جامعة الإسكندرية
ALEXANDRIA UNIVERSITY
كلية الحاسبات وعلوم البيانات

# LEC.3.

# Introduction to Data Science

# Data Scientists

- Data Scientist
  - The most attractive Job of the 21st Century
- They find stories, extract knowledge. They are not reporters
- A data scientist is the key person in acquiring, clearing, representing and analyzing data for business and research purposes

# Data Scientists

- Data scientists are the key to realizing the opportunities presented by big data. They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and decisions

# Essential skills for a data scientist:

- Basic tools
- Basic statistical knowledge
- Machine learning and discovery
- Calculus and linear algebra
- Data mining
- Data visualization and communication
- Software engineering

# Essential skills for a data scientist:

- Mathematics and Applied Mathematics
- Applied Statistics/Data Analysis
- Solid Programming Skills (R, Python, SQL)
- Data Base Storage and Management

# What do Data Scientists do?

- National Security
- Cyber Security
- Business Analytics
- Engineering
- Healthcare
- And more ….

# What is the problem ?!

- The problem is that with this un sorted very large data size , we cant analysis it, more over we cant classify it , it become un-useful if we stored data without any usage .

# How to solve this ?!

The Data Analytics Lifecycle is designed specifically for Big Data problems and data science projects.

The lifecycle has six phases, and project work can occur in several phases at once.

# Data Analytics Lifecycle

# Key stakeholders of an analytics project

- Each plays a critical part in a successful analytics project. Although seven roles are listed, fewer or more people can accomplish the work depending on the scope of the project, the organizational structure, and the skills of the participants. The seven roles follow:
- **Business User:** Someone who understands the domain area and usually benefits from the results. This person can consult and advise the project team on the context of the project, the value of the results, and how the outputs will be operationalized. Usually a business analyst, line manager, or deep subject matter expert in the project domain fulfills this role.

- **Project Sponsor:** Responsible for the genesis of the project. Provides the impetus and requirements for the project and defines the core business problem. Generally provides the funding and gauges the degree of value from the final outputs of the working team. This person sets the priorities for the project and clarifies the desired outputs.
- **Project Manager:** Ensures that key milestones and objectives are met on time and at the expected quality.

- **Business Intelligence Analyst:** Provides business domain expertise based on a deep understanding of the data, key performance indicators (KPIs), key metrics, and business intelligence from a reporting perspective.
- **Database Administrator (DBA):** Provisions and configures the database environment to support the analytics needs of the working team. These responsibilities may include providing access to key databases or tables and ensuring the appropriate security levels are in place related to the data repositories.

- **Data Engineer**: Leverages deep technical skills to assist with tuning SQL queries for data management and data extraction, and provides support for data ingestion.

*The DBA sets up and configures the databases to be used, the data engineer executes the actual data extractions and performs substantial data manipulation to facilitate the analytics. The data engineer works closely with the data scientist to help shape data in the right ways for analyses.*

