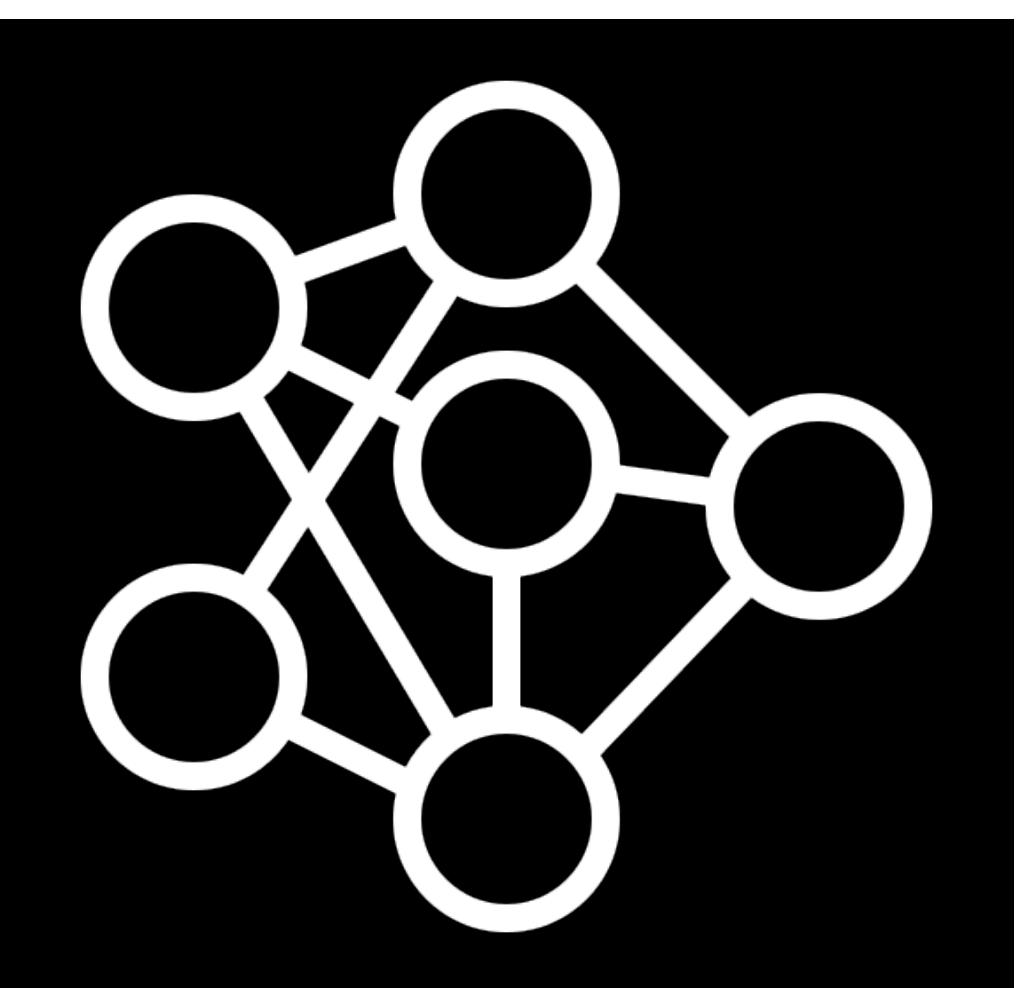
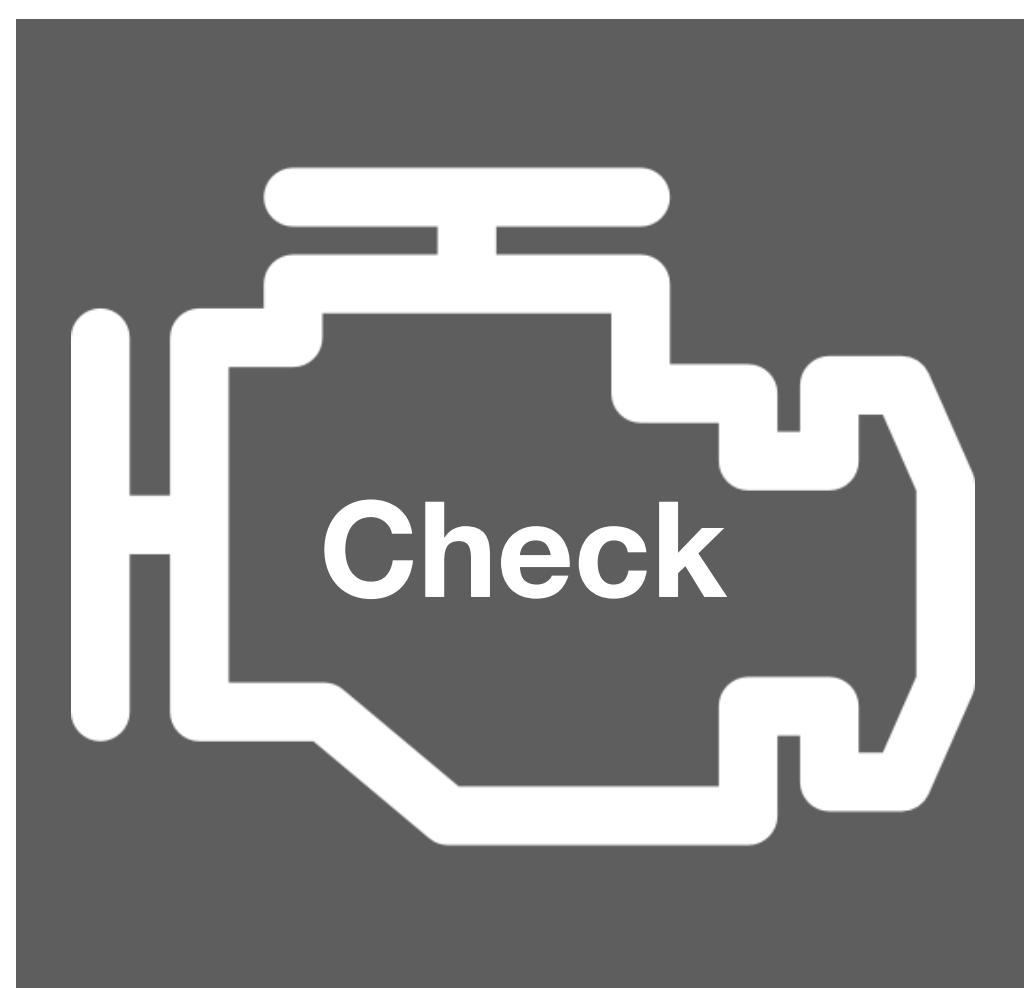


# Explaining Explanations

*Black-box*



*Imprecise*



*System-level*



Leilani H. Gilpin - MIT

# The Need for Explanations



No Explanation

## AI Mistakes Bus-Side Ad for Famous CEO, Charges Her With Jaywalking

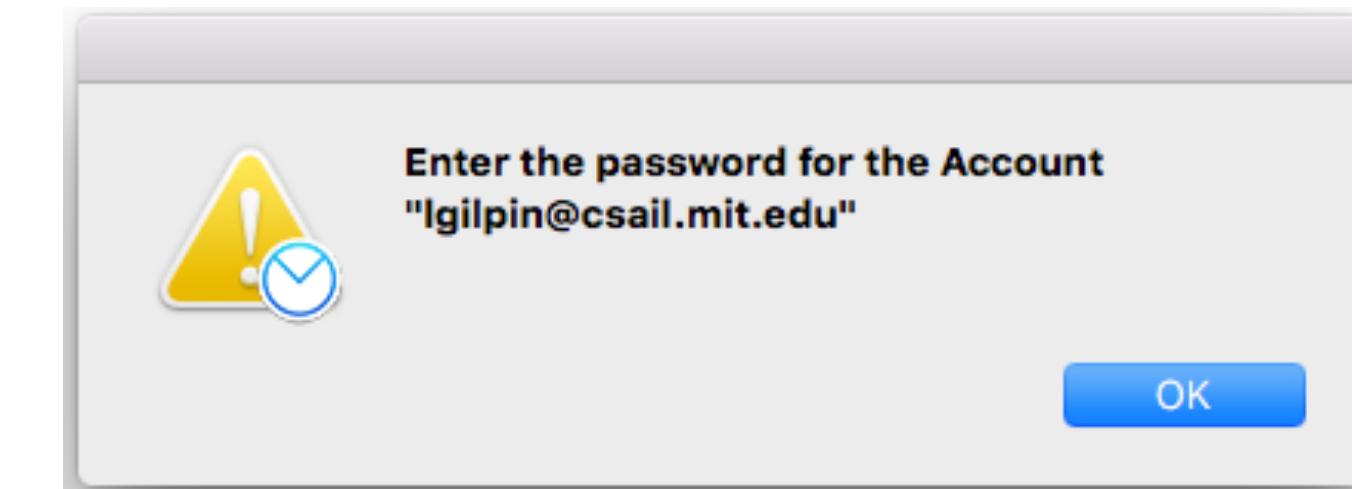
By Tang Ziyi / Nov 22, 2018 04:17 PM / Society & Culture



No Commonsense

```
Igilpin — bash — 80
Last login: Tue Feb 7 15:37:57 on ttys000
30-9-198:~ lgilpin$ sudo mkdir /usr/bin/jemdoc
Password:
mkdir: /usr/bin/jemdoc: Operation not permitted
30-9-198:~ lgilpin$
```

