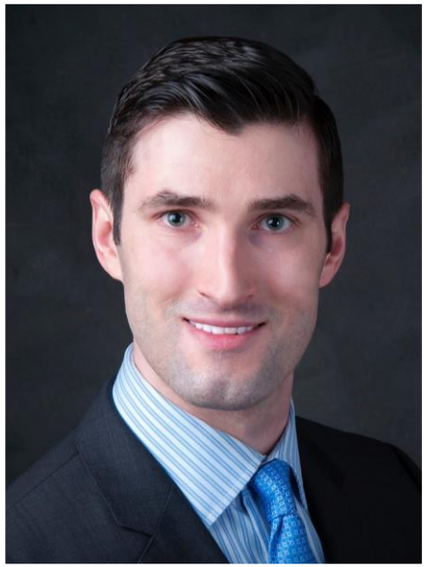
# INTRO TO DATA SCIENCE

## LECTURE 1

# AGENDA

- Producer introduction
- Instructor introduction
- Student introductions
- Course logistics
- Project
- VCS, GIT, GitHub
- What is Data Science?
- Computer setup

# INSTRUCTOR INTRODUCTION



MICHAEL GALVIN

- Data Scientist, GE
- galvin.mj@gmail.com
- @MikeJGalvin
- linkedin.com/in/MikeJGalvin

# STUDENT INTRODUCTIONS

- Your name
- A brief summary of your background (work, school, etc.)
- What do you hope to get out of this class

# COURSE EXPECTATIONS:

- BE PRESENT
- PARTICIPATE
- GET TO KNOW EACH OTHER
- ASK QUESTIONS

# COURSE LOGISTICS

- Syllabus
- Schedule
- Office hours
- Slack
- Homework
- Course Project
- Github

# WHAT IS DATA SCIENCE?

## WHAT IS DATA SCIENCE?

- No Formal Definition

## WHAT IS DATA SCIENCE?

- No Formal Definition
- A set of tools and techniques used to extract useful information from data

## WHAT IS DATA SCIENCE?

- No Formal Definition
- A set of tools and techniques used to extract useful information from data
- An interdisciplinary, problem-oriented subject.

## WHAT IS DATA SCIENCE?

- No Formal Definition
- A set of tools and techniques used to extract useful information from data
- An interdisciplinary, problem-oriented subject
- The application of scientific techniques to practical problems

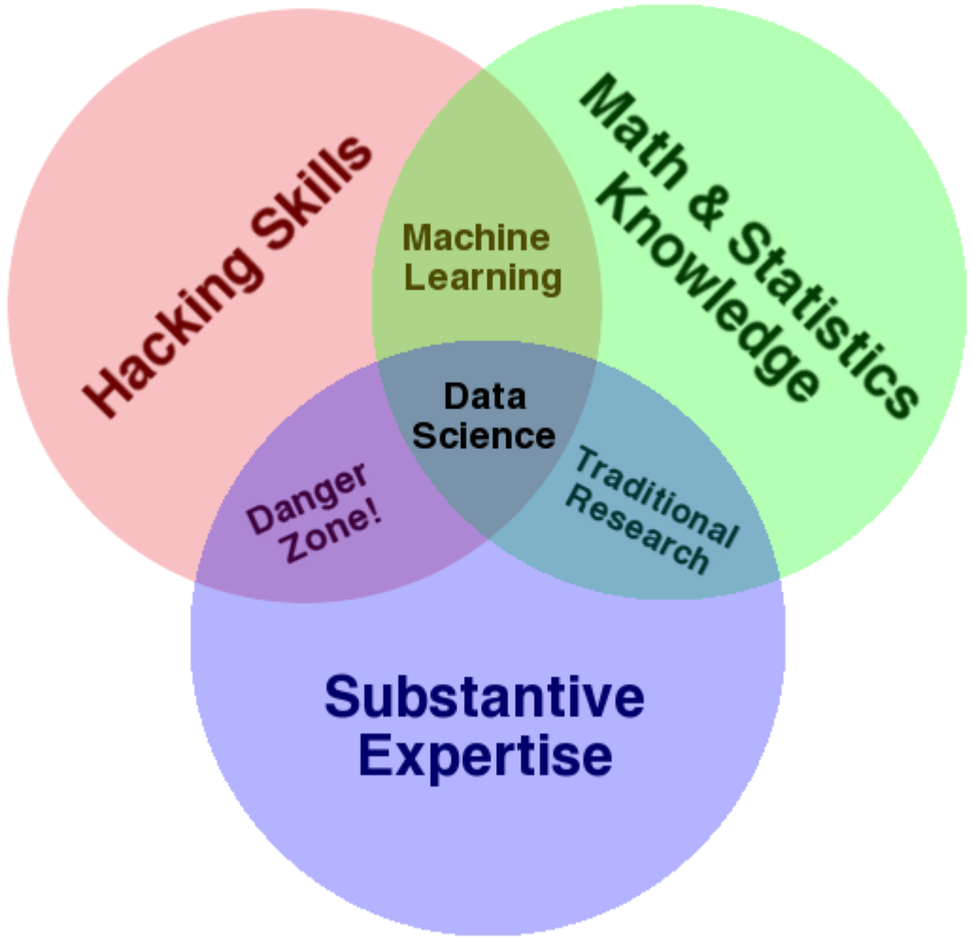
## WHAT IS DATA SCIENCE?

- No Formal Definition
- A set of tools and techniques used to extract useful information from data
- An interdisciplinary, problem-oriented subject.
- The application of scientific techniques to practical problems.
- Analyzes raw data to make impactful decisions

## WHAT IS DATA SCIENCE?

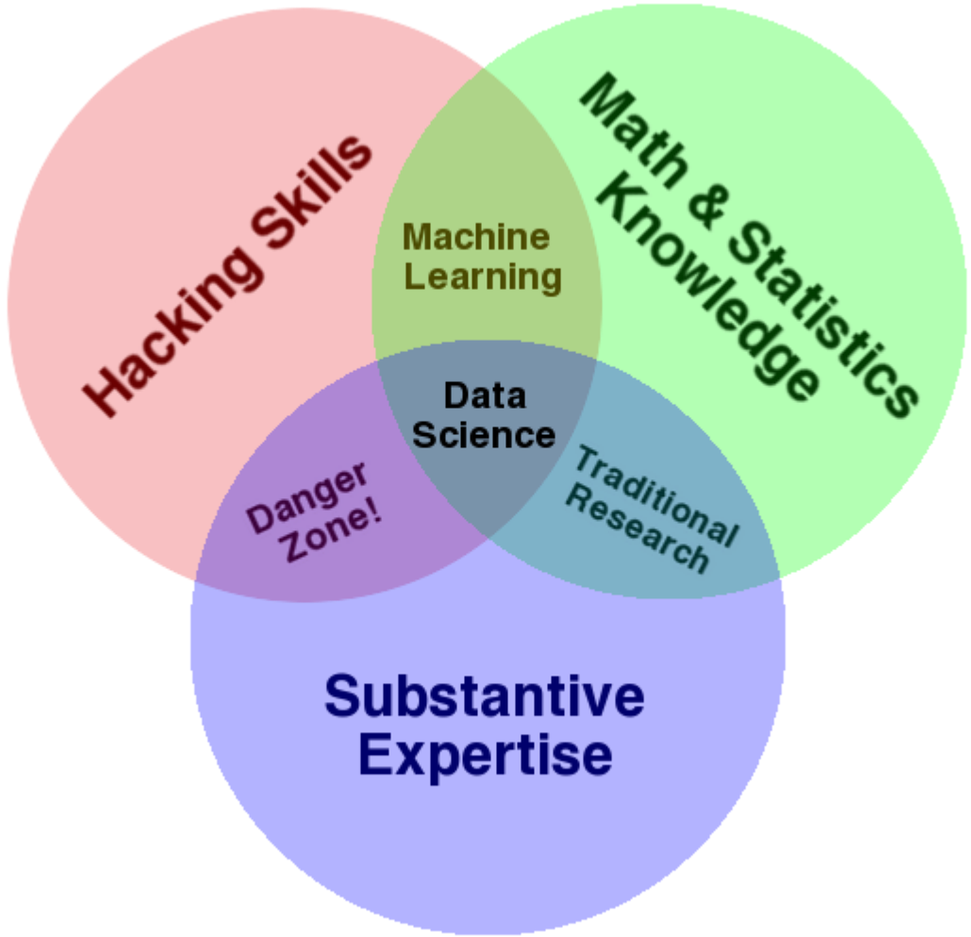
- No Formal Definition
- A set of tools and techniques used to extract useful information from data
- An interdisciplinary, problem-oriented subject.
- The application of scientific techniques to practical problems.
- Analyzes raw data to make impactful decisions
- A rapidly growing field.

