【大数据与金融创新】Intel 吴甘沙：

大数据在金融领域的创新和挑战

（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）

一、金融、互联网金融和大数据金融

所谓金融，就是跨越时空的一种价值交换。具体而言，它主要是三个问题。第一个问题：谁是交换的主体？古代养儿防老，就是一种跨越时空的价值交换，主体就是老人和孩子。第二个问题是交换什么价值？现在花钱养孩子，未来孩子会养老人，这是一种交换价值。第三个问题是怎么进行交换？可能有不同的交换手段。

现在一个非常热的词就是互联网金融。互联网金融针对这三个问题有不同的解答。交换主体方面，互联网金融非常看重长尾人群。传统金融服务高富帅、白富美，而互联网可以覆盖长尾人群。第二个问题，交换价值方面，传统金融往往是利用信息不对称获得价值。我们把钱给银行，但是不知道钱的用途。银行把钱贷给可以支付较高利息的借款人，赚取息差。而互联网金融最小化信息不对称。通过 P2P 借贷，出借人可以了解借款人的情况，决定是否可以借出资金。第三个问题，交换方式方面，互联网金融拓展了原有的交换方式，降低了价值交换的成本，比如余额宝 7×24 小时可以交易，T+0 赎回。因而互联网金融比传统金融较好地解决了这三个问题。

现在大家又在讨论大数据金融。大数据金融是互联网金融的一个必要条件。还是回到那三个问题。在交换主体上，针对缺乏信用基础数据的长尾人群需要多源数据获取。第二个问题，交换什么价值？怎么保证最小化信息不对称呢？需要建立征信体系，一个社会的信用基础设施能够使得这个社会运行成本极大地降低。接着是对信用的实时跟踪、做风险的预测性分析。比如 P2P 网贷，出借人了解借款人的基本情况是在贷出资金之前。贷后借款人的经济状况是否发生变化，需要实时和预测性分析。第三个问题，最小化交易成本也需要实时分析。比如余额宝，通过实时分析，可以监测到账户里的钱被偷出的时刻，能够处理账户被盗取的情况。要使交易更为有效，还需要处方性分析，预测性分析是预测未来要发生的情况，而处方性分析是通过改变现在的行为参数来主观设定未来发生的情况，如果我想要这样的未来发生，需要做一些什么事情。本质上，大数据可以为金融提供这些技术支持。

二、数据获取

数据获取的一个最直接的想法就是利用自有的数据，比如国泰君安做的 3I（Individual Investor Index）指数，能够分析其数十万投资者的样本，了解他们资金账户的活跃程度、资金进出、盈亏比例，了解整个市场投资的景气指数。

第二类方法是自采数据，尤其是针对个人的数据。国外的保险公司 Progressive Insurance 为了获取数据，在其用户的车里面装一个蓝色小盒子，盒子能够去采集开车的行为数据，加速、减速、踩刹车、变道，根据这个行为数据可以知道客户的开车习惯，进而决定保险费率。这是自采的例子。可穿戴智能硬件和量化自我运动的大行其道后面也是自采数据的逻辑。国外的大公司对数据的饥饿感以及不惜成本是让人叹为观止的，比如彭博社专门雇佣一个卫星每周对位于俄克拉何马的美国最大原油储备库拍照，根据油罐浮动顶的阴影长度来判断原油储备量的变化。

第三类方法是利用公开的数据。比如社交网络为金融提供了大量的数据用以分析投资者情绪。格林斯潘说过，“如果我知道这些大众投资者是恐惧还是兴奋……那么在没有其他信息的情况下，我将能比任何人都能预测经济。可惜的是，我无法看清公众的情绪”。他说这句话的时候还没有大数据的技术。美国有几个教授就已经开始利用社交网络做情感分析，将其与标普指数匹配，发现了两者高度吻合。在这里，利用公开的数据也是一种获取数据的手段。

第四类方法是价值链的分享。在一个核心公司的价值链上下游围绕着一批公司。Mastercard 是一家发行信用卡的公司。它的上下游有一些发卡行和零售企业，还有其他的服务性企业。Mastercard 可以在这些企业当中做数据共享，根据数据共享它可以提供很多智能服务。比如用户在加油的地方刷过一次信用卡之后，后台马上知道用户需要用餐了，并及时做相应的推荐。

