

# How does data science happen?

Dr. Ilkay Altintas and Dr. Leo Porter

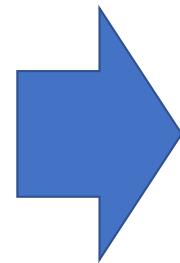
**Twitter:** #UCSDpython4DS

# After this video you will be able to..

- List some of the dimensions of modern data science
- Identify why analyzing these dimensions are important for us as data scientists

Data Science as a multidisciplinary craft that combines an interdisciplinary team with an application purpose.

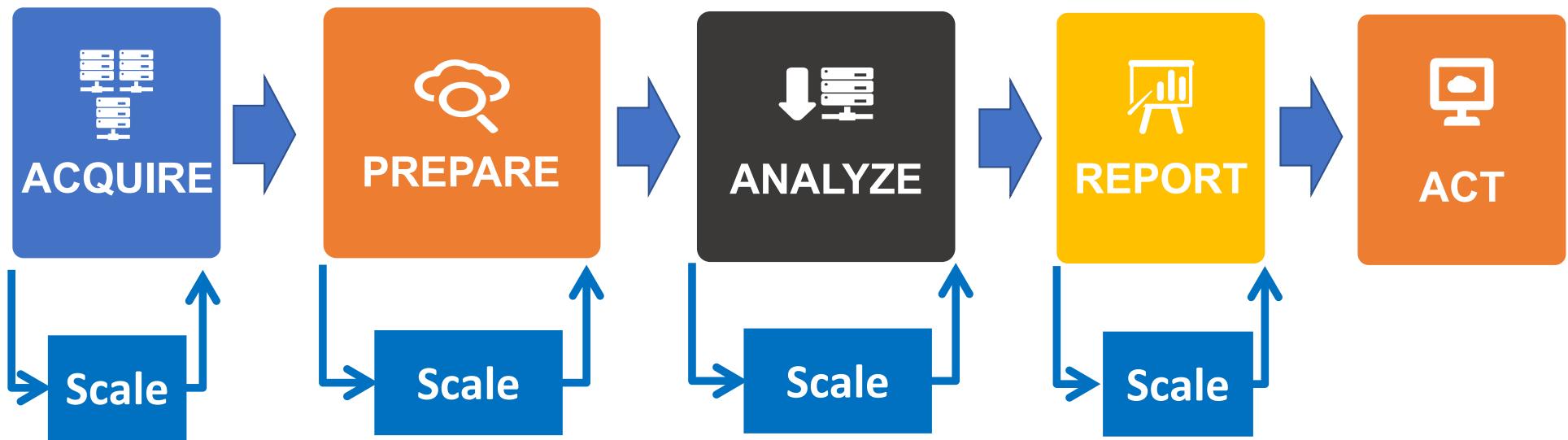


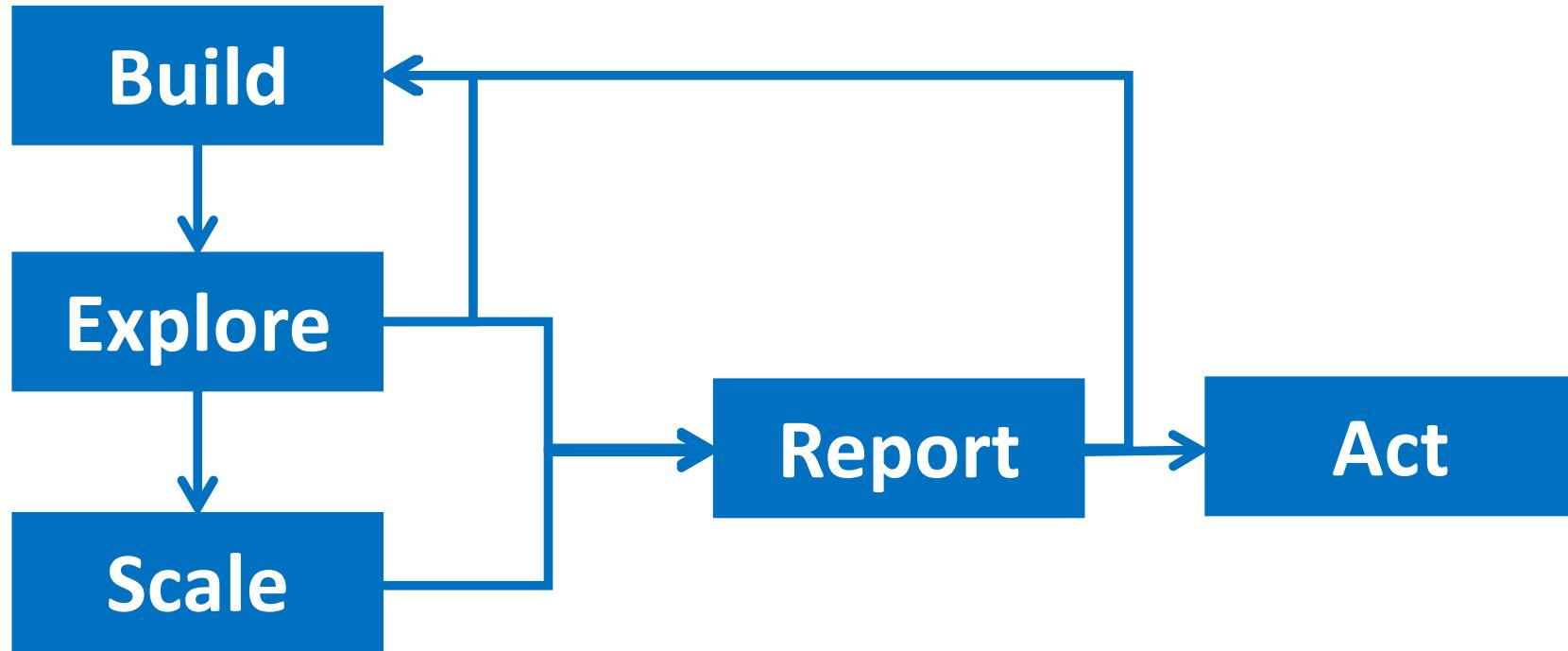


Data Science Process

## Data Engineering

## Computational Data Science





## Data Engineering

## Computational Data Science



Scale

Scale

Scale

Scale

Python for Data Science

Programmability

# Asking the Right Question

Dr. Ilkay Altintas and Dr. Leo Porter

**Twitter:** #UCSDpython4DS

# After this video you will be able to..

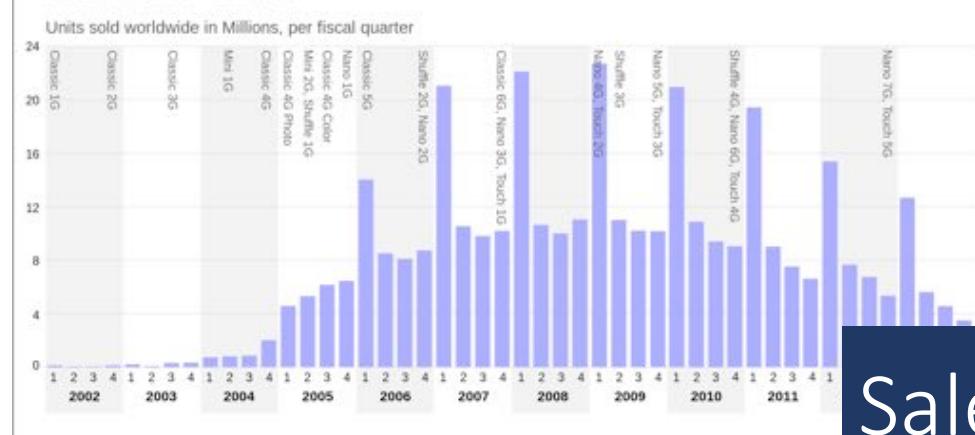
- Describe the ingredients to form a data science problem
- List some questions others asked to get value of their big data
- Formulate the right questions to guide your data science process.



**“A problem well defined  
is a problem half  
solved.”**

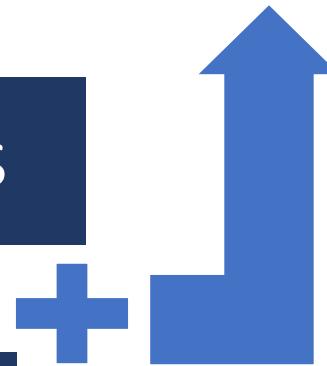
**Charles F. Kettering**

Define the Problem



## Evaluate a new product

Sales figures



## Call center logs



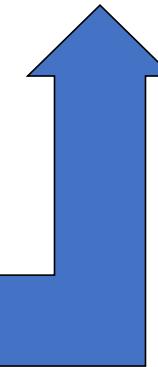
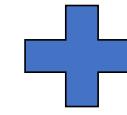


Detect equipment failure

Sensor data

Sensor data

Sensor data





Customer data

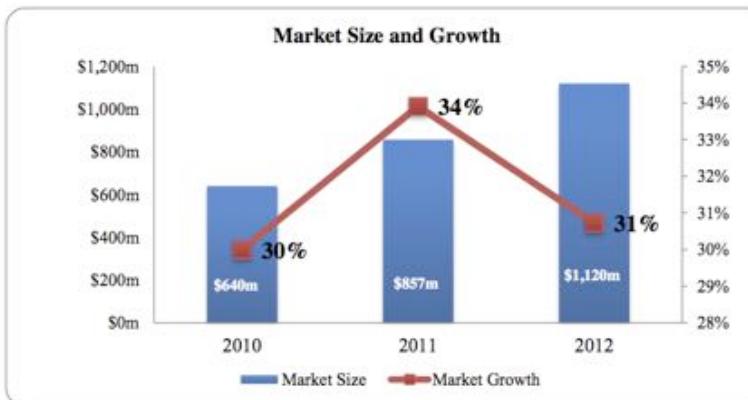


Marketing data

Better targeted  
marketing



## 1 Market Size and Growth



©The LPO Program 2012



Assess the Situation

# Assess the Situation

Risks
Benefits
Contingencies
Regulations
Resources
Requirements



*Answer these questions*

# Define Goals



Objectives

Criteria

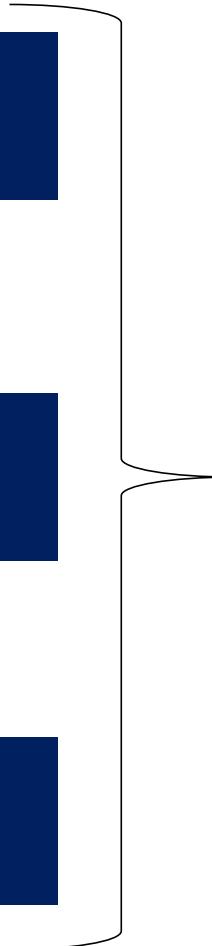
Define the Problem



Assess the Situation



Define Goals



Formulate the Question

# Steps in the Data Science Process

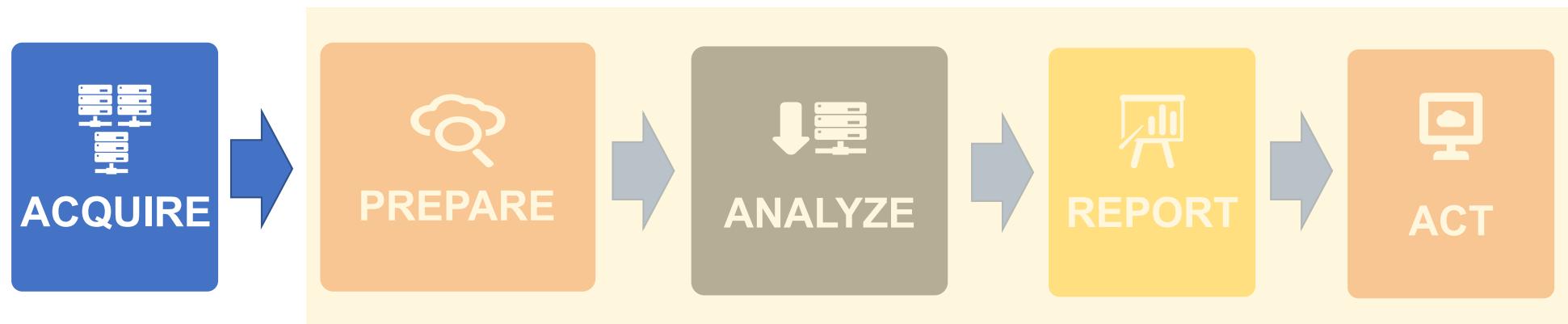
Dr. Ilkay Altintas and Dr. Leo Porter

**Twitter:** #UCSDpython4DS

After this video you will be able to..

- Identify the steps in the data science process
- Understand what each step involves





