THE DATA SCIENCE MINDSET

SIX PRINCIPLES TO BUILD HEALTHY DATA-DRIVEN SKILLS & ORGANIZATIONS

Dr. BRIGHT

The

Data Science Mindset

6 Principles To Build Healthy Data-Driven

Skills & Organizations

Dr Bright

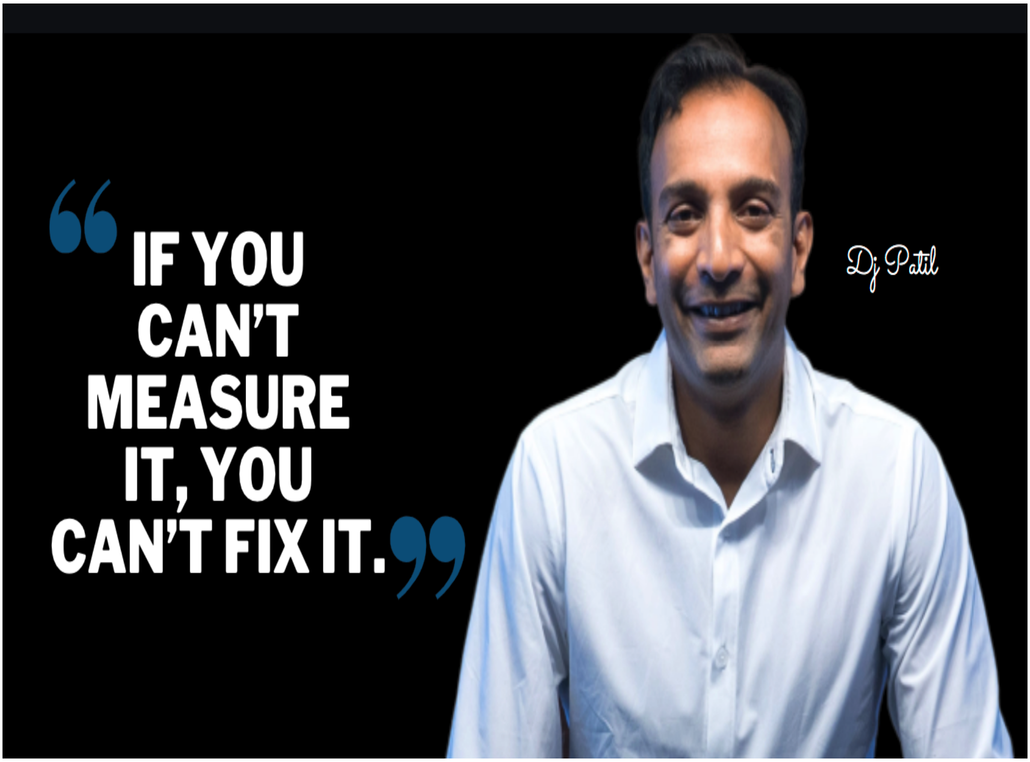
DEDICATION

I dedicate this book to anyone who truly aspires to make the field of Data Science and Artificial Intelligence better.

CONTENTS

ACKNOWLEDGMENT

Thanks to everyone on the Total Data Science team who helped in bringing forth this book. Special thanks to great minds that contributed to the success of this book. Special thanks to Francesca Lazzeri, PhD for her insights.



***DJ Patil** is co-coiner of the term 'Data Scientist'and co-author of the Harvard Business Review article: "Data Scientist: Sexiest Job of the 21st Century." He alsoserved as the Chief Data Scientist at the White House and as Vice President of Product at RelateIQ*

INTRODUCTION

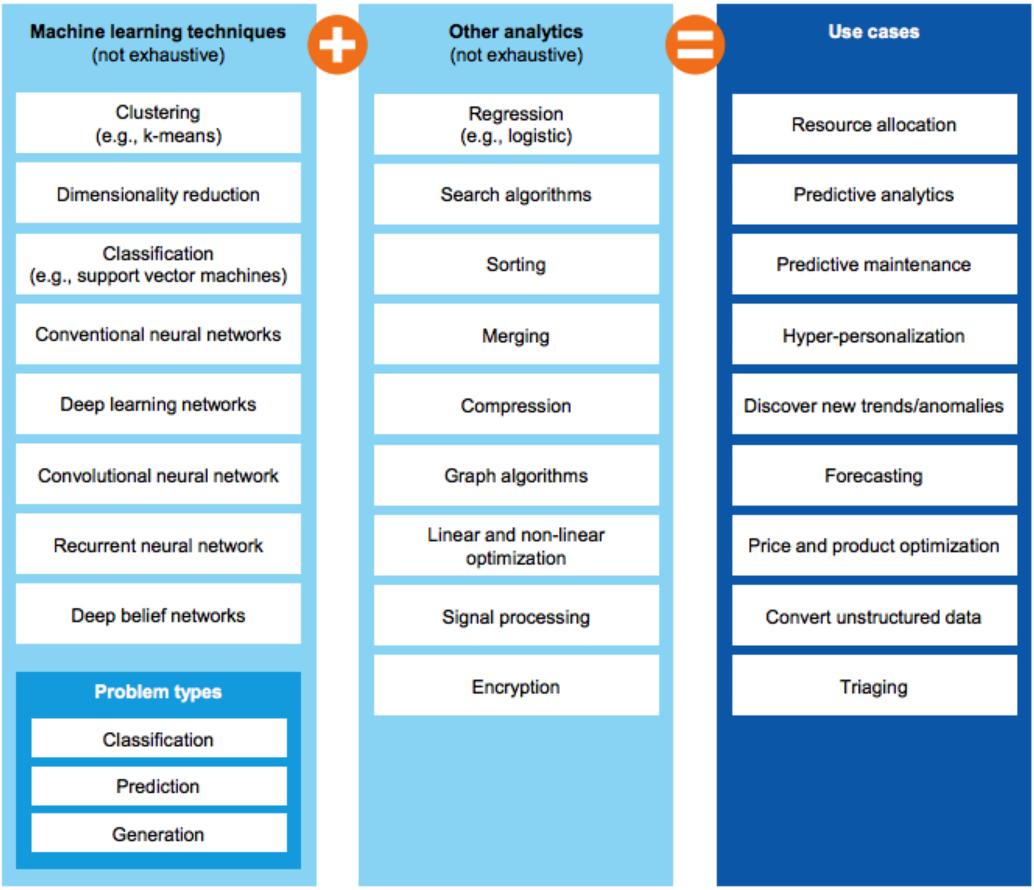
Five years ago, the McKinsey Global Institute (MGI) released *Big data: The next frontier for innovation, competition, and productivity*. In the years since, data science has continued to make rapid advances, particularly on the frontiers of machine learning and deep learning. Organizations now have troves of raw data combined with powerful and sophisticated analytics tools to gain insights that can improve operational performance and create new market opportunities. Most profoundly, their decisions no longer have to be made in the dark or based on gut instinct; they can be based on evidence, experiments, and more accurate forecasts. As we take stock of the progress that has been made over the past five years, we see that companies are placing big bets on data and analytics. But adapting to an era of more data-driven decision making has not always proven to be a simple proposition for people or organizations. Many are struggling to develop talent, business processes, and organizational muscle to capture real value from analytics. This is becoming a matter of urgency, since analytics prowess is increasingly the basis of industry competition, and the leaders are staking out large advantages. Meanwhile, the technology itself is taking major leaps forward—and the next generation of technologies promises to be even more disruptive. Machine learning and deep learning capabilities have an enormous variety of applications that stretch deep into sectors of the economy that have largely stayed on the sidelines thus far.

According to the Harvard Business Review, the biggest obstacles to creating data-based businesses aren't technical; they're cultural, **a kind of mindset** It is simple enough to describe how to inject data into a decision-making process. It is far harder to make this normal, even automatic, for employees — a shift in mindset that presents a daunting challenge.

THE FRONTIERS OF MACHINE LEARNING, INCLUDING DEEP LEARNING, HAVE RELEVANCE IN EVERY INDUSTRY AND WIDE-RANGING POTENTIAL TO SOLVE PROBLEMS

Conventional software programs are hard-coded by humans with specific instructions on the tasks they need to execute. By contrast, it is possible to create algorithms that "learn" from data without being explicitly programmed. The concept underpinning machine learning is to give the algorithm a massive number of "experiences" (training data) and a generalized strategy for learning, then let it identify patterns, associations, and insights from the data. In short, these systems are trained rather than programmed. Some machine learning techniques, such as regressions, support vector machines, and k-means clustering, have been in use for decades. Others, while developed previously, have become viable only now that vast quantities of data and unprecedented processing power are available. Deep learning, a frontier area of research within machine learning, uses neural networks with many layers (hence the label "deep") to push the boundaries of machine capabilities. Data scientists have recently made breakthroughs using deep learning to recognize objects and faces and to understand and generate language. Reinforcement learning is used to identify the best actions to take now in order to reach some future goal. These type of problems are common in games but can be useful for solving dynamic optimization and control theory problems—exactly the type of issues that come up in modeling complex systems in fields such as engineering and economics. Transfer learning focuses on storing knowledge gained while solving one problem and applying it to a different problem. Machine learning, combined with other techniques, could have an enormous range of uses.

