Trend 3

DATA VERACITY

The Importance of Trust

Business is more data driven than ever, but inaccurate and manipulated information threatens to compromise the insights that companies rely on to plan, operate, and grow. Unverified data is a new type of vulnerability—one that every business leveraging digital technologies must address. Left unchecked, with autonomous, data-driven decision-making increasing across industries, the potential harm from bad data becomes an enterprise-level existential threat.

Thirty-five years ago, Soviet watch officer Stanislav Petrov jumped out of his chair.[1] According to the satellite system he was monitoring on September 26, 1983, the United States had launched a nuclear missile at the Soviet Union. Protocol dictated that Petrov notify Soviet leaders, who would order an immediate counterattack.[2]

Fortunately for the world, Petrov wasn't convinced that the alerts were true. He didn't notify his superiors, thereby preventing a global catastrophe. In making his decision, Petrov considered the satellite system's warning data within its larger context. At the time, experts agreed that the scale of any preemptive attack from the United States would be massive, with additional bomber and attack support. With no other alerts to show such attacks were underway, Petrov knew that the data the system was showing didn't match what was expected. That, combined with his understanding of the risks if he followed protocol, informed his ultimate decision.

The Soviets later determined that their satellites had confused the reflection of sunlight off clouds for a missile launch. By questioning the validity of data, Stanislav Petrov had saved the world from nuclear disaster.

Businesses may not make decisions about launching nuclear missiles. However, 82 percent of executives responding to our Technology Vision survey report that their organizations are increasingly using data to drive critical and automated decision-making, at unprecedented scale. Today, the global economy runs on live information: IDC forecasted global revenues of nearly $151 billion for big data and analytics practices in 2017, up 12 percent from the year before.[3]

Without establishing the veracity, or accuracy, of that data, businesses leave themselves open to a new kind of vulnerability—a threat that's critically overlooked. A recent study estimated that 97 percent of business decisions are made using data that the company's own managers consider of unacceptable quality.[4] The result? Business insights and decisions that are of questionable value at best, and corrupted at worst.

Provenance

Verifying the history of data from its origin throughout its life cycle.

Context

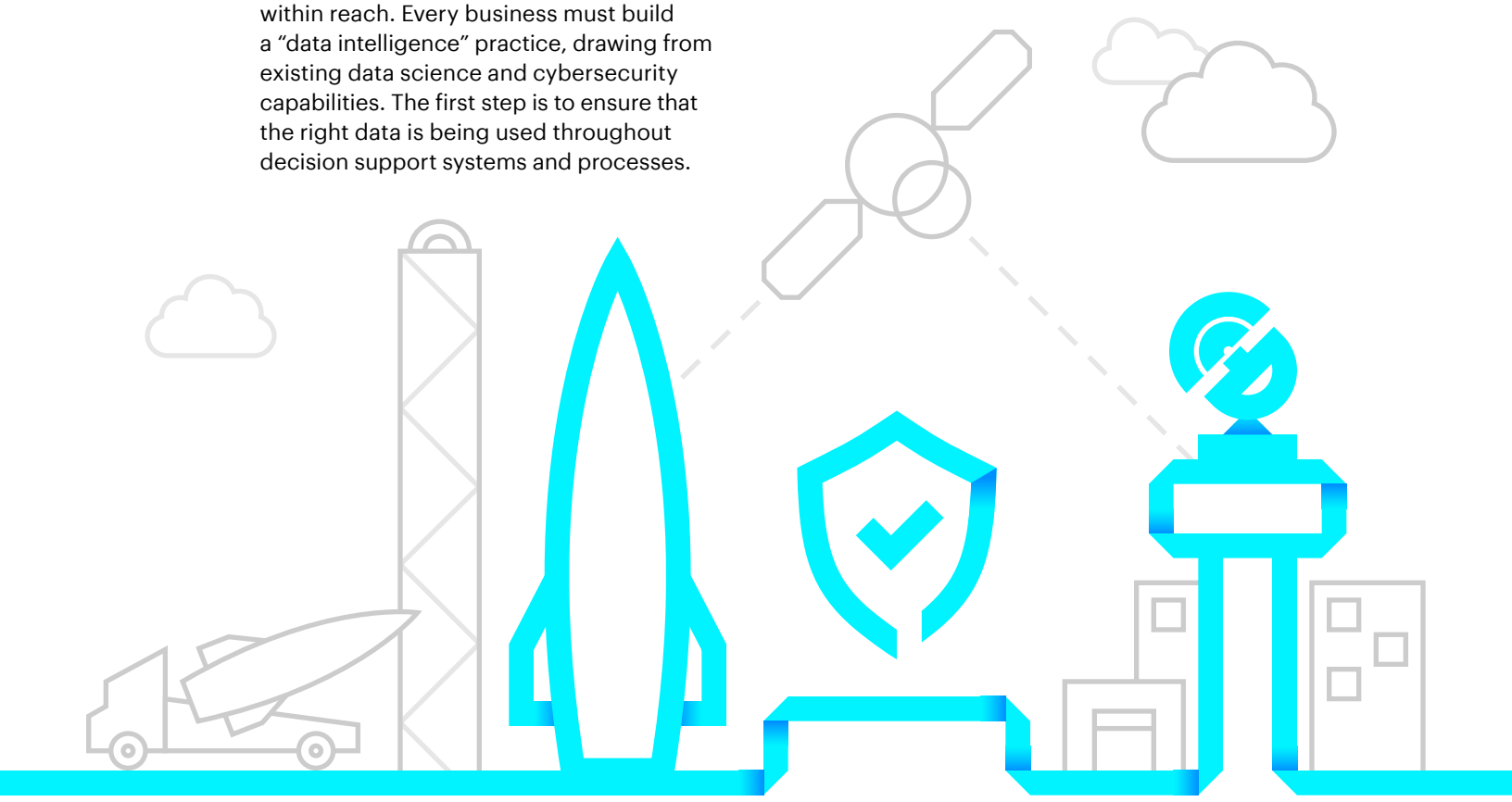
Considering the circumstances around data's use.

