

GitHub curso



WTISC2023

WORKSHOP DE TECNOLOGIA DA INFORMAÇÃO DO SERTÃO CENTRAL | UFC - CAMPUS QUIXADÁ

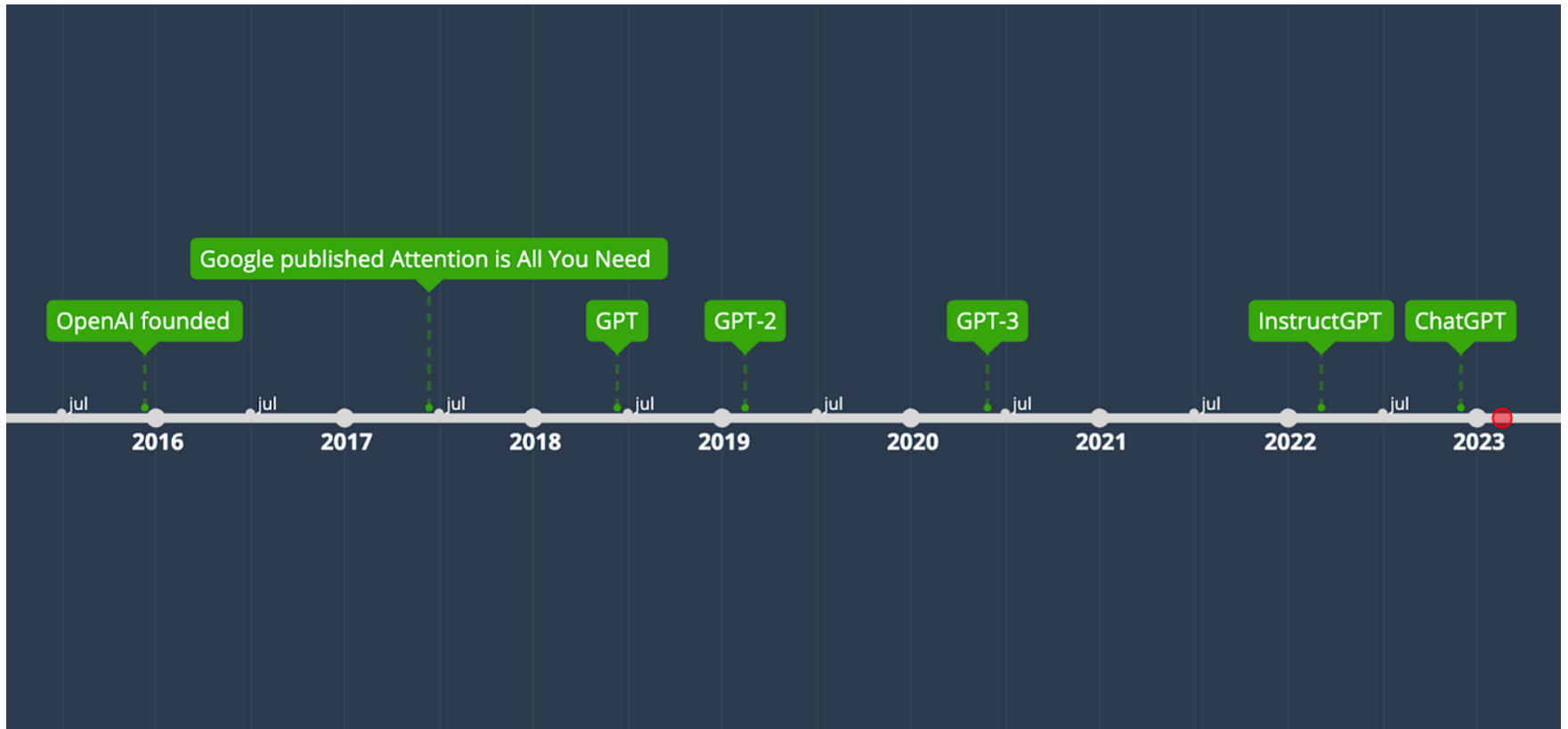
17 A 19 DE MAIO

Introdução à ciência de dados e
aprendizado de máquina usando Python

Regis Pires Magalhães
regis@insightlab.ufc.br



ChatGPT Timeline



Fonte: <https://bootcamp.uxdesign.cc/how-chatgpt-really-works-explained-for-non-technical-people-71efb078a5c9>

O que GPT-4 pensa sobre o futuro do mercado de trabalho?

DidáticaTech: *Nomeie 20 profissões que modelos multimodais como GPT-4 podem substituir no futuro próximo. Faça também uma previsão do percentual de vagas que serão desocupadas em cada profissão pensando em um horizonte de 5 anos. Informe resumidamente as principais características da profissão que permitiram que os modelos de IA tomassem seu papel. Mostre tudo isso em uma tabela com as seguintes colunas: Profissão, % de vagas, Características. Ordene do maior percentual ao menor.*

GPT-4: Aqui está uma tabela de 20 profissões que modelos multimodais como GPT-4 podem substituir no futuro próximo, juntamente com a previsão do percentual de vagas desocupadas e as características que permitiram a IA assumir seu papel:

<https://didatica.tech/o-que-gpt-4-pensa-sobre-o-futuro-do-mercado-de-trabalho/>

Profissão	% de vagas	Características
Atendente de telemarketing	75%	Tarefas repetitivas, script de conversação, resolução de problemas simples
Tradutor	60%	Tradução de idiomas, análise contextual, ajuste a diferentes estilos de escrita
Jornalista	40%	Pesquisa, redação, edição, criação de conteúdo
Assistente administrativo	50%	Tarefas rotineiras, organização, gerenciamento de documentos, comunicação escrita e verbal
Operador de chat	70%	Respostas baseadas em scripts, resolução de problemas, comunicação escrita
Operador de caixa	60%	Transações financeiras, atendimento ao cliente, tarefas repetitivas
Analista de crédito	50%	Avaliação de risco, análise de dados, tomada de decisão baseada em critérios
Escritor de conteúdo	40%	Criação de conteúdo, redação, pesquisa, otimização de palavras-chave
Analista financeiro	35%	Análise de dados, projeções, modelagem financeira, relatórios
Agente de viagens	50%	Pesquisa, planejamento, reservas, atendimento ao cliente
Bibliotecário	35%	Organização, pesquisa, catalogação, gerenciamento de informações
Redator publicitário	30%	Criação de conteúdo, persuasão, marketing, análise de público-alvo
Secretário	50%	Tarefas administrativas, agendamento, comunicação escrita e verbal
Suporte técnico	40%	Resolução de problemas, atendimento ao cliente, conhecimento técnico
Contador	30%	Processamento de dados, cálculos, análise financeira, conformidade tributária
Revisor de texto	40%	Gramática, ortografia, estilo, consistência, edição
Pesquisador de mercado	35%	Coleta de dados, análise, pesquisa, relatórios, identificação de tendências
Planejador de eventos	25%	Organização, pesquisa, logística, comunicação
Assistente pessoal	40%	Tarefas rotineiras, organização, agendamento, pesquisa
Recrutador	30%	Seleção de candidatos, análise de currículos, entrevistas, gerenciamento de processos seletivos

O que GPT-4 pensa sobre o futuro do mercado de trabalho?

DidáticaTech: *Agora crie um gráfico parecido, seguindo a mesma lógica, mas em vez de mencionar vagas que serão substituídas, fale sobre profissões que terão um aumento no número de vagas. Mantenha as mesmas colunas, onde na coluna "Características" você vai mencionar as características que esses profissionais deverão ter para assumir tais vagas.*

GPT-4: Aqui está uma tabela de profissões que podem experimentar um aumento no número de vagas devido ao avanço da IA e as características que esses profissionais devem possuir:

<https://didatica.tech/o-que-gpt-4-pensa-sobre-o-futuro-do-mercado-de-trabalho/>

Profissão	% de aumento	Características
Engenheiro de Machine Learning	60%	Conhecimento em algoritmos de aprendizado, estatística, programação, análise e manipulação de dados
Cientista de dados	50%	Análise de dados, estatística, visualização de dados, programação, machine learning
Especialista em segurança cibernética	40%	Conhecimento em redes, criptografia, detecção de intrusão, legislação, ética
Desenvolvedor de IA	45%	Conhecimento em algoritmos, aprendizado de máquina, programação, sistemas de IA
Designer de experiência do usuário (UX)	35%	Design, psicologia, pesquisa, prototipagem, testes, análise de dados, empatia
Gerente de projetos de IA	30%	Gerenciamento de projetos, liderança, conhecimento em IA, comunicação, planejamento
Analista de ética em IA	25%	Conhecimento em IA, ética, filosofia, política, comunicação, legislação
Treinador de IA	20%	Conhecimento em IA, psicologia, pedagogia, comunicação, empatia, técnicas de treinamento
Especialista em conformidade de IA	20%	Conhecimento em IA, legislação, ética, política, conformidade, comunicação
Consultor de transformação digital	30%	Conhecimento em tecnologia, estratégia, inovação, mudança organizacional, liderança

Machine learning \subseteq artificial intelligence

ARTIFICIAL INTELLIGENCE

Design an intelligent agent that perceives its environment and makes decisions to maximize chances of achieving its goal.
Subfields: vision, robotics, machine learning, natural language processing, planning, ...

