BUILD A DISCIPLINED DATA SCIENCE TEAM

CIVIS ANALYTICS

What does it take to be an effective data science team? This is one of most important questions in the data science community today. Conversations around methods, tools, and skillsets are still valuable, but at Civis Analytics, our focus has always been on impact. Civis Analytics is a company of data scientists who have lived the challenges of being impactful, and now we build a data science platform and consult with organizations who have data science challenges. We believe that data scientists have the potential to transform organizations by consistently getting their work into the hands of decision makers — but the road to success passes through a set of attitudes and behaviors that are hallmarks of disciplined data science teams. This whitepaper will cover the priorities of a disciplined data science team and the payoffs we see from adopting those priorities.

We work with a lot of companies that already have some data science pieces in place. They have a data scientist or three, some analysts helping the data scientists, buy-in from the company leadership, some data (maybe spread across different formats and locations), some models and reports, and … that's it. They're doing data science, but it's not transforming the organization. These teams are looking to take their existing efforts and make them impactful, so their organizations get closer to the "data-driven" ideal.

The data-driven ideal looks a little different for every organization, but generally they share the common theme that the organization has a culture of turning to data for guidance in making important decisions, that the data science team is empowered with the business context to make tools and workflows that are useful, and that the technical and communication bridges between the data scientists and the rest of the organization are robust.

The data-driven organization has a culture of turning to data for guidance and empowers its data science team with the business context to make useful tools.

This whitepaper is for the data scientist who finds themself, or their team, in a spot where they're ready to be more effective. This isn't "getting started" advice: This is advice to turn a nascent and slightly chaotic set of data efforts into a more coherent and professional initiative. It's hard to make this transition — we know because we've been there.

The attitudes and behaviors described in this whitepaper separate the effective data science teams from the teams that feel like they're constantly scrambling to fulfill requests but not making the impact they want to have. These attitudes and behaviors are:

- Discoverability
- Automation
- Collaboration
- Empowerment
- Deployment

DISCOVERABILITY

DATA ANALYSIS RESULTS SHOULD BE DISCOVERABLE BY THE ORGANIZATION

Let's start with an example of a data science result that's not discoverable by an organization: Say there's a company who has built an app for ordering food online, and their data science team decides to build a customer lifetime value model predicting how much food users will order, so the company can predict resource needs.

Often this starts out as a python or R script that a data scientist builds

on their laptop, but getting the results into the hands of their company's operations staff usually involves an email to the data scientist, pointing them toward a customer list, the data scientist scoring the model on the list, and, finally, sending predicted customer values to the stakeholder. It's a communication method that tends toward repeated work, with stakeholders needing to worry about "pulling" information from the data science team, and the data science team having limited impact on the business because their work only gets adopted one person at a time.

A better scenario is one in which the predicted customer values get published to a central location for anyone to find who might need them. This location should be consistent so everyone knows where to look, which suggests some kind of web page or a shared document. We are advocates of the web option: Platforms that support app-like capabilities, such as interactive dashboards, avoid the problem of proliferating documents with names like "model_results_final_final_v2". Whatever option you choose, people across the organization should know where they are supposed to go to get results. When results are available to someone, they are automatically made available to everyone else who might be interested.

People across the organization should know where they are supposed to go to get results.

DISCOVERABILITY

Before Records of data science work exists in emails, chat logs, and in people's brains.

After Centralized location for data science results where teammates and stakeholders access what they need.

AUTOMATION

DATA ANALYSIS RESULTS SHOULD BE RELIABLY PRODUCED AND AUTOMATICALLY REFRESHED

Once the data science team has their model built, and the publication channel is set up, there is little need for the data science team to be producing results by hand anymore.

DISCIPLINED DATA SCIENCE

The workflow should be automated and scheduled so that the data scientist is mostly maintaining the pipeline and monitoring to make sure results look sensible. Automating the workflow minimizes the cognitive overhead for the data scientist, and running the workflow on a schedule allows downstream coworkers, such as salespeople, to get used to a cadence of always knowing when and where to look for up-to-date results.

This means that the pieces are chained together so each successful job kicks off the next step, and that the workflow kicks off regularly.

AUTOMATION

Before Data scientists manually re-run their workflows.

After Complex workflows are programmatically chained together and run on a schedule.

COLLABORATION

THE DATA SCIENCE TEAM SHOULD HAVE A SHARED UNDERSTANDING OF HOW TO COLLABORATE

We've taken a page from our software engineering friends here and are leaning heavily into using tools like git for collaboration and, relatedly, version control.

If you don't see the connection between collaboration and version control, just think of version control as a tool that enables you to collaborate with past/present versions of yourself.

Collaboration is important because it allows the different strengths of everyone on your team to all be brought to bear on a problem. Tools like git allow for rapid iteration as teammates work together to inspect and try out each others' solutions to problems.

