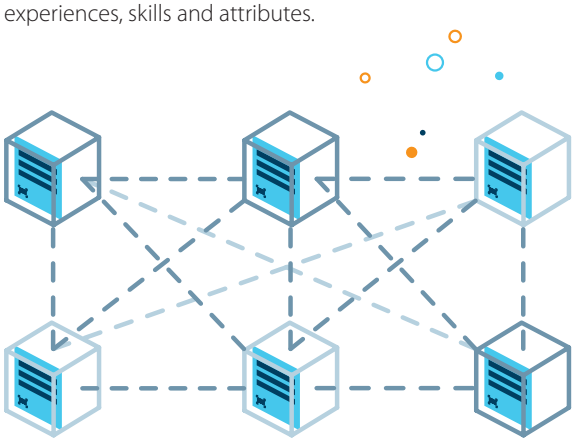
BUILDING A DATA CULTURE

From a business perspective, what does it mean to be data driven? What is a data culture? Why does it matter? What is wrong with the status quo? After all, businesses are surrounded by data and, with experience, surely intuition and gut feel deliver the same result?

Experience and intuition are valuable. What's different in a data driven culture is that insights, hunches, opinions and views are readily challenged, reinforced or enhanced by data. Then they can be leveraged across the entire team – with consistency and scalability. Moreover, a data driven culture allows a team to leverage a broader set of experiences, skills and attributes.



This article draws on the experiences of a 30-person sales team. It is written from a business perspective; not a data perspective. The aim isn't to map out a data strategy, but instead provide an overview of some of the data related challenges, opportunities and terminology that anyone building a data culture faces.

This article doesn't advocate for either a centralised or federated approach to data at an organisational level. What it does argue is that, irrespective of approach, the value your team will extract – along with its ability to manage data governance and ethics – is significantly greater if the team has developed a data culture. And it doesn't matter what size of the team or the type of business.

Our data build is more than two years old and still a work in progress. Encouragingly, though, we haven't had to wait two years for the benefits to appear. In fact, we have been surprised by both the speed and magnitude of results.

Why become more data driven?

For our team, becoming more data driven was the best way to deepen our understanding of our customers, improve engagement (internally and externally), identify automation opportunities, increase productivity and future proof our skills in a rapidly changing industry environment. After careful consideration of various alternatives, including going back for more product investment, we determined that a data driven approach was the most cost efficient way to grow both the business and the people in the business.

What have been the benefits of our data driven approach?

- Deeper understanding of customers and product relevance.
- Improved collaboration across the team, partly through a better understanding of the business, but also through developing and leveraging individual strengths and expertise.
- Considerably higher productivity.

What can't data do?

Data maximises the effectiveness of existing product but it can't replace product gaps. That said, it will help to better identify and prioritise those product gaps.

This article is set out as a user guide of sorts and is broken into three parts:

1. How to start the process;
2. The importance of team dynamics; and
3. Technology requirements.

We also share some of the key questions we needed to answer along the way and some lessons learned.

1. STARTING THE PROCESS

The first step has to be data governance. If you do not have a robust data governance and security framework in place, you'll need to create one.

With a governance framework in place, the best way to become more data driven without losing any real-time business momentum is to add specialised data capability to the team. Reskilling within the team is possible (and encouraged!), but is still best done working alongside specialist capability.

WHAT TYPE OF DATA RESOURCE?

A decade ago the role of data scientist didn't exist; now it's seen as indispensable to a data build. Moreover, as the understanding of data capability has evolved so has the range of data roles. As a result, adding data skills can get confusing.

To this end, it's important that data skills:

- ☑ Match the existing data maturity of the underlying business;
- ☑ Account for any internal resources that can be utilised, such as internal data science teams;
- ☑ Align with available technology resources; and
- ☑ Are consistent with your dominant objective (automation, insight or a combination of both).

