



# Data Quality: Data Quality – From Concept to Report

---

Putting nitrogen fixation to work for smallholder farmers in Africa

# Data Quality and its Dimensions

---



- Data Quality: Anything that alters/changes the ability of data to reflect the 'truth'
- Data Quality Dimension: Measurement or assessment of records, datasets, database, etc in order to understand the quality of data
- Multifaceted nature:
  - Potential problems await at all stages of the process (from design/planning of project to reporting)
  - Many types of data required and applications
  - Complex processes used and entities, complex and context-dependent
- ***The outputs of different data quality checks may be required in order to determine how well the data supports a particular business need***

# Dimensions of Data Quality Checks



## Reliability/Consistency

Extent to which data collection system is stable and consistent across time and space.  
measurement approach is same way every time

The extent to which indicators clearly and directly measure the results intended to be measured (does the data match the rules?)

## Validity

\*The degree to which data represent reality from the required point in time  
Frequency: how often are data collected?  
\*Currency: how recently have data been collected?  
\*Relevance: are data available frequently to support decisions?

## Timeliness

**Completeness:** How comprehensive the data is in relation to requirement (Are all data sets and data items recorded?)

**Conformity/Accuracy:** toeing the line, are data consistent with what is requested/observed?, can we match the data set across data stores?

# Example of Data Quality Assessment Process

---



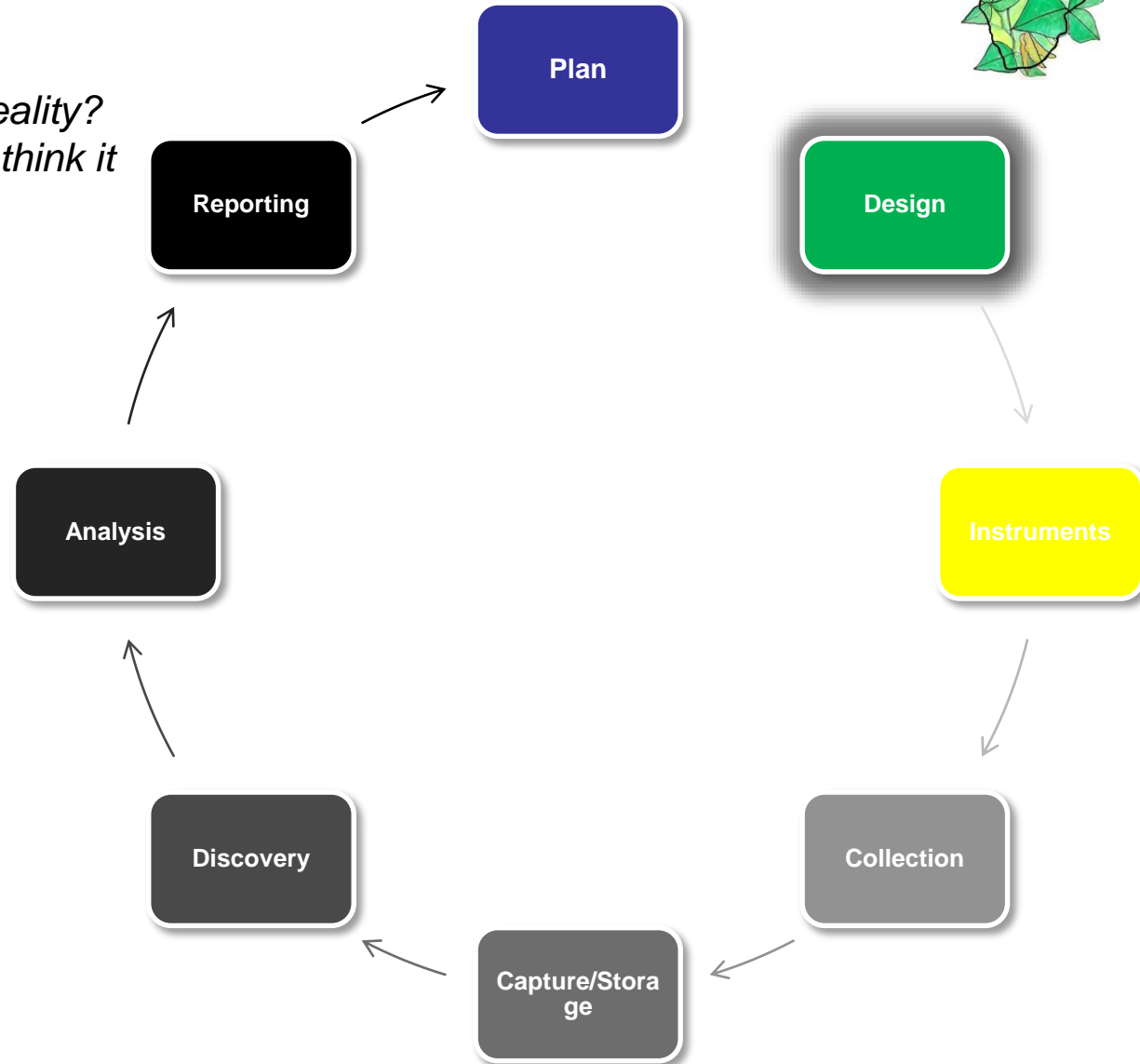
- Identify which data items need to be assessed for data quality, e.g. is data critical to project results (related directly to project indicators)
- Evaluate which data quality dimensions (e.g. completeness) to use and their related weighting (assign weights, e.g. 100% completeness)
- For each data quality dimension, define values or ranges representing excellent, good or bad quality data based on the weightings (e.g. 90% completeness means excellent data).
- Apply the assessment criteria to the data
- Analyze the results and determine if data quality is acceptable or not
- Identify in relation to the dimension, the possible source of the data quality issue
- Take corrective actions e.g. clean the data (this should be based on the sources of data quality issues) and improve data handling processes to prevent future recurrence
- Repeat the above on a periodic basis to monitor trends in Data Quality

