How to Solve Ill-Defined and Open Ended Problems

Daniel W. Engels, PhD

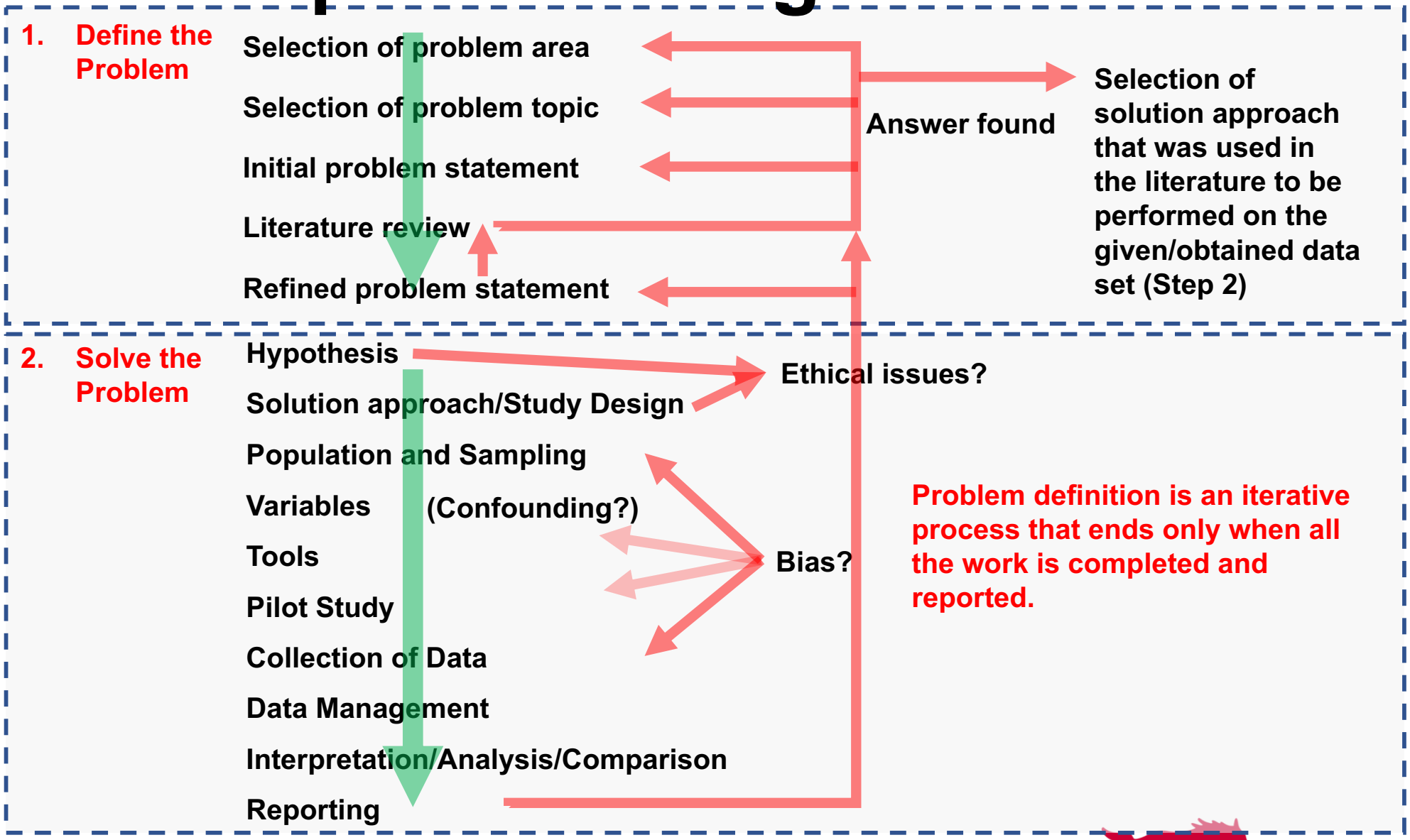
Solving ill-defined and open ended problems is one of the principle jobs of a data scientist*

* Data Scientists are routinely asked to solve a problem, usually stated in the form of a question such as “What can you tell me about the primary influencers in this data set?”

