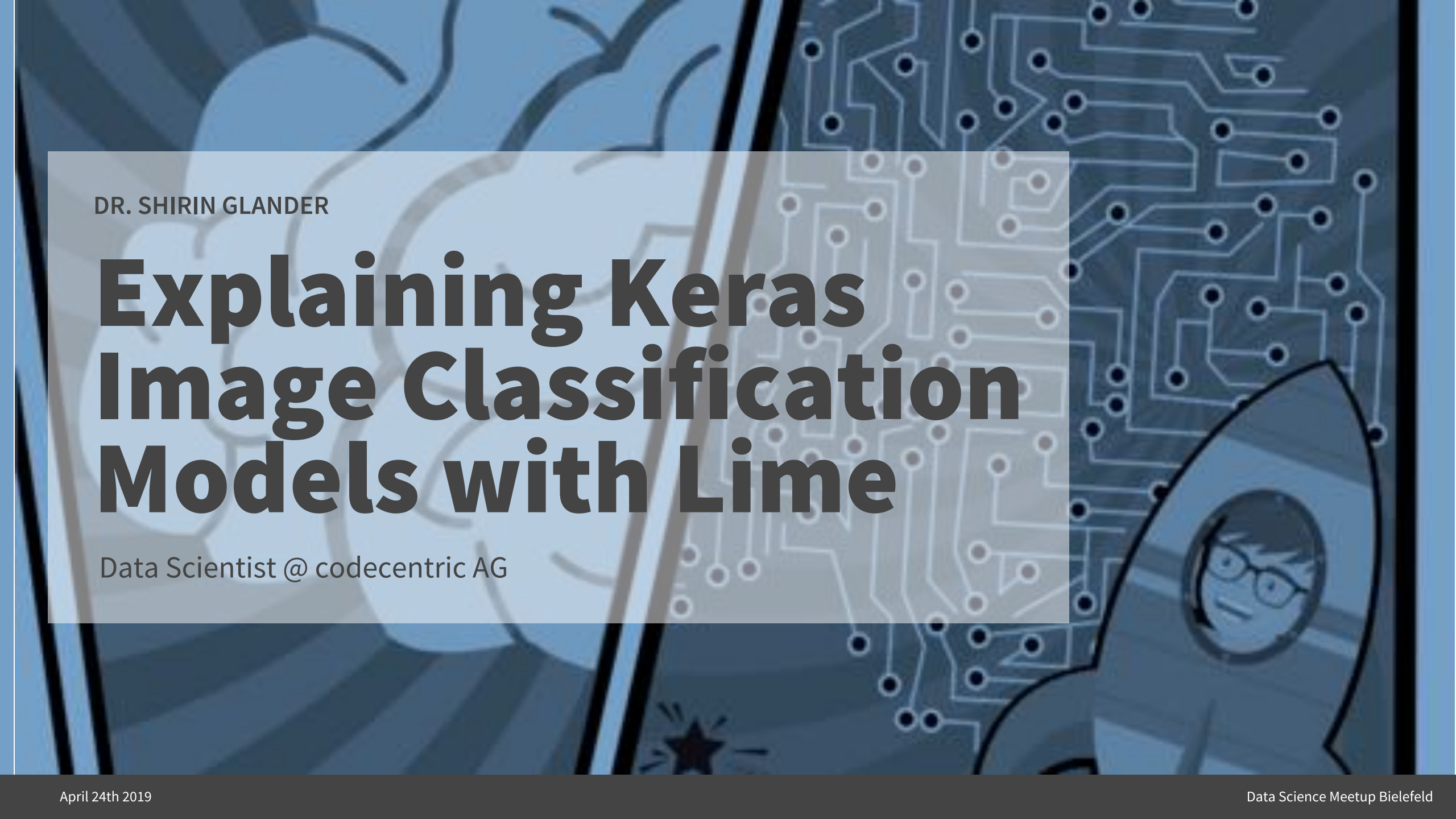
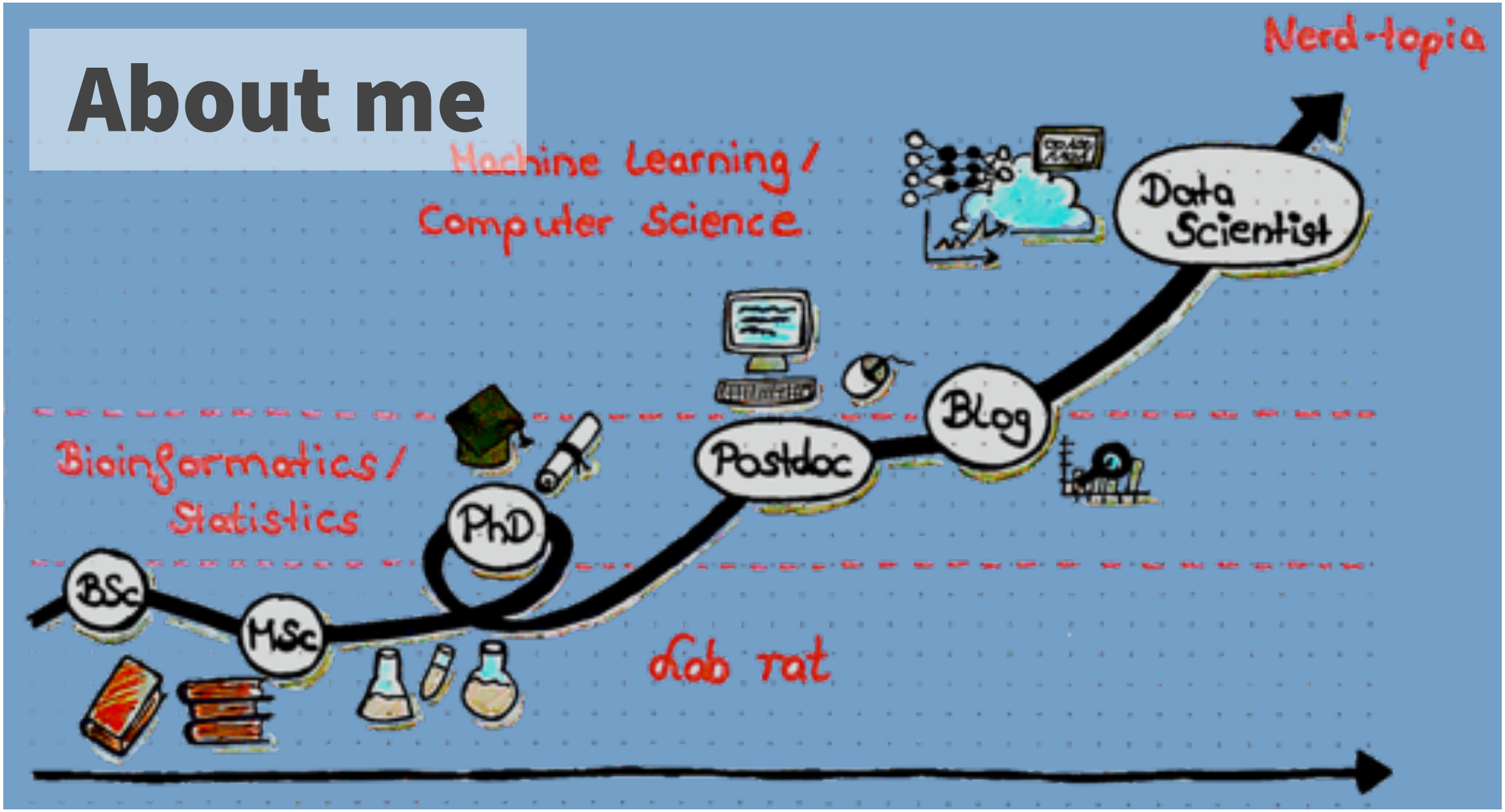


DR. SHIRIN GLANDER

Explaining Keras Image Classification Models with Lime

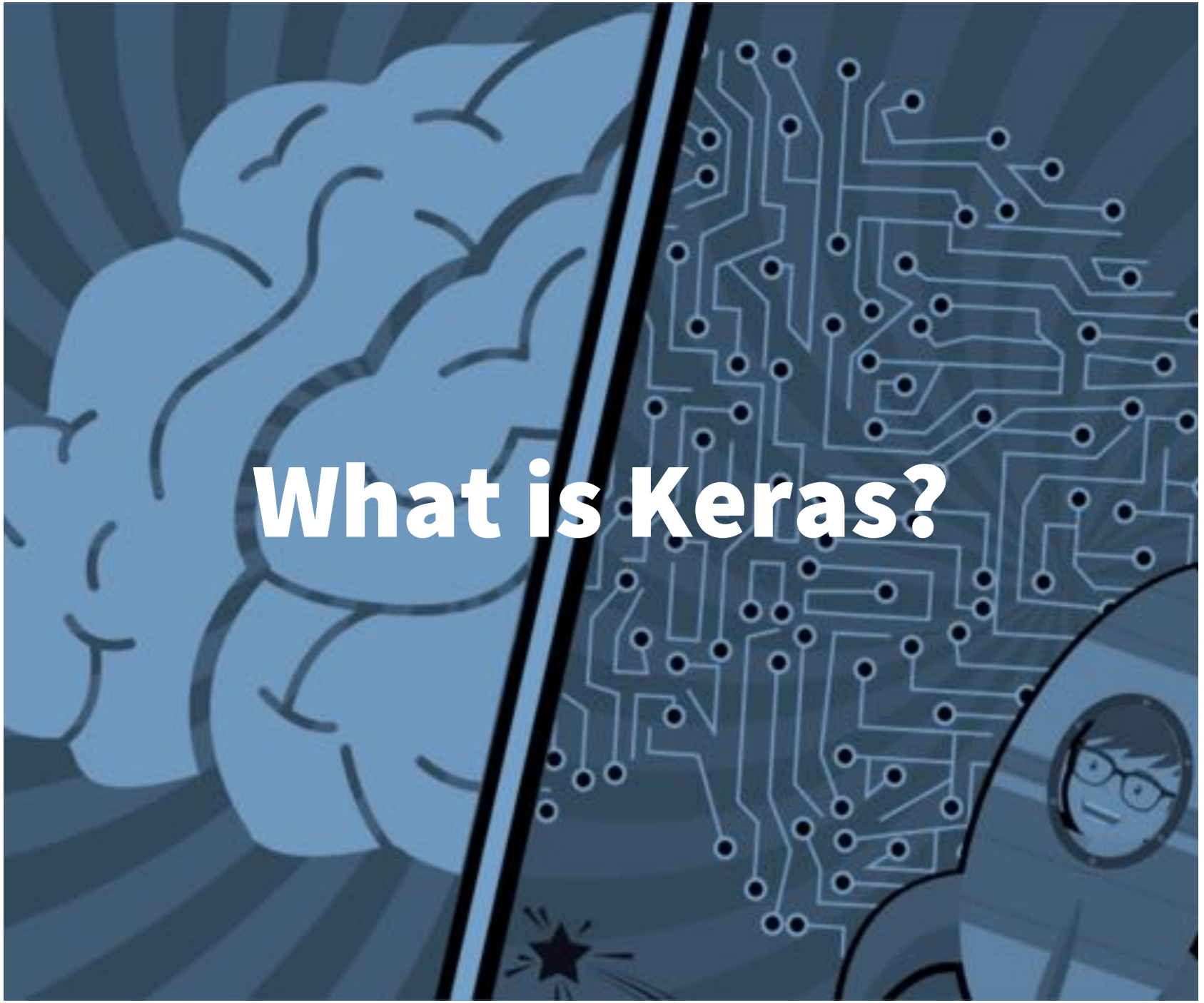
Data Scientist @ codecentric AG

About me


Nerd-topia
Machine Learning /
Computer Science
Data Scientist
Blog
Postdoc
PhD
MSc
BSc
Bioinformatics /
Statistics
lab rat

Overview - what you'll hear today

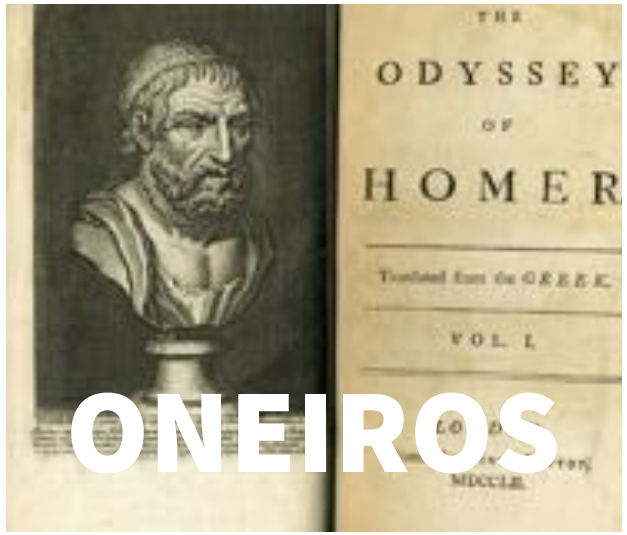
1. What is Keras?
2. Deep Learning & Neural Nets in a nutshell
3. Building an image classifier with Keras
4. Explaining image classification models with LIME

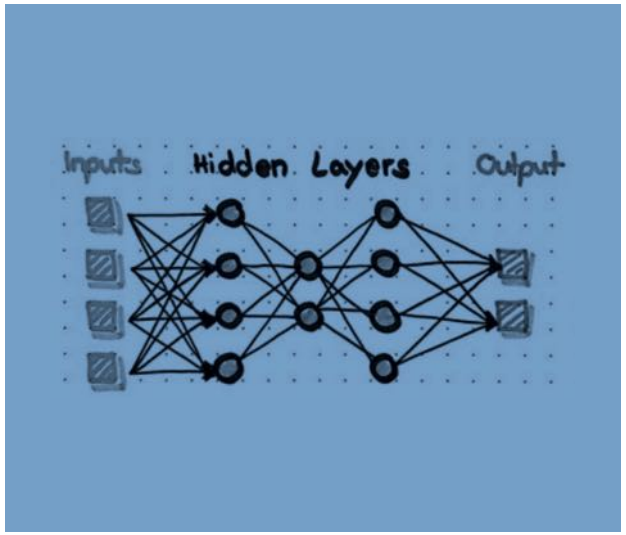

What is Keras?


K

An Open-source DL Framework

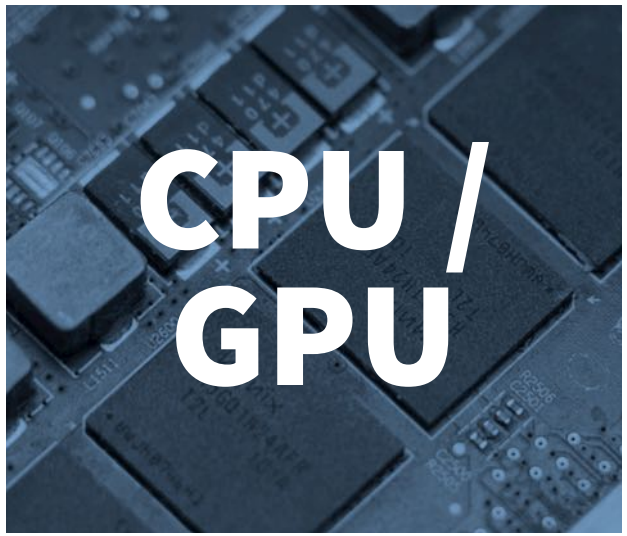
Keras
A deep learning library





"Oneiroi are beyond our unravelling -
who can be sure what tale they tell?
Not all that men look for comes to pass.
Two gates there are that give passage to fleeting Oneiroi;
one is made of horn, one of ivory.
The Oneiroi that pass through sawn ivory are deceitful,
bearing a message that will not be fulfilled;
those that come out through polished horn have truth behind them,
to be accomplished for men who see them."

Homer, Odyssey 19. 562 ff (Shewring translation).

