

THE ANALYTICS EDGE: HOW TO TURN DATA INTO COMPETITIVE ADVANTAGE

Here are innovative practices related to data and analytics from top experts at MIT Sloan, the business school for analytics.



CONTENTS

- Try this data framework for analytics advantage 2
- Data literacy for leaders 8
- In-demand data and analytics skills to hire for now 12
- How to build an effective analytics practice: 7 insights from MIT experts 16
- What is synthetic data — and how can it help you competitively? 21



CREDIT: MARYSIA MACHULSKA

Try this data framework for analytics advantage



by Beth Stackpole

Why It Matters

A framework based on data, models, decisions, and value can help you leverage analytics for better business outcomes.

Data has the power to help drive efficiencies, increase profitability, and bolster innovation.

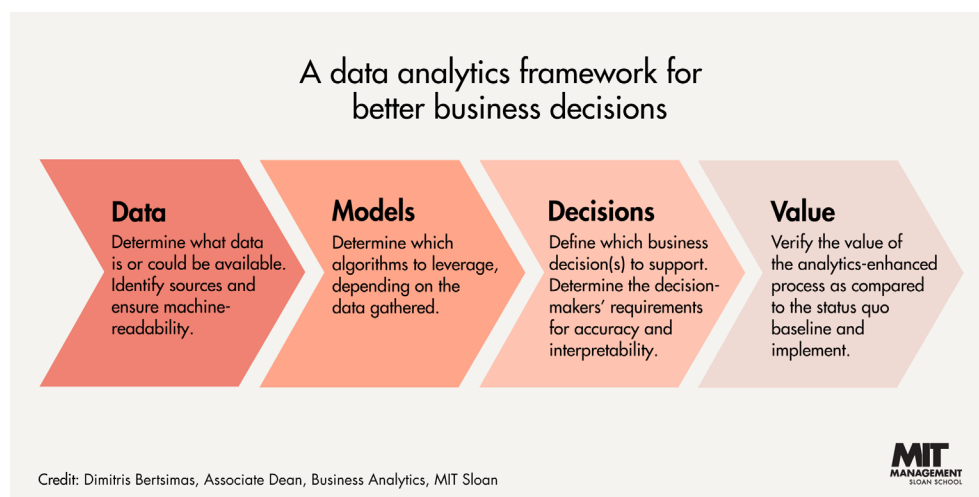
Yet analytics skills remain outside the wheelhouse of most business leaders, who typically have little experience interpreting data from multiple sources or identifying what data sources are necessary to formulate the insights they are seeking.

Shoring up this experience gap among leadership can make all the difference in turning data-driven business aspirations into tangible and decisive actions.

With that goal in mind, [Dimitris Bertsimas](#), the associate dean of [business analytics at MIT Sloan](#), has developed an approach to help nontechnical business users gain confidence using data to improve their decision-making.

In his online course for executives titled [Applied Business Analytics](#), Bertsimas, a professor of operations research, lays out an analytics framework that is designed to help businesspeople determine which analytics approach is best suited for their application and improve their ability to leverage big data for better business outcomes.

The framework is built upon data, models, decisions, and value, with an emphasis on decision-making. “The two protagonists in this process are data and decisions,” Bertsimas says in the introduction to the course. “Analytics leaders may understand the basics of the modeling, but it is their skillful handling of the data and the decisions that gives them an edge.”



The framework at a glance

The framework begins with data, which Bertsimas calls “the most important aspect” of analytics. “Finding the right data, cleaning it, and shaping it so that it works toward your need is a skill that takes experience and understanding,” he says.

Next, organizations use the data they have gathered to decide which models to use; the outcome of those models influences decisions around onboarding a data team, choosing the best tools, selecting the right variables, and outputting effective visuals — all of which should ultimately result in value to the organization.

As an example, here's how the data analytics framework would look when used by an equity firm — let's call it Acme Equity — planning to make a large investment in a growth firm it has identified. Elements would include:

Data. Acme Equity needs to collect relevant data on the earnings of the target firm, including the consistency, repeatability, and quality of the earnings.

Models. To understand how those earnings may grow, Acme needs to develop a predictive model that encapsulates all factors affecting earnings and considers past situations. There's also the question of choosing the appropriate model, which in this scenario could be linear regression, a machine-learning tree-based model, or others.

Decisions. This stage is all about evaluating risk and making informed decisions based on the outcomes from various models, which should be able to predict the drivers that will lead to high or low revenue.

Value. By this stage, Acme Equity should be able to set clear expectations about the potential value of investing in its target company, including the level of return from a potential investment and the ability to maintain a relatively low risk profile.

“Finding the right data, cleaning it, and shaping it ... is a skill that takes experience and understanding.”

DIMITRIS BERTSIMAS

PROFESSOR OF OPERATIONS RESEARCH, MIT SLOAN

Three types of algorithms

Central to any business analytics initiative are algorithms, which Bertsimas defines as “a set of guidelines that describe how to perform a particular task.”

For business leaders still struggling to separate logistic regression from integer optimization (both covered in the class), it's beneficial to think of algorithmic options as three buckets:

Descriptive — for identifying patterns or converting images and text to numbers. A descriptive algorithm might be used to group movies by genre and create a personalized movie recommendation engine — an approach popularized by Netflix.