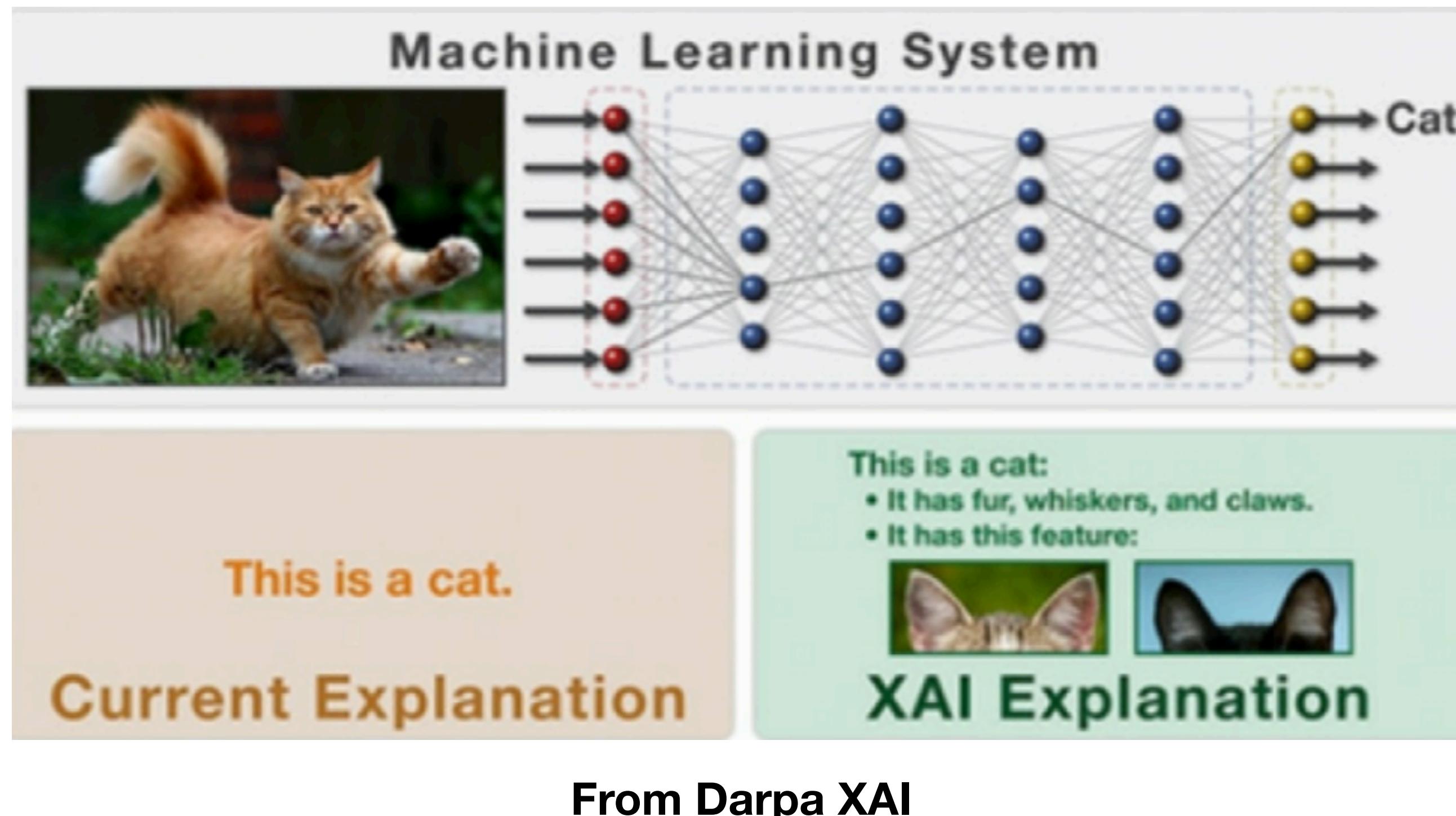
OS Upgrade (Version Skew)



Imprecise (Certificate Missing)

