# Cross Industry Standard Process for Data Mining (CRISP-DM)

by University of California San Diego

# CRISP-DM

## CRoss Industry Standard Process for Data Mining


by University of California San Diego

# CRISP-DM Phases

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

by University of California San Diego

# Phase 1 – Business Understanding

- Define problem or opportunity
- Assess situation
- Formulate goals

by University of California San Diego

# Phase 2 – Data Understanding

- Data acquisition
- Data exploration

by University of California San Diego

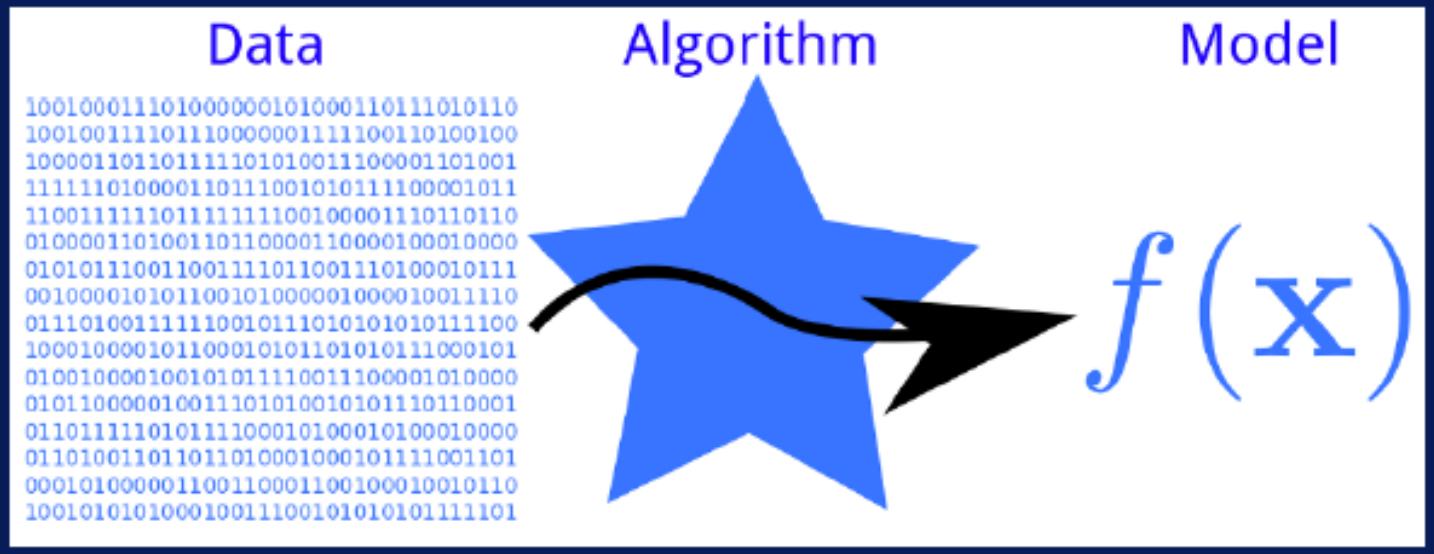
# Phase 3 – Data Preparation

- Prepare data for modeling
- Address quality issues, select features to use, process data for modeling


