Post-hoc Memories of an Unfinished PhD in Interpretable Artificial Intelligence in Medicine

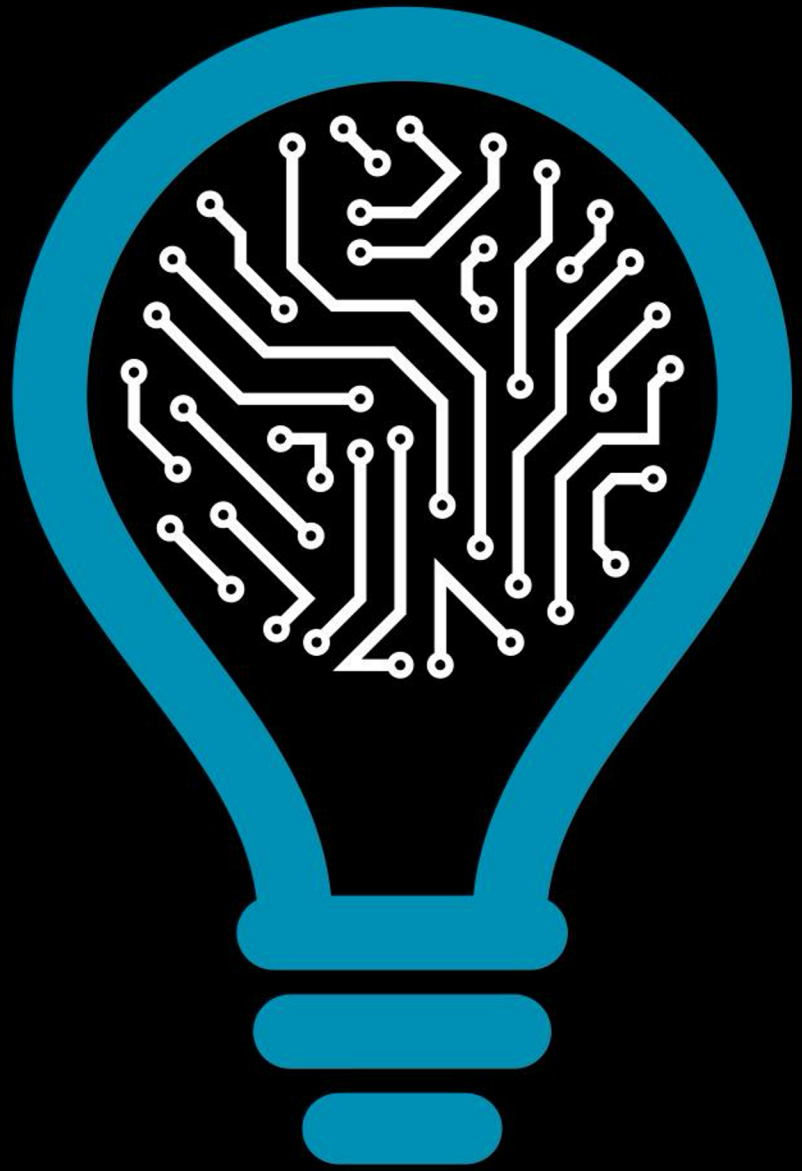
AI for Multi-center Data
Department of Radiology
The Netherlands Cancer Institute

October 27, 2023

Tiago Filipe Sousa Gonçalves
PhD Student at FEUP
Research Assistant at INESC TEC
Visiting PhD Student at MGH

https://tiagofilipesousagoncalves.github.io

INESCTEC
INSTITUTE FOR SYSTEMS AND COMPUTER ENGINEERING, TECHNOLOGY AND SCIENCE



INESCTEC

Outline

1. Data: The New God Emperor of Earth?

2. Throwback to 2020: Thinking about the Open Challenges of Interpretable Artificial Intelligence

3. Attention is All You Need: Is It?

4. A Nexus Point in the AI Timeline: Networks Learning Prototypes

5. Multimodalities of Madness: Combining Disparate Data Sources for the Perfect Spell

6. Days of Future Past: AI and Society

INESCTEC

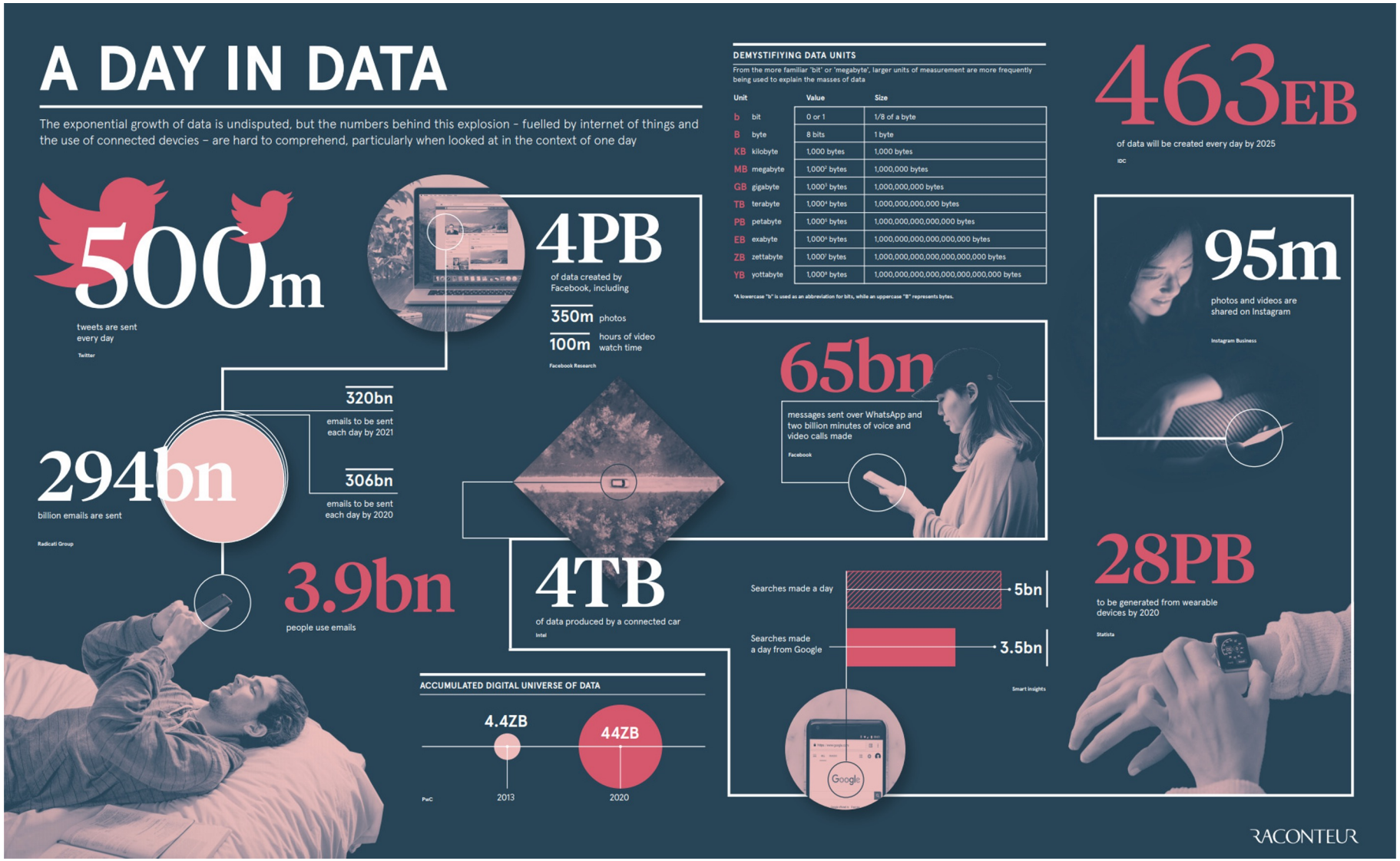
1. Data: The New God Emperor of Earth?

