2

Rethinking the Firm

In June 2018, a record $14 billion fundraising and $150 billion valuation made Ant Financial[1] the largest financial technology (fintech) firm and the most valuable unicorn in the world.[2] Spun out from Alibaba only four years earlier, Ant Financial was already worth more than either American Express or Goldman Sachs.[3]

Based in Hangzhou, China, Ant Financial expanded in only a few years to deliver an unprecedented range of services to more than 700 million users and more than 10 million small and medium enterprises. Ant Financial flourished initially by focusing on financial inclusiveness, offering a comprehensive suite of products to underserved consumers and businesses in China. Ant Financial gradually expanded to the entire market, enabling an increasing range of services from bike sharing to train ticket purchases, and even charitable donations.

At the heart of Ant Financial's success is its ability to leverage data to learn about its users' needs and respond with digital services to address them. The wide adoption of its services across China and, through the Chinese tourist markets, across the rest of Asia, Australia, and Europe provides vast amounts of data, which Ant Financial uses to inform decision making on everything from fraud risk to new product features. The data is assembled into a powerful, integrated platform that uses AI to power such functions as application processing, fraud detection, credit scoring, and loan qualification.

Ant Financial is creating a new template for the twenty-first-century firm—deploying an operating model that leverages digital

scale, scope, and learning to transform financial services and engage in a long-running collision with industry incumbents. Consider the operating model's efficiency: Ant Financial employs fewer than ten thousand people to serve more than 700 million customers with a broad scope of services. By comparison, Bank of America, founded in 1924, employs 209,000 people to serve 67 million customers with a more limited array of offerings. Ant Financial is just a different breed.

This chapter explores three rapidly growing examples of this new template for the twenty-first-century "digital" firm: Ant Financial, Ocado (in grocery delivery), and Peloton (in fitness). Each was created to enable new kinds of business models, with software, data, and AI as the primary operational foundation. Each is in a traditional industry, colliding with incumbent companies, reshaping how firms operate, and transforming the economy around them. The chapter concludes by focusing on Google, a more established firm that has placed AI at the core of its business and operations.

With their new approaches to creating, capturing, and delivering value to customers, these companies are leading the transformation of the economy. To understand how they are doing this, we first break down a firm into its business and operating models and analyze how it has traditionally shaped and executed on its value proposition. We then focus on how these three companies are forging a new path.

Value and the Nature of Firms

There's a well-developed understanding of the nature and purpose of the traditional firm. Economists like Ronald Coase and Oliver Williamson have declared that firms are formed to accomplish tasks that cannot be completed by individuals working through a market structure. We need firms, because coordinating each worker to engage in joint production through markets alone would require prohibitive transaction costs. Instead, firms provide long-term contracts to coordinate tasks without continually incurring the friction of continuous bargaining and negotiation and thus lower the transaction costs needed to create products and services. The value of these "bundles of contracts" is naturally shaped by the range of tasks organized by the firm—by what the firm promises to do and by how the

firm actually gets it done.

The value of a firm is shaped by two concepts. The first is the firm's *business model*, defined as the way the firm promises to create and capture value. The second is the firm's *operating model*, defined as the way the firm delivers the value to its customers.

The business model thus encompasses the strategy of the firm: how it seeks to differentiate itself from competitors by providing and monetizing its unique set of goods or services. Meanwhile, the operating model encompasses the systems, processes, and capabilities that enable the delivery of the goods and services to the firm's customers. The business model defines the theory, and the operating model captures the practice—what the people and resources of the firm actually do every day. And while the business model points to the potential of the firm, in terms of the value it *could* deliver, the operating model is the actual enabler of firm value and its ultimate constraint.

Business Models

A company's business model is therefore defined by how it creates and captures value from its customers. It's important to be precise. There are two elements that come together: first, the company must create value for a customer that prompts her to consume the company's product or service; second, the company must deploy some method to capture some of the value created.

Value creation, then, concerns the reason customers choose to use a company's products or services, and the particular problem the company is solving for customers. This is sometimes known as the *value proposition* or *customer promise*. Think of the car you drive. The auto company's value creation starts with solving your transportation problem. The car allows you to move around in the world. Beyond that, the car company creates value for you by delivering quality (how reliable and safe the car is), styling (how it looks), comfort (how luxurious the interior is), ride quality (how smooth or aggressive the engine and transmission are), cost (how affordable the car is), and the brand (the image of you that it projects). Just think of the value creation differences between, say, a Kia and a Ferrari.

The factors in value creation can, of course, change. For many of us, a car's technology package and its ability to interface smoothly with our smartphone are now important considerations.

Note that the factors you consider in buying a car are very different from those you'd care about in ride-sharing. When was the last time you canceled an Uber ride because a Toyota Prius was picking you up instead of your favorite Cadillac? Value creation in ride-sharing involves the availability of drivers and the wait time, trust in the company's policies on driver certification, customer ratings of drivers, the app's ease of use, and the cost of the ride.

So although both Toyota and Uber provide mobility, the value they create is very different. One makes you buy the car, whereas the other provides you a ride on demand. Thus a company's approach to value creation requires consciously choosing the precise problem it is solving for the customer and its positioning in the marketplace. In the case of ride-sharing companies, value creation also relies on an ecosystem of drivers and riders. The greater the number of drivers available, the more value created for riders, and because drivers are independent contractors who are paid by the ride, the more riders tapping the app, the more value created for drivers.

Value capture is the other side of the coin. Naturally, the value a company captures from a customer should be less than the value it creates for the customer. In our auto company example, the value capture for an auto company rests primarily on the fact that the sales price (P) of the car is greater than the cost (C) of manufacturing the car. So the margin, $P > C$, defines the value capture for an auto company. The company may also capture additional value through its leasing operation; here the company makes money by playing arbitrage in the capital markets by having access to lower interest rates than the consumer, and adds margin by selling spare parts.

The value capture story for a ride-sharing company looks very different; it is based on consumption, or *pay-per-use*. Instead of an upfront investment by the customer, the value capture relies on a customer's choosing to use the ride-sharing service time after time; 70 percent to 90 percent of the customer fee goes to the driver, and the ride-sharing company retains the rest. Margin still matters for ride-sharing, and the price should still be greater than the cost (a point

that seemed to elude both Lyft and Uber in their 2019 initial public offerings).

The new breed of digital firms is all about innovation in the business model, experimenting and recombining various aspects of value creation and value capture. In incumbent companies, value creation and capture are usually straightforward and closely intertwined: value is typically created and captured from the same source (the customer) through a simple pricing mechanism. In a fully digitized business, the options are much broader, because value creation and capture can be separated much more easily and often come from different stakeholders; most of Google's services are free to users, and the company captures value from advertisers across its product portfolio. For the digital firm, underlying all this business model innovation is a very different kind of operating model.

Operating Models

Strategy, without a consistent operating model, is where the rubber meets the air.

—Somewhat famous Italian proverb

Operating models deliver the value promised to customers. Whereas the business model creates a goal for value creation and capture, the operating model is the plan to get it done. As such, the operating model is crucial in shaping the actual value of the firm. A firm could promise to have an online retail business with nearly instant delivery; but to actualize that promise, the firm would need an impressive operating model characterized by an incredibly responsive supply chain. Devising and executing that operating model is where the real work would lie.

Operating models can be very complex, frequently including the activities of thousands of people, sophisticated technology, important capital investments, and millions of lines of code that make up the operational systems and processes that enable a company to achieve its goals. But the overarching objectives of an operating model are relatively simple. Ultimately, the goal of an operating model is to deliver value at *scale*, to achieve sufficient *scope*, and to respond to

changes by engaging in sufficient *learning*. The great business historian Alfred Chandler argued that the two main challenges faced by executives are to drive economies of both scale and scope in order to survive and thrive.[4] Subsequent work in economics and management showed that a third challenge is equally important: learning—the operating capability to improve and innovate.[5] Let's review these three operating challenges.

Scale: Managing scale, simply put, is about designing an operating model to deliver as much value to as many customers as possible at the lowest cost. Classic cases of improving scale involve efficiently increasing production volume or the number of customers served in, say, car production or fast food restaurants. Other examples may involve delivering products of increasing complexity in, say, completing a corporate merger or building an airport. From Ford to Goldman Sachs, firms are structured to make, sell, or provide more (or more complex) goods and services than individuals can, and to do so much more efficiently. A single person cannot efficiently manufacture an entire car in volume, nor can he produce the range of documents that are necessary to complete a complex corporate merger.

Scope: A firm's scope is defined as the range of activities it performs—for example, the variety of products and services it offers its customers. Some assets and capabilities can help an organization reach economies across diverse kinds of businesses. For example, having a centralized research and development organization can confer advantage across multiple product lines. Investing in a brand can deliver benefits for different products under the same brand umbrella. Having a centralized warehouse can achieve efficiencies across multiple product lines.

These economies of scope are important, because they enable corporations to establish multiple lines of business, perhaps managing multiple business units or creating a true conglomerate. With efficiencies of scope, firms can create and deliver a variety of goods and services efficiently and consistently. The Sears catalog operation, for example, was structured to efficiently deliver a wide variety of goods. A

hospital emergency room is designed to handle a variety of emergency conditions more effectively than individual physicians can handle on their own.

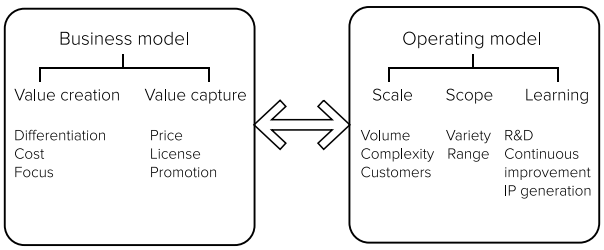
Learning: The learning function of an operating model is essential to driving continuous improvement, increasing operating performance over time, and developing new products and services. From Bell Laboratories' vast R&D impact to Toyota's continuous improvement process, modern corporations have looked to innovation and learning to remain viable and competitive. In recent years, the focus on learning and innovation has increased across the board to deal with threats and capitalize on opportunities.

As firms seek to deliver value and optimize scale, scope, and learning, their operating models should match the direction set by their business models. For many years, scholars in operations strategy have argued that the performance of a firm is optimized by the alignment between strategy and operations—in other words, between business model and operating model.[6] Not surprisingly, the resources of the firm should be deployed to optimize what it seeks to do. Figure 2-1 illustrates the idea of business model and operating model alignment.

