Data Science Process

Date:02/03/2022
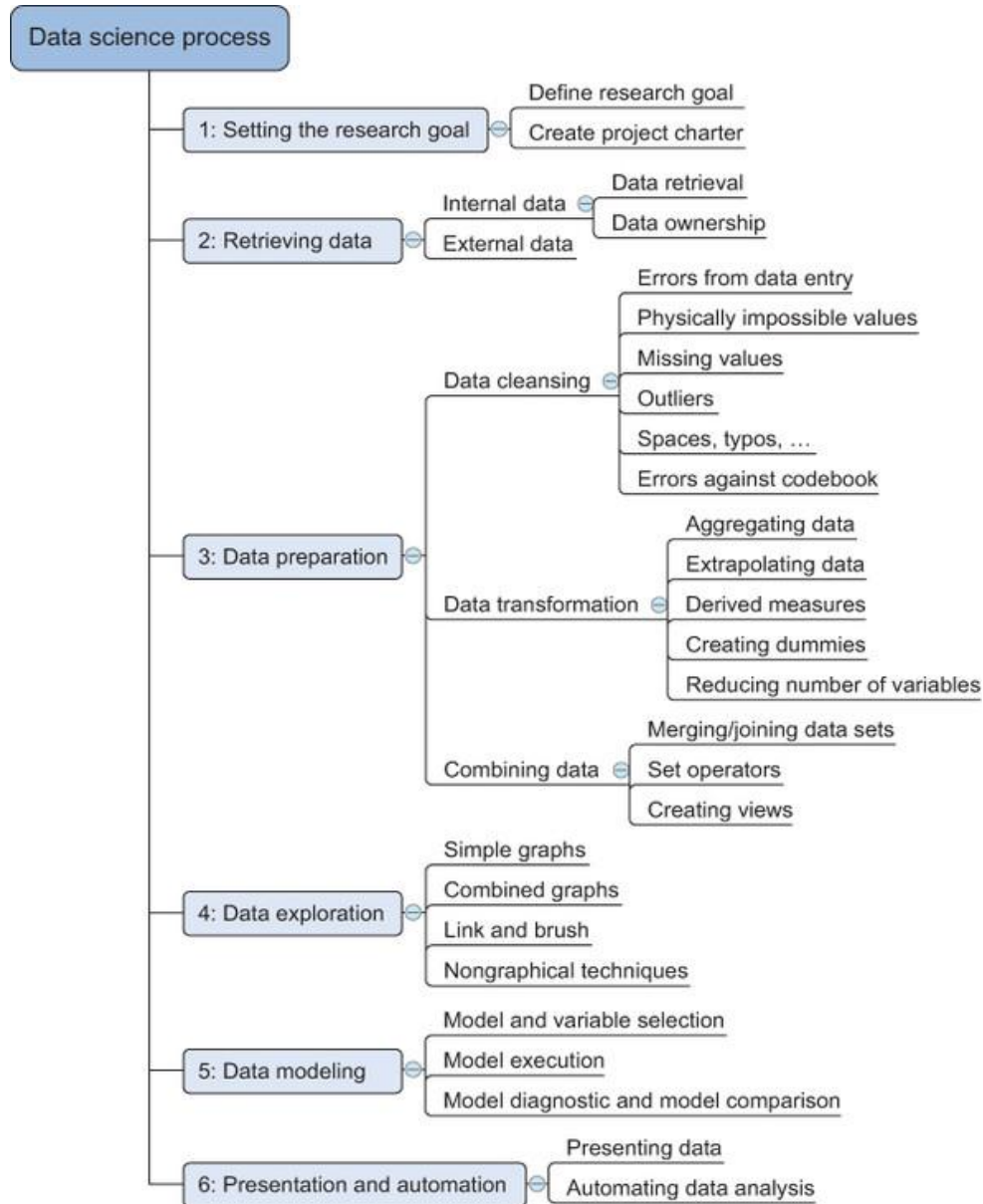
Dr.Inderjit Singh Dhanoa
Assistant Professor, CSE
Guru Nanak Dev Engineering College, Ludhiana

Outline

- Data Science Process

Data science process

- **1: Setting the research goal**
  - Define research goal
  - Create project charter

- **2: Retrieving data**
  - Internal data
    - Data retrieval
    - Data ownership
  - External data

- **3: Data preparation**
  - Data cleansing
    - Errors from data entry
    - Physically impossible values
    - Missing values
    - Outliers
    - Spaces, typos, …
    - Errors against codebook
  - Data transformation
    - Aggregating data
    - Extrapolating data
    - Derived measures
    - Creating dummies
    - Reducing number of variables
  - Combining data
    - Merging/joining data sets
    - Set operators
    - Creating views

- **4: Data exploration**
  - Simple graphs
  - Combined graphs
  - Link and brush
  - Nongraphical techniques

- **5: Data modeling**
  - Model and variable selection
  - Model execution
  - Model diagnostic and model comparison

- **6: Presentation and automation**
  - Presenting data
  - Automating data analysis

Research Goal

- The first step of this process is setting a *research goal*. The main purpose here is making sure all the stakeholders understand the *what*, *how*, and *why* of the project. In every serious project this will result in a project charter.

