

专家洞察

IBM 商业价值
研究院

应对 AI 数据困境

恰当的数据集成方法、治理和工具



主题专家



Michael Haydock

IBM 院士, IBM 全球企业咨询服务部
副总裁兼首席科学家
mhaydock@us.ibm.com
linkedin.com/in/drmikehaydock

Michael 是知名的人工智能专家, 在 IBM 全球企业咨询服务部 (GBS) 从事深度学习研究工作。他专攻数值优化和定量机器学习领域, 应用这些呈指数级发展的技术理解客户关系以及优化企业供应链。Michael 拥有明尼苏达州明尼阿波利斯瓦尔登大学运筹学博士学位。



Steven Eliuk

IBM 全球首席数据办公室 (GCDO)
人工智能与治理自动化副总裁
Steven.Eliuk@ibm.com
linkedin.com/in/m3the01

Steven 在 IBM 的认知型企业数据平台 (数据湖) 中领导 AI 平台和基础架构组件的开发。他还负责将企业中的 AI 应用于内部流程, 保持可靠的治理、安全、隐私和信任水平。他的团队负责开发、维护和运营 IBM 的一系列全球隐私服务, 包括满足 GDPR、CCPA 和其他法规的要求。在加入 IBM 之前, Steven 在三星美国研究院领导设计人工智能高性能计算 (HPC) 基础架构。Steven 拥有阿尔伯塔大学计算机专业博士学位。



Susara van den Heever

云与认知软件专家实验室主任
svdheever@fr.ibm.com
linkedin.com/in/susaravandenheever

Susara 担任欧洲、中东和非洲专家实验室主任一职, 工作重心是使用 IBM Cloud 和认知软件为客户创造价值。她的理想是借助数据和 AI 让世界变得更加美好。她帮助个人和企业提高效率、节省时间、更好地管理资源 — 甚至做出更好的环保选择。她曾供职于许多行业: 帮助啤酒生产商创建更有效的供应链, 确保为电信提供商公平分配带宽, 帮助能源企业建立可持续生产, 以及让城市更加宜居, 等等。Susara 拥有卡内基梅隆大学化学工程专业博士学位。

扫码关注 IBM 商业价值研究院



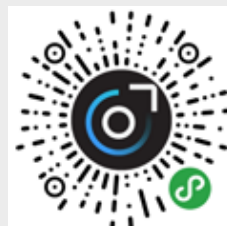
官网



微博



微信公众号



微信小程序

为什么大约一半的 AI 项目以失败告终？一言蔽之，数据使然。

要点

要从 AI 中获得最大价值, 就一定要从业务问题入手。然后寻找多种数据类型 — 结构化和非结构化数据、内部和外部数据、定性和定量数据, 以解决问题并丰富解决方案。

融入基于权限的强有力治理, 建立数据溯源能力, 以形成对数据和 AI 洞察的信任。

制定计划, 应对严格的数据准备挑战以及合并不同数据源带来的复杂性。复用数据, 自动执行流程并采用适当的工具。

—

AI 的独特数据挑战

人工智能 (AI) 不再是新生事物 — 量子计算才是。AI 正广泛应用于各种商业和社会用途。在疫情初期, 84% 的组织预计将保持或提高对 AI 的关注度, 近三分之一的组织因疫情直接增加 AI 投资。¹

最近的一项调研表明, AI、物联网 (IoT) 和云计算是受访 CEO 们认为最有助于实现成果的 3 项技术。² 43% 的 IT 专业人员表示, 他们的企业受疫情影响而加快部署 AI 解决方案。³

这转化为企业对建立战略性 AI 能力 (包括战略、运营模式、人才) 以及将该能力整合到企业的迫切需求。

但即使在完成有前景的概念验证 (PoC) 阶段之后, 许多 AI 项目仍处于停滞状态。90% 的企业难以在整个企业中扩大 AI 的应用范围。⁴ 大约一半的 AI 项目归于失败也就不足为奇了。⁵

