# Responsible AI - Lecture 1

**TAIA - Advanced Topics on Artificial Intelligence**

Tiago Filipe Sousa Gonçalves

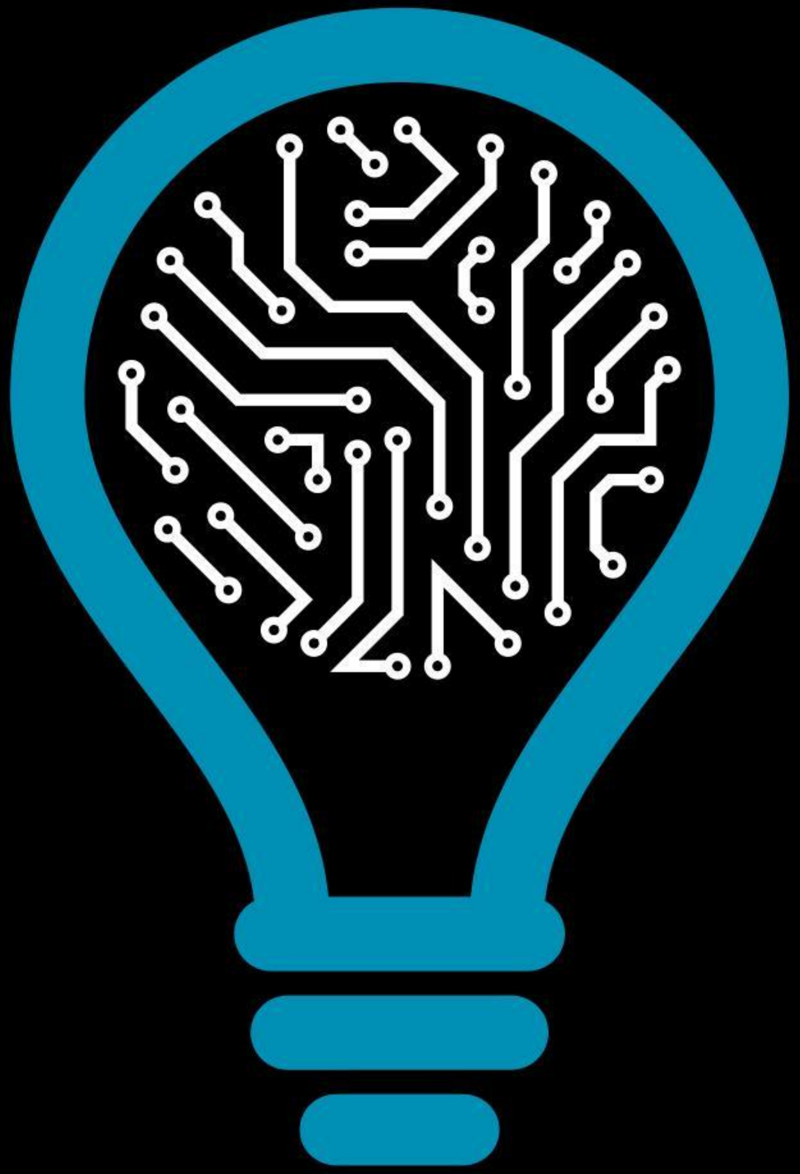
tiago.f.goncalves@inesctec.pt | tiagofs@fe.up.pt

Acknowledgement: Isabel Rio-Torto

isabel.riotorto@inesctec.pt

INESCTEC

INSTITUTE FOR SYSTEMS
AND COMPUTER ENGINEERING,
TECHNOLOGY AND SCIENCE



# Outline

1. Data is the New Black (gold)

2. With great power comes great responsibility

3. Enter the Matrices: can we unveil what neural networks are learning?

4. Take-home messages and further readings

# 1. Data is the New Black (gold)

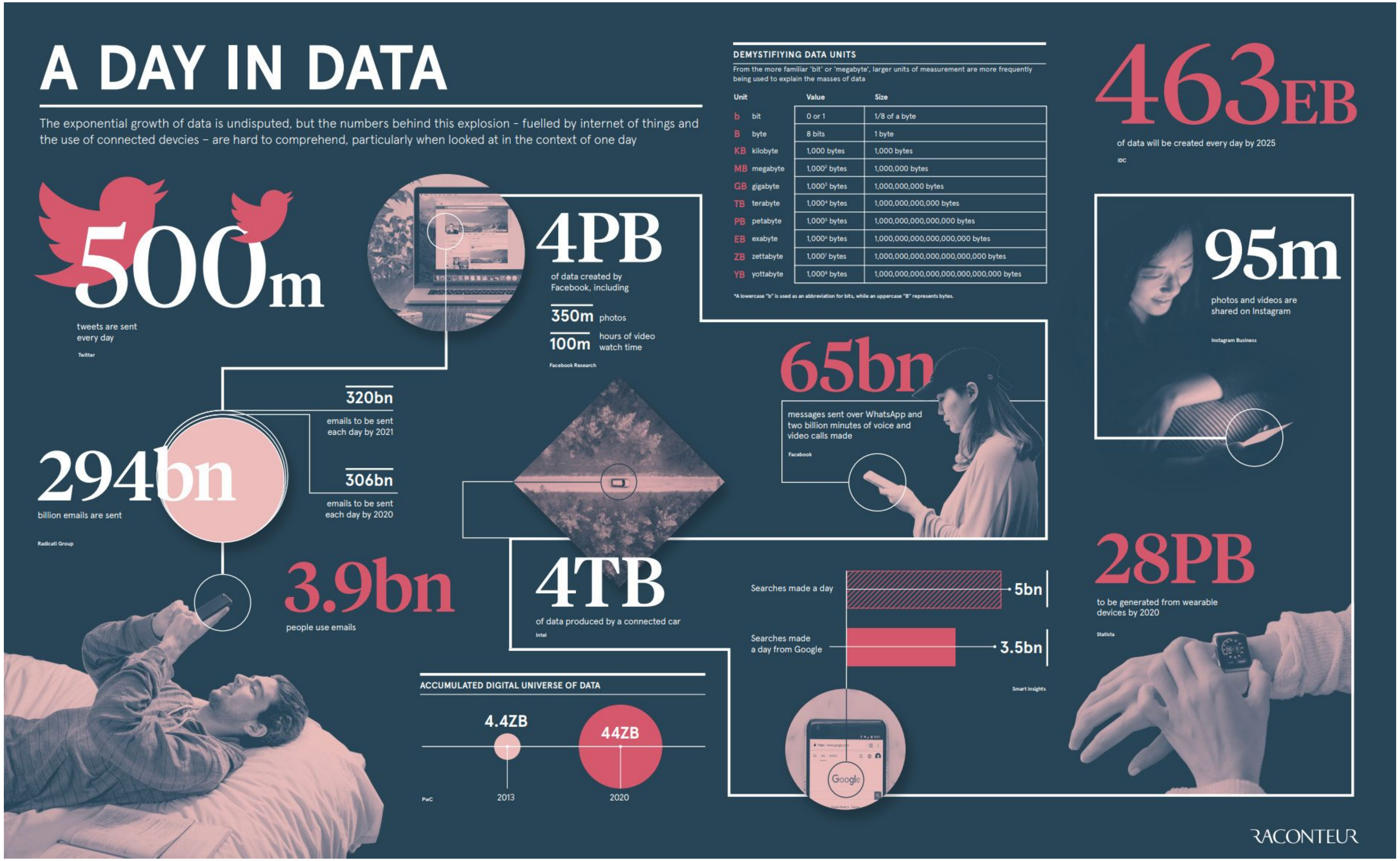
# Nowadays, we are constantly generating data[1]

- **The paradigm is changing:** most of the daily tasks and services can now be performed with the aid of **digital applications** or **gadgets**
- High-tech companies such as Google, Facebook, Netflix or Amazon **have access to huge amounts of data from several data sources and users**:
  - This phenomenon suggests that the *business of data* will become a **significant sector of the global economy**[2]
  - There are several **open-source data sets with millions of entries** (e.g., ImageNet[3])
- Data is referred as **the new oil**[4]
  - The main impact on humanity is related **to the way data can improve our lives**
  - **A proper management process of the "dark side" of data must be implemented**, but the **advances** in data fuels **are worth the effort**



**Sources: [1]** https://techjury.net/blog/big-data-statistics/#gref, **[2]** https://www.forbes.com/sites/gilpress/2020/01/06/6-predictions-about-data-in-2020-and-the-coming-decade/?sh=5cbaadf4fc36, **[3]** http://www.image-net.org, **[4]** https://www.forbes.com/sites/forbestechcouncil/2019/11/15/data-is-the-new-oil-and-thats-a-good-thing/?sh=69bb9a407304

# Take a look: A Day in the Wonderful World of Data[1, 2, 3]




Sources: [1] https://www.raconteur.net/infographics/a-day-in-data/, [2] https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/, [3] https://www.visualcapitalist.com

