Building an open and trustworthy alternative data ecosystem

Alt data report

A report by Refinitiv and the Open Data Institute. Translation by Refinitiv.

February 2020

构建开放和可信的“另类数据”生态系统

A report by Refinitiv and the Open Data Institute. Translation by Refinitiv.

February 2020

A report by Refinitiv and the Open Data Institute. Translation by Refinitiv.		February 2020

Contents 目录

A report by Refinitiv and the Open Data Institute. Translation by Refinitiv. February 2020

Introduction
开篇语

Alternative data, or “alt data”, has been commonly discussed in the offices of hedge funds and other investment firms for years, but has recently seen a spike in popularity and interest in the broader community. Its definition is unclear. Some see it as a new approach to creating investment insights and value from exhaust data, generated as a side effect of other operations and transactions. To others, it is simply a new element of the buzz about big data in investment research.

另类数据（Alternative Data, 或简称为alt data）这个话题，已经在对冲基金和其他类型的投资机构之间广泛讨论了很长一段时间。最近我们看到在更广泛的社区中，它所引起的兴趣和受欢迎程度都有了显著的增长。但到目前为止，’另类数据‘并没有非常明确的定义：一些人认为它是各类活动和交易行为过程中衍生出来的数据’副产品‘，而从这些更广泛而全面的数据中获取价值和投资洞察是一种新兴的手段；对另外一些人来说，它只不过是“大数据加持投资分析”这类”噱头“的又一个新谈资而已。

For many years, analysts have relied on primary research as a method of collecting data that could identify when a stock was potentially mispriced. From using humans counting cars in parking lots of major retailers, to physical inspection of store shelves, these primary research practices first emerged in the coverage of retail stocks. With the rise of online commerce, the proliferation of cellular technology, satellite imagery and the Internet of Things (IoT) it is now possible to track human activity and commerce on a massive scale. This data can now be processed, cleaned and used to create insights that give investors a better understanding of how firms are performing and make decisions based on well-informed predictions. However, the ability to collect and use data on this scale comes with additional risks to privacy, and potential for public fear about the actual or perceived harms from using this data.

多年来，分析师们一直依赖于”一手调研”（Primary Research）所获得的数据来发现一只股票潜在的价格错配。我们常听说的手段包括对大型商家门口的停放车辆进行人工计数，以及实地检验货架上的商品陈列情况等等。这些一手调研的实践始于针对零售行业股票的投资分析，但随着电商的崛起以及移动设备、卫星图像和物联网技术的普及，人们开始有能力对人类的社会和商业活动进行更大规模的追踪和记录。这类收集到的数据可以被处理、清洗和加以利用，从而帮助投资者更好地理解一个公司的实际业绩，并指导人们基于更充分的信息和预期来做出投资决策。然而，如此规模的数据收集能力以及对此类数据的使用，随之而来的是额外的隐私风险；这类数据应用可能造成的实质上或理论上的损害，也会引发公众和相关监管机构的担忧。

Market research estimates that hundreds of investment firms are already using alt data to some extent. In a recent global artificial intelligence/machine learning (AI/ML) survey, conducted by Refinitiv of the top financial institutions using AI/ML today, 70% of firms are using alt data. The U.S. leads the way with 97% adoption. Adoption in Europe (67%) and Asia (53%) is further behind but continues to increase.[1]

据市场调研估算，数以百计的投资机构已经在不同程度上应用了另类数据。路孚特近期发起了一个针对全球范围内应用了人工智能和机器学习技术的顶级金融机构的调查问卷。这份问卷的汇总结果显示，70%的受访机构已经在应用另类数据. 其中美国机构位于全球之首，有着高达97%的采用比例；紧随其后的是欧洲（67%）和亚洲（53%）的机构，而且相关的趋势还在继续上升[1]。

Hundreds of new data providers have entered the alt data space, hoping to build a long-term business selling data to financial services companies, while existing financial information companies have introduced their own offerings, looking to capitalize on the emergence of this new segment of the market.

数以百计的新兴数据提供商已经进入了另类数据的业务领域，他们希望能建立向各类金融服务公司销售数据的长期业务模式；而现有的金融信息提供商也都引入了他们自己的另类数据产品和服务，希望能够从这个细分数据业务市场的发展势头中获利。

With over a thousand purported alt data sources, hundreds of case studies and millions in global spend, the alt data sector seems to be flourishing. Despite its growth however, the market remains a very small part of the multibillion dollar market for access to stock market and other financial data. While the governance and norms for the licensing and distribution of traditional data are well established in many countries, the alt data market is at an early stage. Best practices, codes of practice, regulations and standards have not yet been fully established.

随着数以千计自称为另类数据的数据源不断涌现，随处可见的几百个案例分享，以及全球范围内百万美元级别的相关支出，另类数据市场看上去似乎生气勃勃。但在强劲的增长势头之外，这个领域相对于价值数十亿美元的面向股票及其他类型金融数据的市场来说，还只是很小的一部分。虽然对传统数据的监管以及授权分发机制在很多国家已经成熟并成为常态，但对于另类数据市场来说，这些仍然还处于早期阶段：最佳实践、行为准则，法律法规和相关标准仍然没有完全建立。

Enabling the integration and use of such a disparate set of data sources will require the development of a new set of standards. The alt data industry also needs to adopt codes of practice: ethical and legal frameworks that will build trust in how data is being accessed, used and shared.

另类数据的多样性意味着我们需要新的标准来有效地集成和使用它们。另类数据行业也需要积极采纳道德和法律层次的相关框架和准则，来帮助解决在另类数据获取、使用和分享过程中的信任问题。

These standards and best practices should be developed by the alt data industry in collaboration with regulators. Without taking these steps, users of these new data sources may be exposing themselves to a range of legal risks around the use of personal data, or material nonpublic breach of information. Adoption of these standards will help all alt data industry participants and ensure long-term sustainability of alt data use in financial services.

这些标准和最佳实践应该通过另类数据业界与监管机构的积极合作来进行制定。如果缺少了这些必要的步骤，使用另类数据的用户会被暴露于一系列围绕着使用个人数据或由于其他的重大非公开信息（Material Nonpublic Information）泄露而引发的法律风险。而采用这些标准，能够帮助所有另类数据行业的参与者，确保他们在金融服务业务中长远地、可持续地利用这样的数据。

This report provides a number of recommendations for both alt data providers and users, with the aim of helping to create a more open, trustworthy data ecosystem.

这份报告会对另类数据的提供者和使用者双方提供一系列的建议，以帮助建立一个更开放和可信的数据生态系统。

[1] Refinitiv (April 17, 2019), “Insights from the Refinitiv 2019 Artificial Intelligence/Machine Learning Global Study,” https://refinitiv.com/en/resources/special-report/refinitiv-2019-artificial-intelligence-machine-learning-global-study?utm_source=Refinitivperspective_blog&utm_medium=blog&utm_campaign=107263_AISurveyReport&utm_term=&utm_ content=Reglp&elqCampaignId=6848

About this report
关于这份报告

This report was jointly produced by the Open Data Institute (ODI) and Refinitiv.

这份报告由Open Data Institute (ODI，开放数据学会)和路孚特共同撰写。

Founded in 2012, the ODI is an international, independent and not-for-profit organization based in London, UK. The ODI works with companies and governments to build an open, trustworthy data ecosystem, where people can make better decisions using data and manage any harmful impacts.[2]

成立于2012年，ODI是一家国际性的独立非营利组织，总部位于英国伦敦。ODI与公司和政府机构广泛合作，致力于构建一个开放和可信的数据生态系统，从而帮助人们更好的做出基于数据的决策，以及应对任何可能的负面影响[2]。

Formerly the Financial and Risk business of Thomson Reuters, Refinitiv is a new company built on a unique open platform, high- performance products and best-in-class data. In the face of unparalleled industry change, Refinitiv draws on its deep knowledge and heritage of objectivity to drive performance and innovation with customers and partners.[3]

起源于汤森路透的前金融与风险事业部，路孚特是一家建立于独特的开放平台、高性能产品和业内最好的数据产品之上的全新公司。面对无与伦比的行业变化，路孚特将它深厚的行业知识和有着深刻历史渊源的客观态度，用于同客户以及合作伙伴一起驱动业绩和创新[3]。

This report was written using extensive desk research, user research interviews with 13 professionals in the alternative data space and an expert roundtable of 12 participants. The full list of contributors is in the appendix.

这份报告是基于广泛的桌面研究，针对13个另类数据业内专家的用户调查访问，以及12名参与者的专家圆桌会议撰写的。对本报告做出贡献者的完整名单请参考附录。

What is “alternative data”?
什么是’另类数据’？

Alternative data is a term increasingly being used in the finance and investment sectors. As a relatively new and changing concept, there is currently no definition that is universally agreed upon.

另类数据这个名词越来越多地在金融和投资领域被提起。作为一个相对较新而且不断演化的概念，业界并没有形成具备共识的一个通用定义。

AlternativeData.org defines alternative data as *“data used by investors to evaluate a company or investment that is not within their traditional data sources (financial statements, SEC filings, management presentations, press releases, etc.).”*[4]

AlternativeData.org将另类数据定义为：“*data used by investors to evaluate a company or investment that is not within their traditional data sources (financial statements, SEC filings, management presentations, press releases, etc.).*”[4]， 翻译为中文的话可以理解为：（另类数据是）被投资者用来评估公司或一项投资的数据，而其并不来自公司评估或投资研究专业的传统数据源（例如财务报表，证监会报告，管理层宣讲，以及公司公告等等）。

While some definitions of alt data focus on new data sources, traditional data sets could still be considered alt data if the source or method of analysis was new, as was suggested on a panel at the Financial Management Association Conference in 2018. For example, the use of credit card data to inform investments is now common practice in the industry, but it is still labeled as “alt data.”

虽然一些定义更聚焦于数据源之”新“这个特性，但如果传统数据集的来源和分析方法相对新颖，它们仍然属于另类数据的范畴，在2018年举办的一次国际财务管理协会（Financial Management Association）会议上的嘉宾讨论环节就指出了这一点。例如：使用信用卡的相关数据来支撑投资决策已经是很常见的手段了，但它仍然会被打上“另类数据”分析方法的标签。

J.P. Morgan states that *“the definition of alternative data can also change with time. As a data source becomes widely available, it becomes part of the financial mainstream and is often not deemed alternative.”*[5]

摩根大通（J.P.Morgan）的观点是：“*the definition of alternative data can also change with time. As a data source becomes widely available, it becomes part of the financial mainstream and is often not deemed alternative.*”[5]（另类数据的定义可能会随着时间而变化。随着一个数据源更广泛的应用，它将会融入主流的金融和财务（分析方法），并不再被视为 “另类” ）

“Alternative data” then is largely a sector-specific term that relates to the use of new data sources, combining existing data sets together to create new insights and the application of new analytical techniques. Data sources that provide “traditional data” in the agriculture or retail sectors are considered to be "alt data" in the context of finance and investing. Simply put, alternative data is data that is commonly used and analyzed in a certain domain, put to a different or new use.