From Ford to Sears, and from Bank of America to AT&T and General Electric, there is a long history of firms achieving superior performance by designing and implementing operating models that drive scale, scope, and learning objectives in alignment with their business models. Ultimately, the more the firm can drive scale, scope, and learning, the greater its value.

FIGURE 2-1

Alignment between a company's business model and operating model



At the same time, however, an expansion in each of the three operational dimensions increases the complexity of traditional operating models and makes managing them ever more challenging. This, critically, creates the operational constraints that have traditionally limited the value created and captured by firms. This is exactly where the digital firm differs. By deploying a fundamentally new kind of operating model, this new type of firm is reaching new levels of scalability, achieving a vastly broader scope, and learning and adapting at a much faster rate than does a traditional firm. This is because the digital firm is transforming the critical path in the delivery of value.

When digital technology, in the form of software and data-driven algorithms, replaces labor as the bottleneck in operating activities, the implications reach well beyond the obvious consequences for the workforce. Let's take a look at how three firms are driving business model innovation by transforming operating models and removing traditional operational constraints.

On a Collision Course with Financial Services

Ant Financial is built with scale in mind. There is no way that a human-centric approval process can be deployed here.
—Ming Zeng, Chief Strategy Officer, Alibaba

Ant Financial grew out of the success of Alipay, a payment platform created in 2004 by Alibaba, a then-nascent e-commerce platform, to facilitate payments for its shoppers and merchants.[7] Many of us now take online shopping for granted, but creating this service required Alipay to build a new kind of trust between buyers and sellers.

Many companies at the dawn of internet commerce worked hard to solve the trust problem. For Alibaba, which started as a peer-to-peer marketplace, the challenge was particularly acute: How could buyers trust the quality of the goods on offer, and how could sellers ensure that the buyers had the money to pay if the goods were shipped to them? The solution was to rely on an escrow system, wherein a third party holds payment until a contractual agreement is fulfilled. Alibaba thus invented Alipay as an escrow service for buyers and sellers on its

e-commerce platform. Users connected Alipay to a bank account, and Alipay acted as an intermediary, accepting payment from a buyer, holding it until the buyer confirmed receipt of the item, and then releasing payment to the seller. This system helped alleviate the consumer distrust of online shopping and was instrumental in driving Alibaba's early growth.

Therein lies the initial business model of Ant Financial and Alipay. Value creation is related to offering a substitute for trust in the form of an escrow-based financial payment service that facilitates transactions between merchants and buyers. Ant Financial must create value for two categories of customers: consumers and merchants. Value capture occurs through the 0.6 percent transaction fee charged to merchants; consumers are not directly charged for using the service.

Alipay's growth depends on increasing transaction flow, which can come not only from having existing buyers and sellers engage in more transactions but also from increasing the number of buyers and sellers. In other words, Alipay needs to increase both the *intensive* margin of transactions (how many transactions a user makes) and also the *extensive* margin of transactions by increasing the number of buyers and sellers on the platform.

It is at this point that the second element of value creation kicks in. As the extensive margin increases, the value of Alipay increases to all its users. When the number of merchants goes up, the number of buyers goes up. More buyers, in turn, attract more sellers. And thus a positive feedback loop is created, driving increasing returns to scale. This *network effect* amplifies the value created by trust in the service.

Soon after launch, Alipay made its service available beyond Alibaba's shopping platform to all individuals and businesses in China —a move that led to exponential growth, both contributing to and benefiting from the success of Alibaba's online marketplace. Two years after launch, in 2006, Alipay had 33 million users initiating 460,000 transactions per day. By 2009, that number had grown to 150 million users and 4 million transactions a day.

By 2011, with smartphone usage skyrocketing in China, Alipay gave customers the power to purchase items without cash in person, outside the Alibaba platform, via the Alipay app on their mobile phones. To

facilitate these transactions, Alibaba incorporated an established technology that did not require additional hardware, the QR code. A merchant sets up an Alipay account and displays the store's QR code in the store. Shoppers then open the Alipay app and scan the code to make a purchase, or generate their own QR code for the merchant to scan. Again, Alipay took a 0.6 percent cut of the transaction. Alipay users can use the app to buy coffee, hail cabs, pay utility bills, book medical appointments, split the bill with a friend at a restaurant, even make a donation to a street performer, as long as the vendor, or other party, also has an Alipay account.

Growth and Expansion

Alibaba CEO Jack Ma spun off Alipay because he feared possible government regulation of online payment systems. Alipay became the first product in the portfolio of the new company, Ant Financial, its name chosen carefully to represent the "little guy" the service targeted as customers. Alibaba retained rights to collect 37.5 percent of Ant Financial's pretax profits. Ant Financial's vision was to benefit society by facilitating a myriad of small transactions. Alipay and its rival WeChat Pay, launched by Tencent in 2013 (and discussed in chapter 1), grew rapidly and with no competition from the state-owned banks that dominated China's financial services, in part because they saw the internet payments market as unattractive. Use of Alipay quickly became ubiquitous in China and beyond as consumers and small and micro-enterprise merchants adopted the system. Some did away with credit card payments altogether in favor of Alipay.

Ant Financial did not stop for breath. The company took the data that it had access to and expanded the scope of its services to its clients and to the larger ecosystem. The conservative, traditional Chinese banks had created a massive opportunity for Alipay: only a small fraction of the Chinese population had access to credit, loans, or investment opportunities. Ant Financial jumped in with a sense of purpose and great speed to generate an array of services aimed at this huge market opportunity. Ant Financial extended its financial ecosystem with Yu'e Bao, an investment platform that allows Alipay users to earn interest on money in their accounts. Millions of Alipay

customers can transfer pocket change from their accounts into one of Yu'e Bao's money market funds and get a 4 percent annual return. Users can participate via mobile phone, and there is no minimum deposit required, making the service accessible to a broad swath of the market.

Within the first few days after launch, more than a million people put money into the fund. Eric Mu in *Forbes* described users checking their accounts first thing in the morning to see how much wealth they had accumulated overnight: "Yu'e Bao has created hundreds of millions of ultra-lightweight investors, for whom saving and investing is no more than playing a game, and like all games, this one is slightly addictive."[8] In nine months the fund collected more than 500 billion yuan ($81 billion). By the spring of 2017 Yu'e Bao had become the largest money market fund in the world.

Along with Yu'e Bao, Ant Financial rapidly extended its roster of financial services, adding Ant Fortune, a one-stop personal investment and wealth management platform; Zhima Credit, a social credit scoring system; MYbank, an internet banking services provider; an insurance platform; and a variety of other offerings. Ant Financial launched a number of other applications, all easily accessible from its Alipay app. They included education services, medical services, transportation, social functionality, games, dining reservations, and food delivery, to name a few.

Ant Financial's broad ecosystem of features and services led to dramatic increases in its installed base and in the engagement of each user. In only a few years, Ant Financial and its Alipay services have become ubiquitous in China and beyond, as the massive amounts of data accumulated in each application are integrated, analyzed, and fed back in a relentless effort to improve knowledge about customers, personalization, and innovation.

By 2019, Ant Financial had more than 700 million users and dominated much of the Chinese financial services market even as it faced competition from Tencent. Ant Financial controlled 54 percent of the mobile payments market in China, while Tencent's WeChat controlled 38 percent. As one industry insider told Don Weiland and Sherry Fei Ju of the *Financial Times*, "These companies are like Facebook if it had a bank on top of it and everyone had a bank

account [with Facebook]. There is really nothing like this in the west."[9]

In 2015, Ant Financial began to expand globally with investments in mobile payment systems in Asia, starting with a 40 percent joint stake with Alibaba in India's Paytm. From 2016 to 2018, Ant Financial continued to look for opportunities, pursuing partnerships and acquisitions that allowed the company to follow the needs of Chinese users as they traveled abroad. The company invested in South Korea's mobile payment platform KakaoPay, formed an agreement with Ascend Money (Thailand), Ingenico Group SA (a Paris-based payment system), Wirecard and Concardis (for Chinese travelers in Germany, France, the United Kingdom, and Italy), and acquired US-based biometric authentication technology company EyeVerify. Ant Financial attempted to penetrate the US market with a $1.2 billion purchase of money-transfer company MoneyGram but was thwarted by the US government due to fears about national security.

A New Kind of Operating Model

Alipay's rapidly expanding business model is built on a new kind of digital operating model. Its first foundation is a broad reliance on AI-enabled digital automation. For example, MYbank's hallmark is a 3-1-0 system for processing loans: it takes customers three minutes to apply for a loan, requires one second for approval, and involves zero human interaction. The loan approval and issuance processes rely solely on credit scores and are entirely digital and AI driven: each loan application is run through three thousand risk control strategies. Alibaba Group's Ming Zeng explains: "Our algorithms can look at transaction data to assess how well a business is doing, how competitive its offerings are in the market, whether its partners have high credit ratings, and so on." Zeng notes that Ant Financial's data analysts even feed its algorithms information on "the frequency, length, and type of communications (instant messaging, e-mail, or other methods common in China) to assess relationship quality" before approving a loan.[10] By January 2017, MYbank had served more than 5 million small businesses and individual entrepreneurs; loans averaged about RMB 17,000 and can be as low as RMB 1, with an aggregate loan

volume of more than RMB 800 billion ($18 billion).

The speed and efficiency of Ant Financial's MYbank system demand a huge amount of data processing. Ant relies on cloud computing technologies to keep data processing costs low in order to scale up. The company's computing infrastructure enables it to easily handle billions of transfers per day, with a peak workload capacity of 120,000 transactions every second, and disaster recovery solutions of up to 99.99 percent in place. According to the company, it can process loans at a cost of only RMB 2, compared with RMB 2,000 at a traditional bank. With these digital systems in place, MYbank does not need physical bank locations or a large workforce. In 2018, three years after its launch, the bank still employed only three hundred people, about the same number it started with.

The core of the operating model is a sophisticated, integrated data platform. With hundreds of millions of users making billions of transactions each day on the Alipay app, the platform collects information on everything users do, from the food they eat, to the places they shop, to the kind of transportation they prefer—not to mention how much they spend and how much they save. AI taps in to the data to drive a broad variety of functions, including personalization, revenue optimization, and recommendations, as well as the sophisticated analytics used to understand the value created by potential new products and services.

