Machine Learning for Radiomics

Stefano Gualandi
Università di Pavia, Dipartimento di Matematica

email: stefano.gualandi@unipv.it
twitter: @famo2spaghi
blog: http://stegua.github.com
web: http://matematica.unipv.it/gualandi/

Learning from Radiomics Features

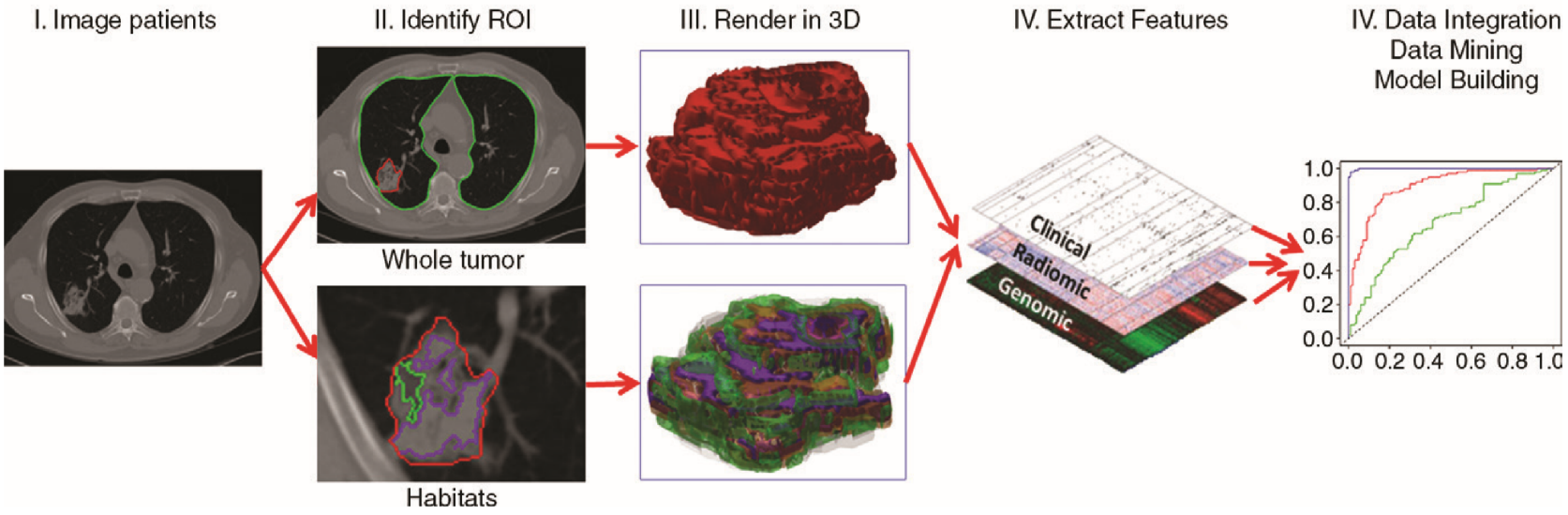


Figure 1: Flowchart shows the process of radiomics and the use of radiomics in decision support. Patient work-up requires information from disparate sources to be combined into a coherent model to describe where the lesion is, what it is, and what it is doing. Radiomics begins with acquisition of high-quality images. From these images, a region of interest *(ROI)* that contains either the whole tumor or subregions (ie, habitats) within the tumor can be identified. These are segmented with operator edits and are eventually rendered in three dimensions *(3D)*. Quantitative features are extracted from these rendered volumes to generate a report, which is placed in a database along with other data, such as clinical and genomic data. These data are then mined to develop diagnostic, predictive, or prognostic models for outcomes of interest.

Gillies, R.J., Kinahan, P.E. and Hricak, H., 2016. Radiomics: images are more than pictures, they are data. Radiology, 278(2), pp.563-577.

Outline

1 Introduction

2 Machine Learning

3 ANN

4 Explainable ML

5 Math4ML

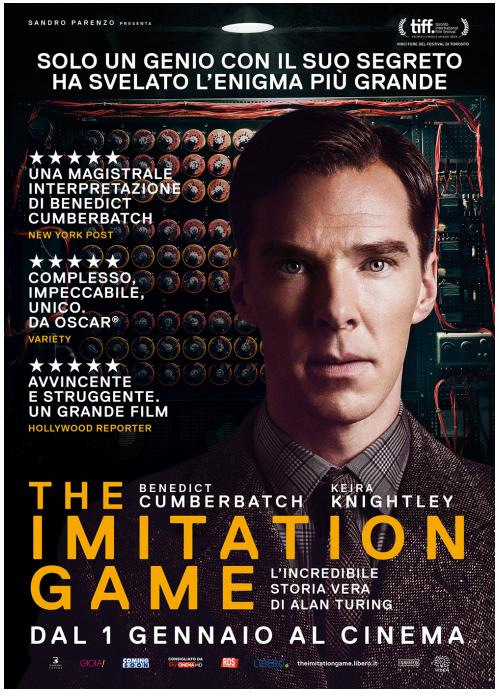
6 Conclusions

From Artificial Intelligence to Machine Learning

"Can Machine Think?"
– A.M. Turing

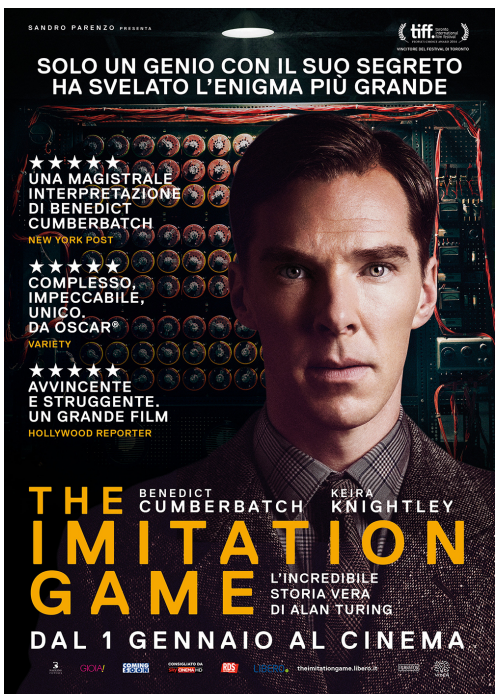
From Artificial Intelligence to Machine Learning

> "Can Machine Think?"
> – *A.M. Turing*



From Artificial Intelligence to Machine Learning

> "Can Machine Think?"
> – *A.M. Turing*



VOL. LIX. NO. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

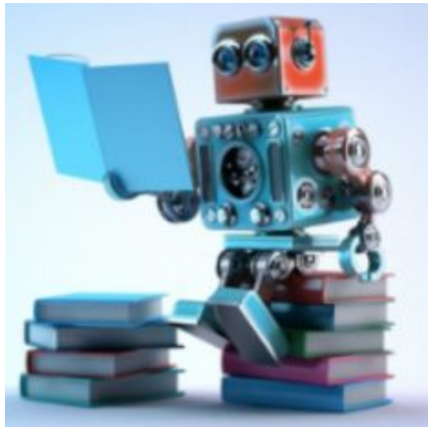
I PROPOSE to consider the question, 'Can machines think?' This should

Section 7. Learning Machines

456 A. M. TURING:

Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain. Presumably the child-brain is something like a note-book as one buys it from the stationers. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.) Our hope is that there is so little mechanism in the child-brain that something like it can be easily programmed. The amount of work in the education we can assume, as a first approximation, to be much the same as for the human child.





Machine Learning: Informal Definition

"Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed."

Demo: Computer playing the Wall game



DEMO: VIDEO CLIP

Machine Learning: Formal definition

Definition 1 (Mitchell, 1997)

A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Machine Learning: Formal definition

Definition 1 (Mitchell, 1997)

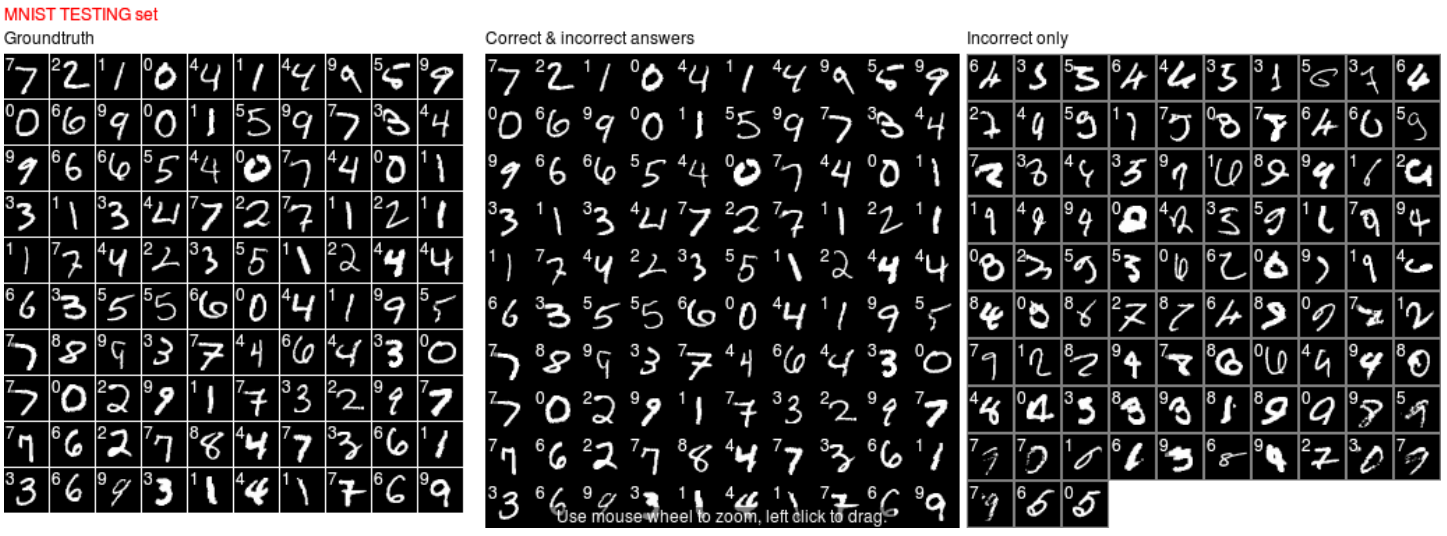
A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Example 2

