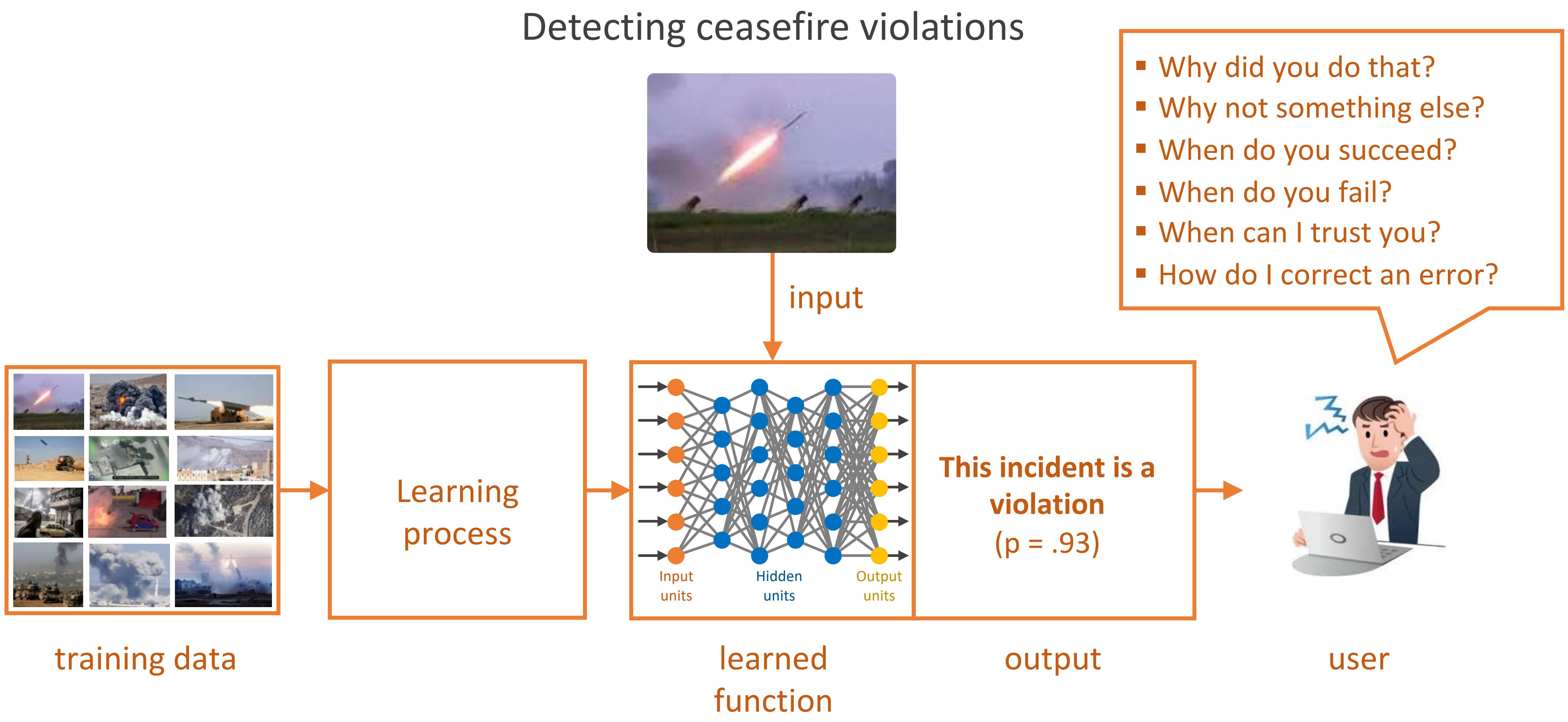
Explainable AI

Matt Turek, I2O

Presenter: John Reeder, Naval Information Warfare Center

How is it done today?



What are we trying to do?

