Cloud Computing and Machine Learning: The Business Case for Big Data

Harvard College Consulting Group

August 2021

1. Table of Contents

2. Executive Summary

The HCCG team mapped out the current technical landscape to evaluate key use cases of big data analytics, assess the challenges organizations face when adopting big data solutions, and break down the core ethical principles governing data usage. This paper presents what analytical tools exist, how big data can be used, why big data solutions are helpful, and when they should be implemented.

TECHNICAL LANDSCAPE 1

Growing technologies like cloud computing provide businesses with easy access to data storage and computing power. Large quantities of data, in tandem with cheaper and more accessible storage, have engendered an increasing trend to build analytical tools that seek to understand and make use of collected data. These tools, including machine learning and Natural Language Processing, open up new insights for companies to make informed, strategic decisions.

2 KEY USE CASES

Regardless of the specific use case and industry, at the core, big data solutions help reduce uncertainties in decision-making. In retail, analytics provide predictive models that allow companies to innovate, optimize pricing, and personalize offerings strategically. Deep learning technologies in healthcare improve patient outcomes, reduce expenses, and increase diagnostic accuracy. Within internal processes, forecasting and analytical tools optimize operations decisions, improve supply chain efficiency, detect fraud, and reduce risks. Armed with big data insights, businesses can strategically innovate both bottom-line and top-line growth.

PRACTICAL & ETHICAL CONSIDERATIONS 3

In adopting big data solutions, organizations face both practical and ethical challenges. Businesses often overestimate their abilities to fund projects, and effective big data implementation requires organizational cultural change and significant human capital. Organizations that are able to integrate analytics into business decisions need to also carefully consider key ethical concerns about data privacy and biases within analytical models.

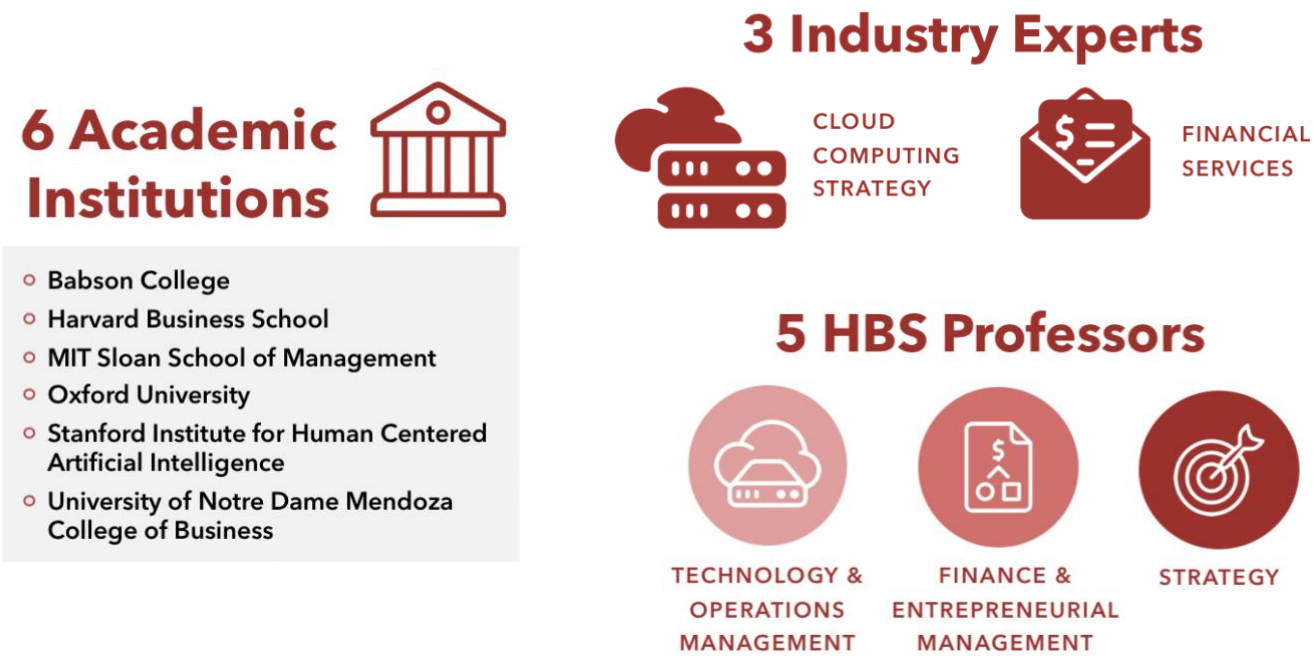
3. Methodology

To examine the business use cases and limitations of big data, the HCCG team conducted extensive secondary research rooted in academia and public literature. The team **reviewed and synthesized over 100 articles, journal studies, and reports.** This research provided background context, case studies, and quantitative insights that informed the trajectory and backbone of this report.

For more nuanced, granular insights, the HCCG team conducted **interviews with 13 academic and industry experts, spanning 9 academic institutions and companies.**

These expert interviews yielded deeper analyses of how businesses can implement big data solutions in practical, effective, and ethical manners.

The synthesis of primary and secondary research allowed the team to assess the current technical landscape of big data, use cases and future trends, and the core ethical principles associated with data governance and privacy.



4. Evolving Landscape of Big Data

4.1 Cloud Computing Overview

The COVID-19 pandemic accelerated the digitalization of markets, companies, and services, pushing consumers and businesses to utilize more digital tools and technologies. The average share of **digital customer interactions has increased from 36% in December of 2019 to 58% in July of 2020.**[1] Across various industries, companies are digitizing 20 to 25 times faster than pre-pandemic estimates. [2]

As digitalization increases, the demand for infrastructure to support these technologies follows. Notably, pandemic restrictions and lockdowns have made it increasingly difficult for companies to manage their own on-site hardware infrastructure despite growing demands for digital communication channels and services. Cloud computing is taking on the challenge to meet this increasing demand for hardware and software infrastructure, led by the major tech companies Amazon, Microsoft, and Google.

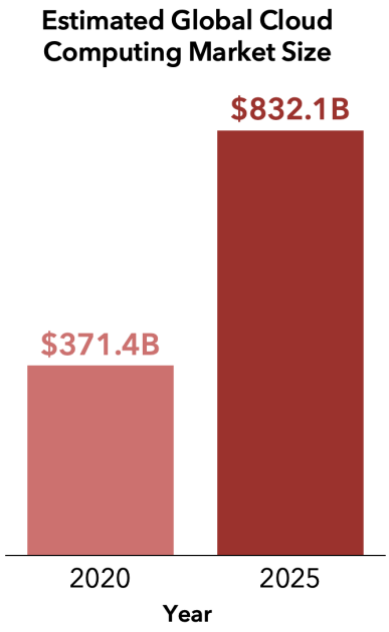


Exhibit 1: *Cloud Computing Market Size (2020-2025)*
Source: Markets and Markets

Cloud computing is "the delivery of computing services – including servers, storage, databases, networking, software, analytics, and intelligence – over the internet."[3] In other words, **cloud computing shifts the responsibility, technical knowledge, maintenance, and upfront investment costs of hardware and software infrastructure to a renewal cloud service.** This allows companies to focus their resources and efforts on other aspects of their business by automating the management of infrastructure and tools necessary to keep their technology online.

In comparison to traditional technology, cloud computing has many benefits, ranging from automation and ease of use to pricing and flexibility. Especially for smaller businesses or new startups, cloud services cut down on the initial investments necessary for supporting a service. Instead of making a large down payment to invest in physical infrastructure and hardware, companies can rent out service on-demand and pay for only what they need.

According to David Linthicum, Chief Cloud Strategy Officer at Deloitte and cloud computing thought leader, one of the biggest advantages of cloud computing is the agility of adding, removing, and adjusting resources. He adds, "there's almost no latency between our desire to get those resources and actually getting those resources in place" because of the on-demand pricing and service.[4] This significantly reduces the lead time of adding new resources since

[1] McKinsey: COVID-19 Digital Transformation
[2] Ibid.
[3] Microsoft Azure: What is cloud computing?
[4] HCCG Interview with David Linthicum

companies no longer need to engage in the tasks of purchasing, configuring, updating, and maintaining the added resource.

But **the most revolutionary advantage that cloud computing has over traditional server architectures and tools is the ability to automate management, deployment, and scaling**. There are three major categories of cloud computing services that are offered: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), which require differing levels of technical knowledge for varying amounts of configurability. IaaS and PaaS require greater technical knowledge and are more targeted for developers and IT. SaaS tools and software can often be used without the need of IT knowledge and support. The infographic below explains these tools in more detail.

4.2 Cloud Computing Trends

Despite the many benefits of cloud computing, established companies face several challenges in the transition to the cloud, including discomfort with change and overwhelming complexity/choices.

> *"There's* **inertia** *there that makes it difficult for people to get past; there's a* **strong incentive to keep things the way they are if it is working**."
>
> **-Tom Krazit, Senior Reporter at Protocol,** on difficulties of cloud adaptation

According to Tom Krazit, senior reporter at Protocol, "one of the biggest barriers to adapting cloud technology is cultural – it's very difficult to get traditional IT folks to think differently."[5] For conservative, risk averse companies, the fear that transitioning to cloud services will break currently working services and applications is commonplace. Many companies are also uncomfortable with moving the core components of their infrastructure to the services. To alleviate the discomfort, some will take the approach of hybridization, using a portion of on-site infrastructure to continue supporting the core components. Offloading non-primary components to the cloud is a common work-around.[6]

Many companies are overwhelmed by the abundance of choice given the many cloud computing companies and services to choose from.[7] This problem is exacerbated by the tendency for cloud companies to lack standardization: different departments often use a different set of applications, tools, or services, which leads to confusion and inefficient adoption of new technology.[8]

[5] HCCG Interview with Tom Krazit
[6] Leading Edge: Types of Cloud Computing
[7] HCCG Interview with David Linthicum
[8] Ibid.

Types of Cloud Computing Services

IaaS
Infrastructure as a Service

IaaS delivers cloud computing and **raw hardware infrastructure** that consist of **servers, storage, network, and operating systems**. Out of the three types of cloud services, this one provides the **greatest amount of customizability** but the client is responsible for the management and configuration of the infrastructure. IaaS is often **used by startups and small companies** to avoid spending time and money on purchasing hardware.

