



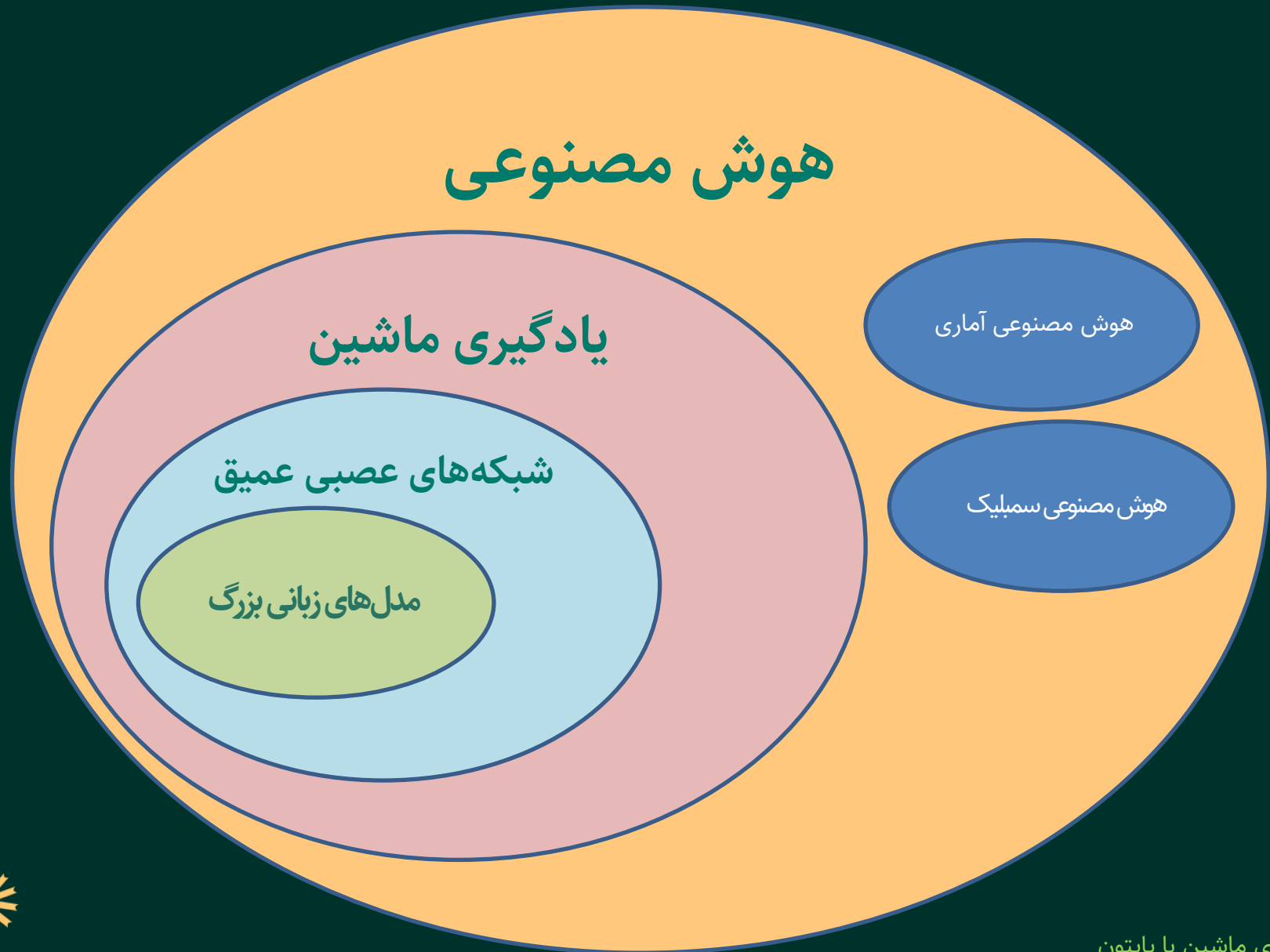
مدرسه پردازش و تحلیل داده دقیقه

مقدمه‌ای بر یادگیری ماشین

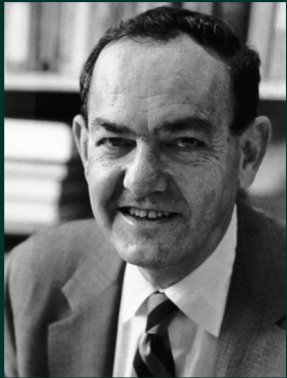
سعید مجیدی

یادگیری ماشین با پایتون

بهار ۱۴۰۴



یادگیری



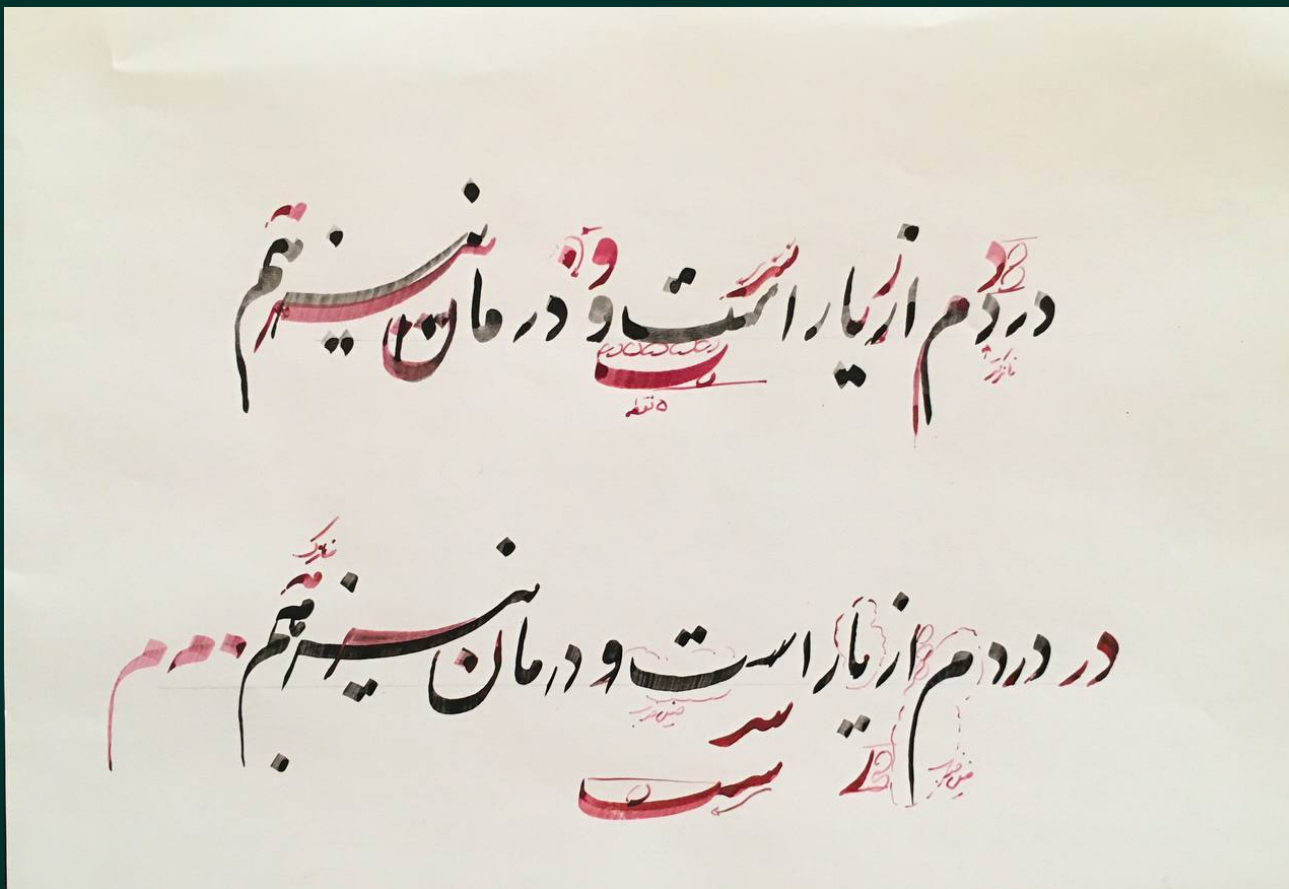
“یادگیری فرآیندی است که در طی آن یک سیستم عملکرد خود را براساس
تجرباتی که کسب می‌کند بهبود می‌بخشد.” **Herbert Simon**

“یادگیری کسب دانش درمورد جهان است.” **Kupfermann [1985]**

فرآیند آموزش

«در دم از بار است و در مان میهم»

فرآیند آموزش



فرآیند آموزش

دردم از بار است و درمان منم

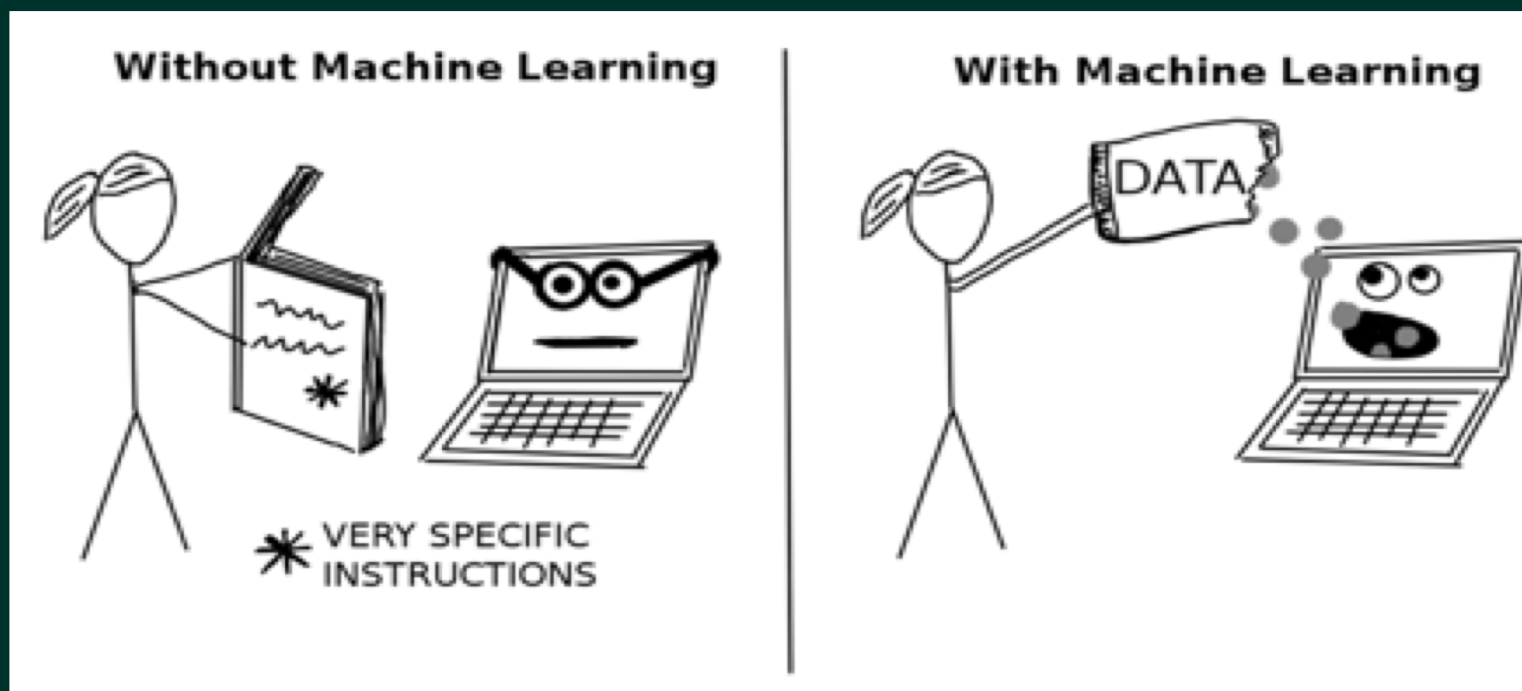
دردم از بار است و درمان منم

یادگیری ماشین



“یادگیری ماشین زمینه مطالعاتی فرشته،
دانش { است که بدون برنامه نویسی به
کامپیوتر توانایی یادگیری را می‌دهد.”

