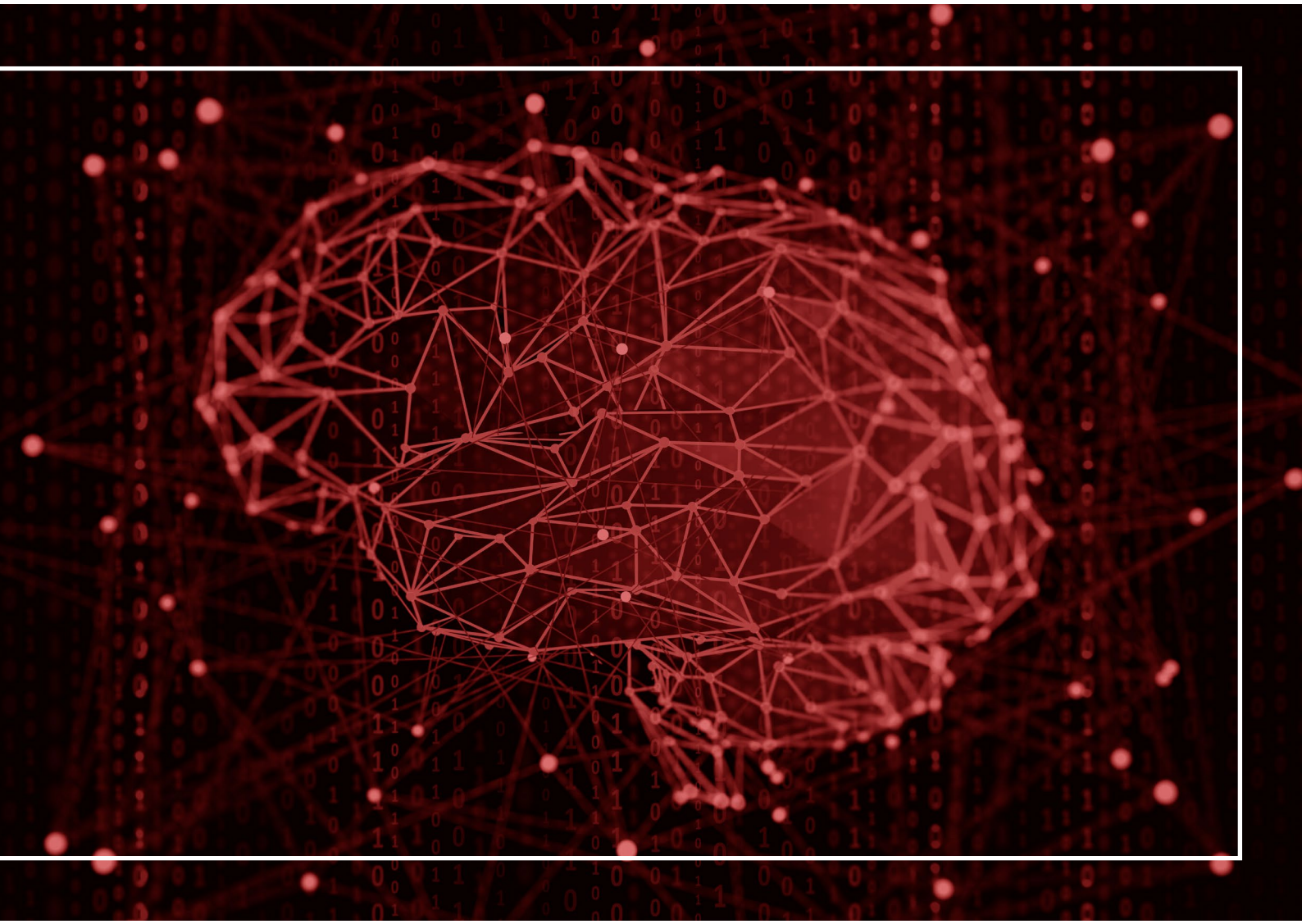
DEMOCRATISING MACHINE LEARNING



CAPCO

THE FUTURE. NOW.