## Step 1: Acquire Data

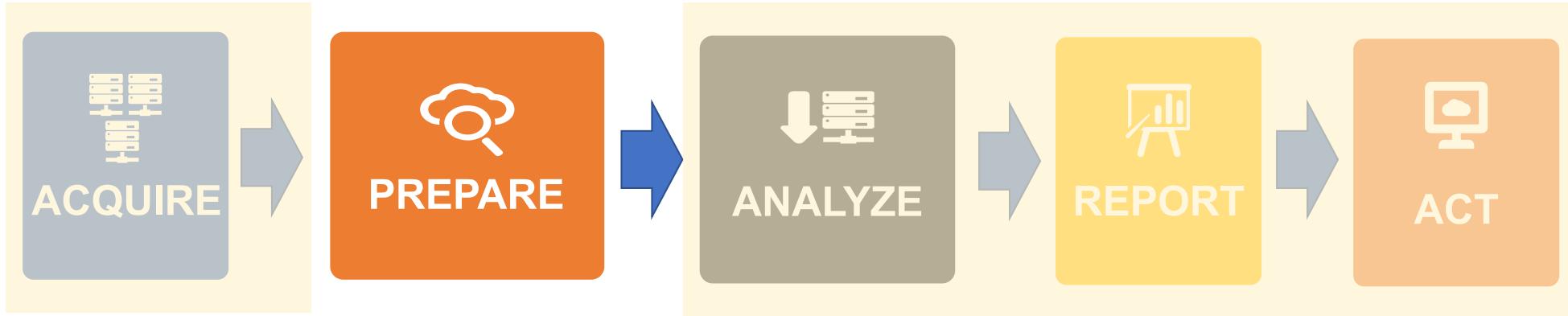


Identify data sets

Retrieve data

Query data

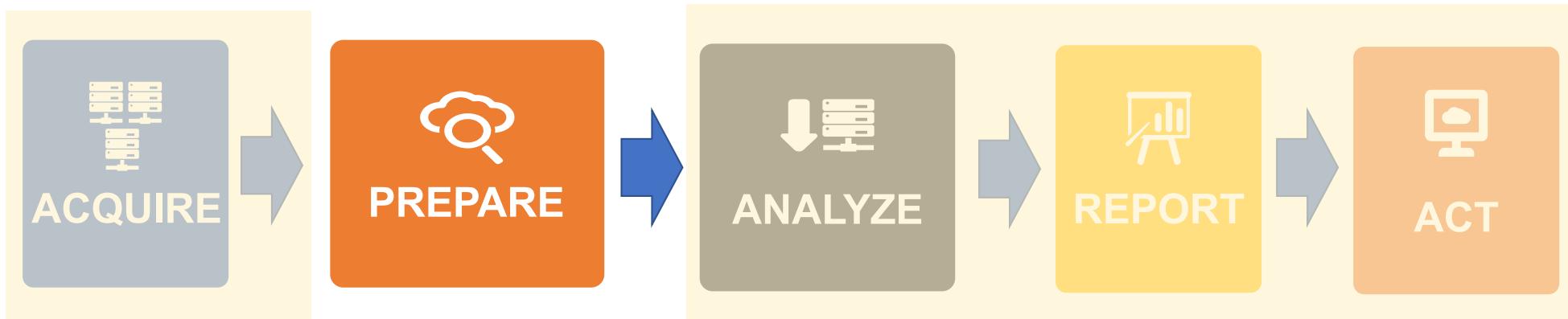
*moving data*



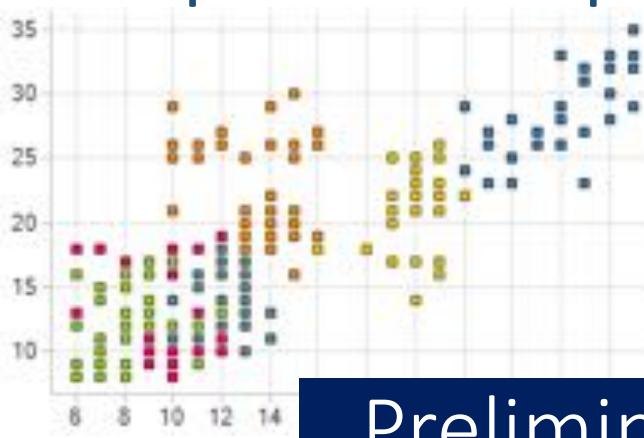
## Step 2: Prepare Data

Step 2-A: Explore

Step 2-B: Pre-process

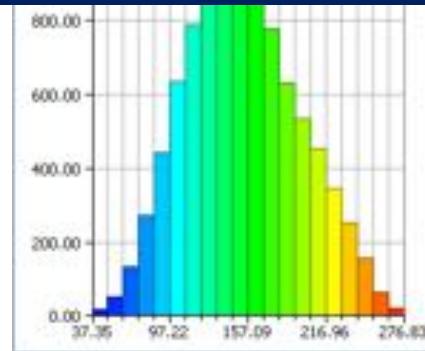


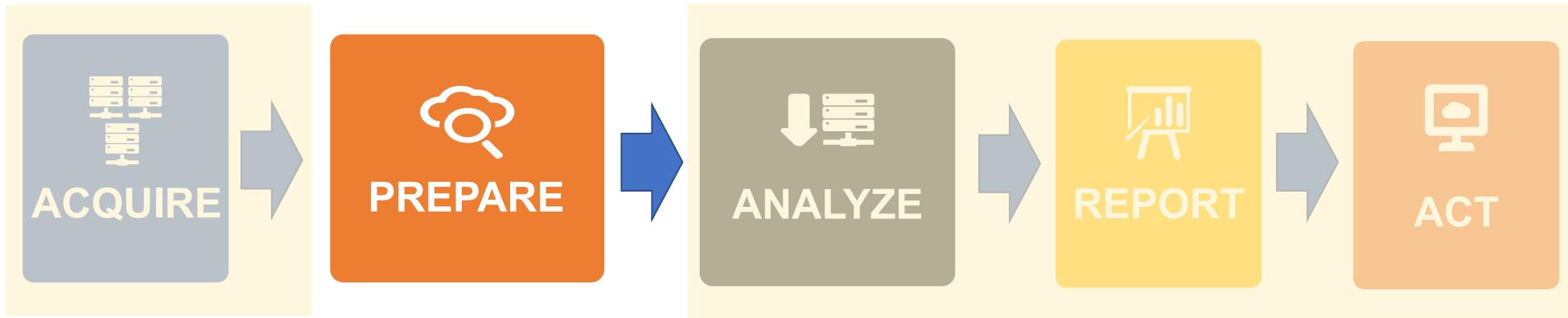
## Step 2-A: Explore Data *(Prepare)*



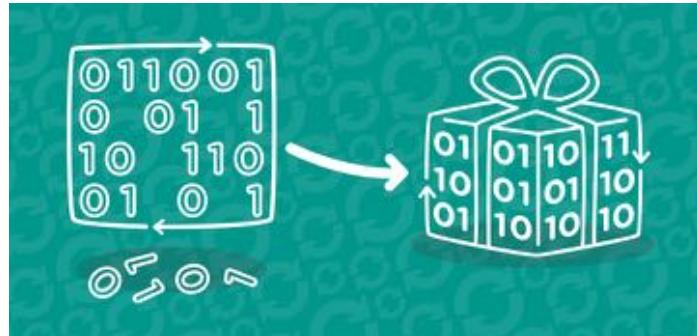
Preliminary analysis

Understand  
nature of data





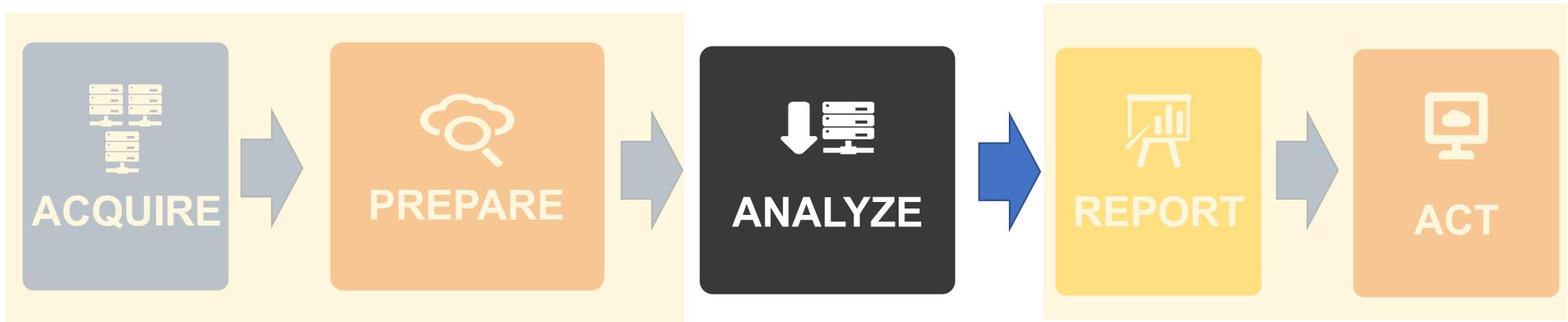
## Step 2-B: Pre-process Data



Clean

Integrate

Package



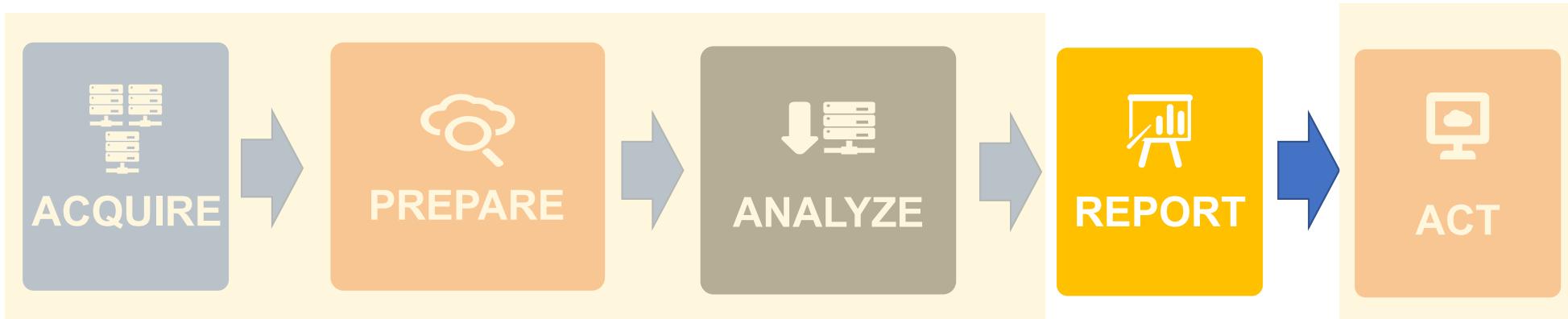
## Step 3: Analyze Data

Select analytical techniques

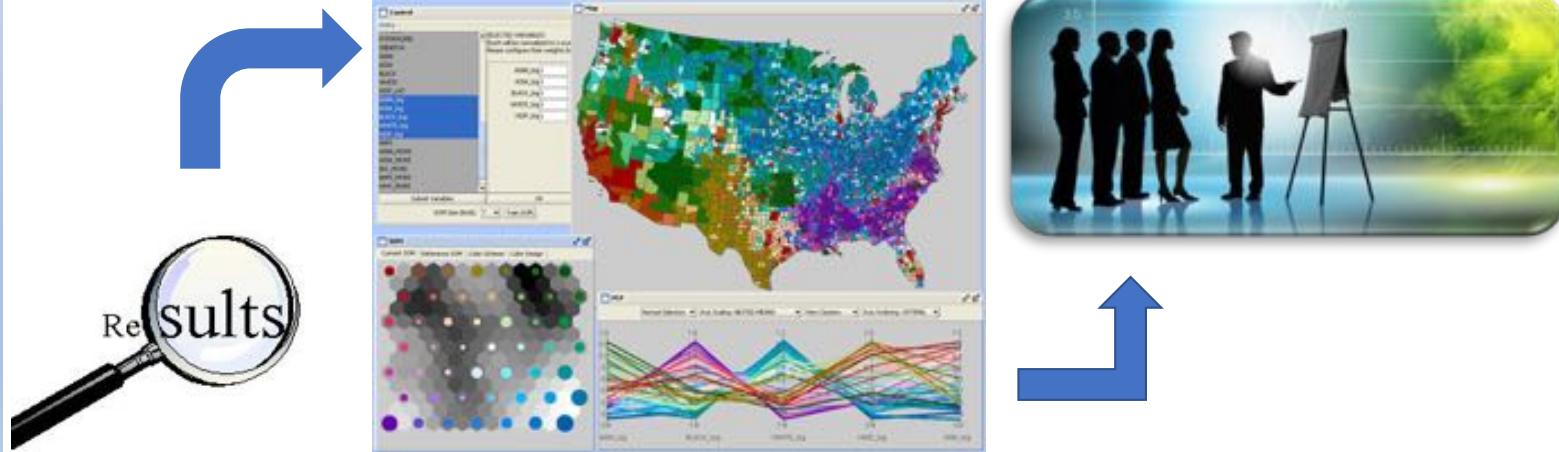
Build models

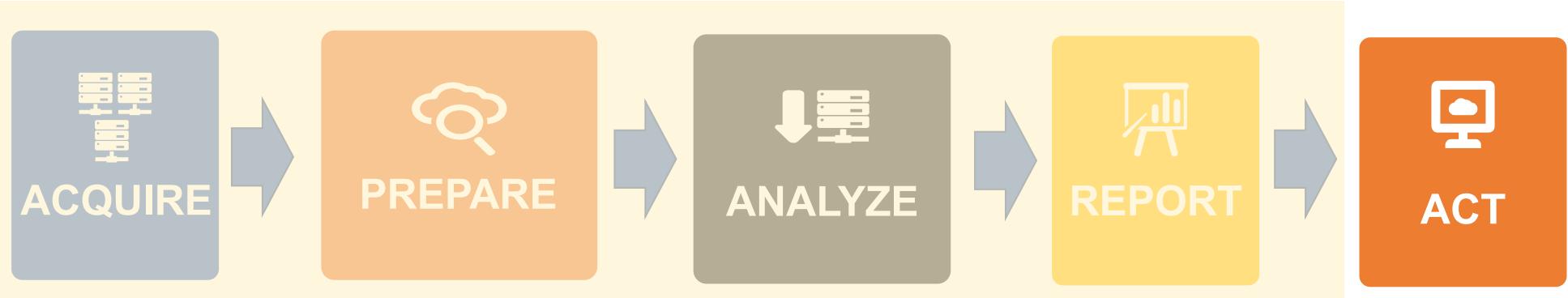


Some time need to  
go back for step 1 & 2

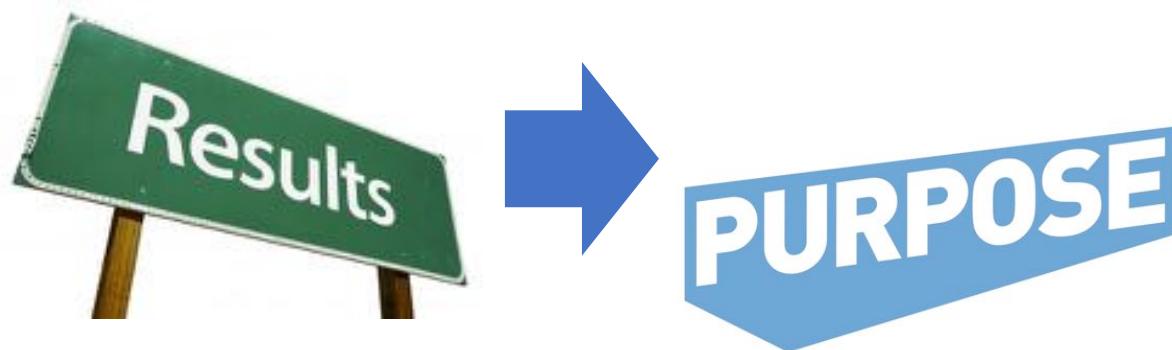


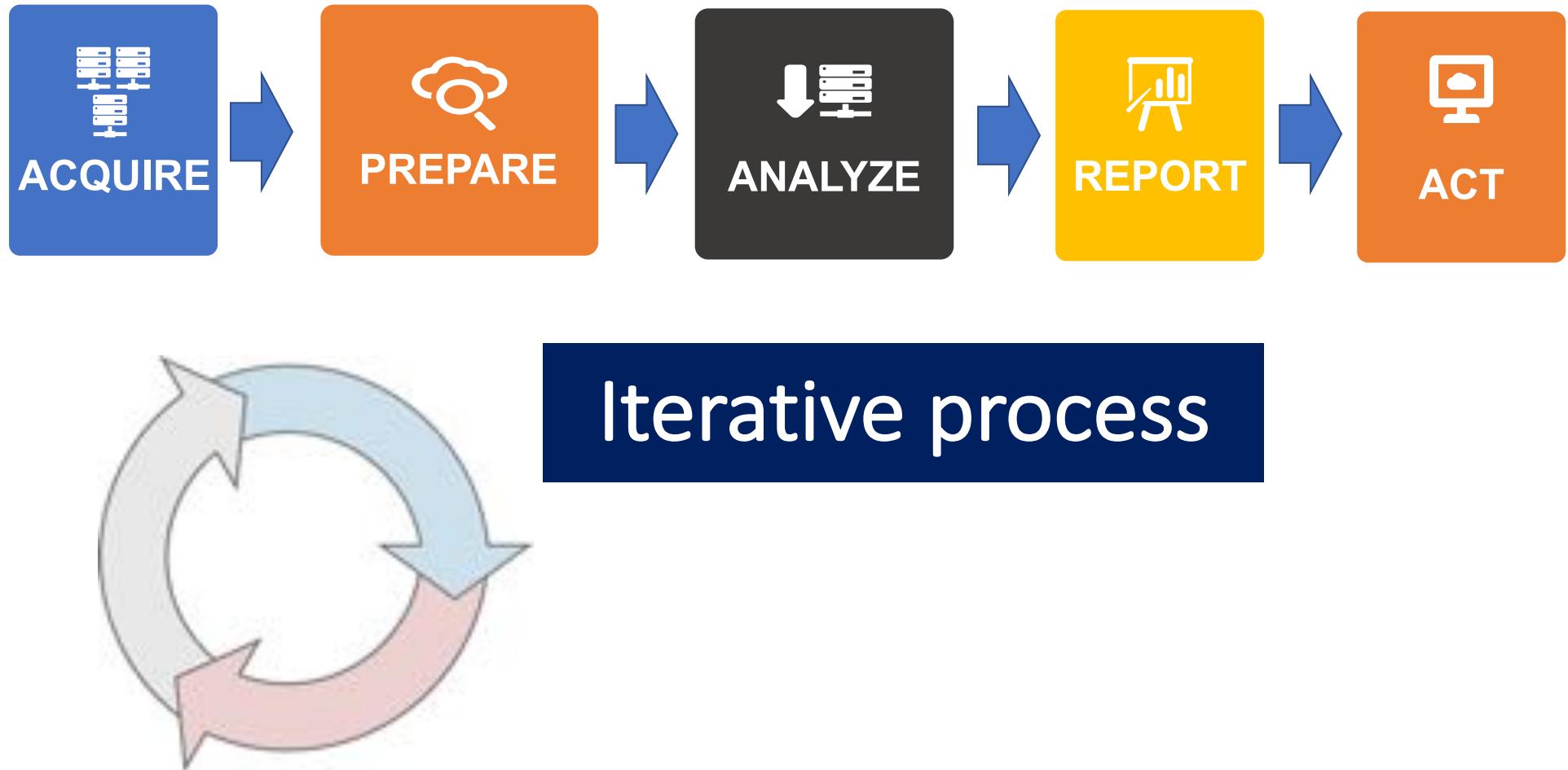
## Step 4: Communicate Results





## Step 5: Apply Results





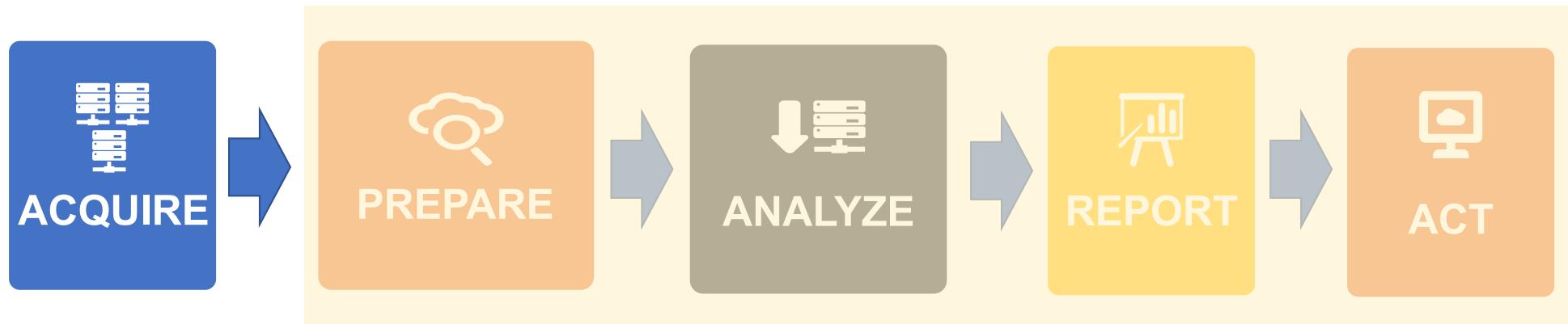
# Step 1: Acquiring Data

Dr. Ilkay Altintas and Dr. Leo Porter

**Twitter:** #UCSDpython4DS

After this video you will be able to..

- List techniques and technologies to access and retrieve the data you need
- Describe an example scenario that accesses data from a variety of sources using different technologies



## Step 1: Acquire Data

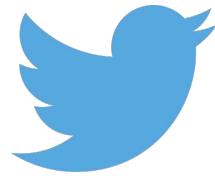
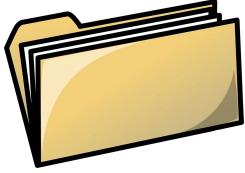


- Identify datasets
- Retrieve datasets
- Query data

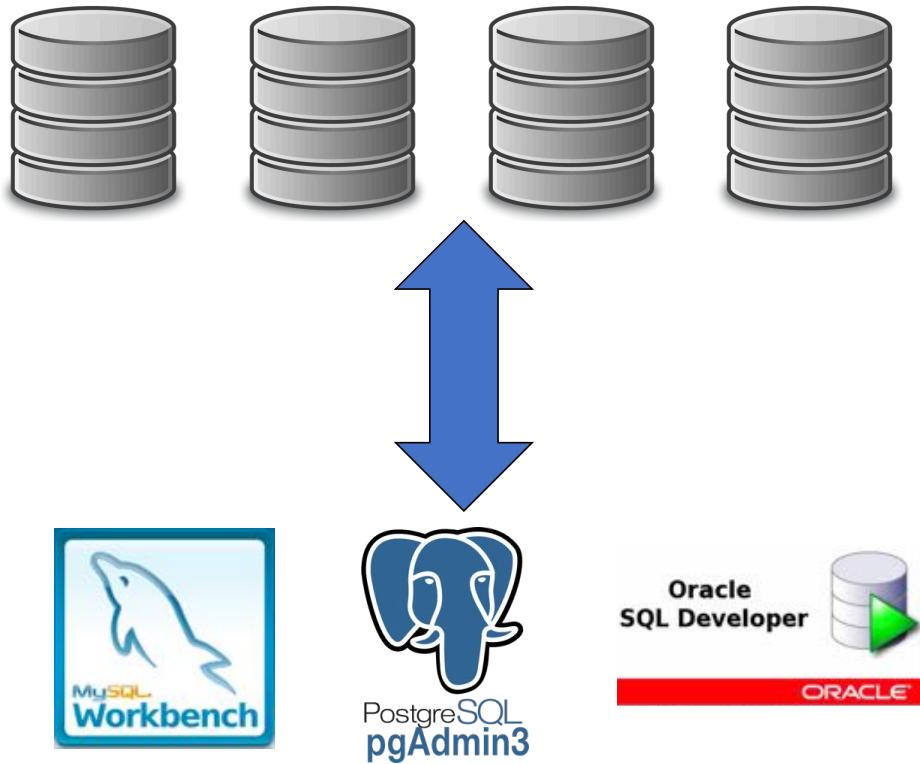
# Where's the data?

- Identify suitable data
- Acquire all available data

Data comes from many places...