## QUALITIES OF A DATA SCIENTIST?



## QUALITIES OF A DATA SCIENTIST?

- Programming Languages
- Hadoop
- Databases
- Spark
- Map-Reduce
- Statistical Computing eg. R

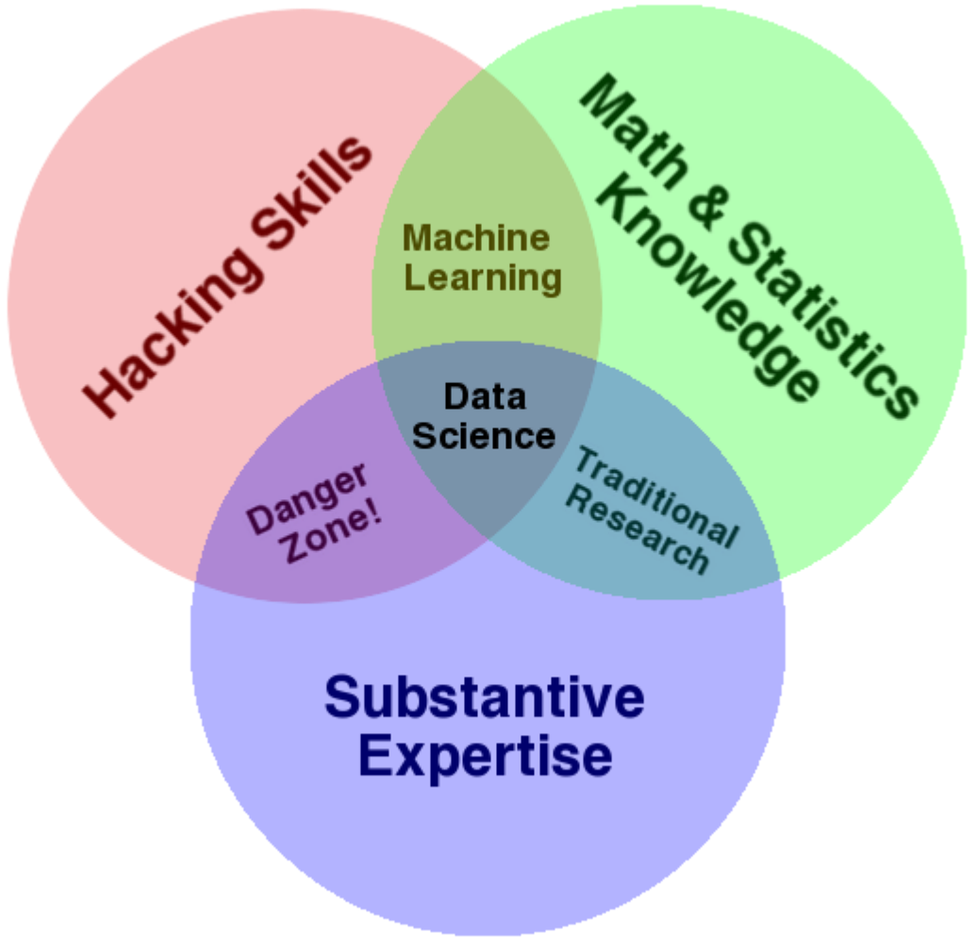



- Statistics
- Machine Learning
- Optimization
- Probability
- Linear Algebra

- Business Understanding
- Influence
- Strategy

## QUALITIES OF A DATA SCIENTIST?

- Programming Languages
- Hadoop
- Databases
- Spark
- Map-Reduce
- Statistical Computing eg. R




- Statistics
- Machine Learning
- Optimization
- Probability
- Linear Algebra

- Business Understanding
- Influence
- Strategy
- Collaboration

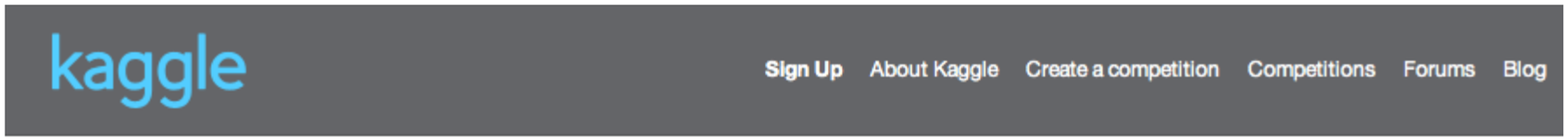
A FEW MORE THINGS!

- Visualization
- Change Management
- Communication

## WHO USES DATA SCIENCE?


- Develop best movie rating prediction
- $50k Progress prize yearly until winner
- $1,000,000 prize to winner (10% improvement)
- Started: October 2, 2006
- Ended: June 26, 2009

# WHO USES DATA SCIENCE?




Microsoft

Allstate.
You're in good hands.

GE

MERCK

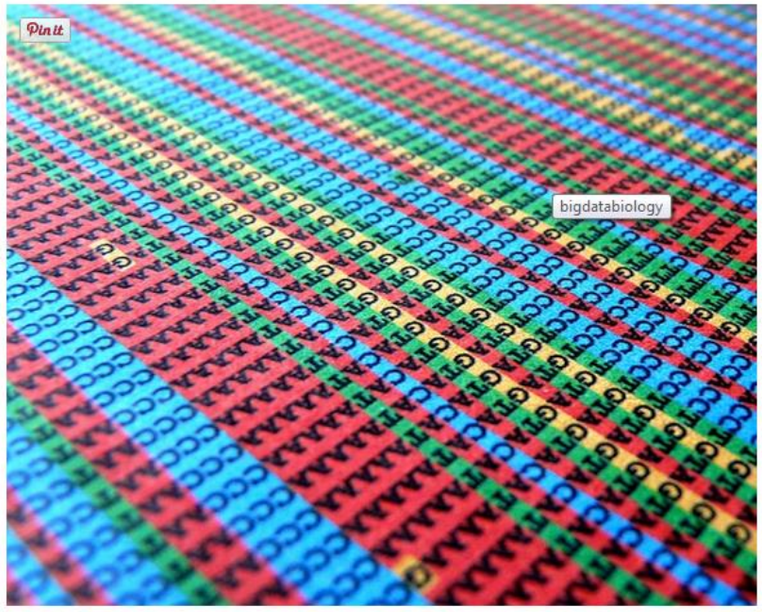
## WHO USES DATA SCIENCE?

Other interesting applications:

- Healthcare
- Defense
- Pharmaceuticals
- Agriculture
- Science
- Marketing
- Engineering

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

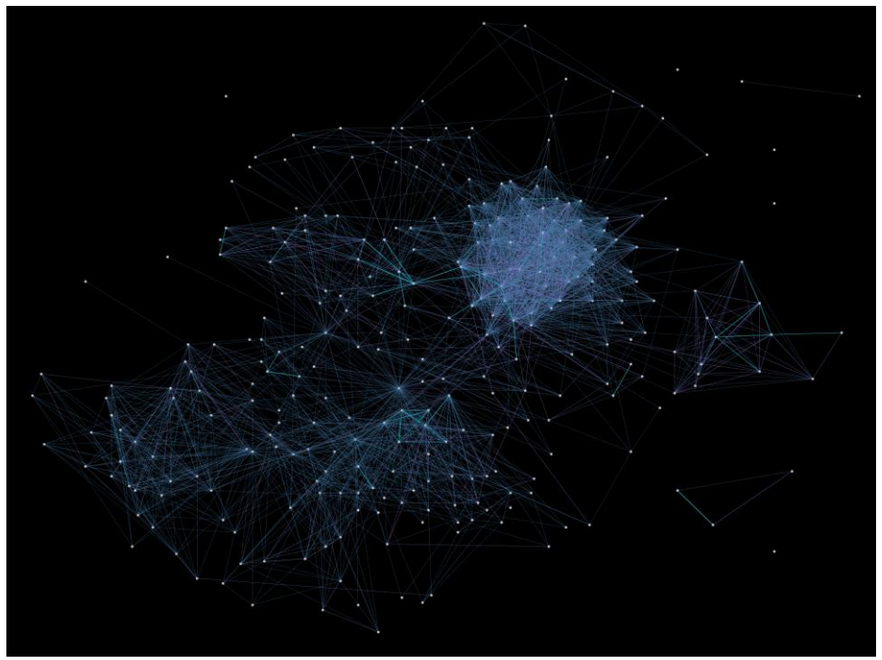
BIOLOGY'S BIG PROBLEM: THERE'S TOO MUCH DATA TO HANDLE

