WEISS

Frontline Data Science: Human + Machine

November 2018

The Chief Data Scientist at Weiss explains the unique role of the firm's Data Science team and provides forecasts on the fast-moving landscape of data analytics and machine learning on Wall Street.

Weiss is a 40-year-old firm[1] known for long/short discretionary investing. Our flagship fund seeks absolute, uncorrelated returns for our investors. The Weiss Data Science (DS) team provides data-driven analytics to the firm. We seek to extract useful insight from raw data, influencing the Allocation Committee and individual investment strategies with their own investment processes. Our goal is to improve risk-adjusted returns across our funds.

As the representative of the Weiss Data Science team, I am asked regularly about the team's role and responsibilities within the firm and how we expect to evolve. Allocators often ask these questions from the perspective of getting to understand our firm and investment process. Their questions are often also accompanied by a hint of curiosity and self-reflection about their firm's own data science practices (or lack thereof). Most modern firms are groping to establish best practices and strengthen the integration between data analytics and decision makers. The playbook is not straightforward, but there are clear practices that work and those that do not. We hope to provide readers with actionable ideas that will help with their organization's own data-driven journey.

This paper begins with details on how we conduct DS at Weiss, what it is (and is not), and provides rationale for our unique design choices. I conclude with (human-generated) forecasts on the evolution of DS on Wall Street.

[1] Through its affiliates.

Data Science the Weiss Way (Your Mileage May Vary)

Weiss began its DS journey just over five years ago. We are a front-office team of investment professionals who interface daily with portfolio managers (PMs), analysts, and traders. Some of us are PMs, some of us manage/enforce risk, some of us model macro regimes and focus particular attention on factor dynamics. We have representation on the allocation and risk committees. We contribute to the development, backtesting and trading of creative new strategies that coexist alongside our flagship fund.

All of us use open source tools to analyze data, model and visualize relationships. We strongly believe in the power of visualization for learning new concepts and deepening our understanding of familiar concepts. As a result, we build interactive web-based dashboards to enable others to prototype ideas, pose their own questions and, ultimately, conduct their own analysis. All of us code with an "agile" mindset using a central version control system. Most of our time is spent prototyping new ideas; the best ones are then automated and put into production across the firm. It's also common for us to become involved with the operational side of the business outside of investments, including but not limited to working with treasury/financing, operations, accounting, IT, legal and marketing.

As an active, long/short multi-strategy fund that typically maintains 1,000 to 1,500 positions (the output from approximately 20 independent discretionary strategies), we capture reams of data on our managers. Our challenge is to extract useful insight from the raw data that influences the Allocation Committee and helps individual teams with their own investment processes.

We build proprietary heuristics and machine learning (ML) models that track factor exposure dynamics across all levels of the firm, measure behavioral biases and assess strategy-level interactions to help avoid surprise concentrations in risk at the fund-level. For example, to increase our probability of generating *persistent* returns, it's critical that active bets (either implicit or explicit) expressed in a discretionary portfolio align with the PM's *edge*[2]. Without care, unintended factor or industry tilts can easily wipe out any idiosyncratic gains – a frustrating experience for managers (and allocators!). Because the skills needed for successful security selection are not the same skills needed for efficient portfolio construction, an opportunity exists for machines to co-exist and provide independent, data-driven advice to a discretionary PM. We are fortunate to have direct access to talented and experienced PMs who can clearly articulate their investment process. This makes our job easier and increases the probability we can design tools that provide tangible impact.

Our Strong Philosophies, Weakly Held

In this section, we will explain our philosophies and best practices that have served our DS team well and may be useful to your organization, ordered from general to

[2] "Edge", itself, can also be quantitatively measured.

technical. Although we have high conviction in these guiding principles, we routinely revisit and revise them as needed as the team and firm evolve. The Weiss DS team is tailored to a discretionary long/short active manager; we would recalibrate some characteristics if we existed within a purely systematic quant fund, for example. While we cannot claim strict adherence and perfect achievement of the following principles, we find it beneficial to have clear ideologies for self-guidance and benchmarking.

Ideology

Less time modeling, more time asking the right question(s). Nailing the right set of questions to ask and explore with the data is critical. Too many DS projects begin without a clear business objective and become a costly excursion with no clear impact to the bottom line. Asking the right questions requires experience and creativity; this is often the underappreciated *human* influence on the initial stage of every successful ML endeavor[3].