Google Compute Engine · amazon web services · Azure · DigitalOcean

PaaS
Platform as a Service

PaaS delivers a **framework for developers to create and host applications** on. It provides developers the freedom to build software with **automatically configured and managed operating system, software updates, storage, and hardware**. The service usually is provided in a "virtualized" format, so resources can be scaled up or down easily. PaaS applications are most frequently **used by companies that utilize a hybrid model**.

Elastic Beanstalk · Google App Engine · HEROKU · OPENSHIFT

SaaS
Software as a Service

Also known as **cloud application services**, this category of services is the **most commonly used cloud tool by all businesses**. Usage of these tools and services are frequently **through the web, eliminating the need for IT staff** and deep technical knowledge to configure and manage applications on individual computers. Although easy to use, SaaS applications are usually l**imited in customizable options** and often provides only **limited support for integration with on-premise apps**, data, and services.

Office · salesforce · Cisco webex

Sources: BMC Software, vXchange

4.3 Big Data and Machine Learning

While tools and services are being digitized, companies have begun to take advantage of the growing trend to build tools capable of collecting large quantities of data for analysis or training data for creating machine learning models.

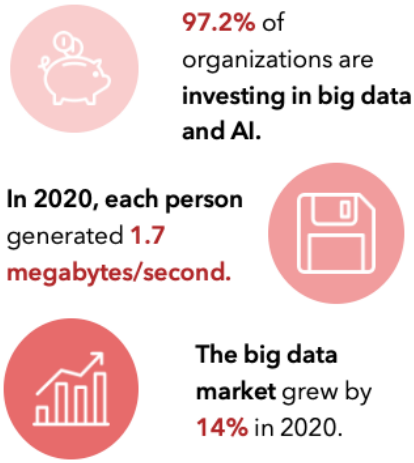


Exhibit 2: *Big Data Statistics*
Source: Statista & New Vantage

The value of big data comes from the insights it can provide for companies to discover opportunities and make better strategic decisions. Big data can create better forecasts. This helps companies evaluate customer behavior and product performance but also detect fraudulent behavior.[9] Netflix, for example, is capable of using big data to forecast customer demand for new products to assist them in the process of product development, design, and launch.[10] As behavioral data in customer purchases and credit activity is tracked, any subtle differences can be automatically flagged as potential fraud and action can be taken to mitigate or prevent further damage.[11] Likewise, information about a customer's purchase and shopping history can allow advertisers to display advertisements relative to the user's interests.

However, **the volume of big data makes it impossible to use traditional methods and data processing software for analysis and management.**[12] One of the most effective ways to extract valuable information out of big data involves the use of AI and machine learning models. Machine learning can do much more than basic statistical analysis or simple regression models. Through a variety of techniques, machine learning can identify patterns in data, categorize and sort unstructured data, act in real time to make automated decisions, or make complex forecasts/estimates.[13] Some of the most used technology include:

- speech recognition (e.g., Siri, Cortana)
- customer service (e.g., FAQ, messaging bots, "customers also purchased" suggestions)
- computer vision (e.g., self-driving cars, photo tagging in social media)
- recommendations (e.g., YouTube, Netflix, Spotify suggestions)
- forecasting (e.g., weather, automated stock trading)[14]

As digitization expands, the amount of big data will exponentially grow over time along with machine learning applications and new methods for analyzing data. The following sections address how companies are able to use collected data and analysis to make faster and better decisions in real time.

[9] MobiDev: 10 AI and Machine Learning Trends to Impact Business in 2021
[10] Oracle: What is Big Data?
[11] Ibid.
[12] Oracle: What is Big Data?
[13] IBM: Machine Learning
[14] IBM: Machine Learning

5. Top-Line Use Cases: Customer Facing

5.1 Big Data Analytics in Retail

The retail industry, with its high turnover and low margins, offers a compelling use case for big data. Retailers use big data to maintain a 360-degree view of their customers, optimize pricing, and create predictive models to forecast demand. Uncertainty in retail demand means that decisions in supply chain, staffing, logistics, marketing, and pricing are difficult to make. Big Data analytics resolve many of these uncertainties, allowing retailers to make more confident and informed decisions.

A. Product Innovation

Companies use big data to launch new products and ensure their success. A key application involves analyzing the relationship between not only their existing products but also their competitors products' attributes and their success in sales figures.[15] Customer feedback, social media, and third-party consumer data all play important roles in indicating potential opportunities. Trend durations, price, color variants, size variants, and use categories may affect success differently and can be altered to create a potentially more favorable product.[16] If the hypotheses and assumptions are supported by the data, gathered through beta product releases, the companies may move forward in releasing beta models and eventually studying customer feedback.[17] Big data ultimately allows for retailers to release innovative products with more certainty and improvement prior to launch.

Case Study: Gap

In an effort to turn around Gap's declining sales in 2017, Art Peck, former CEO of Gap Inc., fired his creative directors and decided to turn to big data to inform his decisions.[18] This new project, called Product 3.0, carefully **combined real-time customer purchase data, seasonal trend data, customer website search engine activity, and geolocations to indicate what kinds of products to launch** with their new line. First, potential items were released in small quantities at select stores to serve as a beta model prior to large quantity production. By **carefully analyzing the feedback from customers,** Gap could continuously improve their product offerings. Lastly, Peck implemented a shortened development cycle of 8-10 weeks which allowed for expedited turnaround time to customer response. Over the next year, GAP net sales increased by 725 million, the largest annual increase since 2011.[19]

[15] Oracle: 22 Big Data Use Cases
[16] BoltGroup: How to Use Big Data to Drive Product Innovation
[17] Cleverism: Big Data and Product Development
[18] HBS Case Study: Predicting Consumer Tastes with Big Data at GAP
[19] Statista: GAP Net Sales

B. Customer Experience

Retailers use big data analytics to create personalized offers, handle complaints quickly and effectively, and improve customer interactions to build long-term customer loyalty. Personalized offers to loyal consumers incentivize customers to continuously make purchases with a brand, especially as it feels like a reward to save more by spending more. E-commerce data assists in improving the customer experience, as it helps retailers create personalized offers and discounts to the customers based on their interests and activity.

Exhibit 3: *GAP's Big Data Model*
Source: HBS Case Study

Quickly and accurately responding to customer complaints significantly improves customer experience. One key difficulty is that customers prefer different mediums of communication, so companies must maintain several platforms and respond quickly. In order to address this, in 2017, Hubspot, a customer relationship management (CRM) software provider, acquired Motion AI, a company that developed chatbots.[20] The chatbots used artificial intelligence to automate responses to customers over Facebook Messenger, social media, email, and many other platforms. By 2020, it was predicted that 90% of customer service issues could be resolved positively using chatbots due to their almost-instantaneous response times and "human-like" interaction. Resolving issues quickly significantly contributes to a positive customer experience, thus playing an important factor in fostering brand loyalty.

Exhibit 4: *Importance of Personalized Offers to Consumers*
Sources: Invesp, Accenture, Statista

[20] HBS Case Study: HubSpot and Motion AI

Case Study: Sephora

Sephora, a French international beauty product retailer, **utilizes numerous outlets of personalized experiences** for its Beauty Insider members both online and in-store. The Sephora mobile application serves as an "in-store companion" helping customers find products and also schedule appointments with the in-store concierges. With every purchase, members receive loyalty points, rewarding those who spend more with exclusive offerings. When shopping online, personalized questionnaires and past purchase data are analyzed to suggest the best products for the individual. Today, Sephora's loyalty program has over **25 million members**, and of Sephora's total transactions in 2018, **80% included Beauty Insider members.**[21]

Exhibit 5: *Sephora's Beauty Insider Program Elements*
Source: Forbes

C. E-Commerce and In-Store Experience

Over 6,300 brick-and-mortar stores closed in 2020 due to COVID-19, which furthered retail's adoption of e-commerce.[22] Professor Kris Ferreira, Harvard Business School's Assistant Professor of Business Administration, emphasizes that with e-commerce, "you can track customers' click stream data to see what they're considering, clicking on, and even how far they're scrolling on the page." This allows retailers to access more data and analyze what may encourage purchase completion online. E-commerce data can be categorized as structured (age, locations, transactions) or unstructured (social media clicks, tweets, and influencers) where both are combined to make conclusions in retail.[23] Combining types of data, such as geolocation and search data, has allowed retailers to determine which in-store locations they should use to deploy new products. Additionally, many retailers have also been improving the online and in-store experience by involving smartphones, which are expected to

Structured	Unstructured
Customer names	Reports
Phone numbers	Audio files
Emails	Video files
Addresses	Images
Transaction information	Social media posts & mentions
Product names & numbers	Customer reviews

Exhibit 6: *Overview of Structured vs. Unstructured Big Data in Retail*
Source: Institute of Applied Informatics at University of Leipzig

[21] Forbes: Sephora Loyalty Program
[22] Business Insider: Store Closures
[23] Institute of Applied Informatics at University of Leipzig: Big Data Analytics in E-Commerce

influence $1.4 trillion in retail sales by 2023.[24] The continuous influx of big data allows for the rapid upkeep with modernizing and streamlining both the e-commerce and in-store experience.

D. Price Optimization

Margin analysis using big data can optimize prices. Professor Rama Ramakrishnan, Professor at MIT's Sloan School of Management, emphasizes that big data in retail pricing is especially useful, as the best applications of big data currently come in influencing "little decisions that have to be made numerous times over the course of a day."[25] Traditional methods of setting prices include production costs, standard margins, and the merchant's "intuition." However, these alone do not maximize profitability.[26] Big data analysis of customer demand, combined with competitor's pricings, advertisements, and feedback, can help retailers set optimal prices. Tracking seasonal customer activity allows for dynamic pricing that can help maximize profitability at different points of the year.[27]

One pricing method in its early stages of adoption is coined the "exploration and exploitation tradeoff."[28] When retailers are unsure of what the demand will be for a new product, they can price it at multiple price points to test the consumer response. After the initial "exploration" phase, retailers then "exploit" the pricing that maximizes revenue for the rest of the product's duration. This strategy is analogous to reinforcement learning in computer science, but has yet to be widely applied in retail.