# We have more computational power than ever

- **The fundamental concepts of artificial intelligence and deep neural networks have been around since 1940**[1]
  - Frank Rosenblatt proposed one of the first approaches to the design and training of artificial neural networks: **the Perceptron**[2]

- The development of **powerful computer processing units (CPUs)** and the leveraging of the **graphical processing units (GPUs)**[3] for computation allowed the training of deep and complex algorithms in **"human time"**

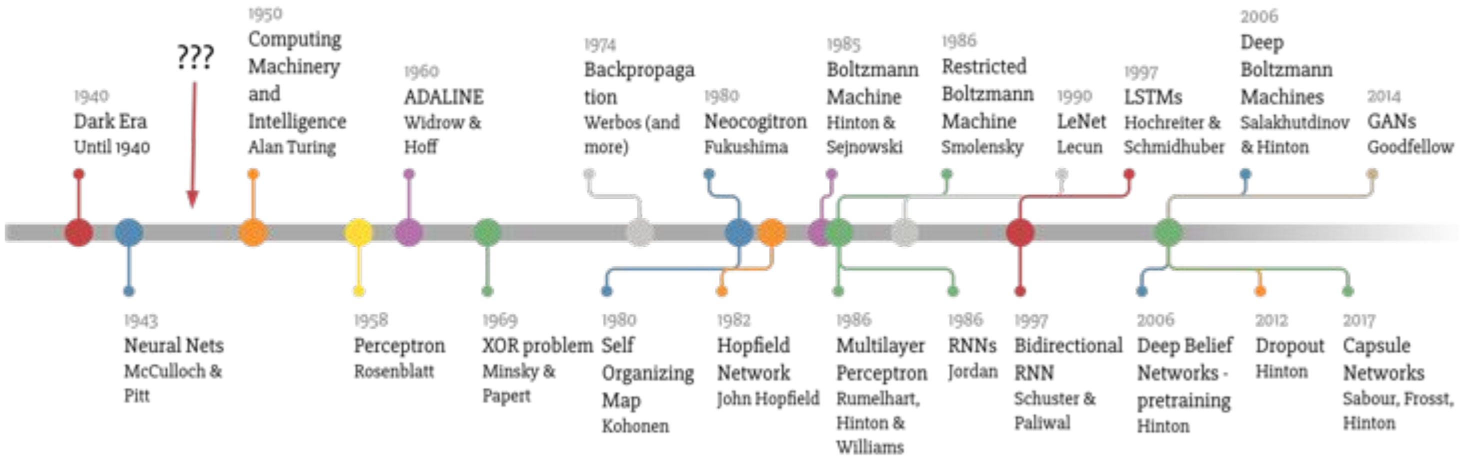



**Figure - A (tentative) deep learning timeline** (Image from [1])

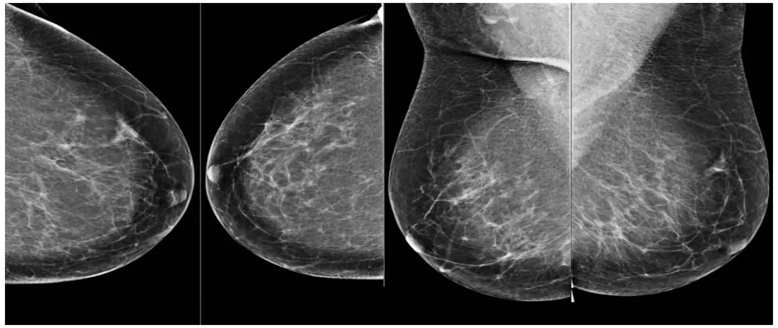
**Sources: [1]** https://towardsdatascience.com/a-weird-introduction-to-deep-learning-7828803693b0, **[2]** Frank Rosenblatt "The perceptron: A probabilistic model for information storage and organization in the brain.", **[3]** https://www.nvidia.com/en-us/

# Technology has been *challenging* human performance...

- There are, at least, two popular events that created a revolution in the History of AI:
  - **In 1997, IBM's Deep Blue beat the Chess World Champion Garry Kasparov**[1]
  - **In 2016, Google's DeepMind AlphaGo learn to play Go alone (i.e., through reinforcement learning policies) and beat the Go World Champion Lee Sedol**[2]

- The two events above are examples of the **(virtually) unlimited boundaries of the application of artificial intelligence** to our daily lives
  - In 2020, Google's DeepMind published a paper in ***Nature*** **suggesting that "its model was able to spot cancer in de-identified screening mammograms with fewer false positives and false negatives than experts"**[3, 4]

**Figure - Medical Image Analysis: Mammograms** (Image from [4])





Sources: [1] https://theconversation.com/twenty-years-on-from-deep-blue-vs-kasparov-how-a-chess-match-started-the-big-data-revolution-76882, [2] https://deepmind.com/research/case-studies/alphago-the-story-so-far, [3] https://www.healthcareitnews.com/news/google-ai-platform-aids-oncologists-breast-cancer-screenings, [4] https://www.nature.com/articles/s41586-019-1799-6

# 2. With great power comes great responsibility

# Do we still remember the “good” old times of AI?

- In the beginning, **artificial intelligence systems were based in algorithms**:
  - An algorithm is a **set of instructions** that the system will follow to **achieve a certain goal** (direct programming)[1]
  - These **explicit** rules were often based on **domain knowledge**
  - Hence, they were “easy” to **explain** and to **understand**

- Nowadays, we use the available data to automatically learn **programs/functions**:
  - In machine learning, we **learn from data and make predictions** (indirect programming)[1]
  - These algorithms work by **optimising an objective function**
  - Hence, the “rules” often are **implicit** and **difficult to understand**

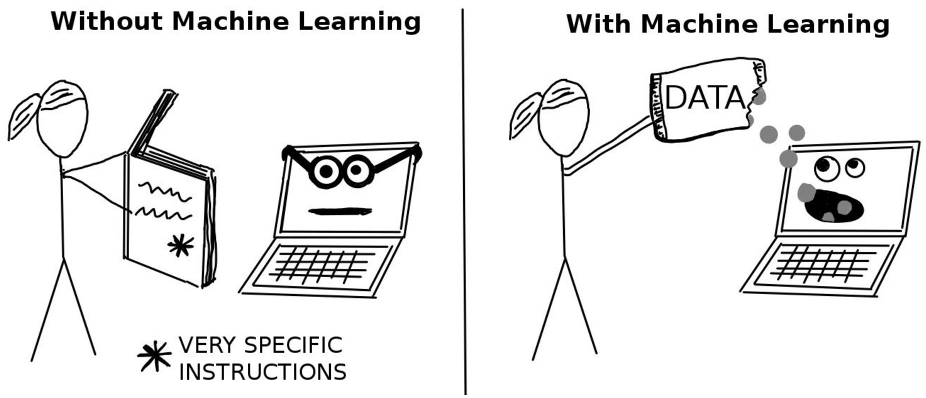



**Figure - Algorithms vs Machine Learning** (Image from [1])

**Sources: [1]** Christoph Molnar “Interpretable Machine Learning A Guide for Making Black Box Models Explainable”

# Deep learning *versus* traditional machine learning[1]

- **Traditional machine learning** required **experts to extract meaningful features** (*i.e.*, domain-specific features) from raw data and feed them into machine learning algorithms to obtain classification/regression models:

- **Deep learning** "only" requires **raw data and labels** to achieve high-performing models, since it **automatically extracts the patterns**
  - Deep learning algorithms are suitable for **representation learning**, i.e., finding the **best representation of the data** that optimises a given optimisation objective

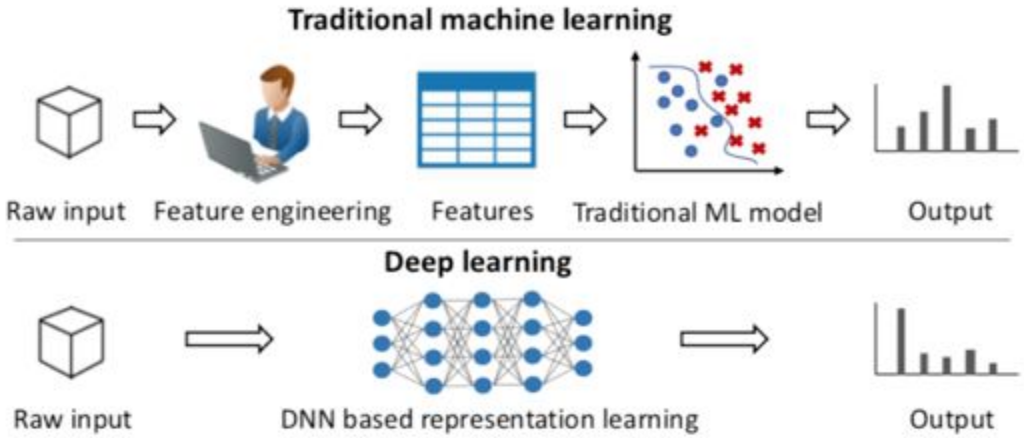



**Figure - Deep learning vs traditional machine learning** (Image from [2])

Sources: [1] Lecun et al. "Deep learning", [2] https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

