How to create business value with AI

12 stories from the field

IBM

How IBM can help

Clients can realize the potential of AI, analytics, and data using IBM's deep industry, functional, and technical expertise; enterprise-grade technology solutions; and science-based research innovations. For more information about AI services from IBM Consulting, visit ibm.com/services/artificial-intelligence.

For more information about AI solutions from IBM Software, visit ibm.com/Watson.

For more information about AI innovations from IBM Research®, visit research.ibm.com/artificial-intelligence.

For more information about the MIT-IBM AI Lab, visit mitibmwatsonailab.mit.edu.

Executive summary

Many common notions about artificial intelligence (AI) are actually misleading myths, churned out in the hype cycle that has become endemic to so many emerging technologies.

A peek behind the AI curtain

The IBM Institute for Business Value (IBV), in collaboration with the MIT-IBM Watson AI Lab, interviewed individuals involved in deep-learning projects from over 35 real-life artificial intelligence (AI) implementations around the globe. We talked to business and technology experts from more than a dozen industries about their AI goals, challenges, and learnings.

Small gains versus scaled transformation

We confirmed that AI uptake continues to increase, but most organizations are not yet using it fully for broad transformation. Instead, many are just addressing discrete business challenges. By the end of 2022, we estimate that just one out of four large companies will have moved beyond pilots to operational AI.[1]

Moving beyond the myths to what's really happening with AI

As enterprises adopt artificial intelligence, it's important that C-suite and other leaders not buy into some of the myths surrounding it, such as "AI shortcuts don't work" or "If it's not deep learning, it's not AI." Instead, they need to make decisions grounded in AI reality.

Learn from your peers across industries

In this piece, we pull back the curtain on five myths, revealing through data and real-world examples the truth about how companies are using AI, so organizational leaders and teams can learn from their peers.

Introduction

It's a matter of perception versus reality.

While headlines herald artificial intelligence (AI) as a panacea for mounting economic malaise, executives are still left wondering: what are companies actually doing with AI? What results are they achieving and how?

The IBM Institute for Business Value partnered with the MIT-IBM Watson AI Lab, interviewing more than 35 organizations to help answer these and other questions. What we learned is how business and technology experts involved in deep learning projects are applying artificial intelligence in the real world of business to drive real value.

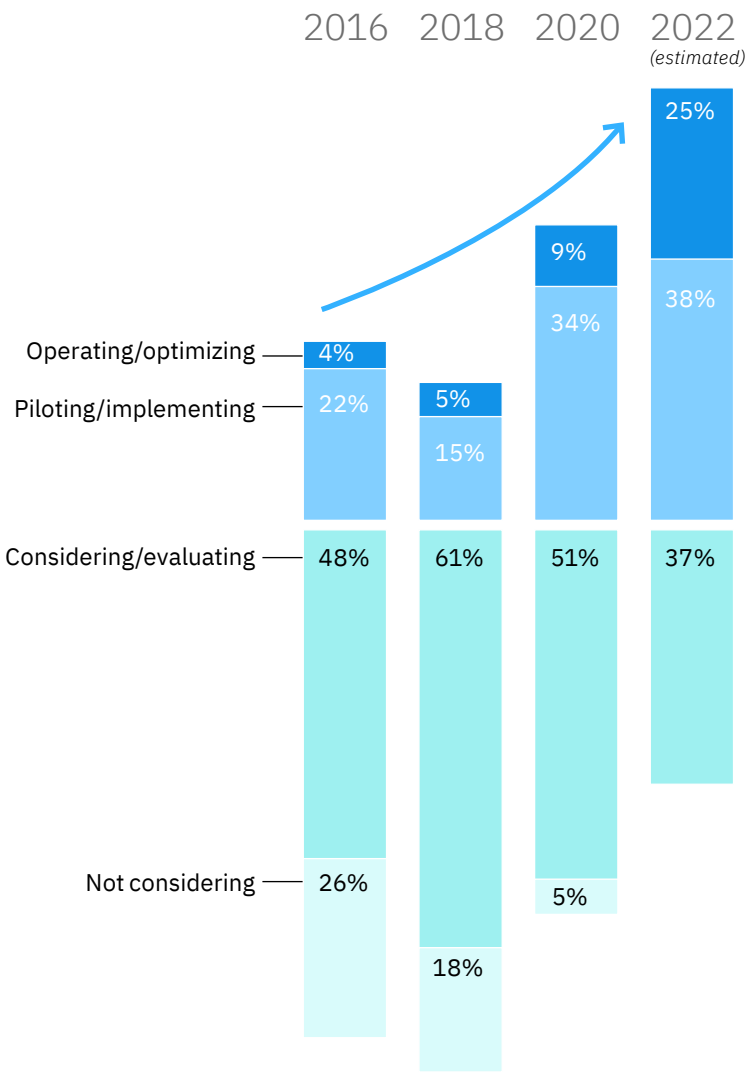
AI: Beyond the numbers—to the stories

Artificial intelligence continues to steadily advance through its technology adoption curve—or hype cycle, for the more cynically minded (see Figure 1).

FIGURE 1

AI adoption* 2016–2022

By the end of 2022, we believe one out of four large companies will have moved beyond pilots to operational AI.



** Note: AI adoption includes piloting, implementing, operating, or optimizing. For details, see endnote 1.*

The pandemic narrowed, then accelerated, organizations' adoption of AI. The number of companies that were piloting AI use cases in the midst of the pandemic more than doubled from 2018—and recent data indicates their ranks continue to grow.[2]

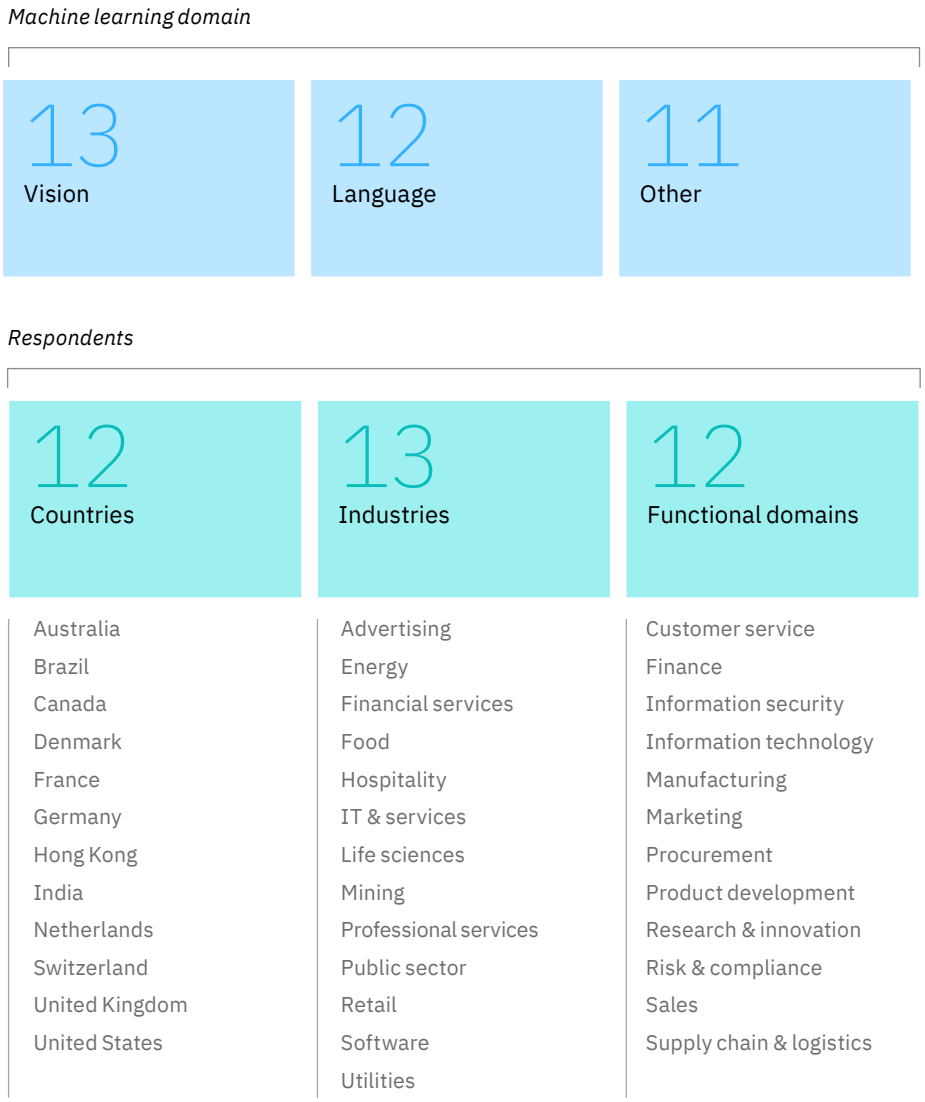
While these numbers show an upward trend, they don't tell the full story that many business and technology leaders need in order to benchmark their own organizations' use of AI.

To get to that story and the challenges AI is helping to solve, we interviewed individuals involved in deep learning projects around the world. From April through August 2021, we talked with business and technology experts in more than a dozen industries about their AI goals, challenges, and learnings (see Figure 2).

FIGURE 2

Scope and scale of our interviews

Our interviews highlight how tailored uses of AI can solve distinct business problems.



What did we learn about the state of AI?

Can AI be an enabler of top-line growth? Absolutely. For some innovative AI adopters, such as NVIDIA, NavTech, and others, AI helps create entirely new offerings and even new business models.

Few enterprises, however, are using it for such broad transformation yet. Instead, they are mainly tackling discrete, tangible business problems. Organizations around the globe are using AI to help reduce costs, enrich customer and employee experiences, increase win rates, optimize supply chain performance, and much more.

We also learned that many common notions about AI are actually misleading myths, churned out in the hype cycle that has become endemic to so many emerging technologies. Unfortunately, these misconceptions often deter and distract organizations from engaging with the more pragmatic realities of AI.

In the pages that follow, we debunk five of the most prevalent AI myths by highlighting relevant insights and practical examples from our interviews. The observations and anecdotes that follow—curated from our discussions with more than 55 professionals at more than 35 organizations on the front lines of AI—can help disentangle fact from fiction. They allow a peek behind the curtain—a virtual check-in with peers—as organizations seek to increase AI's impact and value. (Readers who want additional insight will find 12 detailed case studies in the appendix).

Myth versus reality

Myth 1

AI is a one-size-fits-all proposition

Myth 2

If it isn't deep learning, it isn't real AI

Myth 3

Cost reduction is AI's sweet spot

Myth 4

Shortcuts don't work in AI

Myth 5

AI only delivers value for the problem at hand

Myth 1

AI is a one-size-fits-all proposition

Reality

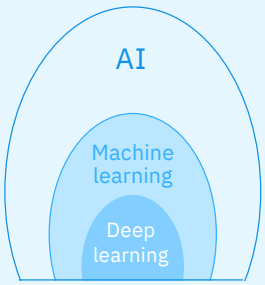
Fit-for-purpose matters. AI-driven business improvements stem from many techniques.

For example, one of those many AI techniques, deep learning, is often best suited to problems related to underlying (often large) data sets in vision, language, and other predictive models. From virtual assistants to fraud detection, deep learning is changing the way we work and play. In these situations, traditional machine-learning techniques may be less effective.

But not every business challenge or desired outcome is a fit for AI, despite the hype that might make it appear so. Organizations first need to determine whether a broader strategic initiative or a particular business problem is a candidate for AI enablement, a topic we address more fully in "Rethinking your approach to AI." Companies can start with an assessment of their overall "data wealth," as well as examine discrete business problems.