# Sources of Data Quality Issues



## How Good is Your Data?

- *Does the data reflect current reality?*
- *Does the data mean what you think it does?*

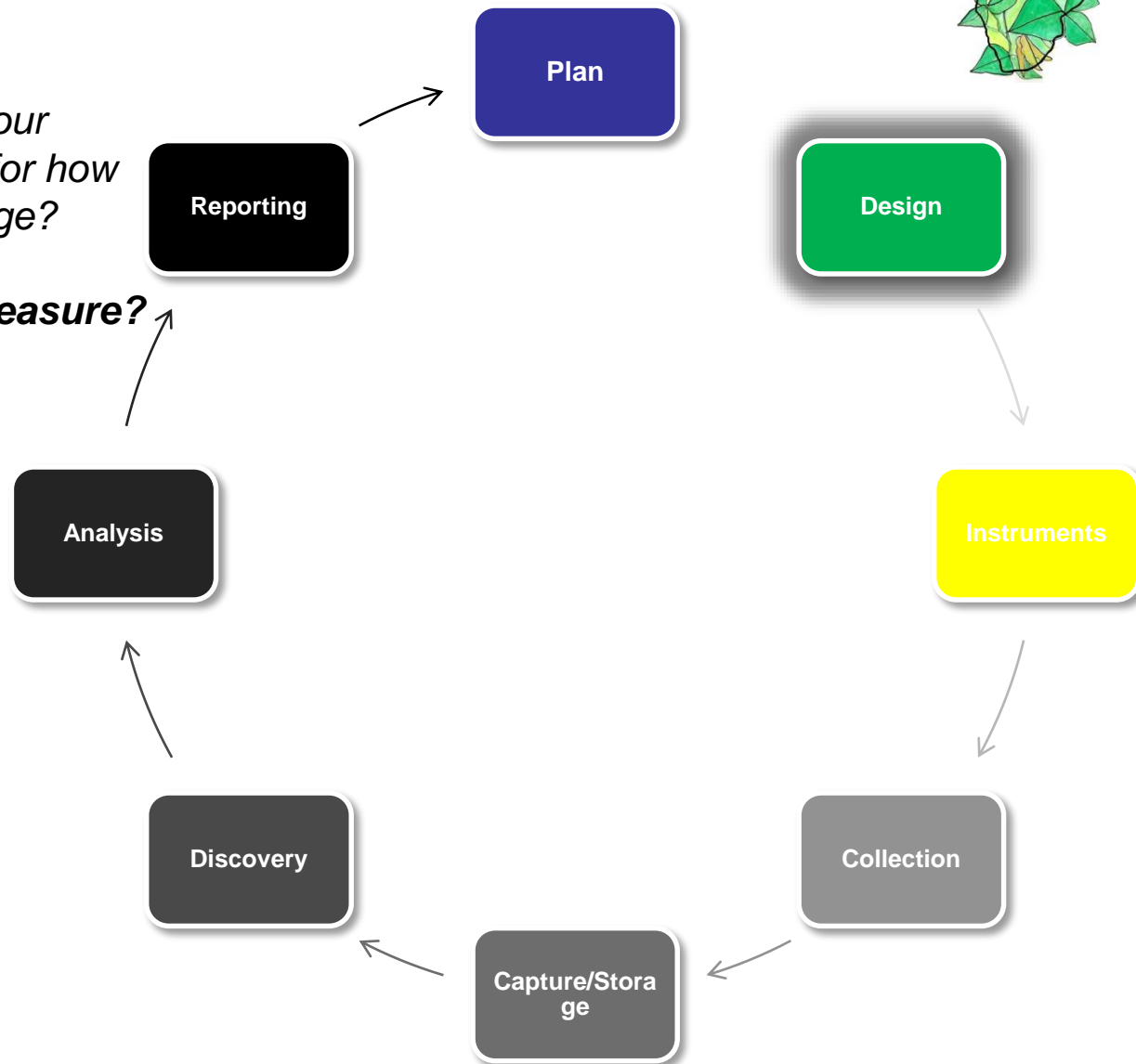


# Sources of Data Quality Issues



## Planning

- *Theory of Change* – does your theory of change pose risk for how well you can measure change?
- **Choices around which outcomes/indicators to measure?**
