

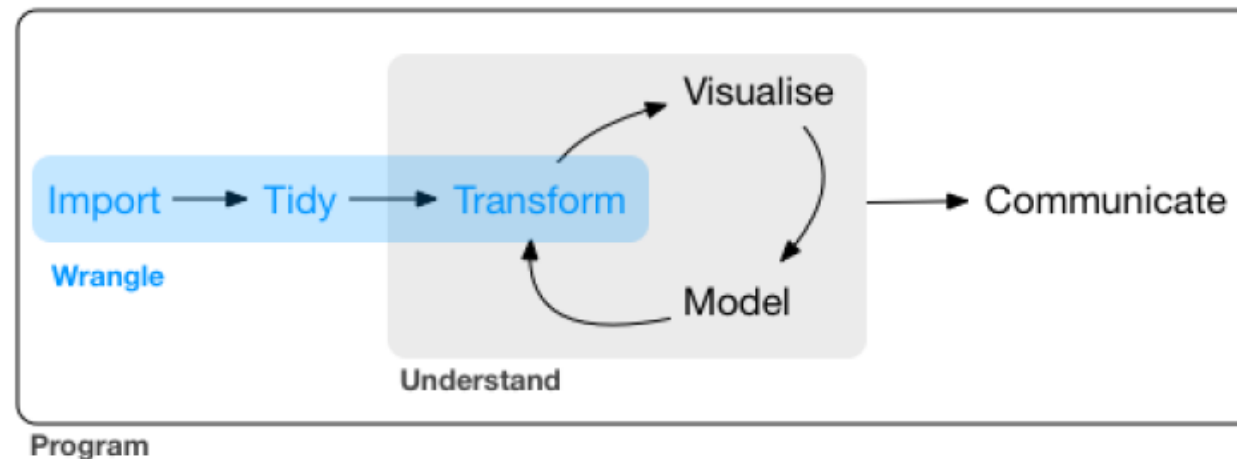


DATA WRANGLING



6.3. Theme 3: Wrangle. (1 week)

The goal of this theme is to master the basic tools of data wrangling.



6.3.1. *Data Import & Tidy Data.* Outcomes: At the end of this study unit, the student should be able to:

- get data from disk and into Python
- understand underlying principles of tidy data, and how to get your data into a tidy form.

3.3 Data Wrangling

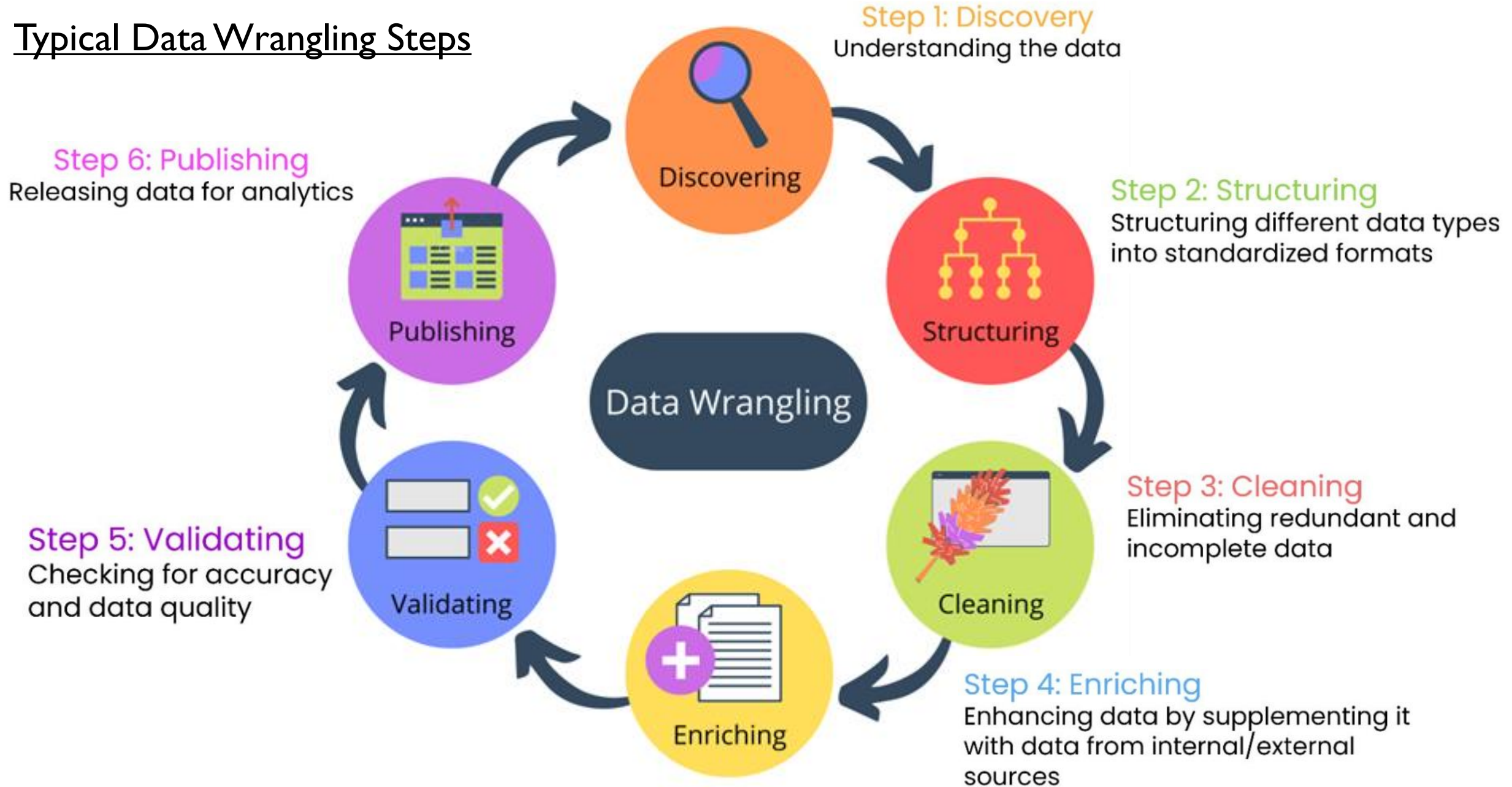
The process of cleaning, organizing, and transforming "raw" data with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.