Alipay uses data and AI to ensure trust. When a user initiates a transaction, her information is passed through five layers of real-time digital checks to ensure that the transaction and the players involved are legitimate. Alipay's algorithms check buyer and seller account information for suspicious activity, look at the devices involved in the transaction, and then aggregate the data to make a decision on the validity of the transaction, much as a human might but much faster. Zeng explains: "The more data and the more iterations the algorithmic engine goes through, the better its output gets. Data scientists come up with probabilistic prediction models for specific actions, and then the algorithm churns through loads of data to produce better decisions in real time with every iteration."[11]

Ant Financial relies on data from four main sources: (1) internal consumer behavior statistics (e.g., records of relocation trends, utility

bills, money transfers, wealth management, purchasing patterns on Alibaba); (2) transaction data from sellers on Alibaba's platforms; (3) public data such as government databases containing criminal records, citizen identification information, and academic profiles; and (4) data from Ant Financial's partners (e.g., merchants, hotel and car rental partners) to power Zhima credit scores. Zeng explains:

> *Ant uses that data to compare good borrowers (those who repay on time) with bad ones (those who do not) to isolate traits common in both groups. Those traits are then used to calculate credit scores. All lending institutions do this in some fashion, of course, but at Ant the analysis is done automatically on all borrowers and on all their behavioral data in real time. Every transaction, every communication between seller and buyer, every connection with other services available at Alibaba, indeed every action taken on our platform, affects a business's credit score. At the same time, the algorithms that calculate the scores are themselves evolving in real time, improving the quality of decision making with each iteration.*

Zhima offers perks to consumers with good credit, such as favorable loan terms, whereas it requires those with low credit scores to put down additional deposits on their purchases, such as hotel rooms and bicycle rentals.

In addition, Ant Financial implemented a comprehensive, AI-driven fraud prevention monitoring system. This system can monitor hundreds of user actions, anything from a user logging in to initiating a transaction. Alipay has trained its software to identify a suspicious action and funnel it through its risk model, which can return a decision on the action almost instantly. Anything the model perceives as low risk is safe enough to proceed, but actions deemed risky require further scrutiny, including possible manual review.

Experimentation to Support Learning

Another component of Ant Financial's operating model is a sophisticated experimentation platform that runs hundreds of experiments daily, enabling the company to learn and understand the opportunities and risks provided by new features and products.

Ultimately, Ant Financial's dramatic expansion came about as a direct result of focusing on the various data sources that could be amalgamated on the existing platform and rapidly recombined by agile teams driving new products and services. Ant Financial's increases in scale and scope were driven by its impressive learning capabilities, combining analytics with agile innovation.

The data and algorithms that Ant Financial deploys in its business are also useful for additional new financial services developed by agile teams. Ant relies on scenario-based prototyping (use cases) to develop new applications (solutions) or opportunities, resting and refining them while attracting a critical mass of consumers and thereby mainstreaming the technology quickly. It also leverages innovations in data mining and semantic analysis to automate customer issue resolution.

Removing the Human Bottleneck

As the Ant Financial example illustrates, the essence of the digital operating model is avoiding direct human intervention on the critical path of the product- or service-delivery process. While employees help define strategies, design user interfaces, develop algorithms, code software, and interpret data (among many other functions), the actual processes that drive customer value are fully digitized. No human organization is a bottleneck in the qualification for individual loans or the recommendation of a specific investment vehicle.

How is this done? The firm anchors these processes in a central repository of data, describing customer and operational needs in an integrated fashion. As the customer interacts with the business process, software modules gather the necessary data, extract and analyze needs, internalize their implications, and interact with the customer to deliver the value as promised. Building customer interaction processes on a centralized data architecture thus operationalizes and automates the idea of customer centricity in a clear, actionable, and scalable way.

Many new operating models, like Ant Financial's, automate data-driven actions and gradually remove human tasks from delivery bottlenecks. Take, for example, shopping on the Amazon mobile app.

As the user browses through the app, offerings are being automatically selected based on data on the user's previous behavior and on the behavior of similar users. Pricing information is processed in real time (or close to it) and merged with the behavioral information to dynamically construct the page the user interacts with. A product manager eventually views aggregated data on transactions and consumer behavior, but almost every human interaction is removed from the actual critical path in service delivery. The only exceptions might be a worker helping pick the item from a largely automated warehouse, and the delivery person leaving the package at your door.

Removing human and organizational bottlenecks from the critical path has a huge impact on the nature of the company's operating model. The marginal cost of serving an additional user on many digital networks is, for all purposes, zero, apart from the small incremental cost of computing capacity, which is easily available from cloud service providers. This inherently makes a digital operating model easier to scale. Growth constraints are much less dependent on human actors, and organizational constraints are rarely a problem, because much of the operational complexity is solved through software and analytics or outsourced to external partners in the operating network.

A digital operating model also fundamentally changes the architecture of the firm. Beyond removing human bottlenecks, digital technologies are intrinsically modular and can easily enable business connections. When fully digitized, a process can easily be plugged in to an external network of partners and providers, or even into external communities of individuals, to provide additional, complementary value. Digitized processes are thus intrinsically multisided. After value is delivered in one domain (e.g., accumulating data about a set of consumers), that same process can be connected to drive value in other applications, thereby increasing firm scope and adding a multiplicative factor to the value it's delivering to the customer.

Finally, digitizing the operating model can also enable much faster learning and innovation. The vast amounts of accumulated data provide critical input to an increasingly broad range of tasks, from instant app personalization to feature innovation and product development. In addition, by digitizing many of the operational

workflows, this model diminishes the overall size of the organization along with the surrounding bureaucracy. The insights provided by analyzing a rich foundation of data can thus be rapidly deployed into actions by a relatively small number of agile product teams.

Ultimately, in a digital operating model, the employees do not deliver the product or service; instead, they design and oversee a software-automated, algorithm-driven digital "organization" that actually delivers the goods. This completely changes the factors involved in management, transforms the growth process, and removes traditional operating bottlenecks constraining scale, scope, and learning in a firm.

Let's look at two more examples.

The Irresistible Digital Bicycle

We see ourselves more akin to an Apple, a Tesla, or a Nest or a GoPro—where it's a consumer product that has a foundation of sexy hardware technology and sexy software technology.

—John Foley, founder and CEO, Peloton

John Foley was reportedly turned down by more than four hundred investors as he was starting his next-generation fitness company, Peloton. Investors could not be convinced that a traditional product like the stationary bicycle, invented more than two hundred years ago, had a digital future. However, Foley had different ideas borne of his experience competing with Amazon as CEO at Barnes & Noble. "The top line when I got there was $500 million. I could have doubled it, and we would still have been losing $100 million," he told *Barron's* in 2014. "As a business guy, I didn't like the value proposition of that."[12] Foley realized that instead of wasting his time chasing another competitor with superior scale, scope, and AI capabilities, he needed to find a traditional category and transform it digitally.

The idea for Peloton grew out of Foley's frustration that he could not get into his favorite indoor spin classes. The studio capacities were so limited that all the choice instructors' classes were booked as soon as they were scheduled. Taking a page from Amazon and Netflix, he envisioned a new fitness company that would take away constraints of

time, space, and capacity.

Founded in 2012, Peloton's main product is a sleek, high-quality indoor bicycle with an integrated 21-inch tablet to display fitness programming. Customers pay about $2,200 for the bike and then an additional $39 monthly subscription for unlimited access to fitness programming. They can choose from more than fourteen hours of daily live studio classes (from New York and London) and an ever-expanding library of more than fifteen thousand previously recorded workouts to access on demand.

Peloton's business model, built on a digital operating model, has turned the fitness industry on its head. People tend to get their exercise either in gyms (how many of us at the start of a new year have bought an annual membership?) or at home (how many of us have treadmills that have become bulky and expensive clothes hangers?). For gyms, the business model consists of making capital investments and charging customers for use through a subscription model (counting on the fact that most won't step into the place after January) and some type of pay-per-use for classes. Home fitness equipment makers sell us the equipment, so we make personal investments and hope to find motivation in working out every day. In contrast, the Peloton business model takes a traditional "analog" product and then transforms it by adding digital content, data, analytics, and connectivity to collide with a traditional industry.

Peloton's initial value creation is straightforward. Customers want the benefit and convenience of an in-home fitness experience without sacrificing access to great instructors and the community of fellow sweat hounds. Peloton brings the fitness studio to the customer's home. Value creation is enhanced by giving users access to an unlimited number of classes, including cycling, treadmill, yoga, meditation, strength training, and even outdoor walking and running workouts. Its more than one million members can binge on workouts the way Netflix subscribers can binge on shows.

Additional value creation mechanisms are the connectivity and community of Peloton members. More than 170,000 members connect through the official Peloton Facebook page, and then there are hundreds of subcommunities that have formed around Peloton instructors (who are celebrities in the Peloton world). There are

countless other tribes who've coalesced around different goals, geographies, and training styles. Taking a live-streamed class is also a communal experience: members can track their performance on a live leader board, virtually high-five each other, connect with each other, and follow each other's workout progress. Instructors name-check live users, calling out their achievements and milestones and reminding them to keep their form and motivation high through the tough parts. The on-demand classes even provide connectivity with riders who might happen to be taking the class at that moment. Peloton has activated voice and video connections among exercisers to bring the fitness class experience to their home. The community also meets face-to-face through regular "home rider invasions," when Peloton members travel from across the United States, Canada, and the United Kingdom to visit the company's Manhattan studios for live classes.

The value capture model for Peloton combines product sales and subscriptions. The bike is relatively useless without a subscription, and the Peloton service has a million subscribers, with a remarkable subscription renewal rate of 95 percent. Peloton fans who don't want to buy the bike can subscribe to the company's digital content and community via the mobile app for $20 per month.

Scaling the fitness experience is at the core of Peloton's operating model. While a typical spin class at SoulCycle might have thirty or forty riders in a studio, a live-streamed Peloton cycling class may have between five hundred and twenty thousand riders sweating simultaneously. After the live class ends, it becomes part of the online library available freely to members. Peloton's leaders also realized that its members needed additional fitness options, so it expanded its scope by offering a range of yoga, strength training, and treadmill sessions (for members who've purchased a sleek Peloton-branded treadmill, of course).

Peloton is in many ways still a product-focused company, but Foley's idea was to design the iPhone of fitness equipment. Peloton built its first bike in 2013, and, in 2014, after a round of investment, it produced an improved bike that could be tested by and sold to consumers. By 2015, the bike had been perfected, and business started to take off.

