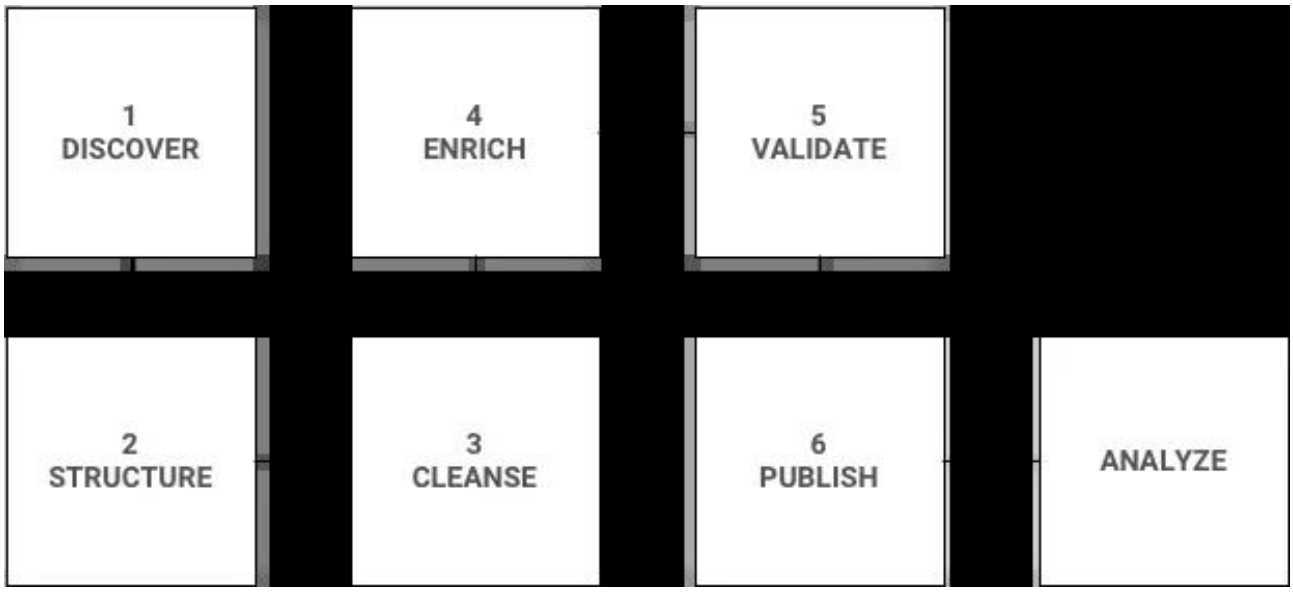
Data Munging

Idea In Short

Whatever you want to call it — data wrangling, data munging, or data transformation, the part of the Data Science Process sitting in between data acquisition and exploratory data analysis (EDA) is one of the core skills a data scientist must have. It includes a set of tasks you have to perform in order to understand your data and prep it for machine learning.

The process of manual data cleansing prior to analysis is known as Data Munging, also known as Data Wrangling. 80% of the time spent on data analytics is allocated to data munging, where the analysts manually clean the data before they could perform any analysis1. Data munging is often a time consuming, laborious, and disjointed process gets in the way of extracting true value and potential from data. According to O'Reilly's 2016 data science Salary Survey, 69% of data scientists will spend a significant amount of time in their day-to-day dealing with basic exploratory data analysis, while 53% spend time cleaning their data2 as part of the data on boarding process.



Data Munging Visualization

The challenges associated with Data Wrangling extend beyond technology:

Users

The core idea of data wrangling is that the people who know the data best should be exploring and preparing that data. This means business analysts, line-of-business users, and managers (among others) are the intended stakeholders that should be involved (directly or indirectly) in the data wrangling process.

Data

A growing variety of data sources can now be analysed, but historically, analysts didn't have the right tools to understand, clean, and organise this data in the appropriate format. Much of the data business analysts must deal with today comes in a growing variety of shapes and sizes that are either too big or too complex to work with in traditional self-service tools such as Excel. Data wrangling process is specifically designed to handle diverse, complex data at any scale. Additionally, a growing amount of analysis occurs in environments where the schema of data is not defined or known ahead of time. This means the analyst wrangling the data is determining how it can be leveraged for analysis as well as the schema required to perform that analysis.

Use Cases

The use cases that require data wrangling tend to be somewhat exploratory in nature and are often conducted by small teams or departments prior to being rolled out across the organisation. Analysts typically try to work with a new data source or new combination of data sources for an analytics initiative. Data wrangling makes existing analytics processes more efficient and accurate as users can always have their eyes on their data as they prepare it. To better cope up with the data wrangling workload, here are the 6 essential steps that any Data Analyst / Scientist should be proficient with:

1. **Discover:** Learn what's in your raw dataset to think ahead about the best approach for your analytic explorations. This allows you to understand unique elements of the data such as outliers and value distribution to inform the analysis process
2. **Structure:** This is a critical step because your data comes in all shapes and sizes, and it is up to you to decide the best format to visualize and explore it. Separating,

blending, and un-nesting are all important actions in this step

3. **Cleanse:** This step is essential to standardizing your data to ensure that all inconsistencies (such as null and misspelled values) are addressed. Other data may need to be standardized to a single format such as state abbreviations
4. **Enrich:** At this point, you've gotten a clear handle on your data – what else could you add to provide more value to your analysis? Enrichment is often about joins and complex derivations
5. **Validate:** Verify if you've caught all of the data quality and consistency issues and go back to address anything you may have missed. Validation should be done on multiple dimensions
6. **Publish:** This is where you can download and deliver the results of your wrangling effort to downstream analytics tools. Once you've published your data it's time to move onto the next step: analytics!

Data wrangling is an essential part of the data science role — and if you gain data wrangling skills and become proficient at it, you'll quickly be recognized as somebody who can contribute to cutting-edge data science work and who can hold their own as a data professional.

Summary

Data wrangling may not be as high profile as other steps in the data science process such as model selection, but this doesn't mean it's not important. Indeed, the importance of data wrangling shouldn't be overlooked. Many projects live or die with this step being completed properly.