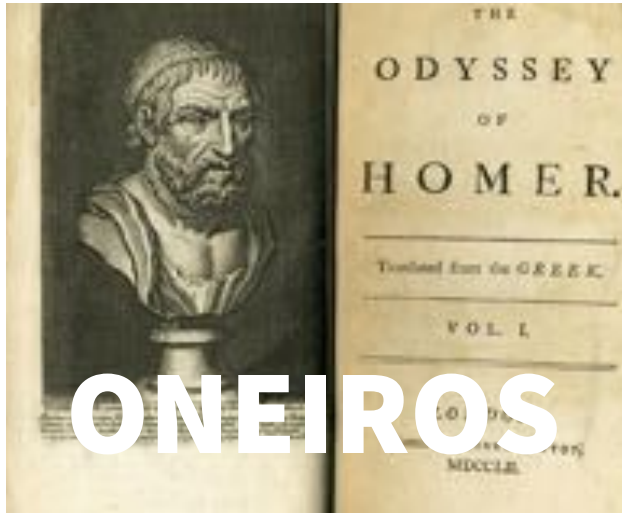


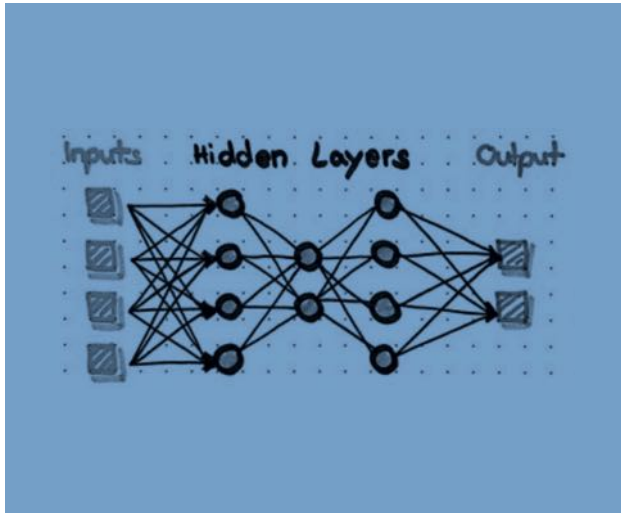
Developed by François Chollet

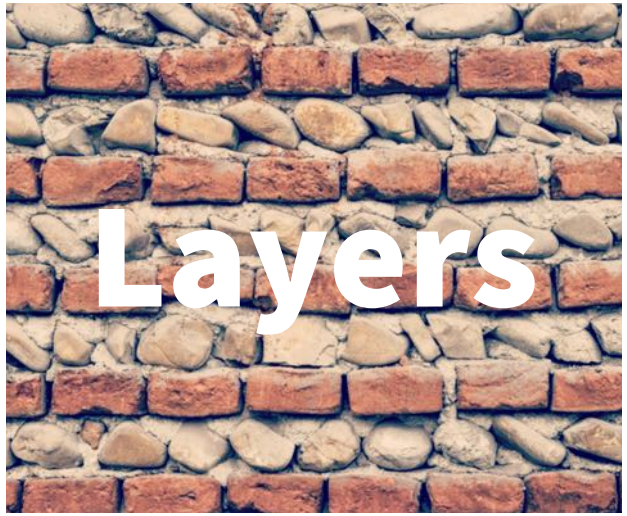
written in Python

Keras

A deep learning library

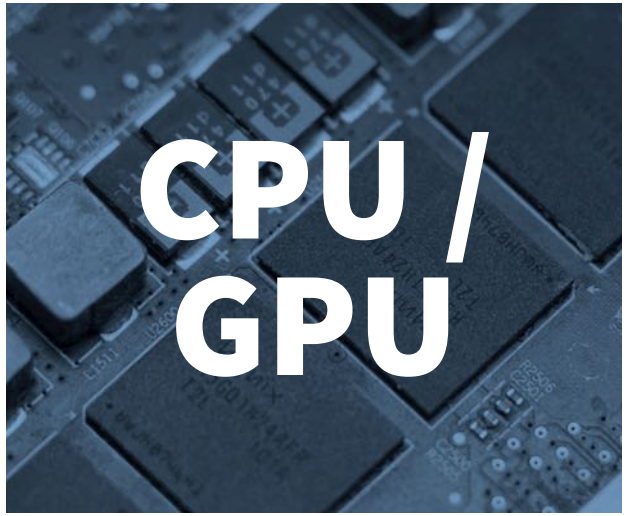






"Oneiroi are beyond our unravelling -
who can be sure what tale they tell?
Not all that men look for comes to pass.
Two gates there are that give passage to fleeting Oneiroi;
one is made of horn, one of ivory.
The Oneiroi that pass through sawn ivory are deceitful,
bearing a message that will not be fulfilled;
those that come out through polished horn have truth behind them,
to be accomplished for men who see them."

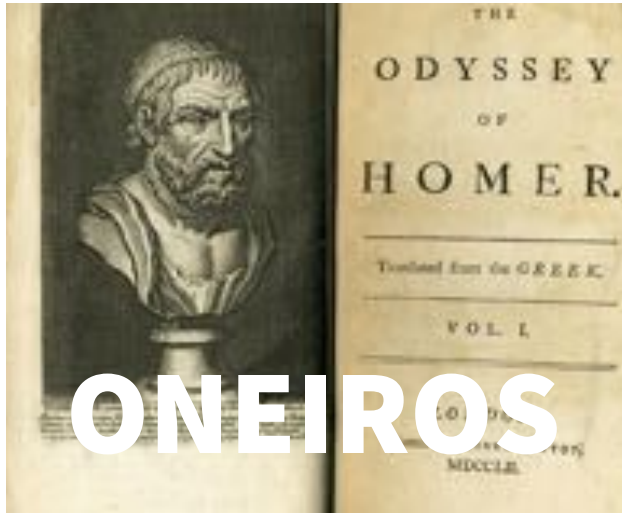
Homer, Odyssey 19. 562 ff (Shewring translation).

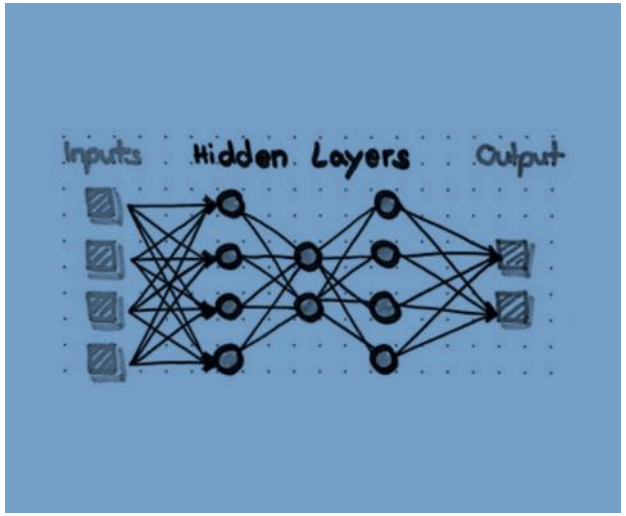


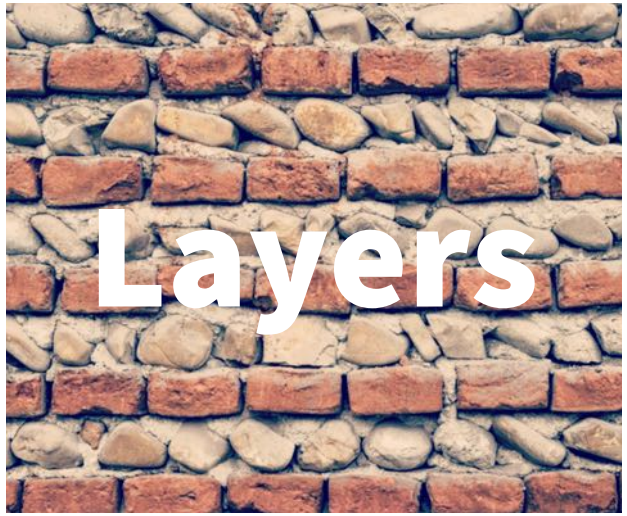
Originally from ONEIROS project

Open-Ended Neuro-Electronic Intelligent Robot OperatingSystem

Keras
A deep learning library

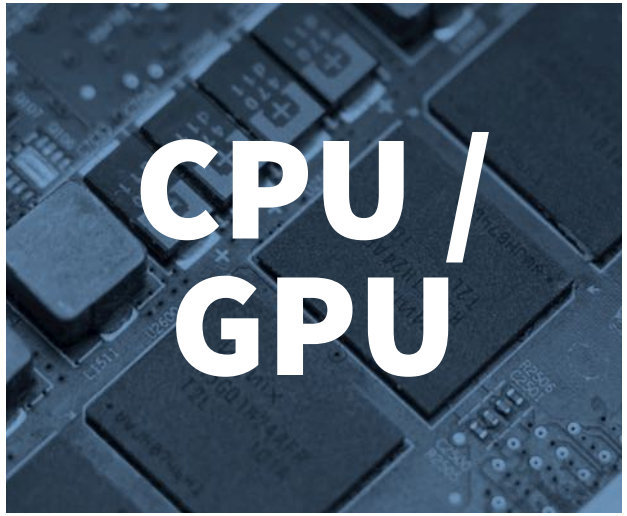






"Oneiroi are beyond our unravelling -
who can be sure what tale they tell?
Not all that men look for comes to pass.
Two gates there are that give passage to fleeting Oneiroi;
one is made of horn, one of ivory.
The Oneiroi that pass through sawn ivory are deceitful,
bearing a message that will not be fulfilled;
those that come out through polished horn have truth behind them,
to be accomplished for men who see them."

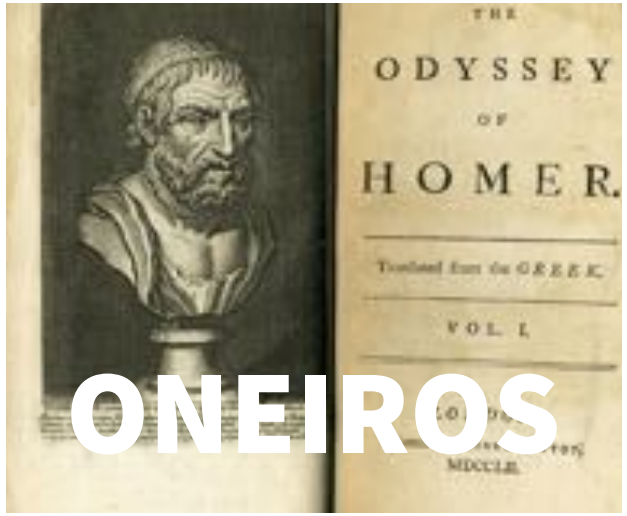
Homer, Odyssey 19. 562 ff (Shewring translation).

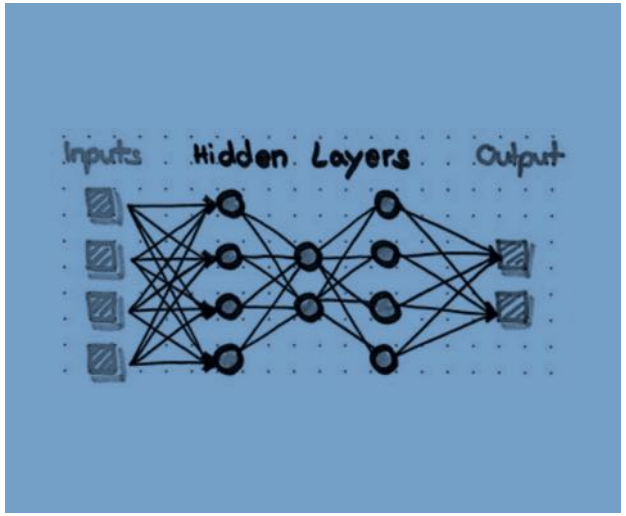


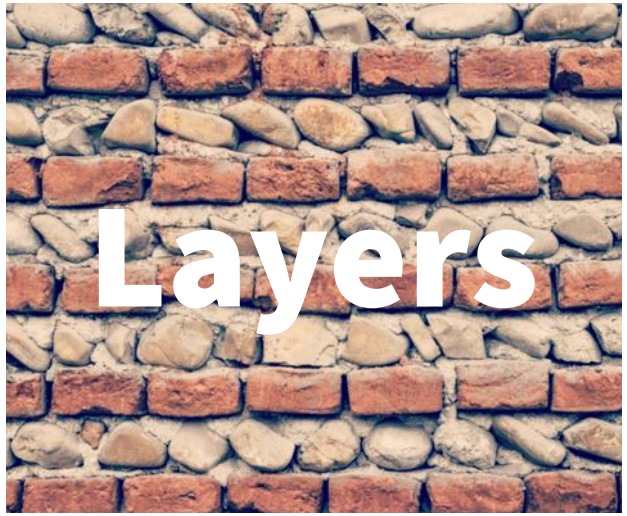
Originally from ONEIROS project

Keras means “horn” in Greek and comes from the Odyssey where it stands for true dreams

Keras
A deep learning library

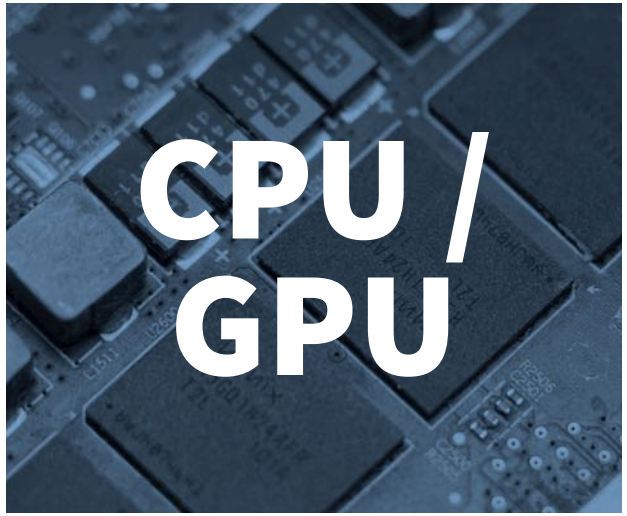






"Oneiroi are beyond our unravelling -
who can be sure what tale they tell?
Not all that men look for comes to pass.
Two gates there are that give passage to fleeting Oneiroi;
one is made of horn, one of ivory.
The Oneiroi that pass through sawn ivory are deceitful,
bearing a message that will not be fulfilled;
those that come out through polished horn have truth behind them,
to be accomplished for men who see them."

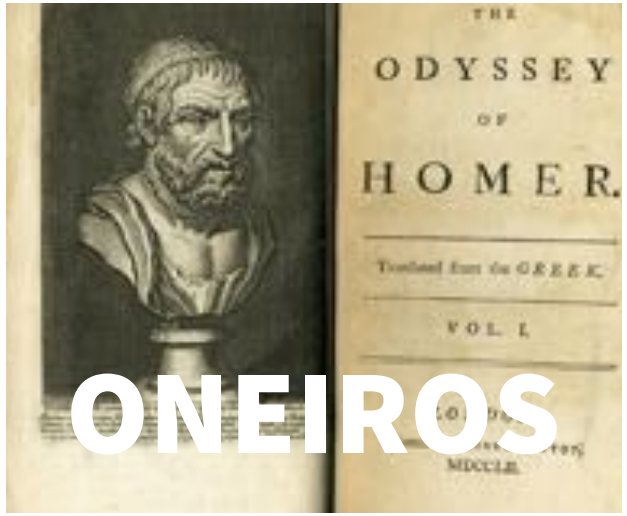
Homer, Odyssey 19. 562 ff (Shewring translation).

