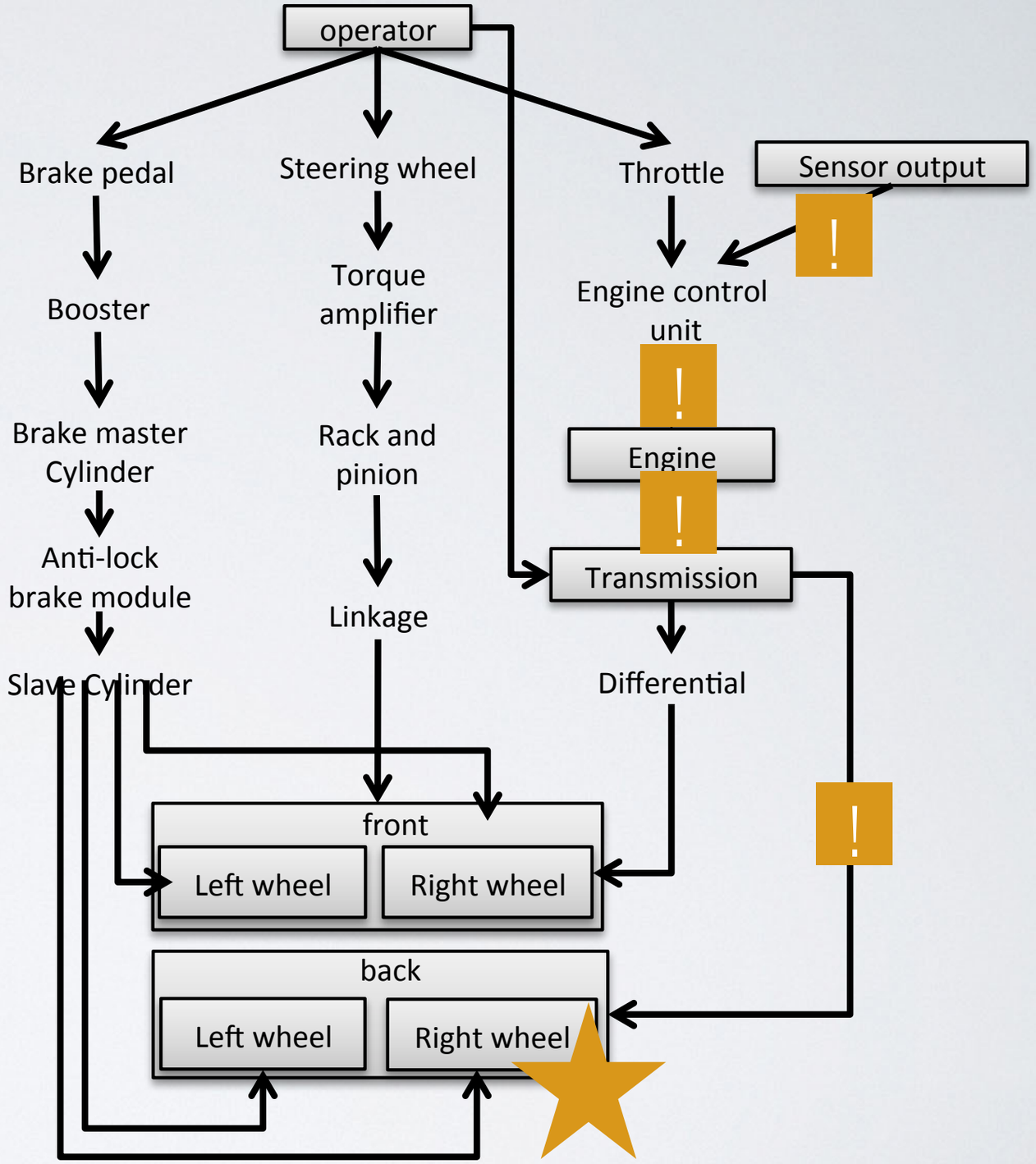
Mechanical systems

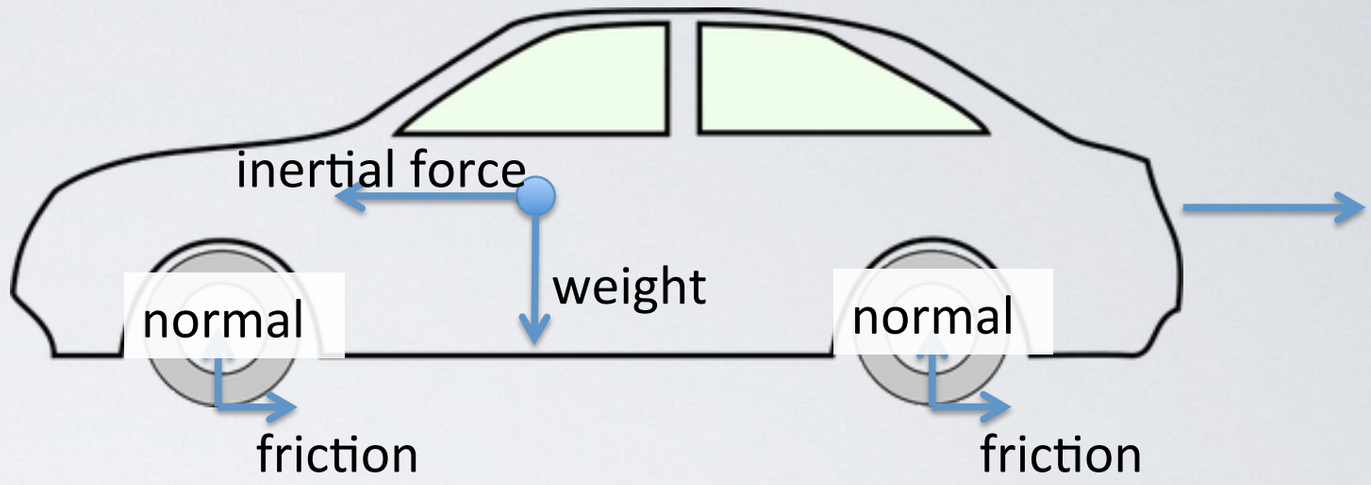


MODELING

Mechanical systems

Tire pressure sensor is anomalous given current state (snow, chains, engine). Check on right back wheel.

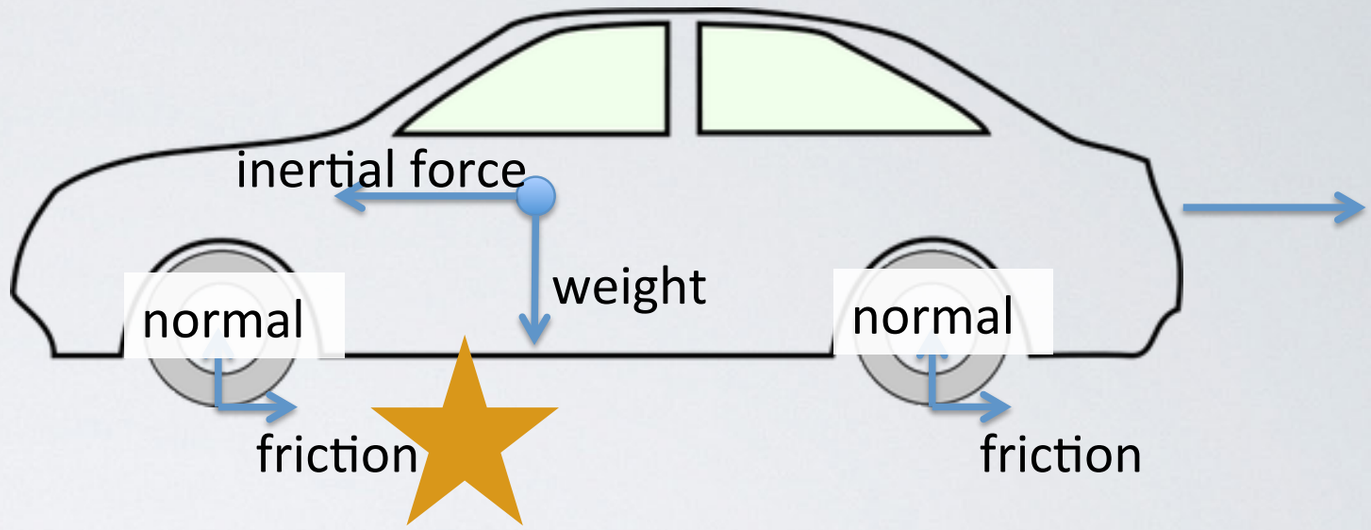




MODELING

Physics systems

```
==> (explain normal-forces)
REASON: rear-wheels-force decreased AND
    its magnitude exceeds the traction
    threshold.
Since the rear wheels lost traction
    the friction of the contact patches
        MUST HAVE decreased;
    so, the normal forces MUST HAVE
    decreased.
Consistent with the accelerometers.
```