第五类方法是跨业共享。Kabbage 是美国一家类似阿里的小微贷款公司。它不像阿里有那么多的数据，但它通过几方面手段进行跨业共享。它从 Google Analytics 获取小企业的网络行为信息，它从 UPS 获取企业物流信息，它还通过美国的云财务管理软件那里知道小企业的运营状况。通过跨业共享能够把一个企业或者个人的画像全面地展现出来。

第六类方法是数据汇聚商。比如 Acxiom、LexisNexis 这些公司，他们向政府、法院、服务类企业买数据，买了原始数据之后会想方设法将其数字化，比如运到菲律宾让人工来做。当把这些数据汇聚起来后，每个人的记录有几万字段，涵盖了生活的方方面面。这些公司通过汇聚数据来提供服务，比如 LexisNexis 为美国保险行业提供个人评分。

第七类方法是通过中间商以少换多。美国有些银行成立一个联盟叫 Cardlytics，收集消费者数据并帮助银行将数据变现。比如可以根据个人消费记录向其推荐一些东西，如果推荐的商品确实产生效益，这些数据的原始提供者，即银行，能够得到分成。

第八类方法是用户授权。聚信立是上海的一家公司。它要帮助信用卡发卡行和贷款机构对个人的信用状况提供评估，但缺乏数据，于是它采用了用户授权的方法。它提供一个云浏览器，让用户自己上到云浏览器里面登陆淘宝、京东、支付宝、移动通信公司等等，在用户的名义下获得大量数据再生成信用报告。这个例子是以用户之名把分散在不同地方的数据汇集在一起。

无论是哪种方法，核心都是数据的开放，表现为无条件的开放、有条件的共享或交易。

对于无条件的数据开放，在保证数据质量的同时（如 Tim Berners Lee 五星标准），还要注意一些问题。

首要问题就是数据权属的明晰问题，数据是属于创建者还是属于采集者，亦或属于被观察的客体？这个权属该怎么来界定？当数据拥有者或者数据涉及主体发生变化的时候，数据的权力相应的该怎么发生变化，自然人离世后谁来继承数据权，离婚之后数据权能不能分割？另外一个问题是敏感数据的界定。不同国家对于敏感数据有不同的定义，比如在欧洲或者在美国，个人位置信息、IP 地址属于敏感数据，但在日本就未必。这需要法律作明确的界定。

其次，要从技术层面做数据的脱敏，第一步是去标识化，去掉跟人名、地址等相关的内容。去标识化不一定能够做得彻底。美国方面做了研究，只要有性别，出生年月以及邮编这三个数据，就有 60-80%的把握能够把个人的信息还原出来。

去标识化要防止重新标识化(re-identification)，比如通过多数据源来重新进行标识。美国在线曾经开放了匿名的搜索记录，但是有人把这个信息跟美国的选举人登记信息一匹配，结果就把人找出来了。Netflix 也一样，它开放了匿名的评论以及打分的信息，但是有人把它跟国际电影数据库 IMDB 匹配，结果把一个有同性恋倾向的人识别出来了。另外一种重新标识的可能性是基于统计。麻省理工学院的研究人员发现，如果获悉一天之中个人在四个不同的时间点所处的不同地点，有 95%的可能把这个人找出来。在 Netflix，如果知道某人区区几个评分信息，就能够有较大的概率将其标识出来。

防止隐私攻击的匿名化技术，比较典型的有 k-anonymity、L-diversity、T-Closeness 等等，但还是有隐私攻击的可能性，特别在敏感属性不够多样化，或攻击者具有背景知识时。更好的一种匿名化技术叫差分隐私(differential privacy)，把噪声加入到数据集中，但仍保持它的一些统计属性。英特尔支持普林斯顿和罗切斯特大学做了这样的研究，现在试图在运营商开放数据中应用。当然，差分隐私需要在隐私安全性和数据可用性之间做好平衡，当噪声太到一定程度，数据可用性会变差。从实践意义上来说，隐私保护的成本一定要低于数据本身的价值。

广义的数据开放不只是包括无条件的数据开放，更普遍的是数据的共享及交易，比如点对点进行数据共享或在多边平台上做数据交易。大数据时代马克思的生产资料所有制已经不再重要，现在更好的一种机制是生产资料的租赁制。在数据时代的场景下，我不一定拥有数据，甚至不用整个数据集，但可以租赁。租赁的过程中要保证数据的权利，数据可以使用，但不可以看见。这也是我们一直在说的“可用不可见，相交不相识”。其实这不是个新问题，姚期智老先生在 1982 年提了个“millionaires’ dilemma”问题，两个百万富翁比富，但谁都不愿意说出自己有多少钱。这就是典型的“可用但不可见”场景。这个场景有很多的实际应用，比如美国国土安全部有恐怖分子的名单，而航空公司有飞行记录。国土安全部要去航空公司要飞行记录，航空公司不愿意给，因为涉及隐私承诺。而国土安全部也不可能给恐怖分子名单，因为这是最高机密。在双方都不能看到对方数据的前提下，怎么能够把两份数据放在一起产生价值？这是现在技术上试图解决的很重要的问题。

