Explainable AI for Fairness and Accountability

Leilani H. Gilpin, Ph.D.
Assistant Professor of Computer Science & Engineering
UC Santa Cruz
lgilpin@ucsc.edu
lgilpin.com

Talk Agenda

Brief Intro

<u>Motivate problem: Systems are imperfect</u>

What is explainability?

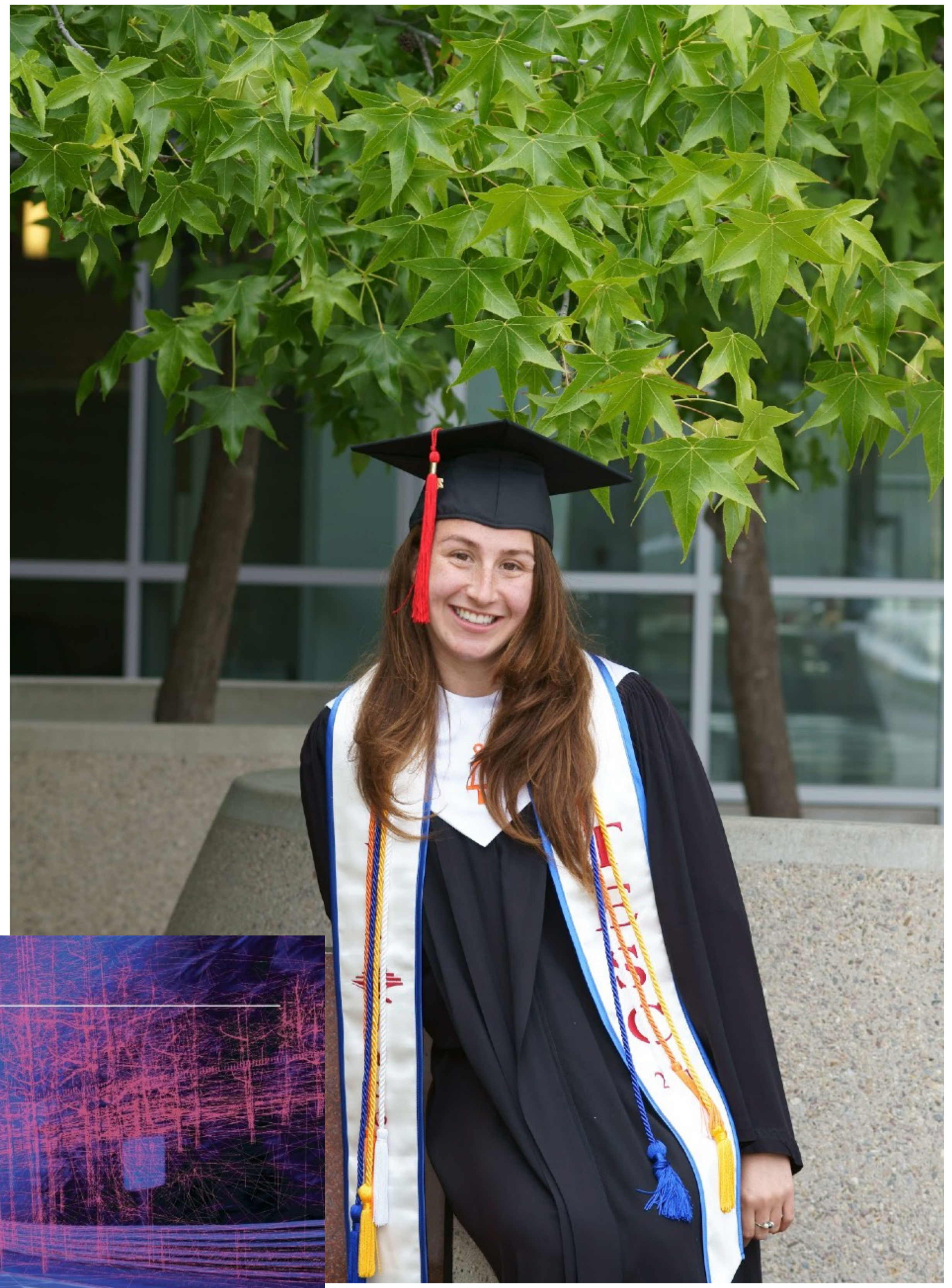
What is *actually* being explained?

How to evaluate explainability?

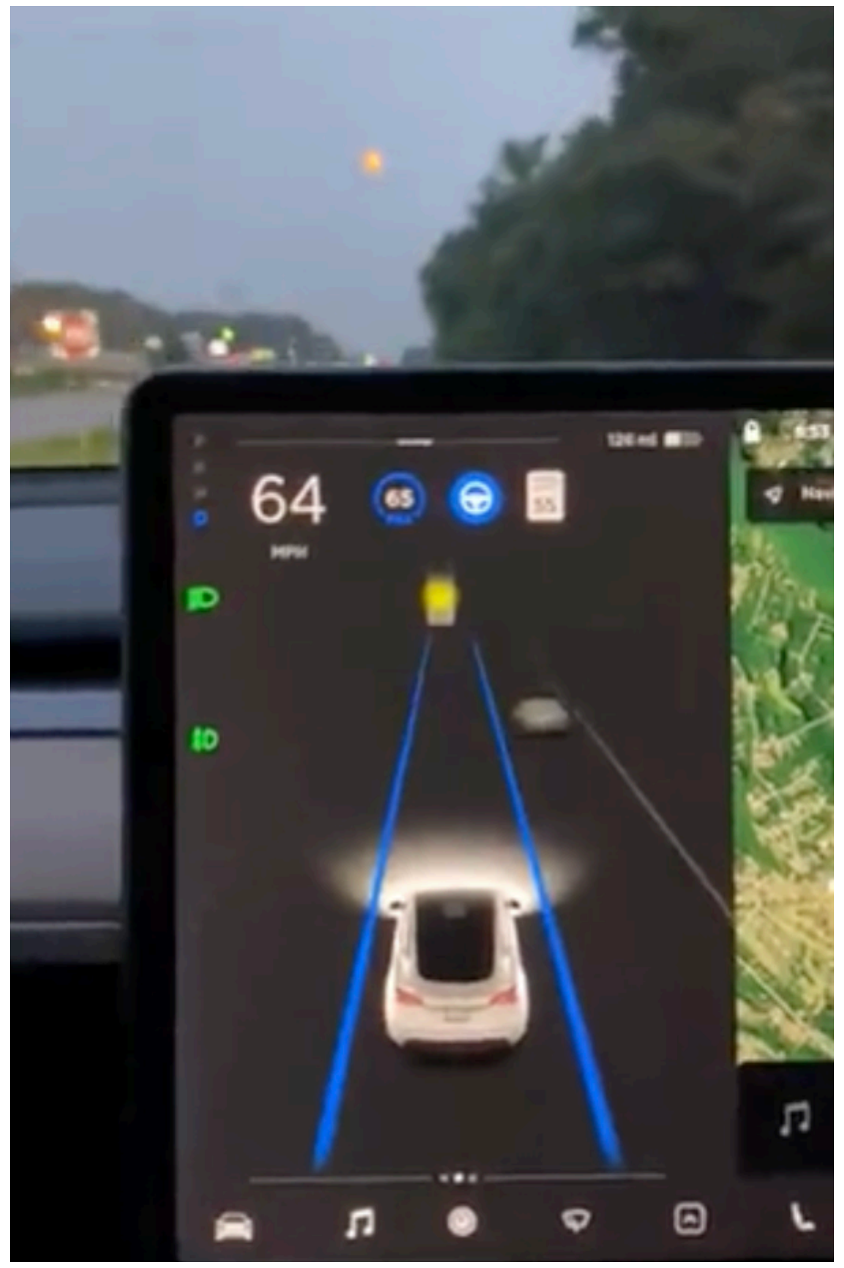
How to explain complex systems? (autonomous driving)

About Me

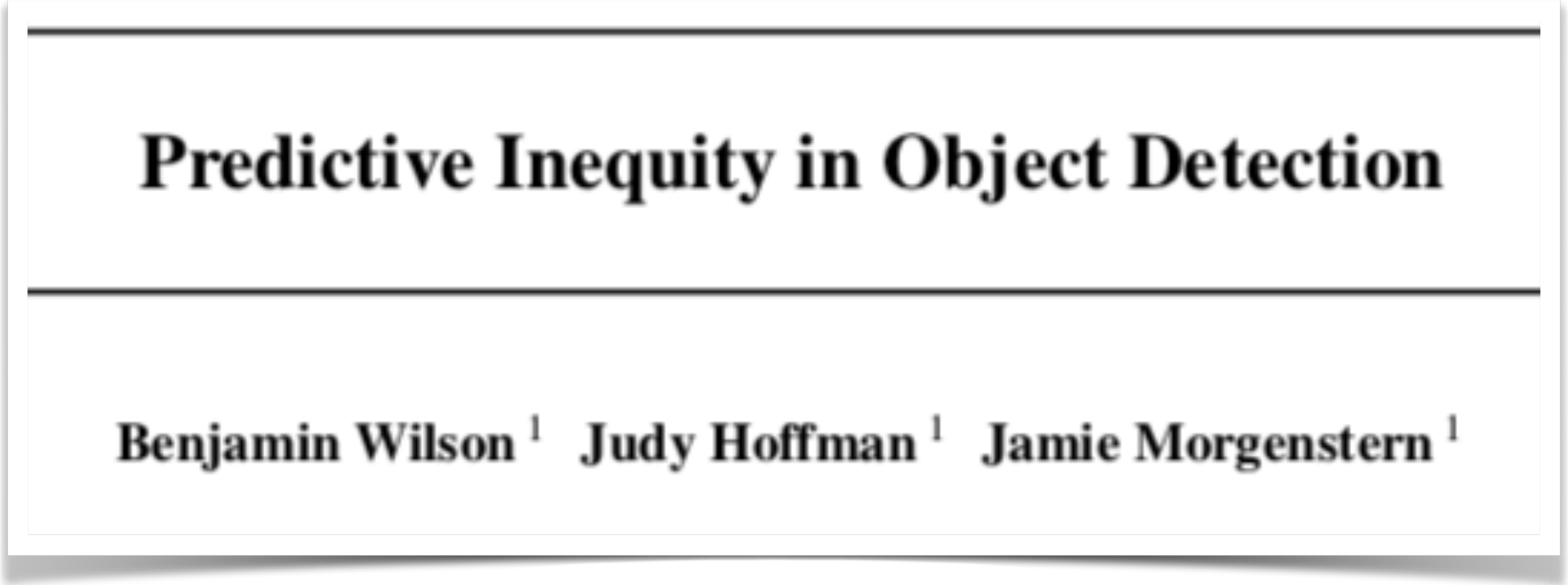
- B.S in Computer Science, B.S. in Mathematics at UC San Diego
- M.S. in Computational Mathematics from Stanford University (2013), Ph.D. in EECS from MIT (2020).
- Industry experience
  - Xerox PARC
  - INRIA (France)
  - Sony AI
- Research: The methodologies and technologies for complex systems to explain themselves.



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]

Societal Need for Explanation

BUSINESS NEWS OCTOBER 9, 2018 / 11:12 PM / 2 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

Business Impact

An AI-Fueled Credit Formula Might Help You Get a Loan

Startup ZestFinance says it has built a machine-learning system that's smart enough to find new borrowers and keep bias out of its credit analysis.

by Nanette Byrnes February 14, 2017

Talk Agenda

Motivate problem: Systems are imperfect

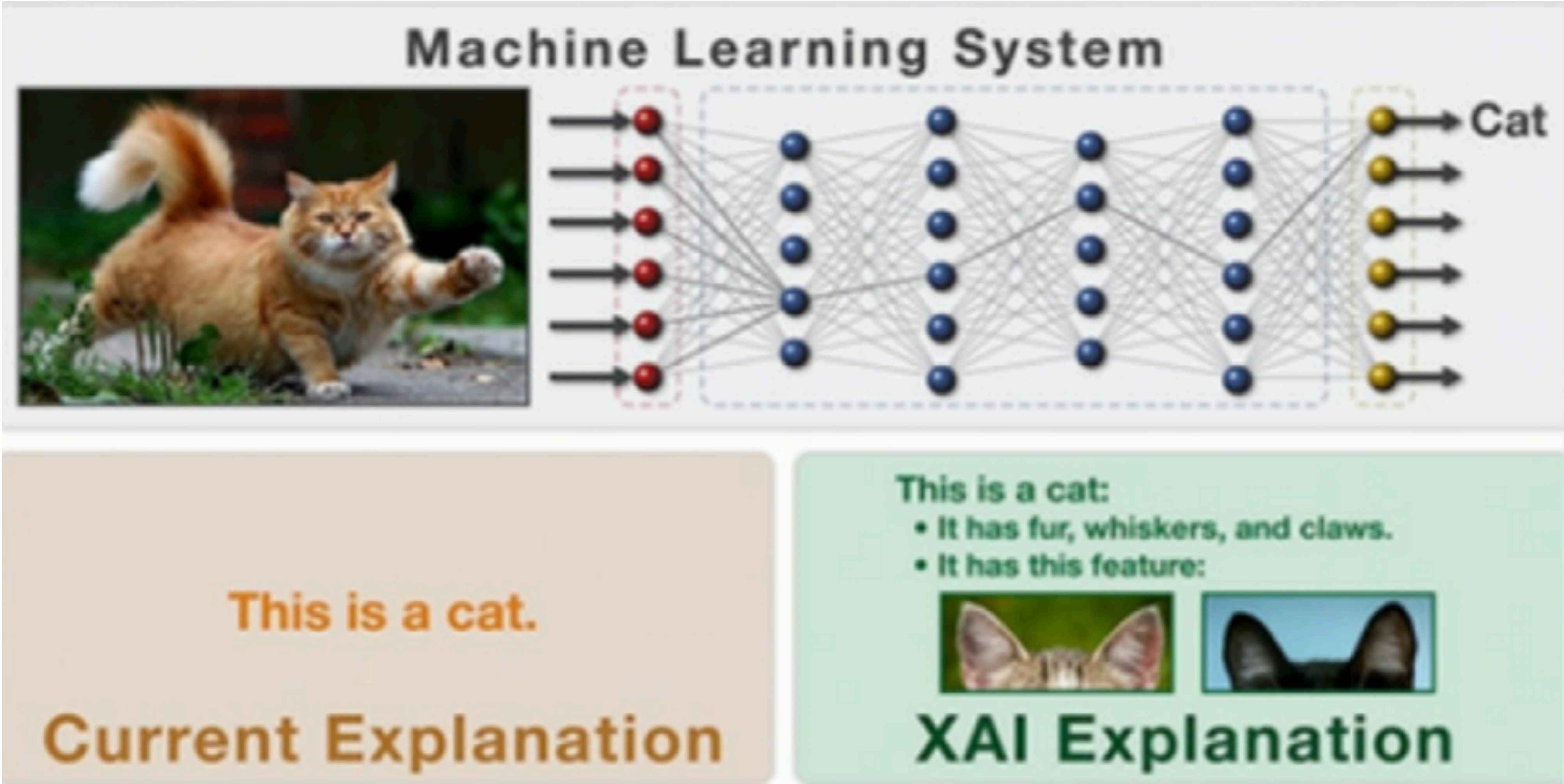
What is explainability?

