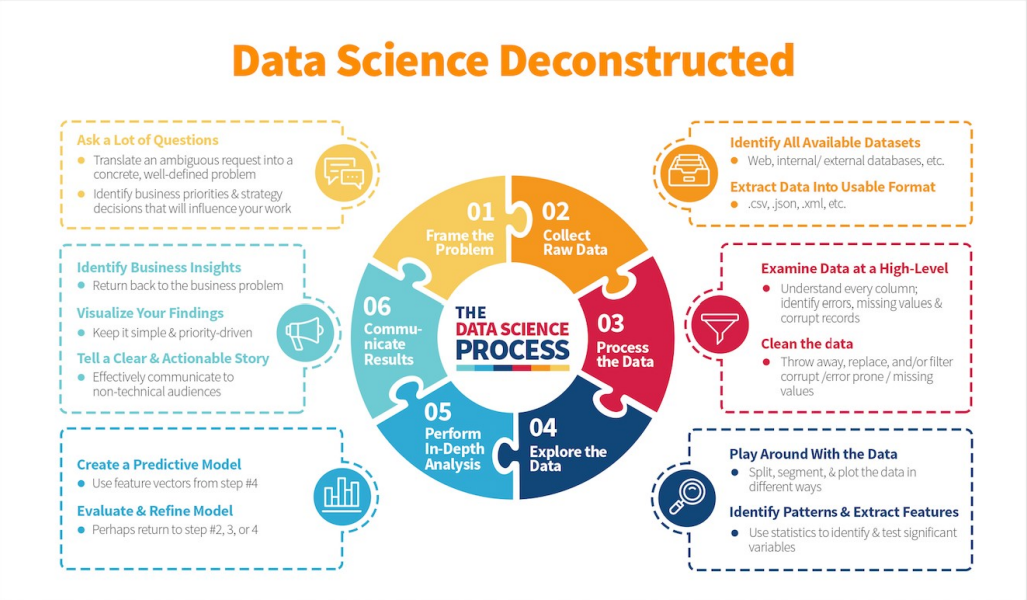
# AJ GOLDSTEIN

Engineering Student. Aspiring Entrepreneur. Part-Time Philosopher.




**NOVEMBER 12, 2017**

# Deconstructing Data Science: Breaking The Complex Craft Into Its Simplest Parts

*This is the SECOND in a series of posts on applying Tim Ferriss' accelerated learning framework to Data Science. My goal is to become a world-class (top 5%) Data Scientist in < 6 months, while open-sourcing everything I find & learn on the way.*

***The purpose of this post is to empower others to start***

***accelerating their own learning by:***

1. *deconstructing the complex craft of Data Science into its simple micro-skills*
2. *identifying the 20% of skills that contribute to 80% of outcomes*

*And if you stick around until the end, you're in for a special treat.*

Estimated reading time: 15 min ( to save you hours of spinning in circles 😉 )

## The Problem


A simple Google search of "how to learn Data Science" returns thousands of learning plans, degree programs, tutorials, and bootcamps. It's never been more difficult for a beginner to find signal in the noise.

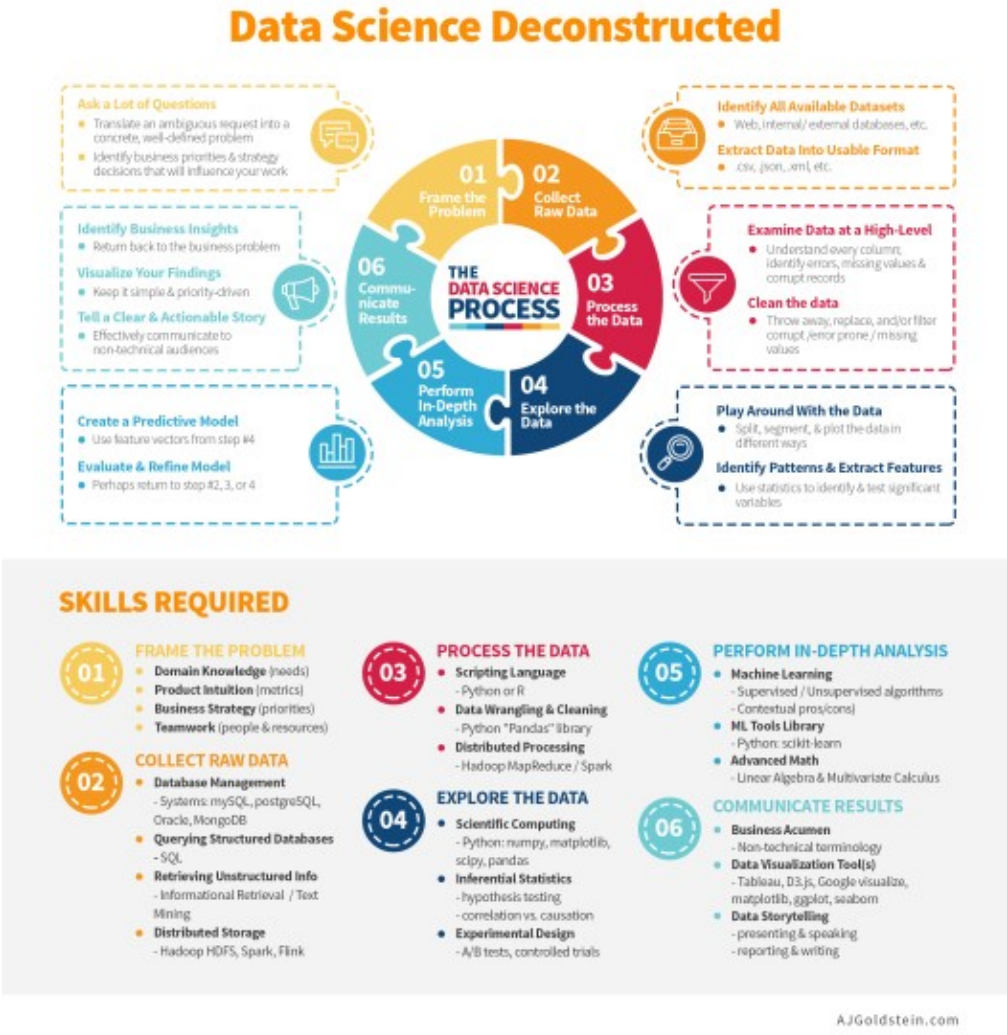
Everyone seems to have a different opinion, and the only common approach appears to be dumping a long list of courses to take and books to read, all the while providing little to no context into how these concepts fit into the bigger picture.

