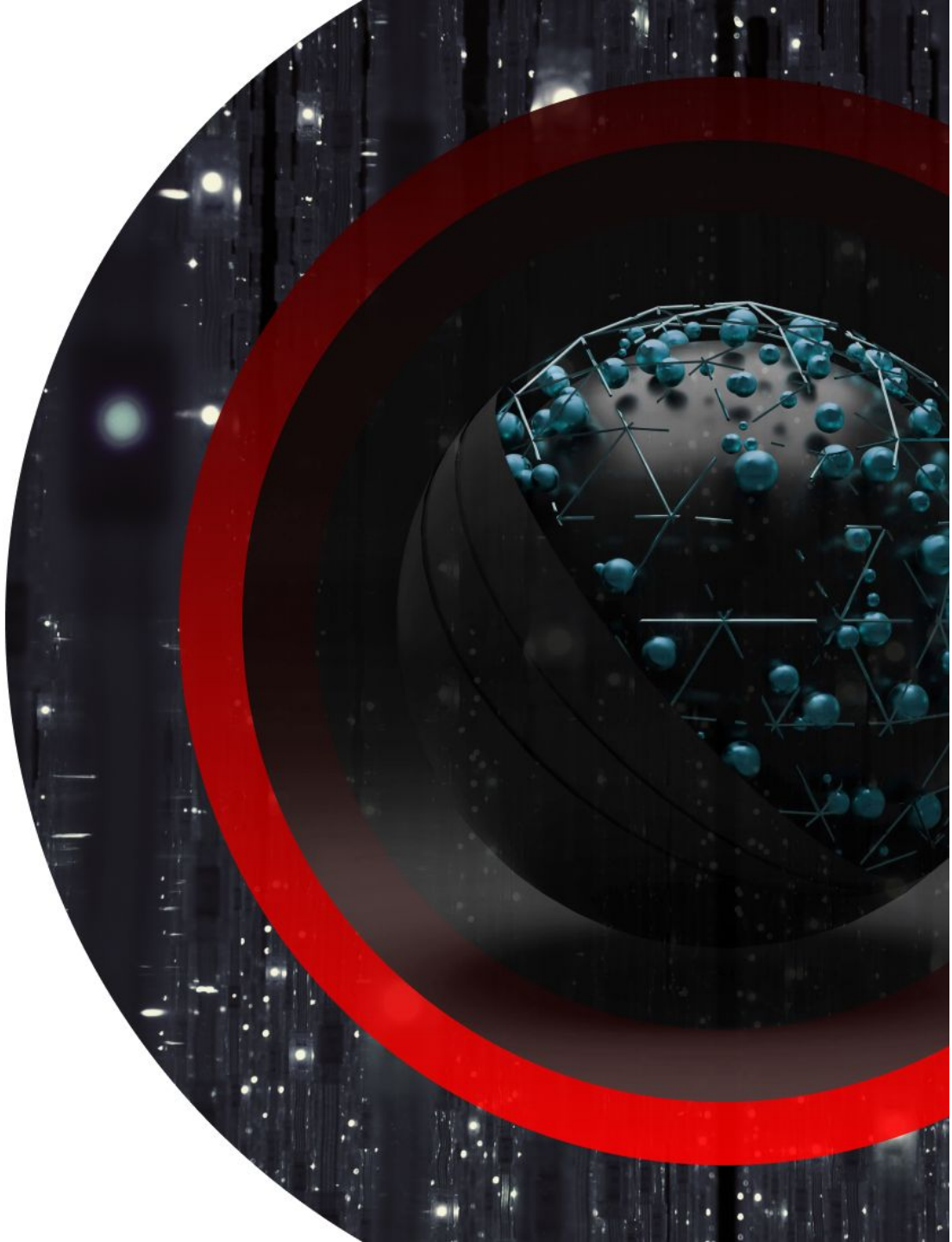
DATA SCIENCE AND ANALYTICS

Introductory Course



CLASS NORMS

COURTESY IN CLASS

Remaining on mute unless called on, exercising courtesy during breakout rooms, and using the chat box for questions only

ATTENDANCE

100% attendance is expected and contributes to success in passing the course and the program.

PARTICIPATION

Keeping an open mind in discussions and sharing experiences, making contributions during team assignments, submitting assignments in Canvas, and participating in discussion boards

USE OF CLASS RESOURCES

Follow along during the lecture with the lesson companion and download any in-class documents prior to class.