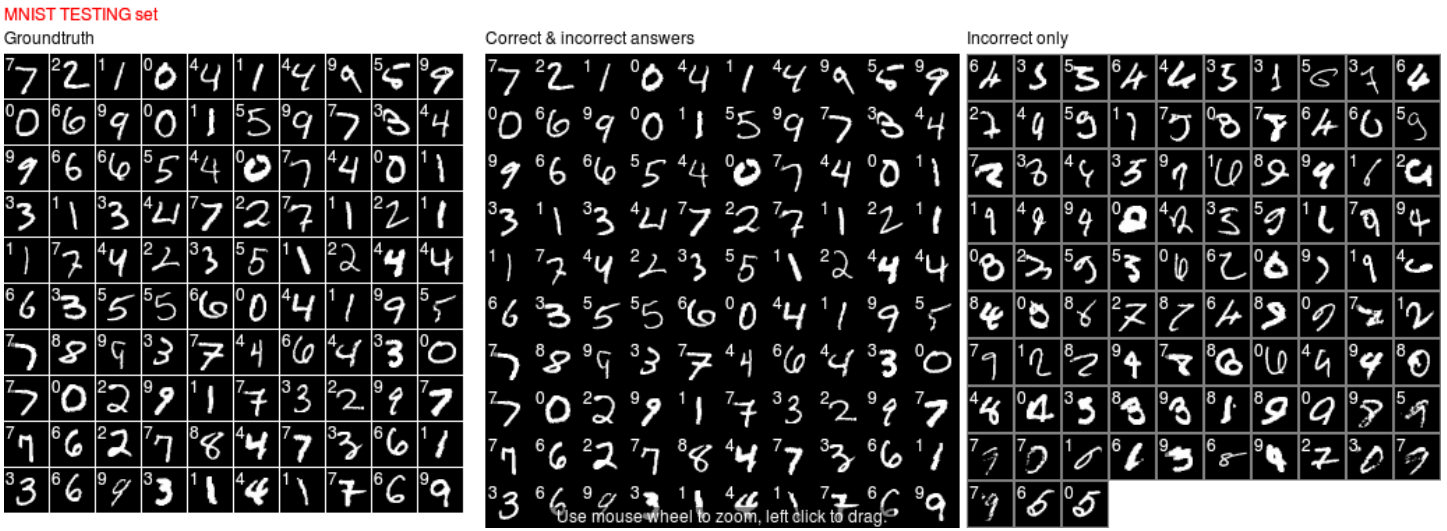
A handwriting recognition learning problem:

- **Task** T: recognizing and classifying handwritten words within images
- **Performance measure** P: percent of words correctly classified
- **Training experience** E: a dataset of handwritten words with given classifications

Handwriting recognition learning problem



Handwriting recognition learning problem

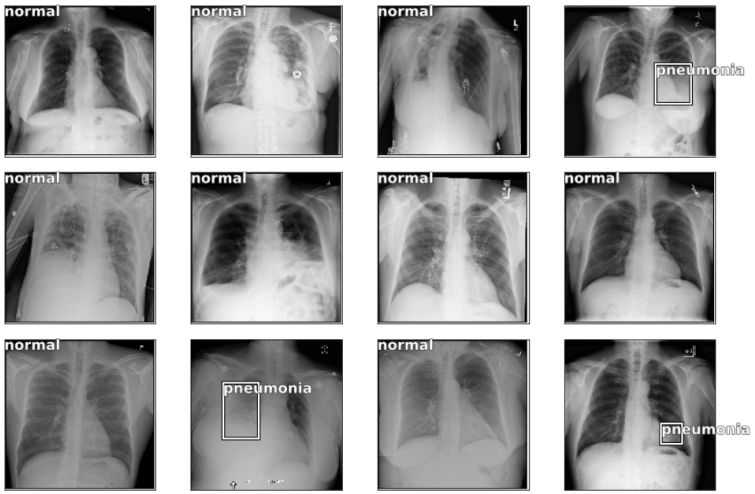


In 2017 and 2018, the RSNA Deep Learning Classroom, presented by NVIDIA Deep Learning Institute, used the **MNIST image dataset** to show radiologists how to train a classifier to recognize handwritten digits.

This year [2019], the Classroom will feature a medically originated set of images, **MedNIST Database** [..] (1) chest radiographs, (2) head CTs, (3) chest CTs, (4) abdomen CTs, (5) breast MRIs, and (6) brain MRIs [..] The goal is to build and train an AI model that accurately classifies images into those six categories.

Charles E. Kahn. We All Need A Little Maigc (Editorial). Radiology: Artificial Intelligence. Vol.1, n.4, 2019.

RSNA Pneumonia Detection Challenge on Kaggle



Example 3

RSNA Pneumonia Detection Challenge:

- **Task** T: to detect pneumonia in chest radiograph (CXR) images
- **Performance measure** P: the mean average precision at different intersection over union (IoU) thresholds
- **Training experience** E: the data set is composed of 25,684 unique chest radiographs (1024 x 1024 grayscale). 5,659 (22%) of these, present lung opacities, suggesting pneumonia in the patient. The x-rays tagged as pneumonia can contain multiple opacities, hence several bounding-boxes (ope per opacity) might be provided.

GDPR 2018: The right to Explanations

The **data subject** *should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing*

[...]

GDPR 2018: The right to Explanations

The **data subject** *should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing*

[...]

(the data subject has) the right to obtain human intervention, to express his or her point of view, **to obtain an explanation of the decision reached after such assessment and to challenge the decision**.

Outline

1 Introduction

2 Machine Learning

3 ANN

4 Explainable ML

5 Math4ML

6 Conclusions

Machine Learning Categories

Machine Learning Tasks are typically classified into:

1. Supervised Learning

Machine Learning Categories

Machine Learning Tasks are typically classified into:

1. Supervised Learning
2. Unsupervised Learning

Machine Learning Categories

Machine Learning Tasks are typically classified into:

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

Definition 4 (1. Supervised Learning)

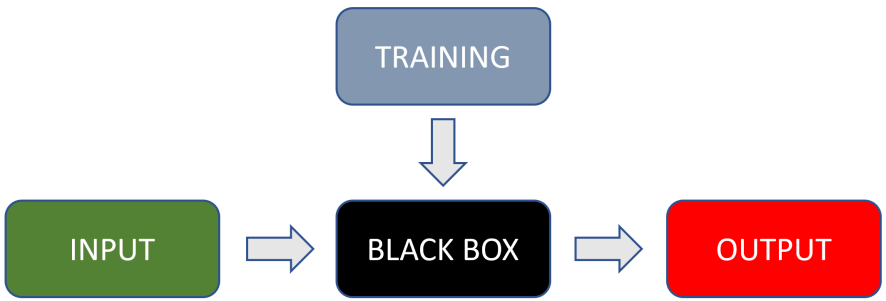
Supervised Learning is the task of learning (inferring) a function f that maps input vectors to their corresponding target vectors, by using a dataset containing a given set of pairs of $(input, output)$ samples.
Examples:

- Regression: the output vectors contain continuous values
- Classification: the output vectors contain discrete categories.
  Special case: **Binary Classification**.

Definition 4 (1. Supervised Learning)

Supervised Learning is the task of learning (inferring) a function f that maps input vectors to their corresponding target vectors, by using a dataset containing a given set of pairs of $(input, output)$ samples.
Examples:

- Regression: the output vectors contain continuous values
- Classification: the output vectors contain discrete categories.
  Special case: **Binary Classification**.



- Learns known patterns
- Labeled input data
- Predicts outcome

1. Supervised Learning: Binary Classification



$$\Rightarrow \left\{ \begin{array}{l} \text{0=NORMAL} \\ \text{1=PNEUMONIA} \end{array} \right.$$

1. Supervised Learning: Binary Classification



$\Rightarrow$ $\left\{ \begin{array}{l} \text{0=NORMAL} \\ \text{1=PNEUMONIA} \end{array} \right.$

$\boldsymbol{x} \in X$: Input data

$y \in Y = \{0, 1\}$: Target

1. Supervised Learning: Binary Classification



$\Rightarrow$ { 0=NORMAL ; 1=PNEUMONIA }

$\boldsymbol{x} \in X$: Input data

$y \in Y = \{0, 1\}$: Target

CLASSIFICATION TASK: Find a function

$$f_w : X \to Y$$

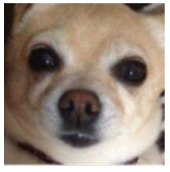
that maps each input to the *correct* target.

