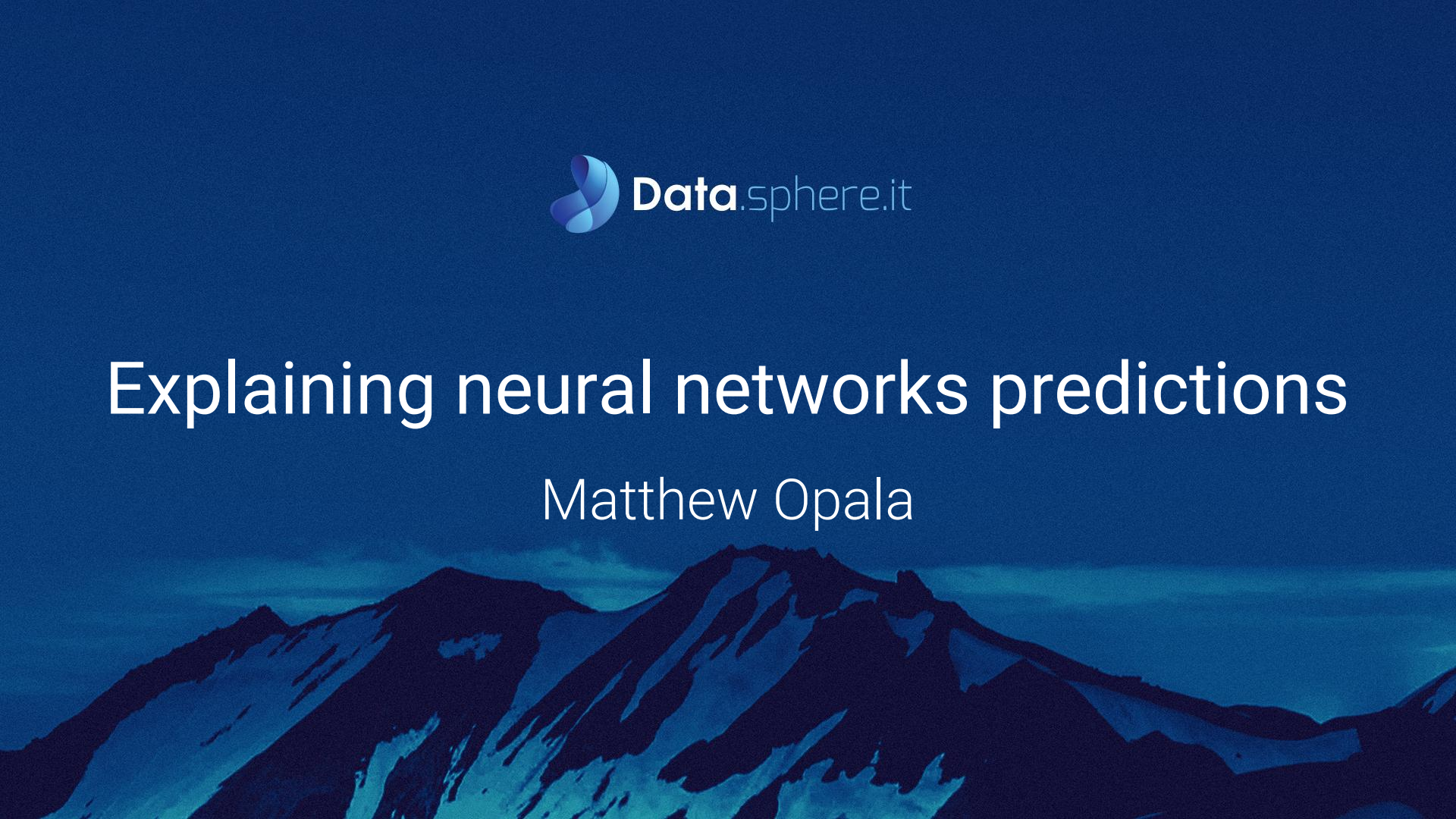

Data.sphere.it
Explaining neural networks predictions
Matthew Opala

https://github.com/plazowicz/datasphere2018

Pre Agenda question

You can choose between AI doctor who is 80% accurate and can explain its diagnoses or AI doctor who is 90% accurate but can’t explain, which one do you pick:

