Some Basic Data Science Questions

Data Literacy

Data literacy is the ability to read, understand, create, and communicate data as information.

What is Data Science

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data

What is big data

Big data typically refers to extremely large data sets that require specialized and often innovative technologies and techniques in order to efficiently “use” the data.

Both Data Science and Big Data work with large data sets and analytics.

What is data mining

Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes.

Using a broad range of techniques, you can use this information to increase revenues, cut costs, improve customer relationships, reduce risks and more.

What is Machine learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves.

What is Artificial Intelligence (AI)

AI is about imparting autonomy to the data model. With Data Science, we build models that use statistical insights. On the other hand, AI is for building models that emulate cognition and human understanding.

Why is Data Science so important for the business?

Data Science enables enterprises to measure, track, and record performance metrics for facilitating enterprise-wide enhanced decision making. Companies can analyze trends to make critical decisions to engage customers better, enhance company performance, and increase profitability.

Which skills are needed to be a good data science practitioner?

- Statistics
- At least one programming language – R/ Python
- Data Extraction, Transformation, and Loading
- Data Cleaning and Data Exploration
- Machine Learning Algorithms

- Advanced Machine Learning (Deep Learning)
- Big Data Processing Frameworks
- Data Visualization

What is an intelligent agent

An intelligent agent is a **program** that can **make decisions** or **perform a service** based on its environment, **user input and experiences**. These programs can be used to **autonomously gather information** on a regular, programmed schedule or when prompted by the user in real time.

Example: AI assistant like **Alexa, Siri** and **Google Voice**

What is data literacy? What a data science developer needs to know about the data she or he works with?

Data literacy is the ability to **read**, **understand**, **create**, and **communicate data as information**. Much like literacy as a general concept, data literacy focuses on the **competencies involved** in working with data.

Levels

Level	Definition	Example
Conversational	Basic understanding of the concepts of data, analytics and use cases; one who "gets it" but cannot explain it to others	A professional who has a basic understanding of an analytics value proposition and the ingredients involved
Literacy	Ability to speak, write and engage in data and analytics programs and use cases	A professional who can explain all aspects of an analytics use case, including the industry problem, business process moment/decision affected, data sources leveraged and analytical methods applied
Competency	Competent of designing, developing and applying data and analytics programs	Experienced data and analytics program managers who have designed and delivered analytical projects from concept through outcome
Fluency	Fluent in all three elements of information language across most business domains within an industry vertical	A smart meter registers kW demand. Over time it creates kWh averages and peak demand. That is interpreted by billing far differently than generation or distribution planning. Fluent speakers can explain all of these use cases
Multilingual	Fluency across all three elements of the information language across multiple business domains, industries and ecosystems	An experienced data analytics strategy consultant who has designed and delivered analytical solutions across multiple industries and business domains, and can explain them to non-native speakers

Which data types Data Science works with

There are 4 main data types

Nominal: Datasets whose values don't possess a natural ordering.

Ordinal: Datasets whose values do possess natural ordering (like our dataset)

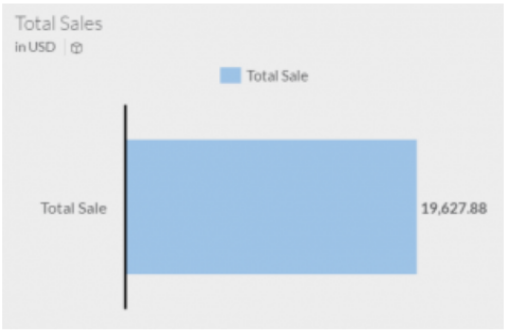
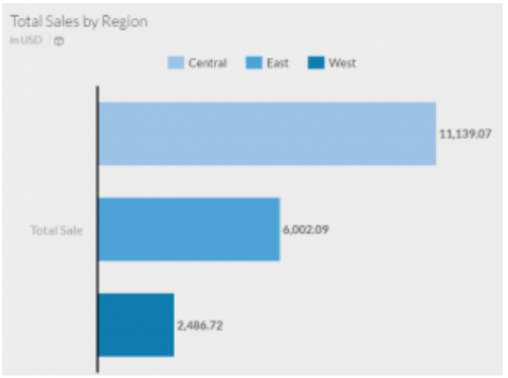
Discrete: Count that can't be made more precise. Typically it involves **integers**. For instance, the number of children (or adults, or pets) in your family is discrete data, because you are **counting whole**, indivisible entities: **you can't have 2.5 kids, or 1.3 pets**.

Continuous: Data that can take **any value**. **Height, weight, temperature and length** are all examples of continuous data. **Some continuous data will change over time - LIKE OUR DATA**

What is a dimension and what is a measure in data science?

Measures are numerical values that mathematical functions work on.
For example, a sales revenue column is a measure because you can find out a total or average the data.

Dimensions are qualitative and do not total a sum
For example, sales region, employee, location, or date are dimensions.

Chart 1 **Measure Only**	Chart 2 **Measure with Dimension**
	

In Data Science applications the variables play different roles. Which?

Independent Variables

- Explanatory variables.
- Intervention variables, or predicting variables.
- Input.

Dependent Variables

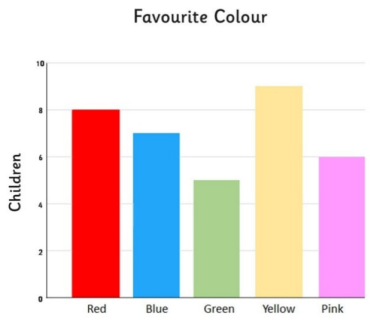
- Outcome.
- Response to other variables.

Other Contributing Variables

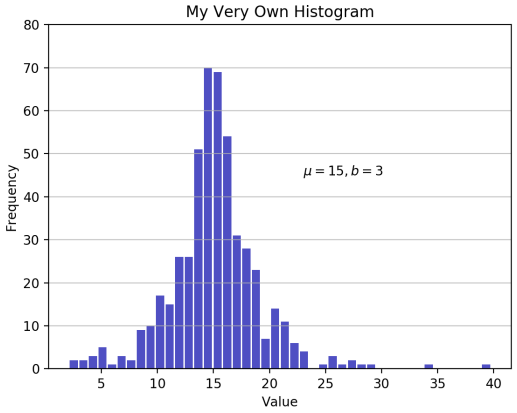
- Other variables that are important in explaining association between dependent and independent.

Discrete and continuous data are represented differently. How?

Discrete Data: Typically represented in a **Bar Chart**.

