

2020: AI Trends for the New Decade



What's Next for the Data-Driven Enterprise



A DATAIKU WHITE PAPER

www.dataiku.com

2019 YEAR IN REVIEW

by Florian Douetteau



Progress in AI is an interesting paradox; on the one hand, there's advancement on the research side in machine learning and deep learning, which is moving incredibly quickly. Every day, there are breakthroughs (e.g., the quick dominance of DeepMind's StarCraft II AI), new research (like on [curiosity-driven learning](#)), and new technologies, making computation cheaper and faster than ever.

On the other hand, there is Enterprise AI, which is undoubtedly moving forward, though perhaps not at the pace the media coverage on AI makes it seem. While the use cases for AI in the enterprise are very promising (from manual process automation to more complex AI systems powering hyper-efficient supply chains or becoming data-powered advisors to humans working in financial institutions), progress is slower than the AI [fantasy the media sells](#).

However, I'd argue that it's unreasonable to expect developments in Enterprise AI move at the pace of AI research - after all, organizations are complex. There are engrained existing processes, revenue targets to hit, and - perhaps most importantly - there are people to consider.

A [study](#) of 500 US-based CEOs and business leaders, conducted by EY, revealed that 33 percent of respondents cited employee trust as one of the greatest barriers to AI adoption, even though CEOs and business leaders completely or somewhat trust AI.

Looking back at some of the topics from the end of 2018 both in the media and in our end-of-year trend review, there are a lot of things that came to fruition (or at least, the seeds were planted).

“Businesses are shifting from having pockets of competency or even brilliance to having a systematic way to pursue machine learning and AI opportunities.”

Hilary Mason, General Manager of Machine Learning at Cloudera, via Venture Beat AI predictions for 2019

“What to expect in 2019: We'll see methods for detecting and mitigating bias and algorithms that can produced unbiased results from biased data.”

MIT Technology Review, Never mind killer robots—here are six real AI dangers to watch out for in 2019

“The biggest trend of 2019 will be companies without AI getting left behind. In 2019, we'll see a marked shift in business reality. Failing to incorporate AI and automation soon will be like sleeping your way through digital transformation.”

Forbes, Top Six AI And Automation Trends For 2019



Here at Dataiku, we have definitely seen shifts to more systematic approaches to AI, though data teams still sit largely in centers of excellence instead of being distributed throughout and weaved into the business - more on that in this piece.

Bias was indeed a big topic for 2019, and customers have started taking note of the risks as well as taking steps to mitigate them. Dataiku, too, has made progress and product updates to make this easier for customers, introducing more robust features to make explainable AI a reality.

As far as a “marked shift”... well, it’s still gradual. Each business is on its own path to Enterprise AI, some are farther behind or ahead, but it’s true that it seems most have at least gotten started in 2019. Though getting started isn’t the same as succeeding with AI, and there’s still a lot to do.

AI initiatives take time and are based in organizational change, both from the top down and from the bottom up. In many ways, investing in the right technology and hiring the right key people are the easy parts; it’s the business-wide transformation (including upskilling existing staff and incorporating data processes at all levels of the organization) that is challenging.

The bottom line is, while enjoying these proposed trends for 2020 and beyond, have a sense of urgency... but don’t get discouraged. In the grand scheme of revolutionizing the enterprise with AI, 2020 is just one year - a drop in the proverbial bucket.



AI TRENDS FOR 2020 AND BEYOND

I. RISE OF THE DATA SCIENTISTS: THE 2ND GENERATION

There have always been different types of data scientists; some are generalists that know a little something about everything; others are experts in specific verticals and the use cases particular to those verticals (like fraud detection in finance or recommendation engines in e-commerce, for example). Some come from a coding and software engineering background and do data science in the context of that role; others are machine learning engineers who bring a strong understanding of model development and deployment.

However, no matter what their speciality, all data scientists will almost certainly see a change in their roles in the coming decade in large part because of:

1. Shifts in the way organizations are approaching data science, including a focus on data democratization and the empowerment of increasingly more people across the company (so-called citizen data scientists) to work with data for business-impacting results.
2. The push to move from just a few to hundreds - or potentially thousands - of machine learning models in production.
3. Technological developments, like investment in AI systems to facilitate AutoML and leverage automation for larger swaths of the model development process.



This is not to say that data scientists are no longer valuable; on the contrary, in this landscape, they are probably more important than ever before. But in the year and decade to come, data scientists will have to redefine their role in the face of these changes, positioning themselves where they add the most value for the company.

