4亿用户的LinkedIn数据产品设计原则和架构实现

李海鹏 AI前线 2016-06-15

在苹果WWDC大会之前，微软宣布了一个大新闻：262亿美元收购企业级社交平台LinkedIn，较之最近一次收盘价溢价50%。收购之后，LinkedIn将与微软Azure云平台和Office 办公套件深度整合，还将与微软其余企业级业务协作，并帮助微软**实现社交梦想**。

早在2003年的时候，LinkedIn创始人因为“人际关系管理将在商业社会发挥巨大作用，而互联网能为其提供最好的工具”这个初心走到一起，创办LinkedIn。

在微软后妈收养LinkedIn这个新闻发布之时，有人借机将并购溢价原因归于“增长”、“数据变现”，让人忽略了最重要的原因：LinkedIn是全球最大的职场社交网站，LinkedIn的社交属性，人才人脉提供了不可估量的价值！

而我们所提供的这篇内容，正好可以让大家看看LinkedIn如何利用4亿注册用户、数十亿用户人脉数据创造数据产品，为人才流动和公司品牌创造价值。在第三部分，李海鹏同时讲了支撑这些产品的技术架构实现。

本文整理自**LinkedIn Sr.Manager李海鹏**于QCon2016北京的演讲：《**创造数据产品驱动商业价值**》。作为现任LinkedIn资深经理，李海鹏带领团队开发的数据产品，对LinkedIn营收的高速增长做出了巨大的贡献。在这次演讲中，李海鹏分享了 LinkedIn 如何从数据中挖掘价值的一些案例。同时也探讨了LinkedIn 是如何把商业、数据和开发相结合，通过数据产品的形式来驱动商业价值的。