Integrity

Securing and maintaining data.

But companies don't need to accept the risks of poor data veracity. They can address this new vulnerability by building confidence in three key data-focused tenets: provenance, or verifying the history of data from its origin throughout its life cycle; context, or considering the circumstances around its use; and integrity, or securing and maintaining data. The skills and tools needed to build this confidence are within reach. Every business must build a "data intelligence" practice, drawing from existing data science and cybersecurity capabilities. The first step is to ensure that the right data is being used throughout decision support systems and processes.

What's more, companies must be vigilant in uncovering and addressing ways stakeholders might manipulate data for their own benefit. As systems from customer-facing apps to robot-run manufacturing floors change their behavior in response to unverified data, every business must answer the question: Where is your Stanislav Petrov?



Risks and Rewards of Data Veracity

Companies around the world are betting big on advances in data-hungry technologies. In 2017 alone, AI investments were projected to reach $12.5 billion, while Internet of Things investments were expected to top $800 billion.[5,6]

Yet without an accompanying push for data veracity, these investments could easily become a sucker's bet. Businesses are spending heavily to determine what they can get out of data-driven insights and technologies, but they also need to invest in what's going *into* them. Even the most advanced analytics and forecasting system is only as good as the data it's given to crunch: as the saying goes, "garbage in, garbage out."

United Airlines realized that inaccurate data was contributing to $1 billion a year in missed revenue. Its seating demand forecasts were based on decades-old assumptions about flying habits, resulting in inaccurate pricing models.[7] The airline highlighted this and other data-driven inaccuracies as key targets for improving operational performance. In an increasingly data-driven world, addressing these risks today will help United ensure that the data underpinning its revenue can be trusted in the future.

The risks around poor data veracity grow as more organizations push toward fully autonomous decision-making, with critical implications for business and society. The US state of Indiana uses an automated system to flag individuals who may be registered to vote in more than one state.[8] It looks at shared names and birthdates: if it finds a "John Smith" born on the same day and year who is registered in both Indiana and Maine, it marks that record as potentially fraudulent. Prior to 2017, these records were submitted for additional review; however, following legislative changes, the system immediately removed flagged individuals from registered voter rolls. With this process, the automated system amplifies data veracity risks: researchers have found that it generates inaccurate fraud alerts 99 percent of the time.[9] The damaging result? The automated removal of legally registered voters, some of whom were simply unlucky enough to have a very common name.

According to our survey, 79 percent of executives agree that organizations are basing their most critical systems and strategies on data, yet many have not invested in the capabilities to verify the truth within it. By making these investments, companies will generate more value from their data, and build a strong foundation for the success of other digital transformation initiatives.

The new "data intelligence" practice will make this possible. Its job will be to grade the truth within data, by establishing, implementing, and enforcing standards for data provenance, context, and integrity.

Creating a Data Intelligence Practice

Businesses don't have to start from scratch to grade the veracity of their data. Some of the most foundational elements of a data intelligence practice revolve around ramping up existing efforts: embedding and enforcing data integrity and security throughout the organization, while adapting existing investments in cybersecurity and data science to address data veracity issues.

