機器學習的基礎圖表!

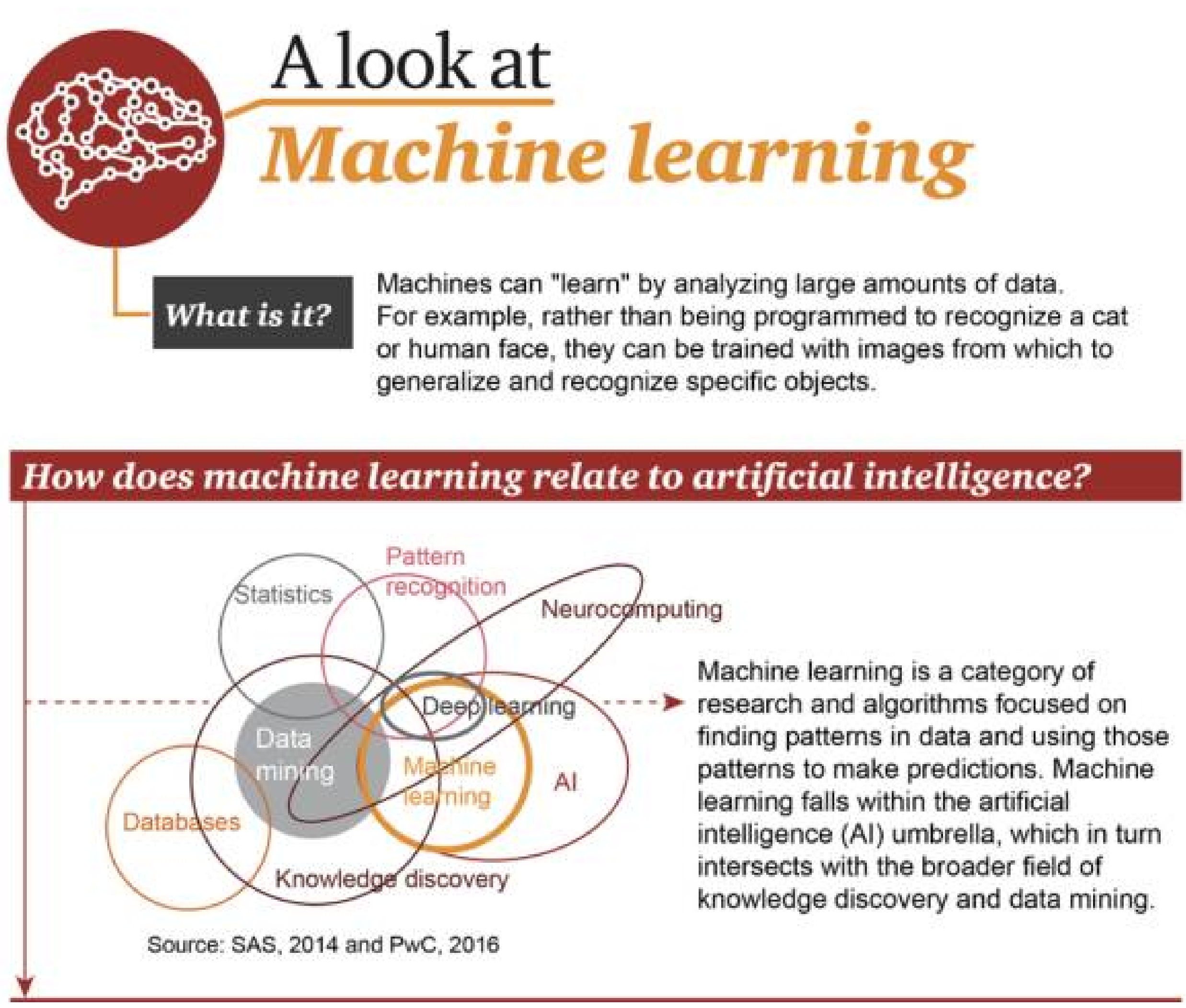
深度學習算法與計算機視覺 2022-03-28 23:49

公眾號關注 “ DL-CVer ”
設為“星標”，DLCV消息即可送達！

作者| Alan Morrison
轉自| 機器之心

> *四大會計師事務所之一的普華永道（PwC）發布了多份解讀機器學習基礎的圖表，其中介紹了機器學習的基本概念、原理、歷史、未來趨勢和一些常見的算法。為便於讀者閱讀，機器之心對這些圖表進行了編譯和拆分，分三大部分對這些內容進行了呈現，希望能幫助你進一步擴展閱讀。*

一、機器學習概覽



A look at

Machine learning

What is it?

Machines can "learn" by analyzing large amounts of data. For example, rather than being programmed to recognize a cat or human face, they can be trained with images from which to generalize and recognize specific objects.

How does machine learning relate to artificial intelligence?

Machine learning is a category of research and algorithms focused on finding patterns in data and using those patterns to make predictions. Machine learning falls within the artificial intelligence (AI) umbrella, which in turn intersects with the broader field of knowledge discovery and data mining.

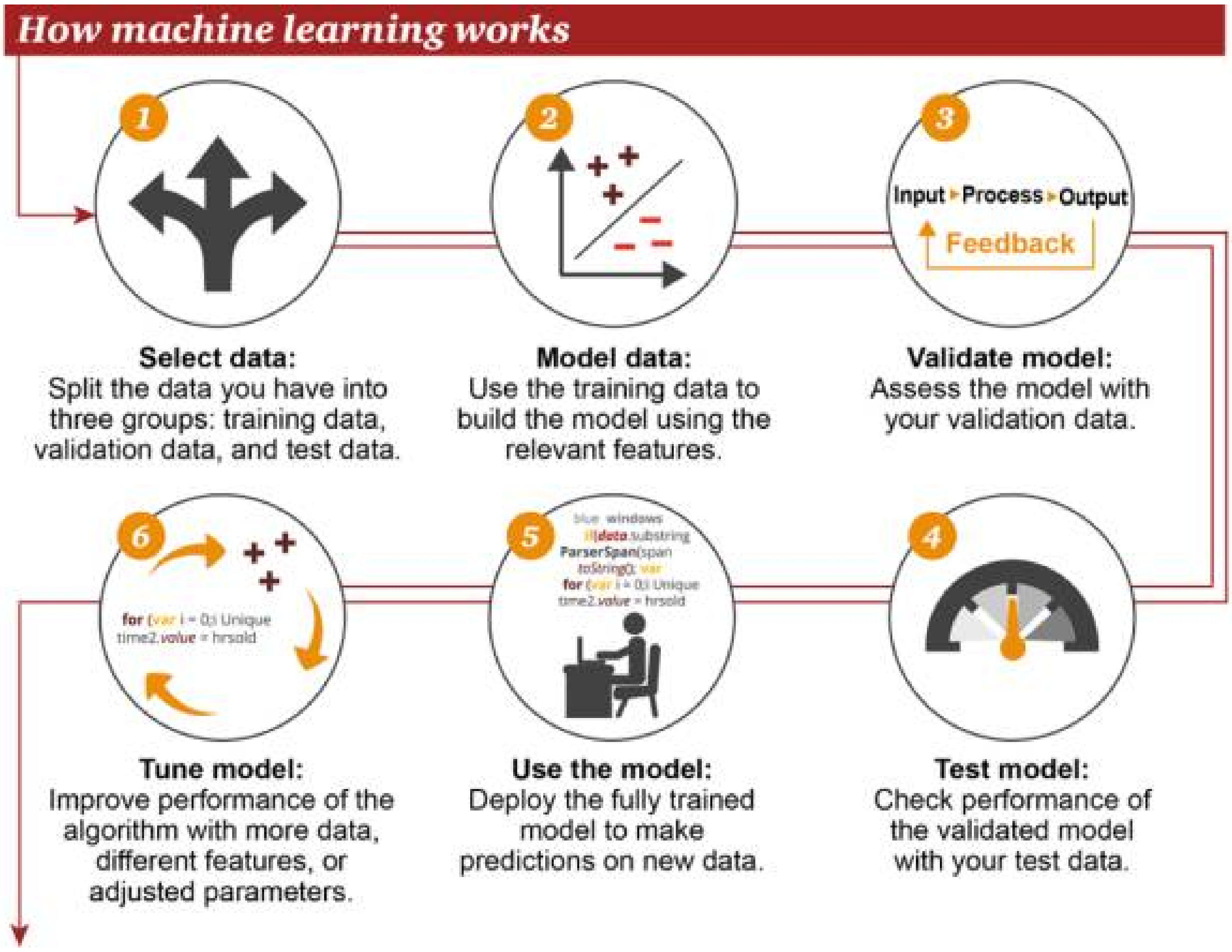
Source: SAS, 2014 and PwC, 2016

1. 什麼是機器學習?