Users Need Explanations

# What is Explainability?



**“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*

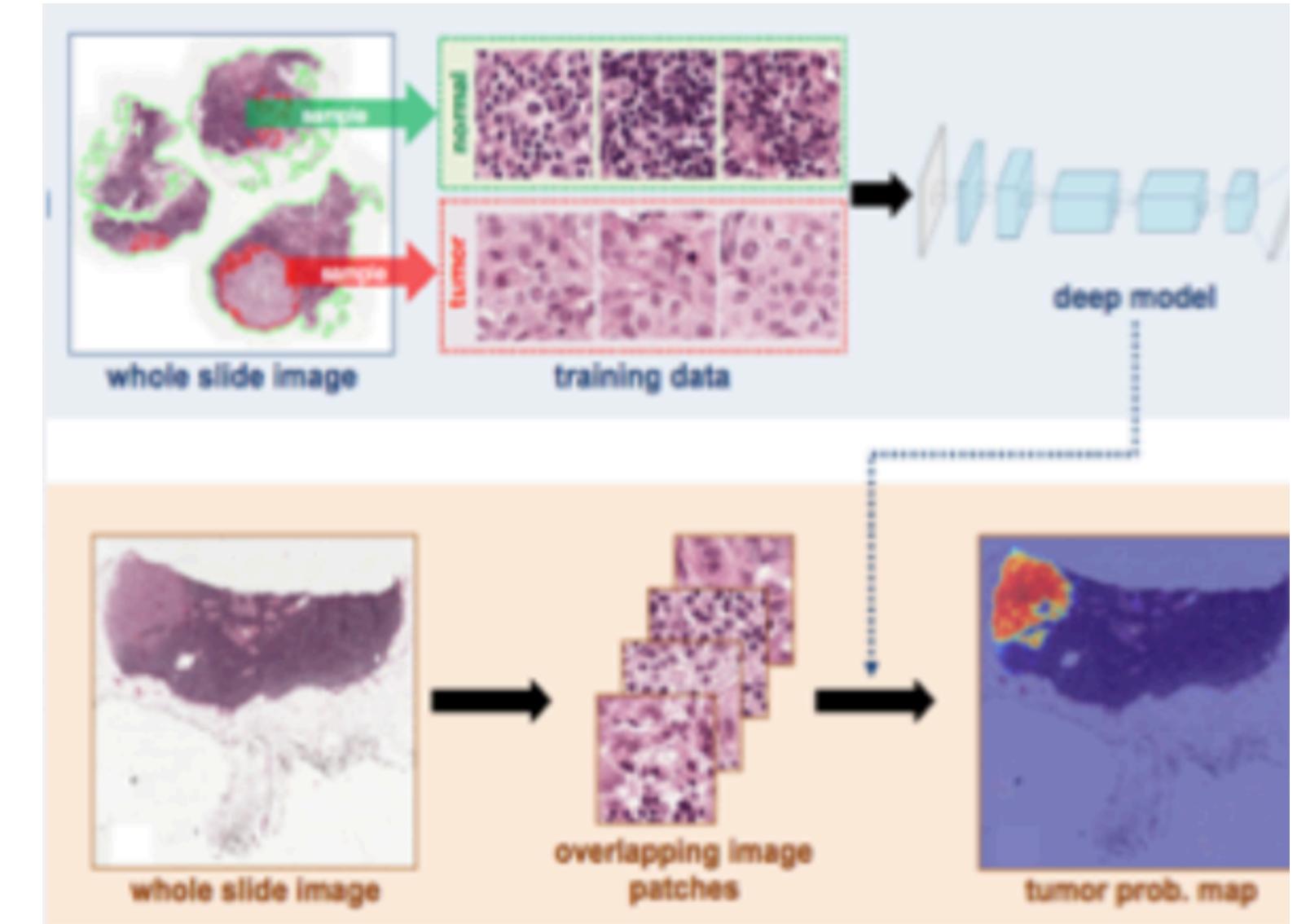
# Deep Nets are Everywhere



Self-driving Cars

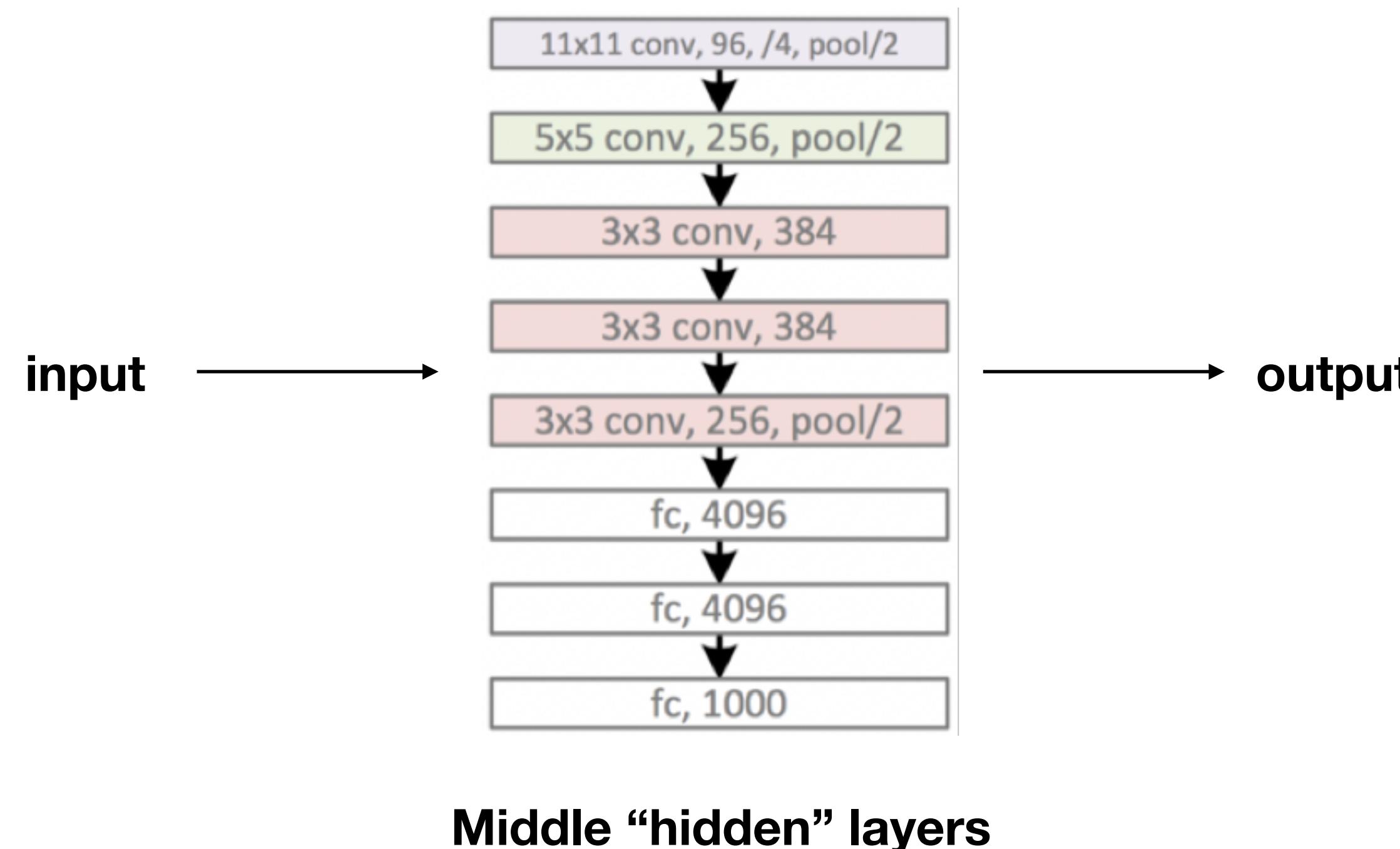


Playing Go



Making Medical Decisions

# Deep Nets are Not Understandable



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

# Overview

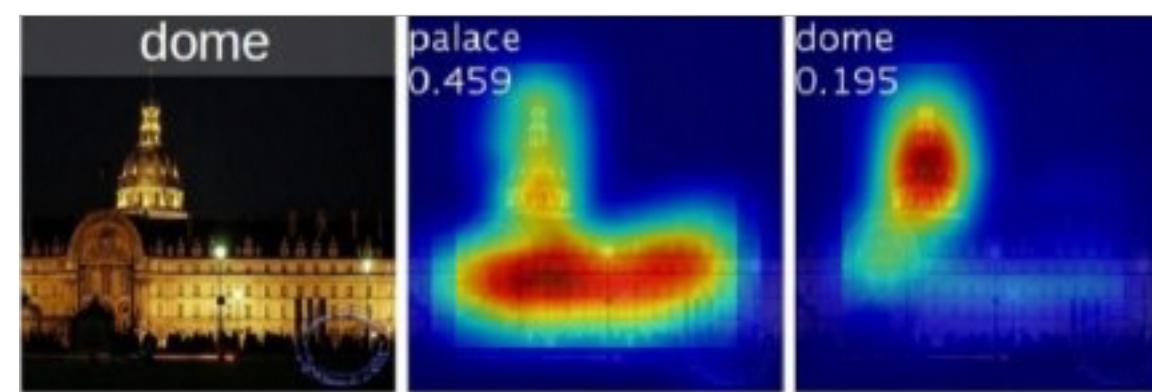
- Definitions
- Taxonomy
  - Survey: Literature review (87 papers) in computer science, artificial intelligence, and philosophy.
  - Recommendations for Evaluation
- How can Explanations Help (e.g. anomaly detection).
- Contributions and Future Work

# Definitions

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

# What we Have

## Visual cues



Interpretable,  
not complete

## Role of individual units



Complete,  
not interpretable

## Attention based

<i>Q: Is this a healthy meal?</i>	Textual Justification	Visual Pointing
<i>A: No</i>	<i>...because it is a hot dog with a lot of toppings.</i>	
<i>A: Yes</i>	<i>...because it contains a variety of vegetables on the table.</i>	

Interpretable,  
not complete

# Why this Matters

## Interpretability

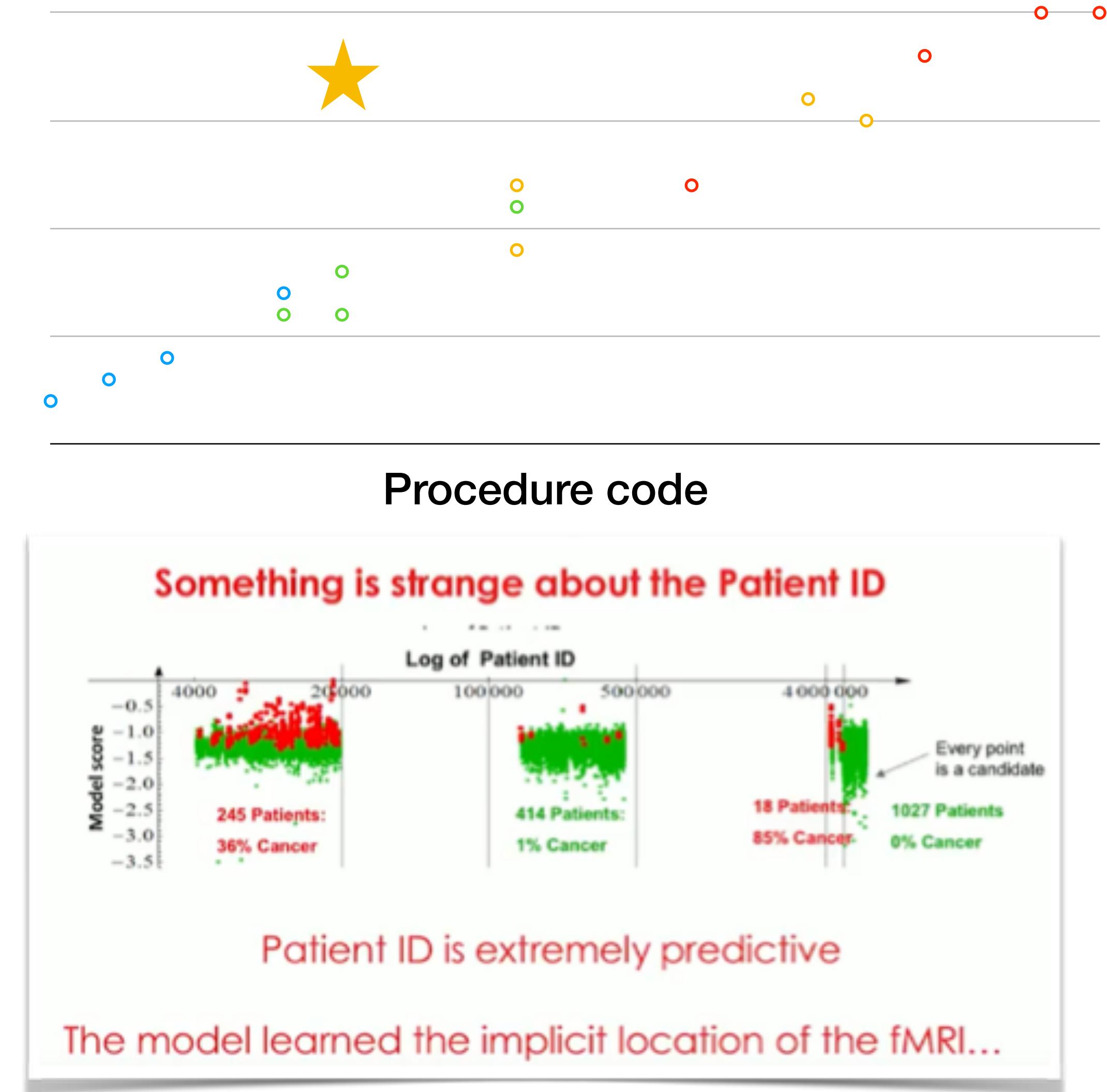
- GDPR
- Liability for decision making



# Why this Matters

## Completeness

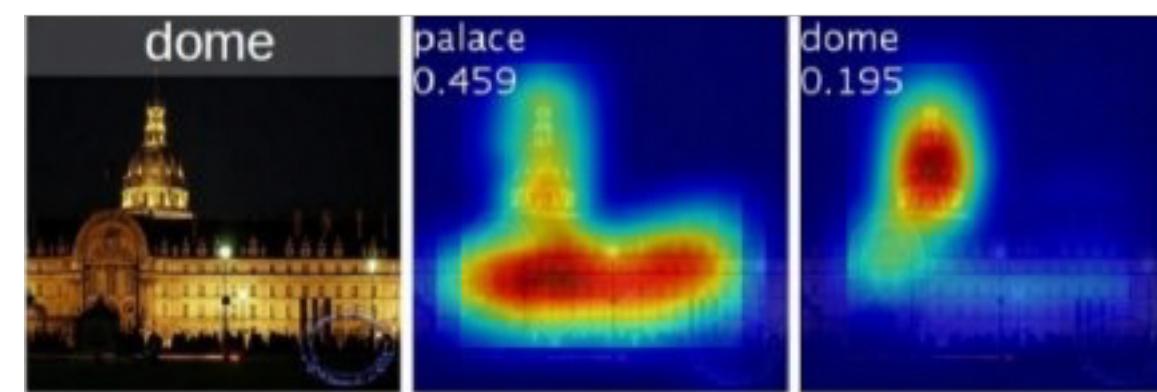
- Explaining the wrong thing.
- Making decisions for the wrong reasons.



