O'REILLY®

Compliments of teradata.

Achieving Real Business Outcomes from Artificial Intelligence

Enterprise Considerations for AI Initiatives

Atif Kureishy, Chad Meley & Ben Mackenzie

Praise for *Achieving Real Business Outcomes from Artificial Intelligence*

"A clear-eyed overview of the past, present, and future of AI in commercial enterprises, with a particular focus on deep learning. Teradata's deep expertise in data and analytics comes to the fore here. A great guide to getting started with AI."

—Thomas H. Davenport, President's Distinguished Professor of IT and Management, Babson College; Fellow, MIT Initiative on the Digital Economy; Senior Advisor to Deloitte's Analytics and Cognitive Practices

"Kureishy, Meley, and Mackenzie have provided concrete, real world examples and case studies that show how AI and machine learning can drive successful outcomes for organizations that are getting started with artificial intelligence. The discussions on challenges and trade-offs will be especially helpful to executives getting started in this exciting area."

—Dave Schubmehl, Research Director, Cognitive/AI Systems and Content Analytics, IDC

"If your company isn't experimenting with AI—or already leveraging it across some disciplines—you are behind your competitors. This book provides a practical framework to understanding AI as a tool positioned to disrupt our data-driven world. It provides great insights on how companies who get AI right use it to predict and meet customer needs."

—Jim Lyski, Chief Marketing Officer at CarMax

"A fantastic list of use cases for prediction machines in practice."

—Avi Goldfarb, Professor at University of Toronto and author of Prediction Machines: The Simple Economics of Artificial Intelligence

"An insightful discussion of AI for the executive, with real examples and practical advice. This book helps you understand why AI is so critical now and how to get started. A quick read you can't afford to miss!"

—Richard Winter, CEO of WinterCorp

"This book cuts through the AI hype, clearly differentiates machine learning and deep learning techniques, and focuses on practical, real-world use cases. It's a must-read for anyone focused on getting to better business outcomes."

—Doug Henschen, Vice President and Principal Analyst, Constellation Research

"In an industry where a significant understanding gap exists between the technology and business, this book provides an easily accessible overview for executive leadership seeking to understand how deep learning can positively augment their enterprise."

—BJ Yurkovich, Principle Investigator, Center for Automotive Research, The Ohio State University

"A useful guide to help executives understand the promise of AI, with concrete examples of how it is being applied now in business, that will leave you with an urge to get started."

—Mike Janes, Former GM of Worldwide Apple Store and CMO at StubHub

"This book provides valuable insight for digital transformation leaders on the impact that AI is having on an organization's strategy, technology, data, and talent."

—Robertino Mera, Senior Director of Epidemiology, Gilead Sciences

Achieving Real Business Outcomes from Artificial Intelligence

Enterprise Considerations for AI Initiatives

Atif Kureishy, Chad Meley, and Ben Mackenzie

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Achieving Real Business Outcomes from Artificial Intelligence
by Atif Kureishy, Chad Meley, and Ben Mackenzie

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://oreilly.com/safari*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Acquisitions Editor: Rachel Roumeliotis
Development Editor: Jeff Bleiel
Production Editor: Justin Billing
Copyeditor: Octal Publishing, Inc.
Proofreader: Sharon Wilkey
Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Court Patton

October 2018: First Edition

Revision History for the First Edition
2018-10-02: First Release

This work is part of a collaboration between O'Reilly and Teradata. See our *statement of editorial independence*.

978-1-492-03818-4

[LSI]

Table of Contents

Foreword vii

Acknowledgments xi

1. Artificial Intelligence and Our World 1
A New Age of Computation 2
The AI Trinity: Data, Hardware, and Algorithms 2
What Is AI: Deep Versus Machine Learning 4
What Is Deep Learning? 5
Why It Matters 7

2. More Than Games and Moonshots 9
AI-First Strategy 9
Where Deep Learning Excels 10
Financial Crimes 11
Manufacturing Performance Optimization 11
Recommendation Engines 12
Yield Optimization 13
Predictive Maintenance 13

3. Options and Trade-Offs for Enterprises to Consume Artificial Intelligence 17
SaaS Solutions: Quick but Limited 17
Cloud AI APIs 18
Building Custom AI Algorithms 19

4. **Challenges to Delivering Value from Custom AI Development and Engineering Countermeasures** 21
Strategy 22
Technology 23
Operations 24
Data 28
Talent 29
Conclusion 30

5. **Artificial Intelligence Case Studies** 31
Fighting Fraud by Using Deep Learning 31
Mining Image Data to Increase Productivity 33
Deep Learning for Image Recognition 34
Natural-Language Processing for Customer Service 34
Deep Learning for Document Automation 35
Conclusion 36

6. **Danske Bank Case Study Details** 37
The Project, the Tools, and the Team 38
Getting the Right Data in Place 39
Ensemble Modeling and Champion/Challenger 40
Working with the Past, Building the Future 40
Moving the ML Models into Live Production 41
From Machine Learning to Deep Learning 41
Visualizing Fraud 42
Visualizing and Interpreting Deep Learning Models 44
A Platform for the Future 45

7. **Predictions Through 2020** 47
Strategy 48
Technology 49
Operations 50
Data 52
Talent 53
What's Next 53

8. **Conclusion** 55
Identify High-Impact Business Outcomes 55
Assess Current Capabilities 56
Build Out Capabilities 56

Foreword

Enterprises are under the impression that they're on their way to using artificial intelligence. They've set up a few machine learning models and have had new algorithms work their way into previously deployed Software as a Service applications. Inside the organization, it feels like they're checking all the right artificial intelligence (AI) boxes.

But the true end goal of AI in the enterprise is something much more sophisticated. Oliver Ratzesberger and Mohanbir Sawhney expressed it succinctly in their book, *The Sentient Enterprise*, noting, "Our objective is to position the enterprise in such a way that analytic algorithms are navigating circumstances and making the bulk of operational decisions without human help."

With the exception of a few Bay Area tech giants, the industry hasn't experienced highly proficient natural-language processing, image-based detection, or other skills that would enable this next generation of AI to drive significant business outcomes instead of just performing basic business tasks.

Imagine if AI platforms could identify and bring together data sources and then explain to their human counterparts the "why" behind the recommendations—something like AI for data engineering and data science. Or, imagine if chatbots could interpret problems and provide solutions using natural language that satisfy buyers more quickly and more effectively than current call centers. Imagine if key business functions were being driven by algorithms with the necessary autonomy to self-learn and change tactics at a level of speed and accuracy that far surpasses any human or team of humans.

These scenarios will one day be mainstream, but how are companies going to get there?

One of the biggest challenges for AI in the enterprise is that each company—even within the same industry—has unique problems. So, for the most part, businesses today need custom AI solutions to drive specific value.

However, the reality for most companies is that homegrown, custom AI solutions aren't feasible for a number of reasons. Not only is it an expensive initiative to take on, but AI development also has a very small talent pool, and it would be difficult to get that kind of brain trust in one organization at an affordable, sustainable rate. The information and opportunity for AI development, however, is out there. To truly accelerate AI, companies should work with partners that have created custom AI solutions before, enabling them to share a vision for how AI will drive business outcomes.

AI is not going to be easy. There is no out-of-the-box AI solution that will transform a company overnight. Instead, building a custom AI solution will take persistent, coordinated effort and deep organizational change. These investments will be necessary not only to develop AI capabilities; they will be necessary for companies to survive.

This book is a thoughtful primer for digital transformation leaders in large enterprises seeking to outpace their competition by embracing the technological and organizational change that comes with AI. In it, the authors review potential enterprise AI use cases and discuss authentic case studies in which companies have realized value from custom AI solutions. For those readers looking for a higher level of engineering detail, the authors include a technical dive into a deep learning solution implemented at Danske Bank.

You will gain insight into the very real challenges that organizations will face as they make this difficult but necessary transition, and various measures that you can implement to approach those challenges. Finally, the book includes a look toward the next several years of AI innovation to give a preview of what organizations can expect to see.

Ultimately, this book provides a practical roadmap for understanding how an enterprise can begin to approach using artificial intelligence to harness its most powerful asset: data.

— Diego Klabjan
Professor, Director of Master of Science in Analytics
Northwestern University

Professor Klabjan's research focuses on developing models and algorithms for machine learning, in particular deep learning, problems with emphasis on the industries of healthcare, marketing, sports, and finance. He has led data science projects with large companies such as Intel, Allstate, HSBC, The Chicago Mercantile Exchange, FedEx, General Motors, and many others.

Acknowledgments

This book incorporates perspectives from a diverse group of people to whom we would like to express our deep gratitude.

Our Customers

We have had the privilege to collaborate with some of the leading data-driven companies across various industries as they take the next step in their analytic evolution, and many have allowed us to share their stories here by name or anonymously. As this is a new field, we have all learned together, and we are grateful for that opportunity.

Our Talented Data Scientists

Data scientists with enterprise-tested deep learning credentials are scarce indeed, and we have been fortunate to call a number of them our colleagues. Their contributions to this book were invaluable in properly documenting the challenges and countermeasures enterprises face when taking on an artificial intelligence (AI) initiative. In particular, we'd like to offer a special thanks to Eliano Marques, Sune Askjaer, Chanchal Chatterjee, Peter Mackenzie, David Mueller, Frank Saeuberlich, and Yasmeen Ahmad.

Industry Analysts

We are voracious consumers of the insights provided by analysts covering the intersection of AI and Enterprise Digital Transformation—notably, those of IDC, Gartner, and Forrester. Their research, responses to our inquiries, and participation at our events has helped shape our point of view on AI market trends. Thank you for sharing your insights with us.

Our Community

Much inspiration has been drawn from associating with NVIDIA and our other industry partners as well as organizations that share our passion for moving the AI industry forward, including Neural Information Processing Systems (NIPS), Stanford DAWN, O'Reilly AI Conference attendees, and various professional meetups and hackathons in North America, Europe, and Asia. Thank you for your enthusiasm, passion, and dedication to the AI industry.

CHAPTER 1

Artificial Intelligence and Our World

Recent advances in artificial intelligence (AI) have sparked increased interest in what the technology can accomplish. It's true that there's hype—as there often is with any breakthrough technology—but AI is not just another fad. Computers have been evolving steadily from machines that follow instructions to ones that can learn from experience in the form of data. Lots and lots of it.

The accomplishments are already impressive and span a variety of fields. Google DeepMind has been used to master the game of Go; autonomous vehicles are detecting and reacting to pedestrians, road signs, and lanes; and computer scientists at Stanford have created an artificially intelligent diagnosis algorithm that is just as accurate as dermatologists in identifying skin cancer. AI has even become commonplace in consumer products and can be easily recognized in our virtual assistants (e.g., Siri and Alexa) that understand and respond to human language in real time.

