

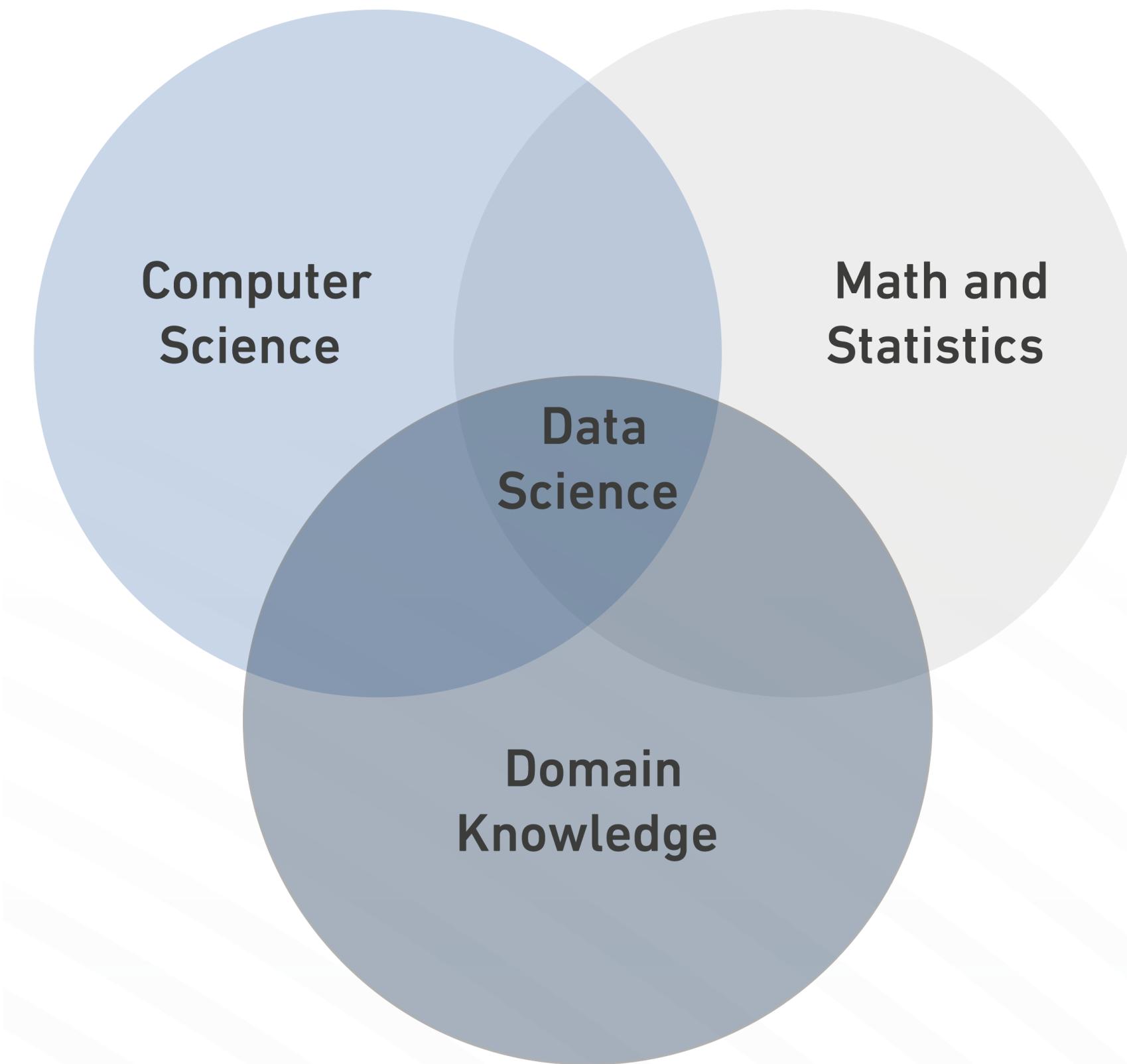


مقدمة في علم السياسات

31-07-2022

ما هو علم البيانات؟

ما هو علم البيانات



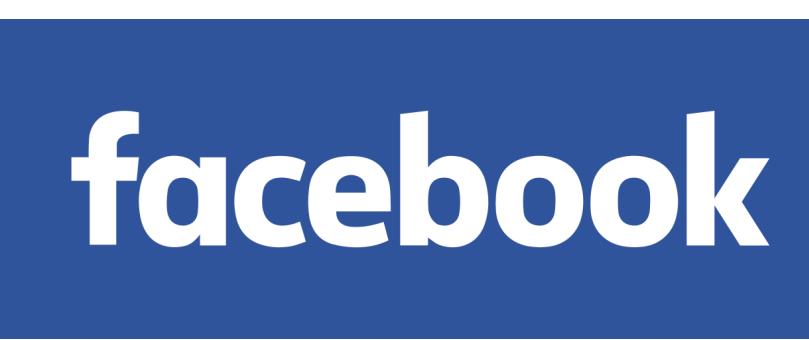
- هو علم يقوم بتطبيق الأساليب الإحصائية على البيانات بغرض الوصول لحل مشكلة معينة، حيث يتم الاستفادة من البيانات التي يتم إنتاجها بشكل يومي من عدة مصادر مثل وسائل التواصل الاجتماعي وغيرها بغرض الوصول لقرارات أو استنتاجات تقوم بحل مشكلة ما.
- ويمكن تعريفه أيضاً على أنه علم قائم على التقاطع بين عدة علوم 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.