What is *actually* being explained?

How to evaluate explainability?

How to explain complex systems? (autonomous driving)

What is Explainability?



From Darpa XAI

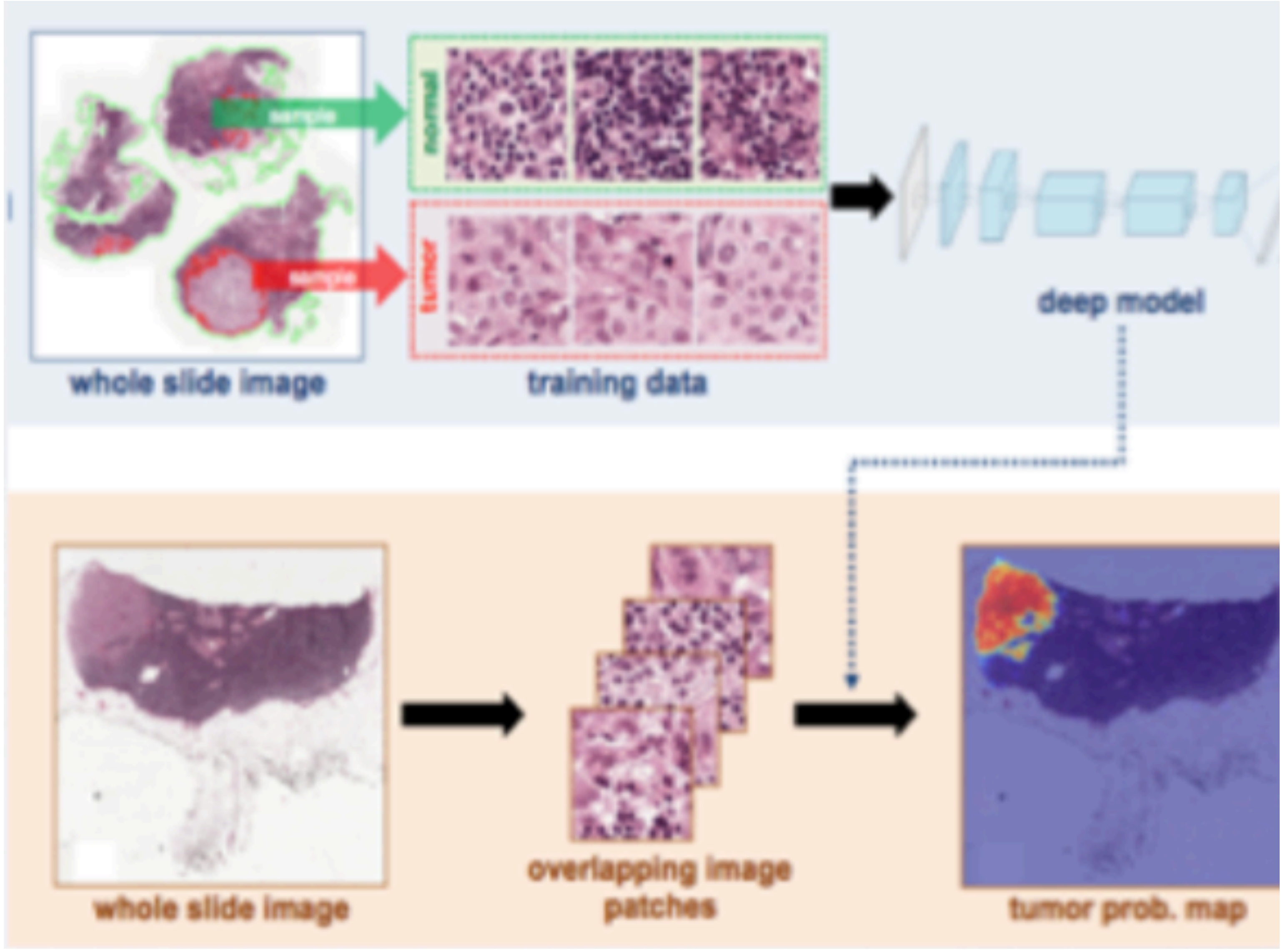
“Explanations…express answer to not just any questions but to questions that present the kind of **intellectual difficulty**…”

Sylvain Bromberger, *On What We Know We Don’t Know*

Deep Nets are Everywhere

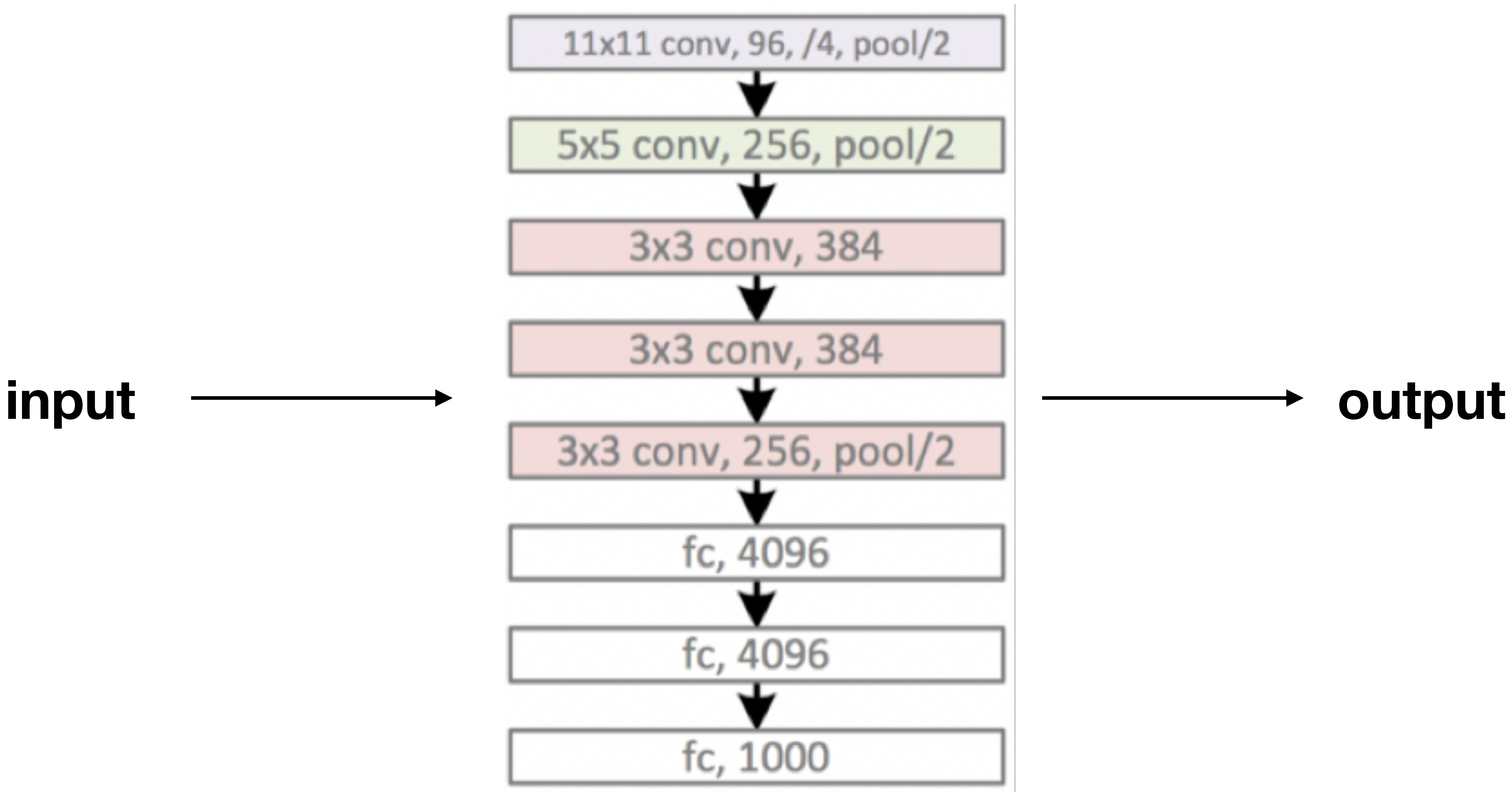
Self-driving Cars

Playing Go



Making Medical Decisions

Deep Nets are Not Understandable



Middle “hidden” layers

Whenever correct: “whatever you did in the middle, do more.”
Whenever wrong: “whatever you did in the middle, do less.”

Review of Research in XAI

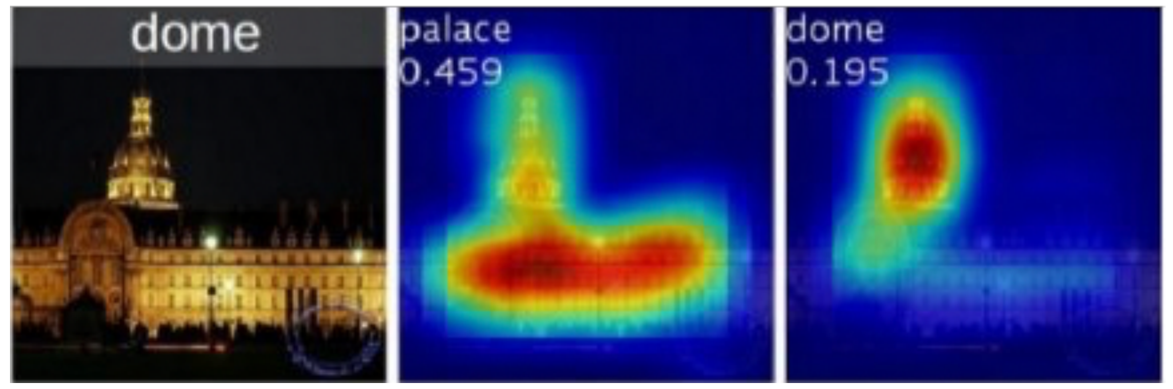
- Definitions
- Taxonomy
    - Survey: Literature review (87 papers) in computer science, artificial intelligence, and philosophy.
    - Recommendations for Evaluation
- How can explanations help (e.g. anomaly detection).
- Contributions and Future Work

Definitions

- Explainability != Interpretability
- **Interpretability** describes the internals of a system that is *understandable* to humans.
- **Completeness** describes operation in an *accurate* way.
- <u>An explanation needs **both.**</u>

