**DATA AND DATA COLLECTION**

1. **What is Data and Data Collection?**

Data collection can be understood as a procedure of collecting, gauging, and analyzing the exact insights to do effective research with the help of best-suited techniques that help researchers evaluate their hypothesis.

Data collection is the process of gathering and measuring information for variables of interest in an established and systematic fashion enabling one to answer queries, state research questions, test hypothesis and evaluate the outcomes.

1. **Where does Data Collection fall in the Data Science Spectrum?**

Data collection is an aspect int the Data science spectrum that is mostly forgotten. We will be looking at different aspects of Data Collection. But where does Data Collection fall in the Data Science Road Map. Here is a list that shows the following.

*a). Understanding the problem* – All Data Science problems start with having a clear comprehension of the problem at hand. This aids to know the data that will needed and probably the type of Data science/Statistics solution that may help solve the underlying problem.

*b). Collecting the Data* – This is the act of gathering information and comes after understanding the need of Data.

*c). Explore and Clean Your Data* – This is the act of turning your data into a more comprehensible format that will fix or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a collected set of Data.

*d). Enrich your data* - is the process of combining first party data from internal sources with disparate data from other internal systems or third-party data from external sources. This is not as common; however, it is a wonderful procedure when you get the right data to enrich your collected data.

*e). Build Visualizations/ algorithms* – This is the creation of solutions using visualizations or Machine Algorithms that will lead to coming up with the resolutions.

*f). Make decisions* – This are the commitments of action after the visualizations and algorithms.

1. **Importance of Data Collection**

*a). The trustworthiness of The Research –* A critical purpose behind data collection from a research perspective is to have honesty, ethics and integrity backing it.

*b). Diminish the probability of blunders or errors* - The utilization of the right data collection methods to different research processes decreases the probability of blunders. Lack of blunders fortifies the trustworthiness of research.

*c). Effective and accurate decision making -* Data collected in the right methods will lead to making the right decision that will impact the need for data collection in the first place.

*d). Have Cost and Time -* Data collection enables one to understand the research being done in a more clear and comprehensible way.

*e). Empowers a new idea or change* - At most times there are notions about the result. Data collection helps changing the narrative and delete predetermined results.

Now that we know the importance of data, we should also discern that data collected might not be the representation of what is on the ground. Hence, collected data should be good data and we will investigate its components.

1. **What are the components of Good data?**

Good data should have the following components.

*a). Accurate* - Accuracy of Data is when the observer values match the true values. Sometimes in Data collection, it is logistically hard to collect data depending on various variables like location, willingness to have respondents share among other reasons. Values amassed should not influenced by external forces, its all about the articulation of the data collection as it.

*b). Reliability* - Reliability of data is same values are determined after repeated measurements all variables held constant. This means if two people were to carry down a research at the same place at the same time asking the same questions to respondents, there shouldn’t be a big variance in results.

*c). Validity* - Validity of Data is when your findings truly represent the phenomenon you claim to measure. Validity indicates how sound your research is.

These three components help to determine if data is good and it suffices to be used to make conclusions.

1. **Different Types of Data**

Its important to know what the formats of data are and how they appear during collection depending on the collection methods.

a). *Structured data*

This is well-organized data. It is data stored in a database. Data stored in databases allows easy access, retrieval, updating and manipulation.   
Most of these databases are Relational.   
**RDBMS** - A database management system (DBMS) that incorporates the relational-data model, normally including a Structured Query Language (SQL) application programming interface. In a relational database, relationships between data items are expressed employing tables. This data may have many tables that are connected using primary keys and foreign keys.

Most of this data is easily refreshable and is collected continuously from business systems like points of sale. So in your spheres, just look into anywhere you can have repetitive data collection that is automated so that it goes into a Database somewhere and through that Database you can get an end point of the data and use it for visualizations that will help to make decisions.

b). *Semi-structured data*

This is data that has some aspects of Structured Data like having a structure of rows and columns but there is no connection of the different tables. Such data is in Excel or CSV data.

Semi-structured data also includes text that is organized by subject or topic or fit into a hierarchical programming language, yet the text within is open-ended, having no structure itself.

Sometimes it is loosely organized into categories using meta tags.

Examples of Semi-structured data could include Emails of inbox, tweets organized by hashtags

c). *Unstructured Data.*

This is data without a structure or a schema. It is mostly text-heavy information not organized in a clearly defined framework or model. The examples include plaintext, tweets, recording voices, videos, sound data.

