Module 2

Data Analytics Lifecycle

Data Analytics Lifecycle Topics

- Data Analytics Lifecycle
- Roles for a Successful Analytics Project
- Case Study to apply the data analytics lifecycle

Data Analytics Lifecycle

Objectives:

- Apply the Data Analytics Lifecycle to a case study scenario
- Frame a business problem as an analytics problem
- Identify the four main deliverables in an analytics project

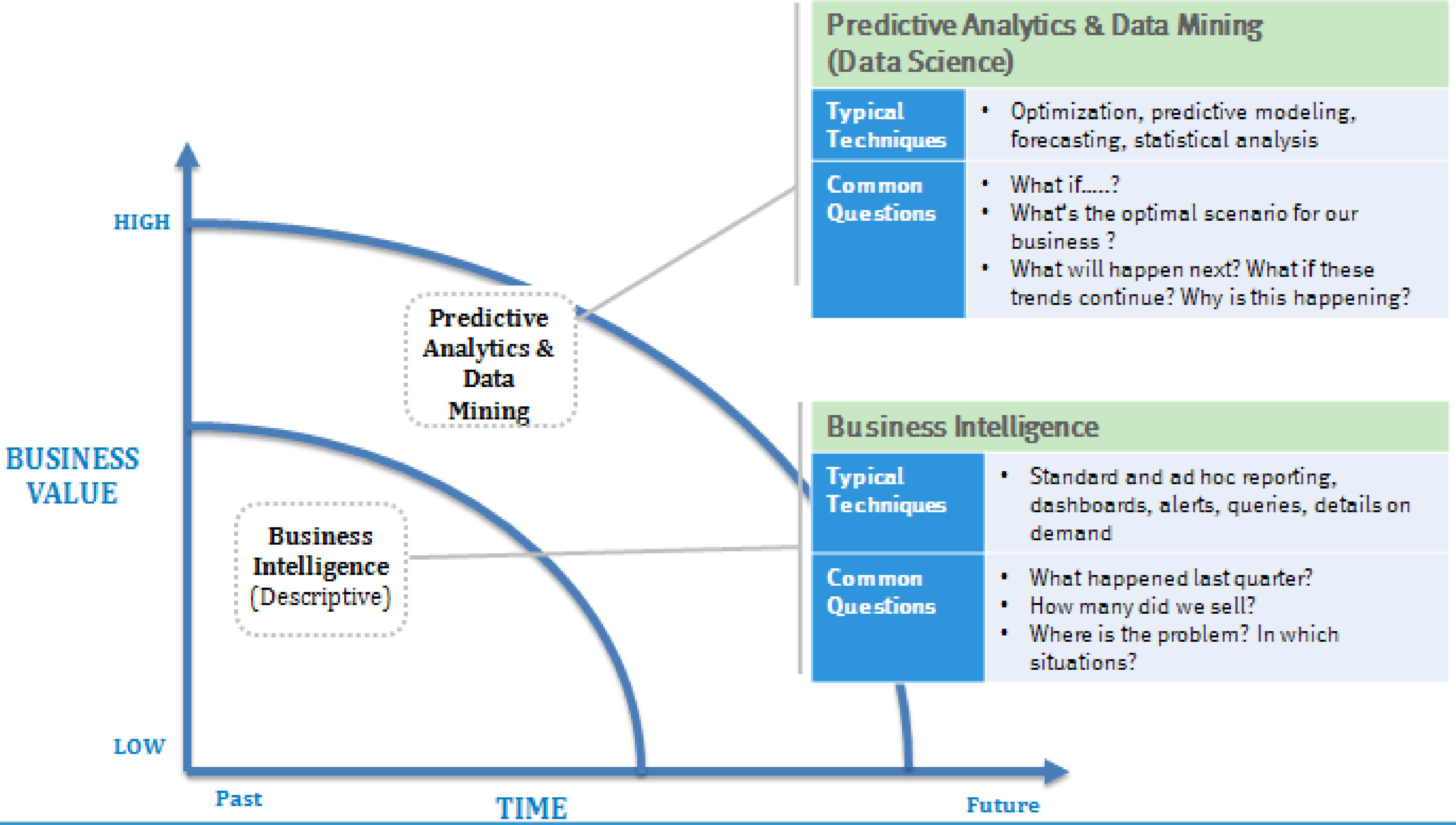
How to Approach Your Analytics Problems

- How do you currently approach your analytics problems?
- Do you follow a methodology or some kind of framework?
- How do you plan for an analytic project?

Advantages of Using the Data Analytics Lifecycle

- Focus your time
- Well defined process enables you to break down complex problems into smaller steps.
- Many times in the rush to begin collecting and analyzing the data, people do not spend sufficient time in planning and scope the amount of work involved or framing the business problem
- Ensure that you establish a comprehensive, repeatable method for conducting analysis.
- Creating and documenting a process will help demonstrate accuracy in your findings.
- Enable better transition to members of the cross-functional analytic teams
  - Repeatable
  - Scale to additional analysts
  - Support validity of findings

Guidance Processes in Data Science Projects


HIGH
BUSINESS VALUE
LOW
Predictive Analytics & Data Mining
Business Intelligence (Descriptive)
Past
TIME
Future

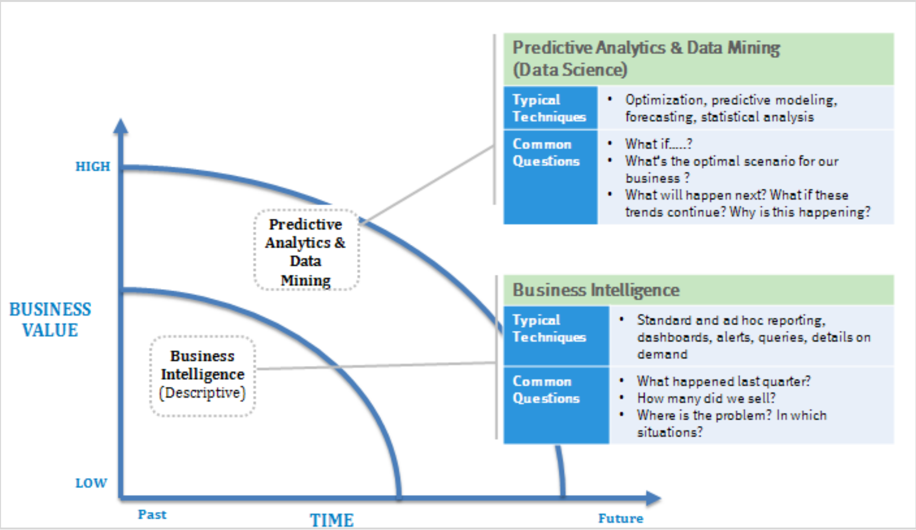
Predictive Analytics & Data Mining (Data Science)	
Typical Techniques	- Optimization, predictive modeling, forecasting, statistical analysis
Common Questions	- What if.....? - What's the optimal scenario for our business ? - What will happen next? What if these trends continue? Why is this happening?

Business Intelligence	
Typical Techniques	- Standard and ad hoc reporting, dashboards, alerts, queries, details on demand
Common Questions	- What happened last quarter? - How many did we sell? - Where is the problem? In which situations?

Guidance Processes in Data Science Projects

1. Well-defined processes can help guide any analytic project

2. Focus on data analytic lifecycle is more suited to data science projects.


HIGH
BUSINESS VALUE
LOW
Predictive Analytics & Data Mining
Business Intelligence (Descriptive)
Past
TIME
Future

Predictive Analytics & Data Mining (Data Science)	
Typical Techniques	• Optimization, predictive modeling, forecasting, statistical analysis
Common Questions	• What if.....? • What's the optimal scenario for our business ? • What will happen next? What if these trends continue? Why is this happening?

Business Intelligence	
Typical Techniques	• Standard and ad hoc reporting, dashboards, alerts, queries, details on demand
Common Questions	• What happened last quarter? • How many did we sell? • Where is the problem? In which situations?

3. Data Science projects tend to require a more consultative approach, and differ in a few ways
   - More due diligence in Discovery phase
   - More projects which lack shape or structure
   - Less predictable data