INSTRUCTIONAL TEAM

INSTRUCTIONAL TEAM

Name	Contact information
Name	Contact information
Name	Contact information

PROGRAM PATH

1 Introductory Course

2 SQL and Databases

3 Statistics and Probability

4 Data Storytelling

Milestone 1: Building and Presenting Data Stories

5 Python Programming

6 Data Wrangling

7 Visual Communications

8 Advanced SQL Programming

Milestone 2: Data Integration, Preparation, Reporting, and Presentation

9 Business Intelligence

10 Big Data

11 Machine Learning

12 Applied AI

Milestone 3: Capstone Project – Delivering Insights and Presentations

INTRODUCTORY COURSE PATH

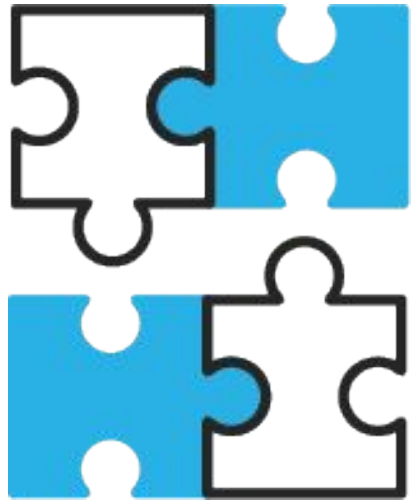
				1 INTRODUCTION TO DATA SCIENCE AND ANALYTICS
2 COMPUTING PRIMER	**3** PROGRAMMING CONCEPTS	**4** **DISCOVERING AND CURATING DATA**	**5** STRUCTURING AND ANALYZING DATA	
6 CLEANING AND ENRICHING DATA	**7** VALIDATING AND PRESENTING DATA	**8** INTRODUCTION TO DATA SCIENCE PROJECTS	**9** ASSESSMENT NIGHT	

MODULE OUTLINE

Lesson 1: The Context of Analysis

Lesson 2: Collecting Data Manually

Lesson 3: Collecting Data Automatically



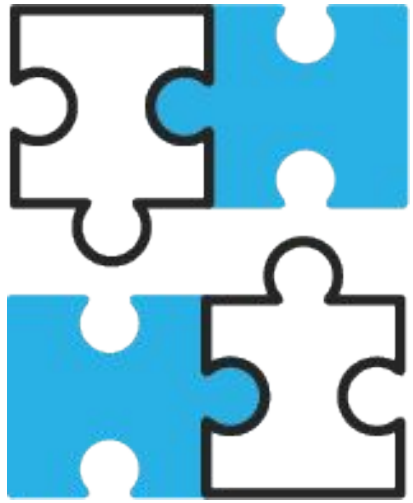
MODULE GOALS

- Create a strategy for determining the problem and goal for analysis.
- Discuss the procedure of manual and automatic data collection process.
- List the phases of the data life cycle.
- Explain how an API endpoint can be used to collect data.

LESSON 1: THE CONTEXT OF ANALYSIS

LESSON OUTLINE

- Process models such as CRISP-DM
- Defining context
- Problem statements
- Strategy



DISCUSSION TIME

How would you define a *project*?

Think about a recent project you completed.

What made it a success?

LESSON GOALS

WHAT ARE THE GOALS?

- Discuss the first phase in the CRISP-DM process: business understanding.
- Create a strategy for determining the problem and goal for an analysis project.

WHY ARE THEY IMPORTANT?

Every successful project begins with a clear understanding of the customer's needs. In a data science and analytics project, it is important to understand the context of what your customer needs and whether it is internal or external.

YOUR TAKE

What do you wish to accomplish by completing this module?

PROCESS MODEL

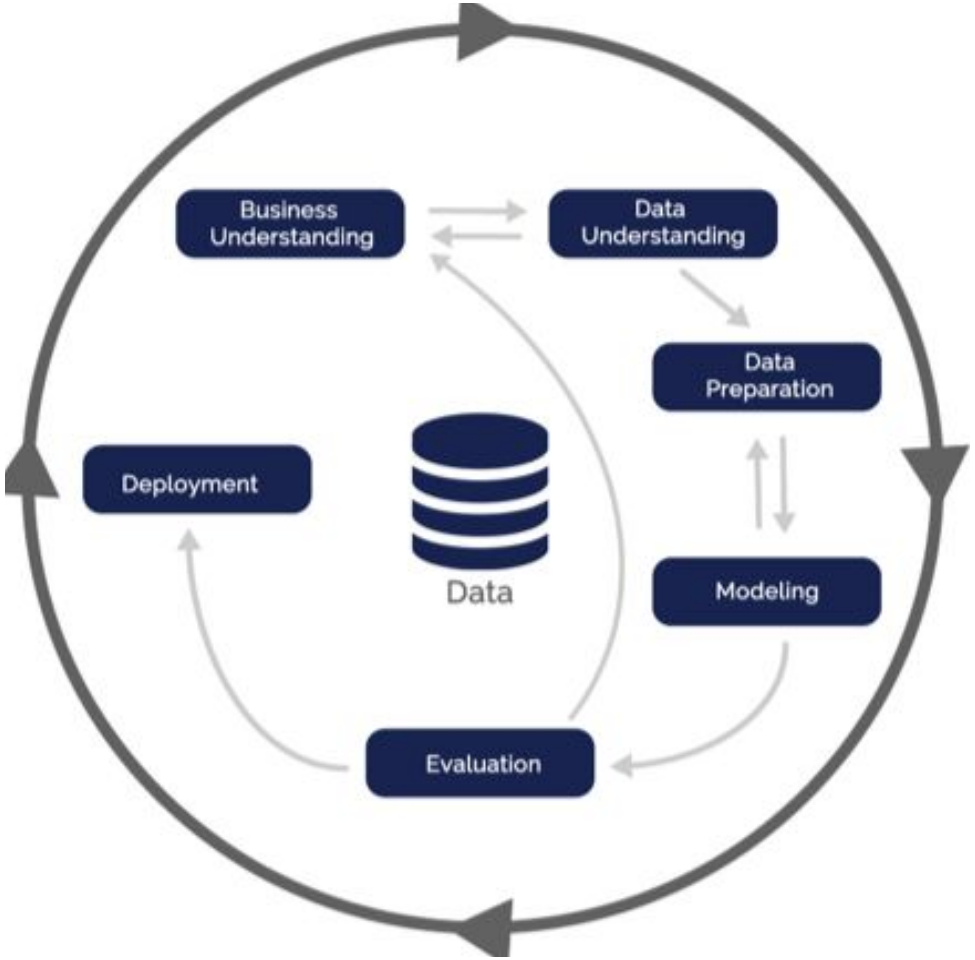
A process model is a set of guidelines to help you organize and plan your projects.

