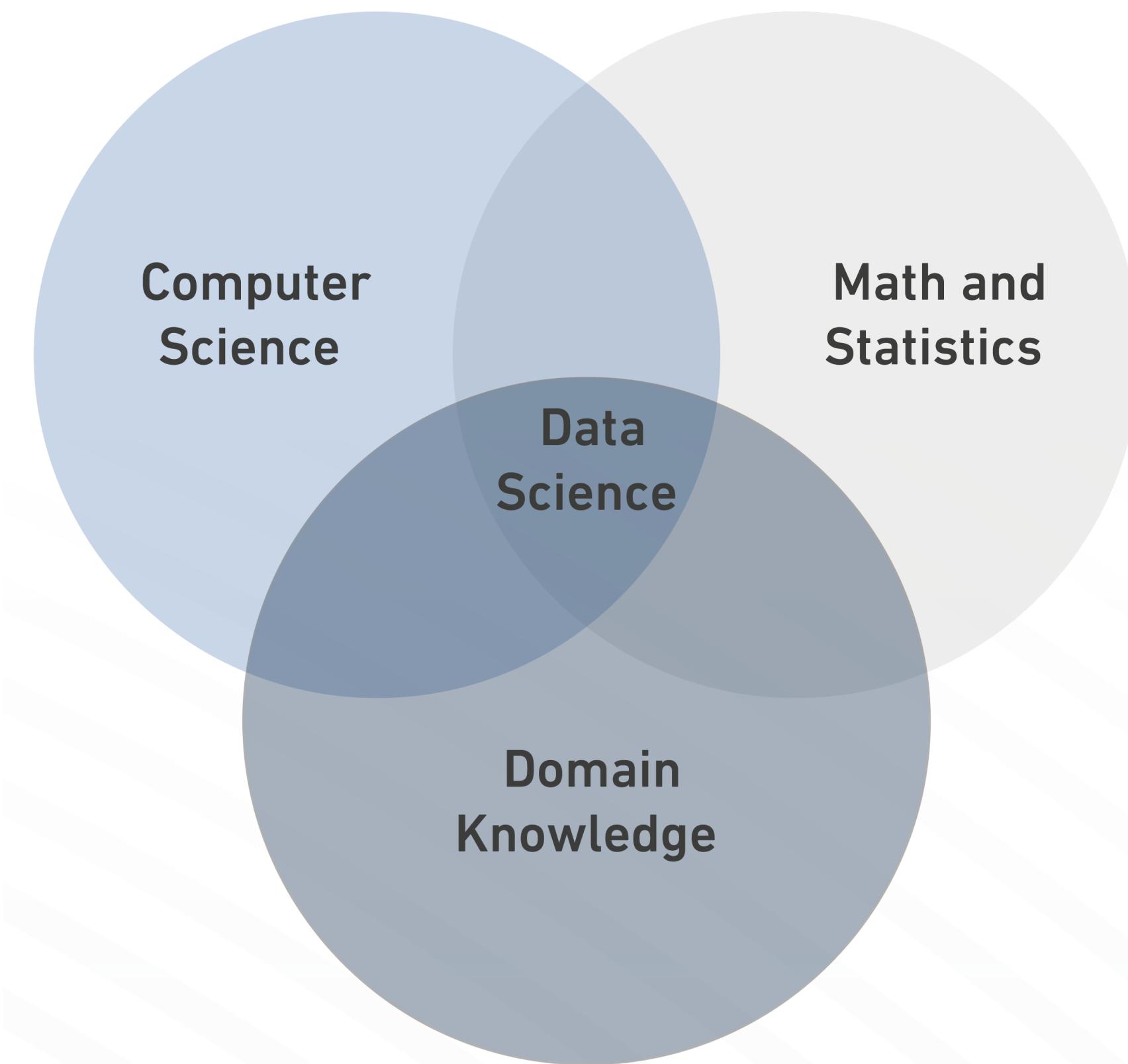




مقدمة في تعلم الآلة

ما هو علم البيانات؟

ما هو علم البيانات

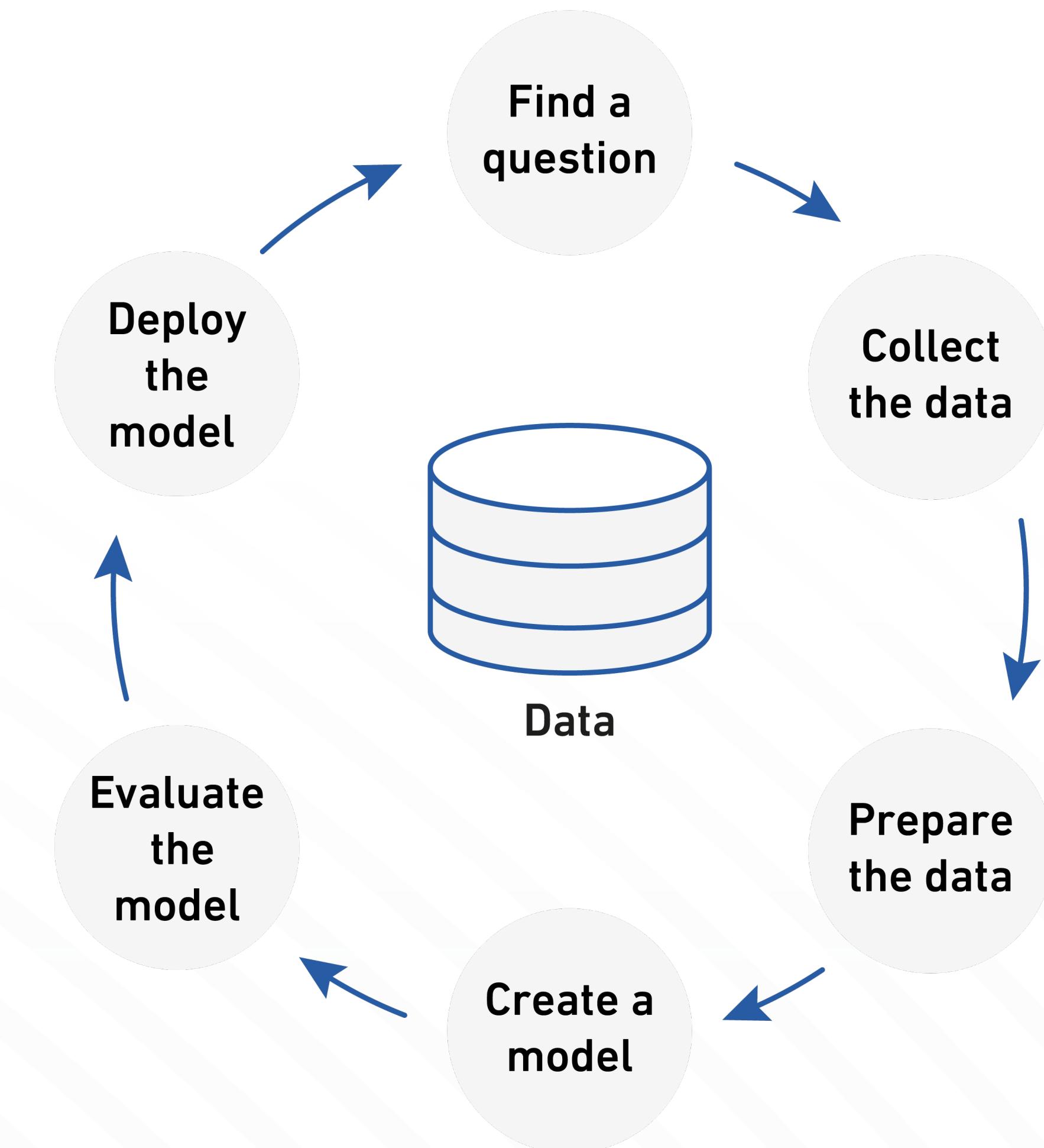


- هو علم يقوم بتطبيق الأساليب الإحصائية على البيانات بغرض الوصول لحل مشكلة معينة، حيث يتم الاستفادة من البيانات التي يتم إنتاجها بشكل يومي من عدة مصادر مثل وسائل التواصل الاجتماعي وغيرها بغرض الوصول لقرارات أو استنتاجات تقوم بحل مشكلة ما.
- ويمكن تعريفه أيضاً على أنه علم قائم على التقاطع بين عدة علوم وهي علم الإحصاء وعلوم الكمبيوتر و(Domain Knowledge) والمقصود به المجال التابع للمشكلة التي نقوم بحلها وهو يختلف بحسب نوع البيانات فمثلاً: بيانات المرضى تتبع المجال الصحي وبيانات الأسهم تتبع المجال المالي وهكذا.
- يهدف علم البيانات لتحويل data إلى Knowledge تساعدنا في اتخاذ القرارات.

