A Relationship Guide

Datapalooza, May 2023

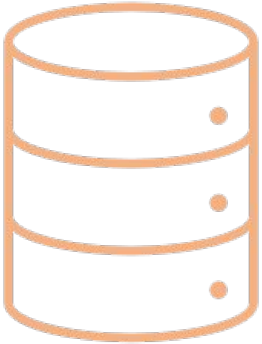
Carolyn Silverman

DISCLAIMER

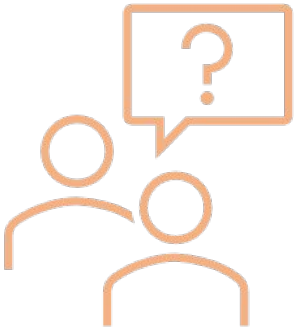
THE FOLLOWING PRESENTATION CONTAINS EVIDENCE-BASED TIPS WHEN APPLIED TO RELATIONSHIPS WITH ***DATA***. APPLY THESE TIPS TO RELATIONSHIPS WITH ***PEOPLE*** AT YOUR OWN RISK.

Motivation

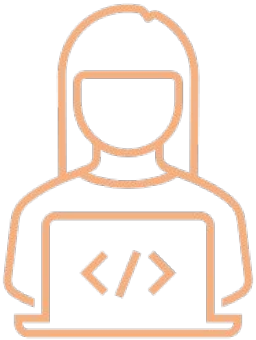
Setup



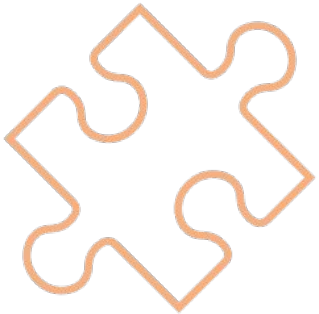
Huge data extract comes in



Immediate questions from research team

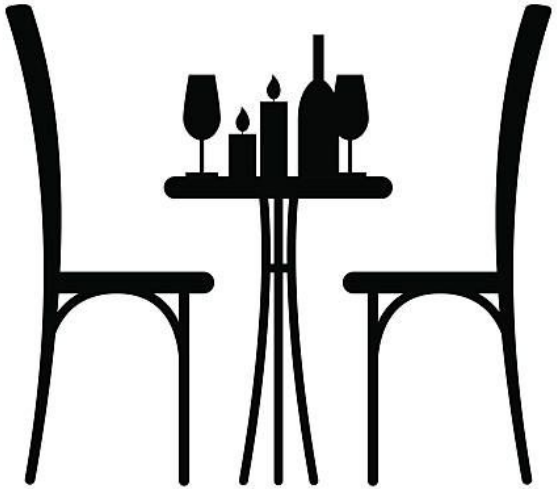


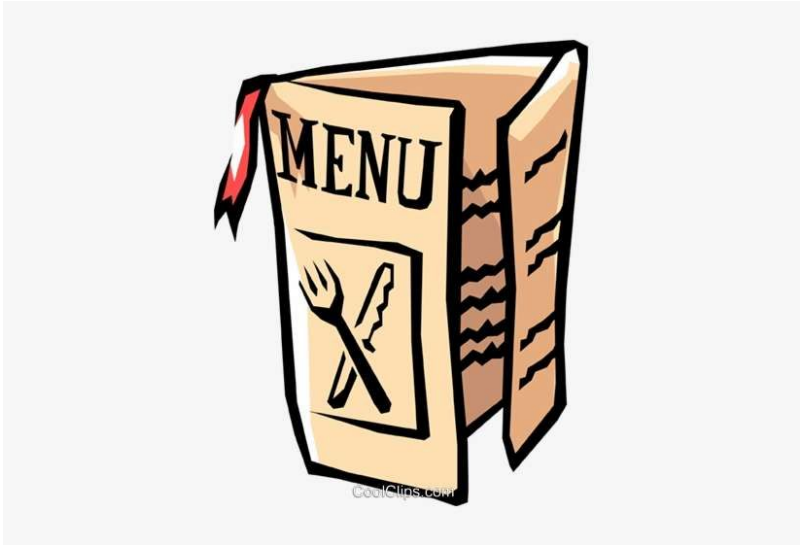
Race to produce answers quickly