Process models have phases that help break down tasks or activities that contribute to the completion of a project.

CRISP-DM

- **CR**oss
- **I**ndustry
- **S**tandard
- **P**rocess for
- **D**ata
- **M**ining

From: KDnuggets



BUSINESS UNDERSTANDING

At the onset of a project, we need to know:

- What are the objectives and goals of the project?
- What is the problem we are trying to solve?
- What resources are needed?

DISCUSSION TIME

What does an analysis look like?

Review the article "Complete Study of Factors Contributing to Air Pollution."

- Introduction
- Problem statement
- Objective/scope

PROBLEM STATEMENTS

> A good data science problem should be relevant, specific, and unambiguous. It should align with the business strategy.
> —Vinita Silaparasetty

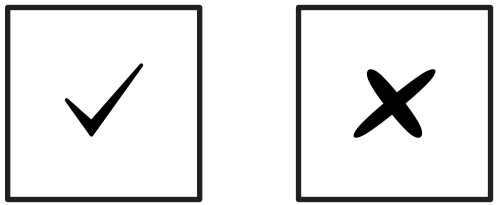
From: Medium

QUALITIES OF AN EFFECTIVE PROBLEM STATEMENT

An effective problem statement has the following qualities:

- Identifies the gap that exists
- Has a time frame, location, and trend
- Quantifies the impact (cost, time, quality, etc.)
- Emphasizes the importance to the organization, individual, or society

QUIZ!



What is the first step of the process model?

A. Business understanding
B. Data understanding
C. Data preparation
D. Modeling
E. Evaluation
F. Deployment

QUIZ FEEDBACK

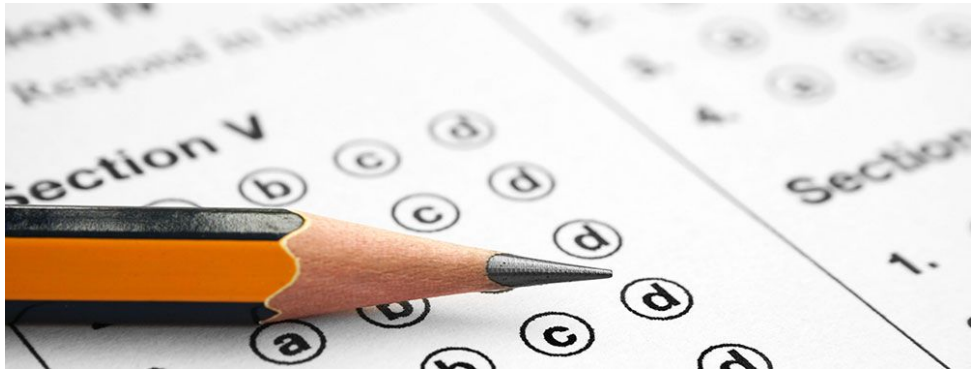
What is the first step of the process model?

A. Business understanding
B. Data understanding
C. Data preparation
D. Modeling
E. Evaluation
F. Deployment

PROBLEM STATEMENTS IN ACTION

75% of the 8th grade students in the Palm Beach School District are having difficulties passing their state assessment exams.

- How do we change state tests to increase passing scores?
- What strategy can we employ to have students pass the state test?