We spend comparatively less time racing to extract alpha signals from the latest expensive "alternative" dataset, and more time helping shape the overall portfolio. This activity has a much greater marginal impact on the organization. There are many ways to integrate alpha within a multi-strategy fund, but the most naïve approach – an incremental strategy slapped alongside 20-plus other strategies – is not going to move the needle and is likely not worth the expense. Later in this paper, I outline a few alternative integration approaches, including 'human-in-the-loop' concepts.

Automate anything that needs to be done more than a few times and enable end-users to answer their own questions by providing interactive dashboards so the DS team is not a bottleneck. This allows our team to scale well beyond our size.

Accessibility / Communication

Be accessible: We are embedded within the trading floor and are widely available. Project ideas and opportunities should not only flow one way from the C-suite and PMs to the DS team; the DS team should also propose projects and initiatives from our unique vantage point. A physically detached DS team, sitting in the corner of a random floor, will give the perception of being accessible only through email, which is highly suboptimal. A major question when constructing an in-house DS team is whether it should be (i) centralized, (ii) fragmented across departments or (iii) hub-and-spoke, a blend of the two. At Weiss, the team reports to the CIO/President and is centralized. This increases resource sharing and communication within the team. However, one approach will not dominate across enterprises, and imitating another organization's integration of DS is destined to be suboptimal. No matter the team's ultimate structure, it is critical that team members be accessible across the firm and able to interact, learn and share tools among one another.

[3] For those interested in this topic, I recommend the book, "A More Beautiful Question: The Power of Inquiry to Spark Breakthrough Ideas," by Warren Berger.

Embrace *effective* communication with non-quants. "Effective communication" is a higher standard than communication. If output from the DS team – often in the form of data-rich presentations/PDFs/web app dashboards – is not self-explanatory to non-quants, it is not actionable and, hence, worthless. Understandably so, most DS job interviews focus a disproportionate amount of time on technical abilities; screening for high emotional intelligence (EI) should not be overlooked[4].

Technical

Keep it simple stupid (KISS). ML modeling used for predictions plays a major role within any DS team, but it is not the only endeavor. Simple descriptive statistics and interactive visualizations (so-called Business Intelligence) can have a great impact on the business, and often require nothing more than high school algebra. These should not be overlooked. Data scientists will usually only admit this after a few drinks, but the majority of impactful calculations we make are remarkably straightforward.

Embrace "Bayesian thinking." What do I know about a given problem? What evidence do I have and how good is it? What alternative explanations might account for the evidence? Like investing, good DS practice embodies nimbleness when constructing one's own "lattice of mental models." That is, expectations formed on past observations (priors) need to be revised to reflect new information to inform our future expectations (posteriors). This mindset embodies how we approach the time-varying markets, as well as the practice of data science itself. This is the hardest principle to adhere to, as old neural pathways can be stubborn to update. Investors are flooded with structured and unstructured data, and if we are honest with ourselves, we'll observe that much of that data will not reconcile nicely with our prior beliefs of how markets *should* work. Models can help here, but if the human design of the models is persistently biased, they become less independent and useful. As a result, wider confidence intervals on all strongly held views are likely prudent.

Some predictive power can be sacrificed for greater model interpretability. We are not black box traders, and it is important for those who use our models to have sufficient transparency to trust results.

Leverage modern tools: This means an open source stack, with R and/or Python being the workhorses for data analysis. We emphasize consistent syntax (tidyverse, pandas), reproducibility (R Markdown, Jupyter Notebooks), version control and collaboration (git). We strive to not reinvent the wheel, aggressively leveraging open source libraries. The ability to quickly customize and prototype using modern technology often outweighs the collective costs of outside commercial solutions[5].

[4] Sometimes referred to as "EQ", comprises four domains: self-awareness, self-management, social awareness, and relationship management.

[5] There is always a trade-off between outsourcing technologies and building proprietary in-house. Decisions must incorporate honest assessments of internal and external comparative advantages and budget constraints. Most of our operating expenses are allocated to raw datasets, not data analysis tools or commercial analytics.