Overlook important aspects of the data

A Rushed Introduction to Your Data



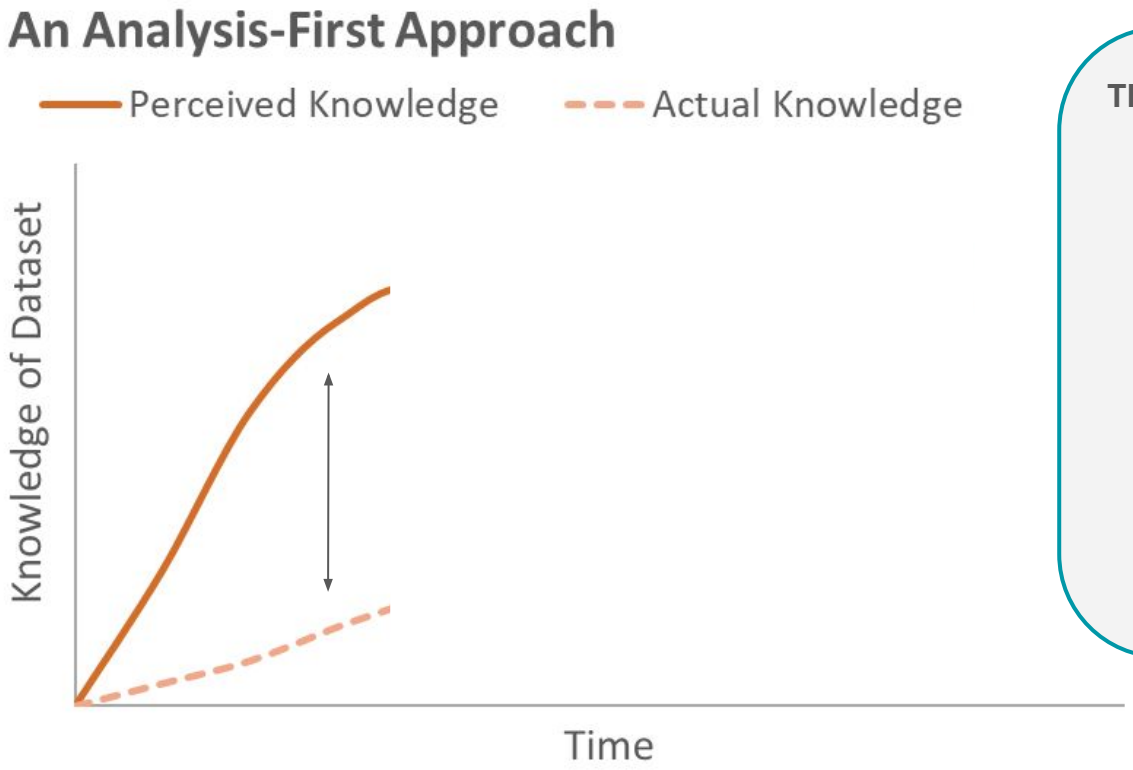


A Rushed Introduction to Your Data

Trying to answer questions about your data before a proper introduction is like going on a first dinner date and ordering for the other person

- You can make educated guesses based on context clues and prior information (*nature and animal lover = vegetarian?*)
- But there are many unchecked assumptions that may not be true (*your date was really craving a burger!*)
- Don’t disrespect your date (or your data) by making assumptions!

