数据科学家

郑 宇 [1, 2]

[1] 京东智能城市研究院，北京市亦庄区经济技术开发区京东大厦 A 座
[2] 京东城市（北京）数字科技有限公司，北京市亦庄区经济技术开发区京东大厦 A 座

摘要：信息时代的到来催生了海量的数据，在各行各业都发挥着重要作用。数据也被我国列为继土地、劳动力、资金和技术之后的第五种生产要素，不仅自身会形成一种产业，也将改变国家命运和世界格局。为了能更好的发挥数据的价值，时代需要一批围绕数据开展工作的数据科学家，用数据科学去研究、探索并解决各种行业问题，并在此过程中不断革新数据采集、清洗、管理、分析、挖掘、展现的理论和方法。然而，这个职业的诸多未知因素和培养难度跟行业的广大需求形成了强烈的反差。因此，本文探讨了培养数据科学家的必要性，定义了数据科学家的内涵和外延，对比了数据科学家跟数据工程师等相关岗位的区别，介绍了数据科学家应该具备的技能和素质，给出了培养数据科学家的路径和方法，并通过具体案例来演绎数据科学家的实战过程。本文有助于增强行业对数据科学家的认知，加快培养数据科学家的步伐，有利于更好的发挥数据要素的效能，为社会和国家创造更多价值。

关键词：数据科学、数据科学家、数据工程师、生产要素

Data Scientists

The proliferation of information technology leads to a digital era, where data has been changing people's lives and the entire industries, calling for data scientists who can apply data science to solving real-world problems and in turn innovate the methodology of data science. While the demands for data scientists are huge, training such a data scientist is very challenging, as it requires the capability of quickly learning domain knowledge, mastering a diversity of data science models, deeply understanding of data and proficiently using big data platforms. To address this issue, this article formally defines the concept of data scientist, describing the soft and hard skills they need to be equipped with. We also compare data scientists with data analysts and AI engineers. In addition, we introduce the way of training a data scientist, and showcase an example where a data scientist solves a real problem. This article can help train quality data scientists who will then provide more values from data to the world.

Keywords: Data scientists, data analysts, data science.

1. 引言

信息时代的到来催生了海量的数据，每个人、每个机构、每个设备既能成为数据产生的源头，也是数据的使用者。数据已经在各行业中得到应用，在降低成本、提升效率和改善用户体验等方面起到了关键性作用。同时，数据在继土地、劳动力、资金和技术之后，被定义成新的生产要素，本身也将成为一个新兴行业，孵化以数据为核心资产的产业，为全球经济发展贡献新的增长动能。

在数据如此重要的时代，如何使用好数据、发挥数据的价值就变得至关重要，将影响到行业发展、国家命运以及世界格局。这样一个新的时代也将培育出一批围绕数据开展工作的机构和从业者，造就一系列以数据为中心的职业来承担时代赋予的使命。数据工程师、数据分析师、数据科学家等职业和岗位应运而生。其中数据科学家（Data Scientist）尤其受到关注，被欧美国家称为 21 世纪最“性感”的工作[1]。当数据这个没有边际、幻影无形、深藏真理、积聚宝藏的土壤

遇上科学家这个神秘尖端、勇于探索、不断实践的职业后，必然会产生惊人的聚变、留给人们无穷的想象空间。

但到底什么是数据科学家？他们应该具备哪些素质和技能、如何开展工作、又如何培养这样的人才，至今还缺乏准确的回答和清晰的思路。这个职业的诸多未知因素和培养难度跟行业的强烈需求形成了巨大反差。因此，作者结合自身15年从业经历，对以上问题做出探讨，希望能帮助行业培养出一批优秀的数据科学家，为社会创造价值、为国家贡献力量。

2. 为什么需要培养数据科学家

需要培养数据科学家的理由包括应用场景需求大、数据要素价值高、人才培养难度大三个原因：

应用场景需求大：数据极大丰富，驱动大量应用，渗透各种场景，催生庞大产业，只要有数据的地方，就需要有人来管理和利用好这些数据，需要大量的数据科学家。

数据要素价值高：数据作为继土地、劳动力、资金和技术之后的第五种生产要素，其创造的价值将超过前四者的总和，原因如下：

- 数据产生门槛低，人人都能产生数据，很多系统和设备还能自动地产生数据；
- 数据不断产生、总量没有天花板；
- 数据被使用后不会被消耗、可重复使用；
- 前四个生产要素都会被数字化；