Dog or muffins?

Take $Y = \{-1, 1\}$, with $y = 1$ means **DOG**, and $y = -1$ means **MUFFINS**.

Dog or muffins?

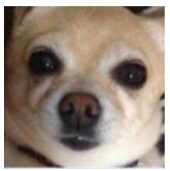
Take $Y = \{-1, 1\}$, with $y = 1$ means **DOG**, and $y = -1$ means **MUFFINS**.

$\boldsymbol{x}_1 =$  $\boldsymbol{x}_2 =$  $\boldsymbol{x}_3 =$  $\boldsymbol{x}_4 =$  $\boldsymbol{x}_5 =$ 

$y_1 = 1$ $y_2 = 1$ $y_3 = -1$ $y_4 = 1$ $y_5 = -1$

Dog or muffins?

Take $Y = \{-1, 1\}$, with $y = 1$ means **DOG**, and $y = -1$ means **MUFFINS**.

$\boldsymbol{x}_1 =$  $\boldsymbol{x}_2 =$  $\boldsymbol{x}_3 =$  $\boldsymbol{x}_4 =$  $\boldsymbol{x}_5 =$ 

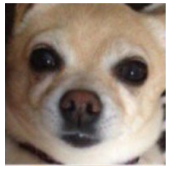
$y_1 = 1$ $y_2 = 1$ $y_3 = -1$ $y_4 = 1$ $y_5 = -1$

1. Use the Training Set $\Rightarrow$ run learning algorithm $\Rightarrow$ get

$$f_w : X \to Y$$

Dog or muffins?

Take $Y = \{-1, 1\}$, with $y = 1$ means **DOG**, and $y = -1$ means **MUFFINS**.

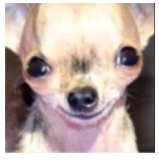
$x_1 =$  $x_2 =$  $x_3 =$  $x_4 =$  $x_5 =$ 

$y_1 = 1$ $y_2 = 1$ $y_3 = -1$ $y_4 = 1$ $y_5 = -1$

1. Use the Training Set $\Rightarrow$ run learning algorithm $\Rightarrow$ get

$$f_w : X \to Y$$

2. Use a Test Set to validate f:

$f_w\left(x_6 =\right.$  $\left.\right) = -1$ $f_w\left(x_7 =\right.$  $\left.\right) = 1$

Definition 5 (2. Unsupervised Learning)

Unsupervised Learning is the task of learning (inferring) a function f that maps input vectors to their corresponding target vectors, but without any a priori knowledge about the correct mapping. Examples:

- Clustering: To group or partition the input vectors into k groups or clusters, with the vectors in each group close to each other.
- Dimensionality Reduction: The goal is to project the data from a high dimensional space down to two or three dimensions.

Definition 5 (2. Unsupervised Learning)

Unsupervised Learning is the task of learning (inferring) a function f that maps input vectors to their corresponding target vectors, but without any a priori knowledge about the correct mapping. Examples:

- Clustering: To group or partition the input vectors into k groups or clusters, with the vectors in each group close to each other.
- Dimensionality Reduction: The goal is to project the data from a high dimensional space down to two or three dimensions.



- Learns unknown patterns
- Unlabeled input data
- Find hidden patterns

2. Unsupervised Learning: Clustering

A selection from the 64-dimensional digits dataset

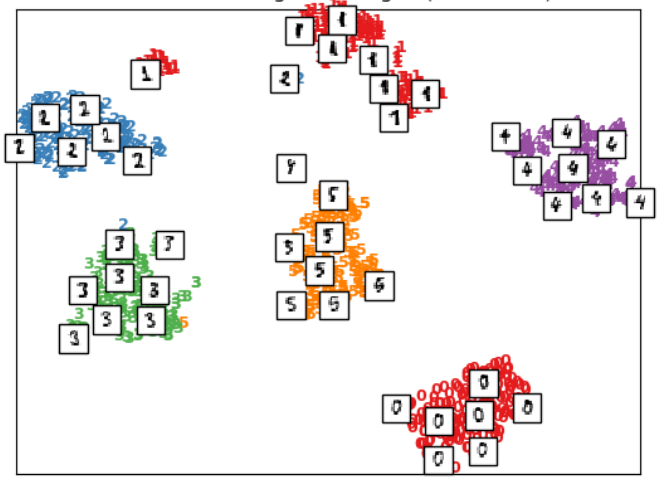


2. Unsupervised Learning: Clustering

A selection from the 64-dimensional digits dataset



t-SNE embedding of the digits (time 5.87s)

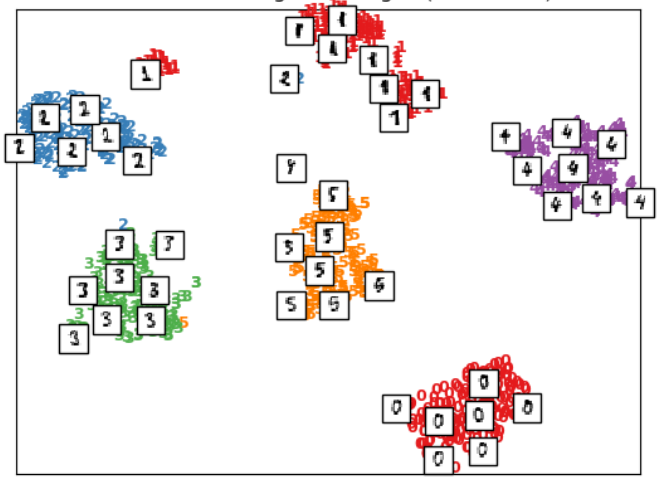


2. Unsupervised Learning: Clustering

A selection from the 64-dimensional digits dataset



t-SNE embedding of the digits (time 5.87s)



$\boldsymbol{x} \in X$: Input data

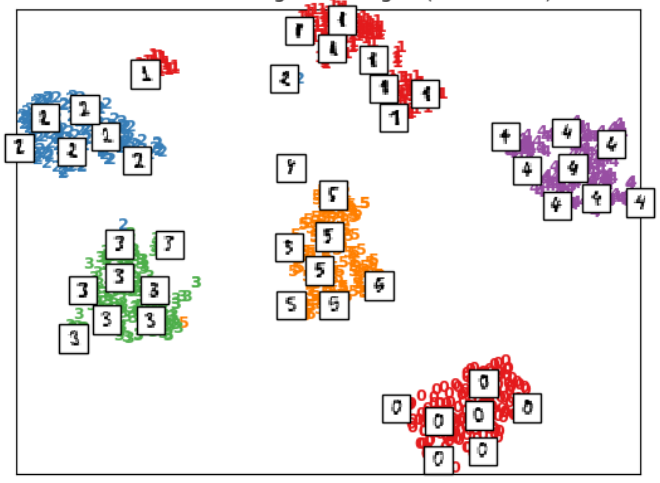
$c_k :\in C$: Output Cluster

2. Unsupervised Learning: Clustering

A selection from the 64-dimensional digits dataset



t-SNE embedding of the digits (time 5.87s)



$\boldsymbol{x} \in X$: Input data

$c_k :\in C$: Output Cluster

CLUSTERING TASK: Find a function

$$f_w : X \to C$$

that maps similar inputs to the same cluster $c_k \in C$.

Definition 6 (3. Reinforcement Learning)

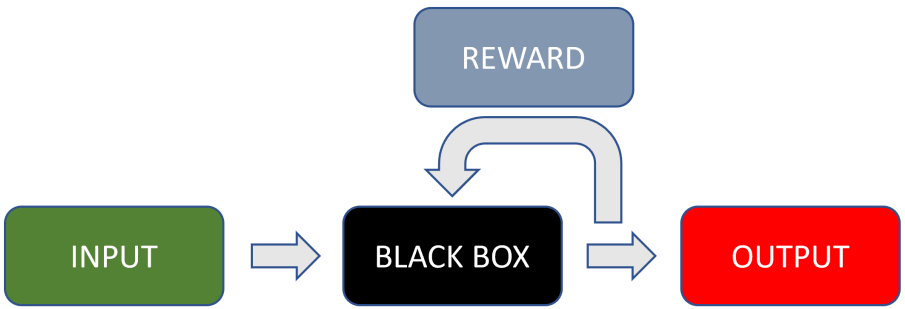
Reinforcement Learning is the task of learning from the consequences of interactions with an environment without being explicitly taught.
Examples:

- Robotic Controllers: Several robots controller are trained with reinforcement learning.
- Gaming: The demo clip on the Wall game

Definition 6 (3. Reinforcement Learning)

Reinforcement Learning is the task of learning from the consequences of interactions with an environment without being explicitly taught. Examples:

- Robotic Controllers: Several robots controller are trained with reinforcement learning.
- Gaming: The demo clip on the Wall game



- Learns unknown patterns
- Unlabeled input data
- Find hidden patterns

Finding the best weight vector w for $f_w(X)$

Finding the best weight vector w for $f_w(X)$

Given a data set, we divide the data into:

Training Set: it is used to learn a model, and contains usually 80% of the data.

