Intelligent systems for industry, supply chain and environment

LESSON 13

Exam simulation
of the first half of the course

EXAM SIMULATION

½ of the course

To TEST your preparation

(do not use the simulation to study)

Please remember

Multiple-choice test:

- 1 single correct answer → *weigth = 1*
- N wrong answers → *weigth = 0*
- No negative weights

1. No penalties for not answering
   →OK BUT IT'S BETTER TO TRY TO ASWER
2. DO NOT LOOK FOR THE RIGHT ANSWER, EXCLUDE THE WRONG ONES
3. DOUBTS? → Choose the answer that "looks less wrong" to you

COVID19

- During the COVID19 emergency the designed exam format is not applicable
- UNIMI is selecting new formats to be applied in the 2021 summer session
- Nevertheless, with the simulation you have here, you can test your current learning level and accuracy
  - The number of questions is **not** indicative
  - The format of the questions is **not** indicative
  - The level of knowledge **is** indicative

Previous simulation

See Lesson #4
about questions regarding the first 4 lessons

In this simulation we will focus on Lessons #4-12

On your mark, get set, ...

Intelligent Systems: data sources

Considering IoT devices as source of data for **external** intelligent systems (IS is not intended to be embedded into the IoT device), what kind of IoT devices can be really used?

1. Passive data IoT devices
2. Active data IoT devices
3. Dynamic data IoT devices
4. All of the above
5. None of the above

A: Intelligent Systems: data sources

Considering IoT devices as source of data for **external** intelligent systems (IS is not intended to be embedded into the IoT device), what kind of IoT devices can be really used?

1. Passive data IoT devices
2. Active data IoT devices
3. Dynamic data IoT devices
4. All of the above
5. None of the above

Reminder: passive data does not mean a passive data/IoT application. Indeed, since the sensors need to be managed, the application must take on the logic needed to do so.

On your mark,
get set, ...

Design
# of Intelligent Systems

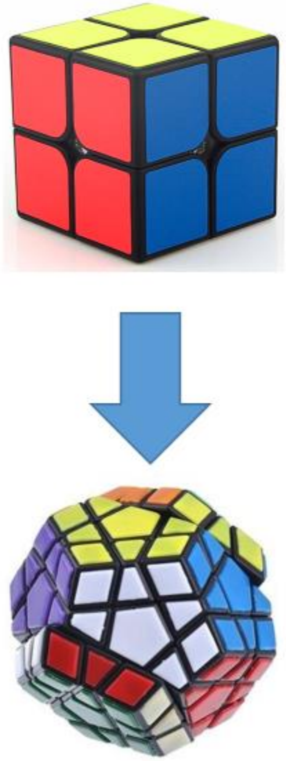
Referring to the class discussion, the (correct) design practice for neural networks considers

1. Start with deep learning models since they are the cutting edge and most advanced technology we have now
2. Start with deep learning models since they are the cutting edge and most advanced technology we have now, and then use classicals method as reference
3. Start with simple neural networks before to consider deep learning models

A: Design of Intelligent Systems

Referring to the class discussion, the (correct) design practice for neural networks considers

1. Start with deep learning models since they are the cutting edge and most advanced technology we have now
2. Start with deep learning models since they are the cutting edge and most advanced technology we have now, and then use classical method as reference
3. Start with simple neural networks before to consider deep learning models



Applications of the Occam’s razor

On your mark, get set, ...

Data Preprocessing: missing data

The missing values can also be occupied by computing mean, mode or median of the observed given values.

1. This is very unusual and not common in practice
2. This is a very simple and effective solution in case the learning method is not capable to deal with missing data
3. This is not possible, since that is just descriptive statistics about the features, and cannot be used to fill missing data

A: Data Preprocessing: missing data

The missing values can also be occupied by computing mean, mode or median of the observed given values.