With huge sales like Black Friday and Cyber Monday, retailers rely on big data to forecast purchases and profitability. Retailers can use items that were kept in consumers' online carts or viewed to predict the demand for that product during the sale period.[29] Data from advertisements also help predict demand, especially with social media clicks and interactions.[30] Additionally, highly rated and reviewed products are often sold in high quantities during these sales, so retailers can adjust these sales by a slim margin. By forecasting the demand for products prior to the sales, companies can adjust their sale prices accordingly to maximize their revenue during the sale period.

These new data-driven developments allow retailers to have more pricing certainty and optimize profitability.

[24] Retail Dive: Retail Sales
[25] HCCG Interview with Professor Rama Ramakrishnan
[26] McKinsey: Using big data to make better pricing decisions
[27] RCG Global Services: Competitive Pricing Insights
[28] Harvard Business School: Online Network Revenue Management
[29] Forbes: The Data Behind Black Friday & Cyber Monday
[30] Think Big Business: Big Data Behind Black Friday

5.2 Social Innovations in Healthcare

Utilizing deep learning assistive analytics in clinical decision making and diagnosis has the potential to improve patient outcomes and reduce expensive labor costs. This section explores three areas in healthcare that are prime for AI disruptions: cancer diagnostics imaging, patient management, and mental healthcare.

A. Cancer Diagnostic Imaging and Triaging

AI is able to consistently out-predict radiologists in screening for cancer by using deep learning analysis at high accuracy levels on training and test data. Although these findings have yet to be replicated in practice, the question of when these assistive machine learning algorithms can begin to supplement expert radiologist care, on a systematic basis, remains. This assistance would be especially valuable for lung cancer detection in particular. Early lung cancer detection is difficult, but it leads to much higher survival rates for patients.[31]

Lung cancer has a 25% 5-year survival rate on average after diagnosis. However, if it is caught early, survival rates increase to 66%. The low average survival rate among lung cancer patients is because 70% of lung cancer cases go undiagnosed until later stages. Universal yearly screening among at-risk patients has been one of the most effective and economical health care interventions in terms of dollars spent per year of life saved. These types of cost savings are why AI is an attractive way to reduce healthcare spend.[32]

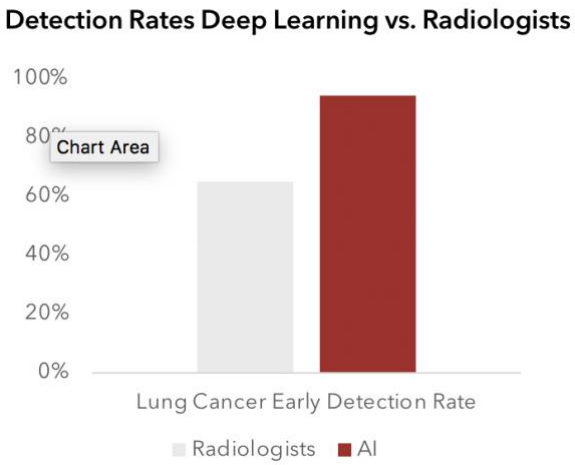


Exhibit 7: *Detection Rates for Radiologists vs. AI*
Source: Nature

New deep learning models have a substantial edge over radiologists' imaging analysis capabilities, particularly in early-stage imaging of lung cancer, presenting a clear opportunity where AI augmented care can vastly improve patient outcomes. New deep learning models can detect early-stage lung cancer in 94% of cases where it is present. Radiologists can only detect it early in 65% of cases.[33] This nearly 30 percentage point difference in detection could reduce the cost of critical care down the line by a minimum of $50,000 per lung cancer patient, reducing the need for more expensive late-stage cancer care. The implications of this detection differential are made more positive by the fact that diagnostic imaging through radiology currently costs the U.S. healthcare system $100 billion dollars a year. The adoption of AI imaging technology may help reduce costs and save lives by improving early detection rates and by assisting radiologists in providing the most accurate diagnoses for their patients.

[31] Nature: AI is improving the detection of lung cancer
[32] Ibid.
[33] Ibid.

Universal screening of the general population for lung cancer, with only high-risk patients triaged and checked by radiologists, could substantially lower long-term critical care costs and become a cheap and effective form of primary care intervention.[34] When put into practice, on a triage level, these deep learning imagining analysis models have the potential to greatly reduce radiological labor time by an estimated 50%, increasing the supply of radiological care and reducing health spend on hours of care needed. New models can triage effectively and send radiologist cases only where there is evidence of cancer. Thus, radiologists can reduce time spent on screening, instead focusing on more pertinent areas of patient care. Radiological imaging is the clearest case where health outcomes can be improved, creating massive efficiency gains in U.S. healthcare.[35]

Similar imaging analysis technologies have been effective for other cancers. A future where cancer screening and imaging are universal and cheap could be in our grasp. However, significant regulatory burdens remain in implementing assistive deep learning imaging analysis technologies. Further, incentives at individual and decentralized radiology practices seem to be misaligned with the speedy adoption of these systems. However, it seems likely that assistive deep learning technologies will improve the productivity and quality of radiological imaging.[36]

B. Deep Learning and Patient Management

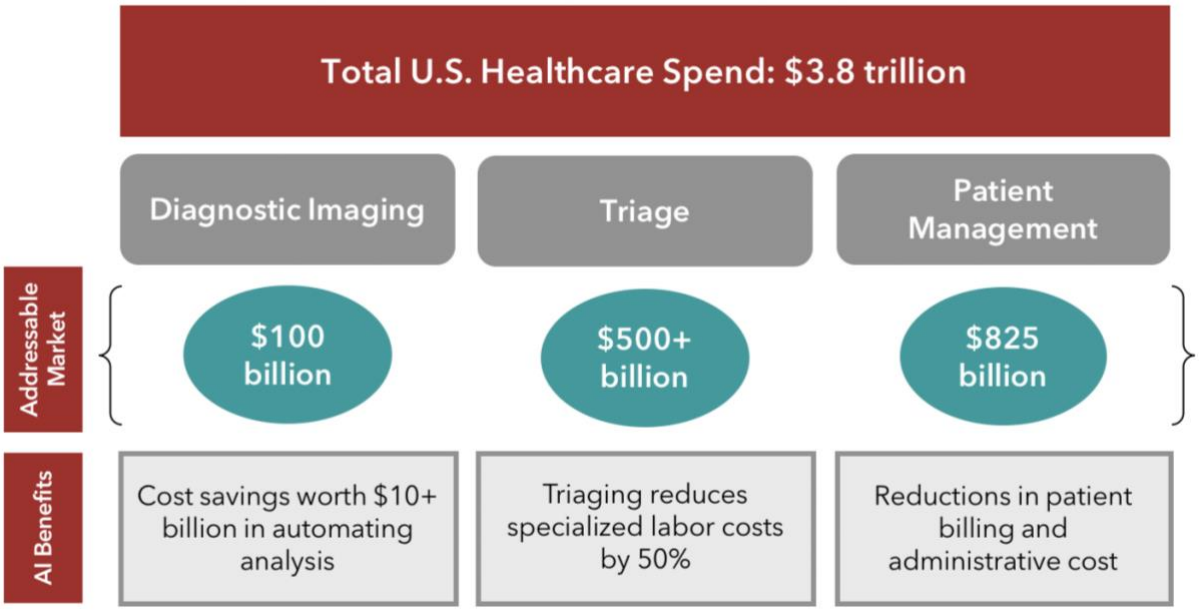


Exhibit 8: *AI Benefits and Use Cases within Healthcare*
Sources: Lancet, Journal of American Medical Association, Nature, British Journal of General Practice

[34] Nature: AI is improving the detection of lung cancer, The Lancet: Effect of AI-based Triaging

[35] The Lancet: Effect of AI-based Triaging, Journal of the American Medical Association: U.S. spends the most on healthcare

[36] The Lancet: Effect of AI-based Triaging

AI and deep learning decision assistive technologies are beginning to be used in primary healthcare and patient management around the world. Assistive patient management technologies have been implemented by the National Health Service (NHS) in the United Kingdom.

The main uses of patient management technologies are clinical decision making and care management around a variety of health conditions. New deep learning models can proactively analyze patient records to predict which patients have currently undiagnosed conditions before they need help. These systems are particularly valuable for disconnected and often non-mobile older patients who are not fully capable of reporting symptoms. As proactive detection continues to progress, experts indicate that the future of healthcare likely involves digital integration of data in patient care. Most saliently, AI has the potential to amplify human strengths, automate diagnoses, and help complement and supplement insufficient numbers of healthcare professionals.[37]

C. Natural Language Processing (NLP) in Healthcare

Natural language processing (NLP) uses "computer-based linguistics and artificial analysis" to extract meaning from text and text-based data. Significant promise has been shown in using NLP to extract data from medical documentation such as "progress notes, procedures and pathology reports, and laboratory test results." These findings indicate application in improving diagnostics for colonoscopies and irritable bowel syndrome, and, most unexpectedly, being able to link seemingly unrelated cases in electronic health records. NLP's promise in healthcare is the largest unknown but could prove to revolutionize how physicians traditionally diagnosis physical and mental health issues by incorporating big data. By improving diagnosis, NLP could increase the quality of patient care and reduce health spend on chronic issues later in life which may have otherwise gone untreated.[38]

Other NLP models such as cTAKES, CLAMP, SemEHR, GATE, and CoreNNLOP have shown promise in diagnostics for mental health related cases. These findings have been made possible by the increased availability of alternative documentation around mental health. Namely, this documentation provides "behavioral, emotional, and cognitive indicators as well as cues on how patients are coping with differ[ing] conditions and treatments." The three main challenges of using NLP to confront mental health are data availability, evaluation workbenches, and reporting standards. These metrics essentially call for increased access to data sets and universal standards as to how data is shared and reported on. Ultimately, a dialogue on how to realize the benefits of making health data more accessible to researchers, whilst preserving patient rights to privacy, must be held. However, NLP's specific promise in terms of mental health could make diagnosis much easier for populations who traditionally do not seek mental health resource. Diagnosis would no longer require a patient to seek treatment initially but instead could be made by looking at patient's long-term data.[39]

[37] British Journal of General Practice: Artificial intelligence in primary care
[38] American Gastroenterology Association: Current and Future Applications of NLP
[39] Journal of Biomedical Informatics: Using clinical NLP for health outcomes research

6. Bottom-Line Use Cases: Internal Processes

6.1 Supply Chain Management

AI's market trend forecasting abilities allow businesses to predict demand for products and determine the proper stock to avoid waste or product unavailability.[40] As these forecasting tools become more accurate, they become an essential part of supply chain management. Professor Ryan W. Buell in the Technology and Operations Management Unit at Harvard Business School emphasizes that, "the companies that are able to get better, more predictive results out of the technology and are able to respond to it quickly are the ones that will thrive."[41] This section examines how AI and cloud computing are used for predictive forecasting, operations management, and performance monitoring along companies' supply chains.