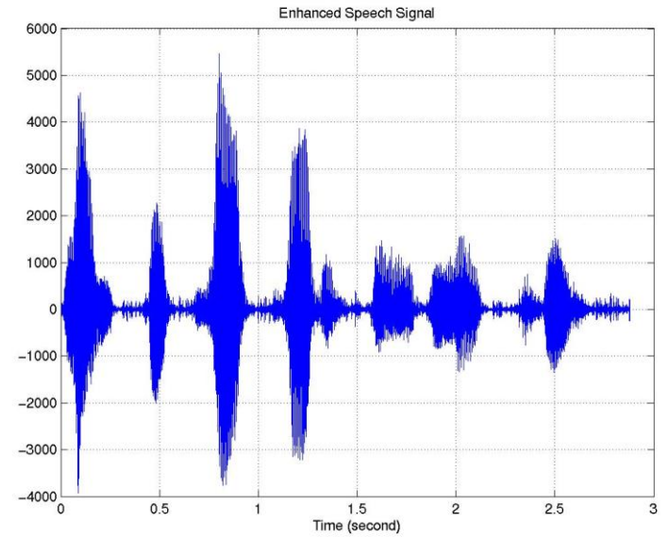


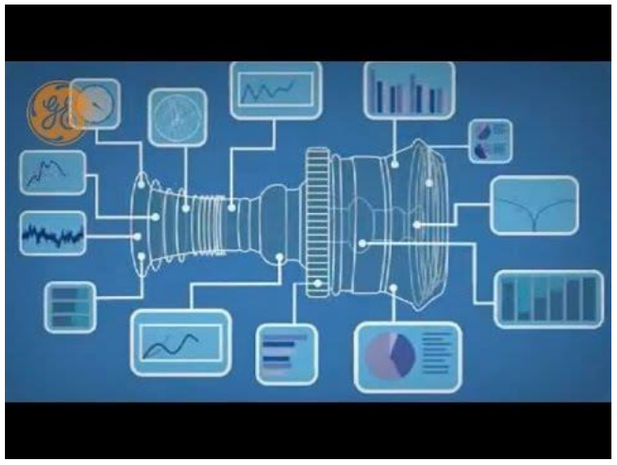

> In the next 10 years, data science and software will do more for medicine than all of the biological sciences together.
>
> — Vinod Khosla —

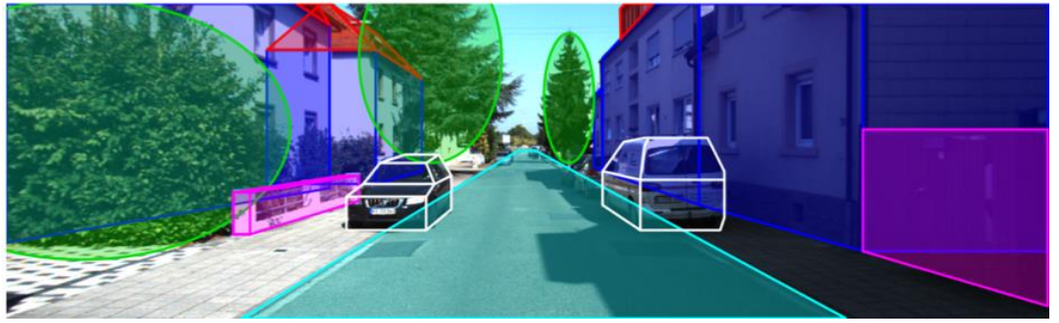
## DATA TYPES

Numeric
Text
Audio
Network
Image
Media
Metadata
Sensor
Weather
Time Series
GPS





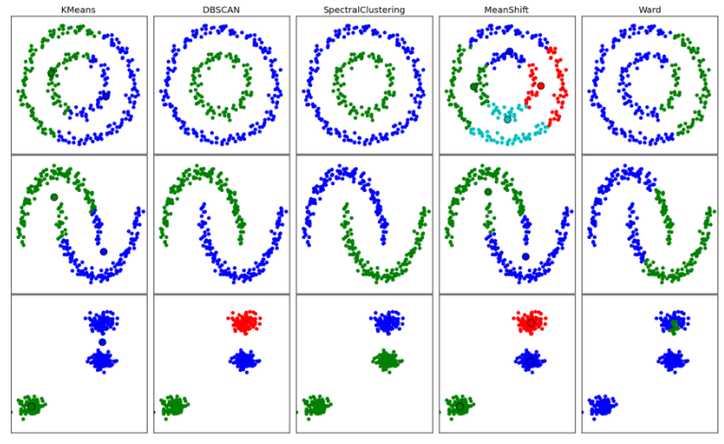


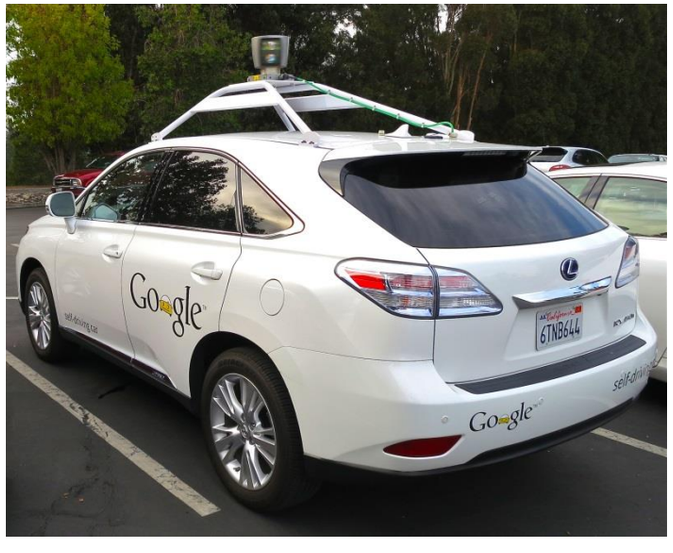
## DATA TYPES?