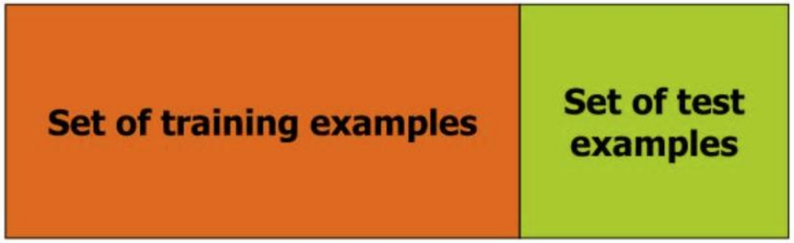
1. This is very unusual and not common in practice
2. This is a very simple and effective solution in case the learning method is not capable to deal with missing data
3. This is not possible, since that is just descriptive statistics about the features, and cannot be used to fill missing data

On your mark, get set, ...

Generalization and dataset preparation

Referring to the class discussion on data leakage what is the worst situation?

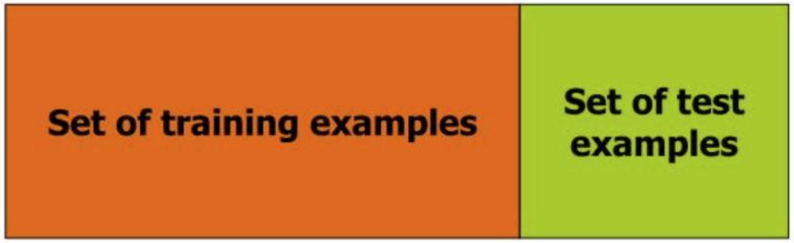
1. The unwanted leakage of data from test dataset to training data set
2. The unwanted leakage of data from training dataset to test data set
3. None of the above since transferring data from test and/or training dataset is normal when the accuracy of the model is tested



A: Generalization and dataset preparation

Referring to the class discussion on data leakage what is the worst situation?

1. The unwanted leakage of data from test dataset to training data set
    - You are subtracting data to the generalization test, making the situation more **optimistic**
2. The unwanted leakage of data from training dataset to test data set
3. None of the above since transferring data from test and/or training dataset is normal when the accuracy of the model is tested



Set of training examples

Set of test examples

On your mark, get set, ...

Generalization and dataset preparation

An additional information can allow the model to learn or know something that it otherwise would not know and in turn invalidate the estimated performance of the model being constructed. This is called

1. Data leakage
2. Data pre-processing
3. Data harmonization
4. Data wrangling

A: Generalization and dataset preparation

An additional information can allow the model to learn or know something that it otherwise would not know and in turn invalidate the estimated performance of the model being constructed. This is called

1. Data leakage
2. Data pre-processing
3. Data harmonization
4. Data wrangling

On your mark, get set, …

#DoF

The degrees of freedom for a given problem are the number of independent problem variables which must be specified to uniquely determine a solution. Hence the #DoF is important to be considered

1. To design the number of vectors in the learning dataset.
2. To avoid overfitting problem in the model
3. All the above
4. None of the above

A: #DoF

The degrees of freedom for a given problem are the number of independent problem variables which must be specified to uniquely determine a solution. Hence the #DoF is important to be considered

1. To design the number of vectors in the learning dataset.
2. To avoid overfitting problem in the model
3. All the above
4. None of the above

On your mark, get set, ...

Basic metrics in data similarity: cosine

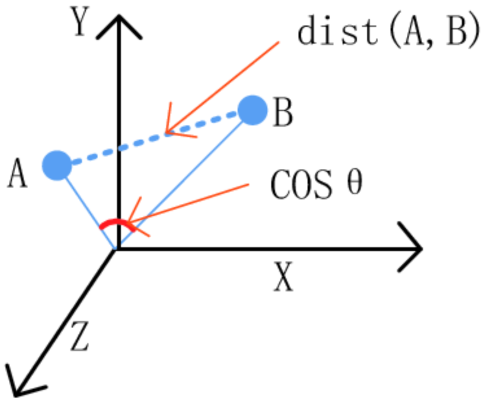
About the cosine metrics it is possible to say that

1. Two vectors with the same orientation have a cosine similarity of 1
2. Two vectors oriented at 90° relative to each other have a similarity of 0
3. All of the above
4. None of the above

Basic metrics in data similarity: cosine

About the cosine metrics it is possible to say that

1. Two vectors with the same orientation have a cosine similarity of 1
2. Two vectors oriented at 90° relative to each other have a similarity of 0
3. All of the above
4. None of the above



On your mark, get set, ...

