

Explainable AI: Opening up the Black Box

Ryan Wesslen, UNC Charlotte

Davidson Machine Learning Group / Oct 16, 2018

The views expressed herein are those of the presenter; they do not necessarily reflect the views of author's employer, organization, committee, or other group or individual.

Agenda

About Me: 5 min

The pros and cons of Deep Learning: 20 min

Explainable AI: 15 min

LIME Application: 15 min

Resources & Questions: 5 min

Agenda

About Me: 5 min

The pros and cons of Deep Learning: 20 min

Explainable AI: 15 min

LIME Application: 15 min

Resources & Questions: 5 min

This presentation is available at this url



UNC Charlotte PhD Candidate

- Computing & Information Systems (Computer Science)
- UNCC Visualization Center, Pacific Northwest National Laboratory, UNCC Data Science Initiative, Project Mosaic
- Computational social science, visual analytics, text-as-data, social media

UNC Charlotte PhD Candidate

- Computing & Information Systems (Computer Science)
- UNCC Visualization Center, Pacific Northwest National Laboratory, UNCC Data Science Initiative, Project Mosaic
- Computational social science, visual analytics, text-as-data, social media

Bank of America (2009-2014) / Publicis Hawkeye (2014-2015)

- Credit risk and marketing analytics and strategy
- Risk rotational program (GRMAP), small business credit risk, auto lending scorecard modeling

UNC Charlotte PhD Candidate

- Computing & Information Systems (Computer Science)
- UNCC Visualization Center, Pacific Northwest National Laboratory, UNCC Data Science Initiative, Project Mosaic
- Computational social science, visual analytics, text-as-data, social media

Bank of America (2009-2014) / Publicis Hawkeye (2014-2015)

- Credit risk and marketing analytics and strategy
- Risk rotational program (GRMAP), small business credit risk, auto lending scorecard modeling

Teaching & R/R Studio enthusiast

- Taught UNCC workshops (<https://github.com/wesslen>) in R for text, social media, data viz.
- Teaching Visual Analytics course for UNCC Data Science program in Spring 2019



Andrew Ng 
@AndrewYNg



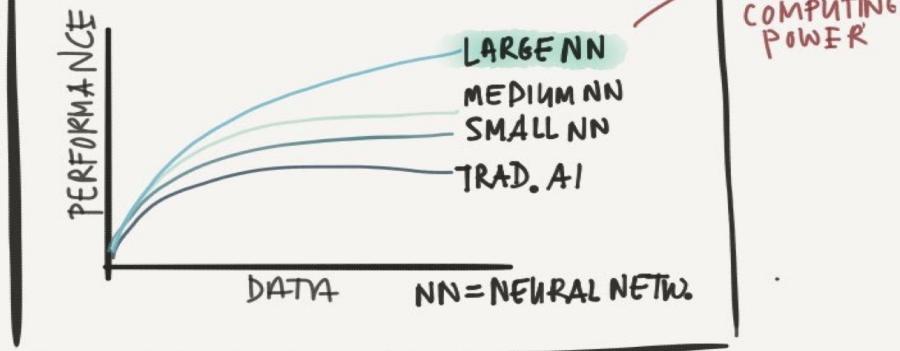
"AI is the new electricity!" Electricity transformed countless industries; AI will now do the same.

361 12:47 PM - May 26, 2016 · Mountain View, CA

346 people are talking about this >

WHY NOW?

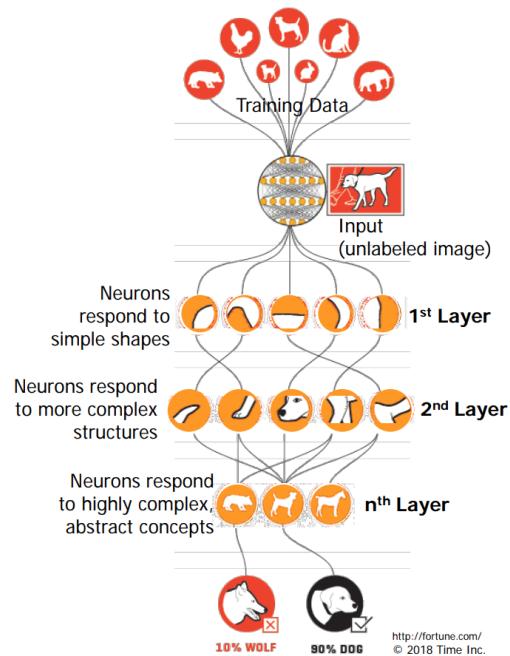
THE PERFECT STORM



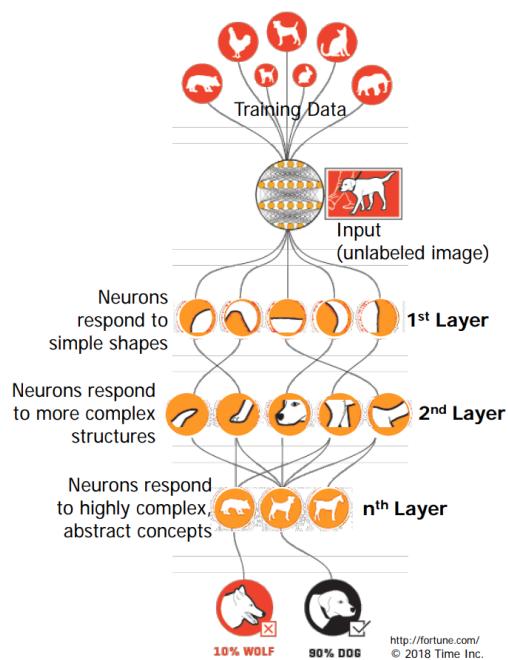
"AI HAS BEEN THROUGH TWO WINTERS BUT NOW I THINK IT HAS REACHED ETERNAL SPRING"

