**Module 5 Questions:**

**Q1.** Explain the difference between data collection and data capture (10mrks)

**Key terms**

d*ata collection is any process where purpose is to acquire or assist in the acquisition of data.*

**Data collection**

* Careful planning of the collection process should include the establishment of roles and responsibilities regarding all aspects linked to collection, including its communication strategy, execution, assessment, monitoring, contingency planning and security.
* Design the collection process in order to reduce respondent burden and collection cost, and to maximize timeliness and data accuracy. Data could be collected through self-enumeration, telephone interviews or personal interviews with either a paper or an electronic questionnaire (e.g. electronic data reporting, Internet, computer-assisted interviewing). To achieve the design objectives stated above more easily, consider using more than one method throughout the collection cycle. For instance, collection may start with self-enumeration using a paper or Internet questionnaire and may finish with a personal interview. For self-enumeration surveys, use multiple events (e.g. pre-collection advertisement card, introduction letter with the questionnaire, reminder card, reminder call or visit) over the collection period in order to stimulate the return of questionnaires. Examine whether some of the data elements could be acquired via administrative records instead of the more costly and sometimes less accurate traditional collection methods. Consider conducting the collection as a supplement to a large-scale survey. This would not only potentially reduce survey costs and respondent burden but also make available a wealth of information for nonresponse adjustment. When feasible, conduct pilot studies or tests to help determine or fine-tune the collection operation.
* Establish appropriate sample control procedures and measures for all data collection operations (e.g. delivery and return of paper questionnaires, follow up on gaps or inconsistencies, follow up on nonresponse). Such procedures track the status of sampled units from the beginning through to the completion of data collection so that data collection managers and interviewers can assess progress at any point in time. This is particularly important for surveys that use many data collection modes and that move cases from one mode to another (or from one collection center to another). Sample control procedures are also used to ensure that every sampled unit is processed through all the steps subsequent to data collection (i.e. capture and coding steps), with a final status being recorded. Sample control measures can be used to evaluate the efficiency of those procedures.
* Establish and maintain good respondent relationships in order to obtain a good response rate. Such measures can include advertising the upcoming survey, an introductory letter to inform the respondents that they will be part of a survey, an informative brochure with key statistics to maintain their interest in participating in the survey (in particular for longitudinal surveys) or procedures facilitating access to publicly available information (for example on a website, a guide to complete the questionnaire or helpline particularly for self-enumeration surveys) or a letter thanking them for their participation. These measures will help to sensitize the units selected in the sample to participate in the survey.
* When collecting data, ensure that the respondent or the appropriate person within the responding household or organization is contacted at the appropriate time. Allow the respondent to provide the data in a method and format that is convenient to the individual and his or her organization. This will help increase response rates and improve the quality of the information obtained from the respondents. Special reporting arrangements should be considered in specific cases in order to reduce respondent burden and to facilitate the collection of information.  For example, consider creating a special collection arrangement for enterprises that are involved in many surveys. For households, when the targeted respondent is not available, establish rules to determine who could act as an appropriate proxy should this be an option.
* For collection by interview, determine the best time to call or to visit survey units based on paradata acquired during previous iterations of the survey or from a similar survey. Manage calls or visits in such a way that respondents are contacted at the best time and that the number of call or visit attempts does not exceed a useful maximum. In addition, respondents should each be assigned a priority level so that they may be contacted or visited for interviews based on order of importance. Assignment priority should be based on the target effective sample size by domain of interest that would lead to estimates accurate enough (having low bias and variance) to be released. For business surveys, this would mean giving higher priority to large or influential units first, possibly at the risk of missing smaller units. For household surveys, priority should be given to units less likely to respond. A score function is a useful tool for prioritization. For telephone interviews, use an automated system to manage case call scheduling. Such a system should also prioritize cases.
* Interviewers are vital to the success of data collection operations. Interviewer manuals and training must be carefully prepared and planned, since they provide the best way to guarantee data quality (e.g. high response rate and accurate responses), the comprehension of survey concepts and subject matter, as well as to ensure proper answers to questions from respondents. Training can use different approaches such as home study, classroom training, mock interviews or live interviews. Interviewing skills of interviewers should be monitored to ensure that they conform to a pre-established list of standards (e.g. reading questions as written in the questionnaire). This monitoring should also be used to identify strengths and weaknesses in the interviewer's skill set, to provide feedback to the interviewers and to focus training on weaker areas. Depending on the interviewing mode and resources, the monitoring may either be done using recordings of the interviews or live. Consultation with interviewers and staff directly responsible for collection operations will help in the development of better training tools. Follow-up interviews with respondents may also be used to get the respondent's point of view on how the interview was carried out.
* Tracing should be conducted to locate and contact respondents when the available contact information on the survey unit is likely to be outdated. Tracing increases response rate and helps in determining if the sampled unit is still in scope. Consider using administrative sources (e.g. telephone files, other survey frames) prior to survey collection and during collection in order to update contact information. During collection, facilitate high quality tracing by obtaining extra information related to the sample unit, for example, the names of other family members, relationship, age, etc. Local knowledge might also be useful. Consider forming a team of tracing experts when the survey is repeated, or its collection period is over several months. In between cycles, facilitate feedback from the respondent to update contact information. For example, provide the respondents with a "change of address" card and ask them to notify the Agency if a move occurs. Collect tracing information (e.g. Internet address, cell phone number) that can be used in the subsequent survey cycles.
* For self-enumeration surveys, once the data is received, verify gaps or inconsistencies related to accuracy of the coverage information and the quality of the data provided. Follow-up interviews may be needed in some cases (e.g. when the questionnaire is missing a large number of items). Assign a follow-up priority based on the statistical importance of these units and of the missing items.
* Given that self-enumeration surveys tend to result in lower unit response rates, consider following up with non-respondents by telephone or in person to obtain their participation or conduct an interview. Ensure that collection staff is informed in a timely fashion of the registration of returned questionnaires in order to avoid unnecessary follow-ups. This type of follow-up is particularly important in the case of longitudinal surveys where the investment is clearly more long-term and the sample is subject to accumulating attrition (and possibly bias) due to nonresponse at each survey occasion. Unit nonresponse follow-ups should also be prioritized with the approach described above for managing interview surveys. Paradata (e.g. number of call or visit attempts) can also be useful to prioritize follow-ups.
* As a last step of the collection operation, consider contacting a sub-sample or all of the non-responding units (including unresolved cases) to determine whether they are in scope or not (e.g. active business or not, occupied dwelling or not); and, if so, a critical data item such as size (e.g. business total income, household size) should be obtained. This information will be useful for the nonresponse adjustment. In some instances, the information can be obtained from or approximated by current administrative data for all of the non-responding units.
* Provide plans and tools to actively manage survey data collection while it is in progress. Productivity measures (e.g. daily and cumulative number of units resolved) and cost indicators (e.g. daily and cumulative interviewer hours and travel expenditures) can be used to assess the relationship between the collection effort and the results (e.g. unit response rate). Compared with planned values, these indicators also help survey managers in their decision-making throughout the collection period. Used in conjunction with the daily response rate, the daily productivity rate and daily average unit cost provide the marginal cost of response rate increase during the course of collection.  Activity and cost indicators (according to selected unit or completed questionnaire) also make it possible to evaluate the additional costs and effort required to increase response rates, particularly towards the end of the collection period.
* Every effort should be made to ensure the confidentiality of the data. Staff handling confidential data must be familiar with best practices regarding the printing, handling and filing of paper documentation, the handling of electronic files, and the rules regarding the dissemination of information.
* Consider implementing a re-interview program to assess the overall accuracy of interviewing operations.
* Use Para data to identify operational efficiency and cost-efficiency opportunities (e.g. sequence of calls, best time to call, optimal limit for calls or visits, etc.) in order to improve current and future collection processes and practices. For example, use average and distribution of interview duration to plan the next survey cycle. Interview duration can also be used to evaluate part of respondent burden. If interview duration is analyzed by interviewer, it can be used as well to identify those potentially requiring additional training (e.g. those with outlying average duration).