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

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"**

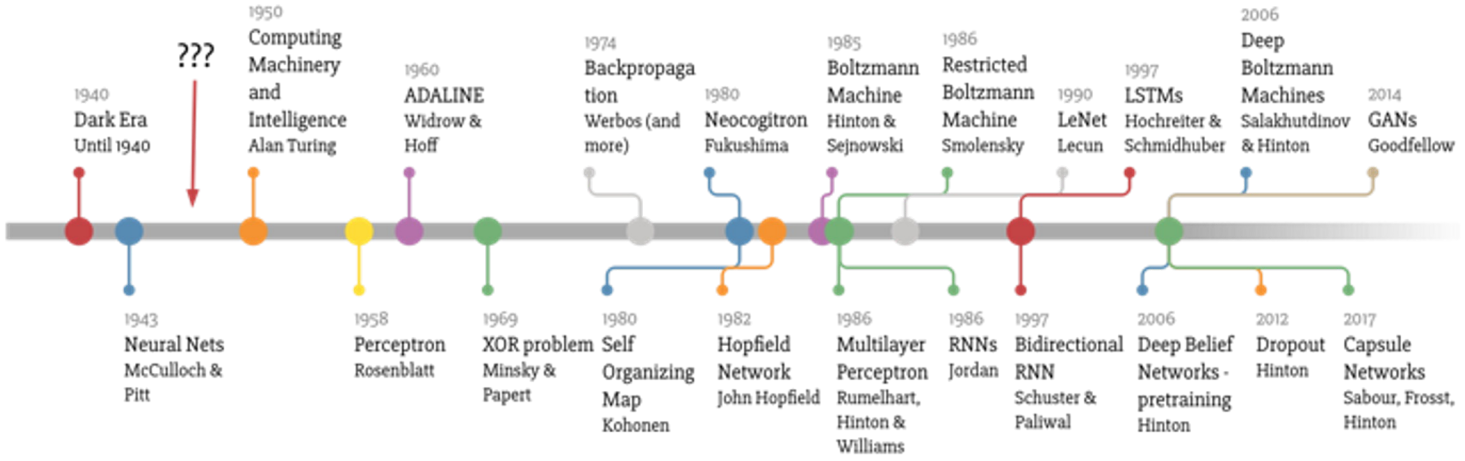


Figure - A (tentative) deep learning timeline (Image from [1])

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 learned 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]

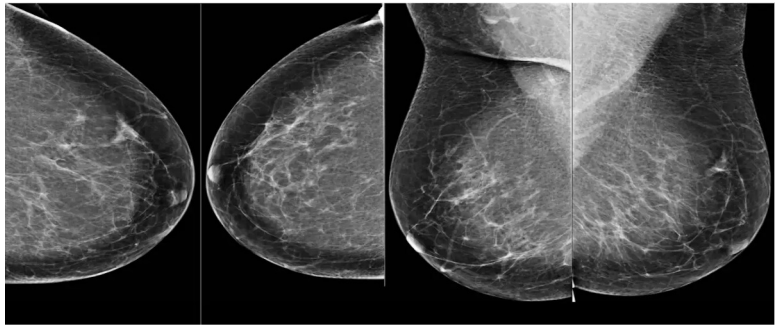


Figure - Medical Image Analysis: Mammograms (Image from [4])

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

INESCTEC

2. Throwback to 2020: Thinking about the Open Challenges of Interpretable Artificial Intelligence

Everything seemed good, but…

- **The increase of available computational power and the democratized access to a huge amount of data** has leveraged the development of novel artificial intelligence (AI) algorithms and their applications

- Deep learning techniques **have been challenging human performance** at some specific tasks such as cancer detection in biomedical imaging[1] or machine translation in natural language processing[2]

- However, most of these models work as black boxes (i.e., their internal logic is hidden to the user) that receive data and output results **without justifying their predictions in a human understandable way**[3]

Sources: [1] McKinney et al. "International evaluation of an AI system for breast cancer screening", [2] Belinkov and Glass "Analysis Methods in Neural Language Processing: A Survey", [3] Cynthia Rudin "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead"

Wait! Are we dealing with black boxes?

- 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** behavior)
  - 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)

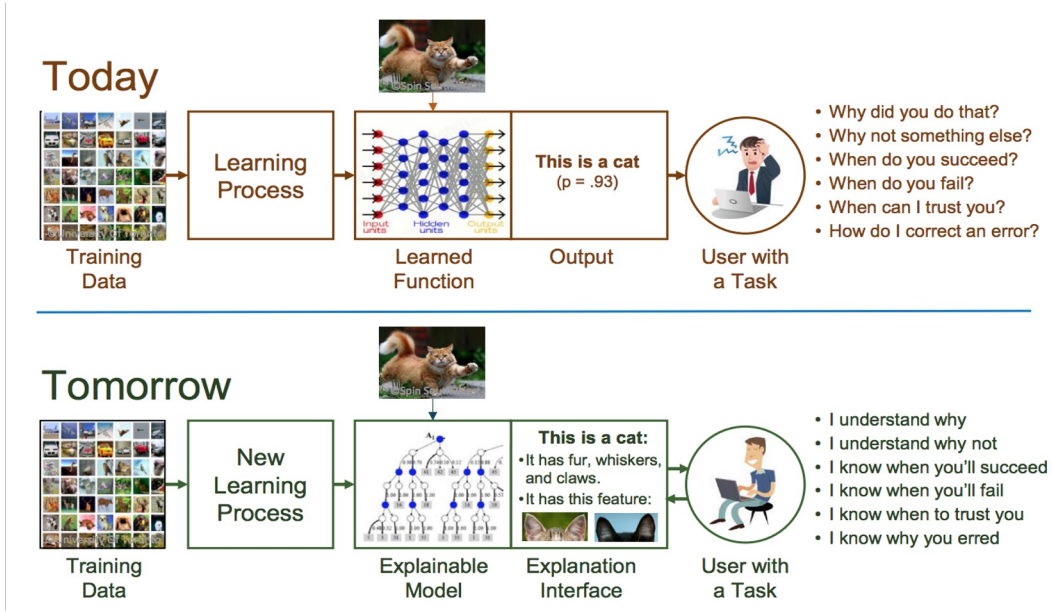


Figure - The future of machine learning algorithms
(Image from [1])

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

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"

Explainable AI will solve all our problems (I thought)

- The topic of explainable artificial intelligence (XAI) appeared intending to contribute to a more **transparent AI**[1]
  - There are three distinct strategies: **pre-, in- and post-model** methods

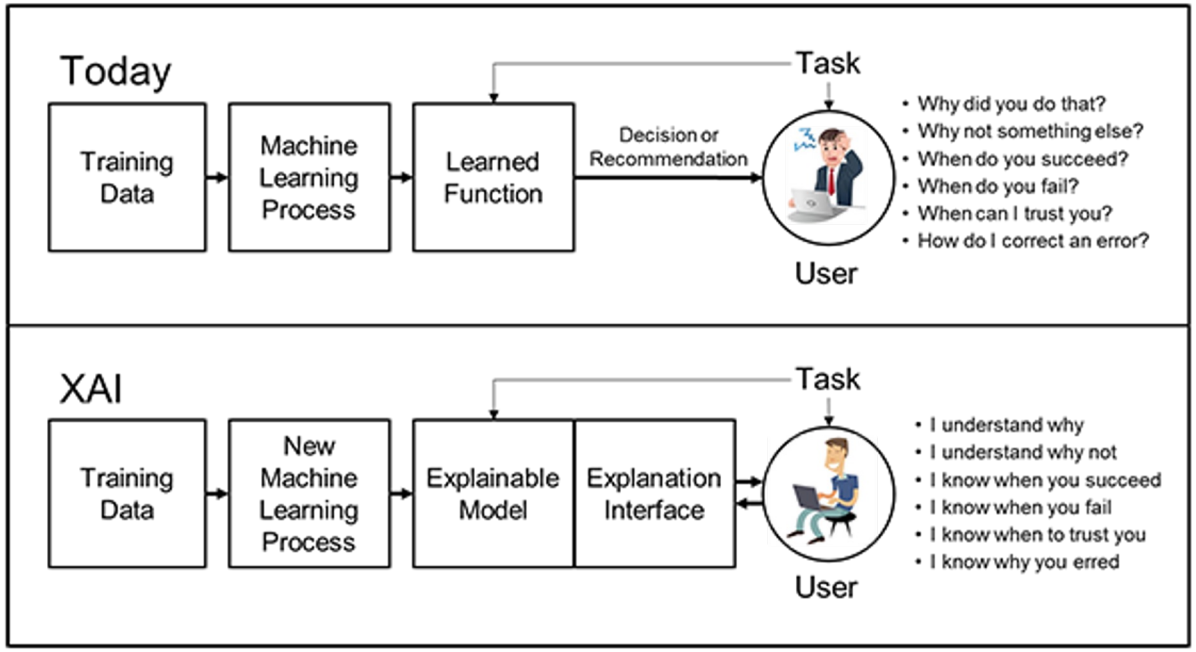


Figure: The concept of XAI, from the Defense Advanced Research Projects Agency (DARPA)[2]

Sources: [1] L. H. Gilpin et al. "Explaining Explanations: An Overview of Interpretability of Machine Learning", [2] https://www.darpa.mil/program/explainable-artificial-intelligence

Why is everyone going post-hoc?

