Chapter 2

Data Analytics Lifecycle

Dr. E. Hamouda
CSCI 398: Introduction to Data Science
Spring 2017

Outline

- Roles of a Data Analytics Project Team
- Data Analytics Lifecycle
- Case Study to apply the data analytics lifecycle

Your Thoughts?

- How to Approach an Analytics Problem -- a Project in General

  - How do you plan for an analytic project?
  - Do you follow a methodology or some kind of framework?

Need For a Process to Guide Data Science Projects

- Well-defined processes can help guide any analytic project
- Break large projects into smaller pieces
- Spend time to plan and scope the work
- Documenting adds rigor and credibility

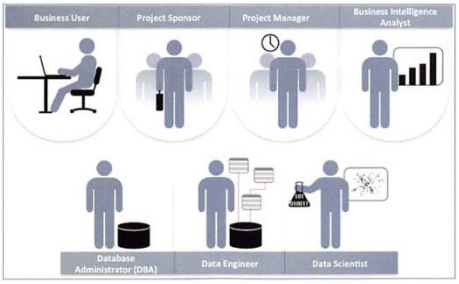
➔ Data Analytics Lifecycle

- Focus of Data Analytics Lifecycle is on Data Science projects

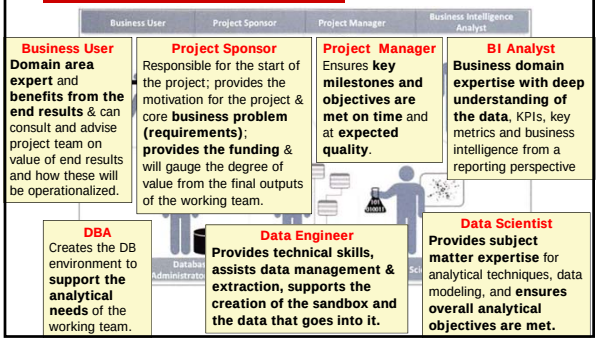
Value of Using the Data Analytics Lifecycle

- 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

Key Roles for a Successful Analytics Project - Team



Key Roles for a Successful Analytics Project - Team



Business User
Domain area expert and **benefits from the end results** & can consult and advise project team on value of end results and how these will be operationalized.

Project Sponsor
Responsible for the start of the project; provides the motivation for the project & core **business problem (requirements)**; **provides the funding** & will gauge the degree of value from the final outputs of the working team.

Project Manager
Ensures **key milestones and objectives are met on time** and at **expected quality**.

BI Analyst
Business domain expertise with deep understanding of the data, KPIs, key metrics and business intelligence from a reporting perspective

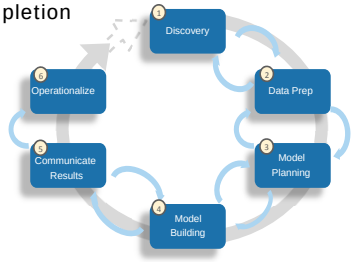
DBA
Creates the DB environment to **support the analytical needs** of the working team.

Data Engineer
Provides technical skills, assists data management & extraction, supports the creation of the sandbox and the data that goes into it.

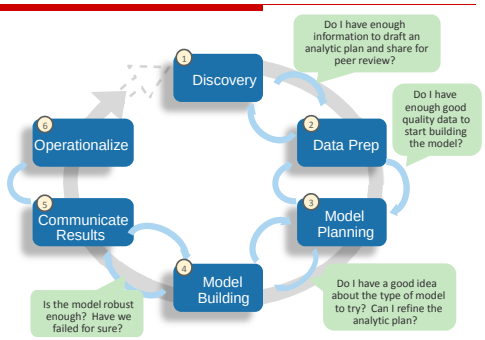
Data Scientist
Provides subject matter expertise for analytical techniques, data modeling, and **ensures overall analytical objectives are met.**

Data Analytics Lifecycle

- Data Analytics Lifecycle defines the analytics process and best practices from discovery to project completion

