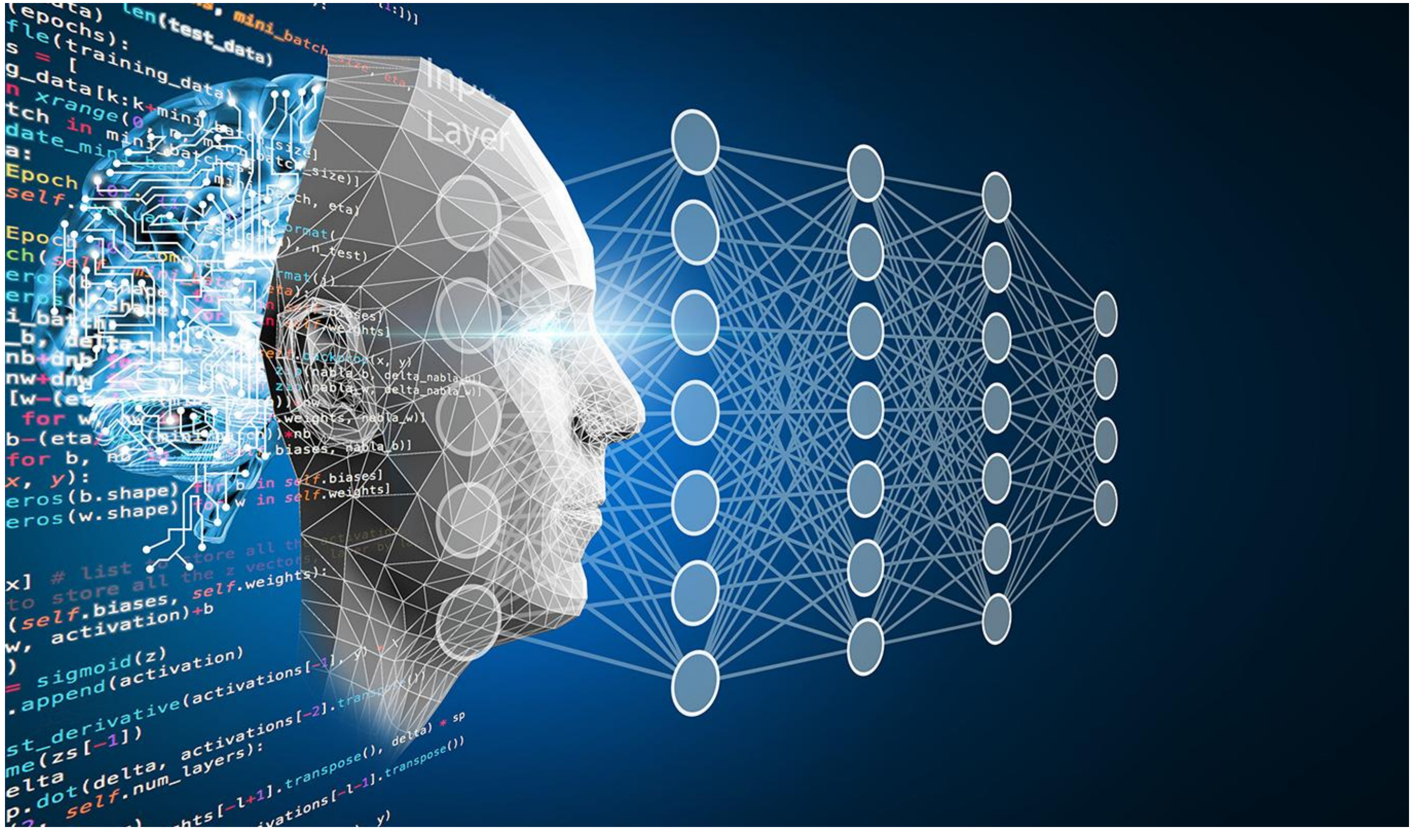


Foundations of Data Literacy and Data Science

Prof. Dr. Hamid Mostofi

mostofidarbani@tu-berlin.de

Learning Objectives

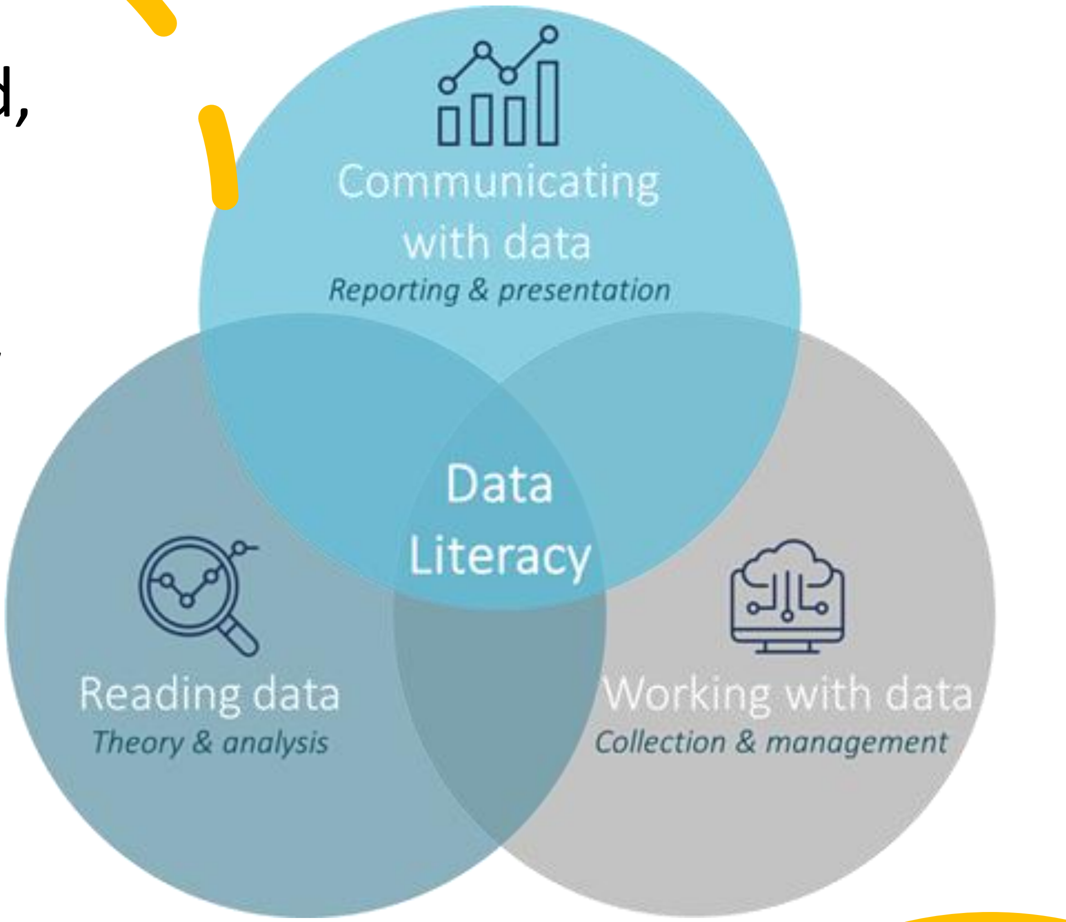
Understanding the basic concepts of data, and ability to work with datasets, AI tools

Learning concepts and skills in scientific programming and statistical inference using Python

System thinking, Computational and inferential thinking skills

Improved decision-making skills using data analytics and AI

- **Data Literacy** refers to the ability to read, understand, analyse, and communicate data effectively. It involves knowing how to collect, interpret, and use data in decision-making.



- **AI Literacy** is the ability to understand, use, and critically evaluate artificial intelligence technologies to make informed decisions about how AI tools impact the work, research, and society.

