Introduction

An overview of data processing and the NumPy library.

In the **Data Manipulation** section, you will learn how to perform data manipulation using NumPy.

A. Data Processing

When asked about Google's model for success, Peter Norvig, the director of research at Google, famously stated,

> *"We don't have better algorithms than anyone else; we just have more data."*

Though probably an understatement (given the amount of talent employed at Google), the quote does provide a sense of just how vital data is to having successful outcomes.

People normally discuss the importance of data in the context of machine learning. No matter how sophisticated a machine learning model is, it will not perform well unless it has a reasonable amount of data to train on. On the other hand, given a large and diverse set of training data, a good deep learning model will significantly outperform non-deep learning algorithms.

However, data is not just limited to machine learning. Companies use data to identify customer trends, political parties use data to determine which demographics they should target, sports teams use data to analyze players, etc.

Player	PA	AVG	OBP	SLG	OPS	K%	BB/K	ISO	Spd	BABIP	wRC	wRC+	wOBA
Joe Mauer	576	0.261	0.363	0.389	0.752	16.1%	0.85	0.128	3.6	0.301	72	102	0.327
Carlos Santana	688	0.259	0.366	0.498	0.865	14.4%	1.00	0.239	3.5	0.258	110	132	0.37
John Jaso	432	0.268	0.353	0.413	0.766	17.1%	0.61	0.145	3.0	0.314	57	111	0.335
Mike Napoli	645	0.239	0.335	0.465	0.800	30.1%	0.40	0.226	3.7	0.296	89	113	0.343
Albert Pujols	650	0.268	0.323	0.457	0.780	11.5%	0.65	0.189	2.5	0.260	83	111	0.331
Miguel Cabrera	679	0.316	0.393	0.563	0.956	17.1%	0.65	0.247	1.3	0.336	125	152	0.399
Mark Teixeira	438	0.204	0.292	0.362	0.654	24.0%	0.45	0.158	2.4	0.238	40	76	0.287
Joey Votto	677	0.326	0.434	0.550	0.985	17.7%	0.90	0.225	3.9	0.366	133	158	0.413
Billy Butler	274	0.284	0.336	0.416	0.752	15.3%	0.50	0.132	1.1	0.320	34	105	0.324
Ryan Howard	362	0.196	0.257	0.453	0.710	31.5%	0.24	0.257	0.8	0.205	37	83	0.298
James Loney	366	0.265	0.307	0.397	0.703	10.1%	0.43	0.131	1.4	0.275	38	89	0.302
Prince Fielder	370	0.212	0.292	0.334	0.626	17.0%	0.51	0.123	0.9	0.235	31	65	0.276
Adrian Gonzalez	633	0.285	0.349	0.435	0.784	18.5%	0.47	0.150	1.4	0.328	83	112	0.334

Example baseball data used in sabermetrics (https://en.wikipedia.org/wiki/Sabermetrics). The concept was popularized by the 2011 film, *Moneyball*.

The universal usage of data makes *data processing*, the act of converting raw data into a meaningful form, an essential skill to have.

B. NumPy

Many scenarios involve mostly numeric datasets. For example, medical data contains many numeric metrics, such as height, weight, and blood pressure. Furthermore, the majority of neural networks use input data that is either numeric or has been converted to a numeric form.

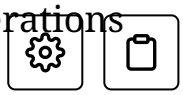
When we deal with numeric data, the best Python library to use is NumPy (http://www.numpy.org/). The NumPy library allows us to perform many operations on numeric data, and convert the data to more usable forms.

```
import numpy as np  # import the NumPy library

# Initializing a NumPy array
arr = np.array([-1, 2, 5], dtype=np.float32)

# Print the representation of the array
print(repr(arr))
```

In the following chapters, you'll learn all the necessary NumPy operations for data manipulation.



Back

Overview

Next

NumPy Arrays

Completed

3% completed, meet the criteria and claim your course certificate!

Buy Certificate

Report an Issue

Ask a Question (https://discuss.educative.io/tag/introduction__data-manipulation-with-numpy__machine-learning-for-software-engineers)