Data Scientists on Wall Street

Author: Ling Cen, Bing Han, Yanru Han, and Chanik Jo

Discussion by:

Varun Sharma (Indiana University)

ABFER, 2026

May 2026

Information and alpha

Alpha = Insights about economy/sectors/firms = Data + Analytics

Information and alpha

Alpha = Insights about economy/sectors/firms = Data + Analytics

Data: Funds can acquire or purchase data, which makes any competitive advantage build just on data short lived

- Literature treats it as an abstract input in funds' asset allocation
- Chi, Hwang, and Zheng (2025); Katona, Painter, Patatoukas, and Zeng (2025); Farboodi, Matray, Veldkamp, and Venkateswaran (2022); Bonelli and Foucault (2023)

Information and alpha

Alpha = Insights about economy/sectors/firms = Data + Analytics

Data: Funds can acquire or purchase data, which makes any competitive advantage build just on data short lived

- Literature treats it as an abstract input in funds' asset allocation
- Chi, Hwang, and Zheng (2025); Katona, Painter, Patatoukas, and Zeng (2025); Farboodi, Matray, Veldkamp, and Venkateswaran (2022); Bonelli and Foucault (2023)

Analytics: Human capital that processes data can be retained (one fund's data scientist cannot simultaneously work for another fund)

- Analyst (Cheng, Liu, and Qian (2006)) and data scientists

Information and alpha

Alpha = Insights about economy/sectors/firms = Data + Analytics

Data: Funds can acquire or purchase data, which makes any competitive advantage build just on data short lived

- Literature treats it as an abstract input in funds' asset allocation
- Chi, Hwang, and Zheng (2025); Katona, Painter, Patatoukas, and Zeng (2025); Farboodi, Matray, Veldkamp, and Venkateswaran (2022); Bonelli and Foucault (2023)

Analytics: Human capital that processes data can be retained (one fund's data scientist cannot simultaneously work for another fund)

- Analyst (Cheng, Liu, and Qian (2006)) and data scientists

This paper treats the human capital that processes data — the data scientists — as the scarce resource used by funds to generate alpha

This paper

- **Question:** What are the implications of data scientists on fund performance, portfolio allocation, and market efficiency
- **Data:** Revelio Lab + Thomson Reuters Global Ownership
- **Identification:** Data-science programs introduction and M&A
- **Empirical findings:**
  - Institutional investors with more data scientists achieve higher trading profitability
  - Investors make portfolio concentrated and increase talent acquisition in response to rivals' hiring
  - When data scientists are concentrated among a few institutional investors, stock prices become less informative about future earnings

This paper

- **Question:** What are the implications of data scientists on fund performance, portfolio allocation, and market efficiency
- **Data:** Revelio Lab + Thomson Reuters Global Ownership
- **Identification:** Data-science programs introduction and M&A
- **Empirical findings:**
  - Institutional investors with more data scientists achieve higher trading profitability
  - Investors make portfolio concentrated and increase talent acquisition in response to rivals' hiring
  - When data scientists are concentrated among a few institutional investors, stock prices become less informative about future earnings

Financial institutions can harness human capital to gain a competitive edge, but this comes at the cost of lower market efficiency

Overall comment

Interesting paper with a timely topic!

- The paper fits well in the literature and has a clean contribution.
- The paper constructs a very rich dataset
  - Sample ends in 2021, but if it can be extended to recent years, can be used to study many other questions
  - AI and analyst productivity (complement or substitute); Skill bundling

Overall comment

Interesting paper with a timely topic!

- The paper fits well in the literature and has a clean contribution.
- The paper constructs a very rich dataset
  - Sample ends in 2021, but if it can be extended to recent years, can be used to study many other questions
  - AI and analyst productivity (complement or substitute); Skill bundling

One main comment: Structure of the paper

- Three, related but still distinct outcomes
- Each is individually interesting, but when combined, it lacks the impact and focus

Overall comment

Interesting paper with a timely topic!

- The paper fits well in the literature and has a clean contribution.
- The paper constructs a very rich dataset
  - Sample ends in 2021, but if it can be extended to recent years, can be used to study many other questions
  - AI and analyst productivity (complement or substitute); Skill bundling

One main comment: Structure of the paper

- Three, related but still distinct outcomes
- Each is individually interesting, but when combined, it lacks the impact and focus

Paper is spread thin due to focus on multiple outcomes, which constrain the execution depth

Comment 1: Focus of the paper (1/2)

Studies how the distribution of data scientists across funds affects:

1. Investor performance
2. Fund behavior (portfolio allocation and hiring)
3. Market efficiency (price informativeness due to information monopoly)

Comment 1: Focus of the paper (1/2)

Studies how the distribution of data scientists across funds affects:

1. Investor performance
2. Fund behavior (portfolio allocation and hiring)
3. Market efficiency (price informativeness due to information monopoly)

Different identification approaches across them

- Investor performance: Data analytics course introduction as IV
- Fund behavior: No causality
- Aggregate market efficiency: M&A as a shock

Comment 1: Focus of the paper (1/2)

Studies how the distribution of data scientists across funds affects:

1. Investor performance
2. Fund behavior (portfolio allocation and hiring)
3. Market efficiency (price informativeness due to information monopoly)

Different identification approaches across them

- Investor performance: Data analytics course introduction as IV
- Fund behavior: No causality
- Aggregate market efficiency: M&A as a shock

Focusing on one outcome will allow more rigorous empirical execution and a more thorough causal claim (more on this later)

Comment 1: Focus of the paper (2/2)

Focus on investor performance: Can data scientists help generate better alpha, and how?

