数据底座洞察 02

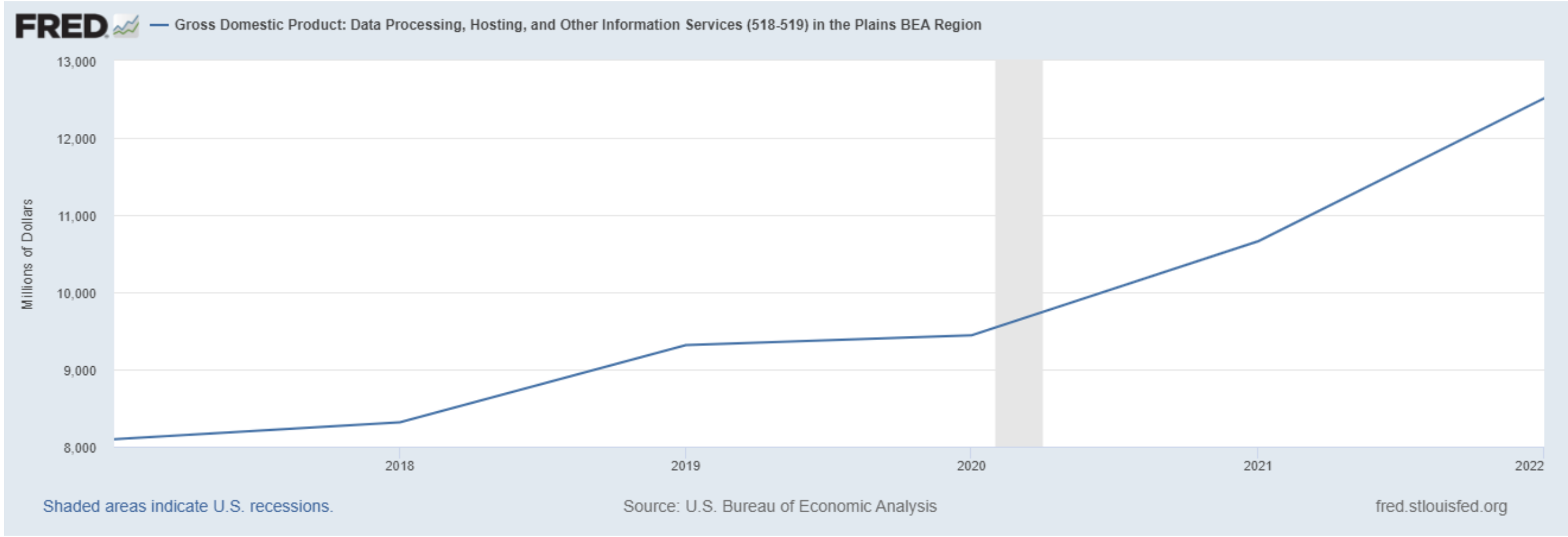
Jiangsheng Yu

02/05/2024

数据科学家的需求增长

- 美国国家教育统计中心报告称，2022 年将授予 897 个数据科学学士学位，而 2020 年仅为 84 个。
- 美国劳工部预测未来 10 年数据科学职位将增加 36%。 研究表明，主修数据科学的女性多于计算机科学和网络安全。
- ACM 教育委员会认为，一些大学正在让学生远离计算机科学，而强调跨学科的数据科学。

