DOMINO

INTRODUCING DOMINO DATA LAB'S

Data Science Maturity Model

Contents

Executive Summary

Many organizations have been underwhelmed by the return on their data science investment. The Data Science Maturity Model assesses how reliably and sustainably a data science team can deliver value for their organization.

Many organizations have been underwhelmed by the return on their investment in data science. This is due to a narrow focus on tools, rather than a broader consideration of how data science teams work and how they fit within the larger organization. To help data science practitioners and leaders identify their existing gaps and direct future investment, we developed a framework called the Data Science Maturity Model (DSMM).

The DSMM assesses how reliably and sustainably a data science team can deliver value for their organization. The model consists of four levels of maturity and is split along five dimensions that apply to all analytical organizations. By design, the model is not specific to any given industry — it applies as much to an insurance company as it does to a manufacturer. The matrix on the next page shows how we map the DSMM.

Level	Structured Processes	Discoverability & Compounding	Analytical Speed & Agility	Breadth & Depth of Impact	Organizational Cohesion
1 Ad Hoc Exploration	Practitioners operate autonomously in a black box	Assets stored locally, emailed around	Limited talent and tools	Ivory tower, no tangible value	Analytics island, purely transactional
2 Repeating, but Limited	Recurring workflows discussed, no enforcement	Assets stored centrally, but lack metadata / permissions	Some tools and talent investment	Static reports in a few business areas	Some collaboration with line managers
3 Defined and Controlled	Formalized process, manually enforced	Assets stored and tagged centrally with metadata and permissions	Ideas rapidly tested with novel methods / tools	Results translated into multiple operational workflows	Analytics are key stakeholders in strategic decisions
4 Optimized and Automated	Best practices codified into infrastructure, transparency for all	All asset versions stored / tagged, searchable, reproducible	Cutting-edge tools, comfortable at the analytical frontier	Data products drive org with robust safeguards	Analytics enmeshed in business; proactively anticipate needs

We believe this map can help executives and practitioners reflect on their own organization's capabilities. There is no turn-key solution to unlock the power of data science. Maturity is achieved over years of discipline and deliberate investment. As analytics teams mature, they can drive increasing value throughout their organization, further reinforcing data science as an essential capability of any business. Ultimately, this is a foundational framework upon which organizations can build their data science strategy.

Data Science Disillusionment

With each passing day, someone proclaims another industry toppled by the power of data scientists. Executives scramble to avoid being caught flat-footed, budget is allocated, heads are hired, and an army of start-ups and consultants materialize with promises of turn-key solutions. Though some firms quickly realize value, many still struggle to show lasting results from the wave of investment over the last decade.

Though some firms quickly realize value, many still struggle to show lasting results from the wave of investment over the last decade.

The problem is that excelling at data science is more than the tools your team uses or the number of PhD's you poach. Tooling and hiring are the easy parts. Many organizations take these steps and are paralyzed by the ambiguity of "but now what?"

The good news is that you're not alone. We've spoken with dozens of companies on this same journey. From those conversations, we developed a framework to help ourselves and our customers figure out where they are and what they should do next.

For this whitepaper:

We use the term "data science" broadly to encompass quantitative research, advanced analytics, predictive modeling, machine learning, etc.

Defining Data Science

Data science focuses on predicting something, prescribing something, or in some cases explaining something, making it distinct from business intelligence (BI), which focuses on backward-looking factual reporting (describing something that happened). "Data science" is also distinct from big data storage and processing technologies like Apache Hadoop and Apache Spark. These tools are valuable inputs into the quantitative research process but are insufficient to realize the full potential of data science. Successful organizations coordinate all three areas (data science, BI, and big data) to achieve maximum value.

We present this framing to reduce ambiguity of terminology and make the remainder of the paper easier to follow. We realize that definitions of these terms are debatable, and we plan to write more on this in the future.

The Data Science Maturity Model

Maturity models are based on the concept of levels or stages that companies traverse as they mature. Defining levels is a way of turning a spectrum of organizational development into discrete and understandable categories.

The DSMM aims to measure how reliably and sustainably an organization can produce the required outcomes from data science.

The Data Science Maturity Model (DSMM) is our attempt to map the evolution of advanced analytics within an organization. It is an adaptation of the Capability Maturity Model (CMM). As early as the 1970s, the CMM was a way for IT organizations to judge how advanced they were and build a roadmap for improvement.

