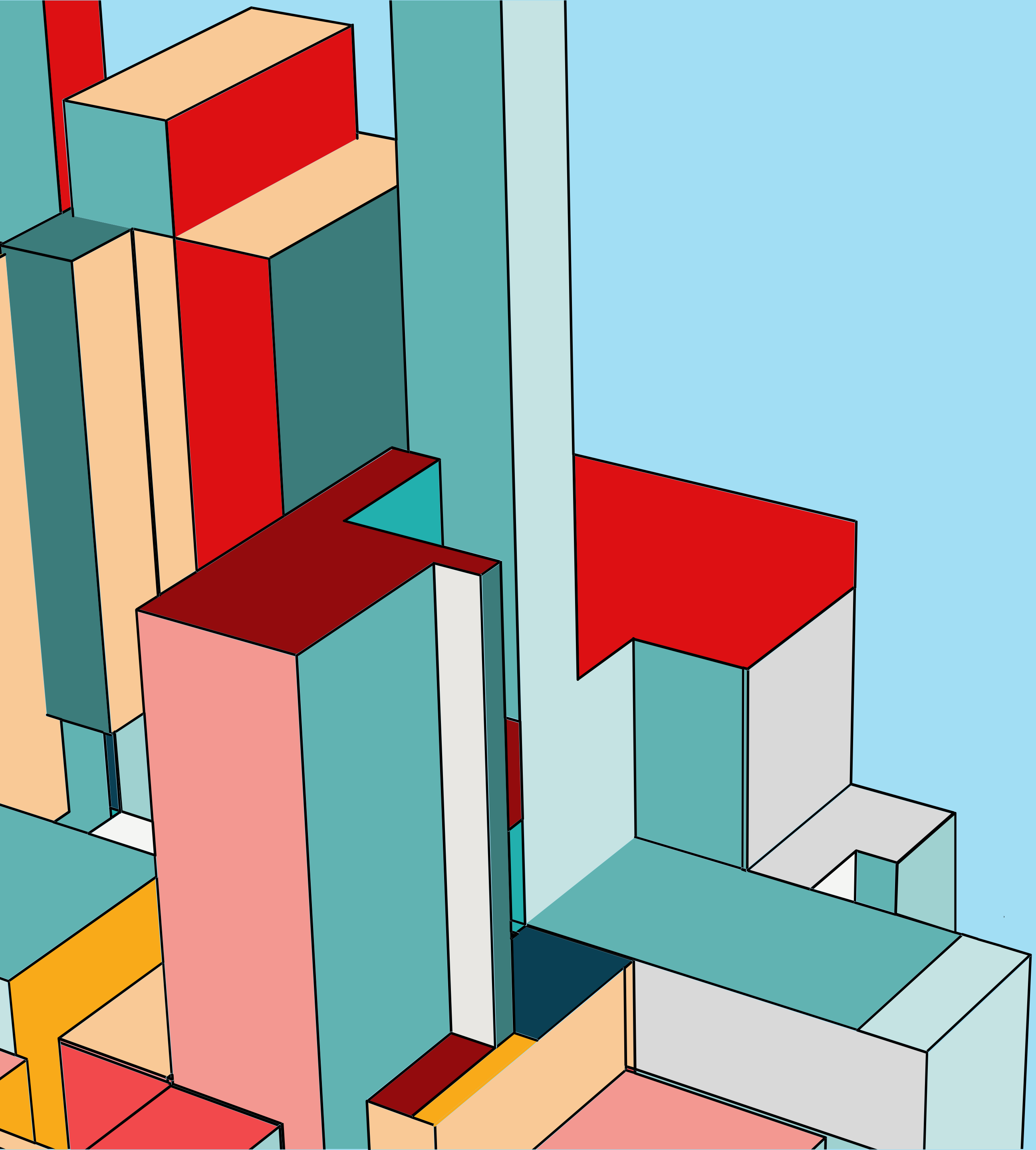


BUILDING A DATA SCIENCE TOOLCHAIN

(WITH R OR PYTHON)

RYAN ORSINGER
TECH DAY 2024

JUST USE SQL

AGENDA

Introduction

Mental Tools

Technical Tools

Selected Readings & Resources

Next steps (Social Tools)