Andrew Ng on YouTube & @TessFernandes

Neural Networks



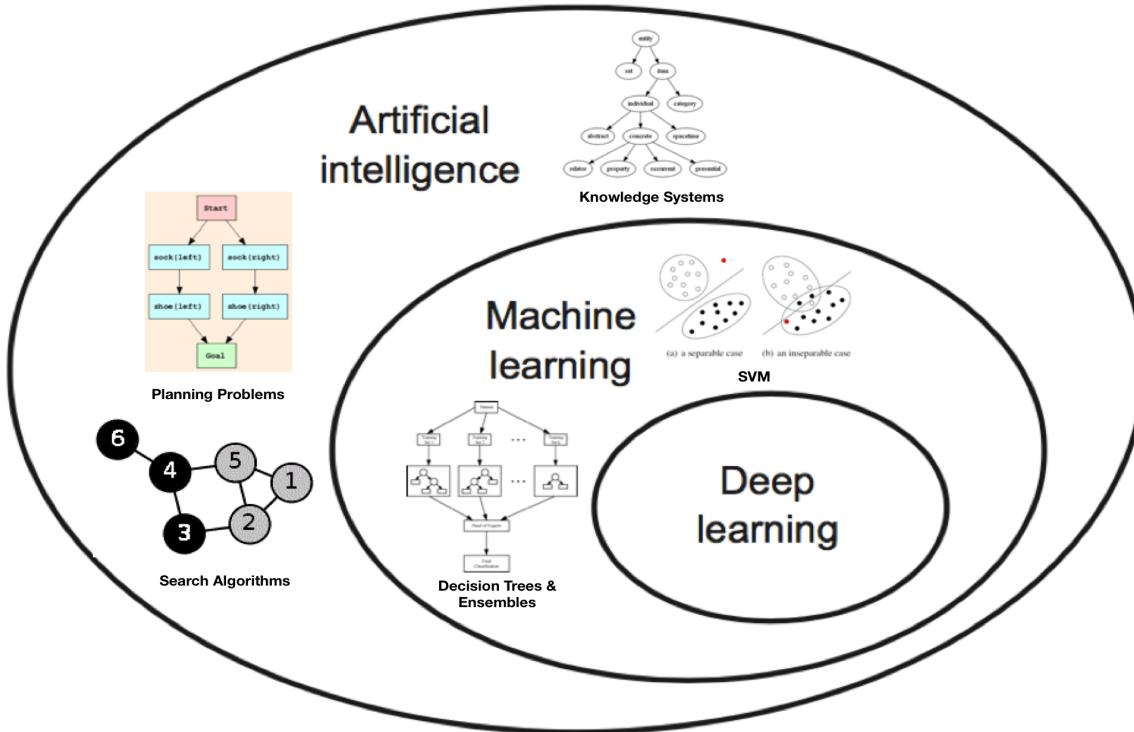
Neural Networks



Large-Scale Computing

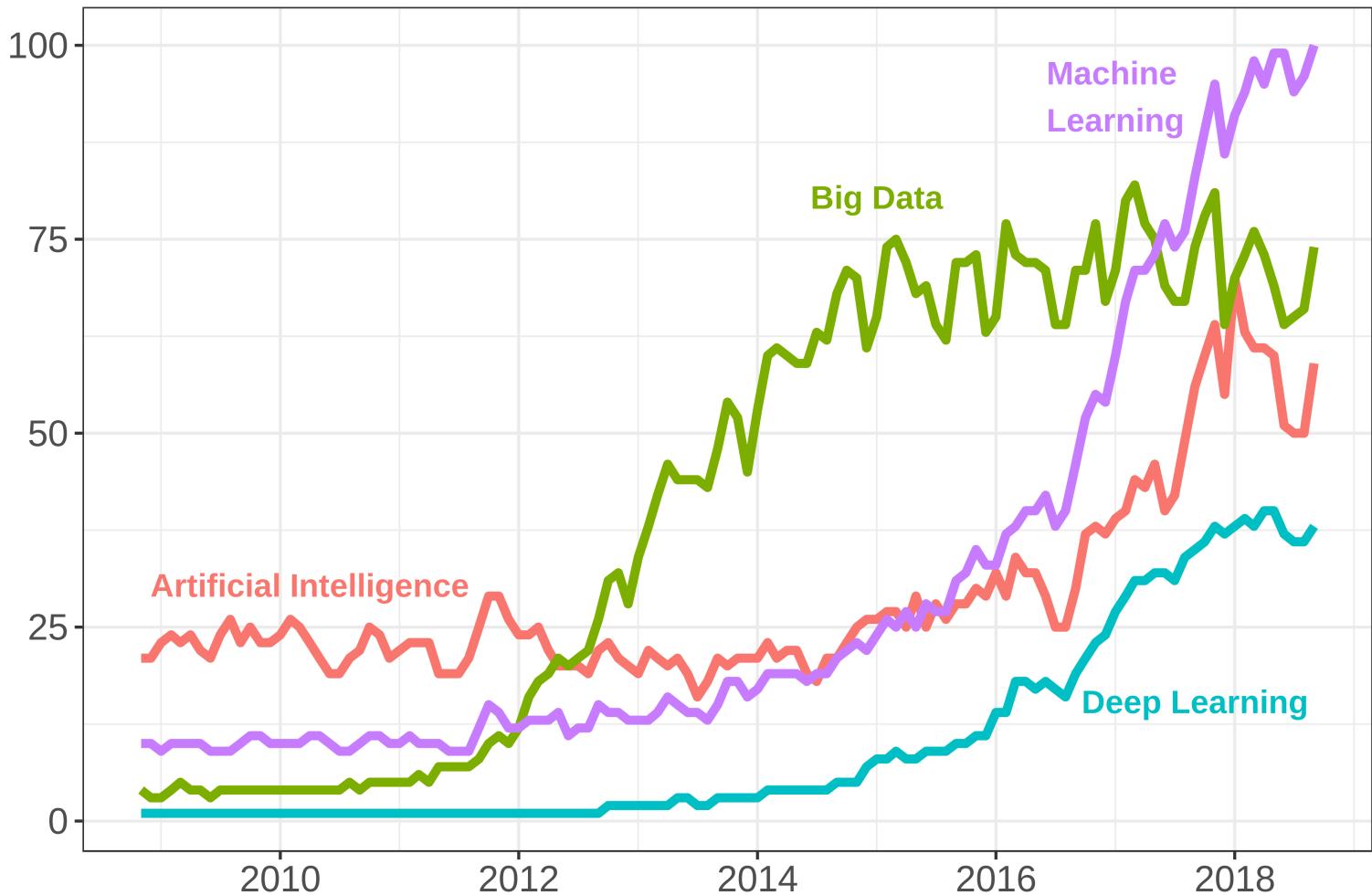


"NVIDIA DGX-2 'The World's Largest GPU' Announced ... with \$399,000 Price Tag"



Adapted from Chollet (2018)

Google Trends Keywords: Last 10 Years



Advances in Deep Learning



xkcd

bitly.com/xai-davidson

Image Classification



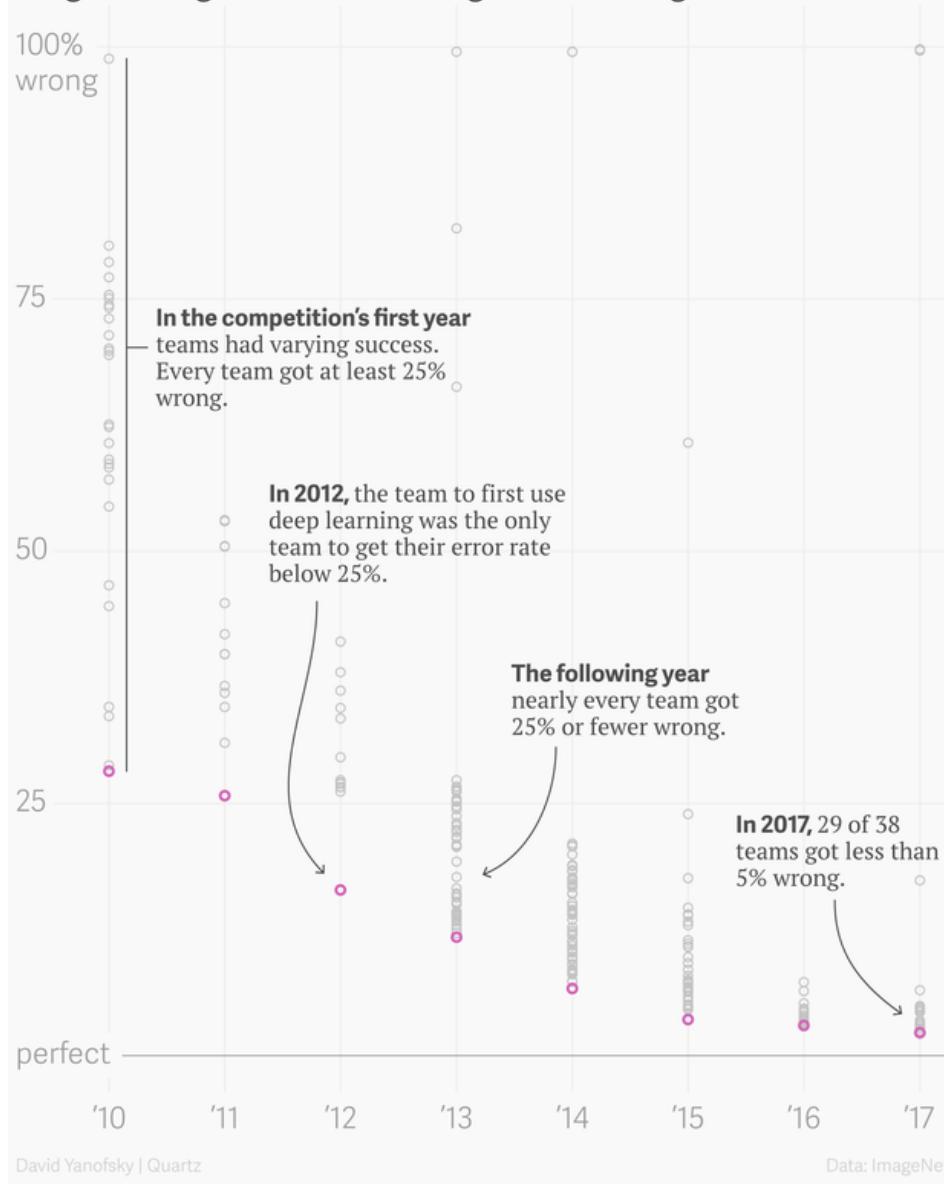
08	02	22	97	38	15	00	40	00	75	04	05	07	78	52	12	50	77	31	66
49	49	99	40	17	81	18	57	60	87	17	40	98	43	69	48	54	56	62	00
81	49	31	73	55	79	14	29	93	71	40	67	58	88	30	03	49	13	36	65
52	70	95	23	04	60	11	42	62	31	68	56	01	32	56	71	37	02	36	91
22	31	16	71	51	63	43	89	41	92	36	54	22	40	40	28	66	33	13	80
24	47	33	60	99	03	45	02	44	75	33	53	78	36	84	20	35	17	12	50
32	98	81	28	64	23	67	10	26	38	40	67	59	54	70	66	18	38	64	70
67	26	20	68	02	62	12	20	95	63	94	39	63	08	40	91	66	49	94	21
24	55	58	05	66	73	99	26	97	17	78	78	96	83	14	88	34	89	63	72
21	36	23	09	75	00	76	44	20	45	35	14	00	61	33	97	34	31	33	95
78	17	53	28	22	75	31	67	15	94	03	80	04	62	16	14	09	53	56	92
16	39	05	42	96	35	31	47	55	58	88	24	00	17	54	24	36	29	85	57
86	56	00	48	35	71	89	07	05	44	44	37	44	60	21	58	51	54	17	58
19	80	81	68	05	94	47	69	28	73	92	13	86	52	17	77	04	89	55	40
04	52	08	83	97	35	99	16	07	97	57	32	16	26	26	79	33	27	98	66
03	06	68	87	57	62	20	72	03	46	33	67	46	55	12	32	63	93	53	69
04	42	16	73	58	35	39	11	24	94	72	18	08	46	29	32	40	62	76	36
20	69	36	41	72	30	23	88	39	52	99	69	82	67	59	85	74	04	36	16
20	73	35	29	78	31	90	01	74	31	49	71	48	66	81	16	23	57	05	54
01	70	54	71	83	51	54	69	16	92	33	48	61	43	52	01	89	19	62	48

What the computer sees

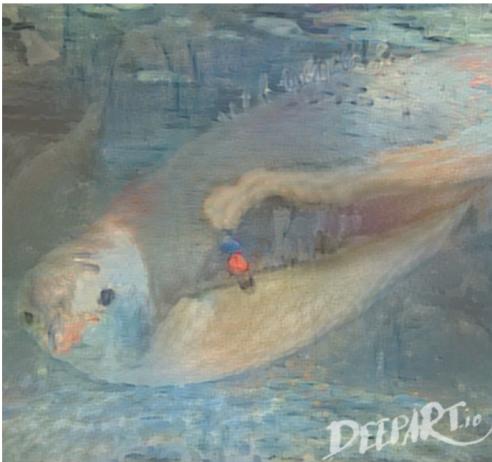
image classification

82% cat
15% dog
2% hat
1% mug

ImageNet Large Scale Visual Recognition Challenge results



Style transfer with DeepArt.io



Style transfer to video



Chan et al., 2018 / video demo

Teaser -- Synthesizing Obama: Learning Lip Sync from Audio



Suwajanakorn, Seitz, and Kemelmacher-Shilzerman (2017)

Don't think it's that easy...



Jake VanderPlas' tweet

bitly.com/xai-davidson

Practical Problems with Deep Learning

Supervised machine learning: labels (y variable) are expensive!

Practical Problems with Deep Learning

Supervised machine learning: labels (y variable) are expensive!

Need **lots** of data

Practical Problems with Deep Learning

Supervised machine learning: labels (y variable) are expensive!

Need **lots** of data

Expensive to train (GPUs)

Practical Problems with Deep Learning

Supervised machine learning: labels (y variable) are expensive!

Need **lots** of data

Expensive to train (GPUs)

Can use pre-trained model, but may need to customize

Practical Problems with Deep Learning

Supervised machine learning: labels (y variable) are expensive!

Need **lots** of data

Expensive to train (GPUs)

Can use pre-trained model, but may need to customize

Architecture & tuning hyper-parameters

Practical Problems with Deep Learning

Supervised machine learning: labels (y variable) are expensive!

Need **lots** of data

Expensive to train (GPUs)

Can use pre-trained model, but may need to customize

Architecture & tuning hyper-parameters

Rare skill (only few years old!)

Practical Problems with Deep Learning

Supervised machine learning: labels (y variable) are expensive!

Need **lots** of data

Expensive to train (GPUs)

Can use pre-trained model, but may need to customize

Architecture & tuning hyper-parameters

Rare skill (only few years old!)

But it gets worse...

Deep fakes...

