- ***Lesson 1:*** **Why collecting data? (the process of KDD)**

In order to understand the importance of creating a base of data, we should proceed at different levels, identifying:

. (I) The meaning, role and finality of data in the decision making process (general-philosophical level);

. (II) The benefits of a big mono-specialistic clinical database (specific-practical level).

Moving across these two different planes, this lesson is aimed at clarifying the science behind the data collection and the following knowledge extraction process.

These activities represent the fundament of any decision-making strategy, which is the real ultimate goal and reason to justify a data collection effort

- ***Lesson 1:*** **Why collecting data? (the process of KDD)**

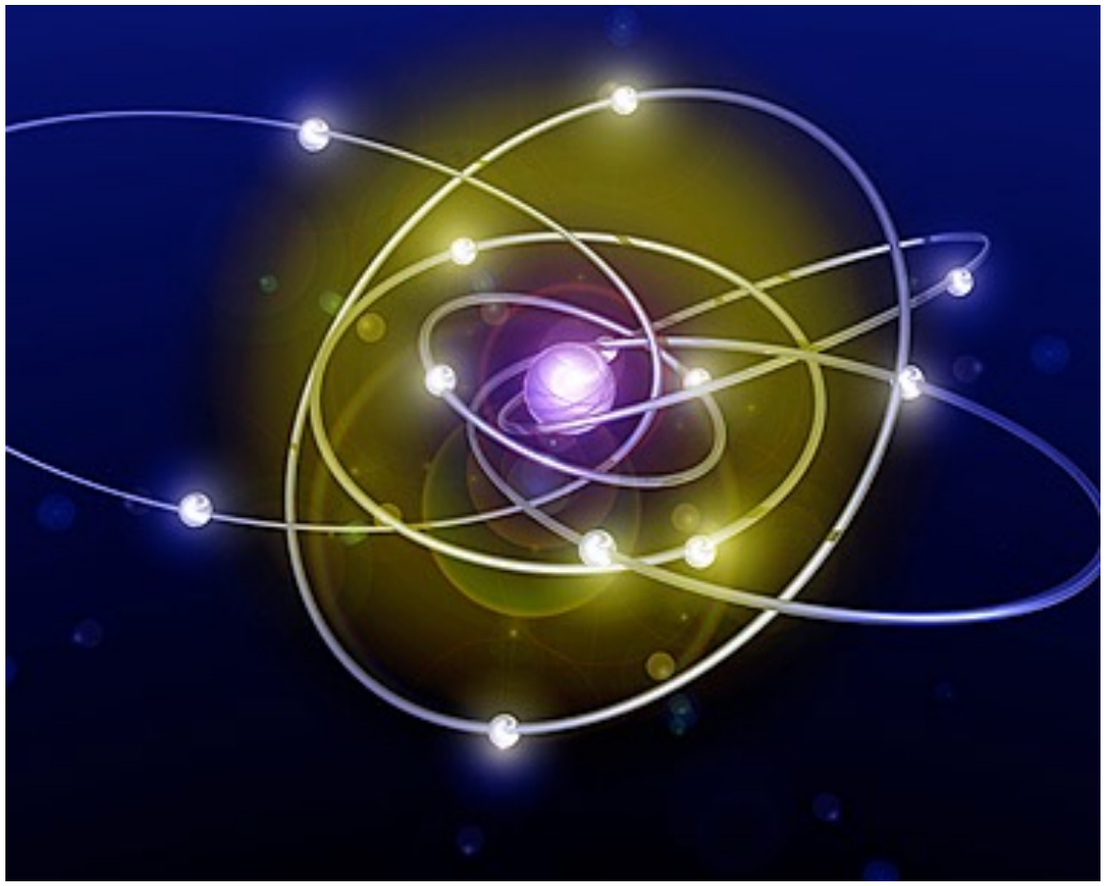
Our ability to analyze and understand massive datasets lags far behind our ability to gather and store the data

However, whether the context is business, medicine, science, or government, the datasets themselves (in raw form) are of little direct value. What is of value is the knowledge that can be inferred from the data and put to use.