- 80% accurate with explanation
- 90% accurate without

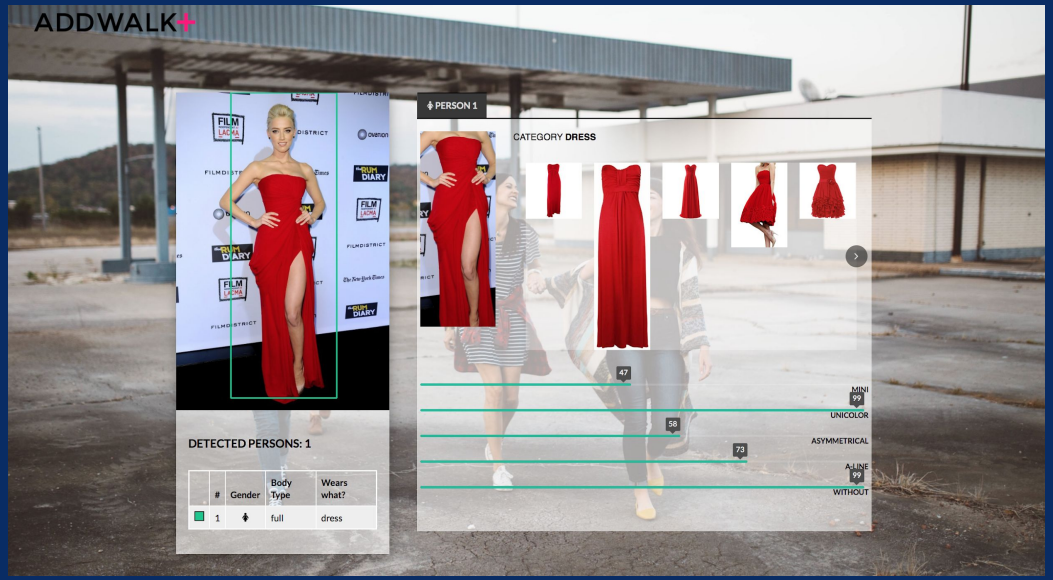
Agenda

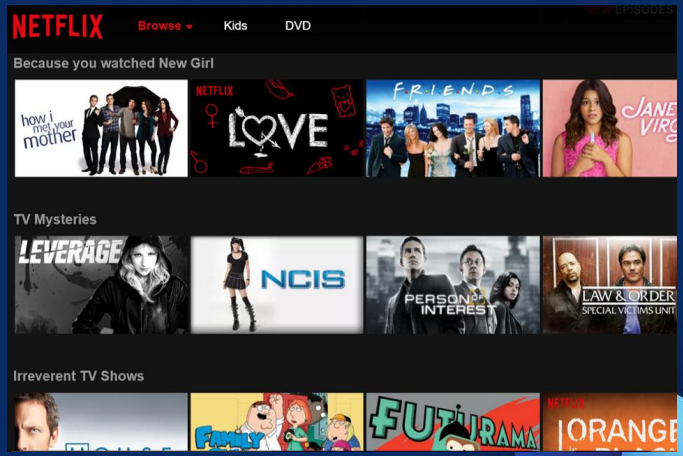
- Why do we need explanations?
- Trust in model
- LIME

“Software is eating the World, but AI is going to eat software”