機器通過分析大量數據來進行學習。比如說，不需要通過編程來識別貓或人臉，它們可以通過使用圖片來進行訓練，從而歸納和識別特定的目標。

2. 機器學習和人工智能的關係

機器學習是一種重在尋找數據中的模式並使用這些模式來做出預測的研究和算法的門類。機器學習是人工智能領域的一部分，並且和知識發現與數據挖掘有所交集。



How machine learning works

Select data: Split the data you have into three groups: training data, validation data, and test data.

Model data: Use the training data to build the model using the relevant features.

Validate model: Assess the model with your validation data.

Test model: Check performance of the validated model with your test data.

Use the model: Deploy the fully trained model to make predictions on new data.

Tune model: Improve performance of the algorithm with more data, different features, or adjusted parameters.

3. 機器學習的工作方式

①選擇數據：將你的數據分成三組：訓練數據、驗證數據和測試數據

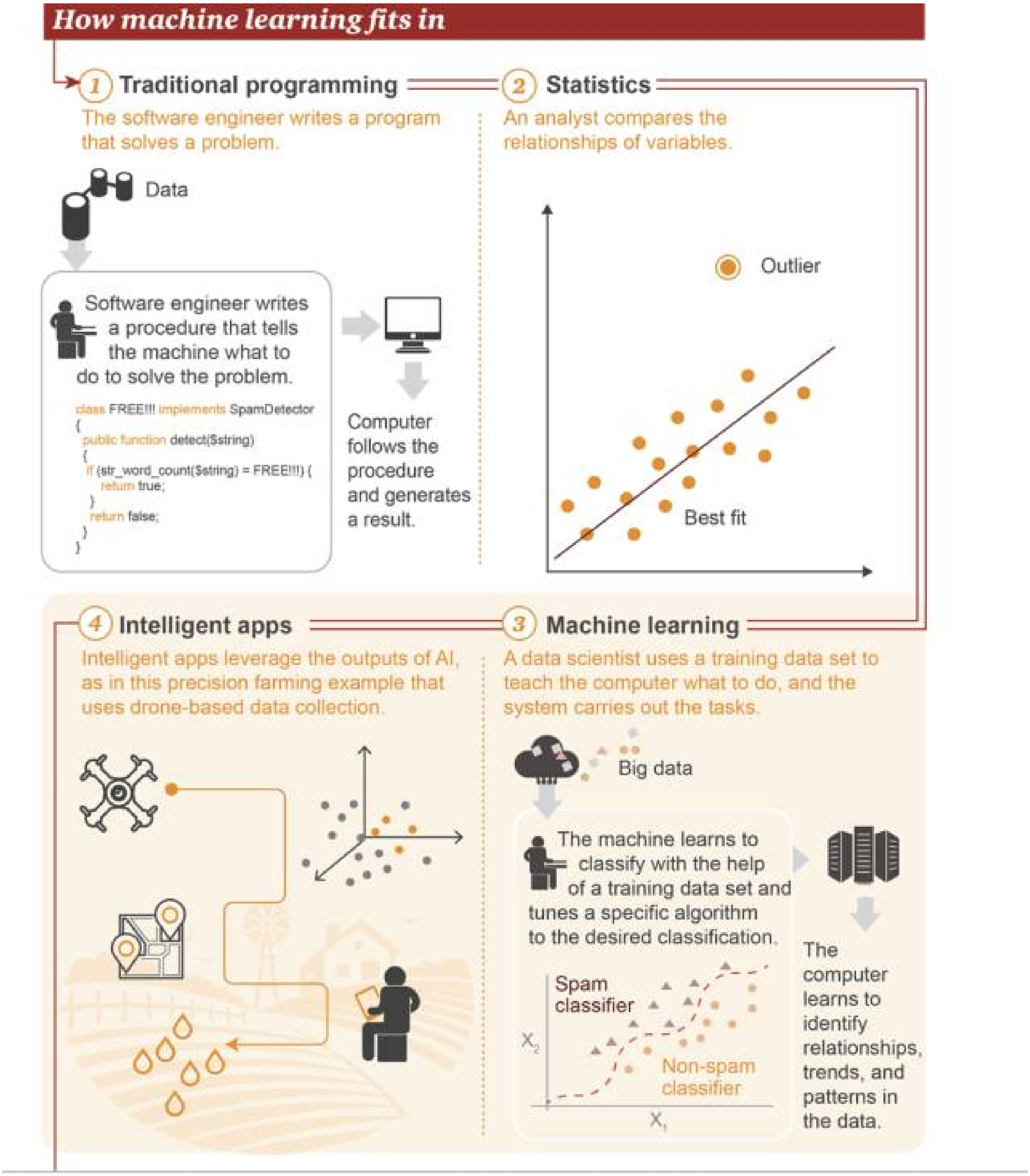
②模型數據：使用訓練數據來構建使用相關特徵的模型

③驗證模型：使用你的驗證數據接入你的模型

④測試模型：使用你的測試數據檢查被驗證的模型的表現

⑤使用模型：使用完全訓練好的模型在新數據上做預測

⑥調優模型：使用更多數據、不同的特徵或調整過的參數來提升算法的性能表現



How machine learning fits in

1 Traditional programming

The software engineer writes a program that solves a problem.

Software engineer writes a procedure that tells the machine what to do to solve the problem.

```
class FREE!!! implements SpamDetector
{
 public function detect($string)
 {
  if (str_word_count($string) = FREE!!!) {
   return true;
  }
  return false;
 }
}
```

Computer follows the procedure and generates a result.

2 Statistics

An analyst compares the relationships of variables.

3 Machine learning

A data scientist uses a training data set to teach the computer what to do, and the system carries out the tasks.

The machine learns to classify with the help of a training data set and tunes a specific algorithm to the desired classification.

The computer learns to identify relationships, trends, and patterns in the data.

4 Intelligent apps

Intelligent apps leverage the outputs of AI, as in this precision farming example that uses drone-based data collection.

4. 機器學習所處的位置

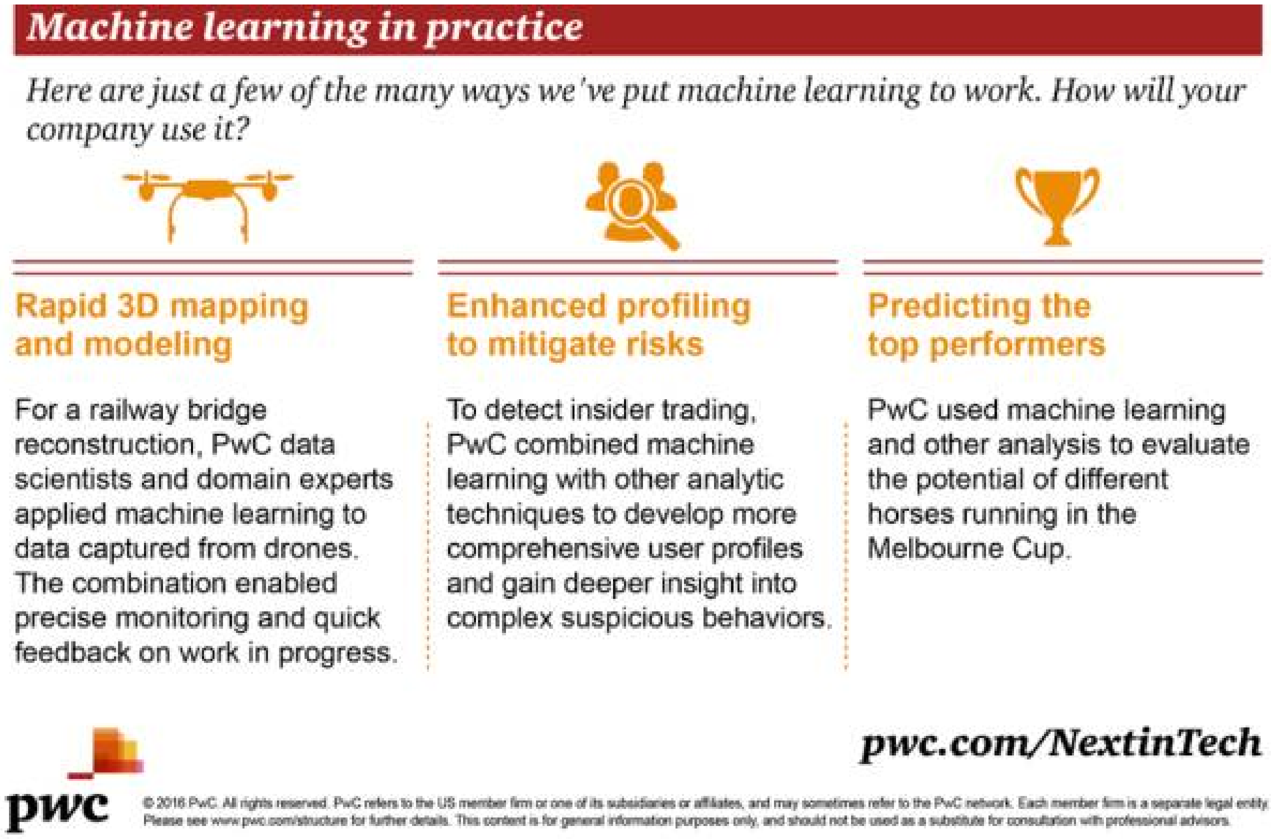
①傳統編程：軟件工程師編寫程序來解決問題。首先存在一些數據→為了解決一個問題，軟件工程師編寫一個流程來告訴機器應該怎樣做→計算機遵照這一流程執行，然後得出結果

