Week1-1 : Introduction

What is Machine Learning?

What is Statistics?

How about AI, Data Science, and Deep Learning?

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2017, VOL. 26, NO. 4, 745–766
https://doi.org/10.1080/10618600.2017.1384734

50 Years of Data Science

David Donoho

Department of Statistics, Stanford University, Standford, CA

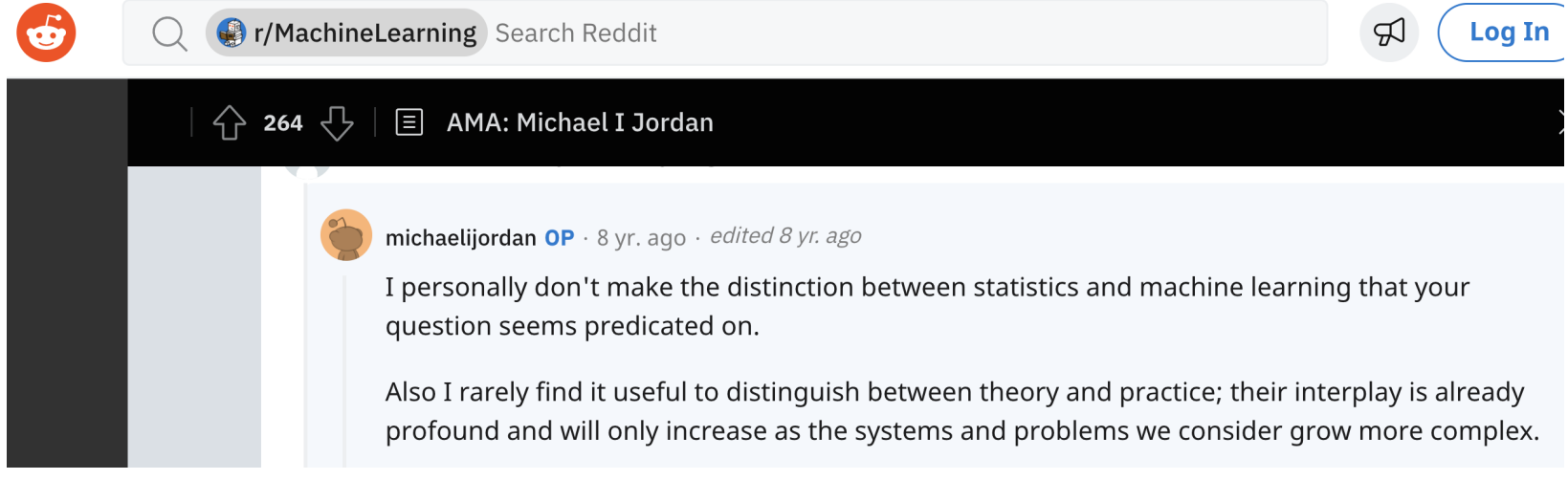
"**Statistics**" means the practice or science of collecting and analyzing numerical data in large quantities.

"**Data Scientist**" means a professional who uses scientific methods to liberate and create meaning from raw data.

"**Machine learning**" is a rapidly growing field at the intersection of computer science and statistics concerned with finding patterns in data.

It is responsible for tremendous advances in technology, from personalized product recommendations to speech recognition in cell phones. This course provides a broad introduction to the key ideas in machine learning. The emphasis will be on intuition and practical examples rather than theoretical results, though some experience with probability, statistics, and linear algebra will be important.

Michael I. Jordan


r/MachineLearning
Search Reddit
Log In
264
AMA: Michael I Jordan
michaelijordan OP · 8 yr. ago · edited 8 yr. ago
I personally don't make the distinction between statistics and machine learning that your question seems predicated on.
Also I rarely find it useful to distinguish between theory and practice; their interplay is already profound and will only increase as the systems and problems we consider grow more complex.

