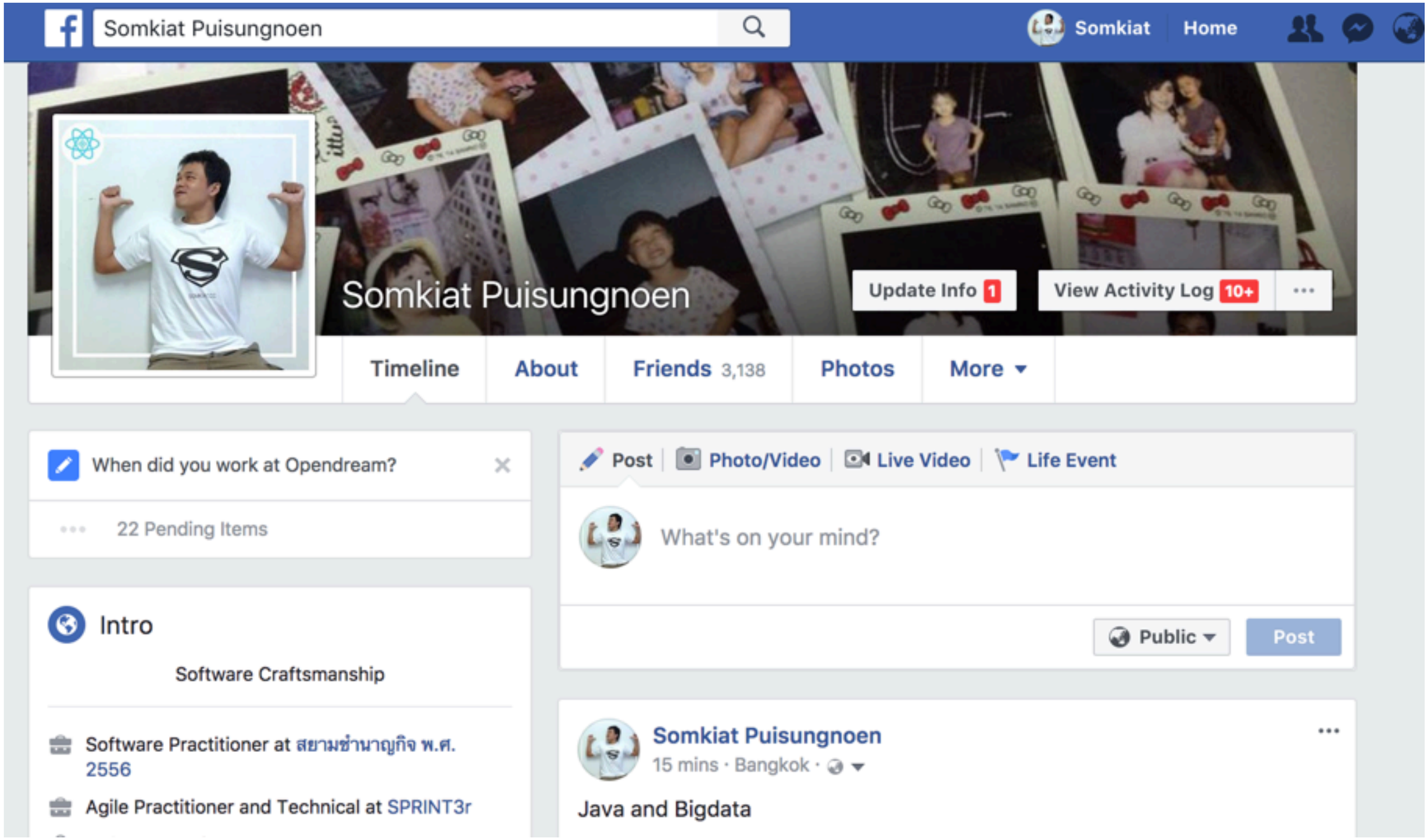
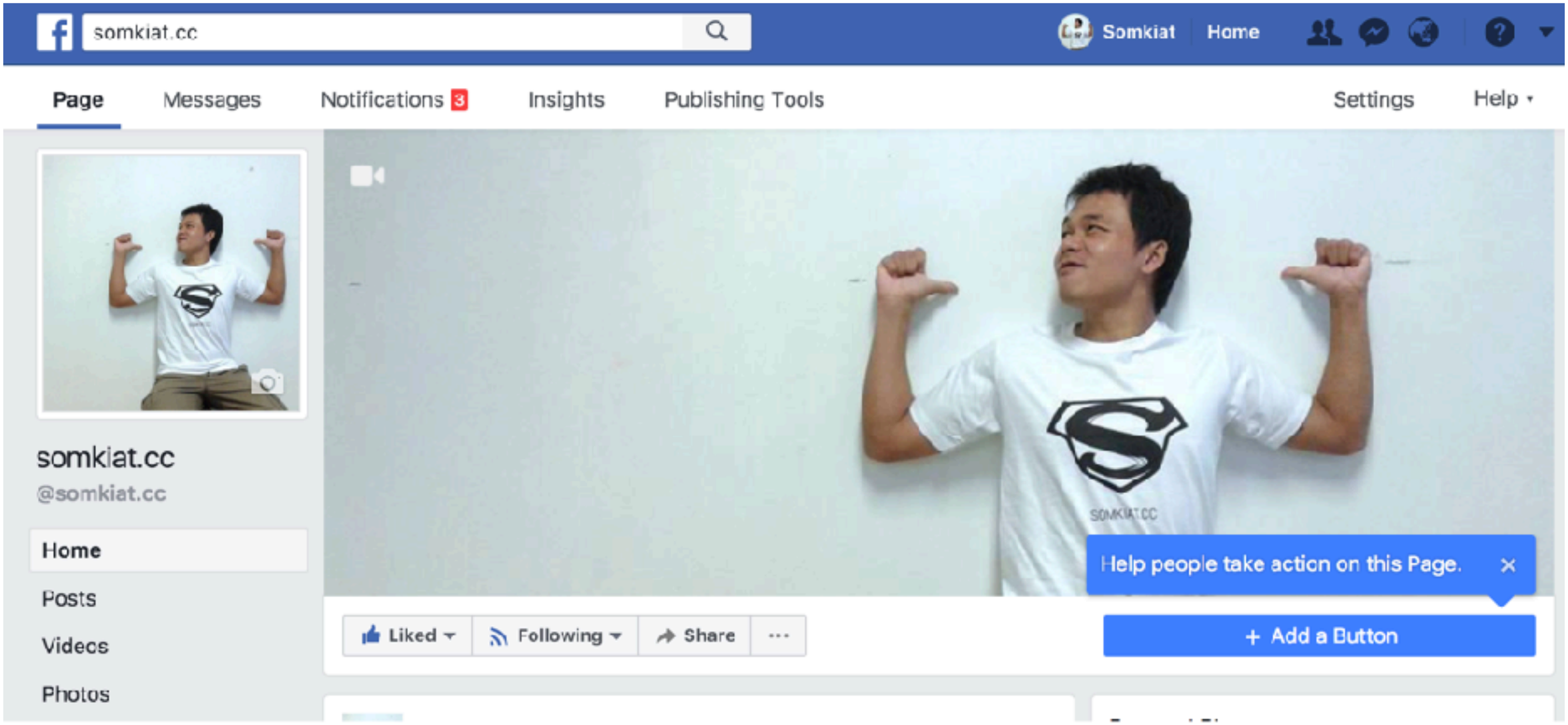
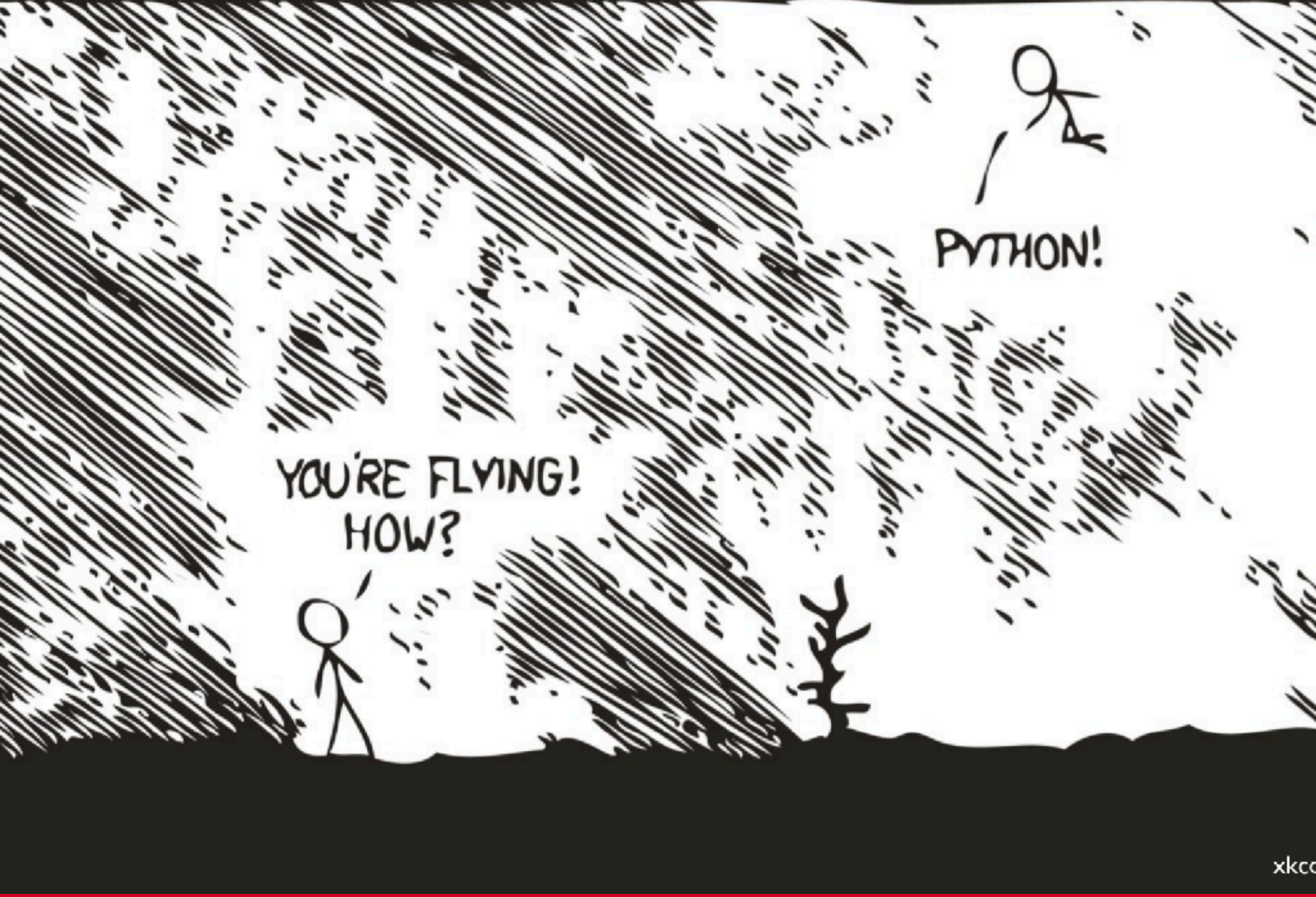
สยามชำนาญกิจ
SIAM CHAMNAN KIT