- *Resource needs?*

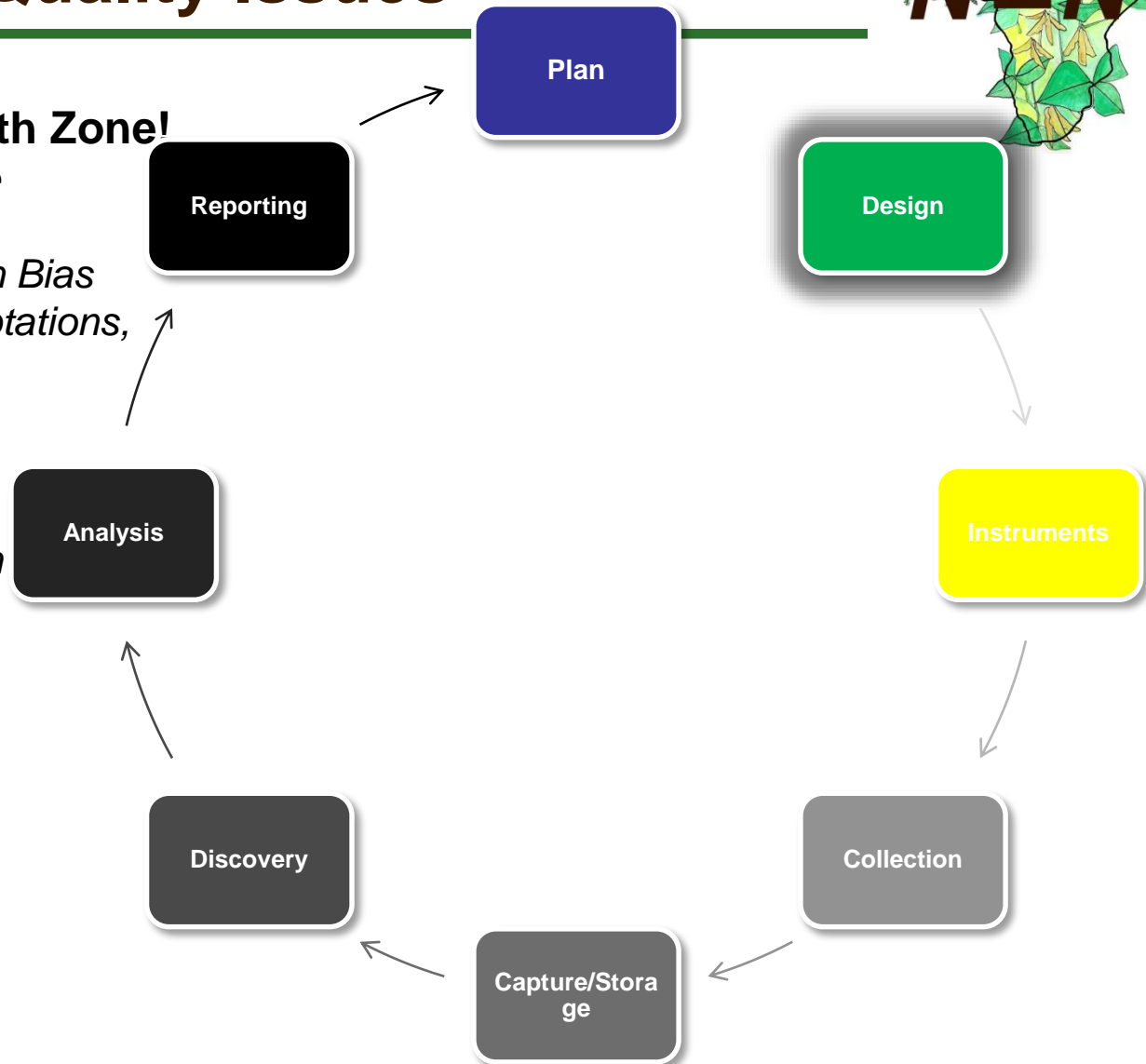


# Sources of Data Quality Issues



## Design – A Potential Death Zone!

- *Qualitative vs. Quantitative*
- *survey*
- *Sampling Frame/ Selection Bias*  
(Sampling Error-focal adaptations, demos)
- **Sampling Strategy – clustering/stratification/aggregation**
- *Sample Size and Precision*