Perspective

AI, machine learning, and deep learning defined



Like stacked Matryoshka dolls, deep learning is a subset of machine learning, which is a subset of artificial intelligence. These techniques are often complemented by robotics, sensors and actuators from the Internet of Things, virtual interfaces, and other adjacent technologies.

What is AI and how did it come about?

AI allows computers to perform tasks that previously could have been done only by humans. But where human capacity begins to plateau in terms of accuracy, speed, and processing power, AI really begins to gain traction.

While AI hype is very much grounded in the 21st century, AI was actually born decades ago in the mid-20th century. In 1955, two mathematics professors (from Dartmouth and Harvard) and two research scientists (from Bell Labs and IBM) suggested that a "2-month...study of artificial intelligence be carried out during the [following] summer." According to the proposal's abstract, "an attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves."[3]

Thus, the first formal definition of artificial intelligence was created, with academia and the enterprise working to build increasingly better AI ever since.

What is machine learning?

Machine learning, as explained by the authors of the MIT Press' "Deep Learning" book: "AI systems need the ability to acquire their own knowledge, by extracting patterns from raw data. This capability is known as machine learning."[4] Put another way, a computer learns from complex data sets, training to become smarter as it learns.

Today, we use machine learning systems for a variety of ends, to select the most relevant results in response to a keyword search, to parse visual images, and more. Increasingly, these AI applications make use of a class of techniques called deep learning.[5]

Deep learning defined

Deep learning is a subset of machine learning, inspired by the way the human brain's own network of neurons functions. Of the machine learning techniques in use today, the most important is deep learning. It can:

- work with unstructured data such as images and free text
- model nonlinear relationships, enabling it to model complex problems
- learn relationships without being pre-programmed on the target task
- improve its predictive power as new data becomes available

For complex problems where enough data is available, deep learning often delivers greater performance than other machine-learning methods.

European clothing retailer uses AI to increase efficiency and sustainability

Demand prediction and selling efficiency have consistently been central to the consumer goods and retail industries; even incremental improvements can impact the business dramatically.

BESTSELLER, a clothing retailer, sought to increase the accuracy of its demand forecasts and predictions to help ensure it sold as much of its clothing as possible. At the time, it already sold 78% of the products it made—a relatively high performance in the volatile world of fashion. But if BESTSELLER could increase the granularity of product attributes used in its forecast algorithms, it could continue to improve efficiency.

When teams determined that traditional analytic techniques had reached their limits, BESTSELLER trained a convolutional neural network (CNN) from images of its clothing. (A CNN is a class of artificial neural network commonly applied to analyze visual imagery.) Doing so allowed BESTSELLER to classify its products based on additional features not otherwise included in its structured data sets.

Feeding these outputs into its core forecasting engine increased selling efficiency to 82% and reduced the design samples needed by 15%—welcome improvements during the broader pandemic-related sales downturn. The change also had a positive impact on sustainability as the company reduced its discounted, donated, or dumped apparel.

McCormick, an American food flavoring company, used AI to supplement the experience of junior food scientists to help them perform equivalently to a senior scientist with 20 years of expertise.

Marketing platform boosts response rate with machine learning techniques

A marketing and advertising agency used machine-learning models to predict consumer receptivity to client campaigns. They included this capability as part of a data and analytics platform that served all clients. A tiny improvement in accuracy at scale can be worth millions (USD) in additional sales, so the stakes are high.

The agency discovered it could raise response rates by 20%-30%, but doing so would also increase compute costs to store, train, and process additional data and model parameters. Fortunately, a move to the cloud enabled more visibility into their costs—providing greater insight into how to manage them. As a result, the team was able to preserve the response-rate lift while increasing the efficiency of their compute utilization and cutting processing costs by about two-thirds.

AI helps call centers, food science, and more

Crédit Mutuel, a French cooperative banking group, used deep learning extensively to assist human call-center agents, saving tens of thousands of hours each month.

In the same vein but a completely different industry, McCormick, an American food flavoring company, used AI to supplement the experience of junior food scientists to help them perform equivalently to a senior scientist with 20 years of expertise.

Other examples from our interviews highlight how tailored uses of AI can solve distinct business problems—across geographies, industries, and even functions. Often, the right approach is clearer after the right data set to solve the problem is chosen—as highlighted by the BESTSELLER and marketing agency examples.

Myth 2

If it isn't deep learning, it isn't real AI

Reality

Large enterprises are solving discrete business problems and attaining meaningful business value with a mix of data science, traditional machine learning, deep learning, and preprocessing techniques.

Many of the advancements in AI research over the past decade have occurred in the field of deep learning. The explosive growth of social media, search, retail, streaming, and other B2C platforms with deep learning embedded throughout their business models has given rise to the fallacy that if it isn't deep learning, it isn't AI.

In reality, deep learning is just one tool among many in an enterprise analytics toolbox that enables AI (see Figure 3 on page 10).

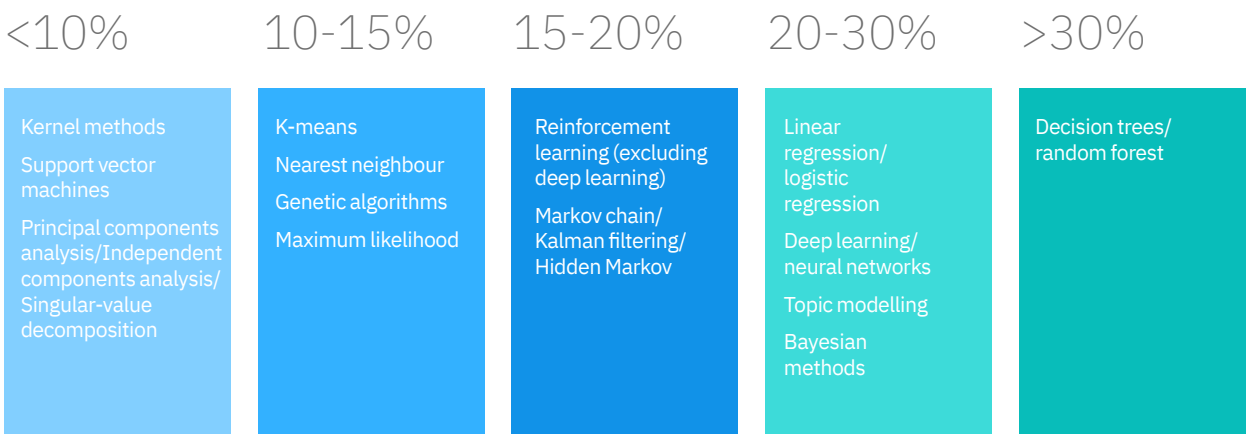
Conceptually, concerns about the costs of deep learning may pose significant challenges to the future direction and nature of AI research (see, "Will cost be deep learning's demise?" on page 11). Pragmatically, determining where deep learning may fall short often occurs by comparing results in practice—usually in a proof-of-concept or pilot.

FIGURE 3

Deep learning is not one-size-fits-all

Organizations employ different machine learning techniques depending upon the business problem.

Percentage of organizations using each machine-learning technique



Source: 2021 IBV AI Capability Survey, unpublished data. Q16A. What Machine Learning (ML) techniques does your organization employ? Choose all that apply.

KPMG uses deep learning and other analytics to help clients save millions

The global tax, audit, and consulting firm, KPMG, initiated an internal hackathon to determine the best approach to reducing the manual effort involved in documenting clients' R&D projects, investments, and tax credits. Doing so could yield tangible business value by helping reduce clients' tax bills each year. The firm found that accuracy ranged from 55% using out-of-the box document discovery software (about as effective as a manual keyword search) to greater than 70% for deep learning.

The best approaches, however, were rule-based machine learning, with accuracy exceeding 85%. Automating these processes translated into a more cost-effective way to save millions (USD) in tax expenses for a given client each year. One client was able to secure an additional 40% tax credit to its R&D spend as a direct result of this approach.

Perspective

Will cost be deep learning's demise?

While artificial neural networks have been around since the 1950s—surviving famines of investment and focus during AI winters[6]—deep learning has been basking in the summer sun since the late 2000s.

A massive increase in compute power to process data, coupled with a rapid explosion of structured and unstructured data, has buoyed the latest phase.

With the continued exponential growth in data—paired with an expectation that Moore's law will reach its end (if it hasn't already)—some AI researchers are concerned about the financial and environmental costs needed to sustain this trend. As Neil C. Thompson, et al put it in a 2021 IEEE Spectrum article: "Clearly, you can get improved performance from deep learning if you use more computing power to build bigger models and train them with more data. But how expensive will this computational burden become? Will costs become sufficiently high that they hinder progress?"[7]

For example, OpenAI's GPT-3 cost $3 million to develop and train, and Alphabet subsidiary DeepMind's AlphaGo is reported to have cost $35 million just to train.

With costs this significant (and rising quickly), the conundrum grows—how to balance the need for bigger models, more data and training, as well as more compute power, with the inherent business realities of budgets and efficiency. Researchers will have to address this conundrum, or progress may languish.[8]

Various research organizations are exploring ways to adapt: different hardware solutions, new AI learning methods, and novel ways of combining powerful data- and parameter-rich deep learning with classical reason- and rule-based symbolic techniques.

An IEEE Spectrum article summed it up: "While deep learning's rise may have been meteoric, its future may be bumpy," as those research efforts unfold.[9]

In the meantime, organizations need to keep a careful eye on their own trade-offs between cost and performance in using deep learning—especially relative to other AI tools.

Zzapp Malaria: Using AI for world good, not just good business

Malaria caused an estimated 627,000 deaths in 2020, with Africa accounting for 96% of all deaths.[10] Zzapp Malaria, 2021's winner of the XPRIZE AI, creates AI-powered approaches to eliminating malaria and delivers them directly to the field through a dedicated mobile application.

In a pilot, Zzapp Malaria's convolutional neural network could analyze visual imagery to detect small bodies of water—potential malaria-carrying mosquito breeding grounds—not readily apparent from existing satellite imagery. It achieved about 75% accuracy but with limited visibility into the drivers underpinning the predictions. While these were good results, they were not good enough to scale to other locations.

Using the CNN, the team extracted 50 topographical and other features from the images, using these features in a traditional linear regression-based approach to determine the likelihood of standing water. The performance was equivalent to earlier results but provided much greater transparency about which factors were driving the prediction. This meant the results were more explainable to the team—and therefore, more transferable to places where terrain differed significantly. They used this AI-driven success to inform adjustments to their approach as they extended their reach to help reduce the incidence of malaria in other locales.

Myth 3

Cost reduction is AI's sweet spot

Reality

Applying AI to business problem-solving can indeed reduce cost, but that's not all it can do. Leading organizations are actively (and strategically) seeking competitive differentiation with AI, achieving top-line-oriented process efficiencies, growth, and business model innovation.