- The **generalised belief that complex models seem to uncover "hidden patterns"** actively contributed to the research and **development of post-model** methods

- There are **several drawbacks of exclusively investing** in a post-model strategy[1]:
  - **Explanations are just an approximation** to what the model computes
  - **Explanations may not provide enough detail** to understand what the model is doing

- It is fundamental to assess the quality of these explanations[1] and to dedicate more effort to pre- and in-model strategies
  - **Pre-model interpretability**: understanding the data distribution that we are dealing with will contribute to an increase of confidence with the posterior decisions and explanations[2]

  - **In-model interpretability**: since models that are inherently interpretable provide their explanations and are faithful to what the machine learning model actually computes[1]

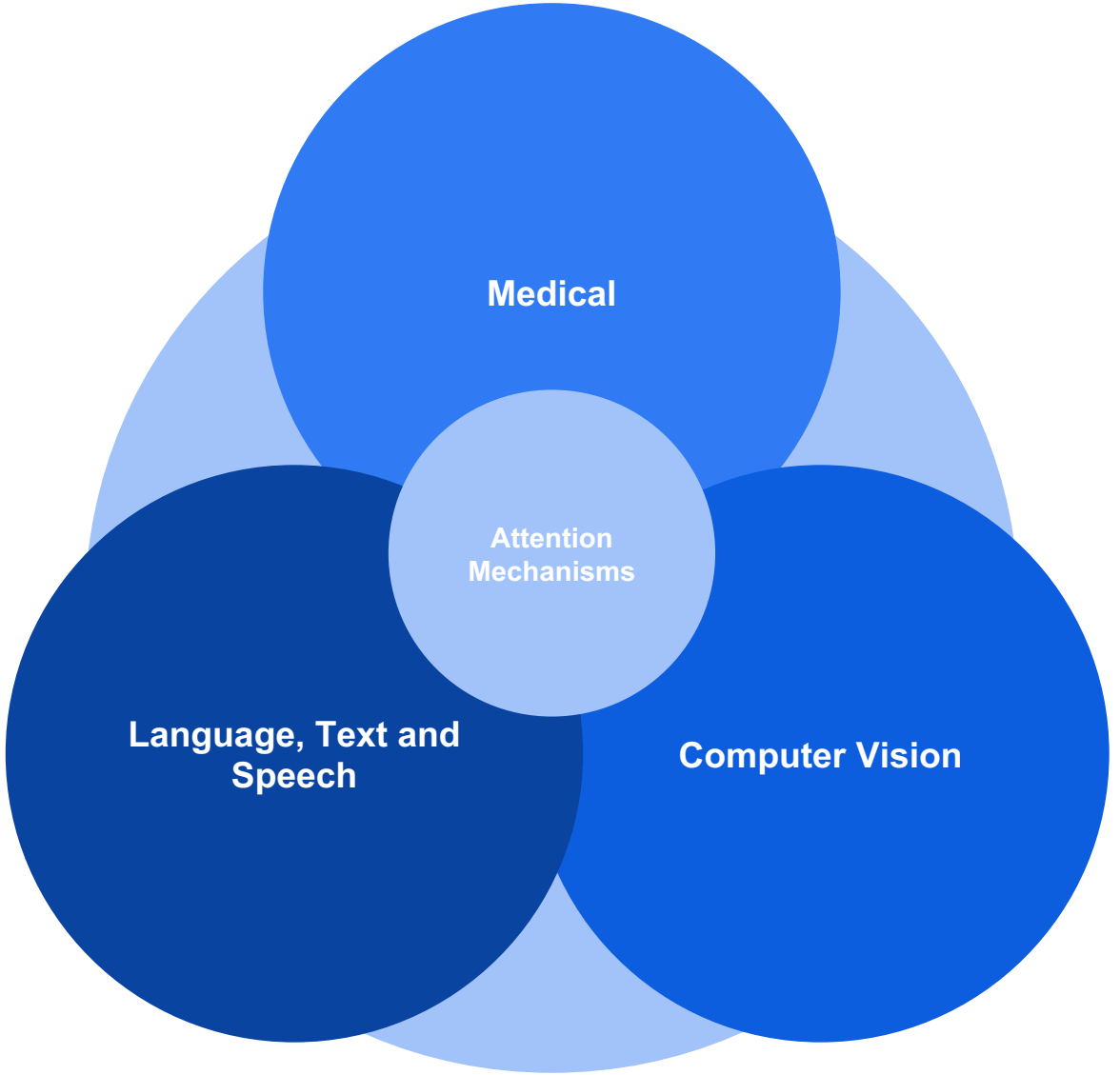
Sources: [1] Cynthia Rudin "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead",
[2] Wilson Silva et al. "How to produce complementary explanations using an Ensemble Model"

INESCTEC

3. Attention is All You Need: Is It?

What if some parts of the data are more relevant than others?

In AI systems, **some parts of the input data are more relevant than others** (e.g., in automatic translation systems, only a subset of words is relevant)[1]



Sources: [1] S. Chaudhari et al. "An Attentive Survey of Attention Models"

What is the expected behavior of attention-based models for medical use cases?[1]

APTOS2019

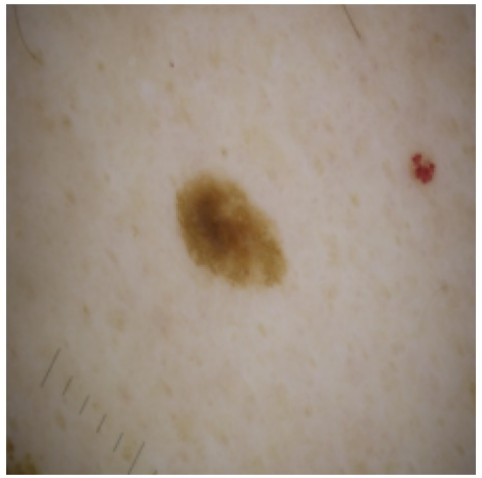
Retinography data related to retinopathy severity score

In our paper, we worked on a binary case: ”Normal” vs “Diabetic Retinopathy”

ISIC2020

Dermoscopic images of benign and malignant skin lesions

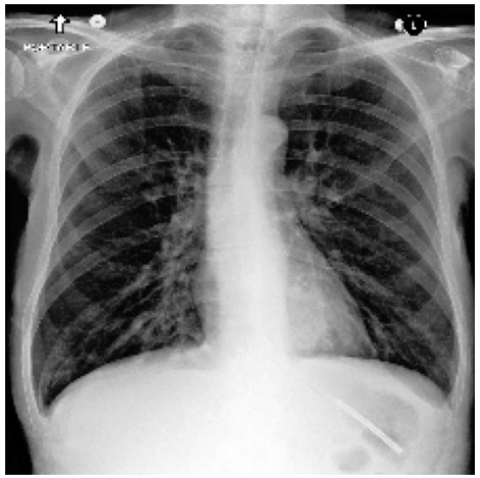
In our paper, we worked on the binary case: “Benign” vs “Malign”



MIMIC-CXR

Chest radiographs database

In our paper, we worked on the binary case: “Normal” vs “Pleural Efusion”

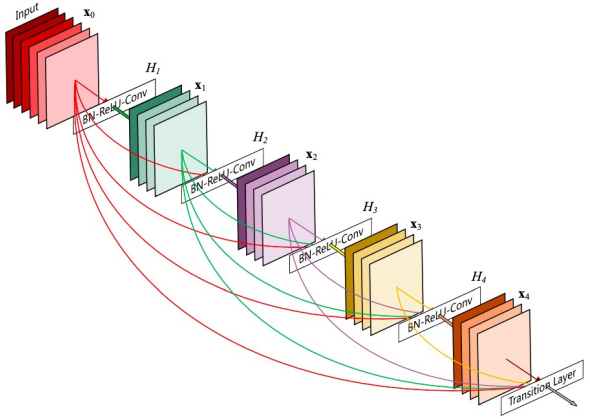


Sources: [1] Tiago Gonçalves et al. “A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"

We started with two different backbone architectures[1]

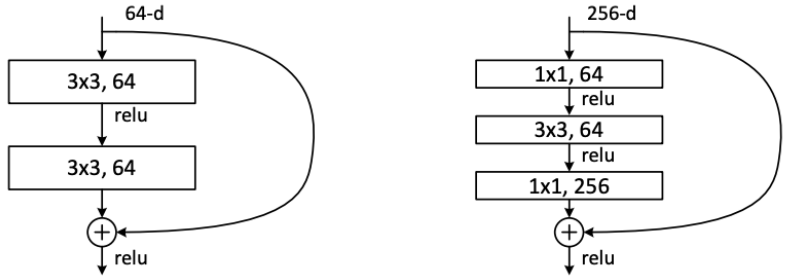
DenseNet-121[2]

Allows connecting all layers directly with each other, thus improving the flow of information and gradients throughout the network, and facilitating their training



ResNet-50[3]