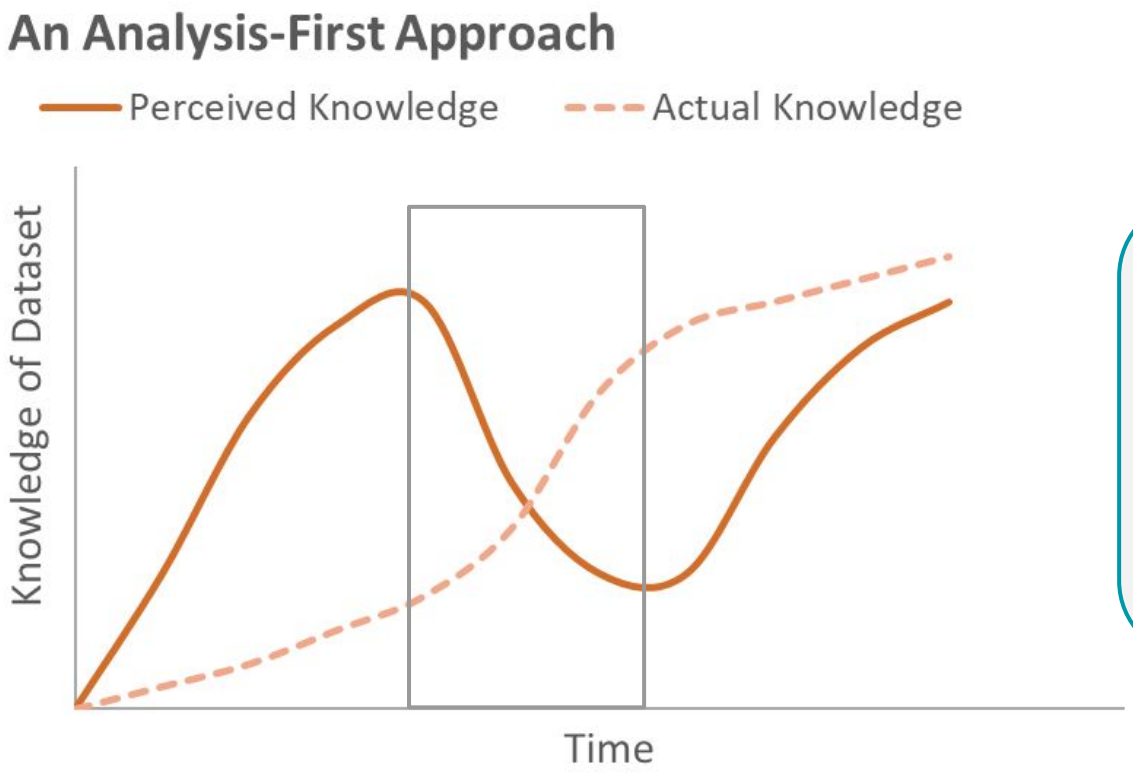
A Recipe for a Toxic Relationship



The Overconfidence Phase

- Thinking you deeply understand your dataset after a very cursory introduction
- Often involves **unchecked assumptions**
- Large gap between perceived and actual knowledge of the data

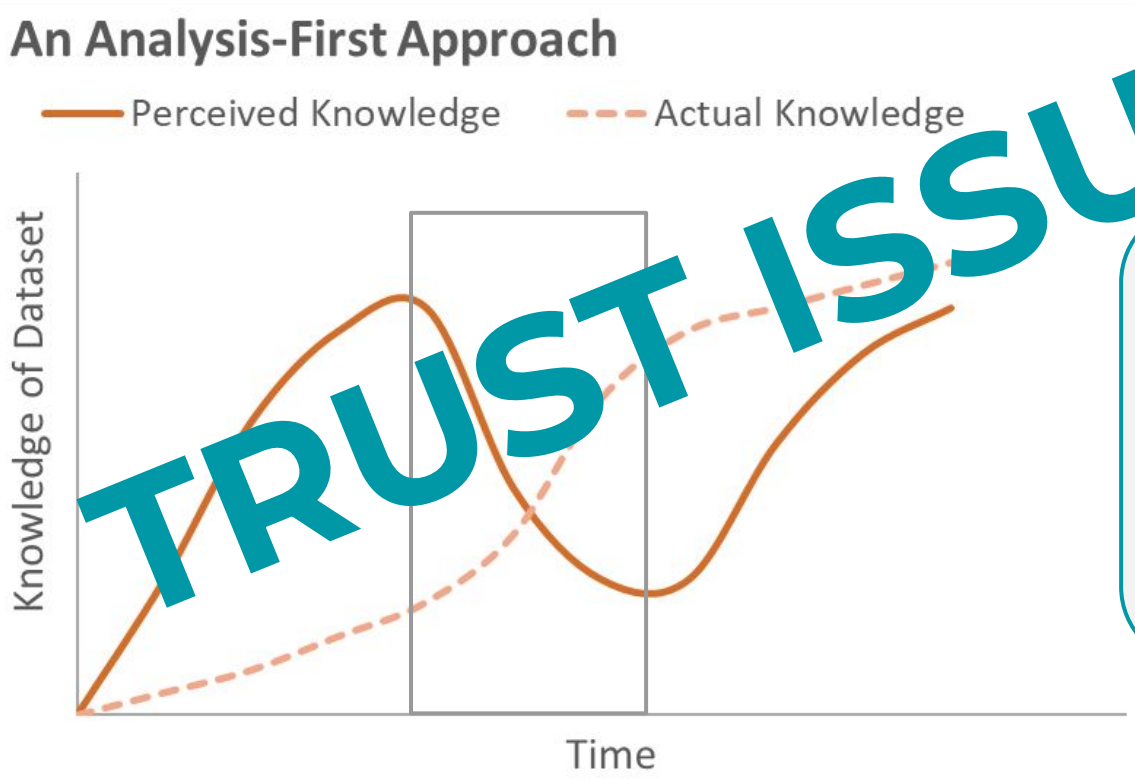
A Recipe for a Toxic Relationship



The Deception Phase

Only after working with your data for some time do you realize that some of your assumptions were incorrect. Many of your analysis results are invalidated. You lose trust in your data and yourself.

