Learn with
Google AI | Explore ML

Aditya Jyoti Paul
@phreakyphoenix
Founder and Team Lead,
Cognitive Applications Research Lab (CARL)
Google AI ExploreML Facilitator

What is interpretability in ML?

Very simply put, it is the ability for the model to explain its output.

Declaration: You identify an object in an image

Justification: You point out to certain features in the object to justify why it’s a tree.


EXCUSE ME
DO YOU HAVE A MOMENT TO TALK ABOUT
INTERPRETABLE MACHINE LEARNING

Do we need a different model? How about rule lists?

If (sunny and hot)	then	go swim
Else if (sunny and cold)	then	go ski
Else if (wet and weekday)	then	go work
Else if (free coffee)	then	attend tutorial
Else if (cloudy and hot)	then	go swim
Else if (snowing)	then	go ski
Else if (New Rick and Morty)	then	watch TV
Else if (paper deadline)	then	go work
Else if (hungry)	then	go eat
Else if (tired)	then	watch TV
Else if (advisor might come)	then	go work
Else if (code running)	then	watch TV
Else	then	go work

Credits:
CVPR: 2018

Maybe rule sets are better?

IF (sunny and hot) OR (cloudy and hot) OR (sunny and thirsty and bored) OR (bored and tired) OR (thirty and tired) OR (code running) OR (friends away and bored) OR (sunny and want to swim) OR (sunny and friends visiting) OR (need exercise) OR (want to build castles) OR (sunny and bored) OR (done with deadline and hot) OR (need vitamin D and sunny) OR (just feel like it)
THEN go to beach
ELSE work

