
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](https://www.linkedin.com/in/MikeJGalvin)
- [linkedin.com/in/MikeJGalvin](https://www.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?

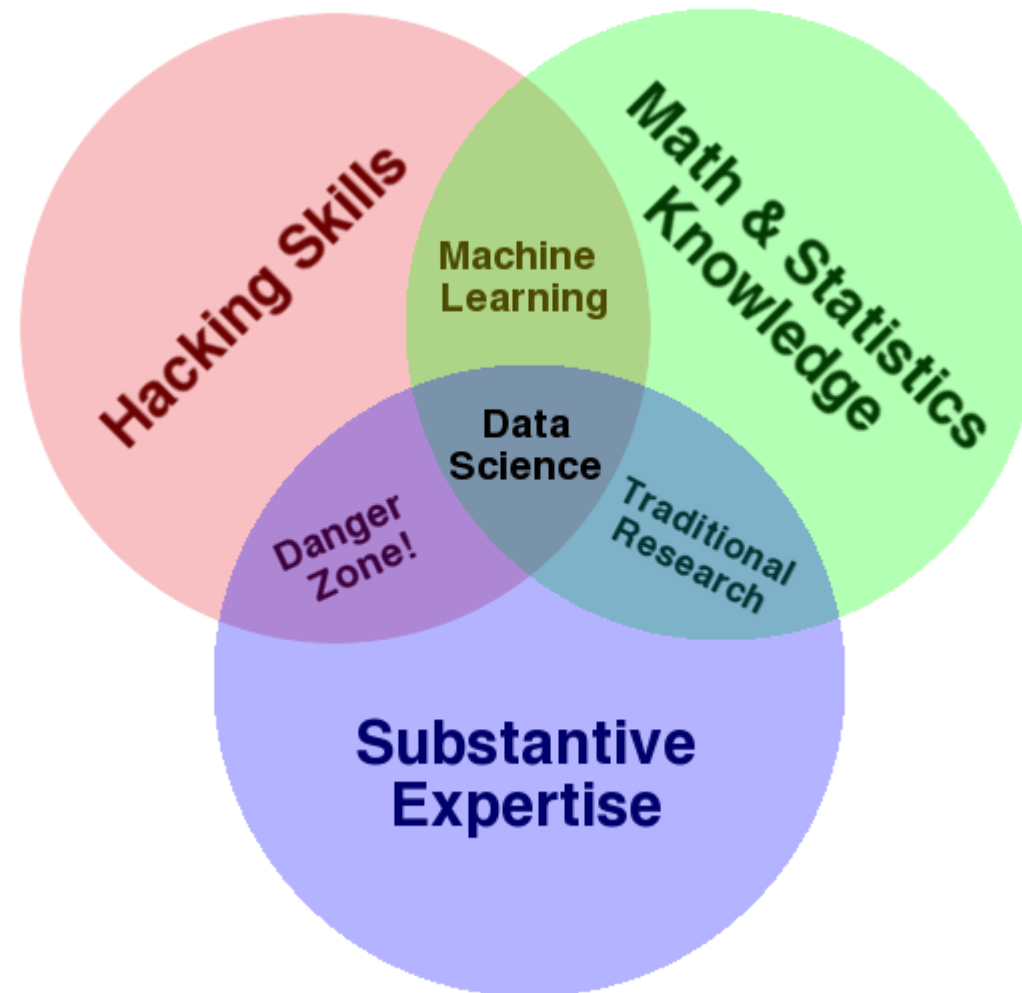
- No Formal Definition
- A set of tools and techniques used to extract useful information from data
- An interdisciplinary, problem-oriented subject.
- The application of scientific techniques to practical problems.
- Analyzes raw data to make impactful decisions

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.

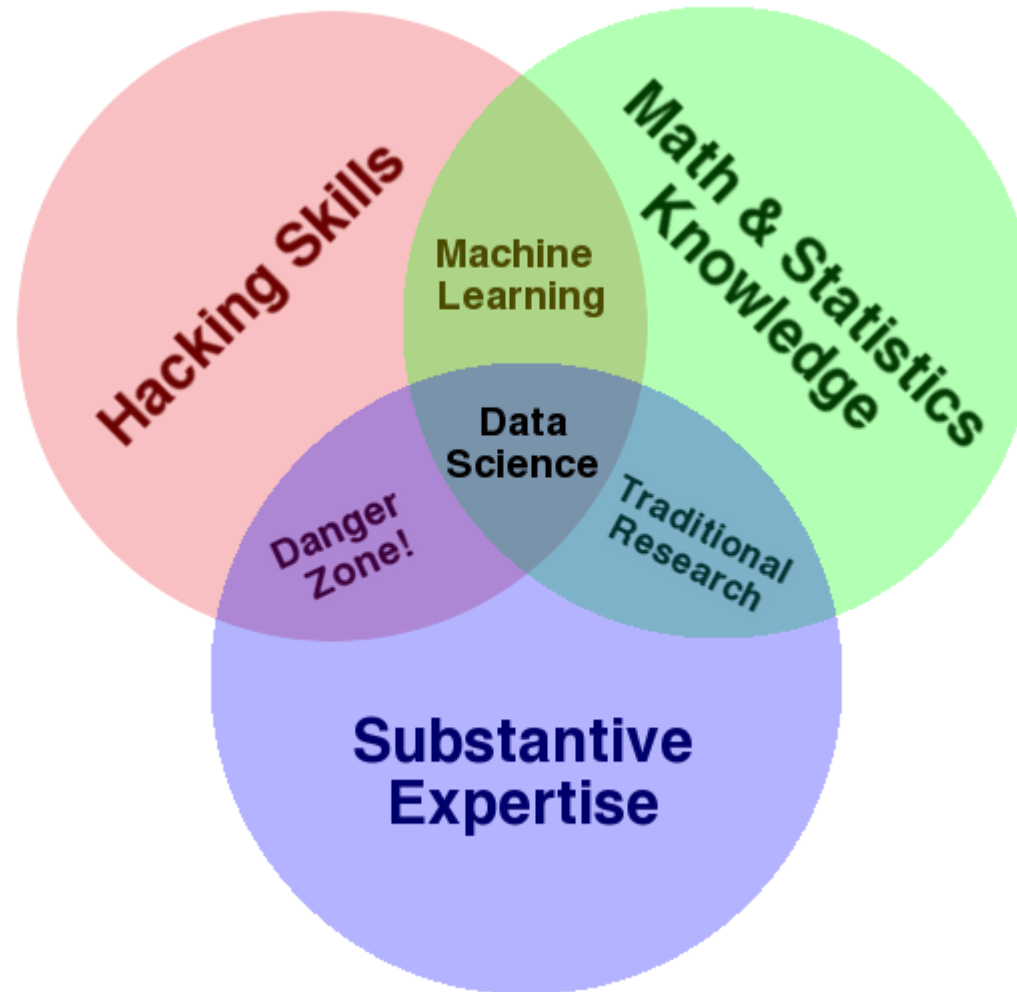


QUALITIES OF A DATA SCIENTIST?



QUALITIES OF A DATA SCIENTIST?

