ICME Stanford Institute for Computational & Mathematical Engineering

SUMMER WORKSHOPS

FUNDAMENTALS OF DATA SCIENCE

# Welcome to Introduction to Python!

(We start at 9am)

1. Python on Google Colab
   `https://bit.ly/3kV6Cyo`

2. Interactive polls (*no account needed*)
   `pollev.com/lc928`

Slides

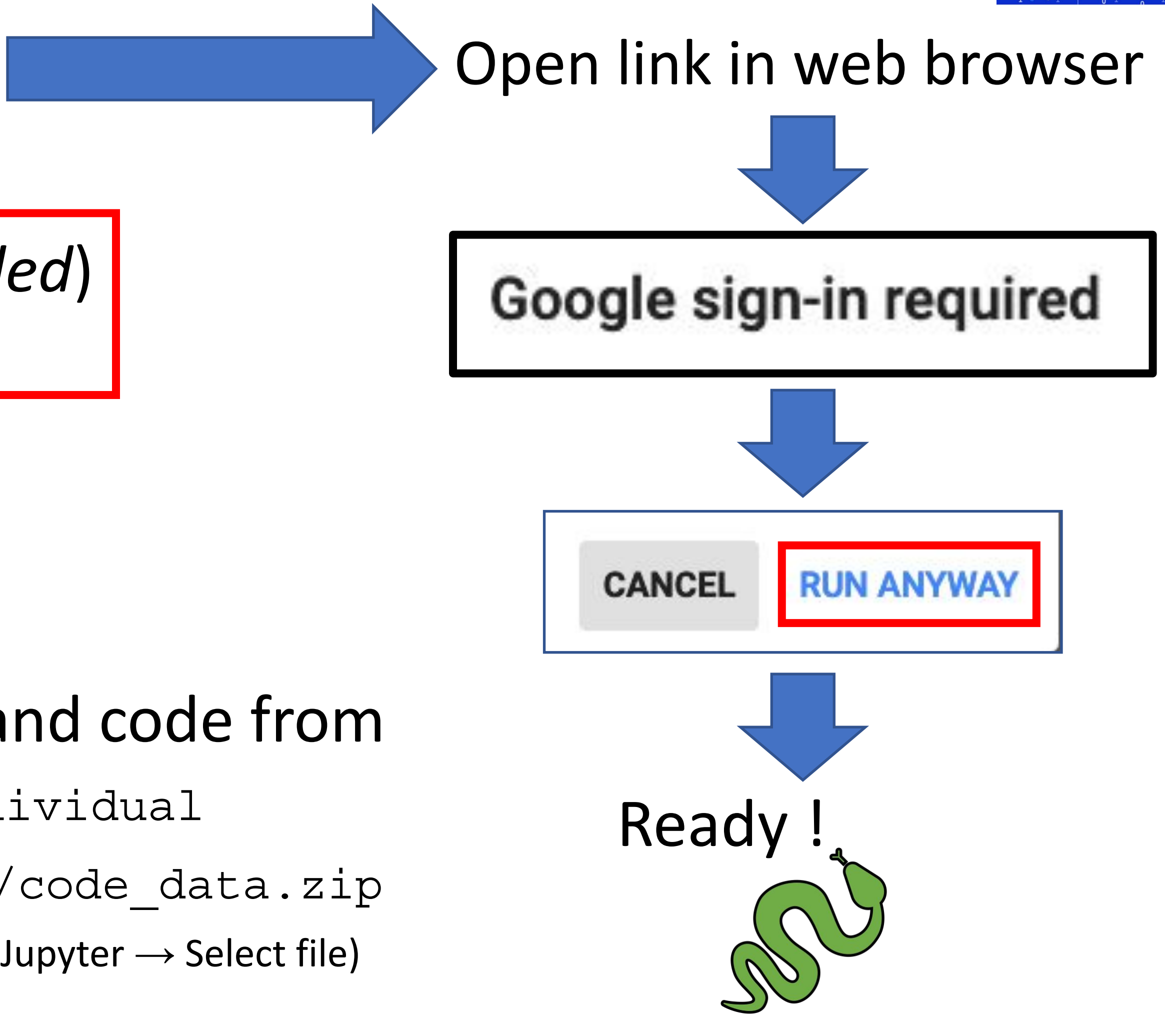
`https://bit.ly/3kYEv1s`

No Colab? Download Anaconda and code from

`www.anaconda.com/products/individual`

`www.stanford.edu/~lcambier/pc/code_data.zip`

Open Python.ipynb in Jupyter (Anaconda Navigator → Jupyter → Select file)



Open link in web browser

Google sign-in required

Ready !

ICME
Stanford | Institute for Computational & Mathematical Engineering

# Introduction to Python

Léopold Cambier
`lcambier@stanford.edu`

ICME Summer Workshops
Fundamentals of Data Science

August 18, 2020

# First and Foremost

python

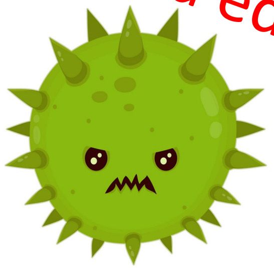
1. Workshop is recorded (audio & video)
   Office-hours (1pm-2pm) will not be recorded.

2. If you are OK, turn on your camera :-)

3. Ask questions in chat.
   Ryan our awesome TA is answering.

4. Stay muted unless you are actively talking

Covid edition



# Some questions for you first!

PollEv.com/lc928

- Mandatory to participate :-)
- Laptop or phone (keep the tab open for later)
- No account needed, anonymous
- The more the better.

# Some info about me

- 5th year PhD in ICME
- Numerical linear algebra (very big sparse “Ax=b”)
- Parallel computing (big computers)