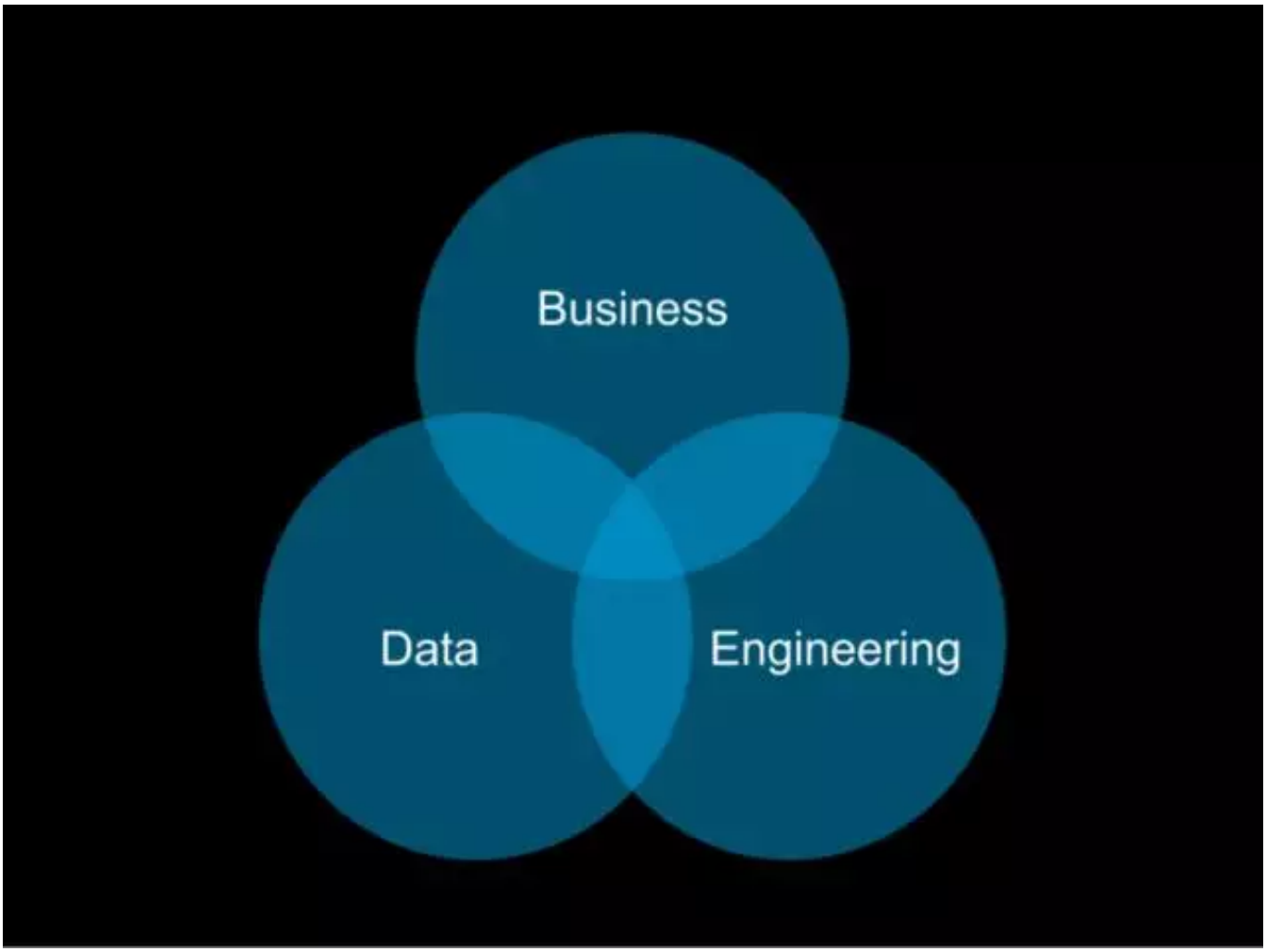
自我介绍

首先做自我介绍，我接触大数据大概是十年以前，是一个被迫的举动，当时做广告分析，广告数据特别多，被逼无奈，当时传统数据库都用不了。后来做亚马逊做了电商广告平台方面的公司，过去四五年在LinkedIn带领团队做数据驱动的产品。

所以我的背景跟变现的东西比较多，跟钱打交道比较广，无论是广告、电商还是现在LinkedIn做的东西，我今天选的话题是想给大家看更多的数据应用，我希望做到的并不是让大家能够把我说的东西运用到你们的工作里面。**因为毕竟没有另外一个微信这样的公**

司，或者很少有第二个LinkedIn，但希望给大家启发，想一想在你的公司数据里面哪些是可以参考我们的例子，运用到系统当中，想一想怎样产生这样的价值。

思维框架



首先讲一下思维框架的事情，有很多很多人跟我说，“Leo Li，LinkedIn的数据应用做的特别好，为什么做的这么好，我们怎么把这个事情做好”。

根据几年前的想法，我把公司里面的人分为**三类**：第一类是Business技能的人。第二类是对数据比较敏感，Data类的人。第三类是Engineering的人。我看到一个好的公司能够把数据应用做的非常好，一定是把这三方面结合起来的。这是我看到硅谷很多成功公司的例子，大家希望把数据应用好的话，不要单纯从一个角度切入。有的人就在数据里挖半天，没有和**业务结合**起来，有可能和业务结合起来也没有按照真正的方法去做，这是方法论的一个看法。

Business:

给大家讲下LinkedIn的Business，会给大家举一些我们做的例子，最后给大家分享我们在Engineering框架下，包括数据系统，包括前后端框架下，到底现在用什么东西，走过哪些坑，真想开发的话起码把我们走过的坑躲过去。

经济图谱：

我们想做的是做**世界经济图谱**，很多人在做着自己不喜欢的工作，并且不知道怎么选择自己更喜欢的工作。

我们有**4亿注册用户，每周大概有20亿用户更新，数十亿用户人脉**，这是LinkedIn的一些数据。

给用户提供价值：

希望他通过LinkedIn建立**职场关系**，对你的职业生涯特别有帮助，其次希望你能够同步职场知识。现在国内很多公众号专门翻译国外硅谷公司的信息，在LinkedIn平台上学习很多职场相关知识和技能。最重要的是希望能够帮助大家找到你们真正找到的一些工作。这是99%的人对LinkedIn的看法和理解，到现在都是2C端的内容。

如何变现？

LinkedIn的变现，首先讲一下数据驱动的概念，**在硅谷任何一个公司都会把自己作为一个数据公司，公司里有很强的数据驱动文化**，我们公司的CEO早上起床第一个看的是公司的KPI，发现问题，并且询问。

