UMBC

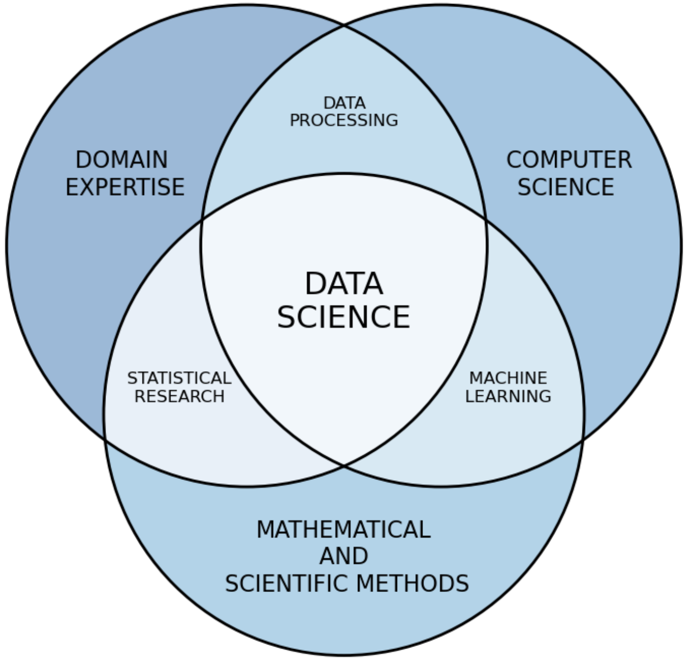
Why & How to Teach Yourself Data Science

Ergun Simsek

Assistant Professor of Computer Science and Electrical Engineering

Director of Graduate Data Science Programs

What is Data Science?



… knowledge discovery from often large and complex data sets

… interdisciplinary by nature, encompassing statistics, computer science, applied mathematics, and domain-specific tools

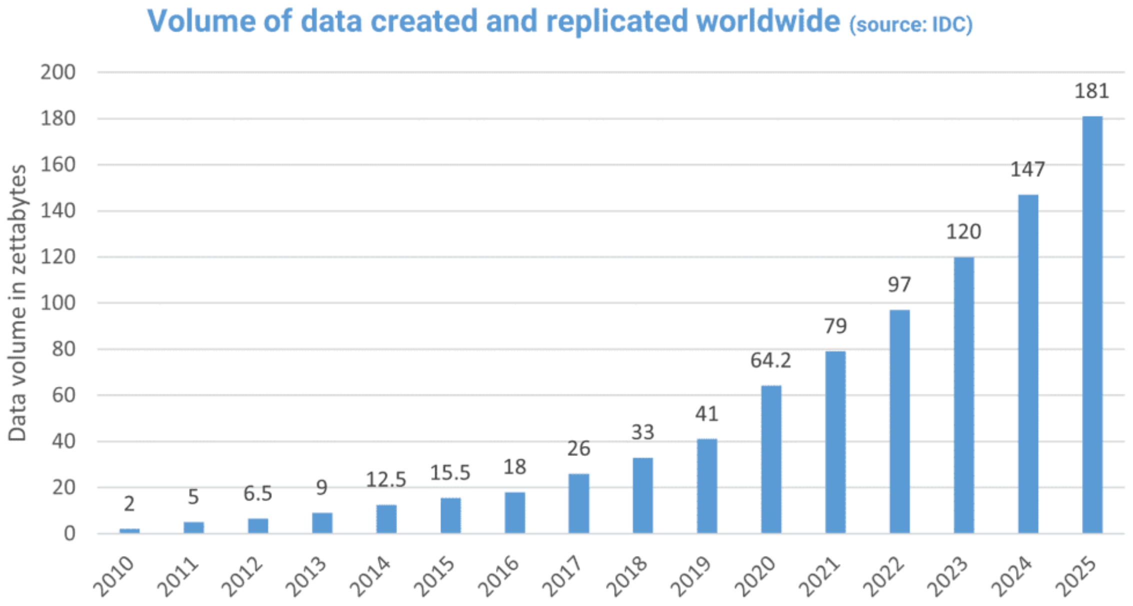
Data Science

- **Focus:** Broadly focused on understanding and interpreting data to gain insights and make informed decisions.
- **Techniques:** Employs various techniques, including statistical analysis, data visualization, and machine learning.
- **Goal:** To uncover patterns, trends, and relationships within data to solve problems and make predictions.
- **Skills:** Requires strong analytical, programming, and communication skills.