Python for Data Science





Advance Python for Data Science



Agenda

TDD for Python

Setup your computer => Python 3 and Jupyter

Summary of Python

List comprehensive

Workshop

Agenda

Data Science

Data Science with Python

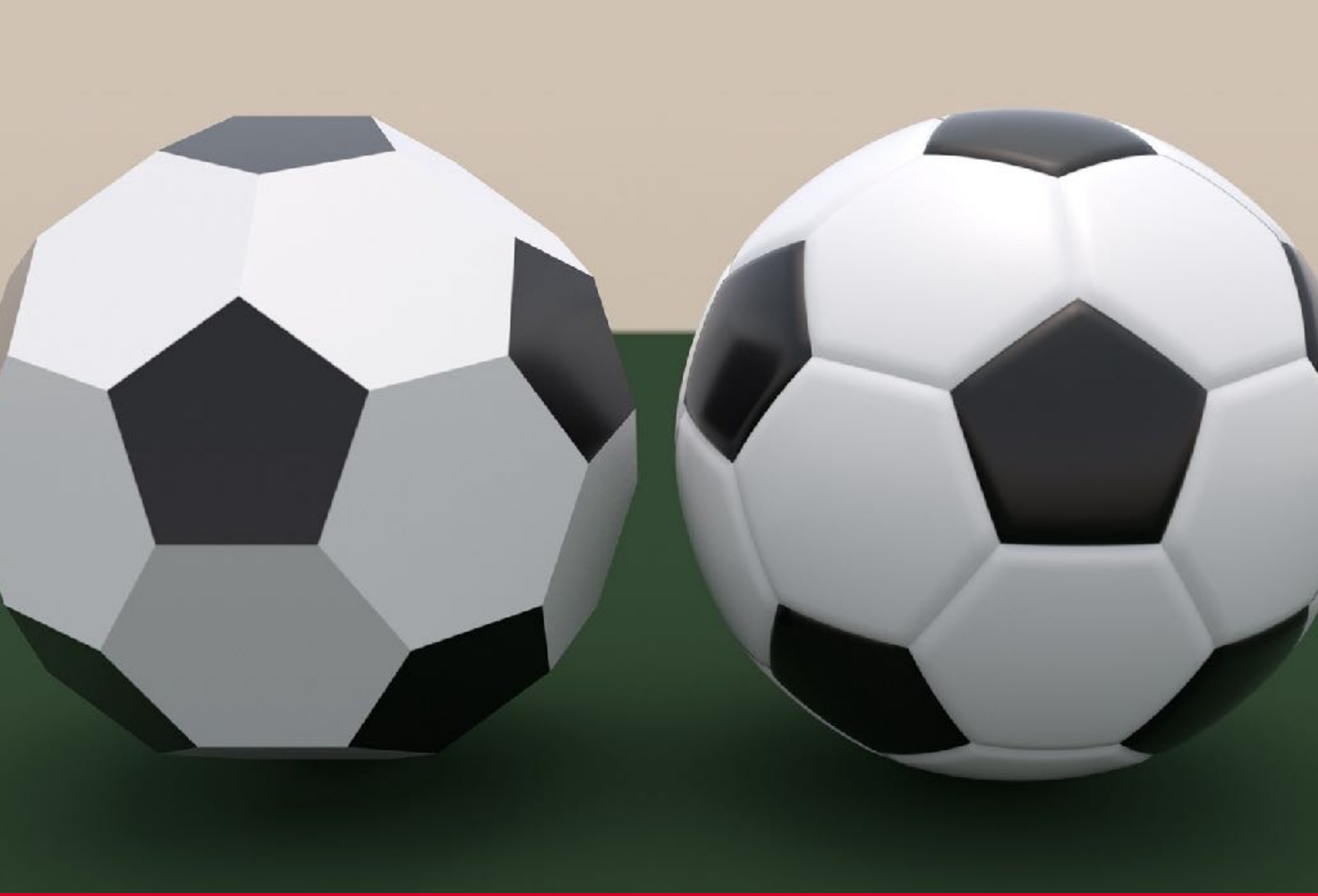
Numpy, Pandas and Matplotlib/Seaborn

Scikit-learn

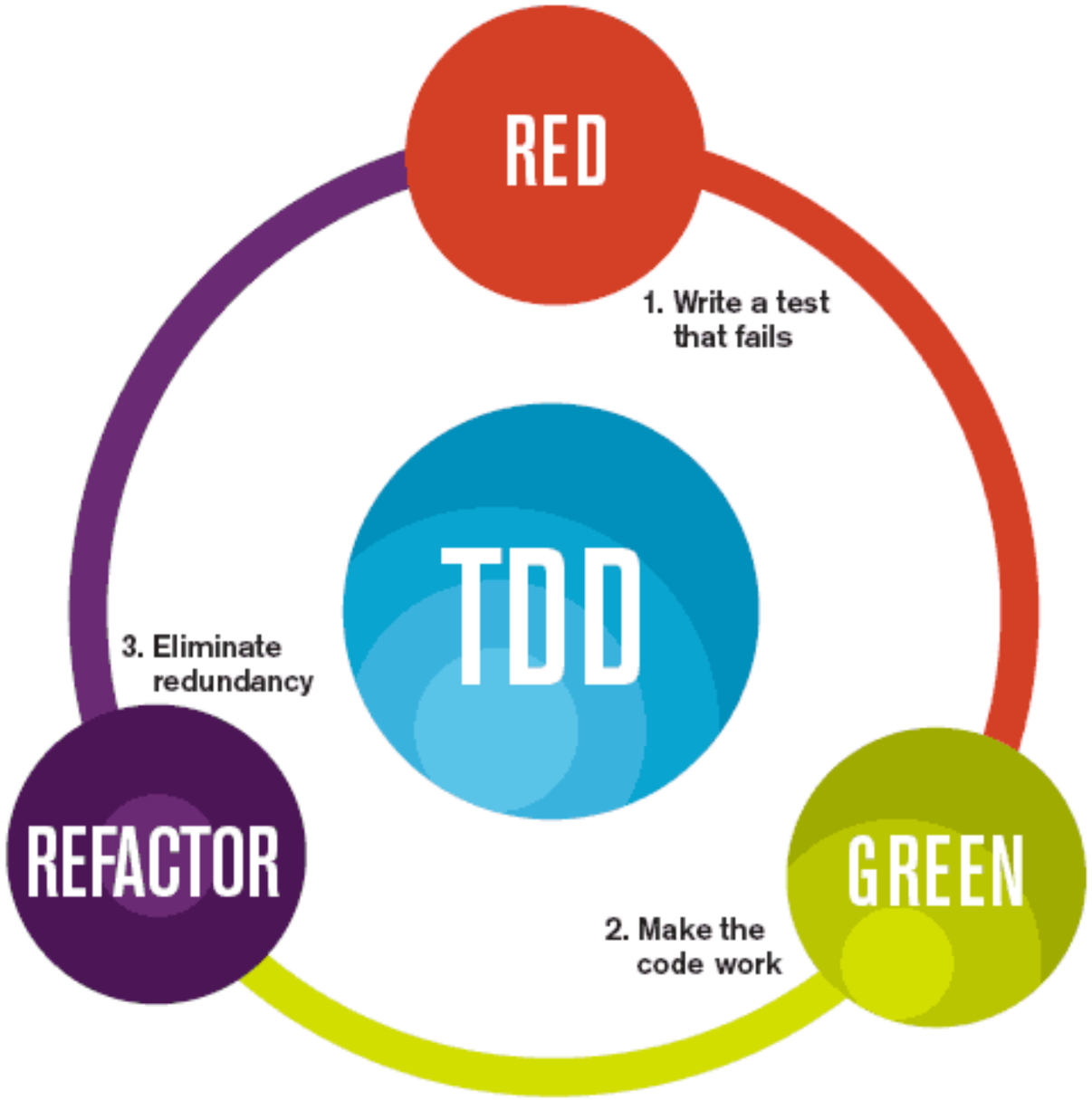
Kaggle :: Home for Data Science

Workshop

Assignment

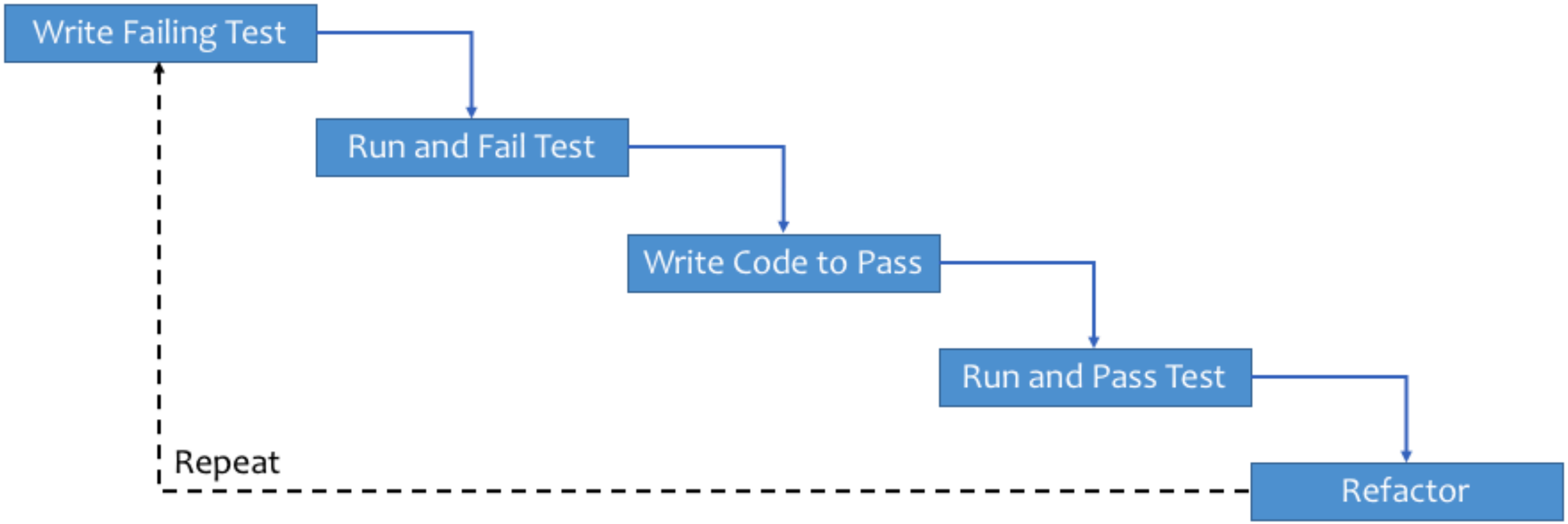


TDD for Data Science (Python)



The mantra of Test-Driven Development (TDD) is "red, green, refactor."

