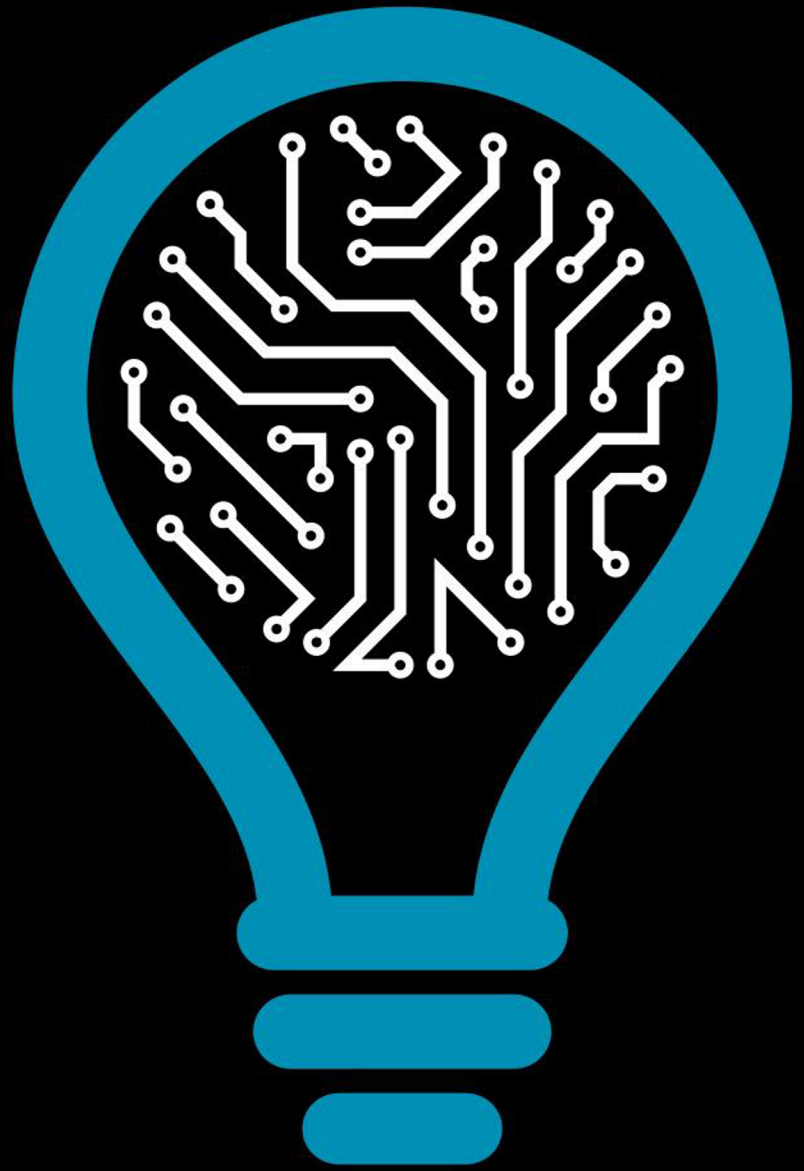
Interpretable Artificial Intelligence in Medicine: What Else?



XIV Symposium on Bioengineering | Porto (Portugal)