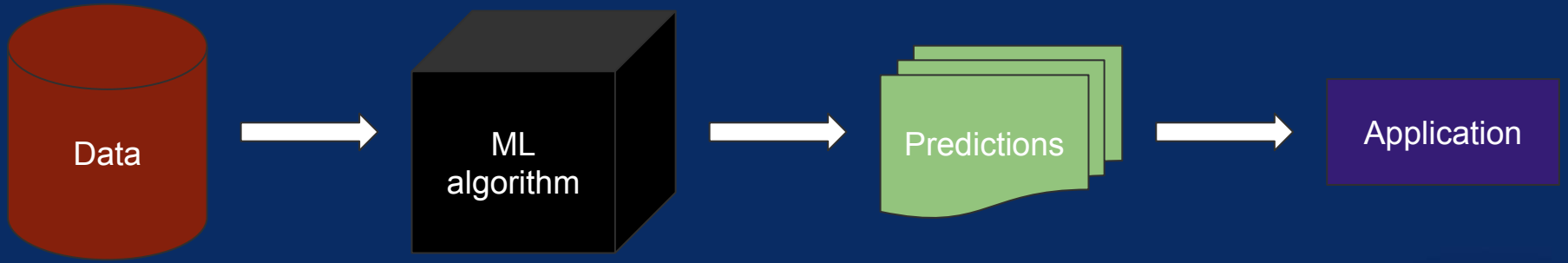
Jensen Huang, NVIDIA CEO

Machine learning everywhere

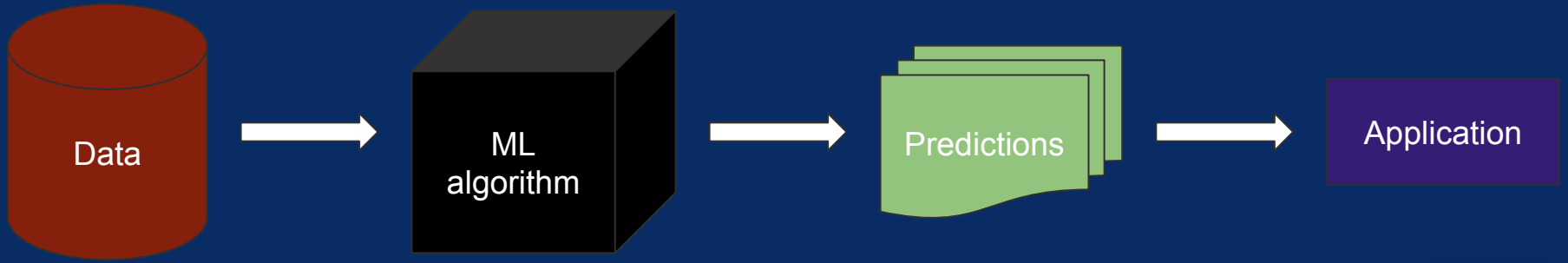




How to train machine learning algorithm?



How to train machine learning algorithm?



Learn Model

Deploy Model

Example: predict if there are football shoes on image

We have already trained shoes detector

Evaluation

- Split into train set, validation set and test set
- Define how we measure performance
- We can use accuracy for classification task
- Let’s use state-of-the art ResNet-50
- Train on train set, tune hyperparameters on validation set and evaluate on test set

95%

95%

Do you trust this classifier?

Eyeball

Football shoes

Not football shoes

Not football shoes

Football shoes

Eyeball

Football shoes

Not football shoes

Football shoes

Not football shoes

Football shoes

Football shoes

Eyeball

Football shoes

Not football shoes

Football shoes

Not football shoes

Football shoes

Football shoes

Eyeball

Football shoes

Not football shoes

Football shoes

Football shoes

Not football shoes

Football shoes

Football shoes

Football shoes


wat

Digression: classic Kaggle pitfall

- Real test set is different that your test-validation set
- Hint: use adversarial validation
  - Learn more here: http://fastml.com/adversarial-validation-part-one/
- If distribution of your train set and test are similar then classifier trained to distinguish training example from test example should achieve about 50% accuracy
- Hint: choose those examples to validation set, that are most certainly classified as test examples in cross-validation mode


Learn model
Deploy model


Learn model
Trust model
Deploy model

How to gain trust?

Interpretable models → For example Decision Trees, but less accurate than Deep Learning models

How to gain trust?

Interpretable models	For example Decision Trees, but less accurate than Deep Learning models
Accuracy	Must have, but could could be unreliable, training data vs real world data

How to gain trust?

Interpretable models	For example Decision Trees, but less accurate than Deep Learning models
Accuracy	Must have, but could could be unreliable, training data vs real world data
A/B Testing	Could be expensive, one may expose bad model to users

How to gain trust?