Key Roles for a Successful Analytic Project

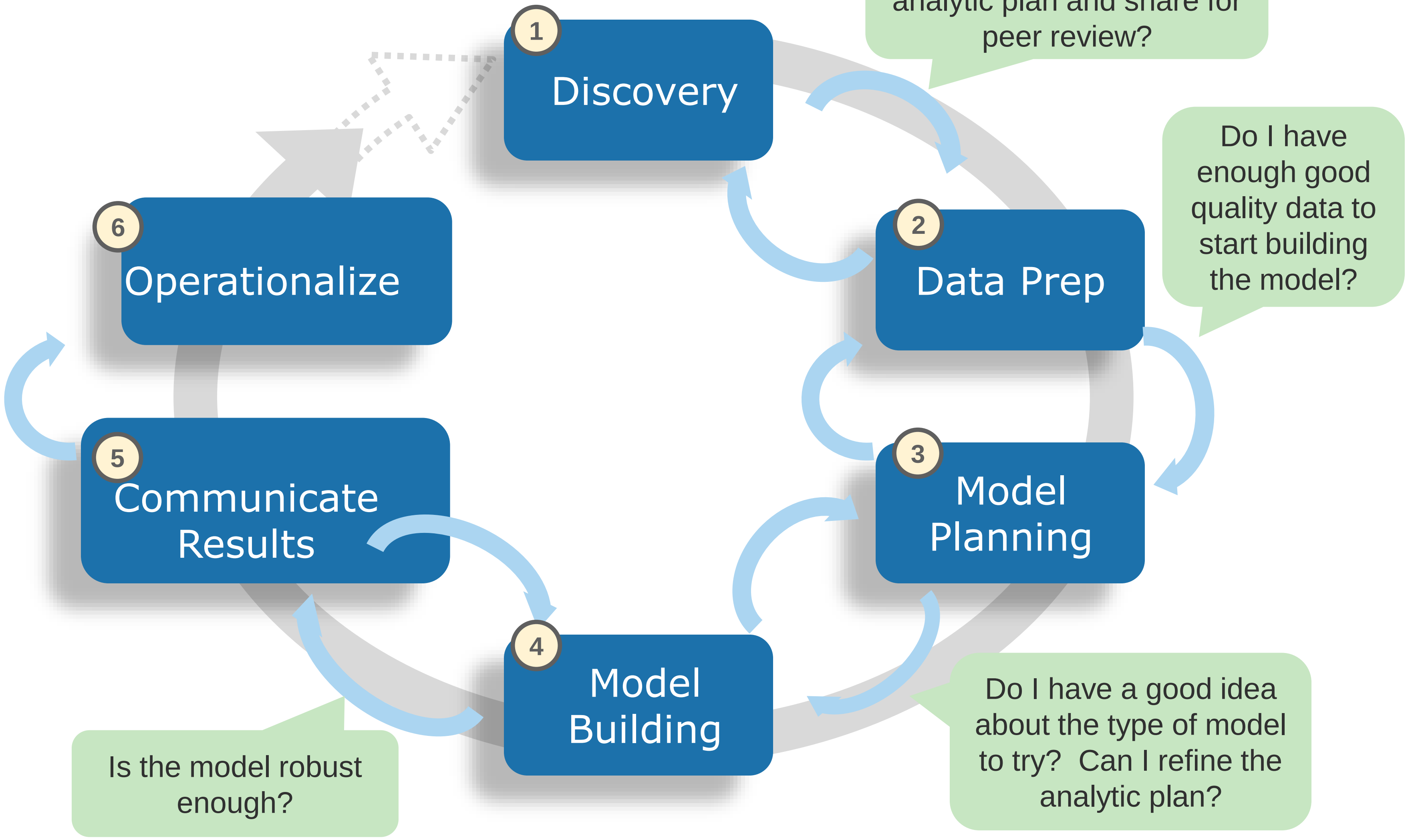
Role	Description
Business User	Someone who benefits from the end results and can consult and advise project team on value of end results and how these will be operationalized
Project Sponsor	Person responsible for the origin of the project, providing the motive for the project and core business problem, generally provides the funding and will measure 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, key performance indicators , 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 consume 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

- The Data Analytics Lifecycle presents a best practices approach for an end-to-end analytics process from discovery to project completion.

- You can move iteratively between phases until you have sufficient information to continue moving forward.

Data Analytics Lifecycle



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

Do I have enough good quality data to start building the model?

Do I have a good idea about the type of model to try? Can I refine the analytic plan?

Is the model robust enough?

Data Analytics Lifecycle
Phase 1: Discovery

1

Discovery

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

Do I have enough good

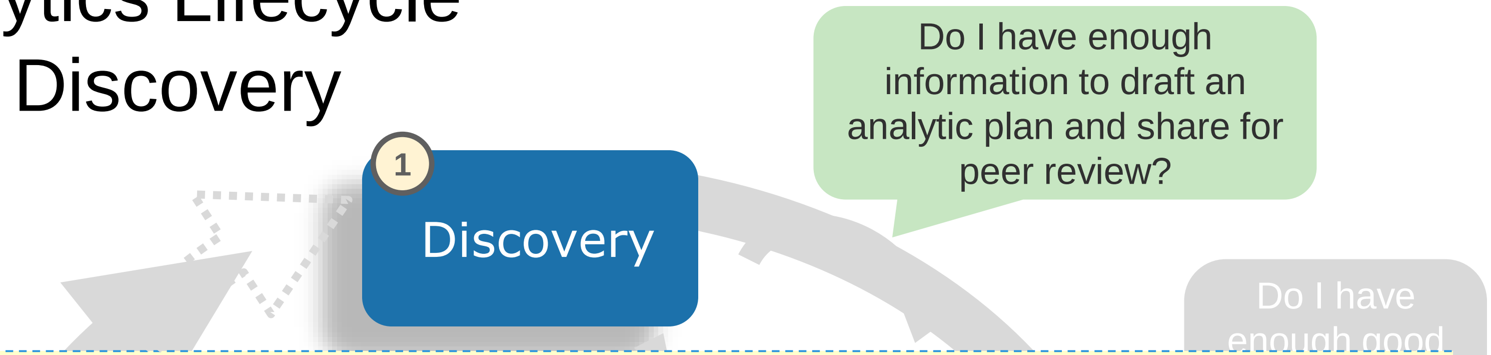
- **Learn the Business Domain**
  - Determine the domain knowledge needed to orient you to the data and interpret results downstream
  - Determine the general analytic problem type (such as clustering, classification)
  - If you don’t know, then conduct initial research to learn about the domain area you’ll be analyzing
- **Learn from the past (Problem History)**
  - 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

1
Discovery
Do I have enough information to draft an analytic plan and share for peer review?
Do I have enough good quality data to start building

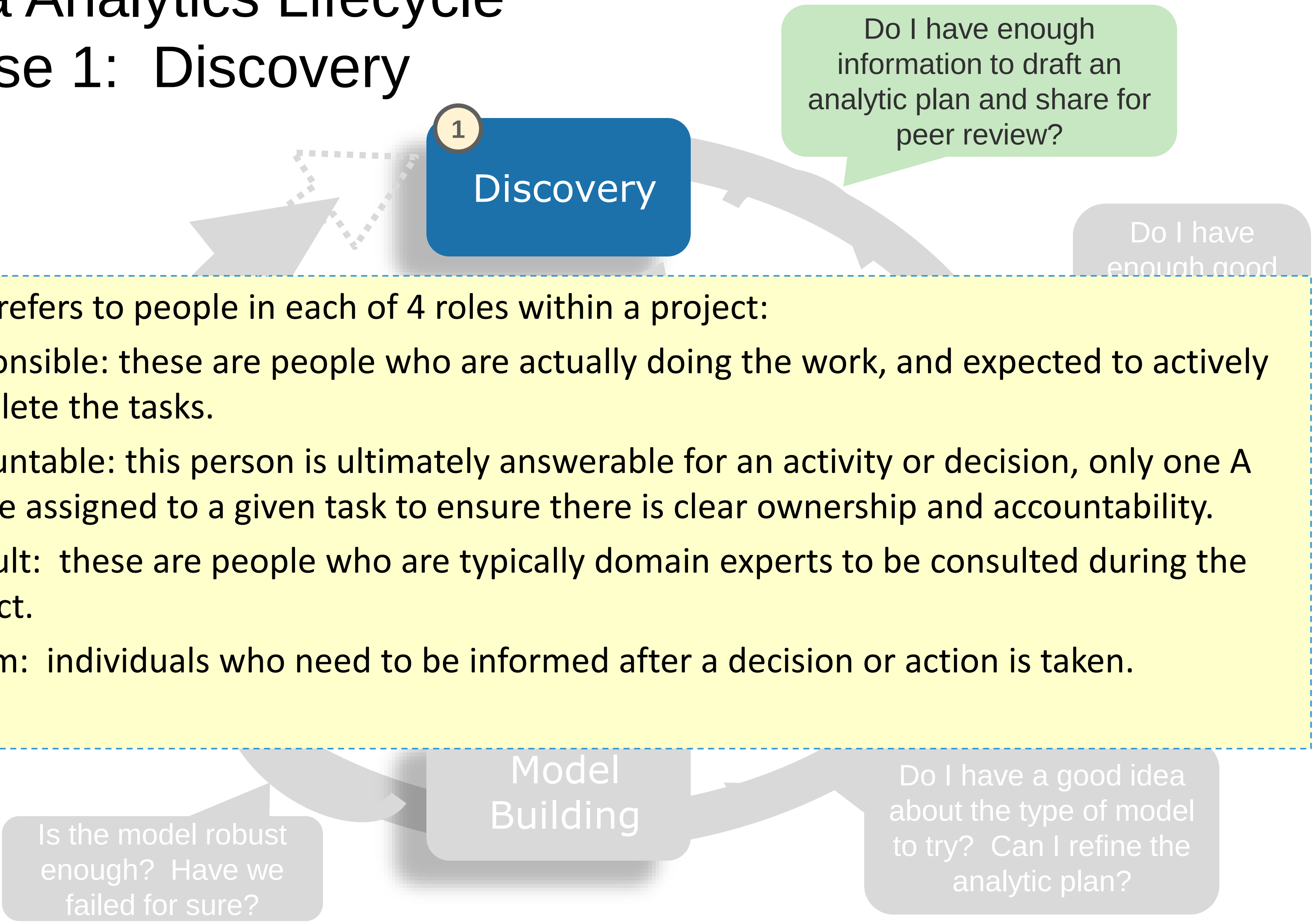
- **Resources**
  - Assess available technology
  - Available data – sufficient to meet your needs
  - People for the working team
  - Assess scope of time for the project in calendar time and person-hours
  - Do you have sufficient resources to attempt the project? If not, can you get more?