# “Python” you said ?

# “Python” you said ?

Monty Python

and the Holy Grail

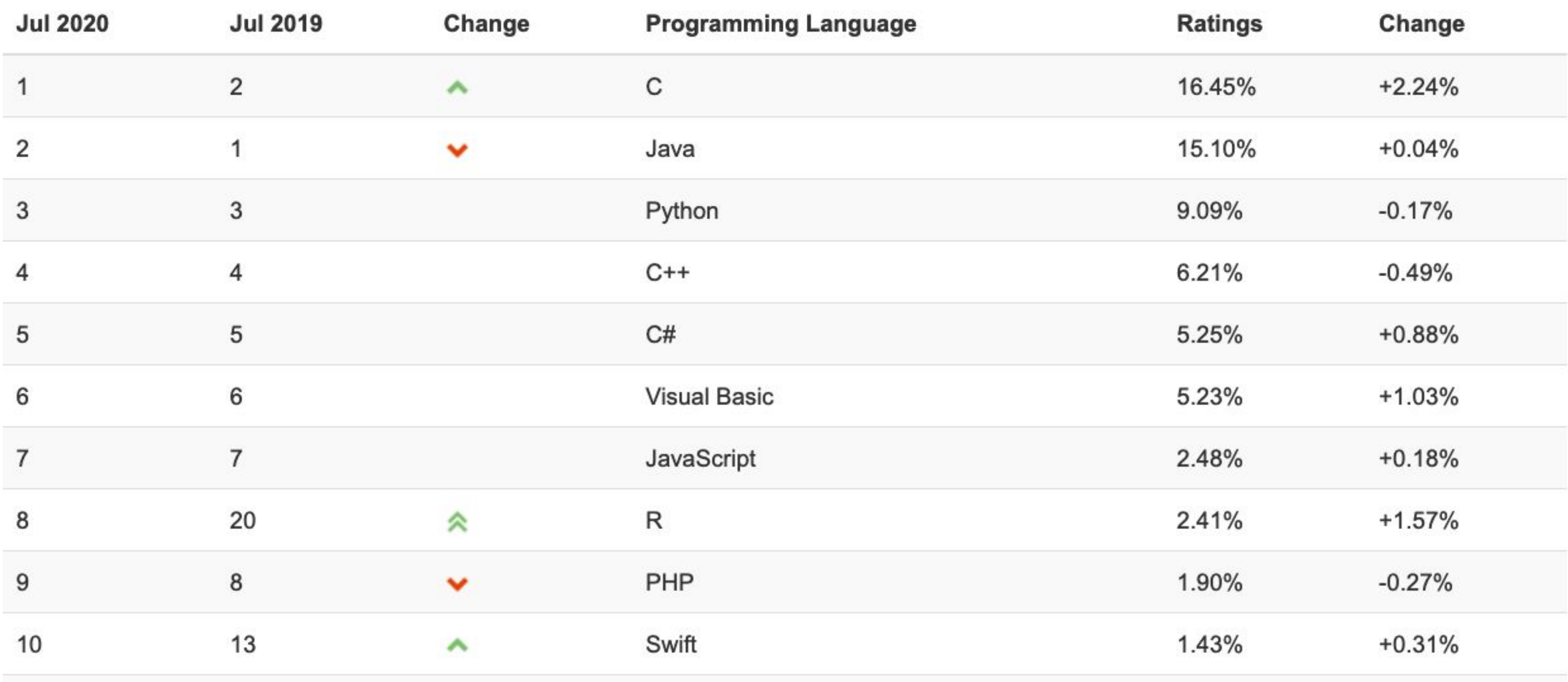
# Python

A very popular programming language

- Web applications
- Scientific computing
- Data science and machine learning
- General purpose applications

# Python

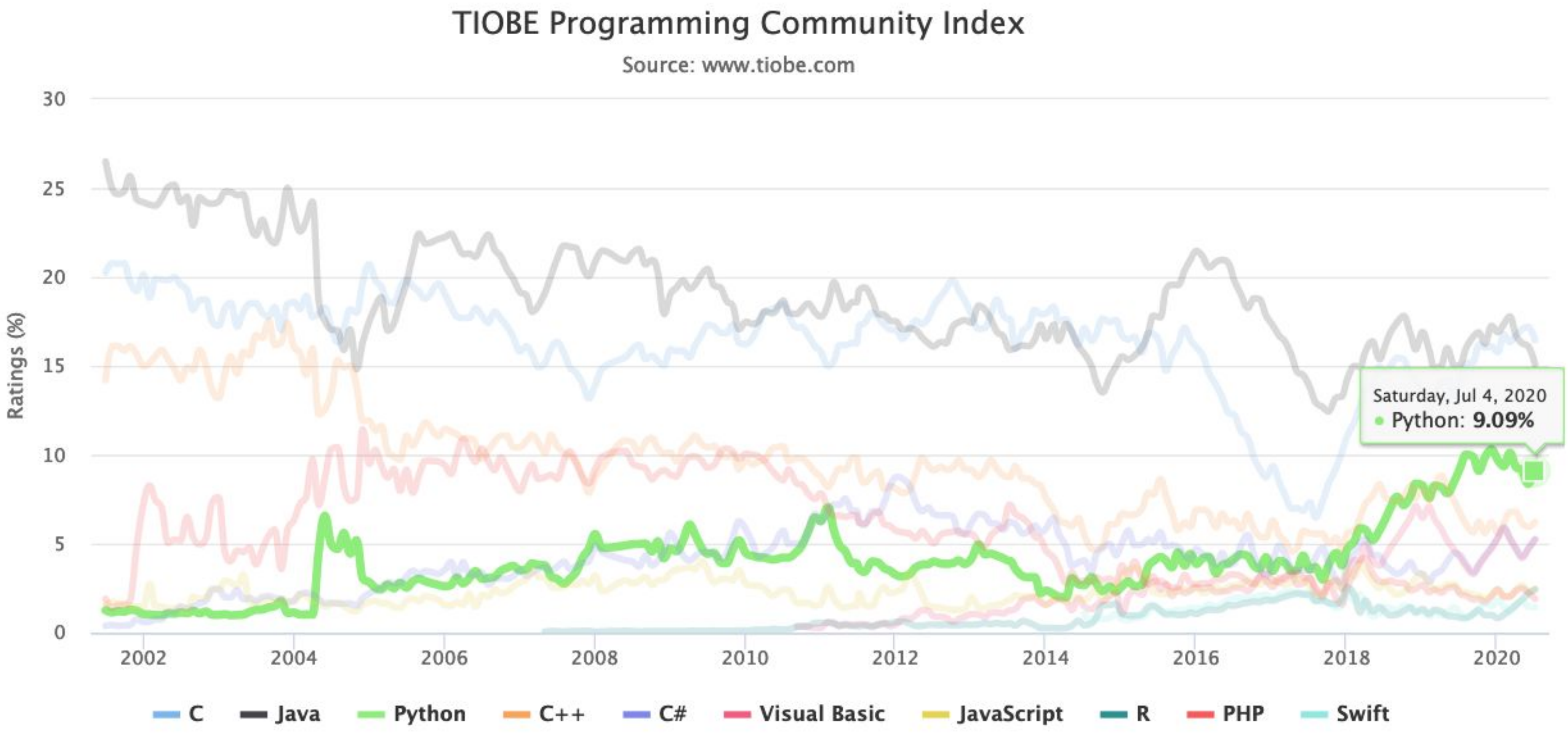
www.tiobe.com



| Jul 2020 | Jul 2019 | Change | Programming Language | Ratings | Change |
|---|---|---|---|---|---|
| 1 | 2 | ▲ | C | 16.45% | +2.24% |
| 2 | 1 | ▼ | Java | 15.10% | +0.04% |
| 3 | 3 |  | Python | 9.09% | -0.17% |
| 4 | 4 |  | C++ | 6.21% | -0.49% |
| 5 | 5 |  | C# | 5.25% | +0.88% |
| 6 | 6 |  | Visual Basic | 5.23% | +1.03% |
| 7 | 7 |  | JavaScript | 2.48% | +0.18% |
| 8 | 20 | ⏫ | R | 2.41% | +1.57% |
| 9 | 8 | ▼ | PHP | 1.90% | -0.27% |
| 10 | 13 | ▲ | Swift | 1.43% | +0.31% |

# Python

www.tiobe.com



# The class

- Python
  - Variables, control-flow, containers, I/O
  - Functions, iterables
  - (Maybe) References, modules
- Numpy + Matplotlib
- Pandas
- Scikit-learn