تطبيقات علم البيانات

- التنبؤ بدرجات الحرارة للأسبوع القادم.
- التنبؤ باحتمالية إصابة شخص بمرض معين.
- معرفة آراء الناس حول حدث معين.

المراحل الأساسية في علم البيانات - Data Science LifeCycle



المراحل الأساسية في علم البيانات - Data Science LifeCycle

أي مشكلة نقوم بحلها عن طريق علم البيانات بشكل عام تمر بعدة خطوات:

أولاً: طرح التساؤلات وهذا قد يكون فرضية نريد اختبارها أو قرار نريد اتخاذه أو منتج نريد إنتاجه

ثانياً: جمع البيانات المتعلقة في المشكلة التي نريد حلها وهذا يعتمد على نوع المشكلة فأحياناً لانحتاج لجمع بيانات ويمكن أن نستخدم بيانات جاهزة

ثالثاً: تجهيز البيانات حتى يتم تحليلها عن طريق تنظيفها (Data Cleaning) وعمل تحويل للبيانات (Data Transforming) إلى شكل يناسب عمل التحليل.

رابعاً: بناء النماذج للبيانات (Statistical Model) أو (Visual Model) أو (Numerical Model) وهناك أشكال مختلفة للنماذج مثل: (Data Modeling) ونقوم باستخدامها لإثبات فرضية معينة أو التنبؤ بنتيجة معينة.

خامساً: تقييم النموذج و في هذه المرحلة نحتاج للتأكد من النموذج هل قام بالإجابة عن التساؤلات التي طرحناها بشكل دقيق أم لا، هل ساعد في اتخاذ القرارات أو التنبؤ بنتائج معينة.

سادساً: نشر النموذج وهذه الخطوة تكون بعد التأكد من دقة النموذج، ثم بعد عملية النشر يمكن استخدام النموذج على أنواع أخرى من البيانات.

أخيراً، Data Science LifeCycle تعتبر iterative Process حيث نقوم بتكرار هذه العملية في كل مرحلة نطرح سؤال معين أو نحاول اتخاذ قرار لحل مشكلة ما وفي كل مرحلة نحسن العملية حتى نصل لنتائج دقيقة، أيضاً هي تعتبر غير متسلسلة Non-Sequential حيث نقوم بالتقدم لخطوات forward أو التراجع لخطوات backward بناء على النتائج التي نحصل عليها.

أشهر المكتبات في علم البيانات



pandas مكتبة

يتم استخدامها لتحليل البيانات (Data Analysis) و هيكلتها (Data Structures).



NumPy مكتبة

تقوم بتوفير إمكانية التعامل مع المصفوفات (Multidimensional Arrays) والعمليات الرياضية (Linear Algebra Functions).



matplotlib مكتبة

يتم استخدامها لتمثيل أو عرض البيانات (Data Visualization) بشكل رسومات بيانية.

أشهر المكتبات في علم البيانات



مكتبة Seaborn

هي مكتبة بُنيت فوق مكتبة matplotlib لتمثيل البيانات بطريقة متقدمة على شكل رسومات بيانية تفاعلية.



مكتبة sikit-learn

تعتبر أحد المكتبات الخاصة بتنفيذ خوارزميات تعلم الآلة (Machine Learning Algorithms).



مكتبة PyTorch

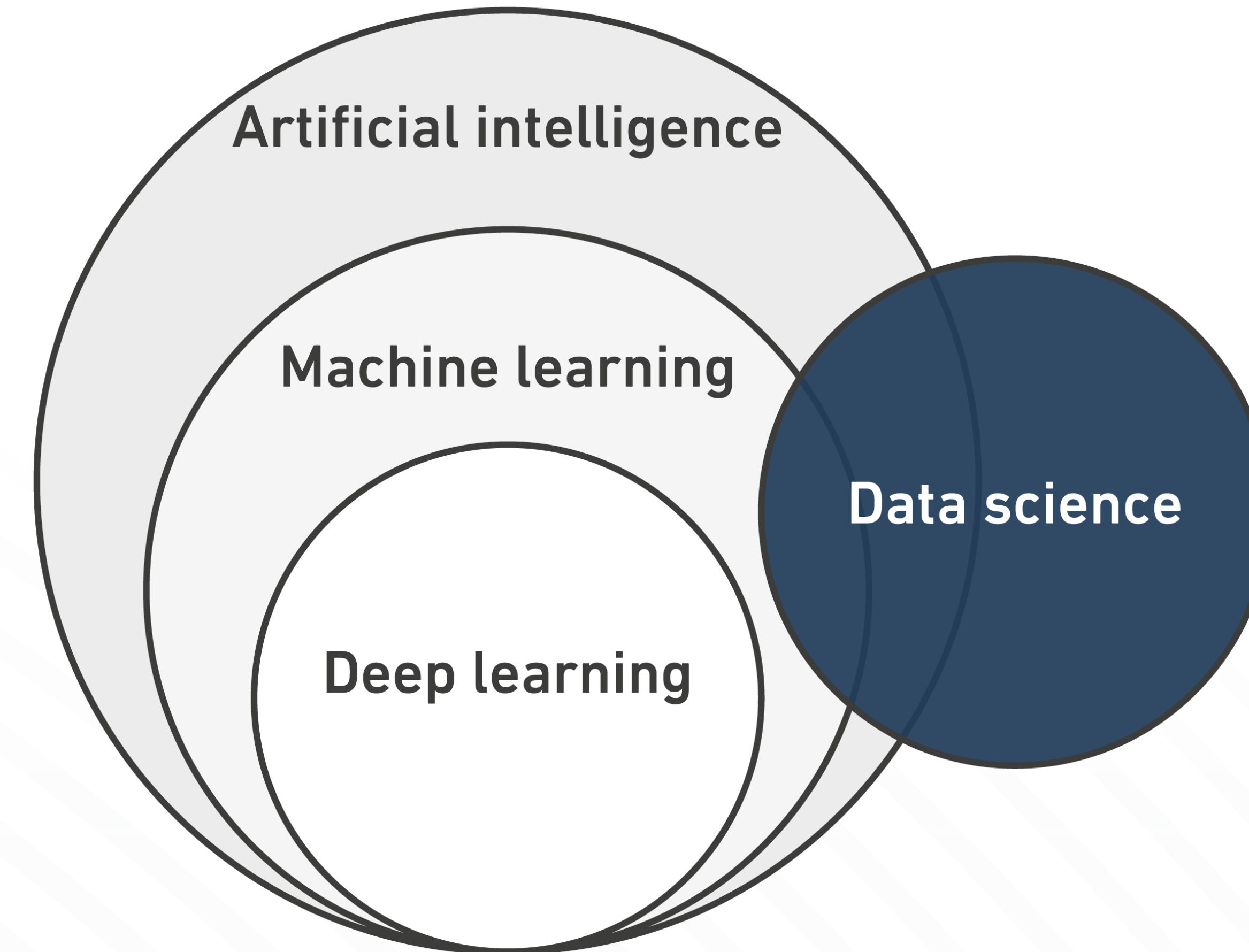
تعتبر أحد المكتبات الخاصة بتنفيذ خوارزميات تعلم الآلة وتستخدم في (Natural Language Processing) و (Computer Vision).



مكتبة TensorFlow

تعتبر أحد المكتبات الخاصة بتنفيذ خوارزميات الذكاء الاصطناعي و تعلم الآلة.