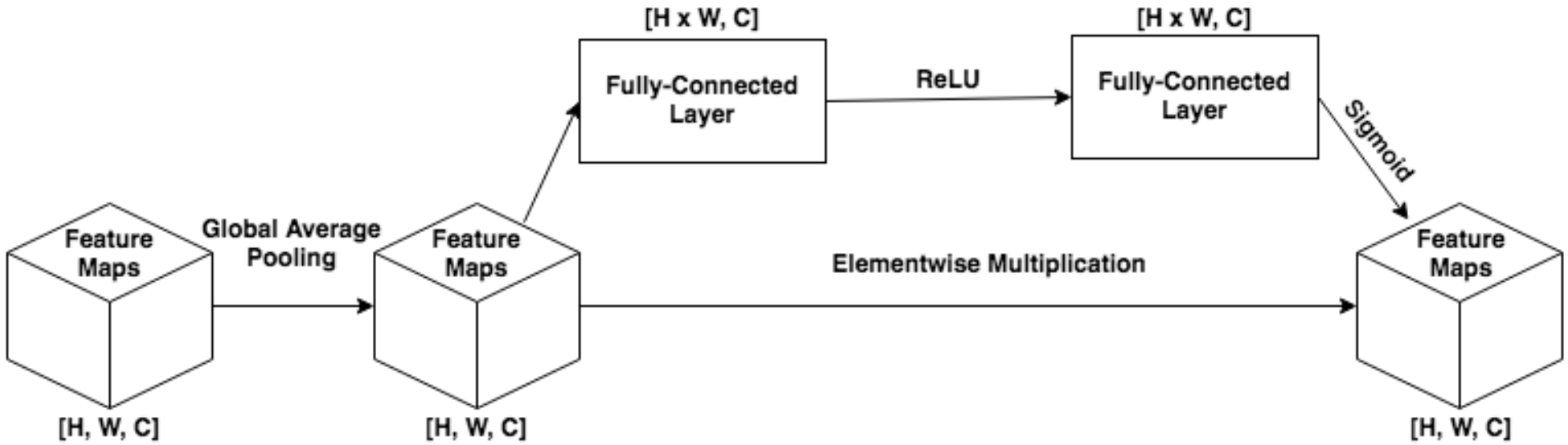
Introduced the deep residual learning framework, which consists of adding skip connections that perform identity mapping and adding their outputs to the outputs of the stacked layers



Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?", [2] Gao Huang et al. "Densely Connected Convolutional Networks", [3] Kaiming He et al. "Deep Residual Learning for Image Recognition"

1/2: Integrated the *Squeeze-and-Excitation (SE)* block into the backbones[1]

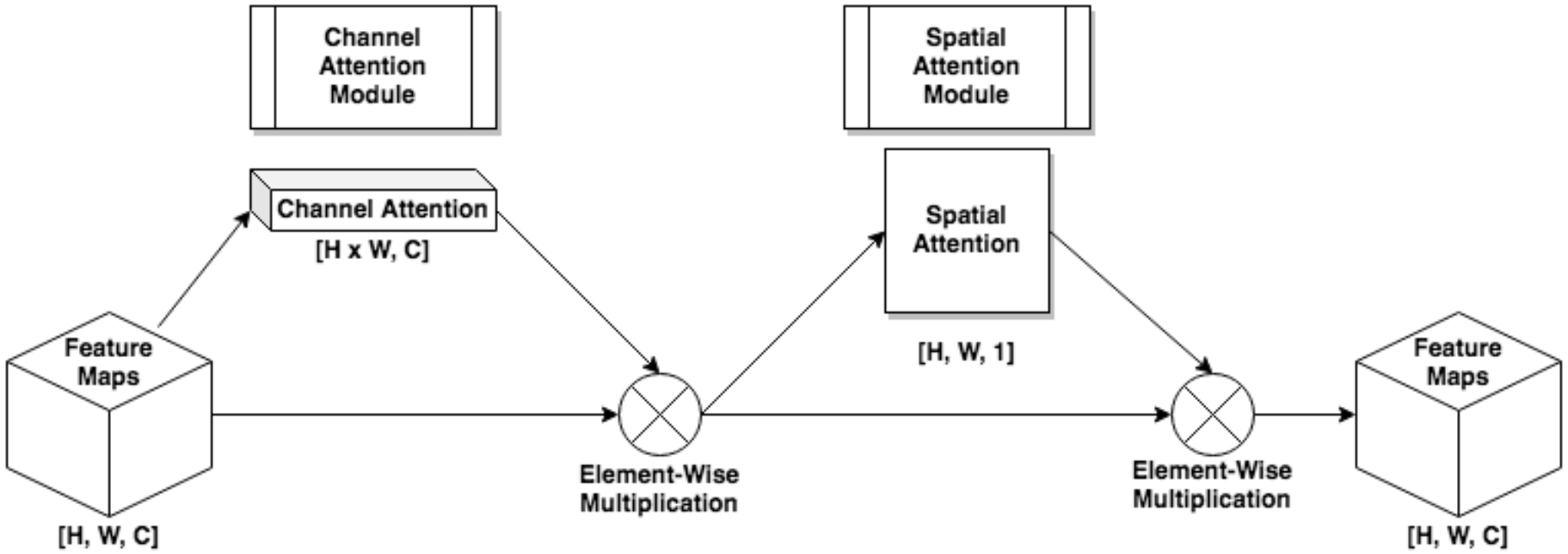
This attention block was designed to adaptively recalibrate channel-wise feature responses by explicitly modeling interdependencies between channels[2]



Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?", [2] Jie Hu et al. "Squeeze-and-Excitation Networks"

2/2: Integrated the *Convolutional Block Attention Module (CBAM)* block into the backbones[1]

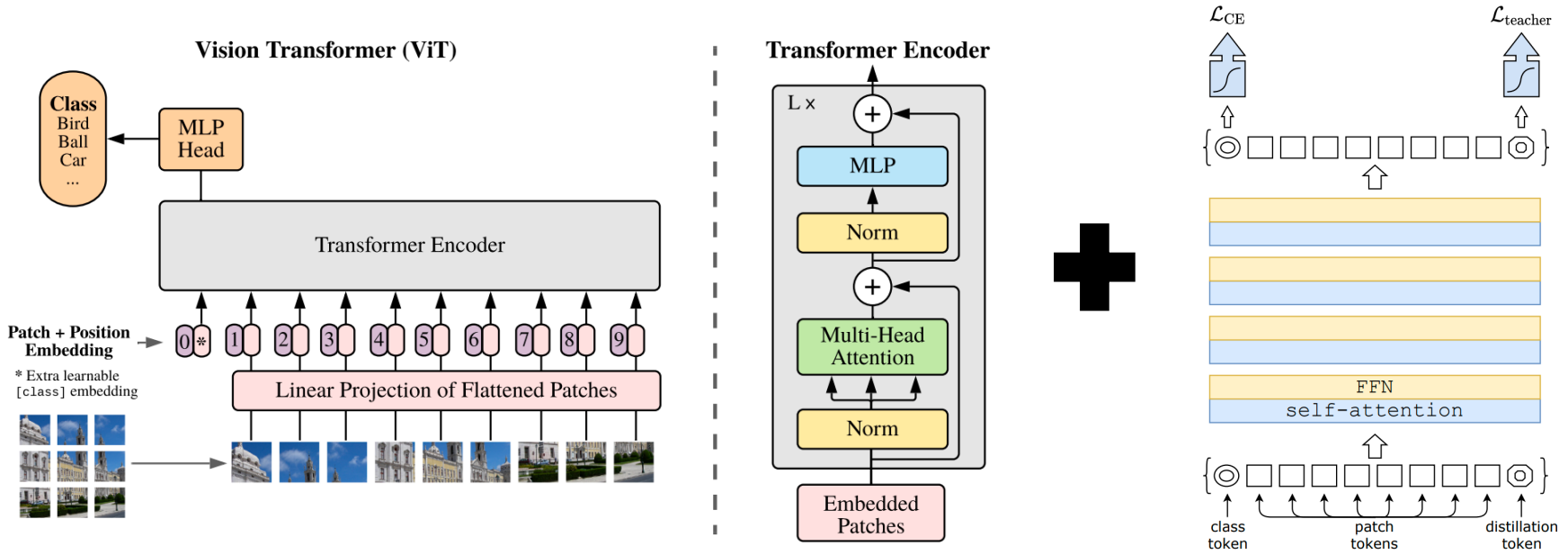
This attention mechanism integrates two specific attention blocks[2]:



Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?", [2] Sanghyun Woo et al. "CBAM: Convolutional Block Attention Module"

Naturally, the Transformer also suffered from our analysis[1]

The Data-efficient image Transformer (DeiT)[2] is an architecture inspired by the Vision Transformer[3] and trained with fewer parameters. In this case, we used the DeiT-Ti variation[2], which has a comparable number of parameters against the chosen CNN backbones



Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?",
[2] Hugo Touvron et al. "Training data-efficient image transformers & distillation through attention", [3] Alexey Dosovitskiy "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale"

The level of interpretability was measured using post-hoc methods[1]

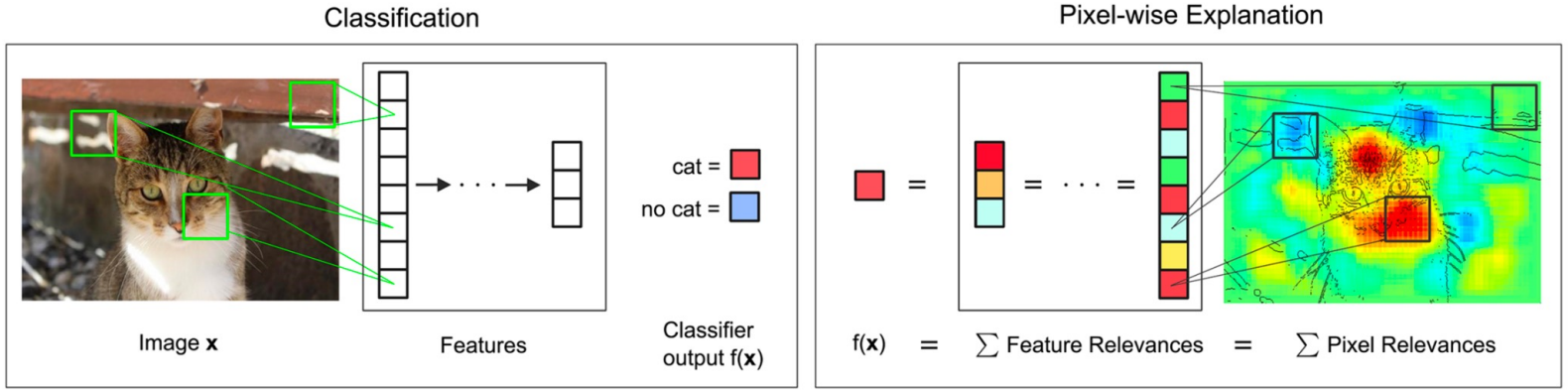
DeepLIFT[2]