那么“另类数据”这个词，常常被与那些将新兴数据源结合于传统的数据集，并应用新的数据分析方法从而获得洞察力的实践联系起来。在农业和零售领域被视为”传统 “的数据一旦和金融与投资结合起来，它们也会被视为”另类数据“。简单来说，另类数据就是那些在某个特定领域经常被用来使用和分析的数据，现在被应用于不同甚至全新的领域。

[2] Open Data Institute (2012), “The Open Data Institute,” theodi.org
[3] Refinitiv (2018), “About Us,” https://refinitiv.com/en/about-us
[4] AlternativeData.org (2018). “Get Started,” https://alternativedata.org/alternative-data/
[5] J.P. Morgan (May 2017), “Big Data and AI Strategies: Machine Learning and Alternative Data Approach to Investing,” https://faculty.sites.uci.edu/pjorion/files/2018/05/JPM- 2017-MachineLearningInvestments.pdf

How is alt data currently being accessed, used and shared?
另类数据被获取、使用和分享的现状是什么？

The current alt data ecosystem is still very fragmented. There are a growing number of alt data providers covering a diverse range
of different data sets. Eagle Alpha, a financial firm specializing in alt data, has created a list of 24 different categories for alt data sets, which is helpful in understanding the breadth of the market.[6] It includes a range of data types including product pricing, employment data, consumer sentiment and sensor data from the IoT. J.P. Morgan has defined nine categories of alt data that are grouped into three broad themes: data from individuals, business processes and sensors.[7]

当前的另类数据生态系统仍然处于非常碎片化的状态。我们能看到越来越多的另类数据提供商覆盖了五花八门的数据集。Eagle Alpha（一个专注于另类数据的金融科技公司）创建了一个包含24个不同类别的另类数据分类表，非常有助于我们理解这个领域的广度[6]。这个列表所涵盖的数据分类包括产品定价数据、就业数据、消费者情绪数据、以及物联网传感器数据等等。摩根大通定义了在三个大类主题之下的9个另类数据分类，这三个主题是：源自个人的数据、源自商业流程的数据，以及源自传感器的数据 (data from individuals, business processes and sensors)[7].

The rest of this section provides some examples of existing uses of alt data for each of these three themes.

随后的章节会列举一些在这三个主题下的另类数据应用案例。

Alt data from individuals
源自个人的另类数据

Alt data from individuals is data generated from online consumer activity that might help inform investment decisions. This category of alt data generally includes social media data (e.g., Twitter®, LinkedIn®, blogs), data from specialized sites (e.g., news media, product reviews) and Web searches and volunteered personal data (e.g. Google® search, email receipts).[8] Users of social media platforms agree to their terms and conditions but few understand the extent to which data about them on these platforms is monetized. Growing awareness of this use could increase the distrust in social media that has already been observed across the UK and Europe.[9] It is not clear whether all of these uses would be legal under the various legislative environments that exist around the world.

源自个人的另类数据是那些消费者在线行为所产生的数据，它们可能有助于指导相关的投资决策。这个类别的另类数据一般包括来自社交媒体的数据（例如Twitter®, LinkedIn®, 博客等），来自特定网站的数据（例如新闻媒体，商品评论）以及来自网络搜索和用户自主产生的个人数据（例如谷歌®搜索，电子邮件发票等等）[8]。一般而言，社交媒体平台的用户会接受各项平台使用条款。但很少有人了解，他们的哪些个人数据会被这些平台用于获利。随着这类利用个人数据获利的故事不断为公众所知，人们对社交媒体的不信任和不安全感也会随之增加，就像我们在英国和欧洲已经看到的那样[9]。而考虑到世界范围内法制环境的多样性，现在很难回答关于这类个人数据的使用是否都是合法的。

Social media can provide a wealth of insights into consumer behavior. An example of an alt data provider in this space is the social media API aggregation company Gnip, which was purchased by Twitter in 2014.[10] Gnip now provides an enterprise API for Twitter that offers a single interface to access Twitter, Facebook®, YouTube, Flickr, Google Buzz, Vimeo and more.

人们从社交媒体可以提炼出非常丰富的消费者行为信息。社交媒体API聚合公司Gnip，就是这个领域另类数据厂商中的一个例子。Gnip在2014年10月被Twitter收购[10]，现在它为Twitter提供了一个企业级API，能够通过一个单一接口来访问众多社交媒体网站的数据包括Twitter, Facebook®, Youtube, Flickr, GoogleBuzz, Vimeo等。

Companies like iSentium collect real time sentiment data from Twitter. By analyzing positive or negative social media posts, investors aim to judge company performance before any traditional data is released from the company. From this, J.P. Morgan constructed an index on the S&P 500 called the JPUSISEN Index based on information obtained from iSentium. Through historical analysis, J.P. Morgan claims that the JPUSISEN Index would have provided a 13.7% return annually since 2013, 1.6% more than the S&P return of 12.1%.[11]

一些公司例如iSentium会从Twitter获取实时的社交媒体情绪数据，并衍生出相关的另类数据产品和服务。投资者们试图通过分析社交媒体相关帖子中传达的积极或消极信号，做到在一个公司发布传统的业绩相关数据之前就能评判它的表现。通过iSentium公司的信息产品，摩根大通针对标普500指数包含的公司构建了一个名为JPUSISEN的指数。通过历史回测，摩根大通宣称这个指数自2013年算起可以提供13.7%的年化回报率，相对于同时期标准普尔指数12.1%的年化回报率增长了1.6个百分点[11]。

From a legal and ethical standpoint, given the prevalence of personal data in this category, social media is one of the more risky types of alt data to work with. Without proper privacy and security practices in place, issues around re-identification and identity theft will exist. This topic will be explored later in the paper.

从法律和伦理的角度，由于大量私人数据的存在，从社交媒体所抽取的数据对使用者来说是风险最高的一种另类数据类型。如果没有合理的隐私和安全措施，用户会被暴露于各类隐私风险之下，比如re-identification（一种基于用户的一组脱敏后信息推演出用户真实身份的过程）和identity theft（身份盗用）。我们会在此报告的后面继续展开讨论这个话题。

Alt data from business processes
源自商业流程的另类数据

Alt data from business processes is data generated from the typical activities of organizations that is not directly related to investing behavior. This category of alt data generally includes *"data made available by public agencies (e.g., federal and state governments), commercial transactions (including e-commerce and credit card spending, exchange transaction data), and data from other private agencies (e.g., industry specific supply chain data)."*[12]

源自商业流程中的另类数据，是指企业日常经营性活动所产生的数据，其并不直接与投资行为相关。这个类别的另类数据一般包括"企业向公众机构（例如联邦和各级政府）公布的数据，商业交易产生的数据（包括电子商务、信用卡消费以及外汇交易），以及源自其他私人机构的数据等等（例如特定行业的供应链数据）[12]"。

Customer transaction data can be a very strong predictor of company performance. Companies involved in payments, such as credit card network companies or point-of-sale terminal companies, can provide this data to investment firms.

客户的交易和消费数据则更是与企业的业绩预期表现紧密相关。那些涉及支付服务的公司如信用卡网络公司或销售终端厂商，都可以为投资机构提供这样的数据。

[6] Eagle Alpha (April 2018), “Alternative Data Use Cases Edition 6,”
https://s3-eu-west-1.amazonaws.com/ea-pdf-items/Alternative+Data+Use+Cases_Edition6.pdf

[7] J.P. Morgan (May 2017)

[8] J.P. Morgan (May 2017)

[9] Open Data Institute (Jul 5, 2018), “Who do we trust with personal data?,”
https://theodi.org/article/who-do-we-trust-with-personal-data-odi-commissioned-survey-reveals-most- and-least-trusted-sectors-across-europe/

[10] Thomson Reuters (April 15, 2014), “Twitter buys social data provider Gnip, stock soars,”
https://www.reuters.com/article/us-twitter-gnip/twitter-buys-social-data-provider-gnip-stock-soars-idUSBREA3E17D20140415

[11] Kolanovic, Marko and Rajesh T. Krishnamachari - J.P. Morgan (May 18, 2017), “Big Data and AI Strategies: Machine Learning and Alternative Data Approach to Investing,”
https://www.ravenpack.com/research/jp-morgan-big-data-ai-machine-learning-alternative-data/

[12] J.P. Morgan (May 2017)

Eagle Alpha have created a model called RevCast which incorporates alternative and traditional data to make forecasts on company performance, including consumer transaction data, online search data and historical financials. By combining these data sets, the RevCast model forecasted the car rental company Hertz had second quarter 2018 revenues of USD 2.45 billion, differing from the market consensus estimate of USD 2.3 billion. A month later, Hertz reported revenues of USD 2.4 billion, a figure that was twice as close to the predicted number that the market consensus produced.[13]

Eagle Alpha 曾经创建了一个名为RevCast的模型，它同时使用传统数据和另类数据包括消费者交易数据、在线搜索数据和历史财务数据来预测企业的业绩表现。他们的RevCast模型曾经通过这些数据预测汽车租赁公司Hertz（赫兹）的2018年第二季度收入为24.5亿美元，与当时的市场共识预测的23亿美元不同。一个月后，赫兹公司报告了实际的营收为24亿美元，相对于实际营收的预测准确程度是市场共识预测的2倍左右[13]。

Business process data, such as consumer transactions, may contain personal data. This means that appropriate privacy and security measures are needed to prevent harm and steps taken to ensure it is used in a trustworthy way.

商业流程数据，比如前面提到的消费者交易数据，也可能会包含个人信息。这意味着恰当的隐私和安全措施对于防止消费者利益受损而言是非常重要的步骤，从而确保此类数据只能以一种可被信任的方式进行使用。

Another major issue regarding business process data is the debate around what is considered to be public information and the social, legal and ethical issues of repurposing data that has been shared for other reasons. Scraping data from public websites, observing store and factory activity and the use of employee information are all examples of where this tension exists.