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

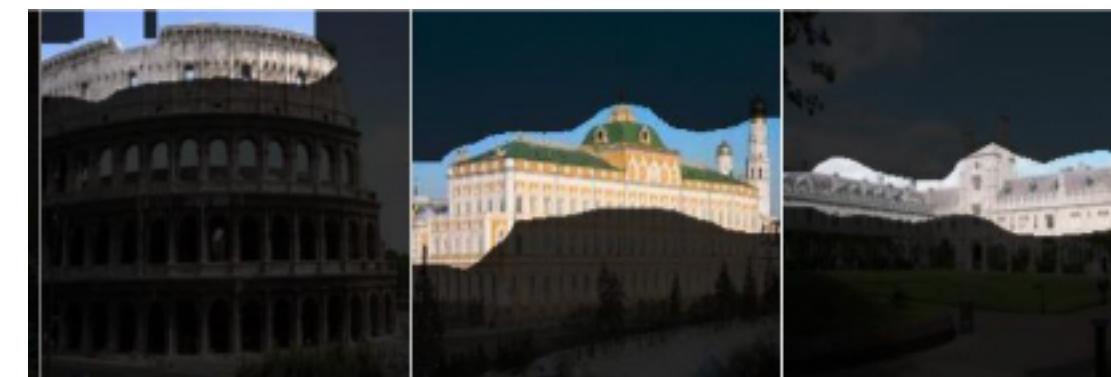
# What is Being Explained?

Visual cues



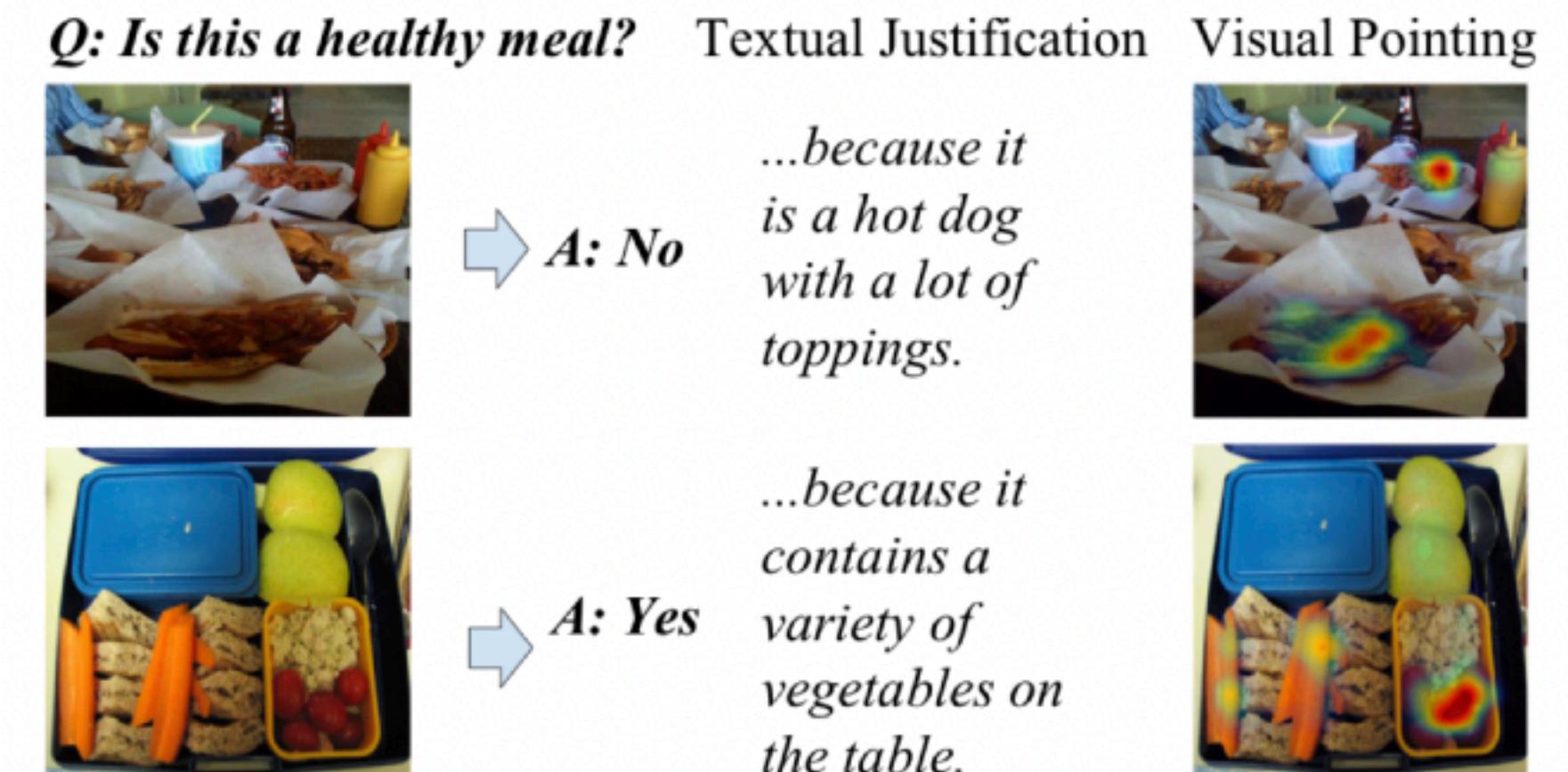
Explain processing

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

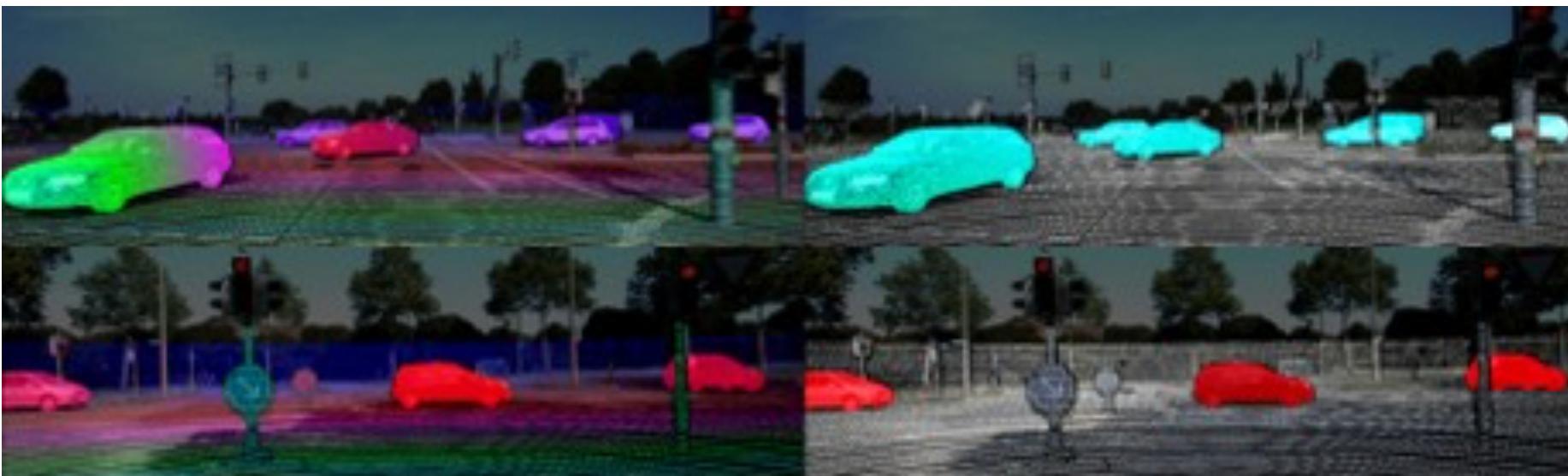
## “Why Should I Trust You?” Explaining the Predictions of Any Classifier

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.*

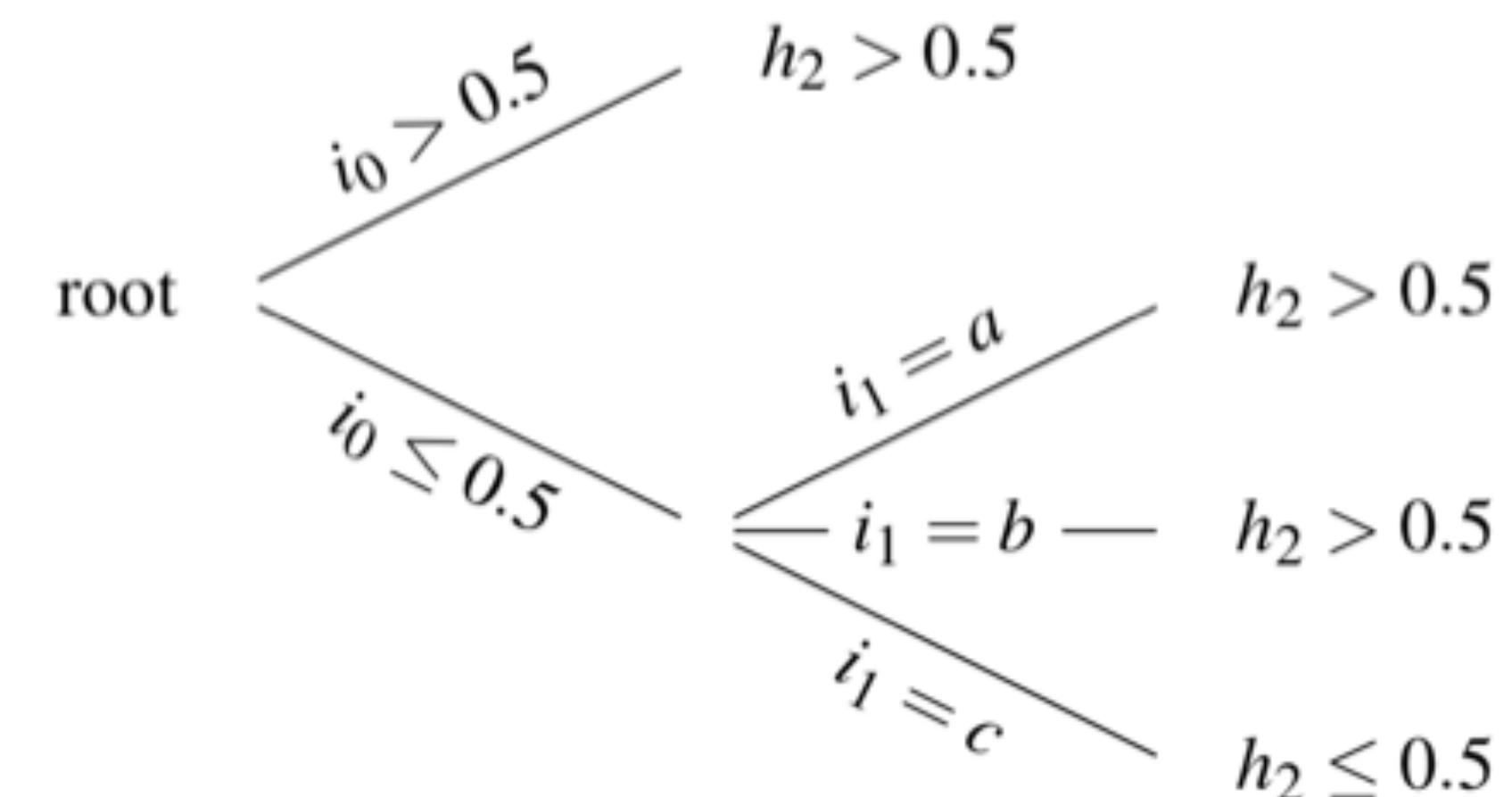
## DeepRED – Rule Extraction from Deep Neural Networks\*