Research Goal

- The outcome should be a clear research goal, a good understanding of the context, well-defined deliverables, and a plan of action with a timetable. This information is then best placed in a project charter. The length and formality can, of course, differ between projects and companies. In this early phase of the project, people skills and business acumen are more important than great technical prowess, which is why this part will often be guided by more senior personnel.

Project Charter

A project charter requires teamwork, and your input covers at least the following:

- A clear research goal
- The project mission and context
- How you're going to perform your analysis
- What resources you expect to use
- Proof that it's an achievable project, or proof of concepts
- Deliverables and a measure of success
- A timeline

Retrieving data

- Data can be stored in many forms, ranging from simple text files to tables in a database. The objective now is acquiring all the data you need. This may be difficult, and even if you succeed, data is often like a diamond in the rough: it needs polishing to be of any use to you.

Start with data stored within the company

- Your first act should be to assess the relevance and quality of the data that's readily available within your company. Most companies have a program for maintaining key data, so much of the cleaning work may already be done. This data can be stored in official data repositories such as *databases*, *data marts*, *data warehouses*, and *data lakes* maintained by a team of IT professionals.

Don't be afraid to shop around

- If data isn't available inside your organization, look outside your organization's walls. Many companies specialize in collecting valuable information. For instance, Nielsen and GFK are well known for this in the retail industry. Other companies provide data so that you, in turn, can enrich their services and ecosystem. Such is the case with Twitter, LinkedIn, and Facebook.

Do data quality checks now to prevent problems later

- Expect to spend a good portion of your project time doing data correction and cleansing, sometimes up to 80%. The retrieval of data is the first time you'll inspect the data in the data science process. Most of the errors you'll encounter during the data-gathering phase are easy to spot, but being too careless will make you spend many hours solving data issues that could have been prevented during data import.

Cleansing, integrating, and transforming data

- The data received from the data retrieval phase is likely to be "a diamond in the rough." Your task now is to sanitize and prepare it for use in the modeling and reporting phase. Doing so is tremendously important because your models will perform better and you'll lose less time trying to fix strange output.

Cleansing data

- Data cleansing is a subprocess of the data science process that focuses on removing errors in your data so your data becomes a true and consistent representation of the processes it originates from.

Common Errors

- By "true and consistent representation" we imply that at least two types of errors exist. The first type is the *interpretation error*, such as when you take the value in your data for granted, like saying that a person's age is greater than 300 years. The second type of error points to *inconsistencies* between data sources or against your company's standardized values. An example of this class of errors is putting "Female" in one table and "F" in another when they represent the same thing: that the person is female. Another example is that you use Pounds in one table and Dollars in another.

Common Errors

Table 2.2. An overview of common errors

General solution	
Try to fix the problem early in the data acquisition chain or else fix it in the program.	
Error description	**Possible solution**
Errors pointing to false values within one data set	
Mistakes during data entry	Manual overrules
Redundant white space	Use string functions
Impossible values	Manual overrules
Missing values	Remove observation or value
Outliers	Validate and, if erroneous, treat as missing value (remove or insert)
Errors pointing to inconsistencies between data sets	
Deviations from a code book	Match on keys or else use manual overrules
Different units of measurement	Recalculate
Different levels of aggregation	Bring to same level of measurement by aggregation or extrapolation

Data Entry Errors

- Data collection and data entry are error-prone processes. They often require human intervention, and because humans are only human, they make typos or lose their concentration for a second and introduce an error into the chain. But data collected by machines or computers isn't free from errors either. Errors can arise from human sloppiness, whereas others are due to machine or hardware failure. Examples of errors originating from machines are transmission errors or bugs in the extract, transform, and load phase (ETL).

Other errors

- **Redundant whitespace**
- **FIXING CAPITAL LETTER MISMATCHES**
- **Impossible values and sanity checks**
- **Outliers**
- **Dealing with missing values**
- **Different units of measurement**