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

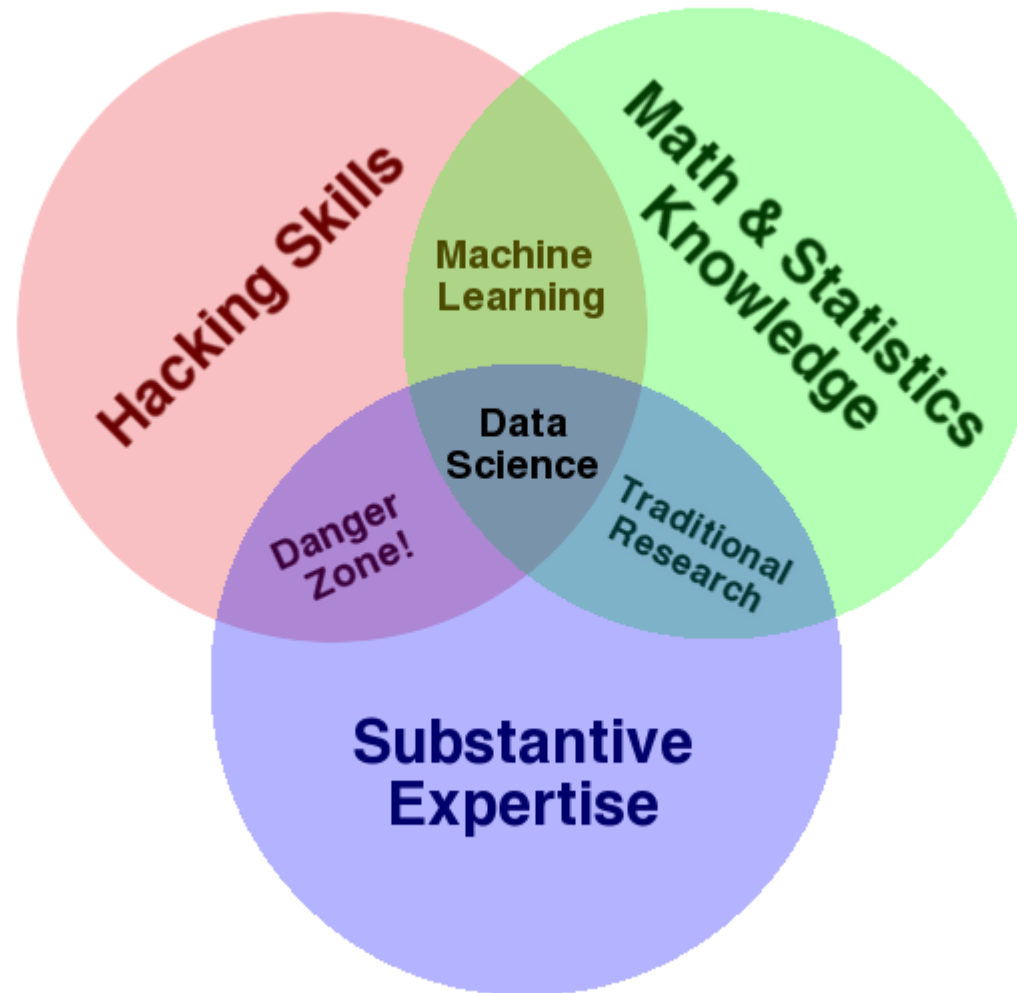


- Statistics
- Machine Learning
- Optimization
- Probability
- Linear Algebra

- 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?

kaggle

[Sign Up](#) [About Kaggle](#) [Create a competition](#) [Competitions](#) [Forums](#) [Blog](#)



We're making data science a sport.

Participate in competitions

Kaggle is an arena where you can match your data science skills against a global cadre of experts in statistics, mathematics, and machine learning. Whether you're a world-class algorithm wizard competing for prize money or a novice looking to learn from the best, here's your chance to jump in and geek out, for fame, fortune, or fun.

[Sign up as a competitor](#)

Create a competition

Kaggle is a platform for data prediction competitions that allows organizations to post their data and have it scrutinized by the world's best data scientists. In exchange for a prize, winning competitors provide the algorithms that beat all other methods of solving a data crunching problem. Most data problems can be framed as a competition.

[See how it works](#)



Allstate
You're in good hands.



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

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

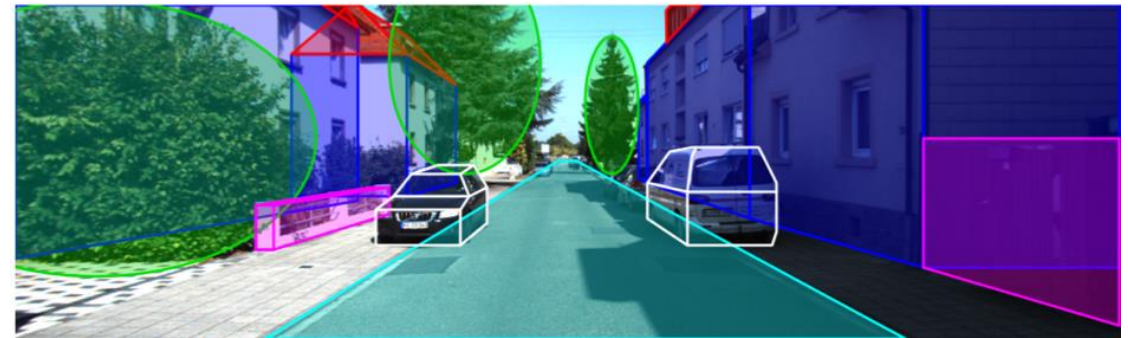
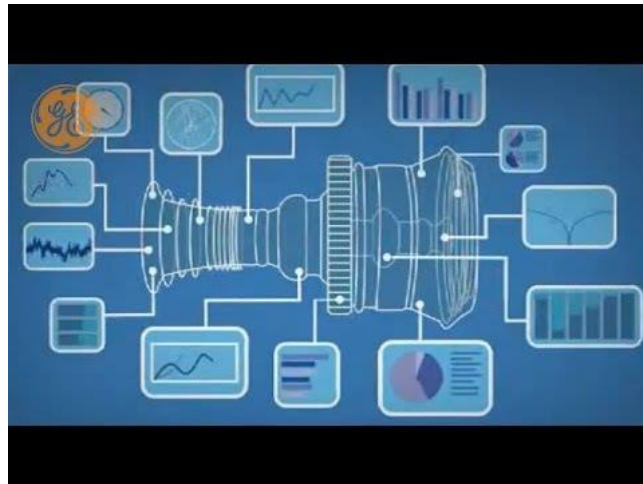
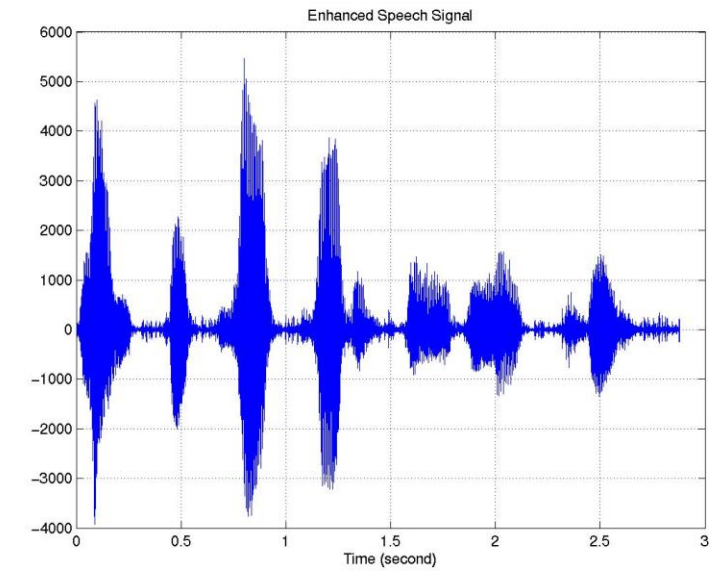
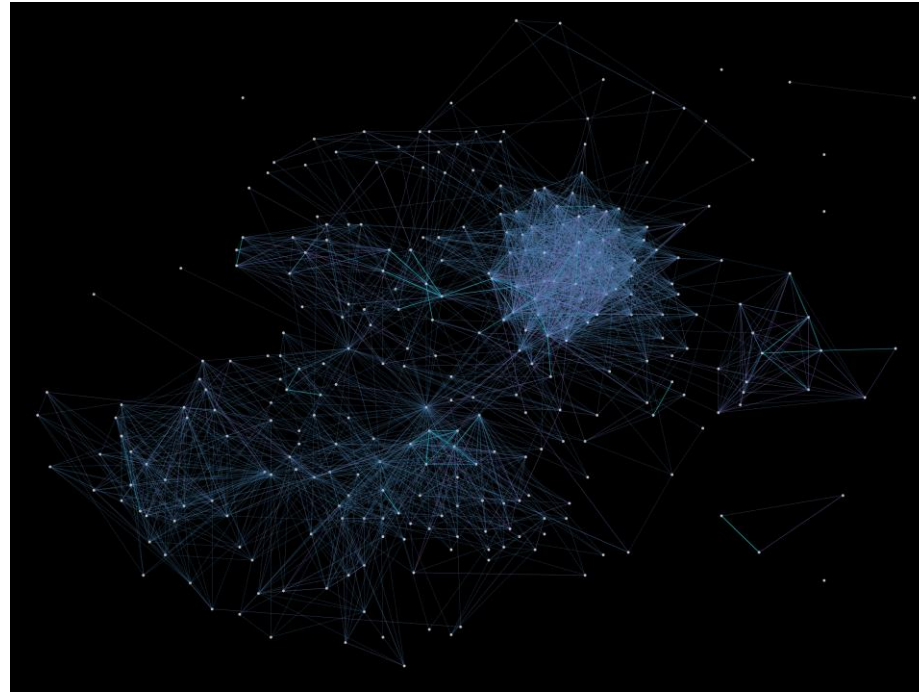
— Vinod Khosla —