These are just a few examples of how AI is affecting our world right now. Many effects of AI are unknown, but one thing is becoming clearer: the adoption of AI technology—or lack of it—is going to define the future of the enterprise.

A New Age of Computation

AI is transforming the analytical landscape, yet it has also been around for decades in varying degrees of maturity. Modern AI, however, refers to systems that behave with intelligence without being explicitly programmed. These systems learn to identify and classify input patterns, make and act on probabilistic predictions, and operate without explicit rules or supervision. For example, online retailers are generating increasingly relevant product recommendations by taking note of previous selections to determine similarities and interests. The more users engage, the smarter the algorithm grows, and the more targeted its recommendations become.

In most current implementations, AI relies on deep neural networks (DNNs) as a critical part of the solution. These algorithms are at the heart of today's AI resurgence. DNNs allow more complex problems to be tackled, and others to be solved with higher accuracy and less cumbersome, manual fine tuning.

The AI Trinity: Data, Hardware, and Algorithms

The story of AI can be told in three parts: the data deluge, improvements in hardware, and algorithmic breakthroughs (Figure 1-1).

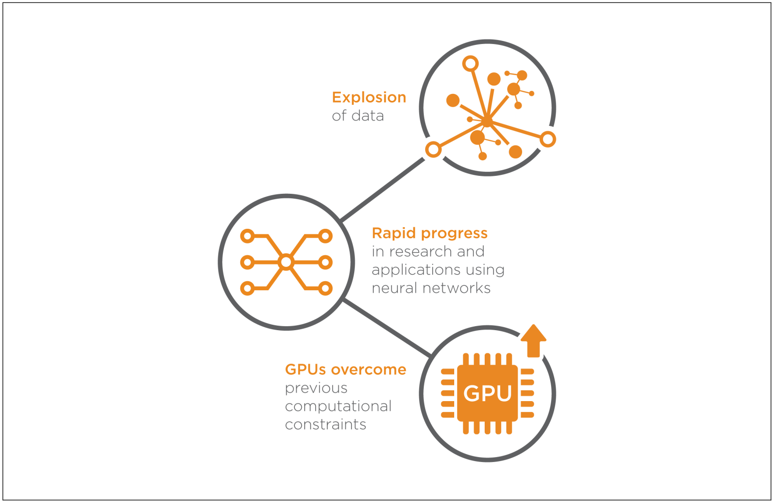

Figure 1-1. Drivers of the AI Renaissance

Exponential Growth of Data

There is no need to recite the statistics of the data explosion of the past 10 years; that has been mainstream knowledge for some time. Suffice it to say that the age of "big data" is one of the most well understood and well documented drivers of the AI renaissance.

Before the current decade, algorithms had access to a limited amount and restricted types of data, but this has changed. Now, machine intelligence can learn from a growing number of information sources, accessing the essential data it needs to fuel and improve its algorithms.

Computational Advances to Handle Big Data

Advanced system architectures, in-memory storage, and new AI-specific chipsets in the form of graphics processing units (GPUs) are now available, advances that overcome previous computational constraints to advancing AI.

GPUs have been around in the gaming and computer-aided design (CAD) world since 1999, when they were originally developed to manipulate computer graphics and process images. They have recently been applied to the field of AI when it was found that they were a perfect fit for the large-scale matrix operations and linear algebra that form the basis of deep learning. Although parallel computing has been around for decades, GPUs excel at parallelizing the same instructions when applied to multiple data points.

NVIDIA has cemented itself as the leader in AI accelerated platforms, with a steady release of ever powerful GPUs, along with a well executed vision for CUDA, an application programming interface (API) that makes it easier to program GPUs without the need for advanced graphic programming skills. Leading cloud vendors like Google, Amazon, and Microsoft have all introduced GPU hardware into their cloud offerings, making the hardware more accessible.

Accessing and Developing Algorithms

Whereas AI software development tools once required large capital investments, they are now relatively inexpensive or even free. The most popular AI framework is TensorFlow, a software library for machine learning originally developed by Google that has since been open sourced. As such, this world-class research is completely free for anyone to download and use. Other popular open source frameworks include MXNET, PyTorch, Caffe, and CNTK.

Leading cloud vendors have packaged AI solutions that are delivered through APIs, further increasing AI's availability. For instance, AWS has a service for image recognition and text-to-speech, and Google has prediction APIs for services such as spam detection and sentiment analysis.

Now that the technology is increasingly available with hardware to support it and a growing body of practice, AI is spreading beyond the world of academia and the digital giants. It is now on the cusp of going mainstream in the enterprise.

What Is AI: Deep Versus Machine Learning

Before venturing further into talking about AI, it will be helpful to discuss what is meant by the terms in this book.

The term *artificial intelligence* has many definitions. Of these, many revolve around the concept of an algorithm that can improve itself, or learn, based on data. This is, in fact, the biggest difference between AI and other forms of software. AI technologies are ever moving toward implicit programming, where computers learn on their own, as opposed to explicit programming, where humans tell computers what to do.

Several technologies have—at various points—been classified under the AI umbrella, including statistics, machine learning, and deep learning. Statistics and data mining have been present in the enterprise for decades and need little introduction. They are helpful for making simple business calculations (e.g., average revenue per user). More advanced algorithms are also available, drawing on calculus and probability theory to make predictions (e.g., sales forecasting or detecting fraudulent transactions).

Machine learning makes predictions by using software to learn from past experiences instead of following explicitly programmed instructions. Machine learning is closely related to (and often overlaps with) statistics, given that both focus on prediction making and use many of the same algorithms, such as logistic regression and decision trees. The key difference is the ability of machine learning models to learn, which means that more data equals better models.

This book, however, focuses on *deep learning*, which is at the heart of today's AI resurgence as recent breakthroughs in the field have fueled renewed interest in what AI can help enterprises achieve. In fact, the terms *AI* and *deep learning* are used synonymously (for reasons we discuss in a moment). Let's delve into that architecture.

What Is Deep Learning?

Building on the advances of machine learning, deep learning detects patterns by using artificial neural networks that contain multiple layers. The middle layers are known as *hidden layers*, and they enable automatic feature extraction from the data—something that was impossible with machine learning—with each successive layer using the output from the previous layer as input. Figure 1-2 briefly summarizes these advancesments over time.

The biggest advances in deep learning have been in the number of layers and the complexity of the calculations a network can process. Although early commercially available neural networks had only between 5 and 10 layers, a state-of-the-art deep neural network can handle significantly more, allowing the network to solve more complex problems and increase predictive accuracy. For example, Google's speech recognition software improved from a 23% error rate in 2013 to a 4.9% error rate in 2017, largely by processing more hidden layers.

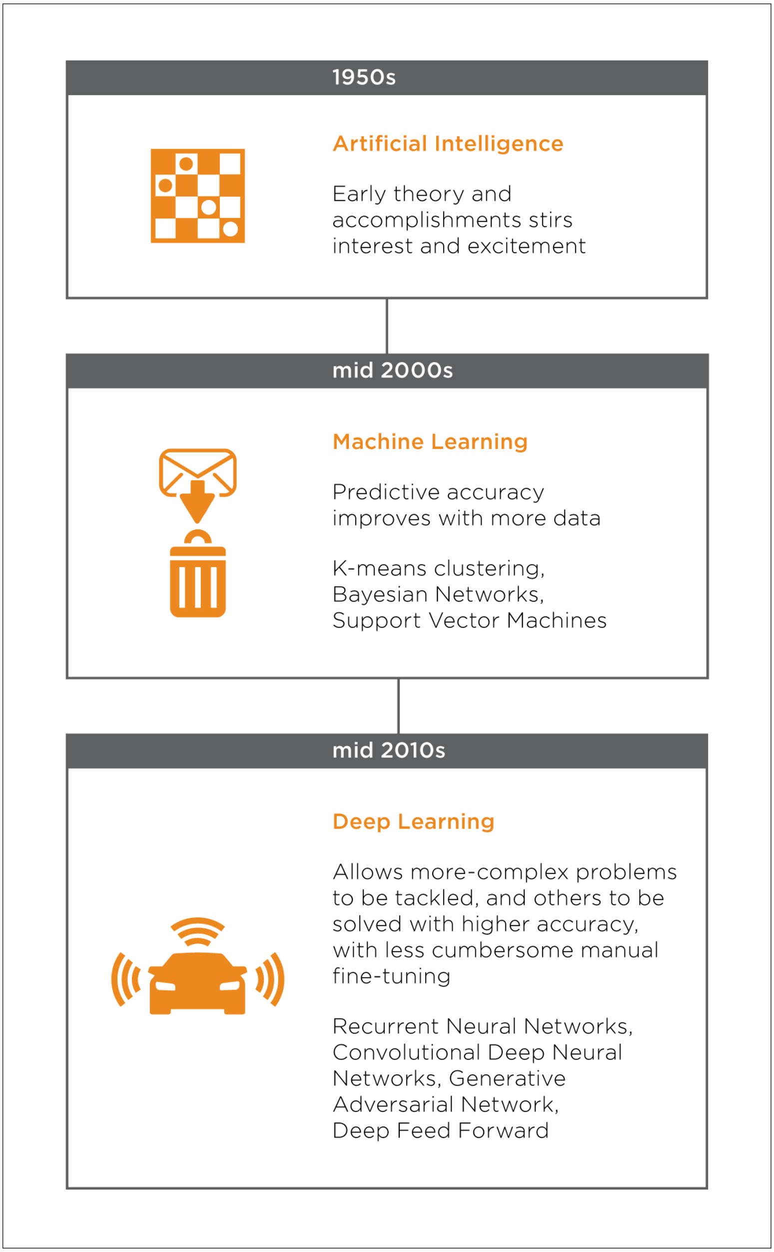

Figure 1-2. Evolution of AI

Why It Matters

Because of its architecture, deep learning excels at dealing with high degrees of complexity, forms, and volumes of data. It can understand, learn, predict, and adapt, autonomously improving itself over time. It is so good at this that in some contexts, deep learning has become synonymous with AI itself. This is how we will be using the term here.

Here are some differentiators of deep learning:

- Deep learning models allow relationships between raw features to be determined automatically, reducing the need for feature engineering and data preprocessing. This is particularly true in computer vision and natural language–related domains.
- Deep learning models tend to generalize more readily and are more robust in the presence of noise. Put another way, deep learning models can adapt to unique problems and are less affected by messy or extraneous data.
- In many cases, deep learning delivers higher accuracy than other techniques for problems, particularly those that involve complex data fusion, when data from a variety of sources must be used to address a problem from multiple angles.

In Chapter 2, we will examine how this technology is changing the enterprise.