تطبيقات علم البيانات

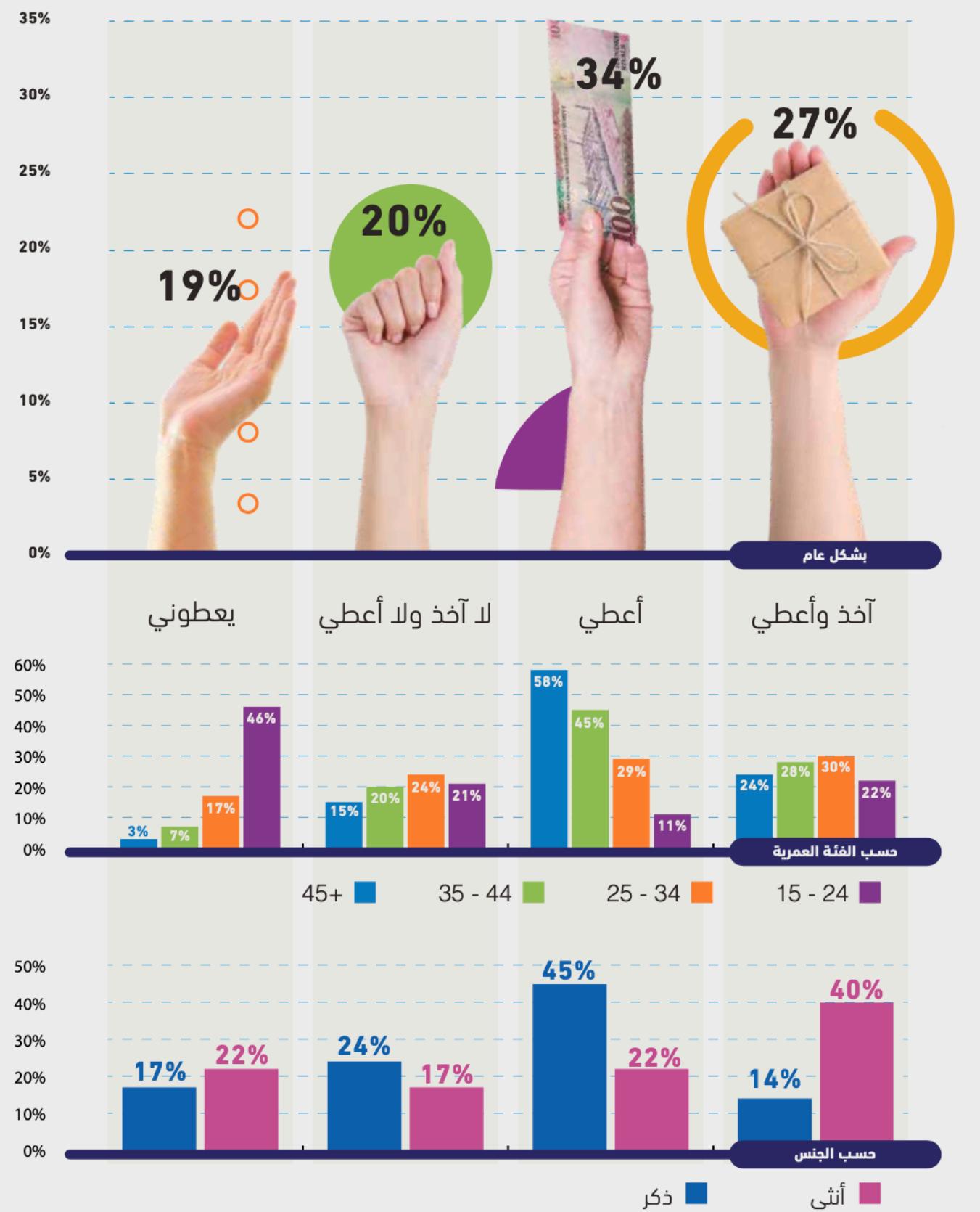


تطبيقات علم البيانات





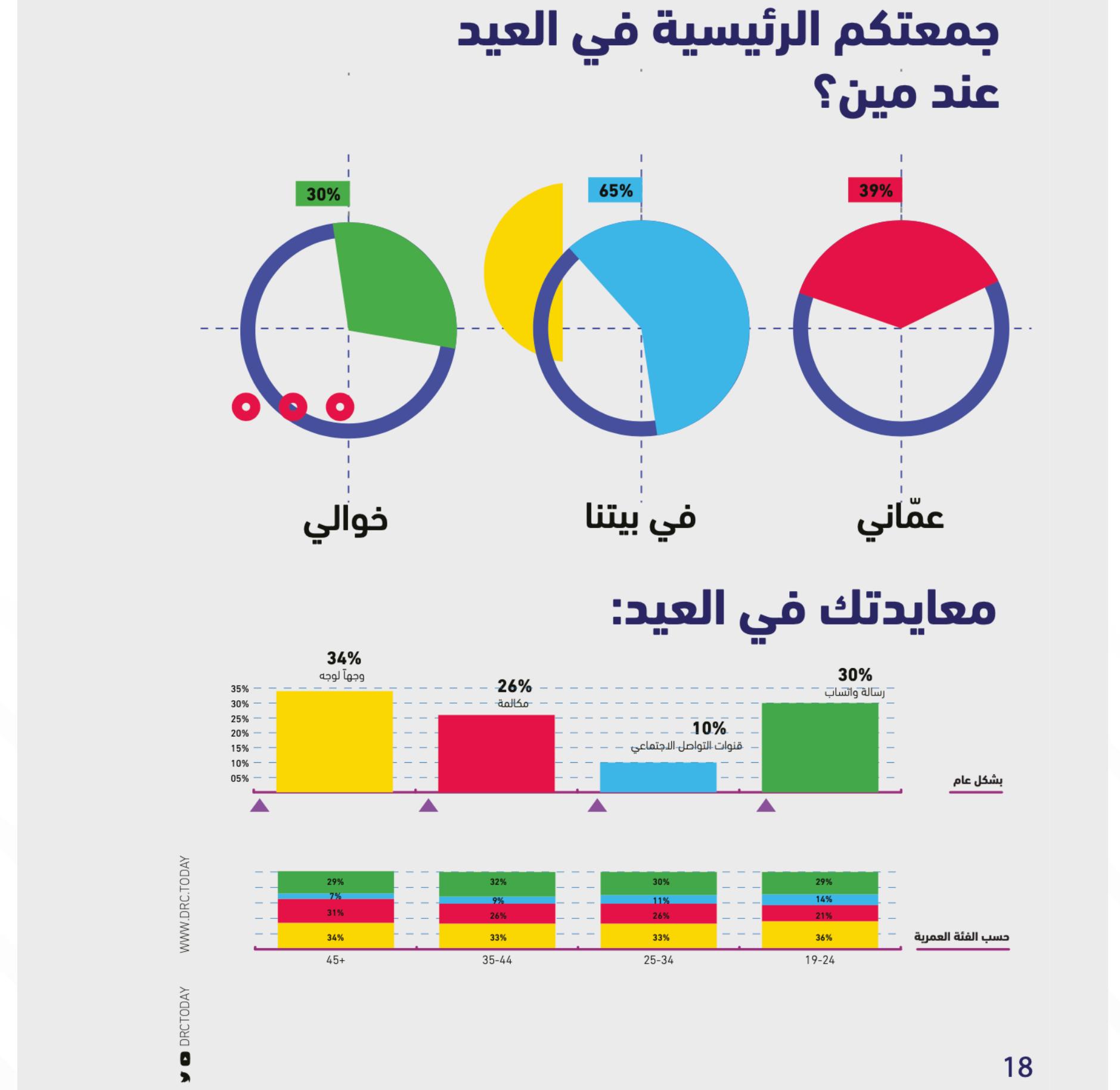
وضعك مع العيديات؟

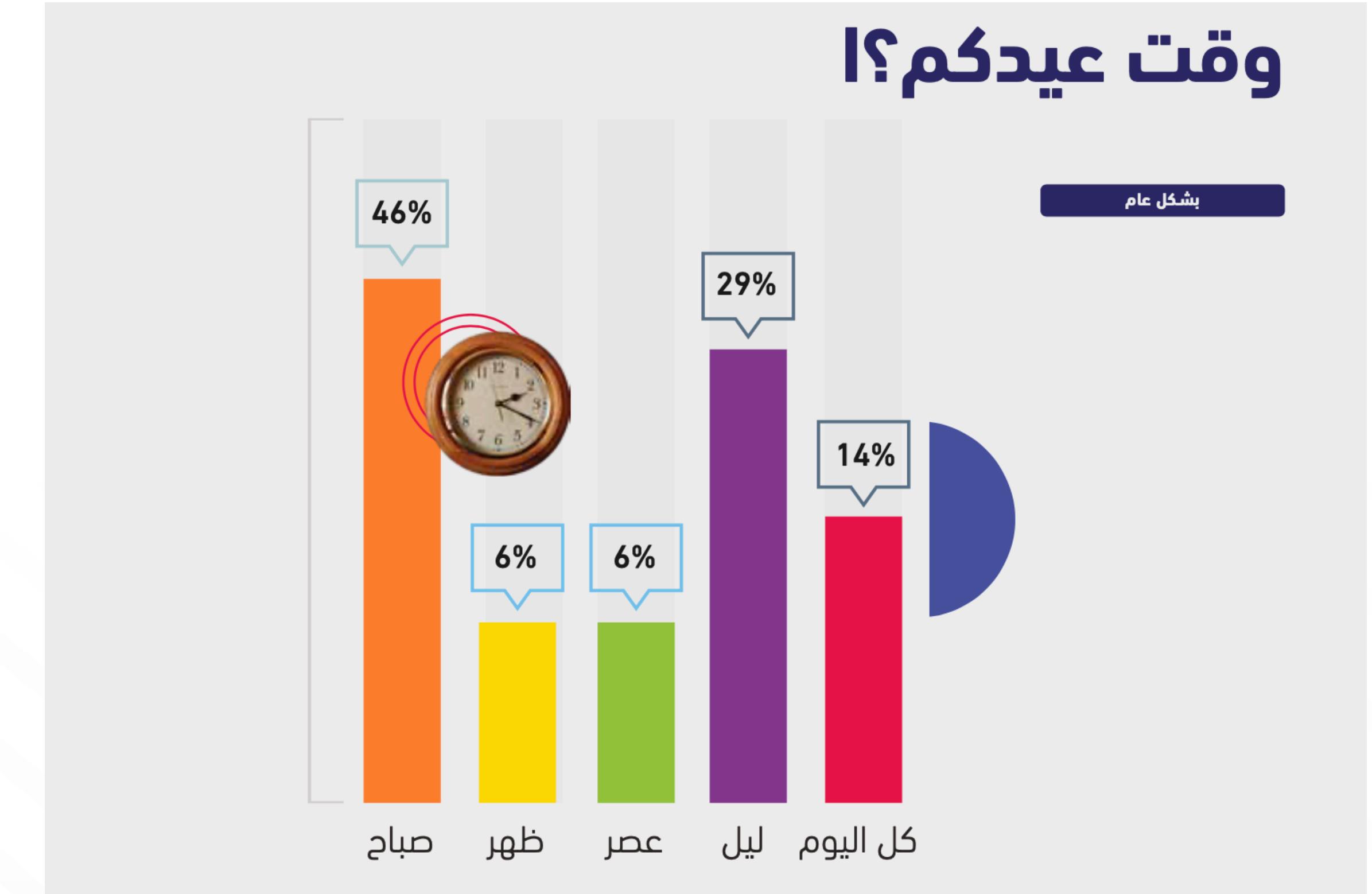
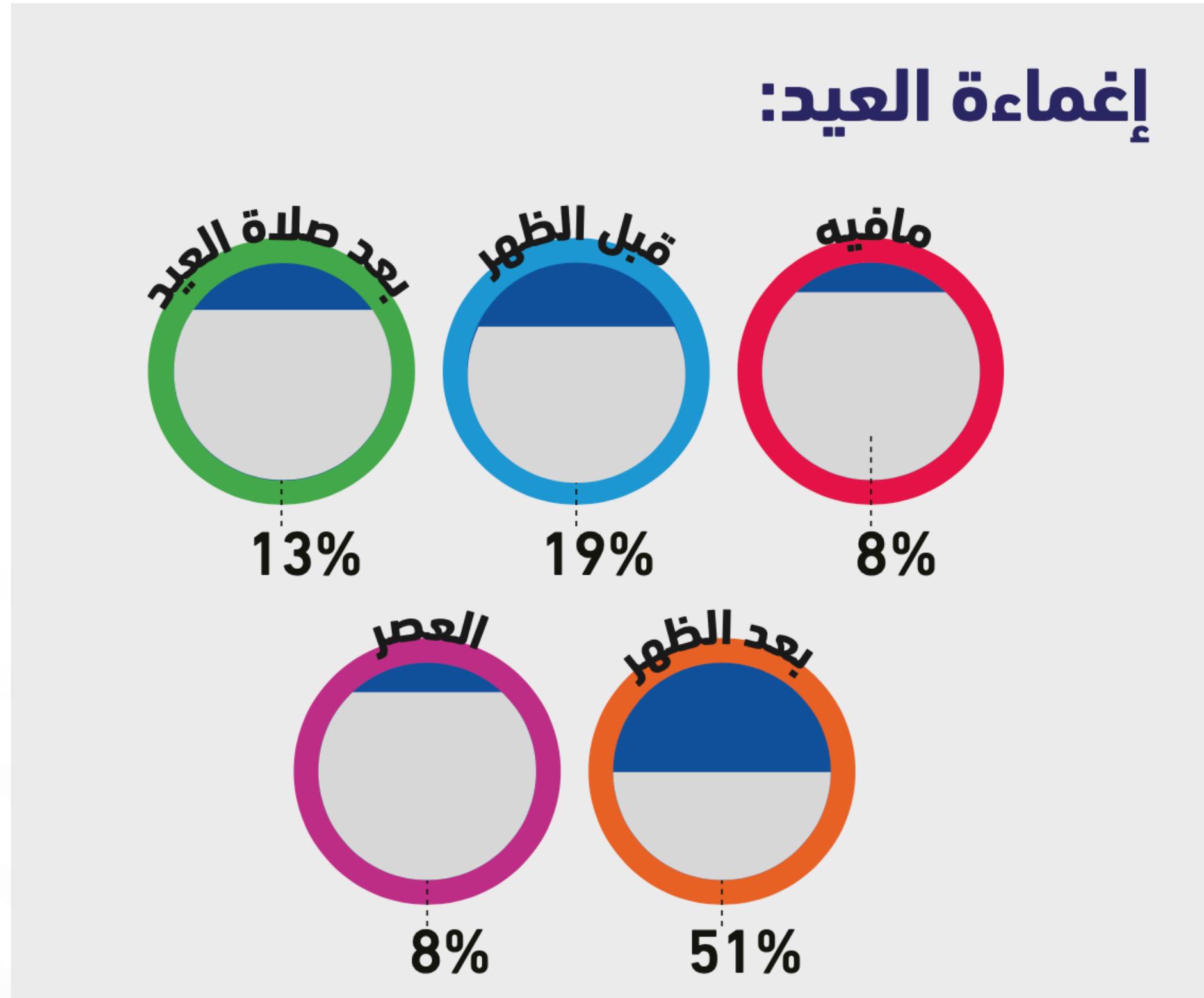


كيف تسلم في العيد؟



جمعتكم الرئيسية في العيد عند مين؟





متى أول مرة سمعت عن علم البيانات؟

تاريخ علم البيانات

- يسمى علم البيانات بالإحصاء التطبيقي أو (Applied Statistics).
- يعود تاريخ علم البيانات للقرن 19
- يعود تاريخ علم البيانات الحديث إلى عام 2000

تاريخ علم البيانات

أهمية علم البيانات:

- قابل للتطبيق باستخدام التكنولوجيا الحالية.
- تكلفة جمع ومعالجة وتحليل البيانات تتناقص سنويا وبالمقابل القيمة المقدمة من هذه البيانات تتزايد.

