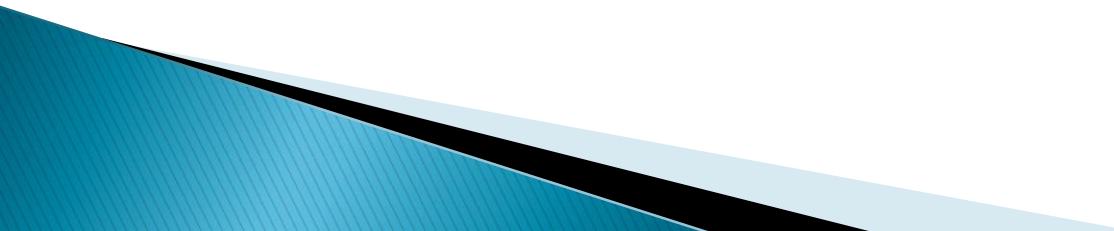
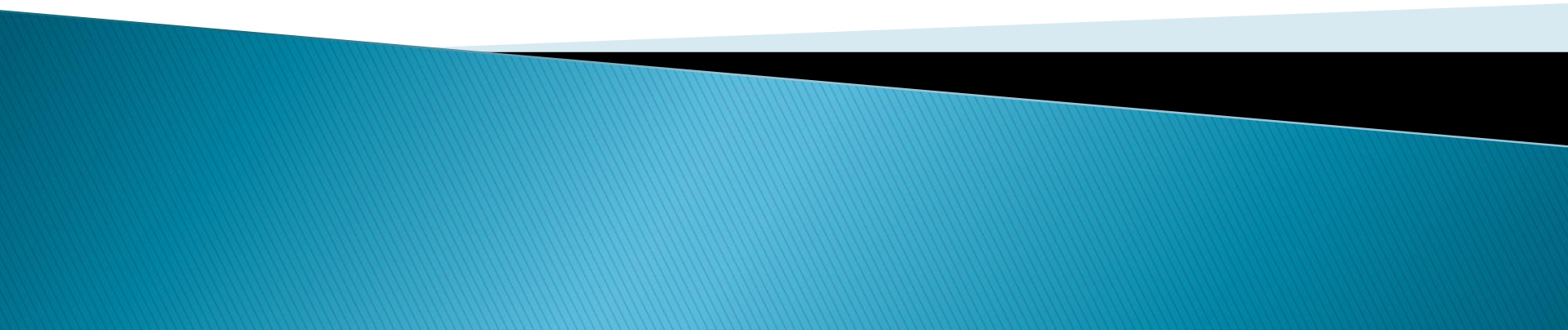


Week 4 agenda

- ▶ Questions from Students
 - ▶ Review of Regression
 - ▶ Iris Example.
 - ▶ Data Integration & collection
 - ▶ Data Pre-processing start
- 

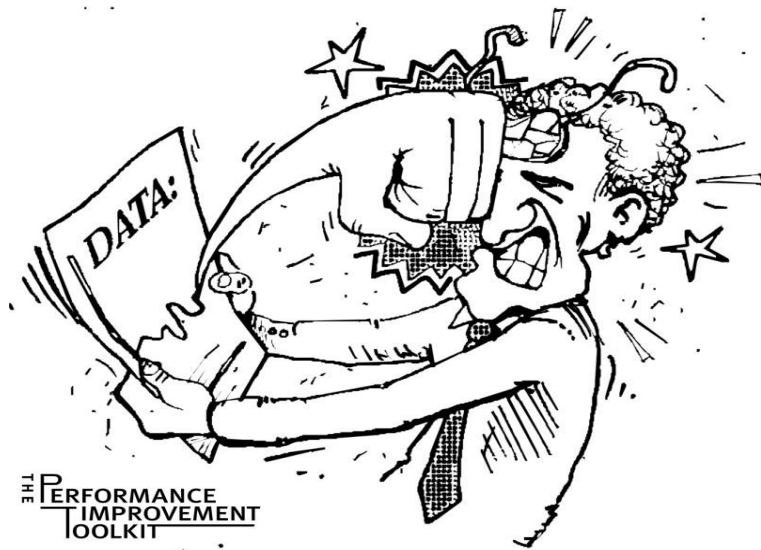
Lecture 4 Data Collection

Andrew McCarren



Why Should There be a Standard Process?