Similarity in images datasets

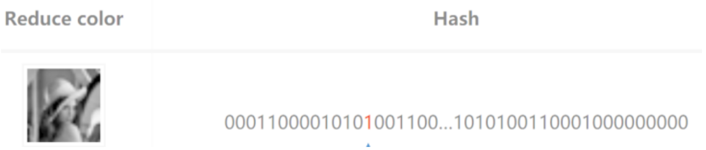
What similarity feature/features discussed in class offers/offer the property to allow a **fast comparison based on a short 1D vector of elements or bits**

1. phash
2. ahash
3. All the above
4. Cross-correlation

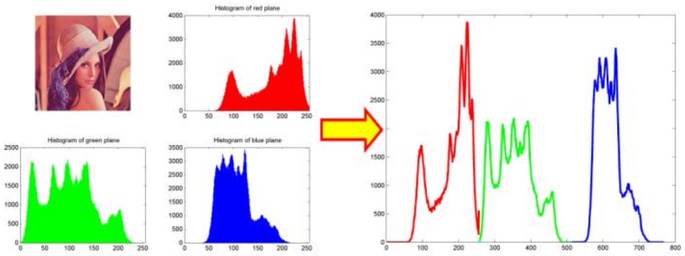
Similarity
in images datasets

What similarity feature/features discussed in class offers/offer the property to allow a **fast comparison based on a short vectors of elements or bits**

1. phash
   (p = Perceptive … Discrete Cosine Transform)
2. ahash
   (a = Average… 0 if gray > below the mean gray)
3. All the above
4. Cross-correlation
   (this is wrong)



Reduce color	Hash
	00011000010101001100...1010100110001000000000



Same with histograms comparison

On your mark, get set, …

Similarity in dataset

In agreement to the class discussion, which description better describes the design activity?

1. Similarity in the dataset
   requires more space and processing time
2. Similarity in the dataset
   can improve generalization
3. Both of the above
4. None of the above

Similarity in dataset

In agreement to the class discussion, which description better describes the design activity?

1. Similarity in the dataset
   requires more space and processing time
2. Similarity in the dataset
   can improve generalization
3. Both of the above
4. None of the above

We have a «dilemma» between case #1 and #2

On your mark,
get set, ...

Dataset preparation: cardinality of checks ops.

In agreement to the class discussion, in a dataset of 1100 labelled images, the search for duplications is typically achieved…

1. by manual exploration of the dataset for better results since the number of images is not critical.
2. by automatic iterations.

Dataset preparation: cardinality of checks ops.

In agreement to the class discussion, in a dataset of 1100 labelled images, the search for duplications is typically achieved…

1. by manual exploration of the dataset for better results since the number of images is not critical
2. by automatic iterations

A basic check→ #iter= N*(N-1)>**1.2M comparisons**

On your mark, get set, …

Dataset preparation: labelling errors

In agreement to the class discussion, what kind of labelling error is generally the worst case for the accuracy of the generalization of the model?

ERR1 = Duplications with same labels

EER2 = Duplications with different labels

1. ERR1
2. ERR2
3. ERR1 = EE2

A: Dataset preparation: labelling errors

In agreement to the class discussion, what kind of labelling error is generally the worst case for the accuracy of the generalization of the model?

ERR1 = Duplications with same labels

EER2 = Duplications with different labels

1. ERR1
2. ERR2
3. ERR1 = EE2

x1

y1 = «cat»

x2

y2 = «dog»

On your mark,
get set, ...

