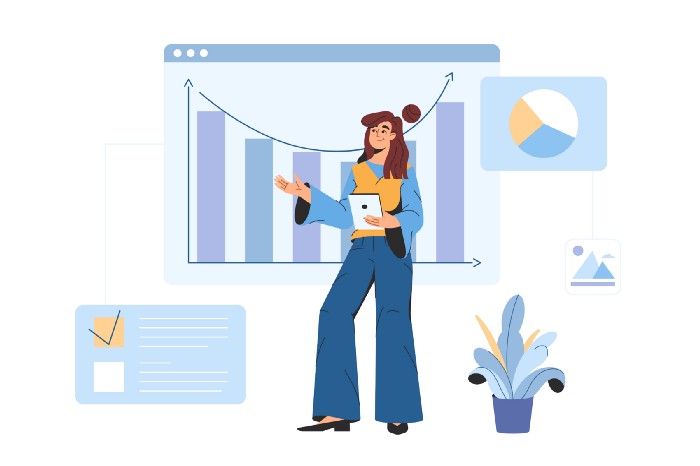
**From Vague to Value — Data Practitioner Insights**

As data professionals, we all want to work on cool data problems and to be successful in those projects. However, what often comes as a surprise is that the definition of cool and measure of success evolves as you go from school to industry. A paradigm shift happens when we go from working on data projects in a controlled environment (like school, bootcamps, etc.) to tackling data projects in the real world. With my years of experience as a data professional, being part of various conversations, and serving as a mentor for aspiring data scientists, I would like to share a few practitioner insights with this article.



**1. Data is the answer, but what is the question?**

Data professionals constantly re-evaluate not only if we are solving the **problem right** but also, more importantly, if we are solving the **right problem**. In school, someone always knows the question; the ask is clear, at least in the teacher’s mind. However, our stakeholders rarely know precisely what needs to be done. They will often come to you with either a concern or a hope and look towards you to provide data context, fill gaps, push back if required, and shape the overall problem statement. Our job is to translate vague ideas into a quantifiable problem statement that can then be translated into mathematical language.

Once you arrive at a quantifiable problem statement, it is not the end of the road either. It is quite possible that the available data does not support the kind of analysis you originally envisioned (missing values, hidden confounding variables, sparse features, sparse data points, etc.). A situation like this will further tune and refine your problem statement.

Adaptability, practicality, and being comfortable with ambiguity are the three most valuable skills in the toolkit of a data scientist.

**2. Value > Accuracy**

The goal of data science is not to find and tune the most accurate machine learning model; the goal is to provide value to an organization. The value can be defined as money, time, customer goodwill, market trust, etc. We work with the business and product stakeholders to understand the business need and quantify ***value***. More often than not, you will find that **simplicity and interpretability** are valued above the accuracy and complexity of a model, as the former is correlated with reduced risk and increased confidence in the success of the deployed model.

Whereas in school, you are encouraged to learn and try increasingly advanced techniques to optimize for accuracy, in solving real data problems, you have to find, and at times advocate, for the tradeoff between accuracy vs. cost or accuracy vs. time spent. The complexity of the technique needs to be balanced with how well the technique solves the problem and what value it provides to the business.

It is essential to recognize that machine learning is **A** solution, not **THE** solution. In many cases, less expensive techniques like multivariate statistical analyses, heuristic-based case statements, and behavioral state machines can provide the results we are looking for.

**3. More != Better**

The sparsity of data presents us with many well-known challenges. However, what is often less discussed is that the abundance of data does not necessarily make the analysis, or the life of an analyst, easier either.

Selecting the right data and the right amount of data is critical in real-world data science applications. In school projects, we usually try to get as much data and extract as much information from that data as possible as (a) that is what gets us extra credit and (b) if it is a research-based project, the more exploration the better.

However, in real-world data projects, we operate under business and product constraints, time and money being two of them, and the **focus is on efficiency** rather than completeness. Limiting the scope of exploration and narrowing down on the required datasets are valuable skills. The more data you add to the analysis, the more complex your analysis becomes. The complexity increases not linearly but exponentially for issues like data cleanliness, completeness, imputation, distributed processing of big data, code complexity, testing requirements, etc. And where the complexity increases exponentially, the value gain is often logarithmic. Recognizing and stopping at the sweet spot is crucial.

**4. Beware of the Sunk-Cost Fallacy**

The sunk-cost fallacy is the phenomenon whereby a person is likely to continue an endeavor if they have already invested in it, even when it is clear that abandonment would be more beneficial.

The problems that data professionals work with are often open-ended, and the conclusions are not always straightforward. For example, you may have to optimize for metrics inversely related to each other. Or your project may be moving in a good direction but not ensuring enough value for the stakeholder to spend another quarter on your work. It is **okay to move on** if you and other primary stakeholders identify a project as no longer feasible. Moving on does not show failure or wastefulness; rather, having the maturity to let go in everyone’s best interest is the hallmark of experts and leaders. In those instances, we critically review the entire project, take notes and learnings, have a retrospective meeting with relevant stakeholders, and move on!

Data science and analytics is a fascinating field where we are tasked to go *from vague to value*. This requires technical expertise, undoubtedly, but more than that, this requires a certain mindset that looks beyond algorithms, code, and confusion matrices. **A data professional’s mindset is one that thrives in the land of ambiguity and tradeoffs; less rigid, more fluid, and open to questioning every assumption.** We hope you will find this article helpful in building that mindset as you embark on your data science and analytics professional journey.