ABOUT RYAN

Haven for Hope®

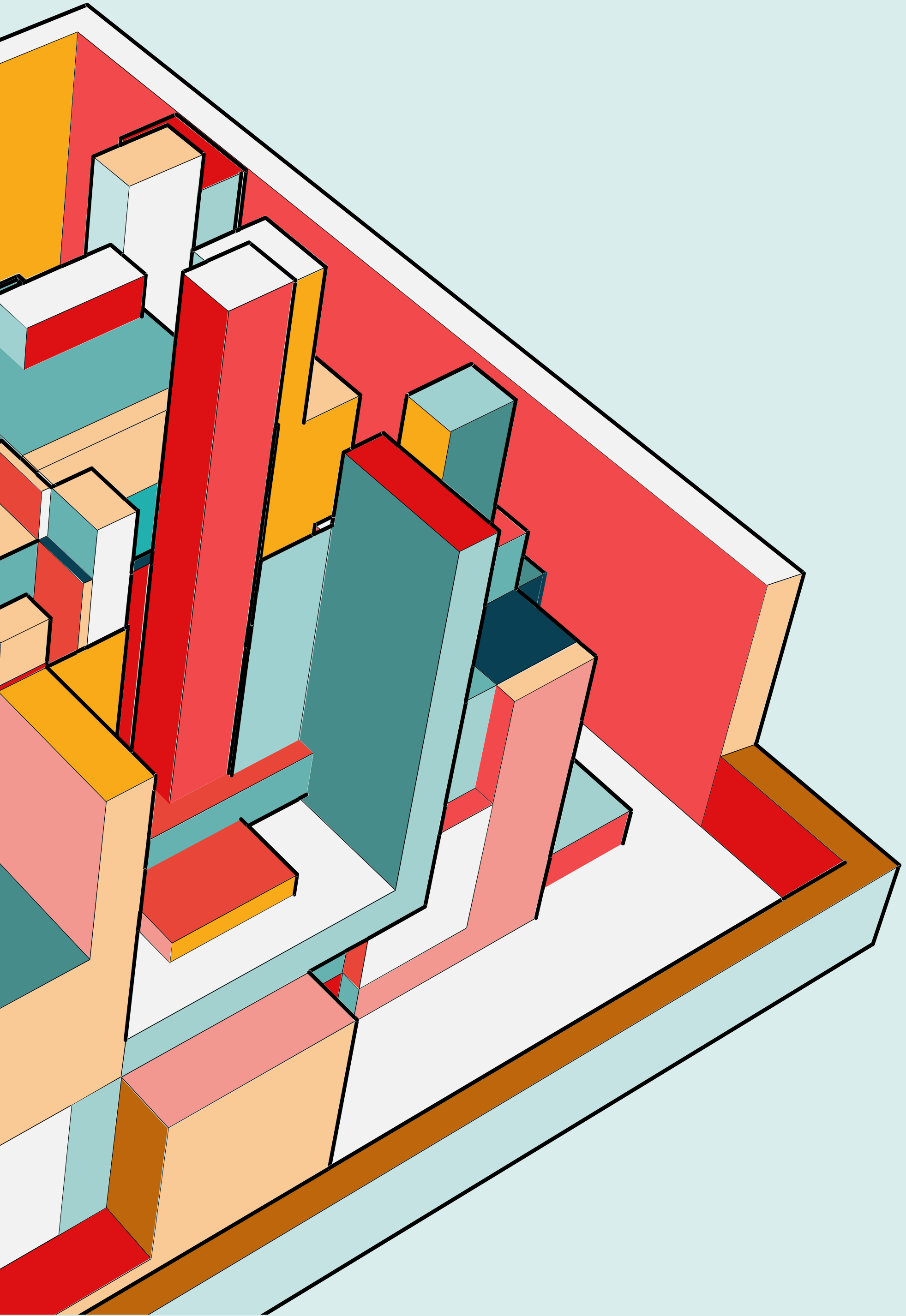
Alamo Regional Data Alliance

codeup

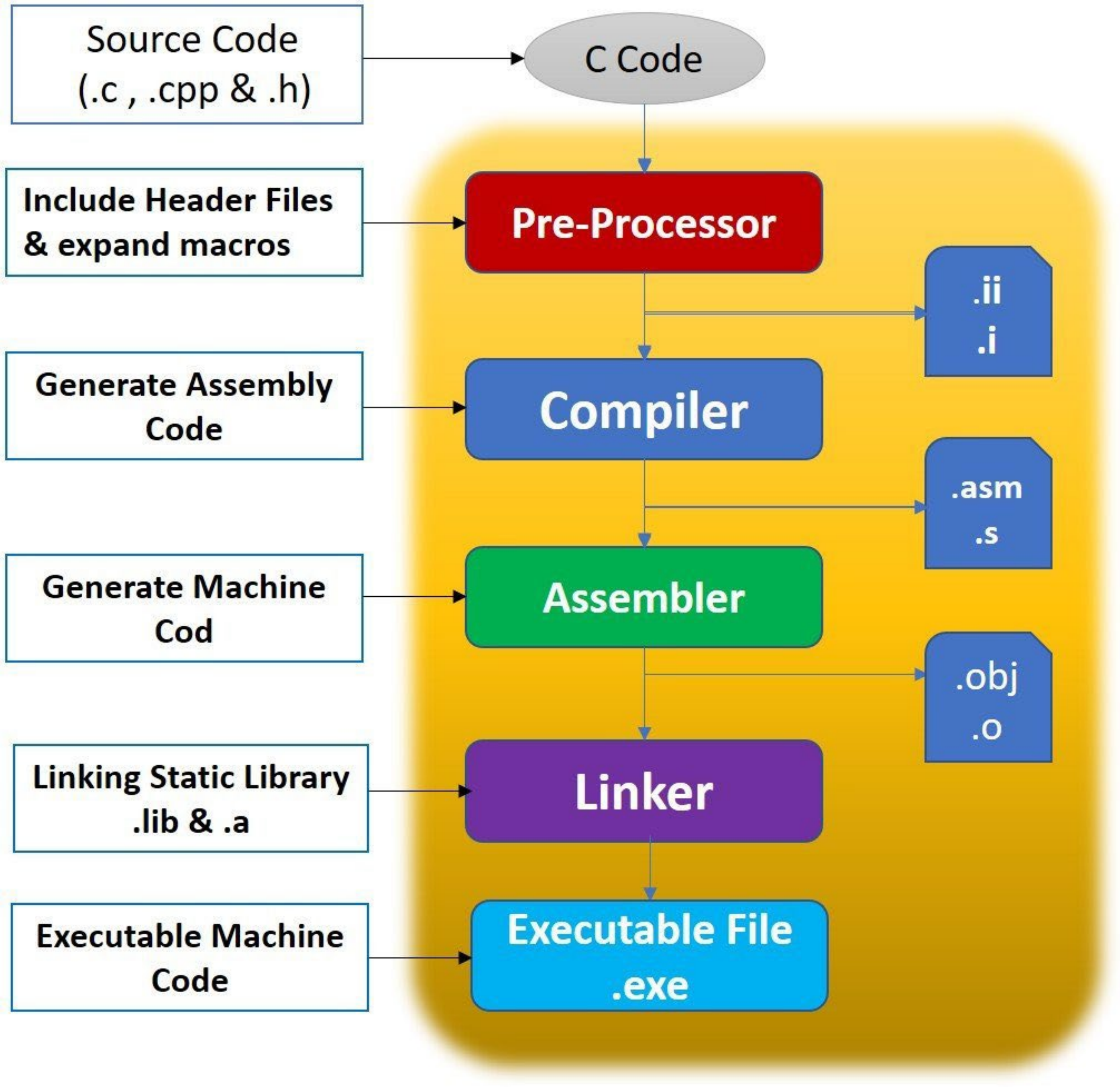