我们很多变现的模式也是从数据里面找到的。比如招聘解决方案，一开始很多人用这个平台，我们去分析到底人们用这个平台干什么。好比发现很多**猎头**在上面找他希望招到的人，我们在上面做招聘解决方案给这些猎头，大概占我们营收的60%以上。以后我们又看到了很多**营销人员**到我们平台上做他的产品推广，尤其是2B的，比如卖云服务，卖服务器之类的，后来我们做了市场的解决方案。一开始大部分是2B的，是不是我们的解决方案都以2B为主？后来我们发现，突然间公司里平台上出现了很多奢侈品的广告，做营销的人嗅觉特别灵敏，因为他发现他有一定购买力的客户都在这个平台里面，所以既是2B，也是2C的混合。最后是销售解决方案，用户互相交互做出来的产品。学习解决方案是我们去年收购Learn，告诉他通过学习那些技能能够达到这方面。

商业模式以及数据的重要性

这是2B端的一些产品。基于这样的业务模式，数据对我们是极为重要的。

首先是有很多的用户和业务增长，伴随着很多用户和业务增长，它们产生了很多数据，现在由于大数据收集比较便宜，无数数据都被收集起来。**最重要的是紫色和红色这块**，如何通过大量数据的收集和积累，创造更多相关的产品和服务，让更多的用户愿意到你这个平台上面来。这是一个闭环的正向积累。很多公司，尤其创业型公司，或者偏小发展型的公司，基本上环在这里闭不上，从蓝色到紫色这里怎么做？买流量，补贴，烧钱，但这个环闭不上，导致很多公司做不下去了。从原始数据的收集到从数据中提取产品价值，这块没有做的很好。

三种主要的数据类型：

第一个是有**用户身份数据**，和微信不一样的地方是，你一上LinkedIn就会有很多信息，里面有学校、职位、技能等等各个方面的信息，我们把信息做深度挖掘，能够提炼出很多有价值东西，一会儿给大家展示。第二个是**用户行为展示**，用户在我们的平台上点击浏览等等，我们可以把这些数据结合起来开发产品。最后一个是**社交数据**的概念。

LinkedIn的数据产品

了解这三种数据类型我就想给大家展示一些干货，让大家看看我们之前做了什么数据产品，产生什么样的价值。

社交关系图谱：

这是LinkedIn的**社交关系图谱**，和刚才国惠分享的比较相似。这是我的社交关系图谱，这个可能是我的工作同事，蓝色是我上一份工作同事，绿色可能是我的大学同学。这张图有很多不同的方法，假如这个数据放进来，看到我跟张三有一个连接，就会把我跟张三的距离拉近一些，看到我跟李四有连接，就会把我跟李四的距离拉近一些，假如看到张三和李四有连接，也会把他们的距离拉近一些，然后投射到二维空间里就会看到这个图。

我们看到这样一个有意思的图，当时非常惊讶，没有用任何的身份数据，只是通过社交数据就可以把很多信息表达出来。数据科学家这个词最早是从LinkedIn走出来的，这张图也是从那个时候开始功不可没的一个成果。

这个图没什么意思，讲讲这个图到底应该怎么用。如果大家想到只通过社交数据就可以找到这张图的话，刚才Randyling给大家看的，不仅仅可以做一个我的一维好友圈子。假如LinkedIn所有公司员工放到这里，甚至把在场所有参与QCon的人都放到这个图里面，自然而然就会产生不同人群之间的关系和距离，可以从一个人的一维空间延伸到任何一群人，非常正向的应用。

很多公司大量用户提供的数据，但是他不会用，就扔在那里不用了。我们也一样，简历给我们了，很多的信息。拿公司举例。假如我在微软公司，有多种写法，写他的总部，也可以写微软中文，甚至拼错了我的公司写法。但后台处理的时候我们希望把这个人的数据在后面做标准化，这个时候这种社交图谱背后的数据就变得非常有意思了。**假如一个人拼错公司名字**，通过

社交图背后的数据看到，这个人其实被包围在这样一群员工里面，他是这个公司的员工，这是一个正向的用法。

聪明的同学就会想到，如果这样用的话，反向也可以用，很多网络公司遇到的一个问题是，很多用户是假的用户，比如今天在LinkedIn说我在京东工作，但实际没有在京东工作过，在类似的图谱里面，我们通过这样的方式找到非常多的**假的用户**，这样的话也能对公司做一个很好的帮助。

