Data Science 101 – a CDO's Viewpoint

Because principles outlive tactics, I offer the following:

Ready, willing & able

The number one criteria for launching a Data Science initiative is that you know the name and title of the executive sponsor who wants to use your insights so they can drive improvement. If you don't know that, you're not ready to take action. Whoever owns the organization's results has accountability for value capture in Data Science. Find them. Talk to them. Align with their agenda.

Target the sweet spot

There are always new opportunities popping up and tempting you. How do you choose the best things to work on? In any organization, the sweet spot is in the overlap of quantifiable value for the customer (for DoD that's the taxpayer), quantifiable value for the employee (for DoD that's the warfighter), quantifiable value for the stakeholders (for DoD that's the President & Congress) and quantifiable value for our own executives (see *Ready, Willing and Able*). There will be a limited number of opportunities that satisfy all requirements, and these can be prioritized / sequenced by value.

Direction > speed

Direction is more important than speed. If you're pointed in the wrong direction, it doesn't matter how fast you're traveling. Inversely, if you're locked on to your desired destination, all progress is positive, no matter how slow you're going. Choosing direction over speed is a simple concept, but in practice it can be difficult to follow (see *Target the sweet spot*).

Better is better

You may have an ultimate goal in mind for each Data Science exercise, but any improvement is better than no improvement. Better is better. Adopt a DevSecOps mindset, with appropriate techniques and tooling. Breaking work into chunks, and then build up the big picture. Produce minimum-viable-products, lots of them. The cost of a prediction is already $0. Iterate your way to the ultimate goal.

Learn by doing

When it comes to Data Science, we're all winging it. Even the experts. Especially the experts; we're dealing with cutting edge problems and there's no playbook. This is applied Research & Development. We're figuring it out as we go. If a material percentage of your Data Science projects aren't failing, you're not taking the risks you need to take in order to increase your knowledge and skills. Foster a data-driven, test-and-learn culture. Let the data speak to the situation. Communicate bad news quickly and without shame. What we learn along the way, and how widely we share that learning, is as important as the immediate outcome. Actively engage in the broader Data Science community to leverage best-practices and learning from completely different fields. Celebrate both the outcome and the learning. Try to figure out what you should do next, not what the right answer is. And remember, there's no compression algorithm for experience.

Walk toward the uncertainty

While the general direction for Data Science is clear, the exact path needed to navigate through uncertainty needs to be discovered organization by organization. No one is smart enough to see clearly through the opportunities and challenges, you have to navigate your way through. You need to work

through the organization's full value chain, use case by use case and data story by data story, to get to value from Data Science. Your executive sponsors don't need help with obvious decisions, they need help making difficult tradeoffs between nuanced alternatives. They want evidenced-based help navigating uncertainty. Change management is critical to the success of many data science initiatives. Give people clear recommendations of the next steps to take.

There is no truth in data

Don't start with the data you've collected; start with the problem you want to solve. Learn to identify when you've got 'good enough' data to proceed, and how to manage the risks of acting on it. You have less data than you think. You certainly don't have any data for things we haven't seen before. Ask yourself, "Does the historical data you've got represent the future you want to create?" It probably doesn't. The old adage "garbage-in → garbage-out" applies to Data Science. Diversity is critical – diversity of data, people, mindsets, backgrounds, models. Diversity helps protect you against hypothesis bias, which often leads you to shape data to an existing agenda.

Data > algorithms

Additional data is more useful than a better algorithm, (Pro-tip – make sure it's the right data). AI/ML algorithms are widely open-sourced, so they're not a source of relative advantage. More data, when carefully chosen, typically leads to better insights. While I'm on the topic of algorithms, they're not products, they're processes. We will never be sure what a process does until we run it. The biggest problem is always misaligned incentives; AI/ML algorithms are incentive-seeking machines.

Buoys not boundaries

You don't want to leave tool choices wide open but there's no value in publishing guidelines you cannot enforce. Forget all about policies, memos, directives, standards, reference architectures. The office building you work in is already wallpapered with those artifacts, and nobody is paying any attention to them. Don't add to the wallpaper. Lead by example instead. Rather than publishing prescriptions – defining a particular process model or a technical architecture for tooling pipeline or a data formatting standard or a policy statement – implement them and make them widely available as a utility or shared service. Speed thrills and people will use your shared service if it's instantly available rather than having to go through a procurement process. The concept is called "buoys, not boundaries" You establish a series of vetted pipelines that are very well tested and very easy to consume. If a Data Science team doesn't know what they want to do in order to create a pipeline, they can leverage the central service very easily. This has the added benefits of creating: freedom of maneuver for data talent (they can go anywhere in the organization without retraining); creating freedom of maneuver for proven data workloads (they can be reused anywhere in the organization without retooling); and bargaining leverage with tools suppliers.

On the other hand, if people do want to explore alternatives, they can go explore in their own environment. That's critical in the Data Science world, because it is changing constantly. You don't want to be a barrier to innovation. You don't want to apply your central team's scarce resources to assessing a queue of tools or techniques. By the time each request makes it to the front of the queue some alternate will already be available in the market. Give your organization the ability to explore and innovate.

Help Wanted

First, let's assuage a concern. Neither Artificial Intelligence nor Machine Learning (AI/ML) will replace people where you work. People who use AI/ML where you work with replace people who do not use AI/ML where you work. AI/ML will assist those that remain do their work better and faster. In order to realize the promise of AI/ML, you're going to need a number of quintessentially human skills. The biggest challenge in AI is knowing if you have the right data. That's a quintessentially human skill. Knowing when the 'right data' shows the wrong things is a quintessentially human skill. Finding 'good enough' data, for AI, is a quintessentially human skill. Generating synthetic data, for the future you want to create, is a quintessentially human skill. Assessing which trained AI model is the most useful and explaining why, these are quintessentially human skills. You need to attract, develop and retain people with these skills.

Pareto was right

Data Science is the easy part. Getting the right data, and getting it ready for analysis, is much more difficult. Most of the costs go to getting data ready for use, not getting value out of the data. Across all industry sectors, CDOs report spending 80% of their resources (people, time, $s, energy) wrangling the data. It's critical work, organizational hygiene, but not high value-adding work. Don't let it take over your team's focus. Leverage the machine learning that comes embedded in your tools. It'll tell you the best ways to visualize the data for improved understanding. It'll tell you how and where to improve data quality. The best time to improve data quality – whether with tagging or formatting or anything else – is at the point of integration, when the data crosses the boundary into your possession. You don't want to depend on anyone else for the quality of your data. Invest in change management. Your audience will adopt modern analytics, visualizations, and models best if you "take them on the journey" of learning with you.

Automate everything

If a task is worth doing, it's worth automating. This is especially applicable to Data Science. You don't have enough people, time or energy to do everything manually. It doesn't scale. What's worse, manual effort decreases quality. The engineering console is not your friend, YAML is. Automate tasks the first time you do them. And share the automation by posting it in the shared repository. We've established a central capability to capitalize on economies of scale for hard assets, share best practices, and improve talent retention by fostering communities of interest.

Fix the production process, not its products

There's a technical term for fixing quality defects in your products. That term is 'rework'. As any quality expert will tell you, rework means unnecessary delays and costs. The accepted best practice is when you see quality defects, to perform root cause analysis then fix the reason for the defect. This approach applies to the Data Science pipeline as well. We'll refine and perfect the pipeline, then let the pipeline refine and perfect our data products.

Michael Conlin
Chief Data Officer, US Department of Defense

michael.j.conlin10.civ@mail.mil