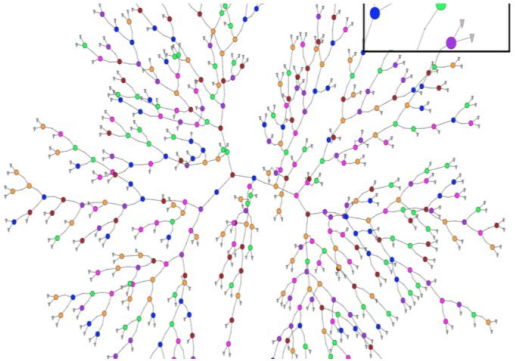
Interpretable models	For example Decision Trees, but less accurate than Deep Learning models
Accuracy	Must have, but could could be unreliable, training data vs real world data
A/B Testing	Could be expensive, one may expose bad model to users
Voodoo	A lot of experience, do similar task as was done before

Explaining individual predictions

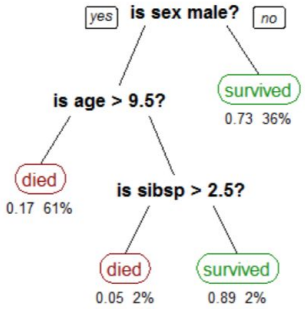
Three must-haves for good explanation

Interpretability