A. Predictive Forecasting

Day-to-day business operations are increasingly incorporating AI and cloud computing. These tools have the potential to reduce costs dramatically; cloud computing only costs around 11% of the cost of comparable on-premises software. [42] Office tasks are also increasingly automated. Oracle reports that the most common ways that artificial intelligence is implemented in the workplace are as digital assistants or chatbots, collecting data on customers as well as employees, conducting training, and processing job applications.[43] Buell explained the value of using big data in staffing, emphasizing how "analytics can [help] inform … how many people we need and when we need those people."[44] Especially during high-volatility periods like the pandemic, these technologies allow us to "leverage real-time information to … make better choices about how many people we need to have around, or strategic decisions about where to locate fulfillment centers, for example, or how we see the markets growing."[45] The ability of AI and cloud computing to cut costs and maximize productivity make them attractive to businesses across industries.

B. Operations Management

Cloud computing is improving the efficiency of each element of the supply chain.

Case Study: FedEx

Before cloud computing, FedEx struggled to analyze large datasets to address customer service requests. After adopting the computing service CloudX in addition to its other software—FedEx® CLI (Critical Inventory Logistics), ROADS (Route Planning and Optimization System), and Salesforce Automation—FedEx was able to reduce its response time by 60%, optimize delivery routes, and fulfill an average of 160,000 orders every month. Its adoption of Salesforce, a hybrid cloud service, allowed it to obtain real-time tracking of shipping and logistics, including inventory statistics globally and status updates from initial order to delivery.[46]

[40] Cloud Computing in Supply Chain Management: An Overview
[41] HCCG Interview with Professor Ryan Buell
[42] Hurwitz Study
[43] Oracle: From Fear to Enthusiasm Artificial Intelligence Is Winning More Hearts and Minds in the Workplace
[44] HCCG Interview with Professor Ryan Buell
[45] Ibid.
[46] Cloud Computing in Supply Chain Management: An Overview

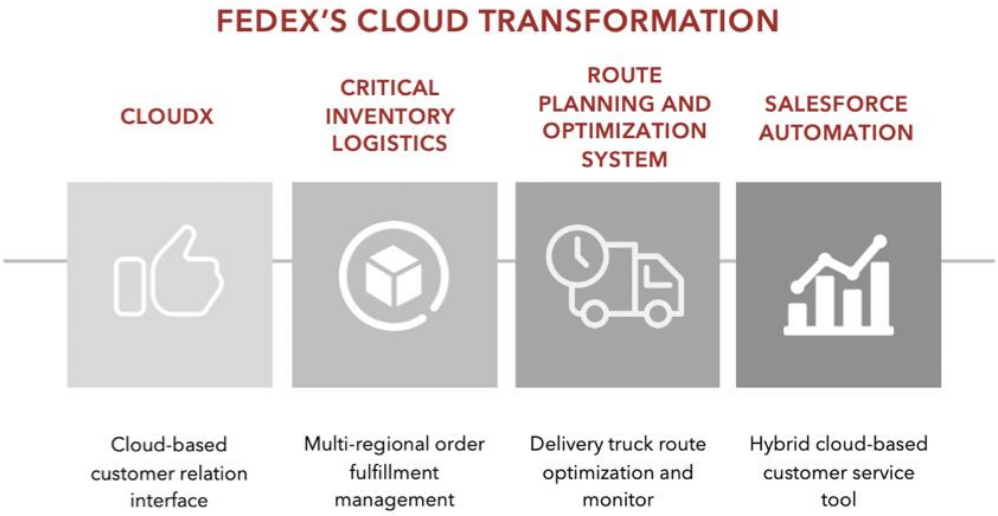


Exhibit 9: *FedEx's Use of Cloud Computing in its Supply Chain*
Source: Cloud Computing in Supply Chain Management

C. Performance Monitoring

As big data powers production at every level of the supply chain, it also gives companies the **ability to monitor performance and adapt its operations** accordingly. Artificial intelligence can identify disturbances in workflow. Utility companies obtain information from sensors and drones that supervise their electrical grids and use machine learning algorithms to identify when equipment starts to break down.[47] Similarly, in a joint program with Accenture called Taleris, General Electric has installed monitoring technology on its airplanes that is able to signal when equipment needs maintenance and give recommendations on operations.[48]

Case Study: Pfizer

Pfizer solved its supply chain traceability issue by requiring all of its 500 suppliers to use a cloud-based framework. This platform was able to handle over 40,000 shipments the first year and a half of its implementation—a dramatic shift from zero traceability. This allowed the company to sell to areas of the world where the industry did not have much of a presence previously. Pfizer could then tap into the pharmaceutical market in Kenya and know exactly when a product touched ground there. Even further, Pfizer was able to prove that its **shipments met temperature requirements** during transportation.[49] **Similar technologies have also been instrumental in delivering COVID-19 vaccines**; in the UK, hospitals have adopted cloud computing technology that monitors refrigerator sensors in real-time to ensure that the temperature-sensitive doses are stored correctly.[50]

[47] Application of Artificial Intelligence in Automation of Supply Chain Management
[48] Accenture: Supply Chain Management in the Cloud
[49] Financial Times: Pfizer Moves Supply Chain to Cloud
[50] CNBC: UK Hospitals are Using Blockchain to Track the Temperature of Coronavirus Vaccines

Predictive forecasting, operational management, and monitoring of performance provide **unity along all parts of production and sales** and allow companies to react quickly to changes or periods of uncertainty.

6.2 Risk Management

Risk management can broadly be viewed as the identification, understanding, and management of individual risk events.[51] Various types of risk exist in all kinds of business scenarios. Systematic risks affect entire markets, whereas unsystematic risks affect specific industries or companies.[52] For instance, loans carry credit risk—the possibility that a borrower will be unable to repay their debt—and the risk of fraud, among other examples, while churn risk, refers to the threat of customer churn, where customers cease to do business with a company. [53, 54] This section focuses on various prevalent examples of unsystematic risk in different industries, and how they may be mitigated by using big data techniques.

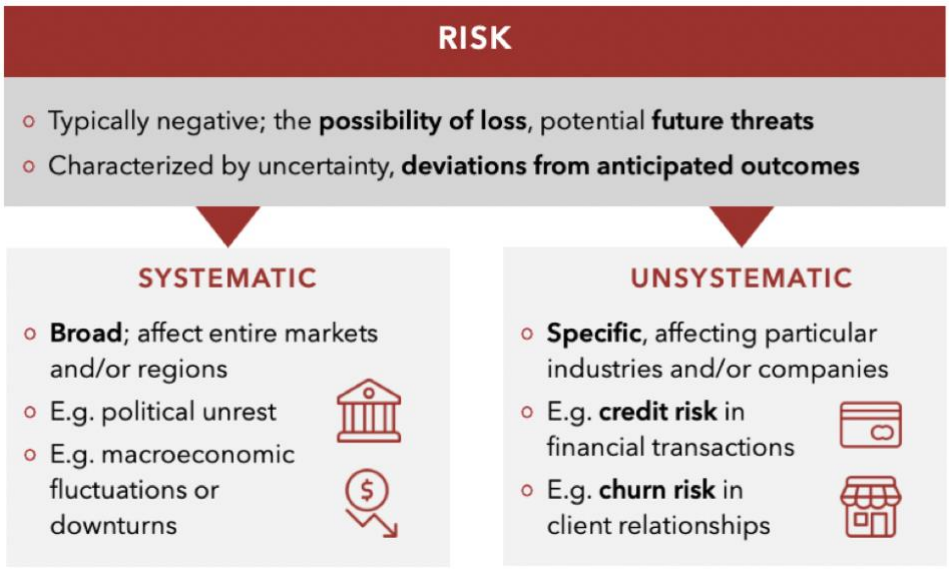


Exhibit 10: *Systematic and Unsystematic Risks*
Sources: Association for Project Management, Investopedia, Allianz

[51] Association for Project Management
[52] Allianz Risk Barometer
[53] Investopedia: Risk
[54] Investopedia: Churn Rate

A. Big Data in Risk Management Settings

Because of the prevalence of risk management in all kinds of business operations, it is no surprise that the subject has been an area of interest for big data. A CEBR report within the UK estimated big data tools to contribute approximately 38 billion pounds in efficiency savings from preventing manageable risk from 2012 to 2017, and 58 billion pounds in efficiency savings from 2015 to 2020.[55] A 2019 survey of over 4,000 organizations worldwide found that 77% of organizations found big data analytics to be important in fraud detection, with 22% finding big data to have "critical" importance.[56]

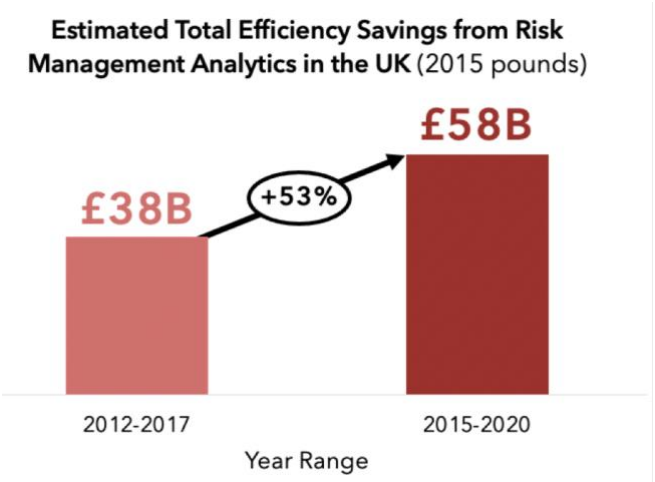


Exhibit 11: *Savings from UK Risk Management Analytics*
Source: CEBR

According to Professor Ramakrishnan, big data applications in risk management are "similar to non-risk applications." However, when analyzing data for fraud detection, the distribution between positive and negative outcomes is "highly imbalanced." That is, the number of data points truly indicative of fraud is extremely small relative to the overall data.[57] Building accurate models, therefore, requires diligence: a model that identifies every transaction as genuine would have a high accuracy rate, but would be useless. For the purposes of detecting fraud in financial services, supervised learning can be employed. Classification or regression models may be built and judged based on their ability to accurately identify fraudulent transactions.[58] Unsupervised learning can also be used, where, as Professor Frank Nagle of Harvard Business School describes, algorithms identify "different types of behaviors" among users and flag them for review. Such techniques have become increasingly prevalent as "the ability to understand an individual customer, their purchasing habits, and [their] credit usage habits has gotten better."[59]