②統計學：分析師比較變量之間的關係

③機器學習：數據科學家使用訓練數據集來教計算機應該怎麼做，然後系統執行該任務。首先存在大數據→機器會學習使用訓練數據集來進行分類，調節特定的算法來實現目標分類→該計算機可學習識別數據中的關

係、趨勢和模式

④智能應用：智能應用使用人工智能所得到的結果，如圖是一個精準農業的應用案例示意，該應用基於無人機所收集到的數據



Machine learning in practice

Here are just a few of the many ways we've put machine learning to work. How will your company use it?

Rapid 3D mapping and modeling

For a railway bridge reconstruction, PwC data scientists and domain experts applied machine learning to data captured from drones. The combination enabled precise monitoring and quick feedback on work in progress.

Enhanced profiling to mitigate risks

To detect insider trading, PwC combined machine learning with other analytic techniques to develop more comprehensive user profiles and gain deeper insight into complex suspicious behaviors.

Predicting the top performers

PwC used machine learning and other analysis to evaluate the potential of different horses running in the Melbourne Cup.

pwc.com/NextinTech

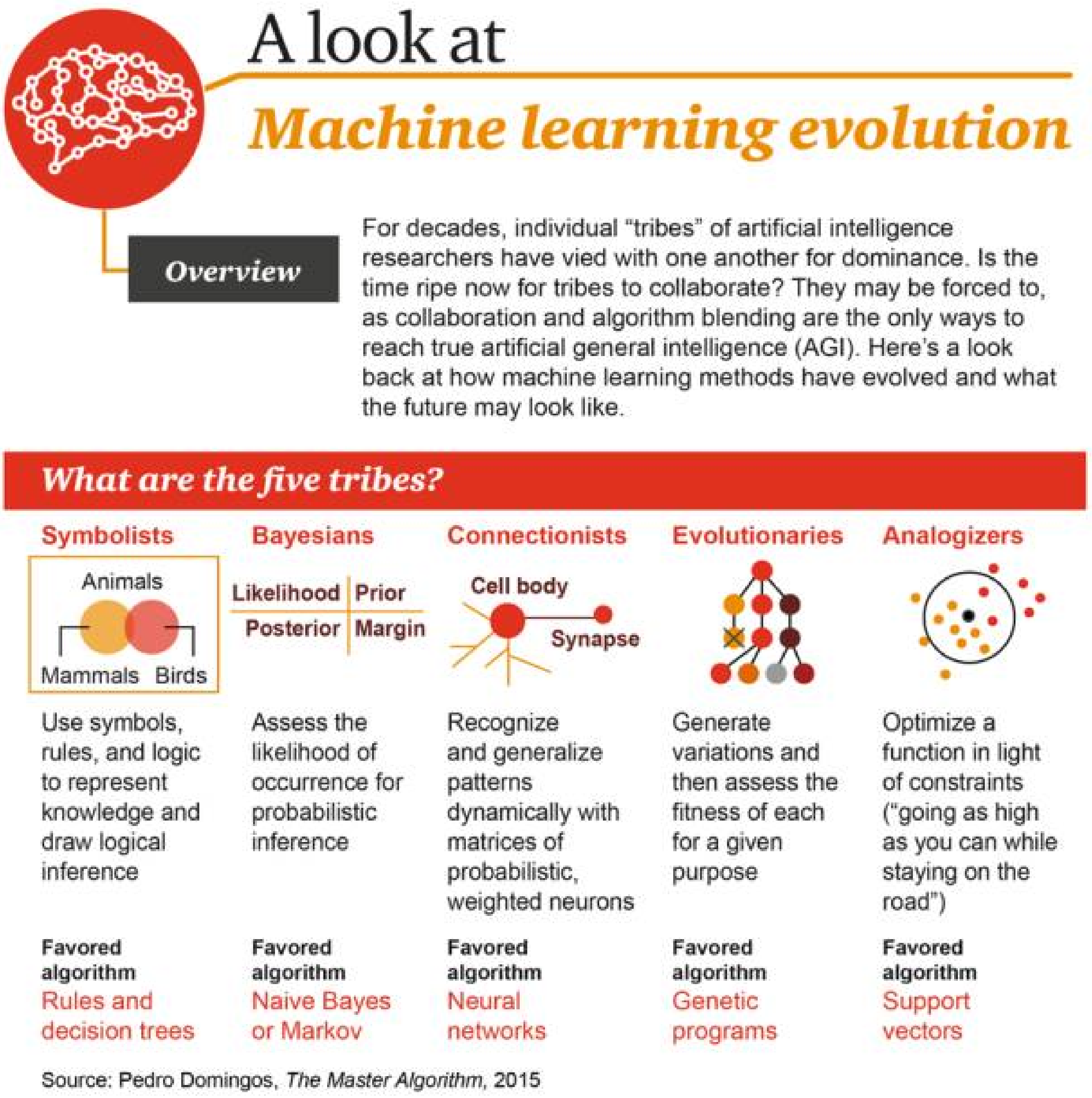
pwc

5. 機器學習的實際應用

機器學習有很多應用場景，這裡給出了一些示例，你會怎麼使用它?

- 快速三維地圖測繪和建模：要建造一架鐵路橋，PwC 的數據科學家和領域專家將機器學習應用到了無人機收集到的數據上。這種組合實現了工作成功中的精準監控和快速反饋。
- 增強分析以降低風險：為了檢測內部交易，PwC 將機器學習和其它分析技術結合了起來，從而開發了更為全面的用戶概況，並且獲得了對複雜可疑行為的更深度了解。
- 預測表現最佳的目標：PwC 使用機器學習和其它分析方法來評估Melbourne Cup 賽場上不同賽馬的潛力。

二、機器學習的演化



A look at
Machine learning evolution

Overview

For decades, individual "tribes" of artificial intelligence researchers have vied with one another for dominance. Is the time ripe now for tribes to collaborate? They may be forced to, as collaboration and algorithm blending are the only ways to reach true artificial general intelligence (AGI). Here's a look back at how machine learning methods have evolved and what the future may look like.

What are the five tribes?

