NPA Data Science

2: Applications of Data Science, The V's of Data and Types of Bias

Topic Outcomes

- Explain the applications of data science and the data science lifecycle
- Explain the V's of Data
- Explain Types of Bias

Applications of Data Science

Consider the following examples of the application in data science:

Personal: Translation apps and websites, image recognition, speech recognition devices, personalised medical treatment plans, self-driving cars, movie recommendations, website sort-order and recommendations.

Business: Fraud detection, financial risk estimation, customer retention, transport logistics, testing marketing approaches, airline route planning, real-time pricing optimisation, personalised advertising

Government: Tax analysis, preventing cyber attacks, detecting terrorist threats, improving national security, improving healthcare, coordinating responses to emergencies.

Data science and its applications have been steadily changing the way we do business and live our day-to-day lives — and considering that 90% of all of the world's data has been created in the past few years, there's a lot of growth ahead of this exciting field.

One example is Airbnb, which has always been a business informed by data. From understanding the demographics of renters to predicting availability and prices, Airbnb is a prime example of how the tech industry is leveraging data science. The company even has an entire section of its blog dedicated to the groundbreaking work its data team is doing.

Faced with a large amount of data from customers, hosts, locations, and demand for rentals, Airbnb used data science to create a dynamic pricing system called Aerosolve, which has since been released as an open-source resource.

Using machine learning techniques, Aerosolve predicts the optimal price for a rental based on its location, time of year, and a variety of other attributes. For Airbnb hosts, it revolutionised the way in which rental owners can best set their prices in the market and maximise returns.

The data science life cycle and the significance of domain expertise

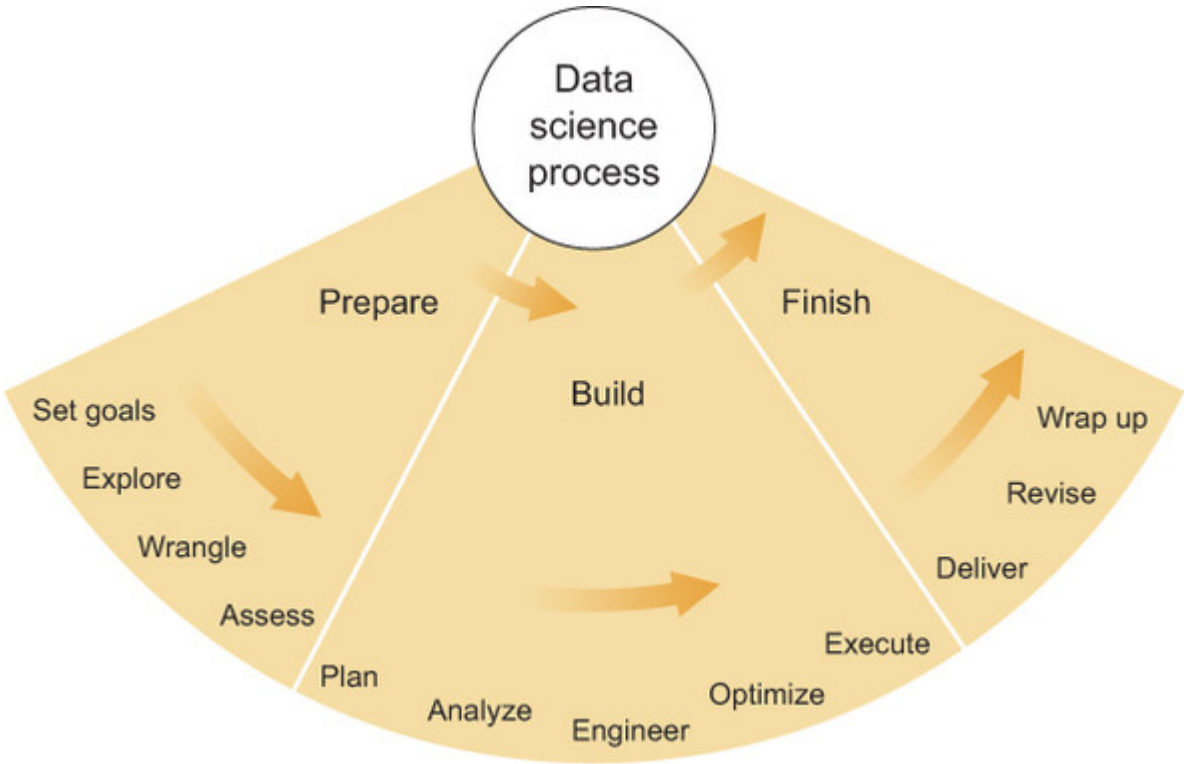
Data science projects do not have a nice clean lifecycle with well-defined steps like software development life cycle (SDLC). People often confuse the lifecycle of a data science project with that of a software engineering project. That should not be the case, as data science is more of science and less of engineering. There is no one-size-fits-all workflow process for all data science projects and data scientists have to determine which workflow best fits the business requirements.