Made in Bio | March 18, 2023

Tiago Filipe Sousa Gonçalves

tiago.f.goncalves@inesctec.pt

INESCTEC

INSTITUTE FOR SYSTEMS AND COMPUTER ENGINEERING, TECHNOLOGY AND SCIENCE

INESCTEC

Outline

1. Interpretable Artificial Intelligence: The Bright Side of the Force

2. What's so interesting about medical data?

3. Towards the development of a Transparent New World

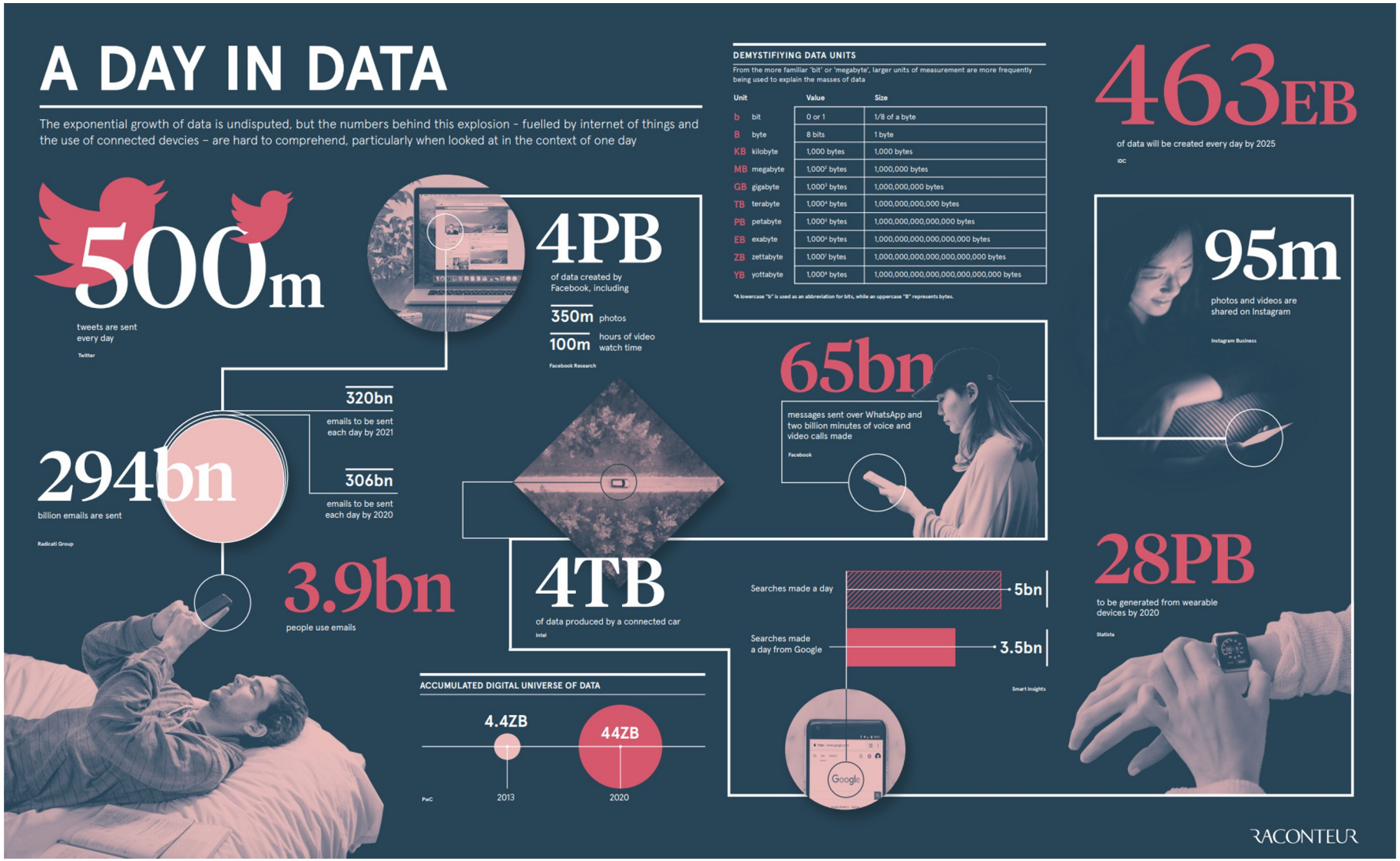
1. Interpretable Artificial Intelligence: The Bright Side of the Force