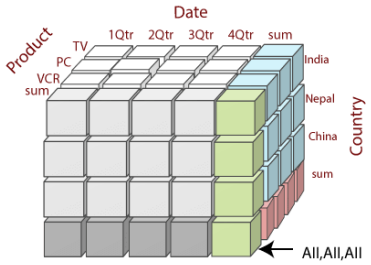


Continuous Data: Typically represented in tally charts, bar charts pie charts or **histogram.**

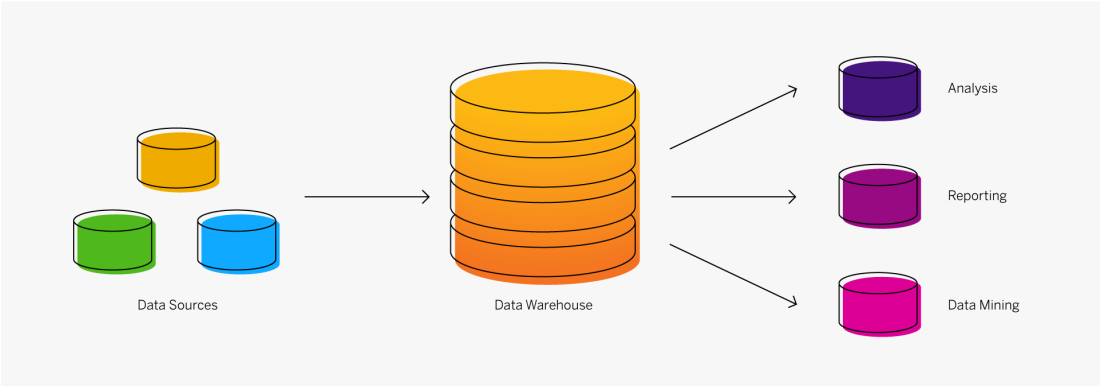


Data cube, data warehouse, data mart and data lake meaning

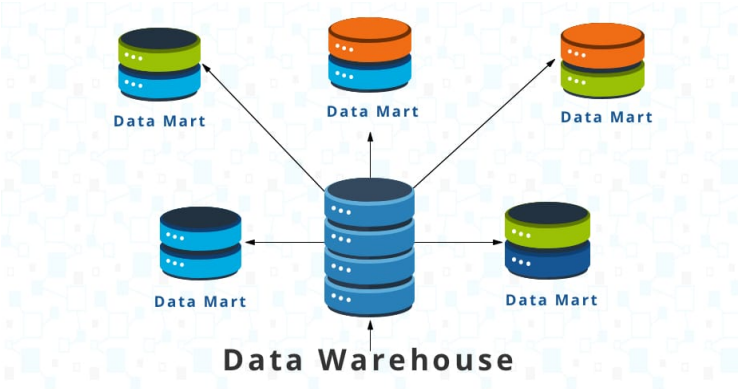
Data Cube: A data cube refers to a **three-dimensional (3D)** (or higher) range of values that are generally used to explain the time sequence of an image's data. It is a data abstraction to evaluate aggregated data from a variety of viewpoints.



Data Warehouse: A Data Warehouse is defined as a central repository where information is coming from one or more data sources. Three main types of Data warehouses are Enterprise Data Warehouse (EDW), Operational Data Store, and Data Mart.



Data Mart: is a subset of a data warehouse focused on a particular line of business, department, or subject area. Data marts make specific data available to a defined group of users, which allows those users to quickly access critical insights without wasting time searching through an entire data warehouse.



Data Lake: A data lake is a storage repository that holds a **vast amount of raw data** in its native format until it is needed. While a hierarchical data warehouse stores data in files or folders, a **data lake uses a flat architecture to store data**.

How do these terms differ?

OLAP – OLTP

OLAP: Online Analytical Processing, a category of software tools which provide analysis of data for business decisions. OLAP systems allow users to analyze database information from multiple database systems at one time.

OLTP: Online transaction processing supports **transaction-oriented applications** in a 3-tier architecture. OLTP administers day to day transactions of an organization.

ETL – ELT

ETL: Extract, Transform and Load - loads data first into the staging server and then into the target system.
Used for

- High amounts of data

ELT: Extract, Load, Transform - loads data directly into the target system.
Used for

- Compute-intensive Transformations
- Small amount of data

Descriptive analytics – predictive analysis

Descriptive Analytics: uses Data Aggregation and Data Mining techniques to give you knowledge about past

Predictive Analytics: uses Statistical analysis and Forecast techniques to know the future.

- This is what we do!

... In a Predictive model, it identifies patterns found in past and transactional data to find risks and future outcomes

Database – snowflake

Database: A database
Snowflake: A data warehouse - Snowflake's cloud data platform is tailored to integrate and support the applications that data scientists use regularly.

Tableau – Tableau Prep

These are two separate tools, for two separate tasks. **Tableau Prep** is for cleaning and organizing your data. **Tableau** Desktop is for the visual analysis of data. You would use **Tableau Prep** to bring a lot of data in from various sources, and clean them, and perform transformations on them.

Quantitative – qualitative

Quantitative data can be counted, measured, and expressed using numbers.
Contrary to qualitative data, quantitative data is statistical and is typically structured in nature – meaning it is more rigid and defined. This data type is measured using numbers and values, making it a more suitable candidate for data analysis.

Used to ask question: “How much” or “How many”.

Quantitative data can be generated through:

- Tests
- Experiments
- Surveys
- Market reports
- Metrics

This can be divided into sub categories such as **Discrete Data** and **Continuous Data**

Qualitative data is descriptive and conceptual.
Qualitative data is **non-statistical** and is typically **unstructured or semi-structured**. This data isn’t necessarily measured using hard numbers used to develop graphs and charts. Instead, it is categorized based on properties, attributes, labels, and other identifiers.

Used to ask question: “Why?”

Qualitative data can be generated through:

- Texts and documents
- Audio and video recordings
- Interview transcripts and focus groups
- Observations and notes

Qualitative data examples:

- Made of wood
- Built in Italy
- Deep brown
- Golden knobs
- Smooth finish
- Made of oak

Discrete - continuous

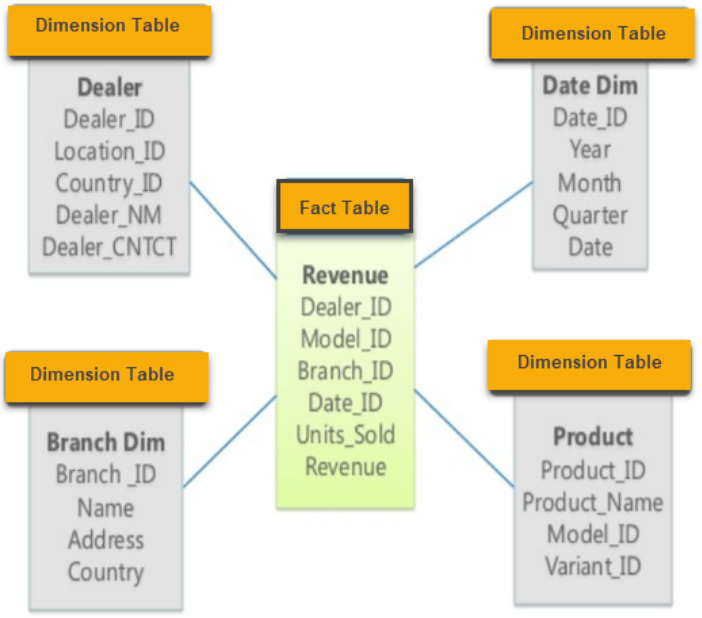
Discrete: Count that can't be made more precise. Typically it involves **integers**. For instance, the number of children (or adults, or pets) in your family is discrete data, because you are **counting whole**, indivisible entities: **you can't have 2.5 kids, or 1.3 pets**.

Continuous: Data that can take **any value**. **Height, weight, temperature and length** are all examples of continuous data. **Some continuous data will change over time - LIKE OUR DATA**

Which types of data stores you are familiar with, what is their implementation?

Explain `star` data structure, `fact tables` and `dimension tables`. Why is this structure good for data analytics? Give an example.

In computing, the **star schema** is the simplest style of **data** mart **schema** and is the approach most widely used to develop **data** warehouses and dimensional **data** marts. The **star schema** consists of one or more fact tables referencing any number of dimension tables



Benefits: Query performance gains – **star schemas** can provide performance enhancements for read-only reporting applications when compared to highly normalized **schemas**. Fast aggregations – the simpler queries against a **star schema** can result in improved performance for aggregation operations.