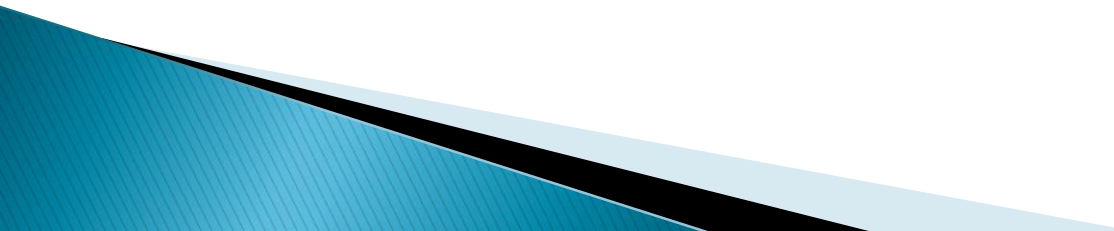
The data mining process must be **reliable and repeatable by people with little data mining background.**



Why Should There be a Standard Process?

- ▶ **Framework for recording experience**
 - Allows projects to be replicated
- ▶ **Aid to project planning and management**
- ▶ **“Comfort factor” for new adopters**
 - Demonstrates maturity of Data Mining
 - Reduces dependency on “stars”

Data Processing Standards

- CRISP (Cross industry Standard for Data Mining)
 - TDS (Team Data Science Process, Microsoft)
 - KDD (Knowledge Discovery in Databases)
- 

CRISP: Process Standardization

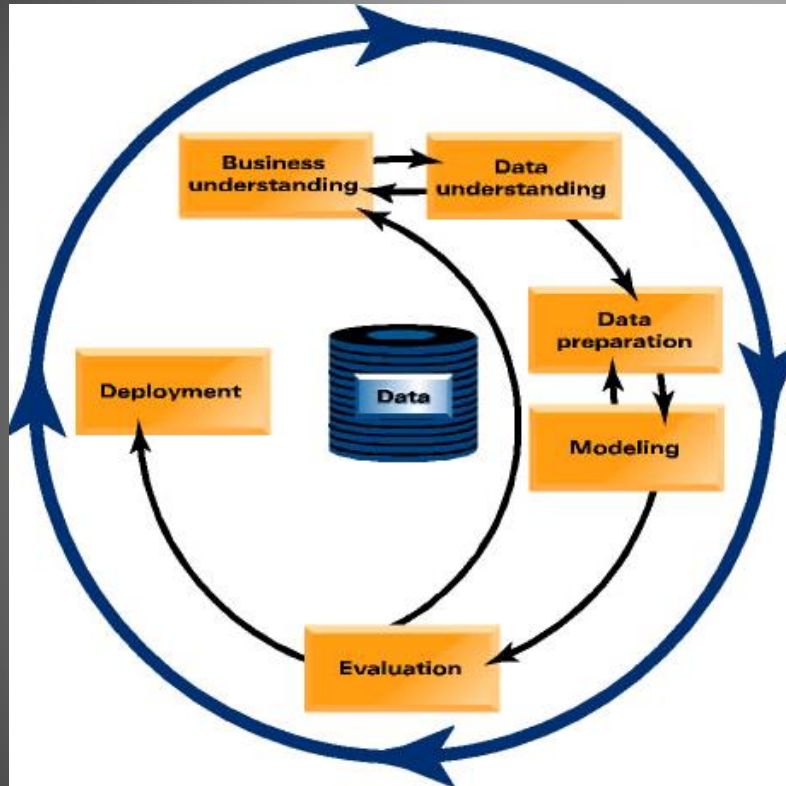
- **Initiative launched in late 1996 by three “veterans” of data mining market.**
 - Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) , NCR
- **Developed and refined through series of workshops** (from 1997-1999)
- **Over 300 organization contributed to the process model**
- **Published CRISP-DM 1.0** (1999)
- **Over 200 members of the CRISP-DM SIG worldwide**
 - - **DM Vendors** - SPSS, NCR, IBM, SAS, SGI, Data Distilleries, Syllogic, etc.
 - - **System Suppliers / consultants** - Cap Gemini, ICL Retail, Deloitte & Touche,
 - - **End Users** - BT, ABB, Lloyds Bank, AirTouch, Experian, etc.