What we Have

Visual cues

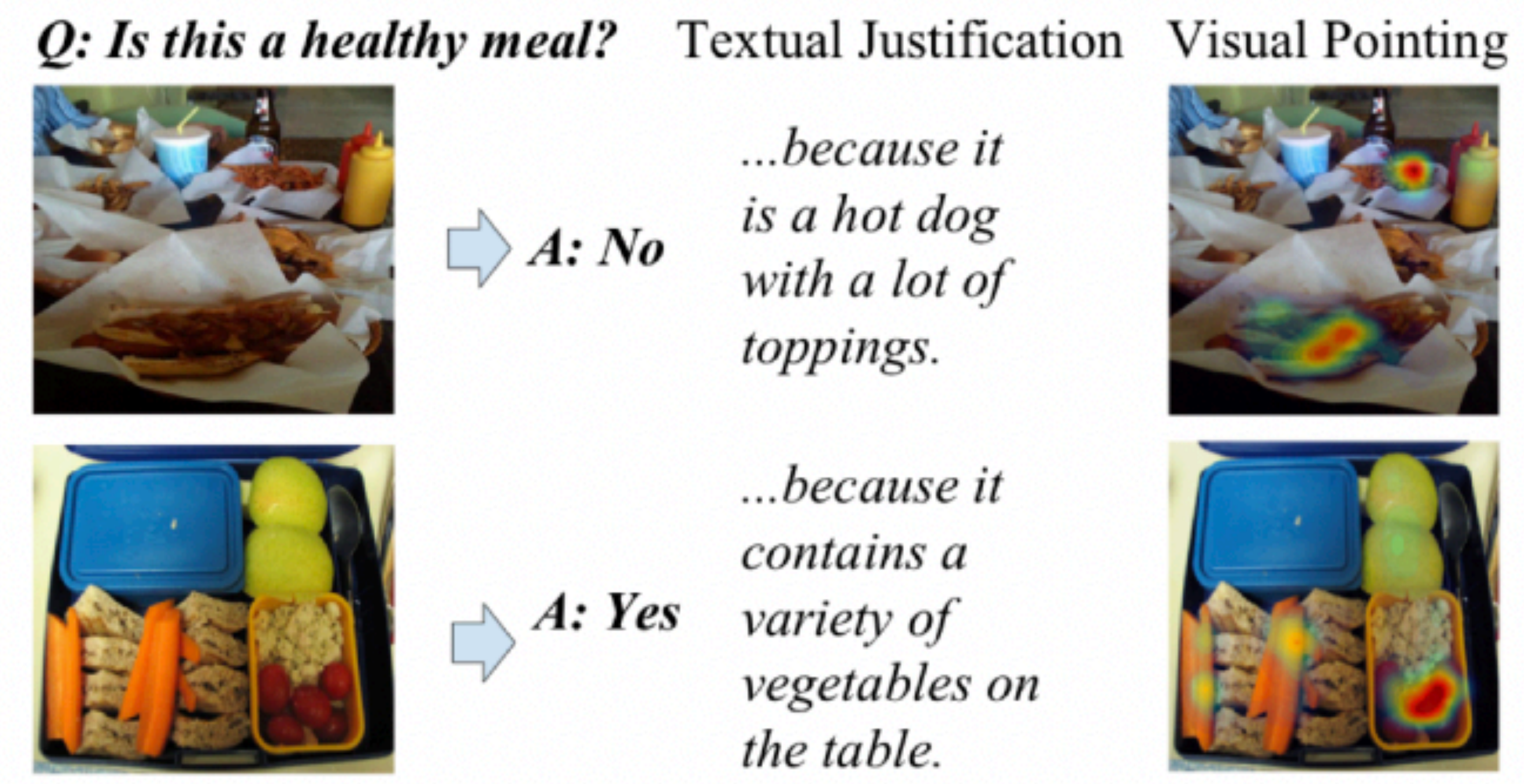


Interpretable,
not complete

Role of individual units

Complete,
not interpretable

Attention based



Interpretable,
not complete

Why this Matters

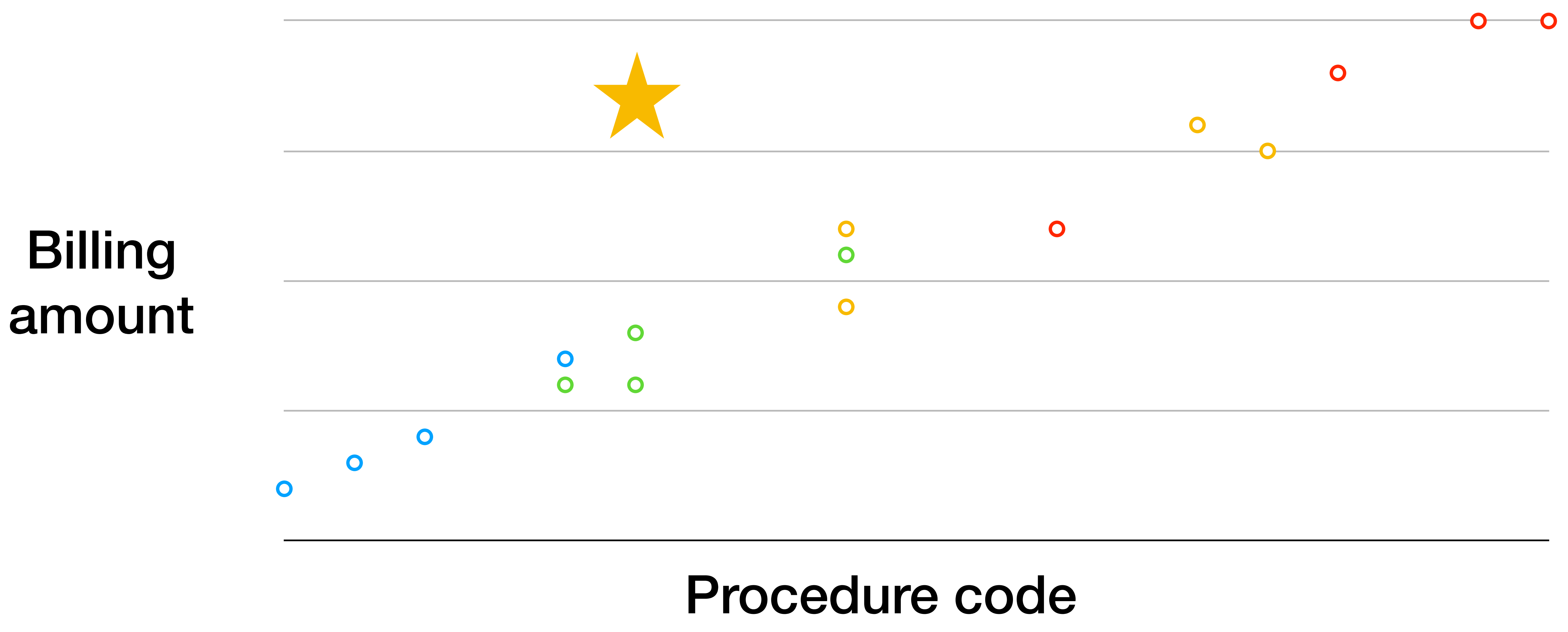
<u>Interpretability</u>

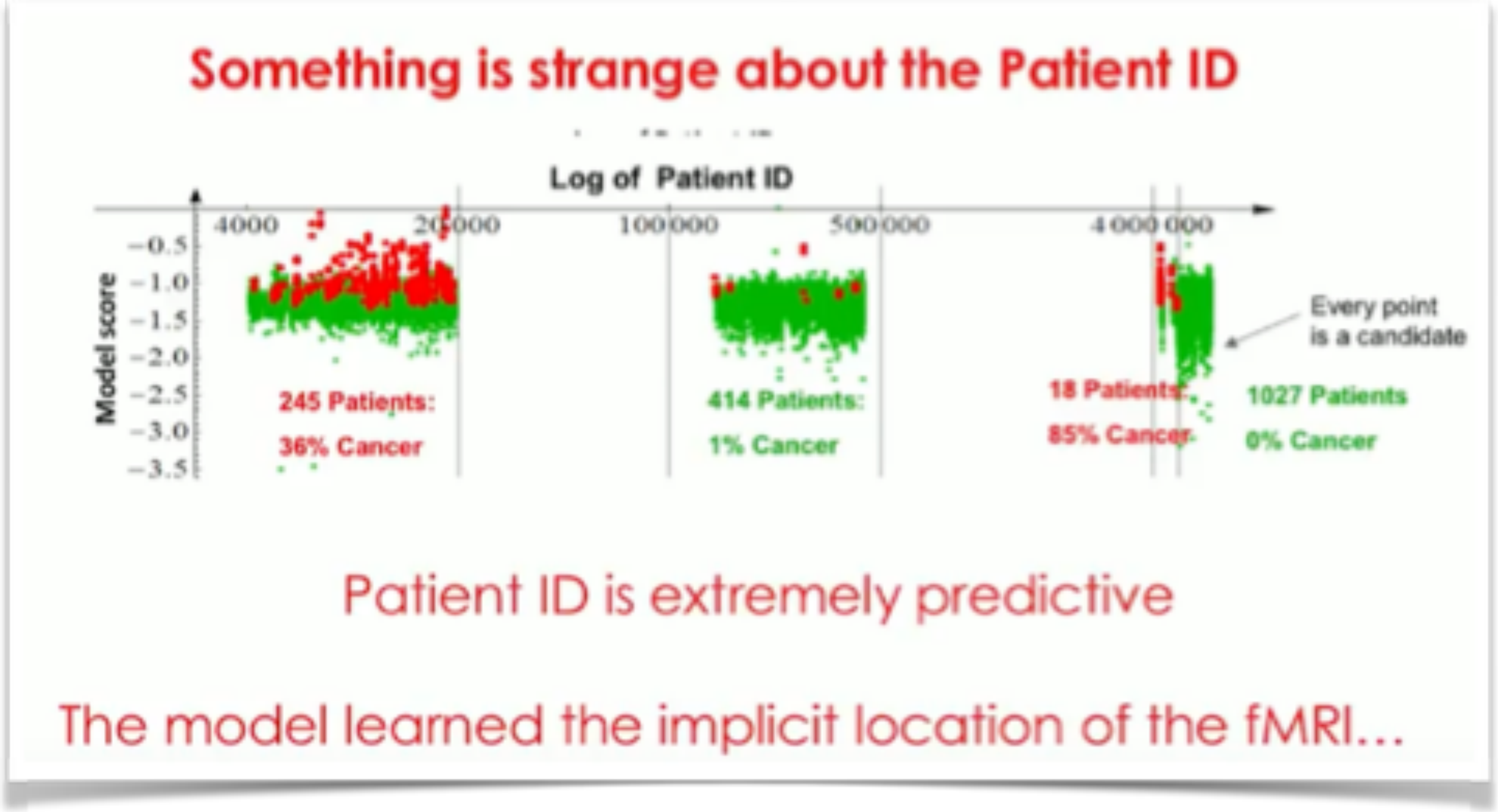
- GDPR
- Liability for decision making

Why this Matters

<u>Completeness</u>

- Explaining the wrong thing.
- Making decisions for the wrong reasons.





From Claudia Perlich at *Women in Data Science* 2018.

Talk Agenda

Motivate problem: Systems are imperfect

What is explainability?

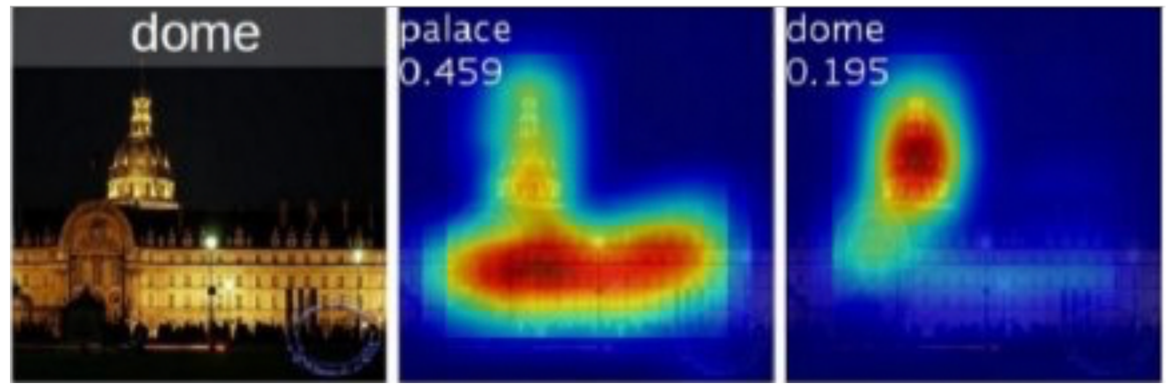
What is *actually* being explained?

How to evaluate explainability?

How to explain complex systems? (autonomous driving)

What is Being Explained?

Visual cues

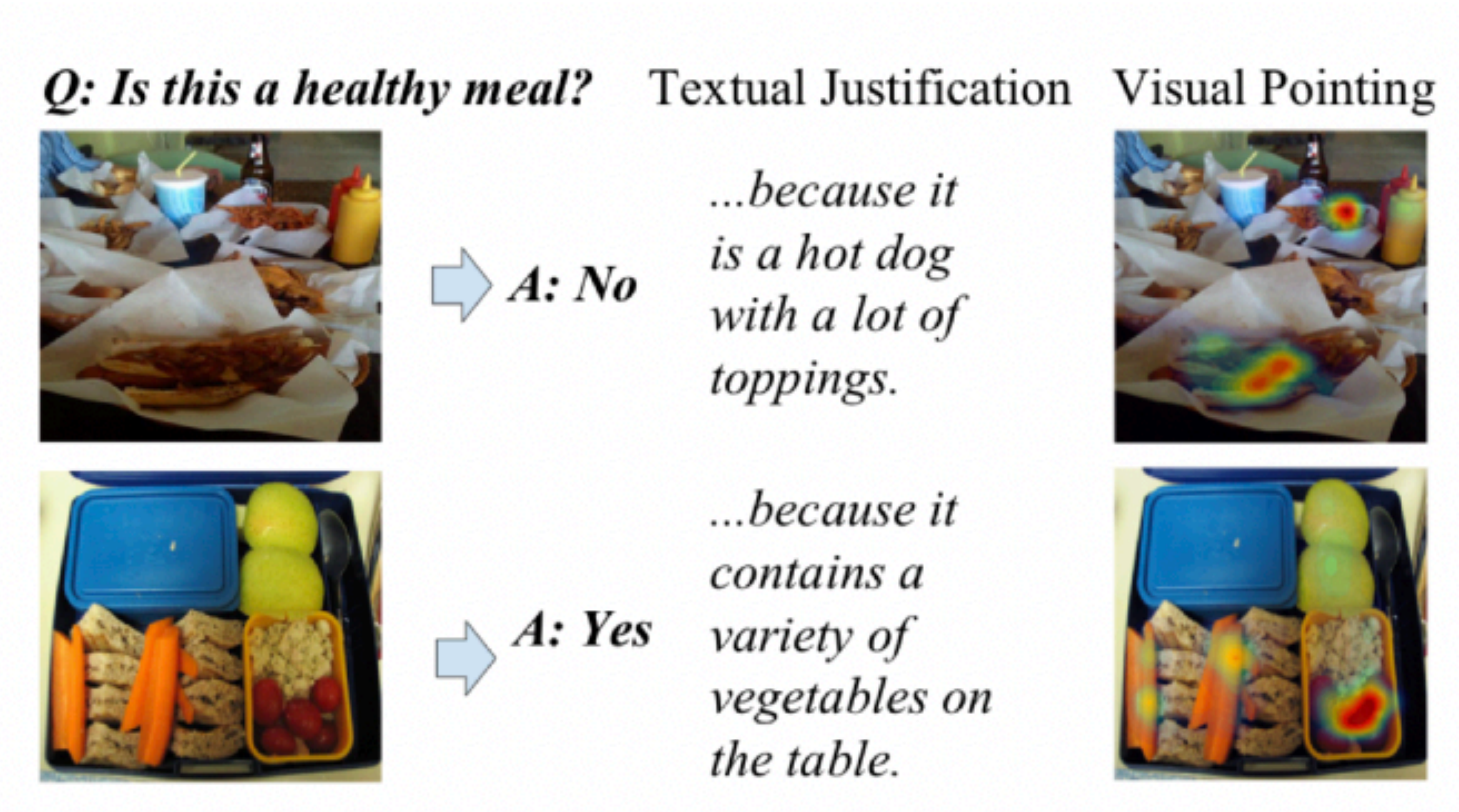


Explain processing

Role of individual units

Explain representation

Attention based



Explanation producing

Taxonomy

	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 representations

Methods that Explain Processing

DeepRED –
Rule Extraction from Deep Neural Networks*

Jan Ruben Zilke, Eneldo Loza Mencía, and Frederik Janssen

Technische Universität Darmstadt
Knowledge Engineering Group
j.zilke@mail.de, {eneldo,janssen}@ke.tu-darmstadt.de

