拉多 · 利普什 （Rado Lipuš）

CFA， 另类数据供应商 Neudata 的创始人兼首席执行官。 在创立 Neudata 之前， 拉多拥有 20 多年在金融科技领导、 销售管理和数据创新领域的买方专业经验。 他曾在 MSCI （Barra） 和 S&P Capital IQ 从事量化投资组合构建和风险管理工作数年， 并为 CITE Investments 筹集资金； 后来在伦敦的 PerTrac 担任董事总经理， PerTrac 是一家领先的金融科技和数据分析解决方案供应商， 为欧洲、 中东、 非洲和亚洲的对冲基金和机构投资者提供服务； 他还曾在 eVestment、 2iQResearch、 I/B/E/S 和 TIMGroup 等金融数据公司工作。 作为公认的另类数据专家， 拉多经常受邀在会议和行业活动中发言。 拉多在奥地利格拉茨大学获得工商管理硕士学位。

达里尔 · 史密斯 （Daryl Smith）

CFA， Neudata 的研究负责人。 他和团队负责为全球范围众多的资产管理公司研究和发现另类数据集。 在加入 Neudata 之前， 他曾在精品投资公司 Liberum Capital 担任股票研究分析师， 涉足多个领域， 包括农业、 化工和多元金融； 在加入 Liberum 之前， 他曾在高盛担任股票衍生品分析师和监管报告策略师。 达里尔拥有巴斯大学机械工程硕士学位。

2.1 导读

大约 20 年前，一群
产管理者使用了另类数
据的基金经理数量与新

我们确定了在过去
Neudata 平台每月会新增
未来几年，我们希望数
来越多的数据创生公司
的和已有的初创企业正

2.1.1 何为“另类”

对不熟悉该领域的
据源，可用作投资管理
基本上指过去 10 年创生
用。在某些情况下，产生
各行各业的实体公司，
况下，另类数据是经济
料”。另类数据主要被买
险投资以及企业非投资

2.1.2 另类数据并不

术语“大数据”与“另
数据的语境中，在某些情

术语“另类数据”最

2.1 导读

大约 20 年前，一群特殊且具有创新思维的对冲基金经理与资产管理者使用了另类数据与机器学习技术。而近年来，使用另类数据的基金经理数量与新的商业可用数据源供应量均大幅增加。

我们确定了在过去几年可商业使用的 600 余个数据集。目前，Neudata 平台每月会新增约 40 个经过彻底审查的新另类数据集。未来几年，我们希望数据集总量能够平稳增长，因为一方面，越来越多的数据创生公司将其现有数据货币化；另一方面，新成立的和已有的初创企业正带来新鲜的、额外的另类数据产品。

2.1.1 何为“另类”？ 与传统相对

对不熟悉该领域的人而言，术语“另类数据”指一种新型的数据源，可用作投资管理分析与量化投资决策。本质上讲，另类数据基本上指过去 10 年创生的数据，直到最近才被投资界所接受和使用。在某些情况下，产生另类数据的本意是为非投资公司——遍布各行各业的实体公司，提供一种可使用的分析工具。在许多其他情况下，另类数据是经济活动的一种副产品，通常被称为“数据废料”。另类数据主要被买方与卖方所使用，同时也被私募股权、风险投资以及企业非投资客户所使用。

2.1.2 另类数据并不总是大数据，大数据并不总是另类数据

术语“大数据”与“另类数据”常常可交换使用，多用在非结构化数据的语境中，在某些情况下也指大量的数据。

术语“另类数据”最初的使用者为美国的数据代理商与顾问，几

Neudata 之前， 拉多拥有
险。 他曾在 MSCI （Bar-
， 并为 CITE Investments
家领先的金融科技和数据
机构投资者提供服务； 他
数据公司工作。 作为公认
在奥地利格拉茨大学获得工

