

当数据迫使理论改道：开普勒、伽利略与牛顿

Agentic Data Scientist 并不是凭空出现的新职业。它试图自动化的，是一条在科学史中反复出现的发现链：提出问题、取得证据、创造解释，再让解释接受检验。下面三个故事不是完整的科学革命史，却分别揭示了这条链上的关键转折。

1601–1609 ·开普勒：不能被忽略的 8 角分

第谷把解释火星轨道的任务交给开普勒。当时连开普勒也相信，天体应沿“完美”的圆运动。他构造的圆轨道模型已经非常接近观测，但在某些位置仍相差 6 到 8 角分。换一个研究者，这可能只是可以修饰掉的残差；开普勒却相信第谷的观测精度，拒绝让漂亮理论压过数据。

这 8 角分迫使他离开圆轨道，穿过大量失败计算，最终得到椭圆轨道与面积定律，并在 1609 年出版《新天文学》。这个故事的核心不是“拟合出了一个更好的曲线”，而是知道什么时候应该怀疑模型背后的基本假设。[H1,H2]

1610 ·伽利略：先创造一种新的观察方式

《星际信使》记录了望远镜下凹凸不平的月面、肉眼不可见的恒星，以及木星附近数个亮点在连续夜晚中的位置变化。尤其是木星卫星的序列观测，给出了“有些天体并不围绕地球运行”的直接实例。

望远镜在这里不只是把旧计算做得更快。它改变了什么可以成为数据，也改变了哪些问题能够被提出。对今天的 ADS 而言，数据源、传感器、检索工具和实验环境同样不是被动输入；创造新的观测方式，本身就是解决问题和创新的一部分。[H3]

1684–1687 ·牛顿与哈雷：发现也需要协作系统

开普勒描述了行星怎样运动，却不知道是什么力量造成这种运动。1684 年，哈雷就轨道问题拜访牛顿，推动他把几页回答扩展成《自然哲学的数学原理》。1687 年出版的《原理》用运动定律和万有引力把地面与天空纳入同一套可推演框架。

出版过程还有一个真实的插曲：皇家学会此前在《鱼类史》上花费过多，难以兑现出版经费，哈雷接手推进并承担费用。科学创新从来不只是“一个天才和一个公式”；好问题、前人成果、计算方法、协作者和传播机制共同构成了发现的基础设施。[H4]

三个故事给出同一条线索：数据科学发现不是“数据进、答案出”。它会因为异常重新提问，因为新工具获得新证据，因为旧理论不足而创造新结构，并把结果重新放回观测中检验。

[H1] NASA, Orbits and Kepler's Laws ; [H2] Stephenson, Kepler, DSB ; [H3] Galileo, Sidereus Nuncius, 1610, Library of Congress ;

[H4] Royal Society, Principia manuscript history。

从验证一项解释，到自动化发现循环

开普勒的 8 角分迫使理论改道，但一个更难的问题仍在后面：当数据似乎支持一个解释时，我们怎样知道自己没有被巧合、偏差或一个动人的故事欺骗？进入 19 世纪以后，发现的主角不再只是新定律，也包括一套逐渐成熟的验证方法。

1854 · 约翰·斯诺：地图不是结论，干预也不是自动证明

1854 年的伦敦苏豪区，死亡记录在宽街水泵附近密集起来。地图让斯诺看见了聚集，却没有替他宣布结论。他继续追问住户从哪里取水，也调查那些“不合群”的病例：为什么有些离水泵很近的人没有发病，有些较远的家庭反而出现病例？这些例外把一幅地图变成了对饮水来源、病例位置和替代解释的调查。

后来流传的版本往往只剩下一个戏剧性动作：9 月 8 日拆掉水泵把手，疫情随即结束。斯诺本人却在 1855 年的报告中写得更克制—此前新发病例已经明显下降，因此无法仅凭先后顺序判断停泵造成了多大效果。真正有力量的不是英雄结尾，而是他把传播机制的证据，与一次干预的因果效果分开讨论。[H5]

1926–1962 · Fisher 与 Tukey：把可靠性写进方法

半个多世纪后，Fisher 面对的是 Rothamsted 的田野。不同地块天然就有差异；如果处理安排得不好，再精巧的检验也分不清收成变化来自肥料还是土壤。他因此把问题向前推了一步：不要等数据到手后才问“用哪个检验”，而要先问怎样安排田块与处理，才能提高精度并得到有效的误差估计。可靠性开始于数据产生之前。[H6]

1962 年，Tukey 又把镜头从试验田转向整个学科。他主张数据分析不应只是为既定模型做收尾，而是一种会在探索、计算和图形中不断发现意外的经验活动。后来，可视化、预测与可复现性继续扩展这条路线：数据科学的对象逐渐从“某个模型”，变成从问题到证据的完整过程。[H7,H8]