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



DATA TYPES

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



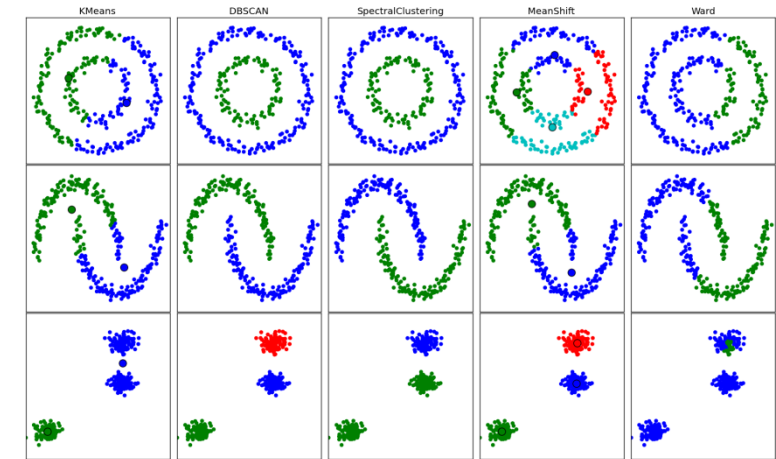
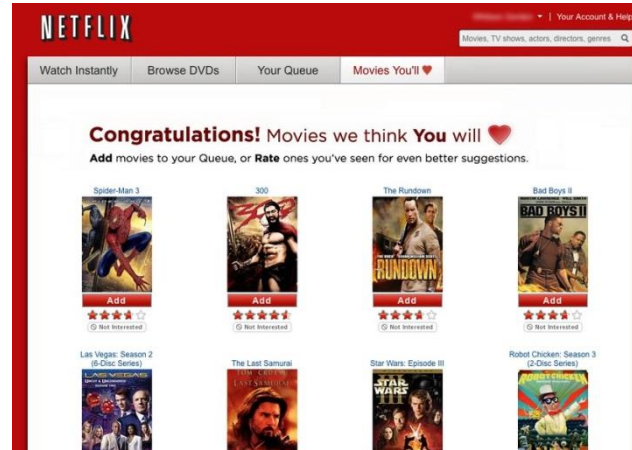
DATA TYPES?



Data Science <> Big Data

PROBLEM TYPES

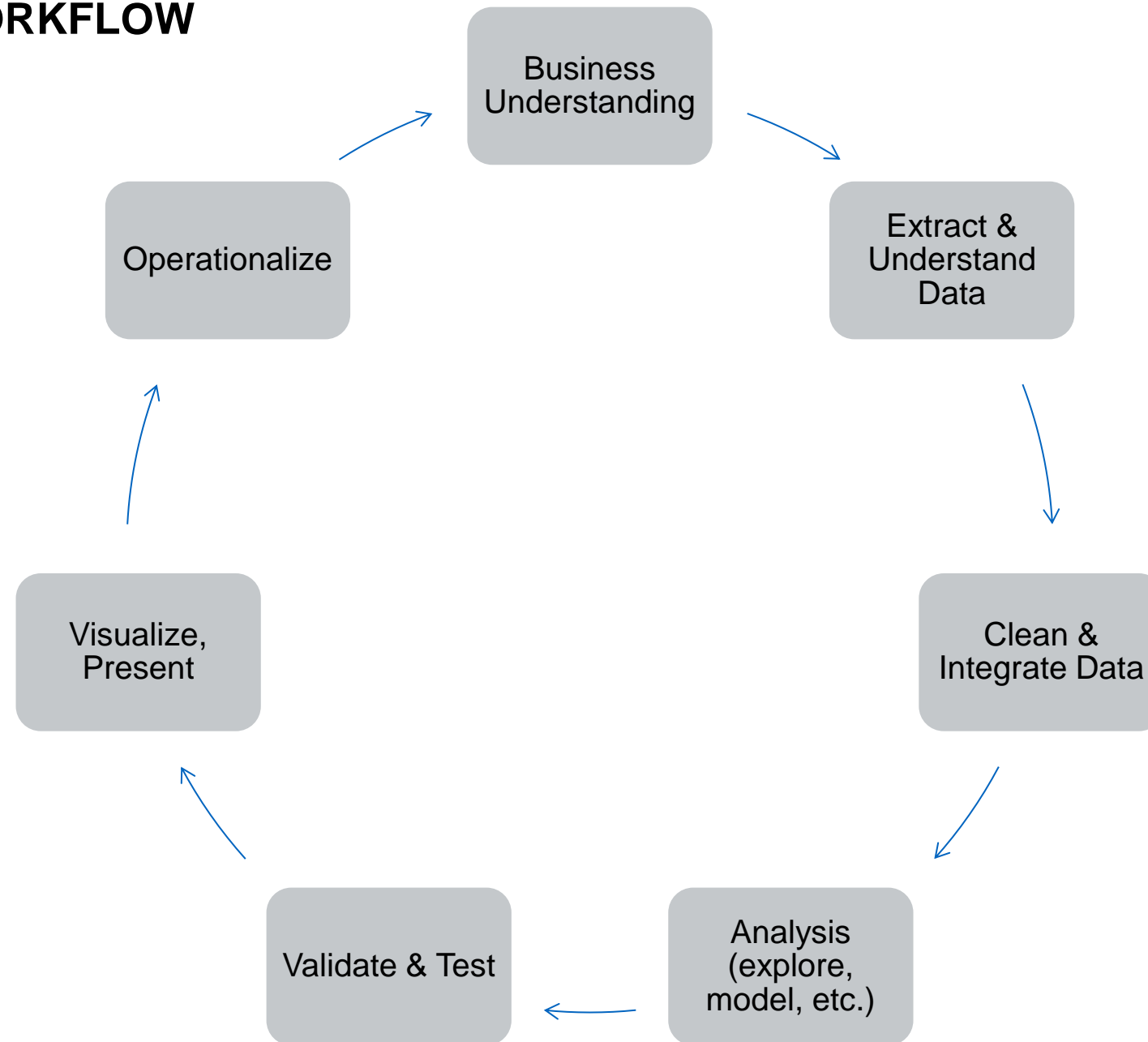
- Recommendations System
- Associations
- Relationships
- Structure
- Predictions
- Forecasting
- Optimization
- Vision
- Speech
- Audio