Predictive — for forecasting outcomes, such as whether a person is likely to default on a loan or how many games a baseball team will win in the next season.

Prescriptive — for recommending next steps, such as suggesting a particular insurance policy or investment selection.

Determining which algorithm is the best fit requires an understanding of the type of data and how it's organized, Bertsimas says.

Structured data is organized in rows and columns, where rows represent data points and columns represent a variable/feature of the data. Unstructured data is any data not organized in that way — for example, voice, video, text, or tweets.

Unstructured data eventually needs to be converted to structured data to ensure a complete and integrated data set — a task that is much on business leaders' minds. According to analytics management firm Komprise, 87% of IT leaders surveyed in 2022 said managing unstructured data growth is a top priority, up from 70% the year previous.

Pro tips for analytics success

Data analytics is a complex and highly specialized endeavor. But with a pragmatic approach, the practice is accessible to more business users, enabling companies of all stripes to reap the rewards of data-driven decision-making.

Setting an objective and mapping a problem statement to the framework are important first steps, Bertsimas says, but equally important is identifying both roadblocks and unanticipated opportunities, and planning for those.

To that end, Bertsimas enlisted the help of [Jordan Levine](#), previously an engagement manager at McKinsey and now a lecturer in operations research and statistics at MIT Sloan, to develop a series of pro tips gleaned from real-world experience.

Among the tips for rolling out an analytics project:

Foster stakeholder agreement. Leveraging analytics to solve complex business problems necessitates some assumptions. But it's also important to ensure that all stakeholders are aligned on a methodology that comprises inclusion and exclusion criteria governing data and models. If stakeholders don't agree on the basics, they aren't likely to accept the recommendations of the analysis, and the effort is for naught.


Enlist the right leader. An analytics leader with political capital can ensure more effective collaboration between junior tech talent and domain experts. The right leader will engage the appropriate subject matter experts to determine what data will be most useful and coach teams to build trust. Ultimately, the choice of leader will help accelerate time to value for analytics efforts and shorten model production cycles.

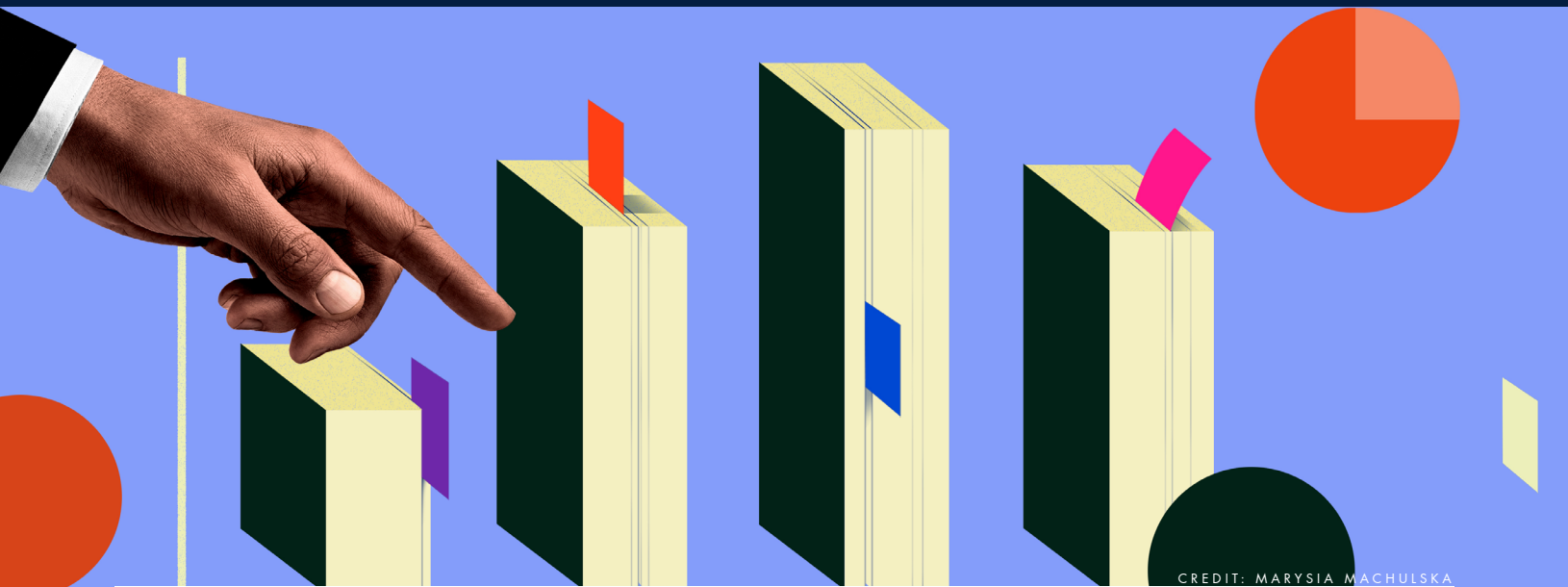
Determine the appropriate data set and size. Data sets can come from numerous places, including data markets, public records, known databases, and self-generated data, as well as the [API economy](#). Analytics leaders should work with both subject matter experts and data scientists to determine the appropriate data sources, along with the right training set and test set data size. Typically, the range for the training set is between 50% and 80% of data, although larger data sets require less data for training models and more for testing. One important rule of thumb: Consider how often your model should be refitted to respond to changes over time.