Symbolists	Bayesians	Connectionists	Evolutionaries	Analogizers
Use symbols, rules, and logic to represent knowledge and draw logical inference	Assess the likelihood of occurrence for probabilistic inference	Recognize and generalize patterns dynamically with matrices of probabilistic, weighted neurons	Generate variations and then assess the fitness of each for a given purpose	Optimize a function in light of constraints ("going as high as you can while staying on the road")
Favored algorithm Rules and decision trees	**Favored algorithm** Naive Bayes or Markov	**Favored algorithm** Neural networks	**Favored algorithm** Genetic programs	**Favored algorithm** Support vectors

Source: Pedro Domingos, *The Master Algorithm*, 2015

幾十年來，人工智能研究者的各個「部落」一直以來都在彼此爭奪主導權。現在是這些部落聯合起來的時候了嗎？他們也可能不得不這樣做，因為合作和算法融合是實現真正通用人工智能（AGI）的唯一方式。這裡給出了機器學習方法的演化之路以及未來的可能模樣。

1. 五大流派

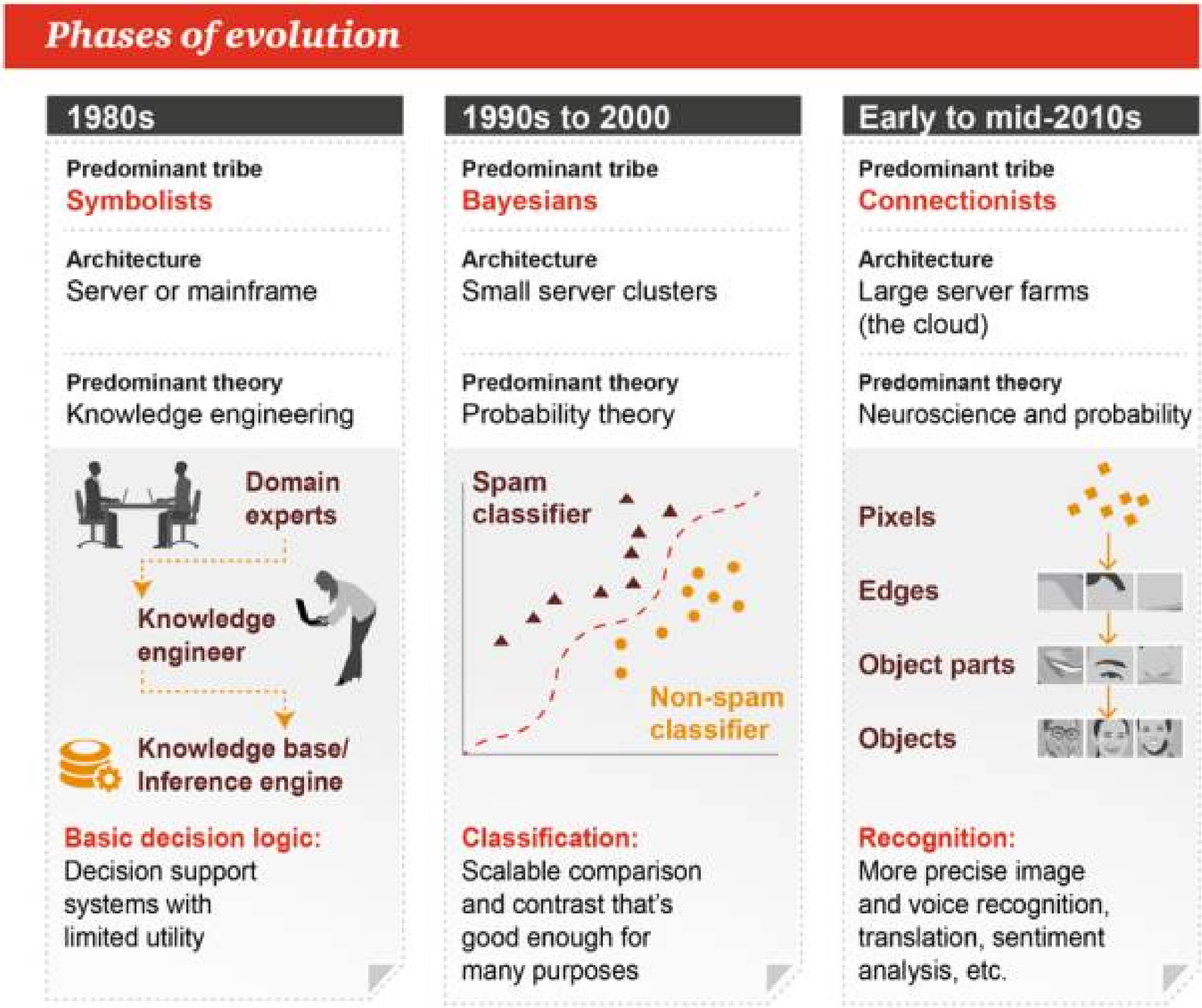
①符號主義：使用符號、規則和邏輯來表徵知識和進行邏輯推理，最喜歡的算法是：規則和決策樹

②貝葉斯派：獲取發生的可能性來進行概率推理，最喜歡的算法是：樸素貝葉斯或馬爾可夫

③聯結主義：使用概率矩陣和加權神經元來動態地識別和歸納模式，最喜歡的算法是：神經網絡

④進化主義：生成變化，然後為特定目標獲取其中最優的，最喜歡的算法是：遺傳算法

⑤Analogizer：根據約束條件來優化函數（盡可能走到更高，但同時不要離開道路），最喜歡的算法是：支持向量機



Phases of evolution

1980s	1990s to 2000	Early to mid-2010s
Predominant tribe Symbolists	Predominant tribe Bayesians	Predominant tribe Connectionists
Architecture Server or mainframe	Architecture Small server clusters	Architecture Large server farms (the cloud)
Predominant theory Knowledge engineering	Predominant theory Probability theory	Predominant theory Neuroscience and probability
Basic decision logic: Decision support systems with limited utility	Classification: Scalable comparison and contrast that's good enough for many purposes	Recognition: More precise image and voice recognition, translation, sentiment analysis, etc.

2. 演化的階段

1980 年代

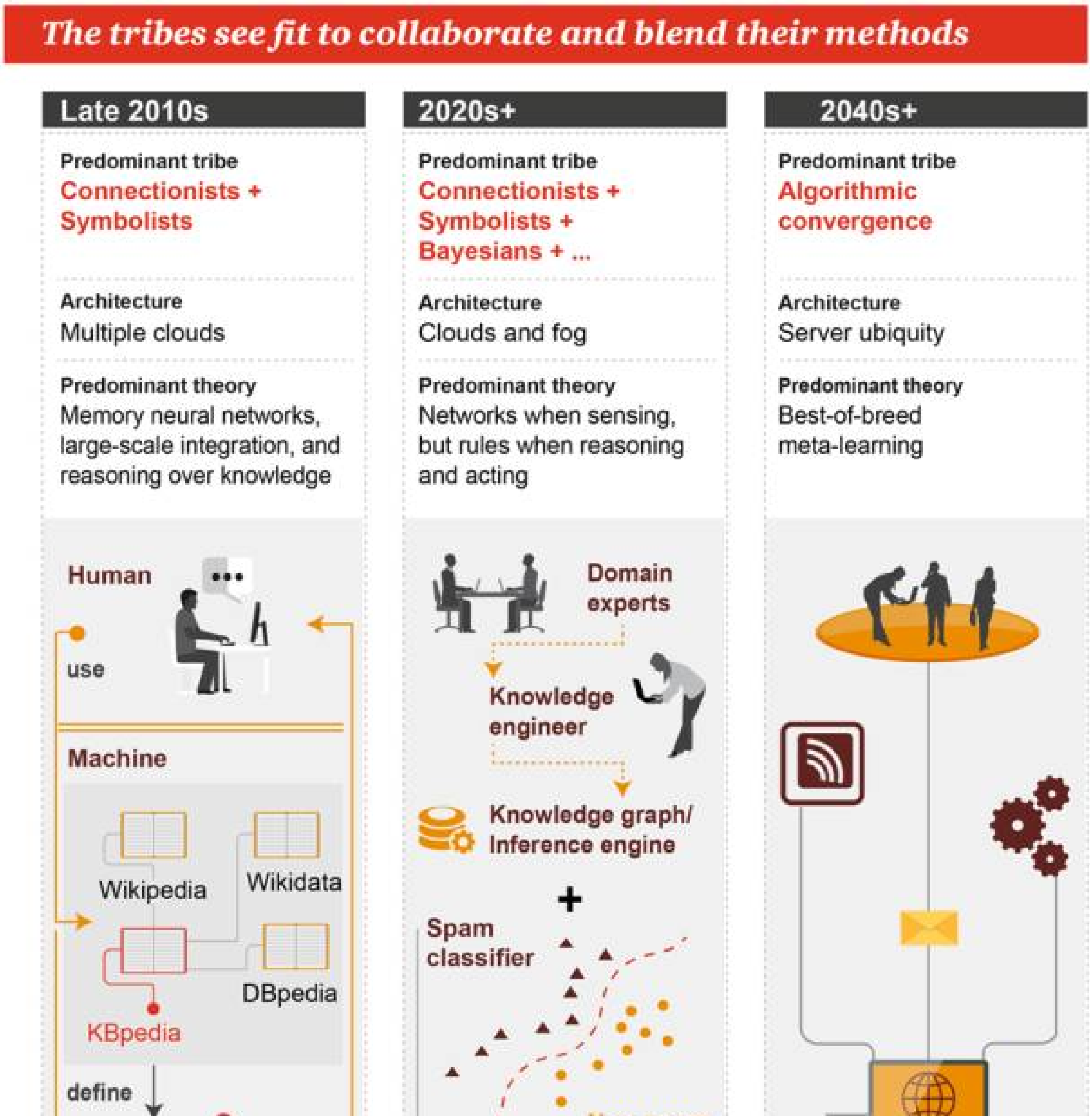
- 主導流派：符號主義
- 架構：服務器或大型機
- 主導理論：知識工程
- 基本決策邏輯：決策支持系統，實用性有限

