数据科学之路第一章：数据科学简介

Hi，大家好，我是李俊，在本次课程中，我将和大家分享FAFU数据科学导论（2018）课程的第一章第二节：数据科学简介的内容。

数据科学的基本内容

作为一门新兴的学科，数据科学的崛起依赖两个因素：一是数据的广泛性和多样性；二是数据研究的共性。现代社会的各行各业都充满了数据，这些数据的类型多种多样，不仅包括传统的结构化数据，也包括网页、文本、图像、视频、语音等非结构化数据。数据分析本质上都是在解反问题，而且通常是随机模型的反问题，因此对它们的研究有很多共性。

进入2012年，大数据（big data）一词越来越多地被提及，人们用它来描述和定义信息爆炸时代产生的海量数据，并命名与之相关的技术发展与创新。作为一门算是比较新兴的领域，它的兴起离不开算法、算力、数据这三大推动力。

- **数据**：伴随着移动互联网的发展，海量的传感器出现在社会的方方面面，潜移默化着影响着人们的生活，其产生的巨量结构化数据（数字）和非结构化数据（图像、视频、文字、语言等）已非人力所能计算。
- **算法**：伴随着深度学习的崛起，CNN、RNN和LSTM等算法领域的创新，科学家们貌似找到了处理非结构化数据的能力。另一方面，在大量数据下，深度学习算法能实现拟合模型效果远高于传统机器学习算法。
- **算力**：随着摩尔定律（CPU性能价格比大约十八个月翻一番）的成功预测，从1971到2015年的44年间，集成电路晶体管数量增长了四百三十五万倍，运算能力得到了大幅度的提升，CPU、GPU、TPU、FPGA等芯片的更新迭代，云计算的快速发展，都为我们处理大数据提供了可能。

数据科学、机器学习、人工智能的区别

这些领域有很多重合的地方，而且各自都有各自的说法，选着哪一个看起来更像是一个市场问题。对于这个问题，我比较认同David Robinson在《数据科学、机器学习、人工智能都有哪些区别》一文中提出的观点：**数据科学产生见解、机器学习做出预测、人工智能产生行为。**数据科学的目标基于人类：能够获得洞察力和理解，其需要有人工介入：有人在理解、洞察，看到数字，或者从结论中受益；机器学习基于给定具有特定特征的实例X来预测Y，所以它是关于预测的领域，对于一个给定的问题，我们可以通过机器学习的方法，利用已有数据匹配出一个模型来进行预测分析，然后用数据科学来构筑结论和可视化结果来验证为什么这个模型有效。人工智能是目前为止这三个类中最古老和最广为承认的，早在二十世纪五十年代，英国数学与逻辑学家阿兰·图灵就已经开始思考这个问题了，并与1950年发表了《计算机与智能》这一传世之作。

为什么要学习数据科学

通过互联网二十多年的发展，我们积累了大量的数据，中国科学院院士、北京理工大学副校长梅宏认为我们正在进入一个以数据的深度挖掘和融合应用为特征的智能化阶段（大数据正在开启第三次信息化浪潮），一次人类历史上的重大事社会的变革正在拉开序幕。生活在当前这个大数据时代中，我们需要有计算思维（现实中的问题能否使用计算的方法解决）和大数据思维（如何获取数据、如何分析数据、如何从数据萃取价值、如何应用数据）。在个这个给我们带来机遇和挑战的信息化、大数据时代中，我们每个人都有可能成为时代的弄潮儿。

强烈推荐大家去看这个演讲，地址：开讲啦：梅宏开讲大数据时代，你准备好了吗?

最后送个大家一点寄语，撒贝宁说的，**大数据也好、网络也好，未来所有的一起，我们今天听到的最新的概念（人工智能、区块链等等），我们不要坐在这等着它们来改变我们的生活。如果你没有办法参与到这其中，那这种等待是痛苦的，你是被动的跟着这个时代在跑；未来是怎么来的，未来不是等来的，未来一定是一步一个脚印用自己的脚丈量出来的。**