ANACONDA®

KING WILLIAM FAIR

TRINITY UNIVERSITY



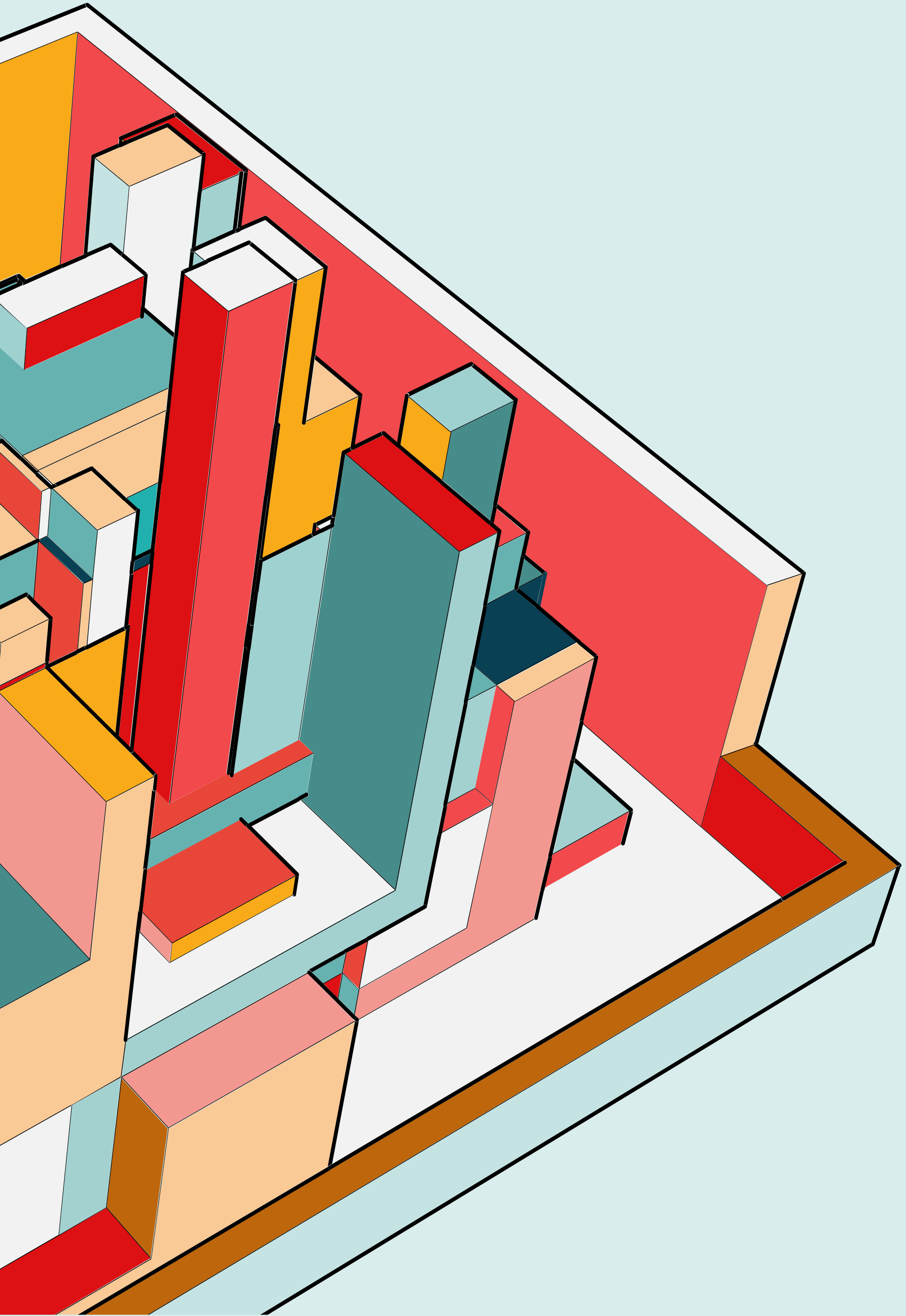
WHAT IS A TOOLCHAIN?


