Data Science – A Definition

Data Science is the science which uses computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze, visualize, interact with data to create data products.

Ben Fry's Model: Visualizing Data Process

1. Acquire
2. Parse (Analyze and put in proper format)
3. Filter
4. Mine (Discovering patterns or knowledge from large datasets)
5. Represent
6. Refine
7. Interact

Jeff Hammerbacher's Model

1. Identify problem
2. Instrument data sources
3. Collect data
4. Prepare data (integrate, transform, clean, filter, aggregate)
5. Build model
6. Evaluate model
7. Communicate results

Databases

Atomicity: When a database processes a transaction, it is either fully completed or not executed at all. If a single portion of the transaction fails, the whole transaction will fail.

Consistency: Is a property ensuring that only valid data following all rules and constraints is written in the database.

Isolation: Is a property that guarantees the individuality of each transaction, and prevents them from being affected from other transactions.

Durability: Durability is a property that enforces completed transactions, guaranteeing that once each one of them has been committed, it will remain in the system even in case of subsequent failures

Data science:

Consistency: All reads receive the most recent write or an error.

Availability: All reads contain data, but it might not be the most recent

Partition tolerance: The system continues to operate despite network failures in case of network failure/slow network: Provides only Availability or Partition tolerance

Area	BI Analyst	Data Scientist
Focus	Reports, KPIs, trends	Patterns, Correlations, models
Process	Static, comparative	Exploratory, experimentation, visuals
Data sources	Pre-planned, added slowly	On the fly, as needed
Transform	Upfront, carefully planned	In-database, on-demand, enrichment
Data quality	Single version of truth	"Good enough", probabilities
Data model	Schema on load	Schema on query
Analytics	Retrospective, Descriptive	Predictive, Prescriptive, Preventative

Applications of Data Scienece

Climate change and weather
- Traffic control
- Agriculture
- Personalized healthcare
- Twitter data analysis
- Facebook information links
- Pollution and Weather

Data Analyst Skills	Data Scientist Skills
Data Mining	Data Mining
Data Warehousing	Data Warehousing
Math, Statistics	Math, Statistics, Computer Science
Tableau and Data Visualization	Tableau and Data Visualization/ Storytelling
SQL	Python, R, Java, Scala, SQL, MATLAB, Pig
Business Intelligence	Economics
SAS	Big Data/Hadoop
Advanced Excel Skills	Machine Learning

Data Science Life Cycle (Draw a Circle)

- Data Collection
- Data Preparation
- Exploratory Data Analysis
- Modelling
- Model Evaluation
- Model Deployment

NumPy/Python

• NumPy is a Python library used for working with arrays.
• It also has functions for working in domain of linear algebra, Fourier transform, and matrices.
• NumPy stands for Numerical Python.
• In Python we have lists that serve the purpose of arrays, but they are slow to process.
• NumPy aims to provide an array object that is up to 50x faster than traditional Python lists.
• The array object in NumPy is called ND array, it provides a lot of supporting functions that make working with ND array very easy.
• NumPy arrays are stored at one continuous place in memory unlike lists, so processes can access and manipulate them very efficiently

Pandas

• Pandas officially stands for 'Python Data Analysis Library', THE most important Python tool used by Data Scientists today.
• Pandas is an open-source Python library that allows users to explore, manipulate and visualize data in an extremely efficient manner. It is literally Microsoft Excel in Python.
• It is easy to read and learn
• It is extremely fast and powerful
• It integrates well with other visualization libraries
• Pandas can take in a huge variety of data, the MOs common ones are csv, excel, SQL or even a webpage.

Amazing real-time Data Science Applications:

Recommendation- Most of the apps and websites like Amazon, YouTube, Flipkart, etc. give recommendation over as per the viewer's interest. Online music applications like Spotify give recommendations as per your taste in music. So these are good examples of data science recommendation applications.

Intelligent Assistant- Google assistant, Siri are examples of intelligent assistants. The advanced machine learning algorithm converts voice input into text output. These smart assistants recognize the voice and provide the required information in both voice and text outputs.

Autonomous driving vehicles- Automobile companies like Waymo and Tesla looking for the next generation of autonomous vehicles. 3D images were taken by the cameras and the information provided to the algorithms for further processing.

Piracy Detection- YouTube is an example of piracy detection using machine learning algorithms. Due to the big database, copied contents cannot be detected manually. So it helps to detect and remove the copied content to reduce human efforts.

Image Recognition- Facebook is the application that uses image recognition by data science and machine learning for the friend suggestion. Even Google lens uses an image recognition algorithm to provide the related information to you.