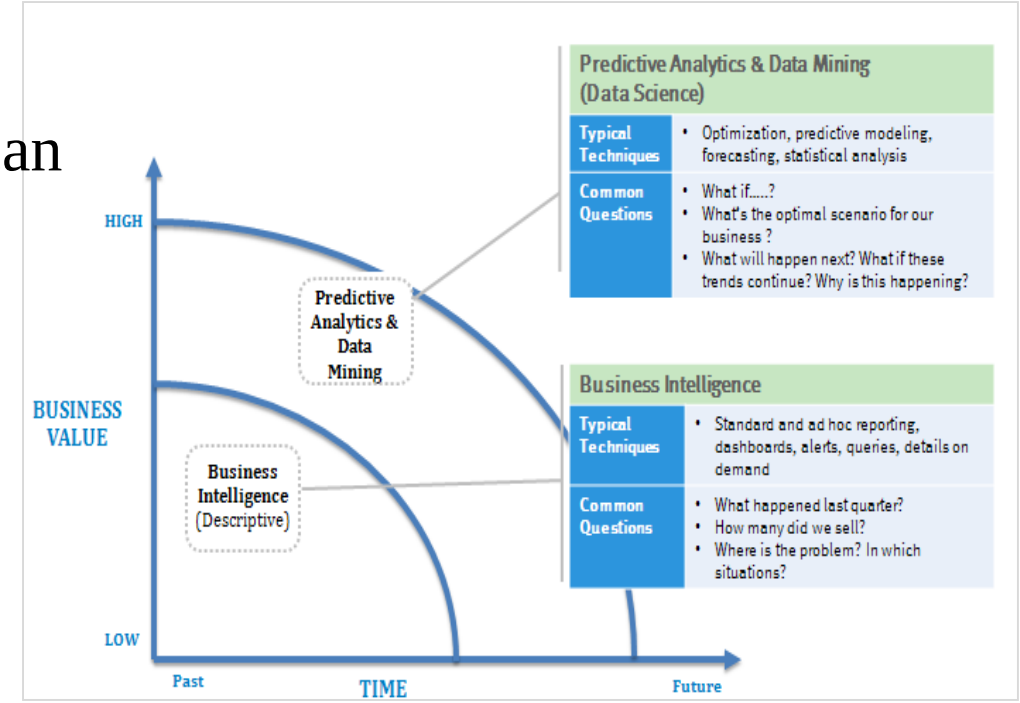
- **Data Scientist:** Provide subject matter expertise for analytical techniques, data modeling, applying valid analytical techniques to given business problems and ensuring overall analytical objectives are met

# Value of Using the Data Analytics Lifecycle

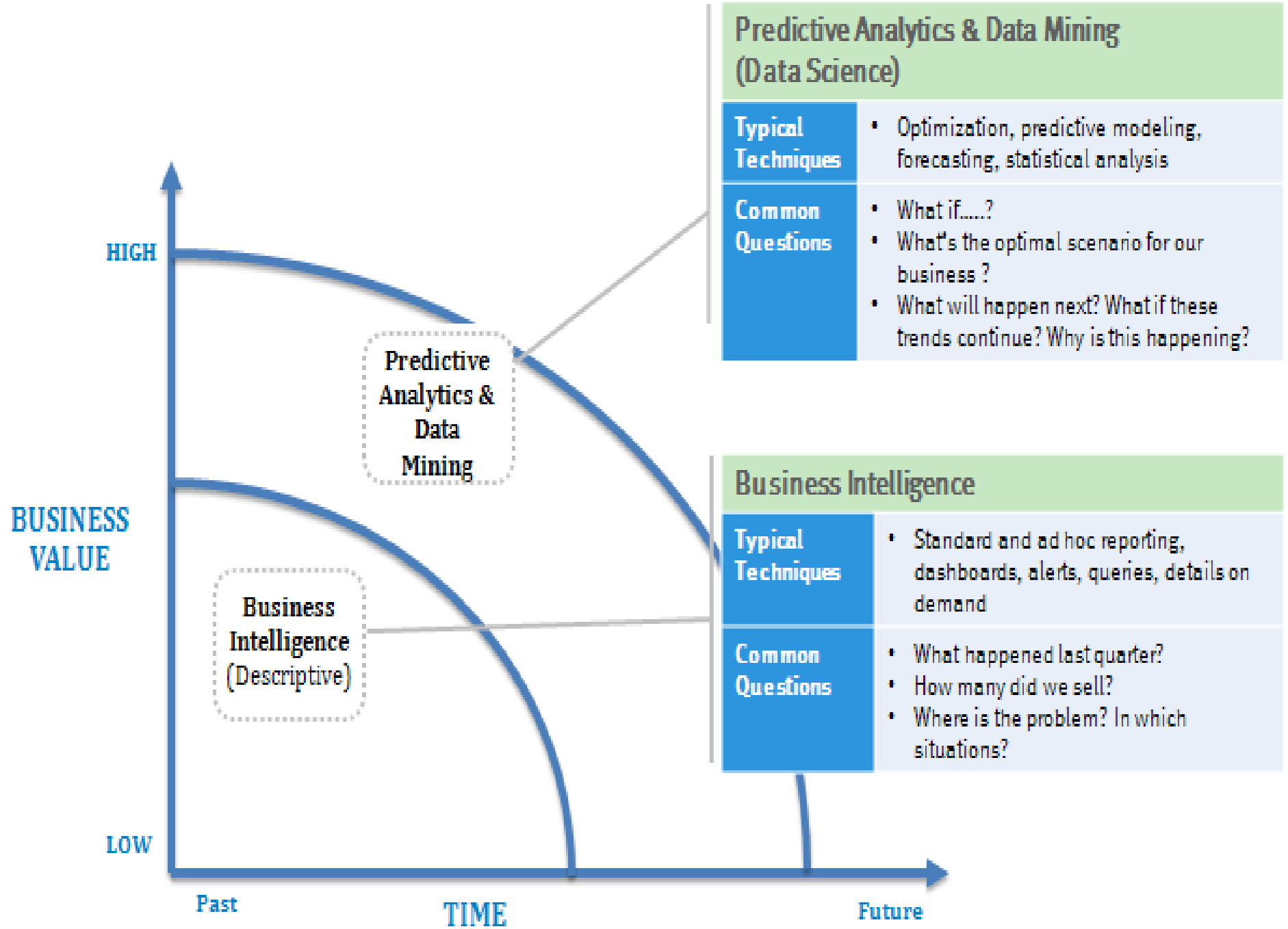
- Focus your time
- Ensure rigor and completeness
- Enable better transition to members of the cross-functional analytics teams
  - Repeatable
  - Scale to additional analysts
  - Support validity of findings