تاريخ علم البيانات

1962

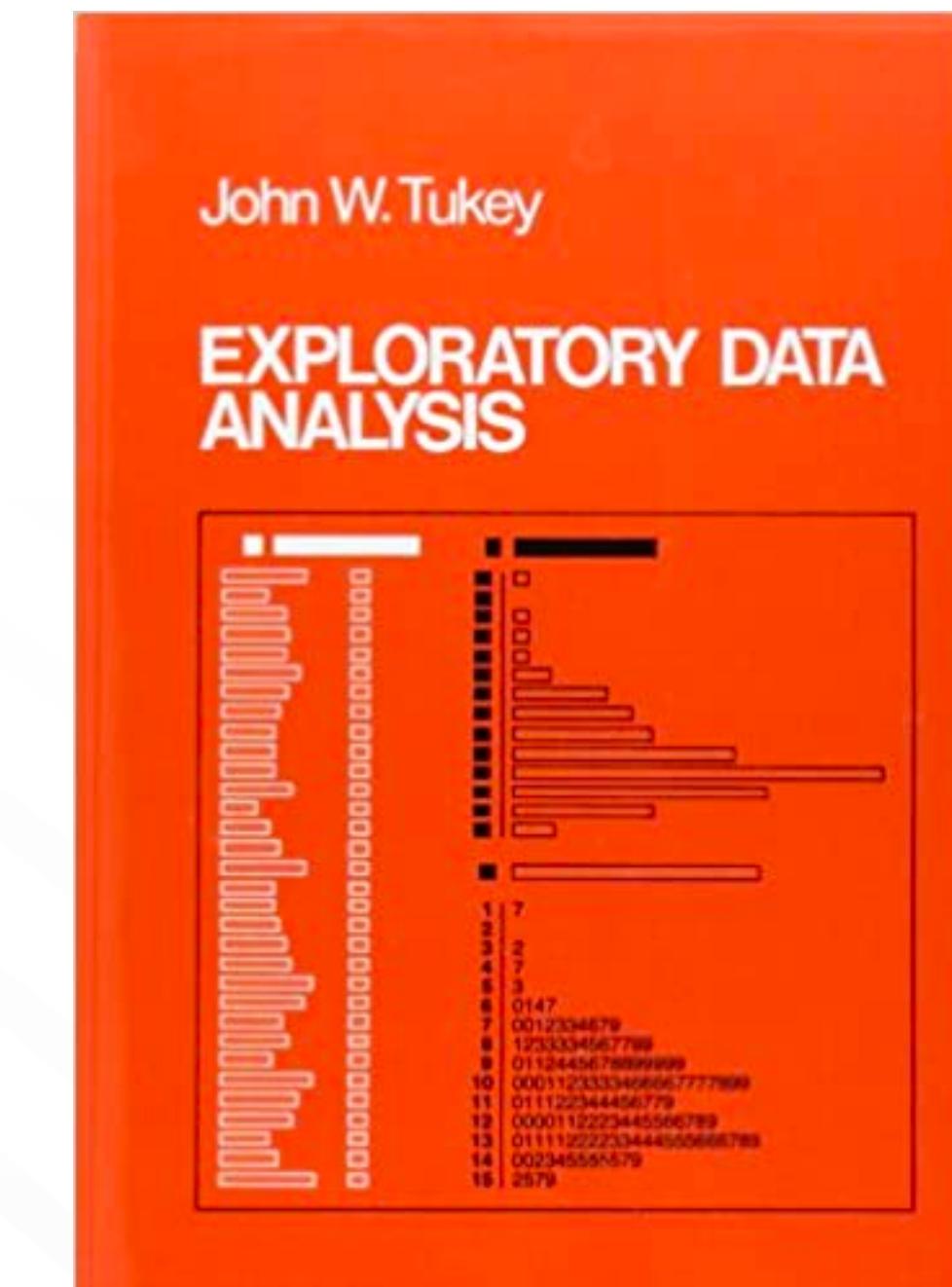


تاريخ علم البيانات

1962



1977

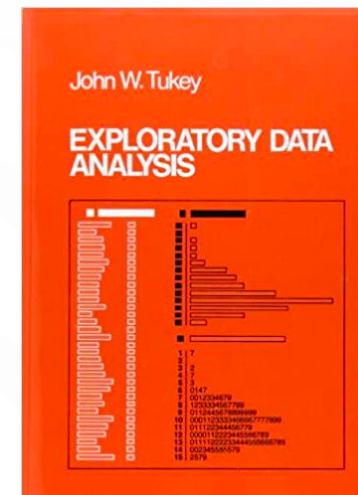


تاريخ علم البيانات

1962



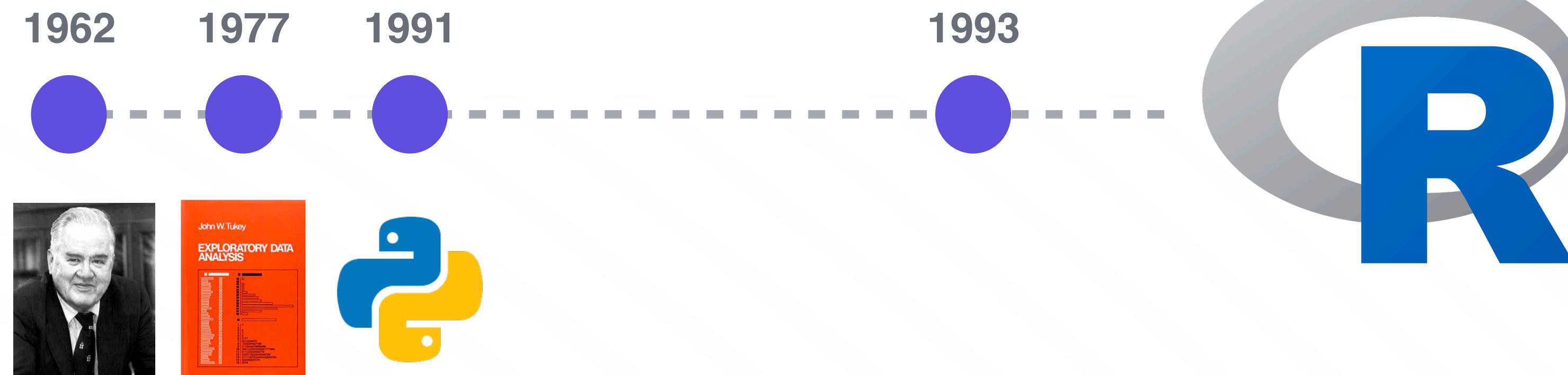
1977



1991



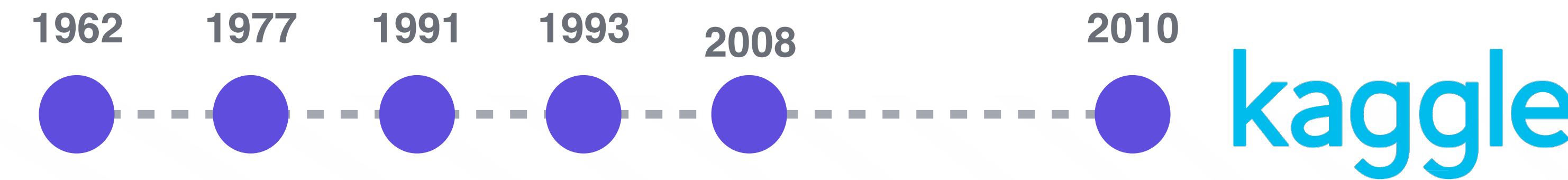
تاريخ علم البيانات



تاريخ علم البيانات

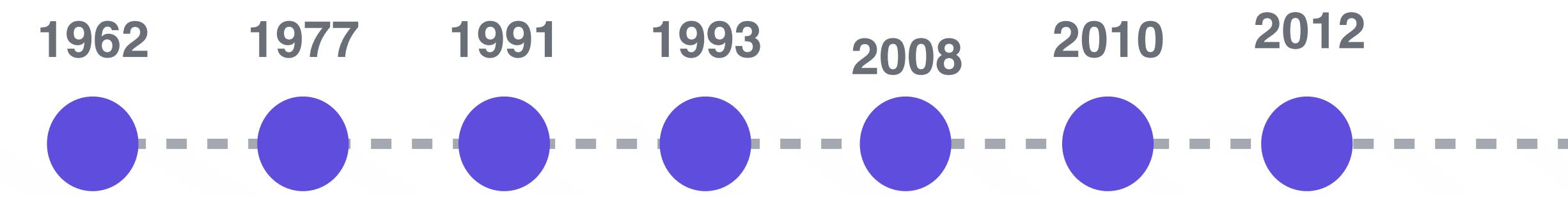


تاريخ علم البيانات

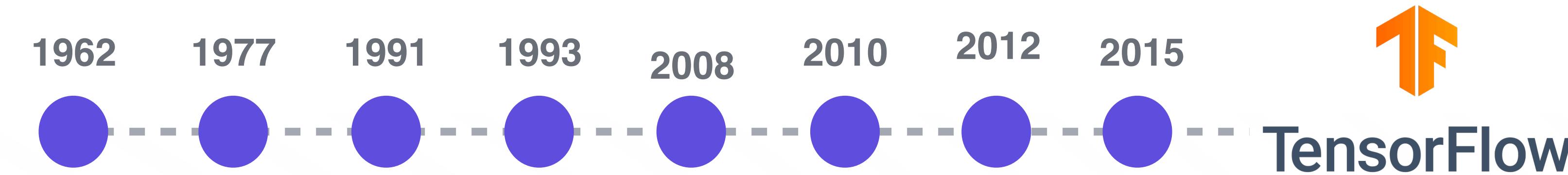




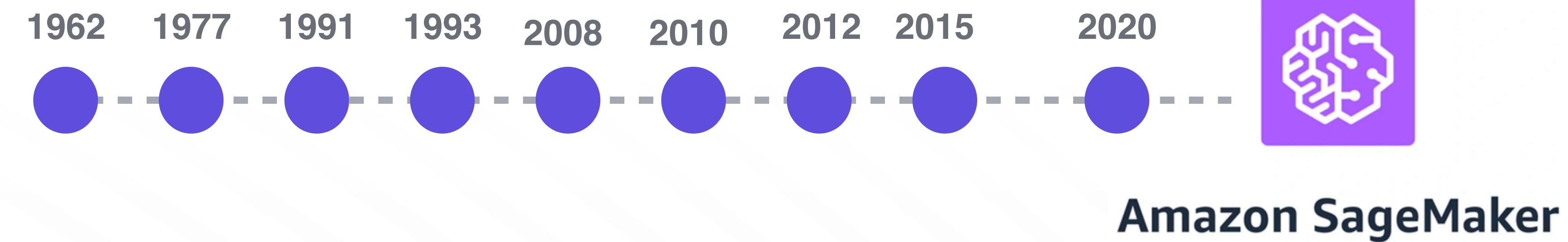
تاريخ علم البيانات



تاريخ علم البيانات



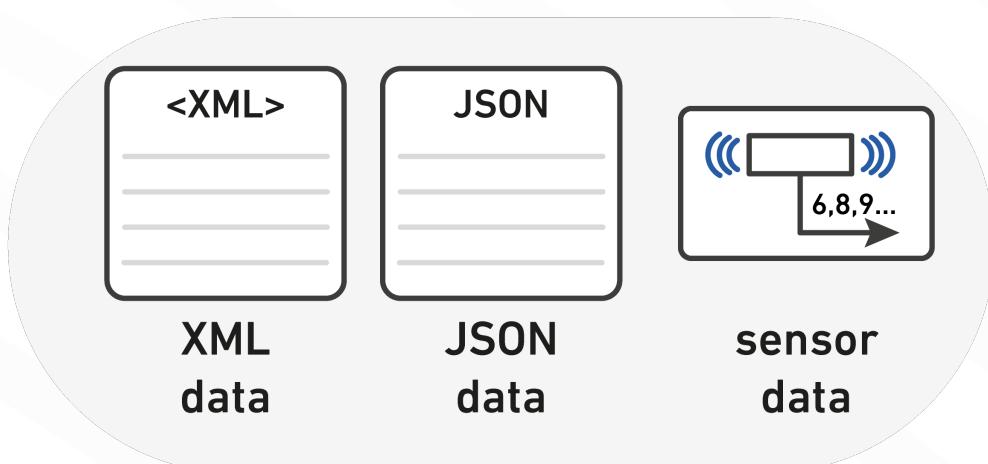
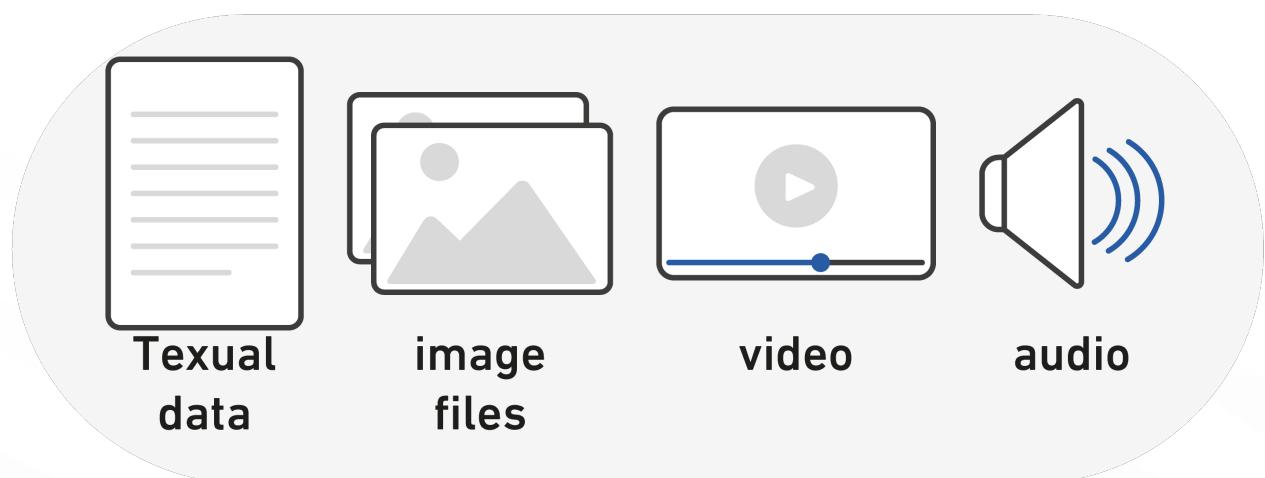
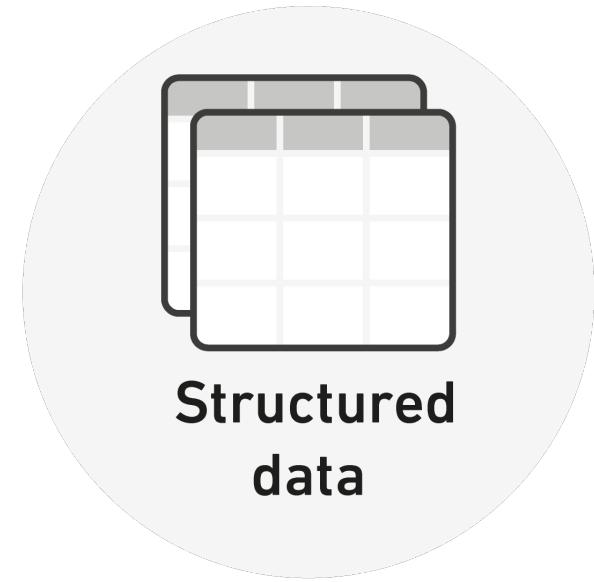
تاريخ علم البيانات



ما هي أنواع البيانات؟

أنواع البيانات

لأننا نعيش ثورة علم البيانات، يوجد لدينا أنواع مختلفة من أشكال البيانات وهي كالتالي:



١- البيانات المهيكلة

- وهي البيانات التي يتم تنظيمها في جداول مكونة من صفوف وأعمدة مثل: ملفات `.xlsx`.

