




The Definitive Guide to Machine Learning for Business Leaders

by Hugo Bowne-Anderson

The Difference Between AI, Machine Learning, and Data Science



The business world is overloaded with buzz terms like artificial intelligence, machine learning, AI transformation, deep learning, and data science. We know that these fields, technologies, and tools are changing the competitive landscape across verticals and are soon to become more table stakes and foundational than disruptive. However, it's possible to know they're important but not understand what they really mean. If you're confused, that's understandable as these are all loaded terms and they're not even used consistently. My goal here is to dispel any confusion by demystifying these questions: What is artificial intelligence, machine learning, and data science? Where do they intersect and where do they diverge?

Defining Artificial Intelligence, Machine Learning, and Data Science

Artificial Intelligence (AI) is a “a huge set of tools for making computers behave intelligently”¹ and in an automated fashion. This includes voice assistants, recommendation systems, and self-driving cars.

Machine Learning (ML) is the “field of study that gives computers the ability to learn without being explicitly programmed”². The lion’s share of machine learning involves computers learning patterns from existing data and applying it to new data in the form of making predictions, such as predicting whether an email is spam or not, whether a customer will churn or not, and diagnosing a particular piece of medical imaging.

Data Science (DS) is about making discoveries and creating insights from data, and communicating these insights and discoveries to non-technical stakeholders.

How are these related?

Machine learning feeds into both artificial intelligence and data science

If the output of your machine learning model is fed into a computational system that performs an action in an automated fashion—such as recommending a movie, decelerating a self-driving car, or serving search results—it can be viewed as a component in your AI system.

If the output of your machine learning model is fed into a human decision making process, it can be considered data science work. For example, when predicting a customer may churn results in a human deciding to incentivize the customer to stay, this insight or discovery informs a data science decision.

Much of machine learning and artificial intelligence relies on high quality data, meaning the most impactful and effective artificial intelligence will stand on the shoulders of robust data science capabilities.

¹ Andrew Ng, co-founder of Google Brain and former Chief Scientist at Baidu.

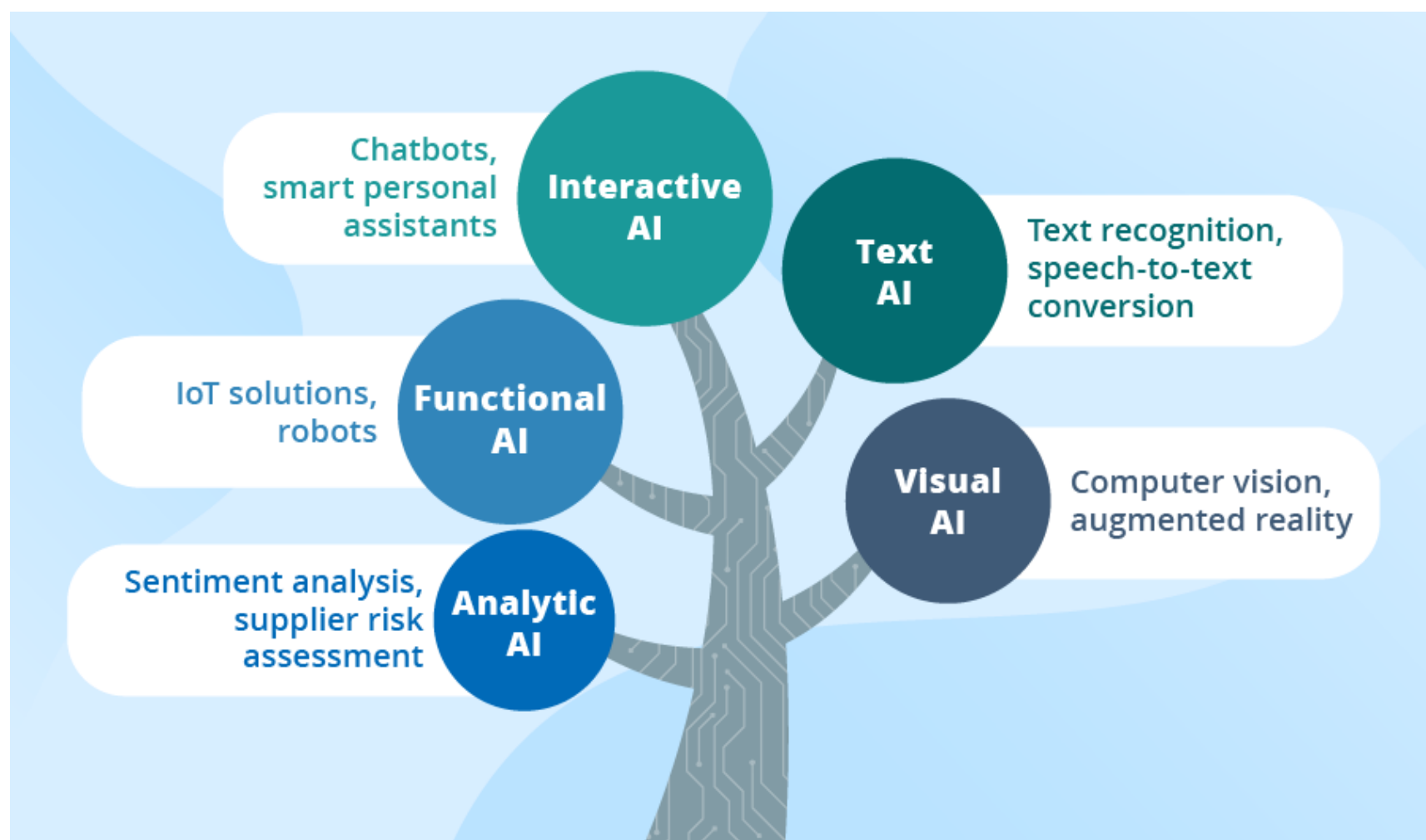
² Arthur Samuel, pioneer in artificial intelligence and computer gaming.

Artificial Intelligence

In his Coursera course [AI for Everyone](#), Andrew Ng, co-founder of Google Brain and former Chief Scientist at Baidu, defines artificial intelligence as “a huge set of tools for making computers behave intelligently.” This definition casts a wide net and it’s worth providing some examples to make clear what “behaving intelligently” means:

- Voice assistants, such as Siri
- Recommendation systems, such as Netflix
- Self-driving cars
- Drones that fly over fields and capture footage to optimize crop yield
- Google Search
- Surfacing algorithms, such as those employed by Twitter and Facebook, that decide what content to show you in your feed

ScienceSoft, an IT consulting company, provides a useful breakdown of the types of artificial intelligence.



Source: [ScienceSoft](#)

It's important to recognize that artificial intelligence means that actions and decisions are automated. It's also key to note that all of these are examples of artificial narrow intelligence, or algorithms that can do one thing well. This is not to be confused with artificial general intelligence, which is a hypothetical, futuristic artificial intelligence that can do anything a human is capable of. Nor is it a superintelligent artificial intelligence, a hypothetical software agent whose intelligence surpasses that of humans.

Both artificial general intelligence and superintelligent artificial intelligence are a long way off, if at all possible, and serve as distractions for real, present, and necessary conversations around the capabilities and limitations of artificial intelligence as we know it today, resulting in headlines such as [An AI god will emerge by 2042 and write its own bible. Will you worship it?](#) This hypothetical intelligence is clearly absurd and distracts from all the current examples of artificial intelligence that allow computers to perform tasks that mimic aspects of human intelligence, such as recognizing stop signs and people in images and videos (self-driving cars), holding basic conversations, retrieving information, performing tasks (voice assistants), and ranking text documents based on their relevance to a particular query (Google Search).

If artificial intelligence is a huge set of tools, what tools are we talking about? Let's explore the tool of central importance to modern artificial intelligence—machine learning.³

"Artificial intelligence is not to be confused with artificial general intelligence, which is a hypothetical, futuristic artificial intelligence that can do anything a human is capable of."

³ Note that there are parts of artificial intelligence that do not leverage machine learning: for instance, logical programming and automated reasoning (which can be used for proving mathematical theorems).

Machine Learning

Machine learning powers recommendation systems, content discovery, search engines, email spam filters, and many other “matching” problems in tech. In healthcare, it’s being leveraged for drug discovery and high throughput diagnostic imaging diagnosis. In finance, machine learning is now foundational for fraud detection, process automation, algorithmic trading, and robo-advisory. In retail, Walmart is at the forefront of using machine learning to reinvent supply chain management. The list goes on. So what actually is it?

Machine learning was a term popularized in 1959 by Arthur Samuel, a pioneer in artificial intelligence and computer gaming. Samuel defined it as the **“field of study that gives computers the ability to learn without being explicitly programmed.”**

Supervised learning

The majority of machine learning involves computers learning patterns from existing data and then applying it to new data in the form of making predictions. Examples include:

- Predicting whether a customer will churn
- Predicting whether an email is spam or not, given the email sender, subject, and body
- Predicting the diagnosis of a particular piece of medical imaging
- Predicting outcomes of sports games

These prediction and classification problems are the most important ML techniques for business leaders to know in the short and medium term—referred to as supervised learning. The pattern of the label that you’re trying to predict—such as spam or not—is said to supervise the learning process.

The power of modern machine learning rests firmly on having good quality data for your algorithm to learn from or be “trained on,” and that such training data needs to be labeled. In the spam classification example, you’ll need many emails labeled with whether they were spam or not. In the diagnostic imaging example, you’d require at least thousands of images labeled with their diagnosis.

Then, your ML algorithms are able to pick up patterns in your training data and generalize those patterns to unlabeled data, where you don't know the outcome that you're trying to predict. It's for this reason that mathematician and Stanford Professor [David Donoho prefers the term recycled intelligence](#) over artificial intelligence for machine learning, as no new intelligence is created, whereas human intelligence, as captured by humans with domain expertise hand-labeling datasets, is recycled and re-applied to new data.

There is a huge and hidden supply chain behind the worlds of machine learning and artificial intelligence. Companies and individuals leverage services such as [Amazon Mechanical Turk](#) to crowd-source labeled data. And [Scale AI](#), a start-up that works with tens of thousands of contractors worldwide to hand-label data, recently raised \$100 million, which speaks to growing market needs.

Unsupervised learning

Unsupervised learning is about discovering general patterns in data. The most popular example is clustering or segmenting customers and users. This type of segmentation is generalizable and can be applied broadly, such as to documents, companies, and genes.

Anomaly detection is usually formulated as an unsupervised learning problem. Anomaly detection algorithms help you pinpoint which observations are out of the ordinary or very different from others. For example, if you have a dataset with a lot of transactions and would like to identify which ones may be fraudulent.

Of course, most data comes unlabeled. This makes unsupervised learning useful in every domain, and it's only grown in importance over the years. There is a lot to be learned from the data in an unsupervised manner. An expert in the field, Yann LeCun, famously compared machine learning to a cake. In his analogy, reinforcement learning is the finishing touch—the cherry on the cake. Supervised learning gives the data some extra depth—it's the icing on the cake. Unsupervised learning is the most substantial part of understanding data—it's the cake itself. Too many people don't perform enough unsupervised learning and apply supervised learning blindly.

In some domains like text and image processing, we have massive amounts of data to learn from, but the catch is that they are usually unlabeled. An effective way to mine this data is to use self-supervised learning techniques. You can think of self-supervised learning as the missing link between supervised and unsupervised learning. In self-supervised learning, humans don't explicitly label the data. Instead, intrinsic labels are added to the data creatively by considering the problem domain. For example, with natural language processing, massive sets of text-based documents are collected from the web, and for each sentence, the middle word is replaced by a blank. It's then used as a label. Finally, a model is trained to predict the middle word from its context. Self-supervised learning enables models to learn intricate patterns without the need for human-labeled data.

Reinforcement learning

Although the vast majority of artificial intelligence and machine learning relies on labeled data and recycling intelligence contained therein, there is a growing sub-field of machine learning called reinforcement learning that relies far less, if at all, on pre-existing training data. With reinforcement learning, which draws on behavioral psychology, software agents are placed in constrained environments and given “rewards” and “punishments” based on their activity. If playing games sounds like a relevant application of reinforcement learning to you, you're spot on: it was how [AlphaGo Zero became the world Go champion in 2017](#), beating AlphaGo, which was trained on human data. More recently, in 2019, [Pluribus beat the best professional players in six-player no-limit Texas Hold'em poker](#). Reinforcement learning also has meaningful applications in other fields like self-driving vehicles and algorithmic trading, and we'll definitely see more applications in coming years.

Is Machine Learning a Form of Artificial Intelligence?

So is machine learning a form of artificial intelligence? It is commonly regarded as a form of artificial intelligence, but if we're thinking of artificial intelligence as a "set of tools for making computers behave intelligently," then ML becomes one of these tools. For example, the Google spam filter is an example of an AI system, and the ML algorithm that classifies a given email as spam or not is one component of this artificial intelligence. Another component is the software that pushes emails classified as spam by the ML algorithm into your spam folder.

Now that we've got a handle on artificial intelligence and machine learning, let's see what data science is all about.

Data Science

In their seminal 2012 Harvard Business Review article [Data Scientist: The Sexiest Job of the 21st Century](#), Thomas Davenport and DJ Patil state unequivocally that "more than anything, what data scientists do is make discoveries while swimming in data." Data science is about creating insights from data, often in a business setting. How do data scientists do this, though? Data science is a multidisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.

How does this play out in practice? There are so many tools and techniques in a modern data scientist's toolbox that it's helpful to partition the space. One way to slice the data science space is into descriptive analytics, predictive analytics, and prescriptive analytics.