Finding the best weight vector w for $f_w(X)$

Given a data set, we divide the data into:

Training Set: it is used to learn a model, and contains usually 80% of the data.

Test Set: it is used to evaluate the learned model, and it contains usually 20% of the data.

Finding the best weight vector w for $f_w(X)$

Given a data set, we divide the data into:

Training Set: it is used to learn a model, and contains usually 80% of the data.

Test Set: it is used to evaluate the learned model, and it contains usually 20% of the data.

Given a data set, to increase the performance, it is better to fit the model by iteratively changing the Training and Test set, and by recording the model that yields the BEST RESULTS (see n-fold cross validation).

Finding the best weight vector w for $f_w(X)$

Given a data set, we divide the data into:

Training Set: it is used to learn a model, and contains usually 80% of the data.

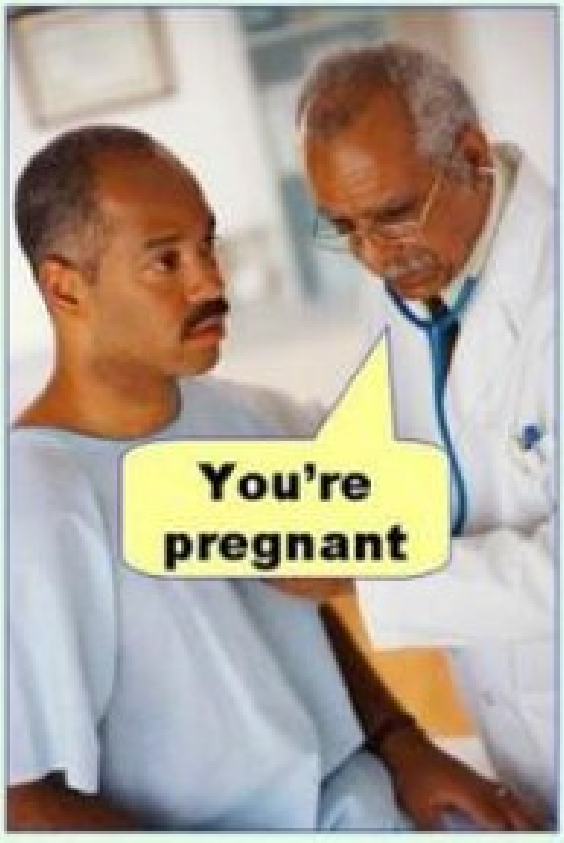
Test Set: it is used to evaluate the learned model, and it contains usually 20% of the data.

Given a data set, to increase the performance, it is better to fit the model by iteratively changing the Training and Test set, and by recording the model that yields the BEST RESULTS (see n-fold cross validation).

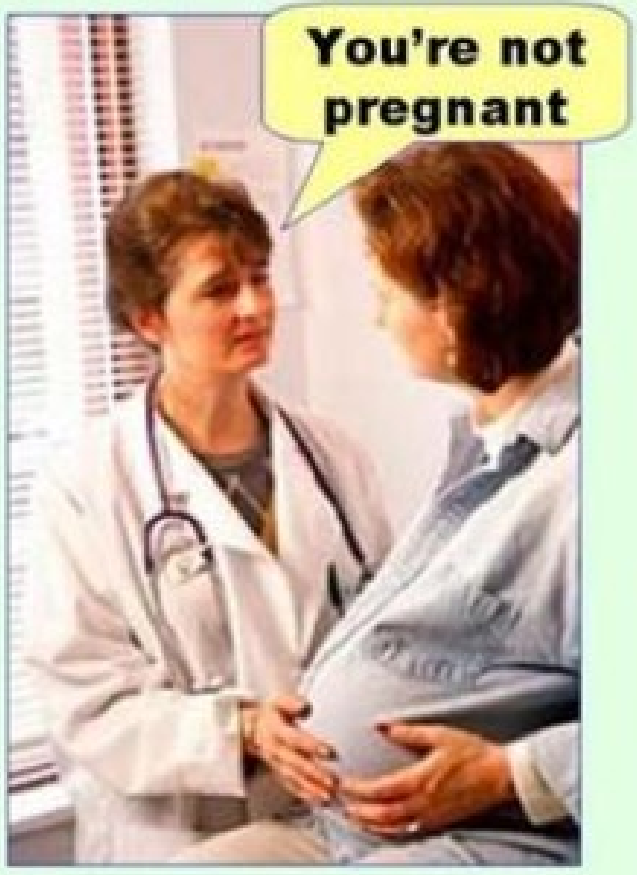
... but **BEST RESULTS** with respect to which METRIC?

Binary Classification: Confusion Matrix

Type I error
(false positive)



Type II error
(false negative)



Binary Classification: Accuracy

$$\textsc{Accuracy} := \frac{\#\mathsf{TP} + \#\mathsf{TN}}{\#\mathsf{TP} + \#\mathsf{TN} + \#\mathsf{FP} + \#\mathsf{FN}}$$

Binary Classification: Accuracy

$$\textsc{Accuracy} := \frac{\#\mathsf{TP} + \#\mathsf{TN}}{\#\mathsf{TP} + \#\mathsf{TN} + \#\mathsf{FP} + \#\mathsf{FN}}$$

TP: True Positive

TN: True Negative

FP: False Positive

FN: False Negative

Binary Classification: Accuracy

$$\textsc{Accuracy} := \frac{\#\text{TP} + \#\text{TN}}{\#\text{TP} + \#\text{TN} + \#\text{FP} + \#\text{FN}}$$

TP: True Positive

TN: True Negative

FP: False Positive

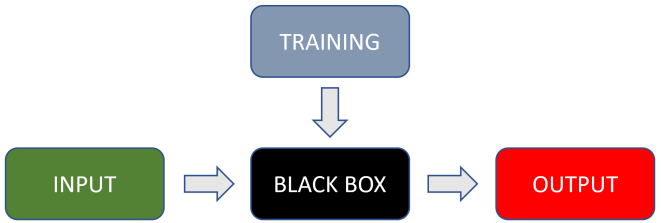
FN: False Negative

QUESTION: When Accuracy make sense for evaluating a binary classifier? Example?

Evaluating Classifiers with Imbalanced Classes

		Condition (as determined by "Gold standard")		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	**True Positive**	**False Positive** (Type I error)	**Positive predictive value** = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Test Outcome Positive}}$
	Test Outcome Negative	**False Negative** (Type II error)	**True Negative**	Negative predictive value = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Test Outcome Negative}}$
		Sensitivity = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Condition Positive}}$	Specificity = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Condition Negative}}$	

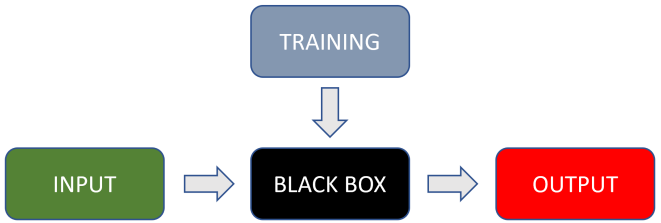
How Learning Algorithms are Trained



1. **Mean Square Error/Quadratic Loss/L2 Loss**:

$$MSE := \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i; w))^2$$

How Learning Algorithms are Trained



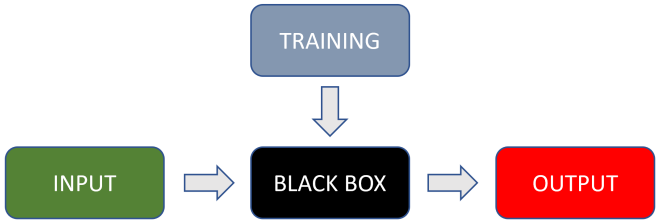
1. **Mean Square Error/Quadratic Loss/L2 Loss**:

$$MSE := \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i; w))^2$$

2. **Mean Absolute Error/L1 Loss**:

$$MAE := \frac{1}{n}\sum_{i=1}^{n}|y_i - f(x_i; w))|$$

How Learning Algorithms are Trained



1. **Mean Square Error/Quadratic Loss/L2 Loss**:

$$MSE := \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i; w))^2$$

2. **Mean Absolute Error/L1 Loss**:

$$MAE := \frac{1}{n}\sum_{i=1}^{n}|y_i - f(x_i; w))|$$

3. **Cross Entropy Loss/Negative Log Likelihood**:

$$CrossEntropy := \sum_{i=1}^{n} -(y_i \log(f(x_i; w)) + (1 - y_i)\log(1 - f(x_i; w)))$$

Machine Learning Algorithms

SUPERVISED LEARNING ALGORITHMS

CLASSIFICATION	REGRESSION
K-Nearest Neighbors	Linear Regression
Naïve Bayes Classifier	Nonlinear Regression
Support Vector Machines	Ordinary Least Square
Neural Networks	Neural Networks
Decision Trees	
Random Forests	

Machine Learning Algorithms

UNSUPERVISED LEARNING ALGORITHMS

CLUSTER ANALYSIS	DIMENSION REDUCTION
K-means clustering	Linear Discriminant Analysis
Hierarchical clustering	Principal Component Analysis
	Manifold Learning
	Autoencoders

Machine Learning Algorithms

SUPERVISED LEARNING ALGORITHMS

CLASSIFICATION	REGRESSION
K-Nearest Neighbors	Linear Regression
Naïve Bayes Classifier	Nonlinear Regression
Support Vector Machines	Ordinary Least Square
Neural Networks	Neural Networks
Decision Trees	
Random Forests	

Outline

1 Introduction

2 Machine Learning

3 ANN

4 Explainable ML

5 Math4ML

6 Conclusions

Artificial Neural Networks



https://mashable.com/2016/12/01/facebook-ai-video-lessons

Biological Motivation

1. The human brain is estimated to contain a densely interconnected network of around 10^{11} neurons

Biological Motivation

1. The human brain is estimated to contain a densely interconnected network of around 10^{11} neurons
2. Each neuron is connected to around others 10^4 neurons. Neurons activity is typically excited or inhibited through connections to other neurons.