美国数据处理、托管和其他信息服务 GDP

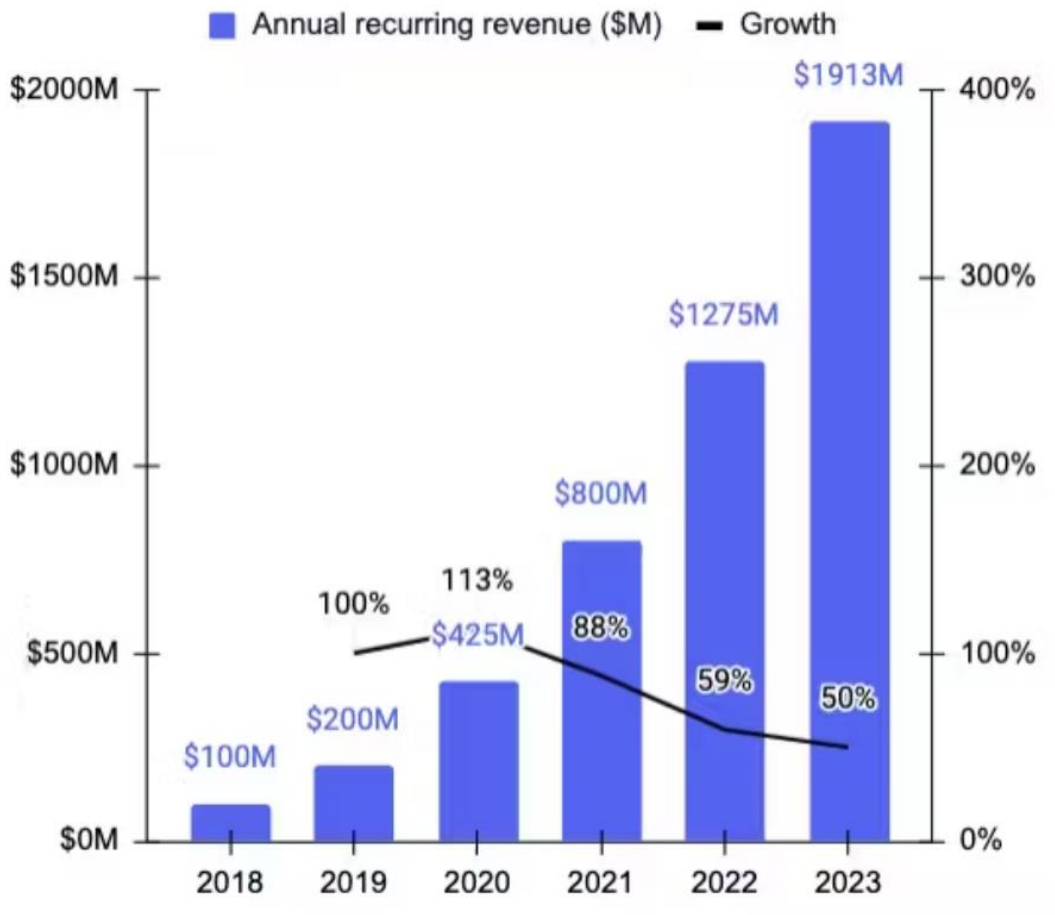


Databricks 增长情况

2023年的情况

收入：$1.91B
估值：$43.00B
年增长率：50%
资金：$4.00B

Databricks已从 Franklin Templeton、Counterpoint Global 和 Andreessen Horowitz 等投资者那里筹集了 40 亿美元。截至 2023 年 9 月，该公司的估值为 43B 美元，使其成为全球最有价值的私营公司之一。



图片来源于 https://sacra.com

Databricks 的商业模式

- Databricks为数据处理和人工智能应用程序构建开源软件，然后提供具有附加专有功能的付费版本，而公司无法自行轻松复制这些功能。虽然开源软件为公司提供了灵活性，使其不被锁定在专有架构中，但大多数公司通常没有工程人才来管理其复杂性。

- Databricks向企业出售其开源软件的完全托管版本，以及用于编写查询的 SaaS 工具和用于连接数据源的连接器等附加实用程序。这方面的Databricks与 AWS 类似，也为开源软件提供托管服务，但是Databricks生产其管理的所有开源软件，这使其比其他软件更具优势。

- Databricks首先使用 Apache Spark 对数据湖中的大型原始数据集运行查询。然后，它通过推出进军相邻市场的产品来扩大收入，例如人工智能生命周期管理/MLOps (MLFlow)、数据仓库 (Delta Lake)、数据可视化 (Redash) 以及 BI 和分析 (数据库SQL）。