Data Analytics Lifecycle
Phase 1: Discovery


1
Discovery
Do I have enough information to draft an analytic plan and share for peer review?
Do I have enough good

- **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 (such as RACI matrix: *Responsible, Accountable, Consulted, and Informed*)

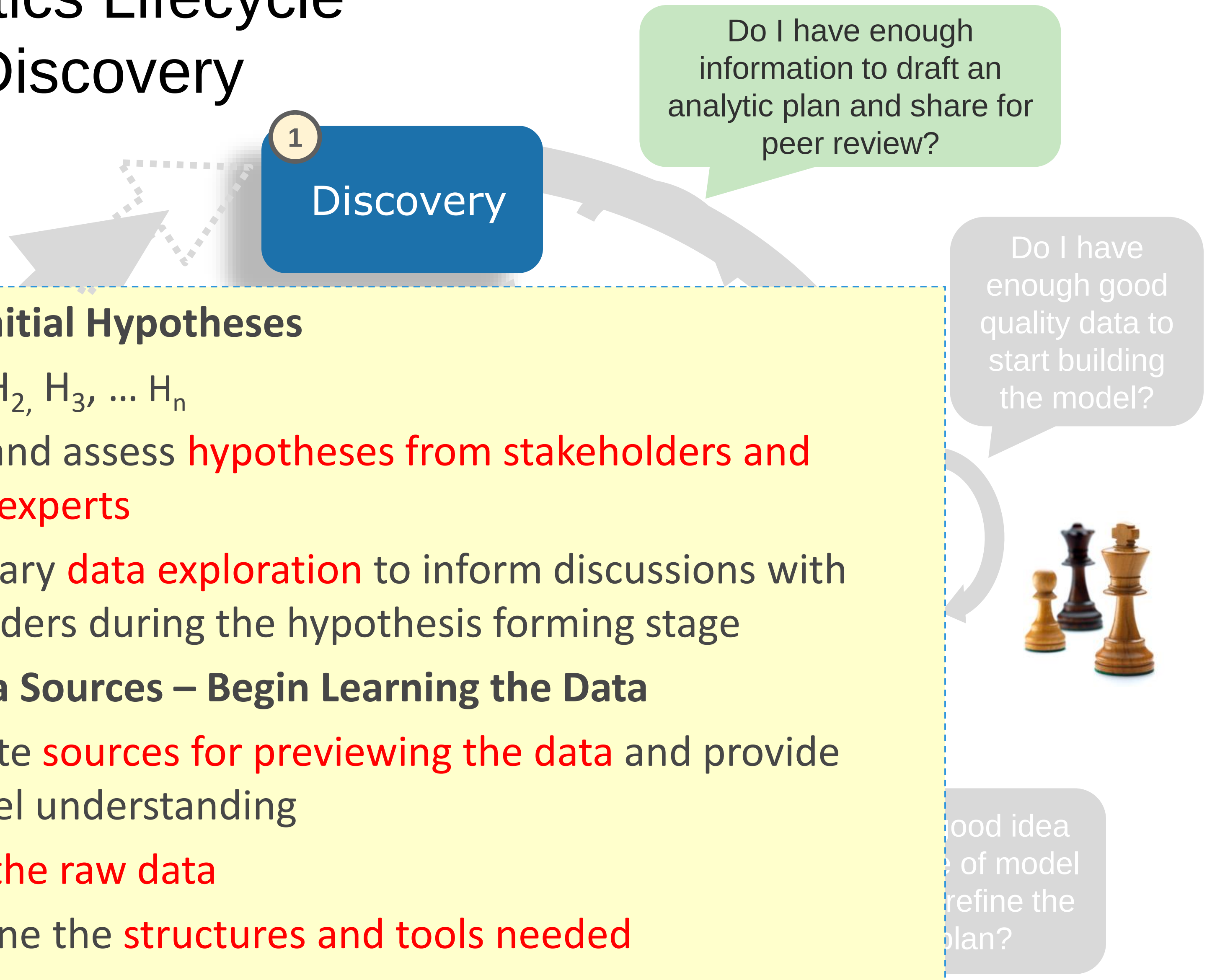
Data Analytics Lifecycle
Phase 1: Discovery



1
Discovery
Do I have enough information to draft an analytic plan and share for peer review?
Do I have enough good
Model Building
Is the model robust enough? Have we failed for sure?
Do I have a good idea about the type of model to try? Can I refine the analytic plan?

- RACI refers to people in each of 4 roles within a project:
- Responsible: these are people who are actually doing the work, and expected to actively complete the tasks.
- Accountable: this person is ultimately answerable for an activity or decision, only one A can be assigned to a given task to ensure there is clear ownership and accountability.
- Consult: these are people who are typically domain experts to be consulted during the project.
- Inform: individuals who need to be informed after a decision or action is taken.

Data Analytics Lifecycle
Phase 1: Discovery



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

Do I have enough good quality data to start building the 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
  - Scope the kind of data needed for this kind of problem

Track the Phases in the Data Analytics Lifecycle

A Case Study

Situation Synopsis

- ABC Retail Bank (RB) wants to improve the Net Present Value (NPV) and Retention Rate (RR)of customers.
  - *RB: The consumer-oriented services offered by commercial banks. These services include checking and savings accounts, mortgages and ...*
  - *NPV: The value in the present of a sum of money, in contrast to some future value it will have when it has been invested at compound interest.*
  - RR: The ratio of the number of retained customers to the number at risk.

Track the Phases in the Data Analytics Lifecycle

A Case Study

Situation Synopsis

- They want to establish an effective marketing campaign targeting customers to reduce the Churn Rate (CR) by at least five percent.

  - *CR: The annual percentage rate at which customers stop subscribing to a service or employees leave a job.*

- The bank wants to determine whether those customers are worth retaining.

- In addition, the bank also wants to analyze reasons for customer escape and what they can do to keep them.

Track the Phases in the Data Analytics Lifecycle
A Case Study

- The bank wants to build a data warehouse to support Marketing and other related customer care groups.

Situation Synopsis

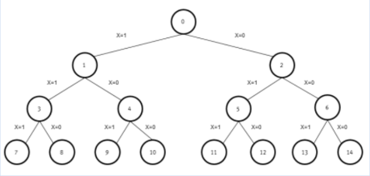
- You have been assigned to lead the analytics team on this project.
- You are in the Discovery phase of the Analytics Life Cycle.
  1. *The first step in this Phase is to learn about the domain of Retail Banking and Marketing.*
  2. *Other steps….. ?*

- What are your initial hypotheses (IH)?
- What data will you need to test the IH?
- What data dependencies will you have?
- Additional information about the data the bank has offered you to assist in your analytical efforts:
  - 250,000 – customers (Pilot Study), 2,500,000 – Final reporting
  - Customer Profile: Salary, age, Number of years as customer
  - Service Indicators (type of accounts, such as credit card, mortgage savings, checking)
  - Customer transactions and associated attributes, such as transaction size (in dollars), count of transactions for credit and debit card

- After initial data exploration, 5 months appears to capture relevant time perioc
- The churn should be determined based on the declining transactions. Churn/no churn situation of any particular customer should be predicted given 5 months of historical data .
- What is Net Present Value?
  - What is 'Net Present Value - NPV'
  - Net Present Value (NPV) is the difference between the present value of cash inflows and the present value of cash outflows.
  - NPV is used in capital budgeting to analyse the profitability of a projected investment or project. (Source: investopedia.com)
- Discuss revenue and cost components for a retail bank customer.
  - **Retail banking**, also known as consumer **banking**, is the typical mass-market **banking** in which individual **customers** use local branches of larger commercial **banks**. Services offered include savings and checking accounts, mortgages, personal loans, debit/credit cards and certificates of deposit (CDs).

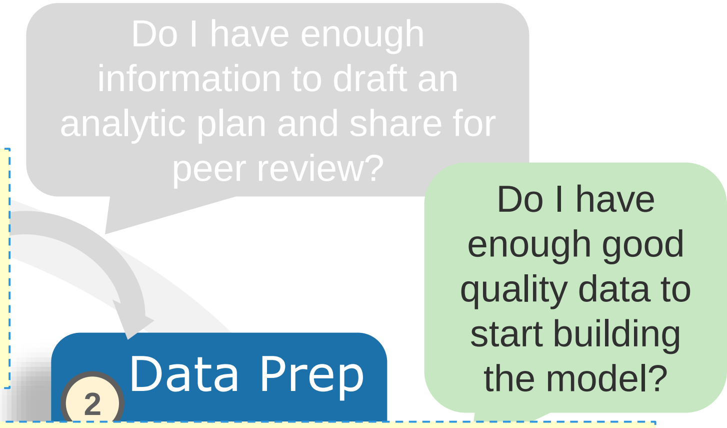
- How do you define “retention rate”?
  - customer retention rate is the percentage of customers you keep relative to the number you had at the start of your period.
  - This does not count new customers.
  - It is the reverse of customer churn.
- What is a churn rate?
- Can we measure the current churn rate? If so, how?
  - *CR: The annual percentage rate at which customers stop subscribing to a service or employees leave a jo*

How to Frame an Analytics Problem

Sample *Business* Problems	Qualifiers	Analytical Approach
• How can we improve on x? • What's happening at real-time? Trends? • How can we use analytics to differentiate ourselves • How can we use analytics to innovate? • How can we stay ahead of our biggest competitor?	Will the focus and scope of the problem change if the following dimensions change: • Time • People – how would x change this? • Risk – conservative/aggressive • Resources – none/unlimited • Size of Data?	Define an analytical approach, including key terms, metrics, and data needed.  0 X=1 X=0 1 2 X=1 X=0 X=1 X=0 3 4 5 6 X=1 X=0 X=1 X=0 X=1 X=0 X=1 X=0 7 8 9 10 11 12 13 14
<u>ABC Bank</u> How can we improve Net Present Value (NPV) and retention rate of the customers?	• **Time**: Trailing 5 months • **People**: Working team and business users from the Bank • **Risk**: the project will fail if we cannot determine valid predictors of churn • **Resources**: EDW, analytic sandbox, Online Transaction Processing(OLTP) system • **Data**: Use 24 months for the training set, then analyze 5 months of historical data for those customers who churned	How do we identify churn/no churn for a customer? Pilot study followed full scale analytical model

Data Analytics Lifecycle
Phase 2: Data Preparation

- **Prepare Analytic Sandbox**
  - Work space for the analytic team
  - To be large at least 10 times EDW


Do I have enough information to draft an analytic plan and share for peer review?
Do I have enough good quality data to start building the model?
2
Data Prep

Define a space where you can explore the data without interfering with live production databases.
For instance, you may need to work with a company’s financial data, but cannot interact with the production version of the organization’s main database ..