Never assume that a data set is accurate. It's important to invest time up front to make a data set readable. Eliminate messy formatting, inconsistent terminology, and the use of abbreviations. Ask for summary tables and [univariate visualizations](#) to ensure that the raw data makes sense. A rookie mistake is failing to plan enough time for this phase; a conservative estimate is that 60% to 80% of time spent on a data analytics project will be devoted to acquiring and cleaning up data. It's also important to make time for data inspection, as it helps teams gain a better understanding of model accuracy.

Consider data security ramifications. Often, when you buy data, it's been sanitized to exclude personally identifiable information such as Social Security numbers, email addresses, or bank account numbers. Make sure legal counsel is involved if you are negotiating to acquire data that does include PII.

Establish a model baseline. It's important to have a baseline to compare against more sophisticated models to help determine whether more data, effort, financial investment, or computational power is required to improve predictions. It's common to see diminishing returns on model improvement in relation to complexity; it's also common to encounter issues that make the model less useful when applying it to new data.

Don't reinvent the wheel. There are software packages available that will take a data set from sparse to dense, which can free up the data science team to focus on other tasks. Spend time exploring the optimization and open-source community to find available packages that can be tapped for broader use so the team doesn't have to create everything from scratch. 



CREDIT: MARYSIA MACHULSKA

Data literacy for leaders



by Sara Brown

Why It Matters

Leaders need to understand data enough to make their best decisions, drive literacy throughout the organization, and create a culture of trust in data.

Data scientists might be in demand, but data literacy starts with leaders. Leaders need to trust and understand data well enough to make good decisions, and they must also drive literacy efforts throughout the organization and create a culture of trust in data.

Data literacy — the ability to work with and understand data to drive business impact — “is as essential a skill as negotiation, communication, management, people management, all of that,” said Piyanka Jain, president and CEO of Aryng, a data science consulting company. Executives need to lead with numbers to gain competitive advantage, she said. “Your competitors are.”

Historically, business school programs have offered courses on data modeling, querying, and architecture, but rarely look at broader data strategy, said Barbara Wixom, a principal research scientist at the MIT Center

for Information Systems Research. “Leaders have to understand and appreciate that void, because it puts even more pressure on them to be literate and drive literacy in their organizations,” she said.

Here’s what leaders need to know about data literacy — from how to establish trust to being appropriately skeptical — and steps to getting there.

Trusting data and promoting its use

Modern leadership skills include trusting in data to guide decisions and knowing when to question results.

Trusting in data can be difficult for leaders, Jain said, because they are good at decision-making but often rely on their intuition. Asking someone to give away some power and accept guidance from analytics is hard. But pairing judgment and intuition with insights from data makes leaders stronger.

Leaders aren’t necessarily involved in creating or analyzing data, but they often make decisions based on analytics. “The goal for a leader, from a data literacy perspective, should be, ‘How can I be a fast but effective consumer of analysis that is produced by my organization?’” said MIT Sloan professor of the practice [Rama Ramakrishnan](#), formerly a senior vice president at Salesforce.

Leaders also are responsible for establishing data literacy in their organizations.

Important steps include defining what data literacy means for your company and for different roles, establishing a baseline of data literacy skills; building a culture of curiosity around using data, and defining success. (For more details, see “[How to build data literacy in your company](#).”)

It’s also important to establish purposeful data vocabulary beyond buzzwords, Wixom said.

Data literacy requires “top leaders committing to not just really learning the language but then using the language in understandable, consumable ways so that everyone across the organization will get on board,” she said. Wixom said her research has found that it’s

“The goal for a leader, from a data literacy perspective, should be, ‘How can I be a fast but effective consumer of analysis that is produced by my organization?’”

RAMA RAMAKRISHNAN
PROFESSOR, MIT SLOAN

particularly important to distinguish between data and a data asset, which has been purposely prepared for future value creation.

Understanding your firm's data processes

Data literacy requires hands-on, ongoing effort. It's not a matter of completing one training session, Jain said — it's a fundamental shift in thinking.

Changing how leaders make decisions takes time. But it's faster and easier with a systematic approach. "If you think about data literacy as training and a checkbox, you will waste your time and your money, and you will waste the time of your organization," Jain said. Creating new business value through data and defining what data literacy means to leaders and their organizations should be part of the process. "It's a commitment to long-term learning and changes," she said.

While reading case studies is a good way to learn about what goes into data initiatives, it's hard to understand data, and data terms, unless you've worked with it, Wixom said. She recommended that leaders get involved with data projects. "Data is such an active kind of concept," she said. Talking about data quality can be vague but trying to work with or remediate a bad data set helps users appreciate why it matters and what's needed to overcome challenges.