另一类围绕商业流程数据收集和使用的争论也值得关注，就是关于如何界定公开与私有信息，以及对于那些偏离了被分享数据原本的目的而进行的数据利用来说，如何看待其所衍生出的社会、法律和伦理问题。一些行为包括从公开网站抓取信息，观察商家和工厂的活动，以及对雇员信息的使用等，都是这类争论经常出现的场合。

Companies like Matchdeck use publicly shared personal data to create products based on analyzing online trends in organizations, such as employee growth and executive turnover and correlating with business performance. Individuals are able to have their profiles removed from these products, and Matchdeck does not crawl password-protected sites, but there is still a lack of understanding in the market regarding these practices. The current position by much of the legal community in the U.S. is that publicly viewable information can be used as alt data, but there are still pending court decisions which may impact the sector.

一些公司例如Matchdeck会将公开可得的个人相关数据用于创建数据产品，即通过分析线上搜集到的企业各类趋势数据，包括雇员增长，管理层变更等，来关联和预测企业的业绩表现。个人可以要求将他们的信息从这些产品中删除，而且Matchdeck并不会抓取密码保护的网站，但市场对于这类数据搜集行为仍然缺乏全面的理解。当前美国法律界对此的共识是，任何公众能浏览的公开信息可以被用于另类数据相关的处理和分析，但仍然存在一些悬而未决的法庭判决，可能会影响到这个领域。

Alt data from sensors and satellite imagery

源自传感器和卫星图像的数据

Alt data in this category includes data generated from smartphones, portable electronic devices, satellite data, geolocation data as well as data from other sensors and “smart” or networked devices.

这一类别的另类数据包括产生于智能手机、便携式电子设备、卫星、地理勘测以及出自于其他类别的传感器或物联网等设备的数据。

The use of satellite and other types of geospatial data for investing is a growing trend in finance. The ODI captured several use cases in a recent paper on geospatial data infrastructure, such as using aggregating activity data via apps in smartphones and GPS devices and matching the points against a street to indicate footfall.[14] The increasing accuracy and timeliness of satellite data collected by commercial organizations also provides the ability to confirm real time data such crop harvests, port usage, parking lot occupancy and factory production. Previously, this type of data was collected by having people physically observe and report on activity in ports, factories and stores, but machine learning is allowing insights to be extracted from satellite images.

将卫星或其他类型的空间地理数据应用于投资分析，在金融领域是一个新的趋势。ODI在最近的一份关于空间地理数据基础设施的报告中，提了若干个应用案例，比如聚合智能手机和GPS设备中记录的活动信息，并将这些聚合后的位置点信息与街道进行匹配，用来识别足球活动[14]。由专门的商业机构所收集的卫星数据所具备的不断增长的准确性和及时性，已经可以用于确认一些关于庄稼收成，港口使用情况，停车场占用率，以及工厂生产活动的实时数据。在此之前，这些类别的数据还是只能通过人工来进行身临其地的观察和汇报，但现在机器学习已经可以从卫星图像上识别和提取相关的洞见了。

The company SpaceKnow has a product that uses satellite image processing technology originally intended for agricultural use to monitor the level of Chinese manufacturing. Through algorithmic imagery analysis, SpaceKnow created the China Satellite Manufacturing Index (SMI) to compete with the current state-run indices – the China Purchasing Managers Index (PMI) and the Caixin PMI. The two PMIs are created by using survey data from managers, collected by the National Bureau of Statistics of China. SpaceKnow uses satellite imagery data sets consisting of over two billion observation points to create an index that they claim more accurately predicts Chinese manufacturing for investors. Analysis of historical data over a 10-year period shows a very strong correlation to both PMIs.[15]

一家名为SpaceKnow的公司提供一种另类数据产品，利用原本应用于农业领域的卫星图像处理技术来监控中国制造业的产量水平。通过算法支撑的图像分析，SpaceKnow设立了“中国卫星制造业指数”（China Satellite Manufacturing Index , SMI），用于和政府编制的中国采购经理指数（China Purchasing Managers Index，PMI）和财新采购经理指数（Caixin PMI）竞争。这两个采购经理指数是通过由中国国家统计局收集的，面向采购经理的调查数据来进行编制的。SpaceKnow使用超过20亿个通过卫星图像数据收集的观察要素创建了这个指数，并宣称能更精确地向投资者预测中国制造业的状况。基于过去10年的历史数据回测表明，这个指数跟前面提到的两个PMI指数都表现出了很强的相关性[15]。

Sensor data collected from personal devices might be useful for measuring footfall in cities and retail districts, but as with other sensitive personal data, it presents significant privacy challenges. Similar to the scraping of data from websites, the ability to remotely observe commercial facilities by satellites raises ethical questions around what is public information and what might be considered trespassing.

从个人设备中采集的传感器数据可能对于衡量城市和零售商区的足球活动程度非常有用，但和其他类别的敏感个人数据一样，它本身带来了很严重的隐私相关的挑战。与从网站抓取数据类似，通过卫星远程观测商业设施也同样带来了伦理方面的一系列问题，尤其是围绕着什么是公开信息，以及什么样的信息的获取和利用会被视为非法行为。

The alt data use cases above showcase the variety of ways data can be sourced and used to provide financial insights. A common feature among the examples is the existence of ethical, and possibly legal, challenges. Organizations in the alt data market need to pay attention to the ethical use of data in their investing decisions to avoid causing societal harm and avoid legal, financial and reputational risks.

以上这些另类数据的应用场景，体现了另类数据的各种来源以及从中洞察财务和投资机会的不同手段。在这些例子中一个共同的特点，就是它们都存在伦理和潜在的法律风险与挑战。投身于另类数据市场上的企业应当注意在他们的投资决策中注意数据使用的伦理问题，从而避免导致可能的社会损害，以及避免相关的法律，财务和声誉风险。

[13] Eagle Alpha (April 2018), “Alternative Data Use Cases Edition 6,” https://s3-eu-west-1.amazonaws.com/ea-pdf-items/Alternative+Data+Use+Cases_Edition6.pdf

[14] Open Data Institute (Feb 2019), “Using geospatial data: a guide to licenses,” https://docs.google.com/document/d/1N_y0Zhc583T8YJ4k2XnhJZpFwnncDIDkUHP8gV3myes/edit#heading=h.4my9b7ojjmma

[15] SpaceKnow (Jan 2018), “China Satellite Manufacturing Index,” https://www.spaceknow.com/case-studies/china/SpaceKnow-China-Satellite-Manufacturing-Index.pdf

Ethical and legal considerations in alt data

另类数据中的伦理与法律考量

Faced with public criticism and debate around the use of data, the individual practitioners and organizations involved in collecting, sharing and working with data are exploring the ethics of their practices. Data ethics is a branch of ethics that evaluates data practices with the potential to adversely impact not only individuals, but businesses and wider society too.[16] While the debate around data ethics often centers on the impacts of the use of data about individuals, it also applies to our growing ability to remotely monitor business operations.

面对公众对于数据使用的各类批评和争论，从事各类数据收集、分发和使用的机构和个人已经开始审视他们在实践中所涉及的数据伦理问题。数据伦理方面的考量涉及将数据相关实践中可能对个人、企业甚至社会产生的不利影响进行评估和应对。这方面的担忧和争论通常聚焦在使用个人数据所造成的影响，但它同样适用于业界不断增长的，远程监控企业运营的技术能力上。

Organizations are using emerging tools, such as the ODI's Data Ethics Canvas, to help identify potential ethical issues associated with a data project or activity. The canvas promotes understanding and debate around the foundation, intention and potential impact of any piece of work, and helps identify the steps needed to act ethically.[17]

各类企业已经开始利用一些新出现的工具，例如ODI的"数据伦理画板"（Data Ethics Canvas），来帮助识别与他们的数据项目和活动相关的潜在数据伦理问题。这个工具提倡对任何数据相关工作从本质，意图和潜在的影响几个角度进行深入理解和推敲，来帮助企业采取必要的步骤来以符合伦理的方式行事[17]。

Trust is equally essential for any organization and for the alt data sector as a whole. When trust in organizations breaks down, they may suffer reputational damage which can lead to loss of business. When trust in a sector as a whole decreases, there is a danger that it cannot realize the full benefits that innovative use of data could bring. Data might not be collected, shared or used to the extent it could be because of concerns it may be misused. Individuals withdrawing consent could lead to data that is biased and misleading. Countries could introduce regulation that limits data collection or use, or makes it significantly more burdensome, presenting challenges to the existing data market.[18]

"信任"是基础，这对于任何企业和整个另类数据行业来说都是这样。如果关于企业的信任被打破，它的声誉会受到破坏，业务也必然受到负面影响；当整个行业的信任被打破，危险之处在于它将无法实现创新型数据应用本可以给这个行业带来的蓬勃发展。对数据滥用的顾虑可能会导致数据无法在应有的程度上进行收集、分发和使用；个人对数据共享的保守和抗拒可能会导致相关数据的偏差和可信度降低；各国政府可能会引入更严格的法规来限制数据的收集与使用，或令数据相关的各类成本和负担大幅增加，从而给整个数据市场带来冲击和挑战。

There are three areas where we feel that the alt data ecosystem needs to focus attention to create an open, trustworthy data ecosystem that benefits everyone: privacy, rights and fairness.

我们认为另类数据行业需要集中关注隐私、权利和公平这三个维度，来创造一个开放、可信的数据生态系统，并惠及生态系统中的所有参与者。

Privacy and personal data

隐私与个人数据

One of the most controversial aspects of the use of alt data in investing revolves around the collection and use, intentionally or otherwise, of personal data. Even if not damaging to the people the data is about, breach of data privacy laws can be financially destructive to firms. In the European Union (EU), the new General Data Protection Regulation (GDPR) legislation stipulates that non- compliant businesses can be fined up to €20 million or up to 4% of their annual worldwide turnover of the preceding financial year, whichever is greater.[19]

在辅助投资决策的过程中，对个人数据的收集和使用（无论是不是有意为之）是最具争议的。即使在这个过程中对数据涉及的个人并没有实际的利益损害，单单违反数据隐私法律也可以让企业付出严重的财务代价例如高额罚款。根据欧盟新版的通用数据保护法案（GDPR）,违规的企业可以被处以高达2000万欧元，或上一财年企业全球营收4%的罚款（以两项之中较高者为准）[19]。

Other activity, although legal, can cause reputational damage through negative press coverage and upset people. Much of this harm comes from the difference between how people expect data will be used and the reality of data sharing between businesses. People may be happy with personal data being used for benefits to society, as noted in a recent poll,[20] but may not want that data being used to assist hedge funds' investment decisions.[21] Being open and transparent when using, accessing and sharing data helps avoid the backlash that comes when unexpected data use gets revealed. In addition to this, assessing the ethical use involves engagement with those who might be affected by the use of data.