CHAPTER 2

More Than Games and Moonshots

Although it might be easy to dismiss artificial intelligence (AI) use cases highlighted by the media as "moonshots" (e.g., curing cancer), publicity stunts (e.g., beating the best human players at Go and Jeopardy), too industry-specific (e.g., autonomous driving), or edge use cases and point solutions (e.g., spam filtering), we can apply deep learning to core strategic initiatives across many verticals. In fact, AI has already begun to demonstrate its value in large enterprises, even outside Silicon Valley and West Coast digital giants. Fortune 500 companies in industries like banking, transportation, manufacturing, retail, and telecommunications have also begun to take advantage of its power.

AI-First Strategy

In an AI-first strategy, AI operates at the core of a company, driving its product and decision-making. In several industries, new challengers are using this kind of strategy to successfully compete against incumbents. One example is in the financial industry, with companies like Citadel, Two Sigma, and Personal Capital maximizing return and reducing risk by creating the best machine intelligence. And in the automotive industry, it's now easy to envision a day when people will decide to buy a new car based on its driving software and not on its engine, body design, or other buying criteria.

Across all industries, the use of deep learning has the potential to increase production, drive down cost, reduce waste, and improve efficiency, as well as push innovation. And, as can be seen in industry leaders like Google, Apple, and Amazon, machine intelligence changes everything and becomes pervasive when an organization pieces together how to use it.

Though we cannot predict its full impact, it's clear that AI represents profound change both in the short and long term, and it is a technology that demands strategic focus and action.

Where Deep Learning Excels

Deep learning is not the best approach for every problem, and more basic tools such as stats and machine learning aren't going away anytime soon. But deep learning is extremely powerful when we have access to plenty of training data, we have many dimensions or features of the data (which would require time-consuming feature extraction in order to conduct machine learning), or we need to process rich media such as images, video, and audio.

In the short term, the techology holds significant promise for dealing with problems like fraud detection, predictive maintenance, recommendation engines, yield optimization, and churn reduction. In these areas, it could produce order-of-magnitude improvements in two ways:

In some cases, deep neural networks will yield better predictability over current models even when using the same dataset
: Later in the book, we discuss how deep learning models were able to detect fraud much more predictably than machine learning ones at Danske Bank, even using the same data as prior-generation models.

Deep learning can allow the enterprise to analyze previously intractable datasets
: For example, companies could use images and audio files for predictive maintenance, as in mining photos of a piston in an engine to spot cracks and other imperfections before they become more serious or using audio from wheel bearings in a train to listen for anomalies that signal a potential derailment.

Let's take a look at some use cases for which deep learning could provide a significant advantage over current prediction methods.

Financial Crimes

Financial crimes cost institutions, consumers, and merchants billions of dollars every year. In the past, fraud was more difficult to perpetrate because banking was personal and channels for crime were more constrained. The internet has changed that. Modern banking is almost completely anonymous and occurs through many avenues. This has enabled a new ecosystem of many kinds of fraud, which are growing increasingly sophisticated and aggressive. Both industries and governments face unprecedented threats from a variety of actors, risking physical loss of money, intellectual property theft, and damage to their reputations.

Financial institutions have long been using machine learning, data mining, and statistics to mitigate risk, and these have certainly provided value. Today's risk landscape, however, demands new tools.

Though deep learning didn't initiate real-time or cross-channel data analysis, it is better at detecting more accurate patterns across all data streams, in addition to its ability to analyze new types of data. Because of this, AI can empower banks and other institutions with insight that keeps up with the pace of modern fraud.

Manufacturing Performance Optimization

Currently, the manufacturing industry suffers from inefficiencies due to "siloed" data and delayed communication of insights across the supply chain, from the acquisition of raw goods through production and sales. Improving this efficiency represents a huge opportunity for manufacturers.

Iterating on manufacturing processes is nothing new—it's something that has been done for decades. However, AI can permit iterations and adjustments to systems in minutes instead of months.

The increased predictive power of AI enables companies to proactively understand their needs and intelligently communicate them across their different branches. This can have a huge impact on every part of the business. According to data from General Electric, smart manufacturing systems using AI can increase production capacity up to 20% while lowering material consumption by 4%.

AI provides data to the business in real time, which can help optimize the supply chain, provide greater economies of scale, and better manage factory and demand-side constraints. GE, for example, saw finished goods buffers reduced to 30% or more by using a smart manufacturing system.

There are many other manufacturing use cases for AI, such as intelligent pricing, ensuring regulatory compliance, improving eco-sustainability, and finding new revenue streams. As the technology develops, more use cases are sure to be discovered.

With its abundance of sensor data and systems that richly reward increased efficiency, manufacturing is an industry poised to be revolutionized by machine intelligence.

Recommendation Engines

Whereas companies used to get to know their customers' buying habits and preferences through face-to-face interactions and relationships built over time within the four walls of a store, they must now infer this same data through online activity. This has made *recommendation engines* essential for many businesses to compete in the online marketplace. By helping customers discover items or content quickly, recommenders increase satisfaction, expenditures, and lifetime value.

Even though recommendation engines predate deep learning, their effectiveness continues to grow for retailers who have made the switch from legacy engines driven primarily by collaborative filtering to ones based on wide and deep learning. For example, Amazon reported that an impressive 35% of sales are a result of their recommendation engine, and 75% of content watched on Netflix comes from such algorithms.

We can use deep learning algorithms at several points for building a recommender. They allow radical personalization, enabling each person to see items particular to their interests and actions, thereby solving the difficult problem of how to show one out of thousands of customers the right product out of thousands of options.

Deep learning can also find unique connections across items that might not be intuitive, such as showing other baby products when someone is searching for children's books, or showing a user novel items, creating the feeling of serendipity. When done well, recommenders essentially scale the record store clerk or the friendly sales associate, helping customers find both what they want and what they didn't know they wanted.

Yield Optimization

A manufacturer must consider many variables when determining how much product to produce. These can include supplier, customer, and team requirements as well as equipment availability and capacity. Unfortunately, factories often perform at less-than-optimal rates due to imperfections and lag-time in communication between devices and management. This includes data streams that are disconnected from one another and variables that often change but require manual processes to take their new values into account.

Manufacturers across industries like aerospace, high-tech, and industrial equipment have been using AI to improve communication across their devices and are seeing gains in yield, creating more efficiency and using more of their production capacity.

Yield optimization presents a huge opportunity because even incremental changes in efficiency can create significant value. For example, leaders at Micron found that each 1% cumulative yield improvement translated to $100 million in annual cost reduction. They were then able to use AI with sensor data to determine the top factors negatively affecting yield, ultimately leading to significant financial and operational improvements.

Predictive Maintenance

According to Statista, *predictive maintenance* is one of the top-ten most valuable use cases for AI, with the potential to generate $1.3 trillion by 2025. This makes sense, as unplanned downtime is extremely expensive, costing manufacturers billions of dollars every year. In the automotive industry alone, each minute of unplanned downtime costs $22,000.

Predictive maintenance with AI is a key part of factory automation. It helps reduce capital expenditure, extends the lifetime value of equipment, and improves safety. With machine intelligence, manufacturers can more efficiently integrate and analyze across all data sources, improving the performance of maintenance, repair, and overhaul (MRO), as illustrated in Figure 2-1. It also manufacturers them greater insight into each component and part, allowing potential points of failure to be identified and repaired before they become a problem.

Figure 2-1. Business outcomes enabled by Industrial IoT predictive maintenance

Traditionally, manufacturers were predicting downtime by building algorithms fed on data sources like maintenance records, parts inventory, and warranties (small data), as well as big data like sensor streams coming from a jet aircraft engine or an MRI machine. These models functioned fairly well, and these algorithms were able to use that data to make much more accurate predictions than models that did not.

[Figure 2-2](#) illustrates how deep learning improves on that technology by enabling the use of new types of data like audio and video that can be incorporated with traditional data sources to enhance prediction capabilities. In essence, deep learning scales the task of someone taking a look at hoses to check for cracks or listening for strange sounds on the shop floor.

Figure 2-2. Deep learning augments traditional data sources and analytic techniques

We have not yet come close to tapping the full potential of deep learning. We will discover more use cases as the technology continues to develop and is implemented across all industries.

In Chapter 3, we discuss ways that enterprises are currently able to consume AI and build deep learning capabilities.

CHAPTER 3

Options and Trade-Offs for Enterprises to Consume Artificial Intelligence

In this chapter, we examine the options available to companies that want to use artificial intelligence (AI) capabilities, which at this time include the following:

- Purchasing Software as a Service (SaaS) solutions
- Using public cloud–based APIs
- Developing custom AI algorithms

SaaS Solutions: Quick but Limited

Perhaps the simplest option for deploying AI within your organization is by taking advantage of SaaS analytics. These are prepackaged, turnkey solutions that are typically in the visual, assistive, or operations space. AI is also being deployed as a feature within existing SaaS offerings such as Customer Relationship Management (CRM) applications.

One example is Everseen, a company that uses deep learning to mine video footage of point-of-sale transactions to detect irregularities. Affectiva operates similarly, deploying advanced video- and audio-mining algorithms to detect emotional patterns. Salesforce

Einstein uses AI models to improve on prior-generation models for things like lead prioritization and personalization.

These platforms address real pain points and opportunities and, as such, can certainly create value for organizations, with the broader market for the platform reducing software development costs and driving feature innovation. However, they do have their downsides.

SaaS solutions tend to be a commodity. By design, they have limited configuration and are constricted to the data exposed to the application. Although that makes them easier, cheaper, and less risky to plug in, they also tend to be disconnected from larger business processes.

Enterprises that deploy these solutions don't need to bother with creating complex algorithms, but this also means that they don't own any unique IP associated with the algorithm. As a point solution available to a broad market, SaaS solutions do not create competitive advantage.

Cloud AI APIs

With AI APIs in the cloud, developers don't need to understand AI technology to benefit from powerful deep learning capabilities. These APIs offer easy-to-use services such as computer vision, speech, and language understanding. Training sets for these kinds of algorithms are widely available, and the cloud vendors have the talent and economies of scale to address these common use cases.

All of the leading cloud vendors (e.g., Microsoft, Amazon, and Google) now offer these kinds of API-based services. Like SaaS solutions, deploying cloud APIs requires no hardware installation and minimal or no AI expertise. Their pay-as-you-go pricing model also makes them relatively low risk.

However, they are trained on publicly available data, not on an enterprise's specific datasets. For instance, a vision API can easily spot a wide range of features in a picture, like wedding dresses, mustaches, celebrities, cats, or swimming pools, but it would be incapable of detecting information that is uniquely valuable to your company, such as hairline cracks in a jet engine component your company produces or operates. So, although cloud AI APIs are valuable in performing common use cases informed by public data, they are not equipped to address insights unique to enterprises such as customer intimacy, operational efficiencies, and risk mitigation.