# Need For a Process to Guide Data Science Projects

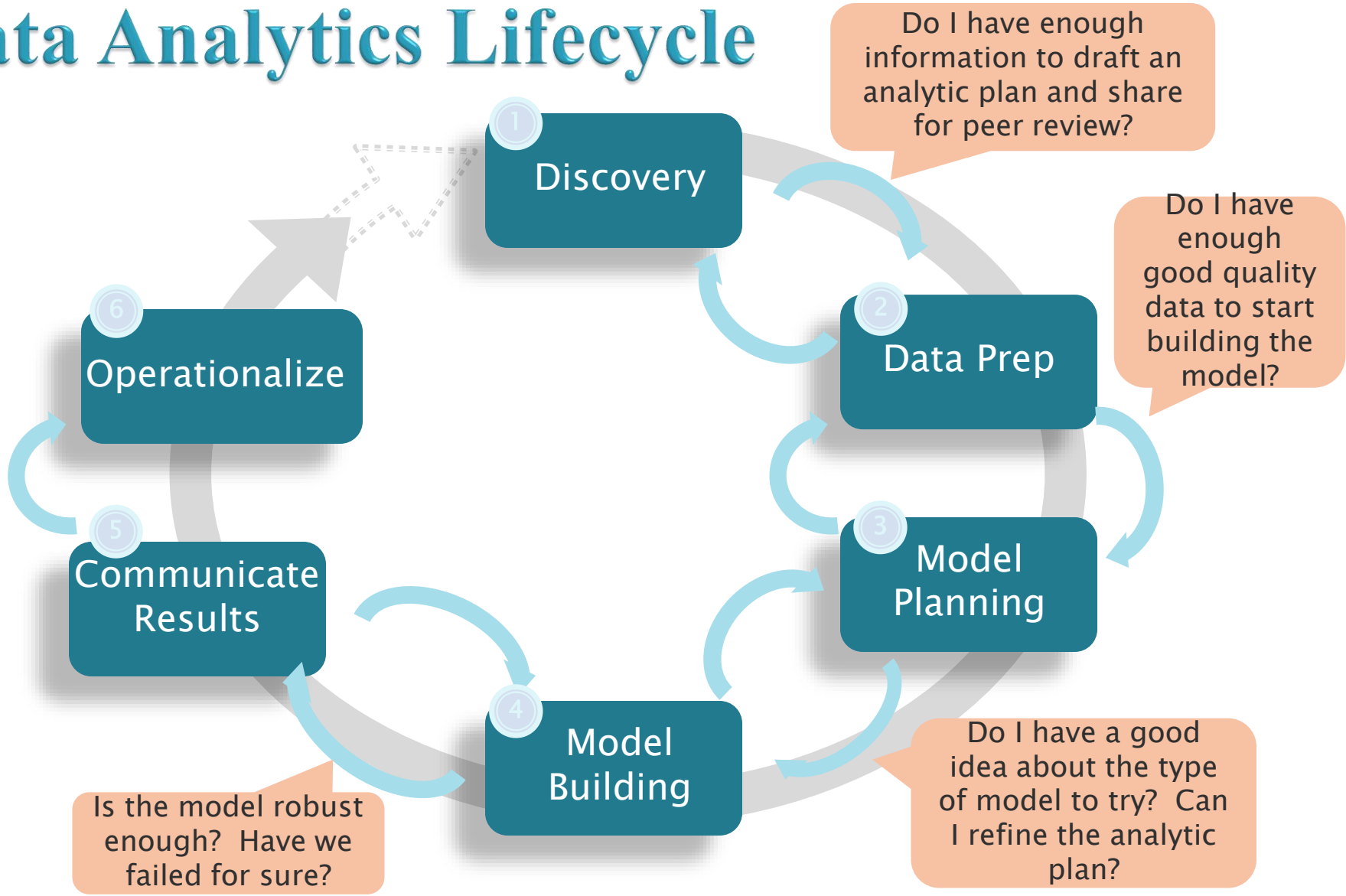
1. Well-defined processes can help guide any analytic project
2. Focus of Data Analytics Lifecycle is on Data Science projects, not business intelligence