Here are some of the top things to know about the second generation of data scientists in 2020 and beyond:

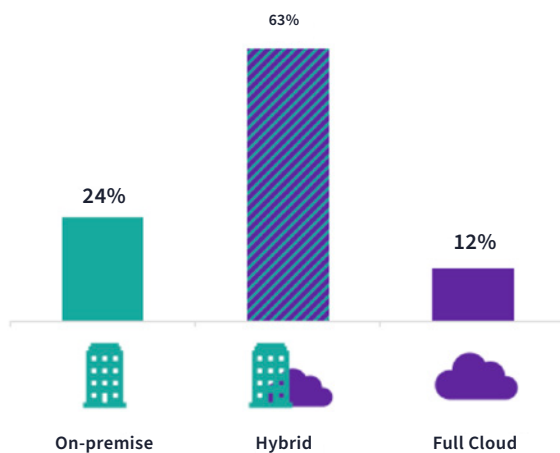
- ***Their skills will be the differentiating factor in the race to AI:*** Soon enough, AutoML will level the playing field in organizations worldwide when it comes to data democratization, and most companies will have already deployed machine learning models that address low-hanging fruit use cases. Once this comes to fruition, how will businesses make the difference with AI? The answer lies in allowing data scientists to unleash their creativity. The next generation of data scientists will be valuable because they are able to apply conventional approaches to unconventional problems - not to mention experimenting with and applying cutting-edge machine learning techniques - in ways that no citizen data scientist could do.
- ***They will probably not be generalists.*** In fact, data science generalists will become less desirable because businesses will be looking for specific skill sets that fill defined holes in their AI deployment strategy (like NLP, model monitoring, operationalization, etc.). This will likely trickle down to the education system as well, driving generalist curriculum in earlier years but finishing with specialization later on to prepare data scientists for the workforce.
- ***Part of their role will be adding structure around data democratization initiatives.*** Citizen data science has been a hot topic for years, but many companies have become disillusioned, not able to actually bring business value through citizen data science-led projects. The second generation of data scientists can address this by building processes around data democratization, incorporating ways to add expertise, approval, and operationalization to grassroots data initiatives.
- ***They will still rely on business analysts.*** Though in an ideal world data scientists are acutely aware of business needs and interests, the reality is that they probably will never have as strong of a grasp on the business as analysts. The second generation of data scientists will understand this weakness and more tightly collaborate with business analysts to increase their effectiveness.



II. MANAGING CLOUD COSTS BECOMING ESSENTIAL

There's no question that across all industries, a great migration to the cloud is underway for businesses worldwide. Dataiku conducted a survey in mid-2019 among IT professionals in large enterprises across industries and found that just less than a quarter of them are storing data used for machine learning projects exclusively on premise:

Where is the data that you use for machine learning projects?



Source: In Q3 of 2019, Dataiku enlisted Gerson Lehrman Group (GLG) to administer an anonymous survey to 200 IT professionals across a range of industries with the goal of uncovering data architecture trends in the enterprise

To be sure, cloud brings several benefits, namely rapid scalability that allows architecture to keep up with the business' data needs. And though there are fewer up-front IT costs associated with cloud platforms, they are generally more expensive overall, and especially so if resources are improperly managed. In early 2019, [The Information](https://www.theinformation.com/articles/as-aws-use-soars-companies-surprised-by-cloud-bills) reported that more and more companies find themselves surprised by their rising cloud costs.

Worldwide spending on AI systems was forecast to reach \$35.8 billion in 2019, an increase of 44% over 2018, and to more than double to \$79.2 billion in 2022.

IDC, Worldwide Artificial Intelligence Spending Guide

In order to keep cloud costs under control in 2020 and beyond, organizations will have to start developing strategies for optimizing cloud use, including considering:

Multi- vs. single-cloud approaches not only to avoid vendor lock-in, but also to optimize for different pricing structures. Adding a layer of transparency around the use of cloud resources so that teams understand - and potentially rationalize, if the task is particularly resource-heavy - their costs. Using AI to automatically monitor, scale, and control IT infrastructure (otherwise known as AIOps).

Indeed, elasticity will be the name of the game in 2020. As organizations' data teams grow and as more staff outside of data teams starts working with data, having a modern approach to architecture that allows for scaling up and down of resources is critical. Even though the cloud is growing in popularity, as the survey results featured above indicate, most companies will take a hybrid approach, investing in AI platforms that sit on top of the underlying architecture to provide a consistent user experience for working with data no matter where it is stored.