Extracting Rules from Artificial Neural Networks with Distributed Representations

Sebastian Thrun
University of Bonn
Department of Computer Science III
Römerstr. 164, D-53117 Bonn, Germany
E-mail: thrun@carbon.informatik.uni-bonn.de

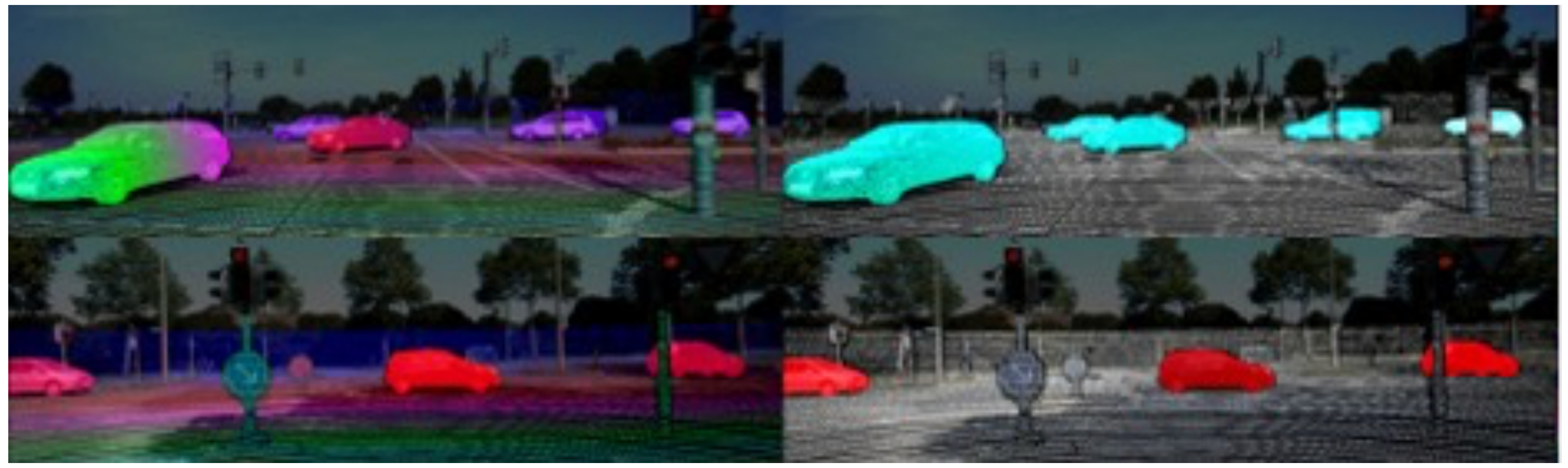
"Why Should I Trust You?"
Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

Examples of Processing Methods

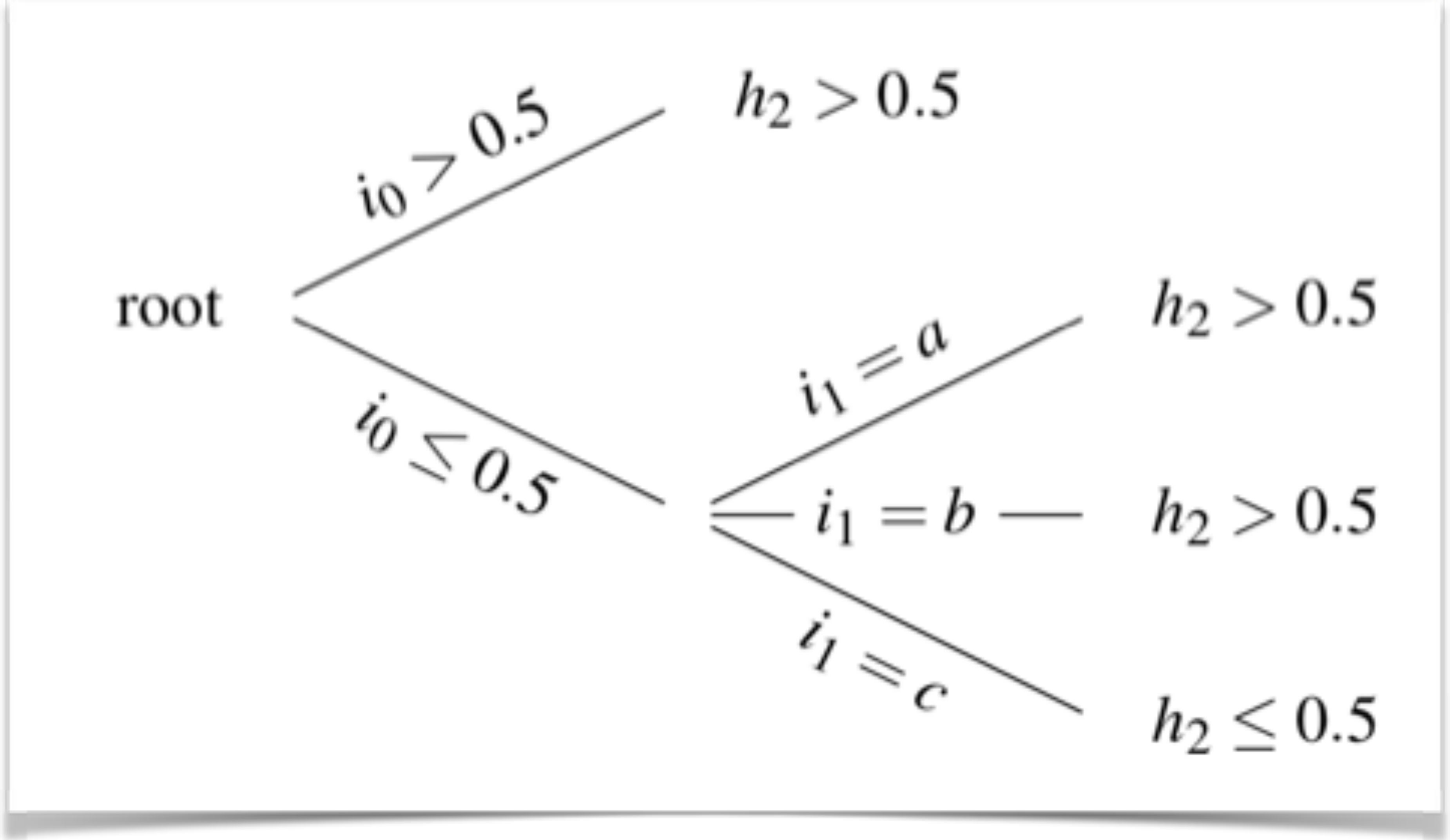


Geiger, Andreas, Philip Lenz, and Raquel Urtasun. "Are we ready for autonomous driving? The kitti vision benchmark suite." *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012.

DeepRED – Rule Extraction from Deep Neural Networks*

Jan Ruben Zilke, Eneldo Loza Mencía, and Frederik Janssen

Technische Universität Darmstadt
Knowledge Engineering Group
j.zilke@mail.de, {eneldo,janssen}@ke.tu-darmstadt.de



Zilke, Jan Ruben et al. "DeepRED - Rule Extraction from Deep Neural Networks." *DS* (2016).

Taxonomy

	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 representations

Methods that Explain Representations

Network Dissection:
Quantifying Interpretability of Deep Visual Representations

David Bau*, Bolei Zhou*, Aditya Khosla, Aude Oliva, and Antonio Torralba
CSAIL, MIT
{davidbau, bzhou, khosla, oliva, torralba}@csail.mit.edu

Interpretability Beyond Feature Attribution:
Quantitative Testing with Concept Activation Vectors (TCAV)

Been Kim   Martin Wattenberg   Justin Gilmer   Carrie Cai   James Wexler
Fernanda Viegas   Rory Sayres

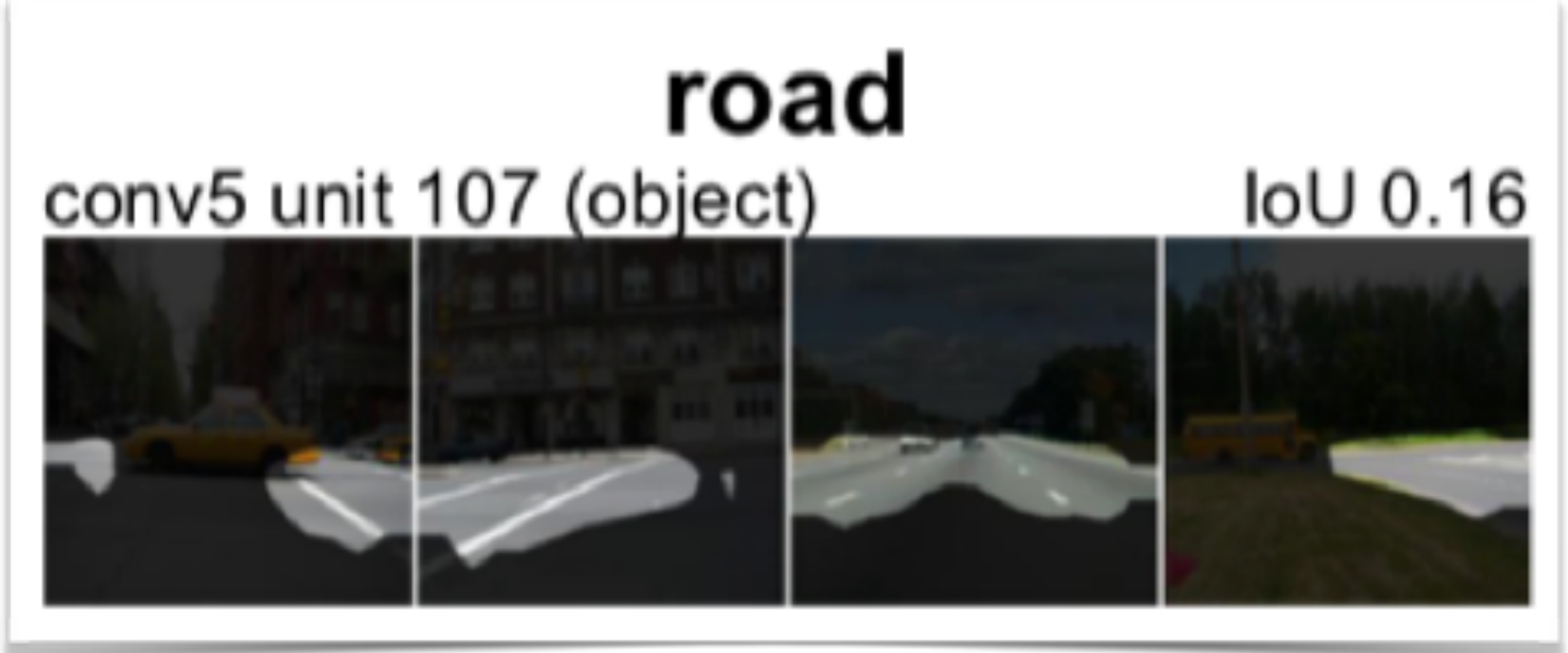
CNN Features off-the-shelf: an Astounding Baseline for Recognition

Ali Sharif Razavian   Hossein Azizpour   Josephine Sullivan   Stefan Carlsson
CVAP, KTH (Royal Institute of Technology)
Stockholm, Sweden
{razavian,azizpour,sullivan,stefanc}@csc.kth.se

