Trustworthy Autonomous Systems

Team members: **Lalana Kagal, Ben Yuan**, Leilani Gilpin, Tianye Chen, Vaikkunth Mugunthan, Ayush Sharma, Doron Hazan, Wanyi Xiao

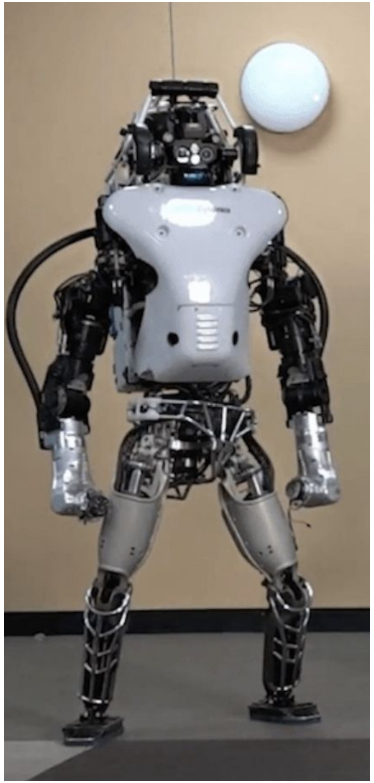
Technical Champions: Alex Abbati, Ali Ayub, Isidoros Doxas, Neta Ezer

