Niclas Isensee

5/10/2025

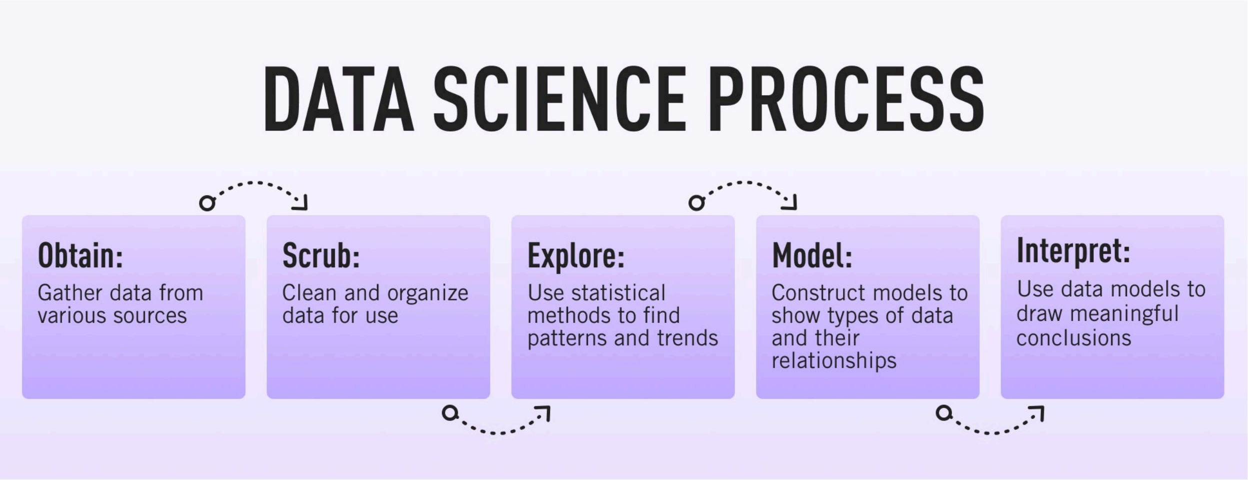
Data Manifesto

Throughout this class, my ability to work with, analyze, and visualize data has not only improved dramatically but it has also allowed me to look at and understand data in new ways. Data, which differs from other things like information, needs to have intent otherwise it is just floating words and characters that have no meaning. As in the DIKW pyramid, data sits at the bottom. It is the structure of all information and knowledge, however on its own it shows us nothing and is meaningless. Data can help you find an answer to a question, explain correlations or even just tell you what song you might like, but it needs to be worked with, controlled and even broken down before it can be understood. Data isn’t harmful, but the things that people can do with it is. Everyday whether you like it or not, you will use, interact with, or create new types of data. Some harmful, some helpful, however what matters is that you know it is happening, and don’t go through day by day without recognizing what is going on around you. Data has been the past, present and future, since the creation of the internet it has grown exponentially however it has always been a part of humanity. As the world begins to evolve, and become more digital, the importance of understanding what's going on with individuals' data will become ever more impactful not only for data scientists but for everyone. Ultimately, data will not be what changes our world, but the way we interact with it will. Which ties directly into my first principle of data science and the foundation of any data science project, Curiosity before Code. Every good

project starts with a meaningful question, not a dataset. People need to think about what they want to understand from this project and what are their goals, before they reach for any tools.

As we step into this new age of Big Data, and AI, the data scientists who take us there will determine its trajectory. As a data scientist it is their job to build a safe, and controllable digital network where data is freely available and yet protected at the same time. As selling personal data becomes more valuable, it is the data scientist's job to stand up for not only themselves but also the individual who may not know what they're giving up. Many people may not have the skill set of a data scientist, yet are still heavily impacted by the data revolution, that is why it is their job to **not only protect but also inform** these individuals of what is happening with their data. Data is often extremely individual, and can highlight things about people that they might not even know about themselves. In Giorgi Lupes Data Humanism piece she highlights this individualism perfectly when she writes, “Data, if properly contextualized, can be an incredibly powerful tool to write more meaningful and intimate narratives.” Data is personal, and just as she says, can write narratives about people in both accurate and inaccurate ways. Which is why I believe the second core principle of my data manifesto is to avoid as much bias as possible. Although data can feel like fact, it is anything but. Data can be skewed, manipulated, or changed to prove any argument, or fight any battle. Data is simply the raw information, but can be used in so many ways. That is why it is the job of a data scientist to include as little bias

in their work as possible in order to make sure that they are presenting their findings in a fair and unbiased setting, that does not try to persuade people in either way. Which is why one of the most important skills to be a data scientist is the ability to see both sides. When performing the data process, which can be broken down into these 5 steps:



DATA SCIENCE PROCESS

Obtain: Gather data from various sources

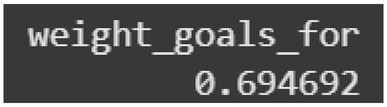
Scrub: Clean and organize data for use

Explore: Use statistical methods to find patterns and trends

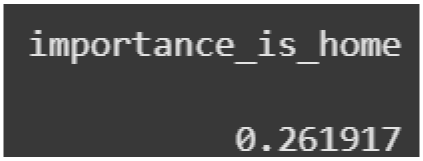
Model: Construct models to show types of data and their relationships