### Key terms

*While data capture is referred to any process that convert the information provided by a respondent into electronic format this conversion is either automated or involves staff keying the collected data.*

Data capture

* Design the capture process in order to reduce capture cost and to maximize timeliness and data accuracy. Data items could be captured during survey collection by the respondents (e.g. Internet, EDR) or the interviewers (e.g.CATI, CAPI). This obviously reduces the cost of capture, increases the timeliness and has the potential of improving accuracy through edit rules being integrated into the computer application. When it is not feasible to integrate capture with collection, the capture is performed either by operators (manual key entry) or in an automated fashion (scanning followed by Intelligent Character Recognition). The latter is preferred as it reduces cost and often enhances accuracy of the data.
* For CATI and CAPI interviewers, who often perform data capture and coding during collection, use standard collection tools and process (e.g. standard screens and standardized questions) to ease interviewer work and limit the risk of introducing a capture error. Integrate edit rules in the collection system to validate the entry of data items and allow for potential corrections of errors (i.e. keying error, response error and missing item) at the time of collection.
* Data capture operators are critical to the success of the capture operations. Ensure that they have appropriate training and tools. Prepare training material and procedures for the keyers and deliver training sessions. This will enhance the skills of the staff and thus ensure accurate capture of data collected. Use quality control methods to verify whether the accuracy of capture performed by operators meets the pre-established levels and provide them with feedback for improvement.
* Manual data capture from paper questionnaires or scanned images is subject to keying errors. Incorporate online edits for error conditions that the data capture operator can correct (i.e. edits that will identify keying errors). Record these cases for later review and analysis. When feasible, the manual operation should be tested prior to conducting the survey.
* For automated data capture, ensure that the questionnaire is designed to ease the scanning and the intelligent character recognition.
* When automated capture is used, some questionnaires cannot be scanned, and others can be scanned but characters cannot be recognized. For damaged or badly scanned questionnaires, use a team of keyers to perform the capture.
* Systems for automated data capture by intelligent character recognition from scanned images should be tested prior to implementation. Such systems may cause relatively high rates of systematic errors in specific data items. It might be possible to improve the algorithms and their parameters to reduce the error rates. For the data items at high risk of systematic error, consider using keyers.
* Keyers should also be used to conduct a sample study assessment of the accuracy of automated capture. The results of such a study can be used to improve the process.
* Institute effective control of systems to ensure the security of data capture, transmission and handling, especially with new technologies such as cell phone and Internet data collection. Prevent loss of information and the resulting decline in quality, and potentially in credibility, due to system failures or human errors. Develop procedures for destroying the data when no longer needed.

*Notes:*

*Data collection is active: you are getting the data by asking question, sending a survey, or having people fill out questionnaire*

*Data capture is passive: you are recording data that is being generated by an activity you didn’t originate, for instance, data on visiting a website.*

***Q2: Explain the benefits of correctly interpreting data in an M&E process. (5 mrks)***

*KEY TERMS*

*Data Interpretation*

*Data interpretation refers to the implementation of processes through which data is reviewed for the purpose of arriving at an informed conclusion. The interpretation of data assigns a meaning to the information analyzed and determines its signification and implications.*

*The importance of data interpretation is evident, and this is why it needs to be done properly. Data is very likely to arrive from multiple sources and has a tendency to enter the analysis process with haphazard ordering. Data analysis tends to be extremely subjective. That is to say, the nature and goal of interpretation will vary from business to business, likely correlating to the type of data being analyzed. While there are several different types of processes that are implemented based on individual data nature, the two broadest and most common categories are “quantitative analysis” and “qualitative analysis”.*