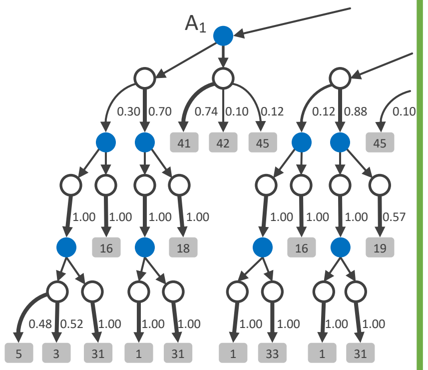
Detecting ceasefire violations

input

- I understand why
- I understand why not
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred

training data

New learning process



explainable model

This is a violation:

These events occur before tweet reports

explanation interface

user

Challenge problems

Data analytics

Getty Images

Explains recommendations to an analyst

Autonomy

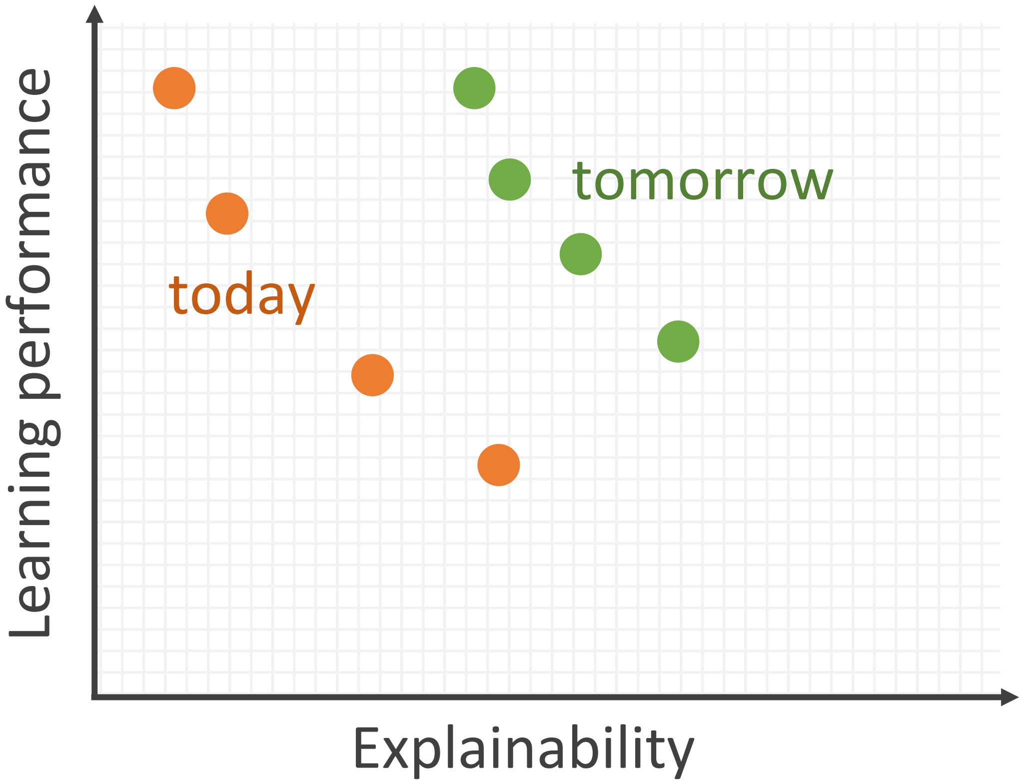
US Army

Explains actions to an operator

Goal: Performance and explainability

Create a suite of machine learning techniques that

- Produce more explainable models, while maintaining a high level of learning performance
- Enable human users to understand, appropriately trust, and effectively manage the emerging generation of AI systems



DARPA

Randomized Input Sampling for Explanation (RISE)