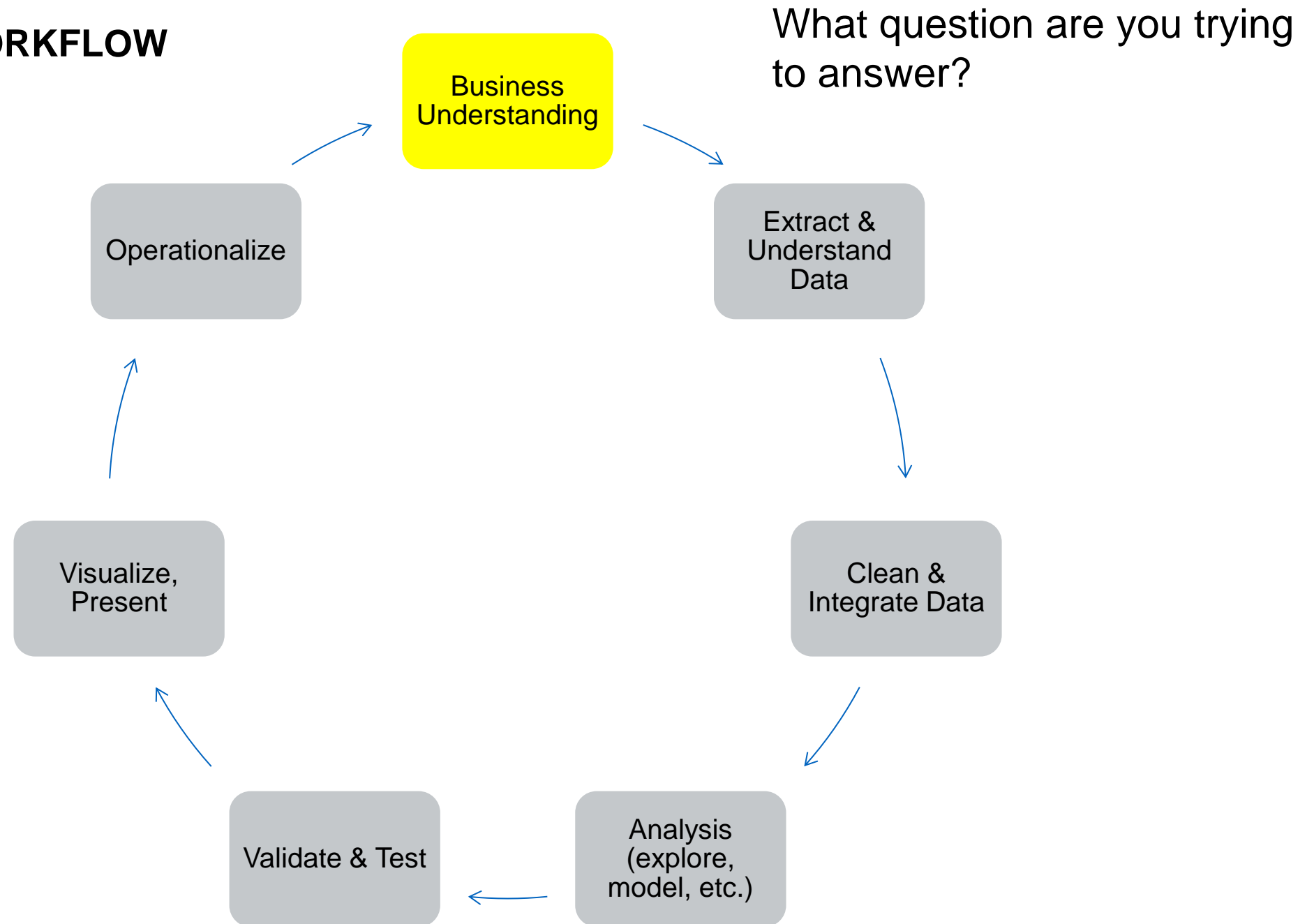
DATA SCIENCE WORKFLOW

A man in a blue checkered shirt is pointing at a whiteboard with handwritten notes and diagrams. Two other people are visible in the foreground, looking at the whiteboard. The whiteboard contains various handwritten notes and diagrams, including a flowchart with boxes labeled 'Profile' and 'Upcoming', a list of 'Issues', and a table with columns 'Sender_Id', 'Receiver_Id', 'Content', 'Stamp', and 'Flag'. The word 'Issues' is written in large blue letters. The word 'primary' is written in purple. The word 'hodo' is written in purple. The word 'primary' is written in purple. The word 'primary' is written in purple.