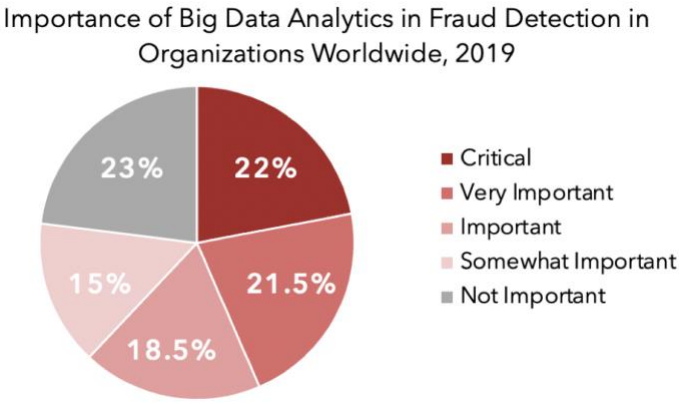


Exhibit 12: *Importance of Big Data in Fraud Detection Operations*

B. Fraud Detection

Models built for internal and external fraud detection have been implemented at several businesses, particularly within the financial services industry. The financial services corporation

[55] CEBR: The Value of Big Data
[56] Statista: Importance of Big Data Analytics Use Cases
[57] HCCG Interview with Professor Rama Ramakrishnan
[58] HCCG Interview with Professor Rama Ramakrishnan
[59] HCCG Interview with Professor Frank Nagle

American Express (Amex), which manages over 100 million credit cards and $1.2 trillion in transactions annually, began to develop AI models in 2010, using data pertaining to their operations in card issuance, merchant management, and their payment network. Amex's current model is capable of returning fraud risk decisions "in real time," contacting card members "within 15 seconds." Instances of fraud among Amex transactions are 50% less common relative to competitors, while disruptions to genuine spending among members are minimized.[60]

The accounting firm BDO also uses big data to identify fraud, focusing on data over interviews during audits.[61] The e-commerce and technology corporation Alibaba also uses a big data fraud risk management software named AntBuckler, which offers merchants the ability to see the riskiness of users on the platform via scores.[62]

The US Internal Revenue Service (IRS) similarly uses a model, the Return Review Program (RRP), to detect tax fraud. Under development since 2009, the RRP uses clustering methods to identify how fraudulent tax returns might be connected. The US Government Accountability Office estimates that the RRP prevented over $6.5 billion in invalid refunds being issued from January 2015 to November 2017.[63] States have followed suit; for instance, Utah, Massachusetts, and Arizona implemented their own models for catching fraudulent returns in the late 2010s.[64]

C. Other Financial Applications

Big data has found further use in other financial risk management applications. Brent Uemura, Senior Director of Technology at the Canadian Imperial Bank of Commerce (CIBC), states that a large use case within CIBC concerns "Know Your Client (KYC)" practices, a part of due diligence put into place to detect nefarious activities such as money laundering, the act of leveraging bank infrastructure to "sanitize" illegitimately obtained money, through monitoring transactions and analyzing behaviors with compromised cards and misappropriated funds. Models have been developed to recognize such behaviors, assign risk scores to clients, and identify insider threats. Furthermore, as Uemura states, big data infrastructure has been used in banks to efficiently manage data for regulatory reporting to comply with legislation like the Dodd-Frank Act, thereby managing compliance risk.[65]

> *"If there's a specific teller that has a touchpoint with a number of clients that have their cards compromised, then… the likelihood that you can* ***triangulate the source*** *to that teller is quite high."*
>
> **–Brent Uemura, Senior Director at CIBC Technology,** on using big data to detect internal fraud

The investment banking and asset management sectors use big data techniques for quantitative analysis. The Goldman Sachs Quantitative Investment Strategies team analyzes information from a variety of structured and unstructured data to understand factors that impact stock prices, assessing risk on a "real-time basis" and finding "insights and connections that aren't as obvious

[60] Forbes: How Amex Uses AI
[61] CIO Dive: 5 ways companies are using Big Data
[62] Jin, Tao, Wang and Chen: Alibaba Fraud Risk Management
[63] CIO: IRS Combats Fraud
[64] Ropes & Gray LLP: States Follow the IRS in Joining the Big Data Revolution
[65] HCCG Interview with Brent Uemura

to other investors" to aid portfolio managers' decisions. For example, natural language processing has been used to gauge sentiment about companies in the news and uncover relationships between companies through analyzing texts like "news articles, regulatory filings or research reports."[66] Morgan Stanley and the United Overseas Bank have also reported launching big data programs to analyze the riskiness of stocks in their portfolio.[67]

D. Churn and Market Risk

As previously discussed, big data can be harnessed to improve a company's customer experience. It can also be used, however, to also predict customer churn and manage **churn risk**. The telecommunications company T-Mobile has implemented big data techniques to predict and tackle churn among their customers, leveraging data from their customer relationship management and billing systems, and analyzing factors such as product usage, area coverage, and customer sentiment.[68, 69] Consequently, T-Mobile's churn rate among postpaid subscribers has more than halved over the past ten years.[70] Recently, after their acquisition of Sprint, the churn rate among Sprint users reduced by more than 60% from the fourth quarter of 2019 to the fourth quarter of 2020.[71] Amex has also used created a model that successfully predicted 24% of accounts that were expected to close in the next four months in Australia. This allowed Amex to take directed marketing measures towards those accounts to improve retention.[72]

Big data can also manage the risk associated with starting new locations, as seen with the café chain Starbucks. When considering new store locations, Starbucks draws from massive amounts of data, including distance to other stores, demographics, and traffic patterns, as factors in the recommendation process.[73] Furthermore, their model also predicts the impact of new stores on existing revenues in neighboring store locations, using geographical information systems.[74]

E. Limitations

The nature of data in risk management as introduced above is not the only limitation of big data in risk management settings. Traditional big data techniques work poorly in situations where results are needed in real-time, as they use "batch processing," where analysis is done after the fact. According to Uemura, this means that actions that are "on-demand" processing, such as disabling a credit card, are harder to perform using traditional big data infrastructure. Moreover, continuously updating data in traditional big data environments is uneconomical; traditional frameworks are better suited to reporting and analytics.[75] Professor Nagle also notes that "false positives" generated by models can result in inconvenience for the customer, which necessitates the development of more refined models to minimize these occurrences.[76]

---

[66] Goldman Sachs Asset Management: Role of Big Data in Investing
[67] MarTech Advisor: Big Data Helps Mitigate Risk
[68] Dataflog: T-Mobile USA
[69] MarTech Advisor: Big Data Helps Mitigate Risk
[70] Statista: T-Mobile US Churn Rate
[71] LightReading: T-Mobile
[72] HBS Digital Initiative: American Express
[73] Forbes: Starbucks
[74] Was Rahman, Medium
[75] HCCG Interview with Brent Uemura
[76] HCCG Interview with Professor Frank Nagle

Another limitation in fraud detection concerns how effectively models can continuously identify fraudulent behaviors. As Professor Ramakrishnan notes, perpetrators of fraud "can also download and use powerful tools like TensorFlow and PyTorch that businesses can," creating a never-ending race between businesses and their adversaries. For instance, bad actors could use machine learning to engineer cyber-attacks that pass existing detection algorithms, allowing them to operate undetected. This creates "new kinds of risk for which [businesses] don't have training data," meaning the business must assemble new data and re-train their detection algorithms to block attacks. One development in big data to accommodate this is to switch to unsupervised learning as mentioned previously, where algorithms do not specifically look for particular patterns and traits, and instead simply flag uncommon behaviors for manual review.[77]

In general, the use of big data is susceptible to overarching logistical and ethical problems regarding big data, concerning limited computing power, incorrectly-trained algorithms, privacy, and inflated expectations. These drawbacks are discussed in the following sections.

[77] HCCG Interview with Professor Rama Ramakrishnan

Big Data Use Cases in Risk Management

DETECTING FRAUD

- **American Express** has been developing AI models since 2010 and uses a model that can return decisions in real-time, achieving **fraud rates 50% lower** than those of competitors **turnover**
- **BDO** and **Alibaba** use fraud-detection big data models to aid their auditing processes and evaluate the riskiness of users, respectively
- The **IRS** has implemented a model to detect tax fraud, the **Return Review Program**, that **prevented more than $6.5B in invalid tax refunds being issued** from January 2015 to November 2017

AMERICAN EXPRESS BDO Alibaba.com IRS

TACKLING MARKET RISK

- **T-Mobile** analyzes factors such as product usage and sentiment to predict and act upon customer churn, **halving the churn rate** among postpaid subscribers over the past decade
- **American Express** used a churn prediction model that successfully **identified 24% of Australian accounts anticipated to close** in the next four months
- **Starbucks** uses big data and information such as store distances and traffic patterns to **estimate the risk associated with starting new locations**, as well as determine new locations' potential impact on existing stores

T-Mobile AMERICAN EXPRESS

OTHER APPLICATIONS

- **CIBC** uses big data infrastructure in **"Know Your Client" practices** to detect activities such as money laundering, assign risk scores to clients, and identify insider threats, and also to manage data for regulatory reporting
- **Goldman Sachs** processes structured and unstructured data using techniques such as NLP to **understand factors that may impact stock prices**
- **Morgan Stanley** and **UOB** are reported to have used big data models to **assess the riskiness of stocks** in their portfolio

CIBC Goldman Sachs Morgan Stanley UOB

7. Big Data Costs and Organizational Challenges

7.1 Fixed Costs of Big Data

Businesses seeking to incorporate big data into their operations must fund large up-front costs for infrastructure and technology. These fixed costs can be generalized into four groups: storage, processing, analytics software, and networking.[78]

Storage for big data manages large data sets, enables real-time data analytics, and utilizes an architecture based on both computation and actual storage.[79] Industry experts further describe big data storage using the three Vs: variety, velocity, and volume of data. Big data storage must be able to handle a wide variety of sources used to generate data imported in different formats (documents, emails, social media posts, etc.). Velocity refers to how quickly storage solutions can import large volumes of data and the speed at which analytic operations utilize it.[80] Volume simply brings attention to the large increase in data in today's era of exponential growth of consumer transactions and analytical variables.