٢- البيانات الغير مهيكلة

- وهذا النوع من البيانات بعكس النوع السابق يوجد بشكل غير منظم ويمثل معظم البيانات مثل: البيانات النصية والفيديو والصور والبيانات الصوتية وغيرها.

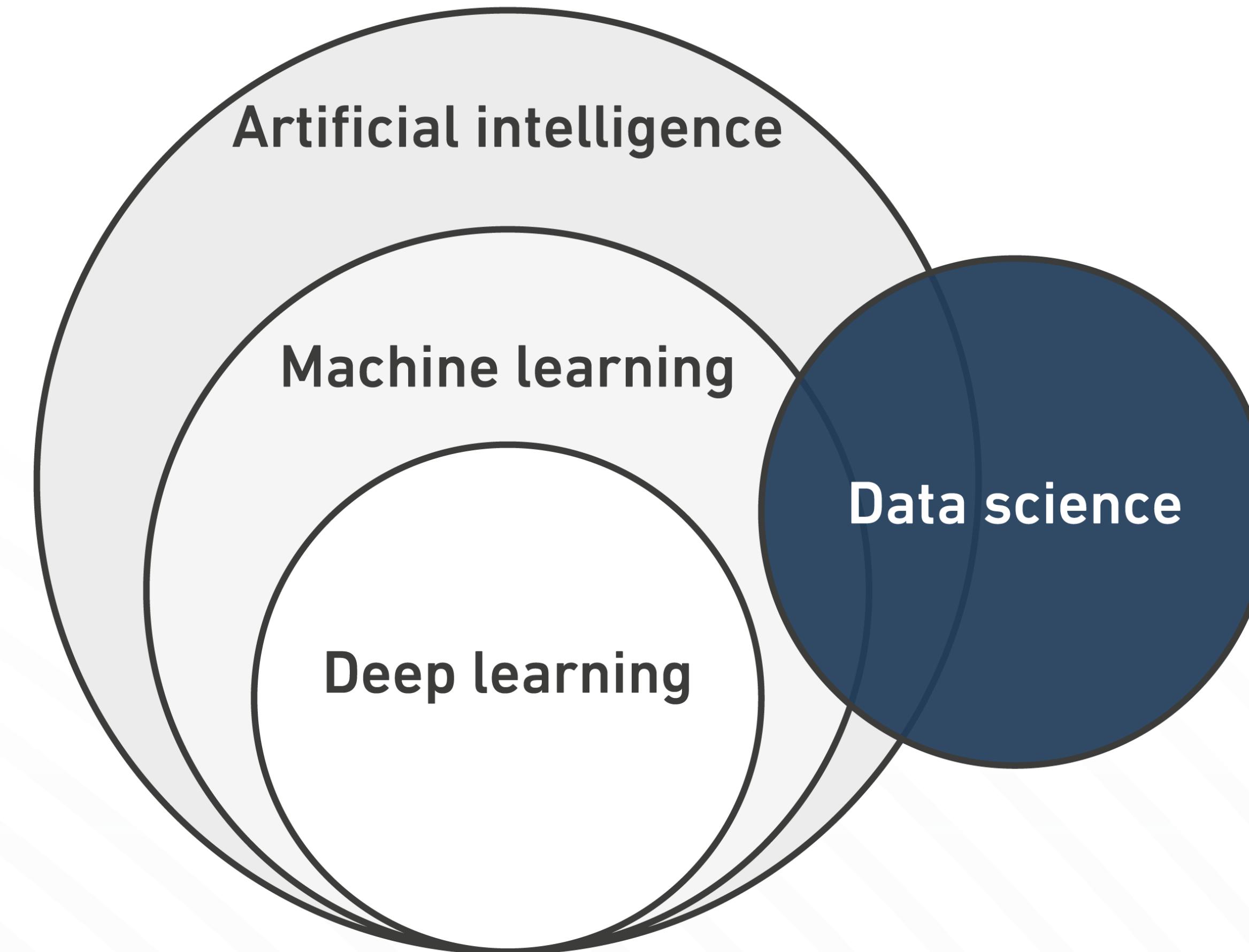
٣- البيانات شبه المهيكلة

- وهذا النوع يجمع بين النوعين السابقين حيث أن البيانات موجودة بشكل شبه منظم مثل ملفات `xml`.

المهارات التي يحتاجها عالم البيانات

- لغات البرمجة مثل: SQL, R, Python.
- العمل على البيانات عن طريق جمعها و تنظيفها و تحويلها (Transform).
- الإحصاء الوصفية (Descriptive Statistics).
- عرض البيانات (Data Visualisation).
- التعامل مع البيانات الضخمة.
- بناء نماذج للتعلم الآلي.

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



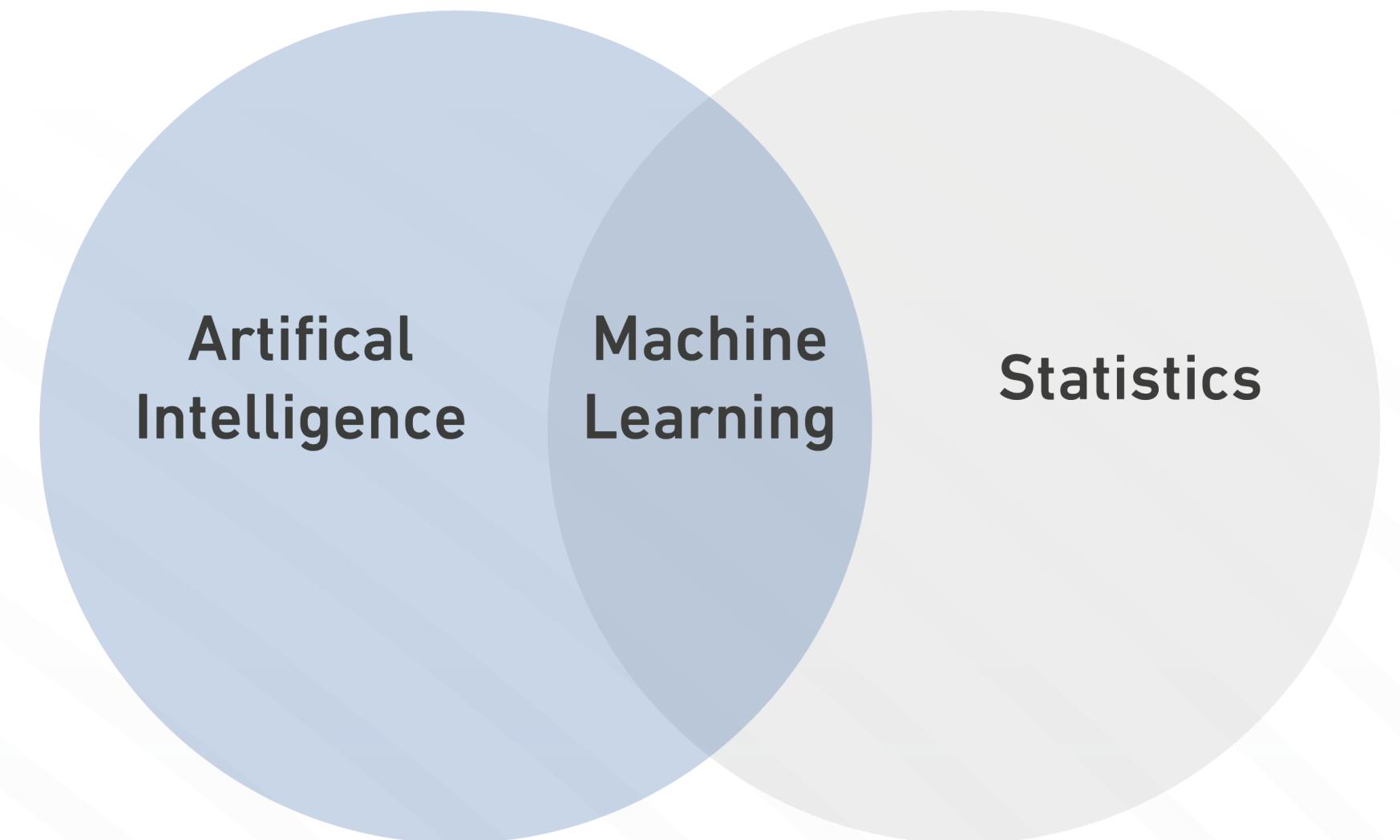
الفرق بين علم البيانات (Data Science) و الذكاء الاصناعي (AI) و تعلم الآلة (Machine Learning) و التعلم العميق (Deep Learning)