Table 1: The Three Most Common Frontline Data Roles Today

Types of Frontline Data Roles			
Title	**Primary Focus of Role**	**When to Use this Role**	**Skills/Tools Required**
Data Analyst/ Visualisation Specialist	• Draw insights from existing data. • Proficient in data visualisation. • Reduced emphasis on technology aspects and machine learning. • More likely to come from a business analyst background.	• When data is reasonably well organised and requires little additional development. • Where the objective is to draw (mainly) visual insights from data rather than develop models or algorithms.	• Excel, pivot tables • Tableau, Qlik, PowerBI • Python and R chart libraries
Data Engineer	• The vast majority of data science is data collection, collating, cleaning and transformation. • Constructing end-to-end data pipelines including "Extract, Transform, Load" (ETL) pipelines. • Reliability and scalability testing. • Delivering models into production. • More likely to have a computer science or technology background.	• Where existing data sources are fragmented and/or missing. • Data is from a variety of sources (API, spreadsheet SQL). • Where the dominant objective is to scale the collection, organisation, cleaning and transformation of data. • Where data science progresses from research to deployment.	• SQL Database, Big Data • APIs • Cloud (e.g. AWS, GCP, Azure, Watson)
Data Scientist	• Operate at the intersection of statistics, mathematics and computer science. • Familiar with using both structured (e.g. numerical, categorical) and unstructured (e.g. voice, text, vision) data. • Should know the full suite of algorithm alternatives, ranging from simple linear regression to neural networks. • Often come from a quantitative engineering or computer science background.	• When the dominant objective is to analyse data to automate processes, discover new relationships, or generate new insights or predictions from data. • To assess different variables and compare different algorithms. • A data scientist will assess the underlying business requirements and help develop a data strategy, including presentation of results and input into product development.	• Python, R, et al • Machine learning, statistics, algorithms • Data pipelines • Data visualisation • Data storytelling • Cloud based machine learning tools

Spend some time deciding on the type of data resource as it could determine the success of your overall effort. Bringing in a data scientist who is ready to build machine learning models when the dominant requirement is to improve data visualisation isn't likely to end well. Neither is asking a visualisation expert to build an end-to-end data pipeline. The three most common mistakes in this space are:

1. Hiring a data scientist when what is really needed is a data analyst or engineer;
2. Trying to develop machine learning and automation without first implementing robust data pipelines; and
3. Having collected and cleaned data, not having the skills to actually deliver automation or insight.

Also keep in mind that there are several ways to access resources: hire a dedicated data resource; use existing internal data resources; or data consultants. The critical point is that they must match business requirements and to get value they must be immersed in the business for a sustained period of time.

WHAT ARE THE MOST IMPORTANT ATTRIBUTES OF A DATA RESOURCE?

Having identified the data capability that is best suited to your needs, the next step is to consider the key skills this resource should have. From a business perspective it's easy to be overwhelmed by the technical requirements of data roles, making it harder to differentiate between must-have and nice-to-have attributes. It's important to stress, though, that beyond a base level of core competencies, the non-data skills are still the most important.

Table 2: Top Five Attributes for Data Resources

Top 5 Data Resource Attributes	Why?
1. Communication/ Collaborations Skills	• The most important attribute of a data resource is the ability to work with the business; understand the business; understand the data relevant to the business. • Must be willing to educate, train and join the dots between data and the business but avoid ending up as the on-site IT help desk.
2. Visualisation/ Story-telling/ Data Evangelism	• Being able explain data, including through data visualisation, significantly increases the likelihood of success of a data driven strategy. • Knowledge of professional visualisation tools such as Power BI, Tableau or Qlik is a clear advantage (and also requires at least some low level programming skills). • A strong data scientist obsesses about data and its potential, which should be apparent. If not question whether they are the right resource to help drive a data culture.
3. Programming	• Two main programming languages used today in data science are Python and R (with SQL being a base language to access relational databases). • R is more widely used by those from a statistics background. • As a general purpose programming language, Python lends itself to better integration with APIs and big data analytics via the use of advanced Python libraries such as Tensorflow, Keras and PyTorch.
4. Ethics/ Governance	• Europe's General Data Protection Regulation (GDPR) has sharpened attention on data governance. • Model bias and explainability is a growing area of focus. • A data scientist must be able to demonstrate a keen awareness of data ethics, privacy, confidentiality and security.
5. Entrepreneur/ Problem Solver	• Ability to spot opportunities and help the team spot opportunities around the data. • Ability to solve problems; and not just of the Machine Learning variety. Probably more Mark Watney (The Martian) than MacGyver, but sometimes hard to tell.

The attribute that occupies a lot of attention (potentially too much) is coding capability. Although coding is important, a stand-alone data scientist scientist won't necessarily be preparing production ready code so it is more useful to focus on data discovery skills.

Moreover, a data scientist should apply the best approach for the problem at hand. That could be a simple linear regression or moving static excel charts into an interactive Tableau dashboard; or it could be a neural network. Be sure to test the breadth of a potential data resource's interest and experience.