...with many ways to access it

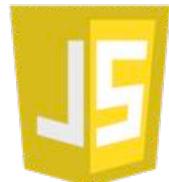


Traditional databases

SQL and query browsers



Ruby



JavaScript



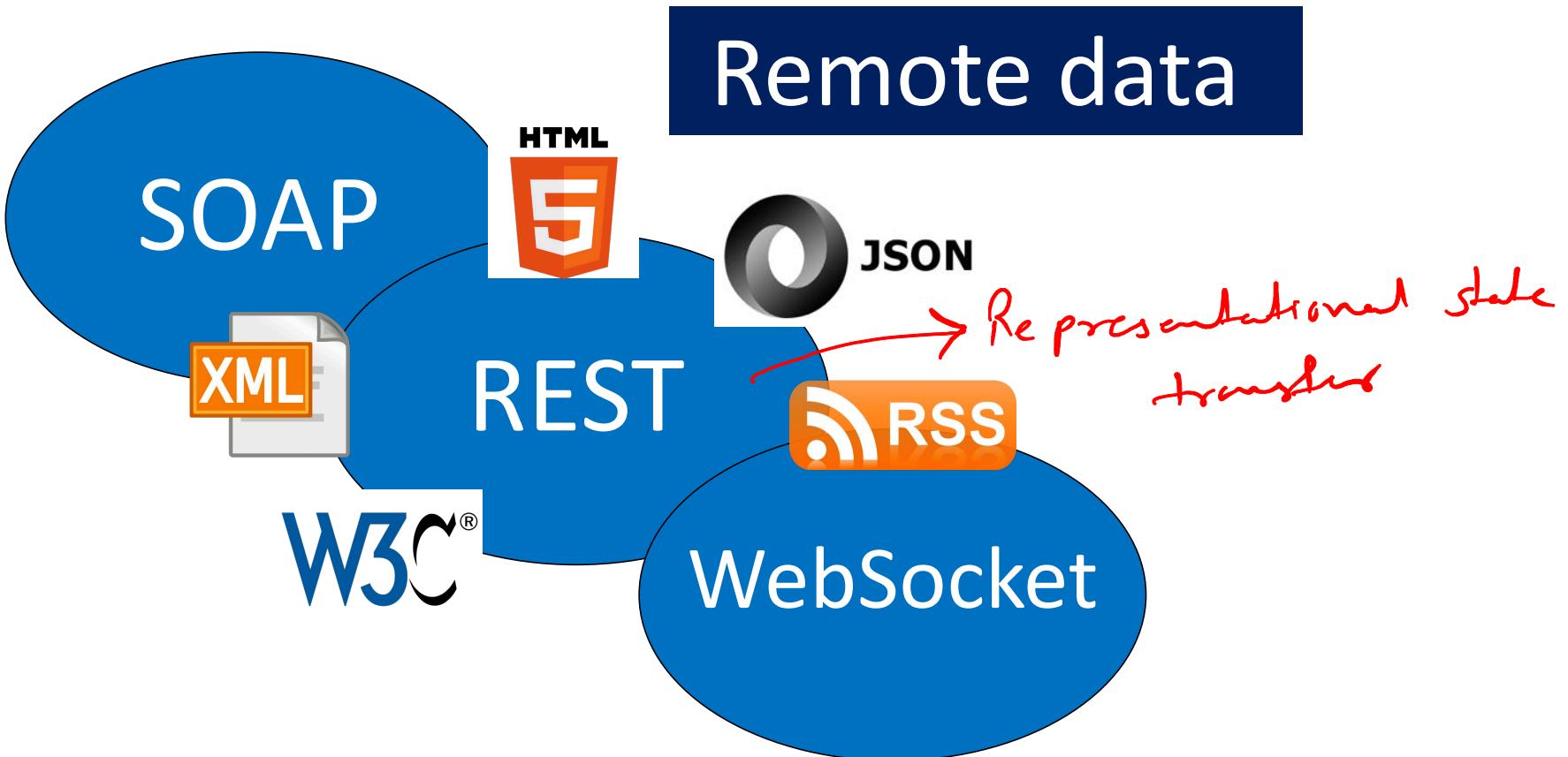
Text files



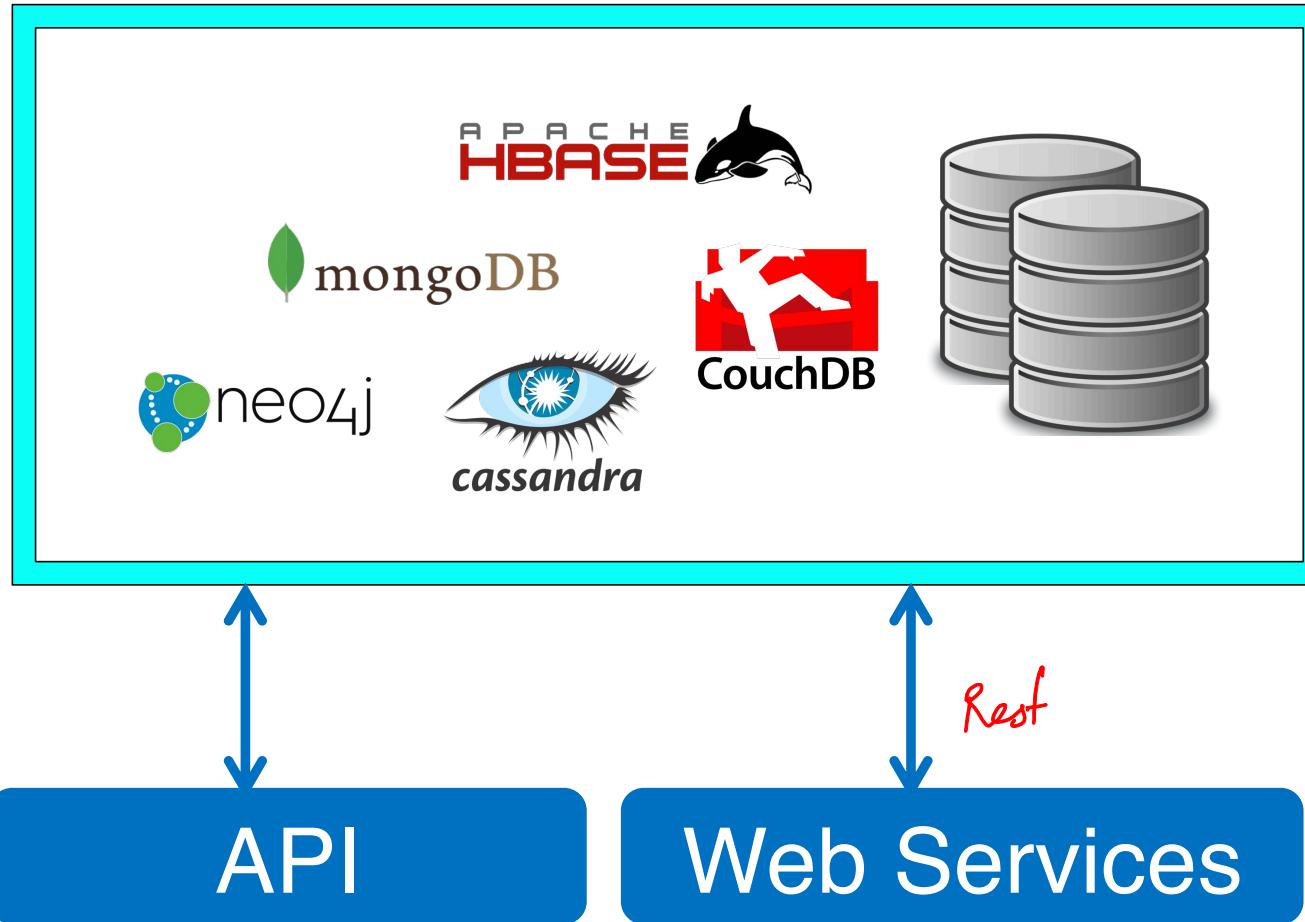
*Octave  
Matlab*

Scripting languages

## Web Services



## NoSQL storage



# Acquiring data related to wildfires...

Historical weather

SQL



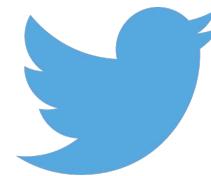
Current weather

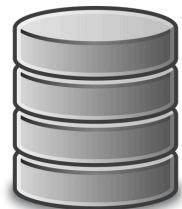
WebSocket



Real-time tweets  
near fires

REST





Traditional databases

SQL and query browsers

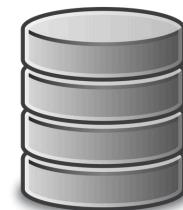


Remote data

Web Services



Text files



NoSQL storage

Scripting languages

Web Services

Programming Interfaces

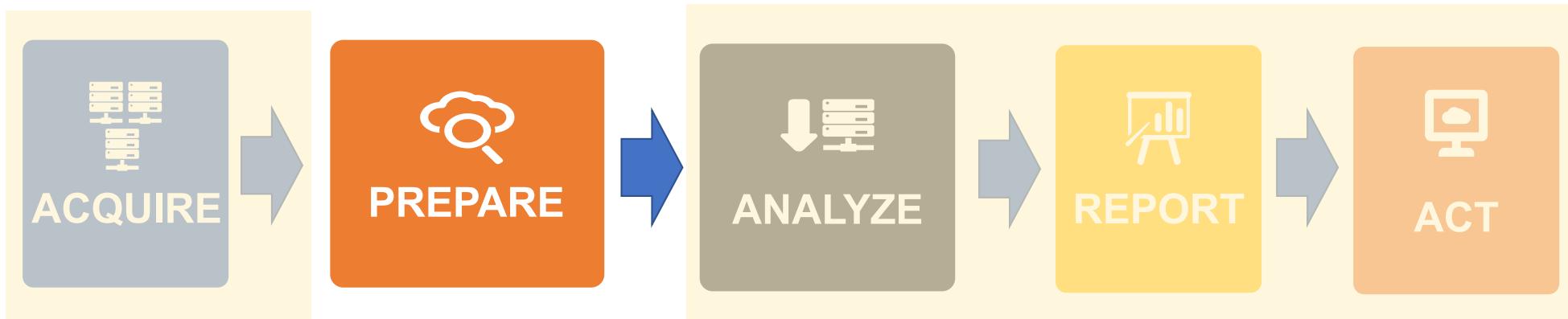
# Step 2-A: Exploring Data

Dr. Ilkay Altintas and Dr. Leo Porter

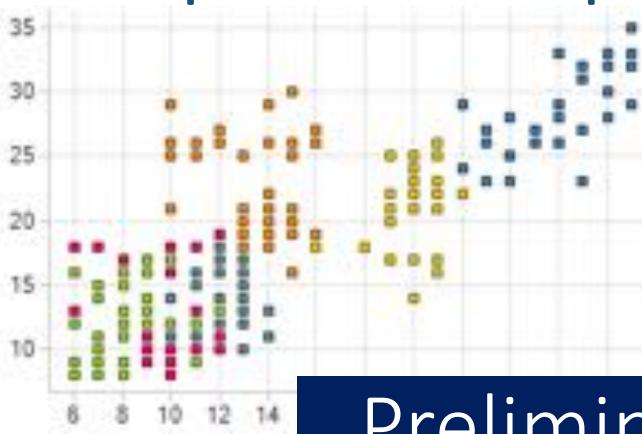
**Twitter:** #UCSDpython4DS

After this video you will be able to..

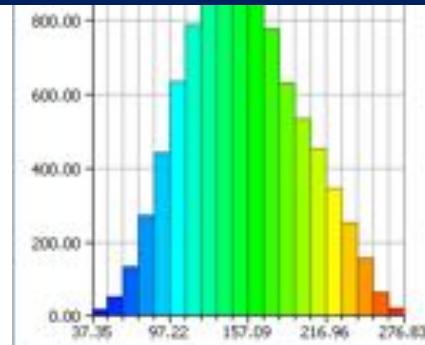
- Explain the importance of exploring data
- Identify methods to perform preliminary analysis of your data



## Step 2-A: Explore Data



Understand  
nature of data



Preliminary  
analysis



## Why explore?

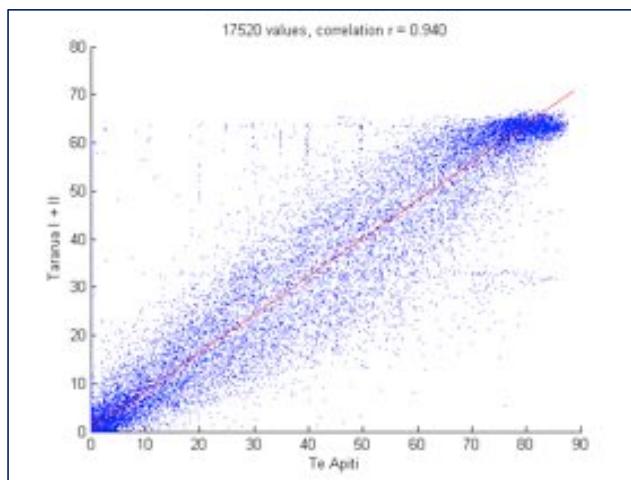
Goal: Understand your data

*Looking correlations, general trends, outliers etc.*



# Why explore?

## Correlations

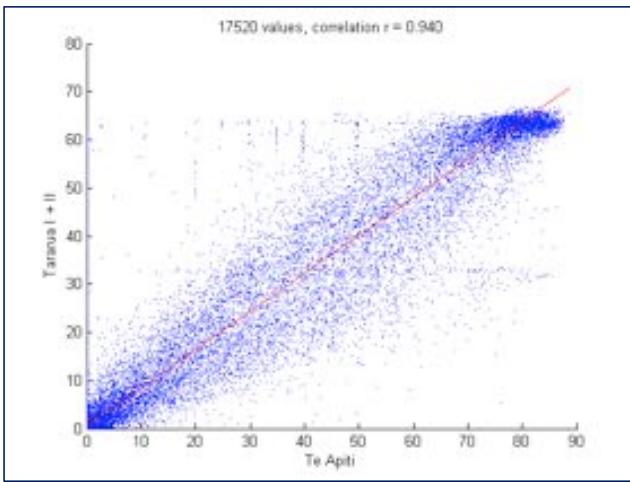


explore dependencies  
between different variables  
in data.

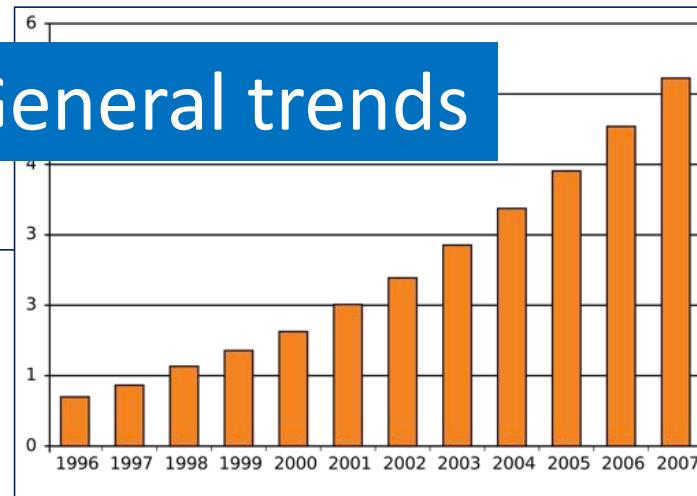


## Why explore?

### Correlations

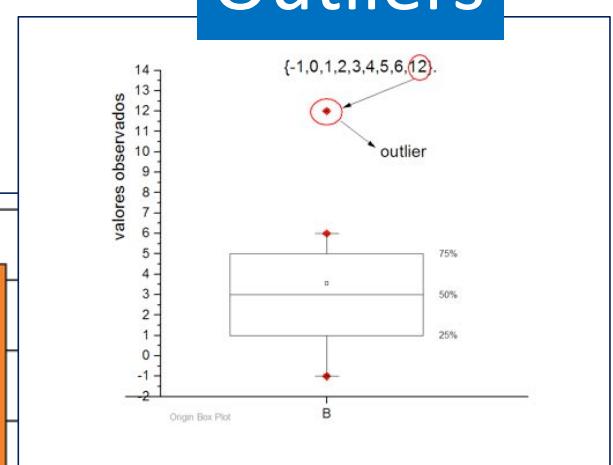


### General trends



how the data is progressing over time

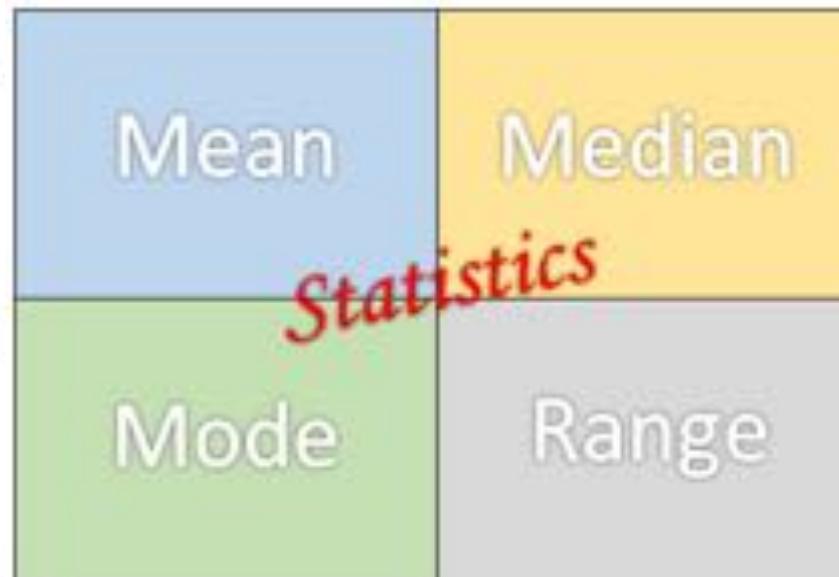
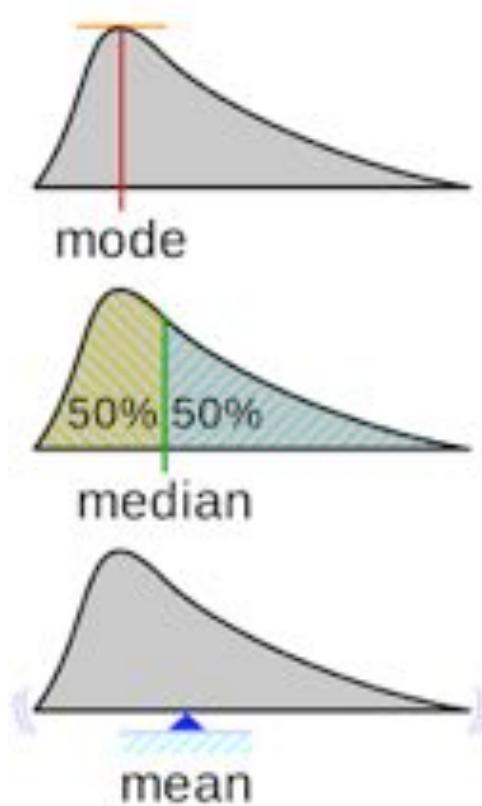
### Outliers



data points that are distant from other data point.

Summary statistics are quantities that capture various characteristics of a set of values with a single number or a small set of numbers.

## Describe Your Data

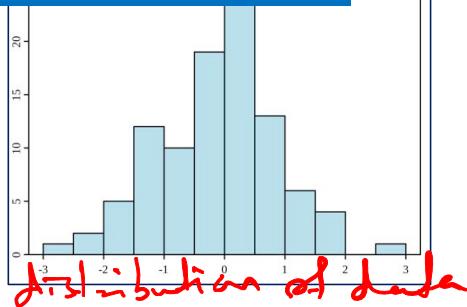


standard deviations.

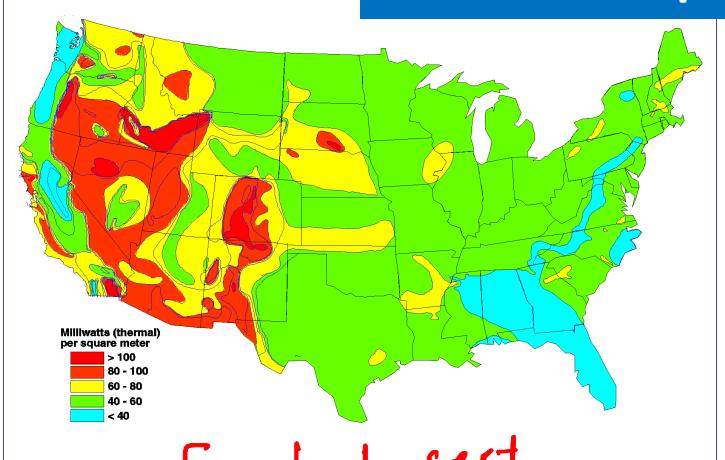
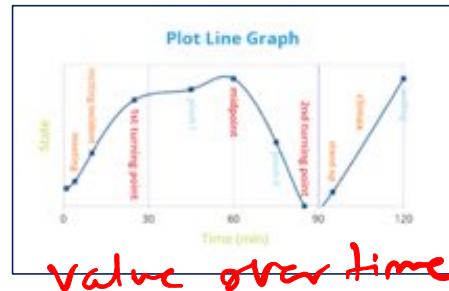
# Visualize Your Data