Arthut Samuel [1959]



• برنامه‌نویسی سنتی

$$X = [1, 3, 4]$$

$$y = 3x^2 + 2x + 1$$



• یادگیری ماشین

$$X = [1, 3, 4]$$

$$Y = [6, 34, 57]$$



چه هنگام از یادگیری ماشین استفاده می‌کنیم؟

- دسترسی به تخصص انسان امکان‌پذیر نباشد (پیمایش روی مریخ)
- نیروی انسانی قادر به توضیح تخصصش نباشد (تشخیص صوت)
- مدل‌ها باید سفارشی و شخصی‌سازی بشوند (داروهای شخصی‌سازی شده)
- مدل‌ها بر مبنای حجم عظیمی از داده‌ها کار می‌کنند (ژنتیک)



تعریف دقیق یک مسئله یادگیری

یک برنامه کامپیوتری

- از تجربه **E (experience)**
 - در مورد وظیفه **T (task)**
 - و با توجه به معیار کارایی **P (performance measure)**
- یاد می‌گیرد** اگر میزان کارایی آن در مورد وظیفه **T**، که با معیار **P** اندازه‌گیری می‌شود، با کسب تجربه **E** بهبود یابد.

Tom Mitchel[1998]

مثال: تشخیص هرزنامه (SPAM DETECTION)

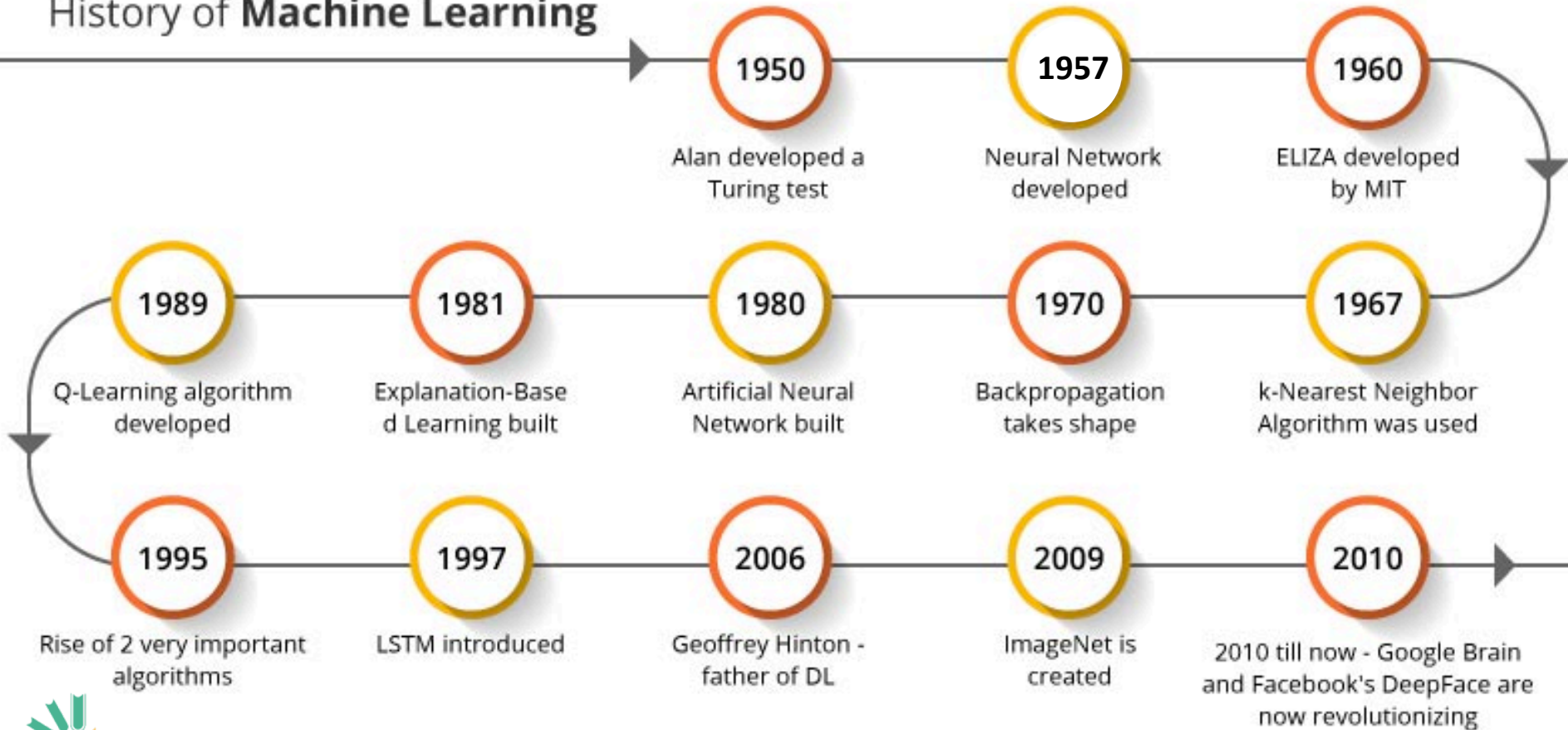
- **وظیفه** (مسئله) **T**: رده‌بندی ایمیل به عنوان *spam* / *not spam*
- **تجربه** **E**: مشاهده انجام این کار توسط یک فرد خبره
- **میزان کارایی** **P**: درصد تعداد ایمیل‌هایی که به درستی رده‌بندی شده‌اند.