Databricks 的商业模式

- 它通过向客户收取使用该平台和帮助他们设置Databricks 的专业服务的费用来赚钱。据估计，其约80%的收入来自平台，其余来自专业服务。2023 年 4 月，他们的数据仓库产品Databricks SQL 在推出一年后年度经常性收入（ARR）达到 1 亿美元。

- Databricks采用即用即付模式，根据客户的级别、使用软件的处理能力以及使用时间向客户收取费用。更昂贵的高级层和企业层提供更多的安全性、治理、更高的速度和数据处理功能。Databricks在 Microsoft Azure、Google Cloud 和 AWS 之上运行，每个层级和计算能力的收费略有不同。

- Databricks主要向大型企业销售产品，每年的合同金额达数百万美元。拥有超过 7,000 名客户，净保留率超过 150%。其一些大客户包括壳牌、CVS Health、Regeneron、T-Mobile、汇丰银行和Comcast。

Databricks 的产品：Spark

- Databricks是由 Apache Spark 的团队创建的。Apache Spark 是一个开源软件，用于在数据湖上运行查询，用于廉价地存储大量原始数据。Spark 于 2009 年推出时，大多数数据湖都托管在 Hadoop（数据中心的第一个操作系统）上。然而，在 Hadoop 上运行大型查询非常麻烦并且需要花费大量时间。Spark 通过让在数据湖上运行查询变得更容易、更快捷，找到了适合其初始产品市场的机会。虽然它最初设计为在 Hadoop 之上运行，但现在可以在任何云存储上运行，例如 AWS、Google Cloud 或 Microsoft Azure。

- Databricks的核心产品是托管 Spark 集群，即用于运行数据分析的机器组。它为数据科学家提供了一个基于 Web 的门户，以创建这些 Spark 集群来运行其数据分析工作负载。该门户还包括一个类似笔记本的工作区，供数据科学家用 SQL、Python 等协作编写查询，以及一个用于定期运行数据管道的调度程序，数据工程师可以将其用作 Airflow 或 Prefect 的替代品。

Databricks 的新产品： MosaicML

2023 年 6 月，Databricks以 14 亿美元收购了 MosaicML，此举旨在增强其训练大语言模型 (LLM) 和图像生成模型的能力。MosaicML 开发了工具和基础设施来简化和降低LLM从数据准备到训练和管理基础设施的运行成本。

- 像 GPT-3 这样的LLM的训练涉及大量成本，但 MosaicML 声称能够为其客户训练 GPT-3 质量模型，费用低至 32.5 万美元（相比之下，Google LaMDA 为 36.8 万美元， Bloom 为 100 万美元，GPT-3 为 84.1 万美元）。

