【大数据与商业创新】传媒大学沈浩：

大数据时代-数据价值与商业创新

（本文选自微金融 50 人论坛演讲汇编系列之一：微金融的基础设施，独家推出，转载请注明出处）

时下，“大数据”是非常时髦、热门的话题。前段时间，全球复杂网络权威、“无标度网络”学者巴拉巴西（Albert-Laszlo Barabasi）写了一本关于大数据的新书——《爆发》。这本书出版的时候，我在推荐语中写道，“这是一个令人兴奋的时代，也是一个大数据的时代，数据科学让我们越来越多地从数据中观察到人类社会的复杂行为模式。以数据为基础的技术决定着人类的未来，但并非是数据本身改变了我们的世界，起决定作用的是我们对可用知识的增加。”

一、大数据时代的爆发

有人说，数据科学是 21 世纪最性感的职业。现在看来，此言非虚。如若想要了解如何从海量数据中挖掘出价值，就需要从数据科学、网络科学、空间地理科学和可视化等方面来诠释。

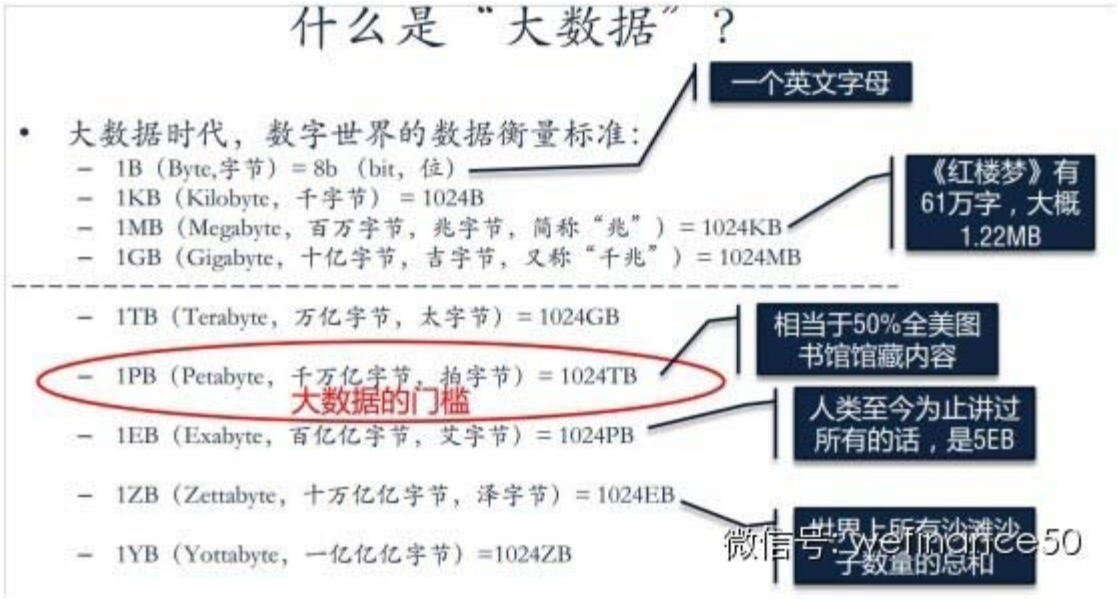

图 1 什么是大数据

《爆发》这本书核心观点是“人类行为的 93%是可预知的”。这是大数据时代背景下预见未来的新思维，阐述了如何从大数据中塑造未来美好世界。基于此背景，可以初步做三个判断：第一大数据时代真的来了，第二我们要热情拥抱大数据，第三大数据真的对我们的社会产生了重要的影响。解释大数据可以有不同的角度，包括隐私问题、个人信息安全问题等等。这些不同的角度思考大数据，但有一个共同的出发点，就是首先要拥抱大数据。资料显示，自人类有数据以来，90%的数据量是最近两年产生的。无需多言，大数据时代真的到来了。

