UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Dipartimento Interateneo di Fisica Michelangelo Merlin

Next_AIM 2023 Milano

Explainable Artificial Intelligence (XAI)

Alessandro Fania, Nicola Amoroso, Roberto Bellotti, Antonio Lacalamita, Alfonso Monaco, Sabina Tangaro

13/02/2023

Bari Applied Physics (INFN Group-5)

Overview

- What is Explainability
- Why Explainability is important
- Grad-CAM
- SHAP
- Application on Editing Data
- Conclusion

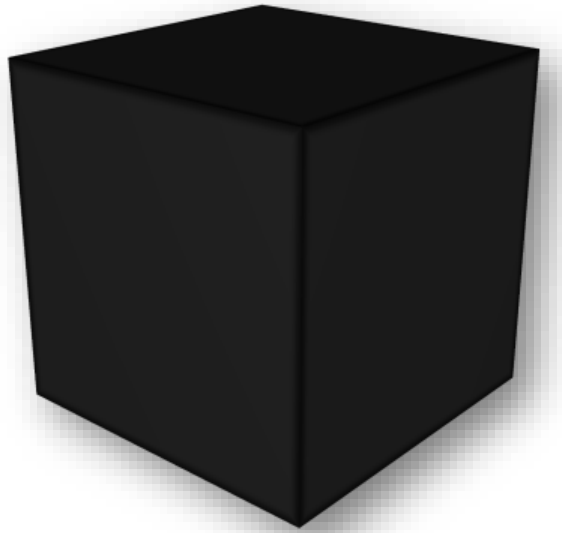
What is Explainability?

Machine Learning and Deep Learning models are difficult to understand

They are often treated as **Black Boxes**

Black box

$$X = \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_n \end{bmatrix}$$



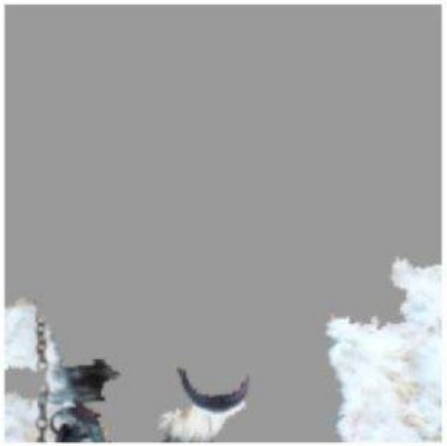
Y

Input data

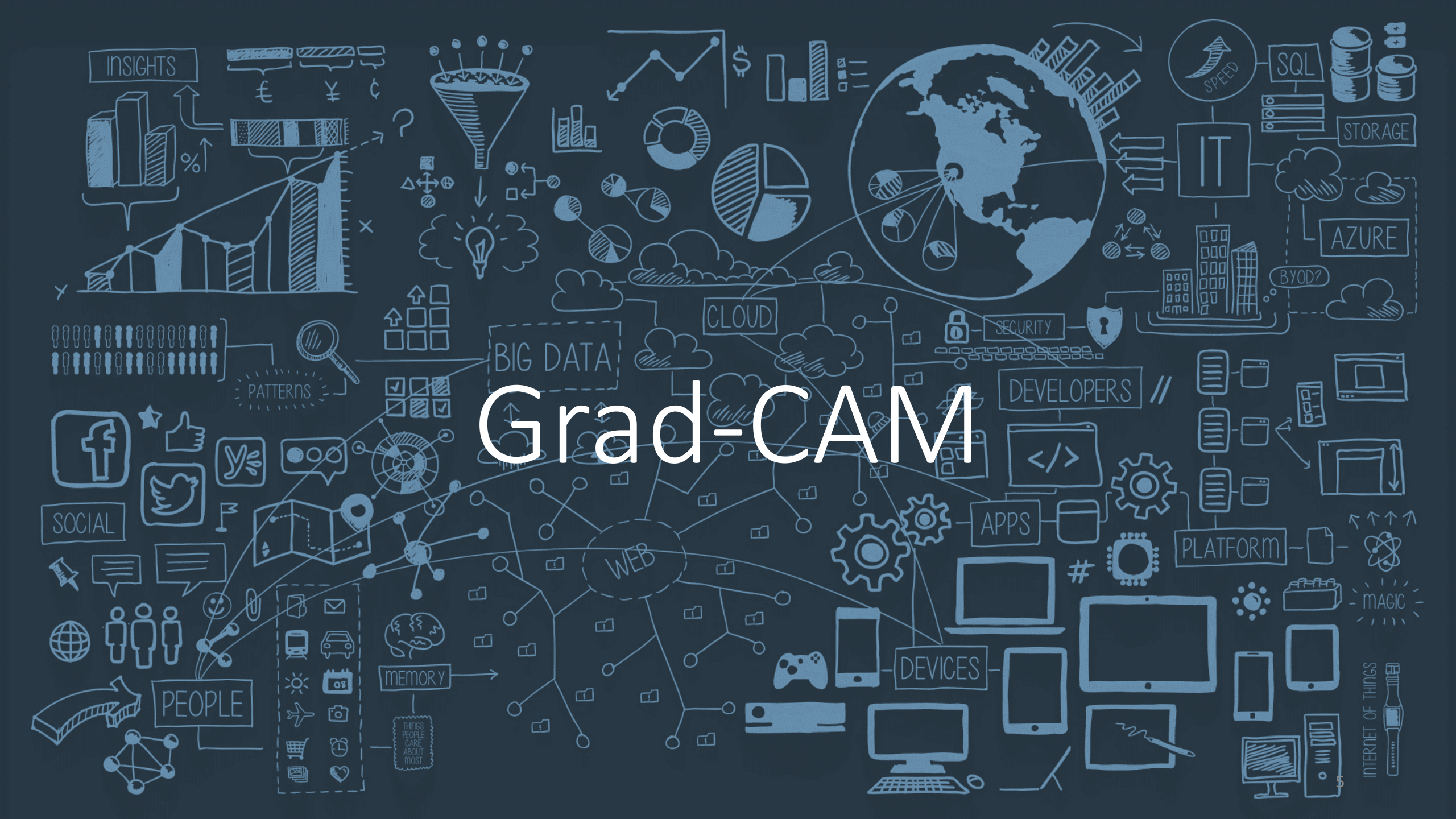
Predicted output

Why Explainability is so important?

- Helps analysts to understand system outputs simply and quickly.
- Explainability can provide recommendations and spot anomalies for analysts to investigate
- Sometimes AI can give an output that's correct but for the wrong reasons.
- Likewise, Explainability makes possible to understand why a mistake was made and even train the system to stop it from happening again.
- This driver for Explainability provides some overlap with the General Data Protection Regulation (GDPR): the customer has the right to obtain explanations.
- The European Commission recently published the first draft of its Artificial Intelligence Regulation which stipulating requirements around AI and Explainability.



-a husky (on the left) is confused with a wolf, because the pixels (on the right) characterizing wolves are those of the snowy background.


INSIGHTS
SPEED
SQL
STORAGE
IT
AZURE
BYOD?
CLOUD
SECURITY
BIG DATA
PATTERNS
DEVELOPERS
SOCIAL
APPS
PLATFORM
MAGIC
WEB
MEMORY
PEOPLE
DEVICES
INTERNET OF THINGS
THINGS PEOPLE CARE ABOUT MOST

Grad-CAM

CAM-model

What is a CAM model?

