Interpreting the Black Box: Explainable Neural Network Models

Marek Rei

UNIVERSITY OF CAMBRIDGE

Interpreting the Black Box

01 Introduction to interpretability

02 Motivation for explainable architectures

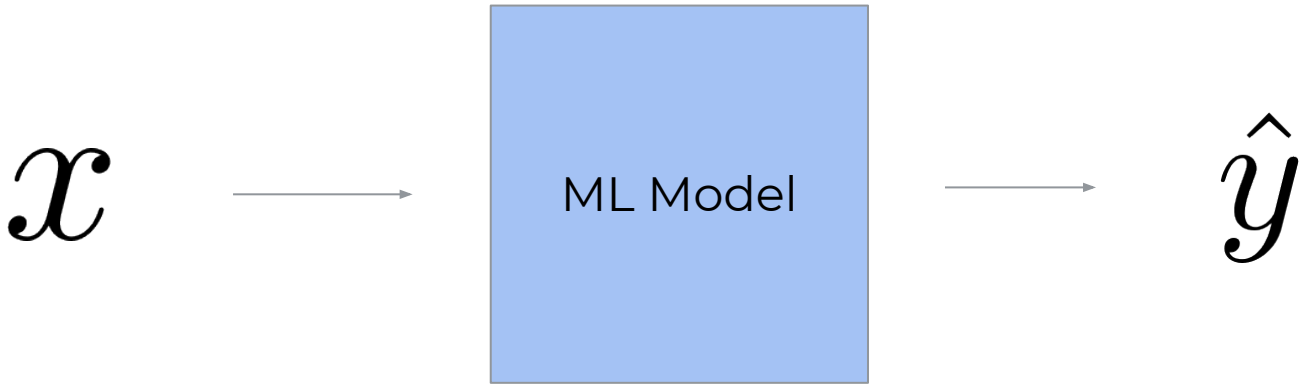
03 Methods of interpreting neural networks

04 Course logistics

Interpretability

Machine Learning Models

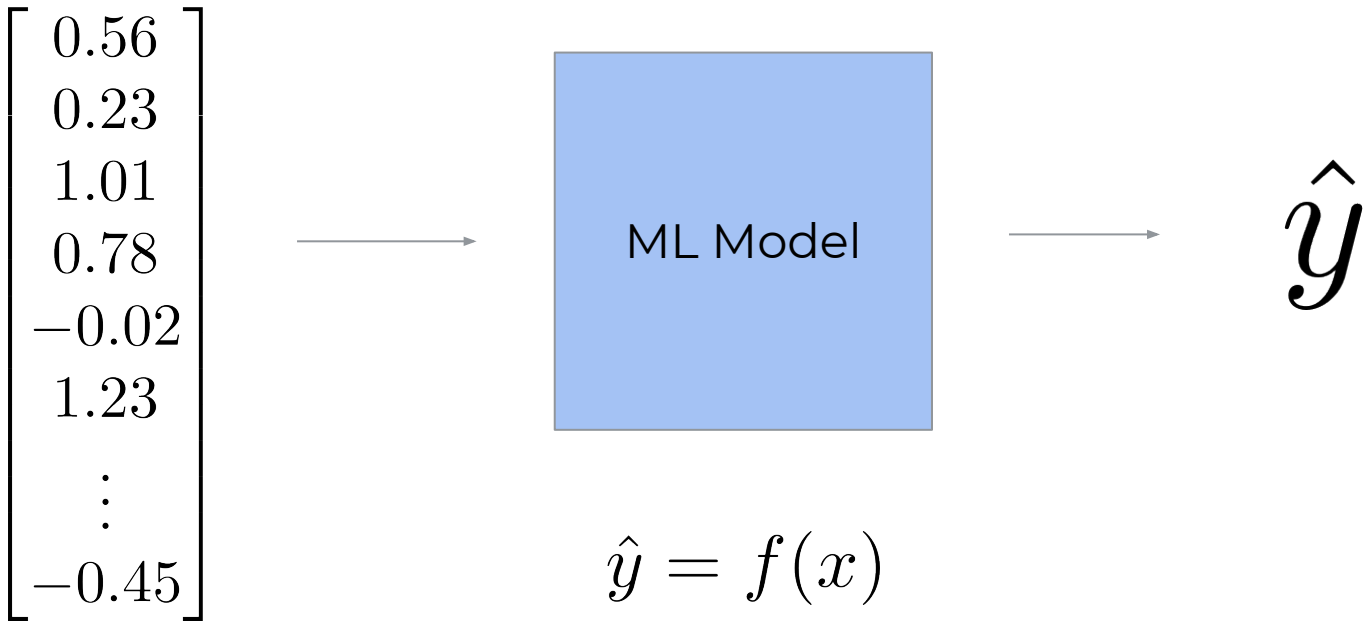
ML models take some **input x and transform it to some output y**, such as a predicted label.



$$\hat{y} = f(x)$$

Machine Learning Models

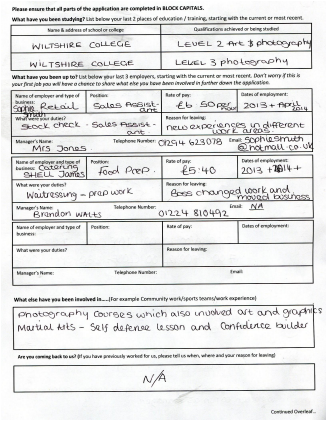
ML models take some **input x and transform it to some output y**, such as a predicted label.



Machine Learning Models

For example, we could train a classifier to predict whether to accept or deny an insurance claim.

If our claim gets denied, we would want to know why.



Please ensure that all parts of the application are completed in BLOCK CAPITALS.

What have you been studying? List below your last 2 places of education / training, starting with the current or most recent.