EVERY ACTION IS BASED ON INFORMATIONS

- SENSORY
- COGNITIVE
- EMOTIONAL

THE “QUANTUM” OF EACH KIND OF INFORMATION IS A “DATUM”



TO PROCEED IN OUR LIVES INDIVIDUALLY OR AS A COMMUNITY

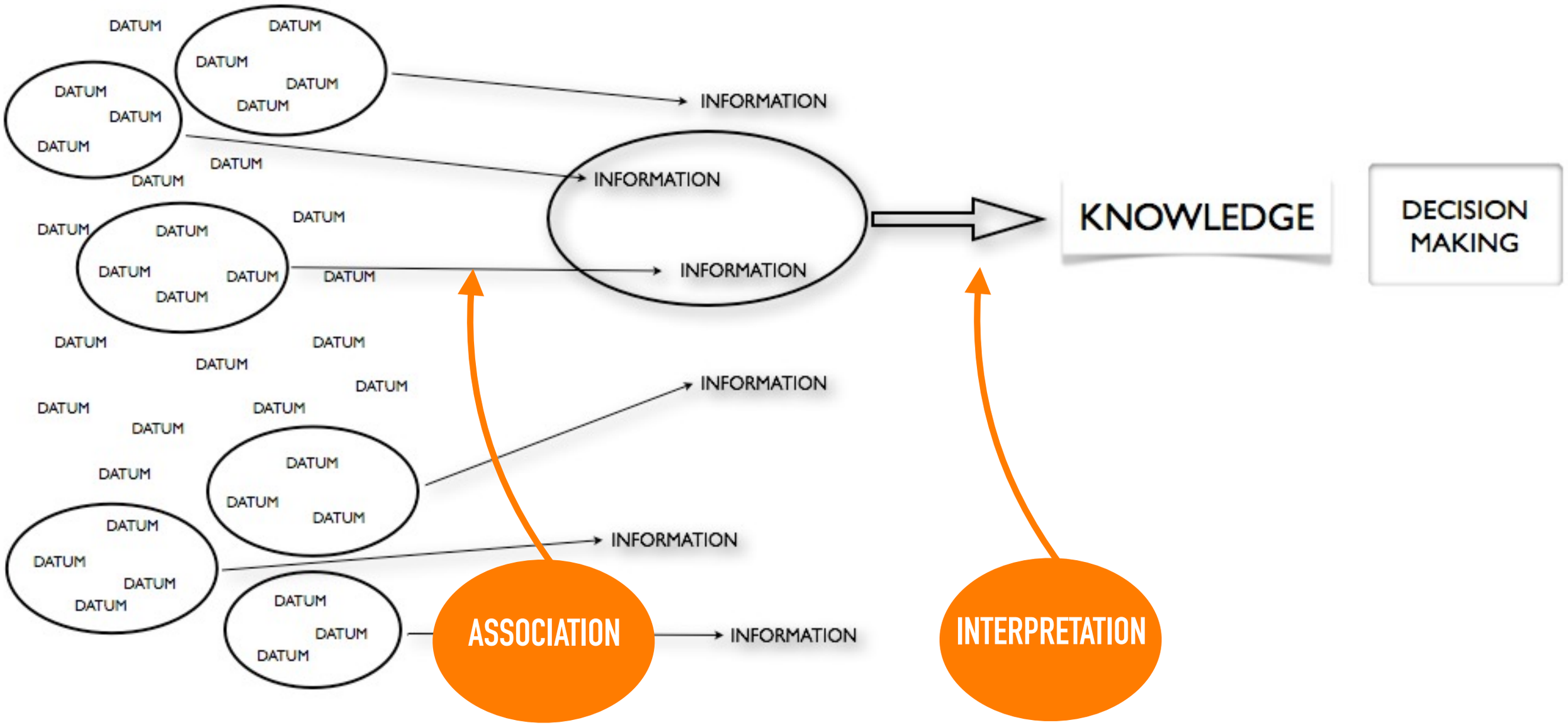
- WE SMELL, TOUCH, EAR, FEEL, ANALYZE, INTERPRET

