DATA 301 Introduction to Data Science Alexander Dekhtyar

The Data Science Process

Data Science Process

The Data Science Process consists of the following steps:

1. Formulation of questions
2. Data Collection/Acquisition
3. Data Cleaning
4. Data Modeling
5. Data Analysis
6. Visualization and Presentation of Results
7. Analysis of Results

Note 1. This process often needs to be considered as a **cycle**, as shows in Figure 1. Upon the completion of Step 7: Analysis of Results, it is often useful to ask the *"Have we learned everything we need?"* question. The answer to this question is almost always a *"no"*.

Note 2. Data scientists often are not the only people working on parts of this process. Step 1: Formulation of questions is often conducted by other professionals working with data scientists (business analysts, executives, project/product managers, Principal Investigators, and so on). Similarly, the last step of the process, Analysis of results is often performed collaboratively by data scientists and the people who they are working for.

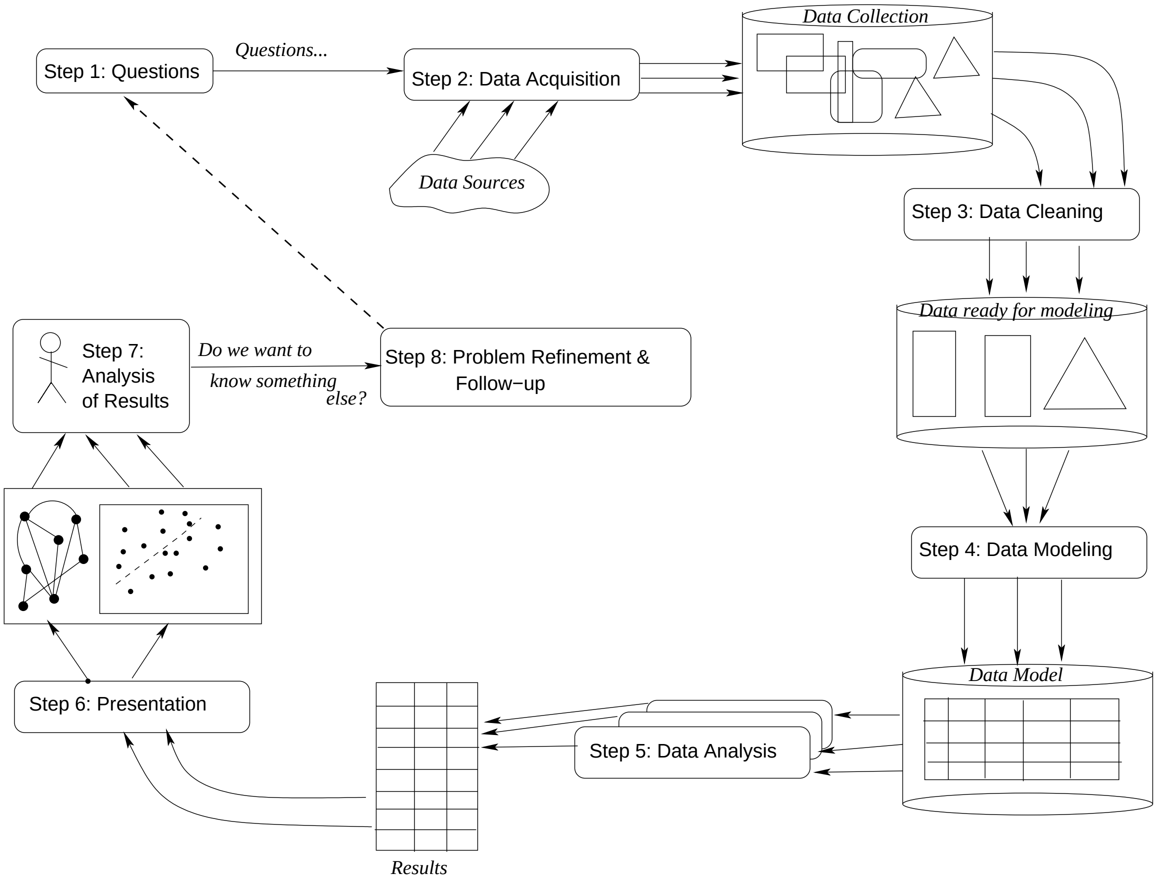

Figure 1: Data Science process as a cycle.

Step by Step

Step 1: Formulation of questions. First step of the data science process. On this step, the specific question that the data scientists must answer is formulated and, if necessary, negotiated.

Step 2: Data Acquisition (Collection). Based on the question given to the data scientists, on this step, the data scientists determine what data is needed to successfully answer it. The data is then collected from a variety of internal (e.g., corporate databases) or external (e.g., world wide web) sources.

Step 3: Data Cleaning. The collected data may contain

(a) data not needed to answer the question(s) studied

(b) data about same objects/events/entities collected from multiple sources

(c) missing data

(d) unrealiable, potentially incorrect data.

The data cleaning step of the data science process identifies these categories of data items in the data collection and performs appropriate manipulations with them. For example, unnecessary data is filtered out; data obtained from multiple sources about same objects can be merged together (integrated), unreliable data, if discovered, can be purged, or labelled as such.

Step 4: Data Modeling. The methods used on the Data Analysis step take input in specific format. Also, for these methods to produce high-quality output (this specifically refers to *machine learning* methods), the input data should contain the correct set of features. The data modeling step takes the cleaned up data, transforms it into the formats necessary for the analytical methods. It also, where needed, includes the feature extraction and feature selection procedures to fill in the inputs for the analytical methods with the *right* data.

Step 5: Data Analysis. The step which converts *data* into *knowledge*. A variety of analytical procedures, from data warehouse operations (roll-up, slice-and-dice, pivot, etc), to statistical methods (t-test, linear regression, factor analysis, multivariate analysis), to machine learning methods (classification, clustering, association rule mining, similarity analysis) can be deployed on the collected data on this stage.

Step 6: Visualization and Presentation of results. Often, the output of the methods used in the Data Analysis step is large and hard to immediately understand. During the Visualization and Presentation step, such output is turned into coherent, easy-to-observe representations.

Step 7: Analysis of results. Final step of the linear data science process - on this step the produced results are observed and discussed. Explanations are given to the observed phenomena, and reports are prepared.

Beyond these seven steps, data scientists need to be aware of two more steps of the business process into which their data science/discovery process is embedded.

Step 8: Goal refinement. (See Figure 1). Based on the results obtained in a single cycle of the data science process, data scientists and other professionals they work with ask the *Have we seen enough?* question. The answer to this question is most often a *"no"*. In such cases, based on the information obtained from the most recent analysis (and possibly from some prior stages), new questions are asked, serving as the starting point for the next round of the data science process.

Step 8': Action items. One of the key question following Step 7: Analysis of the results often is *"What do we do with this information?"*. The actual formulation of action items, i.e., things to do based on the completed data analysis, is usually beyond the scope of responibilities of a data scientist. However, the professionals who do make these decisions may trigger a new round of the data science process with the new sets of questions related to their ability to act upon the information produced on the current (and previous) rounds.