MODELING

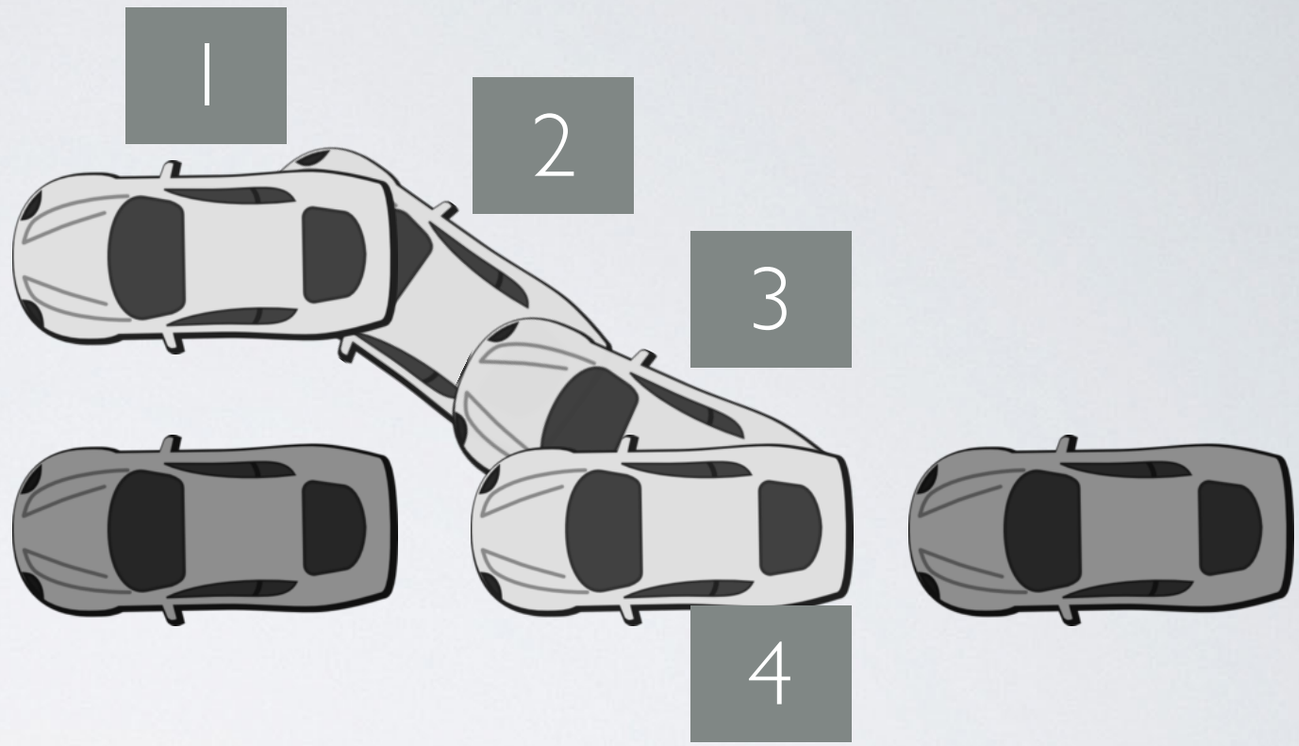
Physics systems

```
==> (explain normal-forces)

REASON:
front-wheels-force decreased AND tire
pressure is low.

Checking on mechanical system for anomalies…
```

MODELING



Explanatory parking

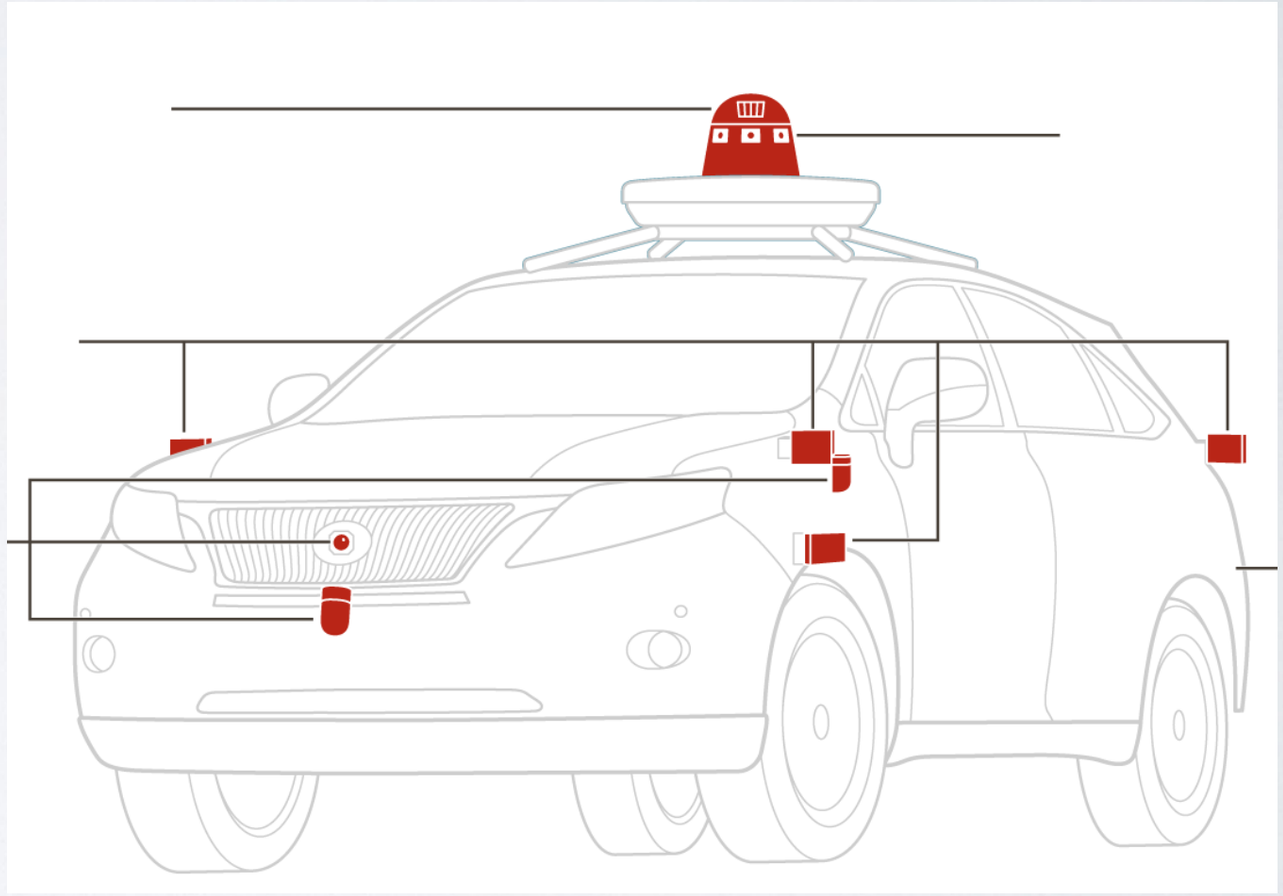
```
==> (explain parking)

Approach - within threshold
Turn - risky, but within threshold.
S-curve complete
Parking complete.
```