There are different ways to conceptualise the data science life cycle. One way shown in the figure below is to think of a Data Science project as having three steps; preparation, building and finishing, which each have their own components.

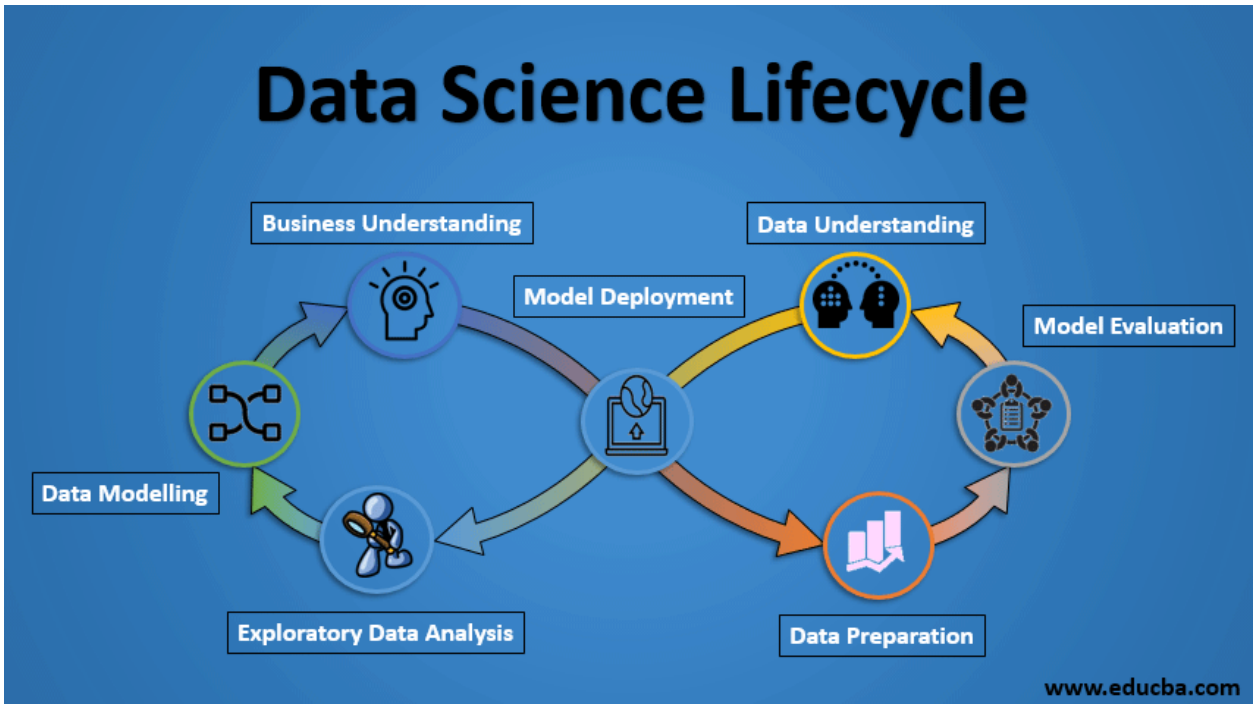
The first phase is preparation—time and effort spent gathering information at the beginning of a project can spare big headaches later.

The second phase is building the product, from planning through execution, using what you learned during the preparation phase and all the tools that statistics and software can provide.

The third and final phase is finishing - delivering the product, getting feedback, making revisions, supporting the product, and wrapping up the project.