- MosaicML 还包括全套 ML Ops 工具，从而进一步减少训练模型所需的专家。

Databricks 的其他新产品

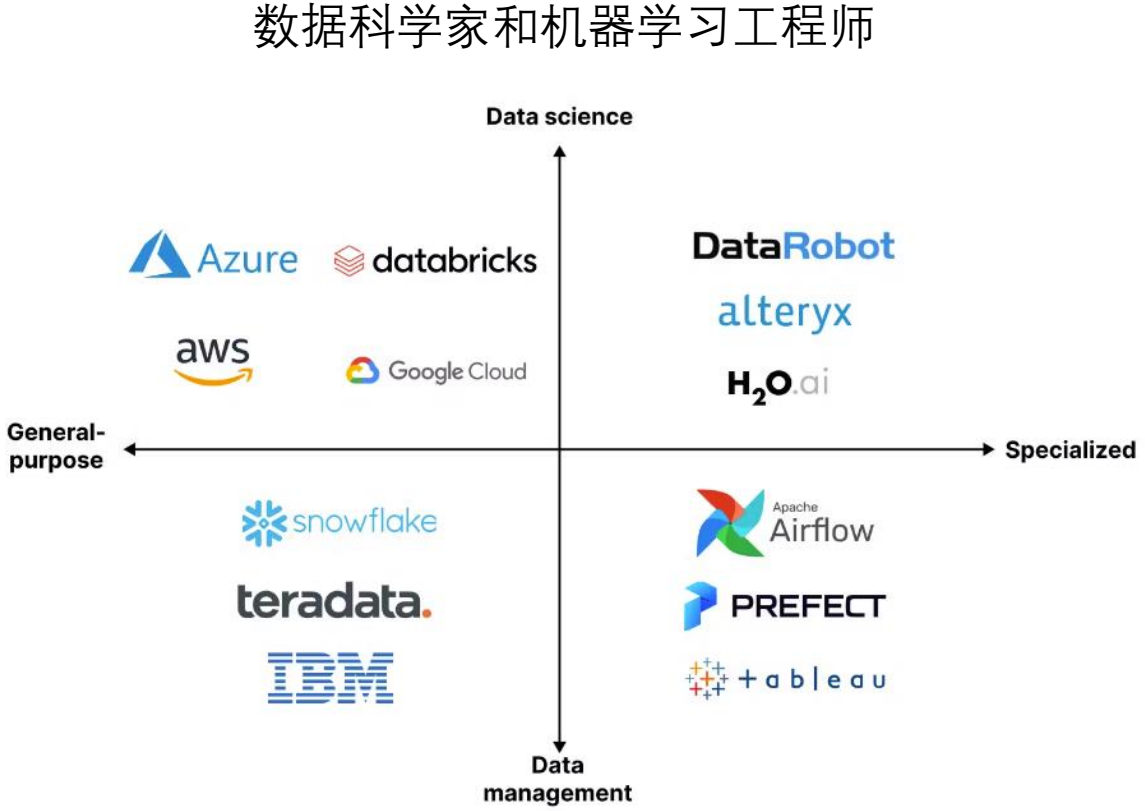
- MLFlow：它是一个机器学习生命周期的平台，数据科学家可以在其中构建机器学习模型、跟踪实验、将其部署到生产中并监控其性能。
- Delta Lake：它是一个运行在数据湖之上的层，可以加速数据查询，类似于数据仓库的查询速度。Delta Lake 的关键组成部分Databricks致力于进军 BI 和数据分析领域，与 SnowFlake、Amazon 等数据仓库公司竞争。
- Databricks SQL：它是一个数据仓库，允许用户在 Delta Lake 之上运行 SQL、创建可视化并构建/共享dashboards，旨在替代 SQL 数据分析师。

Databricks 与 Snowflake 的竞争

- Databricks 的最大竞争对手是 Snowflake（近来，Snowflake添加了数据科学产品，例如 Snowpark、对 Python 的支持）。未来几年，预计 Databricks和 Snowflake 在不失去市场份额的情况下实现增长，类似 AWS、Google 和 Azure 在云市场的情况。

- Databricks还与运行特定任务的数据管理和数据科学领域的专业解决方案竞争。例如，Databricks的调度程序与 Airflow 类似，其 MLFlow 产品与 Datarobot 和 Alateryx 竞争。Databricks 的优势在于拥有从数据传入到部署 ML 模型的整个管道，但它也比专门的应用程序更昂贵。

- Databricks的风险来自数据仓库的粘性：Databricks押注于未来各公司将停止使用单独的数据仓库软件，转而使用满足所有数据处理/存储要求的Databricks 。然而，数据仓库和ERP一样，都是高粘性的产品，对于大企业来说并不容易剥离。这可能会使其销售周期变得非常长，并限制其可用市场。



Databricks 向 AI 进军

- 人工智能：Databricks通过收购 MosaicML，进入了与 OpenAI 交叉的领域。通过向公司提供工具和基础设施来从头开始创建自己的人工智能应用程序。不像OpenAI提供专有模型，Databricks的方法使公司能够利用自己的数据进行 AI 模型训练，从而来吸引热衷于保留对其数据和 AI 资产的控制的组织。

- 营收模式：Databricks依赖对其工具集的订购并根据使用情况收费。添加 MosaicML 经济高效的模型训练基础设施可以吸引更广泛的客户群，以减少训练LLM的费用。训练LLM的有竞争力的定价可能成为推动额外收入的独特卖点。