الفرق بين علم البيانات (Data Science) و الذكاء الاصناعي (AI)
و تعلم الآلة (Machine Learning) و التعلم العميق (Deep Learning)



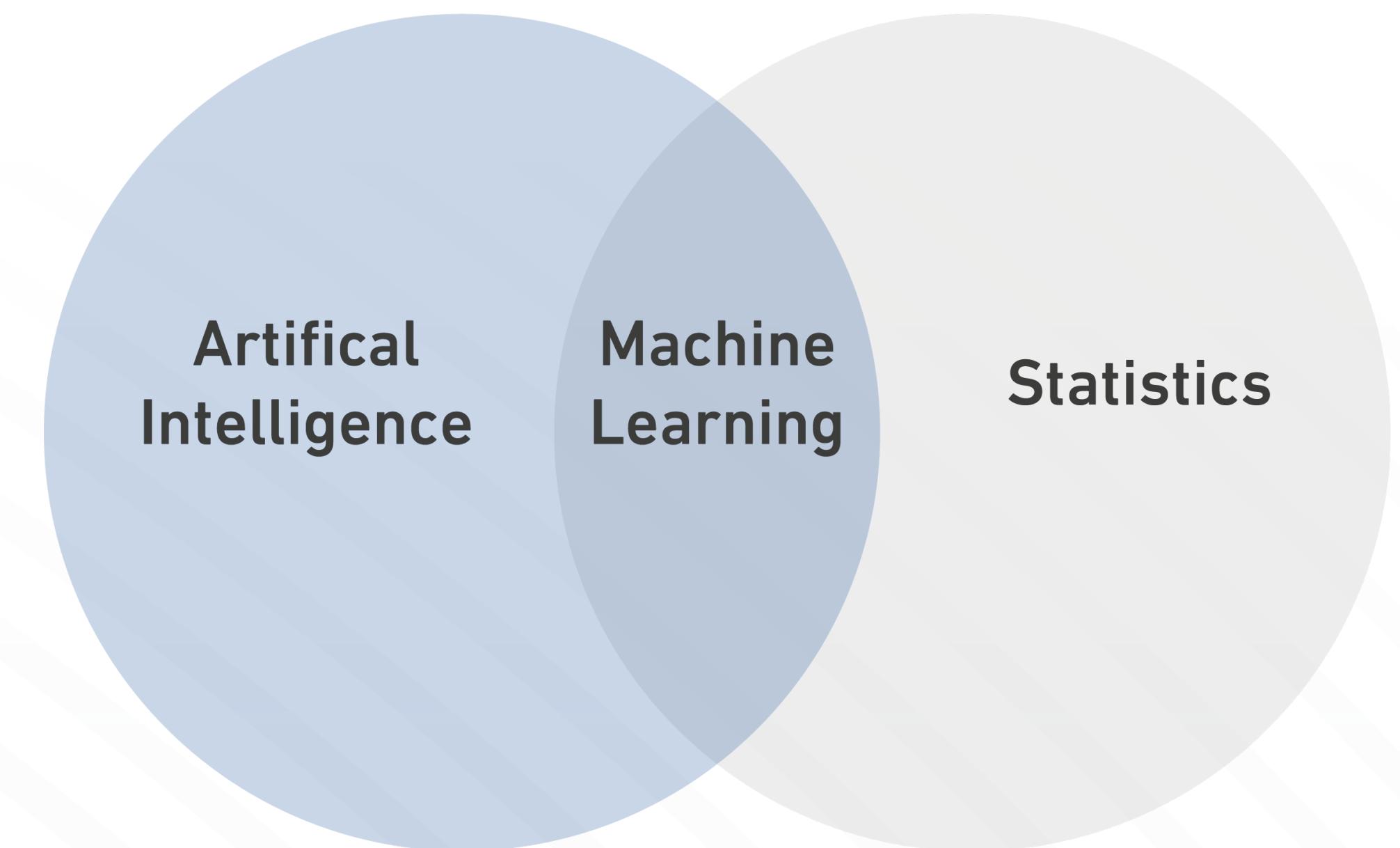
الفرق بين علم البيانات (Data Science) و الذكاء الاصناعي (AI) و تعلم الآلة (Machine Learning) و التعلم العميق (Deep Learning)

الكثير من الأشخاص وخصوصاً المبتدئين يجدون صعوبة في التفريق بين علم البيانات والذكاء الاصناعي وتعلم الآلة و التعلم العميق وربما يعتقدون أن جميع هذه المصطلحات تشير لنفس المفهوم ولكن في الحقيقة يوجد بعض الاختلافات حيث أن:

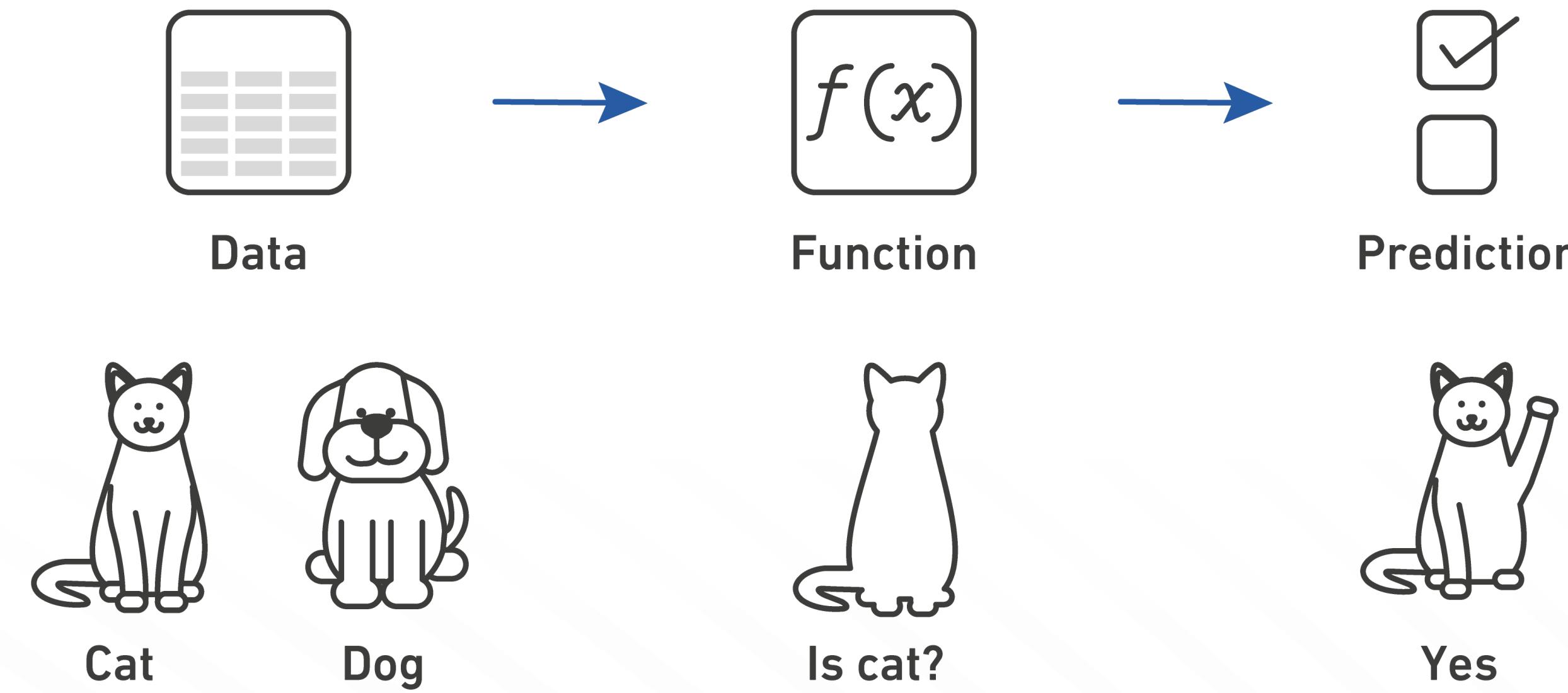
- علم البيانات يقوم باستخدام تقنيات (AI) و (Machine Learning) و تطبيقها على البيانات بغرض الوصول لقرارات أو استنتاجات معينة.
- أما التعلم العميق (Deep Learning) فهو جزء فرعي من تعلم الآلة (Machine Learning) الذي هو أيضاً جزء فرعي من الذكاء الاصناعي (AI) وجميعها تشير للتقنيات التي تساعد الكمبيوتر على التعلم من البيانات لحل المشاكل المعقدة.

تعلم الآلة (Machine Learning)

- هو عملية تعلم الآلة إنجاز مهمة معينة دون كتابة كود صريح أو أوامر صريحة لتنفيذ هذا الأمر، ويمكن التعبير عنه بأنه جزء الذكاء الاصطناعي الذي يحتوي على احصائيات.
- أثناء عملية تعلم الآلة يقوم بإنشاء ما يسمى بالنموذج (Model) الذي يقوم بتزويد مجموعات البيانات و الخوارزمية (algorithm) للتعلم من البيانات.