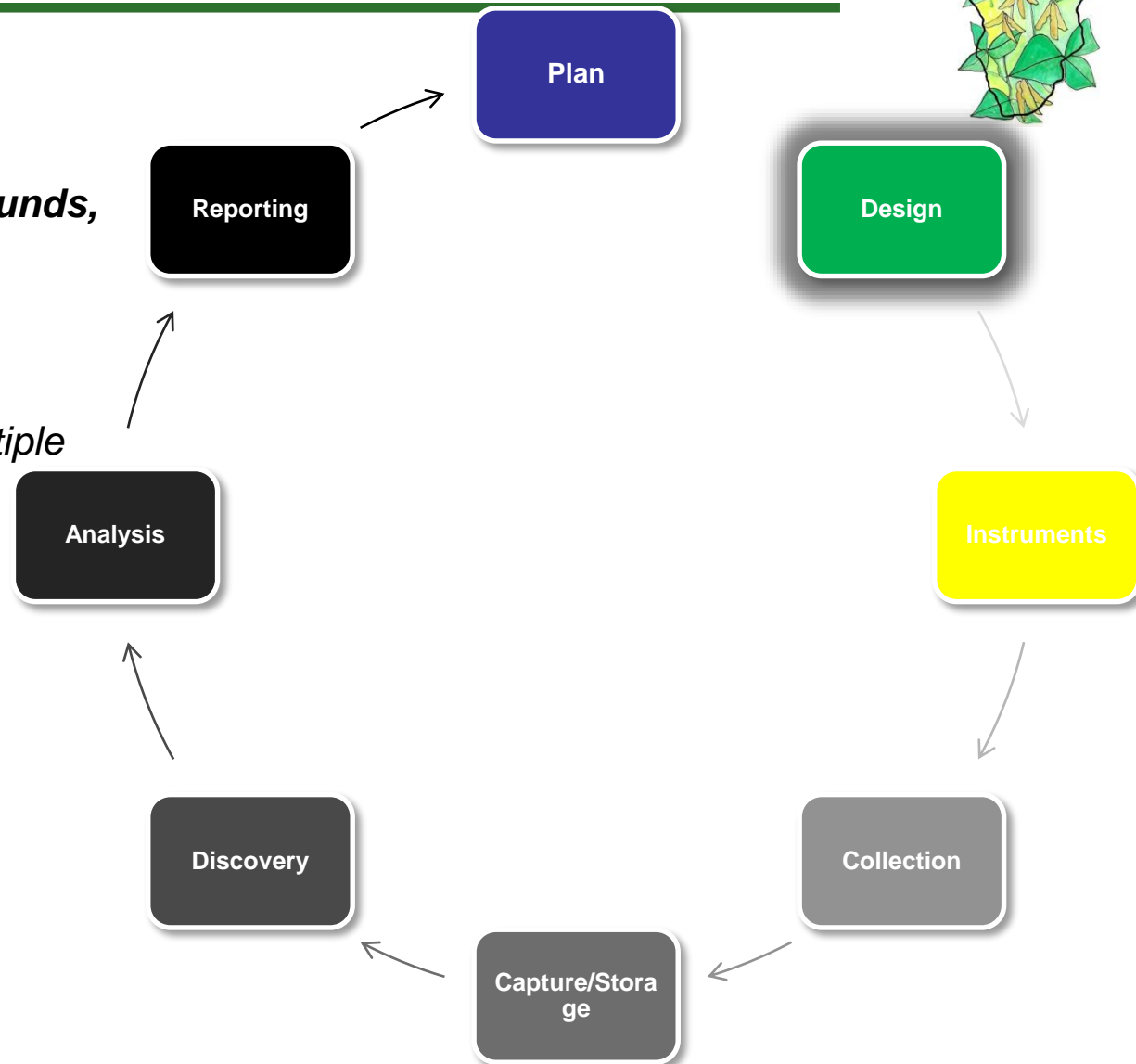


# Sources of Data Quality Issues



## Instruments

- *Instrument design*
- **Paper or Plastic?**
- **Logic control (skip rules, bounds, do loops, etc.)**
- *Wording/vocabulary*
- *Units*
- **Recall**
- *Question format – yes/no, multiple response, etc.*