Wixom suggested that leaders focus on answering three main questions:

How does a data initiative lead to actual financial returns? Data monetization — how data leads to economic returns — is a vital component of a data strategy. "Being familiar with examples of initiatives and how they create value in different ways — I think that's really important for literacy for a leader," Wixom said. "If you're trying to establish a strategy that involves some collection of initiatives, you need to really understand those initiatives and how they contribute to your firm's financial performance."

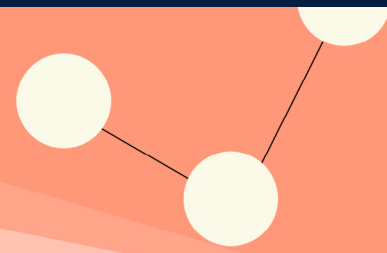
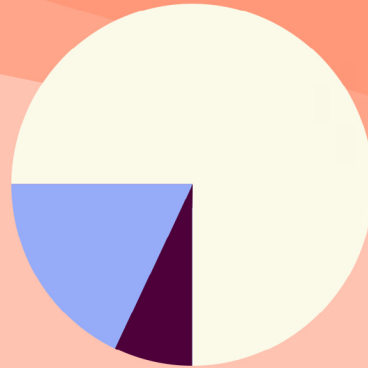
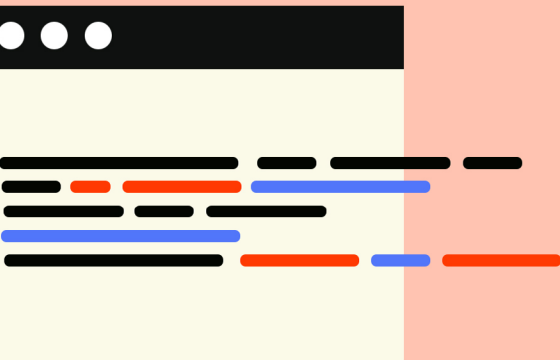
What are the actual practices behind creating data assets? Leaders need to understand the capabilities that support data strategy — things like master data management, metadata, or data catalogs — at a foundational level. "This is stuff that's very tangible," Wixom said. "When you're talking about making investments in these types of practices, leaders need to really pay attention and hear what's involved."

How are people learning about data and from data? Leaders “have to really appreciate how they’re growing the organization in its data literacy,” Wixom said. Consider establishing centers of excellence or embedding data experts in the company.

Being savvy consumers of company data

At the same time, leaders need to be able to evaluate data and be skeptical when appropriate. Leaders should keep the following in mind:

- 1. Before being shown data, think about what you expect to see.** That way, “the contrast between what you expected and what is actually showing up will just jump out at you,” Ramakrishnan said. This will often quickly highlight the most relevant parts of a report. “This is a habit worth cultivating, and you’ll get better at it with practice,” he said.
- 2. Remember that data is uncertain.** There is almost always a degree of uncertainty in data, Ramakrishnan said. He cited statistician John Tukey, who said “The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.” Leaders will need to live with this uncertainty, he said, or ask their team to get more data to provide more information and reduce uncertainty.
- 3. Use the “common sense” test.** If you’re making an important decision based on data, put it to the test. “If something’s true ... chances are, different data paths lead the same truth,” Ramakrishnan said. “If you’re about to make an important decision based on one analysis, try your best to get another team or another data set to be analyzed to see if it points in the same direction.”
- 4. Don’t confuse causation and correlation.** “Whenever you’re looking at an analysis that suggests that some factor is driving some outcome, always try to figure out how much of it is correlational and how much of it is causal,” Ramakrishnan said. “They should really brainstorm with the management team on what else could explain this alleged cause-and-effect relationship.” 🏛️



CREDIT: MARYSIA MACHULSKA

In-demand data and analytics skills to hire for now



by Meredith Somers

Why It Matters

Companies are looking for more than programming language fluency when it comes to hiring for data and analytics roles. Four hiring managers share their must-have lists.

As the quantity and quality of information improve, managers face increasing pressure to leverage their data to be competitive in the digital era, launch new customer strategies, and get the most out of their teams. Without the right talent analyzing that data and translating it for decision-makers, companies will be left behind.

“Everything starts with the data,” said Nadine Kawkabani, global business strategy director at MFS Investment Management and an industry host for MIT Sloan’s business analytics capstone program.

When she joined the program in 2018, the line she often heard was that data was doubling every two years. Now, it’s every 18 months. Because the number



Nadine Kawkabani
Global Business Strategy
Director at MFS Investment
Management

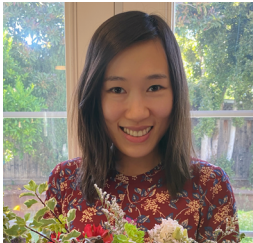
of people hired to work with that data isn't moving as fast, Kawkabani said, the question becomes, "How can we be more efficient at bringing up the trends or what the data is telling us, or how can we rule out the noise?"

Kawkabani and hiring managers from Comcast, Netflix, and Pfizer shared the technical skills that are essential at their companies and the soft skills they look for when it comes to filling today's data-centric roles.

Python, R, SQL, and "the right math"

While acknowledging the proliferation of data and analytics undergraduate and master's programs, the hiring managers said they don't require an advanced degree when looking for a data expert. But they all agreed that fluency in basic

programming is a common denominator for anyone considering a data job.