人才培养难度大：当前学校培养的学生仅仅学习了一些算法和理论，缺乏对业务的理解和实战经验，很难满足市场的需求。传统行业的从业者要学习新的大数据和人工智能技术更加困难。除了掌握行业知识和专业技能外，数据科学家还需具备优秀的基础素质和探索的精神，要求很高。

3. 什么是数据科学家

数据科学家的定义可以从两个维度去理解，一个是“数据”+“科学家”，另一个是“数据科学”+“家”。由于数据和科学家都有清晰的定义，因此，第一个维度可以简单理解为研究数据本身的科学家，即不断革新数据采集、清洗、管理、分析、挖掘、展现的理论和方法的人，这可以被认为是数据科学家的内涵。第二维度是指用数据科学去研究、探索并解决各种实际问题的人，这可以被认为是数据科学家的外延。数据科学家外延的不断扩大驱动其内涵不断深化，两个维度加在一起才构成了对数据科学家的完整诠释。

数据科学家：以数据科学为方法论，利用数据、结合行业知识来认识和探索世界，解决各类实际问题、创造社会价值，并在此过程中，不断研究、创新数据的采集、管理、分析、挖掘、展现的理论和方法，深化数据科学内涵的人。

数据科学：数据科学被称为科学的第四范式，这四大范式分别是：经验法、理论法、计算法和数据驱动法。数据科学利用数据驱动的方法来分析和解决问题，从数据中探寻事物的本质和规律，研究数据获取、管理、分析、挖掘和展示等一系列环节中的理论和方法，并探索其应用。

数据科学家应具备的技能：“一位优秀的数据科学家应当是站在（大数据）平台上，看问题、想数据、关联模型，并把这些模型有机的组合起来部署到大数据平台上，处理鲜活数据、产生知识、解决行业问题”。这句话中蕴含了数据科学家应当掌握的四大技能：

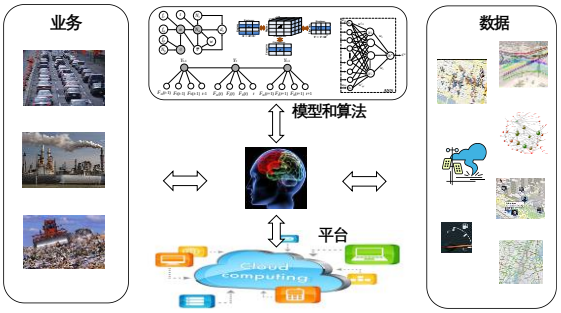

图 1 数据科学家

- **数据科学家要了解行业问题**，知道影响该问题的因素。比如，空气污染即有本地污染源排放，也有外地污染物扩散而来；污染源包括工厂、车辆尾气、餐饮机构等。只有知道导致污染的因素，我们才能去收集相关的数据，并在模型中选择相应的特征。我们还要了解行业里现有的方法，一方面，借鉴其思想和精髓，帮助模型设计；另一方面，也要知道现有方法的不足之处，让数据科学来弥补其缺陷。最后，还要学习行业的语言，以便能够跟行业专家沟通，让其理解和接纳基于数据科学的解决方案。
- **数据科学家要深度理解数据**，不仅仅只是了

解数据的格式、属性和表面意思，更要理解数据背后能够反应的深意。例如，出租车的轨迹数据不仅反应了出租车行驶的速度和去过的地方，也反应了它所在路段的通畅程度。但这还不够，由于轨迹数据还体现了乘客的上下车地点，当把大量的出租车轨迹数据融合在一起后，我们可以洞察一个区域内人们的出行规律。进一步，人们的出行规律又可以反应出这个区域的功能，如住宅区的人们早出晚归，而商务区则相反。这个区域功能又能影射出对空气污染的影响。如，公园的扩散条件好，污染源少，则空气相对会更好；商务区在早晚高峰时段交通拥堵、楼房密集，空气相对较差。有了对数据背后深意的理解，我们就可以用出租车的轨迹数据来推测一个地域的功能[2]和预测其空气质量[3]这样我们就能用领域 A 的数据去解决领域 B 的问题。届时我们就会发现数据其实极大丰富，限制我们的是自己的想象力。