**This post is my attempt to convert all the buzzwords & fluffy terminology into explicitly-learnable skills.** To do this, I'll be walking through my application of the first two steps to Tim Ferriss' accelerated learning framework: Deconstruction & Selection.

Rather than jump right in to a roadmap of my own learning journey (that'll be next post), I want to empower you to begin your own. And if you haven't read my first post, I'd highly recommend starting there: www.ajgoldstein.com/learning-without-limits/

# Deconstruction: The Data Science Process

*'The whole is greater than the sum of its parts." – Aristotle*



*I'll be walking through this infographic step-by-step below*

It's true: Data Science is not a single discipline, but a craft at the intersection of many. So in order to appreciate how the seemingly disparate puzzle pieces fit together, I present to you a story. It's called "The Data Science Process", and it has six parts:

1. **Frame the problem**: who are you helping? what do they need?
2. **Collect raw data**: what data is available? which parts are useful?
3. **Process the data**: what do the variables actually mean? what cleaning is required?
4. **Explore the data**: what patterns exist? are they significant?
5. **Perform in-depth analysis**: how can the past inform the future? to what degree?
6. **Communicate results**: why do the numbers matter? what should be done differently?

But before we begin, a couple quick caveats:

1) In large organizations, "The Data Science Process" is often carried out by an entire team, not a single individual. An individual can specialize in any one of the six steps, but for simplicity, we'll be *assuming a one-person team*.

2) The insights that follow are a compilation of various expert interpretations; *not my original ideas*. I am not (yet) an expert Data Scientist, but over the past 6 weeks I've learned from many. Thus, I'm simply serving as the filter between hundreds of hours of research and the actionable insights you'll find below.

In particular, I'll be pulling from favorite online articles (linked throughout) and conversations with the following 10 experts:

1. Chris Brooks — Director of Learning Analytics at the University of Michigan
2. Andrew Cassidy — Freelance Data Scientist & Online Educator
3. Jim Guszcza — US Chief Data Scientist at Deloitte Consulting

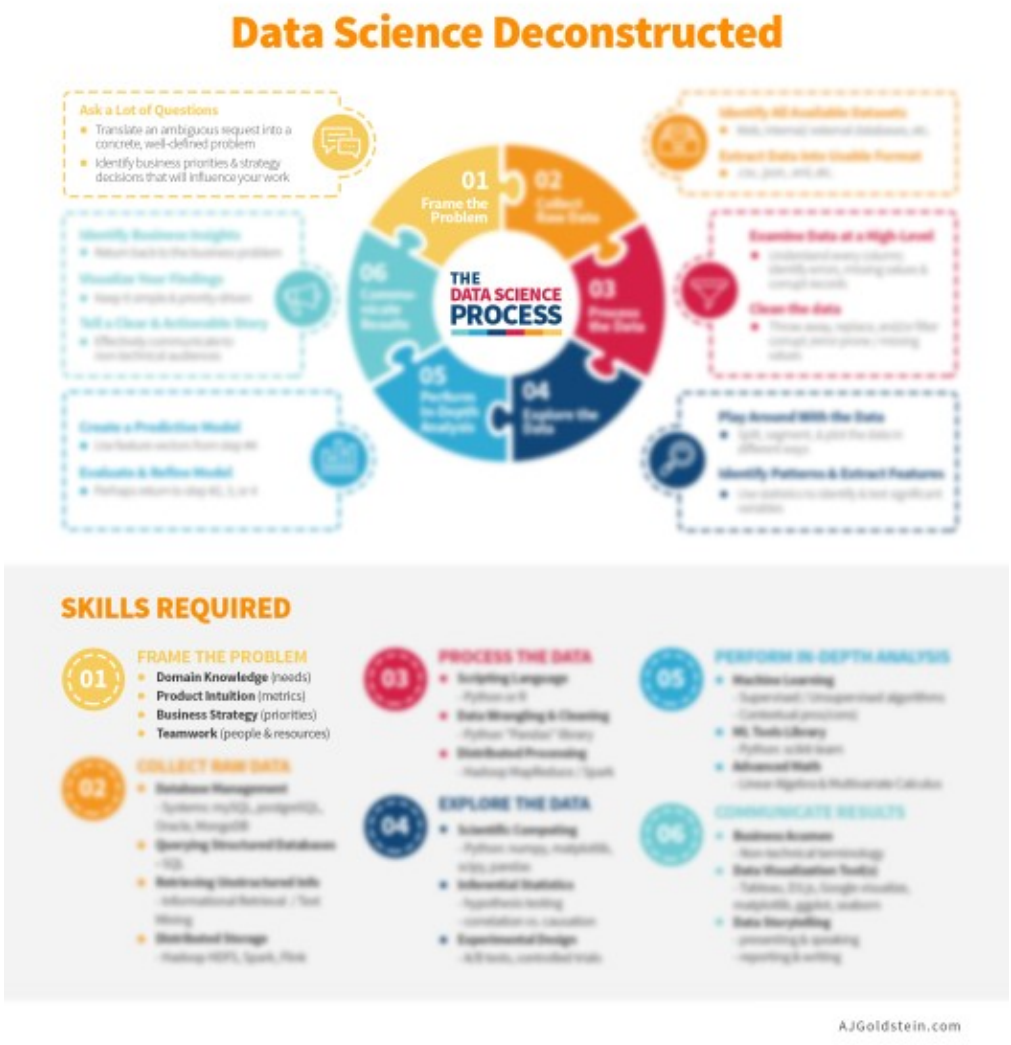
4. Kirk Borne — Principal Data Scientist at Booz Allen Hamilton
5. Michael Moliterno — Data Scientist + Design Lead at IDEO
6. Chris Teplovs — Research Investigator at the University of Michigan
7. Jonathan Stroud — Co-Founder of the Michigan Data Science Team (MDST)
8. Josh Gardner — Data Science Research Associate, Team Leader on MDST
9. Jared Webb — PhD Candidate in Applied Math, Data Manager at MDST
10. Alex Chojnacki — Data Application Manager for Flint-Water-Crisis project

And to bring each step of the process to life, I'll be using my work at Calm.com, Inc. in San Francisco this summer as a real-world case study.