2. TEAM DYNAMICS

Data driven culture is a nebulous concept - easy to spot when it's clearly there or it's clearly not; but it doesn't follow a consistent path. However, there are some consistent, immutable requirements:

- Leadership: tone and actions from the top.
- Team engagement: a lone data resource doesn't equate to a data driven culture.
- Persistence: there is no end date on development of a data driven culture.

And there are some consistent **DOs** and **DON'Ts**.
Our top ones are:

DON'T worry too much about use cases at the beginning – worry more about the data.

DO work with the entire team at the start of the initiative to complete a data inventory. That includes what data the business has, where it is stored, what is manually collected by the team, how it is collected, how accurate it is, what other data should be collected and how it can be collected in a way that is consistent, accurate and achievable.

DO take the time to build data pipelines and any additional governance requirements at the start. DON'T try and retro-fit.

DON'T think you can build a data culture by bringing in a data scientist and have them do their 'data thing' in isolation.

DO make sure that the team has clear and consistent messaging around the strategy and its requirements from the beginning. For us that has meant including data initiatives and engagement in performance objectives, including updates in every team meeting and regularly reiterating the rationale behind the data strategy.

This last **DO** is key for two reasons:

1. If a team is required to adjust their existing habits to collect or provide feedback on data and participation is patchy, data integrity will suffer, potentially to the point where the whole exercise is compromised.
2. If business teams were naturally inclined to immerse themselves in data, then they would already spend their time trawling through spreadsheets. Building a data culture isn't about forcing teams to spend more of their time immersed in data. It's about working with them to better capture and organise data in a way that allows automation of mundane, repetitive tasks. It powers their ability to do what they do best: generate insight from more complex relationships, engage customers and develop product.

HOW CAN THE TEAM GET INVOLVED?

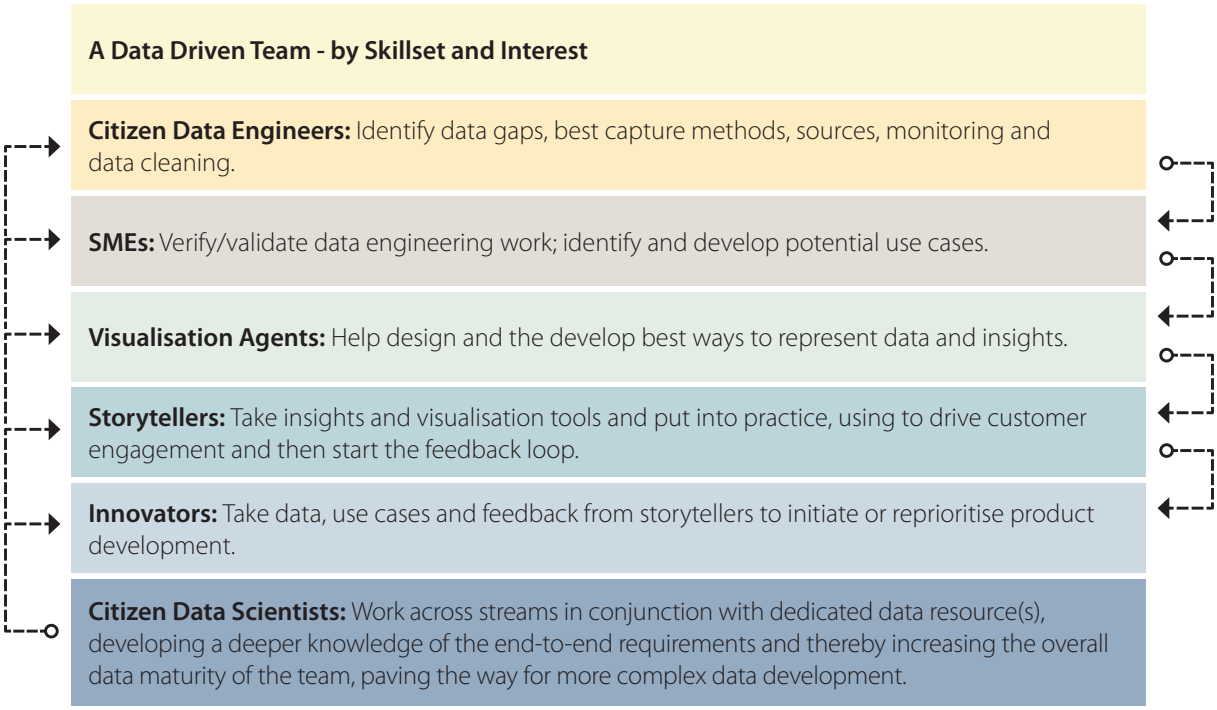
A data driven culture doesn't mean everyone needs to be a quant or a coding expert. There are a range of roles we've seen our team members gravitate towards, which allows them to link their individual strengths and skills to data initiatives.

