“数据科学家”或许不再性感 但“数据团队”的产业化才刚开始

——访领英全球数据科学团队负责人

■作者 魏子敏 夏雅薇 牛婉杨

本文为清华数据科学研究院联合大数据文摘发起的年度白皮书《顶级数据团队建设全景报告》系列专访的第二篇内容。

《报告》囊括专家访谈、问卷、网络数据分析，力求为行业内数据团队的组建和高校数据人才的培养提供指导性意见。

定下“顶级数据科学团队”这个研究话题时，我们第一时间想到了领英（LinkedIn）。

2008 年，正是在这家公司，DJ Patil 建立了全球首个真正意义上的“数据科学团队”，并开始用“数据科学家”（Data Scientist）这个词来描述这些 Data man 们的工作性质。

在这之后，“数据科学家”开始被誉为 21 世纪最性感的工作，也成为全球技术精英们近年来最理想的职位之一。

尽管已经过去了十多年，但当我们请领英全球数据科学团队负责人许亚给数据科学团队下个定义时，她还是表示，这不容易。

的确，尽管数据科学在学术领域的概念 50 多年前就有了，但作为职业，相比业内更多成熟的团队和路径，这依然是个相对很新的概念。

不同公司和团队领导人对于“数据科学团队”的定义范畴大相径庭：

从时间维度来看，当年研发出 Hadoop、Kafka 的人会称自己是数据科学家，但是现在这些大数据底层技术都变成了偏基础设施的内容，在狭义概念上，已经不再属于数据科学团队的范围；

随着这个领域囊括的范围越来越多样，数据对于每家公司的重要性也都只增不减，数据科学的“嵌入”性越来越高，边界也越来越模糊。

尽管如此，谈及领英这些年“数据科学团队”的定位和建设，许亚依然有自己非常清晰的思考。

“对于领英来说，数据科学团队的整体趋势更加走向专业化，他们的职责不再是建立数据基础设施或平台，而是怎样去使用数据科学和工程来最大化数据的价值。”

这是许亚对数据科学团队任务的要求。

那么到底如何让数据的价值最大化呢？从团队运作方式、商业影响力设定和社会责任等角度，许亚给出了领英的答案。

“嵌入式工作，中心化管理”，数据科学团队更加“专业化”“工程化”

和多数互联网公司一样，领英的数据科学团队规模也在近几年飞速增长。许亚表示，仅是近两年来，领英的数据团队扩张了近一倍，从 150 人增加到目前的 300 多人。

许亚提到的数据团队是指领英中心化的数据科学部门。如果用一句话来概括领英的中心数据科学团队的运作方式，那就是“嵌入式工作，中心化管理”。

和国内不少互联网公司将数据分析师归属于业务 BU、向业务主管汇报不同，领英的数据科学团队成员由许亚的中心部门统筹。虽然在项目工作上，数据科学家们依然会在工位分布和职能上与业务部门紧密联系，但是从职级从属上，都直接向许亚汇报，不同领域的数据科学家在工作中会有交集，还会一起开会。

其实领英的数据科学团队的设置也不是一开始就如此，随着领英数据科学团队定位的变化，数据科学团队也从最初的产品组，移到了现在的工程大组。

值得一提的是，目前领英的数据科学和人工智能团队都在同一个大组里，许亚表示，数据团队和人工智能 / 工程团队是紧密相连的。

这也从一个侧面说明，随着对数据科学团队的需求逐渐增大，数据团队的工作会越来越“工程化”。跑的数据会越来越多，对工程团队的需求也会越来越大，需要对工程团队越来越多的要求和技术定位。

近年来，各大公司越来越意识到数据的重要性，已有的数据科学涉入领域在进一步扩张。数据团队之前最常被用到的部门是市场和产品，但是基于领英本身的数据基因，近几年的一些产品也对之前没有用到数据的地方做了数据驱动的尝试。