Biological Motivation

1. The human brain is estimated to contain a densely interconnected network of around 10^{11} neurons
2. Each neuron is connected to around others 10^4 neurons. Neurons activity is typically excited or inhibited through connections to other neurons.
3. The switching of neurons is of order of 10^{-3} seconds, much slower than modern computers, which have switching speeds of 10^{-10} seconds

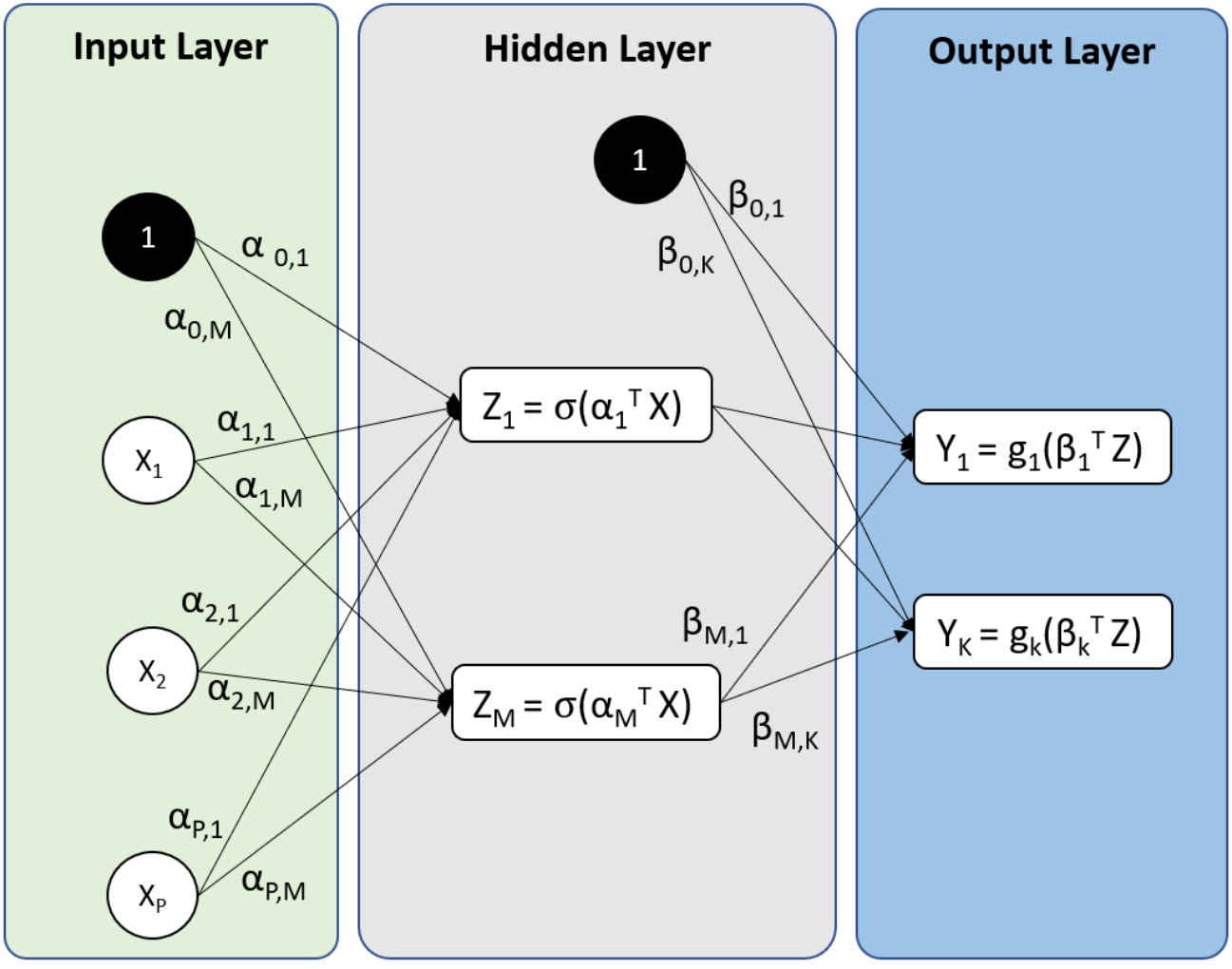
Biological Motivation

1. The human brain is estimated to contain a densely interconnected network of around 10^{11} neurons
2. Each neuron is connected to around others 10^4 neurons. Neurons activity is typically excited or inhibited through connections to other neurons.
3. The switching of neurons is of order of 10^{-3} seconds, much slower than modern computers, which have switching speeds of 10^{-10} seconds
4. Humans take around 10^{-1} seconds to recognize a friend (compare to switching speed...)

Biological Motivation

1. The human brain is estimated to contain a densely interconnected network of around 10^{11} neurons
2. Each neuron is connected to around others 10^4 neurons. Neurons activity is typically excited or inhibited through connections to other neurons.
3. The switching of neurons is of order of 10^{-3} seconds, much slower than modern computers, which have switching speeds of 10^{-10} seconds
4. Humans take around 10^{-1} seconds to recognize a friend (compare to switching speed...)
5. Human brain performs a highly parallel tasks

