HARVARD | BUSINESS | SCHOOL

9-121-006
JULY 7, 2020

SRIKANT M. DATAR
SARAH MEHTA
PAUL J. HAMILTON

Applying Data Science and Analytics at P&G

In December 2019, Guy Peri, Procter & Gamble's (P&G) chief data & analytics officer, slipped into the back of a conference room at the company's headquarters in Cincinnati, Ohio, to catch the last ten minutes of a presentation. He watched as Rachel Breslin, Benjamin D'Incau, and Razi Hyder, colleagues in P&G's oral care business, fielded questions about their data-driven approach to growing P&G's sales in the electric toothbrush market. Using hundreds of economic and demographic variables, D'Incau, a data scientist, had developed a machine-learning model that helped Breslin and Hyder, two business leaders within oral care, become hyper-specific about which dentists to target with P&G's Oral-B® electric toothbrush. Peri was optimistic about the algorithm's potential.

Walking back to his office, Peri reflected on how much progress P&G had made since establishing its data and analytics leadership team in early 2018. This team, led by Peri, helped P&G leverage its data to cut costs and improve outcomes—such as those achieved by the oral care team—across its businesses. But the journey had not been easy. Developing robust data management and governance practices had taken time and investment, and identifying the most effective analytics operating model for P&G had involved trial and error.

Peri and his team had also encountered natural change management issues related to hesitation about changing some long-established work processes within the company. They had tackled this challenge by demonstrating the tangible business results made possible by analytics. "A lot of people get excited about analytics," said Peri, "but we are super clear that this work is all in service of business outcomes. At the end of the day, it's about helping us sustainably deliver total shareholder returns." In pockets of the company, this tactic had proven effective, but reluctance still lingered in several business units. Peri and his team had worked hard to show the value of these new capabilities through top-line and bottom-line business growth. They considered how they might further help P&G transform itself through application of data and analytics capabilities.

Brief Background on Consumer Packaged Goods

The consumer packaged goods (CPG) industry comprised a range of products meant for quick use and regular replacement, such as food, beverages, apparel, and household products. In 2020, consumers in the U.S. were projected to spend more than $700 billion on CPGs.[1] Low switching costs,

---

Professor Srikant M. Datar, Case Researcher Sarah Mehta (Case Research & Writing Group), and Research Associate Paul J. Hamilton prepared this case. The authors received valuable support from Daniel O'Connor. It was reviewed and approved before publication by a company designate. Funding for the development of this case was provided by Harvard Business School and not by the company. HBS cases are developed solely as the basis for class discussion. Cases are not intended to serve as endorsements, sources of primary data, or illustrations of effective or ineffective management.

substantial price competition, and the growth of private label products created a competitive industry. The median revenue growth rate for CPG manufacturers fell from 9.7% in 2011 to 1.2% in 2018.[2]

Procter & Gamble

Founded in 1837 in Cincinnati, Ohio, P&G had become a CPG powerhouse throughout its 182-year history. By 2019, P&G employed 97,000 people and reported net sales of $67.7 billion (see **Exhibit 1** for financials and **Exhibit 2** for a stock chart).[3] Its products were sold in 180 countries and territories to some 5 billion people.[4,5] Nearly half of its brands generated yearly sales of at least $500 million. P&G's portfolio of market-leading brands included Always®, Bounty®, Crest®, Dawn®, Gain®, Olay®, Oral-B®, Pampers®, and Tide®. P&G sold its range of branded products to the end consumer through retailers and distributors, but it was also beginning to experiment with a direct-to-consumer channel. Large retailers were particularly important customers for P&G; sales to Walmart, for instance, accounted for 15% of the company's 2019 sales.[6]

P&G was a market leader in several of its product segments. The company claimed 20% of the global hair care market, 60% of the blades and razors market, 25% of the baby care market, and 25% of the global fabric care market.[7] Given P&G's market leadership and capabilities to help retailers grow their categories, the company had been deemed "category captain" across a number of CPG product types. Category captains, which were often the leading supplier of a given product segment, forged strategic partnerships with retailers to help them manage inventory, merchandising, and product display; in exchange, retailers shared sales data with them.[8]

Internally, P&G organized itself into six Sector Business Units (SBUs), comprising 10 key categories: 1) Baby and Feminine Care; 2) Beauty (Hair Care and Skin & Personal Care); 3) Fabric and Home Care; 4) Family and Ventures; 5) Grooming; and 6) Health Care (Oral Care and Personal Health Care). Two SBUs—Fabric and Home Care and Baby and Feminine Care—generated 60% of P&G's 2019 sales.[9] Market Operations supported the SBUs across the following regions: 1) North America; 2) Europe; 3) Latin America; 4) Greater China; and 5) Asia Pacific, Middle East, and Africa (AMA). Two markets—North America and Europe—accounted for 68% of 2019 sales.[10]

To compete in the crowded CPG industry, P&G aimed to achieve noticeable superiority across product, packaging, brand communication, retail execution, and value proposition.[11] As Peri explained, "If we are noticeably superior in at least four drivers, we consistently win and deliver growth across all business success metrics—sales, profit, value share, household penetration, and category growth. If we are only superior in three or fewer, we are almost universally unsuccessful in driving each of these business success measures." P&G thus invested heavily in these dimensions. The company's 2019 advertising spend of $6.8 billion,[12] for example, ranked it among the top advertisers in the world. Since around 2015, P&G had more intentionally leveraged its data to elevate its performance across these five dimensions (see **Exhibit 3**).

Big Data: The "New Corporate Asset Class"[13]

As the data science and machine learning (ML) revolution took hold through the 2010s, companies began to recognize the untapped value of their data. However, some companies were better equipped to realize the value of that data than were others. Those that originated online (often called "digital natives"), such as Amazon, Facebook, and Google, were further along than their non-digital-native counterparts, including P&G. Digital natives had the added advantage of owning all of their data, whereas many companies, especially those that sold their products through a partner, struggled to access such first-party data.[a] Wrote one observer: "Because CPG brands rely on retailers to sell their goods, they are [. . .] shut out of much of the first-party consumer data that marketers in other industries use to optimize their marketing and demonstrate [return on investment, or] ROI."[14]

Also challenging for non-digital natives, ML models tended to disrupt established work processes. For example, whereas historically, sales strategies might have relied heavily on a given executive's experience, algorithms aimed to supplement executive experience and intuition with objective, data-based perspectives on sales strategies. This shift could breed discontent. Peri explained, "When you introduce something like data science and machine learning into the workplace, natural antibodies emerge." This tension was often exacerbated by the fact that many ML models were "black boxes," meaning their decision-making processes were unknown to the user. People questioned the logic of trusting complex, opaque models rather than their own experience. "At P&G," said Peri, "our approach is to have data and algorithms assist decision-makers. The combination of data insights plus human insights is when we get to our highest-quality decision making."