Main References

Data8 program

at UC Berkeley

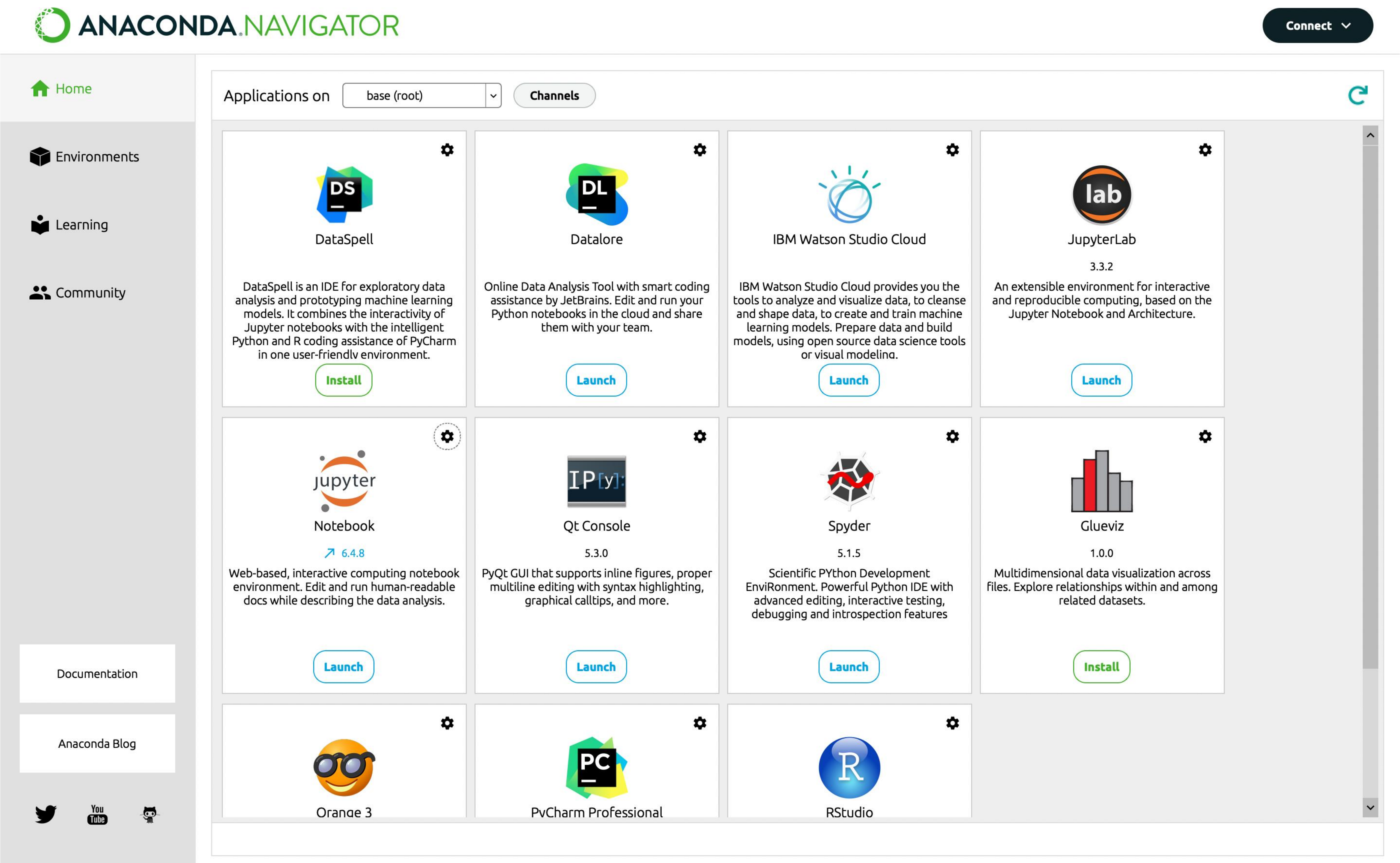
Some Python Libraries like Pandas and Scikit Learn

Data8

Pandas

Scikit Learn

Cloud-based Python Programming Interface



DataSpell

DataSpell is an IDE for exploratory data analysis and prototyping machine learning models. It combines the interactivity of Jupyter notebooks with the intelligent Python and R coding assistance of PyCharm in one user-friendly environment.

Datalore

Online Data Analysis Tool with smart coding assistance by JetBrains. Edit and run your Python notebooks in the cloud and share them with your team.

IBM Watson Studio Cloud

IBM Watson Studio Cloud provides you the tools to analyze and visualize data, to cleanse and shape data, to create and train machine learning models. Prepare data and build models, using open source data science tools or visual modeling.

JupyterLab

3.3.2

An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

Notebook

↗ 6.4.8

Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

Qt Console

5.3.0

PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

Spyder

5.1.5

Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

Glueviz

1.0.0

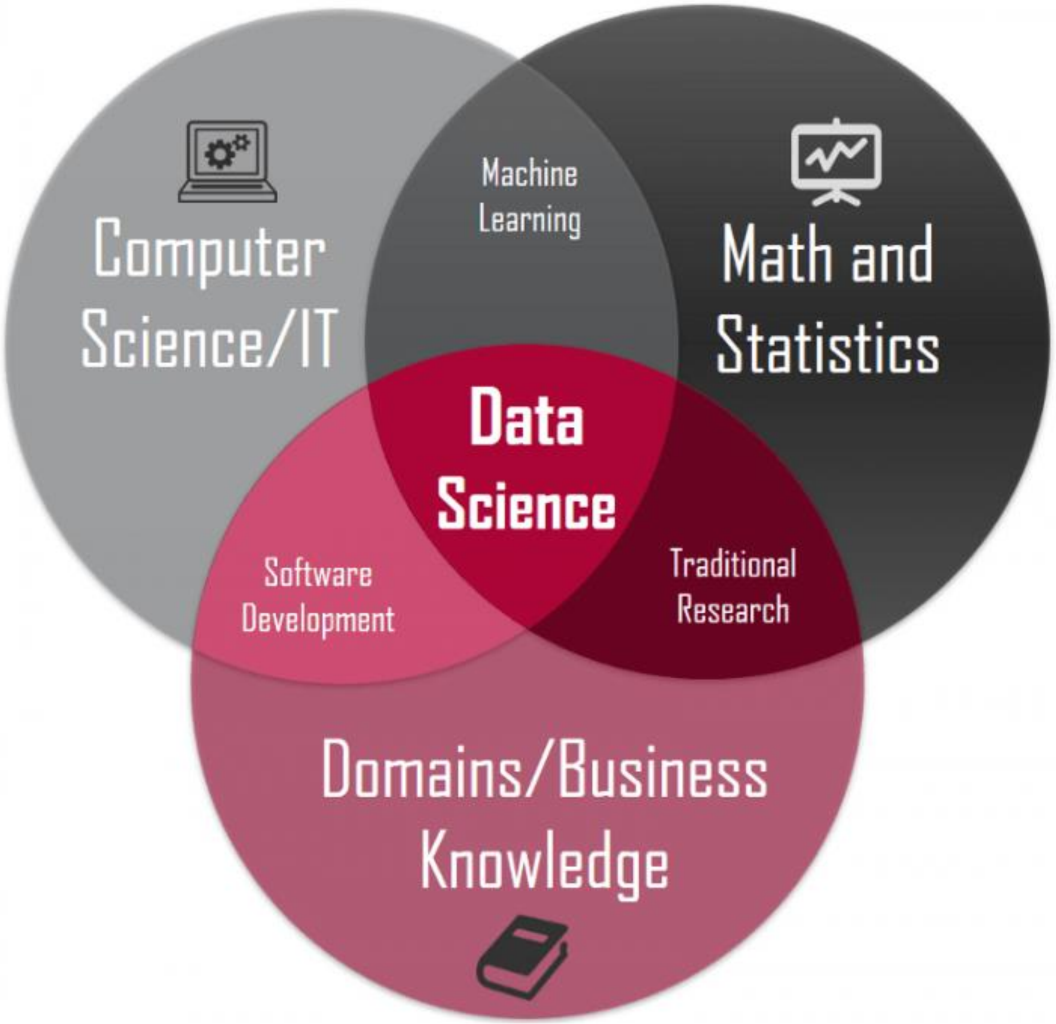
Multidimensional data visualization across files. Explore relationships within and among related datasets.