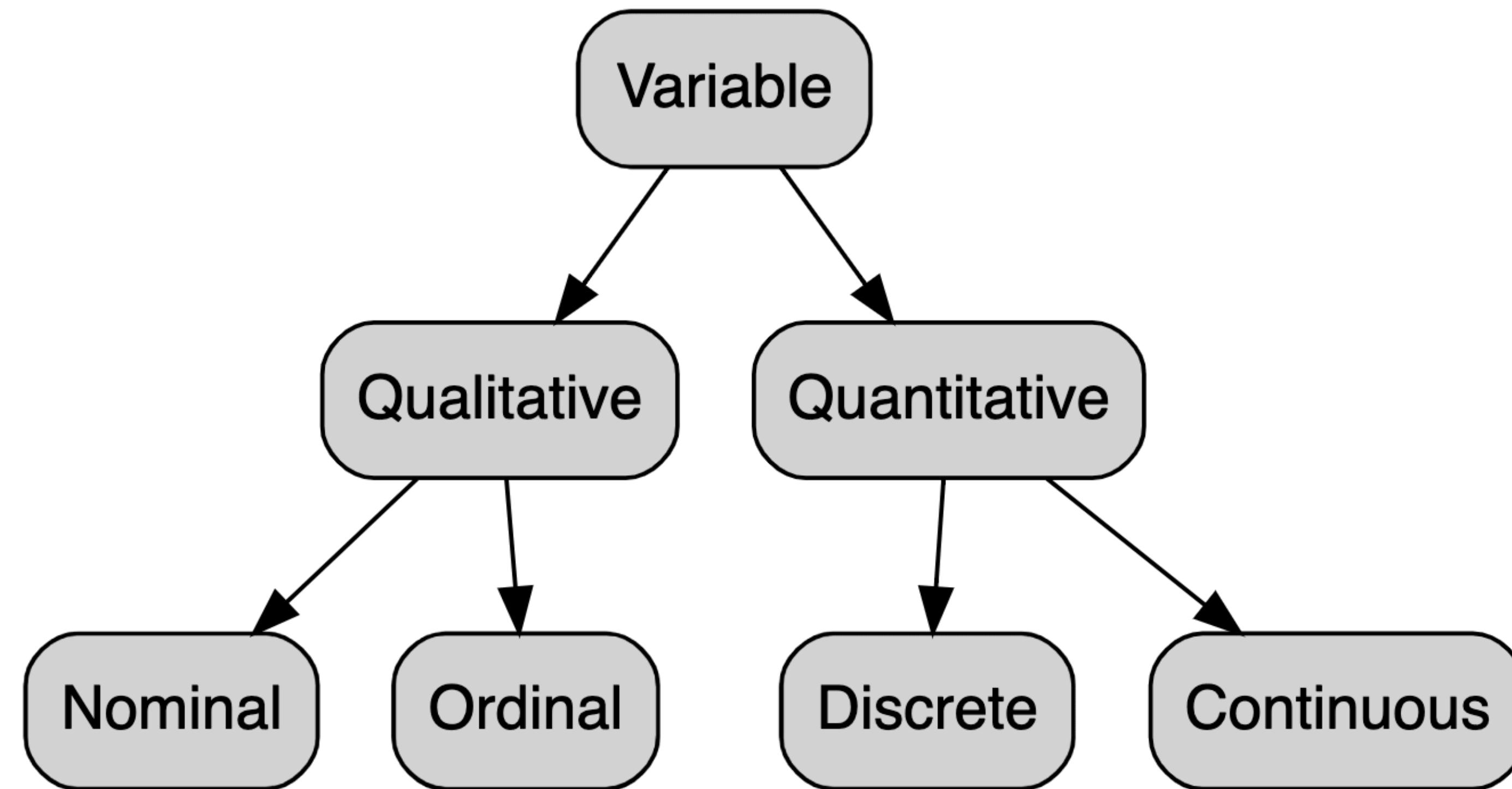


تعلم الآلة (Machine Learning)



- نقوم باستعمال البيانات لتعلم بناء دالة تكون قادرة على التنبؤ بنتيجة البيانات الجديدة على سبيل المثال لنقل أننا نريد بناء دالة تقوم بتحديد هل الصورة تحتوى على قطة أم لا؟
- في البداية سوف نقوم بإنشاء بيانات تحتوى صور للقطط وصور لا تحتوى ذلك.
- ثم نقوم بتطبيق خوارزميات تعلم الآلة على مجموعة البيانات.
- نقوم بهذه الخوارزميات بتعلم الدالة التي تتنبأ بالصور هل الصورة تحتوى على قطة أم لا؟

أنواع البيانات



(Machine Learning Algorithms) تعلم الآلة خوارزميات

Supervised Learning

Labeled Data
Direct Feedback
Classification and Regression

Unsupervised Learning

Unlabeled Data
No Feedback
Clustering & Dimensionality Reduction

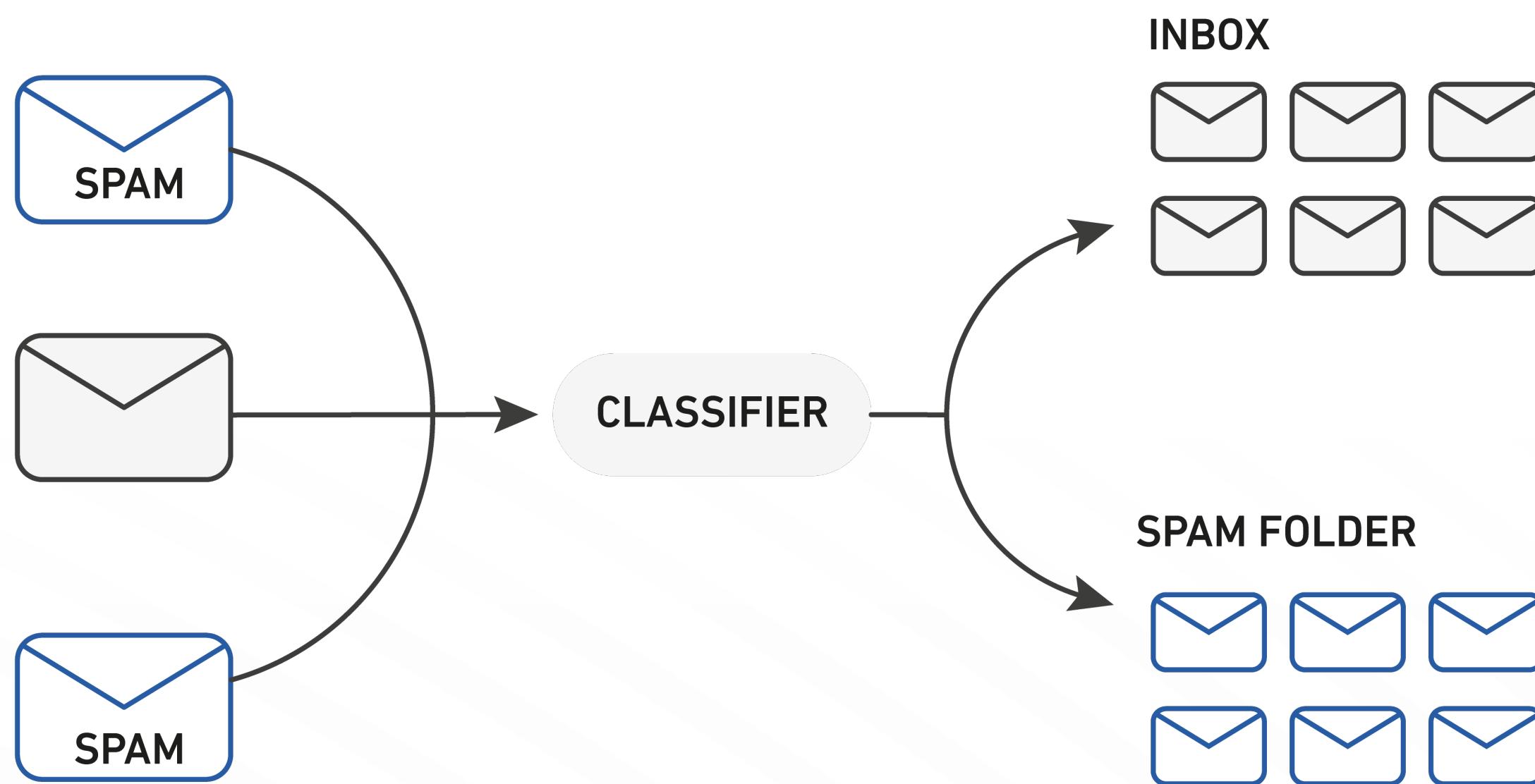
Semi-supervised Learning

Labeled and Unlabeled Data
Some Feedback
Classification and Regression

Reinforcement Learning

Reward Based Learning
Direct Feedback
Learn series of actions

خوارزميات تعلم الآلة (Machine Learning Algorithms)



النوع الأول (Supervised Learning)

يشير هذا النوع إلى العملية التي تتعلم فيها الآلات من مجموعات البيانات المصنفة أو المعروفة (labeled data) وتنتج نموذجاً دقيقاً قادرًا على التنبؤ بملخصات البيانات غير المرئية.