Joint work with S. Lu and B.Z. Yuan.

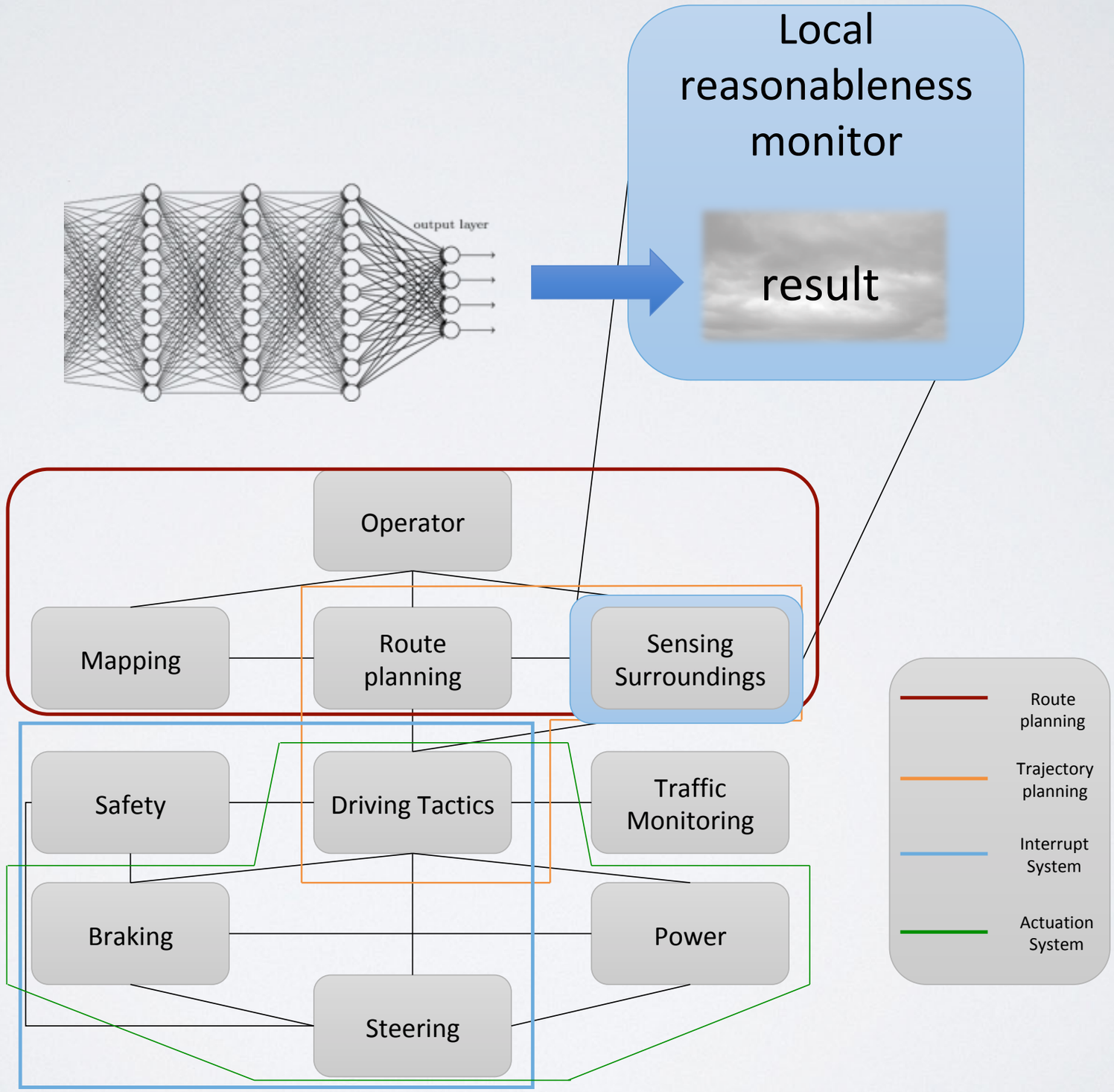
WHAT ABOUT SELF-DRIVING?