Regulatory: GDPR

Article 22. Automated individual decision making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
2. Paragraph 1 shall not apply if the decision:
 - (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
 - (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
 - (c) is based on the data subject's explicit consent.
3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.
4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) apply and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

Figure 1: Excerpt from the General Data Protection Regulation, [26]

Goodman and Flaxman (2016)

Regulatory: GDPR

Article 22. Automated individual decision making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
2. Paragraph 1 shall not apply if the decision:
 - (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
 - (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
 - (c) is based on the data subject's explicit consent.
3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.
4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) apply and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

Figure 1: Excerpt from the General Data Protection Regulation, [26]

Goodman and Flaxman (2016)

Although the question on whether GDPR has a "right to explanation" is hotly debated, e.g. Wachter, Mittelstadt, and Floridi, 2016

Adversarial Examples

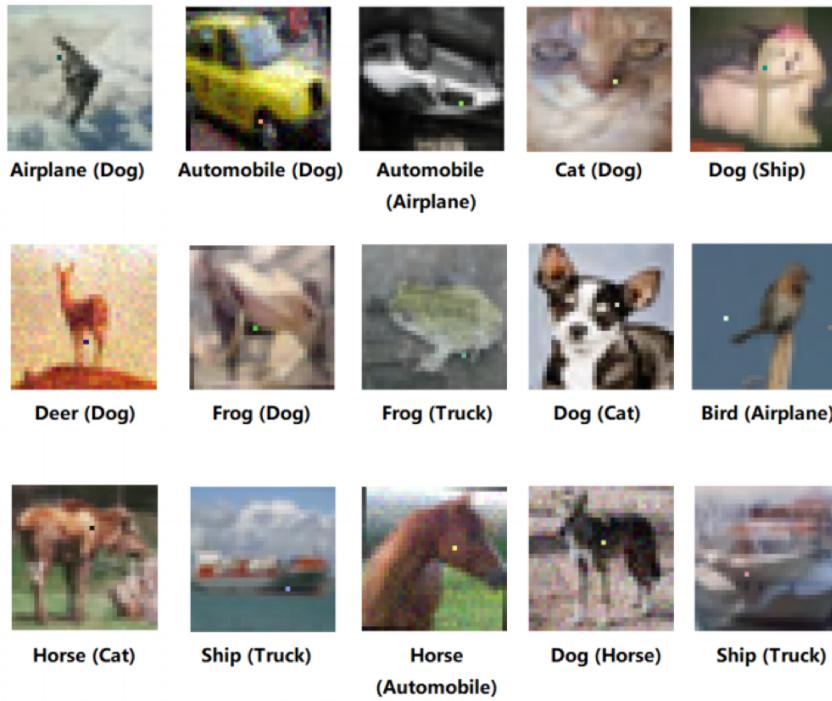


Figure 1. One-pixel attacks created with the proposed algorithm that successfully fooled a target DNN. The original class labels are written below each image with the target class label written inside brackets.

Adversarial Examples

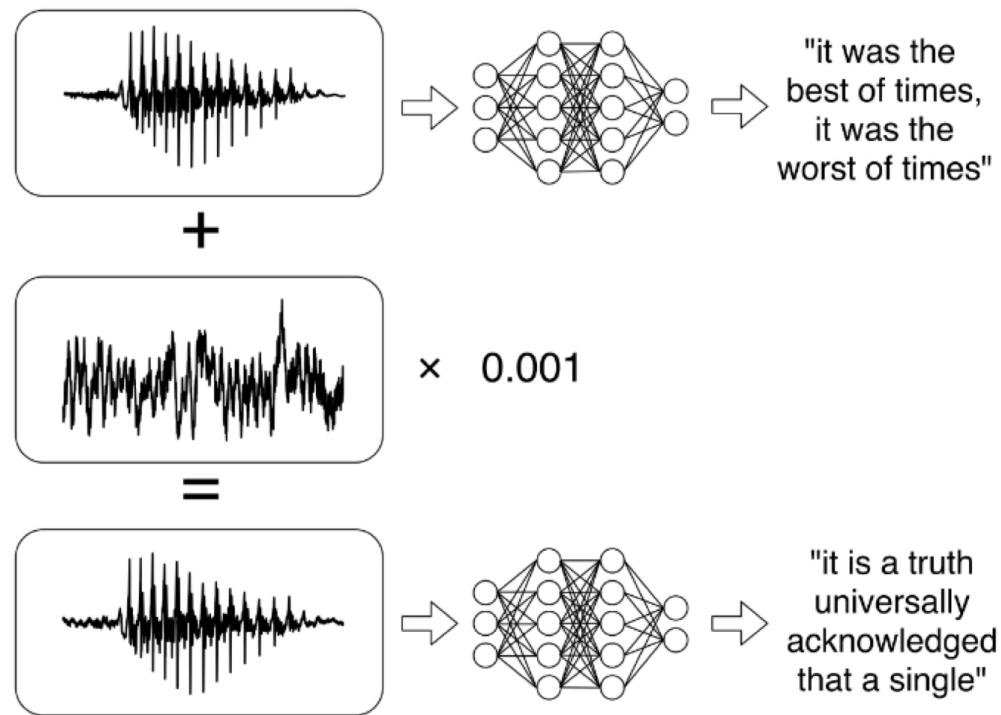


Fig. 1. Illustration of our attack: given any waveform, adding a small perturbation makes the result transcribe as any desired target phrase.

Carlini and Wagner (2018)

bitly.com/xai-davidson

Algorithmic Bias

The image consists of three vertically stacked screenshots from Google Translate and a search interface.

Top Screenshot: A comparison between English and Turkish translations. The English input is "He is a nurse" and "She is a doctor". The resulting Turkish output is "O bir hemşire" and "O bir doktor". This shows a clear gender bias in the machine translation.

Middle Screenshot: Another comparison between English and Turkish. The English input is "O bir hemşire" and "O bir doktor". The resulting Turkish output is "She is a nurse" and "He is a doctor". This illustrates how the algorithm can produce biased results even when the input is in a different language.

Bottom Screenshot: A search interface showing autocomplete suggestions for the query "jews should". The suggestions include:

- jews should be wiped out
- jews should leave israel
- jews should
- jews should get over the holocaust
- jews should go back to poland
- jews should apologize for killing jesus
- jews should all die
- jews should be perfected
- jews should not have a state

This demonstrates how search engines can provide biased or harmful suggestions based on implicit biases in their training data.

Kate Crawford (2017)

bitly.com/xai-davidson

Algorithmic Bias

Color Matters in Computer Vision

Facial recognition algorithms made by Microsoft, IBM and Face++ were more likely to misidentify the gender of black women than white men.



Gender was misidentified in **up to 1 percent of lighter-skinned males** in a set of 385 photos.

Gender was misidentified in **up to 12 percent of darker-skinned males** in a set of 318 photos.



Gender was misidentified in **up to 7 percent of lighter-skinned females** in a set of 296 photos.

Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.

Photos were selected from among those used in Joy Buolamwini's study.

Source: Joy Buolamwini, M.I.T. Media Lab

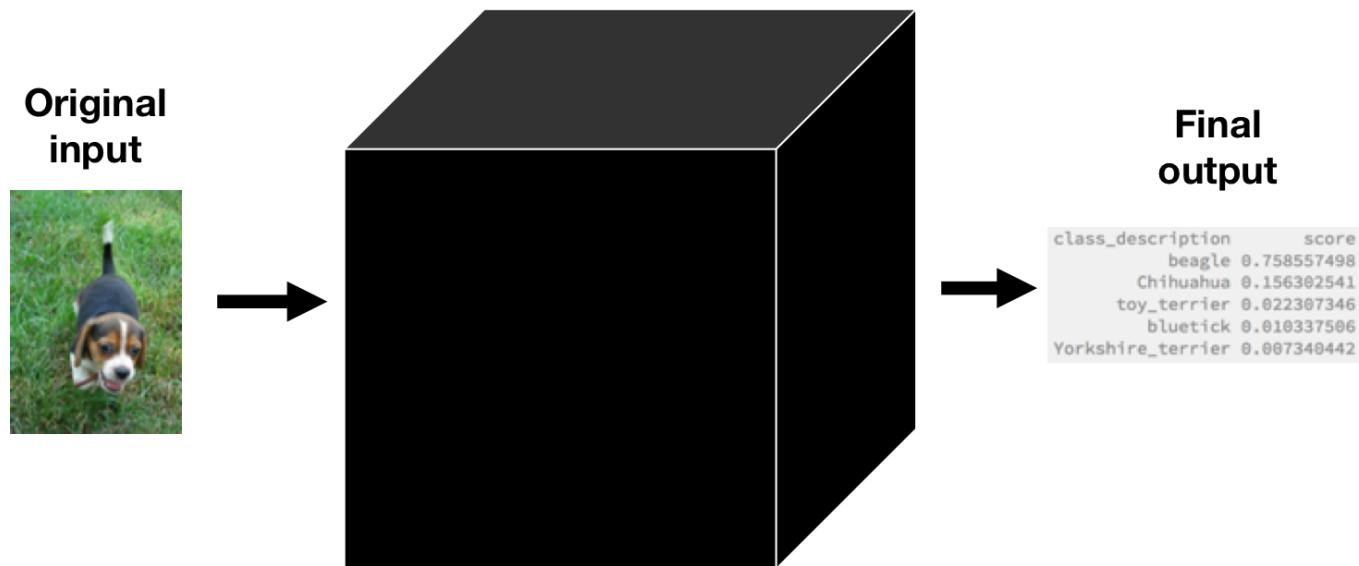
New York Times (2018)

Machine intelligence makes human morals more important | Zeynep Tufekci



| We need to audit our black boxes -Zeynep Tufekci

But it's a black box...



What is Explainable AI?



Explainable AI: DARPA XAI

2015 DARPA (Defense Department Research Arm) program with two goals:

Explainable AI: DARPA XAI

2015 DARPA (Defense Department Research Arm) program with two goals:

- 1) Produce **more explainable** models, while maintaining a high level of learning performance (prediction accuracy)

Explainable AI: DARPA XAI

2015 DARPA (Defense Department Research Arm) program with two goals:

- 1) Produce **more explainable** models, while maintaining a high level of learning performance (prediction accuracy)
- 2) Enable human users to **understand, appropriately trust, and effectively manage** the emerging generation of artificially intelligent partners.

