Project Lifecycle

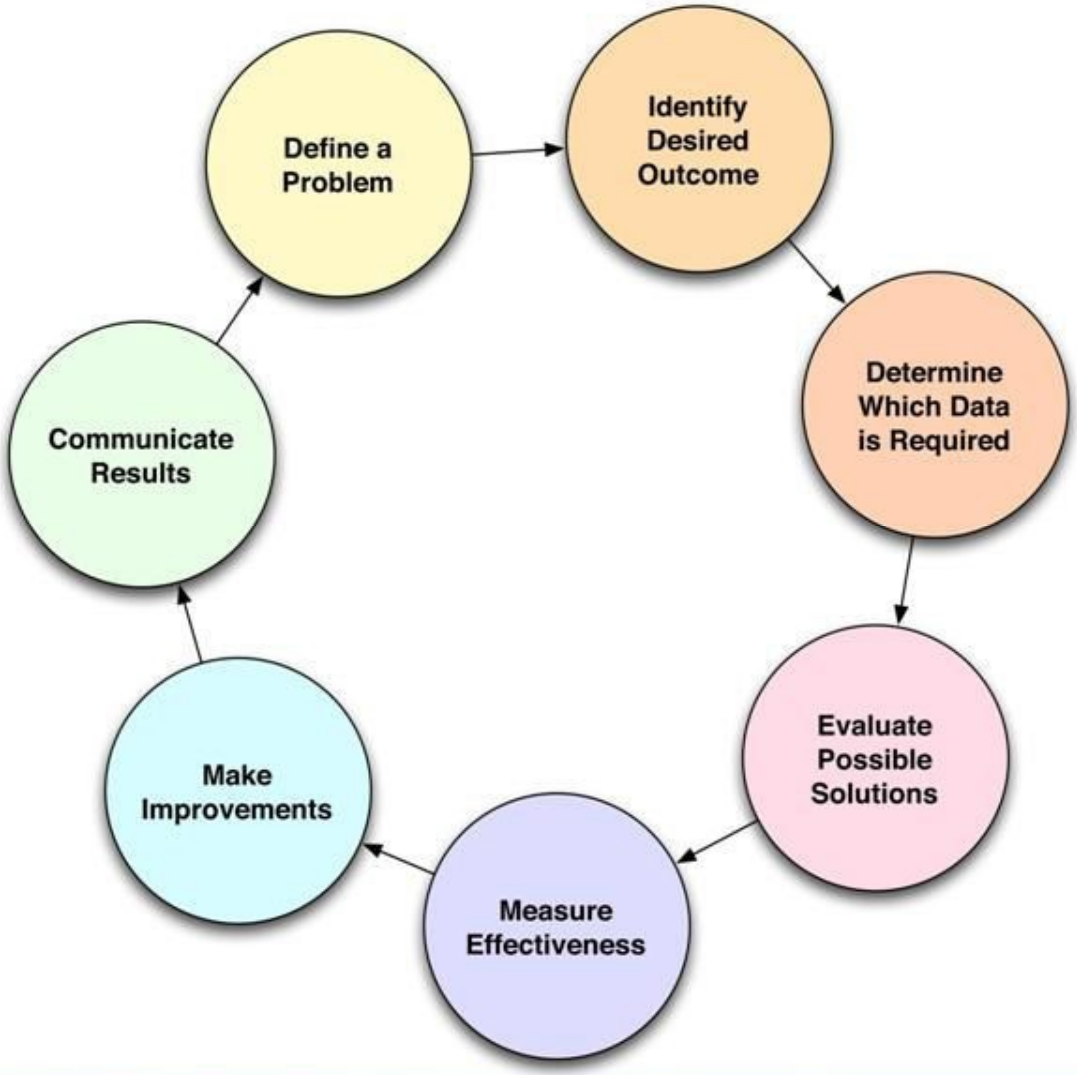
Project Lifecycle

- **How a data scientist approaches a problem**
- **Which steps comprise the lifecycle of a data science problem**