- Same mechanics
- Same physics
- New perception
- More sensors



By Guilbert Gates | Source: Google | Note: Car is a Lexus model modified by Google. Uber's sensing system uses similar technology.

SELF-DRIVING SYSTEM DESIGN

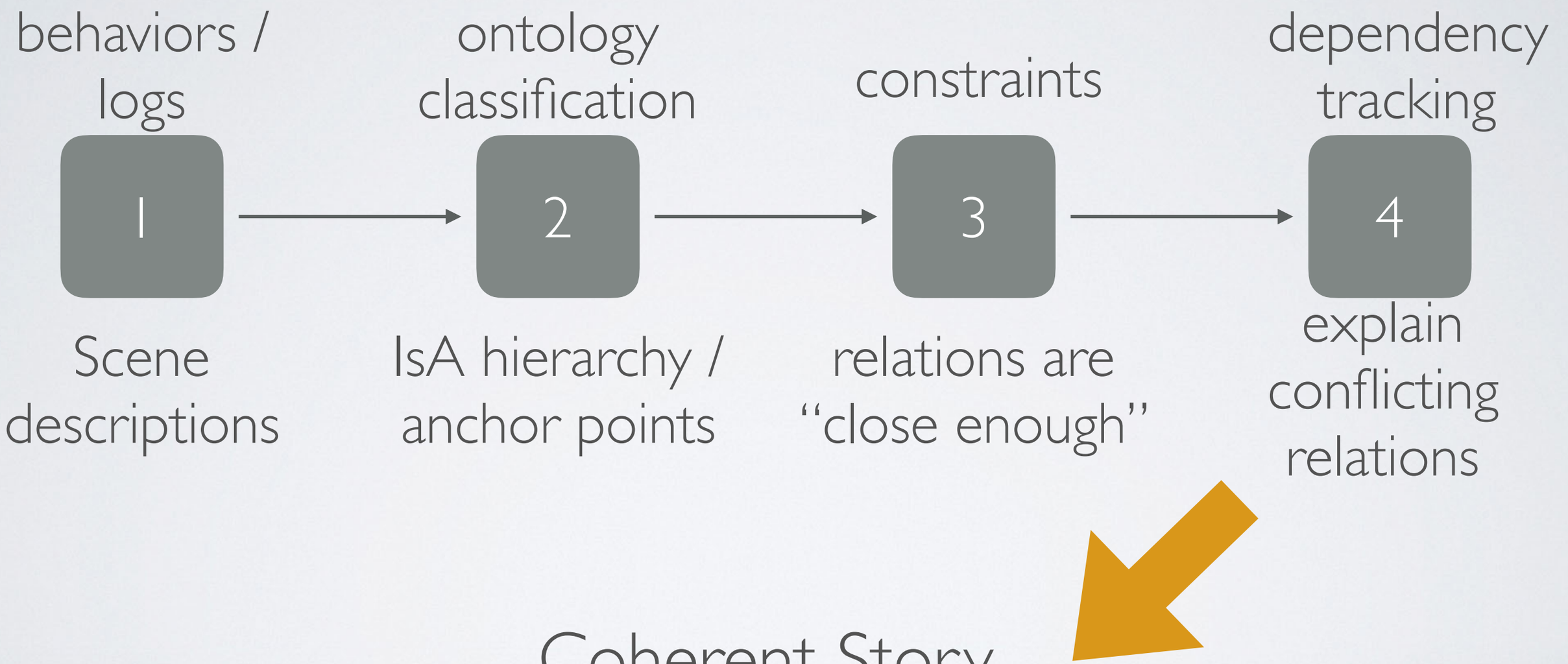


EXPLAINING PERCEPTION TWO WAYS

- Motivation - A first steps towards understanding machine perception is to constrain the output to be reasonable.