by University of California San Diego

# Phase 4 – Modeling

- Determine type of problem
- Select modeling technique(s) to use
- Build model




by University of California San Diego

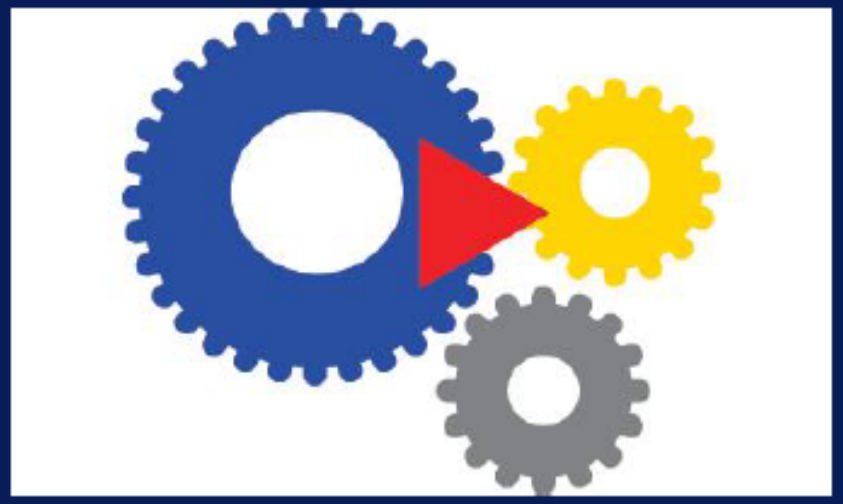
# Phase 5 – Evaluation

- Assess model performance
- Evaluate model results with respect to success criteria