The *Deep Learning Important FeaTures* (DeepLIFT) compares the activation of each neuron to its related reference activation and assigns contribution scores according to the difference

LRP[3]

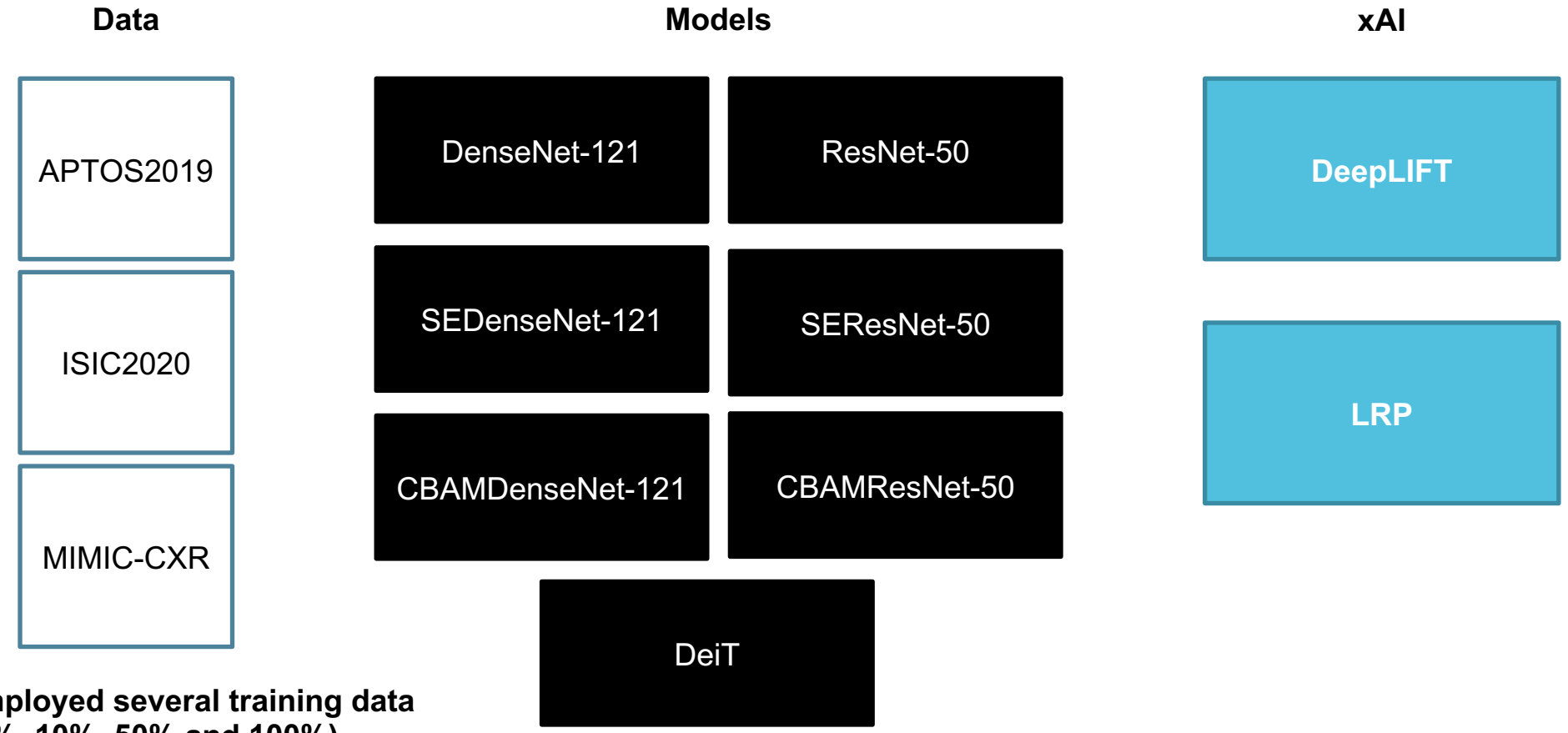
The *Layer-wise Relevance Propagation* (LRP) is a methodology that aims to create visualizations of the contributions of single pixels to predictions

Image below provides an intuition



Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?",
[2] Avanti Shrikumar et al. "Learning Important Features Through Propagating Activation Differences",
[3] Sebastian Bach et al. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation"

Well, after some work, we reached this experimental protocol[1]:

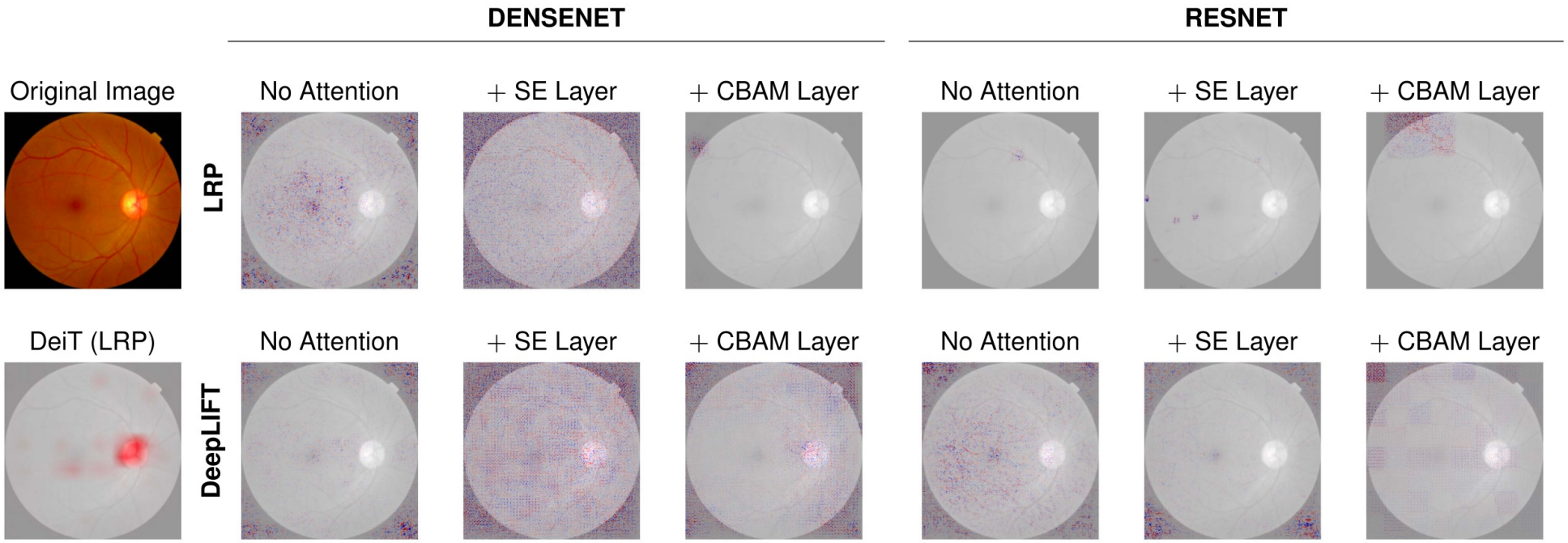


We also employed several training data regimes (1%, 10%, 50% and 100%)

Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"

1/3: What about explainability?[1]