Explainable AI: DARPA XAI

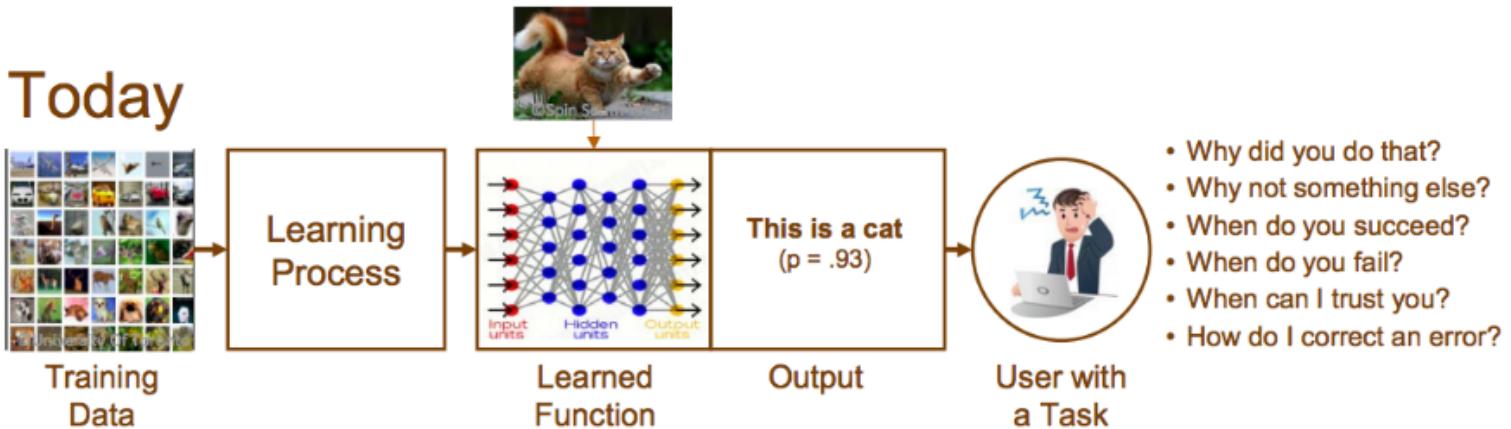
2015 DARPA (Defense Department Research Arm) program with two goals:

- 1) Produce **more explainable** models, while maintaining a high level of learning performance (prediction accuracy)
- 2) Enable human users to **understand, appropriately trust, and effectively manage** the emerging generation of artificially intelligent partners.

Sometimes more generally called **interpretable machine learning**

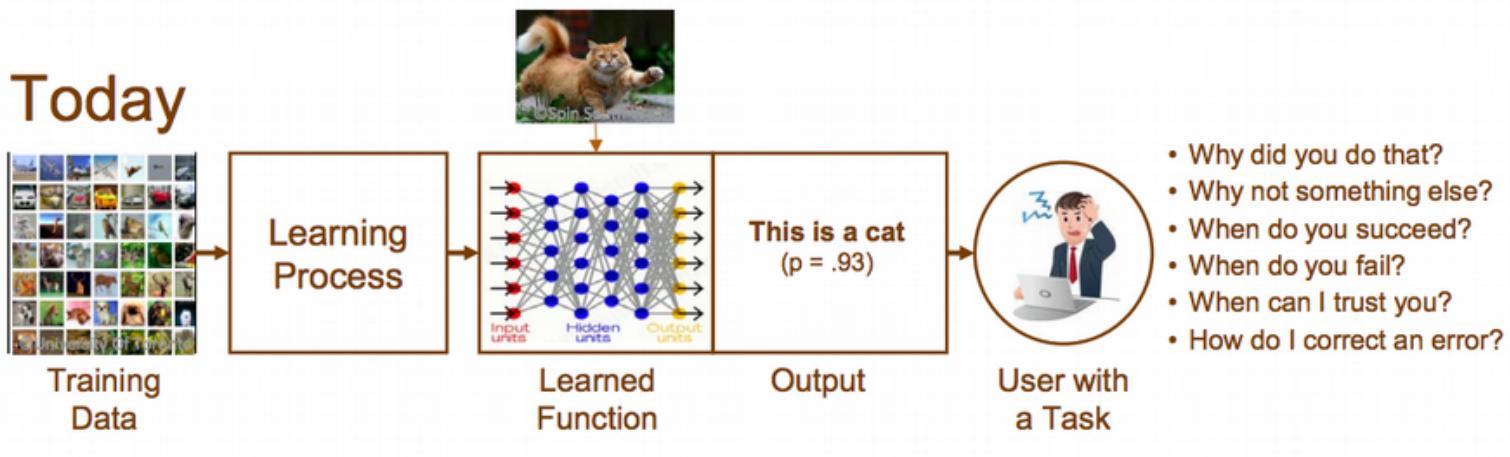
Explainable AI

Today

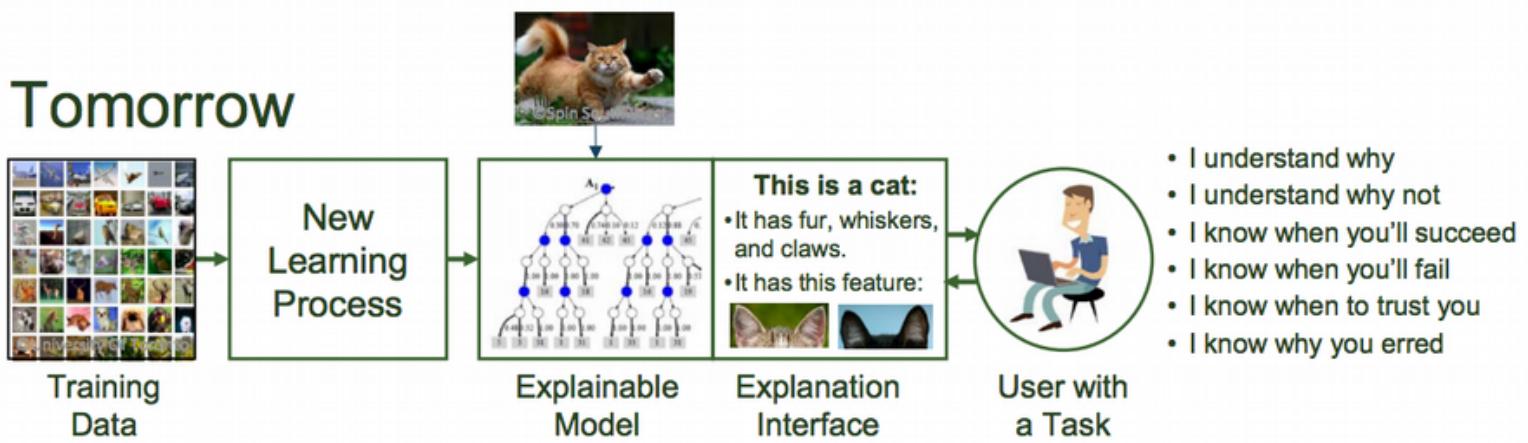


Explainable AI

Today



Tomorrow

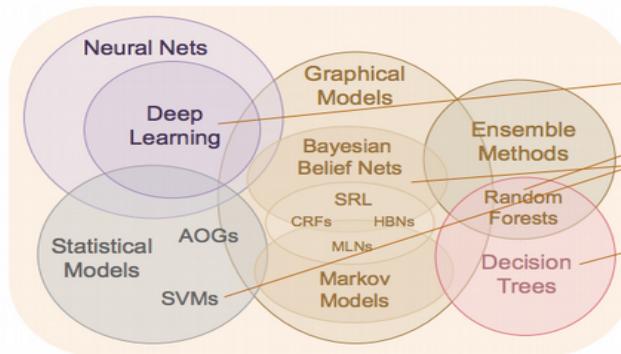


XAI Approaches

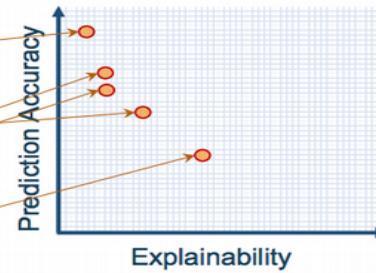
New Approach

Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)

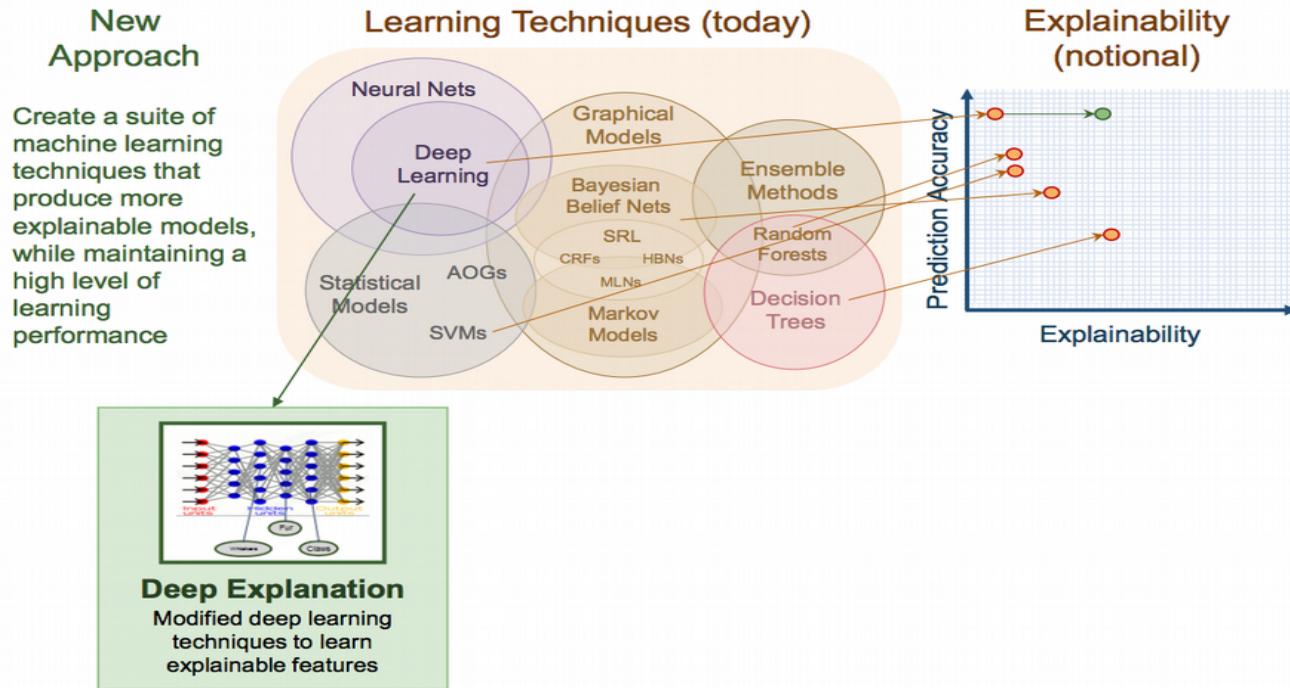


Explainability (notional)



