Overview of Data Analytics Lifecycle

Unit 2

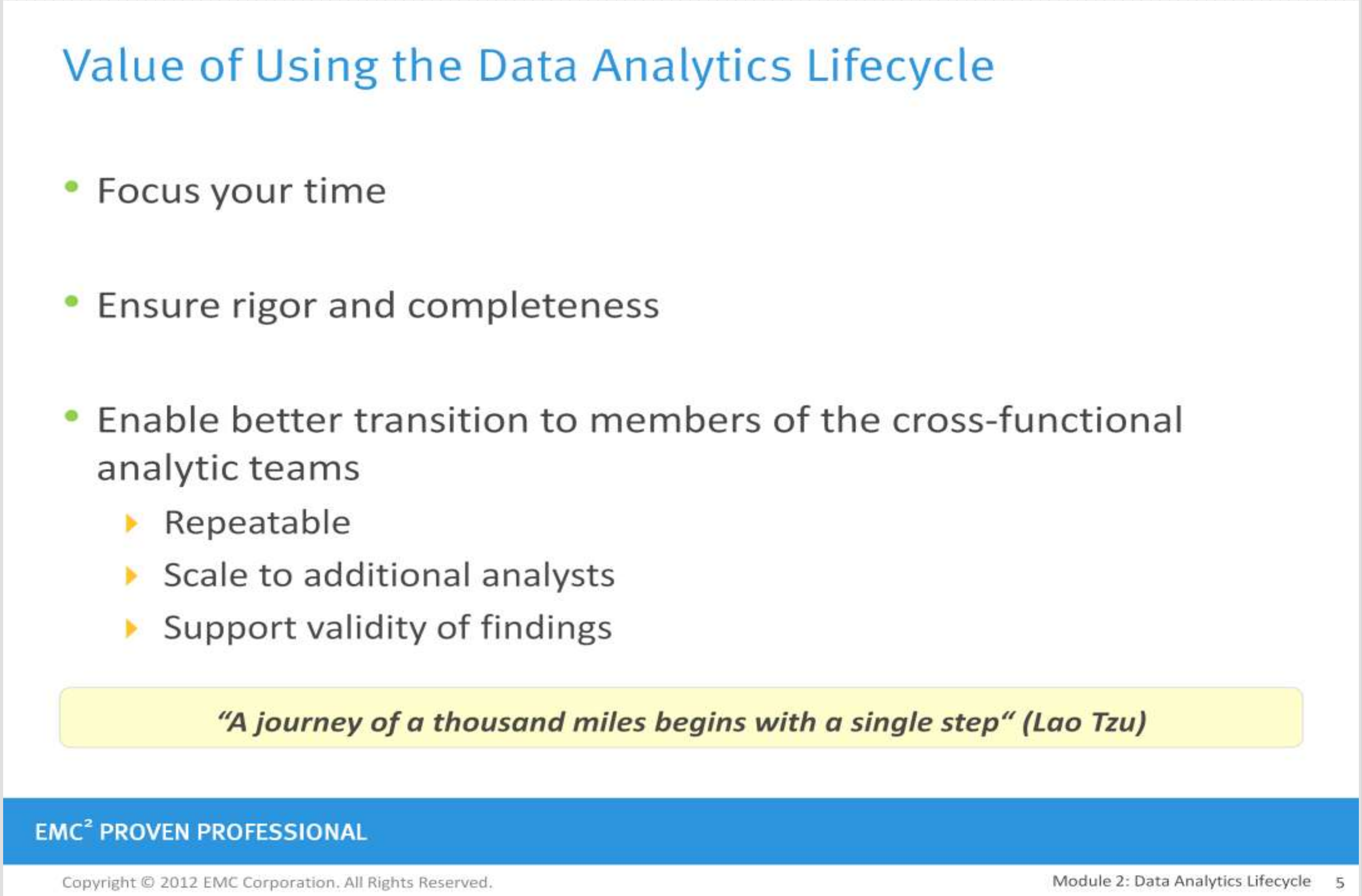
Topics Covered

- Discovery
- Data Preparation
- Model Planning
- Model Building
- Communicating Results and Finding
- Operationalzing

How to Approach Your Analytics Problems

Your Thoughts?

- 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?



Value of Using the Data Analytics Lifecycle

- Focus your time
- Ensure rigor and completeness
- Enable better transition to members of the cross-functional analytic teams
  - Repeatable
  - Scale to additional analysts
  - Support validity of findings

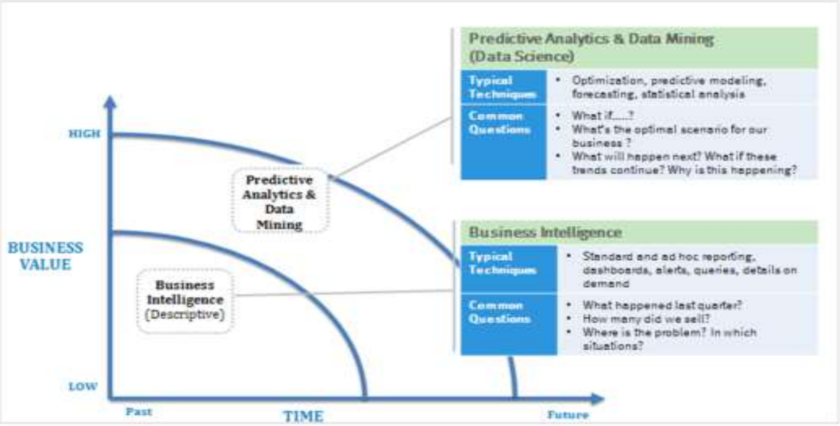
> ***"A journey of a thousand miles begins with a single step" (Lao Tzu)***

- Many problems seem huge and daunting at first, but **a well defined process enables you to break down complex problems into smaller steps** that you can more easily address.
- Having a good process for performing analytics is critical.
- It **will ensure that you establish a comprehensive, repeatable method for conducting analysis**.
- In addition, it will help you focus your time so you spend the requisite energy early in the process getting a clear grasp of the business problem to be solved.
- 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

- As a consequence, participants may discover mid-stream that the project sponsors are trying to solve a different objective, or have are attempting to address an interest that differs from what has been explicitly communicated.
- **Creating and documenting a process will help demonstrate rigor in your findings,** which will provide additional credibility to the project when you share your findings.
- It also enables you to teach others to adopt the methods and analysis so it can be repeated next quarter, next year, or by new hires.

Need For a Process to Guide Data Science Projects

1. Well-defined processes can help guide any analytic project
2. Focus of Data Analytics Lifecycle is on Data Science projects, not business intelligence


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

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?

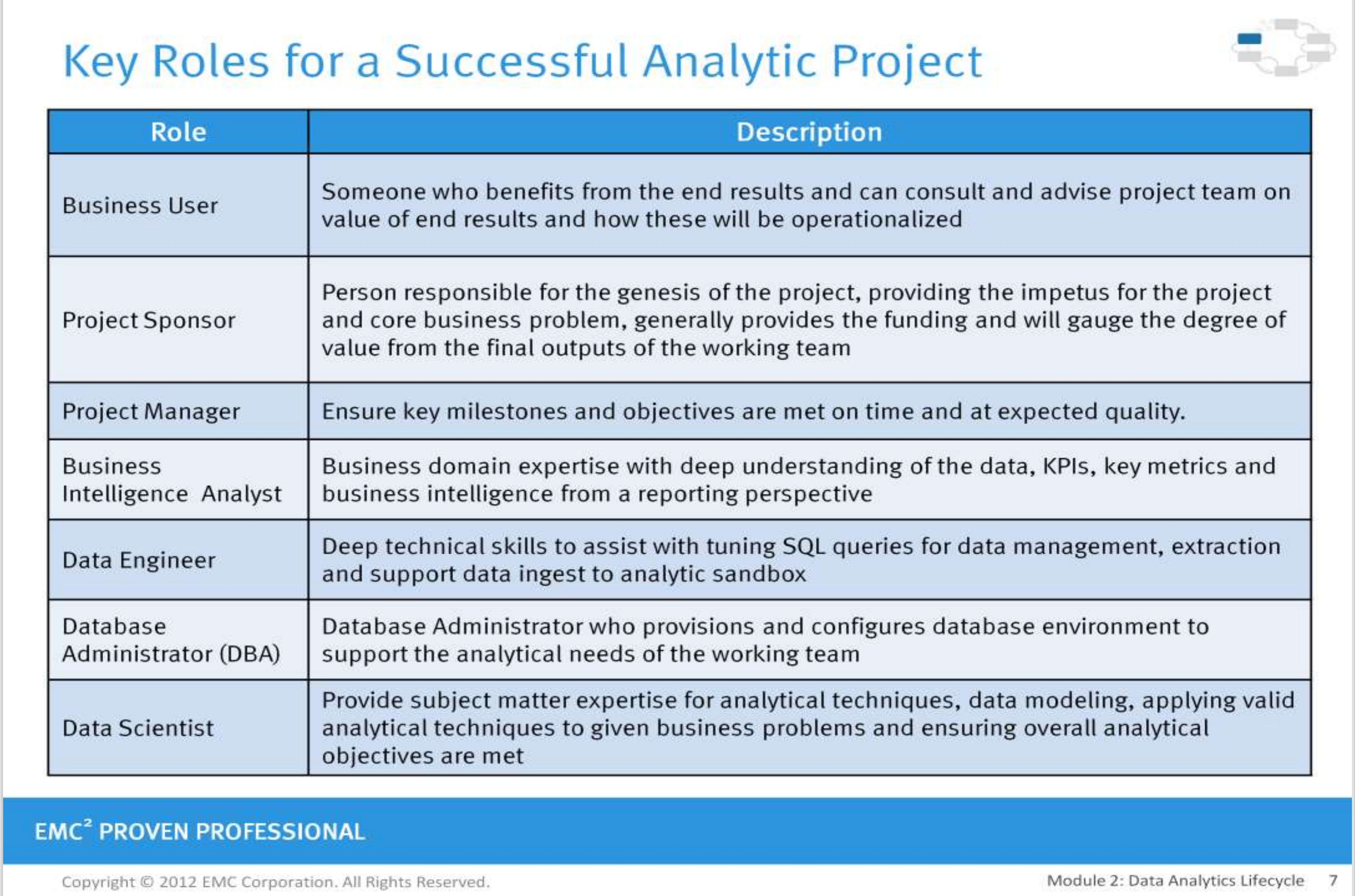
Business Intelligence	
Typical Techniques	• Standard and ad hoc reporting, dashboards, alerts, queries, details on demand
Common Questions	• What happened last quarter? • How many did we sell? • Where is the problem? In which situations?

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

- Although a well-defined process can help guide you with any analytic project, the data analytic lifecycle we will focus on in this module is more suited to data science projects.
- **Data Science projects tend to have a less well-structured approach or scope of the project,** and may require slightly different process than a project focused on deriving KPIs or implementing a dashboard.
- Many of the phases will be similar (for example you would still need to do a Discovery phase for any new project, although the focus would be different), though some would not be needed at all (for example you may not have to create training data

- In addition, because **data science may deal with big data, sparse data sets, or unstructured data, they require more diligence, data preparation and data conditioning,** than a project focused on business intelligence, which tend to leverage highly structured data resident in a data warehouse or OLAP cube.

These are descriptions of the various roles and main stakeholders of an analytics project.



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 genesis of the project, providing the impetus for the project and core business problem, generally provides the funding and will gauge 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

- There are 6 phases in the Data Analytics Lifecycle.
- Work on a project can be done in several phases at once.
- Movement from any phase to another and back again to previous phases occurs throughout the Lifecycle.
- The question callouts represent questions to ask yourself to gauge whether you have enough information and have made enough progress to move to the next phase of the process.
- Translate the results into a language that speaks to the audience that engaged you for the work.
- The Data Analytics Lifecycle shown portrays a best

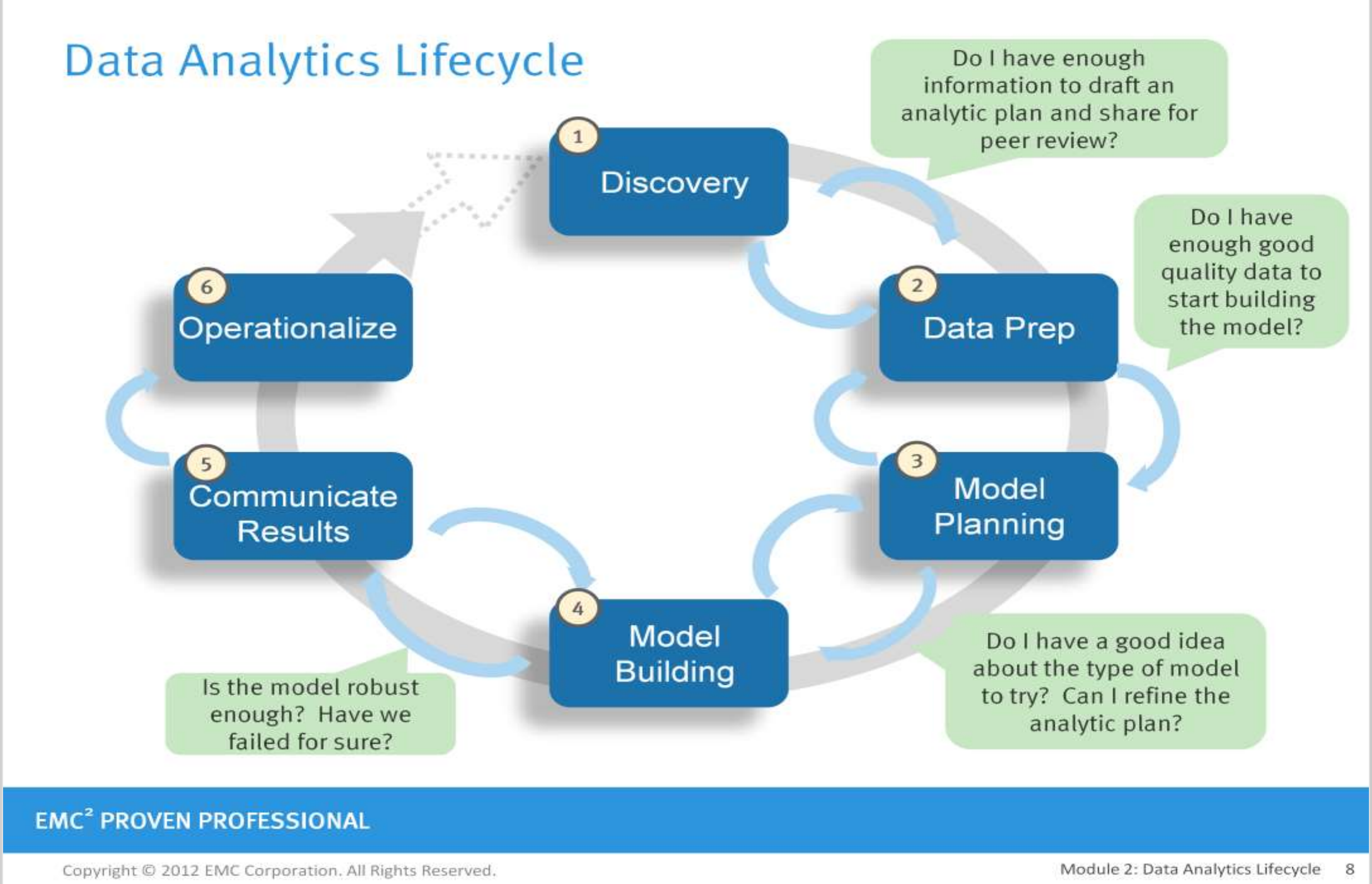
- Also steps to improve the process, drawn from established methods in the realm of data analytics and decision science.
- This synthesis was created after consulting established approaches that provided inputs on pieces of the process, or provided similar types of concepts with differing terminology.
- Several of the processes that were consulted include the following:

➢ Scientific Method, which, although it has been around for centuries, still provides a solid framework for thinking about and deconstructing problems into their principle parts.