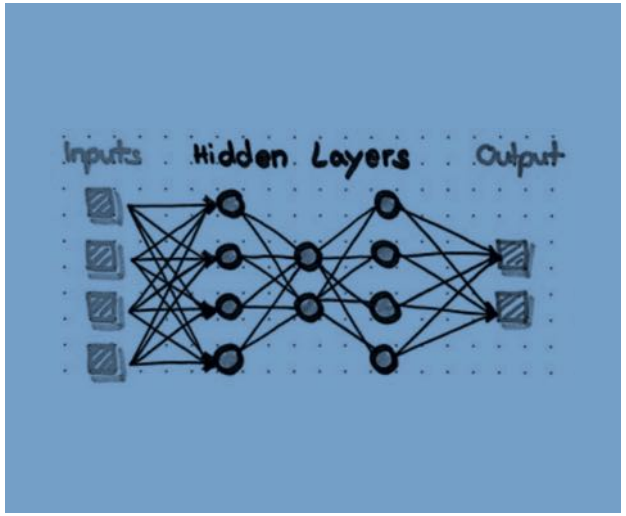


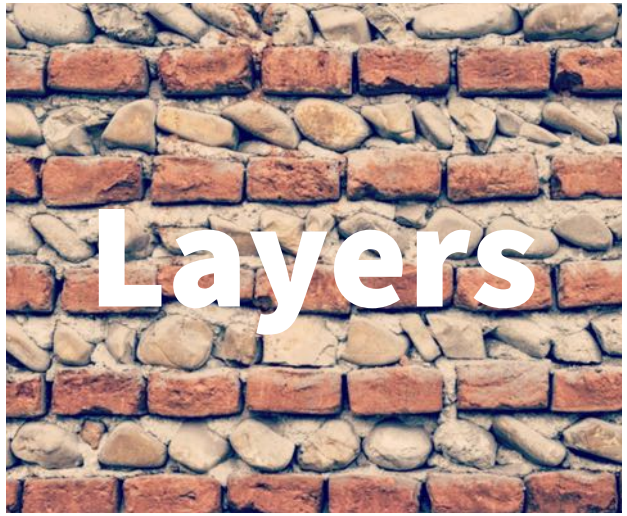
Lets you design neural nets

minimalistic, efficient, highly flexible, framework - independent

Keras
A deep learning library

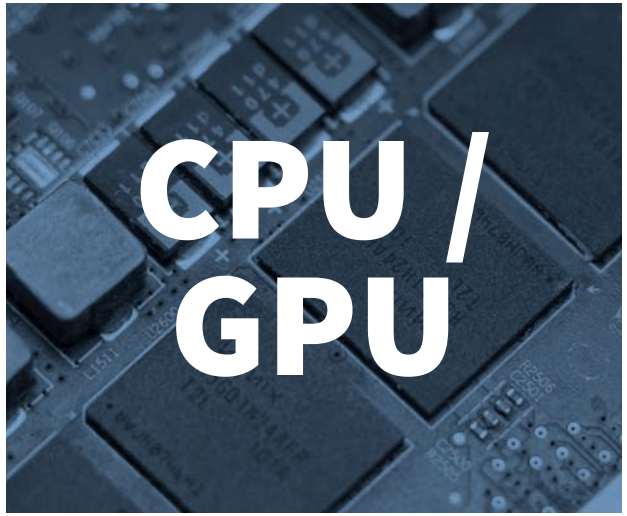






"Oneiroi are beyond our unravelling -
who can be sure what tale they tell?
Not all that men look for comes to pass.
Two gates there are that give passage to fleeting Oneiroi;
one is made of horn, one of ivory.
The Oneiroi that pass through sawn ivory are deceitful,
bearing a message that will not be fulfilled;
those that come out through polished horn have truth behind them,
to be accomplished for men who see them."

Homer, Odyssey 19. 562 ff (Shewring translation).

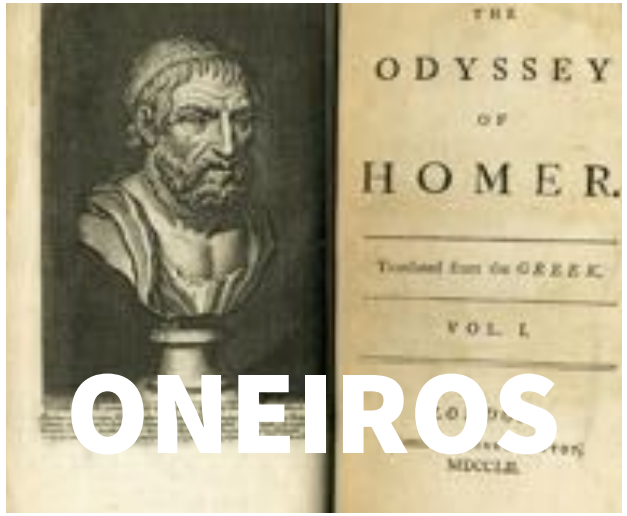


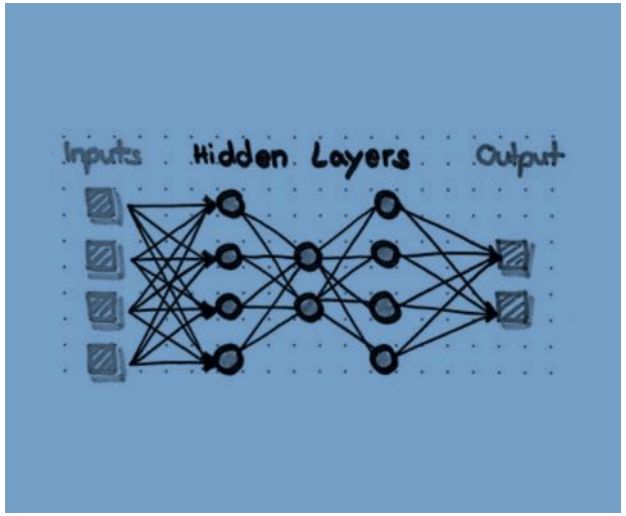
Lets you design neural nets

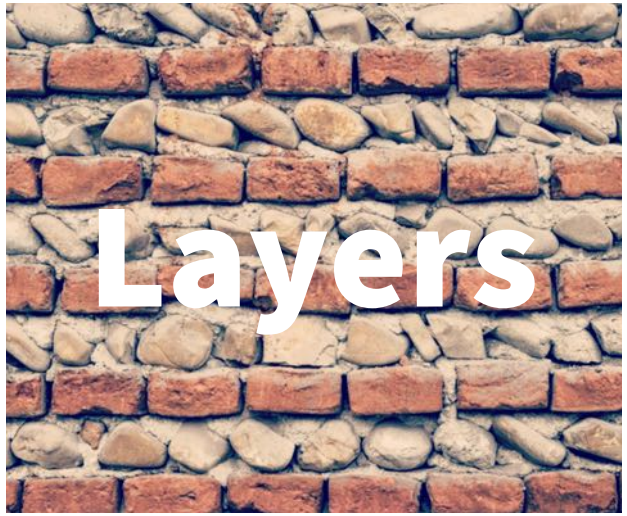
based on a layer system, like LEGO

Keras

A deep learning library

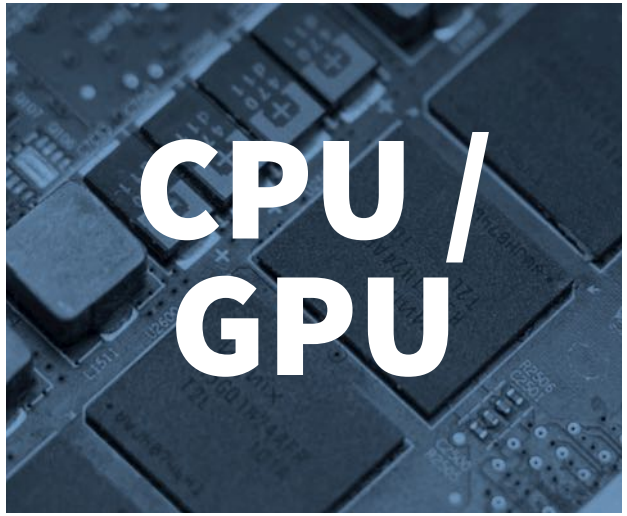






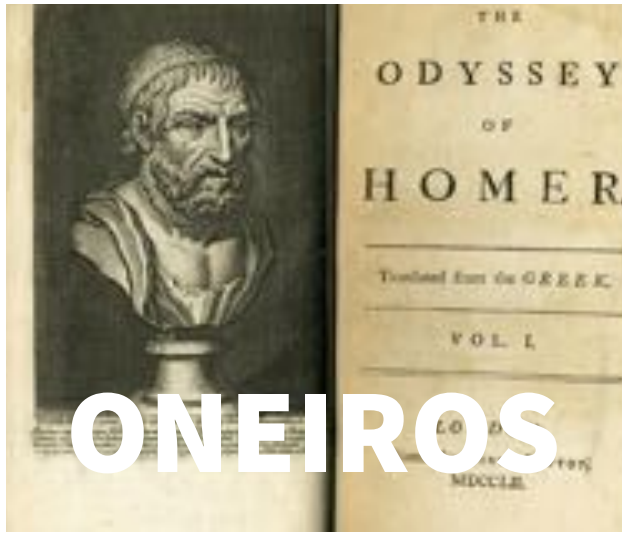
"Oneiroi are beyond our unravelling -
who can be sure what tale they tell?
Not all that men look for comes to pass.
Two gates there are that give passage to fleeting Oneiroi;
one is made of horn, one of ivory.
The Oneiroi that pass through sawn ivory are deceitful,
bearing a message that will not be fulfilled;
those that come out through polished horn have truth behind them,
to be accomplished for men who see them."

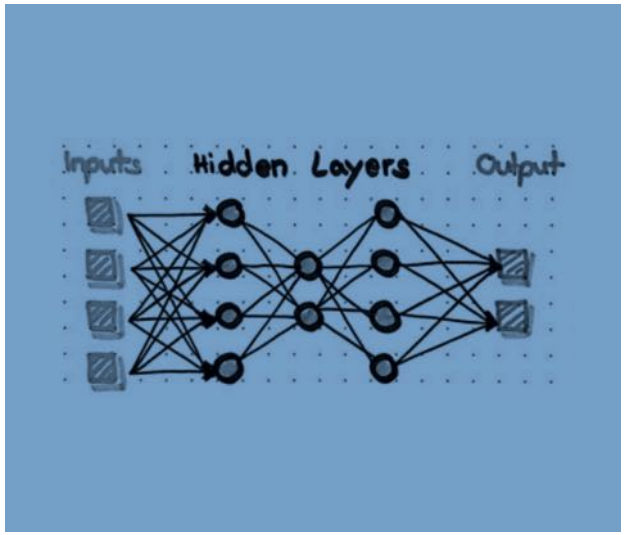
Homer, Odyssey 19. 562 ff (Shewring translation).



Benefit from the community

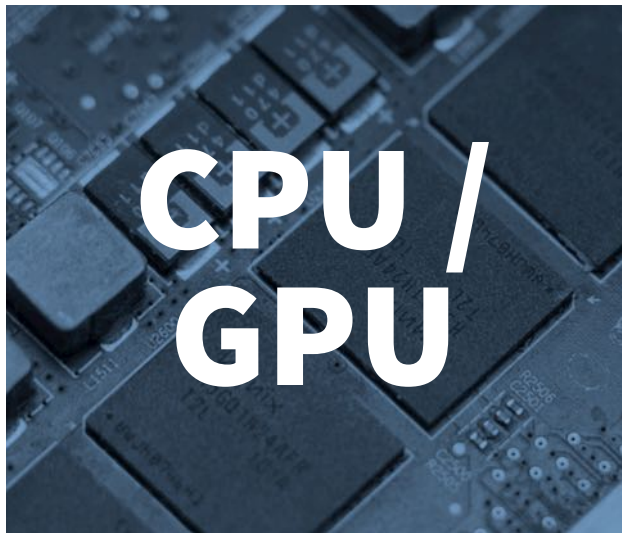
Keras
A deep learning library





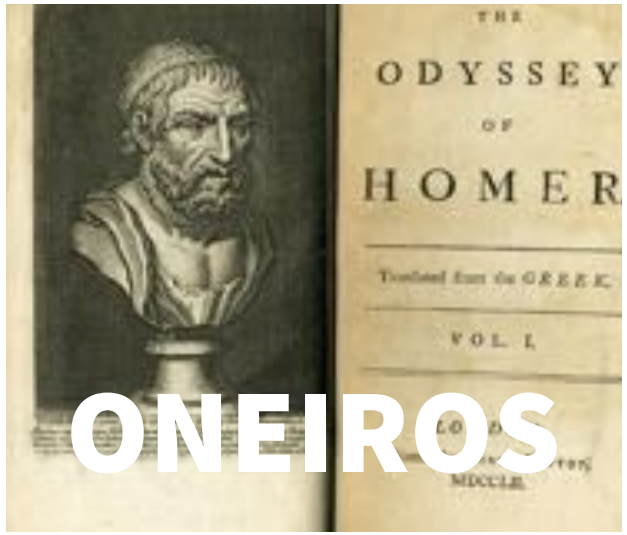
"Oneiroi are beyond our unravelling -
who can be sure what tale they tell?
Not all that men look for comes to pass.
Two gates there are that give passage to fleeting Oneiroi;
one is made of horn, one of ivory.
The Oneiroi that pass through sawn ivory are deceitful,
bearing a message that will not be fulfilled;
those that come out through polished horn have truth behind them,
to be accomplished for men who see them."

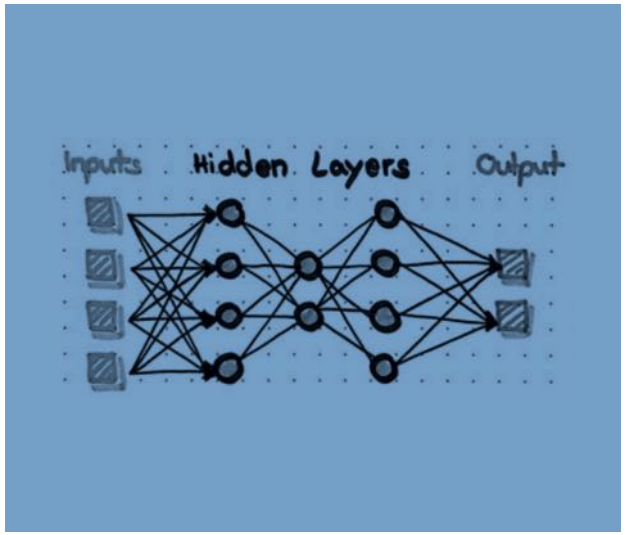
Homer, Odyssey 19. 562 ff (Shewring translation).