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

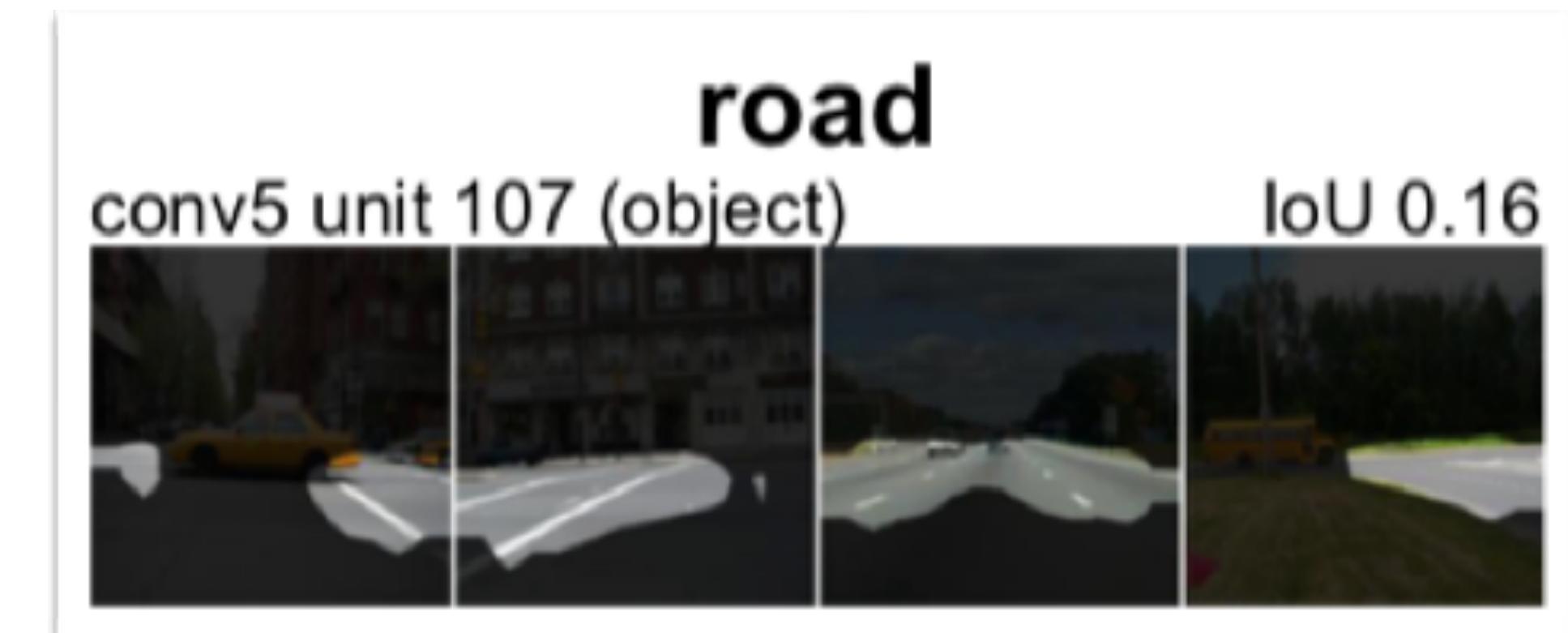
## **CNN Features off-the-shelf: an Astounding Baseline for Recognition**

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

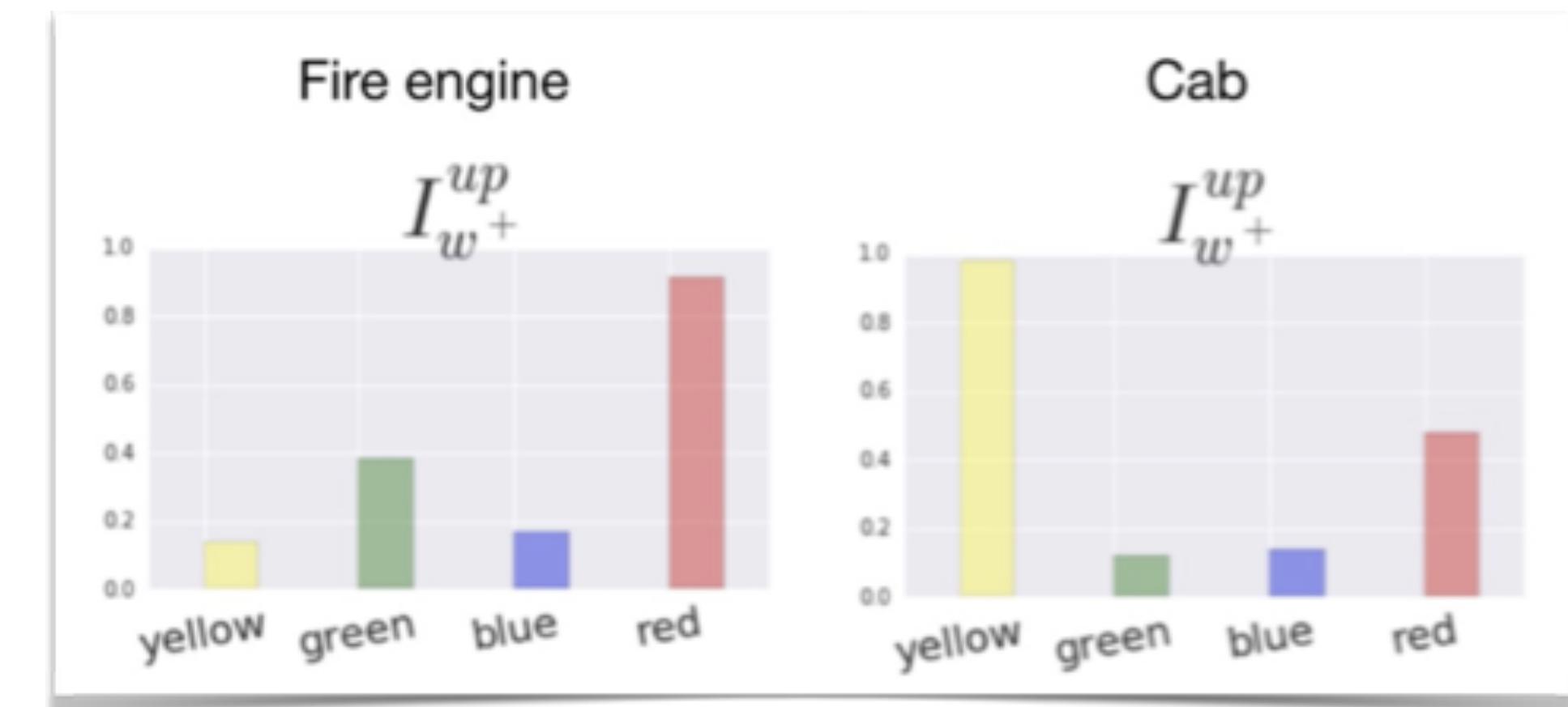
David Bau\*, Bolei Zhou\*, Aditya Khosla, Aude Oliva, and Antonio Torralba  
CSAIL, MIT  
`{davidbau, bzhou, khosla, oliva, torralba}@csail.mit.edu`



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<sup>1</sup>, Lisa Anne Hendricks<sup>1</sup>, Zeynep Akata<sup>2,3</sup>, Anna Rohrbach<sup>1,3</sup>,  
Bernt Schiele<sup>3</sup>, Trevor Darrell<sup>1</sup>, and Marcus Rohrbach<sup>4</sup>

<sup>1</sup>EECS, UC Berkeley, <sup>2</sup>University of Amsterdam, <sup>3</sup>MPI for Informatics, <sup>4</sup>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<sup>1</sup>, Lisa Anne Hendricks<sup>1</sup>, Zeynep Akata<sup>2,3</sup>, Anna Rohrbach<sup>1,3</sup>,  
Bernt Schiele<sup>3</sup>, Trevor Darrell<sup>1</sup>, and Marcus Rohrbach<sup>4</sup>

<sup>1</sup>EECS, UC Berkeley, <sup>2</sup>University of Amsterdam, <sup>3</sup>MPI for Informatics, <sup>4</sup>Facebook AI Research