The data science lifecycle can also be visualised as below:



1. Business Understanding

The entire cycle revolves around the business goal. What will you solve if you do not have a precise problem? It is extremely important to understand the business objective clearly because that will be your final goal of the analysis. After proper understanding only we can set the specific goal of analysis that is in sync with the business objective. You need to know if the client wants to reduce credit loss, or if they want to predict the price of a commodity, etc.

2. Data Understanding

After business understanding, the next step is data understanding. This involves the collection of all the available data. Here you need to closely work with the business team as they are actually aware of what data is present, what data could be used for this business problem and other information. This step involves describing the data, their structure, their relevance, their data type. Explore the data using graphical plots. Basically, extracting any information that you can get about the data by just exploring the data.

3. Data Preparation

Next comes the data preparation stage. This includes steps like selecting the relevant data, integrating the data by merging the data sets, cleaning it, treating the missing values by either removing them or imputing them, treating erroneous data by removing them, and also checking for outliers using box plots. Constructing new data, deriving new features from existing ones. Format the data into the desired structure, remove unwanted columns

and features. Data preparation is the most time consuming yet arguably the most important step in the entire life cycle. Your model will be as good as your data.

4. Exploratory Data Analysis

This step involves getting some idea about the solution and factors affecting it, before building the actual model. Distribution of data within different variables of a feature is explored graphically using bar-graphs, Relations between different features is captured through graphical representations like scatter plots and heat maps. Many other data visualisation techniques are extensively used to explore every feature individually, and by combining them with other features.

5. Data Modeling

Data modeling is the heart of data analysis. A model takes the prepared data as input and provides the desired output. This step includes choosing the appropriate type of model, whether the problem is a classification problem, or a regression problem or a clustering problem. After choosing the model family, amongst the various algorithms amongst that family, we need to carefully choose the algorithms to implement and implement them. We need to tune the hyperparameters of each model to achieve the desired performance. We also need to make sure there is a correct balance between performance and generalisability. We do not want the model to learn the data and perform poorly on new data.

6. Model Evaluation

Here the model is evaluated for checking if it is ready to be deployed. The model is tested on unseen data, evaluated on a carefully thought out set of evaluation metrics. We also need to make sure that the model conforms to reality. If we do not obtain a satisfactory result in the evaluation, we must re-iterate the entire modeling process until the desired level of metrics is achieved. Any data science solution, a machine learning model, just like a human, should evolve, should be able to improve itself with new data, adapt to a new evaluation metric. We can build multiple models for a certain phenomenon, but a lot of them may be imperfect. Model evaluation helps us choose and build a perfect model.

7. Model Deployment

The model after a rigorous evaluation is finally deployed in the desired format and channel. This is the final step in the data science life cycle. Each step in the data science life cycle explained above should be worked upon carefully. If any step is executed improperly, it will consequently affect the next step and the entire effort goes to waste. For example, if data is not collected properly, you'll lose information and you will not be building a perfect model. If data is not cleaned properly, the model will not work. If the model is not evaluated properly, it will fail in the real world.

From Business understanding to model deployment, each step should be given proper attention, time and effort.

The importance of Domain Expertise

The cyclical nature of the data science lifecycle is dependent on topic expertise, which is both the start and end of any data science project. When someone has expertise in a topic, they tend to want to know even more about it, which leads to asking questions. Questions lead to investigation and (hopefully) answers, resulting in even more knowledge of the topic. This, in turn, leads to more questions, which kicks off the whole process once again. This is what data science looks like in action.

Domain Knowledge. It is really just knowledge of the area you are working in. If a financial analyst started analyzing data about heart attacks, they might need the help of a cardiologist to make sense of a lot of the numbers. So why is this so important for data scientists? Simply put, you cannot unlock the full power of an algorithm without proper knowledge about the field where the data comes from.

Domain Knowledge in Data Science

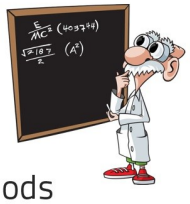
The term “Domain Knowledge” has been in play even before data science became popular. In software engineering, it means the knowledge about the environment in which the target (i.e. software agent) operates.

We can use the same definition in data science to say — “Domain knowledge is the knowledge about the environment in which the data is processed to reveal secrets of the data”. In other words, the knowledge of the field that the data belongs to is known as Domain Knowledge.