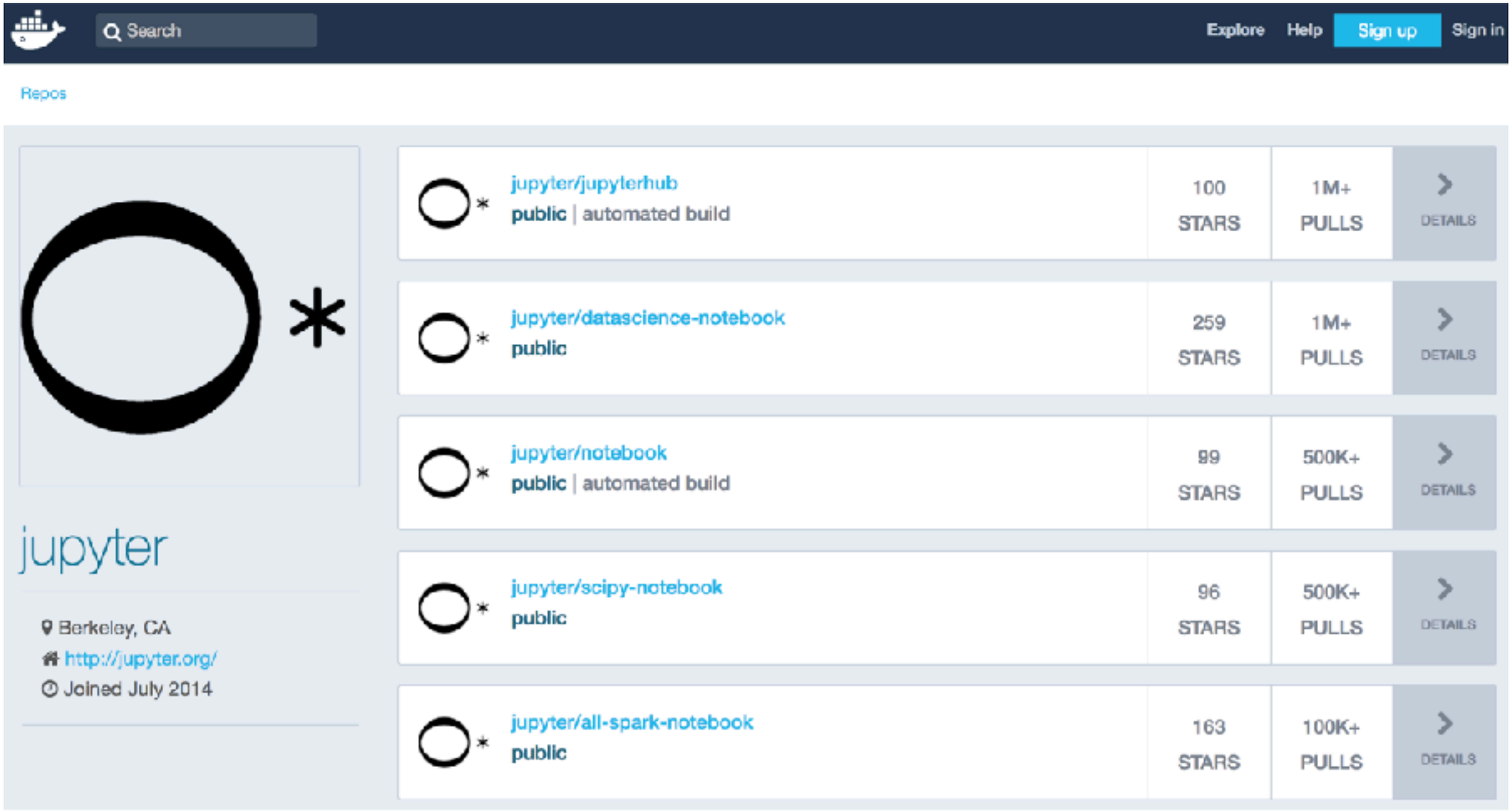
Workflow



Tools

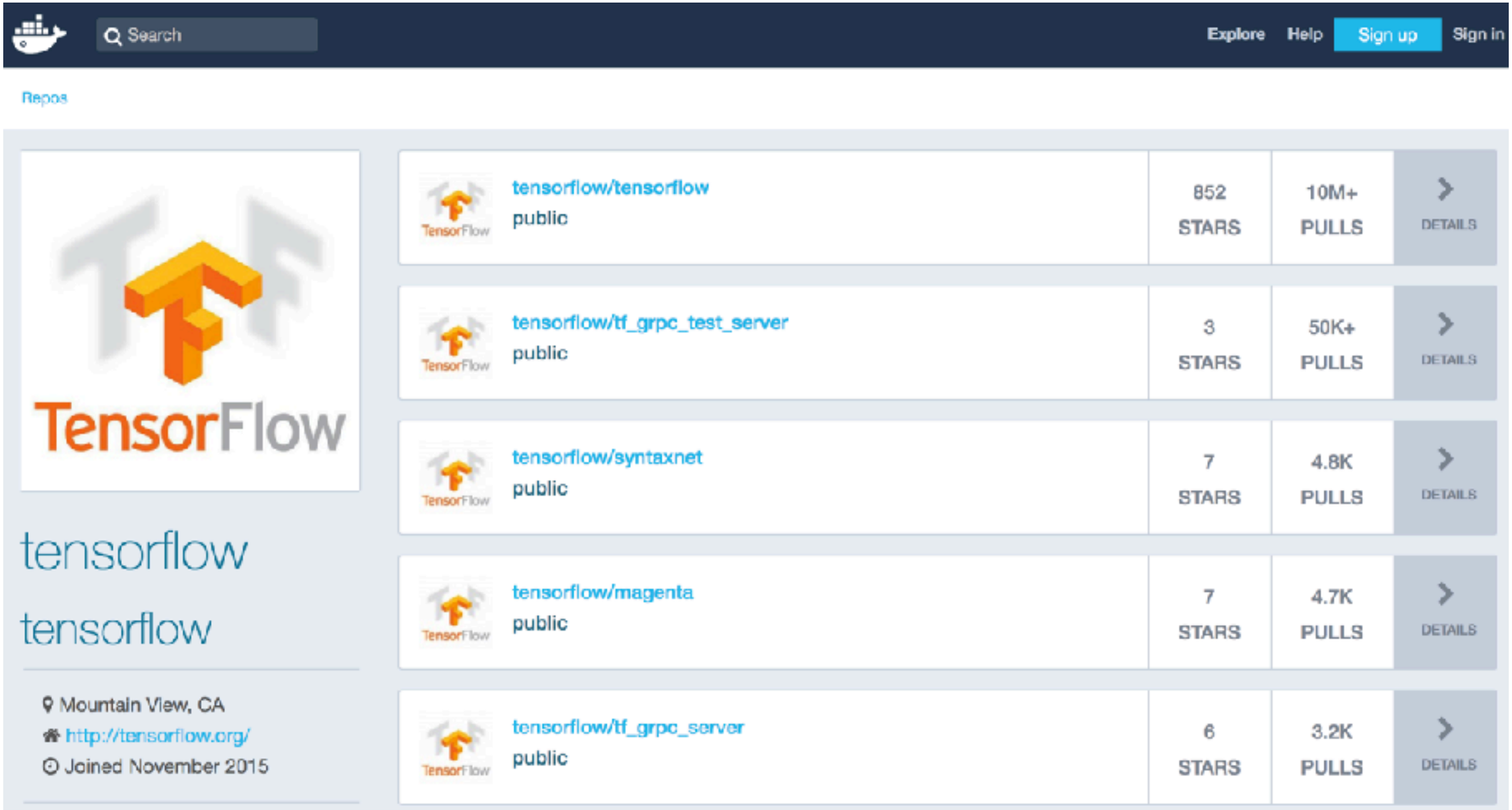
Docker for Data Science

Jupyter Images



https://hub.docker.com/u/jupyter/

TensorFlow Images



https://hub.docker.com/u/tensorflow/

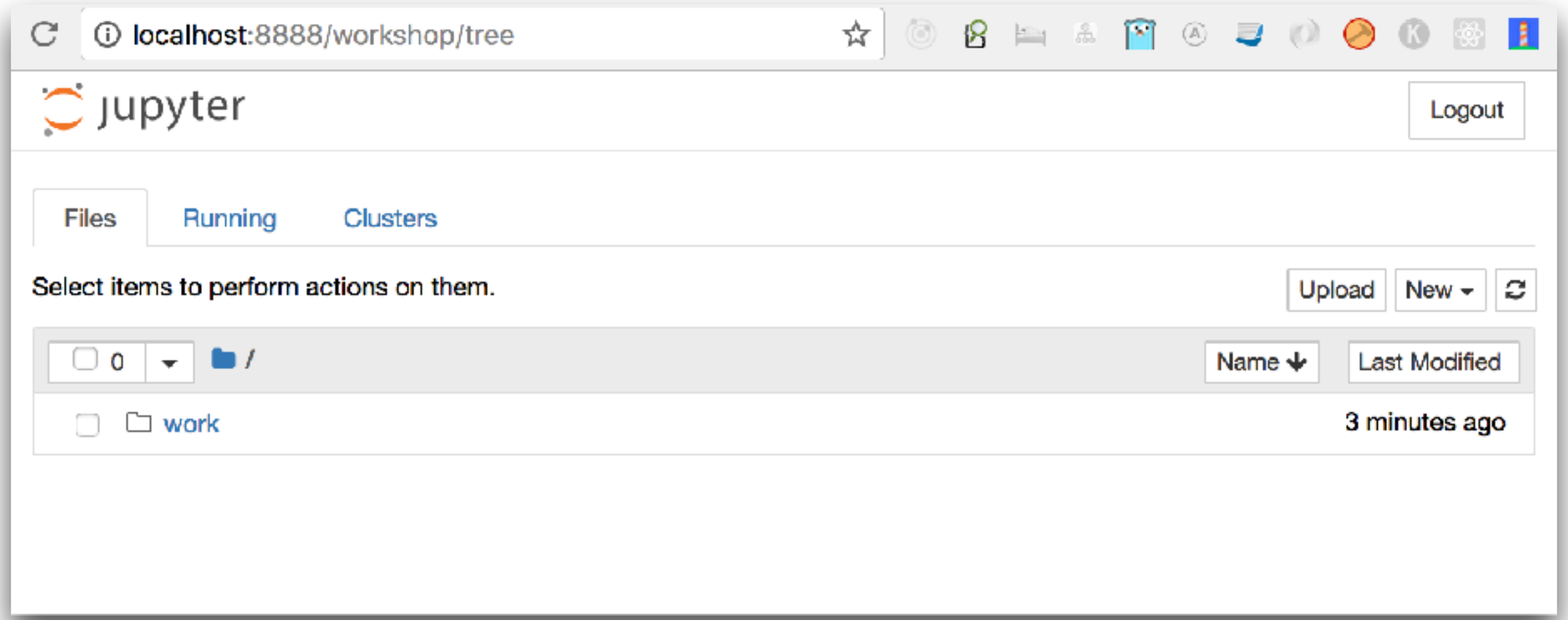
Install Jupyter with Docker

$docker pull jupyter/datascience-notebook

Install Jupyter with Docker

```
$docker container run -d -p 8888:8888
    -v $(pwd):/home/jovyan/work
    jupyter/datascience-notebook
    start-notebook.sh
    --NotebookApp.base_url=/workshop/
```

Hello Jupyter



Basic Data Types

int - Integer value
float - Decimal value
bool - True/False
complex - imaginary
NoneType - null value

Iterable data types

Type	Meaning