- Improve upon the identification, substantiate it with more robustness

Comment 1: Focus of the paper (2/2)

Focus on investor performance: Can data scientists help generate better alpha, and how?

- Improve upon the identification, substantiate it with more robustness

Provide more insights about the channel:

- Use changes in portfolio construction results as an underlying channel
- Try to provide more details about what these data scientists are doing at the fund (IT security vs using alternate data)
- Apply LLMs/NLP on CV and job descriptions

Comment 1: Focus of the paper (2/2)

Focus on investor performance: Can data scientists help generate better alpha, and how?

- Improve upon the identification, substantiate it with more robustness

Provide more insights about the channel:

- Use changes in portfolio construction results as an underlying channel
- Try to provide more details about what these data scientists are doing at the fund (IT security vs using alternate data)
- Apply LLMs/NLP on CV and job descriptions

Keep the market efficiency for the next paper: maybe expand the sample, and check the impact after ChatGPT introduction

Comment 1: Focus of the paper (2/2)

Focus on investor performance: Can data scientists help generate better alpha, and how?

- Improve upon the identification, substantiate it with more robustness

Provide more insights about the channel:

- Use changes in portfolio construction results as an underlying channel
- Try to provide more details about what these data scientists are doing at the fund (IT security vs using alternate data)
- Apply LLMs/NLP on CV and job descriptions

Keep the market efficiency for the next paper: maybe expand the sample, and check the impact after ChatGPT introduction

Make one causal claim supported by a set of tests, then have two causal claims half done

Comment 2: Instrumental variable (1/2)

- **Exclusion restrictions:**
  - Start of data science programs at local universities probably correlates with broader trends
  - For example, regional technology ecosystem development, the presence of tech firms, and the general human capital environment
  - All of these could independently affect the sophistication and performance of nearby financial institutions
- **Local hiring:**
  - Morgan Stanley or Goldman Sachs don't hire data scientists primarily from universities in their headquarters state

Comment 2: Instrumental variable (2/2)

- **OLS vs IV:** OLS estimate (Table 3) is 0.004 percentage points per quarter per additional data scientist, but the IV estimate (Table 4) is 0.095, over twenty times larger

Comment 2: Instrumental variable (2/2)

- **OLS vs IV:** OLS estimate (Table 3) is 0.004 percentage points per quarter per additional data scientist, but the IV estimate (Table 4) is 0.095, over twenty times larger

- **Week instrument?**
  - A much higher IV coefficient indicates a weak instrument
  - In addition, the reported F-stat of 557 does not appear to line up with the first-stage coefficients and standard errors
  - Column 1 of Table 4 has a t-statistic around 6.3 (F-stat $=$ (t-stat)^2)

Comment 2: Instrumental variable (2/2)

- **OLS vs IV:** OLS estimate (Table 3) is 0.004 percentage points per quarter per additional data scientist, but the IV estimate (Table 4) is 0.095, over twenty times larger

- **Week instrument?**
  - A much higher IV coefficient indicates a weak instrument
  - In addition, the reported F-stat of 557 does not appear to line up with the first-stage coefficients and standard errors
  - Column 1 of Table 4 has a t-statistic around 6.3 (F-stat $=$ (t-stat)^2)

- **F-stat:** Is the reported F-statistic the Cragg-Donald F-statistic or the Kleibergen-Paap Wald F-statistic?

Comment 3: Financial institutions: Sell side vs. buy side

- Top three investors with the most data scientists are broker-dealers with large workforce (Morgan Stanley, Credit Suisse, Goldman Sachs)
- Funds usually have a smaller workforce, and p75 for the number of data scientists is 0
- Comparing the alpha of large broker-dealers with informational advantages (example: banking relationships or order flow) with funds
- Take financial institution type $\times$ time fixed effects in tables 3 and 4 (just like in table 2, panel A)
- Why focus on the number of data scientists and not the number of data scientists per dollar of AUM or as a proportion of total workforce

Other comments

- The Introduction states that the merged dataset contains 3,265,145 workers, 7,588 institutional investors, and 326,627 data scientists employed by 3,126 investors.
- In Section 2.1, the figures are 2,908,292 workers, 7,408 investors, and 124,947 data scientists employed by 1,957 investors.

Other comments

- The Introduction states that the merged dataset contains 3,265,145 workers, 7,588 institutional investors, and 326,627 data scientists employed by 3,126 investors.
- In Section 2.1, the figures are 2,908,292 workers, 7,408 investors, and 124,947 data scientists employed by 1,957 investors.
- Mechanical overlap in some regressions
  - Table 5: Dependent variable uses portfolio weight change; independent variable includes current portfolio weight; DS HHI uses holdings
  - Can create possible mechanical links between the left and right-hand side through holdings, portfolio weights, and stock ownership
  - Table 6: Data-scientist-heavy investors hold portfolios with higher portfolio-level DS HHI; can be mechanical because the investor's own data scientists enter the DS HHI of the stocks it holds
  - Try a leave-one-out approach (excludes investor's own data scientists)

Conclusion

- Timely paper that provides good insights about an important topic
- Making the paper more focused and some more robustness tests on the empirical strategy will further enrich the paper
- I look forward to reading the next version