New Constructs®
Diligence | Independence | Performance

DILIGENCE PAYS 1/26/2018

AI Has a Big (Data) Problem (3 of 5)

The total size of all global data hit 20 zettabytes in 2017. For 99% of people, that number probably means nothing, so picture this: if every 64-gigabyte iPhone were a brick, we could build 80 Great Walls of China with the iPhones needed to store all the world's data.

We are awash in an ocean of data that grows bigger by the second. And it's a complete and utter mess.

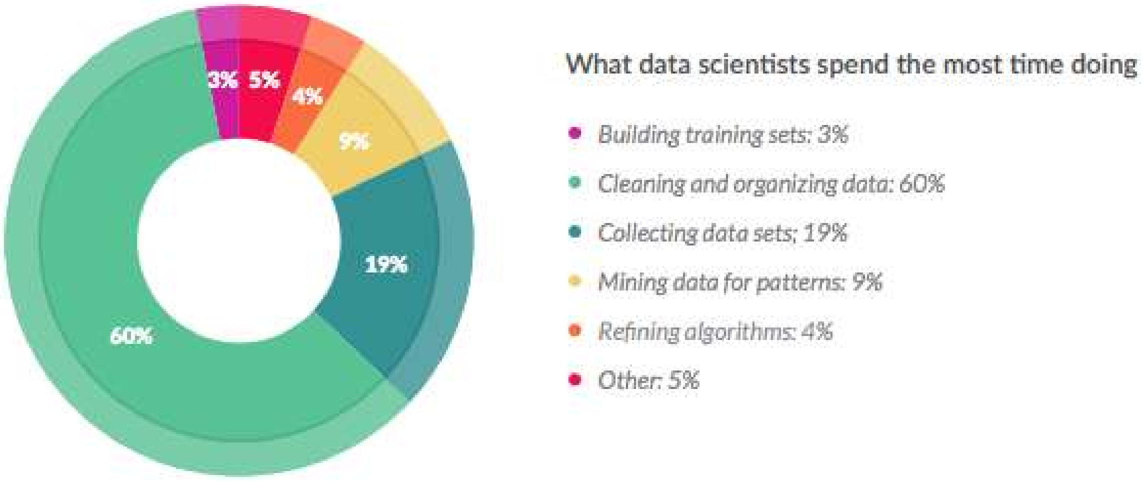
90% of web data is unstructured, meaning it's in a format that cannot be easily searched and understood by machines. Poor data quality costs the US economy $3.1 trillion a year according to IBM (IBM). We have become a society that is excellent at producing, storing and sharing data, but we're lousy at making it useful.

Poor data quality represents the single largest hurdle for developing useful artificial intelligence. It doesn't matter how "smart" machines become if they're fed data that is inaccurate or incomprehensible.

The Size of the Data Management Problem

Poor data quality is a familiar problem for those who analyze data for a living. A recent survey found that 60% of data scientists devote the majority of their time to cleaning and organizing data, as shown in Figure 1.

Figure 1: Cleaning Data Takes the Most Time



Sources: CrowdFlower

Comparatively, just 9% of data scientists devote the bulk of their time to mining data for patterns. Cleaning and organizing data has become such a big task that it leaves precious little time for analysis.

When people predict that AI will make human workers obsolete anytime in the near future, they are ignoring the data quality problem. AI and machine learning may be able to replace those 9% of data scientists who are mining data for patterns, but it will still need the 80% working on collecting, cleaning and organizing data.

More importantly, data scientists need to re-orient their thinking around data quality. More time needs to be spent upfront and on collecting data in a high-integrity manner rather than retroactively "cleaning" it. We are not sure that it is possible to retroactively clean data enough to meet the needs of useful AI. If you cannot validate the data back to its source, then how do you know it is clean? And, if you are going back to the source to validate, then you might as well collect it from the source.[1]

[1] Harvard Business School features the powerful impact of our research automation technology in the case New Constructs: Disrupting Fundamental Analysis with Robo-Analysts.