2024–2026 · 当机器开始生成假设，瓶颈转向验证

时间来到 2024 年，The AI Scientist 已能把研究想法送进代码、实验、论文和模拟评审；2025 年的 AI Co-Scientist 则让候选假设在生成、反思、排序与演化中竞争。机器第一次可以如此快速地推进研究流程的多个环节，但“能完成流程”仍不等于“发现值得相信”。[H9,H10]

设想一夜之间出现上百个听起来合理的假设：现实实验不会因此便宜，独立复现和同行审查也不会自动加速。Google DeepMind 将这种不对称概括为“验证瓶颈”。当候选答案不再稀缺，真正昂贵的能力变成：判断什么问题值得验证，设计什么证据能够淘汰错误答案，并在证据不足时停止。[H11]

现在，四个时代的线索接在了一起：斯诺不让地图替代因果判断，Fisher 把验证前移到数据生成，Tukey 让探索重新打开问题，而研究 Agent 让“提出候选”突然变得廉价。所谓 Agentic Data Scientist，正是要把这些能力组织成一个会被证据改变方向的行动循环。

[H5] Snow, On the Mode of Communication of Cholera, 1855 ; [H6] Fisher, The Arrangement of Field Experiments, 1926 ; [H7] Tukey, The Future of Data Analysis, 1962 ; [H8] Donoho, 50 Years of Data Science, 2017 ; [H9] Lu et al., The AI Scientist, 2024 ; [H10] Google Research, AI Co-Scientist, 2025 ; [H11] Google DeepMind, Validation Bottleneck, 2026.

什么是 Agentic Data Scientist ?

设想团队只交来一句话：“用户正在离开，请找出原因。”一个分析助手可以立刻读取表格、训练流失模型，再把相关性最高的变量排成报告。但数据科学家会先停下来：谁算“用户”？“离开”由哪个时间窗口定义？看到的是行为变化，还是埋点、渠道和用户构成发生了变化？在这些问题说清楚以前，连目标列都还不是一个可靠的研究对象。

第一轮分析也不会结束故事。假如折扣使用与流失同时出现，它可能是原因，也可能只是高风险用户更常收到折扣。为了区分两种解释，研究者可能重组队列、寻找新的数据、设计实验；若证据推翻原假设，就要换方法，甚至回去重问“我们究竟想解释什么”。**问题、解法与证据会在途中共同改变。**

这段往返揭示了“Data Scientist”的专业内核：发现什么问题值得问，决定怎样解决，在旧路径失效时**创造新的假设、数据或方法**，并设计可能证明自己错了的**验证**。[1,2]

本书的定义

Agentic Data Scientist (ADS) 不是一条更长的自动化流水线，而是能够自主推进数据科学发现循环的 AI 系统：面对尚未定义清楚的目标与现象，它能提出有价值且可检验的问题，构造并执行求解路径；当现有数据和方法不足时，它会产生新的假设、证据来源或解法，并主动寻找能够验证或推翻自身结论的证据。

验证不是流程的终点，而是下一步的方向盘：证据支持时继续求解或形成暂时结论；证据反对时改变假设、数据或方法，必要时重新提出问题；证据不足时承认未知并停止。

四种一级能力

能力	真正要完成的事	不等于
提出问题	从现象、矛盾和目标中发现未知，把它表述成重要、可回答、可证伪的问题	改写用户的问题，或补齐任务模板
解决问题	在信息和资源不完备时形成假设，寻找证据，设计路径并执行；结果不符时能够换路	按固定流程调用 SQL、Python 或模型
创新	当已有方案不足时，提出新的假设、代理数据、表示、算法、实验或工具，扩大可解空间	生成更多候选答案，或使用更复杂的模型
验证	主动寻找反例和替代解释，检查识别、稳健性、不确定性与复现性，让结论接受证据约束	只报告一个较高分数，或选择支持自己的指标

为什么“验证”必须单独成为一种能力？

开普勒的 8 角分和斯诺笔下已经开始下降的病例曲线，提醒我们同一件事：一个顺眼的解释可以很快出现，困难的是说明**什么证据会使它不再成立**。如果系统只寻找支持自己的结果，它完成的是论证，不是科学。验证要求系统设计对照、留出集、消融、敏感性分析、外部复现或现实实验，并把失败送回下一轮提问与求解。[3,4]

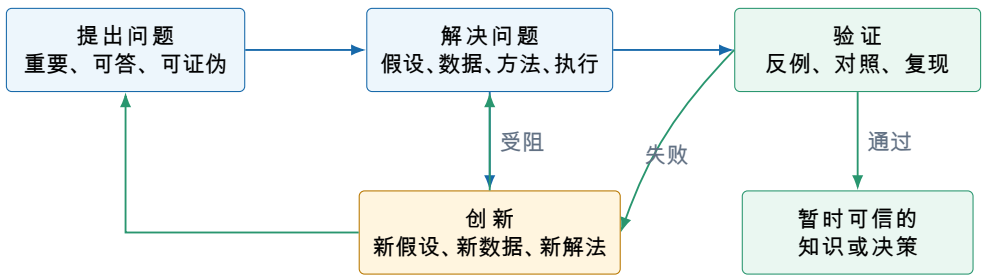
因此，本书不以“是否调用工具”判断 ADS。工具调用只是执行手段；**提出问题、创造解法并检验自己**，才把一个数据 Agent 推向数据科学家。