为什么? 一言蔽之, 数据使然。超过半数的 AI 战略负责人承认, 他们不清楚自己的 AI 数据需求。⁶ 39% 的 IT 专业人员表示, 分析数据以建立和扩展可信的 AI 是组织 AI 之旅中最困难的部分, 32% 表示数据复杂性和数据孤岛是采用 AI 的最大障碍。⁷

难怪超过一半的组织将数据视为 AI 项目停滞的罪魁祸首,他们将数据质量问题视为最主要的因素 (58%), 其次是缺乏整理有序的数据 (45%), 接下来是数据治理问题 (40%)。⁸ 事实上,即便是高级数据能力的常见最佳实践对 AI 而言也是远远不够的。虽然 AI 可能只是数据的一个用例 — 尽管是重要用例,但需要将汲取的经验教训广泛应用于 AI 领域。

AI 有几个特殊的数据考虑因素,会对实践起到重要作用:

- **AI 规模:** 一般来说,对于 AI 而言,可用数据越多,结果的质量和准确性就越好。因此, AI 所需的数据量可能远高于某些高级分析。
- **AI 速度:** 必须使用最新数据,才能达到某些 AI 洞察和最佳预测所需的响应速度。有时甚至需要实时数据或非常接近实时的数据。
- **数据种类:** 数据越多,而且是可增添背景信息的数据越多, AI 结果通常越好。然而, AI 模型的结果可通过操纵数据(恶意或无意)甚至只是广义的“数据漂移”来改变,因此确保所有数据得到正确处理至关重要。
- **数据质量:** AI 对数据高度敏感,因此数据必须能够准确反映基本现实。在某些情况下, AI 密切关注其他传统分析方法可能认为是异常情况的峰值,因此准确性非常重要。
- **人类观点:** 人类如何看待数据 — 包括我们自己经验中的偏见和大脑“黑盒”中的不透明性 — 通常会影响数据使用方法。数据具有背景意义,因此必须结合背景查看和理解数据。如果没有适当的背景信息,可能会无意误用或误解数据。

然而,企业常常被这些复杂性所淹没,在应对数据挑战时受困于动机虽好但不一致的方法。为了满足 AI 的独特数据需求,组织首先必须对试图解决的业务问题有着清晰的认识,然后运用务实的方法来解决这些问题。

密切关注业务问题

一些经过传统培训的数据专家仍然难以应付 AI,他们过分关注数据科学和工程的细节,也就是*如何使用 AI*。如果无法充分理解和考虑较大的业务问题,例如企业*为什么使用 AI*,那么概念证明和研究型项目可能会激增,但不会带来效益。此外,数据科学家和工程师往往倾向于采用“大数据解决方案”来解决业务问题,而这些问题有时可通过高质量、定位精确甚至*定性式*的“小数据解决方案”来解决。

团队首先需要回答两个基本问题:试图解决什么业务问题?如何最恰当地解决?有时,最复杂的 AI 可能并不是最佳答案(见图 1)。根据数据揭示的洞察以及 AI 建议或采取的行动,就可以明确何时以及如何干预业务工作流程。

要快速取得成功, 展示积极的业务成果, 就必须始终关注于业务。

图 1

需求匹配

从 AI/分析“连续体”中选择适当的工具。



实现价值的务实方法

鉴于数据对于 AI 的重要性, 以及在获取、整合、准备和正确管理数据方面存在诸多挑战, 许多组织都急于通过开展单一项目, 整理组织内的所有数据。这通常包括将数据放入大型数据湖中, 尝试一劳永逸地解决所有数据问题。当然, 这基本上不可能取得成功。

这种想法不切实际, 因为其范围通常不明确, 导致无法实现合理的投资回报 (ROI)。更重要的是, 数据和企业的需求变化太快, 无法完成如此庞大的工作。

相反, 具备坚实的 AI 能力的组织则学会如何务实地思考 (请参阅“[The Weather Company: 从预测业务中汲取的经验教训](#)”)。哪些数据随时可用、易于访问而且干净? 要快速取得成功, 展示积极的业务成果, 就必须始终关注于业务。使用存在质量问题或可用性问题的数据会摧毁快速取得成功的可能性, 使组织陷入冗长而低效的数据清理工作泥潭。