Yichen Sun, SM '13
Data Science Manager
at Netflix

"Whether it's an engineer or data scientist or research scientist, SQL or Python are the required programming language — Python or R, depending on the candidate's preference," said Yichen Sun, SM '13, who leads a team of engineers and data scientists at Netflix.

Sun said she also looks for proficiency in basic sampling techniques like A/B testing, and causal inference techniques like difference in differences.

Trace Hawkins, senior vice president of strategic analytics at Comcast, encourages people who know SQL to start learning Python and vice versa, though he said he can find a role for someone regardless of which programming language they prefer. What's nonnegotiable is the way someone interprets and analyzes the data.



Trace Hawkins
SVP, Strategic Analytics
(Enterprise Business
Intelligence) at Comcast

"In Python, you can generate a match pair comparison population, but do you know how to do it right? Do you

actually understand the difference between a good match pair and a bad match pair? And how would you evaluate your clustering algorithm for whether your segments are mutually exclusive and collectively exhausted?” Hawkins asked. “All of the things you might do there, you need to understand methodologically how to validate that your math was the right math.”

In search of unicorns

Hawkins and [Jonathan Lowe](#), the data science lead for Pfizer Global Supply Operations Insights, both said they look for unicorns — not \$1 billion companies in this case, but data experts with coveted second skill sets to apply at their companies.



Jonathan Lowe
Data Science Lead
at Pfizer

Hawkins looks for data workers who can translate their findings to a business audience. Lowe said his “super unicorns” are the data scientists who also happen to have consulting skills and love developing software.

“There’s a fourth category, too, which sometimes we make an exception and hire for even without the other [skills], which is domain expertise,” Lowe said. “If somebody says, ‘I’ve worked in a quality lab for half my career and now, for the last several years, I’ve been learning

more data science,’ we will gobble those people up.”

State-of-the-art technologies aren’t always the best solution in the Netflix production environment, where Sun’s team needs to consider computational cost, consumer experience, privacy requirements, data infrastructure readiness, and more.

“We therefore need someone to be both principled and practical, make the right trade-offs, and to be able to articulate the ‘why’ behind such technical decisions,” Sun said.

“We need someone ... to be able to articulate the ‘why’ behind technical decisions.”

YICHEN SUN, SM '13
DATA SCIENCE MANAGER, NETFLIX

Communication, curiosity, collaboration

Bridging the gap between the business and data sides of a company are top priorities for hiring managers, with each emphasizing the importance of accurately translating the information gleaned from data into actionable business strategies.

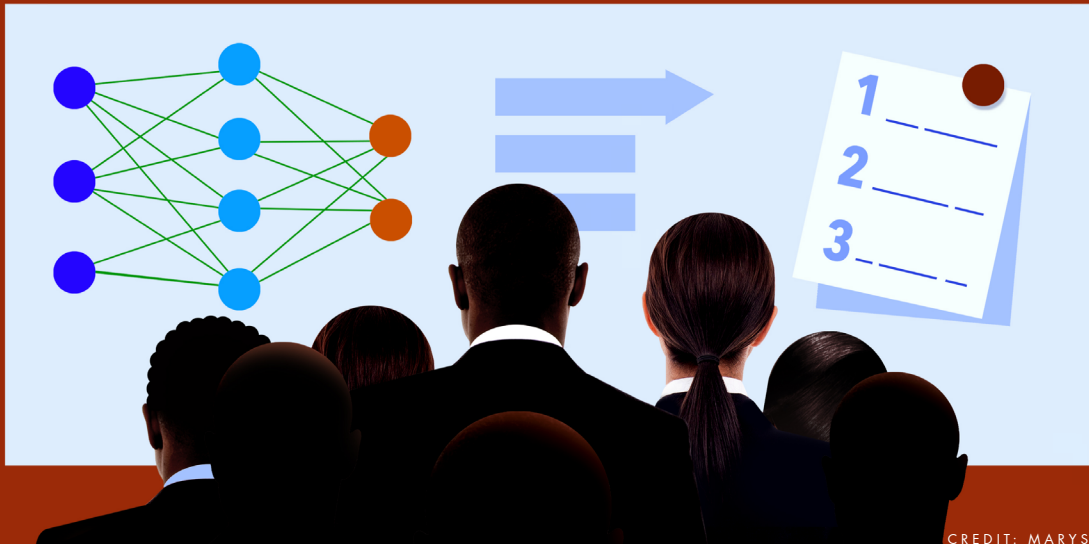
“Storytelling skills would be another way to describe this capability,” Lowe said. “[Don’t] just blurt out a bunch of technical jargon but tell a story around why the business needs this [data] support and what will happen if the business uses what you’ve built.”

Today’s data-centric roles also require curiosity, which contributes to an innovative and problem-first mindset. While a data expert with a solution in search of a problem isn’t a deal breaker, Sun said she will try to coach the person into understanding that their solution might be the right application for a problem but that there might be an “even more elegant or even simpler way to do it.”

Relatedly, Sun also looks for “someone who’s more reflective, who is able to receive this feedback in a very productive way and be adaptable in terms of what approach they use.”