*The activity is*

*A: Mountain Biking*



*A: Road Biking*



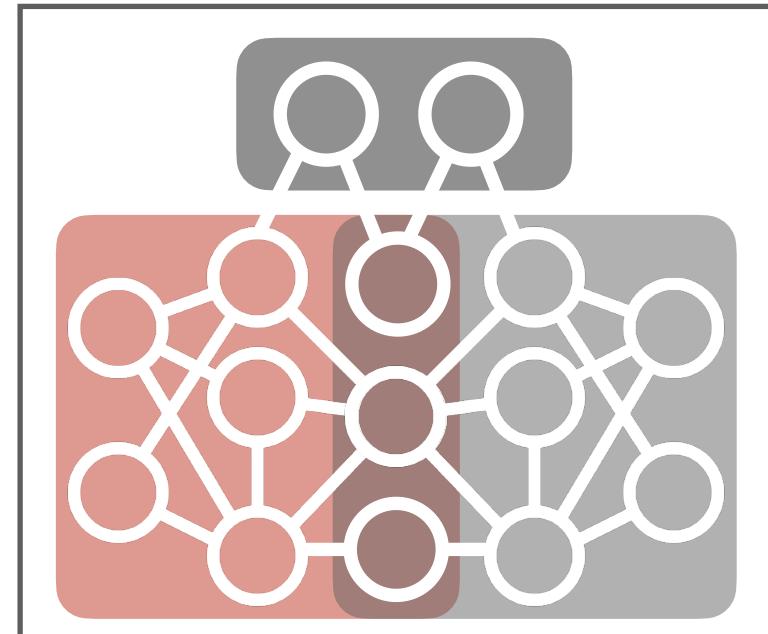
*... because he is riding a bicycle down a mountain path in a mountainous area.*

*... because he is wearing a cycling uniform and riding a bicycle down the road.*

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

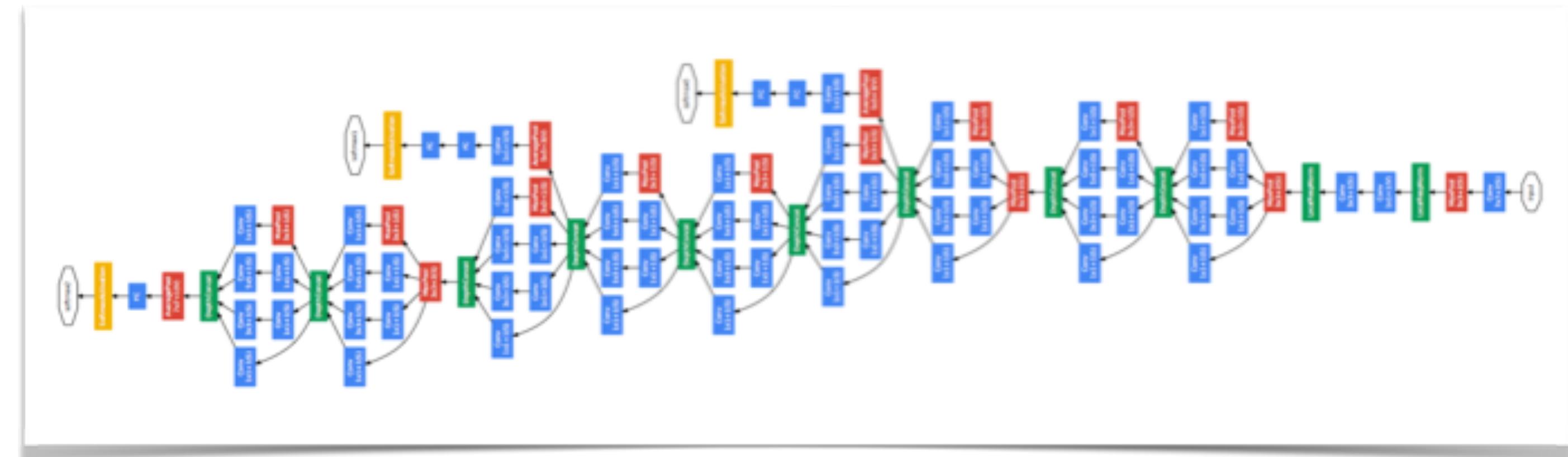
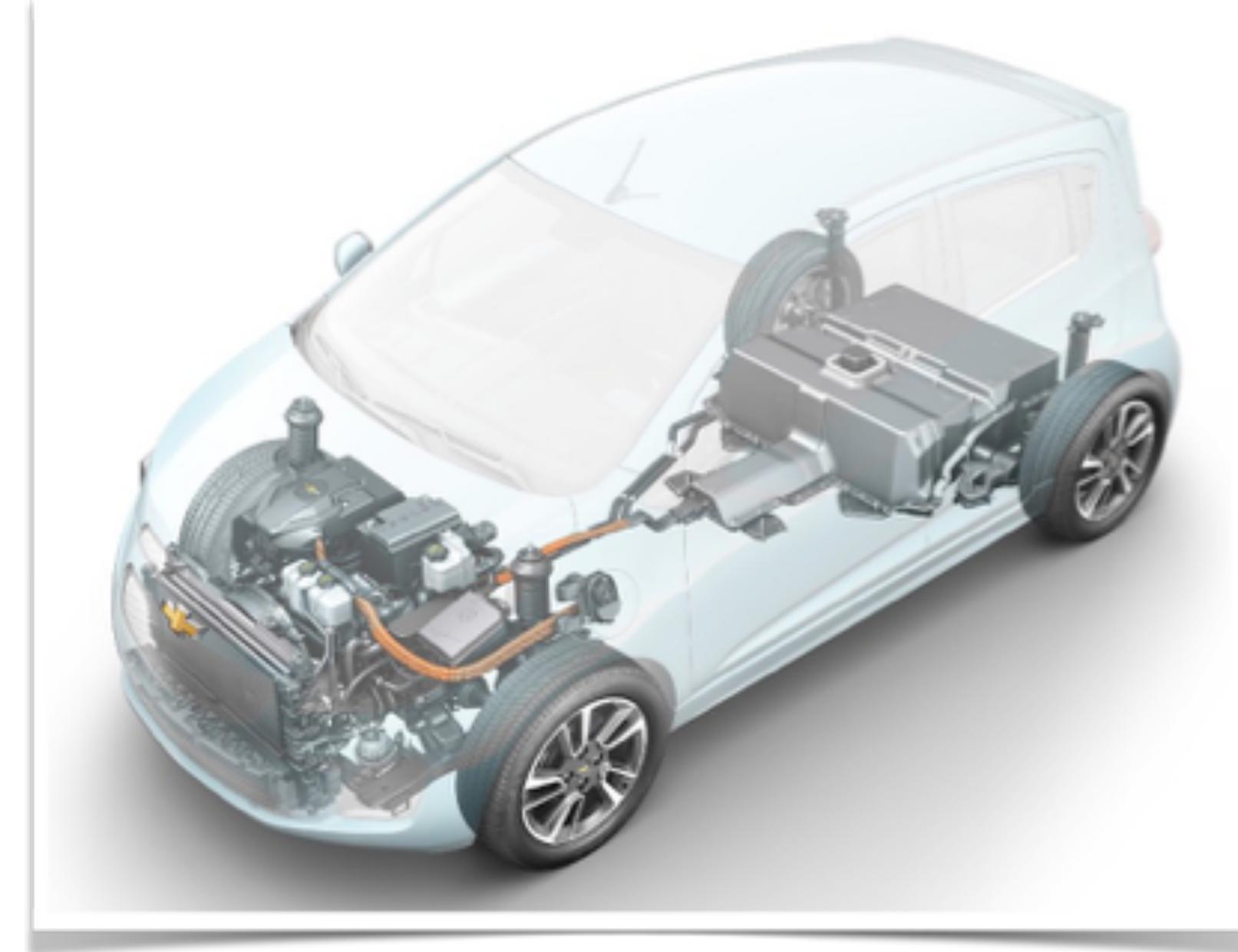
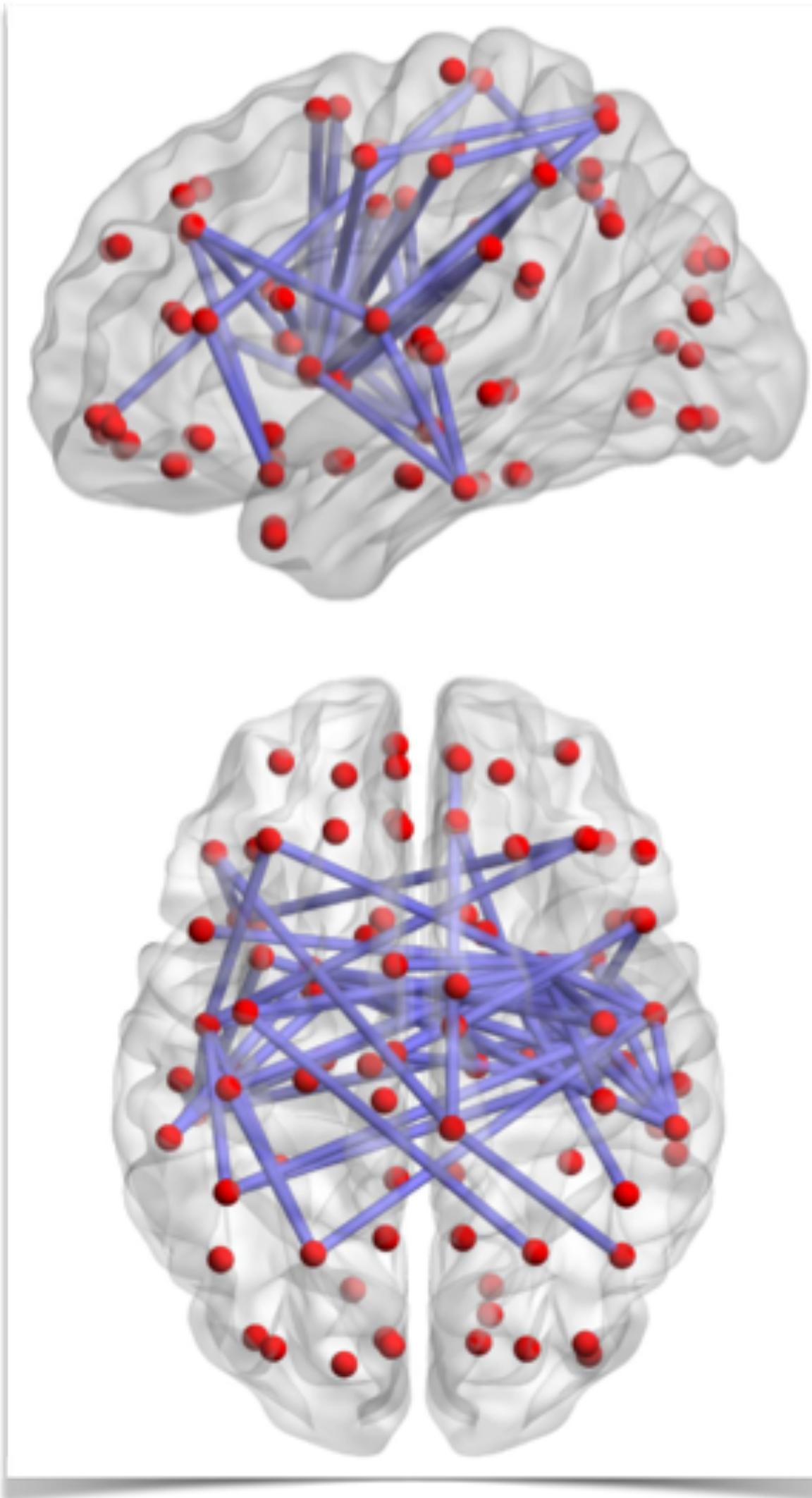
[1] L.H. Gilpin. Explaining possible futures for robust autonomous decision-making. Proceedings of the AAAI Fall Symposium on Anticipatory Thinking, 2019.