Steps of Solving a Problem

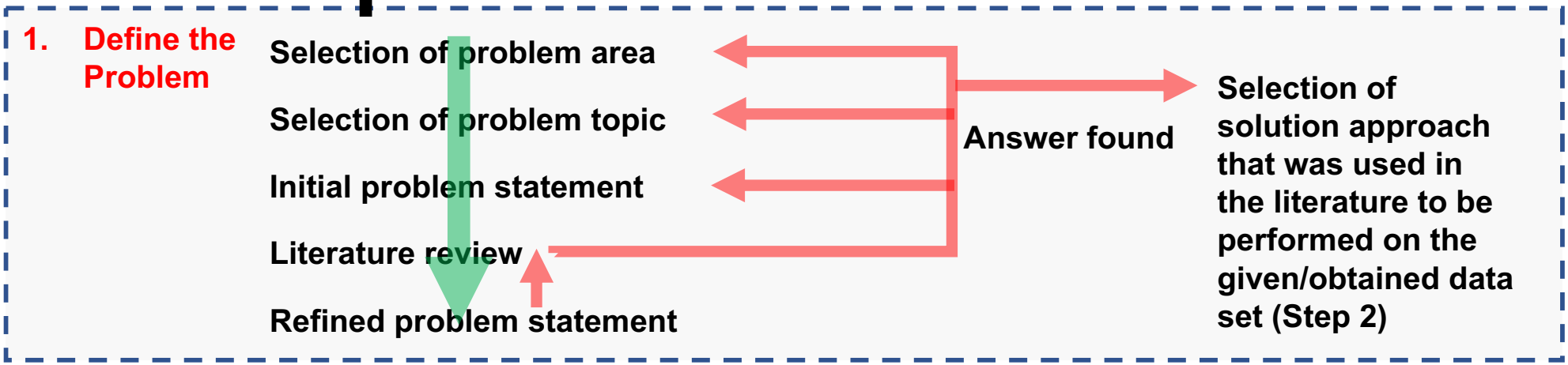
1. Your process begins with the systematic investigation and definition of the actual problem being solved
2. Your solution should involve the systematic collection, analysis, and interpretation of data to solve the problem and the constant narrowing and redefinition of the actual problem being solved.

Steps of Solving a Problem



Adapted from Dr. Rasha Salama

Step 1: Define the Problem



Adapted from Dr. Rasha Salama

Ill-Defined and Open Ended Problems: Coming up with the Initial Problem Statement

- Problem statements typically begin with a question*

  e.g., How do we create World Peace? and

  What are the ethical considerations of the Right to be Forgotten?

* Note that a question such as "Which variables are correlated in the given data set?" is neither ill-defined nor open ended, but is more akin to a statistics homework assignment. Finding an answer to this question is simply an exercise in evaluating correlations between variables. This question is really fairly trivial in terms of your brain power needed to answer. In this talk we are considering non-trivial problems, i.e., problems worthy of a Capstone project only.

What is the Problem in the Question?

- What is the problem that needs to be addressed in order to answer the question?

  How do we create World Peace?

  - Is “World Peace” the problem? Lack of “World Peace”? What is “World Peace”?

Ill-Defined and Open Ended Problems

- What is the problem that needs to be addressed in order to answer the question?

  How do we create World Peace?

  - Is “World Peace” the problem? Lack of “World Peace”? What is “World Peace”?

  What are the ethical considerations of the Right to be Forgotten?

  - What is the “Right to be Forgotten”? How are ethics involved? From whose culture are we evaluating the ethics? Are ethics the problem? Are lack of ethics a problem? Is the Right to be Forgotten the problem?

Ill-Defined and Open Ended Problems

- State the problem

  How do we create World Peace?

  - **The Problem**: Armed conflict, or the threat thereof, creates a world at war.
  - **The Optimization Problem**: Given that armed conflict, or the threat thereof, between nations persists, create a world that minimizes armed conflict or the threat thereof.

Ill-Defined and Open Ended Problems

- State the problem

  How do we create World Peace?

