Steps involved in data science project

1. Defining problem statement
2. Data collection
3. Data Cleaning
4. Exploratory Data analysis
5. Preprocessing
6. Model building
7. Evaluation
8. Visualization
9. Analysis/Deployment

Methods of Data Collection

Presented by
Santosh Borkakati
Asst. Professor
Department of Economics
Mangaldai College

What is Data?

- Data is a existing information /knowledge *represented* or *coded* in some form suitable for better usage or processing.

- Data is a set of values of qualitative or quantitative variables.

Quantitative Vs Qualitative Data

- Quantitative data are anything that can be expressed as a number, or quantified. These data may be represented by ordinal, interval or ratio scales and lend themselves to most statistical manipulation.
- Qualitative data is a categorical measurement expressed not in terms of numbers, but rather by means of a natural language description. In statistics, it is often used interchangeably with "categorical" data.

  For example: favorite color = "blue"

Quantitative Vs Qualitative Data

- Quantitative and Qualitative data can be gathered from the same data unit depending on whether the variable of interest is numerical or categorical. For example:

Data unit	Numeric variable	= Quantitative data	Categorical variable	= Qualitative data
A person	"*How many* children do you have?"	**2** children	"In *which country* were your children born?"	**India**
	"*How much* do you earn?"	**Rs.60,000** p.m.	"*What* is your occupation?"	**Teacher**
	"How many hours do you work?"	**40** hours per week	"Do you work *full-time or part-time*?"	**Full-time**

Primary and Secondary Data

- The task of data collection begins after a research problem has been defined and research design/plan chalked out.

- While deciding about the method of data collection to be used for the study, the researcher should keep in mind two types of data viz., primary and secondary.

Primary and Secondary Data

- Primary Data are collected by the researcher.
- Secondary data collected by someone else and have already been passed through the statistical process.
- A researcher as per requirement of study may decide on use of primary data or secondary data or both.
- Both primary and secondary data have their own pros and cons.

Methods of Collecting Data

- The methods of collecting data mainly refers to collecting primary data.
- As secondary data are already available, we have to carefully choose the sources , relevancy of data and reliability.

Collecting Secondary Data

- Sources of secondary data are existing literature, Reports of professional agencies, Departments, Archives, Internet, etc.
- While collecting secondary data one has to follow legal procedures required and maintain the academic ethics.

Methods of Collecting Primary Data

There are several methods of collecting primary data, particularly in surveys and descriptive research. Important ones are-