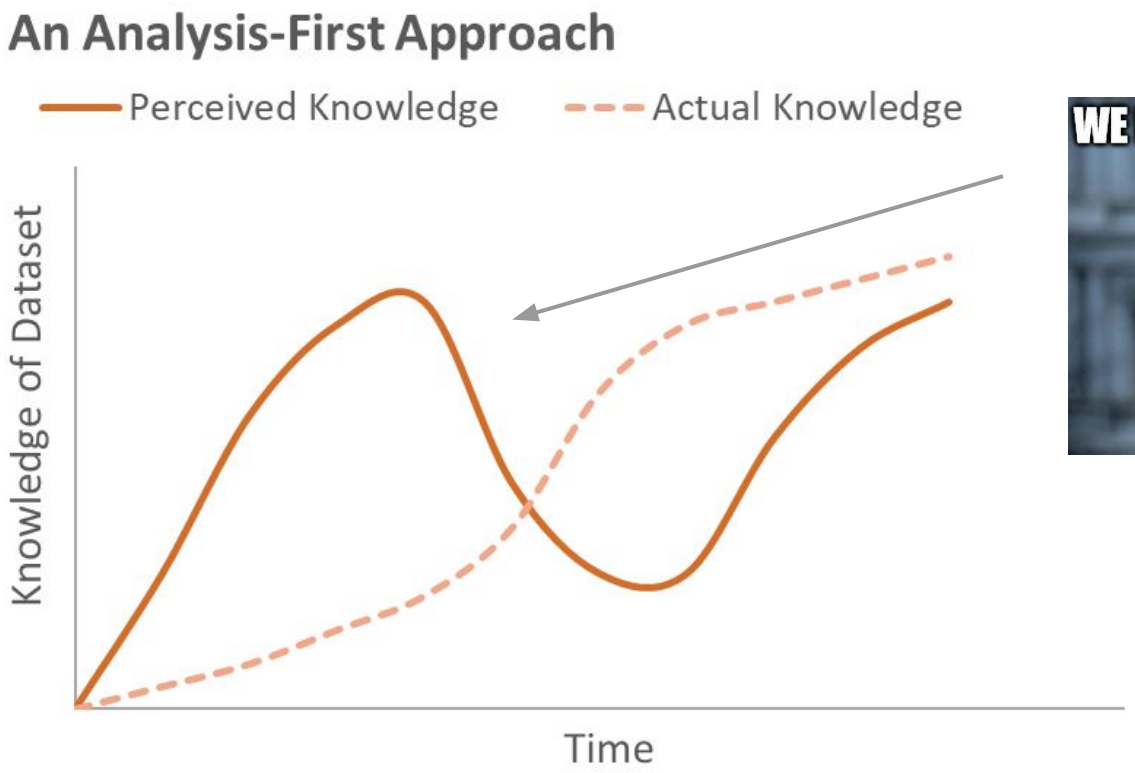
A Recipe for a Toxic Relationship



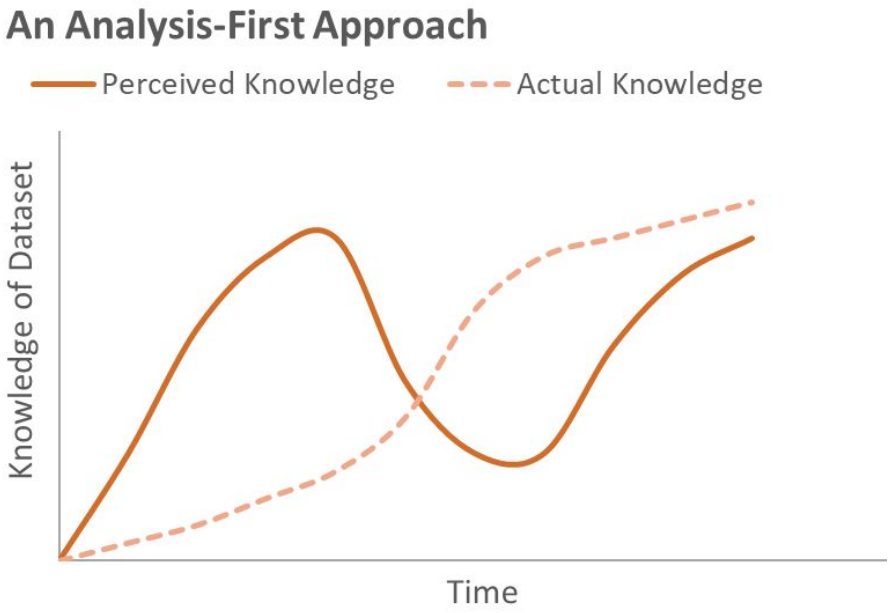
The Deception Phase

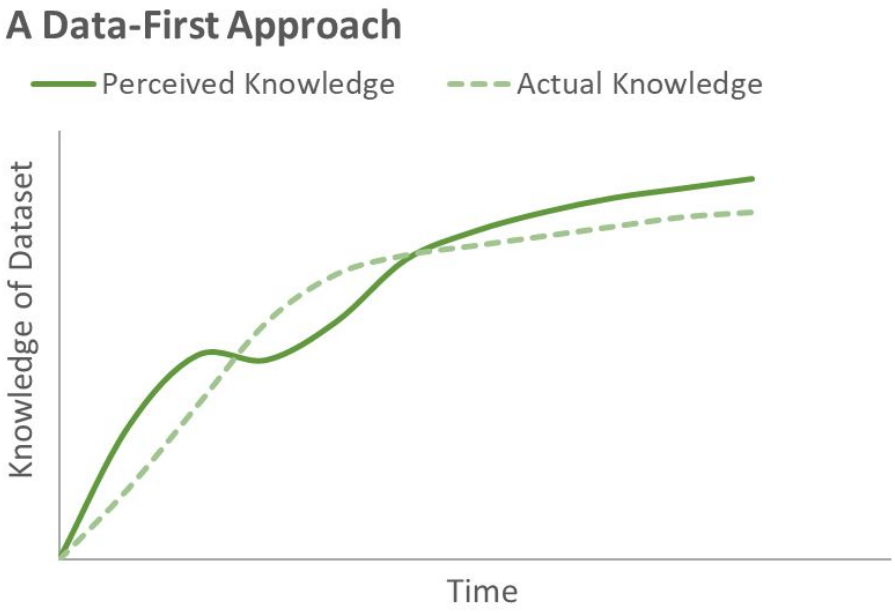
Only after working with your data for some time do you realize that some of your assumptions were incorrect. Many of your