Data Science <> Big Data

# PROBLEM TYPES

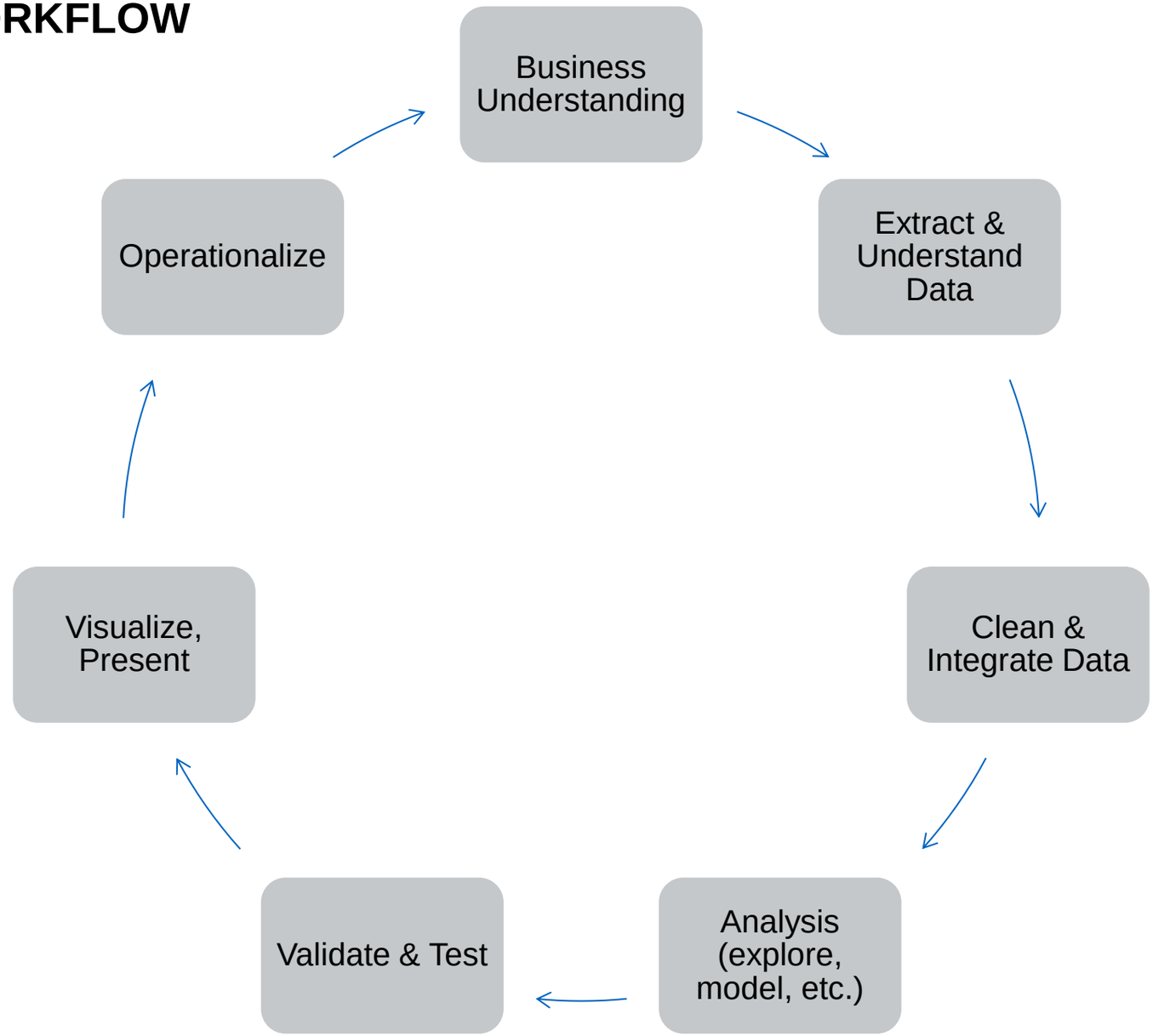
- Recommendations System
- Associations
- Relationships
- Structure
- Predictions
- Forecasting
- Optimization
- Vision
- Speech
- Audio