by University of California San Diego

# Phase 6 – Deployment

- Produce final report
- Deploy model
- Monitor model



by University of California San Diego

# Machine Learning Process

by University of California San Diego

ACQUIRE PREPARE ANALYZE REPORT ACT

# Step 1: Acquire Data

- Identify data sources
- Collect data
- Integrate data

by University of California San Diego

ACQUIRE | **PREPARE** | ANALYZE | REPORT | ACT

# Step 2: Prepare Data

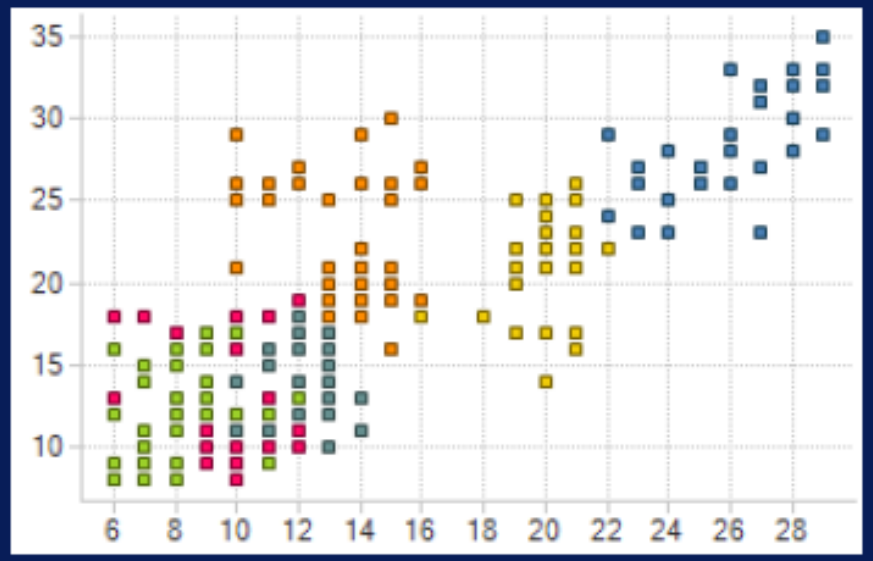
Step 2-A: Explore

Step 2-B: Pre-process

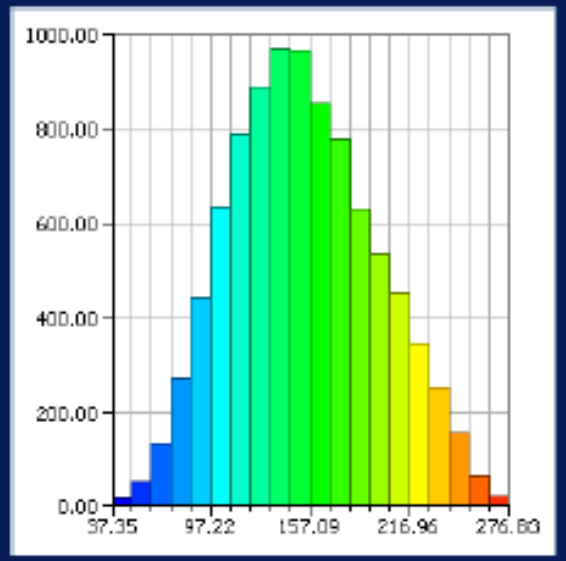
by University of California San Diego

ACQUIRE | **PREPARE** | ANALYZE | REPORT | ACT

# Step 2-A: Explore Data




Understand nature of data

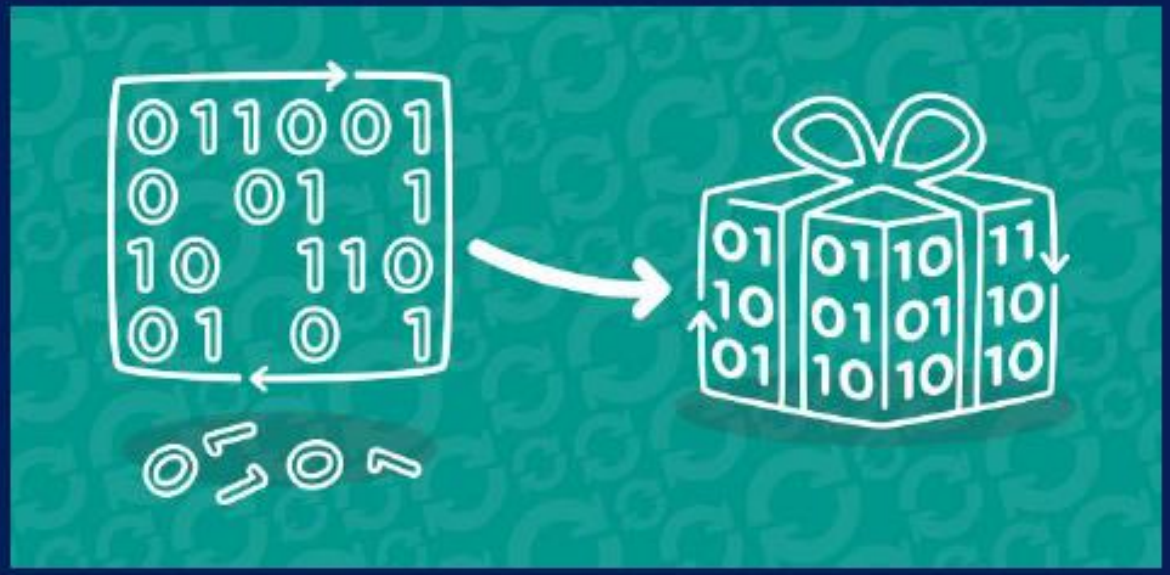



Preliminary analysis

by University of California San Diego

ACQUIRE | **PREPARE** | ANALYZE | REPORT | ACT