*Yet, before any serious data interpretation inquiry can begin, it should be understood that visual presentations of data findings are irrelevant unless a sound decision is made regarding scales of measurement. Before any serious data analysis can begin, the scale of measurement must be decided for the data as this will have a long-term impact on data interpretation ROI. The varying scales include:*

* *Nominal Scale: non-numeric categories that cannot be ranked or compared quantitatively. Variables are exclusive and exhaustive.*
* *Ordinal Scale: exclusive categories that are exclusive and exhaustive but with a logical order. Quality ratings and agreement ratings are examples of ordinal scales (i.e., good, very good, fair, etc., OR agree, strongly agree, disagree, etc.).*
* *Interval: a measurement scale where data is grouped into categories with orderly and equal distances between the categories. There is always an arbitrary zero point.*
* *Ratio: contains features of all three.*
* *Once scales of measurement have been selected, it is time to select which of the two broad interpretation processes will best suit your data need.*

*When interpreting data, an analyst must try to discern the differences between correlation, causation and coincidences, as well as many other bias – but he also must consider all the factors involved that may have led to a result. There are various data interpretation methods one can use.*

*The interpretation of data is designed to help people make sense of numerical data that has been collected, analyzed and presented. Having a baseline method (or methods) for interpreting data will provide your analyst teams a structure and consistent foundation. Indeed, if several departments have different approaches to interpret the same data, while sharing the same goals, some mismatched objectives can result. Disparate methods will lead to duplicated efforts, inconsistent solutions, wasted energy and inevitably – time and money. In this part, we will look at the two main methods of interpretation of data: with a qualitative and a quantitative analysis.*

*Qualitative Data Interpretation*

*Qualitative data analysis can be summed up in one word – categorical. With qualitative analysis, data is not described through numerical values or patterns, but using descriptive context (i.e., text). Typically, narrative data is gathered by employing a wide variety of person-to-person techniques. These techniques include:*

* *Observations: detailing behavioral patterns that occur within an observation group. These patterns could be the amount of time spent in an activity, the type of activity and the method of communication employed.*
* *Documents: much like how patterns of behavior can be observed, different types of documentation resources can be coded and divided based on the type of material they contain.*
* *Interviews: one of the best collection methods for narrative data. Enquiry responses can be grouped by theme, topic or category. The interview approach allows for highly focused data segmentation.*

*A key difference between qualitative and quantitative analysis is clearly noticeable in the interpretation stage. Qualitative data, as it is widely open to interpretation, must be “coded” so as to facilitate the grouping and labeling of data into identifiable themes. As person-to-person data collection techniques can often result in disputes pertaining to proper analysis, qualitative data analysis is often summarized through*[*three basic principles*](ftp://ftp.qualisresearch.com/pub/qda.pdf)*: notice things, collect things, think about things.*

*Quantitative Data Interpretation*

*If quantitative data interpretation could be summed up in one word (and it really can’t) that word would be “numerical.” There are few certainties when it comes to data analysis, but you can be sure that if the research you are engaging in has no numbers involved, it is not quantitative research. Quantitative analysis refers to a set of processes by which numerical data is analyzed. Often, it involves the use of statistical modeling such as standard deviation, mean and median. Let’s quickly review the most common statistical terms:*

* *Mean: a mean represents a numerical average for a set of responses. When dealing with a data set (or multiple data sets), a mean will represent a central value of a specific set of numbers. It is the sum of the values divided by the number of values within the data set. Other terms that can be used to describe the concept are arithmetic mean, average and mathematical expectation.*
* *Standard deviation: this is another statistical term commonly appearing in quantitative analysis. Standard deviation reveals the distribution of the responses around the mean. It describes the degree of consistency within the responses; together with the mean, it provides insight into data sets.*
* *Frequency distribution: this is a measurement gauging the rate of a response appearance within a data set. When using a survey, for example, frequency distribution has the capability of determining the number of times a specific ordinal scale response appears (i.e., agree, strongly agree, disagree, etc.). Frequency distribution is extremely keen in determining the degree of consensus among data points.*

*Typically, quantitative data is measured by visually presenting correlation tests between two or more variables of significance. Different processes can be used together or separately, and comparisons can be made to ultimately arrive at a conclusion. Other signature interpretation processes of quantitative data include:*

* *Regression analysis*
* *Cohort analysis*
* *Predictive and prescriptive analysis*

*Now that we have seen how to interpret data, let’s move on and ask ourselves some questions: what is some data interpretation benefits? Why do all industries engage in data research and analysis? These are basic questions, but that often don’t receive adequate attention****.***

**THE BENEFIT OF CORRECTLY ENTERPRETING DATA**

The purpose of collection and interpretation is to acquire useful and usable information and to make the most informed decisions possible. data collection and interpretation provide limitless benefits for a wide range of institutions and individuals.

Data analysis and interpretation, regardless of method and qualitative/quantitative status, may include the following characteristics:

* Data identification and explanation
* Comparing and contrasting of data
* Identification of data outliers
* Future predictions

Data analysis and interpretation, in the end, helps improve processes and identify problems. It is difficult to grow and make dependable improvements without, at the very least, minimal data collection and interpretation. What is the key word? Dependable. Vague ideas regarding performance enhancement exist within all institutions and industries. Yet, without proper research and analysis, an idea is likely to remain in a stagnant state forever (i.e., minimal growth). So… what are a few of the business benefits of digital age data analysis and interpretation? Let’s take a look!

**1) Informed decision-making:** A decision is only as good as the knowledge that formed it. Informed data decision making has the potential to set industry leaders apart from the rest of the organization pack. [Studies have shown](https://hbr.org/2012/10/big-data-the-management-revolution/ar) that organization in the top third of their projects are, on average of informed data decision-making processes. Most decisive actions will arise only after a problem has been identified or a goal defined. Data analysis should include identification, thesis development and data collection followed by data communication.