While there, I leveraged analytics insights from Calm's database of 11 million users to develop & launch Calm College — the first US platform geared toward using mindfulness to improve college student mental health.

Alright, let's get started!

## Step One: Frame The Problem




The first step of The Data Science process involves asking a lot of questions.

The exact manner in which you do this will depend on the context in which you're working, but whether you're in the private sector, public sector, or academia, the key idea is the same: **before you can start to *solve* a problem, you have to deeply *understand* it.**

Your goal here is to get into the clients' head to understand their view of the problem and desired solution. In the case of a corporation, this will first involve speaking with managers & supervisors to identify the business priorities and strategy decisions that'll influence your work.

It's not uncommon for the first request that a Data Scientists' receives to be entirely ambiguous (i.e. "we want to increase sales"). But it'll be your job to translate the task into a concrete,

well-defined data problem (i.e. “predict conversion rate & return-on-investment across customer segments.”)

This is where domain knowledge and product intuition is crucial. Speaking with subject-matter-experts to cut through confusing acronyms & dense terminology can be incredibly helpful here. And familiarizing yourself with the product/service will be essential to understanding the intuition behind metrics.

**For example…**

With Calm College, the ambiguous request we started with was to establish partnerships with universities to offer the Calm app as a student wellness resource.

To better understand our specific domain, we started by spending two weeks speaking on the phone with as many college administrators as possible.

We asked questions like:

- How would you describe the mental health climate on your campus?
- How high of a priority is improving student mental health?
- What main resources do you currently offer students?
- What have been the greatest challenges?
- Is there precedence for offering 3rd party services?

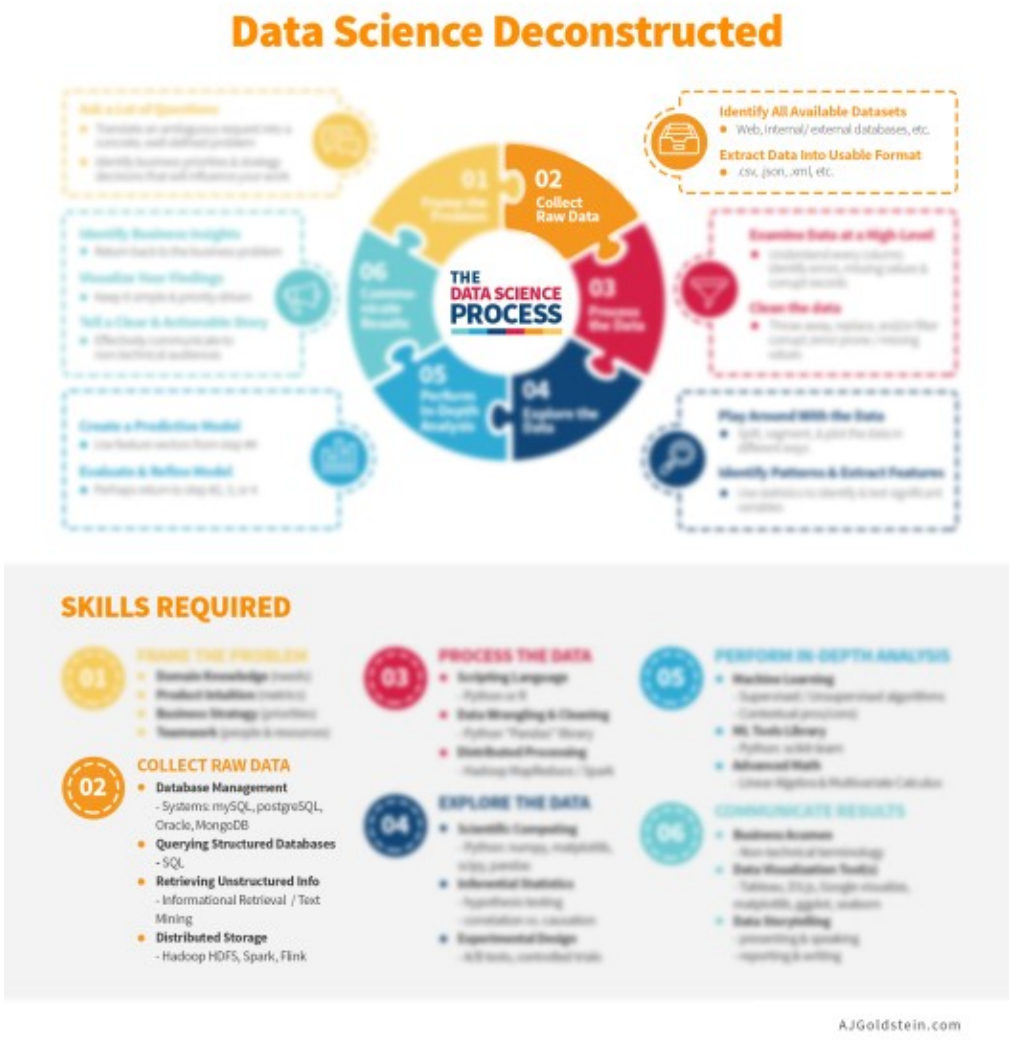
By the time we got to the final question, nearly every administrator had described their campus’ mental health climate as nothing short of “toxic”, and expressed improving it as their #1 priority.

They explained that the greatest challenge to students seeking help has been overcoming logistical issues (i.e. wait-time, transportation, & money) with the counseling services they currently offer.

Finally, here’s where our ambiguous request became a data problem…

Administrators told us that, before a 3rd party service can be adopted, precedence requires evidence supporting its use. In other words, showing that students on campus are already using the Calm app would be crucial to getting a deal done.

## Step Two: Collect Raw Data




The second step of the Data Science Process is typically the most straightforward: collect raw data.

This is where your first technical skill — querying structured databases with SQL — comes into play. But fret not; it's not as complicated as it may sound.

Here's an awesome tutorial by Mode Analytics that'll get you started with SQL in just a couple hours.

More important than the querying itself, however, is your ability to identify all the relevant data sources available to you (e.g. web, internal/external databases) and extract that data into a useable format (e.g. .csv, .json, .xml).

