Data Science Notes

Barclays Research's Data & Investment Sciences blog, where we dive into some of the frequently asked questions about data science and how it is reshaping investing.

BARCLAYS

26 June 2020

Data Science
Adam Kelleher

www.barclays.com

Data Science 101

We frequently field the question of how investors should build a data science capability. In this entry, we share the framework we have used for making business decisions on data science.

Below are the key dimensions to consider:

- **People.** Hiring data scientists is different from hiring investment professionals, and there is a wide range of levels of domain expertise, engineering skill, and quantitative acumen. Still, most investment firms employ plenty of people with the skill and aptitude needed to develop into data scientists. Given the freedom to pick up computer science skills, these people can grow to become data scientists.

- **Infrastructure.** The systems and facilities needed for data science start with advanced versions of what most investors are already using in their IT (with some entirely new capabilities at the very top end), but generally some spending will be required for automated data ingestion and analysis.

- **Data.** Most investors could add some data science capabilities by just doing new things with data they already own. At the more sophisticated (and expensive) end, the data requirements can be an entirely new discipline that requires a dramatically different set of capabilities across the organization, including data scientists, IT engineers, and legal and compliance expertise.

How companies choose to address each one of these dimensions depends on answering the following questions:

- What are you trying to get out of your data science capability, and what is the value to your business?
- What capabilities do you already have?
- What can you reasonably accomplish; put another way, how much are you willing to spend?

This is a high-dimensionality problem, so there are innumerable ways to assemble staff, infrastructure, and data to address the three questions. That said, we think there are a few focal point answers: Figure 1 presents three reasonable options that cover the spectrum of how investors can approach the problem from a resourcing perspective.

It is more efficient to use some data science terms of art in these descriptions, so we define some key ones at the outset. These definitions reflect how we use them in our writing and may differ from how they are used elsewhere. **Data science** is statistics implemented using a heavy dose of computer science (or sometimes statistics as rediscovered by computer scientists). **Data scientists** are its researchers and practitioners (and usually have come to it from a spectrum between pure computer science and pure statistics). **Python** is a scripting language that is popular for data scientists, but one can substitute it for any similar language – R, matlab, etc. Jupyter notebooks are a popular software package that allows one to write Python code and see the results it produces as one goes. **Spark** is a framework for large-scale computing (allocating work across many computers) and can be used within Python. **Production-grade** processes are when computers are able to complete a task beginning to end with no human intervention (unless something goes wrong). **Development** is the process of building production processes, and **ad-hoc analyses** are when humans use data science tools to complete one-off tasks such as answering a question.

FIGURE 1

Our view of some reasonable options for starting a Data Science team at different scales of investment

Scale	People	Infrastructure	Data	Direct Costs	Positives and Considerations
Efficient Can do a lot with limited resources	• Assign existing analysts to integrate some alternative data into the research process	• Bring pre-packaged alternative data signals into Excel • Do some work with Python in Jupyter notebooks with larger datasets • Engineering requirements are limited to building and maintaining data stores, with limited “productionization” of data products.	• Machine-readable foundational data (prices, corporate fundamentals, etc) • Use pre-packaged signals and APIs	• $0 for people, but some lost productivity as they ramp up • $10s of k for database servers, plus some IT support • <$100k (often much less) per dataset	+ Cheap and quick to get running + Making "table stakes" investments in alt data + Integrating organically into culture + Fast delivery - Limited scope for competitive advantage or marketing potential - Reliance on vendors for data processing

Scale	People	Infrastructure	Data	Direct Costs	Positives and Considerations
Innovative Some real competitive advantage with a few more resources	• Hire a dedicated data scientist to support investment and research teams • Allocate dedicated IT resources (people and computers) to production tasks	• Work primarily with Python in Jupyter notebooks • Build and maintain data stores, with some "productionization" of data products. • Add some specialty IT support or cloud services for production engineering • Can use vendors for many intermediate tools	• Machine-readable foundational data (prices, corporate fundamentals, etc) • Make better use of pre-packaged signals and APIs • Purchase some smaller-scale raw sources of data	• A budget for data scientist and IT salaries comparable to your data budget~$[100k] for servers and cloud tools • Typically want 1-3 people • $50-100k per dataset for sector specific data, or $100k->$1mn for each general purpose datasets (transactions, geolocation, text)	+ You can build some of your own custom insights and signals that nobody else might have + You have a data science team, so there is an opportunity for differentiation and marketing + Still lean and pretty fast - Team will have limited capacity - Capabilities depend a lot on who is hired as a data scientist; the best ones are more expensive - Success will depend on integrating a new capability into the team culture
Resourced Hire a full team, and buy new data to do truly novel work	• Hire a data scientist leader, who builds out a full team • Multiple dedicated IT resources to support production tasks	• Python in Jupyter notebooks • Spark on clusters • Full IT build of servers, computing clusters, etc; or full IT support for cloud services, possibly including new service design patterns and large-scale compute models	• Machine readable foundational data (prices, corporate fundamentals, etc) • Consume significant raw data, including the development of custom production algorithms needed to transform data to insights	• A budget for data scientist and IT salaries comparable to your data budget At least $100s of k for servers and cloud tools • 3 people likely the practical minimum • At least $1mn for multiple general use + specific use datasets • Additional need for operational support to build and manage the business cases for different types of investments	+ Potential for novel products that are truly differentiating + Can build entirely novel datasets + This is a real investment area now, so real opportunity for differentiation and marketing - Spending a lot on it, which is hard to justify if results are not produced - Scale, complexity, and difficulty of hiring can result in a slow ramp - A substantial amount of what is produced will be invisible to customers - More potential for culture friction - When things break, you have to fix them yourself. - The more systems you have, the higher the cost – and it grows over time!