- **数据科学家要精通掌握各种数据管理、数据挖掘、机器学习和数据可视化模型，具备数据侧端到端的能力**。这些能力相互关联，决定了数据应用的效果。如只掌握其中某个环节，缺乏对上下游可行性的考虑，设计的方案很难实际落地。
- **数据科学家要熟练运用大数据平台**。真正意义上的大数据不仅体量大、而且高速更新，这就必须有一个大数据平台来承载这些数据以及处理数据的能力。单机模式下的模型设计和工作方式无法应对真实世界的大数据。很多数据处理算法在小数据、单机模式下可以工作，但到了大规模、高动态的数据场景下就完全失效。比如很多数据驱动的空间索引算法（像 R-tree），因为其结构会随着数据的插入而发生巨大变化，不仅更新代价太大，而且会改变空间划分的结构（导致磁盘的映射结构也要不断改变），就不适合作为云计算环境下的空间数据索引结构[4]。

4. 数据科学家与相关岗位对比

为了能让大家更加好的理解数据科学家这个岗位，我们把它跟数据分析师、解决方案架构师、数据工程师和 AI 算法工程师进行对比。

数据科学家 VS 数据分析师：业界经常看到招聘数据科学家的信息，但根据其要求来看，更像是数据分析师（Data Analyst）。虽然数据分析师也很重要，但行业对数据科学家的要求远远高于数据分析师。

数据分析师面对确定性问题，即问题的定义、可以使用的数据源、需要输出的结果都是确定的，然后根据这些确定信息来选择相应的模型，计算结果即可。

例如，在用户申请信用卡时，银行要求申请人填写年龄、职业、房产、收入等个人信息，然后根据这些信息决定是否给申请人发信用卡，如果发，该发多少额度的信用卡才合适。这是一个非常明确的分类问题，模型的输入数据是申报人填报的个人信息，输出的结果就是“不发”、“5000 以下”、“5001-20000”等额度区间。这个模型可以利用已经发出的信用卡持有人填报的个人信息（作为输入特征）以及他们后来的还款记录（如能及时还款则对应额度作为标注）来训练。具体的模型选择和训练方法很多，不是本文重点。当利用历史数据把模型训练好之后，输入一个新申请人的信息，就能自动分类出相应的额度等级结果。

但数据科学家面对的是完全开放的问题，问题没有明确的定义、用什么数据不清楚、输入和输出是什么也不清楚、用什么模型更不清楚，这一切都要靠数据科学家来分析和定义。以下是从业过程中遇到的实际问题样例，这些才是数据科学家需要解决的问题。

例 1：有一条道路上面灰层很大，如何用大数据的办法把灰层彻底根除掉？

例 2：城市里的危化品一旦爆炸后果不堪设想，如何降低危化品带来的隐患，保证城市的安全？

例 3：空气污染严重，如何用最小的经济损失换取更多的蓝天？

例 4：如何抓到违规倾倒垃圾的渣土车？

以上问题没有清晰的定义，没有人告诉你应该用什么数据，期待的输出结果是什么都不知道，更无法归结到数据科学中的聚类、分类、回归等模型问题。另一方面，这些问题也不一定是一个单一模型就能解决的，往往需要把问题拆解成很多环节，然后用一套组合拳来解决。因此，数据科学家不仅要解决完全开放的问题，还需要提供一

套完整的端到端的数据解决方案，而数据分析师只需要解决确定性问题中的一个环节。

数据科学家 VS 解决方案架构师：虽然数据科学家需要提供端到端的数据解决方案，但跟解决方案架构师还是不同。

解决方案架构师针对业务问题，根据客户在特定场景的需求，将产品和能力进行组合、连接并作定制化的封装，解决客户痛点、为客户创造价值。解决方案架构师也不同于技术架构师，后者更加专注于技术的耦合（如 Spark 加 Redis 技术来完成信息处理），而非业务和功能层面的连接（终端录入+信息入库+大屏展示）。

在以数据为中心的应用中，数据科学家可以充当解决方案架构师的角色，但反之不然。解决方案架构师并不一定有数据科学的基础，在很多传统的信息化项目中，更多只是考虑信息的流转，不涉及到数据的分析和挖掘。要解决实际问题，数据科学家要具备解决方案架构师的思维和能力。