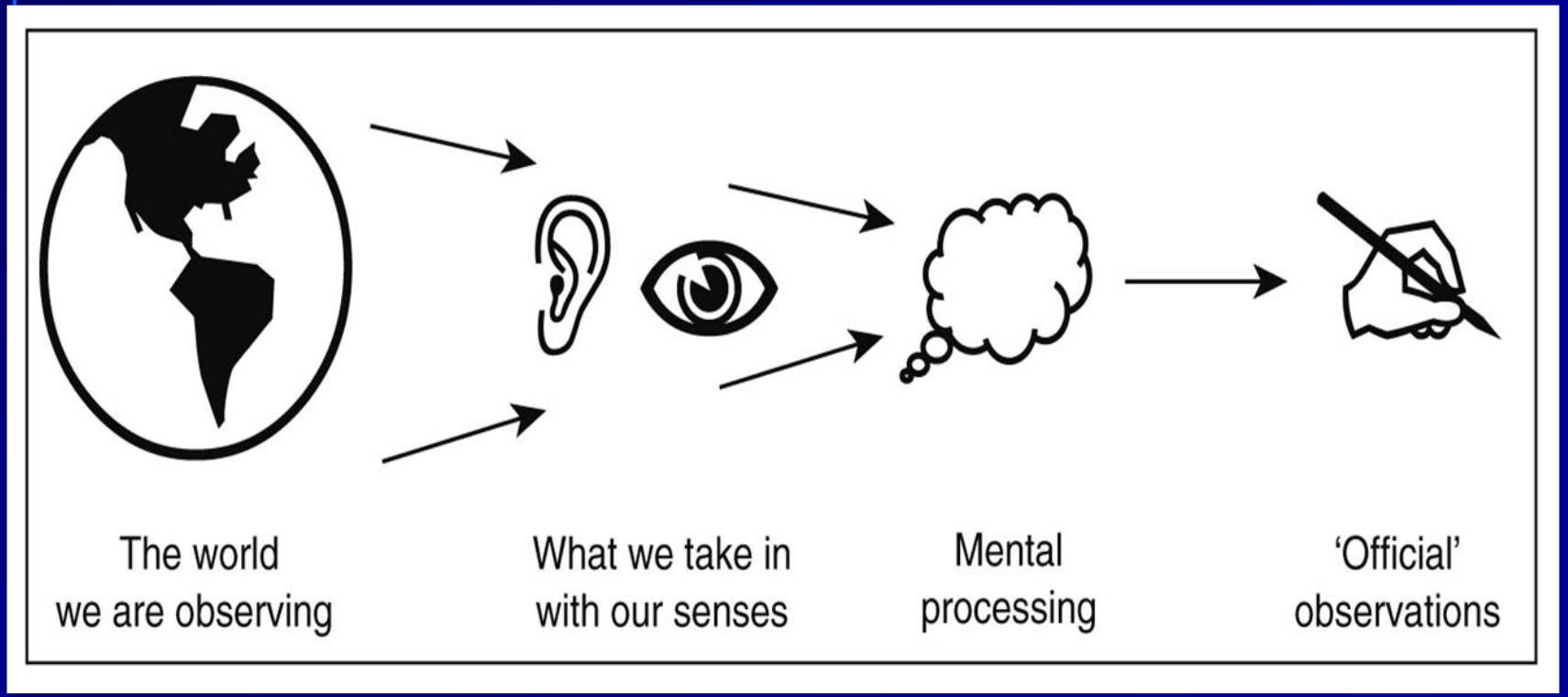
- Observation-ECG, EEG, BP, Rain, temp etc.
- Interview
- Questionnaire
- Schedule
- Other Methods – online, image/video capturing, sound recorcding, text collection, web scrapping etc.

Observation

See what is happening

- traffic patterns
- land use patterns
- layout of city and rural areas
- quality of housing
- condition of roads
- conditions of buildings
- who goes to a health clinic

Filtering Observations



Observation is Helpful when:

- Need direct information
- Trying to understand ongoing behavior
- There is physical evidence, products, or outputs than can be observed
- Need to provide alternative when other data collection is infeasible or inappropriate

Types of Observation

- Participatory and Non Participatory
- Candid and Covert (both are participatory)
- Structured, Semi-structured and Unstructured.
- Controlled and Uncontrolled

Advantages/Disadvantages of Observation

Advantages:

- Subjective bias eliminated
- Researcher gets current information
- Independent of Respondents
- Disadvantages:
- Expensive, Time consuming
- Limited information
- Unforeseen factors may influence observation

Interview

- The interview method of collecting data involves presentation of oral-verbal stimuli and reply in terms of oral-verbal responses.
- This method can be used through personal interviews or telephone interviews.
- Structured, Semi-Structured or Unstructured Interview.

Interview Types

- Personal Interviews: Interviewer asking questions generally in a face-to-face contact to the other person or persons. Direct personal investigation or Indirect oral investigation.
- Focused Interview is meant to focus attention on the given experience of the respondent and its effects.
- Clinical Interview is concerned with broad underlying feelings or motivations or with the course of individual's life experience.
- Non-directive Interview is that where the interviewer's function is simply to encourage the respondent to talk about the given topic with a bare minimum of direct questioning.

Skill of Interviewer

The main game in interviewing is to facilitate an interviewee's ability to answer. This involves:

- easing respondents into the interview
- asking strategic questions
- prompting and probing appropriately
- keeping it moving
- winding it down when the time is right

Merits/Demerits of Interview

Merits:

- More and in depth information obtained
- Personal Information
- Greater Flexibility
- Adaptation as per the respondent

Demerits:

- Bias of Interviewer
- Expensive/Time Consuming
- Need expertise

Questionnaire Method

- A questionnaire is sent (usually by post) to persons concerned with a request to answer the questions and return the questionnaire.
- A questionnaire consists of a number of questions printed in a definite order.
- The respondents have to answer the questions on their own.