What is the difference between `joins` and `blends` of data

Blends

Data Blending allows a combination of **data** from **different data** sources to be linked.

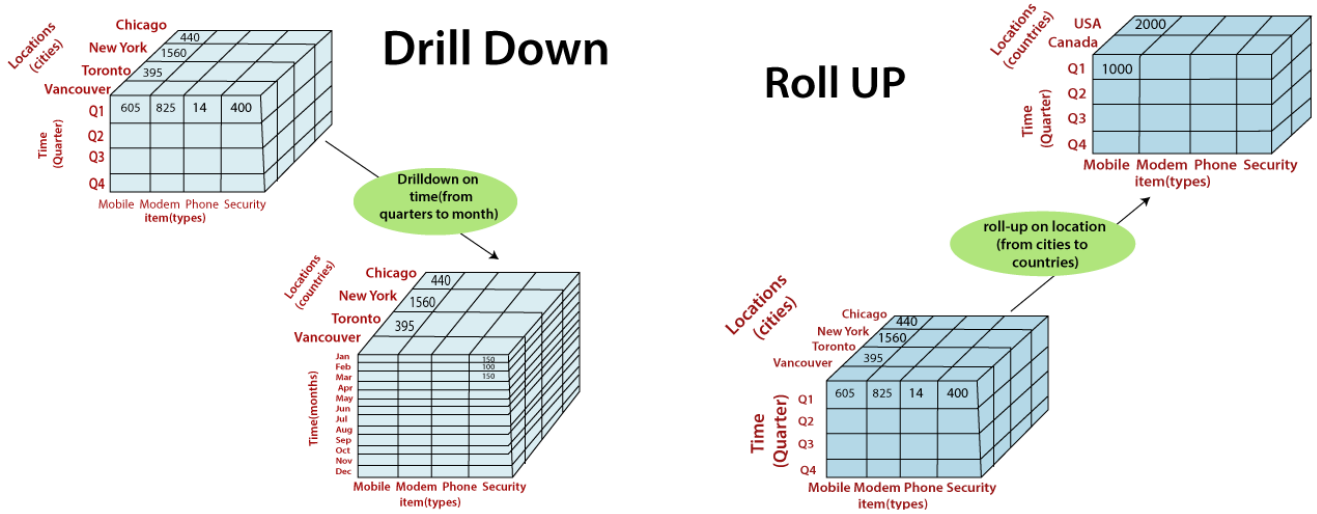
Joins

Data Joining works only with **data** from one and the same source.

For example: If the **data** is from an Excel sheet and a SQL database, then **Data Blending** is the only option to combine the two types of **data**.

What are `drill-down` and `roll-up` operations?

The **drill-down operation** (also called roll-**down**) is the reverse **operation** of roll-up. **Drill-down** is like zooming-in on the data cube. It navigates from less detailed records to more detailed data. **Drill-down** can be performed by either stepping **down** a concept hierarchy for a dimension or adding additional dimensions.



Linear Algebra

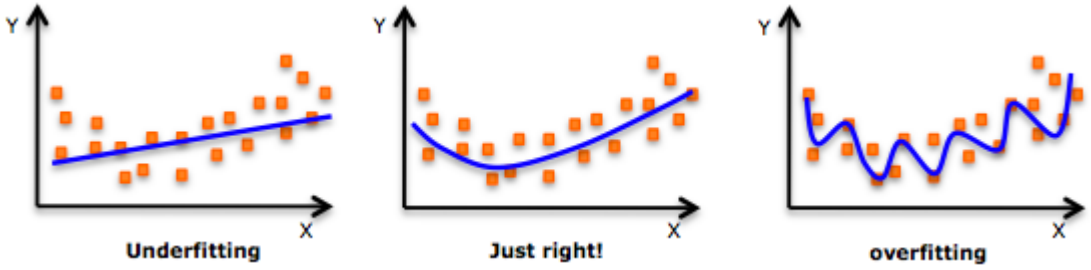
What is linear algebra? Which are the main objects of study by it? How is linear algebra related to data science?

Linear algebra is a sub-field of mathematics concerned with **vectors**, **matrices**, and **linear transforms**. It is a key foundation to the field of machine learning, from notations used to describe the operation of algorithms to the implementation of algorithms in code.

We use it in our **linear-regression** and **polynomial-regression** model.

You must be quite familiar with how a model, say a **Linear Regression model**, fits a given data:

- You start with some arbitrary prediction function (a linear function for a Linear Regression Model)
- Use it on the independent features of the data to predict the output
- Calculate how far-off the predicted output is from the actual output
- Use these calculated values to optimize your prediction function using some strategy like Gradient Descent



Linear Algebra is also used in image recognition and deep learning.

What is a vector? In which different ways can we represent a vector? Give examples. Which operations with vectors can you program?

What is a vector

A vector is a tuple of one or more values called scalars.
A tuple is a finite ordered list of values.

```
Vector created from a list:
[[ 2]
 [ 4]
 [ 6]
 [10]]
```

How can we represent a vector

Vectors are usually **represented** by arrows with their length **representing** the magnitude and their direction **represented** by the direction the arrow points. **Vectors** require both a magnitude and a direction. The magnitude of a **vector** is a number for comparing one **vector** to another

What is a dot product of two vectors? How can it be used for determination of correlation

What is a dot product

The dot product between two vectors is based on the projection of one vector onto another. Let's imagine we have two vectors a and b, and we want to calculate how much of a is pointing in the same direction as the vector b

How can it be used for determination of correlation (r2 score)

The coefficient of determination is a measurement used to explain how much variability of one factor can be caused by its relationship to another related factor. This correlation, known as the "goodness of fit," is represented as a value between 0.0 and 1.0.

Coefficient of determination (r2):

1. Coefficient of determination (r2) = Coefficient of Correlation (r) x Coefficient of Correlation (r)
2. It provides percentage variation in y which is explained by all the x variables together
3. Its value is (usually) between 0 and 1 and it indicates strength of Linear Regression model
4. Higher the R2 value, data points are less scattered so it is a good model. Lesser the R2 value is more scattered the data points.
5.

It can also be calculated as below:
R2 = 1 — (RSS/TSS)

Where
RSS = Residual Sum of Square
TSS = Total Sum of Square (It's square of (actual value — average value))

What is a cosine similarity?

Cosine similarity measures the similarity between two vectors of an inner product space.
It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction.
It is often used to measure document similarity in text analysis.

What is a matrix? Which operations with matrices you can program?

Matrix is a way of writing similar things together to handle and manipulate them as per our requirements easily. In Data Science, it is generally used to store information like weights in an Artificial Neural Network while training various algorithms.

Person ID	HW1 Grade	HW2 Grade
5	85	95
6	80	60
100	100	100

Matrix:

$$A \in \mathbb{R}^{3\times 2} = \begin{bmatrix} 85 & 95 \\ 80 & 60 \\ 100 & 100 \end{bmatrix}$$

Explain the following

Transposition

In linear algebra, the transpose of a matrix is an operator which flips a matrix over its diagonal; that is, it switches the row and column indices of the matrix.

The **rank** of a matrix is equal to the **rank** of its **transpose**. In other words, the dimension of the column space equals the dimension of the row space, and both equal the **rank** of the matrix.