Models can train on CPU & GPU

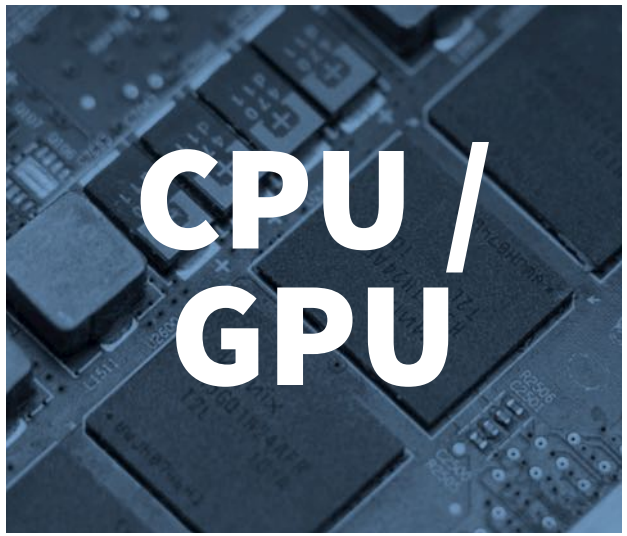
Keras
A deep learning library

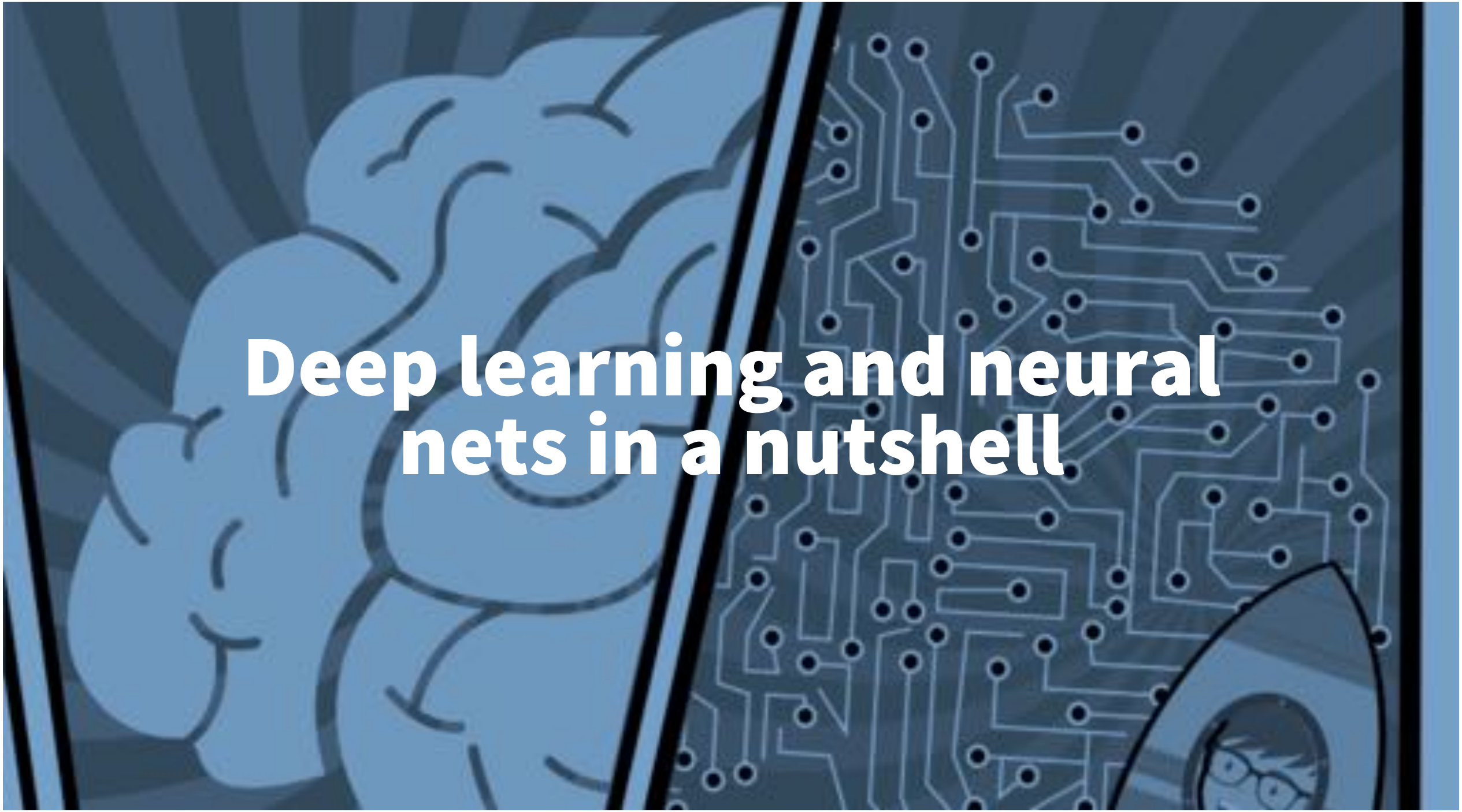




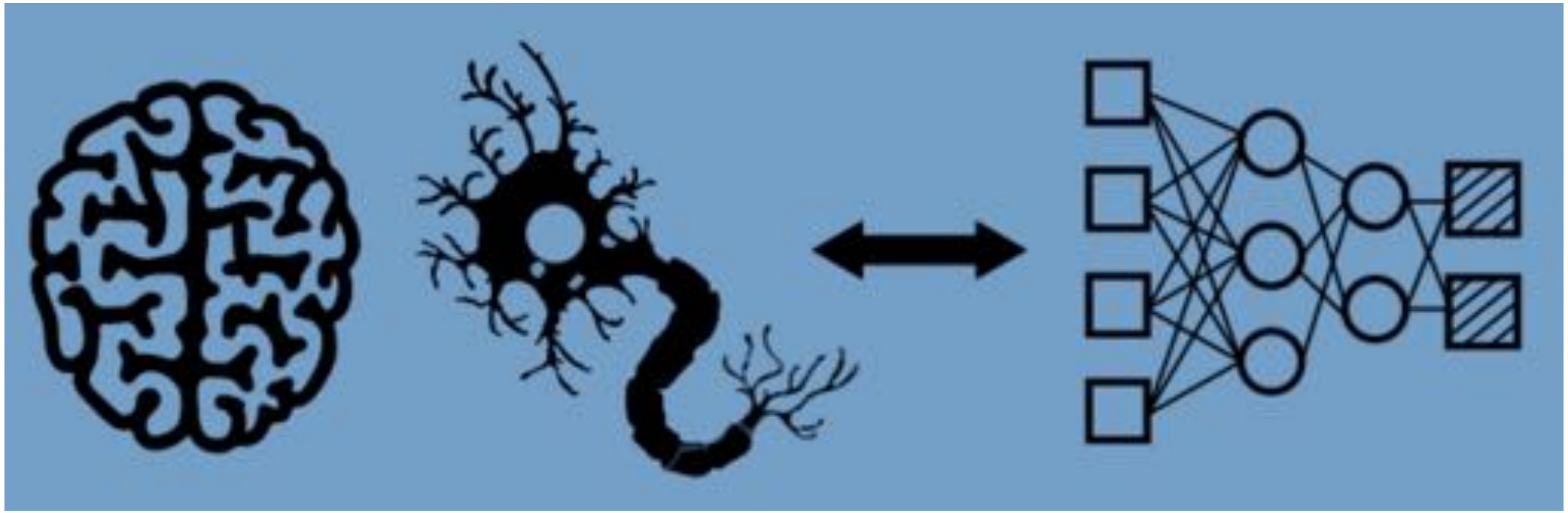
"Oneiroi are beyond our unravelling -
who can be sure what tale they tell?
Not all that men look for comes to pass.
Two gates there are that give passage to fleeting Oneiroi;
one is made of horn, one of ivory.
The Oneiroi that pass through sawn ivory are deceitful,
bearing a message that will not be fulfilled;
those that come out through polished horn have truth behind them,
to be accomplished for men who see them."

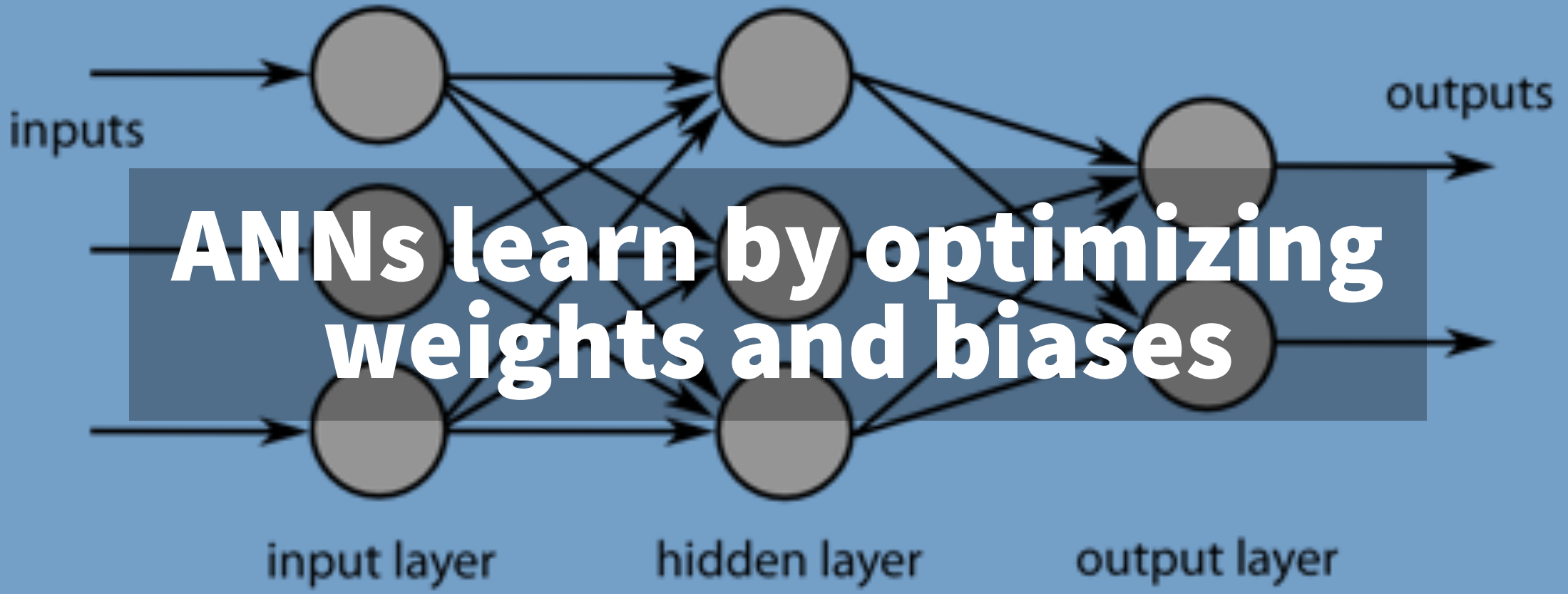
Homer, Odyssey 19. 562 ff (Shewring translation).




Deep learning and neural nets in a nutshell

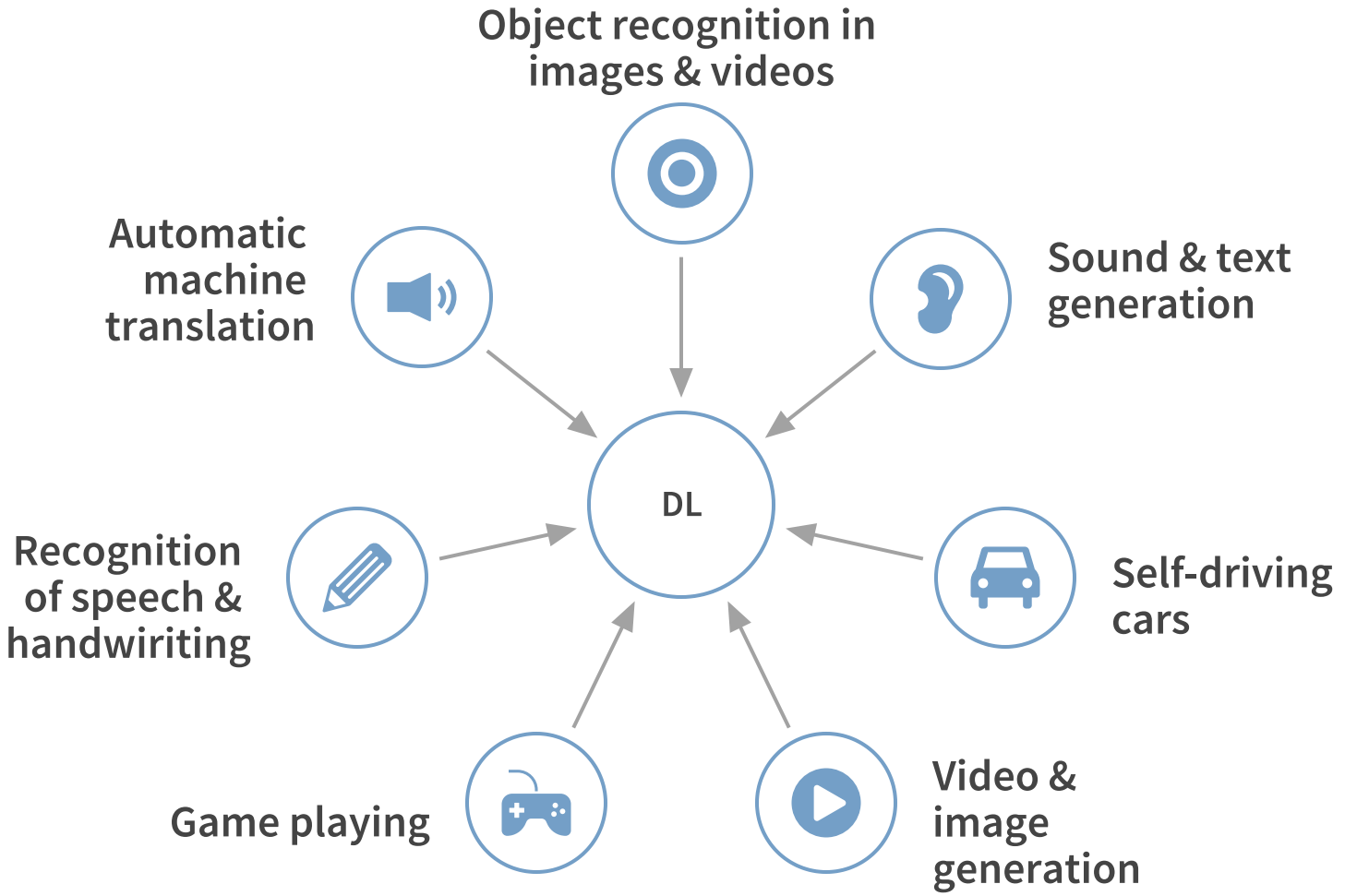
Neural nets are modeled after the human brain.





ANNs learn by optimizing weights and biases

Deep learning solves complex tasks.