MIT Computer Science and Artificial Intelligence Lab

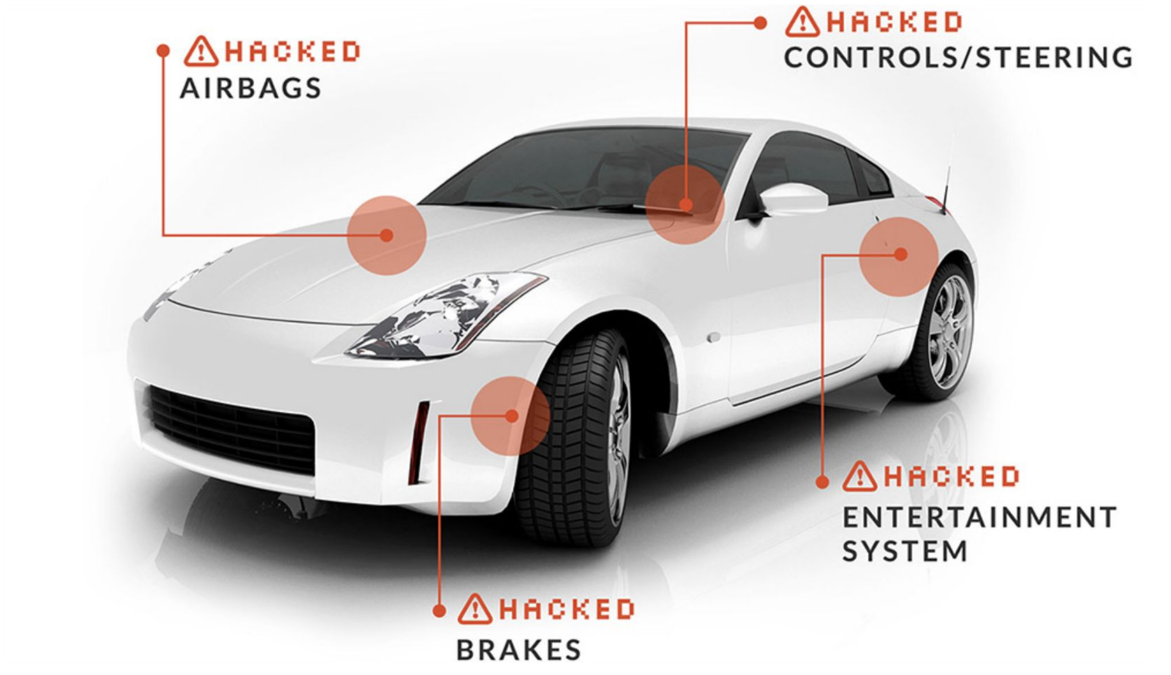
Massachusetts
Institute of
Technology

CSAIL
MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

What problem are we trying to solve

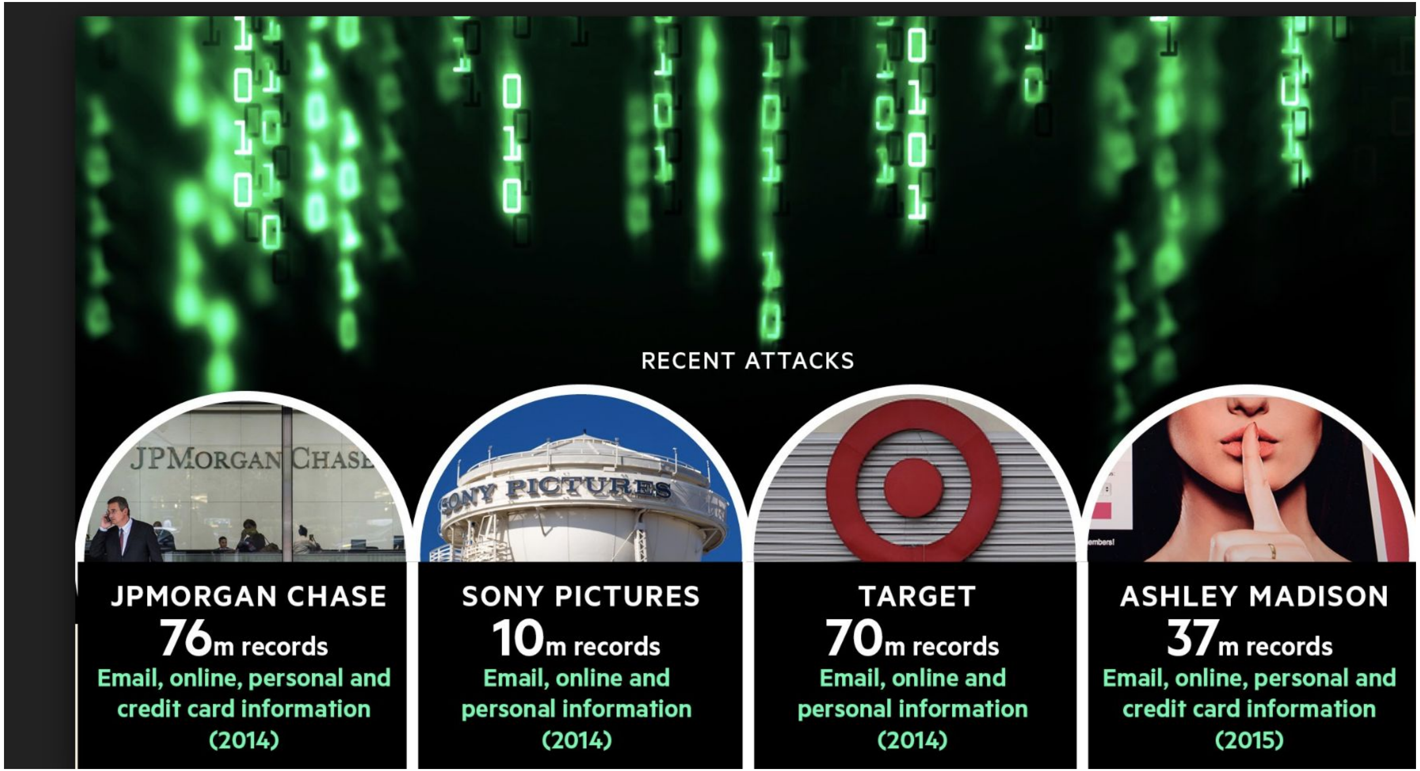


Potential for compromise



Potential for failure

Potential for information leakage



RECENT ATTACKS

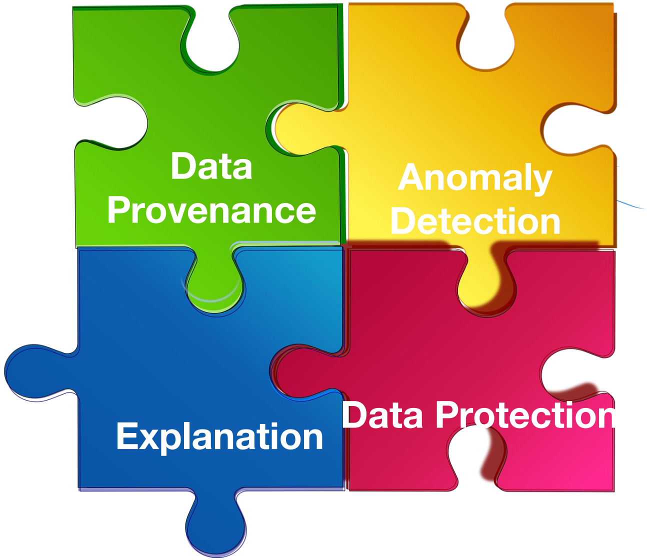
JPMORGAN CHASE
76m records
Email, online, personal and credit card information (2014)

SONY PICTURES
10m records
Email, online and personal information (2014)

