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.

bitly.com/xai-davidson

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

bitly.com/xai-davidson

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

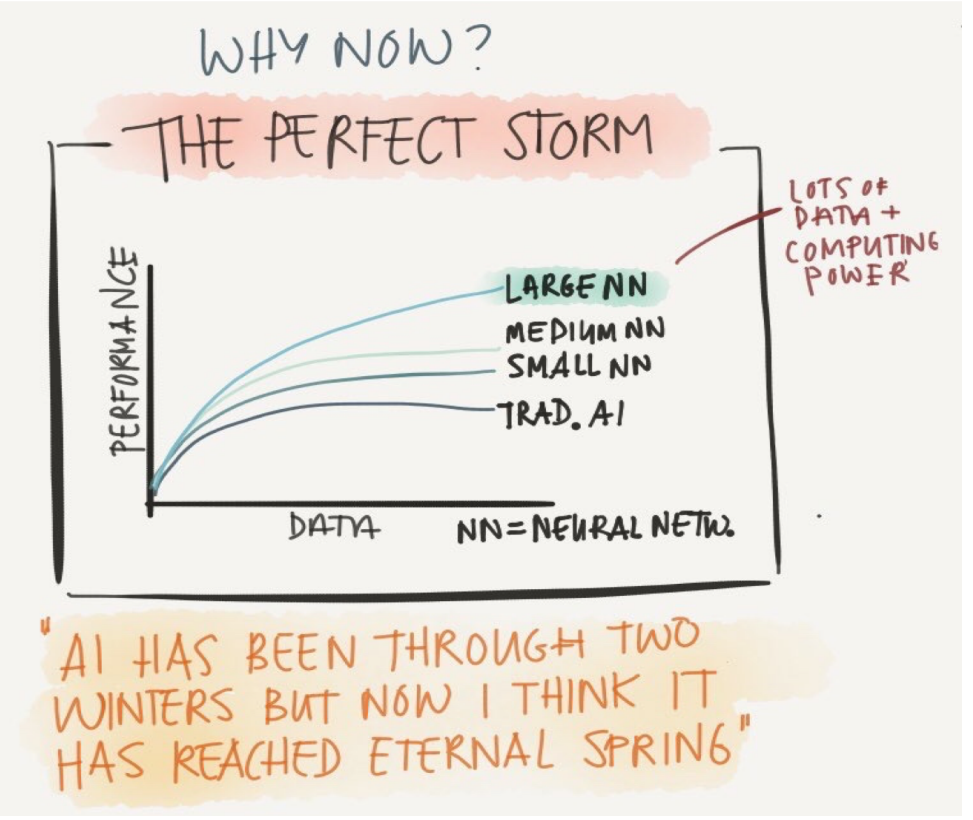
- Taught UNCC workshops (https://github.com/wesslen) in R for text, social media, data viz.
- Teaching Visual Analytics course for UNCC Data Science program in Spring 2019

Andrew Ng
@AndrewYNg

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

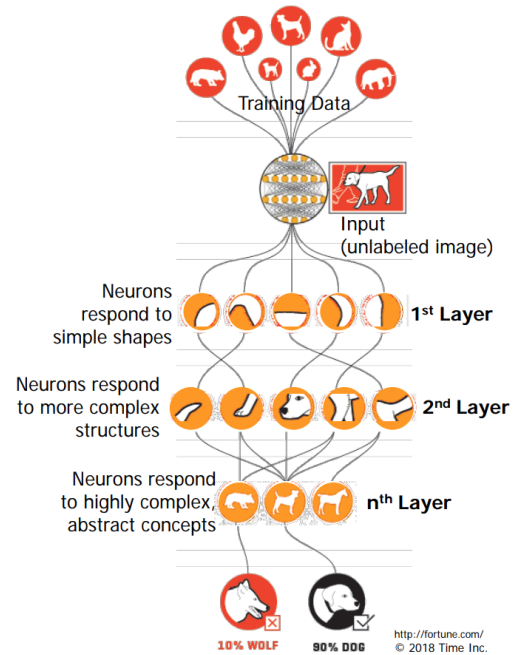
361 12:47 PM - May 26, 2016 · Mountain View, CA

346 people are talking about this

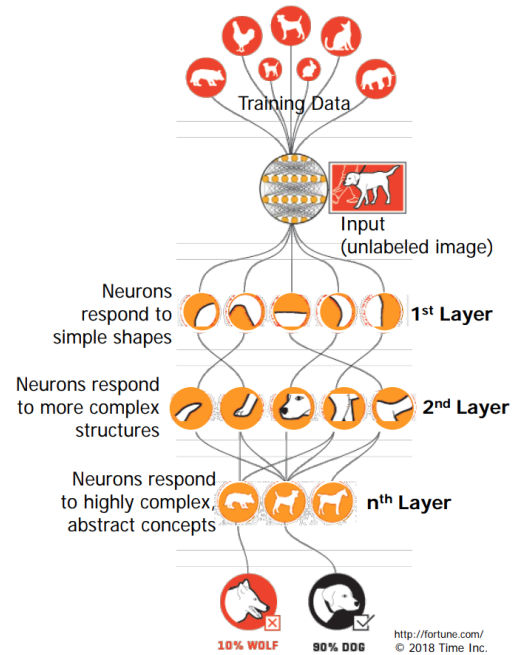


Andrew Ng on YouTube & @TessFernandes

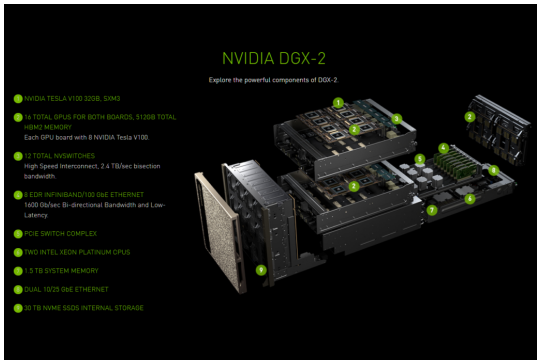
Neural Networks



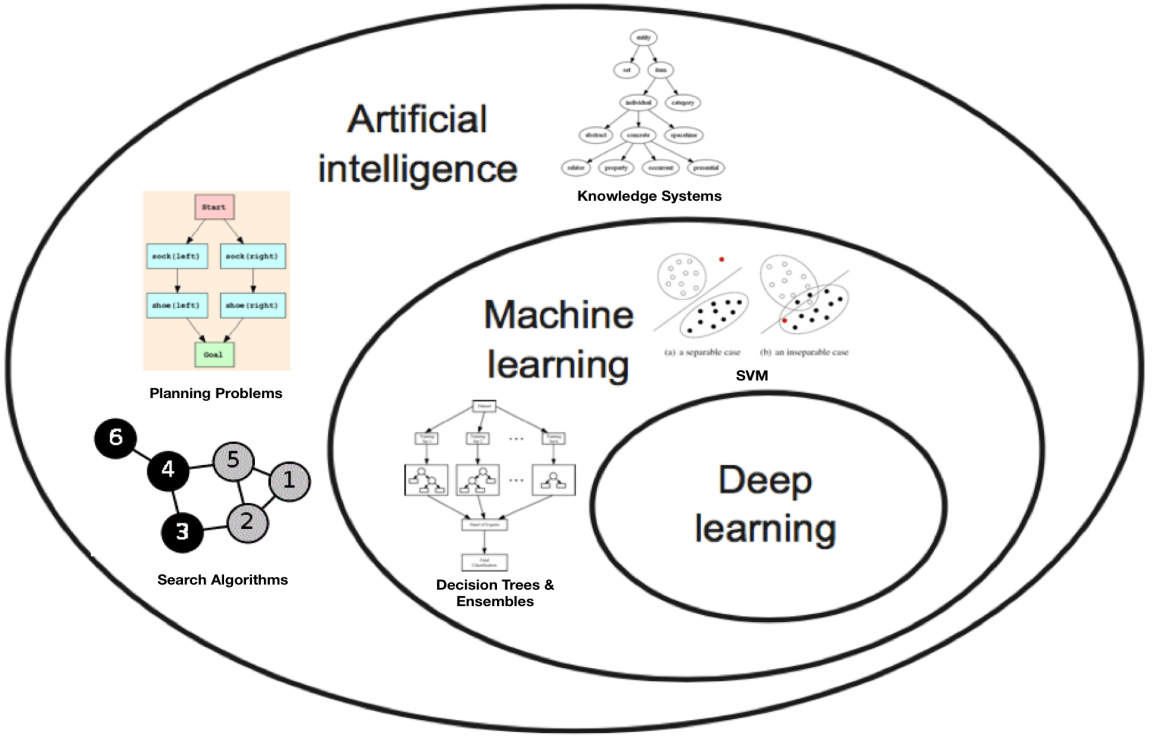
Neural Networks



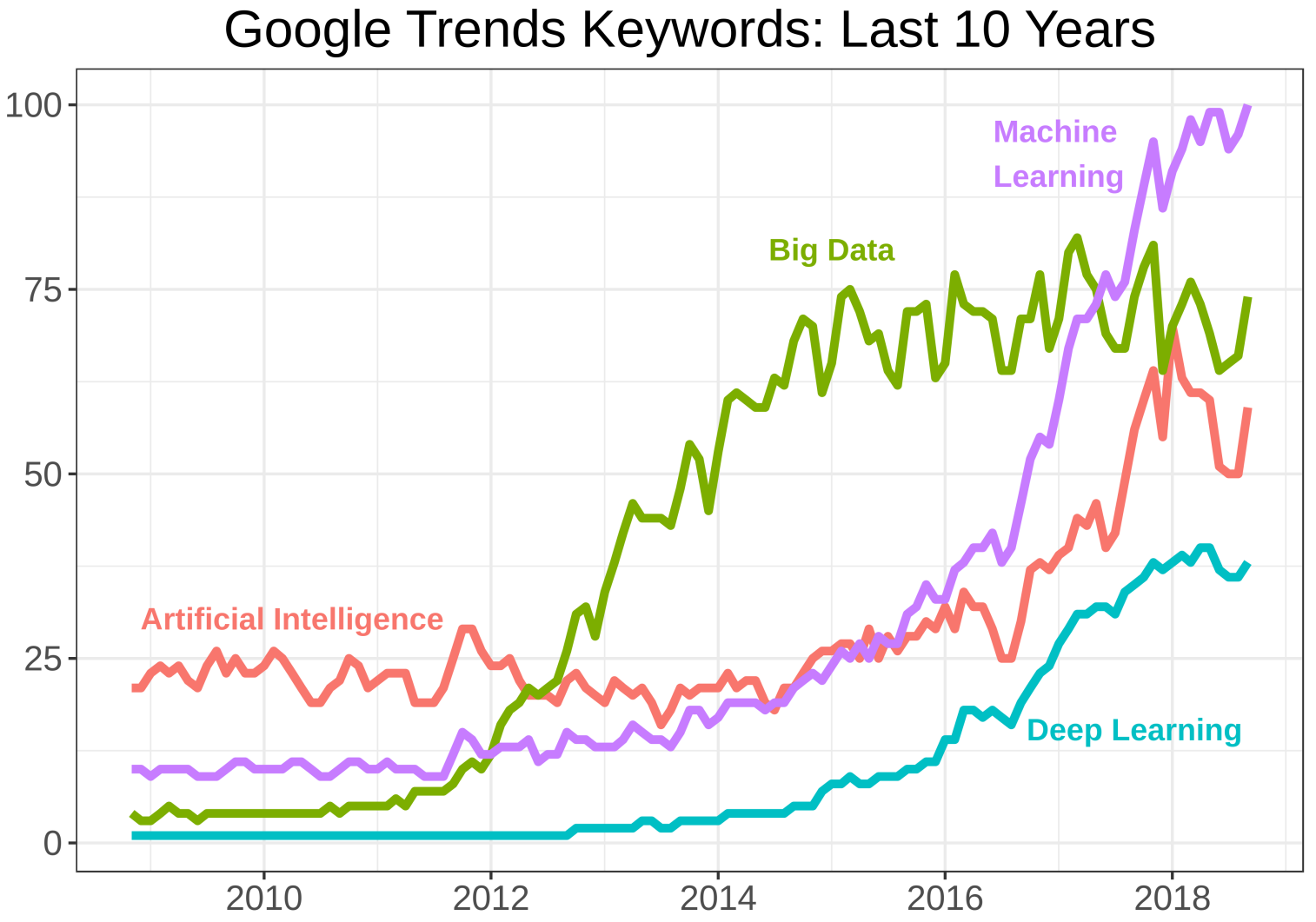
Large-Scale Computing



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



Adapted from Chollet (2018)