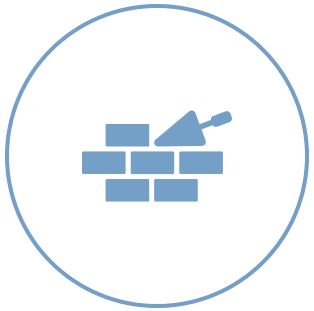


http://www.yaronhadad.com/deep-learning-most-amazing-applications/


Building an image
classifier with Keras

Build neural nets like LEGO systems

Two types of APIs exist for model building



Sequential models

simple

suitable for most cases

linear order of layers

only one direction from input to output

Functional API

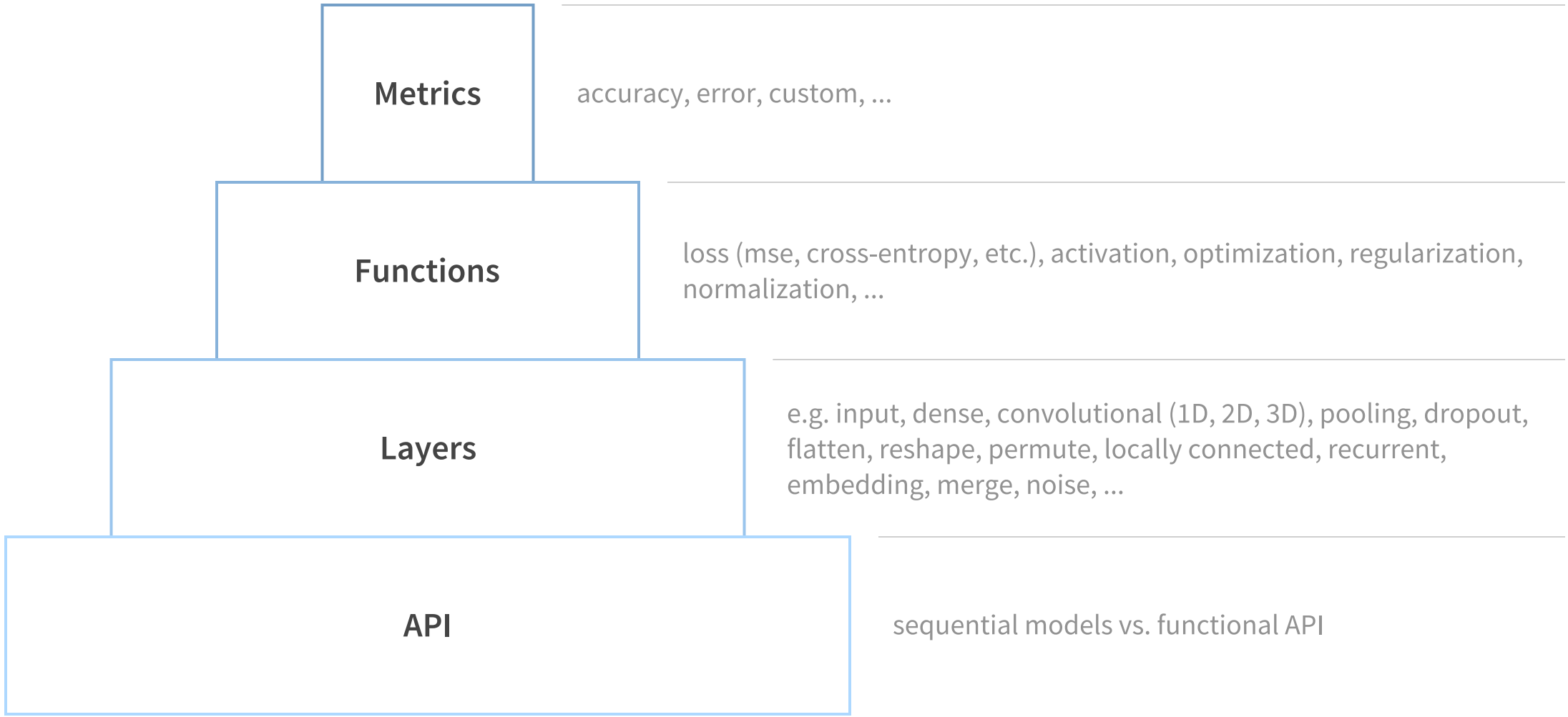
more complex

suitable for complex models

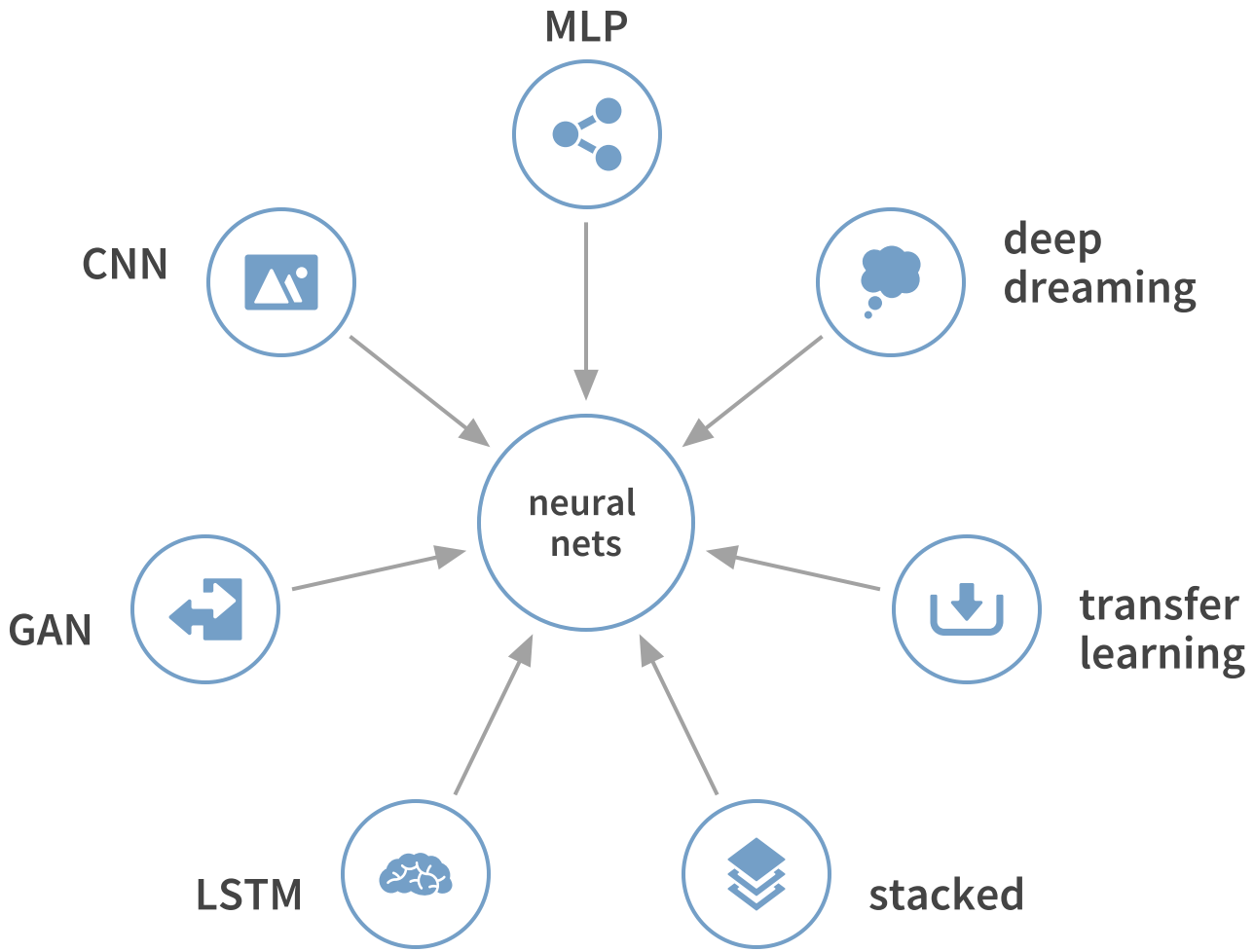
can have multiple in- or outputs

layers can be non-sequential, e.g. LSTM

Modularity of layers



Keras lets you build all kinds of neural nets



There are many options for working with Keras

Train models

- Jupyter notebook
- Python
- R

There are many options for working with Keras

Backend

- TensorFlow
- Caffe
- Theano
- ...

There are many options for working with Keras

Deploy

- TensorFlow object & serving
- JavaScript
- Scala
- ...

There are many options for working with Keras

Scale

- distributed learning with Spark and elephas
- ...

There are many options for working with Keras

Cloud

- AWS
- Azure

Let's train an image classifier

Code: https://shirinsplayground.netlify.com/2018/06/keras_fruits/
Data: https://www.kaggle.com/moltean/fruits/data

- **Classify fruits**

  Kiwi, Banana, Apricot, Avocado, Cocos, Clementine, Mandarine, Orange, Limes, Lemon, Peach, Plum, Raspberry, Strawberry, Pineapple

- **ca 490 images per class**

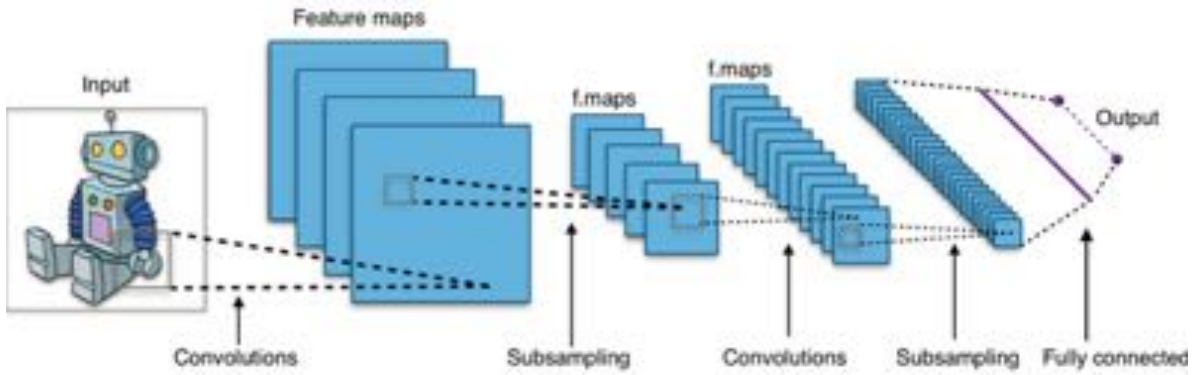
  3 channels

  100 x 100 px reduced to 20 x 20 px

  scaled (divided by 255)

Let's train an image classifier

with a Convolutional Neural Net



- **Convolutional Layer**

  1 - many

  2D or 3D

  sliding window calculates convolutions => filter kernels

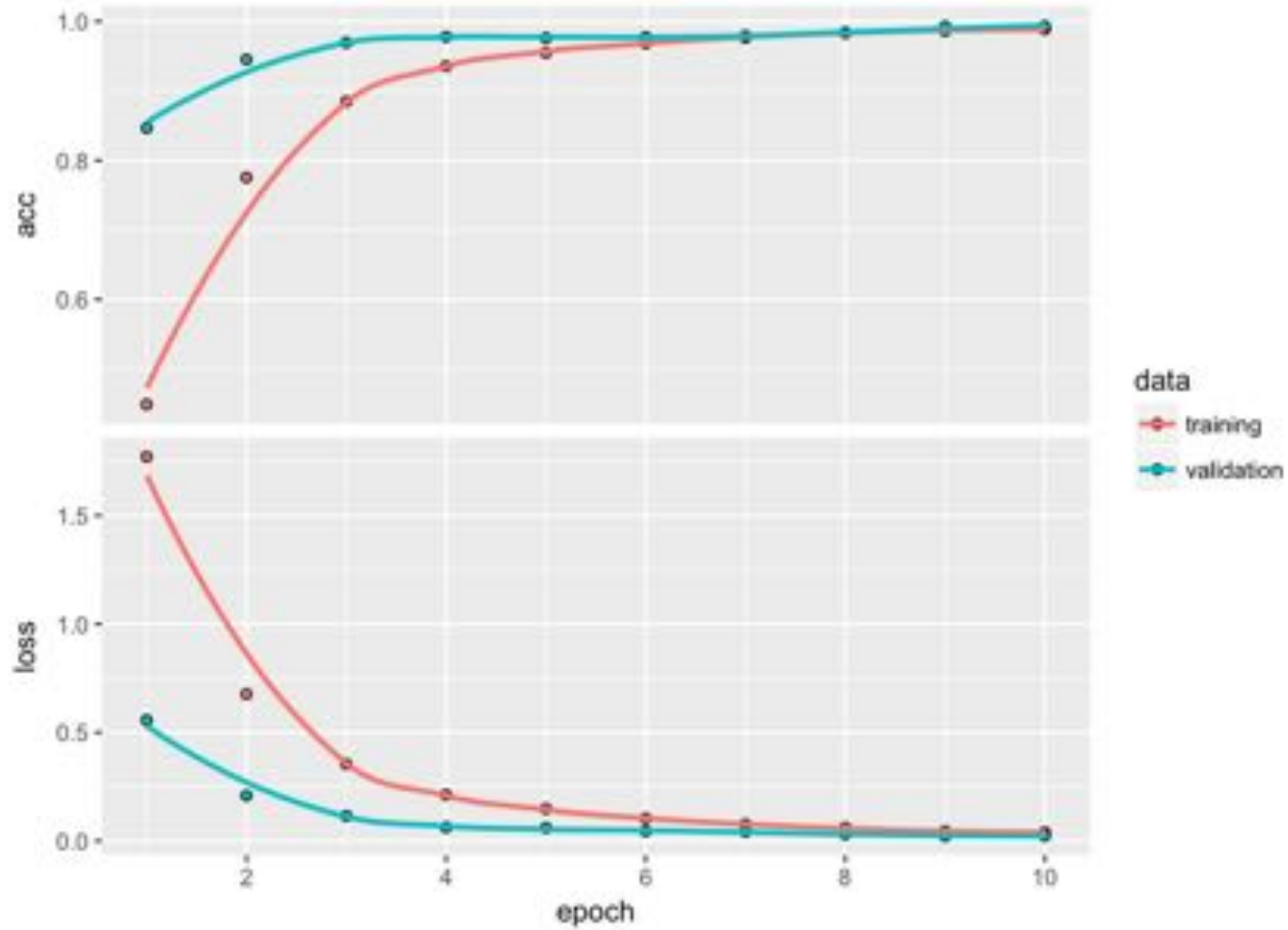
  edge detection => ... => complex features

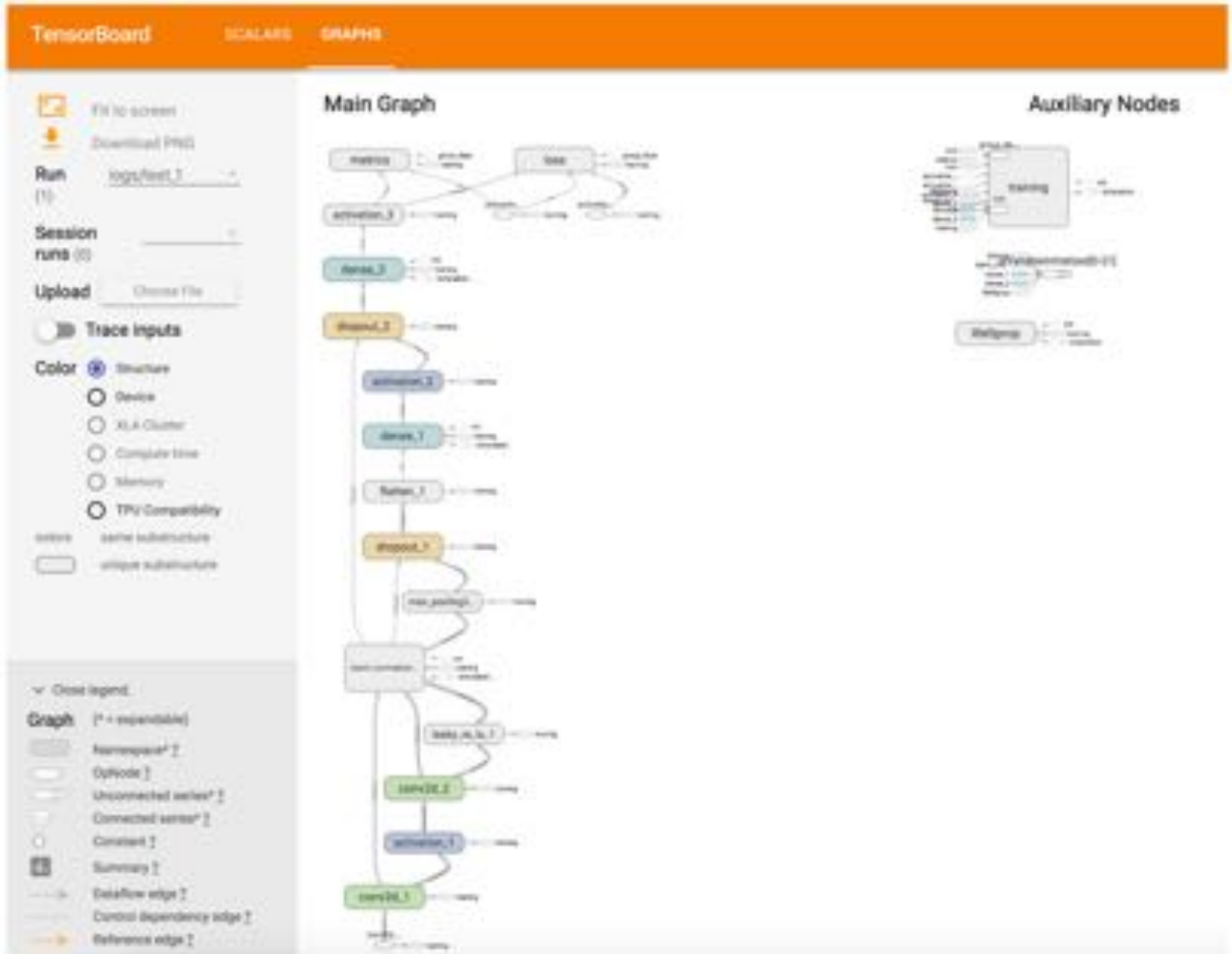
- **Pooling Layer**

  follows a (set of) convolutional layers

  condensing information, e.g. max pooling

```r
# initialise model
model <- keras_model_sequential()

# add layers
model %>%
  layer_conv_2d(filter = 32, kernel_size = c(3,3), padding = "same", input_shape = c(img_w
  layer_activation("relu") %>%

  # Second hidden layer
  layer_conv_2d(filter = 16, kernel_size = c(3,3), padding = "same") %>%
  layer_activation_leaky_relu(0.5) %>%
  layer_batch_normalization() %>%

  # Use max pooling
  layer_max_pooling_2d(pool_size = c(2,2)) %>%
  layer_dropout(0.25) %>%

  # Flatten max filtered output into feature vector
  # and feed into dense layer
  layer_flatten() %>%
  layer_dense(100) %>%
  layer_activation("relu") %>%
  layer_dropout(0.5) %>%

  # Outputs from dense layer are projected onto output layer
  layer_dense(output_n) %>%
  layer_activation("softmax")

# compile
model %>% compile(
  loss = "categorical_crossentropy",
  optimizer = optimizer_rmsprop(lr = 0.0001, decay = 1e-6),
  metrics = "accuracy"
)
```




TensorBoard
SCALARS
GRAPHS
Fit to screen
Download PNG
Run
(1)
logs/fast_1
Session runs (0)
Upload
Choose File
Trace inputs
Color
Structure
Device
XLA Cluster
Compute time
Memory
TPU Compatibility
same substructure
unique substructure
Close legend
Graph (* = expandable)
Namespace*
OpNode
Unconnected series*
Connected series*
Constant
Summary
Dataflow edge
Control dependency edge
Reference edge
Main Graph
Auxiliary Nodes


But the question remains:
Why?
Why does my model make the predictions it makes?