Today's data science movement

Source: http://veronikarock.com/teaching/01_slides.pdf

Just for fun

(Photo by Andrea Piacquadio from Pexels)

HEART HEALTH, INTELLIGENCE, SCIENCE & TECHNOLOGY

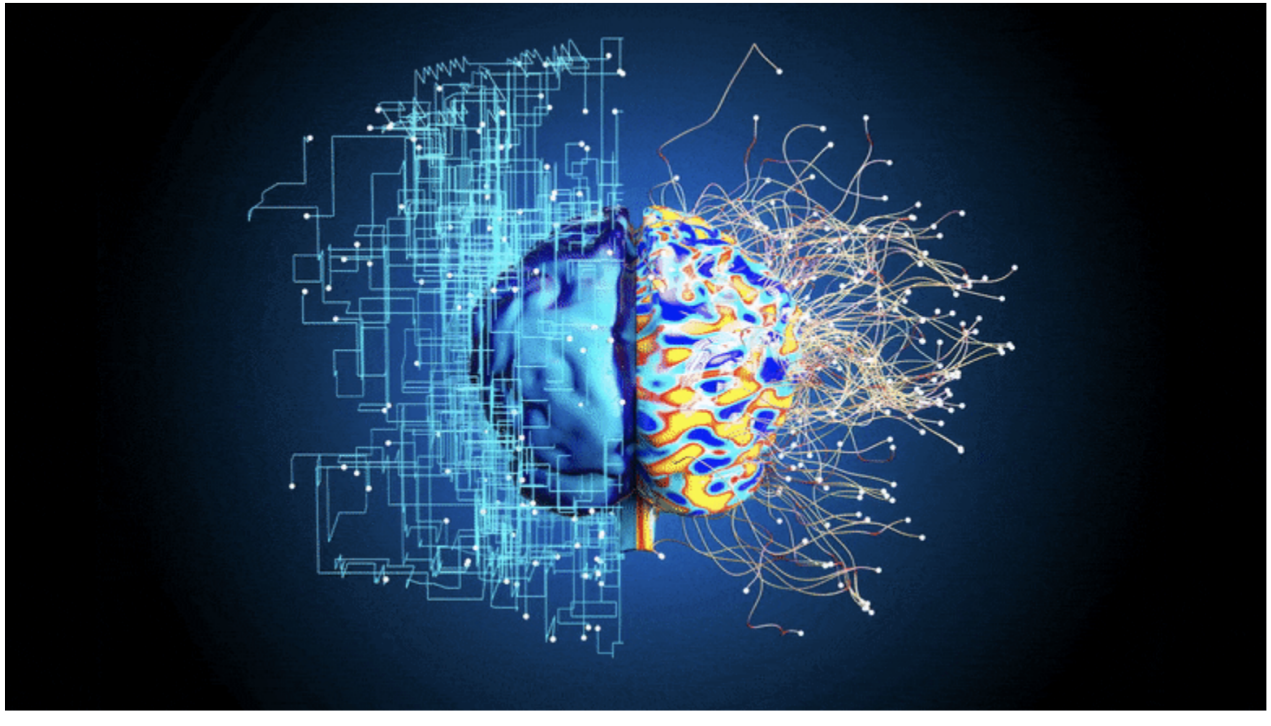
Artificial intelligence can tell if you've got heart problems simply by the sound of your voice

MARCH 24, 2022

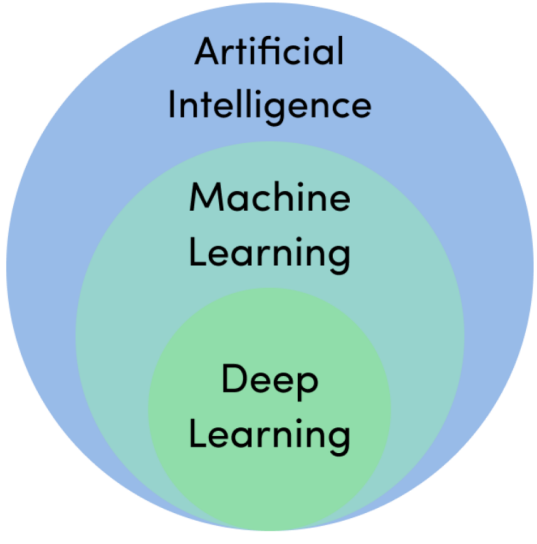
Hidden Signatures of Parkinson's Disease Uncovered by Artificial Intelligence and Robotics