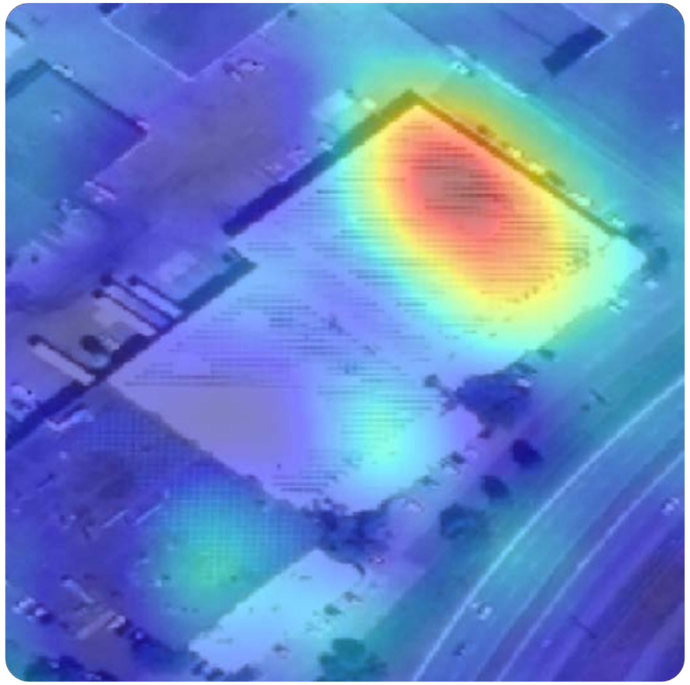
Neural network prediction

solar farm: 63%, **shopping mall**: 23%

FMoW dataset

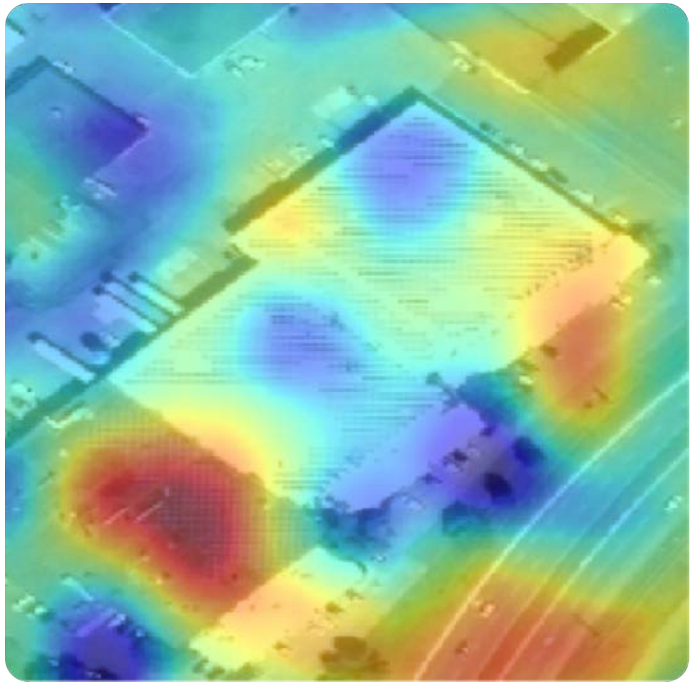
RISE Explanation for solar farm

solar farm: 63%



RISE Explanation for shopping mall

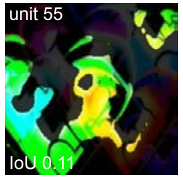
shopping mall: 23%



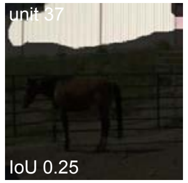
Vitali Petsiuk, Abir Das, and Kate Saenko. *RISE: Randomized Input Sampling for Explanation of Black-box Models*. British Machine Vision Conference (BMVC), 2018.

University of California, Berkeley
Boston University

Network dissection - AlexNet layers for recognizing places



Unit 55 (veined)