If organization only follow that simple order, one that we should all be familiar with from grade school science fairs, then they will be able to solve issues as they emerge in real time. Informed decision making tends to be cyclical. This means there is really no end, and eventually, new questions and conditions arise within the process that need to be studied further. The monitoring of data results will inevitably return the process to the start with new data and sights.

**2) Anticipating needs with trends identification:**data insights provide knowledge, and knowledge is power. The insights obtained from market and consumer data analyses can set trends for peers within similar organization. A perfect example of how data analysis can impact trend prediction can be evidenced in the music identification application, [Shazam](https://www.theguardian.com/technology/datablog/2013/dec/10/shazam-big-data-prediction-breakthrough-music-artists). The application allows users to upload an audio clip of a song they like but can’t seem to identify. Users make 15 million song identifications a day. With this data, Shazam has been instrumental in predicting future popular artists.

When industry trends are identified, they can then serve a greater industry purpose. For example, the insights from Shazam’s monitoring benefits not only Shazam in understanding how to meet consumer needs, but it grants music executives and record label companies an insight into the pop-culture scene of the day. Data gathering and interpretation processes can allow for industry-wide climate prediction and result in greater revenue streams across the market. For this reason, all institutions should follow the basic data cycle of collection, interpretation, decision making and monitoring.

**3) Cost efficiency:** Proper implementation of data analysis processes can provide businesses with profound cost advantages within their industries. A recent data study performed by [Deloitte](http://www2.deloitte.com/content/dam/Deloitte/global/Documents/Deloitte-Analytics/dttl-analytics-analytics-advantage-report-061913.pdf) vividly demonstrates this in finding that data analysis ROI is driven by efficient cost reductions. Often, this benefit is overlooked because making money is typically viewed as “sexier” than saving money. Yet, sound data analyses have the ability to alert management to cost-reduction opportunities without any significant exertion of effort on the part of human capital.

A great example of the potential for cost efficiency through data analysis is Intel. Prior to 2012, Intel would conduct over 19,000 manufacturing function tests on their chips before they could be deemed acceptable for release. To cut costs and reduce test time, Intel implemented predictive data analyses. By using historic and current data, Intel now avoids testing each chip 19,000 times by focusing on specific and individual chip tests. After its implementation in 2012, Intel saved over [$3 million in manufacturing costs](http://www.informationweek.com/software/information-management/intel-cuts-manufacturing-costs-with-big-data/d/d-id/1109111). Cost reduction may not be as “sexy” as data profit, but as Intel proves, it is a benefit of data analysis that should not be neglected.

**4) Clear foresight:**organization that collect and analyze their data gain better knowledge about themselves, their processes and performance. They can identify performance challenges when they arise and take action to overcome them. Data interpretation through visual representations lets them process their findings faster and make better-informed decisions on the future

**To Conclude…**

The importance of data interpretation is undeniable. Dashboards not only bridge the information gap between traditional data interpretation methods and technology, but they can help remedy and prevent the major pitfalls of interpretation. As a digital age solution, they combine the best of the past and the present to allow for informed decision making with maximum data interpretation.

## **Q3.** Explain the main concerns for a data analyst while undertaking the task of data analysis. (10 mrks)

## 

## **Main concern for data analyst while undertaking the task of data analysis**

It is important for an organization to hire a data analyst having skills that are varied as the job of a data analyst is multidisciplinary data analysis

In this digitalized world, we are producing a huge amount of data in every minute. The amount of data produced in every minute makes it challenging to store, manage, utilize, and analyze it. Even large organization are struggling to find out the ways to make this huge amount of data useful. Today, the amount of data produced by large organization is growing, as mentioned before, at a rate of 40 to 60% per year. Simply storing this huge amount of data is not going to be all that useful and this is the reason why organizations are looking at options like data lakes and big data analysis tools that can help them in handling big data to a great extent. Now, let’s take a quick look at some challenges faced in Data analysis:

**1. Need for Synchronization Across Disparate Data Sources**

As data sets are becoming bigger and more diverse, there is a big challenge to incorporate them into an analytical platform. If this is overlooked, it will create gaps and lead to wrong messages and insights.

**2. Acute Shortage of Professionals Who Understand Data Analysis**

The analysis of data is important to make this voluminous amount of data being produced in every minute, useful. With the exponential rise of data, a huge demand for big data scientists and Big Data analysts has been. Another major challenge faced by the organization is the shortage of professionals who understand Data analysis. There is a sharp shortage of data scientists in comparison to the massive amount of data being produced.

**3. Getting Meaningful Insights Through the Use Of Big Data Analytics**

It is imperative for organizations to gain important insights from Data analytics, and it is important that only the relevant department has access to this information. A big challenge faced by the organization in the Data analytics is mending this wide gap in an effective manner.

**4. Getting Voluminous Data Into The Data Platform**

It is hardly surprising that data is growing with every passing day. This simply indicates that organizations need to handle a large amount of data on daily basis. The amount and variety of data available these days can overwhelm any data engineer and that is why it is considered vital to make data accessibility easy and convenient for brand owners and managers.

**5. Uncertainty of Data Management Landscape**

With the rise of Big Data, new technologies and organization are being developed every day. However, a big challenge faced by the companies in the Big Data analytics is to find out which technology will be best suited to them without the introduction of new problems and potential risks.

**6. Data Storage and Quality**

organizations are growing at a rapid pace. With the tremendous growth of the large organizations, increases the amount of data produced. The storage of this massive amount of data is becoming a real challenge for everyone. Popular data storage options like data lakes/ warehouses are commonly used to gather and store large quantities of unstructured and structured data in its native format. The real problem arises when a data lakes/ warehouse tries to combine unstructured and inconsistent data from diverse sources, it encounters errors. Missing data, inconsistent data, logic conflicts, and duplicates data all result in data quality challenges.