ABOUT THIS REPORT

Big Data is here to stay, and it's only going to get bigger. Organisations are investing in tools and resources that can harness the power of that data, but not everyone knows how to capitalise on it. It's a difficult and expensive job, even for the highly-trained data scientist, of whom there are few. To counter the skills-gap, we are seeing an evolution of tools and technologies designed to make it possible for, say, a business analyst to access and manipulate even massively large datasets. For example, Google launched Dataset Search and BigQuery – easy to use platforms that scale to Big Data and can be leveraged by everyone in the organization. Google took this a step further through its release of BigQuery ML, a tool to build and deploy machine learning models through simple, broadly understandable SQL statements.

These self-service tools eliminate the prohibitive skill barrier, offering non-technical users access to an internet of data and the power of machine learning to analyse it. Though some banks have begun to explore their options with respect to large-scale adoption of Google Cloud AutoML solutions, others have been reluctant due to concerns that security and code dependencies remain. In this report, we discuss these concerns in relation to current solutions and highlight key considerations and take-away messages for financial institutions.

I. INTRODUCTION

Machine learning (ML) is the science of getting computers to act without being explicitly programmed. It is based on algorithms that can learn from data without relying on rules-based programming. In the past decade, ML has given us practical effective web search, self-driving cars, speech and face recognition and a vastly improved understanding of the human genome. As the technology evolves and ML becomes more pervasive, we are seeing its tremendous potential for value creation, so much so that it is hard to imagine the future of the financial services without machine learning.

II. TRENDS ACROSS INDUSTRY

MACHINE LEARNING TECHNOLOGY TRENDS

Machine learning (ML) has been the buzzword of the decade and is certainly one of the most disruptive technologies of this century. Though still considered by some to be nascent, it is creating significant value for the world economy. The International Data Corporation (IDC) forecasts that spending on ML and artificial intelligence (AI) will grow from $12b last year to $57.6b by 2021[1], within which the ML market alone is expected to grow from $1.41b to $8.81b by 2022.

Executives in companies around the world are increasingly looking to AI to create new sources of business value. This is especially true for leading adopters of AI. These so-called "pioneers", who are leading the race ahead of investigators, experimenters and passives, are doubling down on AI investments, building competencies, and working to take AI to scale[2]. As well as deepening their commitments to AI, pioneers are also prioritising revenue-generating applications over cost-savings ones – a significant adjustment from the traditional business approach.

A SHORTAGE OF ML EXPERTS

While ML has the potential to create marked business value and efficiencies, for many the potential remains just that, with numerous barriers still existing. These barriers are varied and many, be they a lack of executive-level opt-in, data inaccessibility, or poor data quality. Arguably however, the most substantial obstruction to organisations' efforts has been a lack of skills.

The insatiable demand for creating value from data has given birth to a web of new, highly technical jobs, and the need for ML specialists is on the rise. The wide range of fields is inviting many candidates of varying expertise, a fraction of whom are versed in more advanced programming languages, making them ideal for the job as they can directly apply their skillset to build ML models. On the other hand, the majority candidates either lack the prerequisite level of programming skills or do not possess the necessary understanding of underlying ML theory. Despite a proliferation of specialised courses, titles, and university degrees, there remains a huge imbalance between demand and supply, presenting a significant barrier for organizations eager to unlock value from their data.

1. https://www.idc.com/getdoc.jsp?containerId=prUS43095417
2. 2018 MIT Sloan Management Review and The Boston Consulting Group (BCG) Artificial Intelligence Global Executive Study and Research Report.

II. TRENDS ACROSS INDUSTRY (CONTINUED)

THE RISE OF THE CITIZEN DATA SCIENTIST (CDS)

Rather than attempting to hire finished articles in an increasingly barren marketplace, companies have opted for investing in re-training their staff, and upgrading their existing talent pool. For example, insurers are training actuaries in data analysis skills[3]. The approach has several advantages, most significantly, it advances data skills across the board. This is especially key at a time when a growing number of organizations are attempting to democratize data science capabilities across the workforce, rather than concentrate it within a data science function.

