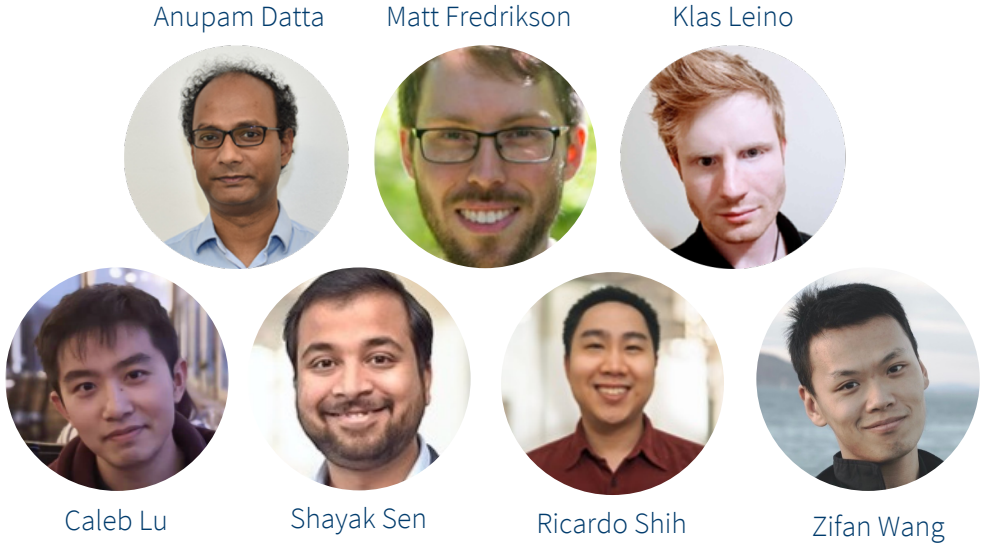
tinyurl.com/trulens-demo

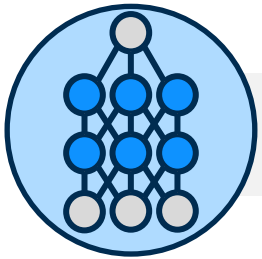
Demo | Neural Information Processing Systems 2021

Exploring Conceptual Soundness with TruLens

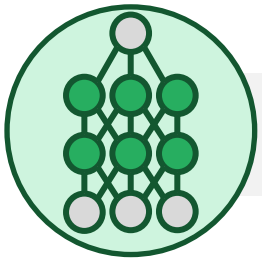
Anupam Datta
Matt Fredrikson
Klas Leino
Kaiji Lu
Shayak Sen
Ricardo Shih
Zifan Wang



Motivating example



model 1



model 2

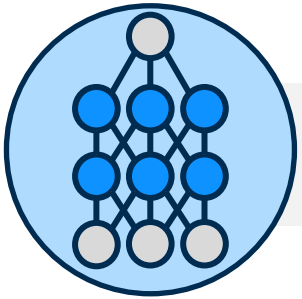
dog

dog

dog

cat

Problem: lack of conceptual soundness



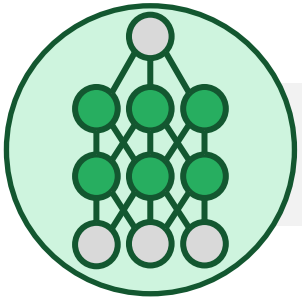
model 1

model 1 is not *conceptually sound*

dog

dog

Problem: lack of conceptual soundness



model 2

model 2 *is* conceptually sound

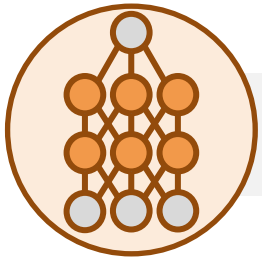
dog

cat

Key Takeaways

accurate explanations can be used to help assess conceptual soundness

Problem: lack of robustness



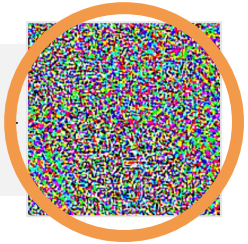
model 3

dog

sports car

how do we explain this prediction?

=



important features?

Key Takeaways: Robust Models, Accurate Explanations

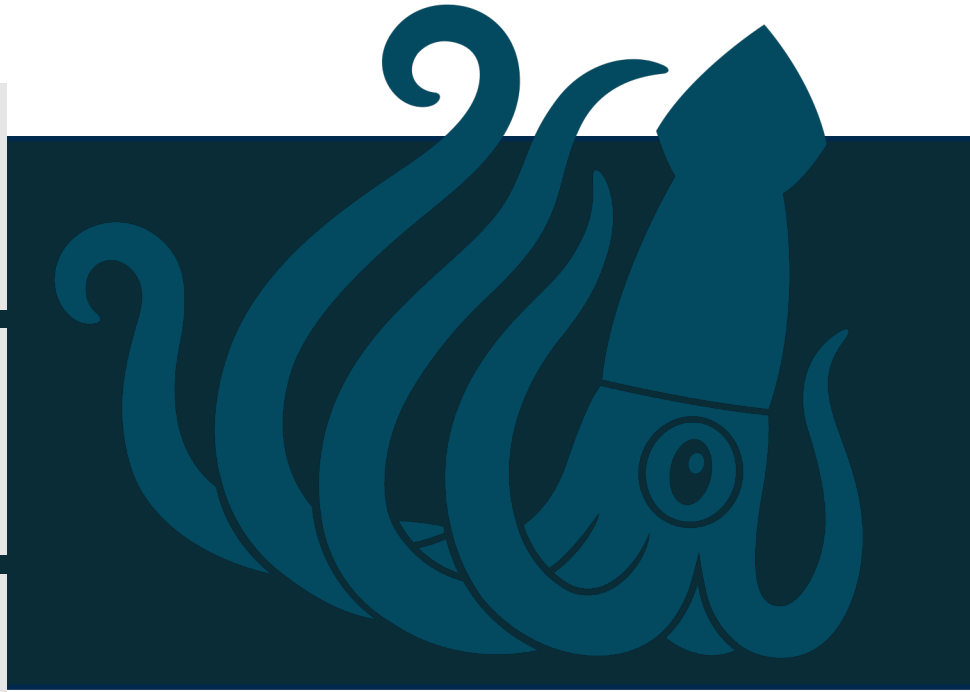
accurate explanations can be used to help assess conceptual soundness

robustness is a key prerequisite for conceptual soundness that helps improve the interpretability of explanations

tinyurl.com/trulens-demo

TruLens

- framework for gradient-based explanations that supports both tensorflow/keras and pytorch
- more flexibility than other explanation libraries
- accompanying materials, usage guide on demo webpage



tinyurl.com/trulens-demo

Demo 1

Conceptual Soundness

Demo 2

Robustness and Explainability

Key Takeaways: Robust Models, Accurate Explanations

accurate explanations can be used to help assess conceptual soundness

robustness is a key prerequisite for conceptual soundness that helps improve the interpretability of explanations