## Heat Maps

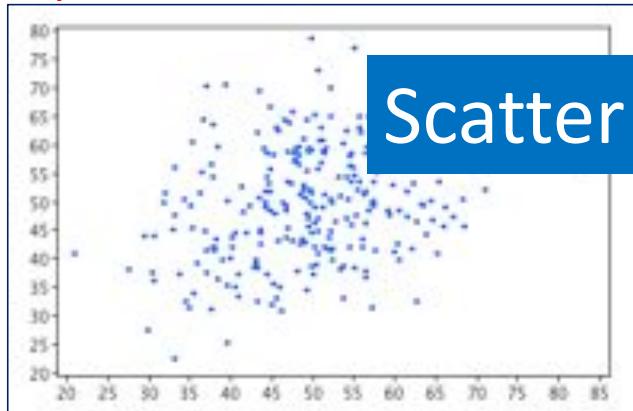
### Histogram



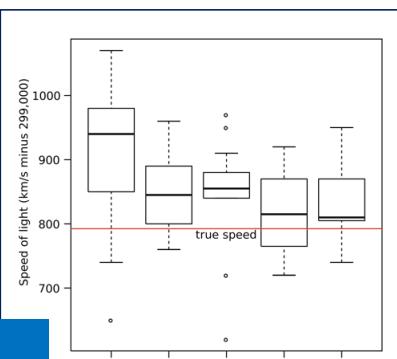
### Line graphs

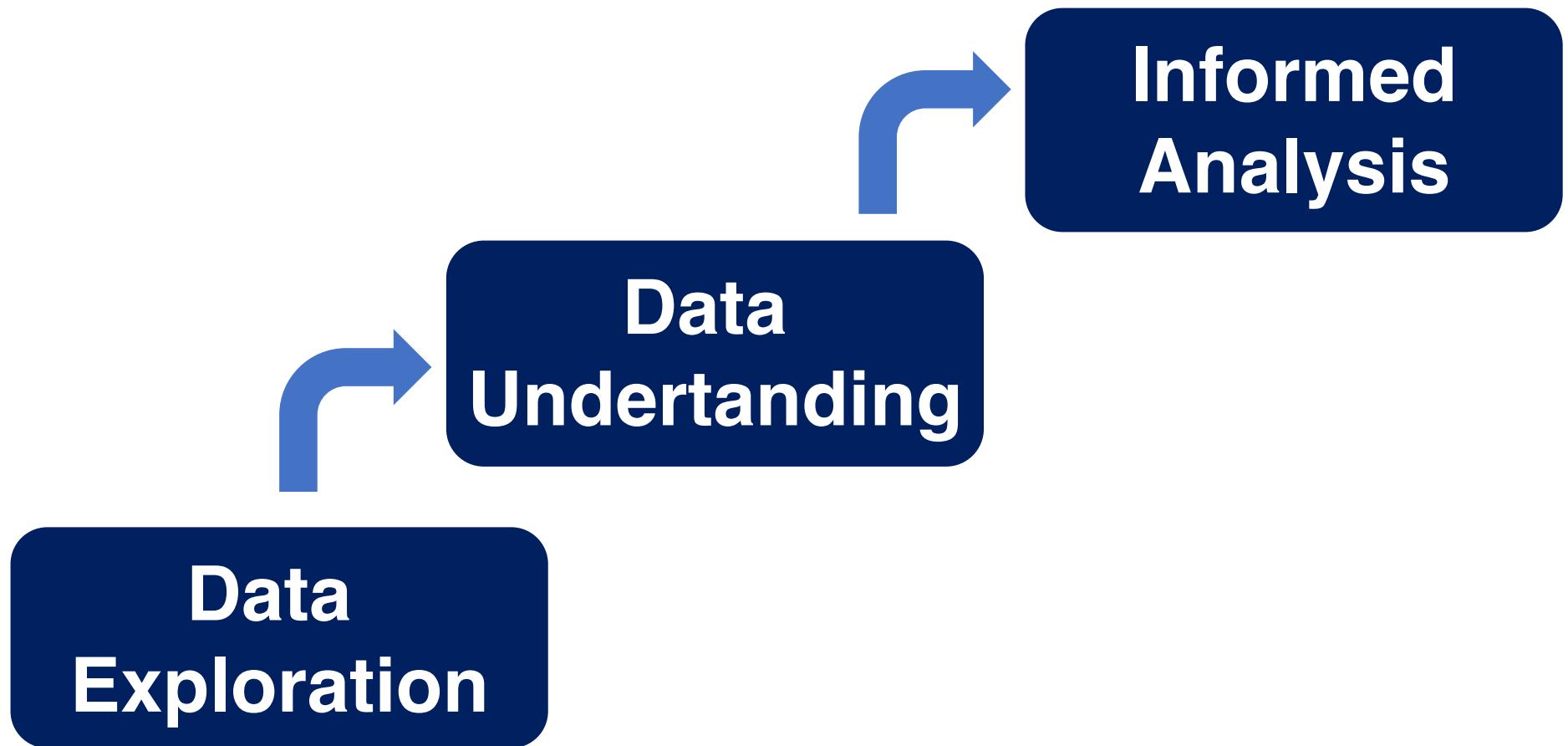


### Scatter plots



### Boxplots





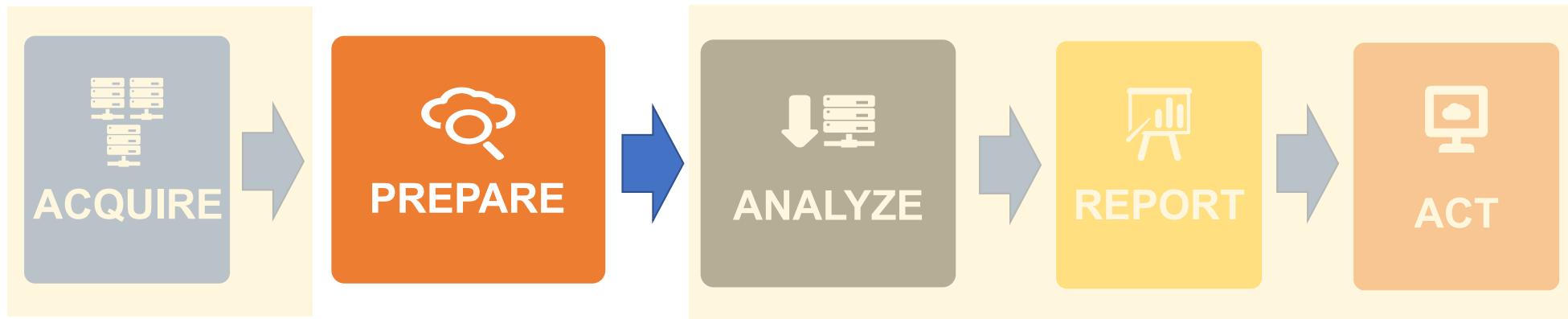
# Step 2-B: Pre-processing Data

Dr. Ilkay Altintas and Dr. Leo Porter

**Twitter:** #UCSDpython4DS

After this video you will be able to..

- Identify some problems with real-world data
- Describe what is needed to transform raw data to data that can be used for analysis



## Step 2-B: Pre-process Data

Clean + Transform

# Real-world data is messy!

- Inconsistent values
- Duplicate records
- Missing values
- Invalid data
- Outliers

# Addressing Data Quality Issues

- Remove data with missing values
- Merge duplicate records
- Generate best estimate for invalid values
- Remove outliers

