The difference between Statistical Modeling and Machine Learning, as I see it  
在我看来，统计建模和机器学习之间的区别

The difference between Statistical Modeling and Machine Learning, as I see it  
在我看来，统计建模和机器学习之间的区别

# The difference between Statistical Modeling and Machine Learning, as I see it 在我看来，统计建模和机器学习之间的区别

Published on June 30, 2016  
2016年6月30日出版

Oliver Schabenberger

Oliver Schabenberger  
奥利弗·沙本伯格

### [Oliver Schabenberger](https://www.linkedin.com/in/oschabenberger?trk=author_mini-profile_title)[Follow](https://www.linkedin.com/uas/login?session_redirect=https%3A%2F%2Fwww%2Elinkedin%2Ecom%2Fpulse%2Fdifference-between-statistical-modeling-machine-i-see-schabenberger&trk=author-info__follow-button)

#### Executive Vice President, Chief Operating Officer and Chief Technology Officer at SAS SAS执行副总裁、首席运营官和首席技术官

* Like1,285  
  比如1285
* Comment83  
  评论83
* Share  
  分享
* 166  
  一百六十六

## I frequently get asked about the differences between Statistics (statistical modeling in particular), Machine Learning and Artificial Intelligence. There is indeed overlap in goals, technologies and algorithms. Confusion arises not only from this overlap, but from the buzzword salad we are being fed in non-scientific articles. 我经常被问到统计学（特别是统计建模）、机器学习和人工智能之间的区别。在目标、技术和算法上确实存在重叠。困惑不仅来自于这种重叠，还来自于我们在非科学文章中被灌输的时髦词汇沙拉。

### Statistical Modeling 统计建模

The basic goal of Statistical Modeling is to answer the question, “\*Which probabilistic model could have generated the data I observed\*\*?\*” So you:  
统计建模的基本目标是回答这个问题，“\*哪个概率模型可以生成我观察到的数据\*\*？”？\*“所以你：

* Select a candidate model from a reasonable family of models  
  从合理的模型族中选择候选模型
* Estimate its unknown quantities (the parameters; aka fit the model to data)  
  估计其未知量（参数；aka将模型与数据拟合）
* Compare the fitted model to alternative models  
  将拟合模型与替代模型进行比较

For example, if your data represent counts, such as the number of customers churned or cells divided, then a model from the Poisson family, or the Negative Binomial family, or a zero-inflated model might be appropriate.  
例如，如果您的数据表示计数，例如客户流失的数量或单元划分的数量，那么来自Poisson族的模型、负二项族的模型或零膨胀模型可能是合适的。

Once a statistical model has been chosen, the estimated model serves as the device for inquiries: testing hypotheses, creating predicted values, measures of confidence. The estimated model becomes the lens through which we interpret the data. We never claim that the selected model generated the data but view it as a reasonable approximation of the stochastic process on which confirmatory inference is based.  
一旦选择了一个统计模型，估计的模型就可以作为调查的工具：测试假设，创建预测值，测量置信度。估计的模型成为我们解释数据的镜头。我们从来没有声称所选择的模型生成数据，而是将其视为基于确认推理的随机过程的合理近似。

Confirmatory inference is an important aspect of statistical modeling. For example, to decide which one of three possible medical devices has the greatest benefit to patients, you are interested in a model that captures the mechanism by which patient benefits are differentiated by treatment. It will often be the case that the model that captures the data-generating mechanism well is also a model that predicts well within the range of observed data—and possibly predicts new observations.  
验证性推理是统计建模的一个重要方面。例如，要确定三种可能的医疗设备中哪一种对患者最有好处，您需要一个模型来捕获按治疗区分患者好处的机制。通常情况下，能够很好地捕获数据生成机制的模型也是一个能够在观测数据范围内很好地预测并可能预测新观测的模型。