➢ CRISP-DM provides some useful inputs on ways of

- Tom Davenport’s DELTA framework from his text “Analytics at Work”,
- Doug Hubbard’s Applied Information Economics (AIE) approach from his work “How to Measure Anything”.
- “MAD Skills: New Analysis Practices for Big Data” provided inputs for several of the techniques mentioned specifically in Phases 3-5 that focus on model planning, execution and key findings.
- As you can see in the graphic, you can often learn something new in a phase to cause you to go back and refine work done in prior phases given new insights and information that you’ve uncovered.

- For this reason, the graphic is shown as a cycle and the circular arrows are intended to convey that you can move iteratively between phases until you have sufficient information to continue moving forward.
- The green callouts represent questions to ask yourself to gauge whether you have enough information and have made enough progress to move to the next phase of the process.
- These are not formal phase gates, but rather serve as criteria to help you test whether it makes sense for you to stay in the current phase or move to the next one.
- Here is a brief overview of the main phases you will go through, as you proceed through the Data



Data Analytics Lifecycle

1 Discovery

2 Data Prep

3 Model Planning

4 Model Building

5 Communicate Results

6 Operationalize

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? Have we failed for sure?

EMC^2 PROVEN PROFESSIONAL

- **Phase 1: Discovery**
- Learn the business domain, including relevant history, such as whether the organization or business unit has attempted similar projects in the past, from which you can learn.
- Assess the resources you will have to support the project, in terms of people, technology, time, and data.
- Frame the business problem as an analytic challenge that can be addressed in subsequent phases.
- Formulate **Initial hypotheses** (**IH**) to test and begin learning the data.
- **Phase 2: Data Preparation**
- Prepare an **analytic sandbox**, in which you can

- Perform **ELT** and **ETL** to get data into the sandbox, and begin transforming the data so you can work with it and analyze it.
- Familiarize yourself with the data thoroughly and take steps to condition the data.

Phase 3: Model Planning

- Determine the methods, techniques and workflow you intend to follow for the model scoring.
- Explore the data to learn about the relationships between variables, and subsequently select key variables and the models you are likely to use.

Phase 4: Model Building

- Develop data sets for testing, training, and production purposes.

- Get the best environment you can for executing models and workflows, including fast hardware and parallel processing.

❑ **Phase 5: Communicate Results**

- Determine if you succeeded or failed, based on the criteria you developed in the Discovery phase, in collaboration with your stakeholders.
- Identify your key findings, quantify the business value and develop a narrative to summarize your findings and convey to stakeholders.

❑ **Phase 6: Operationalize**

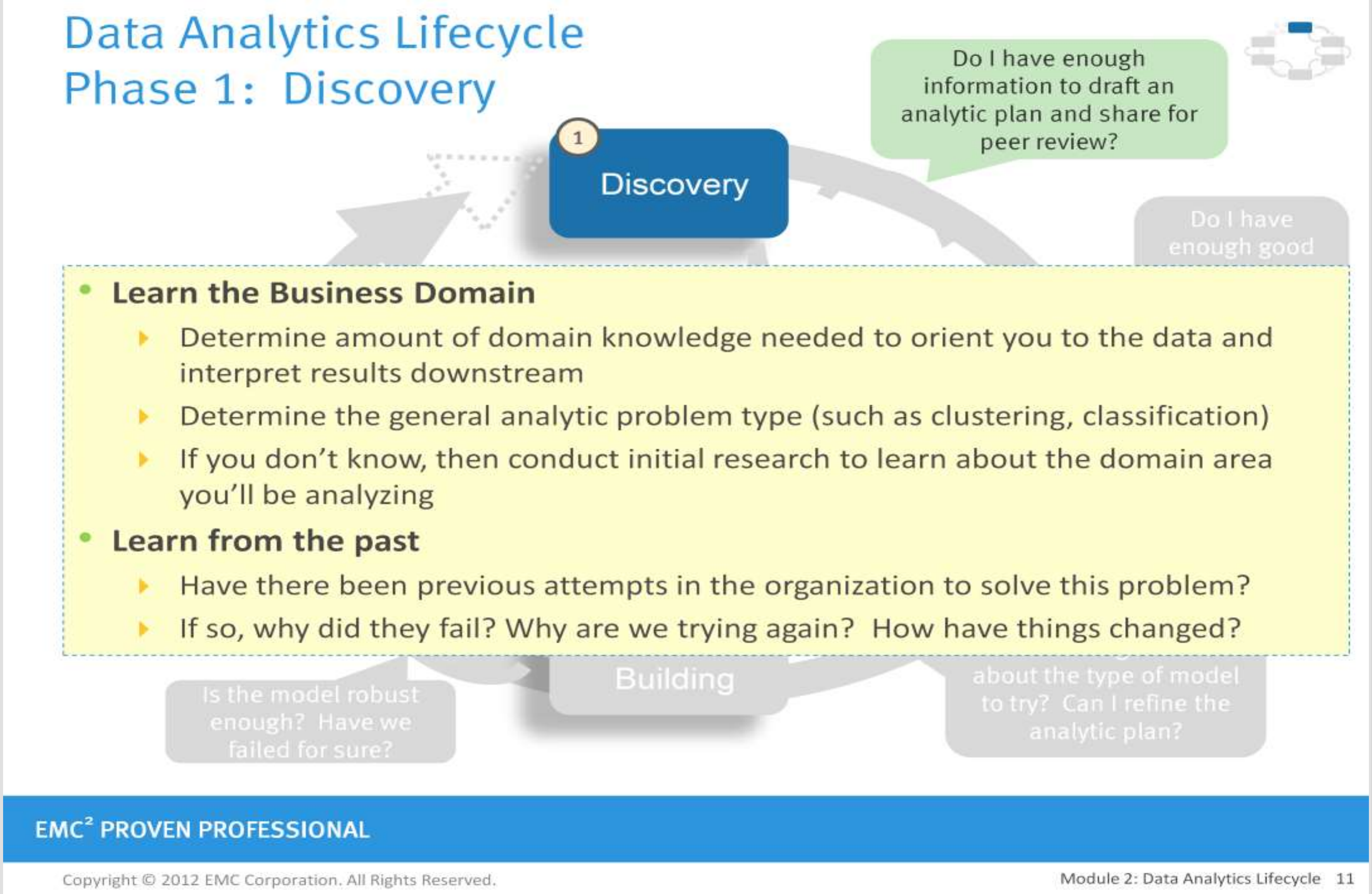
- Deliver final reports, briefings, code, and technical documents.
- Run a pilot project, and implement your models in a production environment.

- It is critical to ensure that once you have run the models and produced findings, you frame these results in a way that is appropriate for the audience that engaged you for the work in a manner that demonstrates clear value.
- If you perform a technically accurate analysis, but cannot translate the results into a language that speaks to the audience, people will not see the value and much of your time will have been wasted.

Phase 1: Discovery

- ❑ **Understanding the domain area of the problem is critical.**
- In some cases, Data Scientists will have deep computational and quantitative knowledge that can be broadly applied across many disciplines.
- An example of this role would be someone who has an advanced degree in Applied Mathematics or Statistics.
- They have deep knowledge of the methods, techniques and ways for applying heuristics to a variety of business and conceptual problems.
- Others in this area may have deep knowledge of a domain area, coupled with quantitative expertise.
- An example of this would be someone who has a Ph.D. in Life Sciences.

- This person would have deep knowledge of a field of study, such as Oceanography, Biology, Genetics, with some depth of quantitative knowledge.
- At this early stage in the process, you need to determine whether the person creating the analytical work products downstream will have sufficient knowledge of the domain (for example Genomics or Financial Services), and how tightly you will need to partner with the business sponsors, who may have deeper domain knowledge, but may have less analytical depth.
- As part of the Discovery phase, you will need to **assess the resources you will have to support the project.**
- With this scoping, **consider the available tools and technology you will be using and the types of systems you will need to interact with in later phases,**



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

about the type of model to try? Can I refine the analytic plan?

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

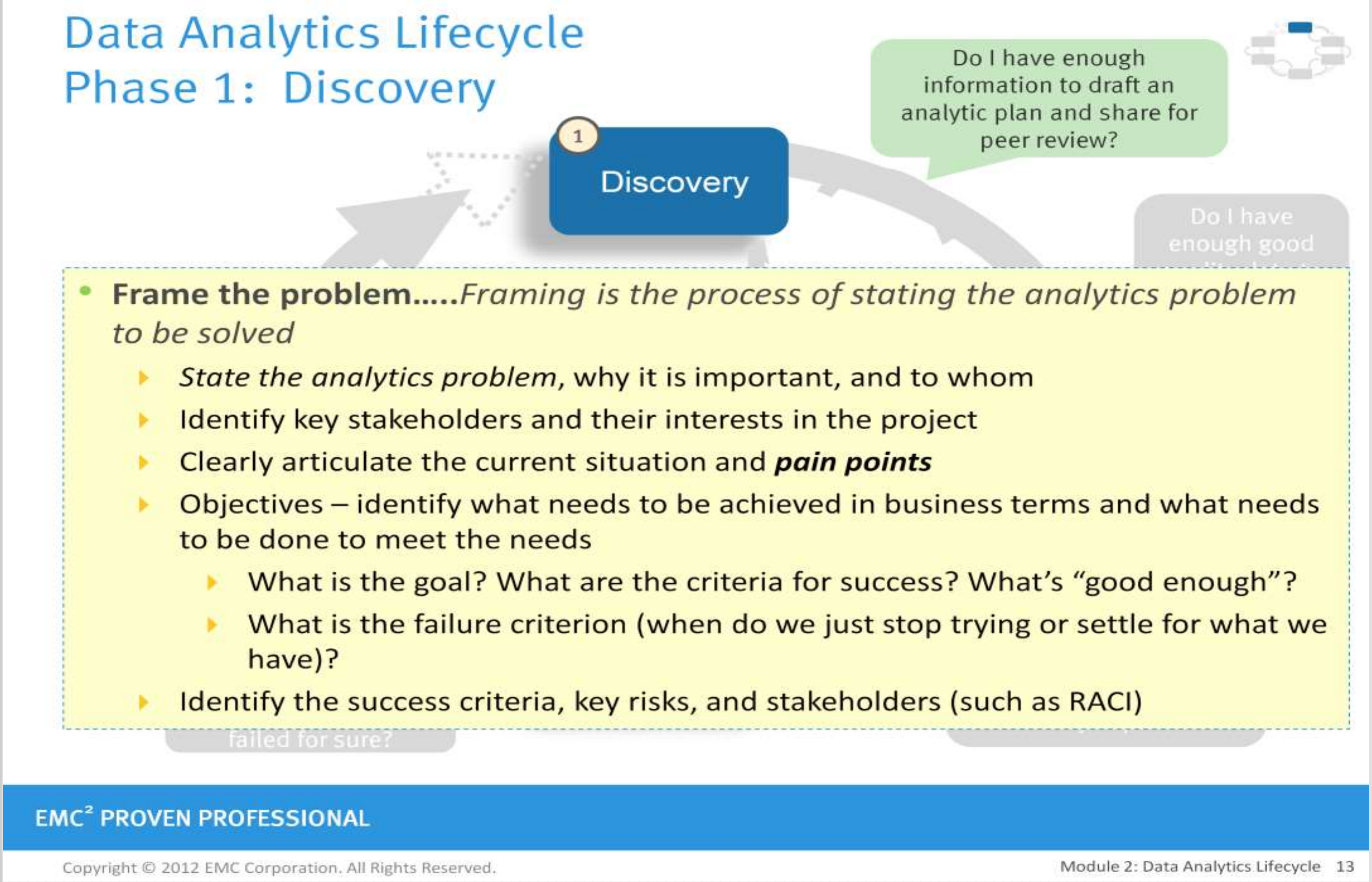
- **Learn the Business Domain**
  - Determine amount of domain knowledge needed to orient you to the data and interpret results downstream
  - Determine the general analytic problem type (such as clustering, classification)
  - If you don’t know, then conduct initial research to learn about the domain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

- In addition, try to evaluate the level of analytical sophistication within the organization.
- What types of roles will be needed for end users of the model you are developing for this approach to be successful.
- Do they exist within the organization? This will influence the techniques you select and the kind of implementation you choose to pursue in later phases.

❑ **Take inventory of the types of data available to you for the project.**

- Consider if the data available is sufficient to support the project's goals, or if you will need to collect data, purchase it from outside sources, or transform

- When considering the **project team, ensure you have the right mix** of domain experts, customers, analytic team, and project management to form an effective team.
- In addition, evaluate how much of their time you will need and if you have the right breadth and depth of skills on the working team.
- After taking inventory of the tools, technology, data and people for the project, consider if you have enough resources to succeed on this project, or if need to request additional resources.
- Negotiating for resources at the outset of the project is generally more useful to do, as you will be scoping the goals, objectives and feasibility of the project, and ensure you have sufficient time to perform it properly.



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

failed for sure?