Name & address of school or college:	Qualifications achieved or being studied:
WILTSHIRE COLLEGE	LEVEL 2 Art & photography
WILTSHIRE COLLEGE	LEVEL 3 photography

What have you been up to? List below your last 3 employers, starting with the current or most recent. Don't worry if this is your first job you will have a chance to share what else you have been involved in further down the application.

Name of employer and type of business:	Position:	Rate of pay:	Dates of employment:
SAINSBURY'S RETAIL	Sales Assistant	£6.50 per hour	2013 + April 2014

What were your duties? Stock check - Sales Assistant

Reason for leaving: need experience in different work areas

Manager's Name: MRS Jones Telephone Number: 01254 623078 Email: SOPHIESMITH@hotmail.co.uk

Name of employer and type of business:	Position:	Rate of pay:	Dates of employment:
SHELL Jones	food Prep.	£5.40	2013 + 2014

What were your duties? Waitressing - prep work

Reason for leaving: does changed work and moved business

Manager's Name: Brandon Wallis Telephone Number: 01224 810492 Email: NA

Name of employer and type of business:	Position:	Rate of pay:	Dates of employment:

What were your duties?

Reason for leaving:

Manager's Name: Telephone Number: Email:

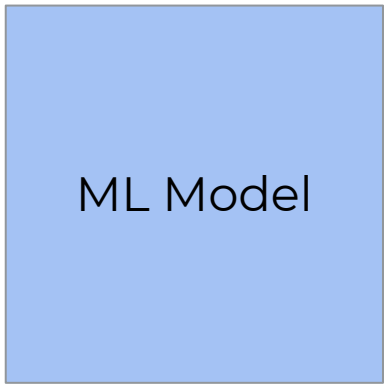
What else have you been involved in.....(for example Community work/sports teams/work experience)

Photography courses which also involved art and graphics
Martial Arts - Self defense lesson and Confidence builder

Are you coming back to us? (If you have previously worked for us, please tell us when, where and your reason for leaving)

N/A

Continued Overleaf...

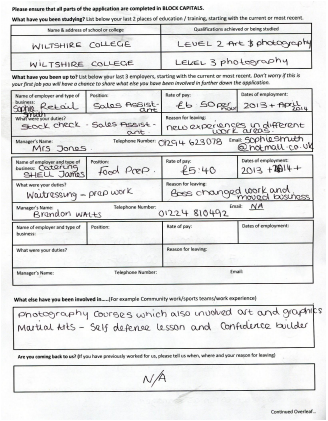


ACCEPT / DENY

Machine Learning Models

If we are using neural network models, interpretation can be very difficult.

Each decision is a combination of thousands of neurons and weights interacting with each other.



Please ensure that all parts of the application are completed in BLOCK CAPITALS.

What have you been studying? List below your last 2 places of education / training, starting with the current or most recent.

Name & address of school or college:	Qualifications achieved or being studied
WILTSHIRE COLLEGE	LEVEL 2 Art & photography
WILTSHIRE COLLEGE	LEVEL 3 photography

What have you been up to? List below your last 3 employers, starting with the current or most recent. Don't worry if this is your first job you will have a chance to share what else you have been involved in further down the application.

Name of employer and type of business:	Position:	Rate of pay:	Dates of employment:
SAVE Retail (Tesco)	Sales Assistant	£6.50ph	2013 + April 2014
What were your duties? Stock check - Sales Assistant	Reason for leaving: need experiences in different work areas		
Manager's Name: MRS Jones	Telephone Number: 01254 623078	Email: SOPHIESMITH@hotmail.co.uk	
Name of employer and type of business: Catering SHELL Jones	Position: Food Prep.	Rate of pay: £5.40	Dates of employment: 2013 + 2014
What were your duties? Waitressing - prep work	Reason for leaving: Boss changed work and moved business		
Manager's Name: Brandon Wallis	Telephone Number: 01224 810492	Email: NA	
Name of employer and type of business:	Position:	Rate of pay:	Dates of employment:
What were your duties?	Reason for leaving:		
Manager's Name:	Telephone Number:	Email:	

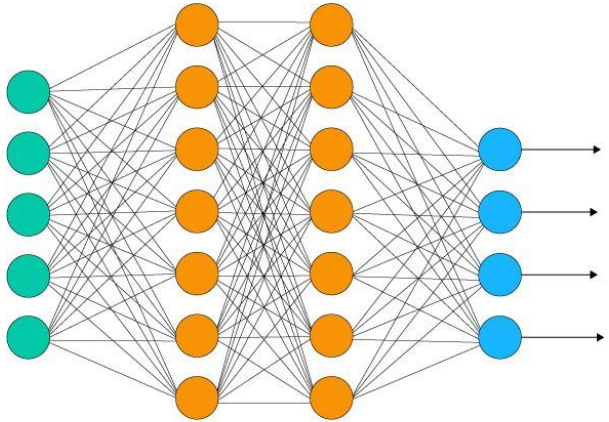
What else have you been involved in.... (for example Community work/sports teams/work experience)

Photography courses which also involved art and graphics
Martial Arts - Self defense lesson and Confidence builder

Are you coming back to us? (If you have previously worked for us, please tell us when, where and your reason for leaving)

N/A

Continued Overleaf...