The company raised around $100 million, enabling it to work closely with its manufacturer in Taiwan to increase capacity, speed up bike

production and delivery, expand its software and analytics team, and dramatically increase the content delivered. The company also built its own supply chain, delivering bikes in Peloton-branded vans and dispatching employees to set up the bikes and advise customers on finding the classes and instructors to suit their tastes.

Although Peloton's success is sparked by a great product, the organization is structured more like a software company. It employs a team of more than seventy software engineers who design the company's systems for a version of Android. Peloton relies on human talent to devise, design, and produce its products and services—everything from the new treadmill to the latest "Power Zone" class. But even though humans are crucial, it is the digital service that delivers the experience in a highly scalable fashion to a rapidly increasing audience of enthusiasts.

There is no limit to the number of consumers who can subscribe to use the Peloton service (as long as its Taiwanese suppliers keep delivering the fitness equipment). As with Ant Financial, growth bottlenecks at Peloton are shifted to internal digitized systems or to resources outside the firm. Peloton, like Ant Financial, is not subject to the most significant traditional operational constraints on growth. In addition, the digital interfaces (the APIs) in Peloton's software easily expand the scope of the business by connecting to a variety of complementary apps (e.g., Apple Health, Strava, and Fitbit), social networks (Facebook and Twitter), and devices (heart rate monitors, smart watches).

Although its AI capabilities are nowhere near the level of Ant Financial's, Peloton has built a sophisticated analytics platform and digitally streamed content to transform fitness training into a new experience. The company gathers extensive data, from rider heart rate to workout frequency to musical taste, from in-studio attendance to social network engagement. It constantly analyzes the data and uses the analytics to implement a variety of improvements, from class selection and design to new product and service optimization. The analytics drive the user experience and greatly enhance engagement while increasing barriers to switching and reducing customer churn.

Unlike other exercise equipment products, loyalty to Peloton is extreme. It's easy to imagine what the company could do with its data

and the type of scope expansion that is possible. For example, Peloton could connect its users to nutrition services, health-care providers, or even insurance products. The company's data stores provide it with a broad range of options to redefine what it means to be a fitness company.

Peloton has enjoyed impressive growth. Its reliance on software, data, and networks has enabled the company to scale fast, reaching more than $700 million in revenue and a $4 billion valuation on an approximately $1 billion investment.

The World's Toughest AI Business

Human beings can do everything that AI can do. They just can't do it to scale.

—Anne Marie Neatham, COO, Ocado Technology

Online grocery delivery must be one of the most challenging businesses ever devised. Imagine promising a million people on-time delivery of more than fifty thousand of the world's lowest-margin and most perishable items through sun, rain, sleet, snow, and the Olympic Games. It is no wonder that it took many years for Ocado to win the respect of financial analysts. After going public in 2010, Ocado was roundly criticized for its business model, its operating model, and even its name ("Ocado begins with an 'o', ends with an 'o' and is worth zero," said Philip Dorgan, an analyst with RFC Ambrian Limited).[13] But in recent years the UK-based company has greatly exceeded expectations and become a darling of the financial markets.

Behind Ocado's success is a surge in AI impact on both its business and its operating model. Ocado delivers groceries, both for its own branded online and mobile service and for a variety of third parties. To do so on time, reliably, and efficiently, it has built a phenomenal foundation of data, AI, and robotics. Ocado is an AI company disguised as a supply chain company disguised as an online grocer. Its capabilities were built by necessity, over time, with painstaking conviction and deep investment.

Originally set up for browser-based commerce, Ocado introduced its first mobile app in 2009. The key to the business is Ocado's centralized

data platform, rebuilt from scratch in 2014, containing unrivaled detail on its products, customers, partners, supply chain, and delivery environment. The data is accumulated in the cloud and is exposed through easy-to-use interfaces for use by agile teams deployed to optimize every kind of application, from delivery routing to robotics, and from fraud detection to spoilage prediction. All this has combined to build a rapidly growing and profitable operation with a record of 98.5 percent on-time delivery.

AI algorithms are in the driver's seat of Ocado's operational execution. Running thousands of routing calculations per second, AI makes sure the company has a highly predictable delivery model, optimized across its fleet of thousands of trucks, delivering in all weather and traffic conditions across the entire United Kingdom. The algorithms optimize truck routing in real time and make sure the products delivered are fresh.

In addition to the routing, the AI actually predicts when customers are likely to order the products in the first place, typically a couple of days ahead of the need for them. Using unusually deep customer preference data, cross-referenced with the constraints of organic farmers in Ocado's supply chain, the algorithms predict when the refrigerated trucks should arrive at Ocado's supplier network of farms to pick up meats, poultry, and produce and bring them to storage in warehouses. And the warehouses are in themselves a masterpiece of AI technology, with thousands of bots that pick, assemble, and transport the groceries; the bots are coordinated and managed by algorithms, which in turn prioritize the most crucial and timely deliveries while minimizing congestion and optimizing overall efficiency.

The warehouses (also referred to as fulfillment centers) are the jewels in Ocado's operating model. A single warehouse can be the size of eleven soccer fields and sport thirty-five miles of conveyors that move hundreds of thousands of grocery boxes every day, around ten thousand simultaneously. Algorithms route every box to avoid traffic jams and ensure freshness and delivery capacity. Other algorithms aggregate and model the entire warehouse system.

The system is flexible and can accommodate an increasing number of locations, customers, and bots as capacity expands with growth, and

as Ocado's technology and operations teams continue to learn, experiment, and innovate, leading to rapidly increasing scale and scope. As COO Anne Marie Neatham notes, "Machine learning never stops. But you'll notice the common theme for the team. Visualize, trial it, iterate, iterate, iterate, iterate in volume."[14]

Over time, Ocado's AI and bot technology has collided with a range of traditional operating processes. Human labor is still used, even in the highly automated warehouses, to perform a number of tasks that bots have a hard time emulating, most notably picking certain difficult grocery items. But as you saw before, the labor is being moved off the critical path, as much as possible, to improve the scalability and reliability of the process. As Paul Clarke, Ocado's chief technology officer, put it, "For us, it's just the same journey we've been on since day one: to look for the next thing to automate, whether that's putting plastic bags in crates, or moving goods around our sheds. We start with the obvious thing and move on to automate the next thing and the next thing. You never get to the end."[15]

Ocado's deep AI and digital capabilities are enabling two different business models. Leveraging the capabilities built on its own UK-based online retail business, Ocado is also offering its technology platform to power third-party retail and delivery services; Marks & Spencer, the venerable UK retailer, is an example. Ocado is also expanding across the ocean, working, for example, with Sobeys in Canada and Kroger in the United States to set up and operate warehouses and customer fulfillment centers.

As part of the partnership, Kroger has increased its stake in Ocado to more than 6 percent and will leverage the Ocado Smart Platform's capabilities in online ordering, omnichannel integration, automated fulfillment, and home delivery. With almost $2 billion in revenue and a valuation around $7 billion, Ocado has come to the United States, and Amazon is watching closely.

Transforming Value Creation, Capture, and Delivery

Ant Financial, Ocado, and Peloton showcase three approaches to

digitizing value delivery, enabling business model innovation, and driving industry transformation. In each case, we witnessed the creation of exceptional consumer value, with scale, scope, and level of innovation that is virtually unprecedented in each industry. The value capture similarities are also striking. In each case, the companies are less transactional and more invested in using digital technology to foster consumer loyalty and engagement. And as long as consumers are deeply engaged with a service, more users will join, and the monetization opportunities will multiply.

The differences among the three firms are also interesting. The three industries they originally targeted could not be more different: financial services, groceries, and fitness. While Ant Financial is exclusively a set of information-based services, Ocado delivers products with a remarkably efficient supply chain, and Peloton provides a tightly integrated product-service combination. Still, in each case, the company digitized critical operating processes, with transformative impact.

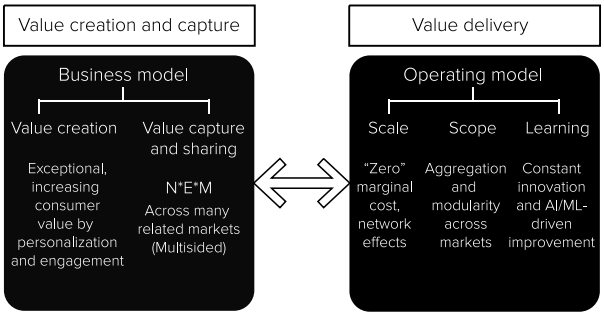
As we look closer, each company used algorithms and networks to transform its markets, but each did so in a unique way, building unique capabilities and employing unique approaches. Ant Financial built impressive capabilities in analytics and AI, and oversees a highly automated system to drive virtually unprecedented scale and scope across financial services and beyond. Ocado also features an operating model that deploys sophisticated AI, founded on an algorithmic core that drives impressive scalability, sustains an increasingly broad scope of product offerings, and enables ongoing learning and innovation. Ocado also emphasizes how its algorithms integrate with the human talent to, for example, aid drivers and product pickers. Peloton, instead, is driven more by networks and community, but it still uses data and analytics to enhance engagement and loyalty. The company takes content created by human talent and greatly amplifies its impact to a broad and expanding community of customers, who stream the service, exercise, and check progress via its increasingly sophisticated analytics. As with Ocado, human skills and labor shift into design, production, and enhancement roles, while the digital technology delivers and sustains the core experience.

Most of all, we are excited about the similarities in the operating

models of these different firms. By digitizing the most critical processes, each operating model removes traditional bottlenecks and enables unprecedented scalability, scope, and learning; once the model is established, most of what these firms need for growth is additional computing power, which is easy enough to access from the cloud. Growth bottlenecks are moved to the technology layer, or to the ecosystem of partners and suppliers. Figure 2-2 illustrates the kind of digital business and operating models at the heart of these three companies.

FIGURE 2-2

Value creation and capture versus value delivery



*Note: N*E*M = (the number of users) * (user engagement) * (monetization)*

Putting AI at the Core

On May 17, 2017, Sundar Pichai, Google's CEO, made a surprise announcement at the Google I/O conference in front of seven thousand attendees, with more than one million people viewing on live streaming. Google's strategic focus, Pichai said, was shifting from mobile to "AI first."[16]

The announcement surprised quite a few people. From its beginning, the company's business and operating models had always been driven by data, networks, and software. After all, Google commercialized the world's best search algorithm, developed leading advertising technology, and turned Android into the world's most popular software platform. The company had already invested heavily in AI, eclipsing most other firms and universities in the number of publications and patents. What did it mean for Google to be AI first?