DATA

IN GOD WE TRUST. ALL OTHERS MUST BRING DATA

W. Edwards Deming

FROM DATA TO ACTION: THE PROCESS OF DECISION-MAKING

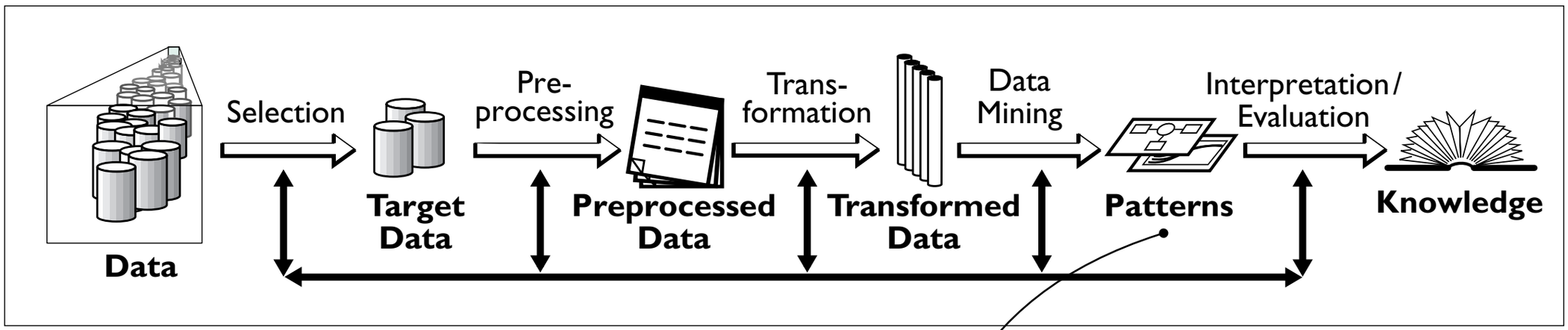


EVERY ACTION IS BASED ON INFORMATIONS

- DATUM
- HR, BP, smocking history
- INFORMATION
- Physical status
- KNOWLEDGE
- Need to reduce the risk of an heart attack

CONSEQUENT ACTION

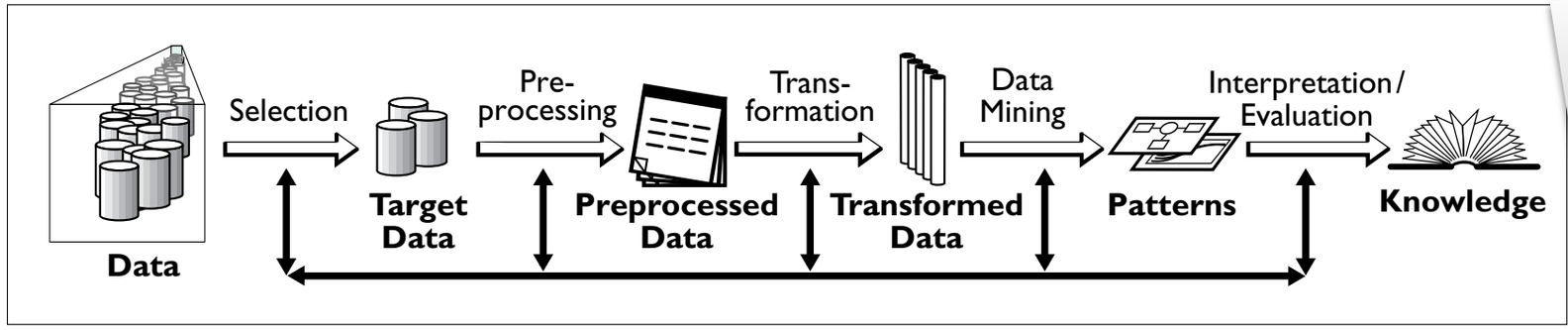
Figure 1. Overview of the steps constituting the KDD process



Interestingness, is usually taken as an overall measure of pattern value, combining:

- validity,
- novelty,
- usefulness,
- simplicity.