These and other soft skills are examples of how data and analytics jobs — and the related culture — have changed, Kawkabani said. It’s no longer about handling data with blinders on; it’s about ensuring that the data makes sense and that the people who are touching the data also understand how they’re impacting the strategy of the firm.

“We’re all relying on each other,” Kawkabani said. “I can put the best strategy out there, but if I don’t have good data, nice graphs, accurate data, and timely, interpretable data, it doesn’t mean anything.” 🏛️



CREDIT: MARYSIA MACHULSKA

How to build an analytics practice: 7 insights from MIT experts



by Tracy Mayor

Why It Matters

MIT Sloan business analytics faculty share mistakes to avoid, strategies to adopt, and the technological developments that excite them most.

Faculty members in MIT Sloan's Master of Business Analytics program don't just teach courses like Hands-On Deep Learning and Econometrics for Managers.

They also have deep experience applying the tools of modern data science, optimization, and machine learning to solve real-world business problems for organizations like Facebook, Salesforce, Booz Allen Hamilton, and the intelligence wing of the Israeli Defense Forces.

Here, they share insights about mistakes to avoid, strategies to adopt, and the developments in analytics and data that excite them most.

Let business decisions drive data strategy

Y. Karen Zheng, Associate Professor, Operations Management

One of the biggest mistakes companies make about analytics is the disconnect between the technology and real business decisions. Companies tend to collect data for the sake of having data, and to develop analytics for the sake of having analytics, without thinking about how they are going to use the data and analytics capabilities to inform business decisions.

Successful analytics organizations are always decision-driven. They start by asking what business decisions they need data and analytics for, then investing resources to collect the right data and build the right analytics.

It is equally important to have the right talents in the organization that speak both the language of analytics and business, so that they can be the bridge between the tech and the business decision-makers. These employees understand the business needs and how analytics can be utilized to satisfy those needs; at the same time, they're able to communicate the tech solution to business decision-makers in an understandable, intuitive way, as opposed to delivering a "black-box solution" that is hardly adopted by humans.

Use deep learning to get value from unstructured data

Rama Ramakrishnan, Professor of the Practice, Data Science and Applied Machine Learning

I am personally most excited by deep learning. Traditional analytics methods are very effective for structured data, but we weren't previously able to get value from unstructured data — images, audio, video, natural language, and so on — without a lot of labor-intensive preprocessing.

With deep learning, this limitation is effectively gone. We can now leverage unstructured and structured data together in a single, flexible, and powerful framework and achieve significant gains relative to what we could do earlier. This is possibly the most significant analytics breakthrough that I have witnessed in my professional career.

Pick use cases that deliver value

Jordan Levine, Lecturer, Operations Research and Statistics

The strategy to build an analytics practice is simple. First, identify three sources of use cases and start to build them. The three sources include:

- Use cases that support C-level metrics (think revenue, cost, and risk).
- Business processes that can be supported by self-serve analytics and dashboards.
- Compliance must-do activities.

I use these three sources because they will be looked at differently by the ultimate scorekeepers — the finance function.

The second important activity is to staff the bench to meet demand once these use cases start driving value. Companies will often erroneously hire for a variety of roles and think the work is done. Given the fluidity of the post-COVID work environment and often short tenures of scarce analytics talent, companies must not only staff up but establish pipelines of talent.

One way companies do this is to partner with a higher-ed organization like the MIT Applied Business Analytics program, work with students on capstone projects, hire those students when they graduate, and then ask them to work with a new crop of students. This virtuous cycle ensures a happy, competent, and staffed bench of analytics talent.

Develop a new organizational language based on data-enabled models

Retsef Levi, Professor, Operations Management

Data and analytics technologies are a critical enabler to create intelligent workflow and decision processes and systems. That said, many companies think about this through a technical lens and miss the fact that this is an end-to-end organizational challenge.

The opportunity to design intelligent decision processes emerges from the ability to sense the organizational environment better than ever. It requires a new organizational language based on data-enabled models. Organizations must deeply understand their existing decision processes and the data they generate,

and then develop layers of data-enabled models to allow the design of innovative intelligent decision processes. To be successful, it's critical that organizations understand and manage required changes in decision rights and workforce role definitions.

Embrace the full analytics pipeline, upstream and downstream

Alexandre Jacquillat, Assistant Professor, Operations Research and Statistics

Most analytics projects in practice are focused on the development of deep learning and artificial intelligence tools. This is the shiny object that any analytics team is trying to build, improve, and deploy, with an emphasis on technical performance indicators — “My accuracy is 87%,” and so forth.

However, these represent only a narrow subset of the full analytics pipeline, which spans data management, descriptive analytics (such as data visualization and pattern recognition), predictive analytics (using machine learning tools, including but not restricted to deep learning), prescriptive analytics (using optimization), and business impact.

Time and time again, analytics projects take shortcuts across that pipeline — upstream and downstream.

At the upstream level, many analytics teams forgo critical steps to ensure the quality of their data, the representativeness of their data, and their own understanding of their data. One remedy for that is systematic exploratory data analysis baked into the analytics pipeline.