Data Preprocessing

Data Preprocessing is a technique that is used to convert the raw data into a clean data set. Data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

Tasks of Data Preprocessing

• **Data Cleaning:** This is the first step which is implemented in Data Preprocessing. In this step, the primary focus is on handling missing data, noisy data, detection, and removal of outliers, minimizing duplication and computed biases within the data.
• **Data Integration:** This process is used when data is gathered from various data sources and data are combined to form consistent data. This consistent data after performing data cleaning is used for analysis.
• **Data Transformation:** This step is used to convert the raw data into a specified format according to the need of the model.
• **Normalization –** In this method, numerical data is converted into the specified range, i.e., between 0 and 1 so that scaling of data can be performed.
• **Aggregation –** This method is used to combine the features into one. For example, combining two categories can be used to form a new group.

PTO

• **Generalization –** In this case, lower-level attributes are converted to a higher standard (e.g. age 20, 40 – may be taken as Young, Old, etac)
• **Data Reduction:** After the transformation and scaling of data duplication, i.e., redundancy within the data is removed and efficiently organize the data.

How we can deal with the missing data

Ignoring the missing record – It is the simplest and efficient method for handling the missing data. But, this method should not be performed at the time when the number of missing values are immense or when the pattern of data is related to the unrecognized primary root of the cause of statement problem.

Filling the missing values manually – This is one of the best-chosen methods. But there is one limitation that when there are large data set, and missing values are significant then, this approach is not efficient as it becomes a time-consuming task.

Filling using computed values – The missing values can also be occupied by computing mean, mode or median of the observed given values. Another method could be the predictive values that are computed by using any Machine Learning or Deep Learning algorithm.

How we can deal with the noisy data

Data Binning: In this approach sorting of data is performed concerning the values of the neighborhood. This method is also known as local smoothing.

Clustering: In the approach, the outliers may be detected by grouping the similar data in the same group, i.e., in the same cluster.

Machine Learning: A Machine Learning algorithm can be executed for smoothing of data. For example, Regression Algorithm can be used for smoothing of data using a specified linear function.

Removing manually: The noisy data can be deleted manually by the human being, but it is a time-consuming process, so mostly this method is not given priority.

What Is Data Wrangling?

Data wrangling is used in step during EDA and modeling to adjust data sets interactively while analyzing data and building a model.

Process of removing errors and combining complex data sets to make them more accessible and easier to analyze.

It is used to convert the raw data into the format that is convenient for the consumption of data

It executed at the time of making an interactive model.

Data is converted to the proper feasible format before applying any model to it. By performing filtering, grouping and selecting appropriate data accuracy and performance of the model could be increased

Data Leakage

• The Leakage of data from test dataset to training data set.
• Leakage of computed correct prediction to the training dataset.
• Leakage of future data into the past data.
• Usage of data outside the scope of the applied algorithm
• In general, the leakage of data is observed from two primary sources of Machine Learning/Deep Learning algorithms such as feature attributes (variables) and training data set.
• Checking the presence of Data Leakage within the applied model

Tasks of Data Wrangling

Discovering: Understand the data and choose the best approach.

Structuring: Organize data from different sources into a consistent format.

Cleaning: Remove data that can degrade analysis performance.

Enrichment: Extract new features to improve model performance.

Validating: Improve data quality and verify transformations.

Publishing: Document the data wrangling steps for future use.

Understanding Data Attribute Types

Qualitative Attributes

1. Nominal Attributes

• Values are name of things or some kind of symbols (without quantitative/numeric value)
• Values of nominal attributes represents some category or state and that's why nominal attribute also referred as categorical attributes and
• There is no order (rank, position) among values of nominal attribute

2. Binary Attributes: Binary data has only 2 values/states. For Example, yes or no, true or false i. Symmetric: Both values are equally important (Gender). No preference on which should be coded as 0 or 1

ii. Asymmetric: Both values are not equally important. Most important outcome is coded as 1

3. Ordinal Attributes:

• Values that have a meaningful sequence or ranking (order) between them
• But magnitude of values is not actually known

Quantitative Attributes

Numeric:

• A numeric attribute is quantitative because, it is a measurable quantity,
represented as integer or real values.
• Numerical attributes are of 2 types, interval and ratio.

