**Data Wrangling Steps**

**Step 1: Gather Data**

Depending on where you find your data and what format it's in, the steps of the gathering process can vary.

Data sources:

* Files (Manually/ Programmatically)
* Database
* Scraped off of website
* **Application programming interface (API)**which allows us to programmatically access data

**Step 2: Assess Data**

### ***Data Quality***

Low quality data is commonly referred to as **dirty data**. Dirty data has issues with its content.

Imagine you had a table with two columns: Name and Height, like below:

| **Name** | **Height** |
| --- | --- |
| Jane | 55 inches |
| Juan |  |
| Amalie | 145 centimetres |
| Kwasi | -50 inches |

### Common Data Quality Issues:

* **Missing data**, like the missing height value for Juan.
* **Invalid data**, like a cell having an impossible value, e.g., like negative height value for Kwasi. Having "inches" and "centimetres" in the height entries is technically invalid as well, since the datatype for height becomes a string when those are present. The datatype for height should be integer or float.
* **Inaccurate data**, like Jane actually being 58 inches tall, not 55 inches tall.
* **Inconsistent data**, like using different units for height (inches and centimetres).

We'll go over more tips and tricks to identify data quality issues and categorize them in the third lesson of the course.

Data quality is a perception or an assessment of data's fitness to serve its purpose in a given context. Unfortunately, that’s a bit of an evasive definition but it gets to something important: there are no hard and fast rules for data quality. One dataset may be high enough quality for one application but not for another.

### Tidiness

Untidy data is commonly referred to as **"messy" data**. Messy data has issues with its structure.

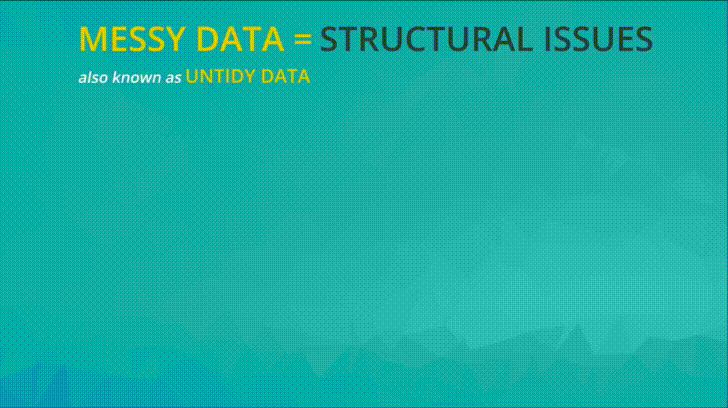
Tidy data is a relatively new concept coined by statistician, professor, and all-round data expert [**Hadley Wickham**](http://hadley.nz/). I’m going to take a quote from his excellent [**paper**](https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html) on the subject:

It is often said that 80% of data analysis is spent on the cleaning and preparing data. And it’s not just a first step, but it must be repeated many times over the course of analysis as new problems come to light or new data is collected. To get a handle on the problem, this paper focuses on a small, but important, aspect of data cleaning that I call data tidying: structuring datasets to facilitate analysis.

...

A dataset is messy or tidy depending on how rows, columns, and tables are matched up with observations, variables, and types. In tidy data:

* Each variable forms a column.
* Each observation forms a row.
* Each type of observational unit forms a table.



Tidy data animation from Lesson 3: Assessing Data

**Step 3: Clean Data**

**Cleaning** means acting on the assessments we made to improve quality and tidiness.

#### Improving Quality

Improving quality doesn’t mean changing the data to make it say something different—that's data fraud.