关于其他类型的数据收集和使用，即使是合法的，仍然有可能因为个人的不满或者媒体的负面报道导致企业的声誉损失。很多这样的损害都源自个人对自身数据合理使用的期待，与企业实际操作中数据分享的范围不相匹配而造成的。人们一般会乐于将个人信息相关的数据用于社会公益 （例如像近期的一个调查所指出的那样），但通常不想自己的数据被用来帮助对冲基金这样的机构进行投资获利相关的决策[21]。在获取、使用和分享数据的时候保持开放和透明，可以帮助避免意料之外的数据使用被曝光后所带来的冲击。另外，评估数据使用的伦理，也涉及到与可能会被数据使用活动所影响的个人进行相关的接触。

With scandals such as Facebook and Cambridge Analytica[22] and the implementation of the GDPR and the California Consumer Privacy Act (CCPA) of 2018, personal data came to the forefront of public and media discussion. Though most of the press regarding alt data
is not negative, there has been growing scrutiny of the risks, with groups such as Big Brother Watch, Privacy International and Open Rights Group identifying serious privacy and human rights risks.[23]

随着Facebook和Cambridge Analytica[22]所涉及的那类丑闻，以及欧盟通用数据保护法案（GDPR）和2018年加州消费者隐私法案（CCPA）的实施，"个人数据"这个话题逐渐出现在媒体和公众讨论的前台。尽管大多数关于另类数据的媒体报道并不是负面的，但对于相关风险的审视与监管正在日渐增强，例如各类组织如Big Brother Watch， Pricacy International 和Open Rights Group对严重隐私与人权相关风险所进行的识别与分析[23]。

These concerns are not unfounded. Sentiment data from social media, credit card transactions from retail stores and geolocation data from sensors can all be used to identify people and it can scare people to think they are being watched. Personal data needs to be properly anonymized, based on a re-identification risk assessment, when shared with or used by alt data providers. If we are to retain trust, it must be clear to people what is happening to data about them, and both regulators and civil society need to be able to scrutinize the claims.

这类担忧并不是毫无根据的。社交媒体的舆情数据，零售店的信用卡交易数据以及传感器物联网所提供的地理信息数据都可以被用于识别到具体的个人，这些令公众时刻担心自己处于监控之中。正因为如此，个人信息相关的数据在被另类数据提供商所使用或分发之前，需要基于对re-identification（一种基于用户的一组脱敏信息推演出用户真实身份的过程）的风险进行评估，并进行必要的匿名化和脱敏处理。如果我们想重拾信任，我们一定要向相关数据涉及到的个人非常清晰的说明，关于他们的数据将会被怎样处理；我们也需要确保监管机构和民间组织能够对这样的声明进行有效的审查。

[16] Open Data Institute (2018), "Data Ethics," https://theodi.org/service/data-ethics/
[17] Open Data Institute (Aug 5, 2017), "The Data Ethics Canvas," https://theodi.org/article/data-ethics-canvas/

[18] Open Data Institute (2018), “Our theory of change,” https://theodi.org/about-the-odi/our-vision-and-manifesto/our-theory-of-change/
[19] European Union (May 2018), “The EU General Data Protection Regulation (GDPR),” https://eur-lex.europa.eu/TodayOJ/index.html
[20] Open Data Institute (Feb 12, 2018), “ODI survey reveals British consumer attitudes to sharing personal data,” https://theodi.org/article/odi-survey-reveals-british-consumer-attitudes-to-sharing-personal-data/
[21] Integrity Research (Sept 12, 2017), “Privacy Watchdogs Worry About Hedge Fund Use of Geolocation Data,” http://www.integrity-research.com/privacy-watchdogs-worry-hedge-fund-use-geolocation-data/
[22] Facebook-Cambridge Analytica scandal (2019) https://www.bbc.co.uk/news/topics/c81zyn0888lt/facebook-cambridge-analytica-scandal
[23] Integrity Research (Jan 2018), “Mitigating Legal Risks Associated With Alternative Data,”

https://www.integrity-research.com/wp-content/uploads/2018/01/Mitigating-Legal-Risks-Alternative-Data-January-2018-2.

A high-profile example of this potential abuse of alt data occured in January 2019, when a scandal concerning the use of people's location data occurred between IBM's The Weather Channel and the City of Los Angeles. At the time of this writing, the municipal government is suing IBM for using location data gathered by The Weather Channel mobile application for commercial purposes despite telling users it was only for localized weather services.[24] IBM have denied this claim.[25]

2019年的一起有着较高曝光度的案例，是发生于IBM与洛杉矶市政厅之间的，关于潜在的个人位置数据滥用的丑闻。在我们撰写这篇白皮书的同时，洛杉矶市政厅正在起诉IBM，指控他们通过名为The Weather Channel的移动应用程序来获取个人位置数据，并在告知用户这些数据只会被用于定制的天气信息服务的情况下，仍然将这些数据用于了商业目的[24]。IBM否认了这项指控[25]。

For both commercial and compliance reasons, users of alt data do not want to access personal data. They want to use aggregated data that may provide insights into overall market trends.[26] Often, to unlock value from the source data sets, alt data providers and users need to combine data sets from multiple sources. This creates potential linkage risks, including the re-identification of individuals, despite efforts to anonymize the data sets. Currently, alt data providers are selling many data sets that can easily be re-identified, such as geolocation data with only the name removed. Using a mixture of suppression, generalization and disruption could improve the anonymization in data sets in the market. It is also important for data providers to understand that anonymization is an in-depth process, involving research, legal and ethical considerations, risk analysis and testing.[27] Adoption of standards and best practices to improve anonymization, such as those found in the UK Anonymization Network's Anonymization Decision-Making Framework, or the ODI guide to Anonymization and open data, and to ensure consideration of the ethical impacts of using personal data, will be important as the alt data ecosystem matures.

出于商业和合规的原因，另类数据的使用者并不想访问个人信息相关数据。他们仅仅需要使用聚合过的、足够为总体市场趋势提供洞察的数据[26]。但为了从获取的数据集中尽可能地发掘价值，另类数据提供者和使用者需要结合多个来源的数据集来达到这一目的。在这个过程中会产生潜在的关联风险，这就包括对个人身份的身份重建（re-identification），即使这些数据集已经被尝试进行了脱敏和匿名化处理。当前在另类数据提供商所出售的很多数据集例如仅仅去除了个人姓名的地理位置数据，都可以被相对容易的用于re-identification，从而定位到具体的个人身份。而使用一系列技术的组合包括隐匿（suppression)、泛化（generalization）和扰乱（disruption）可以提高数据集在市场应用中的匿名化水平。对于数据提供者来说，非常重要的是他们需要了解匿名化是一个需要深化的流程，这包括相关的研究、法律和伦理考量、还有风险分析以及相关的测试。采用有关标准和最佳实践来进行匿名化处理，例如英国匿名化组织（UK Anonymization Network）的Anoymization Decision-Making Framework (可译为”匿名化处理的决策框架“），或者开放数据学会（ODI）关于Anonymization and open data（可译为匿名化与开放数据），并确保使用个人信息相关数据前进行数据伦理相关的考虑，对于另类数据生态系统走向成熟来说非常重要。

Data rights and licensing
数据权利与许可

A significant legal concern regarding alt data is the right to access, use and share the data. A robust data ecosystem needs clear data rights so that companies can operate without high legal or operational risks, while minimizing harm to society. Alt data providers need to be clear about their rights to collect and repackage data. Alt data users need to be clear about the rights that govern the data supplied to them and what they can legally do with it. They also need certainty that their supply of alt data will not be interrupted, for example due to legal action against their alt data provider.

数据访问、使用和分发的权利，是一个关于另类数据的重要法律相关考量。一个健壮的数据生态系统需要有能力描述清晰而明确的数据权利，来确保企业可以在没有高企的法律和运营风险的情况下进行运作，并最小化对社会利益任何可能的损害。另类数据提供商需要明确了解他们对所要收集和打包的数据拥有哪些权利；另类数据的消费者也需要明确了解他们所获得的数据由哪些权利事项所管控，以及他们可以对这些数据进行哪些合法的用途。消费者还需要相关的保障，来确保他们建立的另类数据供应链不会被轻易干扰，例如那些对他们的数据提供商所可能采取的的法律行动等。

Web scraping, a process of extracting data from websites, is a common methodology for collecting alt data from public websites. For example, companies crawl e-commerce websites to compile data on their current inventory, pricing and product reviews. This data is then analyzed through a process called "text and data mining," in order to "discover patterns, trends and other useful information that cannot be detected through usual 'human' reading."[28]

互联网数据抓取，是一类常见的从公开网站收集另类数据的方法。比如，有的公司利用爬虫从电子商务网站抓取相关数据，并编制关于这些网站的当前库存，价格和产品评论的相关数据。这些数据会通过”文本与数据挖掘“的方法来发现模式、趋势以及其他类型的有用信息，特别是那些很难通过人工阅读获取的信息[28]。

A lack of clear licensing or legal basis for using this data is causing concern in the alt data community, specifically in the degree to which this public information can be reused. Many in the industry believe that data collected from public

websites is public data and can be collected and reused for any purpose.[29] Others believe that even data protected by a username and password, can be used if it is amenable to Web scraping.[30]
另类数据社区中对于这种缺乏明确使用许可以及法律依据的数据使用方式，存在着普遍的担忧，特别是有关在何种程度上这类公开信息可以被用于类似的目的。很多业内人士认为从公开网站上收集的数据属于公开数据，可以被收集和应用于任何目的。另一些人认为即使公开网站的数据被用户名和密码所保护，但抓取的数据仍然可以使用（这种观点认为用户名和密码本身并不与数据的使用权限直接关联，除非另有说明）[30]。
The lawsuit at the forefront of this debate is *LinkedIn vs hiQ Labs*. hiQ Labs is a data science company that uses scraped data from public LinkedIn profiles in order to develop tools to help corporate HR departments keep tabs on their workforces.[31] LinkedIn have sued hiQ for breaching their terms of use (ToU). Although from a ToU-perspective there is a breach of terms, it is unclear if this is a breach of the law, as the Web-scraped data is currently considered public information in the United States.[32] So far, the lower courts have supported hiQ and issued an injunction ordering LinkedIn to grant access, however the case is now with an appellate court.[33] The outcome of this case could set a strong precedent for Web scraping businesses in that jurisdiction that could influence the wider market.
在这类争论中比较突出的法律诉讼发生在领英（LinkedIn）和hiQ Labs之间。hiQ Labs是一家数据科学公司，它从领英的公开会员信息中抓取数据，并用这些数据开发了服务于企业人力资源部门的工具来监控他们的员工。领英公司以违反了领英使用条款为由起诉了hiQ。尽管从使用条款的角度，hiQ的做法确实违反了规定，但这样的做法是否违反了法律尚不清楚。因为在美国，通过网络爬虫获得的数据目前还是被认为定为公开数据[32]。到目前为止，一审法院支持hiQ，并且颁布了法庭命令要求领英赋予hiQ必要的访问权限，然而目前这个案子已经移交给了上诉法庭。这个案子的最终判决结果将会在当地法律管辖的范围内，给整个网络数据抓取产业形成一个强有力的先例，并对更广泛的市场产生影响。