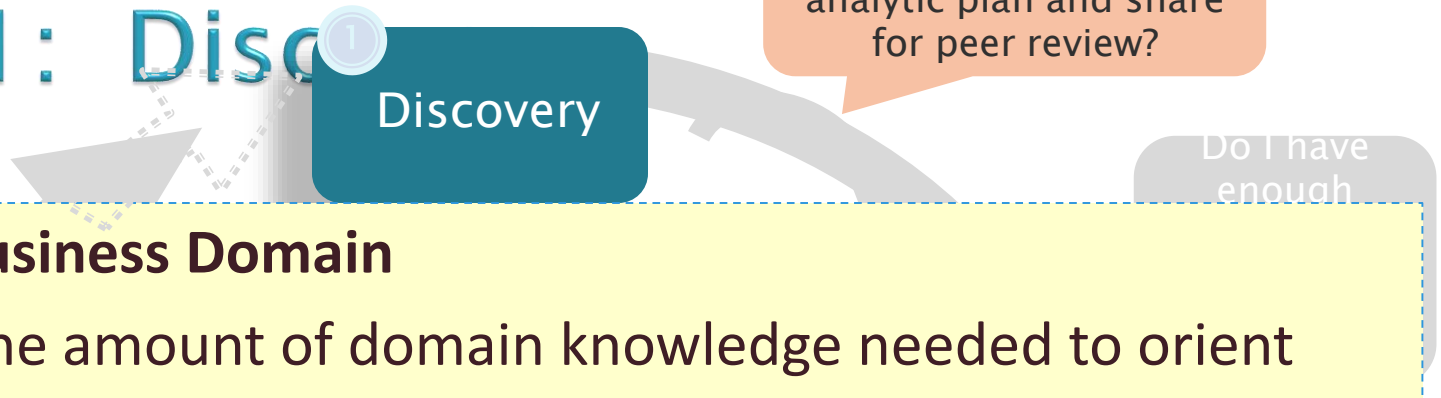

3. Data Science projects tend to require a more consultative approach, and differ in a few ways
   - More due diligence in Discovery phase
   - More projects which lack shape or structure
   - Less predictable data


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?