The basics, however, will only take companies part of the way. Grading data will also require developing an understanding of the "behavior" around it. Whether it's a person creating a data trail by shopping online, or a sensor network reporting temperature readings for an industrial system, there's an associated behavior around all data origination. Companies must build the capability to track this behavior as data is recorded, used, and maintained. With this understanding, they can provide cybersecurity and risk management systems with a baseline of expected behavior around data.

These baselines will empower companies to detect data tampering that predicates poor decisions. Cutting-edge anomaly detection systems like MIT's AI2 identify abnormal patterns of behavior, then categorize them based on experience provided by human experts. AI2 detects 85 percent of cyber-attacks, and presents the most pressing incidents to experts for review.[10] Industrial giant Siemens is offering anomalous behavior detection for industrial systems to oil and gas customers, by comparing aggregate data generated from sensors onboard its industrial equipment with historical norms and trends.[11]

To mitigate risks around data veracity, SpaceX uses a consensus-based system: each Dragon Capsule uses six computers, operating in pairs, to validate calculations.[12] Each pair checks its calculations against the others', and the spacecraft only proceeds when at least two pairs return the same result.[13]

A company's data intelligence practice must also consider given data within available context—the way Petrov responded when he realized that the attack alert didn't fit with accepted knowledge. Some companies are beginning to use data science capabilities to flag data that deviates from a known broader context. An R&D group at Thomson Reuters has developed an algorithm that uses streams of real-time data from Twitter to help journalists classify, source, fact-check, and debunk rumors faster than before.[14]

Meanwhile, Google is using machine learning to remove apps with overreaching permissions from its Play Store. For example, a flashlight app only needs to activate a smartphone's LED; if a purported flashlight app also requests access to a person's contacts, it wouldn't match the accepted "knowledge" around the permissions needed for a flashlight. The system could then mark the app for further review.[15]

Using the right tools to monitor behavior and context around data's provenance will help businesses mitigate risks that threaten data integrity. With this knowledge in hand, companies can begin to address issues that might be incentivizing deceit in the first place.

Incentivize the Truth

Understanding anomalous behavior will help companies address the threat of false data driving faulty decisions. But a data intelligence practice must also be charged with uncovering and addressing the factors contributing to the creation of false data in the first place. It's an uncomfortable realization, but if a business depends on data collection, they are potentially incentivizing data manipulation.

The presence of bad data in a system isn't always the result of malicious intent, but may be a sign that a process isn't working the way it was intended.

Individual instances of manipulated data may have minimal impact, but a bevy of deceptions can skew business outcomes. Researchers at the University of Warwick have studied the way some rideshare drivers organize simultaneous sign-offs to cause a shortage of drivers, and trigger surge pricing.[16] Knowing that they're participating in systems managed by algorithms, these drivers are trying to make the system work in their favor—at the expense of the ridershare company's efficiency.

Dynamic pricing algorithms, and consumer reactions to them, also demonstrate the growing need for companies to understand motives for disclosing—or disguising—data. Online retailers spend hundreds of billions of dollars each year to advertise and price items online to different segments of people, based on zip code or household income.[17,18] Yet this practice sometimes conflicts with consumer preferences toward privacy. If a large percentage of people attempt to trick these algorithms—or perhaps more likely, do so unknowingly while trying to protect their privacy online—businesses will not only lose money, but also collect inaccurate data about their customers. The end result: more distorted insights.

Already, online shoppers can install browser extensions like TrackMeNot or AdNauseam to generate random queries in the background, or robo-click on ads. These tools obscure a person's real search history and misdirect ad networks.[19] On Amazon, product reviews also became subject to data manipulation: third-party sellers were paying people to submit fake reviews to artificially inflate their product and seller ratings.[20] In this case, Amazon responded by giving more weight to verified reviews from customers who had definitively purchased the item from Amazon. They also established an invitation-only incentivized review program, banning reviews from people who received free or discounted products outside the program's curated process.[21] These efforts reduced the incentive to generate fake reviews on the site.

The presence of bad data in a system isn't always the result of malicious intent, but may be a sign that a process isn't working the way it was intended. Uncovering processes that inadvertently incentivize deceit is a key step to improving the truth in data across a system. Incentivizing truth will allow companies to reduce noise in data, so that real threats stand out. Ultimately, it will help ensure the data is trustworthy enough to drive critical decisions in the future.

Conclusion

CONFIDENCE FOR THE FUTURE

Data is the lifeblood for digital companies, fueling complex business decisions that drive sustained growth. Ensuring the veracity of this data, then, becomes a cornerstone of strong leadership.

Failure to do so can have grave consequences—especially as companies invest heavily in autonomous data-driven systems. Already, researchers have developed techniques that cause machine vision systems from mistaking stop signs for other road indicators, like speed limit signs (see Figure 5, page 48); such systems are used in autonomous vehicles, where fraudulent data like this could cause accidents.[22] And as AI is used to make more business-critical decisions, biased data becomes a larger threat, skewing decisions and corrupting business insights.