- **Citizen data engineers:** These members of the team know the existing systems and data inside out. They are invaluable in collating, cleaning and validating the data.
- **Subject matter experts:** Those with deep industry and/or product expertise. These are the members of the team who can identify the best use cases for data.
- **Visualisation agents:** These are your PowerPoint champions. They are already good at presentations and can step it up a notch using data visualisation packages like Tableau, PowerBI or QlikSense.
- **Storytellers:** These team members are naturals at taking insights and visualisations and using them to convey insight to customers.
- **Innovators:** Will take insights and use them to develop new product or improve existing product.
- **Emerging data scientists:** As coding capability becomes more widespread, more team members will develop the ability to design algorithms and generate new insights.

THE TRICK IS IN BRINGING OUT THESE ATTRIBUTES AND HELPING TEAM MEMBERS UNDERSTAND THAT A DATA DRIVEN CULTURE REQUIRES A BROAD RANGE OF SKILLS AND INTERESTS.

From a practical perspective, we have found that the in-team data roles listed above are complementary and often display a natural sequencing. The advantage of this is that different team members will be more involved in a data build at different times, which helps ensure a good balance between the build and real time business momentum. One way to think about it is via swim lanes, with each skill able to progress in parallel, but at a different pace.

DIAGRAM 1: A DATA DRIVEN TEAM



3. TECHNOLOGY

Once you start to scratch the surface of the technology requirements, it's easy to understand why data engineers are in such high demand. While growth in Automated Machine Learning (AutoML) improves access to model development, feature engineering and parameter tuning, the underlying architecture of a data build requires specialist skills.

Internal resources available across technology and dedicated data science teams will determine the degree of specialist skill and investment required. Larger firms often already have a lot of the required technology. The objective here is not to prescribe an architecture, nor recommend a platform or product, but instead provide an overview of some of the more common terms and considerations that emerge.

One point to note is that whereas everything described to date requires full engagement of the team, technology is something that should be dealt with separately.

There are three main areas to navigate.

DATA STORE

All data has to be stored somewhere. In general, the choice is between:

- Traditional, on premise (on-prem) physical servers.
- Cloud (often referred to as Infrastructure as a Service (IaaS) or Network as a Service (Naas)), which includes:
    - Public cloud with shared, on-demand access that is the mainstay offering of most cloud providers today, like Amazon AWS, Microsoft Azure and Google.
    - Private cloud hosted by either the cloud provider (with logical separation of your data) or on premise.

The advantage of on-prem is security, through physical separation of the data from the outside world. The downside is that it is harder to scale and, from a data science perspective, lacks access to many of the value added tools that come with modern cloud infrastructure. The advantage of cloud is the associated services, such as Software as a Service (SaaS) and Platform as a Service (PaaS). Hybrid cloud/on premise solutions are growing, as are initiatives by cloud providers to supply add-on services within private environments.

As data storage and use explodes, this area will remain fast changing. In most cases, though, this infrastructure decision will be determined at a firm level. The point is to engage with the relevant experts in your organisation to ensure that a functional, scalable data plan is put in place at the outset.

In parallel with the decision around the location of data store is the type of data store. This generally falls into two camps:

1. Relational database (e.g. SQL). These are the standard table based, column and row databases defined by a prescribed schema. They are the mainstay of many data stores today.
2. Non-relational (or no SQL) databases. Common types are:
    - Document based: containing all the data for an entity in a flexible manner that a traditional SQL based schema would not allow;
    - Graph based: nodes and edges mapping relationships such as organisational charts;
    - Time series: allowing for quick retrieval of high volume time series data; and
    - Object based: for images, video and audio.

Given their flexibility and efficiency, non-relational databases continue to experience rapid growth. Data sources are growing rapidly so even if your data today is standard, structured data, it may not be in the future. After completing your data inventory, it's worth sitting down with the experts in your business and mapping out what your data needs might look like over the course of the build.

Finally, you also need to think about whether your data requirements will be on a live basis (i.e. updated in real time) or can be processed in batches at discrete intervals. It will depend on the data use case and degree of data maturity already in the business.