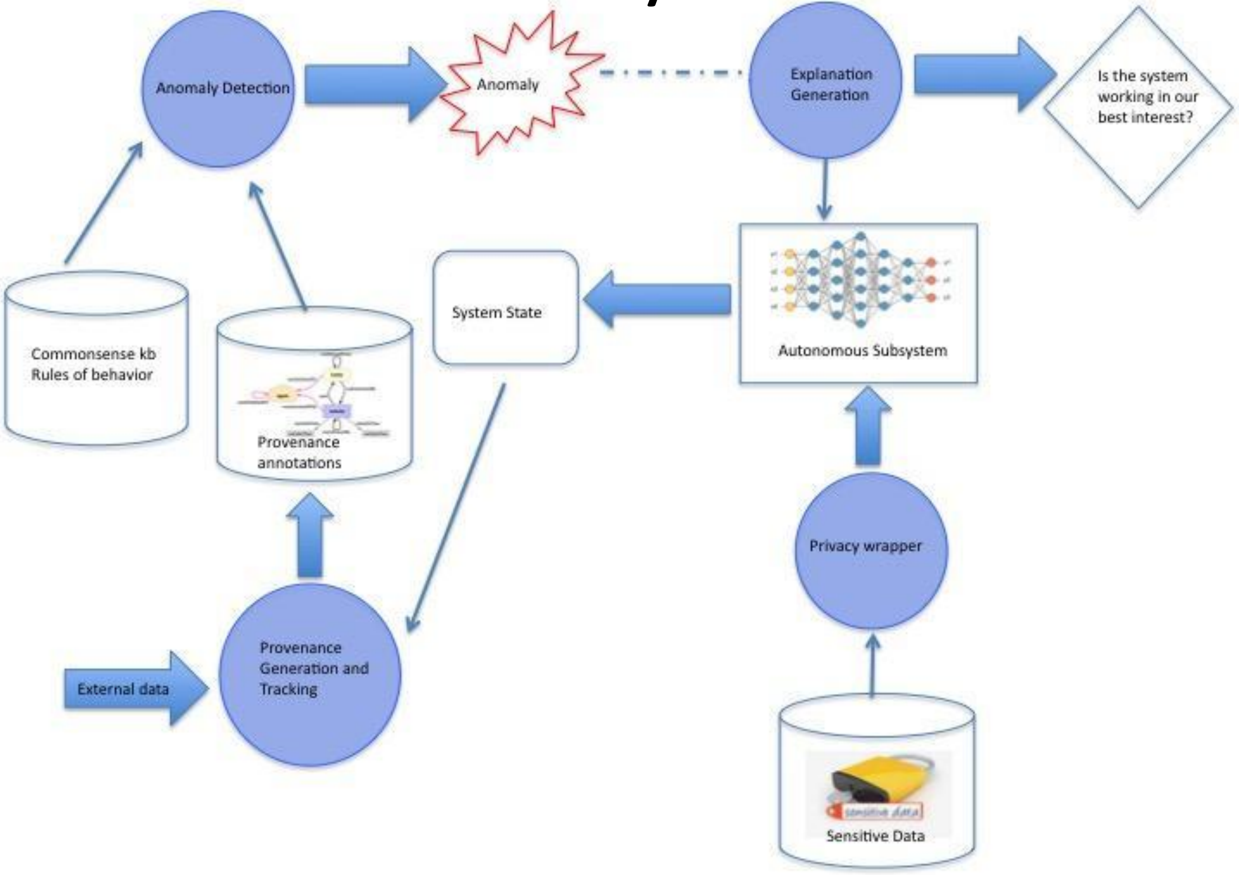
TARGET
70m records
Email, online and personal information (2014)

ASHLEY MADISON
37m records
Email, online, personal and credit card information (2015)

Vision for autonomous systems



Vision for autonomous systems



Anomaly Detection

Anomaly

Explanation Generation

Is the system working in our best interest?