*Domain  
Knowledge*

# Getting Data in Shape

Data  
Munging

Data  
Preprocessing

Data  
Wrangling



data  
manipulation

# Data Munging

*Dimensionality  
Reduction*

*Data  
Manipulation*

*Transformation*

*Feature  
Selection*

*Scaling*

# Scaling



Weight

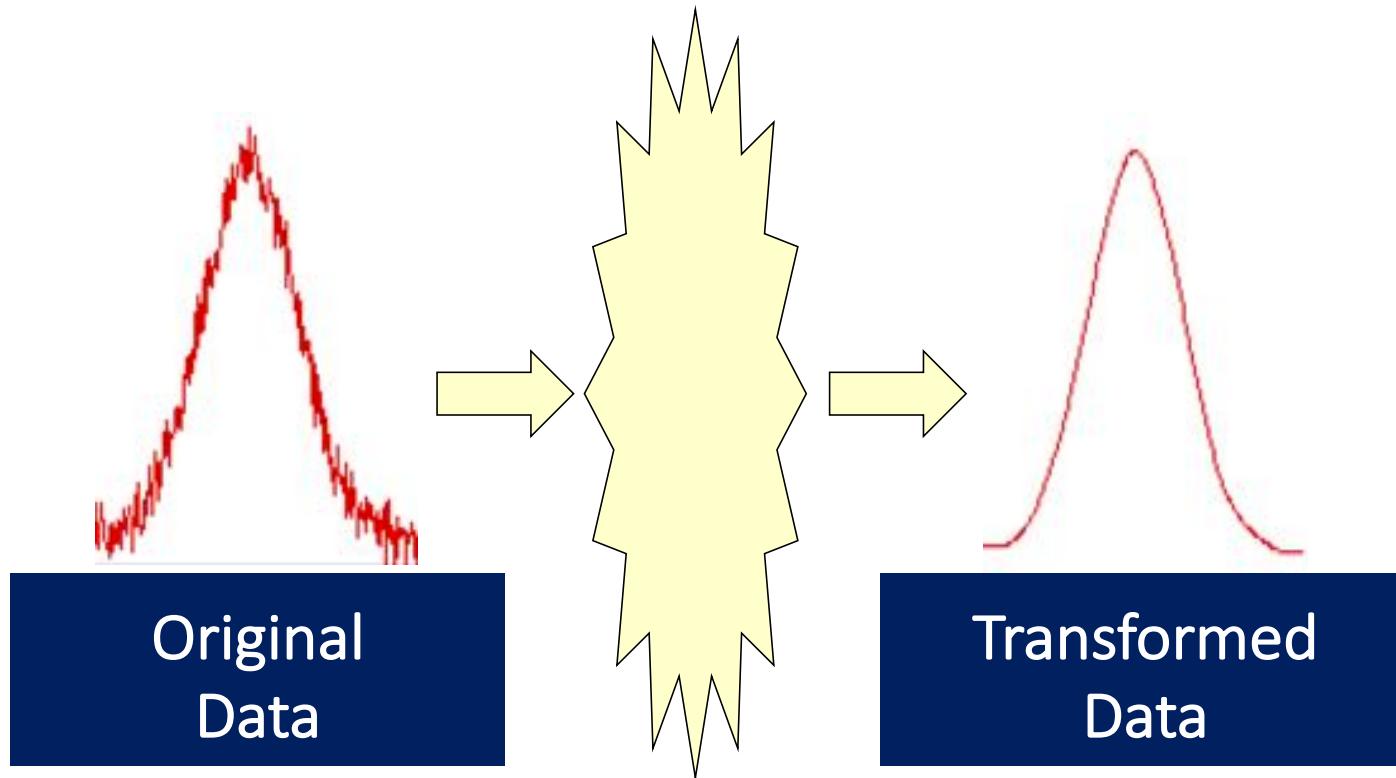
Scaled Values



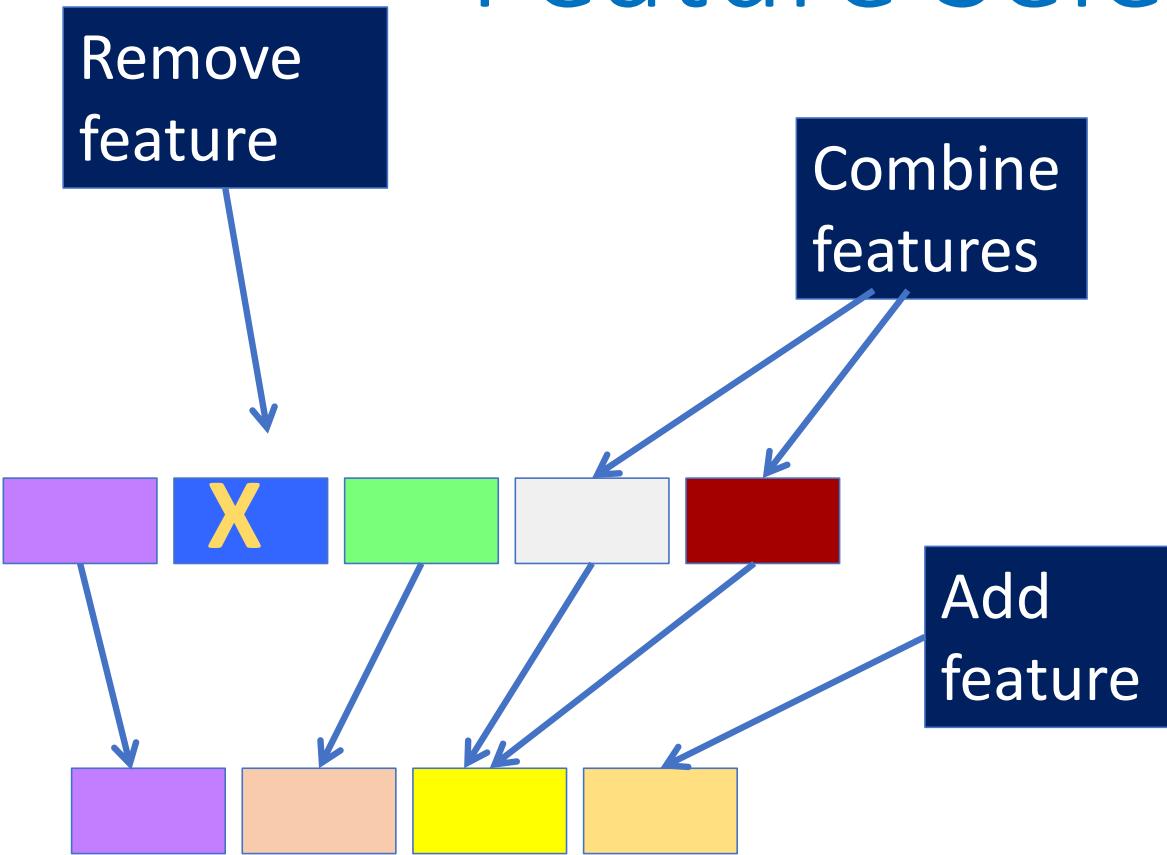
Height

# Transformation

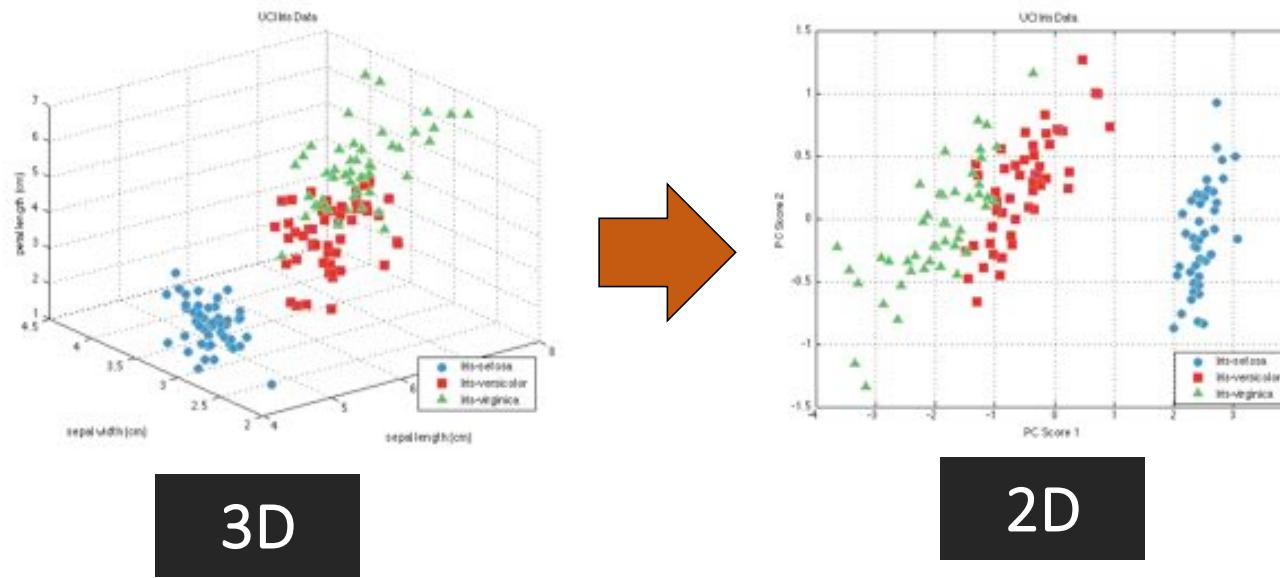
for Reduce  
Scale and noise



# Feature Selection

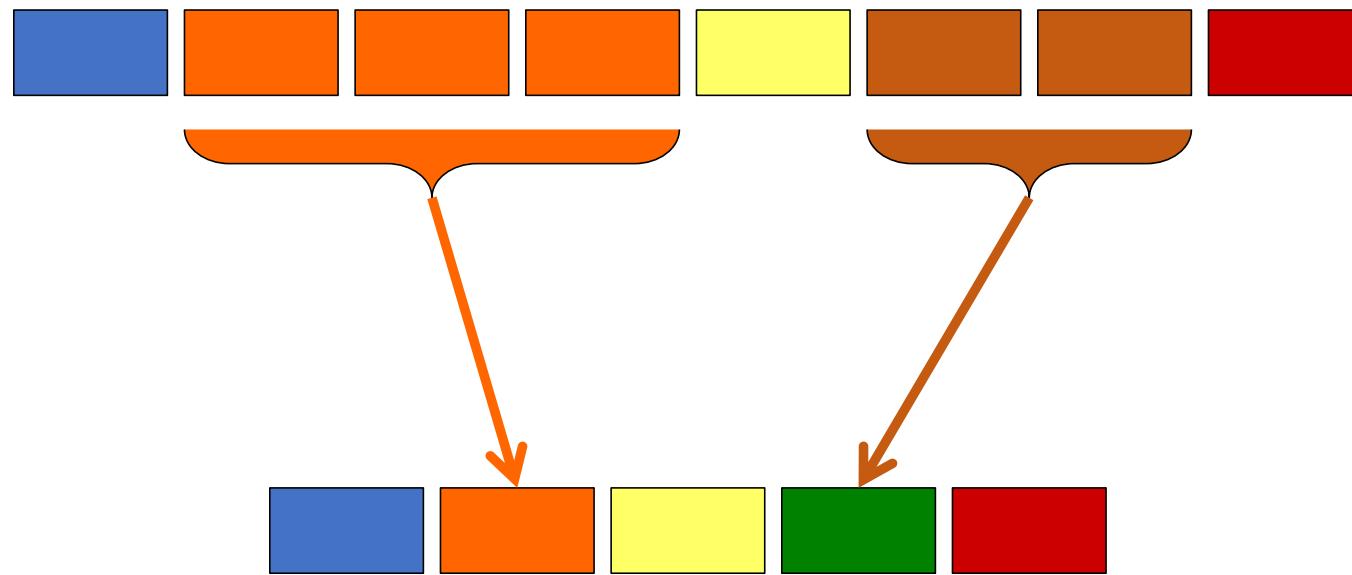


# Dimensionality Reduction



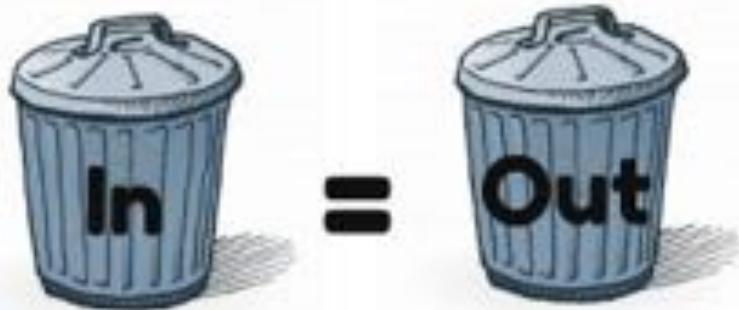
# Data Manipulation

*grouping them  
together*



# Always Remember!

Garbage in = Garbage out



Data preparation is  
very important for  
meaningful analysis!

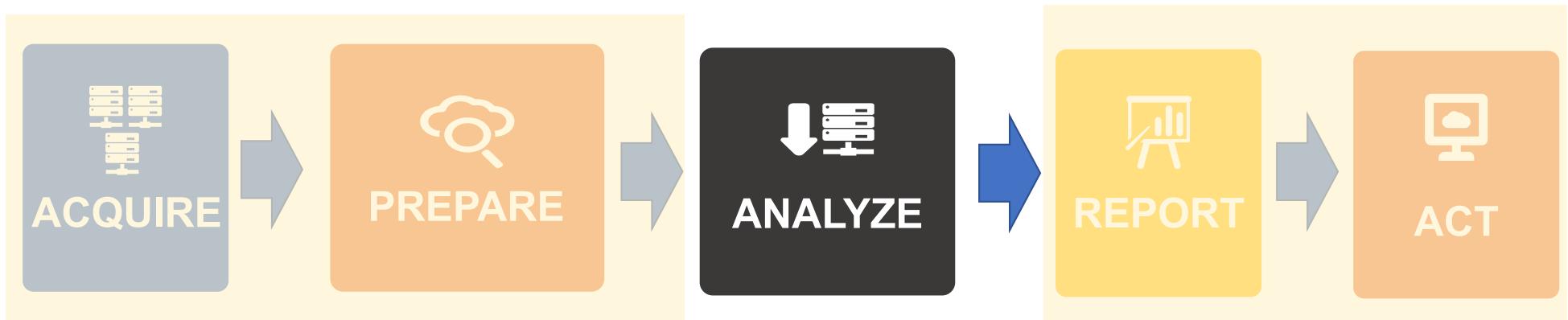
# Step 3: Analyze Data

Dr. Ilkay Altintas and Dr. Leo Porter

**Twitter:** #UCSDpython4DS

After this video you will be able to..

- Describe what is involved in applying an analysis technique to your data
- List three basic analysis techniques



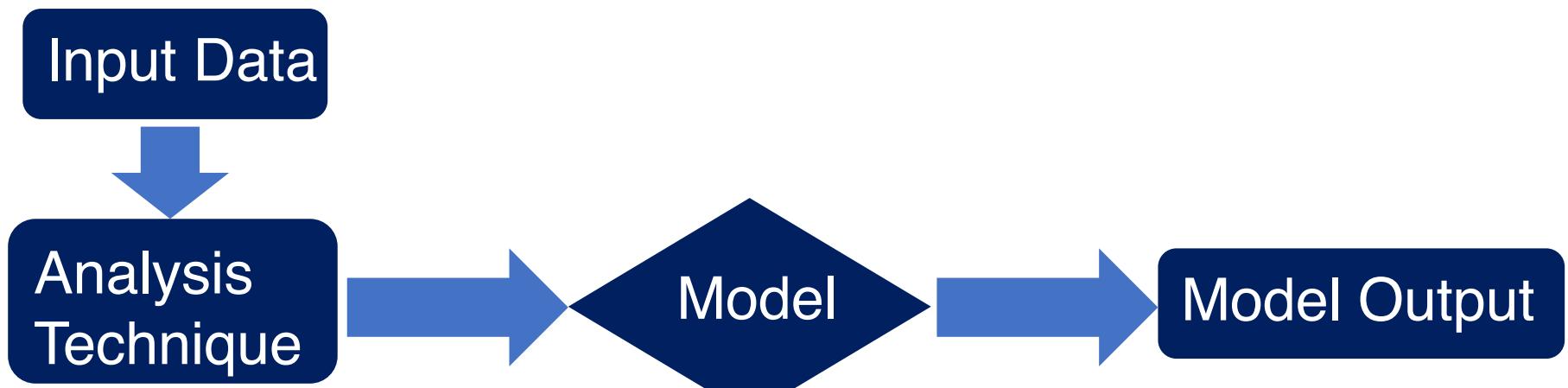