Source Code
(.c , .cpp & .h)
C Code
Include Header Files
& expand macros
Pre-Processor
.ii
.i
Generate Assembly
Code
Compiler
.asm
.s
Generate Machine
Cod
Assembler
.obj
.o
Linking Static Library
.lib & .a
Linker
Executable Machine
Code
Executable File
.exe

YOUR DATA SCIENCE TOOLCHAIN IS NOT A PROGRAMMING LANGUAGE
OR
A SET OF LIBRARIES.

YOUR "DATA SCIENCE TOOLCHAIN" IS A WEB, A NETWORK OF:

- Knowing Your Business Domain
- Mental Models and Habits
- Your Social Ecology
- Feedback & Collaboration
- Technical Tools & Skills
- (Imagine a web with recursion, rather than a linear pipeline)

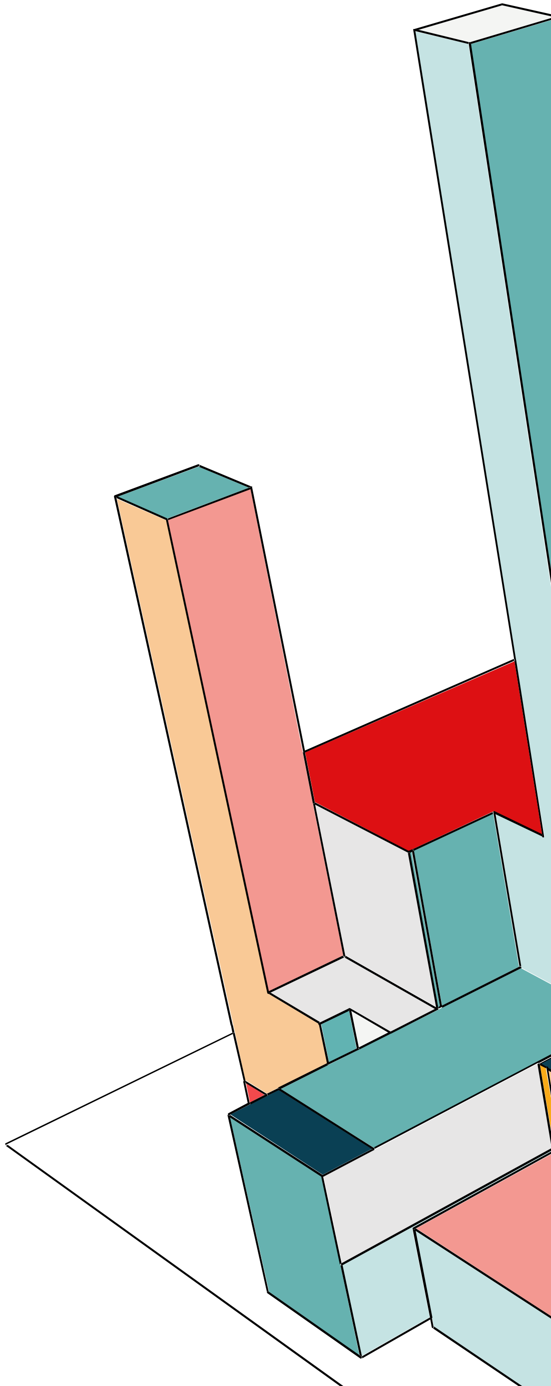


1ST TOOL IS TO KNOW YOUR BUSINESS.

YOU MAY KNOW DATA.
BUT YOU DON'T KNOW YOUR BUSINESS'S
DATA UNLESS YOU KNOW YOUR BUSINESS.

KNOW YOUR BUSINESS DOMAIN

- Learn how your business works. There is no substitute.
- Knowing the biz means people will take your recommendations seriously. People can tell if you're an egghead, even if you mean well.
- Interviews > Shadowing > Doing the work of front-line staff
- Work with front-line staff to discover real data quality issues and how your work can help them and the org.
- Connecting with people is great fuel for your relationships, creativity, and problem solving! (#SocialEcology)
- Part of your unique value proposition and problem-solving process.