成功的组织在中短期内以面向业务的用例为重点, 但在长期内则关注 AI 带来的持续价值。

The Weather Company: 从预测业务中汲取的经验教训

The Weather Company (TWC) 的看家本事是让海量数据发挥作用。更具体地说,就是应用 AI 和各种模型,生成预测和洞察。该公司已经在 AI 领域摸爬滚打二十年,在大规模应用 AI 方面处于全球领先地位。在此过程中,他们汲取了大量关于满足 AI 特殊数据需求的经验教训。

首先是不要低估手头的任务。整合天气与销售数据看似很简单,但绝非如此。在提取数据、了解如何执行必要的数据联接以及了解结果所带来的影响时,需要从经验中获得洞察。例如,模型是需要云端集中运行,还是在最终用户的设备中以“边缘计算”模式运行?模式是接受在线训练并持续更新,还是接受离线训练并定期更新?答案会影响数据处理方式。

事实上,TWC 获得的一个重要经验是,企业需要为数据准备与整合制定切合实际的计划,准确预测所需成本、资源和时间。经验表明,团队可能将 80% 的时间用于数据整理,而用于实际运行模型、调整模型和研究结果的时间则少得可怜。自动化有助于解决这个问题。

想想 TWC 的一个工作领域:预测季节性流感和过敏影响。在整合所有必要数据方面存在独特的挑战,这意味着必须构建自定义数据管道和服务甚至创建用于建模的特殊数据湖。这需要深厚的数据技能,也就是需要回答看似简单的如下问题:我们可以按照自己想要或需要的任何方式进行数据切片吗?幸运的是,在第一次面对这种情况时,TWC 已经拥有能够胜任这项任务的工程师和数据科学家。

其他同样重要的因素呢?该团队得到了高管的大力支持,高管理解、认可该项目,并支持投入大量时间和资源以实现目标和预期结果。

也就是说,许多普遍关注的数据领域并非新生事物;有些甚至已经出现了数十年,但在 AI 时代获得新的紧迫性、重要性和改进机会:

1. **整合。**建立可以从整个组织以及外部为 AI 挖掘数据的能力。
2. **治理。**在管理 AI 数据时应用最新的治理方法,以建立对洞察的信任。
3. **工具。**为需要的团队提供必要工具。

凭借适当的业务敏锐度和务实精神,企业可在持续建立制度化的数据能力方面取得进展,以满足 AI 的一些独特特征。

整合势在必行

新冠病毒疫情生动地展示了“过去的表现不能代表未来的结果”这一标准“免责声明”如何应用于数据以及从中得出的洞察。依赖前几年数据和历史模型进行分析的企业发现,自己的预测毫无用处。⁹ 为了生存,他们必须开始采集短期数据,例如,最近的销售和天气数据,甚至油价,并更新 AI 模型以执行相关性更强的预测。事实上,AI 之所以较快地流行起来,主要是因为它能够采集多变的短期数据以扩充现有数据,快速调整并产生相关预测。

这些可快速更新的短期数据大部分来自企业外部,因此用于整合第三方甚至公共数据的强大方法日益成为 AI 不可或缺的能力。

追踪正在发生的重大事件及其后续影响(如失业、可支配收入减少)的能力,可提供管理洞察。

如果有效利用具有适当颗粒度的数据源,那么从中得到的洞察就可以实现高度本地化,真正发挥作用。追踪正在发生的重大事件(如疫情)及其后续影响(如失业、可支配收入减少以及随之而来的对购物模式的影响)的能力,可提供管理洞察。

证明这种本地化方法的一个例子是 IBM COVID-19 Lockdown Index。该指数根据最新数据每日更新,帮助人们了解疫情对美国各县经济活动的当前影响。企业可将该指数与自己的数据相结合,创建有用、切实可行的预测,然后相应地进行规划和调整(请参阅“观点:整合不同数据源,形成本地疫情洞察。”)。

借助治理和溯源,赢得信任

使用外部或第三方数据进一步表明,组织需要非常了解自己的数据。谁是所有者?谁可以批准使用以及确定时间长度?基于第三方数据的 AI 资产的商业化可能会引发更复杂的问题。如果 AI 资产是用企业不再有权访问或批准使用的数据训练的,那该怎么办?虽然这些问题显然适用于外部数据,但内部数据也可能面临类似的问题。