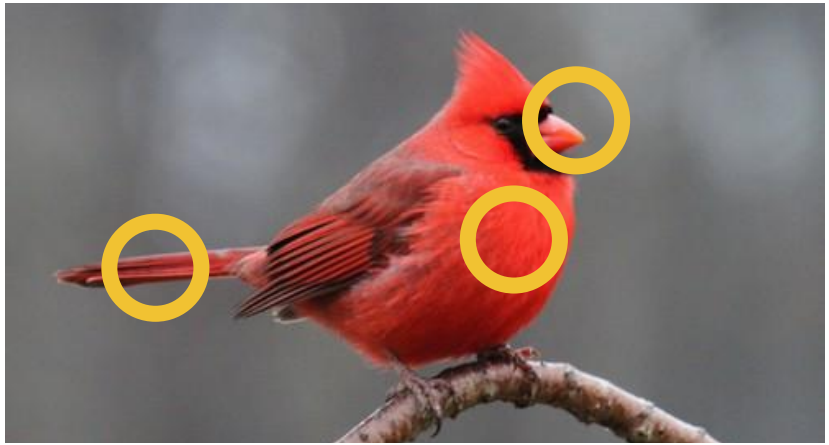


ACCEPT / DENY

Types of Interpretability

1. Why did the model make a specific decision about this datapoint?

   Regardless of whether the prediction is correct, we want to know the reasons!



Types of Interpretability

2. Why does the datapoint have a specific gold label?

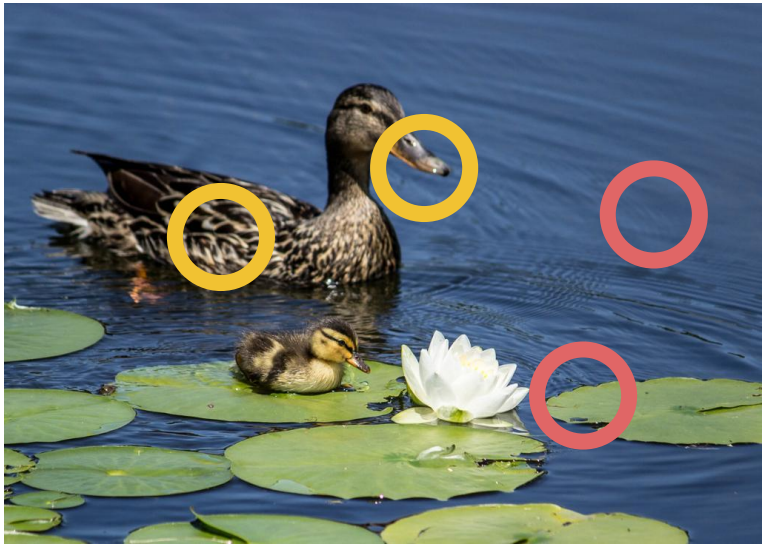
The same as previous ONLY if:

1) the model prediction is correct

AND

2) the prediction is made based on the correct reasons!

Can be difficult to keep separate!



Types of Interpretability

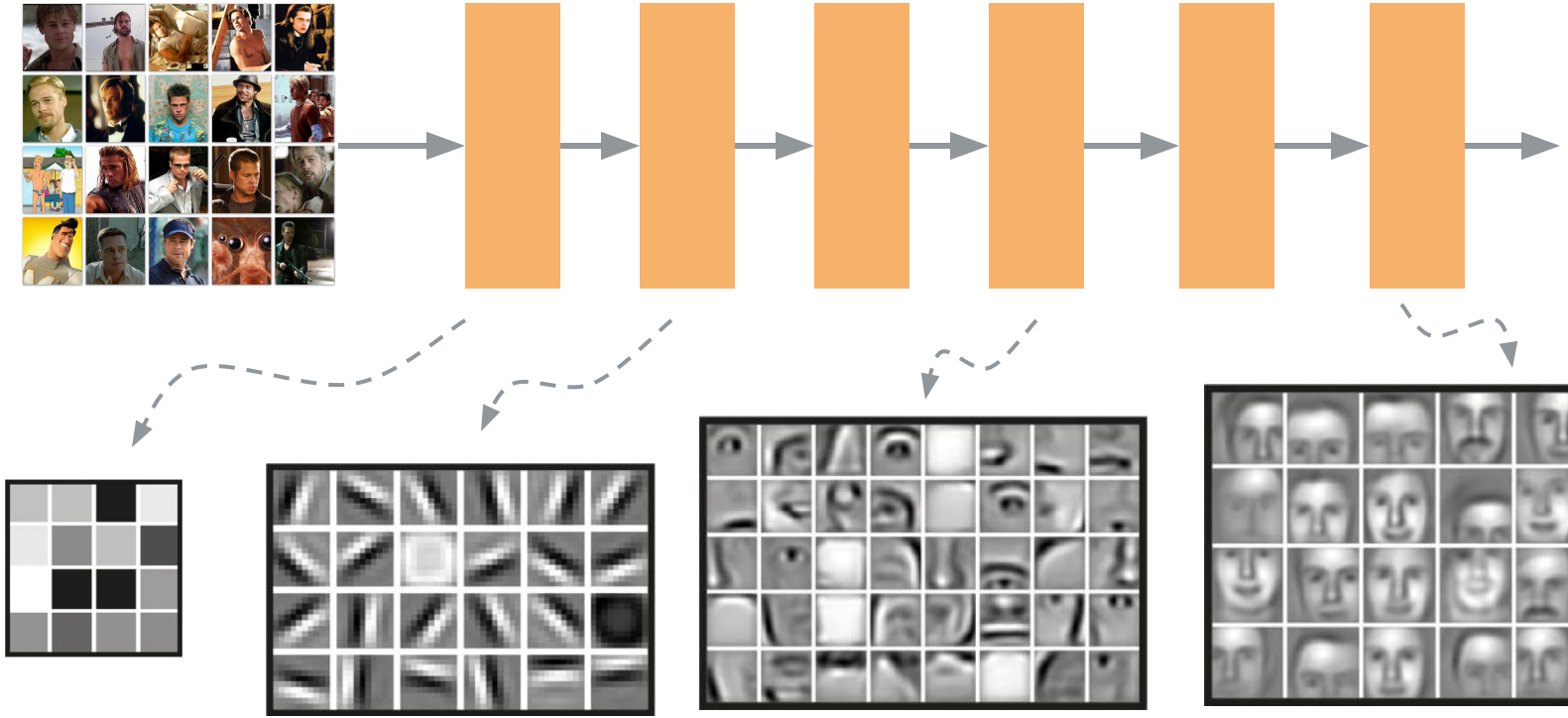
3. What has the model learned?