Machine learning can be combined with other types of analytics to solve a large swath of business problems.



Machine learning techniques (not exhaustive)

- Clustering (e.g., k-means)
- Dimensionality reduction
- Classification (e.g., support vector machines)
- Conventional neural networks
- Deep learning networks
- Convolutional neural network
- Recurrent neural network
- Deep belief networks

Problem types

- Classification
- Prediction
- Generation

Other analytics (not exhaustive)

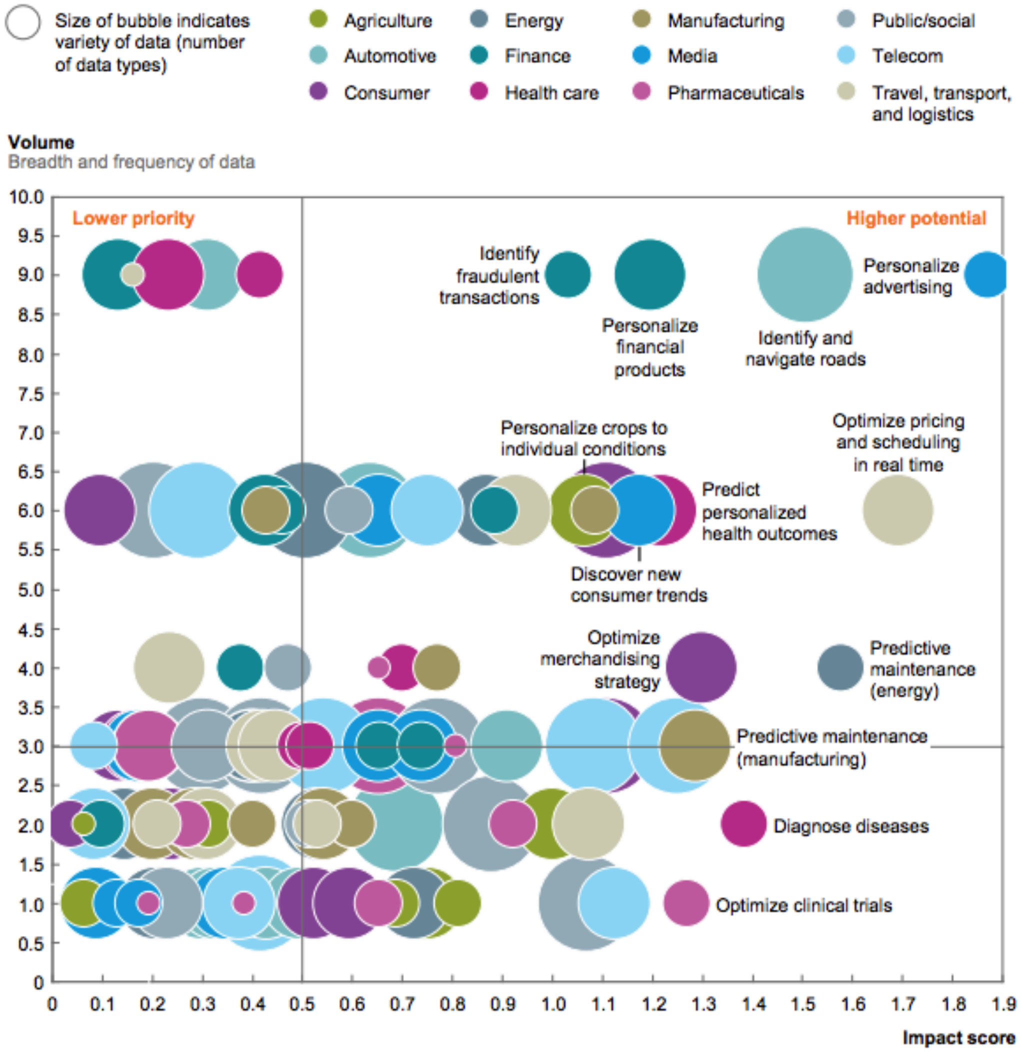
- Regression (e.g., logistic)
- Search algorithms
- Sorting
- Merging
- Compression
- Graph algorithms
- Linear and non-linear optimization
- Signal processing
- Encryption

Use cases

- Resource allocation
- Predictive analytics
- Predictive maintenance
- Hyper-personalization
- Discover new trends/anomalies
- Forecasting
- Price and product optimization
- Convert unstructured data
- Triaging

SOURCE: McKinsey Global Institute analysis

Machine learning has broad potential across industries and use cases.



SOURCE: McKinsey Global Institute analysis

The industry-specific uses that combine data richness with a larger opportunity are the largest bubbles in the top right quadrant of the chart. These represent areas where organizations should prioritize the use of machine learning and prepare for a transformation to take place. Some of the highest-opportunity use cases include personalized advertising; autonomous vehicles; optimizing pricing, routing, and scheduling based on real-time data in travel and logistics; predicting personalized health outcomes; and optimizing merchandising strategy in retail.

The use cases in the top right quadrant fall into four main categories. First is the radical personalization of products and services for customers in sectors such as consumer packaged goods, finance and insurance, health care, and media—an opportunity that most companies have yet to fully exploit. The second is predictive analytics. This includes examples such as triaging customer service calls; segmenting customers based on risk, churn, and purchasing patterns; identifying fraud and anomalies in banking and cybersecurity; and diagnosing diseases from scans, biopsies, and other data.

The third category is strategic optimization, which includes uses such as merchandising and shelf optimization in retail, scheduling and assigning frontline workers, and optimizing teams and other resources across geographies and accounts.

The fourth category is optimizing operations and logistics in real time, which includes automating plants and machinery to reduce errors and improve efficiency, and optimizing supply chain management.

OPPORTUNITIES STILL UNCAPTURED

Industrial companies are embracing artificial intelligence (AI) as part of the fourth digital revolution. AI leverages big data; it promises new insights that derive from applying machine learning to datasets with more variables, longer timescales, and higher granularity than ever. Using months or even years' worth of information, analytics models can tease out efficient operating regimes based on controllable variables, such as pump speed, or disturbance variables, such as weather. These insights can be embedded into existing control systems, bundled into a separate advisory tool, or used for performance management.

Many companies in heavy industry have spent years building and storing big data but have yet to unlock its full value. In fact, research shows that more than 75 percent have piloted some form of AI, yet less than 15 percent have realized meaningful, scalable impact. In these companies, analytics teams typically take an outside-in approach to AI and machine learning, including using various stochastic methods on top of process data that have been engineered with minimal operational insight. That approach can work, but it usually produces models that exhibit a high parameter dependence, require frequent retraining, have a high number of inputs, or give unphysical or unrealistic results. Consequently, these models rarely endure in production or achieve meaningful impact before operators and engineers lose confidence in them.

To succeed with AI, companies should have an automation environment with reliable historian data. Then, they will need to adapt their big data into a form that is amenable to AI, often with far fewer variables and with intelligent, first principles–based feature engineering. McKensey termed the latter format "smart data" to emphasize the focus on an expert-driven approach that improves predictive accuracy and aids in root-cause analysis.

It all boils down to the ***mindset***
of the data scientists and the
leaders in the organisation.

