# CHAP1

"Is the 'Analytic approach' an important step in the process of building a data science product?

- Certainly, it is.
- No, we can skip this step.
- No, it is not related to data science.

What does the 'Analytic approach' step entail in the process of building a data science product?

☑ Transforming a real-world problem into a data science problem.
☐ Selecting an analytic tool to solve a data science problem.
☐ Transforming a data science problem into a real-world problem.

Is understanding the real-world problem ('Business understanding') an important step in the product-oriented data science process?

☑ Yes, of course.
☐ No, we can skip this step.
☐ No, it has nothing to do with Data Science.

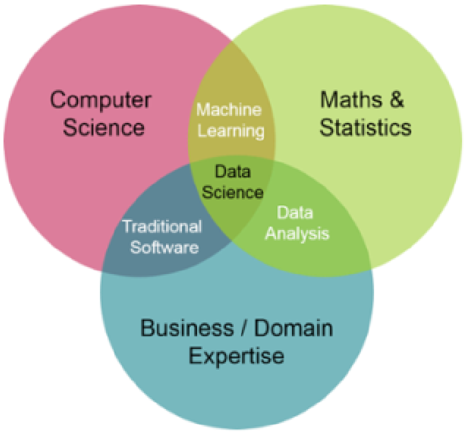
What does the 'Business understanding' step involve in the data science product development process?

☑ Understanding the actual needs that need to be addressed.
☐ Understanding the main business contents of the organization.
☐ Understanding the relationship between business needs and Data Science.

Should the data understanding/visualization phase in the Data Science process be conducted after the modeling phase?

☐ Yes.
☐ No, it should be done before.
☐ All answers are correct.

What does the following figure suggest?



- □ Data science is an interdisciplinary field touching many other areas.
- □ Data science is the core commonality of Computer Science, Mathematics, Statistics, and Domain Knowledge.
- □ Data science is a field that intersects Computer Science, Mathematics, Statistics, and Domain Knowledge.

Is 'Evaluation' a core step in the Data Science process, whether product-oriented or knowledge exploration-oriented?

- □ Yes.
- □ No, it's not necessarily needed when we want to explore new knowledge from data.
- □ No.

What can 'Evaluation' in the Data Science process include?

- □ Analyzing and comparing results from chosen scenarios (including offline and real-life scenarios).
- □ Evaluating the deployment of a system in practice.

In Data Science, what is the main difference between data cleaning and preprocessing?

- □ Data cleaning mainly deals with noisy data, while data preprocessing mainly deals with missing data.

- □ Data cleaning mainly deals with noisy data, while data preprocessing mainly deals with redundant data.
- □ Data cleaning is usually performed before data preprocessing and aims to detect dirty data.
- □ Data preprocessing includes data cleaning steps.

Is 'Prediction' the main task of Data Science?

- □ Yes.
- □ No, it's just one of the tasks in Data Science.
- □ No, it does not belong to the field of Data Science.

What does the following statement say about 'The curse of dimensionality'?

- □ As the dimensionality of data increases, the size of the data space increases so rapidly that the datasets we collect become too sparse. This sparsity creates significant challenges for data analysis methods.
- □ As the dimensionality of data increases, the difficulty in data analysis is not significantly affected.
- □ High dimensions can create many challenges for storage and computation.

'Vagueness' is a challenge in the era of Big Data and it refers to ...

- □ The difficulty of understanding data.
- □ The communication problem between the provider and the user.
- □ The challenge for a non-expert to interpret the analysis results.
- □ The complexity of data patterns in a streaming environment.
- □ The complexity of data analysis algorithms.

'Variability' is a challenge in the era of Big Data and it refers to ...

- □ The high variation in data.
- □ Possible changes in the structure of data sources.
- □ The rate at which data arrives in a streaming environment.
- □ Different rates at which data sources are refreshed.

-> • The possible evolutions in the structure of the data sources
• The different velocities at which these data sources are refreshed

• Poses serious issues for data integration

'Velocity' is a challenge of the big data era and it refers to

- □ The strong changing nature of data.
- □ Large-scale computations.
- □ The continuous and rapid nature of data arrival.
- □ The speed of data analysis.

'Veracity' is a challenge of the big data era and it refers to

- □ The strong changing nature of data.
- □ Large-scale computations.
- □ The continuous flow of data in a streaming environment.
- □ The high degree of uncertainty due to noise, errors, losses, bias... in data.

“Veracity” is a challenge related to big data and refers to

- □ Different types of data that must be handled: structured/unstructured data.
- □ The computing power that big data requires.
- □ Data that arrives continuously and rapidly.
- □ Data with a high degree of uncertainty due to the presence of fake information/noise in some sources (especially on the internet).

Variety is a challenge related to big data and refers to

- □ Different types of data that must be handled: structured/unstructured data.
- □ The computing power that big data requires.
- □ Data that arrives continuously and rapidly.
- □ Data with a high degree of uncertainty due to the presence of fake information/noise in some sources (especially on the internet).

Data science is an interdisciplinary field and goes beyond the scope of Computer Science.

- □ True.

□ False.

# CHAP2+3: Data crawling and preprocessing & Data cleaning and integration

What is an example of a correct xpath?

- □ /node/text()
- □ //Parent[@id='1']/Children/child/@name
- □ span::text
- □ base::attr(href)

What is an example of a correct xpath?