Credits:
CVPR: 2018

Some common misunderstandings

1. **Linear Models and Decision Trees are completely explainable**

Some common misunderstandings

2. Only more data and a clever algorithm will help solve the problem.

Some common misunderstandings

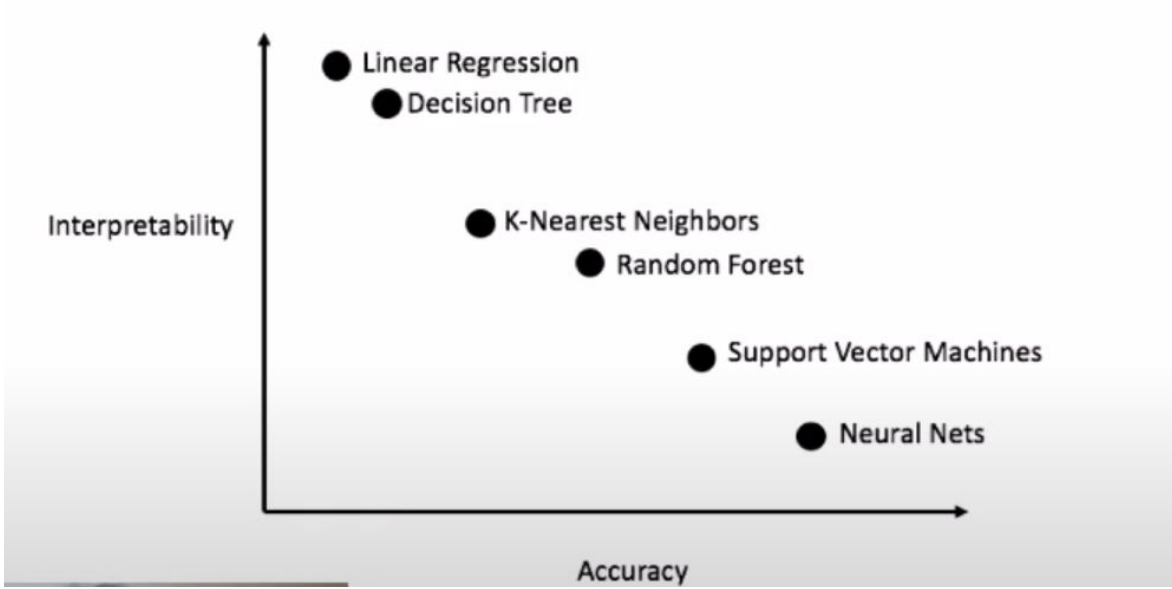
3. **We always need interpretability.**

Some common misunderstandings

4. **Interpretability == (Fairness, Trust, Causality)**


Hmm....
Okay

Accuracy vs Explainability Tradeoff



Faceting

Row-Based Faceting
<NONE>

Column-Based Faceting
Age

1

Positioning

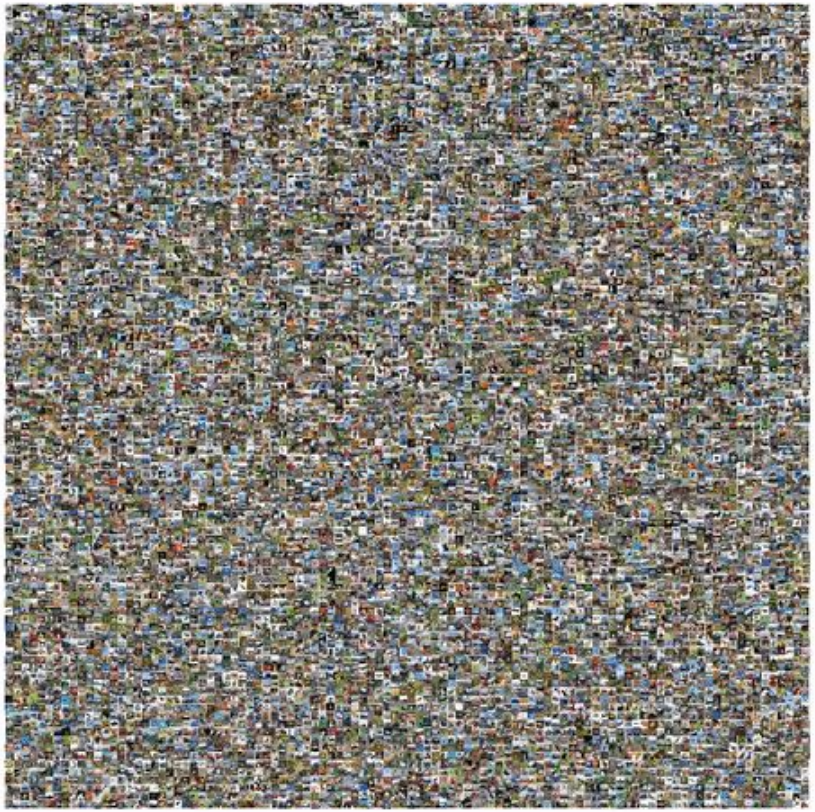
Color

Color By
<NONE>

Display

Legend
Age

17 – 90



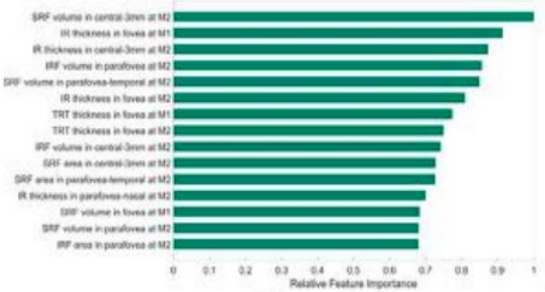
Facets Dive

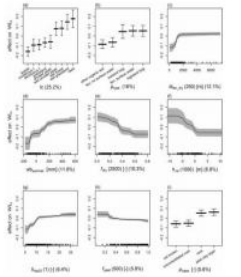

I MADE A VERY POWERFUL MACHINE LEARNING ALGORITHM
MOREOVER I KNOW WHICH FEATURES ARE USEFUL!
BUT... I DON'T KNOW WHY THEY ARE USEFUL...
I DON'T KNOW WHY THEY ARE USEFUL...

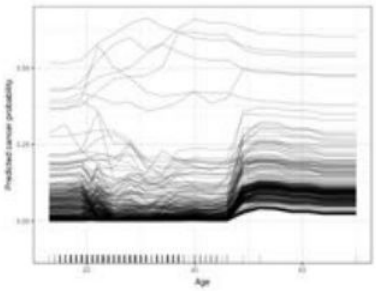
Overview of explanation in different AI fields (1)

- Machine Learning

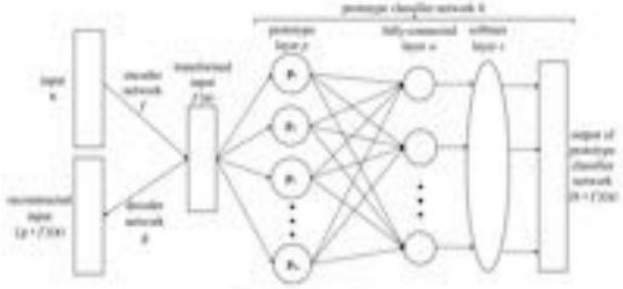
Interpretable Models:
- Linear regression,
- Logistic regression,
- Decision Tree,
- GLMs,
- GAMs
- Naive Bayes,
- KNNs





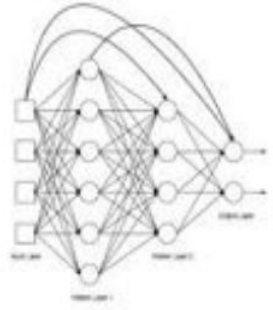


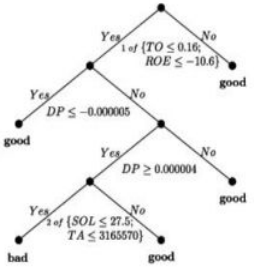
Feature Importance, Partial Dependence Plot, Individual Conditional Expectation