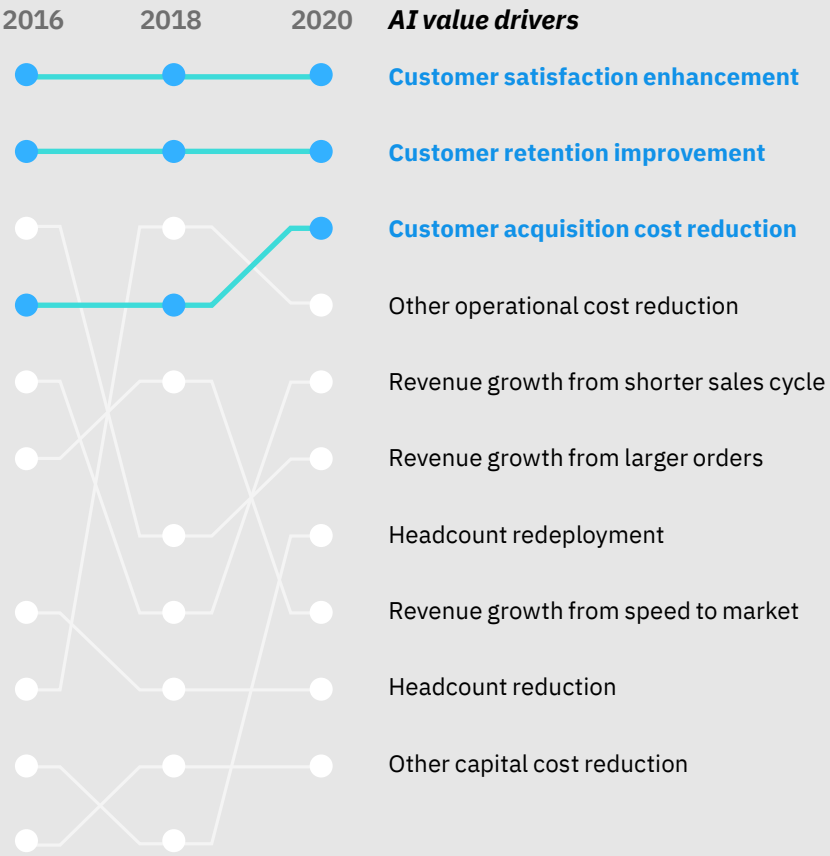
Cost matters, but growth, innovation, and societal good matter more. IBV research shows organizations have consistently ranked customer-centered growth as the top area where they see the greatest business impact from AI (see Figure 4 on page 15).

We wondered whether this top-line emphasis was more wish than reality, but as we talked with executives during our research, we identified several companies who are walking the talk.

FIGURE 4

AI value drivers 2016–2020

Companies are focused on top-line, customer-centered growth.



Source: See endnote 11.

AI drives top-line growth in insurance joint venture

IFFCO-Tokio, an India-based joint venture for general insurance, resolved to enhance its customer experience by paying customers directly for the cost of repairs following an approved claim submission.

Enabling better image capture of the car involved in a crash was the first step. Then teams used deep learning to classify the car model, parts damaged, and damage type. The AI system was able to determine whether the parts could be repaired or would require replacement, and provide a cost estimate, all while keeping a human assessor in the loop to mitigate the risk of fraud.

It was a roaring success: the project paid for itself in less than a year. Settlement costs dropped by 40% and the customer acceptance ratio improved from 30% to 65%. Increased customer satisfaction, retention, and acquisition followed. AI was not just a means to improve efficiency, but a clear driver of top-line growth.

Myth 4

Shortcuts don't work in AI

Reality

While use cases for AI models vary by industry and function, a growing set of "off-the-shelf" foundation and pretrained models can provide a more cost-effective starting point for enterprise data scientists.

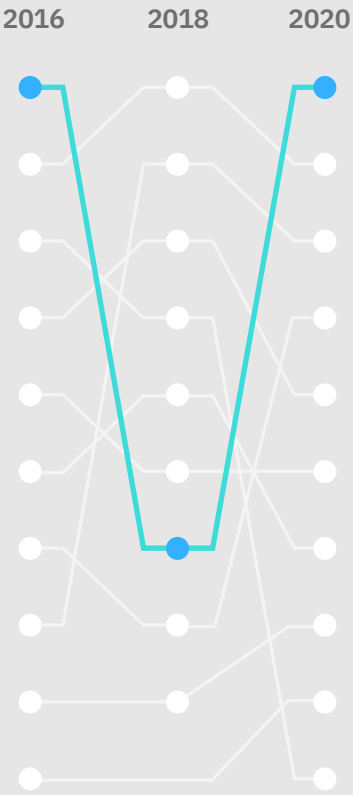
The IBM Institute for Business Value has been taking a systematic approach to quantifying various trends in enterprise AI since 2016. One of the surprises in 2020 was the reemergence of "availability of technology" as a barrier to AI adoption after it ceded top billing to "skills and other factors" in 2018 (see Figure 5 on page 17). We asked ourselves: why is this barrier rearing its head again?

Our interpretation continues to be that organizations are finally realizing what they thought it took to make AI technology work—workers with the right data skills—is necessary but insufficient on its own. The many data scientists companies hired to train various data sets dutifully did as they were asked. But, each business problem—often approached with a different AI model than the last—seemed like starting over. There was no easy way to take advantage of what had gone before.

FIGURE 5

AI barriers 2016–2020

Availability of technology re-emerged as the top barrier to AI implementation in 2020.



AI barriers

Availability of technology

Availability of skilled resources or technical expertise

Regulatory constraints

Degree of executive support

Legal/security/privacy concerns about use of data and information

Data governance and policies for sharing across enterprise boundaries and with external partners

Amount/availability of data to apply and draw context for decision-making

Degree of customer readiness

Degree of partner or stakeholder readiness

Degree of organizational buy-in/readiness/cultural fit

Sources: "Shifting toward Enterprise-grade AI: Confronting skills and data challenges to realize value." IBM Institute for Business Value. September 2018. https://www.ibm.com/thought-leadership/institutebusiness-value/report/enterpriseai. Figure 1, Barriers in implementing AI: 2016 versus 2018, The business value of AI, unpublished data. Q9. What are the top barriers your organization faces in implementing artificial intelligence? Select top 5.

More recently, AI shortcuts are starting to help organizations gain leverage from their AI-driven solutions. What "off-the-shelf" is to software, pretrained and foundation models are to AI. They can be a more effective starting point for new AI projects.

How? By helping organizations get up the curve without generating completely new data sets, instead leveraging AI knowledge gathered from solving one problem to help solve related problems. Key to this approach is transfer learning: repurposing a model originally trained for one task and applying it to a different one—for example, a model for recognizing cars can be applied to recognizing trucks.

Many different types of pretrained models have been designed to solve one or a few specific business problems, and there are a growing number of generalized, gargantuan models (for example, Alphabet's BERT, OpenAI's GPT-3) that can be used to address many challenges. Now even these may have been surpassed by China's Wu Dao 2.0, the world's first trillion-parameter exemplar.

Foundation models can deliver value in a few key ways:

- *Improved economics:* Amortizing costs across multiple use cases
- *Improved results:* Greater accuracy from larger, more robust data sets
- *New capabilities:* Ability to bring together multimodal data more effectively.

But this isn't always true. Adapting pretrained models sometimes results in too large a drop in performance on new data. That is precisely the problem Boston Scientific, a US-based medical device manufacturer, faced—and solved.

Boston Scientific spends $50,000 to save $5 million

Boston Scientific wanted to automate its stent-inspection process to improve accuracy when searching for defects such as broken links or surface imperfections. Accurate inspections are critical for successful clinical outcomes. The US Food and Drug Administration regulates “escape rates” (the proportion of defective parts that might slip through the cracks) based on risk to patients.

Eric Wespi, a data science manager at Boston Scientific, explained, “Human visual inspection is often slow, expensive, and can present unwanted quality risks.” The company has approximately 3,000 experts doing inspections, costing several million dollars each year.

Boston Scientific had already implemented an automated rules-based system that used dimensional measurements and other means to capture common issues. They had tuned the system to be conservative, with a negligible false-negative rate. However, the false-positive rate of 5%-10% was still too high. Too many in-spec parts were being flagged as defective.

Because of their ability to analyze visual imagery, convolutional neural networks would be ideal to tackle this problem, but such models require an enormous amount of data. The team didn’t have enough data to train these models from the ground up. They also recognized that collecting or generating this data would be impractical and cost prohibitive.

The solution? First, they scaled down the problem by focusing on smaller and narrower tasks. Then, they leveraged existing off-the-shelf open source AI models to address the redefined challenge. Finally, they used a smaller data set to fine-tune this system.

The result? The company reaped $5 million in direct savings based on a modest budget of roughly $50,000, as well as accuracy that exceeded existing performance.

Business leaders who want to save time and money with foundation and pretrained models need to remember that they work well for savings—but may not be the best fit when differentiation is the primary goal. With these models available to all—some of them open source—organizations must be careful to select business problems where differentiation matters less. Or they can refocus their efforts on customizing by including additional—often proprietary or proprietarily integrated—data to achieve greater competitive advantage.

Organizations are finally realizing that what they thought it took to make AI technology work—workers with the right data skills—is necessary but insufficient on its own.

Myth 5

AI only delivers value for the problem at hand

Reality

Emergent intra- and inter-company AI network effects are driving real business value across the enterprise.

Proliferating data sources and the increased ability to tap into them provide organizations with a rising wealth of data. Used strategically to fuel thoughtful, ethical AI, companies are reaping not only financial rewards but an uptick in open innovation. This, in turn, continues to deliver other economic benefits of scale—especially among more advanced AI adopters. As we noted in 2020 in our report on the business value of AI:

"Network effects—even if just internal to the enterprise—appear to extend the benefits of AI investments even further. Initial analysis suggests that investing in AI in one area of business operations tends to amplify organizational adaptability and resilience in other areas, resulting in corresponding financial gains. For example, improving data governance and access policies in one function extends to adjacent functions as part of their teaming and collaboration across a workflow. This finding is especially strong for AI investments in core or backbone functions that have particularly strong cross-organization influence or impacts, such as finance, IT, or HR."[12]

For example, rotating AI talent from one department or project to another allows cross-pollination of expertise—and continued organizational learning for people—across a company. This approach helps grow overall AI acumen versus allowing it to stagnate.

Network effects and other synergies between AI and other digital transformation technologies—like cloud, the Internet of Things, security, and data management—add to the value that can be realized.[13]

As with many emerging technologies, we see that what AI was already catalyzing within institutions has also begun to manifest across institutions.

NVIDIA helps foster open innovation in the automotive market

Tech company NVIDIA's approach to business model innovation highlights how the use of AI can spread to customers and business partners. To help address the enormous compute challenge for autonomous vehicles—which some car companies do not have the experience, hardware, and data to develop on their own—the company is creating a shared set of AI-enabled capabilities:

- Common data platform across multiple customers
- Simulation for training and testing
- Common processing of visual tasks.

Depending on their needs and on their existing capabilities, participating carmakers can either lease autonomous vehicle hardware to train their own models based on a larger data set, or use pretrained models from NVIDIA. In either case, instead of making significant capital investments in hardware and AI development capability, carmakers can book the technology as operating expenses, and benefit from improvements as the hardware and software improves.

What we see manifesting across institutions with AI ultimately points to forces that could impact whole economies.

Many of the high-flying B2C platforms—along with hardware/software firms, academic institutions, and governments—have invested significant sums in advancing AI research, often making them available in the public domain. These platforms also have been steadily developing the transferable skills across their executive ranks and knowledge worker base—free agents, all—primarily from their direct experience at the vanguard of AI adoption.

The potential energy of greater professional mobility, accelerated by the dynamic workplace shifts during the pandemic, is being released by the "Great Resignation" into kinetic energy that has the power to transform global economies.

What stands in the way?

Knowledge diffusion—the distribution of knowledge and talent—does not always line up neatly with "absorptive capacity"—the ability of organizations to adapt and integrate those insights and skills.[14] Institutional barriers can get in the way, as can rigid management that resists the discomfort change can bring.