Machine Learning

- **Focus:** Specifically concerned with developing algorithms that enable computers to learn from data without explicit programming.
- **Techniques:** Employs algorithms like regression, classification, clustering, and neural networks.
- **Goal:** To build models that can make predictions or decisions based on learned patterns from data.
- **Skills:** Requires expertise in programming, statistics, and mathematics, as well as knowledge of specific machine learning algorithms and frameworks.

Why did “Data Science” grow so fast recently?



Volume of data created and replicated worldwide (source: IDC)

5MB – $50,000

1 TB - $70

python™

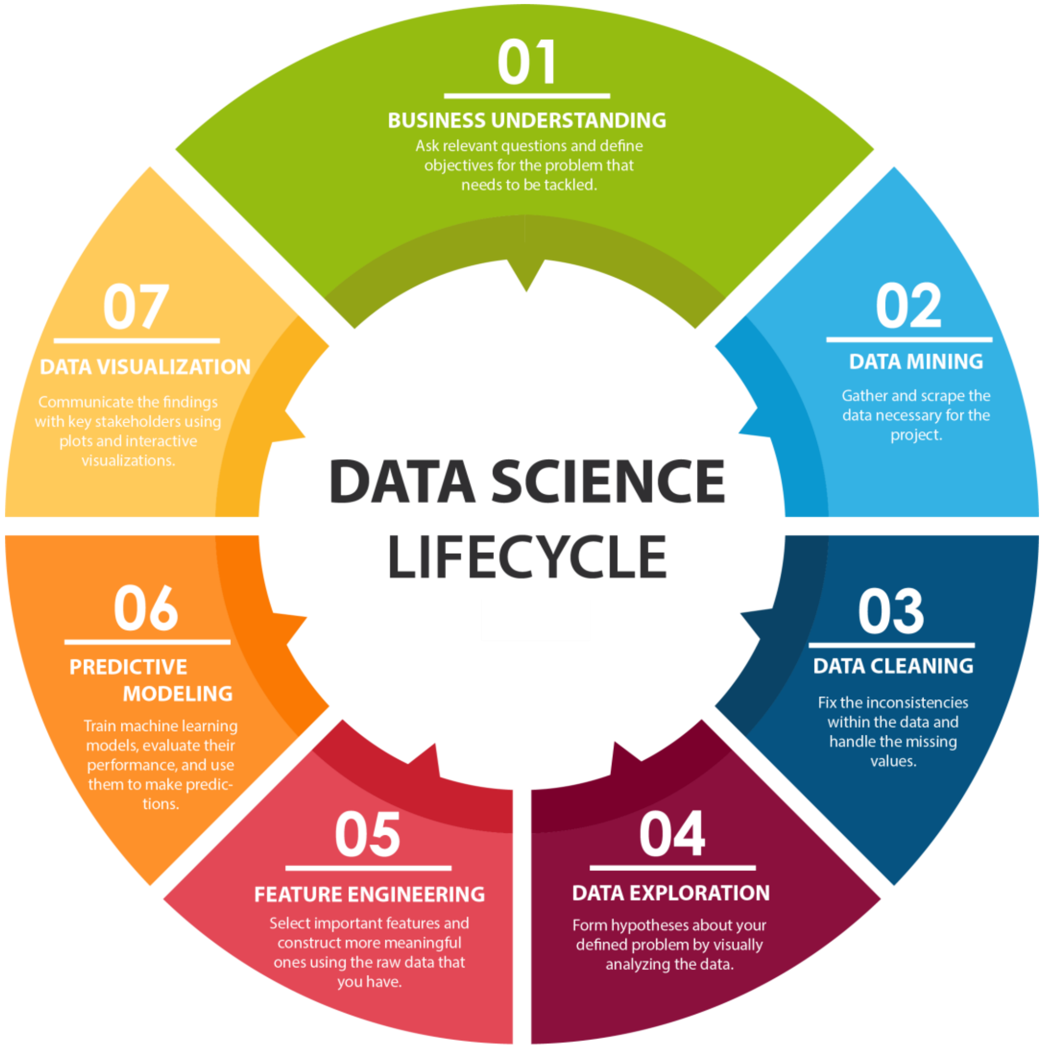
CPU GPU TPU

Where do companies use DS?

Case Study

- Personalized recommendations
- Targeted advertising
- Demand forecasting
- Warehouse and route optimization
- Dynamic pricing
- Fraud and fake review detection
- Sentiment analysis
- Computer vision (Amazon Go)
- Voice recognition (Alexa)
- Robot coordination
- ….

Data Science Lifecycle



DATA SCIENCE LIFECYCLE

01

BUSINESS UNDERSTANDING

Ask relevant questions and define objectives for the problem that needs to be tackled.

02

DATA MINING

Gather and scrape the data necessary for the project.

03

DATA CLEANING

Fix the inconsistencies within the data and handle the missing values.

04

DATA EXPLORATION

Form hypotheses about your defined problem by visually analyzing the data.

05

FEATURE ENGINEERING

Select important features and construct more meaningful ones using the raw data that you have.

06

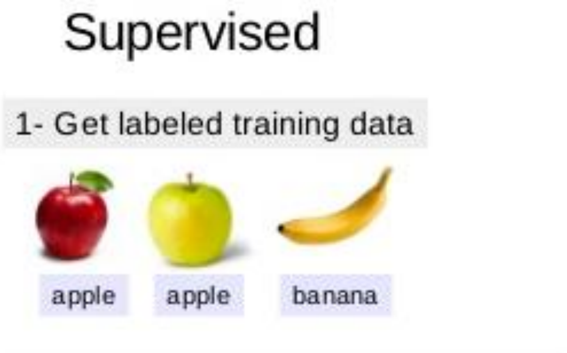
PREDICTIVE MODELING

Train machine learning models, evaluate their performance, and use them to make predictions.