تاریخچه



History of Machine Learning

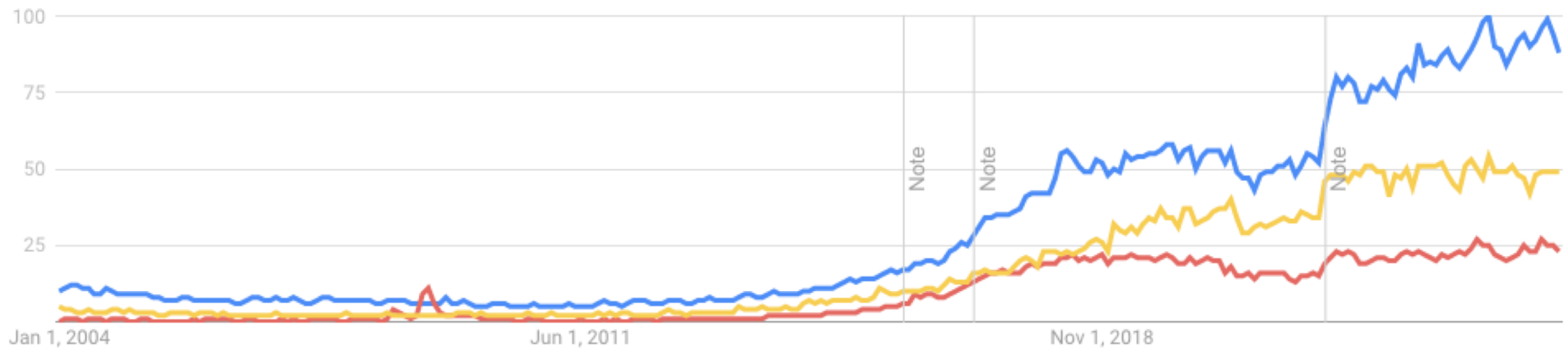


روند رشد

● Machine learning
Field of study

● Deep learning
Field of study

● Data Science
Search term



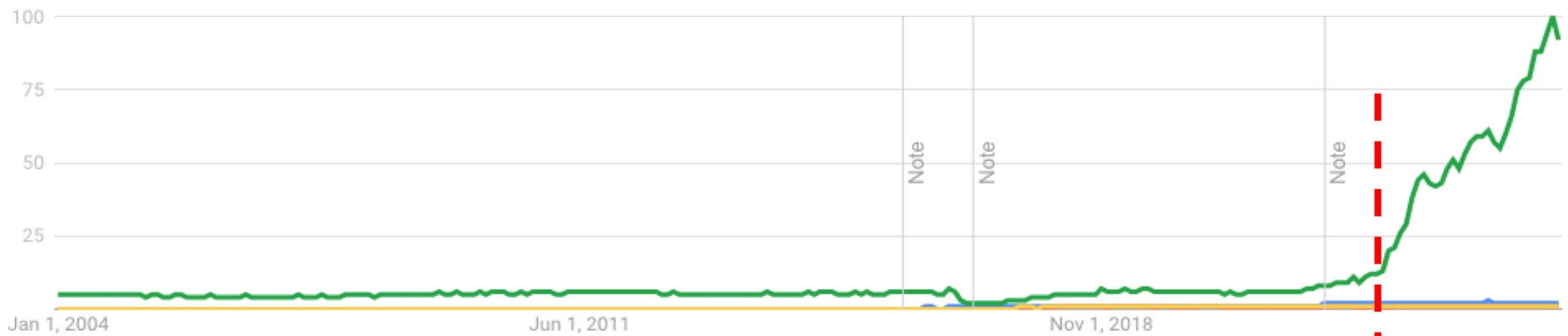
روند رشد

Machine learning
Field of study

Deep learning
Field of study

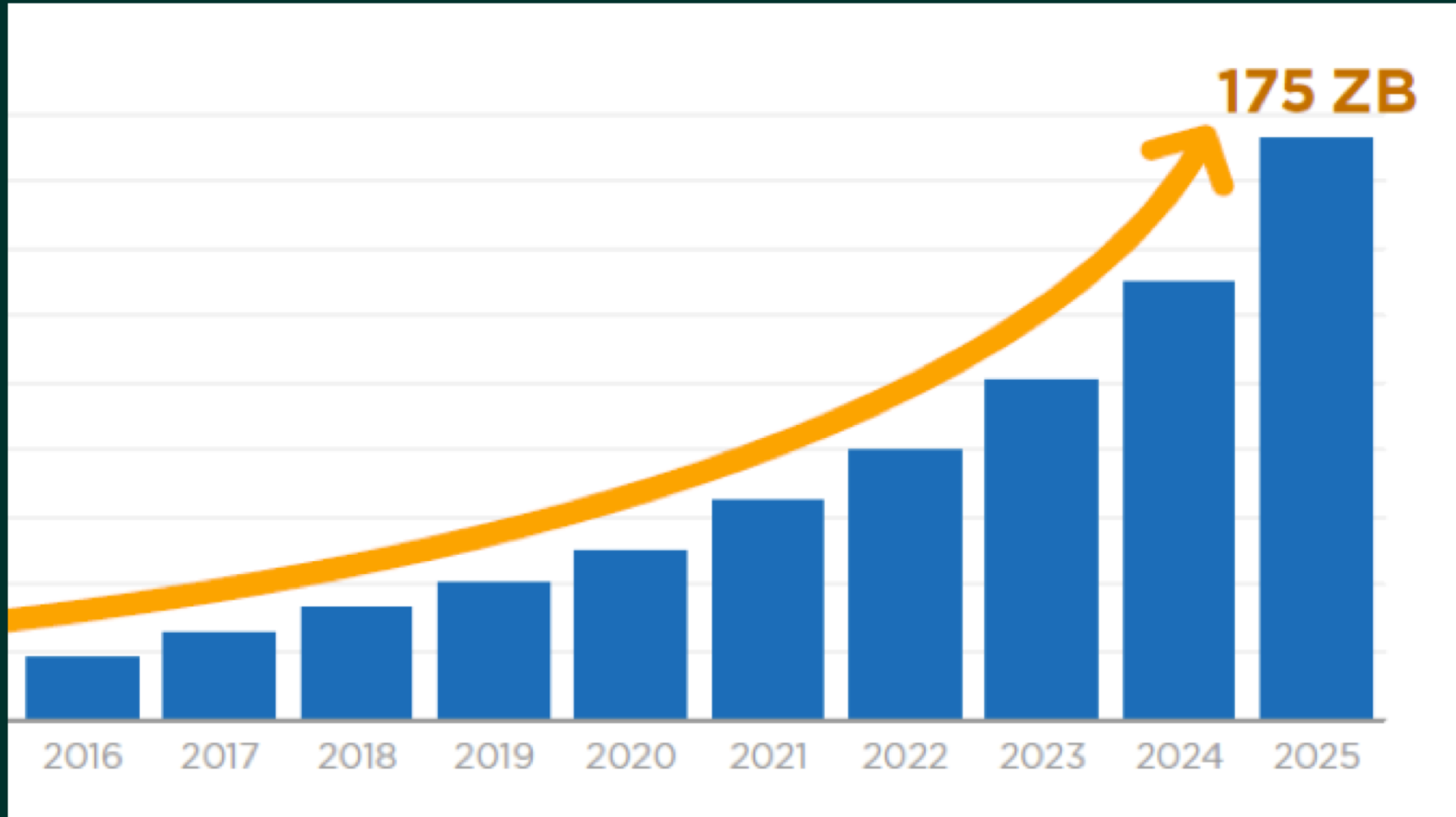
Data Science
Search term

Artificial intelligen...
Topic



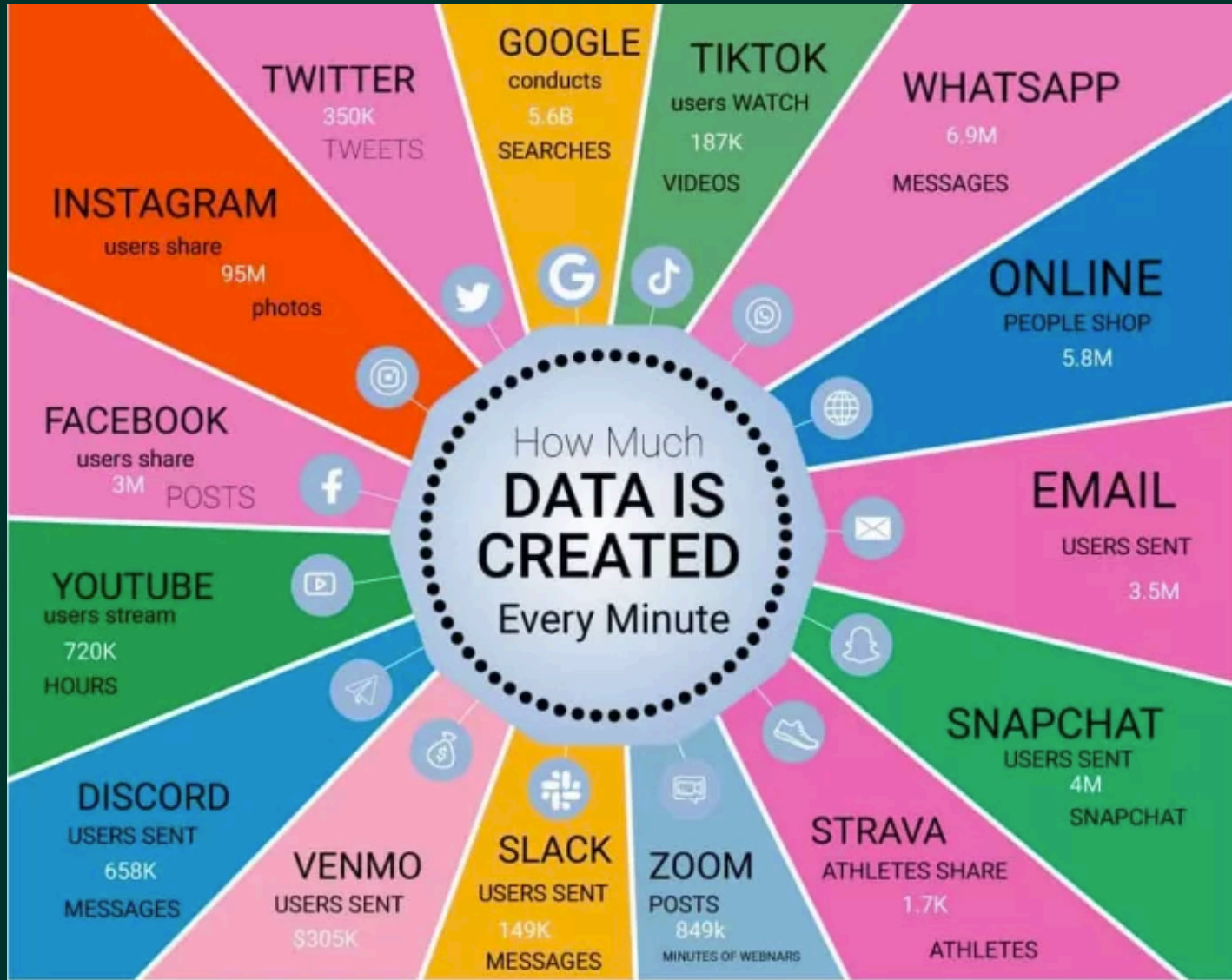
ظهور ChatGPT در نوامبر ۲۰۲۲

چرا حالا؟



رشد تولید داده

تولید بدون توقف داده



تنوع فرمت داده

	A	B	C	D	E	F	G
1	Purchase ID	Last name	First name	Birthday	Country	Date of purchase	Amount of purchase
2	1	Davidson	Michael	04/03/1986	United States	10/12/2016	37
3	2	Vito	Jim	09/01/1994	United Kingdom	02/02/2016	85
4	3	Johnson	Tom	23/08/1972	France	02/11/2016	83
5	4	Lewis	Peter	18/10/1979	Germany	22/11/2016	27
6	5	Koenig	Edward	13/05/1983	Argentina	26/03/2015	43
7	6	Preston	Jack	16/06/1991	United States	06/11/2016	77
8	7	Smith	David	11/03/1965	Canada	15/11/2016	23
9	8	Brown	Luis	03/09/1997	Australia	03/07/2015	74
10	9	Miller	Thomas	07/01/1980	Germany	07/11/2016	13
11	10	Williams	Bill	26/07/1960	United States	20/11/2015	80
12	11	Gemini	Alexia	12/09/1995	Canada	11/03/2017	35
13	12	Bond	James	25/02/1975	United Kingdom	12/08/2017	40
14	13	Burgle	Patricia	01/12/1990	United States	18/01/2015	55
15	14	Reding	Michelle	07/04/1985	Canada	23/02/2017	28
16	15	Harvey	Billy	14/07/1971	United Kingdom	12/01/2016	41
17							

داده‌ی ساختارمند

Structured Data

Introducing one of Australia's greatest treks and one of our newest trips! 🌄

The Cradle Mountain Overland track offers some of Tasmania's most stunning scenery – dramatic valleys, temperate rainforests, beautiful lakes and more. And this 6-day camping trip is the perfect way to experience it, alongside like-minded adventurers and experienced guides.

Why travel there with us?... See More

View Similar Products

116 11 Comments 12 Shares

Jessica Chapman
sigh I miss travel. With COVID travel restrictions here it's going to be a while.

Like Reply 1w

Abderrahman Chafiq
snow in the Atlas mountains from Morocco 🇲🇦🇲🇦🇲🇦🇲🇦

Like Reply 1w

Dian Clayton
Loved it, but snow on Mt Osier in November! 🌨️

Like Reply 1w

Harry Sayed
Beautiful 🌄

Like Reply 1w

Kelly McCarthy
Cradle Mountain is one of my favorite places. So many wild wombats to observe!! 🐨

Like Reply 1w

Most Relevant is selected, so some comments may have been filtered out.

داده‌ی بی‌ساختار

UnStructured Data



ادبیات یک پروژه یادگیری ماشین

- **مجموعه داده (Data Set):** داده‌ای که ماشین از آن یاد می‌گیرد.

- **نمونه داده (Instance):** یک عضو از مجموعه داده. معمولاً یک سطر از داده.

- **ویژگی (Feature):** هر مشخصه‌ای از یک نمونه داده که به فرآیند یادگیری کمک کند. معمولاً به شکل یک ستون در مجموعه داده.

- **برچسب (Label):** یا ویژگی هدف (Target): اطلاعاتی که ماشین یاد می‌گیرد آن را پیش‌بینی کند.

- **مدل (Model):** مجموعه‌ای از اعداد و قوانین که ماشین آن‌ها را یاد می‌گیرد.

Instances: $(1, 6), (3, 34), (4, 57)$

Features: $X = [1, 3, 4]$

Labels: $Y = [6, 34, 57]$

Model: $y = ax^2 + bx + c$
 $a = 3, b = 2, c = 1$



مدل (انگاره)



مواد لازم	مقدار لازم
گوشت سردست	نیم کیلو
پیاز	۱ عدد بزرگ
روغن	۲۵۰ گرم
لیمو عمانی	۵ تا ۶ عدد
لوبیا قرمز	نصف پیمانه
سبزی خورشت	۱ کیلوگرم
زردچوبه	به مقدار لازم
نمک	به مقدار لازم
فلفل	به مقدار لازم