Examples of Explained Representations

Network Dissection:
Quantifying Interpretability of Deep Visual Representations

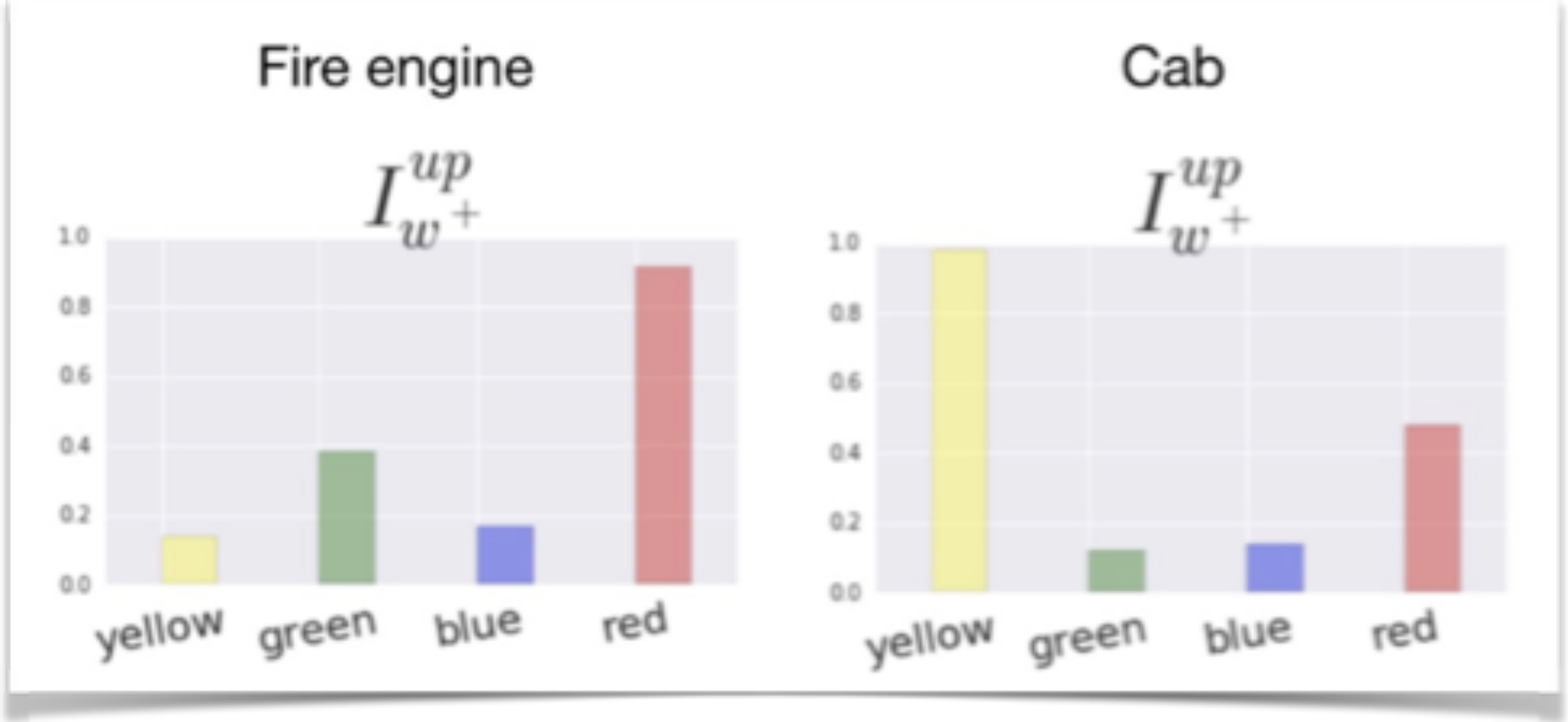
David Bau*, Bolei Zhou*, Aditya Khosla, Aude Oliva, and Antonio Torralba
CSAIL, MIT
{davidbau, bzhou, khosla, oliva, torralba}@csail.mit.edu



D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Computer Vision and Pattern Recognition*, 2017.

Interpretability Beyond Feature Attribution:
Quantitative Testing with Concept Activation Vectors (TCAV)

Been Kim Martin Wattenberg Justin Gilmer Carrie Cai James Wexler
Fernanda Viegas Rory Sayres



Kim, Been, et al. "Tcav: Relative concept importance testing with linear concept activation vectors." *arXiv preprint arXiv:1711.11279* (2017).

Taxonomy

	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 representations

Methods that Produce Explanations

Multimodal Explanations: Justifying Decisions and Pointing to the Evidence

Dong Huk Park[1], Lisa Anne Hendricks[1], Zeynep Akata[2,3], Anna Rohrbach[1,3], Bernt Schiele[3], Trevor Darrell[1], and Marcus Rohrbach[4]

[1]EECS, UC Berkeley, [2]University of Amsterdam, [3]MPI for Informatics, [4]Facebook AI Research

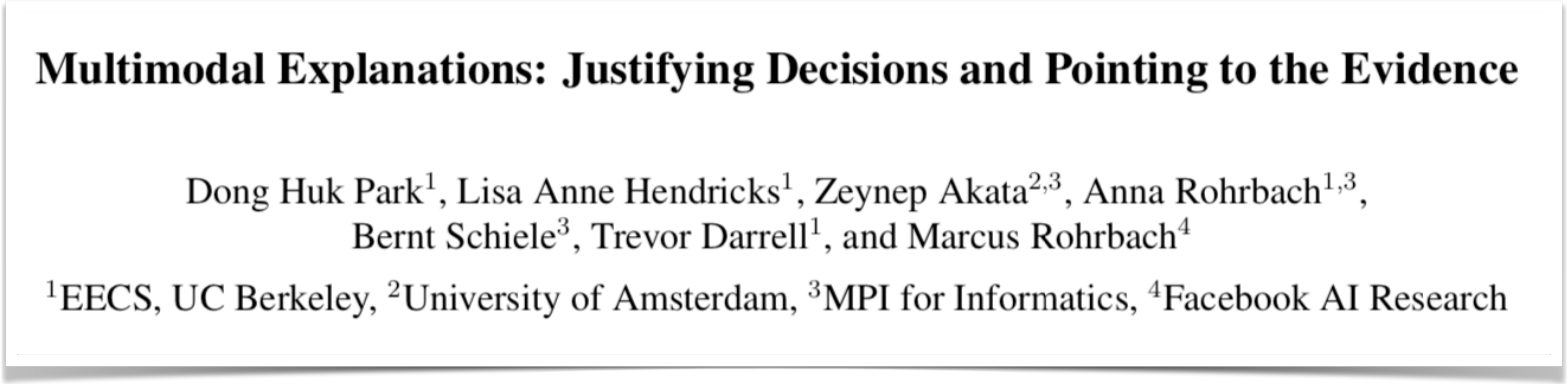
Hierarchical Question-Image Co-Attention for Visual Question Answering

Jiasen Lu[*], Jianwei Yang[*], Dhruv Batra[*†], Devi Parikh[*†]
∗ Virginia Tech, † Georgia Institute of Technology
{jiasenlu, jw2yang, dbatra, parikh}@vt.edu

InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets

Xi Chen[†‡], Yan Duan[†‡], Rein Houthooft[†‡], John Schulman[†‡], Ilya Sutskever[‡], Pieter Abbeel[†‡]
† UC Berkeley, Department of Electrical Engineering and Computer Sciences
‡ OpenAI

Examples that Produce Explanations

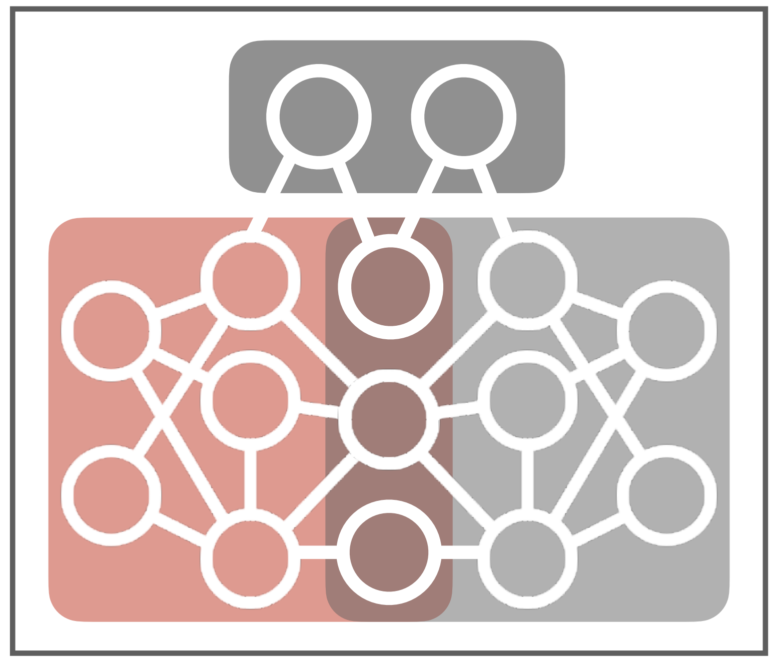

Multimodal Explanations: Justifying Decisions and Pointing to the Evidence

Dong Huk Park[1], Lisa Anne Hendricks[1], Zeynep Akata[2,3], Anna Rohrbach[1,3], Bernt Schiele[3], Trevor Darrell[1], and Marcus Rohrbach[4]

[1]EECS, UC Berkeley, [2]University of Amsterdam, [3]MPI for Informatics, [4]Facebook AI Research

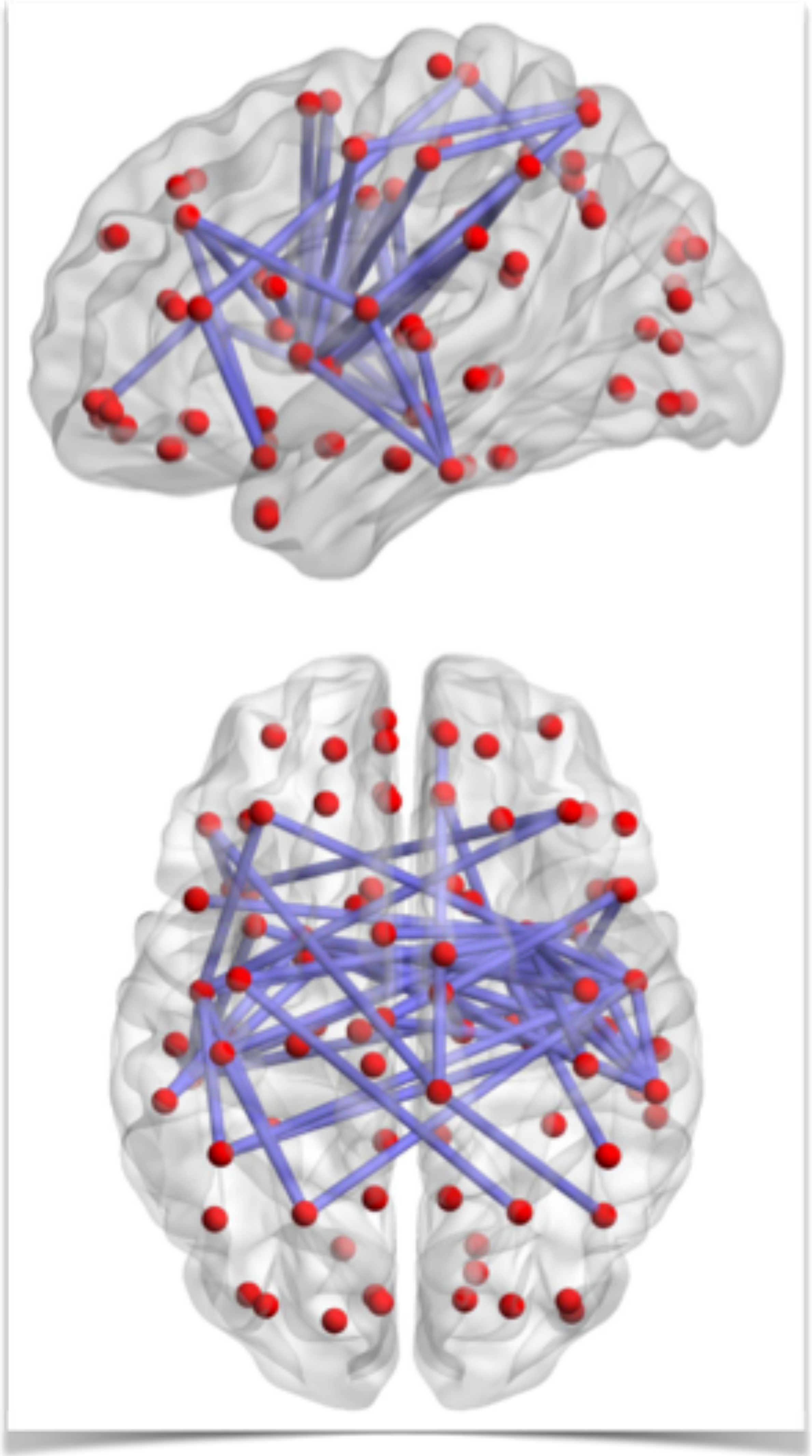
Park, Dong Huk, et al. "Multimodal Explanations: Justifying Decisions and Pointing to the Evidence." *31st IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