从点对点的共享，最后要走到多边的数据交易，从一对多的数据服务到多对多的数据市场，再到数据交易所。如果说现在的数据市场更多的是对数据集进行买卖的话，数据交易所则是一个基于市场进行价值发现和定价的，像股票交易所那样的，小批量、高频率的数据交易集合场所。未来这是一个很大的机会。

怎么做到可用但不可见呢？一种手段是通过同态加密数据库技术。如下图所示，在数据拥有方，甲方的数据库是完全加密的。这防止了现在常常出现的内部人员导致数据泄露问题。加密数据库可以运行乙方的普通 SQL 程序。因为它采用了同态加密技术和洋葱加密法，SQL 的主要语义在密文上也可以执行。

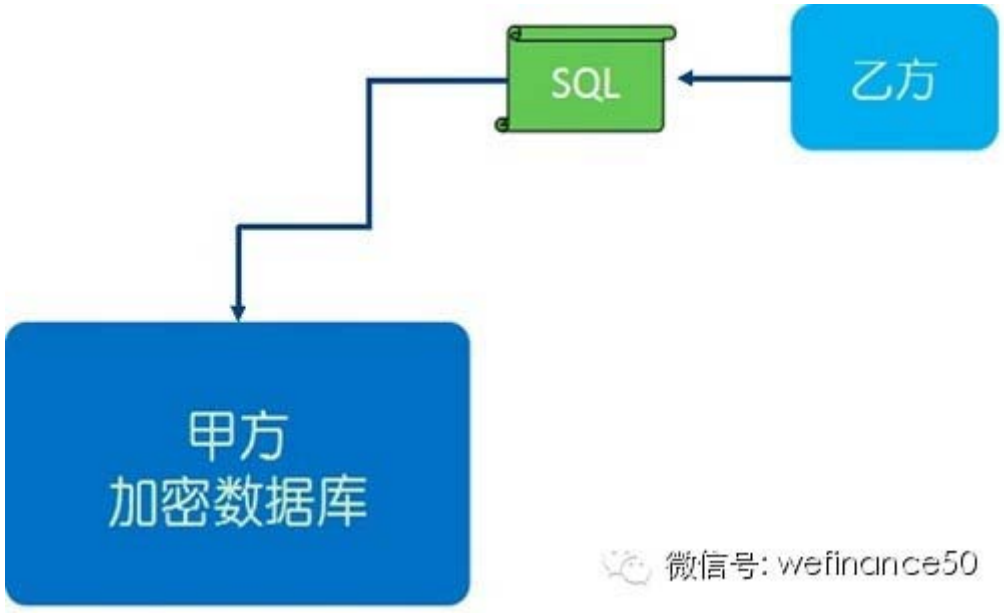

图 1 同态加密数据库技术

我们还探索了另一种实现“可用但不可见”的技术，我们把它叫做数据咖啡馆。“咖啡馆”这个名字来自 Steven Johnson 的 TED 演讲，他指出咖啡馆是开放的、安全的空间，让不同的人凑在一起进行思想的碰撞，产生新的想法。数据咖啡馆让不同的数据碰到一起，产生新的价值。如下图所示，两家电商，一家卖衣服鞋帽，一家卖化妆品，他们对客户的画像都是非常片面的，如果把两者的数据在咖啡馆里相逢，就能够对客户有一个全面的认识。