The transformative value of AI—through its financial, economic, and societal impacts—can only become reality if leaders of more traditional enterprises set aside nostalgic notions of how things have worked in the past. They must fully grasp the opportunities for innovation strategically, thoughtfully, and concretely.

A critical starting point is to separate perception from the emerging reality of AI.

Further reading

Since late 2020, the IBV has been developing a series on building a world-class AI capability. It takes a holistic, enterprise-wide point of view on AI and weaves together many of the relevant themes necessary to realizing financial and economic value from adopting AI.

Each of these pieces has a set of concrete recommendations relevant to a specific theme—also tailored to enterprises that are more or less mature in their adoption of AI business practices.

We suggest consulting these tangible action guides—synthesized from dozens of AI projects and the expertise of hundreds of AI practitioners and other experts—found in each of the following:

- *Strategy and vision:* Rethinking your approach to AI
- *Data and technology:* Dealing with the AI data dilemma
- *Engineering and operations*: Proven concepts for scaling AI

About
the authors

Nicholas Borge

Researcher, FutureTech,
MIT Computer Science and AI Lab
njborge@mit.edu
linkedin.com/in/nicholasborge

Nicholas Borge is a member of MIT's FutureTech project team, where he contributes to research on the economics of AI and the Future of Work. Nick has previously served as Director of Intelligent Automation at Sony Music, founded an AI startup, and has over 11 years of experience in strategy and technology consulting for Fortune 500 companies. Nick has an MS in Engineering and Management from MIT and is a Fellow of MIT's System Design and Management program.

Subhro Das, PhD

Research Staff Member,
MIT-IBM Watson AI Lab
subhro.das@ibm.com
linkedin.com/in/subhrodas/

Subhro Das is a Research Staff Member at the MIT-IBM Watson AI Lab in IBM Research. As a Principal Investigator at the lab, he works on developing novel AI algorithms in collaboration with MIT. His research interests are broadly in the areas of optimization methods for machine learning, reinforcement learning, trustworthy machine learning, and human-centric AI algorithms. He received MS and PhD degrees in Electrical and Computer Engineering from Carnegie Mellon University.

Martin Fleming, PhD

Chief Revenue Scientist, Varicent
martin@fleming41.com
linkedin.com/in/flemingmartin

Martin Fleming is Chief Revenue Scientist at Varicent, a Toronto-based sales performance management software provider. Martin is also a research fellow at The Productivity Institute, a consortium of eight UK universities. His research is at the intersection of technology, productivity, and economics, and he is the author of the upcoming publication, "Breakthrough, A Growth Revolution." Previously, Martin served as IBM's Chief Economist and Chief Analytics Officer.

About
the authors

Brian Goehring

Global Research Lead, AI,
IBM Institute for Business Value
goehring@us.ibm.com
linkedin.com/in/brian-c-goehring-9b5a453/

Brian Goehring is an Associate Partner in the IBM Institute for Business Value, where he leads the AI business research agenda, collaborating with academics, clients, and other experts to develop data-driven thought leadership. He brings over 20 years of experience in strategy consulting with senior-level clients across most industries and business functions. He received an AB in Philosophy from Princeton University with certificates in Cognitive Studies and German.

Neil Thompson, PhD

Director, FutureTech,
MIT Computer Science and AI Lab
neil_t@mit.edu
linkedin.com/in/neil-thompson-5724a614

Neil Thompson is the Director of the FutureTech research project at MIT's Computer Science and Artificial Intelligence Lab and a Principal Investigator at MIT's Initiative on the Digital Economy. Previously, he was an Assistant Professor of Innovation and Strategy at the MIT Sloan School of Management and a Visiting Professor at the Laboratory for Innovation Science at Harvard. He has worked at organizations such as the Lawrence Livermore National Laboratory, Bain & Company, the United Nations, the World Bank, and the Canadian Parliament. He has a PhD in Business and Public Policy, and master's degrees in Computer Science and Statistics from the University of California, Berkeley, as well as a master's in Economics from the London School of Economics.

Contributors

Adam Bogue
Business Development Lead, IBM Research

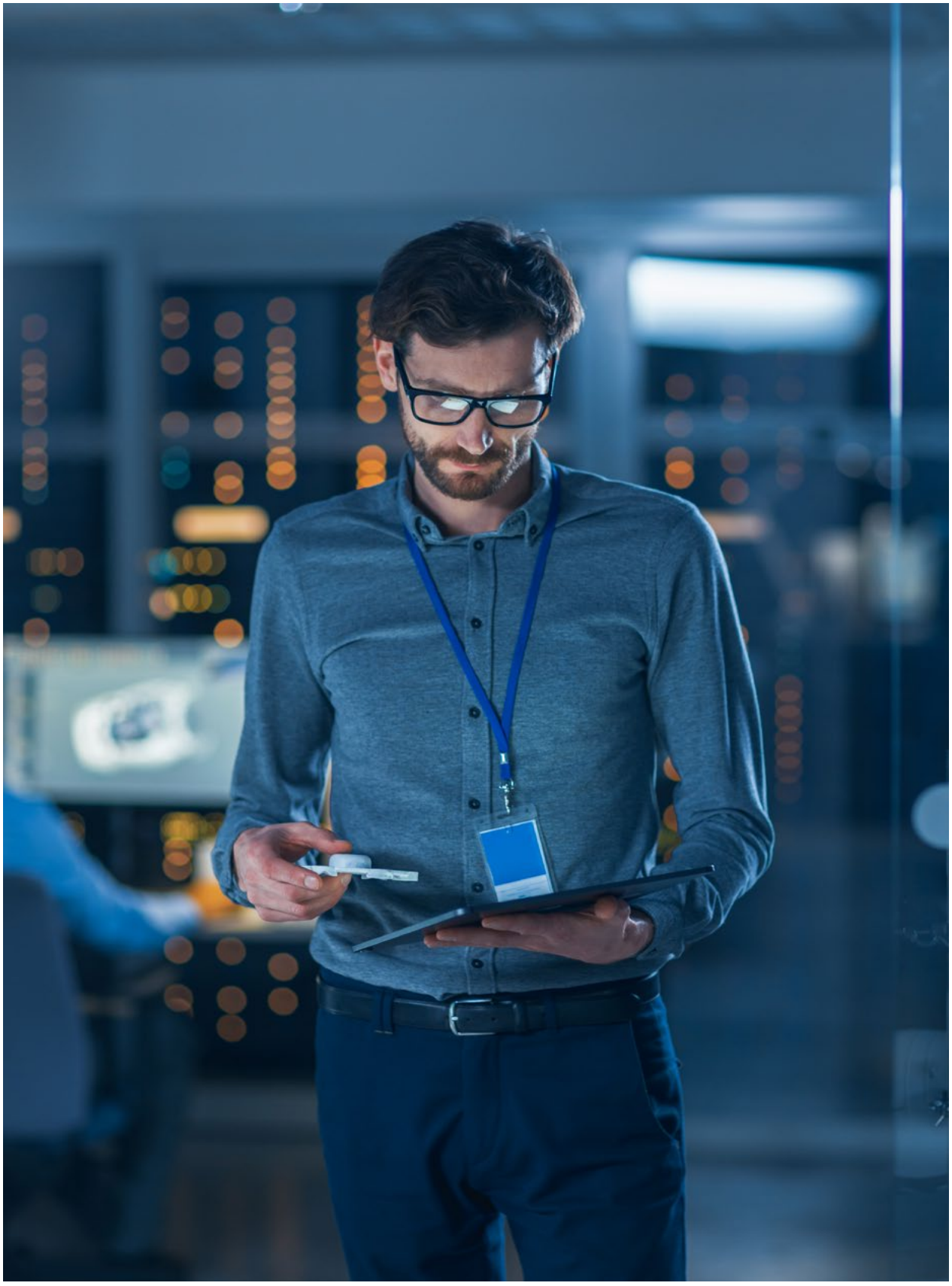
Alex Gorman
Program Director, Client Advocacy, IBM Software

Cathy Reese
Senior Partner, Practice Leader, IBM Consulting

Shannon Todd-Olson
Senior Partner, IBM Consulting

Acknowledgements

The authors and contributors would like to thank the MIT-IBM Watson AI Lab, and its codirectors Aude Oliva and David Cox, for funding this project—as well as Seth Dobrin, Glenn Finch, and Sriram Raghavan for their support.



Appendix

Scope and scale of detailed case studies

Case studies across industries, functions, and machine learning technique.

Company interviews

Name	Industry	Function	Business solution	Machine learning technique
BESTSELLER	Consumer	Fashion design	Better design and selling efficiency by extracting product attributes from catalog images	Vision
Boston Scientific	Industrial	Medical device design	Reduced labor cost by automating visual stent inspection using transfer learning	Vision
Crédit Mutuel	Banking	Customer service	Quicker calls by suggesting better answers for customer advisers using hierarchical NLP	Language
Global Bank	Banking	Internal audit	Increased audit capacity by improving documentation quality using an "instant proof-reader"	Language
IFFCO-Tokio	Insurance	Claims automation	Reduced insurance payouts by issuing directly to claimants using automated assessments	Vision
KPMG	Professional Services	Tax credits	Increased tax credits by surfacing better documentation through document search	Language
Marketing Platform	Professional Services	Ad targeting	Controlled cost of training targeting models by exposing marginal incentives of experiments	Other
McCormick	Consumer	R&D/product design	Improved R&D efficiency by suggesting initial flavor profiles for experimentation	Other
Navtech	Information Technology	Sales	Enabled access to digital product catalogs by building a computer-vision platform	Vision
NVIDIA	Information Technology	Autonomous driving	Unlocked a new business model by pooling data and offering AV tech as a service	Vision
Suncor	Energy	Site operations	Developed early warning of diesel production issues by predicting adverse processing conditions	Other
Zzapp	Information Technology	Public health	Used satellite imagery to identify standing water for antimalarial insecticide treatment	Vision

Source: Neil Thompson, PhD: http://www.neil-t.com/teaching-cases/

BESTSELLER

Using AI to unlock the value in your company's data

Summary

Demand prediction relies on making product characteristics available to algorithms. The more information is available, the more variability in historical demand patterns can be captured, and the better the future predictions are likely to be.

However, assessing more granular attributes of products can be difficult and time consuming. Deep learning provides a solution by rapidly and accurately classifying products with minimal manual intervention, thus increasing the features available for the prediction algorithms. Fashion company BESTSELLER illustrates how this works.

Opportunity: Reducing waste and improving turnaround time

In the fashion industry, around 80% of merchandise is sold across two seasons each year, and everything else is highly discounted, ultimately donated, or dumped. This over-production means sub-optimal profits but presents an enormous sustainability issue as well.

BESTSELLER designs, makes, and sells clothes for the Indian market. For each of its four brands, its team designs and mocks up 3,500 samples, but selects only 1,100 for production. These successful candidates are stratified into 5,000-6,000 SKUs by color, size, and more; 1.5 million pieces are produced. Of these, BESTSELLER can sell approximately 78%, which is relatively good performance in the fashion industry. An opportunity exists, though, to increase this percentage even further by better matching production to customer preferences. With plans to more than double its portfolio of brands from four to nine by the end of the year, improving the sell-through rate can have an outsized impact on profitability.

Challenge: Limited design elements available for analysis

BESTSELLER set out to better understand the factors that drive sales of a particular product. This would inform the design process so the company could improve selling efficiency, more nearly matching the number of products sold with the number produced. It could also potentially improve design efficiency. However, an initial analysis using data on product attributes like color and size, as well as stocking, and location, found that there was simply not enough information about the products to create meaningful inferences. The team needed a richer data set.