The hype associated with ML compounded skepticism. As Peri said, "People think that machine learning will solve all of their problems—and of course it will not. We have found that in some cases, machine learning applied properly with sufficient business context has proven useful in improving the quality of decisions." Machine learning models were still subject to errors, bias, false positives, and false negatives. Many ML models also relied on historical data, and because the past was not always predictive of the future, these models might fail to detect new trends or, worse, perpetuate existing biases. These challenges had stalled many CPG companies' efforts to integrate analytics into their work. A 2013 report found that although CPG leaders widely recognized the need to become more fluent in analytics, just 9% had implemented an analytics operating model.[15]

Strengthening P&G's Analytics Muscle

Around 2015, P&G's leadership began to see vast market opportunities in applying data science and machine learning to improve a number of business outcomes. These included increasing the precision and efficiency of P&G's advertising spend, generating granular insights about consumer behaviors and preferences, and improving the effectiveness of the company's trade spend (i.e., payments to retailers in exchange for desirable placement of P&G products and promotional activities). However, to truly reap these potential benefits, P&G needed to strengthen its data science capabilities as well as its data management and governance processes. This would require large-scale cultural change, as well as a significant investment in skilled data scientists and business analysts. The leadership team thought carefully about whether these investments would be worthwhile.

As a first step, P&G needed to decide whether to build these capabilities in-house or to outsource them. Managing them in-house would require hiring data scientists, re-training business leaders, and effecting a massive culture change. This was especially challenging for a non-digitally native company. The challenge went far beyond reorganizing the company's digital resources; there would need to be a paradigm shift in the company's operating model. It was uncertain whether this change would be worthwhile, given that P&G could outsource much of its analytics work to consultancies that specialized in data science. Why risk directing the company's attention away from its core mission?

Despite these concerns, certain leaders at the company believed that its data should be treated as a strategic asset. Developing internal data science capabilities, they argued, would create a competitive

[a] First-party data was data that a company owned. Second-party data was data that stemmed from a partnership (for CPG companies, this was mostly data from retailers). Third-party data was syndicated data that a company purchased.

advantage over other CPG companies. While external partners could help P&G capitalize on its data, the full potential of data science and machine learning would never be realized unless these capabilities were directly incorporated into individual SBUs. To manage the integration of these capabilities into P&G's business, senior leaders promoted Peri in April 2015 to chief data & analytics officer.

Managing P&G's Data

The next step for P&G was creating a strategy for data management. The company had access to a wide range of disparate data. Its first-party data included product shipments to retailers and information that consumers entered into a P&G mobile app. One example was the popular Pampers® rewards app, wherein consumers received points for every purchase of a Pampers® product, which they could redeem for discounts on future purchases. The types of information P&G gathered from the app included consumers' age and gender, geographic location, purchase frequency, preferred retailer, and method of payment. As P&G began experimenting with data science applications, the company considered ways to provide consumers with meaningful value in exchange for collecting their data.

Peri hoped to grow the company's first-party data over time. "Every coupon we give should come back with some data about the consumer so that we can build out our understanding and insights into their behaviors," said Peri. P&G's second-party data was primarily point-of-sale and promotion data that retail partners shared with the company, as well as distributor sales data and publisher data integrated into P&G's programmatic media ecosystem. Lastly, P&G purchased third-party data from market research companies like Nielsen and IRI. Each of these three data sources presented benefits and challenges, and Peri considered how heavily P&G should rely on each one moving forward.

As of 2015, P&G stored its data locally across dozens of platforms and legacy systems. Thus, to understand a given metric, the SBU needed to aggregate data from many different repositories, which took time. Given the complexity, P&G relied on a cohort of 240 business analysts embedded within the SBUs to collect and analyze data for decision-makers. As Peri began thinking about how the company should organize and manage its data resources, he realized that this operating model created two main problems. The first was that business analysts spent a substantial amount of time gathering and cleaning data from many different internal data sources. While gathering the right data to solve a given business problem was important, the real value that analysts offered was their ability to transform that data into actionable insights. If P&G's data were better organized, analysts could increase their productivity by devoting more of their time to actual analysis.

The second issue was more subtle. Because P&G's data was stored across many different platforms, it took a relatively sophisticated user to access and collect all of the data needed to address a given problem. Because many managers lacked technical expertise, it was virtually impossible for them to manage the data on their own. As a result, they tended to outsource this work to the business analysts on their teams, rather than directly interacting with the data themselves. Peri wondered whether this might hinder the culture change he hoped to foster. Could managers still be effective without developing a first-hand understanding of P&G's data? He recognized the value of specialization; no one person could fill every role, and technical tasks were best left to skilled analysts and data scientists who specialized in working with data. However, Peri also believed that business leaders who did not develop a deep understanding of their data would not be able to manage it properly. Without "hands-on keyboard experience," as Peri referred to it, managers would not be able to fully integrate data science capabilities into their operations. But if P&G could organize its data into a centralized, easy-to-access platform, managers could start to get their hands dirty.

How best to organize P&G's vast data resources was a substantial challenge. Javier Polit, P&G's chief information officer from April 2017 to December 2019, believed that an integrated data strategy should be a core component of P&G's new digital strategy. As captured in the popular adage "garbage in, garbage out," a model trained on low quality inputs produced low quality outputs. Thus, the success of data science at P&G relied on the quality of the company's data. Polit understood that P&G's existing data infrastructure needed to be overhauled. Analysts spent too much time collecting and organizing data from different sources, and the lack of a standard data policy meant that related data sets were not always stored in a common format. However, it was not clear what form the new data infrastructure should take.

Esra Yavuz, P&G's director of data management, recognized one core challenge: where to draw the line between data governance and data management. Broadly speaking, data governance referred to an oversight role that ensured data was managed in line with company guidelines and organizational frameworks. Data management referred to an executional role to manage data to enable business units to make decisions regarding their data assets in line with their business strategy. Yavuz wondered where exactly data governance should end and data management should begin. Should the company develop global data governance and data management practices that all business units would follow? This option was attractive because it would ensure consistency across business units. However, she feared that this would be too restrictive; a single approach would not account for the unique needs of individual business units.

Another option was to make the data governance platform global and place individual business units in charge of data management. The data governance platform would establish a core set of data policies across the entire company, and individual business units could adjust their execution of that policy according to their needs. This option seemed more attractive than managing both data governance and data management at a global level. However, it was unclear how restrictive the global data governance policy should be. If it were too restrictive, the business units would not be able to tailor their data management practices to fit the unique needs of their units. For example, a global policy requiring that all sales data be stored on a weekly basis would negatively affect sectors where daily sales data would be more informative. If the policy were too open-ended, however, data would not be organized and managed in a consistent way across the organization. Using the same example, it would be challenging to combine and harmonize data sets from different business units that recorded their data at different levels of granularity (daily, weekly, monthly, etc.).

A related challenge was how to store P&G's vast quantities of data. How much of the company's data should be stored in a centralized location, and how much should be kept in separate repositories within each business unit? There were certain universal data sets that had applications across the entire organization. For example, multiple business units used data on P&G's market share in different product sectors to inform their overall strategy. There was a clear case for storing this type of data in a central location. This would not only eliminate duplicate work in managing the data, but would also allow all business units to rely on a uniform set of facts to inform their decision-making.

There were also clear cases of data that should not be stored in a centralized location. For example, data from a local retailer that serviced a specific geographic region would not be relevant to the majority of P&G's business units. Putting this type of data in the central repository would needlessly crowd the central resource, making it more difficult to use. Therefore, Polit and his team recognized the need for both a central data repository and individual repositories for different business units and regions. To accommodate these needs, David Dittmann, P&G's director of business intelligence and analytics services, led the transition to a new data ecosystem.