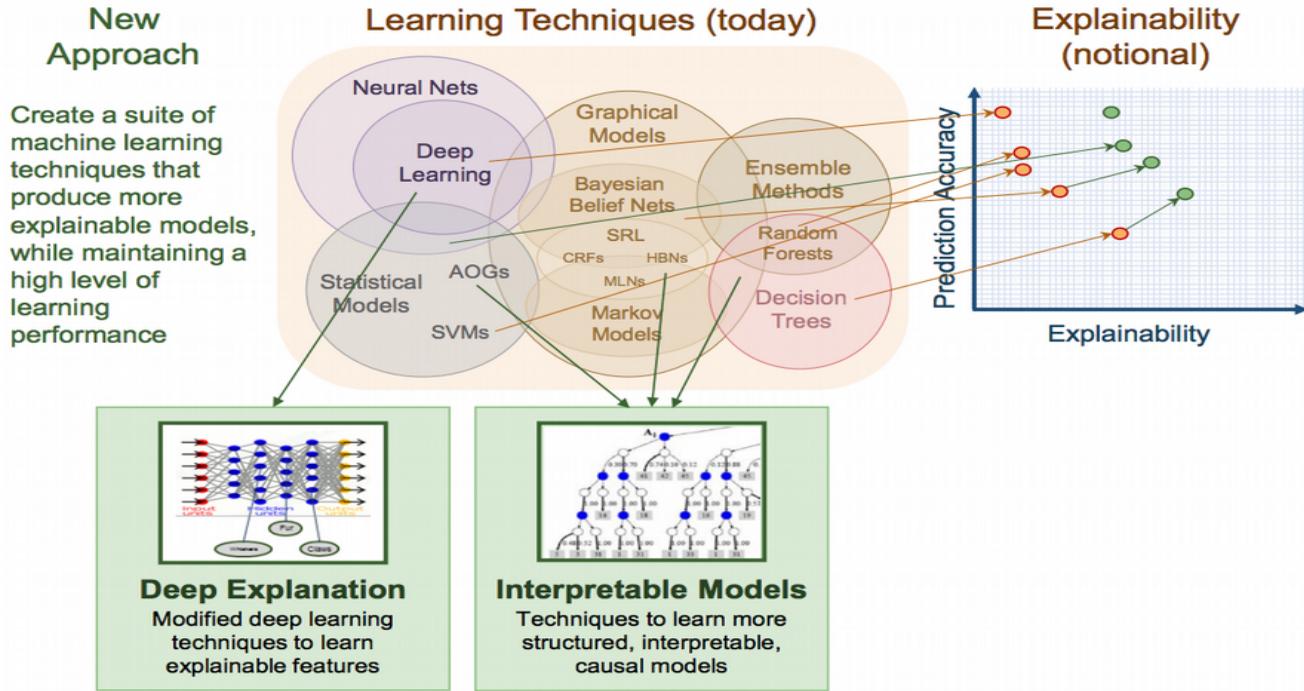
DARPA XAI

XAI Approaches



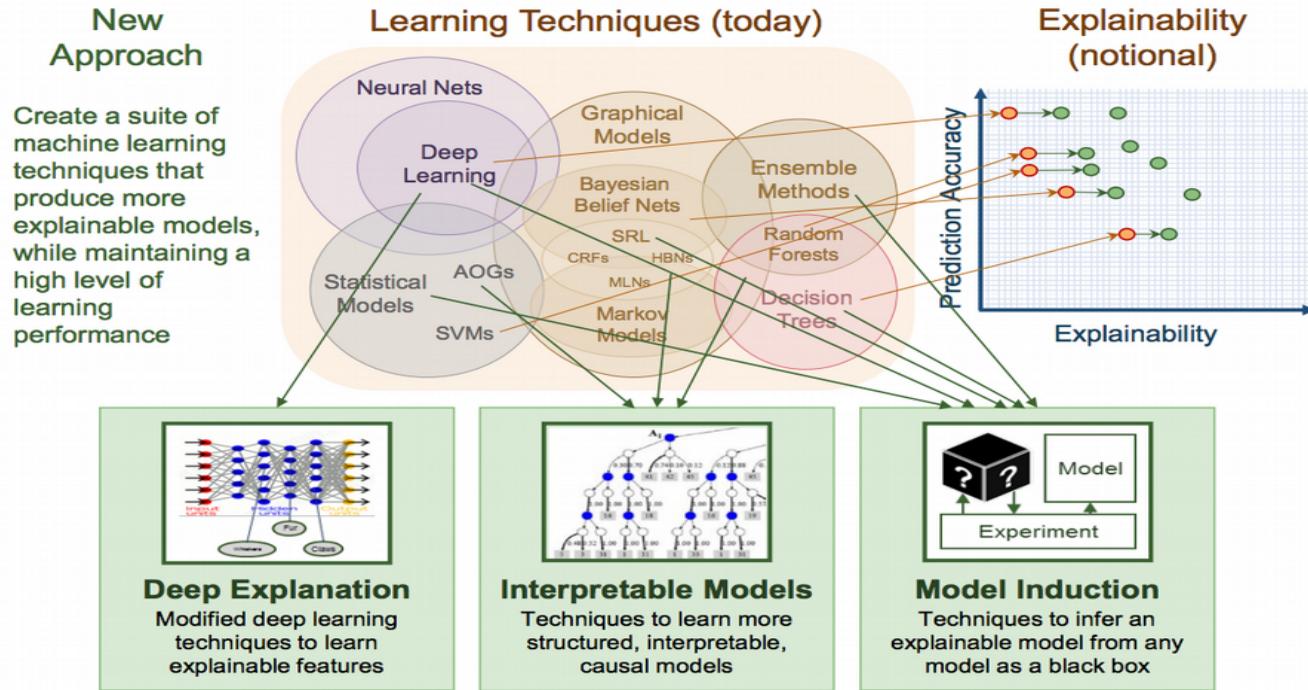
DARPA XAI

XAI Approaches



DARPA XAI

XAI Approaches



DARPA XAI

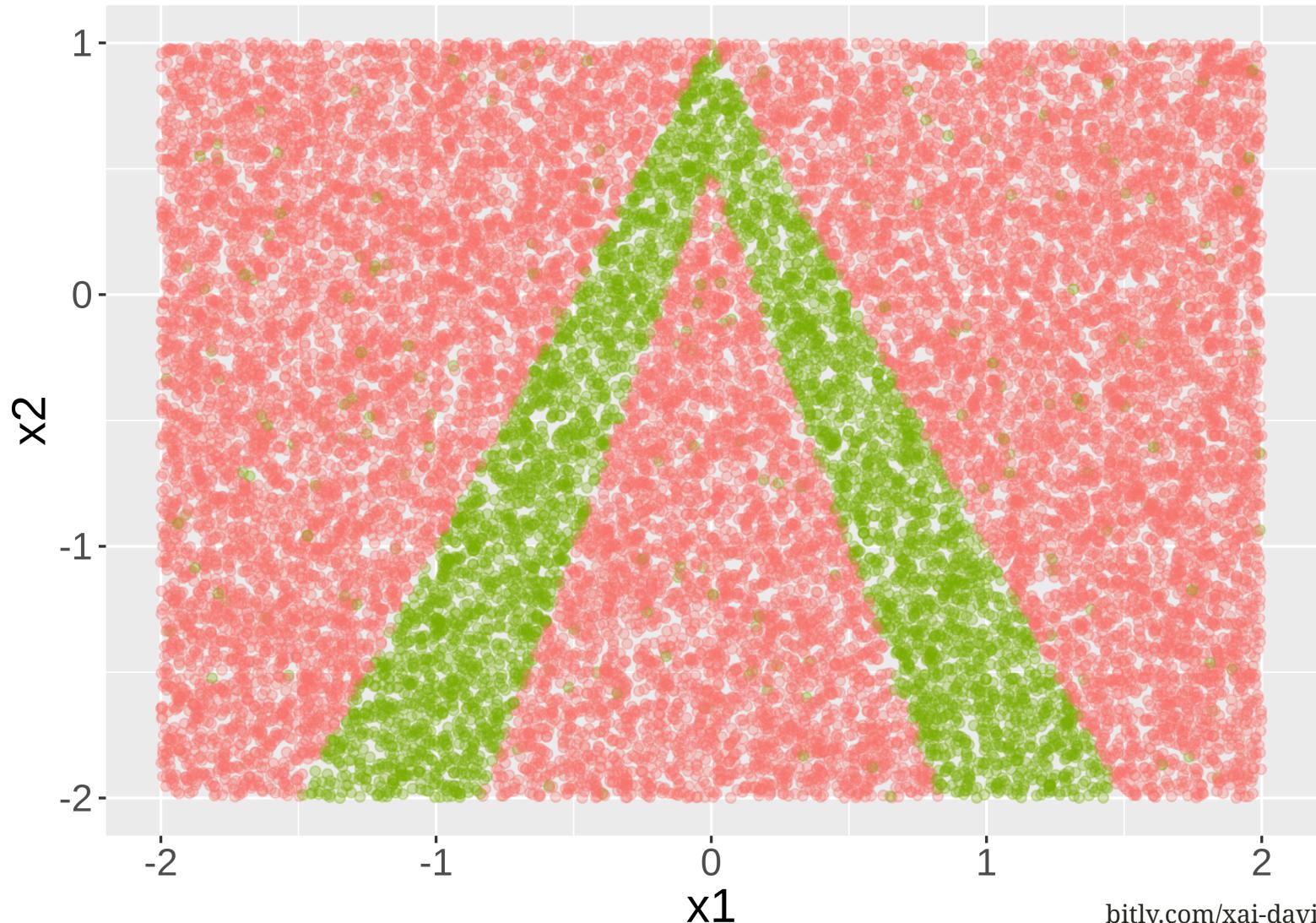
Model Induction (Agnostic): LIME

KDD2016 paper 573