1990 年代到2000 年

- 主導流派：貝葉斯
- 架構：小型服務器集群
- 主導理論：概率論
- 分類：可擴展的比較或對比，對許多任務都足夠好了

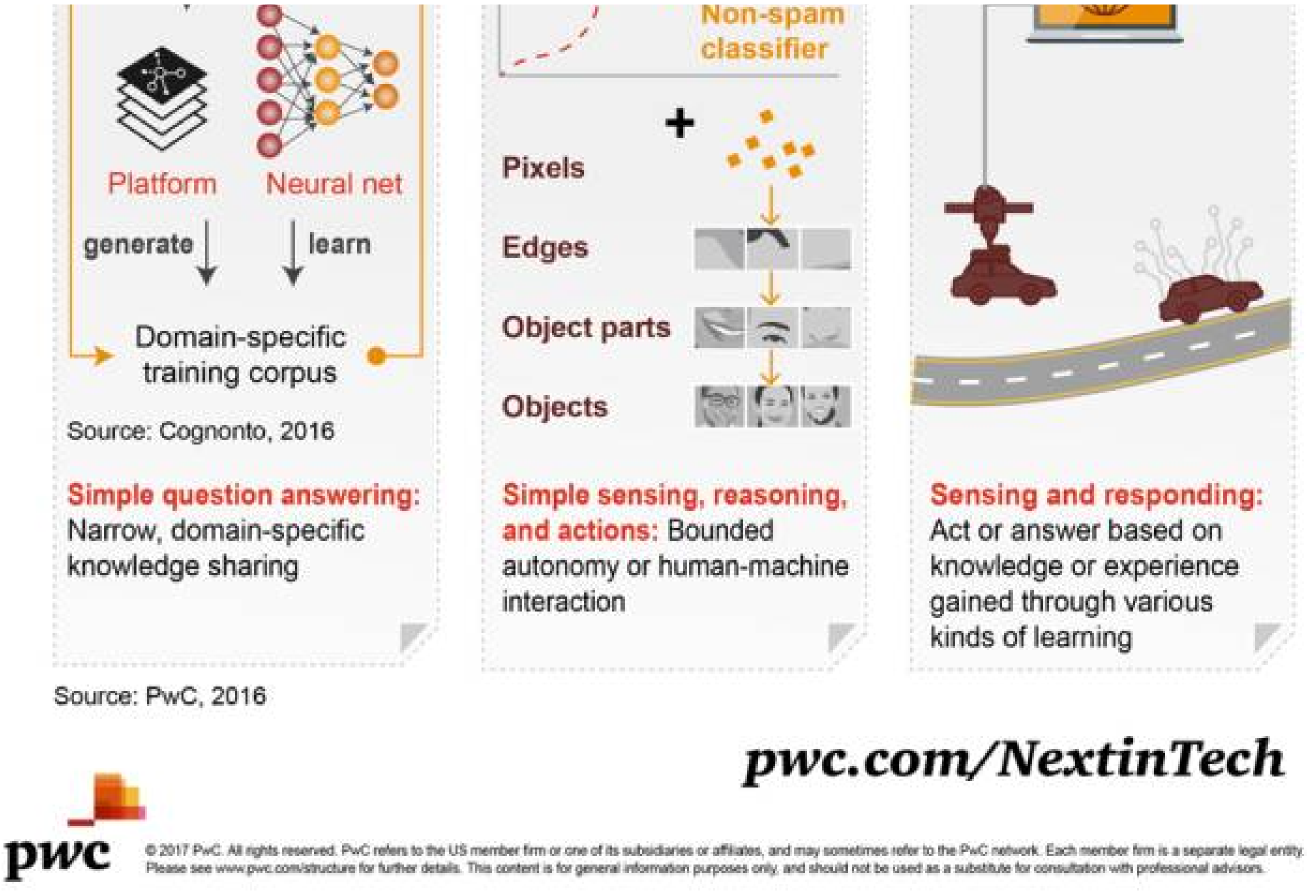
2010 年代早期到中期

- 主導流派：聯結主義
- 架構：大型服務器農場
- 主導理論：神經科學和概率
- 識別：更加精準的圖像和聲音識別、翻譯、情緒分析等



The tribes see fit to collaborate and blend their methods

Late 2010s	2020s+	2040s+
Predominant tribe Connectionists + Symbolists	Predominant tribe Connectionists + Symbolists + Bayesians + ...	Predominant tribe Algorithmic convergence
Architecture Multiple clouds	Architecture Clouds and fog	Architecture Server ubiquity
Predominant theory Memory neural networks, large-scale integration, and reasoning over knowledge	Predominant theory Networks when sensing, but rules when reasoning and acting	Predominant theory Best-of-breed meta-learning


Platform
Neural net
generate
learn
Domain-specific training corpus
Source: Cognonto, 2016
Simple question answering: Narrow, domain-specific knowledge sharing
Non-spam classifier
Pixels
Edges
Object parts
Objects
Simple sensing, reasoning, and actions: Bounded autonomy or human-machine interaction
Sensing and responding: Act or answer based on knowledge or experience gained through various kinds of learning
Source: PwC, 2016
pwc.com/NextinTech
pwc
© 2017 PwC. All rights reserved. PwC refers to the US member firm or one of its subsidiaries or affiliates, and may sometimes refer to the PwC network. Each member firm is a separate legal entity. Please see www.pwc.com/structure for further details. This content is for general information purposes only, and should not be used as a substitute for consultation with professional advisors.

3. 這些流派有望合作，並將各自的方法融合到一起

2010 年代末期

- 主導流派：聯結主義+符號主義
- 架構：許多雲
- 主導理論：記憶神經網絡、大規模集成、基於知識的推理
- 簡單的問答：範圍狹窄的、領域特定的知識共享

2020 年代+

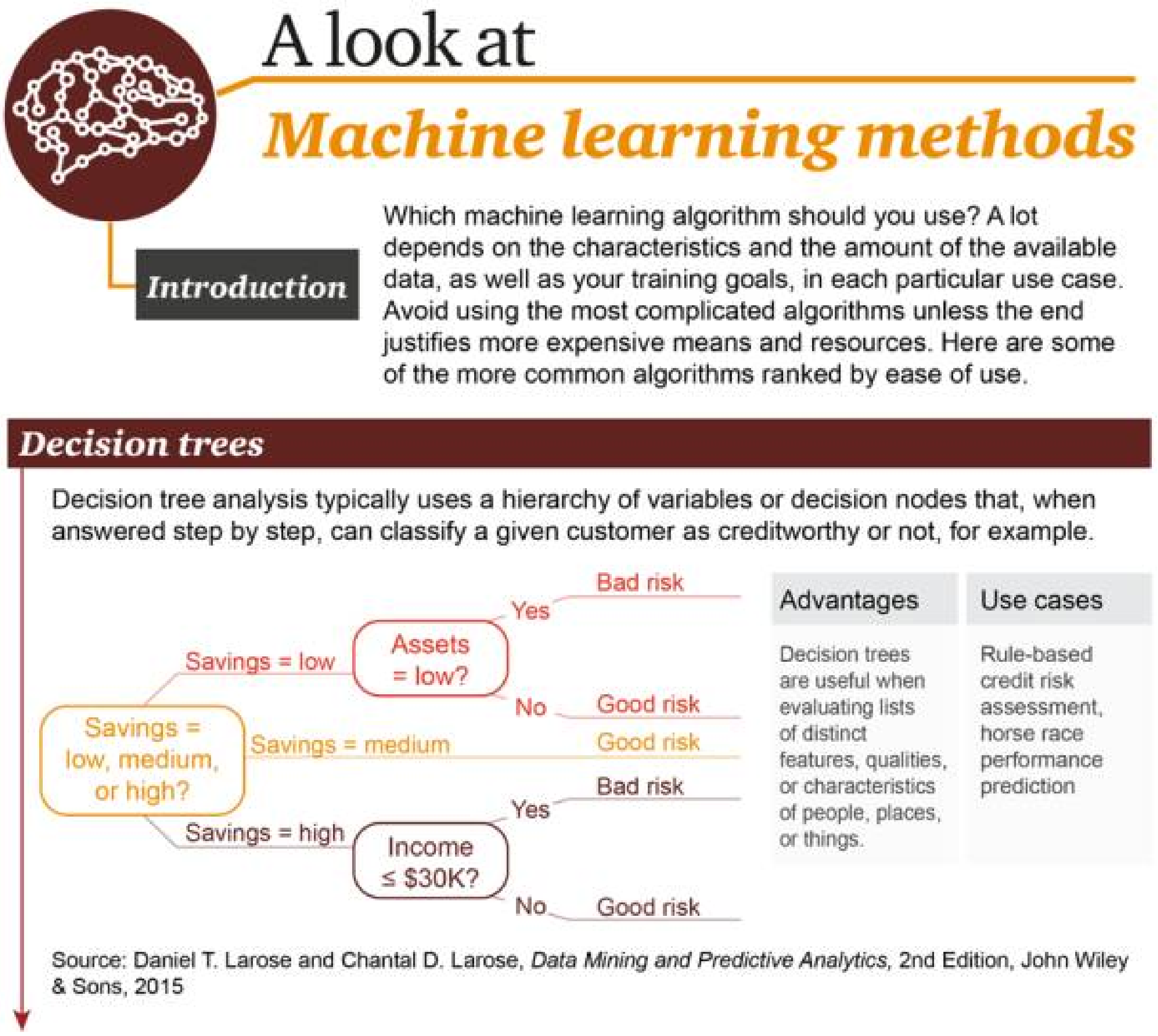
- 主導流派：聯結主義+符號主義+貝葉斯+……
- 架構：雲計算和霧計算
- 主導理論：感知的時候有網絡，推理和工作的時候有規則
- 簡單感知、推理和行動：有限制的自動化或人機交互

