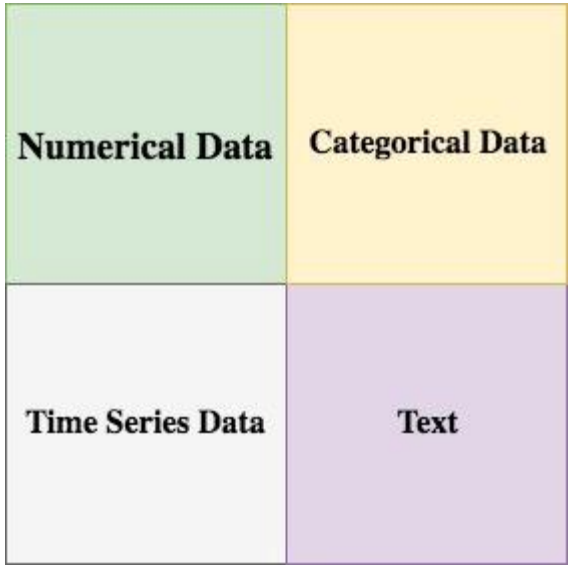
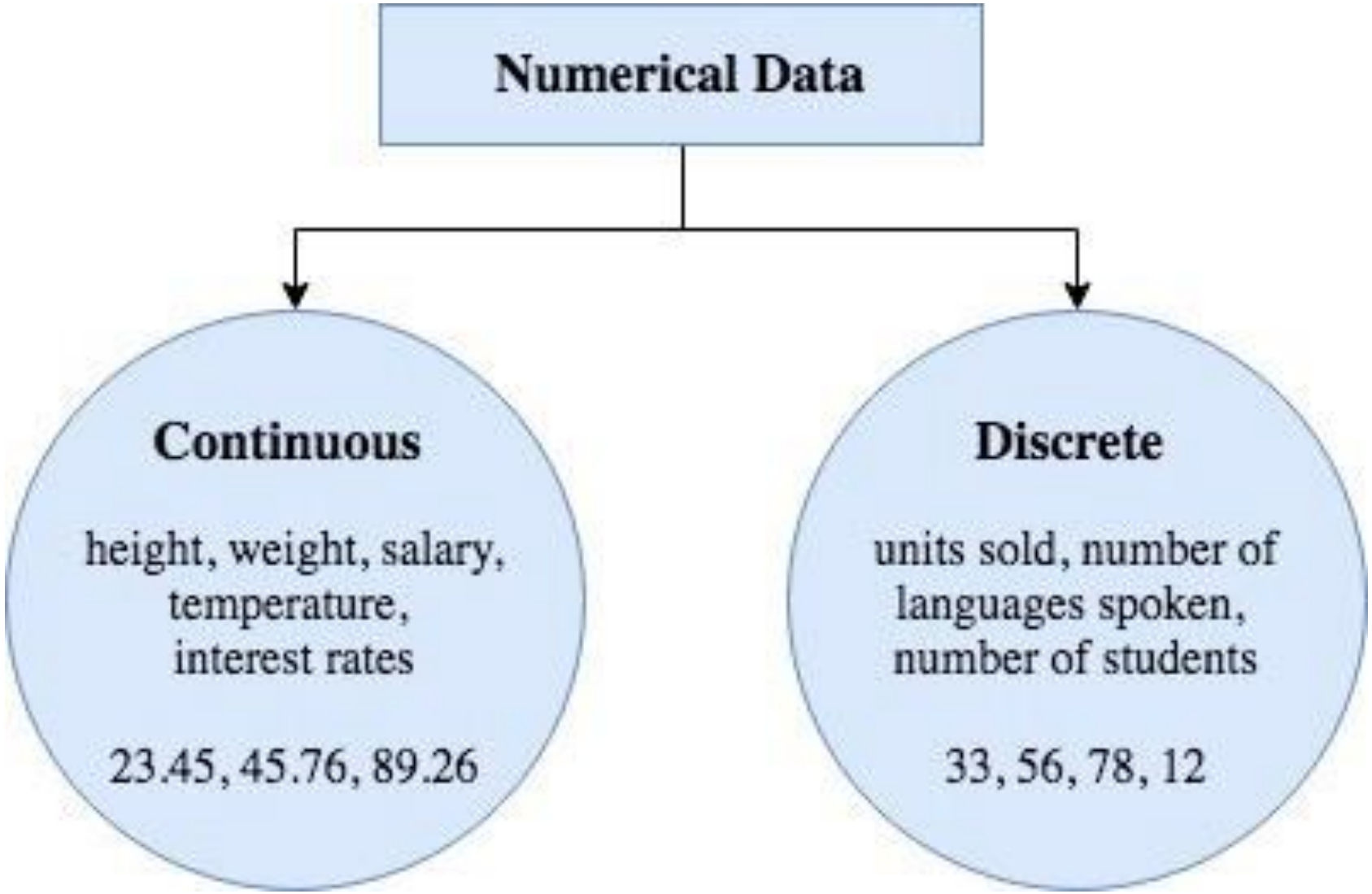
Data Types

- Most data can be categorized into 4 basic types from a Machine Learning perspective: numerical data, categorical data, time-series data, and text.



Numerical Data

- Numerical data is any data where data points are exact numbers. Statisticians also might call numerical data, quantitative data. This data has meaning as a **measurement** such as house prices or as a count, such as a number of residential properties in Los Angeles or how many houses sold in the past year.
- Numerical data can be characterized by continuous or discrete data. Continuous data can assume any value within a range whereas discrete data has distinct values.