CRISP-DM

- ▶ Non-proprietary
- ▶ Application/Industry neutral
- ▶ Tool neutral
- ▶ Focus on business issues
 - As well as technical analysis
- ▶ Framework for guidance
- ▶ Experience base
 - Templates for Analysis

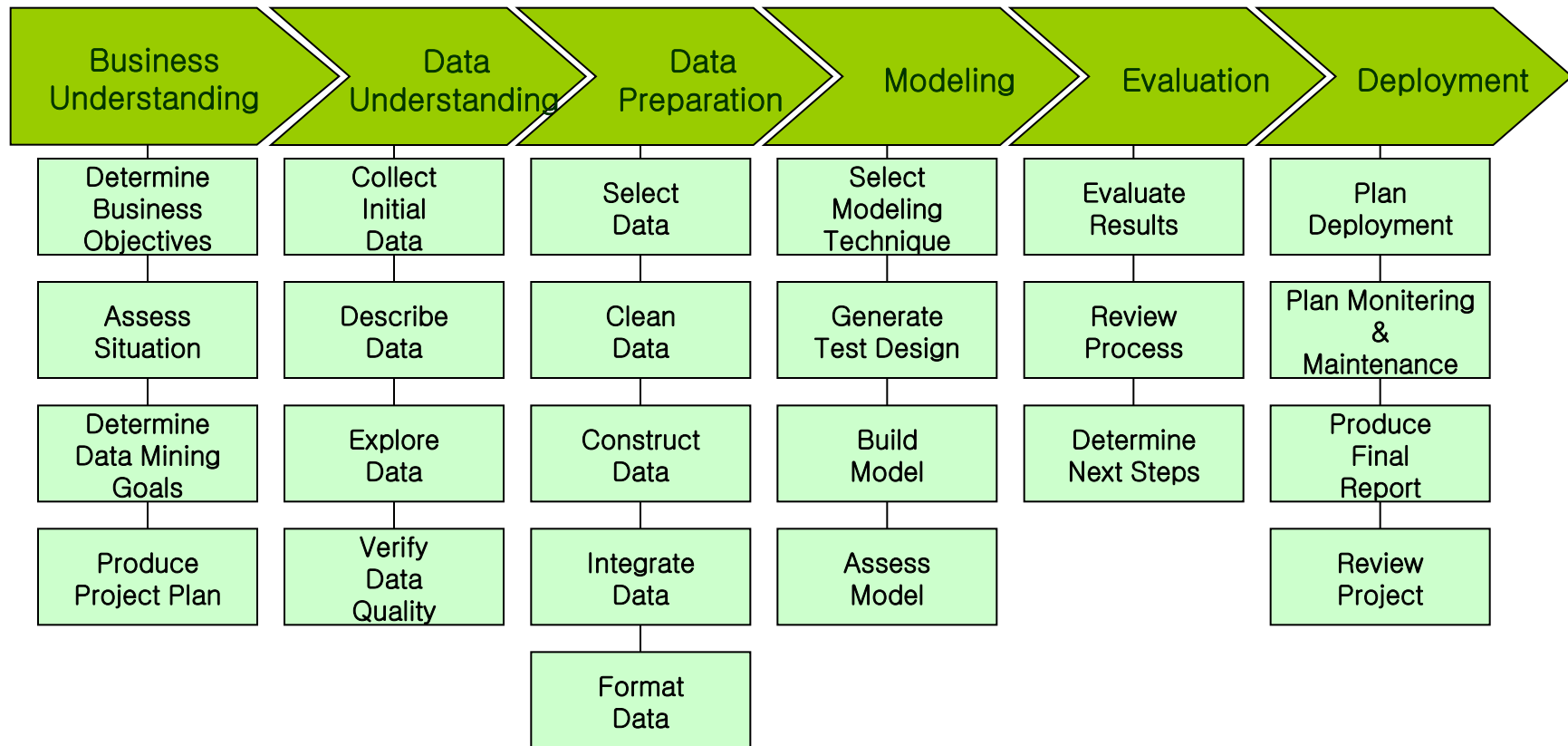


CRISP-DM: Overview

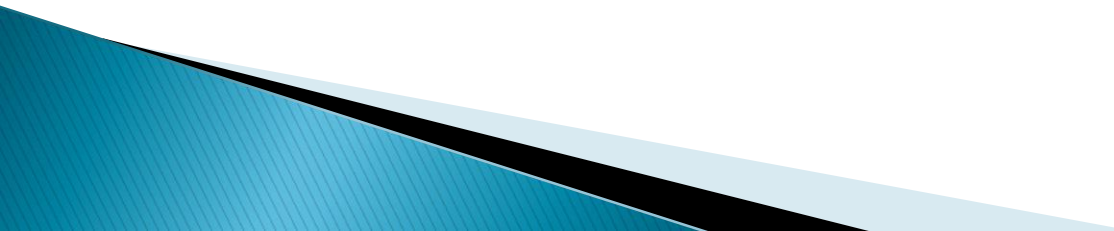



- ▶ Data Mining methodology
- ▶ Process Model for anyone
- ▶ Provides a complete blueprint
- ▶ Life cycle: 6 phases