Ultimately, companies that don't actively work on developing a larger cloud strategy and managing cloud costs in 2020 will face an uphill battle to prove positive return on investment (ROI) with AI projects, racking up a bill that isn't offset by the financial gains or savings from the projects themselves.

¹ <https://www.theinformation.com/articles/as-aws-use-soars-companies-surprised-by-cloud-bills>

² https://www.idc.com/getdoc.jsp?containerId=IDC_P33198



III. SHIFTING EDUCATION AROUND DATA

Starting in 2019, it seemed that everybody was talking about data democratization - the idea that in order to succeed at implementing AI systems, organizations needed to equip people with data skills from the bottom up, no matter what their role or job title.

Unsurprisingly, many organizations found that data democratization is far from being an easy process. It requires fundamentally changing the way people work on top of introducing - in many cases - brand new skills. Even the best tools are only part of the picture when it comes to data democratization, and they can't do the job alone.

The reality is that truly shifting data culture at an organization will be a much longer-term process, and one that starts long before employees walk in the door. With the number of jobs requiring data skills - especially those outside of the role of data scientist or analyst - skyrocketing, change will have to be much wider scale.

That's why 2020 will be the start of reform in the classroom, introducing data literacy curriculum at a much earlier age. It has already started in higher education; the University of California - Berkeley notably opened an undergraduate data science major in the fall of 2019, and it immediately had nearly 1,000 students enrolled. [Dataiku is also working with major universities around the world](#) to bolster data science curriculum.

Developing a keen interest as well as skills in machine learning in children is what will truly boost society's overall data literacy skills. That means not simply producing more data scientists (though that will happen as well), but non-data professions that have the skills to use data extensively in their work.

But what will it take to get data science, machine learning, and AI education into elementary and middle schools? Development of curriculum, of course (something that surely won't happen in 2020 alone). But besides that, at a basic level, teachers will need to understand the basics of machine learning, including the larger picture of how it fits into AI.

This is one part of the equation that can start now, in 2020. Recommended reading: illustrated guides on [Machine Learning Basics](#), [Data Architecture Basics](#), or [Deep Learning Basics](#) for a good foundation.





IV. THE MOVE TOWARD INITIATIVE-DRIVEN TEAMS

In 2019, the data science center of excellence took center stage. Companies around the world and across industries created centralized teams to handle data initiatives as a way to kick-start AI efforts (for just one example, see [Nicholas Bignell](#), Director of Data Science at UBS, talk about spinning up their center of excellence).

The center of excellence model has many pros, including the ability to:

- Scale-up and get started quickly.
- Easily prioritize projects and work on those that are high-value and innovative.
- Support data democratization through the setup self-serve analytics for the rest of the company.

Yet it also has some cons - notably, that it requires quite a lot of discipline to ensure that the business is sufficiently involved in center-of-excellence initiatives. Some centers of excellence also take a consultative approach, where they receive requests from lines-of-business and work on those projects as well in collaboration with the requester.