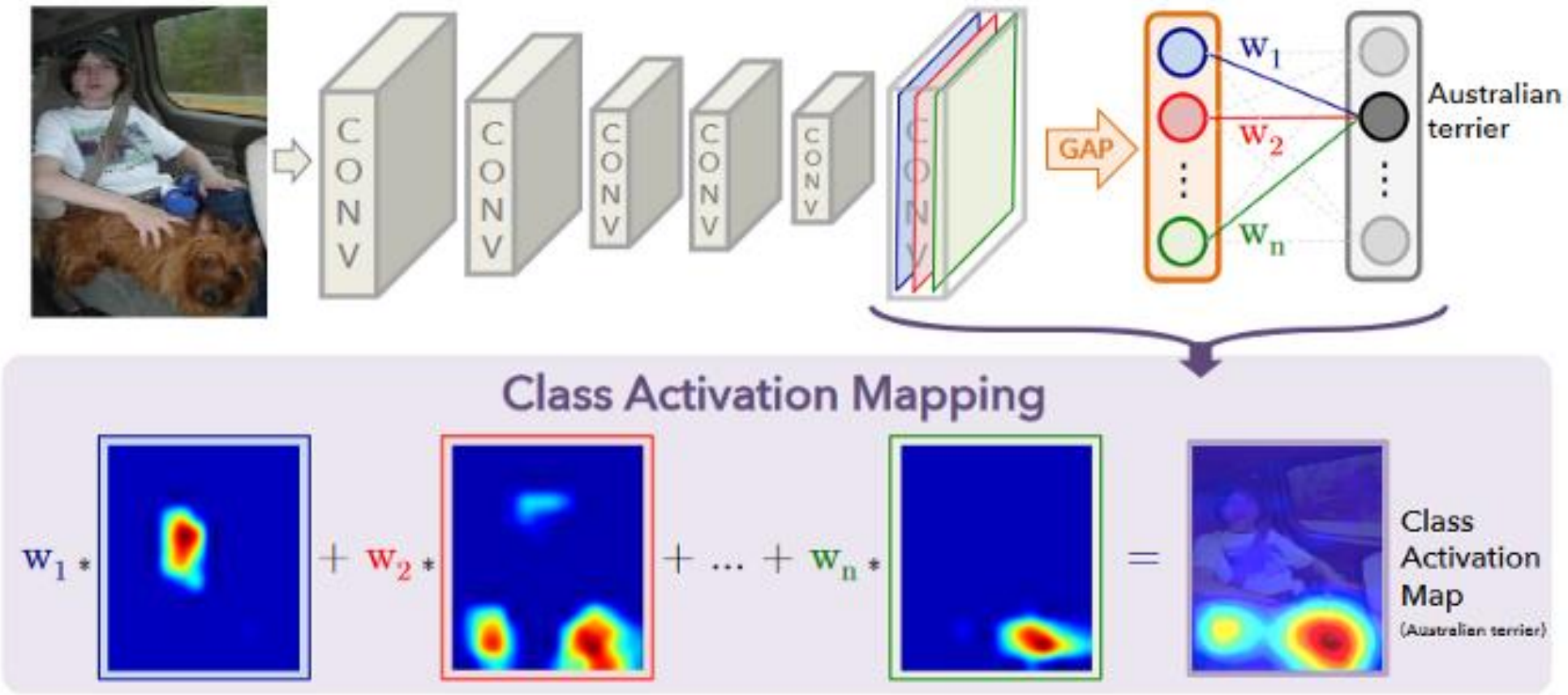
Class Activation Maps (CAM) are a technique to get the discriminative image regions used by a CNN to identify a specific class in the image.

In other words, a class activation map lets us see which regions in the image were relevant to this class.

-Area of the image that explains a prediction "Dog"

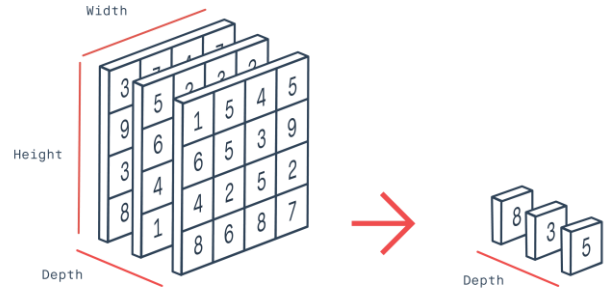
The idea is to use the feature maps of a CNN model as weight to explain a certain prediction

CAM-Architecture



CAM-Algorithm

We can then express the c_{th} output of the network as:



$$y^c = \sum_k \underbrace{w_k^c}_{\text{class feature weights}} \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{A_{ij}^k}_{\text{feature map}}$$

We finally get the *class map*: $L_{\text{CAM}}^c = \underbrace{\sum_k w_k^c A^k}_{\text{linear combination}}$.

- Y_c represents the score of the c_{th} class
- Z is the number of pixels in feature map

CAM-Cons

CAM only works on architectures that have Global Average Pooling (GAP) as a layer before the Dense that deals with the classification:

Limitations:

- The model needs to be modified in order to use CAM.
- The modified model needs to be retrained, which is computationally expensive.
- Since fully connected Dense layers are removed. the model performance will surely suffer. This means the prediction score doesn't give the actual picture of the model's ability.
- The use case was bound by architectural constraints, i.e., architectures performing GAP over convolutional maps immediately before output layer.

Grad-CAM

A possible solution is to use **backpropagation** to calculate the weights of the maps.

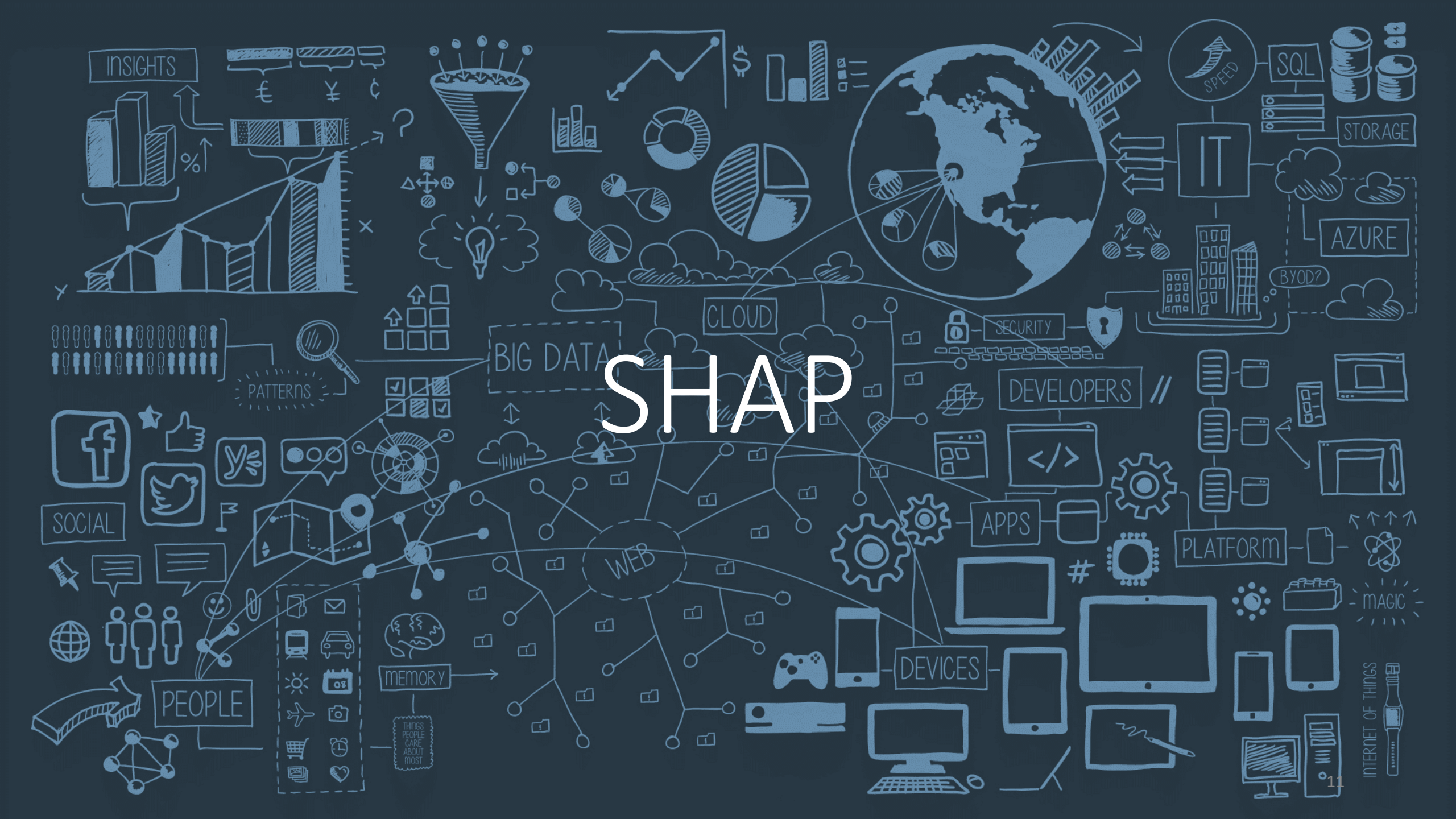
$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