This trend was highlighted back in 2016, when the term 'citizen data scientist' (CDS) was coined[4], referring to a person who can use advanced data analytics capabilities to create models despite their job function being outside the field of statistics and/or data science. Indeed, the approach is an effective path to mitigate the skills gap by avoiding the need for dedicated data analytics or business intelligence experts altogether. The rise of the CDS has generated much interest within organisations, since the CDS can manage both the technical data analysis and the business demand, in turn spreading the data-driven culture that data advocates crave.

TECHNOLOGY THAT IS ACCESSIBLE TO EVERYONE

Concurrently, technology giants have been working on reducing the skill-demands and making technology more accessible. There is now an ever-growing plethora of tools and services designed to facilitate Big Data analytics outside of the IT lab, and across the organization as a whole. Most recently, developments have extended their reach to incorporate easier accessibility to both data and analytics, and tools now exist that can incorporate ML capability to automate data preparation, insight discovery and data science. Vendors such as neptune.ml are already offering pay-as-you-use pre-trained models as a starting point, as well as using built-in assistive features that simplify and accelerate the process.

Large players are expediting this trend, Google in particular. Earlier this year, Google launched Google Dataset Search, a search box akin to the Google Search bar, but focussed wholly on datasets. The launch was followed by the introduction of BigQuery and BigQuery ML, tools also designed by Google to make it easy to access and manipulate large datasets. For example, BigQuery requires knowledge of SQL only, as opposed to traditional data science languages such as R and Python. The launch of these platforms emphasises the value Google sees in making Big Data analysis more accessible.

Democratizing machine learning by making it an accessible commodity to anyone with SQL experience could be a real winner for Google. We predict that even the largest, most complex organisations will seek the technology that can give a technical business analyst the ability to generate ML-level value from massively large datasets.

As the creation of a data model is delegated to the machine, the demand on the skill level of the user is reduced accordingly. This is, perhaps, the natural progression of Big Data analysis.

3. Capco proprietary knowledge.
4. https://www.gartner.com/doc/3539717/pursue-citizen-data-science-expand

III. THE TRANSITION FROM MAN TO MACHINE

Traditionally, the considerable success of ML has relied on human ML experts to perform tasks such as data pre-processing and cleaning, feature selection and model construction, parameter optimization, model postprocessing and analysis. Today however, new ML algorithms can autonomously identify patterns, analyse data, and even interpret data by producing reports and data visualizations. Not only that, but these tools come boxed and wrapped up with an easy-to-use platform, providing an agility unlike that of the coding-heavy, statistical world of traditional ML methods. Much of the technical analysis work is now delegated to the machine.

The big tech players have, to varying degrees, capitalised on this trajectory through their 'machine learning as a service' offerings.

GOOGLE: BIGQUERY – ML WITH SIMPLE SQL

If you haven't heard of BigQuery, it is a massively parallel processing columnar storage data warehouse which Google offers as a service on their cloud. In English, this translates into a platform that scales to Big Data and can be leveraged by everyone in the organization. Within that, Google offers Cloud AutoML, where users can train high-quality custom ML models with minimum effort and ML expertise. Cloud AutoML was a massive leap forward for Google, and a huge appeal to any organisation interested in tapping into the power of deep learning without hiring a data scientist.

More recently, Google added a new capability to BigQuery by introducing BigQuery ML, a tool to build and deploy ML models through simple, broadly understandable SQL statements. Analysts can build and operationalize ML models on large-scale structured or semi-structured data, directly inside BigQuery, using simple SQL — in a fraction of the time.

AMAZON: SAGEMAKER – END-TO-END ML MODEL MANAGEMENT

Amazon Web Services seems to have more to say about end-to-end model management with SageMaker – a fully-managed platform that enables data scientists to quickly and easily build, train, and deploy ML models at any scale. The solution removes all the barriers that typically slow down developers who want to use ML by promoting a visual-centric approach to model development that integrates with MapReduce and other Amazon tools. At its core, SageMaker offers a rapid ML model generation environment whilst seeking to economically weigh ease-of-use against advanced technical capability.

MICROSOFT: AZURE & LOBE – DRAG-AND-DROP ENVIRONMENTS WHERE NO ML EXPERIENCE IS REQUIRED

Microsoft, meanwhile, is addressing data lineage and model management lineage as well as data governance with Azure ML.