例如，与架构工程部门合作的数据团队会去衡量工程架构的建设是否有效率：每年跑大数据的硬件设备花费很高，怎么样在时间上做规划，让硬件 /GPU 等更有效的发挥价值。

在人员构成上，和十年前相比，领英的数据团队也更加专业化了，底层架构人员也从数据科学团队分离了出来。

目前领英的数据科学团队也根据员工不同的专业领域设立了三个工作方向：

工程专家：可以很有效地建立起数据管道（data pipeline）和数据流（data flow）；

算法专家：在预测、算法领域的技术咖；

业务专家：有很强的业务属性，将数据见解和公司战略结合起来；

由于工作侧重不同，在管理的过程中也会有意地区分这三类数据科学家，并且保持各类员工的竞争力。

许亚提到，她的团队内部更多是自下而上的工作文化。她不会给团队指派任务，因为每个组会自觉地告诉许亚他们想达到什么样的目标。对于一些大的项目，一般需要跨部门合作，各部门的领导达成共识，分配资源来一起实现这个目标，是自上而下和自下而上的结合。

三大 KPI 指标，量化数据团队工作

相对复杂的构成和与业务团队的紧密性，给数据团队设定商业影响力和发展路线不是一件容易的事。

许亚表示，两年前她接手领英数据团队后做的第一件事就是拟定了团队成功的三要素。虽然数据团队的价值有时候很难量化，但是有三个指标可以作为探讨的基础。在数据团队内部不同组可能会有不同的侧重，但对大部分组来说这三个因素都很重要。

·数据易得性和工作效率

数据易得性，指的是当外界需要数据的时候，获得这些数据的难易程度；工作效率，指的是一个人的工作是否可以提升整个团队的工作效率。

许亚表示，数据科学家之前被人诟病过于追求新鲜感，喜欢挑战高难度问题，但做完 MVP（Minimum Viable Product）后没有维护迭代的习惯，永远都在追逐下一个新难题。数据团队拥有许多数据资源，比如原始数据，指标数据，数据模型，数据可视化。

当外界对这些资源有需要的时候，如何能够保证这些需求能够随时被满足？软件开发有一系列衡量数据获取难易程度的指标，比如 SLA（Service-Level Agreement）的达标率就是一个很好的量化指标。

有些数据科学家做了一个很不错的分析，但是不太关心怎么把这个分析过程自动化，所以每次有人提需求的时候就需要有人再手动跑一次模型，其实都是重复劳动，不同的人在做相同的重复劳动。如果这个分析实现了自动化，大家都可以享用，其他人就不需要花太多时间精力在这个模型上，整个数据科学团队的集体工作效率都提高了。

以前许亚的团队也缺少这种分析自动化产品化的意识，所以她把这个设置为成功三要素之一，强调这种意识的重要性。

·战略化思维

战略化思维，指的是数据分析结果对公司重要战略性决策是否有指导作用。

许亚的数据团队和公司很多高层会打交道，因为他们团队有一个很重要的职责就是通过数据来确保公司重要决策的大方向是准确的。比如他们需要了解用户在疫情期间是如何使用领英服务，如何通过领英的产品获取价值的。

许亚认为，在疫情后，用户的行为多少会发生一些不可逆转的改变，数据可以帮助团队更好地去学习用户行为变化，从而在战略上指引公司对哪些领域进行重点投资。不管是产品开发还是市场战略的决定，都需要依靠数据。

·直接商业影响力

直接商业影响力（Direct Business Impact），指的是工作成果对公司商业目标的直接影响力。

每个部门的工作开展是和公司要实现的大目标息息相关的，领英有公司层面的四个核心指标，数据部门在计划工作的时候，需要考虑如何对公司的商业目标产生积极影响。

AB Test：用数据来证明一切

我们都知道，企业在做产品 / 功能测试时一般都会用到 A/B test，即分为两组用户，一组对照组，一组实验组。对照组采用已有的产品或功能，实验组采用新功能。要做的是找到他们的不同反应，并以此确定哪个版本更好。