TOPICS: Artificial Intelligence Machine Learning Parkinson's Disease Robotics

By NEW YORK STEM CELL FOUNDATION MARCH 25, 2022

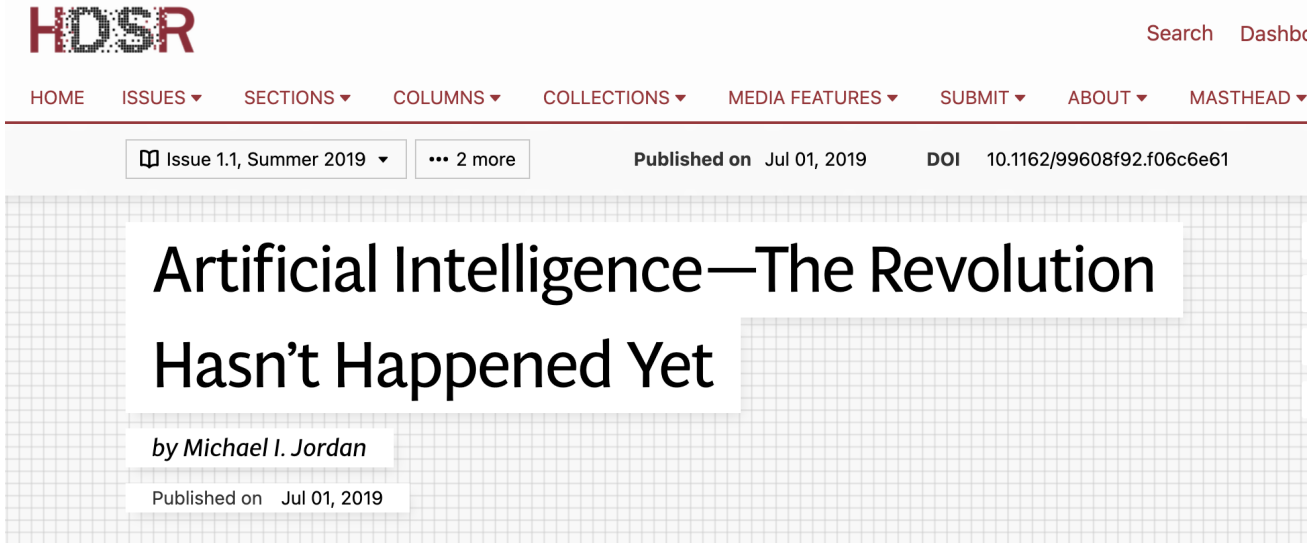


AI-Machine Learning-Deep Learning



Source: https://levity.ai/blog/difference-machine-learning-deep-learning

What is Michael Jordan's opinion?