Figure 1. Overview of the steps constituting the KDD process

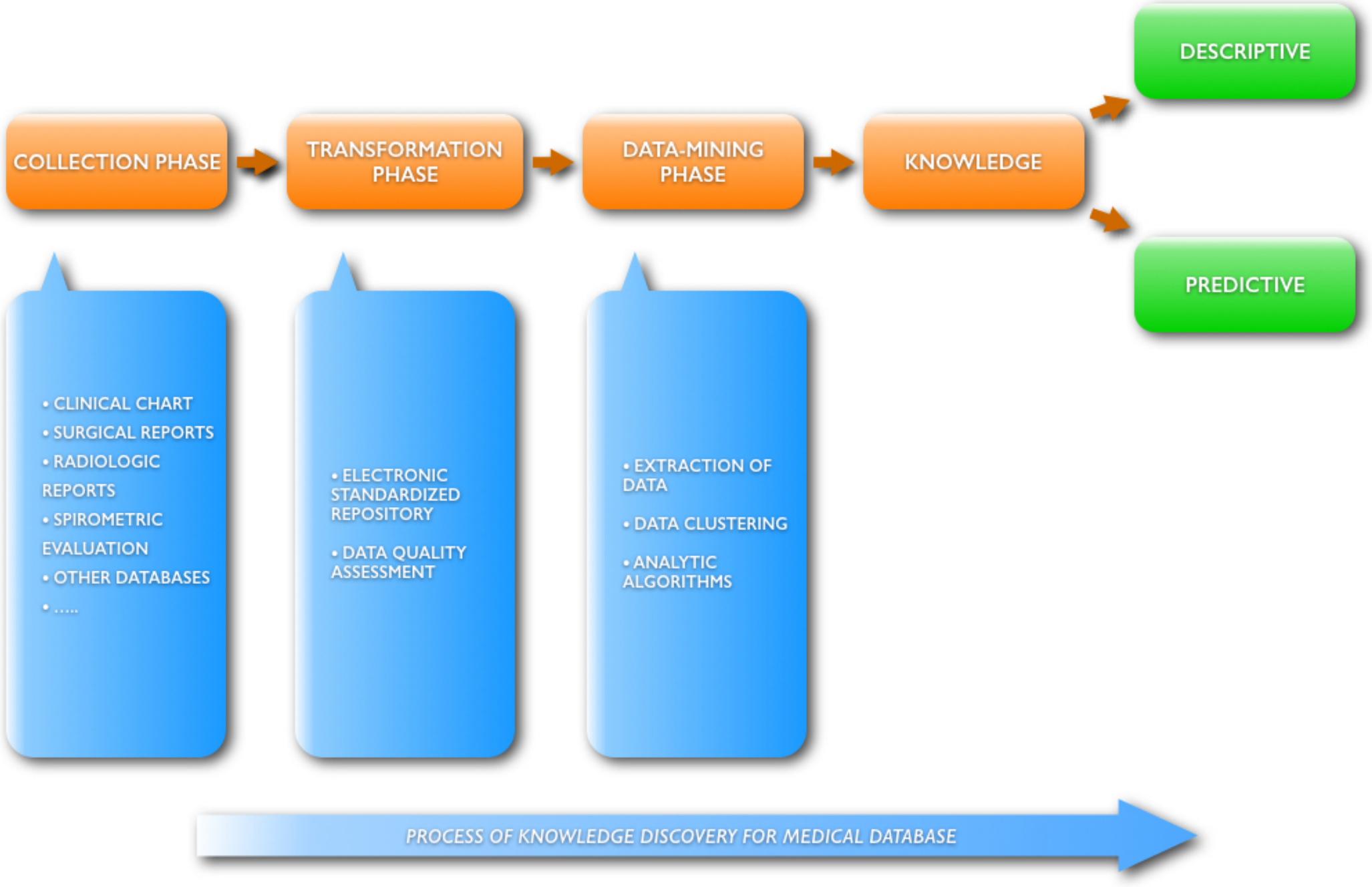


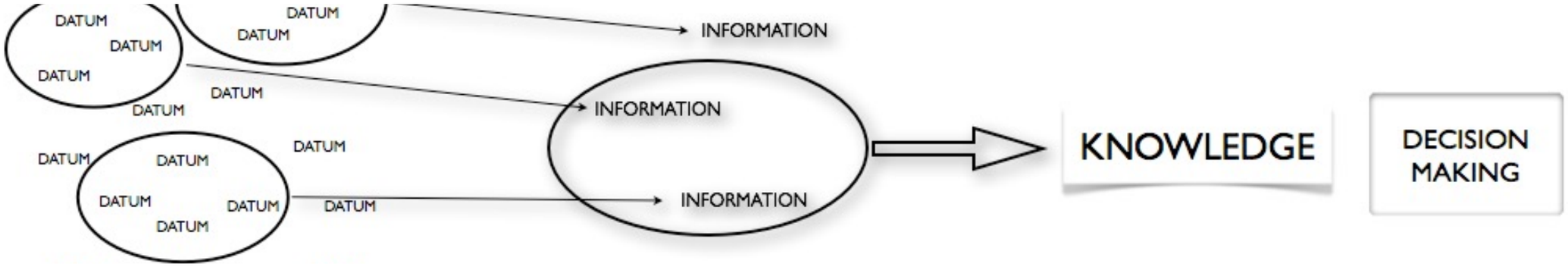
The KDD process is interactive and iterative (with