Orange 3

PyCharm Professional

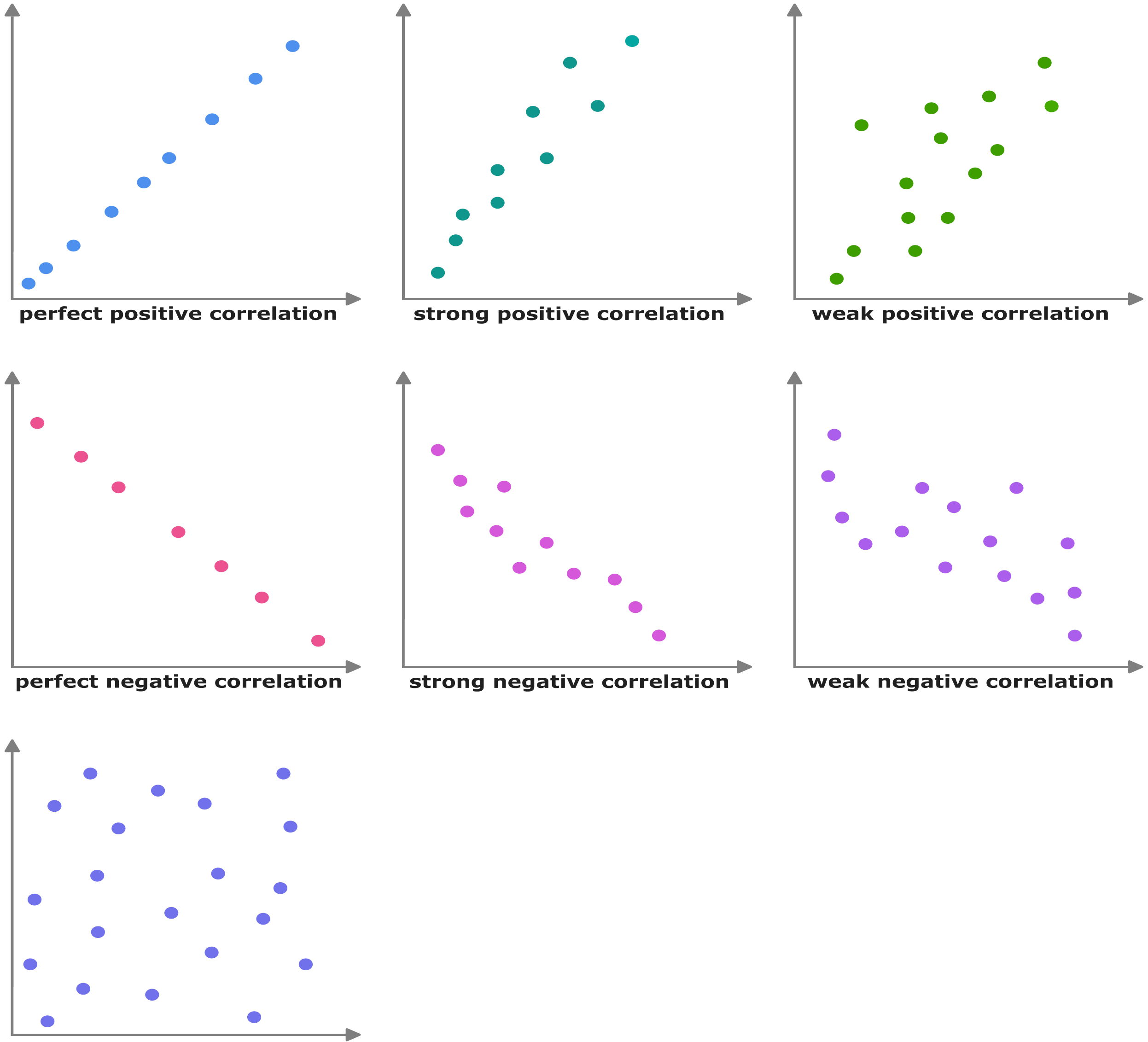
RStudio

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge from data across a broad range of application domains.



Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.

What is correlation?