You should be collecting all kinds of data in your sandbox!

This can include everything from summary, structured data, to raw data feeds, to unstructured text data from call logs or web logs

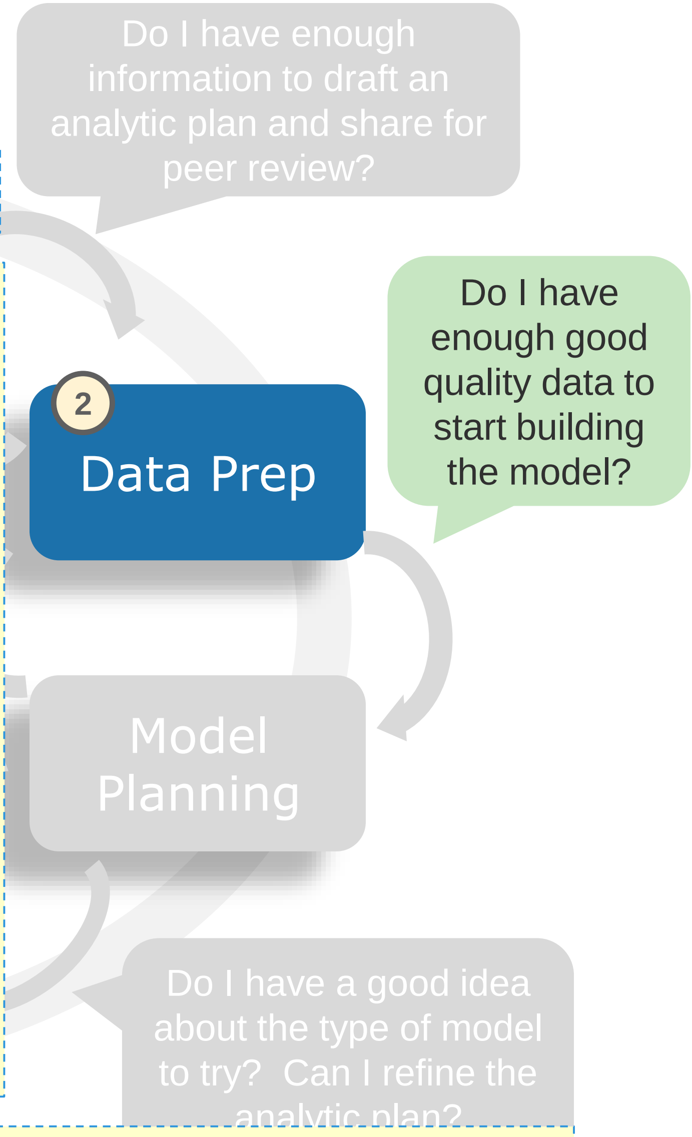
Is the model robust

Remember: enterprise data warehouse (EDW), is a system used for reporting and data analysis, and is considered a core component of business intelligence.

Data Analytics Lifecycle
Phase 2: Data Preparation

- **Perform ELT better than ETL!**

- **In ETL, users perform Extract – Transform – Load processes to get data into a database and perform data transformations before data is loaded into the database.**
- **Using the analytic sandbox approach, we advocate doing ELT – Extract, Load, then Transform.**
- **In this case, the data is extracted in its raw form and loaded into the database.**
- **Then, analysts can choose to transform the data into a new state or leave it in its original, raw condition.**



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

Do I have enough good quality data to start building the model?

Do I have a good idea about the type of model to try? Can I refine the analytic plan?

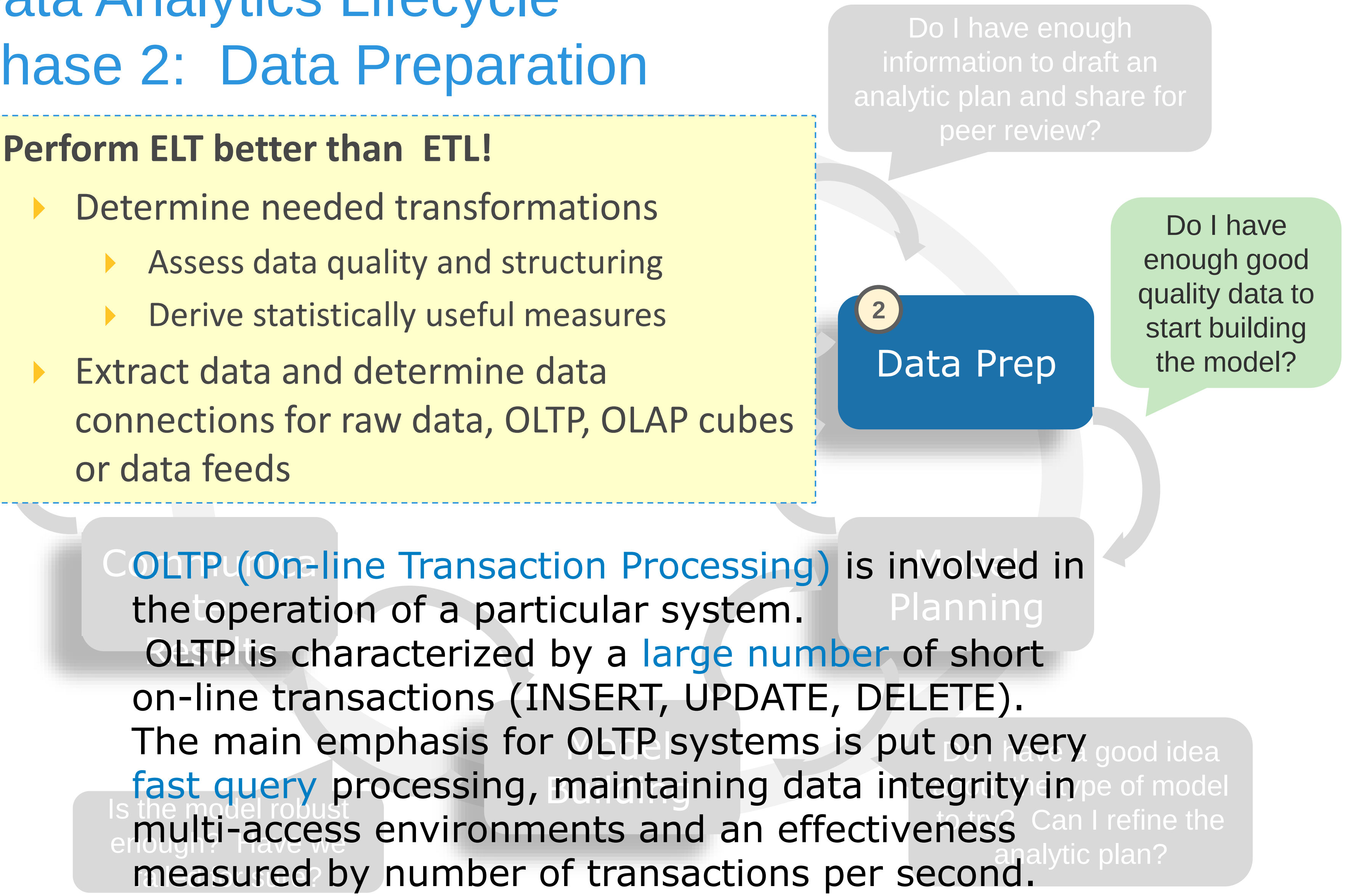
enough? Have we

- **The reason for this approach is that there is significant value in preserving the raw data and including it in the sandbox, before any transformations.**

E

Data Analytics Lifecycle
Phase 2: Data Preparation

- **Perform ELT better than ETL!**
  - Determine needed transformations
    - Assess data quality and structuring
    - Derive statistically useful measures
  - Extract data and determine data connections for raw data, OLTP, OLAP cubes or data feeds



Do I have enough information to draft an analytic plan and share for peer review?
2
Data Prep
Do I have enough good quality data to start building the model?
Communicate Results
Model Planning
Model Building
Is the model robust enough? Have we failed for sure?
Do I have a good idea about the type of model to try? Can I refine the analytic plan?