If code is not inside our version control, it does not exist. This ensures all contributions are transparent and intra-team responsibilities are highly interchangeable. Cutting-edge tools are evolving rapidly, requiring a flexible mindset and willingness to constantly learn. Time is budgeted to revisit working software and, if needed, recode with more efficient libraries. The technology stack used today will not be the stack used in the future. Wall Street is full of stale tools that do not meet modern needs and, rightly so, do not entice young DS talent either.

The team continues to evolve and expand. Current and future projects include but are not limited to (i) extracting greater metadata from discretionary PMs/analysts, (ii) sophisticated behavioral modeling and (iii) adding more strategy-specific ML tools specifically for alpha generation.

Data Scientist != Quant ?

There is no widely agreed distinction between a "data scientist" and a "quant," but I will take a stab. Without a doubt, the Venn diagram of these two honorable cohorts has considerable overlap. In my mind, the difference is small but noteworthy.

A well-constructed DS team should directly report and work closely with the C-suite. Its mandate should be very broad, spanning multiple departments. Wide DS team scope helps identify potential inefficiencies. Providing access to disparate datasets can lead to richer modeling and deeper insight (ones that will not be evident to siloed departments/teams). A Chief Data Scientist needs to understand the trajectory of the firm in order to allocate resources and prioritize tasks within the team. This will, in turn, influence the trajectory of the firm. For example, at Weiss we spend approximately 90% of our time on our investment process and 10% on what I call operational efficiency (which is a catch-all phrase for introducing data driven tools and mindset across other divisions).

Quants, on the other hand, are typically known to have a comparatively narrower role: to generate alpha signals and translate them into a portfolio as a stand-alone strategy or blend signals into a broader portfolio. Both roles are similar in that they require a great deal of curiosity, mathematics, technology and capital markets knowledge but data scientists must be willing to interface with many different personalities and objectives and approach a wider set of problems.

(Human) Predictions

The remainder of this paper is a set of predictions on where the industry is headed. They are output from a human, without any supporting data, formulas or visualizations, so decide for yourself if discounting is needed!

The industry will no longer be "quant vs. fundamental." An improved bifurcation will be "systematic vs. discretionary," with the widespread acceptance that nearly all discretionary investors will be more quantitative. Many of the successful ones already are. Active investing has always been about data analysis, and the rewards have been disproportionally reaped by those who analyze the data most accurately. Data Science is a suite of tools to formalize, automate and expand the process.

Discretionary investors will live or die by their skill to assess which return drivers for a given security are most relevant in the present and future (not the past). Recognizing when the world has changed, and which associated factors will no longer be predictive, regardless of past usefulness, will be a hallmark talent enabling discretionary managers to outperform the disciplined, consistent and often higher-breadth strategies of systematic investors[6]. Allocators will continue to want exposure to both active approaches in their diverse alternatives bucket, as neither approach will dominate the other.

The industry will no longer be "systematic vs. discretionary," for that matter. It is likely the blurring between these two investment approaches will accelerate, becoming less pronounced in the future. Managers who can generate the best forecasts *using all available data* – regardless of how they are branded in the marketplace – will garner the most attention from allocators.

Often, the differences between the two investment approaches are stressed, making it easy to forget just how much the two styles have in common. For one, systematic investors have long overridden their models during extreme times[7]. A data scientist uses *judgment* and imposes strong assumptions when selecting an ML model (e.g. linear relationships). The model is then trained on datasets *from the past*. Human judgment deployed by discretionary investors is simply a (messy, error-prone) amalgamation *of past experiences*.

Many quant models still in production are based on the exact principles used by fundamental analysts, just automated and expanded across a larger universe and applied in a disciplined, consistent manner. In other words, they are not as data- or model-driven, as one might expect. Marketing materials for many systematic strategies stress some underpinning economic theory to bolster investor confidence.

Importantly, we do not think all active investors will become systematic; instead we think that all discretionary investors will become more reliant on analytics to augment their investment process. It's not 'quantamental', it's simply discretionary investing in 2018.

Humans will become more important, not less, and they will become more model- and data- literate. The demand for model builders will increase as the world becomes more model-driven. What gets less attention is the parallel increase in demand for discretionary investors who can *effectively interpret models and exploit their output*. ML will become less "magical" as its basic principles (e.g. how a classification model differs from a regression model) become absorbed and accepted as common knowledge[8]. Nearly all interns interviewed for our firm-wide summer analyst program understand that they need to know how to analyze data in Python or R in order to be successful in whatever role they end up in on Wall Street (in other words, Excel proficiency will no longer cut it).