# Step 2-B: Pre-process Data



Clean | Select | Transform

by University of California San Diego

ACQUIRE PREPARE **ANALYZE** REPORT ACT

# Step 3: Analyze Data


Select analytical techniques

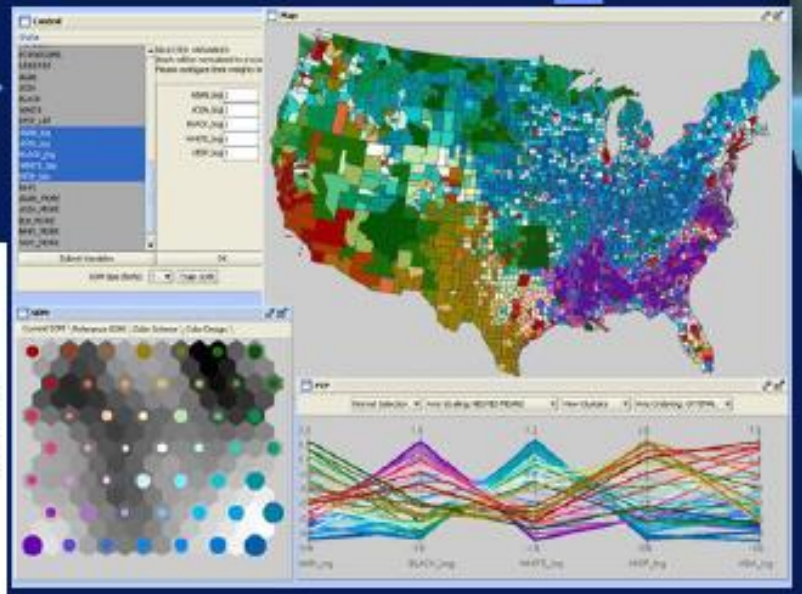
Build models

Assess results

by University of California San Diego

ACQUIRE PREPARE ANALYZE **REPORT** ACT

# Step 4: Communicate Results

by University of California San Diego