perfect positive correlation

strong positive correlation

weak positive correlation

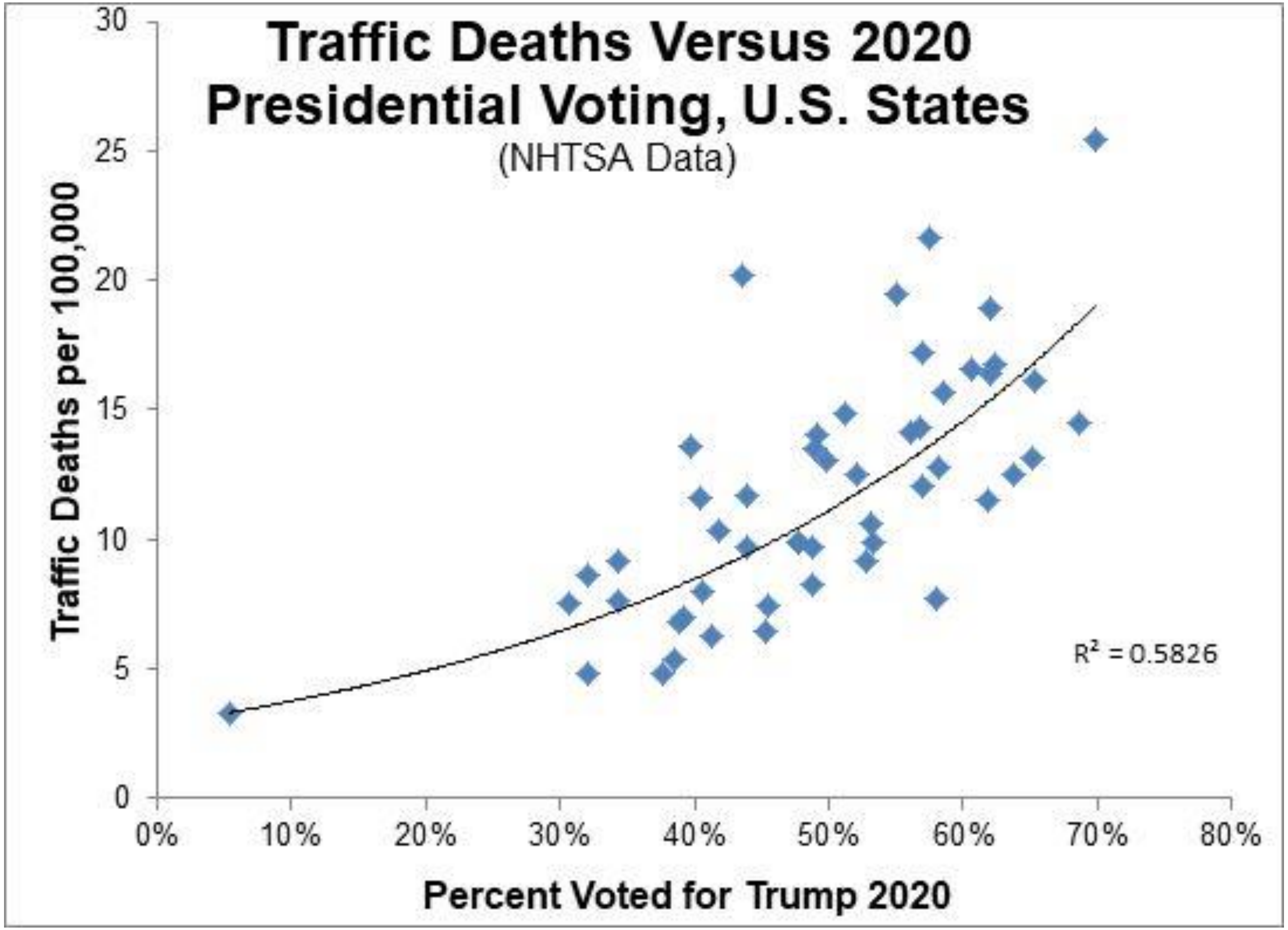
perfect negative correlation

strong negative correlation

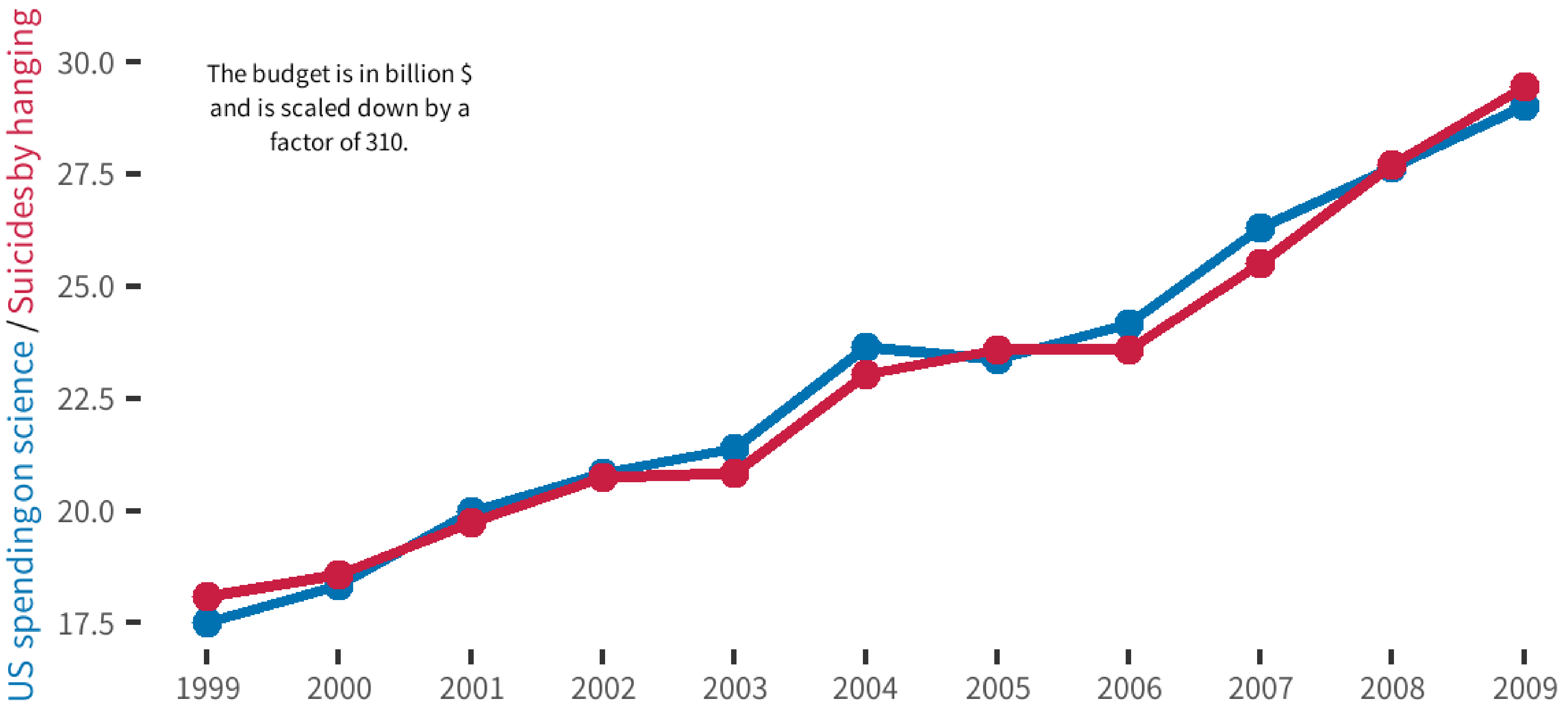
weak negative correlation

no correlation

How do you interpret data ?



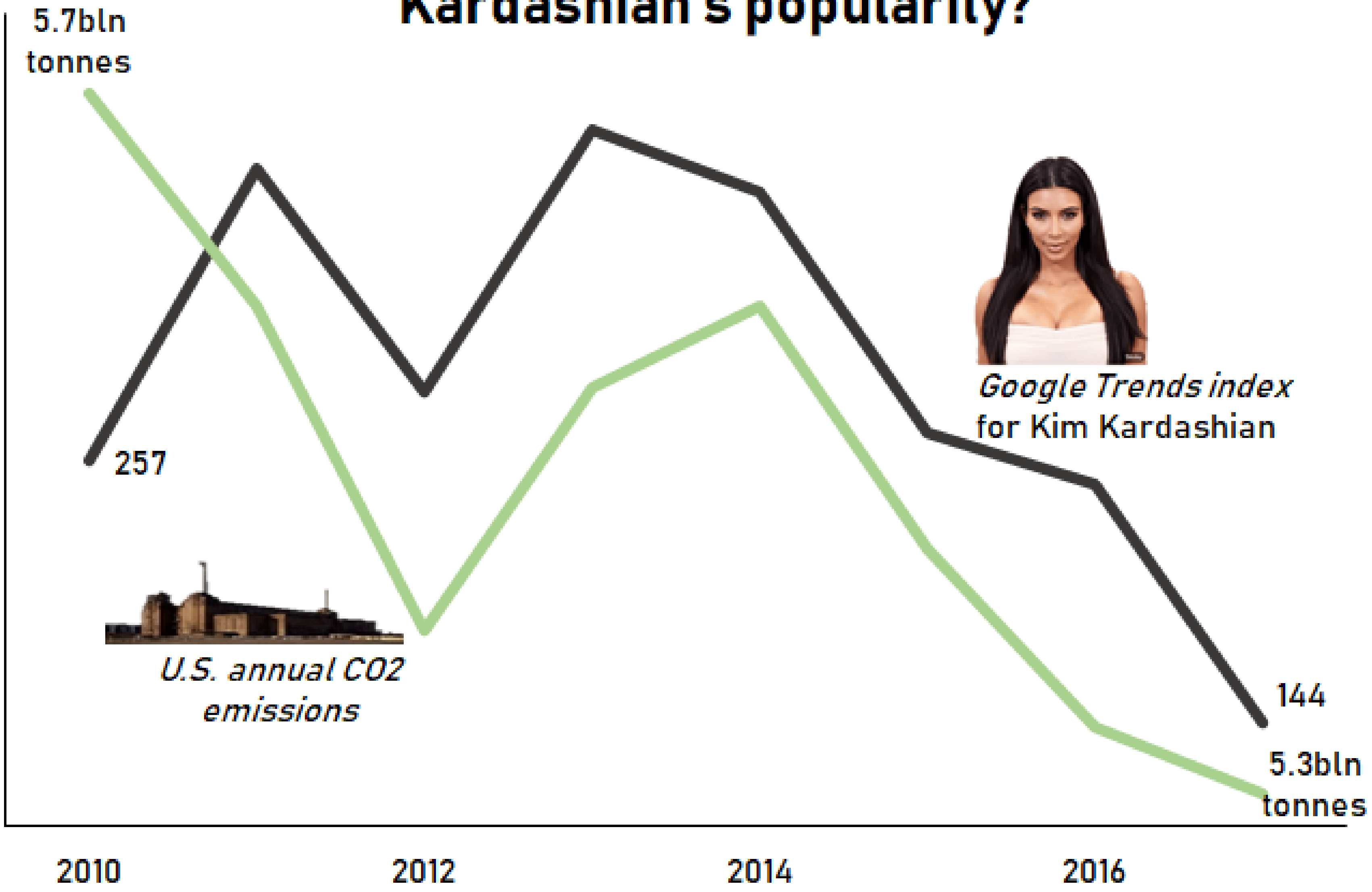
The correlation between US spending on science and suicides by hanging is 99.21%.