- Two ideas
  - Data representation: ConceptNet
  - Structural representation: Conceptual primitives

METHODS (I)


behaviors / logs
ontology classification
constraints
dependency tracking
1
2
3
4
Scene descriptions
IsA hierarchy / anchor points
relations are "close enough"
explain conflicting relations
Coherent Story

PRELIMINARY RESULTS(I)

A mailbox crossing the street

Reasonableness monitor

Perception

A mailbox crossing the street

Premises

(mailbox, IsA, heavy object)
(mailbox, moves, False)
(mailbox, LocatedNear, street)

PRELIMINARY RESULTS(I)

A mailbox crossing the street

```
input : “Mailbox crossing the
street”

This perception is UNREASONABLE
using data from ConceptNet5.

REASONING:
A mailbox is an object typically
found near a sidewalk.

Mailboxes cannot cross a street
because mailboxes are objects
that do not move on their own.
```

LIMITATIONS

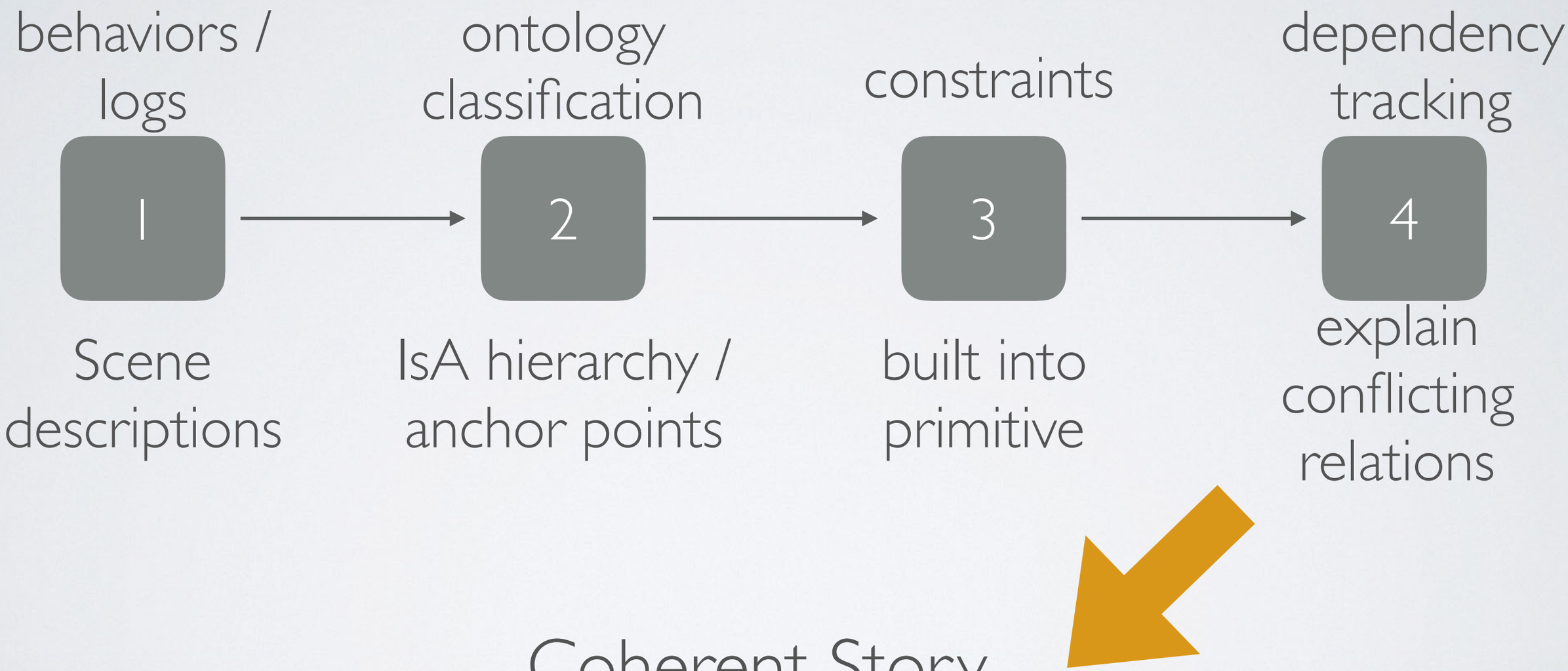
```
input : “A penguin eats food”


This perception is UNREASONABLE


REASONING:
A penguin is an animal that lives in Antartica and eats
enough to eat.  Food is an animal that lives in the
refrigerator and eats food.


So a penguin cannot reasonably be located at the same
location as food.
```

METHODS (II)


behaviors / logs
ontology classification
constraints
dependency tracking
1
2
3
4
Scene descriptions
IsA hierarchy / anchor points
built into primitive
explain conflicting relations
Coherent Story