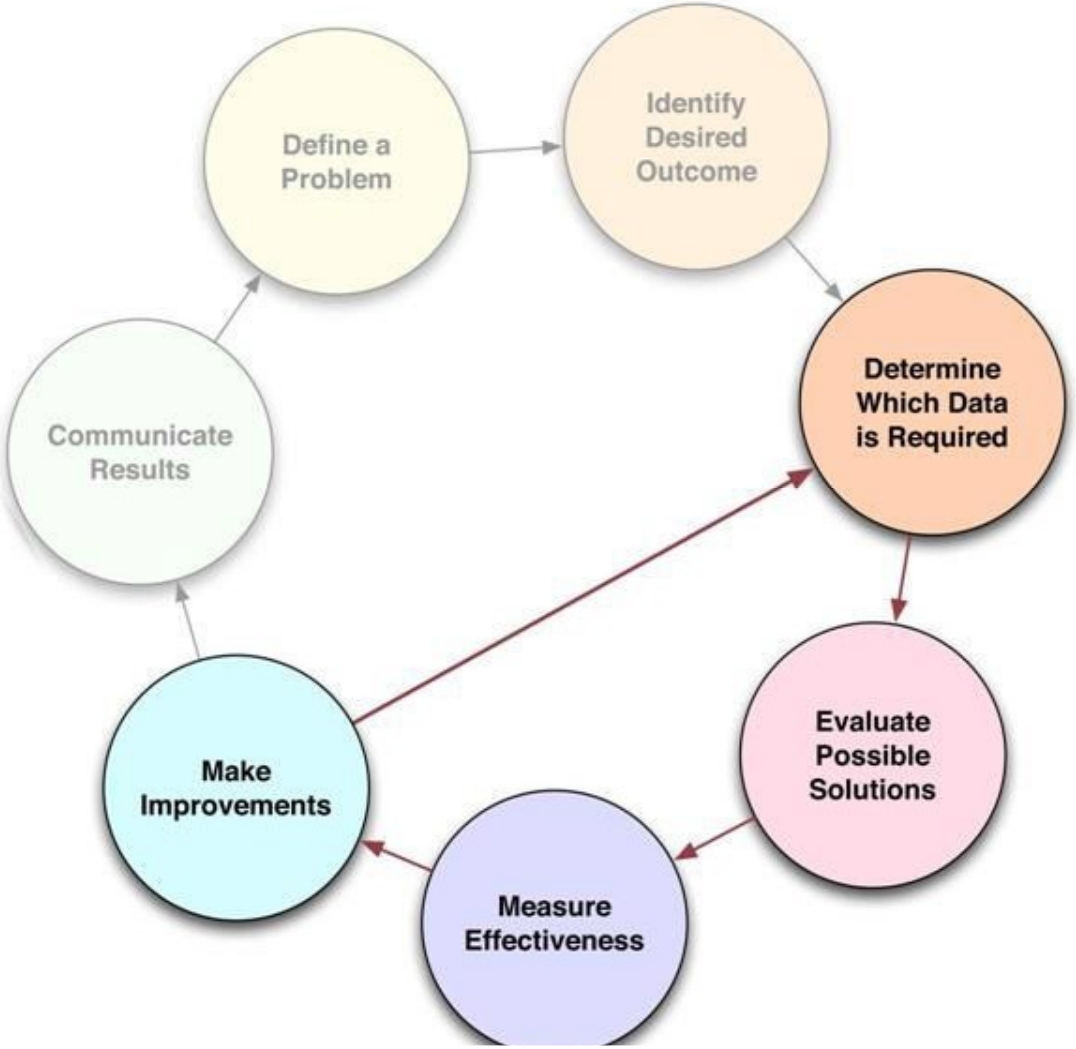
Overview of the Project Lifecycle

- **A typical data science project should follow these steps**
  - Define a problem
  - Identify the desired outcome
  - Determine which data is needed
  - Evaluate possible solutions
  - Measure effectiveness
  - Make improvements
  - Communicate results