The DSMM aims to measure how reliably and sustainably an organization can produce the required outcomes from data science. The concepts of reliability and sustainability are important as immature organizations can occasionally have big wins and mature organizations can sometimes fail.

The key is being able to deliver results over time (sustainable) at a consistently high bar of excellence (reliable) (Figure 1.).

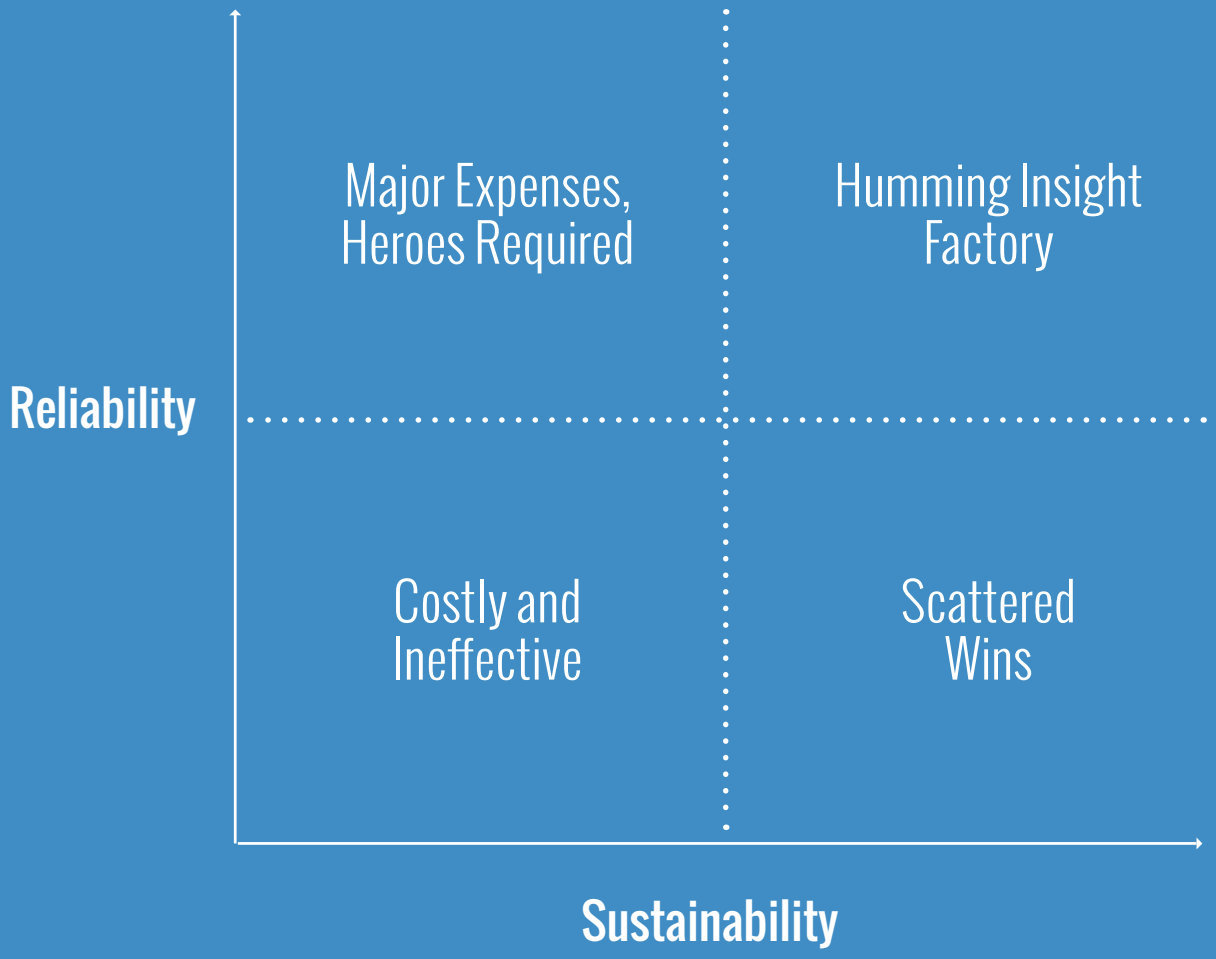

Figure 1.

For practitioners, the DSMM should help identify gaps in your workflows and opportunities for improvement. For executives, the DSMM should arm you with the right questions and give you a holistic framework to guide your investments in people, process, and technology.

Mapping the DSMM

The DSMM consists of levels of maturity and dimensions. Organizations evolve to higher levels of maturity along each of these dimensions.