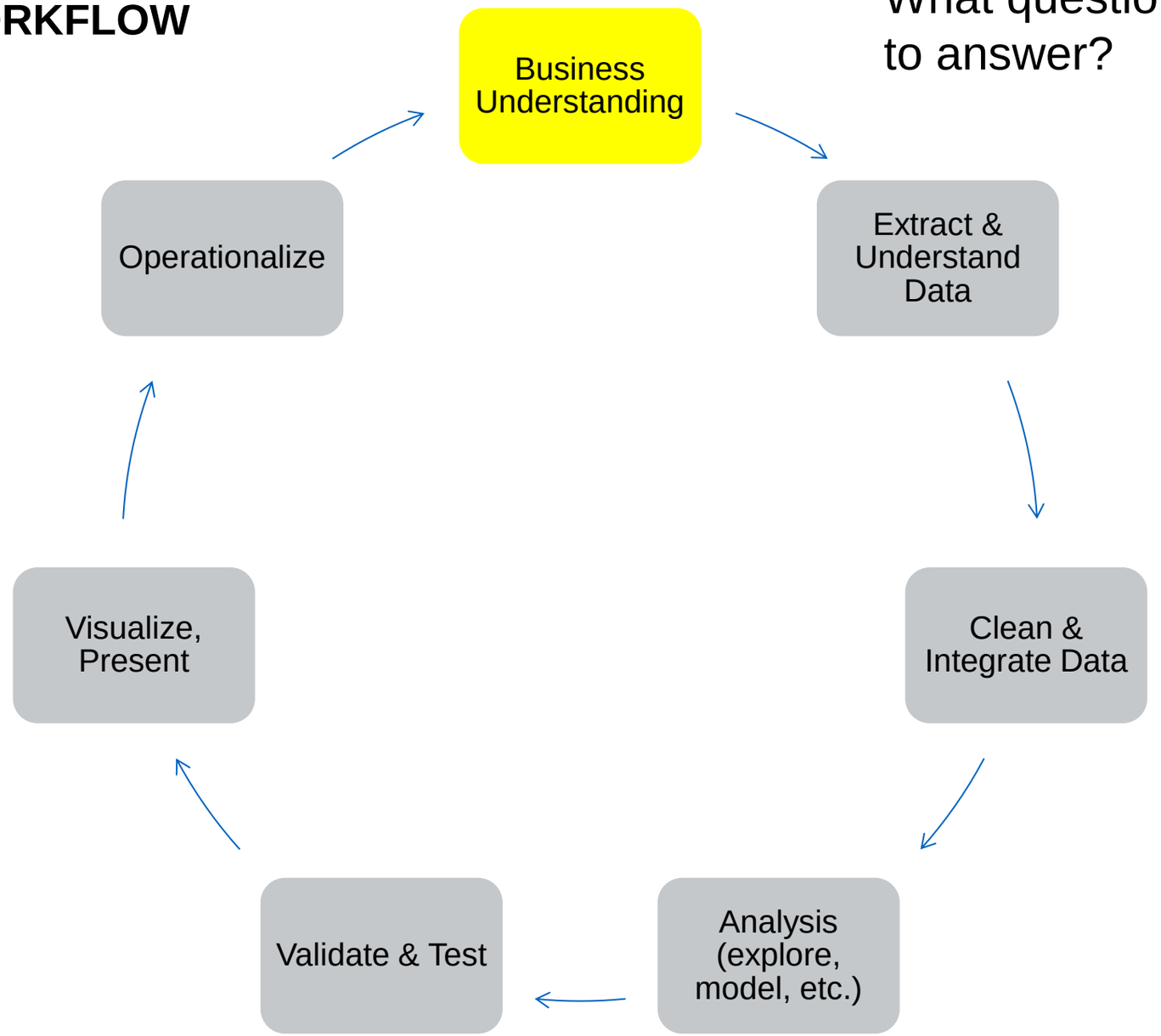

# DATA SCIENCE WORKFLOW

# DATA SCIENCE WORKFLOW

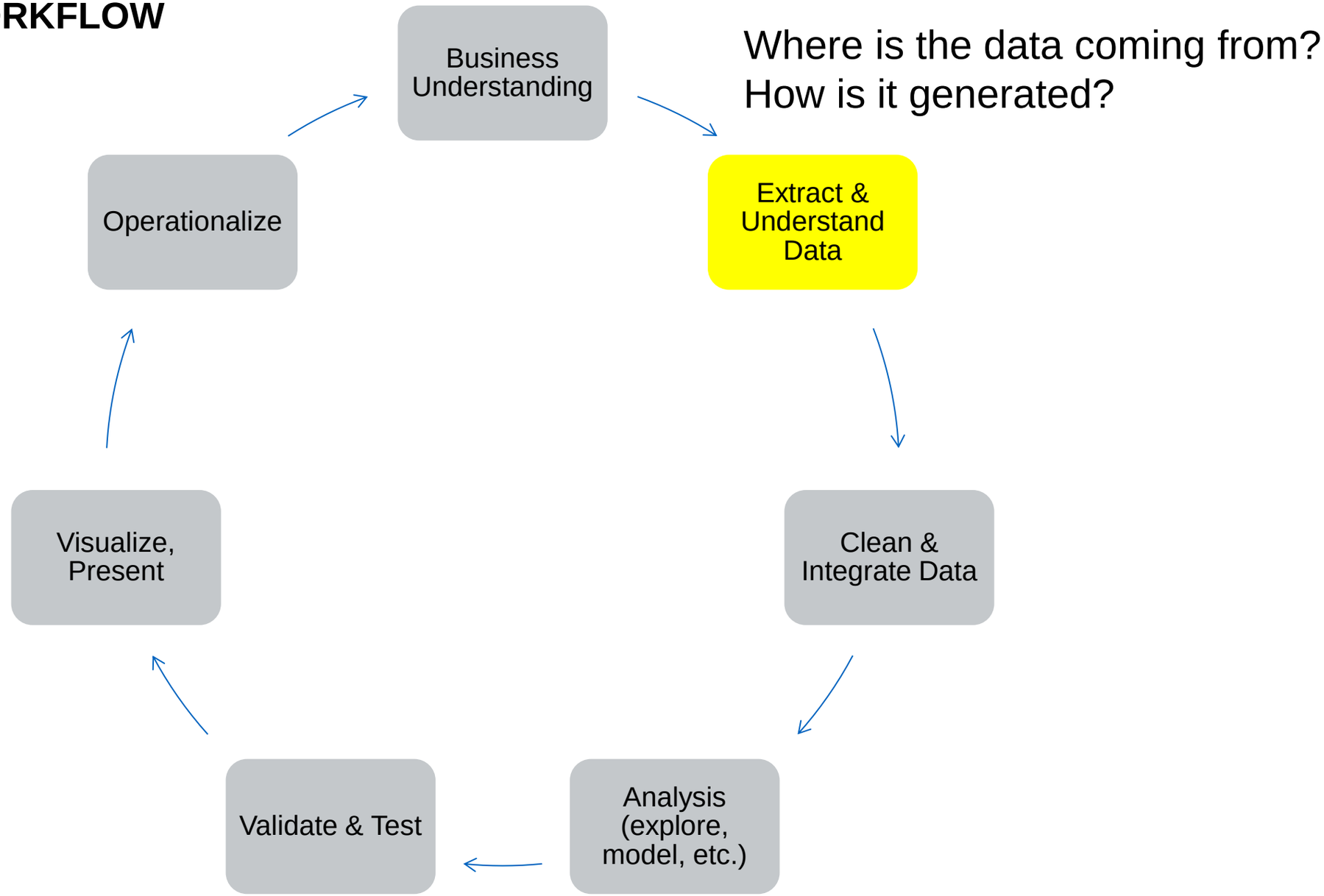


# DATA SCIENCE WORKFLOW

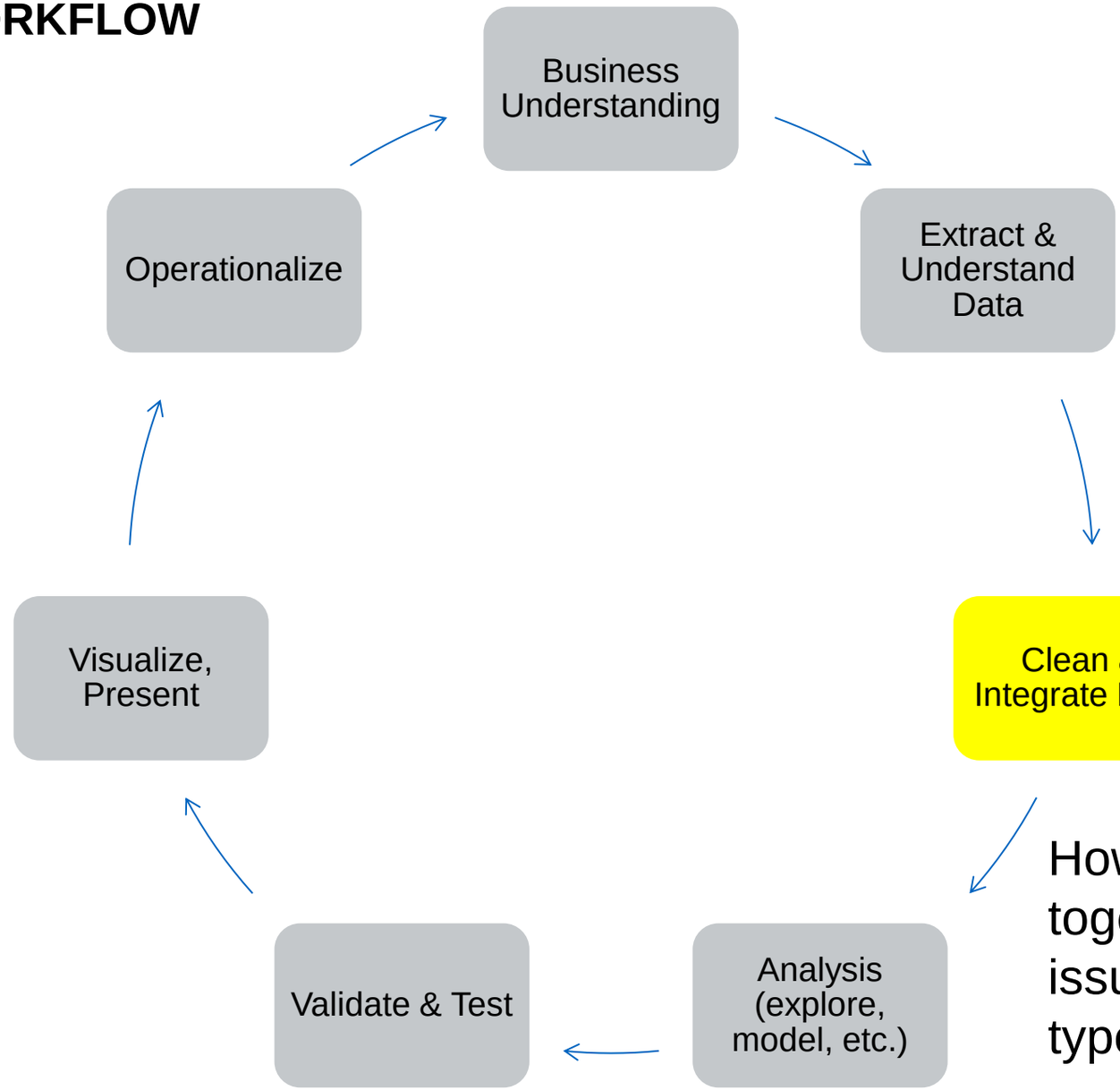
What question are you trying to answer?