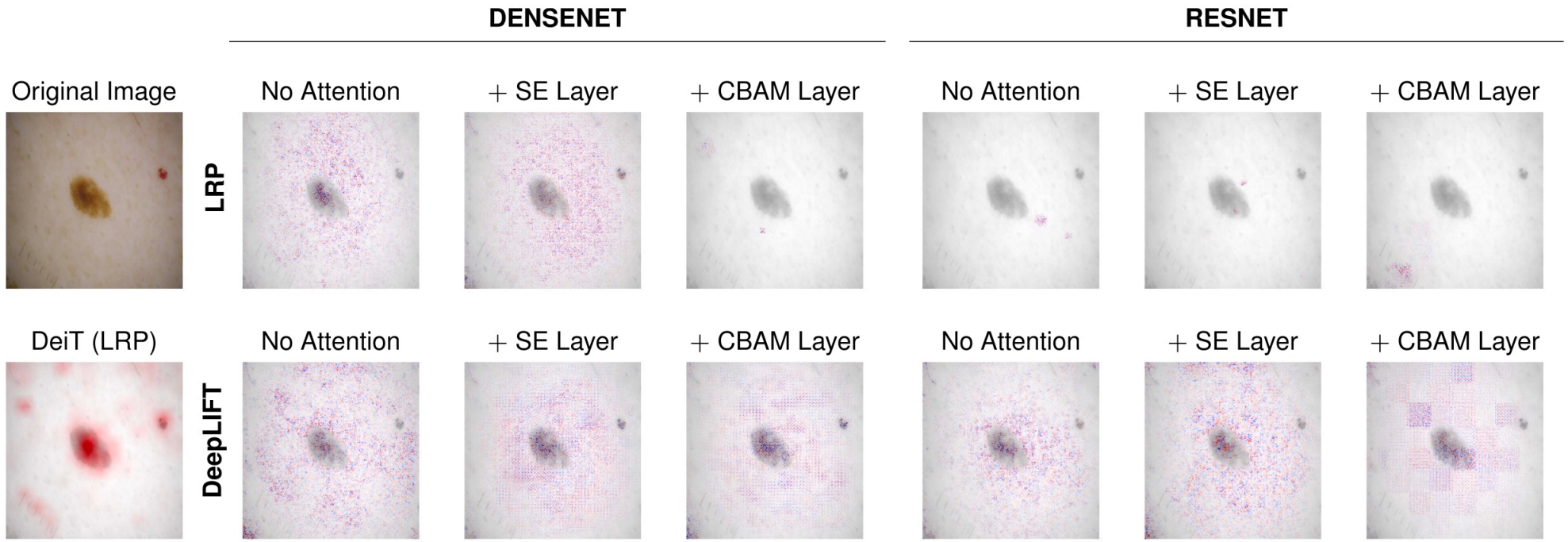
TABLE 5. **Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the APTOS2019 data set with the label 0 correctly classified as 0 by all models.**



Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"

2/3: What about explainability?[1]

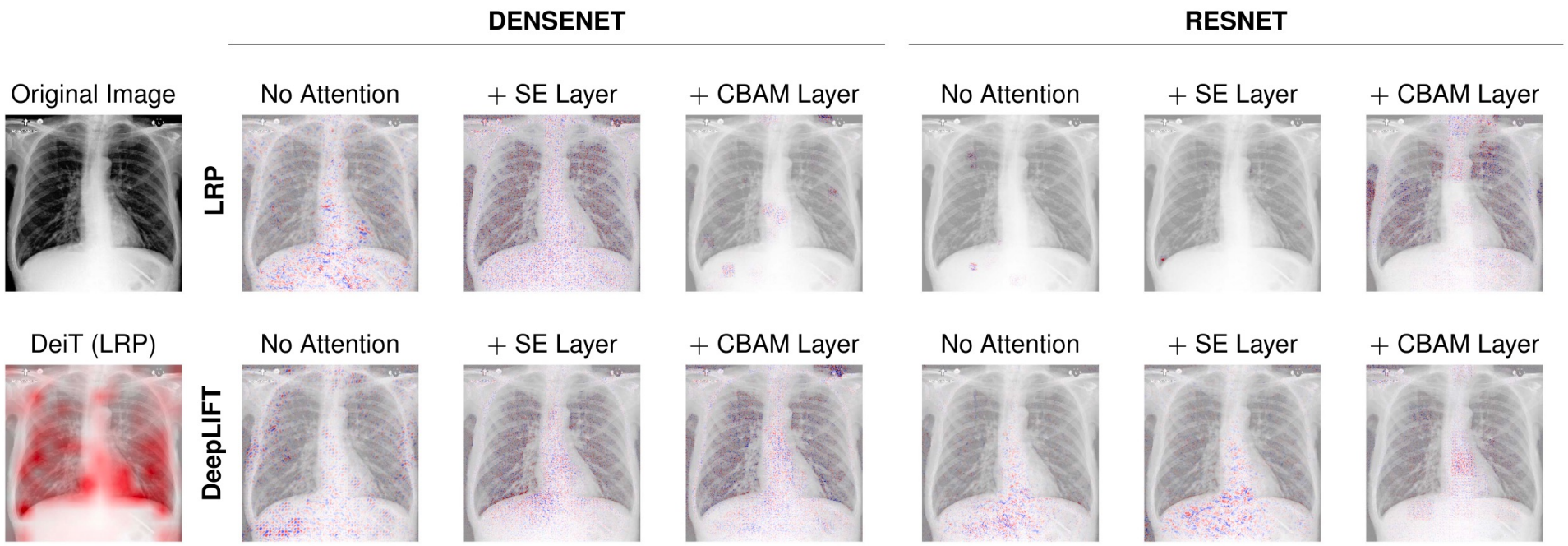
TABLE 8. **Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the ISIC2020 data set with the label 0 correctly classified as 0 by all models.**



Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"

3/3: What about explainability?[1]

TABLE 12. **Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the MIMIC-CXR data set with the label 0 correctly classified as 0 by all models.**



Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"

So… Are we moving towards better algorithms?[1]

- **We found that backbone models can attain equivalent predictive performances to Transformer-based architectures with equivalent model complexity (i.e., number of parameters)**

- **When using a post-hoc framework to visually assess what type of features these models can extract, we can conclude that there is still a high degree of subjectivity in such analysis (i.e., results are very noisy, even for the cases of attention mechanisms, which is counter-intuitive)**

- **The community is moving toward using attention mechanisms (specially Transformer-based ones) and arguing that these frameworks increase the quality, transparency, and interpretability of deep learning architectures. However, we state that this is not true**

Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"

A New Hope (or Future Challenges)[1]

1. **Attention Mechanisms: Past or Future?** Even if attention mechanisms are pushing deep learning algorithms towards the limits of their predictive power, we must start thinking about creating interpretable frameworks that allow us to audit and assess these algorithms concerning the specific conditions of their domains

2. **Design and Integration of Attention Mechanisms** If we look at the topographies of these deep learning algorithms, it is not always clear for the users where they should place these modules, and why it makes sense to put them in a specific place. Another question arises: are these attention modules dependent on the backbone into which they are integrated to?

3. **The Rise of Transformers** While there is hype on the use of these structures, it is not clear whether they are more interpretable or not, or if their generalization power is superior to the other deep models

4. **Interpretability is the Path to Better Algorithms** Even if we intend to keep using visual saliency maps to explain our models, we must achieve a clear standard, validated by the clinical community, of what these maps should look like and what is their effective meaning

Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"

4. A Nexus Point in the AI Timeline: Networks Learning Prototypes

The day the World lost faith in AI

- The **black box behavior of deep learning models does not help decision-makers** to have a **clear understanding of their inner-functioning**, thus preventing them to diagnose errors and potential biases or deciding when and how much to rely on these models[1]

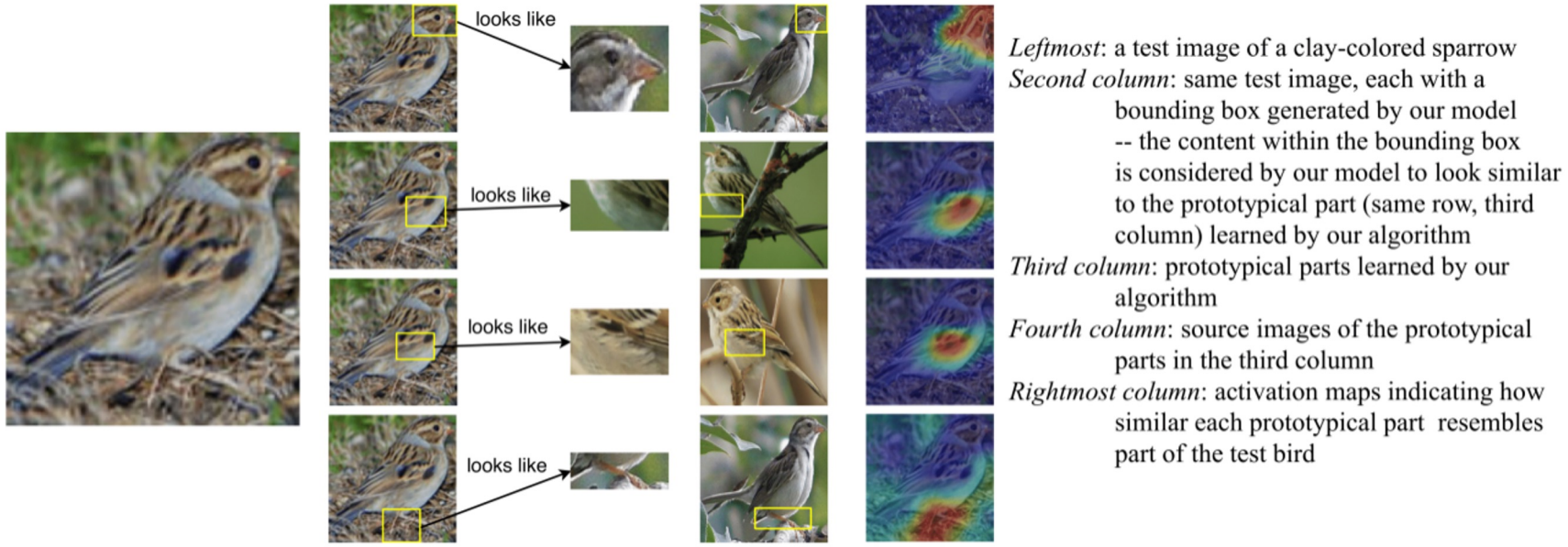
- There has been a huge effort into the **development of post-model strategies** to explain the behavior of black box models, however, the outputs of these algorithms are prone to **subjective evaluation**, **may be misleading**[2] or **fooled**[1]