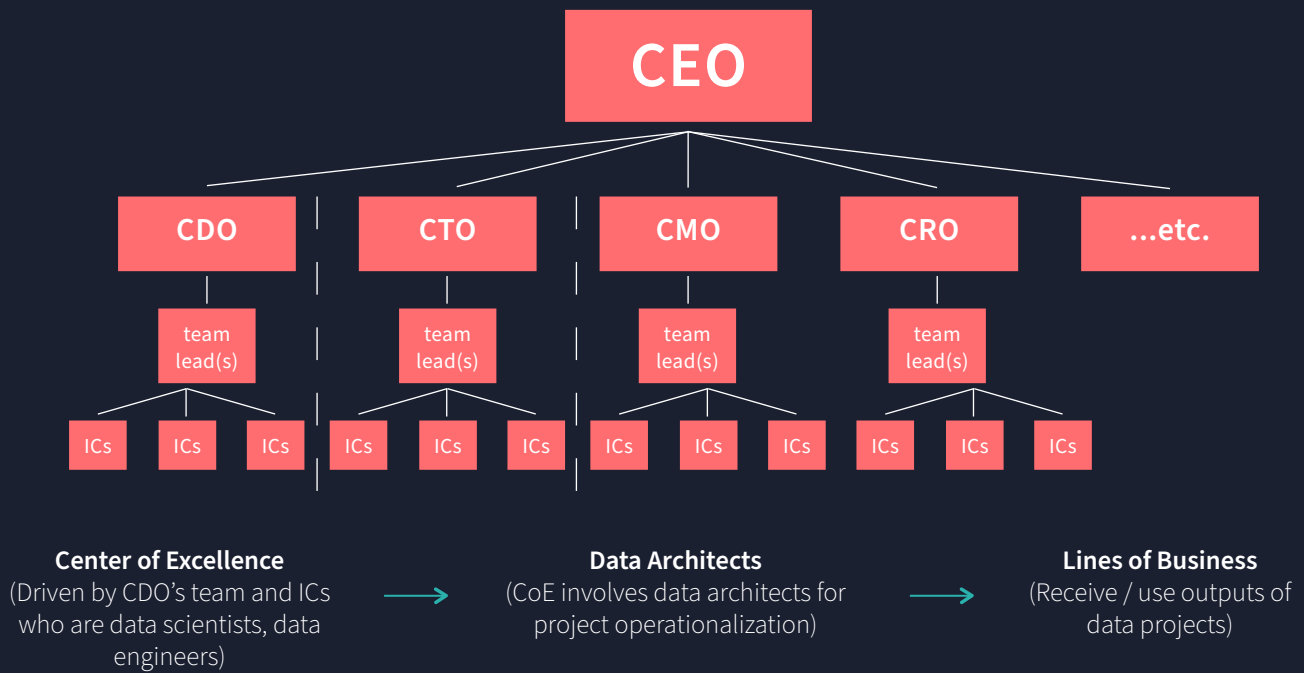
The antithesis of the center of excellence model is the embedded model, where data experts actually sit with the line of business and work on their data and AI projects - for example, there is a marketing data scientist, a risk data scientist, etc. While this model has some positives (notably the facilitation of the data experts' business knowledge), it can also stifle creativity.

The trend for 2020 will be moving toward a more formalized way for data experts and business experts to work together: the initiative-driven teams model. Along the theme of elasticity, this approach allows teams to be spun up or spun down based on the project for tight, expert focus, ensuring that the results align well with business needs and expectations.

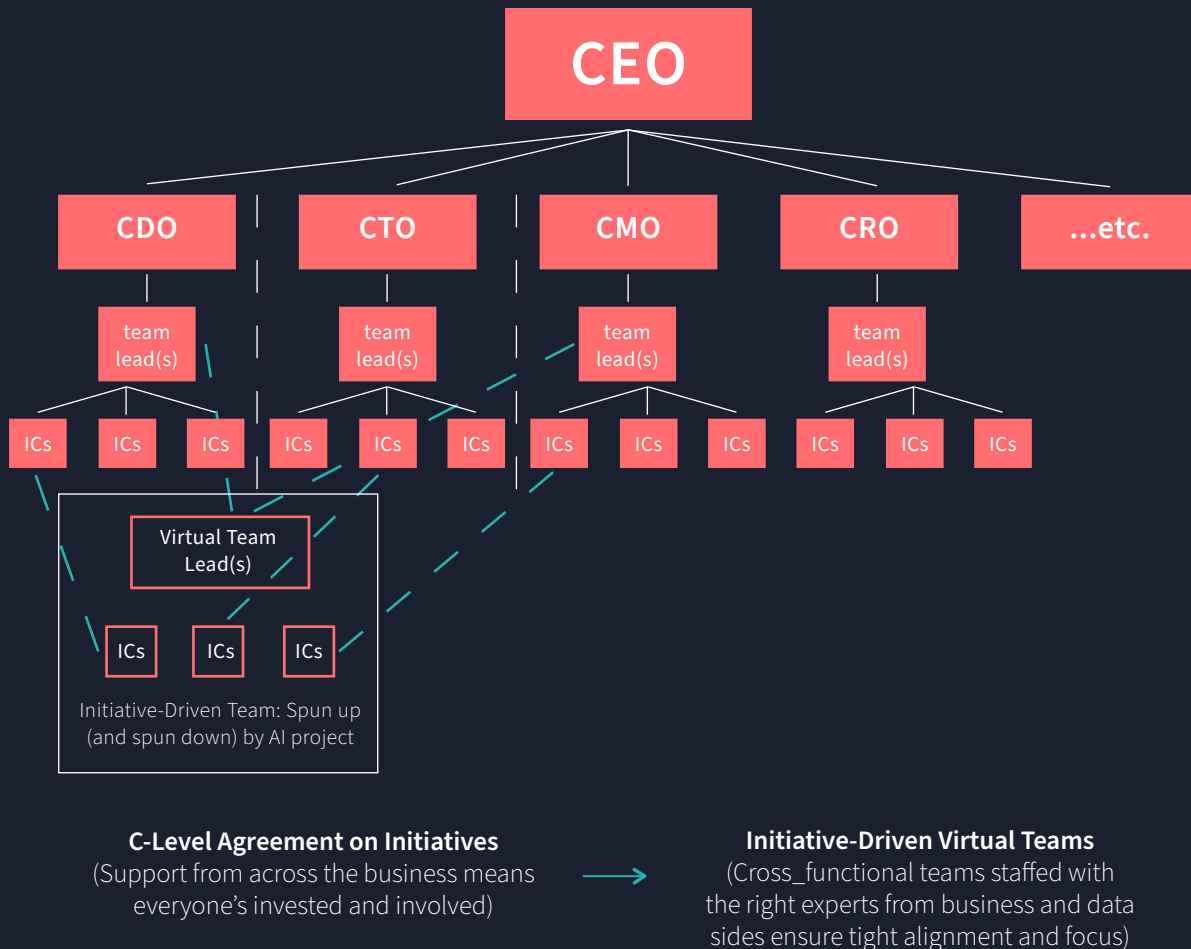
Data experts still work on a variety of projects, so the creativity problem from the embedded model is not an issue. And once the project is finished, there is a large data team that can handle maintenance and monitoring of the project instead of that falling back on the business (which is not their area of expertise and might prevent them from scaling out the number of projects they deploy).