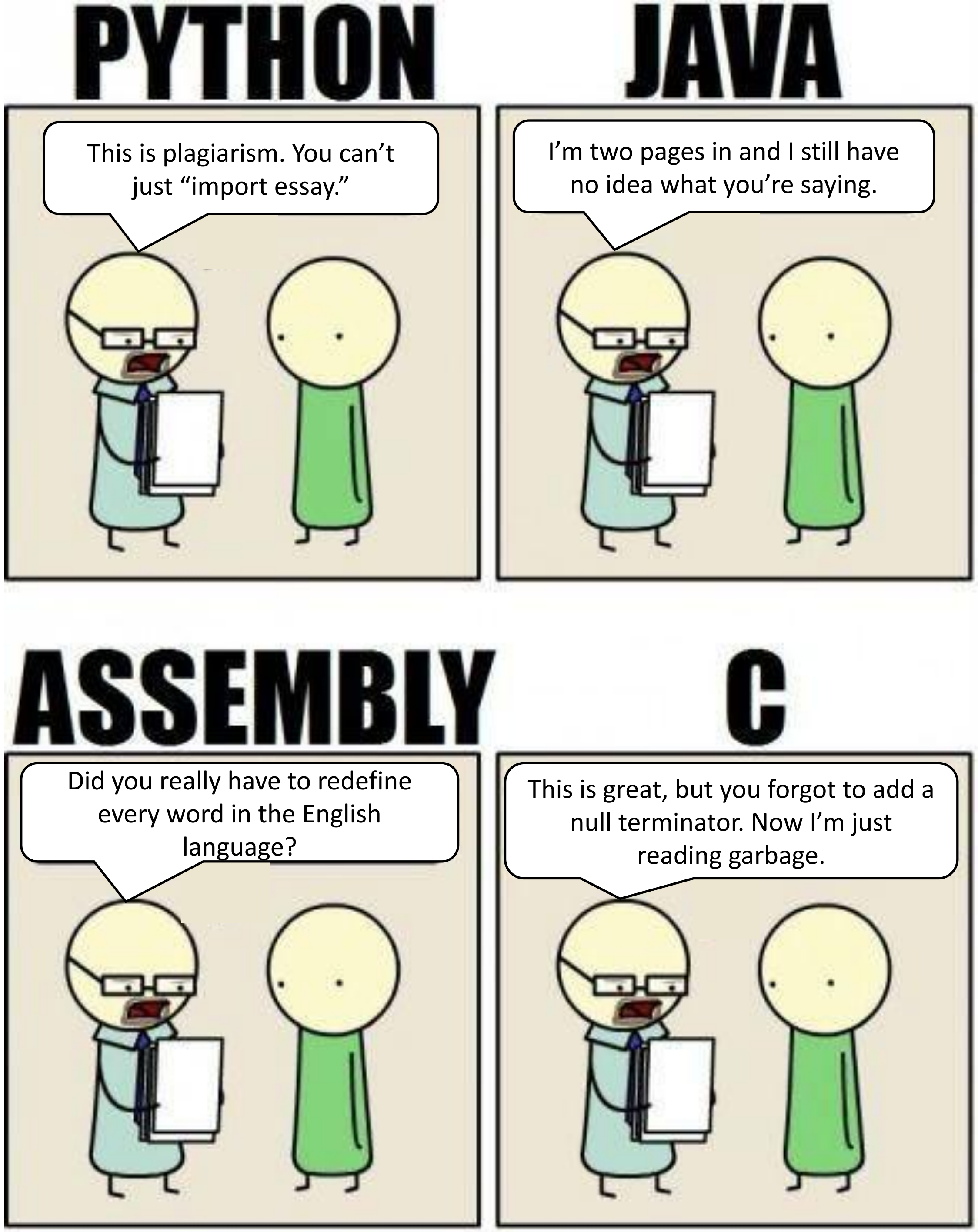
| | |
|---|---|
| 9:00-10:30 | Basic Python |
| | |
| 10:45-12:00 | |
| | |
| 1:00-2:00 | Q&A (optional) |
| 2:00-3:15 | Numpy,<br>Pandas,<br>Scikit-learn<br>(more applied) |
| | |
| 3:30-4:45 | |

# The class

- We will go through code *together.*
- Ask questions in the chat
- Many exercises.
- Goal:
    - Good enough basic Python knowledge to explore on your own.
    - Exposure to various tools used in science. Won’t be an expert.

# Python

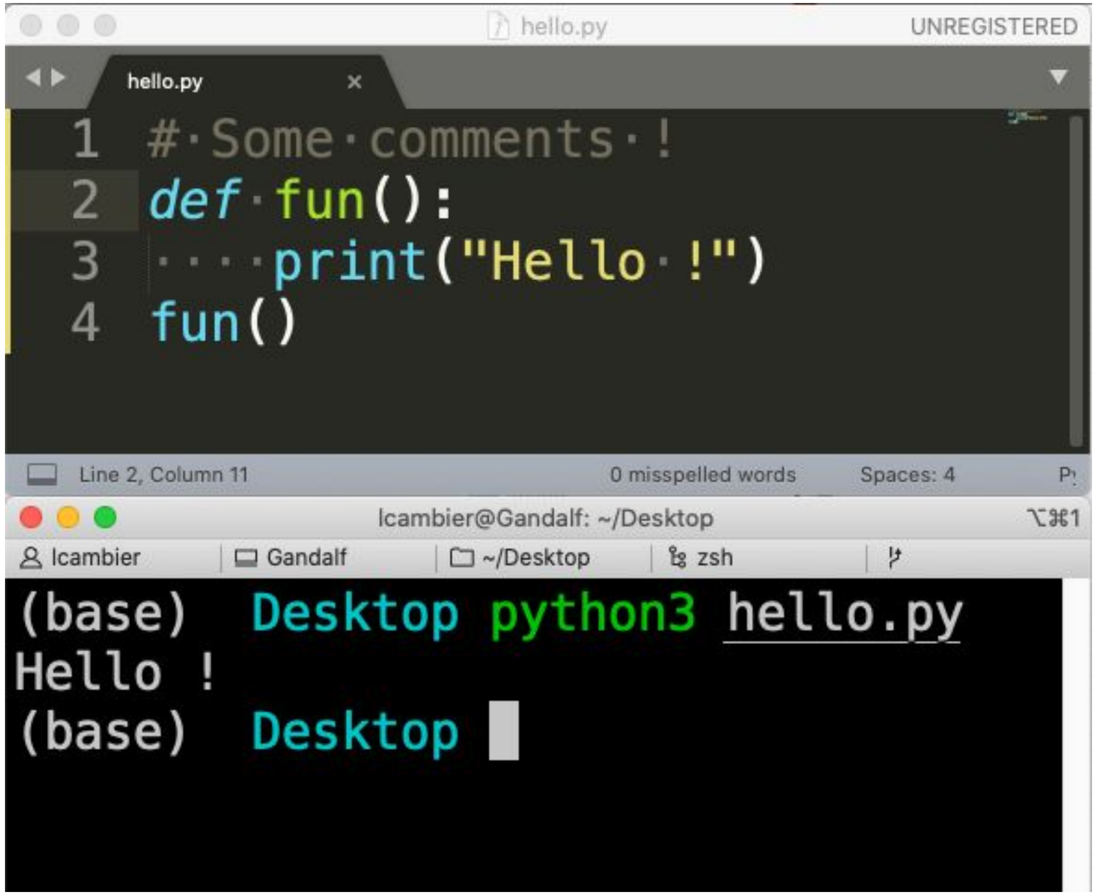
- High-level
- Portable
- Interpreted
- Extensible
- Object-oriented
- Dynamically typed
- Garbage collected


PYTHON
This is plagiarism. You can't just "import essay."
JAVA
I'm two pages in and I still have no idea what you're saying.
ASSEMBLY
Did you really have to redefine every word in the English language?
C
This is great, but you forgot to add a null terminator. Now I'm just reading garbage.