- Besides, we agree that just being able to obtain explanations is not enough and that we rather need to take into account at the development stage that these methods must **respect specific constraints** that give them the **capability of generating human-understandable explanations** and make decisions based on such premises[3]

Sources: [1] Dylan Slack et al. "Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods", [2] Zachary C. Lipton "The Mythos of Model Interpretability", [3] Cynthia Rudin "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead"

What If? 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**

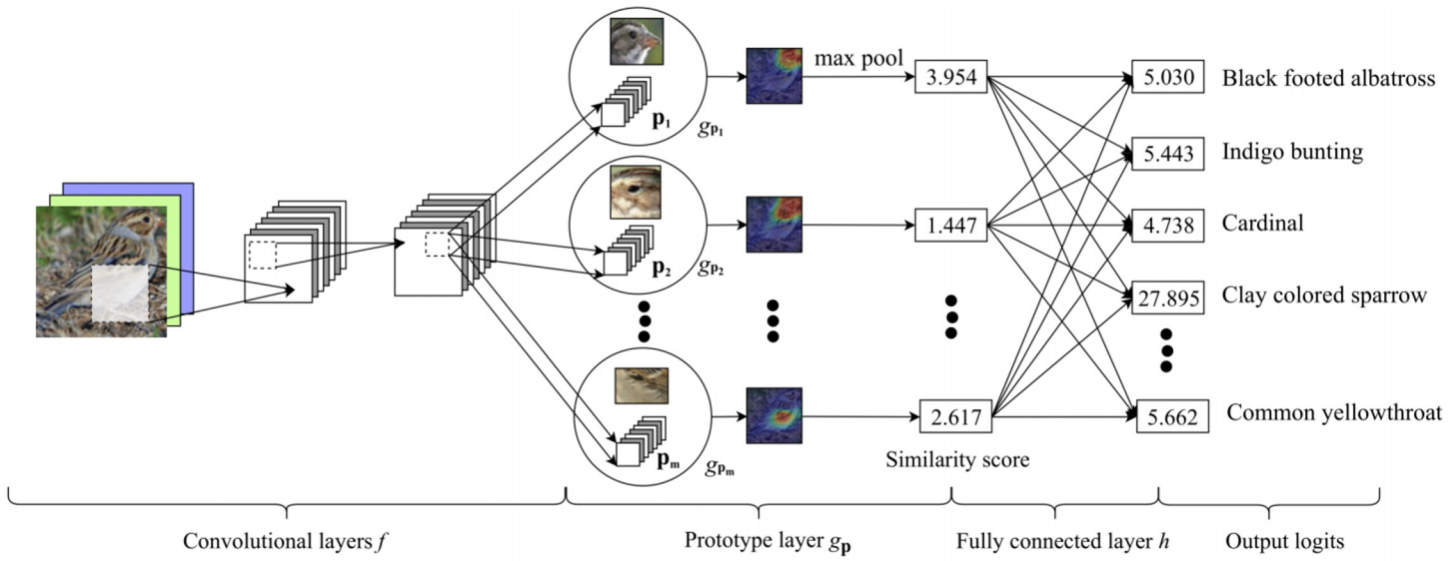


Leftmost: a test image of a clay-colored sparrow
Second column: same test image, each with a bounding box generated by our model -- the content within the bounding box is considered by our model to look similar to the prototypical part (same row, third column) learned by our algorithm
Third column: prototypical 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

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

What If? Learn image *prototypes* and combine them to output a final decision[1]

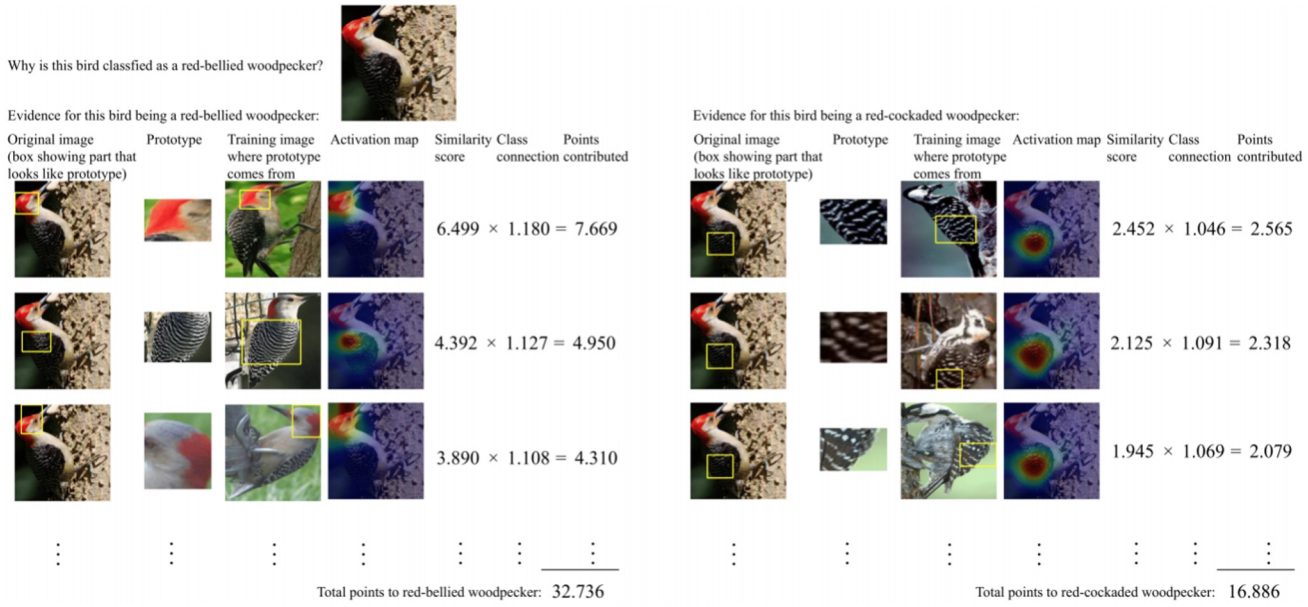
- The proposed model, *prototypical part network* (ProtoPNet), identifies "several parts of the image where it thinks that this part of the image looks like that prototypical part of some class, and makes its prediction based on a weighted combination of the similarity scores between parts of the image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



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

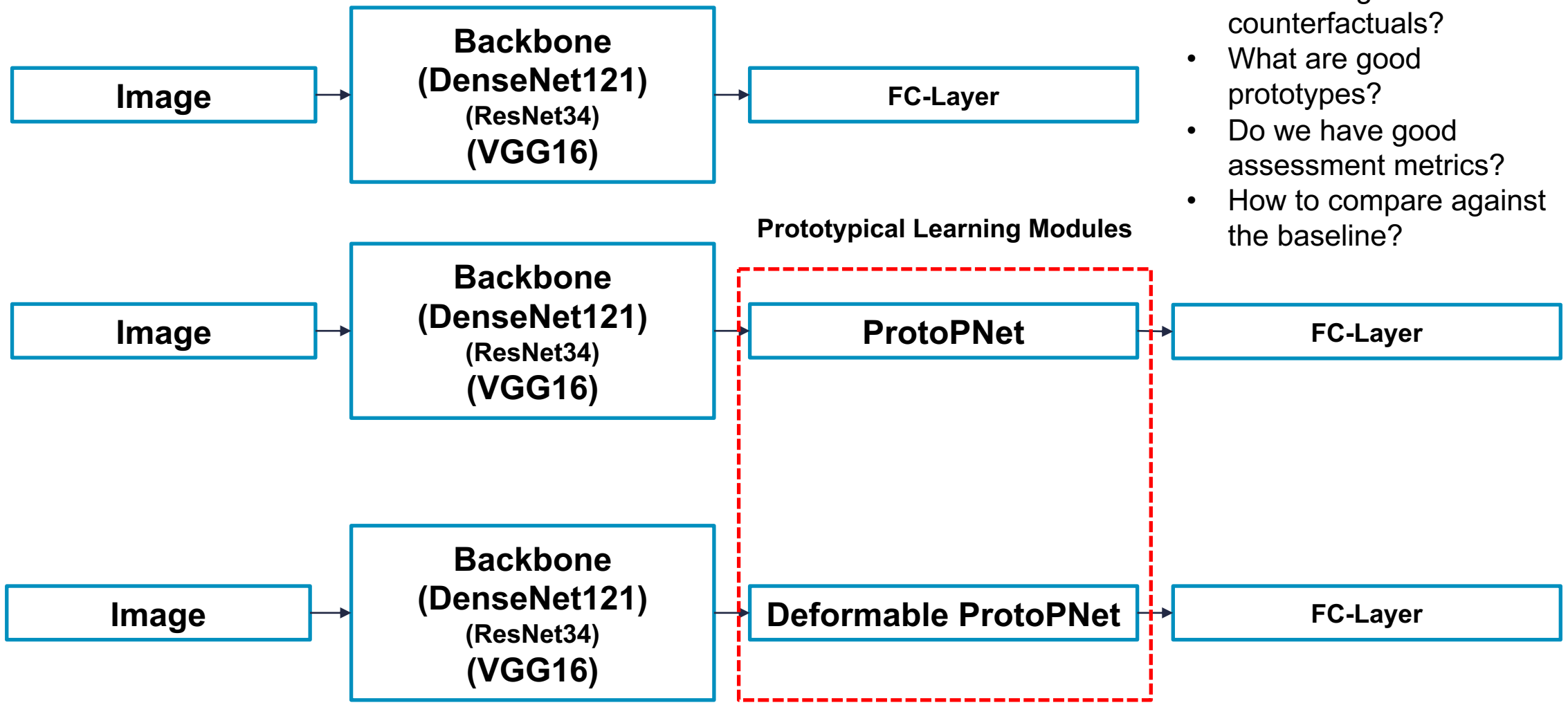
What If? Learn image *prototypes* and combine them to output a final decision[1]