Humans can easily interpret reasoning



Definitely
not interpretable



Potentially
interpretable

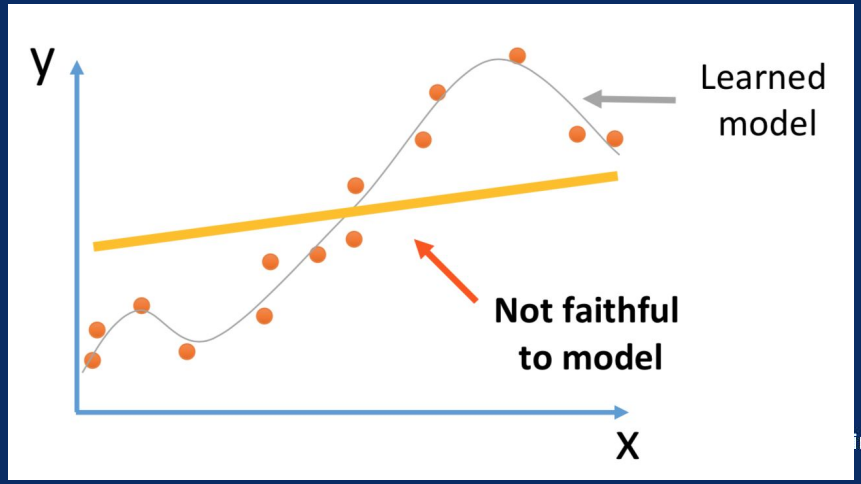
Three must-haves for good explanation

Interpretability

Humans can easily interpret reasoning

Faithful model

Describes how model actually works

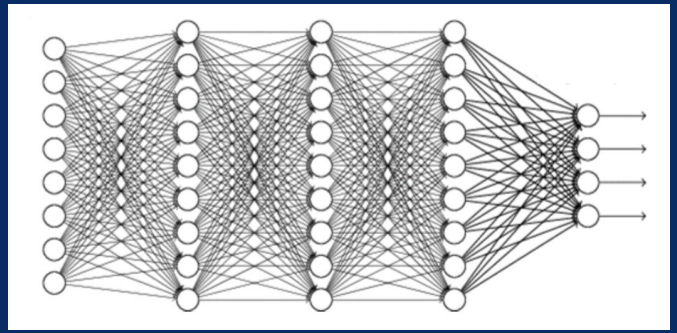


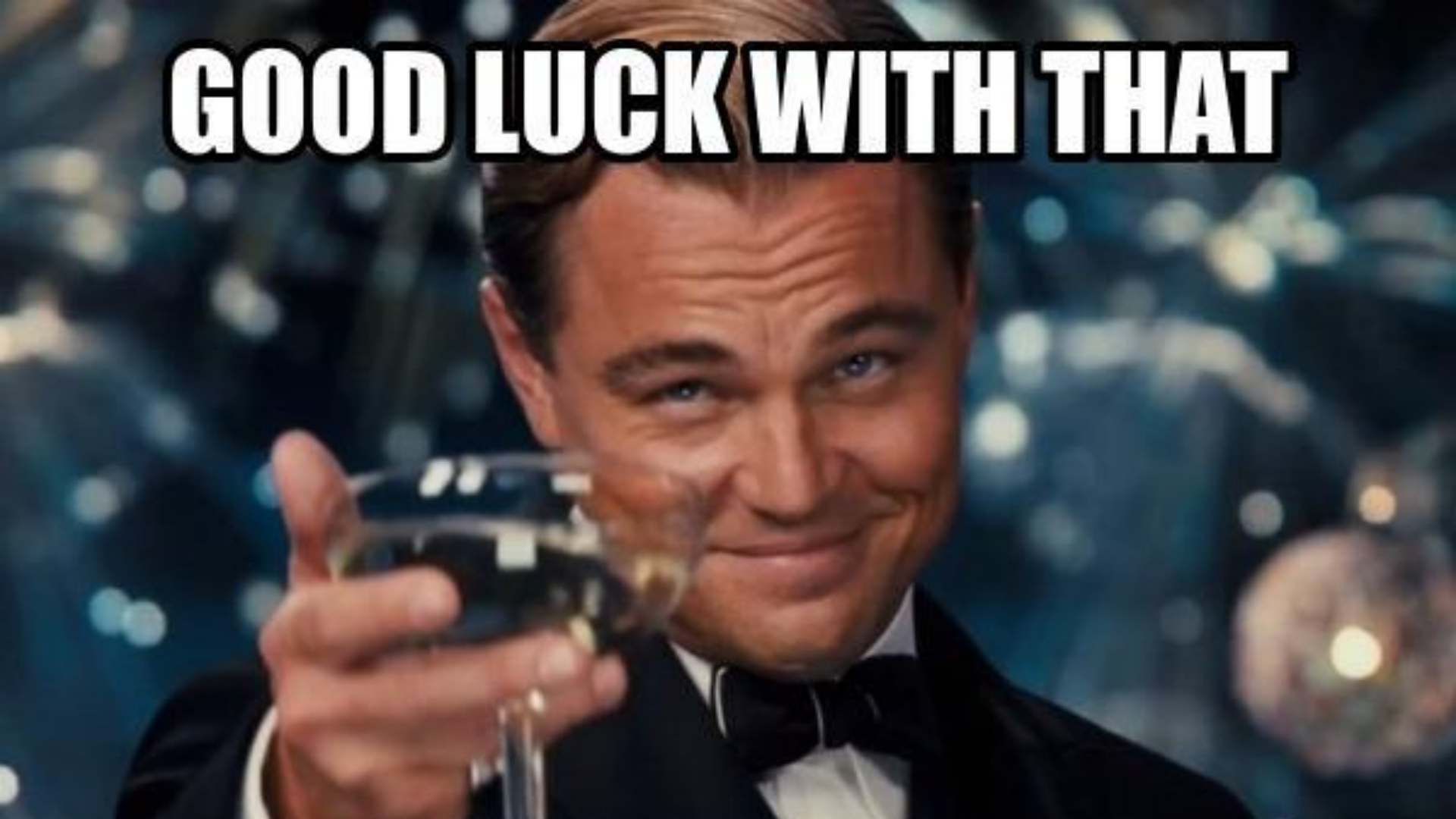
Three must-haves for good explanation

Interpretability
Humans can easily interpret reasoning