### Classical machine learning 经典机器学习

Classical machine learning is a data-driven effort, focused on algorithms for regression and classification, and motivated by pattern recognition. The underlying stochastic mechanism is often secondary and not of immediate interest. Of course, many machine learning techniques can be framed through stochastic models and processes, but the data are not thought in terms of having been generated by that model. Instead, the primary concern is to identify the algorithm or technique (or ensemble thereof) that performs the specific task: Are customers best segmented by k-means clustering, or DBSCAN, or a decision tree, or random forest, or SVM?  
经典的机器学习是一种数据驱动的学习，主要集中在回归和分类算法上，并受到模式识别的驱动。潜在的随机机制往往是次要的，而不是直接的利益。当然，许多机器学习技术可以通过随机模型和过程来构建，但数据并不是由该模型生成的。相反，主要关注的是确定执行特定任务的算法或技术（或其集成）：客户是通过k-means聚类、DBSCAN、决策树、随机森林还是支持向量机进行最佳分割的？

In a nutshell, for the Statistician the model comes first; for the Machine Learner the data are first. Because the emphasis in machine learning is on the data, not the model, validation techniques that separate data into training and test sets are very important. The quality of a solution lies not in a p-value, but in proving how well the solution performs on previously unseen data. Fitting a statistical model to a set of data and training a decision tree to a set of data involves estimation of unknown quantities. The best split points of the tree are determined from the data as are the estimates of the parameters of the conditional distribution of the dependent variable.  
简而言之，对于统计学家来说，模型是第一位的；对于机器学习者来说，数据是第一位的。因为机器学习的重点是数据，而不是模型，所以将数据分为训练集和测试集的验证技术非常重要。解决方案的质量不在于p值，而在于证明解决方案在以前未看到的数据上的性能如何。将统计模型拟合到一组数据，并将决策树训练到一组数据，这涉及到未知量的估计。树的最佳分割点和因变量条件分布参数的估计值由数据确定。

Neither technique can claim to be learning, in my opinion. Training is the process of shaping something. Learning, on the other hand, implies gaining a new skill and training is part of learning. By training a deep neural network—that is, determining its weights and biases given the input data—it has learned to classify, the network morphed into a classifier.  
在我看来，这两种技术都不能说是在学习。训练是塑造事物的过程。另一方面，学习意味着获得新的技能，而培训是学习的一部分。通过训练一个深度神经网络，即在给定输入数据的情况下确定其权值和偏差，将网络转化为分类器。

### Modern Machine Learning 现代机器学习

A machine learning system is truly a learning system if it is not programmed to perform a task, but is programmed to learn to perform the task. I refer to this as Modern Machine Learning. Like the classical variant, it is a data-driven exercise. Unlike the classical variant, modern machine learning does not rely on a rich set of algorithmic techniques. Almost all applications of this form of machine learning are based on deep neural networks.  
如果机器学习系统不是为执行任务而编程的，而是为学习执行任务而编程的，那么它就是一个真正的学习系统。我称之为现代机器学习。与经典的变体一样，它是一个数据驱动的练习。与经典的变体不同，现代机器学习并不依赖于丰富的算法技术。这种形式的机器学习几乎所有的应用都是基于深层神经网络的。

This is the area we now tend to call Deep Learning, a specialization of Machine Learning, and frequently applied in weak Artificial Intelligence applications, where machines perform a human task.  
这就是我们现在倾向于称之为深度学习（Deep Learning）的领域，它是机器学习的一个专门领域，并且经常应用于弱人工智能应用中，机器执行人工任务。

### Role of the data 数据的作用

We can now distinguish statistical modeling, classical machine learning and modern machine learning by the role of the data.  
我们现在可以通过数据的作用来区分统计建模、经典机器学习和现代机器学习。

In statistical modeling, the data guide us to the selection of a stochastic model which serves as the abstraction for making probabilistic statements about questions of interest, such as hypotheses, predictions and forecasts.  
在统计建模中，数据指导我们选择一个随机模型，作为对感兴趣的问题（如假设、预测和预测）进行概率陈述的抽象。

In classical machine learning, the data drive the selection of the analytic technique to best perform the task at hand. The data trains the algorithms.  
在经典的机器学习中，数据驱动分析技术的选择，以最好地完成手头的任务。数据训练算法。

In modern machine learning, the data drive systems based on neural nets that self-determine the regularities in the data in order to learn a task. The process of training the neural network on the data learns the task. As someone put it, “The data does the programming.”