Oftentimes, an analysis requires more than one dataset, so you'll likely need to speak with backend-engineers in your organization who are more familiar with what data is being collected and where it currently resides. **Communication is key.**

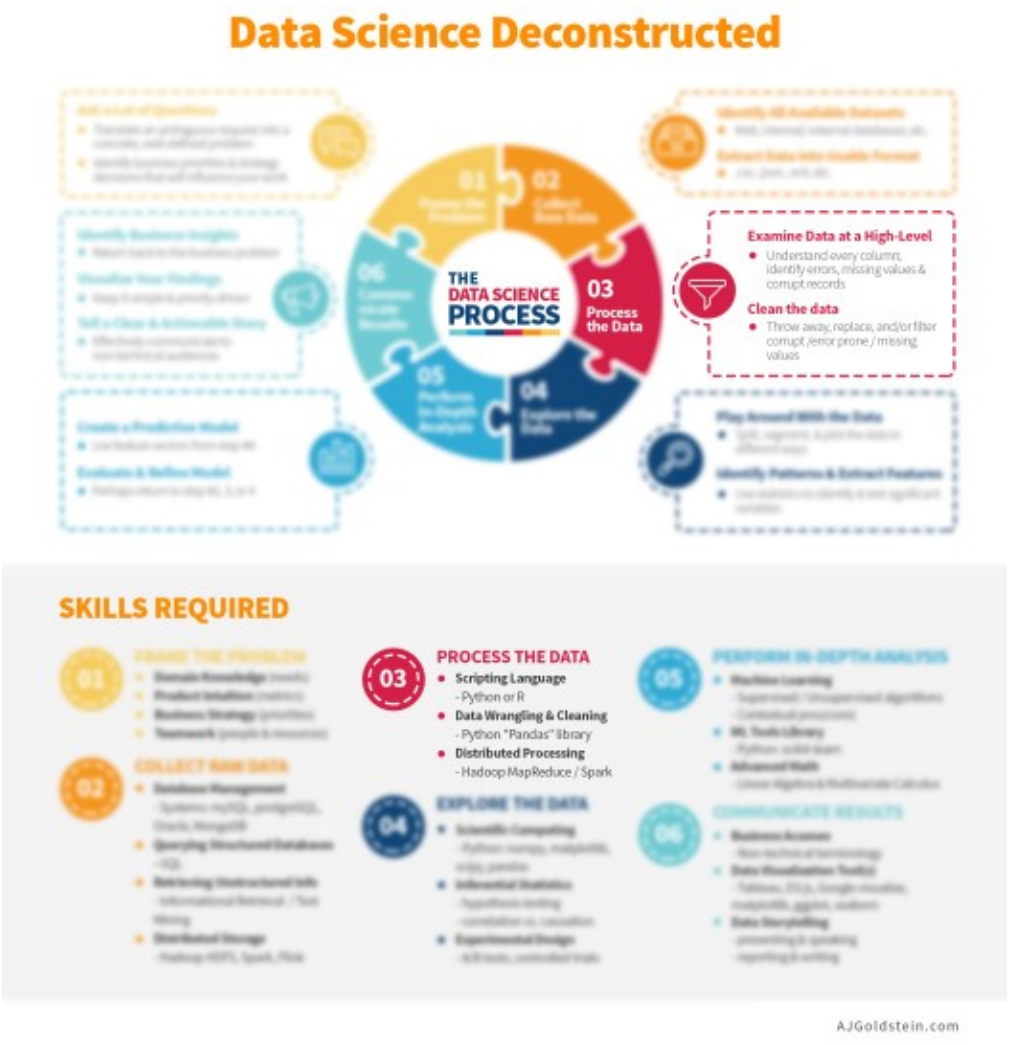
**For example…**

With Calm College, this required me sitting down with Calm's lead engineer and exploring ways to pull usage data for specific college campuses.

Ultimately, I found out that we could simply query user activity by email address and school location. So for the University of Michigan, for example, I simply searched the database for emails ending in "umich.edu" or locations listed as "Ann Arbor, MI".

This approach wasn't full-proof (turns out not all students were using their school email) but it did the job by giving us a representative sample of ~1000 users per college to compare different campuses' activity head-to-head.

## Step Three: Process The Data




The third step of the Data Science Process is the most underrated: process the data.

This is where a scripting language like Python or R comes into play, and a data wrangling tool like **Python's Pandas** is absolutely indispensable.

To get started, here's a breakdown of Python vs. R, intro to Python on Codecademy, 10-minute tutorial to Pandas, and colorful data wrangling cheat-sheet.

Data cleaning is typically the most time-intensive part of data wrangling. In fact, in expert surveys **it's been estimated that up to 80% of a Data Scientists' time is spent here: cleaning & preparing the data for analysis** (more on this below).

The reason this can be so time-consuming is because — before you

can analyze data — you have to go column-by-column, developing an understanding for the meaning of every variable and then checking for bad values accordingly.

The tricky part is that a bad value can be defined as many things: input errors, missing values, corrupt records, etc. And once you've identified a "bad value", you have to decide whether it's most appropriate (given the situation) to throw it away or replace it.

**For example…**

With Calm College, I faced two significant roadblocks here:

1. There was little to no company documentation on database variables
2. I didn't know Python's Pandas and felt too intimidated to try and learn

Each of these presented their own challenge:

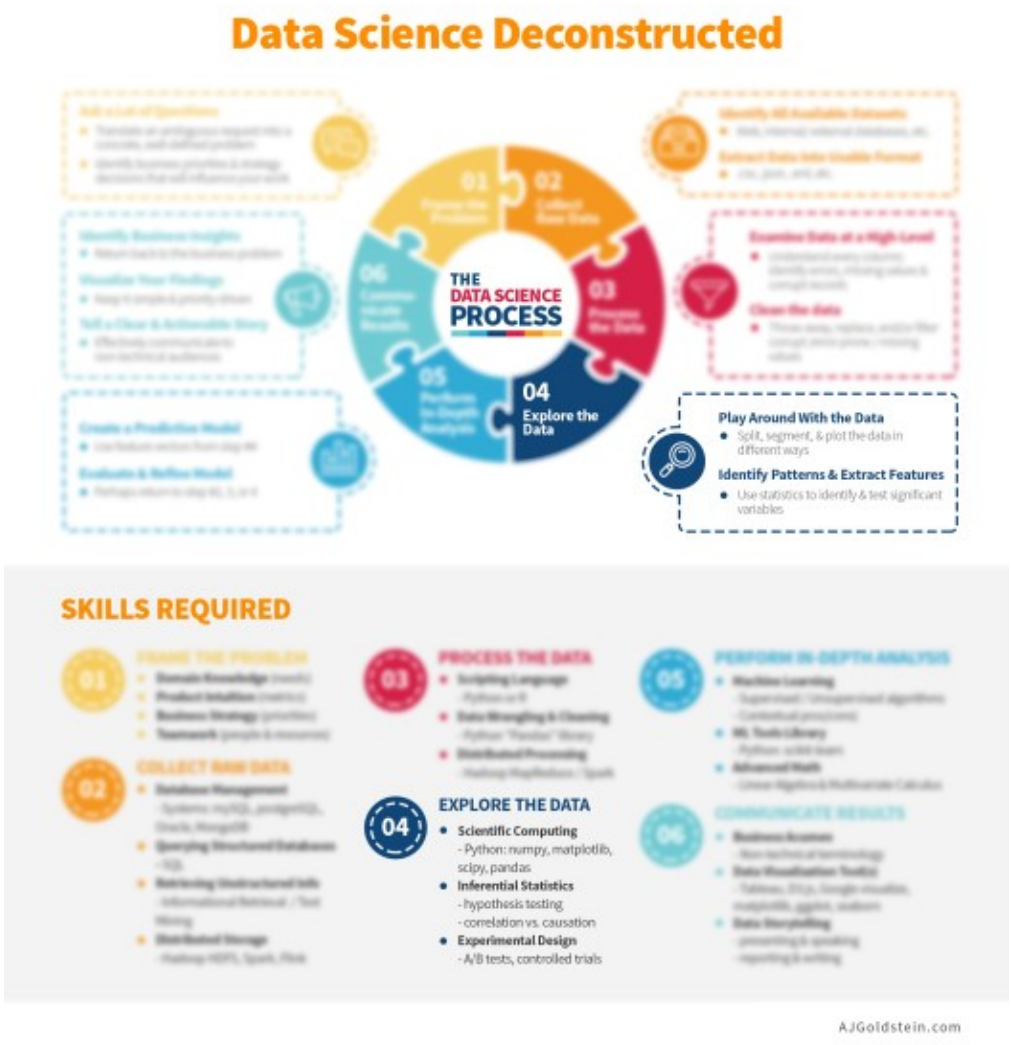
1. It took me several days to figure out how to define an "active user" (i.e. should 'active' mean opening the app, starting a session, or completing a session?)
2. I had to use an analytics tool called Amplitude rather than coding in a script file.

After talking with Calm's Product Manager, I was able to define an active user as someone who "starts a meditation session" and identify the right variables. Then I had to clean the data by filtering out students who hadn't been active in the last 365 days.

The thought process here was that administrators (i.e. our client) would primarily be interested in student activity from the past academic year, and non-active students (i.e. "null" values) were outliers that, if included, would only skew the results.

Noticing a theme here? **It's about your clients' interests, not your own.**

## Step Four: Explore The Data




The fourth step of the Data Science Process is where you explore the data, and the real adventure begins.

This is where the core competency of scientific computing (i.e. Python's numpy, matplotlib, scipy, & pandas libraries) comes into play.

To begin, here's an awesome breakdown of the "SciPy ecosystem" (a collection of libraries in Python), extensive guide to data exploration, and a conceptual handbook of assumptions/principles /techniques.

Using these libraries, you'll split, segment, & plot the data, in search for patterns. Thus, the key is becoming really comfortable with

producing quick & simple bar graphs, box plots, histograms, etc. that'll let you catch trends early on.

Remember that analysts who produce beautiful externally-facing visualizations often have to iterate through hundreds of internally-facing ones first. So playing around with possibilities in this way is more of a guess-and-check art than a hard-and-fast science.

Finally, once you've identified some patterns, you'll want to test them for statistical significance to determine which are worth including in a model. This is where a strong grounding in inferential statistics (e.g. hypothesis testing, confidence intervals) and experimental design (e.g. A/B tests, controlled trials) is essential.

**For example...**

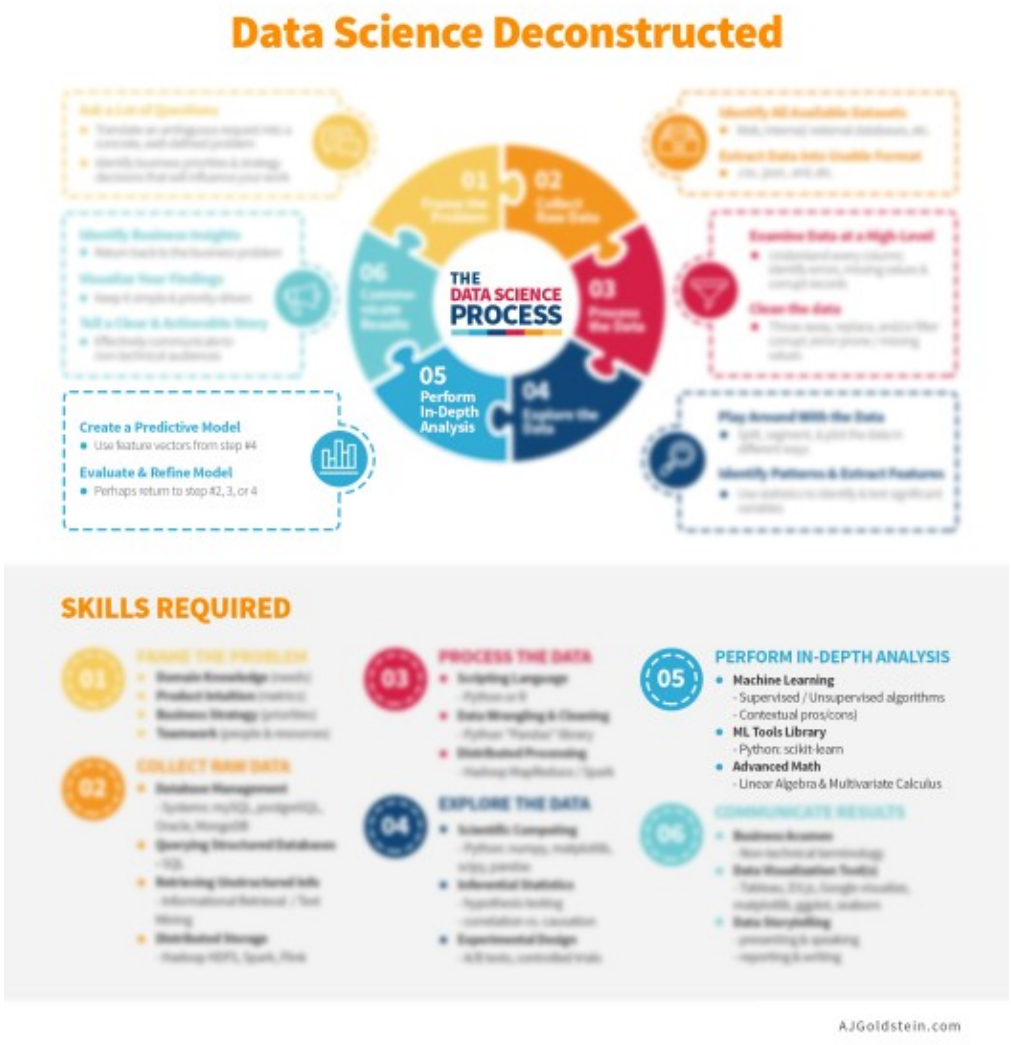
With Calm College, I started by exploring factors that would influence a potential partnership: monthly engagement, week-by-week retention, and subscription rate.