CONNECT YOUR ACTIVITIES TO THE LOGIC MODEL OF THE BUSINESS

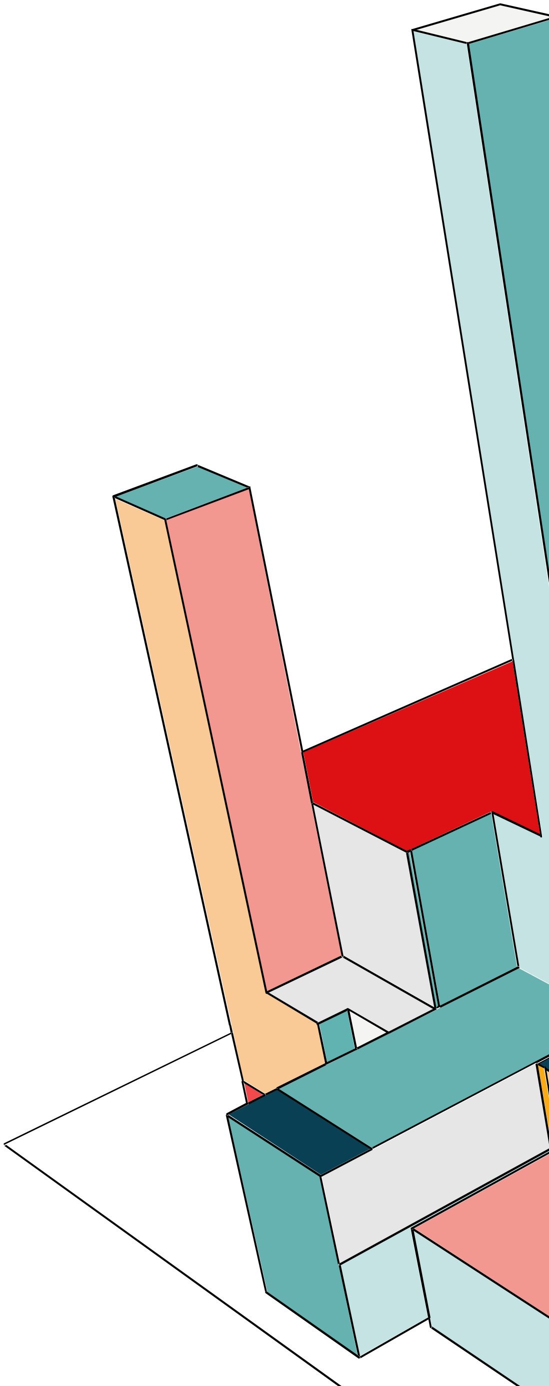
- Knowing how the business *actually* works (on the ground) and how it *should* work opens the door to:
  - Creativity
  - Better problem solving
  - More tangible impact (Know the people you're helping and how your work helps)
  - Opportunities, Leadership, etc…
- Can your role be done remotely by someone crunching purely technical tickets with no domain knowledge? Say hello to offshoring/LLMs/race to the bottom!

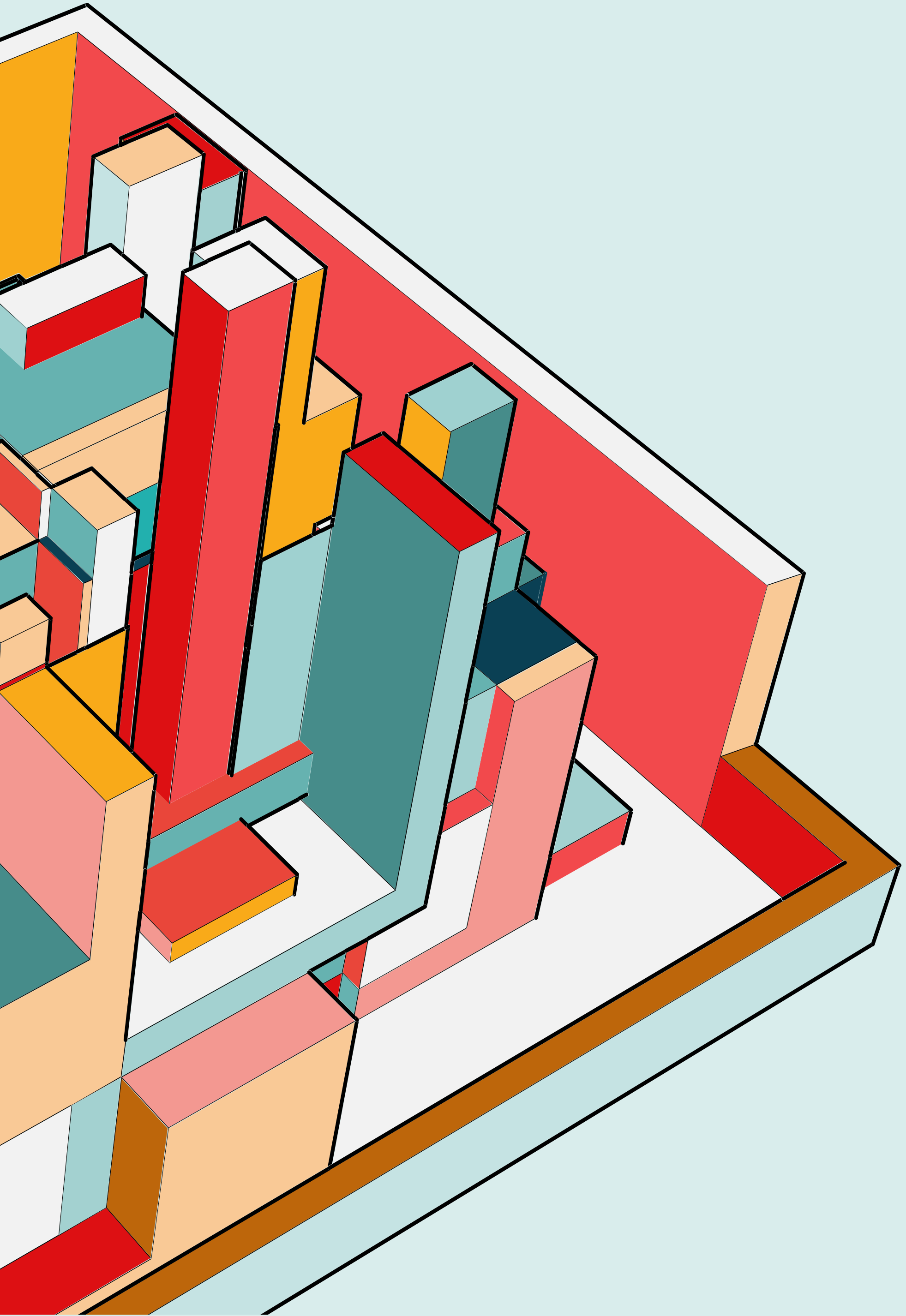
Your work must connect to the Logic Model of how the organization works.