As a business lead, data technology is not something you will need to manage but it will be important to work with your in-house experts to map it out early in your build.

CODING

When it comes to coding, it's easy to get lost amongst the terminology. The two most common code languages used in data science are Python and R. Both are open source i.e freely available. One of the main strengths of a language is the related libraries that support it. A library is a code package that has been written to enable or enhance functionality. In Python, some of the more popular packages are NumPy, Pandas, SciPy, Keras,Tensorflow and PyTorch; and in R, dplyr, tidyr, caret. Libraries tend to support easier data manipulation or different types of machine learning algorithms but there are hundreds of them, designed for almost every task required. There is also a tremendous open source community supporting ongoing development, which is one of the reasons data has become such a powerful driver of business performance.

Aside from code language, the main decision required is the type of environment. Integrated Development Environments (IDEs) enable better code management via library, data and file management; code completion and/or editing (to help finish code, detect code errors or identify the best functions); and in more advanced, professional environments, AutoML, code testing and deployment. There are two main types of IDE to choose from:

1. Open source: For Python the most popular environment is Jupyter Notebooks, but there are many others, all with slightly different features (Google Colab, Spyder, Visual Studio). R-Studio is the most widely used R environment (but there are also many other good environments).
2. Professional: Offered by all the main cloud providers (Amazon, Microsoft, Google), plus Cloudera, H20, Dataiku and DataScience, just to name just a few. The main points of differentiation amongst professional data science platforms (beyond cost) are whether they are cloud only or available on-prem; ease of data connectivity and transformation; visualisation capability; and, increasingly, access to enhanced libraries for a range of tasks, including Natural Language Processing (NLP) and AutoML.

As with technology, work with specialist teams to identify the best tools for your business objectives.

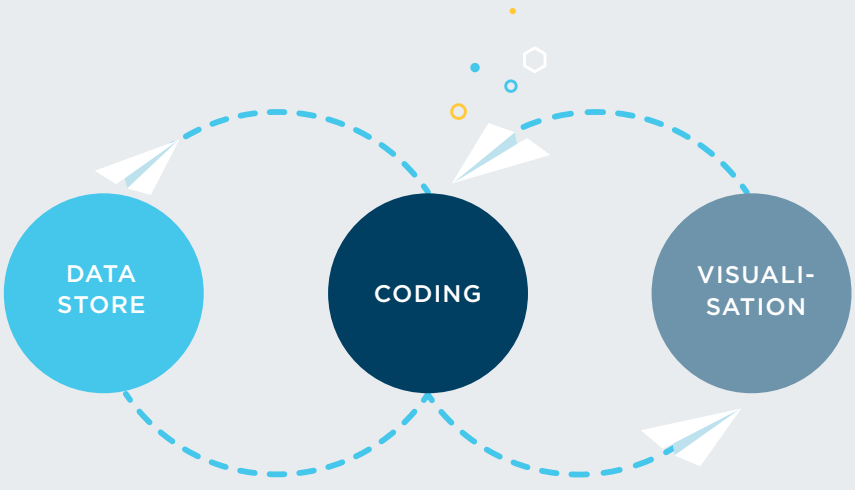
VISUALISATION

It is difficult to overstate the importance of data visualisation, particularly in allowing teams to engage with the data. Strong visualisation conveys a story or identifies relationships between data and is a critical tool in building a data driven culture. There are two main types of visualisation tool:

1. Open source: There are some powerful open source visualisation tools, such as ggplot2 and shiny (in R) and matplotlib, seaborn and bokeh (Python). When combined with widgets they can also provide an interactive experience.

2. Professional: These are available via database providers and data science platforms or via dedicated visualisation tools such as PowerBI, Qlik Sense or Tableau.

The main differences between dedicated visualisation platforms and open source options are scalability, access control, ability to interact and drill down within a dashboard, as well as add-ons such as the ability to distribute dashboards (e.g. web or email).



NEXT STEPS

Most businesses have well entrenched processes for product development. Fewer are as experienced at supporting data led decision processes. Stakeholder education is a significant part of building a data culture. Here are some of the questions and misconceptions that can emerge.

What will the data show?