THERE IS A CONTINUING SHORTAGE OF ANALYTICS TALENT

Across the board, companies report that finding the right talent is the biggest hurdle they face in trying to integrate data and analytics into their existing operations. In a recent McKinsey & Company survey, approximately half of executives across geographies and industries reported greater difficulty recruiting analytical talent than filling any other kind of role. Forty percent(40%) say retention is also an issue. Data scientists, in particular, are in high demand. The McKinsey 2011 report hypothesized that demand for data scientists would outstrip supply. This is in fact what we see in the labor market today, despite the fact that universities are adding data and analytics programs and that other types of training programs are proliferating. Average wages for data scientists in the United States rose by approximately 16 percent a year from 2012 to 2014. This far outstrips the less than 2 percent increase in nominal average wages across all occupations in US Bureau of Labor Statistics data. The scarcity of elite data scientists has even been a factor in some acquisitions of cutting-edge artificial intelligence (AI) startups; deals can command around $5 million to $10 million per employee. This trend is likely to continue in the near term. While we estimate that the number of graduates from data science programs could increase by a robust 7 percent per year, our high-case scenario projects even greater (12 percent) annual growth in demand, which would lead to a shortfall of some 250,000 data scientists. But a countervailing force could ease this imbalance in the medium term: data preparation, which accounts for more than 50 percent of data science work, could be automated. Whether that dampens the demand for data scientists or simply enables data scientists to shift their work toward analysis and other activities remains to be seen. Many organizations focus on the need for data scientists, assuming their presence alone will enable an analytics transformation. But another equally vital role is that of the business translator who serves as the link between analytical talent and practical applications to business questions. In addition to being data savvy, business translators need to have deep organizational knowledge and industry or functional expertise. This enables them to ask the data science team the right questions and to derive the right insights from their findings. It may be possible to outsource analytics activities, but business translator roles require proprietary knowledge and should be more deeply embedded into the organization. Organizations need to build these capabilities from within.

Without a holistic organisational data-driven ***mindset****, everything is just a dream and remains a dream.*

ANALYTICS LEADERS ARE CHANGING THE NATURE OF COMPETITION AND CONSOLIDATING BIG ADVANTAGES

There are now major disparities in performance between a small group of technology leaders and the average company—in some cases creating winner-take-most dynamics. Leaders such as Apple, Alphabet/Google, Amazon, Facebook, Microsoft, GE, and Alibaba Group have established themselves as some of the most valuable companies in the world.8 The same trend can be seen among privately held companies. The leading global "unicorns" tend to be companies with business models predicated on data and analytics, such as Uber, Lyft, Didi Chuxing, Palantir, Flipkart, Airbnb, DJI, Snapchat, Pinterest, BlaBlaCar, and Spotify. These companies differentiate themselves through their data and analytics assets, processes, and strategies.

The relative value of various assets has shifted. Where previous titans of industry poured billions into factories and equipment, the new leaders with data science mindset invest heavily in digital platforms, data, and analytical talent.

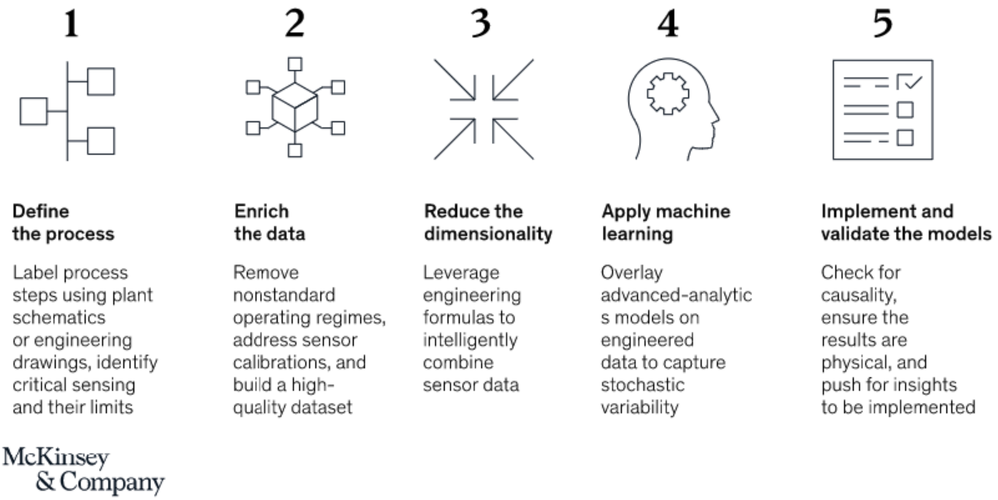
New digital native players can circumvent traditional barriers to entry, such as the need to build traditional fixed assets, which enables them to enter markets with surprising speed. Amazon challenged the rest of the retail sector without building stores (though it does have a highly digitized physical distribution network), "fintechs" are providing financial services without physical bank branches, Netflix is changing the media landscape without connecting cables to customers' homes, and Airbnb has introduced a radical new model in the hospitality sector without building hotels.

But some digital natives are now erecting new barriers to entry themselves; platforms may have such strong network effects that they give operators a formidable advantage within a given market. The leading firms have a remarkable depth of analytical talent deployed on a variety of problems—and they are actively looking for ways to enter other industries. These companies can take advantage of their scale and data insights to add new business lines, and those expansions are increasingly blurring traditional sector boundaries. Apple and Alibaba, for instance, have introduced financial products and services, while Google is developing autonomous cars. The importance of data has also upended the traditional relationship between organizations and their customers since every interaction generates information. Sometimes the data itself is so prized that companies offer free services in order to obtain it; this is the case with Facebook, LinkedIn, Pinterest, Twitter, Tencent, and many others. An underlying barter system is at work, particularly in the consumer space, as individuals gain access to digital services in return for data about their behaviors and transactions.

CREATING SMART DATA WITH A DATA-DRIVEN MINDSET

A common failure mode for companies looking to leverage AI is poor integration of operational expertise into the data-science process. It is advocated to apply machine learning only after process data have been analyzed, enriched, and transformed with expert-driven data engineering. Effective Data Scientists with smart data-driven mindset should be equipped with the following five processes advocated by McKensey:

Five steps can turn process-data into smart-data:



1 Define the process

Label process steps using plant schematics or engineering drawings, identify critical sensing and their limits

2 Enrich the data

Remove nonstandard operating regimes, address sensor calibrations, and build a high-quality dataset

3 Reduce the dimensionality

Leverage engineering formulas to intelligently combine sensor data

4 Apply machine learning

Overlay advanced-analytics models on engineered data to capture stochastic variability

5 Implement and validate the models

Check for causality, ensure the results are physical, and push for insights to be implemented

McKinsey & Company

1. Define the process

Outlining the steps of the process in a data science project with experts and business stakeholders is an important step to the success of the projects at hand.

As an example, a North American mining company endeavored to improve the throughput of its grinding operations, which included seven grinding mills and three cyclone "classifiers," which separate particles based on size. Experts and engineers sat with the data-science team to illustrate the process flow, which was divided into three stages of grinding and separation, each of which was monitored by approximately a dozen sensors. Data tags (or labels) were noted along with sensor redundancies and instrument accuracies. Metallurgical staff provided derivations of Plitt's equation for particle separation and the Bond equation for grinding energy, among others. The result was a unified team with plant experts who understood what to look for in the field that may affect the resulting models and data scientists who recognized operational pitfalls and where to improve data quality.

2. Enrich the data

Raw process data nearly always contain deficiencies. Thus, creating a high-quality dataset should be the focus, rather than striving for the maximum number of observables for training. Teams should be aggressive in removing nonsteady-state information, such as the ramping up and down of equipment, along with data from unrelated plant configurations or operating regimes. Generic methods to treat missing or anomalous data should be avoided, such as imputing using averages, "clipping" to a maximum, or fitting to an assumed normal distribution. Instead, teams should start with the critical sensors identified by process experts and carefully address data gaps using virtual sensors and physically correct imputations.

For example, a European chemical company aimed to apply machine learning to its cracking furnace. Experts indicated that a flow meter was critical to the process, but the data-science team determined it was faulty, and the values were occasionally erroneous because of miscalibration. The operations team proposed pausing the project until a new flow meter was installed. Instead, the existing values were enriched by creating a virtual flow sensor using mass-balance formulas and upstream sensor data for temperature and energy use. With the virtual sensor engineered, the analytics team was able to triangulate and correct the flow values. In total, the project delivered a 20 percent increase in processing throughput.

3. Reduce the dimensionality

AI algorithms build a model by matching outputs, known as observables, to a set of inputs, known as features, which consist of raw sensor data or derivations thereof. Data Scientists with the right data-driven mindset should have this in mind from the start. Generally, the number of observables must greatly exceed the number of features to yield a generalized model. A common data-science approach is to engineer input combinations to produce new features. When this is combined with the sheer number of other observables such as sensors available in modern plants, this necessitates a massive number of observations. Instead, teams should pare the features list to include only those inputs that describe the physical process, then apply deterministic equations to create features that intelligently combine sensor information (such as combining mass and flow to yield density). Often, this is an excellent way to reduce the dimensionality of and introduce relationships in the data, which minimize the number of observables required to adequately train a model.

As an example, a European chemical company observed occasional pressure increases in the feed line to a spray dryer, which necessitated stops or slowdowns in its continuous process. A model was built to predict pressure buildup. Even when all the relevant sensor data were included, the results were unsatisfactory. In response, the team combined details of the pipe geometry with some of the sensor information into the Darcy–Weisbach equation. The result was a reduced number of model inputs and enhanced data quality, which subsequently increased the model performance. Operators were then able to leverage the model to nearly eliminate slowdowns, yielding an 8 percent throughput increase.