Commonsense kb
Rules of behavior

Provenance annotations

System State

Autonomous Subsystem

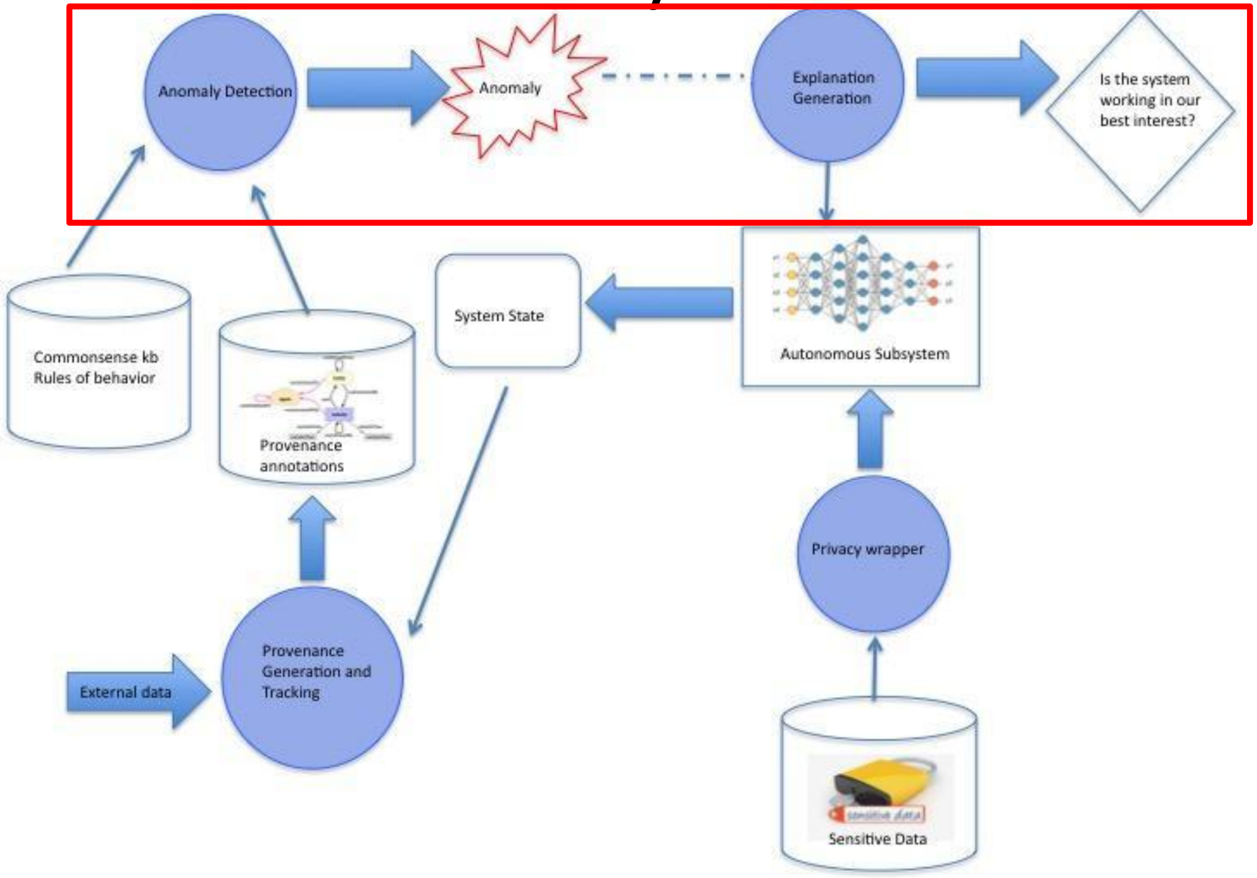
Privacy wrapper

External data

Provenance Generation and Tracking

Sensitive Data

Vision for autonomous systems



Anomaly Detection

Anomaly

Explanation Generation

Is the system working in our best interest?

Commonsense kb
Rules of behavior

Provenance annotations

System State

Autonomous Subsystem

Privacy wrapper

External data

Provenance Generation and Tracking

Sensitive Data

Why explanations?

Explanations can help us…

- Understand motives / causality
- Identify assumptions
- Choose between alternatives
- Predict future system behavior

...and understand if our systems are working **in our best interest**.

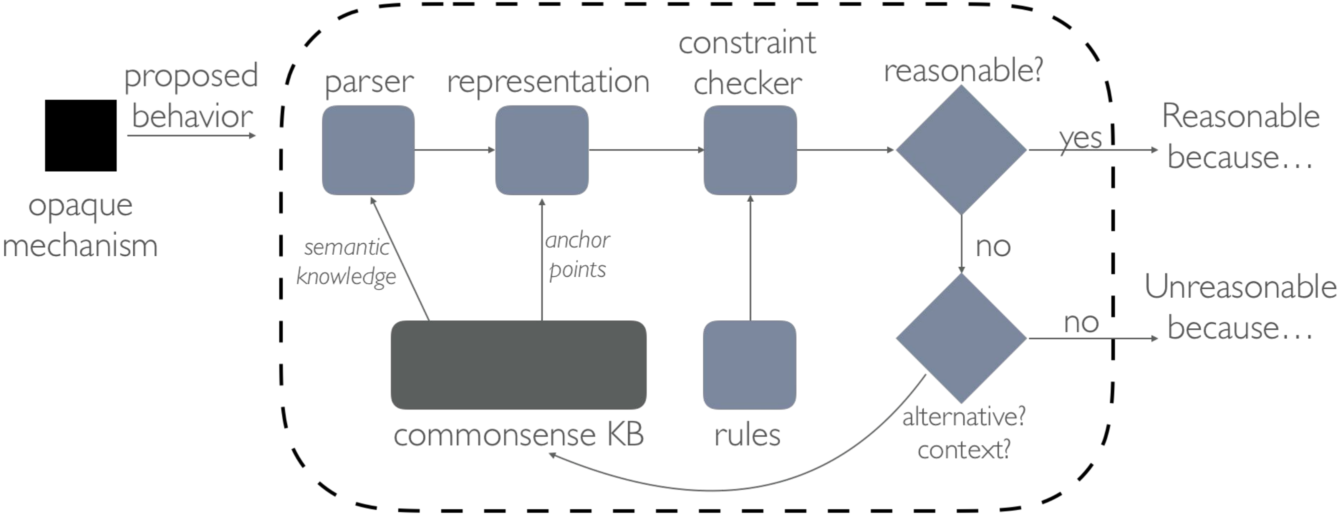
Explanations

Can we…

- Identify system misbehavior using common-sense reasoning?
- Examine decisions made by "black box" methods?
- Measure how well explanations convey useful data?