Faithful model
Describes how model actually works

Model agnostic
Can explain any classifier

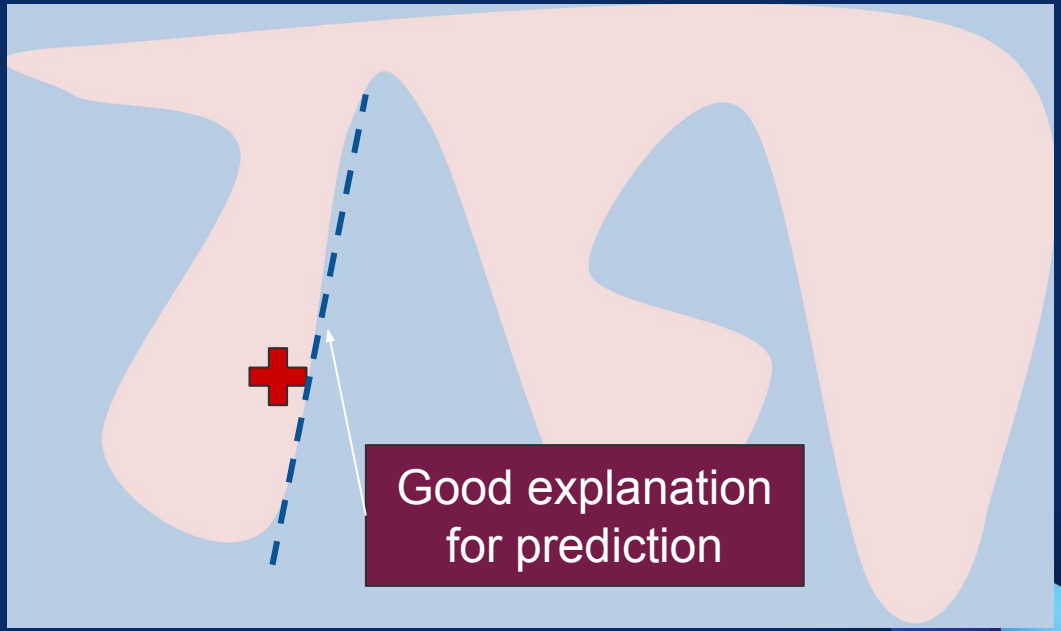



GOOD LUCK WITH THAT

LIME: Local Interpretable Model-Agnostic Explanations

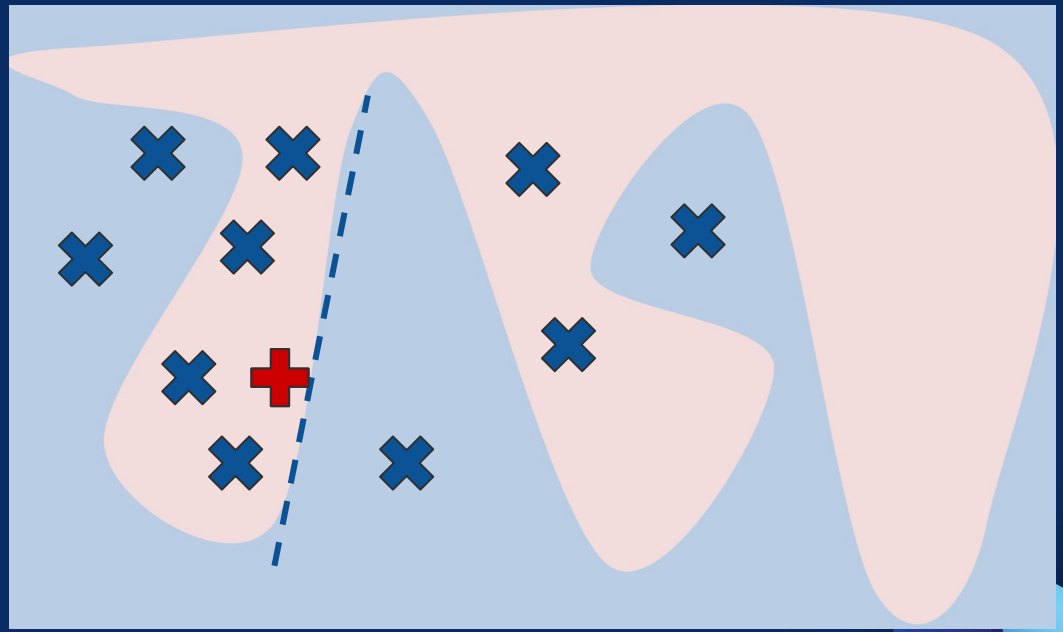
LIME - key ideas

- Pick a model class interpretable by humans
  - Linear regression
  - Shallow decision tree
  - Not globally faithful
- Local approximate black box model
  - Simple model is globally bad, but locally good