07

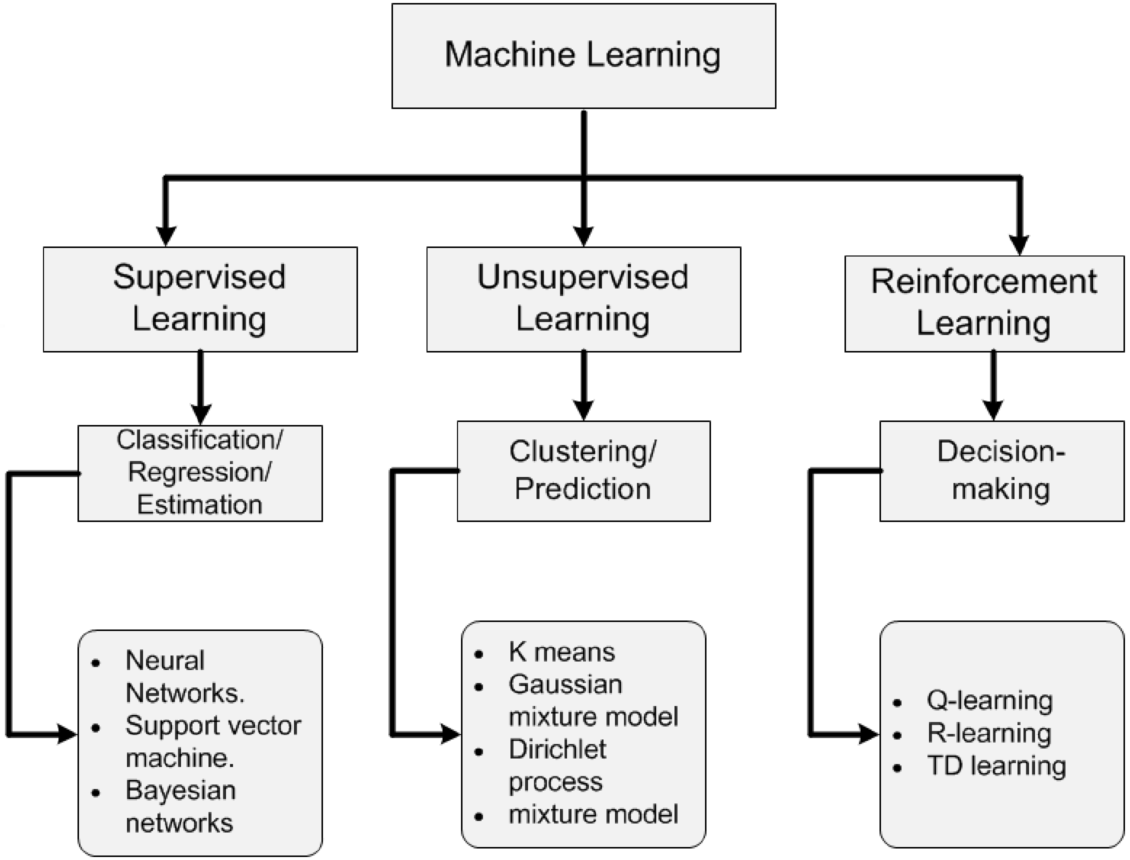
DATA VISUALIZATION

Communicate the findings with key stakeholders using plots and interactive visualizations.