STRATEGY

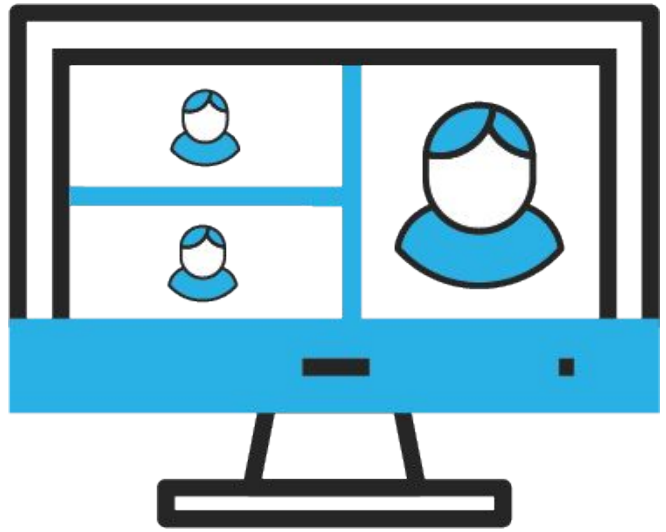
To gain an understanding of the problem you are trying to solve:

- Ask customers clarifying questions.
- Document every stage of the process.

BREAKOUT ROOMS

Help Kenneth determine why his business is suddenly failing.

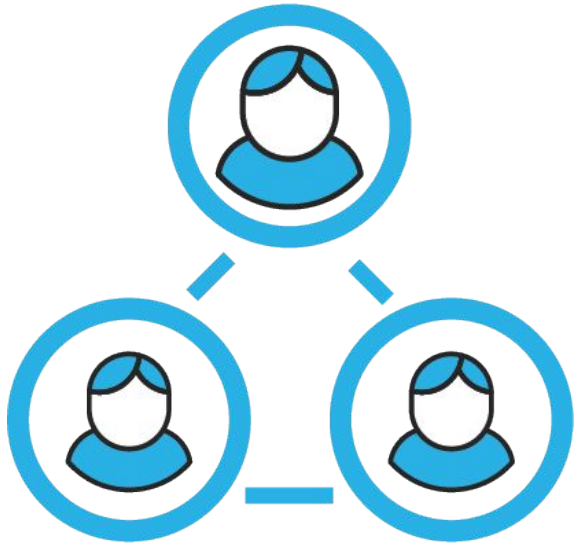
Open the 1.4.1 Activity Google Doc to get started.



BREAKOUT ROOMS.

INTERVIEW QUESTION

What are the elements of an effective problem statement?



INTERVIEW FEEDBACK

What are the elements of an effective problem statement?

An effective problem statement identifies gaps; has a time frame, location, and trend; quantifies the impact of the problem; and emphasizes the importance to the organization, individual, or society.

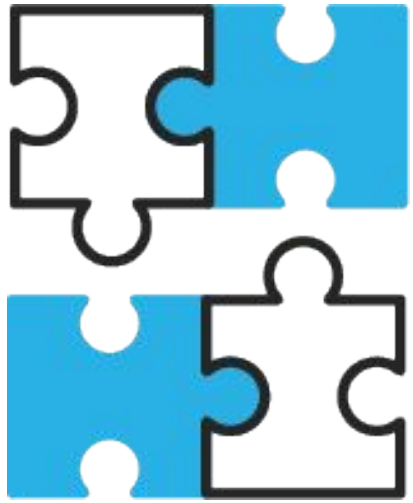
QUESTIONS?

BREAK TIME.