CENTER OF EXCELLENCE MODEL



INITIATIVE-DRIVEN TEAMS MODEL





V. THE RISE OF MLOPS

Once an organization has the ability to quickly operationalize data projects and moves from a handful to hundreds (or thousands) of machine learning models in production, which is very much the case for lots of businesses going into 2020, the question of maintenance and management arise. Enter: MLOps.

MLOps help ensure that deployed models are well maintained, performing as expected, and not having any adverse effects on the business. This role is crucial in protecting the business from risks due to models that drift overtime or that are deployed but unmaintained or unmonitored. At a time when issues like responsibility and bias - two of the hottest topics of 2019 - are at the forefront, MLOps becomes even more important to close the feedback loop between operationalized models and their impact.

The year 2020 will almost certainly be one where MLOps processes are formalized and staffed as well as bolstered up with tools that make the job smoother and more manageable. Key features in MLOps technology include:

- Easy monitoring, including the ability to track and visualize drift over time for all models across the organization in one central location.
- The ability to implement automatic data validation policies and other governance features to ensure models in production aren't presenting organizational risk.
- Providing a common environment between those building models (data scientists) and those responsible for monitoring (sometimes data scientists, but also potentially operations, IT, or the business itself) so that they're all working in the same tool.

Dataiku, for example, already provides these features so that organizations can start getting organized around their MLOps strategies in 2020, with much more to come.

“Technology innovation leaders are keen to apply DevOps principles for AI and ML projects, but they often struggle with architecting a solution for automating end-to-end ML pipelines across data preparation, model building, deployment and production due to lack of process and tooling know-how.”

Gartner, Accelerate Your Machine Learning and Artificial Intelligence Journey Using These DevOps Best Practices, 12 November 2019, Arun Chandrasekaran and Farhan Choudhary



VI. CONTINUED FOCUS ON EXPLAINABLE AI, TRUST, AND BIAS

Explainable AI, trust, and bias were topics that took center stage in 2019 (throughout the six editions of EGG, [The Human-Centered AI Conference](#), around the world in 2019, there were more than 20 talks on the subject). But that by no measure means it's resolved; expect pressure around these topics and for businesses to respond through 2020 and beyond.

What's new for 2020 is that there is pressure to make AI explainable, unbiased, and trusted both:

- Internally, for those designing AI systems as well as other employees who depend on them for their jobs.
- Externally, from customers and end-users of products and services that are affected by AI systems.

Ultimately, building internal trust will provide the foundation for external trust; this starts with trust in the data itself that is being used in AI systems. Data quality is one of the most basic but most important hurdles to overcome in the path to building sustainable AI that will bring business value, not risk.