Gradually, leaders in AI seem to be understanding that their job is as much, if not more, about data management as it is about new technologies. Facebook (FB) just hired Jérôme Pesenti, the former head of IBM's Big Data group, to run its AI efforts. Pesenti replaces Yann LeCun, who will now focus on his core expertise of research.

While research into deep learning and neural networks makes the most headlines, Facebook understands that data management is crucial to delivering bottom-line business improvements.

AI Has a Long Way to Go

Once you understand the limitations imposed by poor data quality, the challenges facing AI become much clearer. Successes like the Go-playing computer that beat the world champion are misleading because they are enabled by completely structured and easily interpreted data points. Accordingly, the subset of tasks that AI can perform effectively remains small, according to LeCun:

> *"In particular areas machines have superhuman performance, but in terms of general intelligence we're not even close to a rat."*

Other researchers have faced similar hurdles. Machines can handle finite and discrete data points well, but even a minor degree of ambiguity can trip them up.

> *"We are still a long way from computers being able to read and comprehend general text in the same way that humans can."*

The last quote comes from Microsoft (MSFT) CTO, Kevin Scott, in a LinkedIn post celebrating the development of an AI that could read and answer questions about Wikipedia pages at the level of an average human. Despite its success, the machine struggled when asked to go beyond simple facts and make intuitive, but still logical, leaps. Contextual clues that a human would understand easily are still incomprehensible to machines.

This problem is even more pronounced when it comes to using machines to read financial filings. One of the biggest issues people face when trying to use natural language processing on financial filings is that the language in these documents is far from natural. If machines can get tripped up by Wikipedia, imagine how they respond to the jargon and legalese that fill your average 10-K.

In the financial world, as in the rest of the economy, AI has had its biggest successes working with data that is structured and standardized. JPMorgan Chase (JPM) has automated hundreds of thousands of hours of work annually by developing machine learning tools to read commercial-loan agreements. These contracts are standardized, which means the machine only has to navigate minor differences.

While machines are replacing humans in these rote tasks, they struggle to make the leap to more sophisticated analysis. Rather than try to tackle the complex task of reading and analyzing financial filings, most applications of AI in the investing business focus on mining patterns out of trading data, alternative data sets, or sentiment and other non-financial indicators. As we discussed in our first piece in this series, these efforts have not yet led to superior returns. It's been easier to apply existing AI and machine learning tools to new data sets than it has been to teach AI to analyze old data sets, especially data directly from SEC filings.

Structuring Financial Data: Not as Easy as Most Think

In theory, financial data in filings would be more structured and standardized, or we could make it that way easily. We have centralized bodies (FASB, SEC) that govern financial reporting standards. Public companies employ teams of accountants and lawyers to conform to these standards.

In reality, the data remains highly unstructured and variable, and we expect that it will only get worse. The most prominent effort to make financial data machine readable, XBRL, remains riddled with errors 10 years after its initial deployment. While companies are required to submit XBRL filings, they're not required to verify them, and only 8% of companies carry out voluntary audits. Until XBRL is strictly enforced by the SEC, it does not stand a chance at being reliable.

Without SEC enforcement, most companies will continue to under invest in the resources necessary to get their XBRL filings right. They will either leave the task in the hands of accountants that lack the technical expertise to do the job, or they outsource the job to a third-party that doesn't fully understand the company's financials.

Again, we come back to the disconnect between people with technical expertise and those with subject matter expertise in finance.

As long as XBRL and other efforts to structure financial data are treated as curiosities that companies can safely ignore, it will be almost impossible to make meaningful progress on this front.

Structuring Data: More About Team Than Technology

As long as financial data remains unstructured, existing machine learning tools cannot process it effectively. Meanwhile, the cost of employing the highly-trained analysts needed to manually structure data remains prohibitive.