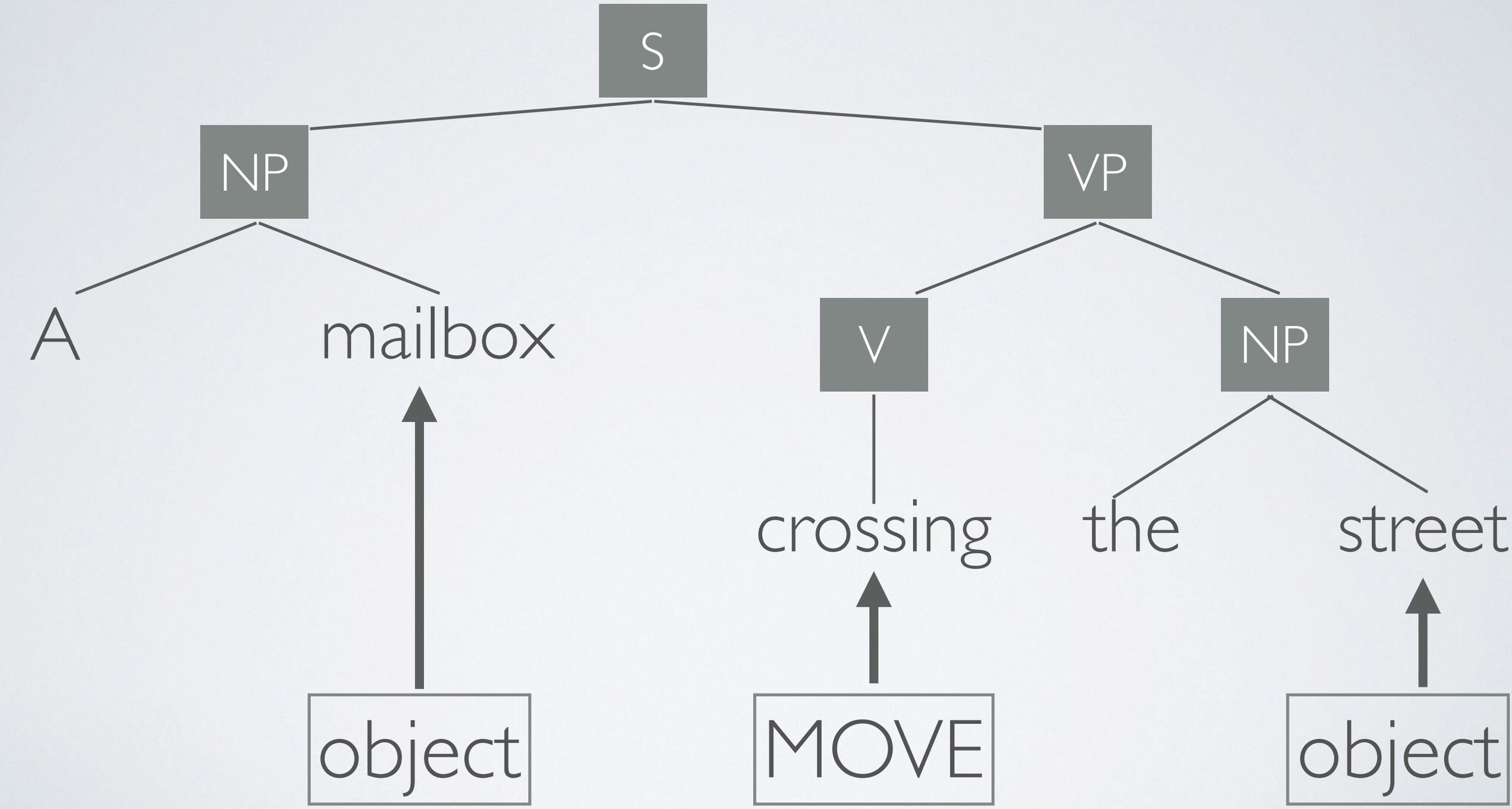
A MAILBOX CROSSING THE STREET

mailbox

MOVE → across street

?

PRELIMINARY WORK



EXPLAINING PERCEPTION (II)

A mailbox crossing the street

```
This perception is unreasonable.

==================================

A mailbox is an object or thing
that cannot move on its own.  So
it is unreasonable for a mailbox
to cross the street.
```

EXPLAINING PERCEPTION (II)

A mailbox crossing the street during a hurricane

```
This perception is reasonable.
==============================
Although a mailbox cannot
move on its own, a
hurricane can propel a
stationary object to move.
So it is reasonable for a
mailbox to cross the street
```

MONITOR IN DEVELOPMENT

Reasonableness of Vehicle Actions

Vehicle Perception

- (●) Red
- ( ) Yellow
- ( ) Green
- [ ] Pedestrian

Driving Tactics

- (●) Wait
- ( ) Go forward
- ( ) Go right

Vehicle State

- ( ) Stopped
- ( ) Moving slowly
- ( ) Moving quickly

The vehicle waits at a red light

Reasonable? ✔

This perception is REASONABLE
==========
A red light means stop.
So it is reasonable for the vehicle to wait