str	String immutable value
list	Collection of elements
tuple	Immutable list
dict	Unordered key-value pairs
set	Unordered collection of unique elements

Iterable data types

Type	How to use ?	Example
str	Defined with quotes	‘ab’
list	Defined with brackets	[‘a’, ‘b’]
tuple	Defined with parentheses	(‘a’, ‘b’)
dict	Defined with braces	{‘a’: 1, ‘b’: 2}
set	Defined with braces	{‘a’, ‘b’}

Control Flows

If-else statements
While loops
For loops

List comprehensive

Use for creating new list from another iterables
Introduced in Python 2.0
Python 3.0 comes with Dict and Set

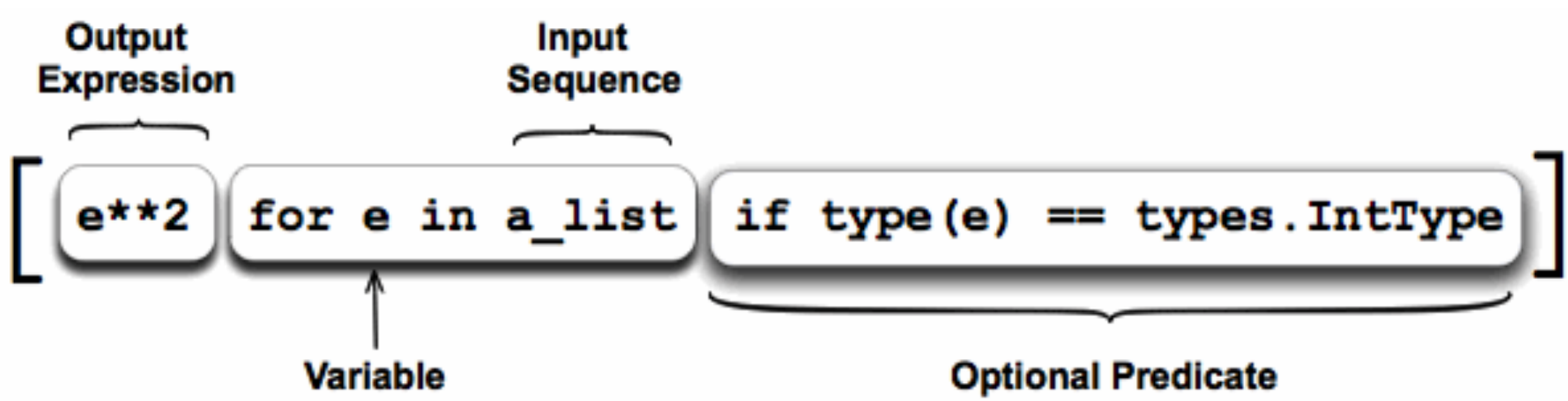
List comprehensive

Try to replace for loops and map(), filter(), reduce()

In Data Science working with List too much !!