Our solution is to leverage our deep financial expertise into software that enables highly-trained analysts to collect and structure data with unrivaled efficiency. In essence, we arm human subject matter experts (SMEs) with technology at every step of the data collection and modeling process. For example:

- Analyst and programmers work together to build the AI. Analysts and programmers anticipate and address problems from multiple perspectives from the outset. Clear communication between financial and technical experts is critical to building machines that work. Anticipating potential problems at the start, along with frequent iteration and joint and rigorous testing, helps build machines that robustly do something small. One step at a time, we teach machines to perform discrete tasks perfectly. However small that step may be, each step means less work for human SMEs. We don't send programmers or data scientists to analyze the data in isolation.

- During data collection, our process leverages a multitude of sophisticated algorithms to validate the data points collected by the machine so humans can transcend most of the banal work. We use our big data experience and financial expertise to automatically identify data that's potentially wrong – from values that are too big or too small, to data points that show up in the wrong places, to data relationships that don't make financial sense. Analysts only have to focus on issues the machines have not already mastered. We also track how analysts address each issue so that if it recurs, the machines can handle it automatically.

- Our data collection process includes sophisticated corporate performance and valuation modeling of the data that produces highly-respected investment ratings and research. Our financial expertise enables us to create quality assurance algorithms that flag modeled results that are unusual or have been linked to errors in the past. This process adds significant integrity to our data collection process compared to traditional data collection processes by humans who are not experts in accounting or finance or do not have a model to help them analyze the impact of the data they collect.

- To teach a machine well, we need to think like machines. Accordingly, every data point in the 120,000 filings parsed into the machine by human SMEs is tagged with 10+ pieces of unique identifying information. These tags include data value, the associated text, the location in the filing, and many other features that are taken for granted by most humans but provide critical context for the machine.

- The scale and efficiency of our process has a virtuous effect on our data validation processes. The more models we can build, the more potential data anomalies or errors we can find and feed back into the machine. The more we do, the more we can teach the machine, and, in turn, rely on it to do more. This approach gives us significant advantage over systems or analysts who can only view a few models at a time.

Working with machines presents many new challenges to our society. It is not something we've done before and, not surprisingly, we have a lot to learn, and so do the machines. One thing we know for sure is that people that are best at teaching machines will have the best machines, and the people with the best machines will have the upper hand.

This article is the third in a five-part series on the role of AI in finance. The first, "Cutting Through the Smoke and Mirrors of AI on Wall Street" highlights the shortcomings of current AI in finance. The second, "Opening the Black Box: Why AI Needs to Be Transparent" focuses on how transparency is crucial to both developers and users of AI. The last two articles will show how AI can lead to significant benefits for both financial firms and their customers.

This article originally published on January 26, 2018.

Disclosure: David Trainer and Sam McBride receive no compensation to write about any specific stock, sector, style, or theme.

Follow us on Twitter, Facebook, LinkedIn, and StockTwits for real-time alerts on all our research.

New Constructs® - Research to Fulfill the Fiduciary Duty of Care

Ratings & screeners on 3000 stocks, 450 ETFs and 7000 mutual funds help you make prudent investment decisions.

New Constructs leverages the latest in machine learning to analyze structured and unstructured financial data with unrivaled speed and accuracy. The firm's forensic accounting experts work alongside engineers to develop proprietary NLP libraries and financial models. Our investment ratings are based on the best fundamental data in the business for stocks, ETFs and mutual funds. Clients include many of the top hedge funds, mutual funds and wealth management firms. David Trainer, the firm's CEO, is regularly featured in the media as a thought leader on the fiduciary duty of care, earnings quality, valuation and investment strategy.

To fulfill the Duty of Care, research should be:

1. **Comprehensive** - All relevant publicly-available (e.g. 10-Ks and 10-Qs) information has been diligently reviewed, including footnotes and the management discussion & analysis (MD&A).
2. **Un-conflicted** - Clients deserve unbiased research.
3. **Transparent** - Advisors should be able to show how the analysis was performed and the data behind it.
4. **Relevant** - Empirical evidence must provide tangible, quantifiable correlation to stock, ETF or mutual fund performance.