مثلاً: عندما نقوم بتدريب الآلة على تصنيف البريد إلى بريد مزعج (spam) أو غير مزعج (not-spam) خلال هذه العملية تتعلم فيها الآلة من مجموعة من البيانات المصنفة إلى بريد مزعج و غير مزعج بهدف إنشاء نموذج (Model) قادر على تصنيف أي بريد جديد إلى بريد مزعج وغير مزعج.

خوارزميات تعلم الآلة (Machine Learning Algorithms)

عادةً ما تكون الخوارزميات المستخدمة في هذا النوع مصنفة لنوعين:

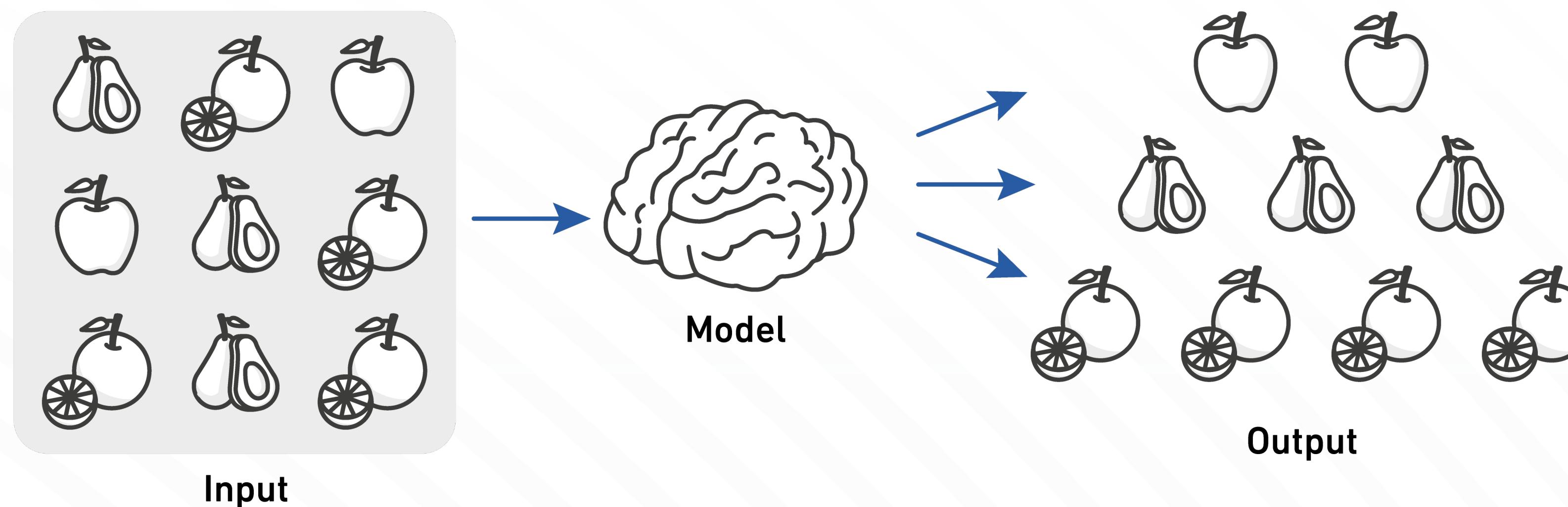
خوارزميات التصنيف (classification) مثل SVMs

خوارزميات الانحدار (Regression) مثل الانحدار الخطى (Linear Regression)

خوارزميات تعلم الآلة (Machine Learning Algorithms)

النوع الثاني (Unsupervised Learning)

يشير هذا النوع إلى العملية التي تتعلم فيها الآلات منمجموعات البيانات غير المصنفة بناءً على التشابه بين مجموعة البيانات. مثال: عندما نقوم بتدريب الآلة على تصنيف الفواكه، في هذه الحالة لا نخبر الآلة عن اسم الفاكهة، إذا كيف تقوم الآلة بالتصنيف؟ تقوم الآلة بالتصنيف بناءً على خصائص الفواكه مثل: اللون والحجم والشكل فمثلاً: إذا كان اللون أحمر يتم تصنيف الفاكهة إلى تفاح وهكذا. ومن أشهر الخوارزميات على هذا النوع: خوارزمية k-Means.



خوارزميات تعلم الآلة (Machine Learning Algorithms)

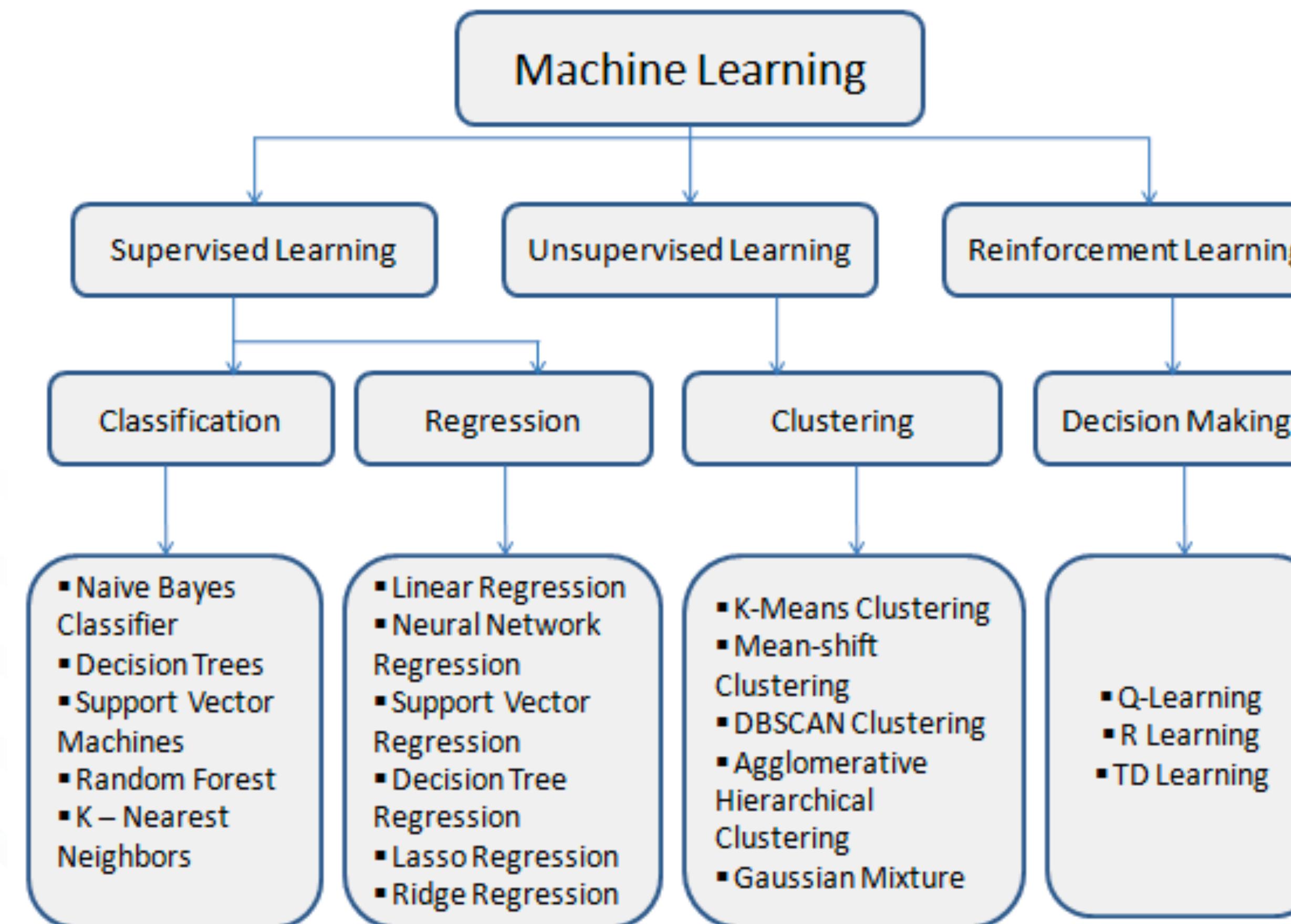
النوع الثالث (Semi-Supervised Learning)

هذا النوع من التعلم نحتاجه بالغالب عندما يكون لدينا مجموعات البيانات الكبيرة ولكن عدد قليل من هذه البيانات قد تم تصنيفها، لذا تقوم الآلة بالتعلم من كلا النوعين البيانات المصنفة والغير مصنفة.