That said, the reality for large enterprises is that developers will use AI APIs in conjunction with tailored and unique enterprise data and algorithms. For example, a developer could use a voice-recognition API to translate spoken words from a customer care call into text and then combine that outside the cloud API model with other data and analytics unique to their enterprise to achieve improved accuracy in churn models.

Building Custom AI Algorithms

The third option for enterprises to consume AI is creating custom algorithms using frameworks, which would allow organizations more flexibility and agility in utilizing AI. A variety of options are available, including popular open source frameworks like TensorFlow, Keras, and PyTorch, and propriety options like Watson. These frameworks are fairly mature with a high potential for insight.

The development of a custom AI solution allows you to use all of the data within your enterprise and fully integrate it within your processes, tailoring the solution to your unique problems. Talented data scientists have access to many taxonomies, providing opportunities to produce greater predictive outcomes when compared to prebuilt solutions or competitors' algorithms.

Designing and training deep learning algorithms on enterprise-specific data is the only way companies can create true competitive advantage with AI. It is this type of AI that we will see grow more and more prevalent over the next several years as organizations begin implementing it and seeing its benefits. Its impact can be transformative, allowing you to build new lines of business and radically increase efficiency. In contrast to the other options, however, developing custom AI algorithms requires significant AI expertise, and presents many challenges to deploy or operationalize at scale. Figure 3-1 summarizes the advantages and disadvantages of the three solutions presented in this chapter.

	SaaS Analytics	AI APIs	Custom AI Development
Models			
	Commodity, limited configuration	One size fits all	Tailored & unique
Data			
	Least common denominator	Not informed by enterprise data	Enterprise data
Deployment			
	Isolated SaaS	Dependent PaaS	Any (Hybrid)
AI Expertise			
	None	Minimal	Significant
Implementation			
	Low cost & low risk	Moderate cost & moderate risk	High cost & high risk
Competitive Advantage			
	Parity "me too"	Tactical	Strategic & game changing

Figure 3-1. Options for enterprise consumption of AI

In Chapter 4, we examine these challenges in detail.

CHAPTER 4

Challenges to Delivering Value from Custom AI Development and Engineering Countermeasures

Deep learning can have a profound impact on the enterprise, but developing and implementing it is not easy, and organizations will face unique challenges that are unlike those that accompany the adoption of other technologies.

Despite amazing breakthroughs in artificial intelligence (AI) software and hardware, organizations must confront the poor interoperability of open source software components and the need to optimize highly specialized hardware, not to mention the challenge of first accessing and then harnessing both high-value and high-velocity data, working across multiple cloud environments, and doing all of this at scale. Further, deep learning methods are a radical departure from traditional statistical and machine learning techniques. As such, they can challenge even advanced data-driven organizations.

In terms of operationalizing, most organizations struggle in the transition from insight to action because of analytical systems that are incapable of reliably serving millions of decisions at the speed of real-time business. Many will also underestimate or discount the governance and risk management aspects of developing AI solutions, elements that must be considered for a successful strategy.

Figure 4-1 summarizes the barriers to AI adoption discussed in this chapter. We will also address some approaches that we can take to address them.

Figure 4-1. Five pillars of enterprise considerations for AI initiatives

Strategy

AI capabilities are not just about data and models. As a cross-functional technology, we will need to implement and socialize AI across the organization, an effort that will require coordination with many areas like legal, line-of-business, security, and compliance and regulation. Though developing an AI solution will be a significant technological challenge, its successful operationalization will ultimately be a human one requiring skillful change management and a deep understanding of your organization's operational processes.

One of the most crucial steps of a successful AI initiative is achieving executive buy-in and support. Here, organizations can encounter a variety of difficulties. The great deal of hype surrounding AI can lead to inflated expectations of what can be achieved, sabotaging a program's success. Even when it isn't overhyped, the value that AI can bring to an enterprise can be poorly understood, leading to overt skepticism and inaction. In addition, you might need to confront indifference, open resistance to change, or the tendency for the organization to focus on short-term emergencies instead of long-term change. Furthermore, even though it will be necessary to express the AI solution in terms of its business value, it can be difficult to communicate its potential and path to production accurately across different kinds of stakeholders in your organization. The resulting lack of a shared AI vision can make it difficult to execute anything beyond an AI science experiment or point solution.

Executive education—at an appropriate level of complexity—will be key to set those expectations and determine the right solution to build. A shared AI vision comes from gaining a mutual understanding across key stakeholders of a series of achievable AI applications that are aligned to the strategic goals of the enterprise. Peer benchmarking and collaboration with experienced partners are particularly helpful in uncovering applications that are currently achievable.

AI is advancing rapidly. As such, your strategy should not only contemplate business outcomes, but also ensure that the decisions that are made lead to a robust and future-proof platform. A good starting place is to conduct an independent review of your current capabilities, identify the most important gaps, and develop a high-level roadmap to address them.

To spark and sustain strategic momentum, it is highly recommended to demonstrate proof of value with AI using an Agile, iterative process to find new, actionable insights in your data. This incremental approach will give stakeholders the chance to experience the AI solution's power and become its advocates as the technology proves its value.

Technology

Working with deep learning means working with open source code, and although there's no doubt that AI owes much of its advancement to open source software, there are some drawbacks.

Much of the cutting-edge research and code is academic grade and must be retooled before being used in the enterprise. The field is still relatively immature, and the code lacks many of the features that enterprises demand (e.g., security and usability).

The fragmentation and lack of interoperability between components can be difficult to navigate, especially because of significant gaps in support. Many vendors have taken a lock-in model, offering support only to those running code in their respective clouds. This is at odds with the hybrid model of cloud and on-premises computing that most enterprises prefer.

Deep learning also requires specialized hardware, and any data science team working with it must know how to manage and share a cluster of graphics processing units (GPUs). Working within these GPU compute frameworks can be challenging because their architectures are much different from those that are CPU-only. Significant engineering is required to optimize the software to ensure efficient parallelism, manageability, reliability, and portability. And because not all learning will occur solely in GPU architectures, they must also integrate with the rest of the analytical ecosystem.

Operations

Whereas many companies are comfortable with analytics in batch processing, deploying neural nets at scale is a completely different type of data analysis. Many companies lack the infrastructure required to use their data in a way that can service fully scaled and productionized deep learning models. In fact, this was cited as one of the top barriers to AI in the enterprise according to a 2017 survey of 260 IT and business decision-makers at a VP level or higher from organizations with a global revenue of more than $50 million per year. Almost all respondents (91%) expect to see barriers to AI realization when trying to implement it across their business. Figure 4-2 summarizes these anticipated barriers with lack of IT infrastructure and access to talent leading the challenges.

Building modern capabilities on top of or alongside legacy systems can be difficult, expensive, and tedious. It can present challenges like maintaining service-level agreements (SLAs) while moving data in and out of a mainframe and through a GPU or integrating data from customer relationship management (CRM), enterprise resource planning (ERP), or other enterprise software. Depending on the industry, there could be additional challenges such as lack of interoperability between diverse sets of equipment, each with its own proprietary control system and data interchange standard.

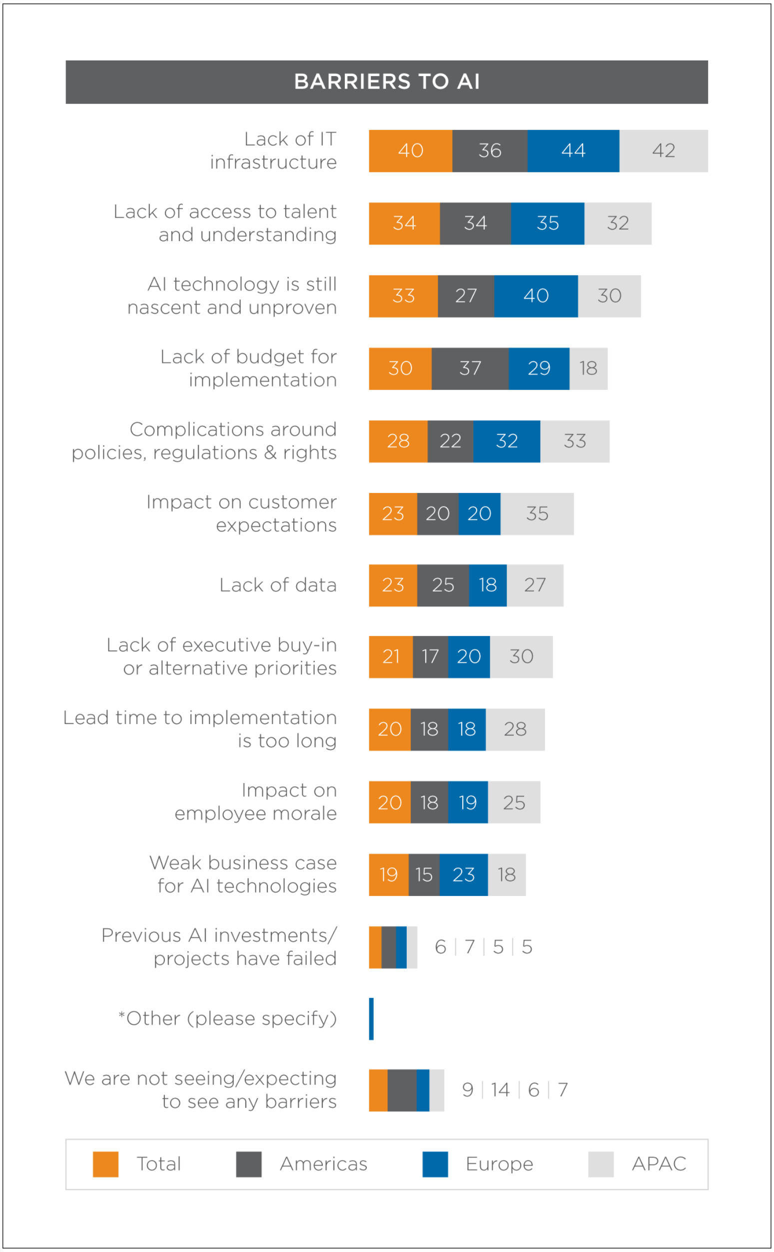

Figure 4-2. Barriers to enterprise AI realization

After you complete the tricky work of scaling the models over multiple servers, datacenters, and GPUs, they must also be managed, monitored, and automatically retrained.

AnalyticOps

Training models that work well in the lab can—in some ways—bear little resemblance to a production environment where they are relied on to make decisions. To maintain, manage, and monitor them, you will need to develop a culture and capability of *AnalyticOps*.

AnalyticOps is a system of best practices that has risen as a response to the demands of supporting automated decision-making. It applies the principles of DevOps and IT operations to data science so that companies can better manage their analytics resources and projects.