# Data Analytics Lifecycle



# Data Analytics Lifecycle
# Phase 1: Disc

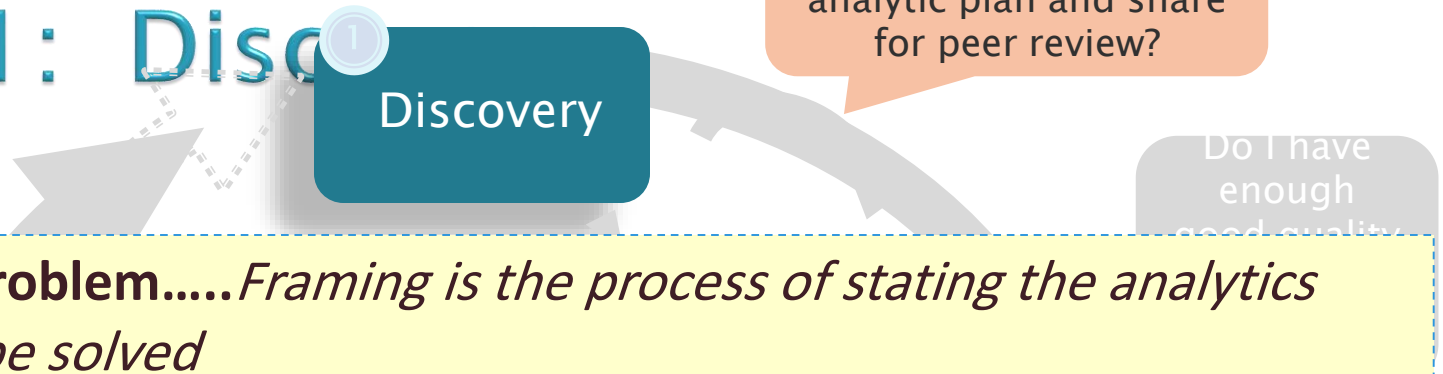



- **Learn the Business Domain**
  - Determine amount of domain knowledge needed to orient you to the data and interpret results downstream
  - Determine the general analytic problem type (such as clustering, classification)
  - If you don't know, then conduct initial research to learn about the domain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

# Data Analytics Lifecycle
# Phase 1: Disc


- **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: Disc




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

# Tips for Interviewing the Analytics Sponsor

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

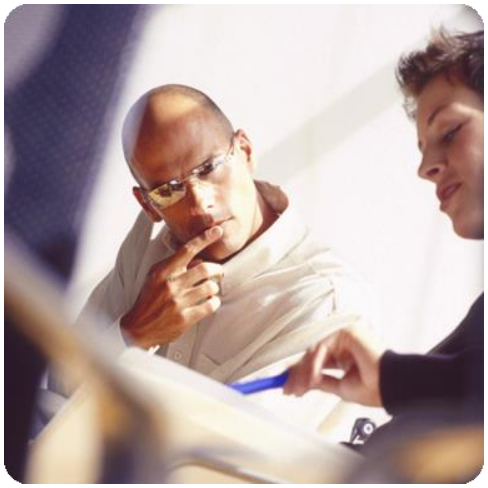
# *Sponsor Interview Tips*

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

# Tips for Interviewing the Analytics Sponsor

- **Interview Questions**
  - What is the business problem you're trying to solve?
  - What is your desired outcome?
  - Will the focus and scope of the problem change if the following dimensions change:
- Time – analyzing 1 year or 10 years worth of data?
- People – how would this project change this?
- Risk – conservative to aggressive
- Resources – none to unlimited (tools, tech, …..)
- Size and attributes of Data

# Interview Questions:



- What data sources do you have?
- What industry issues may impact the analysis?
- What timelines are you up against?
- Who could provide insight into the project? Consulted?
- Who has final say on the project?

# Data Analytics Lifecycle
# Phase 1: Disc

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

good
he type
ry? Can
analytic
?

# 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