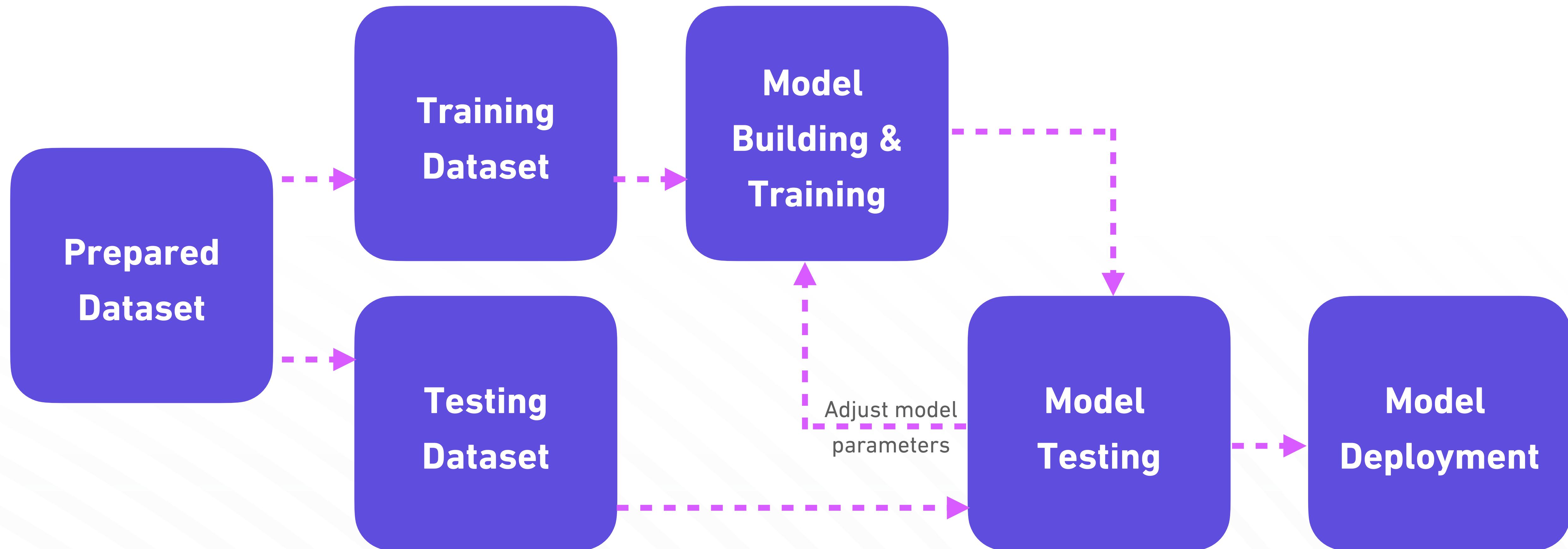
النوع الرابع (Reinforcement Learning)

يشير هذا النوع إلى العملية التي تتعلم فيها الآلات اتخاذ قرارات (actions) بناءً على البيئة الخارجية ثم مكافأة الآلة حتى نصل لتحقيق أقصى قدر من هدف معين.
مثال: بناء الرجل الآلي.

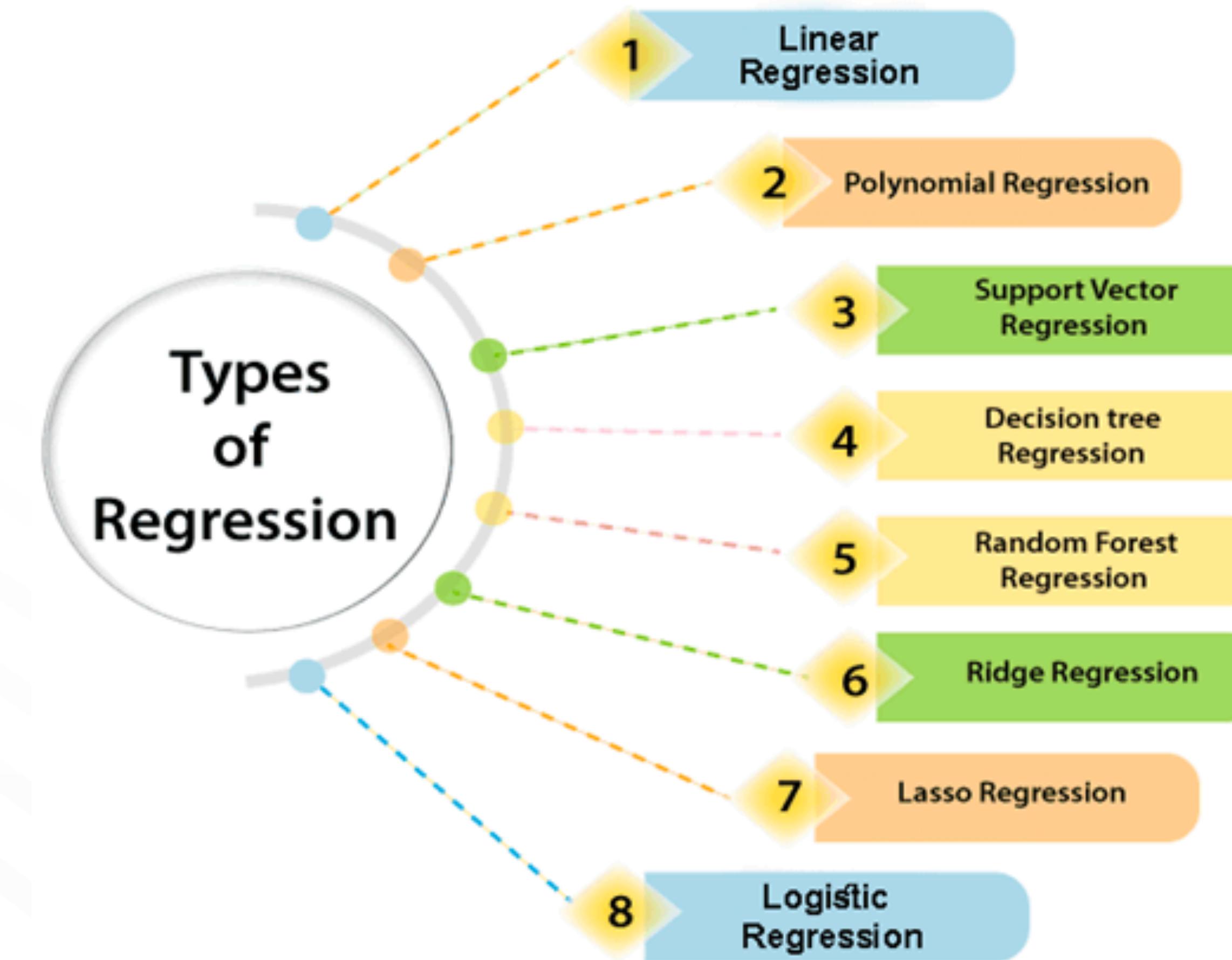
(Machine Learning Algorithms) خوارزميات تعلم الآلة



مراحل بناء نماذج تعلم الآلة (Machine Learning Models)



(Regression Algorithms) الخوارزميات الانحدار



خوارزميات الانحدار (Regression Algorithms)

مفهوم Regression Analysis
كيفية عمل تنبؤات حول الكميات الرقمية (quantities)

أمثلة

- كيف يتغير حجم المبيعات عند تغيير الأسعار؟
- كيف يتأثر حجم المبيعات بالطقس؟
- كيف يؤثر عنوان الكتاب على مبيعاته؟

خوارزميات الانحدار (Regression Algorithms)