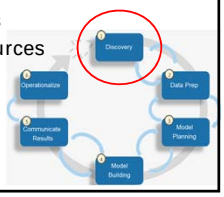


Data Analytics Lifecycle

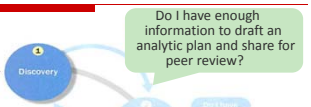


Phase 1: Discovery

1. Learning the Business Domain
2. Resources
3. Framing the Problem
4. Identifying Key Stakeholders
5. Interviewing the Analytics Sponsor
6. Developing Initial Hypotheses
7. Identifying Potential Data Sources

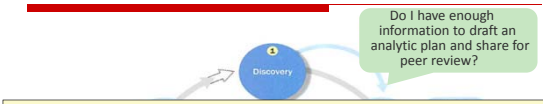


Phase 1: Discovery



- **Learn the Business Domain**
  - Understand the domain area
  - Determine how much business or domain knowledge is needed to orient the data scientist to develop models in Phase 3&4 and interpret results downstream.
- **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 the Model again? How have things changed?

Phase 1: Discovery



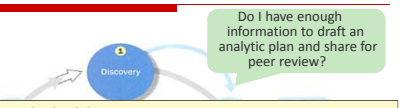
- **Resources**
  - Assess available
    - **technology, tools, system**
    - **data** – sufficient to meet your needs
    - **people** for the working team
    - **time** for the project in calendar time and person-hours
  - Do you have sufficient resources to attempt the project? If not, can you get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

Phase 1: Discovery

Frame the problem:

- State the problem, why it is important, and to whom?
  - Clearly articulate the current situation and pain points
  - Share the problem with the stakeholders.
- Identify what needs to be achieved in business terms and what needs to be done to meet the needs
- Identify the success/failure criteria.
  - 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 & settle for what we have)?

Phase 1: Discovery



Identify key stakeholders

- Identify key stakeholder and their interest in the project
  - Each stakeholder may expect different results from the project
  - Each stakeholder may have different criteria to judge the project
- Even if you are "given" the analytic problem you should work with clients to clarify and frame the problem
- **Interview stakeholders**
  - What is the business problem you're trying to solve?
  - What is your desired outcome? etc.

Phase 1: Discovery

- **Formulate Initial Hypotheses/ideas to be tested**
  - H1 , H2, H3, ... Hn
  - Gather and assess hypotheses from stakeholders and domain experts
  - Preliminary data exploration to discuss with stakeholders during the hypothesis forming stage
- **Identify Data Sources – Begin Learning the Data**
  - Identify the candidate data sources to test the IH
  - Capture aggregate sources for describing the data and providing high-level understanding.
  - Review the raw data.
  - Evaluate the data structures and tools needed.
  - Scope the kind of data needed for this kind of problem.

Case Study to Track the Phases in the Data Analytics Lifecycle

Mini Case Study: Churn Prediction for Yoyodyne Bank

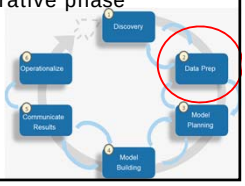
- **Situation Synopsis**
  - Retail Bank, Yoyodyne Bank wants to improve the Net Present Value (NPV) and retention rate of customers
  - They want to establish an effective marketing campaign targeting customers to reduce the churn rate by at least 5%.
  - The bank wants to determine if those customers are worth retaining. In addition, the bank 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

Case Study to Track the Phases in the Data Analytics Lifecycle

Components of Analytic Plan		
Phase 1: Discovery	Learning the Business Domain	equity, fixed income, regulatory, investment banking, risk management (retail/wholesale) etc.
	Resources	Analytic sandbox, OLAP, EDW.
	Framing the Problem	How do we identify customer churn/no? How can we improve the Net Present Value and retention rate of customers? Risk: the project will fail if we can't identify valid predictors of churn.
	Identifying people and Key Stakeholders	Working team and business users from bank
	Developing Initial Hypotheses	Transaction volume and type are key predictors of churn rates.
	Identifying Data Sources	EDW, 5 months of customers history

