- 2- Data Analytics Lifecycle

Lecture 2:
Data Analytics Lifecycle

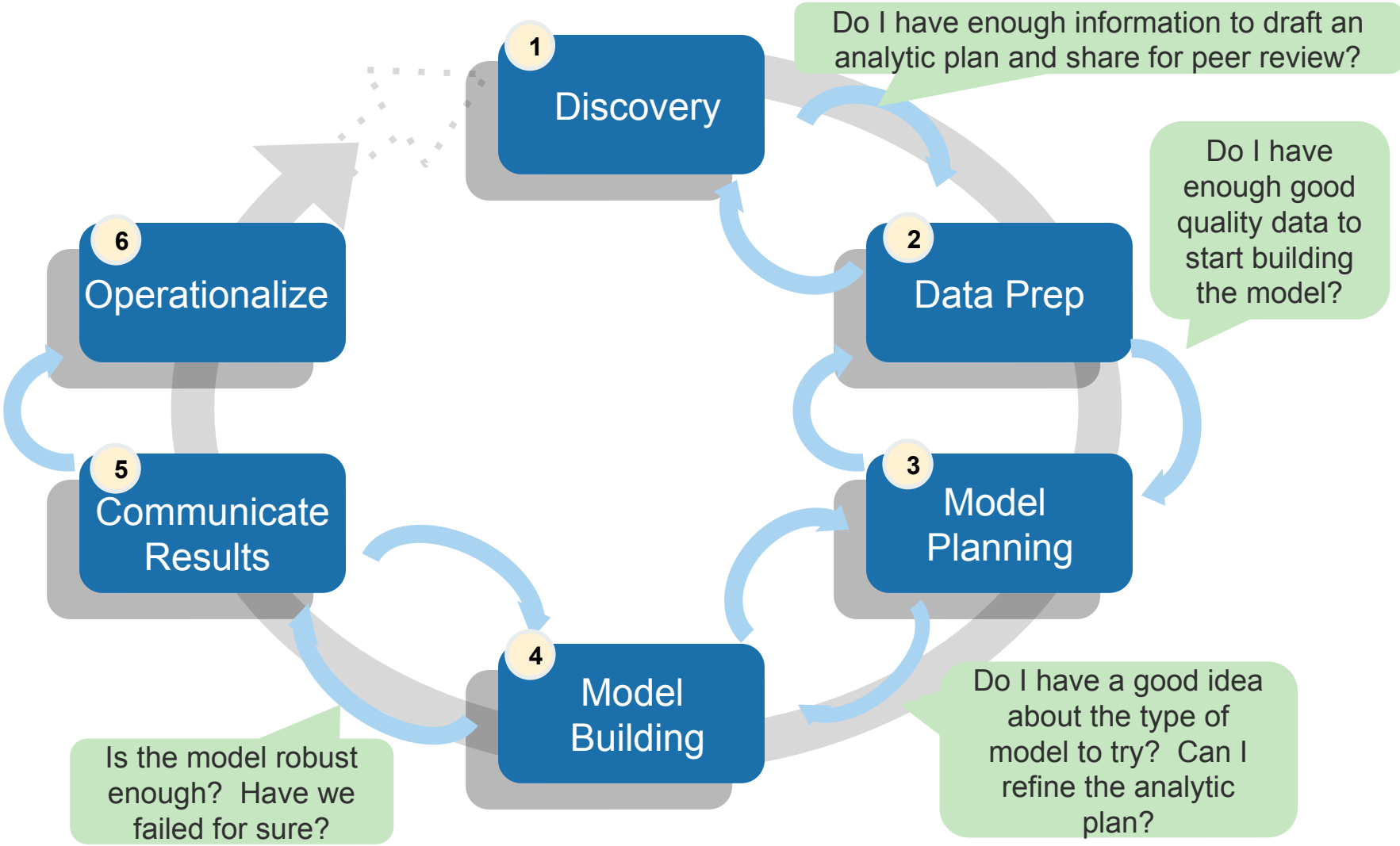
Upon completion of this module, you should be able to know about:

- Data Analytics Life Cycle
  - Discovery
  - Data preparation,
  - Model Planning,
  - Model Building,
  - Communicate Results,
  - Operationalize

Key Roles for a Successful Analytic Project

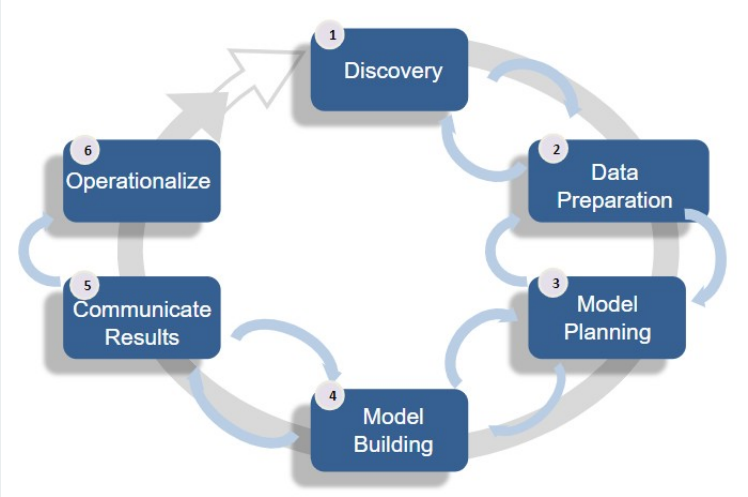
Role	Description
Business User	Someone who benefits from the end results and can consult and advise project team on value of end results and how these will be operationalized
Project Sponsor	Person responsible for the project, providing the motives for the project and core business problem, generally provides the funding and will assess the degree of value from the final outputs of the working team
Project Manager	Ensure key milestones and objectives are met on time and at expected quality.
Business Intelligence Analyst	Business domain expertise with deep understanding of the data, KPIs, key metrics and business intelligence from a reporting perspective
Data Engineer	Deep technical skills to assist with tuning SQL queries for data management, extraction and support data ingest to analytic sandbox
Database Administrator (DBA)	Database Administrator who provisions and configures database environment to support the analytical needs of the working team
Data Scientist	Provide subject matter expertise for analytical techniques, data modeling, applying valid analytical techniques to given business problems and ensuring overall analytical objectives are met