4. Apply machine learning

Data Science teams should evaluate features by inspecting their importance and therefore their explanatory power. Ideally, expert-engineered features that capture, for example, the physics of the process should rank among the most important. Overall, the focus should be on creating models that drive improvement, as opposed to tuning a model to achieve the highest predictive accuracy. Teams should bear in mind that some data naturally exhibit high correlations. In some cases, model performance can appear excellent, but it is more important to isolate the causal components and controllable variables than to solely rely on correlations. Finally, errors in the underlying data should be evaluated with respect to the objective function. It is not uncommon for data scientists to strive for higher model accuracy only to find that it is limited by sensor accuracy.

For example, a North American metal producer wanted to create a model to predict the heat needed to melt a batch of recycled material. The team first created one deterministic feature for “required heat” based on specific-heat equations that utilize the mass, heat capacity, and melting point of each alloy. Subsequently, data from 19 sensors were added as features to capture stochastic behavior, such as loss of heat through the flue or changes in the atmospheric temperature. The resulting model showed excellent performance, with the deterministic feature exhibiting an importance of more than 80 percent. The model output was sent directly to a human-machine interface (HMI) where operators could utilize the predictions to sequence melting. In total, the model has been running every minute for nearly two years, yielding a 10 percent reduction in melt time and a more consistent batch temperature.

5. Implement and validate the models

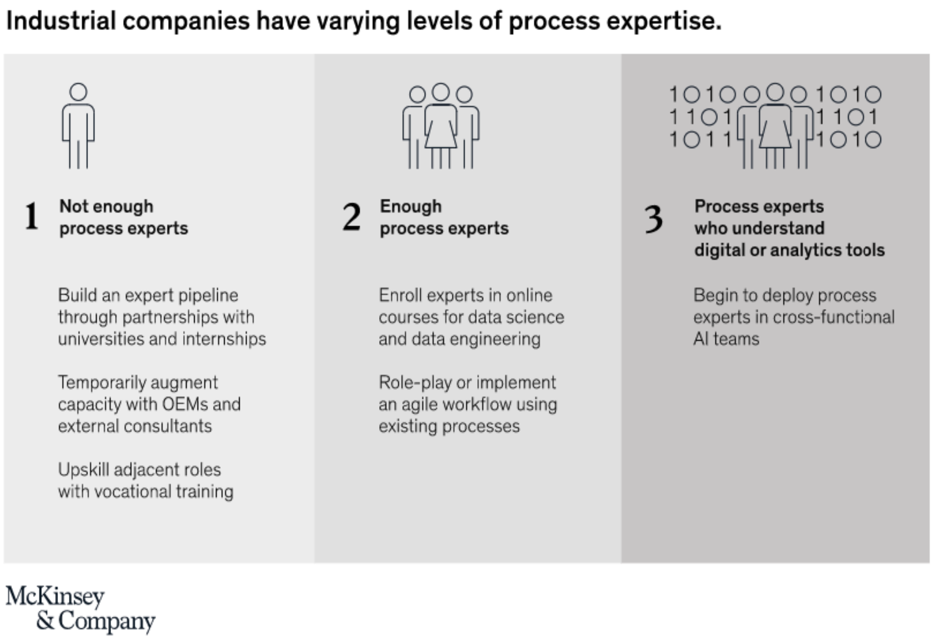
Impact can be achieved only if models (or their findings) are implemented. Taking action is critical. Data Science teams should continuously review model results with experts by examining important features to ensure they match the physical process, reviewing partial dependence plots (PDPs) to understand causality, and confirming what can actually be controlled. Additional meetings should be set up with operations colleagues to gauge what can be implemented and to agree on baseline performance: setting up cross-functional mindset. It is not uncommon for teams to convey model results in real time to operators in a control room or to engage in on-off testing before investing in production-grade, automated solutions.

As an example, a European bioscience player tried to optimize the yield of its fermentation process where data were scarce. After initial modeling efforts, only 40 percent of the variability in throughput could be explained with sensor data and engineered features. The team used insights from the parameter relations in the model to design an experiment in the plant, and these results were used to improve the model and inform operations as to where to place new sensors. The result was consensus between data-science and operations colleagues and a production increase of more than 20 percent.

Impact can be achieved only if models (or their findings) are implemented. Taking action is critical.

BUILDING THE TEAM

Deploying AI in heavy industry requires cross-functional teams made up of operators, data scientists, automation engineers, and process experts. A unified mindset. We often find that companies have (or are hiring to fill) roles for data science, but they face three main challenges regarding experts: there is a dearth or lack of expertise either at a specific facility or across the company; there are sufficient experts, but they are not comfortable with modern digital or analytical tools; or stakeholders don't know how to work effectively on digital teams.



Industrial companies have varying levels of process expertise.

1 Not enough process experts

Build an expert pipeline through partnerships with universities and internships

Temporarily augment capacity with OEMs and external consultants

Upskill adjacent roles with vocational training

2 Enough process experts

Enroll experts in online courses for data science and data engineering

Role-play or implement an agile workflow using existing processes

3 Process experts who understand digital or analytics tools

Begin to deploy process experts in cross-functional AI teams

McKinsey & Company

Process experts

Industrial companies are increasingly facing a shortage of experts due, in part, to the retirement of tenured employees and the lack of younger job candidates with analytical skills. As a result, companies looking to implement AI often need to first rebuild their expert pipeline, typically through partnerships with universities and internship programs. While the pipeline is being reestablished, OEMs and external consultants can be used to augment teams, but “owning” the skills is important in the long term because it is a source of differentiated value.

Concurrently, companies should upskill their existing experts in analytics tools and agile ways of working. Experts typically have engineering or other similar backgrounds; they are accustomed to leveraging formulas to describe physical processes. That type of thinking can be beneficial in creating smart data, but it can also engender distrust for AI-based approaches. Upskilling experts with a combination of classroom training and in-field apprenticeship on cross-functional AI teams can build comfort with the approach and results. With these skills, experts can better support digital teams, including partnering with data scientists to help them understand the problem, create smart data, and pressure-test models to ensure that the models have learned the correct first principle–based behavior. Moreover, upskilling has the added benefit of increasing job satisfaction and retention.

Ways of working

It can be challenging to create high-performing teams using cross-functional roles because of differences in approach. For example, it is common for operations employees to follow unidirectional stage-gated processes—often for safety reasons—whereas data-science colleagues are usually familiar with iterative workflows, such as agile. When deploying AI, our experience shows that iterative, inclusive, and colocated agile teams tend to realize the most impact. As a result, coaching is needed for colleagues unfamiliar with this approach to develop the right data-driven mindset.

Planning out the model development can be a good exercise to solidify a way of working and avoid linear approaches that include exhaustively completing one stage (such as data extraction) before proceeding to the next. Instead, pieces of each stage should be completed concurrently to quickly develop a fully working model with the intention of maturing individual components in future iterations. In practice, this usually means starting with a subset of data, creating a limited list of features, and working with simpler algorithms. Then, the team can decide what to invest in for the next stage. As part of each iteration, there should be a discussion of what the definition of "done" is to align on the outcome and avoid scope creep.

An important part that needs to be reemphasised is to form cross-functional data-science teams that include employees who are capable of bridging the gap between machine-learning approaches and business knowledge. Once these elements are combined with an agile way of working that advocates iterative improvement and a bias to implement findings, a true transformation can be achieved.