# Python 2 vs Python 3

Really no reasons to learn Python 2 in 2020 :-)

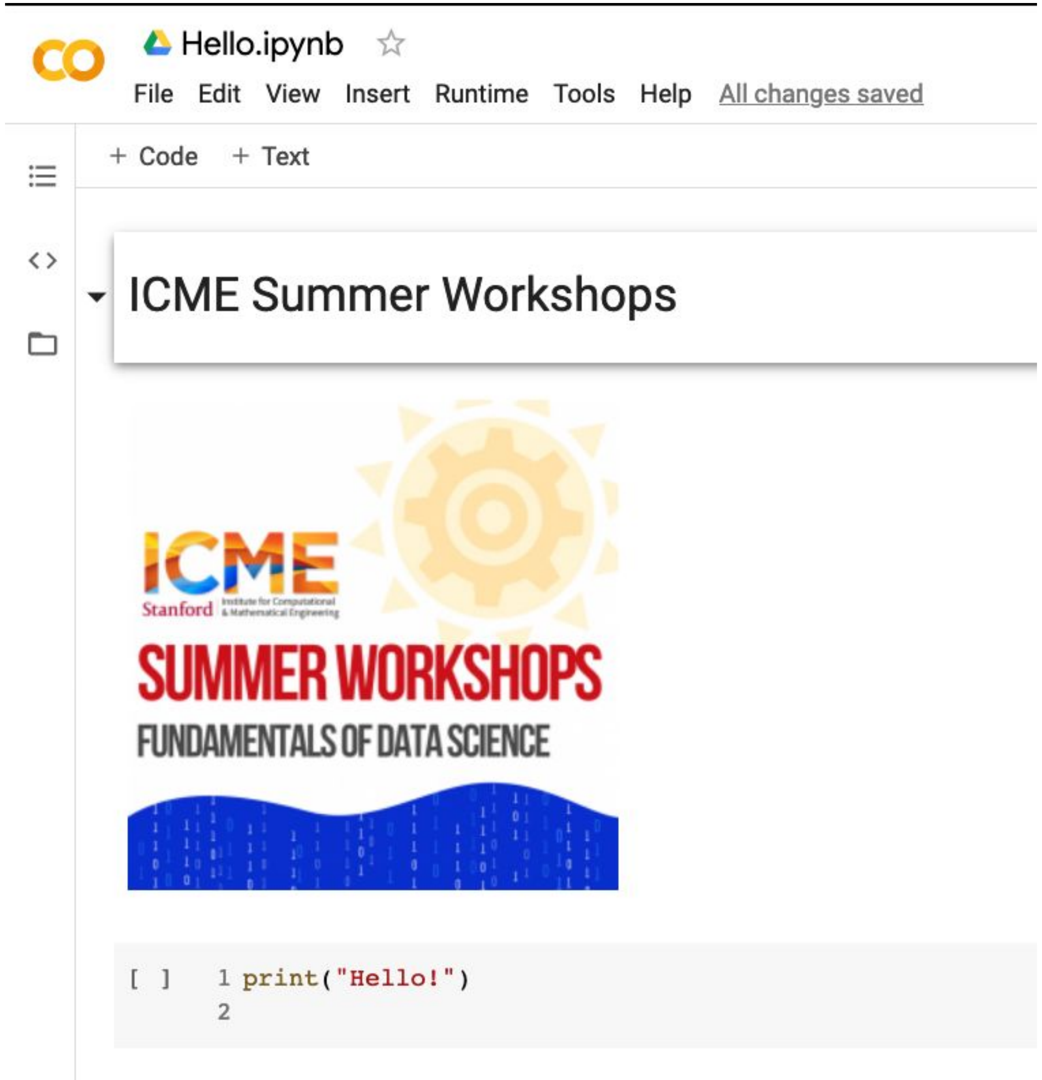
# How to use Python ?



```python
# Some comments !
def fun():
    print("Hello !")
fun()
```

```
(base)  Desktop python3 hello.py
Hello !
(base)  Desktop
```

Scripts and Python text files
(in a text editor, offline, usually)



ICME Summer Workshops

```python
print("Hello!")
```

Notebooks (text + viz + code)
(in web browser, online or offline)

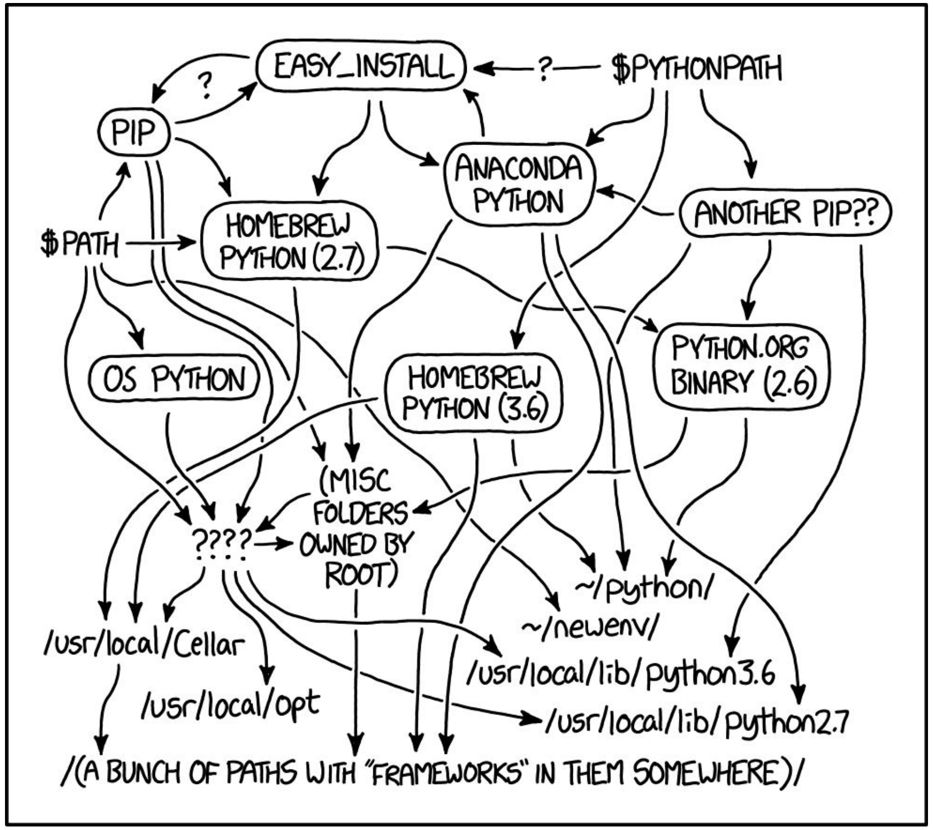
# Running on your laptop

Download Anaconda

[https://www.anaconda.com/products/individual](https://www.anaconda.com/products/individual)

Comes with all you need

- Many modules preinstalled
- Can use for scripts from terminal
- Can use for Jupyter notebooks from browser



MY PYTHON ENVIRONMENT HAS BECOME SO DEGRADED
THAT MY LAPTOP HAS BEEN DECLARED A SUPERFUND SITE.