Data Analytics Lifecycle

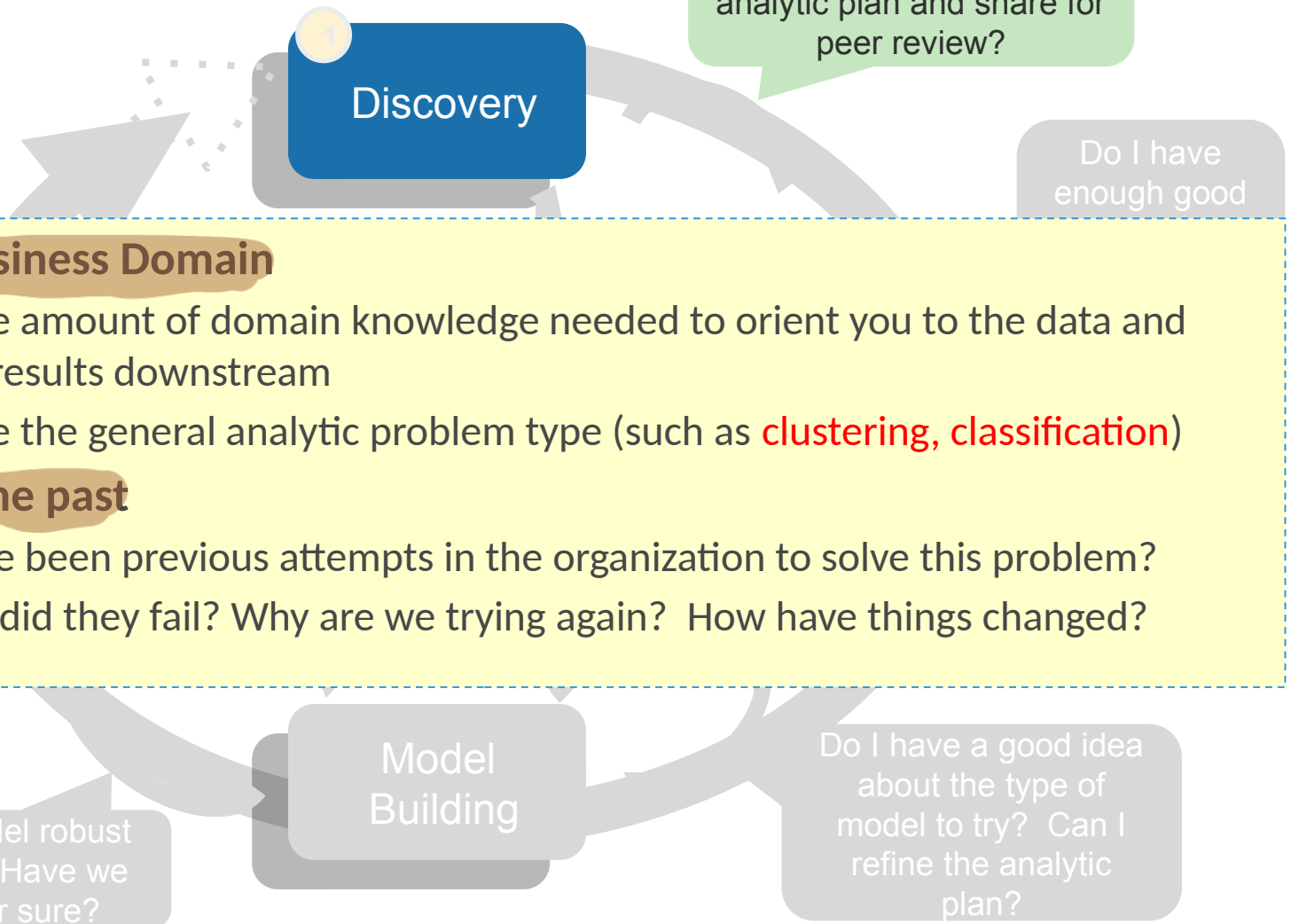


Data Analytics Lifecycle Phases

1. **Discovery**
2. Data Preparation
3. Model Planning
4. Model Building
5. Communicate Results
6. Operationalize

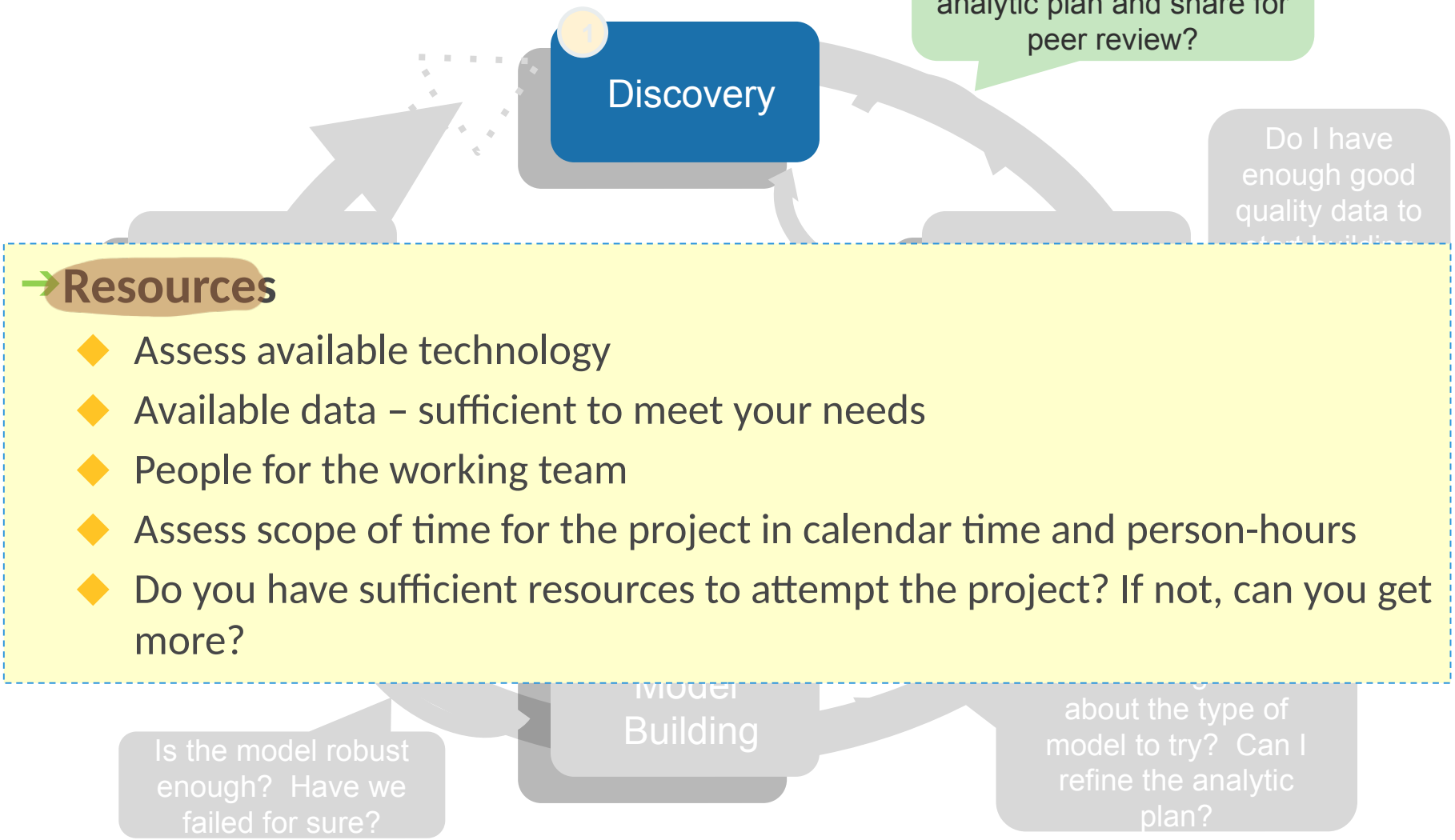


Data Analytics Lifecycle
Phase 1: Discovery 1/5



- **Learn the Business Domain**
  - Determine amount of domain knowledge needed to orient you to the data and interpret results downstream
  - Determine the general analytic problem type (such as clustering, classification)
- **Learn from the past**
  - Have there been previous attempts in the organization to solve this problem?
  - If so, why did they fail? Why are we trying again? How have things changed?

