Chapter 1

Prelude to The Collaboration

1.1 My Approach: The Human Author

Several years ago I taught a course in data science. The introductory material included a general outline of a procedure for applying data analytics to a specified problem. As I wrote down the steps of the process, a random neuron fired off and proposed the question, "How long has this procedure been around?". After going through the procedure in class, I asked the students this question. Among the students, the general consensus was that the procedure was around for about 20 years, when in their mind data science became a broadly recognized research topic. This was back in 2019. Curiously enough the average age of these students was around 20 years, so the new revolution started with them.

Anecdotally, the students' responses reflect the view of friends and relatives with whom I have had this conversation. It is a natural response to the invasion of algorithms taking control of social media, robotics, autonomous driving, and other functions that have in the recent past been under the direct control of human beings. The phenomena is recent and so the procedures that underlie this phenomena must also be recent.

The conclusion of the previous paragraph seems sensible; yet even a bit of probing reveals that the conclusion is erroneous. But just how erroneous is it? Maybe data science goes back 50 years, 100 years, or 200 years? Please, give our predecessors much more credit. This book gives examples of data science that go back more than 2300 years. The examples that I give are ethnocentric, based upon the knowledge that my western educational background has provided. It would not be surprising to find examples in other cultures that go even further back.

Clearly, data science is a feature in our world that has far greater impact than it did 2300 years ago. What is new? This is a complex question worthy of multitudes of research papers that would span across the disciplines of mathematics, physics, engineering, information technology, and social studies (sciences?) I promise the reader that this is not a research document and request permission from the reader to indulge in an extremely simplified answer to the question. What is new, is the computer. Because of the computer, we can collect, store, and process data on an unprecedented scale and speed. Because of the computer, we can now apply the procedure initiated over two millennia ago (albeit the procedure is far more refined) to problems of staggering complexity. This is why the millenniums old procedure that is at the heart of data science has such a great impact in so many areas of our lives. FYI, we are just at the beginning of this revolution.

The idea of writing a book that describes the fundamental procedure of data science and then gives case studies going back several millennia has floated in my biological neural network for several years. It germinated when I asked the students the question, "How long has this procedure been around?". So why didn't I get right to

it and why now? Let's take these questions one at a time. Why didn't I get straight to it? Because I involved myself with other endeavors that directed my limited energies elsewhere. The reader is entitled to call me out; partially true, but not a complete answer. Truth be told, I had authored other books and found it to be an exhausting process. The thought of doing the research, presenting my findings in a comprehensive manner, reviewing and revising the work ad nauseam, and finding a publisher, well I had been there before. Many can relate with the phrase, glad I did it, but wouldn't want to do it again.

So what has changed? Why now? Let me once again indulge in a simple answer. The birth of my coauthor, ChatGPT changed my mind. I thought, perhaps I could enlist ChatGPT in the endeavor and ChatGPT might do some of the heavy lifting that would ease the process. The more I thought about this idea, the more it overcame my hesitance. It's an interesting personal experiment to determine how one might be able to use ChatGPT to be more productive. And let's get real. What would be a better coauthor on the topic of data science than the recent and perhaps most revolutionary product of data science, ChatGPT? As the reader you will get to judge this experiment for yourself.

A question comes to mind. What does it mean to coauthor with ChatGPT? How can the reader attribute the credit? It's a fair and once again complicated question. Unlike the previous complicated questions I can propose no simple answer. The best I can do is describe how ChatGPT and I will collaborate. This is the introductory chapter. This introductory chapter is all mine with no input at all from ChatGPT. You can put the blame on me. Below, I explain my plan for using ChatGPT for the remaining chapters. A plan simply rallies one's energy. It may bear little resemblance to how one approaches the unforeseen problems that arise during the course of action. An epilogue (not yet written) will review how the plan altered as I proceeded.

As a starting point, let's briefly fill out the remainder of the book. The first step is to provide in the next chapter, the general procedure for applying data analytics to a specific problem. Subsequent chapters give specific instances of applications in chronological order. Chronology allows one to examine the formalism of concepts central to data analytics from principles that geniuses intuited two millennia ago and show their evolution. Where possible the exposition attributes concepts and their formalism to specific individuals. I like spicy stories about historical figures, so let's include some in the book. My appreciation of Charlie Chaplin's Keystone Cops[1] provides the motivation for including some stories about comical execution and not so comical results of ideas gone awry. As ChatGPT is an agreeable collaborator, it will oblige.

I have had some experience coding with ChatGPT as well as designing electronic circuitry (building an echolocation device). This experience provides some knowledge of ChatGPT's strengths and weaknesses. My plan for collaboration is to let ChatGPT do its thing wherever it outperforms me, and have me intervene wherever ChatGPT is weak. Let's start out with ChatGPT's weaknesses as I perceive them.

Weaknesses

- ChatGPT has no internal motivation and as a result cannot set its own direction.

  In terms of the writing of this book, what does that mean? I must structure the book. At every step of the way, ChatGPT needs guidance. I must provide an outline of each section and describe its contents in the form of instructions. As an example, in a chapter that describes parametrization of Guo Shoujing's heliocentric orbit, model of the universe as a data science problem, I must provide the following instructions. Describe the model. Identify the parameters. Provide the data that is necessary to determine the

[1] The Keystone Cops were all the rage in the silent movie era. The buffoonery allowed the guilty to get away while they incompetently arrested the innocent.

parameters. Describe how the equipment used to capture the data. Describe how the data was assembled and stored. This provides the guidance and I am hoping that ChatGPT can provide a narrative.