讲到这里很多同学觉得还是没有什么用，因为我们公司不做社交，这就是很多同学没有把这个事情想清楚。

刚才我讲的都定义成狭义的社交概念，实际上我更习惯于把它定义成广义的社交概念，假如说我在跟某一个商家买了一个东西，其实这就是更广义的社交概念，甚至我打了一个Uber或者滴滴，我跟司机一起乘坐一段时间车，这也是社交概念。或者我在豆瓣上评论了一个电影的影评，有可能我跟其他评论人也是同样的社交概念，这样的数据延伸成广义的社交概念，这样的社交图谱概念可以帮助你找到很多信息。你们回去可以想一想，在你们的工作里面有哪些更广义的社交概念，通过社交图谱的方式，更深一步挖掘出它的价值。

Industry Trend：

我们怎么通过用户简历挖掘数据呢？当我们正确知道很多用户信息，首先结合招聘解决方案，我们会看到一个例子，Industry Trend。

这是讲的美国某一个行业，过去一年在美国的发展情况，这个行业总体增长是5.5%，细分到每一个地区，西雅图增长27%，纽约、德州等等，我们都会做一个排列。

这对很多公司做**战略性决定**是非常有帮助的，而且我们不仅仅做这样一个Industry Trend，我们会做出很多，供我们的客户使用。

Talent Intelligence：

我们分享完之后他就会问第二个问题，我知道了行业趋势，我更想知道局部的竞争。

所以我们就给他做了一个**人才情报**的Industry，这个主要是找到具有某种相关技能的人，以及他的数量分布和他的（英文）比较。比如这个技能，在LinkedIn有61000个，有3000人大概过去三年刚毕业的，新加入的，有25%是Managers or above。我们也给出了各地区的竞争比较，有的地区竞争非常高，但是有的地方竞争越来越弱。**有什么好处呢？**从微观角度讲，给公司招聘人才很好的指导，比如你想面对美国招聘人才，你应该知道如果在纽约和旧金山找你的人才，你的竞争是极为大的，达拉斯也有很多这样的人才，你可以去招聘，帮助企业找到他更容易招聘的人才。

结合上面两张图，我们大概类似的东西也有几十个，只是给大家看了很局部比较有代表性的特征，只做两个图挺简单的，但我们如果把所有的技能，大概有几万个乘以他的地区，乘以他不同的行业，**这个数据维度是极大的。**

Hiring and Talent：

这张图讲的是美国某一个科技公司过去一年之内人员招聘的情况。这个公司有1150个人，这是他的分布，有LinkedIn本身用户提供的数据，也有外部进行归纳处理，我们建立了除了中国以外最大的一个数据公司的数据体系，有很多数据为我们其他产品数据进行基石的服务。

人才流动与公司品牌：

这张图特别有意思，这张图给公司带来的价值是非常巨大的，我们把它叫人才流动，具体解释一下什么概念，这是比较机密的数据。

假如这个公司是Google人才流动的图，这张图就展现了Google和其他很多公司人才流动的情况，具体举一个例子。**假如第一个公司是Microsoft，这张图告诉我们的是Google从Microsoft过去一段时间里面招聘了353个人，反过来，从Google去Microsoft的人大概有80人**。假如这个公司是Uber，同样的时间里，Google从Uber招聘了77个人，从Uber去Google有164个人。

所以当你明白这个图的意思以后，我们从其他角度讲一讲这张图有什么用。

比如，我们开始卖招聘解决方案的时候遇到很大困难，我们跟人力资源说我们想卖这个产品，他们不懂。我们没有办法，只能把公司这张图发给他，很多时候第二天第三天就会有很多回应，说我们可以再继续聊一聊，**看看你还知道我们公司什么**。

这是为什么呢？首先，我们一直在硅谷讲的概念，尤其在互联网的创新公司，**人才很重要，所有公司之间的竞争都是人才的竞争**。这张图首先帮你找到很多你潜在的竞争对手，从战略上我们给他洞察和分析。