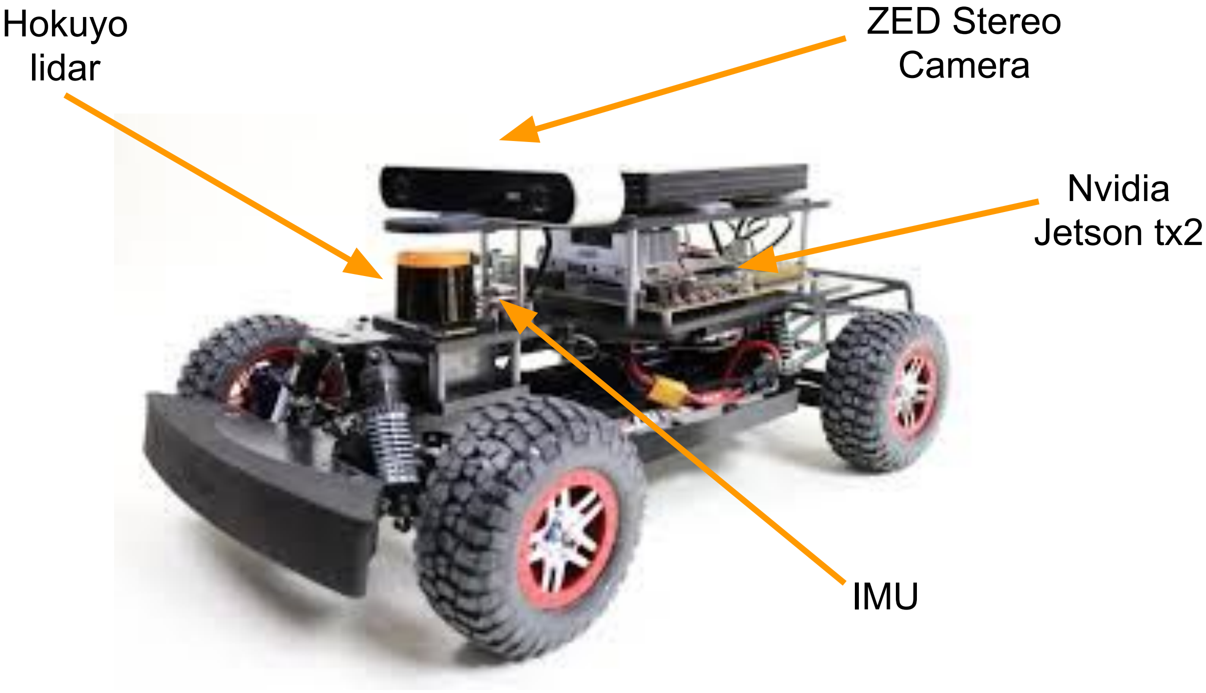
Common-sense reasonableness monitoring

"Is the current behavior reasonable?"



L. H. Gilpin and L. Kagal "An Adaptable Self-Monitoring Framework for Complex Machines", to be presented at AAMAS 2019.

Evaluation with the RACECAR platform



Hokuyo lidar

ZED Stereo Camera

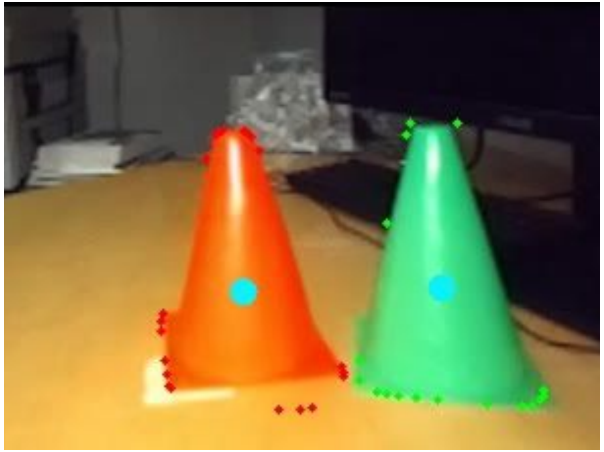
Nvidia Jetson tx2

IMU

Evaluation with the RACECAR platform

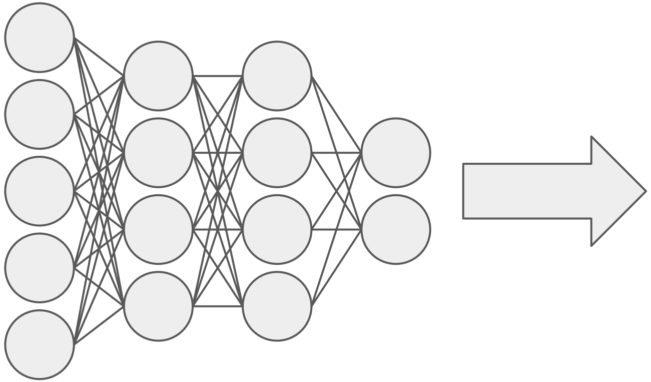
Validate that we can:

- Identify when observations deviate from prior rules (e.g. sequence of cone colors).
- Learn when "anomalies" constitute previously unknown rules.