2040 年代+

- 主導流派：算法融合
- 架構：無處不在的服務器

- 主導理論：最佳組合的元學習
- 感知和響應：基於通過多種學習方式獲得的知識或經驗採取行動或做出回答

三、機器學習的算法



A look at
Machine learning methods

Introduction

Which machine learning algorithm should you use? A lot depends on the characteristics and the amount of the available data, as well as your training goals, in each particular use case. Avoid using the most complicated algorithms unless the end justifies more expensive means and resources. Here are some of the more common algorithms ranked by ease of use.

Decision trees

Decision tree analysis typically uses a hierarchy of variables or decision nodes that, when answered step by step, can classify a given customer as creditworthy or not, for example.

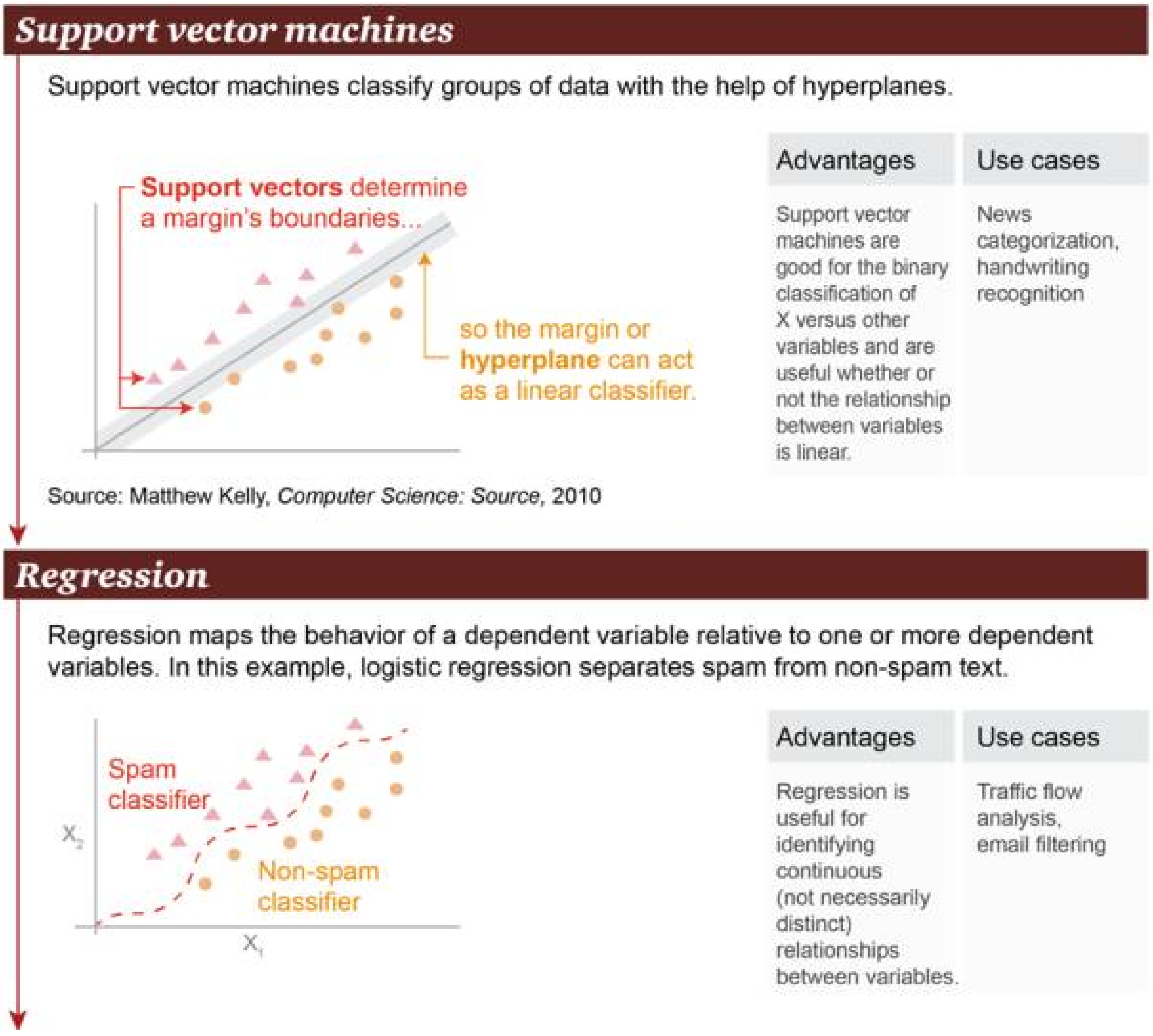
Advantages	Use cases
Decision trees are useful when evaluating lists of distinct features, qualities, or characteristics of people, places, or things.	Rule-based credit risk assessment, horse race performance prediction

Source: Daniel T. Larose and Chantal D. Larose, *Data Mining and Predictive Analytics*, 2nd Edition, John Wiley & Sons, 2015

你應該使用哪種機器學習算法？這在很大程度上依賴於可用數據的性質和數量以及每一個特定用例中你的訓練目標。不要使用最複雜的算法，除非其結果值得付出昂貴的開銷和資源。這裡給出了一些最常見的算法，按使用簡單程度排序。

1. 決策樹（Decision Tree）：在進行逐步應答過程中，典型的決策樹分析會使用分層變量或決策節點，例如，可將一個給定用戶分類成信用可靠或不可靠。

- 優點：擅長對人、地點、事物的一系列不同特徵、品質、特性進行評估
- 場景舉例：基於規則的信用評估、賽馬結果預測



Support vector machines

Support vector machines classify groups of data with the help of hyperplanes.

Support vectors determine a margin's boundaries...
so the margin or hyperplane can act as a linear classifier.

Advantages	Use cases
Support vector machines are good for the binary classification of X versus other variables and are useful whether or not the relationship between variables is linear.	News categorization, handwriting recognition

Source: Matthew Kelly, *Computer Science: Source*, 2010

Regression

Regression maps the behavior of a dependent variable relative to one or more dependent variables. In this example, logistic regression separates spam from non-spam text.