The Entire Project Lifecycle is Iterative



Parts of the Project Lifecycle are Also Iterative



Defining a Problem

- **The process begins by first clearly stating a problem**
- **This is often directly related to revenue**
  - "People browse our site but don't buy anything"
  - "Too many customers abandon their shopping carts"
  - "Subscribers aren't renewing their service"
  - "Sponsors are donating less than ever before"
- **In other cases, the problem is related to costs**
  - "Our employees spend too much time searching for documents"
  - "An increase in support calls cost us $400,000 last year"

Defining a Problem (cont'd)

- **Sometimes the approach is more exploratory**
  - "What can I learn about our users from this clickstream data?"
  - "Why are customers abandoning purchases before checkout?"
  - "How many more finish checkout when offered free shipping?"
  - "Would offering free shipping on all orders increase profits?"

Identifying the Desired Outcome

- **Given the problem, what's the preferred resolution?**
- **Again, these are often tied to revenue or costs**
  - "Increase subscription renewals by 5% within two months"
  - "Decrease shopping cart abandonment by 10% in Q3"
  - "Reduce support call volume by 25% within one year"
- **Be careful what you wish for**
  - **Problem:** "An increase in support calls cost $400,000 last year"
  - **Goal:** "Reduce support call volume by 25% within one year"
  - A reduction in support calls may not mean fewer problems
    - Could indicate customer frustration due to poor support

Determining Which Data is Needed

- **What data must you capture to solve the problem you've defined?**
  - And to determine if your solution meets the goals identified
- **Further refinements may require additional data**
- **Consider the source, format, and quality of this data**
  - Does your organization have everything you need?
  - If not, is it available from external sources?

Evaluate Possible Solutions

- **Consider all solutions that could match desired outcome**
- **This typically involves a hypothesis about the root cause**
  - What prompted the recent increase in support calls?
  - Why are customers abandoning their carts?
  - What causes customers to not renew subscriptions?
- **Given several possible solutions, which should you invest in?**
  - Most can be discounted quickly
  - Small-scale tests can help you choose
- **The simplest solution is usually the best one to pursue first**

Measuring Effectiveness

- **Measuring effectiveness requires two things**
  - Metrics: properties to measure
  - Method: a process for comparing these metrics

Making Improvements

- **Measurement will illustrate how much improvement is required**
- **Consider what you might change**
  - Was your hypothesis about the root cause correct?
  - Could adding an additional data set give more insight?
  - Should you try one of the solutions you originally discarded?
- **Once you've implemented your improvements, test it again**
  - Comparing measurements can help to refine your solution

Communicating Results

- **Communication is an essential part of data science**
- **A data scientist must tell the story found within the data**
  - Like any good story, it must be compelling
  - Be concise and focus on what's important for the audience
- **Dashboards are a common tool for communicating results**
  - Statistics
  - Summaries
  - Visualizations

Scenario Explanation

- **Why are revenues down?**
  - Existing customers aren't renewing their subscriptions
  - Prospective customers aren't joining our service
- **Customers were surveyed when they called to cancel**
  - Reason cited by 79% of customers
    - "You have lots of movies, just none that interest *me*."
- **Movies has hired you to help solve this problem**

Scenario Explanation (cont'd)

- **Problem Definition**
  - "Many customers are choosing not to renew their subscriptions"
- **Desired Outcome**
  - "Decrease cancellation rate by 35% during next quarter"
- **Possible solutions**
  - Decrease subscription cost (discard: price is not the problem)
  - Social media integration (discard: may violate privacy laws)
  - Improve movie recommendations (we'll pursue this one)

Overview

- **There are many prerequisites to building a recommender system**
  - Acquiring the input data from various sources
  - Identifying and correcting errors in the input data
  - Transforming the data into the desired format for analysis
- **The work isn't finished even after you've built the recommender**
  - You need to test it
  - You'll likely find ways to improve it
    - These may require additional data sources
    - The project lifecycle begins again

Essential Points

- **The lifecycle of a data science project is iterative**
- **It's important to clearly state a problem**
- **The problem helps to establish the desired outcome**
- **Success is determined by measuring results**