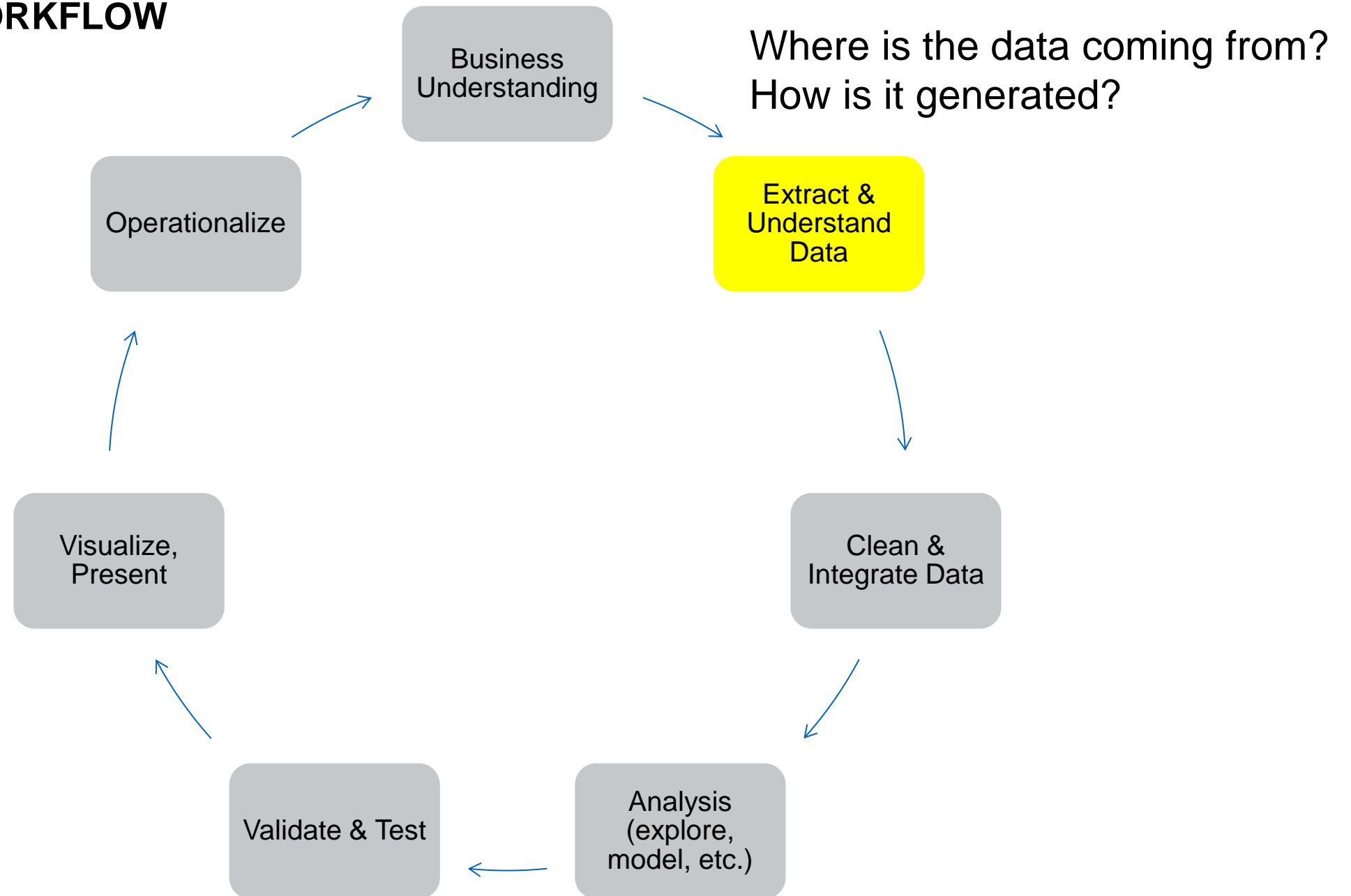
DATA SCIENCE WORKFLOW



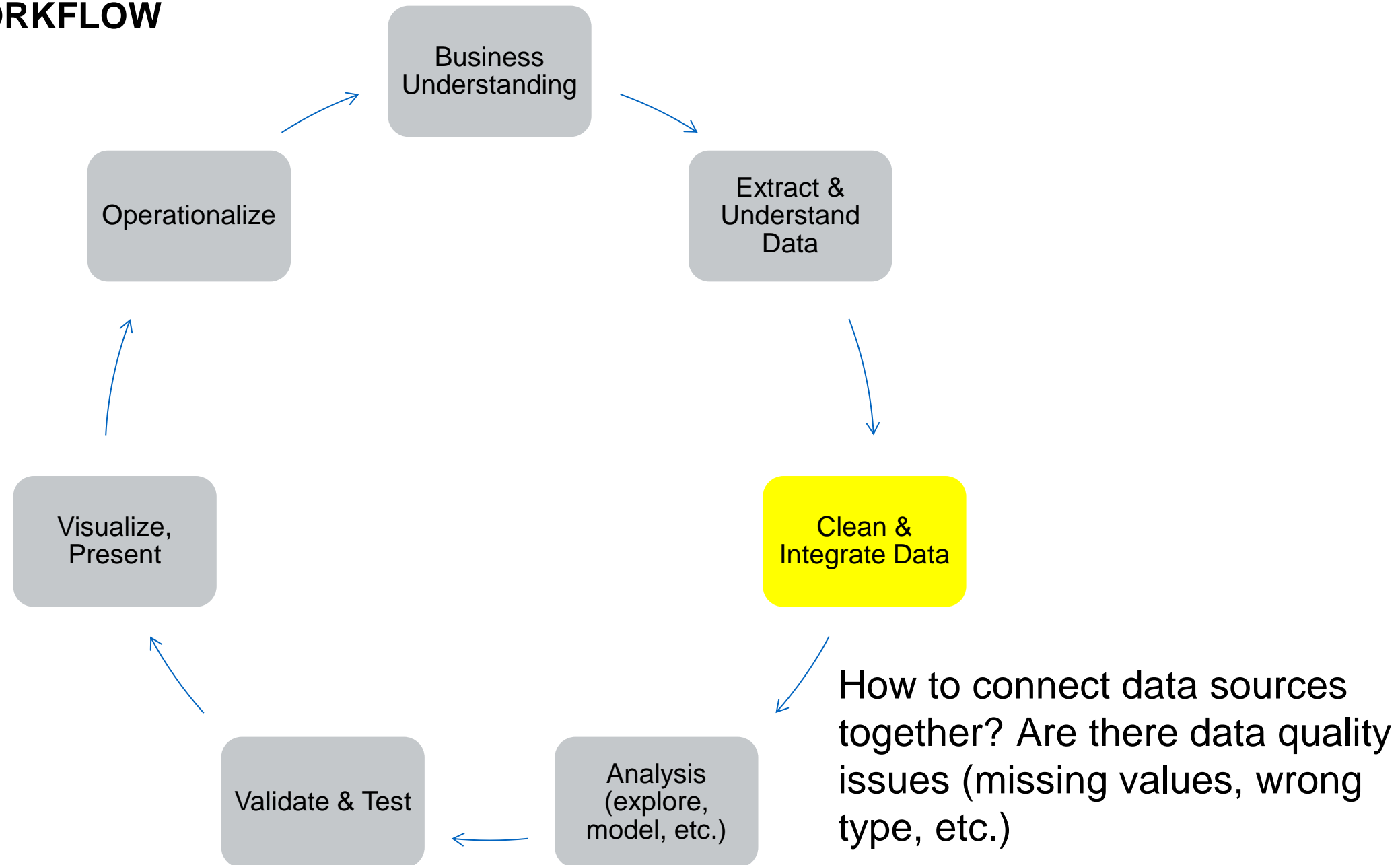
DATA SCIENCE WORKFLOW



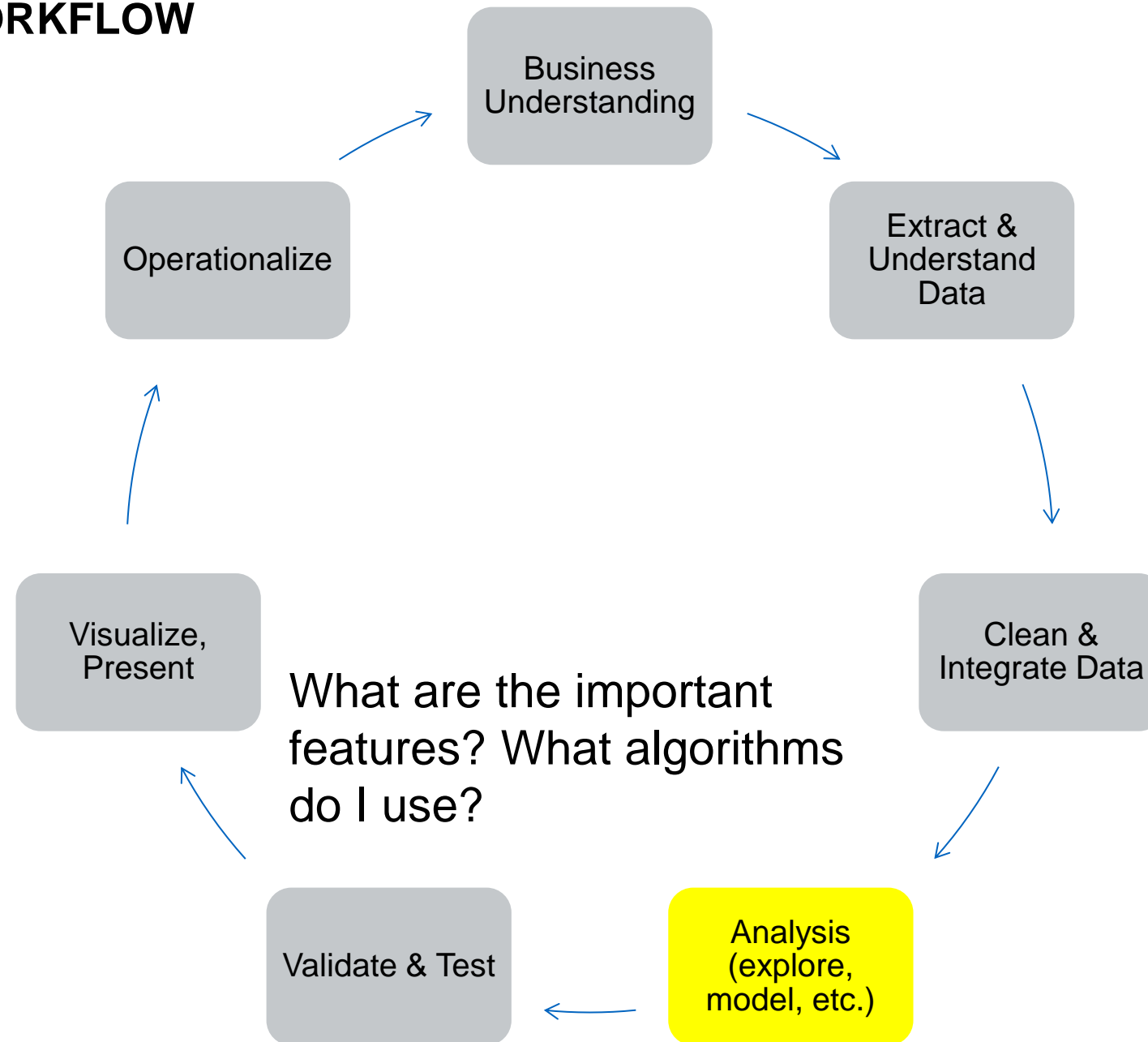