Pichai wasn't talking about introducing a new AI-inspired product or launching a few pilots experimenting with advanced analytics. Rather, his announcement was the real deal, capping two decades of investment in developing software algorithms and AI technologies. It showed that AI had moved to the center of the company, to the core of its operating model. Increasingly, AI would be the common foundation across virtually every operating process. Pichai illustrated the approach with a variety of examples, from novel customer-facing apps (such as the innovative AI-enabled Google Assistant) to the new AI-enabled infrastructure powering Google's data centers and cloud services.

The announcement was a signal to Google consumers, advertisers, external developers, and employees that AI and its associated investments in data and analytics had become essential to the company's business and operating models. Virtually every aspect of Google was going to leverage this core. All of Google's products and services (several with billions of active users) would increase the value they delivered through conversational (speech, text), ambient (in all types of devices), and contextual (understand what you want) AI, and each process would continuously learn and adapt. The embedded AI systems would always be trying to predict what its consumers wanted or needed, updating these models across all interactions. This predictive power would of course be hugely valuable to Google's advertisers as well. An AI-first approach meant that Google's ads would become increasingly personalized and contextualized, ultimately increasing relevance and yielding more clicks.

The Pichai announcement provided a clear message and wake-up call. For Google's employees, technical as well as business focused, this was a signal to develop an in-depth understanding of AI and drive its application across every aspect of the company's value creation, capture, and operating model. For Google's massive ecosystem of partners and developers, it was an invitation to embed AI to improve their own products and services, from exercise apps to TVs. For the rest of us who were listening, it became clear that AI had finally come of age. For literally millions of people, AI was no longer a promising set of innovative technologies; it was becoming the core of the firm.

In the next chapter, we examine how the core of the firm, like

Google's, is a scalable decision factory, powered by software, data, and algorithms.

3

The AI Factory

Through much of history, products were painstakingly and individually crafted in artisanal workshops. That ended when the Industrial Revolution transformed the economy by spawning a scalable and repeatable approach to manufacturing. Engineers and managers became experts at understanding the processes needed for mass production and built the first generation of factories, dedicated to the continuous, low-cost production of quality goods. However, while production was industrialized, analysis and decision making remained largely traditional, idiosyncratic processes.

Now, the age of AI is manifested by companies driving another fundamental transformation. This one involves industrializing data gathering, analytics, and decision making to reinvent the core of the modern firm, in what we call the "AI factory."[1]

The AI factory is the scalable decision engine that powers the digital operating model of the twenty-first-century firm. Managerial decisions are increasingly embedded in software, which digitizes many processes that have traditionally been carried out by employees. No human auctioneer gets involved in the millions of daily search-ad auctions at Google or Baidu. Dispatchers do not decide which car is chosen on DiDi, Grab, Lyft, or Uber. Sports retailers do not set daily prices on golf apparel at Amazon. Bankers do not approve every loan at Ant Financial. Instead, these processes are digitized and enabled by an AI factory that treats decision making as an industrial process. Analytics systematically convert internal and external data into predictions, insights, and choices, which in turn guide or even automate a variety of operational actions. This is what enables the

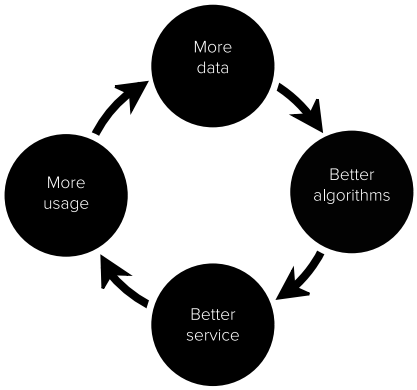
superior scale, scope, and learning capacity of the digital firm.

Digital operating models can take various forms. In some cases, they might only manage flows of information (think Ant Financial, Google, or Facebook). In other cases, operating models guide how the company builds, delivers, or operates actual physical products (think Ocado, Amazon, or Waymo). In either case, AI factories are at the core of the model, guiding the most critical processes and operating decisions, while humans are moved to the edge, off the critical path of value delivery.

In its essence, the AI factory creates a virtuous cycle between user engagement, data collection, algorithm design, prediction, and improvement (see figure 3-1). It integrates data generated from multiple sources (internal or external to the firm) to refine and train a set of algorithms. These algorithms not only make predictions but also use the data to improve their own accuracy. The predictions then drive decisions and actions, either by informing human insights or by enabling an automated response. Hypotheses about changing customer behavior patterns, competitive responses, and process variations are tested through rigorous experimentation protocols that enable causal identification of changes that might improve the system. Data about usage and about the accuracy and impact of the prediction outcomes is then sent back into the system for further learning and predictions. And the cycle continually repeats.

FIGURE 3-1

The AI factory's virtuous cycle



Take, for example, a search engine like Google or Bing. As soon as a

user types a few letters in the search box, algorithms dynamically predict the full search term based on prior search terms and the user's past actions. These predictions are captured in a drop-down menu (the *autosuggest box*), which helps users zero in quickly on the desired search. Every user movement and every click are captured as data points, and every data point gathered improves the prediction for future searches. The more searches, the better the predictions, and the better the predictions, the more the search engine is used.

There are multiple other prediction cycles in a search engine's AI factory. During the natural search process, the search term entered by a user generates a display of organic search results, which are drawn from a previously assembled index of the web and optimized by using the outcomes (the clicks generated) of previous searches. In addition, entering the search term also starts an automated auction for the most relevant ads to match the user's intent, an auction whose results are also shaped by additional learning loops. The search-results page, which combines organic search results and relevant ads, is thus heavily influenced by data on previous search attempts. Any click on or away from the search query or search-results page provides useful data.

In addition, a product manager within the search engine operations might have some new hypothesis—for example, that showing fewer ads might improve revenues on a given page, or that highlighting search results would improve click-through rates. To provide additional fodder for improvement, these hypotheses would be loaded on the experimental machinery and tested on a statistically relevant sample of users.

Clearly there is no way all this data could be analyzed by a few analysts using manual tools, or even by casually assembled code. The AI factory solves this problem by bringing mass production methods to data processing and analytics, thus forming the core of a digital operating model. Let's dig deeper into its nature, using Netflix to anchor the discussion.

Building and Running the AI Factory

Netflix has transformed the media landscape by harnessing the power

of artificial intelligence. The core of Netflix is its AI-centric operating model: it is powered by software infrastructure that gathers data and trains and executes algorithms that influence virtually every aspect of the business, from personalizing the user experience to picking movie concepts to negotiating content agreements.

In its earliest days two decades ago, Netflix displayed movie reviews, generated recommendations based on customers' viewing histories, and shipped DVDs of new releases the day they were made available in stores. Even then, Netflix recognized the importance of using data to improve the customer experience. The company's early efforts were focused on developing a recommendation engine, which suggested movies based on a viewer's history, movie ratings, and the preferences of similar viewers.[2] Netflix not only used this data internally but also shared the reviews with movie studios. Sharing this data helped Netflix negotiate better financial terms in its partnerships with Warner Home Video and Columbia TriStar.[3]

Netflix grew rapidly, hitting eight million subscribers in 2007 when it launched its streaming service. This new offering dramatically increased the company's access to user data, which Netflix analytics teams used extensively. With its mail delivery service, Netflix could track only those titles users requested, the length of time they kept a DVD, and their rating of each title; Netflix could not monitor actual viewing behavior. With streaming, Netflix could track the full user experience—when viewers pause, rewind, or skip during a show, for example, or what device they are using. This behavioral data helped Netflix determine which movie thumbnail image to show a viewer (yes, even these are personalized based on preferences for particular genres, actors, and other such factors), predicting their likely preferences. Through more-advanced analytics, Netflix also predicted drivers of customer loyalty. With the goal of increasing subscriber viewing time and decreasing customer churn rates, Netflix used AI to launch a function that automatically queues the next episode in a series or recommends similar movies. The customization and personalization has become pervasive. As Joris Evers, then chief of communications at Netflix, told the *New York Times* in 2013, "[T]here are 33 million different versions of Netflix," meaning that each user's Netflix experience is personalized and customized.[4]

Netflix also uses data and AI algorithms to decide which content to create on its own. The company's first use of predictive analytics for this purpose was in 2013 to evaluate the potential of *House of Cards*, the fictional account of a senator's rise to the White House, in collaboration with Media Rights Capital (MRC). Cindy Holland, vice president of original content, noted in an interview, "We have projection models that help us understand, for a given idea or area, how large we think an audience size might be, given certain attributes about it. We have a construct for genres that basically gives us areas where we have a bunch of programs and others that are areas of opportunity."[5]

By 2010 Netflix was embracing the AI factory approach to systematically apply analytics and AI to the company's recommendation engine. In 2014, the company expanded the factory to improve the streaming experience by understanding user behavior, creating a personalized streaming experience for each user (based on such factors as connection speed and preferred device), and determining what movies and shows to cache on "edge servers," which are deployed closer to viewers.[6] Now Netflix has about 150 million subscribers in more than 190 countries, has amassed a content library of more than 5,500 shows, and consumes 15 percent of the global internet bandwidth.

Experience from Netflix and other leading firms underlines the importance of a few essential AI factory components (see figure 3-2):

1. ***Data pipeline:*** This process gathers, inputs, cleans, integrates, processes, and safeguards data in a systematic, sustainable, and scalable way.

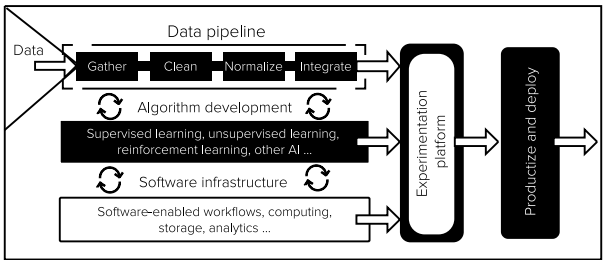
2. ***Algorithm development:*** The algorithms generate predictions about future states or actions of the business. These algorithms and predictions are the beating heart of the digital firm, driving its most critical operating activities.

3. ***Experimentation platform:*** This is the mechanism through which hypotheses regarding new prediction and decision algorithms are tested to ensure that changes suggested are having the intended (causal) effect.