数据科学家 VS 数据工程师、AI 算法工程师：在实际项目中，数据科学家需要带领数据工程师和 AI 算法工程师一起实施方案。数据工程师依照数据科学家设计好的方案，实施数据的采集、接入、治理、管理和展现等工作。AI 算法工程师则根据数据科学家给出的思路完成模型的细化设计（包括模型的内部结构、输入输出的量化、详细参数的选定以及跟其它模型的嵌套组合方式）、模型的训练（训练方法、样本集合等）、测试和发布。虽然 AI 算法工程师并不直接面向客户，但这里有很多具体且重要的工作需要完成。当与设计期望发生偏差时，算法工程师应告知数据科学家，与后者一起迭代模型思路。数据科学家应不断统筹、协调数据工程师和 AI 算法工程师的工作进展，确保方案落地执行。

与数据工程师和 AI 算法工程师相比，数据科学家的工作更加宏观、全面，偏向整体方案的创造和设计，而前两个职位更注重数据科学中某个环节深入具体的工作，偏向于执行和实施。当然，在这些具体环节中仍然有很多需要进一步思考和设计的空间，并不是简单机械的执行。为了确保设计方案的可行性，数据科学家在正式上岗前，必须要有数据工程师和 AI 算法工程师的经验。

5. 如何培养数据科学家

如图 2 所示，培养优秀的数据科学家首先要让其树立正确的数据观，并不断提升其四大基础素质：认知能力、学习能力、创新能力和沟通能力；同时快速学习行业知识，并掌握数据、模型、平台三大专业技能。

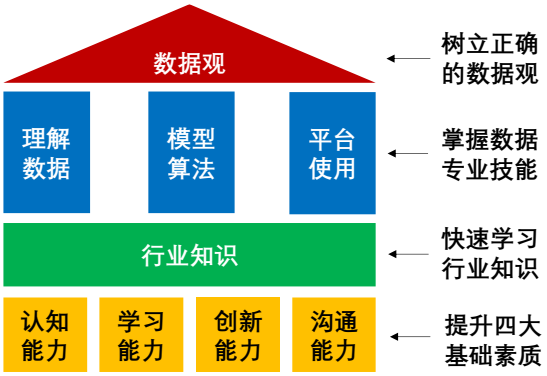

图 2 数据科学家的培养路径

树立正确的数据观：很多学生在面临实际问题时，容易陷入以下误区，这些都是没有树立正确数据观的表现：

- 拿着锤子找钉子，有了一个模型或者方法论，一定要想法设法把它用上去；
- 容易选择过于复杂的模型，以体现自己的技术能力，生怕用的模型简单，被认为工作没有难度和价值；
- 抱怨数据质量太差、想要的数据缺失，或者数据规模太小，因此认为这件事情没法作；
- 认为只有 AI 模型部分最有技术含量，其余部分都不重要；
- 初学阶段，不打好相关基础，直奔 AI 模型，从空中楼阁开始学习。

与之相对的正确的数据观如下：

- 数据科学解法的选择更多是依靠业务驱动（根据问题的特性、数据的实际情况等），解法的价值由业务成果来体现，用不用某种模型不是关键。
- 一个工作的难度由待解决问题的复杂度决定，而不由解法的复杂程度决定。如果能用简单的方法解决复杂问题是非常有价值的工作。因此，面对实际问题，一定从简单方法开始尝试，任何让解法变得复杂的付出，都需在结果侧有性能的提升，否则就是哗众

取宠、浪费资源。

- 在真实世界，最初的数据永远都不会让人满意，永远都会面临数据不足、质量不好等一系列问题。如果数据好到可以直接从中看到结果，数据科学家也就没有存在的必要了。加强对数据的深度理解，学会将领域 A 的数据应用到领域 B 的问题，才能破解数据不足的难题。此外，合理的选择模型，通过“不确定”+“不确定”得到“确定”的思维方式来应对不理想的数据也是解法之一[5]。
- 数据科学链路上的任何一个环节都是同等重要的，AI 算法并不高人一等，任何一个环节的失误都会让我们得不到想要的结果，失去利用数据创造价值的机会。
- 在不同的阶段应该练习好不同的技能。首先应该练好程序设计的基本功、积累软件开发的工程规范经验；然后学习数据管理模型、培养处理数据的动手能力；再尝试数据可视化的常用方法，积累数据展示的经验；之后学习 AI 模型，加强模型训练和部署的实践；最后，面对客户实战，快速学习行业知识，增强业务与数据科学的结合能力，并培养解决方案思维，完善数据侧端到端的能力。数据科学家无法一步到位，必须一步一个脚印的走出来。