I) Interval-scaled attributes

• It is concerned with both the order and difference between your variables. This allows you to measure standard deviation and central tendency.
• Values can be added and subtracted but cannot be multiplied or divided
• e.g. Temperature of 20oC is warmer than 10oC, and the difference between 20 deg and 10 deg is 10oC. The difference between 10 deg and 0 deg is also 10oC.
• Interval data always appears in the form of numbers or numerical values where the distance between the two points is standardized and equal
• Do not have a true zero even if one of the values carries the name "zero
• A true zero has no value, but 0 degrees C definitely has a value: it's quite chilly. You can also have negative numbers.
• If you don't have a true zero, you can't calculate ratios. This means addition and subtraction work, but division and multiplication don't.

II) Ratio-scaled attributes
• Has all properties of interval-scaled
• Have a true zero
▪ A good example of ratio data is weight in kilograms. If something weighs zero kilograms, it truly weighs nothing—compared to temperature (interval data), where a value of zero degrees doesn't mean there is "no temperature," it simply means it's extremely cold!
• Values can be added, subtracted, multiplied & divided
• e.g. Weight, height, etcQuantitative Attributes (Contd.)
• Interval variables, ratio variables can be discrete or continuous.
• A discrete variable is expressed only in countable numbers (e.g., integers)
• A continuous variable can potentially take on an infinite number of values.

Major Tasks in Data Preprocessing
• **Data Cleaning:** Fill in missing values, smooth noisy data, identify or remove outliers and noisy data, and resolve inconsistencies
• **Data Integration:** Integration of multiple databases, or files
• **Data Transformation:** Normalization and aggregation
• **Data Reduction:** Obtains reduced representation in volume but produces the same or similar analytical results
• **Data Discretization** (for numerical data)

Data Cleaning Tasks
• Importance Data cleaning is the first step
• Fill in missing values
• Identify outliers and smooth out noisy data
• Correct inconsistent data
• Resolve redundancy caused by data integration

Python Collections to Store the Data
• Python Collections are used to store data.
• Four types: lists, tuple, dictionaries, sets, and tuples, all of which are built-in
collections
• List is a collection which is ordered and changeable. Allows duplicate members.
e.g. [1, 2, 3, 4, 5]
• Tuple is a collection which is ordered and unchangeable. Allows duplicate members.
e.g. (1, 2, 3, 4, 5)
• Set is a collection which is unordered, unchangeable and unindexed. No duplicate members.
But, you can remove and/or add items whenever you like.
e.g. {1, 2, 3, 4, 5}
• Dictionary is a collection which is ordered (from Python version 3.7) and changeable. No
duplicate members.
e.g. {1: "a", 2: "b", 3: "c", 4: "d", 5: "e"}

How to Handle Noisy Data?
• **Binning:** First sort data and partition into (equal-frequency) bins Then one can smooth by bin mean, smooth by bin median, smooth by bin boundaries, etc.
• **Regression:** Smooth by fitting the data into regression functions
• **Clustering:** Detect and remove outliers
• **Combined:** computer and human inspection: Detect suspicious values and check manually

Regression for Data Smoothing
• The regression functions are used to determine the relationship between the dependent variable (target field) and one or more independent variables. The dependent variable is the one whose values you want to predict, whereas the independent variables are the variables that you base your prediction on.
• A Regression Model defines three types of regression models: linear, polynomial, and logistic regression.
• Linear and stepwise-polynomial regression are designed for numeric dependent variables having a continuous spectrum of values. These models should contain exactly one regression table.
• Logistic regression is designed for categorical dependent variables.

Clustering for Data Smoothing: Outlier Removal

• **Data points inconsistent with the majority of data**
• **Different outliers**
  Noisy: CEO's salary (-10)
  In consistent: One's age = 200
• **Removal methods**
  Clustering
  Curve-fitting
  Hypothesis-testing with a given model

Data Integration

• Combines data from multiple sources into a coherent store
• Careful integration can help reduce & avoid redundancies and inconsistencies
• This helps to improve accuracy & speed of subsequent data mining
• Heterogeneity & structure of data pose great challenges

Redundancy Analysis:

• Redundant data occur often when integration of multiple databases
• Object identification: The same attribute or object may have different names in different databases
• Derivable data: One attribute may be a "derived" attribute in another table, e.g., annual revenue

Correlation Analysis:

• Correlation can help in predicting one quantity from another
• Correlation can indicate the presence of a causal relationship (may not be true in all cases)
• Correlation is used as a basic quantity and foundation for many other modeling techniques
• More formally, correlation is a statistical measure that describes the association between random variables.
• There are several methods for calculating the correlation coefficient, each measuring different types of strength of association

Data Transformation

1) Data Transformation by Normalization

• The measurement unit used can affect the data analysis.
• For example, changing measurement units from meters to inches for height, or from kilograms to pounds for weight, may lead to very different results.
• In general, expressing an attribute in smaller units will lead to a larger range for that attribute, and thus tend to give such an attribute greater effect
• To help avoid dependence on the choice of measurement units, the data should be normalized or standardized.
• This involves transforming the data to fall within a smaller or common range such as [−1,1] or [0.0, 1.0].
• Normalizing the data attempts to give all attributes an equal weight.