Gradient Class Activation Map (Grad-CAM)

Next, a ***ReLU*** function is applied to zero the negative values of the gradient.

$$L_{\text{Grad-CAM}}^c = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

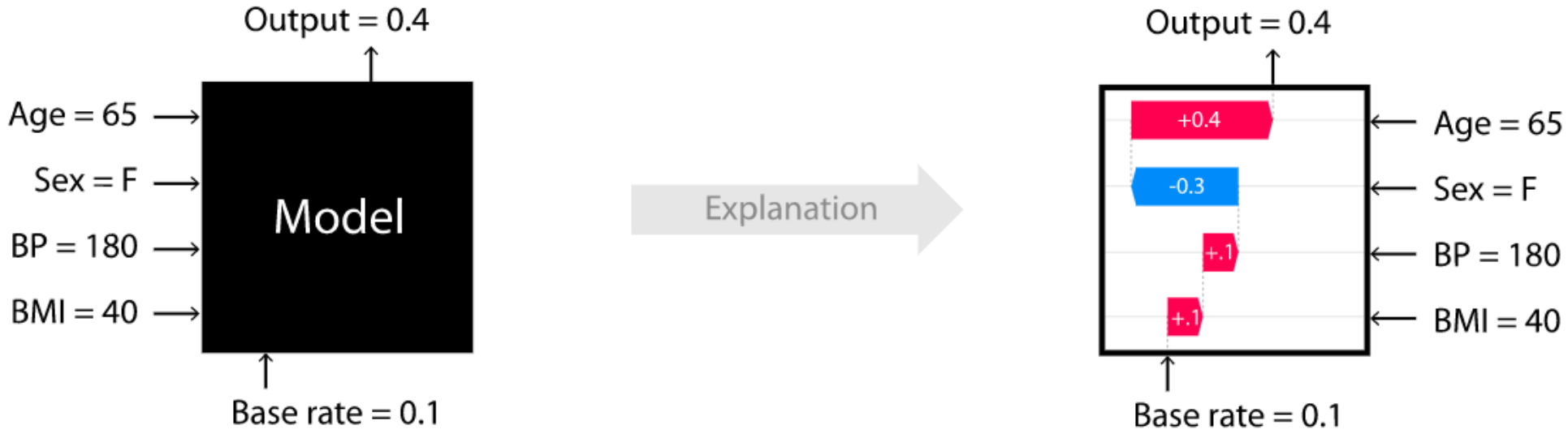

INSIGHTS
SPEED
SQL
STORAGE
IT
AZURE
BYOD?
CLOUD
SECURITY
BIG DATA
PATTERNS
DEVELOPERS
SOCIAL
APPS
PLATFORM
WEB
MAGIC
PEOPLE
MEMORY
THINGS PEOPLE CARE ABOUT MOST
DEVICES
INTERNET OF THINGS

SHAP

SHAP-Introduction

SHAP is a model agnostic explainer:

- Its purpose is to "imitate" the model used.
- It gives an understandable explanation of a **local prediction** of a model by assigning to each feature a value, based on the concept of ***the worth of a coalition*** (Lloyd Shapley in 1951).



Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems* 30 (2017).

SHAP-How it works

- Let's consider a game in co-operative with *m* numbered players and call *F* the set of such players.

$$\mathrm{F} = \{1, 2, 3, 4, \dots, m\}$$

- We then define an *S* coalition as a subset of *F*, which also includes the empty set without players.
- An example of the possible coalitions with 3 players

$$\{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$$

- Let us now define a function ν, called a **characteristic function**, which associates each coalition with a real number. The value *ν(S)* will be called the **worth of the coalition** S and represents the total gain obtained by the coalition if the members act together (example of the calculation of the value of {3})

$$\nu(\{1,2,3\}) - \nu(\{1,2\})$$

SHAP-How it works

Considering the number of permutations and summing up on all other possible combinations we get:

$$\phi_i = \sum_{S \subseteq F - \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} (\nu(S \cup \{i\}) - \nu(S))$$

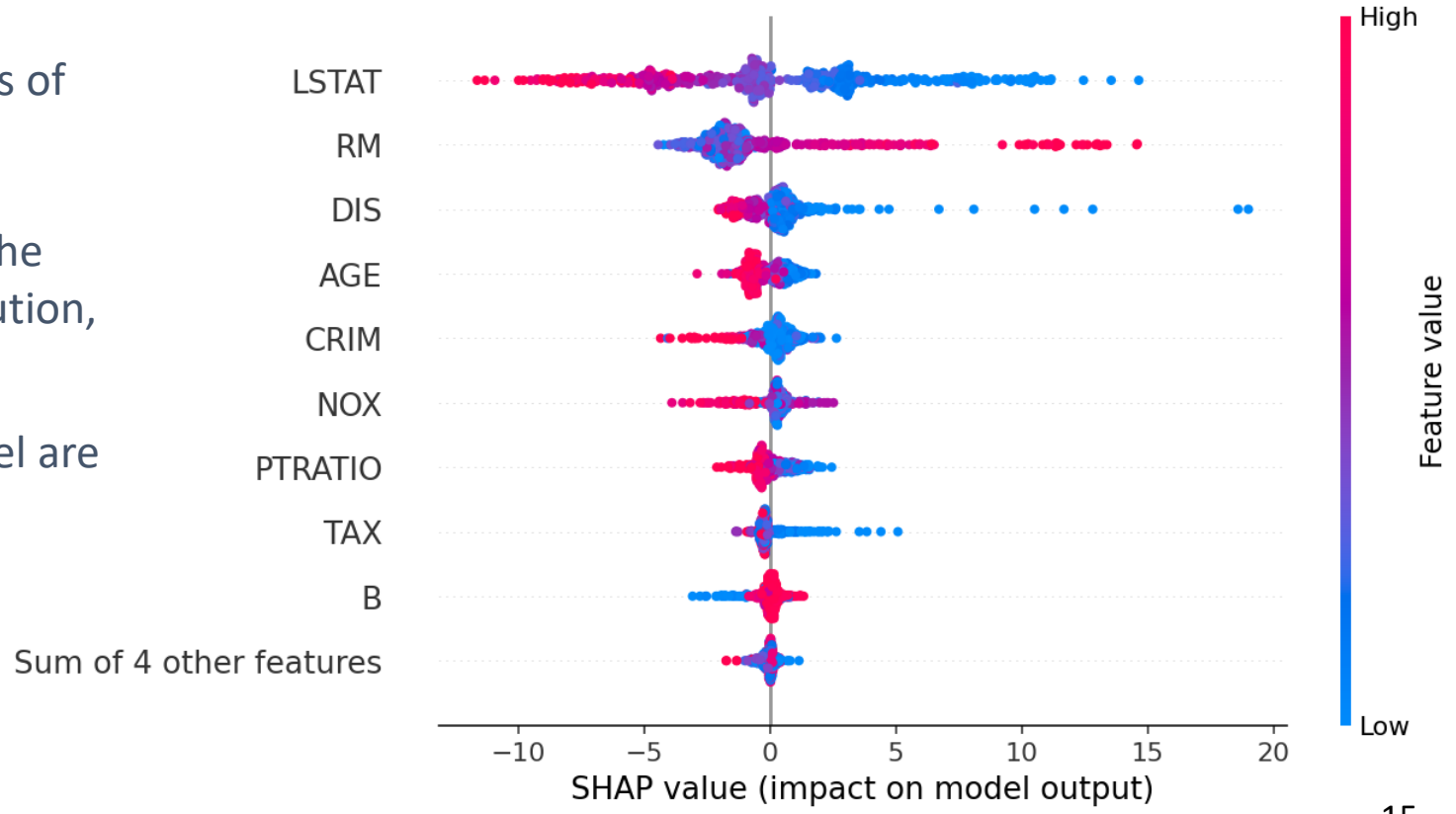
where $|F|!$ is the total number of permutations of the grand coalition, *S* indicates the coalition and $S \cup \{i\}$ indicates the coalition with the addition of *i*.