- 广泛合作：与 Nvidia合作，为 Nvidia 支持的服务器优化其软件。通过 Azure 与 Microsoft 合作。

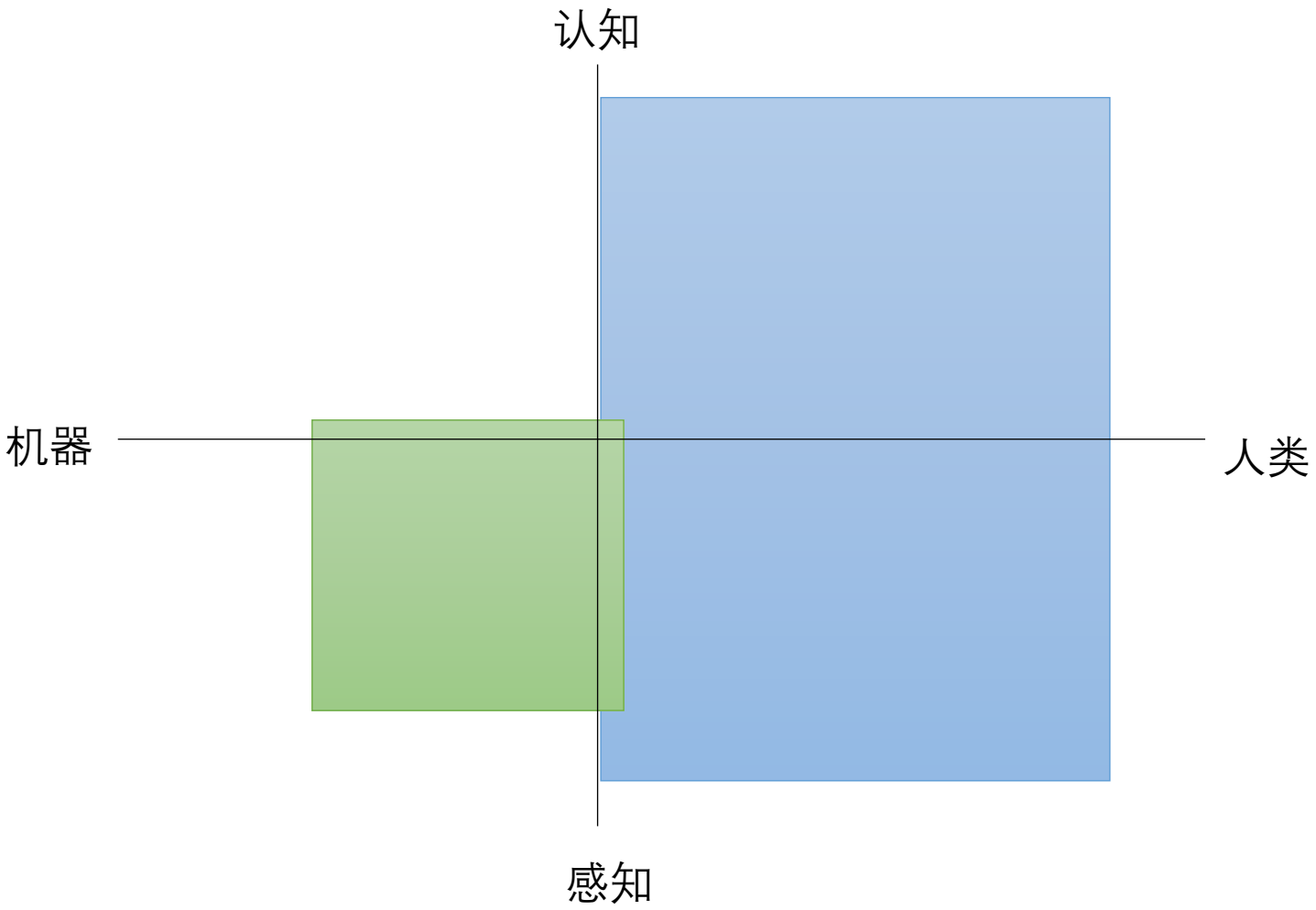
数据集中化

Databricks 认为，**数据集中化**将成为一个大趋势。其原因如下：

1) 凭借廉价的云存储和快速的网络，大多数公司正在从分析 ERP 中的组织数据和 Salesforce 中的客户数据转向将所有数据存储在中央数据存储中。这有助于他们更好地了解发生了什么（**商业智能**）和将要发生什么（**预测分析**）。数据湖技术容纳所有数据，无需担心数据的来源和类型。