FOCUS AREAS FOR DATA SCIENCE AND ANALYTICS



Descriptive analytics

Also called business intelligence (BI), descriptive analytics is essentially about getting the right pre-existing data in front of the right people, typically in the form of dashboards, reports, or emails. This can include both past and real-time time data about revenue, customer engagement, churn, or company and employee performance.

Predictive analytics

Predictive analytics is synonymous with machine learning and is the realm of predicting the future, such as whether a customer will churn or not, and more general classification tasks: Is an email spam or not? Is a tumor in a diagnostic image benign or malignant? Data scientists might also be interested in why a model makes a prediction, not just the prediction itself. A black-box model, or a model that is created directly from data by an algorithm and therefore doesn't reveal its inner workings to humans, might not be as interesting to them as highly interpretable models. There are also key ethical and regulatory considerations to requiring models to be interpretable.

Prescriptive analytics

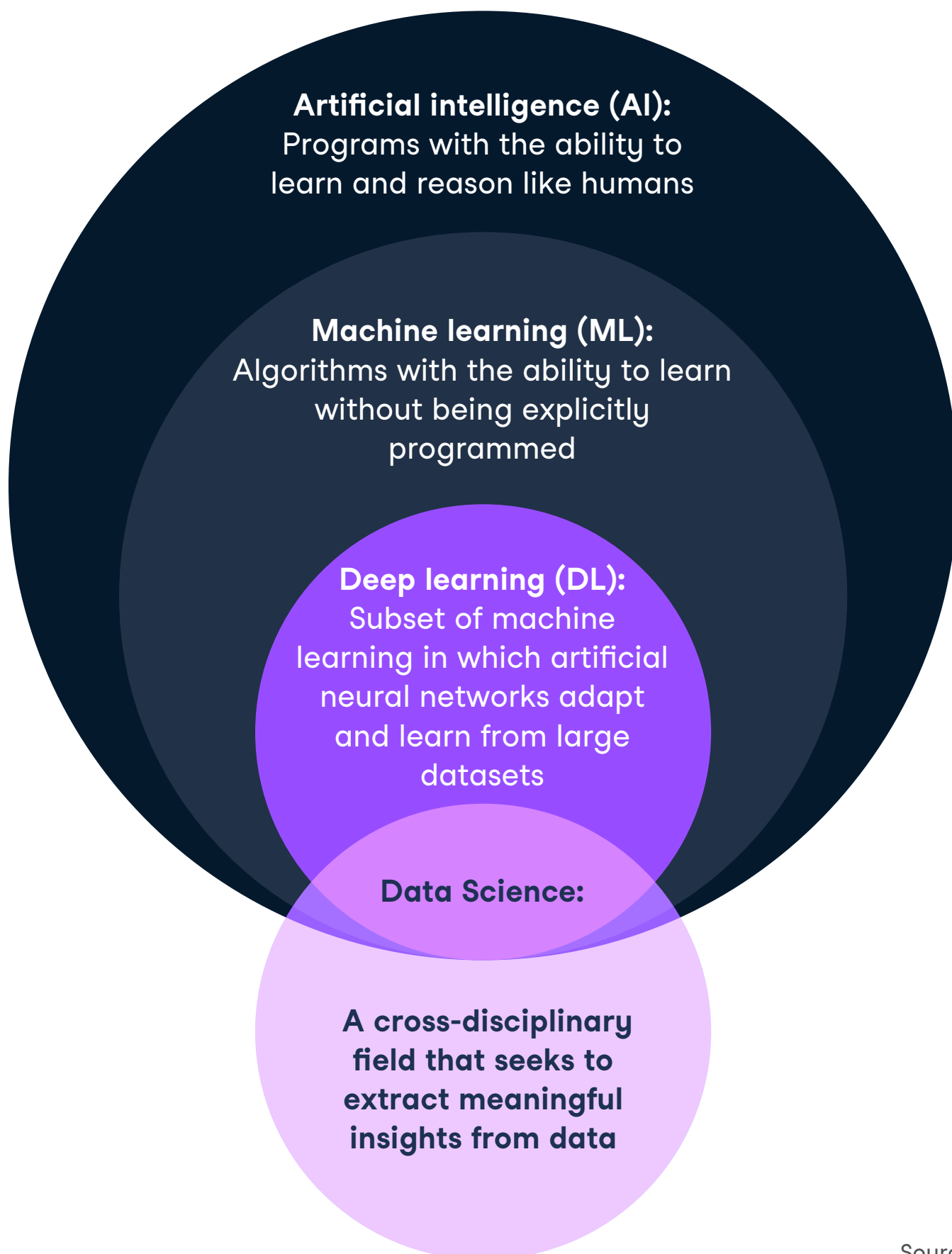
Prescriptive analytics is the realm of decision science and decision analytics, and is concerned with how to make decisions based on data. If, for example, your machine learning model tells you that a particular customer will churn, you don't yet know what to do about it. Prescriptive analytics is concerned with facilitating seamless collaboration between people working with data and those making business decisions—or finding a path from data to insights. Exciting spaces to watch are [data translation](#) (a burgeoning field for those with both domain expertise and technical know-how), advances in reinforcement learning (which bleeds into machine learning, as we've seen), and decision science—especially the work of Cassie Kozyrkov, Chief Decision Scientist at Google Cloud, whose work we discussed on [DataFramed, the DataCamp podcast](#).

So is machine learning part of data science or part of artificial intelligence? And what is the relationship between data science and artificial intelligence? As discussed, these terms are used in a variety of inconsistent ways, but a good rule of thumb is this:

- If the output of your machine learning model is fed into a human decision making process, it can be considered data science work. For example, when predicting a customer may churn results in a human deciding to incentivize the customer to stay, this can be considered insights or discoveries made from the data.
- If the output of your machine learning model is fed into a computational system that performs an action in an automated fashion—such as recommending a movie, decelerating a self-driving car, or serving search results—it can be viewed as a component in your AI system.

The main distinction between artificial intelligence and data science is that although many of the tools, techniques, infrastructures, and processes are the same, data science is often fed into human decision-making processes while artificial intelligence is concerned with automation. However, remember that much of machine learning and artificial intelligence relies on high quality data. This means that the most impactful and effective AI strategies will stand on the shoulders of robust data science capabilities.

THE RELATIONSHIP BETWEEN ARTIFICIAL INTELLIGENCE, MACHINE LEARNING, DEEP LEARNING, AND DATA SCIENCE



Source: corpnce.com

Note that there may be overlaps between these fields not shown in the diagram. If the output of your machine learning model is fed into a human decision-making process, it can be considered data science work. If the output of your machine learning model is fed into a computational system that performs an action in an automated fashion, it can be viewed as a component in your AI system.

CHAPTER 2

Machine Learning For Business Leaders

In chapter one, we discovered that machine learning, deep learning, and artificial intelligence are buzzwords for good reason—these technologies are fundamentally shifting the nature of business, society, and our lives. More importantly, across many verticals, they're shifting from being disruptive technologies to being foundational and table stakes for businesses to remain competitive.

The power of machine learning across verticals

Now, let's look at several examples of machine learning's impact across various verticals.



Tech

Machine learning powers recommendation systems, content discovery, search engines, email spam filters, and matching problems.



Healthcare

Machine learning facilitates drug discovery and diagnostic imaging diagnosis.



Finance

Machine learning is now foundational for fraud detection, process automation, algorithmic trading, and robo-advisory.



Retail

Machine learning is reinventing supply chain management by optimizing supply and demand planning, improving shipping processes and reducing transportation expenses, and improving strategic sourcing.



Other

Burgeoning industries are growing rapidly with machine learning.

- LegalTech is imagining a future in which machine learning is leveraged to predict outcomes of court cases based on natural language analysis of precedents.
- AgTech (agriculture technology) is deploying drones at scale to capture footage, and machine learning is being used to estimate crop yields.

The power of machine learning across teams

There are also many gains in the development of ML algorithms that are vertical-independent, supporting different business functions.



HR

Machine learning helps filter applicants in the hiring flow—but hiring models must be carefully monitored so as not to [perpetuate social biases at scale](#).



Support

Machine learning is used for call center routing and chatbots.



Marketing

ML algorithms are used for paid advertising, customer churn prediction, and targeted nurture campaigns.

In fact, any company that has an app can benefit from leveraging machine learning to determine the most effective push notifications, and any organization that has a website can leverage machine learning to personalize their customer experience by surfacing content and features that are most relevant.

Machine learning improves collaboration and helps teams become more nimble in the way they conduct business.

The 10 Machine Learning Commandments for Business Leaders

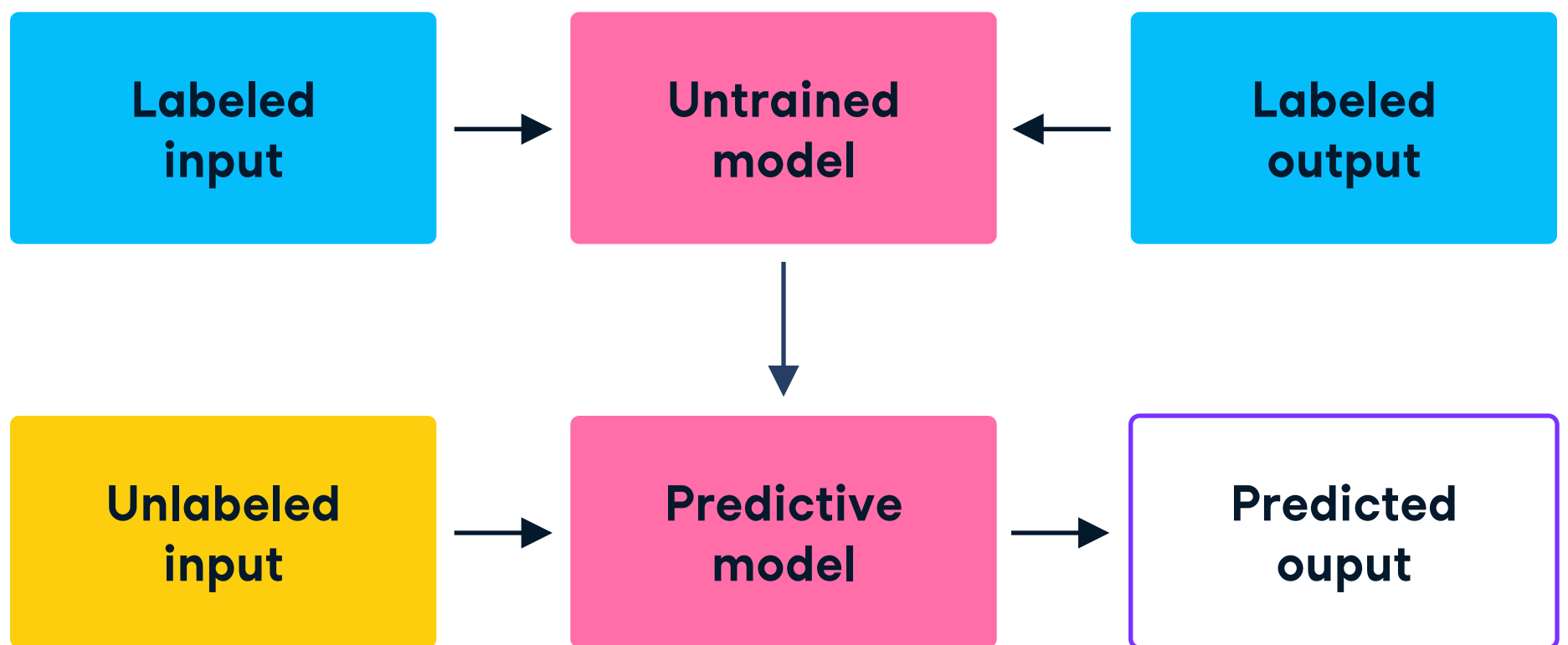
According to [Gartner's Annual Chief Data Officer Survey](#), poor data literacy is the second-biggest internal roadblock to success for chief data officers. Gartner expects that by 2020, 50% of organizations will lack sufficient artificial intelligence and data literacy skills to achieve business value, and 80% of organizations will initiate targeted data literacy initiatives to overcome deficiencies. The data is clear: to keep your competitive advantage, you'll need to leverage machine learning in one form or another. So as a business leader, what do you need to know about it?

1 Embrace the paradigm shift of models learning from data

In fact, any company that has an app can benefit from leveraging machine learning to determine the most effective push notifications, and any organization that has a website can leverage machine learning to personalize their customer experience by surfacing content and features that are most relevant.

In Software 1.0, or code that is written by a human, you would write code that explicitly specifies that, were a tumor above a given size and of a certain texture, among other conditions, it would be classified as malignant. Below a certain size, it's classified as benign. Andrej Karpathy, the director of artificial intelligence and Autopilot Vision at Tesla, [defines Software 2.0](#) as “code written by an optimization function based on an evaluation criterion.”

In Software 2.0, you specify the types of algorithms you want to use and feed them labeled data, that is, images that have already been classified as benign or malignant, and the algorithm discovers patterns in these in order to generalize the classification to new, unlabeled data. As mentioned in chapter one, this training data is often hand labeled by humans and, for this reason, researchers such as David Donoho prefer the term recycled intelligence to artificial intelligence because the machine is merely recycling the human intelligence contained in the labeled examples, and not creating any new forms of intelligence.



Source: [EBC](#)

"The machine is merely recycling the human intelligence contained in the labeled examples, and not creating any new forms of intelligence."

2 Choose your evaluation metric with care

In addition to labeled training data, you need to supply a ML model with an evaluation metric, which tells the algorithm what you're optimizing for. One commonly used evaluation metric is accuracy, that is, what percentage of your data your model makes the correct prediction for. This may seem like a great choice: who would want a model that isn't the most accurate?