Phases and Tasks




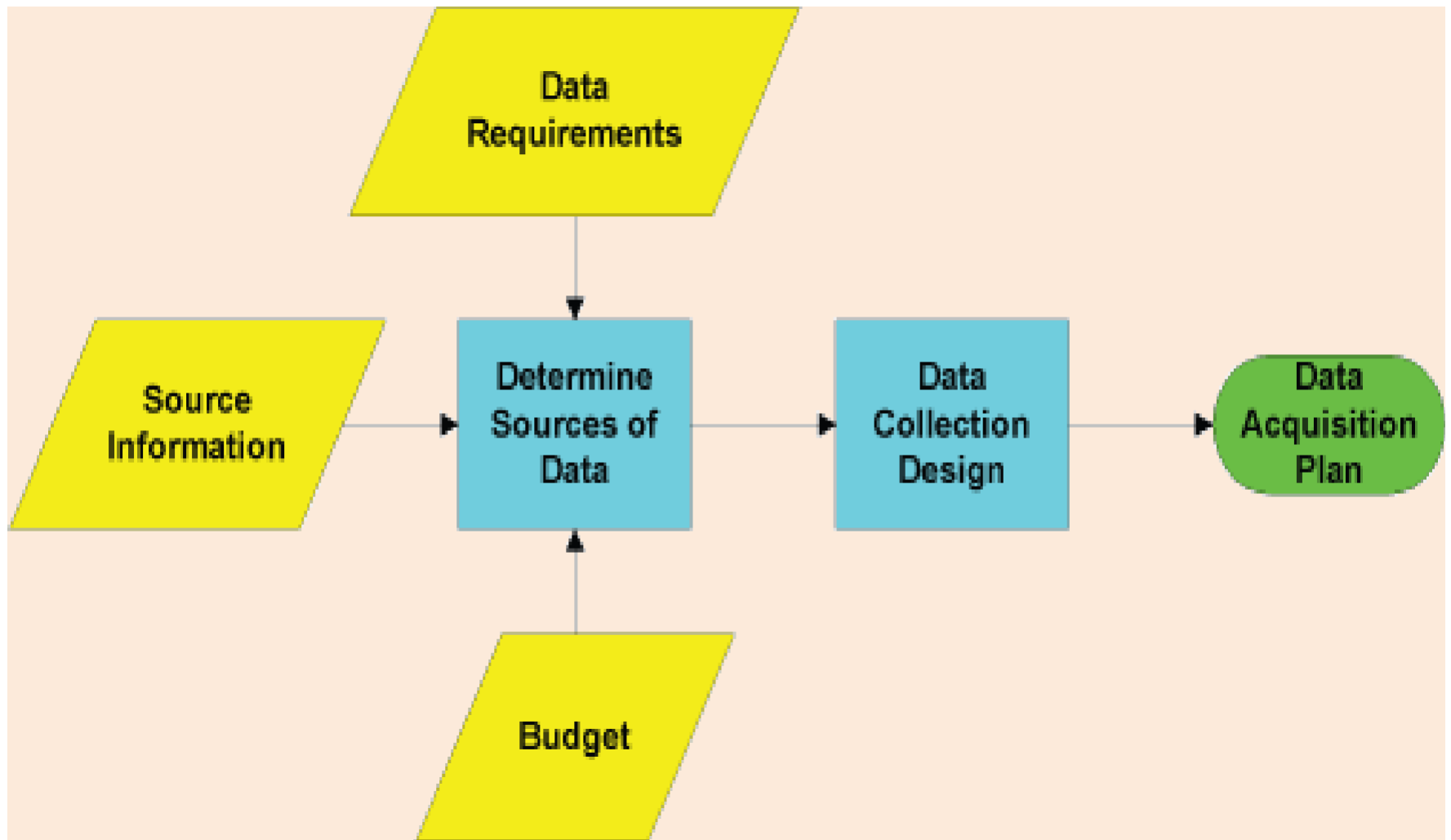
Introduction

- Data can be defined as the quantitative or qualitative value of a variable (e.g. number, images, words, figures, facts or ideas)
 - It is a lowest unit of information from which other measurements and analysis can be done.
 - Data is key to your analysis
- 