Interpret: Use data models to draw meaningful conclusions

It is of the utmost importance to go through each step trying to keep the work as balanced as possible. It may feel easy to remove a few vital pieces of data to prove a point, or formulate an argument, however it is of the utmost importance for data scientists to show as much as possible, even if it doesn't show what they hope it would. For example, as I worked through my final data science project, I ran into an issue where the linear regression analysis method I had chosen was skewing my data so much that in reality my entire data project had no meaning. As you can see, my goals for, had a 69% correlation method, which meant that whatever way this stat fell made every outcome purely based on goals scored, which funnily enough is exactly what real life is based off too. So my entire project became meaningless. My first thought was to simply remove that little piece of info, and then go on with my project as if nothing had happened. However this was the easy way out, if I had removed that data I would have lost some of that intent I had started the project with. So instead I pivoted, I went back and looked for a new analysis method



```
weight_goals_for
        0.694692
```



```
importance_is_home
             0.261917
```

that would better represent the type of data I was working with and ultimately I found one which worked much better as a prediction model. As you can see my most important stat with the Random Forest Algorithm ended up only being 26%, which was important but would not single handedly decide outcomes. This is just one example of a million, where the analysis you choose might cause biases in your work, or not answer the right questions. Those moments might make you feel like choosing the easy way out but it is important to represent all aspects of the data and keep the project as honest as you can possibly make it. Another extremely important skill is the ability to visualize their findings, as highlighted earlier most people do not have the same skill sets to understand the data analysis that most data scientists do. That is why it is so important for data scientists to visualize their data using a variety of tools in order to make their findings as easy to read as possible, and to help non- data scientists to understand what their data is telling them. Just as Giorgia Lupi says "One size does not fit all", and this specifically relates back to visualizations. It is such a skill for data scientists to not only make clear and legible visualizations but just being able to choose which visualization to choose in the first place can be extremely difficult. Which is why it is such an important skill set to have.

Once these visualizations are created, or data is analyzed most people start to look for connections, correlation or something that tells them X=A. However, as my 3rd principle, Patterns are Not Truths. Just because a pattern appears in the data doesn't mean it reflects reality. Correlation is not causation. Data scientists need to validate, triangulate, and challenge results, especially when they confirm their assumptions. The world is constantly changing, so is data. Therefore the data that might be true today, could be untrue the next. It is important to not take patterns in data as fact but as something that might single a connection, but is important to always be rechecking and redoing to make sure that their data stays truthful and as accurate as possible. Often in our class work we used the college scorecard data, which highlighted a bunch of different statistics about different colleges and universities across the country. However the analysis that we may have performed on this data set this year, can not be used as a truth for next year. Things are always changing, schools make more money, or cut a program, or try to expand. Whatever it is, with time things will change and the importance of using up to date data can not be overlooked.

My final and most important principle is People Over Product. Data science is ultimately about people. Behind every row is a person, and behind every analysis is an intention. I keep humanity, equity, and impact at the center of every project. As data becomes increasingly important in our economic well being, and is constantly being sold and analyzed behind our backs. It is of the utmost importance to remember that every single one of those data sets is a living being, someone with their own unique story, not just another data point. Which is why I believe that the most important part of my data manifesto is to remember that the use of data can be harmful, can impact people's lives, and is not to be taken lightly. Being mindful of what your data projects might do, or being mindful of which data should or should not be used, needs to be the first step in any data project. Before you look at which tools

might be helpful, or what analysis method works best for your data. You need to think about the impact your analysis could have also, where are you getting this information from? Do the people know they are using your data? Is this a respectful way to use it? Would I want other people looking at my data in this way? These are all just a few questions you need to be asking yourself before you start any data science project.

In order to show my 4 principles in action, I am going to walk through the process I would take in a data project I am interested in. If I were to begin a new data project focused on soccer—such as analyzing player performance and scouting undervalued talent across different leagues—this is how I would apply my four core principles. I'd start by asking meaningful questions like: what factors best predict future success for players in smaller leagues, and how might those factors be overlooked by traditional scouting? Before touching code, I'd define what "success" means—minutes played, contribution to team wins, or player development trajectory—and carefully choose metrics accordingly. Next, I'd collect and clean data from multiple sources (player stats, transfer history, team styles), while being mindful of biases—like overrepresentation of certain leagues or the subjectivity in scouting reports. During analysis, I might use clustering or regression to identify overlooked player archetypes, but I'd remember that just because a model finds a pattern doesn't mean it reflects reality or guarantees success. Finally, I'd prioritize how this analysis serves people—coaches, players, clubs—by presenting insights in a clear, responsible way that respects players' individuality, avoids reduction to pure numbers, and supports more equitable decision-making in the sport.

Ultimately data science is a field that will continue to change a lot every single day, and will become one of the most important industries in the world. It can be used to save lives, spread awareness, or help people in need, or it can be used to sell products, and dictate war zones. The data itself is not the weapon nor the cure, it is simply something out there in the world. Floating on pieces of paper, written on a wall or floating somewhere in the cloud, the data itself can not hurt or help us. The way that we use it will. Although I have my own Data Manifesto and my own principles, every data scientist needs to create their own as well. Everyone needs to hold themselves accountable for the things they create, and hopefully overtime the positive impacts that data science can have will be felt by people all over the world.