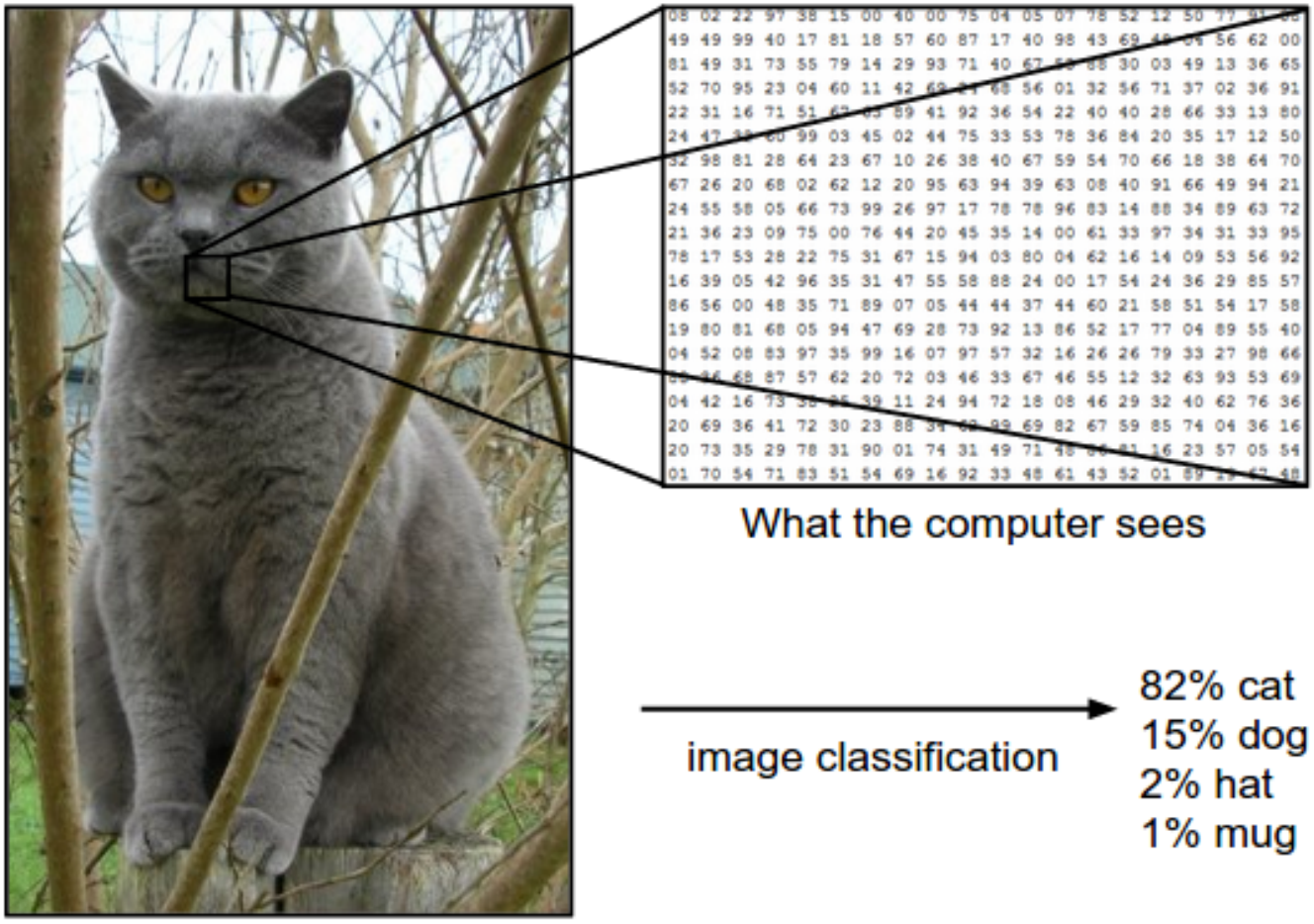


Google Trends

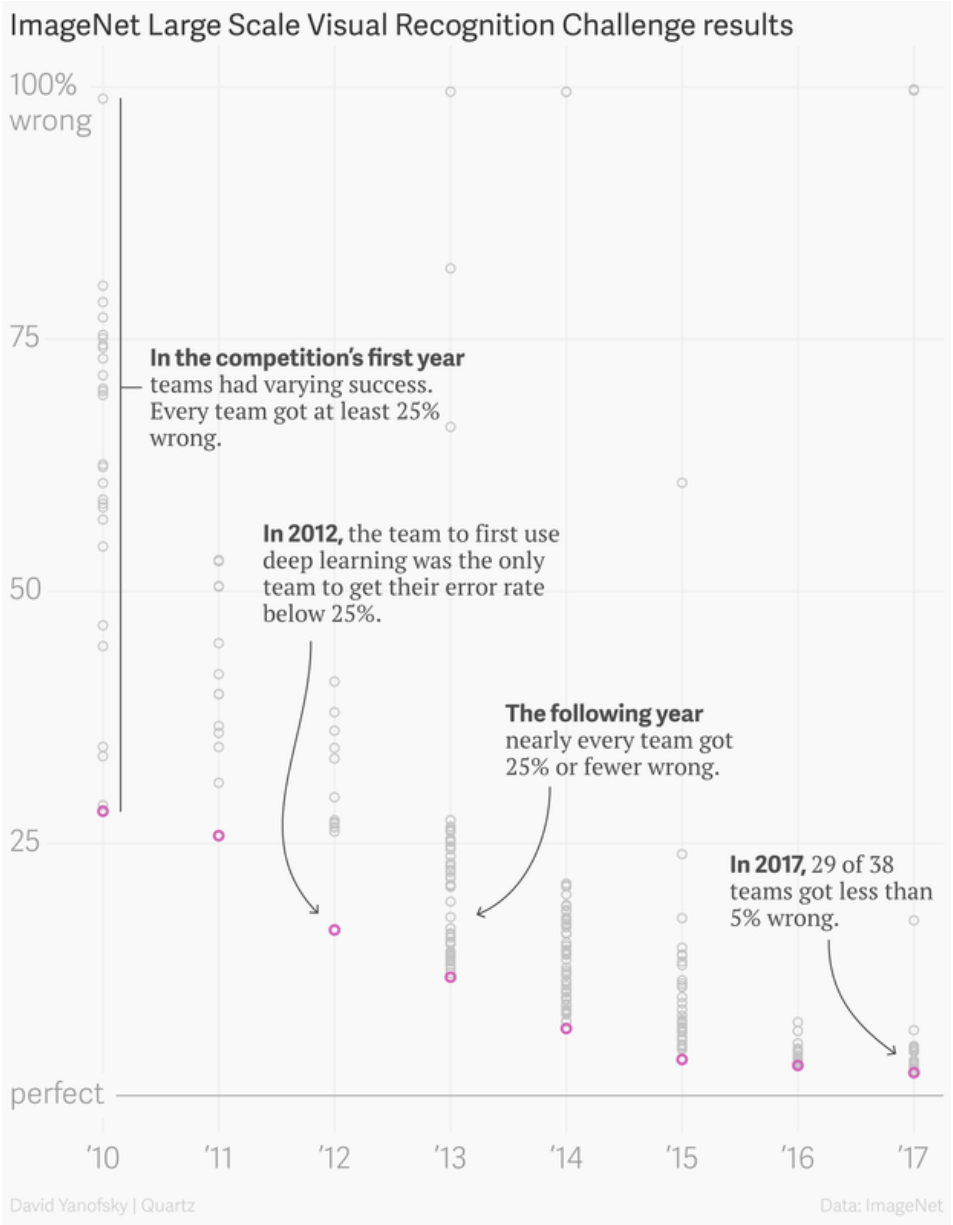
Advances in Deep Learning

xkcd

Image Classification



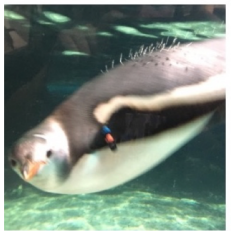
Stanford CS231 Course

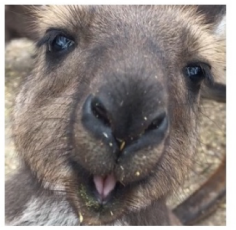


Dave Gershgorn's Quartz Blog on ImageNet

bitly.com/xai-davidson

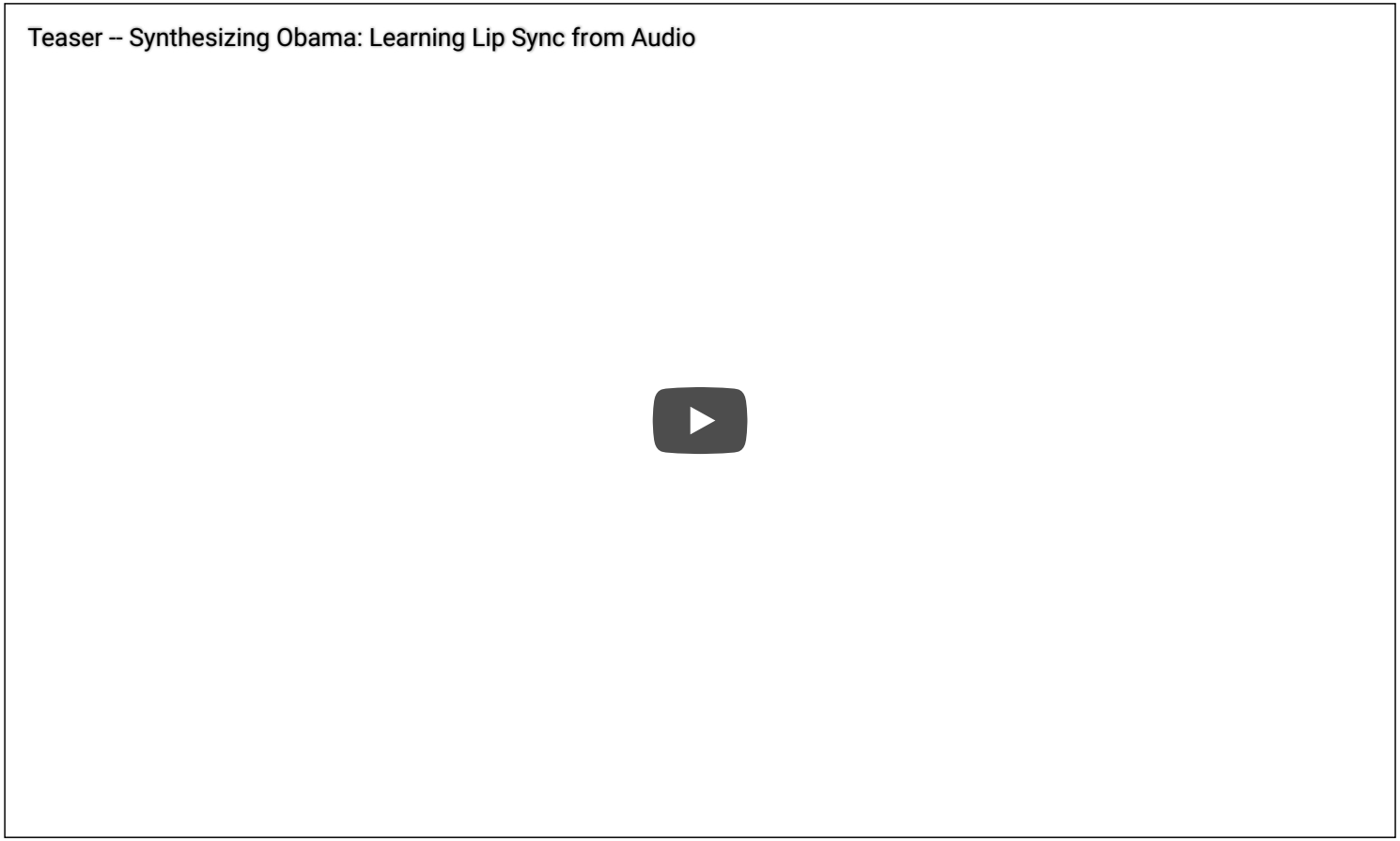
Style transfer with DeepArt.io





Style transfer to video

Chan et al., 2018 / video demo



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

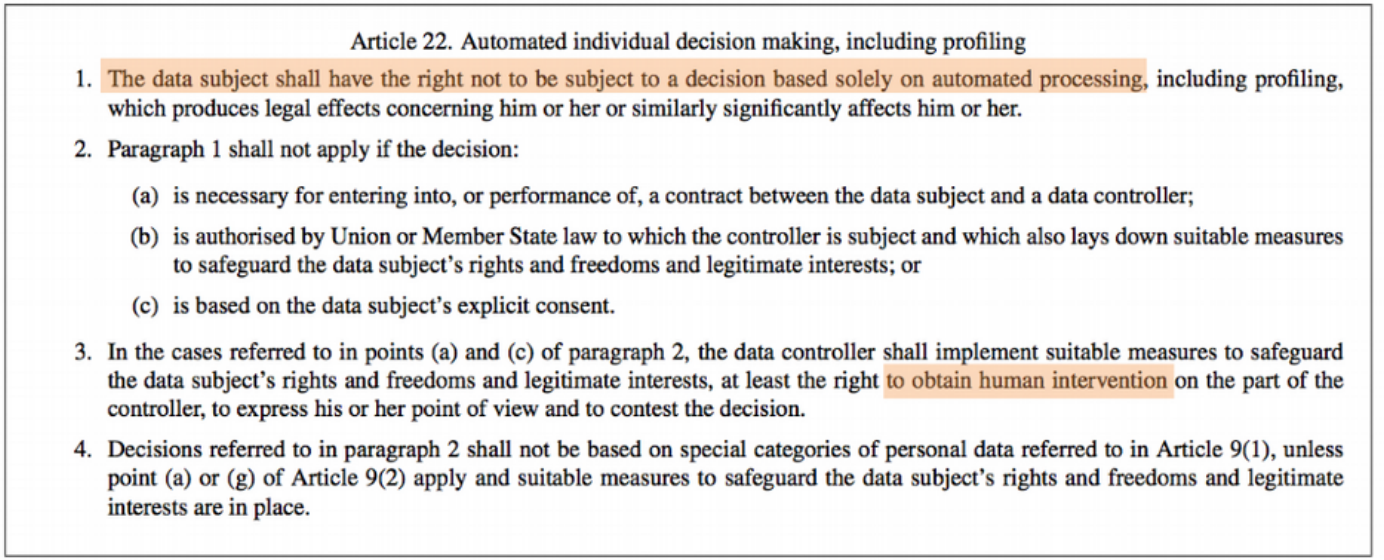
Architecture & tuning hyper-parameters

Rare skill (only few years old!)

But it gets worse...

Deep fakes...

Regulatory: GDPR