مثال

Structured Data

برچسب
Label

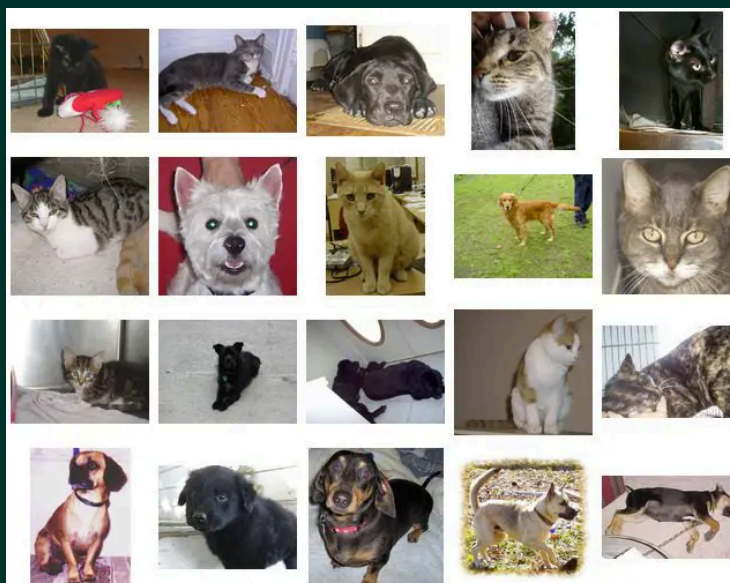
ویژگی
Feature

نمونه
Instance

Creditability	acc_Bal_bin	Pay_Status_bin	Instalment_bin	apartment_bin	Most_Val_Asset_bin	Duration_Credit	Credit_Amount	Age_bin
1	some balance	c. no problem	a. < 20	a. rented flat	c. No assets	18	1028	b. >= 25 yr
1	no balance	c. no problem	a. < 20	a. rented flat	a. Own house / land	24	2825	b. >= 25 yr
1	no balance	b. paid up	a. < 20	a. owner-occ / free	a. Own house / land	18	1239	b. >= 25 yr
0	no account	b. paid up	a. < 20	a. owner-occ / free	b. Savings / Car / Other	18	1216	a < 25 yr
1	some balance	b. paid up	a. < 20	a. rented flat	c. No assets	24	1258	b. >= 25 yr
0	some balance	c. no problem	a. < 20	a. rented flat	c. No assets	18	1864	b. >= 25 yr
1	some balance	b. paid up	a. < 20	a. rented flat	c. No assets	24	1474	b. >= 25 yr
1	no account	c. no problem	a. < 20	a. rented flat	c. No assets	24	1382	b. >= 25 yr
1	some balance	b. paid up	a. < 20	a. rented flat	c. No assets	12	640	b. >= 25 yr
1	some balance	b. paid up	c. (25-35)	a. rented flat	c. No assets	36	3919	a < 25 yr

Unstructured Data

- برچسب: سگ / گربه
- ویژگی: مشخصات مختلفی که از عکس استخراج می‌شود.



یادگیری ماشین با پایتون

مراحل یک پروژه یادگیری ماشین



انواع مسائل یادگیری

- یادگیری بانظارت (Supervised Learning)
- یادگیری بدون نظارت (Unsupervised Learning)
- یادگیری تقویتی (Reinforcement Learning)
- یادگیری نیمه نظارتی (Semi-supervised Learning)
- یادگیری خودنظارتی (Self-supervised Learning)

یادگیری بانظارت

SUPERVISED LEARNING



Getafe

Training



Cádiz

Test



?



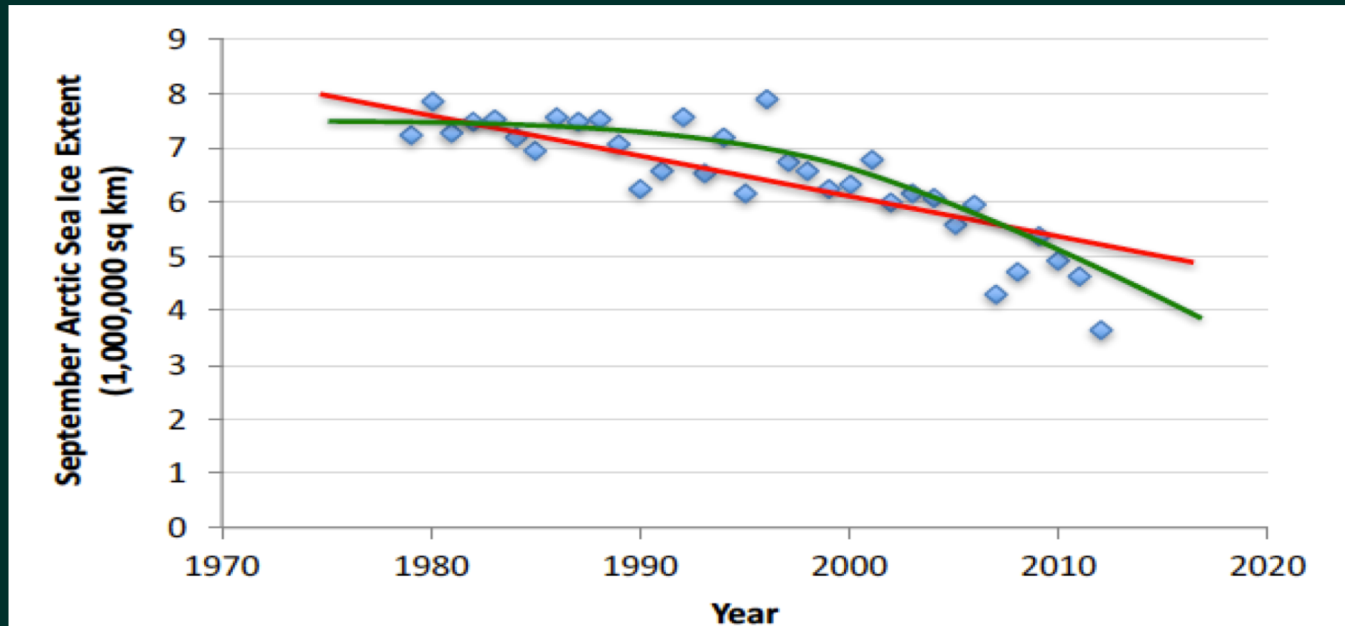
?

ویژگی‌های یادگیری با نظارت

- هدف: ایجاد یک تابع که مقدار خروجی را برای ورودی داده شده تخمین بزند (پیش‌بینی کند).
- در این نوع یادگیری "جواب صحیح" از ابتدا در دسترس است (برچسب یا مقدار هدف)
 - Tag, Label, Target Value, Golden Value
- مسائل این نوع یادگیری به دو دسته تقسیم‌بندی می‌شوند:
 - رگرسیون (Regression): خروجی یک مقدار پیوسته است.
 - دسته‌بندی (Classification): خروجی یک مقدار گسسته است.

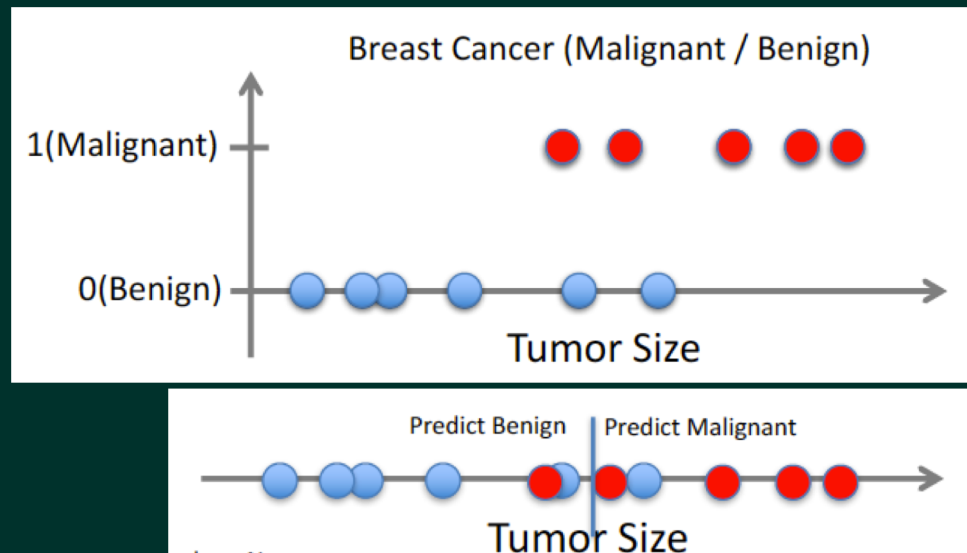
رگرسیون (REGRESSION)

- برای ورودی داده شده به شکل: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- تابع $f(x)$ را یاد می‌گیرد که به‌ازای مقدار ورودی x ، مقدار خروجی y را پیش‌بینی کند.
- مقدار خروجی y یک مقدار حقیقی و پیوسته است.
- پیش‌بینی دمای هوا، قیمت خانه، قیمت سهام، ...

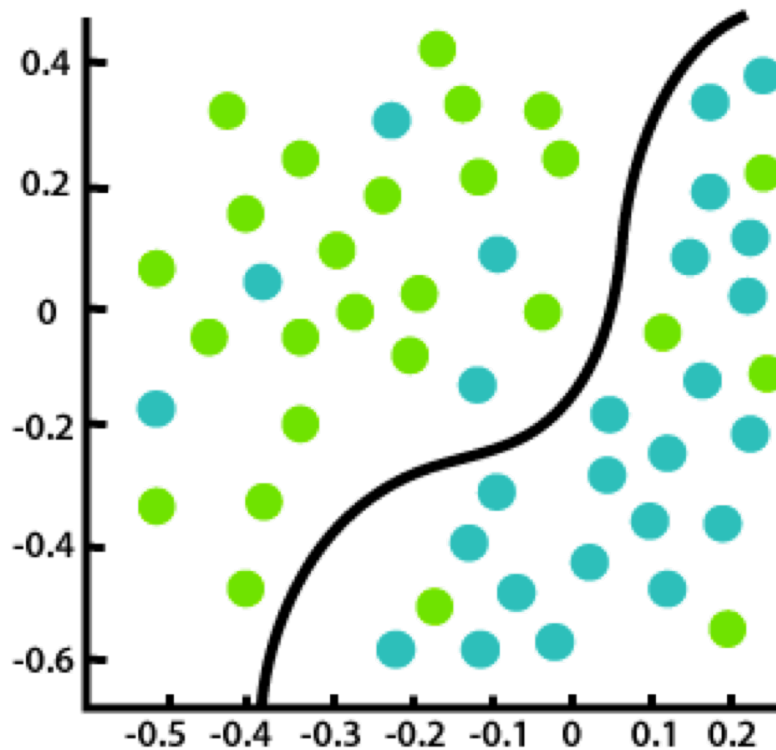


دسته‌بندی (CLASSIFICATION)

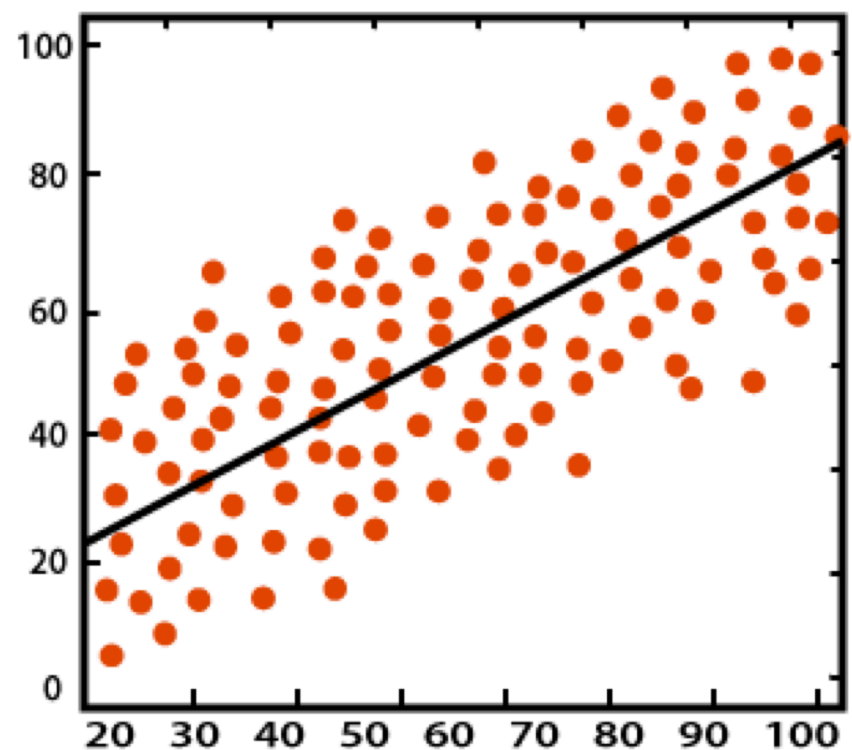
- برای ورودی داده شده به شکل: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- تابع $f(x)$ را یاد می‌گیرد که به‌ازای مقدار ورودی x ، مقدار خروجی y را پیش‌بینی کند.
- مقدار خروجی y شماره یک رده (Class- Category) است.
- تشخیص سرطان، اشیاء در عکس، شماره پلاک، دسته‌بندی اخبار ...