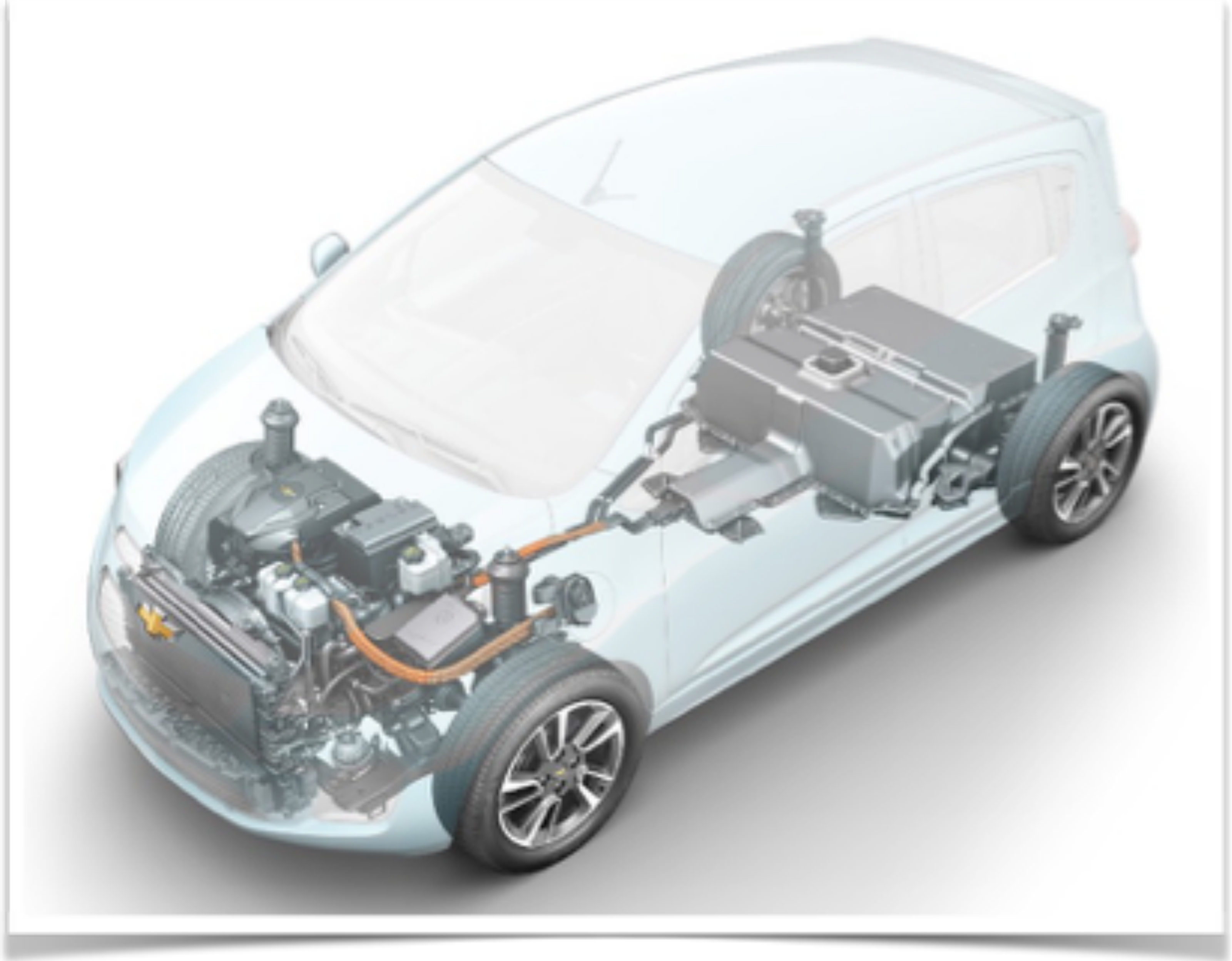
[1] L.H. Gilpin. Explaining possible futures for robust autonomous decision-making. Proceedings of the AAAI Fall Symposium on Anticipatory Thinking, 2019.
[2] L.H. Gilpin, V. Penubarthi, and L. Kagal. Explaining Multimodal Errors in Autonomous Vehicles. DSAA 2021.

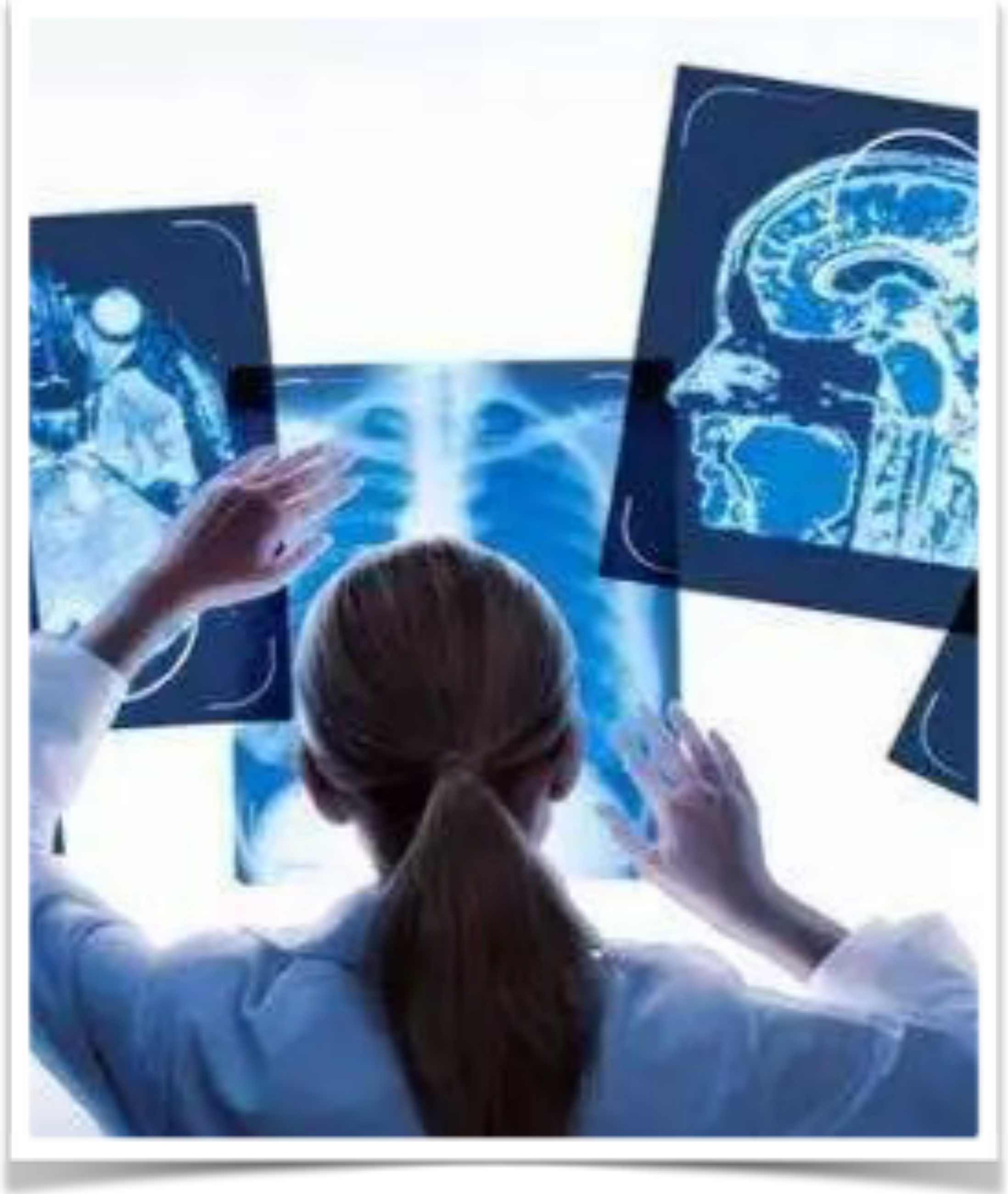


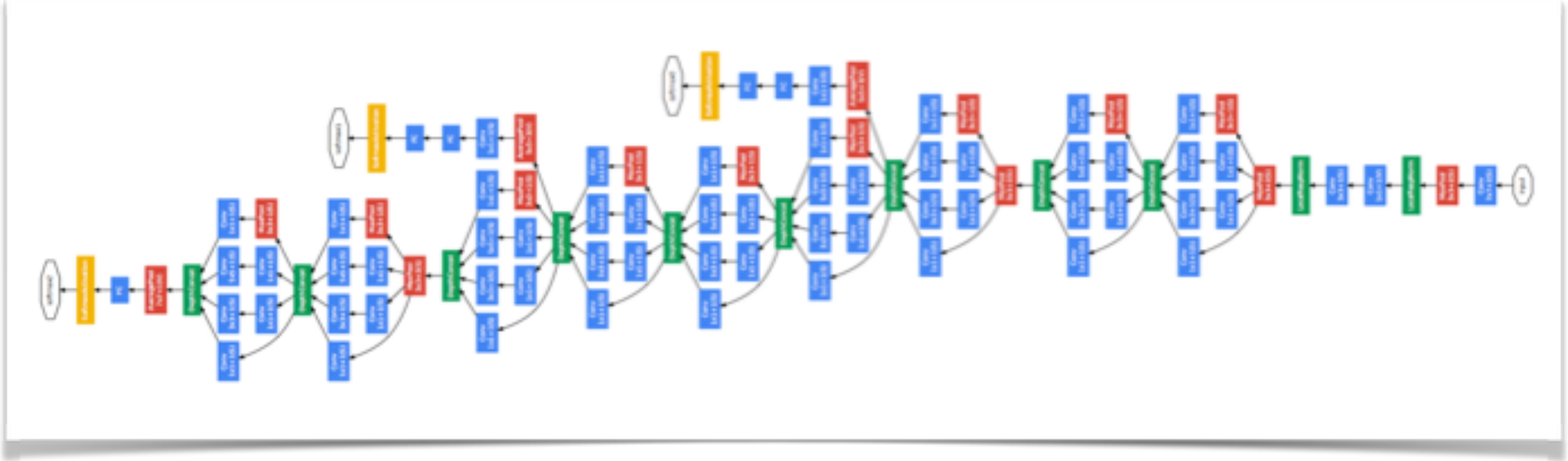
The best option is to veer and slow down. The vehicle is traveling too fast to suddenly stop. The vision system is inconsistent, but the lidar system has provided a reasonable and strong claim to avoid the object moving across the street.

A Problem: Insides Matter





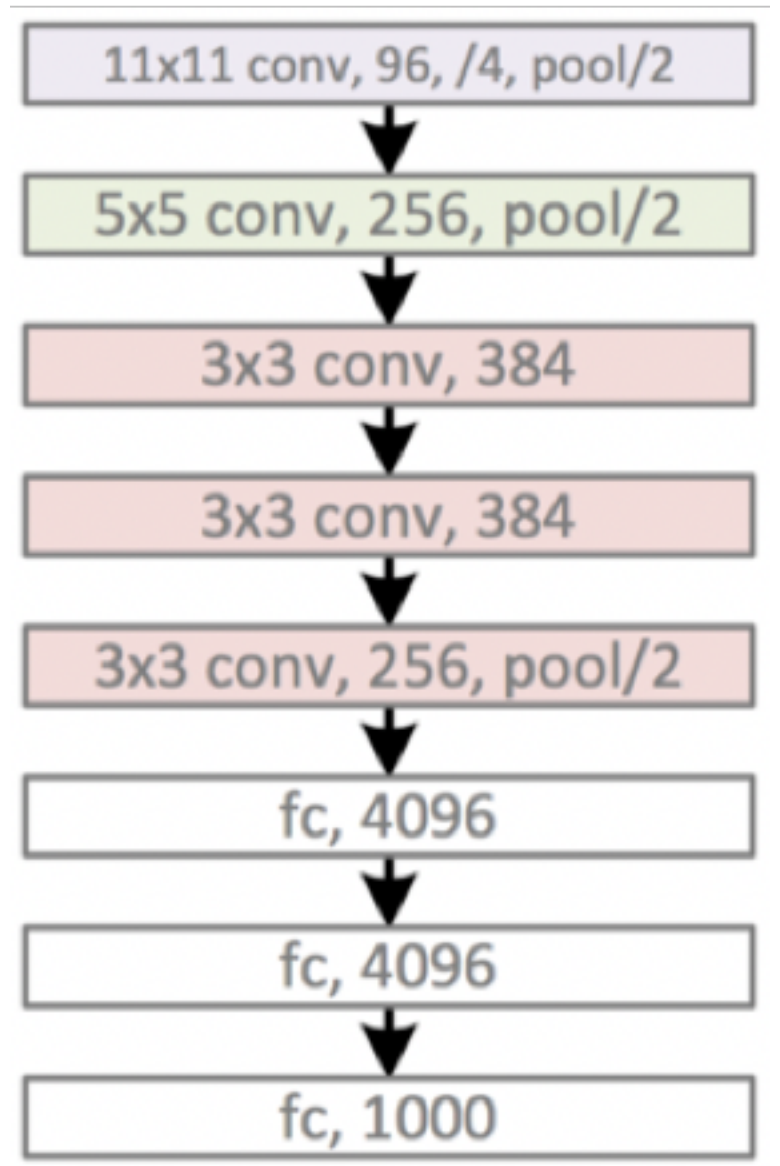




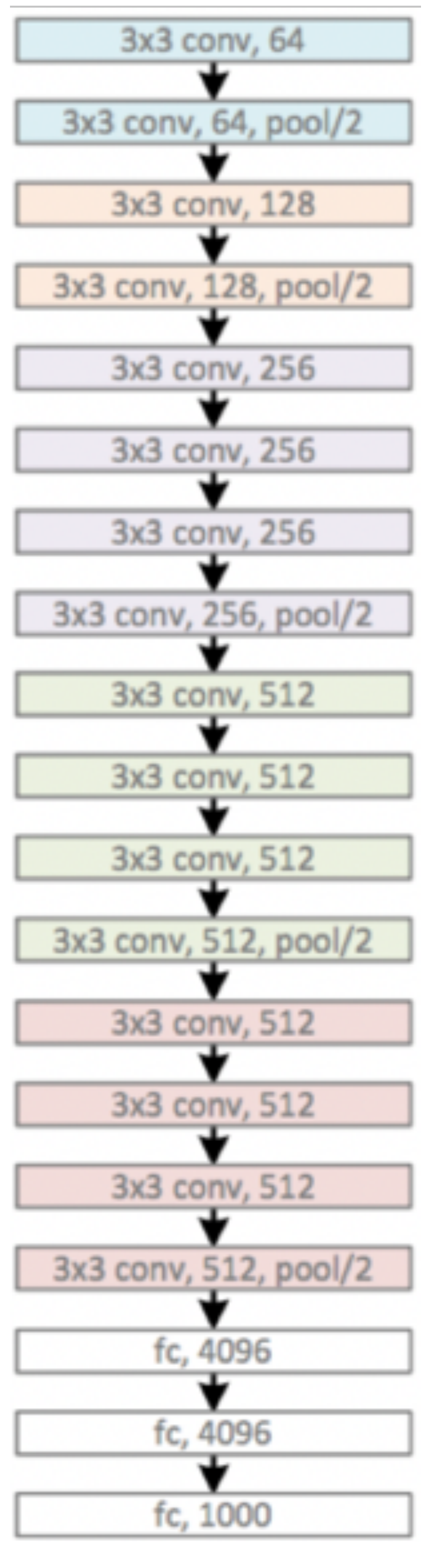
The More Complex (Deeper)