At the downstream level, analytics teams oftentimes fail to address the challenges associated with complex, large-scale decision-making in complex systems. This is where analytics projects could gain an additional edge by systematically embedding predictive tools into prescriptive analytics pipelines and decision-support systems.

Address specific business use cases to improve decision-making

Daniel Freund, Assistant Professor, Operations Management

With all the hype around machine learning, it is easy to forget that predictions are most useful when they inform decision-making. I've seen organizations roll out predictive models that weren't going to inform actual decisions at all.

But even if a predictive model directly feeds into decision-making, improving predictions doesn't always improve the decisions. Instead, new analytics capabilities are most powerful when they're done to address specific business use cases to improve decision-making.

To ensure successful outcomes, it's best for companies to measure these capabilities by the quality of the decisions they produce rather than just the accuracy of the predictions feeding into them.

Establish a centralized system for randomized experiments

Dean Eckles, Associate Professor, Marketing

Firms building an analytics process must have consistent definitions and practices. The foundation for trustworthy analytics is consensus on how basic metrics are defined and how common analyses are conducted.

This is sometimes one more indirect benefit of setting up a centralized system for randomized experiments (A/B tests and beyond): It often requires figuring out what metrics will show up when analyzing a given test, and this requires getting teams to agree on just how particular metrics are defined — be it number of days active, time spent on site, or even ad revenue per user.

These benefits are on top of the more direct benefits of making it easier to run experiments and making their results standardized and trustworthy. 🏛️



What is synthetic data — and how can it help you competitively?



by Brian Eastwood

Why It Matters

Synthetic data — which resembles real data sets but doesn't compromise privacy — allows companies to share data and create algorithms more easily.

Companies committed to data-based decision-making share common concerns about privacy, data integrity, and a lack of sufficient data.

Synthetic data aims to solve those problems by giving software developers and researchers something that resembles real data but isn't. It can be used to test machine learning models or build and test software applications without compromising real, personal data.

A synthetic data set has the same mathematical properties as the real-world data set it's standing in for, but it doesn't contain any of the same information. It's generated by taking a relational database, creating a generative machine learning model for it, and generating a second set of data.

The result is a data set that contains the general patterns and properties of the original — which can number in the billions — along with enough “noise” to

mask the data itself, said [Kalyan Veeramachaneni](#), principal research scientist with MIT's [Schwarzman College of Computing](#).

Gartner has [estimated](#) that 60% of the data used in artificial intelligence and analytics projects will be synthetically generated by 2024. Synthetic data offers numerous value propositions for enterprises, including its ability to fill gaps in real-world data sets and replace historical data that's obsolete or otherwise no longer useful.

"You can take a phone number and break it down. When you resynthesize it, you're generating a completely random number that doesn't exist," Veeramachaneni said. "But you can make sure it still has the properties you need, such as exactly 10 digits or even a specific area code."

Synthetic data: "no significant difference" from the real thing

A decade ago, Veeramachaneni and his research team were working with large amounts of student data from an online educational platform. The data was stored on a single machine and had to be encrypted. This was important for security and regulatory reasons, but it slowed things down.

At first, Veeramachaneni's research team tried to create a fake data set. But because the fake data was randomly generated, it did not have the same statistical properties as the real data.

That's when the team began developing the [Synthetic Data Vault](#), an open-source software tool for creating and using synthetic data sets. It was built using real data to train a generative machine learning model, which then generated samples that had the same properties as the real data, without containing the specific information.

To begin, researchers created synthetic data sets for five publicly available data sets. They then invited freelance data

60%

Gartner has estimated that 60% of the data used in AI and analytics projects will be synthetically generated by 2024.

scientists to develop predictive models on both the synthetic and the real data sets and to compare the results.

In a 2016 [paper](#), Veeramachaneni and co-authors Neha Patki and Roy Wedge, also from MIT, demonstrated that there was “[no significant difference](#)” between predictive models generated on synthetic data and real data.

“We were starting to realize that we can do a significant amount of software development with synthetic data,” Veeramachaneni said. Between his work at MIT and his role with PatternEx, an AI cybersecurity startup, “I started getting more and more evidence every day that there was a need for synthetic data,” he said.

Use cases have included offshore software development, medical research, and performance testing, which can require data sets significantly larger than most organizations have on hand.

The Synthetic Data Vault is freely available on [GitHub](#), and the latest of its 40 releases was issued in December 2022. The software, now part of [DataCebo](#), has been downloaded more than a million times, Veeramachaneni said, and is used by financial institutions and insurance companies, among others.

It’s also possible for an organization to build its own synthetic data sets. Generally speaking, it requires an existing data set, a machine learning model, and the expertise needed to train a model and evaluate its output.

A step above de-identification

Software developers and data scientists often work with data sets that have been “de-identified,” meaning that personal information, such as a credit card number, birth date, bank account number, or health plan number, has been removed to protect individuals’ privacy. This is required for publicly available data, and it’s a cornerstone of health care and life science research.

But it’s not foolproof. A list of credit card transactions might not display an account number, Veeramachaneni said, but the date, location, and amount might be enough to trace the transaction back to the night you met a friend for dinner. On a broader scale, even [health records de-identified against 40 different variables](#) can be re-identified if, for example, someone takes a specific medication to treat a rare disease.