## Step 3: Analyze Data

Select analytical techniques

Build models



# Build Model



# Categories of Analysis Techniques

Classification

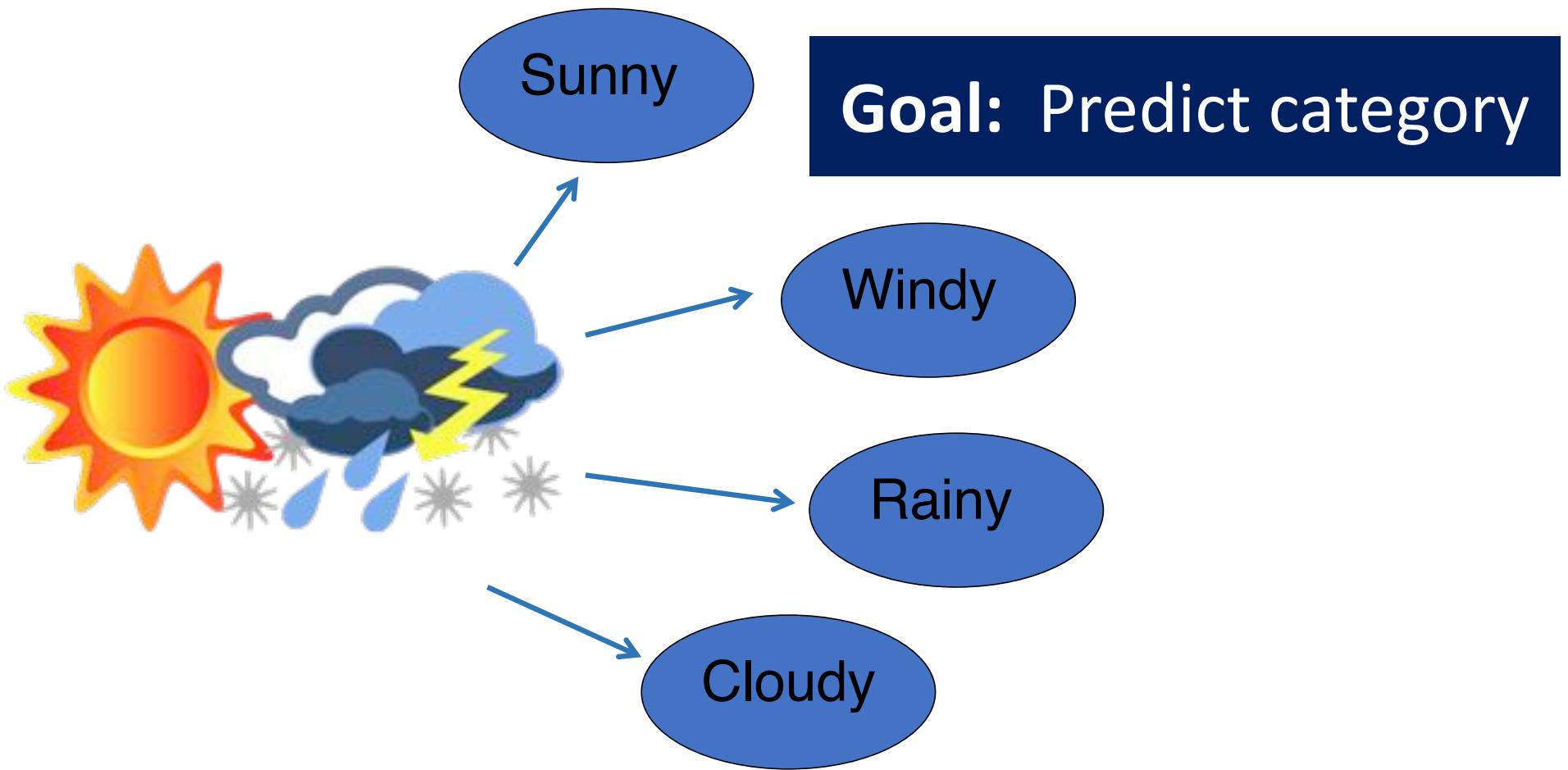
Regression

Clustering

Association  
Analysis

Graph  
Analytics

# Classification



# Regression

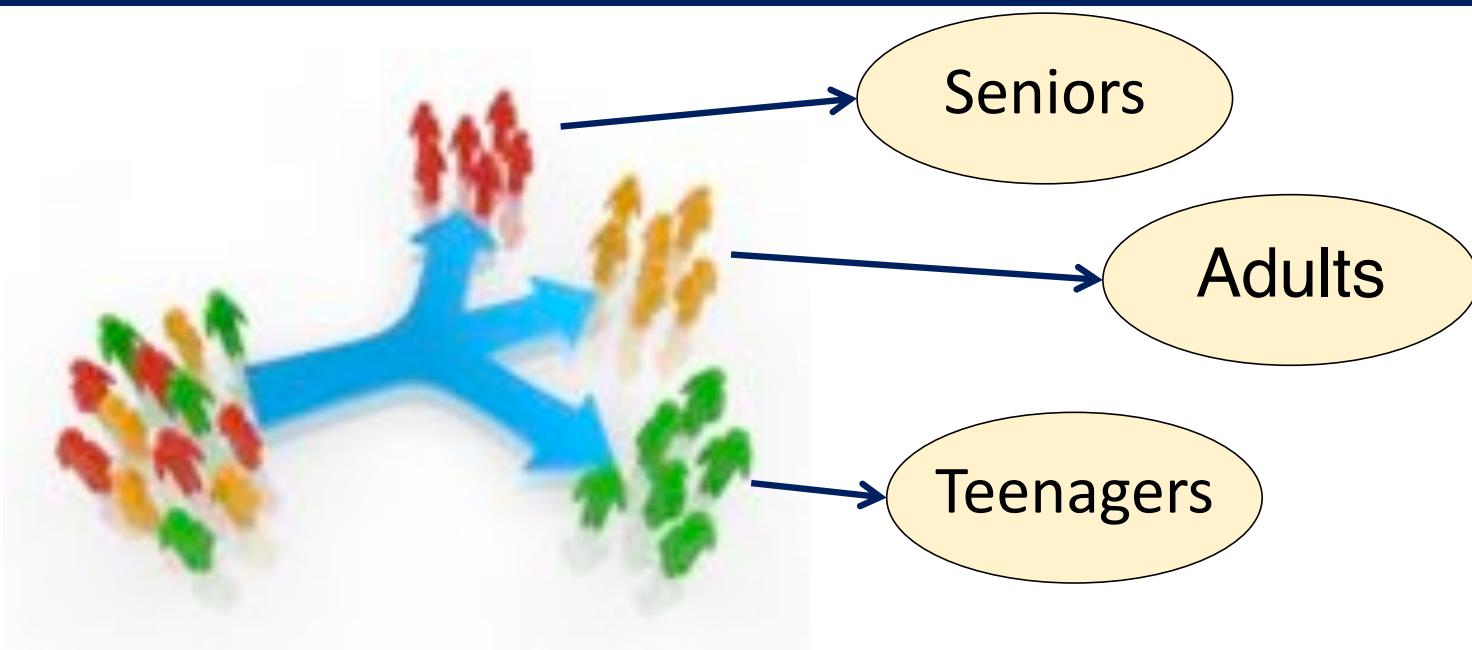
**Goal:** Predict numeric value



- Stock price
- weekly sales
- test score

# Clustering

**Goal:** Organize similar items into groups



# Association Analysis

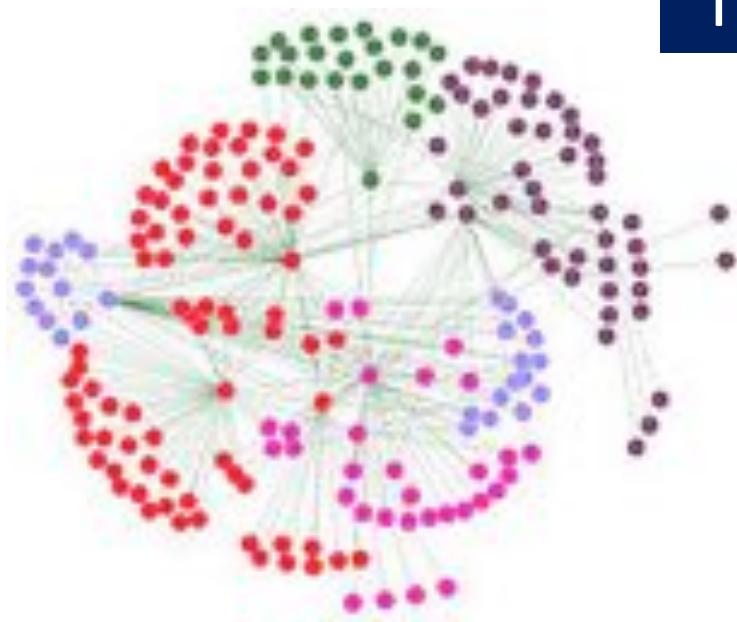


**Goal:** Find rules to capture associations between items



→ market basket analysis

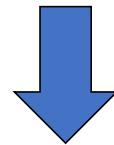
# Graph Analytics



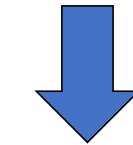
**Goal:** Use graph structures to find connections between entities

- exploring spread of a disease
- identification of security threats

Select technique



Build model



Validate model

Modeling

# Evaluation of Results

# Classification and Regression

Predicted  
Value



Correct  
Value

# Clustering



# Association Analysis and Graph Analytics



Investigate



Validate

# Determine Next Steps



Repeat analysis?