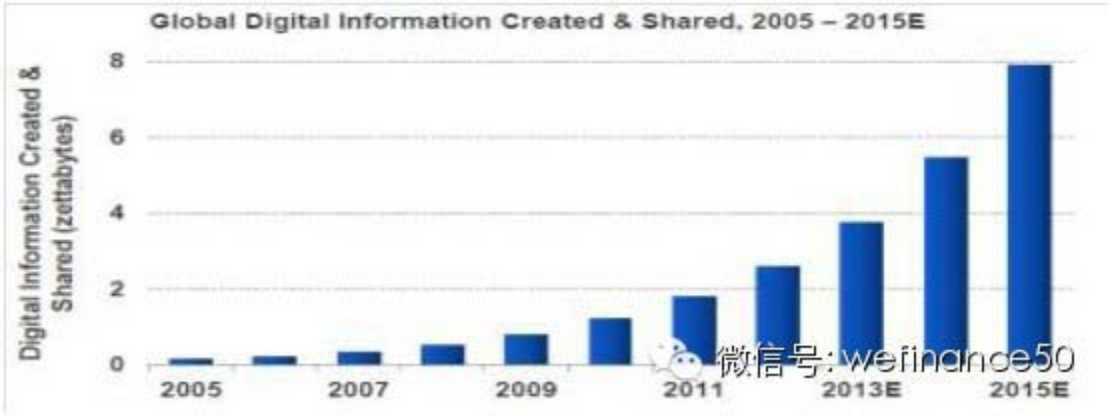

图 2 世界上 90%的数据量产生于近两年

有一本书就叫《大数据时代》。这本书提出了一场生活、工作与思维的大变革。我曾给其推荐语：“我们必须拥抱大数据，我们生活在社会中，就不得不与数据打交道，我们也是数据的一部分，无论我们想不想与大数据牵扯在一起，数据都会找到我们，覆盖我们。大数据时代已经来临，如何从海量的数据中发现知识，寻找隐藏在数据中的模式、趋势和相关性，揭示社会现象与社会发展规律，以及可能的商业应用前景，都需要我们拥有更好的数据洞察力。”

到底如何去挖掘大数据，如何界定大数据，是否有学界、业界统一一致的定义这并不重要，需要特别强调的是大数据对社会产生重要的影响。大数据甚至已经上升到了国家战略，已经被视为一种生产资料、生产要素。美国奥巴马政府率先提出《国家大数据战略》，舆论也将奥巴马称之为大数据总统。不同学科背景的人都在谈论，特别是人文社会学者、商界人士都在谈论大数据，全球已经点燃了迎接大数据时代的热情。

二、大数据的主客体、范畴和特点

美国开放大数据不是简单的开放。它的核心有三点，第一是从大数据中挖掘出价值来支持企业和政府的管理；第二是从大数据中挖掘或者培养更多的能挖掘的人才；第三是要开放大数据。基于此来看《爆发》所说的“人类 93%的行为

是可预知”的观点，这会涉及到个体与大众的关系。何为人类？说的是你还是我，还是其他所有人。过去，大数据对应的所谓小数据，更关注的是一般人和平均值水平。但是到了大数据时代，大数据更关注的是个体。

举个例子，什么是数据挖掘呢？数据挖掘就像挖恐怖分子，只要拉登敢打电话，美国一定会从海量的通话记录中把这个电话抽出来，卫星立刻跟踪，导弹直接就下来了，这就叫精确打击！数据挖掘就是精确打击。这个模式运用到商业领域，就是商业营销领域的精确制导，精确打击。企业完全可以依照抓拉登的方式精确打击每一个消费者，但是企业不会像抓拉登那样付出大的代价！所以当微博出现的时候，我就跟我的学生说，你可要好好写微博，将来你的雇主在雇佣你之前就会通过微博数据精确地了解你的性格、生活方式、消费行为、价值观。每个个体在大数据时代都能被捕捉到。

到底什么是大数据？简言之，不论是数据量还是处理方法都超乎以往的数据可以归入大数据的范畴。维基百科（Wikipedia）提到：大数据就是这样一种数据集，它特指用现有通用软件在可容忍的时间内无法加工、处理和分析的数据就是大数据。今天度量数据存储的大小已经到了 Tb 级和 Pb 级，甚至到了 Eib 级（2 的 60 次方）。但数据量的巨大并不是大数据的唯一特征，在一定条件下，对个人而言是大数据，对企业级应用就是中数据，对移动和淘宝可能就是小数据，对谷歌和百度可能不算数据了。所以有一种说法：大数据就是越来越大的数据。