A synthetic data set doesn't suffer these shortcomings. It preserves the correlations among data variables — the rare disease and the medication — without linking the data to the individual with that diagnosis or prescription. “You can model and sample the properties in the original data without having a problem of data leakage,” Veeramachaneni said.

This means that synthetic data can be shared much more easily than real data. Industry best practices in health care and finance suggest that data should be encrypted at rest, in use, and in transit. Even if this isn't explicitly required in federal regulations, it's implied by the steep penalties assessed for the failure to protect personal information in the event of a data breach.

In the past, that's been enough to stop companies from sharing data with software developers, or even sharing it within an organization. The intention is to keep data in (purportedly) safe hands, but the effect is that it hinders innovation, as data isn't readily available for building a software prototype or identifying potential growth opportunities.

“There are a lot of issues around data management and access,” Veeramachaneni said. It gets even thornier when development, testing, and debugging teams have been offshored. “You have to increase productivity, but you don't want to put people in a situation where they have to make judgment calls about whether or not they should use the data set,” he said.

Synthetic data eliminates the need to move real data sets from one development team to another. It also lets individuals store data locally instead of logging into a central server, so developers can work at the pace they're used to.

An additional benefit, Veeramachaneni said, is the ability to address bias in data sets as well as the models that analyze them. Since synthetic data sets aren't limited to the original sample size, it's possible to create a new data set and refine a machine learning model before using the data for development or analysis.

Access to data means access to opportunities

The ability to freely share and work with synthetic data might be its greatest benefit: It's broadly available and ready to be used.

For Veeramachaneni, accessing synthetic data is like accessing computing power. He recalled going to the computer lab at night in graduate school about

20 years ago to run data simulations on 30 computers at the same time. Today, students can do this work on their laptops, thanks to the availability of high-speed internet and cloud computing resources.

Data today is treated like the computer lab of yesteryear: Access is restricted — and so are opportunities for college students, professional developers, and data scientists, to test new ideas. With far fewer necessary limitations on who can use it, synthetic data can provide these opportunities, Veeramachaneni said.

“If I hadn’t had access to data sets the way I had in the last 10 years, I wouldn’t have a career,” he said. Synthetic data can remove the speed bumps and bottlenecks that are slowing down data work, Veeramachaneni said, and it can enhance both individual careers and overall efficiency.

A 3D dog from a single photograph

Synthetic data can be more than rows in a database — it can also be art. Earlier this year, social media was enamored with [DALL-E](#), the AI and natural language processing system that creates new, realistic images from a written description. Many people appreciated the possibility for whimsical art: [NPR put DALL-E to work](#) depicting a dinosaur listening to the radio and legal affairs correspondent Nina Totenberg dunking a basketball in space.

This technology has been years in the making as well. Around the same time that Veeramachaneni was building the Synthetic Data Vault, [Ali Jahanian](#) was applying his background in visual arts to AI at [MIT’s Computer Science and Artificial Intelligence Laboratory](#). AI imaging was no stranger to synthetic data. The 3D flight simulator is a prime example, creating a realistic experience of, say, landing an airplane on an aircraft carrier.

These programs require someone to input parameters first. “There’s a lot of time and effort in creating the model to get the right scene, the right lighting, and so on,” said Jahanian, now a research scientist at Amazon. In other words, someone needed to take the time to describe the aircraft carrier, the ocean, the weather, and so on in data points that a computer could understand.

As Veeramachaneni did with his own data set, Jahanian focused on developing AI models that could generate graphical outputs based on observations of and patterns in real-world data, without the need for manual data entry.

The next step was developing an AI model that could transform a static image. Given a single 2D picture of a dog, the model can let you view the dog from different angles, or with a different color fur.

“The photo is one moment in time, but the synthetic data could be different views of the same object,” Jahanian said. “You can exhibit capabilities that you don’t have in real data.”

And you can do it at no cost: Both the Synthetic Data Vault and DALL-E are free. Microsoft (which is backing the DALL-E project financially) has said that users are creating more than 2 million images per day.

There are concerns about data privacy, ownership, and misinformation. An oil painting of a Tyrannosaurus rex listening to a tabletop radio is one thing; a computer-generated image of protestors on the steps of the U.S. Capitol is another.

Jahanian said these concerns are valid but should be considered in the larger context of what the technology makes possible. One example is medicine: A visualization of a diseased heart would be a lot more impactful than a lengthy clinical note describing it.

“We need to embrace what these models provide to us, rather than being skeptical of them,” Jahanian said. “As people see how they work, they’ll start to influence how they are shaped and trained and used, and we can make them more accessible and more useful for society.” 🏛️

THINKING ⁺⁺⁺ FORWARD

Want more Ideas Made to Matter?

Sign up here for our
Thinking Forward newsletter.

mitsloan.mit.edu/thinking-forward

Insights from MIT experts, delivered every Tuesday morning.

MIT Sloan Office of Communications
Building E90, 9th Floor
1 Main Street, Cambridge, MA 02142

Questions and Comments: thinkingforward@mit.edu
mitsloan.mit.edu ©2023 MIT Sloan School of Management