The four levels of maturity are:

1. Ad Hoc Exploration
2. Repeating, but Limited
3. Defined and Controlled
4. Optimized and Automated

The key dimensions are:

- Structured Processes
- Discoverability and Compounding
- Analytical Speed and Agility
- Breadth and Depth of Impact
- Organizational Cohesion

Success in one dimension reinforces other dimensions, typically leading to consistent advancement to higher levels of maturity. Though unlike some maturity models, it is possible for an organization to be a Level 2 in some dimensions, and a Level 3 in others.

The DSMM is also deliberately industry agnostic. The framework applies as much to a life sciences company as to an asset manager. We have observed many common pain points across industries that we think the DSMM helps to address. The matrix below describes the dimensions of the DSMM at each level of maturity.

Level	Structured Processes	Discoverability & Compounding	Analytical Speed & Agility	Breadth & Depth of Impact	Organizational Cohesion
1 Ad Hoc Exploration	Practitioners operate autonomously in a black box	Assets stored locally, emailed around	Limited talent and tools	Ivory tower, no tangible value	Analytics island, purely transactional
2 Repeating, but Limited	Recurring workflows discussed, but no enforcement	Assets stored centrally, but lack metadata / permissions	Some tools and talent investment	Static reports in a few business areas	Some collaboration with line managers
3 Defined and Controlled	Formalized process, manually enforced	Assets stored and tagged centrally with metadata and permissions	Ideas rapidly tested with novel methods / tools	Results translated into multiple operational workflows	Analytics are key stakeholders in strategic decisions
4 Optimized and Automated	Best practices codified into infrastructure, transparency for all	All asset versions stored / tagged, searchable, reproducible	Cutting-edge tools, comfortable at the analytical frontier	Data products drive org with robust safeguards	Analytics enmeshed in business and proactively anticipates needs

Over the next few pages we will describe these dimensions in greater detail.

Structured Processes

When a question arises, does your team recognize it as part of a pattern and know "This is another one of those, and so we should do x, y, and z"? At immature organizations, analytics is essentially starting from scratch every morning.

As teams mature, experience builds heuristics, and discipline builds documentation.

Over time, experience builds heuristics, and discipline builds documentation. The most mature teams remove the human dependency altogether and develop workflows into actual infrastructure. Best practices thus happen by default but don't constrain creativity.

Documented best practices also facilitate transparency for the rest of the organization, helping demystify the data science process. This transparency makes outcomes more reliable.

Examples of structured processes include:

- When I get any request, I first check this library for existing work.
- When I select data, I must note the assumptions and limitations of my sample.
- Model validation requires three sign-offs: peer, manager, and business stakeholder.

Mature organizations store and share the full experimental process, ensuring context is preserved.

- Datasets including certain demographic variables need compliance sign-off.
- Models have a pre-defined shelf life and variation tolerance which triggers reviews or re-development.

Discoverability and Compounding

Analytics is inherently collaborative, with one project's output becoming the input of a new project, which informs another project, and so on. This upwards trajectory is only possible if the artifacts created by your team are easy to find, experiment with, and then store again.

Immature organizations work locally, leading to siloing and redundant work. Mature enterprise data science organizations store and share the full experimental process, ensuring context is preserved.

Analytical Speed and Agility

To butcher Tolstoy, "Effective data scientists are all alike; every ineffective data scientist is ineffective in their own way." Data scientists properly equipped with tools and knowledge can rapidly test, validate, and discard ideas.

The most mature enterprise data science teams can operate at the frontier of data science, developing and testing novel techniques and tools.

A data scientist can be ineffective for any number of reasons: underpowered hardware, outdated software packages, limited experience with the tools, etc. Immature teams try, with limited success, to make their data scientists navigate the full data stack: Talking to vendors, managing development environments, data wrangling or writing ETL, building models, and then constructing compelling data narratives.

Mature organizations recognize the limits of a one-size-fits-all approach. They assemble teams with a range of backgrounds, outfit them with the best tools, and provide them access to further learning. The most mature enterprise data science teams can operate at the frontier of data science, developing and testing novel techniques and tools. It's worth noting that many organizations think of speed and agility as the only relevant dimension when assessing their capabilities. We argue it's necessary, but not sufficient.