A/B test 能对大范围的事情进行测试，例如亚马逊对个性化推荐进行 A/B test 后，发现个推能显著提升收益；谷歌在对搜索广告进行排名时也用到了 A/B test。

那么对于领英来说，A/B Test 在领英的产品设计中又扮演着什么角色呢？如何影响产品决策呢？

可以这样说，基本上我们在领英网站上能感知到的更新，领英团队都会做 A/B Test，有些是前端的改变，有些是后端系统的调整。当你打开领英 APP，从搜索栏，搜索引擎算法，底部导航，到页面文字大小，这些都是经过 A/B Test 的。

领英的产品文化以用户为主导，领英自己不会去假设用户喜好，一切都通过数据来说话，而不是靠谁的直觉。除了看得到的东西，后端用户看不到的，领英也会进行 A/B Test。比如打开 APP 要加载内容，需要从后端系统里获取数据，每次获取 20 条数据还是 100 条数据，这个决策就涉及到平衡与取舍，获取数据越多，页面加载时间越长；获取数据越少，用户浏览的时候就需要频繁刷新。所以到底一次获取多少数据，领英还是通过 A/B Test 来决定。

还有一个简单的例子，当领英对一个数据中心的开关做决定时也依靠 A/B Test，比如一个用户发起数据请求，这个请求该发送到哪个数据中心来处理呢？这种情况下用户到数据中心的距离就是一个很重要的考虑因素，最终领英会通过做 A/B Test 来选择最优化的基础设施方案。

虽然数据团队是 A/B Test 方面的专家，在这方面更有经验，但因为领英有非常完备的 A/B Test 平台，可以解决大部分实验需求，包括实验设计、实施和分析，所以数据团队不需要介入到每个 A/B Test。

这对推广实验文化和数据文化很有帮助，因为大家都可以去做实验，享受数据和实验带来的好处。领英内部每天大概有 100 个新实验在进行，数据团队无法关注每个实验，但是会集中关注一些重要的实验，深入参与到研究和分析工作中。

在领英以数据为主导的文化浸染下，长远来看所有人都受益于这样科学的决策机制。也因为有 A/B Test 的文化，所以可以跳过争论，直接做个 A/B Test 就见分晓了。整个过程简单公正，方案落选的组也可以通过这个机会学习到一些关于用户的新知识。

A/B Test 提倡数学引导的创新，这种创新不取决于谁的职位更高，因此任何团队都可以放心大胆地去做测试来发掘新点子。

领英作为一个社交平台的社会责任：给每个人公平的机会

在许亚看来，维护公平是一个很有挑战的课题，因为你很难明确定义公平。

“当我们在说公平的时候，我们在说公平的机会？公平的结果？还是公平的待遇？我之前看过一个有意思的问题，给三个不同高矮的人提供凳子，在公平原则下，你该给他们提供同样高度的凳子？还是提供不同高度的凳子让他们坐上去之后一样高呢？我很难说这个问题有一个绝对正确的答案。”

领英对公平的定义是，拥有同等才能的两个人，应该获得同等的职业机会。而不受到种族或者自身人脉的影响。过去两年时间领英做了很多努力来解决公平问题，取得了不错的成果。

首先，领英很重视可量化、可测量的指标，因为如果一个问题没有被数据抓取到，就很难注意到。

例如，每次领英发布新产品，都需要通过量化的指标来测量这个新产品对用户带来影响是否公平。一开始领英的测量指标比较粗线条，他们会看这个产品平均下来对用户是否有积极影响，但如果细看数据，有可能这个产品只对一部分人有益，但会损害另一部分人的利益。因此，后来领英采用了一个指数来衡量是否在一个群体内无意间引入了不公平因素，也就是对每个新产品，领英想知道其带来的提升是否是公平的。

其次，领英关注现有平台上是否存在公平问题的盲点。

例如一个以男性为主体的数据集，训练出来的模型就更倾向于男性，这是一个隐蔽的不公平点。很多猎头和 HR 用领英产品来招人，如果算法推荐的候选人都是男性，女性就失去了公平的竞争机会。