Similarity

According to the class discussion, about the relationship between the operation of cross-correlation and convolution it is possible to say that:

1. They are very similar in meaning and mathematical expression
2. Despite the mathematical expression is similar, the meaning and their use is completely different
3. There is no specific relationship since they are different in meaning and mathematical expressions

A: Similarity

Cross-Correlation:

$$G = h \otimes F \qquad G[i,j] = \sum_{u=-k}^{k} \sum_{v=-k}^{k} h[u,v]F[i+u, j+v]$$

Just the minus!

Convolution:

$$G = h * F \qquad G[i,j] = \sum_{u=-k}^{k} \sum_{v=-k}^{k} h[u,v]F[i-u, j-v]$$

According to the class discussion, about the relationship between the operation of cross-correlation and convolution it is possible to say that:

1. <u>They are very similar in meaning and mathematical expression</u>
2. Despite the mathematical expression is similar, the meaning and their use is completely different
3. There is no specific relationship since they are different in meaning and mathematical expressions

On your mark, get set, ...

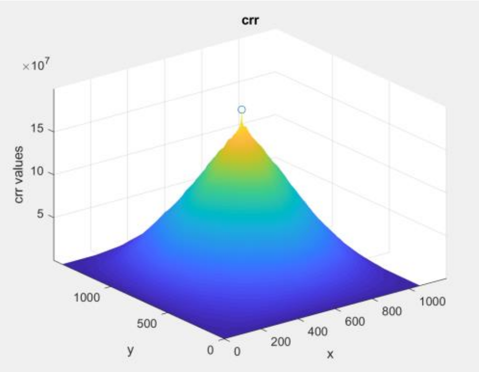
Similarity: autocorrelation

According to the class discussion, what is the characteristic of the self-correlation (O=xcor2(A,A)) map produced by a generic image?

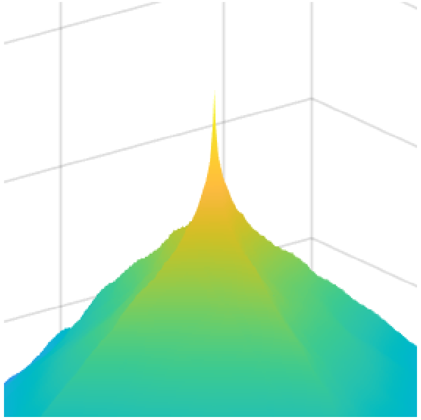
1. A flat and noisy central plateau
2. An evident spike at the center with a very well defined maximum
3. It is not possible to create an autocorrelation map from one single images, two different images are needed

A: Similarity: autocorrelation

According to the class discussion, what is the characteristic of the autocorrelation map produced by a generic image?



1. A flat and noisy central plateau
2. An evident spike at the center with a very well defined maximum
3. It is not possible to create an autocorrelation map from one single images, two different images are needed



On your mark, get set, …

Feature preprocessing/engineering

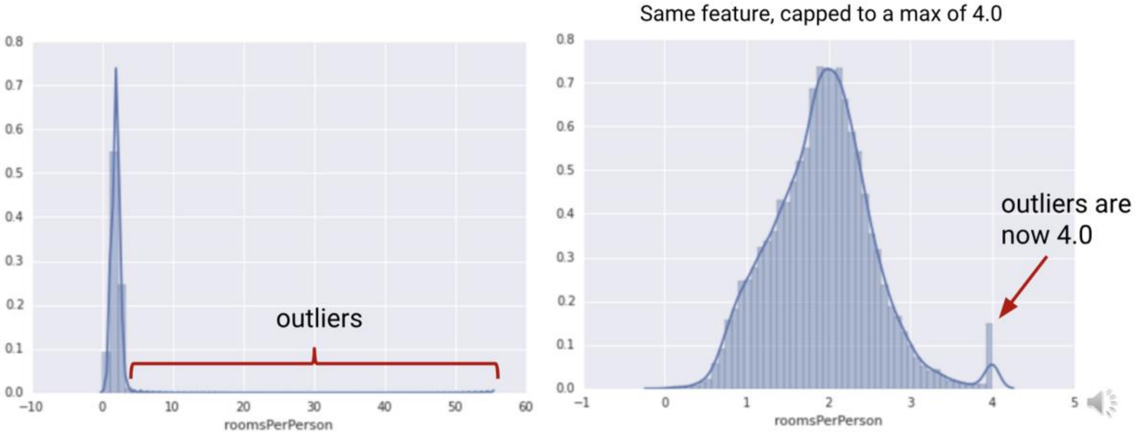
If your data set contains extreme outliers it better to use as preprocessing