Azure's ML Studio is a cloud-based environment that you can access from your browser and use to create ML-based models on any dataset of your choosing. Following the general trend of accessible ML, the unique selling point of this platform is to give a data scientist – without any prior ML experience – the ability to experiment with ML on datasets.

It requires prior knowledge of the R and Python programming languages as a prerequisite.

Other tools also exist to allow those who know the theory of deep learning but have no coding experience to just create a deep learning model within minutes without coding even a single line. Lobe (recently acquired by Microsoft) does exactly this. Lobe offers users a clean drag-and-drop interface for building deep learning algorithms from scratch, without having to know the ins and outs of libraries such as TensorFlow, Keras or PyTorch for example.

IV. SCOPE AND LIMITATIONS: PRACTICAL POINTS FOR ML PRACTITIONERS

Despite the investments the tech industry is making towards making ML more user friendly, there remain limitations of this 'self-service' model.

ONLY SIMPLE INFERENCE MODELS SUPPORTED

While BigQuery ML is arguably the best-designed, in-database ML stack, the current version supports only two types of models: (a) linear regression, which is used to predict numerical values, such as sales forecasts, and (b) binary logistic regression models, which can be used to do simple classification, like classifying loans as good or bad. Moreover, in both cases, the prediction mechanism as it stands cannot be used online and in real-time, as call-associated latencies are larger than those offered (and expected) by typical inference model solutions.

BEING ABLE TO EXPLAIN WHAT THE ML ALGORITHM DOES

Much has been made of 'explainable AI'[5]. This is the concept that we can only trust the results of a ML-generated algorithm if we understand how the algorithm produced them. Since ML does not require precise instructions on how to automate a task, and instead finds the best approach itself, the user needs to figure out the approach themselves. If the user doesn't understand the mechanics behind ML, they may struggle to do this.

5. https://hackernoon.com/explainable-ai-wont-deliver-here-s-why-6738f54216be

IV. SCOPE AND LIMITATIONS: PRACTICAL POINTS FOR ML PRACTITIONERS (CONTINUED)

OTHER TANGIBLE LIMITATIONS

The complexity of ML is not only about choosing the right algorithm, but also about selecting the right architecture (number of layers, number of nodes, sequence of algorithms, etc). For example, regularization and optimization are rarely one-step processes. Instead, they are methodical processes that often necessitate changes in the architecture of the model, such as incorporating new hidden layers and multiple parameter changes.

V. LESSONS LEARNED

With tactical advances being observed in AI technology, ML democratization is the current direction in which data analysis appears to be headed. Google's BigQuery is increasingly being selected by enterprises to drive their data warehouse modernization initiatives. The solution provides extreme scale and performance for organisations; but modernizing your data warehouse requires more than just computer horsepower and unlimited storage; analytics modernization is a journey fuelled by data.

As data science continues to emerge as a powerful differentiator across finance, almost every software platform vendor's goal should now be focused on making simplification through automation of various tasks.

CREATING A MORE PERVASIVE ANALYTICS-DRIVEN ENVIRONMENT

Due to the involved complexity and lack of resources, not all enterprises will be able leverage data science within their organisations, especially since access to data is currently largely uneven. What a lot of organisations do have plenty of are skilled analysts that could, with the availability of these new easy-to-use technologies, perform data analysis and create models to run predictive and prescriptive analytics. Being equipped with the proper tools should enable them to go beyond the analytics reach of regular business users.

The vast amount of analysis produced by CDSs will feed and impact the business/enterprise. Organisations will also have access to more data sources, including more complex data types, while developing a wider and more sophisticated range of capabilities across the firm.

Combined, these factors ought to eventually create a more widespread analytics-driven environment, empowering analysts throughout the organization, with a simplified form of data science. Simultaneously, data scientists can shift their focus onto more complex analysis.

V. LESSONS LEARNED (CONTINUED)

DATA GOVERNANCE AS THE KEY TO ACCESSIBLE MACHINE LEARNING

Interestingly, the launch of BigQuery and AutoML is complemented by the launch of Google Dataset Search, which includes Google's preferred guidelines for dataset providers to create metadata to support their datasets. The purpose of this 'governance' mechanism is to ensure that datasets that the algorithm searches for can be found and indexed. As more datasets are generated and uploaded onto the internet going forward, this ensures the algorithm doesn't become outdated.