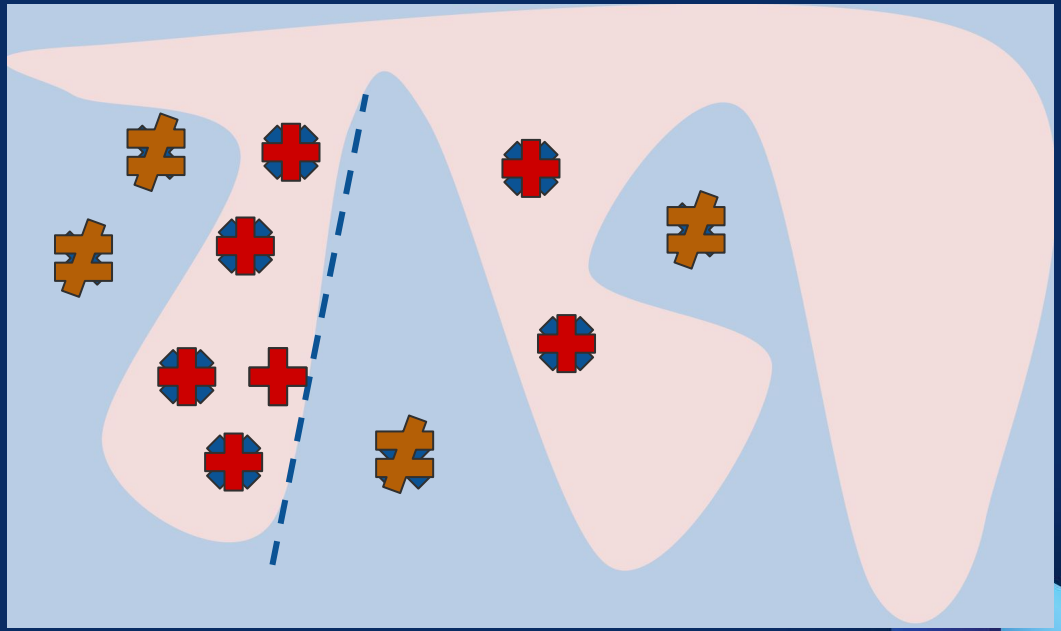
Use LIME to explain a complex model's predictions

- Sample points around example
- Use complex model to predict labels for each sample
- Weigh samples according to distance to example
- Learn simple model on weighted samples
- Use simple model to explain



Use LIME to explain a complex model's predictions

- Sample points around example
- Use complex model to predict labels for each sample
- Weigh samples according to distance to example
- Learn simple model on weighted samples
- Use simple model to explain