**7. Security and Privacy of Data**

Once an organization discover how to use Data, it brings them a wide range of possibilities and opportunities. However, it also involves the potential risks associated with data when it comes to the privacy and the security of the data. The Data tools used for analysis and storage utilizes the data disparate sources. This eventually leads to a high risk of exposure of the data, making it vulnerable. Thus, the rise of voluminous amount of data increases privacy and security concerns.

To overcome these Data challenges in the large organizations, a corporate training program in Data should be organized by the project directors and managers

**Q4.** Describe key measures that are mandatory for data quality assurance at program level and explain the value of data quality assurance. (15 mrks)

.

**Illustrates that data possess 5 key high-quality attributes**

* **Validity**
* **Reliability**
* **Precision**
* **Integrity**
* **Timeliness**

**Validity**

* **Face validity-data must be true representations of the indicator of interest, and the indicator must be a valid measure of the result.**
* **Attribution. Changes in the indicator can be plausibly associated with organization interventions.**
* **Measurement Error. Sampling and non-sampling errors.**

**Reliability**

**Data are considered reliable if the methods by which they are collected and analyzed remain stable over time.**

**Are data collection and analysis methods documented in writing and being used to ensure the same procedures are being followed each time?**

**Precision**

**• Precise data have enough detail to present a fair picture of what is actually happening.**

* **Is the margin of error reported along with the data? – Limited biases**
* **Is the data collection method/tool being used to collect the data fine-tuned or exact enough to register the expected change? (e.g. A yardstick may not be precise enough tool to measure a change of a few millimeters)**

**Integrity**

* **Data that have integrity are protected by a system that reduces the possibility of bias (either by transcription error or deliberate manipulation) Are procedures or safeguards in place to minimize data transcription errors?**
* **Is there independence in key data collection, management, and assessment procedures?**
* **Are mechanisms in place to prevent unauthorized changes to the data?**

**Timeliness**

* **Are data available frequently enough to inform program management decisions?**
* **Are the data reporting the most current practically available?**
* **Are the data reported as soon as possible after being collected?**

**What is a Data Quality Assurance?**

* **Outlines strategies in the routine monitoring system to reduce:**
* **Estimation error and bias**
* **Measurement error and bias**
* **Transcription errors**
* **Data processing error**
* **Describes how/ when internal data quality assessments will be implemented**

**Recipe for successful data quality assurance**

**A. Adequate staff capacity, supervision and accountability**

**B. Complete documentation of processes/ protocols readily available to collectors and processors**

**C. Routine cross checking-mechanisms**

**D. Clear strategy to respond to problems**

**E. Adequate financial and logistical resources to ensure timely performance**

**Essential elements of a data quality assurance**

* 1. **Description of staff capacity, supervision and accountability.**
  2. **Number and qualifications of staff**
  3. **Job descriptions**
  4. **Time allocated to M&E responsibilities**
  5. **Nature and timing of relevant training**
  6. **Provisions for corrective action and accountability**

**Essential elements of a data quality assurance**

**B. Complete documentation of processes and protocols for:**

* **Data collection**
* **Language-appropriate**
* **Clearly defined procedures**
* **Data cleaning**
* **Recording**
* **Aggregation**

**Essential elements of a data quality assurance**

**B. Documentation of processes and protocols for:**

* **Data access**
* **Safeguarding data**
* **Reporting**
* **Regular verification of consistency and compliance with methods and protocols**

**• Data management and safeguard**

**Essential elements of a data quality assurance**

**C. Description of processes for routine crosschecking and verification**

* **What are some effective methods?**
* **Supervisors or M&E officers visit small sample of HH, farmers, mothers, etc.**
* **Systematic review of collected data to compare values collected across time and location**
* **flag outliers**
* **Reasonability checks and comparisons in data entry / processing software**

**Essential elements of a data quality assurance**

**D. Description of the strategy to respond to data quality problems**

* **Quality issues identified during routine cross-checks**
* **Limitations identified during a data quality assessment.**

**Essential elements of a data quality assurance**

**E. Description of financial resources and logistical support to assure timely performance. e.g., for:**

* **Travel**
* **Training**
* **Procurement / reproduction of instruments and tools**

**What must be included in the Data Management and Safeguard Plan?**

* **Database entry procedures**
* **Data management protocol**
* **Data management coordination across partners (for consortium/ partnerships).**

**Q5:** In about 350 words, describe the main challenges to effective data interpretation and analysis. (10 mrsk)

A key challenge is to help users develop a consistent model when they explore data. Actions that cause changes in the visualization of data (for example applying filter to remove points and instantly relocate others as axis scales adjust) tend to disrupt user’s perception of the scene and affect recall and require more attention

* Moreover, users make incremental progress while analyzing. More

often than not, most data exploration systems do not reflect a trace of their past action which

have led them to the current state. Users may not remember the steps that lead to them

current position on data. Another challenge while data scales up is to help users develop models

for information that may surpass their working memory limits. As data scales up, another

challenge is to help users model information that may exceed their working memory limits. For

instance, a user attempting to track a relationship that spans three or four dimensions may have to compare between three or four different charts to find it, due to interface restrictions. This

could get taxing and difficult to manage, even for an expert.

Going by the fact that sensemaking performance is influenced by motivation and efficacy

* interfaces should encourage exploration and extensive data digging. Complex interfaces

may require long periods of training and might be overwhelming and intimidating for new

comers

* Current research provides an insight into not only delivering easy and motivating

interfaces, but also solving the problems of enhancing recall and modeling mentioned in the

previous paragraph. Users are already well trained for interaction with objects in the real world.