For example:

$$A = \begin{bmatrix} 1 & 2 & -1 & 4 & 0 \\ 5 & 12 & 1 & -1 & -3 \\ 7 & 5 & 3 & 2 & 0 \end{bmatrix}$$

Then A transpose, A^T, equals the following 5×3 matrix:

$$A^T = \begin{bmatrix} 1 & 5 & 7 \\ 2 & 12 & 5 \\ -1 & 1 & 3 \\ 4 & -1 & 2 \\ 0 & -3 & 0 \end{bmatrix}$$

Transformation

A **linear transformation** is a function from one vector space to another that respects the underlying (**linear**) structure of each vector space. A **linear transformation** is also known as a linear operator or map. The two vector spaces must have the same underlying field.

Translation

In Euclidean geometry, a translation is a geometric transformation that moves every point of a figure or a space by the same distance in a given direction. A translation can also be interpreted as the addition of a constant vector to every point, or as shifting the origin of the coordinate system.

Think transform-translate in photoshop i guess

Dot product

The dot product between two vectors is based on the projection of one vector onto another. Let's imagine we have two vectors a and b, and we want to calculate how much of a is pointing in the same direction as the vector b

Determinant

In mathematics, the determinant is a scalar value that is a function of the entries of a square matrix. It allows characterizing some properties of the matrix and the linear map represented by the matrix.

Identity matrix

In linear algebra, the identity matrix of size n is the n × n square matrix with ones on the main diagonal and zeros elsewhere. It is denoted by I_n, or simply by I if the size is immaterial or can be trivially determined by the context.

An identity matrix is a square matrix having 1s on the main diagonal, and 0s everywhere else.

For example, the 2 x 2 and 3 x 3 identity matrixes are shown below:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

These are called identity matrices because, when you multiply them with a compatible matrix, you get back the same matrix.

Example:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 4 & 7 \\ -1 & 3 \end{bmatrix} = \begin{bmatrix} 1(4)+0(-1) & 1(7)+0(3) \\ 0(4)+1(-1) & 0(7)+1(3) \end{bmatrix}$$

$$= \begin{bmatrix} 4 & 7 \\ -1 & 3 \end{bmatrix}$$

Correlation matrix

A correlation matrix is simply a table which displays the correlation. The measure is best used in variables that demonstrate a **linear relationship** between each other. The fit of the data can be **visually represented in a scatterplot**. ... A correlation matrix consists of rows and columns that show the variables.

What are the inversion, convolution, pulling, and padding? In which context do they work together?

Exploratory Data Analysis

What is data exploration? What is the purpose of it?

Data exploration is the initial step in **data analysis**, where users explore a **large data** set in an **unstructured way to uncover initial patterns**, **characteristics**, and **points of interest**. ... Data exploration can use a combination of manual methods and automated tools such as data visualizations, charts, and initial reports.

How do we measure the data? Which measurement scales are you familiar with?

In statistics, what are population and sample? How do they differ? How are they used in the statistical calculations

A **population** is the **entire group** that you want to draw conclusions about. A **sample** is the **specific group** that you will collect data from. The size of the **sample is always less than the total size of the population**. In research, a population doesn't always refer to people.

Let’s say we have a dataset with a population of different age groups.
Then a sample of that would be to take only the age group between.. Lets say.. Age 10-25.

We can then use that sample to make specific statical calculations on that age group.

Explain data distribution, statistical methods of central tendency and variability.

Statistical Methods of Central Tendency

- **Measures of central tendency** tell us what is **common or typical** about our variable.
- Three measures of central tendency are the **mode**, the **median** and the **mean**.
- The mode is used almost exclusively with nominal-level data, as it is the only measure of central tendency available for such variables.

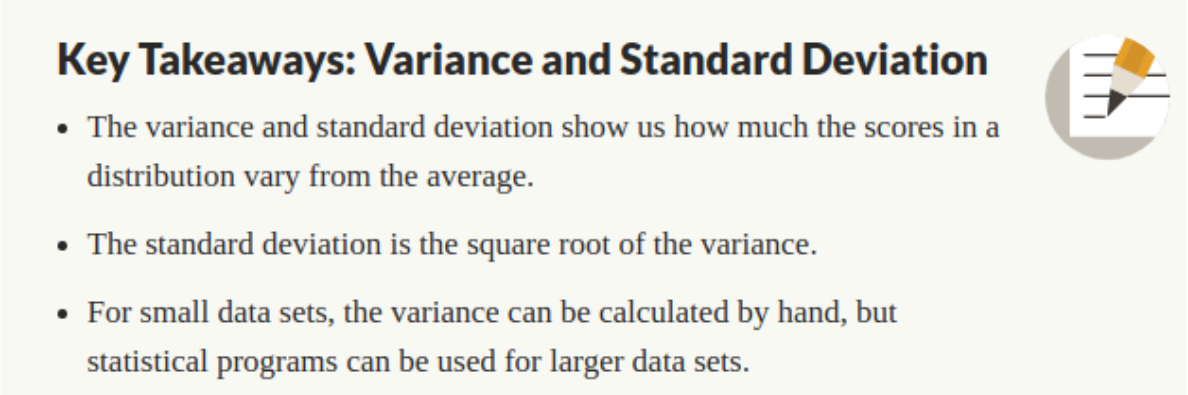
25.Which statistics are related to normal distribution? How are they measured? Give examples.

https://www.youtube.com/watch?v=PwtvDx2_5OY&ab_channel=365DataScience
aliases: bell curve, gaussian distribution
Symmetrical
Mean, median, mode = all equal
Centered around mean

$$N \sim (\mu, \sigma^2)$$

26.What are the variance and the standard deviation? Explain their use, show example.

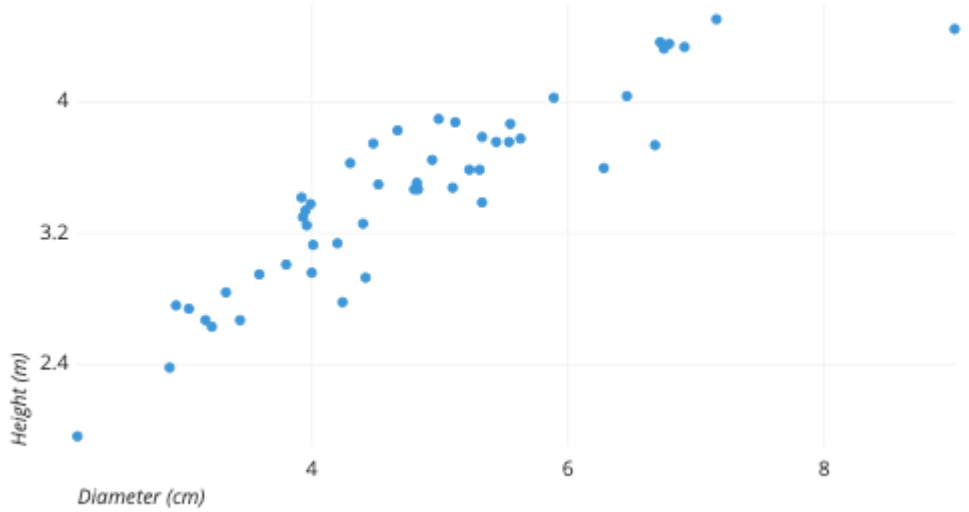
https://www.thoughtco.com/variance-and-standard-deviation-3026711