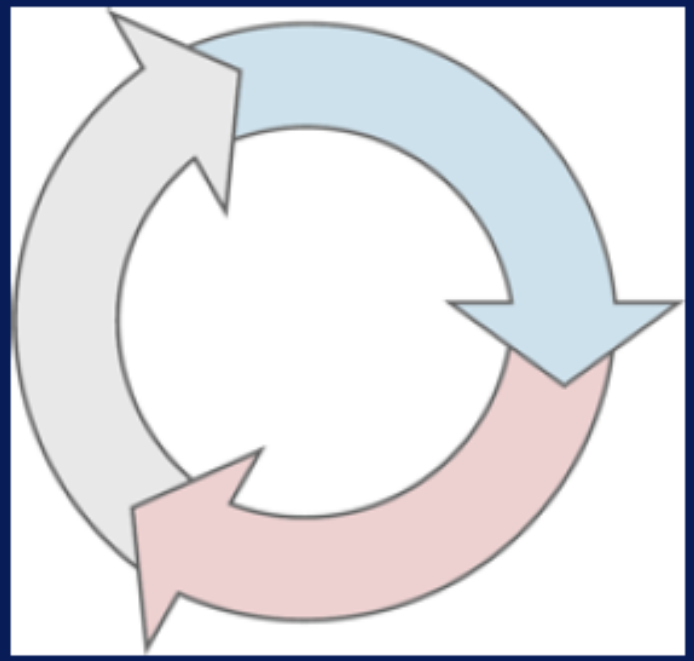
ACQUIRE PREPARE ANALYZE REPORT **ACT**

# Step 5: Apply Results



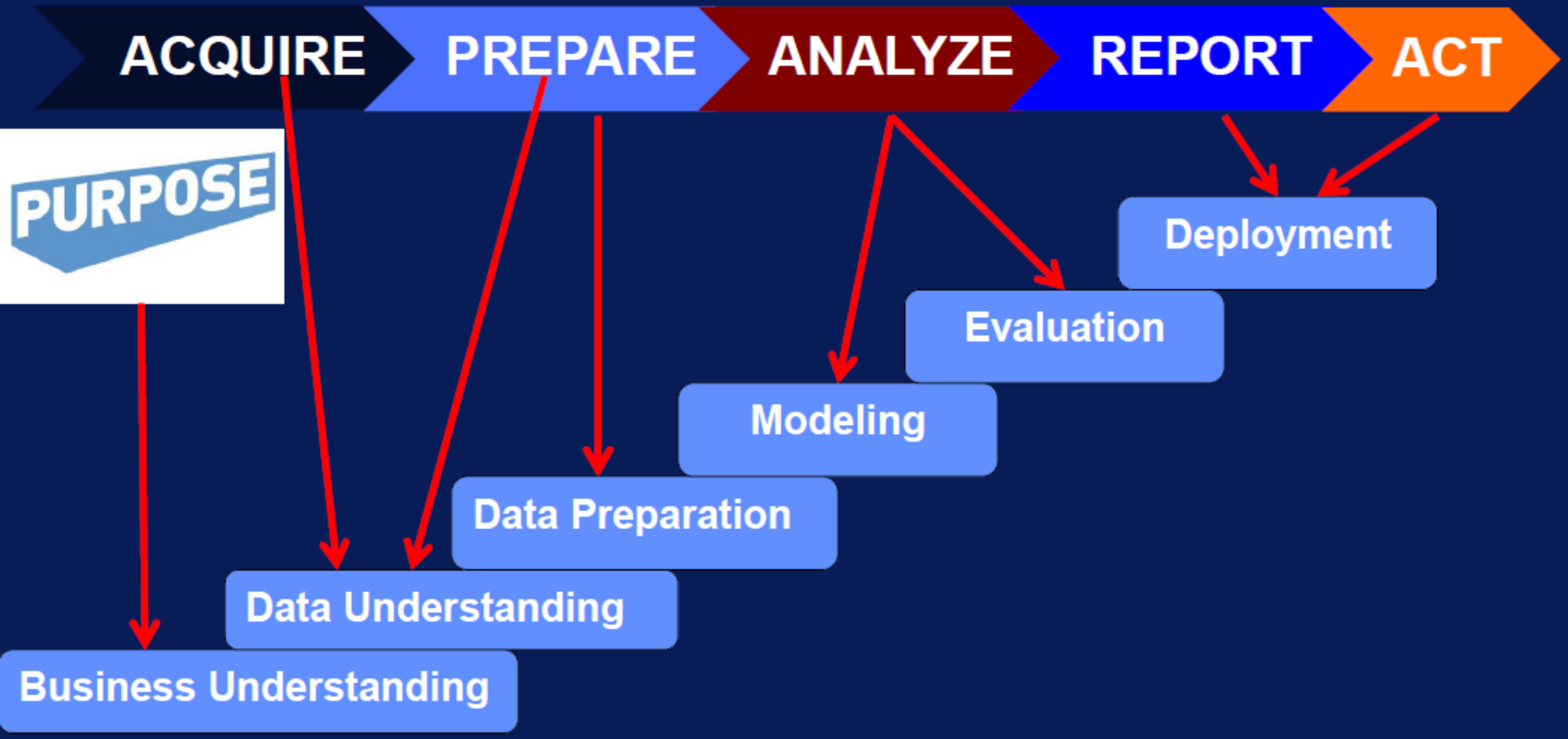
by University of California San Diego

ACQUIRE → PREPARE → ANALYZE → REPORT → ACT



# Iterative process

by University of California San Diego




by University of California San Diego

# Goals and Activities in the Machine Learning Process

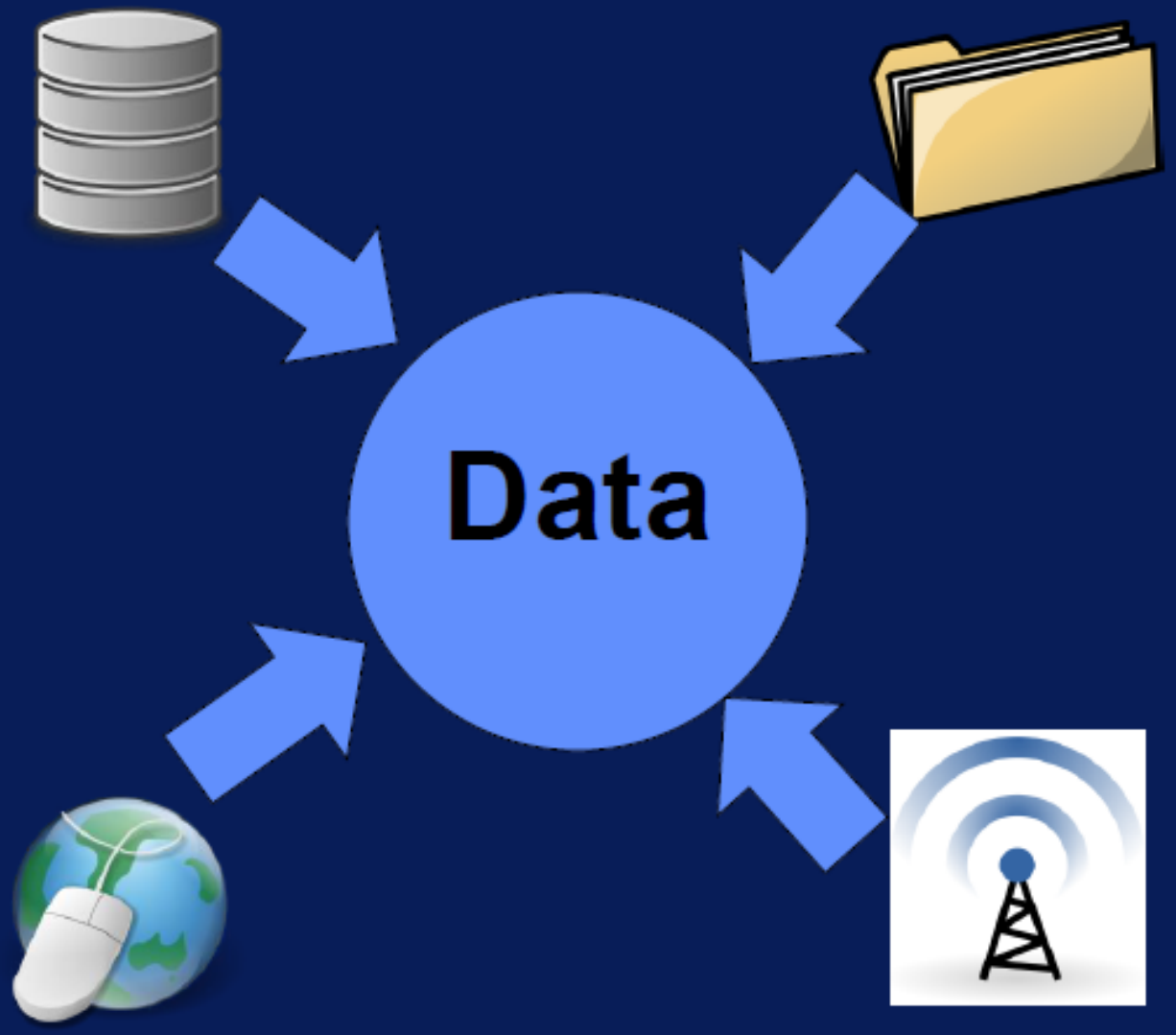