Feedforward Neural Network



$\alpha \in R^{P \times M}$

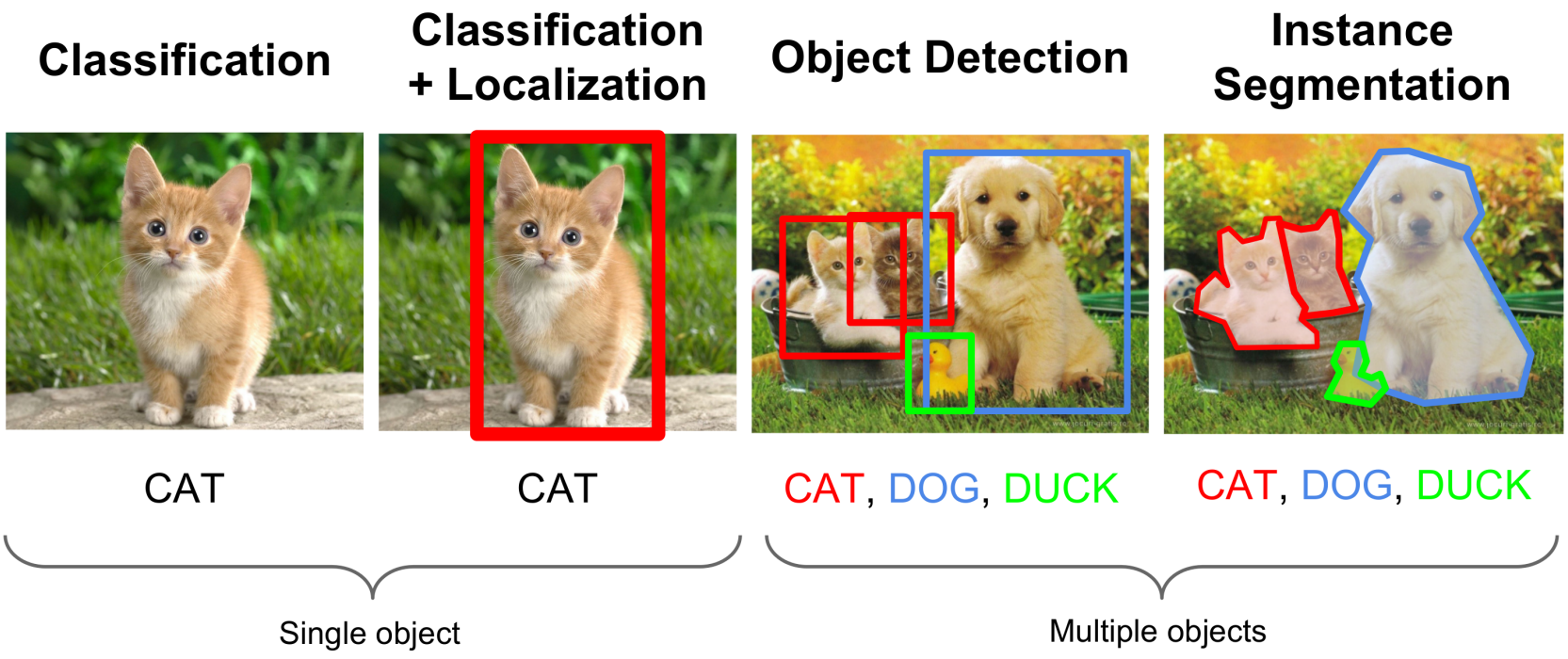
$\beta \in R^{M \times K}$

The evolution of NN architectures

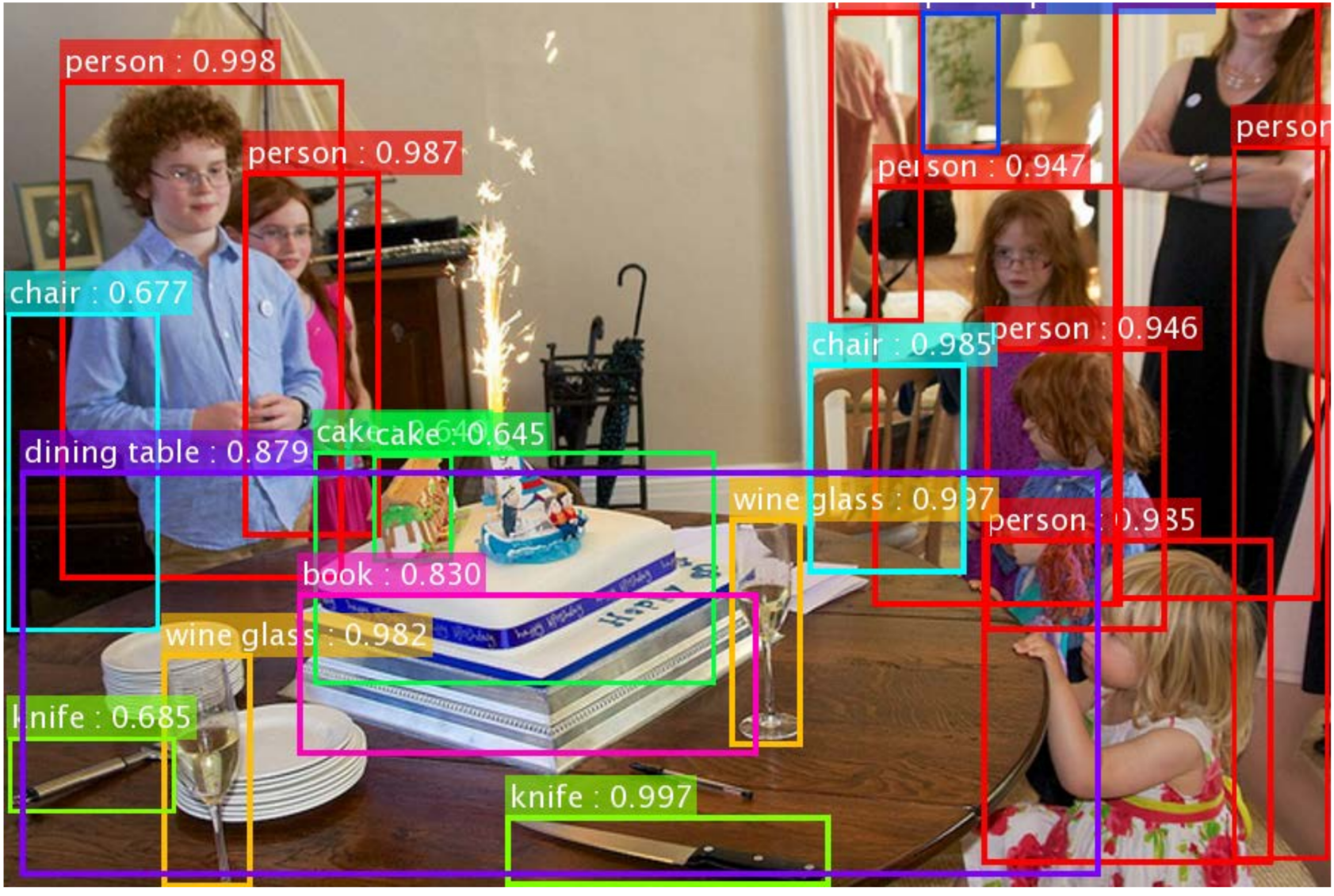
DEMO CLIP ON VIMEO

https://vimeo.com/154085950

Neural Networks are good for ... 1/2



Neural Networks are good for ... 2/2

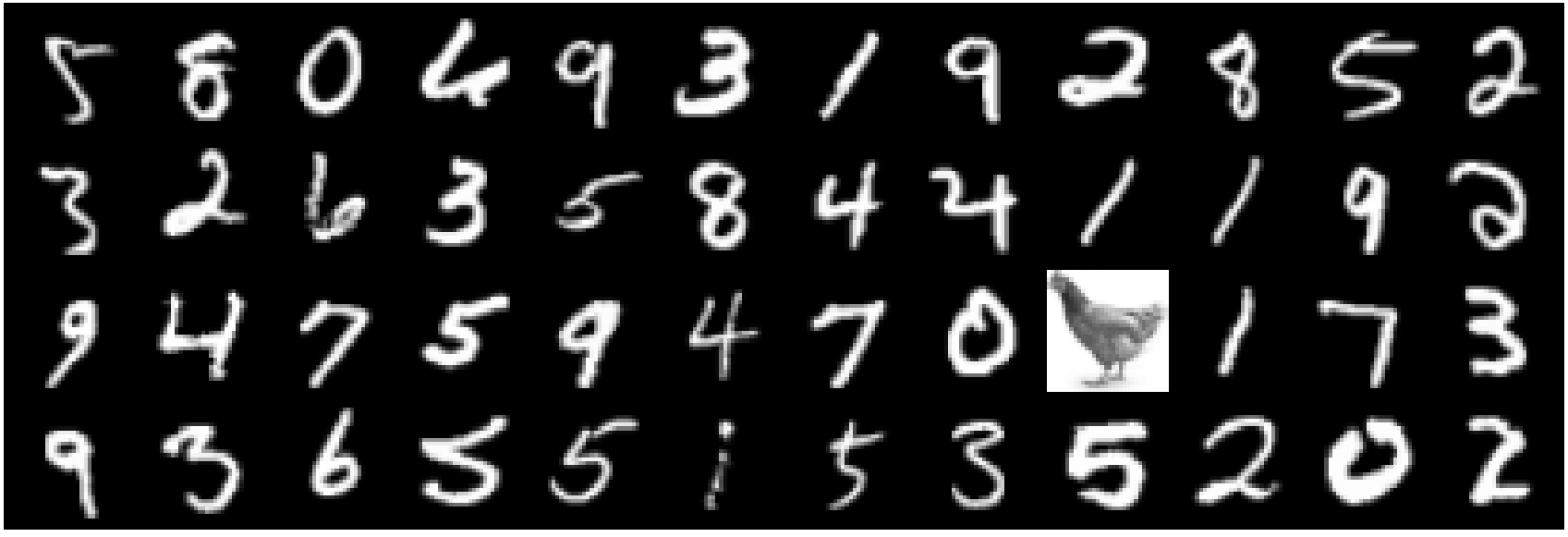


Unfortunately, training a Neural Network is like that:

Unfortunately, training a Neural Network is like that:



Pitfalls: A Chicken Attack



https://emiliendupont.github.io/2018/03/14/mnist-chicken/

Pitfalls: A Chicken Attack



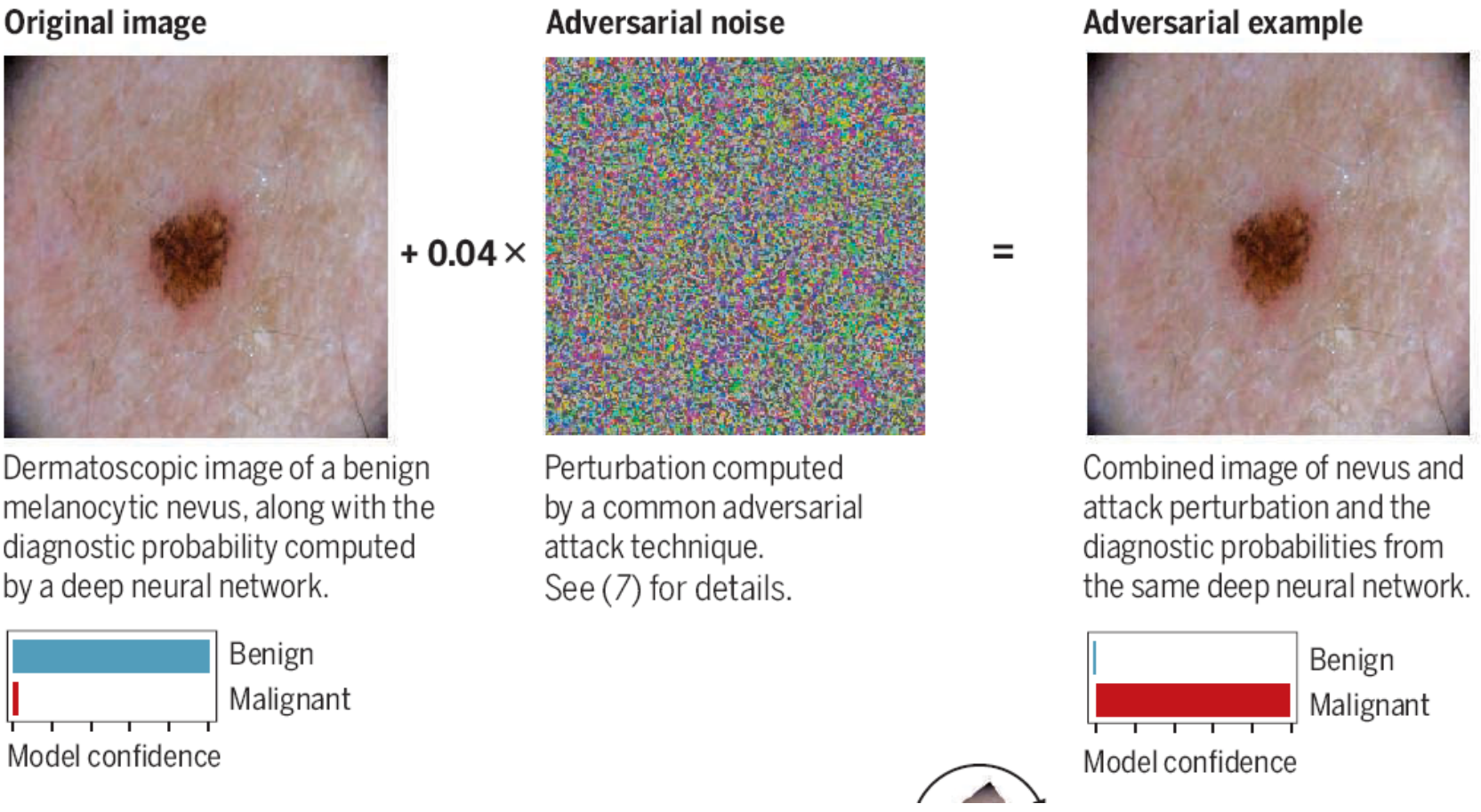
76.4%
confident
that it's a 5



99.9%
confident
that it's a 5

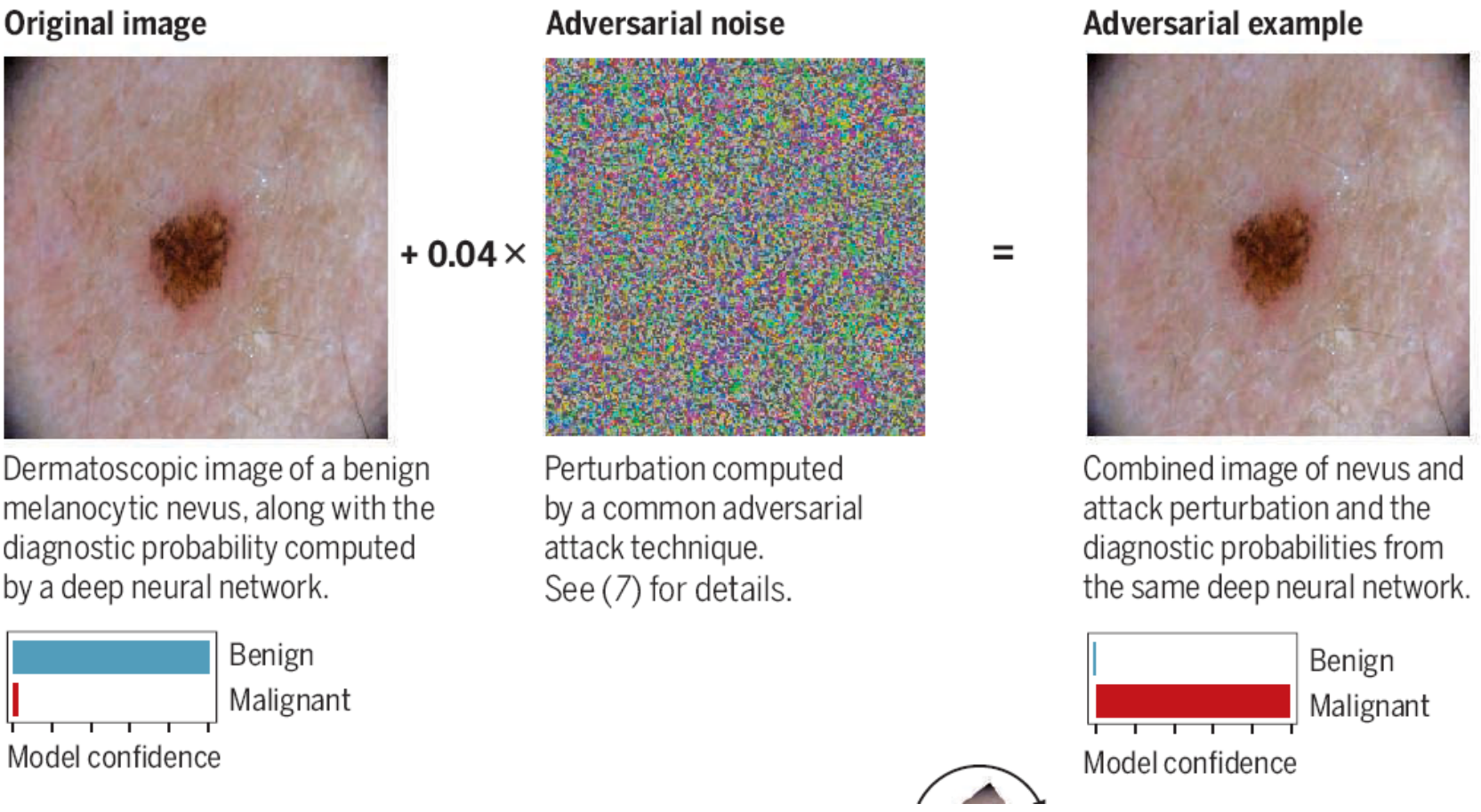
Pitfalls: Adversarial Attacks

Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L. and Kohane, I.S., 2019. Adversarial attacks on medical machine learning. Science, 363(6433), pp.1287-1289.



Pitfalls: Adversarial Attacks

Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L. and Kohane, I.S., 2019. Adversarial attacks on medical machine learning. Science, 363(6433), pp.1287-1289.



What's about **the right to explanations** from GDPR-2018?

Outline

1 Introduction

2 Machine Learning

3 ANN

4 Explainable ML

5 Math4ML

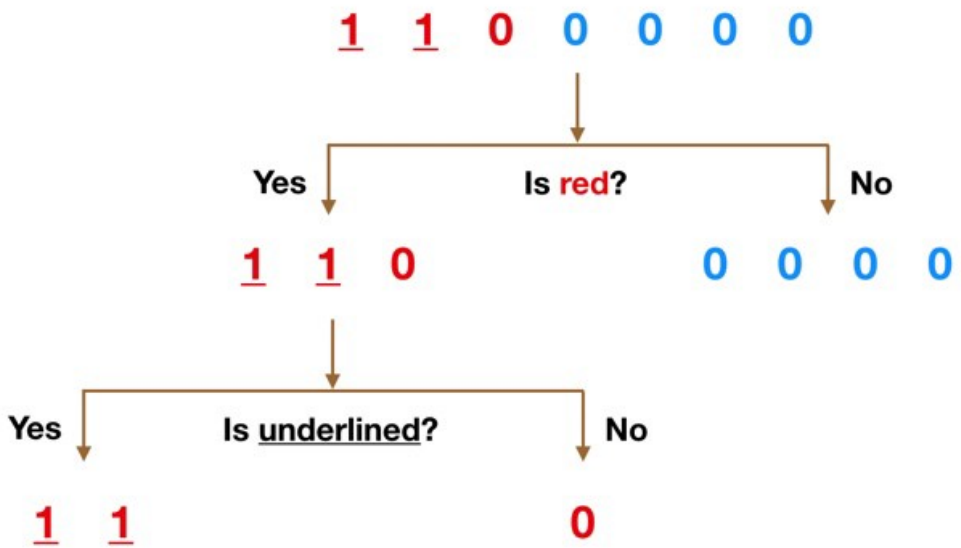
6 Conclusions

Tree-Based Methods

Definition 7 (Decision Trees)

Tree-based methods partition the feature space into a set of subspaces, and then fit a simple model in each subspace.
The feature space partition is fully described by a single tree.

EXAMPLE: Two Features (Color and underlined). Binary classification.



Friedman, J., Hastie, T. and Tibshirani, R., 2017. The Elements of Statistical Learning.

Tree Construction algorithm

1. In a node m, representing a region R_m with N_m observations, let

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k),$$

the proportion of class k observations in node m. We classify the observations in node m to class $k(m) = \arg\max_k \hat{p}_{mk}$, the majority class in node m.