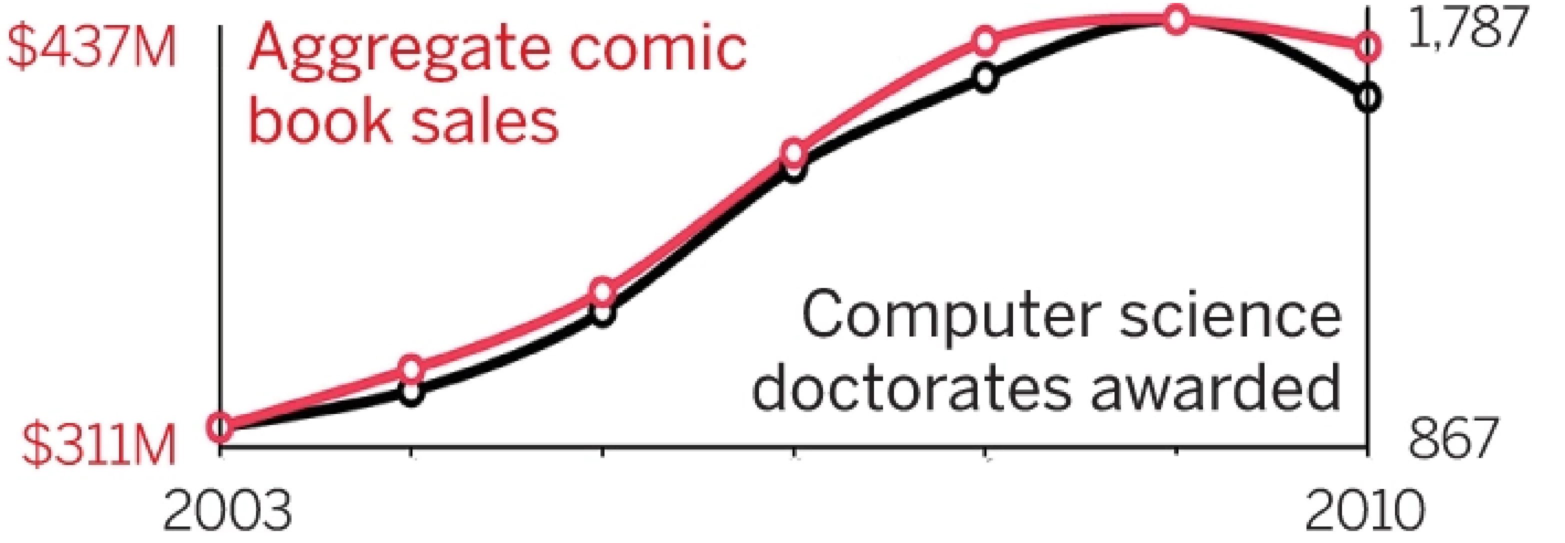


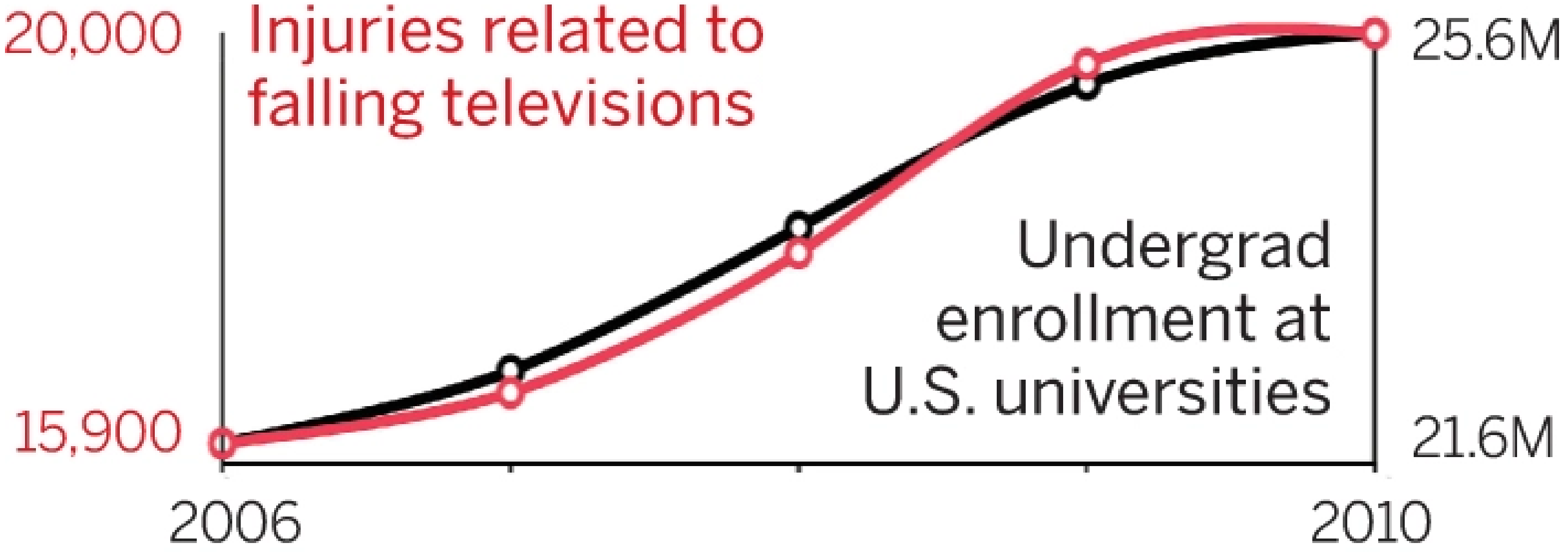
The budget is in billion $ and is scaled down by a factor of 310.

Inspiration: https://www.tylervigen.com
Data: U.S. Office of Management and Budget, Centers for Disease Control & Prevention
Graphic: Abhianv Malasi

Are falling emissions levels impacting Kim Kardashian's popularity?