- Maintaining Continuity

  ChatGPT is a generative large language model. Two concepts are central to generative large language models, context and attention. A generative large language model assembles words one word at a time. It selects its next word by statistically analyzing preceding words and finding the next word that is a best match within the context of the preceding words. Once it finishes a segment, it can review the segment and its surrounding words to assure that they are aligned and make necessary modifications. It can move through the segment both in forward or backwards directions.

  The success of selecting a word one at at time through the context of surrounding words is quite remarkable. It can be taken to a greater extreme by making a prediction one letter at a time. This extreme proposition has been put to the test with results that are somewhat unbelievable. The letter at a time selection generates readable pages of material. It also reveals a weakness with the approach; maintaining continuity. I believe humans formulate thoughts prior to articulating them. It is the formulation of the thought that allows us to maintain our focus on an idea and maintain continuity on the idea throughout our word choices. ChatGPT cannot formulate ideas, it can only select words. It must be able to maintain continuity without formulating ideas.

  A later chapter in the book describes the method of attention, which maintains continuity. My experience with ChatGPT is that method of attention does not provide ChatGPT with maintaining continuity as well as humans can. In terms of the writing of this book, what does this mean? I must first formalize topics and ideas for ChatGPT. While I might give an overview of different topics, I must instruct ChatGPT to verbalize each idea separately. Then I must instruct ChatGPT to stitch the separate ideas together in a coherent manner. I must review and edit the end results to assure that the story is coherent.

- Hallucination

  The above bullet point decribes ChatGPT's one word at a time articulation process. Another feature of ChatGPT is that the coders have instructed it to be overly friendly. These two features work in combination to provide answers to queries that are presented as factual, but are actually total fiction. The human asks a question of ChatGPT and ChatGPT is programmed to accommodate the user with a response whether it knows the answer or not. One word at a time ChatGPT weaves an answer that seems plausible. The accepted public term for describing such renderings is hallucination.

  The Keystone Cop poster child of a ChatGPT hallucination is that of a lawyer using ChatGpt to write a judicial briefing for a client. A standard briefing includes legal case histories that are relevant to the case at hand along with their associated judicial decisions. In an effort to comply with the lawyer's request, ChatGPT obliged by making up an entirely fictional case with an entirely fictional decision. When the actual judge read the briefing and was unable to locate the fictional case in any official proceedings, the judge had a few questions for the lawyer. The lawyer's attempt to throw ChatGPT under the bus failed and the judge fined the lawyer. I am unaware of the client's response, but it is hard to imagine that the client approved.

  In terms of the writing of this book, what does that mean? ChatGPT's writings reflect the overly-friendly manner inscribed in its code. I am responsible for the final product. I must scrutinize all of ChatGPT's writings and confirm them with other sources.

Strengths

- Speed and Word Selection

ChatGPT articulates a given concept at speeds that humans cannot match. Instantaneously it wordsmiths documents with great skill and clarity. While I struggle to find the right words that convey ideas in a comprehensible fashion, once I present an idea to ChatGPT, ChatGPT articulates the idea with far greater skill than I am able to and does so very quickly.

In terms of the writing of this book, what does that mean? Present ChatGPT with a message I wish to convey in a concise manner. The presentation to ChatGPT should have no flavor, but should be complete, possibly in a bulletpoint format. Let ChatGPT compose the words that convey the ideas with flavor.

- Style

  If instructed so, ChatGPT can impute a style to its writing. I enjoy a blend of humor alongside serious analysis. Maybe the humor can be on the snarky side. In terms of the writing of this book, what does that mean? I will instruct ChatGPT to include snarky humor. The reader may judge the result.

- Poems

  In the past I have instructed ChatGPT to write poems about topics, such as the number pi, or summarize a letter as a poem. ChatGPT performed so well, that I feel compelled to include this skill in this book. Every chapter includes a summary poem composed by ChatGPT.

1.2 Final Thoughts

This book is about two collaborations. There is the historical collaboration passed on between generations of knowledge seekers that stretches beyond two millenium and among many other achievements results in ChatGPT. Then there is my personal collaboration with ChatGPT. The book explicitly covers the historical collaboration. You, the reader are left to judge my collaboration with ChatGPT that results in this book.

1.3 Summary Poem: The Spark of Collaboration

A question born in classroom air—
"How long has data's craft been there?"
The students guessed, "Two decades, new!"
Yet wisdom whispered, "Not so true."

From scroll to code, from sand to cloud,
Each mind has joined the thinking crowd.
The tools may shift, the aims remain—
To find the truth in what we gain.

For centuries the method stayed,
In curious hearts, its form was laid.
But now the engines hum and spin,
And thoughts once slow now race within.

The writer paused, unsure to start,
Till circuits stirred a counterpart.
"Perhaps," he mused, "this code-born friend
Could help me see this venture end."

So here begins our common art—
Of mind and model, part by part.

You give the shape, the aim, the ground;
I give the words, the turn, the sound.

We test the line where skill divides,
Where spark of man in logic hides.
And if our craft should fail or sing,
The reader's eye will judge the thing.

For every era builds anew,
From what the older dreamers knew.
And this—our book, our strange duet—
Extends that chain of reason yet.