INTERNAL STORIES

Weather sensor

Hurricane

Premises

(hurricane, has, high winds)

Perception

A mailbox crossing the street

Premises

(mailbox, IsA, heavy object)
(mailbox, moves, False)
(mailbox, LocatedNear, street)

INTERNAL STORIES

Weather sensor

Hurricane

Premises

(hurricane, has, high winds)

Perception

A mailbox crossing the street

Premises

(mailbox, IsA, heavy object)
(mailbox, moves, False)
(mailbox, LocatedNear, street)

INTERNAL STORIES

Weather sensor

Hurricane

Premises

(hurricane, has, high winds)

Perception

A mailbox crossing the street

Premises

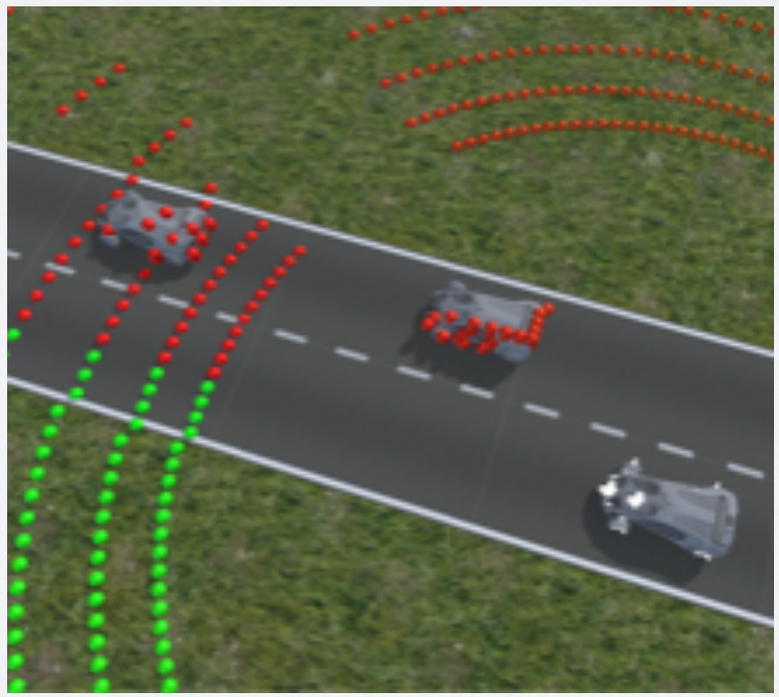
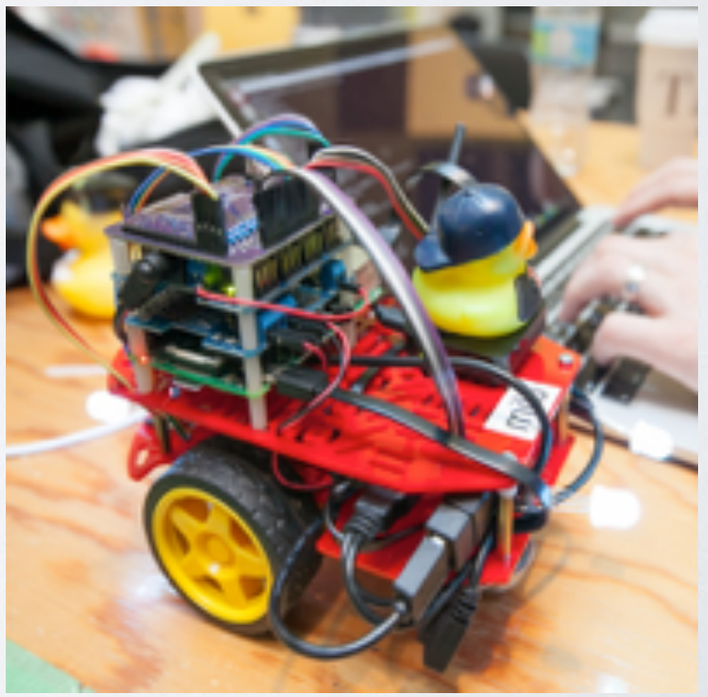
(mailbox, IsA, heavy object)
(mailbox, moves, False)
(mailbox, LocatedNear, street)

internal story: high winds can cause heavy objects to move

FUTURE WORK

- Explaining non-local inconsistencies
- Explaining internal stories and premises
- Incorporating into full-system design

reasonability → relevance

CONTRIBUTIONS

- Ex-post-facto explanations
- Explanations of reasonableness for language descriptions of perception
- Incorporating monitor into a working autonomous simulation.

LESS FRUSTRATION, MORE EXPLANATION

SOFTWARE USED

- Reasoning software
  - MIT/GNU scheme (free software)
    - Art of the Propagator System (free software)
    - Python (open-source)
    - ConceptNet (CC BY-SA 4.0)
- Simulation - Unity game engine and Carla (open-source)