Supervised

1- Get labeled training data



Machine Learning

Supervised Learning

Unsupervised Learning

Reinforcement Learning

Classification/ Regression/ Estimation

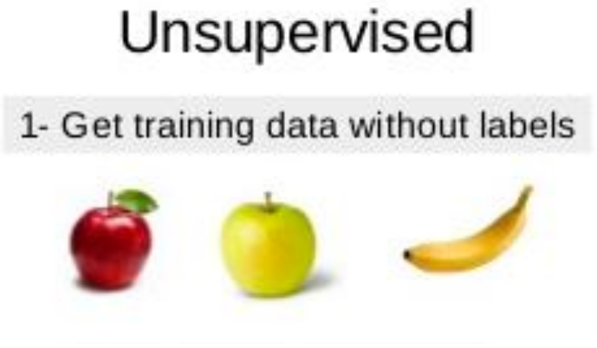
Clustering/ Prediction

Decision-making

- Neural Networks.
- Support vector machine.
- Bayesian networks

- K means
- Gaussian mixture model
- Dirichlet process
- mixture model

- Q-learning
- R-learning
- TD learning



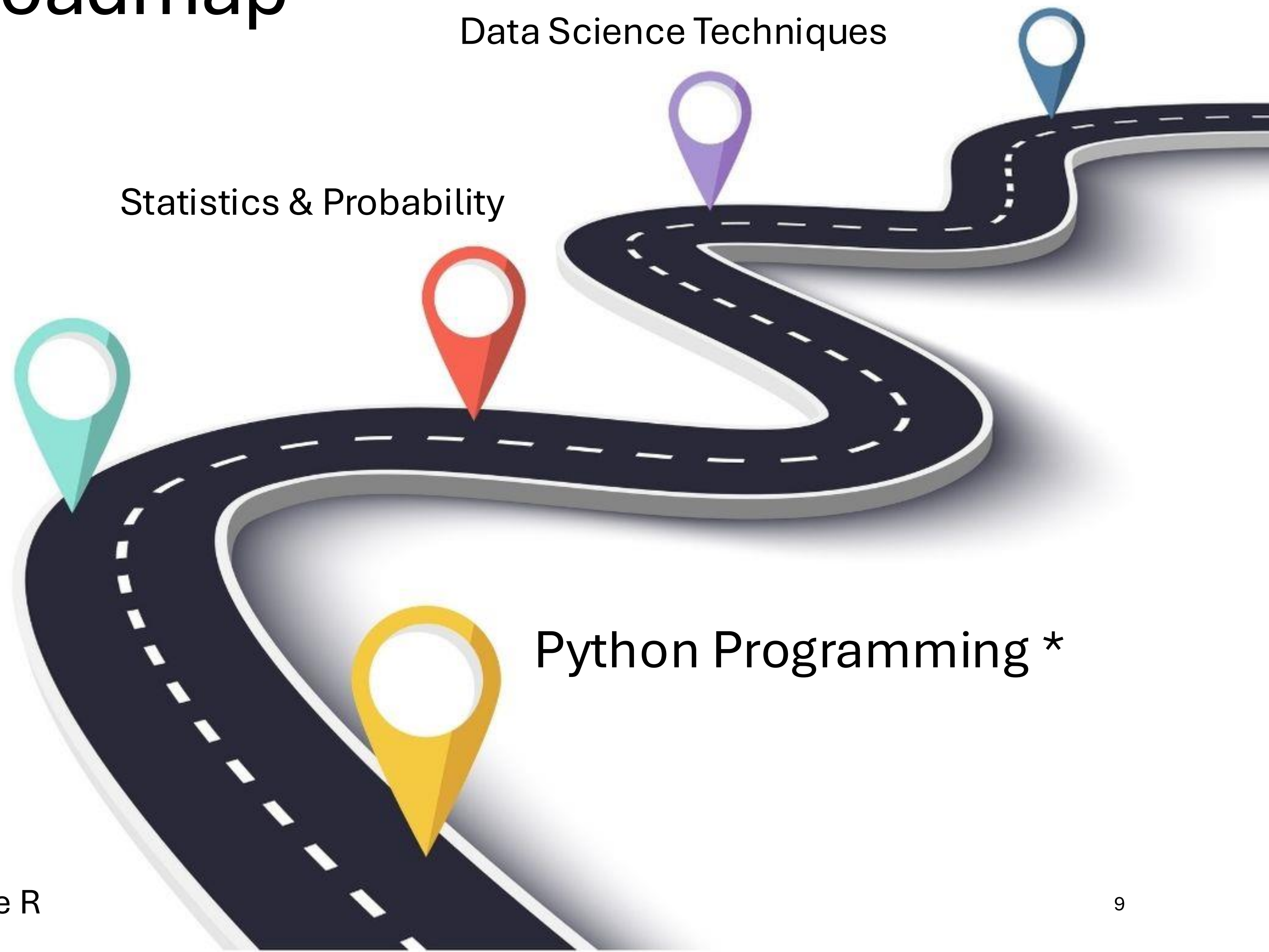
Unsupervised

1- Get training data without labels

How To Self-Learn Data Science

With all the Best Free Resources

Data Science Roadmap



Production

Data Science Techniques

Statistics & Probability

Database Management

Python Programming *

* Healthcare organizations mostly use R

Python Programming

Learning Python

- Free online course (for absolute beginners):
  - https://www.coursera.org/specializations/python

- YouTube tutorials
  - Corey Schafer (well organized)
  - Sentdex (you need to choose what you want to learn)

- Python Documentation
  - https://docs.python.org/3/tutorial/index.html

Python Platforms

- Google Colab: https://colab.research.google.com/
  - Free for anyone with a Gmail account
  - No installation is necessary
  - No need for a fancy computer
  - You can’t run more than 2-3 notebooks simultaneously

- Anaconda: https://www.anaconda.com/download
  - Free
  - A distribution of the Python programming language for scientific computing that aims to simplify package management and deployment

Recommended Path

- Get familiar with Python platforms and the concept of a notebook
- Learn about data types (collections)
- Study conditionals and loops
- Practice how to create and call functions
- Try working with different inputs/outputs and files
- Compute with NumPy
- Visualize with Matplotlib
- Run exploratory data analysis

https://userpages.cs.umbc.edu/simsek/PW/

Test Your Python Programming

- Kaggle: https://www.kaggle.com/learn/python
- Offers Python tutorials and datasets for practicing data science and machine learning.

Note: Kaggle provides tutorials on all aspects of data science (programming, databases, machine learning, etc.)