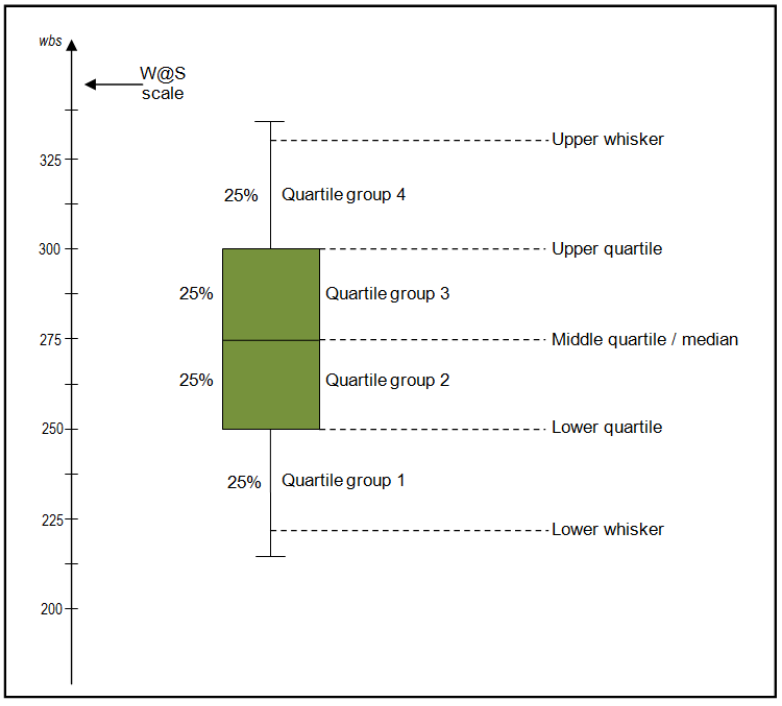
27.What are the scatter plot and the box plot? What is the use of them?

https://chartio.com/learn/charts/what-is-a-scatter-plot/

“Scatter plots are used to observe relationships between variables.”



“In descriptive statistics, a box plot or boxplot is a method for graphically depicting groups of numerical data through their quartiles”

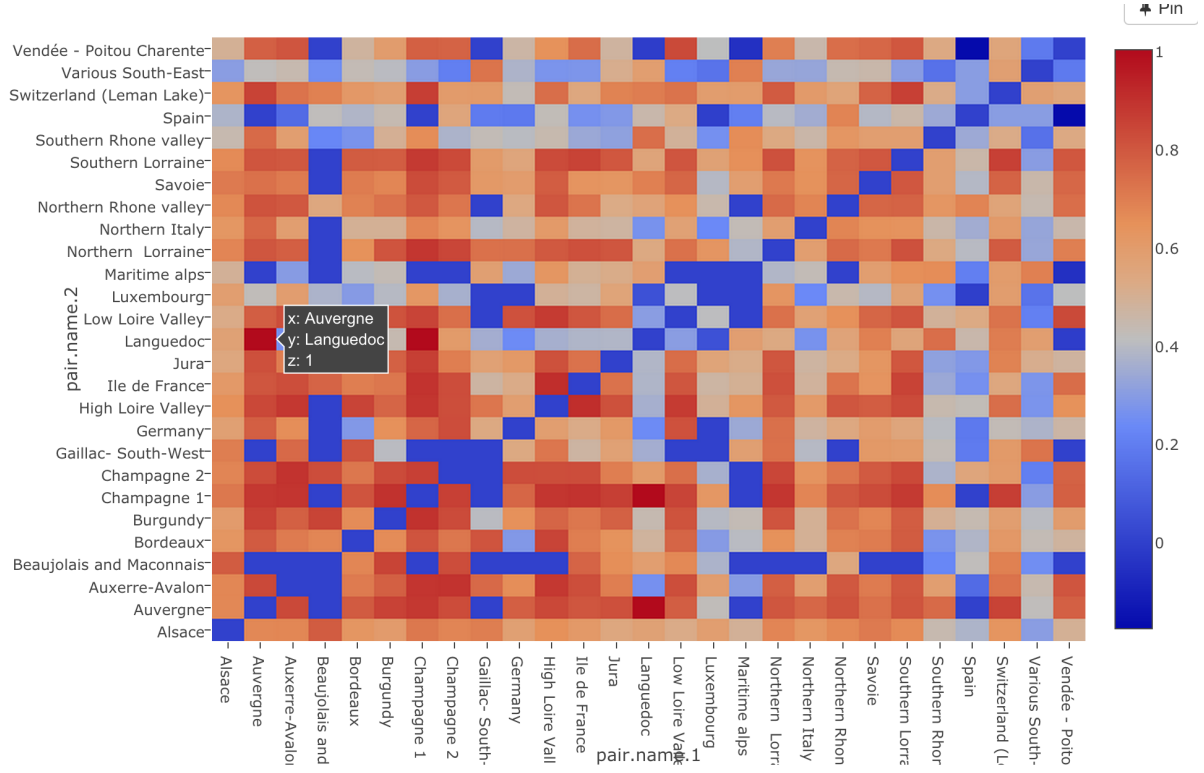


28.What is a heat map? What does it show? How can it be used in data science?

https://en.wikipedia.org/wiki/Heat_map
https://www.techopedia.com/definition/32150/heat-map

“A **heat map** (or **heatmap**) is a data visualization technique that shows magnitude of a phenomenon as color in two dimensions.”

“ Heat maps are used in many areas such as defense, marketing and understanding consumer behavior. Heat maps can be created with the help of software applications such as Microsoft Excel and others.”



29.How do we find the dependencies between the data samples? Which measures do we use? What is their implementation in data exploration?

Linear Regression

30.What are the hypotheses in statistics? Explain the role of null and alternative hypotheses.

https://vitalflux.com/data-science-how-to-formulate-hypothesis-for-hypothesis-testing/

"*a proposition or set of propositions, set forth as an explanation for the occurrence of some specified group of phenomena, either asserted merely as a provisional conjecture to guide investigation (working hypothesis) or accepted as highly probable in the light of established facts.*"

It's something you want to prove.

"In the case where the given statement is a well-established fact which is assumed to be true, one can call it as **Null Hypothesis** (in the simpler word, Nothing New)."

" In case the given statement is a claim and not yet proven, one can call/formulate it as an **Alternate Hypothesis** and accordingly define a Null Hypothesis. "

"One should note that Null and Alternate Hypothesis are mutually exclusive."

Null hypothesis	The housing price **does not** depend upon the average income of people staying in the locality.
Alternate hypothesis	The housing price depends upon the average income of people staying in the locality.

31.How are hypotheses tested? In which context are bootstrap and permutation used with testing hypotheses?

https://vitalflux.com/data-science-how-to-formulate-hypothesis-for-hypothesis-testing/

"say that null hypothesis is set as the statement that housing price does not depend upon average income of people staying in the locality, it would be required to be tested by taking samples of housing prices and, based on the test results, this Null hypothesis could either be **rejected** or **failed to be rejected**."

P-Value gets determined. It can be set to 15% for example. Results from the tests would reject the null hypotheses if the results were >+-15% beyond the expected outcome. Otherwise the hypotheses would fail to get rejected.

32.How is the significance of the result of hypothesis testing measured? Give some test significance measures.

https://vitalflux.com/data-science-how-to-formulate-hypothesis-for-hypothesis-testing/

https://towardsdatascience.com/statistical-significance-hypothesis-testing-the-normal-curve-and-p-values-93274fa32687

“Before starting the hypothesis testing, one would be required to set the significance level (also called as *alpha*) which represents the value for which a P-value less than or equal to *alpha* is considered statistically significant. Typical values of *alpha* are 0.1, 0.05, and 0.01. “

(I don’t get how this should be measured, and not just set.)