مفهوم Regression Analysis

في جميع الأمثلة السابقة، يقوم بالبحث عن إجابة (**response**) يمكن التعبير عنها بمتغير **covariates or independent variables** (وتشتهر باسم **predictors**) أو عدّة متغيرات مستقلة

$$y = a_0 + a_1 x_1$$

intercept parameters of the model
(coefficients)

خوارزميات الانحدار (Regression Algorithms)

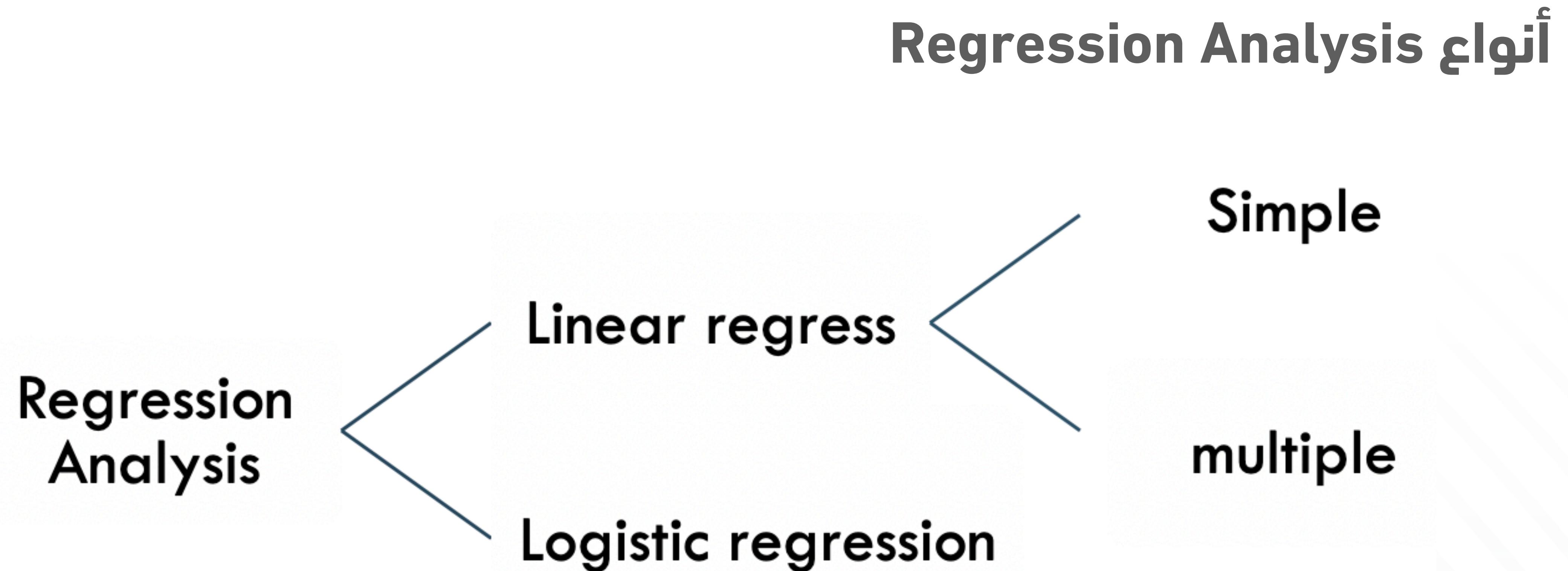
مفهوم Regression Analysis

يمكن القول أن Regression Analysis يشير للطريقة التي نبني بها نموذج (model) للتعبير عن العلاقة بين الإجابة (response) وهي تمثل y و مجموعة المتغيرات المستقلة (independent variables) وهي تمثل x_i

[Regression Role]

Build a model that can predict response from variables.

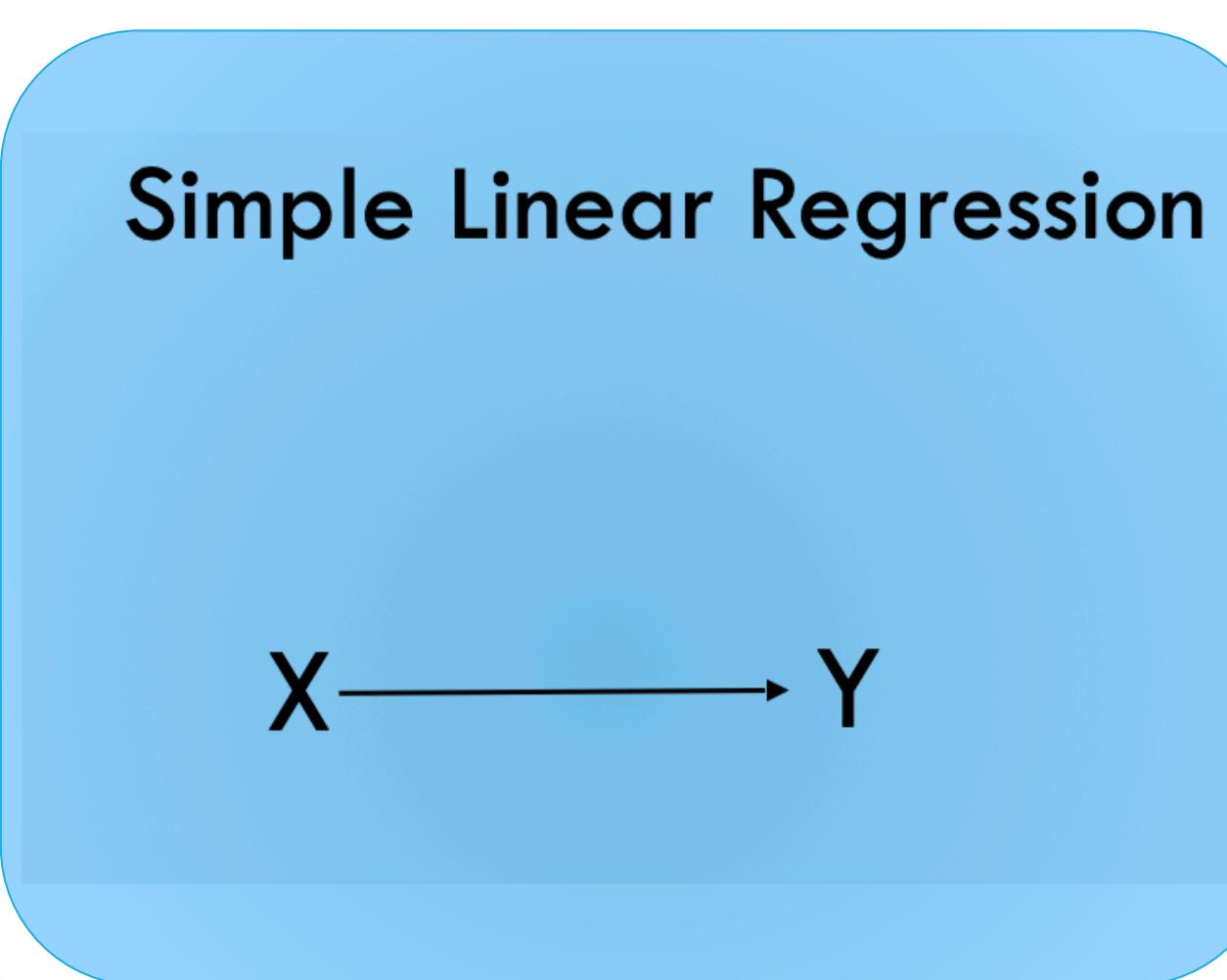
(Regression Algorithms) الخوارزميات الانحدار