DATA SCIENCE WORKFLOW



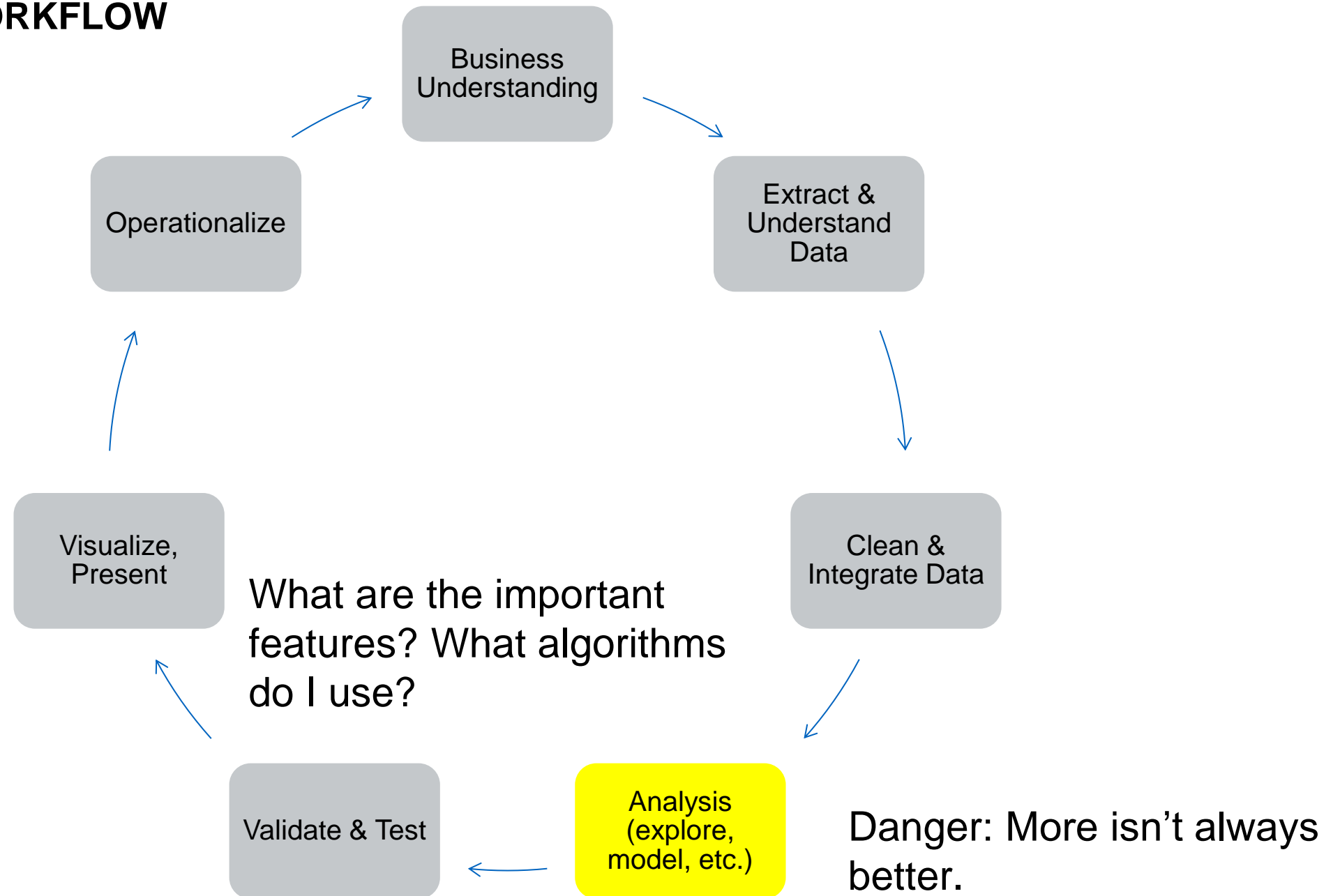
DATA SCIENCE WORKFLOW



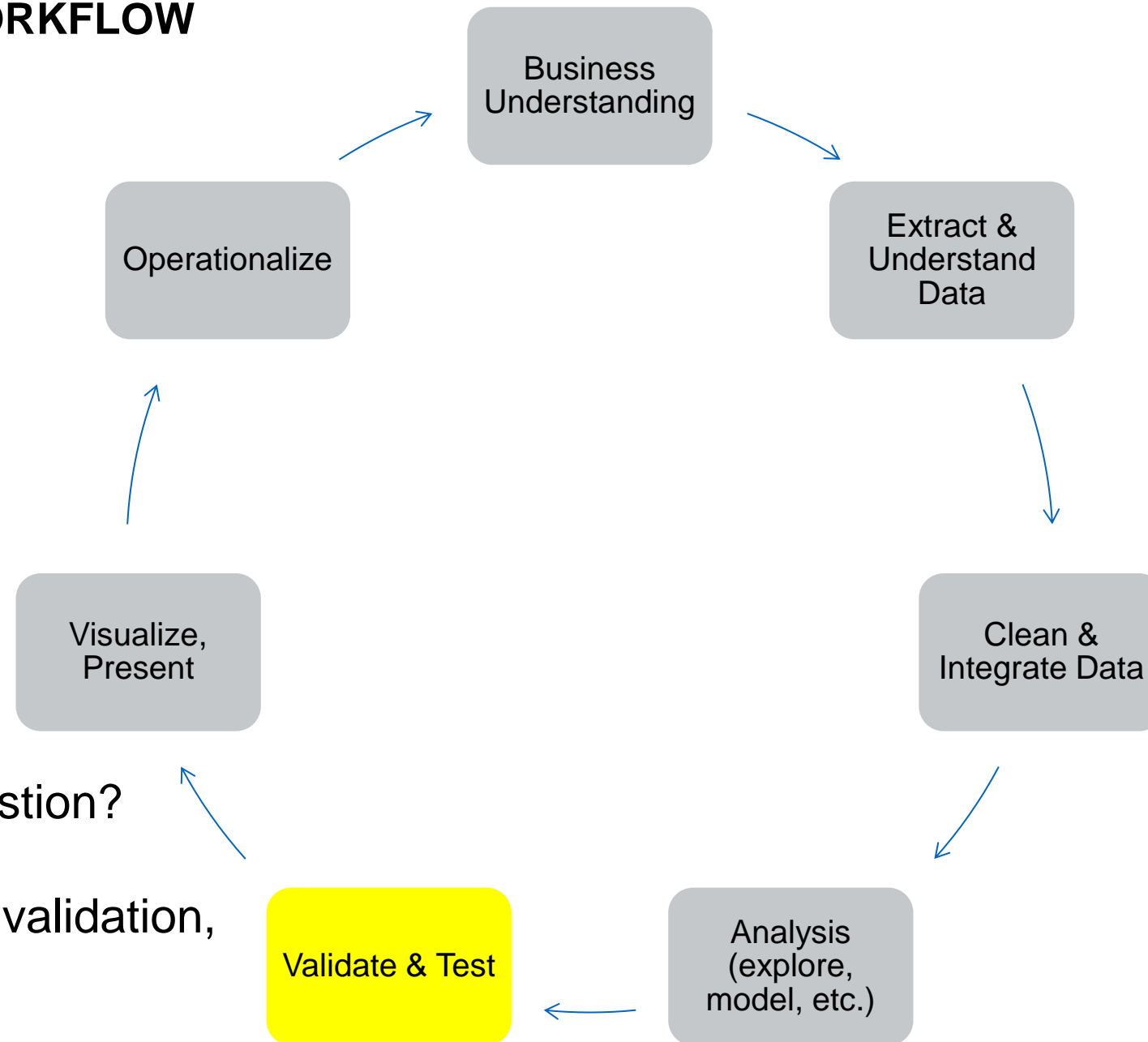
DATA SCIENCE WORKFLOW



DATA SCIENCE WORKFLOW