# Do we understand the features learned by these models?

- Even if the models achieve high performances, **it is not trivial to assure that they are learning features that are relevant for that domain** (i.e., **black box** behaviour)
  - Machine learning models **are good at extracting correlations**

- While this **may not be an issue in several domains** (e.g., recommendation systems), in others, it is of utmost importance that the **system is capable of transparently showing** the reasons behind its decisions (e.g., healthcare)

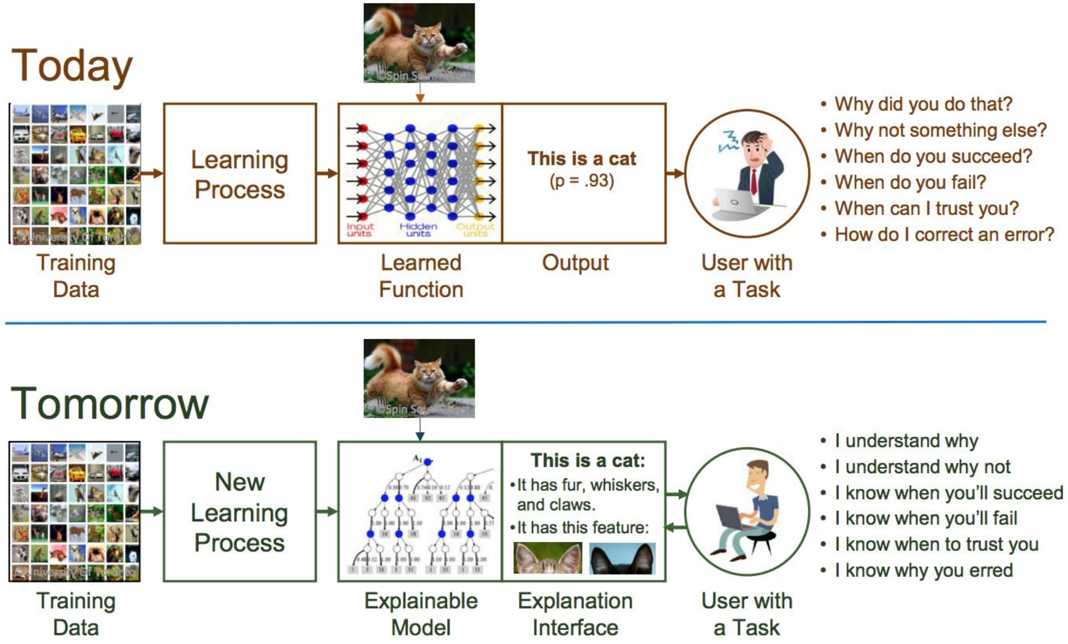



**Figure - The future of machine learning algorithms**
(Image from [1])

**Sources: [1]** https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

# "Who you gonna call?" Responsible AI!

- **Responsible AI** is a framework that guides how we should address the challenges around artificial intelligence from both an **ethical, technical and legal** point of view[1]
  - We must resolve ambiguity for where responsibility lies if something goes wrong!

- This framework relies on fundamental principles[2]:
  - Accountability
  - **Interpretability**
  - Fairness
  - Safety
  - Privacy

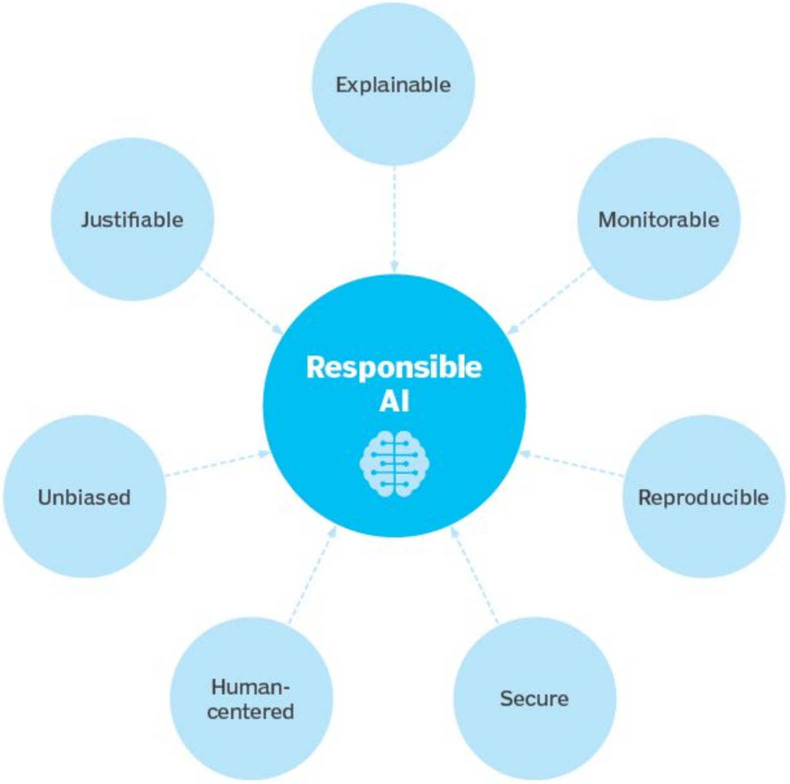



**Figure - Responsible AI** (Image from [1])

Sources: [1] https://searchenterpriseai.techtarget.com/definition/responsible-AI, [2] https://towardsdatascience.com/what-is-responsible-ai-548743369729

# Explain it like a Human: Interpretability is the key!

- **Interpretability** is a concept that results from the interaction between several definitions
  - The degree to which a human can **understand the cause of a decision**[1]
  - The degree to which a human can **consistently predict the model's result**[2]

- **Interpretable machine learning** is also related to the **"extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model"**[3]

- Intuitively, the **higher the degree of interpretability** of a model, the **higher the likelihood of a user comprehending its predictions**[4]

- **"Humans have a mental model of their environment that is updated when something unexpected happens. This update is performed by finding an explanation for the unexpected event"**[4]



**Sources: [1]** Tim Miller "Explanation in artificial intelligence: Insights from the social sciences.", **[2]** Been Kim et al. "Examples are not enough, learn to criticize! Criticism for interpretability.",
**[3]** W. J. Murdoch "Definitions, methods, and applications in interpretable machine learning.", **[4]** Christoph Molnar "Interpretable Machine Learning A Guide for Making Black Box Models Explainable"

# Should we care about Interpretability?

- Why is it important to care about the **inner functioning** of the machine learning models? If a machine learning model attains **good performance**, why not just trust the model and ignore why it made a certain decision?[1]
  - "The problem is that a single metric, such as classification accuracy, is an **incomplete description of most real-world** tasks."[2]

- If one intends to deploy these models into real-world applications, they must be able to **explain their predictions in a human-understandable way**[3]
  - There is an **inherent tension between machine learning performance** and **explainability**: usually, the **best-performing methods** are the **least transparent**, and the ones providing a **clear explanation** are **less accurate**[4]
  - **Law and policy stakeholders** require AI to be transparent, fair and trustworthy[5]

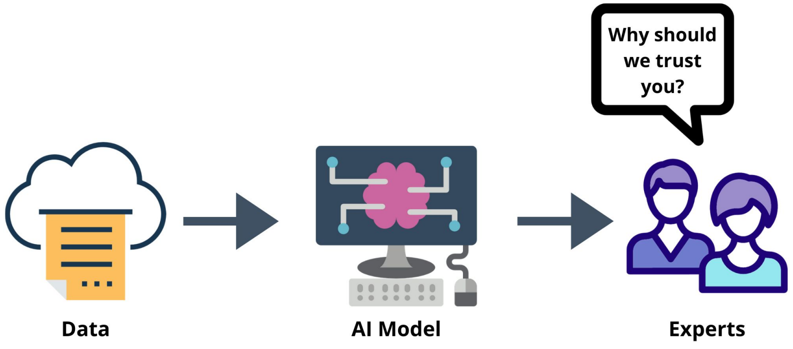



**Figure - Trustworthy AI** (Image from [6])



Sources: [1] Christoph Molnar "Interpretable Machine Learning A Guide for Making Black Box Models Explainable", [2] Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning.", [3] Cynthia Rudin "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead", [4] Holzinger et al. "What do we need to build explainable AI systems for the medical domain?", [5] Kaminsky "The Right to Explanation, Explained", [6] https://inesctec.medium.com/explainable-artificial-intelligence-unveiling-what-machines-are-learning-91b96a63a07a

