Reflections on Big Data

1、 Introduction

大数据的概念，最初起源于美国，是由 EMC、IBM、Oracle 等多家跨国 IT 巨头倡议发展起来的。大约从 2009 年始，“大数据”成为互联网和信息技术行业的流行词汇，进而发展到商业、经济及其他领域。

对于“大数据”一词，国内外曾给出多种不同的定义。2012 年，研究机构 Gartner Group[1] 对其做出诠释，“一种需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力来适应海量、高增长率和多样化的信息资产。”而麦肯锡[2]给出的定义则是，“一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征。”

虽然它们的讲述方式各异，但万变不离其宗的，是对数据规模的海量和复杂性的强调。简单来说，**大数据是指传统数据库管理工具或数据处理应用软件不足以处理的庞大而复杂的数据集**。在我们日常生活中，最常用的存储数据单位是 KB、MB 或 GB，一个电脑硬盘最多也只有几个 TB，而大数据的数据量一般用 EB（10^20 TB）甚至更大的基本单位来表示。面对如此庞大的数据量，我们必然无法用单台计算机进行处理，而必须使用“在数十、数百甚至数千台服务器上同时平行运行的软件”（电脑集群是其中一种常用方式）[3]，或者采用分布式架构，依托云计算的分布式处理、分布式数据库和云存储技术等等。

大数据的重要意义，不在于数字本身，而在于那些海量数字所承载的海量信息。随着人类社会的加速发展，每个人产生的数据量呈指数型增长。据华为董事长梁华[4]采访说，1992 年时，全球人类每天总计只能产生 100GB 数据；而 2021 年时，平均每人每天产生的数据高达 1.5GB，总计 114 亿 GB 数据。数据于每时每刻产生，且存在于社会的每处角落，如社交媒体上用户的行为数据，企业的销售数据，政府的统计数据等等。这些数据如果能够被合理利用和分析，将会为产品推荐、商业决策、风险预测、国家战略等带来巨大的价值。如果我们能够对这些含有意义的数据进行专业化处理，提高对数据的加工能力，也可实现数据的“增值”。

2、 Body

（1）The four fundamental characteristics of big data（4V）

对于大数据的四个基本特征，我们也可以用 4V 来描述:

1. Volume，即体量大。比如以 PB、EB 级别的数据量，这是远超传统数据库处理能力上限的。

2. Velocity，即速度快。数据以前所未有的速度产生和处理，而这就需要实时或者接近实时的处理能力。

---

[1] Douglas, Laney. The Importance of 'Big Data': A Definition. Gartner. [21 June 2012].

[2] Anonymous. (2016, December 12). 科普篇：什么是大数据. 51CTO. https://www.51cto.com/article/527722.html

[3] Jacobs, A. The Pathologies of Big Data. ACM Queue. 6 July 2009 [2015-12-07].

[4] 梁华. 2021 China Mobile Global Partners Conference, 2 Nov. 2021.

3. Variety，即种类多。数据来源广泛，类型多样，包括但不限于结构化数据、非结构化数据（如图片、音频、视频、日志、位置信息等）和半结构化数据。相比传统的单一种类数据，处理过程更为复杂。[5]

4. Veracity，即真实性。随着新数据源的兴起，在大量数据中保证数据的准确性和可靠性至关重要，这是为了能更好地借助大数据的力量，确保分析结果有效，保证决策的安全。[6]

（2）The application of big data

在多数情况下，大数据的具体应用，是通过数据挖掘，以机器学习模型的方式来实现的。目前大数据在各项领域均有广泛的应用。

例如在医疗健康领域，我们可以通过分析海量的医疗数据，利用大数据辅助医生进行疾病诊断。例如，在 Kaggle 网站上就有诸多此类的竞赛，其中有许多医疗图像的数据集，比如某个人体部位的 X 光照片，而参赛者则需要训练出能准确识别病症的机器学习模型。这有助于提高医疗服务质量，减少误诊风险。我们还可以通过共享电子病历收集并分析数据，探索降低医疗成本的策略。现在去医院看病的一个极大的痛点是，不同医院间的数据几乎是不共享的。我去别的医院拍片子，只能带纸质的片子去另一个医院就诊，而电子版只储存在前者的系统里。数据共享能切实地减轻患者重复检查的负担，提升病历的规范化与结构化。[7]我们甚至可以利用大数据技术实时监控并分析相关健康信息，可以预测群体的慢性病风险，例如早期癌症的检测，如果能通过生活中呈现的蛛丝马迹，警示病人、帮助病人尽早地发现患病，而不是等到癌症晚期出现明显不适再就医，那这就是医疗卫生服务极大的进步。虽然这肯定会牵涉到隐私问题，而我在后面会提到。所以，在医疗领域，大数据应用于疾病预测、治疗效果分析和医疗资源优化，能提高医疗效率和患者满意度；然而，这也引发了数据隐私和安全性的担忧。