MACHINE LEARNING

Gives "computers the ability to learn without being explicitly programmed" (Arthur Samuel, 1959)

SUPERVISED LEARNING

Classification, regression

UNSUPERVISED LEARNING

Clustering, dimensionality
reduction, recommendation

REINFORCEMENT LEARNING

Reward maximization

Types of machine learning

Supervised Learning

- > Labeled data
- > Direct feedback
- > Predict outcome/future

Unsupervised Learning

- > No labels/targets
- > No feedback
- > Find hidden structure in data

Reinforcement Learning

- > Decision process
- > Reward system
- > Learn series of actions

Machine Learning

Supervised Learning

Classification

- Naive Bayes Classifier
- Decision Trees
- Support Vector Machines
- Random Forest
- K – Nearest Neighbors

Regression

- Linear Regression
- Neural Network Regression
- Support Vector Regression
- Decision Tree Regression
- Lasso Regression
- Ridge Regression

Unsupervised Learning

Clustering

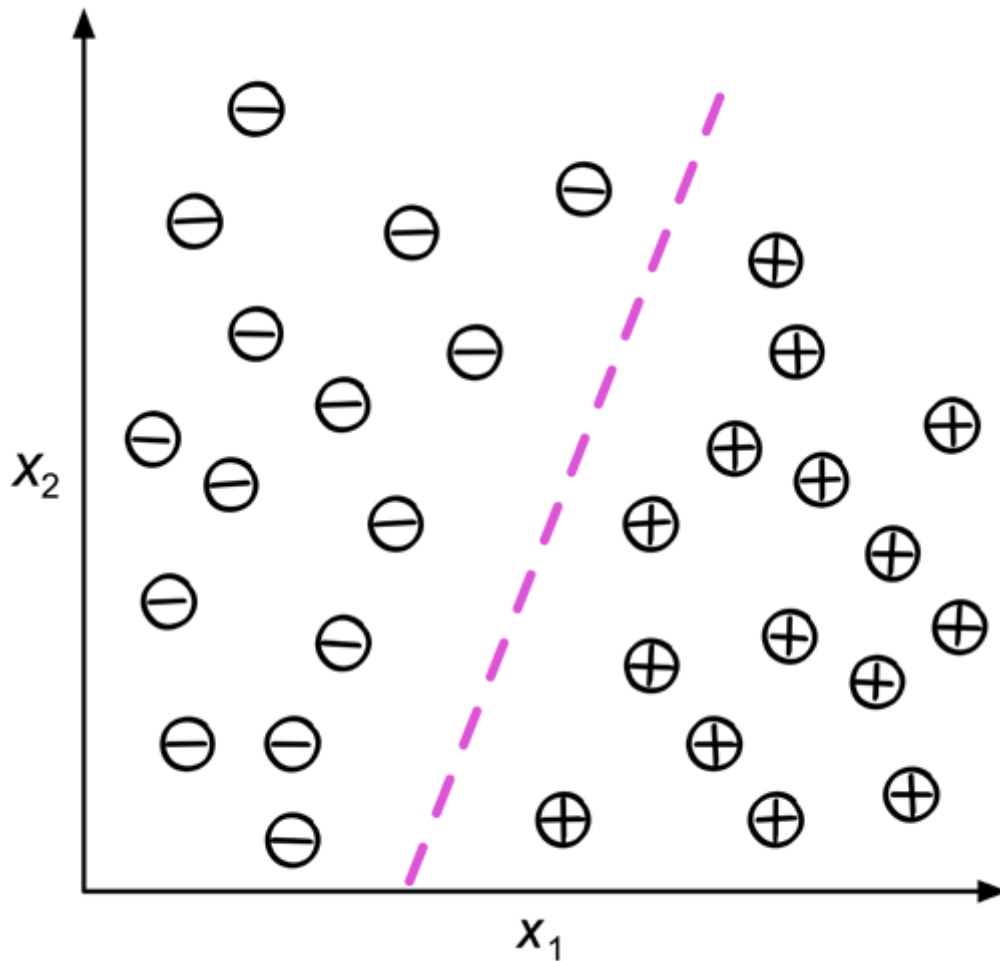
- K-Means Clustering
- Mean-shift Clustering
- DBSCAN Clustering
- Agglomerative Hierarchical Clustering
- Gaussian Mixture

Reinforcement Learning

Decision Making





- Q-Learning
- R Learning
- TD Learning

Binary classification



Supervised Learning: Classification

training set

Observation #	Input image (X)	Label (Y)
1		"dog"
2		"cat"
3		"dog"
...
N		"dog"

test set

1		???
2		???

Supervised Learning: Regression

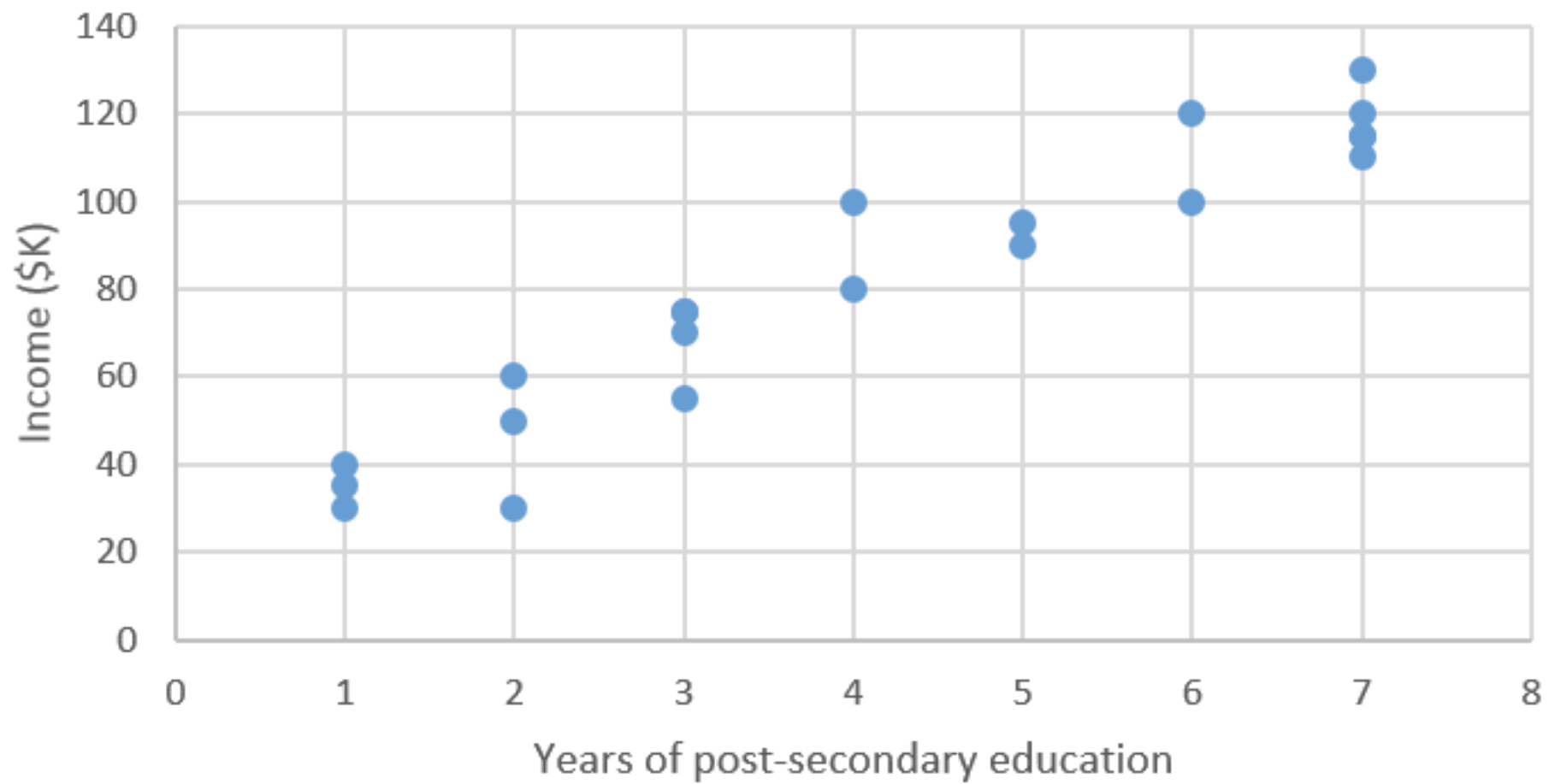
training set

Observation #	Years of Higher Education (X)	Income (Y)
1	4	\$80,000
2	5	\$91,500
3	0	\$42,000
4	2	\$55,000
...
N	6	\$100,000

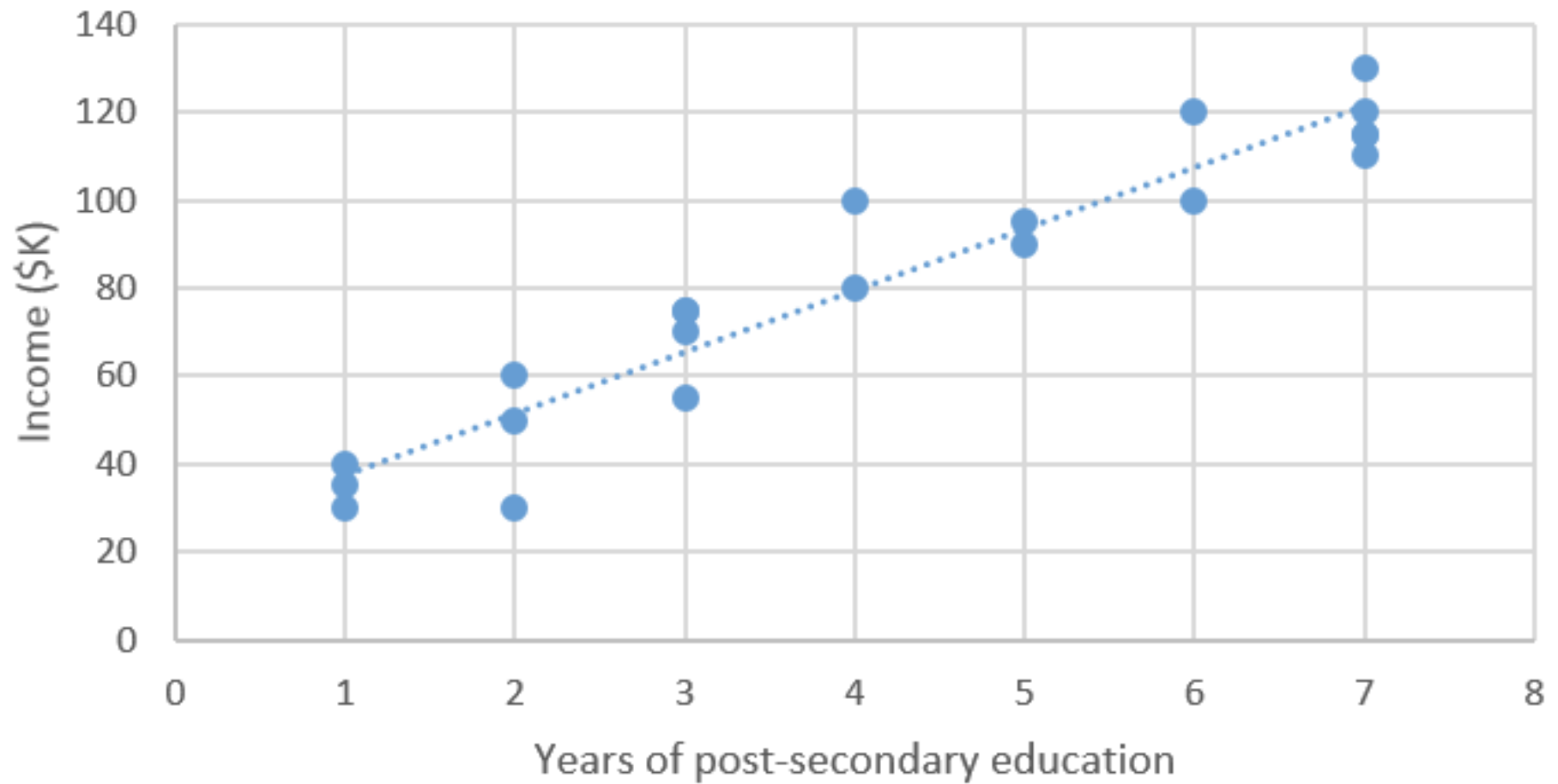
test set

1	4	???
2	6	???