Explaining DNN and CNN behaviour

“How did the system reach this decision?”



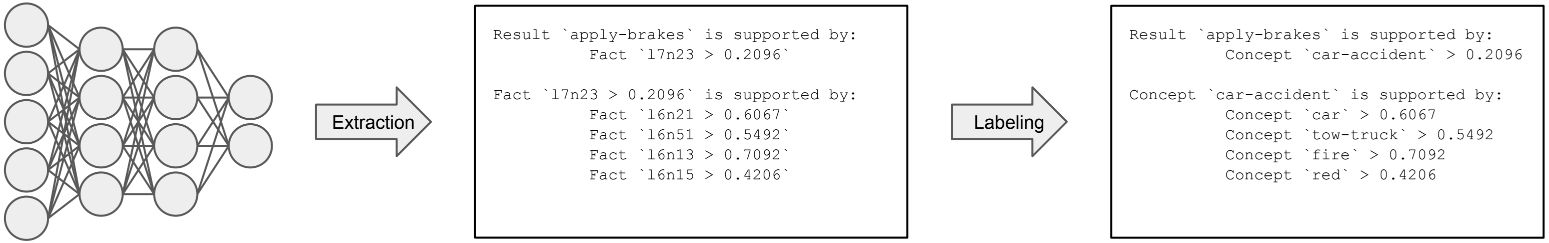
```
Result `apply-brakes` is supported by:
      Concept `car-accident` > 0.2096

Concept `car-accident` is supported by:
      Concept `car` > 0.6067
      Concept `tow-truck` > 0.5492
      Concept `fire` > 0.7092
      Concept `red` > 0.4206
```

Paper in progress.

Explaining DNN and CNN behaviour

Goal: Combine and improve techniques in…

- Rule discovery and extraction
- Semantic concept labeling
- Network pruning



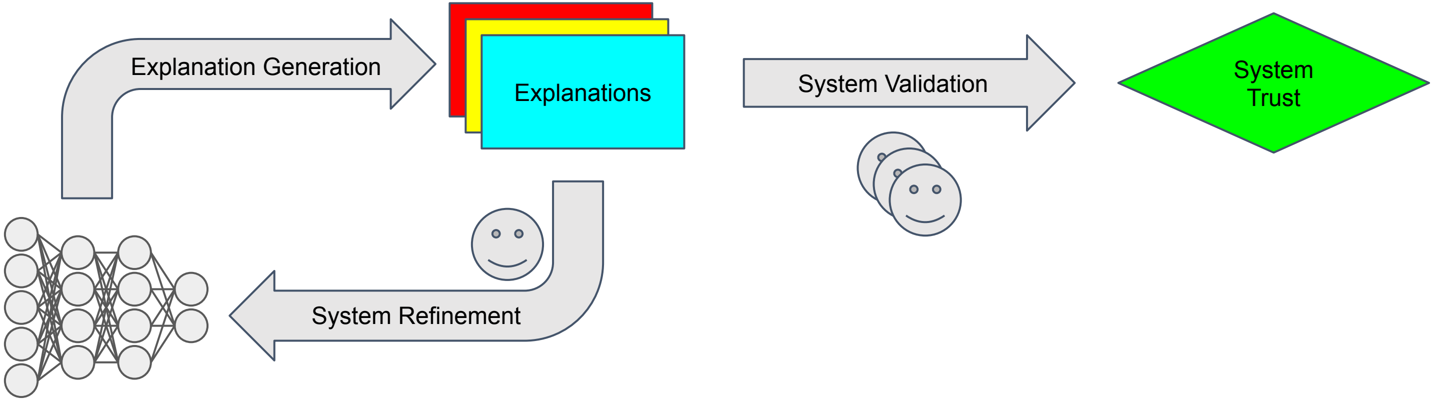
```
Result `apply-brakes` is supported by:
          Fact `l7n23 > 0.2096`

Fact `l7n23 > 0.2096` is supported by:
          Fact `l6n21 > 0.6067`
          Fact `l6n51 > 0.5492`
          Fact `l6n13 > 0.7092`
          Fact `l6n15 > 0.4206`
```

```
Result `apply-brakes` is supported by:
          Concept `car-accident` > 0.2096

Concept `car-accident` is supported by:
          Concept `car` > 0.6067
          Concept `tow-truck` > 0.5492
          Concept `fire` > 0.7092
          Concept `red` > 0.4206
```

...to achieve more useful explanations.

Measuring explanation effectiveness

“Does the explanation effectively highlight the ‘right thing’?”