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"**

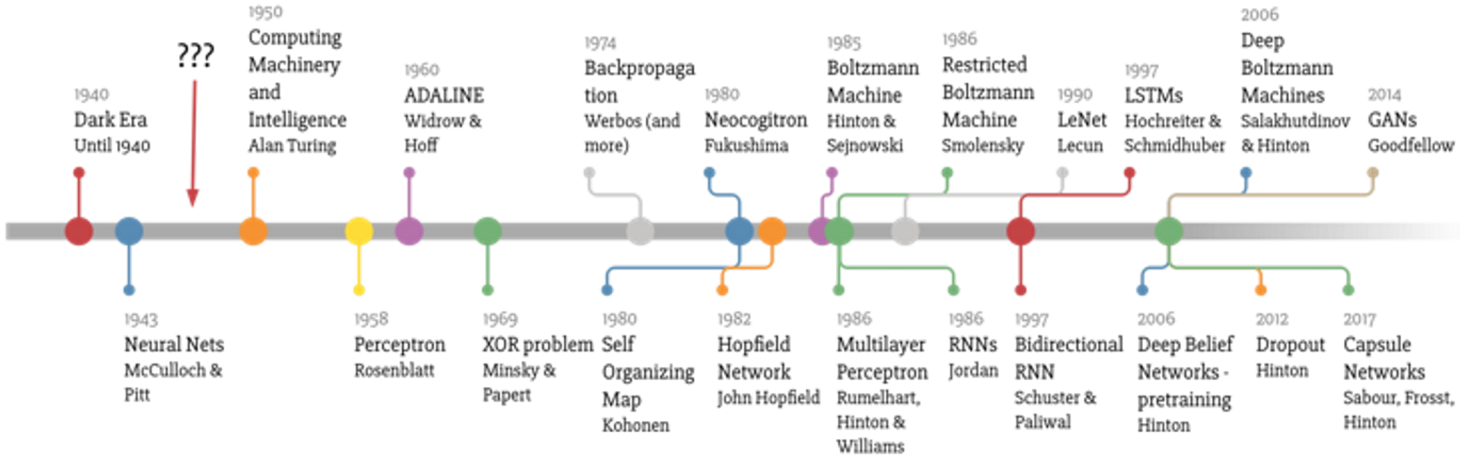


Figure - A (tentative) deep learning timeline (Image from [1])

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 beated 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]

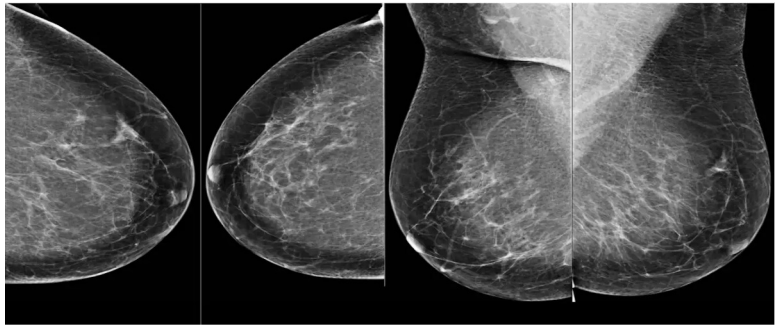


Figure - Medical Image Analysis: Mammograms (Image from [4])

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

Everything seems good except for the lack of transparency

- **The increase of available computational power and the democratised 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"

No worries! We are working on that!

- 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)