Data Analytics Lifecycle
Phase 1: Discovery 2/5



Data Analytics Lifecycle
Phase 1: Discovery 3/5

- **Frame the problem**.....*Framing is the process of stating the analytics problem to be solved*
  - *State the analytics problem, why it is important, and to whom*
  - *Identify key stakeholders and their interests in the project*
  - *Clearly articulate the current situation and pain points*
  - *Objectives – identify what needs to be achieved in business terms and what needs to be done to meet the needs*
    - *What is the goal? What are the criteria for success? What's "good enough"?*
    - *What is the failure criterion (when do we just stop trying or settle for what we have)?*
  - *Identify the success criteria, key risks, and stakeholders*

Tips for Interviewing the Analytics Sponsor

- Even if you are "given" an analytic problem you should work with clients to clarify and frame the problem
  - You're typically handed solutions, you need to identify the problem and their desired outcome

Sponsor Interview Tips

- Prepare for the interview – draft your questions, review with colleague, team
- Use open-ended questions, don't ask leading questions
- Probe for details, follow-up
- Don't fill every silence – give them time to think
- Let them express their ideas, don't put words in their mouth, let them share their feelings
- Ask clarifying questions, ask why – is that correct? Am I on target? Is there anything else?
- Use active listening – repeat it back to make sure you heard it correctly
- Don't express your opinions
- Be mindful of your body language and theirs – use eye contact, be attentive
- Minimize distractions
- Document what you heard and review it back with the sponsor

Tips for Interviewing the Analytics Sponsor

Interview Questions

- What is the business problem you're trying to solve?
- What is your desired outcome?
- Will the focus and scope of the problem change if the following dimensions change:
  - Time – analyzing 1 year or 10 years worth of data?
  - People – how would this project change this?
  - Risk – conservative to aggressive
  - Resources – none to unlimited (tools, tech, ...)
  - Size and attributes of Data
- What data sources do you have?
- What industry issues may impact the analysis?
- What timelines are you up against?
- Who could provide insight into the project? Consulted?
- Who has final say on the project?

Data Analytics Lifecycle
Phase 1: Discovery 4/5

Do I have enough information to draft an analytic plan and share for peer review?

1 Discovery

Do I have enough good ity data to t building e model?

- **Formulate Initial Hypotheses**
  - IH, H_1 , H_2, H_3, … H_n
  - Gather and assess hypotheses from stakeholders and domain experts
  - Preliminary data exploration to inform discussions with stakeholders during the hypothesis forming stage
- **Identify Data Sources – Begin Learning the Data**
  - Aggregate sources for previewing the data and provide high-level understanding
  - Review the raw data
  - Determine the structures and tools needed

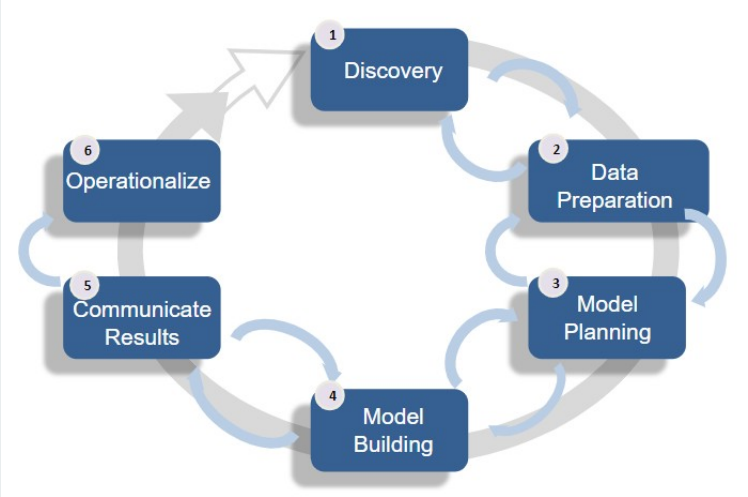
Data Analytics Lifecycle
Phase 1: Discovery Process Example 5/5

- Let's start a simple example starting with the question to discover if there is a relation between Quantity Sold (Output) and Price and Advertising (Input) in our shop.
- Given the data set we recorded below: Can we predict the Quantity to sell in the future if we set the figures/amount for Sale Price and Advertising?

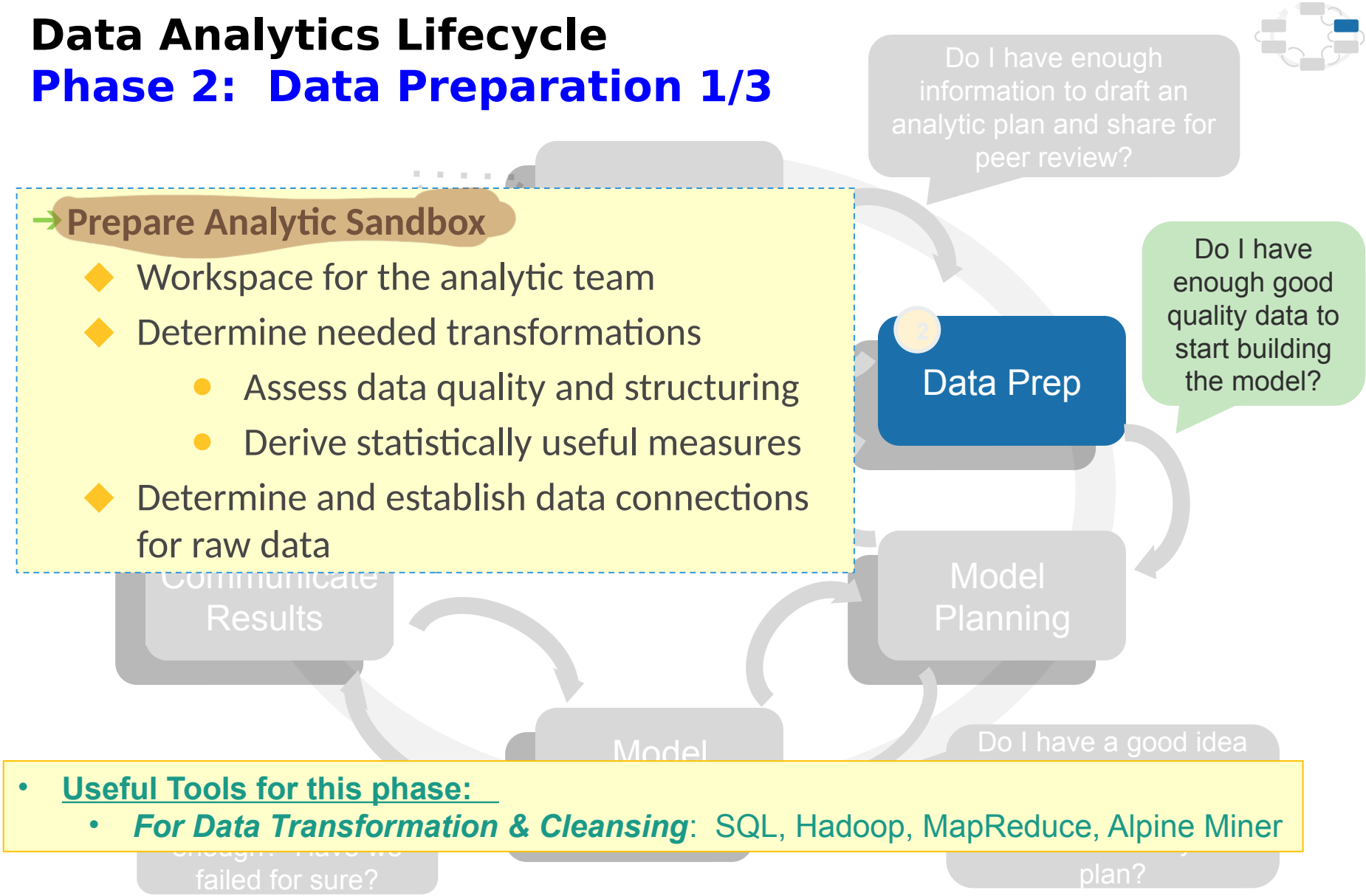
Initial Data Set		
Sales Price	Advertising	Quantity Sold
AED 10.00	AED 2,800.00	8500
AED 25.00	AED 200.00	4700
$4.09	$108.99	5800
AED 10.00		7400
AED 25.00	AED 3,200.00	6200
GBP 3.21	GBP 385.44	7300
AED 20.00	AED 900.00	5600
		?