Building a Data Ecosystem

In the new structure, all of the data across P&G's systems would be harmonized and consolidated into one ecosystem (see **Exhibit 4**). The keystone of this ecosystem was a "core data lake" that housed any data with multiple use cases across P&G's businesses. Dittmann's team then constructed individual "data hubs," or sector- and region-specific repositories. These hubs combined data drawn from the core data lake with bespoke data relevant to each sector or region. The Fabric and Home Care data hub, for instance, contained relevant data for that SBU in that region, as well as higher-level data from the core data lake.

These hubs were more than just repositories that advanced users could access to retrieve data; they also came with a simple interface that anyone could use to interact with the data. To access the hubs, individual P&G users entered their credentials into a company intranet site, which brought them to a data visualization screen that contained a starter template of relevant information to enable ad hoc analysis and standard reporting. While the underlying data lake/data hub structure was complex, the end user saw a simple interface. Any P&G employee could log into their relevant data hub to access a uniform set of information about, for instance, shipments, inventory, and sales data, which they could use to create visualizations that met their needs.

An ongoing debate was the extent to which the data hubs should be standardized and uniform, versus flexible and customized. Before the new data ecosystem, different branches of the organization had their own standards and practices for analyzing data. In Europe, for example, different countries reported different quarterly performance measures to the general manager (GM) of P&G's European operations. This made comparisons between countries difficult. Because the new data hubs provided a common set of analytical tools, it became much easier to standardize these analyses across the different regions. This simplified the work of both the business analysts and the GM. However, this did not mean complete standardization was the perfect solution. If the analytical tools offered by the hubs were too uniform, different sectors and regions would not be able to perform important analyses that were unique to their businesses.

In October 2018, P&G rolled out its first data hubs to Latin America, China, and Japan, with plans to unveil all other data hubs by December 2019. Peri explained that P&G had chosen to pilot the data hubs in these regions based on their interest. "Leadership in both the Latin America and Japan regions really had the interest and commitment to drive their data hubs forward," said Peri. Given that China was the largest e-commerce market in the world,[16] this region was among P&G's most digitally-enabled markets; thus, it also made sense to pilot a data hub there, in addition to early experimentation with data science given emergent platforms and strong business sponsorship. Adoption of the data hubs had occurred more rapidly than expected. In Latin America, one year after launch, 80% of people in the region had used the hub at least once.

Integrating Data Scientists into P&G

With clean, uniform data, P&G was ready to invest in its data science capabilities. The company already had an existing staff of business analysts located throughout the SBUs. These professionals were familiar with analytical methods and had a deep domain-area understanding of their business units. While the company had invested in training business analysts on data science concepts through a new curriculum called "Friends of Data Science" (see **Exhibit 5**), Peri also recognized the need to hire dedicated data scientists who specialized in data mining and machine learning. But attracting data scientists to P&G, a non-digital native, had proved challenging. To aid in recruitment, P&G had stepped up its compensation package and emphasized the opportunity to tackle what Peri called "global, wicked problems"—complex problems such as analyzing trillions of rows of data to optimize media spend at P&G, the largest advertiser in the world. "We have found that the opportunity to touch five billion customers per day is a big draw," said Peri. P&G also equipped data scientists with the latest tools. "Because we partner with all the digital players," said Peri, "we often get alpha access to their new technologies."

Identifying the most effective operating model for integrating data scientists into the company had also been challenging. Said Jeff Goldman, director of data science, "We have experimented with just about every model under the sun." The primary question was whether data scientists should be embedded within individual business units, or placed on a single centralized team that would serve the entire business. Placing data scientists directly into the business units exposed them to mission-critical issues facing those units, but risked isolating them. Conversely, placing them in a central data and analytics team meant that they might develop models divorced from the SBUs' most pressing needs.

P&G chose to place its first few data scientists directly into the business. "But," said Peri, "because data scientists come in with a completely different paradigm, they often challenge the conventional ways of doing things, and that isn't always well-received, especially when our business leaders aren't familiar with the analytic technique and the model's inputs and outputs." For example, one of these early data scientists developed a model to test the effectiveness of P&G's primary approach for media buying, which found several flaws. He proposed a new, data-driven strategy, which ultimately redefined how media testing would be conducted. But when the data scientist presented his findings, the business unit leader rejected his model, leaving the data scientist questioning his role and purpose at P&G. He ultimately resigned.

Additionally, business leaders often tasked data scientists with work that underutilized their skills. Although managers had a deep understanding of P&G's business problems, they often lacked an understanding of how the company's data resources could be leveraged to solve those problems. Missed opportunities occurred when data scientists were assigned to tasks that did not fully capitalize on their skills. Relatedly, data scientists were occasionally asked to solve problems that were unrealistic given the available data. The business analysts embedded in the SBUs sometimes bridged this gap by serving as translators between data scientists and business leaders. However, this was not a perfect solution. Peri realized that tackling these problems was crucial for the success of data science at P&G.

A centralized model seemed to solve some of these problems. Data scientists staffed on a centralized team could report to a manager who had a deep understanding of data science and machine learning capabilities. This ensured that their work would be well-understood and appreciated. It also meant that they would be assigned to projects that challenged their skills without imposing unrealistic demands. Given the competitive nature of the labor market, it was crucial that data scientists felt both stimulated and supported at P&G. A centralized team allowed for a more fully-developed community of data scientists that would not be possible if they were isolated on separate business teams. This had the added benefit of cross-pollination, as data scientists working on similar problems could share insights with each other.

A centralized team also made it easier for P&G to coordinate data science efforts across the entire company. There were many solutions with applications in multiple different SBUs and geographic markets. For example, a machine-learning model that could leverage regional data about demographics and market conditions to predict sales of a new P&G product would have nearly universal appeal. If data scientists were distributed across compartmentalized business teams, it would be very difficult to develop these types of coordinated, global solutions. Only a centralized team that sat above individual business units would have the perspective necessary to recognize these wide-ranging opportunities.

However, Peri understood that a centralized model was not without its flaws. For the true value of data science to be realized, solutions had to be developed in conjunction with a deep understanding of the underlying business problems. Data scientists working on a centralized team would likely not develop the domain-area expertise necessary to help individual teams solve their key business problems. Recognizing the benefits and challenges inherent to both models, Peri thought carefully about how data scientists should be integrated at P&G.

Applying Analytics to P&G's Businesses

To ensure that analytics served business outcomes, Peri and his team worked with the SBUs to create data strategies that aligned with their business strategies. In this process, each SBU identified the priority business problems they wanted to solve and the data needed to solve them (see **Exhibit 6**). As Peri explained, "From a business perspective, the *what* hasn't changed. We are still focused on the five drivers of superiority: product, brand, communication, in-store execution, and the value equation; however, with analytics, pretty much everything about the *how* has changed. Analytics is now embedded into each of these drivers so we can help the business steer across all of our brands." Four examples of the ways in which P&G applied analytics to its work processes follow.

Neighborhood Analytics: Optimizing Oral Care

Around 2015, P&G's data scientists began to build a machine learning model capable of analyzing massive amounts of information about the areas surrounding individual retail stores to predict which products might sell best in those stores. P&G data scientist Dan Ames was the first to conceive of the possibility of combining granular store signals and contextual attributes with sales data to glean insights, which came to be known as the neighborhood analytics capability.