List comprehensive

1. Input sequence
2. Variable of input sequence
3. Optional predicate expression
4. Output expression



Example 1

Square of number

```
def calculate():
    numbers = [1, 2, 3, 4, 5]
    results = []
    for number in numbers:
        results.append(number**2)
    print(results)

if __name__== "__main__":
    calculate()
```

Rewrite with List comprehensive

Square of number

```
def calculate():
    numbers = [1, 2, 3, 4, 5]
    result = [number**2 for number in numbers]
    print(result)

if __name__== "__main__":
    calculate()
```

Example 2

Find the same number in 2 lists

```
def process():
    list1 = [1, 2, 3, 4, 5]
    list2 = [3, 4, 5, 6, 7]
    results = []
    for x in list1:
        for y in list2:
            if x == y:
                results.append(x)

    print(results)

if __name__== "__main__":
    process()
```

Rewrite with List comprehensive

Find the same number in 2 lists of number

```
def process():
    list1 = [1, 2, 3, 4, 5]
    list2 = [3, 4, 5, 6, 7]
    results = [x for x in list1 for y in list2 if x==y]
    print(results)

if __name__== "__main__":
    process()
```

Example 3

Replace number with string (Even and Odd)

```
def process():
    numbers = [1, 2, 3, 4, 5]
    results = []
    for number in numbers:
        if number%2 == 0:
            results.append("Even")
        else:
            results.append("Odd")
    print(results)

if __name__== "__main__":
    process()
```

Rewrite with List comprehensive

Replace number with string (Even and Odd)

```
def process():
    numbers = [1, 2, 3, 4, 5]
    results = ["Even" if number%2 == 0 else "Odd" for number in numbers]
    print(results)

if __name__== "__main__":
    process()
```

Example 4

Remove vowels from sentence

```
def process(sentence):
    vowels = 'aeiou'
    results = []
    for c in sentence:
        if c not in vowels:
            results.append(c)
    return ''.join(results)

if __name__== "__main__":
    print(process('Hello World'))
```

Rewrite with List comprehensive

Remove vowels from sentence

```
def process(sentence):
    vowels = 'aeiou'
    return ''.join([c for c in sentence if c not in vowels])

if __name__== "__main__":
    print(process('Hello World'))
```

Try to practices

Practice python skills

Project Euler.net

Weekly Python Exercise

Project Euler (616 problems)

https://projecteuler.net/archives

ID	Description / Title	Solved By
1	Multiples of 3 and 5	749135
2	Even Fibonacci numbers	602067
3	Largest prime factor	431141
4	Largest palindrome product	383883
5	Smallest multiple	392340
6	Sum square difference	394673
7	10001st prime	337853
8	Largest product in a series	285474
9	Special Pythagorean triplet	288258
10	Summation of primes	264489

1. Multiples of 3 and 5

https://projecteuler.net/problem=1

Problem 1

If we list all the natural numbers below 10 that are multiples of 3 or 5, we get 3, 5, 6 and 9. The sum of these multiples is 23.

Find the sum of all the multiples of 3 or 5 below 1000.

2. Even Fibonacci Numbers

https://projecteuler.net/problem=2

Problem 2

Each new term in the Fibonacci sequence is generated by adding the previous two terms. By starting with 1 and 2, the first 10 terms will be:

1, 2, 3, 5, 8, 13, 21, 34, 55, 89, ...

By considering the terms in the Fibonacci sequence whose values do not exceed four million, find the sum of the even-valued terms.

6. Sum square difference

https://projecteuler.net/problem=6

Problem 6

The sum of the squares of the first ten natural numbers is,

$$1^2 + 2^2 + \ldots + 10^2 = 385$$

The square of the sum of the first ten natural numbers is,

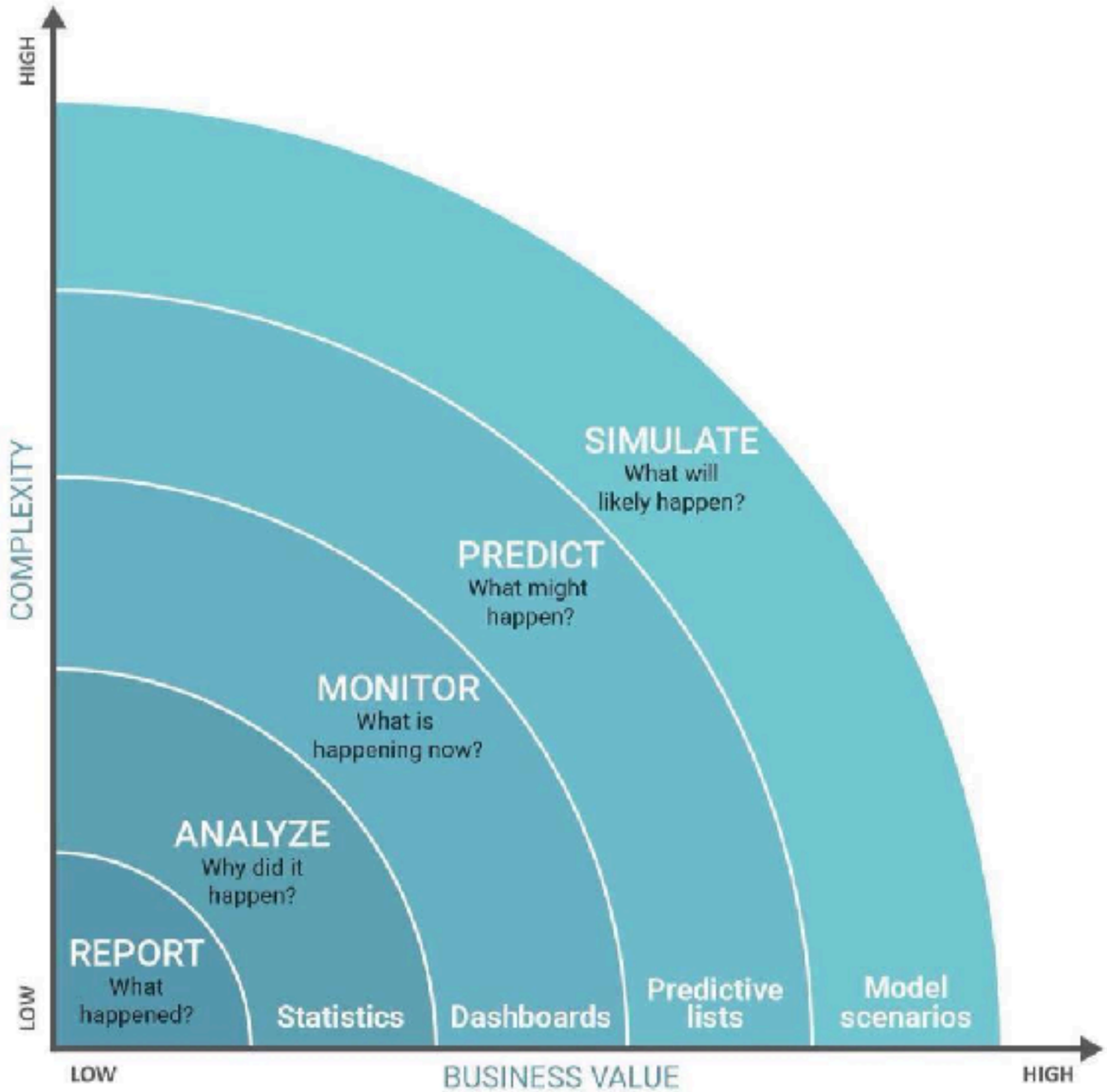
$$(1 + 2 + \ldots + 10)^2 = 55^2 = 3025$$

Hence the difference between the sum of the squares of the first ten natural numbers and the square of the sum is 3025 – 385 = 2640.

Find the difference between the sum of the squares of the first one hundred natural numbers and the square of the sum.