analysis results are invalidated. You lose trust in your data and yourself.

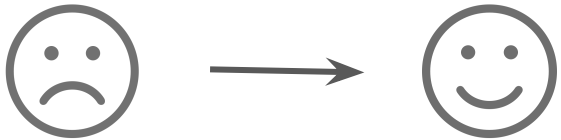
A Recipe for a Toxic Relationship



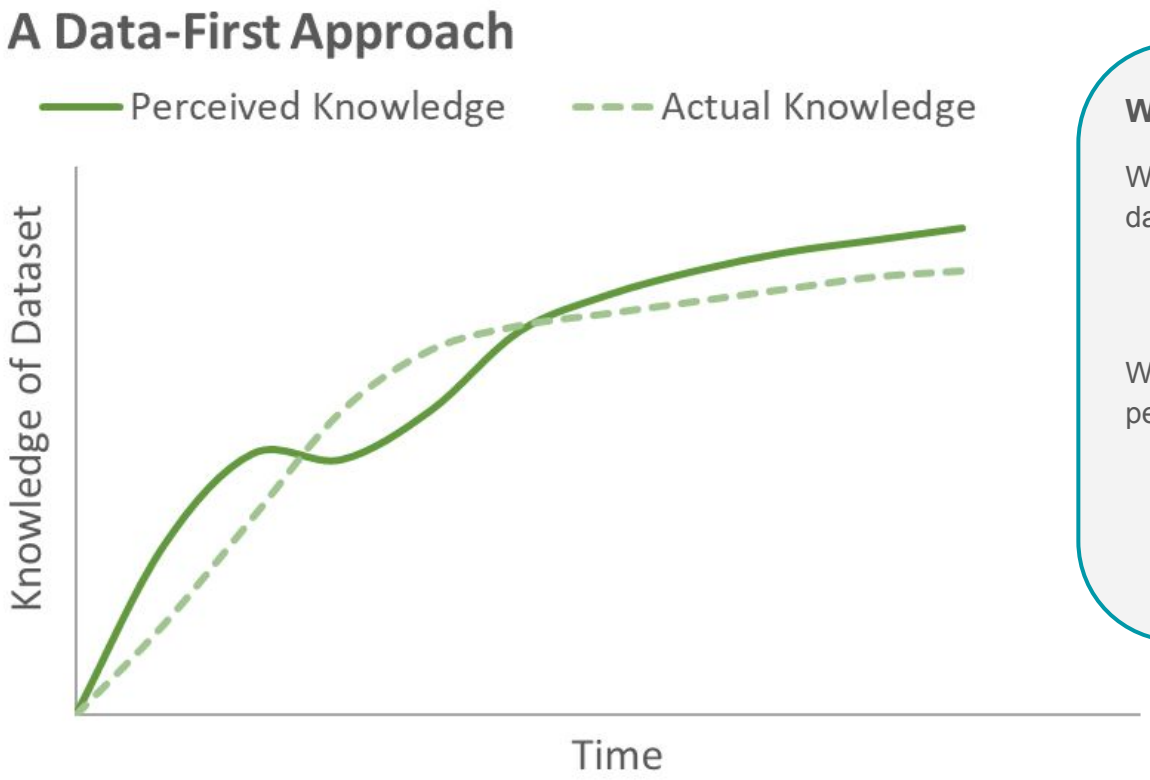
A Recipe for a ~~Toxic~~ *Healthy* Relationship







A Recipe for a Healthy Relationship



Why this approach?