4) Normalization by Decimal Scaling

• We move the decimal point of values of the attribute.
• This movement of decimal points totally depends on the maximum value among all values in the attribute
• Normalized Attribute V= Vi /10j

2) Data Transformation by min-max Normalization

• Min-max normalization performs a linear transformation on the original data.
• Let A be a numeric attribute with n observed values, v1,v2,...,vn.
• Suppose that minA and maxA are the minimum and maximum values of an attribute, A.
• Min-max normalization preserves the relationships among the original data values.
• It will encounter an "out-of-bounds" error if a future input case for normalization falls outside of the original data range for A.

3) Data Transformation by z-score Normalization

• In z-score normalization (or zero-mean normalization), the values for an attribute, A, are normalized based on the mean (i.e., average) and standard deviation of A.
• This method of normalization is useful when the actual minimum and maximum of attribute A are unknown, or when there are outliers that dominate the min-max normalization.

Data Reduction

• Data is too big to work with
• Data reduction
• Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results
• Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.

Data Reduction Strategies

• Data Cube Aggregation
• Dimensionality Reduction
• Data Compression
• Numerosity Reduction
• Discretization and Concept Hierarchy Generation

1) Data Cube Aggregation
• Multiple levels of aggregation in data cubes
• Further reduce the size of data to deal with
• Reference appropriate levels
• Use the smallest representation capable to solve the task
• Queries regarding aggregated information should be answered using data cube, when possible

2) Dimensionality Reduction
• A process of reducing the number of random variables or attributes under consideration.
• Data encoding or transformations are applied to obtain a reduced or compressed representation of the original data.
• Include wavelet transforms and principal components analysis (PCA) which transform or project the original data onto a smaller space.
• Attribute subset selection/Feature subset selection/feature creation: Irrelevant, weakly relevant, or redundant attributes or dimensions are detected and removed.

Heuristic (Greedy) methods for attribute subset selection

1. Stepwise Forward Selection: • Starts with an empty set of attributes as the reduced set • Best of the relevant attributes is determined and added to the reduced set • In each iteration, best of remaining attributes is added to the set

2. Stepwise Backward Elimination: • Here all the attributes are considered in the initial set of attributes • In each iteration, worst attribute remaining in the set is removed

3. Combination of Forward Selection and Backward Elimination: • Stepwise forward selection and backward elimination are combined • At each step, the procedure selects the best attribute and removes the worst from among the remaining attributes

4. Decision Tree Induction: • This approach uses decision tree for attribute selection. • It constructs a flow chart like structure having nodes denoting a test on an attribute. • Each branch corresponds to the outcome of test and leaf nodes is a class prediction. • The attribute that is not the part of tree is considered irrelevant and hence discarded

Numerosity reduction techniques replace the original data volume by
alternative, smaller forms of data.
• Dimensionality reduction and numerosity reduction techniques can be
considered forms of data compression.
• Parametric methods
• Assume the data fits some model, estimate model parameters, store only the parameters,
and discard the data (except possible outliers)
• Non-parametric methods
• Do not assume models
• Major families: histograms, clustering, sampling

Clustering
• Partition data set into clusters, and store cluster representation only
• Quality of clusters measured by their diameter (max distance between any two objects in the cluster) or centroid distance (avg. distance of each cluster object from its centroid)
• Can be very effective if the data is not smeared
• Can have hierarchical clustering (possibly stored in multi-dimensional index tree structures
• There are many choices of clustering definitions and clustering algorithms (further details later)

Sampling
• Simple random sampling
• There is an equal probability of selecting any particular item
• May have very poor performance in the presence of skew
• Sampling without replacement
• Once an object is selected, it is removed from the population
• Sampling with replacement
• A selected object is not removed from the population
• Stratified sampling:
• Partition the data set, and draw samples from each partition (proportionally, i.e.,
approximately the same percentage of the data)
• Used in conjunction with skewed data

3) Data Compression
String compression • There are extensive theories and well-tuned algorithms • Typically lossless • But only limited manipulation is possible
Audio/video, image compression • Typically lossy compression, with progressive refinement • Sometimes small fragments of signal can be reconstructed without reconstructing the whole