Data Science

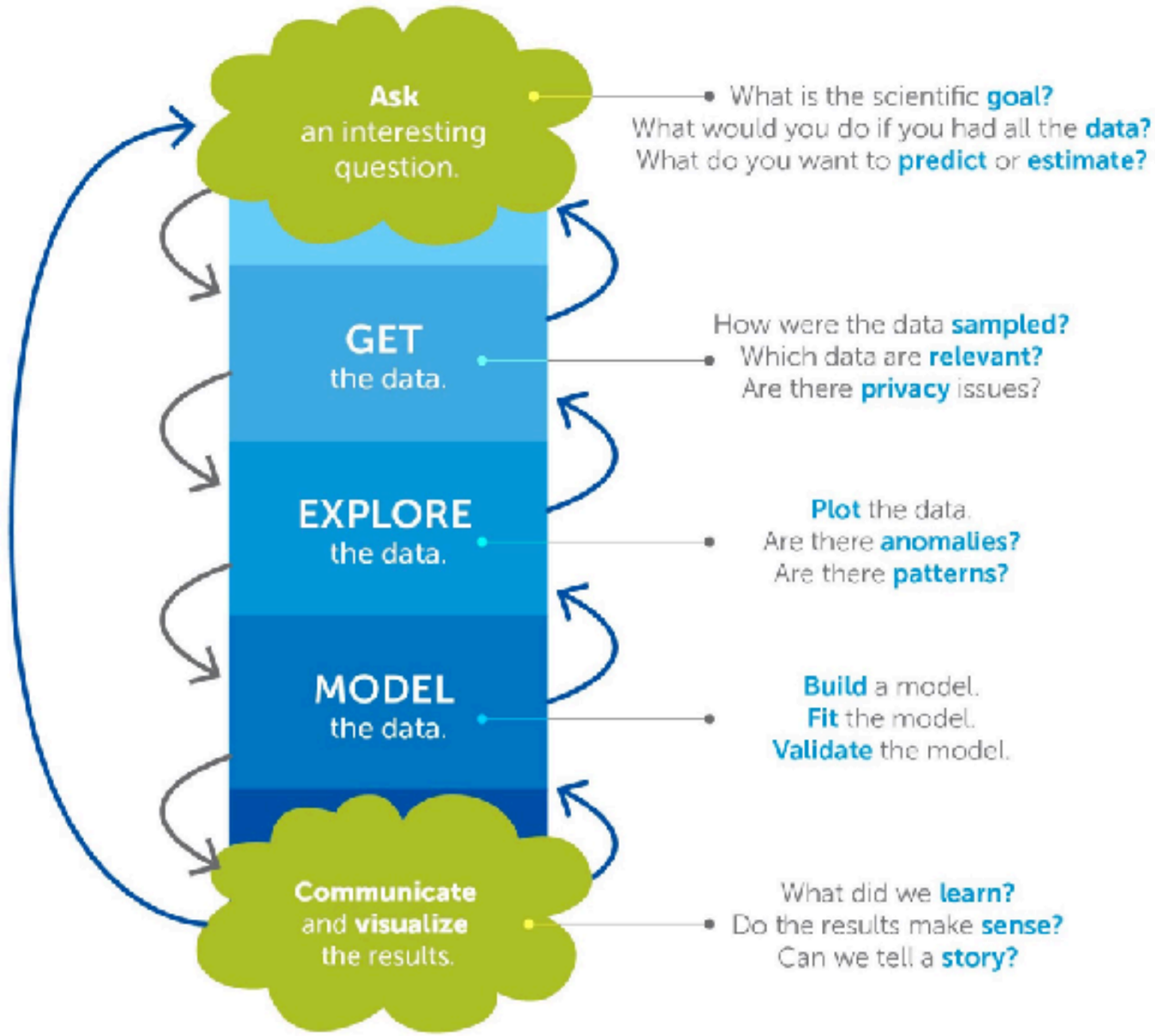
Levels of Data Science



Data Science Process

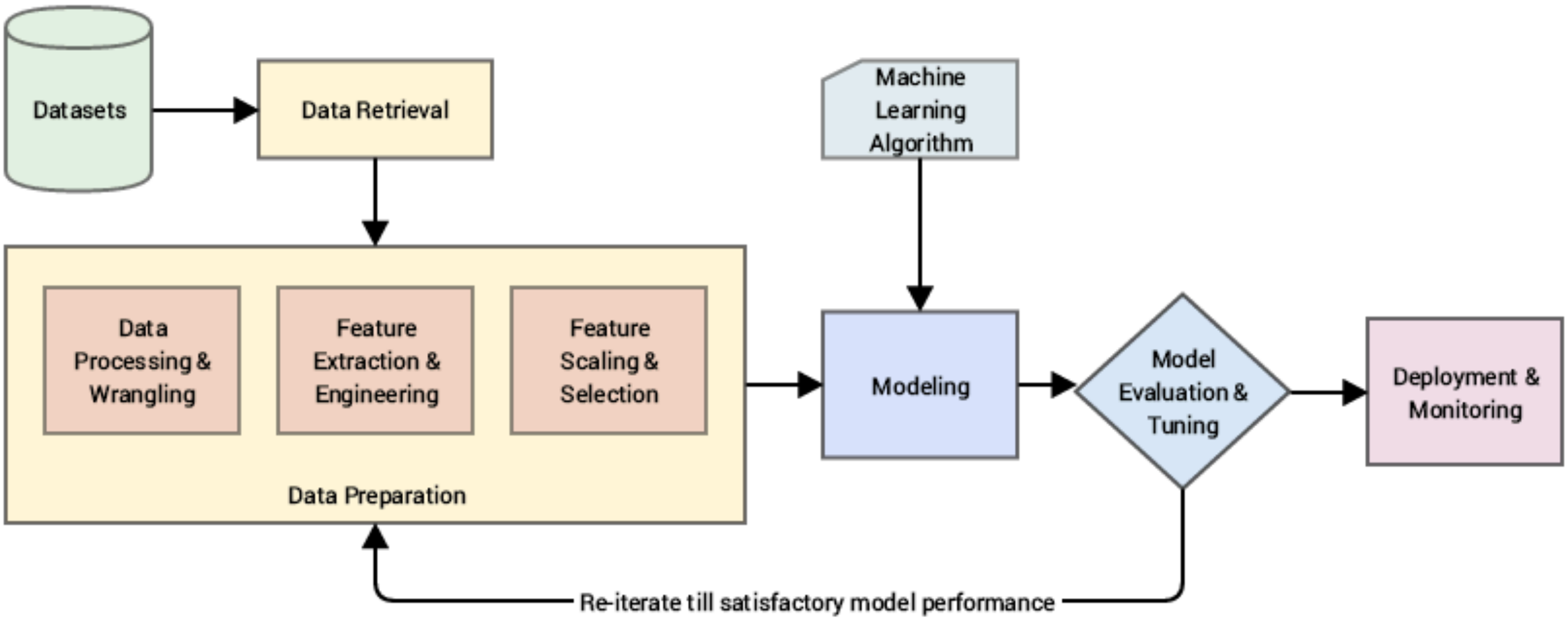
1. Collect the raw data needed to solve problem
2. Process the data (data wrangling)
3. Explore the data (data visualization)
4. Perform in-depth analysis (ML, Statistic, Algorithm)
5. Communicate result of the analysis

The Data Science Process



Derived from the work of Joe Blitzstein and Hanspeter Pfister, originally created for the Harvard data science course http://cs109.org/.

Data Science Process

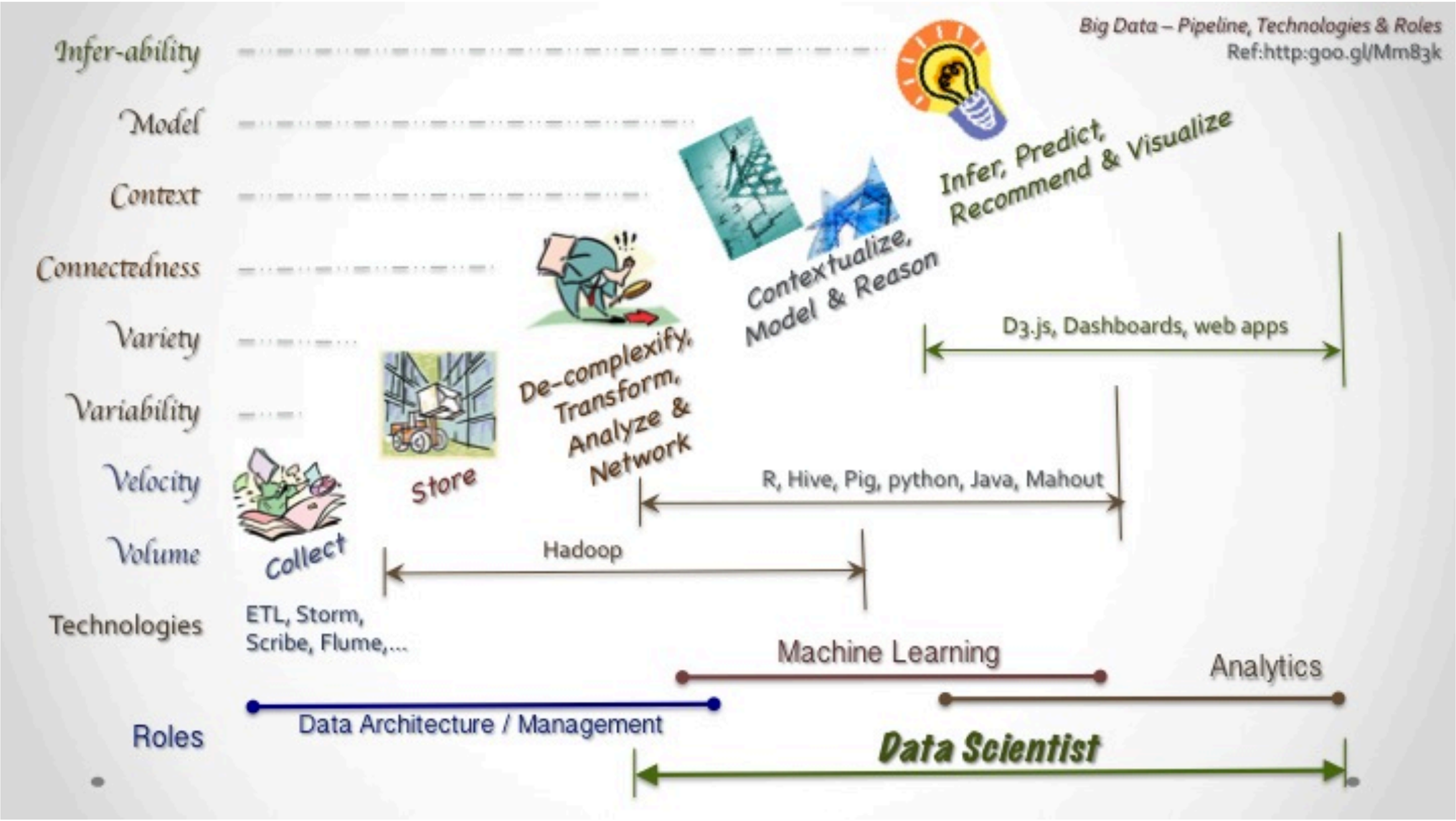


Data Science Skills

	Data Analyst	Machine Learning Engineer	Data Engineer	Data Scientist
Programming Tools	Very important	Very important	Very important	Very important
Data Visualization and Communication	Very important	Somewhat important	Somewhat important	Very important
Data Intuition	Somewhat important	Very important	Somewhat important	Very important
Statistics	Somewhat important	Very important	Somewhat important	Very important
Data Wrangling	Not that important	Not that important	Very important	Very important
Machine Learning	Not that important	Very important	Not that important	Very important
Software Engineering	Not that important	Somewhat important	Very important	Somewhat important
Multivariable Calculus and Linear Algebra	Not that important	Very important	Not that important	Somewhat important

Not that important · Somewhat important · Very important

Operation under Data Science



Exploratory Analysis

Check data how it is scattered
Data dimension
Column name
Unique and grouping values
Missing values

Feature Engineering

Create additional relevance features from the existing features in the raw data.