4. ***Software infrastructure:*** These systems embed the pipeline in a consistent and componentized software and computing infrastructure, and connect it as needed and appropriate to internal and external users.

FIGURE 3-2

AI factory components



If the data is the fuel that powers the AI factory, then infrastructure makes up the pipes that deliver the fuel, and the algorithms are the machines that do the work. The experimentation platform, in turn, controls the valves that connect new fuel, pipes, and machines to existing operational systems.

Let's look first at the data pipeline.

The Data Pipeline

Data is the essential input of the AI factory. One reason for the radical advances made by AI systems in recent years is that the velocity, volume, and variety of data available for analysis has exploded. As far back as 2012, Netflix was using a broad base of data inputs. As described by Xavier Amatriain and Justin Basilico, two Netflix engineers, on the Netflix blog, the inputs are varied.

- *We have several billion item **ratings** from members. And we receive millions of new ratings a day.*
- *We already mentioned item **popularity** as a baseline. But, there are many ways to compute popularity. We can compute it over various time ranges, for instance hourly, daily, or weekly. Or, we can group members by region or other similarity metrics and compute*

popularity within that group.

- *We receive several million stream **plays** each day, which include context such as duration, time of day and device type.*
- *Our members add millions of items to their **queues** each day.*
- *Each item in our catalog has rich **metadata**: actors, director, genre, parental rating, and reviews.*
- ***Presentations**: We know what items we have recommended and where we have shown them, and can look at how that decision has affected the member's actions. We can also observe the member's interactions with the recommendations: scrolls, mouse-overs, clicks, or the time spent on a given page.*
- ***Social** data has become our latest source of personalization features; we can process what connected friends have watched or rated.*
- *Our members directly enter millions of **search terms** in the Netflix service each day.*
- *All the data we have mentioned above comes from internal sources. We can also tap into **external data** to improve our features. For example, we can add external item data features such as box office performance or critic reviews.*
- *Of course, that is not all: there are many **other** features such as demographics, location, language, or temporal data that can be used in our predictive models.*[7]

In 2018, Netflix users had more than 5,600 movie and TV series titles to choose from. Every time users open the Netflix application on their TV, computer, phone, or tablet, the company's systems kick in to make personal recommendations and customize the interface. Virtually every aspect of a user's experience generates data, which then enables Netflix to further fine-tune the customizations it provides. (And certainly, there is much more data available now than when this post was written in 2012.) All of this data is cleaned, integrated, prepared, and used by Netflix to dynamically adapt its service to continuously improve the value it provides to its estimated 300 million users.

The depth and breadth of the Netflix data is the envy of the industry. Part of the company's data and analytics assets includes creating approximately two thousand *microclusters*, or taste communities, which connect viewers having similar tastes. Individual users can fit in to several taste communities, and they defy simple demographic profiles; a sixty-five-year-old grandmother in urban Mumbai may like the same shows as a teenager in rural Arkansas.

Netflix has "datafied" TV entertainment—a term coined by Ming Zeng, Alibaba's strategy chief and academic counsel. The idea of *datafication* refers to systematically extracting data from activities and transactions that are naturally ongoing in any business.[8] The Nest thermostat, for example, invaded a sleepy market by datafying a traditional spectrum of activities—controlling the heating, ventilation, and cooling (HVAC) systems in a home. The addition of a few electronic sensors to monitor temperature and motion in the home, along with computer-based control and Wi-Fi connectivity, enabled Nest to create a brand-new data layer that generates important new value for homeowners. The Nest device, in only a few days, can learn your habits and adjust the temperature automatically in your house, participate in energy reduction programs at your nearby utility, and enable smartphone control.

Similar datafication has happened in almost every setting, from social behavior on Facebook to fitness with an Apple Watch or Fitbit, to sleep and health tracking with the Oura and Motiv rings.[9] Increasingly, as in the Netflix example, the initial process of datafication can be combined with external data sources to provide additional value to the user. The Oura ring's app, for example, combines sleep and heart rate data with the user's activity level monitored by an Apple Watch to coach the user on the level of rest and activity needed for a productive day. Ride-sharing platforms like Uber, Lyft, Grab, DiDi, and GOJEK have built a datafication layer around transportation. The combination of their applications and the smartphone infrastructure has enabled these companies to generate data at an unprecedented level about individual transportation preferences, demand and supply needs, and overall flow of traffic in and out of urban centers. Accurate, real-time data about all this has never existed until now.

Sometimes, innovation is needed to transform traditional activities into sources of useful data. Alipay and WeChat Pay have led the way in economic transactions through their extensive use of QR codes for payments. If data is not readily available or does not exist, it may be worthwhile for a company to invest in technology and services that generate the data in the first place. Even Pitney Bowes, the hundred-year-old provider of postal equipment, has built a datafication strategy around physical addresses in the United States and is augmenting the company's business model by offering data-driven Knowledge Fabric solutions to banks, insurers, social platforms, and retailers—any organization that can use address data for marketing, fraud detection, and other purposes. The company realized that it could create and capture value beyond selling postage.

Many incumbent businesses that are attempting to build AI factories find that the data they possess is fragmented, incomplete, and often siloed within divisions and disparate IT systems. Take, for example, a typical hotel stay for a business traveler. In theory, a hotel chain should have a treasure trove of data on their customers, from home address to credit card information, to frequency of travel, airline, and mode of transportation, location of travel, class of stay, meal selections, local sightseeing favorites, and health and fitness preferences. In practice, though, the data is highly fragmented, resides in various system silos with incompatible data structures, is missing common identifiers, and may not necessarily be very accurate. Executives at many incumbent companies consistently underestimate the challenge and the urgency of the investment they face in cleaning and integrating their data across the enterprise so that they can build an effective AI factory. The first order of business facing these executives is to ensure that the appropriate investments are in place.

We emphasize that after the data is gathered, much work remains to be done in cleaning, normalizing, and integrating it. These steps are quite challenging. Data assets are most often plagued by all kinds of biases and even plain errors, and a significant investment needs to be made in ensuring that the data is checked carefully for inaccuracies and inconsistencies. Moreover, as various streams of data are integrated into a single stream to feed complex analysis, the different kinds of data must be normalized. A particular challenge is making

sure that financial data is being used properly, in a way that is consistent with operational data, so that any insight that comes from analyzing the integrated dataset is accurate. For example, units should be consistent, redundancies eliminated, and variables compatible. These things often sound simple but are not, especially as the datasets reach significant size.

Algorithm Development

After the data is gathered and prepared, the tool that makes the data useful is the *algorithm*—the set of rules a machine follows to use data to make a decision, generate a prediction, or solve a particular problem.

Consider how you would analyze whether a customer is likely to leave a service like Netflix. Here the algorithm would predict customer churn as a function of variables such as usage (frequency and intensity), satisfaction, demographics, and relationships or similarities with other users. The prediction algorithm would be tuned and calibrated with data on past customers, tested for accuracy with past data or with a controlled experiment, and deployed either as an analytical tool for managers or as a step in an operational process—for example, automatically enabling a special offer to retain vulnerable customers.

Ajay Agrawal, Josh Gans, and Avi Goldfarb of the University of Toronto note that data proliferation and advances in AI algorithms have lowered the cost of making accurate predictions, increasing the scope and intensity of the usage of prediction algorithms throughout the economy.[10] Algorithms predict which Google photos include family members or friends, what Facebook content you should read next, how much revenue to expect from giving a Walmart discount to a particular customer, or when a piece of equipment at a Ford manufacturing facility will need maintenance. These kinds of predictions are vital to the success of many organizations, and the algorithms deployed should be geared to provide consistent and robust predictions.

AI algorithms can be used for a broad variety of applications, from generating relatively simple predictions (like a sales forecast) to suggesting stocks to pick for high-frequency trading, to complex image

recognition and language translation tasks that may exceed human capabilities. Some of the most complex applications, such as driving a car, use a variety of different algorithms simultaneously—for example, to identify and track cars and to route a car through heavy traffic.

Although the use of applications has exploded over the past decade, the foundations of algorithm design have been around for quite some time.[11] The conceptual and mathematical development of classic statistical models such as linear regression, clustering, or Markov chains date back more than a hundred years. Although neural networks are now generating a lot of excitement, they were initially developed in the 1960s and are only now being put to use at scale with production-ready outputs. The vast majority of production-ready and operational AI systems use one of three general approaches to develop accurate predictions using statistical models, also known as machine learning. These are supervised learning, unsupervised learning, and reinforcement learning.

Supervised Learning

The basic goal of *supervised* machine learning algorithms is to come as close as possible to a human expert (or an accepted source of truth) in predicting an outcome. The classic case is analyzing a picture and predicting whether the subject is a cat or a dog. In this case the expert would be any human being who could label photos as images of a cat or a dog. The algorithms in this class of machine learning systems rely on an *expert-labeled* dataset of the outcome (the Y) and the potential characteristics or features (the Xs). The operationalization of the algorithm is called a *model*, which takes the general-purpose statistical approach and creates a context-specific instantiation of the prediction problem that needs to be solved.

The first step in supervised learning is to create (or acquire) a labeled dataset. For example, we might acquire a file containing thousands of pictures of cats and thousands of pictures of dogs, with each picture labeled appropriately. The data is then split between training and validation. The *training* dataset is used to determine the parameters of the model that generates the prediction of the outcome (whether a given picture depicts a cat or a dog). After the model is

trained, the *validation* dataset is used to test the accuracy of the model. The model makes its predictions on the validation dataset; we can then compare these predictions to the expert predictions and thereby assess the quality of the model. Supervised machine learning algorithms can be used to predict either a binary outcome (for example, whether a picture shows a cat or a dog) or a numerical quantity (such as the sales forecast for a particular product).[12]

As we compare the algorithmic model's prediction of the outcome to the validated labeled outcomes, we can determine whether we are satisfied with the error rate between model prediction and expert. If we are not satisfied, we can choose a different statistical approach, get more data, or work on identifying other features that may be helpful in making a more accurate prediction. The main challenge here is to keep iterating between data, features, and algorithms until we are satisfied with the error rate between the model prediction and the expert prediction.