1. Feature clipping
2. Min-max normalization
3. Z’ norm

A: Feature preprocessing/engineering

If your data set contains extreme outliers it better to use as preprocessing

1. Feature clipping
2. Min-max normalization
3. Z’ norm



Same feature, capped to a max of 4.0

outliers

outliers are now 4.0

On your mark, get set, …

Feature preprocessing/engineering

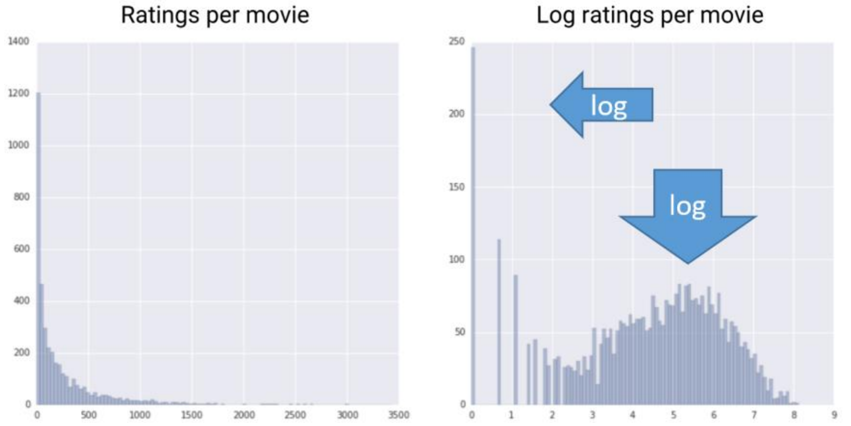
A logarithmic scaling to one feature values is typically applied in a case of

1. Outliers presence
2. Negative values
3. A very large range in the values (>0)

A: Feature preprocessing/engineering

A logarithmic scaling to one feature values is typically applied in a case of

1. Outliers presence
2. Negative values
3. A very large range in the values (>0)



A classical
min-max normalization/z norm
is not effective

On your mark, get set, ...

Plotting of the results

According to the scientific visualization rules presented in class, if you are plotting many figures of merit obtained by your trained neural network on a new dataset, which is the correct ranking of visual attributes to be used?

Left: low accuracy **Right: HIGH ACCURACY**

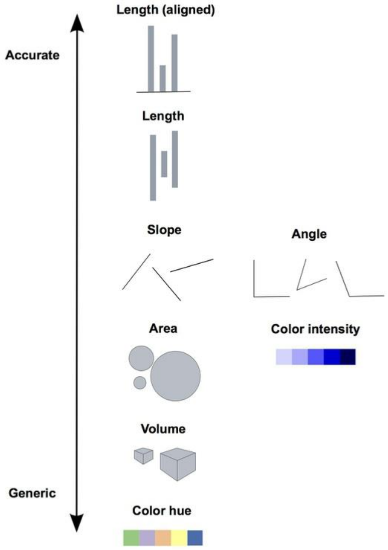
1. Color intensity > Hue > Length
2. Area > Length > Hue
3. Slope > Angle > Volume
4. Hue > Area > Lenght

A: Plotting of the results

According to the scientific visualization rules presented in class, if you are plotting many figures of merit obtained by your trained neural network on a new dataset, which is the correct ranking of visual attributes to be used?