Data Analytics Lifecycle Phases

1. Discovery
2. **Data Preparation**
3. Model Planning
4. Model Building
5. Communicate Results
6. Operationalize

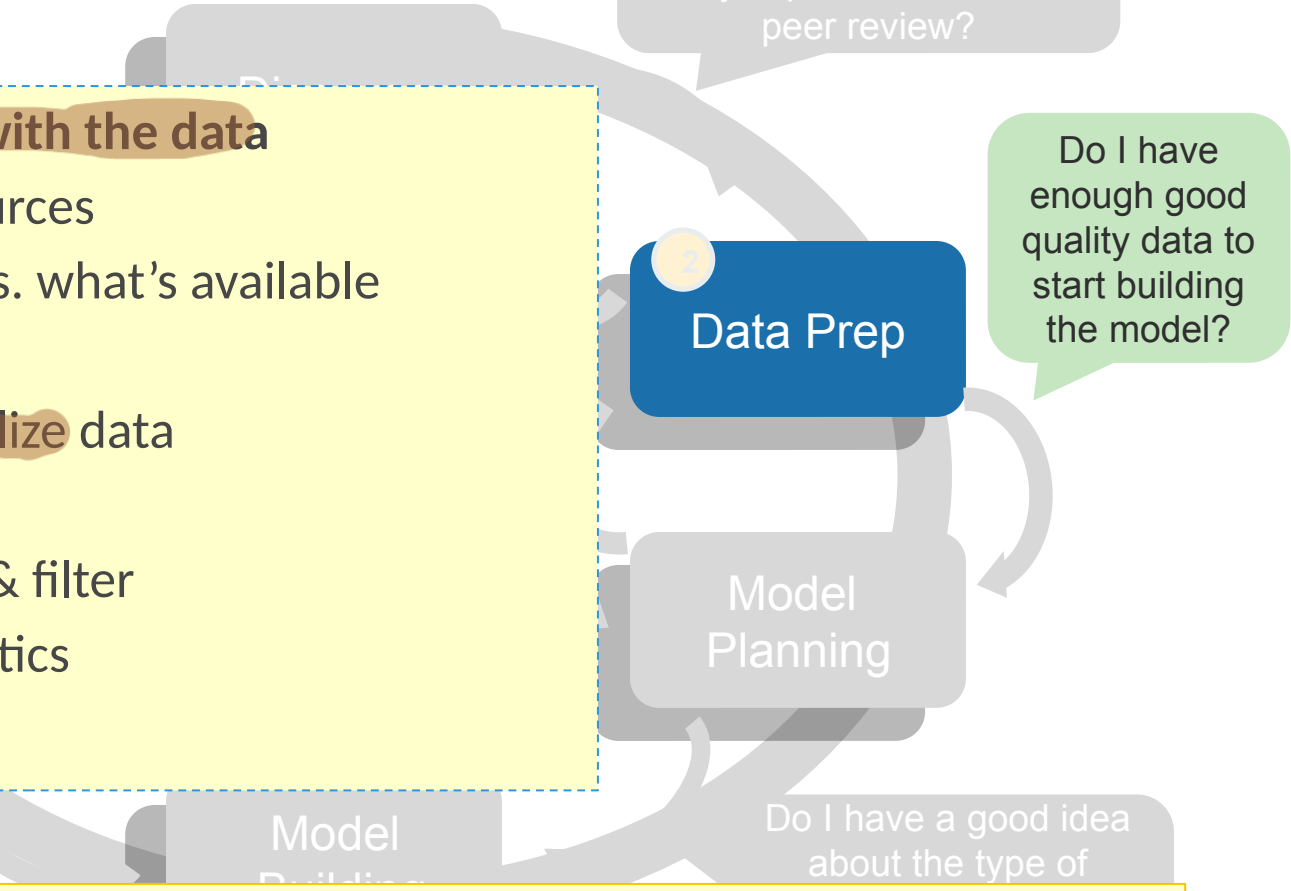


Data Analytics Lifecycle
Phase 2: Data Preparation 1/3



- **Useful Tools for this phase:**
  - ***For Data Transformation & Cleansing***: SQL, Hadoop, MapReduce, Alpine Miner

Data Analytics Lifecycle
Phase 2: Data Preparation 2/3



- **Familiarize yourself with the data**
  - List your data sources
  - What's needed vs. what's available
- **Data Conditioning**
  - Clean and normalize data
- **Survey & Visualize**
  - Overview, zoom & filter
  - Descriptive Statistics
  - Data Quality

- **Useful Tools for this phase:**
  - Descriptive Statistics on candidate variables for diagnostics & quality
  - ***Visualization***: R (base package, ggplot and lattice), GnuPlot, Ggobi/Rggobi, Spotfire, Tableau

Data Analytics Lifecycle
Phase 2: Data Preparation Example 3/3

Let's check the data in the Quantity sold data set:

- Missing Advertising value in Row 4
- Inconsistent currencies Sales Prices in AED, $ and GBP?

Initial Data Set

Sales Price	Advertising	Quantity Sold
AED 10.00	AED 2,800.00	8500
AED 25.00	AED 200.00	4700
$4.09	$108.99	5800
AED 10.00		7400
AED 25.00	AED 3,200.00	6200
GBP 3.21	GBP 385.44	7300
AED 20.00	AED 900.00	5600
		?

Prepared Data

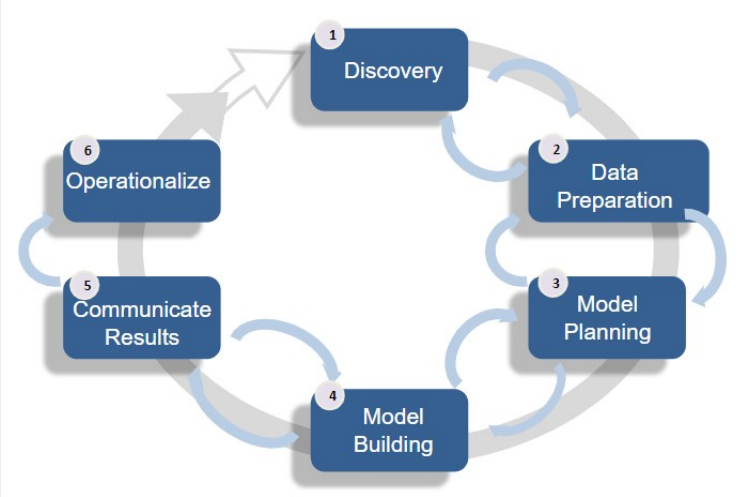
Price	Advertising	Quantity Sold
AED 10.00	AED 2,800.00	8500
AED 25.00	AED 200.00	4700
AED 15.00	AED 400.00	5800
AED 10.00	AED 500.00	7400
AED 25.00	AED 3,200.00	6200
AED 15.00	AED 1,800.00	7300
AED 20.00	AED 900.00	5600
		?

Define the Type of dependent and independent variables:

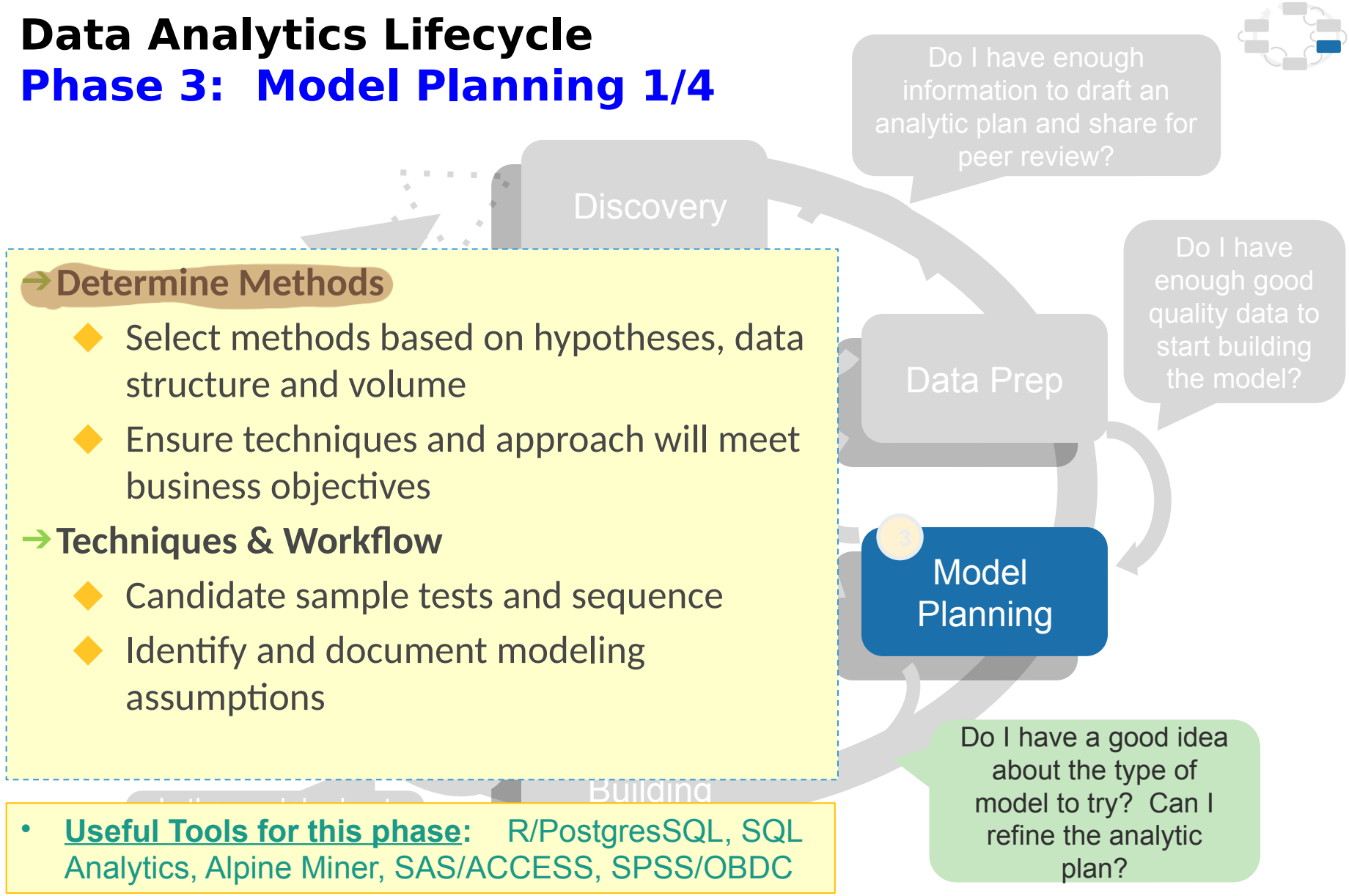
- **Quantity Sold :** Predictor Variable (also called dependent variable)
- **Sales Prices & Advertising:** Explanatory Variable (Also called independent variables).

Data Analytics Lifecycle Phases

1. Discovery
2. Data Preparation
3. **Model Planning**
4. Model Building
5. Communicate Results
6. Operationalize



Data Analytics Lifecycle
Phase 3: Model Planning 1/4

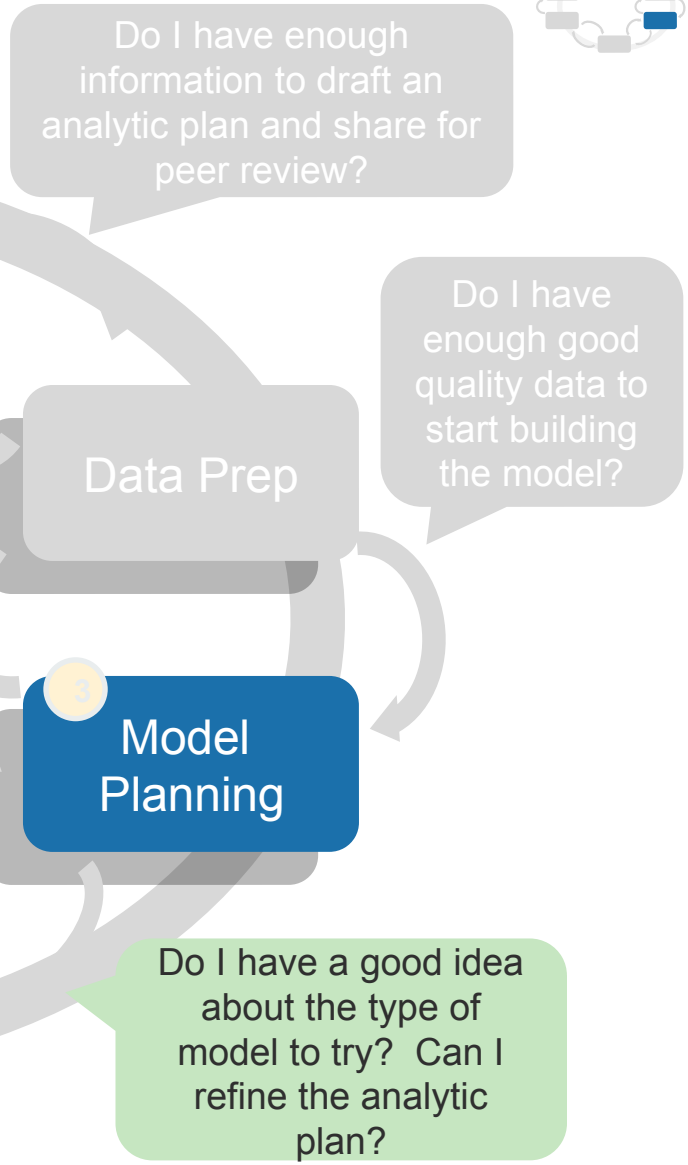


- **Determine Methods**
  - Select methods based on hypotheses, data structure and volume
  - Ensure techniques and approach will meet business objectives
- **Techniques & Workflow**
  - Candidate sample tests and sequence
  - Identify and document modeling assumptions

- **Useful Tools for this phase:** R/PostgresSQL, SQL Analytics, Alpine Miner, SAS/ACCESS, SPSS/OBDC

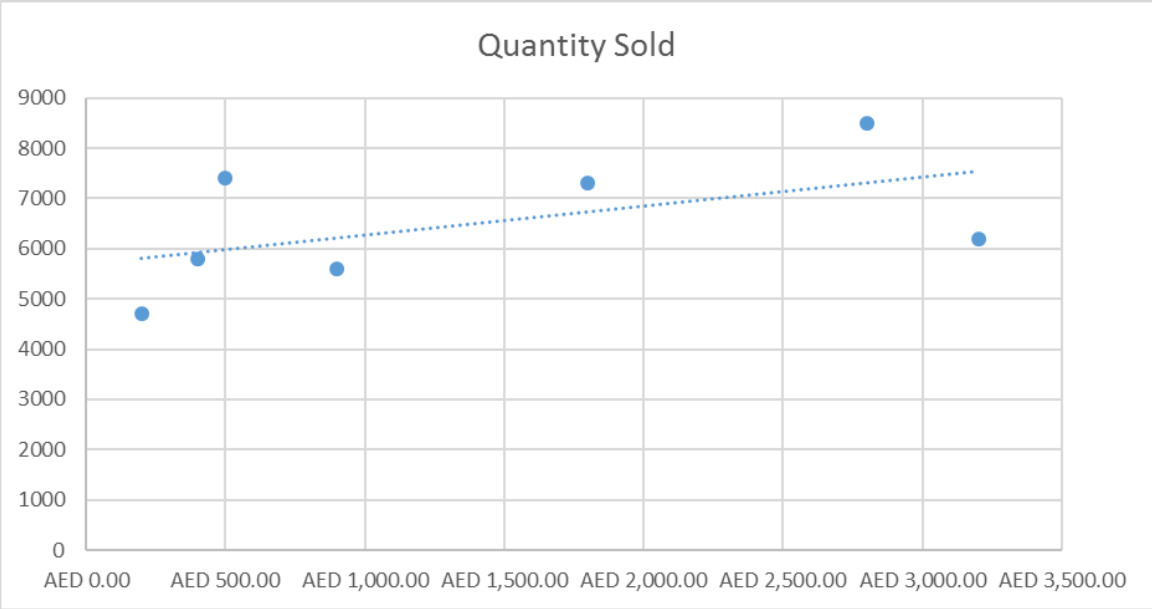
Data Analytics Lifecycle
Phase 3: Model Planning 2/4

