SPRING 2023

DATA MANIFESTO

CS215

Aby Ramata Mbaye

As the field of data science continues to evolve, it's important for aspiring data scientists to approach their work with a set of guiding principles that will help them navigate the complex landscape of data analysis. In this essay, I will detail four key principles that I believe are crucial for any aspiring data scientist to keep in mind: serendipity, breaking conventions, understanding context, and interdisciplinary collaboration.

(Informed) Serendipity

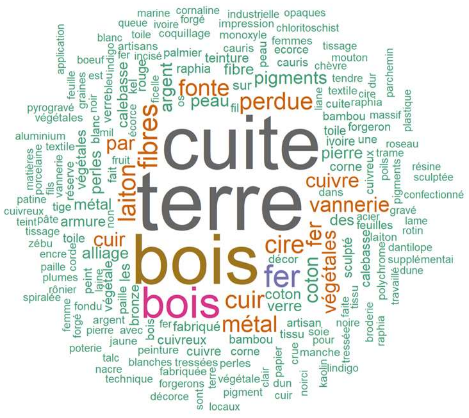
When I first started learning about data science, I was overwhelmed by the sheer amount of information and the complexity of the algorithms and tools. I felt like I needed to follow strict formulas and patterns to achieve success. However, I soon realized that this rigid approach was limiting my potential and causing unnecessary stress.

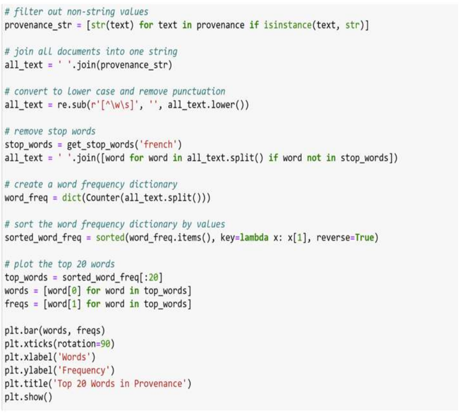
Serendipity, or the ability to embrace unexpected insights and discoveries, is an essential aspect of successful data science. By trusting our instincts and allowing ourselves to explore beyond established norms, we open ourselves up to a world of new possibilities. As Melanie Feinberg writes in Everyday Adventures with Unruly Data, "the magic of serendipity is unexpected insight."

Additionally, informed serendipity is a safeguard against the dangers of blindly following algorithms without a firm understanding of their quality, as Hannah Fry warns in "Power" from Hello World. By taking the time to understand the mechanisms behind a code, we can save time and prevent unnecessary errors, ultimately leading to more successful outcomes.

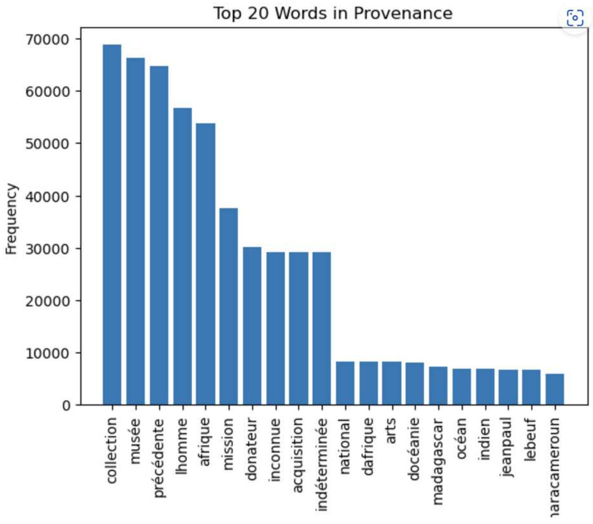
The following quote encapsulates the idea of informed serendipity principle: *'Because trusting a usually reliable algorithm is one thing. Trusting one without any firm understanding of its quality is quite another.'*