Domain knowledge can be divided into four levels. (1) Awareness is a basic level at which we are aware of the nature of the domain. (2) Foundation is knowing what the elements in the domain do, equivalent to a theoretical education. (3) Skill is having practical experience in the domain. (4) Advanced is the level at which there is little left to learn and where skill and knowledge can be provided to other people, i.e. this person is a domain expert.

Data science needs domain knowledge. As it is unreasonable to expect any one person to fulfill both roles, we are necessarily looking at a team effort. The image below captures the necessary elements for two roles, one of a data scientist, the other a domain expert.

data scientist



1. Education in data
2. Experience in data
3. Availability of methods
4. Configuration of tools
5. Model quality
6. Communication with technical staff

1. Education in domain
2. Experience in domain
3. Application of tools
4. Data availability
5. Data quality
6. Communication with intended users

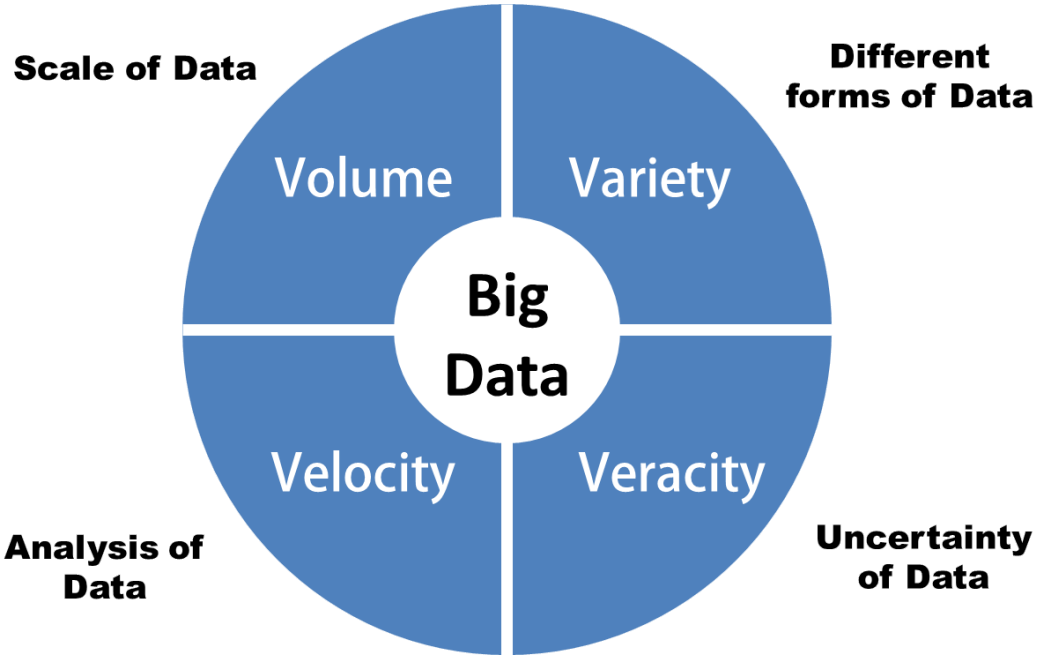
domain expert

So why is this so important for data scientists? Simply put, you cannot unlock the full power of an algorithm without proper knowledge about the field where the data comes from. Try to build a complex data model in an industry that you don't know anything about and you will have a hard time. The less we know about the problem, the more difficult it is to solve.

The V's of Data

The term volume refers to the size of the data, velocity refers to the speed of incoming and outgoing data, and variety describes the sources and types of data. IBM and Microsoft added veracity or variability as the fourth V to define big data. The term veracity refers to the messiness and trustworthiness of data. McKinsey & Co added value as the fourth V to define big data. Value refers to the worth of hidden insights inside big data. Commonly, big data is a collection of large amounts of complex data that cannot be managed efficiently by the state-of-the-art data processing technologies.

Because big data presents new features, its data quality also faces many challenges. The characteristics of big data come down to the 4Vs: Volume, Velocity, Variety, and Value (Katal, Wazid, & Goudar, 2013). Volume refers to the tremendous volume of the data. We usually use Terabytes or above magnitudes to measure this data volume. Velocity means that data is being formed at an unprecedented speed and must be dealt with in a timely manner. Variety indicates that big data has all kinds of data types, and this diversity divides the data into structured data and unstructured data. These multiple types of data need higher data processing capabilities. Finally, Value represents low-value density. Value density is inversely proportional to total data size, the greater the big data scale, the less relatively valuable the data.