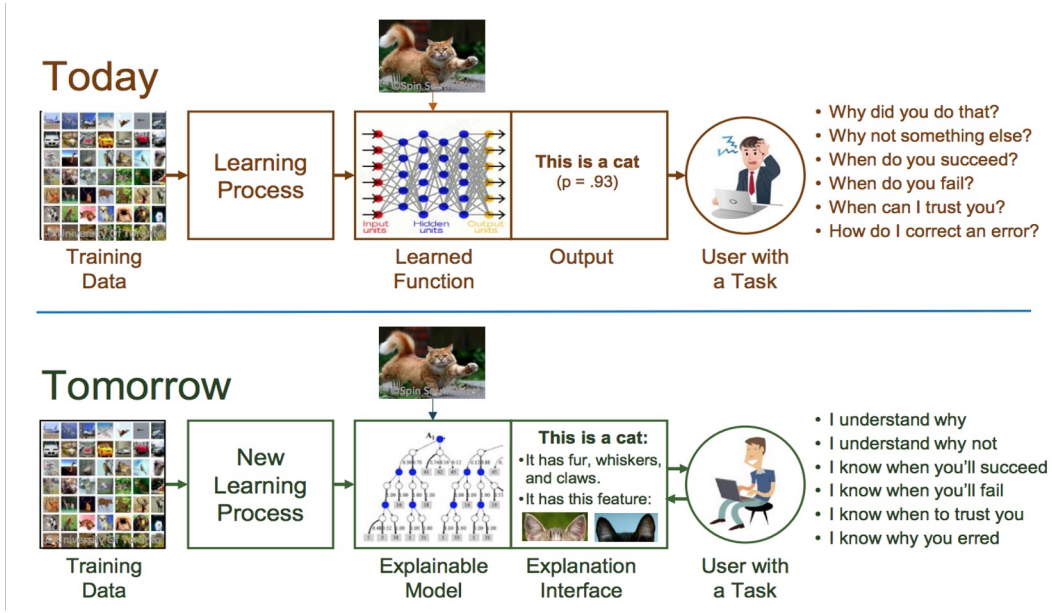


Figure - The future of machine learning algorithms
(Image from [1])

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"