   Visualizing the internal weights and feature detectors of the model.

   Helps analyze the model as a whole

Types of Interpretability

3. What has the model learned?



Lee at al. (2009)

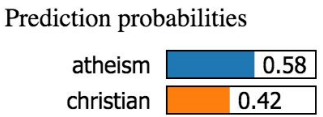
Motivation

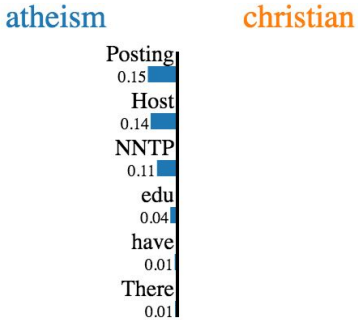
Model selection and analysis

- Understand whether our model is learning what we think it's learning
- Find why it makes mistakes and use that to improve the model
- Reveal hidden biases in the data

Model selection and analysis

- **Understand** whether our model is learning what we think it's learning
- **Find why it makes mistakes** and use that to improve the model
- **Reveal hidden biases** in the data





Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

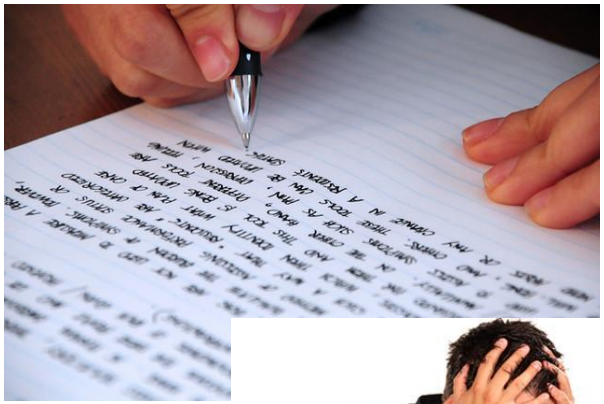
Hello Gang,

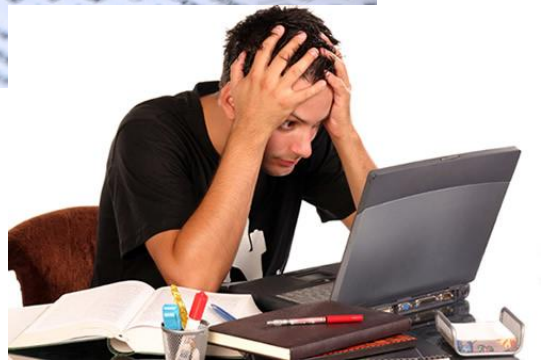
There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the
net. If anyone has a contact please post on the net or email me.

Ribeiro et al (2016)

