

# Text Mining Toolkit

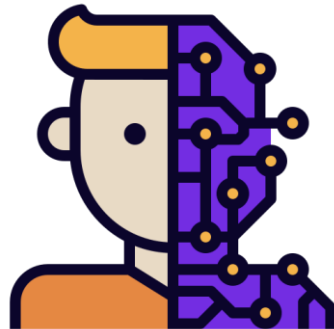
in R



# What is Text Mining?

---

**Text Mining** combines *Machine Learning* and *Natural Language Processing* (NLP) to draw meaning from unstructured text documents

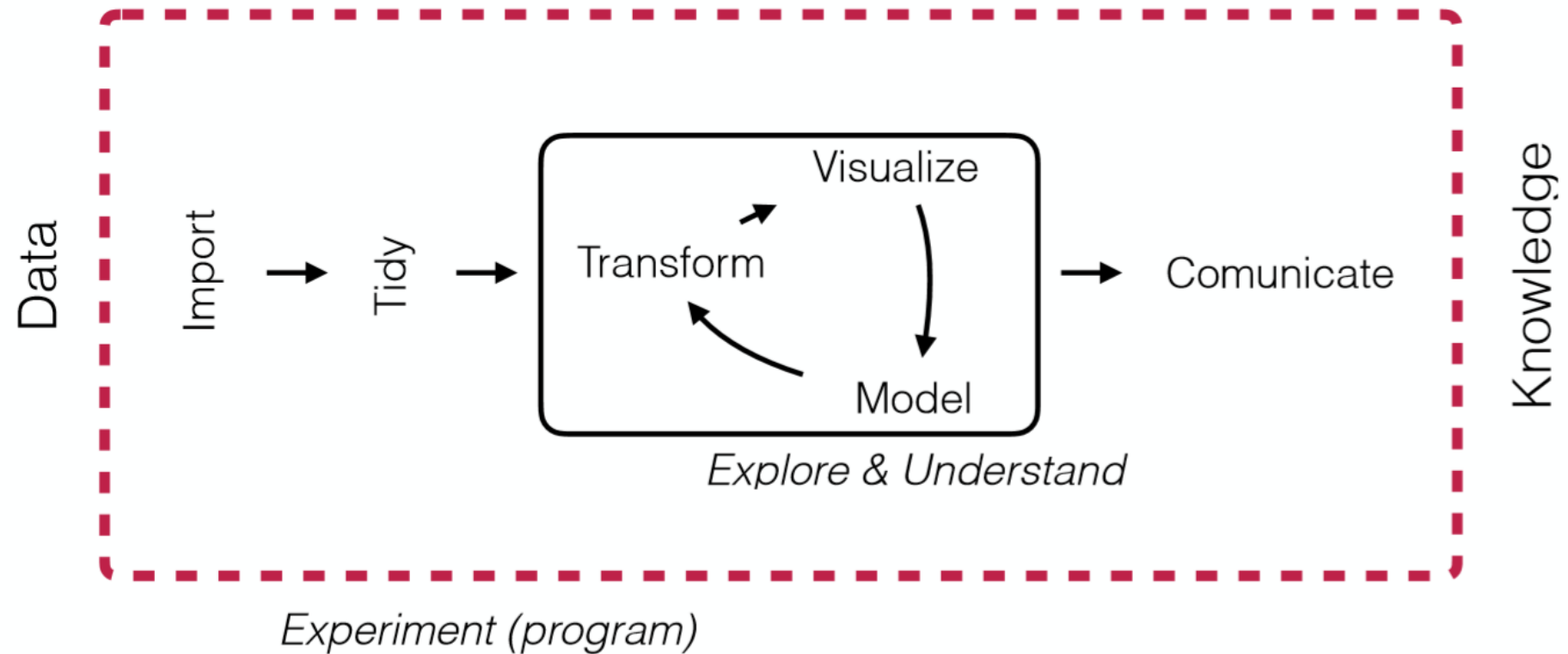


Machine Learning

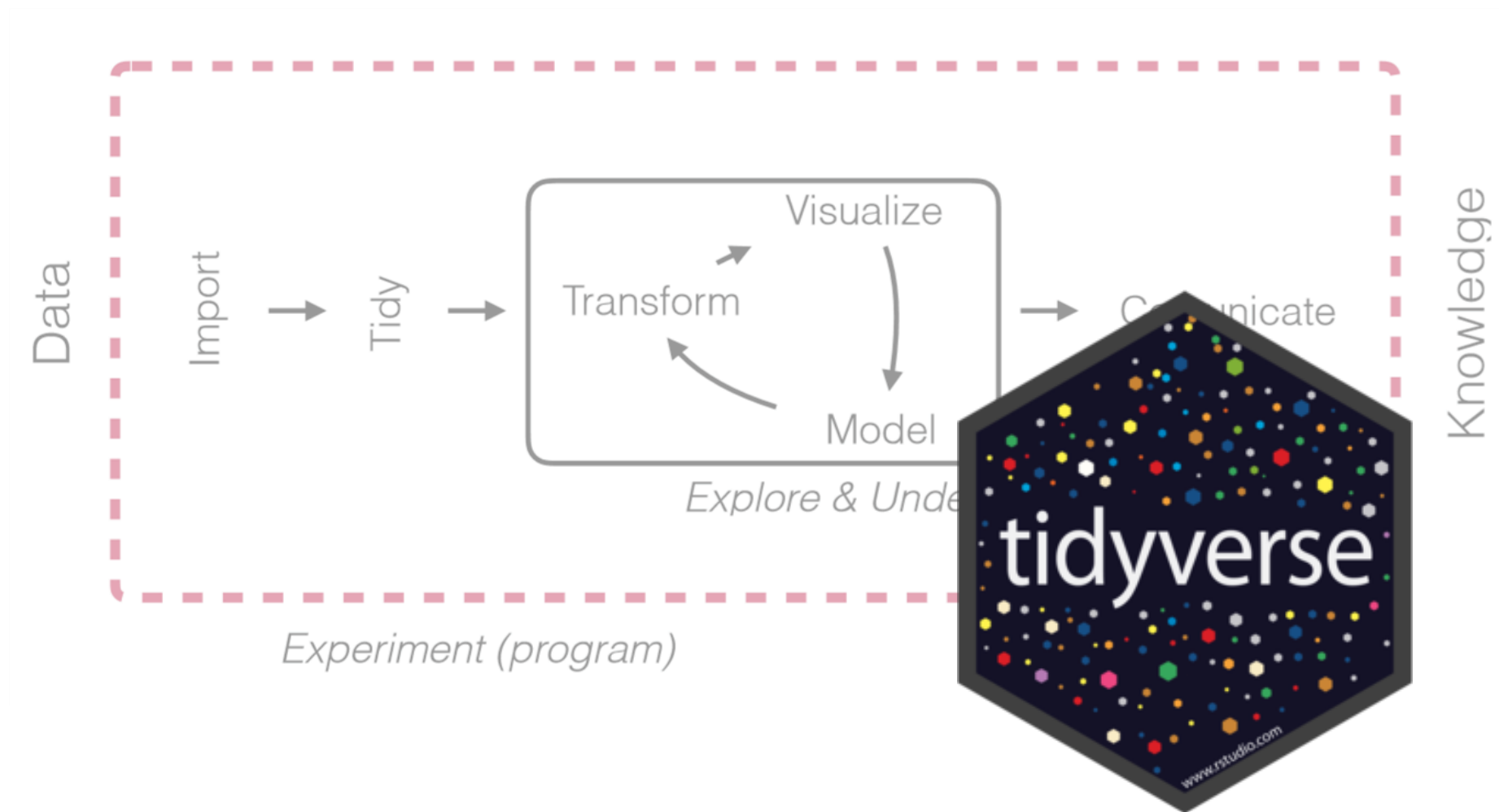


Natural Language Processing

# Data Science Framework



# Data Science Framework



# Tidyverse

---

## Why **tidy** data?

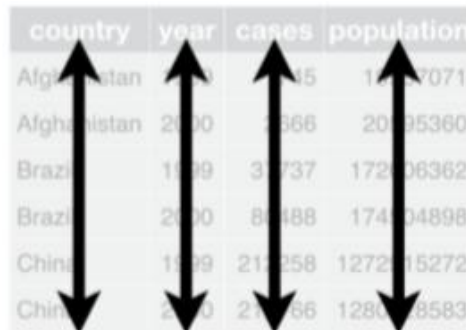
- ✓ Makes data analysis easy
- ✓ Is easy to model, visualize and transform
- ✓ Facilitate the creative proces

# Tidyverse

## How to **tidy** data?


There are three interrelated rules which make a dataset tidy:

1. Each variable must have its own column.
2. Each observation must have its own row.
3. Each value must have its own cell.



country	year	cases	population
Afghanistan	2000	2966	19995360
Afghanistan	2000	2966	20005360
Brazil	1999	31737	172006362
Brazil	2000	80488	174004898
China	1999	212258	1272015272
China	2000	212266	1280008583

variables



country	year	cases	population
Afghanistan	2000	2966	19995360
Afghanistan	2000	2966	20005360
Brazil	1999	31737	172006362
Brazil	2000	80488	174004898
China	1999	212258	1272015272
China	2000	212266	1280008583

observations



country	year	cases	population
Afghanistan	2000	2966	19995360
Afghanistan	2000	2966	20005360
Brazil	1999	31737	172006362
Brazil	2000	80488	174004898
China	1999	212258	1272015272
China	2000	212266	1280008583

values

# Tidyverse

---

## Example

There are three variables in this data set. What are they?

	Pregnant	Not Pregnant
Male	0	5
Female	1	4

# Tidyverse

---

## Example

Pregnant	Sex	Freq
no	female	4
no	male	5
yes	female	1
yes	male	0



# LET'S GET PRACTICAL