2) 目前，大多数公司使用数据仓库来运行实时商业智能操作，并使用数据湖来进行机器学习/数据科学项目。Databricks认为，随着越来越多的事物走向数字化，数据的爆炸式增长将使公司无法运行两个并行的大规模数据存储，而这两个存储将汇聚成一个。 随着 Salesforce 成为客户数据的行业标准，其 Data Lakehouse 产品将成为数据集中的行业标准。

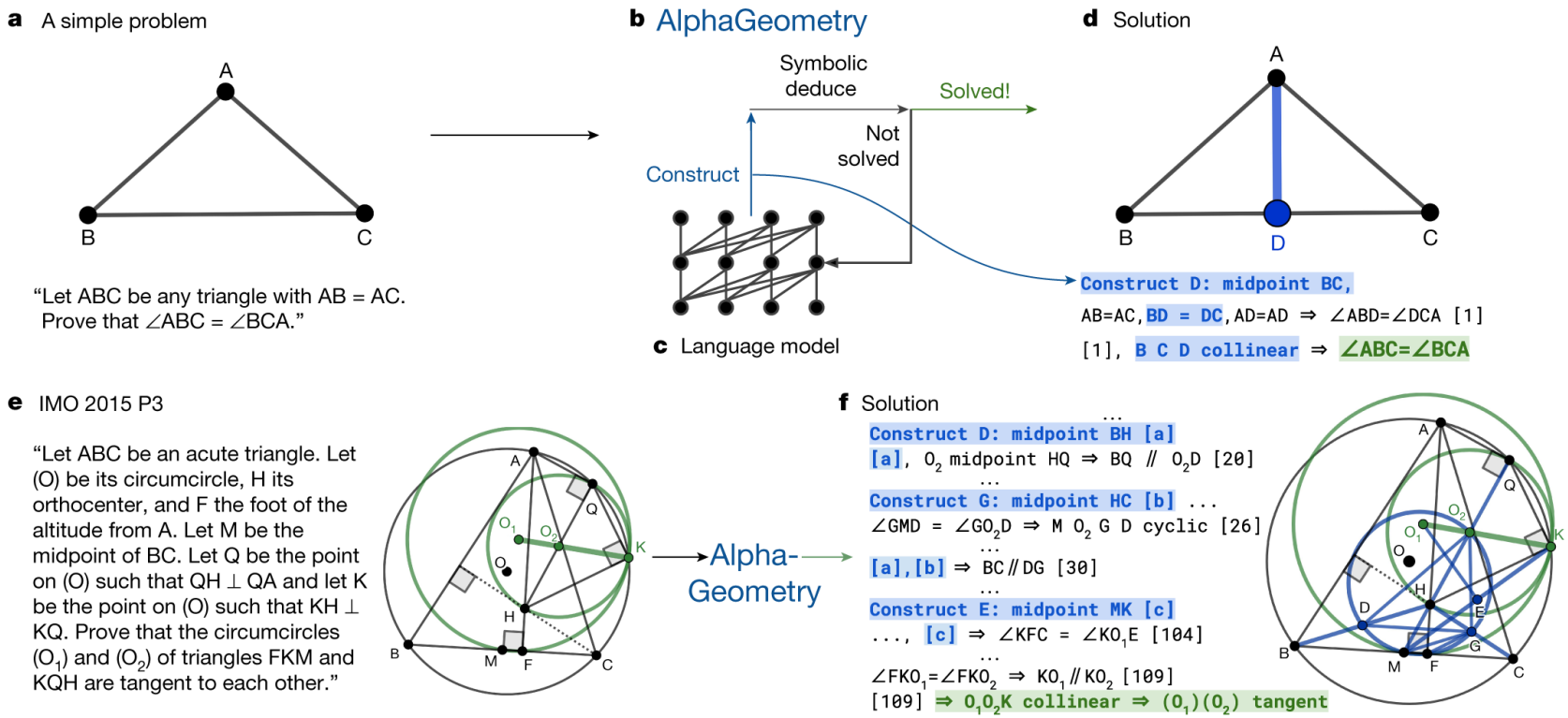
人类数据 vs 机器数据



- 全球每年产生的数据量已经达到了数十Zettabytes（1 Zettabyte = 1万亿GB）。
- 目前，人类数据的总体质量较高，常用作训练数据。但产量低、参差不齐。
- **人机协同**产生数据将是一个过渡阶段（如transformer+RLHF）。
- 机器智能部分地来自人类数据中的语义**关联性**、内在**逻辑**等，缺乏可解释的知识表示与推理。
- **生成式AI**和AGI将使得机器越来越高效地产生高质量的数据。未来数据会出现大爆发，地球变成一个数字星球。
- 因此，数据底座要应对大数据的**3V**（速度、容量、多样性），必须借助AI。

合成数据的应用实例

2024 年 1 月，在 Nature 上谷歌Deepmind团队发表了论文《无需人类演示即可解决奥林匹克几何问题》，提出了欧氏平面几何的定理证明器 AlphaGeometry 来产生人类可读的证明，它是一个基于LLM的神经符号系统，使用神经语言模型，在大规模**合成数据**（数百万个定理和证明）上从头开始训练，引导符号推演引擎搜索证明过程。在包含 30 个最新奥林匹克级别问题的测试集上，AlphaGeometry 解决了 25 个问题，超越了之前仅解决了 10 个问题的最佳方法（吴方法），接近了国际数学奥林匹克 (IMO) 金牌得主的平均表现。