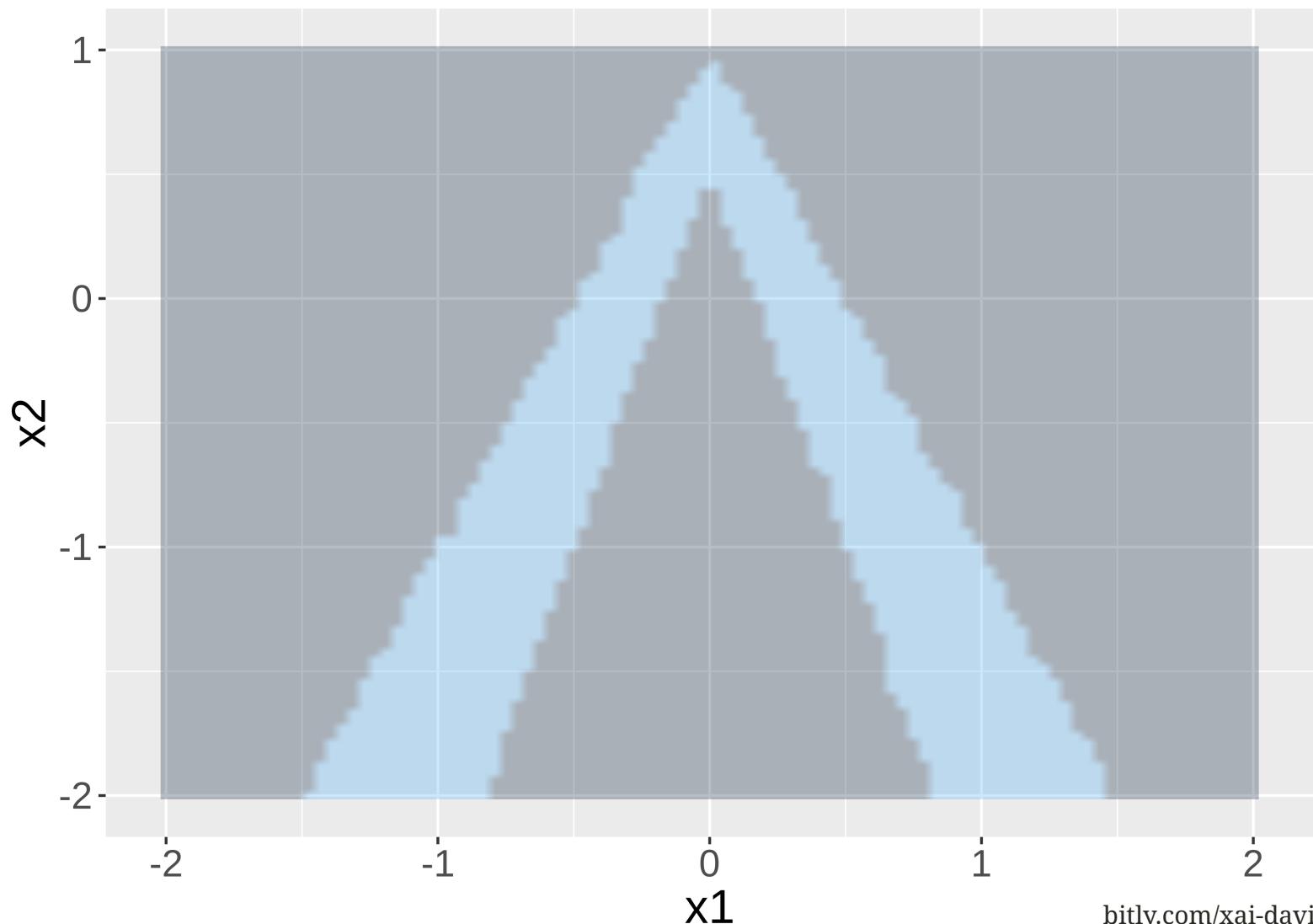


Ribeiro, Singh, Guestrin, 2016

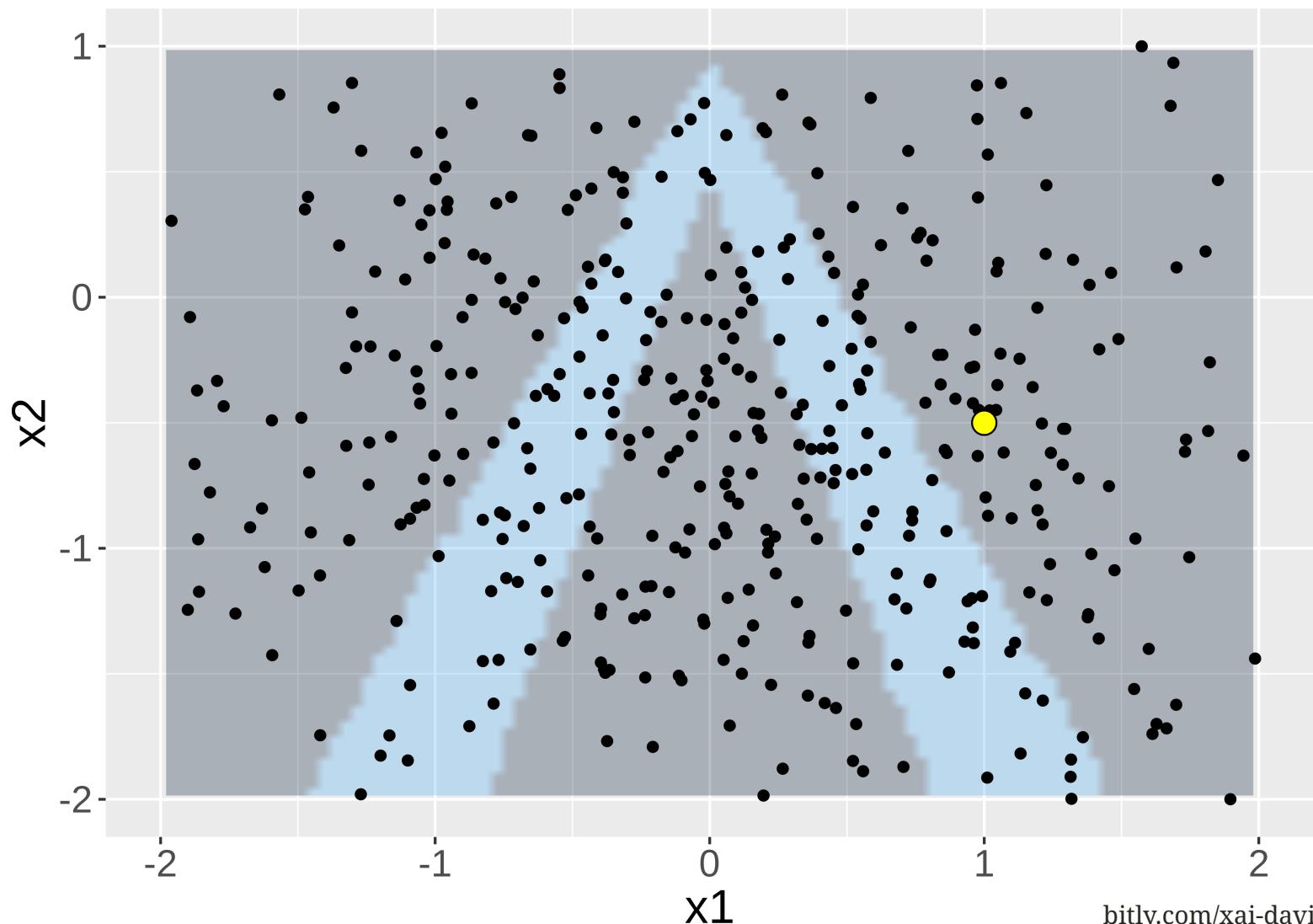
LIME: Intuition / Molnar, 2018 / RStudio.Cloud Project



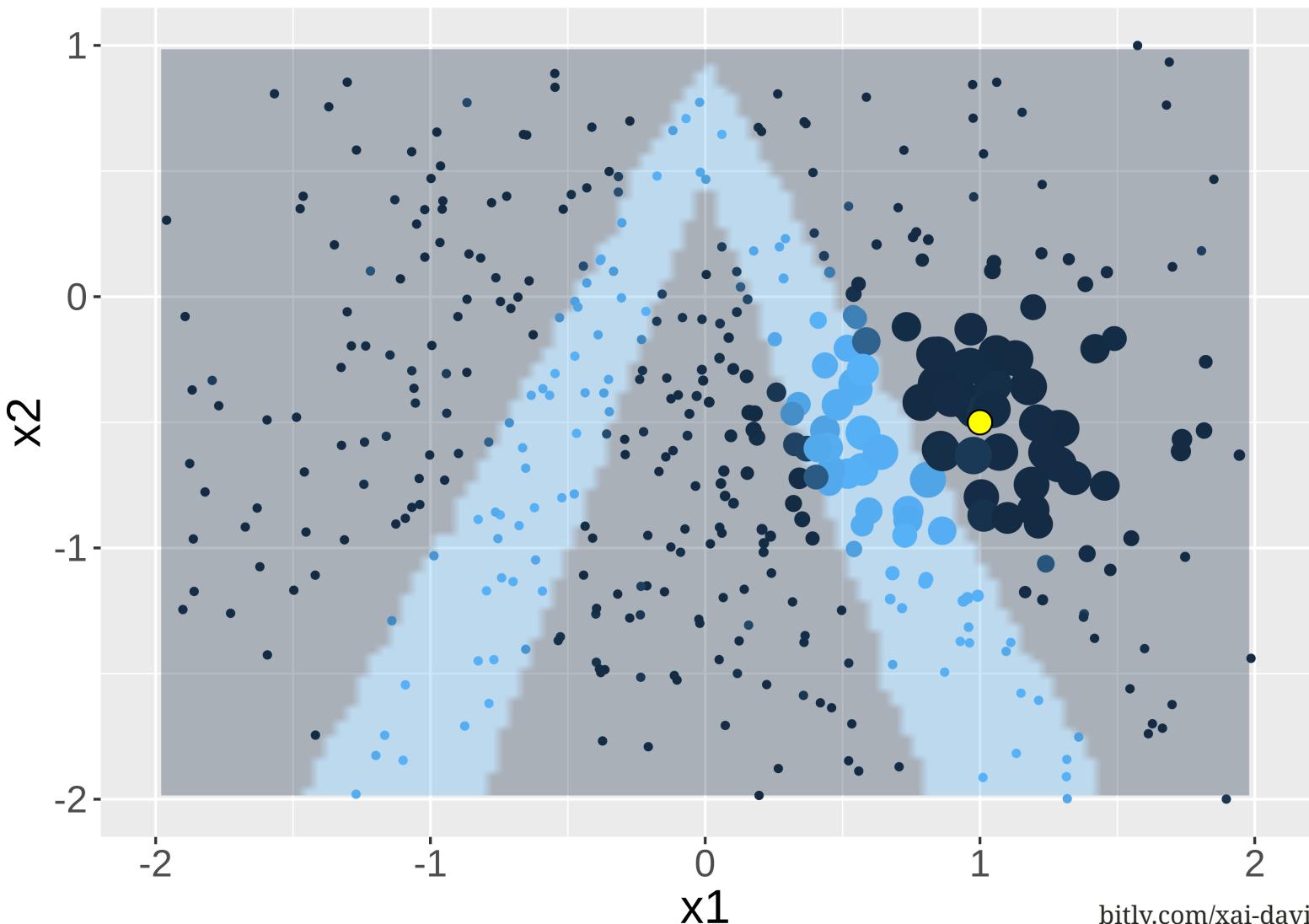
LIME: Intuition / Molnar, 2018 / RStudio.Cloud Project