Take deeper dive?

Act on results?

## Select technique

- Classification
- Regression
- Clustering
- Association
- Analysis
- Graph Analytics

## Build model



## Evaluate



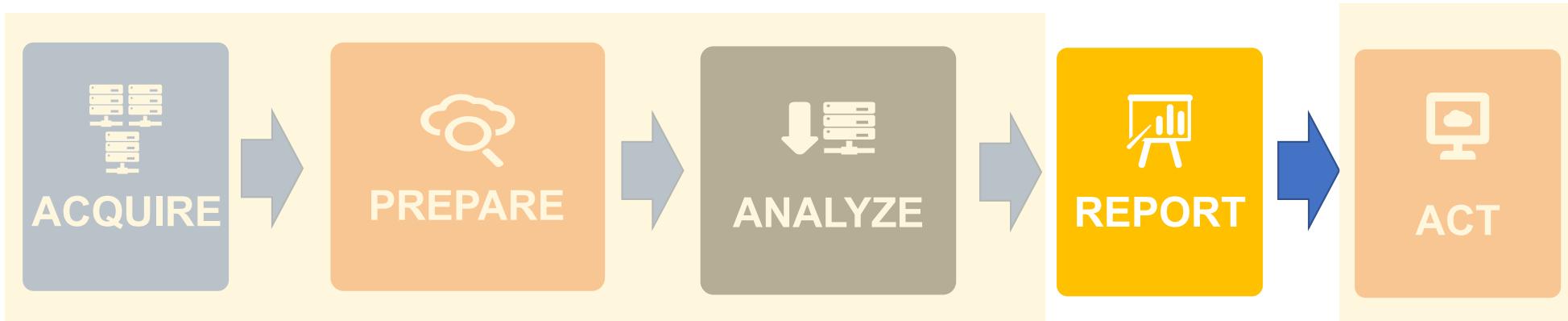
# Step 4: Reporting Insights

Dr. Ilkay Altintas and Dr. Leo Porter

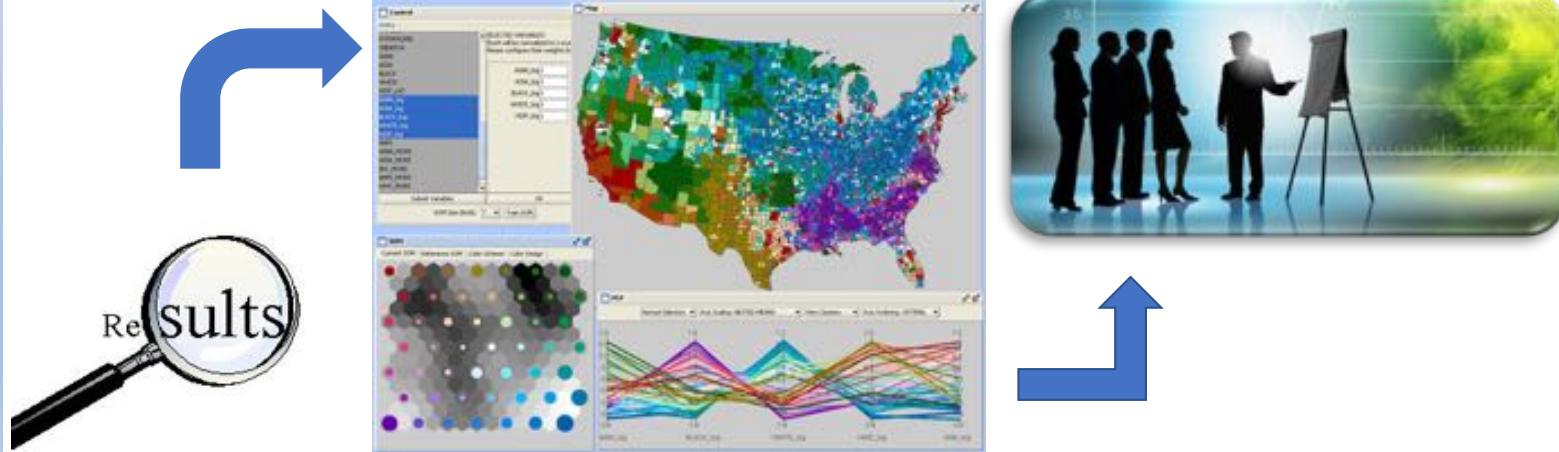
**Twitter:** #UCSDpython4DS

After this video you will be able to..

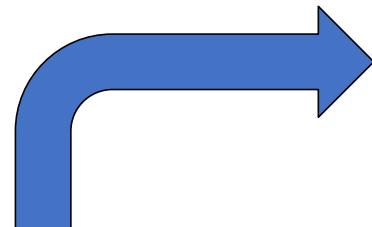
- Determine what to present in reporting your findings
- Identify techniques to communicate your results



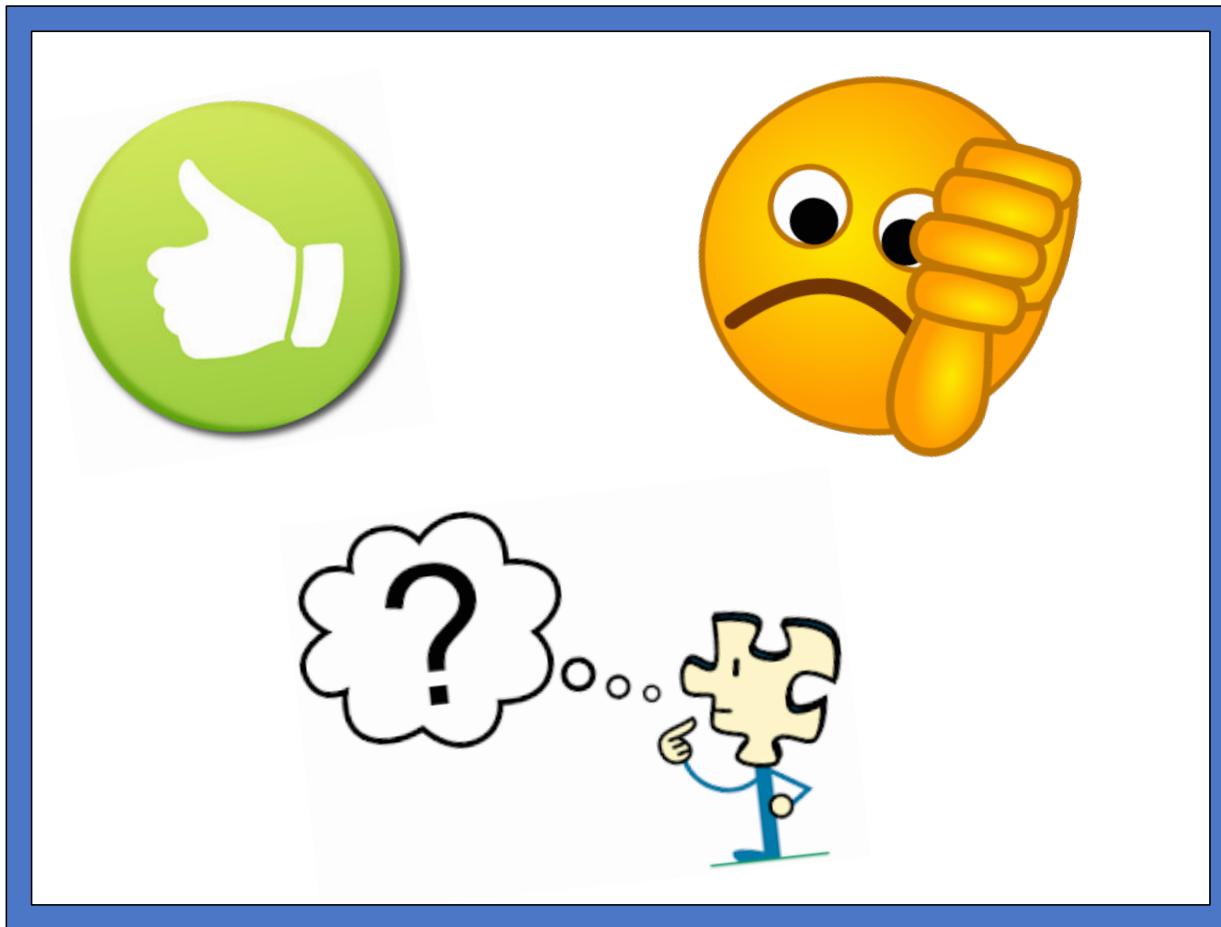
## Step 4: Communicate Results



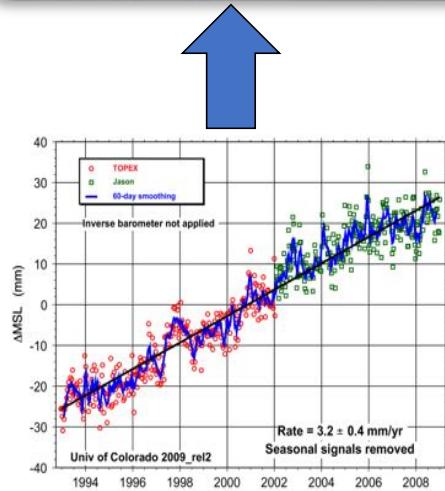
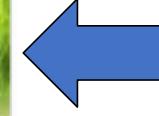
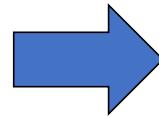
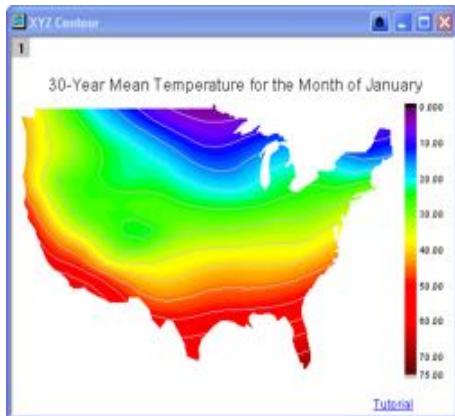
# What to Present



# What to Present



# How to Present



# Visualization Tools



**Timeline<sup>JS</sup>**

Beautifully crafted timelines that are easy and intuitive to use.

(JavaScript)

leaflet  
(JavaScript)