Collaboration is also measurably easier when there's a shared understanding of how code should be written and iterated upon. This understanding isn't fun to nail down, but doing so is worth the effort. It makes the cognitive barrier drastically lower when

> "Data scientists have the potential to transform organizations by getting their work into the hands of decision makers."
>
> KATIE MALONE, DIRECTOR OF DATA SCIENCE

teammates look at each others' code. We suggest adding these concrete tactics to your workflow:

- Everyone lints their code using the same linting standard.
- All code gets documented, including function docstrings and comments. Building in time for more complete and comprehensive documentation—say, one week of documentation time at the end of a major project—is even better.
- All code goes into git (or other version control software) via platforms like Github, Gitlab, or Bitbucket. If it's not in git, it doesn't exist.
- New code is developed in branches or forks on git, not in the production branch. In order for code to be merged into production, it needs to undergo formal code review.
- All code should be tested. Use a continuous integration tool like Travis or Circle to run the tests automatically when you commit the code to git. It can be tricky to know how to test a model, but we find that modeling is only a small piece of most data science workflows anyway. Start small, test what you can, and try to improve incrementally and continuously.

One statistic we like to remind each other about is that code typically gets written once or twice but then read dozens of times

COLLABORATION

Before Each data scientist has their own code style, quality control standards, and versioning habits.

After Data science team has a shared set of quality standards for data science code, and uses standard software engineering tools to enforce those standards.

EMPOWERMENT

THE DATA SCIENCE TEAM SHOULD BE EMPOWERED TO USE THE RIGHT TOOLS FOR THEIR JOBS

Data scientists, and the people on their teams, tend to come from diverse backgrounds. That means that the tools they're most comfortable with can vary. Perhaps the most high-profile example of this is the evergreen python vs. R debate, but whichever your data science team wants to use, the paradox boils down to whether individuals should be allowed to use whatever tools they want, or decide as a team to do their data science together on the same technical stack.

The official Civis position: it's a false dichotomy and they're both great, get back to work.

We are advocates of the latter because it is necessary for discoverability and collaboration, and often also helps data science teams with deployment and automation. These are advantages that more than pay for themselves, although we acknowledge that it comes with costs like any bureaucracy. The trick is to get the right tech stack for the needs of the team (acknowledging that there is no such thing as the perfect stack, so your stack will evolve over time).

We have found that good data science tech stacks tend toward a few biases:

- Use well-supported open source tools whenever possible.
- Be judicious in adoption of new frameworks, languages, etc.
- Spend time being a skeptic of your own logic before making big technical decisions. Write down all the risks of each option, not just the benefits.
- Be diligent in how you store, access, and keep track of data. It's just as important as how you analyze it. Good data management is a hallmark of a great data science team.
- Be thoughtful and disciplined about the interfaces between major pieces of your stack. Be unopinionated about what folks do within their box, but strict about how boxes talk to each other. Communicate repeatedly and consistently with your team about what the expectations are and use code review as a primary means of enforcement.
- Don't be afraid of building out more complex technical architecture as your team matures, but budget time for refactors to keep it from becoming a spaghetti stack.

EMPOWERMENT

Before Ad hoc collection of data storage and analysis tools, without consistency between team members or larger strategy around the tech stack.

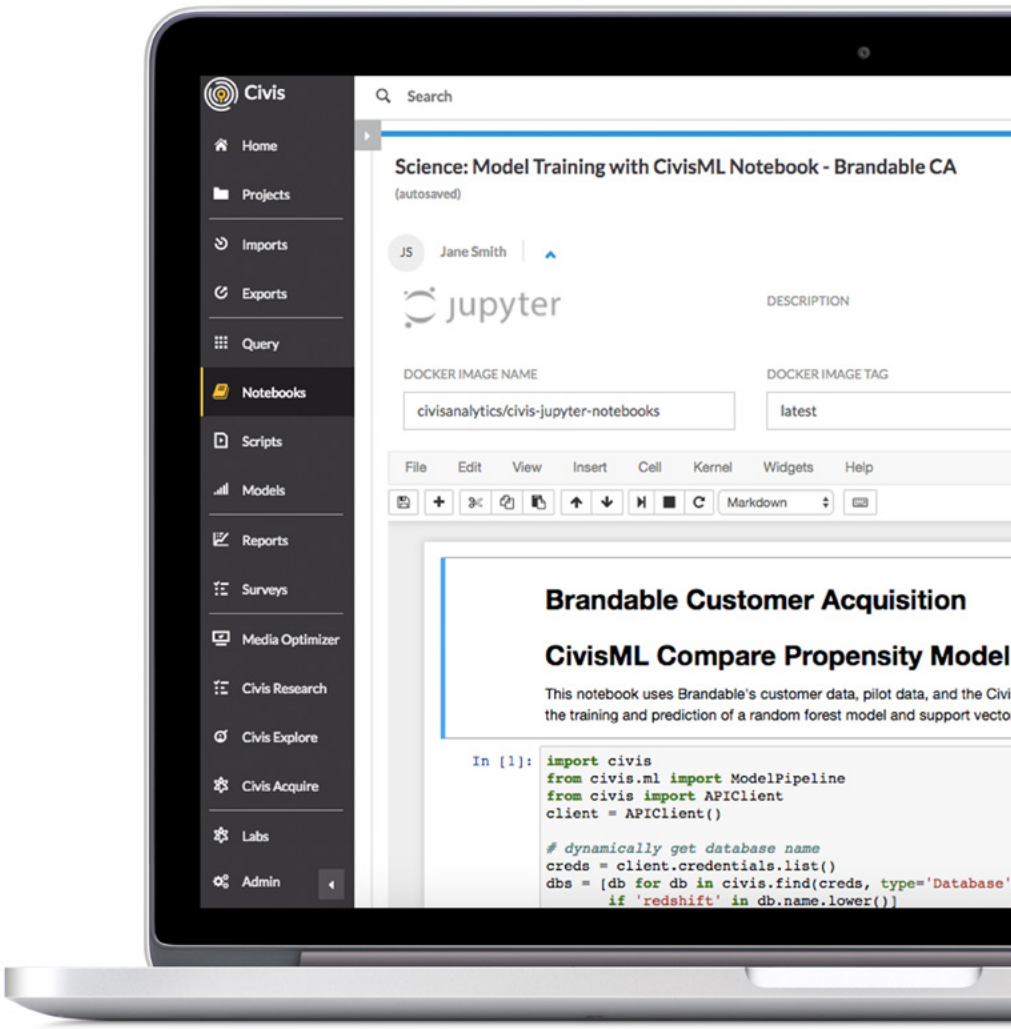
After Data science team has thought carefully about how to make their work high-impact with technical tools optimized to enable those priorities engineering tools to enforce those standards.