资产管理公司研究和发现另
Capital 担任股票研究分析
um 之前， 他曾在高盛担
工程硕士学位。

年前这一术语得以被广泛接受。美国的资产管理行业比其他地区更能广泛理解另类数据的含义，例如，在欧洲，直到2017年这一术语才被广为认可。

2016年与2017年，卖方、传统数据供应商与其他类型的会议主办方举办了大量的会议与活动，确实帮助扩散了对另类数据的认识。此外，过去几年，卖方银行、数据供应商与咨询公司方针对另类数据与人工智能展开的诸多调研与报告，也推动了另类数据在买方与更广泛产业中认知度的提升。

我们所说的另类数据源究竟是什么？有多少数据源可用？哪些数据源是最适用的？

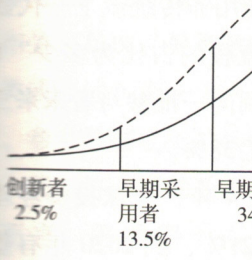
2.2 使用另类数据的驱动因素

2.2.1 创新的扩散：我们处于什么阶段

金融产业仍然处于使用另类数据的早期阶段（见图2.1），从积极寻求并研究另类数据源的买方数量上我们可以证实这一点。然而，对另类数据的使用正处在风口浪尖，正向早期大众阶段过渡，我们观察到大量的资产管理公司、对冲基金、养老基金和主权基金正在建立另类数据的研究能力。

大多数创新者与早期的应用者植根于美国，欧洲所占比例较小，亚洲基金所占比例则更小。大多数创新者与早期的应用者拥有系统化和量化的投资策略，并且在很大程度上拥有聚焦于消费者的主观对冲基金。

2017年，我们观察到使用基本面策略的基金的收益猛增。然而，尽管这些传统的基金经理对使用另类数据越发感兴趣，但量化策略对另类数据的吸收和运用明显更快。我们推测造成这一现象的



资料来源：Rogers（1962）。

主要原因之一是操作方式
究另类数据更具挑战性，
常并不充分，研究团队经
理、确保法律合规性以及
行全面改革，因此会带来

对于大型、成熟的资
提供测试数据的内部流程
供应商进行尽职调查；（
况下是免费的）；（3）获
部流程框架差别很大，因
时间也差别很大。创新型
太注重数据、组织较为低

在投资中使用另类数
的进步也改善了分析不同

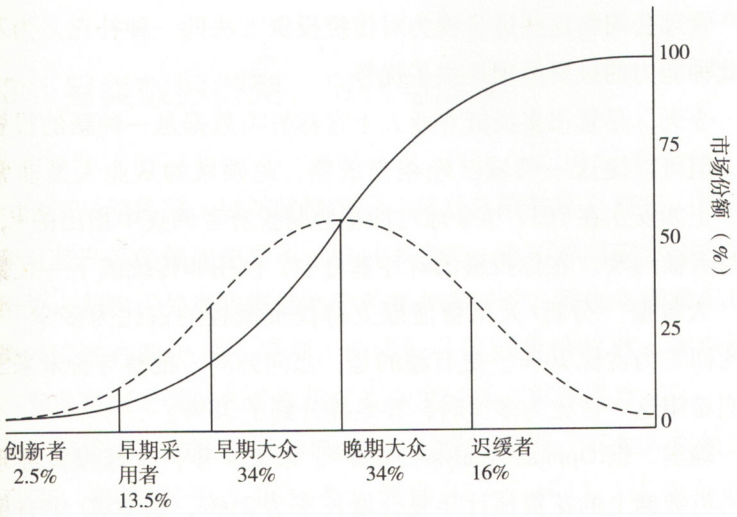


图 2.1　创新扩散规律

资料来源：Rogers（1962）。

主要原因之一是操作方式。简而言之，采取基本面策略的公司要研究另类数据更具挑战性，这是因为其所需的技术与数据基础设施常常并不充分，研究团队经常面临严峻的技术鸿沟。因此，评估、处理、确保法律合规性以及获得大量数据集的任务需要对现有流程进行全面改革，因此会带来组织上的巨大挑战。

对于大型、成熟的资产管理公司而言，一大阻碍是向研究团队提供测试数据的内部流程过慢。这一程序通常需要：（1）对新数据供应商进行尽职调查；（2）签署测试数据的法律协议（大多数情况下是免费的）；（3）获得合规团队批准。资产管理公司的这些内部流程框架差别很大，因此为研究团队组织大量的新数据集所需的时间也差别很大。创新型对冲基金可能需要几天或几周，而一个不太注重数据、组织较为低效的资产管理公司则可能需要几个月。