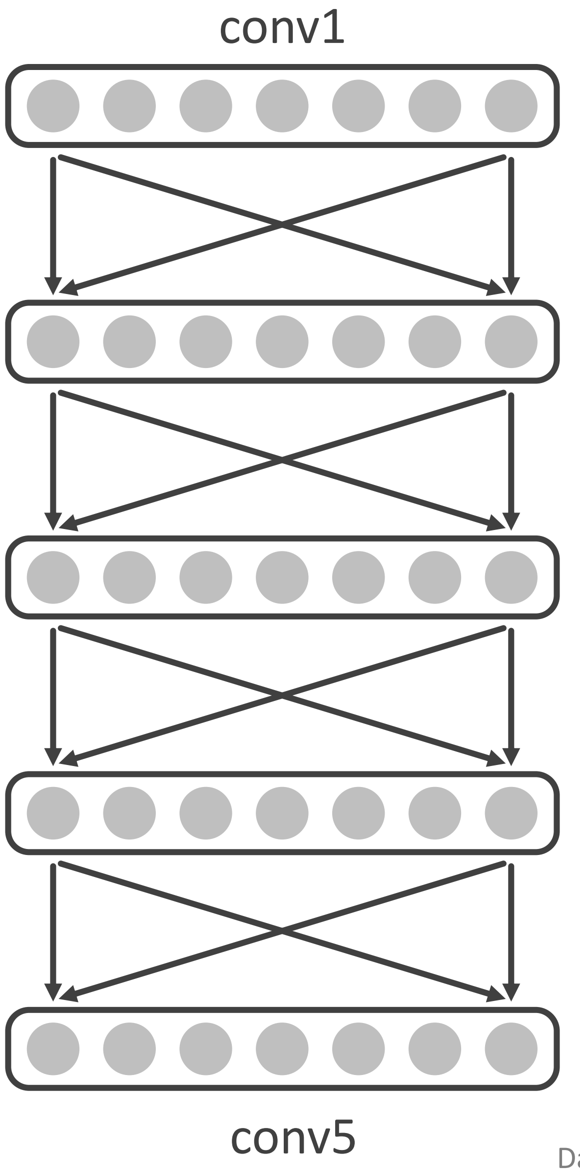


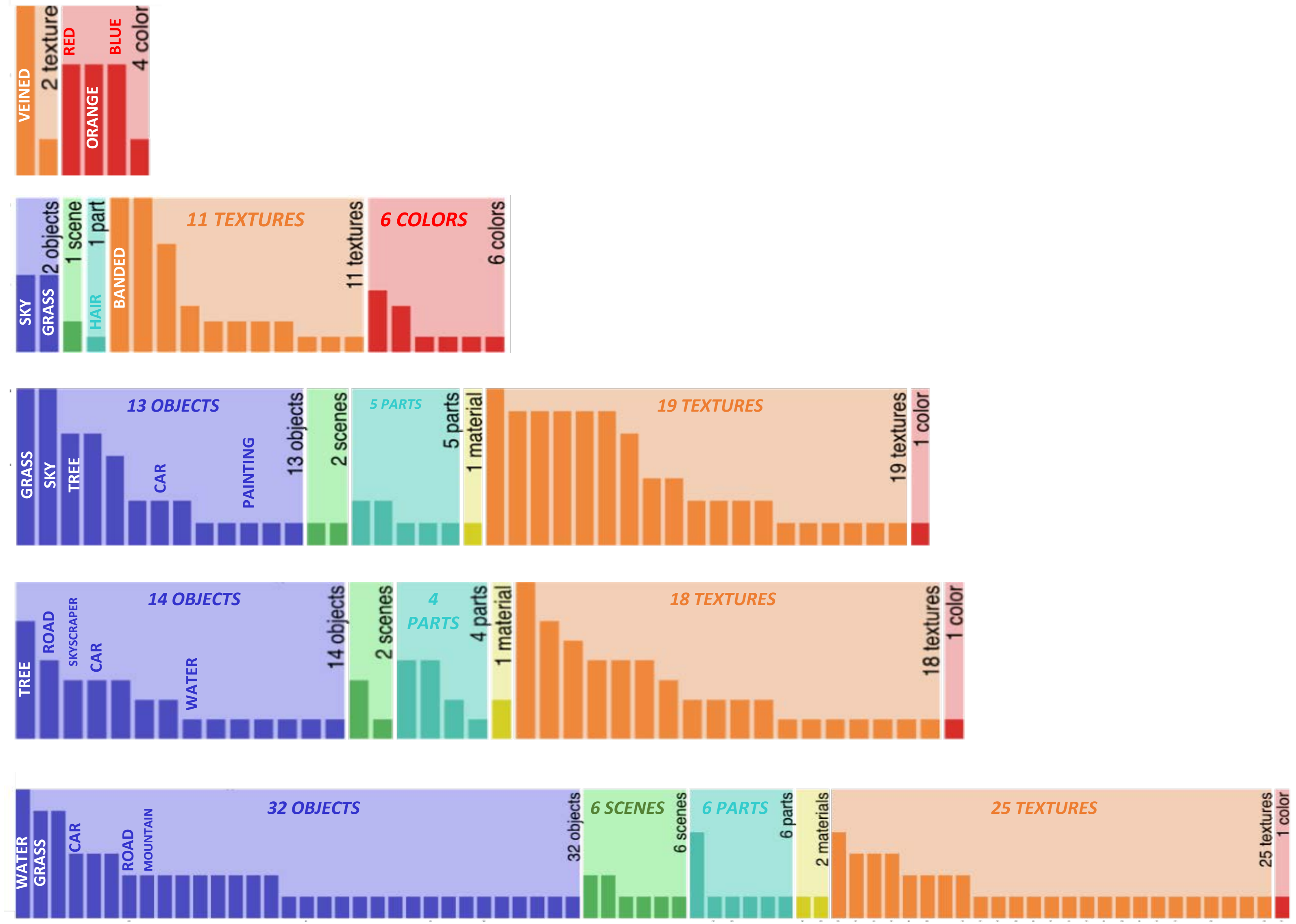
Unit 37 (banded)