To build this capability, P&G's data scientists created a model that used 2 trillion rows of data on people's demographics and their unique tastes and preferences to segment the U.S. into a collection of hundreds of thousands of neighborhoods. They then overlaid point-of-sale data onto this information to make sales predictions. The resulting model allowed P&G to optimize its distribution, sampling, couponing, and advertising. Once the model was built, the central data and analytics team partnered with three SBUs to apply it across 20+ markets.

One of these units was the U.K. oral care team. A leading P&G oral care product in the U.K. was the Oral-B® electric toothbrush. Among P&G's strategies for driving sales of these toothbrushes (and, by extension, improving consumers' oral health) was conducting outreach to dental practices to drive recommendations for the Oral-B® electric toothbrush. A cohort of territory managers across the U.K. regularly called or visited dental practices and provided them with programs and tools to help them improve their recommendation quality.

P&G had limited options for understanding dentists' recommendation practices. Sales Director Razi Hyder explained: "There really is no dataset that we can buy describing what happens at the dental professional level—who is giving more or less recommendations to whom and the impact that ultimately has on the patient's purchasing decisions—so we have historically estimated the effectiveness of our outreach by looking at business results and electric toothbrush usage surveys, which give an indication of recommendation frequency. But that was the only data point we had." Thus, when deciding which dental practices to visit, territory managers had historically chosen practices that were receptive to them, enthusiastic about the product, and used the samples provided. Territory managers tended to visit the same practices a few times each year, dropping a small number of practices each year to add new ones.

In 2016, the oral care business leaders noticed that while sales of replacement heads for the Oral-B® toothbrush were up in the U.K., sales of the actual brush handles had fallen flat, meaning that relatively few new users were adopting the toothbrush. In 2017, to increase sales to new users, P&G decided to apply the nascent neighborhood analytics capability to identify the dental practices with the highest potential for future handle sales.

Using the neighborhood analytics platform, the data science team built a regression model that predicted Oral-B product sales at a given retail store based on the demographic features of the surrounding area. This model provided an understanding of how different demographic features related to sales of dental products. Some important independent variables were standard demographic features, such as age, income, and education level. Others were unique to the dental industry, such as the proportion of residents in the surrounding area who wore dentures.

Data scientist Benjamin D'Incau used the most important demographic features identified by the model to create a profile of every dentist in the U.K. to identify which ones saw patients whose demographics indicated a higher propensity to buy electric toothbrushes. The team also created a proxy variable to estimate the portion of households within the model's patient base of a given dentist that were already using electric toothbrushes. The oral care team used this information to identify the dental practices with the highest potential to convert recommendations into handle purchases by new users.[b]

Once these practices had been identified, the oral care team invited all the territory managers to see the results on a map created via an intuitive visualization tool. Managers could click on any individual dental practice to show granular detail. For instance, clicking on a given practice might show that while the area surrounding that practice seemed promising because of its low handle penetration, the demographics indicated a low propensity to buy, so the practice was not recommended. Business Analyst Rachel Breslin recalled: "That was a powerful experience because the territory manager could think back to that dentist and relate the data to their knowledge of the practice and surrounding area."

Territory managers also helped fill gaps for the analytics team. Breslin explained: "We couldn't understand why this one particular area had poor dental health, and the territory manager said, 'Oh, there is a soda manufacturing plant in that area,' which is a totally plausible explanation for why the area might have poorer dental health. This process built trust and rapport between the analytics team and the territory managers, and also provided a qualitative check to our quantitative data."

Smart Selling: Improving In-Store Sales

P&G had also applied analytics to improve the sales process at the individual store level. This had proven especially useful in the Asia Pacific, Middle East, and Africa (AMA) region. Comprising 105 countries, AMA was an incredibly diverse region with a fragmented retail landscape populated mostly by small owner-operator stores. Whereas in developed markets, P&G had a relatively good understanding of how its products were displayed and promoted at retail stores, it was much harder to access reliable, accurate data across the hundreds of thousands of small stores in the AMA region.

[b] All practices related to this project complied with the General Data Protection Regulation (GDPR), a sweeping set of data privacy guidelines that came into effect across the European Union in 2018. P&G's U.K. team used no identifiable consumer data.

P&G thus contracted with local third-party in-store sellers—small companies with operational skills and knowledge of their markets—to monitor operations. These in-store sellers visited stores and recorded basic information, such as the number of shelves dedicated to P&G products and whether the appropriate point-of-sale merchandising was on display. But they used a number of different systems to report their findings back to P&G, and they had very little ability to provide real-time insights to the retail stores on how they might improve sales. Turnover among in-store sellers was also quite high.

To make selling more effective, P&G built an algorithm that used historical AMA sales data to predict strategies for improving sales at the individual store level. The core model was a collaborative filtering algorithm that identified similar stores based on individual product sales. The model was constructed on a data set in which each column represented a different product (a stock keeping unit, or SKU) and each row represented a different store. To make product recommendations for a given store, the algorithm would first use cosine similarity to identify the other stores that were most similar to the target store based on the sales of each SKU (the "nearest neighbor" stores). The model would then determine the best-selling SKUs among those nearest neighbors. From there, one could identify the SKUs that were being undersold in the target store relative to its nearest neighbors. These SKUs were recommended to the target store as high-priority items.

Explained Sonal Tyagi, data science leader for AMA, "We now have robust data to profile every store and come up with customized recommendations." Sample recommendations included maintaining an optimal mix of products for that particular store, promoting certain products, and increasing or decreasing the frequency of promotional activities. P&G developed an app called Smart Sales App that conveyed these insights; it piloted the app with in-store sellers in the AMA region starting in 2018 (see **Exhibit 7** for a screenshot of the app). Smart Sales App helped in-store selling teams quickly prioritize tasks at the store level and optimize time spent, among other tasks. The app also provided a standardized data collection tool and a place to upload photos of store displays. "The main challenge as we started," added Project Manager Akanksha Singhal, "was to get them to use the tool. However, once they saw the benefit of guided tasks and simplified data collection, they started adopting it."

Prioritizing tasks When in-store sellers visited a store, they often struggled to prioritize tasks. With Smart Sales App, in-store sellers had access to a unique profile of each store with tailored recommendations. "Now," explained Singhal, "when the in-store seller walks into the store on a given day, the app recommends which five tasks they should prioritize." The prioritized tasks were those that had the biggest impact on sales, for example addressing stock-outs of a best-selling item, addressing shrinking shelf space for a given product category, or alerting the store manager of a suboptimal product mix. As Tyagi explained, "Before, in-store sellers were doing their job to the best of their ability, but very few of them had experience with P&G brands and categories. We discovered that most of them sorted their product lists in alphabetical order, so products that start with 'A,' like Ariel® were getting restocked, but products that start with 'P,' like Pampers® were not. A tool like Smart Sales App can be very helpful to them."

Optimizing investment behind the right key business drivers As part of its trade spend, P&G invested in multiple in-store elements, such as Touch & Feel units and testers. Touch & Feel units were display devices that customers could use to physically interact with products; for example, one Touch & Feel unit allowed customers to touch a diaper product before purchasing. Testers were in-store samples of consumable products, such as perfume. With Smart Sales App, P&G's in-store team could track compliance at the local store level by capturing pictures of these in-store elements. By applying analytics to these data sets, P&G could guide retailers on the right investment choices to drive category growth, which the company believed created a win-win for brands and store owners.