# DATA SCIENCE WORKFLOW

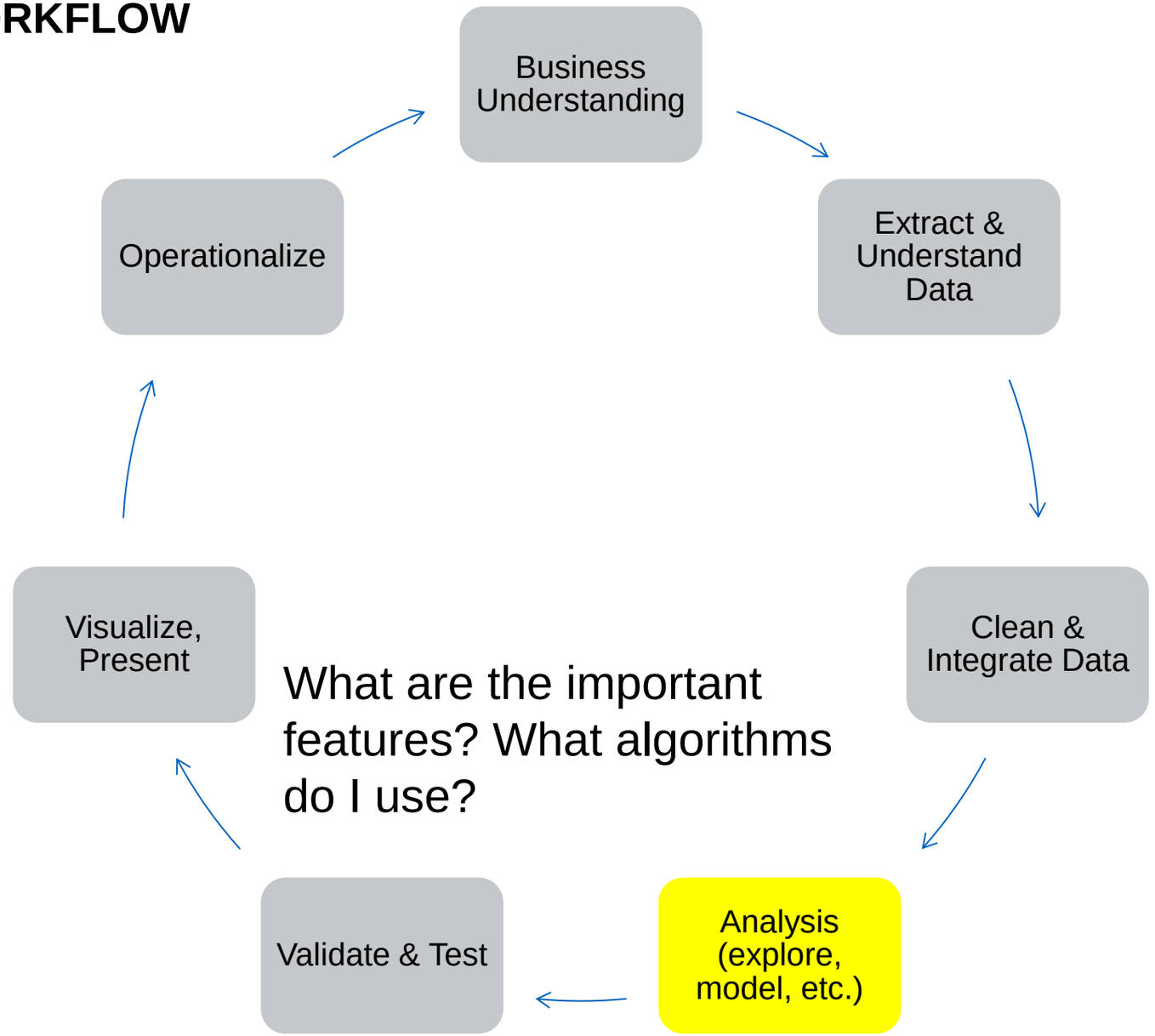


# DATA SCIENCE WORKFLOW

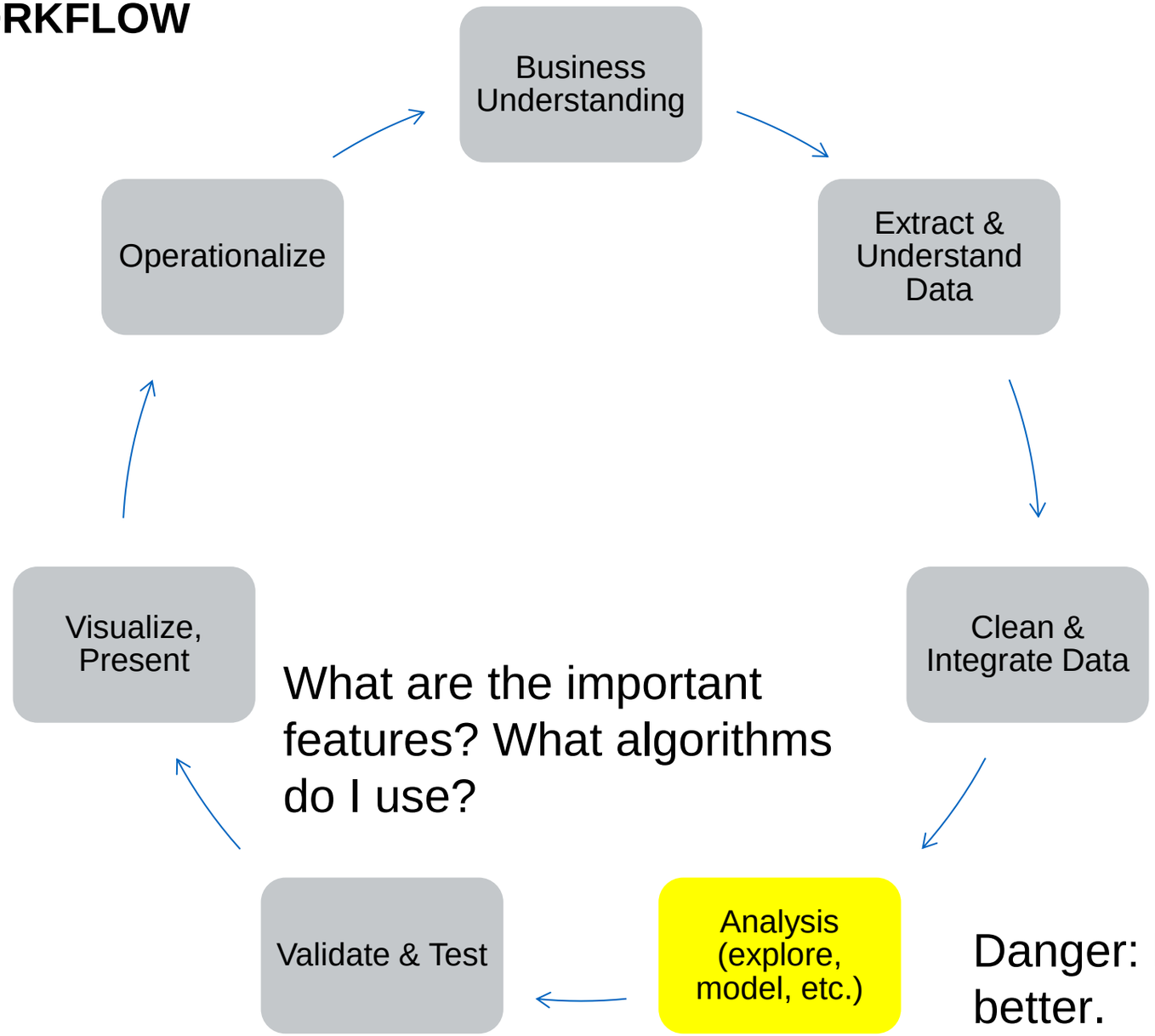



How to connect data sources together? Are there data quality issues (missing values, wrong type, etc.)