- Data Exploration
- Variable Selection (attributes)
  - Inputs from stakeholders and domain experts
  - leverage a technique for dimensionality reduction
  - Iterative testing to confirm the most significant variables
- Model Selection
  - Choose technique based on the end goal

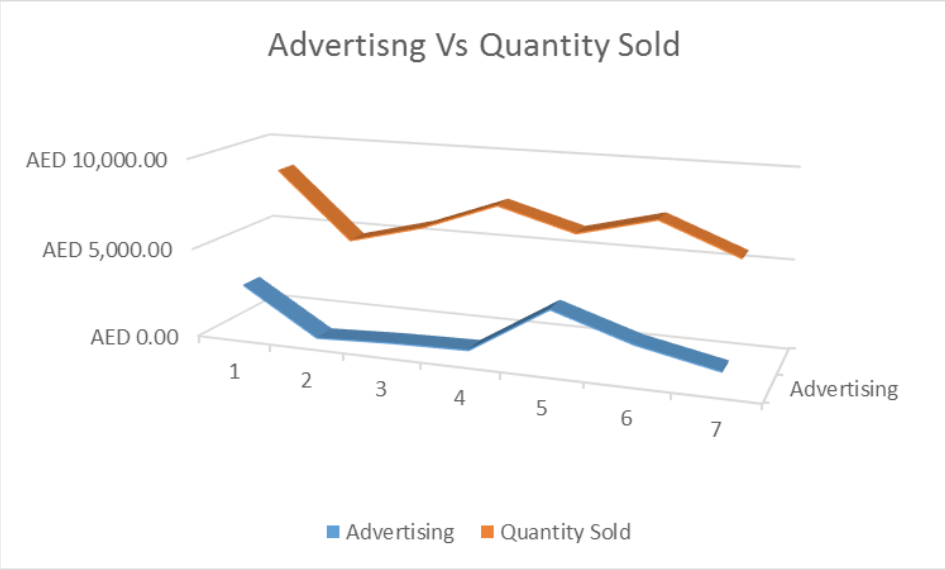


Data Analytics Lifecycle
Phase 3: Model Planning 3/3

Advertising	Quantity Sold
AED 2,800.00	8500
AED 200.00	4700
AED 400.00	5800
AED 500.00	7400
AED 3,200.00	6200
AED 1,800.00	7300
AED 900.00	5600

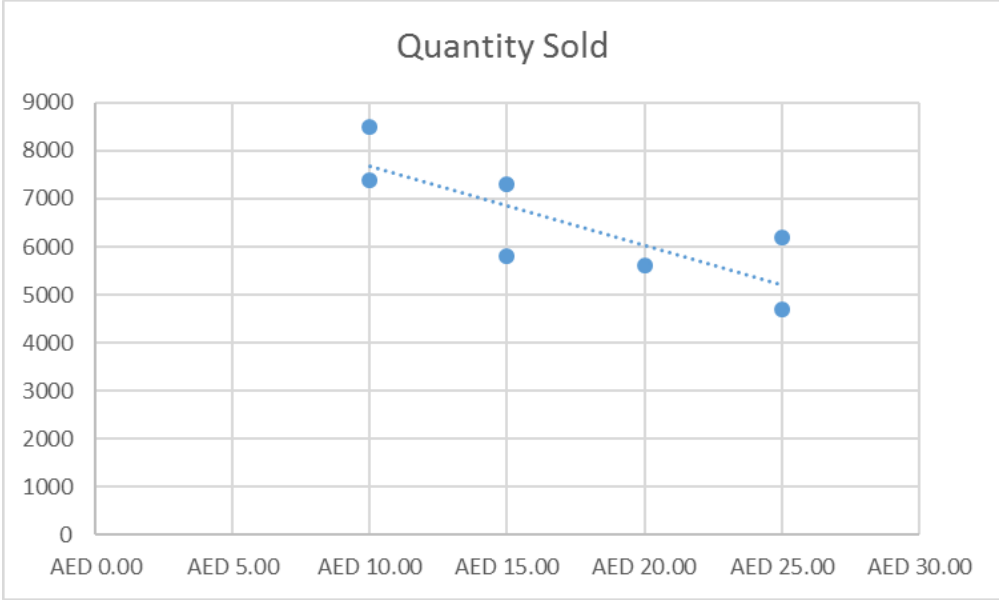


Correlation (Advertising, Quantity sold)=0.537547



Data Analytics Lifecycle
Phase 3: Model Planning 4/4

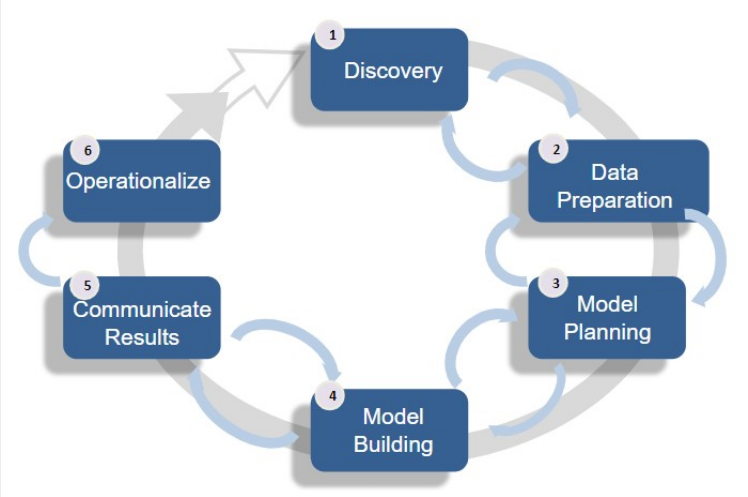
Price	Quantity Sold
AED 10.00	8500
AED 25.00	4700
AED 15.00	5800
AED 10.00	7400
AED 25.00	6200
AED 15.00	7300
AED 20.00	5600