DEPLOYMENT

THERE SHOULD BE A DEFINITION OF A DATA SCIENCE TOOL OR WORKFLOW BEING "IN PRODUCTION," AND THE DATA SCIENCE TEAM SHOULD BE LASER-FOCUSED ON GETTING THEIR WORK INTO PRODUCTION

If your data science team is doing its job correctly, the rest of the organization should be using its output. Getting results into the hands of decision makers is what a great data science team is all about.

At the same time, the data science team needs to be able to explore and experiment, with the expectation that experiments can—and should—fail sometimes. This introduces



the idea of work that's in development (failure is fine) versus production (failure is not ok—it needs to work), and a clear line between the two. Enforcing that line helps the organization know that it's getting work that is vetted and high-quality, which builds trust.

There should be a set of standards for data science results that are in production, and they should be high. Most importantly, the highest-priority metrics for the data science team should include the amount of work in production, and how much use that work gets by people across the organization.

Simply putting tools into production isn't enough. Data scientists should measure themselves by if their tools are actually solving problems for their users.

The highest-priority metrics for the team should include the amount of work in production and how that work is used across the organization.

DEPLOYMENT

Before Data science teams measure themselves by activity and not by output—so their work doesn't get used.

After Stakeholders are using data to make decisions.

WHAT'S NEXT

We've laid out our thoughts about what distinguishes a data science team that is ramping up from a data science team that is fully grown up. But we believe data science teams can go even further. We're actively trying to define (and build!) the next level of standards for mature and effective teams.

- How can we think about versioning and entire analysis, including not just the code but the underlying data, environment, and documentation?
- How do we make data more searchable, understandable, and discoverable?

- How do we build feedback loops into software so data science teams are better able to introspect their own work and more quickly discover bugs, user sore spots, and new best practices?

We have more questions than answers at this point, but we're excited to think about how the next generation of data science tools will even further empower data scientists to transform their organizations. In the meantime, we've been building out our data science platform to emphasize the capabilities we discussed above (for example, we have upcoming features oriented toward helping data scientists put their work into production, and making that work discoverable by anyone else on the platform). If you'd like to take a look, get in touch with us to schedule a demo or get yourself a free trial of platform.

ATTITUDES AND BEHAVIORS FOR A DISCIPLINED DATA SCIENCE TEAM

	Before	*After*
DISCOVERABILITY	Records of data science work exists in emails, chat logs, and in people's brains.	Centralized location for data science results for both teammates and stakeholders
AUTOMATION	Data scientists manually re-run their workflows	Complex workflows are programmatically chained together and run on a schedule
COLLABORATION	Each data scientist has their own code style, quality control standards, and versioning habits	Data science team has a shared set of quality standards and uses standard engineering tools
EMPOWERMENT	No consistency between team members or larger strategy around the tech stack	Careful thought about how to make work data scientists' work high-impact
DEPLOYMENT	Data science teams measure themselves by activity and not by output	Stakeholders are using data to make decisions

ABOUT THE AUTHOR

As Director of Data Science at Civis Analytics, Katie Malone leads a team of data scientists who are responsible for researching and developing new methods to solve complex data problems.

Prior to joining Civis, Katie worked at the Large Hadron Collider at CERN in Geneva, Switzerland, on Higgs boson searches. She also spent a summer working at the online education startup Udacity, where she launched her podcast Linear Digressions.

Katie graduated from Ohio State with a major in Engineering Physics and received her PhD in Particle Physics from Stanford. She spends her free time learning French, running along Lake Michigan, and spoiling her rescue pup, Maeby.

KATIE MALONE

Director, Data Science
Civis Analytics

DISCIPLINED DATA SCIENCE

Learn more about Civis Analytics:
www.civisanalytics.com