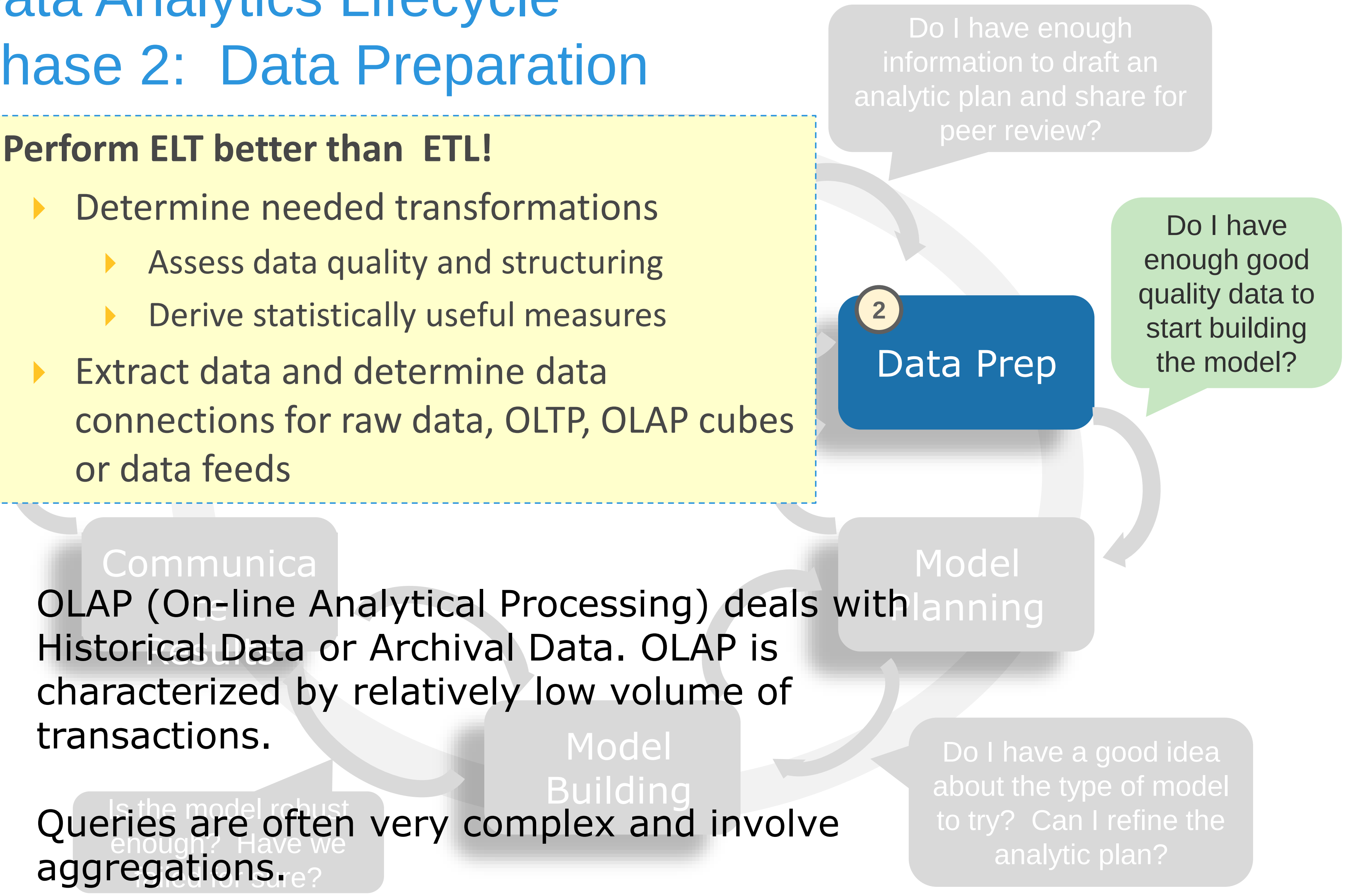
OLTP (On-line Transaction Processing) is involved in the operation of a particular system.
 OLTP is characterized by a large number of short on-line transactions (INSERT, UPDATE, DELETE).
The main emphasis for OLTP systems is put on very fast query processing, maintaining data integrity in multi-access environments and an effectiveness measured by number of transactions per second.

Data Analytics Lifecycle
Phase 2: Data Preparation

- **Perform ELT better than ETL!**
  - Determine needed transformations
    - Assess data quality and structuring
    - Derive statistically useful measures
  - Extract data and determine data connections for raw data, OLTP, OLAP cubes or data feeds

OLAP (On-line Analytical Processing) deals with Historical Data or Archival Data. OLAP is characterized by relatively low volume of transactions.

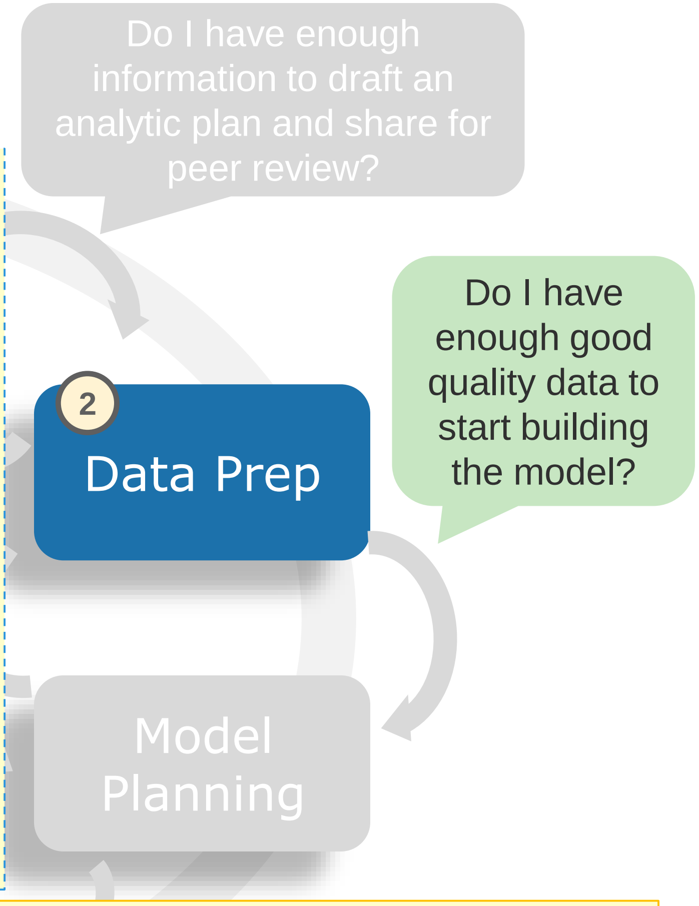
Queries are often very complex and involve aggregations.



Do I have enough information to draft an analytic plan and share for peer review?
2
Data Prep
Do I have enough good quality data to start building the model?
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
Model Building
Is the model robust enough? Have we failed for sure?
Communicate Results

Data Analytics Lifecycle
Phase 2: Data Preparation

- **Prepare Analytic Sandbox**
  - Work space for the analytic team
  - To be large at least 10 times EDW
- **Perform ELT**
  - Determine needed transformations
    - Assess data quality and structuring
    - Derive statistically useful measures
  - Extract data and determine data connections for raw data, OLTP, OLAP cubes or data feeds


Do I have enough information to draft an analytic plan and share for peer review?
2
Data Prep
Do I have enough good quality data to start building the model?
Model Planning
enough? Have we failed for sure?
analytic plan?

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

Hadoob*

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

- Apache Hadoop is an open-source software framework used for distributed storage and processing of big data sets using the MapReduce programming model.
- It consists of computer clusters built from commodity hardware.
- All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common occurrences and should be automatically handled by the framework
- The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part which is a MapReduce programming model.

* from Wikipedia.com

failed for sure?

analytic plan?

Hadoob*

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

- Hadoop splits files into large blocks and distributes them across nodes in a cluster.
- It then transfers packaged code into nodes to process the data in parallel. This approach takes advantage of data locality, where nodes manipulate the data they have access to.
- This allows the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture

* from Wikipedia.com

Is the model robust enough? Have we failed for sure?

to try? Can I refine the analytic plan?

Hadoob*

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

- The base Apache Hadoop framework is composed of the following modules:

- Hadoop Common – contains libraries and utilities needed by other Hadoop modules;
- Hadoop Distributed File System (HDFS) – a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster;
- Hadoop YARN – a resource-management platform responsible for managing computing resources in clusters and using them for scheduling of users' applications; and
- Hadoop MapReduce – an implementation of the MapReduce programming model for large scale data processing.

- * from Wikipedia.com

MapReduce *

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

- MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster.

- A MapReduce program is composed of a Map() procedure (method) that performs filtering and sorting (such as sorting students by first name into queues, one queue for each name)
- and a Reduce() method that performs a summary operation (such as counting the number of students in each queue, yielding name frequencies).
- The "MapReduce System" (also called "infrastructure" or "framework") orchestrates the processing by marshalling the distributed servers, running the various tasks in parallel, managing all communications and data transfers between the various parts of the system, and providing for redundancy and fault tolerance..