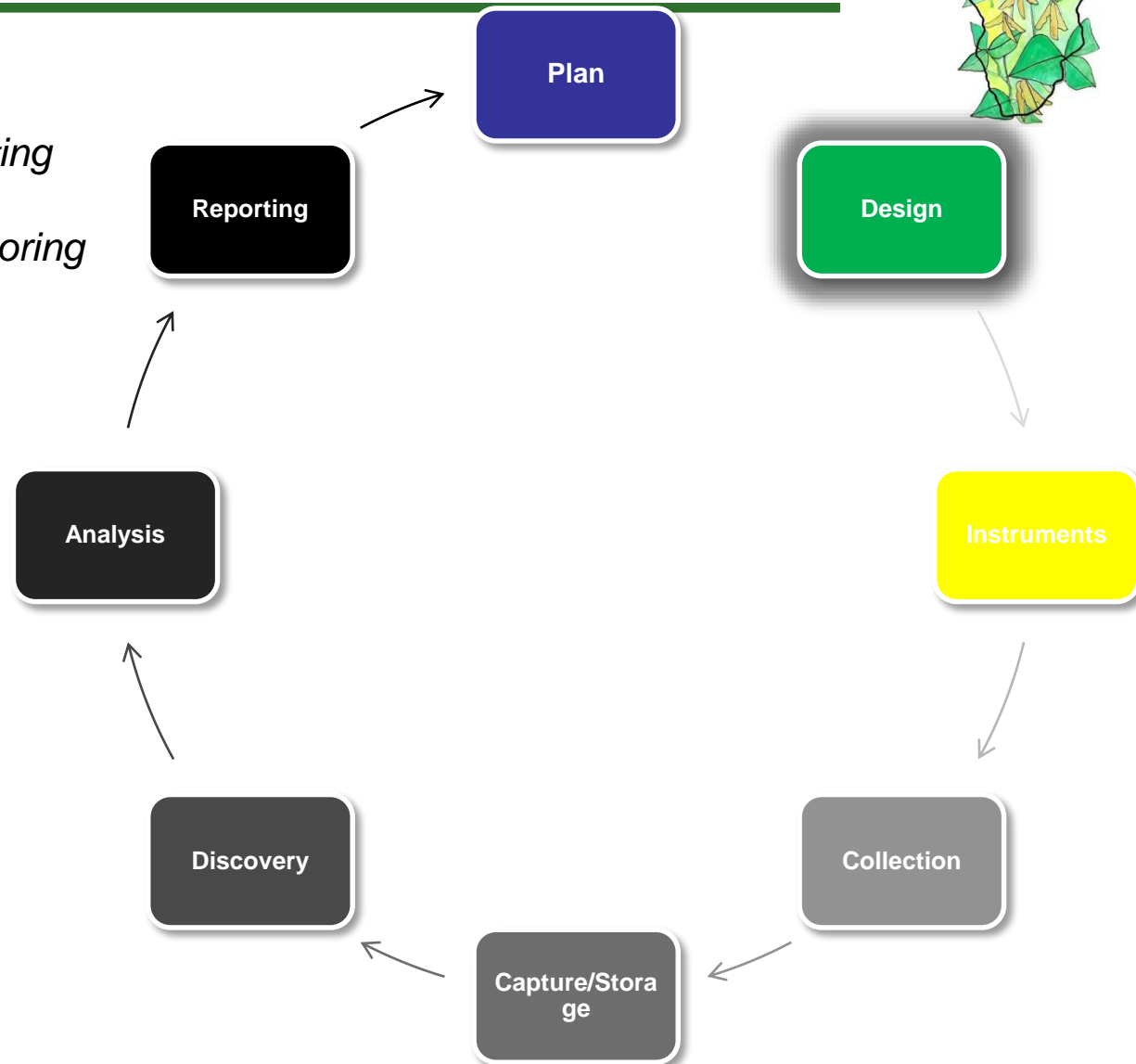


# Sources of Data Quality Issues



## Data Collection Methods

- Enumerator training/pre-testing
- Supervision/ quality control
- Real-time enumerator monitoring
- Timing
- **Measurement error**
- **Respondent error**
- **Enumerator error**
- Degree of difficulty
- Logistics/esprit de corps