INESCTEC

2. What's so interesting about medical data?

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

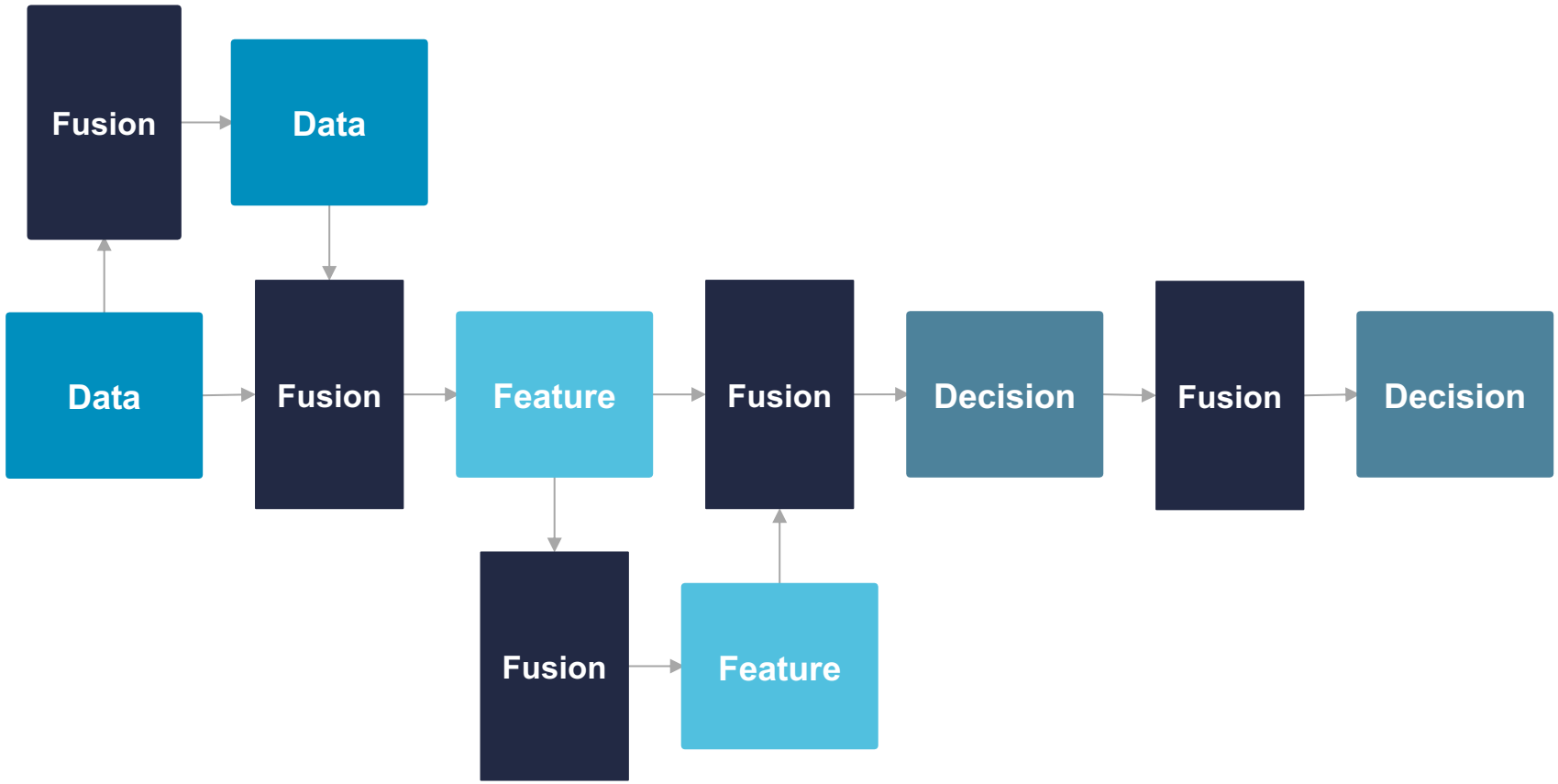


Figure: Different strategies of data fusion[1]

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

The medical world needs human-understandable explanations

- 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]

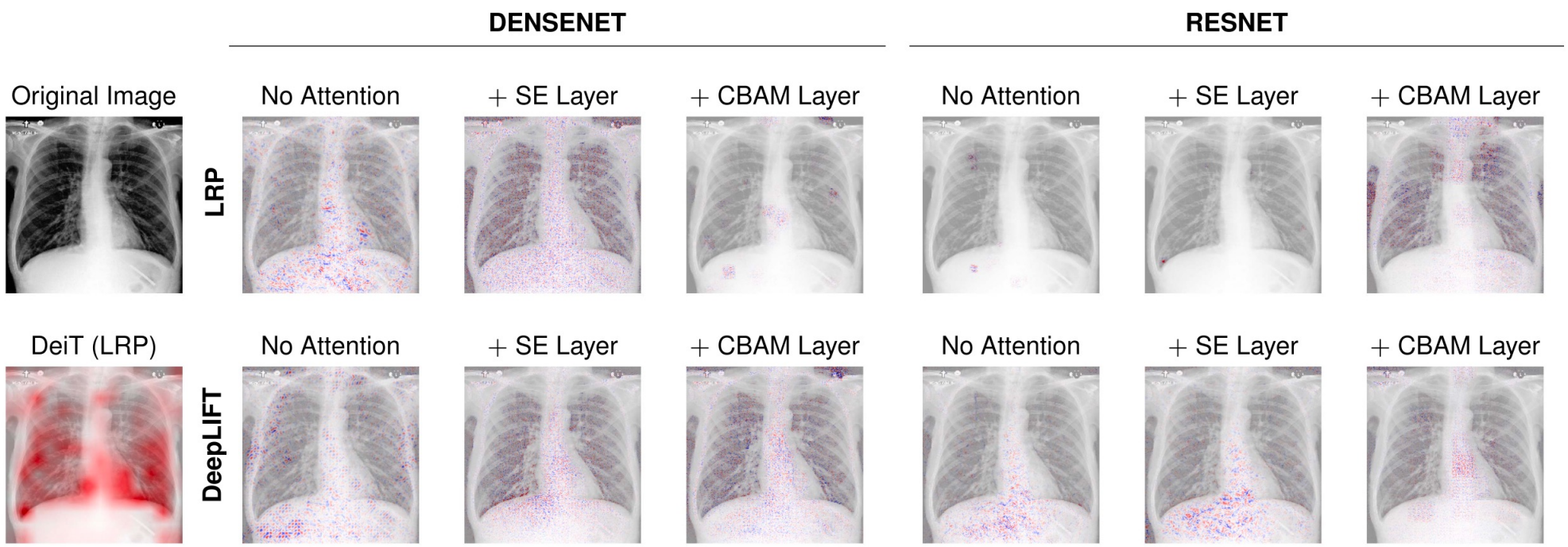
- There has been a huge effort into the **development of post-model strategies** to explain the behavior of black box models, however, the outputs of these algorithms are prone to **subjective evaluation**, **may be misleading**[2] or **fooled**[1]

- 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 do you see?[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?"