Legal issues aside, companies whose websites are being scraped may attempt to stop this by tightening up their terms and conditions, asking companies to cease scraping operations or by applying more sophisticated technical measures – by detecting and blocking the software that is scraping Web pages or deliberately serving them incorrect information, for example. This lack of trust inevitably leads to an arms race between scrapers and websites.
在法律问题之外，那些被数据公司频繁抓取网站数据的企业也会试图阻止这样的行为。他们可能采取的手段包括收紧关于网站的使用条款，要求其他公司停止数据抓取行为，或者应用更复杂的技术手段检测并封堵网页抓取软件的操作，甚至刻意地给此类软件提供错误的信息。信任的缺失将会无可避免地引发数据抓取者和网站之间的军备竞赛。
In a landscape where both legal and ethical frameworks are evolving, there are various ways to mitigate problems and build trust.
在法律和伦理框架都在不断演化的这样一个领域，仍然存在着多种缓解问题和建立信任的方法与手段。

[24] New York Times (Jan 3, 2019), “Los Angeles Accuses Weather Channel App of Covertly Mining User Data,” https://www.nytimes.com/2019/01/03/technology/weather- channel-app-lawsuit.html
[25] Insurance Journal (Jan 7, 2019), “Los Angeles Sues IBM’s Weather Channel for Use of Location Tracking,” https://www.insurancejournal.com/news/national/2019/01/07/514074.htm
[26] Interview with Integrity Research (Feb 20, 2017), Peter Greene (Mar 13, 2019)
[27] The Open Data Institute (April 2019), “Anonymisation and open data: An introduction to managing the risk of re-identification,” https://docs.google.com/document/d/1CoXniaTnQL_4ZyQuji9_MA_YCEElQjx4z1SEdB08c2M/edit#
[28] CopyrightUser (May 18, 2017), “Text & Data Mining,” https://www.copyrightuser.org/understand/exceptions/text-data-mining/
[29] AltData.TV (Sept 20, 2018), “Peter D. Greene on web crawling,” https://altdata.tv/2018/09/20/web-crawling/
[30] Integrity Research (Feb 20, 2019)
[31] hiQ Labs (2018), “Who we are,” https://www.hiqlabs.com/new-who-we-are
[32] Lexology (Dec 3, 2018), “Data Scraping: Theft or Fair Game?,” https://www.lexology.com/library/detail.aspx?g=5e951f2d-55c7-42a3-a539-fbe88165ea5a
[33] Electronic Privacy Information Center (2018), “hiQ Labs, Inc. v. LinkedIn Corp.,” https://epic.org/amicus/cfaa/linkedin/

Proactive publishing of machine-readable, open and aggregate data by organizations will avoid the need for data to be scraped at all, reducing issues with the accuracy and robustness of scraped data and providing more clarity around reuse rights. Data licensing agreements can also help organizations to share data with clear reuse rights where data cannot be published more openly. The Standards Board for Alternative Investments (SBAI) data licensing agreement[34] is one such agreement that exists specifically for alt data.

由企业来主动发布机器可读，开放并且聚合过的数据，可以避免数据被刻意抓取的必要性，从而减少通过抓取而来的数据所存在的的准确性与健壮性相关的一系列问题，并在数据重用许可上提供更大的透明性。“数据许可协议”可以帮助企业在提供了明确的数据使用权利和约束说明的情况下，分享那些无法以完全公开的形式来进行发布的数据。SBAI（The Standards Board for Alternative Investments，另类投资准则委员会）的数据权限协议就是这样的一种特别针对于另类数据的许可协议。

Web scraping companies can be open with sites that are being scraped, by engaging directly with the organizations. For example, maintaining a mail trail showing that websites have been alerted to data being collected and lack of subsequent objections may be useful in a court case. Halting Web scraping, and providing easy ways for that to be requested, can also help to avoid legal issues. Negotiating ongoing access to data – after prototyping data collection using Web scraping, for example – will also help to ensure ongoing access.

网页数据抓取公司也可以通过主动接洽那些被抓取网站所属的企业，来表明自己开放和透明的态度。例如，如果保留了与网站所属企业的相关邮件往来记录，能够证明自己曾经告知了对方关于数据抓取的情况，同时并没有收到后续的反对意见的话，可以令自己在可能的法庭诉讼中占据主动。有能力暂停网页数据抓取活动，并为相关企业提供简单易行的方法来要求采取这样的行动，也能够帮助避免相关的法律问题。另外，如果能够和企业洽谈对数据抓取的许可，例如在进行了相关的数据的抓取试验和演示之后，将会有助于保持自身对相关数据的持续访问。

Organizations like Eagle Alpha are attempting to define best practices for Web scraping to curtail negative behavior, for example by ensuring Web scraping does not harm the business of the website through direct costs from increased traffic load or direct competition.[35] Alt data providers relying on Web scraping are also aligning their approaches to those adopted by the larger search engines such as Google, with a view that these are more accepted behaviors and that there is legal safety in numbers.

像Eagle Alpha这样的公司正在尝试通过定义网页数据抓取的最佳实践来减少具有负面影响的行为，比如确保网页数据抓取活动不会给网站带来了额外的流量相关支出，以及抓取的数据不会被用于和网站进行直接竞争，从而避免给网站的正常业务带来负面影响。依赖于网页抓取的另类数据提供商们也正在尽量让自己的手段与大型搜索引擎例如Google保持一致，寄望于它们的方法是更广为接受和检验过的，并希望自己与业界的大多数厂商站在一起。

Clearly describing the provenance of data will be important in building trust across the alt data ecosystem. The variety of ways in which data is sourced, and the use of multiple data sets to create alt data products, makes this difficult to achieve. Without clarity around how data has been collected, organizations cannot be sure that their use is compliant or understand the risk of an interruption to supply, which increases their operational risks.

明确地描述数据溯源对于在另类数据生态系统中建立信任非常重要。数据来源的多样性，以及另类数据产品生产过程中对多个数据集的依赖，令数据溯源的记录非常困难。企业在不知道数据是如何被收集的情况下，无法确定他们对数据的使用是否合规，也很难弄明白在哪些情况下他们的数据供应链可能会中断，从而增加他们的运营风险。

Organizations in the alt data market need to collaborate to discuss potential risks and work to increase effort around provenance and rights. This will help to build confidence and trust with consumers of the data and reduce risks of additional lawsuits and potentially negative media coverage. Overall, it will help to build a stronger, more sustainable and more trustworthy data ecosystem.

我们认为投身于另类数据市场的企业应该进行合作，来讨论潜在的风险以及必要的工作，并增加关于数据溯源和数据权利的投入。这有助于建立另类数据消费者的信心与信任，并减少额外的法律诉讼与潜在的负面媒体报道的风险。总体来说，这有助于建设一个更强大，可持续发展以及更可信的数据生态系统。

Fairness and material nonpublic information
市场公平与重大非公开信息

A strong data ecosystem provides equitable access to data. Access to data and information promotes fair competition and informed markets and empowers people as consumers, creators and citizens.[36]

一个健全的数据生态系统应能对各方提供平等的数据获取能力。对数据和信息的获取和使用可以促进公平竞争与市场信息透明化，并赋能于市场上的消费者，创造者和普通市民[36]。

The legal consideration about the use of alt data that most concerns hedge fund managers involves insider trading. Even the accusation or investigation into insider trading can financially harm a company. Investing activity is subject to insider trading laws that stipulate that any information used for commercial gain must be publicly available to ensure fairness and competitiveness in the market. Data is material nonpublic information (MNPI) when it earns returns, is clearly not licensed for use and is acquired exclusively. It is illegal for holders of MNPI to use it to their advantage in investing, or doing similarly for others.[37]

对基金经理们来说，'内幕交易'是使用另类数据时最令他们所担心的法律问题。仅仅是关于内幕交易的指控或调查都会对一个公司产生严重的经济影响。投资活动是受反内幕交易相关法律所约束的，它要求任何用于商业获利的信息必须是可以通过公开渠道获取的，这是为了确保市场的公平与竞争环境。当数据可以被用于获利、明确被禁止应用与其他目的且并不为公众所知时应被视为重大非公开信息（Material Non-public Information, MNPI）。掌握重大非公开信息的人不能将它非法地用于为自己或他人进行投资获利。

Differences in the legal systems between the two largest alt data markets, the U.S. and the EU, contribute to uncertainty as operations become international, but also to exclusivity issues.[38] As Point72's chief market intelligence officer, Matthew Granade told the Financial Times regarding acquiring alt data sets: "The great thing about this area is you can arrange deals where you are the only ones who get it."[39] So far, there have been few cases regarding insider trading and the use of alt data. The U.S. Securities and Exchange Commission (SEC) has only successfully prosecuted a single case: the case of *SEC vs Huang* involved the use of MNPI in credit card transactions to inform investment decisions relating to an outdoor goods retailer.[40]

对于美国和欧洲这两个最大的另类数据应用市场来说，他们法律体系的区别导致了这一问题的不确定性（尤其是当他们的运营走向国际化的同时），甚至带来了与排他性相关的问题[38]。就像Point72的首席市场情报官Matthew Granade对财经时报（Financial Times）所表达的关于获取另类数据集的观点所说：“这个领域最微妙的地方在于，如果你是唯一拥有相关数据的人，你就可以主导交易[39]。”到目前为止，已经有了若干起由另类数据的使用与内幕交易联系在一起的案件。美国证监会仅仅成功起诉了其中的一起，就是关于黄姓个人涉及利用信用卡交易数据中的重大非公开信息，对一家户外商品零售商的股票进行投机获利的案件[40]。

Initially, insider trading was a minor concern for firms, as most online and social media data being sold is public information. However, as the data categories of alt data have expanded into credit card transaction or geospatial data, these concerns have increased.[41]

最初，内幕交易相关的担忧并不是一个严重的问题，因为大多数被售卖的网络和社交媒体数据都是公开信息。然而，当另类数据的类别扩展到了信用卡交易和空间地理相关的数据以后，人们的这类担忧逐渐加剧了[41]。

U.S. regulators have been more proactive in the alt data space than their EU counterparts, where the rules for insider trading are broader, making it more difficult to prove illegal activity. Regulatory organizations such as the SEC, the Competition Markets Authority (CMA) and others need to be at the forefront of increasing market fairness by regulating access to data, balancing this against potential consumer harms.