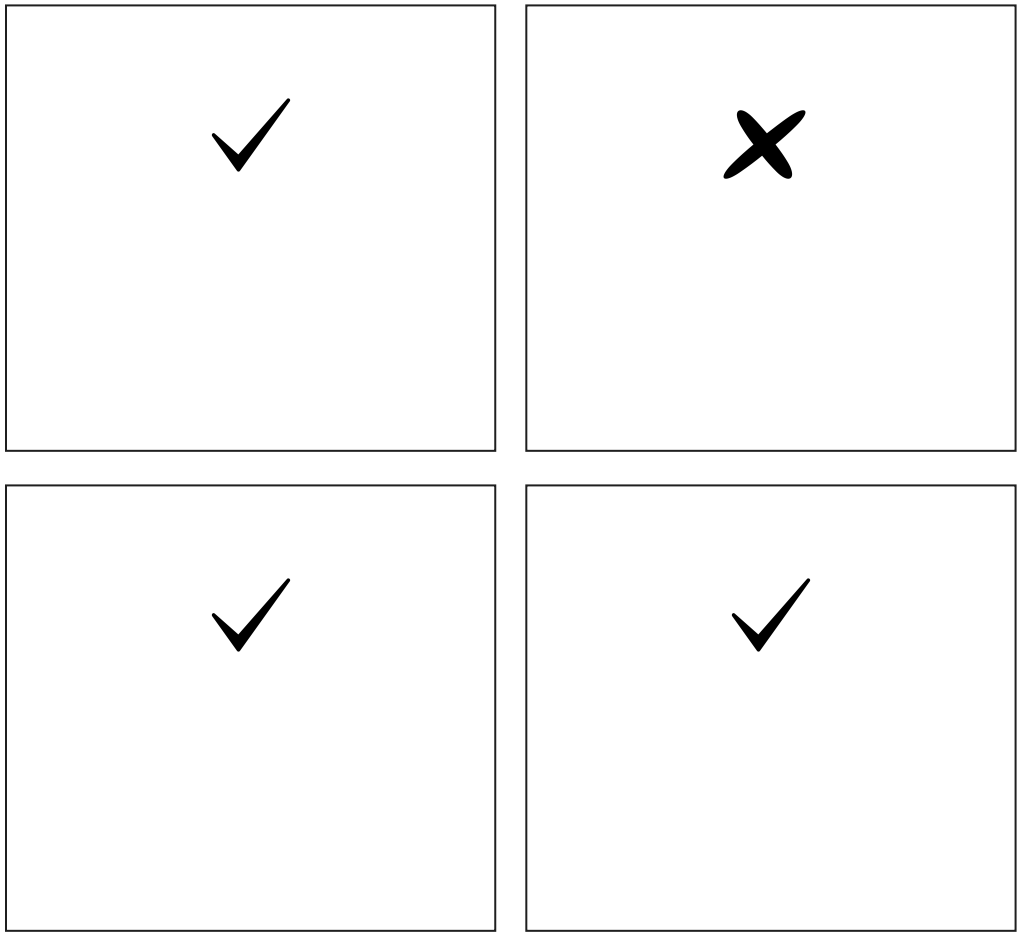
LESSON 2: COLLECTING DATA MANUALLY

LESSON OUTLINE

- Data collection sources
- Data life cycle
- FTP sites
- The housing market



POLL



POLL:

What is the next phase in the CRISP-DM process model after *business understanding*?

A. Data understanding and preparation
B. Modeling
C. Evaluation
D. Deployment

POLL FEEDBACK.

What is the next phase in the CRISP-DM process model after *business understanding*?

A. Data understanding and preparation

B. Modeling

C. Evaluation

D. Deployment

LESSON GOALS

WHAT ARE THE GOALS?

- Describe the discovery phase of the data life cycle.
- Discuss the various ways you can collect data.
- Review popular online sources for collecting data.
- Collect data manually.

WHY ARE THEY IMPORTANT?

Data professionals routinely collect data from various sources, most often programmatically. A combination of manual manipulation and automation will be needed to collect and curate your data for an analysis.

REVIEW

In the last lesson, we discussed the need to establish the context and problem statement for an analysis prior to analyzing data.

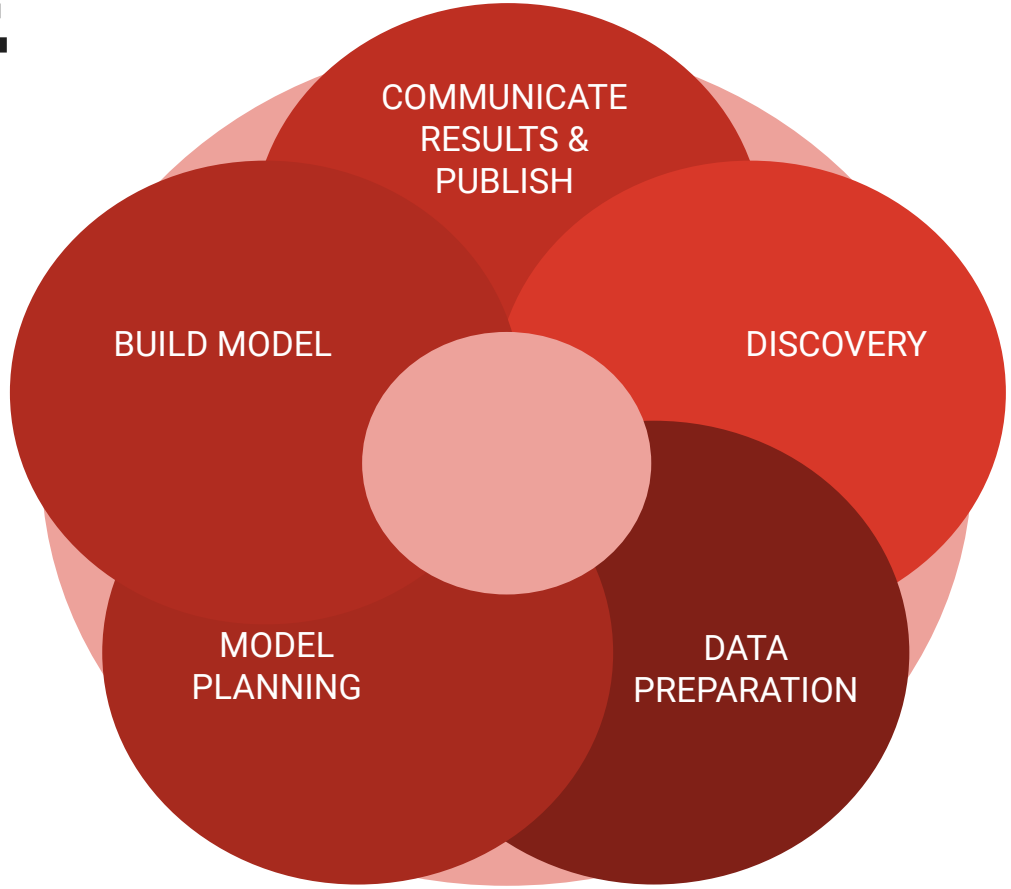
The problem statement and context will guide your data collection. These elements help the team focus its efforts on discovering and collecting the right data for an analysis.

