1-3-cn

1

大家好，我是北京理工大学计算机学院数据科学与知识工程研究所的车海莺，本节我们科学研究的第四范式

2

Jim gray，他是关系数据库的创始人，也是航海运动爱好者。

2007 年 1 月 28 日，他永远消失在海上，就在 11 天前，

2007 年 1 月 17 日在加利福尼亚州山景城举行的 NRC-CSTB，美国国家研究委员会-计算机科学和电信委员会会议上，他发表了他的最后一场演讲，“科学方法的革命”，将科学研究分为四种范式。

（范式，必须遵循的某种规范或每个人都使用的例程），

其次是实验归纳、模型推演、模拟和数据密集型科学发现（Data-Intensive Scientific Discovery）。

其中，最后一个“数据密集型”就是我们现在所说的“科学大数据”。

3

1） 上千年前，科学是经验主义的，人们通过描述自然现象来做科学。和中国古代一样，耕作是通过观察天象来完成的。2）过去几百年，理论分支，使用模型，概括。从原始的钻木取火，发展到以伽利略为代表的文艺复兴时期的科学发展初期。很难对自然现象进行更准确的认识。科学家们开始尝试尽可能地简化实验模型，去除一些复杂的干扰，只留下关键因素（例如，“足够流畅”、“足够长的时间”、“足够稀薄的空气”等。物理学习，就是莫名其妙的条件描述），然后通过微积分来总结，这是科学研究的第二范式。这种研究范式一直持续到 19 世纪末，堪称完美。牛顿三定律成功解释了经典力学，麦克斯韦理论成功解释了电磁学，经典物理大厦宏伟壮观。但在量子力学和相对论出现之后，以理论研究为主，以非凡的大脑思维和复杂的计算超越了实验设计。随着理论验证的难度和越来越高的经济投入，科学研究开始显得力不从心。

4

3）。 20 世纪中叶，冯·诺依曼提出了现代电子计算机体系结构，用电子计算机模拟科学实验的模式迅速普及。人们通过模拟可以推断出越来越多的复杂现象。典型案例包括模拟核试验和天气预报。随着计算机模拟日益取代实验，逐渐成为科学研究的常规方法，是科学研究的第三范式。4）科学未来的发展趋势是随着数据的爆炸式增长，计算机不仅可以进行模拟，还可以进行分析总结，产生理论。数据密集型范式应该从第三范式中分离出来，成为一种独特的科学研究范式。这种科学研究方法被称为第四范式。科学研究的第四范式和科学研究的第三范式。两者都使用计算机进行计算。这两种科学研究范式有什么区别？“什么是科学问题？”，“什么是科学假设？”这是首先提出可能的理论，然后收集数据，然后通过计算验证它们。基于大数据的第四范式是先拥有大量已知数据，然后再计算以前未知的理论。（范式是“人脑+电脑”，人脑是主角，主角，第四范式是“电脑+人脑”，电脑是主角。但是，要发现事物之间的因果关系，在大多数情况下这总是很难的。我们人类推导的因果关系总是基于过去的知识，得到一个“确定性”的机制分解，然后建立一个新的推导模型。但是，这种过去的经验和常识可能是不完整，甚至可能有意或无意地忽略重要变量。）

5

根据现有机理认识，霾天气的形成不仅与源大气化学成分有关，还与地形、风向、温度、湿度和气象因素。

只有这些有限的参数已经超出了常规监测的能力，只能通过简化和人为的方式去除一些看似不重要的因素，

只保留一些简单的参数。

那些看似不重要的参数会在特定条件下发挥至关重要的作用吗？如果考虑不同参数的空间异质性，这些气象站的空间分布是否合理充分？

从这个角度来看，如果我们能获得更全面的数据，或许就能真正做出更科学的预测。

这是第四范式的起点，也许是解决问题最快、最实用的方法。

那么，如何研究第四范式呢？在移动终端迅猛发展、传感器快速发展的时代，未来的趋势似乎在望。

现在，我们的手机可以监测温度和湿度，并且可以定位空间坐标。
很快就会有可以监测大气环境化学和 PM2.5 的传感设备。这些移动监测终端增加了测量的空间覆盖范围，同时产生了大量的数据。
利用这些数据，我们可以分析雾霾的成因，最终做出更好的预测。这种海量数据的出现，不仅超出了普通人的理解和认知能力，也给计算机科学本身带来了巨大的挑战。
因此，当这些大规模计算的数据量超过 1PB 时，传统存储系统已经无法满足海量数据处理的读写需求，数据传输 I/O 带宽的瓶颈越来越突出。
然而，简单地将数据分块并不能满足数据密集型计算的需求，也有悖于大数据分析的初衷。因此，目前许多具体研究面临的最大问题不是缺乏数据，而是数据过多而不知如何处理。
6
另一个放弃因果关系，采用相关性的例子。
2004 年，在沃尔玛购物清单中，沃尔玛记录了所有有用的信息，包括消费金额、购物篮中的物品、具体的购买时间，甚至是购买的天气。
他们发现，每当季节性飓风来临时，不仅手电筒的销量会增加，蛋挞的销量也会增加。
因此，当季节性风暴来临时，蛋挞在飓风用品边上摆放，结果增加了销量。很难解释原因。
换句话说，只要你知道“是什么”，就不需要知道“为什么”。
这颠覆了人类几千年来的思维常规，据说对人类的认知和与世界的交流方式提出了全新的挑战。
因为人类总是思考事物之间的因果关系，对基于数据的相关性不是那么敏感；相反，计算机本身很难理解因果关系，
并且非常擅长相关性分析。所以，我们可以理解这一点。在维克多 · 迈耶-勋伯格所著的《大数据时代》中，
明确指出，大数据时代最大的变化是摒弃了对因果关系的渴望，即因果关系，转而关注相关性。
7
这节课我们学习了科学研究的第四范式，它以大数据为基础，先有大量已知数据，再计算以前未知的理论。 而大数据时代最大的变化就是摒弃了对因果关系的渴望，即因果关系，转而关注相关性。