By designing exploratory visualization tools that use physics and naturalism to match users’ real

life experiences, fluidly move between analysis steps (as real objects do), and utilize touch

modalities that bind users and their interface actions

* exploratory performance can be improved

The oft-repeated mantra of those who fear data advancements in the digital age is “big data equals big trouble.” While that statement is not accurate, it is safe to say that certain data interpretation problems or “pitfalls” exist and can occur when analyzing data, especially at the speed of thought. Let’s identify three of the most common data misinterpretation risks and shed some light on how they can be avoided:

**1) Correlation mistaken for causation:** our first misinterpretation of data refers to the tendency of data analysts to mix the cause of a phenomenon with correlation. It is the assumption that because two actions occurred together, one caused the other. This is not accurate as actions can occur together absent a cause and effect relationship.

* Digital age example:if increased revenue is the result of increased social media followers… there might a definitive correlation between the two, especially with today’s multi-channel purchasing experiences. But that does not mean an increase in followers is the direct cause of increased revenue. There could be both a common cause or an indirect causality.
* Remedy: attempt to eliminate the variable you believe to be causing the phenomenon.

**2) Confirmation bias:**our second data interpretation problem occurs when you have a theory or hypothesis in mind but are intent on only discovering data patterns that provide support, while rejecting those that do not.

* Digital age example: your boss asks you to analyze the success of a recent multi-platform social media marketing campaign. While analyzing the potential data variables from the campaign (one that you ran and believe performed well), you see that the share rate for Facebook posts were great, while the share rate for Twitter Tweets were not. Using only the Facebook posts to prove your hypothesis that the campaign was successful would be a perfect manifestation of confirmation bias.
* Remedy:as this pitfall is often based on subjective desires, one remedy would be to analyze data with a team of objective individuals. If this is not possible, another solution is to resist the urge to make a conclusion before data exploration has been completed. Remember to always try to disprove a hypothesis, not prove it.

**3) Irrelevant data:** the third and final data misinterpretation pitfall is especially important in the digital age. As large data is no longer centrally stored, and as it continues to be analyzed at the speed of thought, it is inevitable that analysts will focus on data that is irrelevant to the problem they are trying to correct.

* Digital age example:in attempting to gauge the success of an email lead generation campaign, you notice that the number of homepage views directly resulting from the campaign increased, but the number of monthly newsletter subscribers did not. Based on the number of homepage views, you decide the campaign was a success when really it generated zero leads.
* Remedy:proactively and clearly frame any data analysis variables and KPIs prior to engaging in a data review. If the metric you are using to measure the success of a lead generation campaign is newsletter subscribers, there is no need to review the number of homepage visits. Be sure to focus on the data variable that answers your question or solves your problem and not on irrelevant data.

* challenges arising from effective analysis and interpretation of large data sets have made the data analyst’s job more challenging than ever. Although data analysts now have access to a variety of statistical and AI tools capable of performing different aspects of data analysis and interpretation, they certainly need further support. For example, they need competent interactive tools for exploring complex real-world data sets; they need powerful graphical techniques for visualizing multi-dimensional data; they need new computational tools for controlling data quality; they need effective means of summarizing large data sets into convenient and relevant forms for analysis and interpretation ; they need data sampling methods with a minimum amount of bias; they need intelligent search methods which will find the most interesting structures; and they need more integrated and "friendly" data analysis environments where "boring aspects" of the analyst’s job may be kept to the minimum so that interesting aspects of the job can be focused on. Of course, many of these needs were there before large data sets came into the picture. However, searching for interesting structures in large data sets has called for better ways of doing these things and suggested new areas for research. Take data quality control as an example. We begin to see "data cleaning" companies being set up to respond to the need to automate (at least partially) the laborious process of managing imperfect data as this often consumes a significant amount of resources. Outliers, or outlying observations, are particularly difficult to handle as some of them are measurement errors, while others may represent something "significant" from the viewpoint of the application domain. In this situation an outright rejection of outliers based on some statistical tests is normally not a very good idea. So, a careful analysis of outliers by data analysts is normally desirable. However, if there are many of them (this is often the case when dealing with large data sets), manual analysis may become insufficient. Some preliminary work has been performed to explore the possibility of automating certain aspects of outlier analysis (Liu et al. 1994). Apart from the need to research into the problems and opportunities brought about by the analysis of large data sets, it is of strategic importance to obtain a deep understanding of how data analysts carry out their task and to see how their analysis procedures may be encoded, extended, or improved. The rapid development of statistical and AI tools has made many aspects of data analysis routine, e.g. there is no longer the need to concern we with the mechanics of how to find a structure in a data set (Hand 1996). Instead we can concentrate on considering the "high-level" issues such as what kind of structures should be sought, what questions should we be asking, what would be the most appropriate method of analysis, and how the results should be interpreted. We can find out how the analyst’s knowledge and strategies can be most effectively captured in IDA tools and applications. We can study the strengths and weaknesses of numerous computational techniques and their potential contributions to the various stages of the data analysis process. We can see how these techniques may be most appropriately used and how we may assist data analysts in their quest to uncover interesting and useful structures from large data sets. In IDA-95, the Intelligent Data Analysis community demonstrated great enthusiasm and resourcefulness in responding to the above challenges by presenting many interesting pieces of research. For example, much interesting work on data exploration was reported; many useful real world IDA applications were described; some fundamental research problems were addressed; and we also begin to see some papers (e.g. St Am ant and Cohen), but not many, on developing intelligent assistants for data analysts. These papers were concerned with human-machine data analysis, and asked: what kind of intelligence is needed to make this type of analysis work effectively? Although IDA-95 was successful in bringing together people with different backgrounds to discuss important issues in intelligent data analysis, there is much to be desired. First, some people presented essentially black-box techniques. The boxes might be intelligent, but it was unclear how these boxes may contribute to the overall data analysis process. Second, there were not many papers on the effective integration of AI and statistical techniques to approach data analysis tasks. For example, papers on statistical analysis tend to have little AI content, whereas AI approaches tend to talk about machine learning and knowledge discovery without referring much to specific data analysis issues. Third, we need more work addressing issues arising from analyzing very large data sets. Last, but not least, IDA-95 was less focused than it should have been. This problem might be addressed in subsequent IDA symposia because the potential areas for IDA are very large and trying to address every possible issue in any given symposium could lead to the possibility of only scratching surfaces and not making serious progress. A survey after IDA-95 supported the idea of making IDA a regular, biennial conference. IDA-97, the second in the series, will be focusing on "Reasoning about Data". We are interested in intelligent systems that reason about how to analyze data, perhaps as human analysts do. Analysts often bring exogenous knowledge about data to bear when they decide how to analyze it; they use intermediate results to decide how to proceed; they reason about how much analysis the data will actually support; they consider which methods will be most informative; they decide which aspects of a model are most - 8 - uncertain and focus attention there; they sometimes have the luxury of collecting more data, and plan to do so efficiently. In short, there is a strategic aspect to data analysis, beyond the tactical choice of this or that test, visualization or variable.