提升基础素质：认知能力、学习能力、创新能力和沟通能力是数据科学家应该具备的四大重要基础素质。如图 3 所示，这四大素质相互连接，不断提升、强化人的知识体系。本质上我们跟客户或者行业专家的交流，就是双方以知识体系为核心，以四大素质为能力支撑的“交锋”，这四大基础素质也在交锋过程中不断历炼、提升。

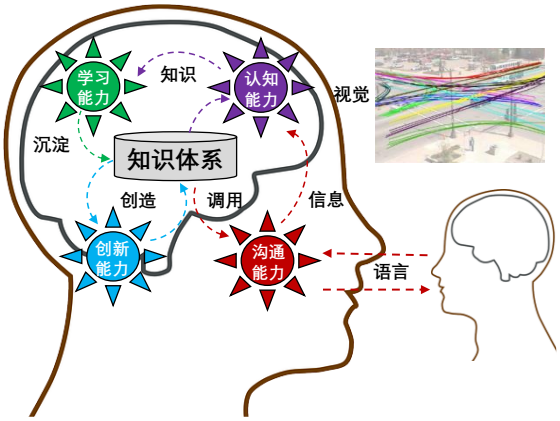

图 3 数据科学家的四大基础素质

首先沟通能力包括准确听懂和正确表达，它把从接收到的语言中提炼的信息传递给认知能力；此外，沟通能力也调用知识体系中的知识，并通过语言正确的表达出来。这个正确既包含意思的准确，也包含表达方式的合理。

认知能力接受来自视觉和沟通能力传递的信息，基于现有的知识体系来判别其深层次含义、认知其中新的知识。这些知识通过学习能力再沉淀回知识体系中，不断充实、壮大知识体系；如果没有学习能力，即便理解了，这些知识也会流失。创造能力基于已有的知识体系不断创新、加工，让知识体系不断自我完善、丰富。

快速学习行业知识：数据科学家必须掌握快速学习行业知识的方法论，并将行业知识跟数据、算法、平台融合。很多精通数据模型和算法的专家无法成为数据科学家，最大的瓶颈也在于此。通过以下四种方法可以快速学习行业知识：

- **从相关行业的文献学习**：通过相关行业高质量的综述、论文以及网络文章，我们可以快速学习整理和提炼好的行业知识。
- **向客户学习**：如在智能城市业务中，政府的主管领导往往对其业务非常了解，可以充当半个产品经理的角色。通过跟他们的沟通和交流，既能了解客户需求，也可以快速学习业务知识。但在这样的交流中，要避免像小白一样只会发问，要能用自己的思想和见解去引导客户，并在交流过程中将从客户那学到的知识快速融合到自己的知识体系中，然后结合自身的知识储备加以深化和拔高，再反馈给客户，让客户有所收获。这种不断思考、互动、深化的学习过程才能快速掌握行业知识。这也非常考验数据科学家四大基础素质的快速联动能力。
- **从国家政策和政府工作报告中学习**：此类报告的内容经过专家学者和政府领导多轮论证和推敲，蕴含专家智慧、条理清晰、高度概括，且反应了一个行业未来的发展动向，是很好的学习材料。
- **从其它案例中学习**：通过新闻报道、参观访问，我们可以学习其它案例中的精华、亮点，吸取经验教训，并感知行业的发展趋势。