Examples of supervised machine learning abound. Every time we label an email as spam, we help our email provider's machine learning algorithms update its models to identify the latest clever scam. Facebook's or Baidu's ability to suggest names of friends who may appear in newly uploaded pictures is based on our prior labeling of photos. Credit card companies or payment platforms decide whether to allow a transaction based on prior purchasing habits, which automatically create labeled data. A Nest thermostat's ability to change the temperature in your living room thirty minutes before you arrive home is based on autogenerated labeled data gathered from your previous arrival and departure times, as well as your prior temperature-setting habits.

Netflix uses supervised learning in a variety of scenarios. For recommendations, Netflix has used labeled datasets made up of actions and results (e.g., movies chosen and liked) by people who are deemed by the algorithm to be similar to a given user. A large dataset of user choices, calibrated by characteristics of the user and of the decision context, can lead to effective recommendations. This kind of *collaborative filtering algorithm* is used for all kinds of recommendations, including Amazon's shopping engine and Airbnb's matching engine.

Many companies may already have vast troves of algorithm-ready labeled data thanks to their investments in systems, technologies, databases, and heavyweight enterprise resource planning (ERP) installations. For example, most large insurance companies have decades of labeled data relating to property damage and could readily implement supervised machine learning models to reduce both fraud and the time it takes to process and resolve claims—especially if the company is equipped for direct photo uploads or drone-based inspection. Similarly, health-care systems are full of labeled datasets. For example, many companies are taking medical data (such as radiology, cardiology, pathology, and EKG results) and correlating it with health diagnoses. Israel-based Zebra Medical Vision now offers technology to help radiologists make better diagnoses from X-ray, CT, and MRI scans.

Unsupervised Learning

Unlike supervised learning models, which train a system to recognize known outcomes, the primary application of *unsupervised* learning algorithms is to discover insights in data with few preconceptions or assumptions. This is what Netflix does when it discovers related groups of customers in analyzed viewing data, when it creates customer segments for marketing campaigns, or when it creates different versions of the user interface that match different usage patterns. Or think of various national security agencies and law enforcement organizations accumulating huge amounts of social media data to look for abnormal patterns and discern potential security threats. In these cases, one does not know exactly what to look for but is searching for related groups or for events that fit or don't fit established patterns.

Unlike supervised learning algorithms, where the data inputs are labeled with a given outcome, unsupervised learning algorithms aim to find "natural" groupings in the data, without labels, and uncover structures that may not be obvious to the observer. Thus the job of the algorithm is to show patterns in data, with humans (or even other algorithms) labeling the patterns or groups and deciding on potential actions. In our example of photos of cats and dogs, an unsupervised

learning algorithm might find several types of groupings. Depending on how the clusters are structured, these groupings could end up separating cats and dogs, or indoor and outdoor photographs, or pictures taken during day or night, or virtually anything else. Again, an unsupervised learning algorithm does not suggest specific labels but rather establishes the most robust statistical groupings. Humans, or other algorithms, do the rest.

Unsupervised learning is useful for gaining insights from social media postings by, say, identifying customer groups and sentiment patterns that can be used to guide product development. Attitudinal and demographic survey responses by customers can be used to create customer segments. The reasons for customer churn could also be categorized through unsupervised learning. In manufacturing settings, one could group instances of machine failure or order delay.

There are three broad types of unsupervised learning. The first relates to algorithms that *cluster* data into groups. A fashion retailer may use this approach to understand how to segment its customers based on the types of products purchased, the pricing and profitability of the items, and the various channels that brought customers to the store. More-sophisticated retailers might have additional data such as social network-based graph data (whom customers are connected to) and their social media postings. All this data then can allow the company to uncover a unique set of segments, well beyond simple demographics.

Netflix microclusters—its taste communities of members with similar movie and series preferences—is a good illustration of the power of such a tool. Cluster analysis in the form of topic modeling is used extensively to find meaning in text-based data and uncover salient topics within and across texts. The technique has been used to analyze news reports, SEC filings, investor calls, customer call center transcripts, or even chat records.

The second broad category is known as *association rule mining*. A common example is the recommendations for additional products an online shopper might want to purchase based on the current set of products in the shopping cart. Amazon has made a science of association rule mining. The algorithms look for frequency and probability of co-occurrence among any set of items and then create

associations that are likely to occur between various types of products. Ocado, for example, learned from its data that there was a strong relationship between diapers and beer. New parents don't get to go out much, so recommending beer and wine to shoppers when they are purchasing diapers turned out to be profitable and also increased customer satisfaction.

The third type of unsupervised learning algorithm is *anomaly detection*. Here the algorithm simply looks at each new incoming observation or datum and makes the judgment whether or not it fits prior patterns. If it does not fit the pattern, then the algorithm flags that item as anomalous. This type of application is often used in fraud detection in financial services, health care for a variety of patient data, and maintenance of systems and machines.

Reinforcement Learning

Although they are still relatively underdeveloped, the potential applications of *reinforcement* learning may be even more impactful than those of supervised and unsupervised learning. Rather than start with data on an expert's view of the outcome, as in supervised learning, or with a pattern-and-anomaly recognition system, as in unsupervised learning, reinforcement learning requires only a starting point and a performance function. We start somewhere and probe the space around us, using as a guide whether we have improved or worsened our position. The key trade-off is whether to spend more time *exploring* the complex world around us or *exploiting* the model we have built so far to drive decisions and actions.

Let's say we take a cable car up a tall mountain and we want to find our way down. It's a foggy day, and the mountain does not have any clearly marked paths. Because we can't see the best way down, we have to walk around and explore different options. There is a natural trade-off between the time we spend walking around getting a feel for the mountain, and the time we spend actually walking down when we believe we have found the best path. This is the trade-off between exploration and exploitation. The more time we spend exploring, the more we will be convinced we have the best way down, but if we spend too long exploring, we will have less time to exploit the information

and actually walk down.

This is close to the way the Netflix algorithm personalizes movie recommendations and the visuals they are associated with.[13] The problem is a bit more complicated, because the Netflix team needs to figure out which movie selection to present and then which artwork to combine it with to maximize the match between user and recommendation. But in a way similar to our finding our way down the mountain, Netflix spends some time exploring options, and some time exploiting the solution offered by its models. To explore visual options, Netflix systematically randomizes the visuals shown to a user, thereby exploring new possibilities and refining the prediction model. Netflix then exploits the improved model to show the user a slew of recommendations with improved visuals.

The Netflix service continues to improve dynamically by automatically cycling between periods of exploration and exploitation, a process designed to learn the most about the preferences of a complex human being and maximize user engagement over the long term. The writer of the Netflix technology blog asked in a 2017 post, "Given the enormous diversity in taste and preferences, wouldn't it be better if we could find the best artwork for each of our members to highlight the aspects of a title that are specifically relevant to them?"[14]

The Netflix challenge is a fancy variant of a common class of models used in reinforcement learning. Known as the *multiarmed bandit problem*, it is named after imagining a gambler playing different slot machines ("one-armed bandits"), each machine characterized by a different (but unknown) reward distribution. The gambler can spend more time exploring which machine seems to give the best rewards or can focus on exploiting the one machine that seems to be the best bet so far. Any deviation from the optimal path (just cranking on the best machine) is expressed as the *regret* measure. Multiarmed bandit problems are useful in the allocation of finite resources across different processes, each associated with different reward distributions. The general idea is to maximize operating performance by minimizing regret.

Multiarmed bandit problems are vitally important to the deployment of AI in operating models. As we strive to optimize and

improve operating performance across processes, managing the trade-off between exploration and exploitation is fundamental. These algorithms are used extensively to manage a variety of operating workflows, from making product recommendations to setting product prices, and from planning clinical trials to selecting digital ads. They can even guide the behavior of actual agents in imagined or real worlds, from the path of Nintendo's Mario Kart video game to the bots in Ocado's warehouses. In essence, multiarmed bandits are set up to make real operating decisions while they optimize the trade-offs between short-term impact and long-term improvement.

Reinforcement learning has captured public attention thanks to a software system called AlphaGo. Created by Google's DeepMind AI research team, AlphaGo has started to beat master players around the world at the ancient Chinese strategy game Go. Although computers have beaten humans at chess (remember Deep Blue by IBM), Go was thought to be too complicated for any program to master it. However, starting in 2016, this started to change as top Go masters kept losing to AlphaGo. These results were stunning—so much so that Kai-Fu Lee, an eminent computer scientist and technology investor, noted in his book *AI Superpowers* that the Chinese government declared its own "Sputnik moment" and made achieving world-class leadership in AI a national priority, with tremendous resources dedicated to achieving this goal.

That was before AlphaGo Zero came on to the scene and started beating AlphaGo at its own game. AlphaGo Zero uses the reinforcement learning approach: unlike prior versions of AlphaGo, wherein data from hundreds of thousands of games was used as input, the AlphaGo Zero system was essentially given the rules of the game and then asked to figure out the best approaches (the "Zero" stands for no external data). Reinforcement learning works by having a software agent interact with the environment and take actions within it to maximize a predefined reward. By giving the rules of the game or environment to the agent, the software system can quickly learn to maximize rewards and achieve superior performance. Google's DeepMind team has applied the lessons from Go to drug discovery and protein folding and has found that its system performs considerably better than the best scientists and their approaches.

The Experimentation Platform

To be reliably impactful, the wealth of predictions generated by data and algorithms in an AI factory requires careful validation. Google runs more than one hundred thousand experiments each year to test a vast variety of potential data-driven improvements to its service. LinkedIn reportedly runs more than forty thousand experiments each year. The experimentation capacity required by digital operating models is such that traditional, ad hoc approaches to experimentation simply cannot handle the scale and impact of what is required. A state-of-the-art experimentation platform will provide the comprehensive set of technologies, tools, and methods required to do experimentation at scale.

To use an experimentation platform, potential significant changes to the business must first be formalized as a hypothesis. Each hypothesis is then typically tested as a *randomized control trial* (also known as an A/B test) in which a random sample of users is exposed to the change (known as a *treatment*) and a second random sample of users experience business as usual (the *control*). The outcomes are then compared, and if the difference between them is statistically significant the treatment is known to actually impact the outcome, instead of just being spuriously correlated. This approach ensures that any prediction being generated by algorithms actually has a *causal* effect on the outcome.

The experimentation platform is a necessary component of the AI factory. Imagine running our algorithm to predict customer churn and learning that churn correlates with a certain age group. We still do not know whether customers in that age group are more likely to churn in general, or whether they would respond positively to some kind of special offer and continue to use our service. Before offering an expensive rebate to millions of customers, it would make sense to try an A/B test on a small fraction of users and gather statistically significant evidence on what portion of customers would remain with our service *because of* that specific offer. The same kind of logic applies to a great variety of potential business improvements recommended by an AI factory at scale.