# Let’s open the first Notebook!

# Strings indexing

```
s = “abcdefgh”
```

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| a | b | c | d | e | f | g | h |
| -8 | -7 | -6 | -5 | -4 | -3 | -2 | -1 |

# Dictionaries

# Everything is a reference (1)

```
a = 2
```

```
a = 3
```

# Everything is a reference (2)

```
a = "Stanford"

b = a

a = "ICME"

print(b)
```

# Everything is a reference (3)

```
a = [1, 2, 3]
b = a
a[0] = "ICME"
print(b)
```

# Everything is a reference (4)

```
a = [1, 2]
t = (a, 1, "String")
a[0] = [3, 4]
print(t)
```

# Everything is a reference (5)

```python
def fun(x):
    # same as x = a
    # ...
a = # Something
fun(a)
```

# Quizzes !

PollEv.com/lc928

# Welcome back to Introduction to Python!

Numpy

`bit.ly/3h844uz`

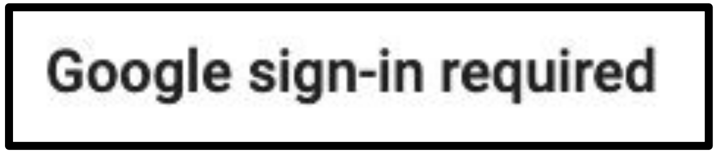
Pandas

`bit.ly/3kYEVoy`

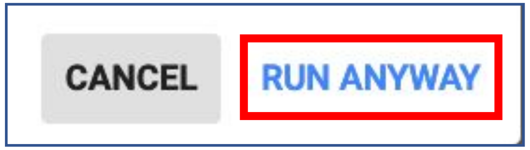
Sklearn

`bit.ly/3280RF8`

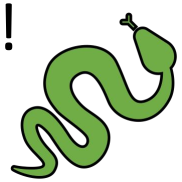
→ Open link in web browser



Google sign-in required



Ready !



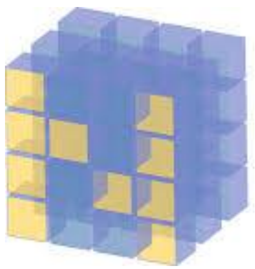
No Colab? Download code from

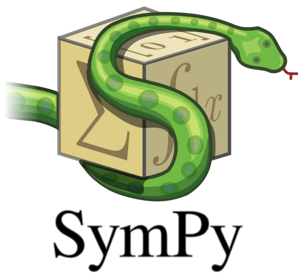
`www.stanford.edu/~lcambier/pc/code_data.zip`

Open Python.ipynb in Jupyter (Anaconda Navigator → Jupyter → Select file)

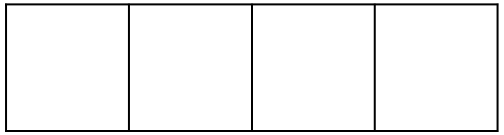
# SciPy

- An ecosystem for Scientific Computing and Data science in Python
- Includes many packages