- 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



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

Can we exploit prototypical learning to achieve good counterfactuals?



Key Ideas

- What are good counterfactuals?
- What are good prototypes?
- Do we have good assessment metrics?
- How to compare against the baseline?

5. Multimodalities of Madness: Combining Disparate Data Sources for the Perfect Spell

Medical data is multimodal, and that is awesome

- In the clinical context, **it is common to combine several image modalities** during the decision-making process (e.g., computed tomography, electroencephalography, magnetic resonance imaging, positron emission tomography)

- Recently, a comprehensive study[1] on data fusion strategies for image classification and segmentation reported that the **network trained with multi-modal images showed superior performance** compared to networks trained with single-modal images, and that **performing image fusion within the network** (e.g., fusing at convolutional or fully connected layers) is generally better than fusing images at the network output (e.g., voting)[1]

Sources: [1] Zhe Guo et al. "Deep Learning-based Image Segmentation on Multimodal Medical Imaging",

Multimodality means that data fusion will play a key role in our lives

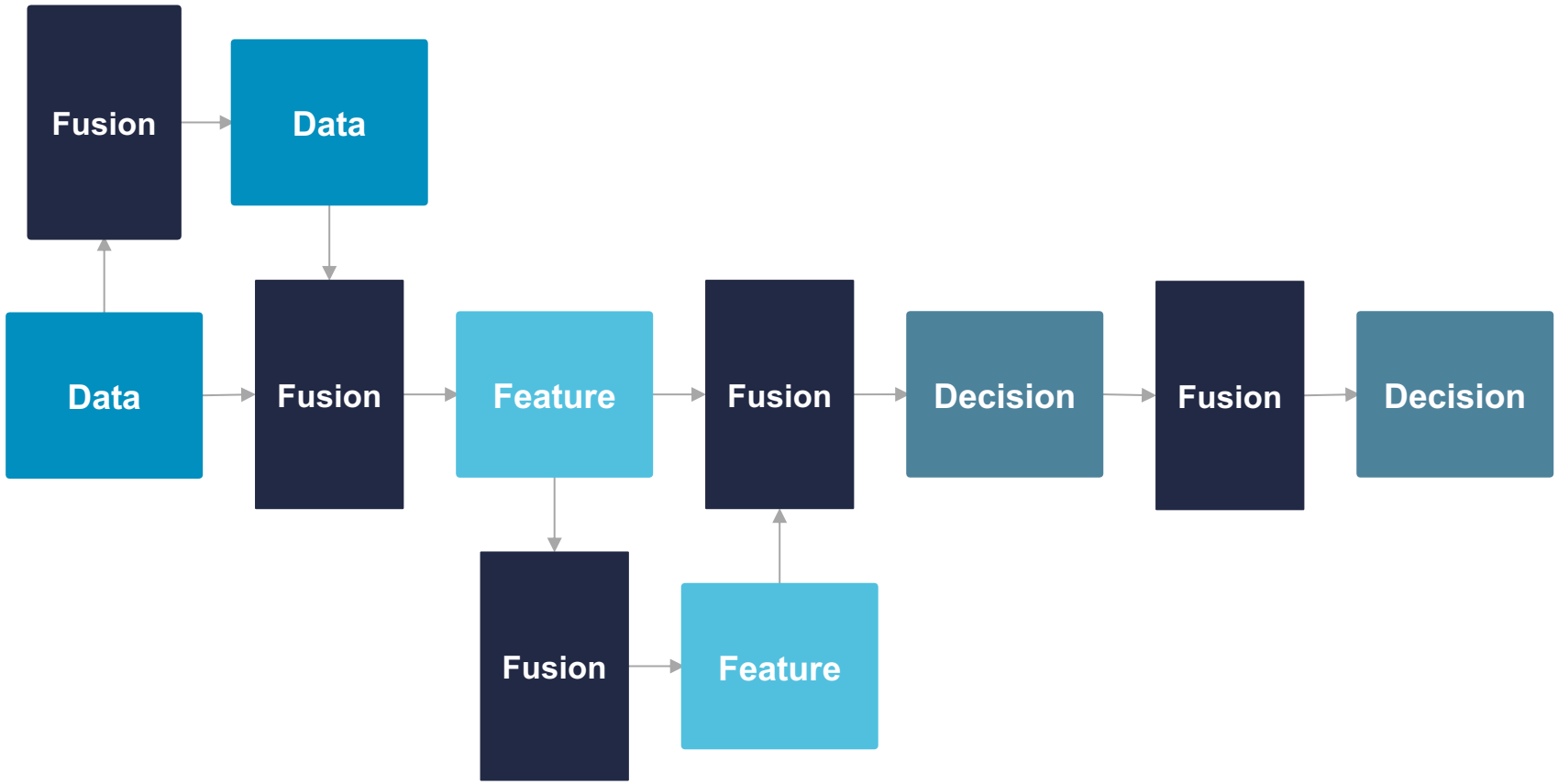


Figure: Different strategies of data fusion[1]

Sources: [1] R. C. Luo et al. "Multisensor fusion and integration: approaches, applications,and future research directions"

Leveraging disparate sources of information to increase interpretability in healthcare

Prostate Cancer Aggressiveness

Leveraging patient metadata and different MRI image modalities (T2, DWI and ADC), where one of the use cases is to predict cancer aggressiveness (based on ISUP/Gleason scores)

Biomarker Detection in Pathology

Derive novel biomarkers related to gene expression quantification using different clinical features of the whole slide image and patient

6. Days of Future Past: AI and Society

Responsible AI relies on fundamental principles

- **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

We need to design ethical and fair algorithms

- To facilitate trust (and increase transparency) in AI algorithms it is important to **ensure a priori that these models are interpretable**, and understand how decisions are made in the clinical context

- On the other hand, it is important to understand what the algorithms are already learning and to **evaluate the quality of such explanations** (e.g., understand if the algorithms are extracting relevant features for the clinical context)

- A different dimension of the application of AI in sensitive domains such as healthcare is the **development of ethical and fair algorithms**[1]
  - This strategy is supported by the new European Union's General Data Protection Regulation (EU-GDPR)[2] which advises that these **algorithms should be able to explain their decisions for the sake of transparency**

Sources: [1] Irene Y. Chen et al. "Treating health disparities with artificial intelligence", [2] Margot E. Kaminski "The Right to Explanation, Explained"

While keeping an attentive eye on the technologies that are shaping our lives

- Many entities are already leveraging their data sources to **optimize their inner processes or to develop new services or products**, thus enabling them to achieve a substantial competitive advantage[1]

- In the healthcare context, systems and algorithms need to go through a continuous **pipeline of validation and error assessment**

- Hence, it is **reasonable to accept that these technologies may need to be calibrated** to the data sources of the institutions that are integrating them into their information systems and that these algorithms **may have a continuous learning policy over time**

- Moreover, to assure **transparency, accountability and accessibility**, new regulatory frameworks will have to be developed to allow model adaptations that enable optimal performance while **ensuring reliability and patient safety**[2]

Sources: [1] Lucas Baier et al. “Challenges in the deployment and operation of machine learning in practice”,
[2] Farhad Maleki et al. “Machine Learning Algorithm Validation: From Essentials to Advanced Applications and Implications for Regulatory Certification and Deployment”

Post-hoc Memories of an Unfinished PhD in Interpretable Artificial Intelligence in Medicine

AI for Multi-center Data
Department of Radiology
The Netherlands Cancer Institute

October 27, 2023

Tiago Filipe Sousa Gonçalves
PhD Student at FEUP
Research Assistant at INESC TEC
Visiting PhD Student at MGH

https://tiagofilipesousagoncalves.github.io

INESCTEC
INSTITUTE FOR SYSTEMS AND COMPUTER ENGINEERING, TECHNOLOGY AND SCIENCE