  - **The Problem**: Armed conflict, or the threat thereof, creates a world at war.
  - **The Optimization Problem**: Given that armed conflict, or the threat thereof, between nations persists, create a world that minimizes armed conflict or the threat thereof.

  What are the ethical considerations of the Right to be Forgotten?

  - **The Problem**: European law has created “The Right to be Forgotten” that may be abused by individuals to effectively erase old or prior misdeeds by eliminating their searchability, and effectively their existence, from the Web.
  - **The Optimization Problem**: Given that ”The Right to be Forgotten” reduces the ease with which information about an individual may be located using the Web, identify (which and explain how) the maximum number of ACM Code of Ethics points that may be violated by exercising one’s Right to Be Forgotten.

Motivation

- Motivation identifies a problem domain and motivates why the problem exists.
- Motivation is essential for placing the problem into context and making it understandable.

Example Motivation and Problem Statement

Overfill has been a serious problem facing Bluffington's city waste facilities for the last decade. By some estimations, Bluffington's city dumps are, on average, 30% above capacity—an unsanitary, unsafe, and unwise position for the city to be in.

Several methods have been proposed in order to combat this. Perhaps the most popular of these is the simplest: building two new landfills on the county outskirts. Others have proposed stronger recycling campaigns and larger per-bag waste disposal costs as a way to lessen the potential damage of our trash situation.

Bluffington is close to drowning in trash. Action is needed if the city is to remain the clean, safe place to live it has always been.

From Wiki How https://www.wikihow.com/Write-a-Problem-Statement

I Have a Problem Statement. Now What?

Literature Review

Where to Begin Your Literature Review?

A quick reminder that the SMU library has significant resources to support your work. You may access and search the library databases at: http://www.smu.edu/cul

Jennifer Sullivan is the Data Science Librarian dedicated to support the MSDS program and the on-campus data science programs and research. You may reach out to her directly via email at: jlsullivan@mail.smu.edu

Jennifer has also put together a Data Science Guide that you may access at: http://guides.smu.edu/datascience

I strongly recommend you start your research into your problems there.

I've Found a Solution to my Exact Problem in the Literature

If your problem has been solved in the literature…

- …**verify** that it's truly **the problem** you want to solve and that your real problem is addressed by the found solution
- …**use the found solution** to solve your problem on your data set and compare your results to the published results (are the results the same?)
- …or, **refine your problem statement** and then repeat the literature search
- …or, refine or **change your** selected topic or area and come up with a new **problem statement**

I've Found a Solution to a Specific Version of My Problem in the Literature

- Your problem statement is too vague if more specific versions of your problem are being solved in the literature