Clothes can be described in terms of shape, cut, fabric, styles, and various design elements. In fact, BESTSELLER was using a taxonomy of more than 7,000 design patterns and 4,000 colors. While many of these features are discernable simply by observing images of the product, very little of this information was tagged in the product master data. BESTSELLER needed a way to extract this information quickly and effectively.

Solution: Parsing images to enrich the features available

The answer was to extract additional features directly from images using computer vision. BESTSELLER took 10,000 images (one season's catalog) and developed a model for each of its four brands. In just three weeks, the company's team was able to develop and train a convolutional neural network (CNN) to classify an image according to various features. These deep-learning derived features could be fed into traditional analysis techniques such as regression or principal component analysis, (for example, regression or principal component analysis) to better understand the factors that drive sales.

Outcomes: Improved design sampling and selling efficiency

Even with the global sales downturn from the pandemic, BESTSELLER saw remarkable improvements in both sales and design efficiency over the past 1.5 years. Selling efficiency rose to 82% (up four percentage points from 78%), and the company reduced the number of design samples created for each brand by 15%, without any decrease in the final number of designs selected. Sampling efficiency increased by enabling designers to focus on a smaller number of designs with higher likelihood of uptake.

Boston Scientific

Avoiding the pitfalls of transfer learning

Summary

Transfer learning involves repurposing a model originally trained for one task to use it for a different task. In this sense, the knowledge gained while solving one problem can be applied to a related problem--for example, a model for recognizing cars can be applied to recognizing trucks.

Transfer learning can save work and help reduce training costs, but it can also come with an outsized (up to 45%) drop in performance, so there are limited applications where it makes sense. Boston Scientific's experience, however, shows an organization can still achieve high performance with transfer learning by "stepping down" the problem, enabling its model to achieve performance of over 99% and labor savings of over $5 million.

Opportunity: Costly stent inspection is critical to patient safety

Boston Scientific produces stents for a range of surgical applications. Teams need to inspect them to ensure there are no defects such as broken links or surface imperfections. Accurate inspections are critical for successful clinical outcomes, and so escape rates (the proportion of defective parts that might slip through the cracks) are regulated by the US Food and Drug Administration based on the risk to patients.

Traditionally, human experts have done much of the inspection, but this is not optimal. As Eric Wespi, a Boston Scientific data science manager, explained: "Human visual inspection is often slow, expensive, and can present unwanted quality risks." This makes intuitive sense; people typically don't perform well on tasks that require focused attention for a long period of time where the probability of an event is infrequent. Moreover, judgement can vary from person to person. Also, experts' time is expensive; Boston Scientific has approximately 3,000 experts performing inspections at a cost of several million dollars each year.

Challenge: Image classification requires a lot of data for training

Boston Scientific had already implemented an automated rules-based system that used dimensional measurements and other means to capture common issues. The team had tuned the system to be conservative, with a negligible false-negative rate. However, the false-positive rate of 5%-10% was still too high. Too many in-spec parts were being flagged as defective for human inspectors.

Convolutional neural networks (CNNs) are particularly well-suited to image classification, but such models require an enormous amount of data to train. In many cases (particularly for newer and rarer defects), the team did not have enough data to train these models from scratch. Collecting or generating this data would be impractical and the cost prohibitive.

Solution: Transfer learning applied to a scaled-down problem

The team wondered if they could do better by starting with a pretrained model. They applied the following approach:

1. *Scale down the problem:* For each defect, inspection could be segmented into smaller and narrower tasks such as, "Does this portion of the image contain a link?" and "Is this link broken or not?".
2. *Customize existing models:* Several open source CNNs were used (for example, VGG16, EfficientNet [B0 through B7], Mask R-CNN, YOLOv3, ResNet-50, and Inception-v3). In each case, the team started with the open-source model's pretrained weights, customized the last couple of network layers, and then retrained the models using their own data.
3. *Test data requirements:* The team found that they needed less data than expected (for example, 100-1,000 examples of each defect and 50,000-60,000 examples of non-defective stents) to exceed human performance.

To improve the robustness of the models, they also augmented the training data by generating additional examples through perturbation (these could be simple adjustments that should not impact classification, such as brightness adjustments or the addition of noise).

All the work was completed within a relatively modest budget of $50,000. Model training was quick and inexpensive, taking 1-2 seconds per image across nine models, 2-10 hours to train each model on a single GPU, and a small team of approximately three people.

Outcomes: Dramatic model performance and reduced labor costs

The resulting accuracy was above 90% for all models, with even smaller networks like VGG16 performing well for simple problems. Accuracy increased for more sophisticated models and with more data--for example, EfficientNet can achieve up to 97% for a B0 network with 100 examples and above 99% for a B7 network with 1,000 examples.).

This level of performance is not what is typically expected with transfer learning. Performance usually drops significantly, requiring more data to offset the deficit. In this case, applying the existing models to a simpler problem appears to have eliminated that need.

Deploying the nine models enabled an equivalent of $5 million in direct labor savings from the reduction in parts being flagged for human inspection, and the opportunity to reassign several experts to other high-value projects.

Boston Scientific's experience suggests that transfer learning works well with the right conditions:

- A generic model exists to be leveraged. In the case of image processing tasks, the early layers of such networks seem to be highly transferable even when the task is notably different.
- The usual drop in performance from transfer learning can be eliminated by using the system on a simpler problem and doing finetuning on the network.

Crédit Mutuel

Using AI to get the right information to customer advisors

Summary

The drive for efficiency in customer service is often at odds with a desire to deepen the customer relationship. Specialists are better able to address questions related to specific products or services but can lack the context necessary to serve customers whose relationship with the company is broader than one domain alone. Credit Mutuel employed AI to scale dedicated points-of-contact for customers across many products by providing curated information to its customer representatives.

Opportunity: Enhance the service provided by human advisors

At Crédit Mutuel, each customer has a dedicated advisor. The advisor acts as a first point of contact, helping customers navigate their relationship with Crédit Mutuel across various products in areas like checking, savings, mortgages, and investments. The quicker and easier it is for advisors to access relevant information, the faster they can respond to customer requests (and the more time available to serve other customers). With approximately 3 million incoming calls and 7 million emails received per month, improvements in resolution time can have a significant impact.

Challenge: Inconsistent documentation across products and groups

The challenge of having a single advisor across many products is the burden that it places on that advisor to have the necessary information at their fingertips. To resolve a customer query, advisers (who are typically generalists) use internal search engines or phone calls to source answers about specific products. But individual banks in the Crédit Mutual network organize their information differently, which complicates the search. Moreover, the language and terminology can also differ. This means that typical, off-the-shelf language models are insufficient for prioritizing the information presented to these advisors.

Solution: Custom word embeddings and hierarchical classification

To create a language search customized for its products, Crédit Mutuel first collected all the questions faced by its customer advisers over a three- to four-month period, and then curated answers to those questions (which took an additional four months), repeating this effort for each of 11 business domains currently in production. Then teams trained a deep learning model for custom word embeddings and used this to train an individual support vector machine (SVM) model for each domain to select the answers most likely to address each question. The company also built tens of thousands of dialog steps to support the collection of any missing information from the initial question. The initial domain classification (which in this setup could focus on only short, simple opening questions) was developed using a FastText[15] model which performed as well as the next-best attempt, BERT, but was much quicker—yielding an F1 score of 90% with only 10-15 seconds weekly training time and 20-30 milliseconds classification time. Splitting in this way helped to minimize the number of classes in each domain-specific SVM model.

Outcomes: Improved answer quality and quicker call resolution

The improved language models enhanced the quality and speed of answers. The Virtual Assistant is now able to provide good answers to 85% of customer cases (and 2 million additional answers to customers each year), while also reducing the time to resolution from 3 minutes to 1 minute on average. The overall time savings (for customers and advisers) was in the order of tens of thousands of hours each month.

This case exemplifies AI being used not to provide a specific answer, but as an integral part of a human-led workflow, generating a smaller and more targeted set of proposed outputs where humans can apply their subjective judgement.

Global Bank

AI can increase capacity by complementing existing processes with minimal disruption

Summary

In highly regulated industries such as banking, it's critically important to maintain process documentation. This enables banks to execute more consistently and repeatably, and to demonstrate compliance during external audits. To help ensure accuracy, completeness, and comprehension, banks conduct internal audits of documentation where their teams review and attempt to replicate process controls. However, the write-ups are often complex and unstructured (free text), and manual audits take time. A global bank shows how deep learning can help scale that manual effort, while complementing, not replacing, existing people and processes.

Opportunity: Audit assurance improves the quality of controls

Banks have many processes, from opening a new savings account to making a transfer, each of which is subject to risks such as fraud or money laundering. These risks are critically important to manage because the financial industry is heavily regulated, with steep penalties for transgressions. Also, with increasing competition from online service providers and others, it's never been easier to switch banks, and trust is a big factor in customer retention.

Banks mitigate these risks through a rigorous system of controls. Some are automated but most are manual. To help ensure that these controls are designed appropriately and continue to be applied effectively, a bank's Internal Audit (IA) department conducts tests by sampling controls and checking to see if they worked. If the banks find an issue (for example, opening an account that should not have been opened), then they implement a corrective action plan (CAP) (such as increasing the frequency of refreshing the list of blocked entities). The more controls that can be checked—and the more often—the greater the assurance provided to the business.

Challenge: Improve documentation quality to facilitate efficient audits

To replicate controls and assess their effectiveness, the IA department relies on good documentation. This should, at a minimum, contain sufficient information to identify what needs to be done, how it should be done, and what the expected outcomes are. If any of this information is missing, then the auditor may need to talk to the control owners/documenters on revisions, and this increases the effort required for the audit. Further, the documentation is critical to support regulators' understanding of the controls being applied, and as such, it should also highlight responsibilities, timings, and other process-level information.

The audit process represents a significant manual effort as is. Global Bank conducts approximately 1,000 audits per year, each of which looks at around 10 controls and takes three hours on average to complete. Global Bank continues to increase its capacity (the number of auditors will grow by 30%), and already has one of the largest internal audit departments in the world, so it's important to optimize the productivity of those resources.

Solution: Using NLP to proactively flag potential gaps in information

Global Bank set out to improve the efficiency of the audit process by improving the quality of documentation using an "instant proofreader" for document writers. The idea was to develop a natural language processing (NLP) model to automatically flag any important information that might be missing from control documentation based on a 5W's test—What, Why, Who, When, Where. Global Bank could use the system when the document was initially written, or could scan across existing documents to identify those where issues may exist.

Global Bank built a proof-of-concept based on a pretrained Bidirectional Encoder Representations from Transformers (BERT) model (a technique for NLP). It used the model for named-entity recognition, attempting to identify terms that represented each of the 5W's). To recognize Global Bank-specific terminology, the model would need to be fine-tuned, and because of new audits coming in and planned rollouts of the capability to other departments, the model would need to be retrained multiple times. However, BERT was a large, complex model and would require significant compute resources to retrain. Additionally, Global Bank was limited to using on-prem hardware for security reasons.

The solution was two-pronged. First, Global Bank built a new model that was easier to retrain. It partnered with IBM to build this model, bootstrapped from a prior engagement, and put in an on-prem implementation using IBM Watson® Studio. Second, Global Bank set out to increase the data available for this new model. Previously, it had built a tagging system in Python directly connected to their internal audit platform, and the system enabled auditors to make new annotations as they worked on audits. The company augmented this system with the original BERT model, which provided auditors instant feedback, and made this data available for the new IBM model.[18]

Outcomes: More efficient audits and increased auditing capacity

The system provides three key benefits. First, by providing immediate feedback on what's missing from the description of controls, it raises the completeness and accuracy of documentation at the time of writing, which also helps new joiners get up to speed much more quickly. Second, there is improved consistency of write-ups against company standards. Third, the improved quality reduces the need for back-and-forth between auditors and control owners, which enables each to be more productive.