My hypothesis going in was that elite schools known for student stress (i.e. Cornell, Harvard, MIT) would have significantly higher numbers across the three statistics. Or, in other words, I suspected that *stressed-out kids need more calm*.

To test this, I began by segmenting universities into their regional groups and then splitting areas into specific college towns. From there, I was able to compare the statistical significance of schools' activity across local, regional, and national averages.

After several iterations of my experimental design (and hundreds of internally-facing visualizations), I found what I was looking for: a list of outlier schools that we would ultimately call "Calm's Most Popular Colleges".

## Step Five: In-Depth Analysis




The fifth step of the Data Science process is where you create a model to explain or predict your findings.

This is where most people lose the forest for the trees, as they enter into the land of shiny algorithms and fancy mathematics. **Creating models is by far the most over-glorified part of Data Science**, which is why most degree programs solely focus on this single step.

But before jumping in to a particular solution, it's important to **pause and return to the bigger picture by asking yourself: "what am I really trying to do and why does it matter?"**.

From here, you'll:

1. apply your knowledge of algorithms' contextual pros/cons to choose one approach best-suited for the situation

2. carry forward statistically significant variables (from the exploratory phase) using what Data Scientists call "feature engineering"
3. use a machine learning library like scikit-learn for implementation.

The overall goal is to use training data to build a model that generalizes to new (unseen) test data. So while building, it's important that you're keenly aware of (and capable of recognizing) overfitting and underfitting.

Here are some amazing free videos from Andrew Ng's Machine Learning course and Harvard's CS109 "Intro to Data Science" class that will teach you how to do this for different algorithm types. A great place to practice is through Kaggle tutorials.

*NOTE: I'd recommend starting by watching just one or two videos on a simple model type like logistic regression or decision trees, and then immediately applying what you've learned on a dataset you care about.*

**For example...**

With Calm College, the model I was building was more "explanatory" than "predictive".

That is, I was simply trying to identify the universities most suitable for a partnership and understand what factors about a school were contributing to that.

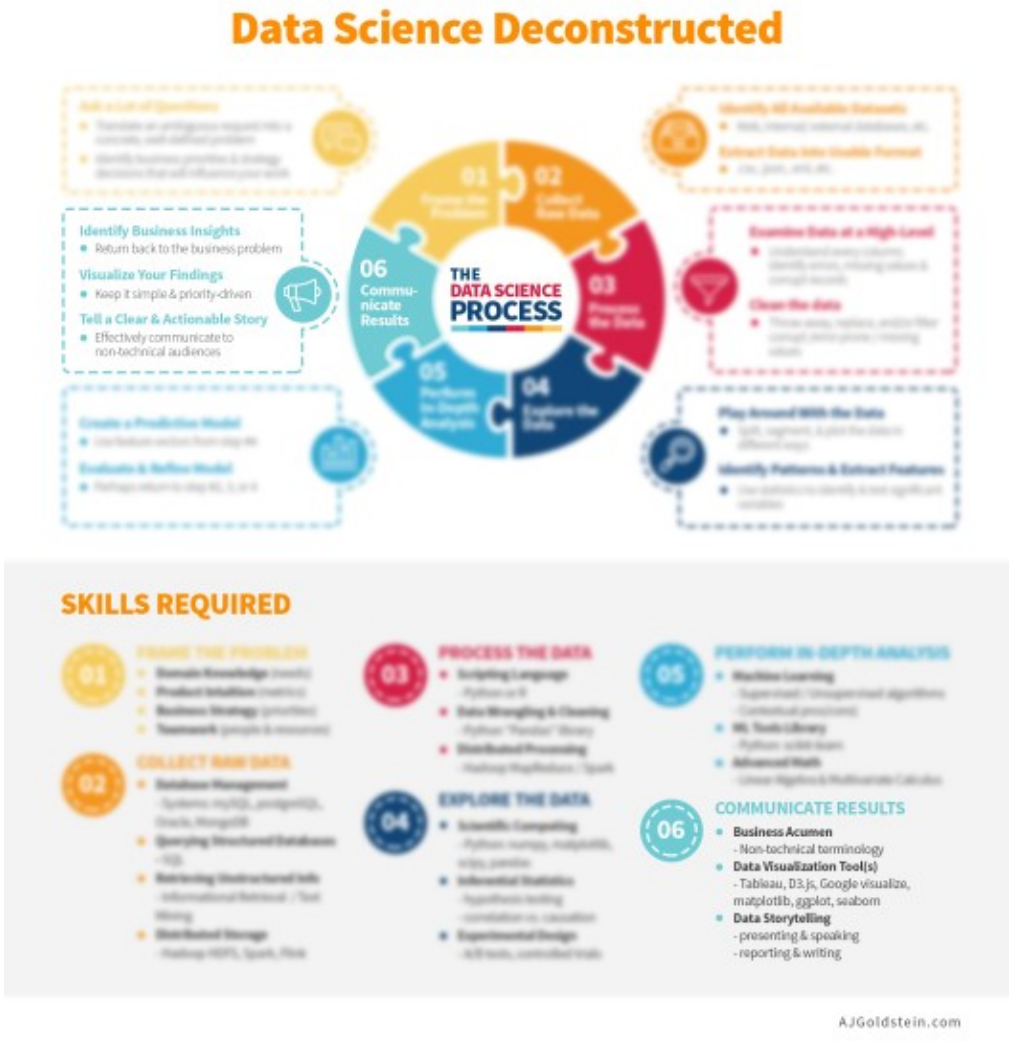
So what I ultimately built was a simple linear regression model (in Excel, no less) that used features like active user count, student enrollment, & university endowment to explain a university's user activity over time.