Whether a new role emerges to fill this need[9] or existing discretionary investors

[6] That said, the persistent underperformance in 2018 of some systematic portfolios that implement well-known, fairly rigid, crowded strategies suggests that this forecast need not exclusively apply to discretionary investors.

[7] Some quants proudly mention these difficult decisions to override as a true benefit to investors; others sheepishly wish the question wasn't asked, as if the admission to overriding implicitly acknowledges a flaw in their underlying model.

[8] "CFA finance exams to grill hopefuls on AI, big data and robo-advice", FT, May 10, 2017.

[9] One example put forth by Jefferies Prime Services is the "Translator" role, who "makes complex issues digestible to multiple consistencies". See "The State of Our Union 2018: Too Much Information, Not Enough Intelligence," February 2018.

adapt and augment their existing skills, the expertise will become more pervasive (creative destruction will not be kind to those without it). Regardless, understanding model limitations is an edge.

Tighter integration between data scientists and discretionary investors. If the above forecast proves accurate, collaboration will become near seamless, looking less like a scene out of "Revenge of the Nerds" and instead simply an elite investment team. In the present, data scientists are often siloed teams, receiving all of the acclaim (or blame!) based on their model's accuracy – whether it's used in isolation or integrated into a bigger investment process. In the future, as a greater number of model-literate, savvy discretionary investors engage more deeply with models, the benefits will be shared among a greater number of investment professionals and end investors.

For example, at Weiss, we continue to refine approaches to best integrate ML models and human judgment (earned through years of investment experience) to optimally allocate capital across internal strategies. How can we exploit the comparative advantages of both humans and ML? This work is far more art than science.

Human-in-the-loop. The current state of model deployment in production is often rigid and uncreative. Often, the model's output is derived and communicated, then acted on directly or fed into other "models" (e.g., formal risk model optimization or informal mental models of a human).

In the future, investment processes – as formal or informal as they may be – will likely integrate models and human judgment more tightly, which makes sense if one believes both cohorts have comparative advantages. This approach is sometimes referred to as human-in-the-loop and will continue to become more prevalent.

Some monolithic models will be replaced by many modular ones. Each sub-model might be individually more modest in scope, but collectively provide superior advice. Some of these intermediate models will take the input from an experienced human. This chain of interactions may continue for a few repetitions before a final actionable output is generated. Attribution to model and human can be made rigorous and will be crucial for making improvements[10]. Both sides aid each other – an activity humans excel at and have been doing for centuries (e.g. a modern take on hunting with eagles in Mongolia!).

Prediction targets will expand. The nature of prediction targets will evolve and expand. There will be less emphasis on predicting forward excess returns of securities (either in the cross section or time series). There will be greater recognition that time spent researching and forecasting attributes related to price action, but not price action directly, can prove fruitful. Examples include revenue, EPS and sentiment from investor call transcripts. This shift has already taken place, but will accelerate.

[10] For example, one can easily imagine a discretionary analyst being aware that an influential industry expert is speaking in two days at noon. This unstructured data is less likely to be captured within a traditional set of factors used in an ML model. The analyst could take this information and widen confidence intervals around model's security in question leading up to the event. The model, in turn, can take this human input (along with others) and revise its own forecasts. An analyst can easily identify structural breaks in a company's business, overriding slower moving factors that update via delayed 10-Qs, for example.

Black box is becoming less black box. There is exciting progress being made in the data science community advancing the modeler's ability to interpret intermediate steps of ML algorithms – particularly those algorithms most criticized by their lack of interpretability[11]. This trend should benefit investors as well as the modelers. Pure systematic strategies will provide greater transparency to their investors. Discretionary investors interfacing to models in a human-in-the-loop framework will likely be able to exploit increased model transparency. Furthermore, models will become more explicit and honest when they are unable to forecast (which can be due to a host of reasons, most commonly in-sample inputs differing materially from a training set). To the end user, receiving an answer of "I do not have a confident answer" is a feature, not a bug. To varying degrees, this has always existed, but it should become more commonplace.