其次，假如Google和Microsoft人才流动是这样的，有300多人从Microsoft去Google，有80个人从Google去Microsoft，说明Google公司更好一些。就像羽毛球比赛打分是一样的，假如今天我跟林丹打一场羽毛球，我赢了林丹，我在国际积分狂涨，林丹赢了我就不会太涨，因为我是无名之辈。

由此我们相等，任何一次人才流动都是两个公司打了一场比赛，从统计意义来讲，我们一直认为人往高处走，水往低处流。我不相信每个人换工作的时候说我想换一个比较差的工作。

所以任何一次人才流动造成两个公司之间一场竞争关系，**这张图后面的数据我们可以把全世界一千万公司做很好的排名**，大家能理解我说的意思吧，这样后面挖掘出来的价值就非常好玩了。

其中有一个有意思的个例，我们每年会发布硅谷25个创新公司，预测一下，**预测都非常准**。这些创新公司要么就获得非常好的投资，或者是IPO做的非常成功。不是因为我们发布他们才成功，而是我们更早预测到他们招聘的这些人才，由于人才的引进给他带来的这些价值。

所以这是公司品牌流动，很多同学聪明的话又会往下想，流动的概念不仅仅限于公司和公司之间，我们可以想到行业和行业之间存在流动的概念，我们可以做一个细分，行业是怎么变化的。

我们也可以做技能流动的变化，很多QCon的人最喜欢讨论哪个语言最好，我到这个讨论一般不会吱声，因为后面有数据会看到是什么样的。所以从点到线到面你会找到很多价值，后面的数据就是一个简历数据，单一维度没有营养的数据，当你做好处理，并且你找到其中一个维度，在我们流动概念里面它是一个时间的维度，虽然看起来是一份简历，但由于每个人在不同时间做了不同的工作，有隐藏的维度在里面，找到这个维度。想想你们公司里面有没有隐藏的维度没有想到，能不能从点到线到面找到更多价值。

Trending Content：

这个是Trending Content，我们的平台上有很多人发布相关内容，市场营销产品有类似的服务。当时我们碰到主要的问题是客户不知道该发什么样的内容，或者该推广什么样的内容在平台上，因为很难知道哪些内容对用户更加重要，长远来看，我们做了这样的产品。

给大家举一个例子，这里面我选了一个行业，在过去一个月里面，这些相关的主题是非常受欢迎的，比如Region、Segment等等，可以找到哪些主题是他们非常关心和关切的，数据量是巨大的，因为我们是实时的，还要从无数的文章里面提取出来。但我们有一个好处，跟微信不太一样的，我们不是只是晒照片，很多人会在上面写文字，帮助我们从这些文字里面提取主题。

Full-funnel Sales Intelligence：

我们过去五六年时间一直做的一个方向，叫**销售智能**。

今天我看到很多中国公司也在开始做这方面的东西，大概两三年以前，这样的东西在美国和硅谷也非常多。前年的一个大会，上面五六家，甚至六七家，做的都是同样的事情，甚至说我们都是差不多的产品。

具体是什么事呢？我们认为任何一个销售的过程都是这样的漏斗过程，从Lead到Customer。

首先，我们做的第一件事是找到哪些公司，卖不同的产品不同的服务，或者提供不同用途，你的用户和公司不太一样，从2B概念来讲，产品公司有哪些，比如一千万公司在我们数据库里面，我们会做很多预测模型，预测哪些公司更愿意购买我们这个产品，这样的话就把一千万公司打分，我们的销售就不会随机打电话，会真正找到我们潜在的客户。

第二，是Who to contact，谁来买我们的产品，通过我们平台上的很多数据，我们可以直接找到用户，比如我想卖服务器给互联网公司，他是IT部门的头，肯定可以做这个决定的。

第三，我们首先找到公司以后就会继续找到到底哪个是联系人，找到联系人还不够，我要知道我们公司哪一个销售应该去跟这个联系人联系，很多时候我看到销售团队建设是按区域来分的，是有一定道理的，但没有很好的优化。因为从我们的社交网络上我们知道哪两个人的关系