[2] L.H. Gilpin et al. Anomaly Detection Through Explanations. Under Review.



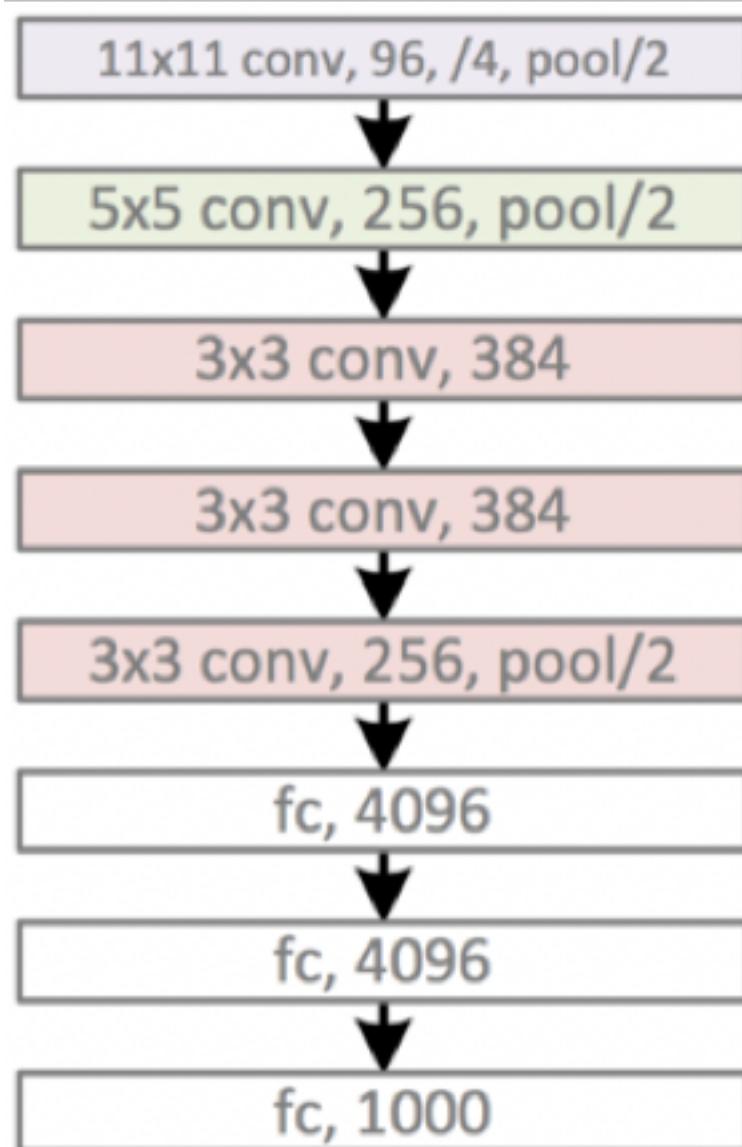
The best option is to veer and slow down. The vehicle is traveling too fast to suddenly stop. The vision system is inconsistent, but the lidar system has provided a reasonable and strong claim to avoid the object moving across the street.

# A Problem: Insides Matter



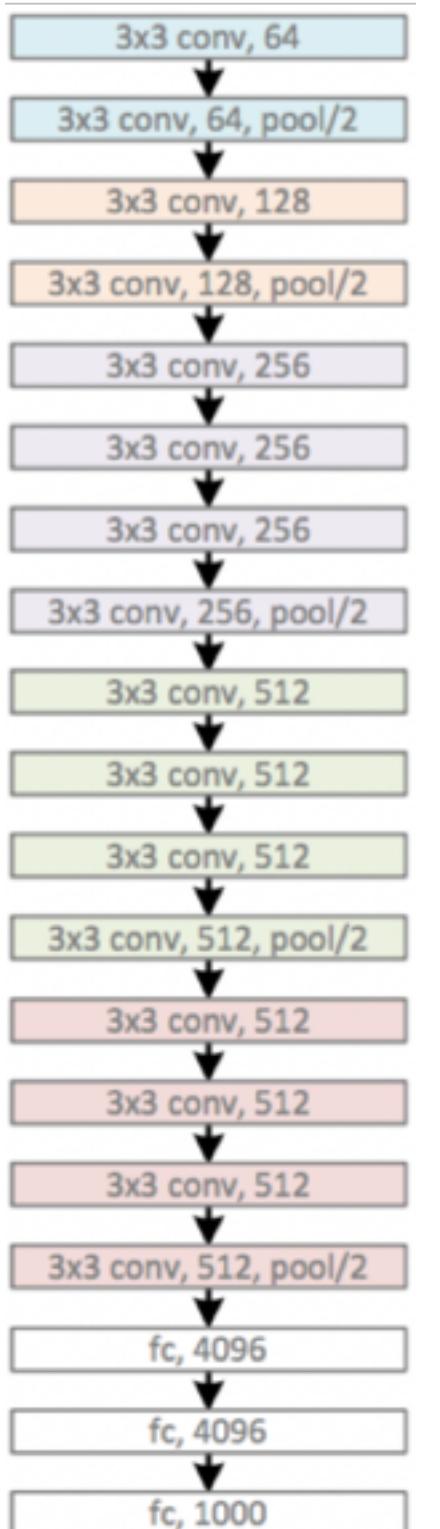
# The More Complex (Deeper) The Deeper the Mystery

The ImageNet logo consists of the word "IMAGENET" in a bold, sans-serif font. The letter "A" is stylized with three colored squares (green, orange, and red) positioned above and to the left of it.



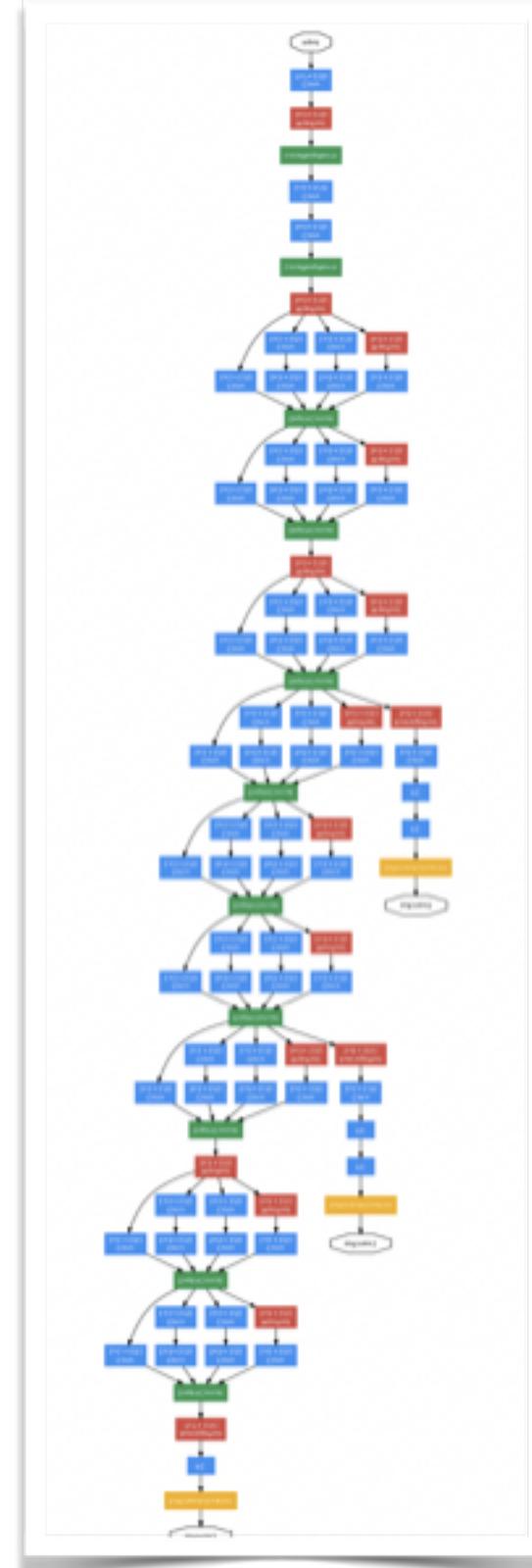
# AlexNet (2012)