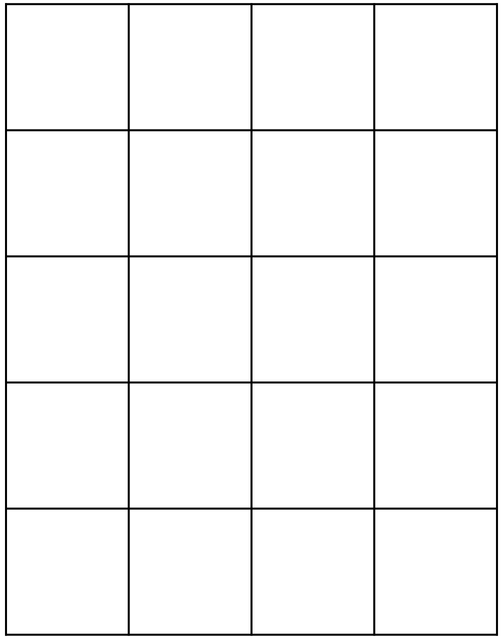




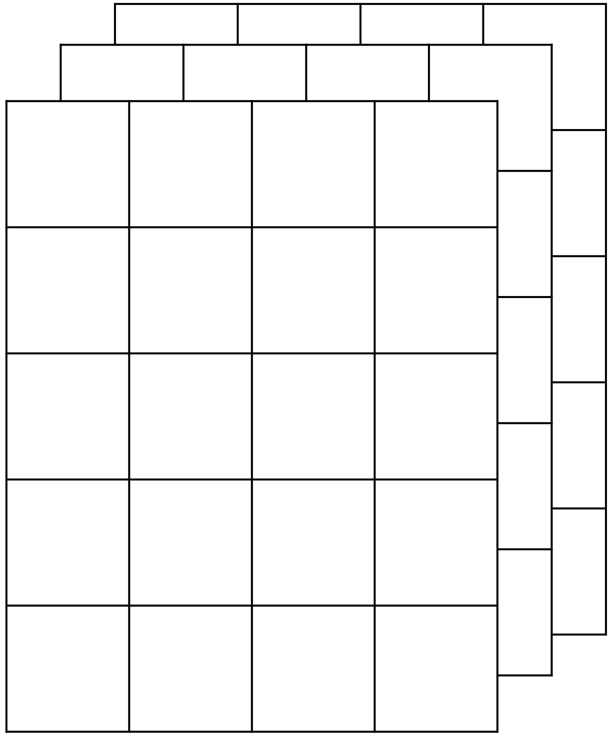
# Numpy: arrays



1-D array

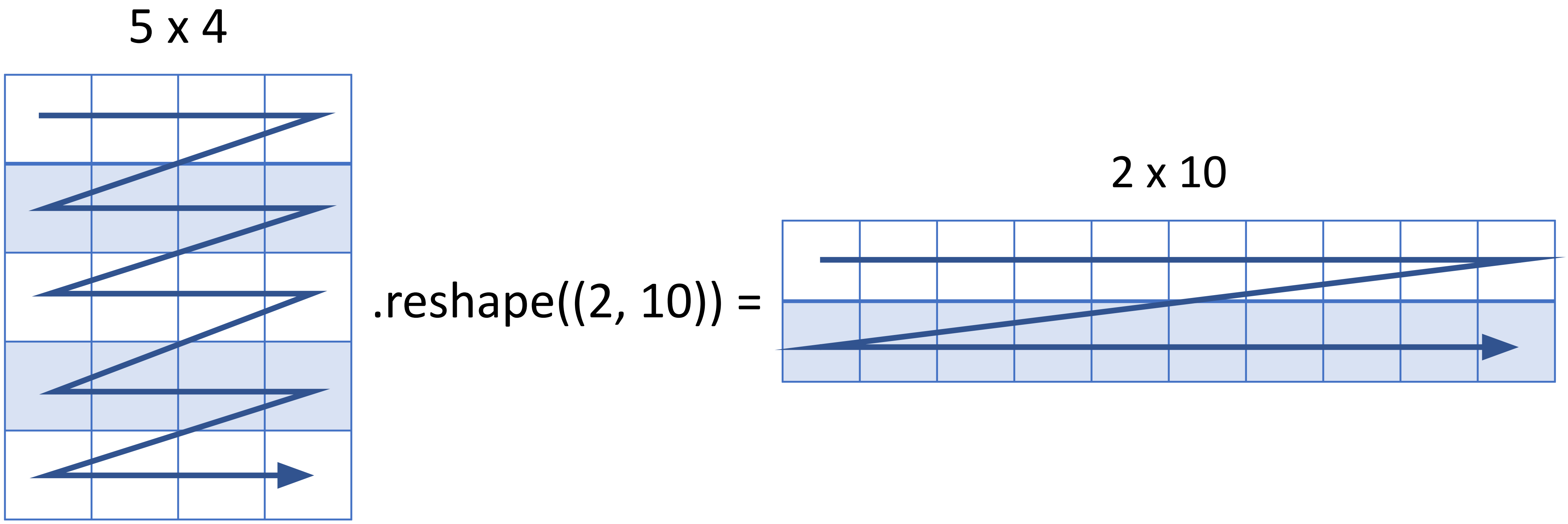


2-D array



3+-D array

# Reshaping

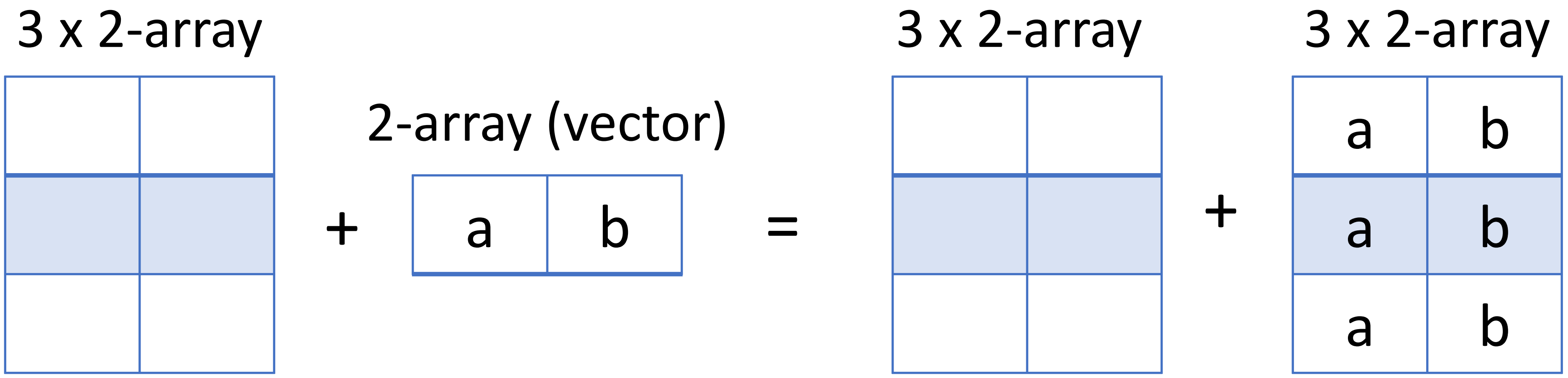


5 x 4

.reshape((2, 10)) =

2 x 10

# Broadcasting



3 x 2-array

+

2-array (vector)

a b

=

3 x 2-array

+

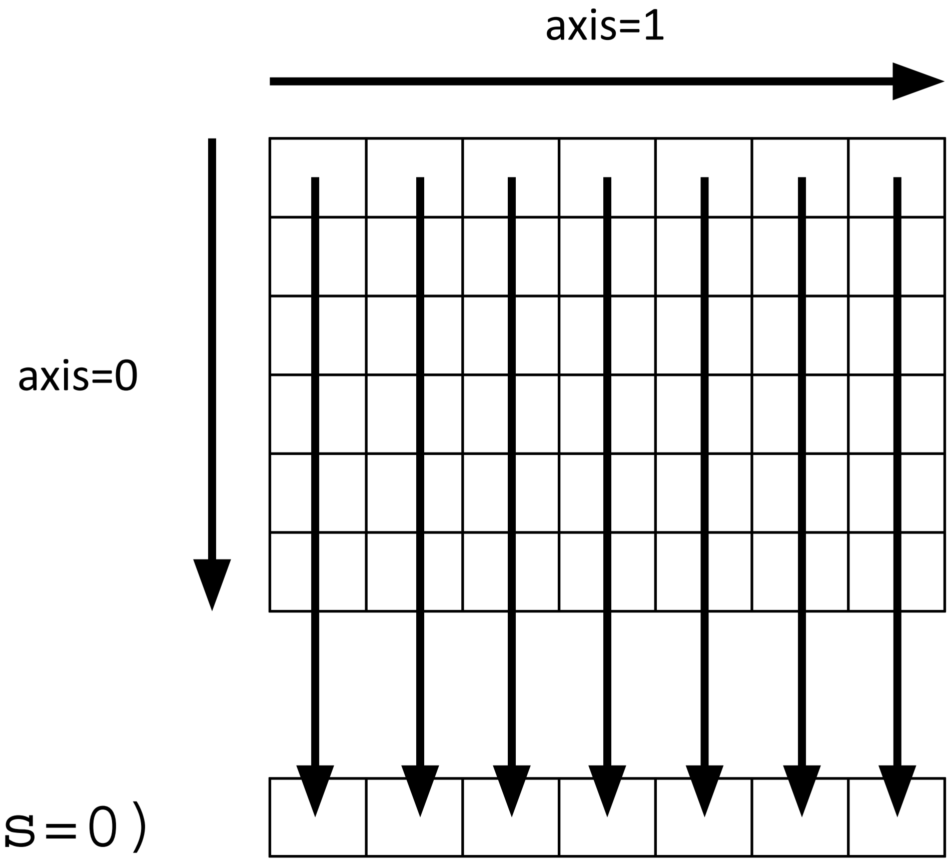
3 x 2-array

a b

a b

a b

# Axis



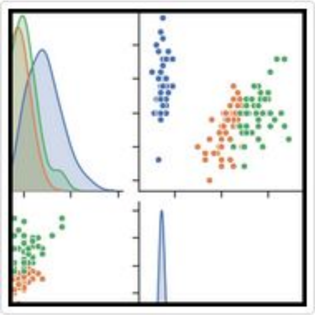
```
a.mean(axis=0)
```

# Plotting Ecosystem

matplotlib

plotly

Bokeh



Seaborn

# Scipy

- Linear Algebra (`scipy.linalg`)
- Optimization (`scipy.optimize`)
- Statistics (`scipy.stats`)

Many more

# Pandas

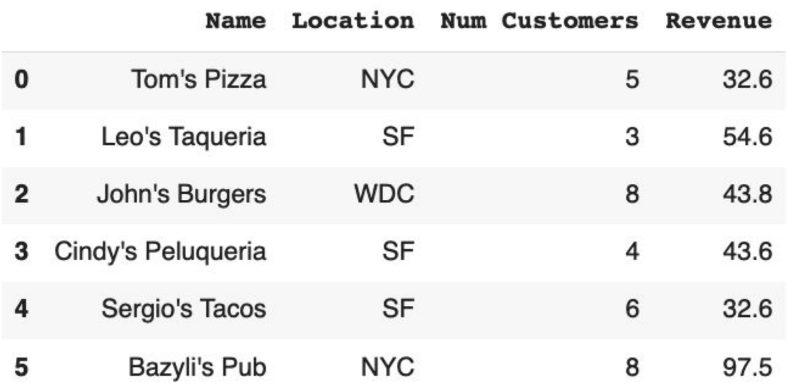
- Open-source, high-performances & easy-to-use data structures
- DataFrame objects
- Aggregation, grouping, reductions, statistics, etc.
- Powerful dates support
- All kinds of read/write functions (csv, HDF5, etc.)