数据底座为AI/ML赋能

人工智能 (AI) 和机器学习 (ML) 为各种流程提供自动化和优化，从而提升**生产力**。其基本驱动力之一是数据底座，它可作为改进和部署AI/ML模型的**原材料**。

- **数据收集**：从各种来源收集数据，允许AI/ML系统实时处理数据。
- **特征工程**：从原始数据中识别和提取有价值的特征来实践AI/ML。数据分析解决方案是AI/ML系统提高准确预测和分类能力的基础。
- **模型训练**：利用数据管道上的大量数据使AI/ML系统能够提高其性能，甚至决定AI模型的伦理。

特征工程：解决数据处理 70% 的工作量

近些年，数据科学发展迅速，尤其在特征工程以及与计算机科学（包括人工智能）的结合上，取得了长足的进步。

- 数据处理的目的是去粗取精、去伪存真，其手段多种多样——有时需要降维对数据进行有损压缩，有时为改善可分性需要把数据变换到更高维度的空间，有时需要探究潜在的不可观测的因子，有时需 要把数据变回原来的样子……。另外，原始数据所在空间的维度非常高时可能会引发复杂度激增、据稀疏（即数据向量或矩阵中绝大多数元素缺失或为零）等问题，进而造成经典统计方法的失效。这便是最优控制、机器学习、数据挖掘等领域中常会遇到的维度之咒 (curse of dimensionality)。特征 工程 (feature engineering) 有助于减轻来自高维的压力。
- 人工智能和大数据分析的发展，更加剧了对应用驱动的实用主义方法的需求。学科的融合以及与计算 机实践更紧密的结合，毋庸置疑，已成为现代数据科学的主要特点。例如，计算统计学 (computational statistics)，也称统计计算 (statistical computing)，是统计学、数值方法和计算机科学的交叉学科。 计算机已成成为数据分析和随机模拟的基本工具，计算统计学必将走入统计学的基本素质教育。