- □ //a[contains(@href, "image")]/@href
- □ a[href*=image]::attr(href)
- □ //base/@href
- □ base::attr(href)

What is an example of a correct css selector?

- □ //a[contains(@href, "image")]/@href
- □ a[href*=image]::attr(href)
- □ a[href*=image]@attr(href)
- □ //a[contains(@href, "image")]::@href

What is the purpose of the Page Rank algorithm?

- □ To find the most relevant results for a query.
- □ To measure the importance of a webpage.
- □ To determine the popularity of a webpage.
- □ To rank the results of a search engine.

How can desired data be extracted when writing a data scraping bot on scrapy?

- □ Using xpath and css selectors to write downloaders.
- □ Using xpath and css selectors to write spiders.
- □ Using xpath and css selectors to write item pipelines.

What is an example of a correct css selector?

- □ /node/text()
- □ //Parent[@id='1']/Children/child/@name

- □ span::text
- □ base::attr(href)

Can a Scrapy bot ignore the information in robots.txt?

- □ Yes.
- □ No.

Does Scrapy support a default incremental crawling strategy?

No.

Yes.

Can using robots.txt block internet data scraping programs?

No.

Yes.

According to Scrapy, where can web page load requests be forced to use a proxy?

- □ Downloader middlewares.
- □ Spider middlewares.
- □ Downloader.
- □ Spider.
- □ Item pipelines.

In Scrapy, where can the user-agent attribute be changed during data collection?

- □ Downloader middlewares.
- □ Spider middlewares.
- □ Downloader.
- □ Spider.

In Scrapy, how can collected data be written into databases?

- □ Write code to add in spider middleware.
- □ Write code to add in item pipelines.
- □ Write code to add in the downloader.

In Scrapy, what is the role of the downloader component?

- □ Receive requests from the engine component, put these requests in a queue for later processing.
- □ Download web pages.
- □ Extract responses.

In Scrapy, what is the role of the spider component?

- □ Extract responses.
- □ Download web content.
- □ Coordinate the data flow between all components of Scrapy.

Which algorithm is used to rank web pages in search engine results?

- □ Webrank.
- □ Pagerank.
- □ Textrank.

Which of the following accurately describes XPath?

- □ XPath is a programming language.
- □ XPath is a query language.
- □ XPath is an XML file structure.
- □ XPath can be read by Microsoft Word.

Which step in the following data cleaning method is not in the correct order?

A. Extract relevant data fields.
B. Repair data quality issues at the value level.
C. Normalize data values.
D. Address data quality issues at the value set level.
E. Address data quality issues at the relational level.
F. Repair data quality issues at the multi-relational level.
G. Gather user feedback.
thứ tự đúng: A -> B -> D -> E -> F -> C

Can Google Openrefine import data from the Internet via URL?

- □ Yes.
- □ No.

Can Google Openrefine be used to automatically group data?

- □ Yes.
- □ No.

What is faceting in Google Openrefine?

- □ Allows seeing an overview of the data.
- □ Allows filtering down to a subset of rows that you want to change in bulk.
- □ Allows making trend judgments from data.

Why is real-world data not clean?

- □ Incomplete.
- □ Noisy.
- □ Inconsistent.

What characterizes data modeling in OLAP?

- □ The database schema needs to be normalized to ensure data consistency.
- □ Usually uses denormalized database schemas.
- □ Usually uses a multidimensional data model.

Which is not a cause of noisy data?

- □ Faulty data collection equipment.
- □ Errors made by people entering data into the system.
- □ Due to different data needs between the time of data collection and the time of data analysis.

Which is not a data quality issue at the value level?

- □ Missing values.
- □ Syntax violations.
- □ Synonyms.

- Missing value, Syntax violation, Spelling error, Domain violation

Which is not a data quality issue at the value set level?

- □ Existence of homonyms
- □ Violation of uniqueness.

□ Violation of integrity constraints.
□ Violation of defined sets.

- Existence of synonyms, Existence of homonyms, Uniqueness violation, Integrity contraint violation

What characterizes OLAP?

□ Mainly transactions of adding, editing, and deleting with short execution times.
□ Queries are often complex and involve aggregate operations.
□ Mainly ad-hoc queries.
□ Often accesses many data records.
□ Supports decision-making.

What characterizes OLTP?

□ Mainly transactions of adding, editing, and deleting with short execution times.
□ Supports the processing of operational transactions for businesses.
□ Often accesses historical data and multidimensional data.
□ Usually involves complex queries.

Which of the following is incorrect about OLAP?

□ Processes historical information (created in the past).
□ Supports business analysis.
□ Scalable to allow millions of users.
□ Stores millions of data records.

What describes a Wrapper in the virtual data integration architecture?

□ A piece of code that transforms data from the source format to the standardized format of the mediator.
□ Can be implemented at the data source side or the mediator side.
□ An essential component of the mediator.

What metadata is included in the Data Source Catalog in the virtual data integration architecture?

□ List of data versions of the source.

- ☐ Query capabilities of the source (e.g., SQL response capability).
- ☐ Data update frequency.
- ☐ Access control and authorization.

-> • Logical source contents, Source capabilities, Source completeness , Physical properties of source and network, Statistics about the data, Source reliability, Mirror sources, Update frequency

What accurately describes Apache Nifi?