On the one hand, Big Data holds great promise for discovering subtle population patterns and heterogeneities that are not possible with small-scale data. On the other hand, the massive sample size and high dimensionality of Big Data introduce unique computational and statistical challenges, including scalability and storage bottleneck, noise accumulation, spurious correlation, incidental endogeneity and measurement errors (Fan et al. 2014, p.239).

Features of Data Quality

In order to perform proper analysis and make the relevant decisions, the maintenance of data quality is important. High-quality data are the precondition for analysing and using big data and for guaranteeing the value of the data

There are seven characteristics that define data quality, these are:

- Accuracy and Precision.
- Legitimacy and Validity.
- Reliability and Consistency.
- Timeliness and Relevance.
- Completeness and Comprehensiveness.
- Availability and Accessibility.
- Granularity and Uniqueness

Accuracy and Precision: This characteristic refers to the exactness of the data. It cannot have any erroneous elements and must convey the correct message without being misleading. This accuracy and precision have a component that relates to its intended use. Without understanding how the data will be consumed, ensuring accuracy and precision could be off-target or more costly than necessary. For example, accuracy in healthcare

might be more important than in another industry (which is to say, inaccurate data in healthcare could have more serious consequences) and, therefore, justifiably worth higher levels of investment.

Legitimacy and Validity: Requirements governing data set the boundaries of this characteristic. For example, on surveys, items such as gender, ethnicity, and nationality are typically limited to a set of options and open answers are not permitted. Any answers other than these would not be considered valid or legitimate based on the survey's requirement. This is the case for most data and must be carefully considered when determining its quality. The people in each department in an organisation understand what data is valid or not to them, so the requirements must be leveraged when evaluating data quality.

Reliability and Consistency: Many systems in today's environments use and/or collect the same source data. Regardless of what source collected the data or where it resides, it cannot contradict a value residing in a different source or collected by a different system. There must be a stable and steady mechanism that collects and stores the data without contradiction or unwarranted variance.

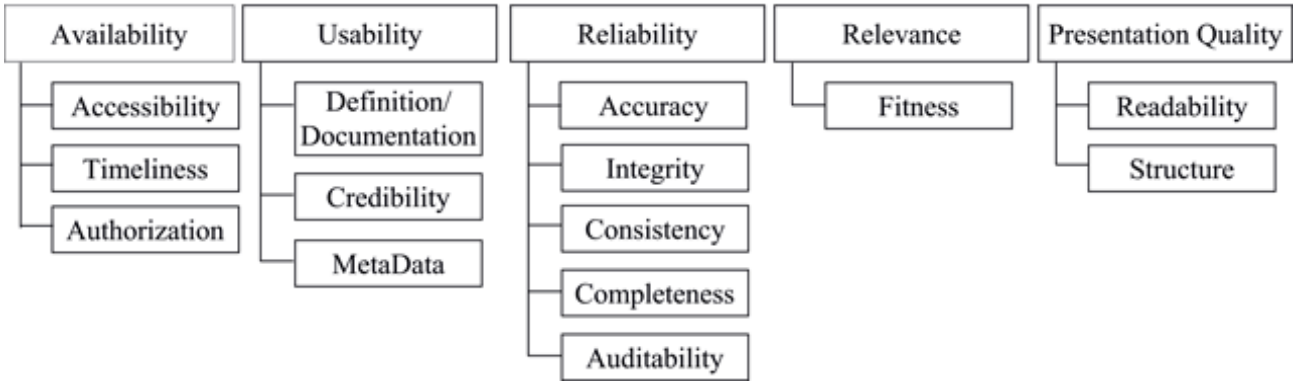
Timeliness and Relevance: There must be a valid reason to collect the data to justify the effort required, which also means it has to be collected at the right moment in time. Data collected too soon or too late could misrepresent a situation and drive inaccurate decisions.

Completeness and Comprehensiveness: Incomplete data is as dangerous as inaccurate data. Gaps in data collection lead to a partial view of the overall picture to be displayed. Without a complete picture of how operations are running, uninformed actions will occur. It's important to understand the complete set of requirements that constitute a comprehensive set of data to determine whether or not the requirements are being fulfilled.