The ϕ_i value is called **Shapley Value** and represents the average contribution of the player, or variable, *i*.

SHAP-Plot

The SHAP library allows the display of SHAP values for each feature of the model.

- In the example opposite, the distributions of the SHAP values of a model are shown.
- The values in red refer to high values of the feature, compared to the starting distribution, vice versa the blue values to low values.
- The most impactful features for the model are placed at the top.
- This plot refers to a test sample but can also be used for a local instance.

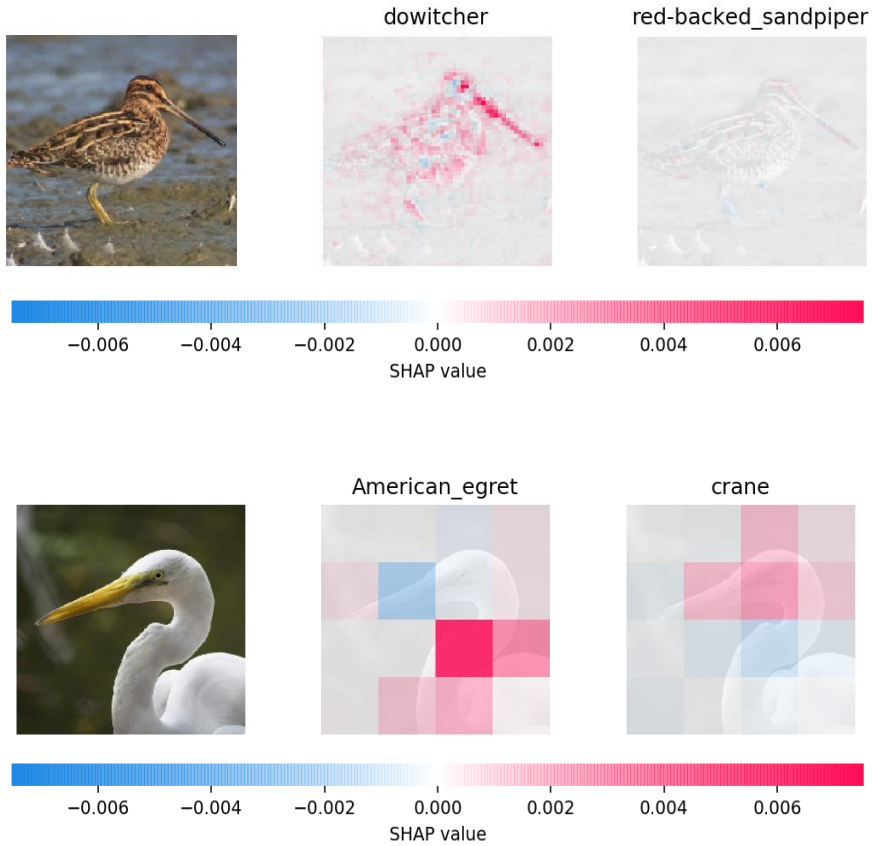


SHAP for Deep Learning: DeepExplainer vs GradientExplainer

The SHAP model is also adapted for the explanation of deep learning models.

- The function in question is **DeepExplainer**:
  - It can have both *tabular* and *image* data as input.
  - In the case of images, *pixels* will be considered as features to which to assign the SHAP values

- Another very similar function is **GradientExplainer:**
  - it is possible to assign importance to the various intermediate layers of the neural network.
  - In the case of a CNN the SHAP value is assigned to the pixels of the feature maps.

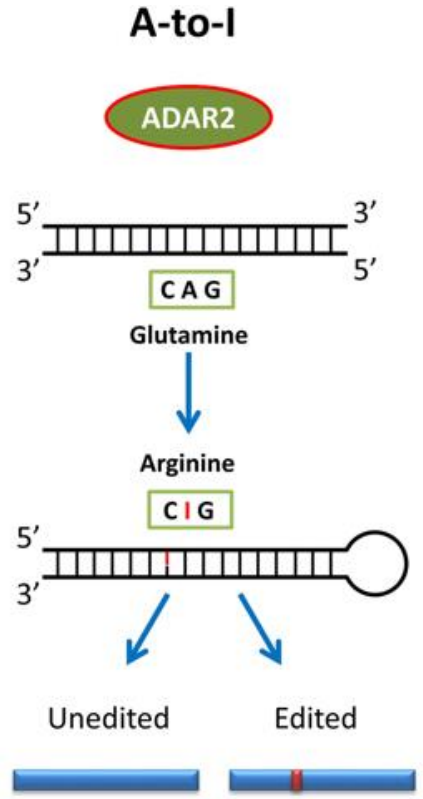


Presented during ML-INFN Hackathon 2022, Bari

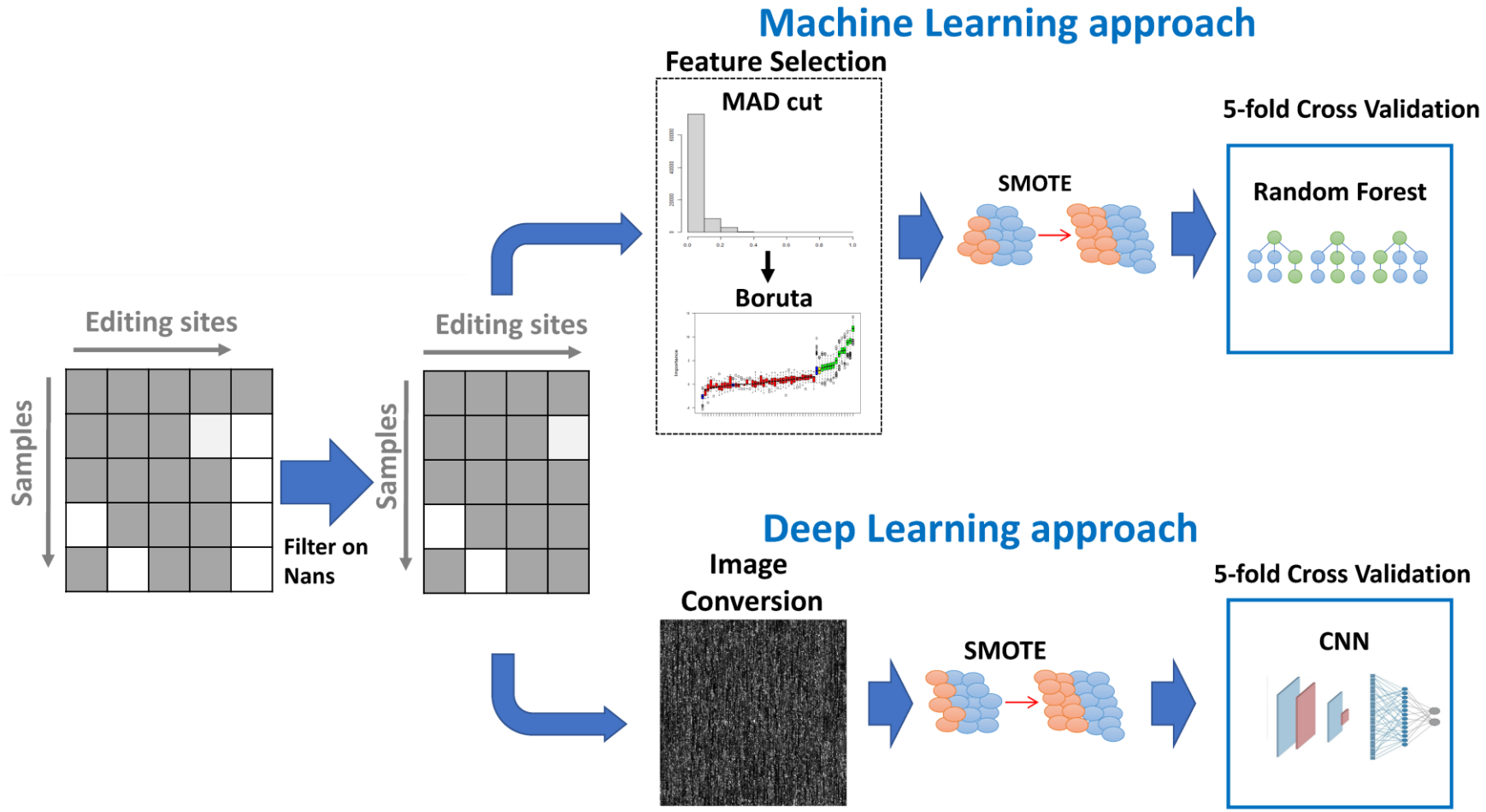
Gender classification based on RNA editing levels through machine and deep learning techniques