在投资中使用另类数据受到了金融科技进步的驱动，金融科技的进步也改善了分析不同数据集的能力。许多投资者、对冲基金与

资产管理公司将这些进步视为对传统投资方法的一种补充，为不具备此种能力的投资经理提供了优势。

今天，尽管很多投资专业人士宣称另类数据是一种新的投资领域，但可以说这一领域已经相当成熟，此领域的从业人员非常常见。正如安永在2017年全球对冲基金与投资者调查中指出的[①]，当参与者被问及“在你投资的对冲基金中，使用非传统或下一代数据及‘大数据’分析/人工智能以支持投资流程的占比为多少”时，平均回答为占比24%。更有趣的是，当问到同一批参与者未来三年他们希望这一占比为多少时，答案提升到了38%。

确实，据Opimas Analysis预测[②]，未来4年，全球投资经理用在另类数据上的花费预计年复合增长率为21%，到2020年有望超过70亿美元（见图2.2）。

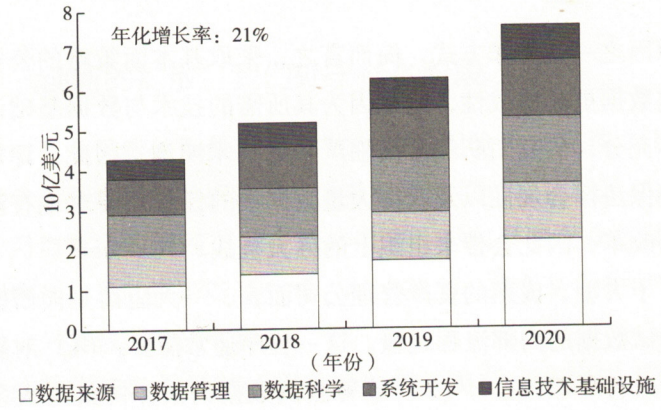


图2.2 另类数据支出

资料来源：Opimas Analysis。https://www.ft.com/content/0e29ec10-f925-11e7-9b32-d7d59aace167.

① http://www.ey.com/Publication/vwLUAssets/EY-2017-global-hedge-fund-and-investor-survey-press-release/$File/EY-2017-global-hedge-fund-and-investor-survey-press-release.pdf

② http://www.opimas.com/research/267/detail

2.3 另类数据类

出于一些原因，另
应商对其产品的描述
分相关。其次，另类
分类或描述为某一单
据、参考数据之类的

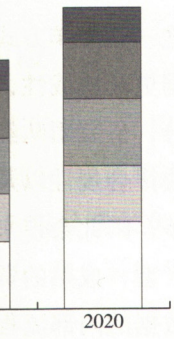
我们将不同的数
据而言，都适用于
（ESG）数据集可能包
（见图2.3）。更复杂
可以以不同形式来使

1. 未经处理的，

资料来源：Neudata。

方法的一种补充，为不具

类数据是一种新的投资领
领域的从业人员非常常
资者调查中指出的[1]，当
使用非传统或下一代数据
流程的占比为多少”时，
到同一批参与者未来三年
了 38%。
来 4 年，全球投资经理用
21%，到 2020 年有望超



发 ■信息技术基础设施
出
/content/0e29ec10-f925-11e7-9b32-

7-global-hedge-fund-and-investor-sur-
-investor-survey-press-release. pdf

投资

2.3 另类数据类型、形式与范围

出于一些原因，另类数据源的分类具有挑战性。首先，数据供应商对其产品的描述通常不一致且不完整，也不都与投资和资管充分相关。其次，另类数据的本质是复杂多面的，数据不能简单进行分类或描述为某一单一类型。而诸如分笔或价格数据、基本面数据、参考数据之类的传统数据的复杂度较低，更容易被定义。