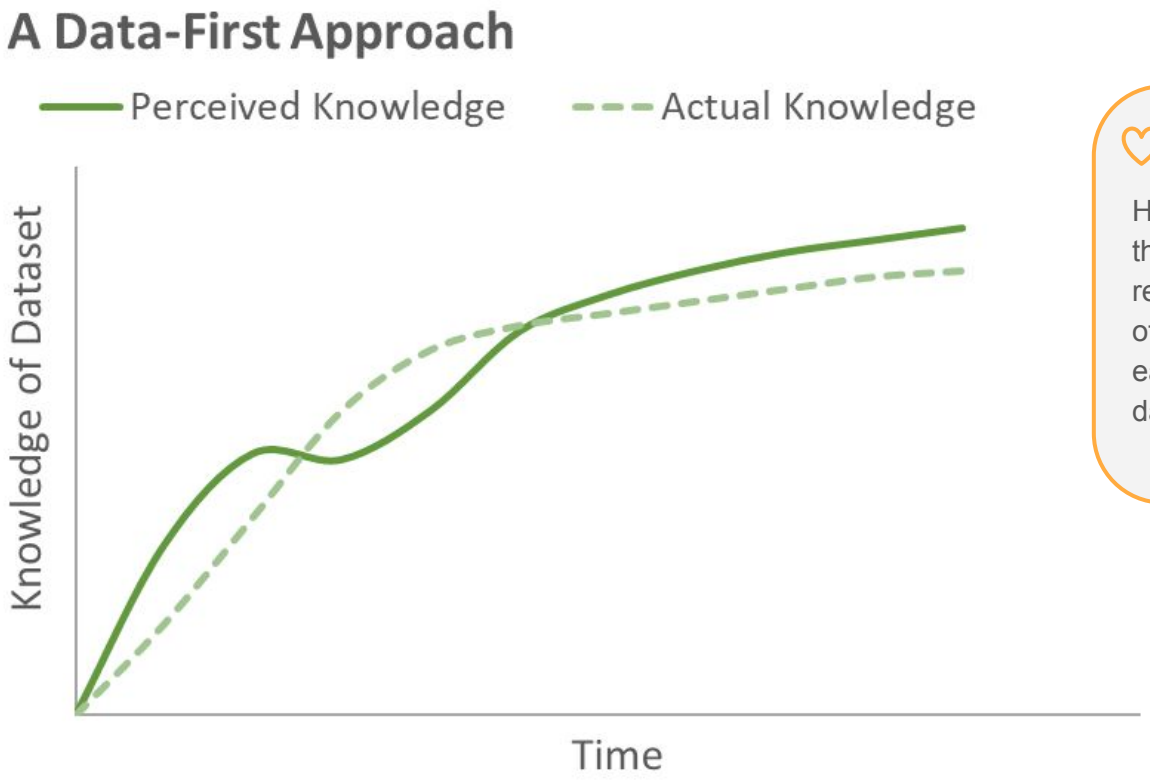
We **frontload** the process of building dataset knowledge

- This reduces the chance of being blindsided later

We minimize the distance between perceived and actual dataset knowledge

- This allows us to put more accurate **confidence bounds** around any early stats that we produce

A Recipe for a Healthy Relationship



It's All About the Foundation

Having a successful first few dates can set the foundation for a good long-term relationship. For this reason, the majority of this presentation will be focused on the early stages of your relationship with your data.

The First Date

Ask Good Questions & Embrace Structure

- Ask a **breadth** of questions instead of going into depth
- Focus early questions on what is important to **you**
  - Define clear **constraints/criteria** upfront: what are must haves vs nice to haves?
- **Structure** helps us avoid the common pitfall of going down rabbit holes too early
  - It doesn’t matter if we can answer a very niche question if we can’t answer something fundamental to the analysis

Ask Good Questions & Embrace Structure

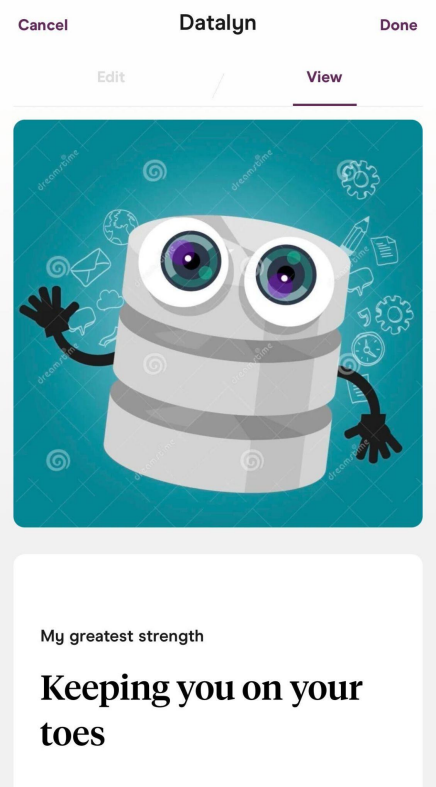
My favorite flavor of questions:

- **What makes you unique?**
  - i.e. what defines a unique row in the dataset?
- **What is your history?**
  - i.e. what is the data generating process?