现在比较流行的一种大数据说法叫 3V 或 4V 理论，强调大数据的数量（Volume）、类型（Variety）、速度（Velocity）、可用性（Veracity）或价值（Value）。当然，大数据的定义，甚至概念界定至今并没有得到学界或业界的统一，不同专业领域，不同学科背景，不同应用场景都有着不同侧重点的阐释。其实大数据这个概念没有流行前，我们就面临着海量数据的处理问题，所以在一定程度上大数据概念落地就是早年的数据挖掘（data mining），是指从海量数据中发现知识的过程，也称为 KDD（Knowledge Discover in Database）。

数据挖掘（Data Mining）就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。进一步狭义的定义就是利用自动或半自动手段，采用统计技术和机器学习方法，从大型数据库中揭示海量数据中有意义的潜在规律和提取人们感兴趣的知识的处理过程。数据挖掘技术经历 20 多年的发展已经基本成熟，有着一套完整的方法论和挖掘软件工具，但是其狭义的定义和解决问题的工具方法并不容易被业界掌握和诠释。

在一定程度上说，大数据概念只是点燃了数据挖掘的社会意义和应用价值，今天的大数据是泛化了的数据挖掘。所以我们更愿意说这是一个大数据时代，但大数据所具有的特征和对社会的影响却是巨大而深远的，特别是在社会科学领域，大数据带来的变革和挑战是颠覆性的，显著特征就是人类社会的数据化生存。社会化媒体使得人们的社会生活，行为态度、交往过程、互动关系都被数据记录并保存下来，这对社会科学研究和预知社会产生革命性影响。大数据带来了社会科学研究的春天。

本质上，大数据研究的是社会人的关系。老师站在讲台，学生坐在下边，老师因为学生的存在才真正成为老师。所以这时候老师和学生都称为社会人。过去说一方水土养育一方人，那是自然人的概念。在经济社会领域，包括金融行业，我们更关注的是哪些是 VIP、哪些客户会流失、哪些是有价值的客户，所以我们建立客户价值模型、流失模型。这是经济人的范畴，人人生来平等，客户生来不平等。大数据更偏爱研究的对象是社会人。大数据是人类社会或者是人的社会行为数据的总和。

三、利用数据流技术分析非结构化数据

以新浪微博的数据为例，我们通过这些数据就可以看到，普通人大概什么时候发的微博，可以精确到日期、分和秒，因此可以计算一天 24 小时平均而言用户都几点发微博，一周 7 天都在哪几天发微博。当然这些数据分析都是非常简单的。我们更关注的是用户说的内容。过去这些内容可能不能分析，但是现在这些内容都变成了数据。当我们拿到数据的时候，可以把数据从结构网络变成结构化的数据，完成对其直接印象的描述。与此相应的，经常有人问，移动、金融行业有大数据吗？没有。为什么？移动如果有大数据，五年前为什么不叫大数据？难道行业形态、业务有根本性的改变吗？移动公司可以分析用户发短信的时间和数量，但是不能分析用户发短信的内容。微博不一样，微博不仅能够知道用户推送消息的时间，还能知道内容。从这个角度来讲，重要的是如何分析这些文本。大数据有一个鲜明的特点，就是可以分析非结构化数据。人类的数据、信息，90%以上都是非结构化数据，不仅仅有文本，甚至还有音频、视频。

分析这些数据，就是要通过技术处理使得数据实现一个流动。先通过自然语言处理技术将一段语料进行分词，再对词汇进行筛选，留下实体词，比如名词、动词、形容词。通过这样的数据预处理之后，将其导出就可以直接对这些数据进行分析了。一个人在微博上经常写名词，说明这个人有专业知识。如果只选出形容词来，就可以做情感分析。如果对

数据进行相应的一种描述，可以做主题模型的分析。通过这种方法就实现了对大量文本进行机器的自动分类和归类。目前，要完全实现这个过程还有很多技术上的难关需要突破。

图 3 非结构化数据分析示例

四、利用大数据实现地理信息匹配、数据可视化

上述之例中，其数据是实时的。大数据可以做实时分析。通过快速提取数据，分析语料，直接完成报表分析。报表能够以可视化的数据形式呈现。大数据时代的数据可视化被称为“大数据时代的梵高”，视觉特效渲染软件可以轻松实现这一功能。