- Accurate and systematic data collection is critical to conducting scientific research.
 - Data collection allows us to collect information about study objects/subjects/participants.
 - Includes documents review, observation, questioning, measuring, or a combination of different methods.
- 

Data Collection plan

- Objectives and scope of the enquiry (research question).
 - Sources of information (type, accessibility).
 - Quantitative expression (measurement/scale).
 - Techniques of data collection.
 - Unit of collection.
- 



Sources of Data

```
graph TD; A[Sources of Data] --> B[External sources]; A --> C[Internal sources]; B --> D[Primary data]; B --> E[Secondary data];
```

External
sources

Internal
sources

Primary
data

Secondary
data

Internal vs. External Sources of Data

Internal

- Data that resides in the institution or company
- Routine surveillance, hospital records
- .

External

- Information collected from outside sources.
- This type of information can be collected by census or sampling.

I – Primary Data

- Data collected from **first-hand experience** is known as primary data. Can be more reliable, authentic. However this depends on how the data is collected.



Methods of collecting primary data

Direct Personal
Investigation
(interviewing)

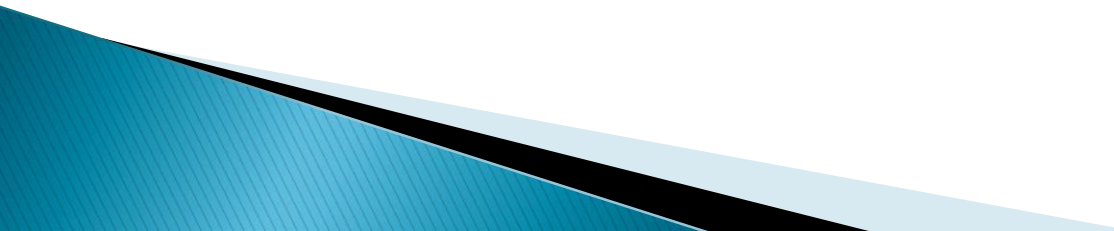
Indirect oral
investigation

Case studies

Measurements
Lab. results
Experimentation

Investigation
through
observation

II–Secondary Data

- ✓ Already been collected by others.
 - ✓ Journals, periodicals, research publication ,official record etc.
 - ✓ May be available in the published or unpublished form.
 - ✓ Resorted to when primary data sources/methods are infeasible -, inaccessible.
- 

Primary vs. secondary data

Primary data


- Real time
- Sure about the sources
- Can answer research question.
- Cost and time
- Can avoid bias
- More flexible

Secondary data

- Past data
- Not sure about sources
- Refining the research problem
- Cheap and no time
- Bias can't be ruled out
- Less flexible

Data Sources for Health Research

Primary or secondary sources

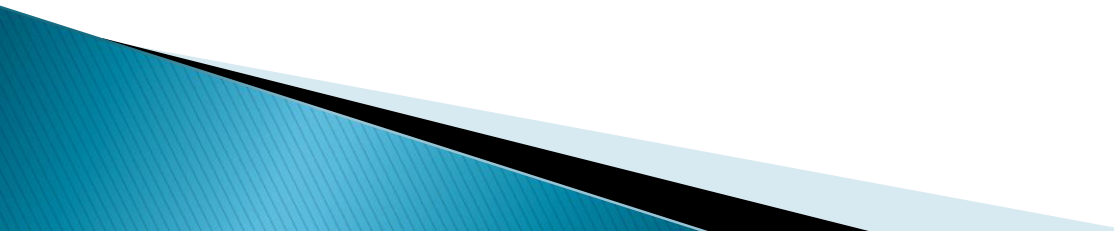
- Birth and death records
 - Medical records at physician offices, hospitals, nursing homes, etc.
 - Medical databases within various agencies, universities, and institutions.
 - Physical exams and laboratory testing
 - Diseases registries
 - Self-report measures: interviews and questionnaires
- 

Research Data: Considerations

Data collection vs. data analysis

- **Poor data** collection and management can render a perfectly executed trial **useless**
- Bad data practices carry **resource and ethical costs**
- Good practices:
 - **What is the data?**
 - **How is it represented?**
 - **How can it be stored for retrieval and use?**