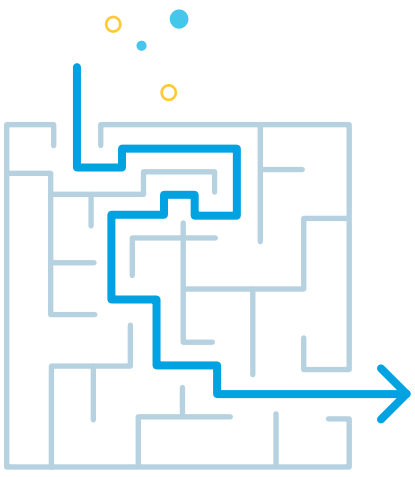
It's an unfortunate truism that before investing in data, many businesses want to know what it will show. The whole point of data is discovery: discover more about customers; more about business processes; and new product opportunities. And sometimes the data will not reveal anything, which is ok too.

Data first or use case first?

There is balance required here. Too often businesses come up with use cases and then when the data doesn't support that use case, the data is questioned and data culture suffers. Alternatively, without spending enough time collecting relevant data, potential use cases narrow to the point of being inconsequential.

At the other extreme, it's easy to keep collecting more and more data and producing more and more dashboards without spending enough time thinking about how the data can be used to automate or add insight. This is dangerous from two perspectives. Firstly, teams will tire of collecting data unless it's clear why it is valuable. Secondly, endless dashboards create 'just more MIS', limit engagement and constrain willingness to develop use cases. The best approach is:

1. Make sure the relevant data is identified and collected;
2. Introduce visualisation early in order to help the team gain early insight; then
3. Start to map out use cases and refine as required.



Where to start?

When it comes to creating a data culture it's important to generate a shared strategic vision, but in implementation, incrementalism is best. As data maturity develops so too will the tools that are used. It's understandable that most businesses would prefer to proclaim the progress they have made developing AI and machine learning algorithms rather than celebrate the fact that their team has spent the past year collecting, cleaning and organising data. But the simple fact is, without doing the latter, there is next to no chance of successfully delivering the former.

Growth or efficiency?

Growth and business efficiency shouldn't be mutually exclusive. Efficiency is often the primary catalyst for starting a data initiative but this shouldn't be the case. In a data driven culture, the combination of data plus team capability lead to both growth and automation opportunities and this needs to be made clear throughout the stakeholder engagement process.

Automation or augmentation?

One of the most common sources of pushback within teams on data/AI/machine learning is the misconception that automation leads to fewer people in the business. The reality is that tasks will be automated and roles will transform. Driving a data culture means being able to convince a team that it, if executed well, their repetitive, cumbersome tasks will be automated and they'll have more time to drive insight, add value and get closer to customers.

IN OTHER WORDS, A GOOD DATA STRATEGY MAKES THE MEMBERS OF A TEAM MORE VALUABLE, NOT LESS.

LESSONS LEARNED

Navigating our data build has been difficult at times. If we could time travel back two years and give ourselves some advice, this is what we would have changed.

1. Spend more time up front on the architecture of the build. Encouragingly, our build has been achieved (re)using existing technology and platforms but we could have reduced the manual workload in collecting, updating and analysing our data pipelines if we had spent more time and effort mapping out the data pipelines first.
2. Introduce data visualisation as early as possible. It's hard to overstate the importance of good data visualisation as a way of developing team engagement and involvement in data.
3. Progress from data visualisation to machine learning (or statistical methods) as soon as practicable. While data visualisation is a great way to explore data and convey insights, it's important to remember that, by design, they often reflect preconceptions about the data. To really unlock discovery, new insight and automation, machine learning or more advanced statistical methods are invaluable.

It is also tempting to add a fourth lesson –that we should have added more dedicated data resources earlier on in the process to amplify the benefits delivered. But the flip side is this may have compromised team engagement if existing team members weren't brought along for the ride. There is no doubt that, given a choice now between more data resources and additional product, data resources would come first (in part because we have learnt how better data leads to better product development). But it can't be at the expense of team engagement.

Development of any team culture is complex. The advantage of developing a data culture is that if executed well, it will deliver better outcomes: for customers, for the business and for the teams involved. There is no doubt a good data strategy, executed well, drives business and productivity growth: and the benefits can be large. But it requires persistence and a willingness to try different approaches.

With growth in data science, data scientists, open source data science tools, cloud infrastructure and related SaaS, it has never been easier to grow a business through data. But it is more immersive than a traditional product led approach. Over the past two years we have found that the key requirements are a team approach and working on the basis of constant, incremental change. Think evolution, not revolution.

GOOD LUCK!

anz.com
Australia and New Zealand Banking Group Limited (ANZ) ABN 11 005 357 522. Item No. 97141 07.2020 WX246492