更近，假如张三是某互联网公司IT部门主管，负责买服务器的事或者云服务，我和张三之间隔着李四，李四是我的大学同学，我就可以很好的通过这个过程联系到张三。

第四，找到销售人以后，你还要讲什么样的故事，不能说所有的用户用同样的模板，很多时候不同的客户在不同时段要采取不同的策略，举一个例子，假如一个公司前两天刚刚裁员20%，如果我们想卖给他招聘解决方案的话，我们要讲不同的故事，一个公司前两天刚刚宣布下一轮融资，肯定要讲不同的故事，很多PPT都是自动生成的，我们的销售团队是非常有效率的团队，很多公司的销售需要培训上岗，很多训练，很多人帮他做PPT。但在我们公司只需要半天，熟悉一下电脑就可以了，点五下十下一个PPT就自动生成了。

Churn Guard：

过去我们把从Lead到Customer都做到了，公司效率极大的提高，LinkedIn过去几年增长是非常迅速的，很多原因就是背后的数据应用和智能化，和Engineering合在一起。**就像一桶水，很多时候我们花很多精力往里注水，但客户流失就像流水一样，如果不把流水减少，无论你怎么去卖，公司的营收增长一定是非常缓慢的**，很重要的是Churn Guard。

第一步从Measure来讲，我们会把它分成不同的地区，让公司的销售团队看到

到底有多少可能性销售的流失。

第二步是做Predict，我们把它分成Low，Medium，Medium High，High，假如你有限的团队去解决客户流失问题，如果最终找到Churn Rate相当于多了三倍的人，这样更有效，非常好的方法。

最后，我们做了自动化的PPT，可以让我们的销售团队很快的点击下载，和客户交流，防止订单的流失。

这个上面是一个Manager的衡量，第二个是假如我是销售，颜色深浅代表Churn的高低，他可能是这个时间流掉的，但你在这个时间就要做了，要知道什么时候应该做什么样的事。最后告诉他为什么我们认为这个客户会流失，尽量给他归类，让他去讲一个很好的故事。有很多可以讲的，**但关键还是背后的这些系统，架构的东西很重要**。

Engineering：

我们面临的问题：

首先我们需要Scalable，需要很好的数据系统，帮你做很好的优化。第二步是可视化，最后是Web Application Framework。

经过演化以后的状态：

这是我们数据系统的图，这是从下到上的顺序。

如果你们从第三方买数据，你需要让它走到你的系统里来。另外是**Kafka**，从LinkedIn走出去的，我们当然要大量用Kafka做数据处理，然后放进来作为主要的存储地方，根据数据的不同用途，我们会把它用不同方法处理，比如有**Map-Reduce**，**Spark**，**Presto**。当我们有一个庞大系统以后，有很多高并发还有性能上的考虑，所以我们有很多专业零散的数据库，这是演化的过程，慢慢发现你需要添加越来越多专业型的数据仓储。

这是LinkedIn开源的，最重要的是你要把每一个都想明白，OLAP的时候我们用的是**Pinot**，**优势**是假如你放几十倍的数据，做前端语句查询，它可能在极短的时间内，几十毫秒时间内，就把结果给我们展示出来。

我讲的理念是，原来很多做技术产品的，先把数据放到一个地方再去查询，问题在于当你有很多数据维度，你的维度极大，不可能再做OLAP，像这种实时的OLAP相对来说就非常好用了。这里我们也有LinkedIn自己的开源项目，速度极快，支持并发。

Data Visualization Platform：

我们希望有一个简单的方式，简单的工具，让他不需要做很复杂的编程，BI这边用的是Reporting，做很多的数据产品。

Web Application Architecture：

Web Application Framework演化最重要的一点是从原来很大的架构走过去，我看也有相关专题分享，大家有兴趣可以多去了解了解，把我们的系统拆的非常细，它支持更好的协作，维护起来相对来说更简单一些。

前端现在用的是Play，我的建议是根据你们已有团队特征特长选择前端架构。

这些是我的分享内容，从三方面分享了我们真正的干货，没有任何的保留，我希望大家可以想一想你的公司里面哪些数据可以用什么样的方式提供价值。

演讲整理：孔伟凡

在后台回复“LL”即可下载完整PPT

大数据杂谈

ID：BigdataTina2016

▲长按二维码识别关注

专注大数据和机器学习，

分享前沿技术，交流深度思考。

欢迎加入社区！