UNDERSTANDING THE HEALTHY DATA SCIENCE ORGANISATION FRAMEWORK

Being a data-driven organization implies embedding data science teams to fully engage with the business and adapting the operational backbone of the company (techniques, processes, infrastructures, and culture). The Healthy Data Science Organization Framework is a portfolio of methodologies, technologies, resources that, if correctly used, will assist your organization (from business understanding, data generation and acquisition, modeling, to model deployment and management) to become more data-driven. This framework, as shown below in Figure 1, includes six key principles:

1. Understand the Business and Decision-Making Process
2. Establish Performance Metrics
3. Architect the End-to-End Solution
4. Build Your Toolbox of Data Science Tricks
5. Unify Your Organization's Data Science Vision
6. Keep Humans in the Loop

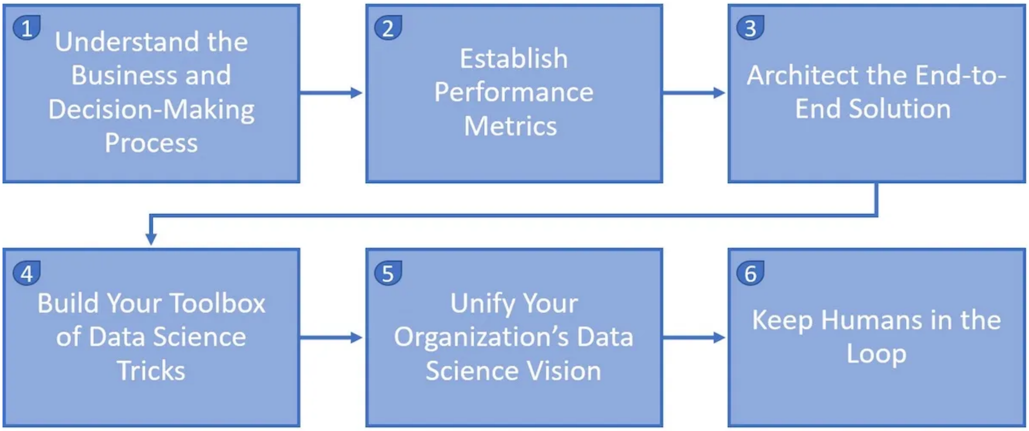

1
Understand the Business and Decision-Making Process
2
Establish Performance Metrics
3
Architect the End-to-End Solution
4
Build Your Toolbox of Data Science Tricks
5
Unify Your Organization's Data Science Vision
6
Keep Humans in the Loop

Figure 1. Healthy Data Science Organization Framework

Given the rapid evolution of this field, data scientists and organizations typically need guidance on how to apply the latest data science techniques to address their business needs or to pursue new opportunities.

In the up-coming pages, we will explore the ***six(6) principles*** that data scientists as well as organisations need to build healthy data-driven skills and organisations.

PRINCIPLE 1:

UNDERSTAND THE BUSINESS AND DECISION-MAKING PROCESS

For most organizations, lack of data is not a problem. In fact, it's the opposite: there is often too much information available to make a clear decision. With so much data to sort through, organizations need a well-defined strategy to clarify the following business aspects:

- How can data science help organizations transform business, better manage costs, and drive greater operational excellence?
- Do organizations have a well-defined and clearly articulated purpose and vision for what they are looking to accomplish?
- How can organizations get support of C-level executives and stakeholders to take that data-driven vision and drive it through the

different parts of a business?

In short, companies need to have a clear understanding of their business decision-making process and a better data science strategy to support that process.

> *With the right data science* ***mindset****, what was once an overwhelming volume of disparate information becomes a simple and clear decision point.*

Driving transformation requires that companies have a well-defined and clearly articulated purpose and vision for what they are looking to accomplish. It often requires the support of a C-level executive to take that vision and drive it through the different parts of a business.

Organizations must begin with the right questions. Questions should be measurable, clear and concise and directly correlated to their core business. In this stage, it is important to design questions to either qualify or disqualify potential solutions to a specific business problem or opportunity. For example, start with a clearly defined problem: a retail company is experiencing rising costs and is no longer able to offer competitive prices to its customers. One of many questions to solve this business problem might include: can the company reduce its operations without compromising quality?

There are two main tasks that organizations need to address to answer those type of questions:

- Define business goals: the Data Science team needs to work with business experts and other stakeholders to understand and identify the business problems.
- Formulate right questions: companies need to formulate tangible questions that define the business goals that the data science teams can target.

USE CASE:

Francesca Lazzeri, PhD (Twitter: @frlazzeri) is Senior Machine Learning Scientist at Microsoft on the Cloud Advocacy team and an expert in big data technology innovations and the applications of machine learning-based solutions to real-world problems.

Last year, the Azure Machine Learning team developed a recommendation-based staff allocation solution for a professional services company. By making use of Azure Machine Learning service, we built and deployed a workforce placement recommendation solution that recommends optimal staff composition and individual staff with the right experience and expertise for new projects. The final business goal of our solution was to improve our customer's profit.

Project staffing is done manually by project managers and is based on staff availability and prior knowledge of an individual's past performance. This process is time-consuming, and the results are often suboptimal. This process can be done much more effectively by taking advantage of historical data and advanced machine learning techniques.

In order to translate this business problem into tangible solutions and results, we helped the customer to formulate the right questions, such as:

1. How can we predict staff composition for a new project? For example, one senior program manager, one principal data scientist and two accounting assistants.
2. How can we compute the Staff Fitness Score for a new project? We defined our Staff Fitness Score as a metric to measure the fitness of staff with a project.

The goal of our machine learning solution was to suggest the most appropriate employee to a new project, based on employee’s availability, geography, project type experience, industry experience and hourly contribution margin generated for previous projects.

These solutions can address gaps or inefficiencies in an organization staff allocation that need to be overcome to drive better business outcomes. Organizations can gain a competitive edge by using workforce analytics to focus on optimizing the use of their human capital.

In the next few paragraphs, we will see together how Francesca and her team built this solution for their customer through a data science mindset.

PRINCIPLE 2:

ESTABLISH PERFORMANCE METRICS

In order to successfully translate this vision and business goals into actionable results, the next step is to establish clear performance metrics. In this second step, organizations need to focus on these two analytical aspects that are crucial to define the data solution pipeline (Figure 2) as well:

- What is the best analytical approach to tackle that business problem and draw accurate conclusions?
- How can that vision be translated into actionable results able to improve a business?

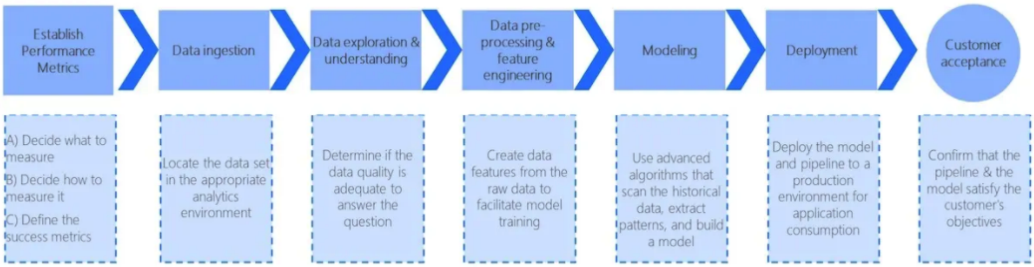

Establish Performance Metrics
Data ingestion
Data exploration & understanding
Data pre-processing & feature engineering
Modeling
Deployment
Customer acceptance
A) Decide what to measure
B) Decide how to measure it
C) Define the success metrics
Locate the data set in the appropriate analytics environment
Determine if the data quality is adequate to answer the question
Create data features from the raw data to facilitate model training
Use advanced algorithms that scan the historical data, extract patterns, and build a model
Deploy the model and pipeline to a production environment for application consumption
Confirm that the pipeline & the model satisfy the customer's objectives

Figure 2. Data solution pipeline

This step breaks down into three sub-steps:

1. Decide what to measure

Let's take Predictive Maintenance, a technique used to predict when an in-service machine will fail, allowing for its maintenance to be planned well in advance. As it turns out, this is a very broad area with a variety of end goals, such as predicting root causes of failure, which parts will need replacement and when providing maintenance recommendations after the failure happens, etc.

Many companies are attempting predictive maintenance and have piles of data available from all sorts of sensors and systems. But, too often, customers do not have enough data about their failure history and that makes it is very difficult to do predictive maintenance – after all, models need to be trained on such failure history data in order to predict future failure incidents. So, while it's important to lay out the vision, purpose and scope of any analytics projects, it is critical that you start off by gathering the right data. If the problem is to predict the failure of the traction system, the training data has to encompass all the different components for the traction system. The first case targets a specific component whereas the second case targets the failure of a larger subsystem. The general recommendation is to design prediction systems about specific components rather than larger subsystems.

Given the above data sources, the two main data types observed in the predictive maintenance domain are: 1) temporal data (such as operational telemetry, machine conditions, work order types, priority codes that will have timestamps at the time of recording. Failure, maintenance/repair, and usage history will also have timestamps associated with each event); and 2) static data (machine features and operator features, in general, are static since they describe the technical specifications of machines or operator attributes. If these features could change over time, they should also have timestamps associated with them). Predictor and target variables should be preprocessed/transformed into numerical, categorical, and other data types depending on the algorithm being used.

2. Decide how to measure it

Thinking about how organizations measure their data is just as important, especially before the data collection and ingestion phase. Key questions to ask for this sub-step include:

- What is the time frame?

- What is the unit of measure?
- What factors should be included?

A central objective of this step is to identify the key business variables that the analysis needs to predict. We refer to these variables as the model targets, and we use the metrics associated with them to determine the success of the project. Two examples of such targets are sales forecasts or the probability of an order being fraudulent.