Your analyses must enable the Business Model of your organization.

HELPFUL HABITS FOR YOUR TOOL-WEB

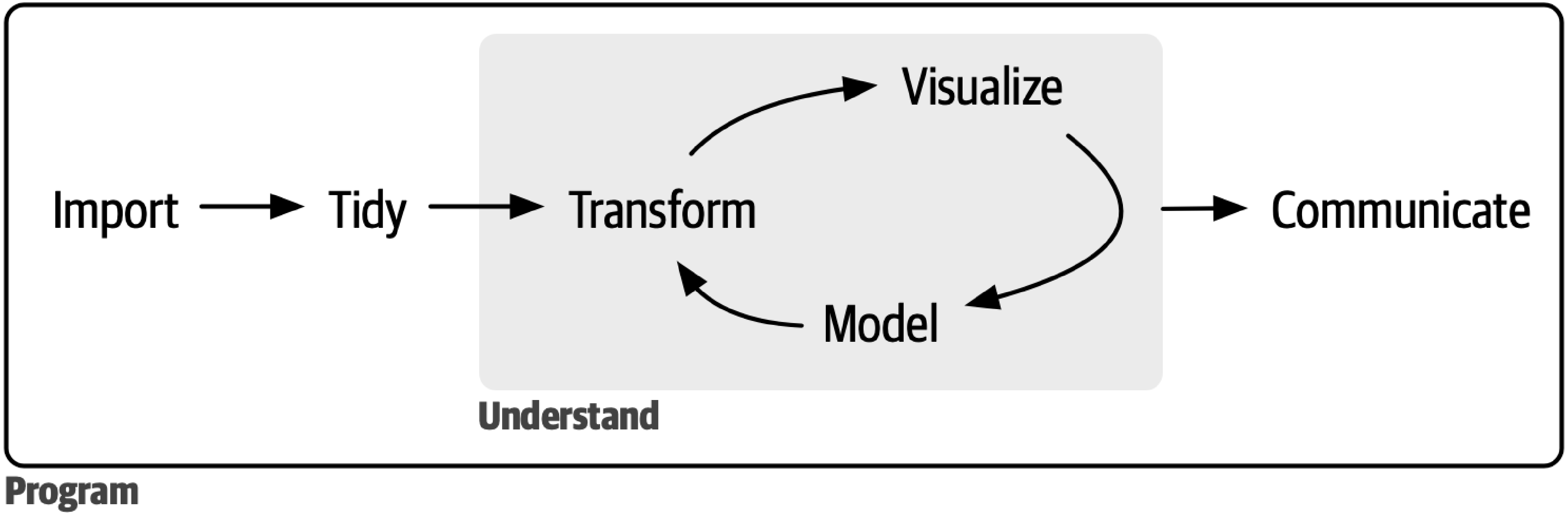
- Rush to a *Minimum* Viable Product (MVP)
  - Rapidly produce low-fidelity prototypes. Done > Perfect. Perfect is the opposite of done. Perfection is asymptotic.
  - Reduces cost of failure. Increases likelihood of success. Easier to pivot.
- Iterate, Seek feedback, Repeat
- Present before you're ready (with a safe environment, when possible)
- When you feel stuck, talk things through!
- Always double check your understanding of things with folks who know the business deeply
- Focus on data definitions, data governance, and the bigger picture.





TECHNICAL TOOLS!

DATA SCIENCE PROCESS IN CODE (WICKHAM)



PYTHON OR R? YES.