AnalyticOps provides a fixed framework for operationalizing models, helping automate your processes to ensure that you don't have more work in progress than you need, that you're monitoring the work you do have, and that people are adhering to your processes.

It helps organizations to be smart about their data science resources, putting applied analytics in the hands of IT operations so that data scientists need to be called in only when something unusual happens. Though creating AnalyticOps might seem like overkill at the start of an AI project, it will be essential for scaling it.

Model Transparency

Model interpretability is another key challenge because deep learning techniques make it more difficult to explain how the model arrived at a conclusion—the information they use to make decisions is hidden away, quite literally in *hidden layers*. Whereas the machines' output might occasionally seem self-explanatory, such as when algorithms correctly identify images, in other cases it might be completely opaque.

The ability to explain these models is often an imperative, however, especially in cases for which laws prevent decisions based on data that can be considered discriminatory (e.g., approving or denying a loan) or cases that involve significant exposure to litigation (e.g., medical diagnosis).

For example, financial institutions in Europe must remain in compliance with the EU's recently enacted General Data Protection Regulation (GDPR), which levies heavy financial penalties on companies that cannot explain how a customer's data was used. In this case, it is not possible—nor is it legal—to tell a customer that their financial transaction was declined simply because the model said so. Even aside from regulation, leaders often need to know the factors that went into a model's decision in order to trust it.

Though the problem of model interoperability is far from solved, enterprises are using several approaches to address it. One approach uses a method called Local Interpretable Model-Agnostic Explanations (LIME), an open source body of research produced at the University of Washington. LIME sheds light on the specific variables that triggered the algorithm at the point of its decision and produces that information in a human-readable way. In the case of fraud, knowing this information can provide security from a regulatory standpoint as well as help the business understand how and why fraud is happening.

The Move to Autonomous Decisions

After the models are built, making AI decisions operational at scale is one of the most significant challenges that companies will face as they adopt deep learning. AI is not like plugging in another app, and it is unlike current forecasting models, in which a report is delivered to a human who then makes a decision. Even for cases in which decisions are currently algorithmically driven, it is a daunting leap between trusting static models and trusting those that learn. This is automated decision-making, and it is different. It can disrupt your enterprise's processes and mean that you need to entirely rethink ones that were designed pre-AI. New types of cognitive platforms driven by voice, perception, natural language, and so on will especially require reimagining how humans interact with machines.

Many organizations will struggle here because of analytical systems that are disconnected from the front line where second-to-second decisions will be made, or because they rely on manual processes that can't keep up with the rapid pace of deep learning capabilities.

Using a model's decision not only touches on explainability and transparency, but also consent, policy, ethics, and security; elements that can make or break the success of an AI solution in an enterprise setting. And, ultimately, the limiting factor of applying these advanced techniques will not be technology considerations—it will be managing the risk associated with pushing a high level of complexity into an operational context.

Just as enterprise risk management currently touches every part of the business, AI risk management and governance will need to incorporate perspectives from across the entire organization to support its adoption. Ultimately, it will be the coordination of these efforts that will build consensus and enable the transformation to a data-driven, analytics-based culture.

Operationalizing an AI solution is not just a question of building a machine that makes the right decision most of the time. It's a question of how to build it with a level of sophistication that enables it to produce a decision based on a blend of considerations, describe the reasons and evidence for its decisions, remain compliant with all relevant policies and regulations, and do this all fast enough to be useful in real time and at an extremely high degree of accuracy.

Because they are crucial to the project's ultimate success, operationalization concerns must be kept at the forefront of the project from the beginning. Leaving them for later risks seeing the project wither.

Executive buy-in and a commitment to redesigning core processes will be crucial to see success and work through some of the larger organizational changes. At the same time, building transparency and trust into your models from the beginning will help facilitate their adoption, enabling humans to move forward and act on model decisions.

Data

Deep learning requires the ability to fuse data from multiple sources of transactions, interactions, and rich media. Deep learning models are particularly adept at finding patterns and correlations across multiple heterogeneous subject areas. As an example, predicting failure of a complex industrial asset has the greatest chance of being accurate when diverse subject areas such as maintenance records, operational usage data, master data, and sensor readings are all factored in. As deep learning algorithms become mission critical to ongoing business operations, the data foundation must be resilient and hardened, facilitating efficient analytic operations.

However, enterprises are at different levels of maturity in being able to connect the dots across their internal and external data sources. Even for the most advanced organizations, deep learning often requires revisiting data management principles, particularly with respect to labeling, in order to properly train new supervised and unsupervised deep learning models.

In fact, 80% to 90% of the total work of a machine intelligence project will be in the data preprocessing—getting access to and cleaning the datasets and setting up data pipelines. This work should not be underestimated and can come with a truly staggering number of potential problems of different varieties.

Some organizations will struggle with even getting access to the high-value structured and unstructured data on which the deep learning project depends. The data itself might be mislabeled or unlabeled—a problem that is especially tricky for unstructured environments. You might also have data that is missing. Or, you might find that your data is imbalanced or that you can't get different features available for analysis in real time.

This is just a sample of the nefarious, subtle, and downright maddening issues that you can confront in data preprocessing. Data snags can persist after processing, as well—one of the most overlooked problems is training your model on data that does not look like the data that will be used in production, so you end up with the AI equivalent of learning to swim in the ocean by practicing in a goldfish bowl.

The antidote to these issues is to implement proven design patterns for both tightly and loosely coupled data management techniques in order to form an Agile data foundation. This will be crucial for success in model training and discovery because deep learning projects must embrace experimentation and rapid iteration to determine the highest-performing network architectures and what techniques can be applied.

Talent

So much of deep learning's success depends on the people behind it and their capabilities for guiding the program through each step. Many problems with deep learning are unique to the domain, such as model selection and the requisite data preprocessing, and require an experienced eye.

Unfortunately, there is a severe shortage of people who have production-level experience with deep learning in the industry, making it another one of the biggest challenges for enterprises that want to develop deep learning capabilities.

Though more and more people are acquiring deep learning knowledge through avenues like massive open online courses (MOOCs), it is rare to find people who can do it in practice, and the difference is important. The classroom is certainly a step in the right direction, but it does not replace real-life experience.

Companies have several choices when looking to acquire the talent that they need for their AI programs. Hiring or developing it is an option, though you should expect competition from digital giants and startups, as these have proven themselves to be magnets for both new grads and experienced practitioners in the field.

Acquiring other companies and their IP and talent is another route, or companies could also bring in a partner to gain external experience. To get the most value here, you should work with a partner that both has access to state-of-the-art information and will engage in knowledge transfer.

Conclusion

The challenges of working with deep learning are both technical and organizational. They begin at aligninng key stakeholders on where deep learning can be applied to address business priorities. Then, to execute a deep learning initiative, you will need to assess and address current capabilities with respect to data management principles, the analytic platform, and talent, and make plans to socialize and operationalize the solution across different business functions.

Over the next two chapters, we examine several custom AI solutions built to harness the power of organization-specific data. We also do a deep dive into how Danske Bank used machine and deep learning for fraud detection and its approach to some of the technical challenges discussed.

CHAPTER 5

Artificial Intelligence Case Studies

In this chapter, we look at some case studies based on real-world client engagements involving custom artificial intelligence (AI) solutions in high-value business contexts. The documentation for these studies was provided by data scientists and industry analysts from Think Big Analytics. In many of the studies, the client requested to remain anonymous.

Fighting Fraud by Using Deep Learning

> **NOTE**
> We go into much more detail about this first case study, Danske Bank, in Chapter 6. If you prefer only a high-level overview, we present this summary here.

Danske Bank is the largest bank in the Nordics, with more than $200 billion of assets in management. As with all financial institutions, one of its biggest priorities is fighting fraud, an increasingly difficult task in today's connected, complex, and anonymous financial transaction market.

The bank had been using a rules-based engine for fraud detection, but it was faltering against modern challenges, catching only 40% of fraudulent transactions. Furthermore, this engine suffered from an extremely high rate of false positives—99.5%. This meant an unnecessarily heavy workload for investigators as well as headaches for customers, and the problem was growing worse because of new payment methods and sophisticated fraud techniques. For Danske

Bank, lowering the rate of fraud represented a huge financial opportunity for preventing monetary loss, reducing investigator workload, and improving customer satisfaction.

The bank decided to update its fraud detection engine by powering it with machine intelligence, a decision that led to impressive results. Danske Bank saw a 50% increase in its fraud detection rate and a 60% decrease in the false-positive rate, a huge improvement over its rules-based engine. Furthermore, the bank did not use any additional data types to achieve these results. Instead, machine and deep learning algorithms were able to detect fraudulent patterns in the same datasets used in the previous solution.

This project was carried out in three interconnected tracks. As a foundation, the team worked to enhance the analytics infrastructure so that it could support automatic decision-making with machine and deep learning while maintaining latency requirements.

When that was ready, the machine learning track began. During this phase, data had to be gathered, cleaned, and routed to train the models. After the models were proven in a testing environment, they were productionized. This immediately created some lift over the manual rules-based engine. However, the rate of false positives was still quite high (98%), and some instances of fraud were yet undetected.

The team decided to start a deep learning track to capture these remaining cases and improve the engine's accuracy. Using the same structured data that had been feeding the rules engine, the team discovered a way to turn it into something that looked like an image. The team was then able to use a convolutional neural network—a type of algorithm typically associated with object recognition—to detect visual patterns in the data that predicted certain transactions to be fraudulent. Using this method, and without using any additional data sources, the team saw substantial improvement over the machine learning output.

Danske Bank was able to win with machine intelligence for a number of reasons—it had the right kind of executive support that made key investments in infrastructure and in developing new processes to accommodate the AI; the company built a solid data foundation that could service a high-performance engineering environment; and the team employed rigorous testing and management for the machine and deep learning models, which ensured accuracy and built trust and credibility.

Mining Image Data to Increase Productivity

In addition to gaining insight into traditional datasets, deep learning can help enterprises dig into data that was previously inaccessible. In this next example, one major logistics company in the United States was able to mine images to increase productivity.

Of the millions of packages that this logistics company ships every day, a small number wind up as orphans—both its shipper and its destination are unknown. For years the problem had been solved manually. After the missing package was reported, someone would go to a warehouse where the orphaned packages were stored and try to match up the claim with one of them.

The process was incredibly expensive, costing the company tens of millions of dollars every year. If the company improved the situation, it could increase customer satisfaction (helping people get their packages sooner) and reduce its liability for reimbursing lost property.