This is a crucial element in implementing machine learning at enterprise scale. A parallel can be drawn with the Airbnb platform, which faced a dearth of available supply of rental apartments when expanding to Paris. To counter this, the platform bought apartments to then offer for rent. This spurred the usage of the platform to the point where the issue was resolved. In similar fashion, data has to be supplied constantly, comprehensively and consistently, for the machine learning platforms to generate business value. Data governance is one method of achieving this.

Private institutions with large quantities of data, such as banks, are investing a lot of money to achieve this. Most large banks are at such a scale that separate departments within the bank operate as independent units; across products, regions, and business functions. Since data-generated insights are generally most powerful when supported by data that includes everything, it is important for banks to find a way to be able to access their data holistically, despite the federated structure with which they operate. Since the operational structure cannot be readily changed, data governance must have an operational structure of its own.

Google's userbase and the opportunities for dataset providers that come with it means that the preferred guidelines will be successfully adopted. Passing the responsibility for maintaining a dataset to the dataset provider, which Google's approach does since it is in the interest of the dataset provider to be easily located and accessed, is the simplest method to ensure compliance. It is the lack of such an enforcement mechanism that is causing banks so much trouble.

Could banks adopt a similar approach to Google? To do that they would need to change their incentive structure to make data governance the responsibility of each employee. They will certainly need to address these concerns soon, as HSBC for example have signed a seven-year partnership with Google. The Google Cloud Platform, with its BigQuery and AutoML capabilities included, will likely be a part of that.

MAKE CULTURE YOUR ENABLER, NOT YOUR ROADBLOCK

To create significant business value out of data and analytics, organisations need not only technology. There is a key additional ingredient, that is, a special mindset, one which grasps the need and importance of strong analytics culture.

Transforming your environment into one which recognises data and insights as the catalyst towards more objective and efficient decision-making processes will mean that data and findings are more easily discoverable. The benefits then of using self-service technologies are immense, as insights are more easily shared with and systematically accessed by the right people.

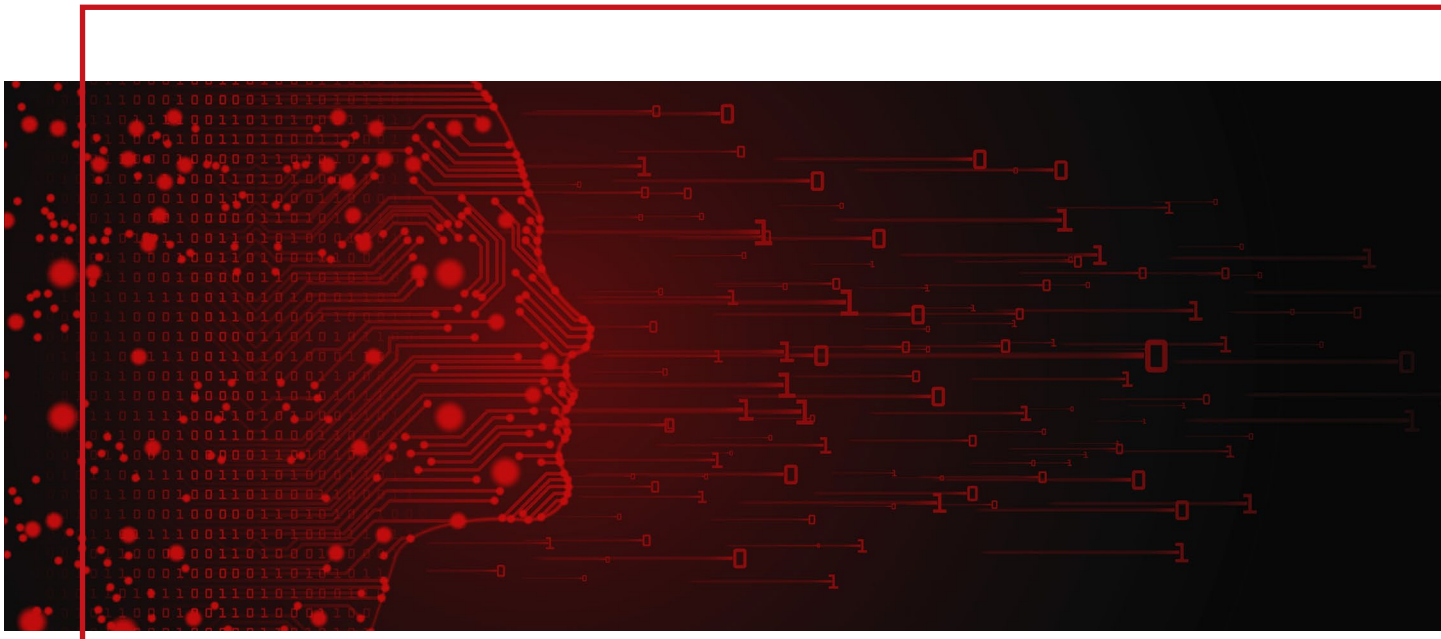
Having this cultural readiness to consume and interpret data will make teams more effective and eventually more successful in discovering business opportunities.