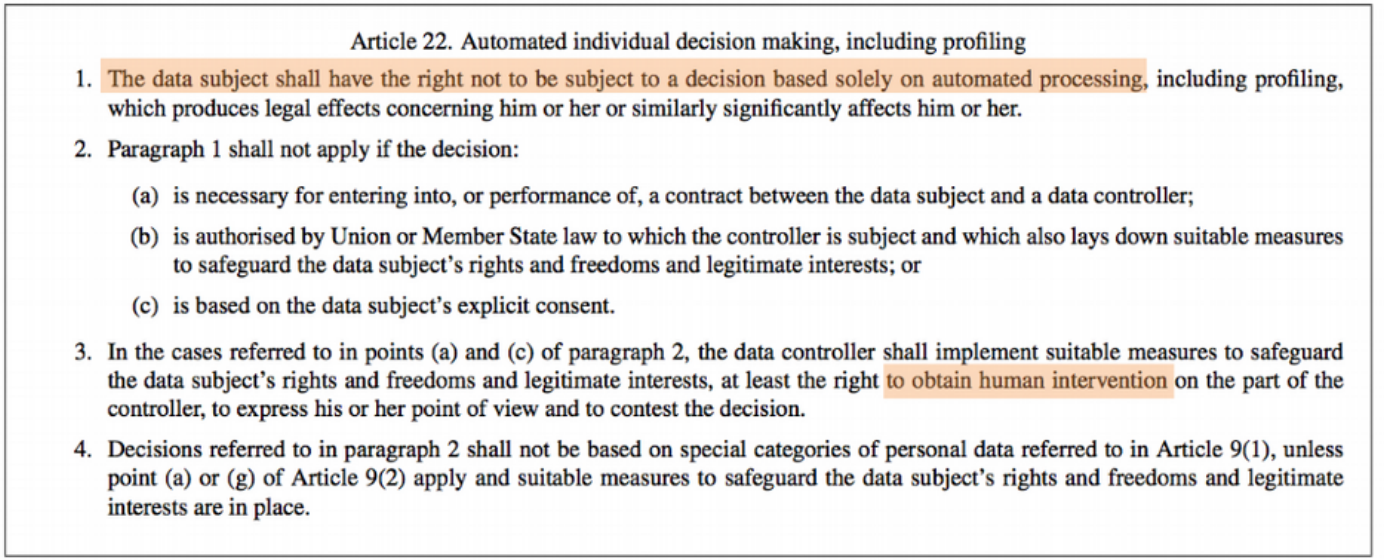
Article 22. Automated individual decision making, including profiling

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

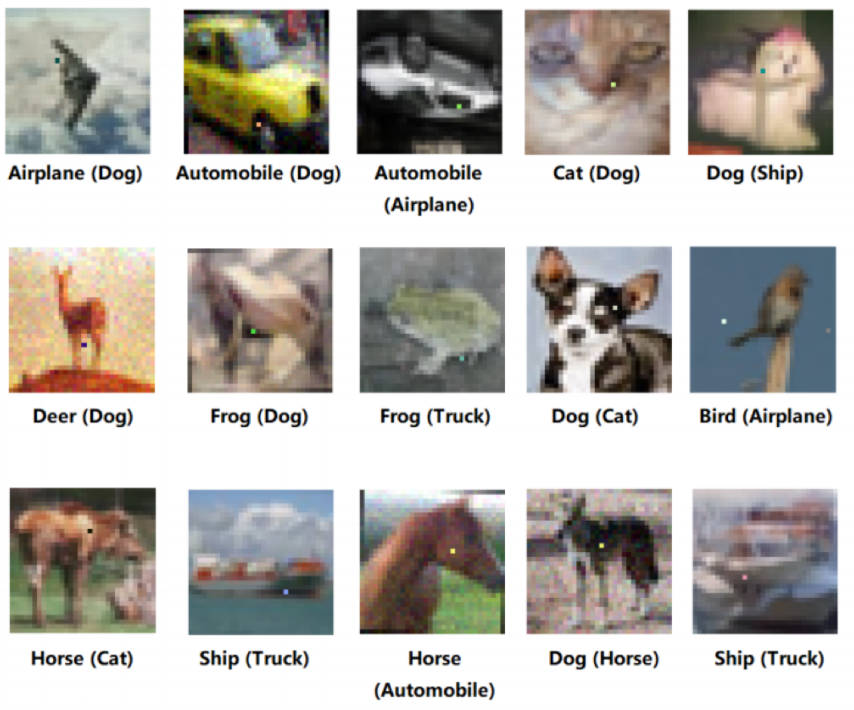


Figure 1. One-pixel attacks created with the proposed algorithm that successfully fooled a target DNN. The original class labels are written below each image with the target class label written inside brackets.

Su, Vargas, & Kouichi (2018)

Adversarial Examples

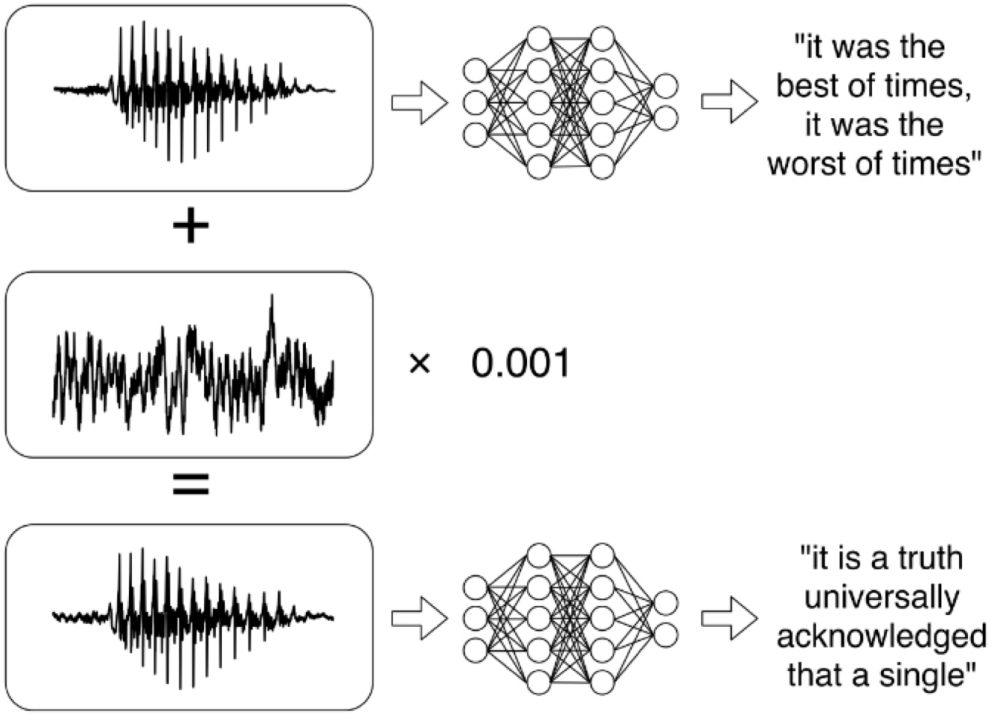
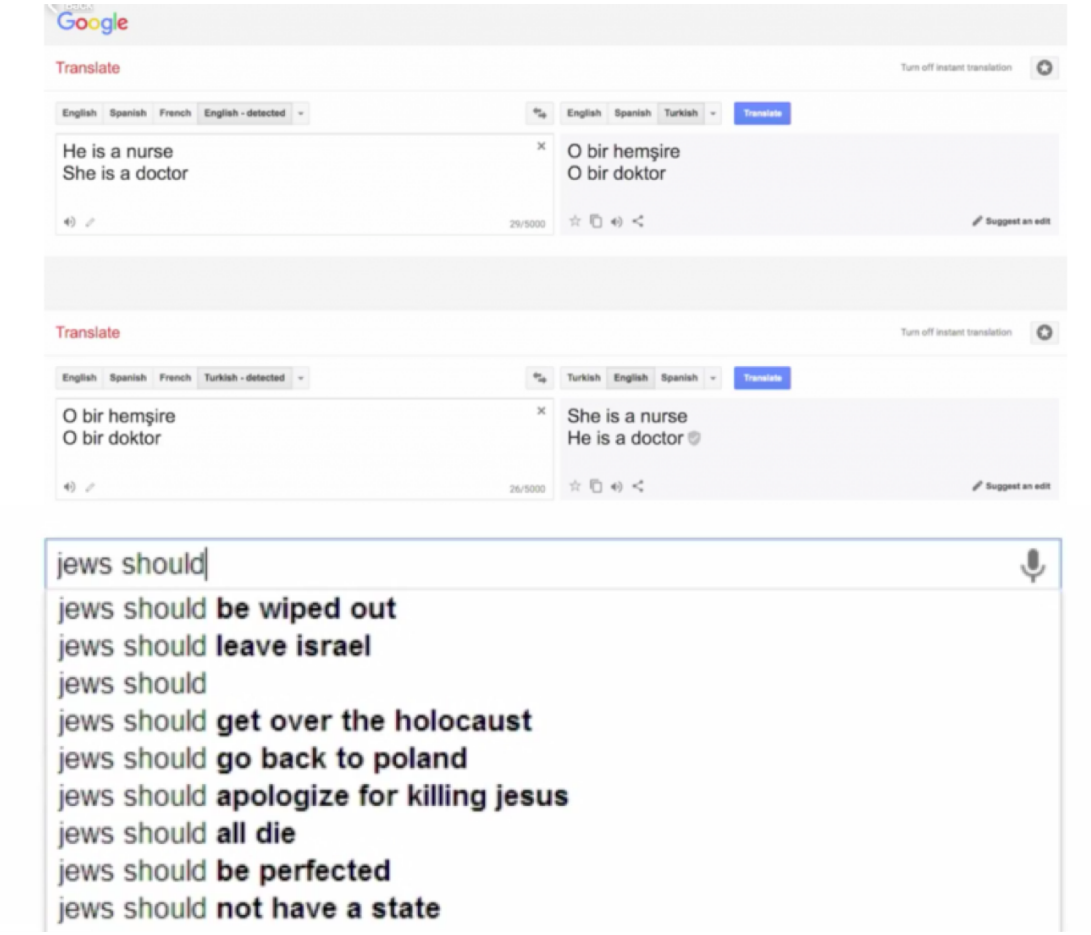


Fig. 1. Illustration of our attack: given any waveform, adding a small perturbation makes the result transcribe as any desired target phrase.