33.Explain the Type 1 and Type 2 errors, and the statistical power.

https://www.datasciencecentral.com/profiles/blogs/understanding-type-i-and-type-ii-errors

https://towardsdatascience.com/a-quick-refresher-of-statistical-power-fe8ae5e0c317

Type 1 Error (False Positive Error):

“A *type I error occurs when the null hypothesis is actually* ***true****, but was rejected as* ***false*** *by the testing*.“

Type 2 error(False Negative):

“A *type II error occurs when the null hypothesis is actually* ***false****, but was accepted as* ***true*** *by the testing.”*

Statistical Power:

No.

34.Explain the Central Limit Theorem and confidence intervals.

https://towardsdatascience.com/central-limit-theorem-clt-data-science-19c442332a32

35.What is the t-test? What is it used for? How?

https://www.analyticsvidhya.com/blog/2019/05/statistics-t-test-introduction-r-implementation/
A T-test is a type of inferential statistic used to study if there is a statistical difference between two groups
"*It helps us understand if the difference between two sample means is actually real or simply due to chance.*"

Types:

- One sample t-test
- Independent two-sample t-test
- Paired sample t-test

36.Which operations clean the data? Explain the difference between data cleaning and data wrangling. Which other types of data preparation are needed? Explain their meaning.

https://www.tableau.com/learn/articles/what-is-data-cleaning
https://www.inzata.com/data-wrangling-vs-data-cleaning-whats-the-difference/
"Data cleaning, also referred to as data cleansing, is the process of finding and correcting inaccurate data from a particular data set or data source. The primary goal is to identify and remove inconsistencies without deleting the necessary data to produce insights."

"Data wrangling, also referred to as data munging, is the process of converting and mapping data from one raw format into another. The purpose of this is to prepare the data in a way that makes it accessible for effective use further down the line. Not all data is created equal, therefore it's important to organize and transform your data in a way that can be easily accessed by others."

"While the methods might be similar in nature, data wrangling and data cleaning remain very different processes. Data cleaning focuses on removing inaccurate data from your data set whereas data wrangling focuses on transforming the data's format, typically by converting "raw" data into another format more suitable for use."

Pandas can do this for you

37.Name at least five types of diagrams for 2D data visualisation. How to choose? How to make diagramming better?

https://matplotlib.org/stable/gallery/index.html

Bar, line, scatter, heatmap, box,

Improve: Labels, histogram, proper dates, color

38.What is the difference between a view, a dashboard, and a story in data exploration context?

What is an actionable dashboard?

39.What are the advantages of dashboards and storytelling?

40.How to organize a data story? Provide some advice for designing good data stories, using

different types of scenarios.

41.What are the advantages of 3D visualization versus 2D visualization in data science?

Providing a z-axis allows for an extra dimension of data to be compared with the two other dimensions. Also makes proper simulation of physical space possible.

42.Do you think virtual reality technologies can help the understanding and cognitive activities related to data exploration in business or in education?

Currently simulations of surgery are being used for educational purposes in the medical industry. War related simulated scenarios are used to train soldiers. Cars are being designed in the car industry via. VR. 3D sculpting.

There can also be found examples of classes which you can take in VR. Many of them aren't in a professional environment.

43.What are the advantages of immersive analytics and immersive visualisation in data science?

immersive analytics

1. Greater Scale. ...
2. Increased **Immersion** = Increased Focus. ...
3. Discover More, In Less Time. ...
4. Human-Centered, Democratized Data. ...
5. Spark Engagement & Joy.

Predictive Data Analysis

Hvad vi gør i vores project

Predictive Data Analysis

What does AI stand for? Give your own explanation of the meaning of it. What is known as the Turing test?

What is AI

AI stands for Artificial Intelligence and it is intelligence demonstrated by machines, unlike the natural intelligence displayed by humans and animals, which involves consciousness and emotionality.

Turing Test

This test that Turing himself called "**the imitation game**" is a method for **judging the intelligence of machines** – and essentially, whether machines are capable of "**thinking**." To pass the test, a computer program must sufficiently **impersonate a human in a written conversation** with a human judge in real-time such that the human judge cannot reliably distinguish between the program and a real human.

What is an intelligent agent in the context of AI? Which are its components?

An intelligent agent is a **program** that can **make decisions** or **perform a service** based on its environment, **user input and experiences**. These programs can be used to

autonomously gather information on a regular, programmed schedule or when prompted by the user in real time.

Example: AI assistant like **Alexa, Siri** and **Google Voice.**

What is machine learning? Give your own explanation of the meaning of it. Compare it to deep learning.

Machine Learning is a set of algorithms which allows for an application to become more accurate at predicting outcomes without being explicitly programmed to do so.
Machine learning algorithms use historical data as input to predict new output values.

47.Which are the basic types of tasks solved by machine learning? How do they differ?

https://machinelearningmastery.com/basic-concepts-in-machine-learning/

(No dice. Too vague of a question?)

48.Describe the process of machine learning. Which activities would you plan to solve a task by implementing machine learning methods? Draw a simple sequence diagram.

https://businessanalyst.techcanvass.com/steps-of-machine-learning/

- Data gathering
- Data preparation
  - Cleansing, scrubbing, loading, fixing, removing duplicates, corrupted, null values, etc
- Data visualization
  - Matplotlib, Tableau, etc
- Model Choosing
  - Classification: KNeighbors classifier, Naïve Bayes, Support Vector Machine, Logistic Regression, and k-Nearest Neighbors, Decision trees, Random Forest
  - Regression: Linear, Logistic, Lasso, Ridge Regression
  - Ensembling: Random Forests, Boosting with XGBoost
  - Clustering: K-means, Affinity Propagation
  - Dimensionality reduction: Principal Component Analysis
    You should choose the model depending on your data type and the outcome you want to achieve.
- Model Training
  - 70% training data, 30% testing data
- Evaluation
  - accuracy, confusion matrix, F-measure, regression metrics, etc
- Parameter Tuning:
- Prediction
  - Model can be used to predict outcome on an unknown dataset

49.What is the difference between supervised and unsupervised machine learning? Give an

example from everyday life.

https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning

Read link for a more precise answer (It's good)

Supervised:

Labeled data. Human intervention. Used to train/"supervise" algorithms into classifying data or predicting outcomes accurately.

Solves two problems:

- Classification:
  - Accurately assign test data into categories. Separate apples from oranges.
  - Linear classifiers, support vector machines, decision trees, random forest
- Regression:
  - Understand the relationship between dependant and independent variables
  - Predict numerical values based on different data points (e.g. sales revenue)

Unsupervised:

Analysis and clustering of unlabeled data sets. No human intervention ("unsupervised")

Used for three tasks:

- Clustering:
  - Data mining.
  - Grouping unlabeled data based on similarities or differences. e.g. K-means assign similar data points into groups, K-value represents size of grouping and granularity
  - Used for: Market segmentation, image compression, etc.
- Association:
  - Find relationship between variables in a dataset
  - Used for: Market basket analysis, recommendation engine,
- Dimensionality reduction:
  - Learning technique for datasets with too many features or dimensions.
  - Reduction of data inputs to manageable size, while preserving data integrity
  - Preprocessing data stage. E.g. noise removal for improving image quality.