The model built by Tyagi and her team used cosine similarity to group stores together based on their sales of different SKUs. Under this method, the rows of the data (i.e., the stores) were grouped together based on the columns (i.e., SKU sales). However, one could also consider taking the opposite approach: grouping SKUs (i.e., the columns) together based on their sales across the different stores (i.e., the rows). This would identify products that were similar to each other based on their sales patterns. Combining the two approaches improved model accuracy in certain regions with low store density.

Smart Products: Oral-B Genius X® and Airia®

P&G recognized opportunities to leverage data through the development of smart, connected products. Such products accomplished three objectives: 1) they enhanced the consumer experience; 2) they generated granular, first-party consumer data; and 3) they enabled P&G to experiment with a model of selling direct to consumers (DTC). As Ron Peri, head of connected product and IT innovation, summarized, "Connected devices allow us to have a real-time digital relationship with our consumers for the first time." Peri saw value in this relationship. The data collected from these smart products would not only enhance the consumer experience with the products; it would also provide P&G with a new source of consumer data that could be used to develop and improve other products. He believed that this opportunity needed to be recognized when assessing the value of P&G's new smart products.

To capitalize on this opportunity, in 2019, P&G built its own Internet of Things (IoT) platform to allow the company to centrally collect and compare data across all of its smart products. Two such products were the Oral-B Genius X®, an electric power toothbrush, and the Airia® Smart Scent product (pronounced "Area"), a connected air freshener.

Oral-B Genius X® As early as 1991, P&G placed a single LED light in the handle of its Oral-B® electric toothbrush, which blinked when the user finished the recommended two-minute brushing cycle. As technology advanced, so did P&G's feedback and coaching features. The Oral-B Genius®, launched in 2018, included an accelerometer and a sensor that communicated with a mobile app to assess brushing habits and provide real-time feedback and position detection so a consumer would know that they covered their entire mouth for the right amount of time. But there were barriers to use of version 1. Because the product used a smartphone as a reference camera, the user had to adhere her phone to the mirror with a suction cup each time she brushed.

To remove these barriers, P&G began developing the Oral-B Genius X®, which would use a camera-less position detection algorithm to detect where the user's toothbrush was located in space, eliminating the need for any other supplemental brushing tools. As the product development team began to work on this new product, however, they faced a major challenge: the cost of connectivity to enable smart interactivity within the handle. "At P&G's scale," said Peri, "every penny counts. Adding connectivity and sensors to every handle adds cost to develop and manufacture." Some P&G leaders questioned the logic of introducing additional cost and complexity to the design process, where no immediate sales benefit was obvious.

Conversely, the product development team saw vast benefit in adding connectivity. The data generated by the chip would allow P&G to identify poor brushing habits and recommend personalized strategies to improve. New business models could potentially emerge from such a focus on outcomes. Whereas historically, P&G had relied on customers buying handles and then replacement heads over time, a new model might emphasize the outcomes associated with personalized brushing recommendations (e.g., fewer cavities and better oral health). "In the future, instead of a toothbrush, we could sell an outcomes-focused oral health package," said Ron Peri. P&G could also use data to inform product innovation. To educate upper management on the potential, the product development team had held dozens of IoT workshops and exploratory design sessions to make the case.

Airia® Another new P&G smart product was the Airia® smart diffuser (see **Exhibit 8**). The product aimed to solve the primary challenge in the air freshener space: that scents from current delivery systems faded over time. Launched in November 2019 at a price point of $250, Airia® used a thermal inkjet printer to deliver scents. "We're essentially printing perfume droplets into the air," explained Thinh Ha, R&D director for smart products, "and because of that, the user can control the strength of the scent. There is no fading." Users could schedule the time, duration, and intensity of their scent experience as well as the ambient lighting through the Airia® app. The product's SmartJet delivery system could cover an area of up to 200 square feet. The Airia® also included a voice control feature that allowed consumers to interact and control their device using their voice. Given the expense associated with developing the Airia® product and its high price point relative to P&G's other scent products, like the Febreeze® plug-in—which retailed for less than $10—some business leaders questioned whether the Airia® could be sold at traditional food/drug/mass channels. From early data, the team learned about consumer use habits, which correlated to increased sales.

More broadly, P&G's growing investment in smart products made some within the company consider engaging in edge computing, which referred to the practice of processing data locally—"at the edge of the network within their home"—before it was transferred to the cloud.[17] Developed in response to the growth of the IoT, edge computing offered companies the ability to prevent the lag in response time that resulted from transfers of real-time data to the cloud. For example, with edge computing, the Oral-B Genius X® algorithm would run locally on an edge processing chip in the brush handle rather than through the cloud, saving time and enabling a better real-time experience.[18] But how these edge processors would interface with P&G's data hubs remained unclear.

Winning with Multicultural Consumers: The Gold Series Collection®

P&G had also used its growing analytics capabilities to identify the most productive retail stores for products, such as The Gold Series Collection®, P&G's newest line of multicultural hair products. This market was growing, with African-American consumers in the U.S. spending $2.6 billion on hair products in 2016.[19] Because the CPG industry had historically underinvested in this market, the majority of suppliers were small businesses.[20] Aaron Steele, associate director for strategy, insights, and analytics within P&G's North America hair-care business, estimated that as much as 70% of multicultural hair products sold through non-traditional retail channels, like independent beauty supply stores, small corner stores, and bodegas. In 2016, a group of African-American scientists at P&G began developing the Gold Series Collection®, a line of products for consumers with textured hair. As Steele recalled, "These scientists had solved the needs of millions of Americans but realized they had not created the perfect solution for themselves and the more than 50 million multicultural consumers who look like them. They created the product, so we had our own founder story."

When The Gold Series Collection® launched in March 2018, Steele worked with data scientist Matt McFadden to consider how they might use data science to identify stores that were the best fit for the product. They had access to lists maintained by retailers, which tagged individual stores according to the demographic features of their customers. In a pre-analytics world, P&G would have used these lists to prioritize launching in stores that served the target consumer. The strategy would have likely been a volume play—blanket the stores serving multicultural consumers with the new product.

But, noted McFadden, many of those retailer lists were outdated. Moreover, the assumption of the past was that race was the most important factor to determine buying patterns. To better identify stores that not only served multicultural customers but also served customers with a propensity to buy the Gold Collection®, McFadden and P&G's data science team built an ensemble model on top of the Neighborhood Analytics platform that capitalized on the strengths of different machine learning techniques. Using a list of the 20,000 possible retail stores where the product might sell, the model ranked each store by order of opportunity for the Gold Series Collection®.

The first challenge when constructing the model was the choice of the target variable. Ultimately, McFadden and his team needed to predict sales of the new Gold Series Collection®. This was a new product line, so no historical sales data existed that could be used as the target feature. However, the data science team recognized that although they had no sales data on the Gold Series, they did have access to historical sales data on similar products that targeted the same market. For example, the African Pride brand was another product in the market designed to meet the needs of African-American consumers. The data science team realized they could use historical sales of this proxy product as the target feature in their model. Although the team hoped that this product was representative of the Gold Series, they also recognized that this solution was not perfect; the unique characteristics of the Gold Series Collection® that differentiated it from similar product lines would not be represented.