# Present



# with



# using

tableau

Google  
Developers  
CHARTS

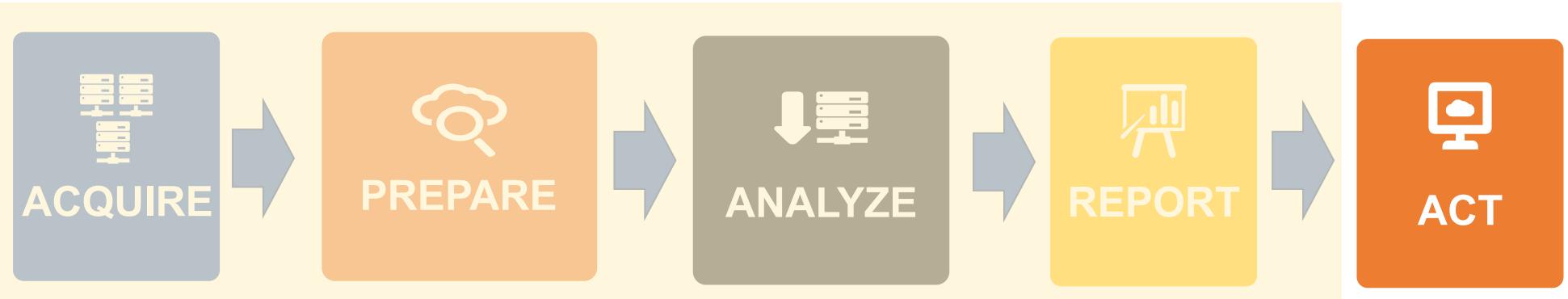
# Step 5: Turning Insights into Action

Dr. Ilkay Altintas and Dr. Leo Porter

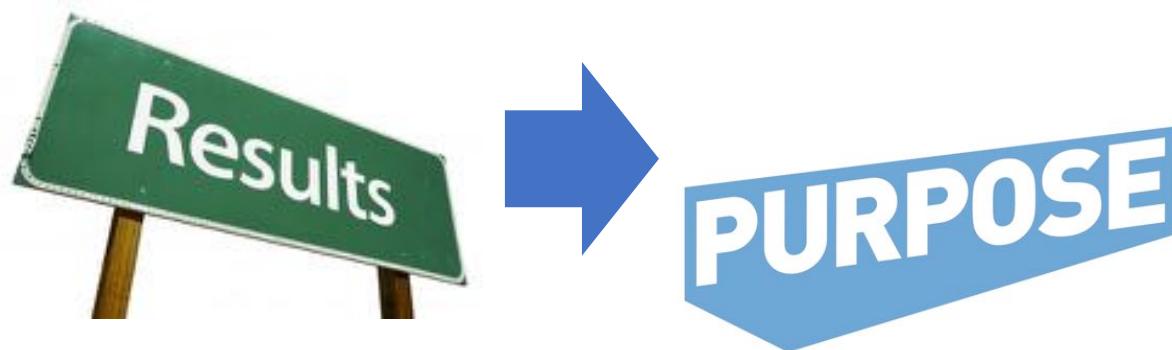
**Twitter:** #UCSDpython4DS

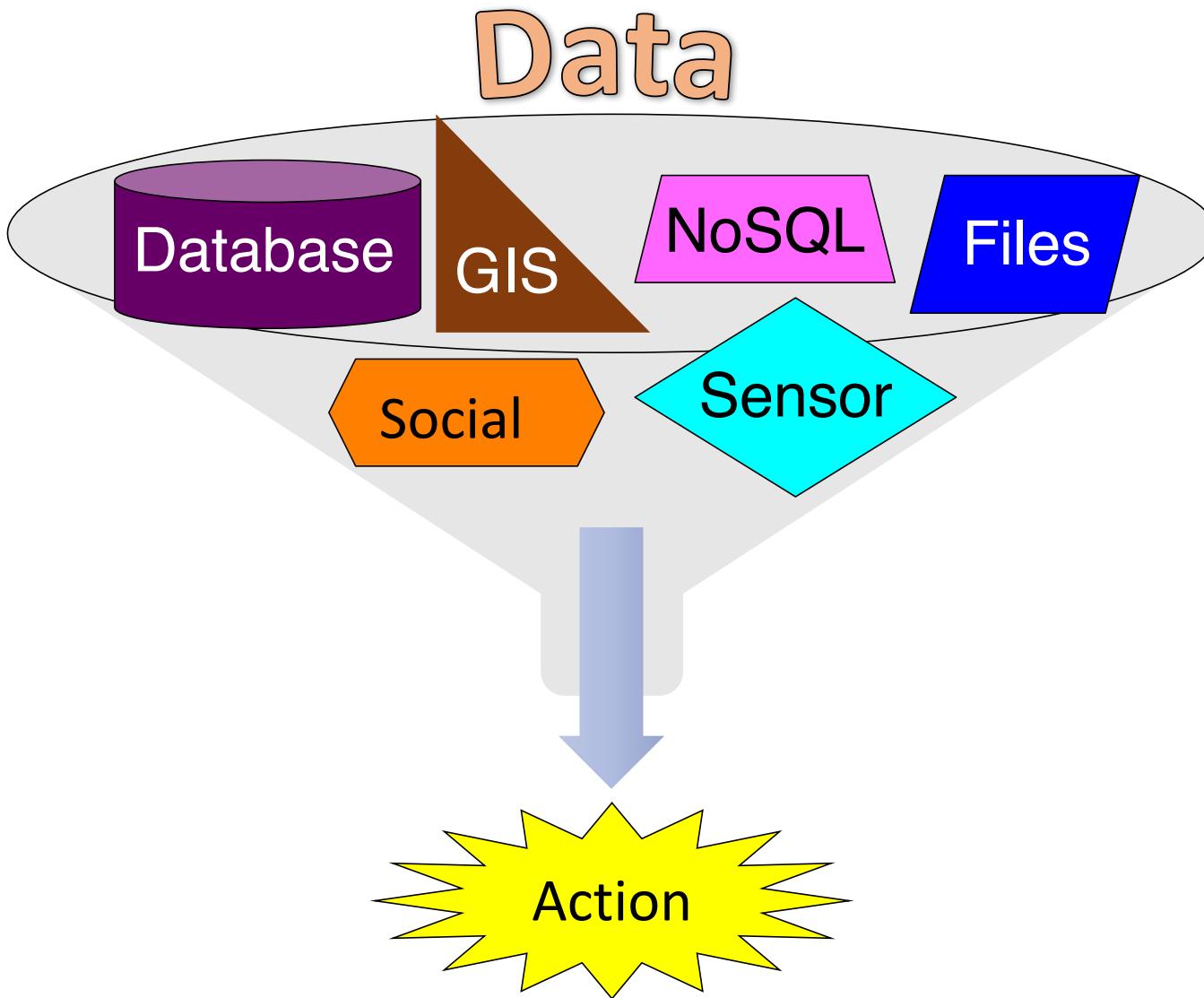
After this video you will be able to..

- Explain what turning insights into action means
- Connect your results with your business question

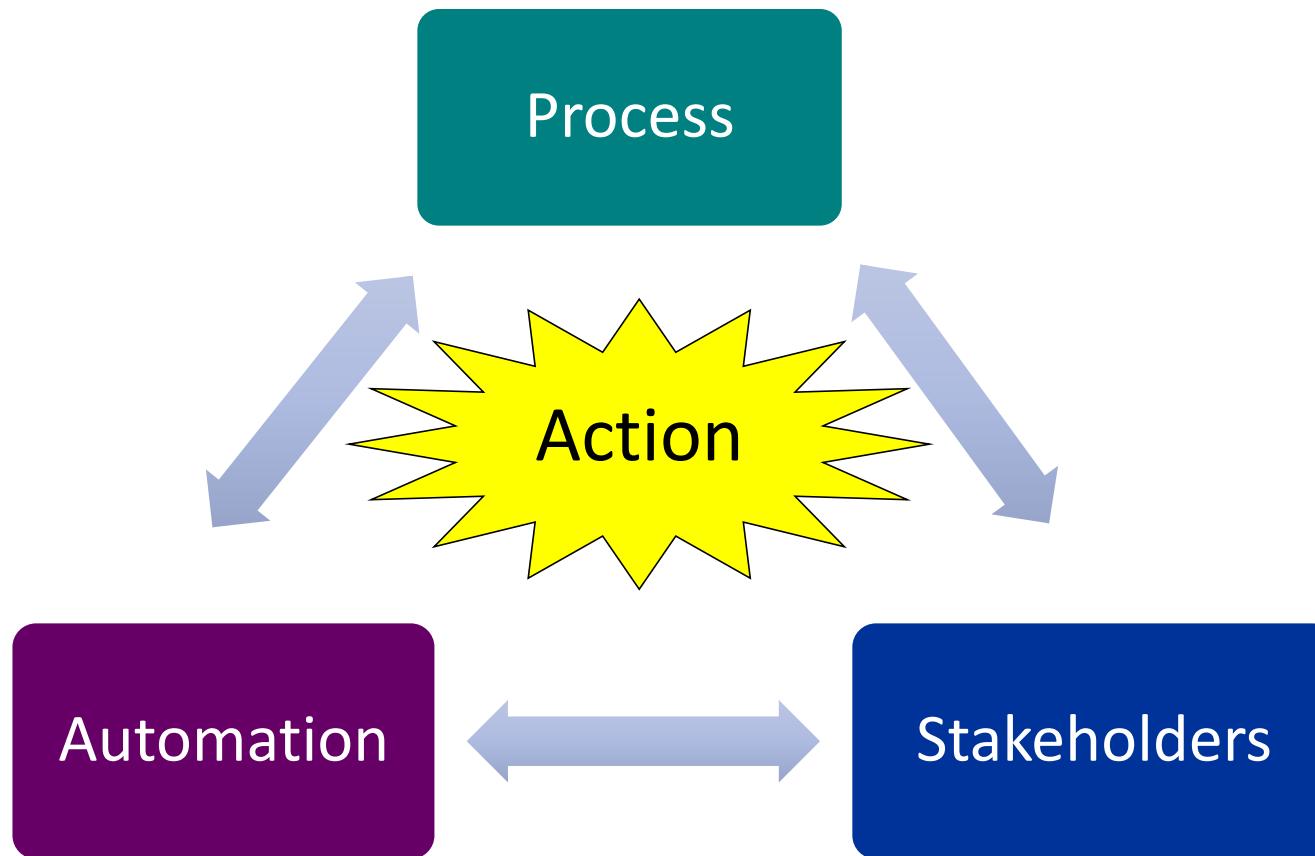


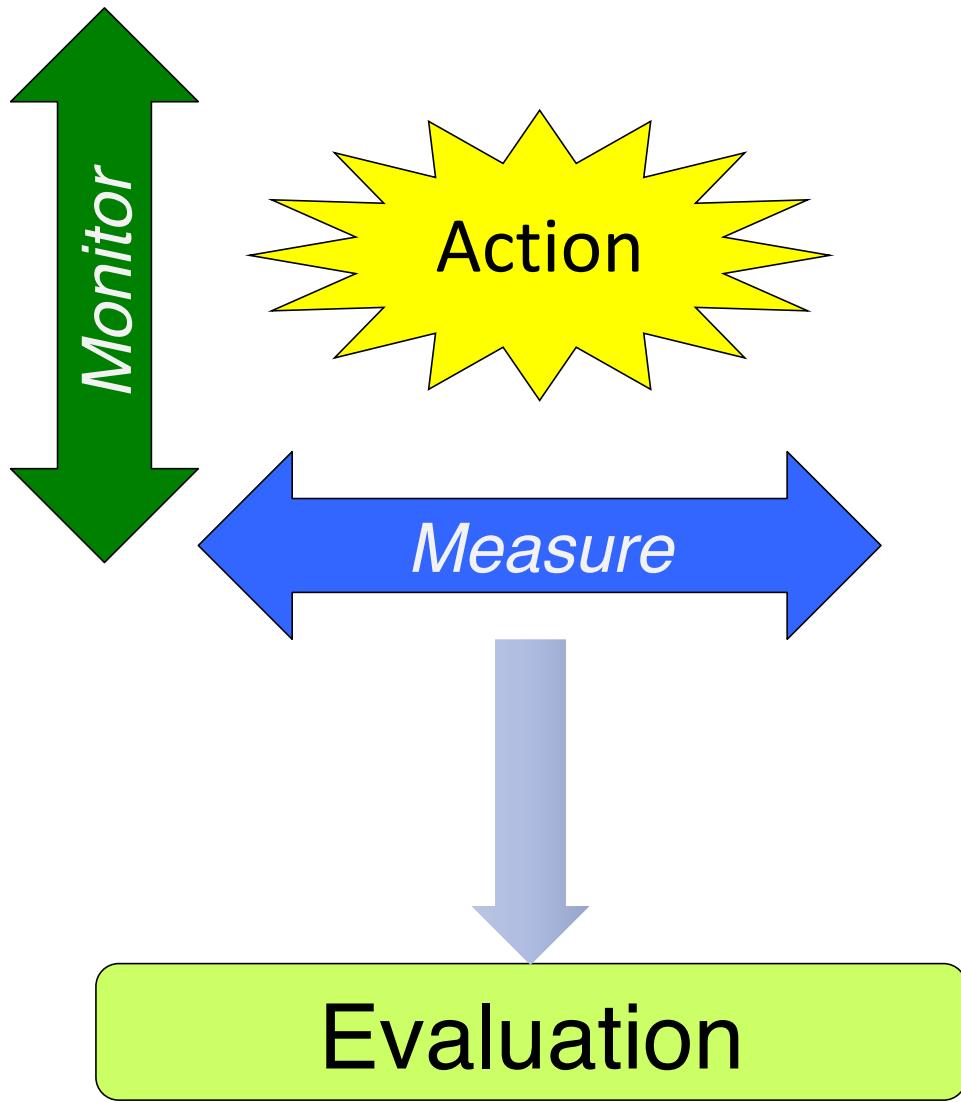
## Step 5: Apply Results



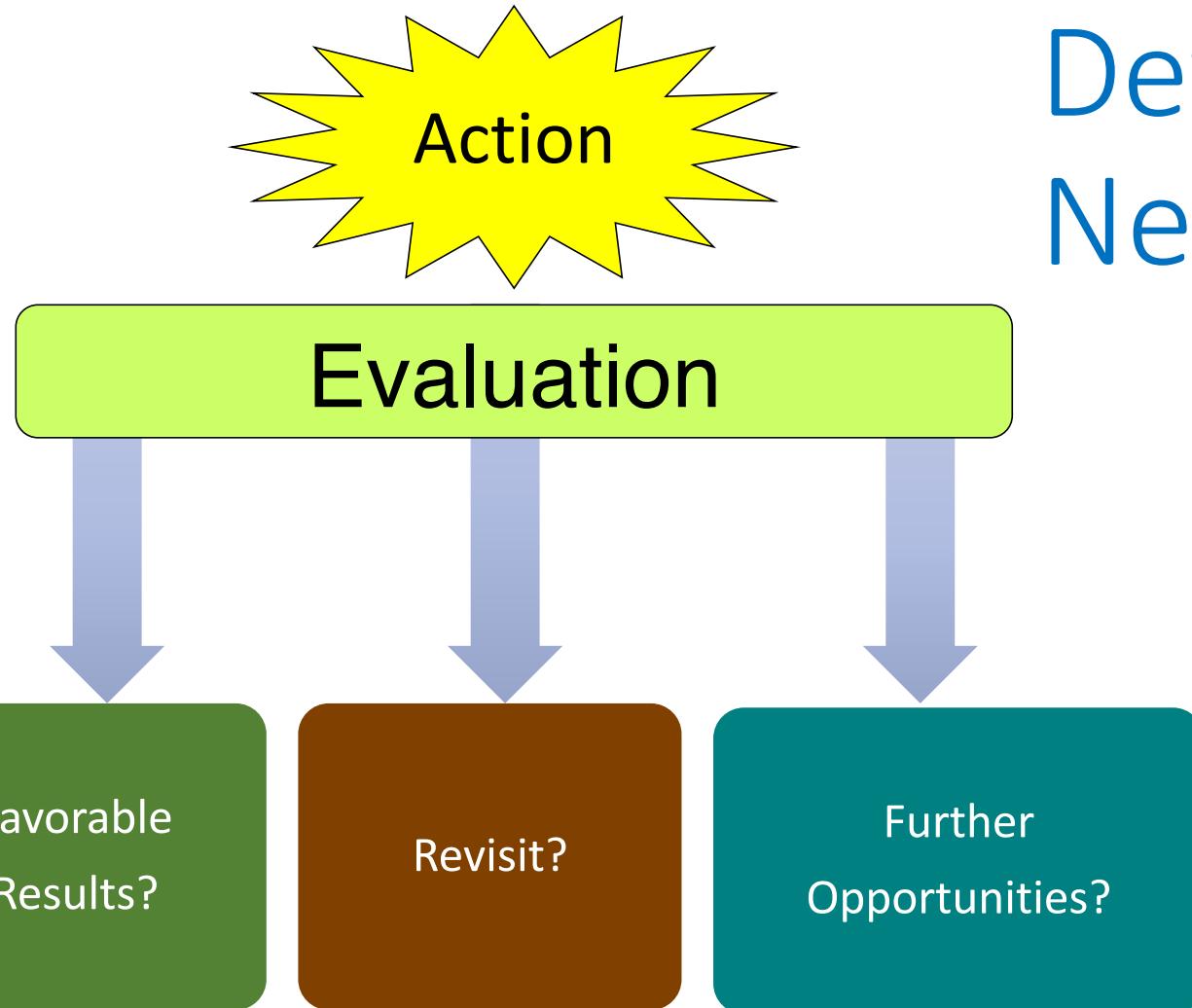


# Implementation





# Assess Impact



Determine  
Next Steps

