

# Data analysis, data science, data engineering

## an overview

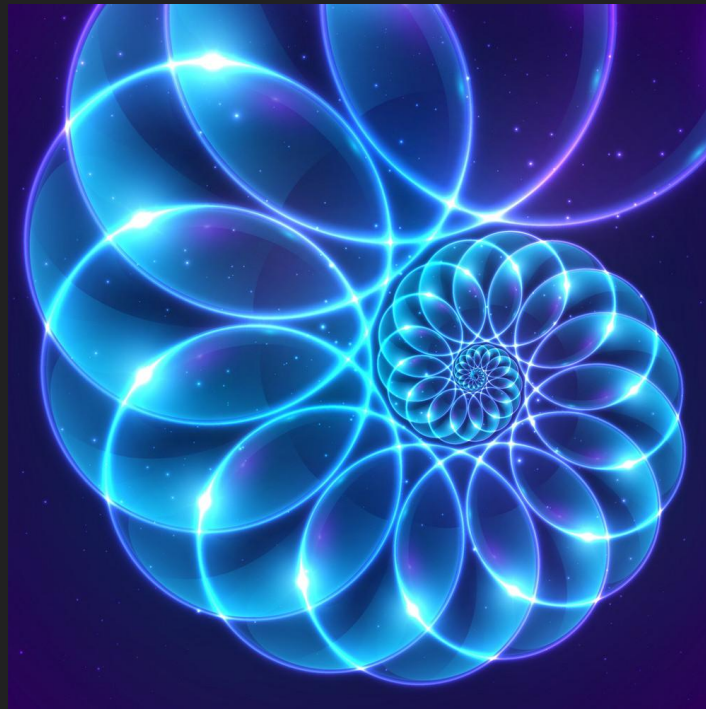
What we talk about when we talk about data work

“All models are wrong, but some are useful”

Box, 1979

# What is data analysis?

1. categorize
2.     -> count
3.         -> compare
4.             -> consider
5.                 -> communicate
6. (loop)
7. a **mundane** skill
8. 30 minute interactive: <https://git.io/datatalk>



# What is data engineering?

- categorize -> count -> compare -> consider -> communicate
- ... -> count -> ...
  
- “Import CSV: there’s your ‘engineering’ Mr. Prententious Hacker.”
  - ... hmmm, this data is kind of messy.
- Hey, Jane knows Python! Maybe she could write a short script ...
- ... Jane is managing a github repo to host the Python library ...
- ... let’s hire another developer to take the load off Jane ...
- ... Jane thinks about starting a company of her own ...

# How should we think about data engineering?

- I am a ~~brilliant code poet~~ cost center
- I am a means—towards what end ...
- ... data analysis?
- How do I know if I have done a good job?

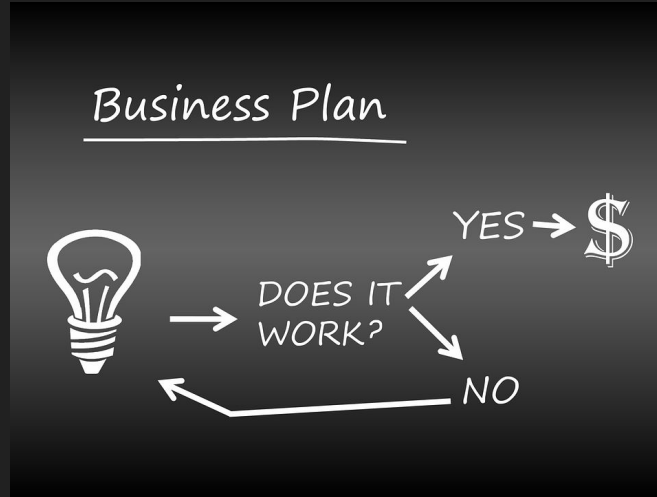
# What is data science?

- categorize -> count -> compare -> consider -> communicate
- ... -> compare -> ... (with higher mathematics)
- ... -> consider -> communicate

# Once more, with feeling

- categorize -> count -> compare -> communicate -> consider
- data eng data science

# Who is data analysis for?





# Data analysis for decision making ...

## ... is an organizational skill

- categorize -> count -> compare -> communicate -> consider -> **CHOOSE**
- client eng      data eng                      data science                      **PRODUCT**
- (cost)              (cost)                      (cost)                      **REVENUE!**

Notice ....

1. Hey, where is “coordination?”
2. The causal flow is **from** means **to** ends
3. We’re “data-driven” ... in the worst way ...

# Reverse the default flow

- decider -> analyst -> engineers
- data, "like fire, will be a good servant, [but] an ill master."

# Data engineering #protips

1. [Airflow](#) is at least half of the right model ...
2. ... as long as you force it to be the decision-makers' servant.
3. Transmit strongly-typed, offset-oriented protocols (i.e. [protobuf](#)) ...
4. ... at scale, JSON becomes an awful serialization format.

# Thank you

[sam.penrose@stem.com](mailto:sam.penrose@stem.com)

<https://git.io/datatalk2> (these slides)

<https://git.io/datatalk>

<https://twitter.com/sampenrose>

<http://www.sampenrose.net>