RNA Editing

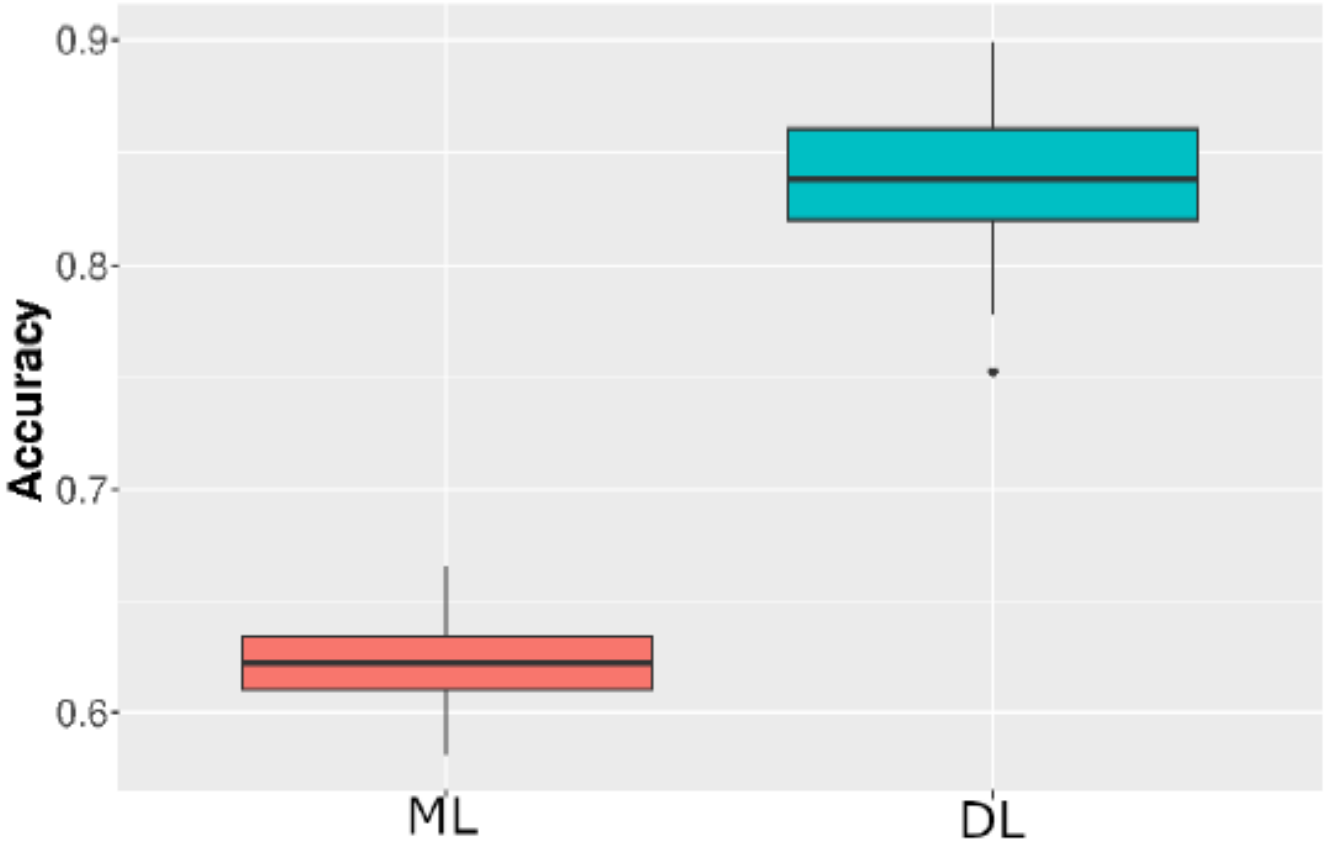
- **RNA editing** is a co/post-transcriptional process that involves nucleotide insertions, deletions and substitutions at specific positions in cellular transcripts
- **A-to-I editing** (contributes to nearly 90% of all RNA editing events)
- Our study is one of the **first** to use editing data and exploit the potential of advanced learning techniques to study the connection between **RNA editing** and **biological sex**
- Our findings could help to shed **light** on A-to-I RNA editing regulation and its connection with human physiology