Unit 20 (grass)

Unit 109 (tree)

Unit 14 (water)





David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. *GAN Dissection: Visualizing and Understanding Generative Adversarial Networks*. arXiv preprint arxiv 1811.10597, 2018.

Raytheon BBN
MIT

Explaining image classifiers by counterfactual generation

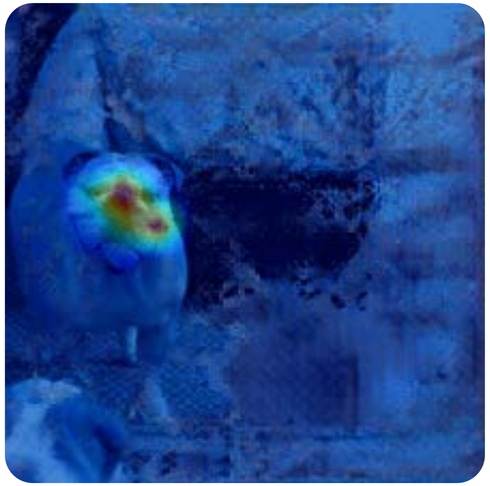
Input image

Foreground retained

Foreground removed

Spatial attention

Spatial attention



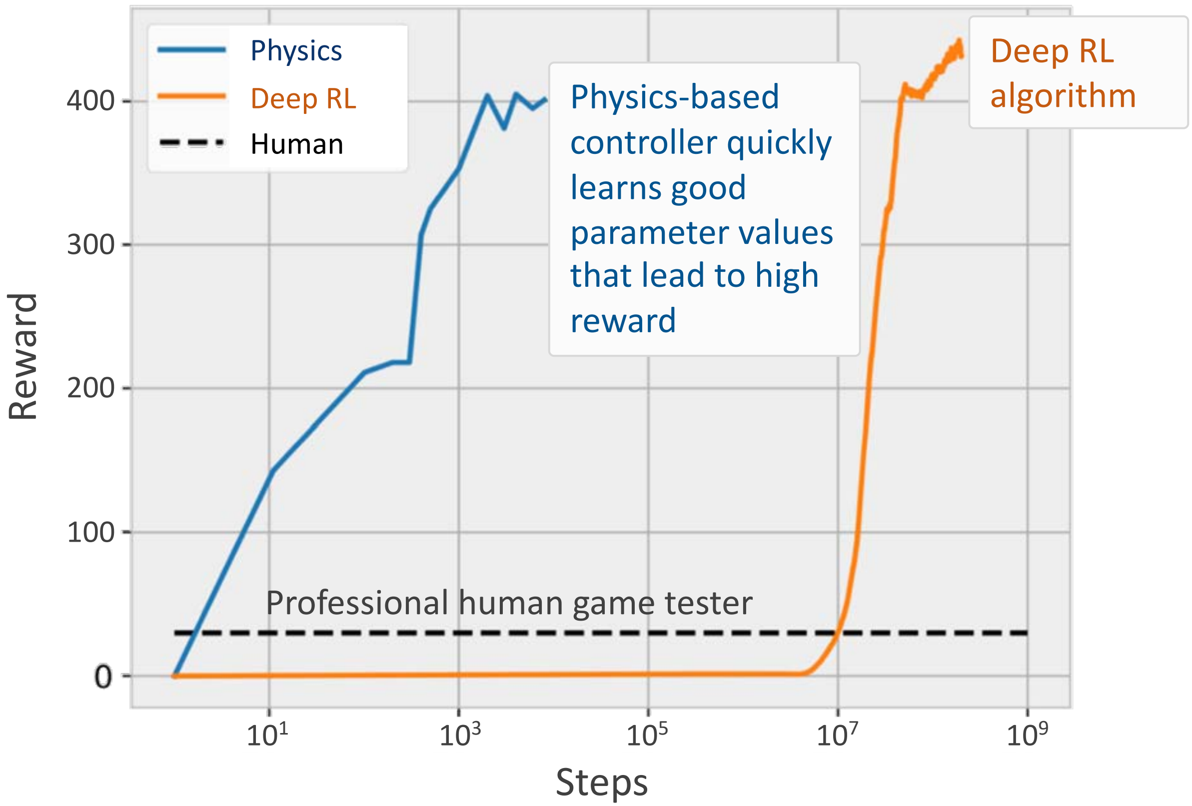
Spatial attention

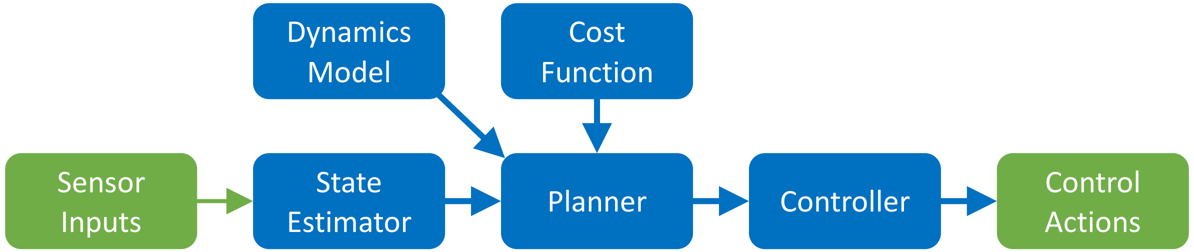
Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. *Explaining Image Classifiers by Adaptive Dropout and Generative In-filling.* CoRR abs/1807.08024 (2018).