- * from Wikipedia.com

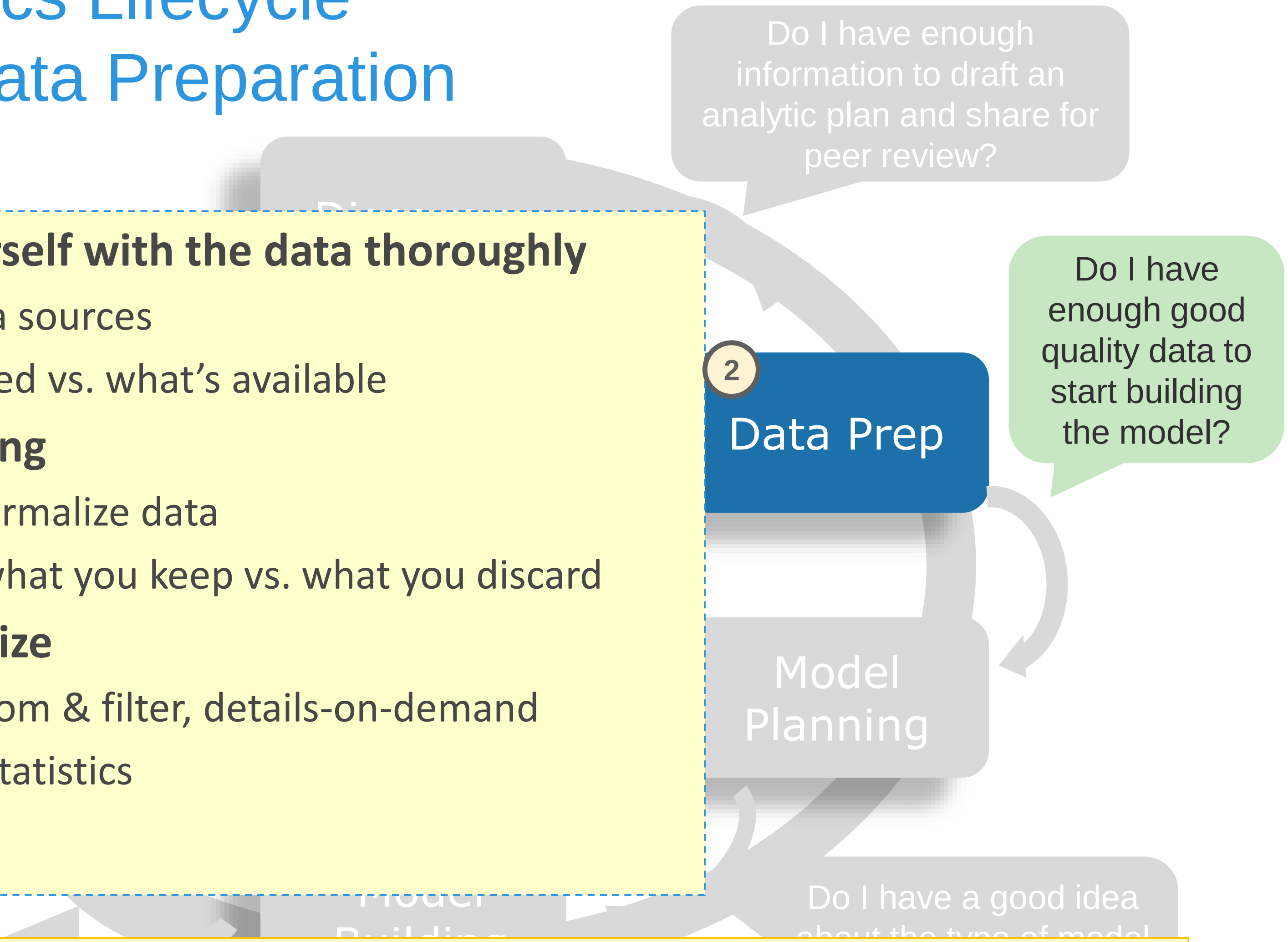
Alpine Data *

Do I have enough information to draft an

- Alpine Data Labs is an advanced analytics interface working with Apache Hadoop and big data
- It provides a collaborative, visual environment to create and deploy analytics workflow and predictive models
- This aims to make analytics more suitable for business analyst level staff, like sales and other departments using the data, rather than requiring a "data scientist" who understands languages like MapReduce
- * from Wikipedia.com

Data Analytics Lifecycle
Phase 2: Data Preparation

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



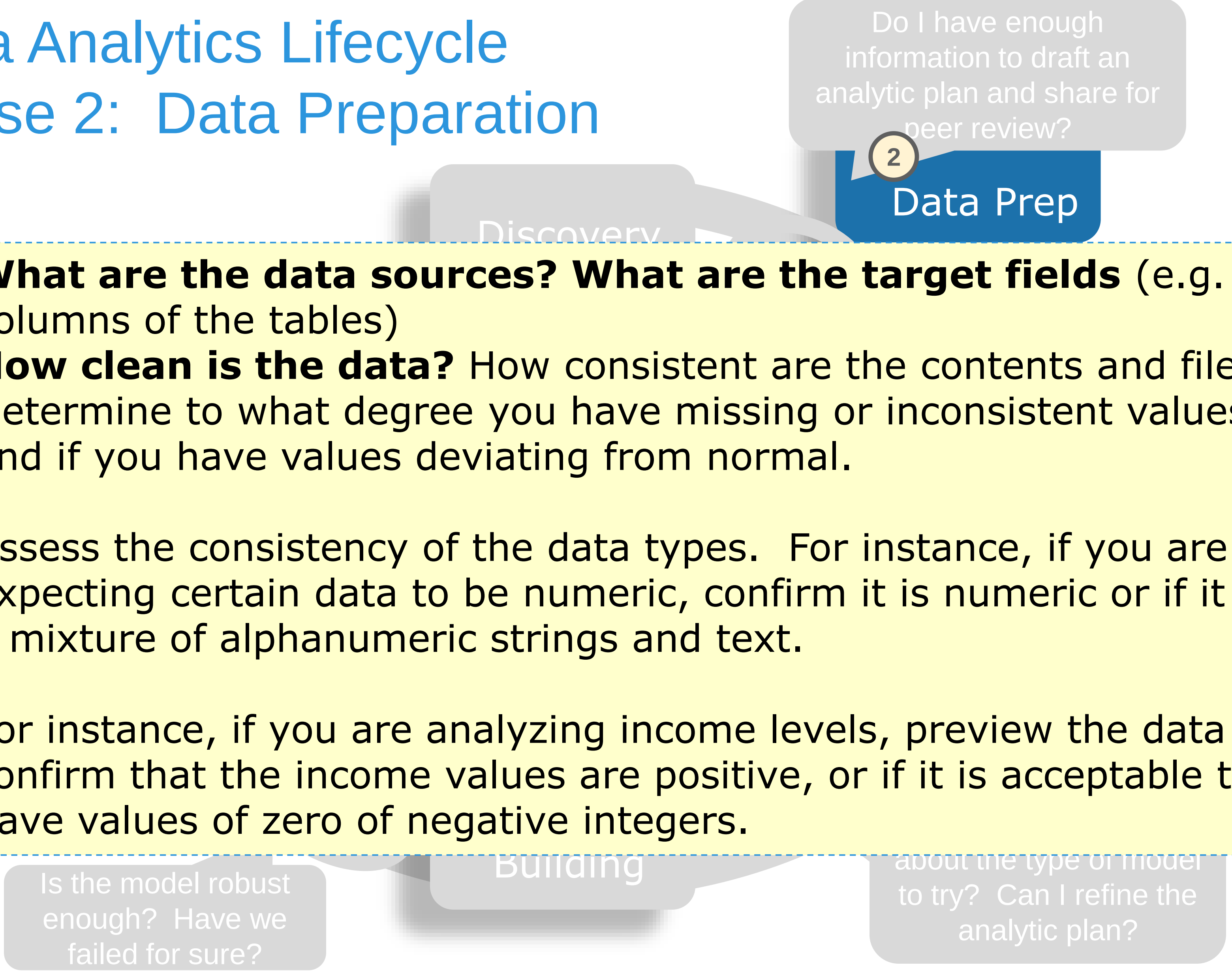
Do I have enough information to draft an analytic plan and share for peer review?

Do I have enough good quality data to start building the model?

Do I have a good idea

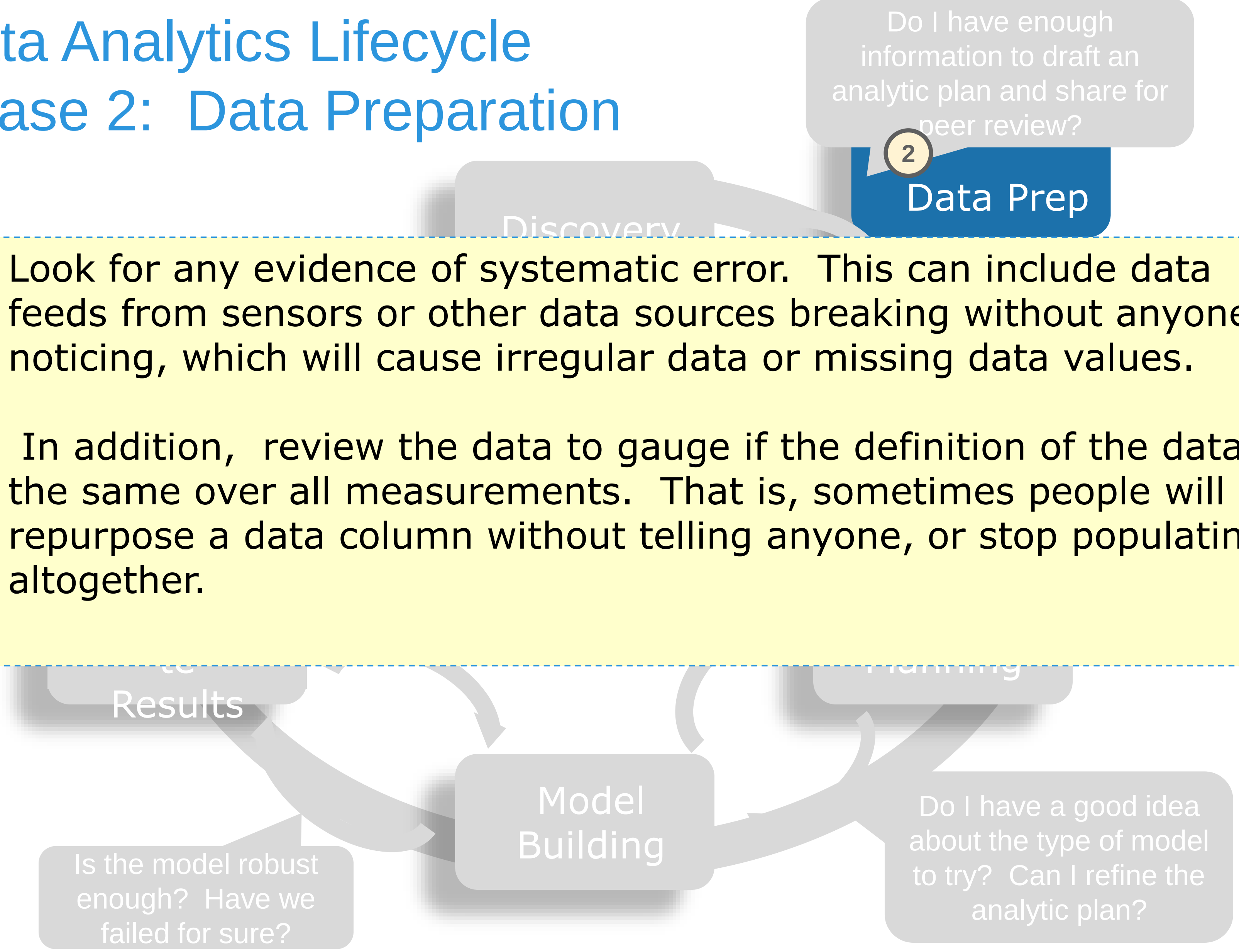
- **Useful Tools for this phase:**
  - Descriptive Statistics on candidate variables for diagnostics & quality
  - ***Visualization***: R (base package, ggplot and lattice), GnuPlot, Ggobi/Rggobi, Spotfire, Tableau

Data Analytics Lifecycle
Phase 2: Data Preparation


Do I have enough information to draft an analytic plan and share for peer review?
2
Data Prep
Discovery
Building
Is the model robust enough? Have we failed for sure?
about the type of model to try? Can I refine the analytic plan?