The independent features of the model included 1,100 demographic (i.e., age, race, gender, income, family size), behavioral (e.g., how people are buying), and environmental (i.e., information about attitudes, weather, etc.) variables. Some of these features were obviously relevant, such as the population density of African-Americans in the area surrounding each store. Others, however, were less obvious to the team. Red wine consumption and interest in golf, for example, ended up being significant predictors in the final model. McFadden also included information about the number of independent beauty supply stores in the surrounding area to estimate the competitive landscape for each store. "Matt helped us understand that while the opportunity is large, it's not the same in every neighborhood," explained Steele.

The data science team trained several different machine-learning models on the data. They began with linear regression, which had initial appeal because of its interpretability. By observing the coefficient assigned to each independent feature of the regression equation, one could understand how the model related each independent feature to the target variable. This made it easier to explain the recommendations of the model to stakeholders who were not familiar with machine learning. "We used regression as a bridge," explained McFadden. After using regression to build a level of trust in machine learning capabilities, McFadden and his team switched to a random forest model, which performed better than linear regression. Although this was a black-box model, the trust that McFadden and his team had built meant that retailers were receptive.

The model ranked the 20,000 possible retail stores in order of opportunity for The Gold Collection®, landing on a list of 400 stores that were likely to be the most productive. The model also provided tailored insights for individual products by retailer down to SKU-level granular insights. For instance, the model might recommend that Retailer X in Neighborhood Y of Atlanta stock two specific P&G multicultural hair SKUs, whereas Retailer Z, down the street from Retailer X, stock just one SKU. McFadden and Steele also recommended how much shelf space stores should devote to multicultural hair products, based on competitor proximity. Prior to these model-enabled insights, retailers were stocking less than 50% of high propensity stores. When developing the model, recalled Steele, "I let Matt's team work through the messy details. Our experience has been that when we unleash our data scientists, who have incredible expertise and capability, then that is when we start to create magic."

Not only could these insights help retailers prioritize, but Steele believed that they helped P&G better serve the consumer. "To see the emotive reaction of a consumer seeing dozens of brands and tons of space dedicated to products for her was amazing," he said. The adoption among retailers had been gradual and measured. Some retailers immediately acted on the insights, while others were just beginning to grapple with how to fuse the insights with their own data and activity systems, which were often equally complex.

Looking Ahead

While Peri and his team had made progress in integrating analytics into certain pockets of P&G, many business leaders remained hesitant. "After all," said Ha, "we have had a very successful history as a chemistry and substrate company." But Peri saw momentum as businesses had started to see the competitive advantage made possible by speed of insights and the resulting in-market actions brought about by data and analytics. Driving activation deeper and further, said Peri, would require ongoing upskilling efforts across all P&G businesses and in every corner of the company.

"We are not interested in turning everyone into a data scientist," said Peri, "but we are interested in being intentional about building P&G's collective literacy in this space." He wanted businesses to ask the right questions of data science, to recognize when a problem or opportunity lent itself to leveraging P&G's analytical capabilities, and for many (if not all) knowledge workers, to build some level of personal applied analytical capability.

Peri's team had also worked to educate leaders about the data science applications at P&G, creating a simplified version of the Friends of Data Science program for executives (see **Exhibit 9**), as well as a "reverse-mentoring program," as a first step. "These upskilling efforts are helping tremendously," he said, "and we are now building on this approach to include problem-specific upskilling." According to Peri, a problem could be how a business developed an effective media plan or promotion plan leveraging data and analytics. His team planned to publish these specific end-to-end "playbooks" to allow other businesses to adopt them.

In addition to upskilling, Peri was committed to addressing the cultural change needed to enable adoption of analytics capabilities. One strategy, he said, was approaching teams with humility. "When we engage a product research team that has a traditional way of working, we never presume that we have the right answer. We say, 'Let's pilot this model on a small piece of the product and see if it even works.' If we're humble and the outcome is great, people want more."

Another important strategy for overcoming organizational resistance to change was balancing the precision of machine-learning models with interpretability. "Machine learning lives and dies based on the human's ability to understand what it does," said Peri. "The number one driver for failure is the black-box syndrome. We try to transform our models into a 'glass box,' providing a very simple explanation of what goes inside the machine. When people start trusting the output, the antibodies diminish. In some cases, we sacrifice perfect accuracy of the model to improve understandability."

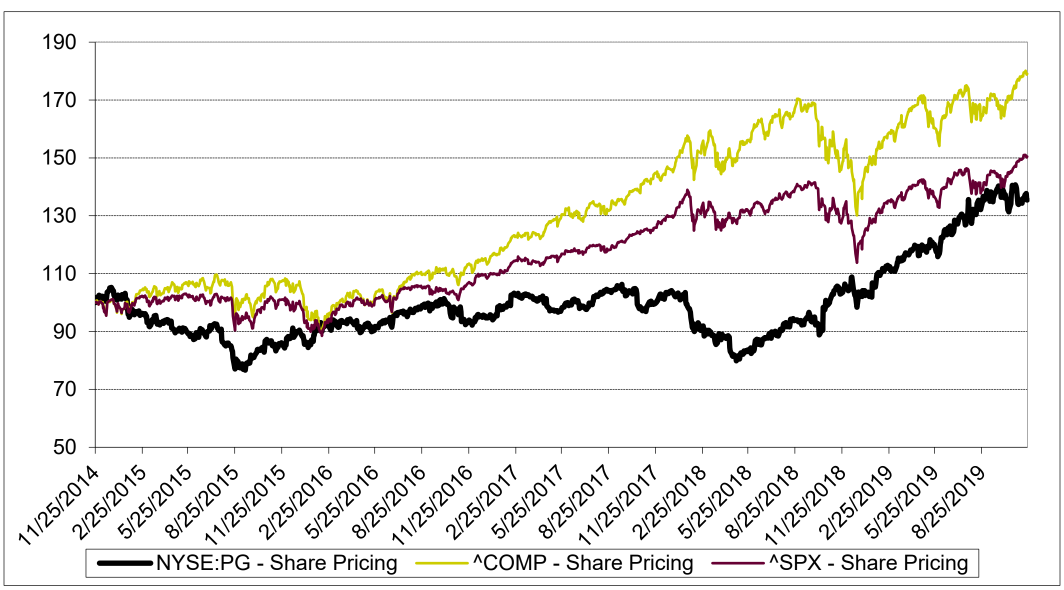
Looking to the future, Peri knew that P&G's customers and consumers were increasingly integrating data and analytics into their daily operations. He wondered what more he and his team could do to help P&G constructively disrupt itself.