相对于欧洲同行，美国的立法者在另类数据领域更为积极主动，那是因为美国关于内幕交易判定的规则比较宽泛，涉猎很广，这使得证明相关的非法行为更加困难。美国证监会，英国竞争与市场管理局（CMA，Competition Markets Authority）等监管机构需要站在促进市场公平的前列，监管对数据的获取与使用，并平抑其可能给消费者带来的潜在利益损害。

As highlighted in the previous section, ensuring clarity around both the provenance of data and its licensing will help to address concerns around insider trading. Data provenance will help to clearly identify sources and adoption of open licenses and/or standard data sharing agreements, and access methods will increase access to data.

正像之前所强调的那样，确保数据溯源和授权许可的透明度，有助于解决围绕内幕交易相关的担忧。数据溯源可以帮助清晰地识别与标注数据来源；采用开放的数据许可和标准的数据分享协议，以及采用建立在这些许可和标准之上的数据访问方法，都有助于促进对数据的访问和使用。

[34] SBAI (Feb. 2019)
[35] Eagle Alpha (June 21, 2016), "Web Crawling as alternative data, a regulatory perspective." https://www.celent.com/system/media_documents/documents/399/944/216/ original/554254227.pdf?1466607837
[36] The Open Data Institute (2018), "Our manifesto," https://theodi.org/about-the-odi/our-vision-and-manifesto/our-manifesto/
[37] Investopedia (Apr 30, 2018), "Material Insider Information," https://www.investopedia.com/terms/m/materialinsiderinformation.asp
[38] Integrity Research (Jan 2018)
[39] The Financial Times, "Hedge Funds See Gold Rush in Data Mining," https://www.ft.com/content/d86ad460-8802-11e7-bf50-e1c239b45787
[40] SEC (Jan 21, 2015), "Securities and Exchange Commission vs. Bonan Huang and Nan Huang," https://www.sec.gov/litigation/complaints/2015/comp23216.pdf
[41] Integrity Research (Jan 2018)

Improving access to alt data
改善另类数据的获取能力

Alt data providers are building new data infrastructure for the investment sector. Data infrastructure consists of data assets such as: data sets; identifiers and registers; the standards and technologies used to curate and provide access to those data assets; the guidance and policies that inform the use and management of data assets; the organizations that govern the data infrastructure and the communities involved in contributing to or maintaining it; and those who are impacted by decisions that are made using it.

另类数据提供商们正在为投资领域构建新的数据基础设施。数据基础设施由这些组成部分构成:

- 数据资产，包括数据集，标识符以及数据注册表。
- 技术与标准，用于对数据资产的编排维护与访问获取。
- 规则与指南，用于指导数据资产的管理和使用。
- 机构与社区，提供对数据基础设施的监理（机构）以及建议，贡献和维护（社区）。
- 以及其他被数据基础设施相关决策所影响的参与者。

In previous sections we highlighted the need for clearer policies and guidelines. In this section we look at other ways to strengthen data infrastructure, specifically through the adoption of open standards.

在之前的章节中我们提到了更明确的规则与指南的必要性。在这一章让我们来看看其他有助于加强数据基础设施的方法，特别是采用开放标准所能起到的作用。

The challenges of standardizing alt data
标准化另类数据的挑战

Open standards for data are reusable agreements that make it easier for people and organizations to publish, access, share and use better quality data.[42] Participants in our research consistently highlighted a lack of standards as an issue when consuming alt data, and reported that a lack of standards is contributing to increased costs when consuming data.

数据开放标准是一种可复用的协议，能够让个人和企业以更容易的方法发布、获取、分享和使用更高质量的数据[42]。参与我们调查的受访者一致提出的问题之一，是缺乏消费和使用另类数据的相关标准，他们还提到这种缺乏标准的现状造成了不断增加的数据消费成本。

Alt data includes vast amounts of non-standardized and unstructured data that takes time, talent and technology to properly analyze. This means that the quality and limitations of the data sets being sold in the market today are often not known by alt data users until this investment has been made.

另类数据包含了大量的非标准以及非结构化的数据，而正确的分析这种数据需要投入时间，合适的技能和相关的技术。这意味着市场上数据集的质量和局限性在用户投入相关代价具体使用之前，是很难得到衡量的。

The use of Web scraping and novel sources of data might allow alt data providers to quickly build up data sets, but this does not give any guarantees that the data is correct, has the necessary detail or coverage to inform decision making or is free from bias.

网页抓取和其他新颖的数据来源可以让另类数据提供商快速地构建出需要的数据集，但这无法提供任何关于数据正确性的保障，也无法保证这样的数据能够包含足够的细节或者覆盖范围来指导决策，也不能保证这样的数据不会包含偏差。

Increasing standardization could help to mitigate some of these issues. By reducing technical integration costs by standardizing data formats, or making due diligence easier by standardizing provenance information, it may reduce the effort needed to explore innovative uses of these new data sources or assess their validity.

提高标准化的程度可以帮助缓解某些这样的问题：标准化数据格式可以降低技术集成的成本；标准化数据溯源的信息可以让数据相关的各类应尽义务变得简单。这些都可以帮助减少围绕这些新数据源进行创新探索或合理性评估所需付出的代价。

However, creating standards for alt data sets is challenging.

然而，为另类数据设立标准充满了挑战。

First, the sheer range of data sets being explored means there is a wide variety of potential areas for standardization. While some of these data types are well understood by the industry, others are still being explored. Satellite and credit card transaction data has been around for over a decade, while Web-scraped employment data is still growing in popularity.[43]

首先，目前被探索和应用的另类数据涉及的种类过于繁多，这意味着潜在的需要标准化的领域非常广泛。虽然一些类别的另类数据已经广为行业所熟悉和理解，但仍有很多正处于探索阶段。例如信用卡交易和卫星相关的数据用例已经存在了超过十年，但通过网页抓取而获得的就业数据正在变得流行[43]。

Second, the data sets produced by alt data providers are often highly bespoke and untested in the market. Potential buyers often do not have a concrete use in mind for alt data sets and are looking to be led by vendors. The process of acquiring and exploring an alt data set, and understanding how it can be combined with existing sources, is currently quite iterative.

其次，另类数据供应商们提供的数据集经常是高度定制化的，并且有效性并没有经历市场的充分验证。潜在的数据买家通常心中并没有非常成型的用例，而是寻求数据供应商的引领。对于另类数据获取和应用的流程，以及探索如何将另类数据结合于现有的数据源的方法，目前还常常是通过迭代的方式来摸索进行的。

Finally, while standards have been highlighted as a general issue, community buy-in and support for creating or setting standards has been inconsistent. While there is interest and agreement in the need for standards, there has so far been little commitment from stakeholders.

最后，虽然相关标准的缺乏已经成为一个公认的问题，不同社区对于建立或设定标准这件事的接纳和支持程度却各不相同。虽然对相关标准的需求和兴趣已经成为共识，但到目前为止并没有足够的来自于各方的承诺和行动。

Premature standardization may hinder the ability to explore new ways to structure and publish these data sets. As the ODI's open standards guidebook highlights, when needs are unclear, then standardization may not be the right approach.[44]

不成熟的标准化也可能会妨害人们探索新的结构化和发布另类数据的能力。就像开放数据学会（ODI）的开放标准指南（Open Standards Guidebook）所指出的那样，当需求仍不明确的时候，标准化也许不是正确的方法[44]。

Prototyping and exploration to understand these needs may be a better short-term focus for the alt data sector, especially when the value of individual data sets may still be unproven in a finance and investing context.

特别是当某些另类数据的价值和作用还没有在金融和投资领域得到验证的话，通过构建原型和来理解针对另类数据的潜在需求，是值得短期投入和关注的重点。

[42] Open Data Institute (2018), “Open standards for data,” http://standards.theodi.org/

[43] Eagle Alpha (April 2018)

[44] The Open Data Institute (2018), “When not to create new standards,” http://standards.theodi.org/introduction/when-not-to-create-new-standards/

Standards in the market today
当今市场上的一些标准

Despite these challenges, the alt data sector has been making some steps towards creating and adopting standards. The initial steps have focused primarily on the processes supporting due diligence and data acquisition, with trial data agreements and due diligence questionnaires (DDQs)[45] being the major focus.

尽管有各种各样的挑战，另类数据行业在制定和采用标准的道路上已经取得了一些进展。早期的努力主要是从支持数据采购和尽职评估的流程入手，而关注点主要集中在数据试用协议(trial data agreements)和尽职评估问卷(DDQs, due diligence questionnaires)上面。

There are several organizations that are trying to coordinate broader activities around standards for alt data. These include the UK- based Standards Board for Alternative Investments (SBAI), and U.S.-based Financial & Information Services Association (FISD) and Investment Data Standards Organization (IDSO).

业界的一些组织也在尝试协调更广泛的另类数据相关的标准化工作。这包括英国的另类投资标准化委员会（SBAI, Standards Board for Alternative Investments），以及位于美国的金融与信息服务联盟（FISD,Financial & Information Service Association）以及投资数据标准化组织（IDSO,Investment Data Standards Organization）所做出的努力。

The SBAI is “the custodian of the standards and brings investors, managers and regulators together to collaboratively improve the standards.” Their standards in the alt data space are more related to conduct rather than technical standards, focusing on disclosure, valuation, risk management, fund governance and shareholder management. Their main contribution to the alt data ecosystem so far has been the Standardised Trial Data License Agreement, published in collaboration with Eagle Alpha. The license agreement has been designed to speed up the process of evaluating data sets before purchasing.