# DATA SCIENCE WORKFLOW

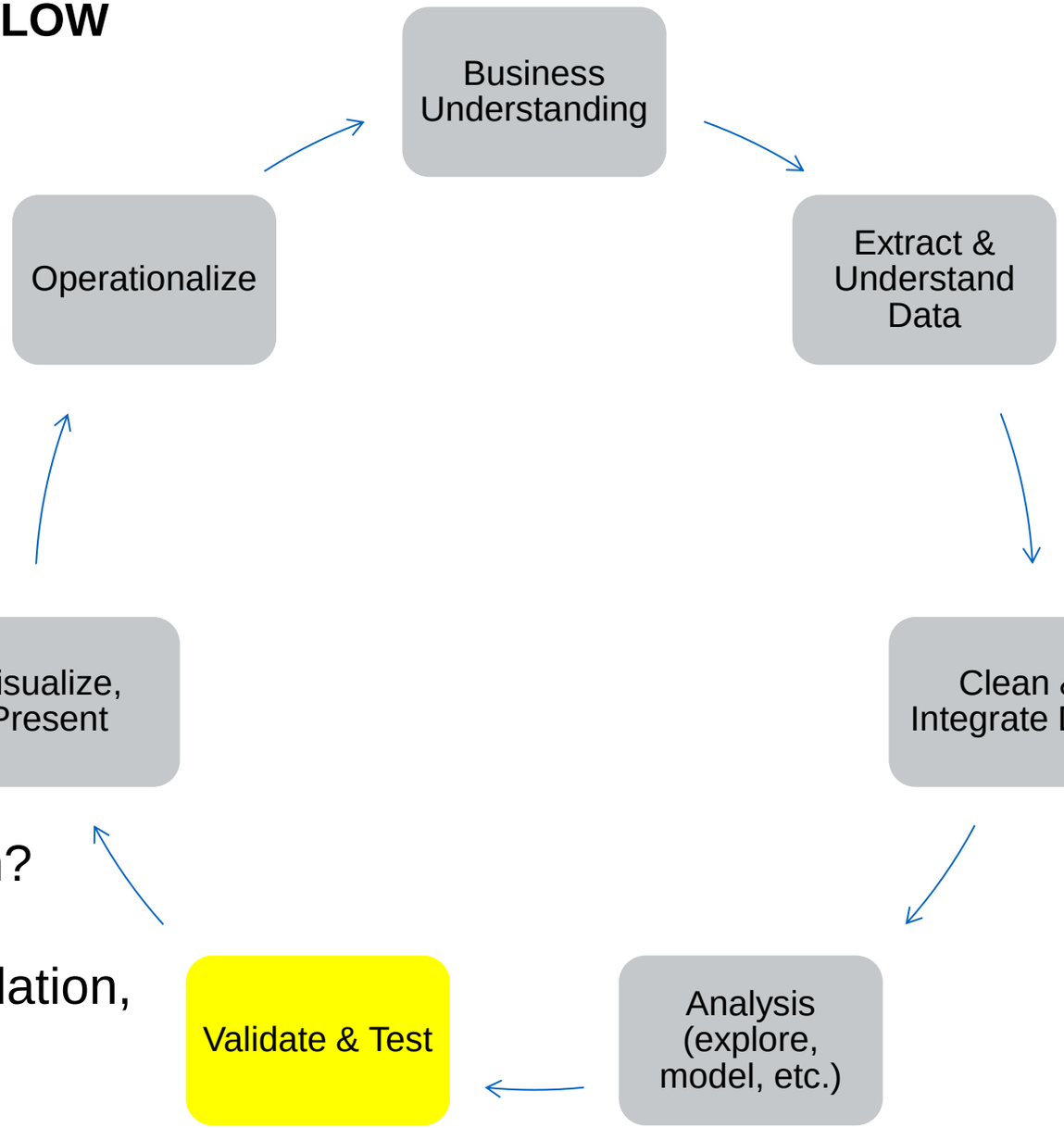


# DATA SCIENCE WORKFLOW



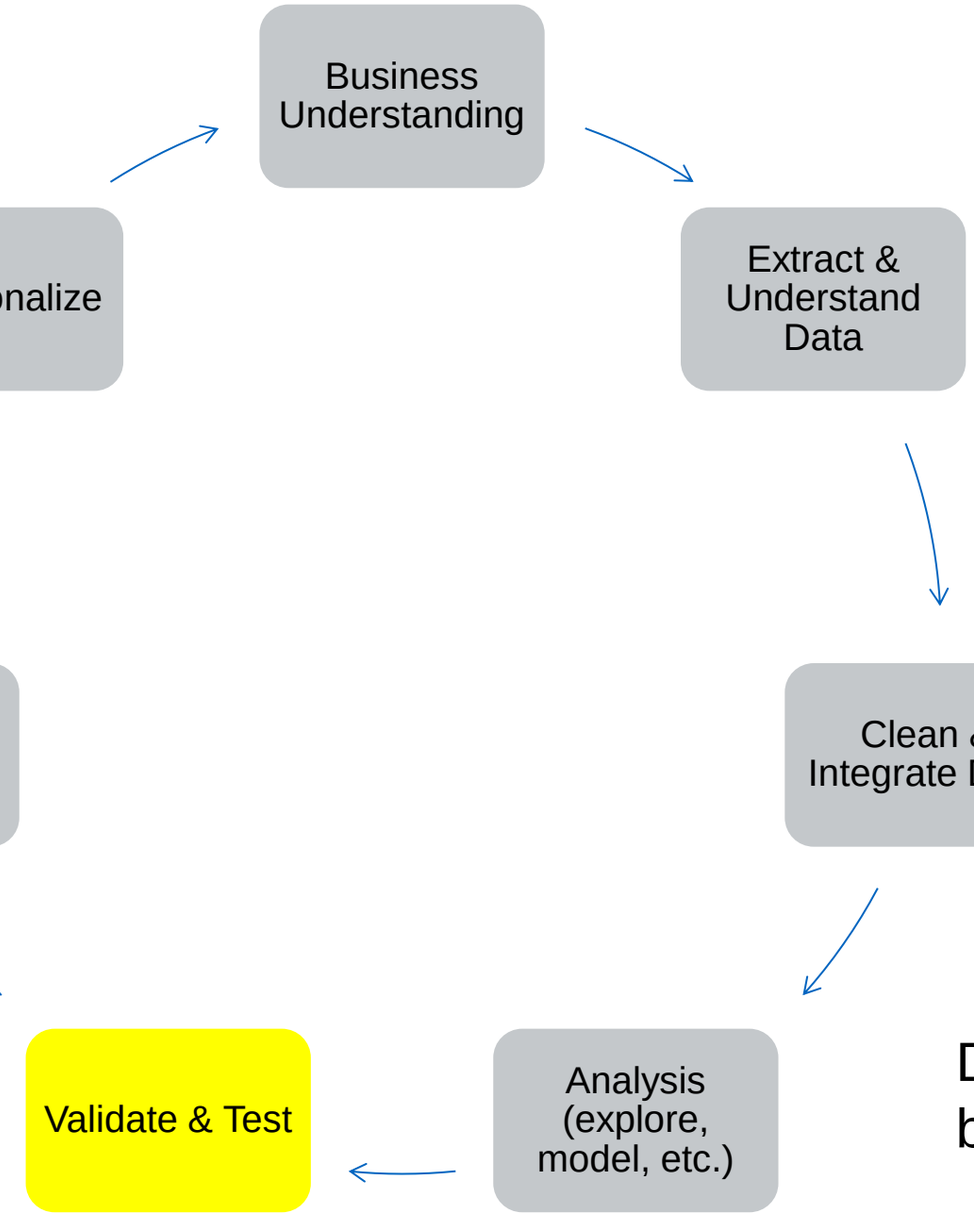

Danger: More isn’t always better.

# DATA SCIENCE WORKFLOW




Did I answer the question?
How well?
Appropriate training, validation, and testing

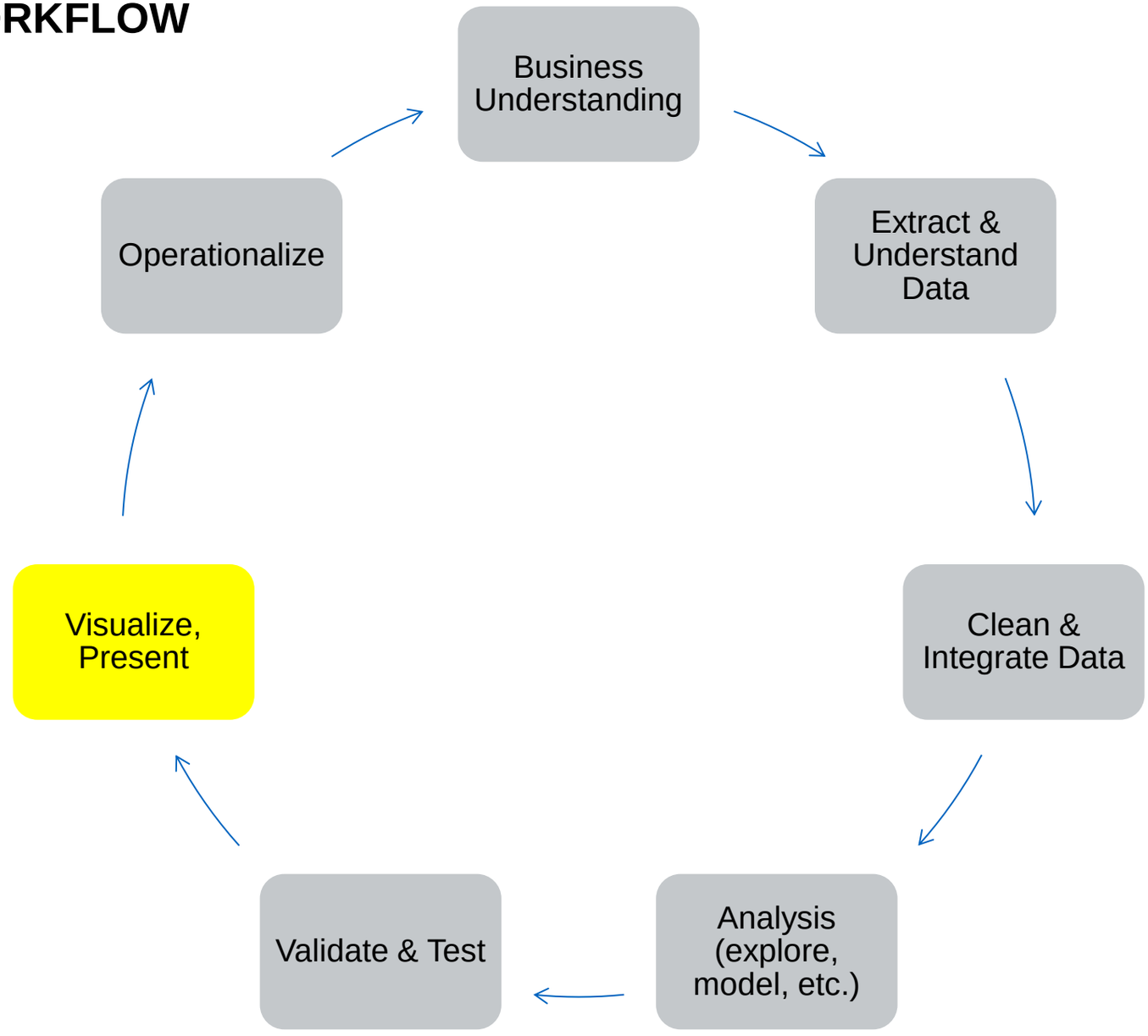
# DATA SCIENCE WORKFLOW




Did I answer the question?
How well?
Appropriate training, validation, and testing

Danger: Metrics can be deceiving.