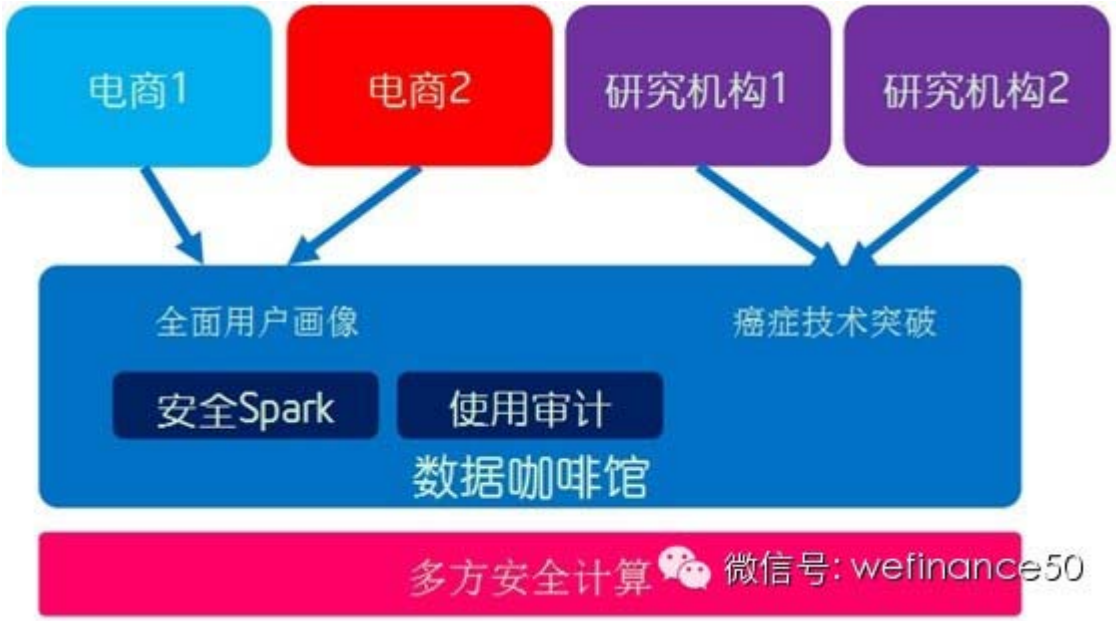

图 2 数据咖啡馆技术

另一个案例：大家知道癌症是一个典型的长尾病症。过去 50 年癌症的治愈率只提升了 8%。为什么呢？因为科研机构拥有的基因样本非常有限。如果能够通过数据咖啡馆，让不同的科研机构把这些样本放在一起，在癌症治疗技术上面就能够有所突破。

数据咖啡馆是一个安全的大数据计算环境，能够保证人（甚至是咖啡馆服务的运用商）看不见数据，只有分析程序看得见数据。为了保证分析程序不做坏事，不暗地里偷数据，要对其数据使用进行审计。

三、数据分析

金融数据分析强调实时性和预测性。原来供应链金融在贷款过程中可能需要提供具体的抵押物品，现在则完全凭借信用。一个企业信用状况是随时变化的，贷前根据过去一些交易记录可以知道企业的信用状况，贷后则需要实时监测和预测信用状况的变化。

金融数据分析需要有可解释性。很多大数据分析结果是要给人看的，要强调可解释性。FICO 是美国信用评分公司，它的数据非常有用。美国一些 P2P 借贷公司根据 FICO 积分设定贷款利率，比如 400-500 分对应 6%的年利率，500-600 分是 5%的年利率。FICO 评分所依赖的参数是公开的。这些参数的变化会产生积分的变化，其结果是可解释的。

还有一类分析是面向机器的。这个更强调分析的精确性而不是可解释性。比如说另一家征信公司 Zest Finance，它从一千个变量当中转换出七万个变量，跑出十个不同的算法对数据进行分析。这十个算法进行一次投票，最后获得个人信用状况。通过这样的组合方式，Zest Finance 给出的结果可解释性非常差，但它的精确性非常好。现在这种算法组合提升精确性的方法学非常重要，在 IBM Watson、Netflix 竞赛中广泛使用。

数据分析要有鲁棒性。谷歌有一个流感指数，最近大家发现它很不准。这个指数跟 CDC 美国疾控中心实际的数据很不吻合。这导致什么问题呢？国家卫生机构会根据谷歌的流感指数预测未来的疫苗需求，并购买疫苗。预测不准则会导致购买的疫苗数量多余或者不足。这个例子说明分析的鲁棒性非常重要。大家都听说过深度学习。普遍认为深度学习是未来解决机器智能的一个主要手段。现在深度学习还存在问题。比如，谷歌发现，一张图片能让深度学习识别出猫来，然后在这张图片上略作一些改动，算法就识别不出来了，说明鲁棒性还需要改进。

传统的机器学习算法一般是基于一种指数分布的假设。指数分布把尾巴都割掉了，因而传统的机器学习算法不在乎长尾，可以针对高频词汇，针对主流信号进行处理。大数据时代要倾听每一个个体的声音，每一个信号都是有用的信号。所以机器学习的算法对于长尾需要有更好的支持。

