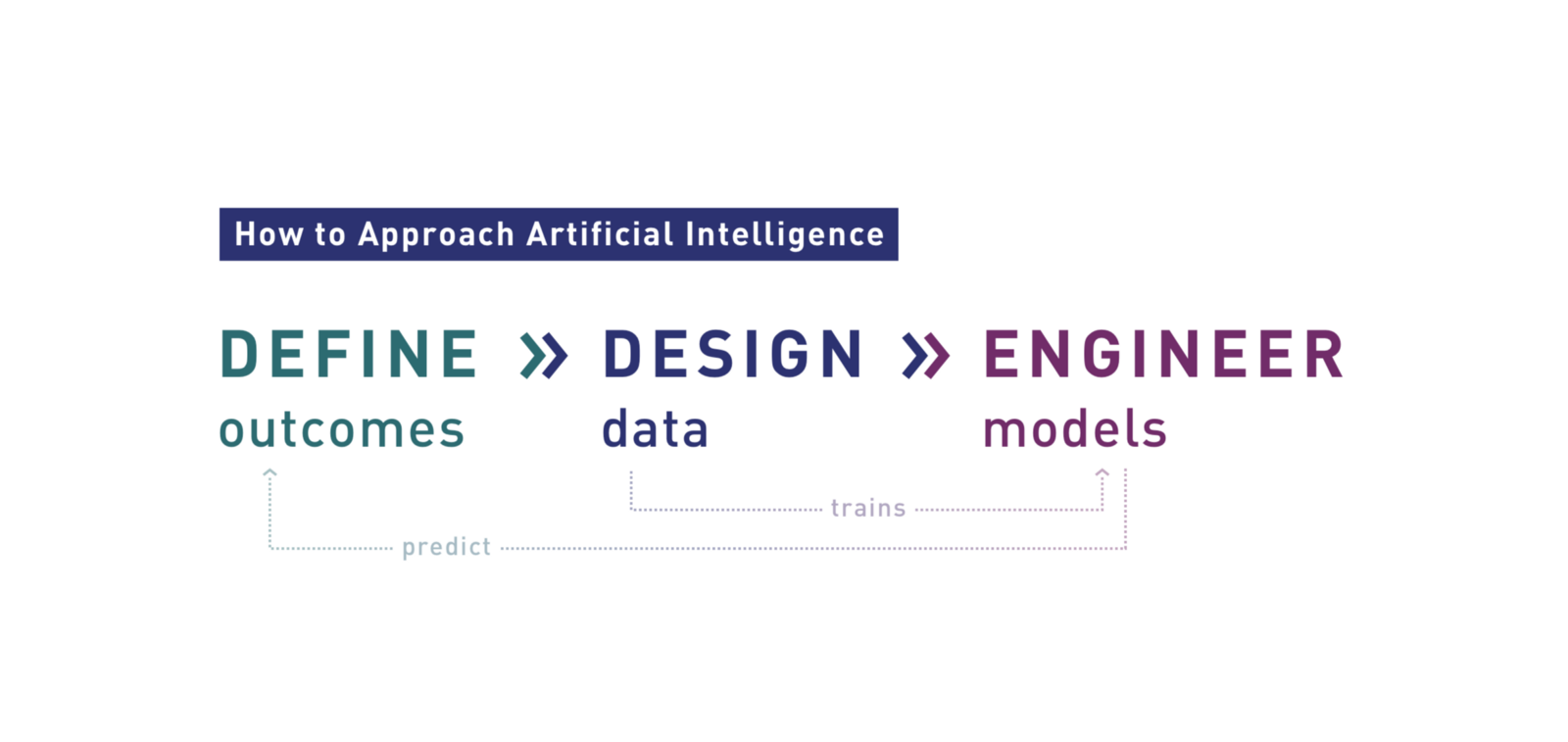
# 应用机器学习时人们忘记采取的两个步骤

对于某些应用，人类设计的人工智能系统没有任何学习能力。专家可以设计这样的系统并证明其正确性；人工智能的这些应用是成本效益和性能的最佳解决方案（例如，在确保人类安全方面）。相比之下，保证机器学习产生的决策策略的正确性是一个重大挑战。我们将深入调查原因。



#### 如何进行机器学习（ML）

随着企业听到更多关于人工智能（AI）和机器学习（ML）如何改变商业实践的故事，数据科学从业者被要求“应用机器学习”，只带模糊或高层次的商业目标。在实现人工智能和从数据中创造商业价值之间存在着巨大的差距，数据科学中的许多失望都是未能弥合这一差距的结果。本文试图向人工智能从业者和业务经理展示如何弥合这一差距。

你接近人工智能的顺序不同于大多数人的预期，应该遵循以下步骤：

1. 目标设想的解决方案
2. 围绕数据进行设计
3. 定义模型

大多数人忘记了前两个似乎会适得其反，不是吗？在对需求以及应用程序和业务约束进行了详细分析之后，您得到了一个非常详细的问题描述。这样，您就可以接近您的ML专家。

#### 一。目标设想的解决方案

你想要的结果应该会推动机器学习，而不是相反。

如果你问的问题不对，我就不能给你答案。我们需要能够明确地定义我们正在努力实现的目标。为机器学习定义和评估业务案例涉及许多考虑因素（我们将在下一篇文章中深入讨论），但目前，以下是定义您所需结果的主要标准：