Measuring explanation effectiveness

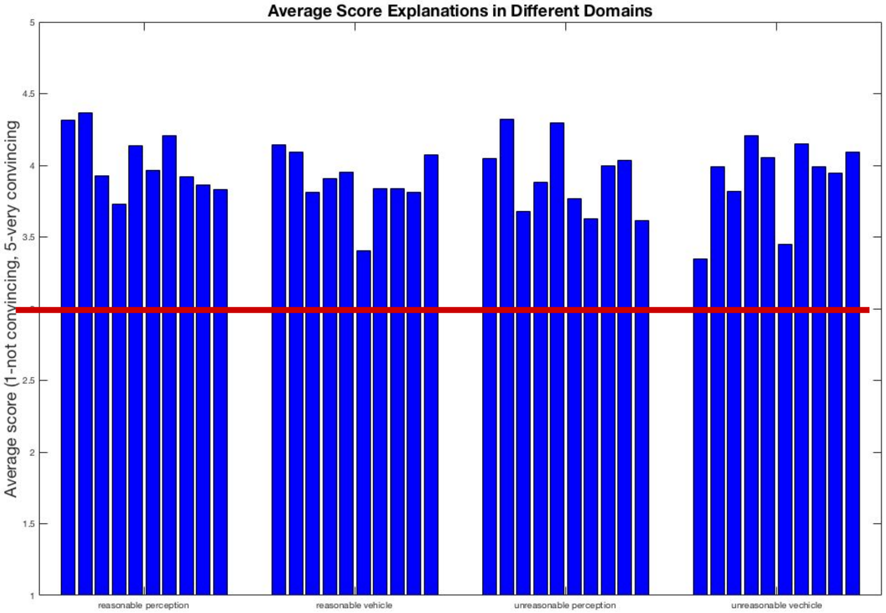
Variety of approaches to the problem of explanation.

	Processing	**Representation**	**Explanation Producing**
Methods	Proxy Methods Decision Trees Salience Mapping Automatic-rule extraction	Role of layers Role of neurons Role of vectors	Scripted conversations Attention-based Disentangled rep.

But how can we tell which approach works best for a given task?

L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal “Explaining explanations: an overview of interpretability of machine learning”, DSAA 2018.

Qualitative explanation evaluation



Prompt:

For example, you might see this explanation:

A cat is an animal and animals eat food. ***(statement)*** *Therefore it is reasonable for a cat to eat food.* ***(conclusion)***

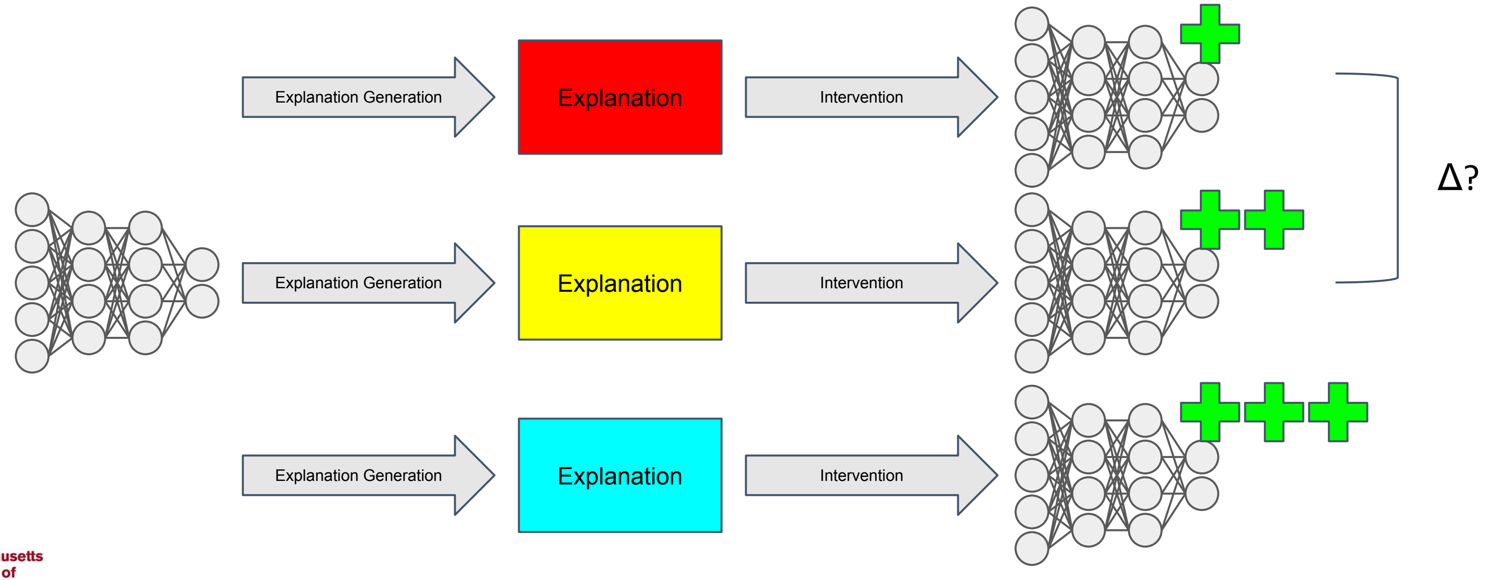
For each explanation, rate it from 1 to 5: 5 being *very convincing*, and 1 being *not convincing*. As you go through and rate each one, think about how convincing they are. Do you believe the statements provide a convincing explanation for the conclusion?