The company decided to approach the problem with machine intelligence, using computer vision and image recognition to design and deploy an AI-assisted solution of image-to-image search with a backend of TensorFlow and Keras. People who had sent packages that had then become lost would submit a picture of the contents of the lost package (if it were available). On the other end, the warehouse would upload pictures of the contents of orphaned packages. Using the search, the company was able to match the pictures that people had sent in with their claims for an orphan package with the actual inventory of the orphan warehouse.

The company was able to achieve a 90% match rate between the images, reducing the warehouse search window from weeks to minutes. This represented a huge reduction in resources as well as the opportunity to improve customer satisfaction due to quick resolution of lost-and-found cases.

Deep Learning for Image Recognition

In another example of using deep learning for image recognition in the enterprise, a state-run postal service was able to improve efficiency on the shop floor by building a custom AI solution that could accurately identify plastic bags.

At this postal service, 115 million parcels are shipped every year. Of those, 7.5 million are plastic bags that need a special sorting process to complete shipping. On the shop floor, these bags must be put into designated bins where they can be sorted for delivery. In the system the postal service had been using, only a small percentage of the plastic bags were detected, with most of them remaining on the conveyor belt, where they needed to be manually picked up and delivered to the correct bin. Employees were responsible for both identifying the size and quality of package (whether they were plastic) and loading and unloading the parcels into the correct containers, a time-consuming task that decreased the system's efficiency significantly.

The postal service needed a way to automate the process, which would reduce labor costs and complete the shipping more effectively. To solve this problem, it trained deep learning algorithms on millions of images from several cameras on the shop floor. After the algorithms could detect plastic bags (not an easy task, as even slight changes in image quality can affect accuracy), the organization was able to build a system to identify the packages in real time and instruct the conveyor belt where to direct the plastic ones.

The solution involved blending traditional data sources with new image data in order to complete the process of package identification and direction. Ultimately, the postal service was able to achieve 80% accuracy on the testing dataset for the identification of plastic bags using an automated pipeline, employing Keras with a Theano backend (because of the more user-friendly interface and limited time for the project).

Natural-Language Processing for Customer Service

Customer communication represented a big challenge for a global communications provider. Not only was the amount of communication significant, but it was also costly to try to maintain high levels of availability for support. The company wanted to find an automated way to answer some of its customers' most common questions with 24/7 availability, reducing the support workload and increasing customer happiness.

Because of the large repositories of customer interactions, the project was a perfect fit for deep learning, which excels at decoding noise and complex data like natural language.

Using Flask, Python, Jupyter, and Microsoft Azure, the company employed machine and deep learning to create several algorithms that replicated 322 types of queries in English, Roman Urdu, and their mixture. This was a significant technical challenge because there are many approaches to choose from regarding natural-language processing techniques, how to analyze and pair up questions and answers, and how to determine when to use unsupervised versus supervised learning. Dealing with questions in two languages also proved to be a significant challenge.

Ultimately, however, the team was able to deliver a user interface in .NET for interacting with the virtual agent to answer these queries. The virtual agent could also be extended to other response types. This had the dual result of reducing the cost of customer-driven communication and increasing customer intimacy and satisfaction.

Deep Learning for Document Automation

As part of the process for opening an account at a European bank, bank employees were responsible for manually validating the application by examining an applicant's submitted documents. These could include pay stubs, copies of photo IDs, proof of address, and so on.

The bank employed about 80 people in this process and was interested in using deep learning to make predictions about the validity of the documents for a number of reasons. It wanted to increase the accuracy of the decisions that were made as well as reduce the cost of the process and speed it up.

The bank was able to use deep learning models to read and predict document validity with up to 90% average accuracy, which represented a significant improvement over the human team in terms of speed, scale, and accuracy. The models themselves were built on a sizeable kit of Amazon Web Services (AWS), cognitive APIs, TensorFlow, Keras, Python libraries, libraries for processing PDFs and images, and GPUs.

There were some challenges in creating this solution, given that working with image detection is never easy. Additionally, many of the forms were completely different from one another (especially the pay stubs), and much of the data needed to be manually annotated before the algorithms could be trained on it.

Conclusion

By creating a custom AI solution, these companies were able to utilize their data and expertise to refine their specific business processes. In these same situations, SaaS solutions or cloud APIs might have partially solved the problem, but they would have fallen short of delivering transformative value.

As you have seen, however, deep learning solutions can involve complex engineering problems. In Chapter 6, we take a closer look at the Danske Bank AI case study and how the bank was able to use AI to fight fraud and see a significant improvement in results over its human-based rules engine.

CHAPTER 6

Danske Bank Case Study Details

In this chapter, we take a closer look at how Danske Bank is achieving high-impact business outcomes by fighting fraud with machine intelligence.

As discussed in Chapter 5, Danske Bank was struggling to mitigate fraud by using legacy detection systems. With a low 40% fraud detection rate and a 99.5% rate of false positives, it was clear the bank needed to modernize its defenses. To do this, it made a strategic decision to apply new analytic techniques, including AI, to better identify instances of fraud while reducing false positives in real time.

In partnership with Think Big Analytics, a Teradata company, the bank was able to develop analytic solutions that take advantage of its unique data and provide a significant improvement over its previous rules-based engine, reducing false-positive detections of fraud by 60% with machine learning (with expectations to reach as high as 80% using deep learning), and increasing the true-positive detection rate by 50%, as illustrated in Figure 6-1.

Figure 6-1. Danske Bank business outcomes enabled by AI

The Project, the Tools, and the Team

The project evolved through several phases. First, the team laid the foundation for the analytics platform and got the data plumbing in place that would serve the machine and deep learning models. Then, the team trained, tested, and deployed machine learning models before moving on to the deep learning track. As of this writing, the deep learning models are showing great promise at Danske Bank, and the team is working on putting them into production.

The data science lab brought together a variety of tools for training, validating, and promoting the machine and deep learning models after they were proven:

- The lab used both CPUs and NVIDIA GPUs to process the data.
- They employed a variety of software frameworks, including the following:
  - — Anaconda Python distribution for some of the simpler models
  - — Spark and Hive for data preparation and wrangling
  - — TensorFlow and Keras for building deep learning models
  - — Tracking and deployment software like Git

Getting the Right Data in Place

When the team kicked off the project, it began by building out the data layer, making sure it had access to the right kind and quality of data that it needed. The team also ensured that it had the right features in place to train the machine learning models.

Right away, the team faced a significant challenge. The bank had only a very limited set of accurate data to work with when the team came on site, and the team quickly established that it needed to improve on that data quality and size before it could develop accurate models.

To get more data, the team first needed to identify and extract historical fraud cases to use as positive examples. Because these were logged in unstructured Excel sheets, the team needed to extract the fraud cases from them in a semi-manual process using regular expression matching and manual processing. Though the work was tedious, one benefit was that the team was able to get a better understanding of what typical fraud schemes looked like as it examined thousands of cases during the extraction process.

Then, the team had to reconstruct all historical transactions within the previous three years between senders and receivers into single transactions from their respective subtransactions (because a single transaction between two parties goes through a variety of intermediate accounts, depending on type and origin). This reconstruction was no small feat, considering that billions of rows of data existed outside the normal business logic of the bank's real-time transactions systems from which the team had to identify relevant transactions and apply the right rules.

As the third step, the team then needed to match the fraud cases to the billions of transactions to ensure that it could train on an accurate dataset that had correct information about which transactions were fraudulent and which were not.

Finally, the team had enough accurate data to work with and could train the models.

Ensemble Modeling and Champion/Challenger

There are a variety of ways to ensure that you're getting the best output from your advanced analytics models. One of those is by using *ensemble modeling*, which is the process of running two or more related but distinct analytical models and then synthesizing the results into a single score. Combining multiple models in this way helps reduce noise, bias, and variance in your output to deliver superior prediction power—the advanced analytics version of "all of us are smarter than one of us." Several well-known techniques (i.e., the use of bagging and boosting algorithms) can further embellish this method to enhance accuracy.

Another time-tested methodology for improving analytics outcomes is using *champion/challenger*, in which advanced analytics systems compare models in real time to determine which one or ones are most effective. At Danske Bank, for example, challenger models process data in real time to see which traits are more likely to indicate fraud. When a model dips below a certain predefined threshold, it is fed more data or augmented with additional features. And when a challenger outperforms another challenger, it is transformed into a champion, leading other models closer to better fraud prediction. Continual retraining helps retain the accuracy of the highest-performing models.

Together, these methods enable increased prediction power and provide other benefits, such as making sure the system always returns an answer, even if one model takes too long to score.

Working with the Past, Building the Future

After the team was able to successfully train the machine learning models—an ensemble of boosted decision tree and logistic regression models—it found another hurdle as it moved to put them into production. As a 146-year-old institution, Danske Bank had decades of transactions running through a mainframe server. It was going to be a challenge to get the models into production and maintain latency requirements using the bank's current infrastructure.

The bank needed an architecture that would allow the models to run across the millions of daily transactions. To make that possible, the team enhanced the bank's infrastructure capabilities with an analytics platform that would be able to elevate the models to production and could be used in future applications for various domains. This was an add-on to what was already running, and the rest of the bank's architecture was left in place.

Moving the ML Models into Live Production

With the advanced analytics platform built and only three months into the engagement, the team was able to go into shadow production of the machine learning models. A shadow production phase was necessary to help stakeholders become familiar with the model and determine whether it needed retraining before it went live.

After three months in shadow production, the machine learning framework was hardened, and it was expanded to run over multiple datacenters with continual monitoring. Nine months after the start of the project, the machine learning model was ready to be put into live production, where it saw impressive results.

The models were a significant improvement over the former rules-based system, with the rate of false positives reduced by 50%. This removed half the workload of investigators. However, many instances of fraud were still going undetected.

To work well, the machine learning models had to view transactions atomically. They could not ingest information about sequences of events, let alone correlation across channels, features, dependencies, and time series—clues that would certainly help pinpoint more instances of fraud. These, however, are areas where deep learning excels.

In the next phase of the project, Danske Bank integrated deep learning software with GPU appliances to try to capture the remaining cases of fraud and achieve an even lower rate of false positives.

From Machine Learning to Deep Learning

As it moved onto deep learning, the team was able to use the analytics platform it had built during the machine learning phase to test and validate different kinds of neural network architectures. Figure 6-2 is a receiver operating characteristic (ROC) that shows the baseline legacy rules engine, the lift associated with classic machine learning, and the even higher lift stemming from an ensemble of deep learning models.