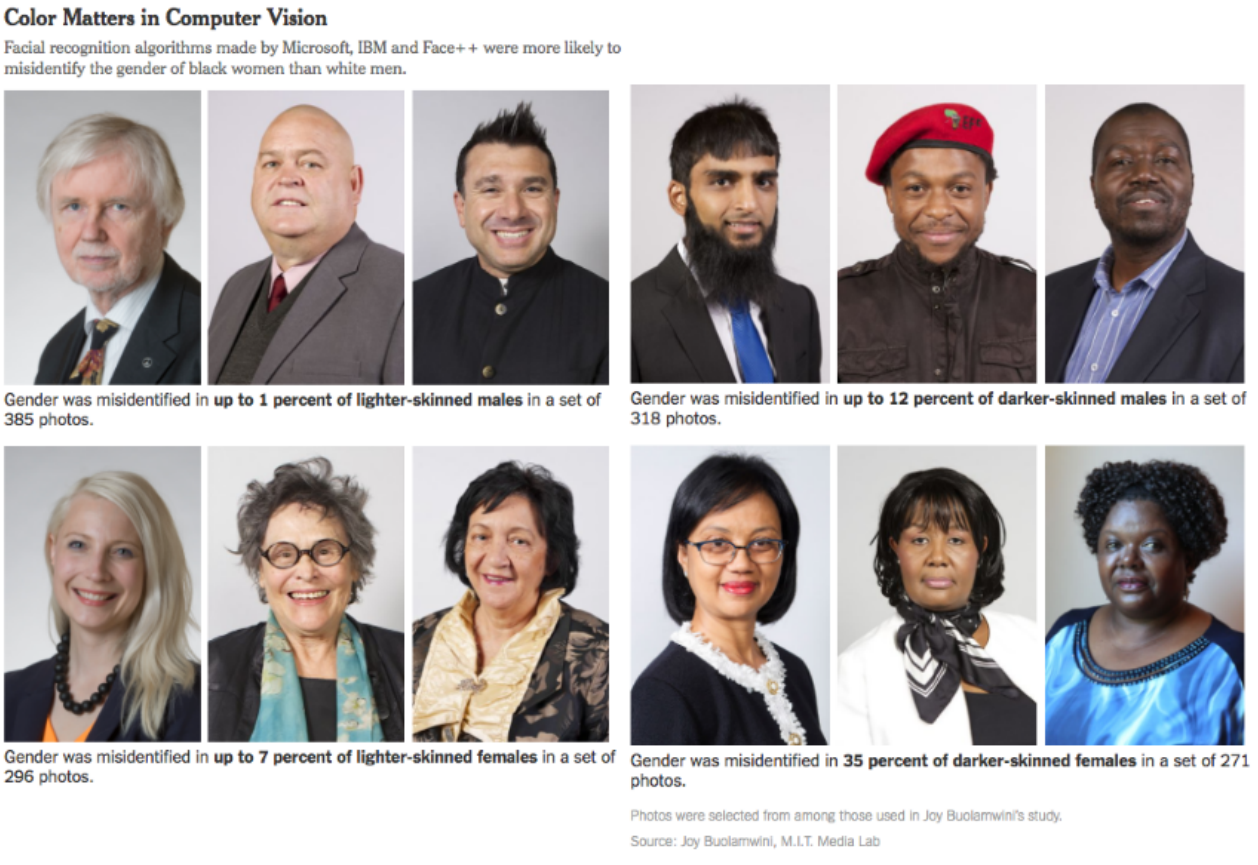
Carlini and Wagner (2018)

Algorithmic Bias

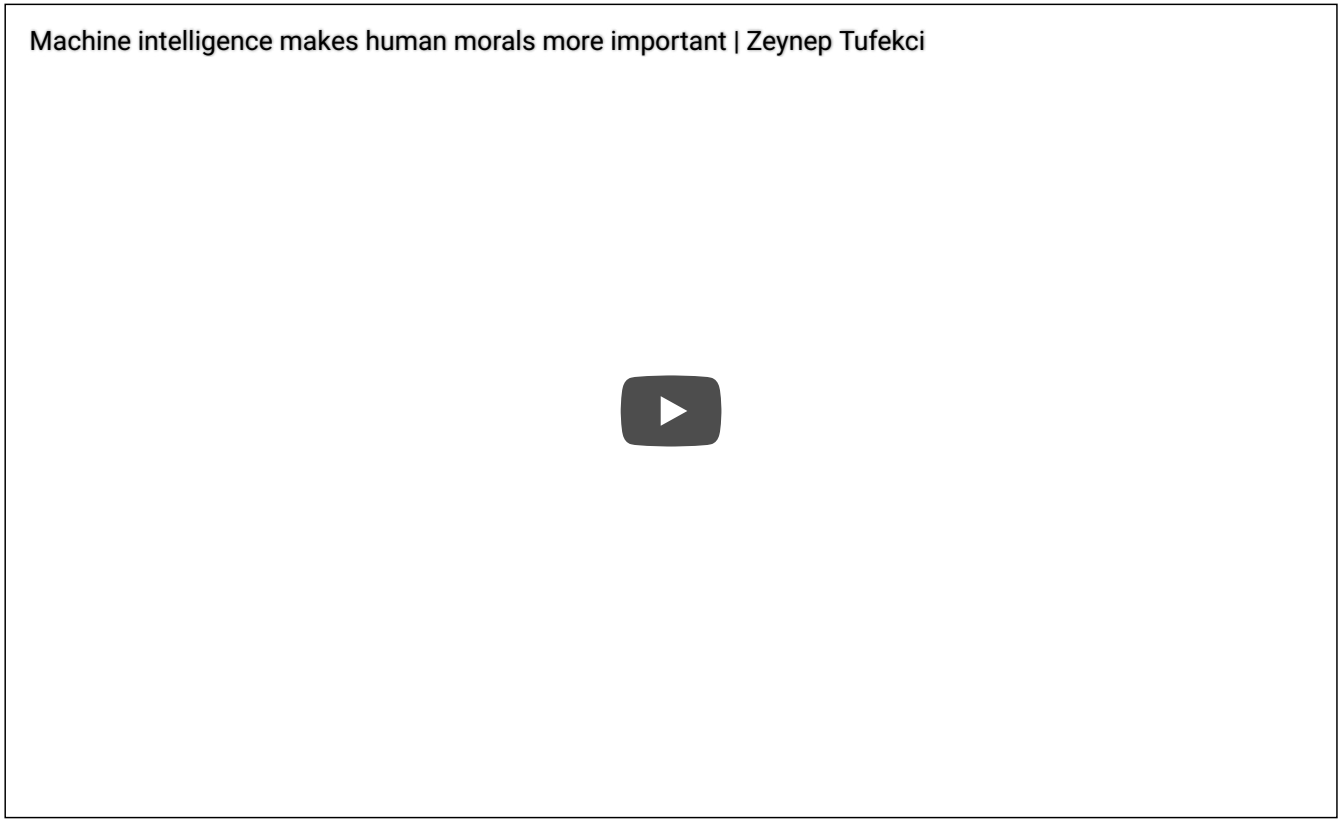


Kate Crawford (2017)

Algorithmic Bias



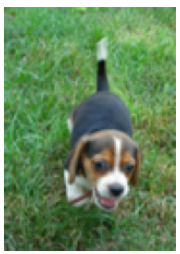
New York Times (2018)

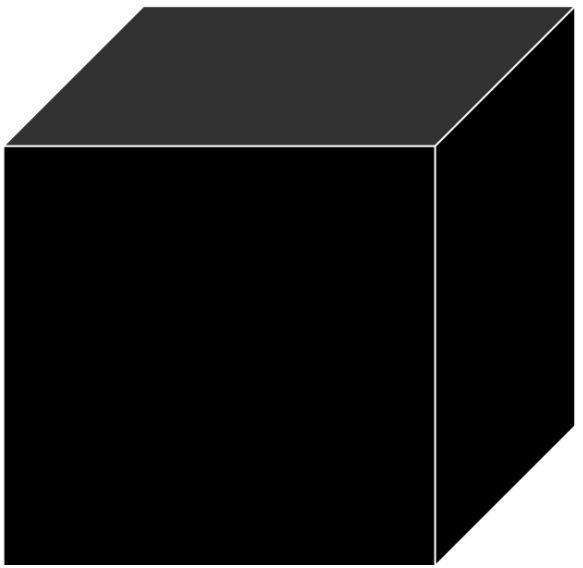


> We need to audit our black boxes -Zeynep Tufekci

But it's a black box...

Original input





Final output

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

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

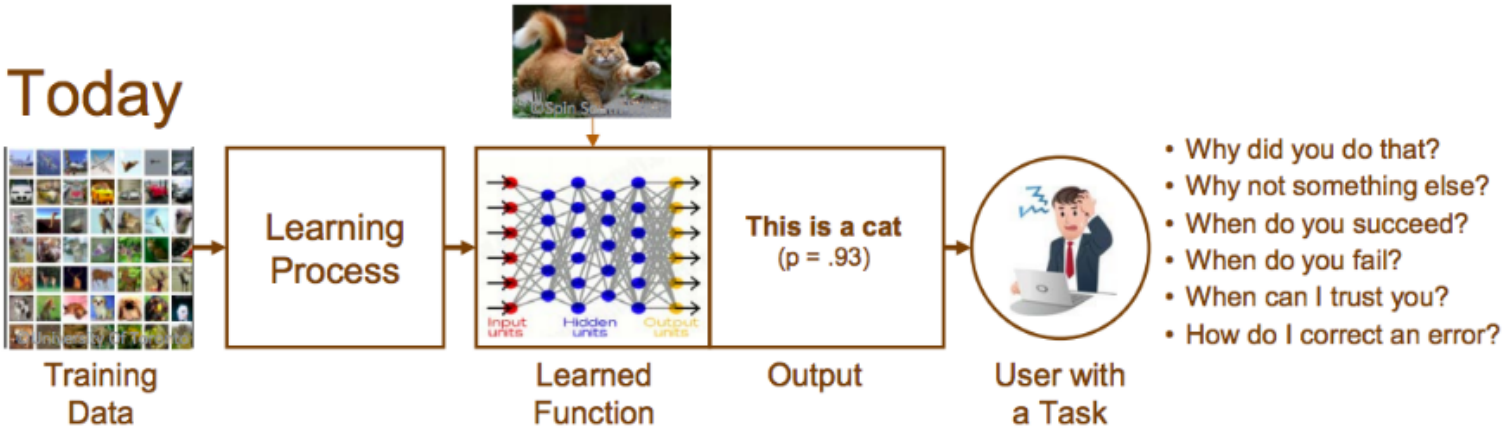
2015 DARPA (Defense Department Research Arm) program with two goals:

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**

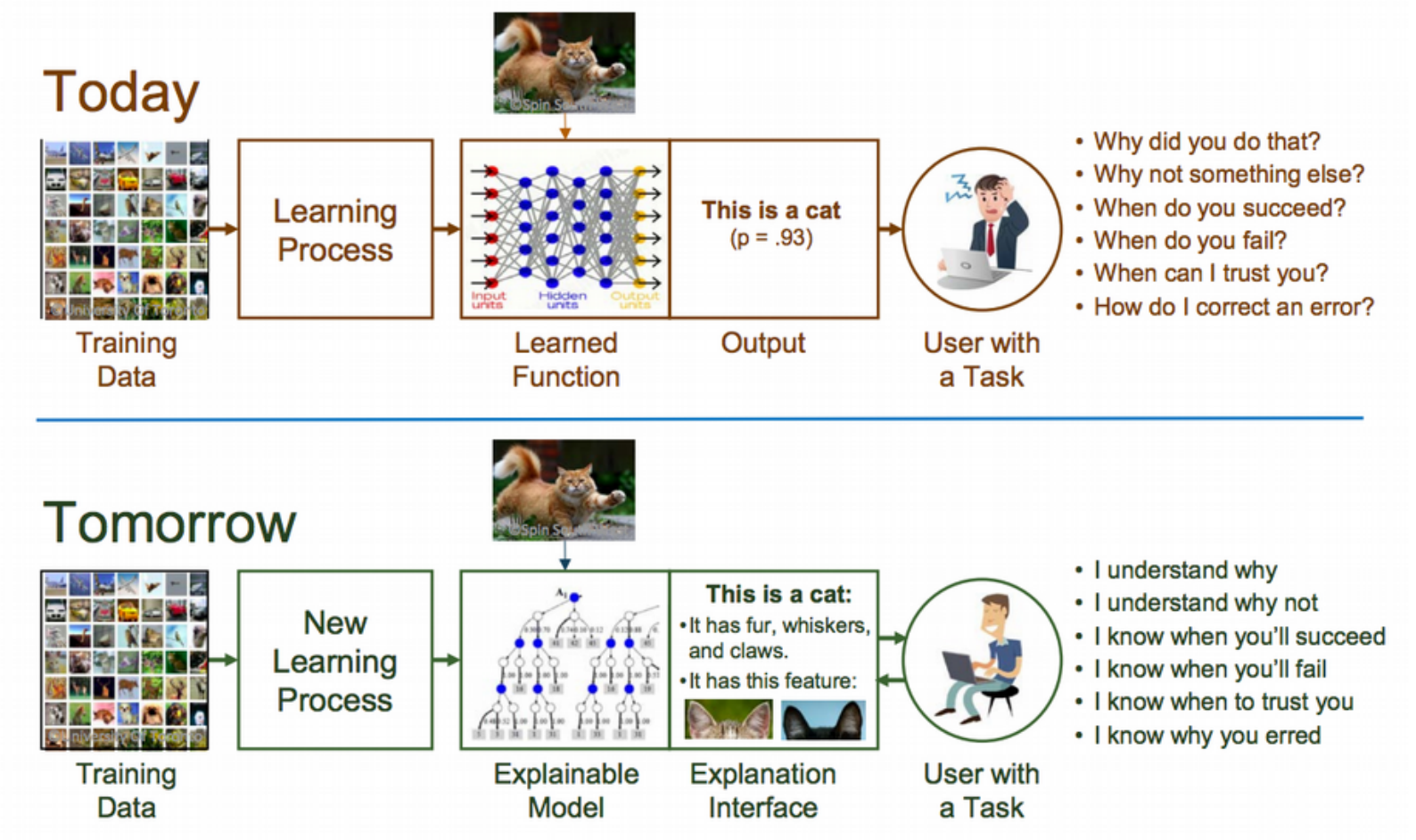
Explainable AI



DARPA XAI

bitly.com/xai-davidson

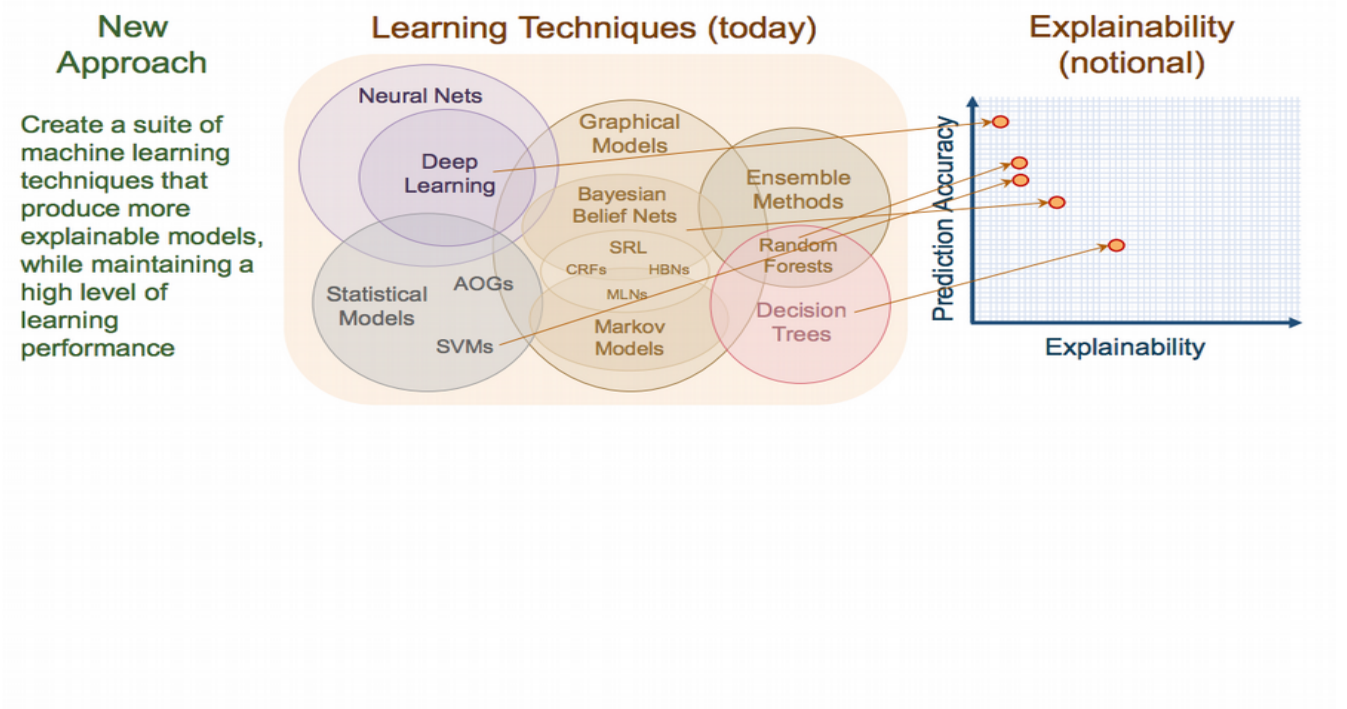
Explainable AI



DARPA XAI

bitly.com/xai-davidson

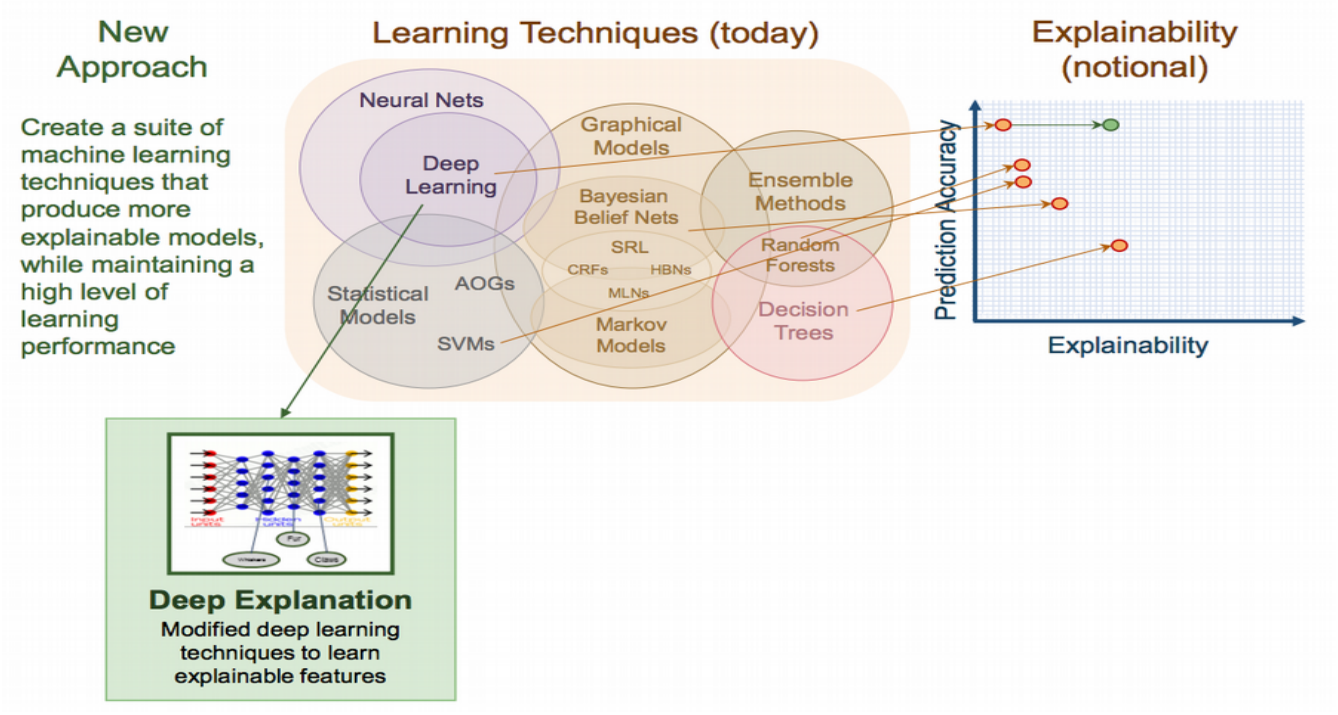
XAI Approaches



DARPA XAI

bitly.com/xai-davidson

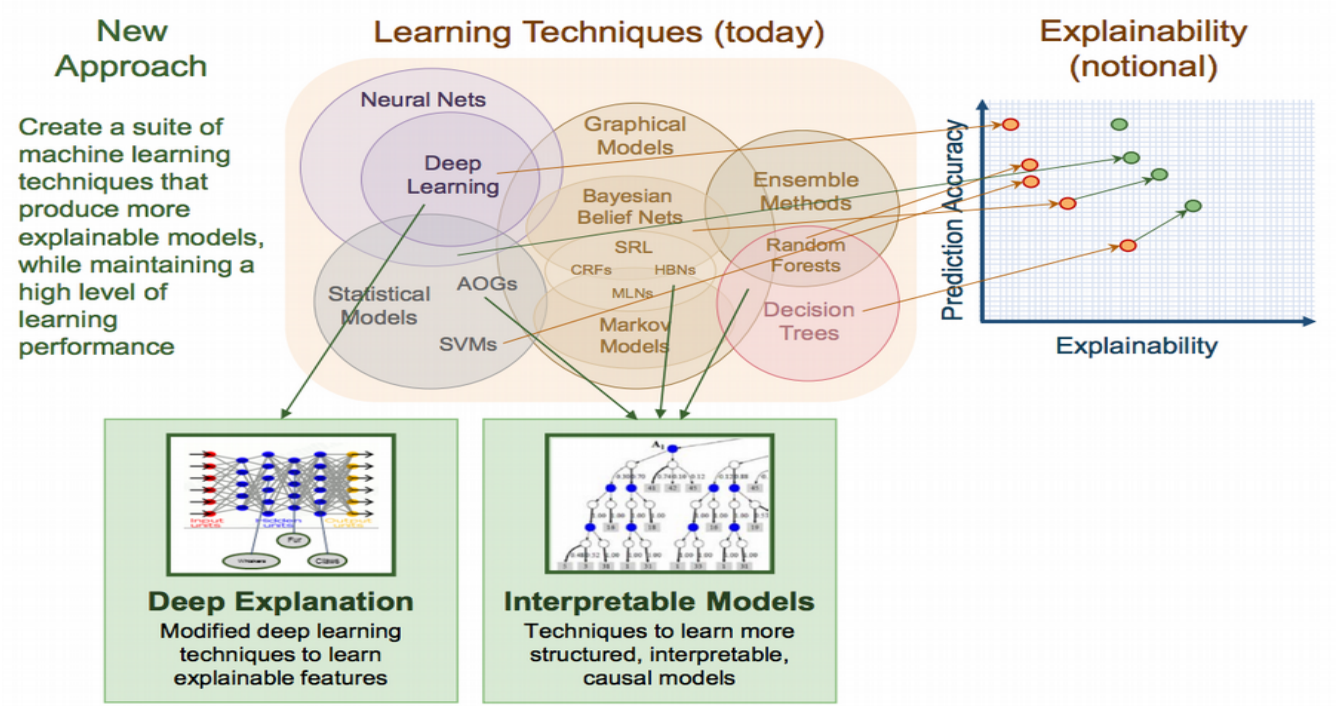
XAI Approaches



DARPA XAI

bitly.com/xai-davidson

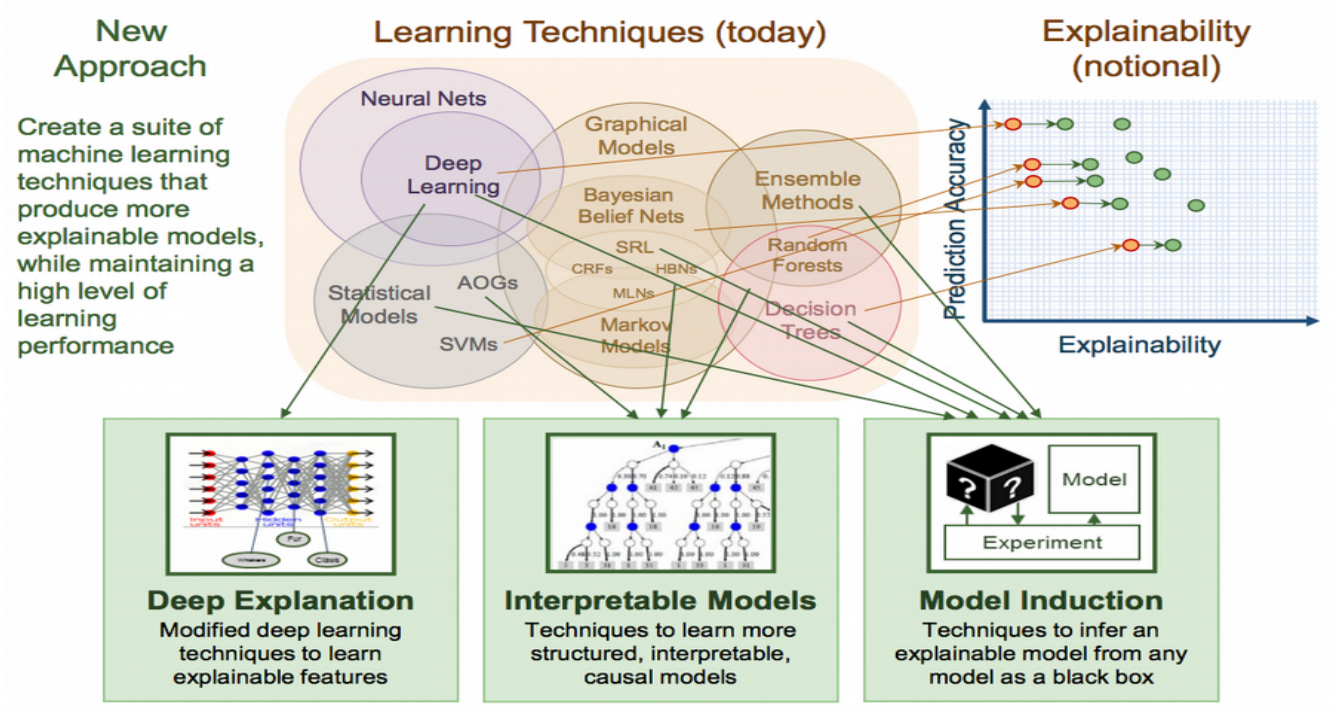
XAI Approaches



DARPA XAI