Storage comes in the form of warehouse-based solutions and cloud-based solutions, the latter being the most cost-effective, secure, and common method as of recent. Warehouses are usually repositories located on-site, whereas cloud-based solutions can be remote and stored in off-site supercomputers. Popular cloud storage providers, such as AWS S3, Microsoft Azure, and Google Cloud have a significant cost advantage. Along with the storage itself are download costs for retrieving data from the cloud server, also called the networking costs.[81] There are also costs associated with the speed at which an organization wishes to retrieve this data, referred to as bandwidth. This can cost anywhere from $1.00 - $1.50 / megabit. Most companies use around 10 gigabits of speed, elevating prices to $10,000 - $15,000 per month. Storage pricing decreases as quantity request increases, while retrieval costs generally increase as quantity increases. Initial storage costs for businesses can land between $10,000 - $60,000.

Another salient cost of big data is data processing. Processing entails sifting through large data sets to extract insightful and relevant information to advance and support decision making. Processing tools utilize programming models, strategies, and algorithms to look for the most comprehensive data. For example, Hadoop is a dominant company that offers data processing tools. Hadoop connects a business's data to their servers to organize, replicate, and parallel process the data.[82] For companies utilizing petabytes (1,000 TB) of data, processing costs can come out to millions of dollars.

Analytical software provides companies with the ability to understand, draw insightful conclusions from, and make predictive decisions about their business's data. Predictive analytics work in tandem with processing – together, they allow for judicious decision making regarding

[78] Nor-Tech: 5 Infrastructure Requirements
[79] TechTarget: Big Data Storage
[80] Ibid.
[81] BackBlaze: Cloud Storage Pricing
[82] PhoenixMap: Hadoop

the project's goals. Software products take the processed data and formulate insights based on patterns (or lack thereof) and statistics embedded within the data. Analytical costs exceed those of processing, ranging anywhere from $15,000 - $30,000 per TB.[83]

The final and most expensive fixed cost of big data is the cost of human capital. Businesses pursuing big data must hire a group of specialists to successfully operate, maintain, secure, and troubleshoot infrastructure and data. The quantity of employees hired scales proportionally to the scaling of a big data venture--the more storage and data purchased, the more specialists needed to ensure success and eliminate error. Human capital is also required for data security, encryption, and segregation. Companies can lose millions of dollars in security breaches to hackers, so it is essential data is closely monitored and protected by specialists.[84]

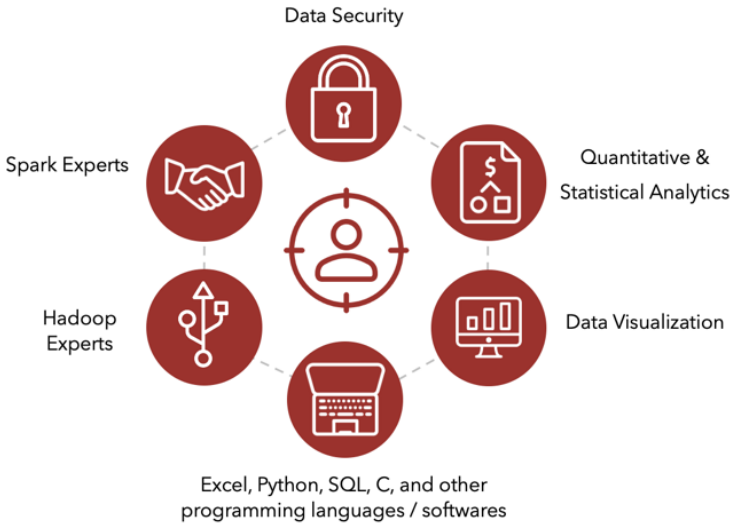


Exhibit 13: *Example of a Business's Top 6 Big Data Skillsets Needed from New Employees*

Exhibit 14: *Overview of Fixed and Variable Costs of Big Data*

[83] Cooladata: True Cost of Building a Big Data Solution

[84] Ibid.

7.2 Management and Variable Costs of Cloud Computing

Upscaling – increasing storage, bandwidth, analytical capacity, and human capital to expand business operations and incorporate more data – raises big data costs enormously.[85] After the initial investments of storage, processing, analytics, and others, upscaling is by far the biggest variable cost of cloud computing. Depending on the scope of the project, most companies will scale – multiply every part of the project by some factor (anywhere from just over 1 to over 100) – their operations at some point, resulting in a new cost that multiplies all sunk costs by that same factor. This is why upscaling is a very costly investment – yet with high upscaling costs come high potential rewards.

While data science has exploded in popularity among career options for college graduates, there still aren't as many professionals as there are in other sectors.[86] In a market tilted towards supply, businesses find it difficult to attract, manage, and retain that talent. According to Professor Lauren Cohen, the L.E. Simmons Professor of Business Administration at Harvard Business School, hiring for big data projects is based on both the value an employee brings and the value needed at the specified position. Retaining talent is difficult, as employees "can leave at any point" and take their newly acquired "special sauce" with them.[87] This "special sauce" refers to the experience, skills, and talent given to the employee by the company – a direct cost for the company. Professor Ryan Buell affirms hiring human capital to be the greatest challenge to companies: "the bottleneck is not the tech – the bottleneck is the people."[88]

However, it turns out that 30% of all capital invested into cloud projects and variable costs goes to waste.[89] This arises due to many reasons, including an improper understanding of big data, data growth issues, confusion with tool selection, and a dearth of data professionals.[90] Most of these issues boil down to either an insufficient understanding of business capabilities or a lack of sufficient funding and infrastructure. To resolve these issues, companies often hire consultants or acquire smaller firms to increase their capabilities.

Case Study: Walmart

For example, to address the shortcomings of their marketing strategies and predictive analytics, Walmart acquired Inkiru Inc, a small startup based in Palo Alto, California. Inkiru specialized in targeted marketing, merchandise, and fraud prevention, and offered a predictive technology platform to enhance analytical personalization.[91] This initial investment allowed for Walmart to jumpstart their project of **revamping their online shopping platform**. Now, Walmart's transformed decision making has resulted in over $1 billion in new revenue and 15% increase in online sales.[92]

[85] BleuWire: Big Data Challenges
[86] SMBCeo: Costs of Big Data
[87] HCCG Interview with Professor Lauren Cohen
[88] HCCG Interview with Professor Ryan Buell
[89] MindInventory: Cloud Computing Challenges
[90] upGrad: Major Challenges of Big Data
[91] Dezyre: Walmart Sales Turnover
[92] Ibid.

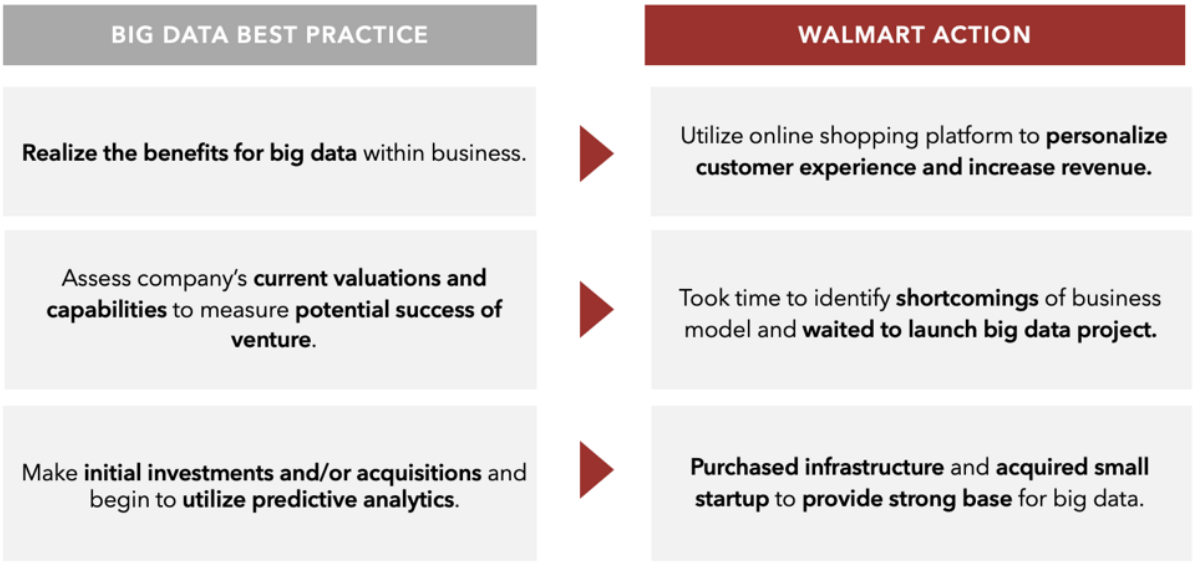


Exhibit 15: *Walmart's Big Data Decisions Compared with Industry Best Practices*
Source: Dezyre

7.3 Integrating Big Data into Business Decisions

Once costs are covered, businesses need to turn their predictive analytics into predictive decision making. While previous case studies exemplify the potential of big data projects, their success is not widely shared in the big data industry. According to multiple studies and surveys, around 85% of big data projects fail entirely, 87% of data science projects never reach full production, and by 2022, only 20% of predictive analytics will deliver business value and outcomes.[93] "You might think that if companies collect all this data, this is where the value is," states Professor Lauren Cohen. Professor Cohen emphasizes that "collecting data is not enough."[94] In fact, the many reasons for failure include mining the wrong data, hiring the wrong talent, or pursuing the wrong business problem.[95] These issues usually stem from weak management and misguided timing. In other words, strong and unified leadership from executives and overseers is essential at every phase of the big data project. This ensures that the proper data is used to support changes implemented at the proper time and to the proper customers.