如果将上述非结构化数据分析的文本材料换成一个城市的角色，其结果自然而然地形成了一个地理信息。举例而言，如果提取人民日报官方微博提到的城市地名，就能统计出相应的地理热点。将统计频数大小对应为地图上的亮点颜色深度，其结果就一目了然地展现在地图之上。相应的，可以改变地图的视觉、背景、地域范围，用以做城市、区域、国家，甚至全球的分析。

图 4 大数据时代地理信息匹配

动态地理信息匹配呈现的是一种贸易上的变化趋势，而这个变化趋势是实时的。贸易过程中发生的交易行为、借贷关系，也可以通过热点的形式表达出来。当重新启动程序时，地理信息系统会实时地显示当地的热点。再举一列，Tweetping 网站开发了一个 Twitter 在线可视化的程序。如图 14-5 显示，全球地图的背景中，每一条实时发送的 Twitter 会在对应的地点形成一个亮点。辅以推送消息的语言形式、设备，将其投影到地理上，立刻可以得出结论，何处人是分居的、何处是群居的；说华语的区域是哪，说西班牙语的区域是哪；用安卓的是什么人、用 iPhone 的是什么人。具体而言，投到大伦敦区，富人区用 iPhone，穷人区用的是安卓系统。还有，像欧洲国家的政府开放了报警数据，通过对每一个报警请求进行统计，公众可以知道整个城市任何一个地方发生的刑事案件或者是火灾，能够建立一套指数并由此感知整个大伦敦区各个住房板块可能的价格和社会情况。这些都是大数据时代地理信息匹配的运用。

图 5 全球 Twitter 消息发布实时统计

五、利用大数据反映信息传播路径

这些研究让我们从大的方向上说明大数据不仅能够让我们感知到全体，也能让我们感知到个体。大数据可以分析个人，可以实现精准的个性化推荐，也可以关心整个全局到底是什么。图 14-6 是一条微博在转发到十万的时候被提取出来的传播路径图，这个数据总共有十万条左右的记录。该图表现的是这条微博转发的人群、时间、设备等基本数据元素的集合。这些都是结构化数据。雅安地震的时候红十字会发过一条微博，要大家共同赈灾，共度难关，大家说“滚”。通过把这条微博抓取下来，可以数数有多少个“滚”，什么人在那说“滚”，有没有团伙在说“滚”，最后形成一张可视化的路径图。这时候我们看到，微博中转的过程，是大家“织”微博的过程。我们利用算法去感知这种结果，这个感觉其实就是转发中你起了什么作用。只要用户转了微博，理论上就可以知道在转发过程当中用户的角色、位置和所起的作用。