Availability and Accessibility: This characteristic can be tricky at times due to legal and regulatory constraints. Regardless of the challenge, though, individuals need the right level of access to the data in order to perform their jobs. This presumes that the data exists and is available for access to be granted.

Granularity and Uniqueness: The level of detail at which data is collected is important, because confusion and inaccurate decisions can otherwise occur. Aggregated, summarised and manipulated collections of data could offer a different meaning than the data implied at a lower level. An appropriate level of granularity must be defined to provide sufficient uniqueness and distinctive properties to become visible. This is a requirement for operations to function effectively.

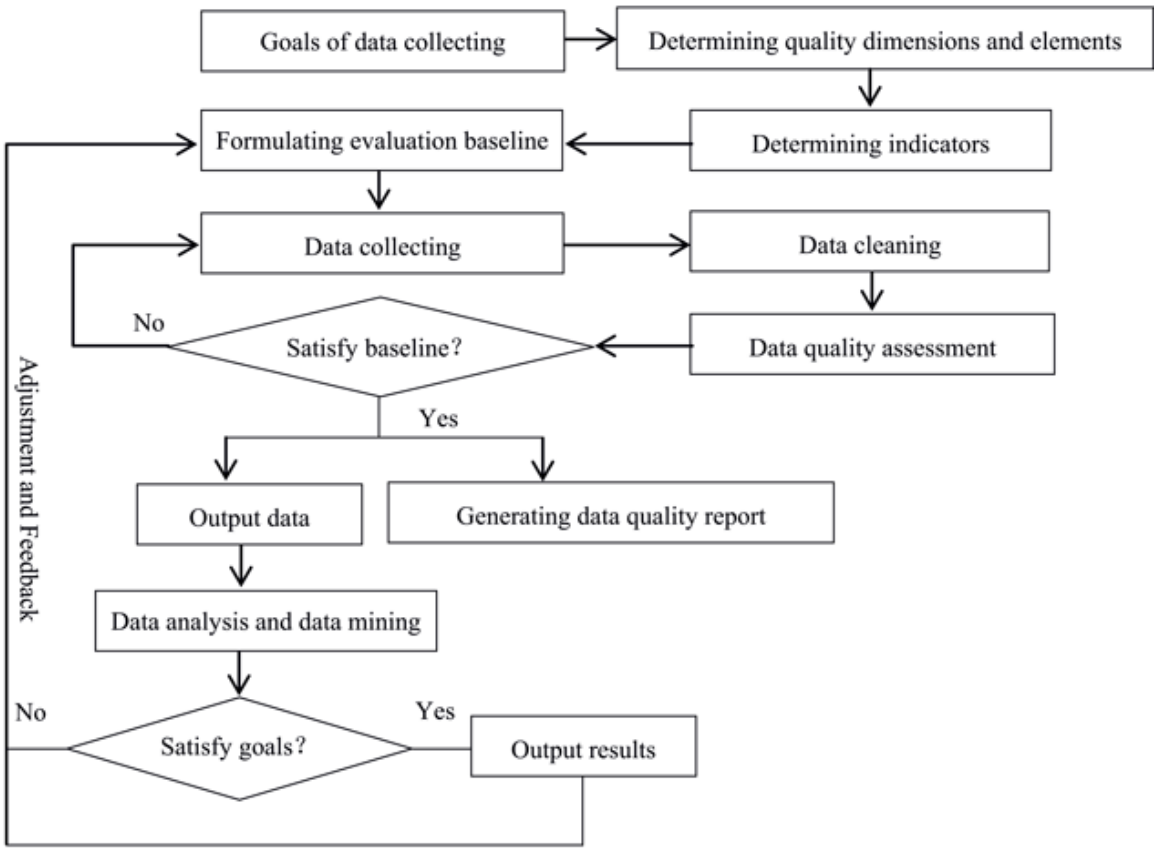
The figure below is another framework which can be used to capture the dimensions of data quality.