Try to increase the predictive power of the learning algorithm.

Data Manipulation

The process of changing data in an effort to make it easier to read and organize.

Exploratory Data Analysis (EDA)

Seeing what the data can tell us beyond the formal modeling or hypothesis testing tasks.

Exploratory Data Analysis (EDA)

The approach to analyzing datasets to summarize their main characteristics, **often with visual methods**.

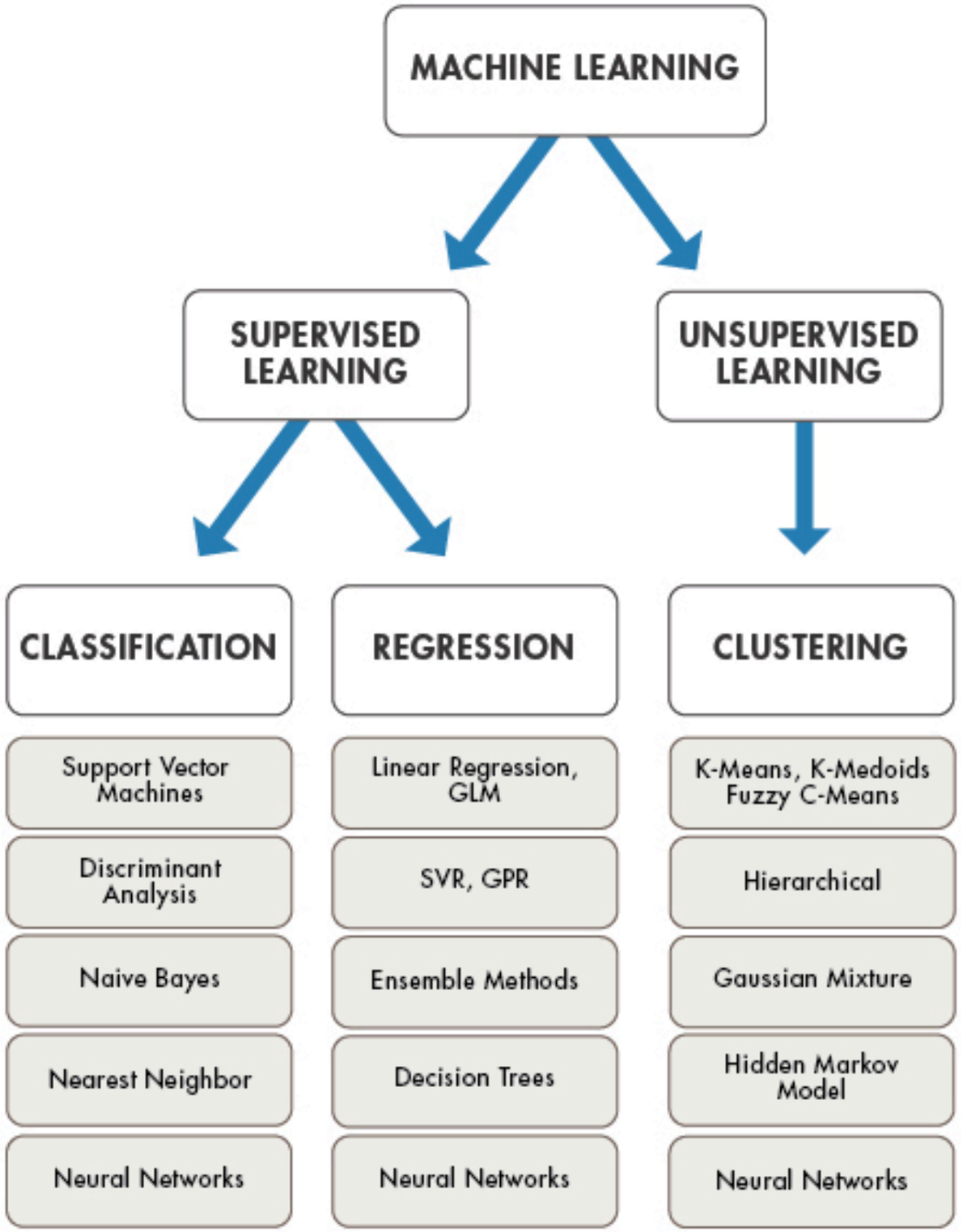
Machine Learning (ML)