我们将不同的数据源分成 20 种不同的类型，对大多数另类数据而言，都适用于多重分类。例如，一个环境、社会与治理（ESG）数据集可能包含“众包”“网页抓取”“新闻”“社交媒体”（见图 2.3）。更复杂的是，一个数据集可能也会是一种衍生产品，可以以不同形式来使用：

1. 未经处理的，28%。

图 2.3 另类数据类型

资料来源：Neudata。

2. 结构化的或聚合的，35%。

3. 单一的（派生指标），22%。

4. 报告，15%。

2.3.1 另类数据分类与定义

另类数据的分类与定义见表2.1。

表2.1 数据分类类型

数据集分类	定义
众包	从一大群贡献者处收集的数据，尤其是利用社交媒体或智能手机应用
经济	所收集的数据与某一特定地区的经济相关。例如贸易流、通货膨胀、就业或消费性开支数据
ESG	所收集的数据旨在帮助投资者明确不同企业的环境、社会与治理风险
事件	能够提醒用户对权益价格敏感事件的任何数据集，例如收购公告、催化剂事件日程或交易预警
金融产品	任何与金融产品相关的数据集，例如期权定价、隐含波动性、交易所交易基金（ETF）或结构性产品数据
资金面	任何与机构或散户投资相关的数据集
基本面	源自专有分析技术的数据，与公司基本情况相关
物联网	源自相互关联的实体设备的数据，如无线基础设施以及具有嵌入式网络链接的设备
位置	通常源自移动电话位置数据的数据集
新闻	源自新闻源的数据，包括公开的新闻网站、新闻视频频道或公司特定的资讯供应商
价格	源自场内或场外交易的价格数据
调查与民意测验	利用调查、问卷调查或分组收集的基础数据
卫星及航空	利用卫星、无人机或其他航空设备收集的基础数据
搜索	包含或源自网络搜索数据的数据集
观点	源自自然语言处理、文本分析、音频分析或视频分析手段的输出数据

数据集分类
社交媒体
交易
天气
网页抓取
网页与应用程序追踪

资料来源：Neudata。

2.3.2 有多少另类

我们预测，如今
21%属于网站与应用
观经济数据包括若干
产、经济指标等（见

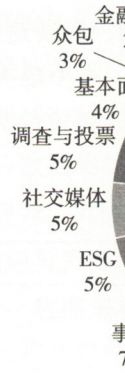


图2.

资料来源：Neudata。

（续表）

数据集分类	定义
社交媒体	利用社交媒体源收集的基础数据
交易	源自收据、银行对账单、信用卡或其他交易源的数据集
天气	源自天气的相关数据，如源自地面站及卫星的数据
网页抓取	从网站中定期收集特定数据的某一自动程序收集的数据
网页与应用程序追踪	数据源自归档现有网站与应用程序，追踪每一网站特定变化的自动程序；监控网站访客行为

资料来源：Neudata。

2.3.2 有多少另类数据集

我们预测，如今买方使用的另类数据源有 1 000 余个。其中 21% 属于网站与应用程序相关数据，8% 为宏观经济数据，这些宏观经济数据包括若干子类，如就业、国内生产总值、通货膨胀、生产、经济指标等（见图 2.4）。

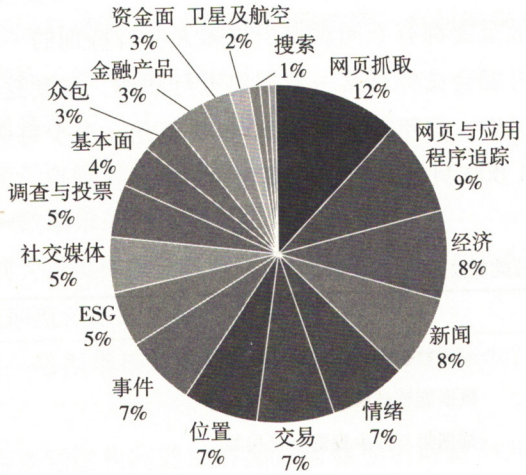


图 2.4 买方使用的另类数据源分类

资料来源：Neudata。

前6个数据种类占据所有数据源的50%。有必要指出，一个数据集可以分为多个种类。一个数据集可以包含多个来源，可适用于不同的使用实例。

然而，在投资管理中利用这些数据源的方式并不统一，也不能反映数据源的供应方情况。

2.4 如何判断哪些另类数据有用

对很多基金经理而言，终极问题在于选择何种数据源用于研究或回测。其中一个关键问题在于，哪种数据集较易操作？需要进行多少数据清理、数据对应及准备工作，从而将一个数据集与一个研究数据库整合？