Actually, there are many cases where you wouldn't want to optimize for accuracy—the most prevalent being when your data has imbalanced classes. Say you're building a spam filter to classify emails as spam or not, and only 1% of emails are actually spam (this is what is meant by imbalanced classes: 1% of the data is spam, 99% is not). Then a model that classifies all emails as non-spam has an accuracy of 99%, which sounds great, but is a meaningless model.

There are alternative metrics that account for such class imbalances. It is key that you speak with your data scientists about what they're optimizing for and how it relates to your business question. A good place to start these discussions is not by focusing on a single metric but by looking at the [confusion matrix](#) of the model, which contains:

- False negatives (e.g., real spam incorrectly classified as non-spam)
- False positives (non-spam incorrectly classified as spam)
- True negatives (non-spam correctly classified)
- True positives (spam correctly classified)

CONFUSION MATRIX

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

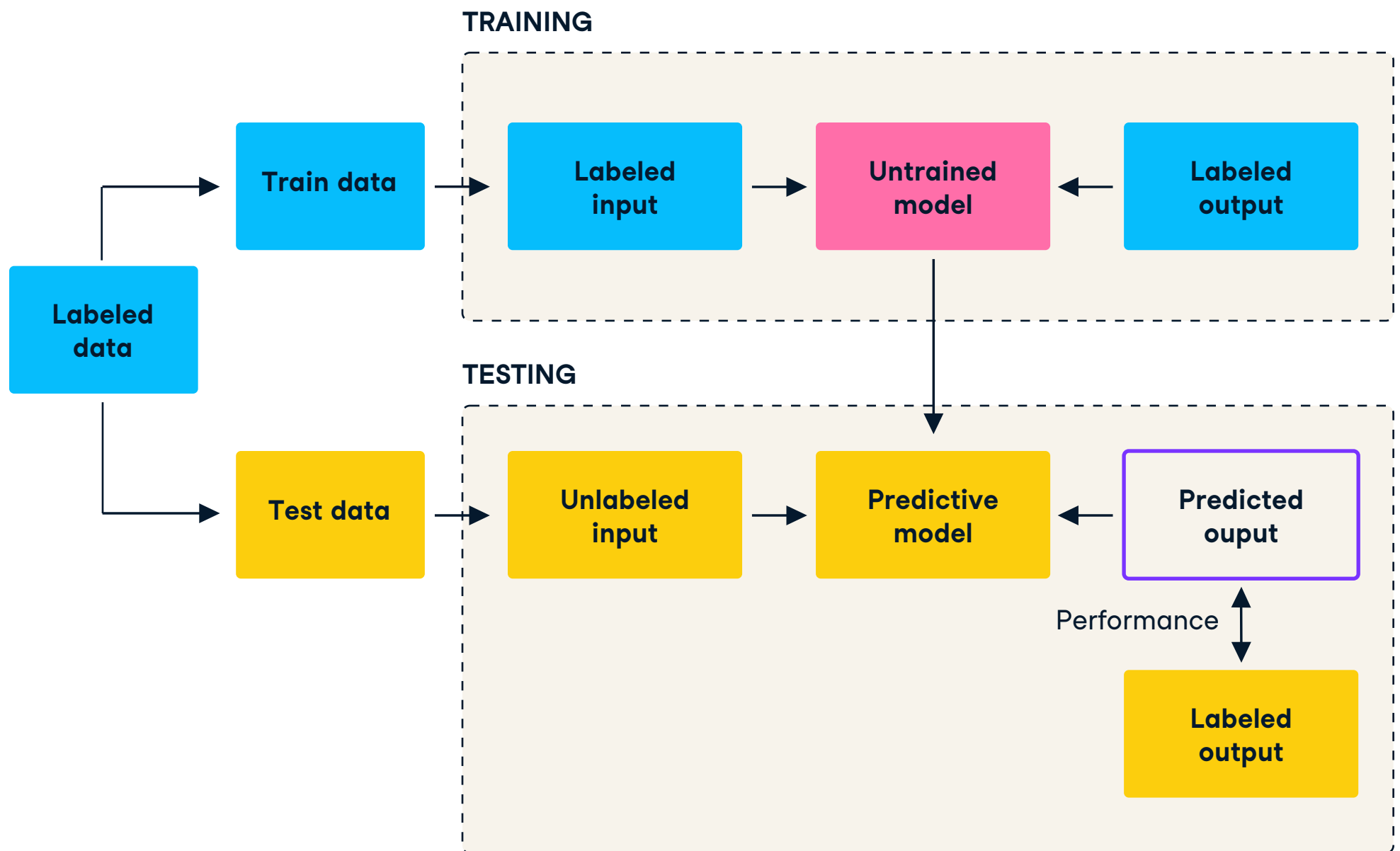
A lot of attention is currently focused on the importance of the data you feed your ML models and how it relates to your evaluation metric. YouTube had to learn this the hard way: When they optimized for revenue based on view time (how long people stay glued to videos), this had the negative effect of recommending more violent and incendiary content, [along with more conspiracy videos and fake news](#).

An interesting lesson here is that optimizing for revenue—since viewing time is correlated with the number of ads YouTube can serve you, and thus, revenue—may not be aligned with other goals, such as showing truthful content. This is an algorithmic version of Goodhart's Law, which states: "When a measure becomes a target, it ceases to be a good measure." The most well-known example is a Soviet nail factory, in which the workers were first given a target of a number of nails and produced many small nails. To counter this, the target was altered to the total weight of the nails, so they then made a few giant nails. But algorithms also fall prey to Goodhart's law, as we've seen with the YouTube recommendation system.

"An interesting lesson here is that optimizing for revenue—since viewing time is correlated with the number of ads YouTube can serve you, and thus, revenue—may not be aligned with other goals, such as showing truthful content."

3 Remember to split your data

The attentive reader may be asking: How do we calculate the accuracy of the model if we've already used all our labeled data to train it? This is a key and crucial point: If you train your model on a dataset, you'd expect it to perform better on that data than on new data. To get around this, before even training the model, you can split the data into a training set and a test set—this procedure is called train test split. We split the data into a training and test set so that we can estimate the model on the training data and evaluate its performance on the test data. This prevents overfitting, which occurs when a model tries to predict a trend in data that's too noisy, and makes the model more generalizable.

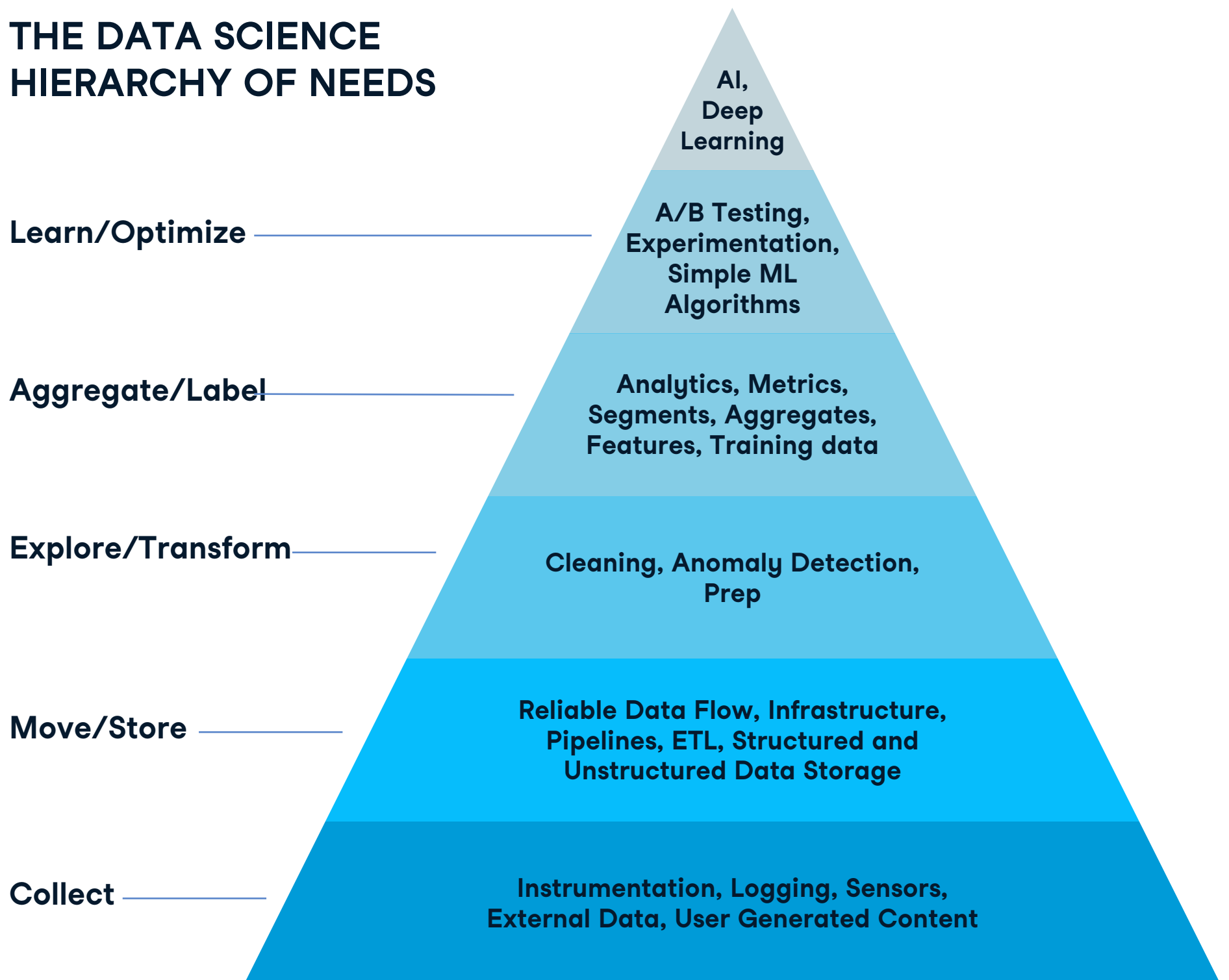


Source: [EBC](#)

4 Focus on solid data foundations and tooling

Having good quality data is a huge challenge in itself! This is why when executives ask me how they can make their companies AI-driven, I respond by showing them Monica Rogati's [AI Hierarchy of Needs](#), which has machine learning close to the top as one of the final pieces of the puzzle. This hierarchy illustrates that before machine learning can happen, you need solid data foundations and tools for extracting, loading, and transforming data (ETL), as well as tools for cleaning and aggregating data from disparate sources.

THE DATA SCIENCE HIERARCHY OF NEEDS



Source: [Hackernoon](#)

5 Beware of bias in your data and algorithms

Machine learning can only be as good as the data you feed it. If your data is biased, your model will be too. For example, [Amazon built a ML recruiting tool](#) to predict the success of applicants based on resumes with ten years' worth of training data that favored males due to historic male dominance across the tech industry—which caused the ML tool to also be biased against women. As [Cassie Kozyrkov has analogized](#), a teacher is only as good as the books they're using to teach the students. If the books are biased, their lessons will be too.

6 Pry open the black box of your model

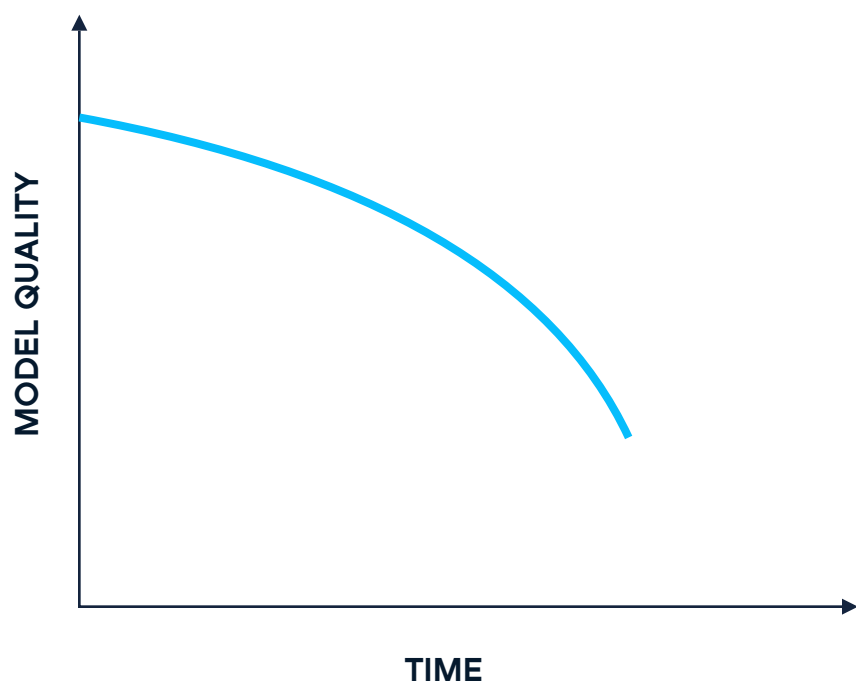
We've seen that accuracy—the percentage of your data that your model predicts or classifies correctly—is not always the best metric to measure the success of your model, such as when your classes are imbalanced (for example, when 99% of emails are spam and 1% non-spam). Another space where metrics such as accuracy may not be enough is when you need your model to be interpretable.

Interpretability is the characteristic of being able to say why your model makes the predictions it does, which is necessary for many models deployed in financial markets due to regulation. It is also essential for algorithms that impact stakeholders' lives, such as Northpointe's [COMPAS recidivism risk model](#), which is used by judges to make decisions during parole hearings. There's an inherent trade-off between accuracy and interpretability, in that the most accurate models are generally black box and not interpretable since they're created directly from data by an algorithm—so even the humans who design them can't understand how variables are being combined to make predictions. Interpretability is an important part of the conversation to be aware of even if your particular industry isn't concerned with it at the moment, since regulation will bring it to many other industries in the coming years.