Value Investing 2.0: Diligence Matters: Technology is Key to Value Investing With Scale

Accounting data is only the beginning of fundamental research. It must be translated into economic earnings to truly understand profitability and valuation. This translation requires deep analysis of footnotes and the MD&A, a process that our robo-analyst technology empowers us to perform for thousands of stocks, ETFs and mutual funds.

DISCLOSURES

New Constructs®, LLC (together with any subsidiaries and/or affiliates, "New Constructs") is an independent organization with no management ties to the companies it covers. None of the members of New Constructs' management team or the management team of any New Constructs' affiliate holds a seat on the Board of Directors of any of the companies New Constructs covers. New Constructs does not perform any investment or merchant banking functions and does not operate a trading desk.

New Constructs' Stock Ownership Policy prevents any of its employees or managers from engaging in Insider Trading and restricts any trading whereby an employee may exploit inside information regarding our stock research. In addition, employees and managers of the company are bound by a code of ethics that restricts them from purchasing or selling a security that they know or should have known was under consideration for inclusion in a New Constructs report nor may they purchase or sell a security for the first 15 days after New Constructs issues a report on that security.

DISCLAIMERS

The information and opinions presented in this report are provided to you for information purposes only and are not to be used or considered as an offer or solicitation of an offer to buy or sell securities or other financial instruments. New Constructs has not taken any steps to ensure that the securities referred to in this report are suitable for any particular investor and nothing in this report constitutes investment, legal, accounting or tax advice. This report includes general information that does not take into account your individual circumstance, financial situation or needs, nor does it represent a personal recommendation to you. The investments or services contained or referred to in this report may not be suitable for you and it is recommended that you consult an independent investment advisor if you are in doubt about any such investments or investment services.

Information and opinions presented in this report have been obtained or derived from sources believed by New Constructs to be reliable, but New Constructs makes no representation as to their accuracy, authority, usefulness, reliability, timeliness or completeness. New Constructs accepts no liability for loss arising from the use of the information presented in this report, and New Constructs makes no warranty as to results that may be obtained from the information presented in this report. Past performance should not be taken as an indication or guarantee of future performance, and no representation or warranty, express or implied, is made regarding future performance. Information and opinions contained in this report reflect a judgment at its original date of publication by New Constructs and are subject to change without notice. New Constructs may have issued, and may in the future issue, other reports that are inconsistent with, and reach different conclusions from, the information presented in this report. Those reports reflect the different assumptions, views and analytical methods of the analysts who prepared them and New Constructs is under no obligation to insure that such other reports are brought to the attention of any recipient of this report.

New Constructs' reports are intended for distribution to its professional and institutional investor customers. Recipients who are not professionals or institutional investor customers of New Constructs should seek the advice of their independent financial advisor prior to making any investment decision or for any necessary explanation of its contents.

This report is not directed to, or intended for distribution to or use by, any person or entity who is a citizen or resident of or located in any locality, state, country or jurisdiction where such distribution, publication, availability or use would be contrary to law or regulation or which would be subject New Constructs to any registration or licensing requirement within such jurisdiction.

This report may provide the addresses of websites. Except to the extent to which the report refers to New Constructs own website material, New Constructs has not reviewed the linked site and takes no responsibility for the content therein. Such address or hyperlink (including addresses or hyperlinks to New Constructs own website material) is provided solely for your convenience and the information and content of the linked site do not in any way form part of this report. Accessing such websites or following such hyperlink through this report shall be at your own risk.

All material in this report is the property of, and under copyright, of New Constructs. None of the contents, nor any copy of it, may be altered in any way, copied, or distributed or transmitted to any other party without the prior express written consent of New Constructs. All trademarks, service marks and logos used in this report are trademarks or service marks or registered trademarks or service marks of New Constructs.