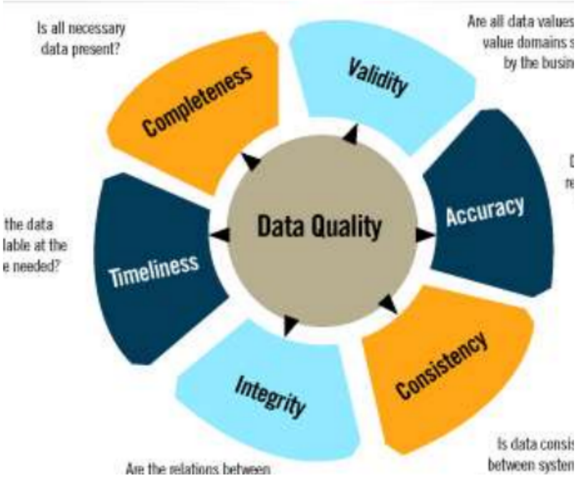
- ☐ An ETL tool.
- ☐ A data warehouse platform allowing the storage of large-sized data.
- ☐ A tool for data cleaning and preprocessing.

What concepts are included in Apache NiFi?

- ☐ FlowFile
- ☐ FlowFile Processor
- ☐ Scheduler
- ☐ Process Group
- ☐ Flow Controller

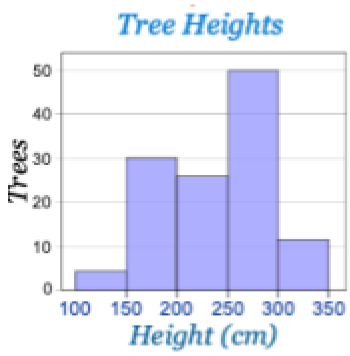
What are the dimensions of measurement when discussing data quality?

- ☐ Completeness, Validity, Integrity
- ☐ Timeliness, Accuracy, Consistency
- ☐ Timeliness, Isolation
- ☐ Completeness, Integrity, Value


Is all necessary data present?
Are all data values value domains s by the busin
Completeness
Validity
Data Quality
Accuracy
the data lable at the e needed?
Timeliness
Integrity
Consistency
Are the relations between
Is data consis between systen

## Chap4: Exploratory data analysis

What does this chart represent?



- □ A bar chart showing the height of trees and the corresponding number of trees for each height.
- □ A histogram showing the distribution of tree heights
- □ A bar chart drawn incorrectly, needs to rename the median and x-axis.

What conclusions can be drawn from a box plot in exploratory data analysis?

- □ Are there any important features (variables)?
- □ Is there a difference in location concentration between subgroups?
- □ Is there a difference in variation between subgroups?
- □ Are there any outliers?

• Is a factor significant? • Does the location differ between subgroups? • Does the variation differ between subgroups? • Are there any outliers?

What conclusions can be drawn from a histogram in exploratory data analysis?

- □ Examine the distribution of a set of observations.
- □ Examine the concentration of data.
- □ Examine the dispersion of data.
- □ Is the data distribution symmetrical or skewed?
- □ Are there any outliers in the data?

• What kind of population distribution do the data come from? • Where are the data located? • How spread out are the data? • Are the data symmetric or skewed? • Are there outliers in the data?

What conclusions can be drawn from a scatter plot in exploratory data analysis?

□ Is there a relationship between variable X and Y?
□ Is the relationship linear?
□ Does the variation of variable Y depend on variable X?
□ Which variable, X or Y, is more important?

What is Exploratory Data Analysis (EDA)?

□ EDA is not a set of techniques but a philosophy on how we should proceed when we want to understand data.
□ EDA is a set of techniques that allows us to understand data, including the use of charts and statistical techniques.
□ EDA involves using charts to understand data.

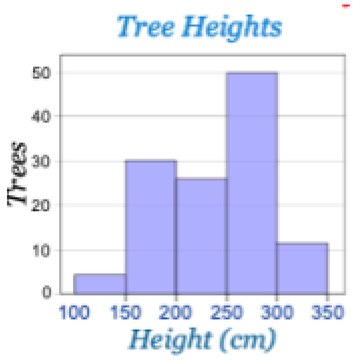
How is Exploratory Data Analysis (EDA) performed?

□ Examine descriptive properties, central measures, and dispersion measures of data.
□ Examine data distribution.
□ Examine relationships among variables in data.
□ Examine data structure characteristics.

What is the focus of Exploratory Data Analysis (EDA)?

□ EDA focuses on the structure, exceptions, and patterns from data.
□ EDA focuses on all data points in the dataset.
□ Visualization and data cleaning.
□ EDA focuses on tools that allow examining the structure and exceptions from data.

Which statement is incorrect regarding the chart below?



- □ The number of trees with heights from 250 to 300 is the highest.
- □ The number of trees with heights from 100 to 150 is the smallest.
- □ There are 30 trees with a height of 150.
- □ There are fewer than or equal to 50 trees with a height of 300.

For the same dataset, what factors does the result of the K-means clustering algorithm depend on?

- □ The way of measuring distance.
- □ The way of merging clusters.
- □ The configuration of the initial number of clusters K.
- □ An initial guess for the centroids.

For the same dataset, what factors does the result of the hierarchical clustering algorithm depend on?