Netflix engineers and data scientists have built an extensive

experimentation platform that is fully integrated within its algorithm development and execution process.[15] Every significant product change at Netflix goes through A/B testing before it becomes a standard part of the product experience. The experimentation platform is also utilized to improve video streaming and content delivery network algorithms (the service supports hundreds of devices and a vast range of bandwidth conditions) as well as image selection, user interface changes, email campaigns, playback, and registration.

Indeed, the company tries to bring scientific rigor to all of its decision making by embracing experimentation as an integral component. The fully automated experimentation platform enables Netflix employees to run experiments at scale. The platform allows them to kick off the experiment, ensures there are no other blocking experiments or overlapping subject pools, recruits subjects from its audience, and creates reports to analyze and visualize results both during and after the experiments are completed.

Software, Connectivity, and Infrastructure

The data pipeline, the algorithm design and execution engine, and the experimentation platform should all be embedded in software infrastructure to drive the operating activities of the digital firm.

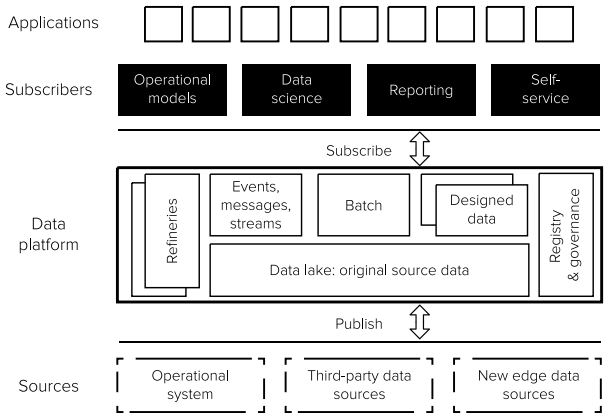
Figure 3-3 depicts an example of a state-of-the-art data platform powering an AI factory, with data flowing from bottom to top. The data platform provides a structure for software developers to build, deploy, and execute AI applications. The basic idea behind the pipeline is a *publish-subscribe* methodology for APIs (application programing interfaces). The purpose is to make clean, consistent data available to applications; think of it as something like a data supermarket.

After the data is aggregated, cleaned, refined, and processed, it is made available through consistent interfaces (the APIs), allowing applications to rapidly subscribe, sample what they need, test, and deploy. All this lets an agile development team build a new application in weeks, sometimes even days. Without these assets, a traditional IT custom-built process takes orders of magnitude more time and cost and becomes a nightmare to maintain and update. And in becoming an AI-driven company like Netflix, the idea is not to build one AI

application. Rather, the idea is to build thousands, of them—indeed enough to help make as many different types of predictions as possible.

FIGURE 3-3

A state-of-the-art data platform



Source: Keystone Strategy

Concurrent with investments in data and software are strategic investments in connectivity and infrastructure to integrate with the data platform. As we discuss in detail in the next chapter, most enterprises, even today, operate in separate silos. Even though customers view the enterprise as a unified entity, internally the systems and data across units and functions are typically fragmented, thereby preventing the aggregation of data, delaying insight generation, and making it impossible to leverage the power of analytics and AI.

Data platforms, and the organizations that work with them, should avoid siloed structures and instead should be designed in a modular fashion. The design of interfaces is critical in ensuring modularity in both code and organization. Clear interfaces therefore allow for decentralized innovation at the module level; as long as there is a standard for the sharing of data and functionality, each module can improve its core function independently. APIs compartmentalize the innovation problem and enable independent agile teams or individual developers to focus on specific tasks without destroying the consistency of the whole.

Building a consistent (and secure!) data platform is even more

important if the data is exposed to external partners. Taobao, Alibaba's online mall, is a good example, listing more than one billion items, all supplied by third-party providers. The only way for the company to satisfactorily share data with its internal and external users is through clear and secure APIs that enable the required range of functionality.

A typical internal Alibaba developer or external Taobao seller may be subscribing to more than one hundred different data platform software modules to enable them to upload inventory information, set pricing (manually or automatically), track consumer reviews, handle shipments, and the like. The development of well-designed APIs not only frees Taobao's engineers to keep developing and advancing internal systems to serve billions of users and millions of merchants but also unleashes creativity by an ecosystem of software vendors to offer a wealth of additional services.[16]

Finally, building a state-of-the-art AI factory with a well-designed data platform improves the organization's ability to focus on the crucial challenges of data governance and security. The massive amount of data that is increasingly captured from users, suppliers, partners, and employees is extremely valuable, sensitive, and private. It simply should not be stored in an ad hoc fashion. An organization needs to build a secure, centralized system for careful data security and governance, defining appropriate checks and balances on access and usage, inventorying the assets carefully, and providing all stakeholders with the necessary protection.

As part of the essential data governance challenge, carefully defining clear and secure APIs is essential to the AI factory. After all, APIs throttle the flow of data in and out of AI factory systems. Think of it as a way for the company to control all the data and functionality that it is willing to offer to internal and to external developers. As such, APIs control access to some of the most critical and private assets within the organization. They force the company to define, ahead of time, which of these critical assets it wants to make available within the enterprise and which it may be willing to offer to anyone outside the company. The data that can flow through an API can make or break a digital company. The Cambridge Analytica scandal happened because developer and manager errors apparently caused a

critical hole in the Facebook platform's graph API, allowing external application developers to access much more data than may have been originally intended by the company.

Ultimately, the data, software, and connectivity underlying an AI factory must reside within a secure, robust, and scalable computational infrastructure. Increasingly this infrastructure is on the cloud, is scalable on demand, and is built using standard off-the-shelf components and open source software. In addition, it needs to be seamlessly connected to the many individual processes and activities that constitute the company's operating model. Ultimately, these are the core digital processes that shape the delivery of value, such as creating, recommending, selecting, and delivering Netflix content, billing Netflix customers, or tracking the performance of Netflix content partners.

Building an AI Factory

You don't have to be Netflix to build an AI factory. The Laboratory of Innovation Science at Harvard (LISH), where we are faculty directors, in collaboration with colleagues from Harvard Medical School and the Dana-Farber Cancer Institute, demonstrated the development of an AI system that maps the shape of lung cancer tumors based on CT image scans. Deployed in only ten weeks and on an academic budget, the system is as good as a Harvard-trained radiation oncologist.

To develop the system, we leveraged the LISH AI factory, itself built to create a data pipeline and platform architecture for solving a variety of problems, usually with the help of crowdsourced algorithm design contests on Topcoder. LISH has partnered with leading organizations like NASA, Harvard Medical School hospitals, Broad Institute of Harvard and MIT, and Scripps Research to take some of their toughest computational and prediction challenges.

Outlining a lung cancer is critical in developing an effective therapy for patients. Oncologists therefore spend much time mapping the exact volumetric shape of any tumor that is to receive radiation therapy. Correctly outlining the tumor is particularly critical so that the therapy does not miss cancer cells or damage healthy tissue. The LISH team worked with Raymond Mak, from the Dana-Farber Cancer

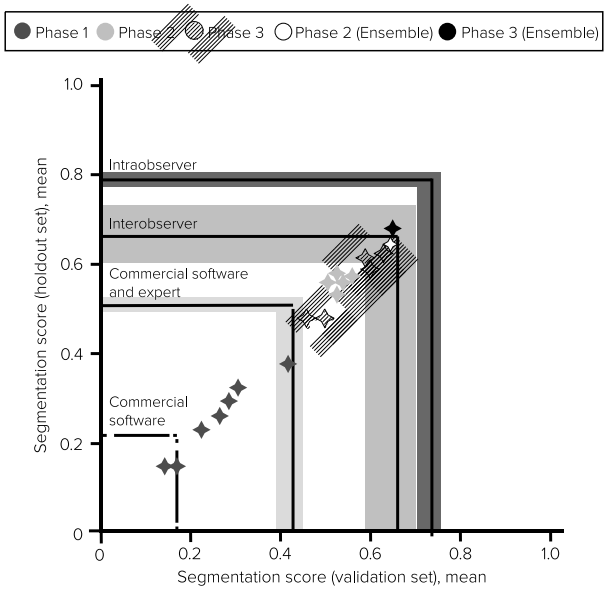
Institute, on the possibility of automating this task, leveraging data from 461 patients consisting of more than 77,000 CT image slices.

Using Dr. Mak's data, cleaned and prepared by our lab-based AI factory, two data scientists (physicists with no background in medical imaging) designed a series of contests to find the best algorithm to outline a tumor. We embarked on three sequential contests over ten weeks and had thirty-four contestants submit forty-five algorithms. We gave our contestants a "training" dataset consisting of scans from 229 patients, with the cancer fully outlined across the images by Mak. We held back the remaining dataset to see how accurate the algorithms would be in mimicking Mak's work.

The top five contestants used a variety of approaches, including convolutional neural networks (CNNs) and random forest algorithms. Surprisingly, none of our contest participants had any prior experience with medical imaging or cancer diagnostics. The solutions they developed involved both custom and published architectures and frameworks to perform the tasks of object detection and localization, with open source algorithms originally developed for facial detection, biomedical image segmentation, and road scene segmentation for research on autonomous vehicles. The phase 3 algorithms produced segmentations at rates between fifteen seconds and two minutes per scan—substantially faster than a human expert, who took eight minutes per scan. The ensemble of the five best algorithms performed as well as a human radiation oncologist (interobserver), and better than existing commercial software, as shown in figure 3-4.

FIGURE 3-4

Results of LISH analysis contest using data from the Dana-Farber Cancer Institute



We cite this example not only because we're proud of it but also to demonstrate that an organization doesn't have to be rich in data, IT resources, or AI talent to construct an AI factory. To create ours we tapped resources that are available to everyone. And the benefit we got from it is invaluable. We shared our findings in the *Journal of the American Medical Association Oncology*—not where you'd expect to find the work of business school faculty.[17]

We admit that it's relatively easy to tap the power of AI within a small laboratory. We did not have to deal with large, siloed organizations or complex, outdated, and mismatched IT systems. As AI enables more of the operating processes in complex corporations, the way it is embedded and architected in the broader operating model becomes increasingly critical. This is why a firm's operating architecture has become a strategic consideration that should be thought through at the most senior levels. This is the topic of the next chapter.