TRANSFORMATION



Typical Data Wrangling Steps



The 3 Fundamental Attributes of Tidy Data



A diagram illustrating the concept of variables in tidy data. It shows a table with four columns: 'country', 'year', 'cases', and 'population'. Each column has a vertical double-headed arrow next to it, indicating that each column represents a variable.

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174604898
China	1999	212258	127291272
China	2000	213766	128042583

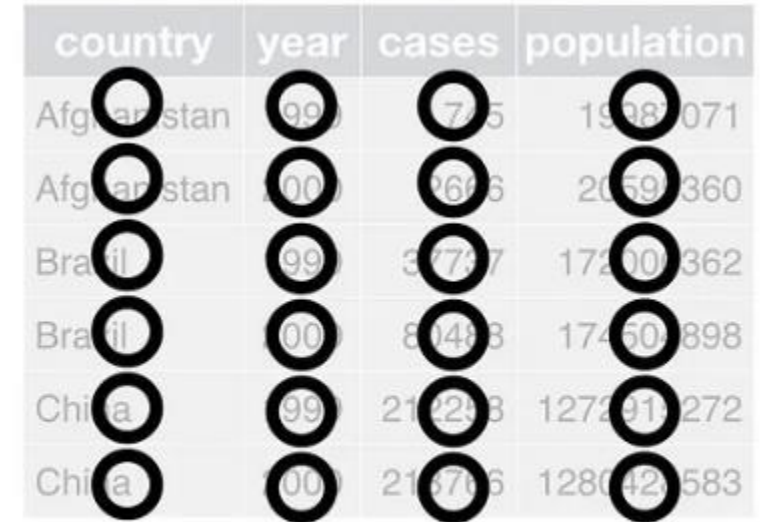
variables



A diagram illustrating the concept of observations in tidy data. It shows a table with four columns: 'country', 'year', 'cases', and 'population'. Each row has a horizontal double-headed arrow next to it, indicating that each row represents an observation.

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174604898
China	1999	212258	127291272
China	2000	213766	128042583

observations



A diagram illustrating the concept of values in tidy data. It shows a table with four columns: 'country', 'year', 'cases', and 'population'. Each cell in the table contains a value, represented by a circle around the text.

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174604898
China	1999	212258	127291272
China	2000	213766	128042583

values

1. Each variable is a column
2. Each observation is a row
3. Each cell contains a single value

Example from Wikipedia

Starting data

Name	Phone	Birth date	State
John, Smith	445-881-4478	August 12, 1989	Maine
Jennifer Tal	+1-189-456-4513	11/12/1965	Tx
Gates, Bill	(876)546-8165	June 15, 72	Kansas
Alan Fitch	5493156648	2-6-1985	Oh
Jacob Alan	156-4896	January 3	Alabama
John, Smith	445-881-4478	August 12, 1989	Maine



Result

Name	Phone	Birth date	State
John Smith	445-881-4478	1989-08-12	Maine
Jennifer Tal	189-456-4513	1965-11-12	Texas
Bill Gates	876-546-8165	1972-06-15	Kansas
Alan Fitch	549-315-6648	1985-02-06	Ohio



Let's dive into the notebook, `Data Wrangling (part I).ipynb...`