Python for data science

- Recommend if you already know a little Python or a “curly brace” language
- Install Anaconda. This will install Python, too. Batteries included for Mac/Windows/Linux.
- Can run in PowerBI, Tableau, and Excel
- Use conda environments if you want/need isolated environments/sandboxes

R for data science

- Recommend if you’re flying solo or starting from scratch
- Install R, Rstudio, and Tidyverse
- Can run in PowerBI and Tableau
- Use renv if you want isolated environments/sandboxes
- Community can be more beginner friendly

A NOTE ON PYTHON “VS.” R

- Use whatever your business already uses. Go with the grain.
- If you’re starting from scratch, R code/textbooks get you to the data science faster and with less overhead.
- R is built for statistical computing. Historically, Python has trailed R’s development.
- It can be *very* helpful to learn something you already know in another programming language or paradigm.
- More Python jobs, higher median salaries for Python, but much more competition. R could be a way to “backdoor” some applications.
- What matters most for skill-building is building TONS of projects.
- Real world projects > Toy Projects > Practice > Study

WHAT ABOUT DEPLOYING YOUR DATA PRODUCTS TO THE CLOUD?

Need to deploy your dashboards, APIs, and models?

If you’re using Python, recommend Anaconda’s suite of tools.

If you’re using R, recommend using Posit (formerly RStudio)

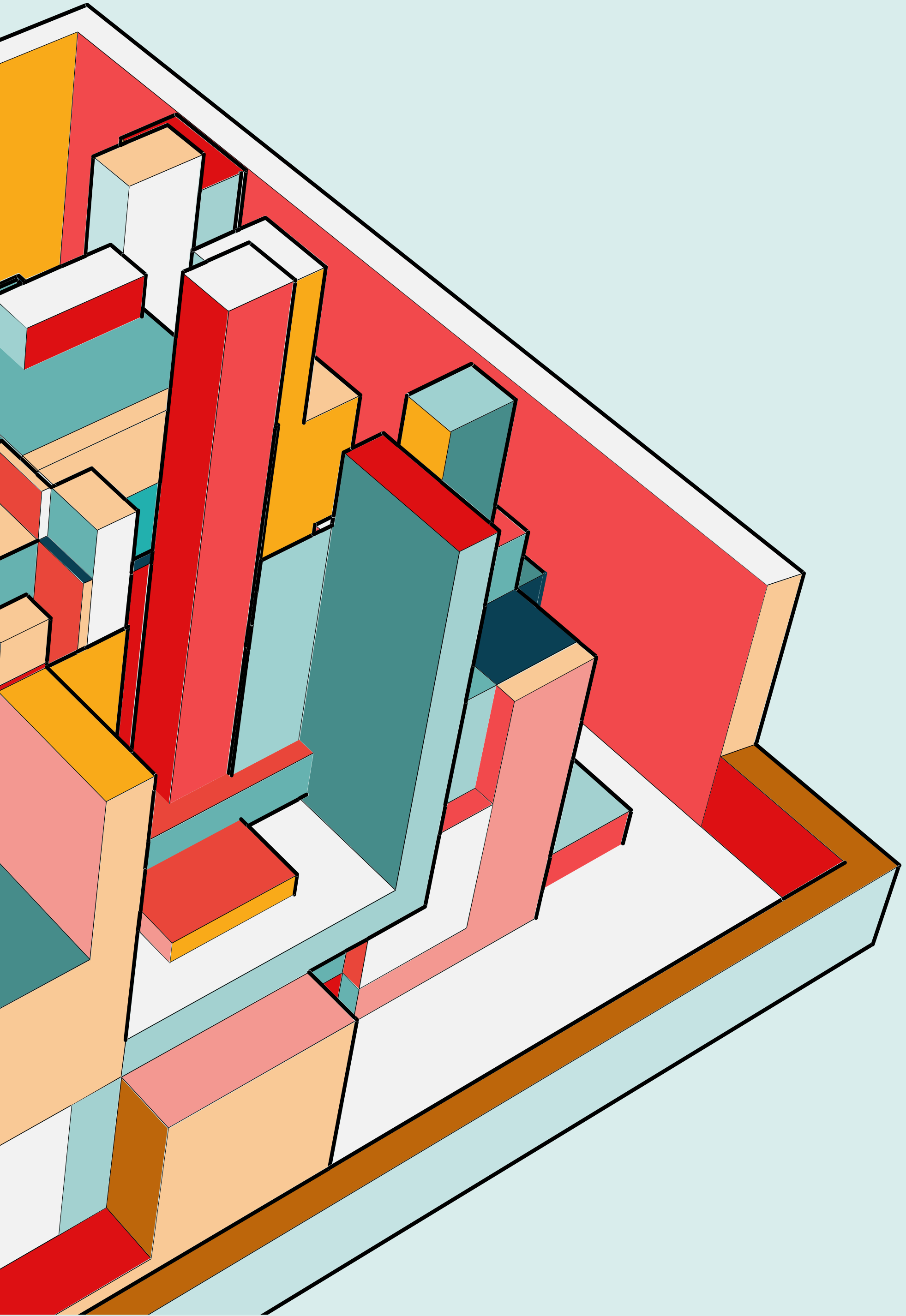
If you have an IT or DevOps team who will set up environments for you, so you’re compliant, secure, etc… that’s great, but no need to do this yourself, from scratch.