Left: low accuracy **Right: HIGH ACCURACY**

1. Color intensity > Hue > Length
2. Area > Length > Hue
3. Slope > Angle > Volume
4. Hue > Area > Length



On your mark, get set, …

Plotting of the results (2)

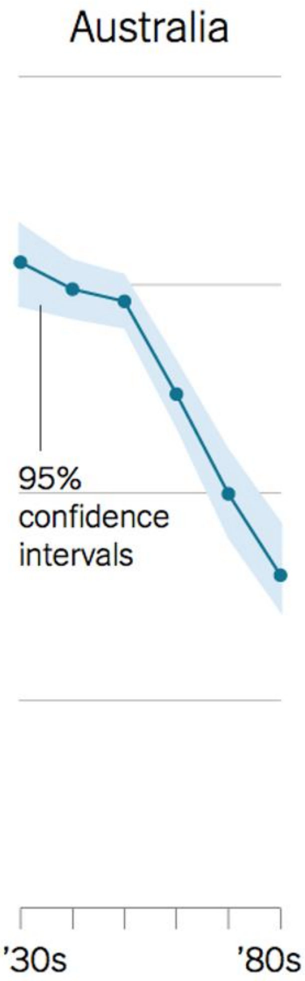
According to the scientific visualization rules presented in class, is it possible to plot a graphical representation of the confidence level of your figures of merit of your trained model?

1. No, it is a statistical index with different units and meaning and hence can not be represented in the same plot
2. Yes, the confidence interval data have the same units and meaning and they can be represented in the same plot

A: Plotting of the results (2)

According to the scientific visualization rules presented in class, is it possible to plot a graphical representation of the confidence level of your figures of merit of your trained model?

1. No, the confidence intervals data have different units and meaning and hence can not be represented in the same plot
2. Yes, the confidence interval data have the same units and meaning and they can be represented in the same plot



On your mark, get set, ...

Plotting of the results (3)

According to the discussion presented in class about the data visualization, and considering the following steps of the design workflow
1) Get Data, 2) Clean Manipulate Data,
3) Train models, 4) Test Data, 5) Improve the design, which are the main step/steps where data visualization should be involved?

A. #1

B. #5

C. #1 and #5

D. #2, #3 and #5

A: Plotting of the results (3)

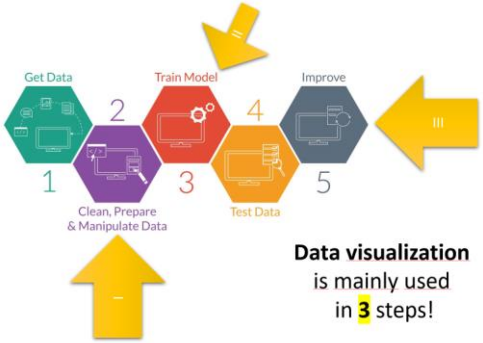
According to the discussion presented in class about the data visualization, and considering the following steps of the design workflow
1) Get Data, 2) Clean Manipulate Data,
3) Train models, 4) Test Data, 5) Improve the design, which are the main step/steps where data visualization should be involved?

A. #1

B. #5

C. #1 and #5

D. #2, #3 and #5



Data visualization is mainly used in 3 steps!

On your mark, get set, …

Similarity

According to the discussion presented in class about the similarity,

Consider an image A(x,y) with internal similarity (repetitions of patterns).

What happens to the output of the self crosscorrelation (O = xcorr2(A,A))

A. It is not possible to apply the crosscorrelation to the same image
B. Output O tends to be a flat plateau with one clear central peak
C. Output O tends to have many peaks and one evident maximum
D. Output O tends to have many equivalent peaks with the same maximum value

Similarity

According to the discussion presented in class about the similarity,

Consider an image A(x,y) with internal similarity (repetitions of patterns).