超过半数的组织在数据整合方面苦苦挣扎,而三分之一的组织对自己连接多个数据源的能力缺乏信心。¹¹

接下来是受监管数据的问题。国家边界可能会影响数据可用性及其使用方式。数据质量也可能因国家/地区而异,这意味着整个地理区域中的数据可能不一致,或无法以相同的方式使用。例如,在一些国家/地区,由于部署了先进的微芯片和读卡器,可以捕获非常详细的信用卡交易数据;而在其他缺乏这种硬件的国家/地区,信用卡机器收集到的交易信息就比较少。

适当的数据治理方法有助于解决这些问题,包括了解并追踪数据来源。简单而言,数据治理是指管理不同类型数据所采用的规范、原则和规则。数据治理不仅适用于局部层面,比如组织内部,以帮助妥善处理数据,在数据生命周期的各个阶段保持其完整性和有效性;它还适用于组织之间、生态系统之间甚至国家/地区之间的合作,确定如何共享数据。

为了使数据更易于访问和管理,也许适当的规则和文化比适当的技术更为重要。熟悉标准条款和条件以及组织处理数据的方式的数据文化,更有可能帮助组织有效共享数据并充分利用数据。

观点： 整合不同数据源，形成本地疫情洞察

一般来说，如果知道一场巨大的风暴即将袭击某个国家/地区的东半部，人们可能会予以关注；但如果知道它会袭击自己所在的州、县或城镇，或者从你身边经过，那么会引起更大的关注。本地化的洞察可能作用很大。

IBM COVID-19 Lockdown Index 数量化了每个地区当前的中断程度。它将纵向信息与每日实时更新的疾病蔓延、病床和感染统计数据、当地社区限制以及美国整体市场波动措施结合在一起。然后预测各县何时到达疫情曲线的高峰，以及每个地区独特的曲线下降趋势，之后分配风险评分并确定当地解除封锁的速度。

它的用户友好界面直观说明美国各县的封锁程度以及未来四周最可能出现的封锁走势(见图 2)。

该指数使用公开可用的数据，包括：州医院的流行病学措施要求；各县疫情病例数；病床使用情况；重症监护病房床位使用情况；使用中的呼吸机数量。该系统还抓取当地新闻报道，以获取有关学校开放/关闭日期和其他潜在的社区限制方面的更新信息。

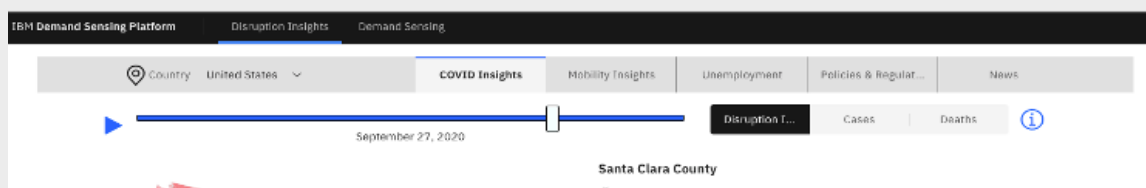
另一个类似的计划是 The Emergent Alliance，这是 IBM、劳莱斯、微软和数十家全球企业组成的非营利合作联盟。该联盟致力于提供关于新冠病毒病例的准确和最新的地区状况，以帮助地方当局更有效地应对疫情爆发。

本地化的风险指数结合感染率、社交媒体、新闻、Airbnb 数据等多种数据。分析内容包括疾病对健康的影响、政府的响应及公众行为的变化及其整体经济影响。

但该项目的优点不仅限于应对疫情。关于新闻主题的情绪数据可以帮助预测新的行为模式。例如，关于户外运动的新闻或山区的 Airbnb 预订增加，可能会引起关于徒步旅行或其他相关商品和服务的活动。¹⁰

图 2
直观洞察

COVID-19 Lockdown Index



预览已结束，完整报告链接和二维码如下：

https://www.yunbaogao.cn/report/index/report?reportId=1_46721