The Deeper the Mystery

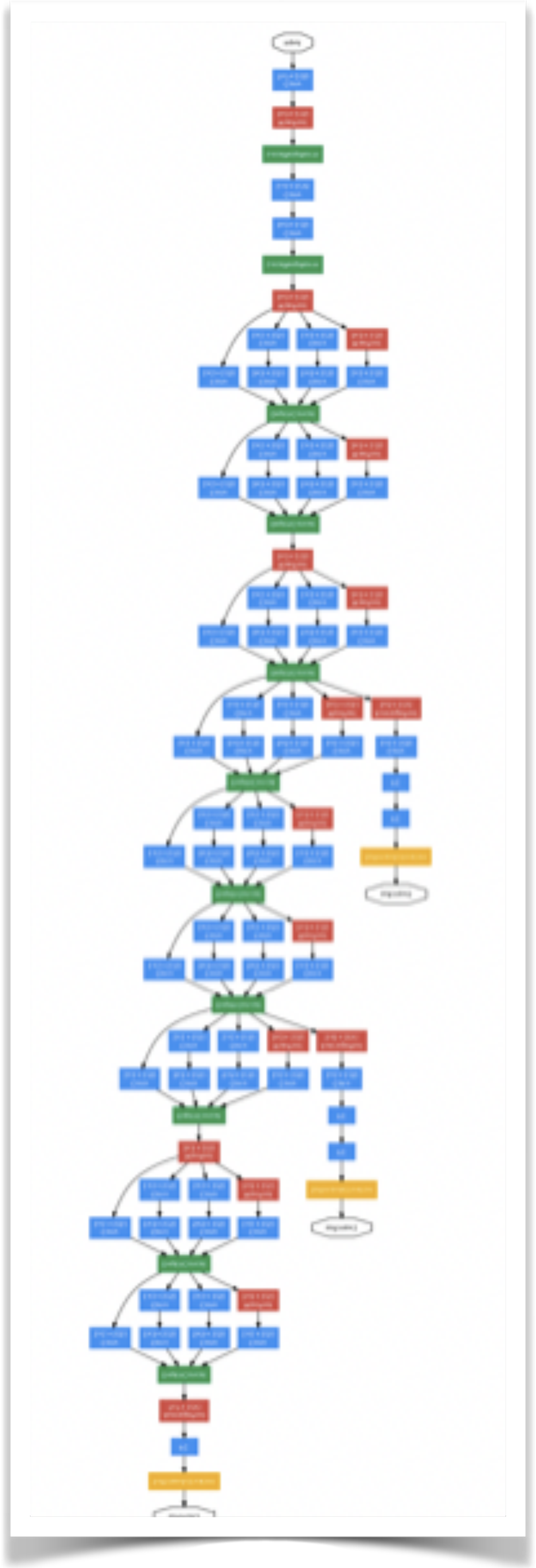
IMAGENET



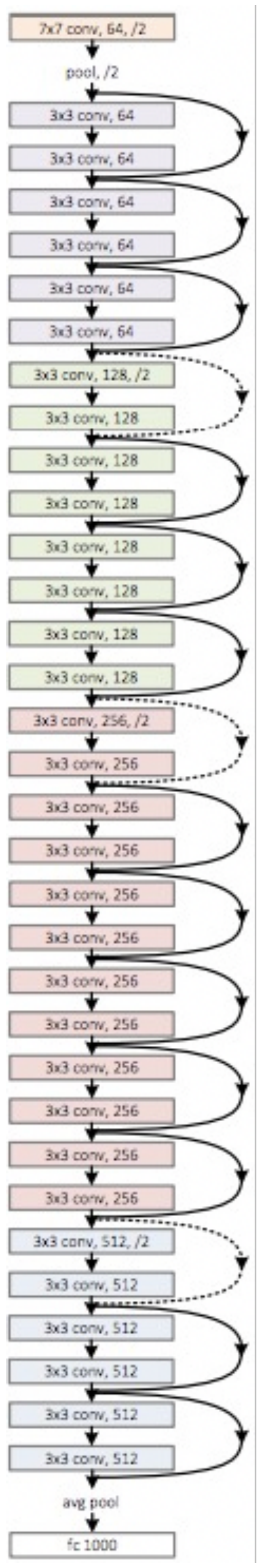
AlexNet (2012)
8 layers; acc 84.7%



VGG (2014)
19 layers; acc 91.5%



GoogLeNet (2015)
22 layers; acc 92.2%



ResNet (2016)
152 layers; acc 95.6%

Talk Agenda

Motivate problem: Systems are imperfect

What is explainability?

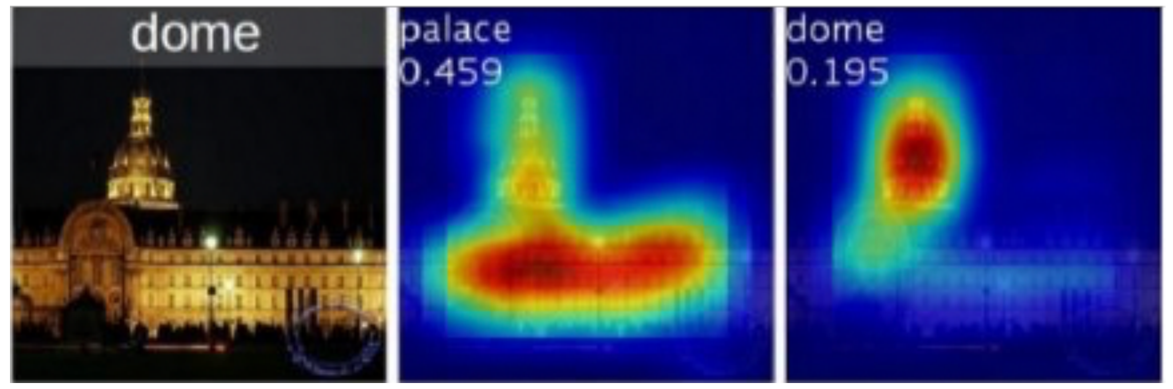
What is *actually* being explained?

<u>How to evaluate explainability?</u>

How to explain complex systems? (autonomous driving)

What is Being Explained?

Visual cues

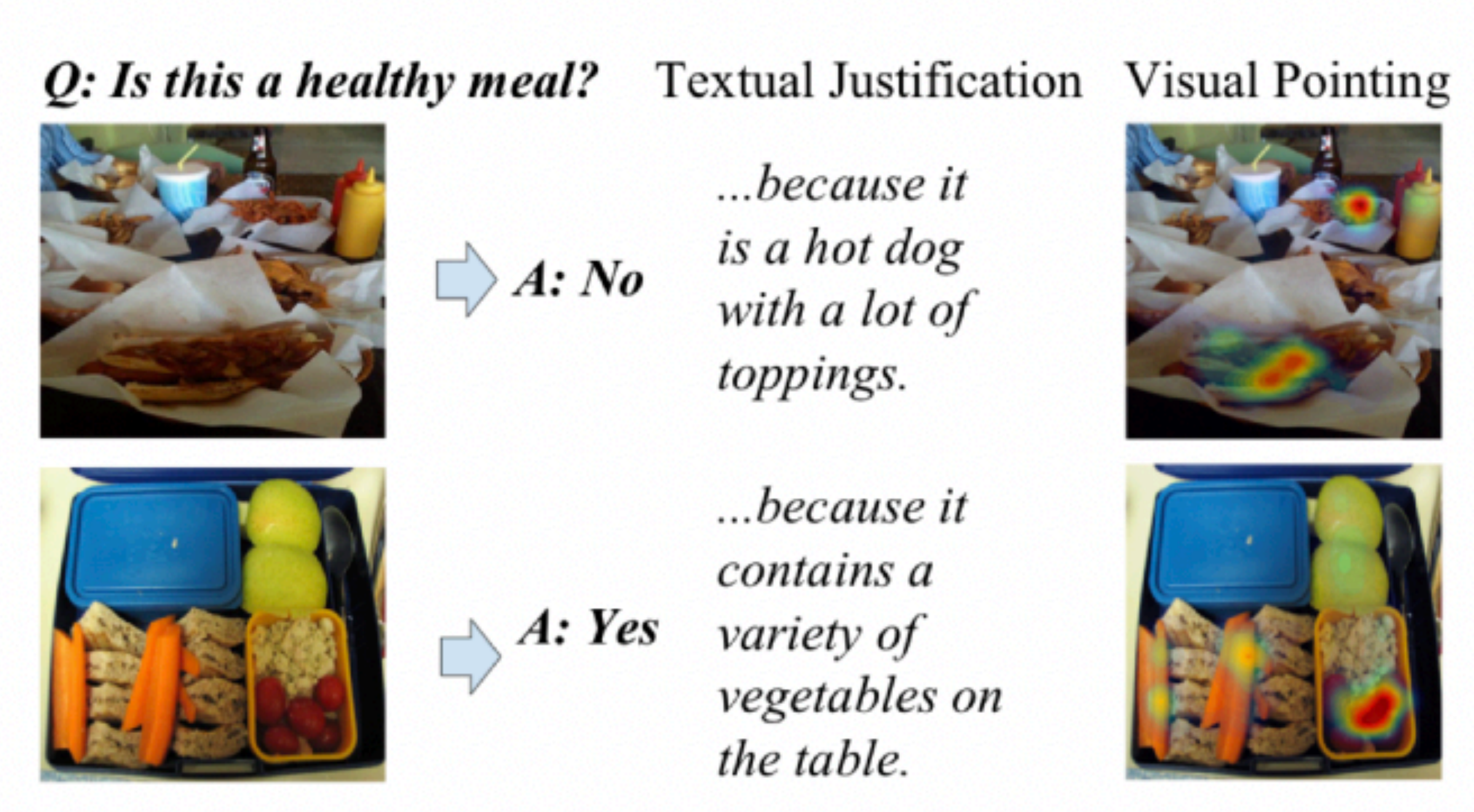


Completeness to model

Role of individual units

Completeness on other tasks

Attention based



Human evaluation

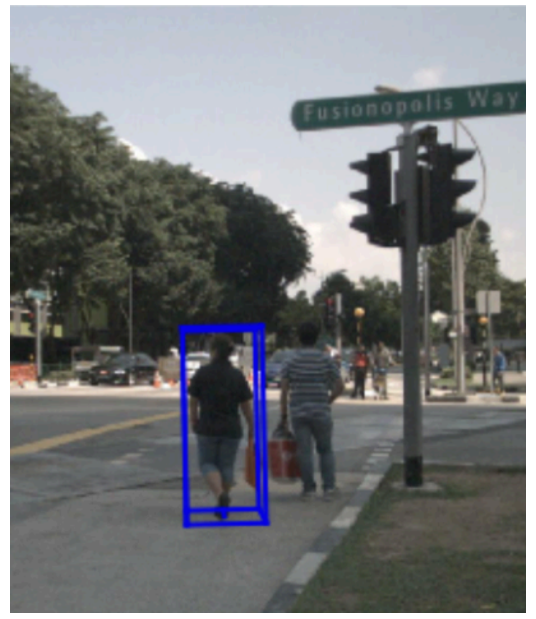
Taxonomy

	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 representations

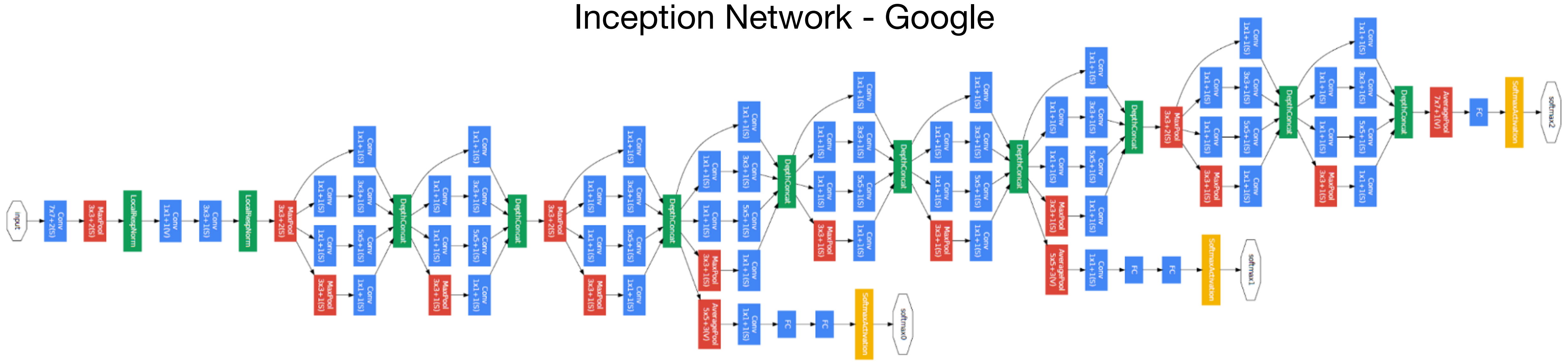
Challenges in Explainability

- Standards and metrics for explanations
    - How to **evaluate** explanations?
- Current metrics of evaluation are “fuzzy”
    - User based evaluations are not *always* appropriate
- Benchmarks for safety-critical and mission-critical tasks.