There is no single blueprint for successful big data projects, but external parties can bolster a business's ability to make sound decisions when delivering new products and services to customers. For example, companies can hire consulting firms to help identify the best markets, goods, and customers to gear big data projects towards. Like Walmart, companies can even acquire small firms or start-ups in industries specific to their project. Either way, predictive decision making has incredible potential for a business's profit margin and future. However, big data projects can easily fail, which is why meticulous planning and foresight are crucial to successful big data projects.

[93] Data Science Project Management
[94] HCCG Interview with Professor Lauren Cohen
[95] Ibid.

8. Ethical Implications of Using Big Data

Whether it be in retail, healthcare, or financial services, the potential of big data to streamline supply chains and realize overall efficiency increases cannot be overstated. However, as with all new and emerging innovations, **it is important to take a step back and examine not only the potential risks behind the exploitation of big data, but questions of whether we should even be employing big data in certain situations.**

One of the major challenges in practicing ethical usage of big data is the obscurity behind data ethics itself. Luciano Floridi of the Oxford Internet Institute defines data ethics as the study of "moral problems related to data, algorithms and corresponding practices,"[96] but there remains considerable ambiguity on what these moral issues constitute. According to Professor Viktor Mayer-Schönberger, Professor of Internet Governance and Regulation at the University of Oxford, there is significant variability in how different firms interpret what ethical practices entail that setting a standard becomes difficult.[97] It has become commonplace for firms to establish ethics review boards without acting upon their recommendations, effectively practicing ethics whitewashing by relegating them to a relatively tokenistic position. Thus, whilst it is important to understand what data ethics entails, it is perhaps even more important to examine real-world data issues and how ethical usage of data may be practically applied**.**

8.1 Principles of Data Ethics

At a high level, there are a few principles that are universally accepted as integral to the ethical usage of data and have driven policy from regulators based in the European Union and the United States.[98]

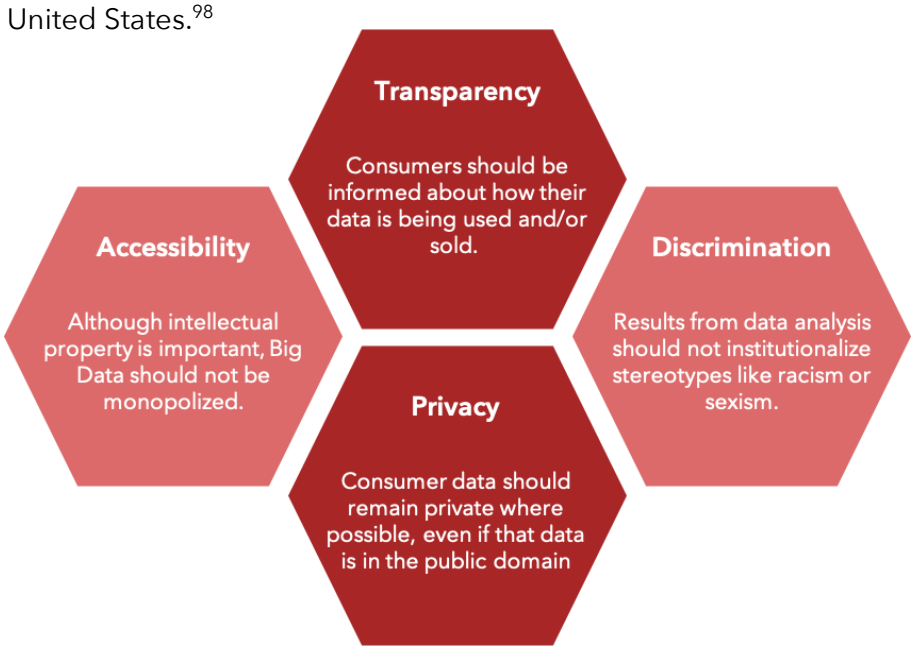


Exhibit 16: *Four Major Data Ethics Principles: Privacy, Transparency, Accessibility, and Discrimination. Source: Towards Data Science*

[96] Luciano Floridi: What is Data Ethics?

[97] HCCG Interview with Professor Viktor Mayer-Schönberger

[98] Towards Data Science: 5 Principles for Big Data Ethics

These principles correspond to the "Three Paradoxes of Big Data" presented by the Stanford Law Review.[99] Firstly, whilst big data collects copious amounts of information, the actual process behind that collection is "shrouded in legal and commercial secrecy." Secondly, using big data to fulfill consumer preferences inadvertently leads to the undermining of individual and collective identity. That is, practices like targeted advertising straddle the line between simple marketing and excessive manipulation of consumers' mindsets. Finally, although big data is "characterized by its power to transform society," its own power effects tend to disproportionately privilege large government and corporate entities.

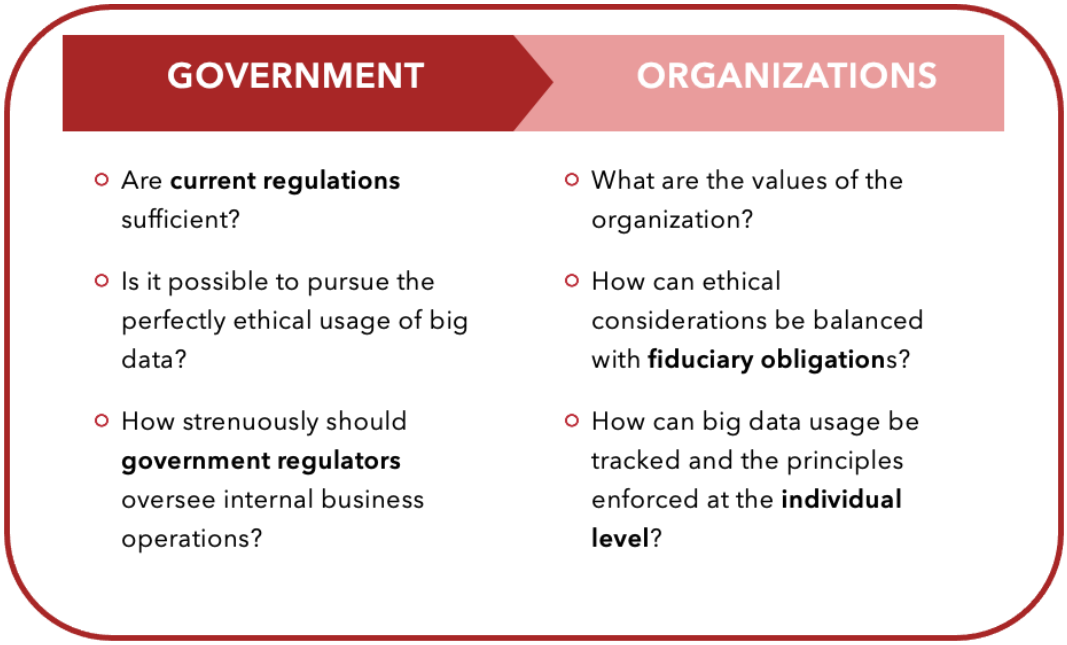


Exhibit 17: *Key Questions in Two Levels of Big Data Policy and Governance: Government and Organizations*
Source: HCCG Interview

Given this background information, specific issues within data ethics may now be explored.

8.2 Data Privacy

Over the past few years, Data Privacy has become the most prominent ethical concern when it comes to big data usage. There have been multiple high-profile cases, from Amazon[100] to Salesforce,[101] and widespread media coverage. However, with this attention comes major policy reforms and regulations to ensure consumer security. The European Union implemented the General Data Protection Regulation in 2018, which not only set up independent supervisory authorities within each member state, but also dictated several principles restricting the processing of personal data.[102] Similarly, the United States Congress

[99] Neil M. Richards and Jonathan King: Stanford Law Review, Three Paradoxes of Big Data
[100] Politico: Amazon Data Security
[101] Infosecurity: Oracle and Salesforce to Face GDPR
[102] Intersoft Consulting: General Data Protection Regulation

has held multiple hearings on "Consumer Privacy in the Era of Big Data"[103] and has heard several proposed bills, such as the Consumer Online Privacy Rights Act of 2019.[104]

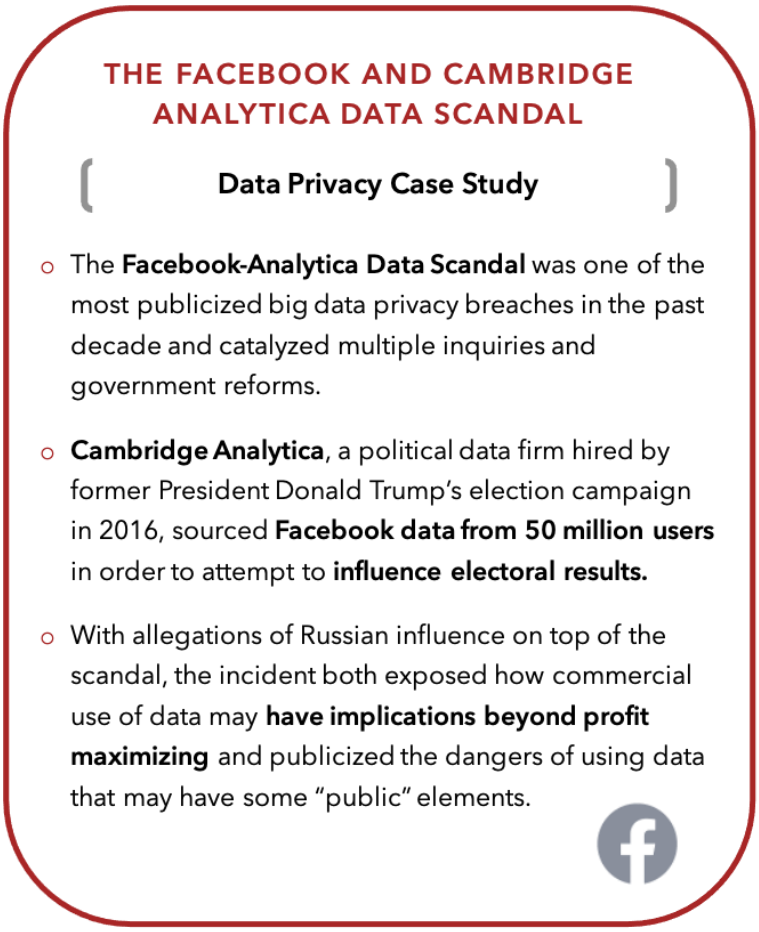


Exhibit 18: *2016 Facebook and Cambridge Analytica Data Breach*
Source: New York Times

These efforts have prevented most instances of overt breaches of data privacy; businesses have realized that focusing on subverting privacy regulations is neither feasible nor commercially advantageous. When one of Apple's largest marketing points emphasizes its premium on user privacy,[105] it is clear that big data is trending towards protecting users.