Document your findings!

- Documentation keeps all knowledge easily accessible and helps you avoid mixing up any information between datasets (rude!)

Take Stock: Do Expectations Meet Reality?



Cross-reference the data with any prior information you have (DSAs, codebooks, etc)

- Is the expected population in the data?
- Are the expected variables in the data?
- Do the variables represent what they are supposed to capture?

Internet stalking your data is encouraged!

The Second Date

Take Your Date(a) for a Spin

You've established a certain comfort level with your data on date #1. Now it's time to let loose and test out some questions that you've been eager to ask

- Start with simple summary statistics and see if they match our intuition

Look out for any 🚩🚩🚩

- Is your dataset giving you **consistent answers**?
- NJ Courts example: counting the number of arrests per year in NJ using two different sources gave very different results

Take Your Date(a) for a Spin

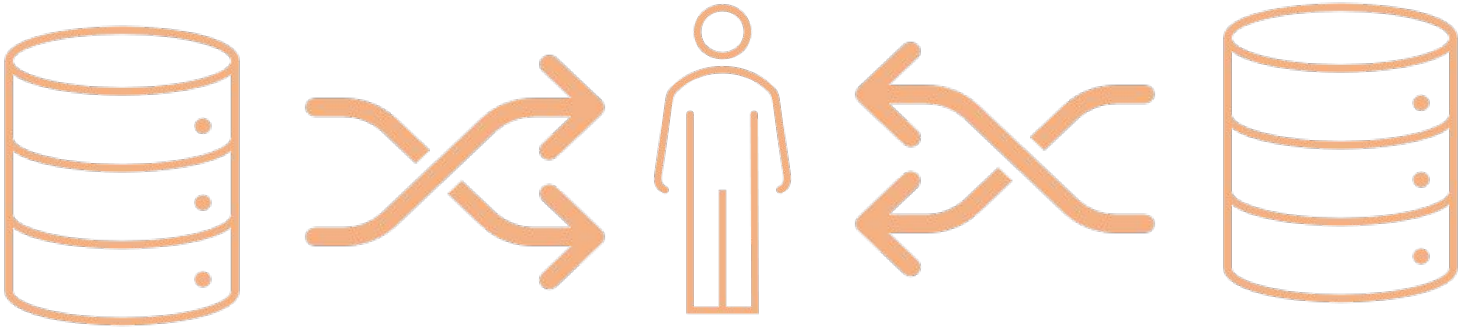
Remember that this is a **trial period**:

- Proceed with a healthy dose of skepticism but allow yourself to have fun
- You are not committed to anything yet!
  - Make sure any early results that you share are appropriately caveated

Dates 3-5(ish)

Get Specific

- One great way to get to know a date is by listening anecdotes about their past experiences
- Now it’s time to dig into the specific stories of your data
  - Sample all records for a unique entity
  - Can you put together a cohesive narrative of that entity across many data sources?

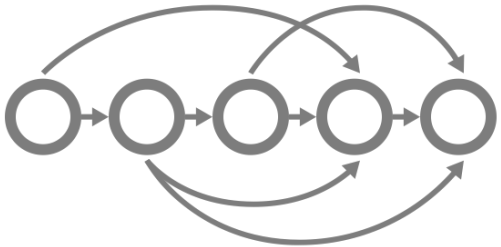


Embrace Curiosity

- The initial dates were focused on what you wanted to get out of your data
- After the data passes those tests, take off your blinders and see what else the data has to offer
- Who knows what unexpected gems may be hidden