Differences:

- Labeled data - Unsupervised doesn't use it.
- Goals:
  - Supervised goal: Predict outcomes for new data. Expected outcome
  - Unsupervised goal: Gain insight from large volumes of new data
- Application:
  - Supervised: Spam detection, sentiment analysis, weather forecasting

    - Unsupervised: Anomaly detection, recommendation engines, customer personas
- Complexity:
    - Supervised: Simple. Use R or Python libraries
    - Unsupervised: Computationally complex. Large amounts of data required.
- Drawbacks:
    - Supervised: Time-consuming to train, in/out labels require expertise.
    - Unsupervised: Possibly wildly inaccurate results without human intervention.

50.Which data structures are used to hold the data needed for machine learning?

https://blog.statsbot.co/data-structures-related-to-machine-learning-algorithms-5edf77c8bbf4

(assuming data structures as in programming data structures)

Array, linked list, associative arrays, Binary tree, Balanced tree, heap, stack, queue, set, extensible arrays

51.In machine learning what is a feature and what is a label? Illustrate with appropriate

examples.

https://stackoverflow.com/questions/40898019/what-is-the-difference-between-a-feature-and-a-label#40899529

“Briefly, feature is input; label is output. This applies to both classification and regression problems.
A feature is one column of the data in your input set. For instance, if you're trying to predict the type of pet someone will choose, your input features might include age, home region, family income, etc. The label is the final choice, such as dog, fish, iguana, rock, etc.
Once you've trained your model, you will give it sets of new input containing those features; it will return the predicted "label" (pet type) for that person.”

52.How would you proceed, if you do not have sufficient data for building a reliable model?

seppuku

53.What is scatter plot and how can it help you in training models?

A scatter plot is a diagram type. You can use the scatter plot for visualizing your results. Presumably including training.

54.The model built of specific data can suffer from bias or variance. Explain that.

“Bias describes how well a model matches the training set. A model with high bias won’t match the data set closely, while a model with low bias will match the data set very closely. Bias comes from models that are overly simple and fail to capture the trends present in the data set.”

“Variance describes how much a model changes when you train it using different portions of your data set. A model with high variance will have the flexibility to match any data set that’s provided to it, potentially resulting in dramatically different models each time. Variance comes from models that are highly complex, employing a significant number of features.”

What is the difference between a method and a model in machine learning? And algorithm?

To summarize, an **algorithm** is a method or a procedure we follow to get something done or solve a problem. A **model** is a computation or a formula formed as a result of an **algorithm** that takes some **values as input** and **produces some value as output**.

Which machine learning libraries and frameworks you are familiar with? Which range of functions each of them provides

https://light-it.net/blog/top-10-python-libraries-for-machine-learning/
https://scikit-learn.org/stable/getting_started.html

- Tensorflow
- scikit-learn
    - (un)supervised learning. Model fitting, data preprocessing, model selection.
- Keras

Compare classification and regression. Give examples of appropriate cases for each.

Fundamentally, classification is about predicting a label and regression is about predicting a quantity.

The main difference between Regression and Classification algorithms that Regression algorithms are used to **predict the continuous** values such as price, salary, age, etc. and Classification algorithms are used to **predict/Classify the discrete values** such as Male or Female, True or False, Spam or Not Spam, etc

Regression - Continuous - Example: “What is the price of a house in California?”
Classification - Discrete - Example: “Is the email spam or not?”

58.In Regression, we search for a function that best represents the relation between the input and the output. If the input sequence is, for example, x = {1, 4, 3, 5, 2}, and the output sequence is y = {3, 9, 7, 11, 5}, which function would be appropriate?

59. One of the methods discussed in class is based on the theory of probabilities. What is the name of this method? Is it supervised? Explain the logic behind it. Which type of cases is this method good for?

60.If you program a robot that has to sort out big potatoes from small potatoes, which machine learning method would you use? Explain how it works.

61.What is clustering in machine learning? Name some methods for clustering. How do they differ? Give examples of appropriate implementations.

- **K-means Clustering**: The K-means clustering algorithm is used to find groups which have not been explicitly labeled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets

62.What are KNN and K-Means? What are they used for? What are the major differences in their implementations? What does K stay for in each of the abbreviations? What is the optimal value of K in the various implementations?

KNN = K-Nearest-Neighbour
K is the number of nearest neighbours.

KNN is used to solve classification and regression problems.

63.How would you test the validity of the model you have created? If the accuracy of a model is not good enough, what would you try to improve it?

We test the accuracy of our model by calculating the R2 score.
If the accuracy is not good enough, you can try adding more data to the model or find a more refined dataset.

64.GIGO is one of the most important metaphoric principles in machine learning. What is it associated with?

Garbage-in-garbage-out
GIGO is the idea that the output of an algorithm, or any computer function for that matter, is only as good as the quality of the input that it receives.

65.What is meant by cleaning the data? How many ways of cleaning are you familiar with? Give examples. Give recommendations for improving bad data.

Data Cleaning is when you clean your data so that it works well with your ML algorithms.

Techniques:

- Remove Irrelevant data
- Get rid of duplicate values
- Correct typos
- Convert datatypes (keep numeric numeric and string as string)
- Take care of null-values / missing values
- Input missing values
    - Imputation: replace missing data with substituted data

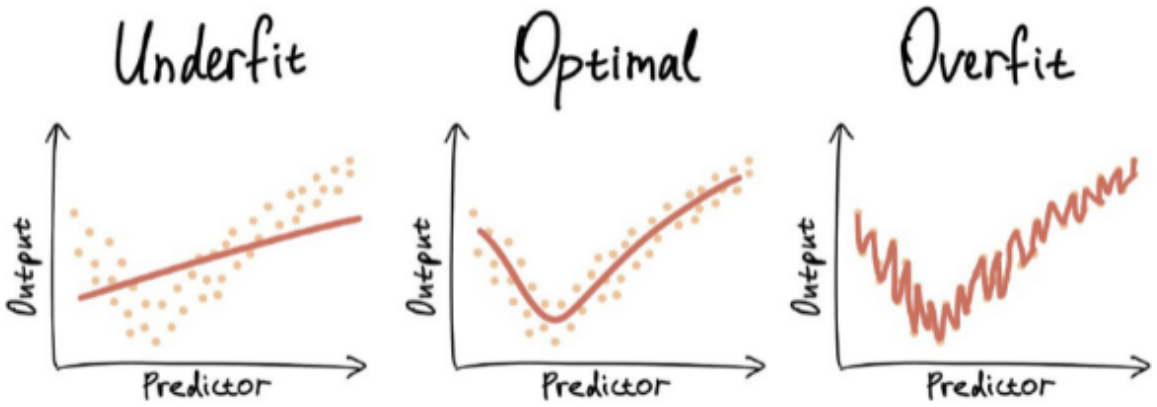
66.tIf the training data set contains missing values, can it still be used? Would you modify the data, and if yes, how?

Yes, it can still be used, however it depends on how much data is missing.
You can substitute missing data by adding an average between the values that are missing, or by combining different datasets together to make up for the missing data.

67.If there are too many features of objects available, how would you decide which are more valuable than others? Is creating a new feature an option?