Sure, building a predictive model would've been the "cool" thing to do, but the goal wasn't to predict sales leads for the future; it was to establish partnerships with universities NOW.

Lesson learned: **the job of a Data Scientist is NOT to build a fancy model; it's to do whatever it takes to solve a real-world human**

**problem.**

## Step Six: Communicate Results




The sixth step of the Data Science Process is where you bring it all together and communicate results.

This is where you practice the most underrated skill in the Data Science toolbox; the X-factor that separates the good Data Scientists from the great ones: data storytelling.

Speaking with experts, I heard it time and time again: **your worth as a Data Scientist will be ultimately determined by your ability to convert insights into a clear and actionable story.**

In other words, the ability to create and present simple, effective

data visualizations to a non-technical audience is the most sought after skill in business today.

For a perfect example of how to do it right, here's the most well-put-together data story I've ever seen on "Wealth Inequality in America".

And here's a lecture by Harvard's CS109 that's a brilliant encapsulation of the art of data storytelling. The professor covers everything from understanding your audience to providing memorable examples. If you don't have time to watch the lecture, you can check out my Evernote notes that sum it all up.

Finally, to create beautiful data visualizations, I'd recommend going beyond Python's basic matplotlib library and checking out seaborn (statistical) and bokeh (interative).

**For example...**

With Calm College, we had to weave our findings on student activity into an actionable story for campus administrators.

First, I used our list of "Calm's Most Popular Colleges" to generate sales leads, by reaching out to 50 schools that the model identified as most suitable for a partnership.

Then, for each of the 50 schools, I crafted a personalized story about their students' activity on the Calm app.

For example, with Harvard, we reached out to the head of campus wellness to let her know that Harvard's campus was a top 5 most popular college for the Calm app. Then we included 4 graphs depicting the following insights:

1. 6% of the Cambridge, Massachusetts population (17,000+ people) are Calm users.
2. More than 82% of Harvard users are active on a monthly basis, with an average of 15 (fifteen!) sessions/month!
3. Week-by-week retention amongst Harvard users is 3x that of the

average Calm user.

4. Yet, despite all of this, Harvard student's subscription rate is still well below average.

The first 3 graphs told a story of extraordinary interest in the Calm app on Harvard's campus. But what really drove home our program was the last point:

*'despite all this amazing interest, it's clear that your students cannot afford Calm's $60/year subscription. That's why you need Calm College: to make the Calm app a FREE wellness resource for your students."*

**Rather than sell our product, we were selling their students' *past and present use* of our product.** And it worked like a charm.

Repeating this approach for other colleges, we were able to successfully get our foot-in-the-door at many of the most elite institutions in the country.

And eventually, thanks to this application of The Data Science Process, we were able to launch the program at 8 schools this Fall:


*the 8 schools Calm College launched at this Fall*

## Selection: The Core 20%

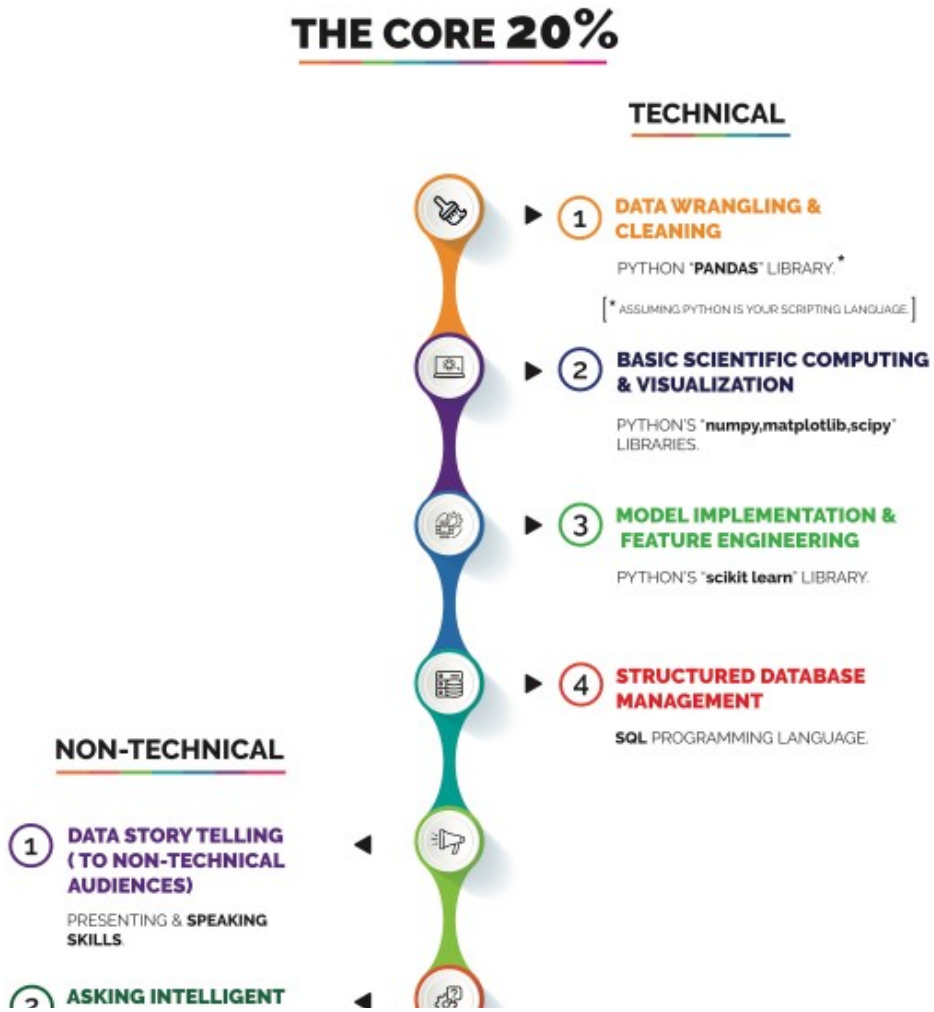
*'You are not flailing through a rainforest of information with a machete; you are a sniper with a single bull's-eye in the cross-hairs." — Tim Ferriss, The Four Hour Chef*