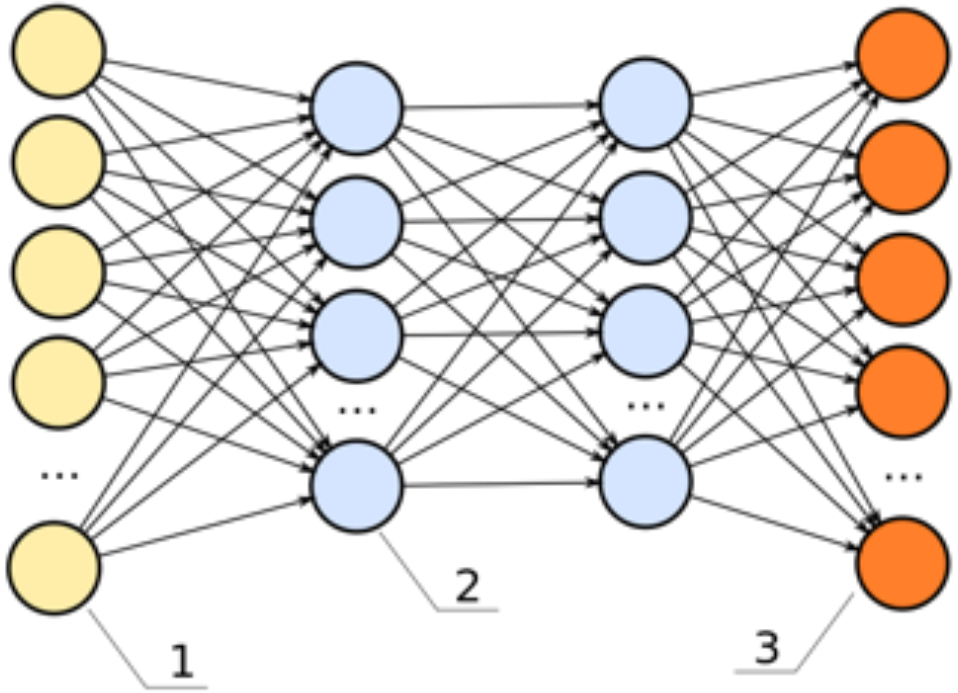

Explaining image classification models with LIME

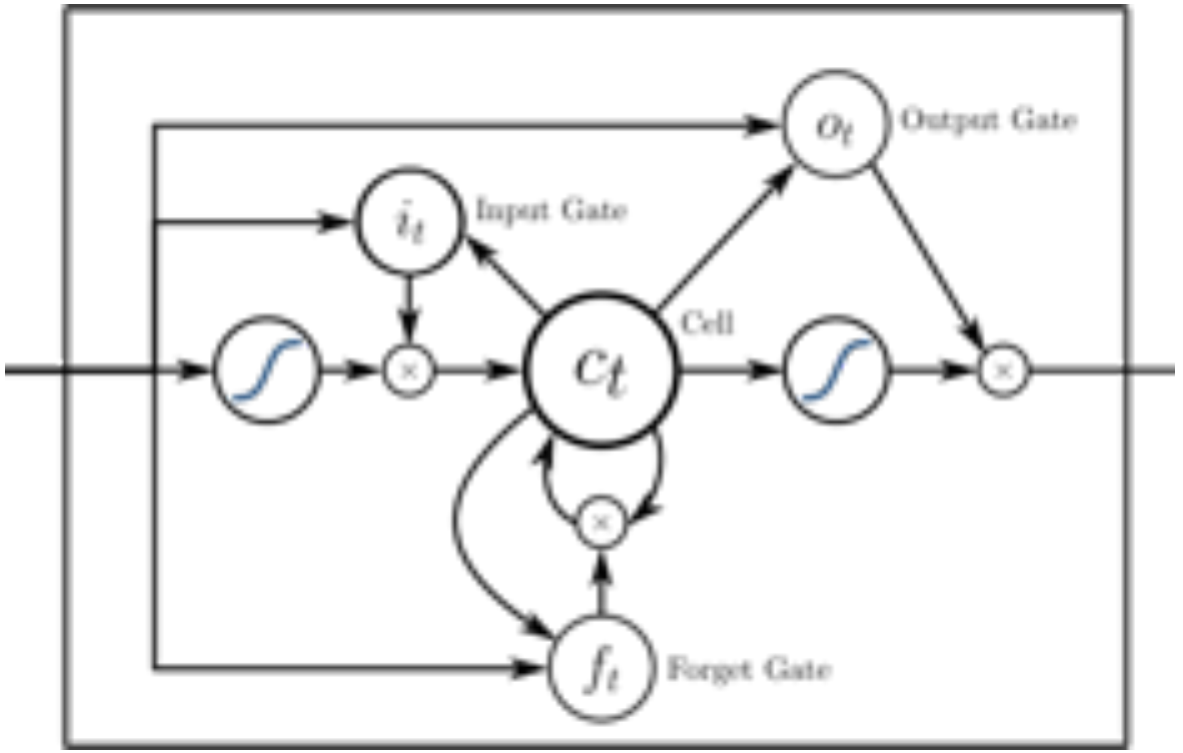
A typical ML workflow


But can we
trust
our models?
Are they fit for the wild?

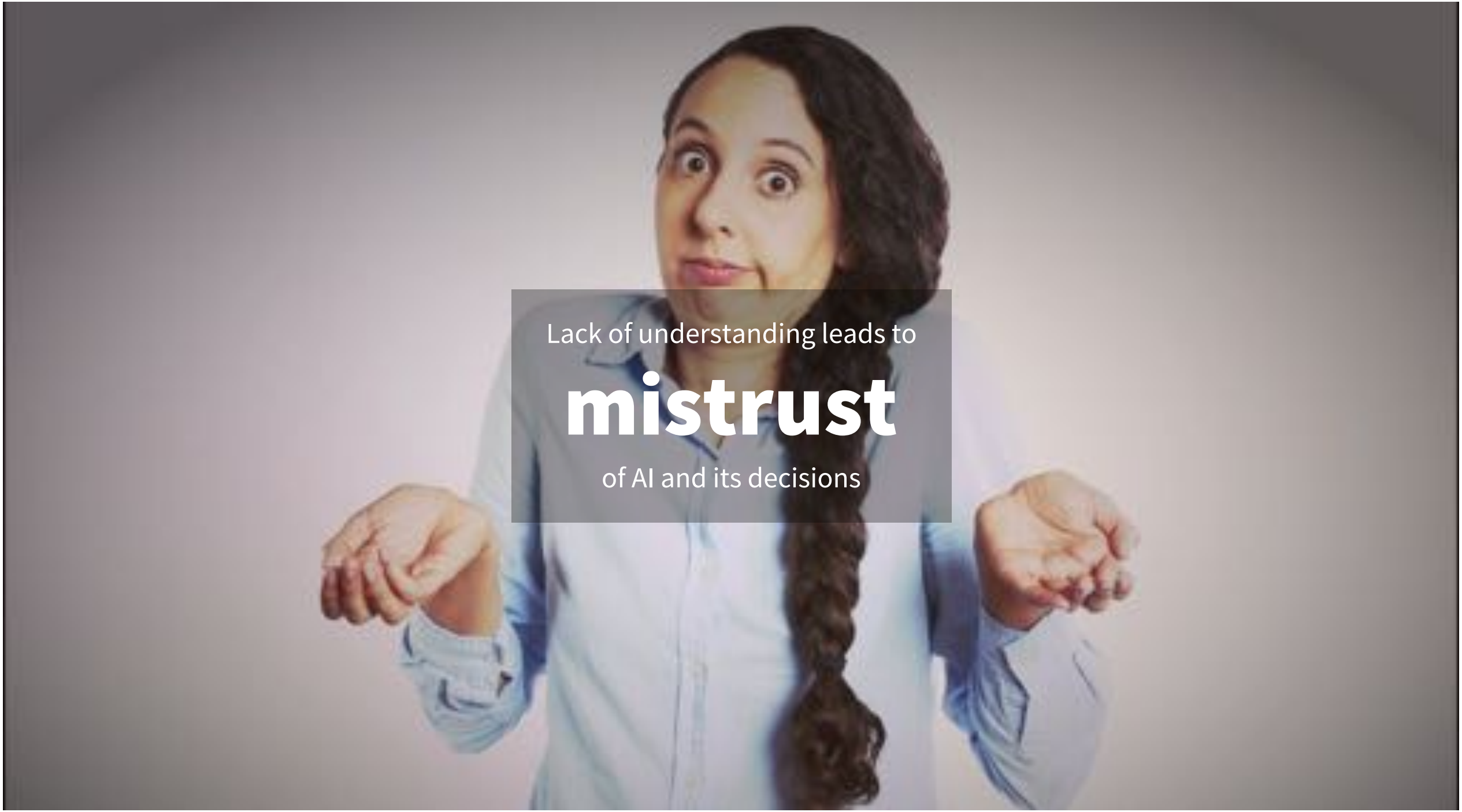
Trade-off between interpretability & complexity

Complexity tends to make models perform better but harder to understand.





Source: Wikipedia


Lack of understanding leads to
mistrust
of AI and its decisions

General Data Protection Regulation (GDPR)

- **“[...] the controller shall provide [...] the following information: the existence of automated decision-making and [...] meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.”**

Art. 13-15 & 22 Regulation (EU) 2016/679

https://dsgvo-gesetz.de/

Why should we improve our understanding of ML models?

... if technically it isn't necessary ...

Improving our models

Generalisability

“Sanity Check”

Prevent wrong conclusions &

potentially adversarial attacks

Trust and transparency

Can I trust my model's decisions?

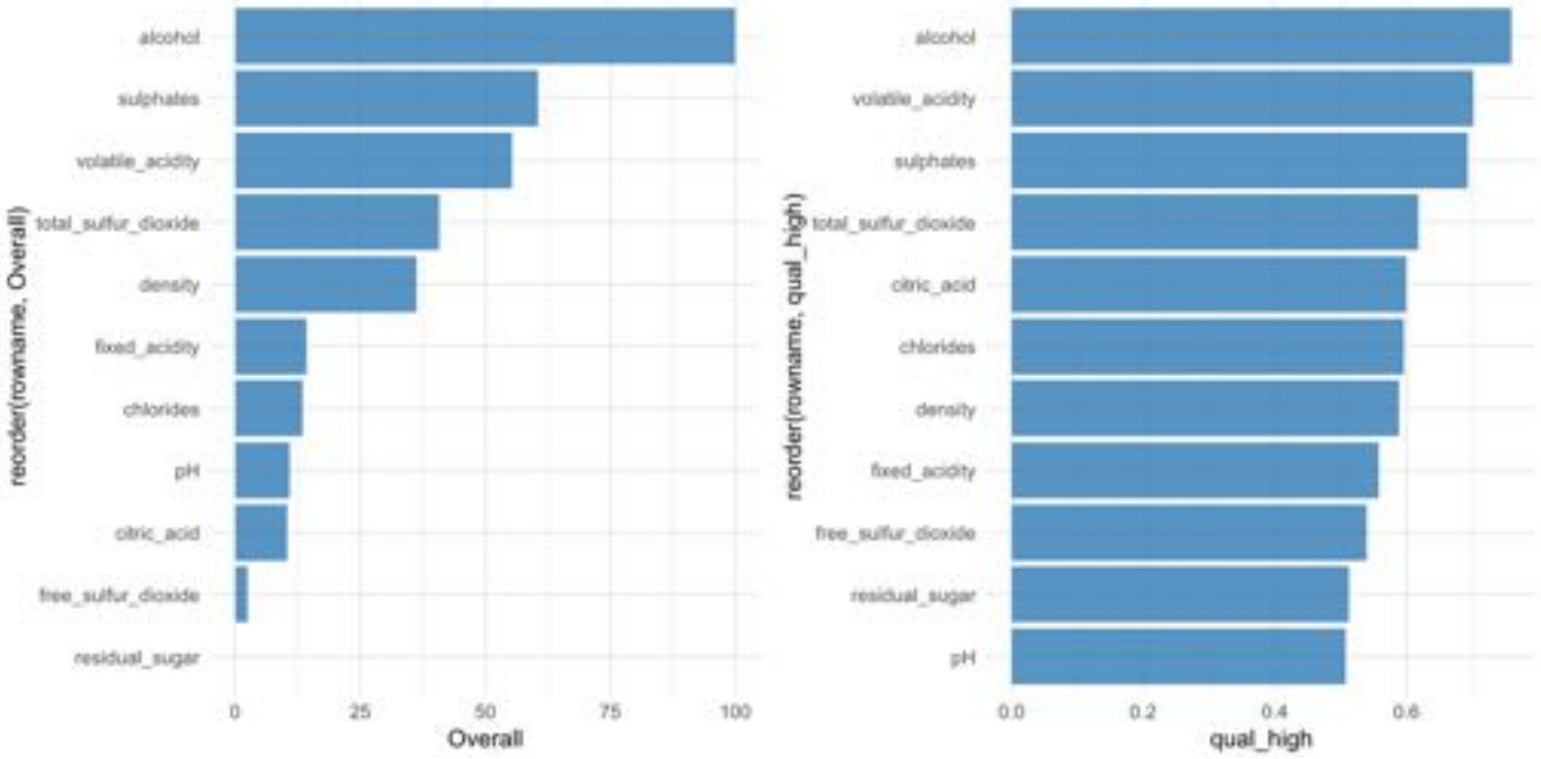
Why does my model make the predictions it makes?

Prevent Bias

Fairness

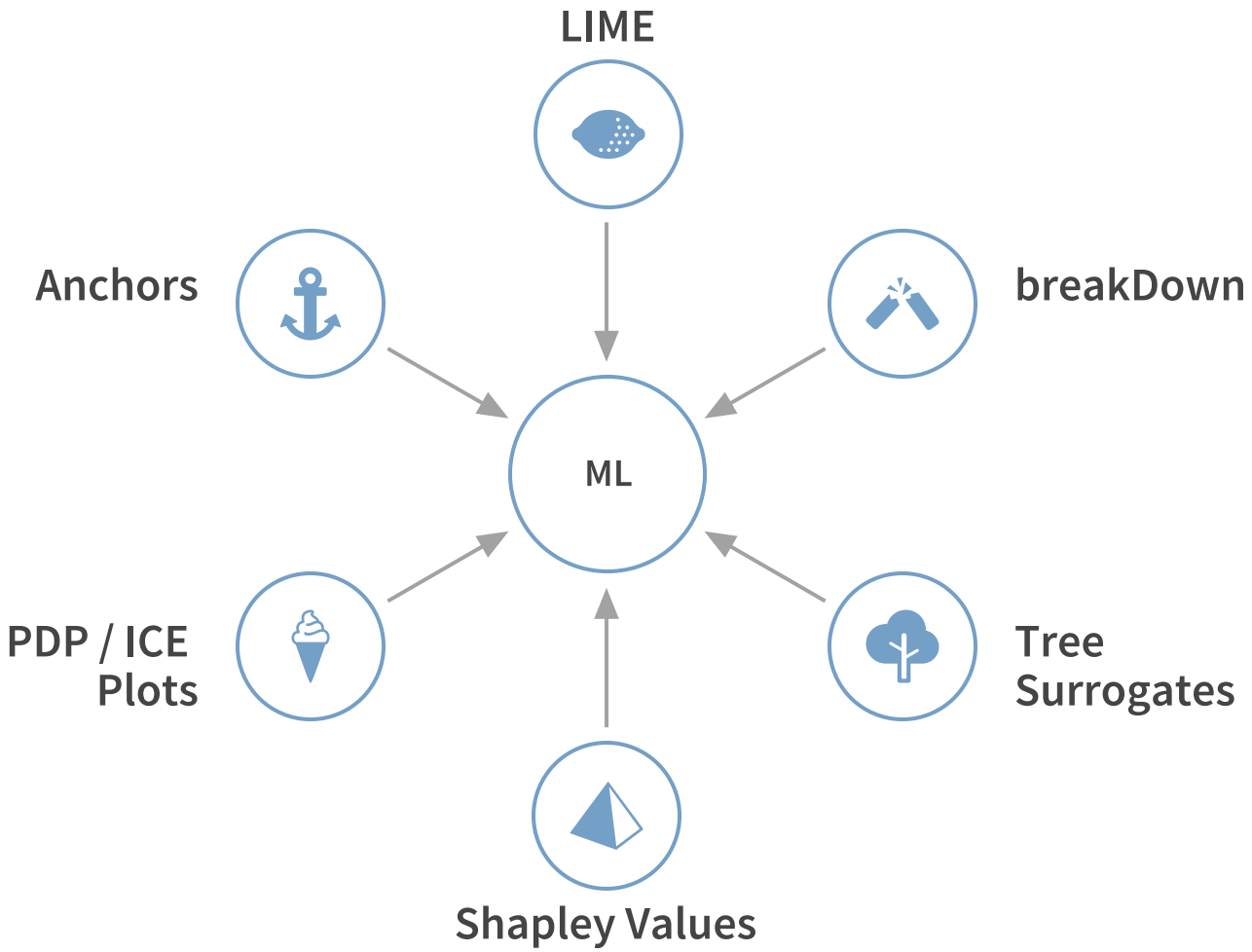
Identify and prevent bias

Interpretability is more than feature importance!



https://shirinsplayground.netlify.com/2018/07/explaining_ml_models_code_caret_iml/

Approaches to help explain and interpret ML models



https://shirinsplayground.netlify.com/2018/07/explaining_ml_models_code_caret_iml/

Local Models: LIME

Local Interpretable Model-agnostic Explanations

- **“Why Should I Trust You?”: Explaining the Predictions of Any Classifier**

  Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin.