مسائل یادگیری با نظارت



Classification



Regression

الگوریتم‌ها (روش‌ها)ی معروف یادگیری بانظارت

رگرسیون

- رگرسیون خطی (Linear Regression)
- رگرسیون محلی (Local Regression)
- درخت تصمیم (Decision Tree)
- شبکه‌های عصبی (Neural Networks)
-

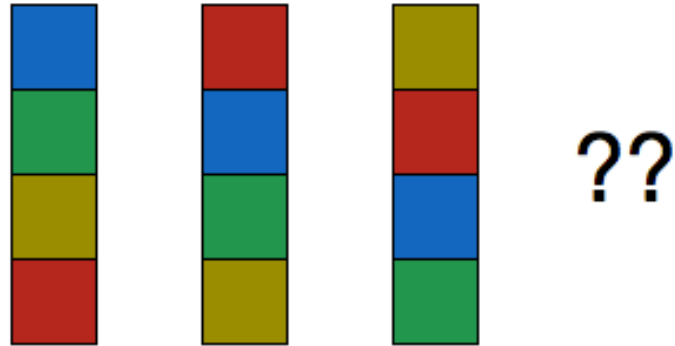
دسته‌بندی

- رگرسیون لاجیستیک (Logistic Regression)
- درخت تصمیم (Decision Tree)
- K-نزدیک‌ترین همسایه (KNN)
- ماشین‌های بردار پشتیبان (SVM)
- بیز ابتدایی (Naive Bayes)
- شبکه‌های عصبی (Neural Networks)
-

یادگیری بدون نظارت

UNSUPERVISED LEARNING

زنگ تفریح



a)



b)



c)



d)

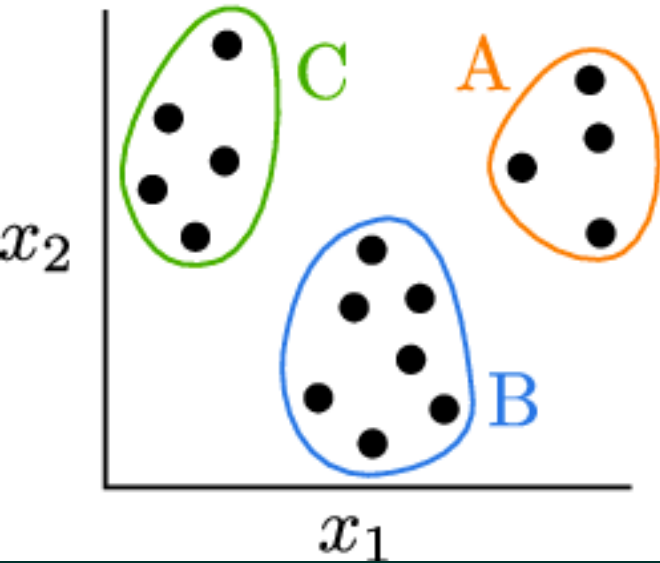




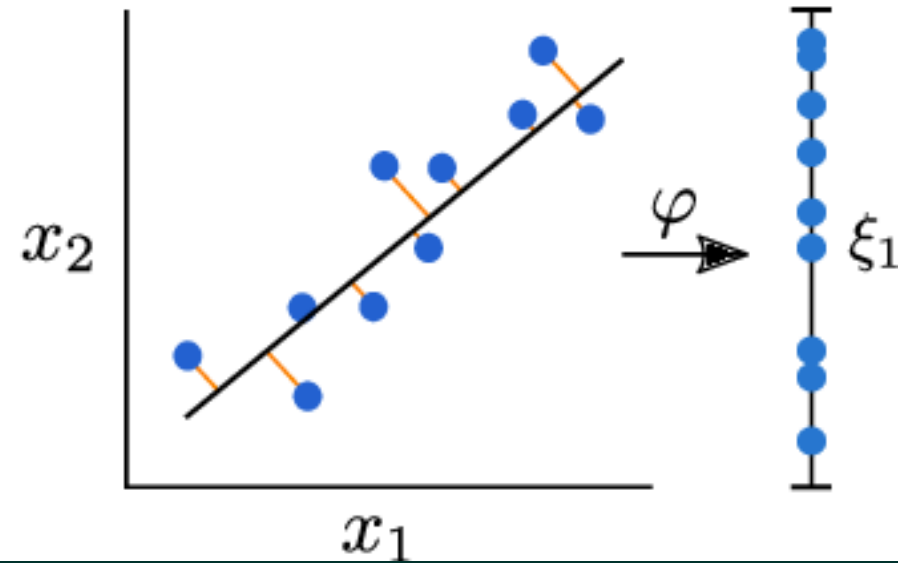
ویژگی‌های یادگیری بدون نظارت

- یادگیرنده تنها به داده‌های بدون برچسب دسترسی دارد.
- در این نوع یادگیری "جواب صحیح" از ابتدا در دسترس نمی‌باشد.
- تعداد صحیح رده‌ها از قبل مشخص نیست و الگوریتم باید به‌تنهایی ساختار پنهان حاکم بر داده‌ها را تشخیص دهد.
- مسائل این نوع یادگیری را می‌توان به دسته‌های زیر تقسیم‌بندی کرد:
 - خوشه‌بندی (Clustering)
 - کاهش ابعاد (Dimensionality Reduction)

مسائل یادگیری بدون نظارت

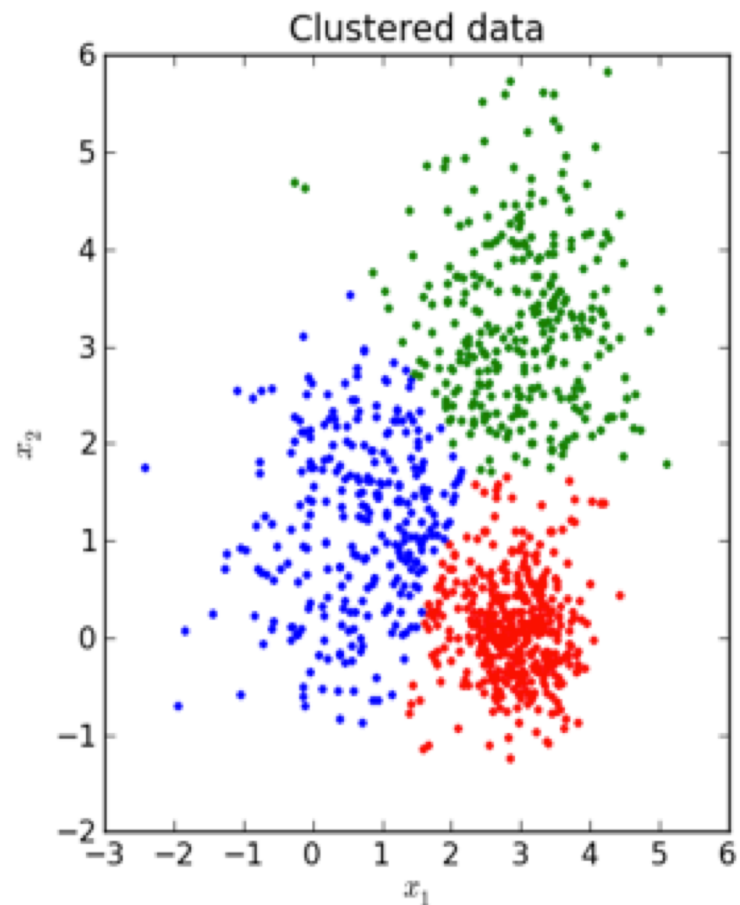
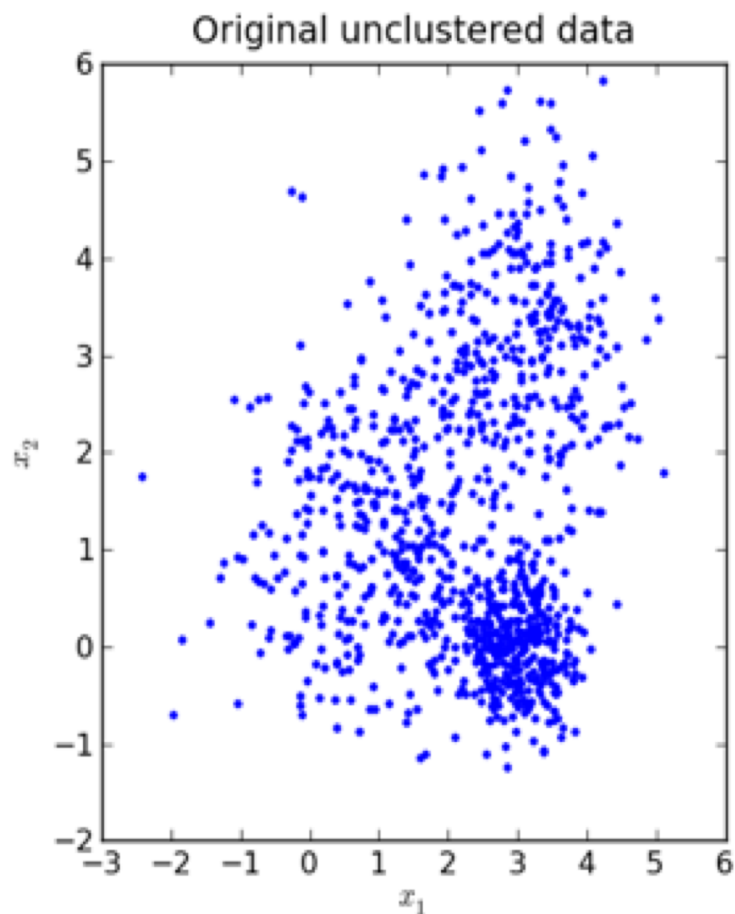


خوشه‌بندی (Clustering)



کاهش ابعاد (Dimensionality Reduction)

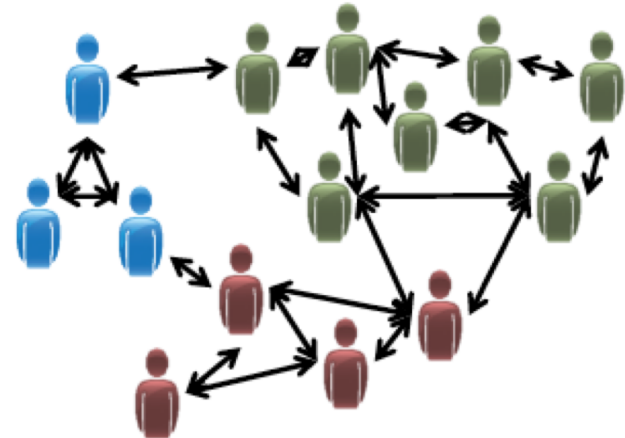
خوشه‌بندی (CLUSTERING)