#### 定义的问题

有针对性的结果要求应该足够详细，以便进行工作。你应该问自己这样的问题：

* ML模型应该输出什么？这是对你能观察到的事物的预测，还是对在特定情况下如何行动的建议？例如，假设你需要一个骑自行车的机器人部件，以防止它在湿滑的路面上摔倒。你想让模型预测在给定的环境条件和倾角下自行车是否会打滑吗？或者你想让你的模型预测最快的方法来绕过曲线而不打滑？
* 您需要模型以多快的速度计算其输出？什么样的硬件和软件环境可以作为一个平台来执行ML模型？
* 你的模型应该有多大的可转换性？回到自行车机器人的例子：模型应该只适用于一个特定的自行车还是来自不同制造商的稍有不同的自行车？

#### 可衡量的问题

假设有人给你一个预先训练过的模型。我们需要考虑你如何评估它的性能。对于给定的输入，您应该能够生成模型应该输出的所需目标值，并对大量输入执行此操作，从而预测生成目标输出的成本。

你还应该考虑如何向机器发出信号，表明结果是好的。这可以像使用true或false机制一样简单（例如，某个操作是否成功），也可以根据应用程序的不同而迅速变得复杂得多。

#### 有用的问题

从商业角度来看，这个问题值得解决吗？

* ML是否会提供客户愿意支付的福利？
* 或者ML可以降低制造成本？
* 也许，使用ML只会提供额外的营销优势？（想想苹果的Siri）

ML的应用将需要在ML工程专业知识方面的资源投资，以及基础设施的设置和维护。确保您正在考虑使用ML的好处是否会被数据收集和基础设施的成本所抵消。

#### 训练与回忆

当你训练一个模型时，你的目标是提高它的性能。在回忆过程中，您只需评估一个预先训练的模型来生成一个预测，但该模型并没有学到任何新的东西。

重要的是要记住，回忆的计算要求明显低于训练的要求。要训练模型，您需要访问所有的训练数据和许多小时的训练时间。根据您的ML任务，更好的游戏计算机可能提供足够的计算能力来训练您的模型；在某些情况下，可能需要计算集群。训练模型后，通常可以删除大部分训练数据并将模型保存为小文件。许多ML模型只需要很少的资源，比如嵌入式设备或手机，就可以调用预测。这已经建议将培训和召回分开。

另一个优点是软件基础设施所需的复杂性和敏捷性，这对于培训比召回要高得多。这种分离提供了将培训作为服务的业务模型（也称为订阅模型）。

#### 2。围绕数据设计

人工智能看不到角落，所以模型需要数据。很多数据。主要考虑因素包括数据设计、数据可防御性和收集、数据结构和清理、隐私或法律问题。

良好数据的标准如下：

可量化：数据需要可量化的结果，才能成为有用的培训 - ，可以是简单的是或否，也可以包括更详细的衡量标准。

相同分布：所有的数据点都应该从相同的概率分布中取样。通常，这意味着要确保您使用的数据集是一致的测量和记录。当你收集更多的数据时，你需要确保测量的一致性。

多样性：你应该有足够大的样本来覆盖你的整个潜在结果空间。这应该包括对罕见事件的足够覆盖，使您的模型能够区分随机错误和罕见（但很重要）事件。例如，如果你试图构建一个人工智能来预测下一次股市崩盘，那么仅使用2009年和2017年的数据将过于严格。

捕获外部驱动因素：外部驱动因素是影响您的结果的特征，需要捕获这些特征才能生成良好的预测。举个例子，如果你只知道方向盘的宽度和油箱的大小，那么很难预测汽车的价格，但如果你知道车型和里程数，就容易多了。

#### 三。成功的典范

在考虑人工智能时，组织往往从这里开始。，ML科学家通常在这方面积累专业知识。然而，ML的成功实现需要从探索之初就集成深入的战略和设计思想。

简而言之，只有在已知预期结果和数据约束之前，才能设计ML技术（模型），只有这样，才能将这些问题分解为机器可学习的部分。以下是优秀模特的标准：

可行性和可用性：在给定数据、技术约束和设计约束的情况下，使用此模型是否可以解决问题？如果不进行测试，很难解决先验问题。一篇即将发表的文章将讨论加速人工智能原型的策略。

可伸缩性：这是模型的计算效率。考虑到当我们增加数据量时计算和模型训练成本的增加，这些技术在规模上可用吗？在规模上，基础设施成本可能是不可忽略的。

人工智能实现的障碍主要是业务和产品战略可以消除的问题。这并不是说技术部分是容易的，但是技术约束通常是已知的数量；我们知道当前的人工智能能力是什么，以及它可以用来解决什么样的问题。

人工智能和人工智能的未来仍存在一些不确定性。一般来说，即使是一位经验丰富的数据科学家，也很难预测训练一个模型需要多少数据。此外，很难预测需要对模型的训练过程进行多少调整，以及训练过程对数据中的异常值（异常）可能有多敏感。

最后，将培训过程从早期原型转移到成熟产品所需的时间具有很高的不确定性。在这种情况下，必须记住，大多数机器学习算法都是随机的；当在同一个数据集上重复执行训练时，模型会有所不同，更重要的是，当在不同的数据集上训练时，它们的性能会有所不同。一般来说，随着时间的推移会获得新的数据，对于成熟的应用程序，需要自动化的模型培训。这通常需要自动化一些或所有的过程步骤，这些步骤是由人机学习专家在初始原型阶段执行的：数据清理、根据特定数据集调整培训过程以及自动选择模型。

So it is wise to treat each machine learning project, to some extent, as a research project with the associated higher risks and potentials.

Machine learning isn’t magic — it’s an incredibly powerful tool, but only if applied intelligently.

Written by [Nick Chow](https://www.axiomzen.co/about/nick) and [Alex Hentschel](https://www.axiomzen.co/about/alexh)Edited by [Yasmine Nadery](https://www.axiomzen.co/about/yasmine)