# Accessing a DataFrame

- By Labels
  - `df[column]` `# Get one column`
  - `df[rows]` `# Get multiple rows`
  - `df.loc[cols,rows]` `# End-points INCLUDED`

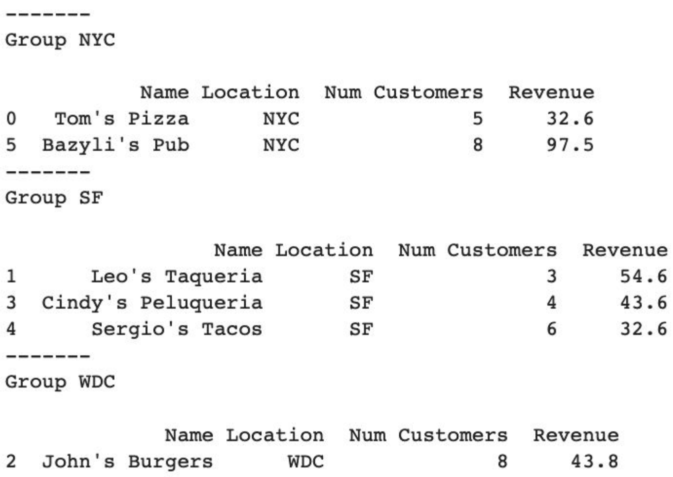
- By position
  - ``df.iloc[cols,rows]`` ``# End-points NOT INCLUDED w/ `:` ``

# Groupby



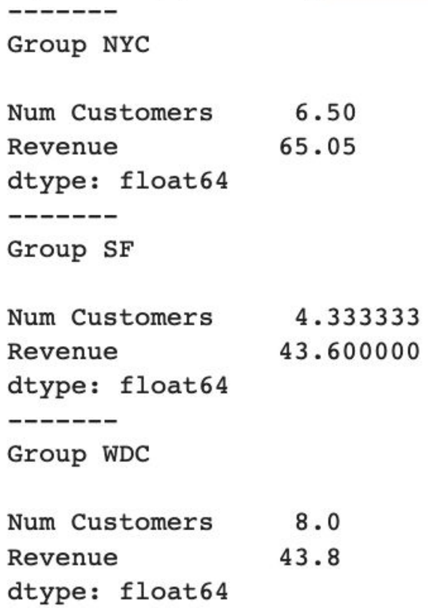
| | Name | Location | Num Customers | Revenue |
|---|---|---|---|---|
| 0 | Tom's Pizza | NYC | 5 | 32.6 |
| 1 | Leo's Taqueria | SF | 3 | 54.6 |
| 2 | John's Burgers | WDC | 8 | 43.8 |
| 3 | Cindy's Peluqueria | SF | 4 | 43.6 |
| 4 | Sergio's Tacos | SF | 6 | 32.6 |
| 5 | Bazyli's Pub | NYC | 8 | 97.5 |

Split



```
-------
Group NYC

             Name Location  Num Customers  Revenue
0    Tom's Pizza      NYC              5     32.6
5   Bazyli's Pub      NYC              8     97.5
-------
Group SF

                   Name Location  Num Customers  Revenue
1       Leo's Taqueria       SF              3     54.6
3   Cindy's Peluqueria       SF              4     43.6
4       Sergio's Tacos       SF              6     32.6
-------
Group WDC

               Name Location  Num Customers  Revenue
2   John's Burgers      WDC              8     43.8
```

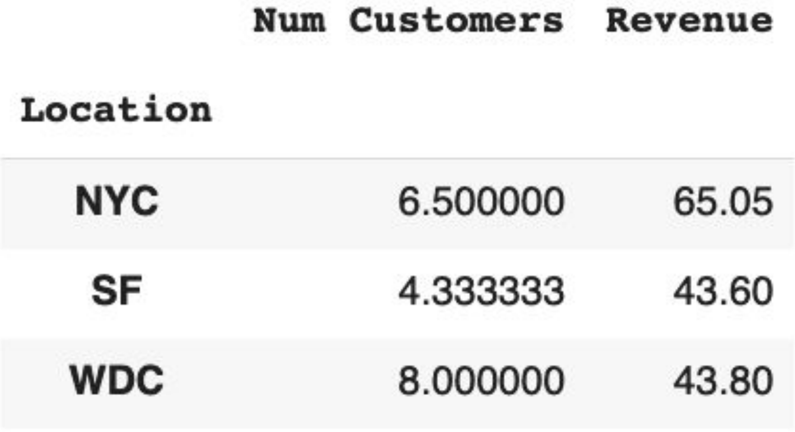
Transform



```
-------
Group NYC

Num Customers     6.50
Revenue          65.05
dtype: float64
-------
Group SF

Num Customers     4.333333
Revenue          43.600000
dtype: float64
-------
Group WDC

Num Customers     8.0
Revenue          43.8
dtype: float64
```

Combine

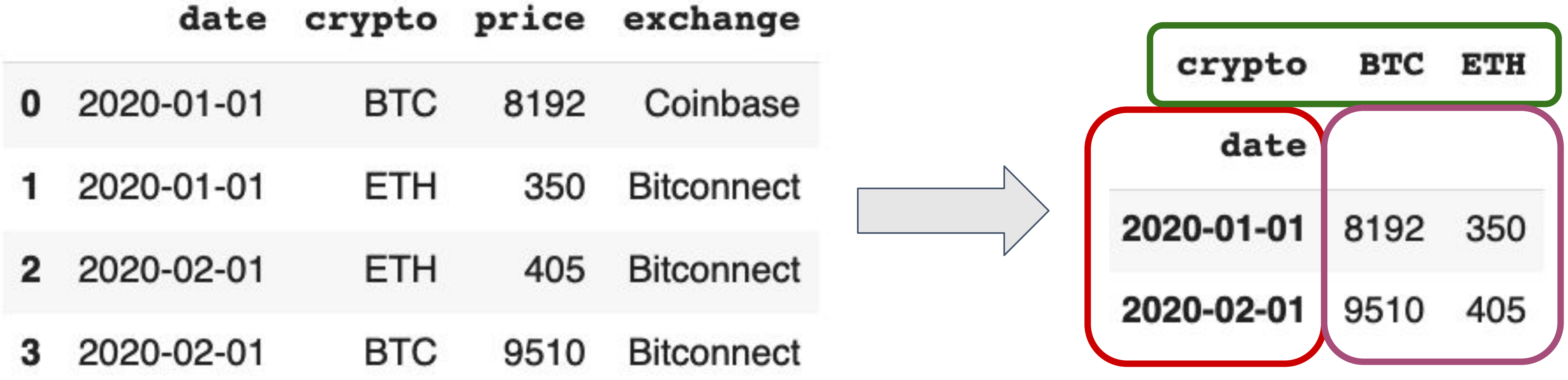
```
df.groupby('Location’).mean()
```



| | Num Customers | Revenue |
|---|---|---|
| Location | | |
| NYC | 6.500000 | 65.05 |
| SF | 4.333333 | 43.60 |
| WDC | 8.000000 | 43.80 |

# Pivot

```
df.pivot(index=’date’, columns=’crypto’, values=’price’)
```



|   | date | crypto | price | exchange |
|---|---|---|---|---|
| **0** | 2020-01-01 | BTC | 8192 | Coinbase |
| **1** | 2020-01-01 | ETH | 350 | Bitconnect |
| **2** | 2020-02-01 | ETH | 405 | Bitconnect |
| **3** | 2020-02-01 | BTC | 9510 | Bitconnect |

| **crypto** | **BTC** | **ETH** |
|---|---|---|
| **date** |  |  |
| **2020-01-01** | 8192 | 350 |
| **2020-02-01** | 9510 | 405 |

# Scikit-learn

- A package for machine learning

- Supervised learning
  - Classification
  - Regression

- Unsupervised

# Scikit-learn

A typical supervised learning problem

- Given a dataset

$$S = \{x_i, y_i\}$$

- Learns a function (mapping)

$$y = F(x)$$

Lots of kinds of models !