Steps in questionnaire construction

- Preparation
- Constructing the first draft
- Self-evaluation
- External evaluation
- Revision
- Pre-test or Pilot study
- Revision
- Second pre-testing
- Preparing final draft

Essentials of a Good Questionnaire

- Questionnaire should be short and simple
- Question arranged in from simple to difficult.
- Personal and intimate questions should be left to the end.
- Technical term and vague expression should be avoided.
- Questions should be answered in yes or no ; multiple choice.
- Control question to cross check the information of the responded.

Advantages of Questionnaire

- Lower cost
- Time saving
- Accessibility to widespread respondents
- No interviewer's bias
- Greater anonymity
- Respondent's convenience
- Standard wordings
- No Variation

Disadvantages of questionnaire

- Questionnaires can be used only for educated people.
- Sometimes different respondent's interpreted questions differently
- Questionnaires do not provide an opportunity to collect additional information
- Researchers are not sure whether the person to whom the questionnaire was mailed has himself answered the questions.
- Many questions remain unanswered
- The respondent can consult other persons before filling in the questionnaire.

Collection of Data Through Schedule

- Schedules like questionnaires contain a set of questions.
- Researcher /Enumerators appointed collect data through schedules.
- Enumerators go to the field, put questions to the respondents and fill the schedules.
- Enumerators need to be trained.

Questionnaire Vs. Schedule

Questionnaire	Schedule
Mailed, filled by Respondent	Direct contact , filled by Researcher or Enumerator
Economical	Expensive
Non-Response high	Non-Response low
Time Consuming	Time bound
Literate, co-operative respondents	No such pre condition
Success depends on quality of questionnaire	Success depends on quality of enumerator

Some Other Methods

- Warranty Cards Post card size cards sent to customers and feedback collected through asking questions.
- Distributor or Store Audits are performed by manufacturer/distributor through salesmen. Information so obtained are used to estimate market size, market share, seasonal sales pattern, etc.
- Pantry Audits From the observation of pantry of customer to know purchase habit of people (of which product, what brand, etc.). Questions may be asked at the time of audit.

Some Other Methods

- Consumer Panels Pantry audit approach on a regular basis is known as 'consumer panel', where a set of consumers are arranged to come to an understanding to maintain detailed daily records of their consumption and the same is made available to investigator on demands.
- Projective techniques developed by psychologists to use projections of respondents for inferring about underlying motives, urges, or intentions which are such that the respondent either resists to reveal them or is unable to figure out himself.

Some Other Methods

- Use of Mechanical Devices Eye Camera is used to record the focus of eyes of a respondent on a specific portion of a sketch or diagram or written material. Psychogalvanometer is used for measuring the extent of body excitement as a result of the visual stimulus. Motion picture camera is used to record movement of consumer at time of purchase. Audiometer is used to know the preferences to TV channels, programmes.

Some Other Methods

- Depth interviews are those interviews that are designed to discover underlying motives and desires and are often used in motivational research. Indirect question or projective technique are used to know the behaviour of respondents.

- Content Analysis Analyzing the contents of documentary materials such as books, magazines, newspapers and the contents of all other verbal materials which can be either spoken or printed.

Selection of Appropriate Method of Data Collection

- Nature, Scope and Object of enquiry
- Availability of Fund
- Availability of Time
- Degree of Precision Required

Precautions in Data Collection

- The data must be relevant to the research problem.
- It should be collected through formal or standardized research tools.
- The data should be such as these can be subjected to statistical treatment easily.
- The data should have minimum measurement error.

Precautions in Data Collection

- The data must be tenable for the verification of the hypotheses.
- The data should be collected through objective procedure.
- The data should be accurate and precise.
- The data should be reliable and valid
- The data should be complete in itself and also comprehensive in nature.

THANK YOU