而在金融领域，大数据最著名的应用就是量化交易。量化交易利用大数据技术分析市场数据，包括价格变动、交易量、新闻报道等，通过算法模型预测市场走势并自动执行交易。这种方法能够处理庞大的数据集，快速做出交易决策，提高交易效率。然而，量化交易也可能放大市场波动，引发系统性风险，并且对市场的预测并非总是准确的，需要持续优化算法以适应市场变化。除此之外，大数据可用于分析客户的信用记录、收入和支出等信息，去评估客户的信贷风险，从而帮助银行等金融机构做出更好的决策，但这可能会造成歧视的问题；或是用于分析客户的消费行为和偏好，可以实现理财产品的精准营销，提高营销效果，而这涉及到个人信息获取权限；还有就是通过大数据分析，可以识别出交易欺诈行为，帮助金融机构和个人减少损失。[8]综上，在金融领域，大数据用于量化交易、风险管理、客户行为分析和欺诈检测，增强了金融机构的服务能力和市场竞争力，但也可能导致市场波动、数据歧视和个人信息泄露问题。

---

[5] Douglas, Laney. 3D Data Management: Controlling Data Volume, Velocity and Variety (PDF). Gartner. [2001-02-06].

[6] What is Big Data?. Villanova University. [2015-12-08].

[7] 杨彦帆. (2023, December 29). 信息共享加速 患者诊疗减负（暖闻热评）. 人民日报

[8] 大数据在金融领域的典型应用研究(PDF). 信通院. [March 2018]

（3）The challenges posed by big data

大数据是一把双刃剑——它带来无数的机遇，但与此同时公民或机构的隐私权也受到了极大的冲击。当我们在使用各种手机 app 时，需要授权多种个人信息如相册照片、地理位置、麦克风等，甚至不给予授权就完全无法使用该 app，这给我们日常的隐私保护带来了极大的困扰。例如抖音、小红书等，其内容推荐算法已经过于强大而达到了恐怖的境地，有时在日常对话中，或是微信聊天里无意间提及的事物竟都反映在了其推荐页面上，让人不得不怀疑这样的 app 是否有监控和监听的举动。相同的问题还有诸如淘宝和京东的商品推荐。一些厂商对个人数据无所不用其极，隐私滥用、贩卖个人数据的现象时常发生，有些人甚至公然宣称，“中国人没有隐私”，这无不反映了大数据时代的症痛。

大数据包含各种个人信息数据，现有的隐私保护法律或政策貌似无力解决这些新出现的问题。有人提出，大数据时代，个人是否拥有“被遗忘权”，即是否有权利要求数据商不保留自己的某些信息，大数据时代信息为某些互联网巨头所控制，但是数据商收集任何数据未必都获得用户的许可，其对数据的控制权不具有合法性。2014 年欧盟法院就“被遗忘权”一案对谷歌作出裁定，判决谷歌应根据用户请求删除不完整的、无关紧要的、不相关的数据以保证数据不出现在搜索结果中。[9]这说明在大数据时代，加强对用户个人权利的尊重才是时势所趋的潮流。当然，监控也有相对较好的应用场景，例如前面提到的癌症的早期检测和预防，对于这一方面，需要立法完善的同时，提升公民意识，才能真正地将大数据用于正途。

机构在大数据时代的浪潮中也不能独善其身。大面积的数据安全事件时有发生，前几年，超星学习通就曾曝出几千万学生的个人账户信息全部遭遇泄露，共计 1.7 亿条学生数据库信息被公开售卖。[10]可以说，在未来，每个企业都会面临数据攻击；所有企业，无论规模大小，都需要重新审视今天的安全定义。在世界 500 强企业中，超过 50%将会设置首席信息安全官这一职位。[11]企业需要从新的角度来确保自身以及客户数据，所有数据在创建之初便需要获得安全保障，而并非在数据保存的最后一个环节。