$437M
Aggregate comic book sales
$311M
1,787
Computer science doctorates awarded
867
2003
2010

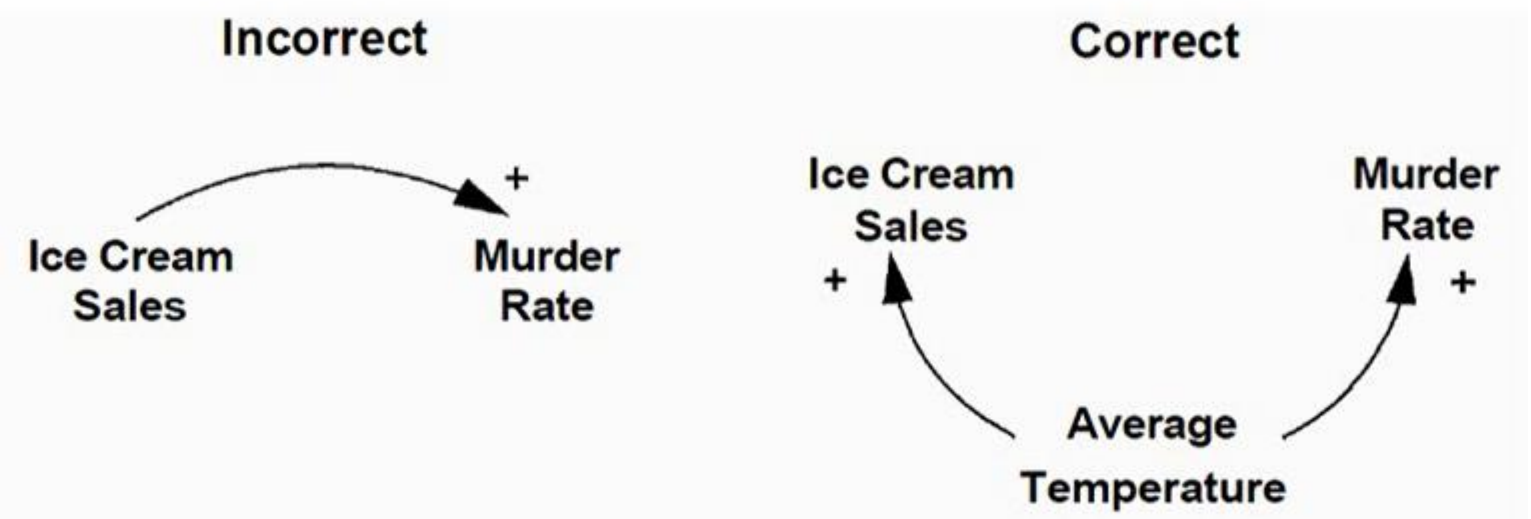

20,000
Injuries related to falling televisions
15,900
25.6M
Undergrad enrollment at U.S. universities
21.6M
2006
2010

Correlation

- Correlation is a statistical term describing the degree to which two variables move in coordination with one another. If the two variables move in the same direction, then those variables are said to have a positive correlation. If they move in opposite directions, then they have a negative correlation.

Causality vs. Correlation

- A **correlation** between variables does not automatically mean that the change in one variable is the cause of the change in the values of the other variable.
- **Causation** indicates that one event is the result of the occurrence of the other event; i.e. there is a causal relationship between the two events.



- **All causation have correlation**
- **All correlation can not be causation**
- **If there is no correlation , for 100% there is no causation**

Learn Together Python

- **Python** is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability.
- **Python** is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly procedural), object-oriented and functional programming.
- It is often described as a "batteries included" language due to its comprehensive standard library.

Interpretive vs compiled languages

Python is an interpretive language.

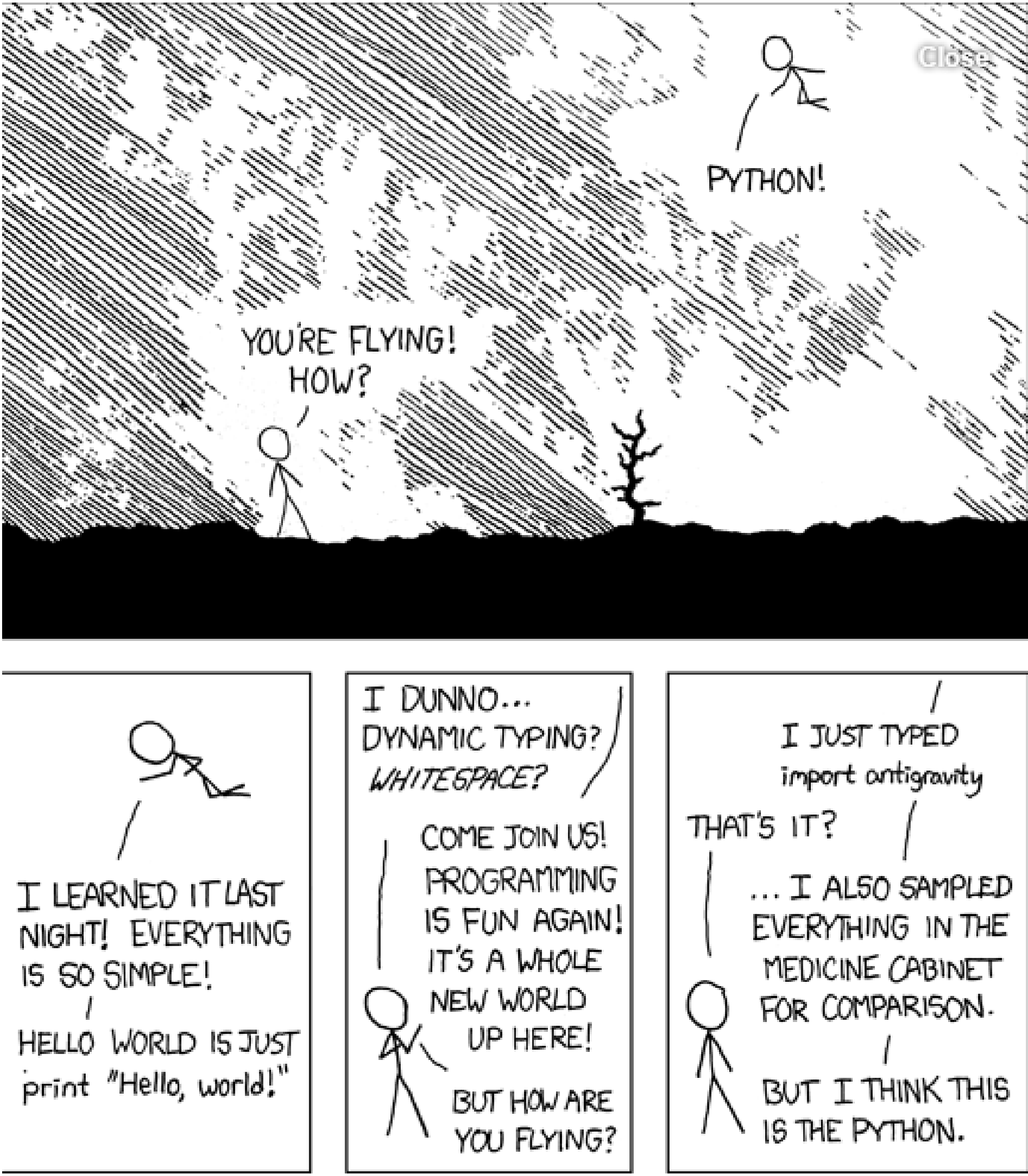
This means that your code is not directly run by the hardware. It is instead passed to a *virtual machine*, which is just another programme that reads and interprets your code. If your code used the '+' operation, this would be recognised by the interpreter at run time, which would then call its own internal function 'add(a,b)', which would then execute the machine code 'ADD'.

This is in contrast to compiled languages, where your code is translated into native machine instructions, which are then directly executed by the hardware. Here, the '+' in your code would be translated directly in the 'ADD' machine code.

Advantages of Python?

Because Python is an interpretive language, it has a number of advantages:

- Automatic memory management.
- Ease of programming.
- Minimises development time.
- Python also has a focus on ***importing* modules,** a feature that makes it useful for scientific computing.