- □ The way of measuring distance.
- □ The way of merging clusters.
- □ The configuration of the initial number of clusters K.

What libraries and tools can be used to perform exploratory data analysis?

- □ NLTK, Spacy
- □ Requests, Scrapy, BeautifulSoup
- □ Tensorflow, Keras, Scikit-learn
- □ SciPy, NumPy, Matplotlib, and Pandas

## CHAP5+6+7: Data visualization Machine learning

Generalization and Overfitting are two opposing aspects of machine learning models.

- □ True.
- □ Not necessarily opposing each other.
- □ They are two independent features.

Assuming you want to use a machine learning method to analyze hidden knowledge within a dataset without any notion of that knowledge. Which of the following problems is most suitable?

- □ Unsupervised learning.
- □ Supervised learning.
- □ Regression.
- □ Multiclass classification.

The learning process of a decision tree by the ID3 algorithm stops if

- □ The tree has completely classified the training data correctly, or at any path from root to leaf, all attributes have been used.
- □ The tree has completely classified the training data correctly.
- □ The tree is large enough.
- □ The tree cannot completely classify the training data correctly.

What is the role of "information gain" in the ID3 algorithm when learning a decision tree?

- □ To measure the discriminating capability of attributes to find a testing attribute at each vertex.
- □ To see how good an attribute is after the training process.
- □ To measure errors at each vertex in the tree.
- □ It has no role.

What is K-means?

- □ A clustering method.
- □ A classification method.
- □ A supervised learning method.
- □ A method for computing the arithmetic mean from data.

What can a machine learning method learn?

- □ A function that can map an input data point to an output.
- □ New knowledge to predict outputs.
- □ To simulate human capabilities.
- □ Anything.

What is the difference between supervised and unsupervised learning?

- □ The training set in supervised learning typically requires labels/outputs for each data sample.
- □ The type of output in supervised learning usually involves real numbers.
- □ The way we train a supervised learning model usually requires step-by-step instructions on how to learn.
- □ The goal of unsupervised learning algorithms usually does not make any predictions.

The least squares method learns a function

$$\Leftrightarrow \boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{M} (y_i - w_0 - w_1 x_{i1} - \cdots - w_n x_{in})^2$$

The role of the empirical loss function is:

- □ To measure the error in predictions in some sense and is often used as the objective function when training a model
- □ To measure the error in future predictions
- □ Has no role

Machine Learning (ML) provides methods to Analyze data and make predictions for future data

- □ True
- □ False, it provides foundations for computational expansion
- □ True, it also provides platforms to accelerate computation

The "No-free-lunch" theorem states:

- □ No algorithm can outperform another algorithm across all problem domains
- □ There is no free lunch for anyone
- □ Without maximum effort, an algorithm cannot outperform other algorithms

Overfitting refers to which situation?

- □ A method that produces a low error rate on the training set but a high error rate on future data.
- □ A method that can make incorrect predictions about the behavior of another method.
- □ Too little training data.
- □ There is so much training data that a computer can learn easily.

Which of the following techniques can help reduce overfitting?

- □ Using regularization, a technique that often helps limit the search space when training a model.
- □ Using a new method/model.
- □ Removing some data if there is too much.

Where does machine learning appear in a data science process?

- □ The modeling step, where we use a specific method to analyze the data.
- □ The data understanding step.
- □ The step of choosing an approach to solve the current problem.

# CHAP8: BIG DATA ANALYSIS

How to analyze scalable data for big data?

- □ Parallelize machine learning algorithms.
- □ Use real-time processing architectures.
- □ Use Principal Component Analysis (PCA).
- □ Use deep neural network models.

Indicate the correct statement:

- □ Hadoop needs to run on high-configuration specialized hardware to process big data.
- □ Hadoop 2.0 and above allow running jobs that are not MapReduce jobs.
- □ In the Hadoop programming framework, the result files are divided into lines or records.
- □ None of the answers is correct.

Choose the correct statement:

- □ MapReduce brings data to computing nodes.
- □ MapReduce brings computing to nodes containing data.
- □ Data for MapReduce must be on HDFS.
- □ All of the answers.

Which of the following statements is incorrect about Apache Hadoop?

- □ It processes distributed data with a simpler and more user-friendly programming model like MapReduce.
- □ Hadoop is designed to scale out by increasing the number of servers.
- □ It is designed to operate on common hardware with the ability to withstand hardware failures.
- □ It is designed to operate on supercomputers with high configuration and reliability.

Which tool can be used to support import and export of data into and out of the Hadoop ecosystem?

- □ Oozie.
- □ Flume.
- □ Sqoop.
- □ Hive.

What is the role of YARN?

- □ Manages and distributes resources in the Hadoop cluster.
- □ Provides a high-level user interface transforming queries into MapReduce jobs.
- □ Provides high-reliability distributed coordination functions like cluster membership, leader election, and system state monitoring.

Hadoop is an ecosystem consisting of which components:

- □ MapReduce, YARN.
- □ MapReduce, MySQL.
- □ MapReduce, Skykeeper.
- □ MapReduce, Heron.

Hadoop achieves reliability through data replication across multiple servers, thus not requiring ...... on these server nodes.

- □ RAID.
- □ Local file system.
- □ Operating system.

The ...... function is responsible for aggregating results from Map() tasks.

- □ Reduce.
- □ Map.
- □ Sort.
- □ None of the options.