# 3. Enter the Matrices: can we unveil what neural networks are learning?

# Delving deeper into the field of explainable AI (xAI)

- **Explainability and interpretability definitions are often used interchangeably** (i.e., there is no clear distinction between these two terms)[1]

- xAI can be seen as a **three stage process**:
  - Pre-Model
  - In-Model
  - Post-Model

| Pre-Model | In-Model | Post-Model |
| --- | --- | --- |
| (Aim to understand the data before building the model) | (Seek to integrate interpretability inside the model) | (Perform posterior analysis of the model predictions) |

Sources: [1] Lipton "The Mythos of Model Interpretability", [2] Doshi-Velez and Kim "Towards A Rigorous Science of Interpretable Machine Learning"

# Pre-model methods rely on data *exploratory analysis*[1]

- **We aim to perform an analysis of the data distribution**
  - This comprehension of the data may contribute to **higher confidence** with the **posterior decisions** that a model can provide

- One may think of “K-Means Clustering”, “K-Nearest Neighbours” and, more recently, **“Prototypes & Criticisms (MMD-critic framework)”**[2]


**Figure - MMD-critic framework** (Images from [6])

Sources: [1] Tukey “Exploratory data analysis”, [2] Kim et al.“Examples are not Enough, Learn to Criticize! Criticism for Interpretability”

# Post-model: the posterior analysis of model predictions

- In computer vision, one may think of methods based on **“Gradients”, “Decomposition”, “Optimisation” and “Deconvolution”**[1, 2]

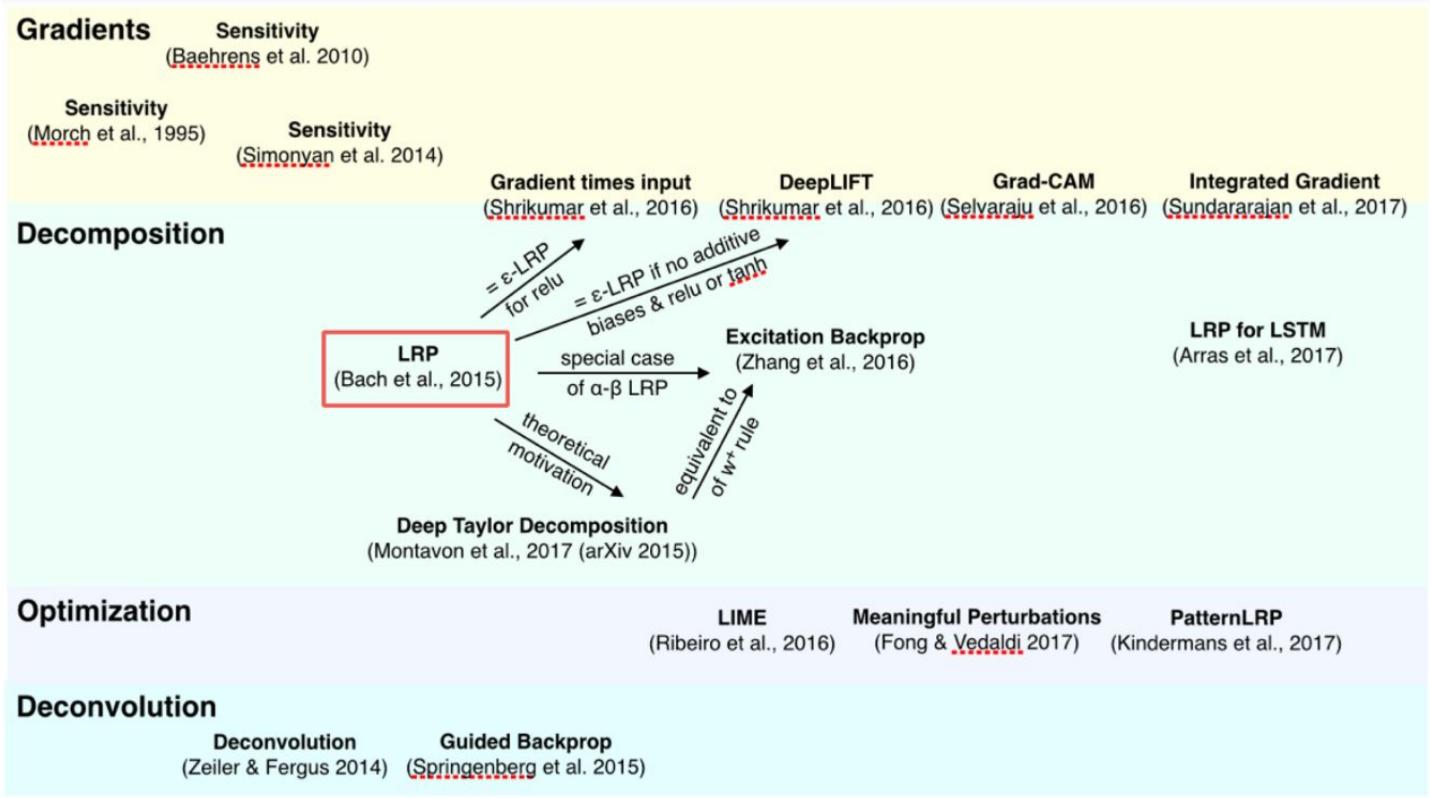



**Figure - Post-model methods for computer vision** (Image from [1])

Sources: [1] Samek “Interpreting Deep Neural Networks and their Predictions”, [2] Alber et al. “iNNvestigate Neural Networks!”

# Post-model: are we really visualising models?

- Each post-model method has intrinsic properties, hence, it is of utmost importance **that we understand what we want to visualise**!

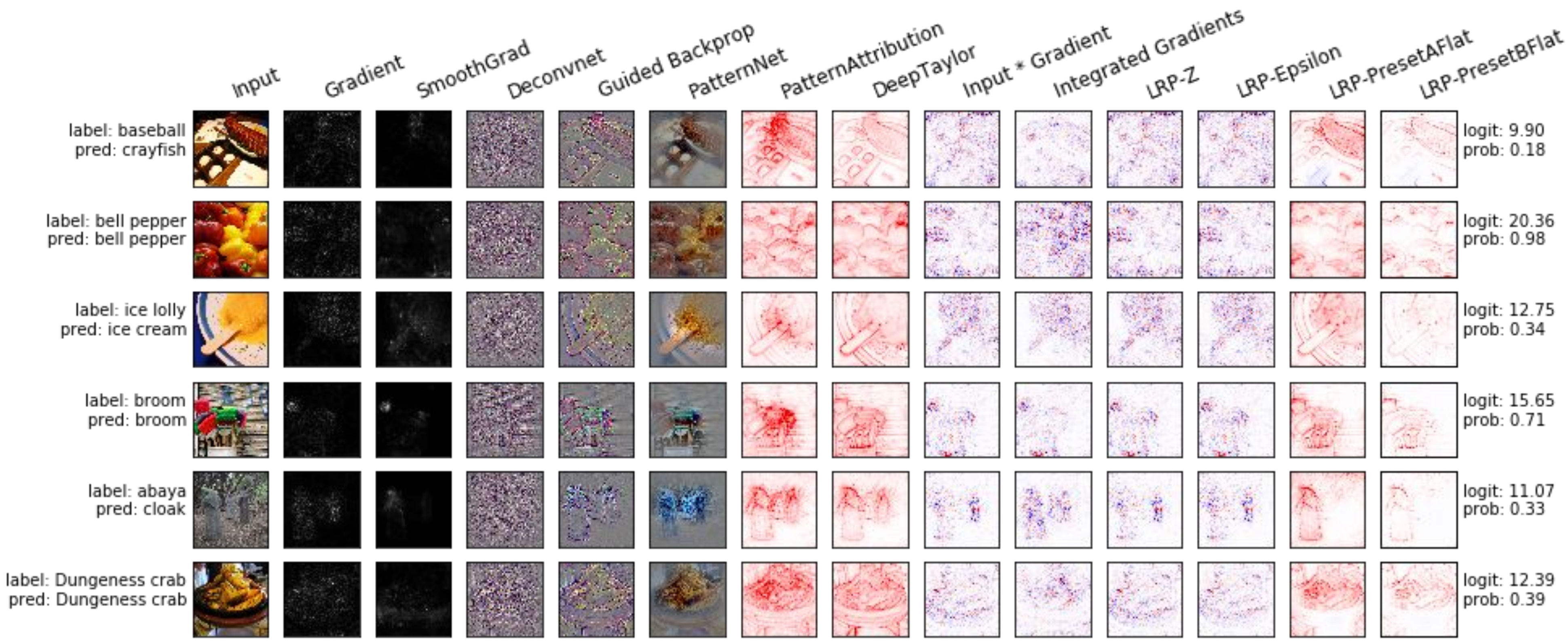



**Figure - The behaviour of several post-model methods for computer vision** (Image from [1])

**Sources: [1] Alber et al. "iNNvestigate Neural Networks!"**

# A different approach to post-model explanations

- **Testing with Concept Activation Vectors**[1]
  - **Explanations: given in terms of human-friendly concepts, by quantifying the degree to which a concept is important to a classification outcome**
  - **Example: how sensitive a prediction of "zebra" is to the presence of stripes**