掌握数据专业技能：这里的专业技能包含对数据的深刻理解，以及设计模型和使用平台的能力。

要学好以上专业技能，必须练好基本功、深入一线、做到应用闭环，并为业务直接创造价值。

- **练好基本功**：在学校，除了程序设计、软件工程等基础课程（计算机相关专业必修课），学生可以按照如下顺序学习简单的数据管理、数据挖掘、机器学习和数据可视化课程，并参照大数据平台教程作一些实验。以上任何一门课程，如果想深入了解，都需要花费数年的时间。因此，建议在完成初步学习后，依托一个具体项目，边做边学、逐步深入，这样印象会更加深刻，动力也会更足。《Urban Computing》一书就是按照以上思路编写[6]，以满足大部分学生快速入门的需求。对于信息科学相关专业的高年级大学生学生来说，学习这部分知识不会有太大难度。
- **深入一线**：学生普遍缺乏应用数据科学的实战经验，而数据科学家需要用真实的项目和数据来培育。因此，掌握算法模型的学生一定要尽快去有数据、有行业需求的一线历练，多跟客户和行业专家沟通学习，多观察、多动手处理数据，逐步建立起对数据的深入理解，熟悉对平台的操作以及对模型特性的直观感受。可以在课程完成后选择去工业界实习，或者参与高校与工业界的联合项目。
- **应用闭环**：要经历数据的采集、接入、管理、分析、展现、决策和反控的全链路，避免只做其中的模型设计环节。如果前面数据处理不当，会让本该有效的模型失效。另外，如果只作其中的模型环节，可能会理想化地脱离实际约束，使得模型无法工作。最后，如果不能将结果有效呈现给客户，数据科学家也不能得到反馈，导致模型不能迭代优化。
- **价值体现**：数据科学家设计的解决方案一定要在业务关注的领域，在成本、效率、用户体验中的至少一个方面直接创造价值。如在智能城市领域，政府关注城市的安全、稳定和发展，那数据科学家设计的方案就应该在保障城市安全方面降低成本、或提高管理者的效率，或改善工作人员的体验；也可以在促进城市发展方面提高政府资金的利用效率、降低资源投入等。避免只做到中间结果，看不到直接的业务价值。

6. 实战案例

下面以用大数据治理空气污染为实战案例，来剖析数据科学家如何结合行业知识和数据科学来解决开放式问题。

看问题：首先，弄明白这个问题为什么重要？大约十年前，由于环境、经济和人们对健康的重视程度等因素的变化，空气质量（尤其是 PM2.5 浓度）备受关注，一度还成为指引交通出行、厂矿工作和学校运行的指挥棒。空气污染如不能治理好，不仅影响人民的健康，还容易导致高端人才流失、吸引发展要素困难的局面，并造成社会舆论，关乎国家稳定。

其次，搞清楚导致问题的因素有哪些，即污染物从哪来？为什么会积聚？根据环境学的相关文献以及跟多位环保学专家和政府管理人员交流，得知污染源包含厂矿排放、交通尾气、餐饮排烟、烧煤供暖、土壤挥发等。污染物产生方式有三种：本地排放、外部扩散而来、以及在大气中发生二次化学反应而产生的污染物。导致污染物积聚的原因是污染物的产生强度要大于其被自然界消化（如扩散开或被吸附）的速度。因此，除了污染源和污染物产生的形式，扩散和吸附条件也是一个很重要的因素。

然后，了解行业的解题思路。要彻底根治空气污染就要理清现状、预知未来和回溯历史。理清现状指实时监测细粒度的空气质量，知道城市中各个角落的空气质量的现状；预知未来指能够预测未来空气质量的变化；回溯历史指搞清楚问题的根源，即从哪来、如何治。后续，我们以理清现状为例来介绍其分析过程。

为了做到实时监测，环保部门在城市中安装了一些高精度的空气质量监测站点，但由于价格昂贵、且需占据地理空间、后续维护成本也较高，此类站点的数量有限。由于污染源的分布和大气扩散条件在城市的各个角落均不相同，城市中的空气质量也存在巨大差异，非常不均匀。有可能全城的平均值还是优，但某个社区的空气质量已经达到轻度污染。没有细颗粒度的空气质量作为支撑，后续的预警、整治等工作将无法精确开展。因此，政府需要知道每平方公里甚至更细粒度的空气质量。由于不可能安装如此之多的监测站点，传统方案只能结合机理模型做一些假设推测。

再者，深入学习具体方法，吸取其精华，补充其不足。传统的方法有基于物理学的机理模型，也有基于化学的成份分析模型。但由于空气污染既有排放和扩散产生（物理过程），也有二次化学反应（化学过程），单纯的物理模型和化学模型都无法准确模拟空气污染这一过程。此外，物理机理模型需要预知污染源信息，并对风场作简化假设，这两点在真实世界很难成立。排放污染的工厂为了躲避惩罚会隐藏其排放行为，汽车尾气和餐饮排烟更是无法收集；大气在城市楼群中的流动更是异常紊乱，与简单模型的假设相差甚远。虽然这些方法有不足之处，但为我们后续设计模型提供了很好的思路。