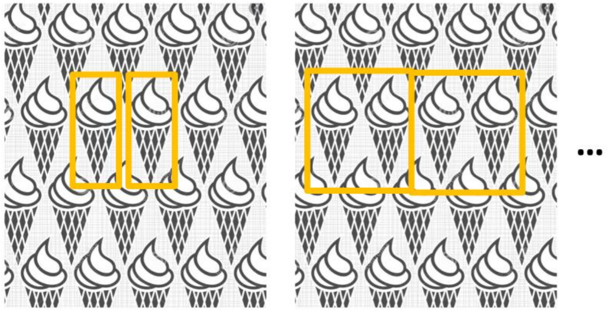
What happens to the output of the self crosscorrelation (O = xcorr2(A,A))

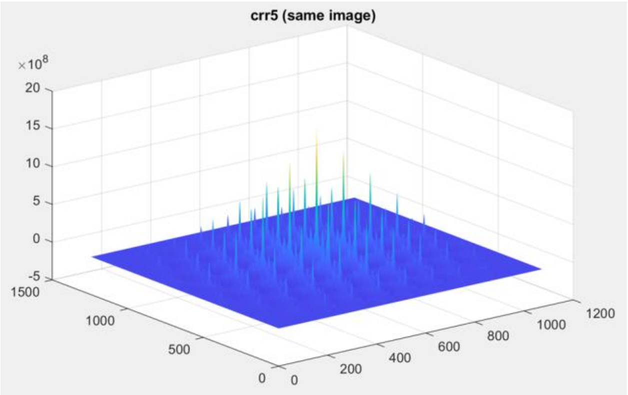
A. It is not possible to apply the crosscorrelation to the same image

B. Output O tends to be a flat plateau with one clear central peak

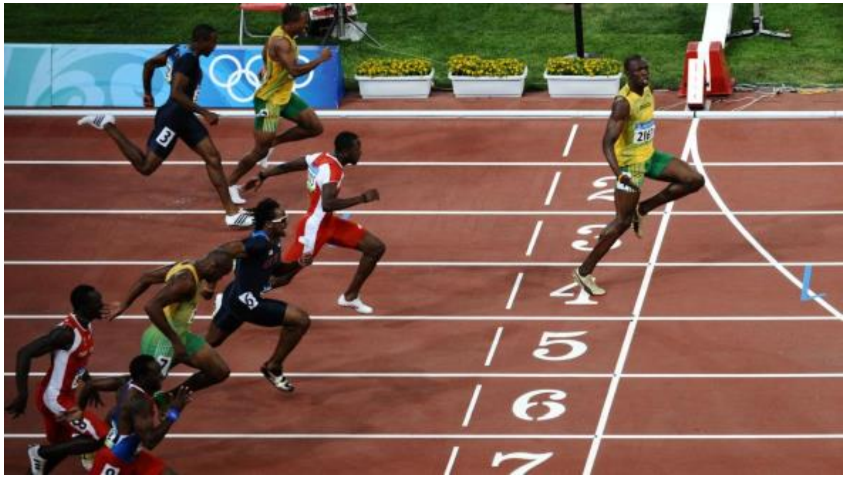
C. Output O tends to have many peaks and one evident maximum

D. Output O tends to have many equivalent peaks with the same maximum value





The End



How is your accuracy?

Do not hesitate to contact the teacher in case of doubts

What about topics not present in the simulation?

- They are equally important
    - The absence in the simulation does not mean they are not relevant
- You can find them in the exam
- Use the simulation not to study but to understand you are profitably attending the course and, in case, change the study method
- Try to create your own simulation by browsing all slides of the course and ask yourself: what can be asked in this slide?

COVID19

- During the COVID19 emergency the exam format for remote exams is almost the same of exams in presence
- With the simulation you have here, you can test your current learning level and accuracy
  - The number of questions is **not** indicative
  - The format of the questions is **not** indicative
  - The level of knowledge **is** indicative