SRI International
University of Toronto

End-to-end learning of differentiable physics

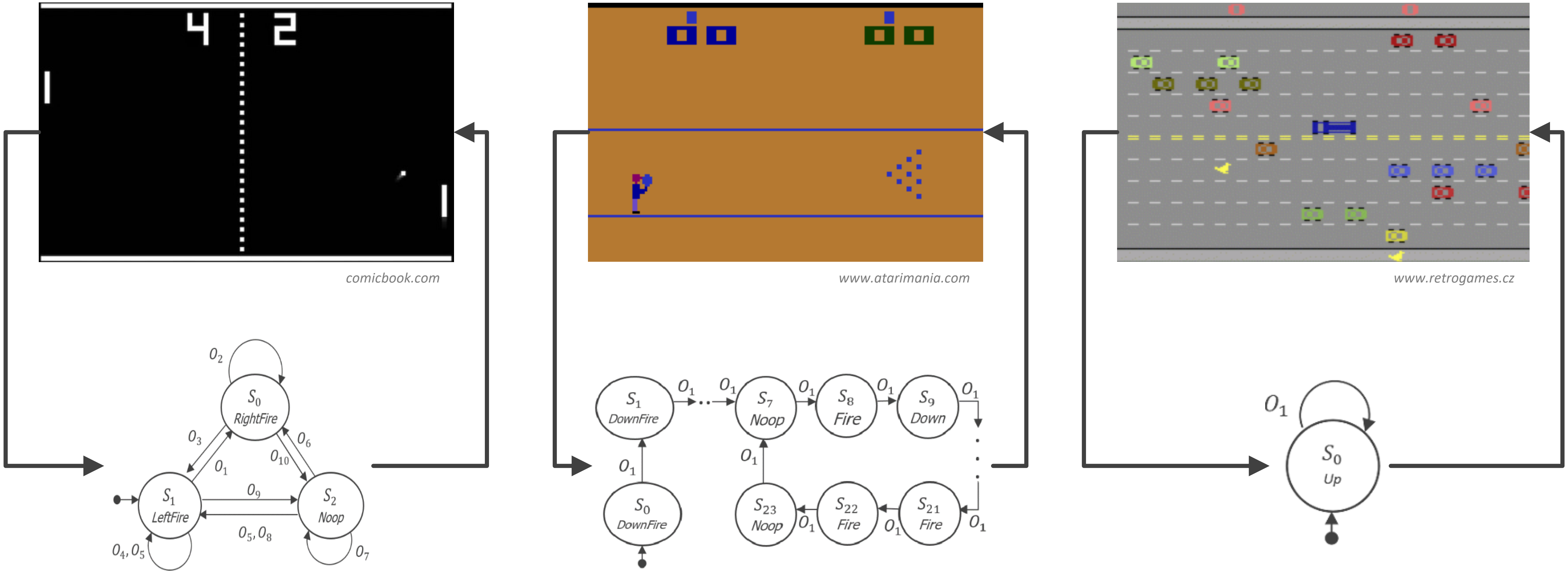
github.com



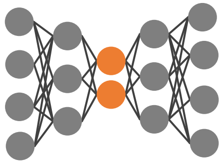


Filipe de A. Belbute-Peres, Kevin A. Smith, Kelsey R. Allen, Joshua B. Tenenbaum, and J. Zico Kolter. *End-to-End Differentiable Physics for Learning and Control*. NeurIPS 2018: 7178-7189.

Learning finite state representations of recurrent policy networks



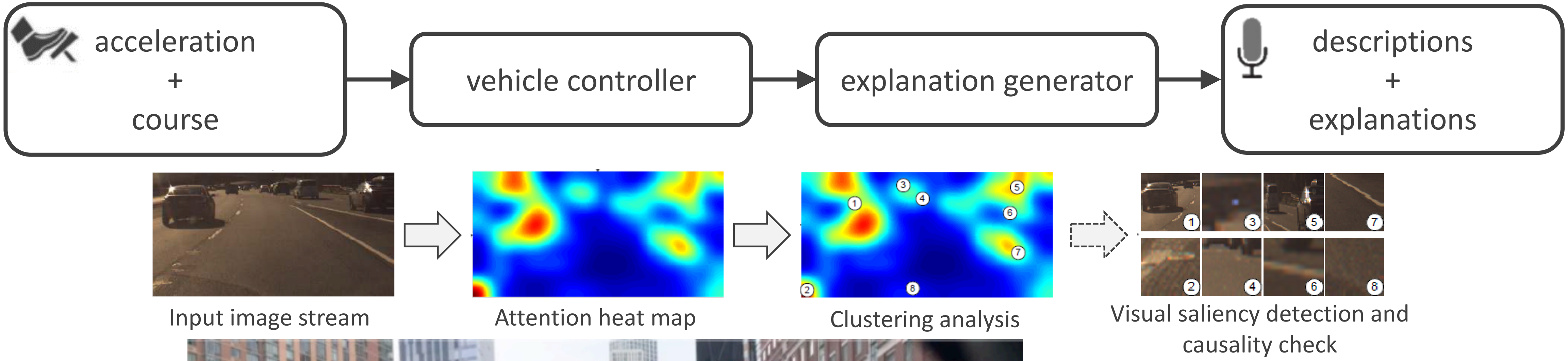
Quantized Bottleneck Networks (QBNs)