many decisions made by the user)

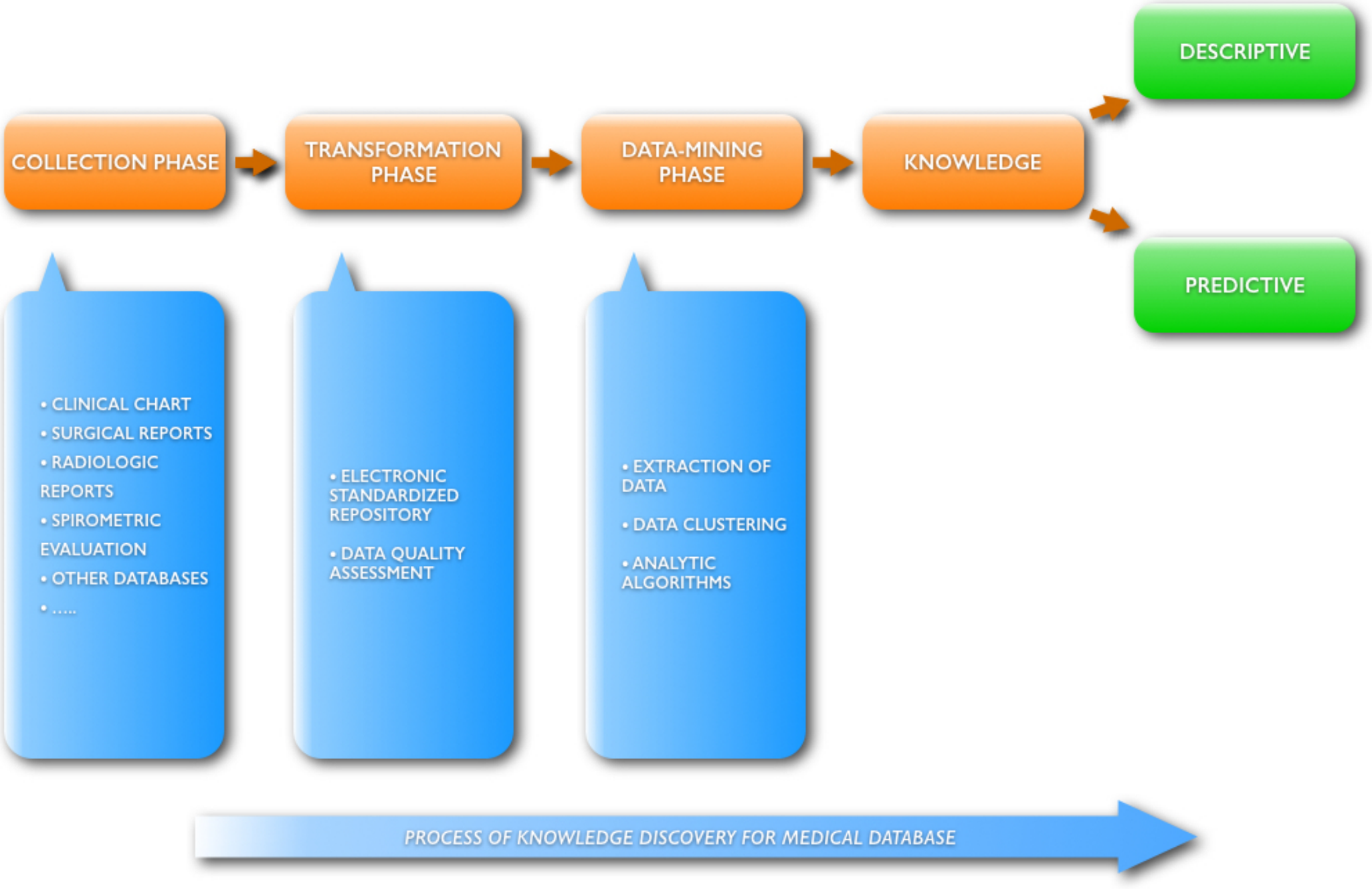
The KDD process is interactive and iterative (with many decisions made by the user), involving numerous steps, summarized as:

1. *Learning the application domain: includes relevant prior knowledge and the goals of the application*
2. *Creating a target dataset: includes selecting a dataset or focusing on a subset of variables or data samples on which discovery is to be performed*
3. *Data cleaning and preprocessing: includes basic operations, such as removing noise or outliers if appropriate, collecting the necessary information to model or account for noise (deciding on strategies for handling missing data fields, and accounting for time sequence information and known changes, as well as deciding DBMS issues, such as data types, schema, and mapping of missing and unknown values)*
4. *Data reduction and projection: includes finding useful features to represent the data, depending on the goal of the task (and using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data)*
5. *Choosing the function of data mining: includes deciding the purpose of the model derived by the data mining algorithm (e.g., summarization, classification, regression, and clustering)*
6. *Choosing the data mining algorithm(s): includes selecting method(s) to be used for searching for patterns in the data, such as deciding which models and parameters may be appropriate (e.g., models for categorical data are different from models on vectors over reals) and matching a particular data mining method with the overall criteria of the KDD process (e.g., the user may be more interested in understanding the model than in its predictive capabilities)*
7. *Data mining: includes searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, clustering, sequence modeling, dependency, and line analysis*
8. *Interpretation: includes interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualization of the extracted patterns, removing redundant or irrelevant patterns, and translating the useful ones into terms understandable by users*
9. *Using discovered knowledge: includes incorporating this knowledge into the performance system, taking actions based on the knowledge, or simply documenting it and reporting it to interested parties, as well as checking for and resolving potential conflicts with previously believed (or extracted) knowledge.*

THE KNOWLEDGE DISCOVERY USING BASE OF DATA




DATUM
DATUM
DATUM
DATUM
DATUM
DATUM
DATUM
DATUM
DATUM
DATUM
DATUM
DATUM
DATUM
DATUM
DATUM
INFORMATION
INFORMATION
INFORMATION
KNOWLEDGE
DECISION MAKING


COLLECTION PHASE
TRANSFORMATION PHASE
DATA-MINING PHASE
KNOWLEDGE
DESCRIPTIVE
PREDICTIVE
• CLINICAL CHART
• SURGICAL REPORTS
• RADIOLOGIC REPORTS
• SPIROMETRIC EVALUATION
• OTHER DATABASES
• .....
• ELECTRONIC STANDARDIZED REPOSITORY
• DATA QUALITY ASSESSMENT
• EXTRACTION OF DATA
• DATA CLUSTERING
• ANALYTIC ALGORITHMS
PROCESS OF KNOWLEDGE DISCOVERY FOR MEDICAL DATABASE