DATA LIFE CYCLE

The data life cycle encompasses the process of discovering data until implementation.

Data can be stored, used, shared, and archived.

Within the data understanding and preparation phase (CRISP-DM), there is a data discovery phase.



KEEP IN MIND

Data can be in one of the following states:

- At rest
- In process
- In transit

Data is messy: meaning that it might contain blanks, null values, and/or incorrect values.

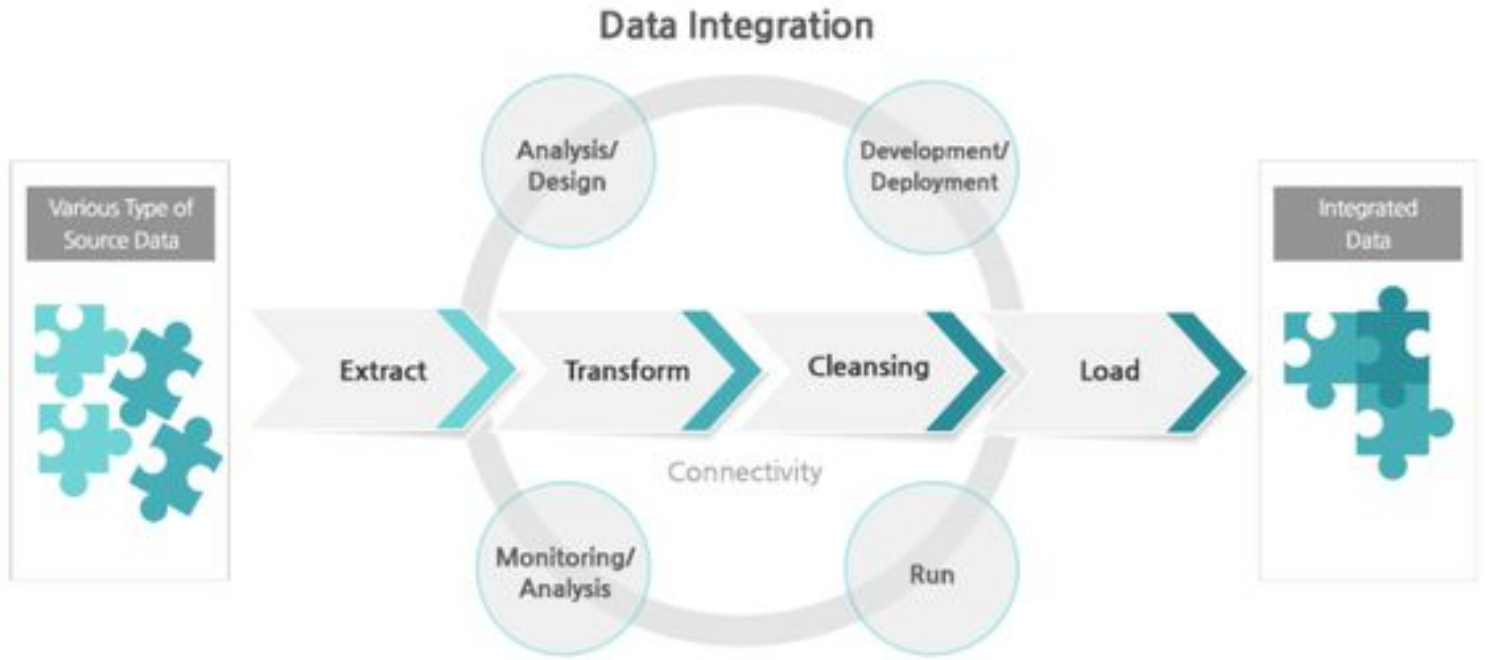
We must verify that our data is clean, or clean it up ourselves!