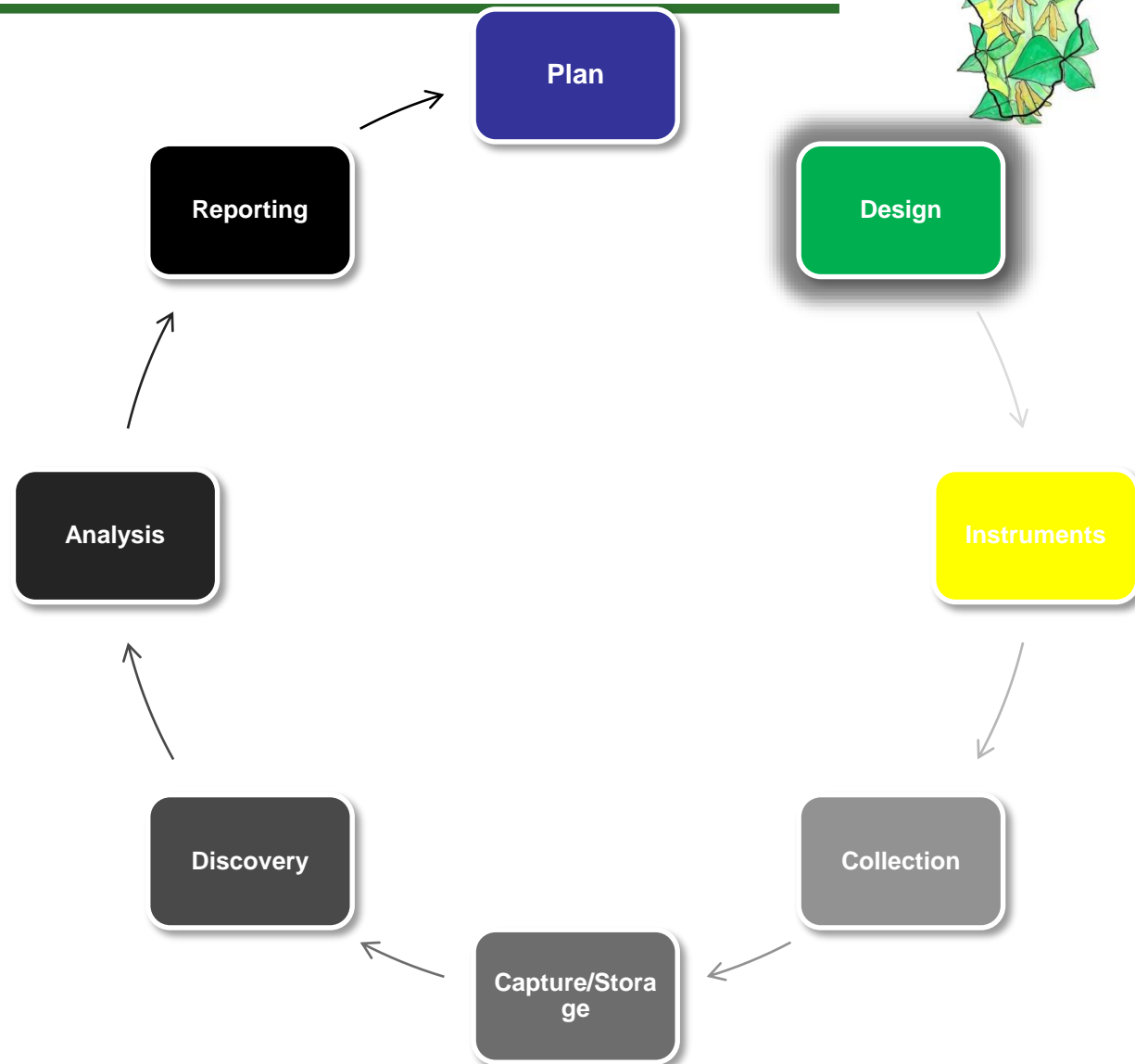


# Sources of Data Quality Issues



## Capture/Storage

- *Input error*
- *Quality of data processing*
- *Security*



# Sources of Data Quality Issues

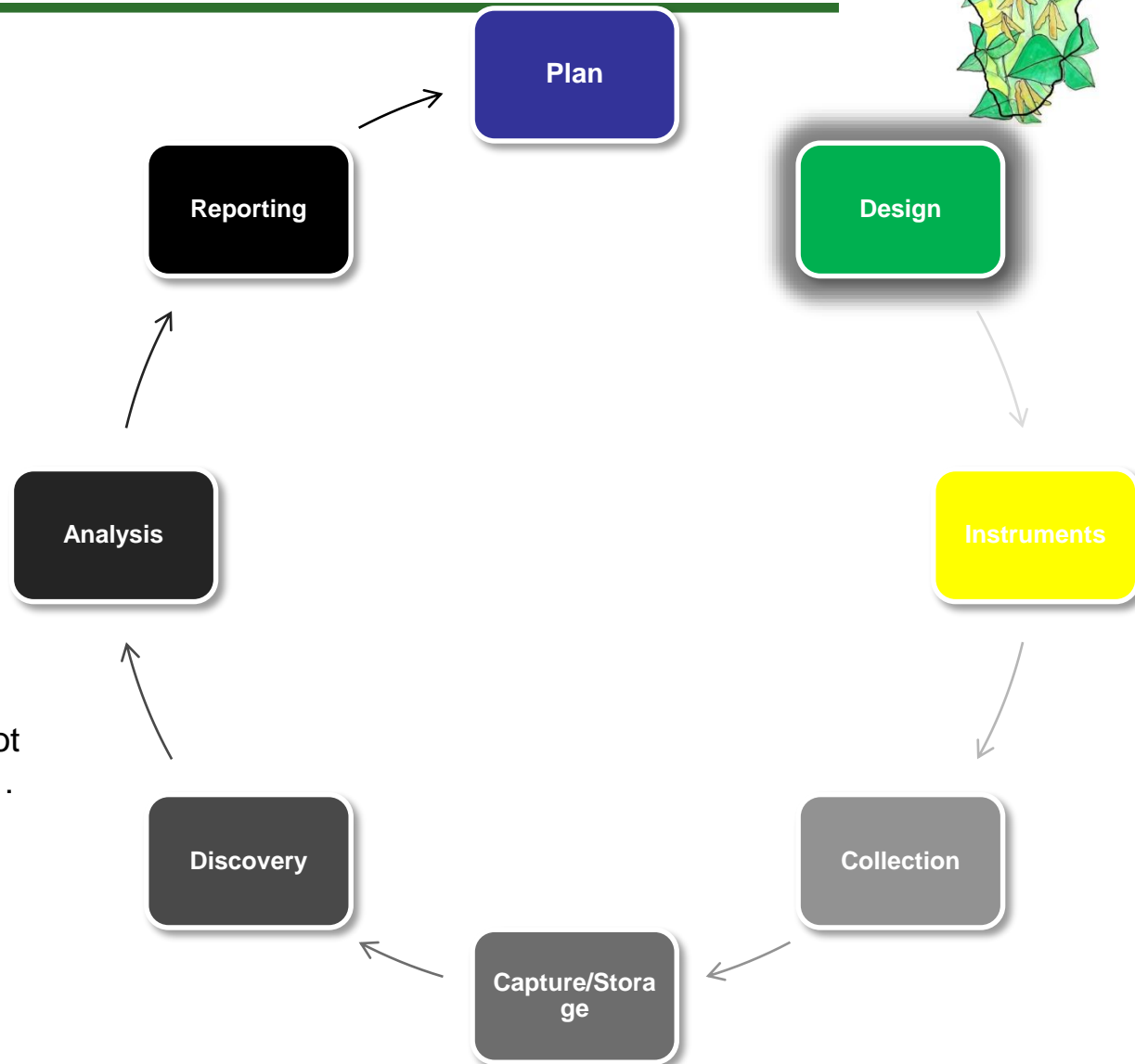


## Discovery

- *Lineage*
- *Meta-data*
- **Data anomaly**
- *Weighting*
- *Cleaning/outliers*

## Data anomaly

- Changes in data layout / data types
  - Integer becomes string, fields swap positions, etc.
- Changes in scale / format
  - Dollars vs. shillings
- Missing and default values
  - Application programs do not handle NULL values well ...
- Gaps in time series
  - Especially when records represent incremental changes.