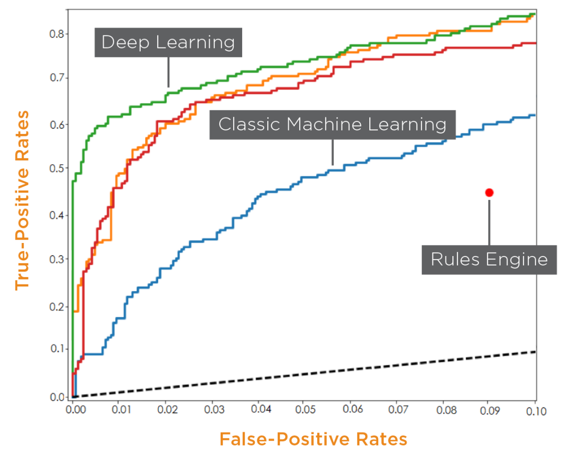

Figure 6-2. ROC showing dramatic lift resulting from Deep Learning models when compared to Classic Machine Learning and legacy Rules Engine

Over the course of four two-week sprints, the team tested a variety of deep learning models, including convolutional networks, Long Short-Term Memory (LSTM), and generative models. One was found to be especially effective—the ResNet model architecture is a version of a convolutional neural network and an algorithm that is typically used for object detection and computer vision.

This might be surprising, considering the team was not working with pictures or video. Rather, it was using traditional table sets and structured data. So how (and why) did the team make this work?

Visualizing Fraud

The deep learning architectures that specialize in visual detection and object recognition are some of the most advanced, even surpassing human performance at identifying and labeling objects, and doing it much more quickly and reliably. Another benefit of convolutional neural networks is that they use relatively little preprocessing compared to other image-classification algorithms, so people need to do less work in order for the algorithm to determine what's relevant.

The team was interested in using these architectures because of their maturity. To do so, it discovered a way to turn a table set of data into an image by taking raw features as input and clustering correlated features along the x-axis, with the y-axis representing time, providing a two-dimensional view of the transaction. The team then fed this view into the model, which was then able to detect new relationships and complex patterns in the data. Figure 6-3 demonstrates how features such as frequency of transactions, merchant location, relative size of transaction, and others are transformed from a tabular layout into a matrix.

Figure 6-3. Converting tabular data to a matrix in order to conform to deep neural network input requirements

This matrix is similar to a digital image, meaning it better conforms to the input required by neural networks. Figure 6-4 shows the model output, with fraudulent transactions appearing more red in hue when compared to bona fide transactions.

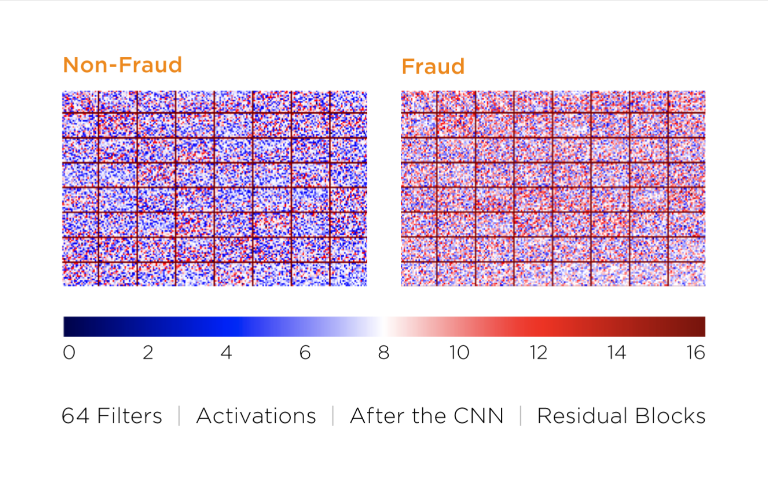

Figure 6-4. Visualizing the convolutional neural network output for fraud detection

The net result was a 20% reduction in false-positive rate—a significant improvement over traditional machine learning models.

Visualizing and Interpreting Deep Learning Models

Deep learning can provide a significant advantage over machine learning in some domains. However, it does come with its own challenges. In particular, it can be difficult to understand how deep learning algorithms make decisions.

In using deep learning for financial transactions, however, model interpretability is crucial for a number of reasons:

- Investigators have less work to do if they understand why a model made a particular decision, because they know what to look for as they are examining possibly fraudulent transactions. They can also gain insight into why the fraud is happening, which is information that can be very useful. Finally, interpretability increases trust in the model's results, helping with its adoption and integration into current processes.

- For Danske Bank, interpretability is also necessary for compliance with the EU's General Data Protection Regulation (GDPR). If it is found that the bank is not in compliance with these regulations, which requires that any financial institution be able to provide information about how it used a customer's data, it could face very heavy financial penalties.
- It is also important for building customer happiness and trust, so the customer can have a satisfactory answer as to why their transaction was denied.

To approach the problem of interpreting its deep learning models, the team deployed open source work out of the University of Washington called Local Interpretable Model-Agnostic Explanations (LIME).

LIME (introduced in Chapter 4) is a system that allows you to examine the key characteristics of a model at the point of decision so that you can see which specific data points triggered the model. If you have multiple models running, LIME is also helpful in comparing them and finding out which features were triggered in order to judge performance.

Though LIME is certainly a step forward, this problem is far from solved. Visualizing and interpreting deep learning models is important and ongoing work. For fraud detection, it is crucial to be able to see that fraud events match human expectations based on experience and history, and that the model is treating fraudulent transactions using different mechanisms than nonfraudulent ones.

A Platform for the Future

Through their partnership with Think Big Analytics, Danske Bank was able to build a fraud detection system that made autonomous decisions in real time that were aligned with the bank's procedures, security, and high-availability requirements. The solution also provided new levels of detail, such as time series and sequences of events, to better assist the bank with its fraud investigations. With it, the bank's engineers, data scientists, lines of business, and investigative officers were able to collaborate to uncover fraud.

Though this chapter emphasizes the technology behind the solution, the organizational and change management considerations of this project were equally essential for the solution's success. The project leaders fully understood and embraced change-management best practices, beginning with small wins to prove the value and viability of the AI solution, socializing it across stakeholders, and moving from there to full implementation and operationalization.

For Danske Bank, building and deploying an analytic solution that met its specific needs and used its data sources delivered more value than an off-the-shelf model could have provided for a number of reasons, not the least of which is that no off-the-shelf product would be able to provide fraud detection techniques at the level of its custom solution.

With its enhanced capabilities, the solution is now ready to be used across other business areas of the bank to deliver additional value, and the bank is well-poised to continue using its data in innovative ways to deliver value to its customers.

CHAPTER 7

Predictions Through 2020

As we've seen, artificial intelligence (AI) is already producing high-impact business outcomes, and it will most certainly continue to advance rapidly. In this chapter, we take a look at some trends in the field and provide informed predictions to help enterprises prepare for the next two years and beyond. Figure 7-1 shows the five pillars of enterprise considerations for AI initiatives.

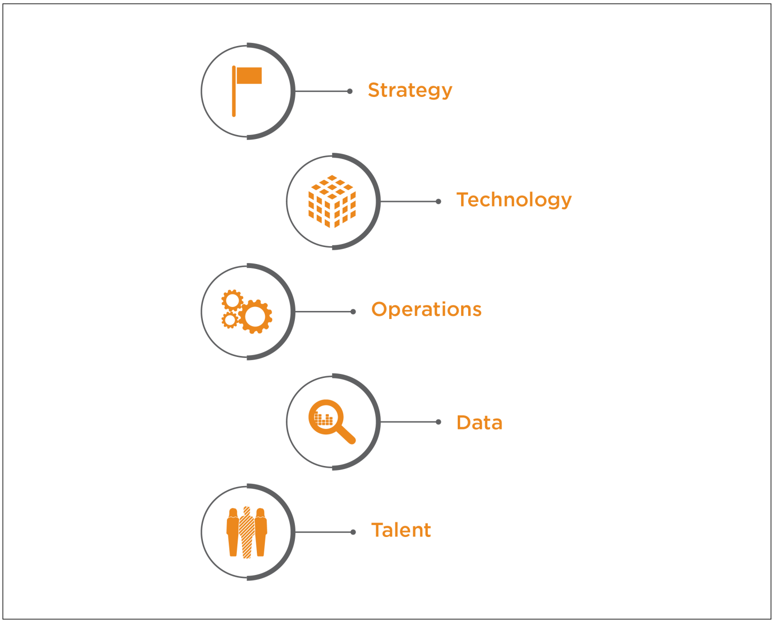

Figure 7-1. Five pillars of enterprise considerations for AI initiatives

Strategy

The future of AI will increasingly be a winner-takes-all proposition, with the first companies that are able to harness AI abilities in an effective way dominating their industry or subindustry segment and then expanding. In this landscape, the companies with the best models and the best data will take the lead and do so quickly. This can be seen currently with Amazon, which has harnessed AI's predictive powers in retail and logistics and is now pivoting into healthcare and finance. Even though retailers like Walmart and Target will compete in an AI-first future, they are being challenged like never before.

A successful AI strategy will ensure that AI initiatives are tied to business outcomes and channel investment and energy accordingly. In telecom, for example, this might mean identifying AI projects that directly affect business outcomes, like increasing network productivity by intelligently detecting anomalies, or improving annual revenue per user through product personalization.

Other AI use cases, although not as obviously attached to business outcomes, might be just as important. For example, deploying an AI solution to automate cumbersome data integration processes and costs could improve a company's potential for later AI and analytics projects that directly facilitate business goals. The strategic challenge will be identifying the correct initiatives and the right timeline for their implementation.

The strategy must be multidisciplinary, addressing technical, cultural, and corporate interests. This includes elements like investing in the technology to unlock data value and drive information yield as well as creating a culture of experimentation and facilitating the democratization of analytics, enabling their self-service consumption across the enterprise. The strategy must also account for making sure that the AI solution is compliant, secure, and ethical, with appropriate developments made in terms of risk management and business continuity.

Many enterprises will invest in AI, but not all of these investments will be aligned with the desired business outcomes dictated by corporate strategy. In contrast, those that win at AI will ensure each initiative is making an impact in a discrete and measurable way and that company culture and processes are conducive to continual innovation.

Technology

Though some are attempting to develop proprietary AI technology, open source frameworks will continue to dominate the field, benefitting from the intellectual contributions of experts from around the world. Of these deep learning frameworks, the ones backed by digital giants (e.g., TensorFlow and Google) represent the best options for enterprises. These frameworks have the benefits of scale as well as sponsors with vested interests in improving and supporting them.

Gaps in support, however, will remain for open source software as these cloud vendors fight to grow their businesses by embracing a vendor lock-in model. This might change at some point, but for now, enterprises that want more flexibility in their environments will need to seek other options.

Unlike Hadoop, which quickly became the industry standard for big data and generated several vendor distributions, framework fragmentation for AI will remain a way of life. Although out-of-the-box AI solutions are on the horizon, they are not here yet. Interoperability between open source tools is a serious issue that the industry must address as well as developing productization features like orchestration, security, and user-friendly interfaces.