Within just four months, rapid adoption of the system had already yielded results. Fifty active users had collectively input 12,000 entries (custom annotations) across 5,000+ controls and were adding an additional couple hundred entries each week. The increased efficiency of the review process enabled Global Bank to release an estimated 30,000 hours of effort and deploy it towards additional assurance that would not otherwise have been possible.

IFFCO-Tokio

AI process improvements generate better customer incentives

Summary

Like many industries, insurance companies face a challenge in providing the right incentives. In general, insured people are not motivated to seek the best price for services if an insurance company is footing the bill. But sometimes AI can make possible the process changes that enable them to unlock this value.

Opportunity: Pay customers directly

IFFCO-Tokio settles approximately 500,000 motor-vehicle damage claims in India each year. Traditionally, customers sought repairs at privately owned repair workshops, which provided quotations for IFFCO-Tokio to approve, but there were several issues with this process. Workshops are incentivized to inflate their estimated costs, and customers do not have incentives to seek out the best prices. This led to disagreements about repair charges and delays. Customers, many of whom depended on their vehicle for their livelihood, were severely disadvantaged by delays of up to 20-30 days per settlement. Some could not afford to wait, forgoing repairs and driving vehicles that were unsafe.

IFFCO-Tokio resolved to pay customers directly for the cost of repairs. This would provide customers better incentives and empower them to take charge of the process. But there were two challenges: providing fast turnaround on quotations and getting estimates correct without a quote from the repair shops.

Challenge: Manual effort and poor-quality data

Initially, IFFCO-Tokio did this work manually. It developed a smartphone app that customers could use to upload photos of the damage, receive a quote for estimated cost, decide for themselves whether the quote was acceptable, and receive the payment independent of the timeline for repair. Now the customer was in charge.

The new approach proved to be immensely popular with customers, but it was time consuming and imprecise. The turnaround time for this process was up to 5 hours per claim, much of which was assessor time to inspect parts and populate forms with “repair/replace” decisions, as well as cost estimates for each part. Their efforts were also complicated by poor-quality images. Previously, pictures of the damage had been taken in the controlled environment of a professional garage, but customers frequently submitted images with incorrect angles, poor lighting, or glare.

Solution: Guided image capture and variable training data size

IFFCO-Tokio had hoped that machine learning could speed up the process by automatically providing a first estimate for each part, but its team knew the quality and consistency of the images were key. To improve the quality of image capture, the team augmented the app with camera-stencils to guide customers on composition and added additional instructions. They also increased the volume of training data for part types where glare or reflection made damage particularly difficult to discern (for example, 3x the number of images for metallic parts, and 5x for glass parts). This combination of better images for inference and for training enabled the use of deep learning techniques to classify the car model, the parts damaged, and the damage type. Based on this, the system was then able to determine whether the parts could be repaired or would need replacing, along with an estimate for the cost.

When automating a decision-making process, human oversight is generally reduced. This allows potential for abuse, such as duplicate claims for previously claimed damage. In this case, IFFCO-Tokio anticipated an increase in fraud and built a fraud detection engine to identify previously used images. However, it found minimal exploitation in the overall system, as ultimately an experienced human assessor was always in the loop.

Outcomes: Reduced time and settlement costs, increased customer retention/acquisition

The system was a roaring success in ways intended and unintended. It reduced assessor effort significantly. End-to-end time decreased to 30 minutes per claim on average (including negotiation with customers), and the new system paid for itself in less than a year. More surprisingly, IFFCO-Tokio also experienced a reduction in settlement price by 40%, and an increase in the acceptance ratio from 30% to 65%. Moreover, the system offered increased resiliency, as customers were still able to receive their settlements when garages were closed during the pandemic. Finally, and perhaps most importantly, the new system was directly linked to increased customer satisfaction, retention, and even acquisition. AI had become not just a driver of increased efficiency, but also a driver of top-line growth.

KPMG

Task complexity, not data availability, drives the choice of machine learning method

Summary

Natural language processing (NLP) and text-mining approaches often rely on deep learning for named-entity recognition (NER). These approaches enable meaning to be extracted (for example, from sentences and passages) and are often able to perform well on complex tasks. However, KPMG's experience with document search and classification illustrates that when a particular sub-task is simple enough, traditional machine learning may be the best approach.

Opportunity: Better substantiation for R&D tax credit claims

Tax incentives for research and development (R&D) in the US can be significant. They can be up to 11%-15.8% of eligible incremental R&D expenditures—and the benefit can further increase with many states also offering a state R&D credit. For smaller research organizations in particular, these tax credits can be decisive in making an R&D project commercially viable, or in securing investor funding to initiate projects in the first place. KPMG is hired by businesses of all sizes to document the R&D they've done and to help ensure that the business gets the maximum tax credit for that R&D.

Challenge: Intensive manual effort

In the US, the Internal Revenue Service (IRS) assesses the merit of R&D tax-credit claims using a four-part test. The test checks whether the activities being claimed for:

- Involve the creation of a new business component or improvement of an existing one
- Are technological in nature
- Discover new information that eliminates uncertainty related to the methodology, capability, or design of the business component
- Involve a process of experimentation through simulation, modelling, or testing

Clearly there is subjectivity in this assessment, so providing strong evidence is key to a good outcome.

The evidence is typically collected from an organization's documentation and can take many forms. It could include presentations, emails, meeting minutes, lab reports, test records, and engineering drawings. Content is often unstructured, of intractable volume, and/or stored in various repositories. In some cases it may be very limited, such as in agile or continuous improvement environments. In any case, regulations do not specify what qualifies as "sufficient" evidence, so it's important to review as much information as is available and present as much of it as possible that is relevant and high quality.

KPMG supports clients in their audit readiness and is experienced in managing the discovery process. Traditionally this has involved a top-down approach, starting with a list of projects, trawling through document repositories related to those projects, manually searching through the documents using keywords, and reading and tagging specific document sections that satisfy each of the four tests. This is a significant manual effort and necessitates some prioritization— which risks excluding valuable evidence. It also takes valuable time from client scientists and engineers to support the effort. KPMG wondered if machine learning could help them do better.

Solution: Rules-based approaches outperform machine learning

KPMG initiated an internal hackathon with four teams to compete on solving a subset of the problem (document chunking) using alternative methods. The teams were given 1,000 documents with labeled sections and asked to present a confidence score for each document's relevance to each of the four tests.

Documents were chunked into sections by tokenizing words and sentences. Teams tried a range of approaches, including statistical learning, such as regular expressions, support vector machines, decision trees, and random forest; deep learning for named entity recognition (NER), and rules-based approaches. They found that accuracy ranged from 55% using out-of-the box document discovery software (about as effective as a manual keyword search) to above 70% for deep learning. However, the best approaches were rules-based with accuracy exceeding 85%. This was likely due to a relatively high degree of standardization between document formats, making the document chunking task relatively straightforward.

Outcomes: Better evidence provides increased ability to claim credits

The system is now in operation with several KPMG clients to great effect. Each month, it processes upwards of 5,000 documents, and—critically—the search is transformed from a selective, top-down approach to a bottom-up, exhaustive approach. Tax law is relatively static over time, so the system requires minimal maintenance and improvement while providing greater leverage for human expertise. Anecdotal evidence suggests significant impact. For instance, one KPMG client was able to secure an additional 40% tax credit for its R&D tax-credit study because it utilized machine learning to review R&D project documentation in order to determine eligibility for the tax credit.

It is worth reflecting on the relative performance of deep learning to other methods. The results reinforce the assertion that even if there is a large enough data set, deep learning tends to be superior only when the data or the problems are extremely complex. In this case, simple rules and keywords were sufficient to identify the relevant information for each test, while also providing greater explainability.

Marketing Platform

Controlling AI cost through better visibility

Summary

With on-prem infrastructure, the cost of training and running machine learning models is usually hidden from the business. For example, there can be a disconnect between the value generated from the model and what it takes in terms of engineering and compute resources to get there.

This case study with Marketing Platform shows that bringing transparency to the link between computational demand and cost can help in creating incentives and in reducing costs in a meaningful way.

Opportunity: Moving to the cloud to better leverage existing data

Marketing Platform helps retailers and nonprofits to improve the return on their marketing efforts by predicting who will be most receptive to each campaign. A tiny improvement in accuracy, at scale, can be worth millions of dollars in additional sales or donations, so the stakes are high.

The data sets available for these models are enormous. Marketing Platform runs a data cooperative with thousands of members, with 25%-40% of these members regularly contributing data on things like transactions, donations, or subscriptions. It combines this data with compiled third-party data on everything from demographics, to census data, to household incomes. After feature engineering, the data comprises 12,000 variables and covers almost the entire US population.

With such a rich data set, Marketing Platform was at the limits of what its existing on-prem infrastructure could handle. The team was able to train models only on internal data, and even then, only a small portion of it (for example, a 50,000-100,000 sample) at a time. Marketing Platform knew that if it could use more data, there was a huge opportunity to generate additional value.

Challenge: Initially cloud came with a significantly higher price tag

Moving to the cloud (IBM Cloud Pak® for Data) increased the organization's ability to manage its data and leverage all its data assets, both offline and online. The additional scalability of compute resources also enabled training on 600,000 records (up from a max of 100,000) and 800 features (up from 150-200). This, together with machine learning tools like XGBoost,[16] helped deliver a 20%-30% lift in response rate, a dramatic increase in return for clients.

Initially, the scaling of compute resources (along with the initial switching cost and learning curve) meant that total costs increased. In a fully on-prem world, the cost of infrastructure had been independent of utilization, so data scientists were able to run whatever they wanted, constrained only by compute availability. With effectively unlimited scalability, experiments would have to have to be designed more thoughtfully.

Solution: Connecting to marginal incentives helps offset extra costs

Fortunately, moving to the cloud also enabled platform leaders to better understand and ultimately optimize spend. The team could now generate a per-model cost of compute and incorporate these marginal incentives when structuring their data exploration and analysis.

The team was also able to increase the efficiency of their compute utilization by improving cluster allocation, data flow, and the overall modeling pipeline. For the machine learning model itself, they ran tests to optimize the approximately 100 required model parameters and fixed several of them based on what worked well to minimize the number that would need tweaking each time.

Outcomes: Reduced training costs and a mandate for wider rollout

The result is a dramatic reduction in training costs from $1,500 per training run to hundreds of dollars per training run, even with the 30% uplift in model performance.

The platform's success in this area is now fueling a transformation of its data science capabilities, bringing the number of practitioners in the US up from 40 to 4,000 within the year.

Key takeaways from suggest that:

- Visibility of training costs at a per-model run level of granularity can help make the cost of moving to new machine-learning techniques less expensive than would be expected.
- Volume matters: even small gains in accuracy can have an enormous impact to an organization when it can be applied at scale.

McCormick

Augmenting creativity in R&D through AI-directed exploration

Summary

AI is often used to make recommendations based on what has worked well in the past, but this can result in solutions that are simply more of the same. McCormick's experience shows that AI can also be used to explore a solution space and result in new, creative combinations that may not otherwise have been attempted. In this way, AI can help augment and accelerate a process of creative experimentation.