Actionable information for the users

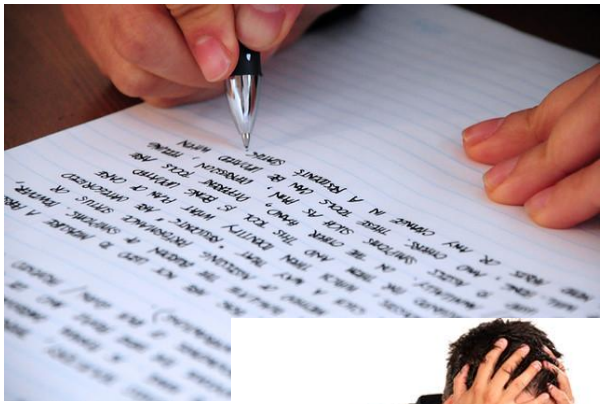
Automated essay scoring

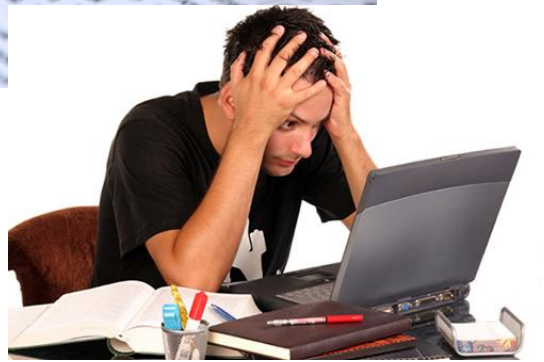




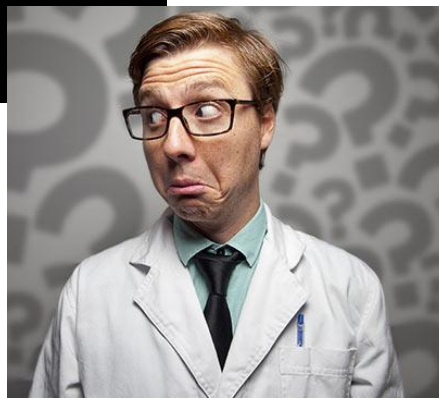
Actionable information for the users

Automated essay scoring



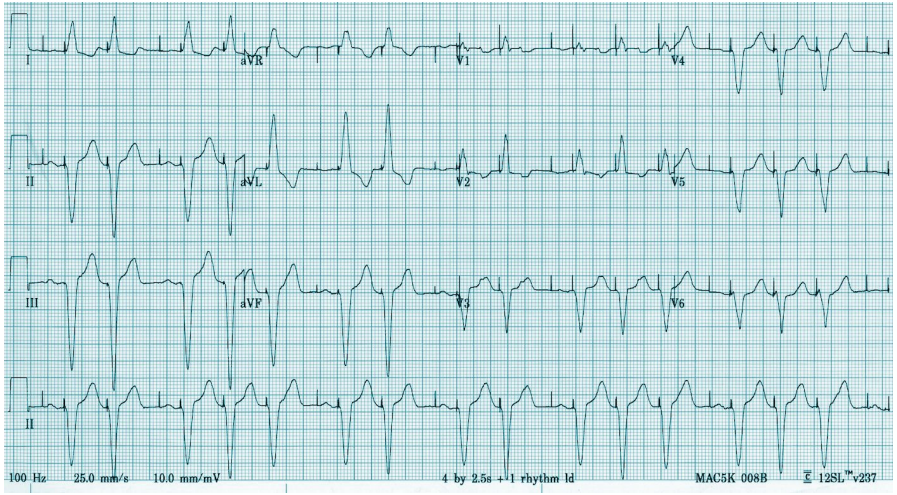


Medical imaging analysis



Learn about the data

- **Scientists want to know** how the model achieves high accuracy
- **Revealing novel informative features** in the data can advance our understanding of the field



Neuroscience
Biology
Cardiology
Sociology
Linguistics
Economics
...

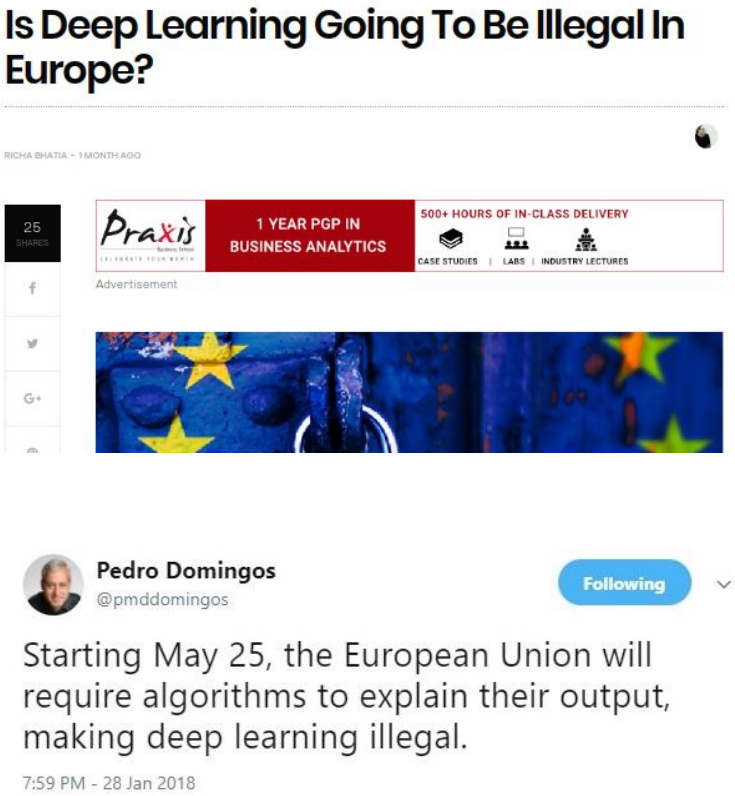
Ensuring transparency

The **General Data Protection Regulation** (GDPR) came into effect on 25 May 2018.

Requires machine learning algorithms to be able to **explain their decisions**.

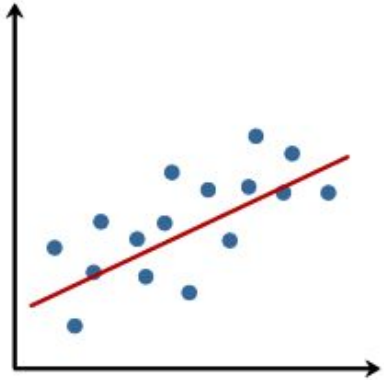
Replaces the former Data Protection Directive **across the EU, with global effects**.

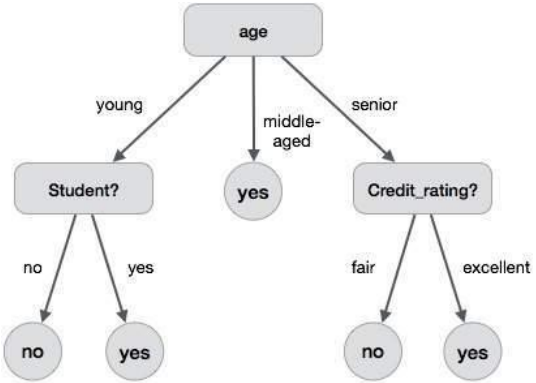
Maximum fine: 20M euros or 4% of global revenue, whichever is greater.