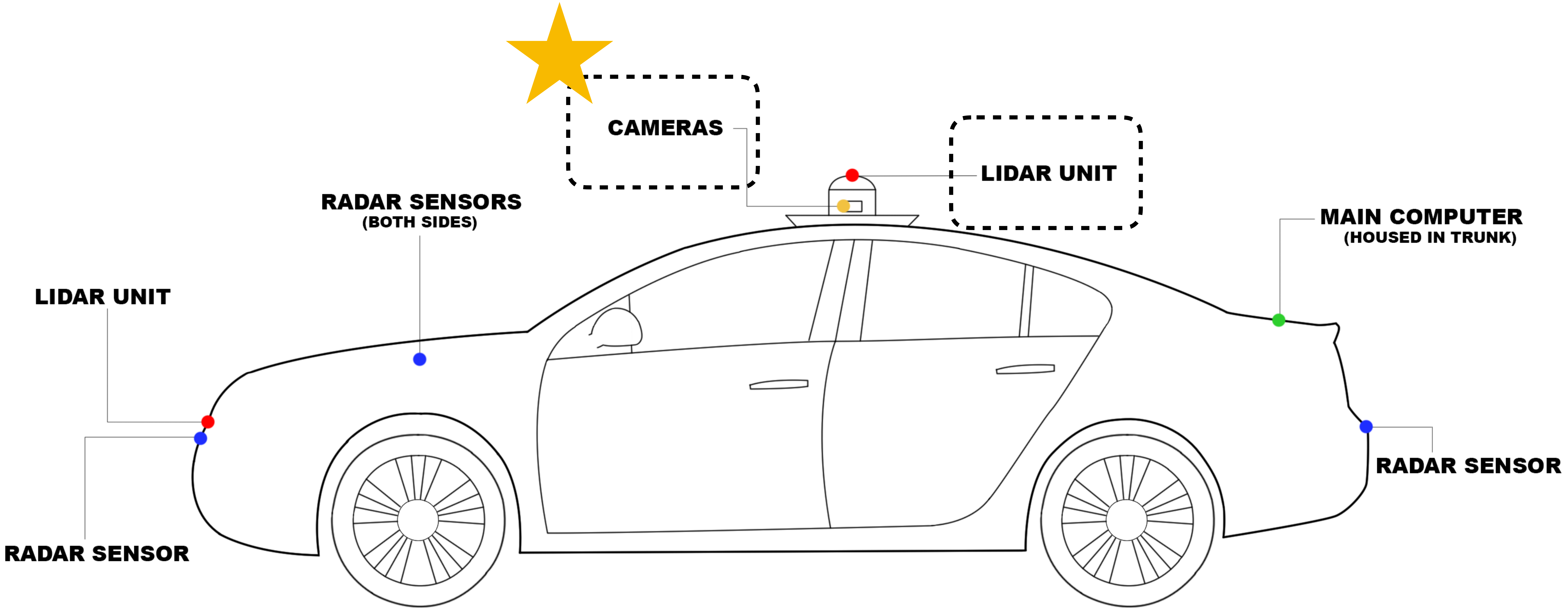
A Neural Network Labels Camera Data



Inception Network - Google

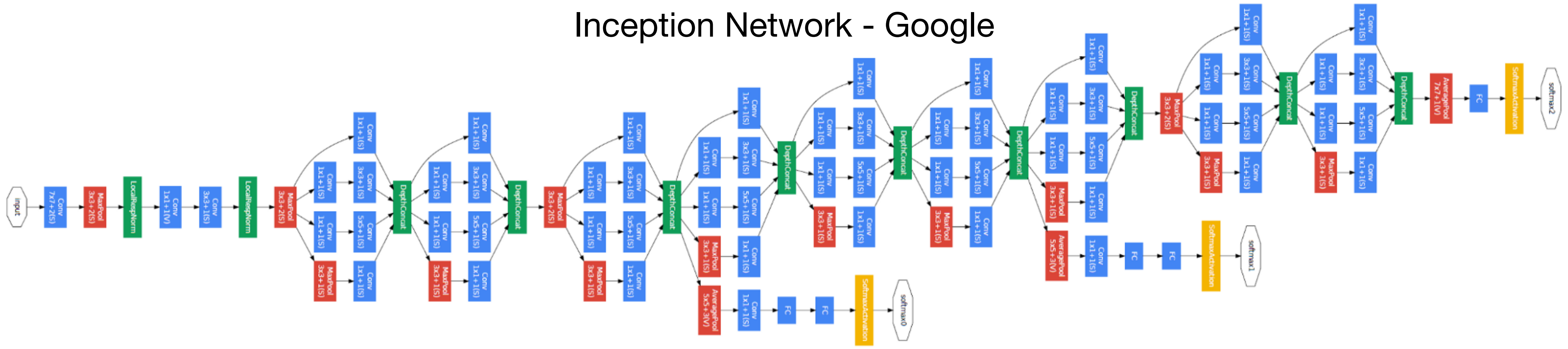


Label
e.g. pedestrian



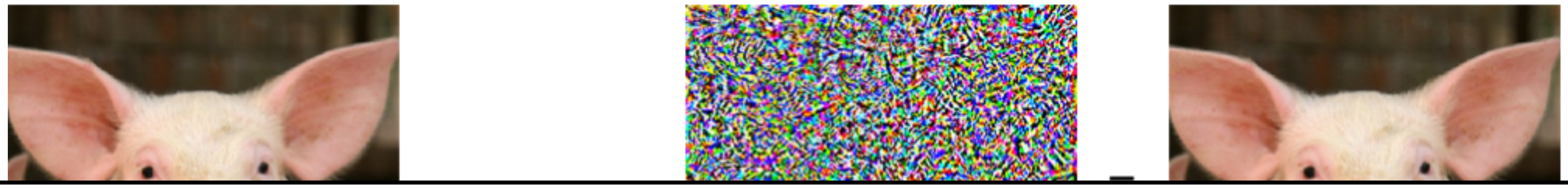
Problem: Neural Networks are Brittle

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.”

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

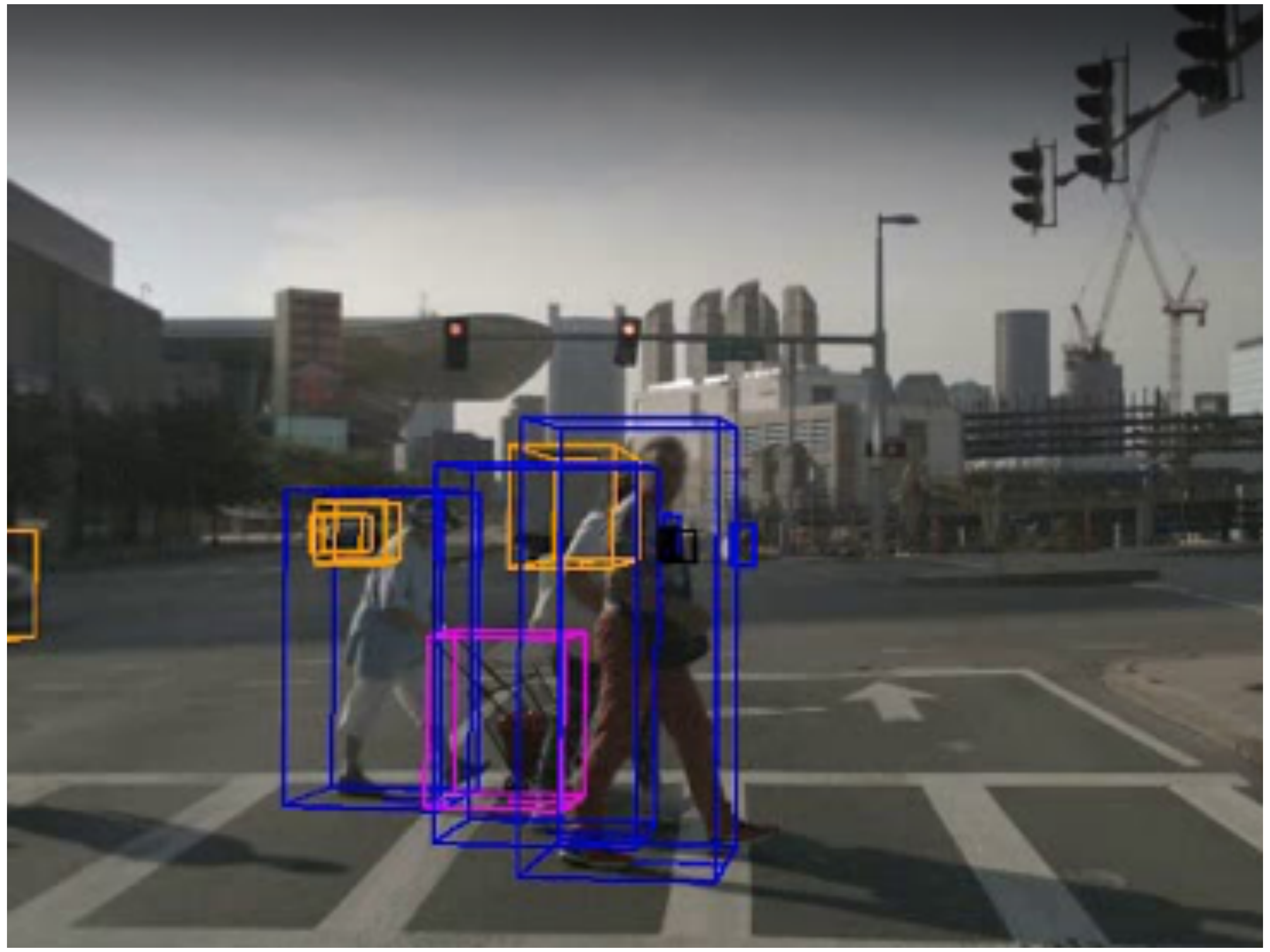
Anticipatory Thinking Layer for Error Detection

“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

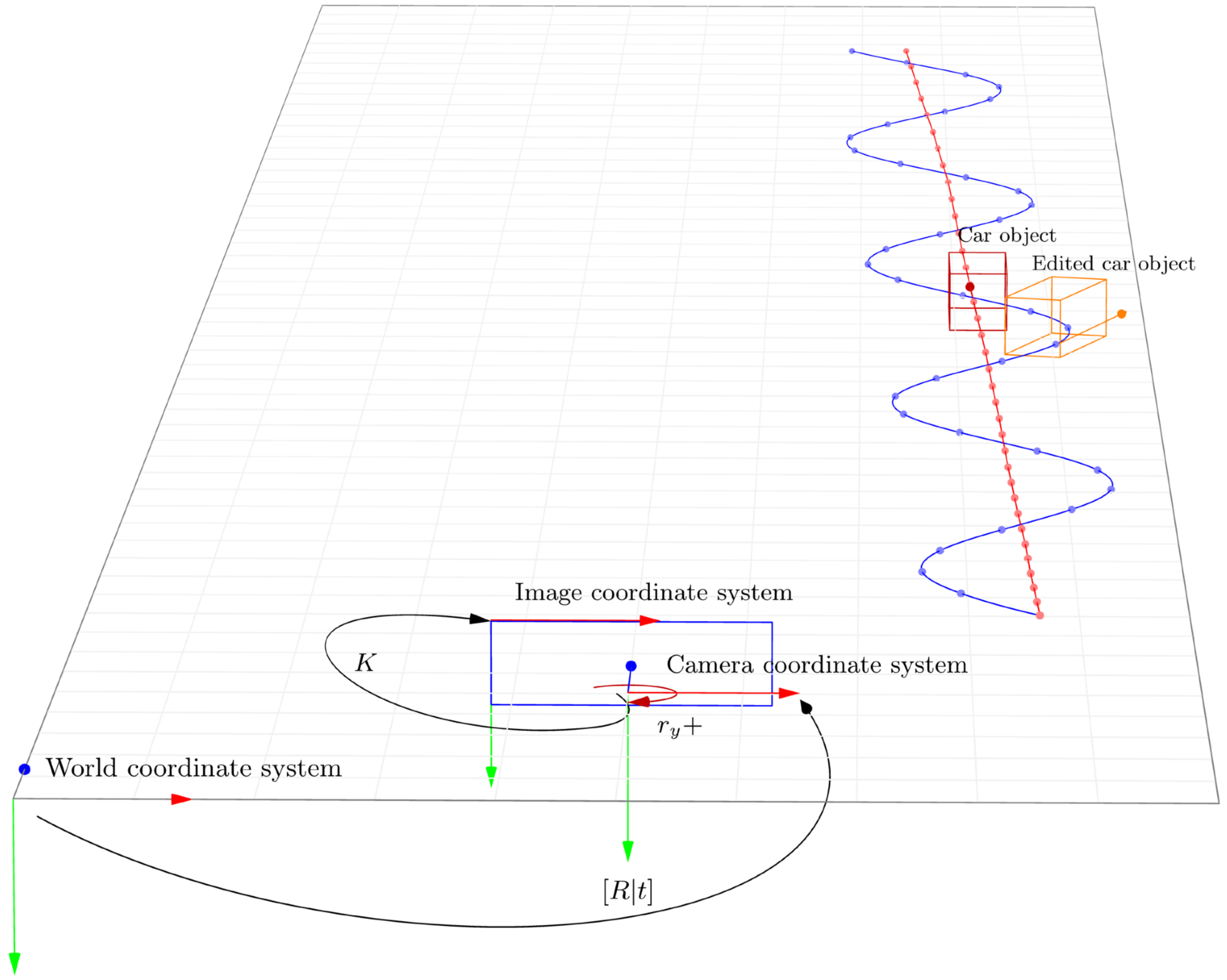
Approach: Content Generation

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 Layer for Error Detection



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

Brief Intro

<u>Motivate problem: Systems are imperfect</u>

What is explainability?

What is *actually* being explained?

How to evaluate explainability?

How to critical systems? (autonomous driving)