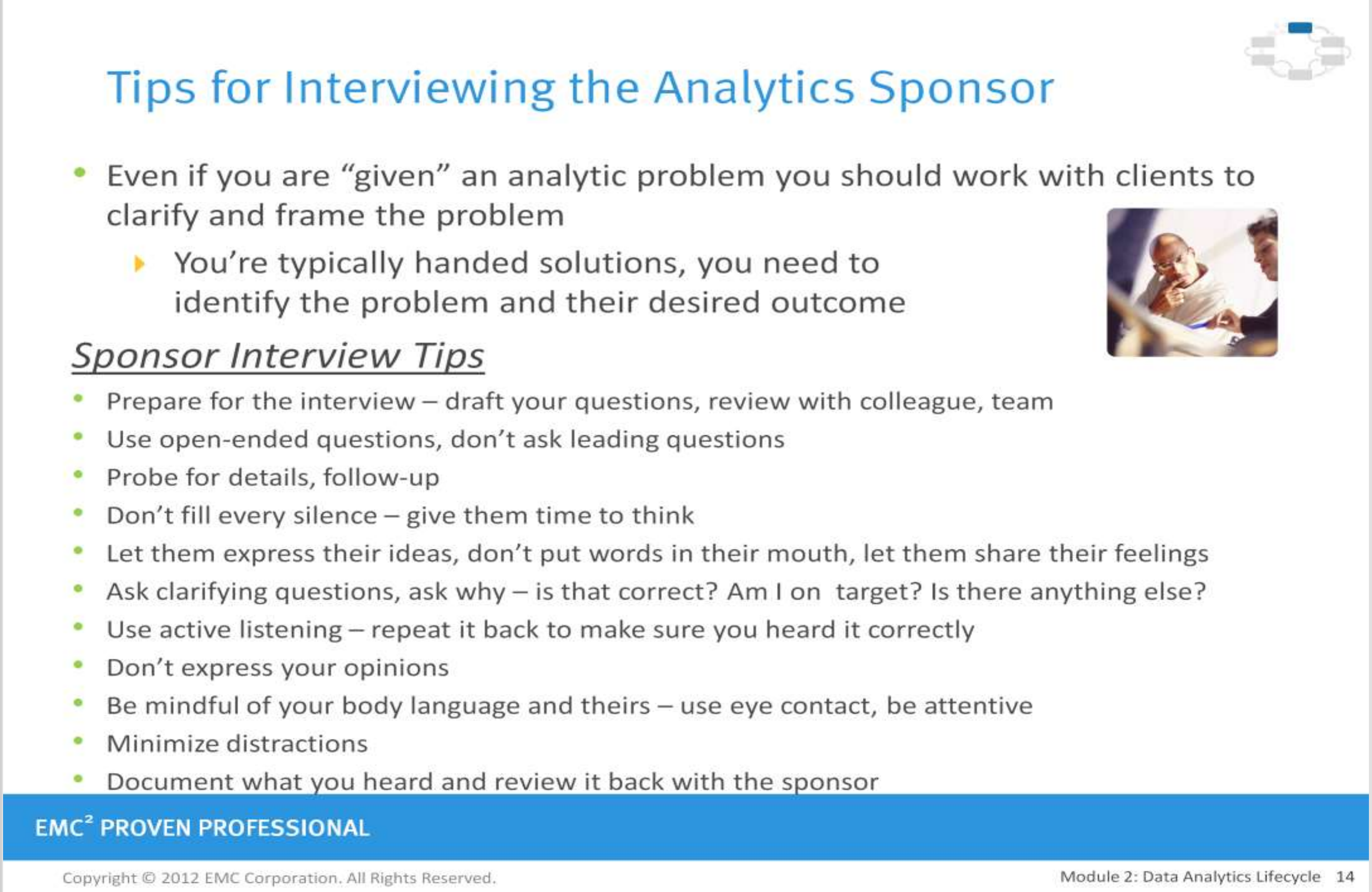
- **Frame the problem…..***Framing is the process of stating the analytics problem to be solved*
  - *State the analytics problem*, why it is important, and to whom
  - Identify key stakeholders and their interests in the project
  - Clearly articulate the current situation and ***pain points***
  - 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)

- Now that you have interviewed stakeholders, learned about the domain area and any relevant history on similar analytical projects, you can begin framing the business problem and the analytics problem.
- At this point **identify the key stakeholders and their interests.**
- For example, identify the results each stakeholder wants from the project and the criteria they will use to judge the success of the project.
- This analytical project is being initiated for a reason, and **it is critical to articulate the pain points as clearly as possible** so you can make sure to address them and be aware of areas you should pursue or avoid as you get farther along into the analytical process.

- Depending on the number of stakeholders and participants, you may consider outlining the type of activity and participation you expect from each stakeholder and participant.
- This will set clear expectations with the participants, and avoid delays later when you may feel you need you to wait on approval from someone, who views themselves as an adviser, rather than an approver of the work product.
- One instrument that is useful in this regard is a RACI matrix. RACI is a way to chart responsibilities and compare what a person thinks is his/her role in a project, what others in the project think their role is, and what the person actually will do within the project.
- RACI refers to people in each of 4 roles within a

- **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.
- Creating a framework, such as a RACI matrix, will ensure you have accountability and clear agreement on responsibilities on the project, and that the right people are kept informed of progress.

Here are some tips for conducting an interview with the project sponsor.



Tips for Interviewing the Analytics Sponsor

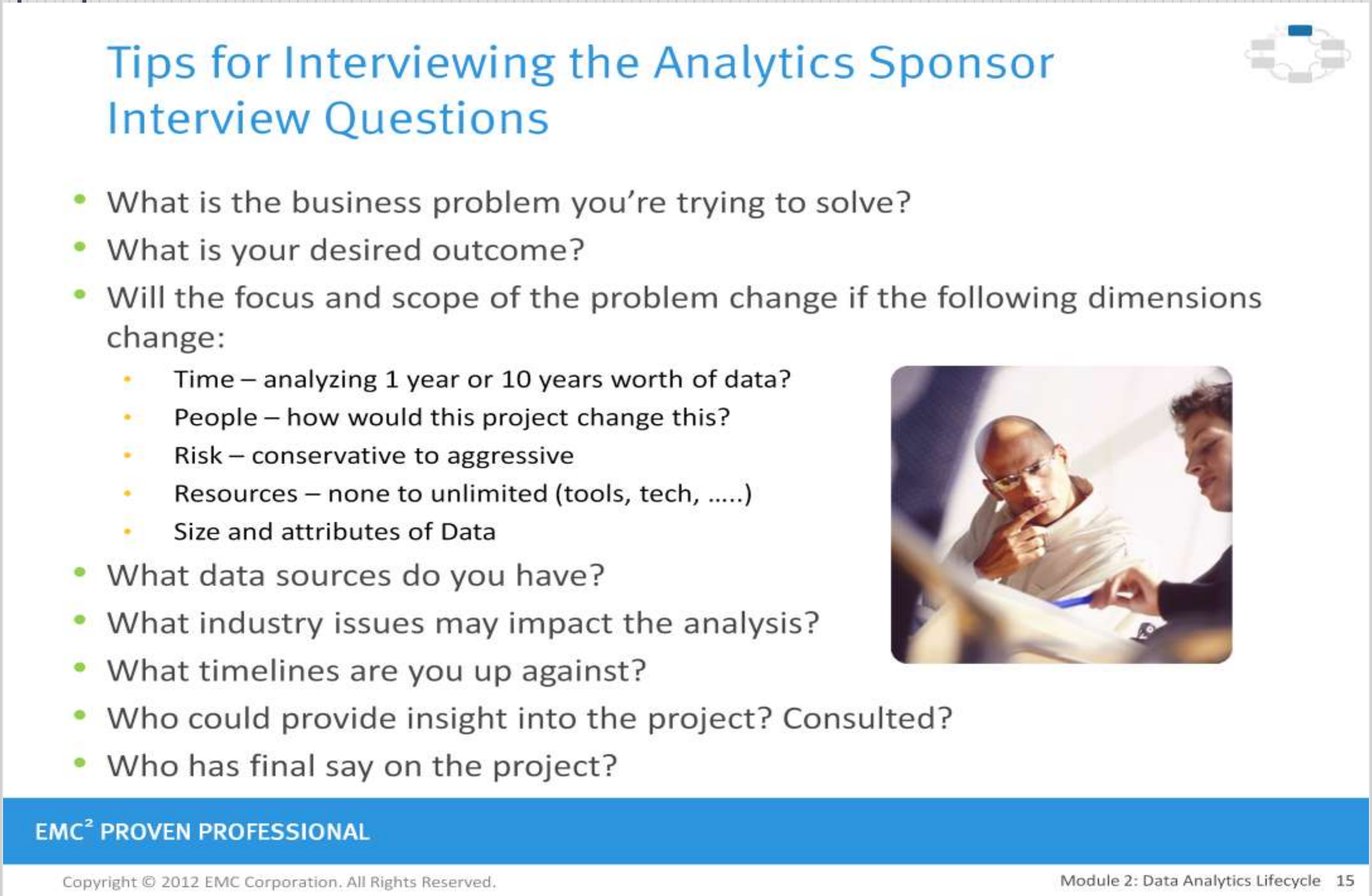
- Even if you are “given” an analytic problem you should work with clients to clarify and frame the problem
  - You’re typically handed solutions, you need to identify the problem and their desired outcome

<u>Sponsor Interview Tips</u>

- 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

EMC^2 PROVEN PROFESSIONAL

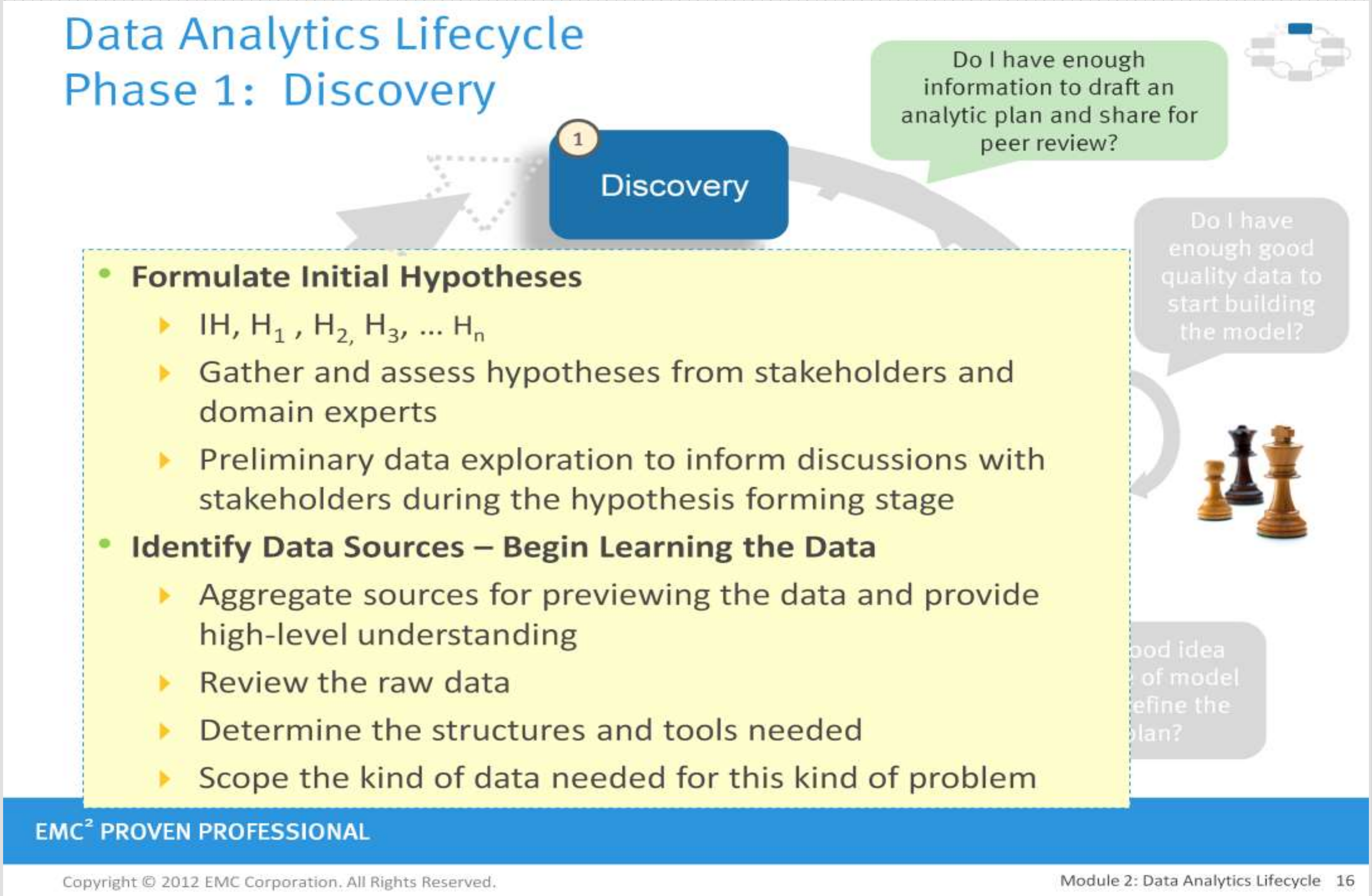
Here are some sample interview questions to be used with the sponsor in framing the core business problem and project constraints.



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?

EMC² PROVEN PROFESSIONAL



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 the model?

ood idea
of model
efine the
lan?

- **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

EMC² PROVEN PROFESSIONAL

- **Form initial hypotheses (IH) that you can prove or disprove with the data**.
- I’d encourage you to come up with a few primary hypotheses to test, and be creative about additional ones.
- These IH’s will form the basis of the tests you will analyze in later phases, and serve as the basis for additional deliberate learning.
- Hypothesis testing will be covered in greater detail in Module 3.
- As part of this initial work, **identify the kinds of data you will need to solve the problem**.
- Consider the volume, type, and time span of the data you will need to test the hypotheses.

- Also keep in mind the data sources you will need, and ensure to get access to more than simply aggregated data.
- In most cases you will need the raw data in order to run it through the models.
- Determine whether you have access to the data you need, since this will become the basis for the experiments and tests.
- Recalling the characteristics of big data from Module 1, assess which characteristics your data has, with regard to its structure, volume, and velocity of change.
- A thorough diagnosis of the data situation will inform the kinds of tools and techniques to use in

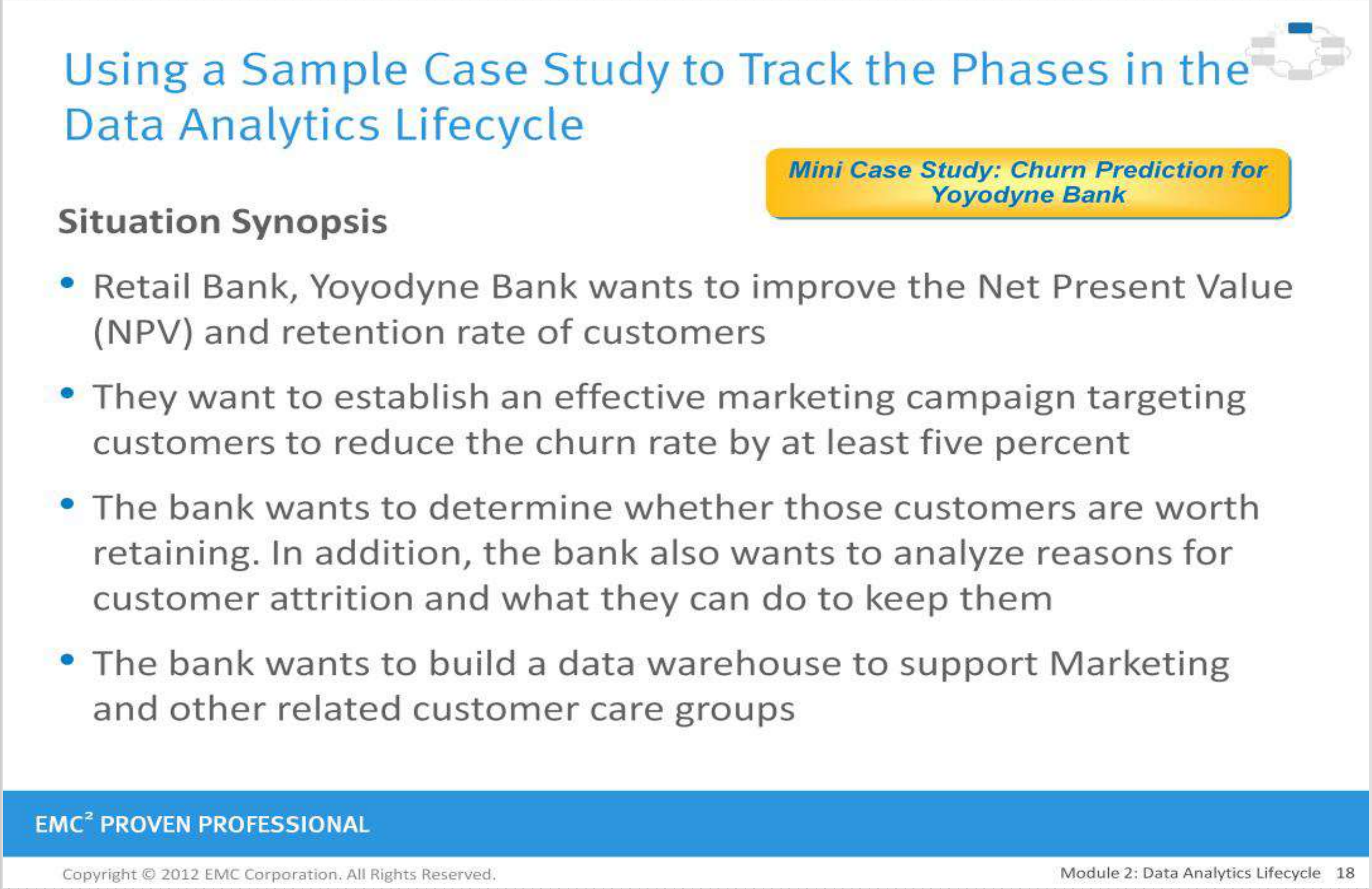
- In addition, performing data exploration in this phase will help you determine the amount of data you need, in terms of the amount of historical data to pull, the structure and format.
- Develop an idea on the scope of the data and validate with the domain experts on the project.
- For building expertise, it is critical to design experiments by first considering possible answers to a question before asking for the answer.
- In this way, you will come up with additional possible solutions to problems.