我们试着回答这些问题的其中一个方式是根据表2.2中的8个因素对每一数据集打分。可以理解的是，每一个基金经理对表2.2中哪种因素最重要都有不同观点。很多人会有特别的“硬性要求”。例如，有人可能会选择对某一数据集进行回测，只要这一数据集有至少5年的历史，每年花费少于50 000美元，至少每天更新一次，并且与至少1 000只公开上市股票相关。

表2.2 评估另类数据有用的关键标准

因素	描述
数据历史长度	时间数据可获得的最早历史点
数据频率	数据能够被推送的频率
覆盖范围	数据集与多少投资公司相关
市场熟知度	Neudata对某一数据集有多少机构投资者熟知进行的评估
拥挤度	Neudata对有多少对冲基金与资产管理客户使用这一数据集的估计

（续表）

因素	描述
独特性	Neudata 对某一特定数据集独特性的评估
数据质量	对数据完整性、结构、准确度与及时性的评估
年度价格	数据供应商收取的年度订阅价格

资料来源：Neudata。

当然，上述这些因素只是一种初步概述，以使机构投资者确定一个数据集与另外一个数据集到底有什么不同。除此之外，有众多质量因素需要考虑，以判断某一数据集是否值得进一步调查。可以进行全面调查，试着回答 80 到 100 个问题，其中包含了我们从投资界最常收到的提问。例如：

1. 数据的原始来源是什么？
2. 数据是如何收集的，又是如何呈现的？
3. 三年前的数据是否如今天一样完整？
4. 数据面板规模大小如何随时间改变，有何偏误？
5. 数据推送的时效性如何？
6. 数据是否在“某一特定时间点”？
7. 数据是否映射到证券标识或代码上，如果有，如何映射？
8. 这一数据集如何区别于类似产品？
9. 目前为止机构投资者对产品的兴趣如何？（如果有兴趣的话）
10. 地理覆盖范围如何？会不会扩大？
11. 这一数据集相关的投资公司具体列表是什么？

我们为找到这些问题的答案，与数据供应商召开了多次会议，审查样本数据（通常与感兴趣的客户共享），并调查了独立的信息来源（如学术论文）。在采取这些措施时，不但创建了一个全面和

独特的数据集概况，也提供了一些参考使用案例，可应用于回测中。

2.5 另类数据需要多少成本

对数据供应商和另类数据购买者而言，最具挑战性的问题之一在于，如何决定一个数据集的价格。

对很多进入金融服务行业的新数据供应商而言，定价是一件非常困难的事，主要出于以下两个原因：第一，在很多情况下，新供应商对同业或类似数据订阅定价的理解和认知非常有限；第二，数据供应商不清楚买方将以何种方式使用他们的数据，也不清楚一个数据集能为资产管理公司带来多大价值或超额收益。在资产管理公司看来，数据集的附加值取决于多个因素，例如投资策略、投资期、投资领域、投资规模等，以及许多对某一特定基金经理的投资策略具有独特意义的其他因子。如果新数据源与已经被某一特定基金经理使用的数据集高度相关的话，新另类数据集的边际超额收益可能会很小。

对那些开始研究另类数据的资产管理公司而言，挑战来自用于数据订阅的预算。基于数据格式（如2.3节所述）、数据质量和其他数据供应商的特质，年度数据订阅价格会有很大差别。另类数据集的价格从免费到年度订阅费250万美元不等。在所有数据集中，约70%的定价区间在每年1～150 000美元。也有若干免费的另类数据集。然而，使用免费的数据源，可能会带来数据检索、数据清理、数据规范化、数据标注及其他准备工作带来的间接成本，只有做了这些准备工作，数据源才能够被基金经理应用于研究和生产（见图2.5）。