SBAI定位自己的角色是“另类投资行业的准则制定机构，并扮演标准则托管人角色。并致力于协调投资人，投资管理者以及监管机构共同完善这些标准”。他们在另类数据领域关注得更多的是与信息披露、估值、风险管理、基金监管和股东管理相关的操守及行为方面的标准化，而不是技术的标准化。到目前为止，他们对另类数据生态系统的主要贡献就是与Eagle Alpha共同发布和撰写的“数据试用许可证标准化协议”（Stadardised Trial Data License Agrement）。这是为了帮助使用者加快其在购买数据集之前所需要进行的数据评估流程而设计的。

A standardized agreement will lower the risk on trialing data and improve the efficiency of the current trial negotiation process. This will reduce the time spent in these processes, ideally allowing for a more efficient industry and a higher level of innovation through more rapid testing processes.[47] SBAI is focusing on standardizing one aspect of reusing third-party data sets, as opposed to more technical specifications around file formats and schemas.

通过标准化的协议，机构可以降低数据试用过程中的风险，并提高与另类数据提供者商讨相关条款的效率。它希望通过减少在这些流程上花费的时间来提升整个行业的效率，并通过更敏捷的数据试用流程来促进更高层次的创新型应用[47]。SBAI的关注点并不是从技术层面的标准化（比如文件格式和文档结构），而是更聚焦于解决第三方数据利用过程中的某一个特定的方面。

The FISD is “the global forum of choice for industry participants to discuss, understand and facilitate the evolution of financial information for the key players in the value chain including consumer firms, third-party groups and data providers.”[48] FISD developed the Market Data Definition Language (MDDL) metadata standard[49] as an open industry standard for securities market data[50] that is considered for use with alt data. MDDL is already used for financial instruments, corporate events and market-related data,[51] so it seems natural to extend its use as a metadata standard into alt data. The information it covers “includes pricing, descriptive and reference information, and statistics about financial instruments, exchanges and the organizations that trade through them, the economy in general, and other related economic and business factors.”[52]

FISD是“专注于金融与信息行业的全球化论坛，帮助价值链上的消费者企业，第三方组织以及数据提供者讨论、理解并推动金融信息服务的不断演化”[48]。FISD针对证券市场数据开发了名为“市场数据定义语言” 的一个元数据标准（MDDL, Market Data Definition Language[49]）[50]。作为一个开放的行业标准，MDDL已经被应用于金融工具、公司事件以及其他类型的金融市场相关数据[51]，人们很自然地会考虑将它扩展并应用于另类数据领域。它所覆盖的信息包括“价格行情、描述和参考信息；关于金融工具，交易所和交易参与机构的统计数据；经济形势信息；以及其他金融市场相关的经济和商业因素的信息[52]”。

IDSO seeks to improve standardization by publishing best practices, checklists and technical specifications. Their Web crawling checklist touches on regulation and compliance, website assessment criteria and risk management amongst other areas.[53]

IDSO则致力于通过发布最佳实践、检查清单（checklist）和技术文档来提高标准化水平。他们发布的“网页抓取检查清单”就涉及了监管与合规、网站评估标准以及风险管理等若干领域[53]。

However, their main focus has been on the handling of personal data where they have recommended the use of existing privacy standards.[54] IDSO has defined a non-exhaustive list of personal data categories and examples such as name, address, phone numbers, account info, personal characteristics, linked information and more.[55] IDSO has also defined a three-tiered personally identifying information (PII) identifiability scale, ranging from direct identification (Level 1), to the ability to contact or impersonate (Level 2), to personal information that does not identify an individual (Level 3). There are numerous methods recommended for anonymization in the guide such as pseudonymization, hashing and swapping.

然而，IDSO的主要关注点在于对个人信息数据的相关处理，并推荐对一些现有的隐私相关标准加以采纳和利用[54]。IDSO定义了一个关于个人信息数据分类和示例的清单。虽然这个清单的目的并不是穷尽所有可能的类别，但已经包括了姓名、地址、电话号码、账户信息、个人特征、关联信息等一些常见的分类[55]。IDSO还定义了一个三个层级的个人识别信息(PII, Personal Identifying Information）可识别度等级体系，其等级范围包括：可用于直接识别个人对象的信息（一级）、可用于联系或伪装个人对象的信息（二级），以及虽属于个人信息但无法用于识别个人对象的信息（三级）。相关的指南还包括相当数量的对个人信息数据进行匿名化处理的方法，例如假名化（pseudonymization），哈希处理（hasing）、信息置换（swapping）等。

[45] Alternative Investment Management Association (Oct 13, 2017), “AIMA launches new due diligence template,” https://www.aima.org/article/aima-launches-new-due-diligence- template.html

[46] SBAI, “About Us,” https://www.sbai.org/about-us/

[47] SBAI (Feb. 2019), ‘SBAI Publishes Standardised Trial Data License Agreement,” https://www.sbai.org/wp-content/uploads/2019/02/SBAI-Press-Release-SBAI-Publishes-Big-Data-Trial-Agreement-6-Feb-2019.pdf

[48] Financial & Information Services Association (2019), “About FISD,” http://www.siia.net/Divisions/FISD-Financial-Information-Services-Association/About

[49] MDDL (2014), “The Market Data Definition Language,” https://web.archive.org/web/20130512024014/http://v3-beta.mddl.org/

[50] Finextra (May 30, 2007), “FISD launches market data definition language 3.0 beta,” https://www.finextra.com/pressarticle/15193/fisd-launches-market-data-definition-language-30-beta

[51] MDDL (2014)

[52] Finextra (2007)

[53] IDSO (Jan 2018), “IDSO Best Practises: Web Crawling,” https://docs.wixstatic.com/ugd/c6ff57_43135666c05c49f88ff7c2763ef846f0.pdf

[54] The National Institute of Standards and Technology (April 2010), “Guide to Protecting the Confidentiality of Personally Identifiable Information (PII),” https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-122.pdf

[55] IDSO (May 2018), “IDSO Best Practises: Personally Identifiable Information (PII),” https://docs.wixstatic.com/ugd/78bb6b_8c273efa9b934df796fd309bb5fb5de8.pdf

Increasing adoption of standards
促进行业标准的采用

As the current activities highlight, while there are challenges around standardizing some aspects of the technical infrastructure around alt data, there are clearly some areas where adoption of common standards and best practices could be beneficial.

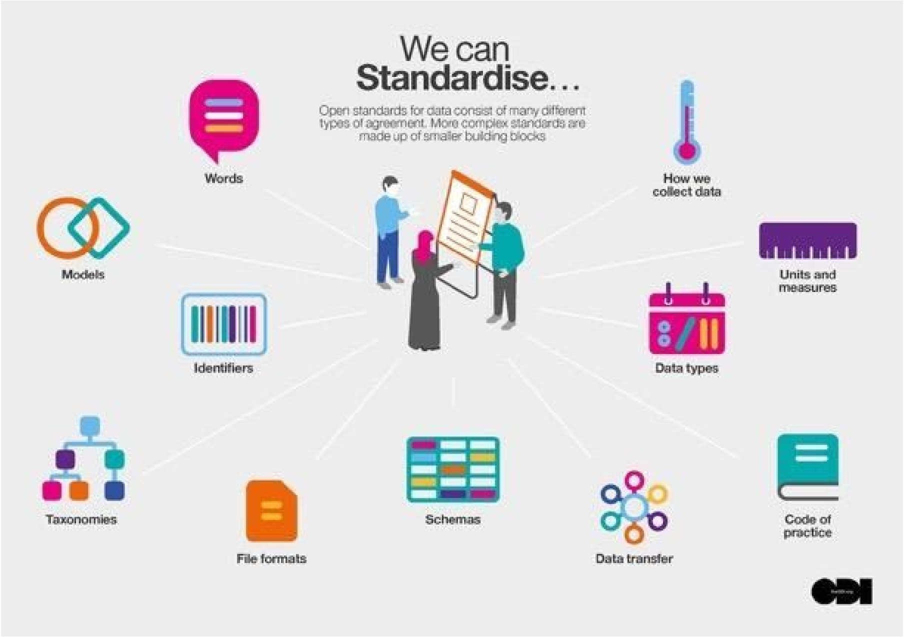
从目前我们观察到的一些活动来看，虽然在标准化技术基础设施的某些方面还存在一些挑战，但非常明显的是，采用一些现有的常见标准和最佳实践，已经能给这个行业带来很多好处。

Data sets are often mapped and combined at great expense in terms of time and money, and every standardized identifier and practice could contribute to market efficiency.[56] The sector should continue to focus on agreeing to use common standards for areas that are already well-defined, and where there needs to be broad agreement to help reduce risks and unlock benefits.

将多个数据集相互关联和集成在一起，经常需要耗费很多时间和成本，而任何标准的数据标识符体系与相关实践的采用都有助于提高市场的效率[56]。这个行业应该继续关注和推进对常见标准的采用，特别是在那些已经成熟的应用领域，以及那些需要形成更广泛的协议与共识来降低风险和释放更大潜能的领域。

As the figure below shows, we can standardize a variety of different components of data infrastructure.[57]

如下图所示：我们可以对很多数据基础设施的组成部分进行标准化[57]。



图片来源于Open Data Institute。

In the context of alt data, the sector could choose to create and adopt standards for:

- Representing basic data types, for example dates, and use of common data formats to help reduce unnecessary friction when analyzing data
- Identifying entities such as organizations, geographic areas, products. Using existing identifier schemes like PermID® could make it easier to combine data from different sources, even if the content and structure of data sets varies
- Codes of practice and technical methods for Web scraping and other forms of data collection
- Documenting data sets, including key metadata standards, that will help to describe the provenance of data sets and the processes by which data has been collected and analyzed
- Describing the quality, coverage and known limitations of a data set

行业可以制定和采用相关的数据标准并用于：

- 表达基础的数据类型例如日期，从而减少在数据分析中遇到的麻烦。
- 标识常见实体包括组织、地理位置、产品等。使用现有的一些通用标识符例如路孚特的PermID®，可以简化将不同来源的数据进行整合的过程，即使这些数据的结构和表达方式各不相同。
- 为网页数据抓取和其他形式的数据搜集行为定义实操准则和技术方案。
- 为数据集创建所需的文档，包括那些用于辅助描述数据溯源和数据收集、分析过程的元数据标准。
- 描述数据集的质量、覆盖范围和已知的各种约束。