How is data organized in Datanode chunks in HDFS?

- □ Chunks are files in the local file system of the datanode server.
- □ Chunks are continuous data areas on the hard drive of the datanode server.
- □ Chunks are reliably stored on datanode using RAID mechanism.

What is the data replication mechanism in HDFS?

- □ Namenode decides the location of chunk replicas on datanodes.
- □ The primary datanode decides the location of chunk replicas at secondary datanodes.
- □ The client decides the storage location for each chunk replica.

HDFS is programmed in which language?

- □ C++.
- □ Java.
- □ Scala.
- □ None of the answers is correct.

The ...... task is responsible for processing one or several data chunks and returning intermediate results.

- □ Map.
- □ TaskTracker.
- □ All of the options.
- □ Reduce.

The ...... component is responsible for executing tasks assigned by the JobTracker.

- □ MapReduce.
- □ Mapper.
- □ TaskTracker.
- □ JobTracker.

Which scenario might not be suitable for HDFS?

- □ Random read and write into a file.
- □ Storing data related to applications requiring low-latency data access.

- □ Storing small-sized files.
- □ None of the answers is correct.

State the correct statement:

- □ A MapReduce job typically divides the input data set into independent chunks processed by map tasks in a completely parallel manner.
- □ MapReduce views data as key-value pairs.
- □ Applications usually implement Mapper and Reducer interfaces to execute map and reduce methods.
- □ MapReduce only works with data on Hadoop HDFS.

State the correct answer:

- □ Hive is not a relational database but a SQL query engine for querying data.
- □ HBase is a large-scale database supporting SQL.
- □ Pig is a relational database supporting SQL.
- □ All of the options.

What is an authority page on a topic?

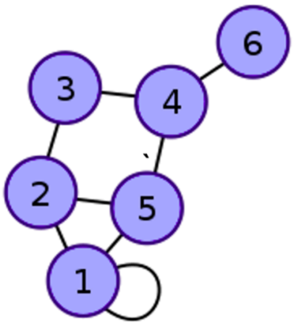
- □ A page that is linked to by many good hubs.
- □ A page that is linked to by many authority pages.
- □ A page that links to many good hubs.

Considering matrix P obtained by adding 0.1 to all elements of the transition probability matrix P above, does P create an ergodic Markov chain?

$$\boldsymbol{P} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

- □ Yes.
- □ No.
- □ We cannot say anything about the ergodic nature.

What is the value of cell [11] in the adjacency matrix of the following graph? 2



What is an ergodic Markov chain?

- □ A chain that allows us to gradually move from any state to any other state with positive probability.

đi dần dần

- □ A chain that allows us to move directly from any state to any other state with positive probability.
- □ A chain in which there exists a pair of states that cannot reach each other.

## Lecture 10+11: Text, image, graph analysis

How does the PageRank algorithm rank web pages?

- □ PageRank uses the long-term visit rate of each web page, and this rate is calculated from the transition probability matrix.
- □ PageRank uses the number of inbound links to each web page.
- □ PageRank uses the number of outbound links from each web page.
- □ PageRank ranks randomly.

Which task is not included in link analysis?

- □ Graph ranking.
- □ Community detection.
- □ Link prediction.
- □ Sentiment analysis.

The Power method can...

- □ calculate the long-term visit rate for each web page.
- □ calculate the steady-state probability distribution for a Markov chain.
- □ use a Markov chain to predict a sequence of pages that will be visited.
- □ calculate a sequence of pages that will be visited given a starting point.

Given the transition probability matrix P above, does P create an ergodic Markov chain?

$$\boldsymbol{P} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

- □ Yes.
- □ No.

What is the difference between the base set (S) and the root set (W) in the HITS algorithm?

- ☐ The base set is built from the root set.
- ☐ The root set is built from the base set.
- ☐ The base set is the basis for evaluating the quality of pages found by HITS based on the root set.

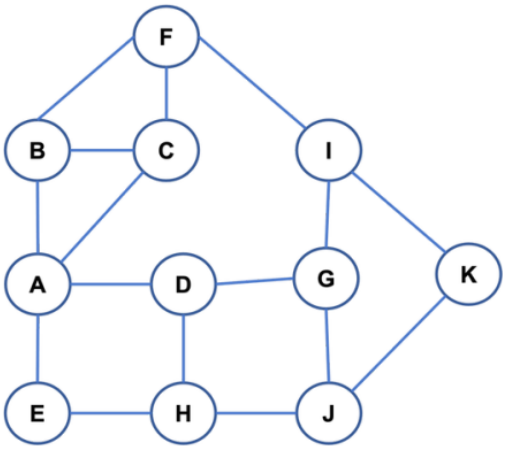
How does the HITS algorithm rank web pages?

- ☐ HITS finds a small set of hubs and authority pages using an iterative algorithm to calculate scores for the pages.
- ☐ HITS finds a small set of authority pages using an iterative algorithm to calculate the steady-state probability distribution.
- ☐ HITS finds a small set of authority pages using an iterative algorithm to calculate the long-term visit rate.