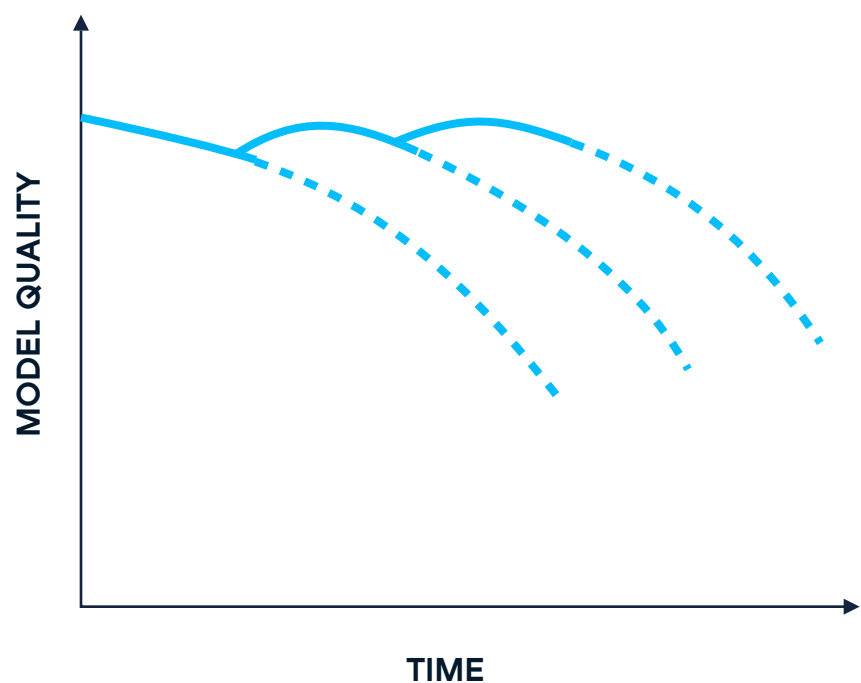
7 Keep tabs on your model and improve it

Remember that the job of machine learning doesn't end when your model is in production, making predictions, or performing classifications. Models that are deployed and doing work still need to be monitored and maintained. If you have a model predicting credit card fraud based on transaction data, you get useful information every time your model makes a prediction and you act on it. On top of this, the activity you're trying to monitor and predict—in this case, credit card fraud—may be dynamic and change over time. This process, where data that's generated is constantly in flux, is called data drift—and it proves how essential it is to regularly update your model.

Static models



Refresh models



Source: [DataBricks](#)

8

Delve into applications for deep learning

Deep learning is a form of machine learning that uses models called artificial neural networks, which are loosely inspired by biological neural networks in human brains. As we’ve noted in chapter one, this is the extent to which the analogy holds—deep learning is not equivalent to human intelligence. And deep learning only applies to artificial neural networks that have several hidden layers—not all artificial neural networks perform deep learning. This distinction is important because, in practice, very deep models require several parameters to be tuned well. Problems can arise when an artificial neural network becomes too deep, like exploding or vanishing gradients and the amount of storage required for the model and its huge dataset.

Many deep learning applications occur in the supervised learning world, in the form of image classification (self-driving cars, drone footage utilized to estimate crop yield in AgTech, facial recognition⁴).... There are other applications in time-series prediction, such as financial prediction problems. Deep learning systems are rarely good at more than one task: an algorithm that is built for self-driving cars will not be any good at classifying legal documents. Although you may like to call deep learning a form of artificial intelligence, it is only so in the sense of narrow artificial intelligence.

Deep learning doesn’t always require labeled data. Deep learning for natural language processing or image classification often works in a self-supervised way. For instance, NLP systems are trained on bodies of raw text where it is trained to predict the middle word from the context around it. It is self-supervised in the sense that it makes its own labels by taking out words and marking them as labels.

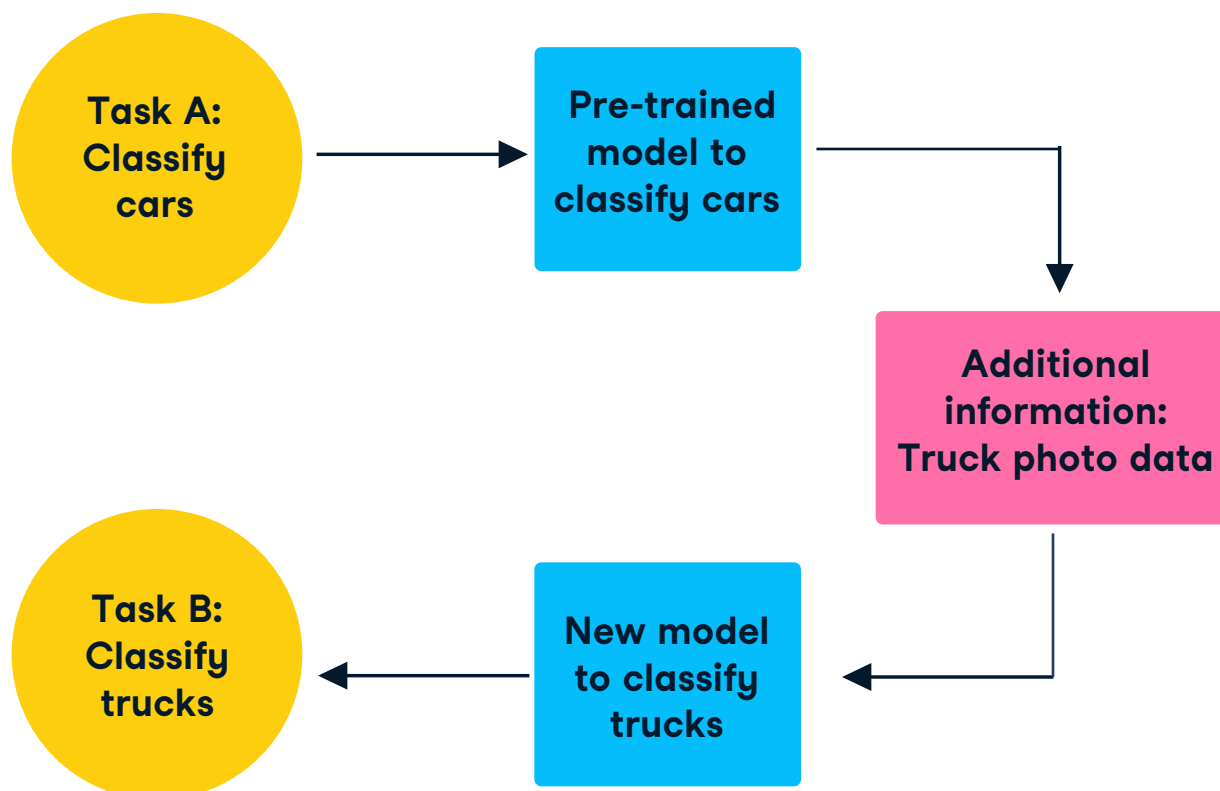
Examples of deep learning

IMAGE CLASSIFICATION	NATURAL LANGUAGE PROCESSING
Self-driving cars	Google translation
Drone footage	Document classification
Facial recognition	Sentiment analysis

⁴I personally don't see any productive use cases for facial recognition that don't introduce [far more challenges than they solve](#) and we need to think as a society whether we want to use them at all.

9 Explore pre-trained models with transfer learning

In a world where building competitive ML models relies on state-of-the-art, domain-specific data, you might be concerned about not having enough data yourself or the ability to collect it. But much of the future of machine learning will involve using pre-trained models, or models already trained on other data. This is the world of transfer learning. For example, you can buy a pre-trained image classification model that recognizes and classifies cars. With transfer learning, you can fine-tune pre-trained models for your particular question and domain. For example, if you want an algorithm that classifies trucks, you could take one that classifies cars and train it further on truck photo data.



We are currently seeing the emergence of algorithm marketplaces, such as [Booz Allen's Modzy](#), where you can buy and sell pre-trained models. There are key concerns, however, such as how to think about data and algorithmic bias, along with model governance, when trading algorithms without having access to the datasets that they're trained on. The space is ripe for growth, but it's also ripe for abuse and regulation.

10 Embed machine learning into your decision making

The final point that I cannot stress enough is that machine learning—and all data work—needs to be directly embedded in the decision function. Machine learning doesn't exist in a vacuum, it's there to serve decision making. This can be automated (as is Google Search), embedded in a scientific process (as when an algorithm flags an MRI for a specialist to look at), or embedded in organizational processes (such as when decisions are made around what to do with customers who are predicted to churn).

You want to make sure that the data work always reflects your real-world concerns and that you avoid [Type III errors](#), where you get the right answer but to the wrong question. This is why the data translation space is heating up and why it's so important to establish a culture of data work in your organization. This will require the workforce to understand what data science and machine learning can and can't do, along with how to ask good questions from data professionals. The goal is to establish healthy, productive lines of communication between the data function and the rest of the company at large.

Is a given drug effective for treating a particular disease?


TYPE I ERROR (FALSE POSITIVE)	TYPE II ERROR (FALSE NEGATIVE)	TYPE III ERROR (CORRECT ANSWER, WRONG QUESTION)
An ineffective drug is recommended to treat a particular disease.	An effective drug is rejected as a treatment for a particular disease.	An effective drug is recommended to treat an unrelated disease.

To recap, the 10 machine learning commandments you should follow are:

- 1 Embrace machine learning as a new paradigm of software development in which your models learn from data.
- 2 Choose your evaluation metric (e.g., accuracy or revenue) with care. Remember to split your data into training and test sets.
- 3 Focus on solid data foundations and tooling upon which to build your machine learning initiatives.
- 4 Watch out for any bias in your data that may produce biased algorithms.
- 5 Pry open the black box of your ML models, especially if you're in a regulated market like finance.
- 6 There's an inherent trade-off between accuracy and interpretability, but interpretability—the characteristic of being able to say why your model makes the predictions it does—is becoming more important.
- 7 Monitor and maintain ML models that have been deployed to prevent data and model drift.
- 8 Delve into applications for deep learning for your business, which can be very good at specific supervised learning tasks like image classification and natural language processing.
- 9 Explore the growing field of transfer learning, which allows you to retrain pre-trained models—especially if you're concerned about collecting enough appropriate data for a specific business problem.
- 10 Tie machine learning and all data work to decision making. This will surface real-world concerns and establish productive lines of communication between the data function and the rest of the company at large.

CHAPTER 3

Five Things Business Leaders Need to Know About Data Strategy



In the first two chapters, we demystified buzz terms like artificial intelligence, machine learning, and data science, and took a deep dive into what business leaders need to know about machine learning. Now let's look at what business leaders need to know about their data strategy.

The five things that business leaders need to know about data strategy are:

- 1 The 80/20 rule for data science.
- 2 Big data ain't all that big.
- 3 The future of data work isn't just in coding—it'll include even more point-and-click.
- 4 Data culture is an important piece of your data strategy.
- 5 Data strategy requires considering the different people who are impacted.

1 The 80/20 rule for data science

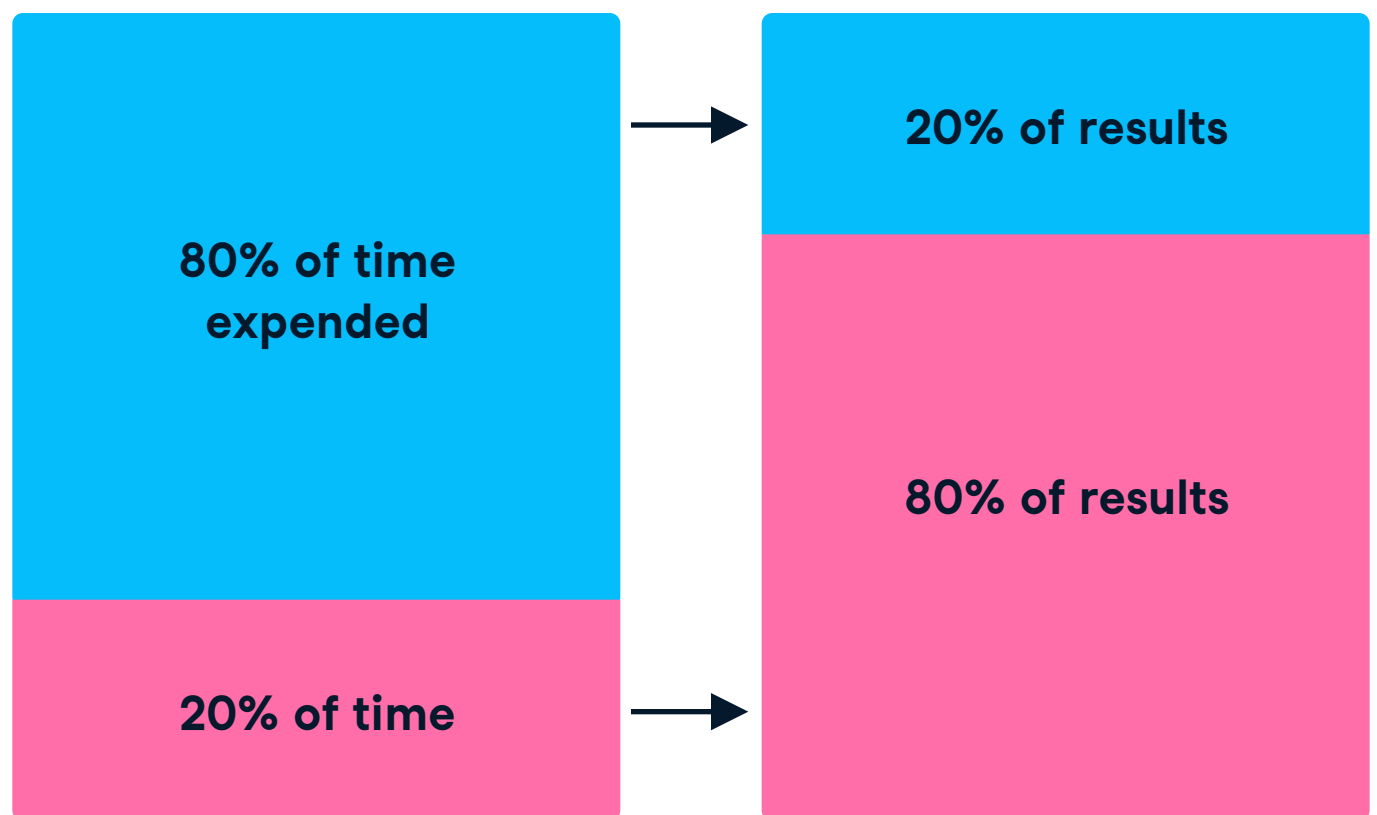
Focus on the 20% that matters

Vilfredo Pareto (pictured below), while thinking about land ownership in Italy, discovered that 80% of land was owned by 20% of people. This 80/20 rule, also known as the Pareto principle, states that for many events, roughly 80% of the effects come from 20% of the causes.