最后，我想从数据本身的局限性来说。首先我想的是，一个活生生的，有血有肉，富有情感的人，是绝不可能被几个数字所定义的，是万不可被数据完整概括的。我们关注数据的价值的同时，不能忽视了其对人的异化。其次，对于数据的分析研究而言，我们不可尽信数据。在任何情况下，仅仅依靠数据不能反映全貌的；甚至真实的数据有时候也会误导人，欺骗人。对于任何一件事情，会有起因经过结果，他可能是由多种因素造成，其经过和结果也会有很多种理解诠释的方式。事情的每一个环节，我们都无法用一个个数字去准确地还原。有些时候数据与结果确实有相关性，但是有些时候看似的关联仅仅只是巧合而已，却被人为赋予了特别的价值。在这里，确认偏误（confirmation bias）是很难避免的，即有时候我们会不由自主地预设立场和观点（通常是先前形成的既有成见，或是一套政治正确的话术），

---

[9] Streitfeld, D., Kanter, J., & Scott, M. (2014, May 14). 欧洲法院: 谷歌应根据用户要求删除搜索结果. The New York Times. https://cn.nytimes.com/technology/20140514/c14google/

[10] 央视网. (2022, June 21). 学习通用户数据泄露？公司声明：公安机关介入调查. https://news.cctv.com/2022/06/21/ARTIc7r5k53PUR9YgKKKuE0Z220621.shtml

[11] 腾讯云. (2018, April 20). 【聚焦】"数据探索年" 2015 年是大数据发展八大趋势. https://cloud.tencent.com/developer/article/1105273

然后为了论证，而去寻找“合适的”数据。所以说，如何对数据“去伪”，是一件重要的事情。“伪”可指数据本身，亦指利用数据的过程。而这其中，批判性的实证、讨论和分析是关键。作为数据研究者，常会纠结于细微的数据之中，因为追求细枝末节的精准而忽视了一些本质的东西。这些本质因素，通常是决定存亡的关键，必须经过实地调研考察，才可能对其有所领会。

3、 Conclusion

综上所述，大数据在医疗和金融等众多领域展现了巨大潜力，用于优化决策过程、提高效率并引领创新。然而，其发展也伴随着侵犯隐私、数据泄露等安全和伦理方面的挑战。面对这些问题，我们唯有加强法律法规建设，进一步提升公众意识，并且构建更为安全的数据处理技术。此外，我们应保持对数据分析的客观批判性，确保科技进步同时，人的价值和个性不被忽视。在未来，大数据将继续在技术创新与社会责任间寻求平衡，为人类社会带来更深远的影响。

4、 References

1、Douglas, Laney. The Importance of 'Big Data': A Definition. Gartner. [21 June 2012].

2、Anonymous. (2016, December 12). 科普篇：什么是大数据. 51CTO. https://www.51cto.com/article/527722.html

3、Jacobs, A. The Pathologies of Big Data. ACM Queue. 6 July 2009 [2015-12-07].

4、梁华. 2021 China Mobile Global Partners Conference, 2 Nov. 2021.

5、Douglas, Laney. 3D Data Management: Controlling Data Volume, Velocity and Variety (PDF). Gartner. [2001-02-06].

6、What is Big Data?. Villanova University. [2015-12-08].

7、杨彦帆. (2023, December 29). 信息共享加速 患者诊疗减负（暖闻热评）. 人民日报

8、大数据在金融领域的典型应用研究(PDF). 信通院. [March 2018]

9、Streitfeld, D., Kanter, J., & Scott, M. (2014, May 14). 欧洲法院: 谷歌应根据用户要求删除搜索结果. The New York Times. https://cn.nytimes.com/technology/20140514/c14google/

10、央视网. (2022, June 21). 学习通用户数据泄露？公司声明：公安机关介入调查. https://news.cctv.com/2022/06/21/ARTIc7r5k53PUR9YgKKKuE0Z220621.shtml

11、腾讯云. (2018, April 20). 【聚焦】"数据探索年" 2015 年是大数据发展八大趋势. https://cloud.tencent.com/developer/article/1105273