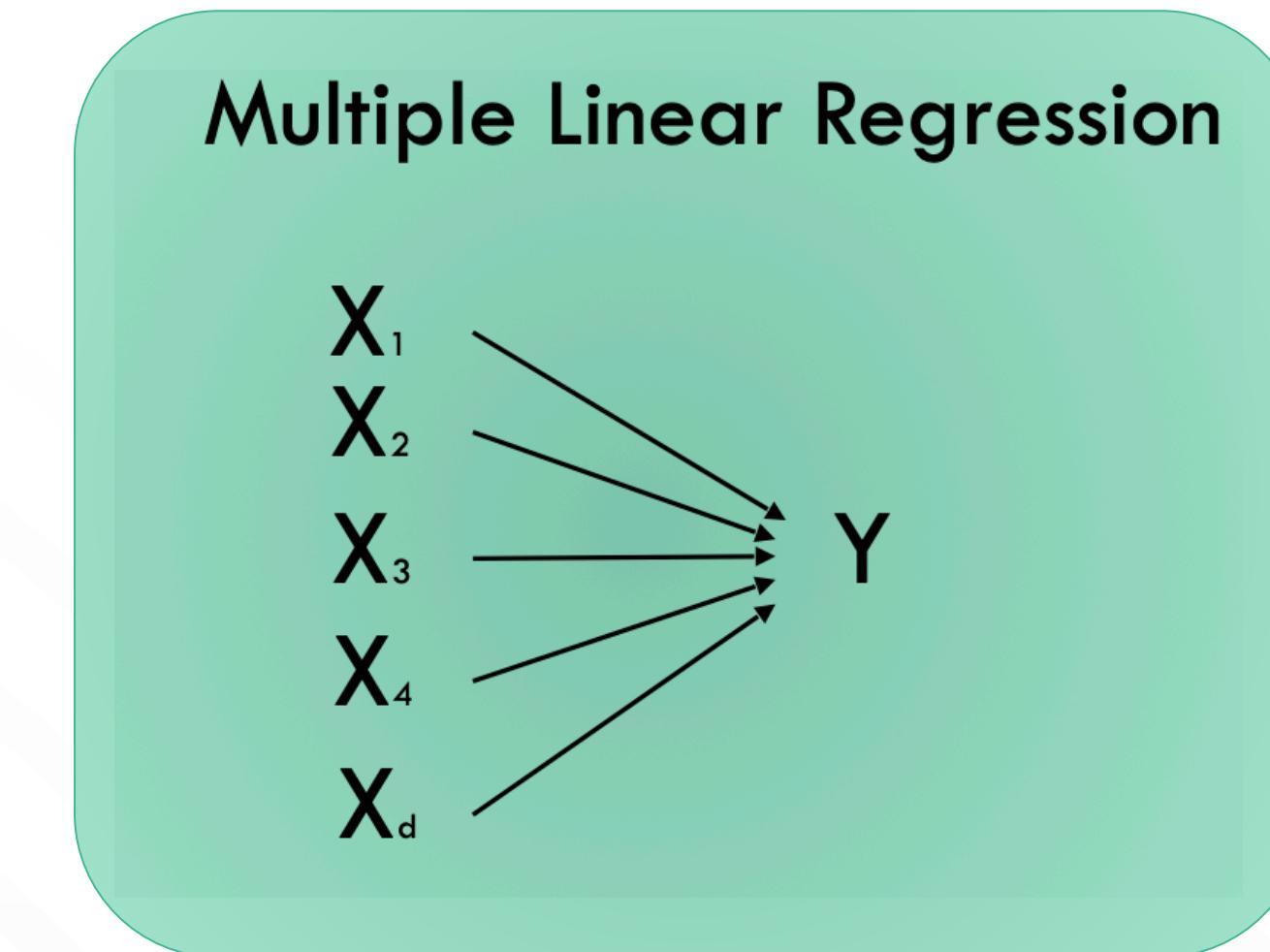
(Regression Algorithms) الخوارزميات الانحدار

أنواع Regression Analysis

- يُعتبر امتداد لخوارزمية simple linear regression باسم Polynomial Regression.
- يمكن تسميتها Multiple Linear Regression.



$$\mathbf{y} = a_0 + a_1 \mathbf{x}_1$$

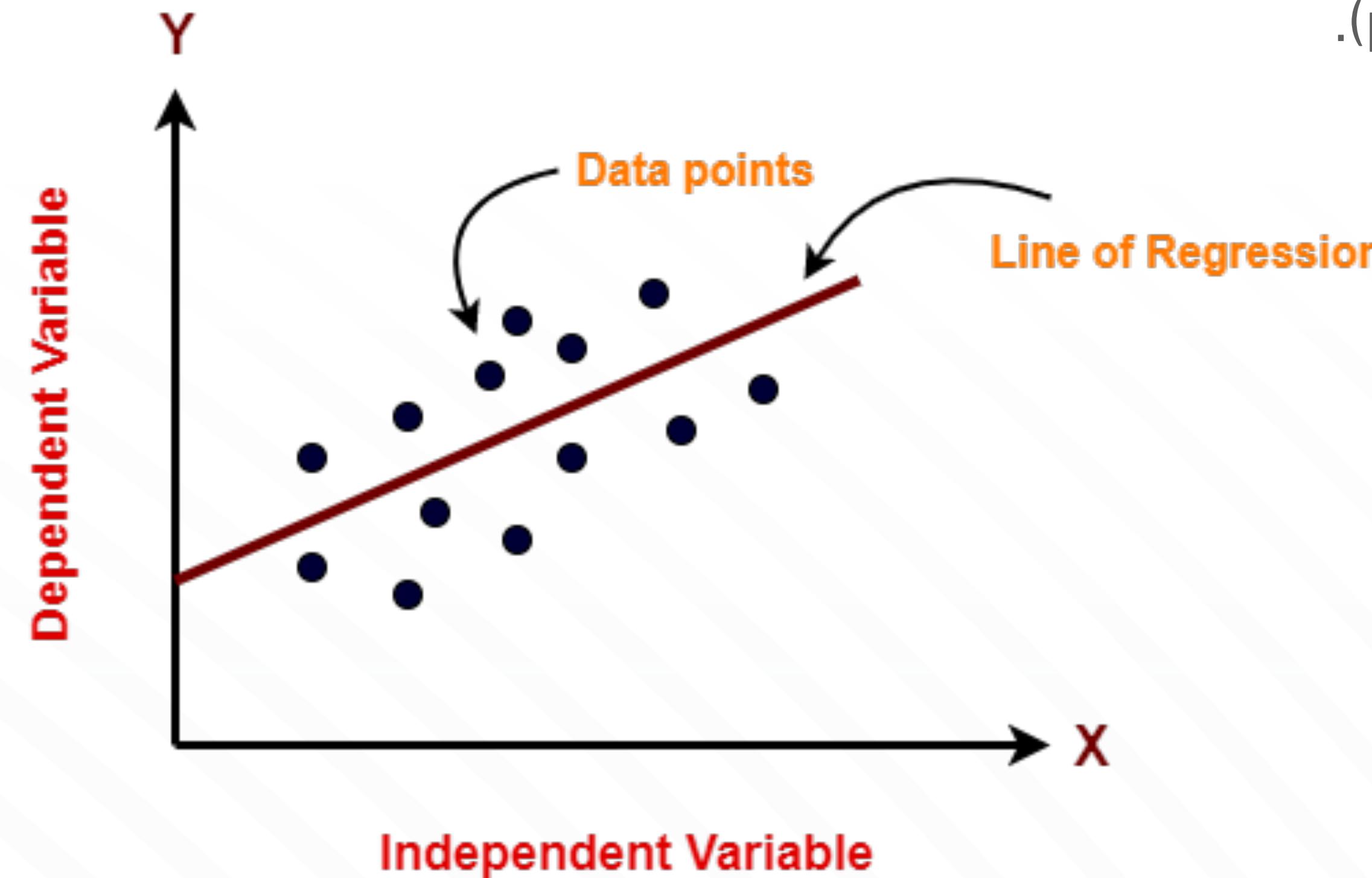


$$\mathbf{y} = a_1 \mathbf{x}_1 + \cdots + a_m \mathbf{x}_m$$

(Regression Algorithms) الخوارزميات الانحدار

- لِيُجَاد (simple linear regression) بين مجموعة البيانات نقوم باستخدام:
Ordinary least squares (OLS) method

وهي تعتمد على اختيار المتغيرات (a's) التي تقلل من مربع المسافة بين القيم الحقيقية (actual values) والقيم التي يتنبأ بها النموذج (predicted values).



تقييم نماذج الانحدار (Regression Evaluation Metrics)

- النوع الأول: Mean Squared Error (MSE)
- كلما انخفضت قيمة MSE كلما زادت دقة التنبؤ

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

MSE = mean squared error

n = number of data points

Y_i = observed values

\hat{Y}_i = predicted values

تقييم نماذج الانحدار (Regression Evaluation Metrics)

• النوع الثاني: Mean Absolute Error (MAE)

$$\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

• النوع الثالث: Root Mean Squared Error (RMSE)

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

تقييم نماذج الانحدار (Regression Evaluation Metrics)

- أفضل قيمة تساوي 1.0
- كلما قلت قيمة R² يعتبر النموذج غير دقيق
- يمكن أن تكون قيمة R² سالبة

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$$