Source: Barclays Research

Disclaimer:

BARCLAYS

This communication has been prepared by Barclays.
“Barclays” means any entity within the Barclays Group of companies, where “Barclays Group” means Barclays Bank PLC, Barclays PLC and any of their subsidiaries, affiliates, ultimate holding company and any subsidiaries or affiliates of such holding company.

CONFLICTS OF INTEREST

BARCLAYS IS A FULL SERVICE INVESTMENT BANK. In the normal course of offering investment banking products and services to clients, Barclays may act in several capacities (including issuer, market maker and/or liquidity provider, underwriter, distributor, index sponsor, swap counterparty and calculation agent) simultaneously with respect to a product, giving rise to potential conflicts of interest which may impact the performance of a product.

NOT RESEARCH

The information provided does not constitute ‘investment research’ or a ‘research report’ and should not be relied on as such, although it may contain references to views or information published by the Barclays Research department. Investment decisions should not be based upon the information provided.

BARCLAYS POSITIONS

Barclays may at any time acquire, hold or dispose of long or short positions (including hedging and trading positions) and trade or otherwise effect transactions for their own account or the account of their customers in the products referred to herein which may impact the performance of a product.

FOR INFORMATION ONLY

This information has been prepared by the Research Department within the Investment Bank of Barclays. The information, analytic tools, and/or models referenced herein (and any reports or results derived from their use) are intended for informational purposes only. Barclays has no obligation to update this information and may cease provision of this information at any time and without notice.

NO OFFER

Barclays is not offering to sell or seeking offers to buy any product or enter into any transaction. Any offer or entry into any transaction requires Barclays’ subsequent formal agreement which will be subject to internal approvals and execution of binding transaction documents. The products mentioned may not be eligible for sale in some states or countries, nor suitable for all types of investors.

NO LIABILITY

Neither Barclays nor any of its directors, officers, employees, representatives or agents, accepts any liability whatsoever for any direct, indirect or consequential losses (in contract, tort or otherwise) arising from the use of this communication or its contents or reliance on the information contained herein, except to the extent this would be prohibited by law or regulation.

NO ADVICE

Barclays is not acting as a fiduciary. Barclays does not provide, and has not provided, any investment advice or personal recommendation to you in relation to any transaction and/or any related securities described herein and is not responsible for providing or arranging for the provision of any general financial, strategic or specialist advice, including legal, regulatory, accounting, model auditing or taxation advice or services or any other services in relation to the transaction and/or any related securities described herein. Accordingly Barclays is under no obligation to, and shall not, determine the suitability for you of the transaction described herein. You must determine, on your own behalf or through independent professional advice, the merits, terms, conditions and risks of any transaction described herein.

NOT A BENCHMARK

The information provided does not constitute a financial benchmark and should not be used as a submission or contribution of input data for the purposes of determining a financial benchmark.

INFORMATION PROVIDED MAY NOT BE ACCURATE OR COMPLETE AND MAY BE SOURCED FROM THIRD PARTIES

All information is provided “as is” without warranty of any kind. Because of the possibility of human and mechanical errors as well as other factors, Barclays is not responsible for any errors or omissions in the information contained herein. Barclays is not responsible for information stated to be obtained or derived from third party sources or statistical services. Barclays makes no representation and disclaims all express, implied, and statutory warranties including warranties of accuracy, completeness, reliability, fitness for a particular purpose or merchantability of the information contained herein.

PAST & SIMULATED PAST PERFORMANCE

Any past or simulated past performance including back-testing, modelling or scenario analysis contained herein is no indication as to future performance. Past performance is not necessarily indicative of future results. No representation is made as to the accuracy of the assumptions made within, or completeness of, any modelling, scenario analysis or back-testing and no representation is made that any returns will be achieved.

OPINIONS SUBJECT TO CHANGE

All opinions and estimates are given as of the date hereof and are subject to change. The value of any investment may also fluctuate as a result of market changes. Barclays is not obliged to inform the recipients of this communication of any change to such opinions or estimates.

NOT FOR RETAIL

This document is being directed at persons who are professional investors and is not intended for retail customer use.
Not For Further Distribution or Distribution To Retail Investors.
For Discussion Purposes Only.

IMPORTANT DISCLOSURES

For important regional disclosures you must read, visit the link relevant to your region. Please contact your Barclays representative if you are unable to access.
EMEA
https://www.home.barclays/disclosures/important-emea-disclosures.html.
APAC
https://www.home.barclays/disclosures/important-apac-disclosures.html.
U.S.
https://www.home.barclays/disclosures/important-us-disclosures.html.

CONFIDENTIAL

This communication is confidential and is for the benefit and internal use of the recipient for the purpose of considering the securities/transaction described herein, and no part of it may be reproduced, distributed or transmitted without the prior written permission of Barclays.

ABOUT BARCLAYS

Barclays Bank PLC is authorised by the Prudential Regulation Authority and regulated by the Financial Conduct Authority and the Prudential Regulation Authority and is a member of the London Stock Exchange. Barclays Bank PLC is registered in England No. 1026167 with its registered office at 1 Churchill Place, London E14 5HP.

COPYRIGHT