Numerical Data
Continuous
height, weight, salary, temperature, interest rates
23.45, 45.76, 89.26
Discrete
units sold, number of languages spoken, number of students
33, 56, 78, 12

Categorical Data

- Categorical data represents characteristics, such as a hockey player’s position, team, hometown. Categorical data can take numerical values. For example, maybe we would use 1 for the colour red and 2 for blue. But these numbers don’t have a mathematical meaning. That is, we can’t add them together or take the average.
- In the context of super classification, categorical data would be the class label. This would also be something like if a person is a man or woman, or property is residential or commercial.

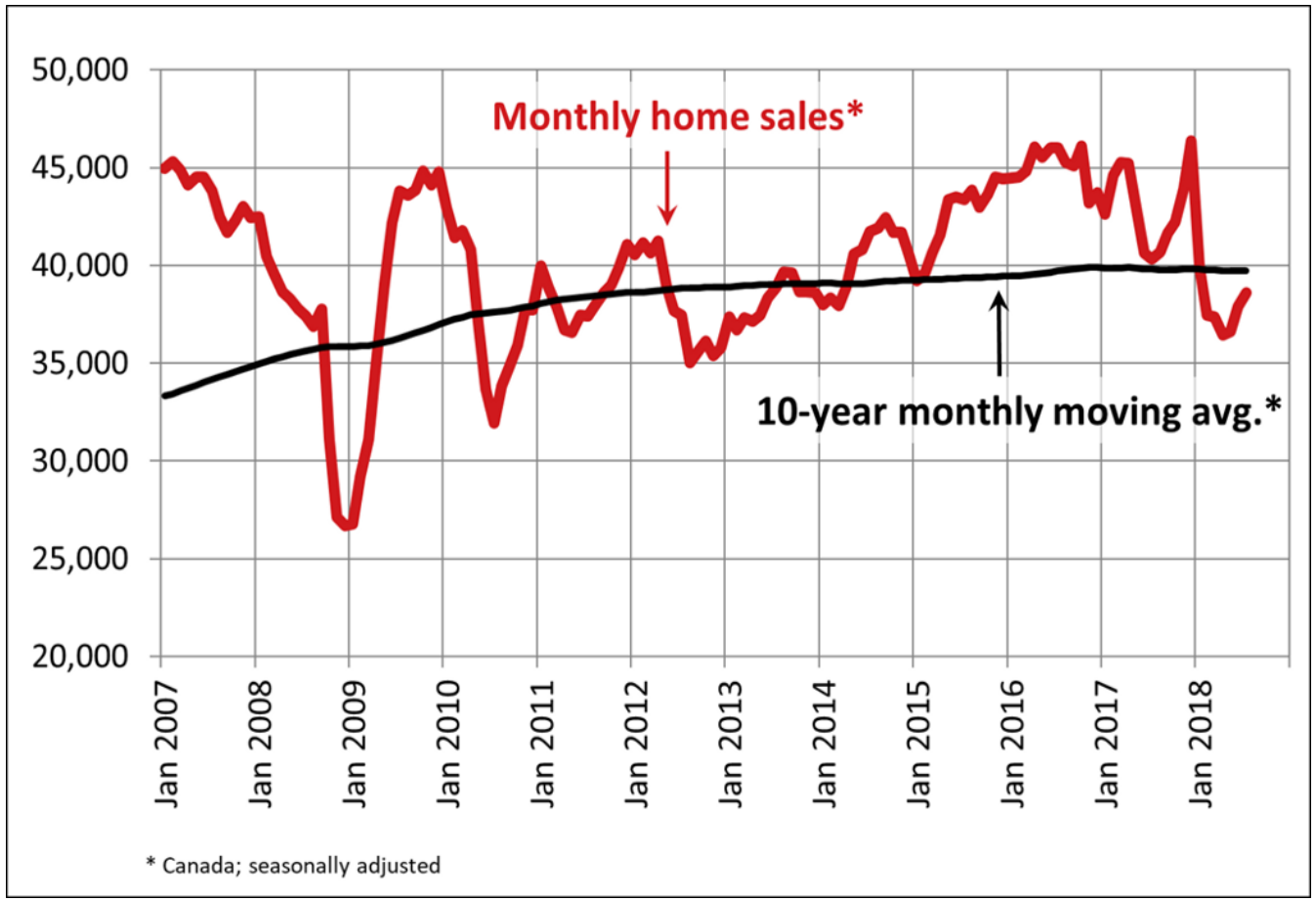
beginner

intermediate

advanced

Time Series Data

- Time series data is a sequence of numbers collected at regular intervals over some period of time. It is very important, especially in particular fields like finance. Time series data has a temporal value attached to it, so this would be something like a date or a timestamp that you can look for trends in time.


50,000
45,000
40,000
35,000
30,000
25,000
20,000
Monthly home sales*
10-year monthly moving avg.*
Jan 2007
Jan 2008
Jan 2009
Jan 2010
Jan 2011
Jan 2012
Jan 2013
Jan 2014
Jan 2015
Jan 2016
Jan 2017
Jan 2018
* Canada; seasonally adjusted

Text

- Text data is basically just words. A lot of the time the first thing that you do with text is you turn it into numbers using some interesting functions like the bag of words formulation.

Plotting data with Python :

- As mentioned above, Python has several good packages to plot the data and among them **Matplotlib** is the most prominent one. **Seaborn** is also a great package which offers a lot more appealing plot and even it uses matplotlib as its base layer.