Close
PYTHON!
YOU'RE FLYING! HOW?
I LEARNED IT LAST NIGHT! EVERYTHING IS SO SIMPLE!
HELLO WORLD IS JUST print "Hello, world!"
I DUNNO... DYNAMIC TYPING? WHITESPACE?
COME JOIN US! PROGRAMMING IS FUN AGAIN! IT'S A WHOLE NEW WORLD UP HERE!
BUT HOW ARE YOU FLYING?
I JUST TYPED import antigravity
THAT'S IT?
...I ALSO SAMPLED EVERYTHING IN THE MEDICINE CABINET FOR COMPARISON.
BUT I THINK THIS IS THE PYTHON.

Disadvantages

- Interpreted languages are slower than compiled languages.
- The modules that you import are developed in a decentralised manner; this can cause issues based upon individual assumptions.

The Anaconda IDE…

- The Anaconda distribution is the most popular Python distribution out there.
- Most importable packages are pre-installed.
- Anaconda Distribution equips individuals to easily search and install thousands of Python/R packages and access a vast library of community content and support.

Variables

- Variables in python can contain alphanumerical characters and some special characters.
- By convention, it is common to have variable names that start with lower case letters; but you can do whatever you want.
- Some keywords are reserved and cannot be used as variable names due to them serving an in-built Python function; i.e. `and`, `continue`, `break`. Your IDE will let you know if you try to use one of these.
- Python is dynamically typed; the type of the variable is derived from the value it is assigned.

Variable types

- Integer (int)
- Float (float)
- String (str)
- Boolean (bool)
- Complex (complex)
- […]
- User defined (classes)

- The print() function is used to print something to the screen.

Arithmetic operators

The arithmetic operators:

- Addition: +
- Subtract: -
- Multiplication: *
- Division: /
- Power: **

Boolean operators

- Boolean operators are useful when making conditional statements, we will cover these in-depth later.
- and
- or
- not

Comparison operators

- Greater than: `>`
- Lesser than: `<`
- Greater than or equal to: `>=`
- Lesser than or equal to: `<=`
- Is equal to: `==`

Working with strings

Strings are amongst the most popular types in Python. We can create them simply by enclosing characters in quotes. Python treats single quotes the same as double quotes. Creating strings is as simple as assigning a value to a variable.

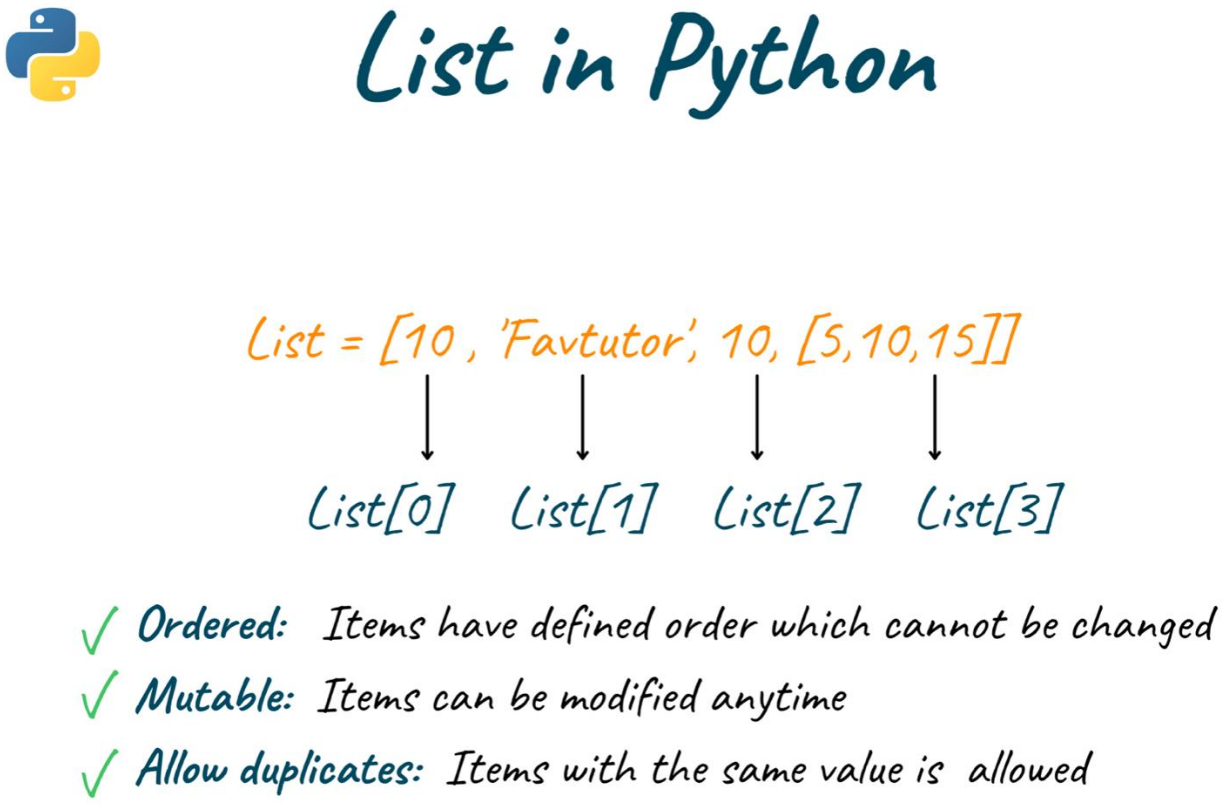
For example –

```
var1 = 'Hello World!’

var2 = "Python Programming"
```

What are Lists in Python?

Python possesses a list as a data structure that is an ordered sequence of elements and mutable in nature. Each item or value that is inside of a list is called an element. Just as strings are defined as characters between quotes, lists are defined by having values between square brackets ([ ])separated by commas.



List in Python

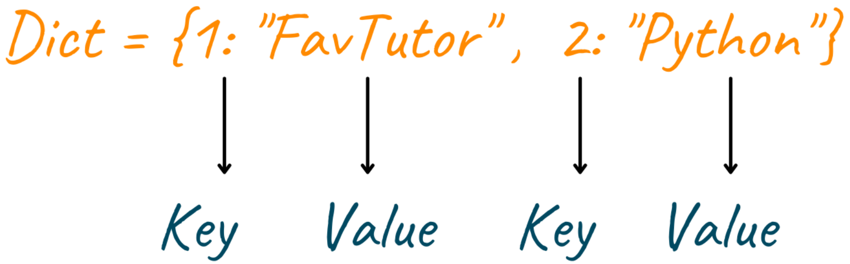
List = [10 , 'Favtutor', 10, [5,10,15]]

✓ **Ordered:** Items have defined order which cannot be changed

✓ **Mutable:** Items can be modified anytime

✓ **Allow duplicates:** Items with the same value is allowed

Dictionary in Python


Dict = {1: "FavTutor", 2: "Python"}

Key Value Key Value

✓ *Ordered:* Key-value pairs have defined ordered and cannot be changed

✓ *Mutable:* Dictionaries are mutable but keys are immutable

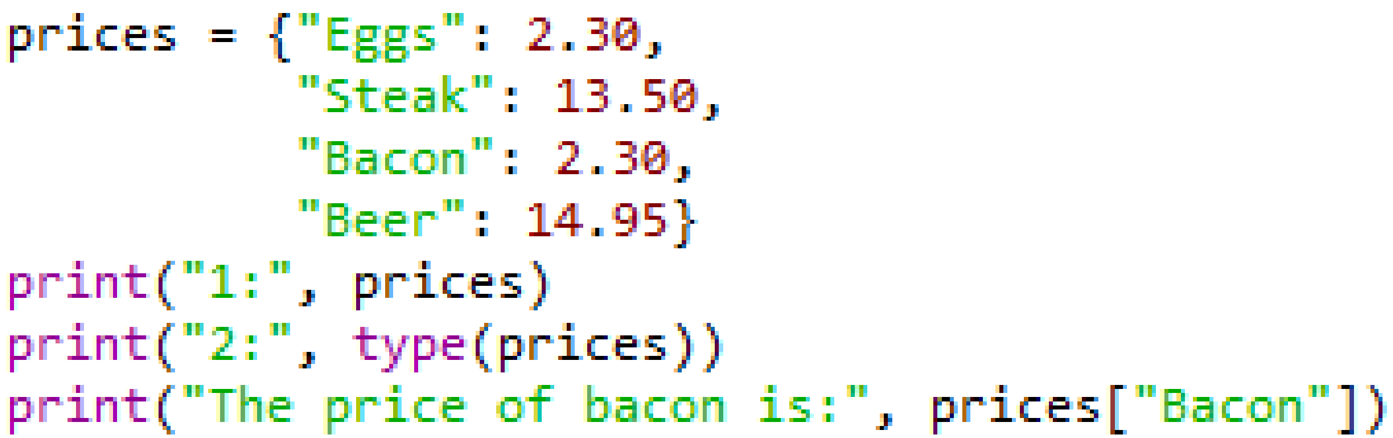
✗ *Duplicates:* Do not allow duplicates Items with same keys