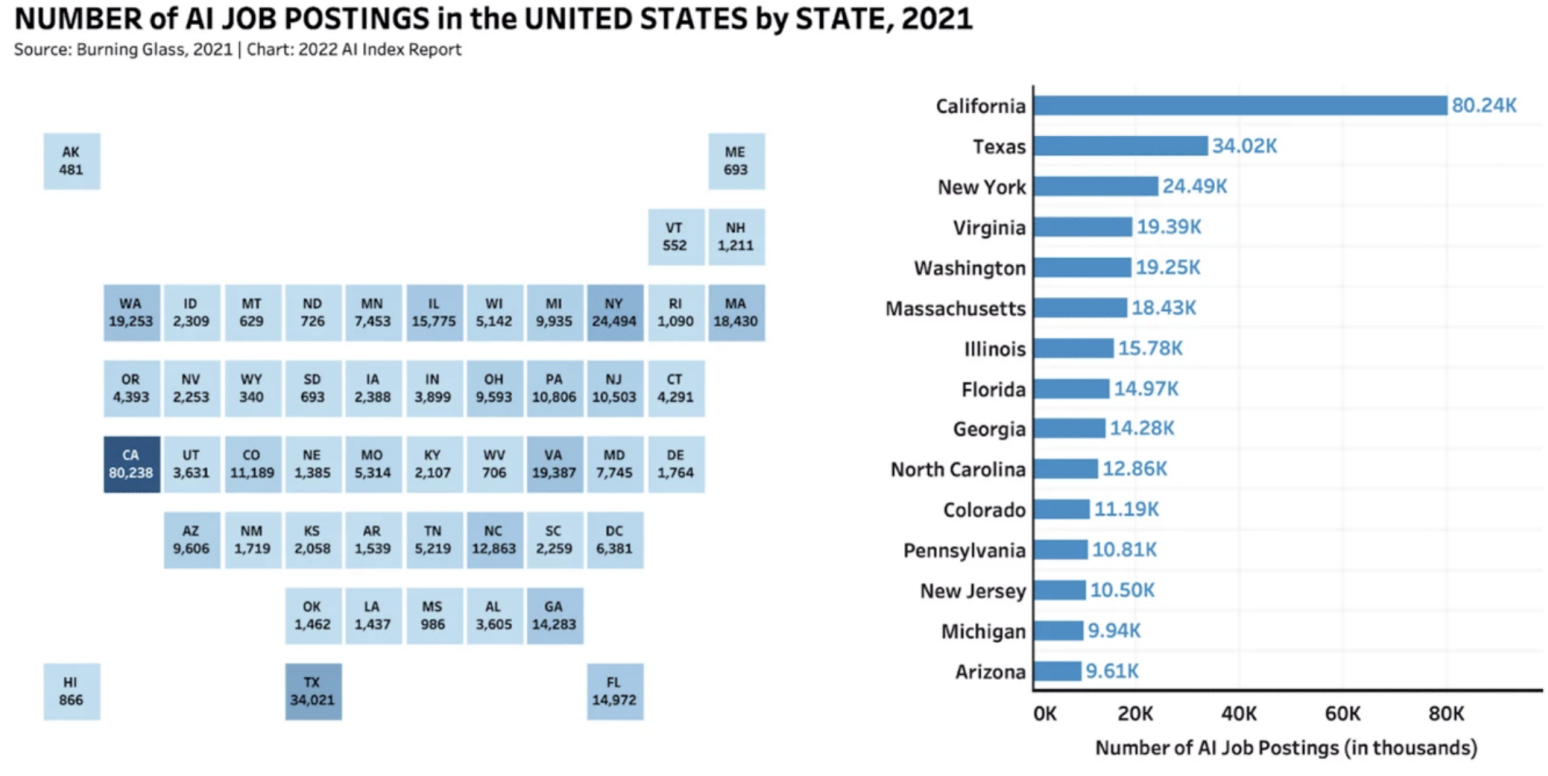


Source: https://hdsr.mitpress.mit.edu/pub/wot7mkc1/release/9

"We should embrace the fact that we are witnessing the creation of a new branch of engineering. "

"In the current era, we have a real opportunity to conceive of something historically new: a human-centric engineering discipline."