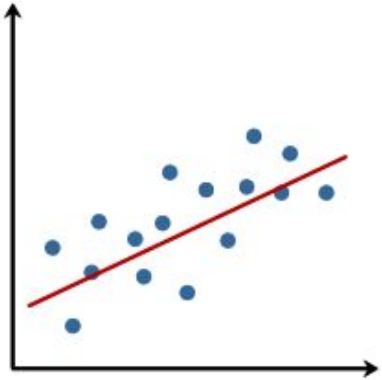
Methods

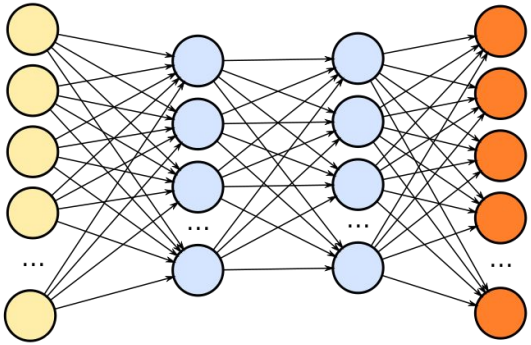
1. Use simple models

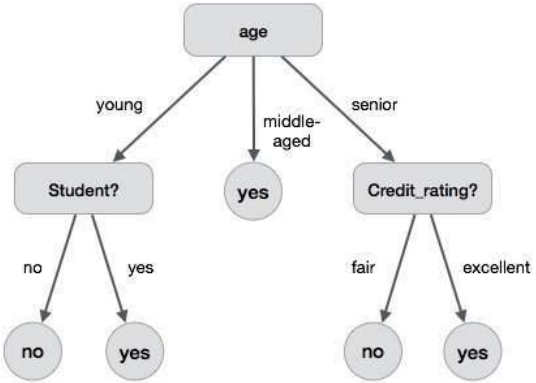


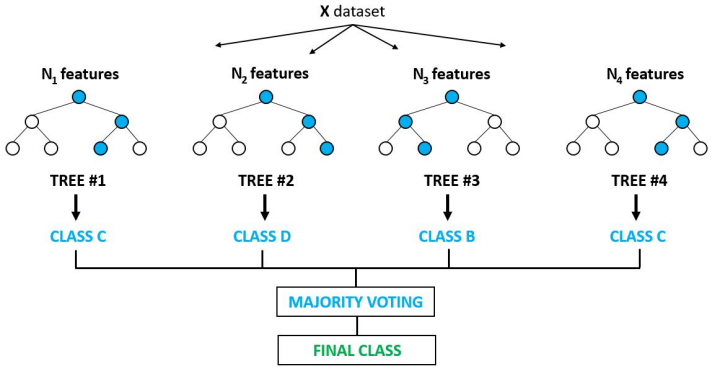


1. Use simple models

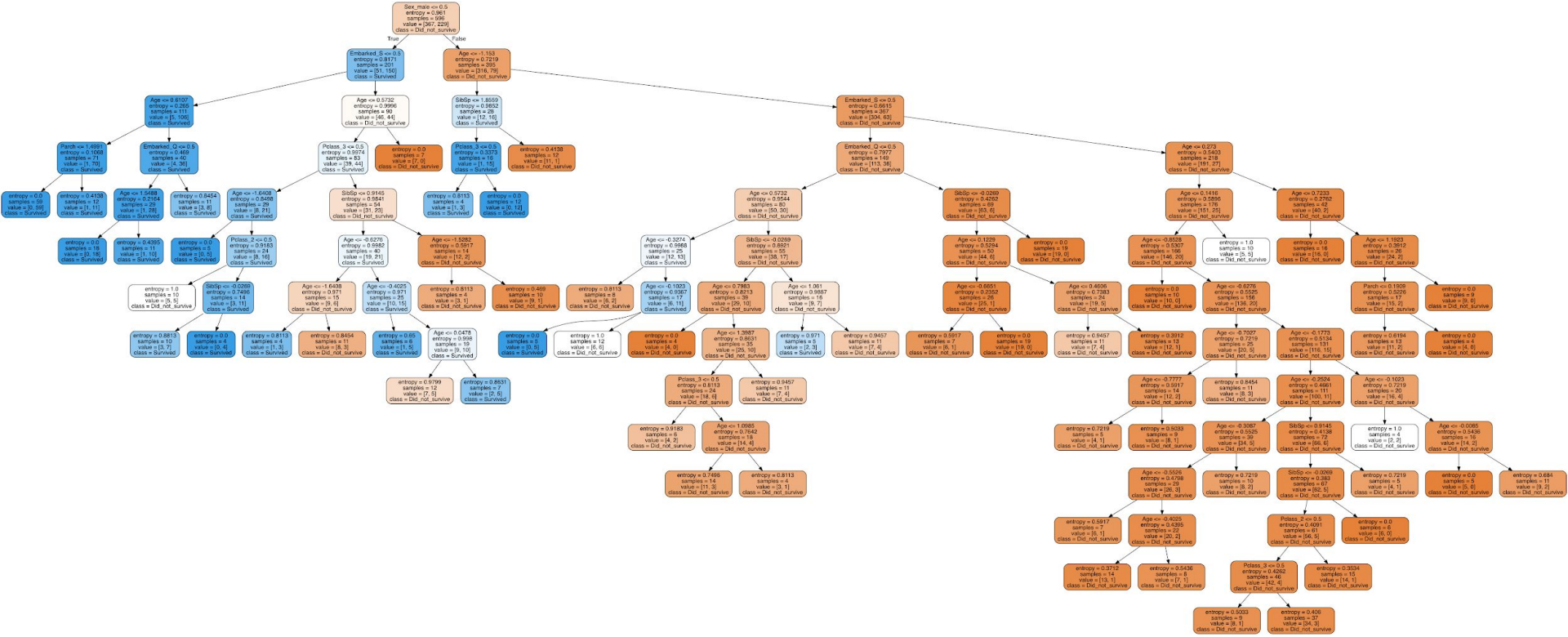








1. Use simple models



2. Measure the gradient

- **Calculate the gradient** of each input feature with respect to a certain label.
- **The magnitude of that gradient** shows the importance of the feature.