1. **Data Collection Methods**

*a). Interviews* - Interviews are used to collect data from a small group of subjects on a broad range of topics. You can use structured or unstructured interviews. Structured interviews are comparable to a questionnaire, with the same questions in the same order for each subject and with multiple choice answers.

*b). Observations* - Observation is the way of gathering data by watching behavior, events, or noting physical characteristics in their natural setting. Through observation, you can monitor or watch a process or situation that you are evaluating as it occurs. When you are gathering data on individual behaviors or interactions between people.

*c). Documents and Records* - Document- and records-based research uses existing data for a study.

*d). Focus Groups* - Focus groups are a qualitative data collection method, meaning that the data is descriptive and cannot be measured numerically. When should you use focus groups for evaluation? To get more in-depth information on perceptions, insights, attitudes, experiences, or beliefs.

*e). Oral Histories* - Oral history is a method of conducting historical research through recorded interviews between a narrator with personal experience of historically significant events and a well-informed interviewer, intending to add to the historical record.

*f). Questionnaires* - A questionnaire is a research instrument that consists of a set of questions or other types of prompts that aims to collect information from a respondent. A research questionnaire is typically a mix of close-ended questions and open-ended questions.

**7. What is Digital Data collection?**

Digital data collection is the process of collecting data electronically using existing technology such as smartphones, tablets, and other digital devices.

**a). What are the advantages of Digital Data Collection?**

1. *Is efficient and accurate.*
2. *Data quality and specification.*
3. *Entry limits.*
4. *Specifying compulsory and non-compulsory questions.*
5. *It is cost effective.*
6. *Easy monitoring.*
7. *Secure.*
8. *Most is already in good shape during collection.*

**9). What are the digital data collection methods?**

a). ***Kobo toolbox -*** *Is a field data collection form. Forms are designed on a Computer web interface but can be filled on offline and online mobile devices.*

*b).* ***Survey monkey*** *- Is an online survey tool used for sending creating, sending surveys and getting feedback.*

*c).* ***ODK -*** *Is an open-source software for collecting, managing, and using data. It is easy to use, modify and scale.*

*d).* ***MS & google forms -*** *You can use these online tools to create questionnaires, get feedback and see the analytics of the feedback.*

*e).* ***Sentiment analysis data collection -*** *This involves getting data from different sites e.g., social media and one can tell the opinion about a certain discussion.*

*f).* ***Web-Scrapping -*** *Collecting and extracting data from websites. One can use requests with beautiful soups or Selenium.*

*g).* ***Survey cto –*** *Is a reliable secure mobile data collection app for researchers.*

*h).* ***Phone data collection -*** *Using phone calls, messaging, and WhatsApp to collect data from target audience.*

*10.***What is Data enrichment? -** Is the process of getting more data related to the collected data.

- It is more of merging third-party data from an external authoritative source with an existing database of first-party customer data.  
  
**Explorium** – This is a site that contains third- party data and one can use it to increase model accuracies and have more useful features for research.

**11. What are the ethics in Data collection?**

Data ethics encompasses the moral obligations of gathering, protecting, and using personally identifiable information and how it affects individuals.

12**. What are the principles of Data Ethics?**

***a). Privacy /Ownership***

Need people consent. as PII, personally Identifiable Information relates to the information that identifies with the individual and can be used to identify who you are.

***b). Transparency***

Be clear on the reason for Data Collection

***c). Intention***

Know the reason why you need the data**.**

**13. What are the main Secondary sources that you can obtain data?**

1. *GitHub -* [*link*](https://github.com/)
2. *Data World -* [*link*](https://data.world/)
3. *UCI Machine Learning Repository -* [*link*](https://archive.ics.uci.edu/ml/index.php)
4. *Kaggle -* [*link*](https://www.kaggle.com/)
5. *Zindi -* [*link*](https://zindi.africa/competitions)
6. *Open Data -* [*link*](https://www.opendata.go.ke/)
7. *Data. Gov -* [*link*](https://www.data.gov/)
8. *Google Datasets -* [*link*](https://datasetsearch.research.google.com/)

I hope you have gained much about Data collection and this will help in your data collection journey. In case you need slides for the same Data Collection check it on my GitHub repository [here](https://github.com/DerrickKuria/Data-Collection).