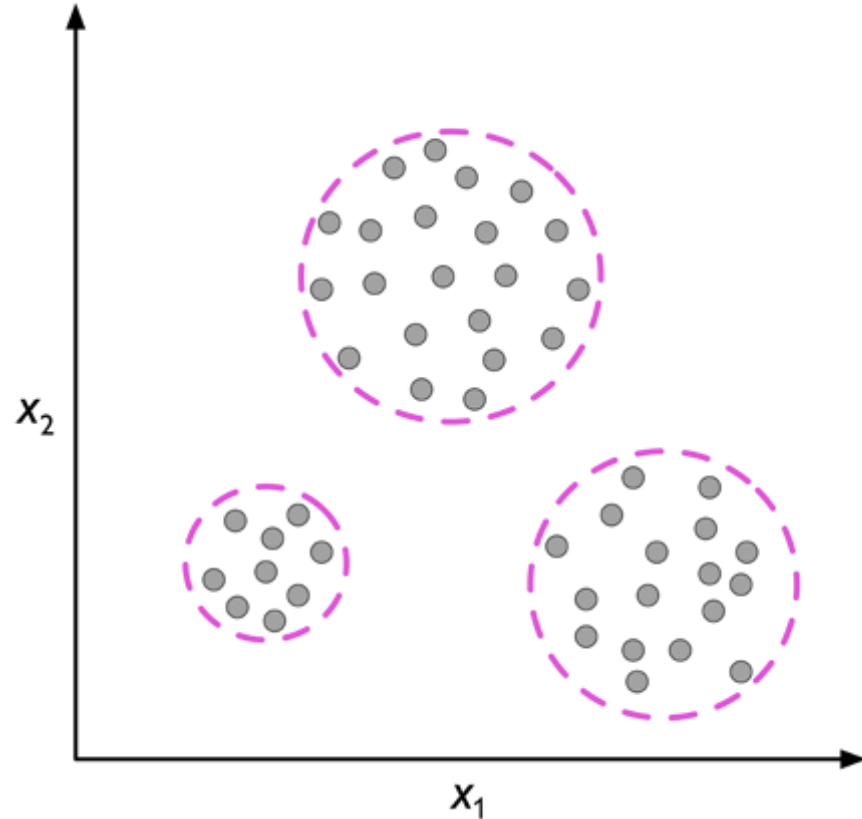
Income



Income



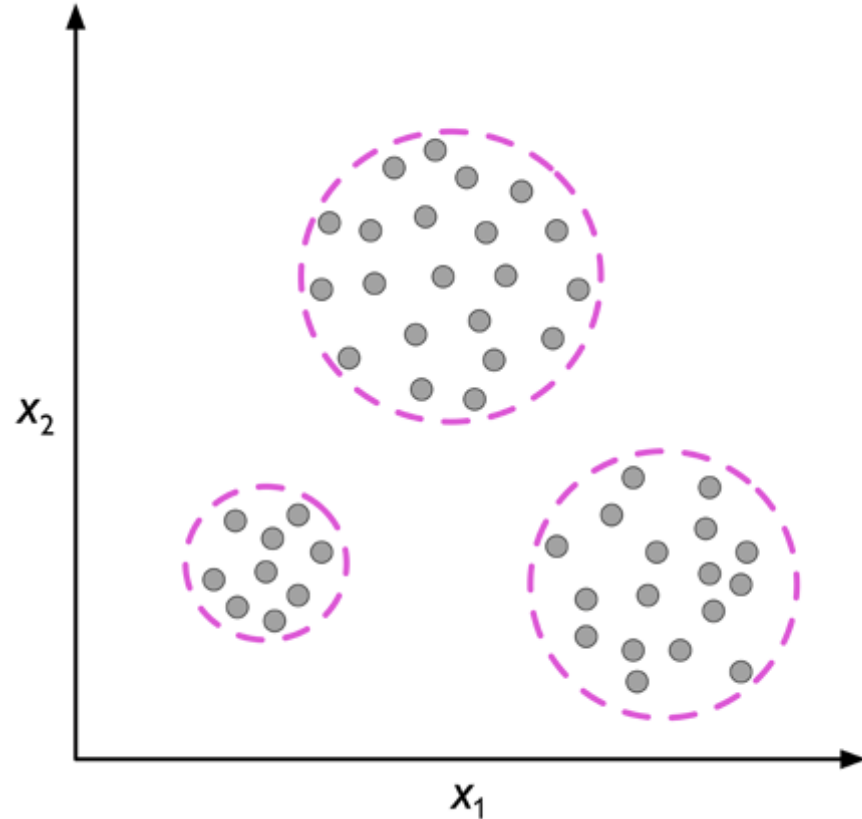
Clustering



Exploratory data analysis technique that allows to organize a pile of information into meaningful subgroups (clusters) without having any prior knowledge of their group memberships.

Each cluster defines a group of objects that share a certain degree of similarity but are more dissimilar to objects in other clusters.

Clustering

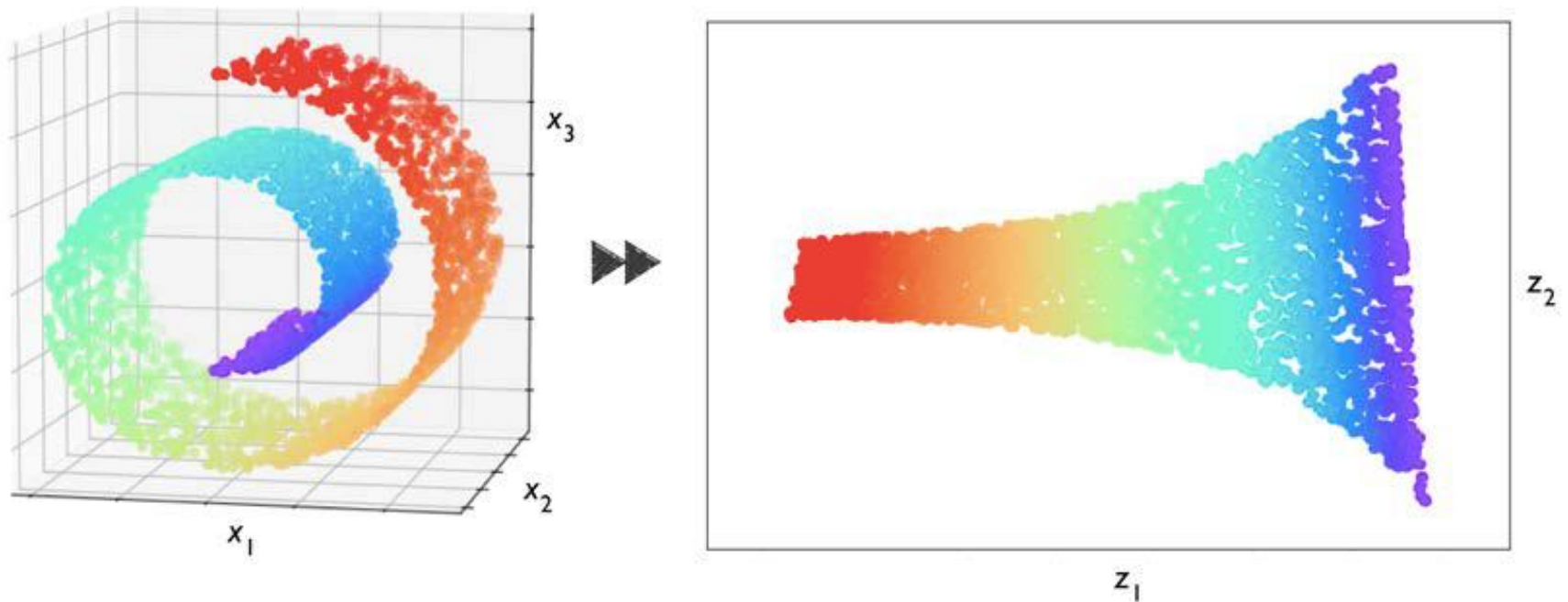


Clustering can be applied to organizing unlabeled data into three distinct groups based on the similarity of their features x_1 and x_2 .