**Reference**

Brodley, C.E. and Smyth, P. (1996) "Applying Classification Algorithms in Practice", Statistics and Computing, (In Press). Chatfield, C. (1988) "Problem Solving: a Statistician’s Guide", London: Chapman and Hall. Cohen, P., (1995) "Empirical Methods for Artificial Intelligence", Cambridge: MIT. Elder IV, J. and Pregibon, D. (1996) "A Statistical Perspective on Knowledge Discovery in Databases", in U M Fayyad, G Piatetsky-Shapiro, P Smyth and R Uthurusamy (eds): Advances in Knowledge Discovery and Data Mining, AAAI/MIT. Hand, D.J., (1996) "Intelligent Data Analysis and Deep Understanding", Proc. of Intelligent Data Management 96, 26-39, Unicom, London. Liu, X., Cheng, G., and Wu, J. (1994) "Noise and Uncertainty Management in Intelligent Data Modeling", Proc. of AAAI-94, 263-8, Seattle, Washington. Michie, D., Spiegelhalter, D. J., and Taylor, C.C. (eds) (1994) "Machine Learning, Neural and Statistical Classification", London: Ellis Horwood. Nakhaeizadeh, G. (1995) "What Daimler-Benz Has Learned as an Industrial Partner from the Machine Learning Project StatLog?", Proc. of the Workshop on Applying Machine Learning in Practice, 22-26. Piatetsky-Shapiro, G., Brachman, R., Khabaza, T., Kloesgen, W., Simoudis, E. (1996) "An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications", in E Simoudis, J Han and U Fayyad, (eds). Proc, of the Second International conference on Knowledge Discovery and Data Mining", AAAI Press. Tukey, J. W. (1977). "Exploratory Data Analysis", Addison-Wesley. Weiss, S.M. and Kulikowski, C.A. (1991) "Computer Systems that Learn", California: Morgan Kaufmann.

Auerbach, C. F., & Silverstein, L. B. (2003). Qualitative data: An introduction to coding and analysis. New York: New York University. Bybee, R. W. (2011). Scientific and engineering practices in K-12 classrooms: Understanding A Framework for K-12 Science Education. The Science Teacher, 78(9), 34–40. Committee on Integrated STEM Education, National Academy of Engineering, & National Research Council. (2014). STEM integration in K-12 education: Status, prospects, and an agenda for research. In M. Honey, G. Pearson, & H. Schweingruber (Eds.). Washington, DC: National Academies Press. Cramer, K., Ahrendt, S., Monson, D., Wyberg, T., & Miller, C. (2017). Making sense of third-grade students’ misunderstandings of the number line. Investigations in Mathematics Learning, 9(1), 19–37. Gal, I. (2004). Statistical literacy: Meanings, components, responsibilities. In D. Ben-Zvi & J. Garfield (Eds.), The Challenge of Developing Statistical Literacy, Reasoning and Thinking (pp. 47–78). New York: Kluwer Academic Publishers. Garfield, J. (1995). How students learn statistics. International Statistical Review, 63(1), 25–34. Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. International Statistical Review, 75(3), 372–396. http://doi.org/10. 1111/j.l751-5823.2007.00029.x Garfield, J., Ben-Zvi, D., Chance, B., Medina, E., Roseth, C., & Zieffler, A. (2008). Research on teaching and learning statistics. In Developing students’ statistical reasoning: Connecting research and teaching practice (Vol. 75, pp. 21–43). Springer. Hjalmarson, M. A., Moore, T. J., & delMas, R. (2011). Statistical analysis when the data is an image: Eliciting student thinking about sampling and variability. Statistics Education Research Journal, 10(1), 15–34. Hofstein, A., & Lunetta, V. N. (2004). The laboratory in science education: Foundations for the twenty-first century. Science Education, 88(1), 28–54. http://doi.org/10.1002/sce.10106 Kukliansky, I., & Eshach, H. (2013). Evaluating a contextual-based course on data analysis for the physics laboratory. Journal of Science Education and Technology, 23(1), 108–115. http://doi.org/10.1007/ s10956-013-9456-6 Lehrer, R., Kim, M.-J., & Jones, R. S. (2011). Developing conceptions of statistics by designing measures of distribution. ZDM: The International Journal for Mathematics Education, 43(5), 723–736. http:// doi.org/10.1007/s11858-011-0347-0 Lohr, S. (2012, February 11). The age of big data. The New York Times. New York. Moore, D. S. (1990). Uncertainty. In L. A. Steen (Ed.), On the shoulders of giants: New approaches to numeracy. Washington, DC: National Academy Press. Moore, D. S. (1998). Statistics among the liberal arts. Journal of the American Statistical Association, 93(444), 1253–1259. National Council of Teachers of Mathematics. (1989). Curriculum and evaluation standards for school mathematics. Reston, VA: NCTM. NGSS Lead States. (2013). Next generation science standards: For states, by states. Washington, DC: The National Academies Press. Shaughnessy, J. M. (2006). Research on students’ understanding of some big concepts in statistics. In Thinking and reasoning with data and chance: Sixty-eighth yearbook (pp. 77–98). Reston, VA: National Council of Teachers of Mathematics. Watson, J. M. (2006). Statistical literacy at school: Growth and goals. Mahwah, NJ: Lawrence Erlbaum Associates. Wilkins, J. L. M. (2000). Preparing for the 21st century: The status of quantitative literacy in the United States. School Science and Mathematics, 100(8), 405–418. Yin, R. K. (2009). Case study research: Design and methods (Fourth). Los Angeles, CA: Sage Publications