Breadth and Depth of Impact

Data science is only as valuable as its impact on the organization as a whole. Immature data science teams are often stuck in ivory towers while the rest of business continues as usual. Then they typically find traction with a particular group (e.g. Fraud). Over time, mature organizations expand their impact to all areas of the business, both internal (e.g. HR, Compliance) and external (Sales, Marketing). Moreover, they deepen their impact, moving from sending static reports to building dashboards to sophisticated data applications for end-users and predictive APIs.

Organizational Cohesion

What would happen if your data science team disappeared tomorrow? Would anyone notice? A data science team should engage with the rest of the organization in an evolving way: First, transactionally, then collaboratively, and eventually anticipatively.

Immature teams lack business context and respond as if they're servicing help desk tickets. Immature organizations struggle to articulate the value that data science should provide. Executives and managers follow a "hire and hope" strategy.

Mature enterprise data science teams are integral to the organization, incorporating feedback from the business while paving the way forward. The C-suite talks less about data science as an initiative because it has become part of the company's core capabilities.

This dimension is the hardest to measure but can make or break the overall reliability and sustainability of impact on the organization.

Hypothetical Examples

Below we walk through a hypothetical description of an organization at each level of maturity. This should make the types of interactions and issues more tangible for readers who are unsure where they fall. We focus on a few different industries, but mature and immature data science organizations can be found across all industries, including within non-profits and government. Note these are not based on real organizations.

Level 1
Ad Hoc Exploration

Company W has a few data scientists who report to the CTO of a small enterprise SaaS company. The team has no clear mandate other than to "create value." When a business user asks the team to help her predict churn, they each debate who will do the analysis, and then start from scratch.

Each person works in their own way, using their own local machine and preferred tools. One person uses GitHub, but the others just email their code and datasets. File names like customer_funnel_data_clean_vSally2.csv are not uncommon. Though they will run logistic regressions and random forests, most of the time business users just want dashboards on what happened last week.

Someone quits after complaining that they didn't sign up to hack through old Excel files all day, and their knowledge is lost forever. If you ask someone on the marketing team what the data science team does for them, they may reply, "They gave me a prediction for April's SEO conversion rates once. It was kind of interesting, but I was focused on Facebook ads."

Start to define the key questions that analysis can help answer to bring value to the business.

Things to think about:

- Start to define the key questions that analysis can help answer to bring value to the business.
- Discuss best practices for data cleansing, preferred algorithms, and validation amongst the team.
- Talk to stakeholders about what goes into research projects (e.g. data and other resources), what the results could be mean, and how they can use results to improve the business.
- Show value quickly: Data science can be expensive, and people want to know the investment is worth it.

Level 2
Repeating, but Limited

Company X is a medium-sized lender focused on consumers and small businesses. The team of quantitative researchers and data scientists report to the Chief Risk Officer. They focus on building probability-of-default models and have an informal process of code review where the junior modelers present their work at a weekly deep-dive.

The longest-tenured team member often stops people before they start projects saying, “Someone worked on something like that last year… I’ll see if I can dig up the PPT.”

The longest-tenured team member often stops people before they start projects saying, “Someone worked on something like that last year… I’ll see if I can dig up the PPT.” The IT team gave them a powerful AWS machine to run their analyses, but sometimes it lacks the right packages / software, rendering it useless.

The underwriting team relies on the data scientists’ models to complement their discretionary process. Every week the executive team gets a PDF report on the expected default rate for the lender’s portfolio. The sales team once asked if they could help them with lead scoring, but the analytics team didn’t know where to start. If the team disappeared, the underwriting process would be at risk, but the business would endure.

Things to think about:

- Document your process in a shared location, like a wiki, so that it's easily consulted by team members.
- Think about the types of metadata that are helpful to capture. In particular, documentation and commentary about how decisions were made.
- Share results with the whole org so they can see the types of problems the team is capable of solving and start inquiring about new applications.
- Track how data science results are incorporated into the business. Is data science doing more BI-type work? Make sure the information you are providing is used effectively and that you can show the value your work brings.

Level 3
Defined and Controlled

Company Y is a large property and casualty insurance company operating across the US. They have dozens of data scientists reporting into the Chief Analytics Officer.

Each step of the analytical process is documented on the intranet as a set of best practices that each new member of the team must learn.

Each step of the analytical process is documented on the intranet as a set of best practices that each new member of the team must learn.

There's a formal control process where a second team member "double-does" the work to ensure consistent results. Once done, they save their source data, code, and the executive report to a designated folder tagged with their names and the date.