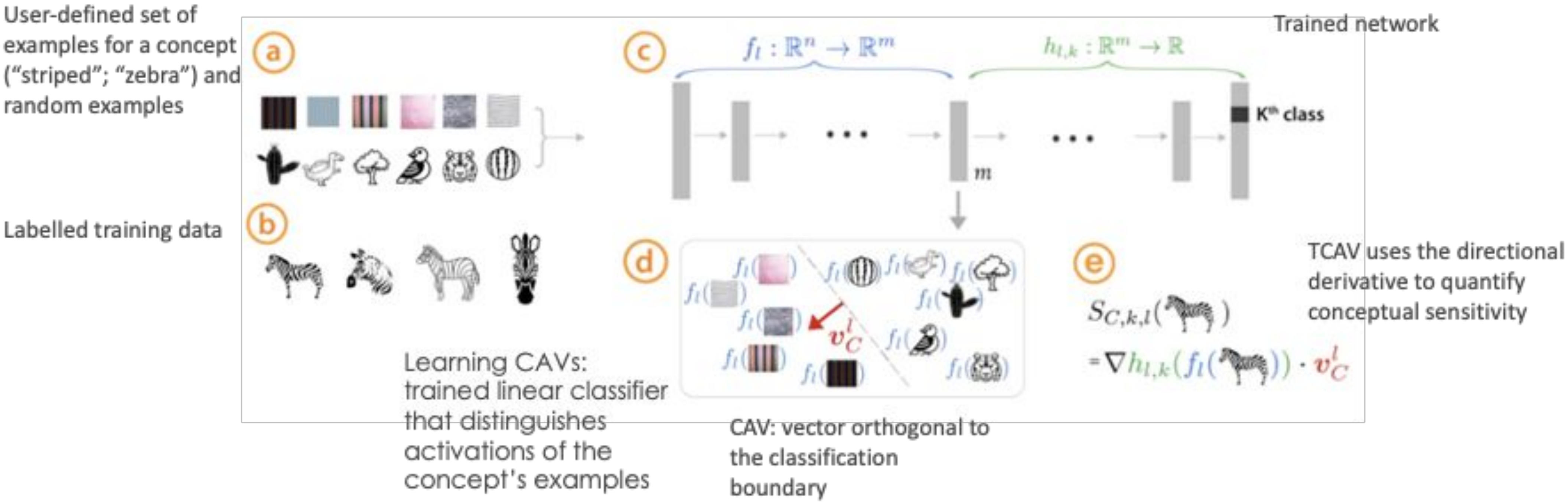



**Figure - Testing with Concept Activation Vectors** (Image from [1])


**Sources: [1]** [Kim et al. "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)"]

# But... Why are we investing so much in post-model?

- **“There is a widespread belief that more complex models are more accurate, meaning that a complicated black box is necessary for top predictive performance”**[1]

- However, this is **not necessarily true**: in problems that **deal with structured data**, one **can often extract meaningful features** for the training of simpler classifiers **without jeopardising performance**

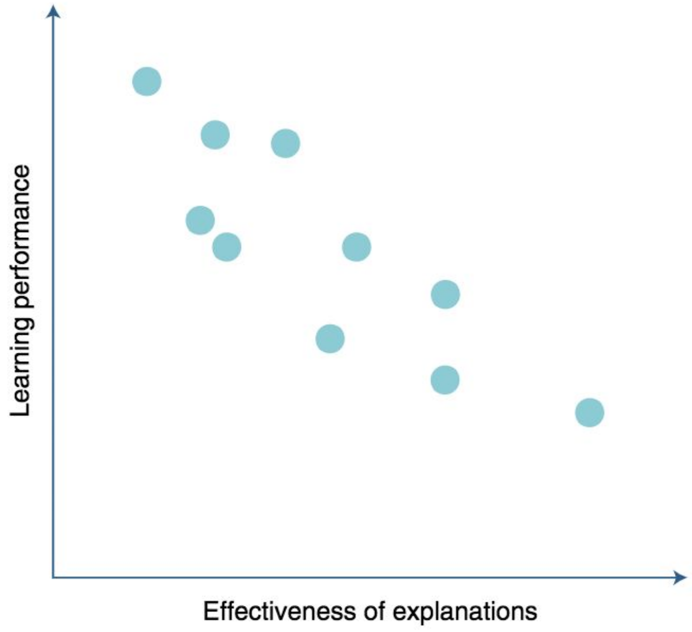



**Figure - The belief that complex models should be more accurate** (Image from [1])

Sources: **[1]** [Cynthia Rudin “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”]

# Post-model explanations often do not make sense in a *human-understandable* manner[1]

- Let's recall the post-model methods presented for computer vision
  - One way or another, most of them produced some kind of ***saliency-maps***

- **Are the outputs considered meaningful explanations for humans?**


**Figure - Do post-model explanations make sense?** (Image from [1])



**Sources: [1]** Cynthia Rudin "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead"

# *Black box* models may be hard to troubleshoot[1]

- Assume that you have an overly **complex model that has *unknown* flaws**
  - **How would you *debug* such a model?**

- Let's say we use an **algorithm that generates post-model explanations to understand the behaviour** of your model
  - Since your model is flawed, **the generated explanations may be impacted by these flaws**

- **Therefore, you could end up with two models to debug: the original model and the explanation model**

**Table 1 | Machine learning model from the CORELS algorithm**

| | | |
|---|---|---|
| IF | age between 18–20 and sex is male | THEN predict arrest (within 2 years) |
| ELSE IF | age between 21–23 and 2–3 prior offences | THEN predict arrest |
| ELSE IF | more than three priors | THEN predict arrest |
| ELSE | predict no arrest | |

This model from ref. 39 is the minimizer of a special case of equation (1) discussed later in the challenges section. CORELS' code is open source and publicly available at http://corels.eecs.harvard.edu/, along with the data from Florida needed to produce this model.

**Table 2 | Comparison of COMPAS and CORELS models**

| COMPAS | CORELS |
|---|---|
| Black box; 130+ factors; might include socio-economic info; expensive (software licence); within software used in US justice system | Full model is in Table 1; only age, priors, gender (optional); no other information; free, transparent |

**Figure - Complex vs Simple. What is the best?** (Image from [1])

Sources: [1] Cynthia Rudin "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead"

# Do *black box* models uncover "hidden patterns"?[1]

- **Inherently-interpretable machine learning models may not be easy to optimise**

- This may contribute to **the *belief*** that the **complex black-box models "have the ability to uncover subtle hidden patterns in the data about which the user was not previously aware"**

- Assume that this is **true**
  - **Are these models extracting meaningful features?**

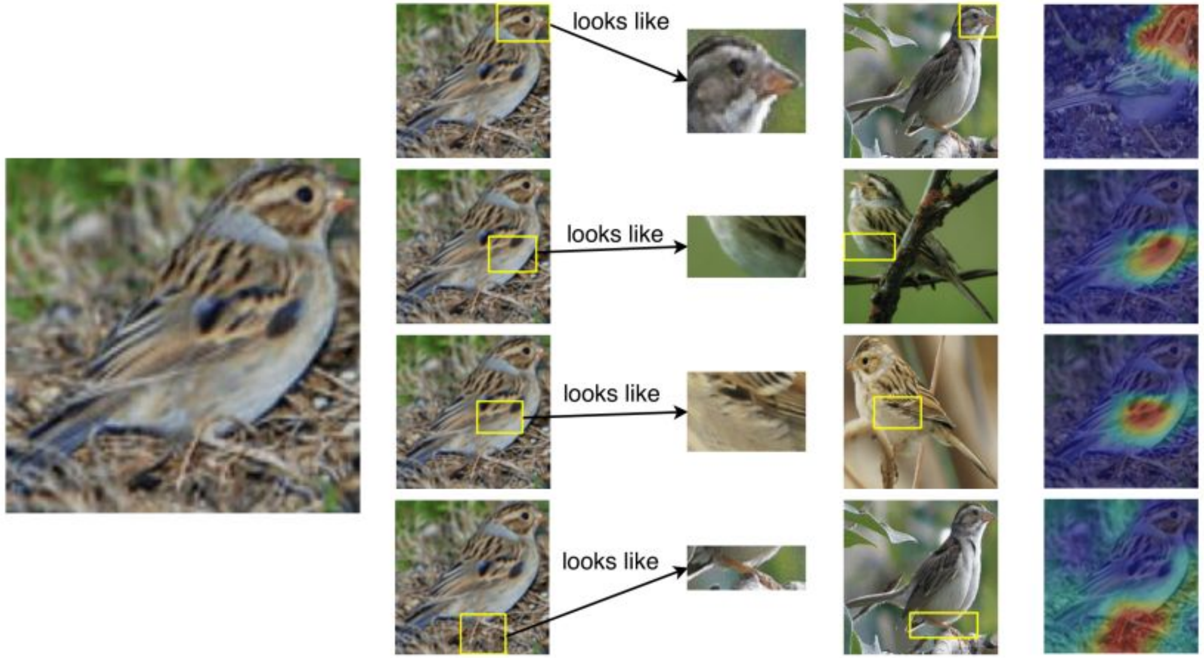
- Assume that this is **false**
  - **Therefore, one should be capable of building a transparent interpretable model that achieves similar performances**
  - **Can we do this for computer vision, where deep learning is widely used?**