VI. CONCLUDING THOUGHTS

Banks have a mix of legacy systems running along more modern, virtualised applications, and like any large organisation, many will have managed a very large and diverse portfolio of suppliers for decades. With this dependence, the financial sector is one of the more cautious industries when it comes to adopting the cloud and associated data analytics tools.

Many banks nonetheless have recognised the need to establish the ability to use machine learning as a core competency within their organisations. For example, one of the world's largest banks, HSBC, is a much-vaunted Google Cloud Platform customer[6]. HSBC has adopted the use of the BigQuery data warehouse for real time anti-money laundering analytics and is now also running machine learning models on top of time series data. The bank is also adopting self-service tools, analysing data and generating models on platforms such as Amazon Web Services and Microsoft Azure.

Easy-to-use machine learning tools serve as a catalyst for integrating the technology at scale, retraining existing staff, and creating a working data governance process. The emerging ecosystem, consisting of marketplaces for data, tools, platforms, algorithms and computing infrastructure, will make it easier for organisations to maximise the benefits of machine learning. This increased accessibility will also be particularly powerful in several use cases (e.g. financial risk assessment) that have been traditionally cost prohibitive. With a lower barrier, and with banks starting to recognise the need to establish machine learning as a core competency, we foresee a strong future for machine learning within the financial services industry.



6. https://www.computerworlduk.com/cloud-computing/hsbc-looks-ramp-up-machine-learning-usage-with-google-cloud-3681316/

AUTHORS

Nadir Basma, Associate consultant
nadir.basma@capco.com

CONTACT

Jibran Ahmed, Head of Research & Development
jibran.ahmed@capco.com

ABOUT CAPCO

Capco is a global technology and management consultancy dedicated to the financial services industry. Our professionals combine innovative thinking with unrivalled industry knowledge to offer our clients consulting expertise, complex technology and package integration, transformation delivery, and managed services, to move their organizations forward. Through our collaborative and efficient approach, we help our clients successfully innovate, increase revenue, manage risk and regulatory change, reduce costs, and enhance controls. We specialize primarily in banking, capital markets, wealth and investment management, and finance, risk & compliance. We also have an energy consulting practice. We serve our clients from offices in leading financial centers across the Americas, Europe, and Asia Pacific.

To learn more, visit our web site at **www.capco.com, or follow us on Twitter, Facebook, YouTube, LinkedIn and Instagram.**

WORLDWIDE OFFICES

APAC
Bangalore
Bangkok
Hong Kong
Kuala Lumpur
Pune
Singapore

EUROPE
Bratislava
Brussels
Dusseldorf
Edinburgh
Frankfurt
Geneva
London
Paris
Stockholm
Vienna
Warsaw
Zurich

NORTH AMERICA
Charlotte
Chicago
Dallas
Houston
New York
Orlando
Toronto
Washington, DC

SOUTH AMERICA
São Paulo

WWW.CAPCO.COM

CAPCO
THE FUTURE. NOW.