Advantages	Use cases
Regression is useful for identifying continuous (not necessarily distinct) relationships between variables.	Traffic flow analysis, email filtering

2. 支持向量機（Support Vector Machine）：基於超平面（hyperplane），支持向量機可以對數據群進行分類。

- 優點：支持向量機擅長在變量X 與其它變量之間進行二元分類操作，無論其關係是否是線性的
- 場景舉例：新聞分類、手寫識別。

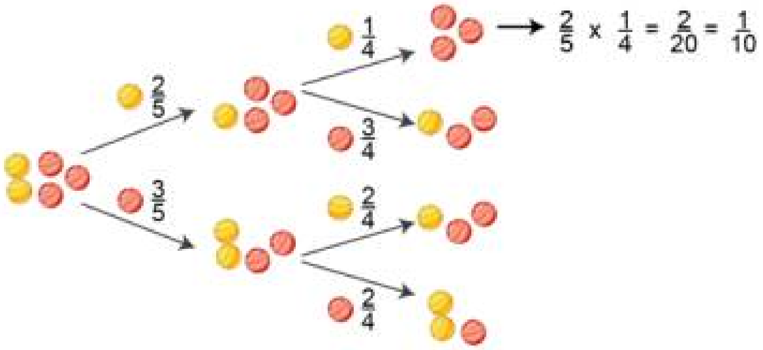
3. 回歸（Regression）：回歸可以勾畫出因變量與一個或多個因變量之間的狀態關係。在這個例子中，將垃圾郵件和非垃圾郵件進行了區分。

- 優點：回歸可用於識別變量之間的連續關係，即便這個關係不是非常明顯

- 場景舉例：路面交通流量分析、郵件過濾

Naive Bayes classification

Naive Bayes classifiers compute probabilities, given tree branches of possible conditions. Each individual feature is “naive” or conditionally independent of, and therefore does not influence, the others. For example, what’s the probability you would draw two yellow marbles in a row, given a jar of five yellow and red marbles total? The probability, following the topmost branch of two yellow in a row, is one in ten. Naive Bayes classifiers compute the combined, conditional probabilities of multiple attributes.

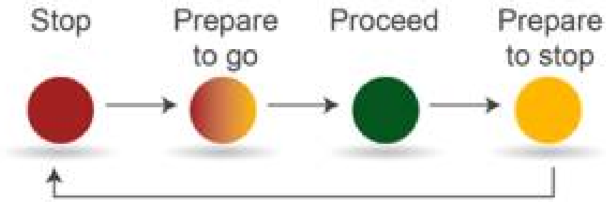


Advantages	Use cases
Naive Bayes methods allow the quick classification of relevant items in small data sets that have distinct features.	Sentiment analysis, consumer segmentation

Source: Rod Pierce, et al., *MathIsFun*, 2014

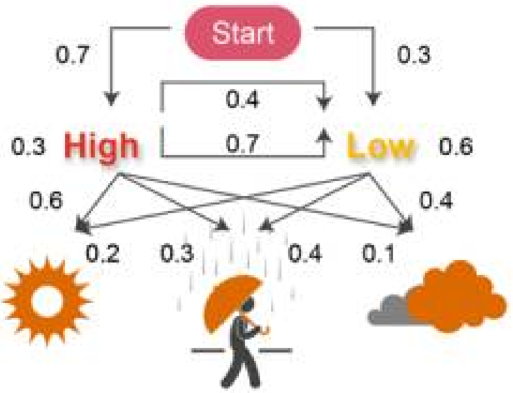
Hidden Markov models

Observable Markov processes are purely deterministic—one given state always follows another given state. Traffic light patterns are an example.



Source: Derek Kane, 2015

Hidden Markov models, by contrast, compute the probability of hidden states occurring by analyzing observable data, and then estimating the likely pattern of future observation with the help of the hidden state analysis. In this example, the probability of high or low pressure (the hidden state) is used to predict the likelihood of sunny, rainy, or cloudy weather.



Advantages	Use cases
Tolerates data variability and effective for recognition and prediction.	Facial expression analysis, weather prediction

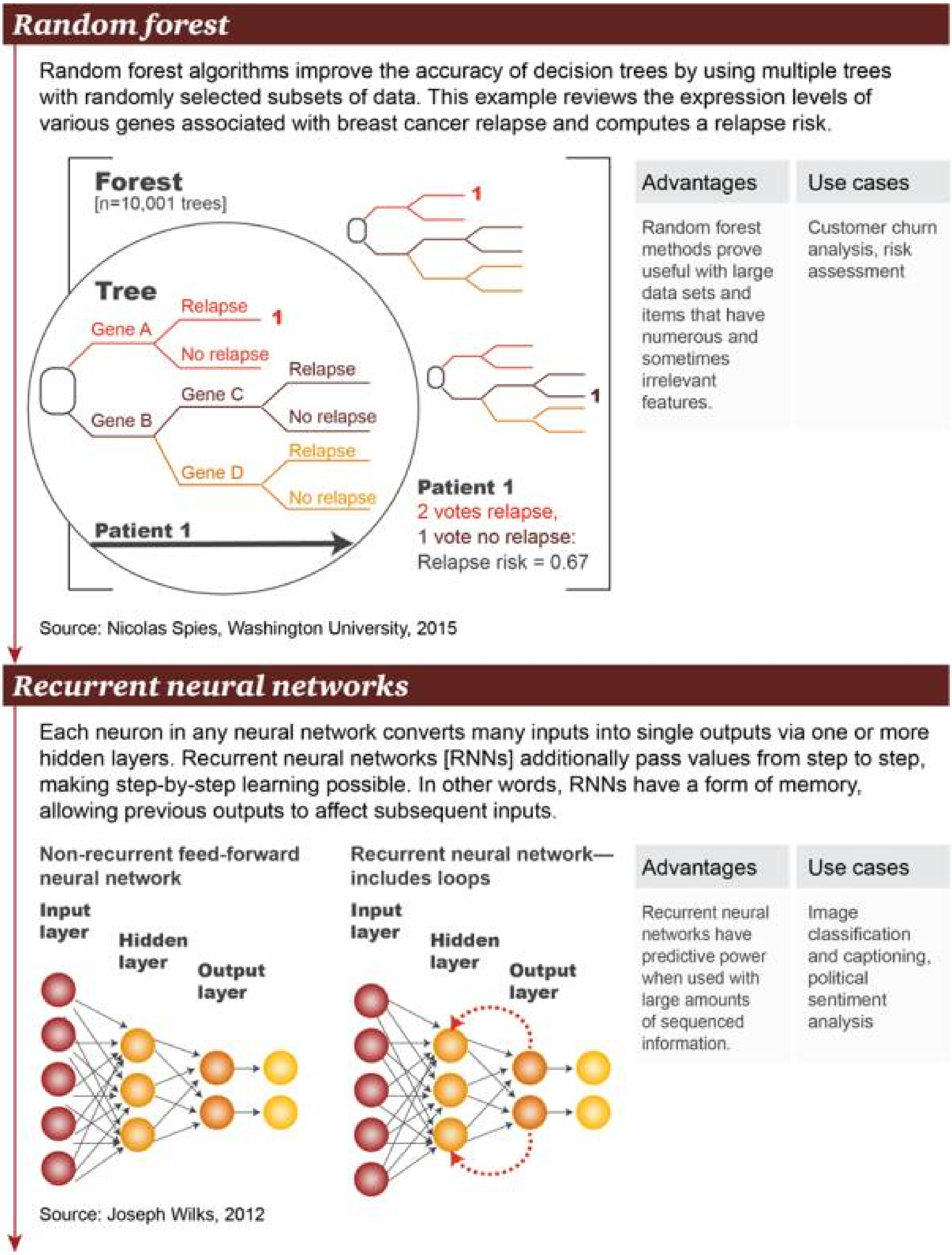
Source: Leonardo Guizzetti, 2012

4. 樸素貝葉斯分類（Naive Bayes Classification）：樸素貝葉斯分類器用於計算可能條件的分支概率。每個獨立的特徵都是「樸素」或條件獨立的，因此它們不會影響別的對象。例如，在一個裝有共5 個黃色和紅色小球的罐子裡，連續拿到兩個黃色小球的概率是多少？從圖中最上方分支可見，前後抓取兩個黃色小球的概率為1/10。樸素貝葉斯分類器可以計算多個特徵的聯合條件概率。

- 優點：對於在小數據集上有顯著特徵的相關對象，樸素貝葉斯方法可對其進行快速分類
- 場景舉例：情感分析、消費者分類

5. 隱馬爾可夫模型（Hidden Markov model）：顯馬爾可夫過程是完全確定性的——一個給定的狀態經常會伴隨另一個狀態。交通信號燈就是一個例子。相反，隱馬爾可夫模型通過分析可見數據來計算隱藏狀態的發生。隨後，借助隱藏狀態分析，隱馬爾可夫模型可以估計可能的未來觀察模式。在本例中，高或低氣壓的概率（這是隱藏狀態）可用於預測晴天、雨天、多雲天的概率。

- 優點：容許數據的變化性，適用於識別（recognition）和預測操作
- 場景舉例：面部表情分析、氣象預測



Random forest

Random forest algorithms improve the accuracy of decision trees by using multiple trees with randomly selected subsets of data. This example reviews the expression levels of various genes associated with breast cancer relapse and computes a relapse risk.