**Sources: [1]** Cynthia Rudin "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead"

# Case-Study 1: Learn image *prototypes* and combine them to output a final decision[1]

- **The intuition behind this work is related to the human reasoning method: when we want to classify an image, we may rely on specific parts of the image to justify our final decision**




*Leftmost*: a test image of a clay-colored sparrow
*Second column*: same test image, each with a bounding box generated by our model -- the content within the bounding box is considered by our model to look similar to the prototypical part (same row, third column) learned by our algorithm
*Third column*: prototypical parts learned by our algorithm
*Fourth column*: source images of the prototypical parts in the third column
*Rightmost column*: activation maps indicating how similar each prototypical part resembles part of the test bird

**Figure - Learning images prototypes** (Image from [1])

**Sources: [1]** [Chen et al. "This Looks Like That: Deep Learning for Interpretable Image Recognition"]

# Case-Study 1: Learn image *prototypes* and combine them to output a final decision[1]

- The proposed model, ***prototypical part network*** **(ProtoPNet),** identifies **"several parts of the image where it thinks that this part of the image looks like that prototypical part of some class, and makes its prediction based on a weighted combination of the similarity scores between parts of the image and the learned prototypes"**

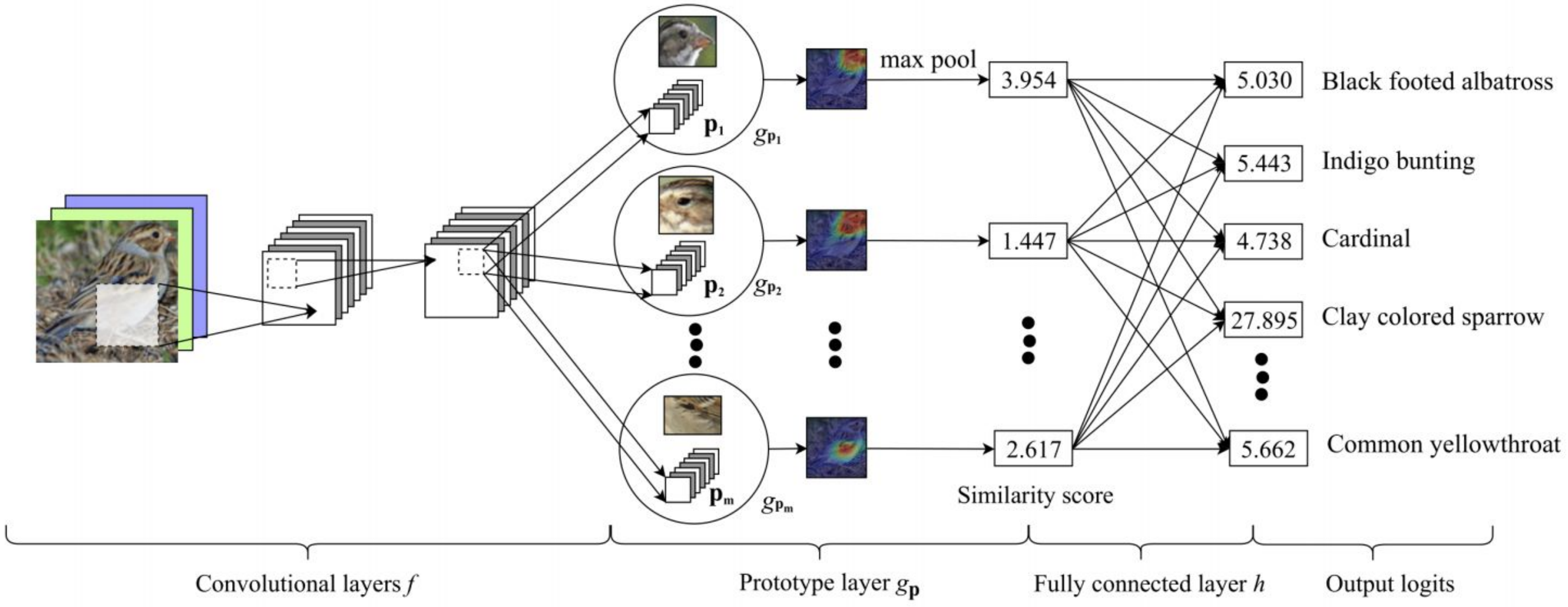



**Figure - ProtoPNet architecture** (Image from [1])



**Sources: [1]** Chen et al. "This Looks Like That: Deep Learning for Interpretable Image Recognition"

INESCTEC

# Case-Study 1: Learn image *prototypes* and combine them to output a final decision[1]

- This model may be considered **"interpretable, in the sense that it has a transparent reasoning process when making predictions"**

- **It is transparent since it can output its explanations in a human-understandable manner**


**Figure - ProtoPNet outputs** (Image from [1])

Sources: [1] Chen et al. "This Looks Like That: Deep Learning for Interpretable Image Recognition"

# Case-Study 2: Jointly learn to classify and explain[1]

- **Joint training of a Classifier and an Explainer in a three-phased training process**

- **Custom loss function that ensures the explainer is justifying the classification decision**

- **Advantages:**
  - Unsupervised training
  - No additional labelling costs
  - Same classification performance
  - Adaptable to different CNNs

- **Disadvantages:**
  - Longer training time
  - Hyperparameter tuning

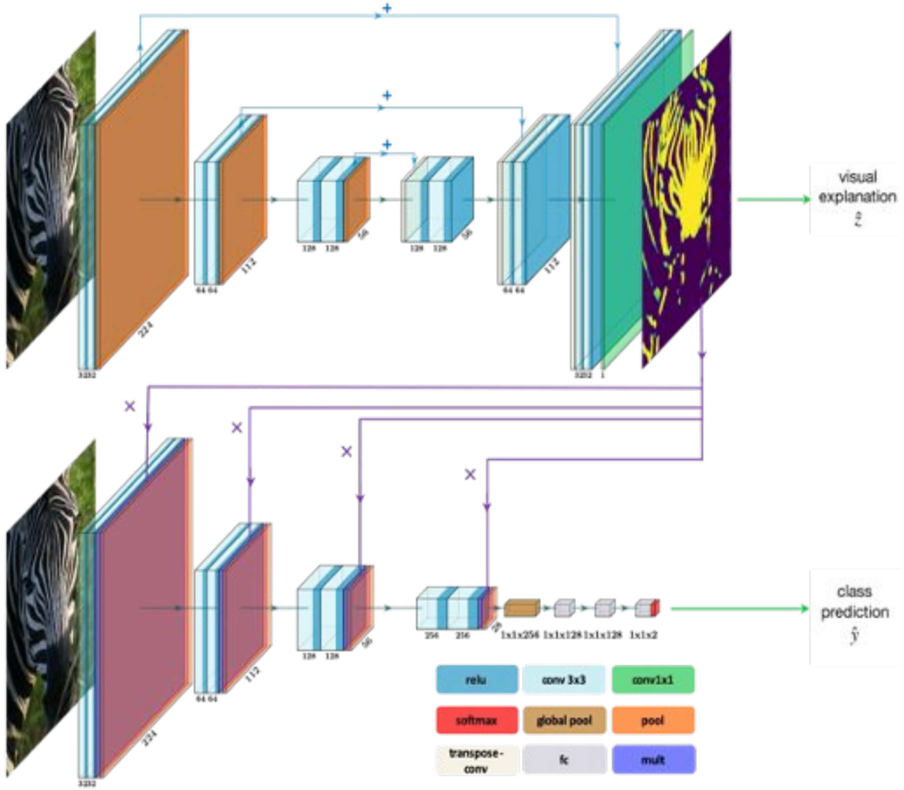



**Figure - Jointly learn to classify and explain** (Image from [1])

Sources: [1] Rio-Torto et al. "Understanding the decisions of CNNs: An in-model approach"