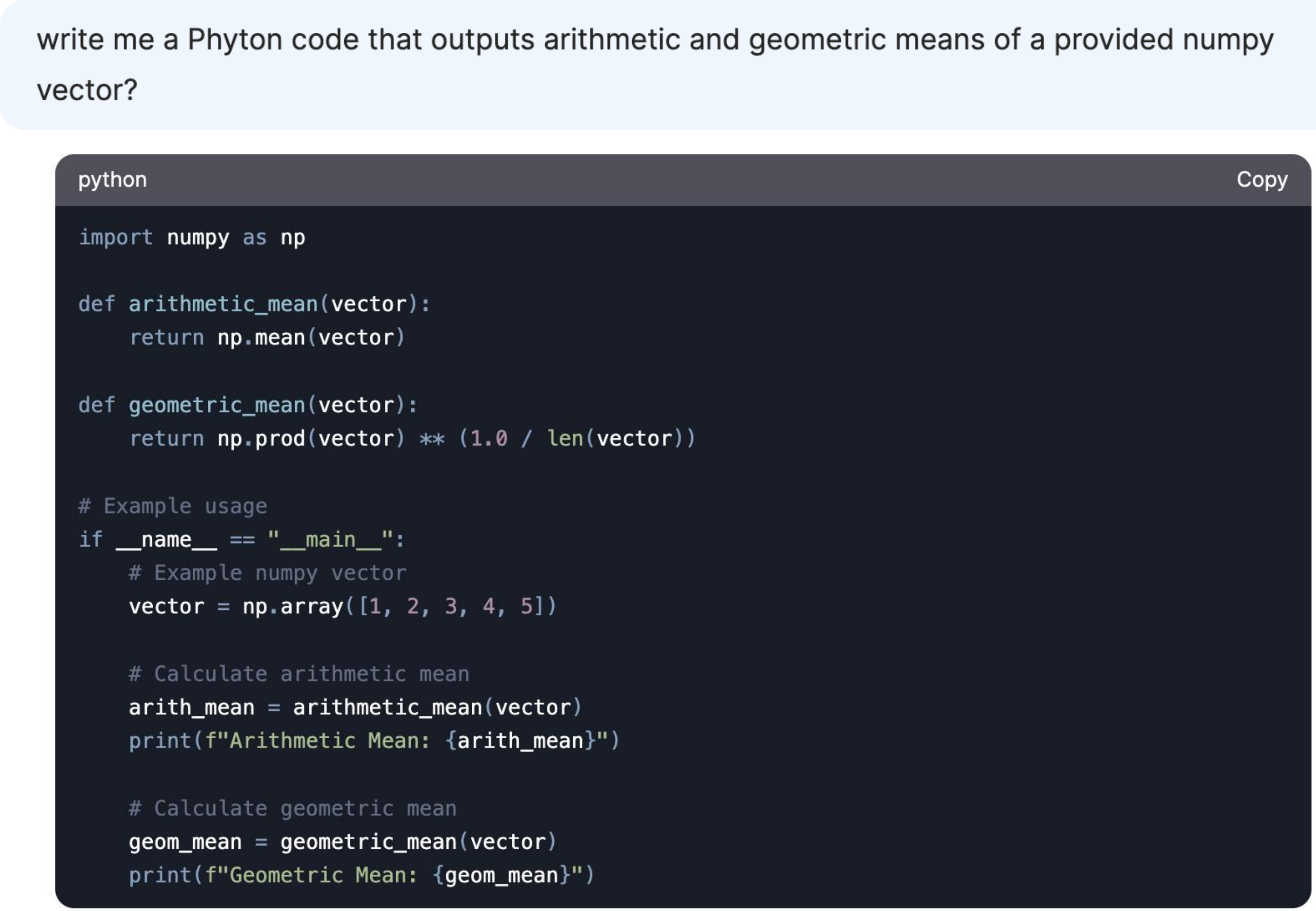
https://www.kaggle.com/learn

Get help from AI

AI platforms can write code for you and help with debugging your codes.