Auto-encoder

Oscar Li, Hao Liu, Chaofan Chen, Cynthia Rudin: Deep Learning for Case-Based Reasoning Through Prototypes: A Neural Network That Explains Its Predictions. AAAI 2018: 3530-3537



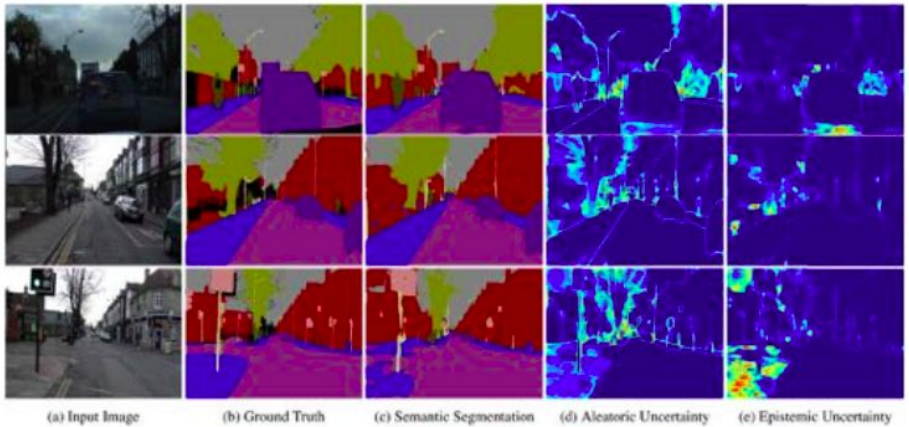


Surogate Model

Mark Craven, Jude W. Shavlik: Extracting Tree-Structured Representations of Trained Networks. NIPS 1995: 24-30

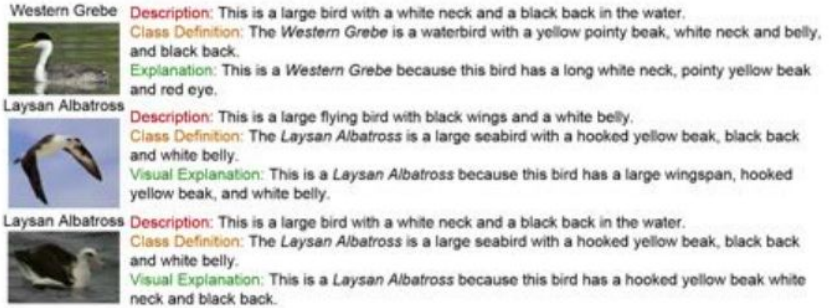
Overview of explanation in different AI fields (2)

- Computer Vision



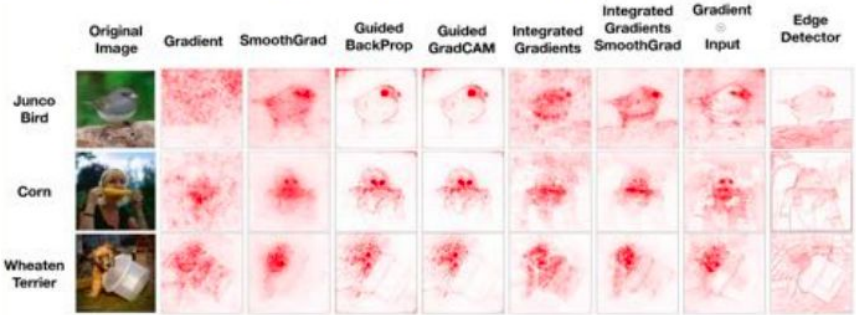
Uncertainty Map

Alex Kendall, Yarin Gal: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? NIPS 2017: 5580-5590