Many trivial tasks

Few vital tasks



For example, in 2018, roughly 80 to 90% of taxes in the US were paid by the top 20% of earners. It isn't always 80/20, though: 1% of Wikipedia users generate 99% of their content.

We've observed with the companies we work with that when applying the Pareto principle to data science, work done in 20% of the time generates approximately 80% of the results. So when prioritizing work in your data strategy, you should focus on what really drives value.

To be clear, this might not be work where you see an immediate return on investment. For example, setting up a robust data infrastructure is essential work that doesn't drive immediate return, but will end up demonstrating results.

It is key to recognize that we're not talking about large and sexy wins for data involving the hottest deep learning, artificial intelligence, or machine learning. We're talking about what has business impact and what can really move the needle. Let's look at some examples.

Unify your data

In any organization, data is often siloed in different departments. If you have customers who are serviced by different departments—for example, the marketing team tracks customers in one database and the customer support team tracks them in another—there are often all types of inconsistencies. Customer names or addresses may be stored using different conventions. Developing a unified data source will move the needle a great deal for future data initiatives. This essentially comes down to data infrastructure.

Create accessible dashboards

Generating data views and dashboards so that everybody in your organization has access to the data in the way that they need it can be very impactful.

Build customer-centric processes

Call center routing and customer churn models can be impactful, as can sales propensity models, which will be incredibly important for your sales and marketing teams.

Leverage conversational AI

The growing field of conversational AI can have a huge impact even in its most basic form of simple chatbots for customer support. This example is not only a huge win for customer service and support—it's also something you can pilot internally and iterate on before releasing into the wild.

Make sure your data is actionable

Recall that an instructive way to think about data work is by breaking it down into descriptive analytics, predictive analytics, and prescriptive analytics.

A huge amount of the value that data can create comes in the form of descriptive analytics—if it's strongly tied to the decision function. So when thinking about the 20% of work that can deliver 80% of the value for most businesses, it won't be through AI—it will be through descriptive analytics and getting the right data in the hands of the right people. That's taking data the company already has and getting it to the right people in whatever form they can consume it and utilize it, typically via dashboards, reports, and emails. As we'll see later, tying the decision function to this data generation and descriptive analytics function is also a huge cultural challenge, as is making sure that people actually use it.

The purpose that data scientists serve in organizations is to analyze data to facilitate answering business questions. Ideally, the decision function is tied to the data function as tightly as possible.

When I discussed this with Renee Teate, Director of Data Science at HelioCampus, [on the DataFramed podcast](#), she framed it as follows: We want to go from a business question to a business answer, and we want to factor this through a data question and a data answer. Your stakeholders, managers, and C-suite executives will likely not be concerned with the data question on the data answer. What they want is the business answer with logic to support how you arrived there. This is what it means to tie the data function to the decision function.

CALL TO ACTION

List three to five projects based on descriptive analytics that could inform your decision making, and then order them in terms of what projects could have the greatest business impact.

2 Big Data Ain't All That (Big)

Defining big data

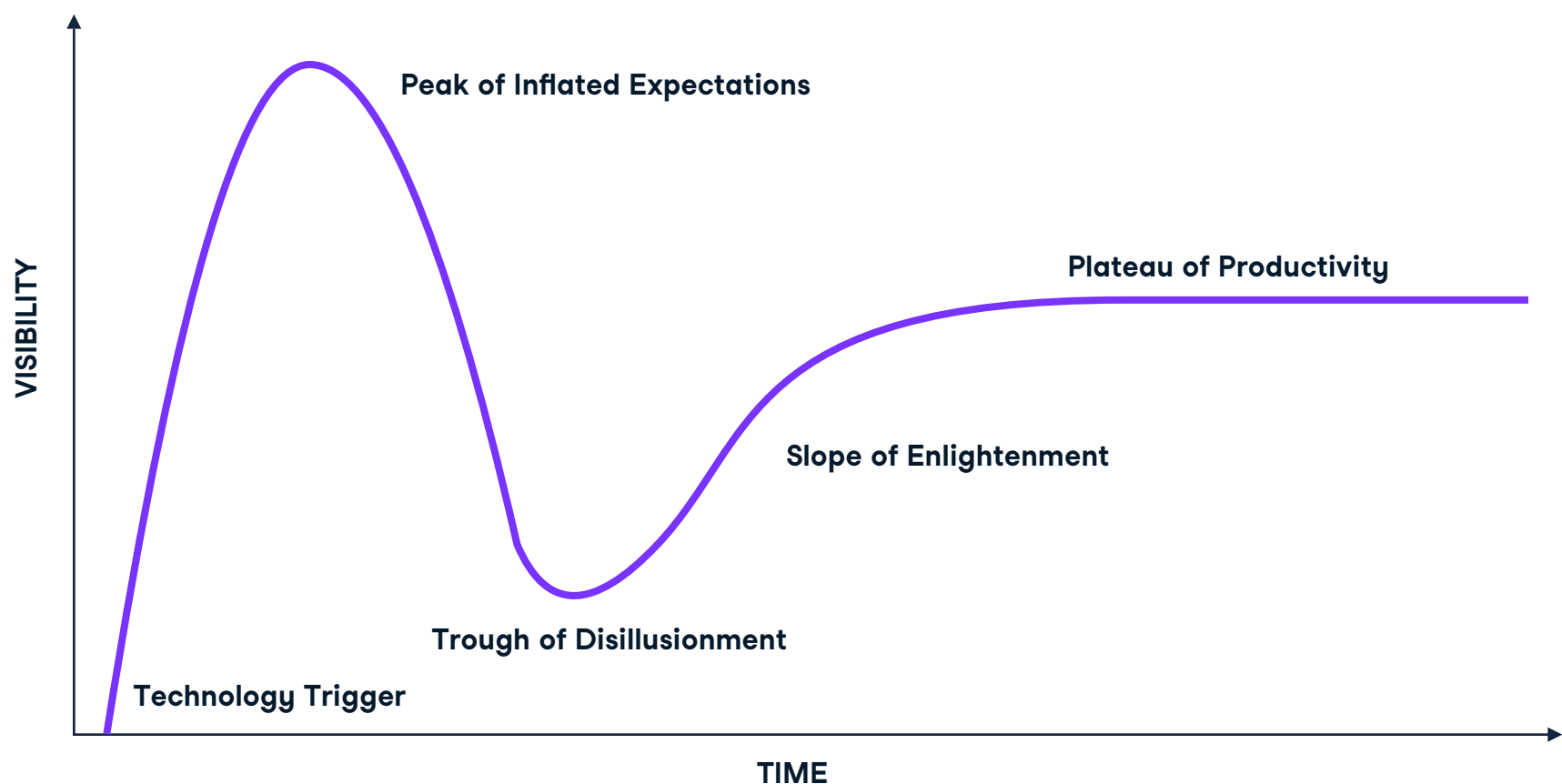
Big data is one of the buzziest words in the data space. But just how big is big? One way to think about it is in terms of volume, or the amount of data you have. Based on volume, we can define big data as data that is far more than you can store on a modern laptop or hard drive. Thus, it needs to be distributed across clusters of computers to work with, transmit, and analyze. We can extend this definition of big data to cover [the three Vs](#): volume, velocity, and variety.

Is big data the end of theory?

In 2008, Chris Anderson wrote a provocative article in Wired called [The End of Theory, The Data Deluge Makes the Scientific Method Obsolete](#). The premise was that we had enough data to make satisfactory predictions about the world that we didn't need theory to understand the world. Part of the impetus for such arguments was what we were seeing happen with Google—they were able to operate on huge amounts of data and then provide predictive models in the forms of data products for programmatic ad buying. Moreover, they were able to do so with sufficiently advanced predictive analytics without needing to understand or theorize about the system under study.

This was all about using captured human behavior in order to build better products and services. Google enabled those who wanted to buy an ad to easily purchase one on Google AdWords, and the model behind AdWords didn't rely on any theory behind whether someone would click or not. Google had just enough data to make a “good enough” prediction. Chris Anderson's provocative hypothesis is that big data contains so much information that we don't need to model the world anymore, and we don't need to understand the theory behind it or what's actually happening.

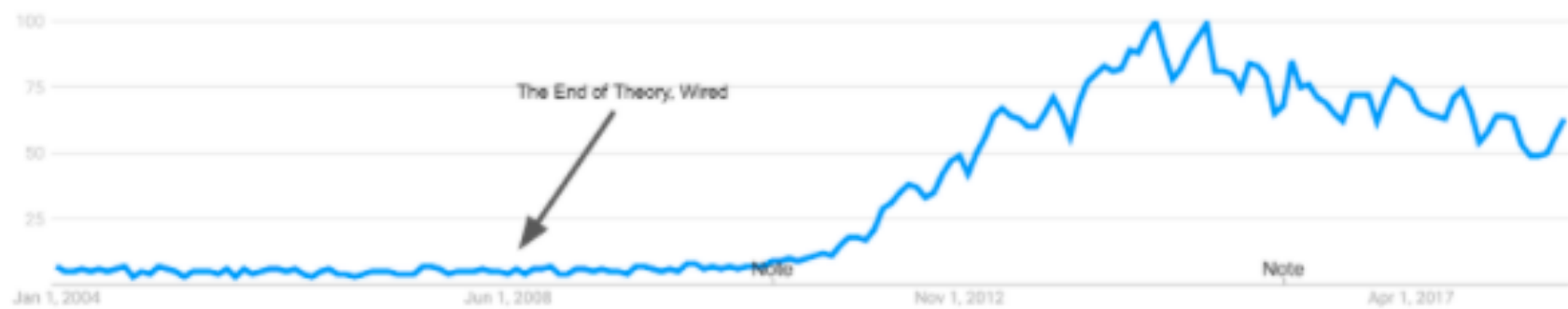
Has big data actually fulfilled its promises to us? One way to think about this is modeling it on the Gartner Hype Cycle.



Source: [Wikipedia](#)

The Gartner Hype Cycle tells us about a technological innovation and the expectations around it as a function of time. We begin with an Innovation Trigger, which in the case of big data was the ability to store, transmit, and analyze large amounts of data. Then people get buzzed about it, leading to Inflated Expectations. After that, we don't see the value delivered against expectations, so we enter the Trough of Disillusionment. Only after this do we see the actual real value start to be delivered across several different verticals—and we enter the Slope of Enlightenment.

Where is big data currently in the Gartner Hype Cycle? One way to think about expectations is to see what people have searched for on Google as a function of time. Below are the Google trends for “big data” since 2004. You can see that Chris Anderson was ahead of his time—but he was wrong about big data being the end of theory due to the importance of small data and thick data.



● big data
Search term

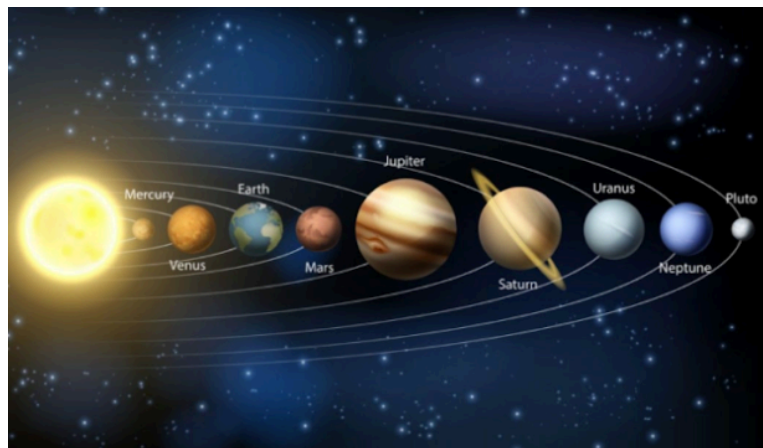
Google Trends

If we accept the Gartner Hype Cycle as a valid model for thinking about big data, and if we accept Google trends as a proxy for expectations, we can see that the Peak of Inflated Expectations was around 2014 and we haven't necessarily reached the Trough of Disillusionment yet.

Small data is also powerful

While a lot of the recent innovations in data science have centered on our ability to efficiently handle increasingly large volumes of data, it's important to recognize that a vast majority of the data analyzed in the real world does fit in the memory on a modern laptop. As a business leader, you should carefully consider the needs of your data organization before deciding which tools and architectures to adopt.