还有就是数据分析要社会化。马云有句话说，在信息时代我要比别人聪明，我要从数据提取出信息卖给别人。数据时代，别人可以比我聪明，我只要采集到数据，可以雇佣更聪明的人来替我分析。Dunnhummby 就是一家分析公司，专门帮乐购做分析，他们形成一个非常紧密的合作。Palantir 是美国一家给美国政府和金融机构做分析的公司。他们之间的合作相对松散一些。Kaggle 是社会化分析平台，一边是十万个分析师，一边是企业的分析需求。通过这个平台，分析师可以帮企业去做分析。

数据分析算法之间会形成一种战争。华尔街请了很多聪明人，尤其是很多物理学家，仅东欧的物理学家就有几千个。这些人原来可能是分析如何从雷达信号发现美国隐型飞机的，现在跑到华尔街在浩瀚的交易中发现隐藏的竞争对手及其算法，就造成了算法之间的竞争，传统的高频交易实际上并不是大数据，现在的策略交易更需要大数据的分析。

四、数据社会的基础是数据经济

互联网时代一个很重要的规律就是梅特卡夫定律。梅特卡夫是以太网的发明者。这个定律是互联网经济的基石。梅特卡夫定律认为一个网络的价值与其节点数的平方成正比。只有一台电话机没有价值，两台电话机之间形成一个连接，三台电话机有三个连接。其连接数是 N×(N-1)的关系，其价值也就跟网络里面节点数的平方成正比。在 90 年代，互联网公司没有盈利模式、岌岌可危的时候，另外一个人出现了，她叫玛丽·米克尔，这位现在被称为互联网女皇的股票分析师根据梅特卡夫定律，推导出互联网公司估值的规律。互联网公司估值跟用户数平方成正比，大量互联网公司获得资本青睐，并不是追求其市盈率，而是追求其用户规模。

大数据时代也需要一些经济规律，比如数据定价。数据和信息的一个重要区别就是，信息是具有特定意义的数据，因为特定意义存在，所以信息价格很好估计。

但数据还没使用的时候它的价值是不确定的，就像赌玉石，不把它剖开来谁也不知道它的价值。更重要的是，汽油的价值在其燃烧的一瞬间消失，但数据跟传统物质资源不同，它是可以反复使用的。而且数据具有外部性，它应用于不同领域时可能产生超出预期的价值。因此，数据的定价原则是：不为买断式的产权变更定价，而是为一次使用、价值变现定价，先使用后定价。数据市场上，要改变现在买卖数据集产权的方式，而是让租用和买卖一站式发生。

同样，企业的数据资产，也可以根据使用来估值。某些数据被用得多，它也就估值更高。

个人数据也可以定价。个人数据通过中间商出售给广告商可以变现，这个过程完成了个人数据的定价。原来这些数据已经被 Facebook、谷歌拿去赚钱了，个人没有获得任何的收获。现在把数据交给中间商至少一个月可得 8 美金。

数据定价的另一个原则是在共享和交易中要防止劣质数据。两边数据碰在一起，一边放了一些劣质数据，一定会影响最终的结果，所以要防止劣质数据。斯诺登的文件曾经揭露，英国情报局会在互联网插入大量虚假数据，左右舆情。这就属于劣质数据的例子。

数据经济需要跨界思维。举例而言，金融数据和电商的数据放在一起能够产生很多的价值，比如小微贷款。同时，电商交易活动数据和现金流数据能够反映各个细分领域的真实状况。所以说马云是最早预测 2008 年金融危机的人，因为他有 B2B 的数据。金融数据和医疗数据放在一起也能够产生很多价值。比如现在国内骗保问题很严重，用户并没有发生医疗服务行为却在大量刷社保卡。通过两边数据的结合能够解决这一问题。还有金融数据和物流数据放在一起就产生供应链金融。诸如此类，都说明了大数据时代不同行业的数据结合将产生乘法效应，实现价值放大。另外也能产生外部效应，将某一个行业的数据用于其他行业。比如智能电表的数据能够了解房地产的景气状况。总而言之，大数据符合德尔菲气象定律的特征——只需在气象上面投入 1 块钱可以产生 98 块钱的社会价值。

总之，对于金融行业来说，即使它本身并不是那么开放，在大数据时代需要拥抱开放才能破局，才能实现量子跳跃。具体而言，要做开放的数据共享和交易，开放的社会化分析，开放的数据定价，以及开放的跨界思维。