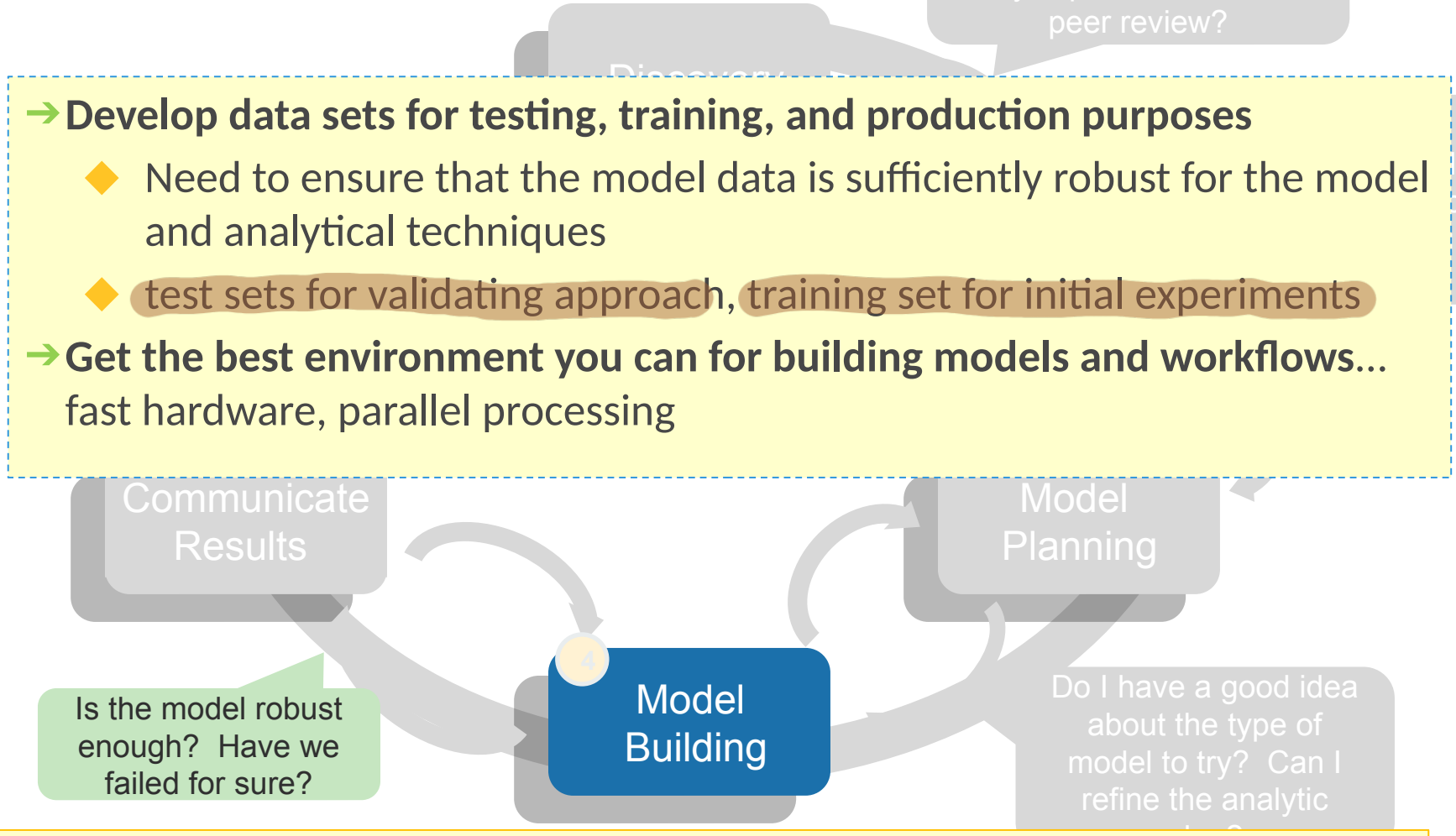
Correlation (Price, Quantity sold)= -0.80845

Data Analytics Lifecycle Phases

1. Discovery
2. Data Preparation
3. Model Planning
4. **Model Building**
5. Communicate Results
6. Operationalize



Data Analytics Lifecycle
Phase 4: Model Building

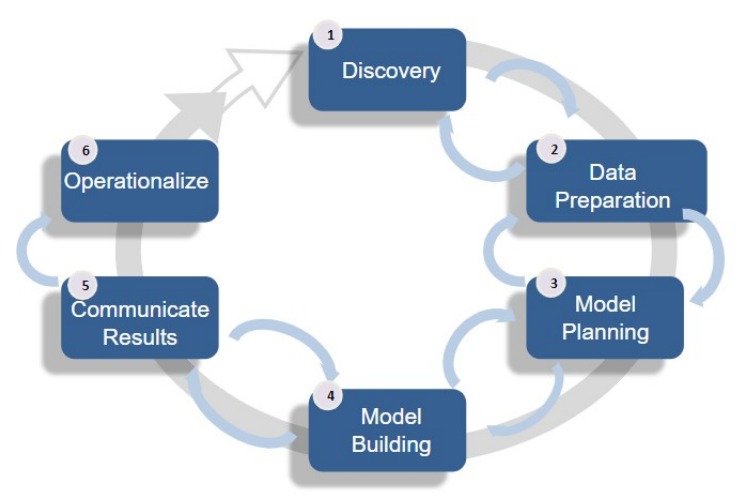


- **Develop data sets for testing, training, and production purposes**
  - Need to ensure that the model data is sufficiently robust for the model and analytical techniques
  - test sets for validating approach, training set for initial experiments
- **Get the best environment you can for building models and workflows…** fast hardware, parallel processing

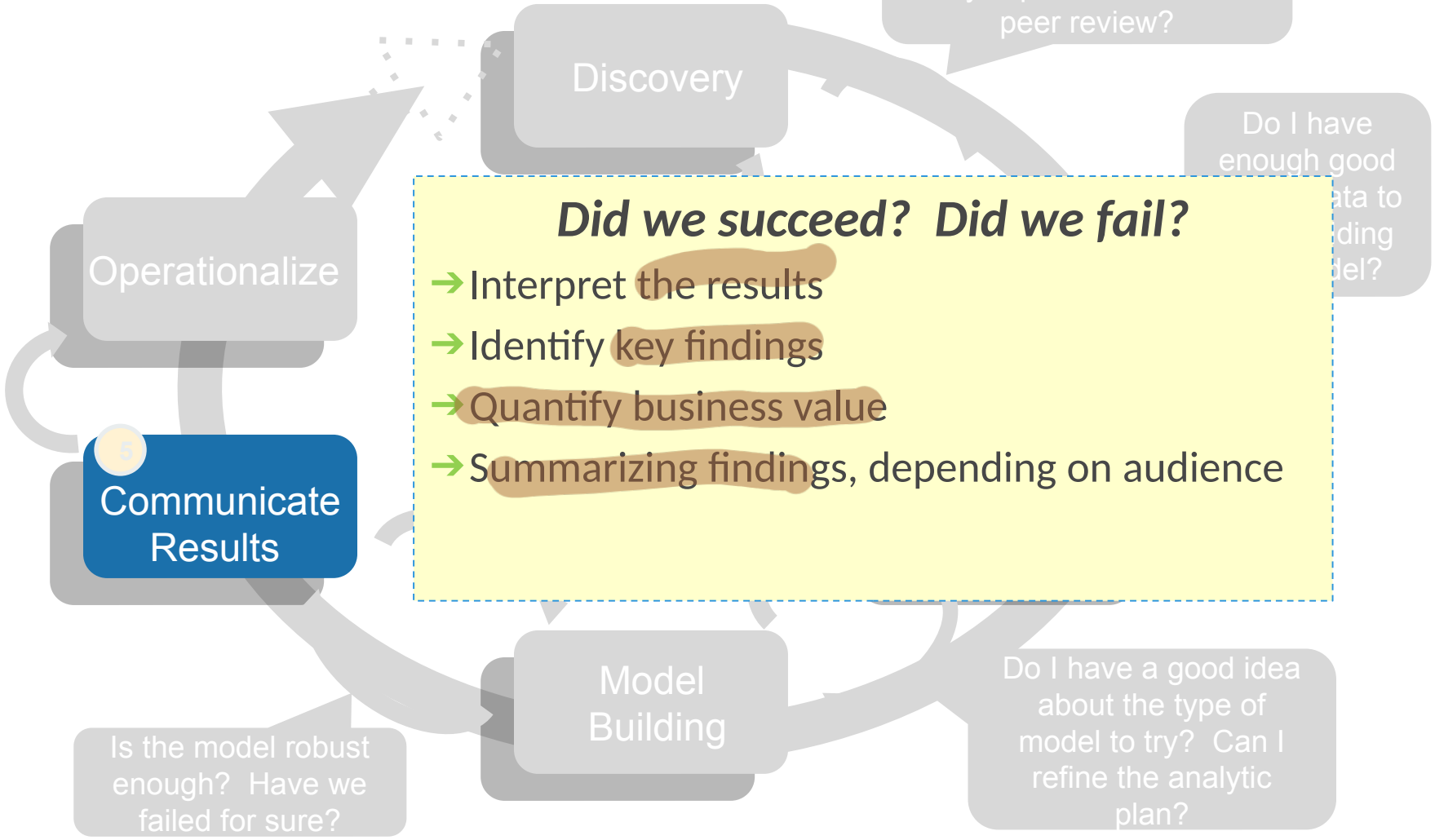
- **Useful Tools for this phase**: R, PL/R, SQL, Alpine Miner, SAS Enterprise Miner