8 layers; acc 84.7%



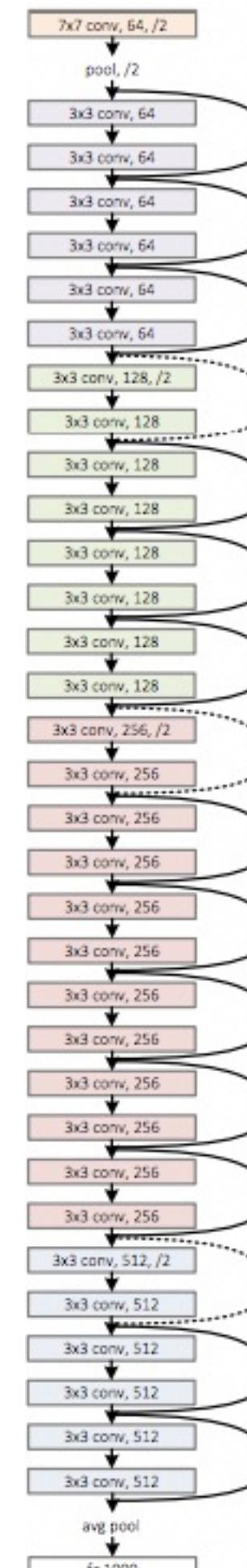
# VGG (2014)

19 layers; acc 91.5%



# GoogLeNet (2015)

22 layers; acc 92.2%

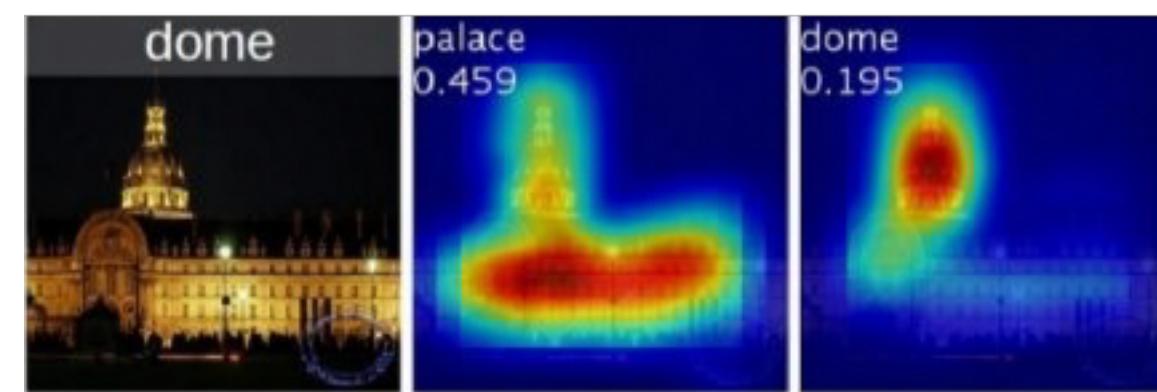


# ResNet (2016)

152 layers; acc 95.6%

# What is Being Explained?

Visual cues



Completeness to  
model

Role of individual  
units



Completeness  
on other tasks

Attention based

<i>Q: Is this a healthy meal?</i>	Textual Justification	Visual Pointing
<i>A: No</i>	<i>...because it is a hot dog with a lot of toppings.</i>	
<i>A: Yes</i>	<i>...because it contains a variety of vegetables on the table.</i>	



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
Evaluation	Completeness to model Completeness on a substitute task	Completeness on a substitute task Detect biases	Human evaluation Detect biases

# Challenges in Explainability



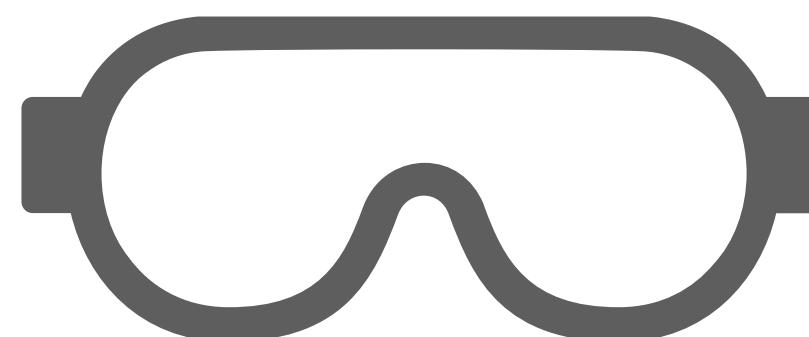
- Standards and metrics for explanations

- 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.



# But How Can Explanations Help?

## Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter and Lalana Kagal

Computer Science and Artificial Intelligence Laboratory

Massachusetts Institute of Technology

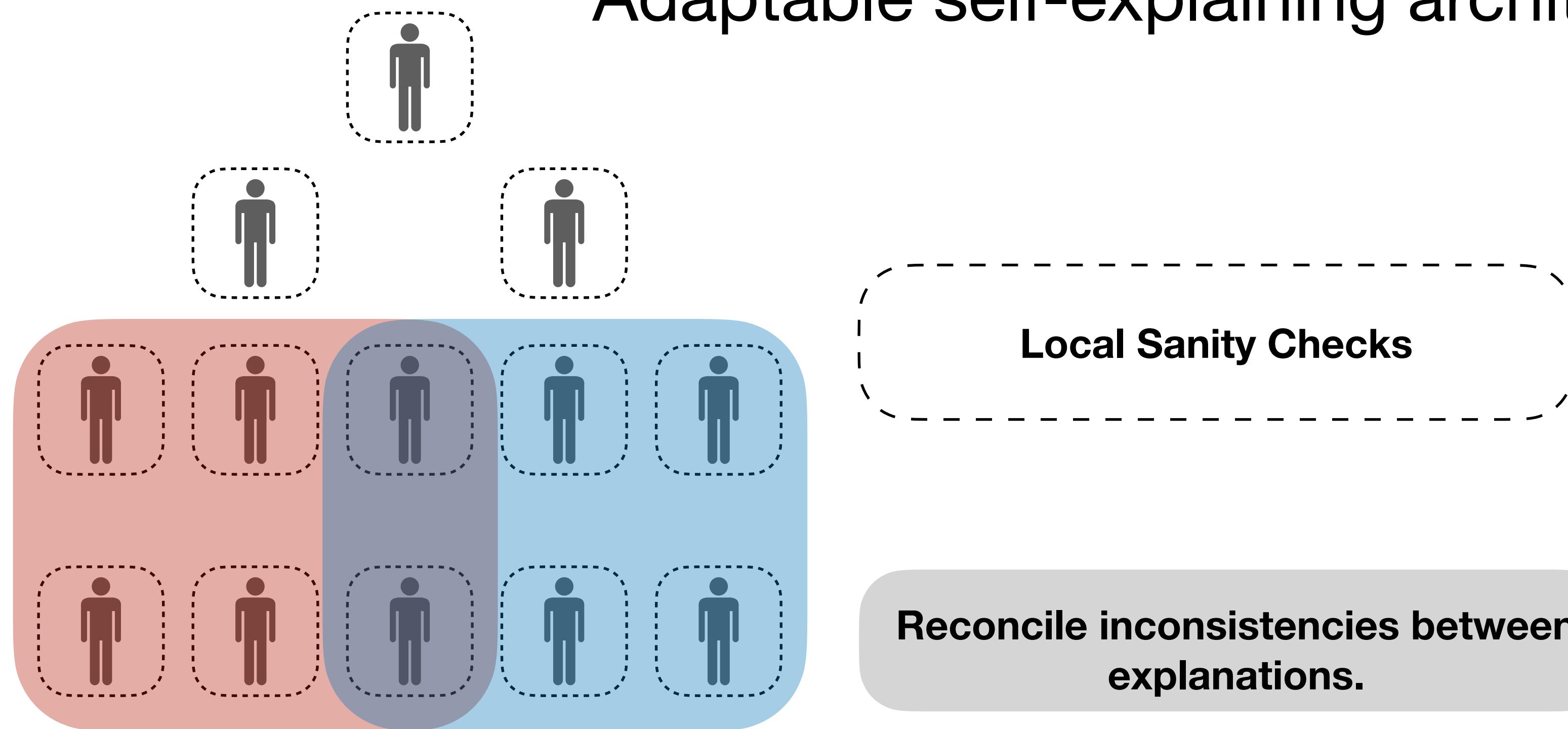
Cambridge, MA 02139

{lgilpin, davidbau, bzy, abajwa, specter, lkagal}@ mit.edu

- Ex-post-facto
- Static
- Dynamic
- Self-explaining architectures.

# Explanatory Anomaly Detection

## Adaptable self-explaining architectures



1. Hierarchy of overlapping self-explaining committees.
2. Continuous interaction and communication.
3. When failure occurs, a story can be made, combining the member's explanations.

[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 et al. Anomaly Detection Through Explanations. Under Review.

# Explanations can Mitigate Common Problems

Reconcile inconsistencies between explanations.



Reconcile conflicting explanations

Local Sanity Checks

## The Trollable Self-Driving Car

Humans are pretty good at guessing what others on the road will do. Driverless cars are not—and that can be exploited.

Reason about new examples.  
Utilize commonsense knowledge.

# Contributions and Future Work

- A taxonomy and best practices for explanations via completeness and interpretability
  - What [part or parts] is being explain?
- Future directions
  - How can a network explain itself?
  - How to incorporate explainable methods?
  - Is there a provable trade-off between completeness and interpretability?
  - What explanations are best suited for policy?
    - See our follow-up paper: “Explaining explanations to society”