bitly.com/xai-davidson

XAI Approaches



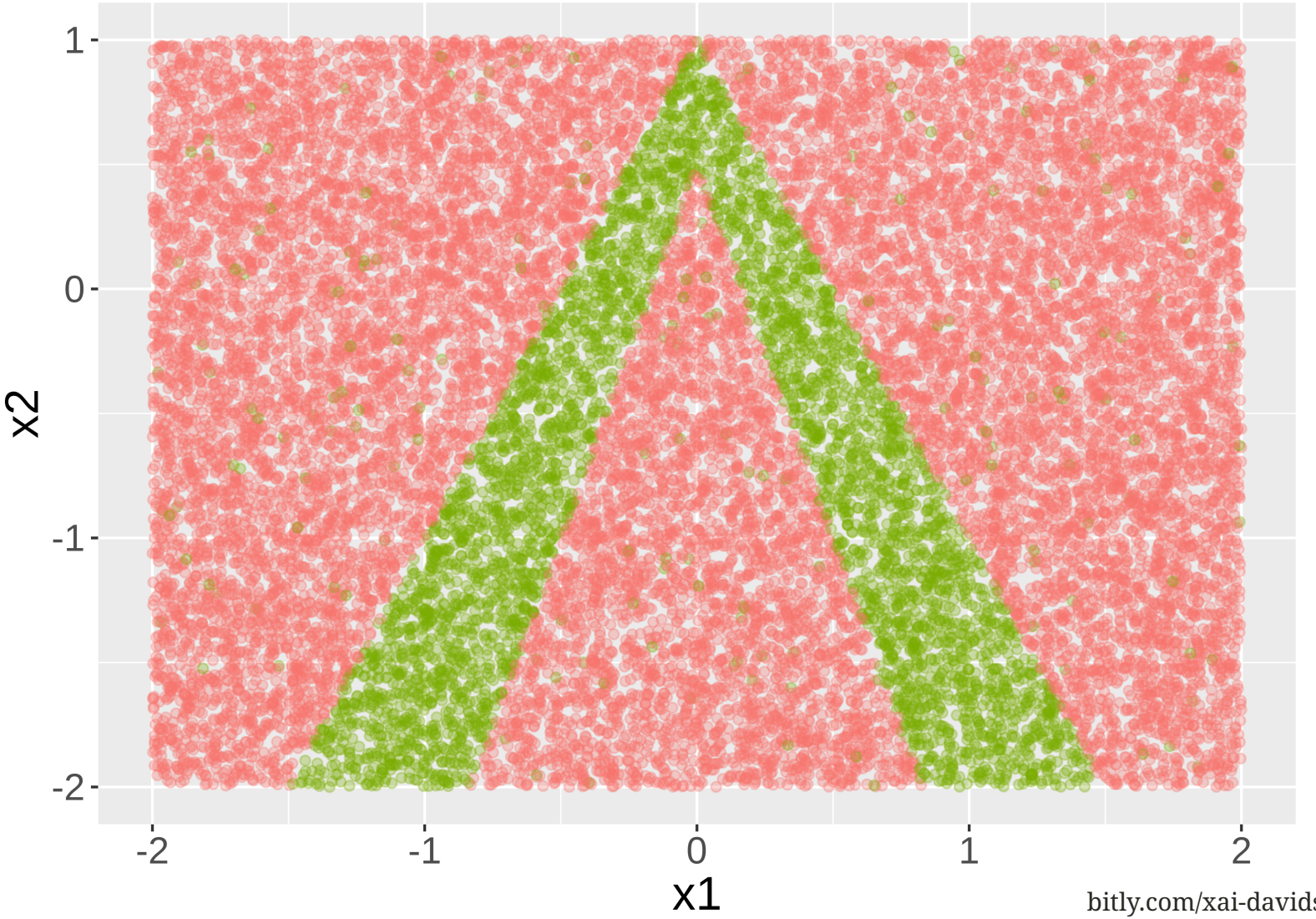
DARPA XAI

bitly.com/xai-davidson

Model Induction (Agnostic): LIME

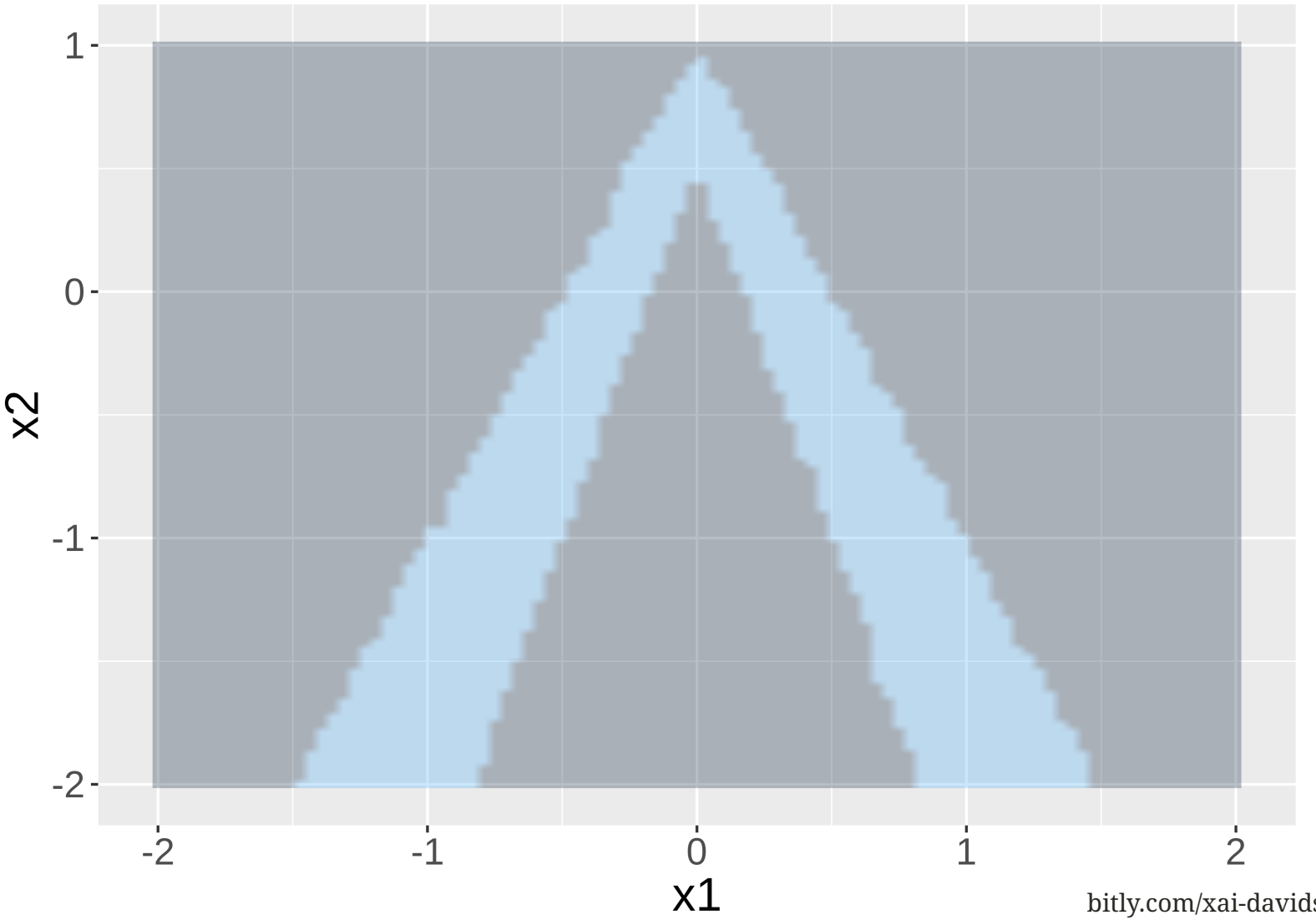
Ribeiro, Singh, Guestrin, 2016

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

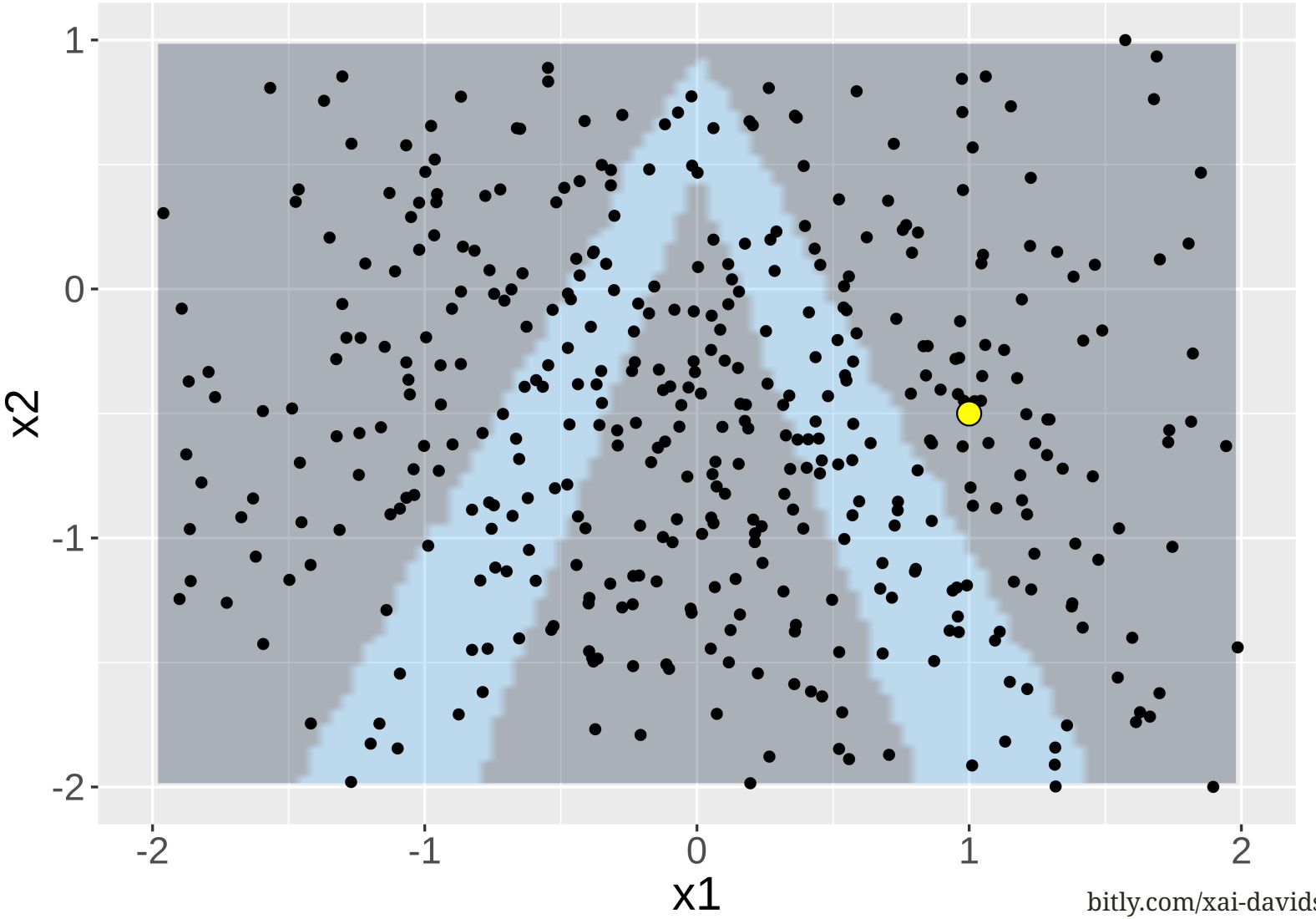


bitly.com/xai-davidson

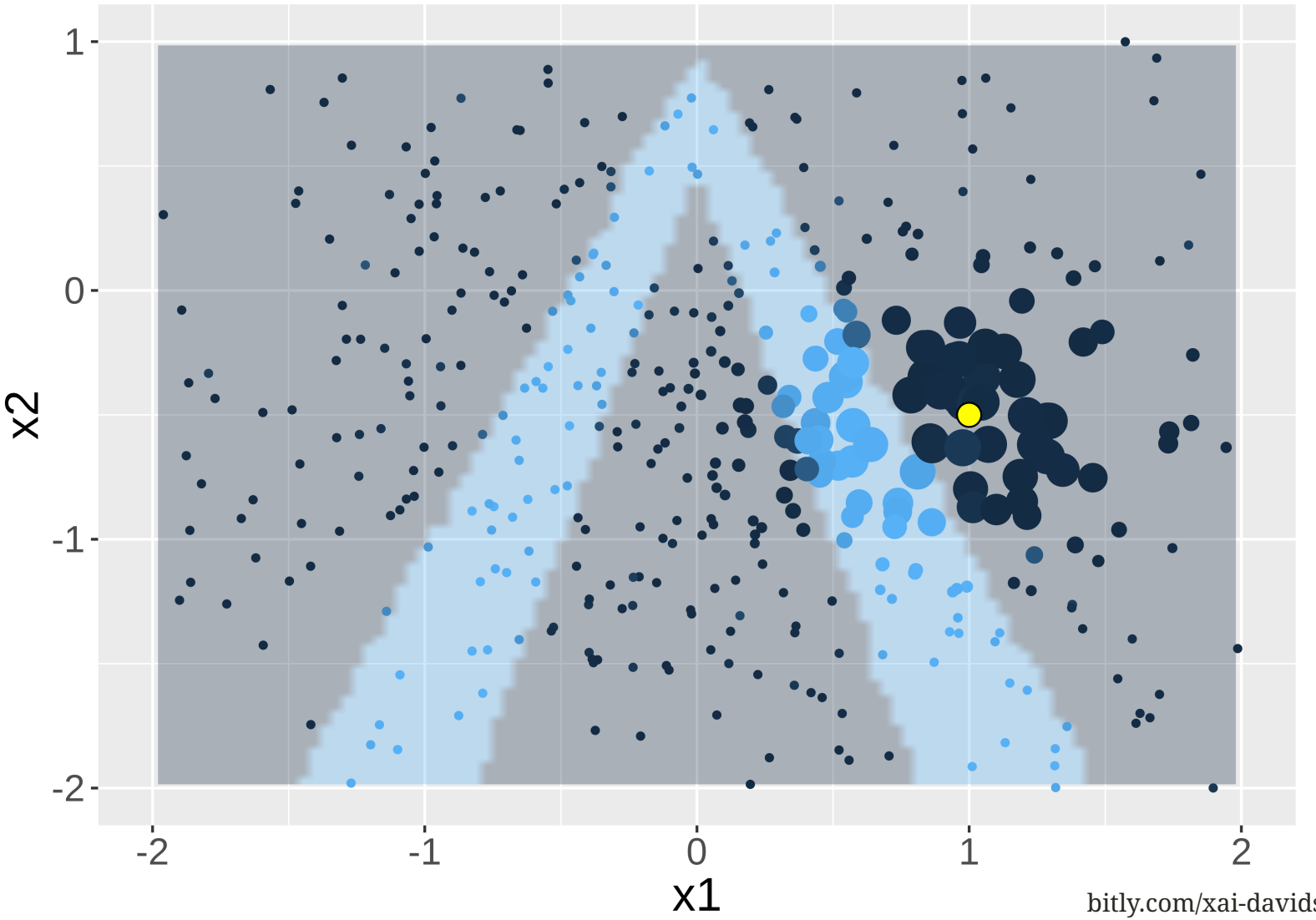
LIME: Intuition / Molnar, 2018 / RStudio.Cloud Project



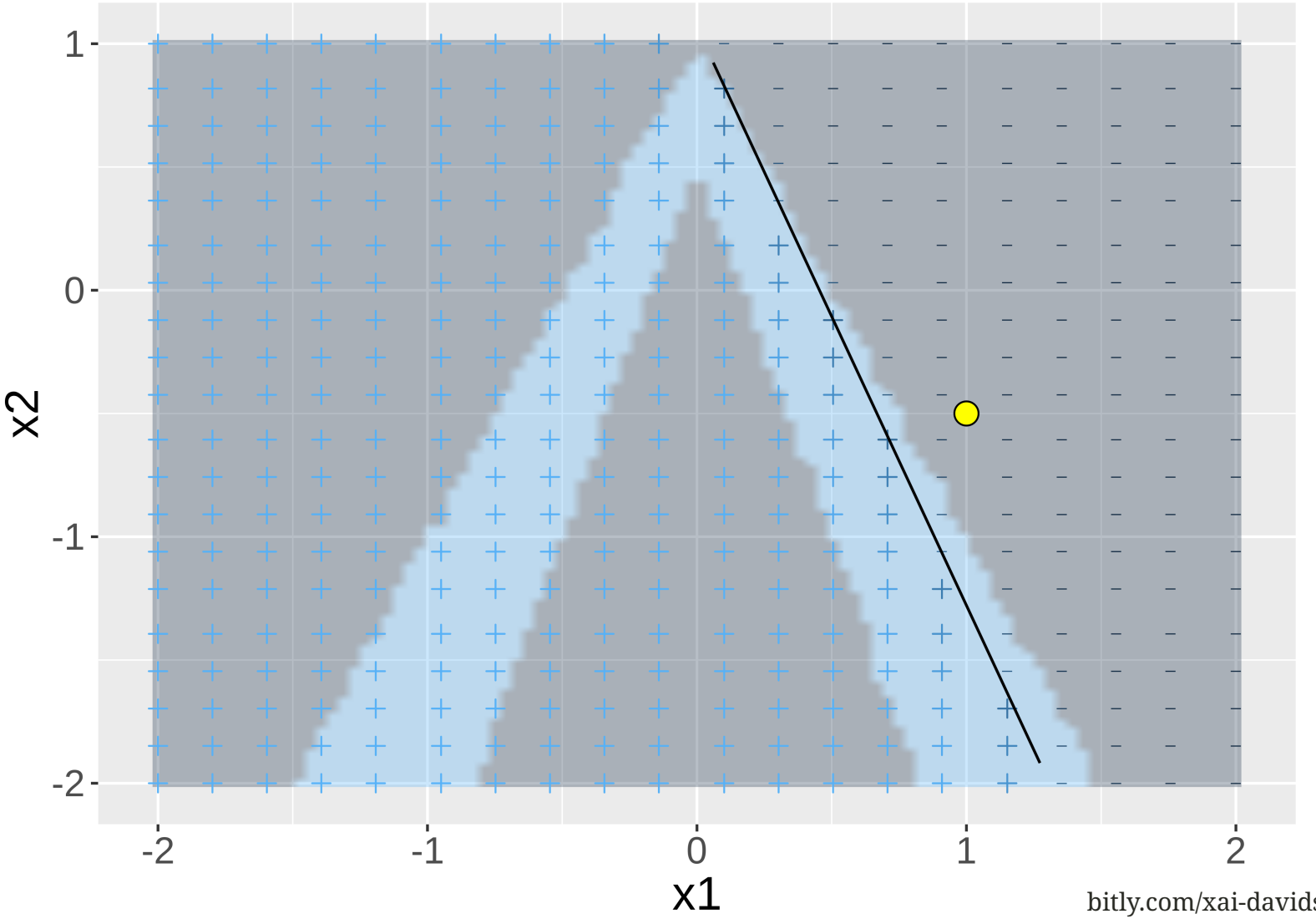
