Example - Data Imputation

- Readings from five sensors (X_1, X_2, X_3, X_4, X_5) are made available to you (for 100 different tests, check the file, *GTPvar.csv*). The readings are not arranged according to any order.

- There are some records, though, where there are a few missing readings that are marked *NA*.

- Your supervisor has asked you if there are any ideas that can be employed to rationally fill the missing values. Can you develop a data analytic approach to answer this question ?

Example - Data Imputation

- STEP 1: Problem Definition
  - Fill in missing data records

- STEP 2: Problem Characterization
  - Given part of the information, fill the missing information
  - Relate missing information with known information
  - Function approximation problem
  - $x_{\text{unknown}} = \text{f}\,(x_{\text{known}})$

Example - Data Imputation

- STEP 3: Solution Conceptualization
  - Need complete data set for identifying the function
    - Collect records without missing data
  - Assumption: All variables are independent of each other

    $\Rightarrow$ no relation exists between the variables
  - For each variable, fill the missing data with the most likely value
- Step 3a: Verify assumption
  - Assumption not satisfied
- STEP 3: Solution Conceptualization
  - Assumption: Variables are inter-related
  - Step 3a: Assumption cannot be verified a priori

Example - Data Imputation

- STEP 4: Method Identification
  - Identify relationships using null space
  - Fill in missing values using the notion of pseudo-inverse
- STEP 5: Actualization
  - Implement in R programming language
- STEP 6 : Assess assumptions
  - Use it in intended application to check performance ?
- Solution realized (OR)
- STEP 3: .....
- STEP 4: …..
- STEP 5: ….
- STEP 6: ….

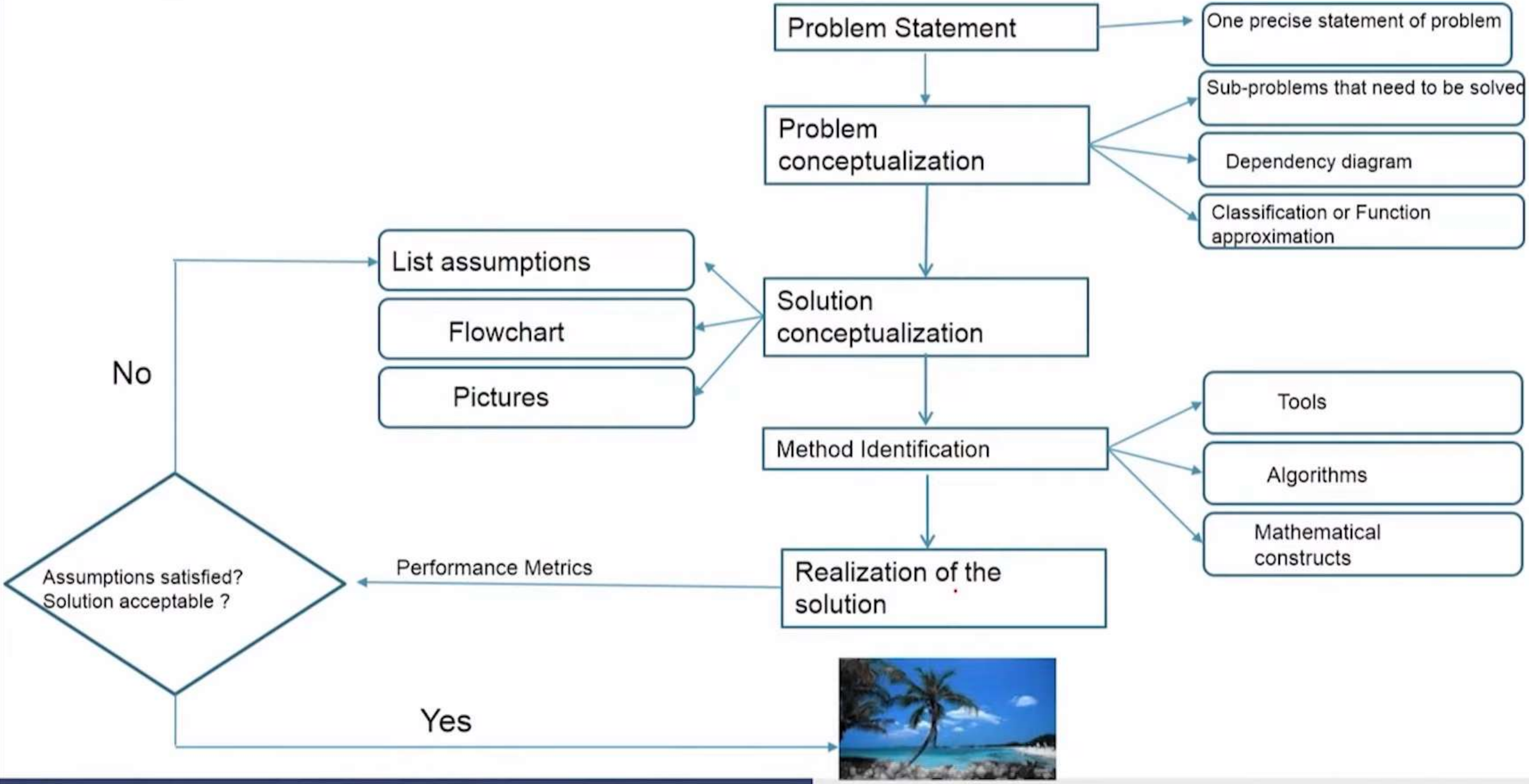
Conceptual Framework for Solving Data Analysis Problems

- START: Problem Arrival – Whole lot of words. Diffuse problem statement
- STEP 1: Problem Definition – Convert the loose words in to one problem statement (as precise as possible)
- STEP 2: Problem Characterization
    - Define high-level problems and sub-problems that need to be solved maintaining a high-level granularity
    - Develop a dependence diagram
    - Identify the problems and sub-problems as either function approximation or classification problems

Conceptual Framework for Solving Data Analysis Problems

- STEP 3: Solution Conceptualization – Visualization of the solution process through two conceptual devices
  - List assumptions (3a – Assumptions that can be verified a priori)
  - Flowchart
  - Pictures
- STEP 4: Method Identification – Map the elements of the flowchart and pictures into mathematical modules
  - Identify mathematical constructs/algorithms for the elements in the flowchart/picture
  - Identify lacunae – Data scientist to conceptualize method development
  - Develop the solution method map
- STEP 5: Actualization
  - Realize the solution method map in a software environment of choice
- STEP 6 : Assess assumptions and go through steps 3 to 6 if necessary

Conceptual Framework for Solving Data Analysis Problems



Problem Statement

One precise statement of problem

Problem conceptualization

Sub-problems that need to be solved

Dependency diagram

Classification or Function approximation

Solution conceptualization

List assumptions

Flowchart

Pictures

Method Identification

Tools

Algorithms

Mathematical constructs

Realization of the solution

Performance Metrics

Assumptions satisfied? Solution acceptable ?

No

Yes