The healthcare tech market is full of opportunities and requires interpretable artificial intelligence

- Interpretability is already playing its role in the pipelines of machine learning deployment: researchers and developers are **using interpretability techniques to validate and debug** their models before deployment[1, 2]

- Regarding the availability of end-user software that contain **machine learning algorithms for medical applications**, we point to the popularity of:
  - Software for the analysis of volumetric medical images
  - Software for the development and creation of **DICOM pipelines and servers**
  - Software for the **annotation** and **segmentation** of medical images
  - Software for the **automatic classification** of medical images
  - Software for the **automatic analysis of electronic health records** of patients to generate **diagnosis** and **recommendations**

Sources: [1] Umang Bhatt et al. "Explainable Machine Learning in Deployment", [2] Isaac Lage et al. "An Evaluation of the Human-Interpretability of Explanation"

3. Towards the development of a Transparent New World

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]

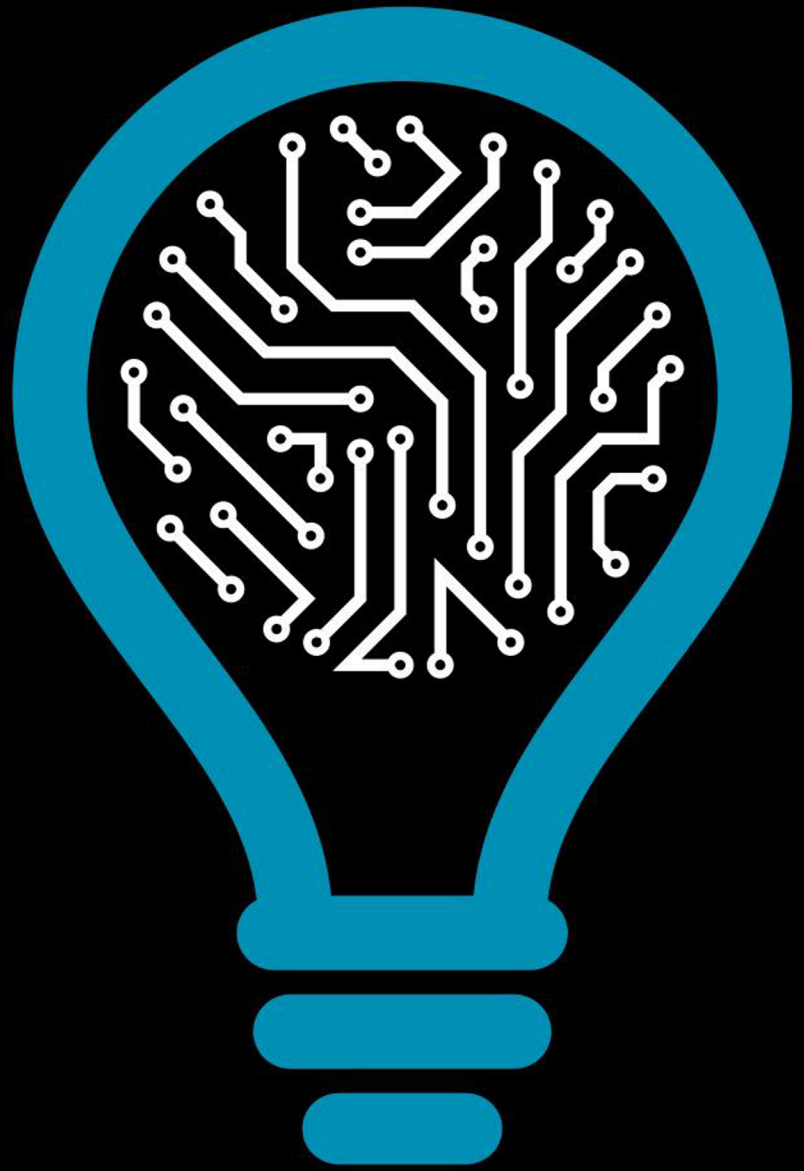
- In the healthcare context, systems and algorithms need to go through a continuous **pipeline of validation and error assessment**

- 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"

Interpretable Artificial Intelligence in Medicine: What Else?



XIV Symposium on Bioengineering | Porto (Portugal)

Made in Bio | March 18, 2023

Tiago Filipe Sousa Gonçalves

tiago.f.goncalves@inesctec.pt

INESCTEC

INSTITUTE FOR SYSTEMS AND COMPUTER ENGINEERING, TECHNOLOGY AND SCIENCE