I want to take you back in time to Johannes Kepler, who discovered three laws of planetary motion, and Tycho Brahe, a Danish astronomer who collected the data that Kepler eventually analyzed to build his three laws of motion, which then informed Newton's theory of gravitation. We have a huge amount of scientific knowledge developed from the data that Brahe collected, which consisted of around just 2,000 data points. This is a tiny dataset compared to datasets we talk about today, which sometimes contain hundreds of millions of data points. But the data Brahe collected was of high quality. If you have good, properly collected data, strong analytical and principled theoretical models, and a way of doing statistical modeling, you can get a huge amount out of your data.



"We have a huge amount of scientific knowledge developed from the data that Brahe collected, which consisted of around just 2,000 data points. This is a tiny dataset compared to datasets we talk about today, which sometimes contain hundreds of millions of data points."

Let's look at another example explored in Andrew Gelman's article, [How can a poll of only 1,004 Americans represent 260 million people with only a 3 percent margin of error?](#) Gelman takes you through the math to show that when you increase the amount of data, you get seriously diminishing returns on the reduction of the margin of error. Disclaimer: This result is true only if you have some sort of representative sampling of your population, which definitely isn't the case in polls. But statisticians now have sophisticated correction methods for non-representative samples, which they can use to learn about the voting preferences of a larger population.

Again, we see a significant result from a small amount of data, and it tells us about the nature of human behavior and human preferences. The same principle is true of business—it's about understanding and being able to predict future human behavior, especially in terms of your stakeholders. So why do we talk about big data so much when it's not necessary? A primary reason is that it's readily accessible and so computable these days.

The bottom line is Harvard Business Review's point in [Sometimes "Small Data" Is Enough to Create Smart Products](#) that "it's not about mountains of data, it's about small, high precision data."

"We see a significant result from a small amount of data, and it tells us about the nature of human behavior and human preferences. The same principle is true of business—it's about understanding and being able to predict future human behavior, especially in terms of your stakeholders."

Don't underestimate the power of thick data

Now that we've discussed the power of small data, I want to move onto another type of data called thick data, or qualitative data. Thin data involves numbers and tables, whereas thick data, a term from sociology and anthropology, is more qualitative and descriptive. A consulting group called ReD Associates has done fantastic work using thick data to help people build analytic models and machine learning models.

One example that I want to mention is their work in [detecting credit card fraud](#). This is a huge challenge and machine learning has been used to detect credit card fraud in the past, using features such as the amount of transaction, the time of transaction, location, and so on. ReD Associates attempted to collect thick data to solve this problem taking a sociological approach.

To do so, ReD Associates sat down with credit card fraudsters to find out what they actually do and what their processes are like. They found a community of credit card fraudsters on the dark web and met them in real life to learn about their processes and habits.

They discovered that the point at which credit card fraudsters have the highest likelihood of getting caught is when they actually have to do something in the real world—like picking up deliveries. They're tech savvy enough to rarely be detectable online. And fraudsters are also careful in the physical realm—they typically don't send parcels to their own address, their work address, or their friends' addresses. Instead, they send deliveries to addresses of properties that are abandoned or on the real estate market.

"Thin data involves numbers and tables, whereas thick data, a term from sociology and anthropology, is more qualitative and descriptive."

Equipped with this knowledge, ReD Associates built a credit card fraud detection model using the location that the parcel was being sent to as a feature, and joined that with publicly available data around abandoned houses and houses on the market. They observed that this model based on qualitative data obtained a significant lift in accuracy when compared with more traditional models for fraudulent transactions. This is a wonderful example of the importance of thick data and how taking a sociological approach can provide increased value. I recommend these two articles to explore how good quality data is more important than big data in further detail: [The Power of 'Thick' Data](#) and [Big Data Is Only Half the Data Marketers Need](#).

"They observed that this model based on qualitative data obtained a massive increase in accuracy when compared with more traditional models for fraudulent transactions."

CALL TO ACTION

Choose a data source that's valuable to your business and think about how much of this data you really need to inform decision making. One way you can gauge this is by considering the added value of increasing the amount 2X, 5X, and 10X, particularly with respect to the investment of collecting, storing, and analyzing it. Then, answer this question: What thick data could you use to enhance the quality of this data source?

3 The Future of Data Work Includes Even More Point-and-Click

One of the main reasons behind the rapid innovation and advances in data science has been the proliferation of open-source code-centric tools. This has allowed data scientists and software engineers to build powerful abstractions that make it really easy to analyze data.

For example, R packages like `dplyr` and `tidymodels` allow users to manipulate data and build sophisticated machine learning models with a few lines of code. Similarly, Python packages like `tensorflow` and `pytorch` allow anyone to build a deep learning model with little effort.

While code-driven tools have served data work well, I believe that the future of data work will be much broader, and involve graphical user interfaces (GUIs) that will let a broader audience engage in data science using point-and-click tools. Note that this does not by any means imply the death of the coding data scientist. Similar to the industrial revolution, this move will allow coders to redirect their focus to solving more complex problems, and building interfaces to tackle problems, leaving the domain-specific work to the domain experts. This is already happening across different slices of data work: descriptive, predictive, and prescriptive analytics.

Descriptive analytics is accessible with business intelligence

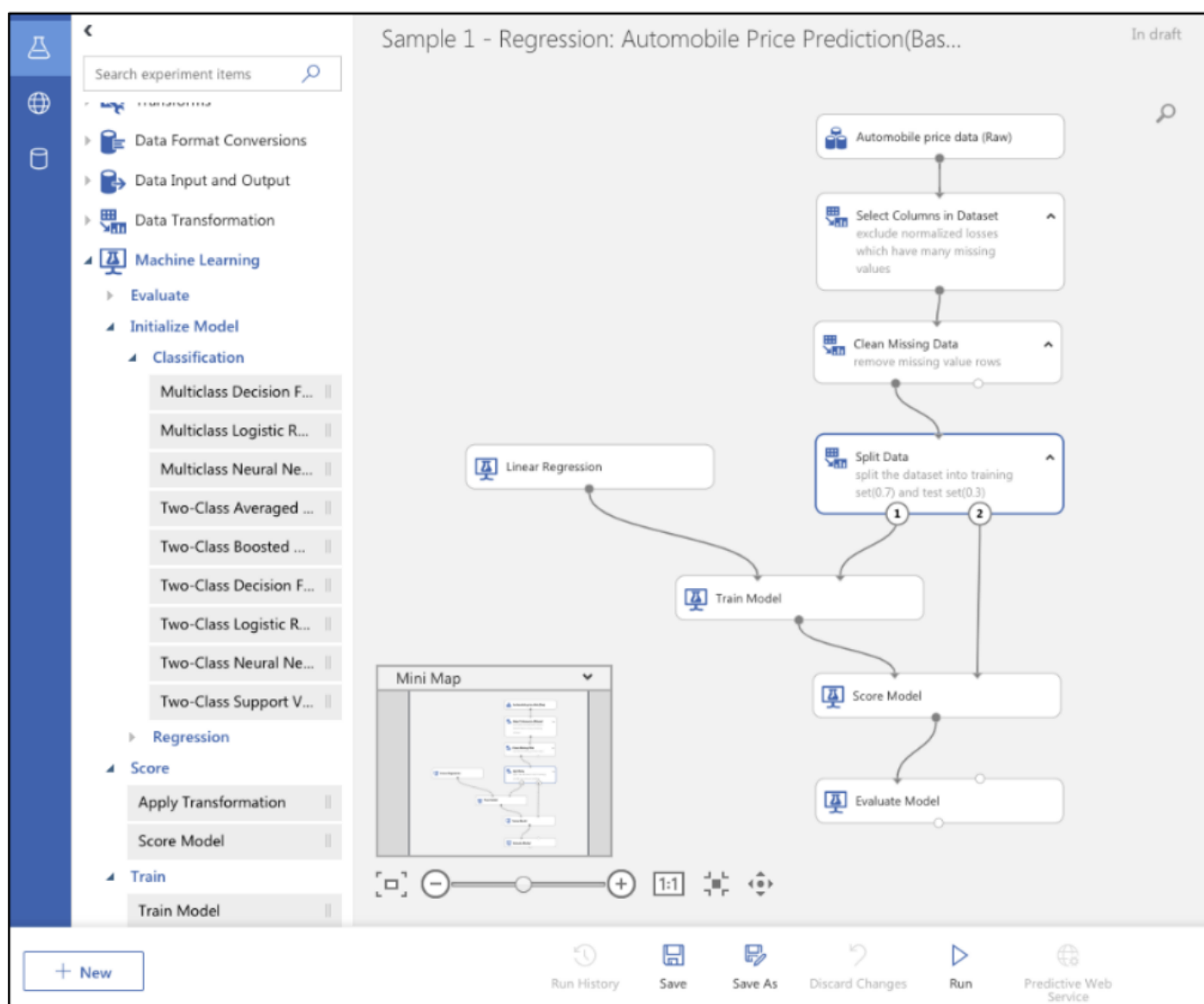
Descriptive analytics has been made accessible to a wider audience by the drag-and-drop business intelligence (BI) tools that are everywhere now. Just look at last year's \$2.6 billion acquisition of Looker, a very popular BI tool, by none other than Google. Also last year, Salesforce completed the acquisition of Tableau, another very popular BI tool. Furthermore, Microsoft has invested in its own tool, Power BI. (DataCamp offers courses on both [Tableau](#) and [Power BI](#).)

Business intelligence tools make data discovery accessible for all skill levels—not just advanced analytics professionals. They are one of the simplest ways to work with data, providing the ability to collect data in one place, gain insight into what will move the needle, forecast outcomes, and much more.

How drag-and-drop interfaces reduce the barrier to entry for predictive and prescriptive analytics

The democratization of data work is evolving to also include predictive and prescriptive analytics. For example, [Alteryx](#) is a \$10 billion dollar company that provides a drag-and-drop data science platform that allows business analysts to manipulate data and build predictive models using a drag-and-drop interface.

Google is investing a huge amount in their AutoML tool, which makes “AI as simple as drag and drop.” We see that [Microsoft is also entering the space](#):



Above is an example of a basic regression model with a data pipeline. You can see that the data flows from rectangle to rectangle. First data is ingested, then columns are selected, the data is cleaned, and the model splits and trains the data.

There are many potential use cases for drag-and-drop machine learning. This can be reframed as: what type of people in an organization would want to use ML tools that aren't coders? Perhaps your customer success team wants to model customer churn. They'll be able to do this using automated, drag-and-drop tools to forgo writing new models and new code every time a new customer comes in. Similarly, your Chief People Officer and HR team will be using more and more automated machine learning models in hiring flows, including screening resumes (we'll get to the potential perils of this in a minute). And your marketing team may be interested in using such tools for the marketing funnel, which changed with the advent of ML tools for programmatic ad buying. One other telling example, is supply chain optimization, which can greatly reduce costs within an organization and is relatively low-hanging fruit.

This trend can also be observed in complex spaces like deep learning. For example, [Lobe](#) is a startup that makes deep learning accessible to all, allowing users to build models using a GUI. Lobe was acquired by Microsoft even before they came out of a beta, which is testament to the importance of GUI tools in shaping the future.

Another example is an [image search engine](#) built by Google that lets medical doctors explore model predictions and fine-tune them further with human intelligence. Such tools can increase adoption with great effect by opening up the black box and supplementing ML models with human intelligence.

We're going to see more drag-and-drop interfaces that will reduce the barrier to entry. This will allow us all to think more about which types of tools we want to build in-house and which ones we want to source from vendors.

"There are many potential use cases for drag-and-drop machine learning. This can be reframed as: what type of people in an organization would want to use ML tools that aren't coders?"

Algorithmic marketplaces allow non-coders to purchase models

In chapter two, we mentioned that we'll see more and more marketplaces for these ML products, like Booz Allen's Modzy. Modzy is a point-and-click marketplace where you can buy models that others have created and incorporate them into your tech stack.

There are dangers to allowing non-coders and non-statisticians to use these tools. Namely, we must ensure that everyone knows of the inherent risks associated with these tools and how to mitigate them. I hope that increasingly, product development advancements in point-and-click tools will focus on helping people recognize when they're given biased results, and not to take them at face value.

Beware of the danger zone of drag-and-drop ML

I encourage you to check out an interesting educational game out of Mozilla Labs, [Survival of the Best Fit](#). It's a 10-minute educational game about hiring bias in AI where you get to play a CEO who attempts to scale the hiring process with an automated tool that causes hiring practices to go wrong.

To recap: we'll see more data work done in GUIs, via drag-and-drop and point-and-click interfaces. The space is getting more and more competitive with key investments by big players like Google and Microsoft, with much more to come. And even if a lot of your data science work is currently done in-house, more data work is increasingly shifting to vendors—which means there needs to be widespread education of potential risks and consequences to adopting AI.