Alt data practitioners should also look to other parts of the broader data ecosystem to understand best practices in other domains. Statistical agencies, for example the Office of National Statistics (ONS) in the UK, provide good examples of how to document the provenance of data sets and identify the limitations of their use in other contexts.[58]
另类数据的践行者也应该从整个数据生态系统中的其他领域学习相关的最佳实践。英国的国家统计局(ONS, Office of National Statistics)作为一个数据统计机构就提供了一个很好的例子，关于如何记录数据集的溯源，以及标注它们在其他领域的适用范围与限制[58]。
Creating standards should be done through open processes. Collaborative models build trust, reduce cost and create more value than other approaches. Being open improves quality, as more people can contribute to the outcome, and increases the number of connections that can be made.[59] This approach has worked in adjacent industries before, the best example of which is the development of the Open Banking Standard. The CMA convened the nine largest consumer banks in the UK, and supported by organizations like the ODI, was able to implement an industry-wide set of standards around open APIs.[60]
制定标准也需要通过开放的流程来进行。相互合作的模式能够构建信任，相比其他模式也能够通过更少的成本来创造更多的价值。因为可以有更多人参与达成最终的结果，拥抱开放也会提高标准制定的质量，并在这个过程中增加与生态系统的各种联系[59]。这样的办法曾经在相近的行业中被采用过，其中很好的一个例子就是关于开放银行标准(Open Banking Standard)的制定。英国竞争与市场管理局（CMA）通过汇集英国九家最大的消费者银行，并得到了ODI这样的组织的协助，最终得以在行业内推广了一套开放编程接口相关的技术标准[60]。

[56] Amelia Axelsen (Feb 20, 2019), “Crowded Alt Data Market Makes Standing Out Difficult for Providers,” https://www.waterstechnology.com/data-management/4162636/ crowded-alt-data-market-makes-standing-out-difficult-for-providers
[57] The Open Data Institute (2018), “Types of open standards for data,” http://standards.theodi.org/introduction/types-of-open-standards-for-data/ [58] Office for National Statistics (2011), “Methodology and variables,” https://www.ons.gov.uk/methodology/geography/geographicalproducts/areaclassifications/2011areaclassifications/methodologyandvariables
[59] Open Data Institute (Aug 31, 2016), “Principles for strengthening our data infrastructure,” https://theodi.org/article/principles-for-strengthening-our-data-infrastructure/
[60] Open Data Institute (2016), “Open banking: setting a standard and enabling innovation,” https://theodi.org/project/open-banking-setting-a-standard-and-enabling-innovation/

Recommendations and next steps
我们的建议，以及下一步的措施

As we have shown, there are a number of issues around privacy, rights, ethical uses of data and the role of standards. Based on that, we recommend that the alt data industry focuses on several areas to create a more open, trustworthy data ecosystem, as set out below.
正像我们所展示的那样，在另类数据生态系统中存在着关于隐私、权利、数据伦理以及相关标准的角色的一系列问题。基于这些分析，我们建议另类数据行业集中关注下面这些领域，来开创一个更加开放和可信的数据生态系统。
Both alt data providers and users need to ensure a strong commitment to ingraining legal and ethical practices across the alt data sector. Legal compliance and ethical behavior serve as a differentiator, a premium value proposition that helps ensure low-risk and high- quality products and services. Adhering to the law and demonstrating ethical behavior will reduce harm to the company and to society, and it will demonstrate the high value of one’s investment services.
另类数据的提供商和使用者需要确保在这个行业中植入强大的法律和数据伦理观念。“合理合规”是另类数据产品与服务打造自身高质量和低风险的独特价值，并区别于其他竞品与对手的关键特性。对投资机构来说，在法律框架下操作并展现自己的合理合规，可以降低另类数据的使用对企业和社会产生的副作用，并展现其所提供的投资服务的优越性。
Organizations in the alt data space also need to collaborate better to improve access to data. Although there is currently some agreement across the industry on best practices, primarily related to the due diligence process and provenance, this remains limited. More collaboration around standards for technical aspects and transparent processes is both needed and achievable by an industry that is currently desiring it.
尽管目前在业界已经取得了关于某些最佳实践的共识 (主要是针对数据采取应尽义务的相关流程以及数据溯源方面)，但这方面的进展到目前为止还非常有限。另类数据领域的企业之间还需要进一步的合作来促进对数据的利用。在技术相关领域以及流程的透明性方面，通过更多的合作来推行标准化对于一个有很多这方面需求的行业来说，是必要而且十分可行的。

Recommendations for alt data providers
给另类数据提供商的建议：

- Alt data providers need to help their customers manage and mitigate legal and operational risks associated with the use of their data sets. This can be done by ensuring they have appropriate rights/permissions to collect and distribute data, conducting risk assessments to help customers understand potential ethical or legal issues, providing clear provenance information, and, where rights are unclear, helping their customers stay on top of changing terms.

- Alt data providers need to be transparent with users about the quality and limitations of their data sets. This is important so that decisions based on their data sets are appropriate to their accuracy, coverage and timeliness. If provided, case studies and sample data need to be carefully documented so that users can make a fair assessment of the value of the data.
- Alt data providers should be open with the organizations, individuals and communities impacted by the data they are scraping/ repurposing. This will help create more trust and transparency across the data ecosystem.

- 另类数据提供商需要主动帮助他们的客户管理和降低那些与使用另类数据所相关的，在法律与操作层面的风险。例如，帮助确保他们有合适的授权与许可来收集和向客户分发数据、通过风险评估来帮助客户理解潜在的伦理和法律风险、提供明确的数据溯源信息，以及当数据权利并不明确的时候，帮助客户及时掌握相关的变化和动态等等。
- 另类数据提供商需要对他们的客户保持透明，使得客户能够确切理解他们所提供数据集的质量和局限性。这对客户来说非常重要，关乎他们基于这些数据集的准确性、覆盖范围和时效性而所做出的决策是否合理。如果对客户提供了相关案例和试用数据的话，一定要确保相关的文档能够准确地描述和传达信息，来确保客户能够对数据的价值和适用性做出合理的判断。
- 另类数据提供商如果进行了网站数据抓取，或者将某些他们掌握的数据用于额外用途的话，他们需要对那些可能受这些行为影响的组织机构、个人和社区团体保持坦诚和开放。这有助于在整个数据生态系统中打造透明与可信的氛围。

Recommendations for alt data users
给另类数据使用者的建议

- Alt data users should require data suppliers to provide detailed metadata, provenance information and documentation about the data sets they are supplying. By setting expectations with suppliers, the users of alt data can help to encourage good behavior and create a more transparent and sustainable data ecosystem.
- Alt data users should build robust processes to assess whether alt data sources are properly licensed. Licenses are likely to be nonstandard but need to describe key areas such as usage rights (rights over redistribution), representations and warranties about the authenticity of the data, and the rights of the seller to license such data.

- 另类数据用户应该要求数据供应商对所提供的数据集提供详尽的元数据、溯源信息以及各种文档。通过对供应商提出明确的期望，另类数据的使用者可以帮助在行业中提倡正确的行为方式，从而打造更为透明和可持续发展的数据生态系统。
- 另类数据用户应该推行可靠的流程，来评估供应商所提供的另类数据是否对自己进行了正确的使用授权。相关的授权许可很可能无法全都通过统一的标准来描述，但仍然需要描述关键的几个部分例如数据的使用权利（数据提供商否有权再分发数据）、关于数据真实性的表述和证据、以及数据提供商是否有权对使用者进行数据使用的再授权等。

Recommendations for both alt data providers and users
给另类数据提供商和使用者共同的建议

- All organizations in the alt data market should implement ethics assessments that go beyond compliance and legal issues.
  Use tools such as the ODI’s Data Ethics Canvas to help identify and make decisions about potential ethical issues associated with alt data-informed investment activity. Publish your findings openly, if you can.
- Convene the key industry players and regulators in the alt data market to agree on a road map for standard adoption to improve technical integration and data set due diligence. Similar to the implementation of the Open Banking Standard in the UK, have regulators and neutral third parties help decide on a framework for implementation.[61] Standards will need to be stewarded by an independent organization and organizations may need support in implementing them. In the banking sector, this role is fulfilled by the Open Banking Implementation Entity (OBIE).[62]
- Identify where existing (open) standards in the market can be used and where new (open) standards need to be created. Consider existing standards such as PermID, Market Data Definition Language (MDDL), DDQs and the Standardized Trial Data License Agreement before creating new standards. Identify if the SBAI, FISD, IDSO or other groups have other standards to be leveraged. Where new standards are required, follow open processes, such as those described in the ‘Open Standards for Data’ handbook. Starting small, for example by agreeing on common identifiers and basic data formatting, may help to build momentum.

- 另类数据市场的所有参与各方，都应该进行超越法律和合规要求的额外的数据伦理评估。相关工具（例如开放数据学会所提供的“数据伦理画板”）可以帮助识别那些基于另类数据的投资活动中所包含的潜在数据伦理风险，并指导相关的决策。另外如果可能的话，请分享在这个过程中发现和学到的东西。
- 建议汇集关键的另类数据业内参与者和监管机构并一起制定关于相关标准采纳的路线图，以便改善另类数据的技术集成和针对数据的各类应尽义务。与前面提到的英国制定“开放银行标准”的过程类似，请联合监管机构和中立的第三方一起来制定帮助落实的框架。相关标准需要被独立的组织来管理和协调，而各个企业实施相关标准时也需要其他组织来帮助落地。在英国的银行业领域，这个角色是由Open Banking Implementation Entity (OBIE，可译为“开放银行实施机构”）来承担的。
- 识别市场上哪些现有的标准可以被推广使用，而哪些领域需要制定新的开放标准。在制定新的标准之前，请考虑使用现有的标准，例如前面提到的PermID，Market Data Definition Language (MDDL，“市场数据定义语言”），DDQs （Due Diligence Questionnaire，关于对数据应尽义务的标准问卷），以及”数据试用许可证标准化协议”（Standardised Trial Data License Agreement）等。请考虑是否SBAI，FISD，IDSO以及其他一些组织是否有其他的可以加以利用的相关标准。当需要制定新的标准时，请遵循一个开放的流程，例如在ODI的“开放数据标准”手册(Open Standards for Data）中描述的那样。通常情况下，从简单的领域入手有助于工作的稳步开展，例如将通用的数据标识符和基础的数据格式形成为行业标准等等。

[61] Open Data Institute (2016), “Open banking: setting a standard and enabling innovation,” https://theodi.org/project/open-banking-setting-a-standard-and-enabling-innovation/

[62] Open Banking Ltd. (2016), “About Us,” https://www.openbanking.org.uk/about-us/