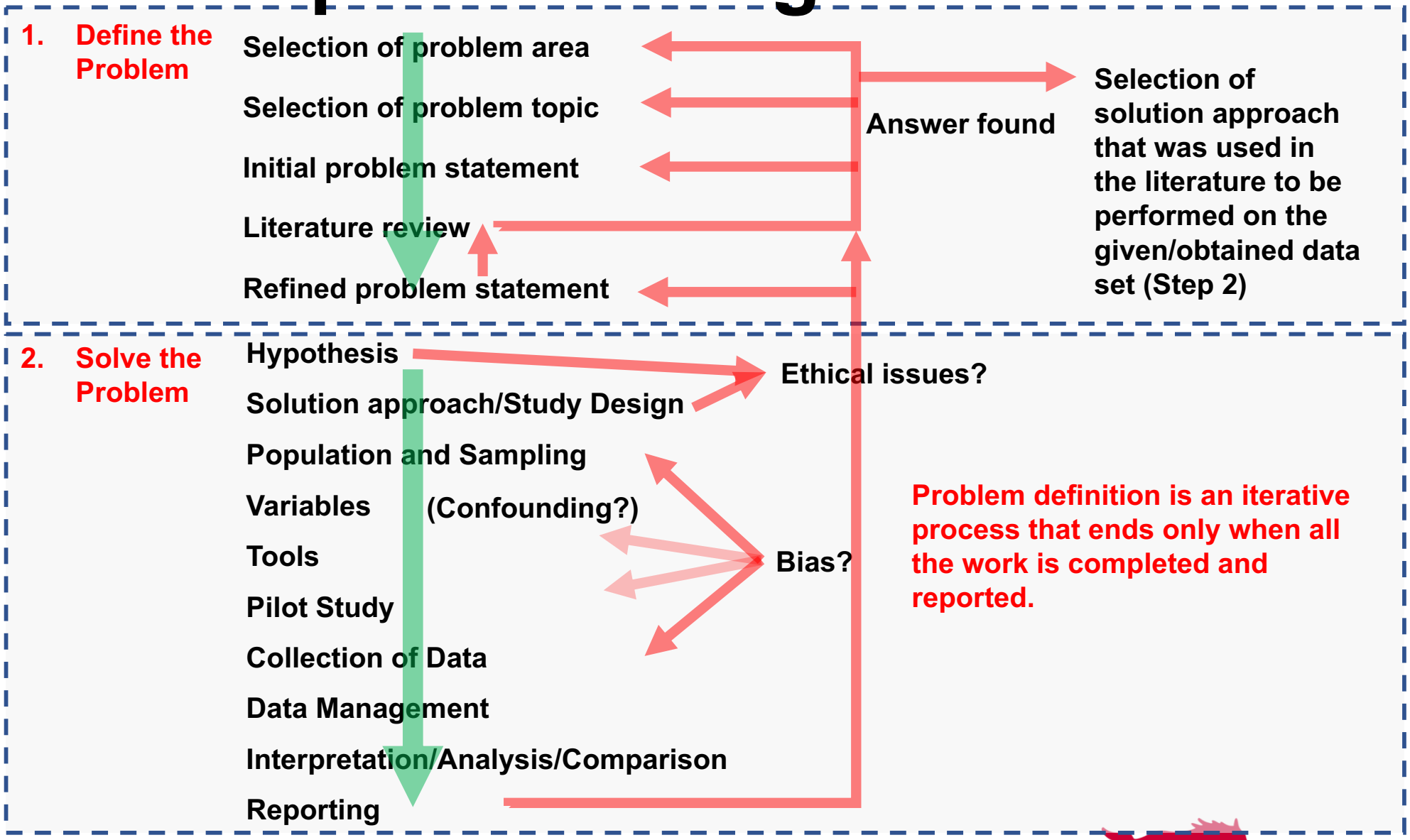
  …refine your problem statement to make it more specific