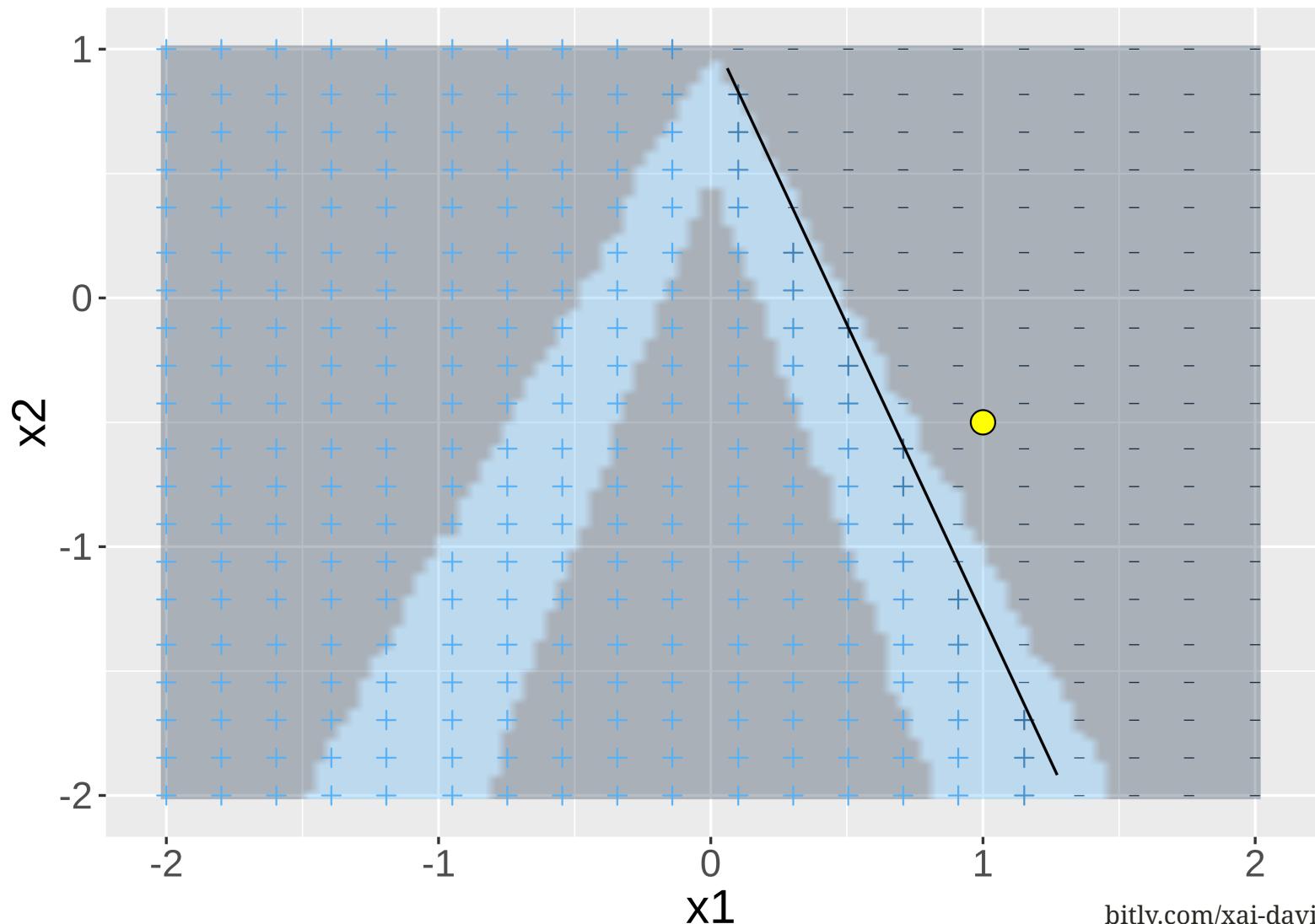
LIME: Intuition / Molnar, 2018 / RStudio.Cloud Project



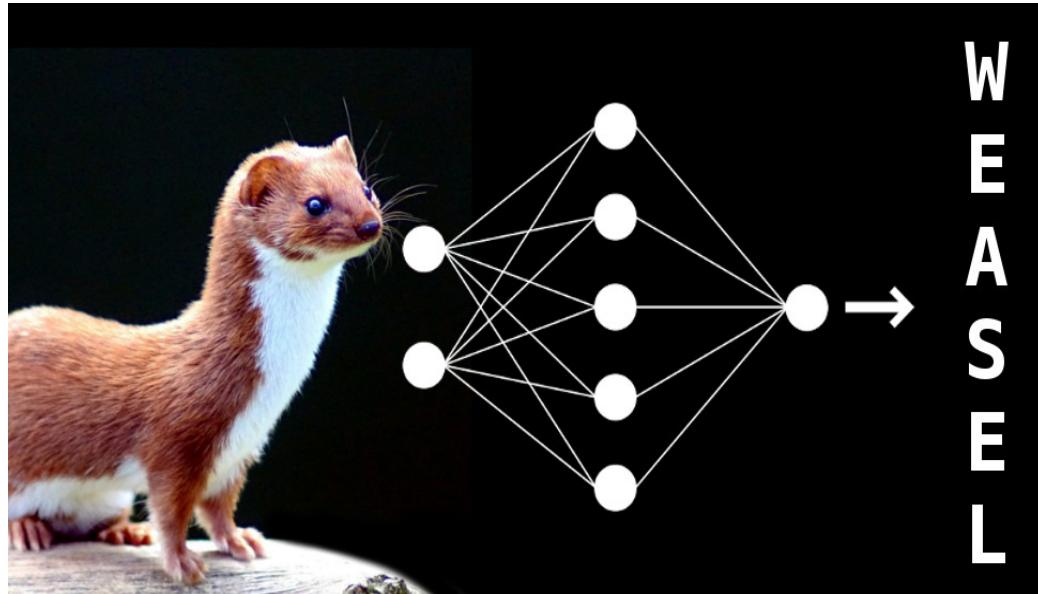
LIME: Intuition / Molnar, 2018 / RStudio.Cloud Project



LIME: Intuition / Molnar, 2018 / RStudio.Cloud Project



Application: LIME on image classification in R

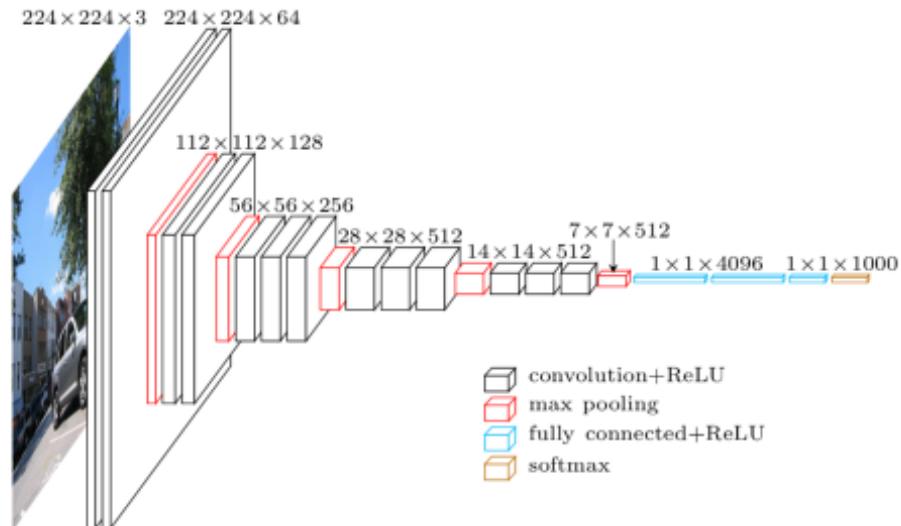


Thomas Lin Pederson's Blogpost on TensorFlow for R Blog

Use pre-trained vgg16

```
# see https://keras.rstudio.com/
library(keras)

# create pre-trained vgg16 as model
model <- application_vgg16(
  weights = "imagenet",
  include_top = TRUE
)
```





bitly.com/xai-davidson

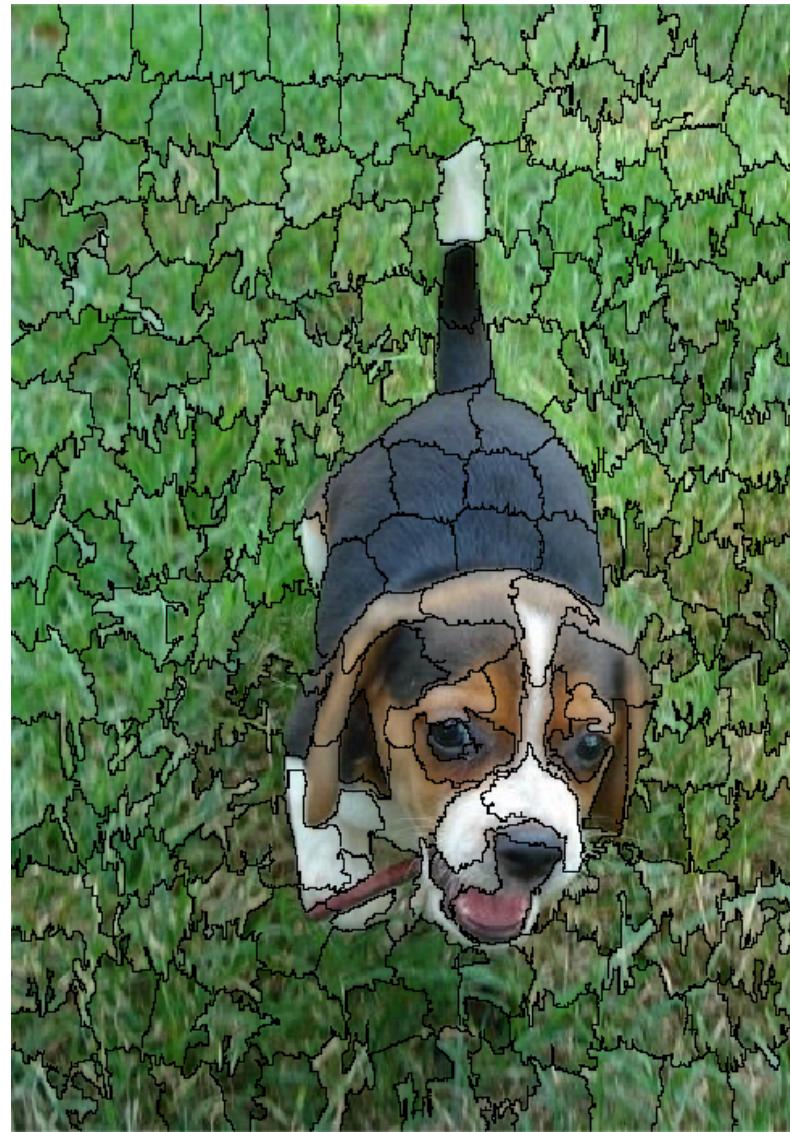
Predict Image using vgg16

```
# set image's local file path  
img_path <- file.path('rusty-puppy.jpg')  
  
# create prediction (res) of the image (after prep-) and the model  
res <- predict(model, image_prep(img_path))  
  
# get top 5 predictions  
imagenet_decode_predictions(res)
```

```
# class_description      score  
#           beagle 0.758557498  
#       Chihuahua 0.156302541  
#      toy_terrier 0.022307346  
#      bluetick 0.010337506  
# Yorkshire_terrier 0.007340442
```



```
plot_superpixels(img_path, n_superpixels = 200, weight = 40)
```



```
library(lime)

# get model labels
model_labels <- system.file(
  'extdata',
  'imagenet_labels.rds',
  package = 'lime') %>%
  readRDS() # read in rds file

# create classifier
classifier <- as_classifier(model, model_labels)
```

```
library(lime)

# get model labels
model_labels <- system.file(
  'extdata',
  'imagenet_labels.rds',
  package = 'lime') %>%
  readRDS() # read in rds file

# create classifier
classifier <- as_classifier(model, model_labels)

# create explainer for given image
explainer <- lime(img_path, classifier, image_prep)

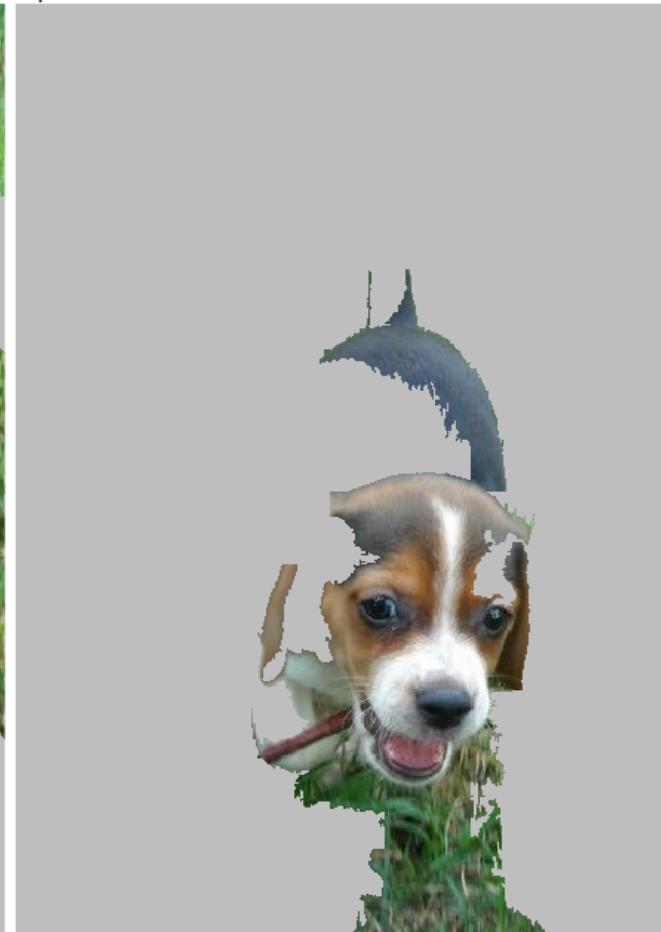
# takes 10+ min on CPU (ideally use GPU(s)!)
explanation <- explain(img_path,
                        # image
                        explainer,
                        # explainable model
                        n_labels = 2,
                        # choose top 2 classes
                        n_features = 20) # use 20 features

plot_image_explanation(explanation,
                       # explanation
                       display = 'block', # block-mode
                       threshold = 0.01)
```