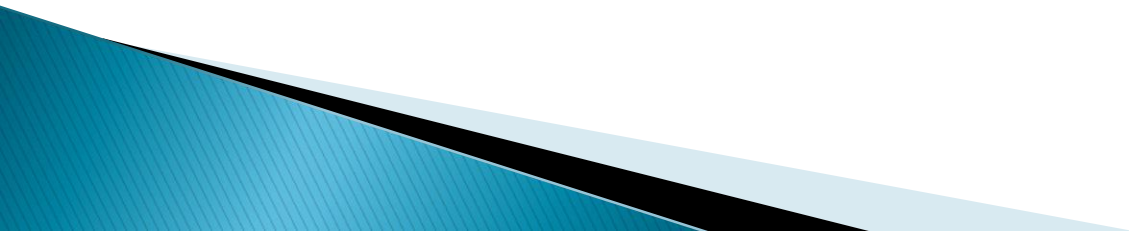
Considerations in collecting clinical data:

- Data (specimens) are all that remain after the active phase of a clinical trial
 - Data represents the objects and events in the trial
 - Understanding how the data is captured and recorded affects interpretation
 - Improper collection or interpretation lead to mistakes.
- 

Objectivity, Subjectivity and Reproducibility

Objectivity: the degree to which recorded data may be influenced by the individual thought of the observer

- Data reported by subjects (such as symptoms of a disease) can be objectively recorded (statements themselves are subjective)
- Objective observations (such as signs of a disease) can be made by outside observers

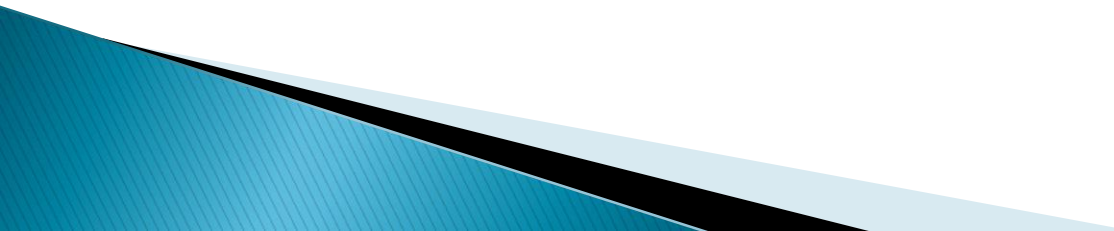


Objectivity is a process by which the observation is represented in the data

- Human observer is never completely free of influences
- **Controlling of subjectivity:**
 - Use an unbiased device to record information
 - Employ rating systems and train the observers
 - Consider limits to data precision
- **Reproducibility** – corroborating findings in a subsequent study requires knowing how observations were made (metadata)

The concept of METADATA

“Data about the data” (e.g., method of collection, relationship of data to the events in the research protocol, etc.)

- Temporal metadata require particular attention
 - Understand the implications of a time (e.g., if a blood specimen is drawn to measure a drug level, we must know the time that the specimen was drawn and the time the drug was administered)
 - Need to choose when to measure and how often
- 

Types of Data

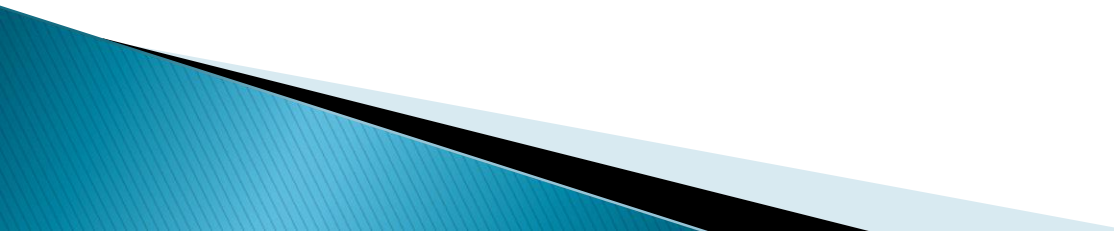
I- **Quantitative data** - measurements that can be manipulated mathematically

- **Precision** - body temperature, serum chloride , absolute eosinophil count.

II- **Qualitative data** - **conceptual entities** rather than numeric values (subject gender and race, signs and symptoms, diagnoses) May represent concepts that relate to quantitative data

III. **Ordinal data** look like numbers (e.g., urine protein measurements “0”, “1+”, “2+”, etc.)

IV. **Signal data** - quantitative in nature but are treated as qualitative (e.g., electrocardiogram tracings)



Data Sources

- IOT (Internet Of Things)
 - Sensors
 - Image data from Drone
 - Edge devices
 - Web Scrapping
 - Urllib2, BeautifulSoup
 - CRMs & ERPs
 - Salesforce, SAP etc
 - Business Data Warehouse
 - Usually internal.
- 