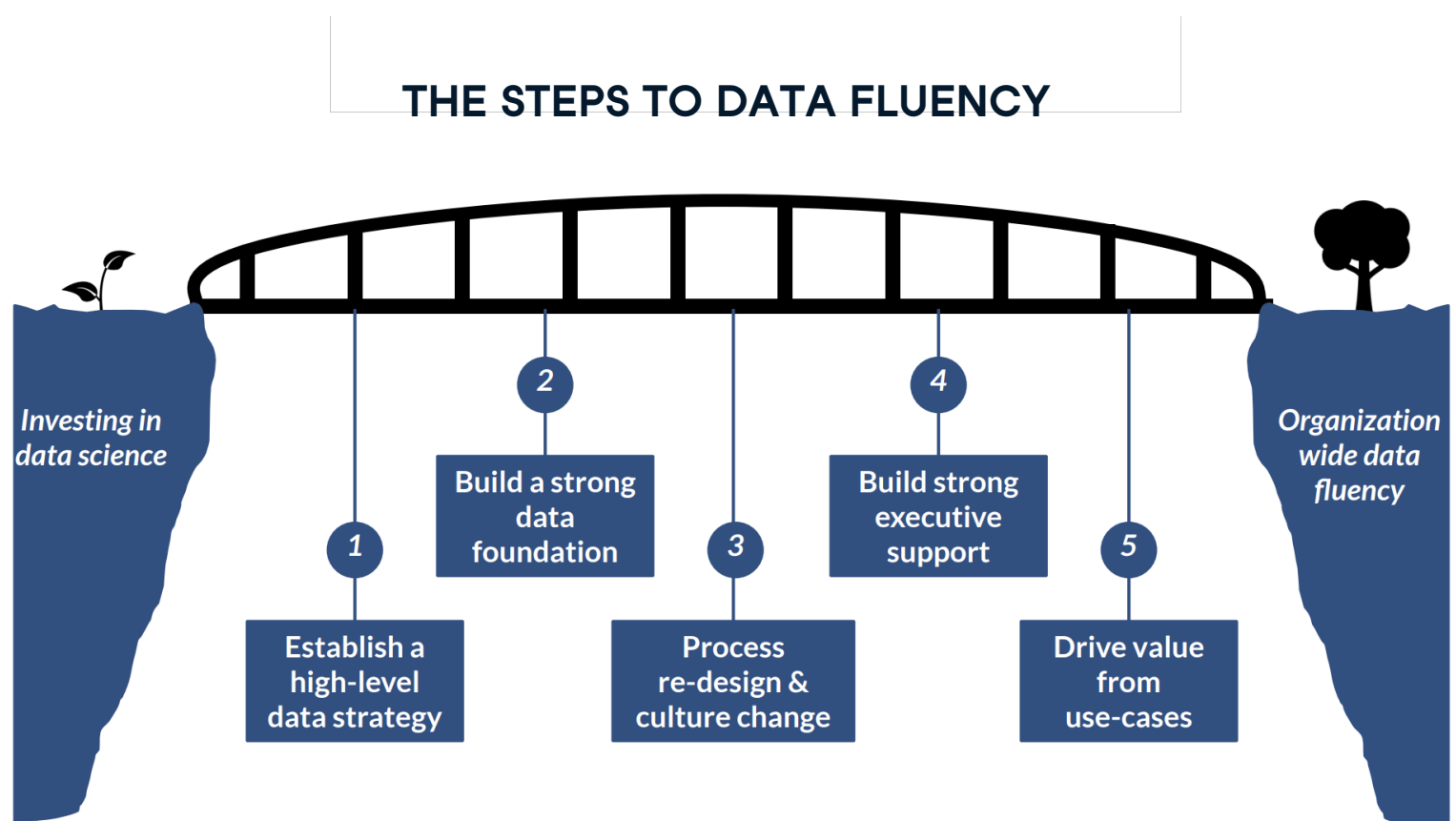
CALL TO ACTION

Write down how much of the day-to-day work in your organization currently happens in code. You could write this in human hours or how much it actually costs your organization. Then calculate how much is happening in GUIs, point-and-click, and drag-and-drop interfaces. How do you see this changing over the next 24 months?

Data Strategy Means Data Culture

Now it's time to dive into how data culture is a key component of data strategy. When I spoke with Taras Gorishnyy about [data science at McKinsey](#), I asked him to identify the key moving parts for data science, AI, and data strategy within an organization. The key moving parts Taras identified were:

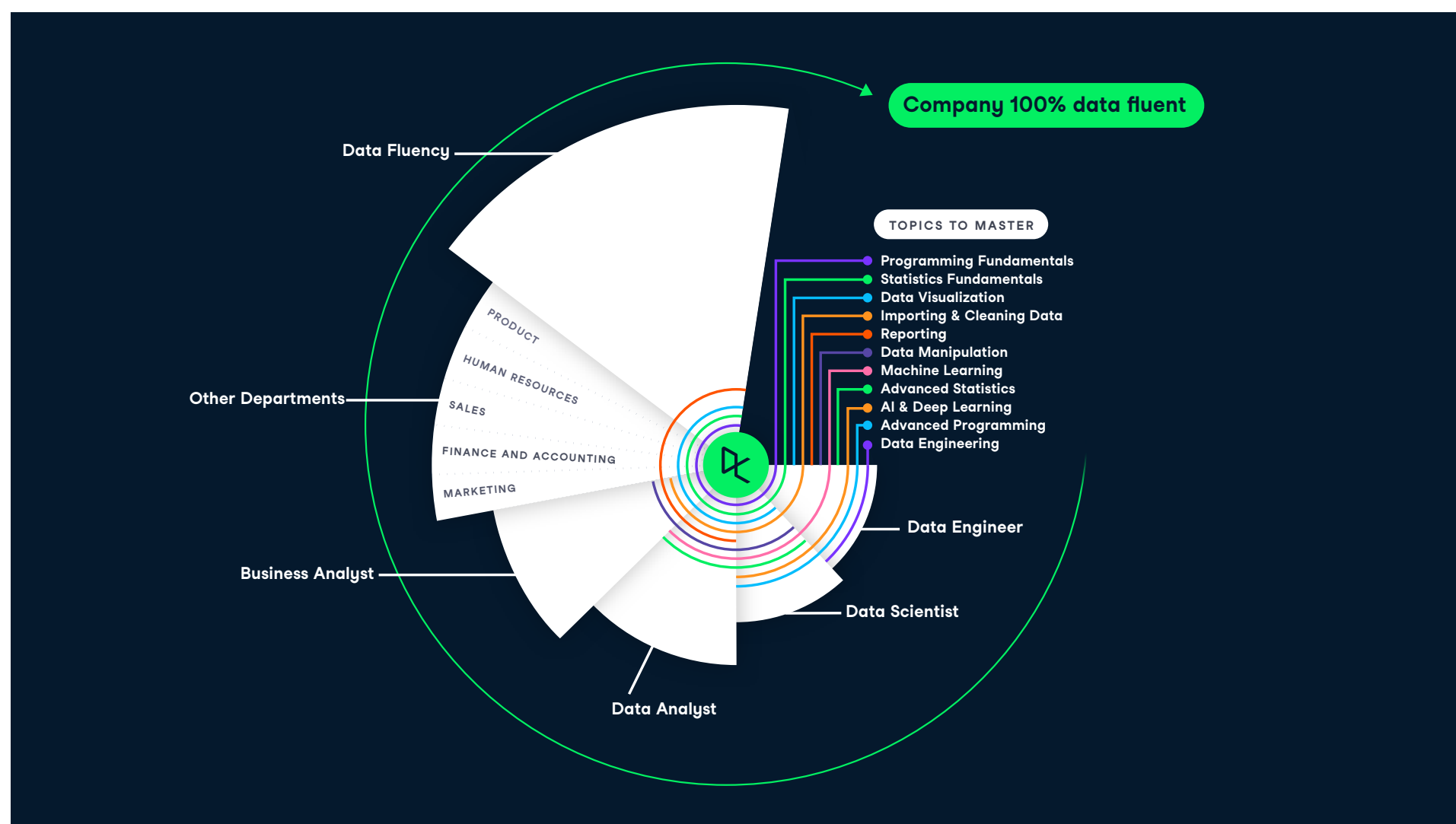
- Vision for analytics
- Robust data foundations
- A distribution of data skills and a data culture
- Executive support
- Establishing the impact of analytics early on



What do we mean when we talk about establishing a data culture? As [Tanya Cashorali](#) said to me on DataFramed, "Everyone at any level, whether it's C-level or entry level, should be looking and diving into data the same way you were expected to start using email 20 years ago."

Build data fluency across your organization

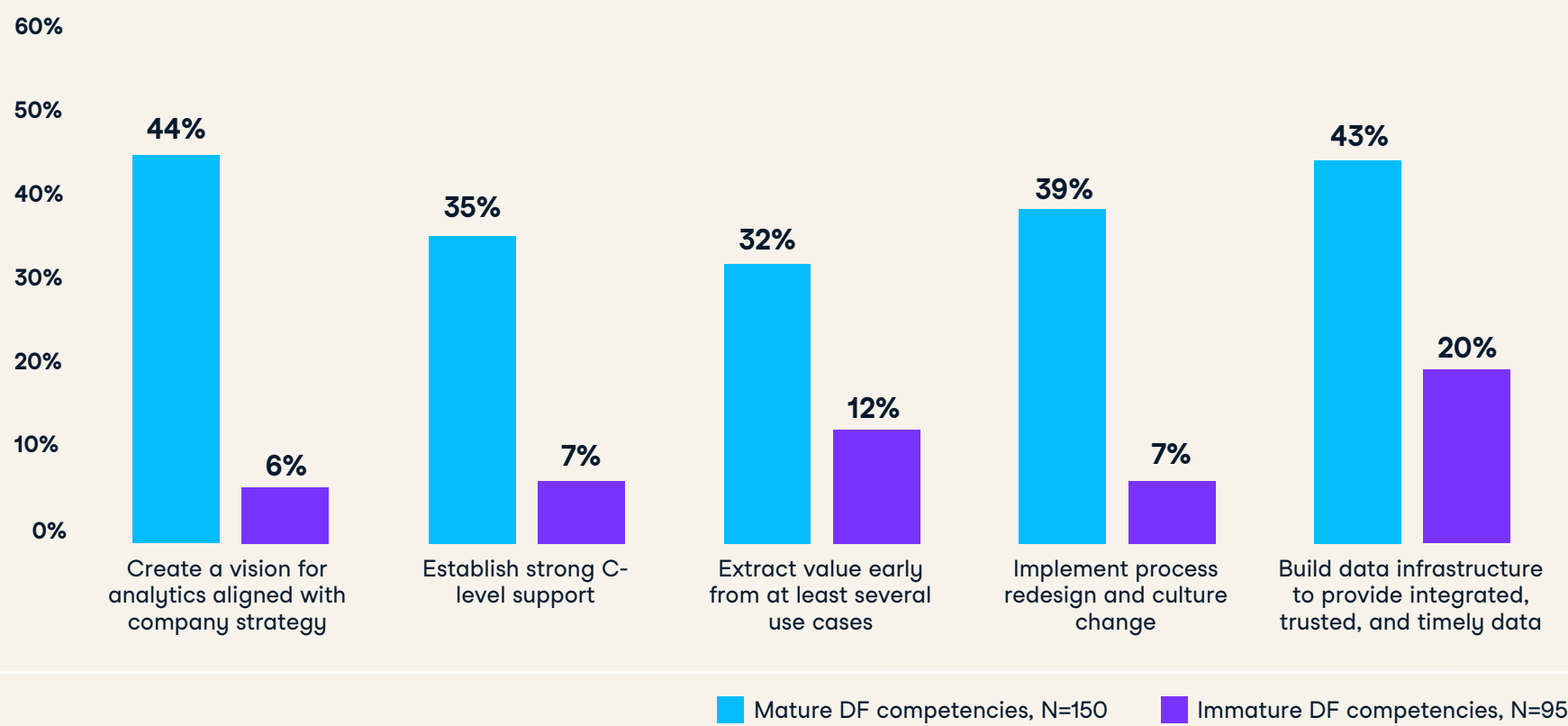
What implications does this have on data culture and data fluency? A data-fluent organization is one in which everybody knows how to dive into the data that they need to do their job. For example, our VP of Marketing might not need to write SQL code or any code at all, but to do her job well, she does need to know how to access the analytics dashboards and interact with them. She also needs to know how to ask the right questions of data scientists and how to use the results that they give her. The way we think about this at DataCamp is summarized in the figure below:



A strong data foundation requires starting with a data engineer, and then hiring data scientists and data analysts. Then, we can bring business analysts into the fold, along with other departments. This process helps everyone become data literate, leading to the company becoming 100% data fluent. Basic understanding of data tools and resources across a company greatly improves the quality of interaction among colleagues, allows teams to make better requests, and empowers everyone to make decisions autonomously.

In order to gauge the state of data fluency across industries, we conducted a survey in which we asked over 300 organizations to what extent they have taken actions to become data fluent, summarized in the figure below:

To what extent has your company taken (or plans to take) the following actions to become data fluent?



On the left, we have the mature companies with mature data fluency competencies, and on the right, immature competencies. What I want to highlight is that of the mature companies, only 39% had implemented process redesign and culture change with respect to data and only 7% of the immature companies had. There's a huge amount of improvement to be made here.

"Of the mature companies, only 39% had implemented process redesign and culture change with respect to data and only 7% of the immature companies had."