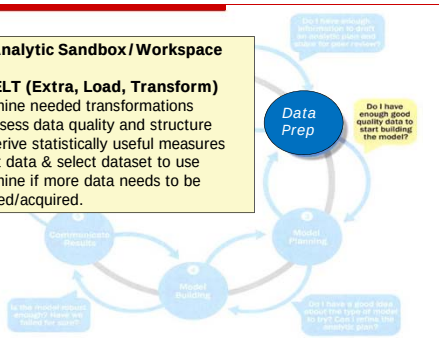
Phase 2: Data Preparation

- Includes steps: to explore, preprocess, and condition data.
- Tends to be the most labor-intensive step in the analytics lifecycle
  - Often at least 50% of the data science project's time
- It is generally the most iterative phase



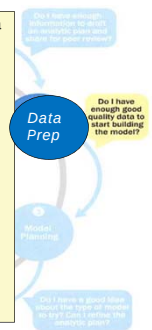
Phase 2: Data Preparation

- **Prepare Analytic Sandbox / Workspace**
- **Perform ELT (Extra, Load, Transform)**
  - Determine needed transformations
    - Assess data quality and structure
    - Derive statistically useful measures
  - Extract data & select dataset to use
  - Determine if more data needs to be collected/acquired.



Phase 2: Data Preparation

- **Learning about the data:** Becoming familiar with the data is **critical**
  - List your data sources
  - List what's needed vs. what's available
  - Highlights gaps – identifies data not currently available
  - Identifies data outside the organization that might be useful
- **Data Conditioning**
  - Clean and normalize data
  - Discern what you keep vs. what you discard
- **Survey & Visualize**
  - Overview, zoom & filter then maintain data of interest
  - Descriptive Statistics
  - Use data visualization tools to gain an overview of the data
    - Does the data contains unexpected values or indicator of dirty data?



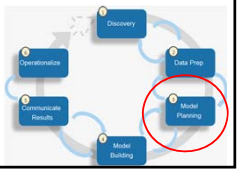
Phase 2: Data Preparation

- Learning about the Data: Sample Dataset Inventory

Dataset	Data Available and Accessible	Data Available, but not Accessible	Data to Collect	Data to Obtain from Third Party Sources
Products shipped	●			
Product Financials		●		
Product Call Center Data		●		
Live Product Feedback Surveys			●	
Product Sentiment from Social Media				●

Phase 3: Model Planning

- Identify candidate models to apply to the data set (clustering, classifying, finding relationships in the data, etc.)
  1. Data Exploration
  2. Variable Selection
  3. Model Selection



Phase 3: Model Planning

- **Data Exploration**
  - Assess the data to understand the relationship between variables
  - Assess the structure of the data – this dictates the tools and analytic techniques for the next phase
- **Variable Selection**
  - Explore the data to select the variables and methods
  - A common way to do this is to use data visualization tools
  - Inputs from stakeholders and domain experts
  - Iterative testing to confirm the most significant variables
- **Model Selection**
  - choose an analytical technique, or several candidates, based on the end goal of the project

Phase 3: Model Planning

The Problem to Solve	The Category of Techniques
I want to group items by similarity. I want to find structure (commonalities) in the data	Clustering
I want to discover relationships between actions or items	Association Rules
I want to determine the relationship between the outcome and the input variables	Regression
I want to assign (known) labels to objects	Classification
I want to find the structure in a temporal process I want to forecast the behavior of a temporal process	Time Series Analysis
I want to analyze my text data	Text Analysis

Mini Case

Phase 3: Model Planning

- After conducting research on churn prediction, you have identified many methods for analyzing customer churn.
- At this point, a Data Scientist would assess the methods and select the best model for the situation
- Example of other analysts approaching a similar problem

Market Sector	Analytic Techniques/Methods Used
Consumer Packaged Goods	Multiple linear regression, automatic relevance determination (ARD), and decision tree
Retail Banking	Multiple regression
Retail Business	Logistic regression, ARD, decision tree
Wireless Telecom	Neural network, decision tree, hierarchical neurofuzzy systems, rule evolver, logistic regression

Case Study to Track the Phases in the Data Analytics Lifecycle

Components of Analytic Plan	
Phase 3: Model Planning – Analytics Techniques	Logistic regression to identify most influential factors predicting churn