Schematic overview of the proposed analyses

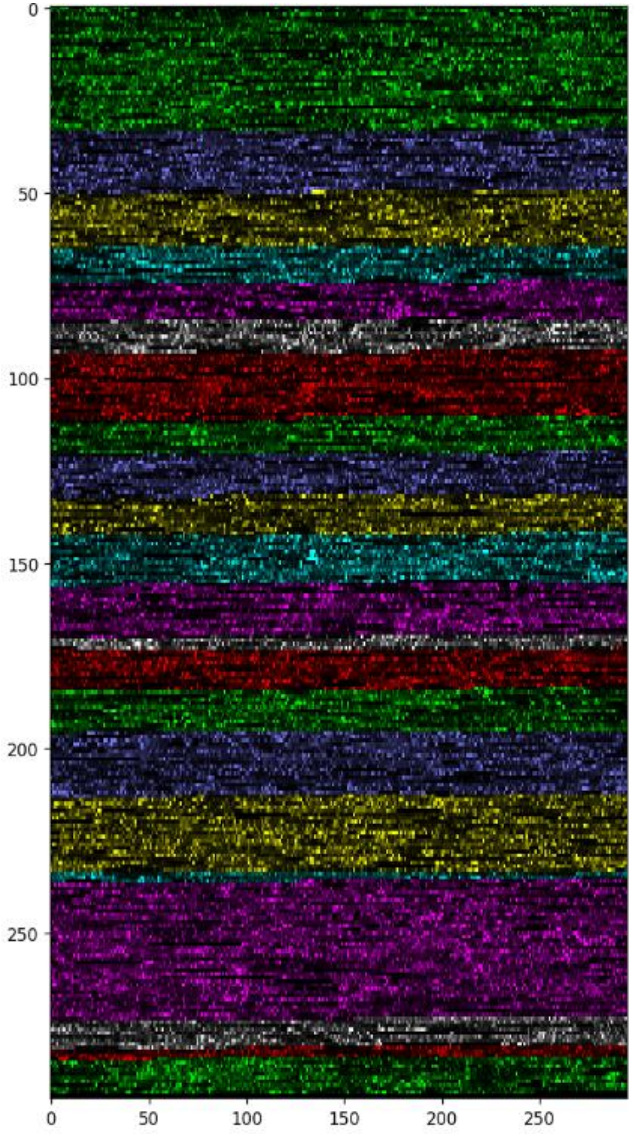


Machine Learning vs. Deep Learning: Results

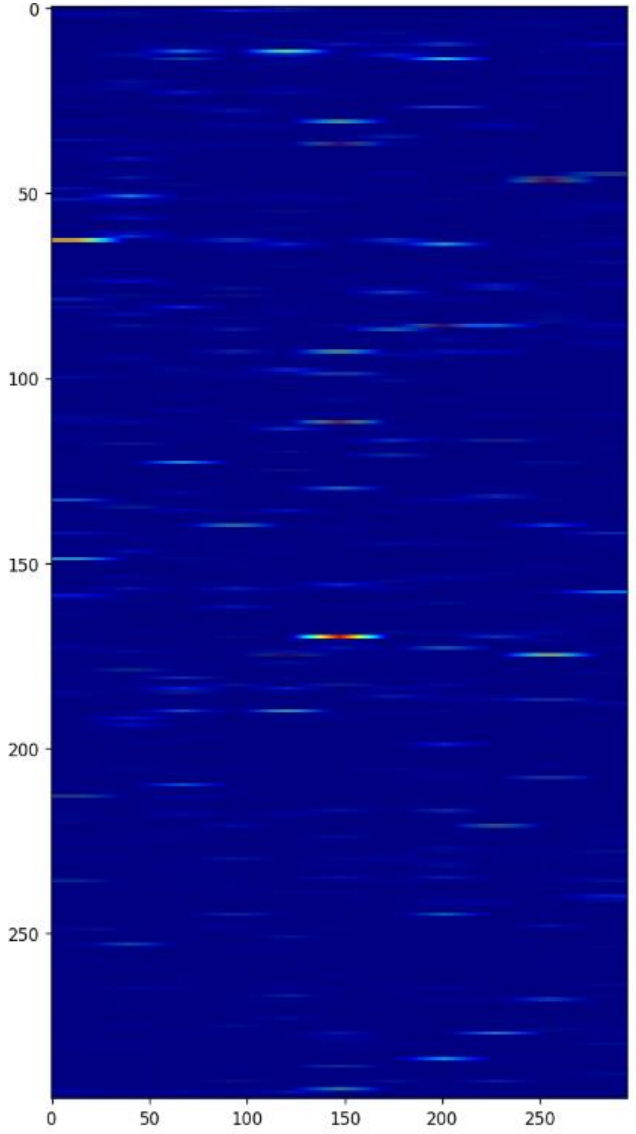


Grad CAM: Results

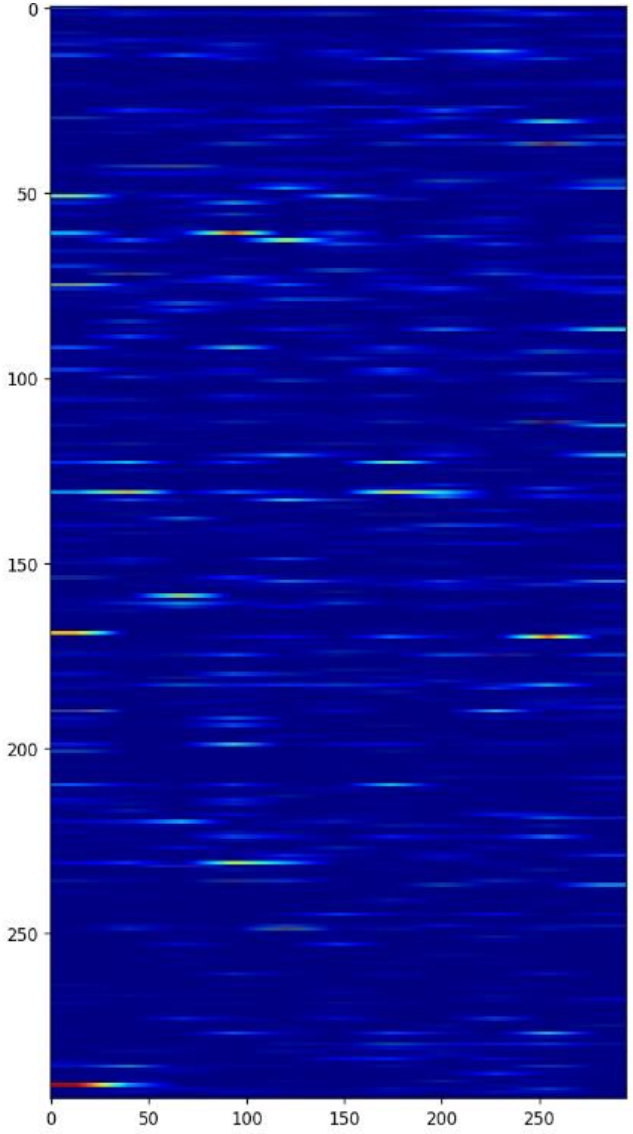
Original image



Female

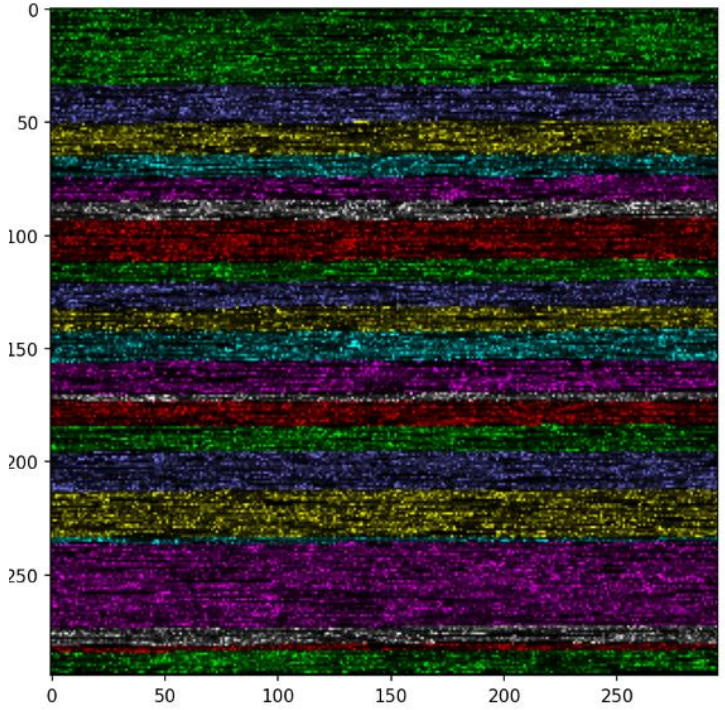


Male

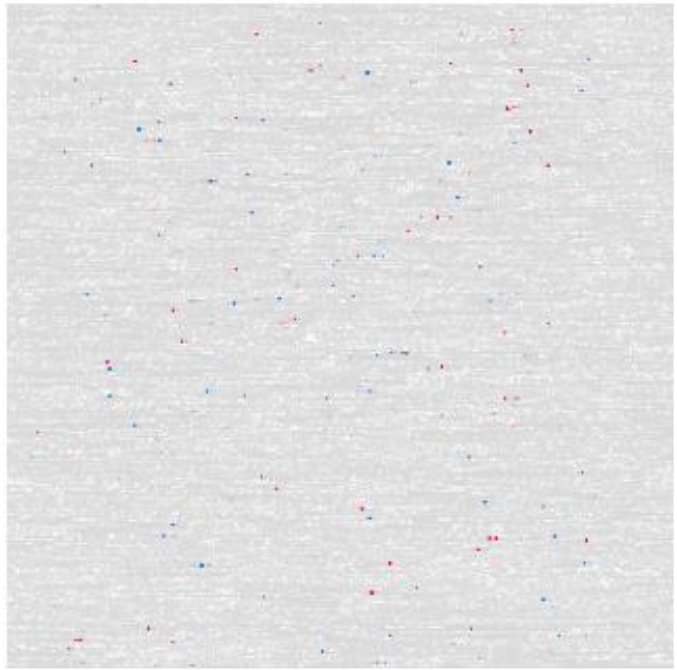


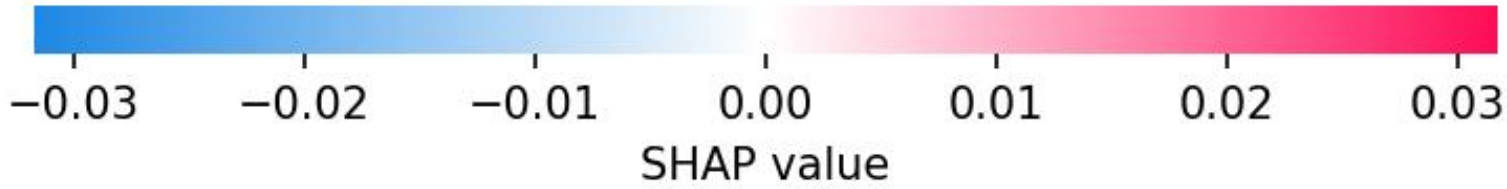
DeepExplainer : Results

Original image



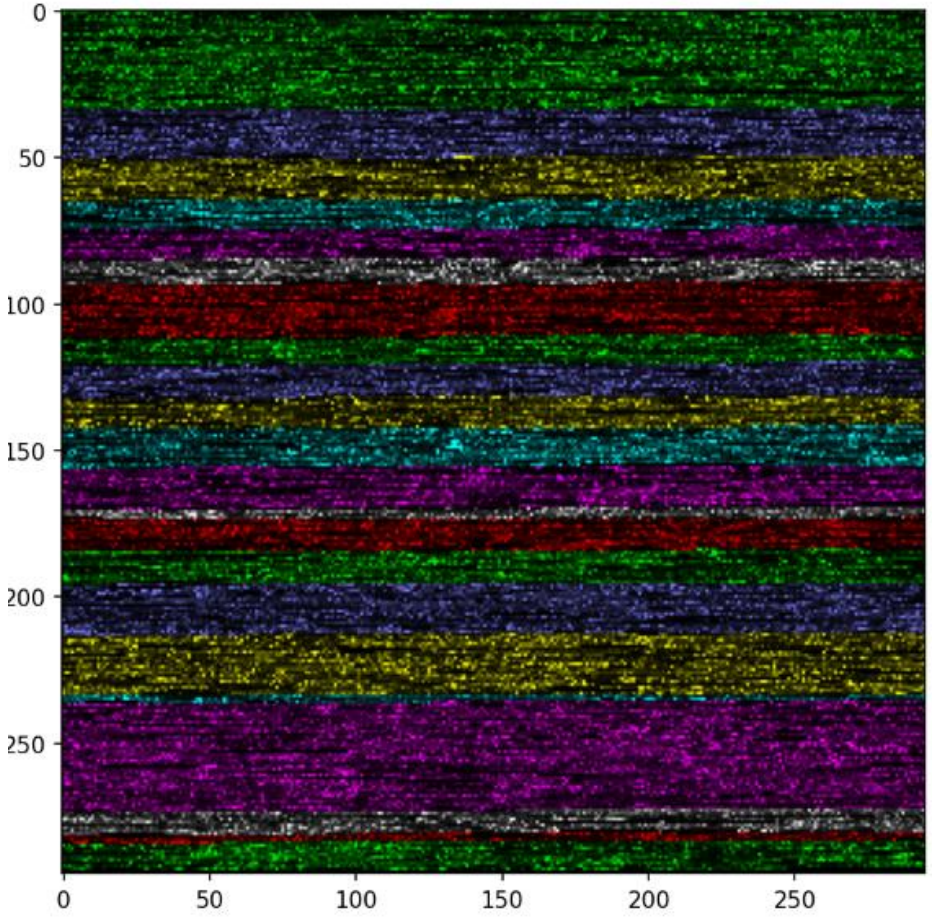
Male vs Female



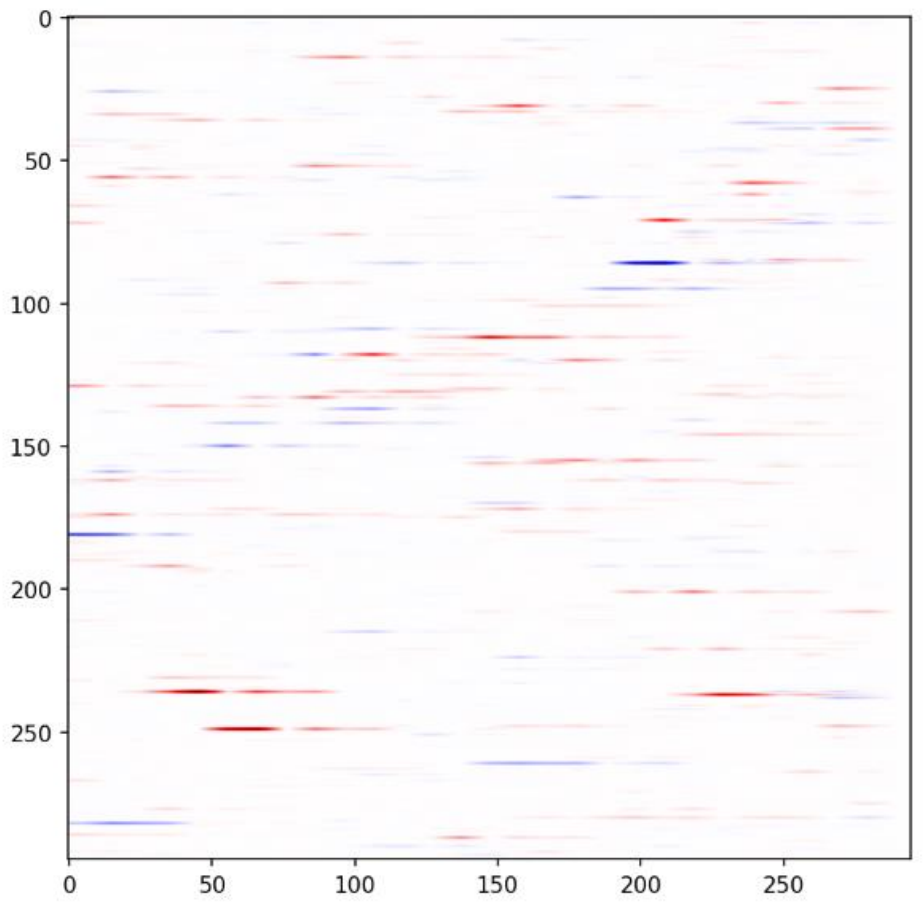


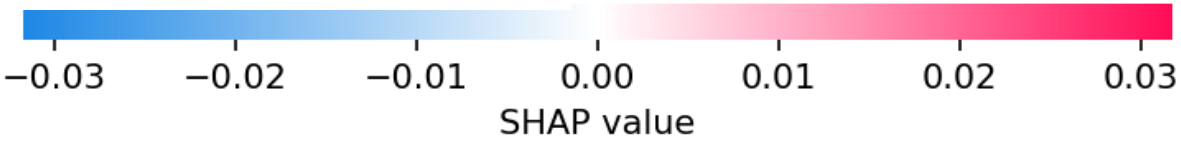
GradientExplainer: Results

Original image



Male vs Female





Grad-CAM vs SHAP?

Grad-CAM and SHAP are different explainable methods:

Grad-CAM:
- Based on Feature maps
- Consider only positive gradients.
- Applicable only on CNNs.

SHAP:
- Pixel based.
- Give explanation of how the value of the feature influenced the prediction.
- Applicable on several models.

In conclusion, which is better?
A possible answer:
- The best way is to consider the *consensus method*.
- If not possible, consider to use the most reasonable one.