Anurag Koul, Sam Greydanus, and Alan Fern. *Learning Finite State Representations of Recurrent Policy Networks*. International Conference on Learning Representations (ICLR), 2019.

Oregon State University

Textual explanations and visualizing causal attention



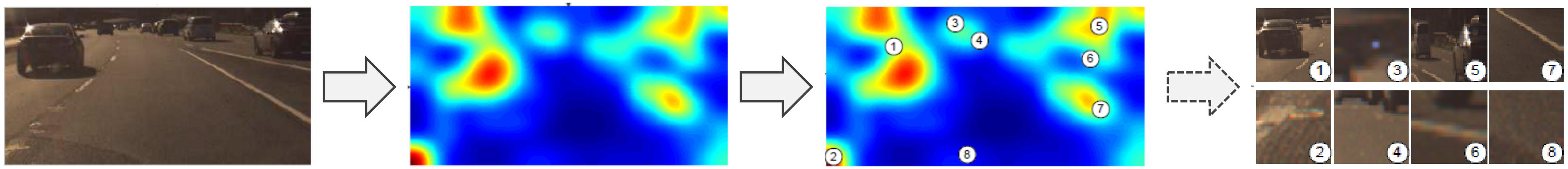
Jinkyu Kim and John Canny. *Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention*. In Proceedings of IEEE International Conference on Computer Vision (ICCV), 2017.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. *Textual Explanations for Self-Driving Vehicles*. In Proceedings of European Conference on Computer Vision (ECCV), 2018.