The example above is a convincing explanation for a cat eating food, so most people would rate it a 4 or 5.

L. H. Gilpin and L. Kagal "An Adaptable Self-Monitoring Framework for Complex Machines", to be presented at AAMAS 2019.

Future: quantifying explanation effectiveness

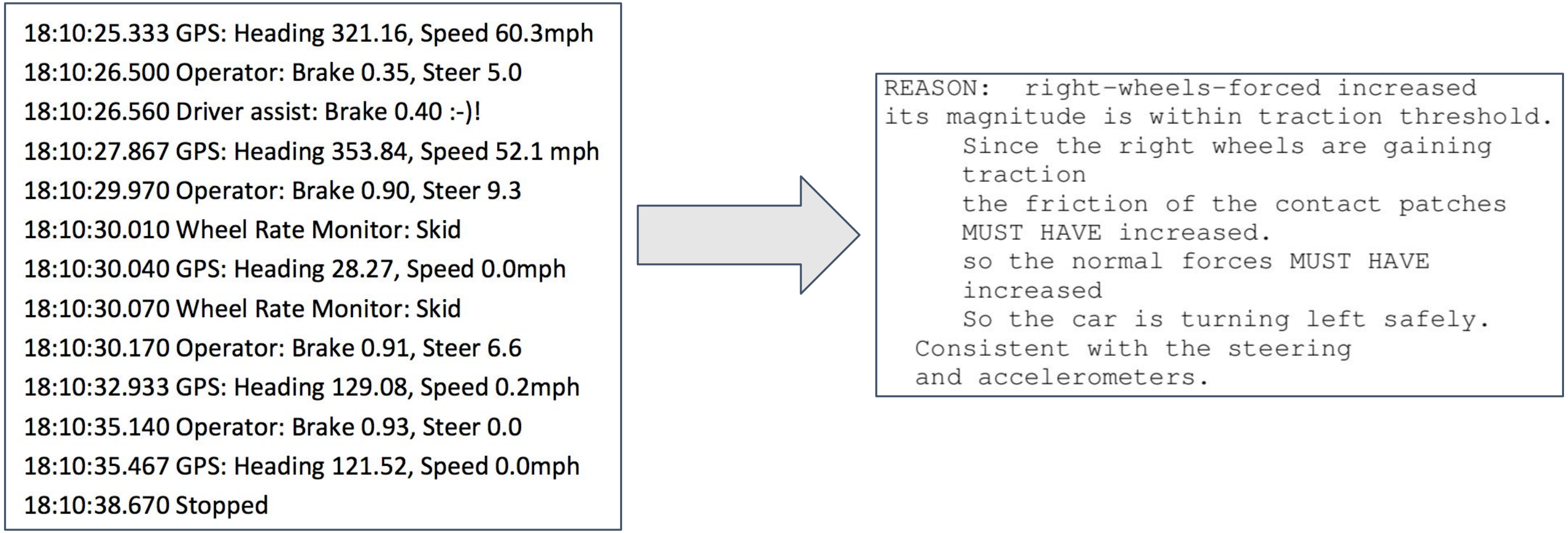
Goal: Measure the contribution of an explanation method to NN "repair".



Summary

Explanations from time-series data

“What happened? Why?”



```
18:10:25.333 GPS: Heading 321.16, Speed 60.3mph
18:10:26.500 Operator: Brake 0.35, Steer 5.0
18:10:26.560 Driver assist: Brake 0.40 :-)!
18:10:27.867 GPS: Heading 353.84, Speed 52.1 mph
18:10:29.970 Operator: Brake 0.90, Steer 9.3
18:10:30.010 Wheel Rate Monitor: Skid
18:10:30.040 GPS: Heading 28.27, Speed 0.0mph
18:10:30.070 Wheel Rate Monitor: Skid
18:10:30.170 Operator: Brake 0.91, Steer 6.6
18:10:32.933 GPS: Heading 129.08, Speed 0.2mph
18:10:35.140 Operator: Brake 0.93, Steer 0.0
18:10:35.467 GPS: Heading 121.52, Speed 0.0mph
18:10:38.670 Stopped
```

```
REASON:  right-wheels-forced increased
its magnitude is within traction threshold.
     Since the right wheels are gaining
     traction
     the friction of the contact patches
     MUST HAVE increased.
     so the normal forces MUST HAVE
     increased
     So the car is turning left safely.
  Consistent with the steering
  and accelerometers.
```

L. H. Gilpin and B. Z. Yuan “Getting up to speed on vehicle intelligence”. In “Papers from the 2017 AAAI Spring Symposia”, 2017.