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:

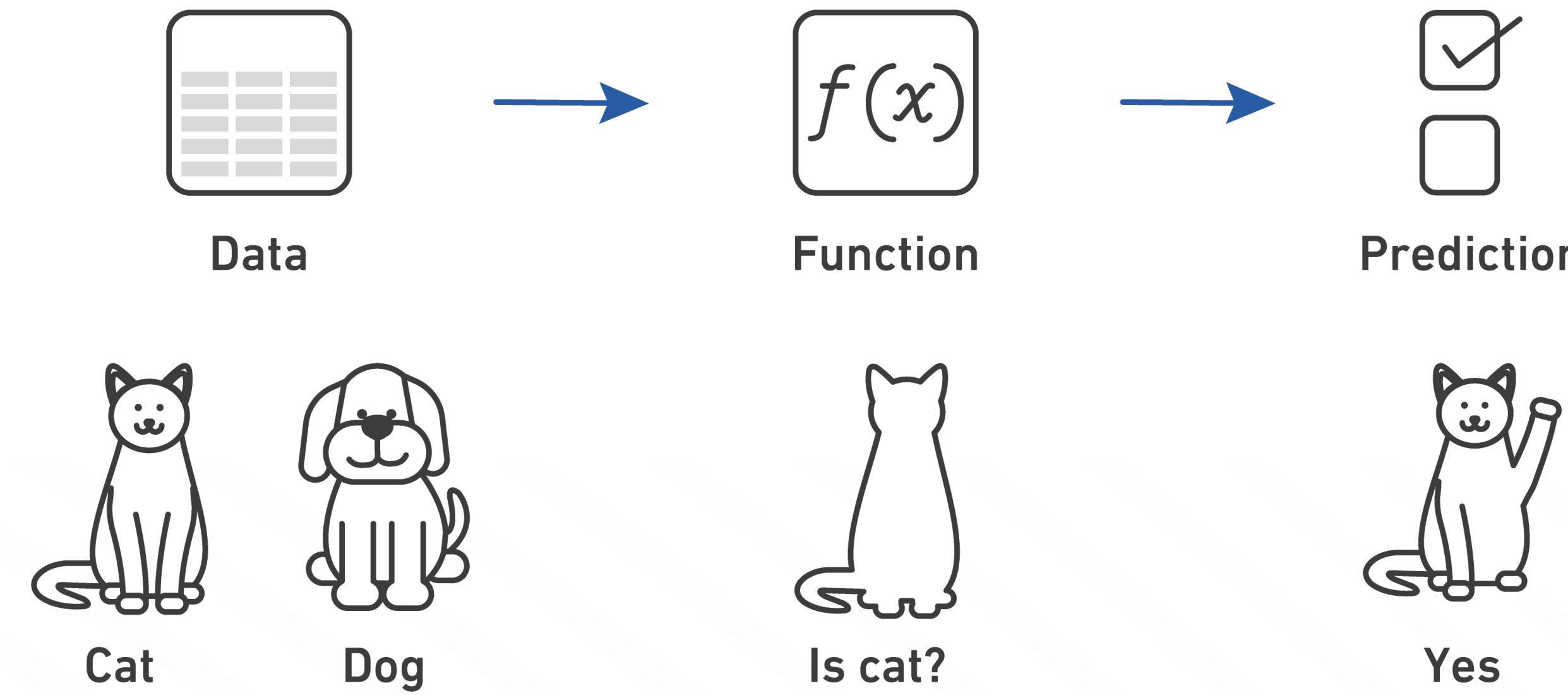
- علم البيانات يقوم باستخدام تقنيات (AI) و (Machine Learning) و تطبيقها على البيانات بغرض الوصول لقرارات أو استنتاجات معينة.
- أما التعلم العميق (Deep Learning) فهو جزء فرعي من تعلم الآلة (Machine Learning) الذي هو أيضاً جزء فرعي من الذكاء الاصناعي (AI) وجميعها تشير للتقنيات التي تساعد الكمبيوتر على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



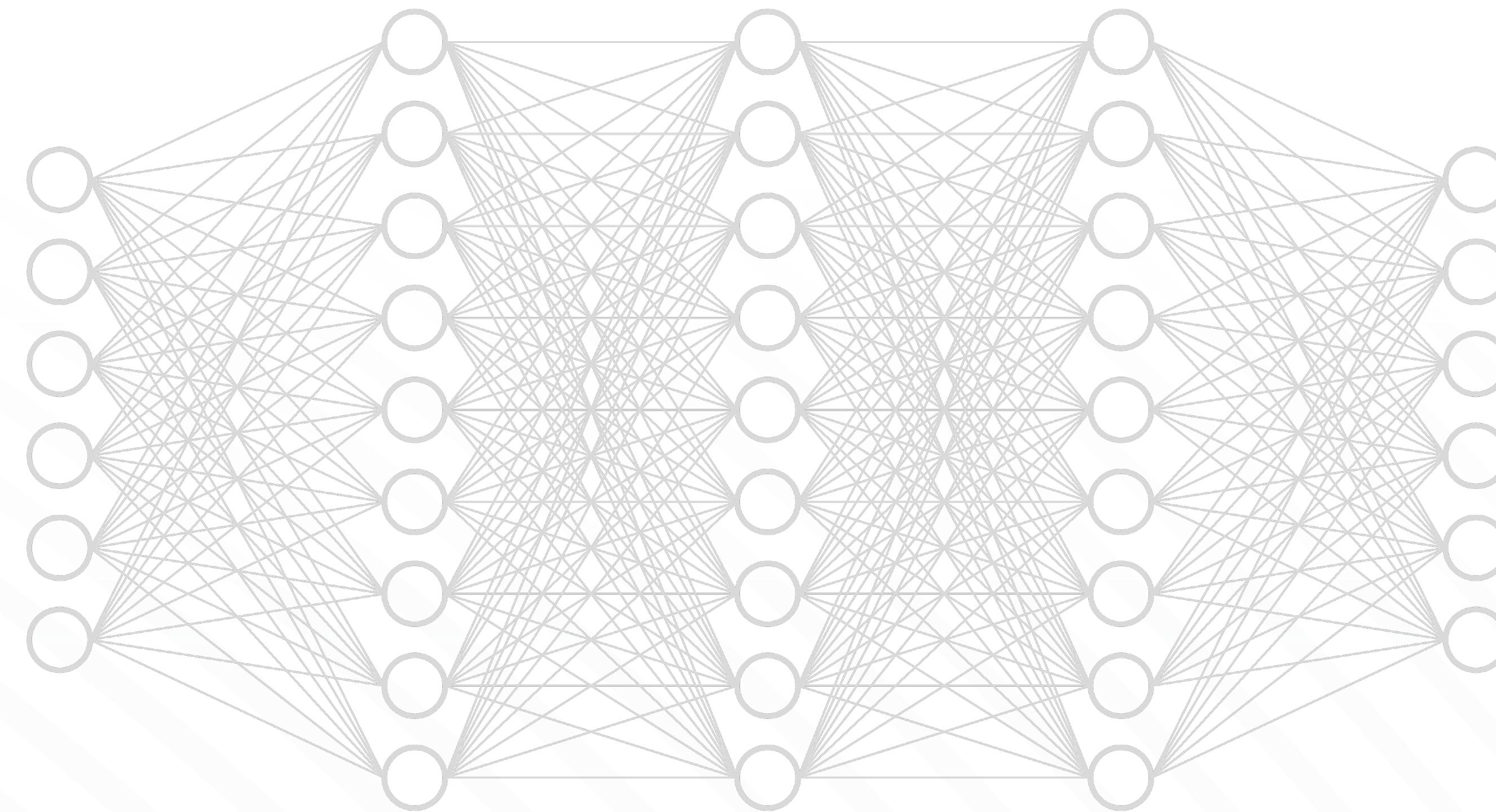
تعلم الآلة (Machine Learning)



- 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؟

التعلم العميق (Deep Learning)

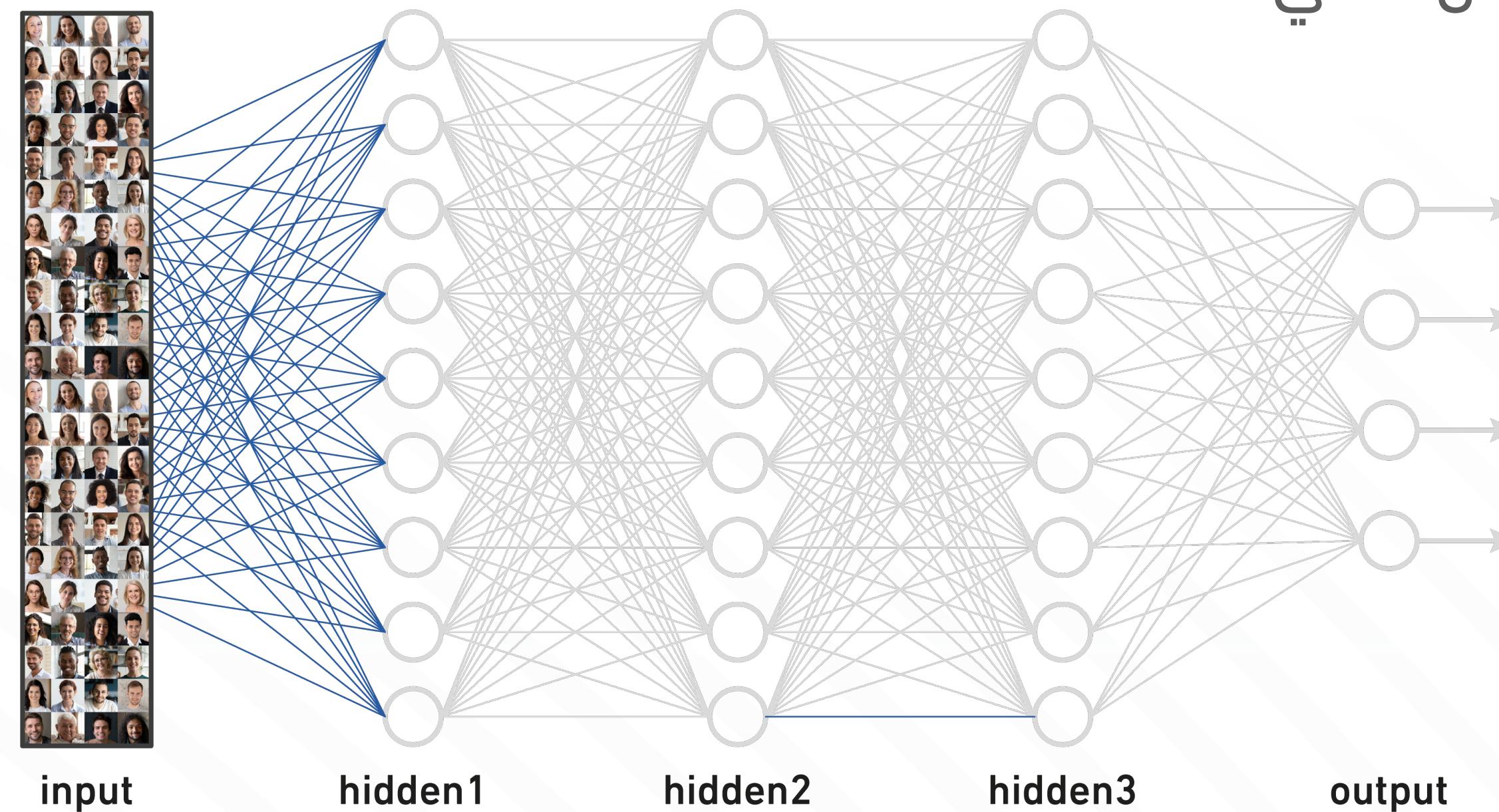
- التطور الكبير الذي حدث في تعلم الآلة أدى لظهور التعلم العميق
- يعتبر التعلم العميق عبارة عن عدة طبقات من نماذج Machine Learning مبنية فوق بعضها البعض، كما في الشكل التالي:



التعلم العميق (Deep Learning)

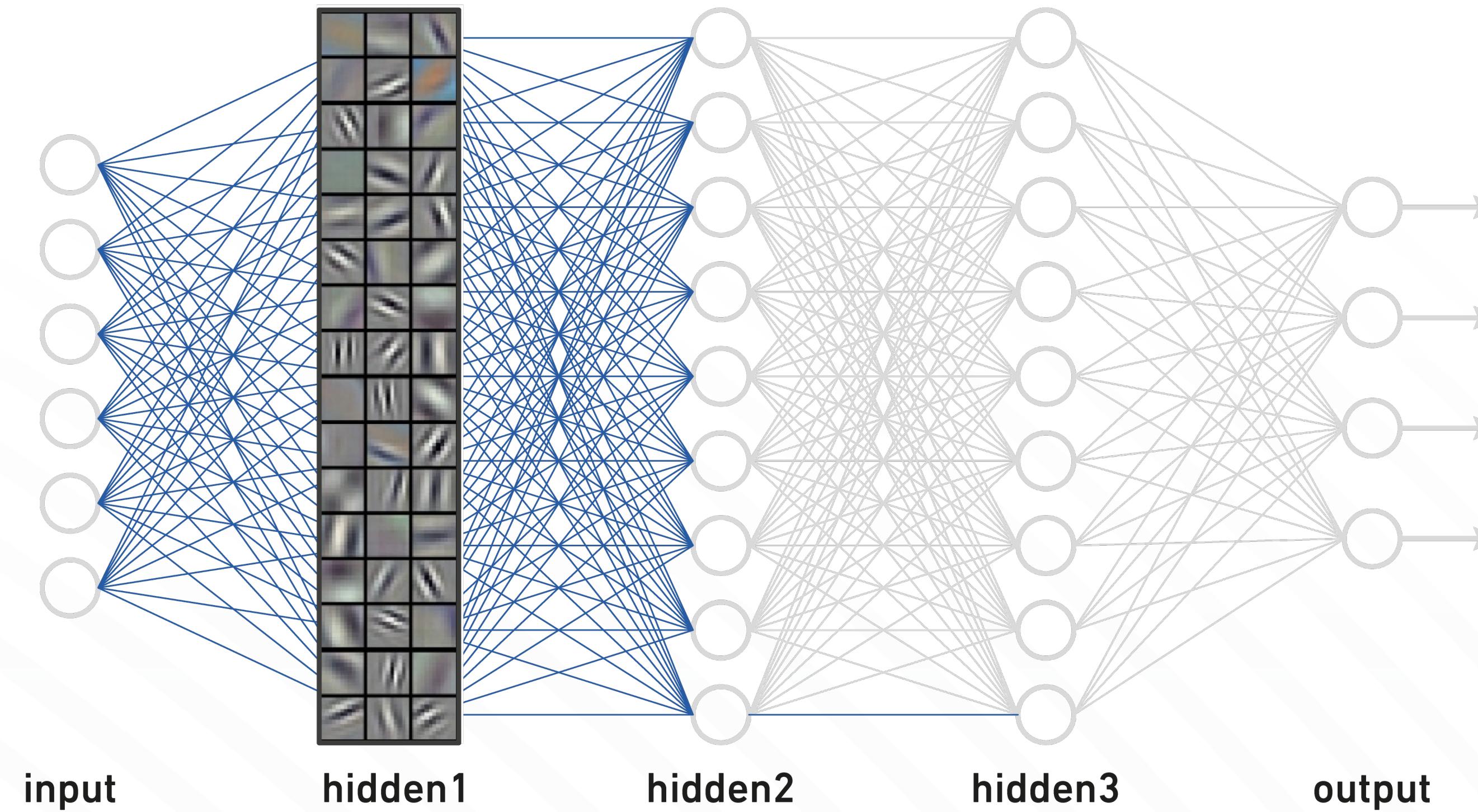
مثال توضيحي:

- لو كنا نريد تعلم خوارزمية (Deep Neural Networks) التعرف على ما إذا كانت الصور تحتوي وجه إنسان أو لا ؟
- في البداية سوف نقوم بعرض عدة صور لأشخاص مختلفين بهدف تعلم الخوارزمية كيف تبدو أشكال مختلفة من الوجوه البشرية كما في الشكل التالي:



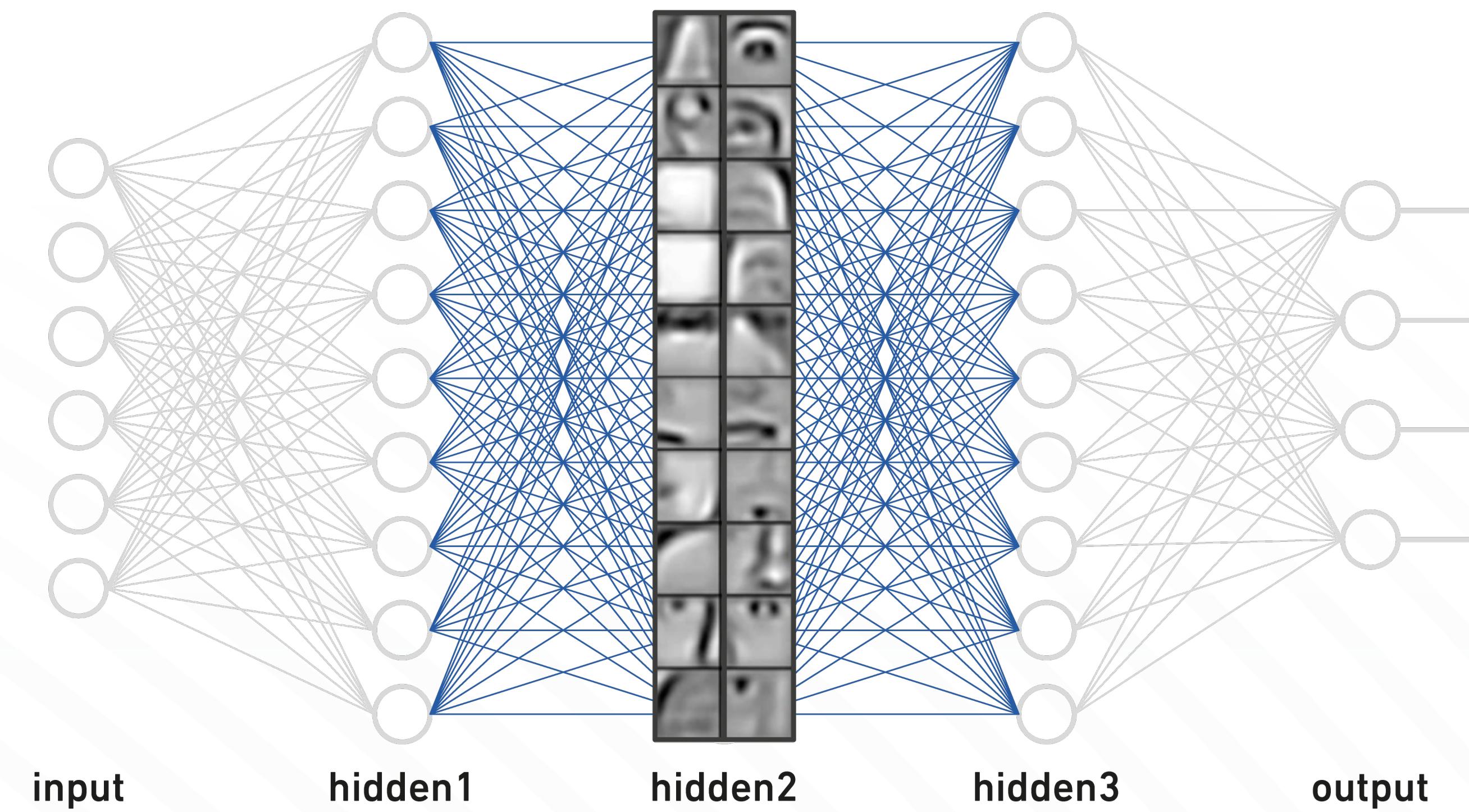
التعلم العميق (Deep Learning)

- في الطبقة الأولى سوف نقوم باستخراج بعض الخصائص الأفقية والعمودية كما في الشكل التالي:



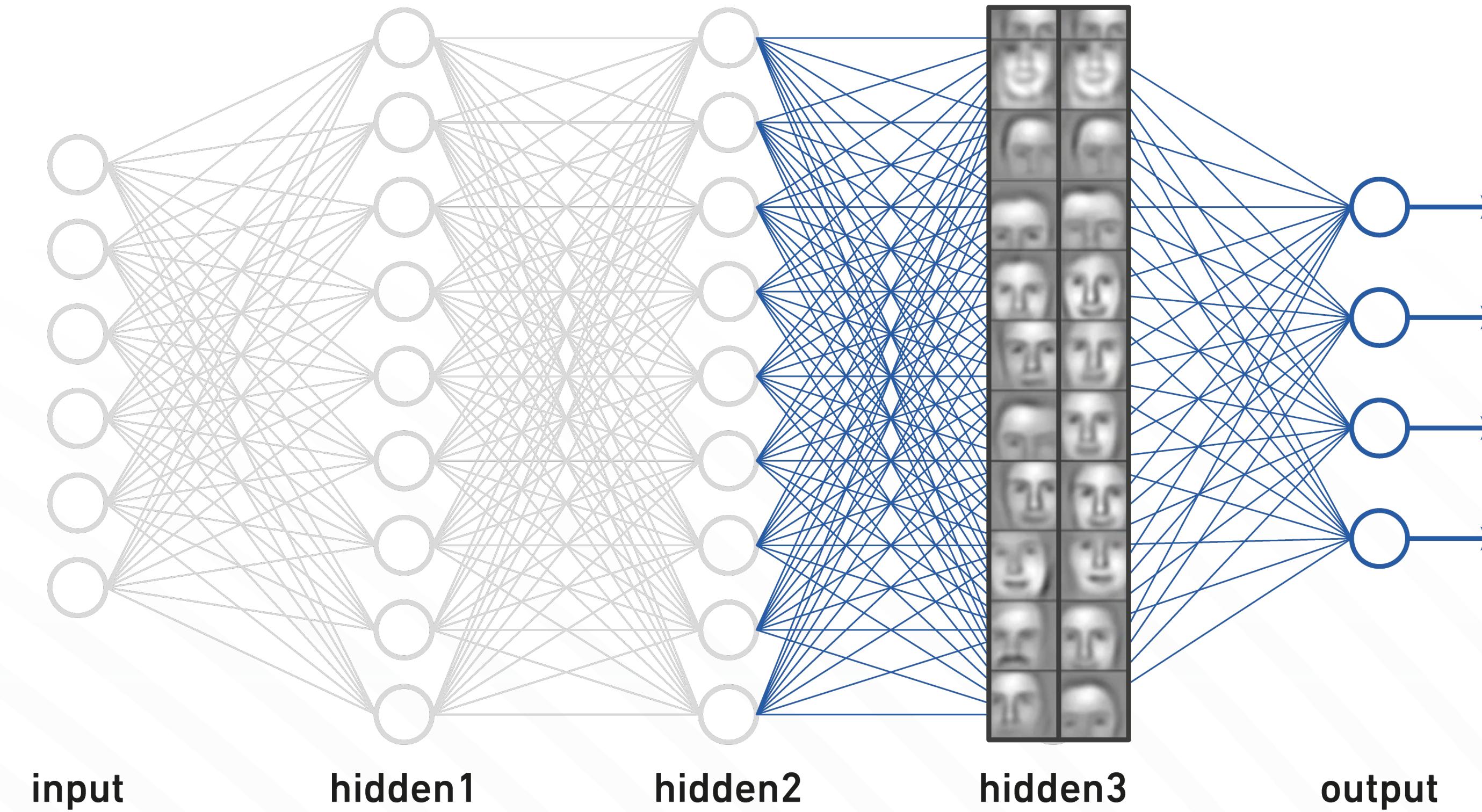
التعلم العميق (Deep Learning)

- في الطبقة الثانية سوف نقوم باستخراج خصائص الوجه الأكثر تعقيد مثل: الأنف و العين، كما في الشكل التالي.



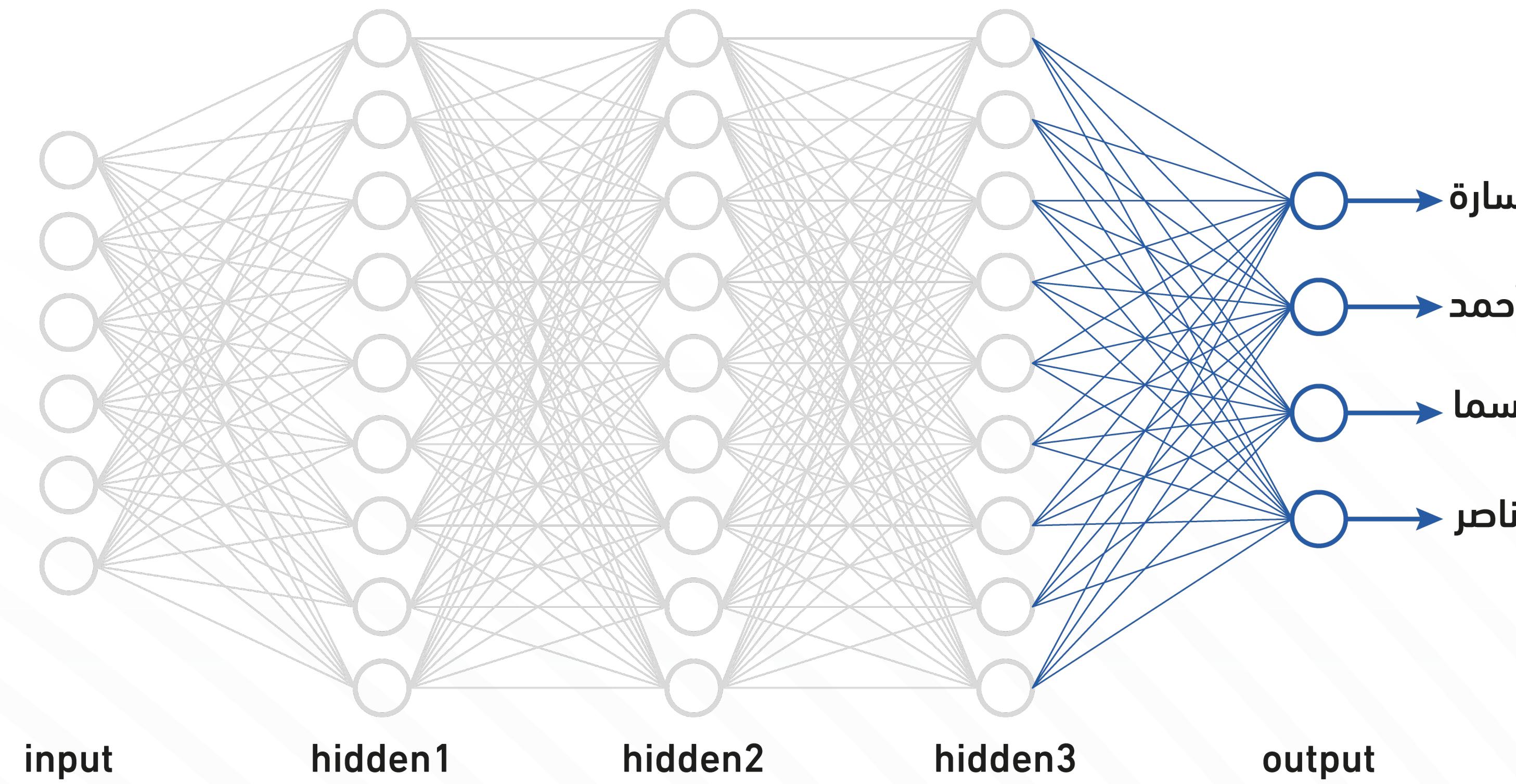
التعلم العميق (Deep Learning)

- في الطبقة الثالثة سوف نقوم باستخراج الخصائص العامة ل كامل الوجه، كما في الشكل التالي.



التعلم العميق (Deep Learning)

- وفي آخر طبقة سوف نكون قد تعرفنا على وجوه الأشخاص وعرض الأسماء كما في الشكل التالي.

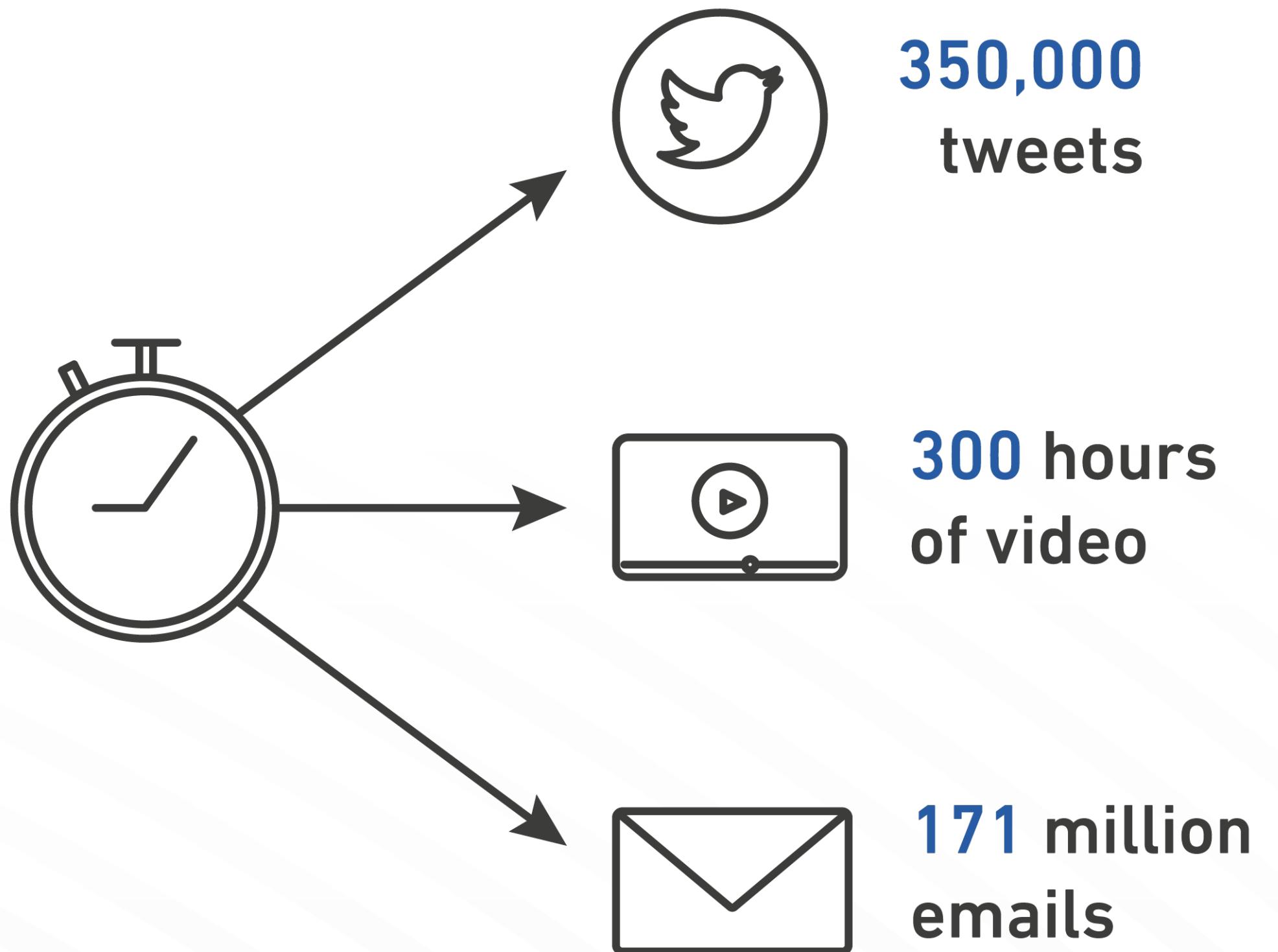


ما هي التدريبات المرتبطة بعلم البيانات؟

الخصائص المرتبطة بعلم البيانات

- تعلم الآلة (Machine Learning)
- الذكاء الاصناعي (AI)
- التعلم العميق (Deep Learning)
- ذكاء الأعمال (Business Intelligence)
- هندسة البيانات (Data Engineering)
- البيانات الضخمة (Big Data)
- معالجة اللغة الطبيعية (Natural Language Processing (NLP))