Thanks for your attention

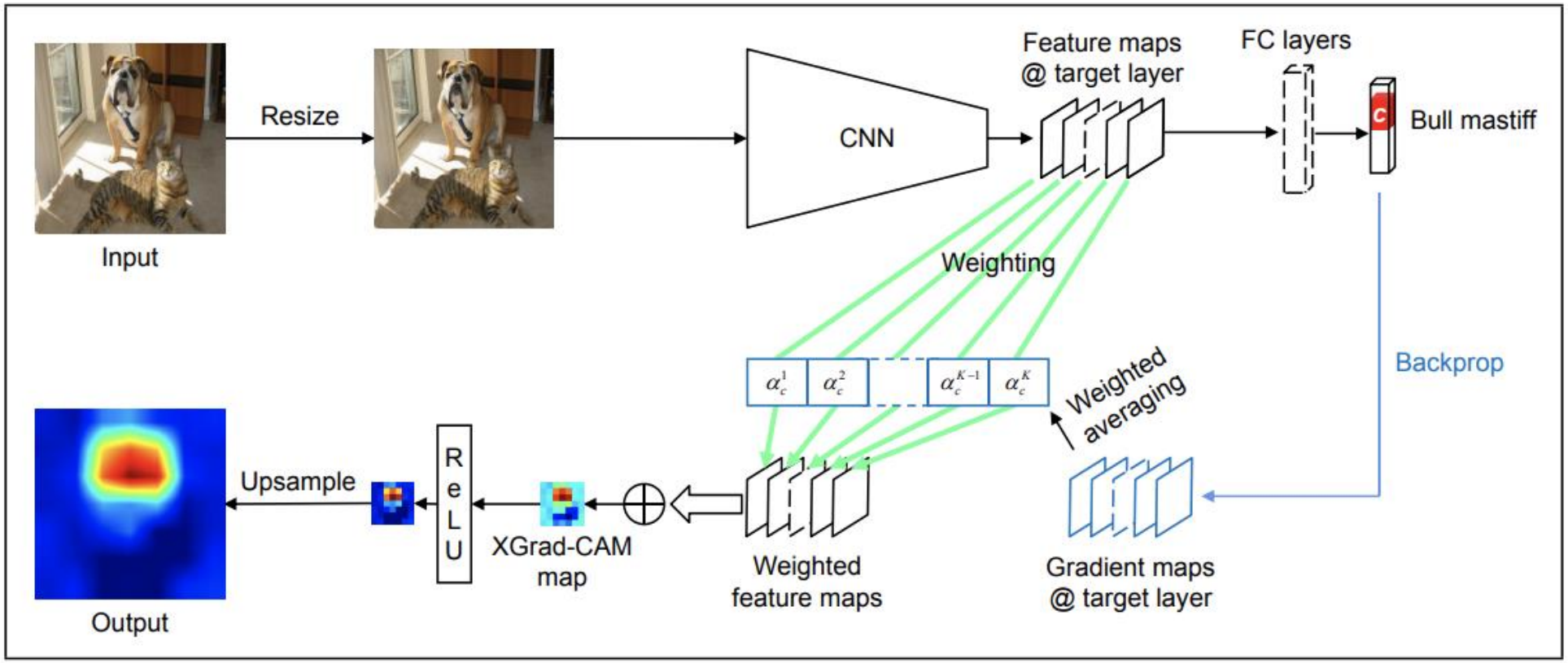
Questions?

What is Explainability?

«Explainability is the degree to which a human can understand the cause of a decision»

Tim Miller-Explanation in artificial intelligence: Insights from the social sciences

Grad-CAM Architecture



RNA Editing Problematics

- RNA Editing data are classifiable as **Big Data**: millions of sites are known (**REDIportal**)
- RNA Editing datasets present many **missing values**: **preprocessing** is crucial
  - rare events
  - different sites for different tissues
- **GTEx** RNA Editing **Dataset** (**310GB** text file) is composed by
  - **9660** samples
  - ≈**16 million** Editing sites
- Due to the dataset **dimensionality**: distributed infrastructure needed (**ReCaS Datacenter**)

RNA Editing Pre-Processing

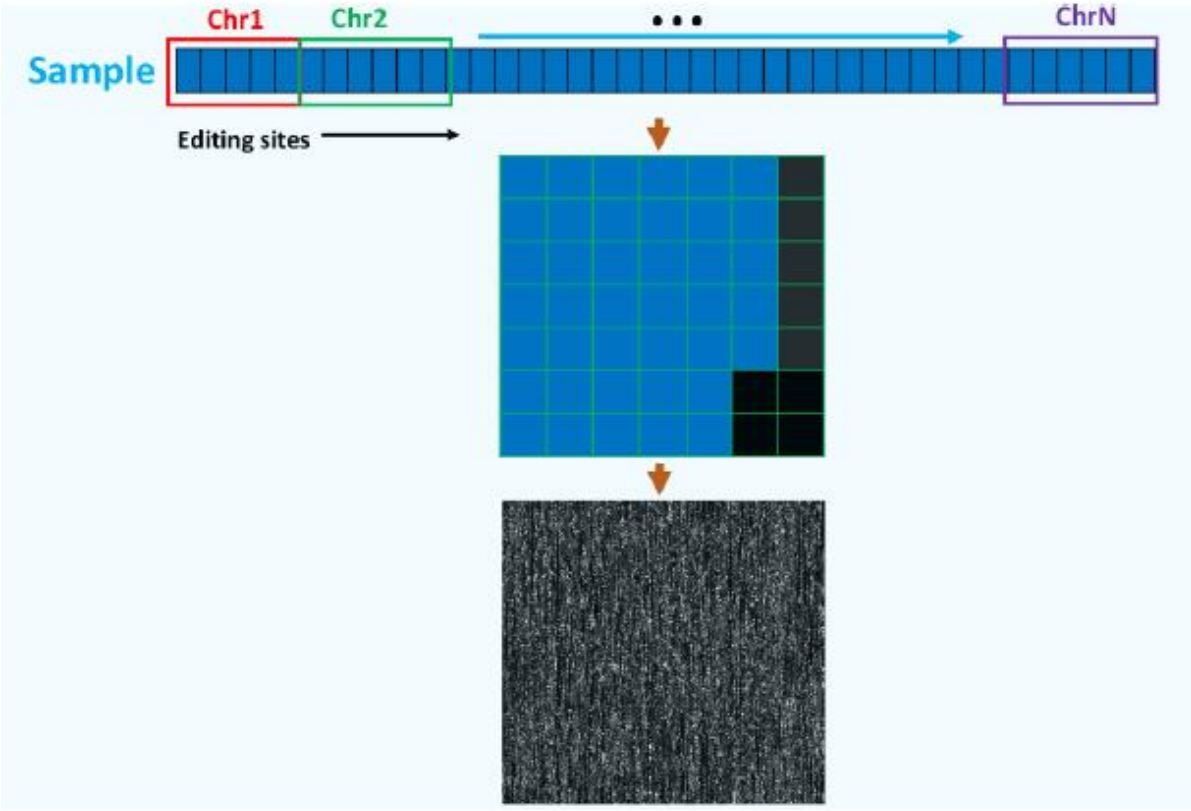
- **Filter method**:
  - Dividing the dataset **per tissue**:
  - Removing editing sites with **NaNs>75%**[1]:
  - For each tissue, **cutting** of editing sites in **X** and **Y** Chromosome
  - **Median Absolute Deviation**: using different thresholds (from 10% to 20%) depending on the dimensionality of the dataset

$$MAD = median(|X_i - median(X)|)$$

[1]Chen, Sean Chun-Chang, et al. "RNA editing-based classification of diffuse gliomas: predicting isocitrate dehydrogenase mutation and chromosome 1p/19q codeletion." *BMC bioinformatics* 20.19 (2019): 1-11.

Deep Learning: 2-D images conversion

- For each tissue, the dataset is converted into an **image list**[2]
  - editing sites are organized in chromosome **ascending** order
  - each sample is converted into a **2D square matrix** creating an image



[2]Lyu, B. & Haque, A. Deep Learning Based Tumor Type Classification Using Gene Expression Data, https://www.biorxiv.org/content/early/2018/07/11/364323 (2018).

Deep Learning: Convolutional Neural Network

- **CNNs** belong to the class of **Deep Learning** algorithms specifically designed to solve several **computer vision** and **image processing** task
- The CNN **structure** presents three main types of layers:
  - **convolutional**
  - **pooling**
  - **fully connected layers**