  …repeat the literature search

I Haven't Found any Solutions to my Specific Problem

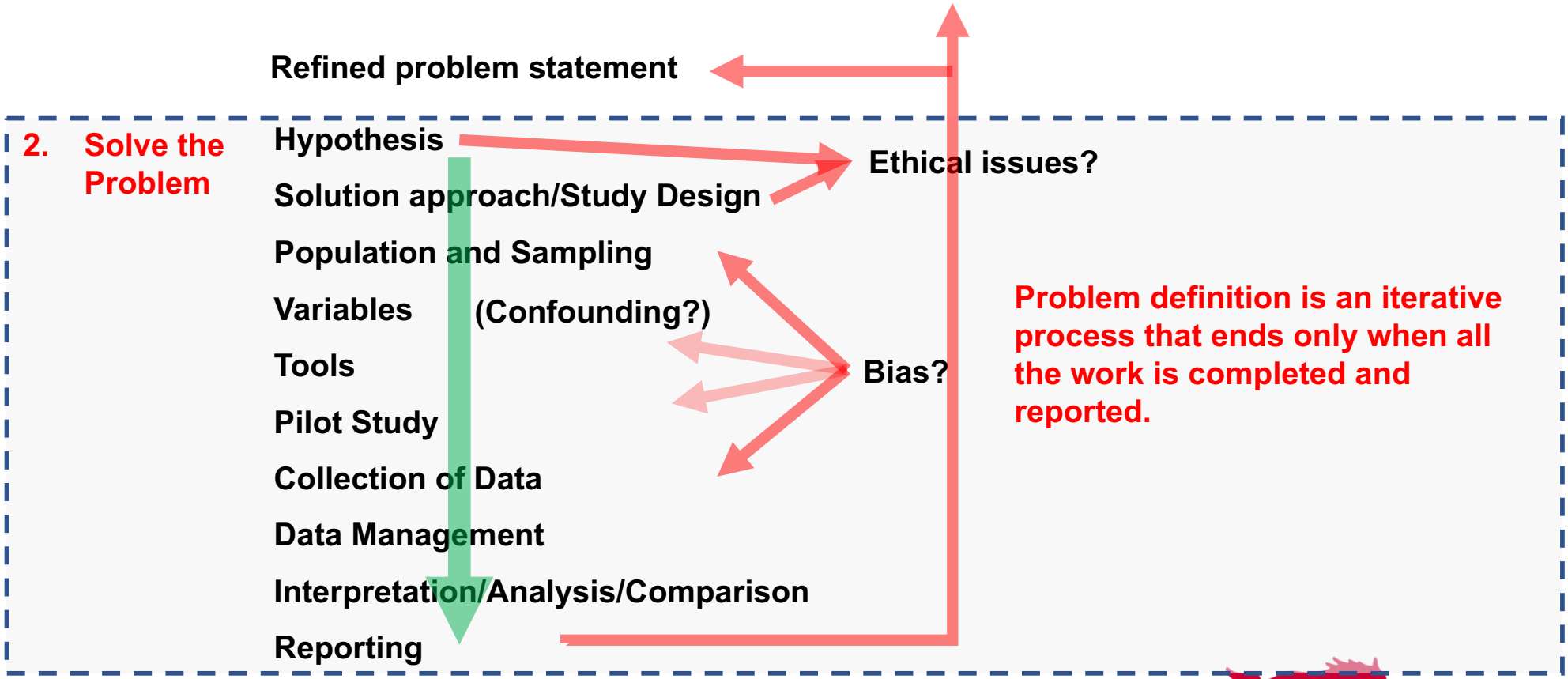
- Verify that your problem is stated correctly
- Identify related problems and problem variations and perform literature search for them
- Begin Step 2: Solving your problem