Dazbrowski, M. J., M. Pilot, M. Kruczyk, M. Zmihorski, H. M. Umer, and J. Gliwicz. 2014. Reliability assessment of null allele detection: inconsistencies between and within different methods. Mol. Ecol. Resour. 14:361–373. Dawson, D. A., A. D. Ball, L. G. Spurgin, D. Martın-Galvez, I. R. K. Stewart, G. J. Horsburgh, et al. 2013. High-utility conserved avian microsatellite markers enable parentage and population studies across a wide range of species. BMC Genom. 14:176. De Mita, S., and M. Siol. 2012. EggLib: processing, analysis and simulation tools for population genetics and genomics. BMC Genet. 13:27. Defaveri, J., H. Viitaniemi, E. Leder, and J. Meril€a. 2013. Characterizing genic and nongenic molecular markers: comparison of microsatellites and SNPs. Mol. Ecol. Resour. 13:377–392. Dharmarajan, G., W. S. Beatty, and O. E. Rhodes. 2013. Heterozygote deficiencies caused by a Wahlund effect: dispelling unfounded expectations. J. Wildl. Manag. 77:226– 234. Di Rienzo, A., A. C. Peterson, J. C. Garza, A. M. Valdes, M. Slatkin, and N. B. Freimer. 1994. Mutational processes of simple-sequence repeat loci in human populations. Proc. Natl Acad. Sci. 91:3166–3170. Do, C., R. S. Waples, D. Peel, G. M. Macbeth, B. J. Tillett, and J. R. Ovenden. 2014. NeEstimator v2.0: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. Mol. Ecol. Resour. 14:209–214. Duchesne, P., and J. Turgeon. 2012. FLOCK provides reliable solutions to the “number of populations” problem. J. Hered. 103:734–743. Dufresne, F., M. Stift, R. Vergilino, and B. K. Mable. 2014. Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. Mol. Ecol. 23:40–69. Duran, C., R. Singhania, H. Raman, J. Batley, and D. Edwards. 2013. Predicting polymorphic EST-SSRs in silico. Mol. Ecol. Resour. 13:538–545. Epperson, B. K., B. H. McRae, K. Scribner, S. A. Cushman, M. S. Rosenberg, M.-J. Fortin, et al. 2010. Utility of computer simulations in landscape genetics. Mol. Ecol. 19:3549–3564. Eriksson, A., and A. Manica. 2011. Detecting and removing ascertainment bias in microsatellites from the HGDP-CEPH Panel. Genes Genom. Genet. 1:479–488. Eschbach, E., and S. Schoning. 2013. Identification of € high-resolution microsatellites without a priori knowledge of genotypes using a simple scoring approach. Methods Ecol. Evol. 4:1076–1082. Estoup, A., P. Jarne, and J.-M. Cornuet. 2002. Homoplasy and mutation model at microsatellite loci and their consequences for population genetic analysis. Mol. Ecol. 11:1591–1604.

Freedman, D. A. (1991). Statistical models and shoe leather. In P. V. Marsden (Ed.), Sociological methodology (Vol. 21). Washington, DC: Blackwell. Furlong, A., & Cartmel, F. (1997). Young people and social change: Individualization and risk in late modernity. Philadelphia: Open University Press. Gillies, V., Holland, J., & Ribbens McCarthy, J. (2003). Past/present/future: Time and the meaning of change in the “family.” In G. Allan & G. Jones (Eds.), Social relations and the life course. Houndmills, UK: Palgrave Macmillan. Goldthorpe, J. (2000). On sociology: Numbers, narratives and the integration of research and theory. Oxford, UK: Oxford University Press. Hakim, C. (1996). Key issues in women’s work: Female heterogeneity and the polarisation of women’s employment. London: Athlone Press. Hakim,C.(2000).Work–lifestylechoicesinthe21stcentury: Preferencetheory.NewYork:OxfordUniversityPress. Hammersley, M. (1992). What’s wrong with ethnography? London: Routledge. Hammond,C.(2005).Thewiderbenefitsofadultlearning: An illustration of the advantages of multimethod research. International Journal of Social Research Methodology, 8(3), 239–255. Hattery, A. (2001). Women, work and family: Balancing and weaving. Thousand Oaks, CA: Sage. Irwin, S. (1995). Rights of passage: Social change and the transition from youth to adulthood. London: UCL Press. Irwin, S. (1999). Resourcing the family: Gendered claims and obligations and issues of explanation. In E. B. Silva & C. C. Smart (Eds.), The new family? London: Sage. Irwin, S. (2004). Attitudes, care and commitment: Pattern and process. Sociological Research Online, 9(3). Retrieved from www.socresonline.org.uk/9/3/irwin.html Irwin, S. (2005). Reshaping social life. London: Routledge. Irwin, S., & Morris, L. (1993). Social security or economic insecurity? The concentration of unemployment (and research) within households. Journal of Social Policy, 22(3), 349–372. Joshi, H. (1984). Women’s participation in paid work: Further analysis of the Women and Employment Survey (Department of Employment RP 45). London: Department of Employment. Kelle, U. (2001). Sociological explanations between micro and macro and the integration of qualitative and quantitative methods. Forum: Qualitative Social Research, 2(1). Retrieved December 10, 2006, from qualitative-research.net/fqs/fqs-eng.ht