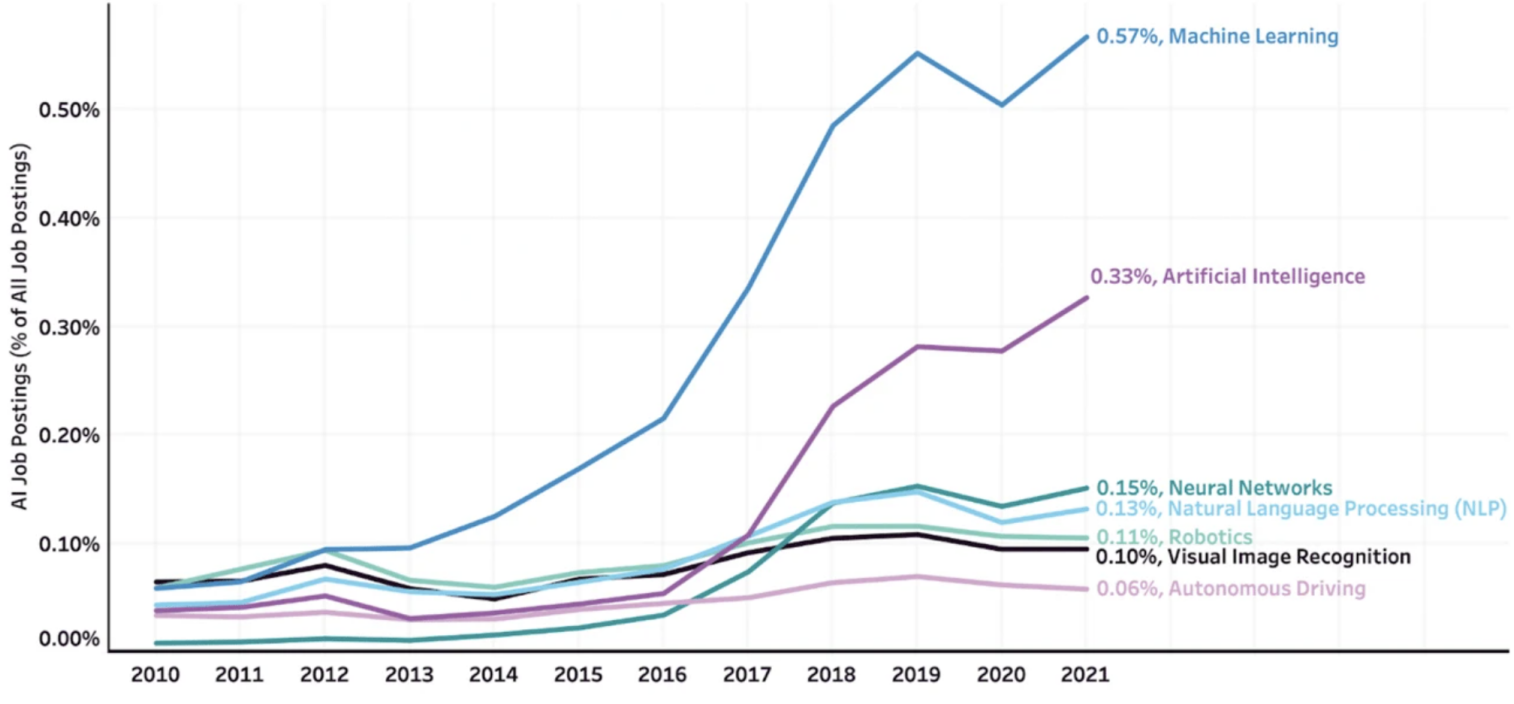
From the 2022 Stanford AI Index Report

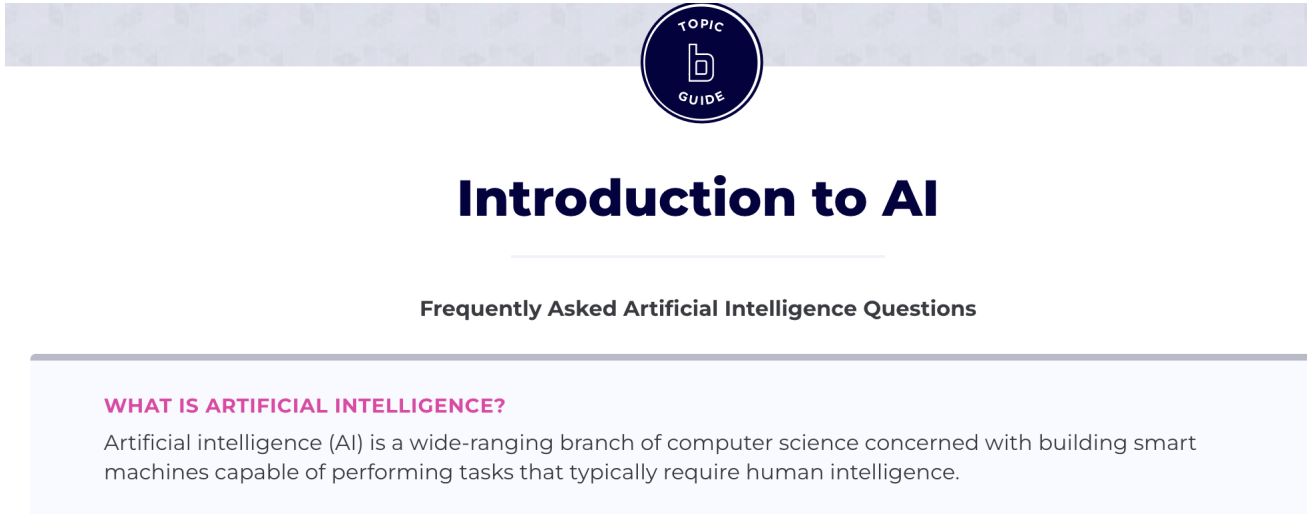


Source: https://www.economicmodeling.com/2022/03/16/where-are-the-artificial-intelligence-jobs/?fbclid=IwAR3-WY-PyuuvSHn3uuu3FSYsxGcza2BLvJ_eZ05c7wwhFg5b7--jmfrhK7g

AI JOB POSTINGS (% of ALL JOB POSTINGS) in the UNITED STATES by SKILL CLUSTER, 2010-21

Source: Burning Glass, 2021 | Chart: 2022 AI Index Report





Source: https://builtin.com/artificial-intelligence

Back to ML and Statistics

1 Statistics versus ML

Statistics and ML are overlapping fields. Both address the same question: how do we extract information from data? But there are differences in emphasis. In particular, some topics get greater emphasis than others. Here are some examples:

More emphasis in ML	More emphasis in Stat	Common Areas
Bandits	Confidence Sets	Prediction (Regression and Classification)
Reinforcement Learning	Large Sample Theory	Probability Bounds (Concentration)
Efficient Computation	Statistical Optimality	Clustering
Deep Learning	Causality	Graphical Models

However, the lines between the two fields are blurry and will become increasingly so.