More data, but different data. Unique datasets have always been valuable to managers. Sometimes this information is *unstructured* (e.g. the number of cocoa beans counted on an African farm relayed over the phone), sometimes it is *structured* and delivered seamlessly into your local database (the latter is less likely to be "unique"!). Many expensive commercial datasets labeled as "alternative" will just be called "data." In fact, many of them are already commoditized when used in isolation. Managers will look inward for data that is truly unique to their organization. As a higher turnover multi-strategy fund with many positions, we have a lot to work with. Metadata (data describing data) will continue to flourish.

Women's representation on DS teams will continue to grow. This is the easiest forecast to make, given the dismal baseline. To meet the wide ranging scope of DS teams, diversity of thought and experience is critical. Inclusive online groups such as R-Ladies[12] and PyLadies suggest trends are moving in the right direction.

Asset owners will all have their own DS teams. Greater resources will be allocated to the creation and growth of in-house DS teams within allocators, even resource strapped ones. These teams will work directly with the CIO and Investment Committee. This will come at the expense of some traditional roles that will be automated within the organization or outsourced to third-party providers. Furthermore, consultants will also increase quantitative analysis to extract maximum insight from the (often low-frequency) data at their disposal.

Less linear modeling, more non-linear modeling. Cornerstone theory in finance assumes linear relationships. The Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT) are probably the two most widely studied examples. Multivariate ordinary least squares (OLS) linear regression has been the workhorse for practitioners for both alpha generation and risk management (and will remain in production for decades buried deep within legacy software across Wall Street). Linear models are not going away, but ML is helping dramatically improve

[11] Deep Reinforcement Learning is one example that immediately comes to mind. See, "Toward Interpretable Deep Reinforcement Learning with Linear Model U-Trees," by Liu et al., July 16, 2018.

[12] R-Ladies is a "world-wide organization to promote gender diversity in the R community."

these models in many contexts (e.g. see robust/lasso/ridge regression techniques). However, using greater amounts of data and computing power, nonlinearities can be better captured. Non-linear models may have better success capturing complex dynamics, particularly in the extremes where linear can fast become non-linear (e.g. liquidity degradation is not linear; there are tipping points when liquidity declines dramatically). Two examples of non-linear methods that will continue to grow in importance are (i) ensemble learning methods using decision trees and (ii) deep learning in the time series domain. Non-linear approaches are not a panacea, however. If the underlying process changes from the past, there is little reason to believe their forecasts will outperform their linear cousins.

The pain has yet to set in as big data becomes bigger. When pressed, many data scientists will sheepishly admit that most of their work continues to be done on small-ish datasets (< ~10 GB). Many traditional DS development stacks in R or Python begin to fail in the medium data range (~10 to ~100 GB). Truly big data (>~100 GB to 1 TB) requires distributed tools like Hadoop and Spark. To be clear, there is already lots of pain felt modeling big data, but it's soon going to get much worse with the exponential growth of data.

Formalization and unification of DS tools will continue. DS is a rapidly changing space. Tools learned within the last two years commonly become obsolete due to new technologies. In the open source world, the barrier to entry for emergent, disruptive technologies is low, and the adoption of new innovations is quick. This makes the space fun and challenging, but difficult for businesses to adequately support.

While this is taking place, the vocabulary of DS continues to formalize across different data wrangling[13] tools. Essentially, the core data structure for most analysis is the data frame or tabular object that supports different datatypes, such as dates, strings and real valued numbers (think Excel workbook without any of Excel's flaws). Up to now, data frames have been largely non-portable[14] across technologies. That is quickly changing thanks to nascent open source technology such as Apache Arrow, which is helping to "defragment" data access. Apache Arrow is a cutting-edge effort that is a cross-language development platform for in-memory analytics. It will likely become the native format for in-memory analytics.

We are witnessing similar trends when it comes to applying ML algorithms to large data sets (exceeding the size of local memory). Historically, the majority of ML algorithms applied to small and medium-size datasets were implemented in the same language used for data munging and ran locally within memory. For example, if you used Python, you used scikit-learn; if you used R, you used any of the thousands of ML packages on CRAN[15]. Now modelers can stick to their favorite languages for data wrangling and visualization and farm out analysis into clusters that use frameworks like Apache Spark ML for general ML, or Google's

[13] Sometimes referred to as "data munging," this action includes many of the common transformations applied to raw data to alter it for seamless downstream purposes, such as visualization and modeling.