68. How do we measure the quality of a model? Which statistics can be used?

https://towardsdatascience.com/fitness-navigator-7ca25de7757

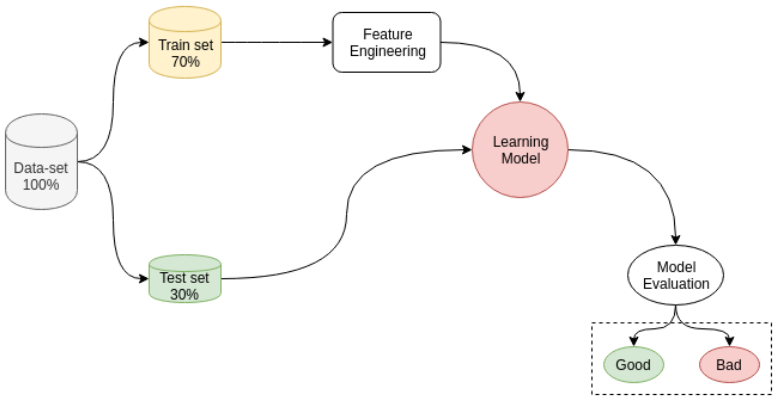


(1) Optimal fit. Image by author

(I'm literally too stupid and tired to answer that question)

69. What is called cross-validation and in which cases it is recommended for use?

We have done this by dividing our code into a training set and a test set.
Then calculate the R^2 score from these values.



Prescriptive Data Analysis

70.Neural Network is a metaphor from neurology used in machine learning. How does a

neuron relate to a model? Draw a sketch to illustrate your meaning.

71.What is the difference between a perceptron and a deep neural network? Do you get

better results by using deep neural networks, compared to other methods? Why or

why not?

72.Explain the back propagation of errors in an artificial neural network.

73. What is an activation function, how does it affect the work of a neuron? Which are the

most common functions used as an activation function?

74.What is a convolutional neural network? What kind of tasks are they good for solving? How does CNN differ from any other type of ANN?

75.Are you familiar with any technique used in processing text in natural language?

Which?

76.How does an image recognition system work? Explain some of the approaches used

for processing images by AI instruments.

77.What is TensorFlow? Where does it get its name from? Which are its application areas? What is Keras? How do TensorFlow and Keras contribute to the AI development process?

https://github.com/tensorflow/tensorflow

https://keras.io/about/

https://towardsdatascience.com/quick-ml-concepts-tensors-eb1330d7760f

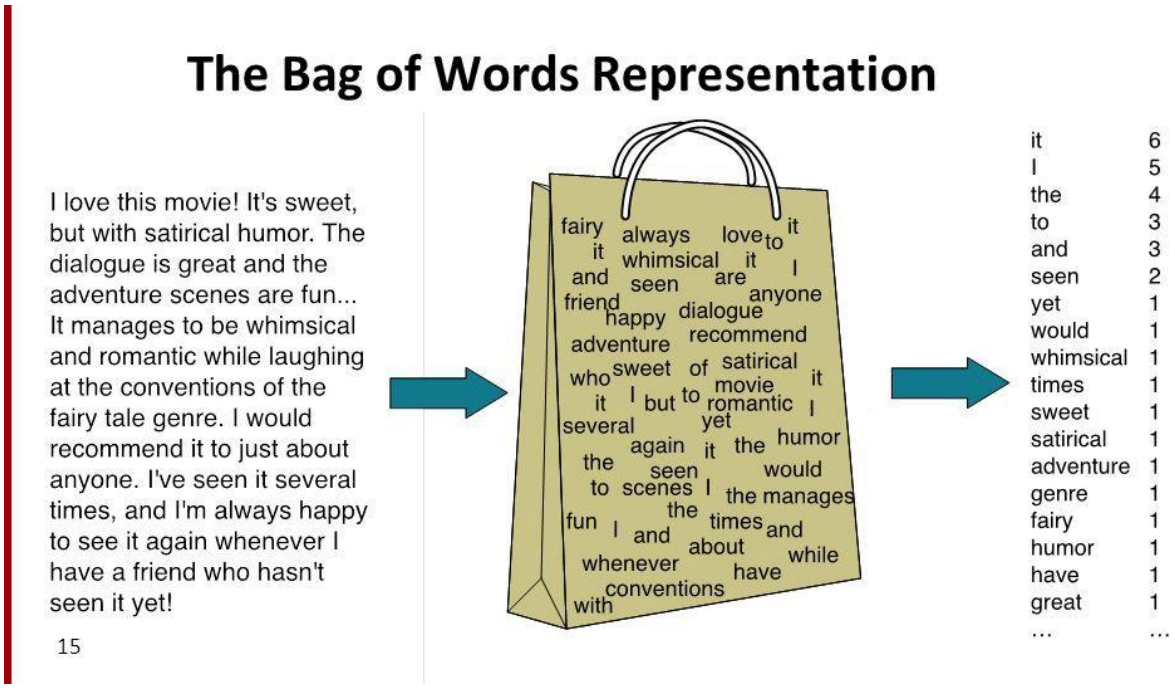
“Tensorflow is an end-to-end open source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries, and community resource that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML-powered applications”

“Keras is the high-level API of TensorFlow 2: an approachable, highly-productive interface for solving machine learning problems, with a focus on modern deep learning. It provides essential abstractions and building blocks for developing and shipping machine learning solutions with high iteration velocity.”

“Tensors are mathematical objects that generalize scalars, vectors and matrices to higher dimensions.”

78.What is called ‘one hot encoder’? How is it used in data science? Give an example.

What is called a ‘bag of words'? How is it used in data science? Give an example.



80.What does PCA stand for? Explain how it works? What is it valuable for?

Data Science Ethics

What are some of the ethical problems commonly discussed in Data Science

There are **3 main ethical challenges** related to data and data science:
Unfair Discrimination
Reinforcement of Human Biases
Lack of Transparency.

What is a protected attribute

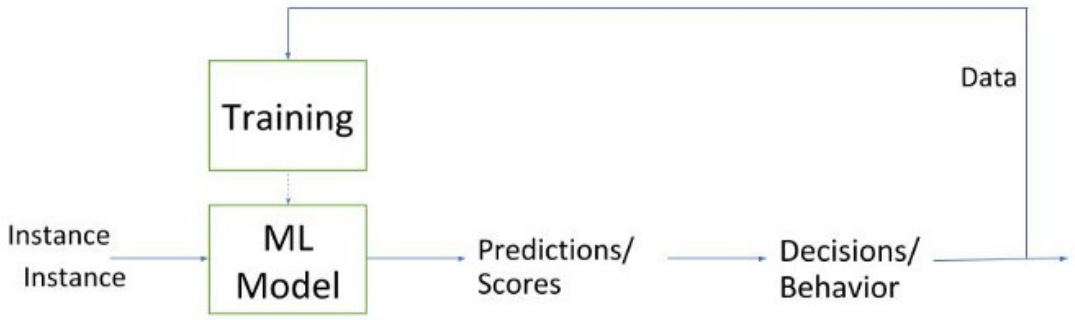
Protected attributes are those qualities, traits or characteristics that, by law, cannot be discriminated against.
These are race, religion, national origin, gender, marital status, age, and socioeconomic status.

What is a bias? How can it result in feedback and what is the problem with feedback

Bias in the algorithm is a measure of how well the algorithm “fits” the **data**. If the algorithm is overfitted then the result is lots of false negatives. If it is underfitted the result is false positives. Algorithm **bias** is more of a **data science** mathematical measure than an **ethical** issue.

Let’s take the example of twitter cropping out images.
We have a white male and a black male in an image.
The while male is at the top and the back male is at the bottom.
Due to poor feedback, the while male is almost always shown when the image is cropped, whereas the black male is not.
This is due to the machine learning algorithm being fed data which favored the while male.

The issue occurred due to an error in the feedback loop, where it received negative data.



Note

The questions are provided as a source of orientation only. You can use them for supervised learning!