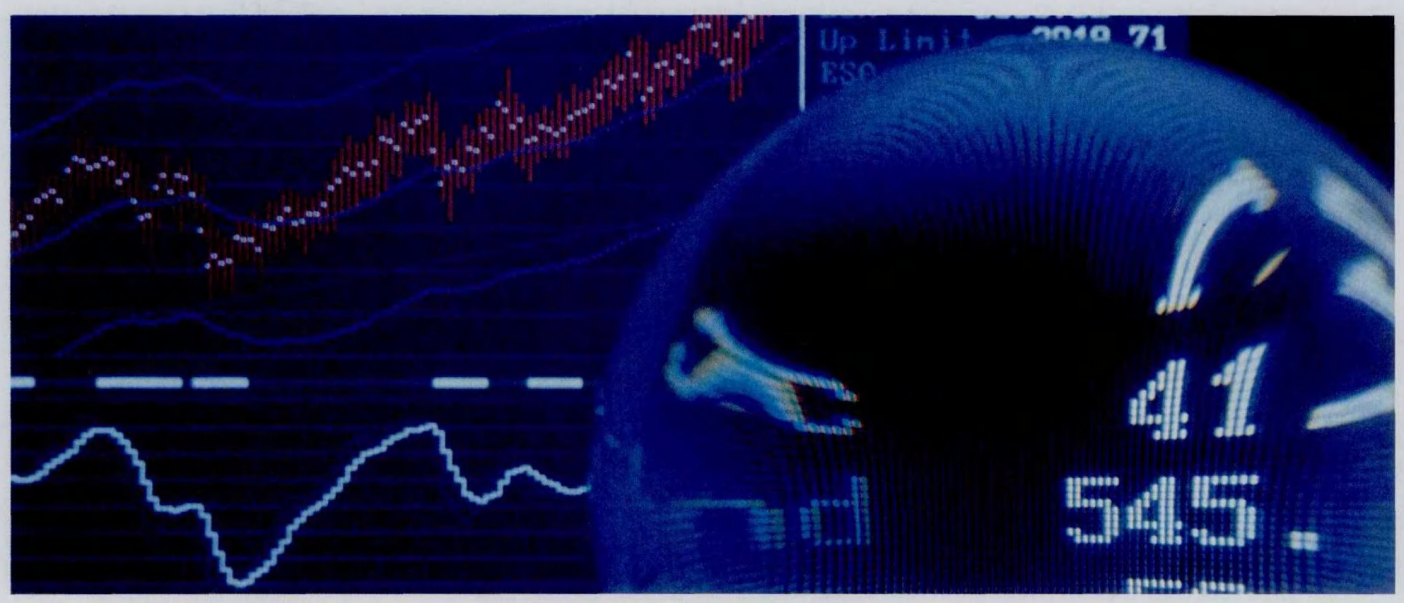
# MICROSOFT
# 大数据与人工智能

**大数据加上机器学习，代表了软件产业一个新时代来临：不再是人写软件，而是数据加算法，在数字化之后，以计算机驱动，用光速来推进人类社会。**

微软亚洲研究院常务副院长 马维英




大数据不是一个单一的现象，在过去5 ~ 10年，借着越来越强大的计算能力，加上知识挖掘等算法上的突飞猛进，我们可以构建更大的模型。

我们都知道，其实很多的人工智能、机器智能，或者驱动一个数字世界的自动化，都需要模型。在过去的很长一段时间内，这些模型因为数据的不足，通常只能做一些小的模型或者浅的模型，但特别在过去的两三年里，在机器学习领域有了突飞猛进的发展，可以构建更大、更深度的模型。

另外是知识挖掘，知识挖掘是怎样在互联网上大规模进行的？在这些结构化、半结构化的数据中构建人类最完整的知识表达，一旦我们掌握了这样的表达，我们就可以对很多的数据做更深度的理解。

## 软件产业新时代

微软为什么看重大数据？微软是从一家软件企业开始的，在30年前，微软就知道软件会改变全世界，软件几乎可以做所有的事情。但今天，软件的时代已经不一样

# 069

了。过去是程序员写，接下来的软件是靠数据，靠机器学习，自动写出软件。这意味着什么？当你拥有更大的数据的时候，软件就更强大，软件的性能就更好。大数据加上机器学习，这代表了软件产业一个新时代来临：不再是人写软件，是数据加算法，在数字化之后，以计算机驱动，用光速来推进我们人类社会的方方面面。因为都是靠数据和算法，我们能够产生一个非常大的一个信息量，所以我们看重下一代所有的应用和服务。

今天，为什么各个大公司都在不断地在人才、技术方面大量投入？就是因为我们已经在过去的几年不断地对这些数据进行加工，我们已经越来越接近从数据到信息，到知识到科技。

语音识别过去十几年一直没有突破，但就在这两年实现了很大进展。其中重要原因就是因为有一种新的机器学习，可以从数据里面学表达方式，做很多模式识别。大家知道做影像、语音等最难的是怎么找出特征向量表示数据，过去的二三十年的研究都在研究怎么找特征，大数据来到的时候我们发现，特征也可以直接从数据去学。而且在这个过程中我们发现一个非常有趣的现象，就是越大的数据表达方式越好。因为数据大，信息就增加，所以技术上的突破就是计算能力和数据大了，而且自动学出来的发现比过去人设计的特征向量更好。

## 大数据的城市应用

今天的技术非常令人兴奋，但今天很多城市里面的数据都是很低阶的，这么大的数据如何表达？今天深度学习、机器学习带来了革命性的机会。过去两年我们也把这样的机器学习开始应用在城市，所以微软有一个城市计算的项目。在北京，我们收集了很多方面跟城市有关的数据，例如北京交通路网的数据、北京商业各方面的历年数据，我们可以发现很多现象，比如北京过去 10 年酒吧在哪里越来越多，电影院也在一些地方增加了，这些其实代表了这个城市的发展。还有空气的数据、气象的数据等。我们还搜集了北京的 30000 辆出租车，特别装有定位系统的数据，这样便可以用出租车当做传感器实时检测城市的脉动，甚至交通的状况，可以算出更好的开车路径。

过去两年，微软与北京市政府及中国的高校合作，在城市计算这一领域实现了很多科研成果。大数据可以分析城市问题，改善城市规划。

空气质量是今天的城市居民最关注的一个话题。北京这么大规模的城市，现在只有 15 个空气检测站点，非常稀疏，每个站点的投入和花费与运营非常高。这些站点在任何时刻给的数据都是非常不一样的，说明一个城市里面空气质量分布是不均匀的。今天我们的问题是能不能用大数据，用机器学习的方法预测那些没有空气站点的地点的空气质量。我们也利用了大数据把历史上所有这些我们可以收集到的，关于侦测带里的历史的数据、气象数据、交通数据、人员流动等数据，建一个非常大的模型，包含空间和时间的预测，能够在有限的 15 个站点之上，进一步预测所有的位置。这就是大数据在智能城市里的应用，其中既有数据分析，而且是海量和一致性的数据。

实验证明，虽然我们用的很多数据从某个角度来讲是比较弱的信号，但是把大量相关的相对弱的信号收集起来，居然比原来最好的模型还好 20%。利用大数据对整个城市的空气做预测的模型是非常有价值的，这预示了这样的应用将不只在城市，而是会进到人类社会方方面面，各个产业都会被大数据带动，用更智能的算法，让过程变得更有效。