  CoRR 2016

- **Written in Python**

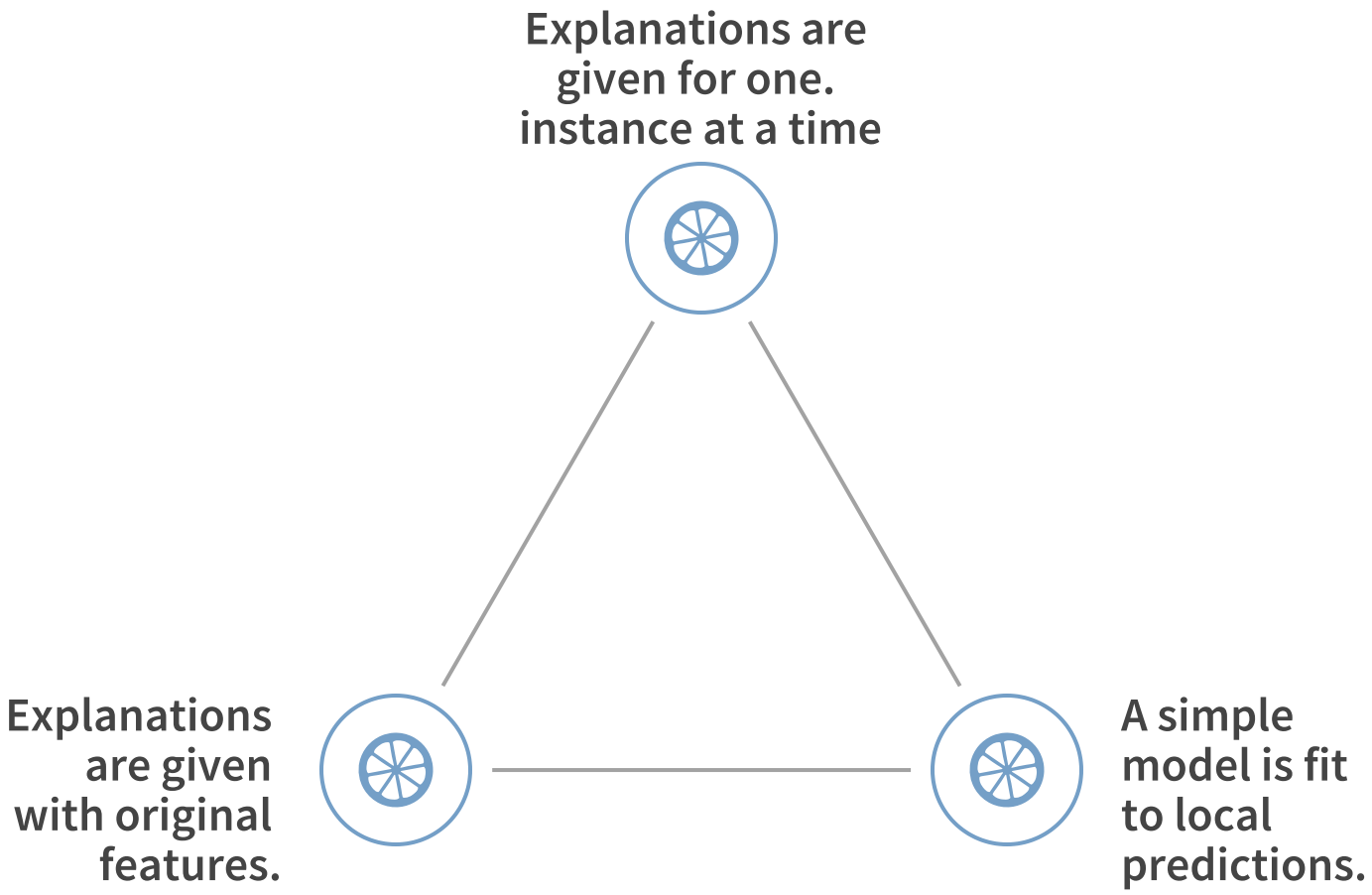
  https://github.com/marcotcr/lime

- **Adapted for R**

  https://github.com/thomasp85/lime

How does LIME work?

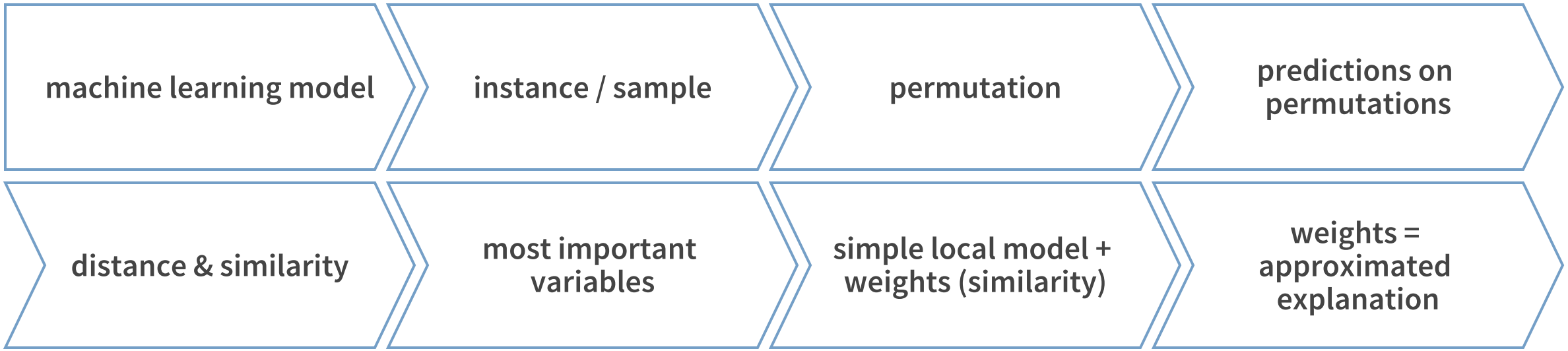
3 principles




YouTube DE
Suchen
"Why should I trust you?"
Explaining the predictions of any classifier
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin
University of Washington
https://youtu.be/hUnRCxnydCc

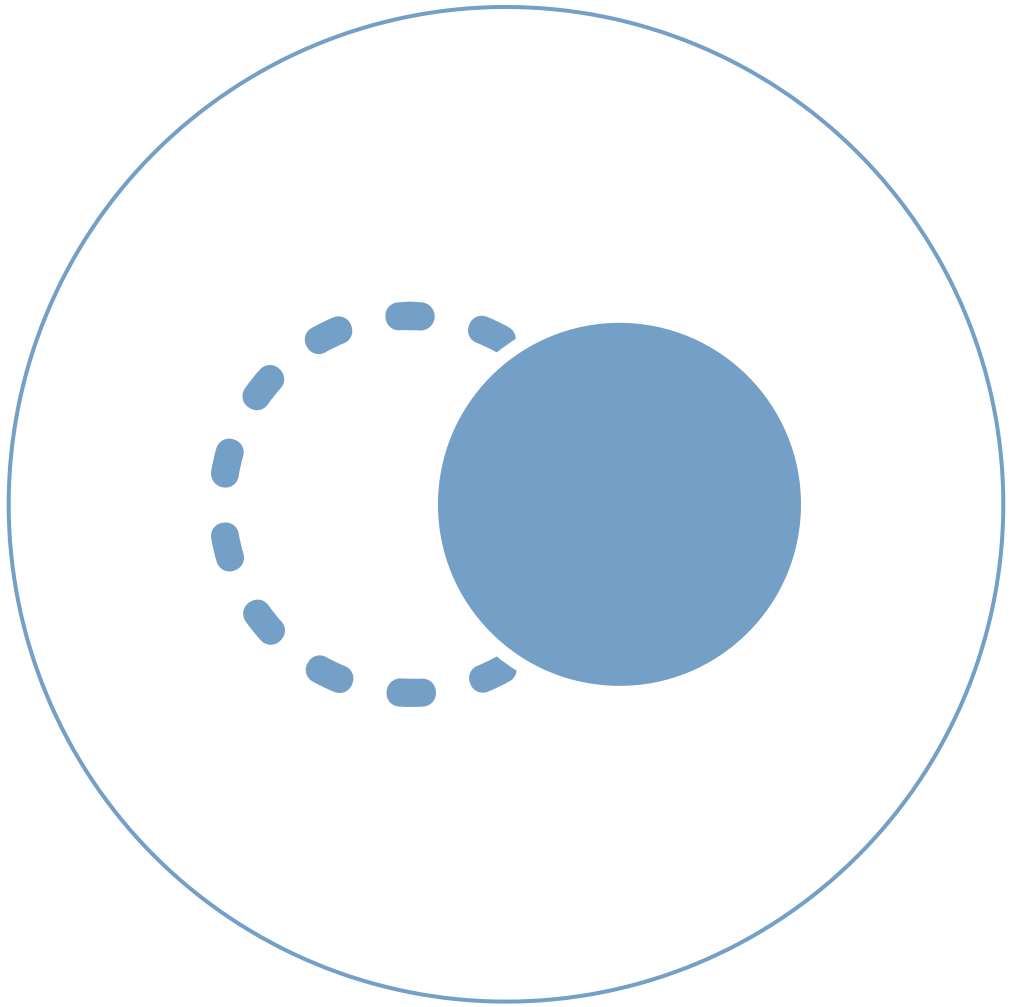
How can LIME give explanations of complex models?

With the following workflow:



https://cran.r-project.org/web/packages/lime/vignettes/Understanding_lime.html

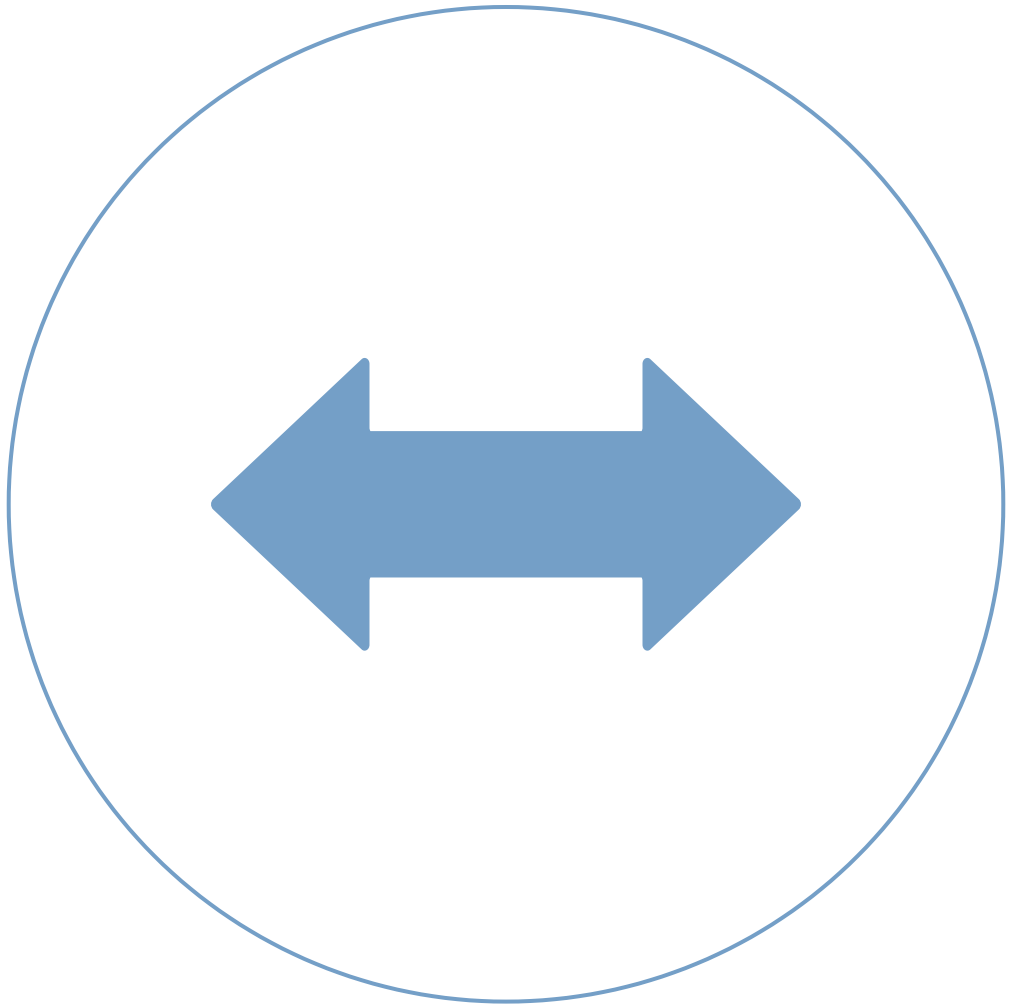
Steps of the LIME algorithm and workflow



Permutation

- Sampling from distribution curve or kernel density function
- Or by randomly leaving out words
- Predictions on every permutation with original (complex) machine learning model

Steps of the LIME algorithm and workflow



Distance & Similarity

- Distance measure for each permutation with original instance
- Converted to similarity metric
- E.g. cosine similarity (text), gower (default for tabular data), euclidean distance or any other

Steps of the LIME algorithm and workflow

Most important variables

- Choose number of features
- Forward selection & ridge regression fit
- Highest weights in ridge regression fit
- Lasso regularization
- Decision tree splits

Steps of the LIME algorithm and workflow

Simple local model

- Fit on output from machine learning model
- Output used as probability of each possible class
- Ridge regression (default in lime package)
- But principally, any model could be used

Applying LIME to our image classification model

Code: https://shirinsplayground.netlify.com/2018/06/keras_fruits_lime/

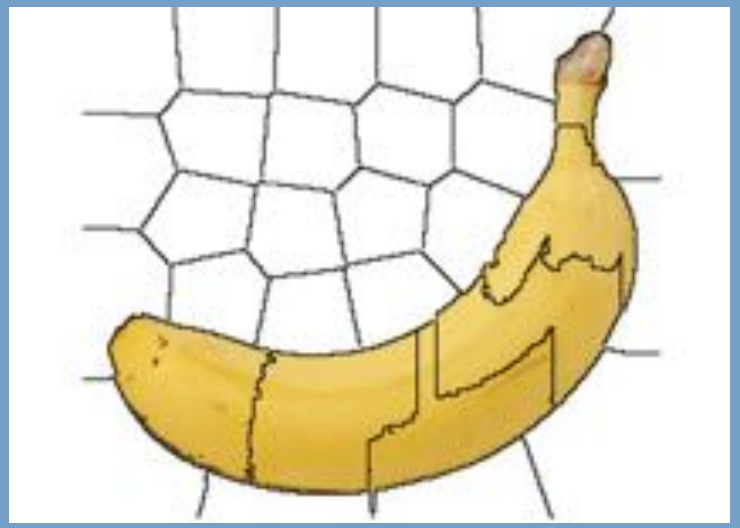
- **Explaining predictions of**

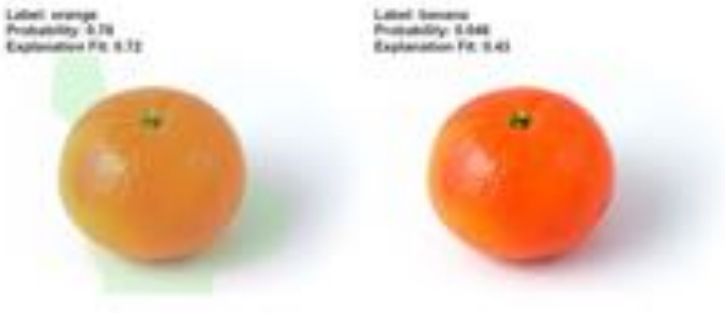
  - our Keras model

  - a pretrained Imagenet model (VGG16)

- **On two new images**

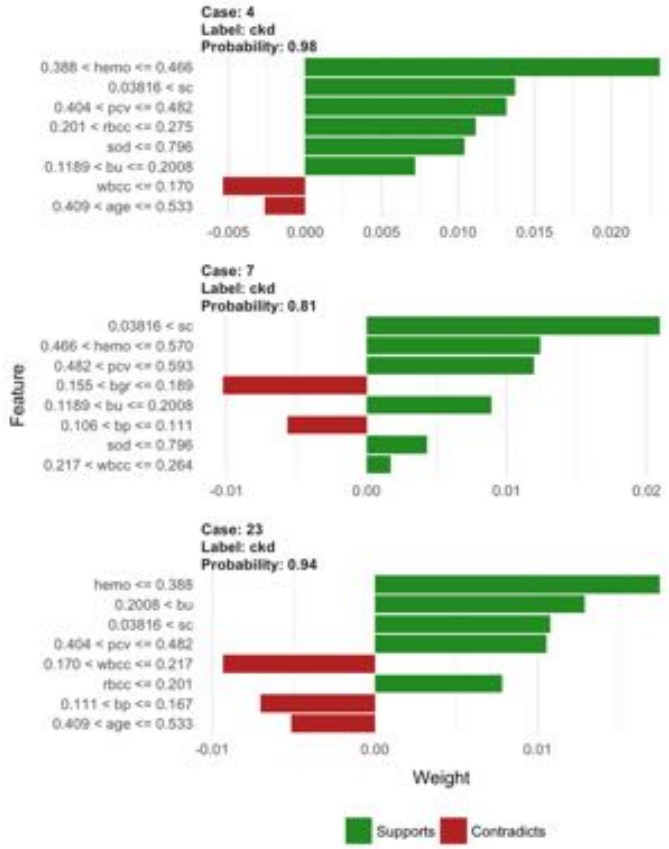
  downloaded from the internet





Explanations of tabular models

Chronic Kidney Disease (ckd) model trained with caret



https://shirinsplayground.netlify.com/2018/01/looking_beyond_accuracy_to_improve_trust_in_ml/

Explanations of text models

Women's clothing reviews model trained with xgboost

Problems and Pitfalls with LIME

- Uses linearly weighted combinations of features
- Depending on the goodness of fit linearity does not apply to new instances.
- Coverage is not clear

https://homes.cs.washington.edu/~marcotcr/aaai18.pdf

Anchors

If-then rules locally anchor predictions with sufficiently high probability.