Exhibit 1 P&G Consolidated Financials, in US$ Millions, except margins and per share, 2015-2019

	2019	2018	2017	2016	2015
Net sales	**67,684**	**66,832**	**65,058**	**65,299**	**70,749**
Gross profit	32,916	32,400	32,420	32,275	33,649
Operating income	5,487	13,363	13,766	13,258	11,056
Net earnings from continuing ops.	3,966	9,861	10,194	10,027	8,287
Net earnings/(loss) from discontinued operations	-------	-------	5,217	577	(1,143)
Net earnings attributable to P&G	**3,897**	**9,750**	**15,326**	**10,508**	**7,036**
Net earnings margin from continuing operations	5.9%	14.8%	15.7%	15.4%	11.7%
Basic net earnings per common share:					
Earnings from continuing operations	$1.45	$3.75	$3.79	$3.59	$2.92
Earnings/(loss) from discontinued operations	-------	-------	$2.01	$0.21	($0.42)
Basic net earnings per common share	**$1.45**	**$3.75**	**$5.80**	**$3.80**	**$2.50**
Diluted net earnings per common share:					
Earnings from continuing operations	$1.43	$3.67	$3.69	$3.49	$2.84
Earnings/(loss) from discontinued operations	-------	-------	$1.90	$0.20	($0.40)
Diluted net earnings per common share	**$1.43**	**$3.67**	**$5.59**	**$3.69**	**$2.44**
Dividends per common share	**$2.90**	**$2.79**	**$2.70**	**$2.66**	**$2.59**
Research and development expense	1,861	1,908	1,874	1,879	1,991
Advertising expense	6,751	7,103	7,118	7,243	7,180
Total assets	115,095	118,310	120,406	127,136	129,495
Capital expenditures	3,347	3,717	3,384	3,314	3,736
Long-term debt	20,395	20,863	18,038	18,945	18,327
Shareholders' equity	47,579	52,883	55,778	57,983	63,050

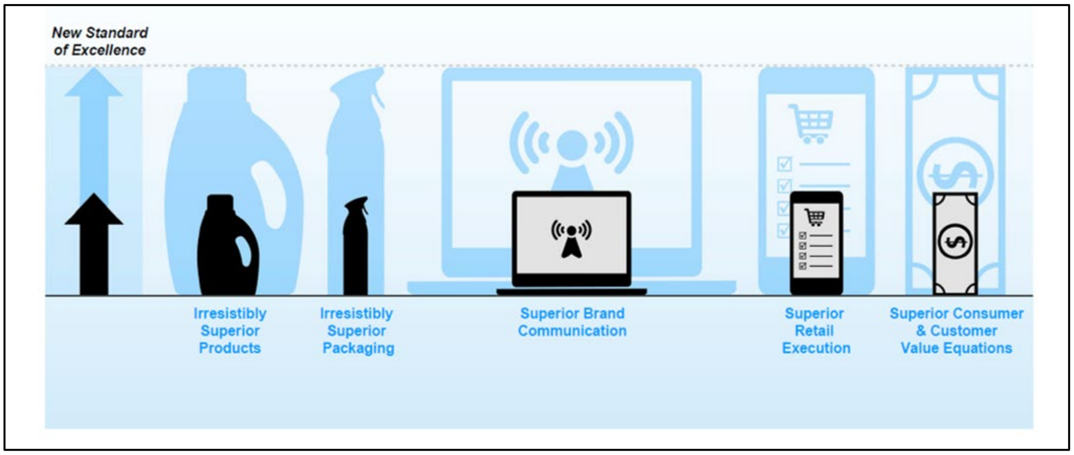
Source: P&G, June 30, 2019 Form 10-K, http://d18rn0p25nwr6d.cloudfront.net/CIK-0000080424/6bc79675-4ef0-4940-bed2-7650e689353e.pdf, p. 10, accessed November 2019.

Exhibit 2 P&G Stock Price, As compared to S&P 500 and Nasdaq Indices, 2015-2019



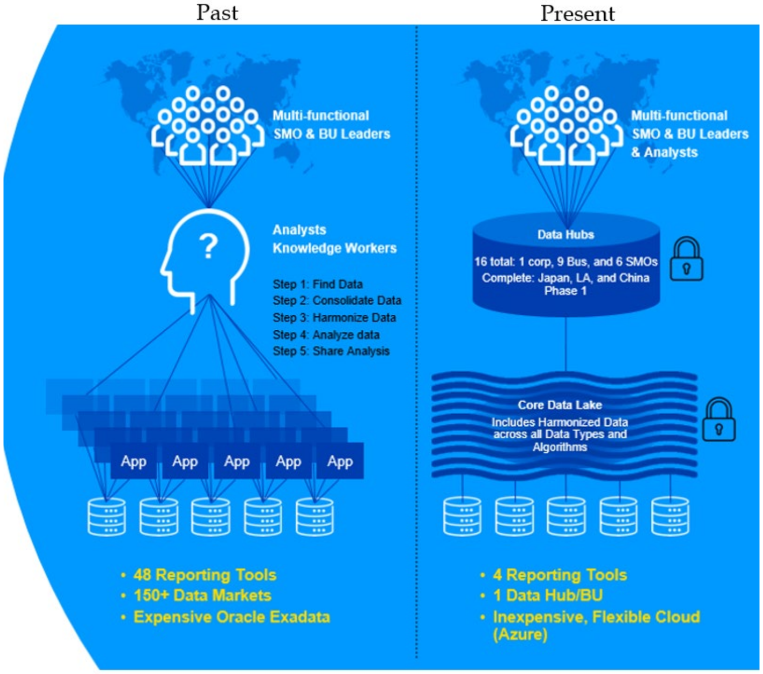
Source: Capital IQ.

Exhibit 3 P&G Using Analytics to Drive toward a New Standard of Excellence



Source: Company documents.

Exhibit 4 P&G's Historical vs. Current Approach to Data Storage and Use



Source: Company documents.

Note: SMO = Selling and Market Operations (these preceded the focus and enterprise markets); BU = Business Unit; LA = Latin America.

Exhibit 5 Friends of Data Science Curriculum for Business Analysts

Friends of Data Science Level 1	Friends of Data Science Level 2
(1) Intro to SQL for Data Science	(1) Data Engineering 2 weeks
(2) Joining Data in PostgreSQL	(2) Decision Tree 1 week
(3) Intro to Python for Data Science	(3) Neural Network 1 week
(4) Intermediate Python for Data Science	(4) Time Series 1 week
(5) Python Data Science Toolbox (Part 1)	(5) Recommender System 1 week
(6) Statistical Thinking in Python (Part 1)	(6) Text Analytics 2 weeks
(7) Statistical Thinking in Python (Part 2)	(7) Graph Analytics 1 week
(8) Hadoop Starter Kit	(8) Deep Learning 1 week
(9) Hadoop – Ecosystem	
(10) P&G Big Data Platform On-boarding	

Source: Company documents.

Notes: Python is a programming language. Hadoop is a tool used to process big data. Decision trees and deep learning are types of machine learning.

The key objective of Level 1 is for business analysts to learn key tools (SQL, Python), and gain a basic understanding of the Hadoop environment. Level 1 is delivered as a 30-hour e-learning course. The Level 2 training focuses on the fundamentals of machine learning with critical focus on not just the models but their limitations. Level 2 is a 13-week course with a certification exam. As of November 2019, 100% of P&G's business analysts had completed these two courses.