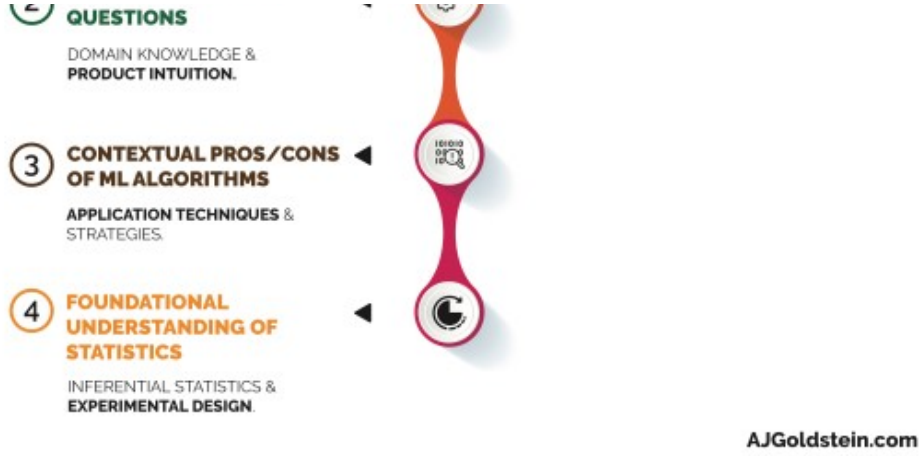
The greatest mistake you can make in accelerated learning is trying to master everything. **This is not Pokémon. You are not going to catch 'em all.**

Instead, the key is being relentlessly focused with the micro-skills you choose to develop. Through rigorous application of the 80/20 rule, it's possible to cut down a long list of possibilities to the highest frequency material. Then, once you've cleared your plate, it's depth over breadth all the way.

In his book, the "Four Hour Chef", Tim Ferriss discusses this selection process by introducing the idea of a "Minimum Effective Dose" (MED). Simply put, an MED is the smallest dose that will produce a desired outcome.

Here, I've broken down the MED for all 6 steps of The Data Science Process:






*the 20% of Data Science skills that result in 80% of outcomes*

In conversations with experts, these 8 skills continuously came up as the most essential.

In particular, Data Wrangling (i.e. Python's Pandas) was said to be the #1 skill (in terms of time spent doing) by every Data Scientist I spoke with. **Data cleaning is not sexy, but it encapsulates up to 80% of the job**.

You may be wondering where big data tools like Hadoop & Spark, or modeling techniques like neural networks & deep learning fall into all this. The answer: surely outside the core 20%.

To my surprise, many Data Scientists I spoke with emphasized that only a small percentage of companies have data that even requires something as complex as a neural network!

Instead, an overwhelming majority of employers need more simple services like data cleaning, exploratory analysis, and logistic regression models (as recently reflected in an industry-wide survey by Kaggle).

When choosing what to learn, remember: you can always revisit the heavier topics later, but don't weigh yourself down at the start. The goal is to *accelerate* learning. So wait until your house of expertise has a strong foundation before adding the shiny stuff.

If you're looking to master the fundamentals of Data Science in 6 months or less, you'll want to simply focus on the core 20%.

## Next Steps

*'Live as if you were to die tomorrow. Learn as if you were to live forever." — Mahatma Gandhi*

I do not believe knowledge is useful for the sake of knowledge; only if you use what you've learned to improve your life, or the lives of others. So I would encourage you to pause, reflect, & ask yourself: **"what's the *smallest* possible action I can take *right now* with what I've learned?".**

For instance, a great place to start would be picking one of the six steps you're most interested in and exploring the skills/resources associated with it. Then find a dataset that's of interest to you and start *learning by doing* through a mini-side-project.

The key is trusting yourself by following the path that you're instinctually most drawn to... because that's where you're find the most short-term motivation & long-term fulfillment.

Personally, after deconstructing data science and identifying the core 20%, I decided to enroll in Springboard's Data Science Intensive online bootcamp (recently renamed to "Intermediate Data Science"). I chose this program because it was **the only curriculum I could find that covered all 6 steps of the data science process while focusing in on all 8 skills of the core 20%**.

For more information on the program, I'd recommend checking out Raj Bandyopadhyay's brilliant Quora answers (here and here) on the methodology behind Springboard's approach to Data Science education. And **here's a discount code for $100 off any Springboard course.**

Whatever you choose to do with this information, the important thing is that you *do* something. Getting started is always the hardest part, so I challenge you to turn *intention* into *action*.

## Final Thoughts

Over the past few weeks, the power of the internet has sure become apparent. In just the first 7 days, my first post — Learning Without Limits — had 3000+ views from 66 countries around the world. Never did I expect it to spread so far and wide, but I guess I have all of you to thank for that.

**So as long as you all continue to pay it forward, I'll continue to be an open book. As promised, I've complied and will continue to open-source all my favorite resources, insights, and findings via this new page: ajgoldstein.com/resources.**

All I ask of you is that you share this with people you think would benefit. That's my call-to-action. Share. Why? Because we're all in this together and true happiness comes from other people.

---

***To follow along this journey, feel free to drop your email in the sign-up bar below.*** *By signing up, you'll receive one (just one) email when I've posted a new update.*

*And don't hesitate to leave any questions, thoughts, or feedback you have in the comments box below. I'd love to hear from you.*