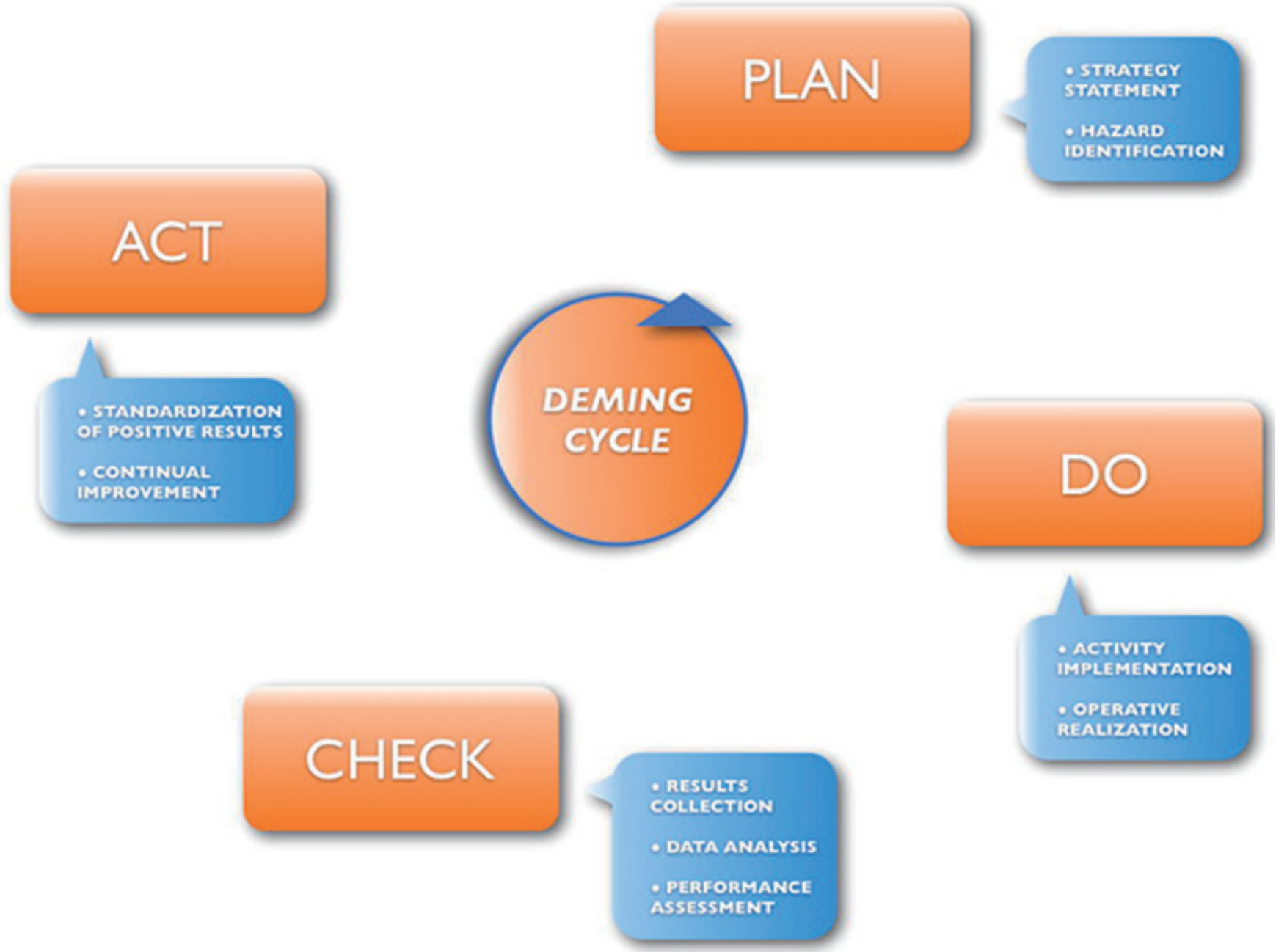


Figure 2 Deming Cycle.