最后，用行业的语言告诉行业专家，为什么基于数据科学的方法比传统方法好。其实无论是基于数据科学的方法还是传统机理模型都是在用模型拟合数据，思路是一致的。如表 1 所示，对于简单问题（如重力加速度等），根据少量数据样本，加上人的经验，便可构造出经典模型来很好的拟合问题（如 V=gt）。这些经典模型通常都可以用一些比较简洁的公式来表达。

表 1. 经典模型和数据科学对比

问题	因素	样本	手段	结果
简单	少	小	经验假设	经验公式
复杂	多	多	机器学习	AI 模型

当问题变得复杂，涉及的因素非常多，需要的数据量也变得越来越大时，依靠人的观察和经验来设计模型拟合数据就越来越难了。此时，用机器学习这个工具从数据中学出一个复杂的公式来精确打击这个问题，其本质还是在用模型拟合数据。用数据驱动的方法，通过对数据和特征的选择，即借鉴了经典模型的思想精华，又避免了依靠与现实有较大偏差的经验假设。

想数据：由于知道了大气污染要考虑污染源、污染物产生方式和扩散条件等，因此，我们选择的**数据应该尽量涵盖或间接反应这些因素，同时还要考虑获取这些数据的可行性**。这里我们选取了空气质量监测站点的历史和实时数据、兴趣点（如楼房、加油站、公园、厂矿、商场等）、路网数据、出租车的轨迹数据、天气预报和实报数据。

兴趣点、路网反映了一个区域的地貌、功能，出租车的轨迹数据蕴含了区域内人们的出行规律（第 3 节已做解释），进一步强化了对于区域功能的推断。因此，兴趣点、路网和出租车轨迹数据就隐含了区域污染源的分布和扩散条件。同时，出租车的数据间接反映了路面的交通流量，但由于其数量远小于私家车，因此不能用这个轨迹数据来直接推断全量尾气排放量，而要结合路网和兴趣点来补足其信息的缺失，共同隐含、关联整体交通流量和尾气排放。这些都是思路，不需要、也不可能确切的把每个指标都计算出来，而是要借助大数据“不确定”+“不确定”推出“确定”的思想，用领域 A 的数据去解决领域 B 的问题。

关联模型：这部分注重**基于行业知识来选择特征和设计模型结构**。根据之前的分析，特征方面，从兴趣点数据中提取了厂矿、公园、学校等重要类别兴趣点的数量，以及建筑密度、空旷度等反映扩散条件的特征；从路网中提取了交叉路口个数、路网密度、不同等级道路长度等影响交通流量的特征；从出租车的轨迹中提取区域内、不同时间段上车和下车的人数、行驶速度、速度的方差等。这些信息隐含了交通尾气排放的情况，如红绿灯很多（路口数），车辆拥堵（车速）、车辆走走停停（车速方差），此时尾气排放最为严重，车辆越多（车道数、道路总长度）排放越多。

模型方面选择了基于 Co-Training 的多视角学习模型。从污染物的产生角度来理解，一个空间分类器模拟外地扩散，一个时序分类器模拟本地排放，Co-Training 的迭代近似二次化学反应。从空气质量的相关性来看，一个地方的空气质量即有空间相关性，会受到周边地域空气质量的影响；同时也有时间相关性，受过去一段时间空气质量的影响。

从数据科学的角度来理解，空间分类器接受路网、兴趣点等空间特征，拟合空气质量的空间相关性，在地理空间进行非线性插值；即根据一个地域周边地区的空气质量信息来判断该地域此时的情况。时序分类器接受气象、交通和人们出行等跟时间相关的动态特征，拟合空气质量的时序相关性；即根据一个地域过去一段时间的情况来推断现在的情况。两个分类器从不同的角度来判断一个地区的空气质量，互相补强各自的弱点。选择这个模型的另一个原因是因为已有站点数量有限，训练样本有限，必须采用半监督学习的方法