Though tools will continue to be developed that abstract complexity and make deep learning more accessible, these will not have the effect of turning business laypersons into data scientists. Rather, these tools will enable current practitioners to further their productivity and move toward value much more quickly. For example, better tools might make optimizing CPUs and GPUs go from taking up 25% of the work of the project to something more like 5%, or even 0%.

Look for advances in hardware that will make training and inference of deep learning models easier, more integrated into the enterprise hybrid cloud platform, and available at a lower total cost of ownership.

Training deep learning models in the cloud is currently the most popular way to take advantage of GPUs; that is the path of least resistance because of ready-to-run servers that have a pricing model conducive to ephemeral experimentation. However, the majority of large enterprises are not committed to moving 100% of their data to the public cloud (let alone a single cloud vendor). As such, the market will respond to enterprise demands to provide ready-to-run, enterprise-grade, deep learning hardware that can be deployed on premises in a way that removes usage friction between deployment options.

Like past breakthroughs in database hardware (and most recently with in-memory processing), GPU databases will ultimately be absorbed into enterprise-class hardware platforms that offer the benefits of both GPUs and CPUs, which will ultimately work together smoothly and synchronously.

General-purpose hardware and chips from NVIDIA and competitors will dominate the field as their price-to-performance ratio continues to fall in the race to create the fastest chip. In the future, however, we'll see application-specific integrated circuits (ASICs) and field-programmable gate arrays (FPGAs) used for AI engines in the post–Moore's law era. These types of hardware will be critical for embedding AI in systems that need to run in real time. The systems will include custom chips for AI workloads, edge-cloud systems for efficient data processing at the edge, and techniques for abstracting and sampling data.

Operations

Though significant, the tech challenges surrounding AI will almost certainly be solved within a few years by leading analytics-forward companies. Past that point, the biggest hurdles to operationalizing AI revolve around issues of risk management, ethics, compliance, security, and trust, as companies update business processes and ethical codes to match the sophistication of their analytics.

Trust in model output and transparency in decision-making will become increasingly important as AI solutions touch more areas of the business—affecting the customer base, core products, corporate strategy, and ultimately, human lives. Furthermore, it's true that AI is pivotal when it comes to gaining competitive advantage, but if not done correctly (and especially in high-risk industries), it could also be disastrous. Given that public opinion is extremely sensitive to the failures of AI, industry front-runners should prepare for special scrutiny with regard to their AI programs.

Facing these challenges, it will be important to move toward a quantitative, numbers-based risk-management system to automate or assist AI-powered decision-making in an operational environment. Instead of a hand-assembled team of experts that manually reviews documents and conducts interviews to evaluate riskiness, over the next few years we will see chief risk officers looking to sophisticated analytics to determine risk exposure with regard to AI.

Enterprise crisis management and disaster preparedness must also begin to encompass managing AI events and supporting highly available AI. For example, if an AI solution critical to the business needed to be recalled for some reason, could it be rolled back? Could humans take over temporarily until the models were restored?

In terms of security, it will be important to develop ways to defend against possible malicious assaults on analytics models made with the intent of corrupting output or otherwise affecting a program. One naive example of this could be what happened with Tay, the Twitter chatbot that Microsoft deployed, that was quickly taught to be a racist xenophobe. These kinds of problems (i.e., intentional data manipulation) become more serious when they touch core business processes or even national security.

Data privacy and consent are well-known ethical issues surrounding the development and use of AI technologies, but it is certain that unforeseen ethical questions will arise from the use of these technologies as well, and companies must have a strategy in place for when they do. Model transparency will be a key element here in order to make sure the AI solution is operating in line with regulatory, moral, and practical concerns. Successful organizations will design models with explainability and trust in mind, with the right level of transparency and interaction with humans.

The importance of trust with AI output cannot be overstated. The public still struggles to trust AI consumer products like virtual assistants, and it will be a challenge of quite a different magnitude to have a business rely on AI. A well-thought-out risk-management plan in place alongside transparent models and a robust security strategy will be key elements to building the trust that will enable the successful operationalization of an AI solution.

Data

The community must also continue researching ways for algorithms to learn more quickly, whether through learning in an unsupervised manner or with less supervision or scaling the process of labeling data (which can be very time-consuming and costly). With continued research, unsupervised and transfer learning might go mainstream in two years, allowing people to get value from deep learning sooner.

Although currently machine intelligence models learn on training data before operating in the "real world," it will be important to design AI systems that learn continually by interacting with a dynamic environment while making decisions that are timely, robust, and secure. Doing so will help them adapt to changing environments and patterns of data, which is crucial to retaining accuracy. This requires them to be able to respond to situations that they have never encountered before.

Shared best practices around AI will help success breed more success. Proven design patterns, code, IP, and proven deep neural network taxonomies will accelerate time to value and de-risk domain-specific use cases.

There will be a push toward AI becoming a group effort, with companies increasingly using third-party data to augment their AI-powered services. The proliferation of such applications will lead to a transition from data silos to data ecosystems, in which applications learn and make decisions using data owned by different organizations.

To reach this stage of collaboration, it will be necessary to design AI systems that can train on datasets owned by different organizations without compromising their confidentiality. This way, organizations can learn from one another and, in the process, provide AI capabilities that span the boundaries of a potentially competing organization (e.g., banks sharing data to combat fraud).

We will also see increasing automation of processes that are used to build AI itself; that is, AI building AI. Many time-consuming and manual tasks surrounding data integration and cleaning will soon become the domain of advanced programs that perform the tasks much more quickly and accurately than humans across a wide variety of data sources (e.g., ETL.ai).

Talent

The talent landscape has huge implications for the AI industry and is affected by a number of trends, most notably its uneven distribution. Currently, leaders like Google and Amazon are absorbing a huge amount of the available talent, which goes on to serve their business models (i.e., advertising and retail). The result is that some industries (e.g., oil and gas, and manufacturing) are left behind despite their desperate need for innovation due to changing economies and job markets.

Even though there will be more and more practitioners as a result of expanding graduate and undergraduate programs, expect for talent to remain unevenly distributed, with the lion's share going to industries where significant innovations are already occurring. As part of the same trend of talent scarcity, AI startups will continue to be gobbled up in acqui-hires.

Edged out by digital giants and startups, Fortune 500 companies will continue to starve for AI expertise. To acquire it, and left with less-attractive options, they will attempt using cloud APIs or point solutions and end up falling short of the transformative power of the technology. Far from an indictment of the potential of AI, these results will be a direct realization of the talent situation.

To remain competitive—and in addition to training and preparing the current workforce for AI—enterprises must develop a robust talent strategy, either through hiring graduates, acquiring IP through mergers and acquisitions, or interacting with academia or a partner that believes in knowledge transfer.

Gaining momentum and demonstrating wins in the AI space right now will lay the foundation for attracting talent in two or three years, when it will be even more crucial.

What's Next

Deep learning will soon become a pervasive technology, transforming organizations, enabling new types of businesses, and creating new industry leaders. Looking toward the future, it is the companies that are investing in AI capabilities right now that will position themselves to attract and retain valuable talent, capitalize on the industry's ever-improving tools, and maintain an edge over their competitors.

CHAPTER 8
Conclusion

Companies are seeing value from artificial intelligence (AI) initiatives right now, and this is just the beginning of the transformational change we are soon to experience across many industries. In this book, we examined some real-life examples in which companies were able to address the challenges associated with AI initiatives in order to realize their value. To get started with AI at your own organization, we recommend the strategies outlined in this chapter.

Identify High-Impact Business Outcomes

You should begin introducing AI into your organization by first identifying the high-impact business outcomes that AI can address. Follow these guidelines:

- Identify the areas where AI can make an impact in your organization.
- Avoid deploying technology for technology's sake. Instead, develop an understanding of what can—and should—be accomplished with AI in your organization based on the kinds of problems that state-of-the-art AI research is currently solving in other spaces.
- Educate stakeholders to ensure realistic expectations for the technology and alignment on what is possible.
- Clearly express your organization's policies on AI and address head-on your staff's fears and concerns regarding displacement or disruption due to automation.

- Demonstrate the viability of the technology to solve problems in the areas you've identified, proving its value to stakeholders across different functions.

Assess Current Capabilities

The next step is to make an honest assessment of your organization's current capabilities. The following are steps you can take:

- Start by assessing your current capabilities and examining the platform, data governance, and data science abilities within your organization as well as your organization's culture—will it facilitate an environment that rewards AI successes and learns from its failures?
- Benchmark with peers and against your competitors.
- Prioritize addressing capability gaps across the pillars of data, talent, analytics, and technology.
- Consider engaging an independent third party in this assessment to ensure its thoroughness and veracity.

Build Out Capabilities

Finally, you should build out capabilities with the end in mind, as follows:

- Right from the start, think about how analytics should be made operational. Build and scale your solution following AnalyticOps best practices with continual retraining, monitoring, and governance, refining the system until it is hardened and fully operational.
- A successful AI program requires robust design thinking and bold experimentation. For this, you will need to move quickly, building agility into your processes so you can easily pivot based on new information gained from your successes and failures.
- Resist departmental rogue efforts that deviate from the strategy and add complexity and costs. Siloed efforts are usually a symptom of inadequate communication or program funding. Moreover, they effectively kill the AI inspiration because

transformation can occur only when data, analytics, and processes are aligned across departments toward one outcome.

The challenges of developing and deploying a custom AI solution are not insignificant, especially given that the field is still maturing. It will require strategic focus, a willingness to discover new ways to approach your business, and a commitment to innovation.

That said, there is too much to be gained by implementing an AI solution and becoming an analytics-forward organization to wait to start. This is an era of exponential change—your organization's evolution is ready to begin now.

About the Authors

Atif Kureishy is VP of global emerging practices and AI/deep learning at Think Big Analytics. His teams are trusted advisors to the world's most innovative companies to develop next-generation capabilities for strategic, data-driven outcomes in areas of artificial intelligence, deep learning, and data science. You can connect with Atif on LinkedIn and Twitter @AtifKureishy.

Chad Meley is VP of product marketing at Teradata, responsible for its Teradata Analytical Ecosystem, IoT, and AI solutions. Chad understands trends in the analytics and big data space and leads a team of technology specialists who interpret the needs and expectations of customers while also working with Teradata engineers, consulting teams, and technology partners. You can connect with Chad on LinkedIn and Twitter @chad_meley.

Ben Mackenzie is director of AI engineering at Think Big Analytics, a business outcome–led global analytics consultancy, where he leads the team helping enterprise customers build and deploy deep learning models to drive business value. In addition to a solid hands-on experience and theoretical understanding of deep learning practices, Ben draws on years of experience building solutions using big data and public cloud technologies for a broad array of enterprise and startup customers. You can connect with Ben on LinkedIn.