مثال



Organize computing clusters



Social network analysis



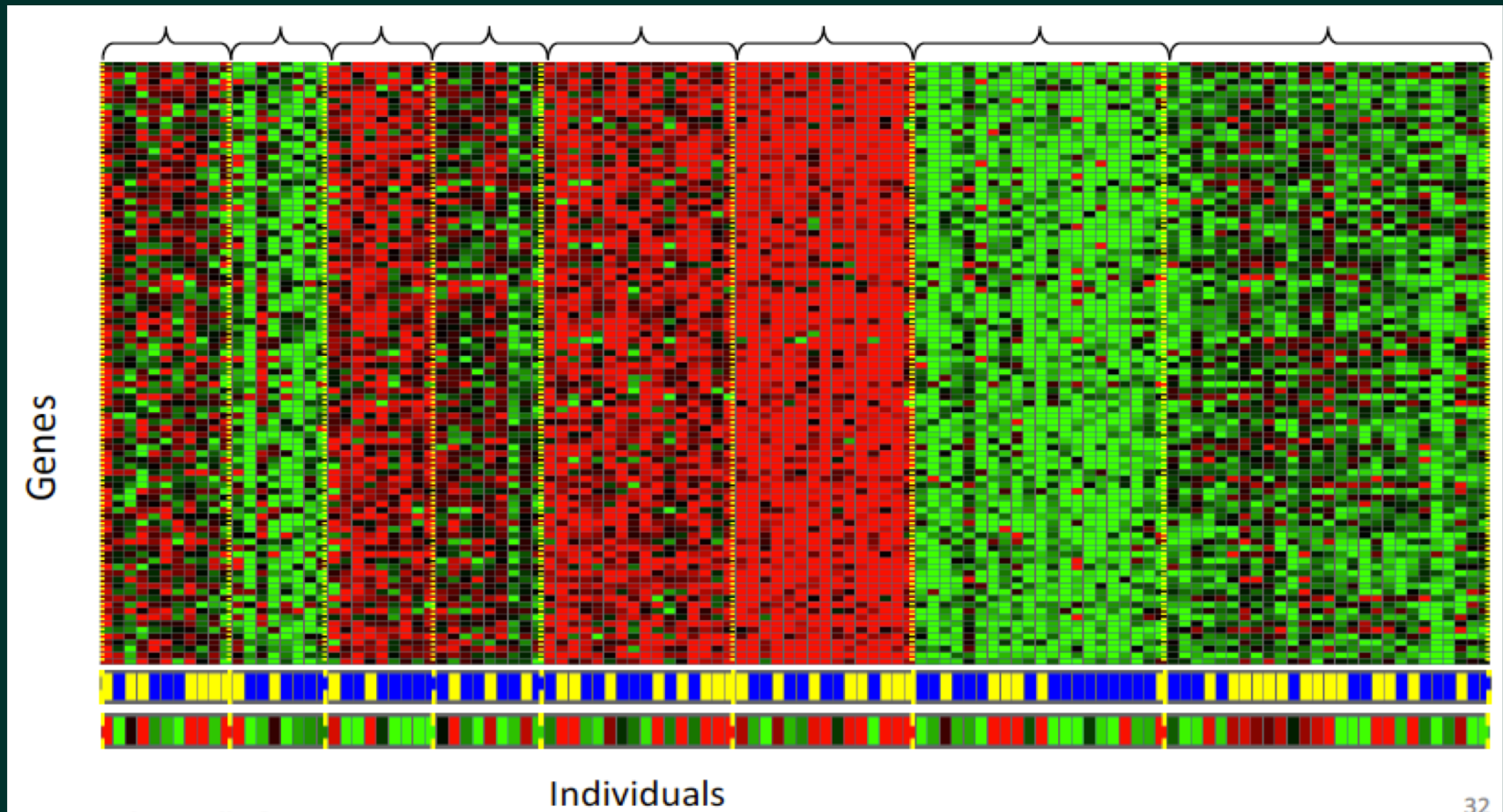
Market segmentation



Astronomical data analysis

مثال

- گروه‌بندی افراد براساس تشابه ژنتیک



الگوریتم‌ها (روش‌ها)ی معروف یادگیری بدون نظارت

خوشه‌بندی

کاهش ابعاد

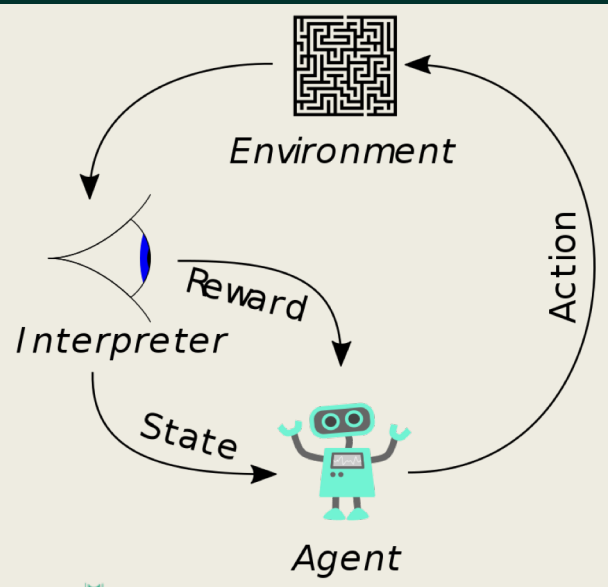
- خوشه‌بندی K-میانه‌ها (K-Means Clustering)
- خوشه‌بندی سلسله‌مراتبی (Hierarchical Clustering)
- DBSCAN
- Principle Component Analysis
- Linear Discriminant Analysis

یادگیری تقویتی

REINFORCEMENT LEARNING

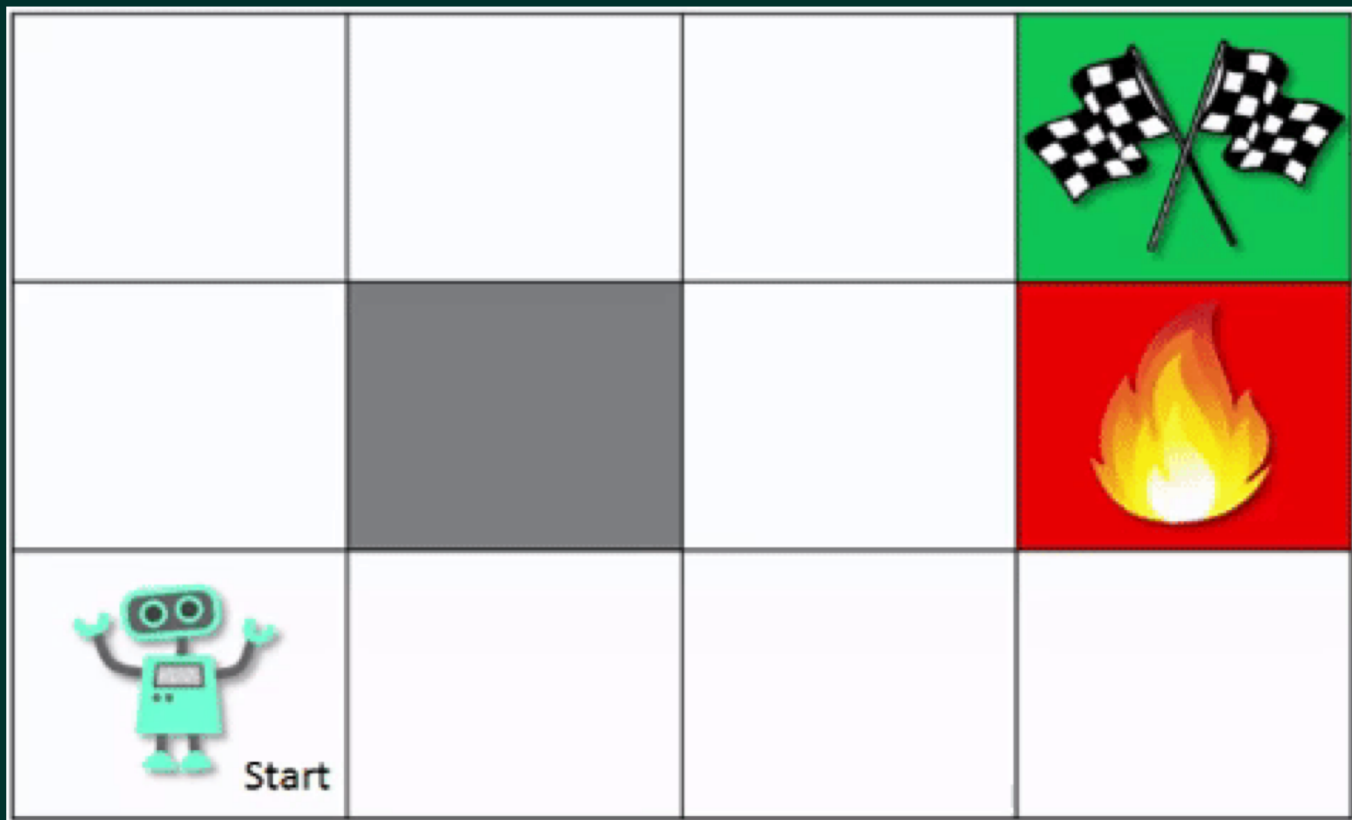
ویژگی‌های یادگیری تقویتی

- **ورودی:** مجموعه‌ای از حالات (states)، اعمال (actions) انجام شده در آن حالت، و پاداشی (reward) که آن عمل دریافت کرده است.
- **خروجی:** سیاست (Policy) بهینه برای رسیدن از یک حالت شروع به حالت بهینه



- سیاست نگاشتی از حالات به اعمال است
- بیان می‌کند که در هر حالت چه عملی باید انجام شود.
- مثال: بازی کامپیوتری، ربات جاروبرقی،

یادگیری تقویتی



یادگیری خودنظارتی

SELF-SUPERVISED LEARNING

ویژگی‌های یادگیری خودنظارتی

- نیازی به داده‌های برچسب‌گذاری شده توسط انسان وجود ندارد.
- تمامی داده‌های دردسترس بدون برچسب هستند.
- فرایند یادگیری معمولا ۲ مرحله‌ای است:

1. تولید برچسب مورد نیاز از داده‌های بدون برچسب

2. آموزش مدل با استفاده از داده برچسب‌دار تولید شده

مثال

Masked Language Modeling (MLM)

A quick brown fox jumps over the lazy dog

OH SNAP!



A quick brown ____ jumps over the ____ dog

Next Sentence Prediction (NSP)

Sentence 1

Sentence 2

Next Sentence?

I am going outside.

I will be back after 6.

YES

I am going outside.

You know nothing John snow.

NO

فرآیند یادگیری

- آموزش (Training)
 - مشاهده‌ی داده‌های آموزشی
 - مدل‌سازی، استخراج قوانین و الگوها
- تعمیم (Generalization)، آزمون (Test)
 - اعمال قوانین یادگرفته‌شده بر روی نمونه‌های جدید
- ارزیابی عملکرد (Performance Evaluation)

ادبیات

مجموعه‌های آموزش، اعتبارسنجی، آزمون

- مجموعه آموزشی (Training Set)

– زیر مجموعه‌ای از داده‌ها که برای آموزش مدل از آن استفاده می‌شود.