- ChatGPT: https://chatgpt.com/
- DeepSeek: https://www.deepseek.com/
- Gemini: https://gemini.google.com/app

Example



write me a Phyton code that outputs arithmetic and geometric means of a provided numpy vector?

```python
import numpy as np

def arithmetic_mean(vector):
    return np.mean(vector)

def geometric_mean(vector):
    return np.prod(vector) ** (1.0 / len(vector))

# Example usage
if __name__ == "__main__":
    # Example numpy vector
    vector = np.array([1, 2, 3, 4, 5])

    # Calculate arithmetic mean
    arith_mean = arithmetic_mean(vector)
    print(f"Arithmetic Mean: {arith_mean}")

    # Calculate geometric mean
    geom_mean = geometric_mean(vector)
    print(f"Geometric Mean: {geom_mean}")
```

Database Management

Get Familiar with SQL

- IBM’s free course on SQL: https://www.coursera.org/learn/sql-practical-introduction-for-querying-databases

- YouTube tutorials: https://www.youtube.com/playlist?list=PLavw5C92dz9Ef4E-1Zi9KfCTXS_IN8gXZ

- Get help from AI: https://www.sqlchat.ai/

Statistics and Probability

Learn Statistics & Probability

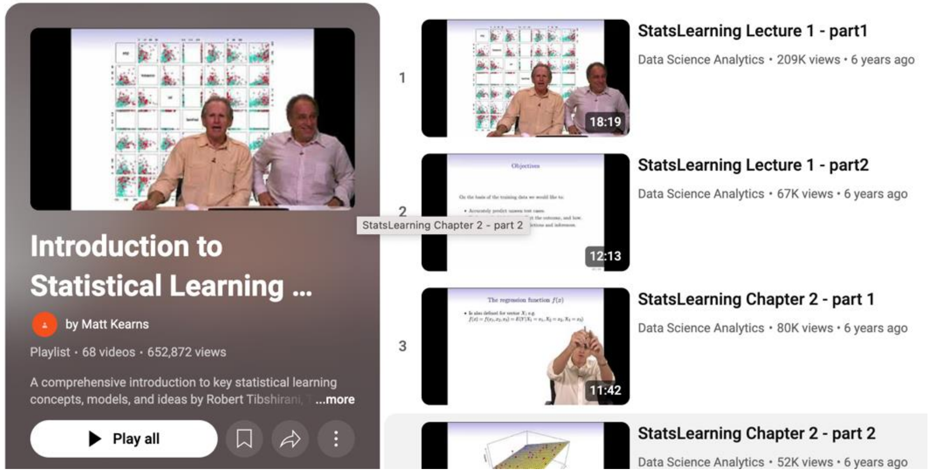
- Best online course for beginners:
  - Khan Academy: https://www.khanacademy.org/math/statistics-probability

- Short YouTube tutorials for beginners:
  - StatQuest: https://www.youtube.com/watch?v=qBigTkBLU6g&list=PLblh5JKOoLUK0FLuzwntyYI10UQFUhsY9

Merging Statistics with Data Science

Ultimate Source: ISL: https://www.statlearning.com/

- You can download the entire book for free
- You can obtain the Python codes: https://github.com/JWarmenhoven/ISLR-python
- You can watch the lectures on YouTube: https://www.youtube.com/playlist?list=PLOg0ngHtcqbPTlZzRHA2ocQZqB1D_qZ5V
  - Not only super educative but also tremendously funny



Introduction to Statistical Learning ...

by Matt Kearns

Playlist • 68 videos • 652,872 views

A comprehensive introduction to key statistical learning concepts, models, and ideas by Robert Tibshirani, T ...more