# Examples of default values



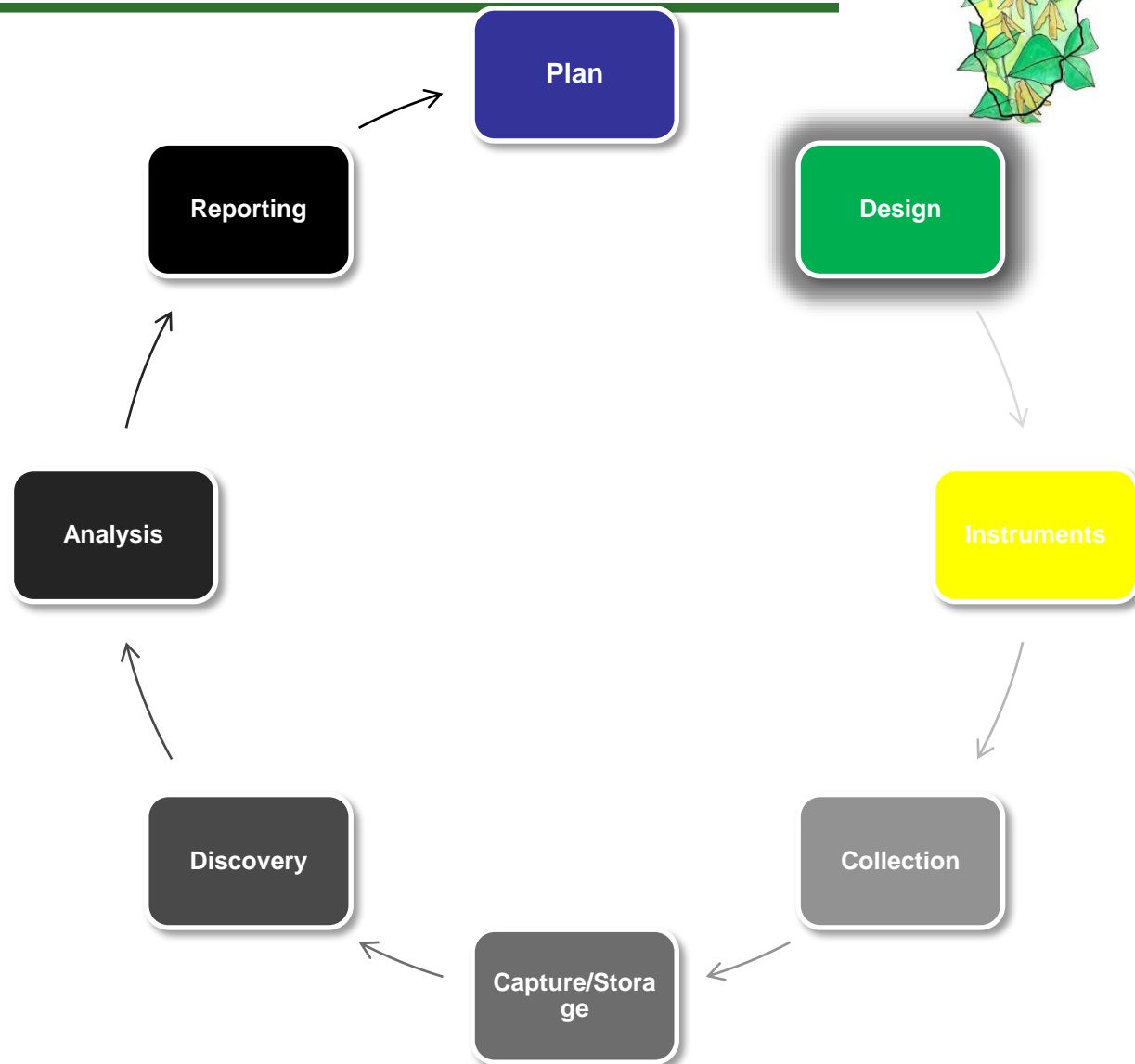
Farm01	M	Mar-11
Farm02	female	2/20/2012
Farm03	1	14/2/2012
Farm04	male	2011
Farm05	0	Feb-11
Farm06	0	3/1/2012
Farm07	F	1/14/201
Farm08	female	2-Jan-13

# Sources of Data Quality Issues



## Analysis

- *Analytic skills*
- *Missing values versus zeros*
- *Appropriate tests*



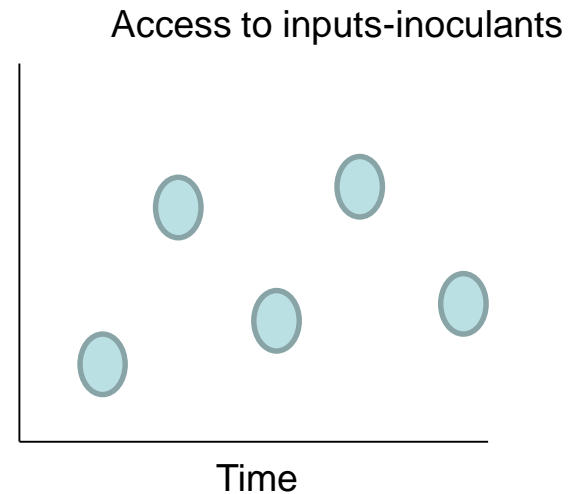
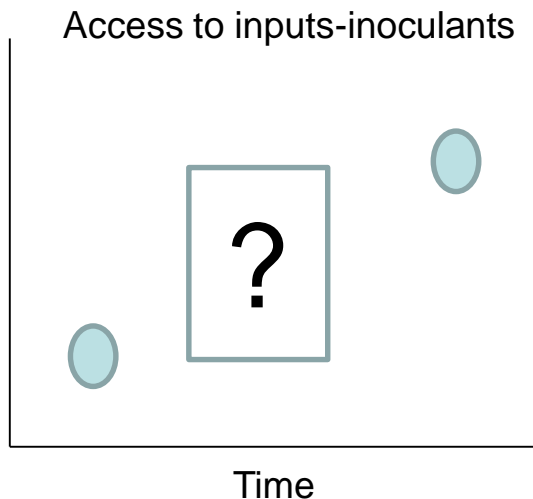


# **Analysing Data for Performance Results**

# Analysing Performance Data



- Focus attention on specific indicators
- Use the indicator definitions to have precise analysis
- Examine changes over time-trend analysis
- Compare present to past data to look for trends and other changes based on the indicators
- The more data points there are, the more compelling the trends.





THANK YOU