Strong cybersecurity and data science capabilities are prerequisites for building a data intelligence practice to ensure data veracity. Among other things, this group will determine the embedded risks across a portfolio of data supply chains, and set standards for how much risk is acceptable based on business priorities and implications of automated decisions. As such, the data intelligence practice should report up to the Chief Digital Officer, and collaborate closely with the Chief Information Security Officer.

Organizing in this way, with a dual mandate to maximize veracity and minimize incentives for data manipulation, will support a business that can be confident in its insights, and alert to new potential threats. Now, every company has a new challenge: ensuring truth in the data that powers its enterprise.

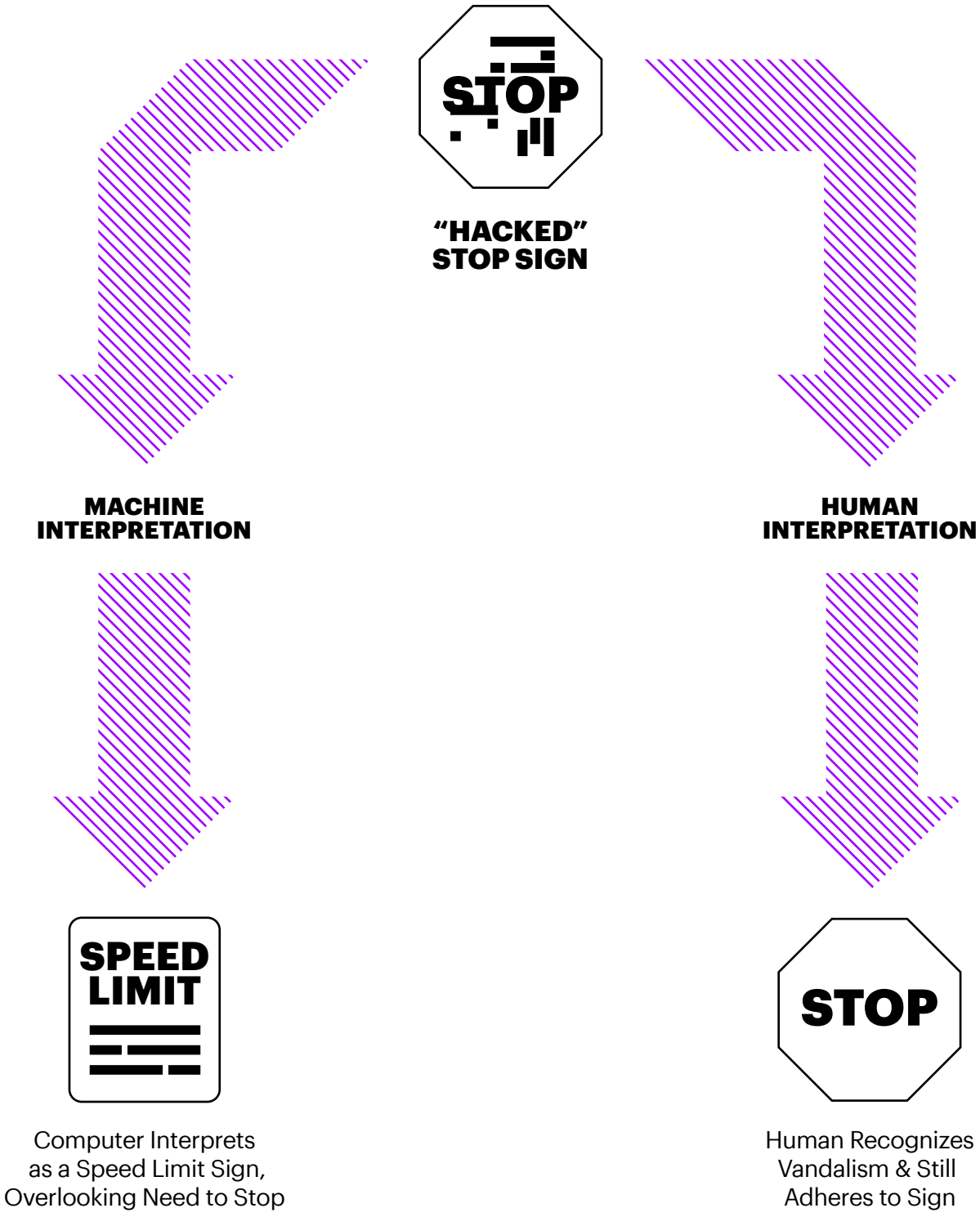


Figure 5—New Technologies Present Threat-vectors that Businesses have never Considered.