Boosting the data startup

[cross-post from Medium]

Some thoughts from working in data-driven startups.

Data scientists, like consultants, aren't needed all the time. You're very useful as the data starts to flow and the business questions start to gel, but before that there are other skills more needed (like coding); and afterwards there's a lull before your specialist skills are a value-add again.

There is no data fiefdom. Everyone in the organization handles data: your job is to help them do that more efficiently and effectively, supplementing with your specialist skills when needed, and getting out of the way once people have the skills, experience and knowledge to fly on their own. That knowledge should include knowing when to call in expert help, but often the call will come in the form of happening to walk past their desk at the right time.

Knowing when to get involved is a delicate dance. Right at the start, a startup is (or should be) still working out what exactly it's going to do, feeling its way around the market for the place where its skills fit; even under Lean, this can resemble a snurfball of oscillating ideas, out of which a direction will eventually emerge: it's too easy to constrain that oscillation with the assumptions and concrete representations that you need to make to do data science. You can (and should) do data science in an early startup, but it's more a delicate series of nudges and estimates than hardcore data work. As the startup starts to gel around its core ideas and creates code is a good time to join as a data scientist because you can start setting good behavior patterns for both data management and the business questions that rely on it, but beware that it isn't an easy path: you'll have a lot of rough data to clean and work with, and spend a lot of your life justifying your existence on the team (it helps a lot if you have a second role that you can fall back on at this stage). Later on, there will be plenty of data and people will know that you're needed, but you'll have lost those early chances to influence how and what data is stored, represented, moved, valued and appraised, and how that work links back (as it always should) to the startup's core business.

There are typically 5 main areas that the first data nerd in a company will affect and be affected by (h/t Carl Anderson from Warby Parker): data engineering, supporting analysts, supporting metrics, data science and data strategy. These are all big, 50-person-company areas that will grow out of initial seeds: data engineering, analyst support, metric support, data science and data governance.

- Data engineering is the business of making data available and accessible. That starts with the dev team doing their thing with the datastores, pipelines, APIs etc needed to make the core system run. It'll probably be some time before you can start the conversation about a second storage system for analysis use (because nobody wants the analysts slowing down their production system) so chances are you'll start by coding in whatever language they're using, grabbing snapshots of data to work on 'til that happens.

- To start with, there will also be a bunch of data outside the core system, that parts of the company will run on 'til the system matures and includes them; much of it will be in spreadsheets, lists and secondary systems (like CRMs). You'll need patience, charm, and a lot of applied cleaning skills to extract these from busy people and make the data in them more useful to them. Depending on the industry you're in, you may already have analysts (or people, often system designers, doing analysis) doing this work. Your job isn't to replace them at this; it's to find ways to make their jobs less burdensome, usually through a combination of applied skills (e.g. writing code snippets to find and clean dirty datapoints), training (e.g. basic coding skills, data science project design etc), algorithm advice and help with more data access and scaling.
- Every company has performance metrics. Part of your job is to help them be more than window-dressing, by helping them link back to company goals, actions and each other (understanding the data parts of lean enterprise/ lean startup helps a lot here, even if you don't call it that); it's also to help find ways to measure non-obvious metrics (mixtures, proxies etc; just because you can measure it easily doesn't make it a good metric; just because it's a good metric doesn't make it easy to measure).
- Data science is what your boss probably thought you were there to do, and one of the few things that you'll 'own' at first. If you're a data scientist, you'll do this as naturally as breathing: as you talk to people in the company, you'll start getting a sense of the questions that are important to them; as you find and clean data, you'll have questions and curiosities of your own to satisfy, often finding interesting things whilst you do. Running an experiment log will help here: for each experiment, what was the business need, the dataset, what did you do, and most importantly, what did you learn from it. That not only frames and leaves a trail for someone understanding your early decisions; it will also leave a blueprint for other people you're training in the company on how data scientists think and work (because really you want everyone in a data-driven company to have a good feel for their data, and there will be a *lot* of data). Some experiments will be for quick insights (e.g. into how and why a dataset is dirty); others will be longer and create prototypes that will eventually become either internal tools or part of the company's main systems; being able to reproduce and explain your code will help here (e.g. Jupyter notebooks FTW).
- With data comes data problems. One of these is security; others are privacy, regulatory compliance, data access and user responsibilities. These all fall generally under 'data governance'. You're going to be part of this conversation, and will definitely have opinions on things like data anonymisation, but early on, much of it will be covered by the dev team / infosec person, and you'll have to wait for the right moment to start conversations about things like potential privacy violations from combining datasets, data ecosystem control and analysis service guarantees. You'll probably do less of this part at first than you thought you would.

Depending on where in the lifecycle you've joined, you're either trying to make yourself redundant, or working out how you need to build a team to handle all the data tasks as the company grows. Making yourself redundant means giving the company the best data training,

tools, and structural and scientific head start you can; leaving enough people, process and tech behind to know that you've made a positive difference there. Building the team means starting by covering all the roles (it helps if you're either a 'unicorn' — a data scientist who call fill all the roles above — or pretty close to being one; a 'pretty white pony' perhaps) and gradually moving them either fully or partially over to other people you've either brought in or trained (or both). Neither of these things is easy; both are tremendously rewarding and will grow your skills a lot.