Steps of Solving a Problem



Adapted from Dr. Rasha Salama

Step 2: Solve the Problem

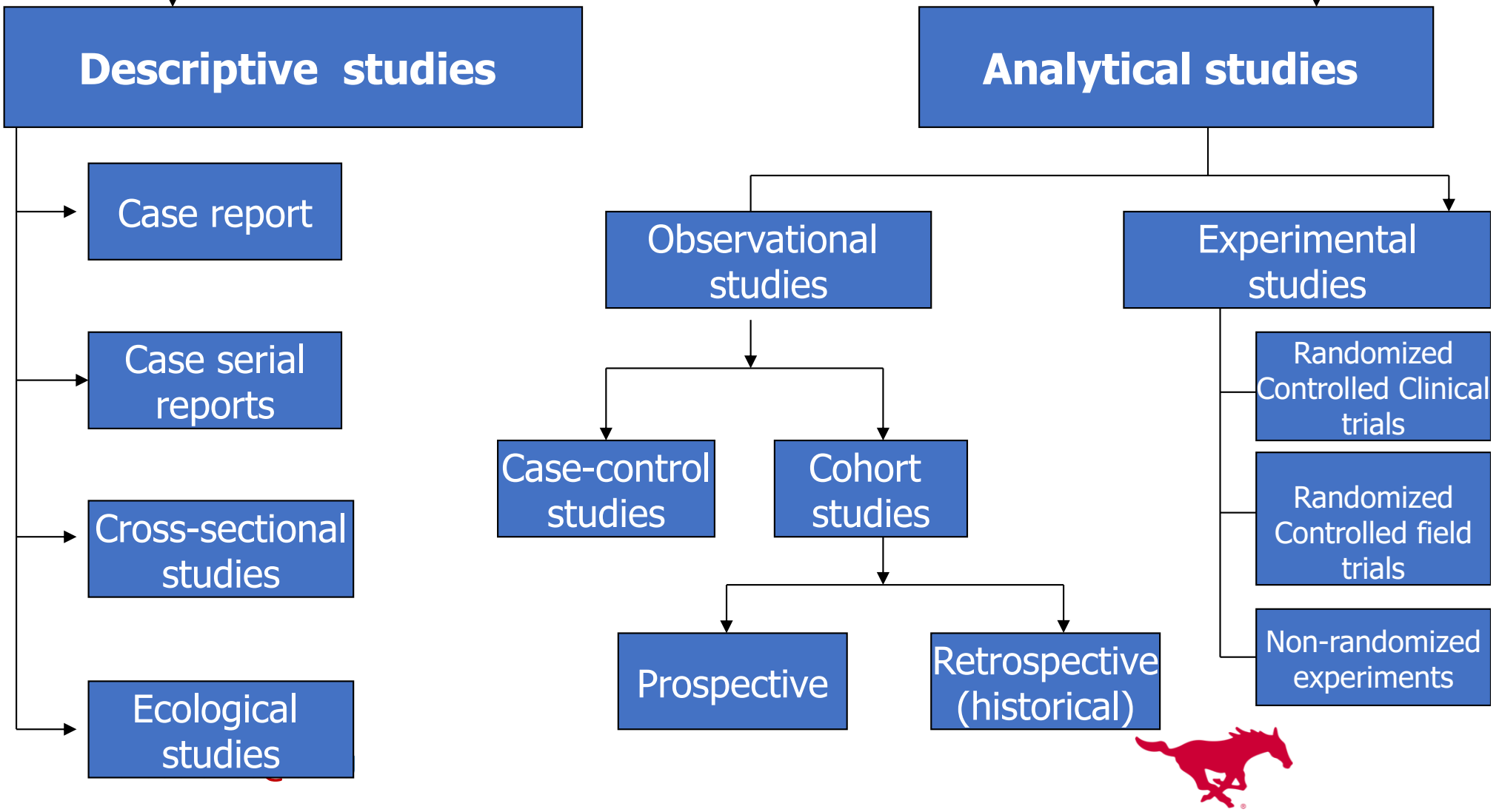


Adapted from Dr. Rasha Salama

Hypothesis

Form a hypothesis as to what/how you can solve your problem.

Solution Approach/ Study Design



Selecting Study Design

- Purpose of the study
- State of existing knowledge (in relation to study question)
- Characteristics of the study variables
- Latency
- Feasibility

Population and Sampling

Sampling is the process of selection of a number of units from a defined study population.

The process of sampling involves:

1. Identification of study population
2. Determination of sampling population
3. Definition of the sampling unit
4. Choice of sampling method
5. Estimation of the sample size

Identification of Study Population

- The study or target population is the one upon which the results of the study will be generalized.

- It is crucial that the study population is clearly defined, since it is the most important determinant of the sampling population