来解决样本不足的问题。可见，当把问题分析透彻后，数据科学可以跟经典模型思想很好的融合，既能提高结果的精度，也能获得行业认可。

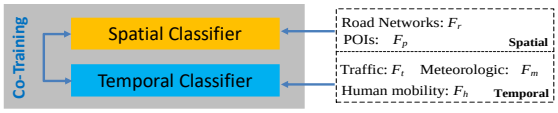

图 3 基于多视角的空气质量推断模型

利用已有站点的空气质量数据训练好一个初步模型，就可以基于路网、兴趣点、气象、出租车轨迹，开始对没有建设监测站点的任意地域开始推断。之后，还需要考虑到预测结果的展现。如以 1 平方公里为最小区域，展示全国的空气质量则需要 960 多万个格子，浏览器无法直接显示。因此，这里又涉及到基于四叉树的数据管理算法和可视化技术的结合，根据不同的视野层级来高效、动态聚合空气质量信息。

平台部署：利用平台实时接入各种数据，部署设计好的管理、挖掘和可视化模型，并把这些模型有机的组合起来，为全国 300 多个城市提供服务。服务可以在政府侧的大屏、电脑端展示，也可以为各类移动应用提供接口。为了保证性能，哪些内容需要放到缓存（如 Redis）、哪些需要用到分布式计算环境、哪些内存数据需要用到索引结构、哪些内容放到磁盘上、要用多少虚机服务器等，这都需要对平台的性能和使用方式非常熟悉，否则之前设计的数据科学解决方案根本无法运行。

结束语

数据时代已经来临，如何发挥数据的价值将关乎行业发展、国家命运以及世界格局，需要一批优秀的数据科学家来承担时代赋予的使命。数据科学家需要快速学习行业知识、深度理解数据、精通各类数据模型、熟练运用大数据平台，并具备数据侧端到端的解决方案能力。同时，数据科学家还要树立正确的数据观，并不断提升认知能力、学习能力、创新能力和沟通能力四大基础素质。数据科学家以数据科学为方法论来认识和探索世界，解决各类行业问题、创造社会价值，不断扩大数据科学的外延，并在此过程中，不断研究、创新数据的采集、管理、分析、挖掘、展现的理论和方法，深化数据科学的内涵。

参考文献：

1. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century.
2. Nicolas Jing Yuan, Yu Zheng, Xing Xie. Discovering regions of different functions in a city using human mobility and POIs. In proc. of 18th SIGKDD conference on Knowledge Discovery and Data Mining, 2012
3. Zheng, Y., Liu, F., Hsieh, H. P. U-Air: When Urban Air Quality Inference Meets Big Data. In proc. of 19th SIGKDD conference on Knowledge Discovery and Data Mining, 2013.
4. Li, R., He, H., Wang, R., Huang, Y., Liu, J., Ruan, S., He, T., Bao, J., Zheng, Y. JUST: JD Urban Spatio-Temporal Data Engine. In proc. of the 36th IEEE International Conference on Data Engineering.
5. Zheng, Y., 2015. Methodologies for cross-domain data fusion: An overview. IEEE transactions on big data, 1(1), pp.16-34.
6. Zheng, Y., Urban Computing. MIT Press, 2019.

郑宇 博士、京东集团副总裁、京东科技首席数据科学家、IEEE Fellow、美国计算机学会杰出科学家，具有超十五年中美领先科技公司的管理和产品研发经验，是城市计算领域的先驱和奠基人，也是大数据、人工智能领域的领军人物和实践者。他还是上海交通大学讲座教授、南京大学、香港科技大学等多所知名高校的客座教授。他担任人工智能顶尖国际期刊 ACM TIST 的主编、IEEE 智能城市操作系统国际标准组主席、国家重点研发计划项目首席科学家、总负责人。加入京东后，他开创了京东智能城市业务板块，从 0 到 1 搭建了业务体系，为全国 60 多个城市提供了技术服务。他带领团队设计和研发的城市操作系统成为雄安智能城市建设的数字基石；他作为总负责人在南通建设了中国第一个市域治理指挥中心，成为市域社会治理现代化的国家级标杆。2013 年他被 MIT 科技评论评为全球杰出青年创新者；2014 年，被美国《财富》评选为中国 40 位 40 岁以下商界精英。2021 年，被授予首都劳动奖章。同年 8 月，当选 KDD China 主席。根据 AI2000 公布的 2021 年排名，他在数据挖掘领域影响力位列中国第一、全球第八。