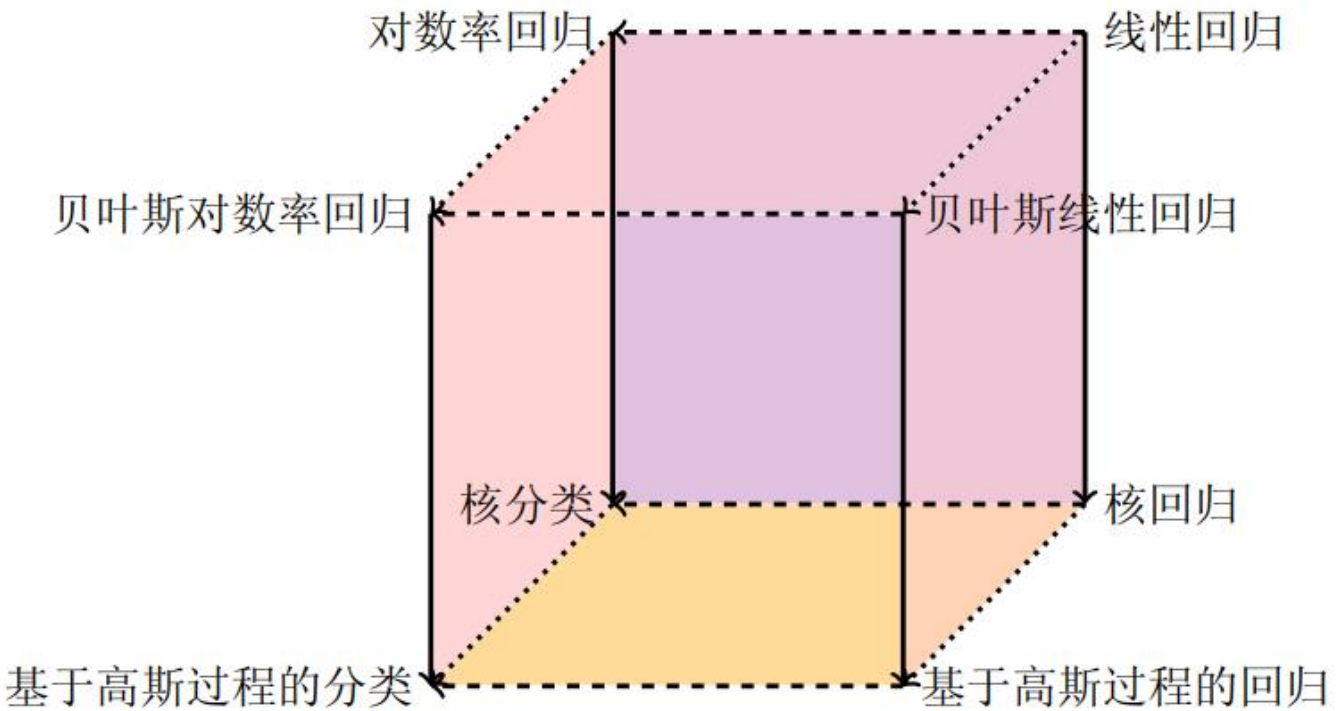
特征工程的方方面面

- **数据整理**：对数据进行汇总、分组、归类、编码、审核等，使之更加条理化，包括
  - ❑数据清洗：例如，独立成分分析、去噪、去伪存真。
  - ❑缺失数据分析（missing data analysis）：参考Donald Rubin的工作。
  - ❑诱导特征：由已知的特征构造新的、有意义的特征（譬如，由距离和时间定义平均速度）。
  - ❑数据压缩与重构：有损压缩和无损压缩（例如，主成分分析等）。
  - ❑数据合并：不同来源的同类数据（例如，不同实验室采集的某个癌症的质谱数据）如何合并成 一个更大的数据集?

- **关系发现**：如关联规则 (association rule)、知识图谱、因果分析，等等。

- **异常检测**：从多种角度（例如，分布、预测、模式识别等）判定异常点，然后找到引起异常的原因（根因分析更重要），特别应用于健康监测、故障诊断、欺诈检测、干扰分析等。

例子：回归/分类模型之间的关系



回归/分类模型之间的关系