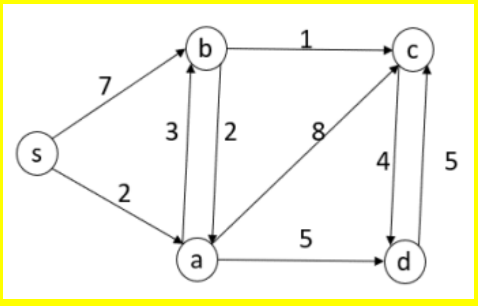
Among the following centrality measures, which one only depends on the adjacent vertices of the vertex being considered?

- ☐ Closeness centrality.
- ☐ Betweenness centrality.
- ☐ Degree prestige.
- ☐ Proximity prestige.

How many shortest paths are there from A to K in the following graph? 6



Using Dijkstra's algorithm, what is the length of the shortest path from s to c? 6



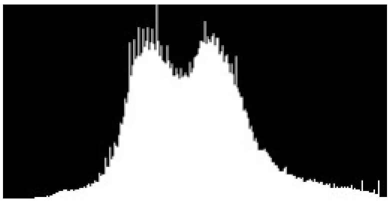
If the brightness of a multi-level gray image is 255, which of the following characteristics does the image have?

- □ The image is entirely white.
- □ The image is entirely black.
- □ The image has some black blocks and some white blocks.
- □ Nothing special, the pixels can take a variety of values within its value domain.

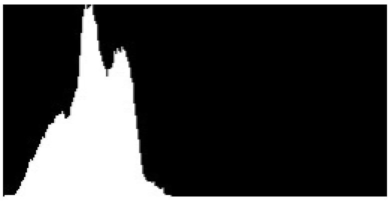
If the brightness of a multi-level gray image is 0, which of the following characteristics does the image have?

- □ The image is entirely white.
- □ The image is entirely black.
- □ The image has some black blocks and some white blocks.
- □ Nothing special, the pixels can take a variety of values within its value domain.

Given 2 histograms corresponding to 2 images like the picture below, which of the following statements is correct?



Histogram của ảnh I1



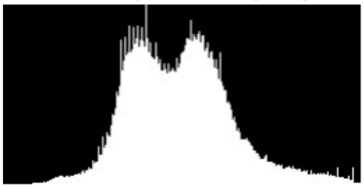
- □ The brightness of image I1 is higher than the brightness of image I2.
- □ The brightness of image I1 is lower than the brightness of image I2.
- □ The brightness of the two images is similar.
- □ The brightness of the two images cannot be compared.

For a 256-level grayscale image that is uncompressed, how many bytes are needed to store each pixel? 1

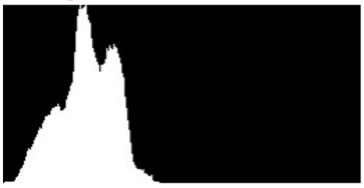
In which color space are the color component and brightness not encoded separately in the channels?

- □ RGB.
- □ HSV.
- □ Lab.
- □ YCbCr.

Given 2 histograms corresponding to 2 images like the picture below, which of the following statements is correct?



Histogram của ảnh I1



Histogram của ảnh I2

The contrast of image I1 is better than the contrast of image I2.
The contrast of image I2 is better than the contrast of image I1.
The contrast of image I2 is similar to the contrast of image I1.
The contrast of image I1 and I2 cannot be compared.

Which of the following statements is correct?

- □ The histogram of two different images can be the same.
- □ The histogram of two different images is always different.
- □ The histogram of an image always has 256 levels (256 bins).
- □ If the objects in the image are shifted to the left by 10 pixels, the histogram of the image is also shifted to the left.

What is the purpose of histogram equalization?

- □ To enhance the contrast of the image.
- □ To increase the brightness of the image.
- □ To represent the content of the image.
- □ To reduce noise.

For a 256-level grayscale image, in what range do the pixel brightness values fall?

- □ \([0, 255]\)
- □ \([0, 100]\)
- □ \([0, 256]\)
- □ \([1, 256]\)

If we take pictures of the same object under different lighting conditions and represent them in the Lab color space, which color channel will show the biggest difference between the two images?

- □ L.
- □ a.
- □ b.
- □ a and b.
- □

How many channels are in an RGB image? 3

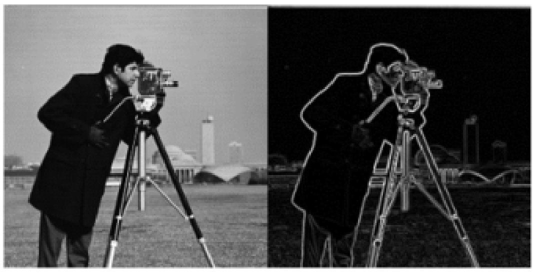
What is the purpose of the Canny detector?

- □ Edge detection.
- □ Local feature extraction.
- □ Global feature extraction.
- □ Noise removal.

What order of derivative does the Canny detector use on an image?

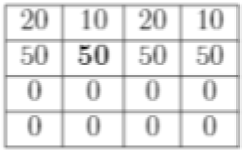
- □ First-order derivative.
- □ Second-order derivative.
- □ Both first and second-order derivatives.
- □ Does not use first or second-order derivatives.

Given the original image on the left, which filter was used to obtain the result image on the right?

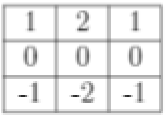


- □ Sobel filter.
- □ Median filter.
- □ Gaussian filter.
- □ Average filter.