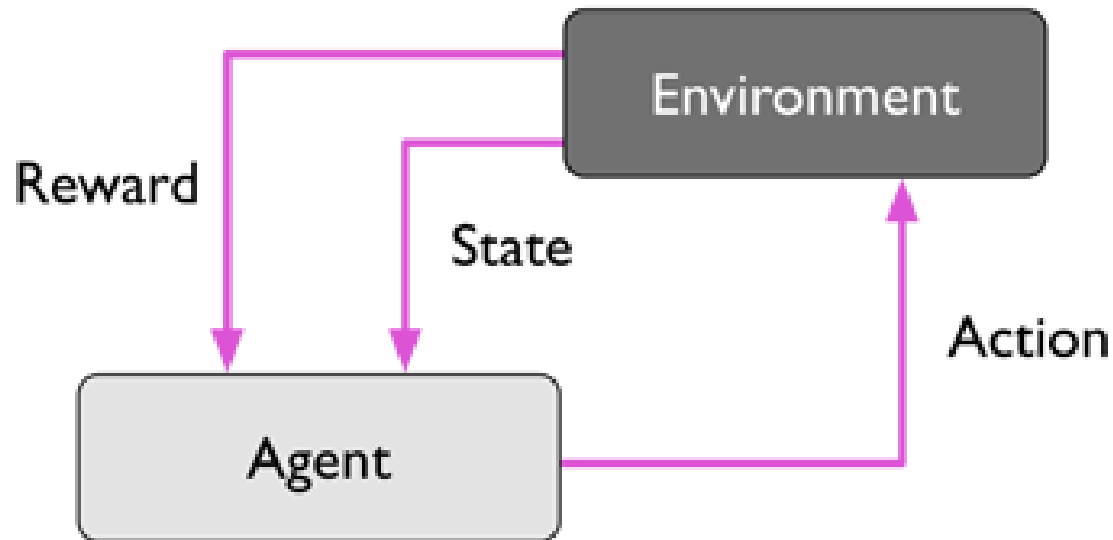
For example, it allows marketers to discover customer groups based on their interests, in order to develop distinct marketing programs.

Dimensionality reduction

Nonlinear dimensionality reduction applied to compress a 3D Swiss Roll onto a new 2D feature subspace:

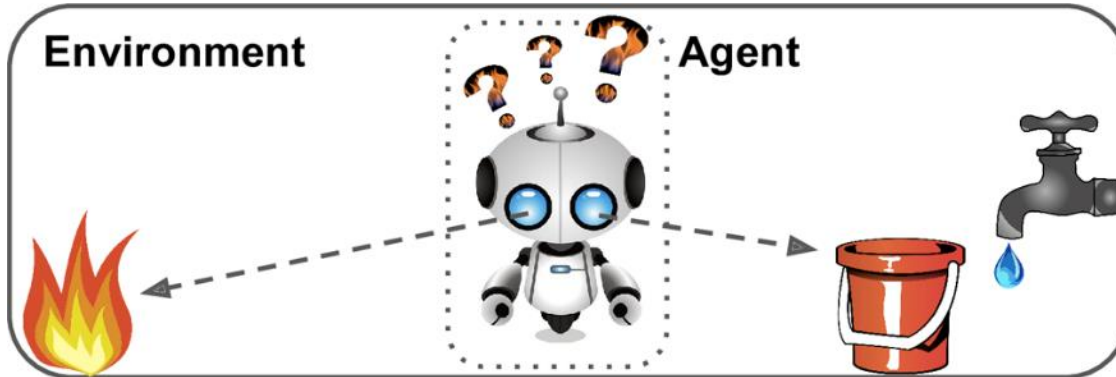


Reinforcement learning



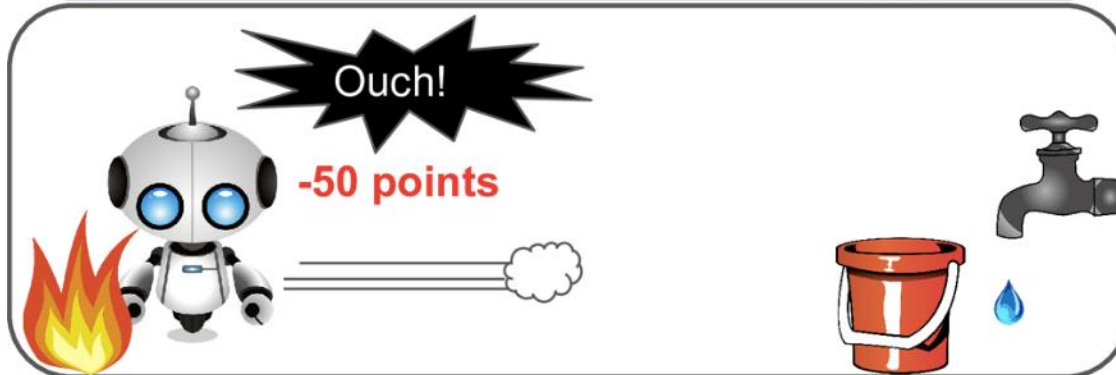
Reward can be defined as win or lose at the end of the game.

Reinforcement learning



1 Observe

2 Select action using policy



3 Action!

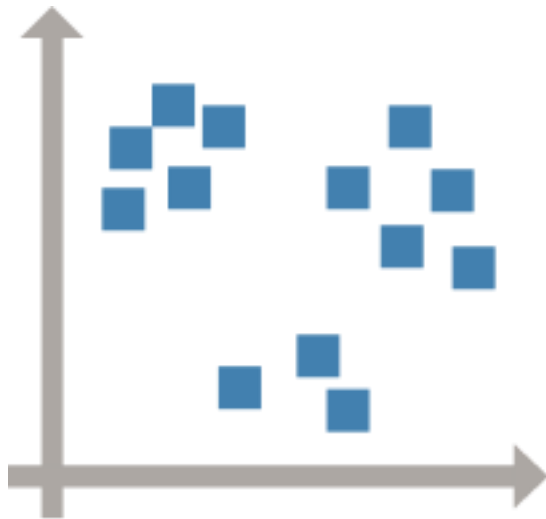
4 Get reward or penalty



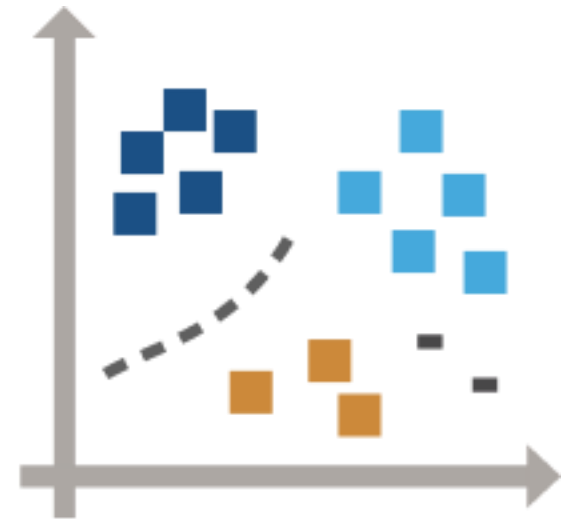
5 Update policy (learning step)