Data Analytics Lifecycle Phases

1. Discovery
2. Data Preparation
3. Model Planning
4. Model Building
5. **Communicate Results**
6. Operationalize

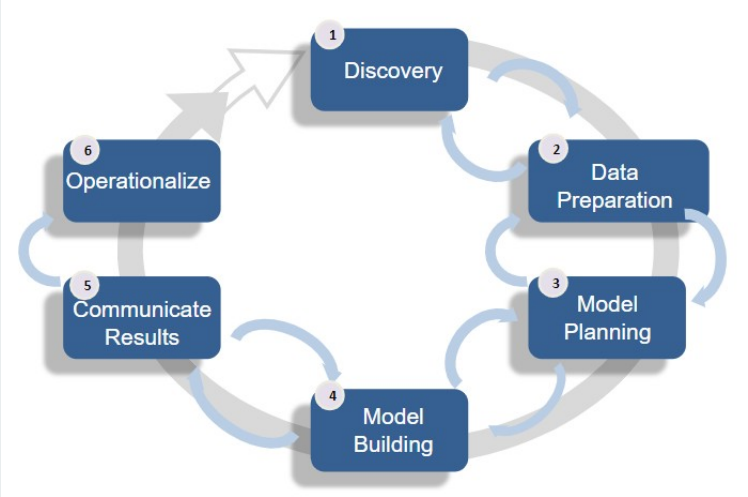


Data Analytics Lifecycle
Phase 5: Communicate Results

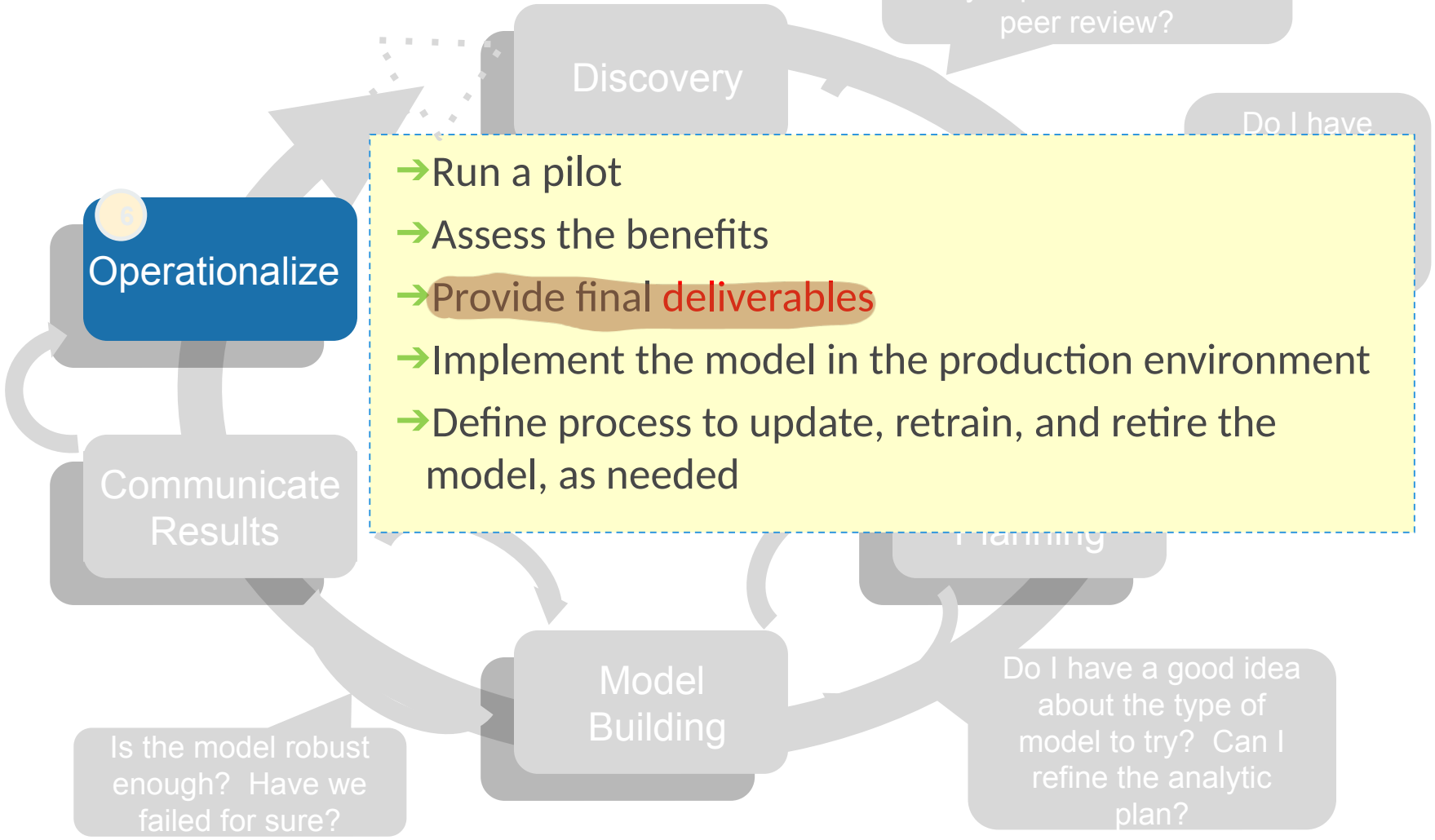


Data Analytics Lifecycle Phases

1. Discovery
2. Data Preparation
3. Model Planning
4. Model Building
5. Communicate Results
6. **Operationalize**



Data Analytics Lifecycle
Phase 6: Operationalize



- ➔Run a pilot
- ➔Assess the benefits
- ➔Provide final deliverables
- ➔Implement the model in the production environment
- ➔Define process to update, retrain, and retire the model, as needed

Data Analytics Project

Mini Case Study:
Churn Prediction for Retail Banking

Analytic Plan

Mini Case Study: Churn Prediction for Retail Banking

Components of Analytic Plan	Retail Banking: Yoyodyne Bank
Phase 1: Discovery Business Problem Framed	How do we identify churn/no churn for a customer?
Phase 2: Data prep	5 months of customer account history.
Phase 3: Model Planning - Analytic Technique	regression to identify most influential factors predicting churn.
Phase 4: Model Execution	Apply the model on data
Phase 5: Result & Key Findings	Once customers stop using their accounts for gas and groceries, they will soon erode their accounts and churn. If customers use their debit card fewer than 5 times per month, they will leave the bank within 60 days.
Business Impact	If we can target customers who are high-risk for churn, we can reduce customer attrition by 25%. This would save $3 million in lost of customer revenue and avoid $1.5 million in new customer acquisition costs each year.

Check Your Knowledge

- In which phase would you expect to invest most of your project time and why? Where would expect to spend the least time?
- What are the benefits of doing a pilot program before a full scale rollout of a new analytical methodology? Discuss this in the context of the mini case study.
- What kinds of tools would be used in the following phases, and for which kinds of use scenarios?
  - Phase 2: Data Preparation
  - Phase 4: Model Execution
- Now that you have completed the analytical project at Yoyodyne, you have an opportunity to repurpose this approach for an online eCommerce company. What phases of the lifecycle do you need to focus on to identify ways to do this?