[1] 郑宇,《数据科学家》, CCCF 17(12), 2021 ; [2] Donoho, *50 Years of Data Science*, 2017 ; [3] Popper, *The Logic of Scientific Discovery*, 1959 ; [4] Shmueli, *To Explain or to Predict?*, 2010。

Agentic：让四种能力形成发现闭环

“会提问、会建模、会检查”仍可能只是四个彼此分离的功能。**Agentic** 的含义是：系统保存研究状态，观察行动结果，并依据证据决定接下来调用哪一种能力。验证失败后，它不只是重跑程序，而能回到假设、数据、方法甚至原问题本身。

图 1.1 数据科学发现不是直线流水线



Agentic 的最低含义

系统必须能够根据新证据改变研究控制流：选择继续求解、调用新的数据与工具、创造候选解法、加强验证、退回重提问题，或停止。若所有步骤和分支都由人预先写死，它是 Workflow；若只在既定目标和搜索空间内优化，它更接近 AutoML。[5,6]

不要再把三个层次混在一起

层次	内容	它回答的问题
科学家能力	提出问题、解决问题、创新、验证	为什么它不只是分析工具？
数据科学工作对象	行业问题、数据、模型与算法、计算平台	它在哪些对象上思考和行动？
Agentic 运行机制	状态与记忆、规划与行动、工具与环境、反馈与反思	四种能力如何持续接力并根据证据改道？

郑宇所说的“站在平台上看问题、想数据、关联模型”，刻画的是数据科学家的**端到端工作对象**；它非常重要，但不是最高层能力分类。LLM、Agent、Harness、数据与计算环境，则是实现这些能力的技术条件，而不是“科学家”一词的定义。[1,7]

一个可以被反驳的判断

如果系统面对失败只会换提示词、重试工具或继续调参，却不能重新审视“问题是否问对、证据是否足够、现有方法是否需要创新”，那么无论调用多少工具，它都还不是完整意义上的 ADS。

[5] Lu et al., *The AI Scientist*, 2024；[6] Gottweis et al., *Towards an AI Co-Scientist*, 2025；[7] Chen et al., *LLM-based Data Science Agent: A Survey*, 2025。

系统边界：能做到，不等于可以去做

四种能力回答“系统是否像一名数据科学家”；系统边界回答另一个问题：**这名数据科学家在当前任务中可以观察什么、改变什么，何时必须停下，以及谁承担后果。**能力越强，不代表权限越大，更不代表责任可以交给模型。^[8,9]

能力、自治、权限与责任是四件不同的事

概念	它回答什么	一个系统可能出现的情况
能力 capability	它能否提出、解决、创新和验证？	能设计可靠实验，但不会自主执行
自治 autonomy	它能连续决定多少步，何时需要人介入？	可在沙箱里自行探索，发布前必须审批
权限 authority	它被允许读取哪些数据、调用哪些工具、改变哪些状态？	可读脱敏数据，不能接触生产库
责任 accountability	谁批准、监督并对结论及外部影响负责？	研究负责人批准实验并承担发表责任

一个可执行的系统边界必须写清五件事
观察：允许读取哪些数据、文献、日志与反馈；行动：允许查询、写入、实验、部署或联系谁；资源：时间、算力、费用、凭据和重试上限；升级与停止：哪些风险、证据或预算条件触发人工批准或终止；责任：谁对数据使用、研究结论和现实后果负责。

定义边界：什么还不能叫 ADS？

系统形态	它已经能做什么	距离 ADS 还缺什么
分析助手	回答问题，生成 SQL、代码、图表或解释	问题、行动与验证主要由人完成
Workflow / AutoML	执行固定流程，或在给定目标与搜索空间内优化	不能重提问题，也通常不能创造搜索空间之外的解法
数据 Agent	调用工具完成多步数据任务，并依据结果调整计划	若问题始终由人给定，且不具备创新与主动证伪，仍只是执行型 Agent
Agentic Data Scientist	自主组织提问、求解、创新和验证，并依据证据改道	仍受明确权限、批准和责任边界约束

四个可检验的最低判据
提问：它能提出用户没有直接给出的新问题，并说明为什么值得研究吗？求解：当基线失败时，它会改变假设、数据或方法，而不只是重试吗？创新：它能提出已有方案之外的新解法，并与强基线比较吗？验证：它能事先说明什么证据会推翻自己的结论，并真正执行检查吗？
四问都能留下可审查的证据，才接近本书所说的 ADS。高自治、长轨迹或大量工具调用，都不能替代这四项判据。

[8] Feng et al., *Levels of Autonomy for AI Agents*, 2025 ; [9] NIST, *AI Risk Management Framework 1.0*, 2023 ; 参见 Kapoor et al., *AI Agents That Matter*, 2024。