DISCOVERY PHASE

During the discovery phase we must:

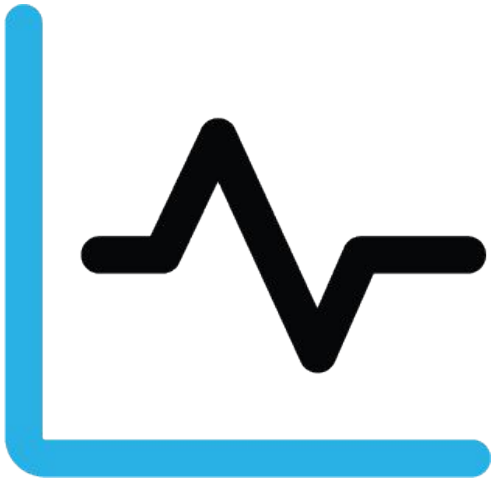
- Ask guiding questions to prepare for analysis.
- Think about where we will collect the data we need.
- Determine how we will collect the data we need.

WHERE'S THE DATA?



From: OmniSci

WEBSITES AS A DATA SOURCE



There are many open-source websites that contain free datasets for a variety of uses:

- Industry data for market research and analysis
- Data for personal or passion projects
- Forum communities for help solving analysis or coding issues

DISCUSSION TIME

Spend the next few minutes exploring the following websites:

- Kaggle
- KDnuggets
- UCI Machine Learning
- Data.gov
- Dataset Search

Characterize each website by the type of data available.

DISCUSSION FEEDBACK

- **Kaggle:** Variety of datasets; competition hosting; portfolio for personal projects
- **KDnuggets:** Resources for data science and analytics (blogs); courses; datasets for machine learning, data science, visualization, etc.
- **UCI:** Datasets for machine learning
- **Data.gov:** Public information for research, app development, and data visualizations
- **Dataset Search:** Google's version of data sources; similar to Google Scholar

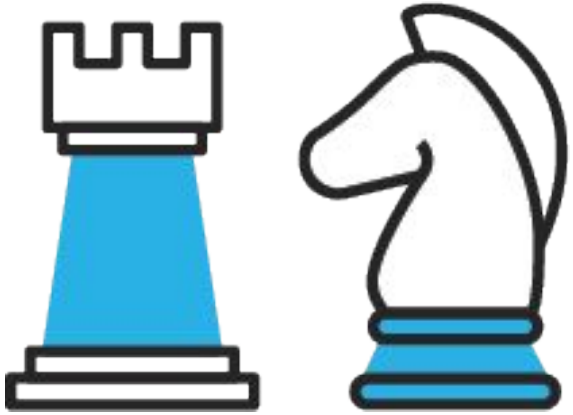
FTP SITES AS A DATA SOURCE

File Transfer Protocol (FTP) is the protocol used to transfer files over a network (often the internet, but internally as well).

FTP sites allow businesses to securely store files in a structured directory, much like on your computer.

From: WindowsReport

ACTIVITY: COLLECTING DATA

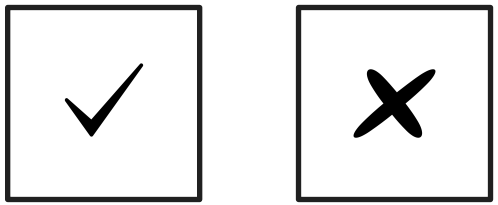


In this activity, you will use Data.gov to manually download a dataset and perform a basic inspection.

You will need the 1.4.2 Activity Google Doc and the Excel program to get started.

BREAKOUT ROOMS.

QUIZ!



1. What is an easy way to examine a CSV file after you download it?
2. In which phase of the data life cycle do we conduct planning and determine the context for our analysis?

QUIZ FEEDBACK.

1. What is an easy way to examine a CSV file after you download it?

   Open the file with Excel, Google Sheets, or Notepad.

2. In which phase of the data life cycle do we conduct planning and determine the context for our analysis?

   The discovery phase

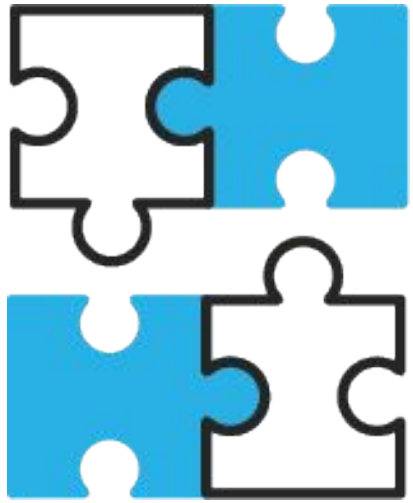
QUESTIONS?

BREAK TIME.

LESSON 3: COLLECTING DATA AUTOMATICALLY.

LESSON OUTLINE

- Two v's of big data: velocity and volume
- Batch vs. real-time processing
- API endpoints



DISCUSSION TIME

What do you think *automation for data collection* means?

Make an inference about how data professionals can achieve automatic data collection.

LESSON GOALS

WHAT ARE THE GOALS?