- Likewise, if you spend time formulating several initial hypotheses at the outset of a project, you will be able to generate more conclusions and more expansive findings after executing an analytic model than you otherwise would if you only began your interpretation after receiving the model’s results.

You can move to the next Phase when….

- …you have enough information to draft an analytic plan and share for peer review.

- This is not to say you need to actually conduct a peer review of your analytic plan, but it is a good test to gauge if you have a clear grasp of the business problem and have mapped out your approach to addressing it.
- This also involves a clear understanding of the domain area, the problem to be solved, and scoping the data sources to be used.
- As part of this discussion, you may want to identify success criteria for the project.
- Creating this up front will make the problem definition even more clear, and help you when it comes to time make choices about the analytical methods being used in later phases.



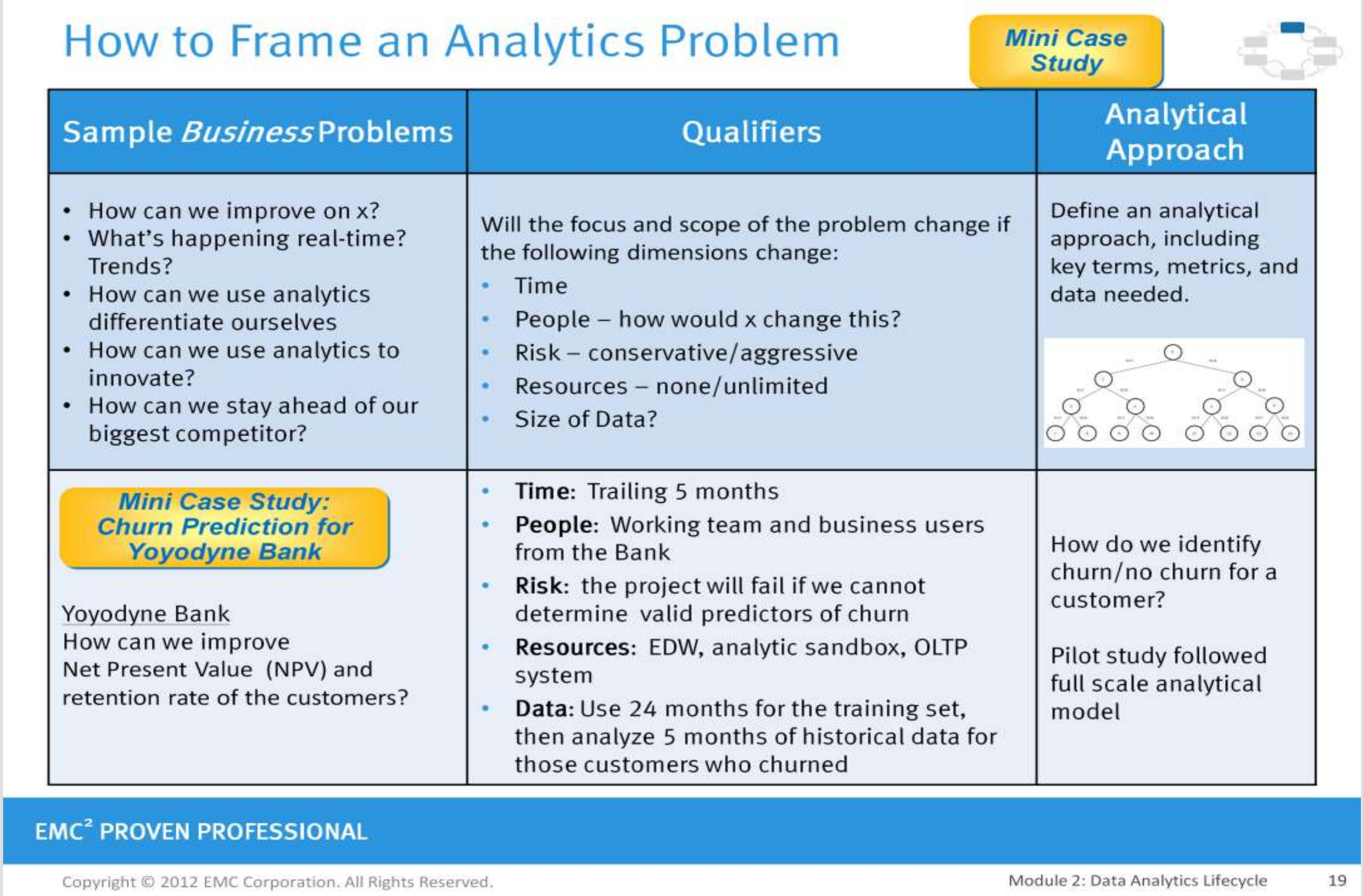
Using a Sample Case Study to Track the Phases in the Data Analytics Lifecycle

Mini Case Study: Churn Prediction for Yoyodyne Bank

Situation Synopsis

- Retail Bank, Yoyodyne Bank wants to improve the Net Present Value (NPV) and retention rate of customers
- They want to establish an effective marketing campaign targeting customers to reduce the churn rate by at least five percent
- The bank wants to determine whether those customers are worth retaining. In addition, the bank also wants to analyze reasons for customer attrition and what they can do to keep them
- The bank wants to build a data warehouse to support Marketing and other related customer care groups

EMC² PROVEN PROFESSIONAL



How to Frame an Analytics Problem

Mini Case Study

Sample *Business* Problems	Qualifiers	Analytical Approach
• How can we improve on x? • What's happening real-time? Trends? • How can we use analytics 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.
Mini Case Study: Churn Prediction for Yoyodyne Bank Yoyodyne Bank 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, 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

1. What are your initial hypotheses (IH)?
2. What data will you need to test the IH?
3. 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

- After initial data exploration, 5 months appears to capture relevant time period
- 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 .

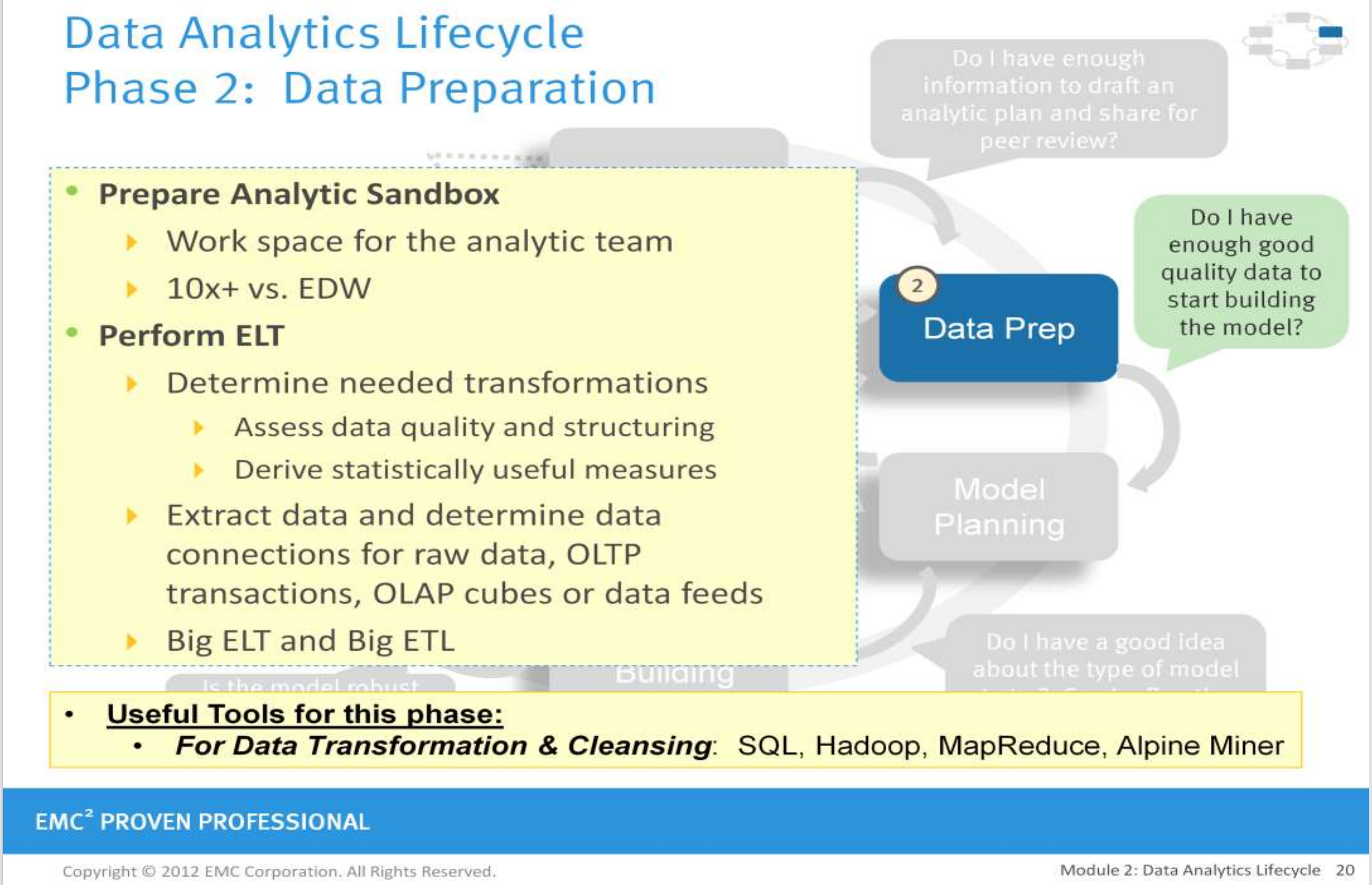
1)What is Net Present Value?

2)Discuss revenue and cost components for a retail bank customer.

3)How do you define “retention rate”?

4)What is a churn rate?

5)Can we measure the current churn rate? If so, how?



Data Analytics Lifecycle
Phase 2: Data Preparation

- **Prepare Analytic Sandbox**
  - Work space for the analytic team
  - 10x+ vs. 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 transactions, OLAP cubes or data feeds
  - Big ELT and Big ETL

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
Building
Is the model robust

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

EMC² PROVEN PROFESSIONAL

- Shown is an overview of the Data Analytics Lifecycle for Phase 2, with its associated sub-phases.
- **Of all of the phases, the step of Data Preparation is generally the most iterative and time intensive.**
- In this step, you will need to 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 since that will be tightly controlled and needed for financial reporting.

- You should be **collecting all kinds of data in your sandbox,** as you will need access to high volumes and varieties of data for a Big Data analytics project.
- This can include everything from summary, structured data, to raw data feeds, to unstructured text data from call logs or web logs, depending on the kind of analysis you are looking to undertake.
- Expect this sandbox to be large, at least 10 times the size of a company’s EDW.
- Make sure you have strong bandwidth and network connections to the underlying data sources so you can quickly do transformations on the data or extractions from data sets you will need.

- Note that the graphic above indicates doing “ELT”, rather than ETL, which is a more typical approach to approaching data extractions.
- 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, where analysts can choose to transform the data into a new state or leave it in its original, raw condition.

- 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.
- For instance, consider the example of an analysis for fraud detection on credit card usage.
- Many times, the outliers in this data population can represent higher-risk transactions that may be indicative of fraudulent credit card activity.
- Using ETL, these outliers may be inadvertently filtered out or transformed and cleaned before being loaded into the database.
- For this reason, ELT is encouraged so that you have data in its raw state and also the ability to transform it

- **This approach will give you clean data to analyze after its in the database and the data in its original form for finding hidden nuances in the data.**

❑ **Tools**

- Hadoop can perform massively parallel ingest and custom analysis for parsing web traffic, GPS location analytics, genomic analysis and for combining massive unstructured data feeds from multiple sources.
- Alpine Miner provides a GUI interface for creating analytic workflows, including data manipulations and a series of analytic events such as staged data mining techniques (eg., first select top 100 customers, then run descriptive statistics and clustering) on Postgres SQL and other big data

- **People**

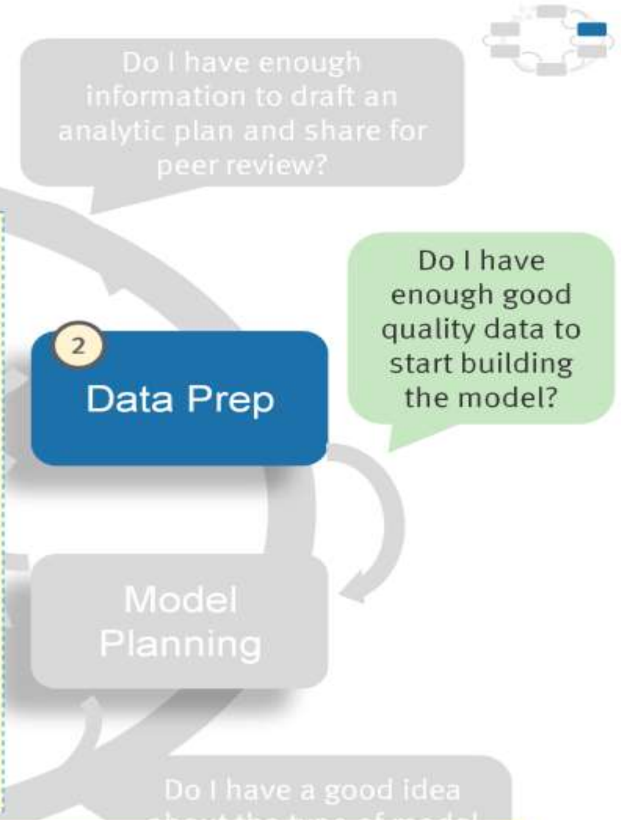
- For Phase 2, you will need assistance from IT, DBAs or whoever controls the EDW or data sources you will be using.
- Phase 2 is critical within the analytics lifecycle.
- **If you do not have data of sufficient quality or cannot get good data, you will not be able to perform the subsequent steps in the lifecycle process.**

In addition to what's shown in the graphic above, here are additional aspects to consider during this phase of a project, and common pitfalls to avoid…