[14] In an acceptably efficient manner, at least.

[15] Technically, nearly all of these packages serve as wrappers to lower-level languages like C/C++ or, gasp, Fortran.

TensorFlow for deep learning. The implementation of the underlying ML algorithm has been entirely abstracted away.

In other words, "front ends" for data analysis will continue to decouple, and users will converge to the best available tool that suits their preferences. Open source languages are borrowing the best approaches from each other. This means more DS time can be freed to think about the problem, less time with code. From a Chief Data Scientist's view, this also means easier integration within a team and greater accommodation of different preferences from top data scientists.

Finally, formalization of roles within a Data Science team will lead to increased specialization. A DS team consisting of one employee is a thing of the past. As technologies advance and complexity grows, I see a trend toward specialization within the world of DS. The following set is a non-exhaustive list of skills that are needed in a modern DS team: big data storage, calibration of Spark clusters, mathematical theory of ML (includes formal knowledge of algorithms, statistics, probability, linear algebra, optimization), Linux system administration (locally and in the cloud), interactive data visualization, UX design, *not to mention actual domain specific knowledge* (e.g. capital markets dynamics, investment strategies, portfolio optimization techniques, etc.).

Most DS practitioners possess many of these skills, but as each skill becomes more demanding and mature, it is more efficient to delegate responsibilities to specialized teammates. Recently, we have seen the emergence of a formal distinction between a "data scientist" and "data engineer," for example, but roles within a DS team will become far more granular. As responsibilities are efficiently distributed within a DS team, it is critical that all DS teammates maintain a curiosity and high level understanding of the other skills represented across the team, as decisions made in one area often impact all other areas within a successful data-driven endeavor. Furthermore, it is the responsibility of the chief data scientist to make sure that this increased specialization does not produce isolation and decrease communication within a DS team. Instead, the greater diversity should bolster spirited debate and increase overall quality control, improving the team's output.

Conclusion

We hope this paper provides some visibility into our approach and thoughts on the direction on the industry. Despite all the hype[16], the supply of data scientists has not yet met demand[17]. While our tools continue to formalize, expand and improve rapidly, the industry is still in its early innings. Firms continue to struggle with how best to integrate a data-driven process. Each enterprise is unique and warrants a custom solution that is tailored to meet its biases, history and objectives[18].

[16] "Data Scientist: The Sexiest Job of the 21st Century," Harvard Business Review, October 2012.
[17] "August LinkedIn Workforce Report: Data Science Skills are in High Demand Across Industries," LinkedIn, August 2018.
[18] There are many wonderful (and free) resources available for those that want to follow DS/ML trends. In particular, I recommend joining the following newsletters: "O'Reilly Data Newsletter", "R Weekly - Weekly Updates from the R

We like our approach, and it has proven successful; but we also firmly believe we must continue to evolve and revisit its efficacy. Like the pursuit of alpha, we must continue to adapt and improve; otherwise we will be run over.

Yes, Wall Street will look different in the future (this has always been the case). There will likely be less overall head count (but not as little as some are calling for as new roles are being created). There will undoubtedly be a shift in skillsets. Everyone will have an increased comfort with data, multidimensional visualizations and the basics of modeling. That's not a bad thing for nimble managers or asset owners.

Acknowledgments: I'd like to thank the members of the Weiss DS team for their thoughtful feedback and suggestions on this paper. I also want to acknowledge their immense contributions within the team and firm at large. All predictions that prove correct will likely have come from them, while the embarrassingly wrong ones will surely have originated from me. I would also like to thank Bill Poutsiaka for his valuable comments.

Community," and "Data Elixir." Contact me directly for a list of my favorite podcasts, which are too numerous to list here.

AUTHOR
Charles S Crow IV
ccrow@gweiss.com

NEW YORK OFFICE
320 Park Avenue, 20th Floor
New York, NY 10022

HARTFORD OFFICE
One State Street, 20th Floor
Hartford, CT 06103

INQUIRIES
Gillian Tullman
Director of Investor Relations
and Marketing
gillian.tullman@gweiss.com
+1 212 390-3451

WEBSITE
http://www.gweiss.com