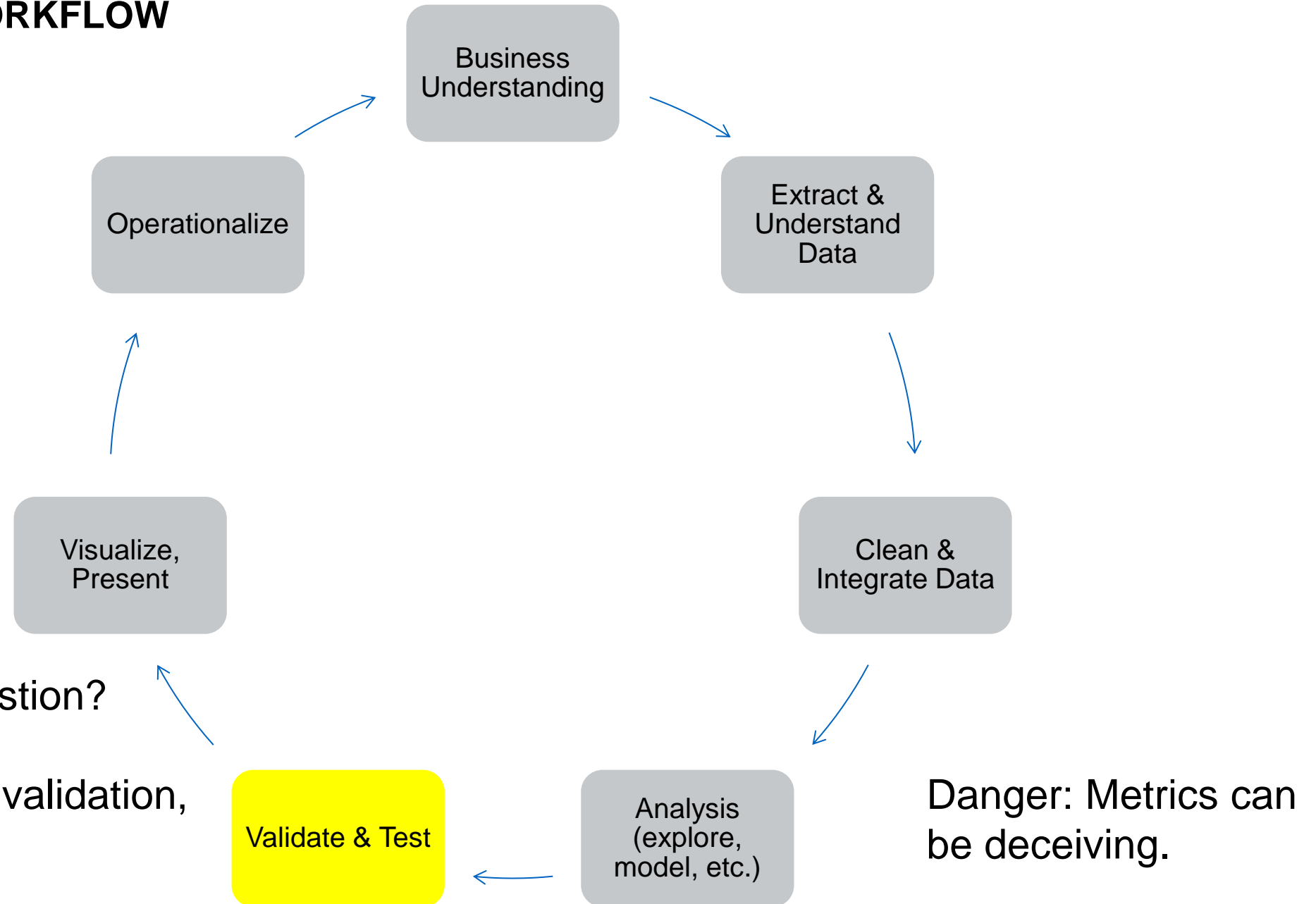


DATA SCIENCE WORKFLOW

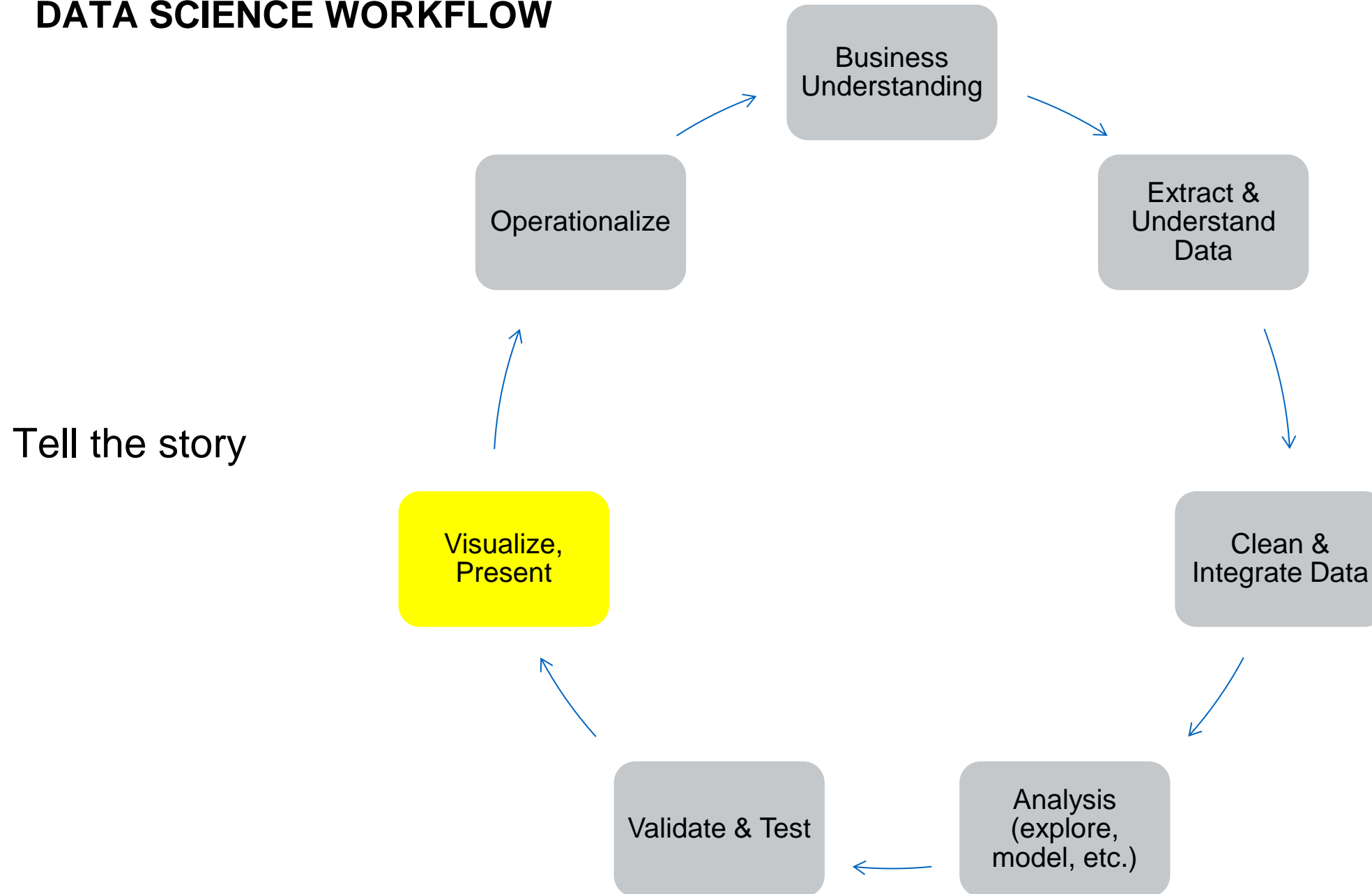


Did I answer the question?
How well?
Appropriate training, validation,
and testing