6 Iterate until an optimal policy is found

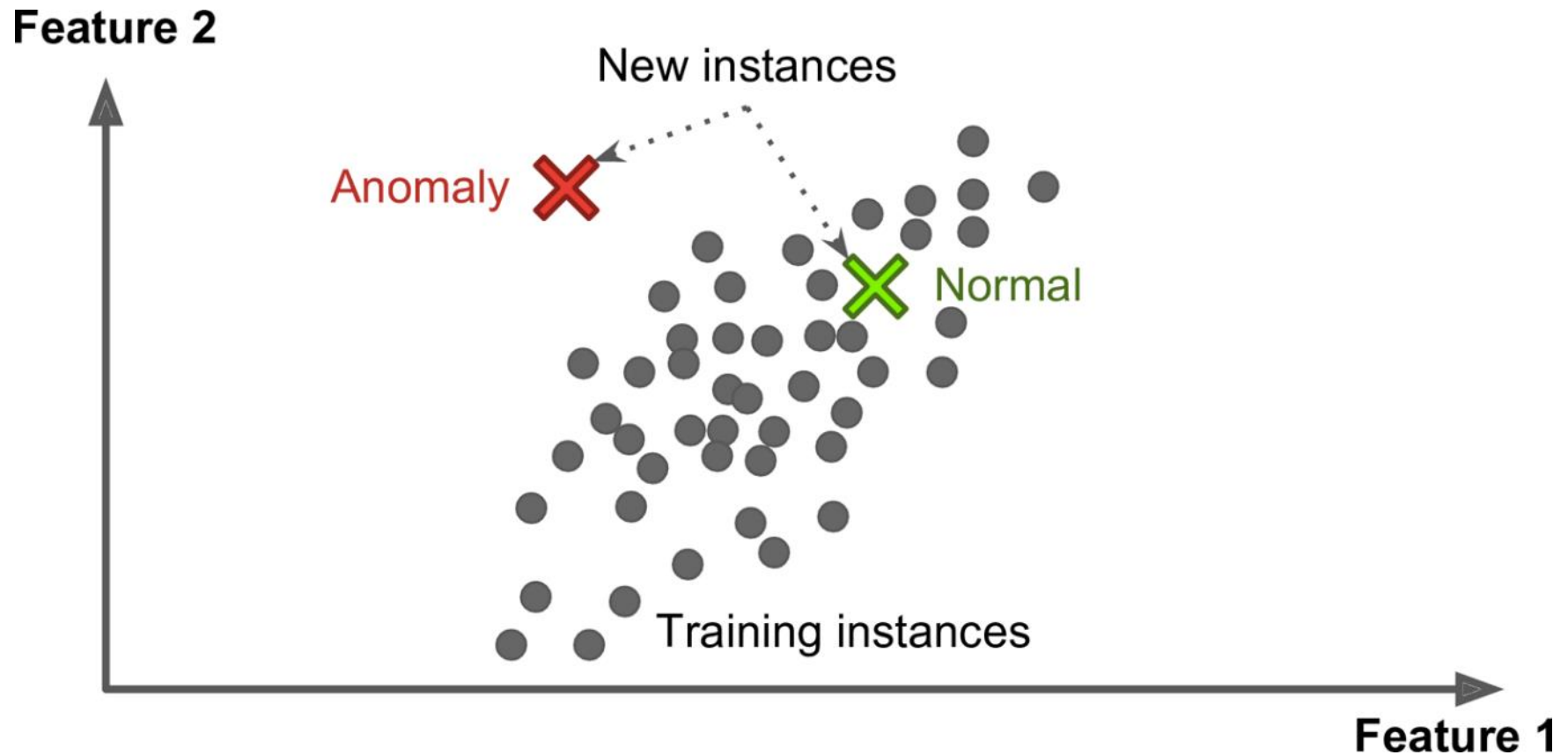
Aprendizagem não supervisionada



Clustering
Patterns in
the Data



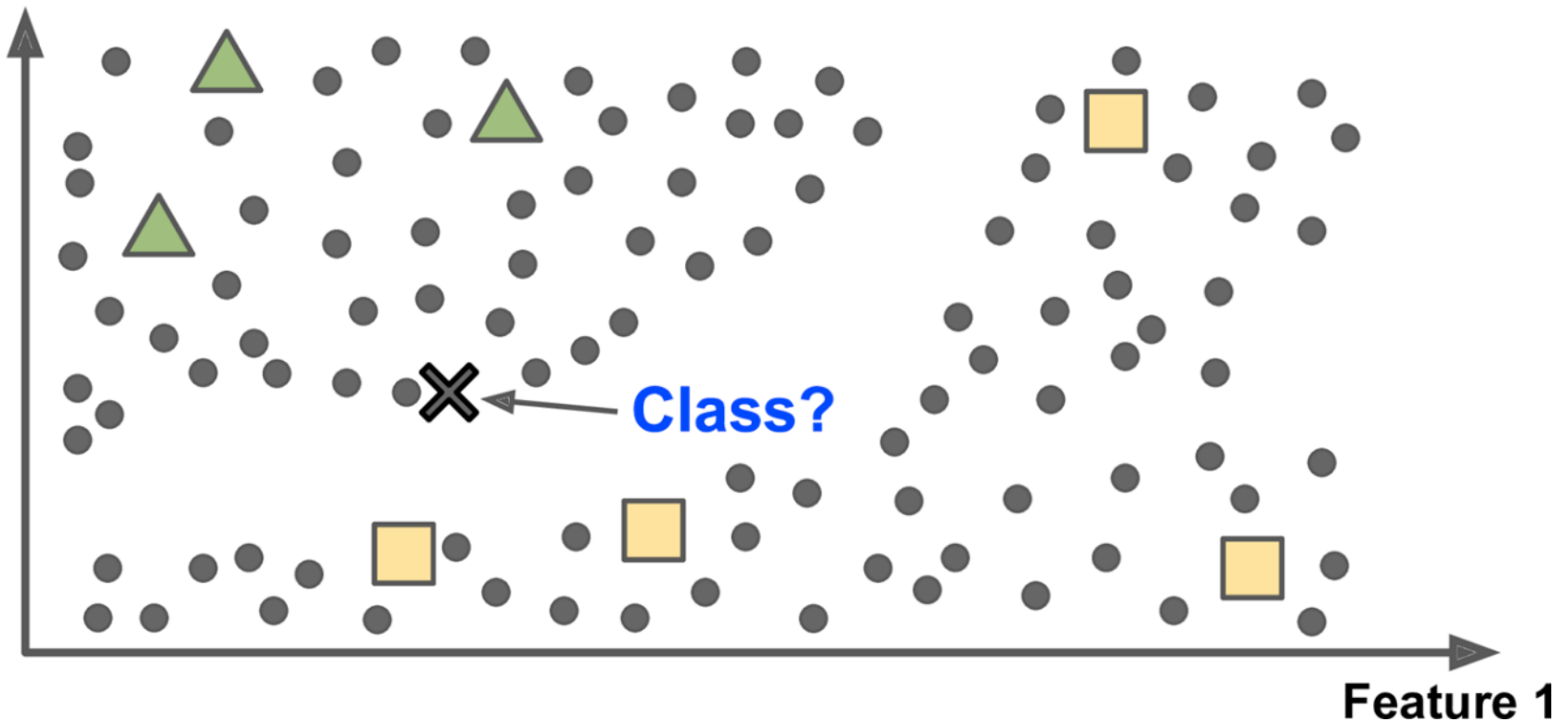
Unsupervised task - anomaly detection



Semisupervised learning

Some algorithms can deal with partially labeled training data, usually a lot of unlabeled data and a little bit of labeled data.

Feature 2

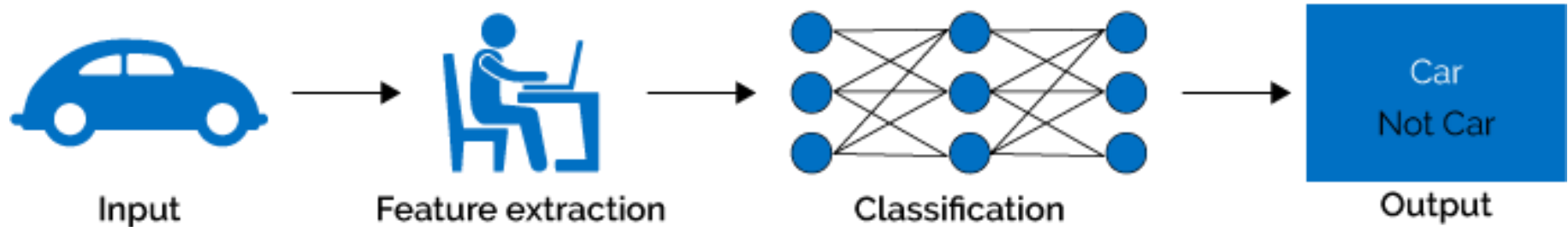


Semisupervised learning

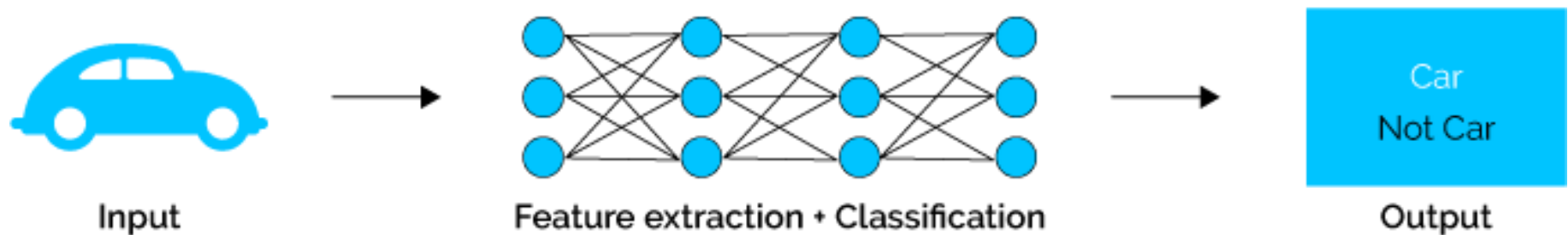
Photo-hosting services, such as Google Photos:

- **Unsupervised learning (clustering)**
 - Once you upload all your family photos to the service, it automatically recognizes that the same person A shows up in photos 1, 5, and 11, while another person B shows up in photos 2, 5, and 7.
- **Supervised learning**
 - Tell the system who these people are. Just one label per person, and it is able to name everyone in every photo, which is useful for searching photos.

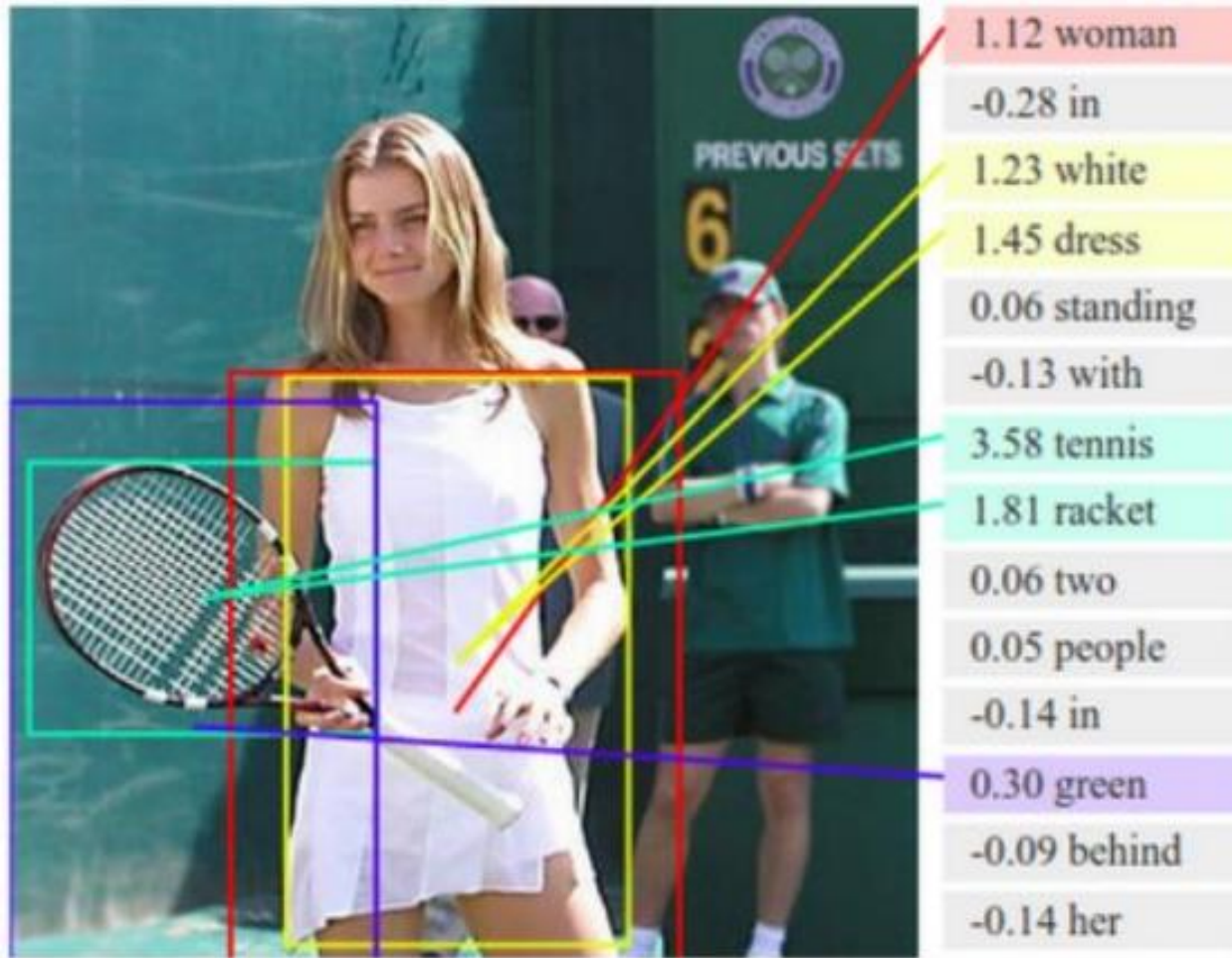
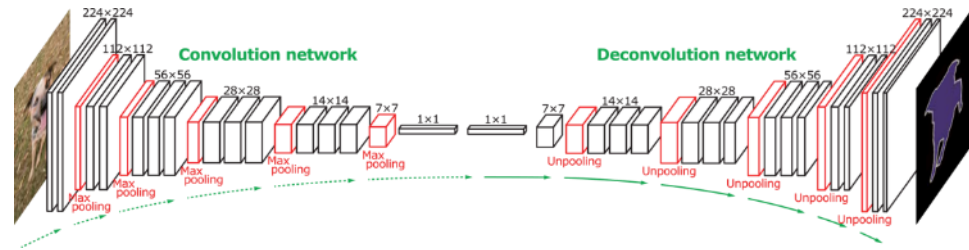
Machine Learning