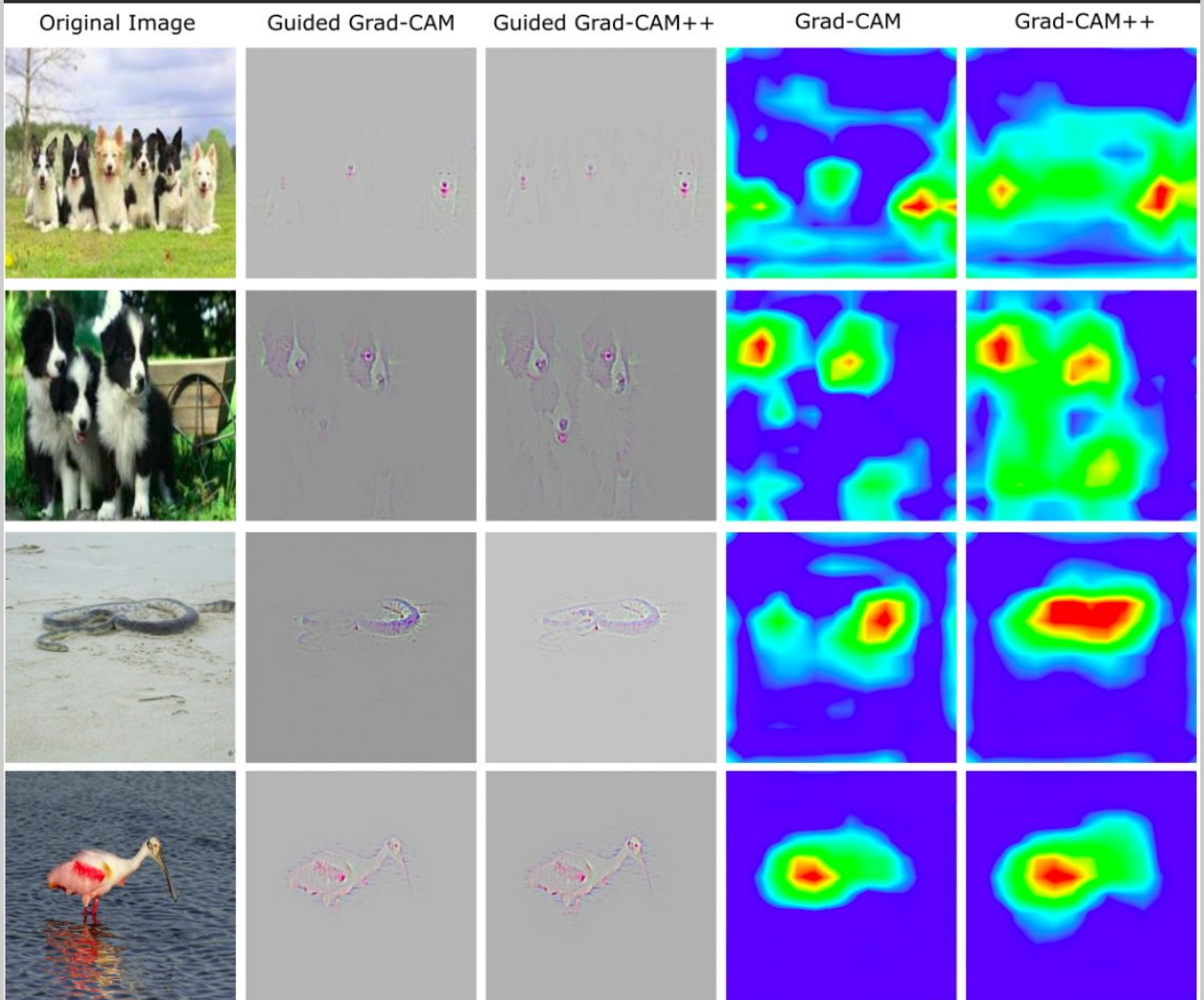
Visual Explanation

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, Trevor Darrell: Generating Visual Explanations. ECCV (4) 2016: 3-19



Saliency Map

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, Been Kim: Sanity Checks for Saliency Maps. NeurIPS 2018: 9525-9536


Original Image
Guided Grad-CAM
Guided Grad-CAM++
Grad-CAM
Grad-CAM++

Categorizing the Efforts

1. Post-Hoc vs Ante-Hoc
2. Model-agnostic vs Model-Dependent
3. Global vs Local

tl;dr

- Explanations and interpretability are required for better human trust, system debug, and legal compliance.
- No monolithic, agreed upon definition of Explainable AI
- Adoption spans multiple AI fields
- Explainability, interpretability come at a cost
- Design with humans and task in mind
- Human-based evaluation is essential

Resources:

1. Explainable AI: xaitutorial2019.github.io
2. adityac94/Grad_CAM_plus_plus: A generalized gradient-based CNN visualization technique
3. insikk/Grad-CAM-tensorflow: tensorflow implementation of Grad-CAM (CNN visualization)
4. Lecture Video on XAI: https://www.youtube.com/watch?v=2nUiVJiVchw


THANK YOU FOR YOUR ATTENTION
NOW IT'S TIME FOR QUESTIONS

If I missed any of your questions, find me on:

1. LinkedIn: https://www.linkedin.com/in/phreakyphoenix/
2. Twitter: https://twitter.com/phreakyphoenix
3. Email : adityajyotipaul007@gmail.com
4. Github: https://github.com/phreakyphoenix

These slides will soon be made available on https://slides.com/phreakyphoenix