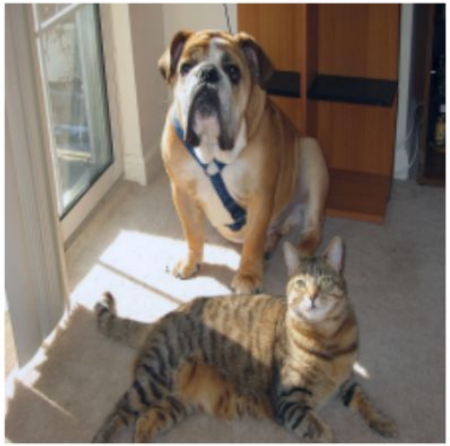
$$\hat{y} = f(x)$$

$$f'(x) = \lim_{\Delta x \to 0} \frac{(x + \Delta x) - x}{\Delta x}$$

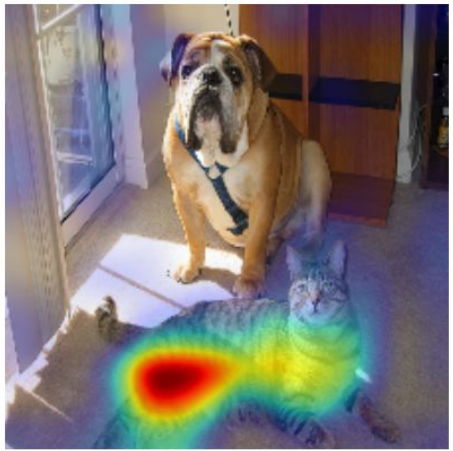
The derivative shows how much the prediction would change if we changed the input feature by a small amount.

2. Measure the gradient

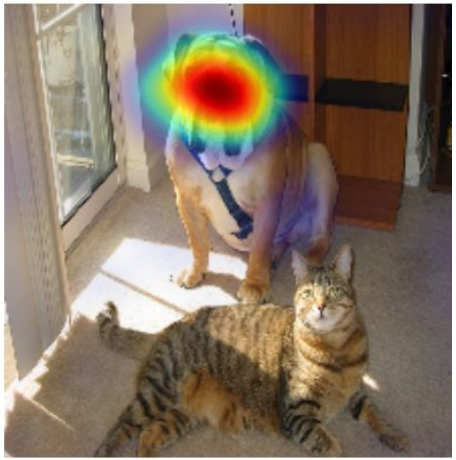
- **Calculate the gradient** of each input feature with respect to a certain label.
- **The magnitude of that gradient** shows the importance of the feature.



(a) Original Image



(c) Grad-CAM ‘Cat’

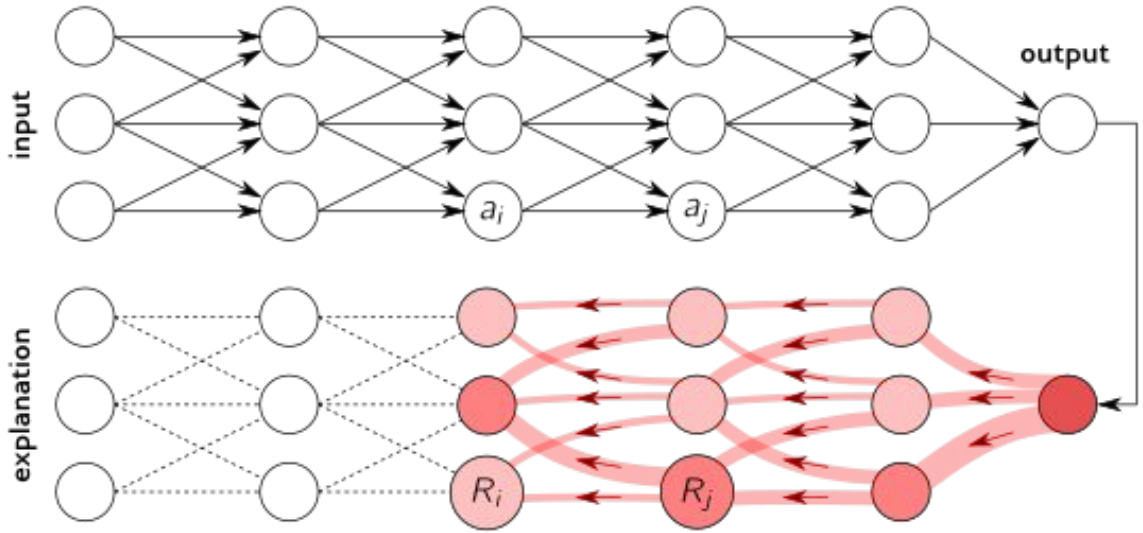


(i) Grad-CAM ‘Dog’

Selvaraju et al (2016)

3. Layer-wise relevance propagation

- Very similar method to the gradient calculation
- Redistributes the prediction confidence in the opposite direction, based on the connecting weights



$$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_k a_k w_{kj}^+} R_j$$

Bach et al. (2015)
http://www.heatmapping.org/

4. Explain using examples

- Show examples that the model finds most similar or useful
- Can essentially calculate the gradient with respect to each training image to find its usefulness