Opportunity: Use AI to accelerate development of new flavor profiles

McCormick creates a variety of products including seasonings, sauces, and flavors (some of which may also be sold B2B to clients for incorporation into third-party products). A formula is a combination of specific ingredients in precise proportions, standardized for consistency, that delivers a flavor profile which describes the eating experience. Flavor creation for B2B products can be a competitive process, with several companies producing formulas in response to client requests. To improve the success of new flavors, McCormick looked at two important factors:

- Scaling the experience of food scientists. A junior flavorist typically apprentices for seven years, during which they build experience and glean valuable insights. It is this cumulative experience that enables creativity, through an understanding of what works, what doesn't, and what are the degrees of freedom in between.
- Improving the efficiency of the experimentation process. To produce a candidate flavor for a client, several steps are involved. Flavorists formulate a range of different flavor profiles, produce samples of the flavors, test them (both in isolation and after cooking in a test kitchen), and iterate until they have something worth submitting to the client.[17] Improving the efficiency of this process means a quicker time to market and better leverage for food scientists' time.

McCormick wondered if AI could help. If flavorists can intuit insights from empirical experience, then there was probably more that could be extracted, more rapidly, through analytics. If these insights could be captured and systematized, then it could help flavorists better explore the flavor space in two ways. The first is finding optimal flavors within an area around what is known well. The second is by finding promising new areas in the flavor space that have not been explored. Ultimately, this would result in faster development and better quality.

Challenge: Credit assignment and large search space

The company leveraged data on approximately 350,000 formulas created over more than a decade, covering product attributes such as category (baked goods, salty snacks), format (seasoning, condiment, wet sauce, dried), amounts, type, and success metrics such as product tasting scores. It also captured functional attributes like shelf stability and flow rate, as well as non-functional attributes like grain size, sodium content, and FEMA[18] numbers (attributes on 40,000 raw materials). With such a high dimensionality in the data set, the team needed a way to condense the problem to keep it manageable.

Solution: Graph representations and reduced dimensionality

A new deep learning system, SAGE, was developed to generate the new flavor profiles. It accepts two principal user-defined inputs: (1) a seed formula (for example, a flavor profile for Korean BBQ) and (2) any specific constraints desired in the output formulas (such as "must have mango"). The system then generates formulas with varying levels of deviation from the seed—four with only minor tweaks that optimize for anticipated performance, four with greater freedom but still subject to the constraints, and four that differed significantly. This gave the flavorists a range of options to iterate from, depending on the desired level of novelty.

To enable this effort, the team needed a few tricks. First, they reduced the dimensionality of the data by aggregating 40,000 distinct raw materials into 3,000 groupings. Second, they sampled 3,000 formulas for training from the 350,000 total, each labeled with a "success" rating. Last, they formulated the model as a graph problem, defining distance metrics between materials where each formula was represented as a vector.

Outcomes: Performance equivalent to 20 years of experience

McCormick observed that junior food scientists using the system could achieve performance similar to that of a food scientist with 20 years of experience, significantly reducing the number of trials required. It also found that the system enabled greater use of global knowledge. In one case, the system recommended a flavor profile from Canada to a flavorist in the US who had no prior experience with the Canadian market—enabling an increase in creative output, while also better targeting the experimentation effort.

Navtech

The new platforms being created by deep learning

Summary

Some innovations can only be economically feasible when a centralized provider can make the up-front investments required and spread the costs over a wide enough customer base. This is increasingly true for machine learning , particularly where the tasks being automated are perception based (for example, image recognition), because these can require significant data and compute resources to develop and maintain. For many individuals and businesses, this technology is simply out of reach.

Navtech identified an opportunity to bring advanced computer vision to individual diamond retailers across the globe by creating a model and delivering it as a service. This approach can be a win-win and is an excellent example of where AI can not only deliver improved efficiency and performance, but also unlock new capabilities and business models in the process.

Opportunity: Digitizing catalogs helps increase sales

India alone has an estimated 300,000 diamond jewelry retailers. Many of these are smaller companies, with limited inventory capacity, who typically increase their offerings through custom-made jewelry. Relative to offering their own inventory alone, adding custom-made jewelry options can result in a doubling of conversion rate (twice as many customers who find something they like and make a purchase).

Visual catalogs are a key part of the sales process. Each retailer maintains one for its own inventory and supplements it with images of other jewelry as inspiration for customers looking for bespoke pieces. Traditionally, these have been physical booklets or magazines, but these formats can only present a limited number of items and can't be refreshed often.

Digital catalogs remove many of these constraints but introduce other challenges. Staff compile images from various sources like inventory photographs, the web, and manufacturers' catalogs, and categorize them manually into folders. The process is slow (30-60 seconds per image, up to a million images), prone to error (there are lots of duplicate images and it's difficult to remember what you've seen already), and results in only very high-level categorization (for example, rings versus necklaces). Ideally there would be a way to automatically create catalogs and enable customers to search based on additional criteria.

Challenge: Computer vision is too expensive for individual retailers

Computer-vision systems that leverage deep learning to classify images could help improve speed and accuracy, but the reality is that they are out of reach for most retailers. Deep learning is resource intensive, requiring enormous amounts of data and compute both to train and implement. The system may not be used frequently enough to justify that expense, particularly when the comparative cost of labor for manual classification is low. For example, in India, retail worker salaries might start at around $100 per month. The business case for computer vision for any one retailer is thus unlikely to be attractive.

Solution: Build once and provide as a service

Dr. M.I.M. Loya, general manager of emerging technologies at Navtech, had an idea: build a computer-vision system and offer it as a service. This could be a win-win. Navtech had the resources to make the initial investment and offer access to the system for a modest fee, and retailers could benefit from low-cost access to the system on an ongoing basis.

Navtech selected three attributes for its pilot and created a deep learning model for each:

- For product category (for example, rings, bracelets) and style, it used a VGG16[19] network to classify images. The open source ImageNet-trained backbone was fine-tuned by custom training the head and the first and second layers of the network.
- For diamond cut (for example, round, square), it instead used Mask RCNN for object detection and classification (having achieved only 55-56% accuracy using VGG16). The training data for this model was labelled by an intern, who manually drew a polygonal mask around the shape of each diamond.

Outcomes: Broader access to cutting-edge machine learning at a manageable cost

The system enables retailers to build larger digital catalogs much more rapidly. It can classify up to 100 images per minute (versus one to two per minute with manual classification), at an accuracy of 90%-93% for product category and style and 85%-86% for diamond cut (versus 80% for manual labeling). The team was also able to achieve this using a relatively small data set (only 3,000 labeled images for each model), which is somewhat surprising. Navtech did perform some post-processing, where the results of one model (such as style) were used to increase the confidence of predictions for another model (such as diamond cut). It's possible, though, that the high leverage from this small data set was also due to a step-down in problem complexity relative to ImageNet (images of jewelry vary less than images of cats, for example).

This experience illustrates the cost-benefit tradeoffs of deep learning, where some use cases can only be enabled by larger, centralized players with the ability to serve wider audiences. Such systems, because they are delivered at scale, must be delivered as part of a wider product and service architecture, underpinned by traditional software development, but are much more cost-effective.

NVIDIA

Building a compute and data platform for self-driving cars

Summary

To build safe and reliable AI models for autonomous vehicles (AVs), enormous amounts of compute power and training data are needed, along with the skills, resources, and expertise required at that scale. Under such conditions, we are likely to see an emergence of larger platforms that can pool data from multiple participants, aggregate sufficient demand to justify the large investments required, and ultimately enable a new business model where AV software can be offered as a service to carmakers and fleet operators.

Opportunity: Solving complex problems to drive mass adoption of AV

Driver-support systems are becoming more widespread and can now perform parking, emergency braking, lane changing, and other functions. Once vehicles are fully self-driving, the applications will be numerous, from freight and mass transit, to on-demand transportation like robo-taxis. The global autonomous vehicle system market was $82 billion in 2021 and is forecast to grow to $770 billion by 2030—a compound annual growth rate (CAGR) of 39.1% (from 2022 to 2030).[20]

Challenge: Enormous data and compute requirements

The data requirement for fully autonomous vehicles is enormous because of the range of planning and control tasks that need to be performed, such as finding pedestrians, as well as detecting road markings and traffic lights. These functions must be robust in varying environmental conditions like weather and varying locales. They also must be able to handle transient, rare events (for example, being cut off unexpectedly). Current systems are improving on the number and effectiveness of these functions, but there is a high bar to demonstrate their reliability in terms of reduced fatalities and injuries. The RAND corporation has estimated that matching a human-level error rate could take 11 billion miles of testing, equivalent to 100 vehicles being test-driven continuously for over 500 years.[21] NVIDIA itself estimates that good performance on certain AV tasks requires training examples in the order of one million scenes.

Considering that each of these scenes will entail data from numerous sensors, the compute challenge for AVs is enormous. To create the perception models needed for a full AV stack, NVIDIA estimates that a productive development team could require in the order of 5,000 dedicated GPUs.[22] A single model can take three to six days to run on 32 GPUs, and there can be 25-50 deep learning experiments for each task. Individual car companies typically don't have the resources in terms of skills, experience, hardware, and data to develop these systems on their own.

Solution: Shared data and compute platform across multiple customers

NVIDIA is addressing these challenges through the following:

- Extending a common data platform across multiple customers: Pooling data between several carmakers increases the data available for training and can enable greater model performance, particularly with edge cases. The quality of the data can be enforced through a reference architecture that specifies the standards for sensor specifications and placement.
- Simulation for training and testing: Hundreds of millions of driving scenarios can be simulated to supplement real-world data and help bootstrap models for silent on-street testing and iteration—running AI in the vehicles to compare what it would have done relative to the driver's actual behavior.
- Common processing of visual tasks: NVIDIA was able to minimize the compute required by jointly training multiple tasks on a single ResNet-based model architecture. Once the full model is trained, the heads (later layers) of the model can be optimized for each given task, without the need to retrain the trunk (earlier layers) of the model. Noting that the compute was not much greater than that required for a single task alone suggests that there is a lot of common processing possible, which makes intuitive sense for the domain of computer vision.

Outcomes: New business model that enables competition at different levels of the stack

Centralizing the management of data in this way enables new possibilities for AV technology. Depending on their needs and on their existing capabilities, participating carmakers can either lease AV hardware to train their own models based on a larger data set or use pretrained AV models from NVIDIA. In either case, instead of making significant capital investments in hardware and development capability, carmakers can book the AV technology as operating expenses, and benefit from improvements as the hardware and software improves.

It also represents the start of a new market dynamic. On one side are vertically integrated carmakers such as Tesla), that can codesign their software and hardware for more seamless experiences. On the other are increasingly modularized carmakers who compete on the quality of their hardware and buy their software from centralized players like NVIDIA (which greatly reduces cost of entry to the AV market and is likely to stimulate greater competition as a result). The success of either of these two paradigms depends on how important the quality of AV software is in terms of the overall experience.

Suncor

Performance and explainability aren't always a tradeoff

Summary

Deep learning often performs well at predicting processes that are nonlinear (small changes can have an outsized impact) and highly coupled (there is a lot of dependency between factors). Often there is sufficient performance improvement from adopting deep learning that firms are willing to have their models be less explainable (that is, black box). But as Suncor's experience shows, when the stakes are high enough, explainability is at a premium.