- مجموعه اعتبارسنجی (Validation Set)

– زیر مجموعه‌ای از داده‌ها که برای تنظیم پارامترهای مدل استفاده می‌شود.

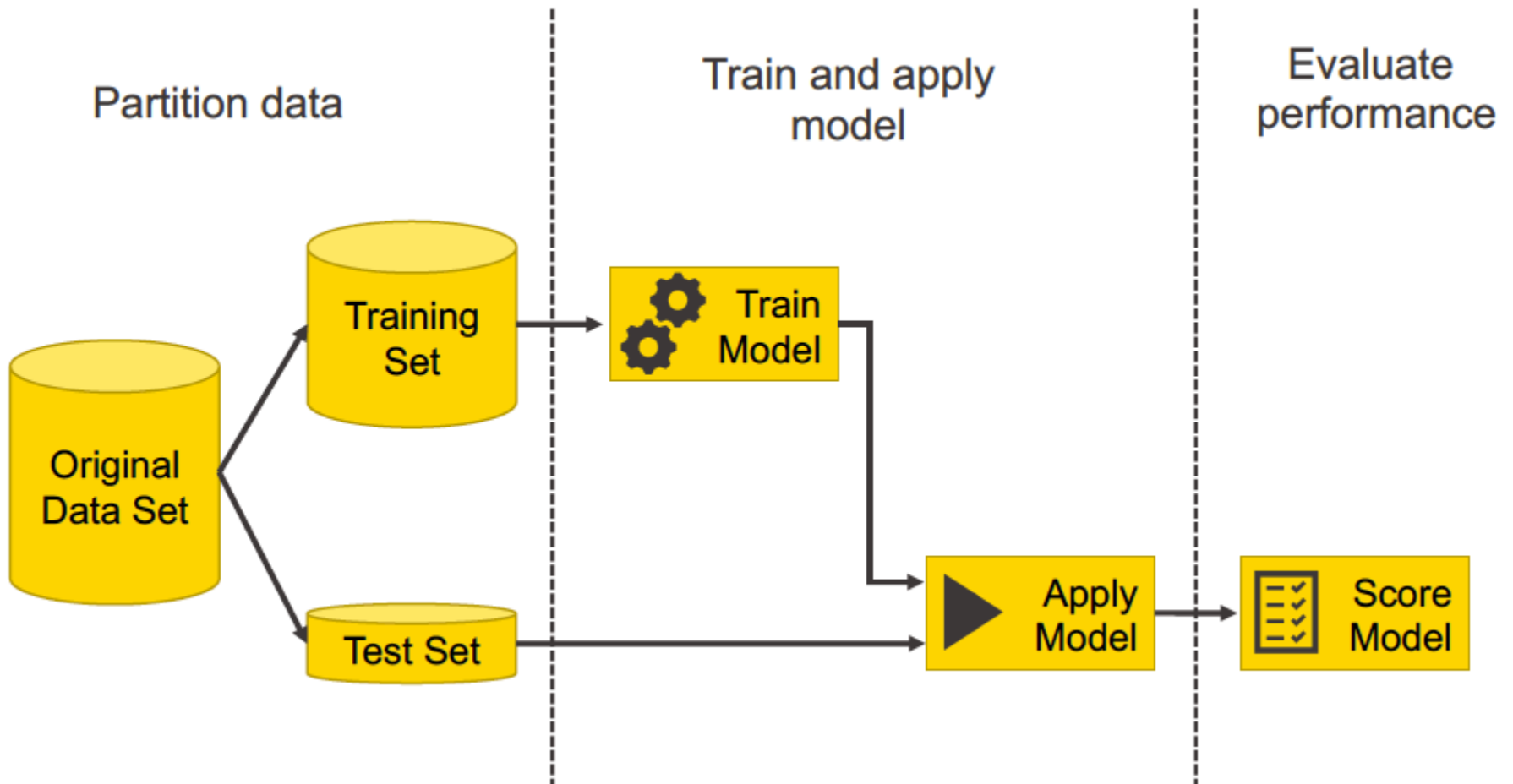
- مجموعه آزمون (Test Set)

– زیر مجموعه‌ای از داده‌ها مستقل که قبلاً توسط مدل دیده نشده‌اند و برای ارزیابی مدل به کار می‌روند.

- **نسبت:** کل داده‌ها معمولاً به نسبت ۲۰/۱۰/۷۰ و یا ۱۰/۱۰/۸۰ بین مجموعه‌های آموزش، اعتبارسنجی و آزمایش تقسیم می‌شود.



مراحل فرآیند یادگیری



ارزیابی عملکرد

PERFORMANCE EVALUATION

معیارهای ارزیابی

- **دقت (Accuracy):** درصد درست پیش‌بینی ($\text{Accuracy} = 1 - \text{Error}$)
- Precision and Recall
- F-Score
- Squared Error
- AUC
- Likelihood
- Posterior probability
- Cost / Utility
- Margin
- Entropy
- K-L divergence
-

ادبیات

خطای آموزش، خطای آزمون

- خطای آموزش (Training Error)

– خطا در تشخیص نمونه‌ای از مجموعه آموزش که قبلاً دیده شده است.

- مجموعه آزمون (Test Error)

– خطا در تشخیص نمونه‌ای از مجموعه آزمایش که تا کنون دیده شده است.



Getafe

Training



Cádiz

Test



~~Getafe~~

Training Set Error

~~Getafe~~

Test Set Error



یادگیری ماشین با پایتون

یادگیری نمونه‌محور در مقابل مدل‌محور

یادگیری نمونه‌محور (Instance-based Learning)

- الگوریتم بدون ساختن مدل و تنها با استفاده از تمامی داده‌های آموزشی اقدام به تصمیم‌گیری در مورد نمونه‌های جدید می‌کند
- به این‌گونه الگوریتم‌ها حافظه‌محور (Memory-based) گفته می‌شود.
- **k-nearest neighbors algorithm, kernel machines and RBF networks**

یادگیری مدل‌محور (Model-based Learning)

- الگوریتم ابتدا یک مدل از روی داده‌ی آموزشی می‌سازد
- سپس از مدل ساخته‌شده برای پیش‌بینی/رده‌بندی نمونه‌های جدید استفاده می‌کند
- **SVM, Decision Trees and Logistic Regression**

کاربردهای یادگیری ماشین

There are many applications for machine learning, including:

- Agriculture
- Anatomy
- Adaptive website
- Affective computing
- Astronomy
- Automated decision-making
- Banking
- Bioinformatics
- Brain-machine interfaces
- Cheminformatics
- Citizen science
- Climate science
- Computer networks
- Computer vision
- Credit-card fraud detection
- Data quality
- DNA sequence classification
- Economics
- Financial market analysis^[76]
- General game playing
- Handwriting recognition
- Information retrieval
- Insurance
- Internet fraud detection
- Knowledge graph embedding
- Linguistics
- Machine learning control
- Machine perception
- Machine translation
- Marketing
- Medical diagnosis
- Natural language processing
- Natural language understanding
- Online advertising
- Optimization
- Recommender systems
- Robot locomotion
- Search engines
- Sentiment analysis
- Sequence mining
- Software engineering
- Speech recognition
- Structural health monitoring
- Syntactic pattern recognition
- Telecommunication
- Theorem proving
- Time-series forecasting
- User behavior analytics
- Behaviorism

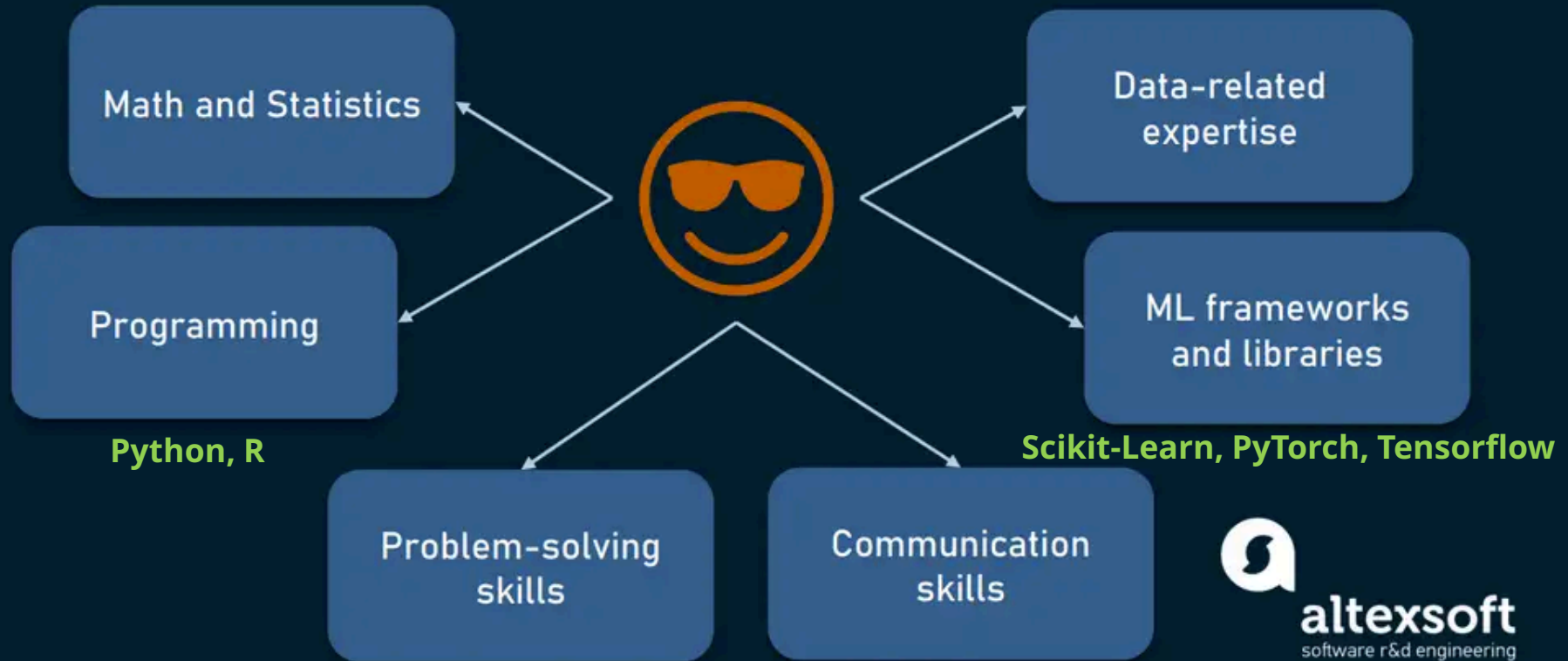
en.wikipedia.org/wiki/Machine_learning#Applications