The challenges of quality data in the era of big data:

- The diversity of data sources brings abundant data types and complex data structures and increases the difficulty of data integration.
- Data volume is tremendous, and it is difficult to judge data quality within a reasonable amount of time.
- Data change very fast and the “timeliness” of data is very short, which necessitates higher requirements for processing technology
- No unified and approved data quality standards have been formed in China and abroad, and research on the data quality of big data has just begun

Given the complexity of the issues relating to data quality, one might consider using a tool such as the one shown in the figure below to assess the quality of data for any given project.



Types of Bias

Bias is important, not just in statistics and machine learning, but in other areas like philosophy, psychology, and business too.

Generally, bias is defined as “prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair.” Data scientists need to be aware of the potential for bias, stay vigilant and do everything they can to minimise it in their models and analysis.

Types of bias which are important in data science are statistical, data and sampling bias, alongside algorithmic and programming bias.

Bias reflects problems related to the gathering or use of data, where systems draw improper conclusions about data sets, either because of human intervention or as a result of a lack of cognitive assessment of data. An increasing number of researchers, practitioners and policy makers are realising that much needs to be done to deal with bias in data and algorithms, and to promote transparency of Artificial Intelligence (A.I.) models. Only in this way can the proper use of A.I. can be ensured and benefits to people’s lives and support for fundamental human rights can be expected.

Statistical, Data and Sampling Bias

Statistical bias is essentially when a model or statistic is unrepresentative of the population, and there are several sources of bias that cause this.

The most common sources of bias include:

- Selection bias
- Survivorship bias
- Omitted variable bias
- Recall bias
- Observer bias
- Funding bias

The common definition of data bias is that the available data is not representative of the population or phenomenon of study. But it can also be used more broadly to denote when:

- Data does not include variables that properly capture the phenomenon we want to predict
- Data includes content produced by humans which may contain bias against groups of people

Based on this definition, except for data generated by carefully designed randomised experiments, most organically produced datasets are biased. Data bias occurs due to structural characteristics of the systems that produce the data.

Sampling bias occurs when a sample statistic does not accurately reflect the true value of the parameter in the target population, for example, when the average age for the sample observations does not accurately reflect the true average of the members of the target population. Typically, sampling bias focuses on one of two types of statistics: averages and ratios (Lavrakas 2008).

For example, consider the image below. To give an example, imagine that there are 10 people in a room and you ask if they prefer grapes or bananas. If you only surveyed the three females and concluded that the majority of people like grapes, you'd have demonstrated sampling bias.

Algorithmic & Programming Bias

Algorithmic bias can be defined in a variety of specific, technical ways, but is increasingly being used in reference to fairness and discrimination. As our societies become increasingly dependent on algorithms, we are seeing our age-old prejudices, biases and implicit assumptions reflected back at us in digital form. But the algorithmic systems we use also have the potential to amplify, accentuate and systemise our biases on an unprecedented scale, all while presenting the appearance of objective, neutral arbiters. Therefore, a reasonable definition of algorithmic bias in the sense we are using here is the unfair treatment of a group (e.g. an ethnic minority, gender or type of worker) that can result from the use of an algorithm to support decision-making. The Equality Act 2010 prohibits discrimination against people on the basis of certain protected characteristics, while the GDPR and Data Protection Act 2018 have introduced privacy restrictions which must be considered when making assessments for algorithmic bias.

The ways in which algorithmic bias is likely to be expressed, and the consequences for individuals and groups, is highly context-specific. In some areas there may be severely detrimental consequences for relatively small numbers of individuals, while in others there may be relatively minor consequences which are distributed across large subsections of society.

In the US, similar concerns have even led to the cancellation of comparable programmes. In 2017, for example, both the Illinois Department of Children and Family Services and the County of Los Angeles Office of Child Protection terminated the use of predictive analytics programmes, in part due to perceptions of inaccuracy (including high false positive and false negative rates), the poor quality of data being used and the difficulty of verifying their decisions.

If bias is introduced even subconsciously by a computer programmer, when developing a system, this is known as programming bias.

Opacity in Machine Learning, the so-called black-box effect, is often mentioned as one of the main impediments for transparency in Artificial Intelligence.

Removing the algorithmic black-box will not eliminate the bias. You may be able to get a better idea of what the algorithm is doing but it will still enforce the biased patterns it ‘sees’ in the data.