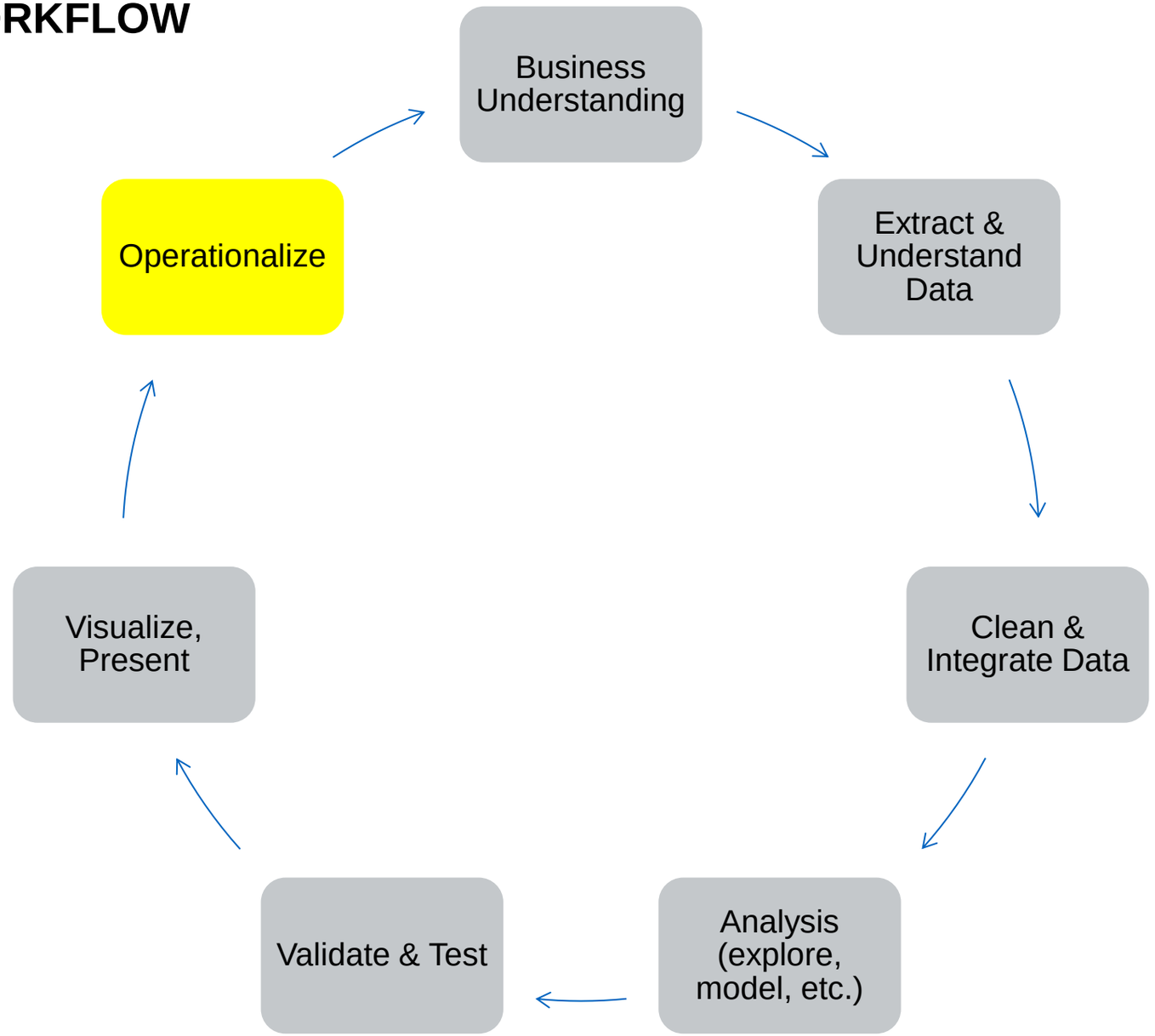
## DATA SCIENCE WORKFLOW




Tell the story

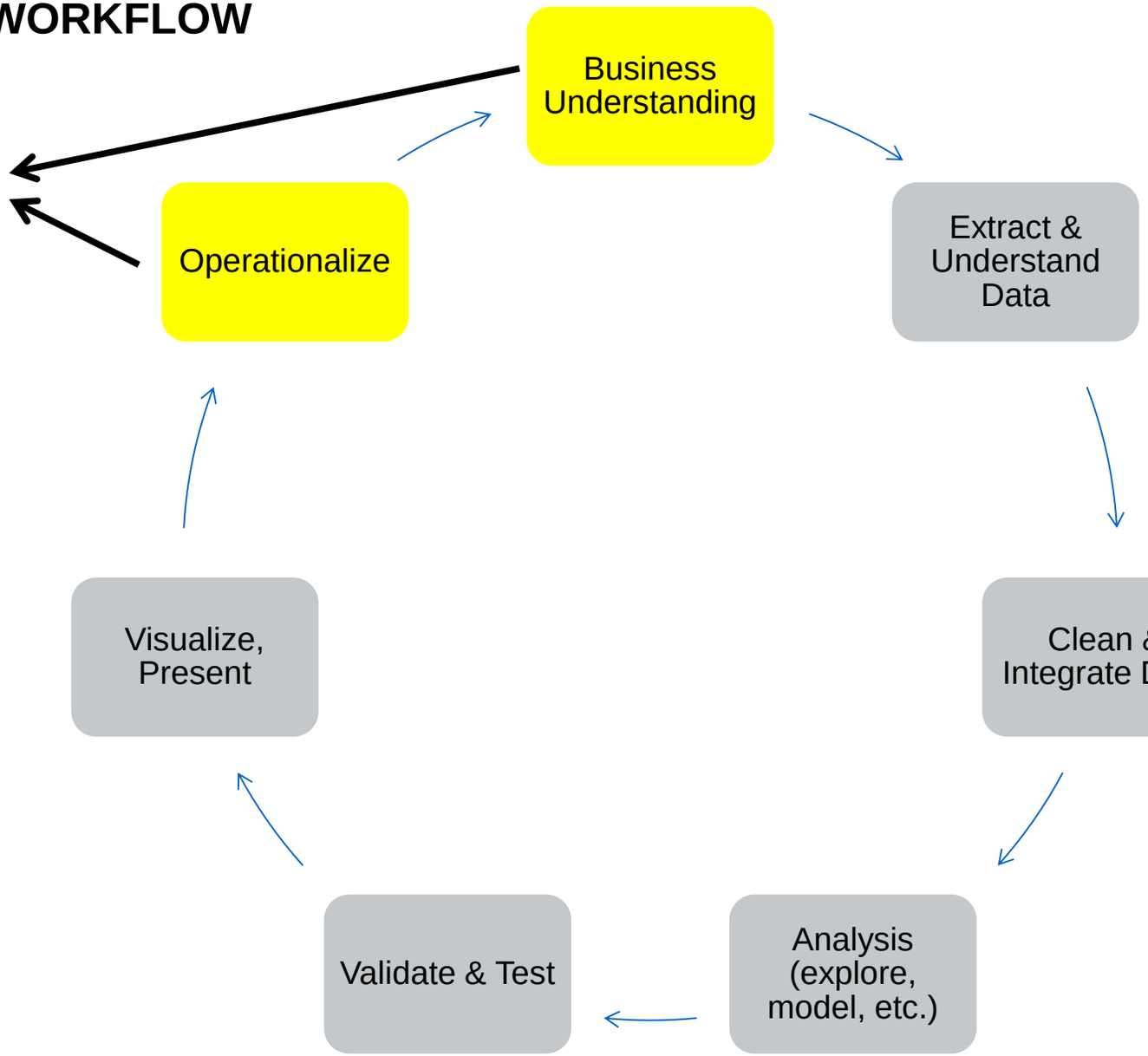
## DATA SCIENCE WORKFLOW

How do I make this impactful?



## DATA SCIENCE WORKFLOW

Make this connection early



# DISCUSSION Q&A