“This movie is not bad.” “positive review” A={“not”, “bad”}, A(x)=1

Instance (x) → **Prediction on x (f(x))** → **Rule (A)**

Conditional distribution for when A applies (D & sample from D = z) → **Anchor IF A(x) = 1 AND IF A is a sufficient condition**

similar reviews w A = 1 high probability

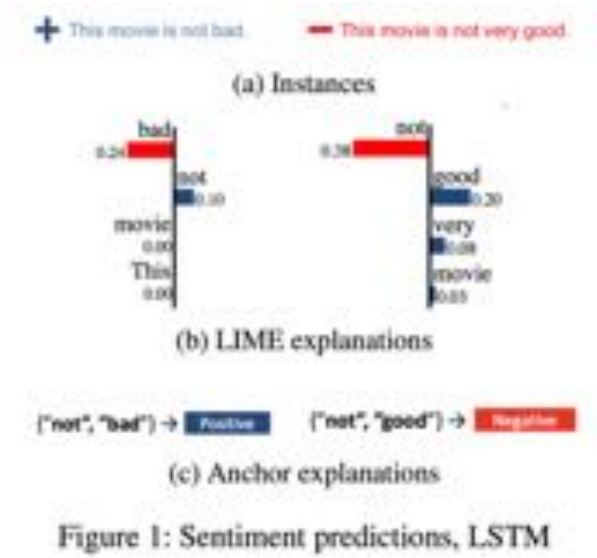


Figure 1: Sentiment predictions, LSTM

https://homes.cs.washington.edu/~marcotcr/aaai18.pdf & https://github.com/marcotcr/anchor

Advantages: Easy to understand & clear coverage!

Anchor explanations

Figures 3 & 5 and Tables 2 & 3 from Ribeiro, Singh and Guestrin 2018

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

English	Portuguese
This is the **question** we must address	Esta é a questão que temos que enfrentar
This is the **problem** we must address	Este é o problema que temos que enfrentar
This is what we must address	É isso que temos de enfrentar

Table 2: Anchors (in bold) of a machine translation system for the Portuguese word for "This" (in pink).

	If	Predict
adult	No capital gain or loss, never married	$\leq$ 50K
	Country is US, married, work hours $>$ 45	$>$ 50K
recid	No priors, no prison violations and crime not against property	Not rearrested
	Male, black, 1 to 5 priors, not married, and crime not against property	Re-arrested
lending	FICO score $\leq$ 649	Bad Loan
	649 $\leq$ FICO score $\leq$ 699 and \$5,400 $\leq$ loan amount $\leq$ \$10,000	Good Loan

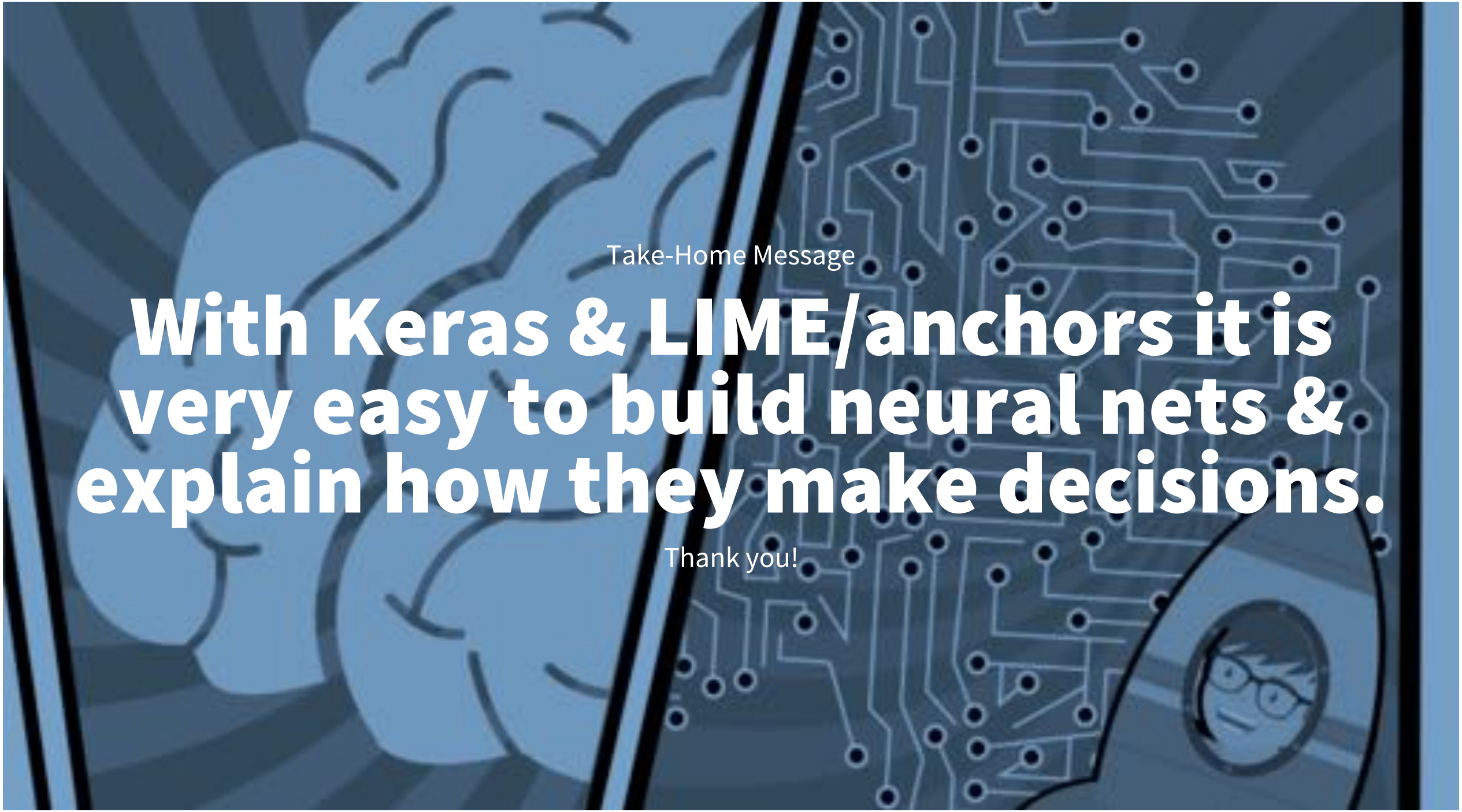
Table 3: Generated anchors for Tabular datasets

IF Country = United-States **AND** Capital Loss = Low
AND Race = White **AND** Relationship = Husband
AND Married **AND** 28 $<$ Age $\leq$ 37
AND Sex = Male **AND** High School grad
AND Occupation = Blue-Collar
THEN PREDICT Salary $>$ \$50K

https://homes.cs.washington.edu/~marcotcr/aaai18.pdf

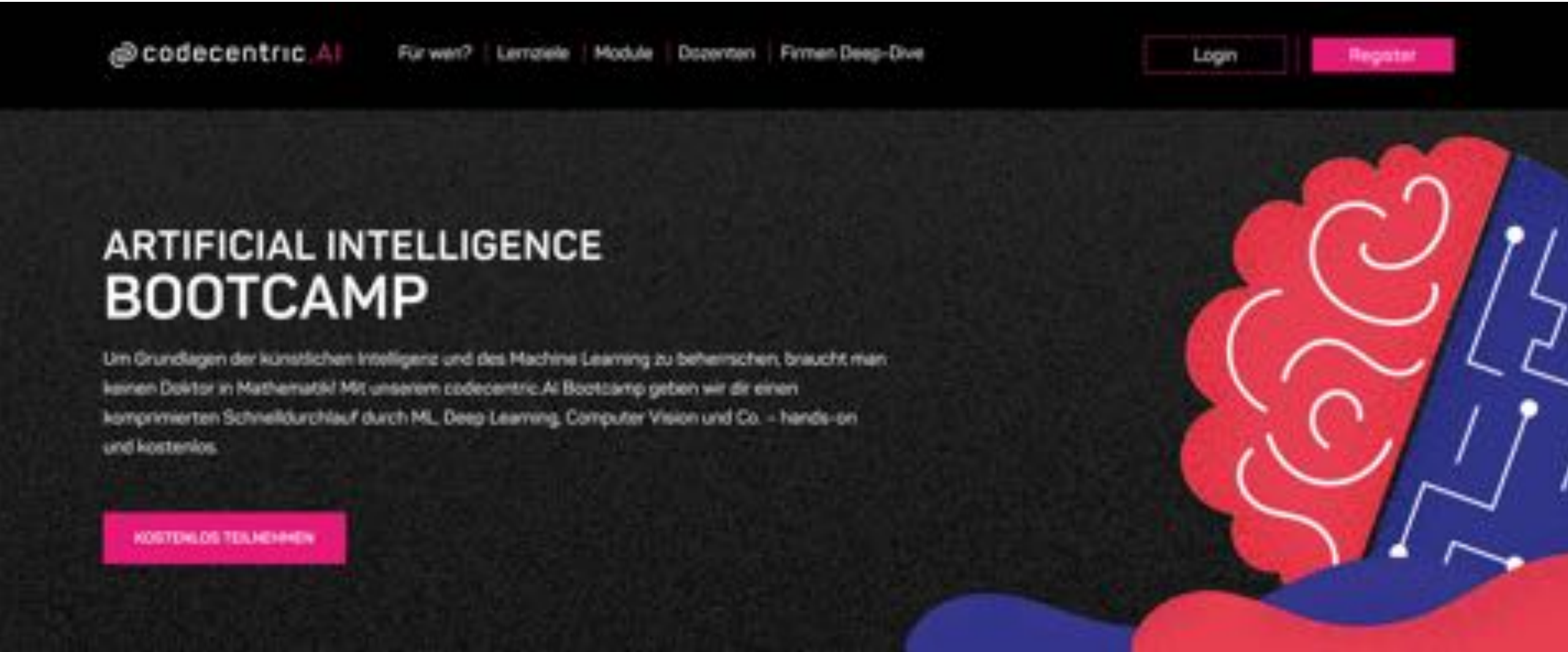
Take-Home Message

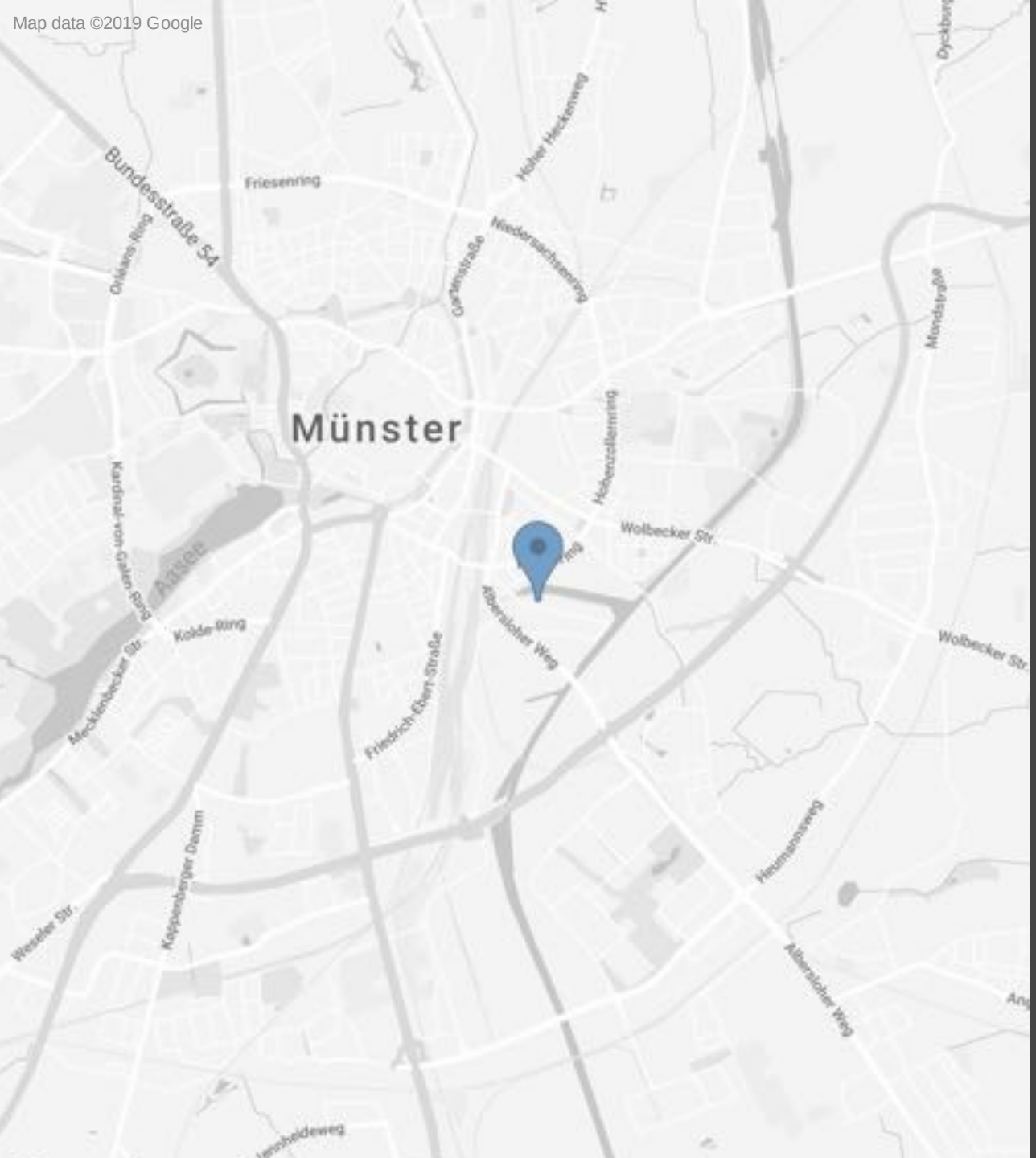
WE need to make sure that our models are not going to cause harm to people or society!


Take-Home Message
With Keras & LIME/anchors it is very easy to build neural nets & explain how they make decisions.
Thank you!

codecentric.AI Bootcamp

https://bootcamp.codecentric.ai/





Thank you! And stay connected

Am Mittelhafen 14, Münster

ShirinGlander

shirin.glander@codecentric.de

www.shirin-glander.de

Additional information

https://youtu.be/RcUdUZf8_SU

https://youtu.be/CY3t11vuuOM

https://github.com/marcotcr/lime

https://github.com/thomasp85/lime

https://github.com/marcotcr/anchor