The team has a wide range of backgrounds including statistical experts and computer science-types working on a dedicated private network with powerful computing resources. They stay abreast of industry developments and occasionally try new techniques like deep learning to augment their traditional tools.

If the data science team were to disappear, it would have a major impact on many parts of the business.

The group functions as a "shared service" and has recurring meetings with Marketing, Risk, Finance, and Operations. Once results are finalized, they work closely with the software engineers to translate their models from R into the production systems. Shelf lives are predetermined, and models get reviewed depending on compliance requirements. If the data science team were to disappear, it would have a major impact on many parts of the business.

Things to think about:

- How can you remove the risk that people don't follow the standard process, particularly in heavily regulated industries? Manual enforcement can only take you so far.

- Data science is inherently exploratory, with many variations of features and models being discarded before settling on a final result. How can you preserve and expose this iterative process, so that it's highly transparent to team members and consumers?

- Are there untapped data sources that could be used to deliver better insights?
- Look at causes of variation. When there are surprises in outcomes, could they have been prevented?
- Look for other areas where analytics can help. Are you supporting all the groups that could benefit from data science?

Level 4
Optimized and Automated

Company Z is an e-commerce firm with data scientists distributed throughout the organization. They have ingrained their ways of working into the infrastructure of the team. Before anyone starts a project, their system searches the metadata of all previous projects and flags the most relevant prior work. There is a single repository of all analytical work done throughout the organization.

Results can't be published back to the library unless necessary metadata is populated. When a data scientist wants to explore a recommendation engine for the

checkout process, she not only sees the prior modelers' efforts, but a full record of their exploratory process is available, including discarded model variants.

These practices support rapid iteration and faster compounding of knowledge.

The team easily attracts top talent based on their contributions to industry open source projects and unfettered access to powerful tools. These practices support rapid iteration and faster compounding of knowledge. Results pour into the central knowledge hub from all departments, from fraud detection to facilities.

Dozens of "helper apps" are available for internal users to enhance their efficiency, such as a defect forecasting tool for the QA team. These update on a weekly basis and easily plug into other operational systems through a scalable system of APIs. If the team were to disappear, the business would endure for a while thanks to their robust infrastructure.

Educated business users identify more opportunities for data science, bring valuable context to the analytical process, and act as advocates for the data science team.

Things to think about:

- The impact of staying on the leading edge of data science. The best data science teams have a process to identify and evaluate potential innovations quickly while minimizing the cost of exploring dead ends.

- How to educate and improve process beyond the data science organization. It is important that data scientists are working to help improve the entire analytical lifecycle from data discovery to business action. Educated business users identify more opportunities for data science, bring valuable context to the analytical process, and act as advocates for the data science team. True data democratization requires deep, ongoing education for the organization as a whole.

Conclusion

The DSMM is a conceptual map to identify where an organization's gaps and opportunities lie. There are many ways to conceptualize the world of data science. If you have feedback or suggestions on how we can improve the framework, do not hesitate to contact us at dsmm@dominodatalab.com. We hope to iterate on this first draft and incorporate insight from the data science community.

In the future, we believe the competitive edge in data science will derive from process and structure, more so than tools alone.

Finally, it is important to remember there is no turn-key solution to becoming a mature data science organization. Anyone pitching such a solution is naive or misleading. The teams we've observed operating at the highest level had visionary champions and took years of disciplined investment across people, process, and technology. Moreover, as the field evolves, it's not enough to rest on one's laurels. Technology has reduced costs and barriers to entry in data science, eliminating a moat relied on by visionary firms. In the future, we believe the competitive edge in data science will derive from process and structure, more so than tools alone. Practitioners and executives alike should consider if their current strategy is building towards a mature organization.

Domino Data Lab

Domino is the enterprise data science management platform trusted by over 20% of the Fortune 100. Domino's data science platform enables data scientists to develop better medicines, grow more productive crops, build better cars, or simply recommend the best song to play next.

Data scientists are being called upon to solve ever more complex problems across every facet of business and civil life. Domino allows them to develop and deploy ideas faster with collaborative, reusable, reproducible analysis.

Further Reading & Resources

Below are a few links to resources that we've found particularly helpful in shaping our thinking on this topic, as well as related resources we've developed.

The Data Science Lifecycle Assessment

The Practical Guide to Managing Data Science at Scale

AirBnB on Scaling Data Science

UC Berkeley Understanding Science

Data Science for Business Book

Harvard Business Review - Big Data: The Management Revolution