DATA SCIENCE WORKFLOW

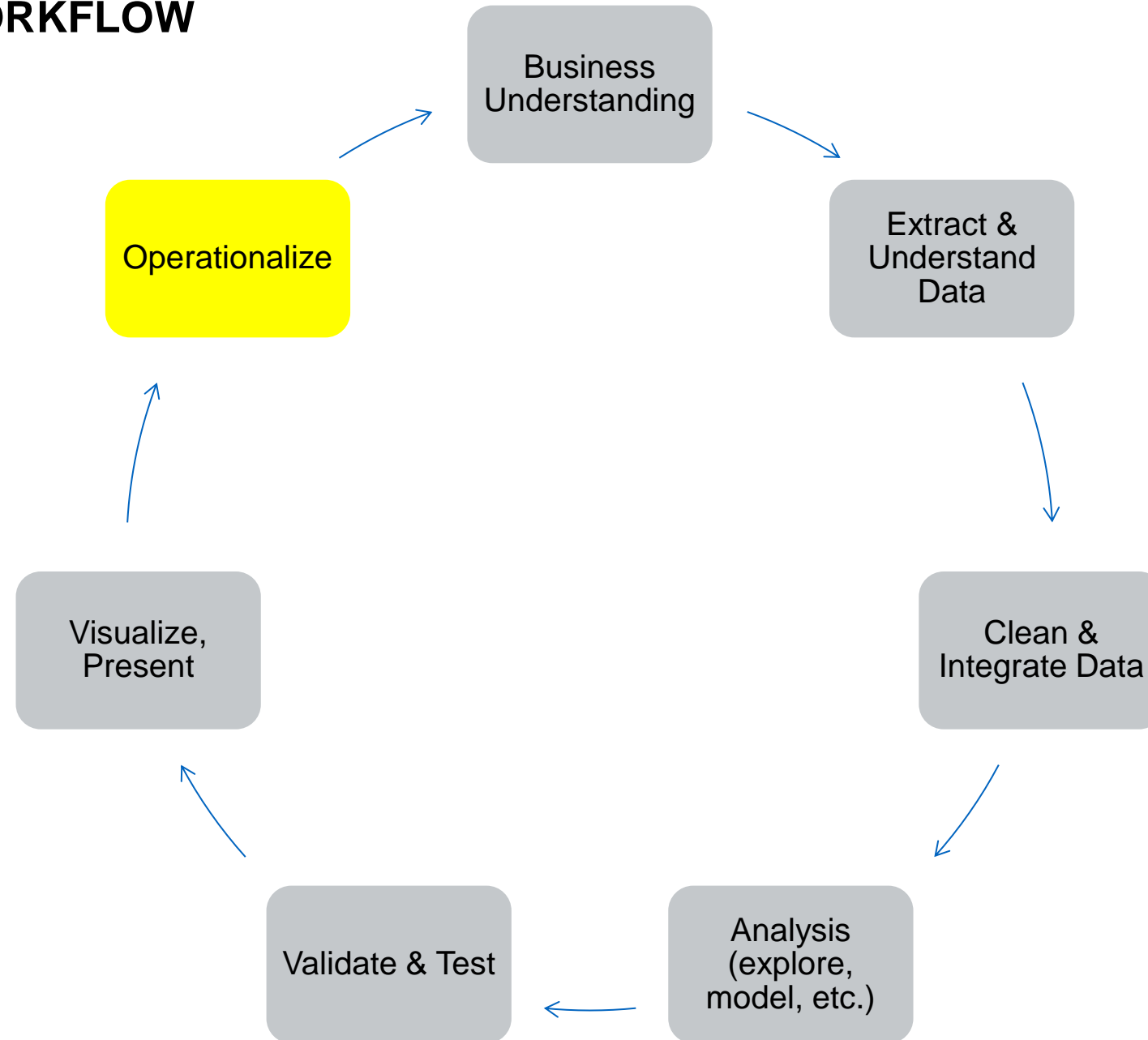


DATA SCIENCE WORKFLOW



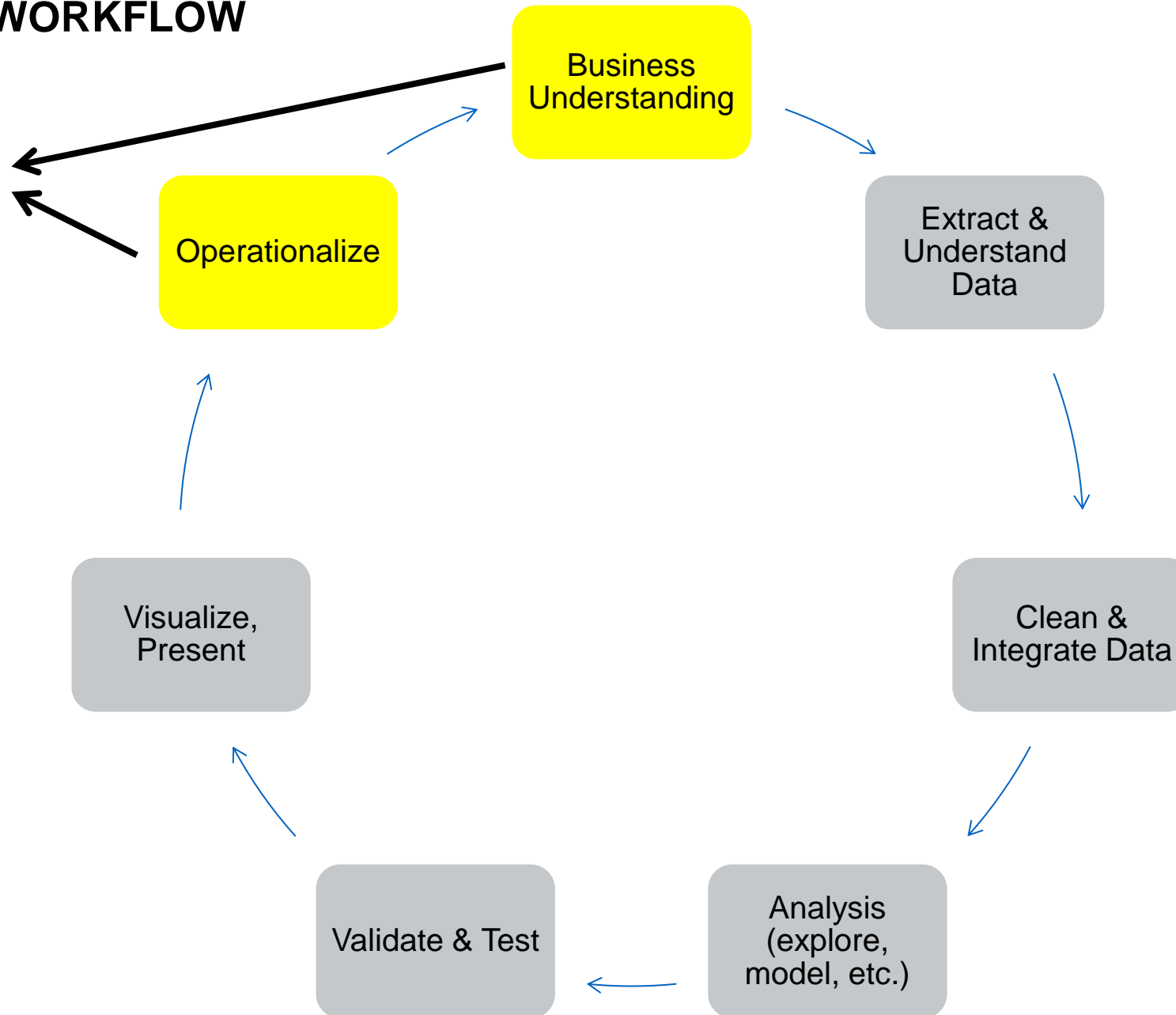
DATA SCIENCE WORKFLOW

How do I make this impactful?



DATA SCIENCE WORKFLOW

Make this connection
early



DISCUSSION Q&A