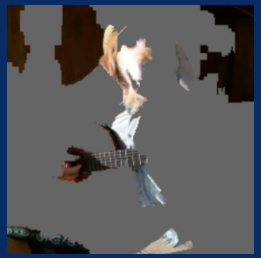


Superpixel segmentation

Sampling example - images

Original Image
P(labrador) = 0.21

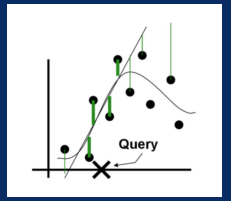
P(labrador) = 0.92



P(labrador) = 0.01

P(labrador) = 0.34

Locally weighted regression



Explanation

Explaining Inception V3

P(electric guitar) = 0.32

P(acoustic guitar) = 0.24

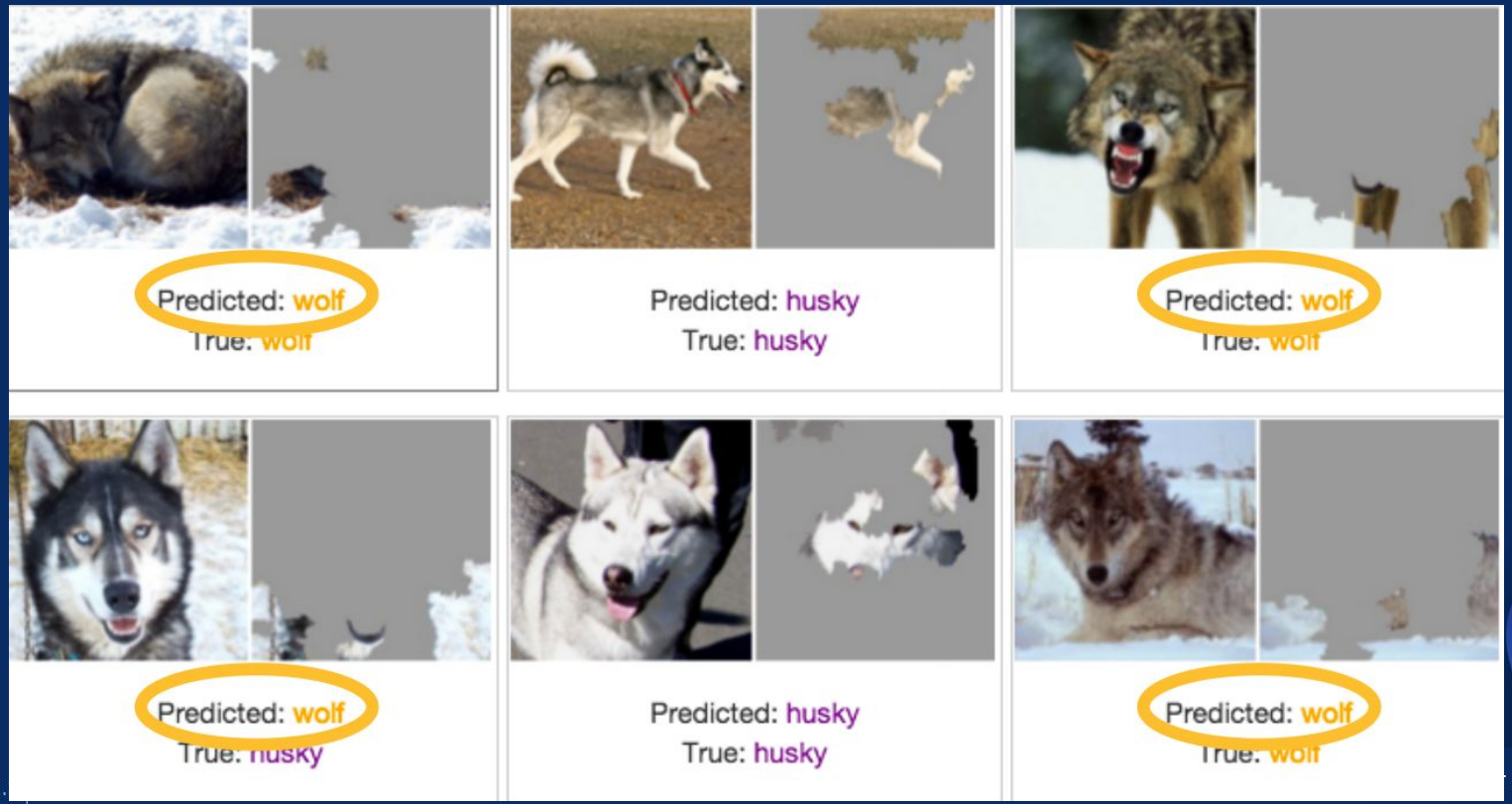
P(labrador) = 0.21

Predict wolf vs. husky

Predict wolf vs. husky

Only 1 mistake! Do you trust this classifier?

Explanations for predictions with LIME



Explanation should match our intuition

Original image

Bad classifier

Good classifier

Otherwise...

Otherwise...

MAN TAKES HOME ADORABLE FREE PUPPY ONLY TO FIND IT'S ACTUALLY A WOLF

Whoops.

BRANDON FRIEDERICH · OCT 12, 2016

6.2K SHARES

[Photos: Wolf Connection]

If you're on the hunt for a new canine companion, a sign reading "free puppy" is a

Learn more

- Project Lime - https://github.com/marcotcr/lime
  - pip install lime
  - Go to doc/notebooks to see simple examples
- Recent post by C. Olah on building blocks of interpretability
  - https://distill.pub/2018/building-blocks/
- Successor of lime: Anchors
  - https://github.com/marcotcr/anchor
- https://github.com/dongyp13/Robust-and-Explainable-Machine-Learning#interpretability

Credits:

- Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin
  - “Why should I trust you?” - explaining predictions of any classifier
- http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf
- https://www.youtube.com/watch?v=KP7-JtFMLo4
- https://www.fatml.org/schedule/2016/presentation/why-should-i-trust-you-explaining-predictions

After presentation question

You can choose between AI doctor who is 80% accurate and can explain its diagnoses or AI doctor who is 90% accurate but can’t explain, which one do you pick:

- 80% accurate with explanation
- 90% accurate without

Pedro Domingos
@pmddomingos

Obserwowany

Given the choice between an AI doctor that's 80% accurate and can explain its diagnoses and one that's 90% accurate but can't, I'd pick the latter.

Przetłumacz z języka: angielski

02:17 - 26 sty 2018

56 podań dalej **173** polubienia

28 56 173

“It ain’t what you don’t know that gets you in trouble. It’s what you know for sure that just ain’t so.”

Mark Twain

Q & A

matthew.opala@craftintiy.com
twitter/@matthewopala

Data.sphere.it

Join the conversation #sphereIT #DataSphere