Deep Learning

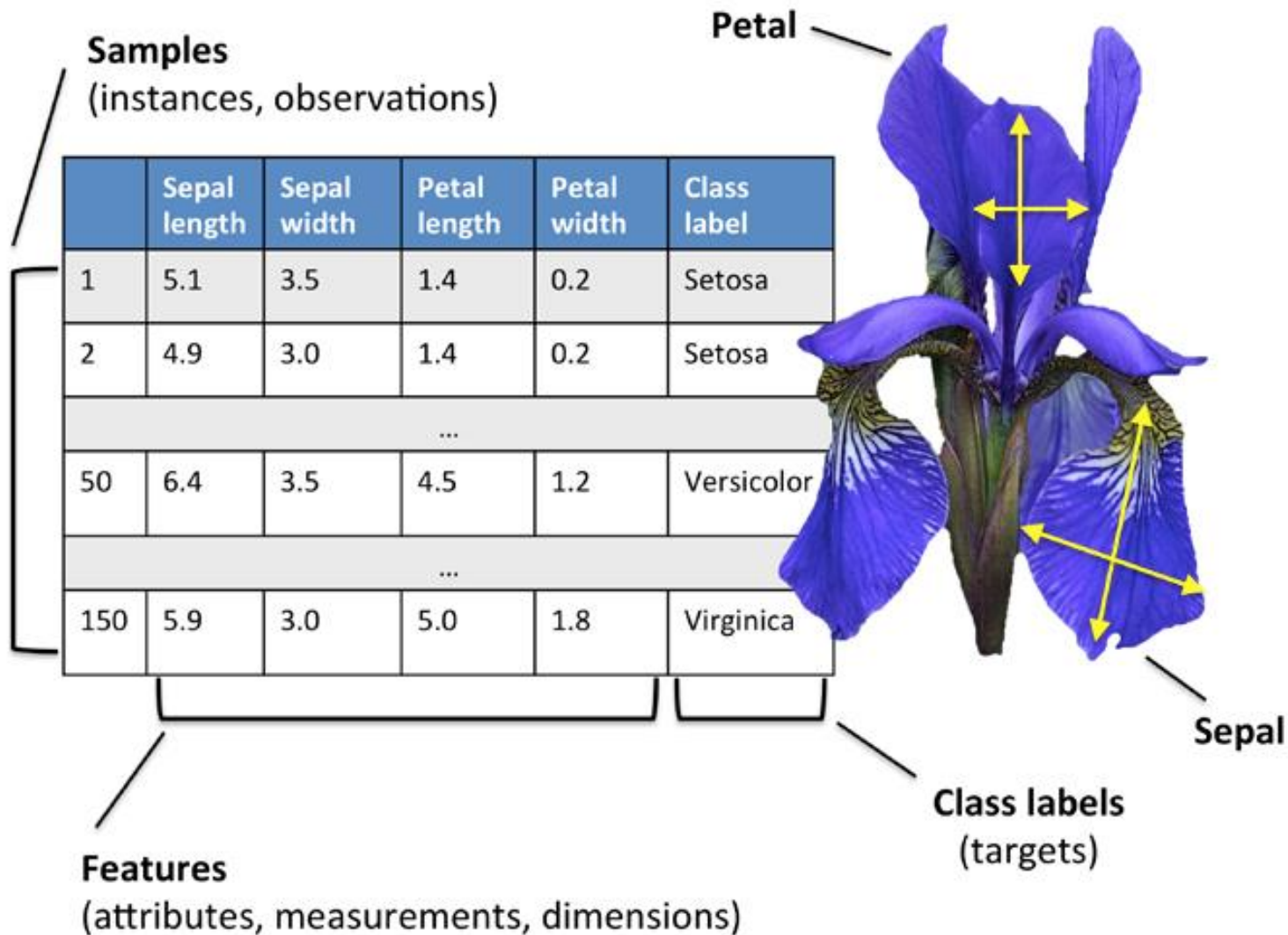


Deep Learning



Image,
automatically
annotated by
Deep Learning.

Terminology and notations



Iris dataset

150 samples and four features:

$$X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & x_4^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & x_4^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{(150)} & x_2^{(150)} & x_3^{(150)} & x_4^{(150)} \end{bmatrix}$$

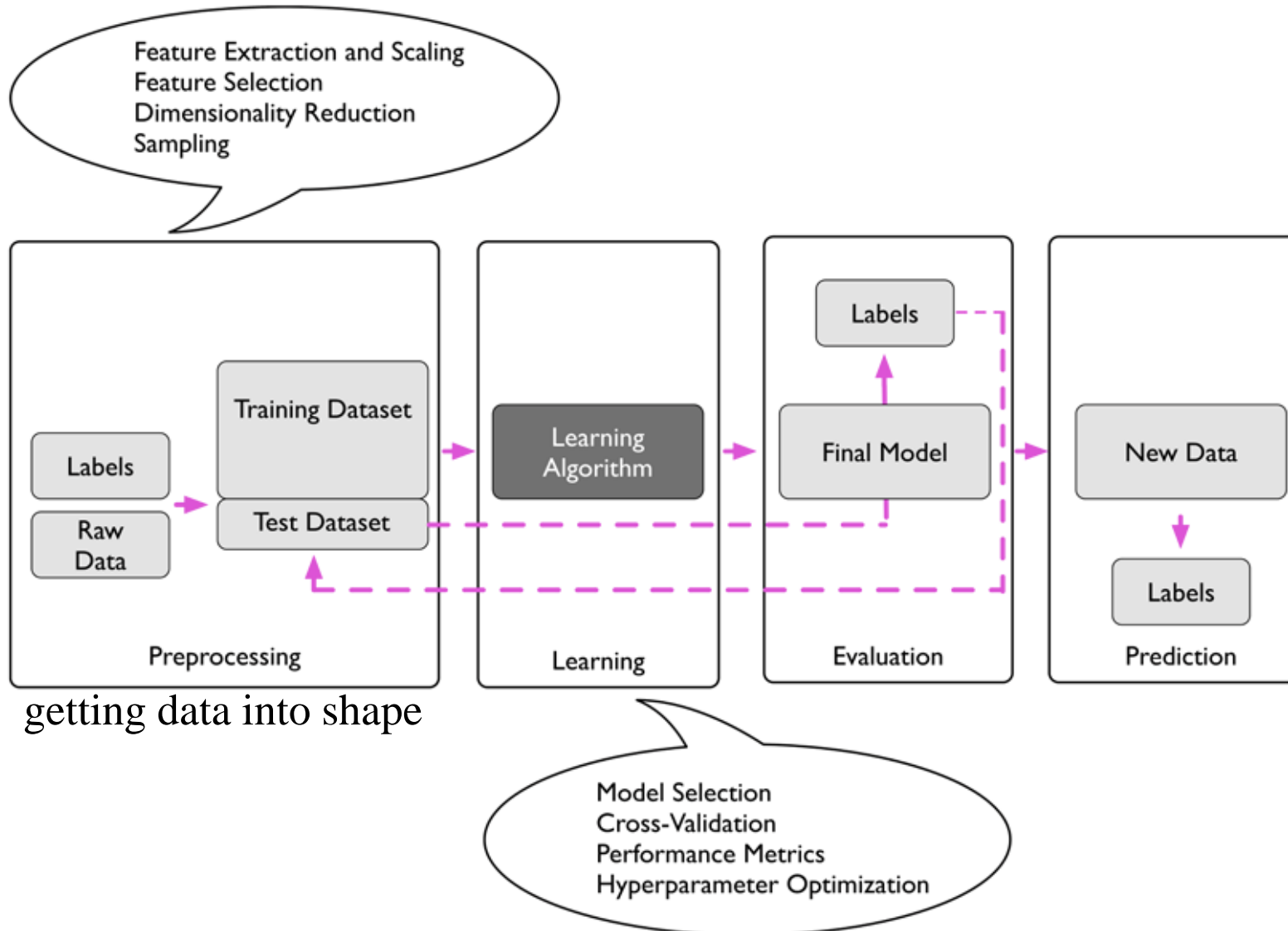
Row vector:

$$\mathbf{x}^{(i)} = \begin{bmatrix} x_1^{(i)} & x_2^{(i)} & x_3^{(i)} & x_4^{(i)} \end{bmatrix}$$