- Describe volume and velocity in relation to big data.
- Differentiate between batch and real-time processing.
- Identify an API endpoint you can use to collect data.

WHY ARE THEY IMPORTANT?

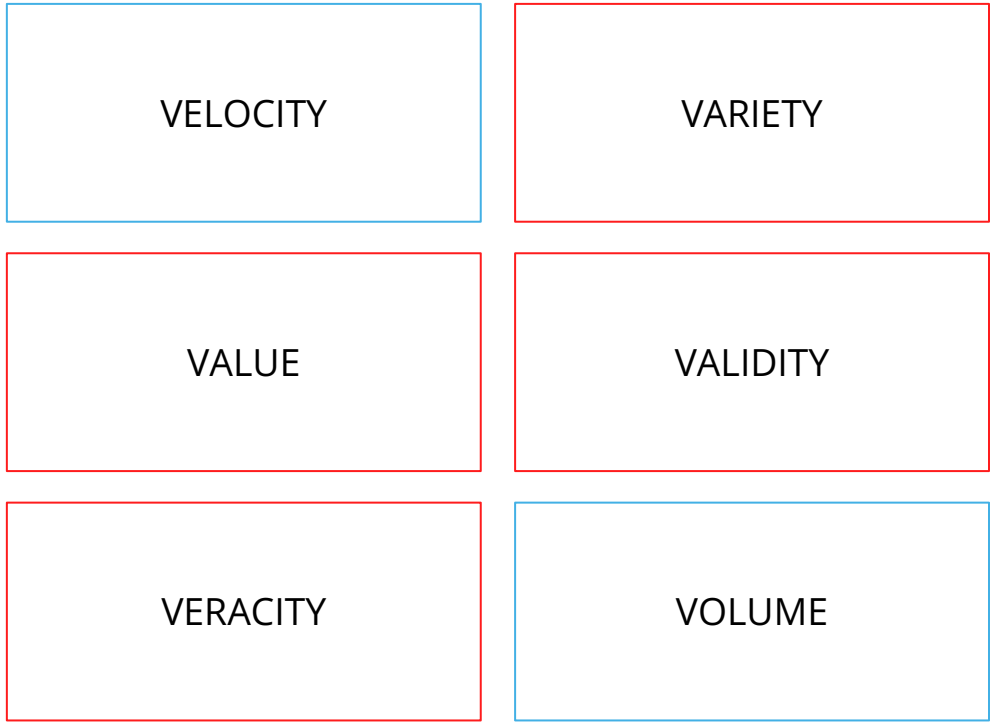
Scripts can be written to automatically ingest, explore, analyze, and report an analysis. The ability to acquire data automatically and programmatically will save you time and increase productivity.

6 V'S OF BIG DATA

We can use the 6 V's to describe big data.

Velocity is the speed at which data moves and is ingested to a system.

Volume is the amount of data to scale (i.e., gigabytes).



BATCH OR REAL-TIME?

Batch processing occurs periodically (i.e, daily, weekly, or monthly).

Real-time processing occurs within seconds.

API ENDPOINT

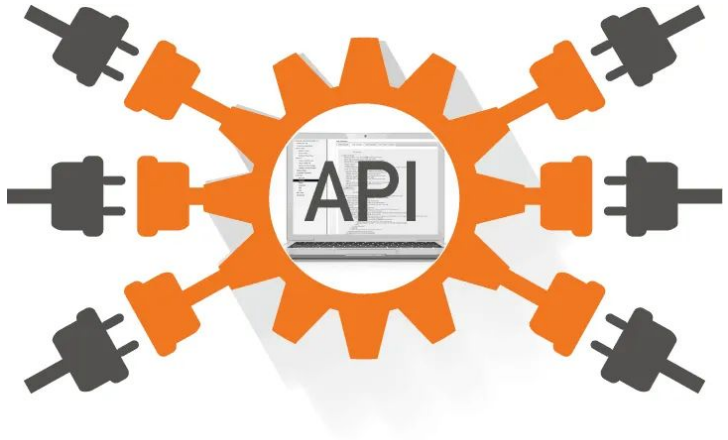
Application programming interfaces (APIs) allow programmatic access to data and functions without having to go to a webpage and download them.

Image Source: Netreo

HOW DO WE USE APIs?

We can use programs like Python to access APIs programmatically with code.

Function libraries are built-in modules that help Python access APIs that wouldn't be accessible otherwise.



From: Learn Steps

ENDPOINT EXAMPLE

REVIEW AND WRAP-UP

Today you learned about:

- The role of process models in data science analytics projects
- The elements of an effective problem statement
- The discovery phase in the data life cycle
- Different data sources for manual collection
- The characteristics of batch and real-time processing
- How APIs contribute to automatic data collection

YOUR TAKE

- Reflect on what you have learned so far.
- Share key takeaways.

QUESTIONS?

NEXT STEPS

- [x] Assigned Activities
- [x]
- [x]