Data Collection

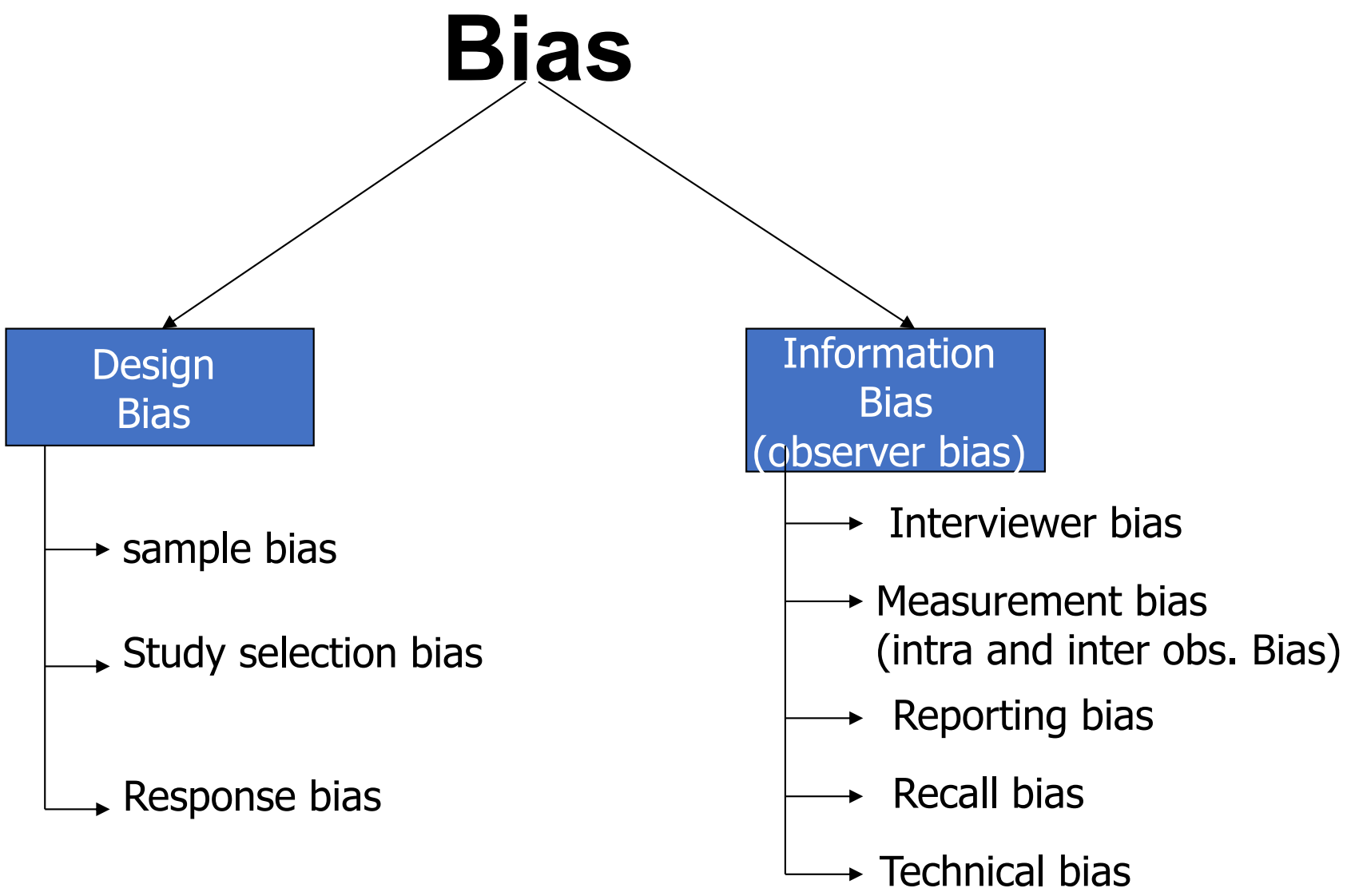
- Data collected are “variables”
- Variables are classified according to their:
  - Type:
    - QT (continuous, discrete)
    - QL (ordinal, nominal)
  - Role in the study:
    - Dependent
    - Independent
  - Relationship with other study factors:
    - Main study variables
    - Confounding variables
    - Effect modifiers
    - Intermediate factors

Methods of Data Collection

- Selection of the suitable technique depends on:
  - The availability of information
  - The type of data
  - The resources available
  - The characteristic of the tool

Research Tools

- Most important techniques:
  - Using available information (records)
  - Observation (checklist)
  - Self-administered questionnaire
  - Interviewing (individual/group)
  - Measuring (all lab tests and other investigations)


Bias
Design Bias
sample bias
Study selection bias
Response bias
Information Bias (observer bias)
Interviewer bias
Measurement bias (intra and inter obs. Bias)
Reporting bias
Recall bias
Technical bias

Data Management

- Data management is the whole process of dealing with data from the very beginning of the study. Data analysis is just the last part of it.
- It can be divided into the following phases:
  - Preparation of data entry
  - Data entry
  - Data storage and retrieval

Analysis

Analysis of the data so as to solve the problem/prove or disprove your hypothesis

Interpretation

Discussion of the results in a way that relates data and analysis obtained to each other clarifying the associations and other findings including the solution to the problem.