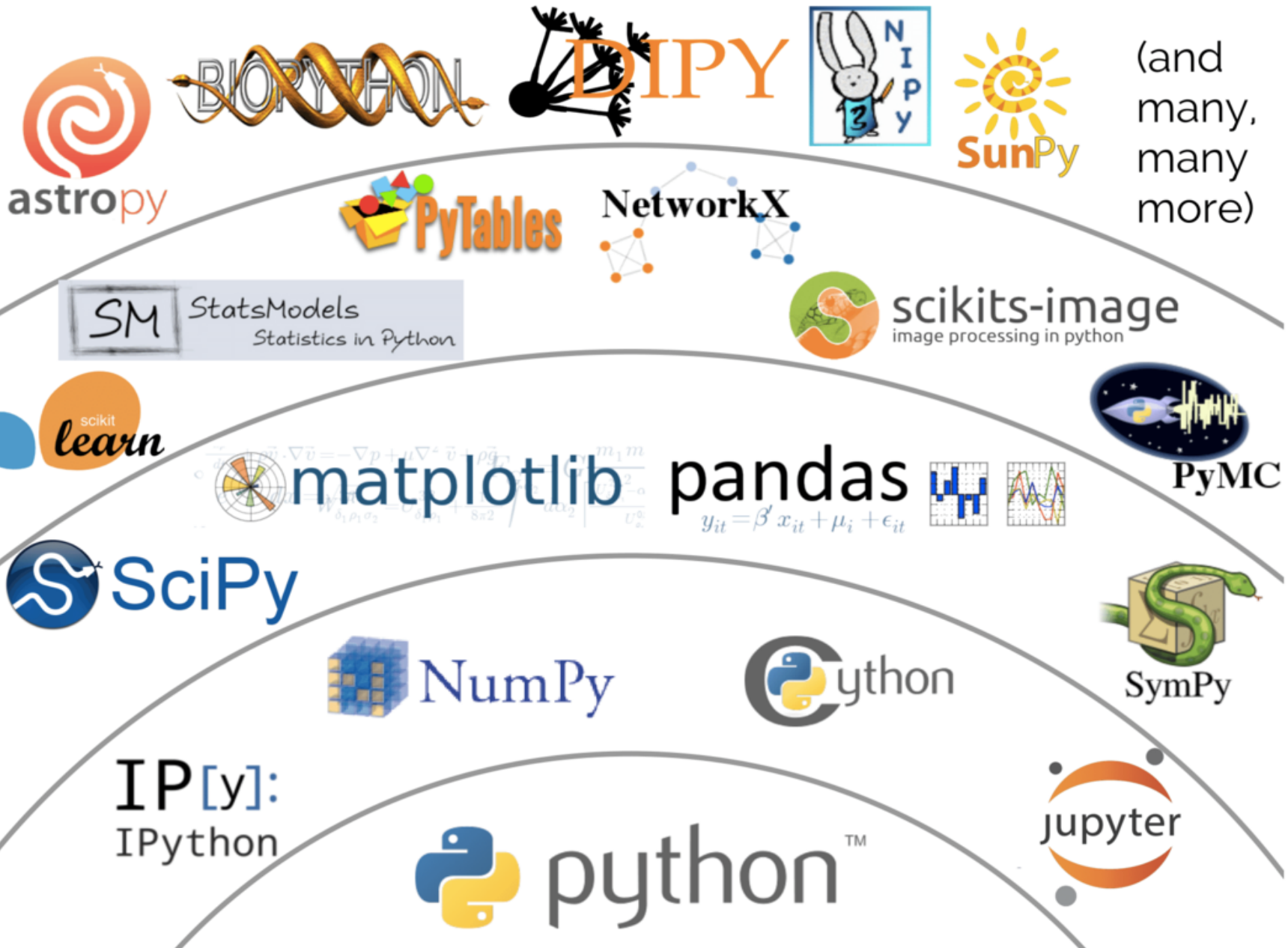
Column vector:

$$\mathbf{x}_j = \begin{bmatrix} x_j^{(1)} \\ x_j^{(2)} \\ \vdots \\ x_j^{(150)} \end{bmatrix}$$

Roadmap for building machine learning systems



Ecosystem Python



GitHub Student Developer Pack

Learn to ship software like a pro. There's no substitute for hands-on experience. But for most students, real world tools can be cost-prohibitive. That's why we created the GitHub Student Developer Pack with some of our partners and friends.

Sign up for Student Developer Pack



About DataCamp

DataCamp helps companies and individuals make better use of data. Our users build data fluency while learning from the world's top data scientists.

Benefit

Free 3-month individual subscription for students.

<https://education.github.com/pack>

Datacamp

Career Tracks

Our career tracks are hand-picked by industry experts. You will learn all you need to start a new career in the data science field.

All	Power BI	Python	R	SQL	Tableau
-----	----------	--------	---	-----	---------

? Which technology should I choose?



Data Scientist
with Python
ENROLLED

CERTIFICATION AVAILABLE



19%



Python Programmer

20%



Data Analyst
with Python

CERTIFICATION AVAILABLE




⌚ 36 hours 📁 9 courses





Machine Learning Scientist
with Python


⌚ 93 hours 📁 23 courses

Datacamp



 [Home](#) **[Learn](#)** [Workspace](#) [Certification](#) [Jobs](#) [For Business >](#)


 Progress


 Bookmarks


 Leaderboard


CATALOG


 Tracks 


 Courses

 Practice

 Projects

 Assessments






 Competitions

 Live Events

INTERACTIVE COURSE

Understanding Machine Learning

[Continue Course](#) [Bookmark](#)

 2 hours  12 Videos  36 Exercises  154,726 Participants  2,350 XP

Course Description

What's behind the machine learning hype? In this non-technical course, you'll learn everything you've been too afraid to ask about machine learning. There's no coding required. Hands-on exercises will help you get past the jargon and learn how this exciting technology powers everything from self-driving cars to your personal Amazon shopping suggestions. How does machine learning work, when can you use it, and what is the difference between AI and machine learning? They're all covered. Gain skills in this hugely in-demand and influential field, and discover why machine learning is for everyone!

Grupo de Práticas

Ciência de Dados Prática / GEAM

- Grupo de Estudos em Aprendizagem de Máquina
 - <https://t.me/joinchat/B74i-BMgvXniaMTf6PJCeQ>
 - <http://bit.ly/cdpgithub>
 - <http://bit.ly/cdpvideos>
 - <https://bit.ly/geam-videos-insight>

Canal InsightLab no YouTube

- <https://bit.ly/videosinsight>
- Curso de Imersão em Ciência de Dados
 - <http://bit.ly/curso-icd>

Plataforma Kaggle

- Início em 2010.
- Março de 2022 mais de 9 milhões de usuários.
 - <https://www.kaggle.com/regispires/unique-kaggle-users/>
- Maior e mais diversa comunidade de dados do mundo.
- Competições geralmente atraem mais de mil equipes e indivíduos.
- 3.500 submissões de soluções de competições por dia.
- Competidores altamente qualificados.
- Março/2017 – Google anunciou a aquisição do Kaggle.
- Diversos níveis de problemas.



Getting Started Prediction Competition

Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics



Kaggle · 6,174 teams · 3 years to go

[Overview](#)[Data](#)[Kernels](#)[Discussion](#)[Leaderboard](#)[More](#)[Submit Predictions](#)[Overview](#)<https://www.kaggle.com/c/titanic>[Description](#)[Evaluation](#)[Frequently Asked Questions](#)[Tutorials](#)

Start here if...

You're new to data science and machine learning, or looking for a simple intro to the Kaggle prediction competitions.

Competition Description

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

<https://www.kaggle.com/titericz>

Giba

Senior Data Scientist at NVIDIA RAPIDS

Curitiba, State of Paraná, Brazil

Joined 11 years ago · last seen in the past day

[GitHub](#) [Twitter](#) [LinkedIn](#) <https://rapids.ai/>

Followers 11521

Following 32



Competitions
Grandmaster

[Home](#) [Competitions \(259\)](#) [Datasets \(20\)](#) [Code \(120\)](#) [Discussion \(1,025\)](#) [Followers \(11,521\)](#)

[Contact User](#)

[Follow User](#)

Competitions
Grandmaster



Current Rank
13
of 201,789

Highest Rank
1

62

47

31

[PetFinder.my - Pawpu...](#)

· a year ago
Top 1%

1st
of 3537

[Santander Value Predi...](#)

· 5 years ago
Top 1%

1st
of 4463

[Melbourne University ...](#)

· 6 years ago
Top 1%

1st
of 477

Datasets
Contributor



Unranked

0

0

2

[Homecredit Best Solu...](#)

· 5 years ago

9
votes

[segmentation-models...](#)

· 2 years ago

8
votes

[pytorch_pretrained_1](#)

a year ago

7
votes

Notebooks
Grandmaster



Current Rank
65
of 275,377

Highest Rank
23

19

21

29

[The Property by Giba](#)

· 5 years ago

389
votes

[Better XGB Baseline](#)

· 4 months ago

378
votes

[Building and Visualizin...](#)

· 4 years ago

218
votes

Discussion
Grandmaster



Current Rank
50
of 342,124

Highest Rank
6

113

127

537

[Data Scientist Hero](#)

· 7 years ago

534
votes

[1st PLACE - WINNER ...](#)