3. Define the success metrics

After the key business variables identification, it is important to translate your business problem into a data science question and define the metrics that will define your project success. Organizations typically use data science or machine learning to answer five types of questions:

- How much or how many? (*regression*)
- Which category? (*classification*)
- Which group? (*clustering*)
- Is this weird? (*anomaly detection*)
- Which option should be taken? (*recommendation*)

Determine which of these questions companies are asking and how answering it achieves business goals and enables measurement of the results. At this point it is important to revisit the project goals by asking and refining sharp questions that are relevant, specific, and unambiguous.

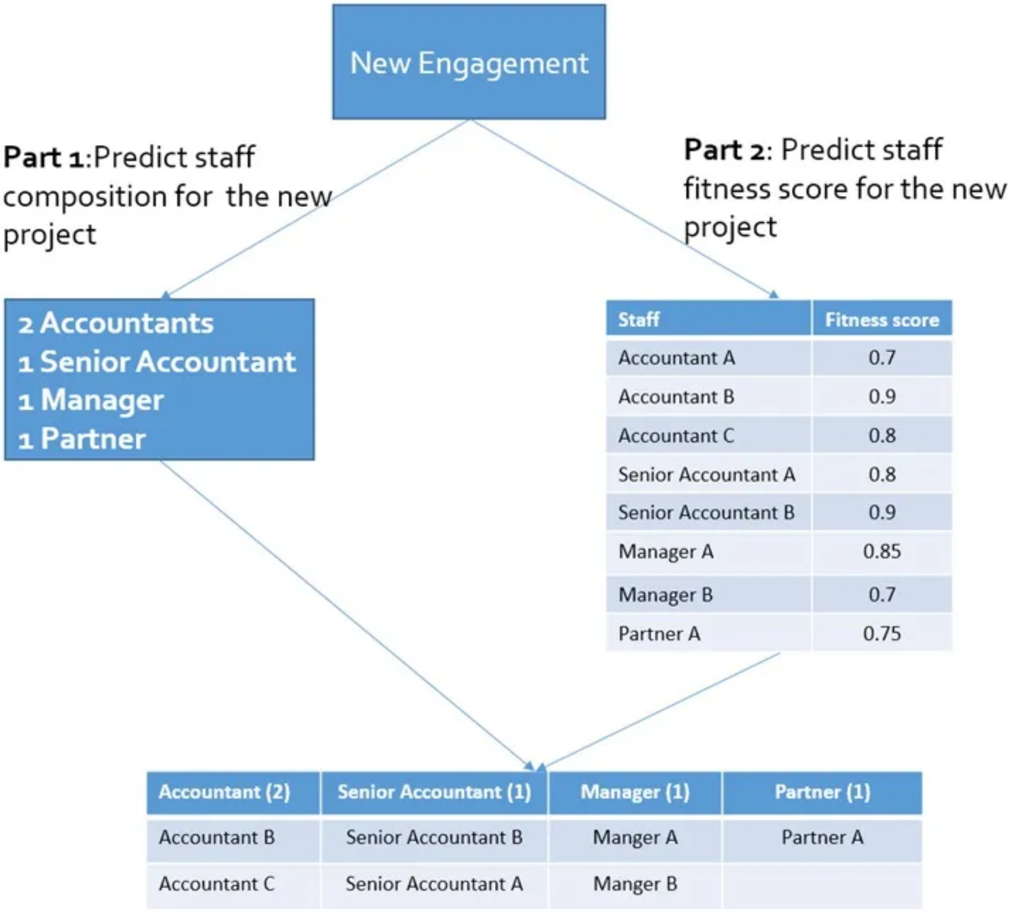
For example, if a company wants to achieve a customer churn prediction, they will need an accuracy rate of "x" percent by the end of a three-month project. With this data, companies can offer customer promotions to reduce churn.

In the case of our professional services company, we decided to tackle the first business question (How can we predict staff composition, e.g. one senior accountant and two accounting assistants, for a new project?). For this customer engagement, we used five years of daily historical project data at individual level. We removed any data that had a negative contribution margin or negative total number of hours. We first randomly sample 1000 projects from the testing dataset to speed up parameter tuning. After identifying the optimal parameter combination, we ran the same data preparation on all the projects in the testing dataset.

USE CASE:

By *Francesca Lazzeri, PhD and team.*

Below (Figure 3) is a representation of the type of data and solution flow that we built for this engagement:


New Engagement
Part 1:Predict staff composition for the new project
Part 2: Predict staff fitness score for the new project
2 Accountants
1 Senior Accountant
1 Manager
1 Partner

Staff	Fitness score
Accountant A	0.7
Accountant B	0.9
Accountant C	0.8
Senior Accountant A	0.8
Senior Accountant B	0.9
Manager A	0.85
Manager B	0.7
Partner A	0.75

Accountant (2)	Senior Accountant (1)	Manager (1)	Partner (1)
Accountant B	Senior Accountant B	Manger A	Partner A
Accountant C	Senior Accountant A	Manger B	

Figure 3. Representation of the type of data and solution flow

We used a clustering method: the k-nearest neighbors (KNN) algorithm. KNN is a simple, easy-to-implement supervised machine learning algorithm. The KNN algorithm assumes that similar things exist in close proximity, finds the most similar data points in the training data, and makes an educated guess based on their classifications. Although very simple to understand and implement, this method has seen wide application in many domains, such as in recommendation systems, semantic searching, and anomaly detection.

In this first step, we used KNN to predict the staff composition, i.e. numbers of each staff classification/title, of a new project using historical project data. We found historical projects similar to the new project based on different project properties, such as Project Type, Total Billing, Industry, Client, Revenue Range etc. We assigned different weights to each project property based on business rules and standards. We also removed any data that had negative contribution margin (profit). For each staff classification, staff count is predicted by computing a weighted sum of similar historical projects' staff counts of the corresponding staff classification. The final weights are normalized so that the sum of all weights is 1. Before calculating the weighted sum, we removed 10% outliers with high values and 10% outliers with low values.

For the second business question (How can we compute Staff Fitness Score for a new project?), we decided to use a custom content-based filtering method: specifically, we implemented a content-based algorithm to predict how well a staff's experience matches project needs. In a content-based filtering system, a user profile is usually computed based on the user's historical ratings on items. This user profile describes the user's taste and preference. To predict a staff's fitness for a new project, we created two staff profile vectors for each staff using historical data: one vector is based on the number of hours that describes the staff's experience and expertise for different types of projects; the other vector is based on contribution margin per hour (CMH) that describes the staff's profitability for different types of projects. The Staff

Fitness Scores for a new project are computed by taking the inner products between these two staff profile vectors and a binary vector that describes the important properties of a project.

We implemented this machine learning steps using Azure Machine Learning service. Using the main Python SDK and the Data Prep SDK for Azure Machine Learning, we built and trained our machine learning models in an Azure Machine Learning service Workspace. This workspace is the top-level resource for the service and provides a centralized place to work with all the artifacts we have created for this project.

In order to create a workspace, we defined the following configurations:

Field	Description
Workspace name	Enter a unique name that identifies your workspace. Names must be unique across the resource group. Use a name that's easy to recall and differentiate from workspaces created by others.
Subscription	Select the Azure subscription that you want to use.
Resource group	Use an existing resource group in your subscription, or enter a name to create a new resource group. A resource group is a container that holds related resources for an Azure solution.
Location	Select the location closest to your users and the data resources. This location is where the workspace is created.