Fussing with infrastructure can slow down the velocity of your data insights.

COMPLIANCE NOTE ON LLMS

LLM's are very helpful, but they won't *understand* things for you. AGI ain't here.

- Some folks are creating data breaches each time they use an LLM. (HIPPA, for example)
- Be careful about sending material, non-public data to a 3rd party…
- 11% of data provided to ChatGPT is confidential, material, non-public
- Microsoft Copilot for Work advertises security, but your mileage may vary.
- Generate synthetic data with synthpop, Synthetic Data Vault, or LLM itself
- Recommend only feeding synthetic data into an LLM, as examples, to generate code to use in your secure environment.



SELECTED READINGS AND RESOURCES

SELECTED READINGS

Think Python by Downey

Elements of Data Science by Downey

Python for Data Analysis by McKinney

Think Stats by Downey

Think Bayes by Downey

Causal Inference for the Brave and True

R for Data Science by Wickham

Regression and Other Stories by Gelman

Bayes Rules! by Johnson, Ott, Dogucu

Statistical Rethinking by McElreath

ggplot2:: Elegant Graphics for Data Analysis by Wickham

Tidy Modeling with R by Kuhn and Slidge

ONE PATH TO DATA SCIENCE SKILL BUILDING

- Read lots of books!
- Conduct lots of analyses
- Complete lots and lots of projects!
- Your mileage may vary due to motivation and self-accountability

Whether if you learn better on your own or with others, you'll need to do lots of projects to build new skills and make new connections!

OR LEAN INTO SOCIAL ECOLOGY FOR SKILL BULIDING

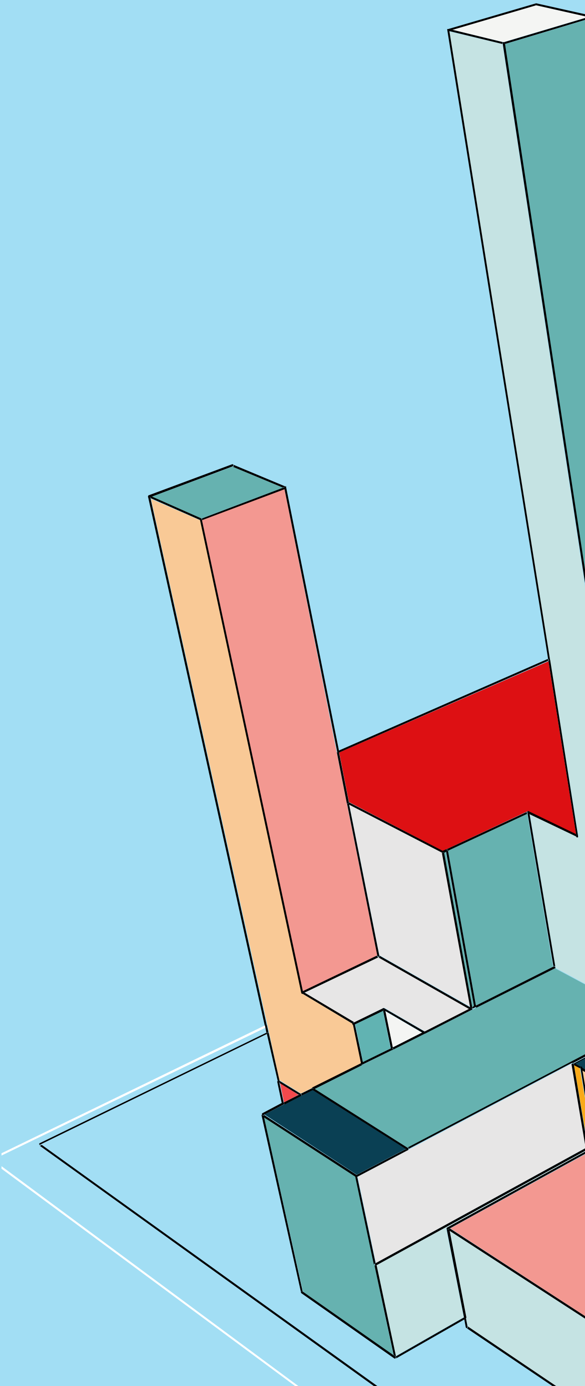
Community Creates Accountability

Community Creates New Connections

- Data Science Book Club
- Statistics Study Group
- Bayesian Inference Club
- Data Science Project Sharing Group
- Mentoring others, seeking mentors

FIND A GROUP THAT WORKS FOR YOU!

1. Basic Python Practice Group
2. Basic R for Data Science Book Club
3. "Think Stats" Python Book Club
4. "Elements of Data Science" Python Book Club
5. Causal Inference study group (Python)
6. Bayesian Inference in R book club



OTHER CONTRIBUTIONS

101 Exercises for Python and JS

90 Minutes to Machine Learning video and code

67 Minutes to NumPy video and code

Finding the Story in the Data

Hypothesis Testing in a Nutshell

Contact:

Ryan.Orsinger at gmail dot com