- **What are the data sources? What are the target fields** (e.g. columns of the tables)
- **How clean is the data?** How consistent are the contents and files? Determine to what degree you have missing or inconsistent values, and if you have values deviating from normal.
- Assess the consistency of the data types. For instance, if you are expecting certain data to be numeric, confirm it is numeric or if it is a mixture of alphanumeric strings and text.
- For instance, if you are analyzing income levels, preview the data to confirm that the income values are positive, or if it is acceptable to have values of zero of negative integers.

Data Analytics Lifecycle
Phase 2: Data Preparation



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

2

Data Prep

Discovery

te

Results

Planning

Model Building

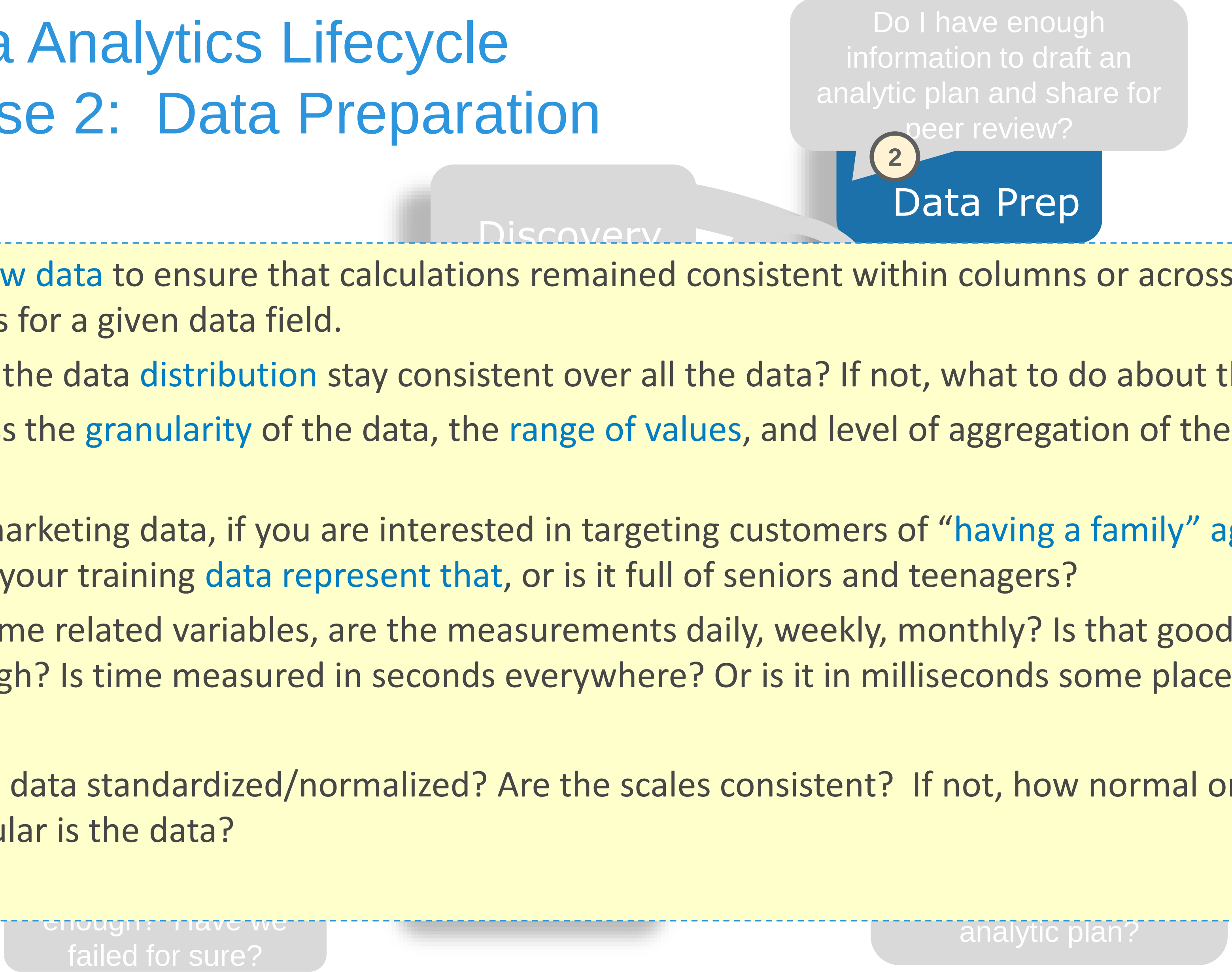
Is the model robust enough? Have we failed for sure?

Do I have a good idea about the type of model to try? Can I refine the analytic plan?

- Look for any evidence of systematic error. This can include data feeds from sensors or other data sources breaking without anyone noticing, which will cause irregular data or missing data values.

- In addition, review the data to gauge if the definition of the data is the same over all measurements. That is, sometimes people will repurpose a data column without telling anyone, or stop populating it altogether.

Data Analytics Lifecycle
Phase 2: Data Preparation

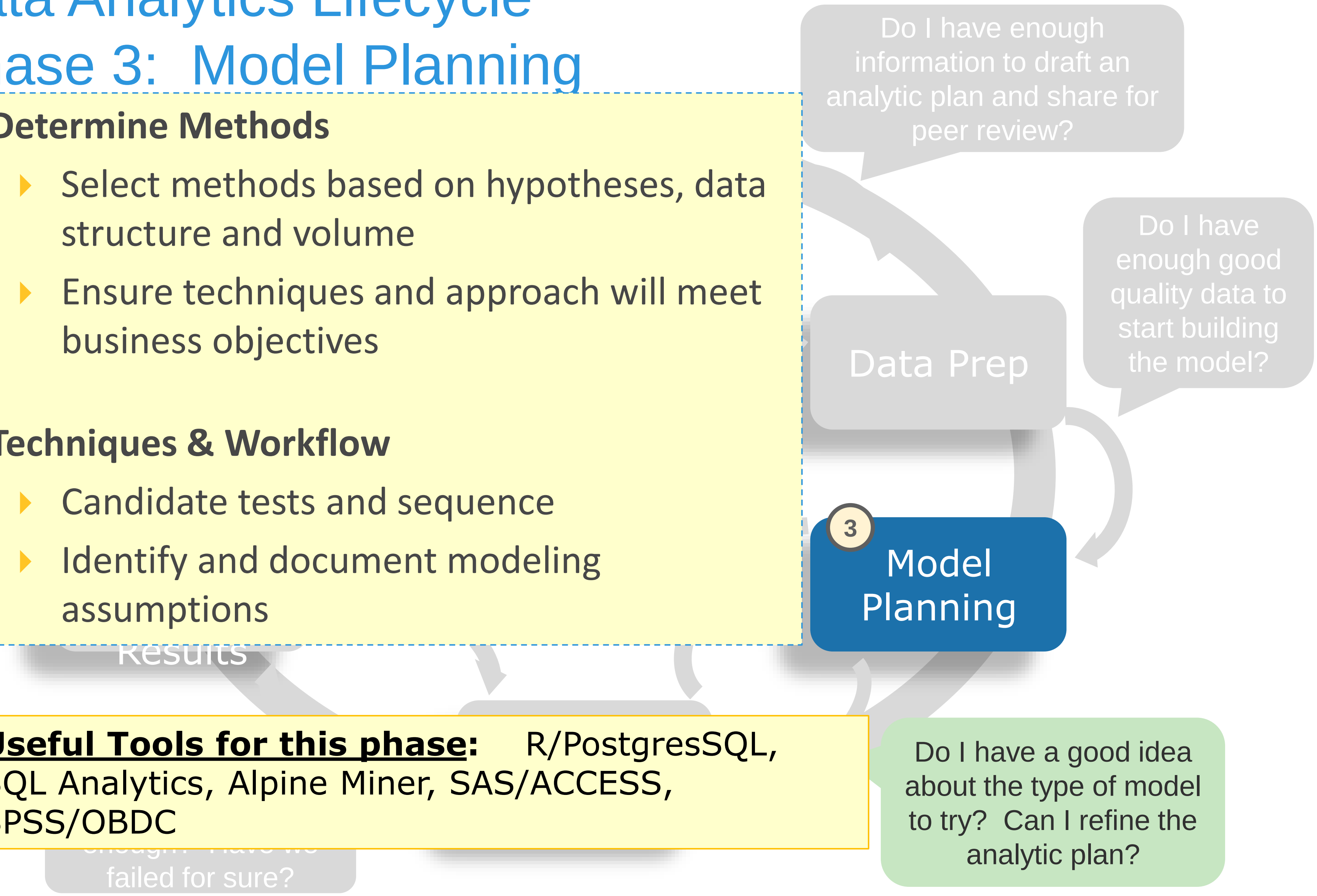

Do I have enough information to draft an analytic plan and share for peer review?
2
Data Prep
Discovery
enough? Have we failed for sure?
analytic plan?