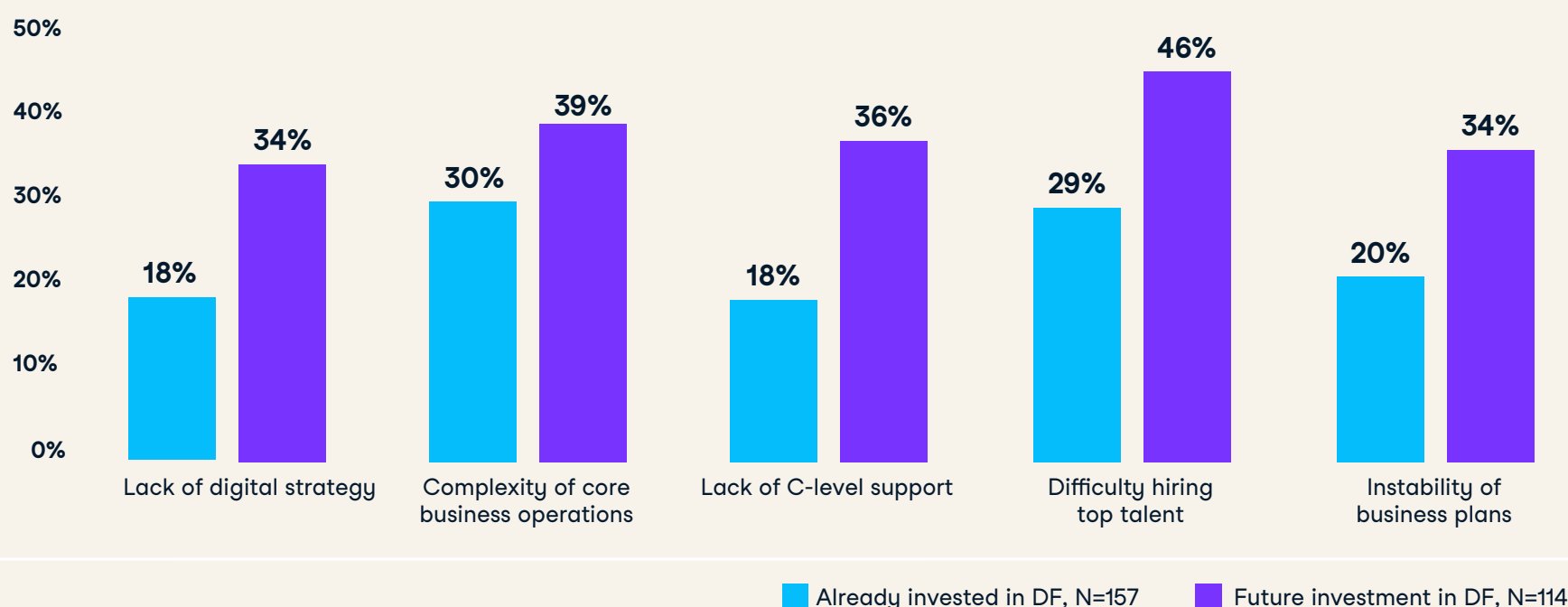
Involve your employees

It is important to prepare your employees for the future of work when crafting your data strategy. This requires ongoing conversations with all of your employees about what type of tasks you think will be automated and what won't, how your employee base can coexist with the tasks that are automated, and what that human-machine interface actually looks like. For some roles, we will see job automation, but the bigger challenge across the board will be task automation, since certain parts of someone's job may be automated away. I think it's far better to figure out how to upskill and re-skill these employees to transition them to do more high-level, meaningful work.

Prioritize upskilling employees

Another question we asked in our survey of 300 organizations was, “What type of business challenges prevent companies from building or improving data fluency?” The majority of organizations flagged their difficulty in hiring top talent, as so much data talent chooses to work in tech these days. This means you won't be able to hire as many unicorns as you'd like, so you should prioritize upskilling and reskilling everyone in your organization.

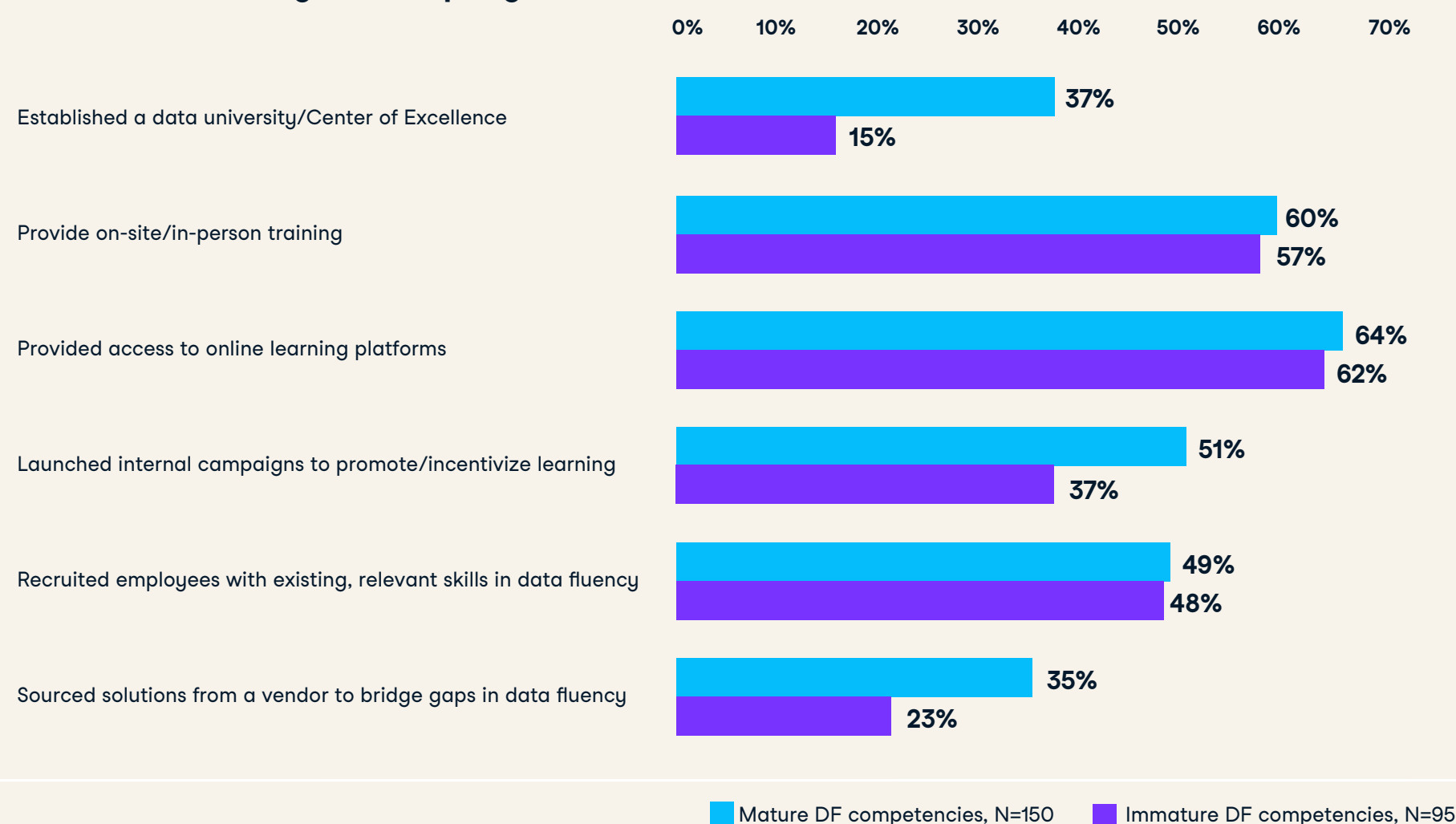
To what extent do the following business challenges prevent your company from building/improving data fluency?



To build data skills, many of the organizations we surveyed have considered the following options:

- Establishing data universities or center of excellence
- In-person training
- Online learning platforms
- Internal campaigns to promote and incentivize learning
- Recruiting
- Sourcing solutions from vendors

What actions has your company taken to build data skills?



In polls I've conducted on DataCamp webinars, over 50% of respondents have said that less than 50% of their data work was actually used to inform decision making. This leakage is due to culture!

CALL TO ACTION

List three to five outputs of data work in your organization and audit them with respect to how much your organization actually uses them to inform decision making. Too often, organizations hire a lot of great data analysts and data scientists to build dashboards and machine learning models, only to see them not be utilized in the intended way due to cultural challenges. All this good work can be wasted, resulting in a poor return on investment.

4 Data Strategy Means Considering Your Stakeholders

As with any other business strategy, data strategy requires you to consider who you're impacting and how you are impacting them. This is especially important with data strategy, since products are built and business decisions are implemented at a large scale and can impact many stakeholders. The three elements to carefully explore while evaluating your data strategy are bias, fairness, and privacy.

Explore vulnerability to bias and acknowledge your blind spots

It's tempting for businesses to use AI models to maximize profits, but they must consider whether their models are biased before deploying them. Responsible corporate citizens make efforts to mitigate any biases in their models. Tech companies are especially vulnerable to AI bias. They create a lot of value when building products and services at scale, but they can also create significant damage.

For example, last year the [New York Times](#) reported that the Apple Card was discriminating against women applying for credit, and there was little transparency on how these algorithms were built to decide which offers to extend to individual applicants. This led New York State regulators to investigate the algorithm that Apple used to determine creditworthiness. Whether intentional or not, discriminatory treatment of protected classes may violate laws and damage a company's brand, and such algorithms will increasingly become subject to regulation.

Thinking about how to implement different models in different contexts is incredibly important. Data is not one-size-fits-all—insights must consider who the stakeholders are and whether vulnerable groups may experience harmful consequences. Abeba Birhane's [The Algorithmic Colonization of Africa](#) uncovers the misappropriation of “mining” people for data, which harkens back to “the colonizer attitude that declares humans as raw material free for the taking.”

ProPublica’s [Machine Bias](#) is now a notorious example of a model that purportedly predicted recidivism rates for people on parole. These risk assessments were used in parole hearings and are becoming increasingly common in courtrooms across America. ProPublica demonstrated that the model was actually biased against Black people, assigning them disproportionately high risk scores, and that this was happening at scale.

As a result of trends like these, Cathy O’Neil, author of [Weapons of Math Destruction](#), said, “Data science doesn’t just predict the future, it causes the future.” This machine bias example illustrates that you have data science work as an input into the future of whether somebody gets parole or not. When we discussed this, O’Neil mentioned that she is working in the space of algorithmic audits and proposing a tool called the Ethical Matrix, in which the rows are the stakeholders and the columns are the concerns.

	Efficiency	Fairness	False +s’	False -s’	Transparency	Predictive Parity	Consistency	Data Quality
Court								
Black Defendants								
White Defendants								
Public								

Northpointe

Source: [KDnuggets](#)

Red denotes areas of big concern—for example, data quality seems to be a major concern across stakeholder groups, while predictive parity and fairness seem to be impacting Black defendants disproportionately. Understanding these metrics across groups of stakeholders is important to better empathize across groups and build unbiased models.

There will inevitably be groups that are impacted that we overlook. To protect against this, we must also consider who decides who the stakeholders are and what the concerns are. It’s a good idea to consult as many people as possible, especially for companies building products at scale and developing large-scale data strategy. Listing your stakeholders and concerns is the first step.

Measure fairness implications for your algorithms

Since much of machine learning is supervised learning, where previous data is used to train models, this can perpetuate systemic biases and stereotypes that hinder fairness. For example, consider a machine learning model that aids banks in their decision to provide credit or loans. Such models are typically built to maximize economic profits. However, profit maximization can result in a model that unfairly lowers the percentage of loans provided to marginalized groups. There are several examples from the popular press that highlight these issues, like Google AI's visual exploration of [Measuring Fairness](#).

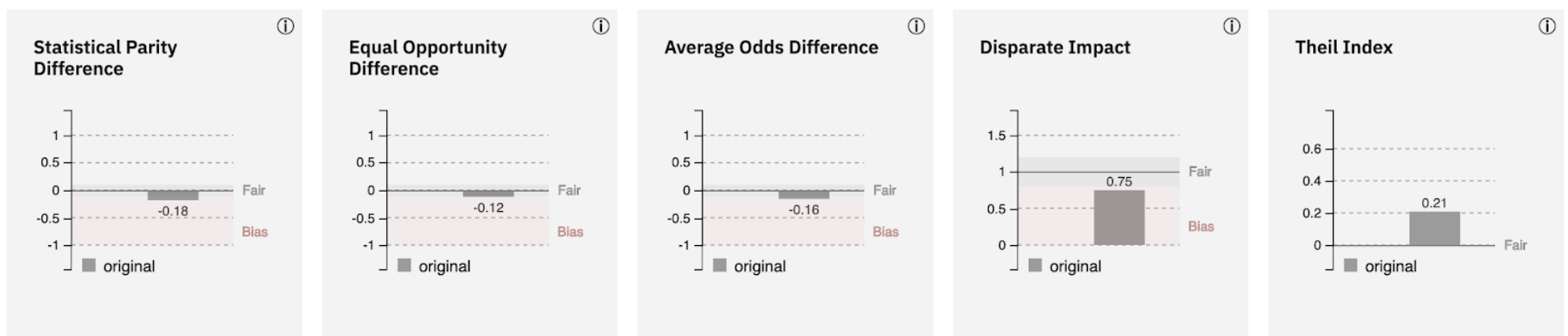
Another example is IBM's [AI Fairness 360](#) toolkit, which computes various bias metrics for a machine learning model and suggests modifications to the algorithm to mitigate bias.

Protected Attribute: Race

Privileged Group: **Caucasian**, Unprivileged Group: **Not Caucasian**

Accuracy with no mitigation applied is 66%

With default thresholds, bias against unprivileged group detected in 4 out of 5 metrics



Source: [AI Fairness 360](#)

"Profit maximization can result in a model that unfairly lowers the percentage of loans provided to marginalized groups."

Prioritize privacy to safeguard your stakeholders

One blind spot that's easy to overlook is the stakeholders behind data that you collect. This group requires protection from abuse in regard to privacy rights and consumer protection—which will increasingly take the form of regulation. In recent years, we've seen landmark data privacy regulations become implemented, like GDPR and CCPA. [GDPR](#) has substantially changed the game in Europe, and [CCPA](#) is a similar standard adopted in the state of California.

Businesses owe it to their stakeholders to adopt principled thinking around how to build their products and services and construct their data strategies.

CALL TO ACTION

Choose a current data initiative at your company and list all the stakeholders impacted by it. Compare your notes with a colleague to examine who you may have missed. Then list two ways that each stakeholder could be positively impacted by the data initiative and two ways they could be negatively impacted.