Test image

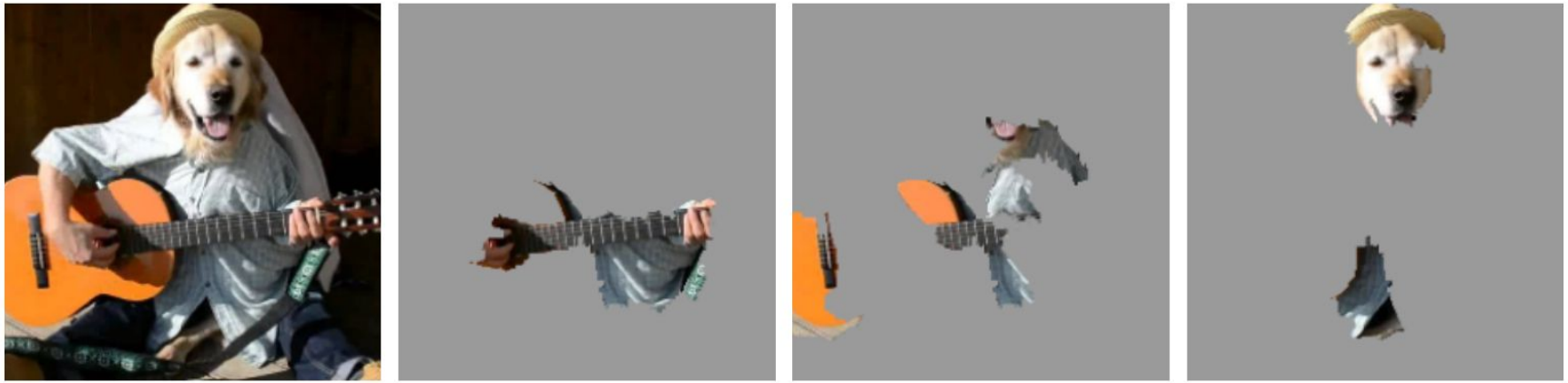
Most helpful images of the same class

Most helpful image of a different class

Koh & Liang, (2017)

5. Train surrogate models

- Train a **simple model** to predict the output of the complex model
- **Block out parts** of the input space for training and visualisation.



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Ribeiro et al (2016)

6. Generate explanations

- Can train a supervised component to generate explanations
- The output can easily end up explaining the data, as opposed to the model

This is a ***Baltimore Oriole*** *because* this is a small bird with a black head and orange body with black wings and tail.
This is a ***Cliff Swallow*** *because* this bird has a black crown a black throat and a white belly.
This is a ***Painted Bunting*** *because* this is a colorful bird with a red belly green head and a yellow throat.

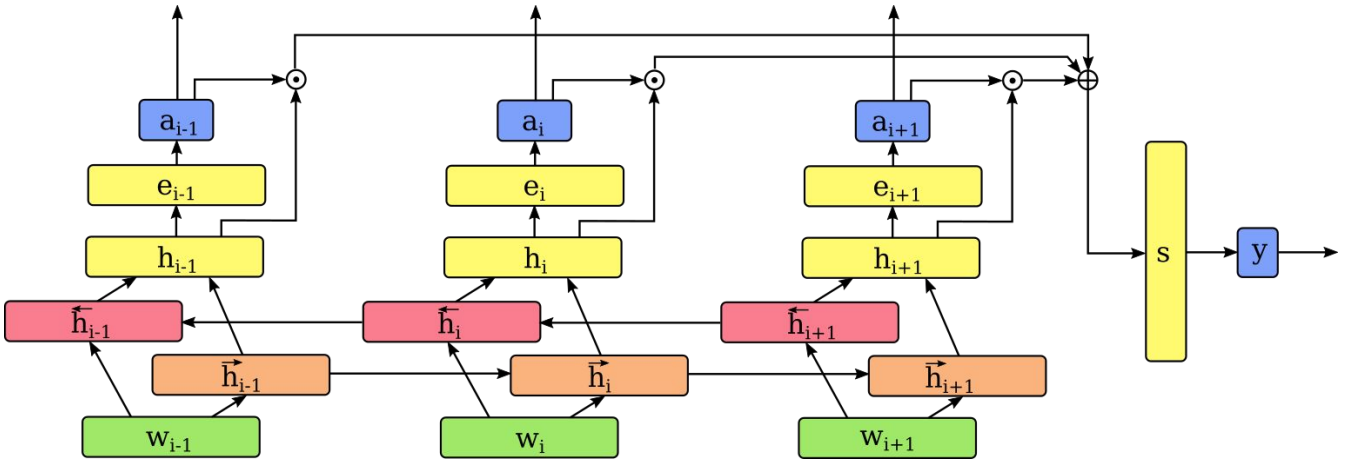
This is a ***Baltimore Oriole*** *because* this is a small bird with a black head and a small beak.
This is a ***Cliff Swallow*** *because* this bird has a black crown a brown wing and a white breast.
This is a ***Painted Bunting*** *because* this is a small bird with a red belly and a blue head.

This is a ***Baltimore Oriole*** *because* this is a small orange bird with a black head and a small orange beak.
This is a ***Cliff Swallow*** *because* this is a black bird with a red throat and a white belly.
This is a ***Painted Bunting*** *because* this is a colorful bird with a red belly green head and a yellow throat.

Hendricks et al (2016)

7. Design explainable models

- **Design the model to make finer-grained decisions**, which it uses to make the overall prediction.
- **Provides an inherent explainability** in the architecture.



Rei & Søgaard (2018)

7. Design explainable models

- **Design the model to make finer-grained decisions**, which it uses to make the overall prediction.
- **Provides an inherent explainability** in the architecture.

Rei & Søgaard (2018)

Course Logistics

Course structure

1 introductory lecture

3 sessions of paper presentations

6 papers to present, 15-20 min each

Everybody is expected to read all the papers and participate in the discussion

Assessment:

- Paper presentation (5%)
- Attendance and contribution to discussion (5%)
- Project report or essay (90%)

Course structure

Everybody does a project in one of their chosen topics.

This could be:

- Replication of a method from previous work
- Small novel experiment
- A survey of existing methods

Report length up to 5000 words.

Deadline for sending me a 500 word project proposal:
11 March

Deadline for submitting the report to the graduate education office:
24 April 16:00

Presentation schedule

Date	Presenter	Title
Monday 25 Feb	dk525	Why should i trust you?: Explaining the predictions of any classifier (*)
Monday 25 Feb	jbb50	Generating visual explanations
Wednesday 27 Feb	el476	Show, attend and tell: Neural image caption generation with visual attention
Wednesday 27 Feb	ktwo2	Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization
Monday 4 March	acs207	Explainable Prediction of Medical Codes from Clinical Text
Monday 4 March	jv401	Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement

Any questions?