However, one facet of big data that presents immediate concerns to user privacy is Predictive Analysis, a practice that uses historical data and machine learning to predict future trends and consumer behaviors.[106] By making conclusions based upon data that may be only tangentially related to the analytical target, corporations can essentially turn public data into private data.

One of the most famous examples of this is Target using historical purchase data to predict which customers were likely to be expectant parents, sometimes even before immediate family members knew.[107] Although Target's usage of this data was relatively benign, providing the parent with coupons that conveniently highlighted infant clothes, the potential for more malicious usage exists. The dynamic nature of how and where consumer data is analyzed raises questions on whether users are truly able to consent to disclosing their data being used, even if they initially provided permission.

[103] Library of Congress: Protecting Consumer Privacy
[104] The Verge: All the Ways Congress is Taking on the Tech Industry
[105] Apple: Privacy
[106] PostFunnel: Should Predictive Analytics Be Subject to Government Regulation?
[107] Forbes: How Target Figured Out a Teen Girl Was Pregnant Before Her Father Did

8.3 Data Transparency and Accessibility

Although the protection of privacy is important, some corporations have been accused of using the issue to act unethically through excessive data protection and a lack of data transparency, a phenomenon that demonstrates the complexity of data ethics. This centralization of data may be the most immediate ethical issue that the data industry faces today.

Professor Mayer-Schönberger identifies concerns in data monopolization on three levels.[108]

1. **Loss of market competition:** Big data requires copious amounts of infrastructure in order to process and utilize properly, favoring large companies with existing resources and crowding out smaller businesses.
2. **Loss of innovation:** Businesses with otherwise innovative ideas cannot efficiently make products without data that is usually licensed by larger companies. One of the only possible methods of recourse constitutes being bought by larger companies.
3. **Loss of security:** Security breaches become far more devastating if data is monopolized by a few large corporations. Single points of failure present high vulnerability to attacks, potentially releasing private information.

THE GOOGLE ANTITRUST LAWSUIT

Data Accessibility Case Study

- It was widely reported in June 2021, that **Google** was once again the subject of a **European Union Antitrust Investigation.**
- **The investigation** is focusing on the company's abuse of its own advertising platforms in order to block competing technologies and restrict third party access to user data for advertising.
- This case demonstrates the inherent tension between **User Privacy and Fair Competition** in the utilisation of Big Data. It seems that European Union regulators are favoring the idea of **decentralization**, although there is uncertainty around if lawmakers are choosing the right side in this **trade-off.**

EU

Exhibit 19: *2021 European Union Antitrust Investigation into Google's Advertising Policies*
Source: Wall Street Journal

As data becomes more valuable, large corporations will seek to retain the value of the data they have by restricting access to it. This not only leads to the above three harms, but also means that corporate data usage is susceptible to becoming a black box due to the exponential level of usage complexity as scale increases.

[108] HCCG Interview with Professor Viktor Mayer-Schönberger

8.4 Data Discrimination and Bias

Finally, in many cases, corporate usage of Big Data involves decision making processes that are either hidden from, or cannot be understood by, the general public. This lack of transparency leads to a lack of scrutiny, especially concerning potential cases of data discrimination, such as Amazon's erroneous machine learning experiments. To analyze the suitability of potential recruits, Amazon employed numerous Artificial Intelligence models to evaluate their resumes based on their similarity to those of past hires. However, given the large proportion of men in previous Amazon hiring rounds, the AI algorithm began placing a premium on resumes that possessed masculine characteristics like being on all-male sports teams, or using aggressive verbs to describe previous employment experiences.[109]

Similarly, concerns have begun to emerge around the notion of "proxy variables," public data that can be utilized to predict private behaviors, as in Predictive Analysis. Specifically, there are currently wide-scale concerns about the ability of anti-LGBTQ+ governments to identify closeted LGBTQ+ members. One facial recognition technology model from Stanford was able to identify homosexuality with 81% and 71% accuracies for men and women respectively. Another model, which analyzed behavior on Facebook, was able to determine whether a social media account belonged to a gay or lesbian user with a similarly concerning degree of accuracy.[110] As technology continues to improve, models like these will increase in their predictive capabilities, further exacerbating the potential for data misuse.

Unfortunately, these are not the only social concerns facing big data. As Kirsten Martin, Professor of Technology Ethics, IT, Analytics, and Operations at the University of Notre Dame, states, "More and more powerful companies are using their influence through data to disenfranchise people."
Whether it be racial discrimination or religious intolerance, the ability of data analysis to internalize social justice is currently underdeveloped.

For big data to provide sustainable and ethical social utility, businesses must listen "closely to their customers" and, above all, implement "credible and clear transparency policies for data management."[111]

[109] ACLU: Why Amazon's Automated Hiring Tool Discriminated Against Women
[110] HBR: The Legal and Ethical Implications of Using AI in Hiring, Cornell: Bias Mitigation in AI Systems, BBC
[111] Gry Hasselbalch and Pernille Tranberg, Data Ethics: The New Competitive Advantage

9. Conclusion

In an increasingly digital world, companies and organizations now have a wealth of data available to them, both in structured and unstructured forms. However, advanced infrastructure is required to manage, store, and process the sheer magnitude of data generated. Advancements in cloud computing services provide companies with the IT resources and processing power necessary to keep up with the demands of widespread digitization. As organizations weigh the up-front and variable costs of migrating to the cloud, **companies need to also consider the cultural change and human capital required to successfully integrate business operations into cloud services.**

Crucially though, the value of big data does not come from simply amassing large volumes of data but rather from abilities to systematically extract and analyze datasets for hidden patterns and insights. Through capitalizing on nascent and developed technologies like machine learning and Natural Language Processing, **companies can use data strategically to reduce uncertainties, discover new opportunities, and make informed decisions.**

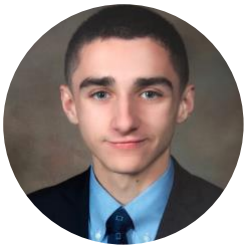
The primary use cases of big data manifest in various industries and currently revolve around prediction, innovation, and monitoring. For instance, in retail, predictive models allow retailers to analyze consumer data to personalize offers, optimize pricing, and improve customer experiences. Machine learning in healthcare revolutionizes diagnostics and screening, improves patient outcomes, and reduces health spend. In internal operations, analytics allow companies to reduce risks, increase efficiency, and monitor metrics at each point in the supply chain. Although this paper only presents the developments and trends found within a few use cases, the future of integrating big data into business decisions is clear: **it is more salient than ever before that companies combine qualitative business expertise with data-driven insights. Big data solutions, when implemented properly, grow business capabilities, and improve outcomes for both companies and consumers.**

However, ultimately, the centralized nature of big data presents a number of concerns in terms of fair competition, privacy, security, and transparency. As with all issues, there has been substantial progress made with the trend towards corporate social responsibility. **As consumers become more conscious of their digital footprint, businesses are realizing the competitive advantage in pledging to act ethically in all data related decisions.** While it is infeasible to suggest that it is possible to completely eliminate any ethical tensions when using big data, firms must learn to balance their fiduciary obligations with their duty to protect consumers.

Lauren Yang

Originally from Sugar Land, Texas, Lauren is a rising Junior studying Applied Mathematics and Sociology. Outside of HCCG, she has interned for the UN Foundation, conducts research on the limits of self-defense laws, and volunteers for organizations focused on addressing gender-based violence. In her free time, she enjoys spending the day at art museums and trying new foods.

Alex Fleury



Alex is a rising Sophomore from Massachusetts living in Quincy House. At Harvard, he studies Computer Science with a secondary in Economics, and is interested in pursuing a career in finance. Outside of HCCG, Alex is a proud member of the Varsity Track and Cross Country teams.

Annabel Cho

Originally from Minnesota, Annabel is a Junior in Quincy House studying Bioengineering with a secondary in Computer Science. Outside of HCCG, she is on Harvard Korean Association's board and conducts biomedical imaging research at Harvard Medical School's Martinos Center.

Charlie Yang

Born in New Zealand but raised in Sydney, Australia, Charlie is a rising Sophomore living in Dunster House and studying Applied Mathematics and Economics. At Harvard, Charlie debates for the Harvard College Debate Union, edits the Economics Review, and helps direct Model UN conferences. He is interested in pursuing a career in consulting or finance and loves to hike, listen to audiobooks, and play basketball.

Edward Dan

Originally from Ohio, Edward now lives in Winthrop House and studies Computer Science and Economics at Harvard. Outside of HCCG, he is also a member of the Harvard Computer Society and has interned for multiple early-stage tech startups. In his free time, Edward likes to code Discord bots, watch movies, and play the piano.

Eric Shen

Eric is a Canadian studying Applied Mathematics. Taking interest in a variety of topics in Mathematics and Computer Science, outside of HCCG, Eric is also involved in the Harvard Computer Society and Harvard Open Data Project.

Litsa Kapsalis

Litsa is from the Chicago, IL area. She is currently studying Biomedical Engineering. Outside of HCCG, Litsa is an innovation team lead in the Harvard Global Alliance for Medical Innovation and a summer biomaterials research intern with the Stupp Laboratory at Northwestern University. In her free time, she enjoys running along the Charles River and playing piano.

Maxwell Dostart-Meers

Max is a rising Junior at Harvard College studying Economics and Government. He is interested in the intersection of big data and labor economics and is currently an Emergent Ventures Progress Fellow.