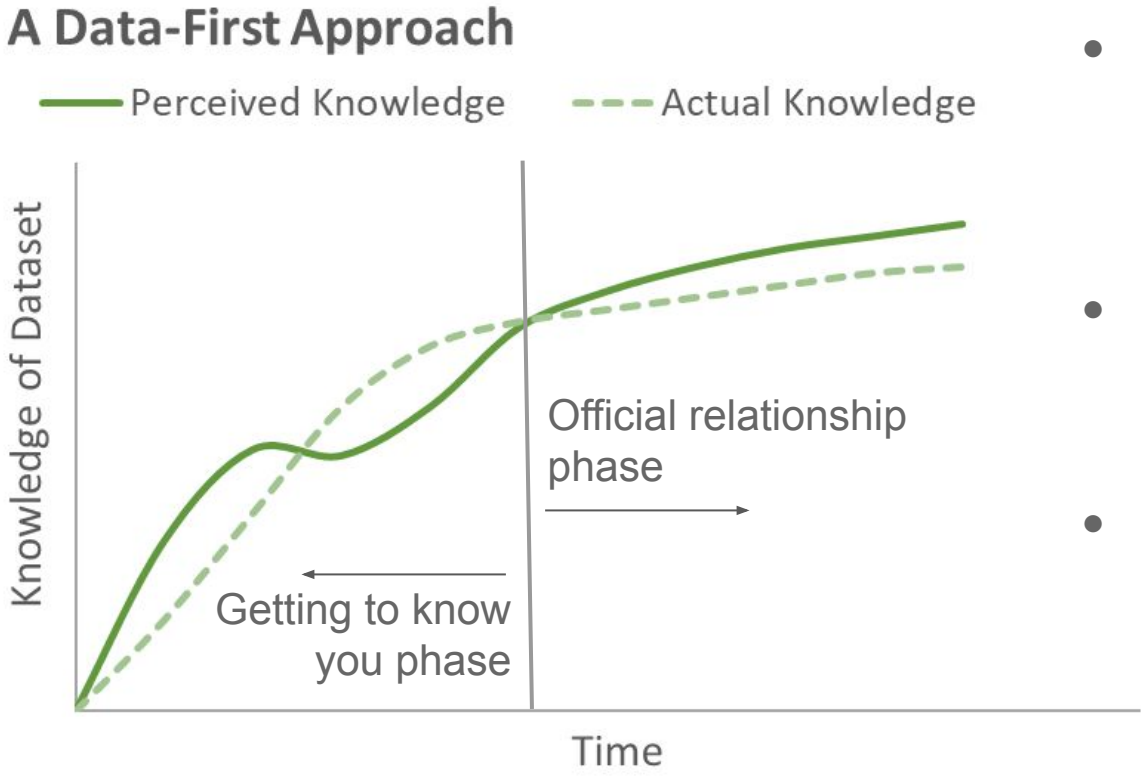
Set Yourself Up for a Sustainable Future

- Build a **pipeline** so you don't have to start from scratch every time you meet up with your data; instead, pick up from where you left off
- Your knowledge of the data is constantly growing and evolving. Try to bake this knowledge into the dataset itself, e.g.:
  - Use clear column naming conventions
  - Delete (or have some indication of) bad/untrustworthy measures in the final analysis file



Your LTR

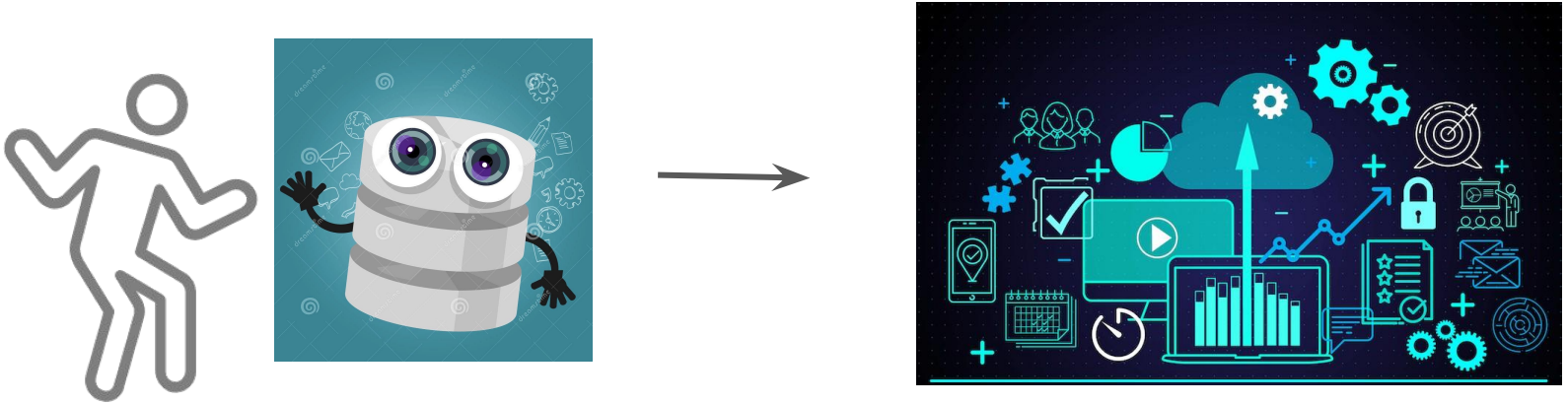
Official Relationship Status



- The foundation is set and you're ready to try new things together (e.g. take a ride through a new heterogeneous treatment effect model)
- Stay humble as your data can continue to introduce twists, even in later relationship stages
- When asking a new question, don't be afraid to go back to the basics

Takeaways

Embrace a Data-First Approach



Prioritize a good foundational relationship with your data. Then you can be more confident in your base as you trek into the turbulent territory of more complex modeling and analysis tasks.

Testimonials from Data Relationship Gurus

"I have never regretted spending more time looking at the data."

– *Greg Stoddard*

"On average, people should be more skeptical when they see numbers. They should be more willing to play around with the data themselves."

– *Nate Silver*

"Data scientists not only are adept at working with data, but appreciate data itself as a first-class product."

– *Hilary Mason*