- **What are the data sources? What are the target fields (e.g. columns of the tables)**

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
  - Discern 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?
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

- **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

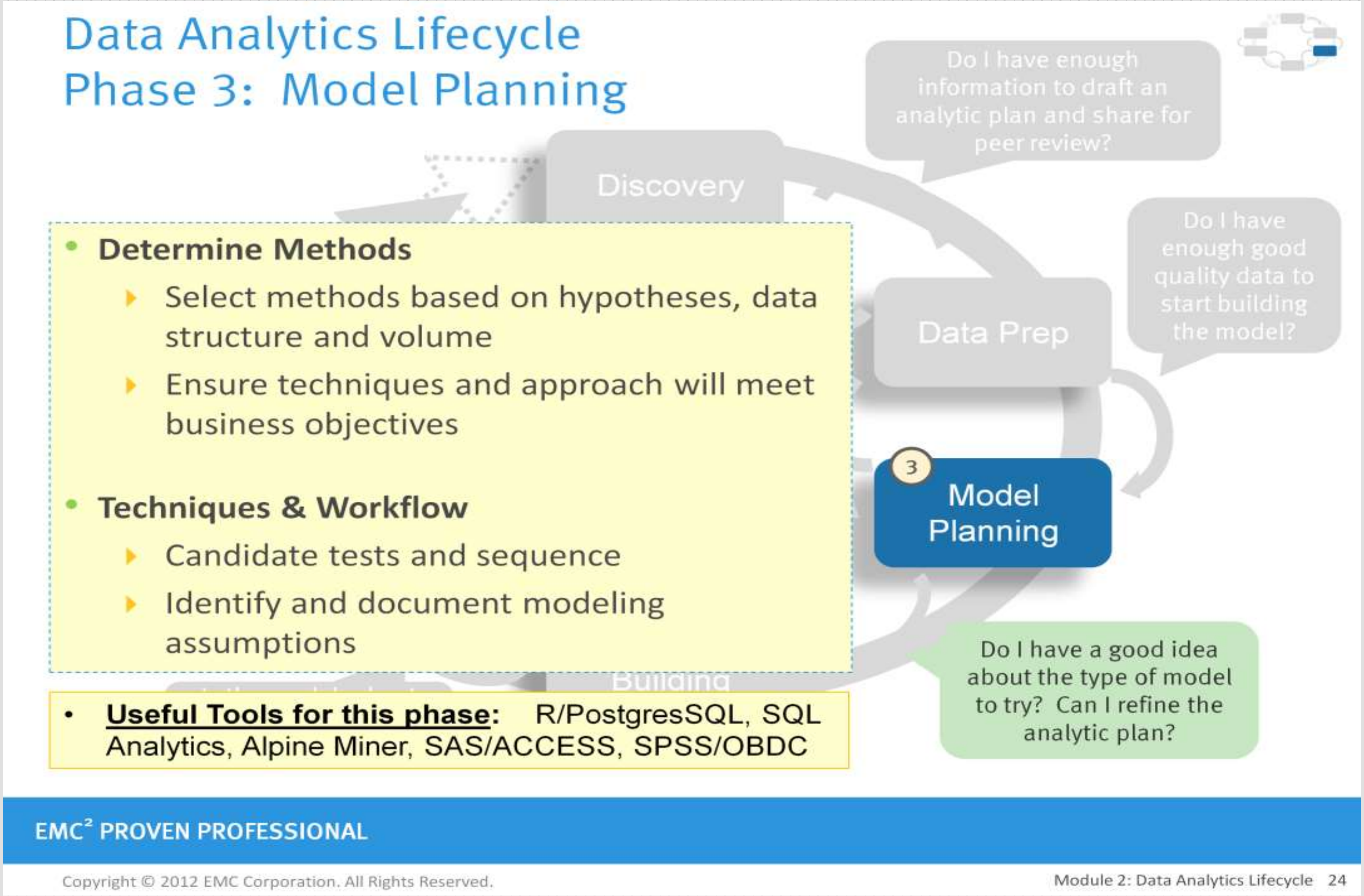
- **How clean is the data?** How consistent are the contents and files? Determine to what degree you have missing or inconsistent values, and if you have values deviating from normal. Assess the consistency of the data types. For instance, if you are expecting certain data to be numeric, confirm it is numeric or if it is a mixture of alphanumeric strings and text. Review contents of data columns or other inputs and check to ensure they make sense. 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.

- 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.

- Review data to ensure that calculations remained consistent within columns or across tables for a given data field. For instance, did customer lifetime value change at some point in the middle of your data collection, or if working with financials, did the interest calculation change from simple to compound at the end of the year?

- 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

- For geospatial data sets, are state abbreviations consistent across the data? Are personal names normalized? English units? Metric units?
- These are some typical considerations that should be part of your thought process as you assess the data you have to work with. Getting deeply knowledgeable about the data will be critical when it comes time to construct and run your models later in the process.
- **You can move to the next Phase when….**you have enough good quality data to start building the model.



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?

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?

EMC² PROVEN PROFESSIONAL

- Phase 3 represents the last step of preparations before executing the analytical models and, as such, requires you to be thorough in planning the analytical work and experiments in the next phase.
- 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.
- These hypotheses will help you frame the analytics you’ll execute in Phase 4, and choose the right methods to achieve your objectives.

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 (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.**

- **Determine if your situation warrants a single test** (eg., Binomial Logistic Regression or Market Basket Analysis) or a series of techniques as part of a larger analytic workflow.

- A tool such as Alpine Miner will enable you to set up a series or steps and analyses (eg., select top 100,000 customers ranked by account value, then predict the likelihood of churn based on another set of heuristics) and can serve as a front end UI for manipulating big data sources in Postgres SQL.
- In addition to the above, consider how people generally solve this kind of problem and look to address this type of question.
- With the kind of data and resources you have available, consider if similar approaches will work or if you will need to create something new

- Many times you can get ideas from analogous problems people have solved in different industry verticals.

Tools

There are many tools available to you, here are a few….

- **R** has a complete set of modeling capabilities and provides a good environment for building interpretive models with high quality code.

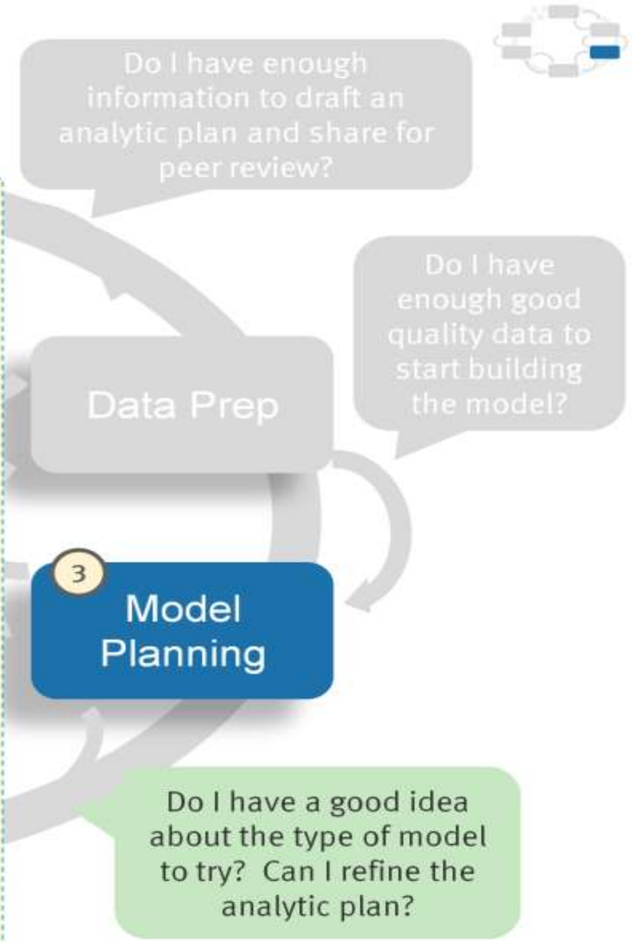
- In addition, it has the ability to interface with Postgres SQL and execute statistical tests and analyses against big data via an open source connection (R/PostgresSQL).
- These two factors make R well suited to performing statistical tests and analytics on big data.

- R contains over 3,000 packages for analysis and graphical representation.
- New packages are being posted all the time, and many companies are providing value add services for R (training, instruction, best practices), as well as packaging it in ways to make it easier to use and more robust.
- This phenomenon is similar to what happened with Linux in the late 1980s and early 1990s, where companies appeared to package and make Linux easier for companies to consume and deploy.
- Use R with file extracts for off-line analysis and optimal performance.
- Use R/Postgres SQL for dynamic queries and faster development.

- **SQL Analysis** services can perform in-database analytics of common data mining functions, involved aggregations and basic predictive models.
- **SAS/ACCESS** provides integration between SAS and external database via multiple data connectors such as OBDC, JDBC, and OLE DB.
- SAS itself is generally used on file extracts, but with SAS / ACCESS, you can connect to relational databases (such as Oracle or Teradata), and data warehouse appliances (such as Greenplum or Aster), files, and enterprise applications (eg., SAP, Salesforce, etc.).

Data Analytics Lifecycle
Phase 3: Model Planning

- **Data Exploration**
- **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?

EMC² PROVEN PROFESSIONAL

Data exploration

- There is some exploration in the data prep phase, mostly for data hygiene reasons and to assess the quality of the data itself. In this phase, it is important to explore the data to **understand the relationships among the variables** to inform selection of the variables and methods, and to understand the problem domain.
- Using tools to help you with data visualizations can help you with this, and aid you in previewing the data and assessing relationships between variables.

Variable Selection

- In many cases, stakeholders and subject matter experts will have gut feelings about what you should be considering and analysis

- Likely, they had some hypothesis which led to the genesis of the project.
- Many times stakeholders have a good grasp of the problem and domain, though they may not be aware of the subtleties within the data or the model needed to prove or disprove a hypothesis.
- Other times, stakeholders may be correct, but for an unintended reason (correlation does not imply causation).
- Data scientists have to come in unbiased, and be ready to question all assumptions.
- Consider the inputs/data that you think you will need, then examine whether these inputs are actually correlated with the outcomes you are trying

- Some methods and types of models will handle this well, others will not handle it as well.
- Depending on what you are solving for, you may need to consider a different method, winnow the inputs, or transform the inputs to allow you to use the best method.
- Aim for capturing the most essential predictors and variables, rather considering every possible variable that you think may influence the outcome.
- This will require iterations and testing in order to identify the most essential variables for the analyses you select.
- Test a range of variables to include in your model, then focus on the most important and influential variables.

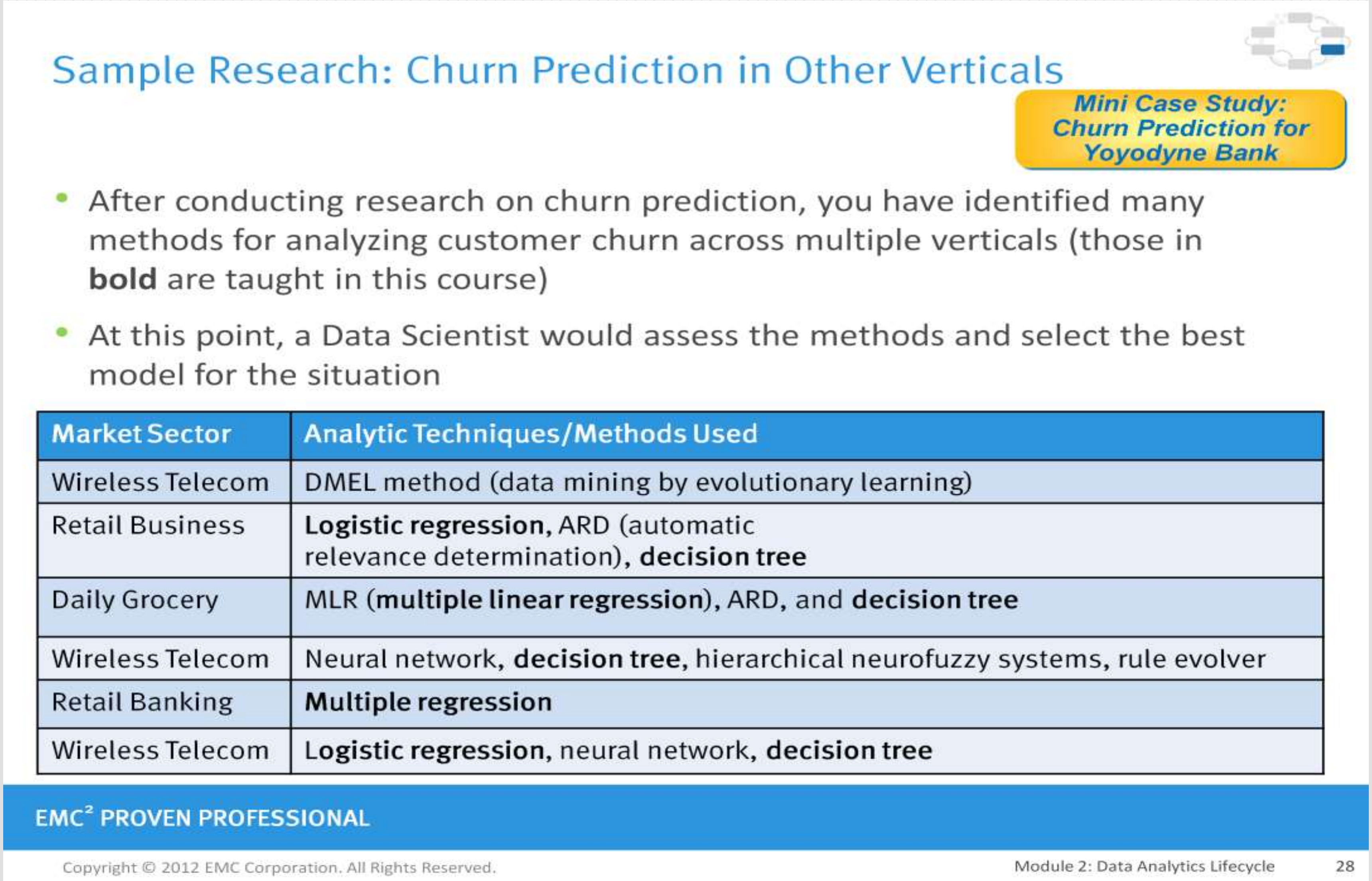
- If running regressions, identify the candidate predictors and outcomes of the model.
- Look to create variables that will determine outcomes, but will provide strong relationship to outcome rather than to each other.
- Be vigilant for problems such as serial correlation, co-llinearity and other typical data modeling problems that will interfere with the validity of these models.
- Consider this within the context of how you framed the problem. Sometimes correlation is all you need (“black box prediction”), and in other cases you will want the causal relationship (when you want the model to have explanatory power, or when you want to forecast/stress test under situations a bit out of your range of observations – always dangerous).

Model Selection

- **Converting the model created in R or a native statistical package to SQL will enable you to run the operation in-database, which will provide you with the optimal performance during runtime.**
- Consider the major data mining and predictive analytical techniques, such as Categorization, Association Rules, and Regressions.
- Determine if you will be using techniques that are best suited for structured data, unstructured data, or a hybrid approach. You can leverage MapReduce to analyze unstructured data.