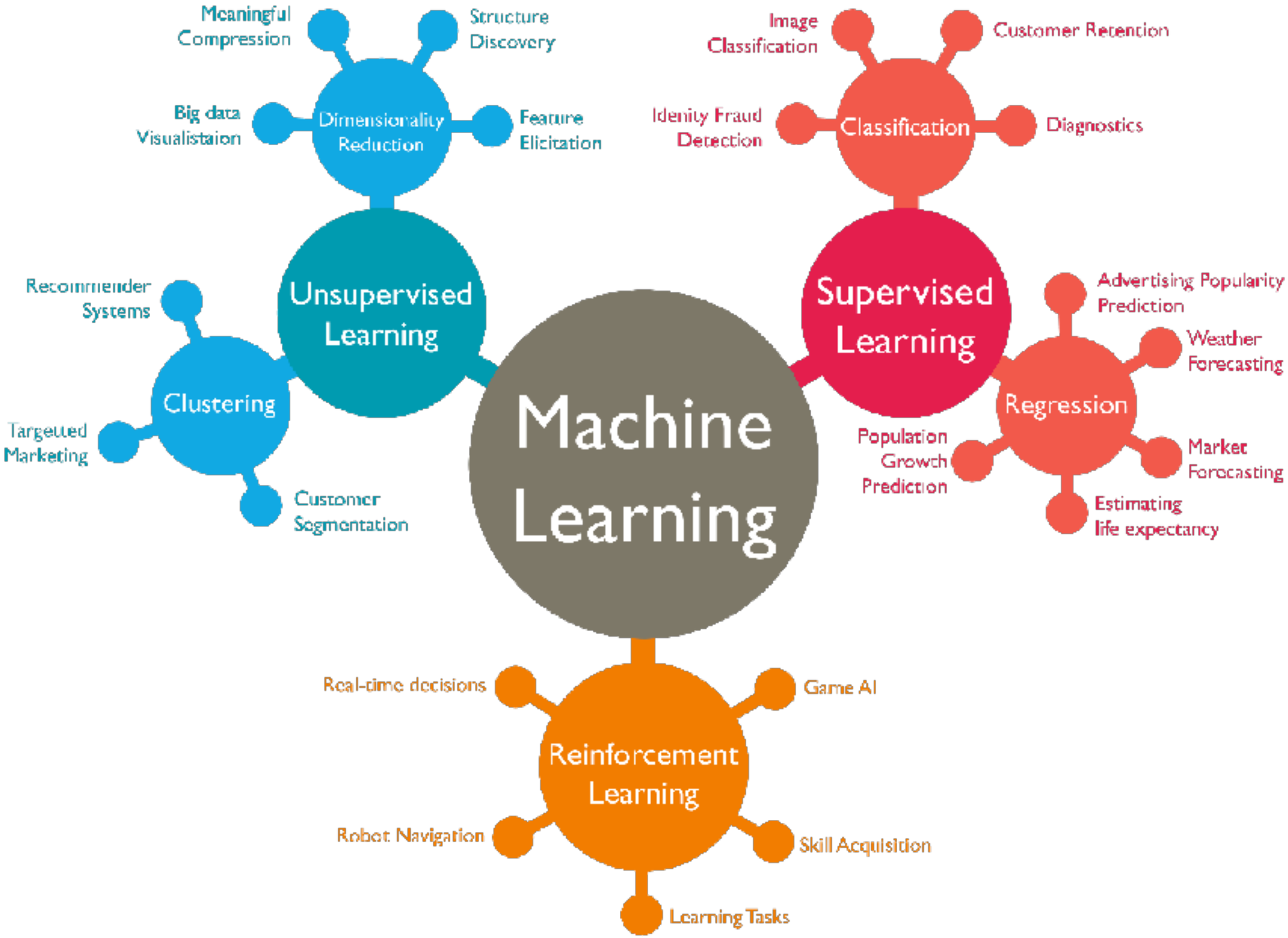
Machine Learning

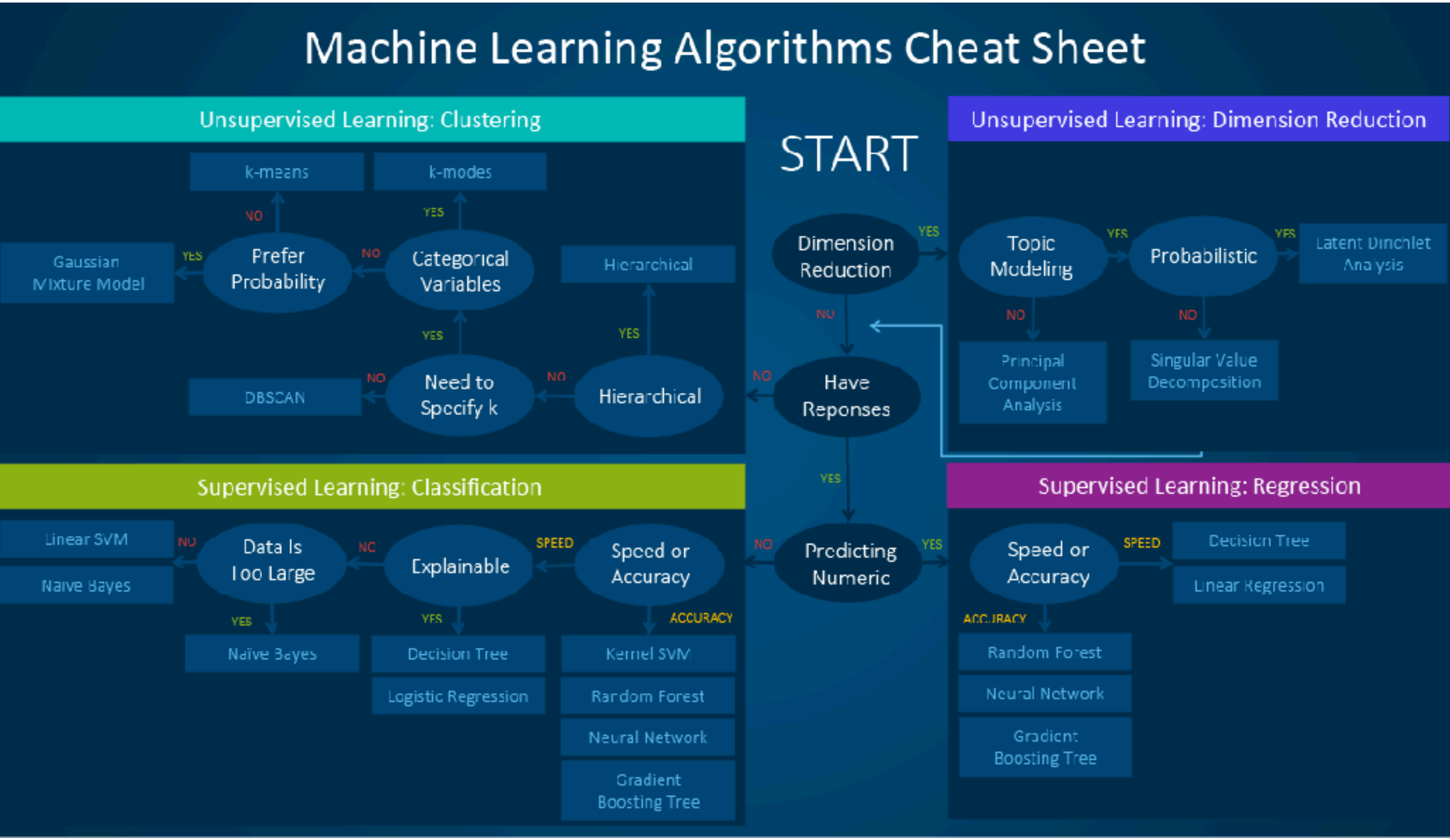
The application of AI that provides system the ability to **automatically learn and improve** experience without explicitly programmed.

Machine Learning

Focus on the development of computer program that can access data and use it to learn themselves.





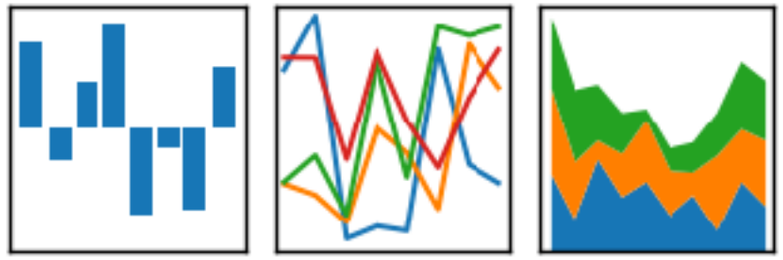


Libraries for Data Science

NumPy

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

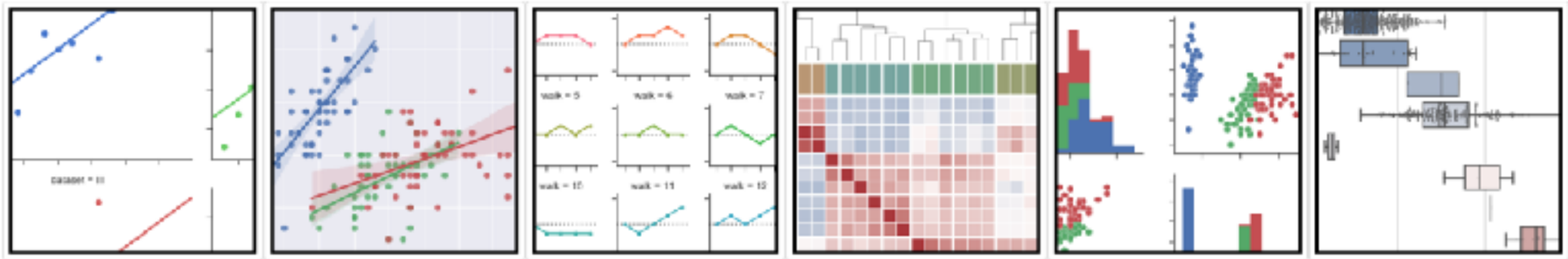


matplotlib

scikits learn

machine learning in Python

seaborn: statistical data visualization

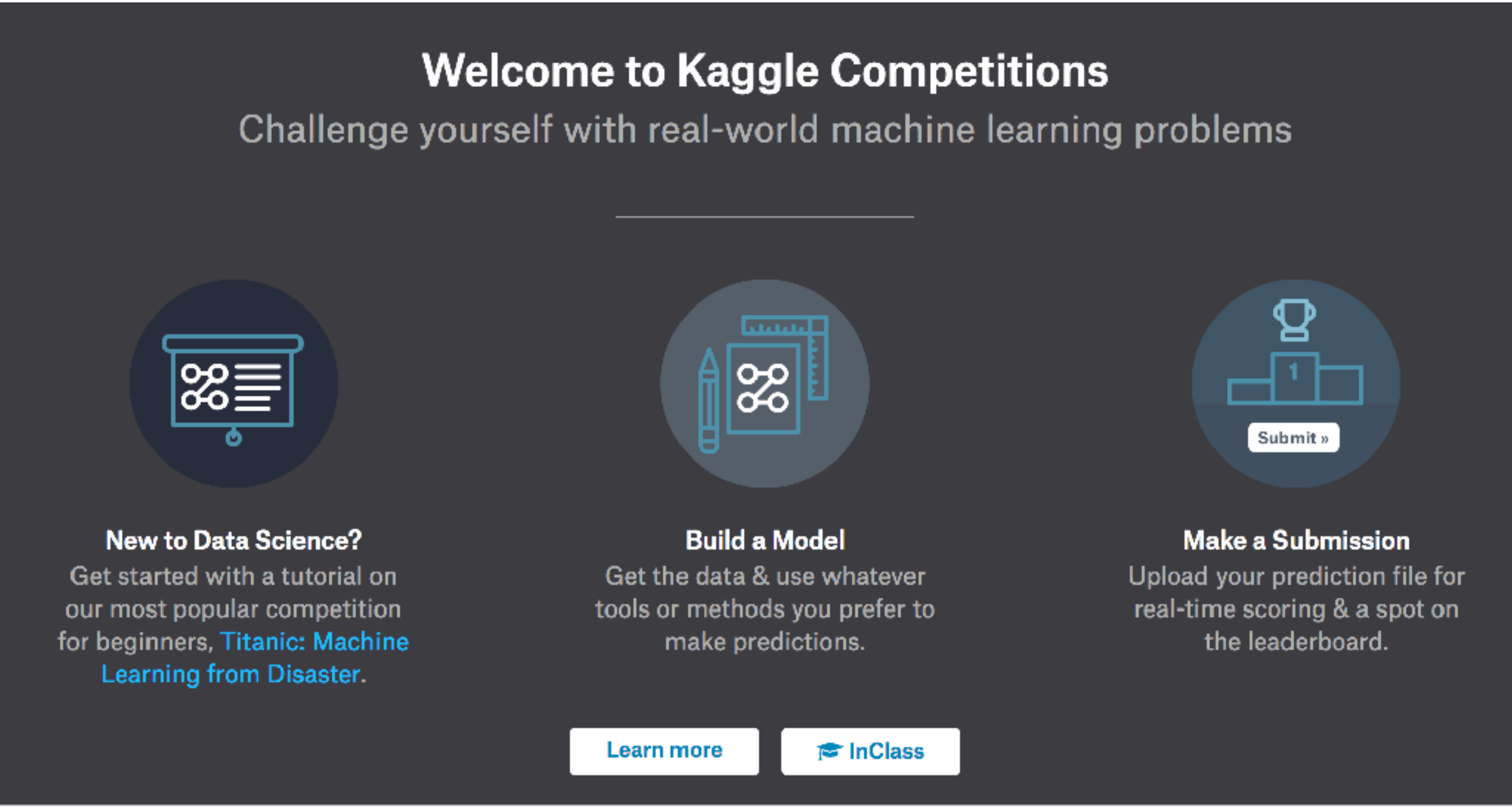


Working with data

Python for Data Science Cheat Sheet

https://s3.amazonaws.com/assets.datacamp.com/blog_assets/PythonForDataScience.pdf

Home of Data Science



https://github.com/up1/course-python-for-data-science