· 8 years ago

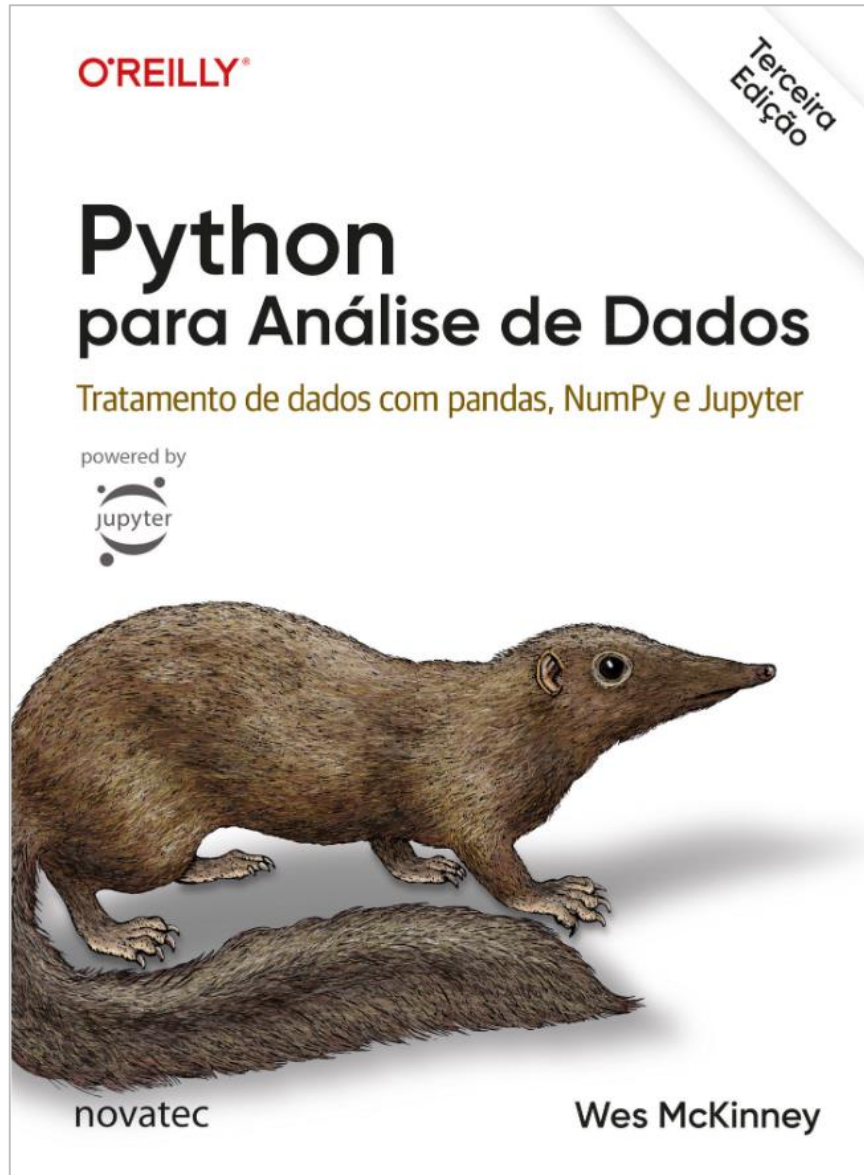
481
votes

[1st Place Removed So...](#)

· 3 years ago

390
votes

Bibliografia



MCKINNEY, Wes; **Python Para Análise de Dados: Tratamento de Dados com Pandas, NumPy e Jupyter.** 3ª Ed. Novatec, 2023.

Bibliografia



GÉRON, Aurélien; **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn, Keras & TensorFlow: Conceitos, Ferramentas e Técnicas para a Construção de Sistemas Inteligentes.** 2ª Ed. Alta Books, 2021.

PARTE I - Os conceitos básicos do aprendizado de máquina

1. O Cenário do Aprendizado de Máquina
2. Projeto de Aprendizado de Máquina Ponta a Ponta
3. Classificação
4. Treinando Modelos
5. Máquinas de Vetores de Suporte
6. Árvores de Decisão
7. Aprendizado Ensemble e Florestas Aleatórias (Bagging, Random Forests, Boosting, Stacking)
8. Redução de Dimensionalidade (PCA, Kernel PCA, LLE)
9. Técnicas de Aprendizado Não Supervisionado (Clusterização, Misturas de gaussianas)

PARTE II - Redes Neurais e Aprendizado Profundo

10. Introdução às Redes Neurais Artificiais com a Biblioteca Keras
11. Treinando Redes Neurais Profundas
12. Modelos Customizados e Treinamento com a Biblioteca TensorFlow
13. Carregando e Pré-processando Dados com a TensorFlow
14. Visão Computacional Detalhada das Redes Neurais Convolucionais
15. Processamento de Sequências Usando RNNs e CNNs
16. Processamento de Linguagem Natural com RNNs e Mecanismos de Atenção
17. Aprendizado de Representação e Aprendizado Gerativo com Autoencoders e GANs
18. Aprendizado por Reforço
19. Treinamento e Implementação de Modelos TensorFlow em Larga Escala

Bibliografia

RASCHKA, Sebastian; MIRJALILI, Vahid. [Python Machine Learning](#), 3rd Ed. Packt Publishing, 2019.

EXPERT INSIGHT

Python Machine Learning

Machine Learning and Deep Learning
with Python, scikit-learn, and TensorFlow 2

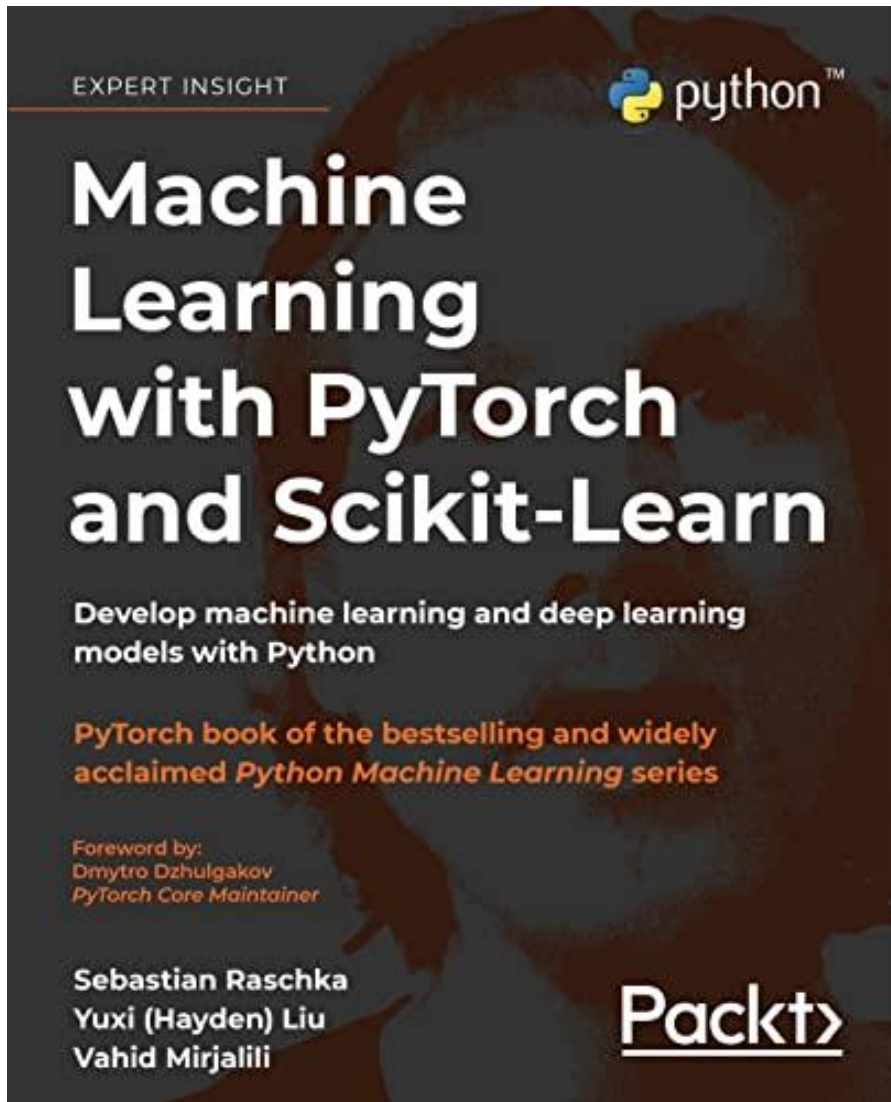
**Third Edition - Includes TensorFlow 2,
GANs, and Reinforcement Learning**

**Sebastian Raschka
& Vahid Mirjalili**

Packt

1. Giving Computers the Ability to Learn from Data
2. Training Simple Machine Learning Algorithms for Classification
3. A Tour of Machine Learning Classifiers Using Scikit-Learn
4. Building Good Training Sets - Data Preprocessing
5. Compressing Data via Dimensionality Reduction
6. Learning Best Practices for Model Evaluation and Hyperparameter Optimization
7. Combining Different Models for Ensemble Learning
8. Applying Machine Learning to Sentiment Analysis
9. Embedding a Machine Learning Model into a Web Application
10. Predicting Continuous Target Variables with Regression Analysis
11. Working with Unlabeled Data - Clustering Analysis
12. Implementing a Multilayer Artificial Neural Network from Scratch
13. Parallelizing Neural Network Training with TensorFlow
14. Going Deeper - The Mechanics of TensorFlow
15. Classifying Images with Deep Convolutional Neural Networks
16. Modeling Sequential Data using Recurrent Neural Networks
17. Generative Adversarial Networks for Synthesizing New Data
18. Reinforcement Learning for Decision Making in Complex Environments

Bibliografia



RASCHKA, Sebastian; MIRJALILI, Vahid. **Machine Learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python**, 1st Ed. Packt Publishing, 2022.

Bibliografia

CHOLLET, François; **Deep Learning with Python**. 2nd Ed. Manning Publications, 2021.

DEEP LEARNING with Python

SECOND EDITION

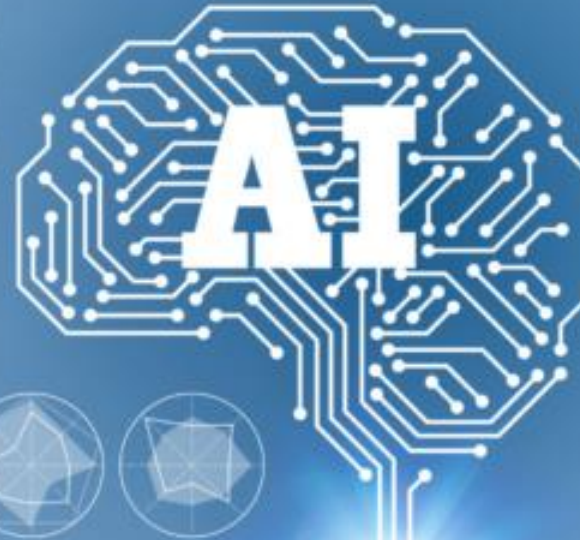
François Chollet

 MANNING



1. What is Deep Learning?
2. The Mathematical Building Blocks of Neural Networks
3. Introduction to Keras and Tensorflow
4. Getting Started with Neural Networks: Classification And Regression
5. Fundamentals of Machine Learning
6. The Universal Workflow of Machine Learning
7. Working With Keras: A Deep Dive
8. Introduction To Deep Learning For Computer Vision
9. Advanced Computer Vision
10. Deep Learning For Timeseries
11. Deep Learning For Text
12. Generative Deep Learning
13. Best Practices For The Real World
14. Conclusions

DEEP LEARNING



www.deeplearningbook.com.br

<https://www.deeplearningbook.com.br/>

Obrigado!

Dúvidas, comentários, sugestões?



GitHub curso



Regis Pires Magalhães - regismagalhaes@ufc.br