LIME: Intuition / Molnar, 2018 / RStudio.Cloud Project



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

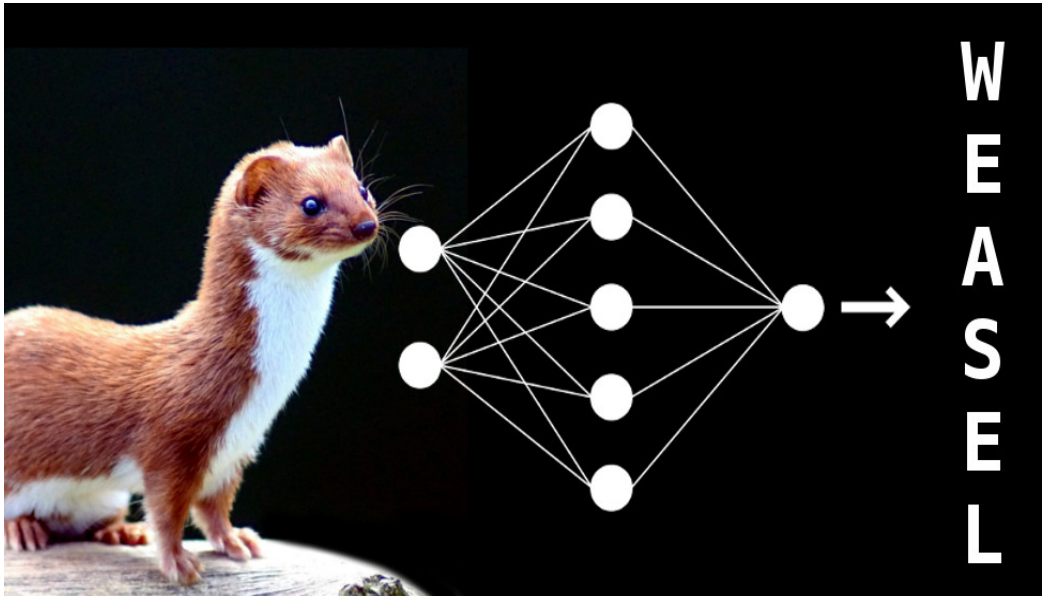


LIME: Intuition / Molnar, 2018 / RStudio.Cloud Project



bitly.com/xai-davidson

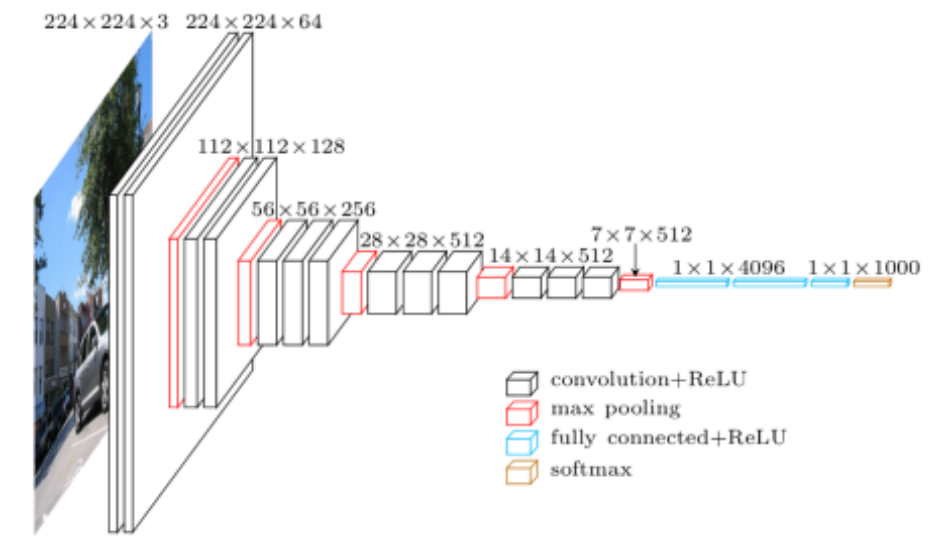
Application: LIME on image classification in R



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
)
```

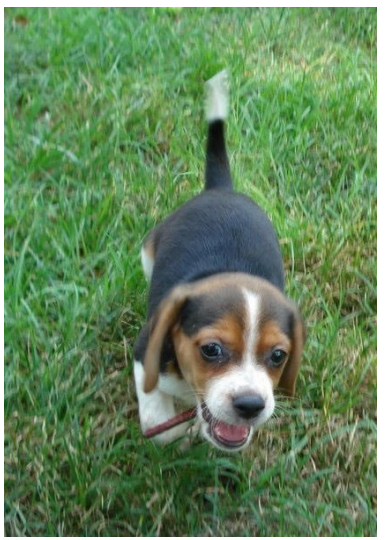


bitly.com/xai-davidson

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
```