When we created a workspace, the following Azure resources were added automatically:

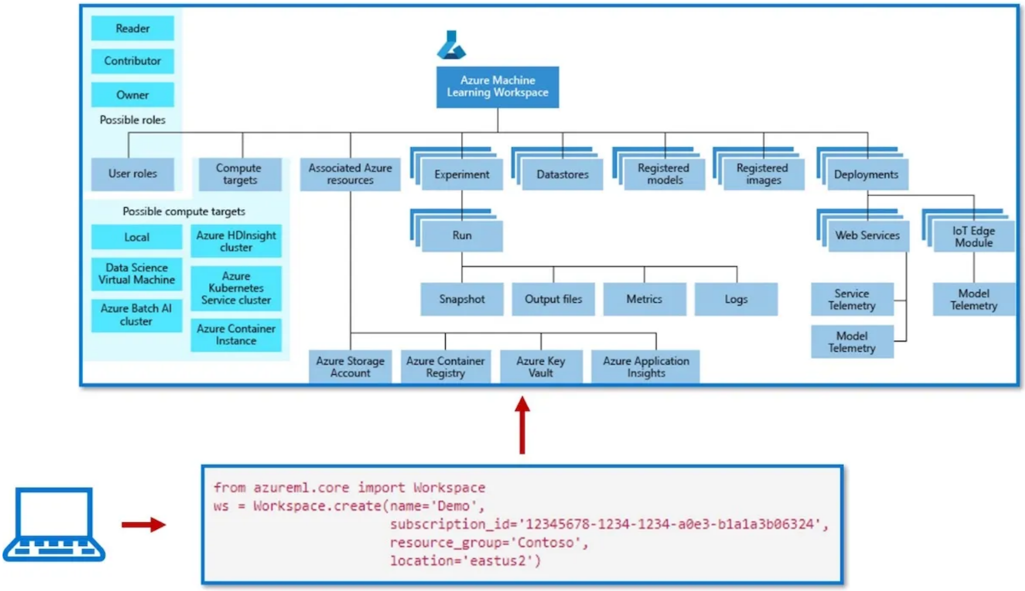
- Azure Container Registry
- Azure Storage
- Azure Application Insights
- Azure Key Vault

USE CASE:

By *Francesca Lazzeri, PhD and team.*

The workspace keeps a list of compute targets that you can use to train your model. It also keeps a history of the training runs, including logs, metrics, output, and a snapshot of your scripts. We used this information to determine which training run produces the best model.

After, we registered our models with the workspace, and we used the registered model and scoring scripts to create an image to use for the deployment (more details about the end-to-end architecture built for this use case will be discussed below). Below is a representation of the workspace concept and machine learning flow (Figure 4):



```
from azureml.core import Workspace
ws = Workspace.create(name='Demo',
                      subscription_id='12345678-1234-1234-a0e3-b1a1a3b06324',
                      resource_group='Contoso',
                      location='eastus2')
```

Figure 4. Workspace concept and machine learning flow

PRINCIPLE 3:

ARCHITECT THE END-TO-END SOLUTION

In the era of Big Data, there is a growing trend of accumulation and analysis of data, often unstructured, coming from applications, web environments and a wide variety of devices. In this third step, organizations need to think more organically about the end-to-end data flow and architecture that will support their data science solutions, and ask themselves the following questions:

- Do they really need this volume of data?
- How do they ensure its integrity and reliability?
- How should they store, treat and manipulate this data to answer my questions?
- And most importantly, how will they integrate this data science solution in their own business and operations in order to successfully consume it over time?

Data architecture is the process of planning the collection of data, including the definition of the information to be collected, the standards and norms that will be used for its structuring and the tools used in the extraction, storage and processing of such data.

This stage is fundamental for any project that performs data analysis, as it is what guarantees the availability and integrity of the information that will be explored in the future. To do this, you need to understand how the data will be stored, processed and used, and which analyses will be expected for the project. It can be said that at this point there is an intersection of the technical and strategic visions of the project, as the purpose of this planning task is to keep the data extraction and manipulation processes aligned with the objectives of the business.

USE CASE:

By *Francesca Lazzeri, PhD and team.*

After having defined the business objectives (*Principle 1*) and translated them into tangible metrics (*Principle 2*), it is now necessary to select the right tools that will allow an organization to actually build an end-to-end data science solution. Factors such as volume, variety of data and the speed with which they are generated and processed will help companies to identify which types of technology they should use.

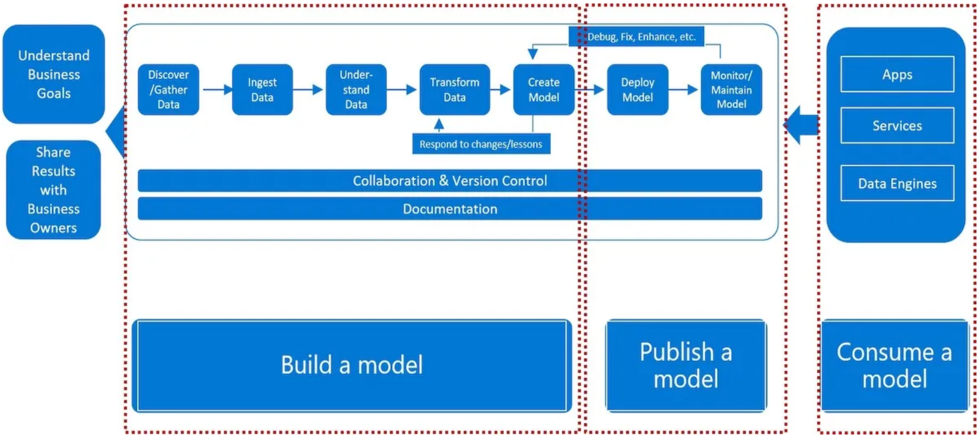
Among the various existing categories, it is important to consider

- Data collection tools, such as Azure Stream Analytics and Azure Data Factory. These are the ones that will help us in the extraction and organization of raw data.
- Storage tools, such as Azure Cosmos DB and Azure Storage: These tools store data in either structured or unstructured form, and can aggregate information from several platforms in

an integrated manner

- Data processing and analysis tools, such as Azure Time Series Insights and Azure Machine Learning Service Data Prep With these we use the data stored and processed to create a visualization logic that enables the development of analyses, studies and reports that support operational and strategic decision-making
- Model operationalization tools, such as Azure Machine Learning service and Machine Learning Server: After a company has a set of models that perform well, they can operationalize them for other applications to consume. Depending on the business requirements, predictions are made either in real time or on a batch basis. To deploy models, companies need to expose them with an open API interface.

The interface enables the model to be easily consumed from various applications, such as:



- Online websites
- Spreadsheets
- Dashboards
- Line-of-business (LoB) applications
- Back-end applications

The tools can vary according to the needs of the business but should ideally offer the possibility of integration between them to allow the data to be used in any of the chosen platforms without needing manual treatments. This end-to-end architecture (Figure 5) will also offer some key advantages and values to companies, such as:

Accelerated Deployment & Reduced Risk: An integrated end-to-end architecture can drastically minimize cost and effort required to piece together an end-to-end solution, and further enables accelerated time to deploy use cases

- Modularity: Allows companies to start at any part of the end-to-end architecture with the assurance that the key components will integrate and fit together
- Flexibility: Runs anywhere including multi-cloud or hybrid-cloud environments
- End-to-End Analytics & Machine Learning: Enables end-to-end analytics from edge-to-cloud, with the ability to push machine learning models back out to the edge for real-time decision making
- End-to-End Data Security & Compliance: Pre-integrated security and manageability across the architecture including access, authorization, and authentication

- Enabling Open Source Innovation: Built off of open-source projects and a vibrant community innovation model that ensures open standards

USE CASE:

By *Francesca Lazzeri, PhD and team.*

In the case of our professional service company, our solution architecture consists of the following components (Figure 6):

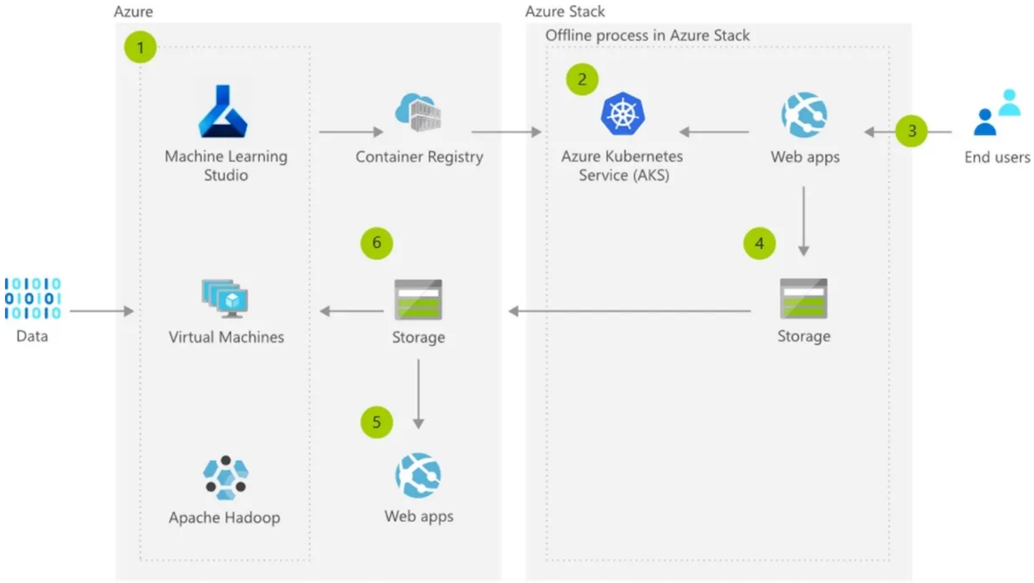


Figure 6. End-to-end architecture developed by Microsoft Azure ML team

1. Data scientists train a model using Azure Machine Learning and an HDInsight cluster. Azure HDInsight is a managed, full-spectrum, open-source analytics service for enterprises. HDInsight is a cloud service that makes it easy, fast, and cost-effective to process massive amounts of data. The model is containerized and put into an Azure Container Registry. Azure Container Registry allows you to build, store, and manage images for all types of container deployments. For this specific customer engagement, we created an Azure Container Registry instance using the Azure CLI. Then, use Docker commands to push a container image into the registry, and finally pull and run the image from your registry. The Azure CLI is a command-line tool providing a great experience for managing Azure resources. The CLI is designed to make scripting easy, query data, support long-running operations, and more.