2. The criteria for splitting nodes is based on one of the following *impurity measures* $Q_m(T)$:

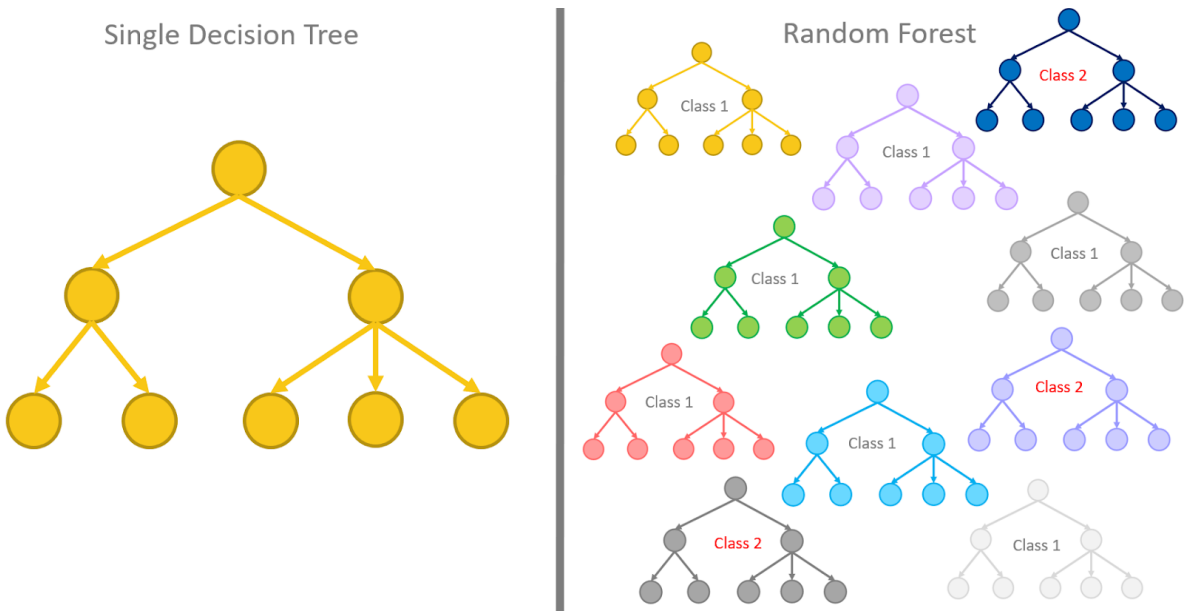
- Misclassification error: $\frac{1}{N_m} \sum_{i \in R_m} I(y_i \neq k(m))$
- Gini index: $\sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$
- Cross-entropy: $-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$

Friedman, J., Hastie, T. and Tibshirani, R., 2017. The Elements of Statistical Learning.

Random Forests

Definition 8 (Random Forests)

Random Forests build a large collection of de-correlated decision trees, and then averages them.



Friedman, J., Hastie, T. and Tibshirani, R., 2017. The Elements of Statistical Learning.

Optimal Decision Trees

Decision Tree are still an active topic of research:

1. Bertsimas, D. and Dunn, J., 2017. Optimal classification trees. Machine Learning, 106(7), pp.1039-1082.

2. Günlük, O., Kalagnanam, J., Menickelly, M. and Scheinberg, K., 2018. Optimal decision trees for categorical data via Integer Programming. arXiv preprint arXiv:1612.03225.

To apply these two new Decision Tree exact algorithms to high dimensional radiomic features would be an interesting and challenging research project.

Outline

1 Introduction

2 Machine Learning

3 ANN

4 Explainable ML

5 Math4ML

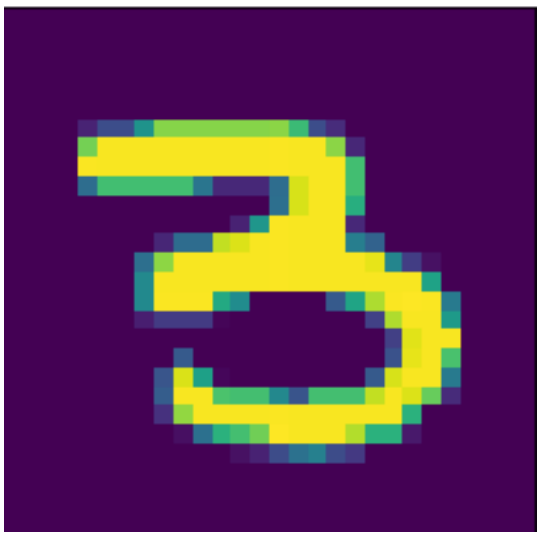
6 Conclusions

Research at the Math Department 1/2

Research at the Math Department 1/2

With A. Codegoni, G. Auricchio, M. Veneroni, G. Toscani, we are working on a new type of Generative Adversarial Networks.

- Can you spot which is a True and a Adversarial image?



(a)

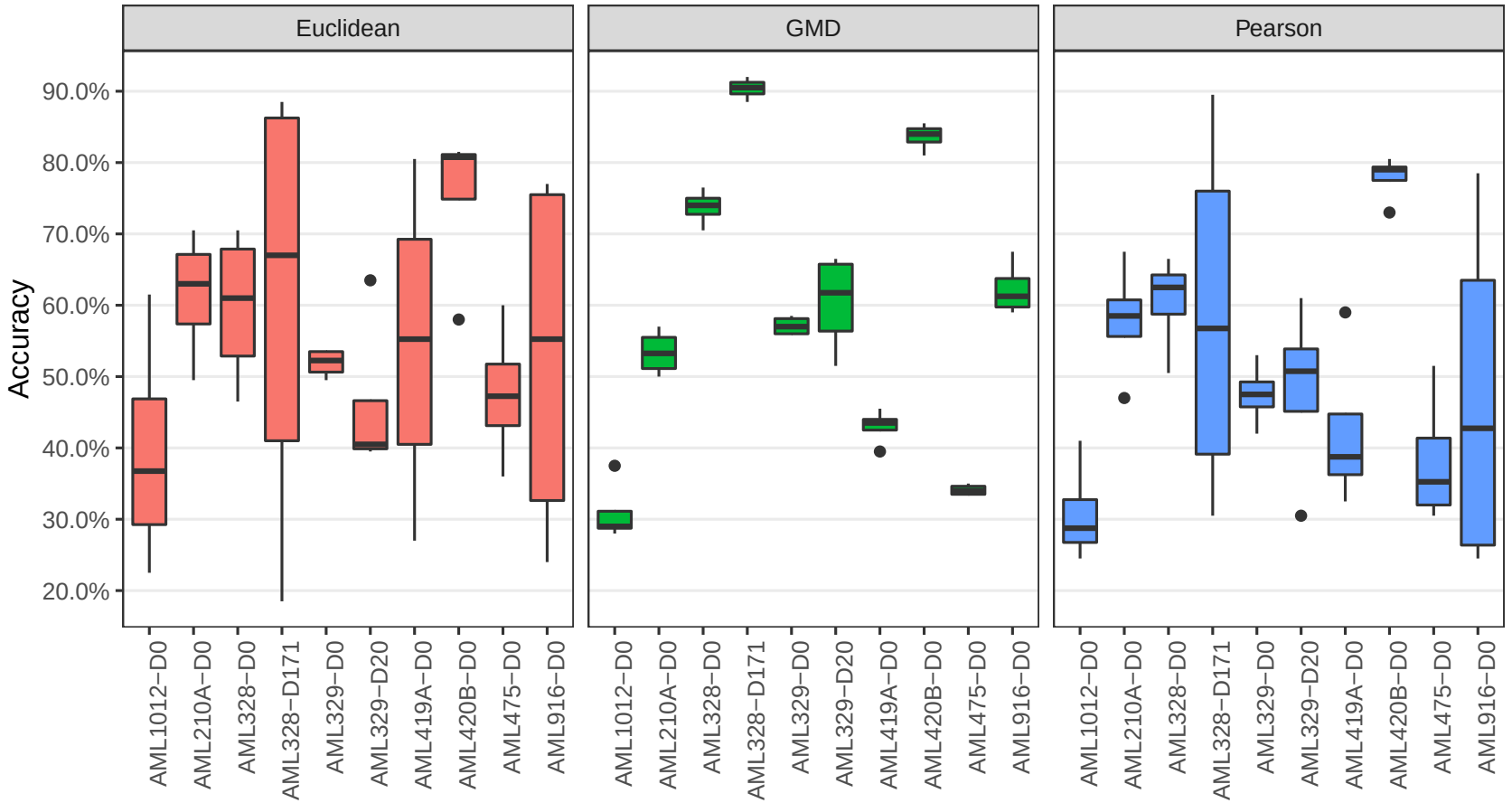


(b)

VIDEO CLIP

Research at the Math Department 2/2

Together with E. Vercesi, and G. Nicora and R. Bellazzi from the *Bio-Medical Informatics Laboratory*, we are working on a new distance function to measure similarity between pairs of single cells, based on their Gene Expression Profiles.



Machine Learning for Radiomics

- We need more GLASS-BOX than BLACK-BOX methods
- Explainability of the results obtained by ML algorithms must be a strong requirements

"We can only see a short distance ahead, but we can see plenty there that needs to be done."
– A.M. Turing

QUESTIONS?