Label: beagle
Probability: 0.75
Explanation Fit: 0.63

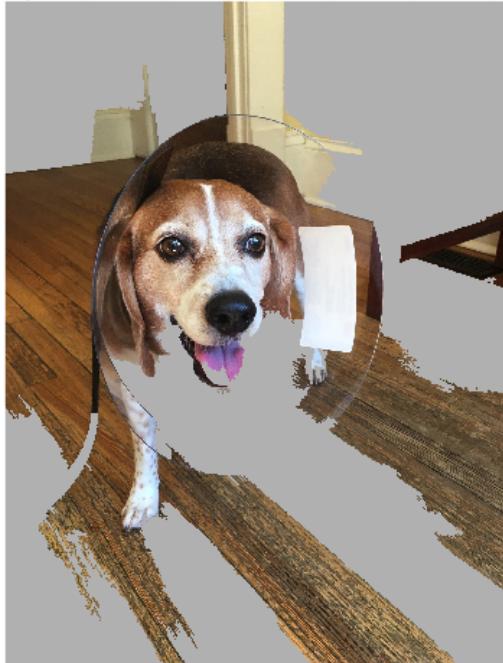


Label: Chihuahua
Probability: 0.16
Explanation Fit: 0.33

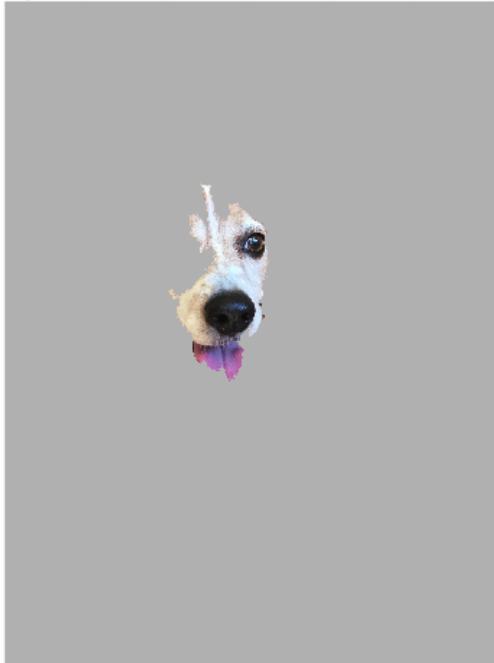


The cone of shame...

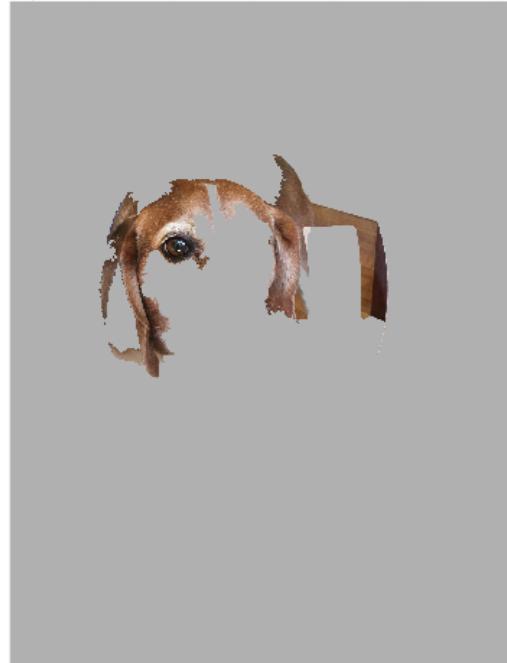
Label: beagle
Probability: 0.94
Explanation Fit: 0.64



Label: bluetick
Probability: 0.013
Explanation Fit: 0.56



Label: basset, basset hound
Probability: 0.0063
Explanation Fit: 0.26



Or in disguise...



```
plot_image_explanation(explanation, threshold = 0.01)
```

Label: cradle
Probability: 0.6
Explanation Fit: 0.33



Label: bassinet
Probability: 0.081
Explanation Fit: 0.29



Label: pajama, pyjama, pj's, jammies
Probability: 0.064
Explanation Fit: 0.20



```
plot_image_explanation(explanation, threshold = 0.01)
```



```
plot_image_explanation(explanation, threshold = 0.001)
```



@smironchuk's Pinterest Ewok Costumes

bitly.com/xai-davidson

```
interactive_text_explanations(cfpbExplanation) # run shiny app
```

LIME: CFPB Comments to Predict Monetary Award (1 = Yes, 0 = No)

The bank charged a fee for me to check my balance on another bank's ATM. This is ridiculous and I have never seen this fee for doing this before.

Quantity of permutations to generate

5000

Word selection strategies

auto

Number of words to select

1 3 20

Text Weights

The bank charged a fee for me to check my balance on another bank's ATM. This is ridiculous and I have never seen this fee for doing this before.

Label predicted: 1 (85.48%)

Explainer fit: 0.96

lime Package

bitly.com/xai-davidson

Caveats to LIME

Possibly slow for images (less for text or tabular data).

Good for local explanations, not for global explanations.

Cognitive psychology: what makes a good explanation?

- Can cognitive theories on visual attention provide clues on how person understands explanation?
- What happens if someone's prior knowledge conflicts with explanations (e.g., cognitive biases)?

Other Approaches

- Anchor approach by Ribeiro, Singh, and Guestrin, 2018 (see appendix)
- Global surrogate models

Explainable AI Resources

Christoph Molnar's Interpretable Machine Learning Book

Deep explanation

- Olah et al.'s "Building Block of Interpretability"

Bayesian networks and causal inference

- Pearl & Mackenzie's "The Book of Why"

Cognitive science take on deep learning:

- Gary Marcus' "Deep Learning: A Critical Appraisal"
- 2017 Lecun & Marcus NYU Debate

Questions & Discussion: Thanks!

 github.com/wesslen

 wesslen.github.io

Anchor-based Explanations

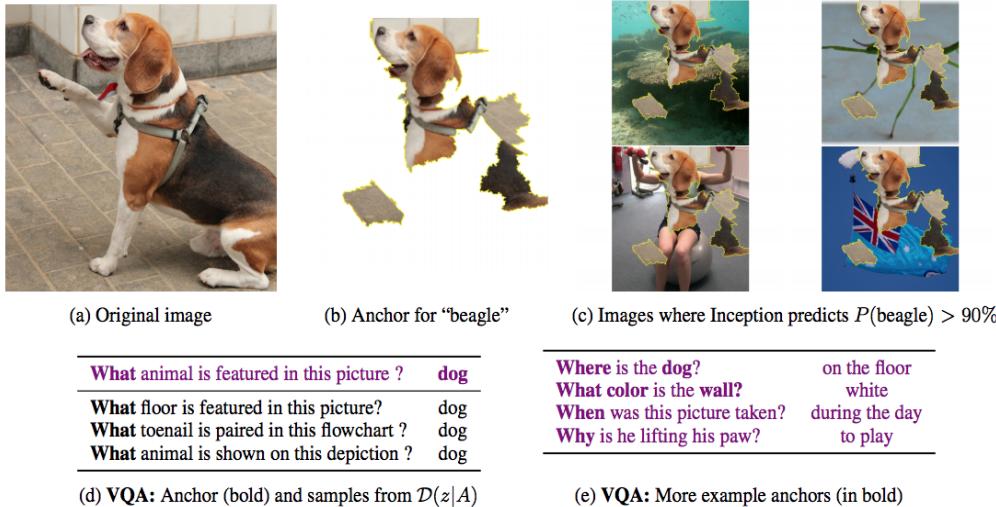


Figure 3: Anchor Explanations for Image Classification and Visual Question Answering (VQA)

Ribeiro, Singh, and Guestrin, 2018