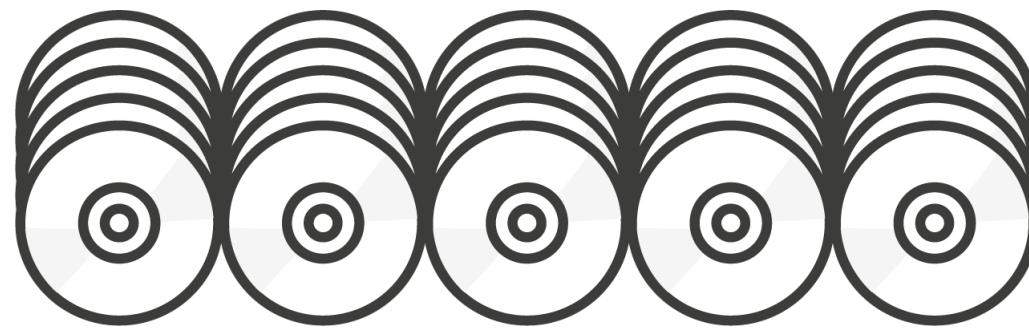
خصائص البيانات



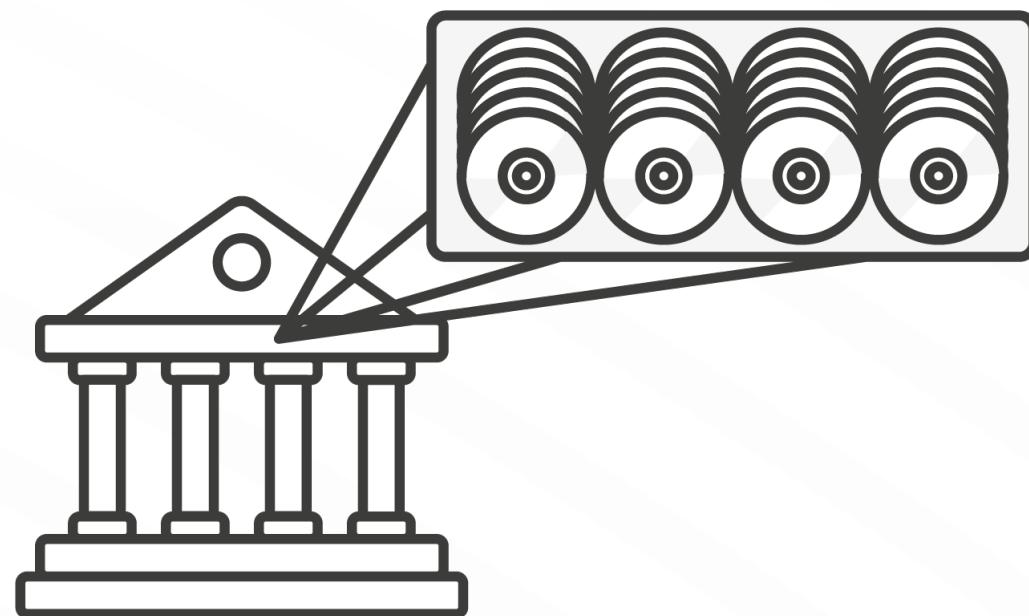
• **Velocity: السرعة**

من خصائص البيانات إنتاج البيانات بسرعة هائلة جدا حيث أنه خلال ثواني ينتج لدينا كم كبير من البيانات مما يتطلب حلول سريعة ومرنة لهذا التدفق السريع للبيانات. يوضح الشكل التالي السرعة الهائلة في تدفق البيانات فخلال دقيقة واحدة يتم كتابة 350.000 تغريدة و تحميل 300 ساعة من لقطات الفيديو على YouTube و إرسال 171 مليون بريد إلكتروني.

خصائص البيانات



5 billion DVDs
=
2.5 exabytes



65,000 DVDs
=
300 terabytes

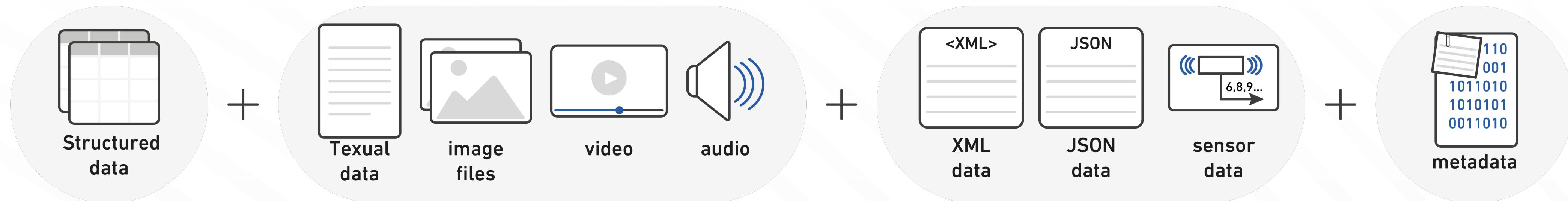
• ثانياً: الحجم Volume

يتزايد حجم البيانات باستمرار مما ينتج لدينا بيانات ضخمة ناتجة من عدة مصادر مثل: وسائل التواصل الاجتماعي و أجهزة الاستشعار مثل أجهزة استشعار GPS ، RFIDs . المعاملات المصرفية عبر الانترنت وغيرها. ويوضح الشكل التالي مدى ضخامة البيانات التي يتم انتاجها خلال اليوم الواحد حيث تعادل مامقداره (EBs 2.5) مقارنة بجميع البيانات في مكتبة الكونجرس (TBs 300).

خصائص البيانات

- **ثالثاً: التنوع Variety**

نظراً لكثرة البيانات من حولنا فهي تختلف من ناحية الأنواع والتنسيقات وتشمل بيانات منظمة في جداول مثل بيانات المعاملات المالية ، وبيانات شبه منتظمة مثل سائل البريد الإلكتروني وبيانات غير منتظمة مثل الصور ويوضح الشكل التالي الأنواع المختلفة للبيانات.



خصائص البيانات

• رابعاً: الصحة والموثوقية Veracity

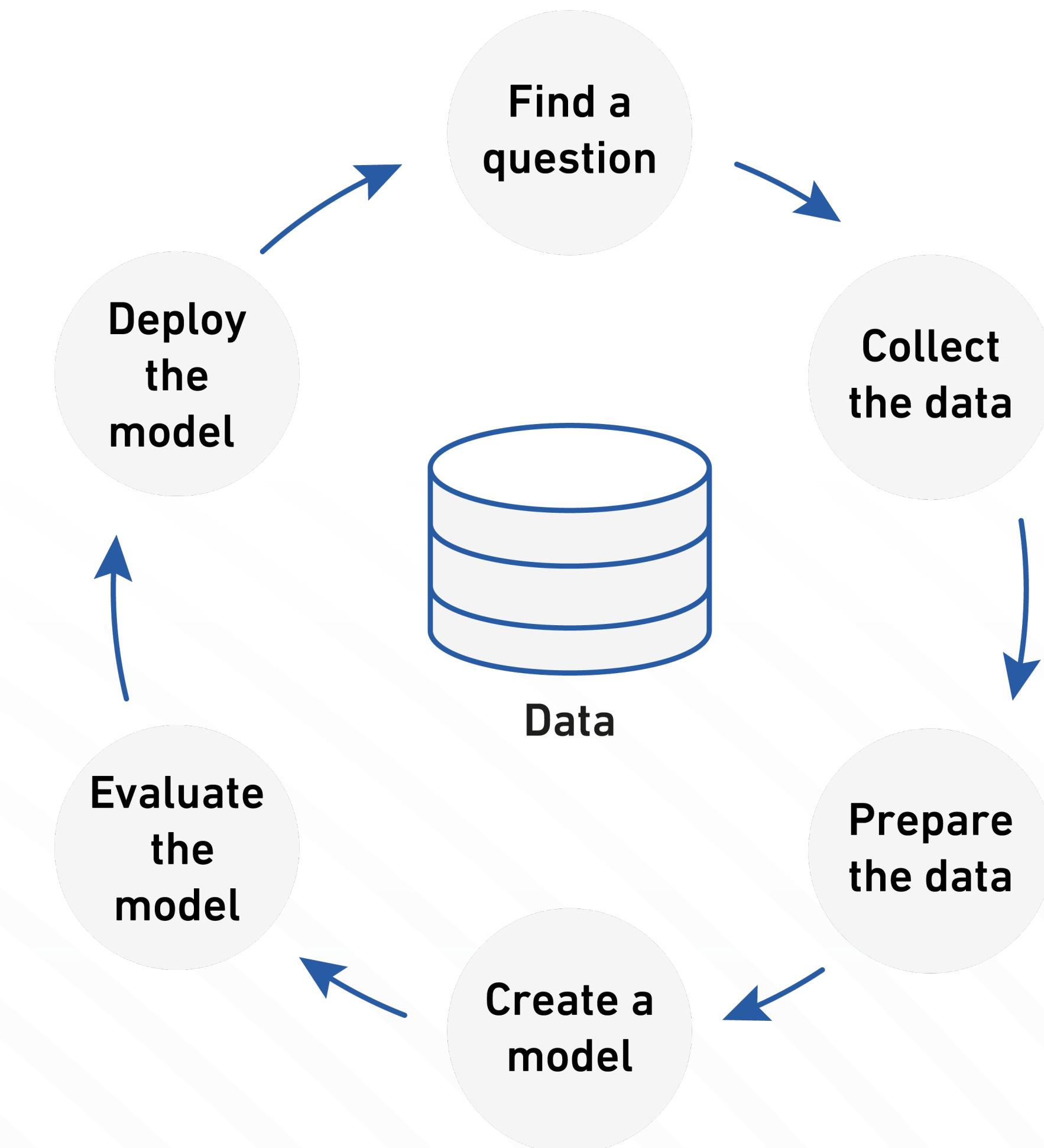
أحد خصائص البيانات التي ينبغي التأكد منها هي جودة ودقة البيانات ونقصد بذلك خلوها من البيانات التي لا يمكن تحويلها لمعلومات وبالتالي لاتكون لهذه البيانات قيمة. وتعتمد جودة البيانات على مصدر البيانات ونوعها فمثلاً: البيانات التي يقوم بالتحكم في المدخلات التي يقوم بإرسالها المستخدم تكون أعلى جودة وأكثر موثوقية مثل البيانات التي نحصل عليها من بعكس مثل البيانات التي لاتتحكم بالمدخلات مثل: (online customer registrations) .(blog postings)

خصائص البيانات

- خامساً: القيمة Value

ترتبط قيمة البيانات بالخاصية السابقة (Veracity) حيث تزيد قيمة البيانات كلما زادت دقتها، أيضاً تعتمد قيمة البيانات على مدة التي تحتاجها لمعالجتها هذه البيانات لأن قيمة النتائج لها عمر افتراضي فمثلاً: التأثر بمعالجة بيانات أسعار الأسهم لن يعطي قيمة أو فائدة تساعدنا باتخاذ قرار لتداول الأسهم أم لا.

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.



شكراً لكم