- **You can move to the next Phase when….**you have a good idea about the type of model to try and you can refine the analytic plan.
- This includes a general methodology, solid understanding of the variables and techniques to use, and a description or diagramming of the analytic workflow.
- Here are *sample* analytical methods used across multiple market segments for churn prediction.
- Some of these methods we will cover later in Module 4.
- Several of these methods are out of scope for this course.
- Within the context of the case study, at this stage, you have conducted research and interviewed select

These discussions should provide you with ideas for appropriate analytical methods for the pilot study.

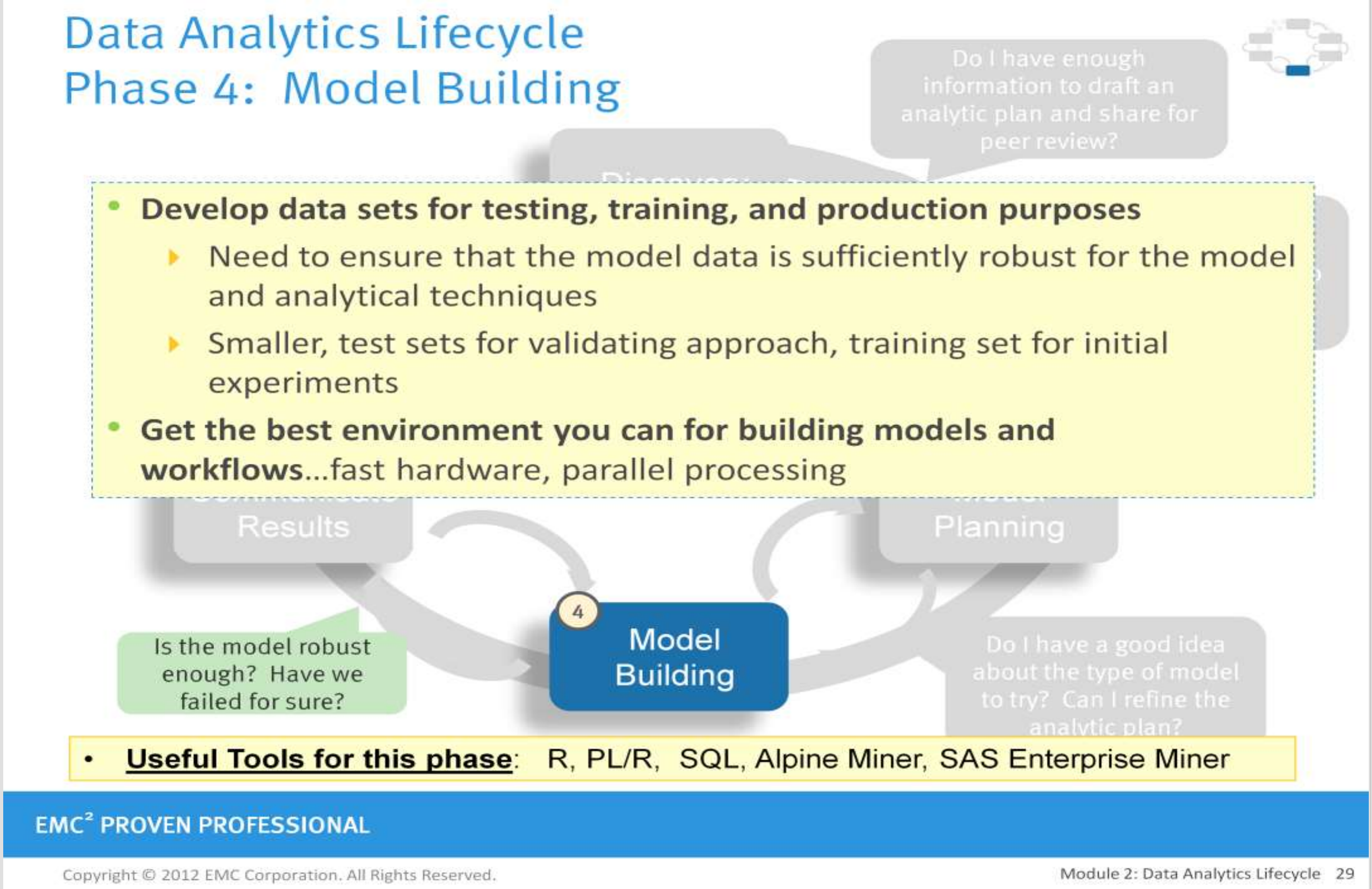


Sample Research: Churn Prediction in Other Verticals

Mini Case Study: Churn Prediction for Yoyodyne Bank

- After conducting research on churn prediction, you have identified many methods for analyzing customer churn across multiple verticals (those in **bold** are taught in this course)
- 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)
Retail Business	**Logistic regression**, ARD (automatic relevance determination), **decision tree**
Daily Grocery	MLR (**multiple linear regression**), ARD, and **decision tree**
Wireless Telecom	Neural network, **decision tree**, hierarchical neurofuzzy systems, rule evolver
Retail Banking	**Multiple regression**
Wireless Telecom	**Logistic regression**, neural network, **decision tree**



Data Analytics Lifecycle Phase 4: Model Building

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

- **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
  - Smaller, 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

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?

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

EMC² PROVEN PROFESSIONAL

- In this phase, the model is fit on the training data and evaluated (scored) against the test data.
- Generally this work takes place in the sandbox, not in the live production environment.
- The phases of Model Planning and Model Building overlap quite a bit, and in practice one can iterate back and forth between the two phases for a while before settling on a final model.
- Some methods require the use of a training data set, depending on whether it is a supervised or unsupervised algorithm for machine learning.

- **Although the modeling techniques and logic required to build this step can be highly complex, the actual duration of this phase can be quite short,** compared with all of the preparation required on the data and defining the approaches.
- In general, plan to spend more time preparing and learning the data (Phases 1-2) and crafting a presentation of the findings (Phase 5), where phases 3 and 4 tend to move more quickly, although more complex from a conceptual standpoint.
- As part of this phase, you’ll need to conduct these steps:

1) Execute the models defined in Phase 3

2) Where possible, convert the models to SQL or similar, appropriate database language and execute as in-database functions, since the runtime will be significantly faster and more efficient than running in memory. (execute R models on large data sets as PL/R or SQL (PL/R is a PostgreSQL language extension that allows you to write PostgreSQL functions and aggregate functions in R).

- SAS Scoring Accelerator enables you to run the SAS models in database, if they were created using SAS Enterprise Miner.

3) Use R (or SAS) models on file extracts for testing and small data sets

4) Assess the validity of the model and its results (for instance, does it account for most of the data, and does it have robust predictive power?)

5) Fine tune the models to optimize the results (for example modify variable inputs)

6) Record the results, and logic of the model

While doing these iterations and refinement of the model, consider the following:

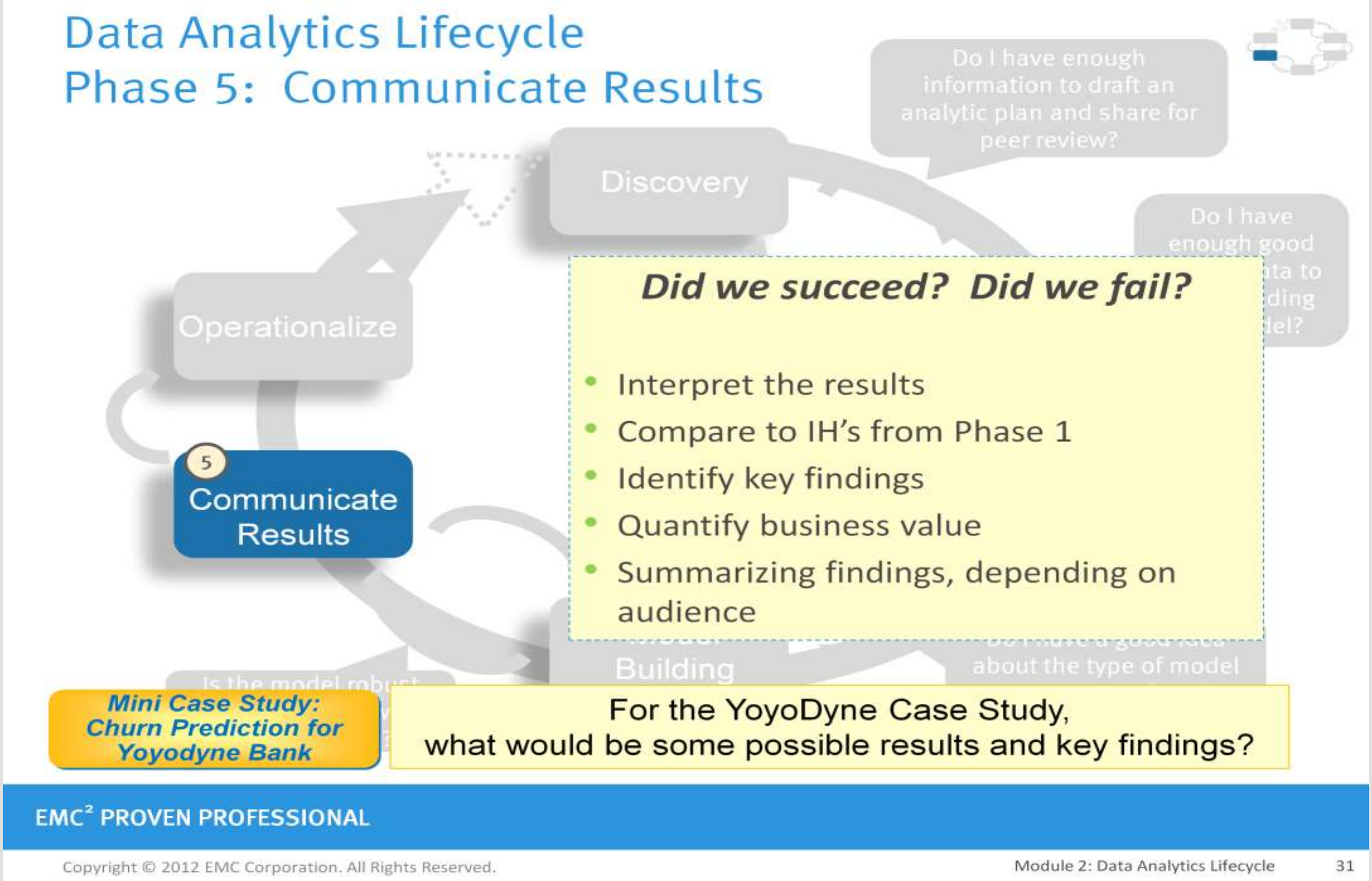
- Does the model look valid and accurate on the test data?
- Does the model output/behavior makes sense to the domain experts? That is, does it look like the model is giving “the right answers”, or answers that make sense in this context?

- Is the model accurate enough to meet the goal?
- Is it avoiding the kind of mistakes it needs to avoid? Depending on context, false positives may be more serious or less serious than false negatives, for instance. (False positives and negatives will be discussed further in Module 3)
- Do the parameter values of the fitted model make sense in the context of the domain?
- Do you need more data or more inputs? Do you need to transform or eliminate any of the inputs?
- Do you need a different form of model? If so, you’ll need to go back to the Model Planning phase and revise your modeling approach.

You can move to the next Phase when…. you can gauge if the model you've developed is robust enough, or if you have failed for sure.

Phase 5: Results & Key Findings

- Now that you've run the model, you need to go back and **compare the outcomes to your criteria for success and failure.**
- Consider how best to articulate the findings and outcome to the various team members and stakeholders. Make sure to consider and include caveats, assumptions, and any limitations of results. Remember that many times the presentation will be circulated within the organization, so be thoughtful of how you position the findings and clearly articulate the outcomes.



Data Analytics Lifecycle
Phase 5: Communicate Results

Did we succeed? Did we fail?

- Interpret the results
- Compare to IH's from Phase 1
- Identify key findings
- Quantify business value
- Summarizing findings, depending on audience

Mini Case Study: Churn Prediction for Yoyodyne Bank

For the YoyoDyne Case Study,
what would be some possible results and key findings?

Discovery
Do I have enough information to draft an analytic plan and share for peer review?
Do I have enough good ...ta to ...ding ...del?
Operationalize
5
Communicate Results
Building
about the type of model
Is the model robust

- **Make recommendations for future work or improvements** to existing processes, and consider what each of the team members and stakeholders need from you in order to fulfill their responsibilities.
- For instance, Sponsors have to champion the project, Stakeholders have to understand how the model affects their processes (for instance, if it's a churn model, marketing has to understand how to use the churn model predictions in planning their interventions), and Production engineers need to operationalize the work that’s been done.
- In addition, this is the phase where you can underscore the business benefits of the work, and begin making the case to eventually put the logic into a live production environment.

Now that you have run the model, you can do the following:

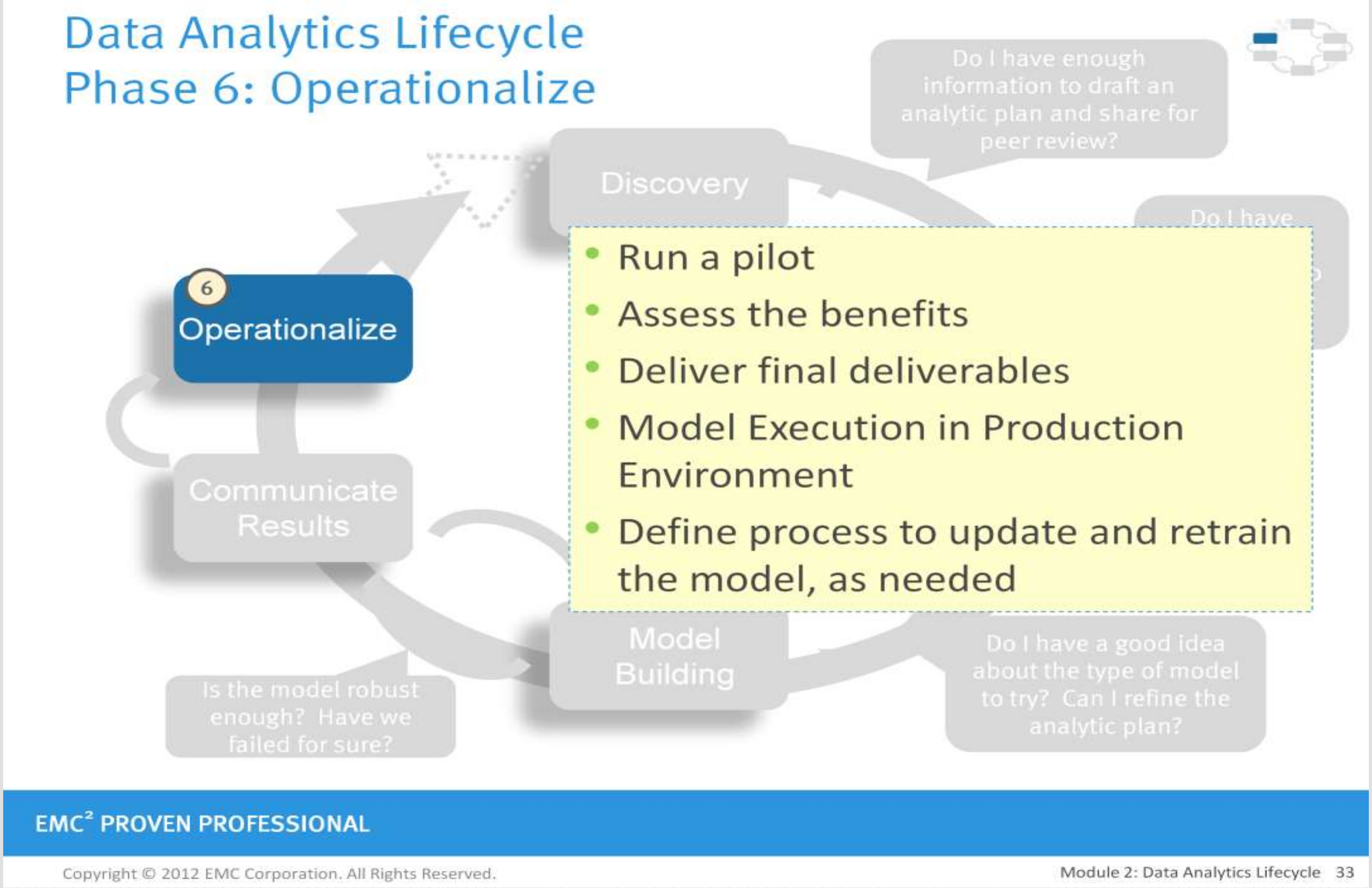
1) Assess the results of the models.