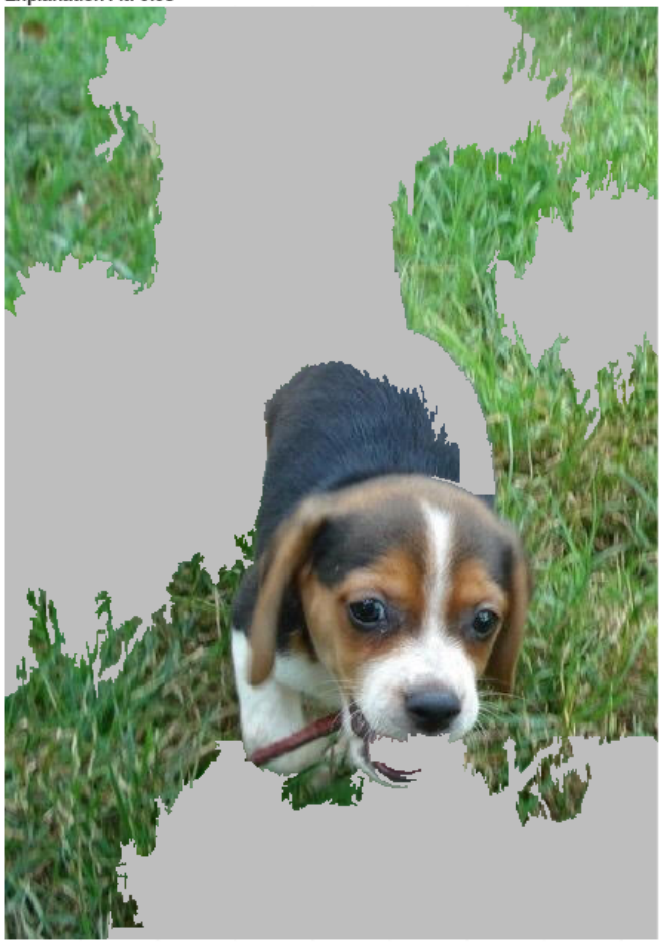
```
plot_superpixels(img_path, n_superpixels = 200, weight = 40)
```

```r
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)
```

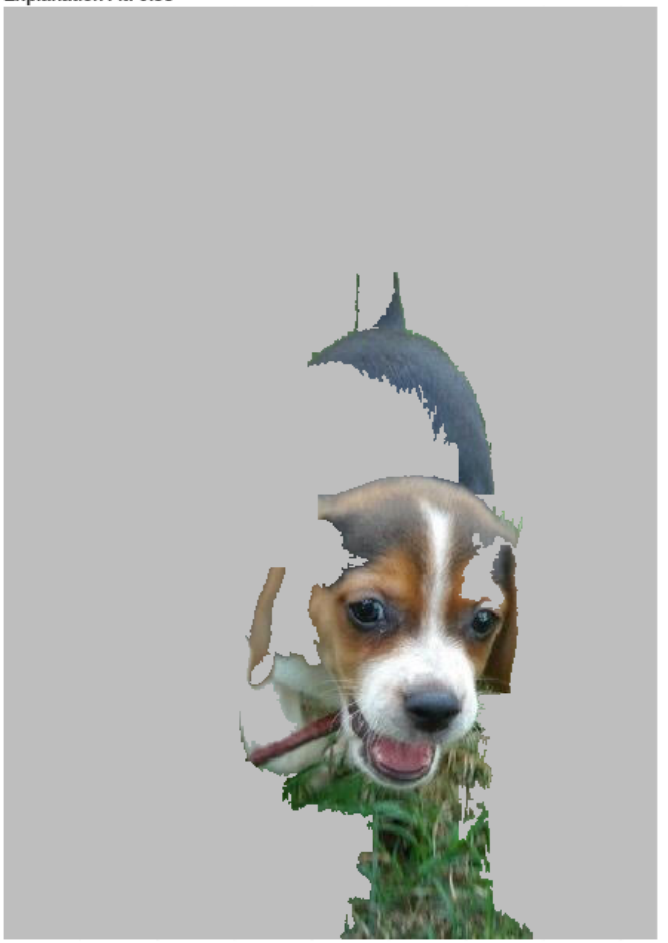
```r
library(lime)

# 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)
```

```r
# 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



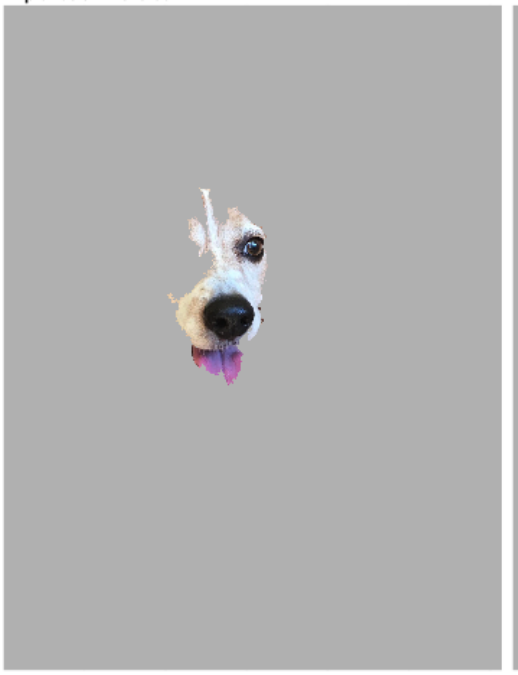
Label: Chihuahua
Probability: 0.16
Explanation Fit: 0.33



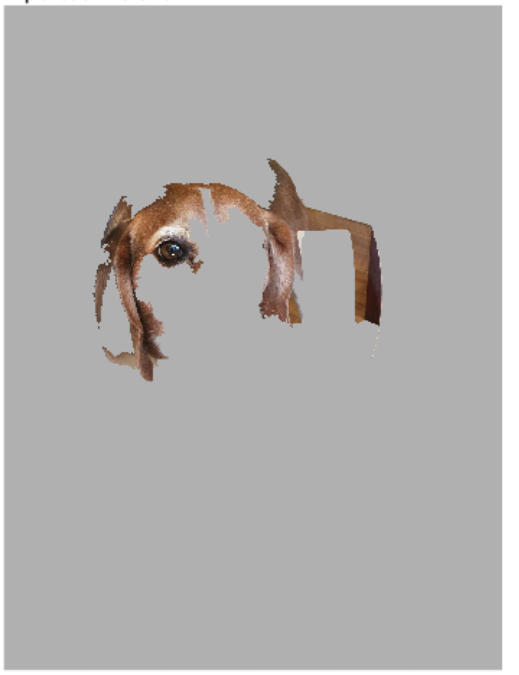
The cone of shame...

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



bitly.com/xai-davidson

Or in disguise...

bitly.com/xai-davidson

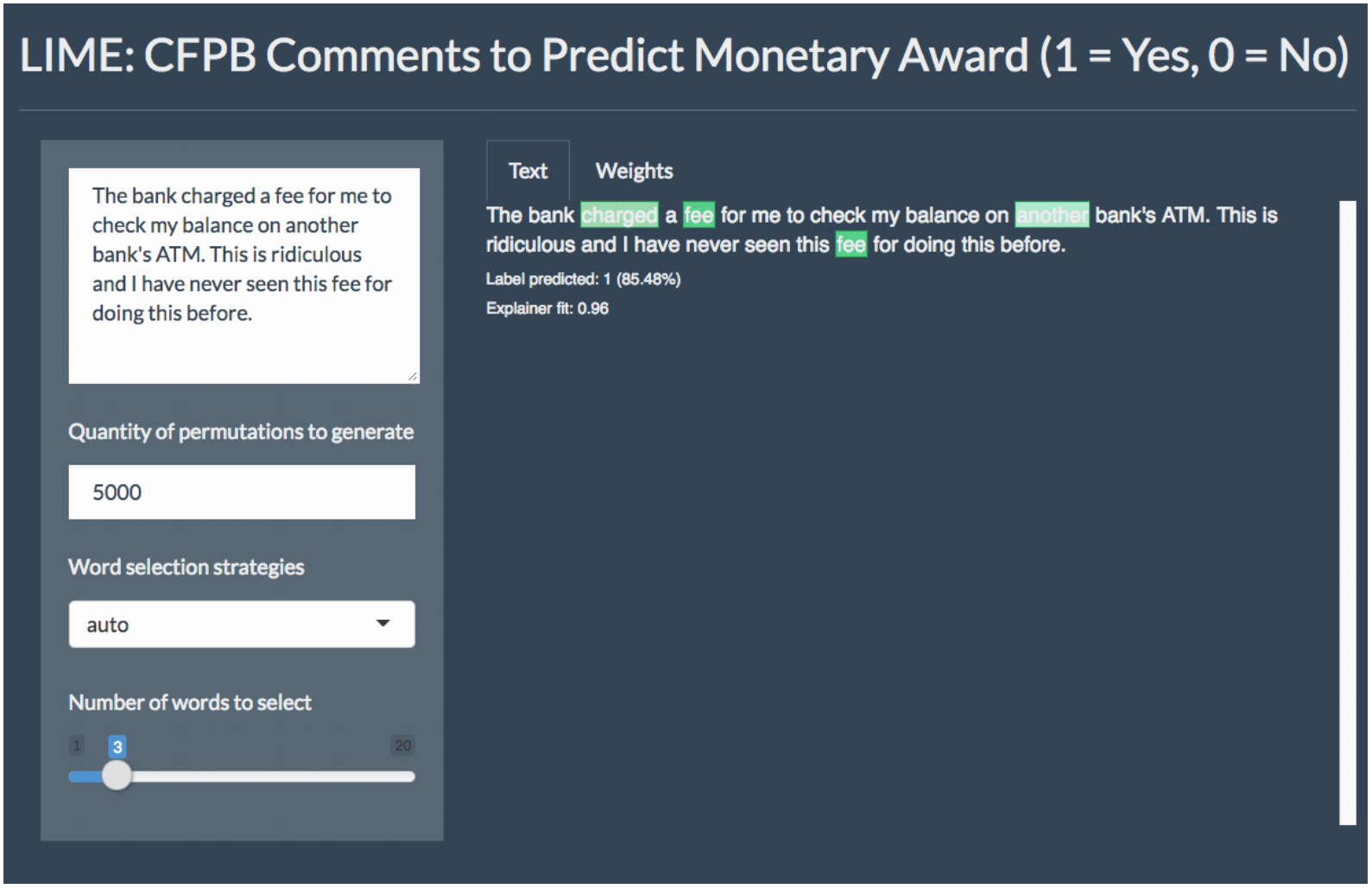
```
plot_image_explanation(explanation, threshold = 0.01)
```

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


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

@smironchuk's Pinterest Ewok Costumes

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



lime Package

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 "Buiding 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

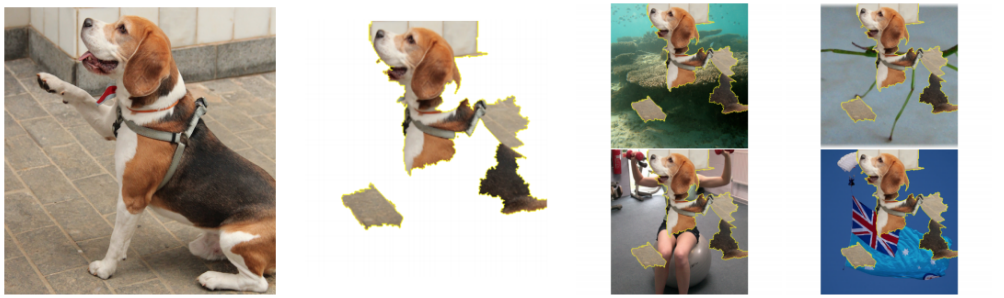
Questions & Discussion: Thanks!

github.com/wesslen

wesslen.github.io

bitly.com/xai-davidson

Anchor-based Explanations



(a) Original image (b) Anchor for "beagle" (c) Images where Inception predicts $P(\text{beagle}) > 90\%$

What animal is featured in this picture ?	**dog**
What floor is featured in this picture?	dog
What toenail is paired in this flowchart ?	dog
What animal is shown on this depiction ?	dog

(d) **VQA:** Anchor (bold) and samples from $\mathcal{D}(z|A)$

Where is the **dog**?	on the floor
What color is the **wall?**	white
When was this picture taken?	during the day
Why is he lifting his paw?	to play

(e) **VQA:** More example anchors (in bold)

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

Ribeiro, Singh, and Guestrin, 2018