Opportunity: Better prediction of issues to manage output quality

Suncor Energy specializes in the production of synthetic crude from oil sands. For diesel in particular, this involves removing sulfur and nitrogen by hydrotreating, mixing the straight-run diesel with hydrogen (and a solid metal catalyst such as cobalt) at high temperature and pressure. The process is complex, with several variables (pressure, temperature, flow rates) that interact to affect the resulting quality of the output. These factors must be tightly monitored and controlled to minimize "upsets", where product quality deviates outside acceptable rates and cannot be sold. With diesel production averaging 43,000 barrels per day, there is a strong commercial incentive to avoid upsets as much as possible.

Challenge: High-impact decisions require high explainability

Ultimate accountability for the output quality rests with sitewide leads, who oversee the production sites and make key operational decisions that influence output quality. Understanding which adjustments to make and the resulting effects can often come down to individual experience and judgement. To be defensible, these high-impact decisions need to be articulated with a clear rationale, which means that any analytical technique for decision support must be transparent and well understood.

Solution: Principal component analysis has greater explainability and comparable performance

Suncor set out to improve its prediction capability by developing an "upset-flagging" model where emerging suboptimal conditions could be identified with enough time for corrective action to be taken. The data science team identified 11 factors associated with product quality that needed to be assessed in real time and built models that incorporated 30 different measurements from sensor data. Initially, they explored a wide range of sophisticated, but hard-to-interpret machine learning techniques, including neural networks, long short-term memory, random forests, gradient boosting, and decision trees—with XGBoost (an ensemble technique combining decision trees with gradient boosting) performing the best.

However, when the team compared the performance to simpler traditional statistical techniques, they observed performance that was much better than expected—for example, principal component analysis (PCA) performed at only a 10% performance deficit relative to XGBoost, while being much easier to interpret.

Outcomes: Early warning with transparent, defensible rationale

After thoroughly testing both approaches with its sitewide lead stakeholders, Suncor decided that the greater explainability more than outweighed the performance tradeoff in this case. Together with the predictions, using PCA enabled a readout of the associated weightings behind each factor, such as a ranking of which factors were most important in driving the prediction. The resulting system was able to predict upset events up to an hour in advance, with an accuracy of 80%, every five minutes.

Zzapp Malaria

Learning from satellite images to fight malaria

Summary

With satellite imagery, computer vision is commonly used to identify visible objects, with convolutional neural networks (CNNs) most often being the default choice. Even when the objects themselves are not clearly visible, predictive models using CNNs can sometimes infer the presence of objects based on other characteristics (for example, the region surrounding the object in question). However, as the experience of ZzApp Malaria shows, this is not always the case, and in such situations, traditional techniques such as linear regression may be good enough.

Opportunity: Preventing malaria through treatment of standing water

Malaria caused an estimated 627,000 deaths in 2020, with Africa accounting for 96% of all deaths. Vector control is the main way to prevent transmission through bites from malaria-carrying Anopheles mosquitos. The primary focus thus far has been on commodities like bed nets or indoor spraying of insecticides, but these are only partially effective and don't work outdoors. An alternative approach is to treat standing water within the community (where mosquitos breed and multiply) directly, but such programs have not been systematic or comprehensive enough to be effective at scale.

Challenge: Small water bodies not visible in satellite imagery

The difficulty with water treatment lies in identifying standing water so it can be managed. Larger water bodies are easily visible in satellite imagery, and computer-vision algorithms have been developed to identify them automatically. Smaller bodies, however, are difficult to detect, even with sophisticated satellite-imaging techniques, and they can be covered or only present season to season. If standing water locations could be better identified, then it would be possible to better direct spraying and better control the mosquito population.

Solution: Infer presence of standing water through topography

Zzapp Malaria was founded to address this problem and began by looking at malaria hot spots in Sao Tome. It created an app that enabled on-the-ground inspectors to log the location of water bodies encountered, to track water treatments over time, and to establish a training set of positive examples, such as locations where standing water was present. It also collected satellite imagery (photographs, infrared, and radar), and used this to train a CNN-based object detection algorithm, which performed well for large water bodies, but poorly for small ones (particularly when they are obscured).

As an alternative, the team extracted 50 topographical and image-based features from the images and used these in a traditional linear regression-based approach to determine the likelihood of standing water in each segment of a map. At 75% accuracy, the performance was equivalent to that of the CNN but enabled much greater transparency into which factors were driving the prediction. The team also found that topographical determinants were highly dependent on locale, and that the linear regression was more transferable to other locales than neural network approaches.

Outcome: A transparent and transferable approach to other locales

The relatively high performance of the regression models may have been because they were able to take advantage of characteristics about how water pools as a function of topological features of the data, as opposed to having to infer them with CNN. In any case, the additional transparency and transferability of the regression model is essential to Zzapp's ambition to expanding the approach beyond Sao Tome and to other locales—Ghana, Zanzibar, and beyond—where the terrain can differ significantly.

About Research Insights

Research Insights are fact-based strategic insights for business executives on critical public- and private-sector issues. They are based on findings from analysis of our own primary research studies. For more information, contact the IBM Institute for Business Value at iibv@us.ibm.com.

The right partner for a changing world

At IBM, we collaborate with our clients, bringing together business insight, advanced research, and technology to give them a distinct advantage in today's rapidly changing environment.

IBM Institute for Business Value

For two decades, the IBM Institute for Business Value has served as the thought leadership think tank for IBM. What inspires us is producing research-backed, technology-informed strategic insights that help leaders make smarter business decisions.

From our unique position at the intersection of business, technology, and society, we survey, interview, and engage with thousands of executives, consumers, and experts each year, synthesizing their perspectives into credible, inspiring, and actionable insights.

To stay connected and informed, sign up to receive IBV's email newsletter at ibm.com/ibv. You can also follow @IBMIBV on Twitter or find us on LinkedIn at https://ibm.co/ibv-linkedin.

Notes and sources

1 Sources: “Fast Start in cognitive innovation: Top performers share how they are moving quickly.” IBM Institute for Business Value. January 2017. https://www.ibm.com/blogs/internet-of-things/fast-start-cognitive/ Unpublished data. C&A8. In general, where is your organization in its adoption of cognitive computing? Select the most advanced level for your organization; “Shifting toward Enterprise-grade AI: Confronting skills and data challenges to realize value.” IBM Institute for Business Value. September 2018. https:// www.ibm.com/thought-leadership/institutebusiness-value/report/enterpriseai Unpublished data. AI1. In general, where is your organization in its adoption of artificial intelligence? Select the most advanced level for your organization; “The business value of AI: Peak performance during the pandemic.” IBM Institute for Business Value. November 2020. https://www.ibm.com/thought-leadership/institute-business-value/report/ai-value-pandemic# Unpublished data. S6. In general, where is your organization overall and your particular function in terms of adoption of artificial intelligence? 2022 Omdia AI Market Maturity survey https://omdia.tech.informa.com/OM023919/AI-Market-Maturity-Survey--2022-Database Q1. What is the state of AI deployment in your company? The rating scale in Omdia survey has been assumed to equivalent to IBM IBV rating scale in the following way: investigating technology and use cases = considering; Identified at least one use case and developing pilot = Evaluating; Currently piloting AI in at least one function or business = Piloting; Live AI deployment in at least one function or business unit = Implementing; Scaling AI deployment across multiple business functions or units = Operating/optimizing.

2 “The business value of AI: Peak performance during the pandemic.” IBM Institute for Business Value. 2020. https://ibm.co/ai-value-pandemic

3 “McCarthy, J; M.L. Minsky; N. Rochester; C.E. Shannon. “A proposal for the Dartmouth summer research project on artificial intelligence.” Accessed on July 13, 2022. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

4 Goodfellow, Ian; Yoshua Bengio, Aaron Corville. “Deep Learning.” The MIT Press. 2016. https://www.deeplearningbook.org

5 LeCun, Yann; Yoshua Bengio; Jeffrey Hinton. “Deep Learning.” Nature. May 28, 2015. https://www.nature.com/articles/nature14539.pdf

6 Burns, Ed. “Timeline of AI winters casts a shadow over today’s applications.” TechTarget. Accessed on July 13, 2022. https://www.techtarget.com/searchenterpriseai/infographic/Timeline-of-AI-winters-casts-a-shadow-over-todays-applications

7 Thompson, Neil C.; Kristjan Greenewald; Keeheon Lee; Gabriel F. Manso. “Deep Learning’s Diminishing Returns.” IEEE Spectrum. September 24, 2021. https://spectrum.ieee.org/deep-learning-computational-cost

8 Ibid.

9 Ibid.

10 World Health Organization malaria fact sheet. April 6, 2022. https://www.who.int/news-room/fact-sheets/detail/malaria

11 “Fast Start in cognitive innovation: Top performers share how they are moving quickly.” IBM Institute for Business Value. January 2017. https://www.ibm.com/blogs/internet-of-things/fast-start-cognitive/ Unpublished data. QC&A10 What are the important value drivers for cognitive computing? Select the top 5.“Shifting toward Enterprise-grade AI: Confronting skills and data challenges to realize value.” IBM Institute for Business Value. September 2018. https://www.ibm.com/thought-leadership/institutebusiness-value/report/enterpriseai Unpublished data. AI2. What are the important value drivers for artificial intelligence/cognitive computing? Select top 5. “The business value of AI: Peak performance during the pandemic.” IBM Institute for Business Value. November 2020. https://www.ibm.com/thought-leadership/institute-business-value/report/ai-value-pandemic# Unpublished data Q8. What are the most important value drivers for artificial intelligence? Select top 5.

12 “The business value of AI: Peak performance during the pandemic.” IBM Institute for Business Value. 2020. https://www.ibm.com/thought-leadership/institute-business-value/report/ai-value-pandemic

13 Payraudeau, Jean-Stéphane; Anthony Marshall; Jacob Dencik. “Unlock the business value of hybrid cloud: How the Virtual Enterprise drives revenue growth and innovation.”IBM Institute for Business Value. 2021. https://ibm.co/hybrid-cloud-business-value. Payraudeau, Jean-Stéphane; Anthony Marshall; Jacob Dencik. “Extending digital acceleration: Unleashing the business value of technology investments.” IBM Institute for Business Value. 2021. https://ibm.co/hybrid-cloud-business-value

14 Fleming, Martin. “Breakthrough: A Growth Revolution.” Business Expert Press. 2022

15 An open source NLP library developed by Facebook AI. https://fasttext.cc

16 XGBoost is a decision-tree-based ensemble ML algorithm that uses a gradient boosting framework. https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

17 McCormick has observed that each 5-10 years of experience halves the number of iterations.

18 The Flavor Extract Manufacturer’s Association of the United States. FEMA numbers refer to ingredients generally recognized as safe and allowed in the United States. https://www.femaflavor.org/

19 VGG16 (also called OxfordNet) is a convolutional neural network architecture named after the Visual Geometry Group from Oxford. https://blog.keras.io/how-convolutional-neural-networks-see-the-world.html

20 Report Ocean press release. “Autonomous Vehicle System Market |(CAGR) of 39.1%| by Product Type, End-User, Application, Region – Global Forecast to 2030.” July 14, 2022. https://www.marketwatch.com/press-release/autonomous-vehicle-systemmarket-cagr-of-391-by-product-type-end-user-application-region-global-forecast-to-2030-2022-07-14?mod=search_headline

21 Kalra, Nidhi and Susan M Paddock. “Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?” Rand Corporation. 2016. https://www.rand.org/content/dam/rand/pubs/research_reports/RR1400/RR1478/RAND_RR1478.pdf

22 GPUs are arranged into purpose-built deep learning systems (for example, the NVIDIA DGX, which comprises 8 GPUs per server).

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America | August 2022

6PQKYZ12-USEN-01