The use of Black boxes can lead to machine bias, which is the effect of erroneous assumptions in machine learning processes. Machine learning is only as smart as the AI that has been programmed, and the data that can be used to learn from. Where bias comes in machine learning is that humans have a certain built-in bias, and by nature, some of that bias will exist in machine learning unless it is looked for. When the bias is found, internal/open tools work to understand how to correct it.

An example of bias in automated decision making is Amazon's 'prime-lining', similar to historical red-lining, where services were denied to more disadvantaged neighbourhoods and Prime delivery services were concentrated in predominantly white neighbourhoods.

Hidden biases have a serious impact on society and in many cases the divisions that have appeared among us.

One such example is found here at: http://gendershades.org/. The research to determine if there were any biases in the algorithms of three major facial recognition AI service providers— Microsoft, IBM and Face++— was conducted by providing 1270 images from a mix of individuals originating from the continent of Africa and Europe. The performance of accuracy was good overall, however the algorithms performed poorly when classifying dark skinned individuals, particularly women. Clearly, any decisions that one makes based on the classification results of these algorithms, would be inherently biased and potentially harmful to dark skinned women in particular.

Summary of the causes of bias:

There are a variety of causes of bias, including sample bias, exclusion bias, measurement bias, stereotype bias, survivorship bias, simpson's paradox, confirmation bias and correlation bias.

Exclusion bias refers to when predictive variables are removed prior to the modelling process. Measurement bias arises from systematic issues with the collection of data, often through calibration or faulty detectors. Stereotype bias is related to cultural or gender stereotypes which can lead to an uneven distribution of modelling data. Survivorship bias arises from concentrating on the successful output of a process and ignoring those that don't survive. Simpson's paradox refers to a statistical phenomenon that leads to incorrect inferences when trends for subgroups reverse the overall norm. Correlation bias arises from inferring a correlation where it doesn't exist due to confounding variables that have not been accounted for. Confirmation bias arises from cherry-picking data or variables that confirm existing beliefs and expectations. A simple example of this would be the use of specific search terms that bias the results:

If you were to search "Are cats better than dogs?" in Google, all you will get are sites listing the reasons why cats are better. However, if you were to search "Are dogs better than cats?" google will only provide you with sites that believe dogs are better than cats. This shows that phrasing questions in a one-sided way (i.e. affirmative manner) will assist you in obtaining evidence consistent with your hypothesis.

As a data scientist, try to be aware of as many causes of bias as possible and be mindful to avoid bias when working with data. Reflect on the ways in which bias is built into different digital products and services. What other examples of bias can be identified beyond those presented here.

References

Anand. 2019. Why Domain Knowledge is Important in Data Science. Medium. Available at: https://medium.com/@anand0427/why-domain-knowledge-is-important-in-data-science-anand0427-3002c659c0a5

Banerjee, D. 2019. Data Science : Brief understanding of Typical Project Life-cycle, Tools, Techniques and skills. Data Science Foundation. Available at: https://datascience.foundation/sciencewhitepaper/data-science-brief-understanding-of-typical-project-life-cycle-tools-techniques-and-skills

Cai, L. and Zhu, Y., 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14: 2. Available at: https://datascience.codata.org/articles/10.5334/dsj-2015-002/

EDUCBA. n.d. Data Science Lifecycle. Available at: https://www.educba.com/data-science-lifecycle/

Fan, J. et al. 2014. Challenges of Big Data analysis. National Science Review, 1(2): 293–314.

Godsey, B. 2017. Think Like a Data Scientist: Tackle the data science process step-by-step. New York: Manning Publications.

Krishnamurthy, P. 2019. Understanding Data Bias: Types and sources of data bias. Towards Data Science. Available at: https://towardsdatascience.com/survey-d4f168791e57

Lavrakas, P. 2008. Sampling Bias In: *Encyclopedia of Survey Research Methods.* Available at; https://methods.sagepub.com/reference/encyclopedia-of-survey-research-methods/n509.xml

Ortega, D. 2017. Seven Characteristics That Define Quality Data. Blazent. Available at: https://www.blazent.com/seven-characteristics-define-quality-data/

Shin, T. 2020. What is statistical bias and why is it so important in data science? Towards Data Science. Available: https://towardsdatascience.com/what-is-statistical-bias-and-why-is-it-so-important-in-data-science-80e02bf7a88d