INESCTEC

# Case-Study 2: Jointly learn to classify and explain[1]

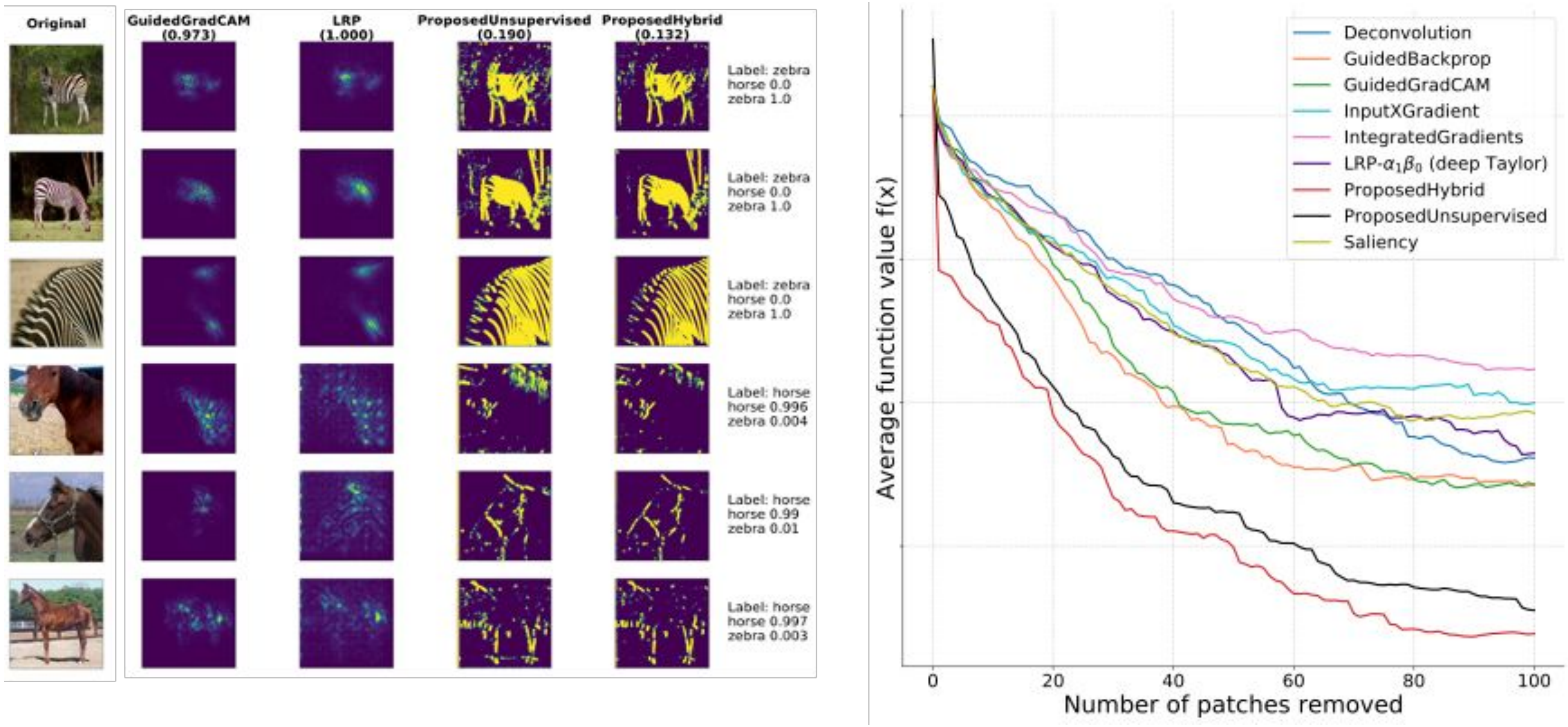



**Figure - Examples of results obtained with the "Jointly learn to classify and explain" architecture** (Image from [1])

Sources: [1] Rio-Torto et al. "Understanding the decisions of CNNs: An in-model approach"

# Case-Study 3: Can we show that post-model methods generate misleading explanations?

- **What if post-model approaches (LIME and SHAP) can be fooled using adversarial attacks?[1]**

- **LIME and SHAP "explain individual predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model (e.g., linear model) locally around each prediction"**

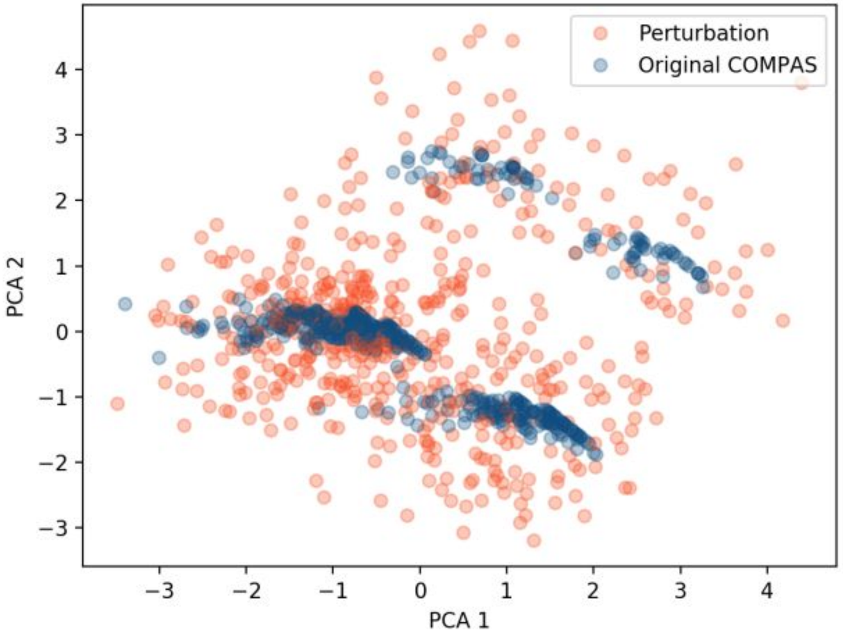



**Figure - Fooling LIME and SHAP** (Image from [1])

Sources: [1] [Slack et al. "Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods"]

# Case-Study 3: Can we show that post-model methods generate misleading explanations?

- **Can we exploit this type of characteristics to achieve a *biased* (racist) classifier which is capable of hiding its inner biases from these post-model strategies that generate explanations based on perturbations?**[1]

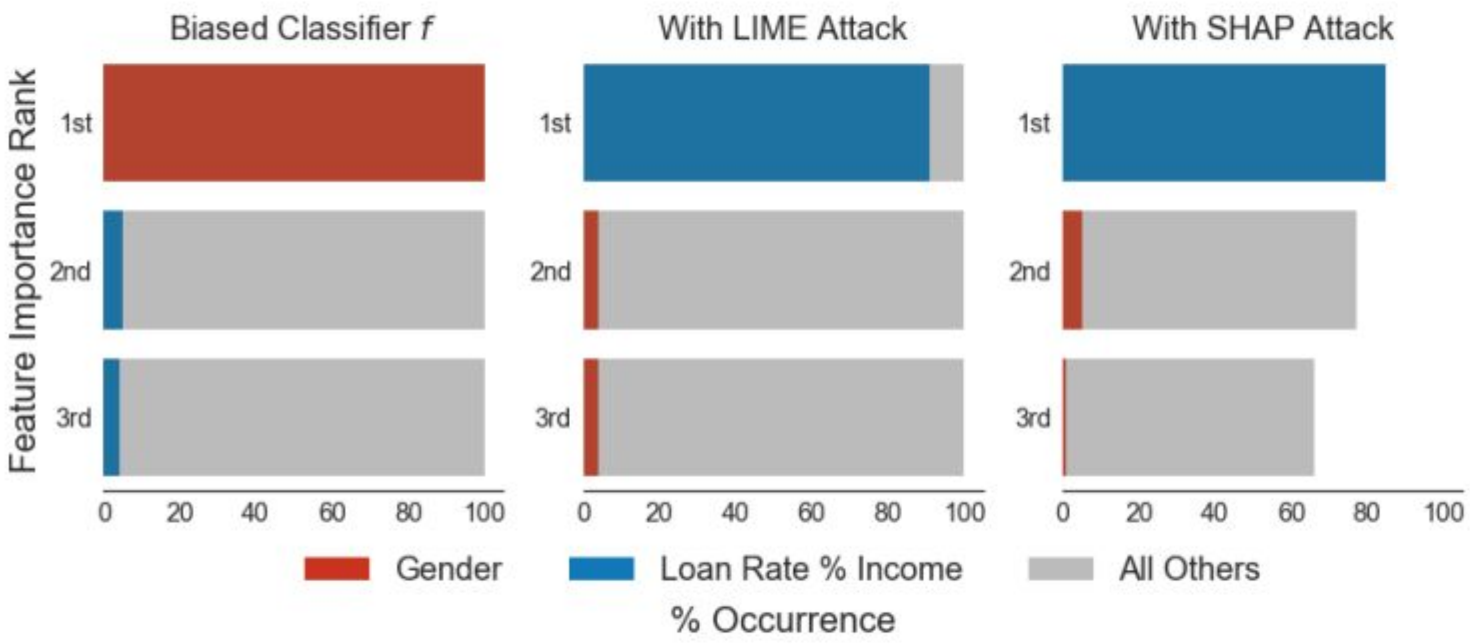



**Figure - Fooling LIME and SHAP** (Image from [1])

**Sources: [1]** [Slack et al. "Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods"]

# 4. Take home messages and further readings

# It is important to contextualise interpretability

- **How can we assess the quality of the explanations?** The ideal system should be:[1]
  - **Flexible**
  - **Robust**
  - **Capable of explaining its reasoning in different modalities, exploring their complementarity and ensuring adaptability to audiences with varying levels of expertise and different use-cases**