1. Are the results statistically significant and valid? If so, which aspects of the results stand out? If not, what adjustments do you need to make to refine and iterate on the model to make it valid?
2. Which data points are surprising, and which were in line with your incoming hypotheses that you developed in Phase 1? Comparing the actual results to the ideas you formulated early on typically produces additional ideas and insights that would have been missed if you did not take time to formulate IHs early in the process.

2) What do you observe in the data as a result of the analytics?

1. Of these, what are the 3 most significant findings?
2. What is the business value and significance of these findings? Depending on what emerged as a result of the model, you may need to spend time quantifying the business impact of the results to help you prepare for the presentation. For further reading, see Doug Hubbard’s book: “How to Measure Anything”, which provides an excellent resource for teaching people how to assess “intangibles” in business, and quantify the value of seemingly un-measurable things.

- **As a result of this phase, you will have documented the key findings and major insights as a result of the analysis.**
- The deliverable as a result of this phase will be the most visible portion of the process to the outside stakeholders and sponsors, so take care to clearly articulate the results, methodology, and business value of your findings.



Data Analytics Lifecycle Phase 6: Operationalize

- Run a pilot
- Assess the benefits
- Deliver final deliverables
- Model Execution in Production Environment
- Define process to update and retrain the model, as needed

Discovery

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

Do I have

6 Operationalize

Communicate Results

Model Building

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

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

EMC² PROVEN PROFESSIONAL

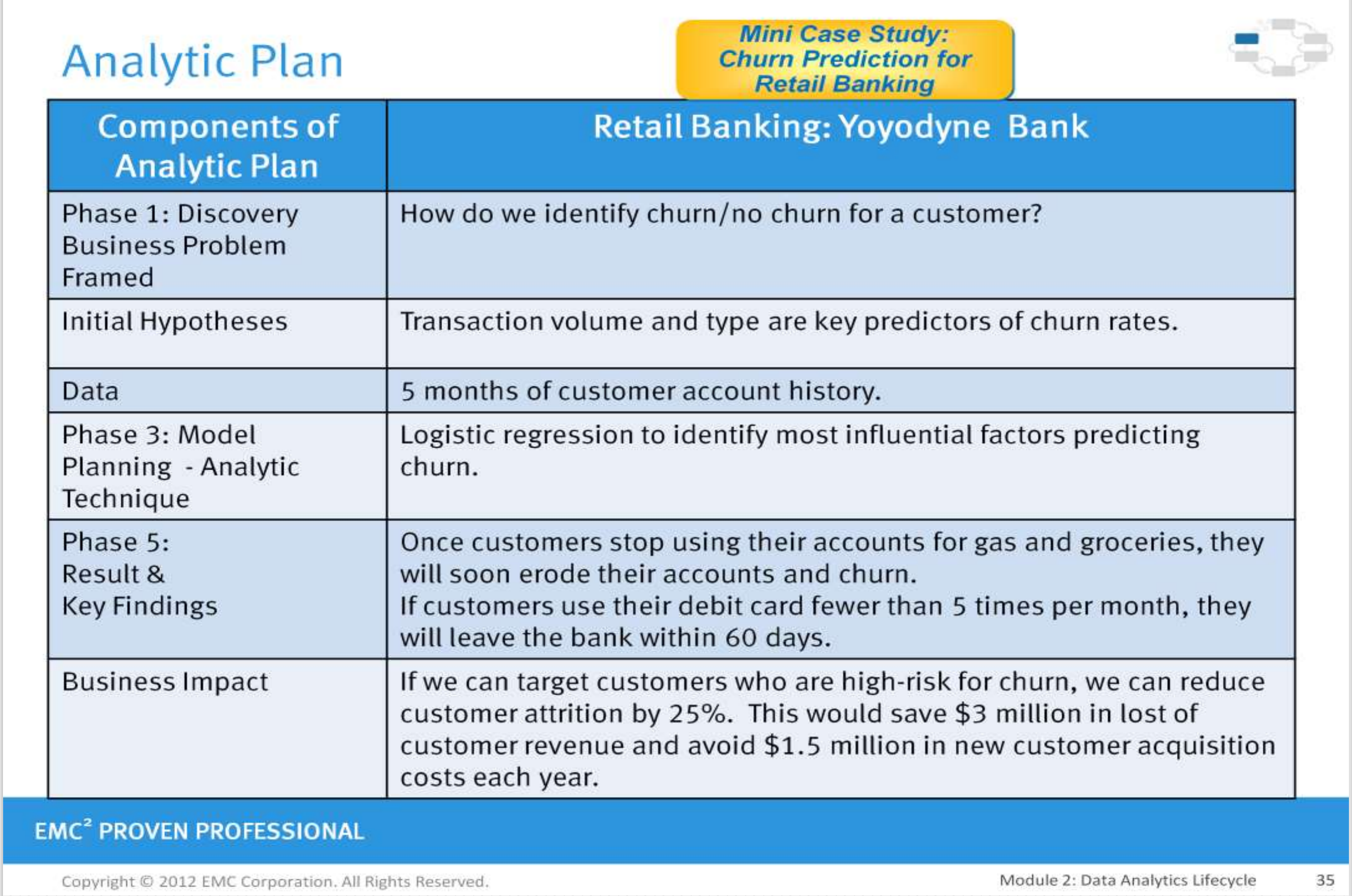
- In this phase, you will need to assess the benefits of the work that’s been done, and setup a pilot so you can deploy the work in a controlled way before broadening the work to a full enterprise or ecosystem of users.
- In phase 4, you scored the model in the sandbox, and phase 6 represents the first time that most analytics approach deploying the new analytical methods or models in a production environment.
- Rather than deploying this on a wide scale basis, we recommend that you **do a small scope, pilot deployment first.**
- **Taking this approach will allow you to limit the amount of risk** relative to a full, enterprise deployment and learn about the performance and related constraints on a small scale and make fine

- As you scope this effort, consider running the model in a product environment for a discrete set of single products, or a single line of business, which will test your model in a live setting.
- This will allow you learn from the deployment, and make any needed adjustments before launching across the enterprise.
- Keep in mind that this phase can bring in a new set of team members – namely those engineers who are responsible for the production environment, who have a new set of issues and concerns.
- They want to ensure that running the model fits smoothly into the production environment and the model can be integrated into downstream processes.

- **While executing the model in the production environment, look to detect anomalies on inputs before they are fed to the model.**
- **Assess run times and gauge competition for resources with other processes in the production environment.**
- **After deploying the model, conduct follow up to reevaluate the model after it has been in production for a period of time.**
- Assess whether the model is meeting goals and expectations, and if desired changes (such as increase in revenue, reduction in churn) are actually occurring.

- If these outcomes are not occurring, determine if this is due to a model inaccuracy, or if its predictions are not being acted on appropriately.
- If needed, automate the retraining/updating of the model. In any case, you will need ongoing monitoring of model accuracy, and if accuracy degrades, you will need to retrain the model.
- If feasible, design alerts for when model is operating “out-of-bounds”.
- This includes situations when the inputs are far beyond the range that the model was trained on, which will cause the outputs of the model to be inaccurate.
- If this begins to happen regularly, retraining is called

This slide represents an outline of an analytic plan for the mini case study.



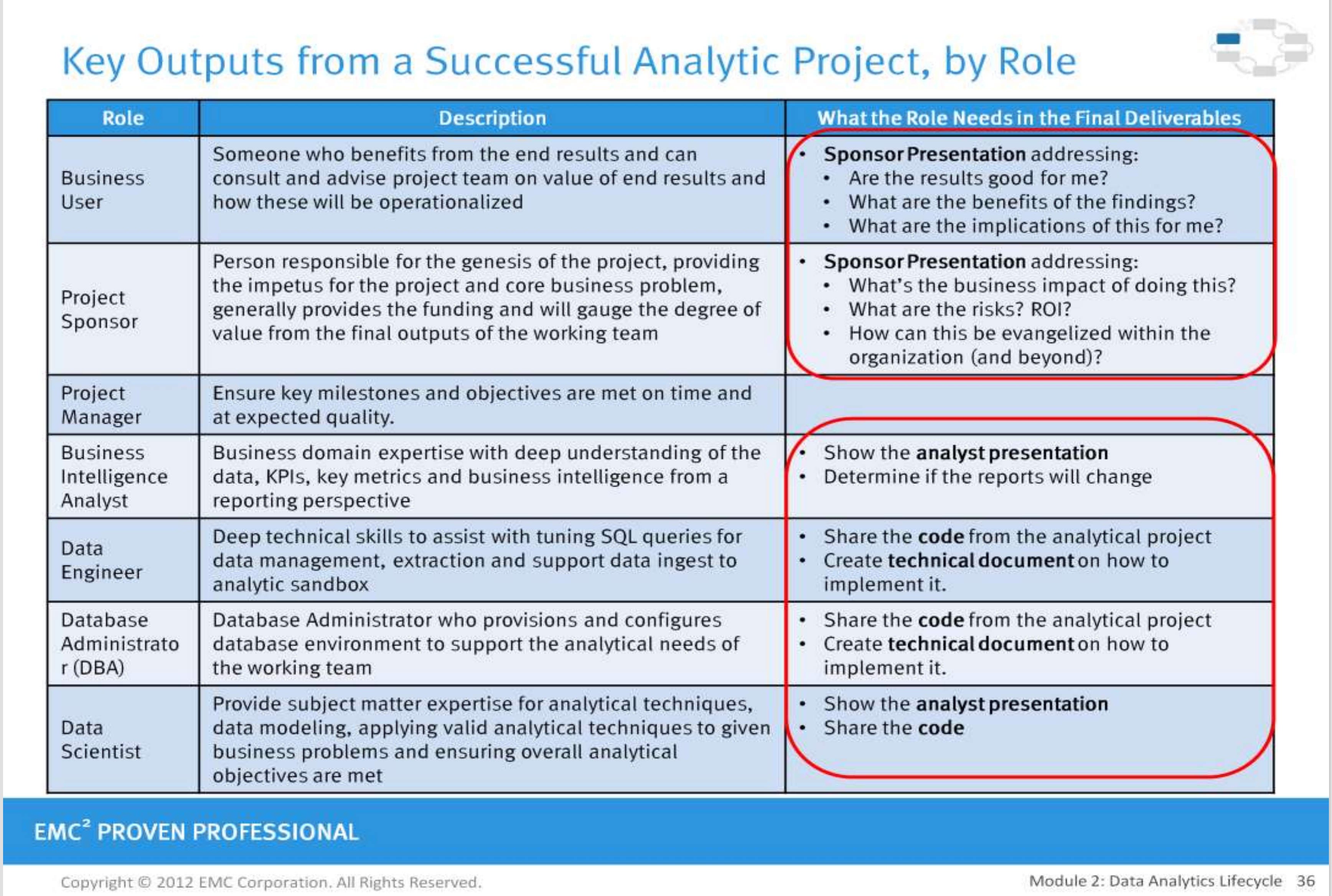
Analytic Plan

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?
Initial Hypotheses	Transaction volume and type are key predictors of churn rates.
Data	5 months of customer account history.
Phase 3: Model Planning - Analytic Technique	Logistic regression to identify most influential factors predicting churn.
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.

EMC² PROVEN PROFESSIONAL

These are the key outputs for each of the main stakeholders of an analytic project.



Key Outputs from a Successful Analytic Project, by Role

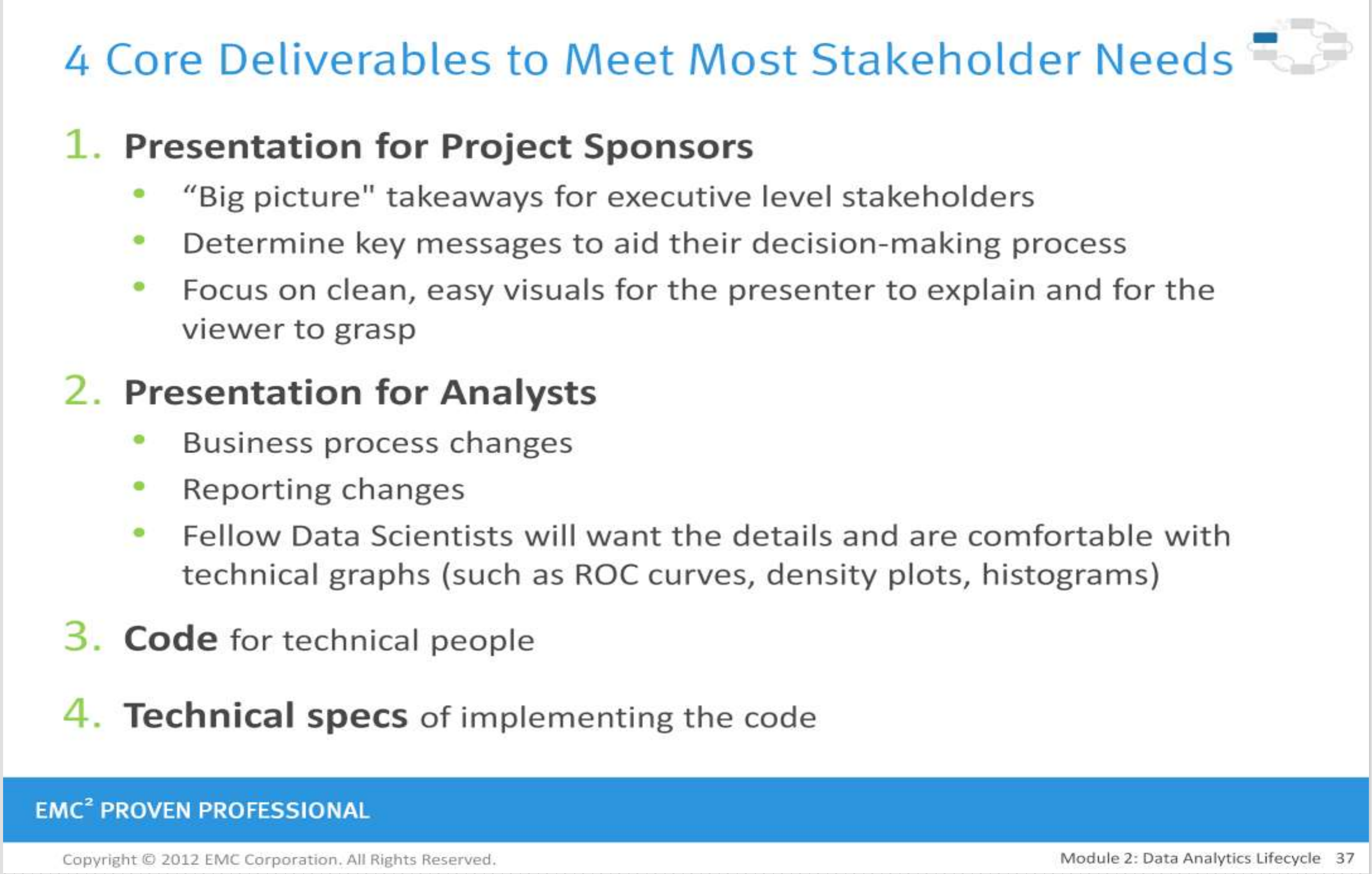
Role	Description	What the Role Needs in the Final Deliverables
Business User	Someone who benefits from the end results and can consult and advise project team on value of end results and how these will be operationalized	• **Sponsor Presentation** addressing: • Are the results good for me? • What are the benefits of the findings? • What are the implications of this for me?
Project Sponsor	Person responsible for the genesis of the project, providing the impetus for the project and core business problem, generally provides the funding and will gauge the degree of value from the final outputs of the working team	• **Sponsor Presentation** addressing: • What's the business impact of doing this? • What are the risks? ROI? • How can this be evangelized within the organization (and beyond)?
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	• Show the **analyst presentation** • Determine if the reports will change
Data Engineer	Deep technical skills to assist with tuning SQL queries for data management, extraction and support data ingest to analytic sandbox	• Share the **code** from the analytical project • Create **technical document** on how to implement it.
Database Administrator (DBA)	Database Administrator who provisions and configures database environment to support the analytical needs of the working team	• Share the **code** from the analytical project • Create **technical document** on how to implement it.
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	• Show the **analyst presentation** • Share the **code**

EMC² PROVEN PROFESSIONAL

- Many times analytical projects yield new insights about a business, a problem, or an idea that people may have taken at face value or thought was impossible to big into.