تمرین: در زمینه تخصصی کار شما، یادگیری ماشین کجا می‌تواند استفاده شود؟

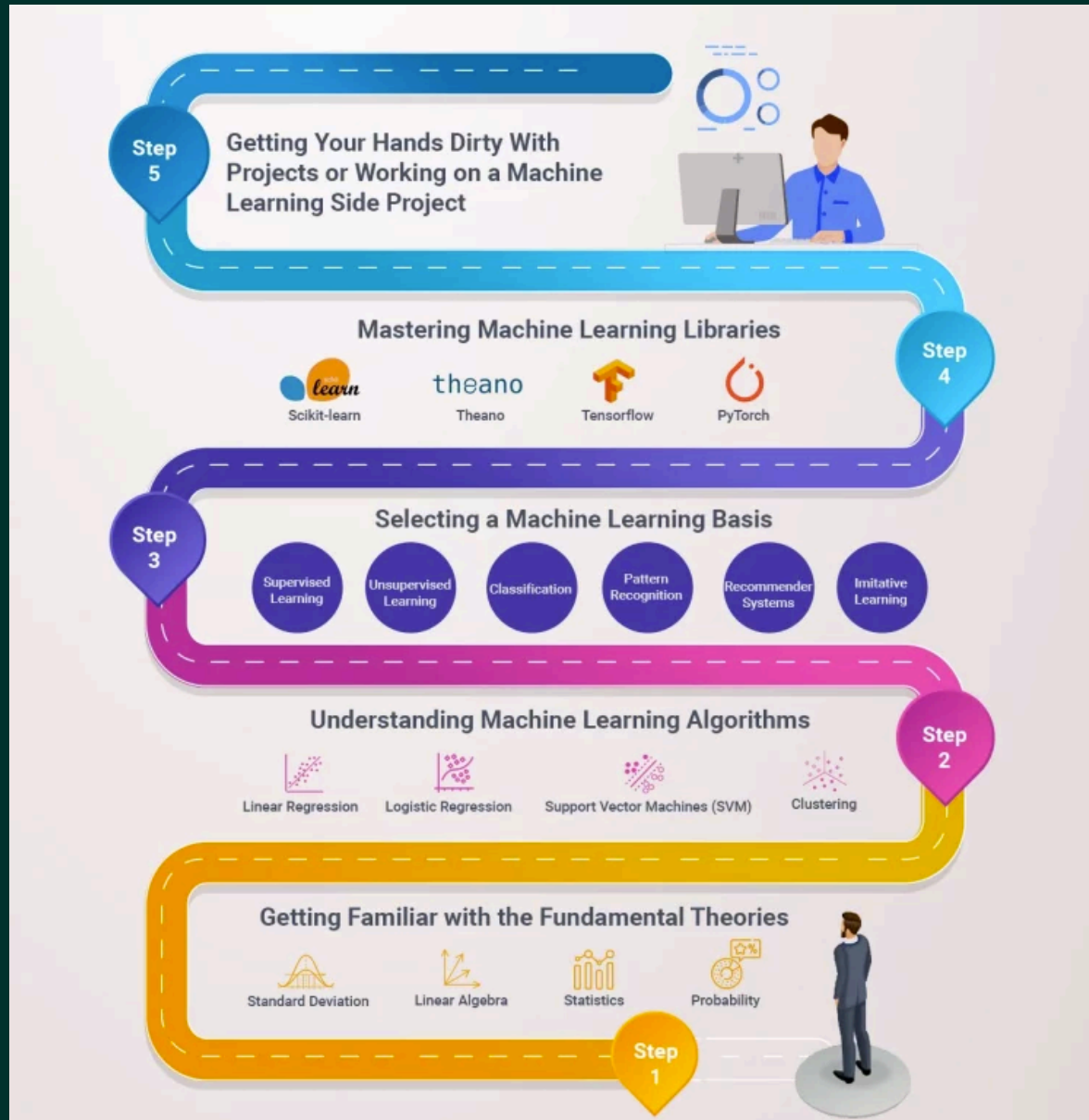
مهارت‌های لازم

Calculus, Linear Algebra, Statistics

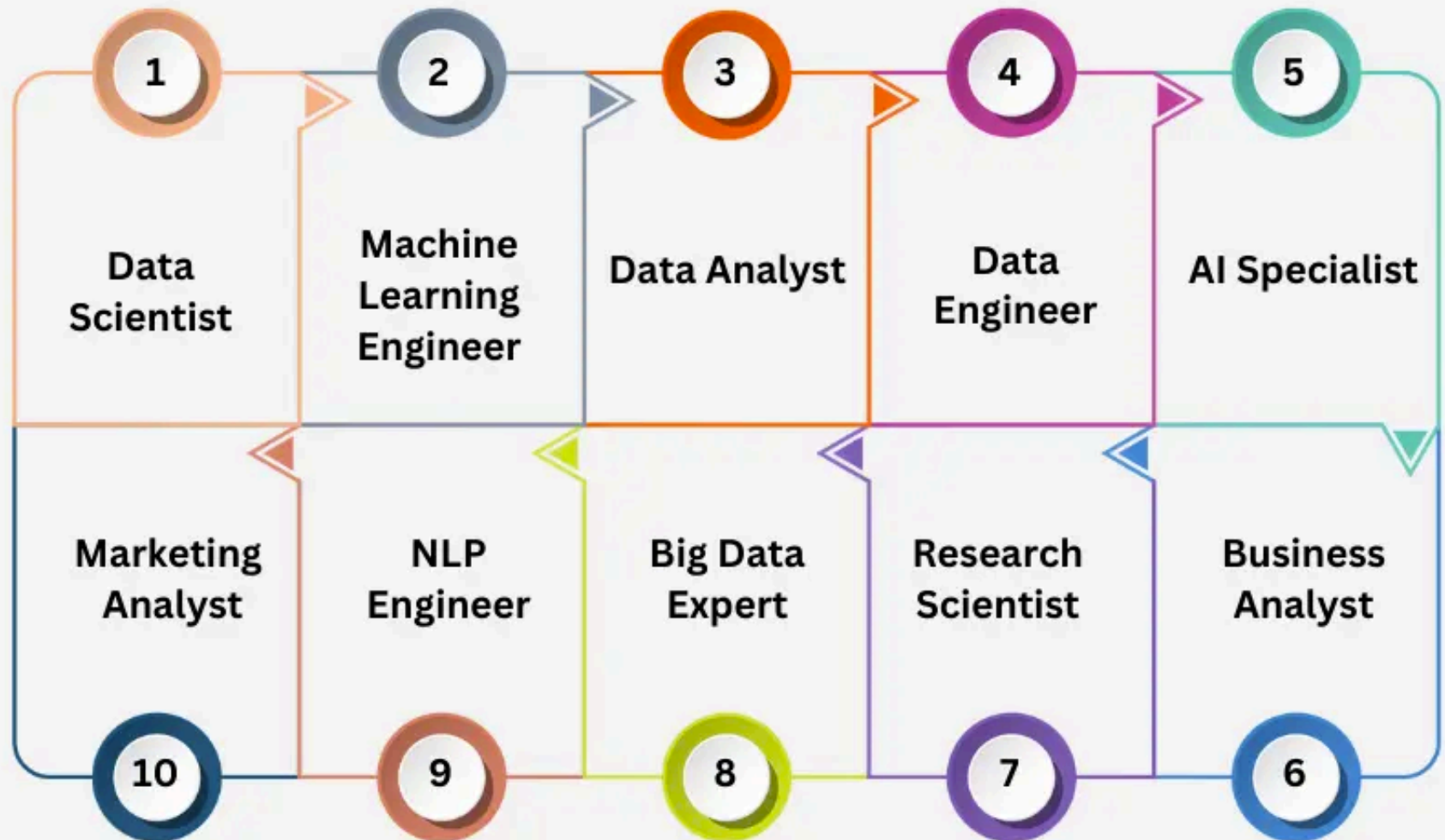
Pandas, NumPy, Matplotlib, Seaborn



مسیر یادگیری

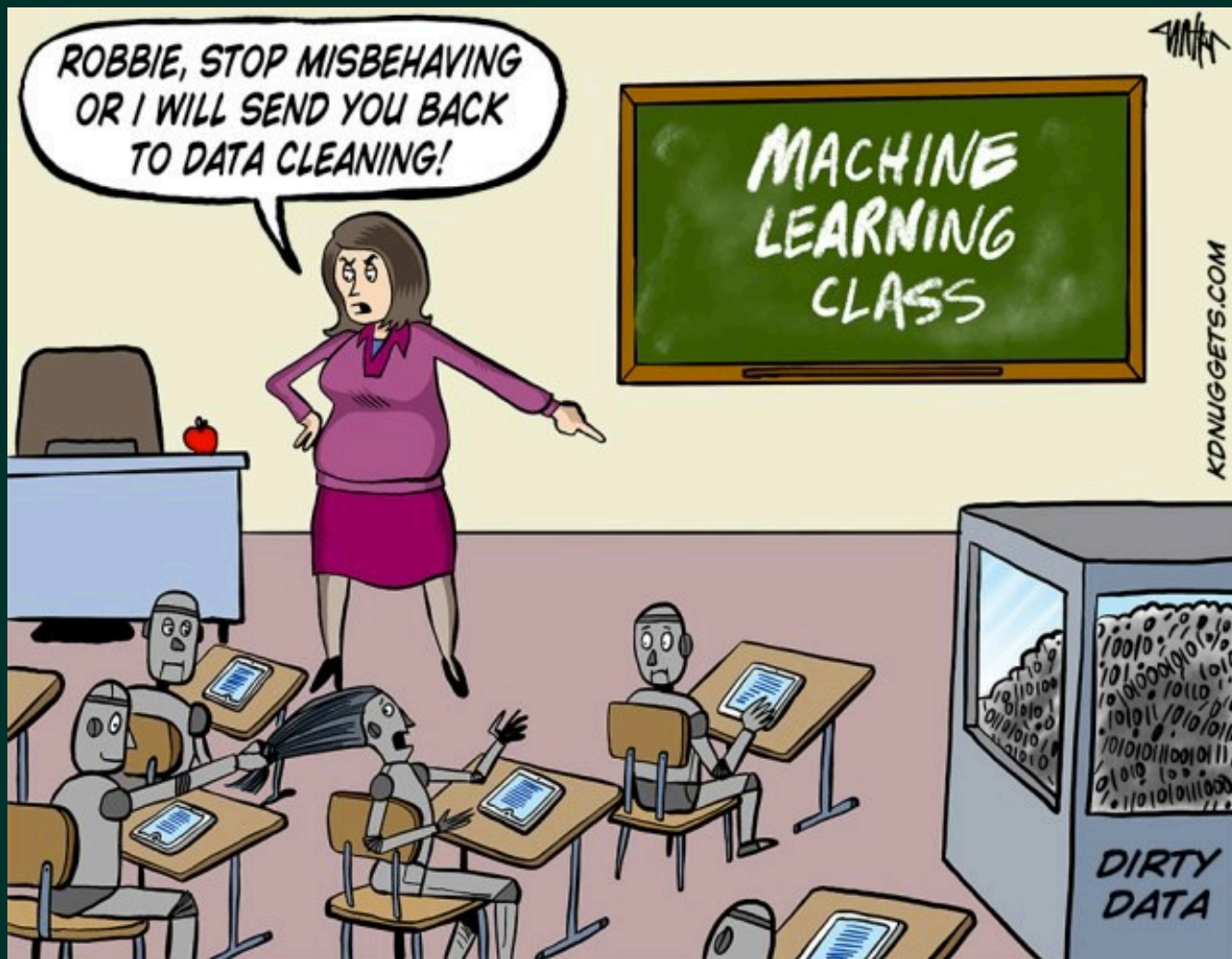


عناوین شغلی



SPRINTZEAL
Empowering Engineers





سوال ؟