- **What is Dictionary in Python?**
- Dictionary is a default python data structure used to store the data collection in the form of key-value pairs.
- Dictionaries are written inside the curly brackets ({}), separated by commas.
- However, the key and value of the data are separated by placing a semi-colon between them(:).

Dictionaries

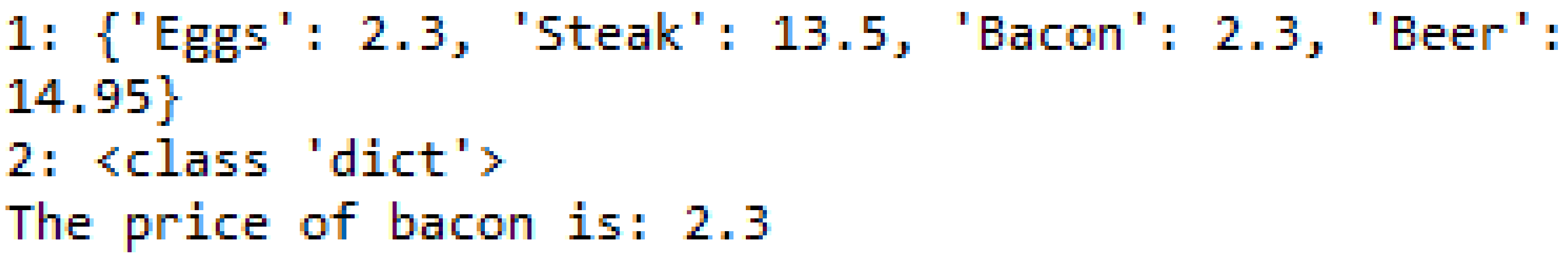
- Dictionaries are lists of key-valued pairs.

In:



```
prices = {"Eggs": 2.30,
          "Steak": 13.50,
          "Bacon": 2.30,
          "Beer": 14.95}
print("1:", prices)
print("2:", type(prices))
print("The price of bacon is:", prices["Bacon"])
```

Out:



```
1: {'Eggs': 2.3, 'Steak': 13.5, 'Bacon': 2.3, 'Beer':
14.95}
2: <class 'dict'>
The price of bacon is: 2.3
```

What is an Array?

An array is a data structure that holds fix number of elements and these elements should be of the same data type. Most of the data structure makes use of an array to implement their algorithm. There is two important part of the array: one is an Element: Each item store in the array is called an element and second is an Index: Every element in the array has its own numerical value to identify the element. These elements allocate contiguous memory locations that allow easy modifications in data.

The most important difference with the list is the elements inside the list is not compulsorily be of the same data type along with negative indexing.

Python Library

- A Python library is **a collection of related modules**.
- It contains bundles of code that can be used repeatedly in different programs. It makes Python Programming simpler and convenient for the programmer.
- As we don't need to write the same code again and again for different programs.

Pandas
PyTorch
NumPy
plotly
matplotlib
theano
TensorFlow
scikit learn
Keras
python™
Natural Language Analyses with NLTK

Python Libraries for Data Science

Many popular Python toolboxes/libraries:

- NumPy
- SciPy
- Pandas
- SciKit-Learn

Visualization libraries

- matplotlib
- Seaborn

and many more …

Python Libraries for Data Scie

NumPy:

- introduces objects for multidimensional arrays and matrices, as well as functions that allow to easily perform advanced mathematical and statistical operations on those objects

- provides vectorization of mathematical operations on arrays and matrices which significantly improves the performance

- many other python libraries are built on NumPy

Link: http://www.numpy.org/

Python Libraries for Data Science

SciPy:

- collection of algorithms for linear algebra, differential equations, numerical integration, optimization, statistics and more
- part of SciPy Stack
- built on NumPy

Link: https://www.scipy.org/scipylib/

Python Libraries for Data Science

Pandas:

- adds data structures and tools designed to work with table-like data (similar to Series and Data Frames in R)

- provides tools for data manipulation: reshaping, merging, sorting, slicing, aggregation etc.

- allows handling missing data

Link: http://pandas.pydata.org/

Python Libraries for Data Science

SciKit-Learn:

- provides machine learning algorithms: classification, regression, clustering, model validation etc.

- built on NumPy, SciPy and matplotlib

Link: http://scikit-learn.org/

Python Libraries for Data Science

matplotlib:

- python 2D plotting library which produces publication quality figures in a variety of hardcopy formats
- a set of functionalities similar to those of MATLAB
- line plots, scatter plots, barcharts, histograms, pie charts etc.
- relatively low-level; some effort needed to create advanced visualization

Link: https://matplotlib.org/

Python Libraries for Data Science

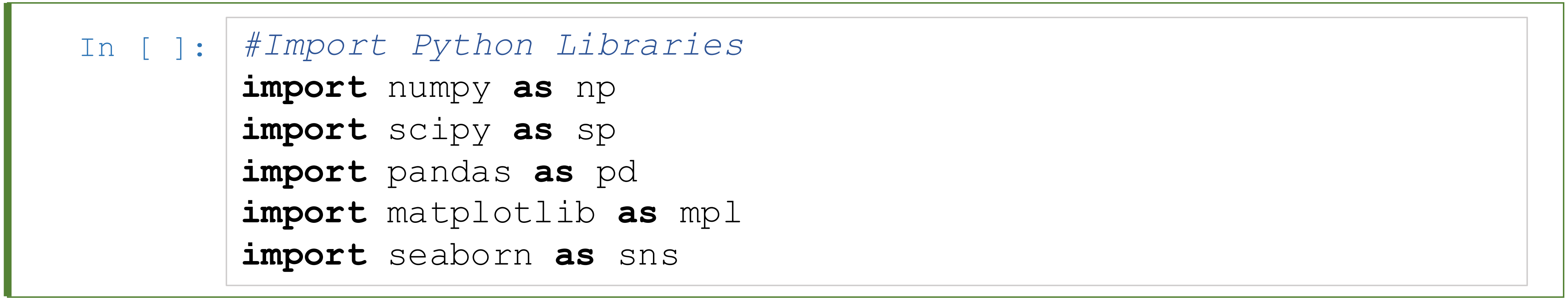
c:

- based on matplotlib
- provides high level interface for drawing attractive statistical graphics
- Similar (in style) to the popular ggplot2 library in R

Link: https://seaborn.pydata.org/

seaborn

Loading Python Libraries



In [ ]:

```
#Import Python Libraries
import numpy as np
import scipy as sp
import pandas as pd
import matplotlib as mpl
import seaborn as sns
```

Press `Shift+Enter` to execute the *jupyter* cell