Comparison

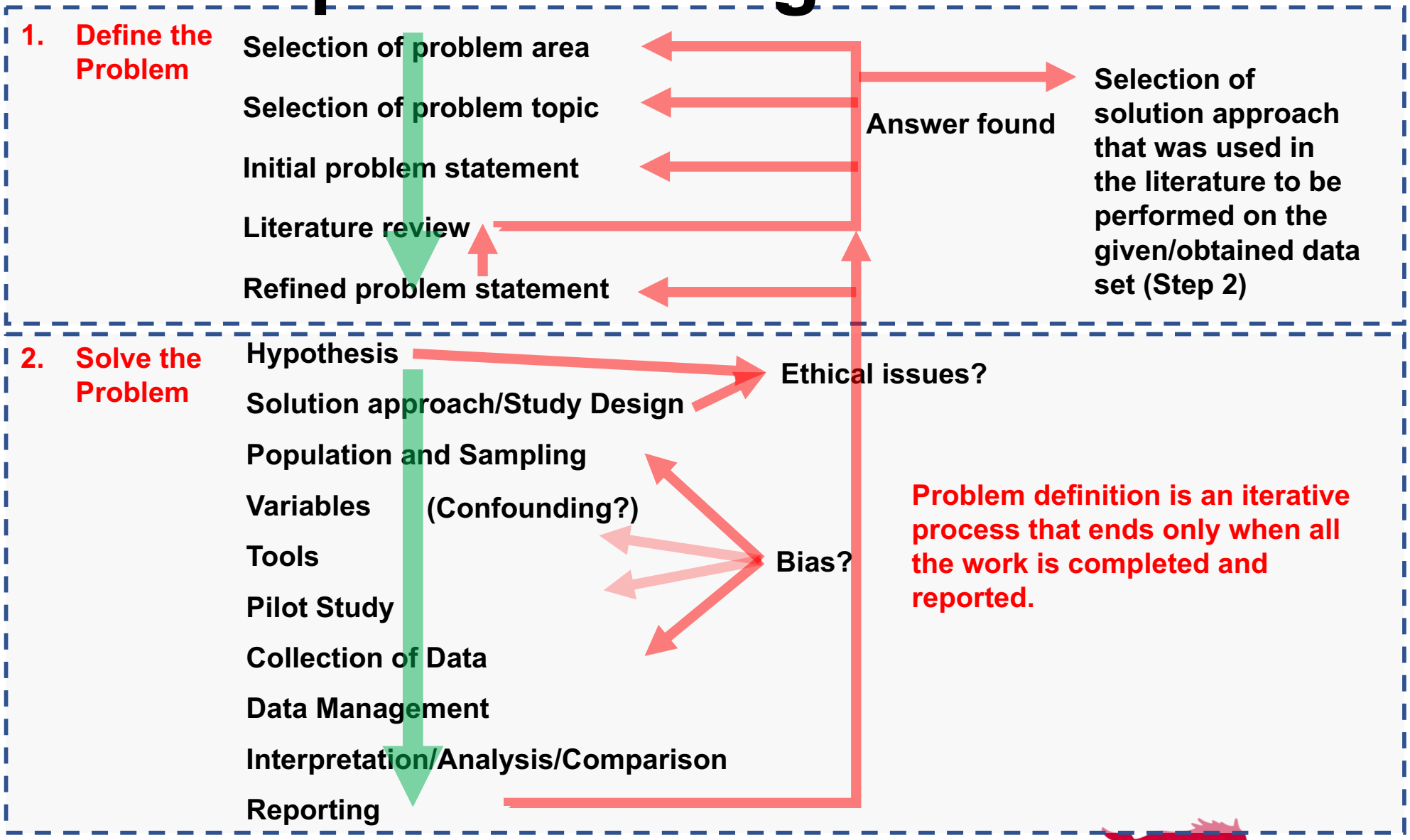
For the very simple reason that YOU have defined the problem for which YOU have come up with an answer, it is incumbent upon YOU to compare your solution/approach/whatever is appropriate to other solutions/approaches/problem variations/whatever is appropriate to show your solution is robust and good.

Reporting

Report your findings…

…for Capstone this includes writing a technical paper, a lightning presentation, and a poster presentation

Steps of Solving a Problem



Adapted from Dr. Rasha Salama

Questions?