- Review data to ensure that calculations remained consistent within columns or across tables for a given data field.
- Does the data distribution stay consistent over all the data? If not, what to do about that?
- Assess the granularity of the data, the range of values, and level of aggregation of the data
- For marketing data, if you are interested in targeting customers of “having a family” age, does your training data represent that, or is it full of seniors and teenagers?
- For time related variables, are the measurements daily, weekly, monthly? Is that good enough? Is time measured in seconds everywhere? Or is it in milliseconds some places?
- Is the data standardized/normalized? Are the scales consistent? If not, how normal or irregular is the data?

Data Analytics Lifecycle
Phase 3: Model Planning

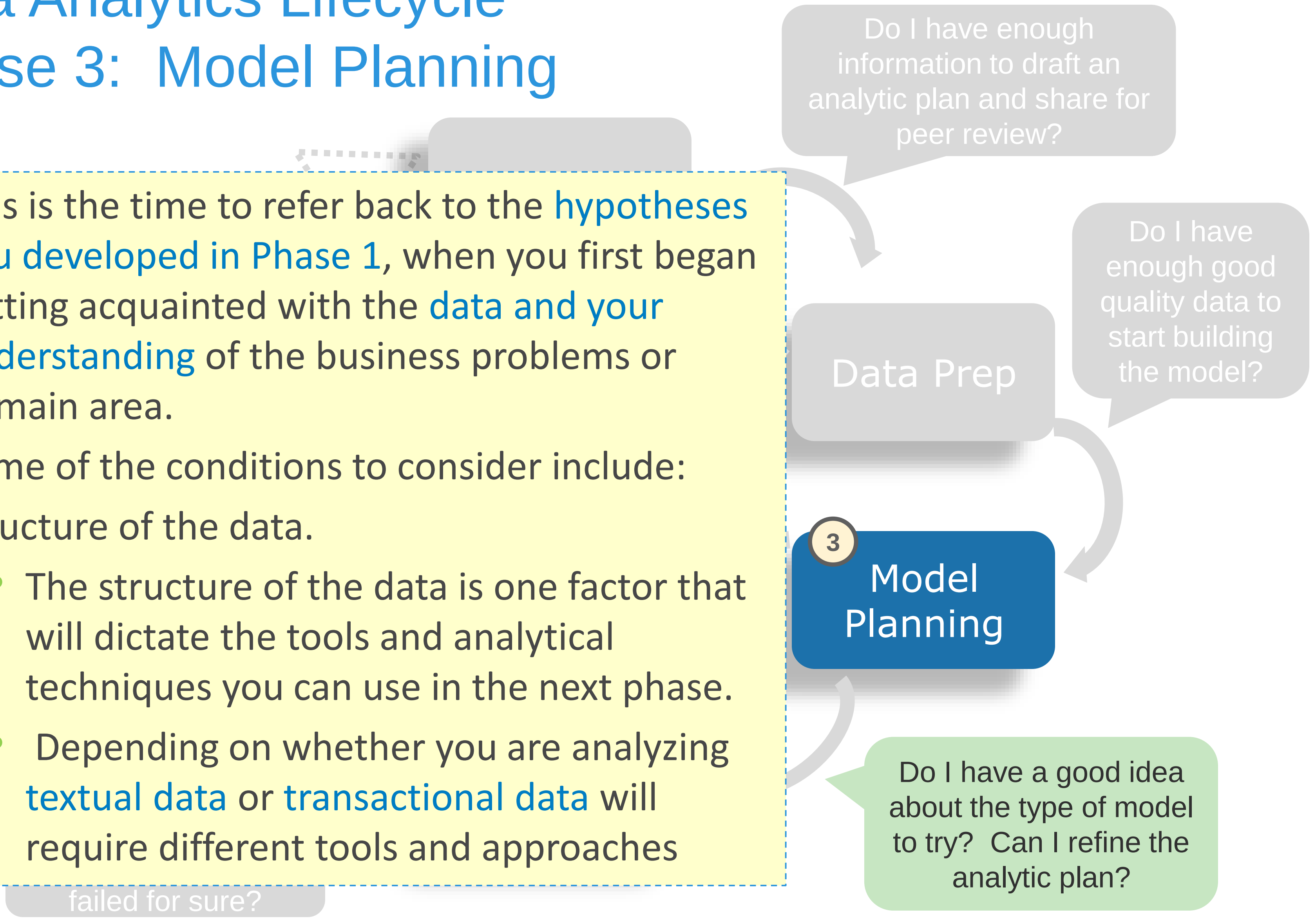
- **Determine Methods**
  - Select methods based on hypotheses, data structure and volume
  - Ensure techniques and approach will meet business objectives
- **Techniques & Workflow**
  - Candidate tests and sequence
  - Identify and document modeling assumptions

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


Do I have enough information to draft an analytic plan and share for peer review?
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
Results
enough? Have we failed for sure?

Data Analytics Lifecycle
Phase 3: Model Planning

- This is the time to refer back to the hypotheses you developed in Phase 1, when you first began getting acquainted with the data and your understanding of the business problems or domain area.
- Some of the conditions to consider include:
- Structure of the data.
  - The structure of the data is one factor that will dictate the tools and analytical techniques you can use in the next phase.
  - Depending on whether you are analyzing textual data or transactional data will require different tools and approaches


Do I have enough information to draft an analytic plan and share for peer review?
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
failed for sure?

Data Analytics Lifecycle
Phase 3: Model Planning

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

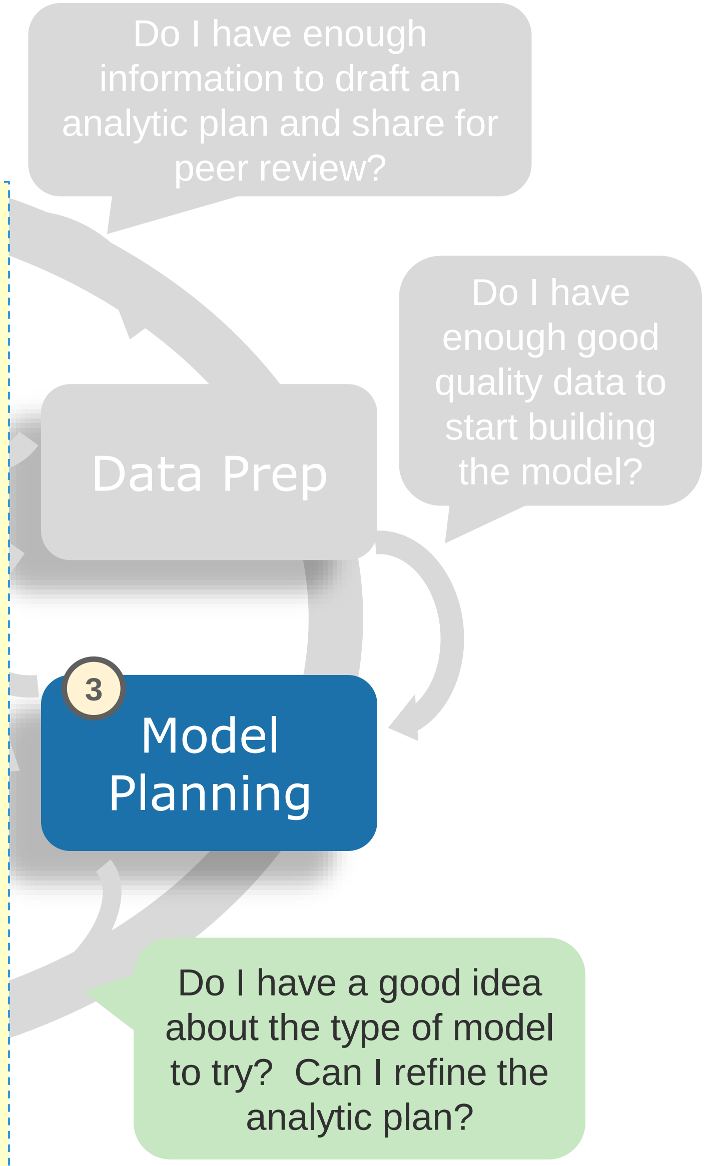
- eg., Sentiment Analysis using Hadoop) than forecasting market demand based on structured financial data (for example revenue projections and market sizing using regressions).
- Ensure that the analytical techniques will enable you to meet the business objectives and prove or disprove your working hypotheses.

ze

- Sentiment Analysis
- Sentiment analysis (sometimes known as opinion mining or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.
- Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine

Data Analytics Lifecycle
Phase 3: Model Planning

- **Data Exploration**
  - Understand the relationships among the variables
- **Variable Selection**
  - Inputs from stakeholders and domain experts
  - Capture essence of the predictors, leverage a technique for dimensionality reduction
  - Iterative testing to confirm the most significant variables
- **Model Selection**
  - Conversion to SQL or database language for best performance
  - Choose technique based on the end goal


Do I have enough information to draft an analytic plan and share for peer review?
Do I have enough good quality data to start building the model?
Data Prep
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?

Sample Research: Churn Prediction in Other Verticals

- After conducting research on churn prediction, you have identified many methods for analyzing customer churn across multiple verticals.
- At this point, a Data Scientist would assess the methods and select the best model for the situation

Market Sector	Analytic Techniques/Methods Used
Wireless Telecom	DMEL method (data mining by evolutionary learning), Neural network, **decision tree**, hierarchical neurofuzzy systems, rule evolver, **Logistic regression**.
Retail Business	**Logistic regression**, ARD (automatic relevance determination), **decision tree**
Daily Grocery	MLR (**multiple linear regression**), ARD, and **decision tree**
Retail Banking	**Multiple regression**