“A myth plaguing early AI exploration is that you need massive amounts of data to build successful AI models. In fact, many successful use cases can be achieved using a reasonable amount of data, as long as that dataset is of quality (i.e., normalized, complete, diversified). The lack of volume can always be compensated for through a reduction in project scope, but a lack of data quality invariably leads to POC failure. Probabilistic reasoning techniques such as ML rely heavily on data to deliver insights; therefore, this is where data quality problems are most acute — throughout the ML life cycle.”

Gartner, 5 Steps to Practically Implement AI Techniques, Erick Brethenoux and Frances Karamouzis, 25 April 2019





VI. THE NEXT-GENERATION DATA LAKE

Data lakes are traditionally associated with Hadoop, but in 2020 with the move to more elastic data architecture - including hybrid approaches combining on-premise and in the cloud, as discussed in section II - this could change.

Imagine the next-generation data lake as a service that goes beyond its classic purpose and becomes a layer of data-oriented services for the enterprise. These services would include things like data preparation, model design, and model deployment - all critical elements of the data pipeline that are currently being increasingly transformed by automation.

The next-generation data lake would leverage scalable compute through Kubernetes - continuing the theme of elasticity - as well as databases like Snowflake and BigQuery as a service. But what would it take to get there?

Notably, this layer of data-oriented services would rely heavily on metadata, which unlocks the power of data by assigning information that describes what the data is, where it comes from, what it should be used for, and more. This again circles back to the overall theme of the coming year in trust - namely trust in data at the most basic, fundamental level..

In order to move in this direction in 2020 and beyond, companies will have to put a serious focus on metadata, potentially by leveraging modern solutions like Collibra for effective data cataloguing. With strong metadata, and possibly the next-generation data lake, enterprises will be able to move forward with even more automation and efficiency gains in data processes.



CONCLUSION

In looking forward to 2020 and beyond, starting a new decade in the age of AI, the aforementioned trends have several overarching themes. If there are any areas for organizations to focus on in the years to come to succeed in the path to Enterprise AI, they are:

- 1. Elasticity.** From architecture and scaling up and down resources to teams, businesses have to be ready for whatever the future of AI holds. That means not investing in (and getting locked into) one technology, one way of doing things, etc., but being able to pivot at any turn. Whether it's a move to the cloud, embracing the next-generation data lake, new data team organizational structure, new types of data scientists, or new mindsets from employees born into the age of AI, the industry will never stop moving.
- 2. Trust.** Employees - whether in a data role or not - need to trust data from the bottom up, from its raw format to the AI systems it powers. This is an essential component to organizational change that powers data democratization. Similarly, customers or users need to be able to trust the business, and that means deploying AI that is explainable and responsible. Businesses that ignore trust (whether internally or externally) in the coming decade are sure to lose the AI race.





300+
CUSTOMERS

30,000+
ACTIVE USERS



FOX
NETWORKS GROUP



KUKA

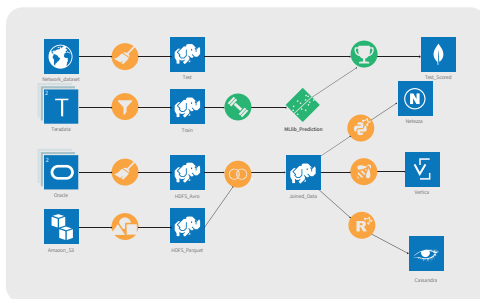


PREMERA |



SEPHORA

Name	Sex	Age
Network Log	Gender	Image
Braund, Mr. Owen Harris	male	22
Allen, Mr. T. Jones	male	38
Hicks, Mrs.	Remove this containing Mr.	26
Arnold, N.	Keep only ones containing Mr.	35
Allen, Mr. T.	Split column on Mr.	35
McCarthy	Apply Ms. to	25
Howland, N.	Remove rows equal to Moran, Mr. James	29
	Keep only ones equal to Moran, Mr. James	
	Clear (or edit) Moran, Mr. James	
☐ Filter on Mr. James		
☐ Filter on Mr.		
☐ Toggle row highlight		
☐ Show complete value		

[illegible]

The screenshot displays the QGIS desktop environment. The main map window shows a map of Mexico with green circles of varying sizes representing COVID-19 cases by state. A legend on the left indicates a color scale from white to green. A toolbar on the right contains various map navigation tools. The status bar at the bottom shows the map is in 'Normal' mode and the scale is 1:100,000,000.



WHITE PAPER

www.dataiku.com