2. The model is deployed via an offline installer to a Kubernetes cluster on Azure Stack. Azure Kubernetes Service (AKS) simplifies the management of Kubernetes by enabling easy provisioning of clusters through tools like Azure CLI and by streamlining cluster maintenance with automated upgrades and scaling. Additionally, the ability to create GPU clusters allows AKS to be used for high-performing serving, and auto-scaling of machine learning models.

3. End users provide data that is scored against the model. The process of applying a predictive model to a set of data is referred to as scoring the data. Once a model has been built, the model specifications can be saved in a file that contains all of the information necessary to reconstruct the model. You can then use that model file to generate predictive scores in other datasets

4. Insights and anomalies from scoring are placed into storage for later upload. Azure Blob storage is used to store all project data. Azure Machine Learning Service integrates with Blob storage so that users do not have to manually move data across compute platforms and Blob storage. Blob storage is also very cost-effective for the performance that this workload requires.

5. Globally-relevant and compliant insights are available in the global app. Azure App Service is a service for hosting web applications, REST APIs, and mobile back ends. App Service not only adds the power of Microsoft Azure to your application, such as security, load balancing, autoscaling, and automated management. You can also take advantage of its DevOps capabilities, such as continuous deployment from Azure DevOps, GitHub., Docker Hub, and other sources, package management, staging environments, custom domain, and SSL certificates.

6. Finally, data from edge scoring is used to improve the model.

PRINCIPLE 4:

BUILD YOUR TOOLBOX OF DATA SCIENCE TRICKS

USE CASE:

By *Francesca Lazzeri, PhD and team.*

When working on the recommendation-based staff allocation solution for our professional services company, we immediately realized that they were limited in time and didn't have an infinite amount of computing resources. How can organizations organize their work so that they can maintain maximum productivity?

We worked closely with our customer's data science team and helped them develop a portfolio of different tricks to optimize their work and accelerate production time, for example:

- Train on a subset much smaller than the whole data set you have first: Once data science teams have a clear understanding of what they need to achieve in terms of features, loss function, metrics, and values of hyperparameters, then scale things up.

- Reuse knowledge gained from previous projects: Many data science problems are similar to one another. Reusing the best values of hyperparameters or feature extractors from similar problems other data scientists solved in the past will save organizations a lot of time.
- Setup automated alerts that will inform data science teams that a specific experiment is over: This will save data science teams time in case something went wrong with the experiment.
- Use Jupyter notebooks for quick prototyping: Data scientists can rewrite their code into Python packages/classes once they are satisfied with the result.
- Keep your experiment code in a version control system, such as GitHub.

- Use pre-configured environments in the cloud for data science development: These are virtual machine images (such as Windows Virtual Machines and Azure Data Science Virtual Machine), pre-installed,

- configured and tested with several popular tools that are commonly used for data analytics, machine learning training.

- Have a list of things to do while experiments are running: data collection, cleaning, annotation; reading on new data science topics, experimenting with a new algorithm or a framework. All those activities will contribute to the success of your future projects.

PRINCIPLE 5:

UNIFY YOUR ORGANISATION'S DATA SCIENCE VISION

Right from the first day of a data science process, data science teams should interact with business partners. Data scientists and business partners get in touch on the solution non-frequently. Business partners want to stay away from the technical details and so do the data scientists from business. However, it is very essential to maintain constant interaction to understand implementation of the model parallel to building of the model. Most organizations struggle to unlock data science to optimize their operational processes and get data scientists, analysts, and business teams speaking the same language: different teams and the data science process are often a source of friction. That friction is what defines the new data science iron triangle and is based on a harmonic orchestration of data science, IT operations, and business operations.

USE CASE

By *Francesca Lazzeri, PhD and team.*

In order to accomplish this task with our customer, we implemented the following steps:

- Request the support of a C-level executive to take that vision and drive it through the different parts of the business: Where there is a clear purpose, vision and sponsorship, the taste of initial success or early wins spurs further experimentation and exploration, often resulting in a domino effect of positive change.

- Build a culture of experimentation: Even where there is a clearly articulated purpose, that alone often doesn't lead to successful business transformation. An important obstacle in many organizations is the fact that employees just aren't empowered enough to bring about change. Having an empowered workforce helps engage your employees and gets them actively involved in contributing towards a common goal.

Involve everyone in the conversation: Building consensus will build performance muscle. If data scientists work in silos without involving others, the organization will lack shared vision, values and common purpose. It is the organization's shared vision and common purpose across multiple teams that provide synergistic lift.

PRINCIPLE 6:

KEEP HUMANS IN THE LOOP

Becoming a data-driven company is more about a cultural shift than numbers: for this reason, it is important to have humans evaluate the results from any data science solution. Human-data science teaming will result in better outcomes than either alone would provide.

USE CASE

By *Francesca Lazzeri, PhD and team.*

For instance, in the case of our customer, using the combination of data science and human experience helped them to build, deploy and maintain a workforce placement recommendation solution that recommends optimal staff composition and individual staff members with right experience and expertise for new projects, which often led to monetary gains.

After we deployed the solution, our customer decided to conduct a pilot with a few project teams. They also created a v-Team of data scientists and business experts whose purpose was to work in parallel with the machine learning solution and compare the machine learning results in terms of project completion time, revenue generated, employees and customer satisfactions from these two pilot teams before and after using Azure Machine Learning's solution. This offline evaluation conducted by a team of data and business experts was very beneficial for the project itself because of two main reasons:

1. It confirmed that the machine learning solution was able to improve ~4/5% of contribution margin for each project;
2. The v-Team was able to test the solution and create a solid mechanism of immediate feedback that allowed them to constantly monitor the results and improve the final solution.

After these pilot projects, the customer successfully integrated our solution within their internal project management system.

There are a few guidelines that companies should keep in mind when starting this data-driven cultural shift:

- Working side by side: leading companies increasingly recognize that these technologies are most effective when they complement humans, not replace them. Understanding the unique capabilities that data science and humans bring to different types of work and tasks will be critical as the focus moves from automation to the redesign of work.
- Recognizing the human touch: It is important to remember that even jobs with greater levels of computerization have to maintain a service-oriented aspect and be interpretive in order to be successful for the company — these roles, like data scientists and developers, still need essential human skills of creativity, empathy, communication, and complex problem-solving.

- Investing in workforce development: A renewed, imaginative focus on workforce development, learning, and career models will also be important. Perhaps most critical of all will be the need to create meaningful work—work that, notwithstanding their new collaboration with intelligent machines, human beings will be eager to embrace.

The human component will be especially important in use-cases where data science would need additional, currently prohibitively expensive architectures, such as vast knowledge graphs, to provide context and supplant human experience in each domain.

CONCLUSION

By applying these six principles from the Healthy Data Science Organization Framework on data analysis process, Data Scientists as well as organizations can make better decisions for their business. Their choices will be backed by data that has been robustly collected and analyzed.

These six transformational data science mindset principles should be implemented as a cross-functional tool with everyone's hand on deck.

Francesca Lazzeri's customer was able to implement a successful data science solution that recommends optimal staff composition and individual staff with the right experience and expertise for new projects. By aligning staff experience with project needs, they helped project managers perform better and faster staff allocation.

With the right data science mindset, data science processes will get faster and more accurate – meaning that organizations will make better, more informed decisions to run operations most effectively.

ABOUT THE AUTHOR

Dr. Bright holds a bachelor's in Mathematics, a master's in Data Science, and a Ph.D. in Data and Information Systems. He has worked as a Data Scientist and Product Manager in many Startups and MNCs including Amazon, Deloitte, Microsoft, ACBSP, and Synacor. He is an entrepreneur, multinational business owner, and career coach.

For **MENTORSHIP** contact: drbriit@gmail.com.

Sign-up to get a FREE copy whenever there is an update to this book.