- **"For post-hoc interpretability, papers ought to fix a clear objective and demonstrate evidence that the offered form of interpretation achieves it"**[2, 3]

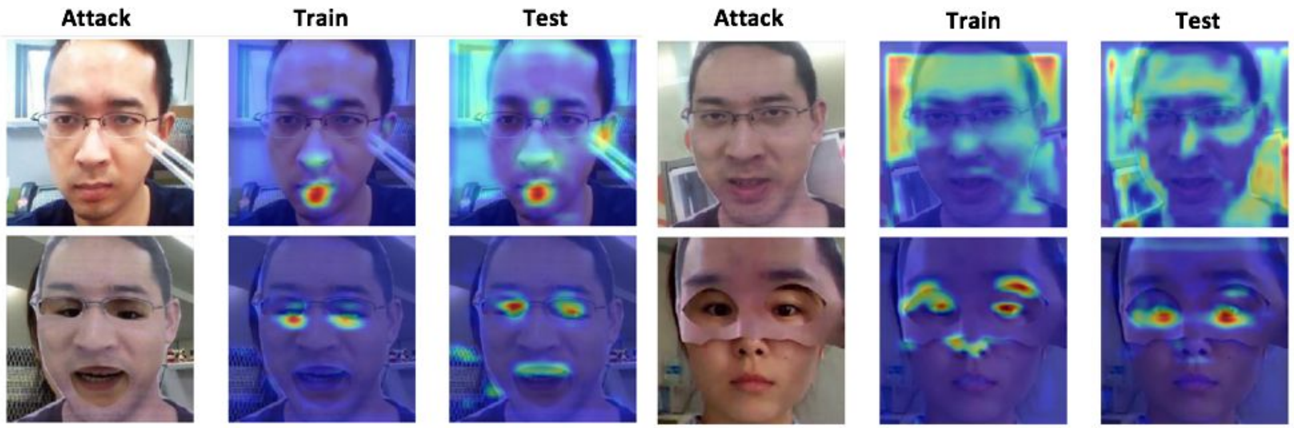



**Figure - Interpretable Biometrics** (Image from [3])

**Sources:** [1] Silva et al. "Towards Complementary Explanations Using Deep Neural Networks", [2] Lipton "The Mythos of Model Interpretability", [3] Sequeira et al."Interpretable Biometrics: Should We Rethink How Presentation Attack Detection is Evaluated?"

# Fairness, transparency, privacy and causality…

- In **high-risk applications** (e.g., justice, healthcare, finance), **fair and transparent** algorithms may be **preferable**
  - If this preference **impacts the predictive power of our model,** it is **important to assess "that the desire for transparency is justified and isn't simply a concession to institutional biases against new methods"**[1]

- **About the data…**
  - **Can we assure that we are not harming the privacy of the subjects present in our datasets?**[2]
  - **Can we be sure that the distribution of our data is not hiding systemic biases?**[3]

- **Current models are quite good at extracting correlations**
  - **Can they be trained to only look at causal events**[4, 5]**?**
  - Some machine learning applications already apply some of these concepts (e.g., reinforcement learning)



**Sources:** **[1] Lipton "The Mythos of Model Interpretability", [2] Yu et al. "Differentially Private Model Publishing for Deep Learning", [3] Shalini Kantayya "Coded Bias", [4] Bernhard Schölkopf "Causality for machine learning", [5] Dana Mackenzie and Judea Pearl "The Book of Why"**

# A (tentative) fair and accurate summary of this lecture

- **Data is the new oil:** as people are generating more data everyday, it will be our duty to use it to create positive impacts on Society

- The democratised access to computational power leveraged the development of novel deep learning algorithms, however, **with great power comes great responsibility**

- **Responsible AI** framework was created to help stakeholders in the implementation of rules and good practices regarding the correct usage of data

- **Interpretability** (aka explainable artificial intelligence aka **xAI**) can be seen as a **three-stage process** composed of **pre-model, in-model, post-model methods**

- Although we are investing in post-model methods, the **future of applications that require high-stake decisions** will rely on **pre-model and in-model** methods!

# Further readings...

- Wilson Silva and Tiago Gonçalves "Explainable artificial intelligence: unveiling what machines are learning"
- Carvalho et al. "Machine Learning Interpretability: A Survey on Methods and Metrics"
- Sequeira et al. "An exploratory study of interpretability for face presentation attack detection"
- Chen et al. "Concept whitening for interpretable image recognition"
- Silva et al. "Interpretability-Guided Content-Based Medical Image Retrieval"
- Sundararajan et al. "Axiomatic Attribution for Deep Networks"
- Bach et al."On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation"

INESCTEC

# Responsible AI - Lecture 1

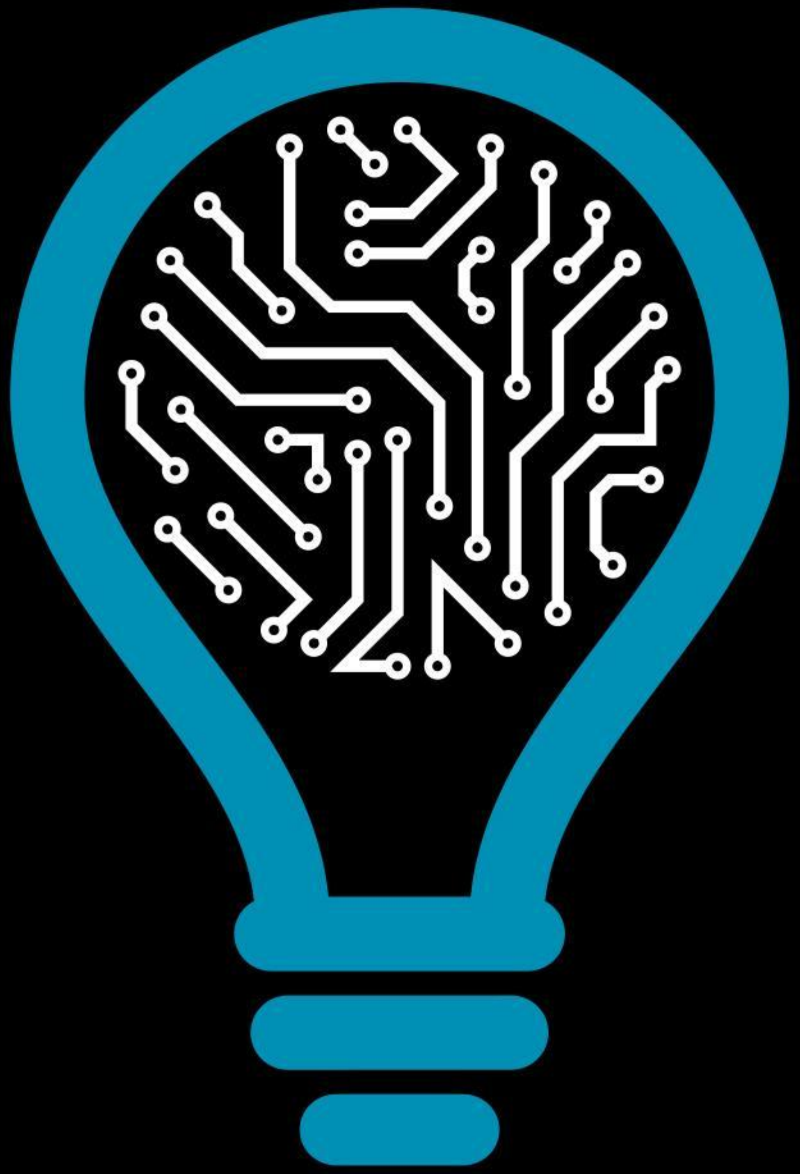
**TAIA - Advanced Topics on Artificial Intelligence**

Tiago Filipe Sousa Gonçalves

tiago.f.goncalves@inesctec.pt | tiagofs@fe.up.pt

Acknowledgement: Isabel Rio-Torto

isabel.riotorto@inesctec.pt



INESCTEC INSTITUTE FOR SYSTEMS AND COMPUTER ENGINEERING, TECHNOLOGY AND SCIENCE