Breaking Conventions: Brainstorming before browsing.





```
# filter out non-string values
provenance_str = [str(text) for text in provenance if isinstance(text, str)]

# join all documents into one string
all_text = ' '.join(provenance_str)

# convert to lower case and remove punctuation
all_text = re.sub(r'[^\w\s]', '', all_text.lower())

# remove stop words
stop_words = get_stop_words('french')
all_text = ' '.join([word for word in all_text.split() if word not in stop_words])

# create a word frequency dictionary
word_freq = dict(Counter(all_text.split()))

# sort the word frequency dictionary by values
sorted_word_freq = sorted(word_freq.items(), key=lambda x: x[1], reverse=True)

# plot the top 20 words
top_words = sorted_word_freq[:20]
words = [word[0] for word in top_words]
freqs = [word[1] for word in top_words]

plt.bar(words, freqs)
plt.xticks(rotation=90)
plt.xlabel('Words')
plt.ylabel('Frequency')
plt.title('Top 20 Words in Provenance')
plt.show()
```



Breaking conventions is a crucial principle for aspiring data scientists. To be truly successful in this field, we must be willing to explore beyond traditional data sources and methods. For example, in Seth Stephens-Davidowitz's book "Data Reimagined," he gives an example of Jeff Seder's horses, which illustrates the importance of exploring unconventional data sources. For years, experts relied on a method that yielded inaccurate results until Seder's data analysis revealed the true predictors of horse racing success.

After reading Georgia Lupi's book "Data Humanism," I had a moment of self-reflection about my approach to data visualization in my internship. Lupi warns against blindly relying on out-of-the-box charts and tools to make sense of data without properly framing the underlying questions. This is precisely what I had been doing, including my use of word clouds. However, during a presentation at the Whitman Undergraduate Conference, I discovered the limitations of word clouds. While they look visually appealing, my audience struggled to grasp the context behind the words in a concise way. After further research, I found numerous articles detailing the shortcomings of word clouds, such as accessibility issues and inaccuracies resulting from longer words taking up more space. As I continue to work on my projects, I am reflecting on how to effectively communicate my results in a way that goes beyond aesthetics. Ultimately, I replaced the word cloud with a simple bar graph in my report.

By challenging established norms and experimenting with unconventional approaches, we can achieve breakthrough insights and make significant contributions to the field of data science.

Context

Understanding context is another essential principle that aspiring data scientists must keep in mind. It's not enough to simply analyze data without considering the cultural and social norms of the communities we serve. For example, consider the story my father shared with me about a rural community where the United Nations invested millions of dollars to provide potable water to the people. Despite the impact analysis showing that the implementation of the water pump project would bring tremendous benefits to the community, the pump was barely used. Upon investigating the matter, my father discovered that the women, who were the primary users of the water pump, were experiencing miscarriages from the movements of handling the pump between their legs. This crucial piece of information was not included in the data analysis, and it had severe consequences. This story is a reminder that data science is not just about numbers and algorithms; it's about understanding the people behind the data and their unique contexts. As data scientists, we must approach our work with empathy, an open mind, and a willingness to learn about the communities we serve. Only then can we make informed decisions that benefit everyone involved.

Not just a data scientist

Finally, it's important for aspiring data scientists to remember that they are not just data scientists. While data science is certainly an exciting and rewarding field, it's important not to lose sight of the broader context and purpose of our work. Our last reading about the 'Henry Higgins' effect exemplifies the importance of collaboration in data science. The chapter titled "The Henry Higgins Effect" describes how the lack of diversity in data science teams has led to significant gaps in knowledge and understanding. These gaps have resulted in biased algorithms and models, which ultimately perpetuate systemic inequalities. By working collaboratively across disciplines, data scientists can bring a broader range of perspectives to their work, ultimately leading to more comprehensive and accurate results. Additionally, collaboration with experts in other fields, such as social scientists and policy makers, can help data scientists better understand the real-world implications of their work and ensure that their models and algorithms are ethical and responsible.

As I reflected on my own career goals, I realized that what drew me to data science was a desire to understand the challenges faced by vulnerable communities and provide evidence-based solutions to them. By collaborating with experts from different fields, we can gain a more comprehensive understanding of the problems we're trying to solve and develop more effective solutions.

In conclusion, data science is not just about numbers and codes; it's about embracing serendipity, breaking conventions, understanding context, and fostering interdisciplinary collaboration. By incorporating these principles into our work, we can make a positive impact on society and contribute to a more equitable future.