- https://scikit-learn.org/stable/supervised_learning.html
- https://scikit-learn.org/stable/unsupervised_learning.html
- https://scikit-learn.org/stable/model_selection.html
- … https://scikit-learn.org/stable/user_guide.html

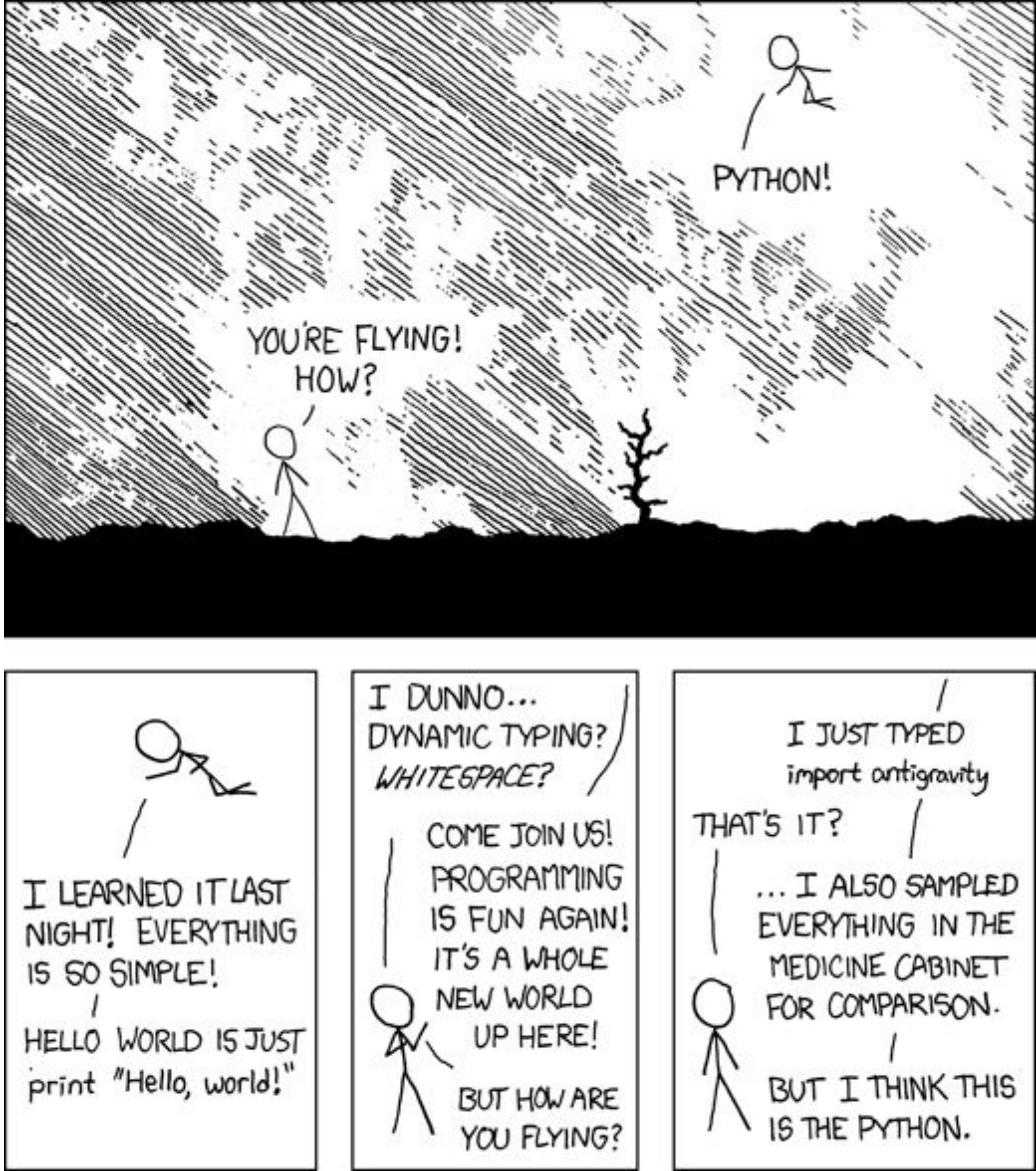
# Scikit-learn

```
# Pick a model
from sklearn import model
m = model.somemodel
# Train
m.fit(X_train, y_train)
# Predict
y_pred = m.predict(X_pred)
```

# Recap

```
import antigravity
# Try it on your laptop,
# In a Python interpreter
```

(https://xkcd.com/353/)


YOU'RE FLYING! HOW?
PYTHON!
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.

# More easter eggs

Try those in a Python interpreter

```
>> from __future__ import braces

>> import this

>> import __hello__
```

# Recap: What did we learn?

- Basic Python
- `Numpy` for arrays
- `Scipy` for linear algebra, optimization, statistics
- `Matplotlib` for simple plotting
- `Pandas` for data analytics
- `Scikit-learn` for machine learning

# The Python Ecosystem (a tiny subset)

NumPy

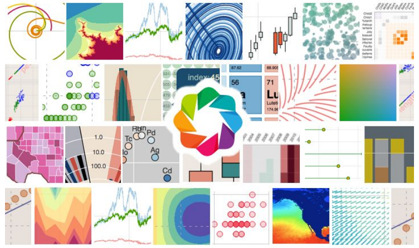
SciPy

jupyter

matplotlib

scikit learn

pandas



plotly

Seaborn

python

TensorFlow

Keras

PyTorch

theano

And many, many more (mpi4py, numba, joblib, …)

# References

- Python
  - Google & Stackoverflow
  - https://docs.python.org/3/
  - https://developers.google.com/edu/python/
  - https://www.learnpython.org
  - http://www.practicepython.org/
  - https://dabeaz-course.github.io/practical-python/Notes/Contents.html
- Numpy & Scipy
  - https://docs.scipy.org/doc/numpy/user/quickstart.html
- Pandas
  - https://pandas.pydata.org/pandas-docs/stable/10min.html
  - https://github.com/jvns/pandas-cookbook
- Scikit-learn
  - http://scikit-learn.org/stable/tutorial/basic/tutorial.html

# At Stanford

- CME211
  - Software Development for Scientists and Engineers
- CS106AP
  - Programming Methodology in Python
- CS102, CS131, CS230, CS231N, CS375: Machine Learning (using Python)
- CME302: Numerical Linear Algebra (using Python)

Best class at Stanford! Really!

# Any question after the class?

lcambier@stanford.edu

-> Stanford Continued Education