大概一年前左右领英推出了一个代表性指数来衡量推荐结果对整体数据集的代表性。比如所有可能候选人的男女比例是 1:1，那领英给猎头推送的前 100 位候选人的男女比例也应该是 1:1。有了这些量化指标，领英可以更好地规范和规避不公平的举措。

许亚还给我们举了一个例子。之前领英有一个内推功能，当某个人想申请 Google 的工作，会收到提示说我的一位好友在 Google 工作，我可以找他要个内推。

上线初期，领英内部对这个新功能很满意，因为可以帮助那些有广泛人脉资源的人更快找到工作，后来领英意识到这个功能会让那些没有人脉资源的人更难找到工作，所以就关闭了这个功能。取而代之的是领英推出了一个新工作快速提示功能，一个新职位刚发布出来，领英会立刻给所有对此类职位感兴趣的用户推送提示。这个功能不仅能帮助所有用户更快找到工作，对那些关系少的人尤其有帮助，因为他们的消息相对更闭塞一点，所以这个功能能让更多的人受益。

最近领英也开源了这套技术，希望能助力其他公司去构建一个更公平的社会环境。

随着近年来数据泄漏事件频频爆发，数据隐私和安全问题被推上了风口浪尖。许亚也跟大数据文摘聊了聊领英在保护用户的数据隐私方面都做了什么。

领英全球有超过 6.9 亿用户和 5000 万家企业，领英的愿景是为全球劳动力市场中的每一位创造经济机会，通过将所有在领英平台发生的行为数据可视化，进而打造全球“经济图谱”。因此用户数据对领英至关重要，如果没有用户的信任，领英就没有办法去实现他们的愿景和使命。

所以在 GDPR 这些开始之前，领英在保护用户隐私上已经有了很多投资。许亚提到，除了实现规定里的要求，领英也用一些很前沿的技术去确保不泄露隐私，比如现在认为是数据隐私保护的“Gold Standard”——差分隐私（Differential Privacy）。

大家经常说到保护隐私，比如说把一些个人信息隐去了，其他人看不见，我就没有隐私泄露了，其实不是这样的。

差分隐私只是一种保证。假设你的信息在一堆数据里面，如果把这些信息删掉，再运行同样的一些算法，从数据当中得到的两个的结果都是一样的。相当于你的数据在或者不在这个数据库里面，最后对于得到的信息没有影响。这样用户就不需要担心他们的数据隐私被泄露。

领英三年前就开始针对数据隐私问题进行一些重要的研究，同时也有一些比较成功的应用，例如最近一个针对广告商的产品，客户想要用领英的 API 去获得一些信息，比如用户互动量前十的文章，像这样一些集合的信息，领英也用差分隐私去确保用户的信息不泄露。

最后，从整个公司文化上面来说，许亚透露，除了去实现数据保护条例的一些要求，领英也用到了一些很前端的技术，来确保用户的隐私不被泄露。另外，领英也十分重视在数据分享方面的问题，并表示会对此加强技术防护。

采访过程中，许亚多次提到领英的社会责任。今年，一场突如其来的疫情，全球的劳动力市场都受到了不同程度的影响，不论是就业还是工作方式都迎来了一种新常态。领英利用数据优势，实时展现劳动力市场的趋势变化，帮助个人更好地应对当下的不确定性。在分析数据时，领英还发现不同分组内的用户受到的影响程度不一样，比如刚入职场的新人会受到更大的冲击，疫情对女性的负面影响可能大于男性。

通过数据观察到这些问题后，领英数据科学团队和业务部门迅速沟通，快速响应，针对各个市场及时提供了一系列有针对性的服务来帮助这些人，让每个人都能在自己能力范围内获得平等的工作机会。

“这是领英作为一个职场社交平台的社会责任。”E

BIG DATA DIGEST 大数据文摘

（本文由《大数据文摘》杂志授权转载）