Source: https://www.stat.cmu.edu/~larry/=sml/Review.pdf

What is this course NOT about?

This course is not about:

- teaching you how to write code in Python, R, etc.
- teaching you how to fit packages
- learning as many methods as possible
    - Machine learning is a field, it is impossible to cover everything in a 10-week course
    - Methods can change and adapt to new tasks
- theory and large sample properties
- (For fun!) definitely not about teaching you how to create an human intelligence-like creature

What is this course about?

- Inference at large scale
- Principles - although we don't study theory, some important results will be mentioned. They will have important application implications
- Ideally: Ability to replicate an exisiting method
- Statistical machine learning methods

- Topic 1: Linear methods for regression (3 weeks)
    - Linear regression, bias-variance trade-off, subset selection, ridge regression, lasso, lasso in high-dimension, variations of lasso methods
- Topic 2: Classification (2 weeks)
    - Linear method for classification, logistic regression, unsupervised learning, support vector machine, multiclass classification, K-means, nearest neighbor classifiers, mixture models, PCA, factor analysis, PCA in high-dimension
- Topic 3: Nonparametric methods (2 weeks)
    - Kernel density estimation, tree-based method, basis expansions
- Topic 4: Bayesian methods (2 weeks)
    - Intro to Bayesian methods, Bayesian nonparametrics - Gaussian processes, Bayesian high-dimensional analysis - sparse priors
- Topic 5: Other topics if time permits
    - Causal inference, graphical models, neural networks, variational inference

Syllabus

Python

The course is based on python. We will mainly use **numpy**, which is the fundamental package for scientific computing with Python. **scikit-learn** is another useful package we will use frequently.

Install Python and Jupyter Notebook

- Download python from the webiste: https://www.python.org/downloads/
  We use Python 3 for this class, do not download Python 2

- Install Python notebook
    1. intall conda https://conda.io/projects/conda/en/latest/user-guide/install/index.html
    2. go to **Terminal**
    3. Type conda install -c conda-forge notebook
    4. In Terminal, type **jupyter notebook** to open the notebook
    5. use pip

Python basics

check out the python document: https://docs.python.org/3/
GOOGLE is always your best friend!

In [5]:
```
print("Hello, World!")
```

```
Hello, World!
```

In [1]:
```
print('5')
```

```
5
```

In [3]:
```
print(5.0)
```

```
5.0
```

In [4]:
```
# <-- this starts a comment
# print a number
5
```

Out[4]:
```
5
```

In [6]:
```
# calculate 2 + 3
2+3
```

Out[6]:
```
5
```

In [12]:
```
import numpy as np
```

In [ ]:
```
np.
```

In [11]:
```
from FUNCTION NAME import numpy
```

Out[11]:
```
array([0., 0., 0., 0., 0.])
```

In [18]:
```
# use package
np.random.seed(2022)
x = np.random.normal(loc=0.0, scale=1.0, size = 5)
print(x)
```

```
[-5.27899086e-04 -2.74901425e-01 -1.39285562e-01  1.98468616e+00
  2.82109326e-01]
```

In [25]:
```
y = np.random.normal(loc=0.0, scale=1.0, size = 5)
y
```

Out[25]:
```
array([-0.09021319, -2.30594327,  1.14276002, -1.53565429, -0.86375202])
```

In [26]:
```
x + y
```

Out[26]:
```
array([-0.09074109, -2.58084469,  1.00347446,  0.44903187, -0.58164269])
```

In [20]:
```
x[1]
```

Out[20]:
```
-0.27490142489105457
```

In [ ]:

In [15]:

```
# for loop
# k = 0
# for i in range(1,100):
#     k = k + i
#     print(k) # indentation matters in Python!

k = 0
for i in range(1,100):
    k = k + i

print(k)
```

```
4950
```

In [ ]:

```
# range(a, b) means start with a and end with b-1
# range(b) means start with 0 and end with b-1
```

In [22]:

```
# addition, Subtraction, Multiplication, Division, Modulus, Exponentiation, Floor division
print(3 + 5)
print(3 - 2)
print(3 * 2)
print(3 / 2)
print(3 % 2)
print(3 ** 2)
print(3 // 2)
```

```
8
1
6
1.5
1
9
1
```

In [17]:
```
range(100) # python starts at 0, this is different from R!
```

Out[17]:
```
range(0, 100)
```

In [18]:
```
range(2, 100)
```

Out[18]:
```
range(2, 100)
```