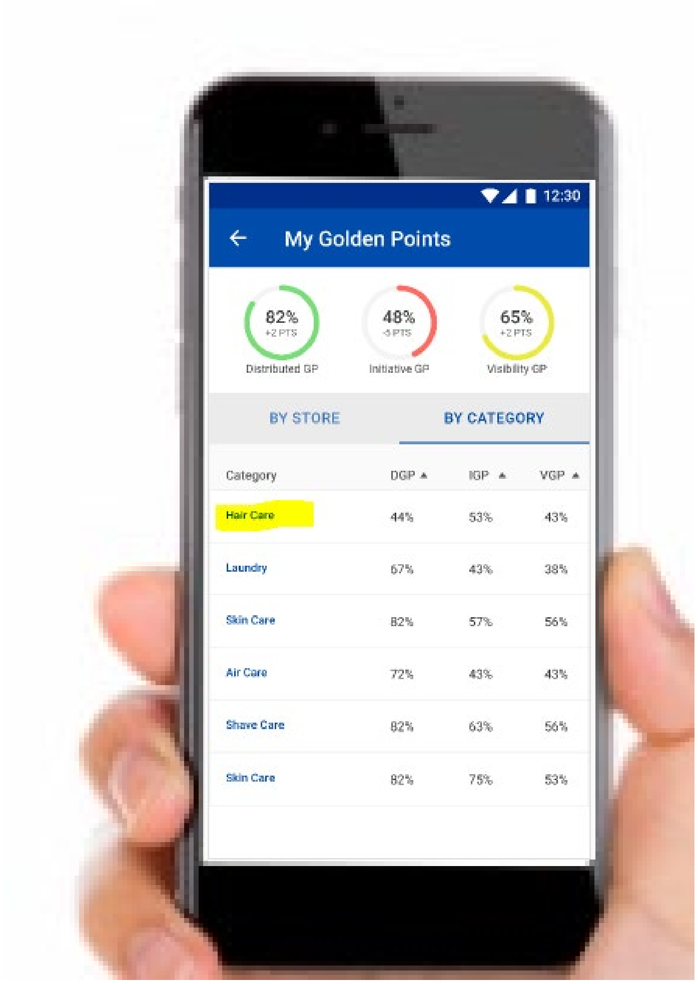
Exhibit 6 Example Data Strategy Priorities

Sample Priorities		
Key Business Drivers	Media	Trade Spend
Pricing	Omni-Channel	Post-event Analytics
e-Commerce	Retail Execution	

Source: Casewriter, adapted from company documents.

Note: Key Business Drivers were product, brand, communication, in-store execution, and the value equation.

Exhibit 7 P&G's SmartApp for Smart Selling



Source: Company documents.

Exhibit 8 Airia®, P&G's Smart Diffuser

Source: Company documents. AIRIA, "AIRIA Device," https://shop.airiasmartscent.com/airia-device/, accessed March 2020.

Exhibit 9 Friends of Data Science Curriculum for Executives

Friends of Data Science Curriculum for Executives
(1) Overall introduction of the importance of data and algorithms and where to apply them in P&G (2) Data capture, governance, and engineering (3) Machine learning algorithms (4) Common tools (SQL, Python, KNIME) (5) Latest technologies including cloud and GPU (6) How to activate – questions executives should be asking

Source: Company documents.

Endnotes

[1] "Consumer packaged goods' (CPG) industry sales in the United States in 2015 and 2020 (in billion U.S. dollars)," June 22, 2016, accessed via Statista.

[2] Kari Alldredge, Brandon Brown, Lindsay Hirsch, and Travis Reaves, "'Power partnerships': Manufacturer-retailer collaborations that work," McKinsey & Company, May 2019, https://www.mckinsey.com/industries/consumer-packaged-goods/our-insights/power-partnerships-manufacturer-retailer-collaborations-that-work, accessed December 2019.

[3] P&G, June 30, 2019 Form 10-K, http://d18rn0p25nwr6d.cloudfront.net/CIK-0000080424/6bc79675-4ef0-4940-bed2-7650e689353e.pdf, accessed November 2019.

[4] P&G, June 30, 2019 Form 10-K, http://d18rn0p25nwr6d.cloudfront.net/CIK-0000080424/6bc79675-4ef0-4940-bed2-7650e689353e.pdf, p. 11, accessed November 2019.

[5] "About P&G," P&G, https://news.pg.com/about/core_strengths, accessed December 2019.

[6] P&G, June 30, 2019 Form 10-K, http://d18rn0p25nwr6d.cloudfront.net/CIK-0000080424/6bc79675-4ef0-4940-bed2-7650e689353e.pdf, p. 1, accessed November 2019.

[7] P&G, June 30, 2019 Form 10-K, http://d18rn0p25nwr6d.cloudfront.net/CIK-0000080424/6bc79675-4ef0-4940-bed2-7650e689353e.pdf, p. 13, accessed November 2019.

[8] "What is a Category Captain?" DotActiv, August 25, 2016, https://www.dotactiv.com/blog/what-is-a-category-captain, accessed November 2019.

[9] P&G, June 30, 2019 Form 10-K, http://d18rn0p25nwr6d.cloudfront.net/CIK-0000080424/6bc79675-4ef0-4940-bed2-7650e689353e.pdf, p. 12, accessed November 2019.

[10] M. Shahbandeh, "Percentage of Procter & Gamble's net sales worldwide in 2019, by region," August 27, 2019, via Statista.

[11] P&G, June 30, 2019 Form 10-K, http://d18rn0p25nwr6d.cloudfront.net/CIK-0000080424/6bc79675-4ef0-4940-bed2-7650e689353e.pdf, p. 1, accessed November 2019.

[12] P&G, June 30, 2019 Form 10-K, http://d18rn0p25nwr6d.cloudfront.net/CIK-0000080424/6bc79675-4ef0-4940-bed2-7650e689353e.pdf, p. 10, accessed November 2019.

[13] Niko Mohr and Holger Hürtgen, "Achieving business impact with data," McKinsey & Company, April 2018, https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20analytics/our%20insights/achieving%20business%20impact%20with%20data/achieving-business-impact-with-data_final.ashx, p. 3, accessed December 2019.

[14] Rachel Haberman, "CPG Has a First-Party Data Problem," jebbit, March 11, 2019, https://www.jebbit.com/blog/cpg-has-a-first-party-data-problem, accessed December 2019.

[15] Julio Hernandez, Bob Berkey, and Rahul Bhattacharya, "Building an Analytics-Driven Organization," Accenture, 2013, https://www.accenture.com/us-en/~/media/accenture/conversion-assets/dotcom/documents/global/pdf/industries_2/accenture-building-analytics-driven-organization.pdf, p. 10, accessed November 2019.

[16] Jing Wang, "China Is World Leader In E-Commerce, New Retail Innovation: EMarketer," Alizila, July 30, 2019, https://www.alizila.com/china-is-world-leader-in-ecommerce-new-retail-innovation-emarketer/, accessed December 2019.

[17] Keith Shaw, "What is edge computing and why it matters," NetworkWorld, November 13, 2019, https://www.networkworld.com/article/3224893/what-is-edge-computing-and-how-it-s-changing-the-network.html, accessed December 2019.

[18] Shaw, "What is edge computing and why it matters."

[19] "Money flowing into the natural hair industry is a blessing and curse for those who built it up," *Los Angeles Times*, August 10, 2017, https://www.latimes.com/business/la-fi-natural-hair-industry-20170809-htmlstory.html, accessed December 2019.

[20] "Money flowing into the natural hair industry is a blessing and curse for those who built it up," *Los Angeles Times*.