Consider the animals DataFrame, which has headers for name, body weight (in kilograms), and brain weight (in grams). The last five rows of this DataFrame are displayed below:

animals.tail()

|  | **Animal** | **Body Weight (kg)** | **Brain Weight (g)** |
| --- | --- | --- | --- |
| 24 | Chimpanzee | 52.160 | 440.0 |
| 25 | Mouse | 230.000 | 0.4 |
| 26 | Apple | 0.100 | NaN |
| 27 | Brachiosaurus | 87000.000 | NaN |
| 28 | Mole | 0.122 | 3.0 |

### Examples of Improving Quality

* Correcting when inaccurate, like correcting the mouse's body weight to 0.023 kg instead of 230 kg
* Removing when irrelevant, like removing the row with "Apple" since an apple is a fruit and not an animal
* Replacing when missing, like filling in the missing value for brain weight for Brachiosaurus
* Combining, like concatenating the missing rows in the more\_animals DataFrame displayed below

more\_animals

|  | **Animal** | **Body Weight (kg)** | **Brain Weight (g)** |
| --- | --- | --- | --- |
| 0 | Cat | 3.3 | 25.6 |
| 1 | Giraffe | 529.0 | 680.0 |
| 2 | Gorilla | 207.0 | 406.0 |
| 3 | Human | 62.0 | 1320.0 |
| 4 | African Elephant | 6654.0 | 5712.0 |

**Improving Tidiness**

Improving tidiness means transforming the dataset so that each variable is a column, each observation is a row, and each type of observational unit is a table. There are special functions in pandas that help us do that. We'll dive deeper into those in Lesson four of this course.

### The Programmatic Data Cleaning Process

1. Define
2. Code
3. Test

**Defining** means defining a data cleaning plan in writing, where we turn our assessments into defined cleaning tasks. This plan will also serve as an instruction list so others (or us in the future) can look at our work and reproduce it.

**Coding** means translating these definitions to code and executing that code.

**Testing** means testing our dataset, often using code, to make sure our cleaning operations worked.

**Step 4: Reassess & Iterate**

### Data Wrangling is an Iterative Process

We've gathered, assessed, and cleaned our data. Are we done? No. After cleaning, we always reassess and then iterate on any of the steps if we need to. If we're happy with the quality and tidiness of our data, we can end our wrangling process and move on to storing our clean data, or analyzing, visualizing, or modeling it.

* Sometimes we realize we need to gather more data.
* Sometimes we miss assessments. It's hard to catch everything on the first go, and it's also very common to find new issues as you're fixing the ones you've already identified.
* Sometimes our cleaning operations don't work as we intended.

Once we go through each step once, we can revisit any step in the process at any time.

### Storing Data (Optional)

After reassessing your data and revisiting any steps of the data wrangling process deemed necessary, storing your cleaned data can be the next logical step. Storing data is important if you need to use your cleaned data in the future.

Storing data isn't always necessary, though. Sometimes the Jupyter Notebook that you gathered, assessed, cleaned, analyzed, and visualized your data in, plus the original data files is good enough. Sometimes the analysis and visualization are the final products and you won't be using the cleaned data any further. If you ever want to reproduce the analysis, the Jupyter Notebook suffices.

Storing your data, in files and databases, for example, will be covered in detail in a later lesson in the course.

### Data Wrangling is an Iterative Process

We've gathered, assessed, and cleaned our data. Are we done? No. After cleaning, we always reassess and then iterate on any of the steps if we need to. If we're happy with the quality and tidiness of our data, we can end our wrangling process and move on to storing our clean data, or analyzing, visualizing, or modeling it.

* Sometimes we realize we need to gather more data.
* Sometimes we miss assessments. It's hard to catch everything on the first go, and it's also very common to find new issues as you're fixing the ones you've already identified.
* Sometimes our cleaning operations don't work as we intended.

Once we go through each step once, we can revisit any step in the process at any time.

### Storing Data (Optional)

After reassessing your data and revisiting any steps of the data wrangling process deemed necessary, storing your cleaned data can be the next logical step. Storing data is important if you need to use your cleaned data in the future.

Storing data isn't always necessary, though. Sometimes the Jupyter Notebook that you gathered, assessed, cleaned, analyzed, and visualized your data in, plus the original data files is good enough. Sometimes the analysis and visualization are the final products and you won't be using the cleaned data any further. If you ever want to reproduce the analysis, the Jupyter Notebook suffices.

Storing your data, in files and databases, for example, will be covered in detail in a later lesson in the course.