1 StatsLearning Lecture 1 - part1
Data Science Analytics • 209K views • 6 years ago

2 StatsLearning Lecture 1 - part2
Data Science Analytics • 67K views • 6 years ago

3 StatsLearning Chapter 2 - part 1
Data Science Analytics • 80K views • 6 years ago

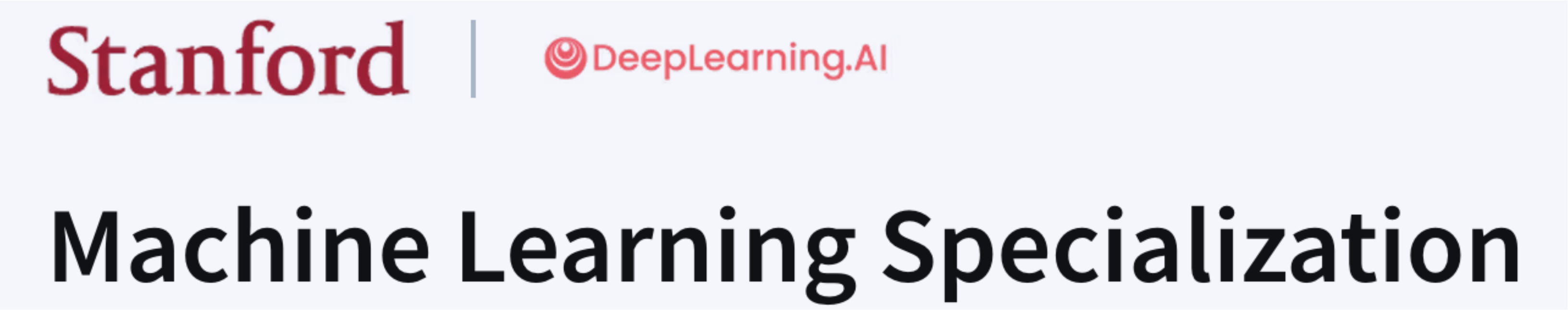
StatsLearning Chapter 2 - part 2
Data Science Analytics • 52K views • 6 years ago

Play all
18:19
12:13
11:42
StatsLearning Chapter 2 - part 2
Objectives
The regression function $f(x)$

Data Science Techniques

Learn Machine Learning (ML) Methods

- Andrew NG's famous free course: https://www.coursera.org/specializations/machine-learning-introduction
  - Supervised
  - Unsupervised
  - Advanced


Stanford | DeepLearning.AI

Machine Learning Specialization

Recommended Path for advancing from Data Analysis to Machine Learning

- **Linear Regression** – Predicting a continuous output using a linear relationship.
- **Logistic Regression** – Classification method for binary/multiclass outcomes.
- **K-Nearest Neighbors (kNN)** – A simple, instance-based classification/regression method.
- **Decision Trees** – Basic tree-based learning method for classification/regression.
- **Naïve Bayes** – Probabilistic classifier using Bayes' theorem.
- **Support Vector Machines (SVM)** – Finds an optimal hyperplane for classification tasks.
- **Principal Component Analysis (PCA)** – Dimensionality reduction method.
- **Random Forest** – Ensemble method using multiple decision trees.
- **Gradient Boosting Machines (GBM)** – A boosting approach to improve tree-based models.
- **XGBoost / LightGBM / CatBoost** – Optimized gradient boosting frameworks.
- **Deep Learning**
  - Neural networks: Convolutional, Recurrent, Graph
  - Transformers
  - Reinforcement Learning
  - ….

Where to find datasets to practice DS/ML?

- UCI ML Repository: https://archive.ics.uci.edu/

- Kaggle Datasets: https://www.kaggle.com/datasets

- Google: https://datasetsearch.research.google.com/

- Government: https://data.gov/
- World Bank: https://data.worldbank.org/
- United Nations: https://data.un.org/
- NOAA: https://www.ncdc.noaa.gov/cdo-web/datasets
- NASA: https://data.nasa.gov/
- Zillow: https://www.zillow.com/research/data/

Production

How to showcase your data science skills?

- GitHub: https://github.com/
  - A free online platform where you can store your codes in "repositories" on GitHub, which can be public or private.
  - Basic usage can be learned in 10 minutes
    - https://www.youtube.com/watch?v=iv8rSLsi1xo
  - Advanced use of git and GitHub
    - https://www.youtube.com/watch?v=RGOj5yH7evk
  - Create an attractive profile
    - https://x-team.com/magazine/stand-out-with-a-github-profile

Thank you very much for your attention.

I will be more than happy to answer your questions.