Patient 1
2 votes relapse,
1 vote no relapse:
Relapse risk = 0.67

Advantages	Use cases
Random forest methods prove useful with large data sets and items that have numerous and sometimes irrelevant features.	Customer churn analysis, risk assessment

Source: Nicolas Spies, Washington University, 2015

Recurrent neural networks

Each neuron in any neural network converts many inputs into single outputs via one or more hidden layers. Recurrent neural networks [RNNs] additionally pass values from step to step, making step-by-step learning possible. In other words, RNNs have a form of memory, allowing previous outputs to affect subsequent inputs.

Non-recurrent feed-forward neural network
Input layer
Hidden layer
Output layer
Recurrent neural network—includes loops
Input layer
Hidden layer
Output layer

Advantages	Use cases
Recurrent neural networks have predictive power when used with large amounts of sequenced information.	Image classification and captioning, political sentiment analysis

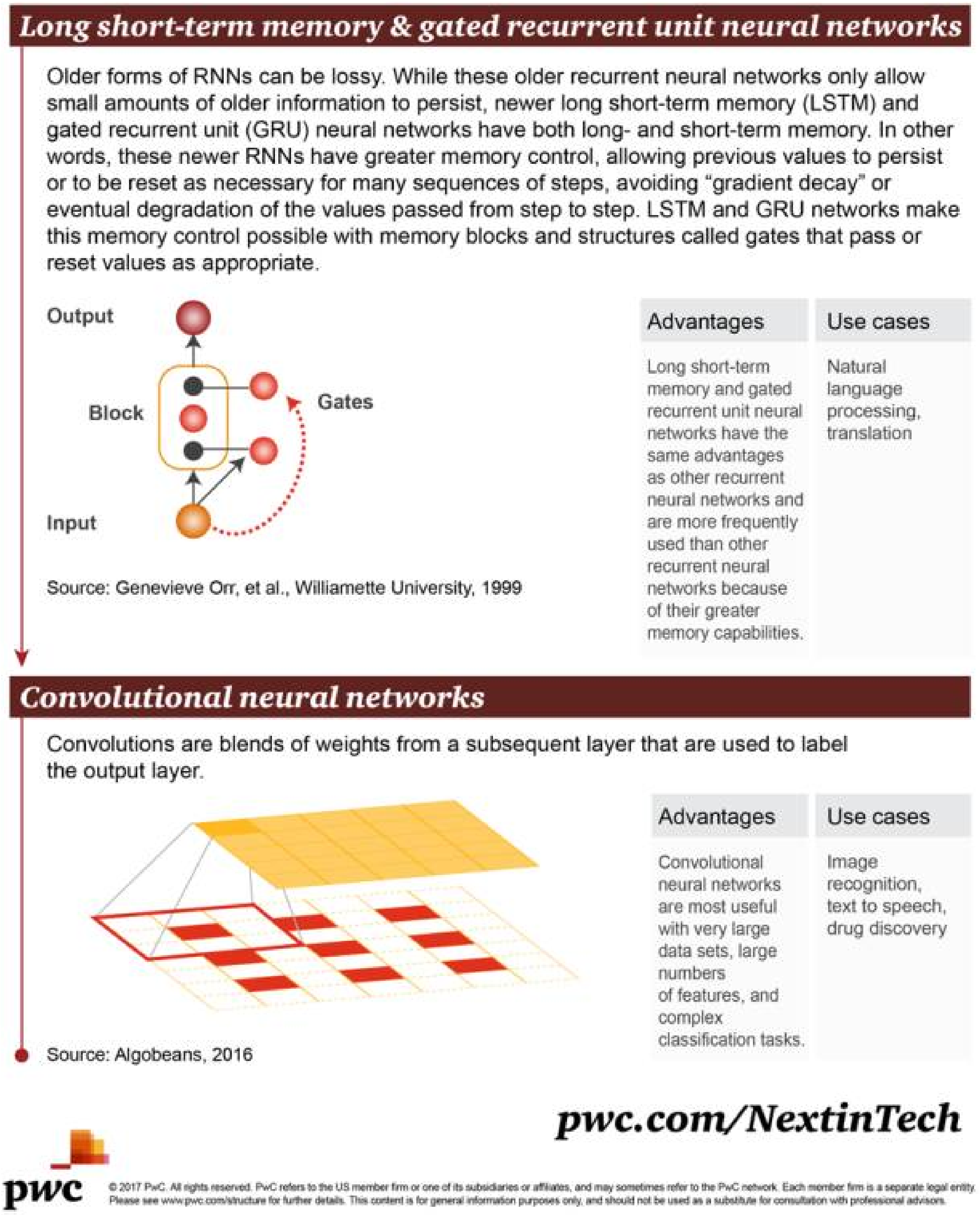
Source: Joseph Wilks, 2012

6. 隨機森林（Random forest）：隨機森林算法通過使用多個帶有隨機選取的數據子集的樹（tree）改善了決策樹的精確性。本例在基因表達層面上考察了大量與乳腺癌復發相關的基因，併計算出複發風險。

- 優點：隨機森林方法被證明對大規模數據集和存在大量且有時不相關特徵的項（item）來說很有用
- 場景舉例：用戶流失分析、風險評估

7. 循環神經網絡（Recurrent neural network）：在任意神經網絡中，每個神經元都通過1 個或多個隱藏層來將很多輸入轉換成單個輸出。循環神經網絡（RNN）會將值進一步逐層傳遞，讓逐層學習成為可能。換句話說，RNN 存在某種形式的記憶，允許先前的輸出去影響後面的輸入。

- 優點：循環神經網絡在存在大量有序信息時具有預測能力
- 場景舉例：圖像分類與字幕添加、政治情感分析



Long short-term memory & gated recurrent unit neural networks

Older forms of RNNs can be lossy. While these older recurrent neural networks only allow small amounts of older information to persist, newer long short-term memory (LSTM) and gated recurrent unit (GRU) neural networks have both long- and short-term memory. In other words, these newer RNNs have greater memory control, allowing previous values to persist or to be reset as necessary for many sequences of steps, avoiding "gradient decay" or eventual degradation of the values passed from step to step. LSTM and GRU networks make this memory control possible with memory blocks and structures called gates that pass or reset values as appropriate.

Source: Genevieve Orr, et al., Williamette University, 1999

Advantages	Use cases
Long short-term memory and gated recurrent unit neural networks have the same advantages as other recurrent neural networks and are more frequently used than other recurrent neural networks because of their greater memory capabilities.	Natural language processing, translation

Convolutional neural networks

Convolutions are blends of weights from a subsequent layer that are used to label the output layer.

Source: Algobeans, 2016

Advantages	Use cases
Convolutional neural networks are most useful with very large data sets, large numbers of features, and complex classification tasks.	Image recognition, text to speech, drug discovery

pwc.com/NextinTech

pwc

8. 長短期記憶（Long short-term memory，LSTM）與門控循環單元神經網絡（gated recurrent unit nerual network）：早期的RNN 形式是會存在損耗的。儘管這些早期循環神經網絡只允許留存少量的早期信息，新近的長短期記憶（LSTM）與門控循環單元（GRU）神經網絡都有長期與短期的記憶。換句話說，這些新近的RNN 擁有更好的控制記憶的能力，允許保留早先的值或是當有必要處理很多系列步驟時重置這

些值，這避免了「梯度衰減」或逐層傳遞的值的最終degradation。LSTM 與GRU 網絡使得我們可以使用被稱為「門（gate）」的記憶模塊或結構來控制記憶，這種門可以在合適的時候傳遞或重置值。

- 優點：長短期記憶和門控循環單元神經網絡具備與其它循環神經網絡一樣的優點，但因為它們有更好的記憶能力，所以更常被使用
- 場景舉例：自然語言處理、翻譯

9. 卷積神經網絡（convolutional neural network）：卷積是指來自後續層的權重的融合，可用於標記輸出層。

- 優點：當存在非常大型的數據集、大量特徵和復雜的分類任務時，卷積神經網絡是非常有用的
- 場景舉例：圖像識別、文本轉語音、藥物發現