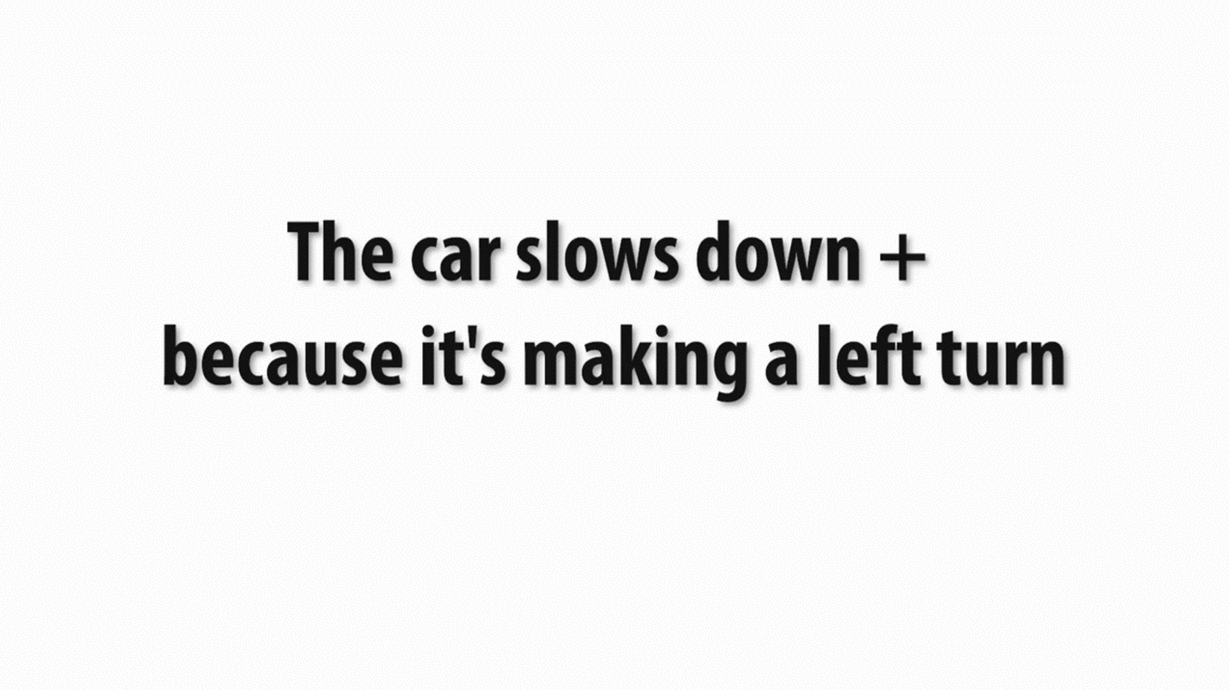
University of California, Berkeley
University of Amsterdam

Textual explanations and visualizing causal attention

acceleration + course → vehicle controller → explanation generator → descriptions + explanations



Input image stream → Attention heat map → Clustering analysis → Visual saliency detection and causality check



Jinkyu Kim and John Canny. *Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention*. In Proceedings of IEEE International Conference on Computer Vision (ICCV), 2017.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. *Textual Explanations for Self-Driving Vehicles*. In Proceedings of European Conference on Computer Vision (ECCV), 2018.

University of California, Berkeley
University of Amsterdam

What we've learned

- Promising approaches to explainability

CP	Performer	Explainable Model
Both	UC Berkeley	Deep Learning
	Charles River	Causal Modeling
	UCLA	Stochastic And-Or-Graphs
Autonomy	Oregon State	Deep Adaptive Programs
	PARC	Cognitive Modeling
	CMU	Explainable RL (XRL)
Analytics	SRI International	Deep Learning
	Raytheon BBN	Deep Learning
	UT Dallas	Probabilistic Logic
	Texas A&M	Mimic Learning
	Rutgers	Explanation by Example

- Initial results from measuring explanation effectiveness
  - Users preferred explanations
  - Explanations engendered appropriate trust
  - Explanations sometimes improved mental model predictions
  - Incorrect explanations negatively impacted these measures

Thank You