Given a (4x4) pixel matrix and a convolution mask, what is the value of the pixel (11) (bolded point) after convolution?



| | | | |
|---|---|---|---|
| 20 | 10 | 20 | 10 |
| 50 | **50** | 50 | 50 |
| 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 |



| | | |
|---|---|---|
| 1 | 2 | 1 |
| 0 | 0 | 0 |
| -1 | -2 | -1 |

Mask

-60.

Which of the following statements about 2D convolution is NOT correct?

- □ The new value of a pixel is calculated by the weighted sum of pixel values in its neighborhood.
- □ The same function is applied to all pixels.
- □ 2D convolution can be used for noise removal, enhancing image sharpness, or edge detection.
- □ None of the mentioned answers.

Which of the following statements is correct?

The Laplace mask can be used to calculate the second derivative of an image.

The Laplace mask can be used to calculate the first derivative of an image.

The second derivative of an image cannot be approximated by convolution.

None of the mentioned answers.

How are edge points determined?

- □ Find zero-crossings on the second derivative.
- □ Find local extremes on the first derivative.
- □ Find zero-crossings on the first derivative.
- □ Find local extremes on the second derivative.

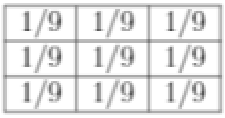
Which of the following statements about image features is correct?

- □ Local features describe the content of a specific region in an image.
- □ Local features represent information of the entire image.
- □ The histogram of an image is a local feature.
- □ None of the mentioned answers.

Which of the following statements about image features is correct?

- □ Global features represent information of the entire image.
- □ Local features describe the content of a specific region in an image.
- □ SURF is a global feature.
- □ None of the mentioned answers.

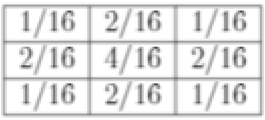
Which mask below is used for which filter?



| | | |
|---|---|---|
| 1/9 | 1/9 | 1/9 |
| 1/9 | 1/9 | 1/9 |
| 1/9 | 1/9 | 1/9 |

- □ Average filter.
- □ Median filter.
- □ Gaussian filter.
- □ Edge enhancement filter.

Which mask below is used for which filter?



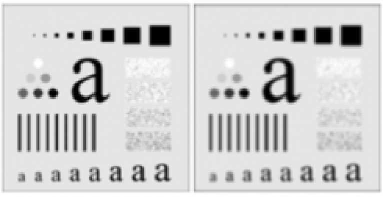
| | | |
|---|---|---|
| 1/16 | 2/16 | 1/16 |
| 2/16 | 4/16 | 2/16 |
| 1/16 | 2/16 | 1/16 |

- ☐ Average filter.
- ☐ Median filter.
- ☐ Gaussian filter.
- ☐ Edge enhancement filter.

How are regions for calculating local features determined?

- ☐ Using image segmentation methods.
- ☐ Dividing the image into pieces using a predefined grid.
- ☐ Detecting feature points and defining local regions around those points.
- ☐ All of the mentioned options.

The two images below are results obtained when applying average masks of different sizes on the same image. If the left image is the result of a 9x9 filter, what is the corresponding size of the mask for the right image?



- ☐ 15x15.
- ☐ 9x9.
- ☐ 5x5.
- ☐ 3x3.
- ☐

What is SIFT?

- ☐ A local feature.
- ☐ A global feature.
- ☐ A contrast enhancement method.
- ☐ An edge detection tool.

## CHAP 12: Evaluation of analysis results

Suppose you use K-means to analyze data from Facebook to find specific user groups. Increasing the number of groups K always reduces clustering error on the training set. You might have difficulty choosing K to get the best clustering result. What should you do?

Find one (or a few) experts in that field to evaluate the quality of the found groups/clusters.

Choose K with the smallest clustering error

"Is hold-out a method for preprocessing and understanding data?"

- ☐ No, it is a strategy for evaluating a model.
- ☐ Correct, of course.
- ☐ No, it is a method for training a model from a given dataset.

Suppose you have built a system to detect network attacks and are certain that its accuracy on the test set is 99%. However, your boss says the system is not usable in reality. What could be the reason?

- ☐ Your assessment of the system might be incorrect.
- ☐ The accuracy might not reflect what the boss expects.
- ☐ The boss may not have enough knowledge to understand the system and your efforts.
- ☐ The training set might be too simple.
- ☐ You are just unlucky.

Model evaluation is:

- ☐ The process of evaluating the effectiveness (quality) of a model or data analysis method using one or more datasets.
- ☐ The process of evaluating the effectiveness (quality) of a model or data analysis method using only real-world scenarios.
- ☐ The process of exploring a learned model to discover new knowledge.

Suppose you train a classification model on a training set of 10,000 points and achieve 99% accuracy on that set. However, when you submit it to Kaggle, you get 67% accuracy. Which of the following approaches is likely to help you improve performance on Kaggle?

- □ Set the regularization coefficient (if any) to 0.
- □ Train on more data.
- □ Use a step to optimize parameters.
- □ Randomly drop data when training.

Which of the following statements is FALSE?

- □ Model evaluation and model selection in Machine Learning are two independent tasks.
- □ Model evaluation usually requires performing a model selection step.
- □ Model selection is a mandatory step when comparing different machine learning models (or methods).

Which of the following statements is the most appropriate about model selection?