✓ If appropriate, hold a post-mortem with your analytic team to discuss what about the process or project that you would change if you had to do it over again.

- Module 6 will present further details on how to create the deliverables for each type of audience.
- Here are a few general guidelines about preparing the results of the analysis for sharing with the key sponsors….



4 Core Deliverables to Meet Most Stakeholder Needs

1. **Presentation for Project Sponsors**
   - “Big picture" takeaways for executive level stakeholders
   - Determine key messages to aid their decision-making process
   - Focus on clean, easy visuals for the presenter to explain and for the viewer to grasp
2. **Presentation for Analysts**
   - Business process changes
   - Reporting changes
   - Fellow Data Scientists will want the details and are comfortable with technical graphs (such as ROC curves, density plots, histograms)
3. **Code** for technical people
4. **Technical specs** of implementing the code

1)The more executive the audience, the more succinct you will need to be.

- Most executive sponsors attend many briefings in the course of a day or a week.
- Ensure your presentation gets to the point quickly and frames the results in terms of value to the sponsor’s organization.
- For instance, if you are working with a bank to analyze cases of credit card fraud, highlight the frequency of fraud, the number of cases in the last month or year, and how much cost or revenue impact to the bank (or the focus on the reverse, how much more revenue they could gain if they address the fraud problem).
- This will demonstrate the business impact better

- You will need to include supporting information about analytical methodology and data sources, but generally only as supporting detail or to ensure the audience has confidence in the approach you took to analyze the data.

2)If presenting to other analysts, focus more time on the methodology and findings.

- You can afford to be more expansive in describing the outcomes, methodology and analytical experiment with a peer group, as they will be more interested in the techniques, especially if you developed a new way of processing or analyzing data that can be reused in the future or applied to similar problems.

3)Use imagery when possible.

- **People tend to remember mental pictures to demonstrate a point more than long lists of bullets.**
- Additional references to learn more about best practices for giving presentations:

✓ Say It With Charts, by Gene Zelazny.

- Very simple reference book on how to select the right graphical approach for portraying data, and for ensuring your message is clearly conveyed in presentations.

✓ Pyramid Principle, by Barbara Minto.

- Minto pioneered the approach for constructing logical structures for presentations in threes.
- 3 sections to the presentations, each with 3 main

- This will teach you how to weave a story out of the disparate pieces that emerge from your work.

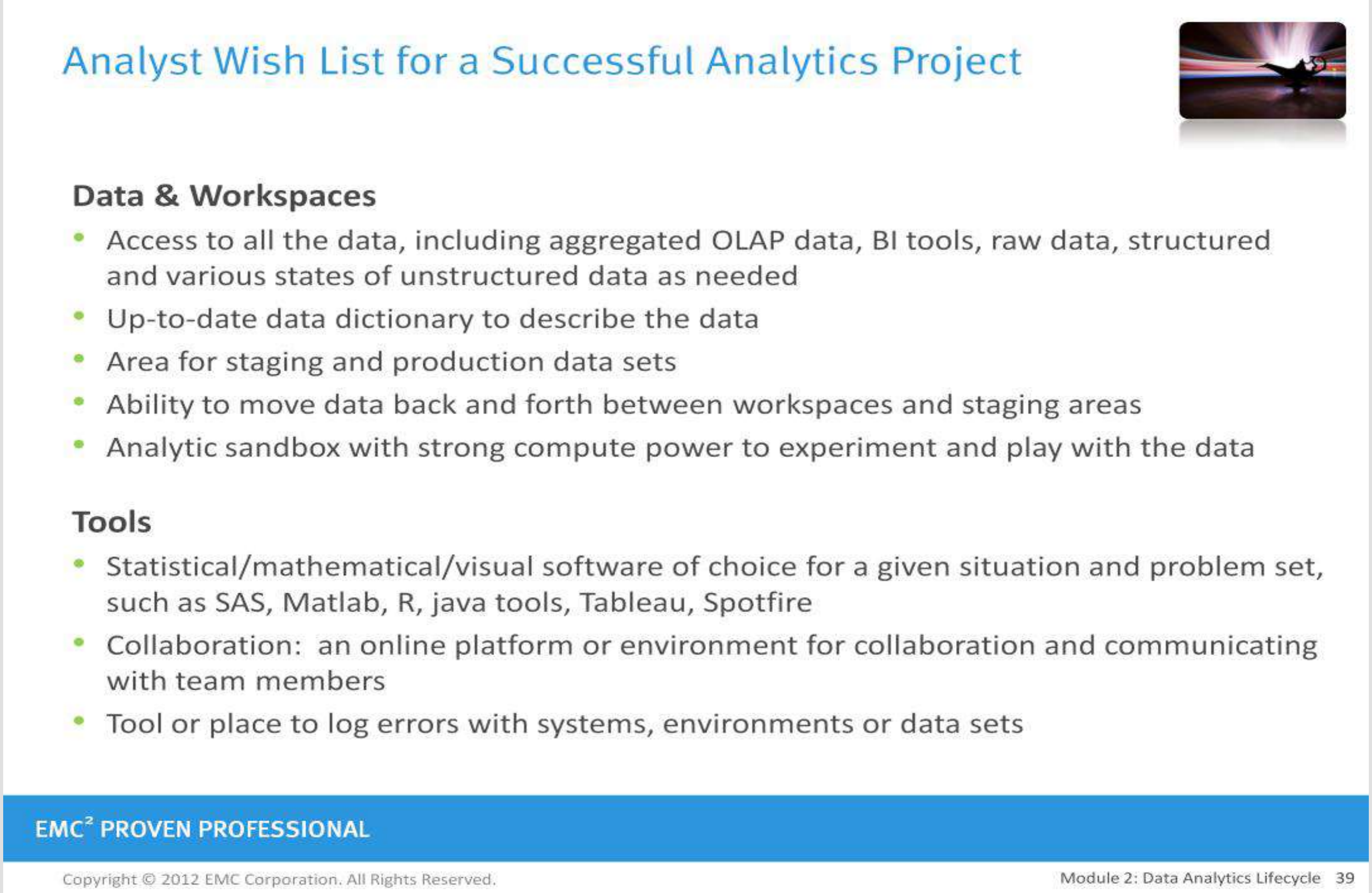
✓ Presentation Zen, by Garr Reynolds.

- Teaches you how to convey ideas simply and clearly, use imagery in presentations, and shows many Before and After versions of graphics and slides.

- The “analyst wish list”, as it pertains to recommendations for tools, data access, and working environments to ensure people are efficient on the project and increase your likelihood of a successful project.

- These needs reflect the need for more flexible environments for storing data and conducting sophisticated and iterative analysis.
- In addition to the technical wishes listed above, easy access to key stakeholders and domain experts would enable to the project to progress smoothly.

- A brief overview of the **MAD approach to analytics.**
- This is one way organizations are responding to the need of analysts and Data Scientists to have more control over the data, and establishing an analytic sandbox in which to perform more sophisticated types of analyses in a flexible manner.



Analyst Wish List for a Successful Analytics Project

Data & Workspaces

- Access to all the data, including aggregated OLAP data, BI tools, raw data, structured and various states of unstructured data as needed
- Up-to-date data dictionary to describe the data
- Area for staging and production data sets
- Ability to move data back and forth between workspaces and staging areas
- Analytic sandbox with strong compute power to experiment and play with the data

Tools

- Statistical/mathematical/visual software of choice for a given situation and problem set, such as SAS, Matlab, R, java tools, Tableau, Spotfire
- Collaboration: an online platform or environment for collaboration and communicating with team members
- Tool or place to log errors with systems, environments or data sets

Here is additional elaboration on the key areas of the "MAD" approach to analytics:

- ❑ **Magnetic:** Traditional EDW approaches "repel" new data sources, discouraging their incorporation until they are carefully cleansed and integrated.
- Given the ubiquity of data in modern organizations, a data warehouse can keep pace today only by being "magnetic": attracting all the data sources that crop up within an organization regardless of data quality niceties.
- Analytic sandbox that attracts all kinds of data with an organization, regardless of quality, structure or size.
- ❑ **Agile:** Data Warehousing orthodoxy is based on long range, careful design and planning.

- Given growing numbers of data sources and increasingly sophisticated and mission-critical data analyses, a modern warehouse must instead allow analysts to easily ingest, digest, produce and adapt data at a rapid pace.
- This requires a database whose physical and logical contents can be in continuous rapid evolution.
- Flexible data structures whose physical and logical contents can support elasticity and rapid, iterative evolution as new analyses demand changes to data structures.

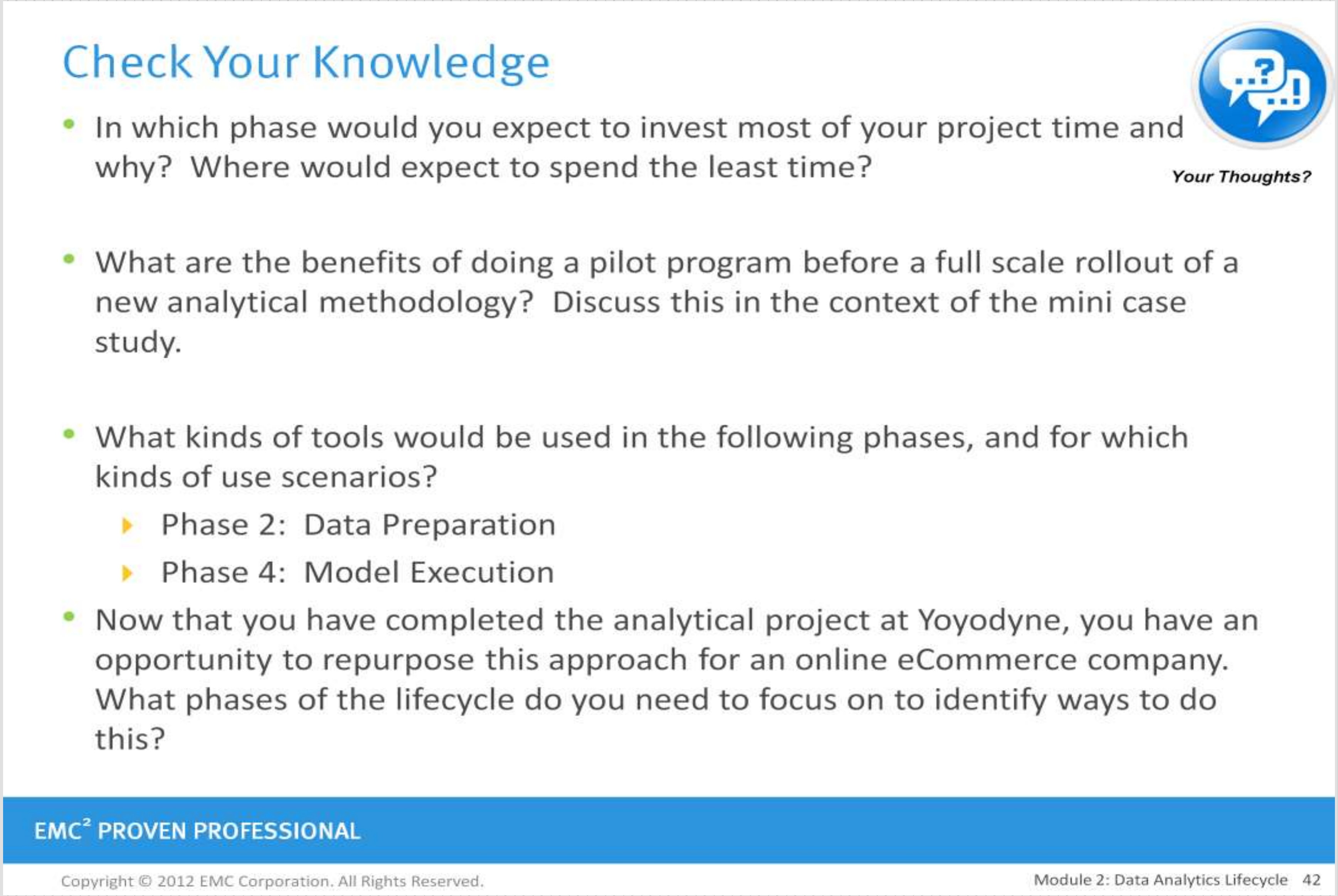
❑ **Deep:** Modern data analyses involve increasingly sophisticated statistical methods that go well beyond the rollups and drilldowns of traditional BI.

- Moreover, analysts often need to see both the forest and the trees in running these algorithms -- they want to study enormous datasets without resorting to samples and extracts.
- The modern data warehouse should serve both as a deep data repository and as a sophisticated algorithmic runtime engine.
- To perform sophisticated analyses on big data, you will need a data repository that can store big data and enable complex analytic algorithms to run with high performance.

- For more information on the MAD approach, see the paper http://db.cs.berkeley.edu/jmh/papers/madskills-

- Sometimes the results of your analysis may change the course of your inquiry….your initial hypothesis needs adjusting and you may need to consult your project sponsor if the results significantly change the focus, or suggest an unexpected outcome.

"The pessimist –
complains about the wind
The optimist –
expects it to change
The leader –
adjusts the sails
John Maxwell
(Leadership Author)



Check Your Knowledge

Your Thoughts?

- 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?

EMC² PROVEN PROFESSIONAL