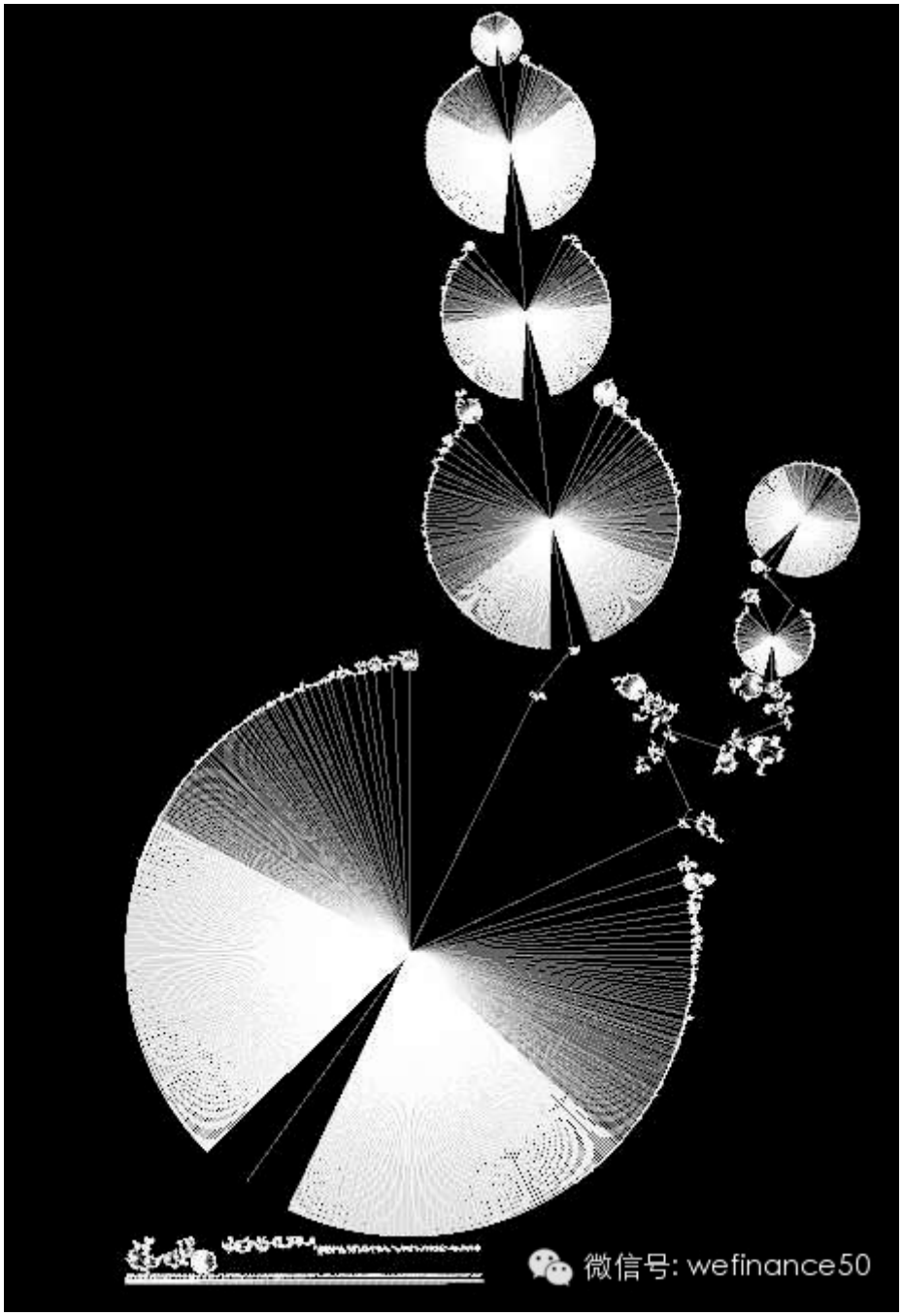

图 6 上帝的指纹——微博传播路径的可视化

由此可以看到在一个信息传播过程中，意见领袖这种微博大号的价值。引申开去，就知道如何阻断谣言，如何有效地营销，如何让用户有所感知。这些都让大众传播能够落地到每一个行为上。大数据挖掘不是挖那些因果逻辑，而是要挖那些事先不知道的甚至是违背直觉的联系。它是没有理论，没有假设的，仅仅是从海量的数据中寻找在这个数据集中存在的模式、趋势和相关性。

如果我们知道一个人的社会网络、社会关系，知道了这个人的价值，如果还能够把他对应落回到地理上，又可以产生新的价值。在地图上，随意标注二十个人，就可以实时测算这二十个人是否落在某一个指定区域之中。如果这个算法能够匹配出来，意味着在北京市锁定十万人的目标群体，以及 100 个银行的营业点，就能实时知道距离某个银行网点最近的那些人。对用户而言，他可以看到距其位置最近的营业厅。他也可以任意指定区域进行匹配。对商业数据使用者而言，可以在数据图层上标识或者是圈出这些人来，可以感知到他的 LBS，给予其个性化推荐。把一个社会人的行为、经济人的价值回归到一方水土养一方人的自然属性上，肯定能对消费者、对智慧城市管理产生更多更好的体验价值。这种

体验数据的处理全部在云端。当然云计算是必备的一种方式，每一个 APP 都是一个云计算。总而言之，大数据时代，数据价值与商业创新将彻底改变我们的工作、生活和思维方式。