- □ Model selection is concerned with finding the best settings for the (hyper) parameters within a model when training it from a dataset. Sometimes it also refers to choosing one among several available models.
- □ Model selection is only concerned with choosing the best model from a set available.
- □ The other statements are incorrect.

When using a method to analyze data, two different runs might produce different results even though the same settings are used for the parameters. What could be the reason?

- □ Due to the randomness in splitting the existing dataset into two subsets for training and validation.
- □ Due to using different settings for the parameters.
- □ Due to using different datasets.
- □ Due to incorrect use of the method.
- □ Due to the randomness of the learning/analysis algorithm.

When exploring data, you discover that attribute A is strongly correlated with the class label. However, when training a machine learning model with dataset A, it significantly reduces accuracy. Why might this situation occur?

- □ A is a noisy attribute.
- □ A has a negative correlation with the class label.
- □ Your evaluation might not be thorough.
- □ A might be common and not distinctive.
- □ This situation cannot occur.

When exploring data, you find that attribute A has a very low correlation with the class label. However, when training a machine learning model with dataset A, it often increases accuracy. Why might this situation occur?

- □ A is a noisy attribute.
- □ A has a negative correlation with the class label.
- □ The way you measure correlation might not accurately describe the hidden dependency between A and the class label.
- □ A might provide additional knowledge for the model.
- □ A might be common across all class labels.
- □ This situation cannot occur.

The 3-layer architecture of ...... includes backend, artist, and scripting.

- □ Seaborn
- □ Pyplot
- □ Matlab
- □ Matplotlib

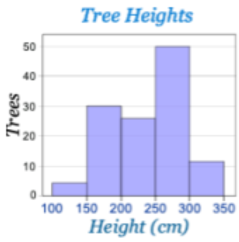
What would you do when you want to analyze and explore data?

- □ Statistically describe the data (min, max, avg, std...)
- □ Calculate the frequency of occurrence of data values.
- □ Draw a histogram of the data.
- □ All the other answers are correct.

Indicate the correct statement about scatter plots.

- □ A collection of points plotted in both the vertical and horizontal dimensions.
- □ A collection of points plotted randomly in the coordinate system.
- □ A collection of points clustered around a straight line.
- □ None of the statements are correct.

Indicate the correct statement about the following figure:



- □ A bar chart of tree height data.
- □ A histogram of tree height data.
- □ A graph showing data on the number of trees.
- □ A graph showing data on tree heights.

When analyzing the histogram of data, what information about the data are we looking for?

- □ Correlation
- □ Asymmetry
- □ Statistical information
- □ Outliers

Which of the following chart types best visualizes hierarchical data?

- □ Treemap
- □ Population pyramid
- □ Bar chart
- □ The other choices are incorrect.

Which chart type is suitable when we want to track changes over time?

- □ Line graph
- □ Column Graph
- □ Bar Graph
- □ All the other choices

Which visualization tool would be used to represent the complexity of software source code?

- □ Scientific visualization
- □ Mathematical visualization
- □ Information visualization

Which type of chart is the least ambiguous and often the best choice to start exploring data?

- □ Table chart
- □ Pie Chart
- □ Radial column chart
- □ Bar chart

A version of scatter plot that allows displaying 3-dimensional data?

- □ A heatmap
- □ A scatter map
- □ A bubble plot
- □ The other choices are incorrect.

What is an object that explains the color symbols and pattern shapes used in charts called?

- □ Legend
- □ Chart title
- □ Axis title
- □ Data label

What type of data is temperature among the following?

- □ Discrete, unordered data
- □ Continuous, ordered data
- □ Discrete, ordered data
- □ Continuous, unordered data

What features of data can be visualized in scatter plots?

- □ Correlation
- □ Associations

- □ Skewness
- □ Dispersion

What is the most accurate statement about pie charts?

- □ Pie charts are used when we want to show the composition of different parts within the data.
- □ Pie charts are a circular graph divided into different segments, each segment representing a change over time.
- □ Pie charts are used when comparing data categories.
- □ The other statements are incorrect.

What information can we derive from observing a box plot?

- □ Lower/upper quartile
- □ Gap
- □ Probability distribution
- □ Skewness

Which library should be used if one wants to visualize data with Python?

- □ Numpy
- □ Pandas
- □ Seaborn
- □ Pyplot, pandas, seaborn

Which Python library is commonly used for data visualization?

- □ NLTK, Spacy...
- □ Requests, Scrapy, BeautifulSoup...
- □ Tensorflow, Keras, scikit-learn...
- □ SciPy, NumPy, Matplotlib, and Pandas...

Among the following statements, which is the most accurate regarding choosing the appropriate visualization technique for a type of data?

- □ Gathering data, Organizing data, and Analyzing data
- □ Using bar charts is suitable for all types of data.
- □ Generating questions from a data visualization technique
- □ All the other statements are correct.

What is the correct combination of function and parameter to create a box plot in Matplotlib?

- □ Function = plot and Parameter = type with value = "box"
- □ Function = boxplot and Parameter = type with value = "plot"
- □ Function = plot and Parameter = kind with value = "box"
- □ Function = plot and Parameter = kind with value = "boxplot"

What type of graph does the following code represent?
question.plot(kind='barh')

- □ Line graph
- □ Column Graph
- □ Bar Graph
- □ The other choices are incorrect.