by University of California San Diego

ACQUIRE PREPARE ANALYZE REPORT ACT

Goal: Identify and obtain all data related to problem

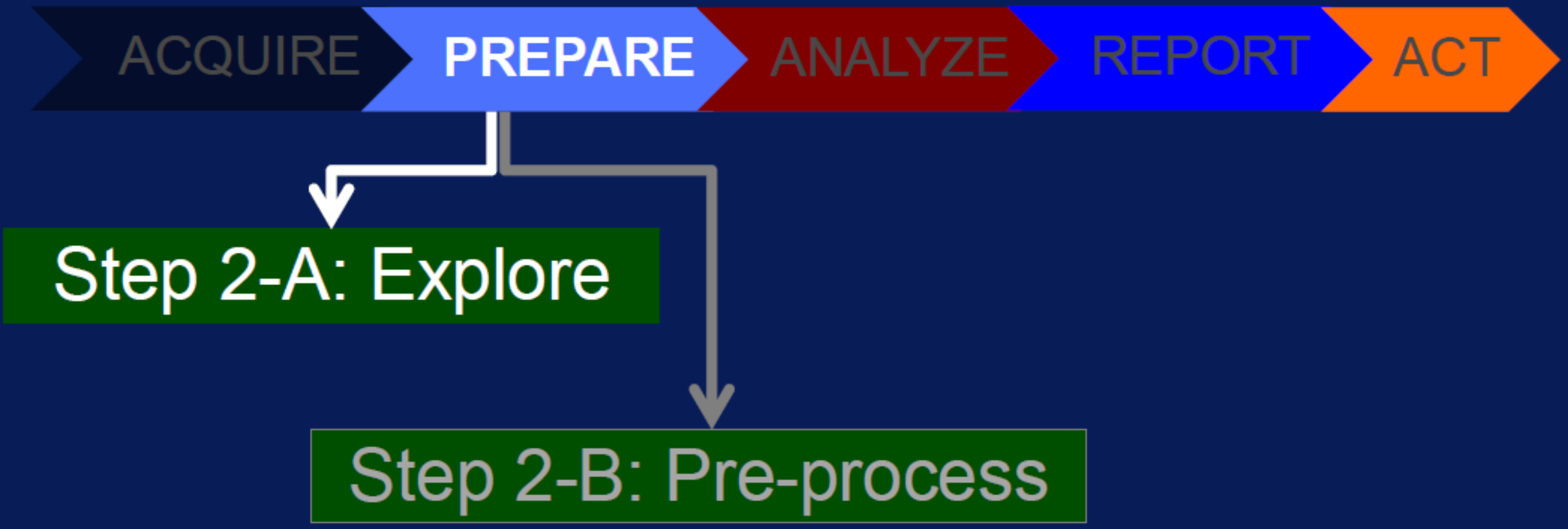
by University of California San Diego

# Acquire Data




Identify data sources
Collect data
Integrate data

by University of California San Diego




by University of California San Diego

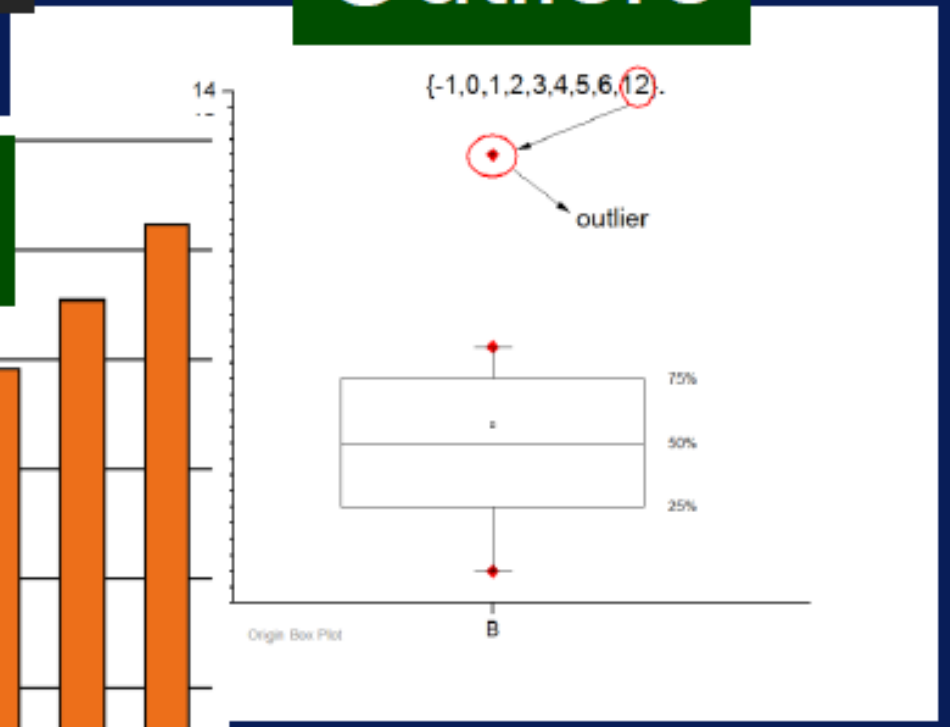
# Why Explore?

Goal: Understand your data

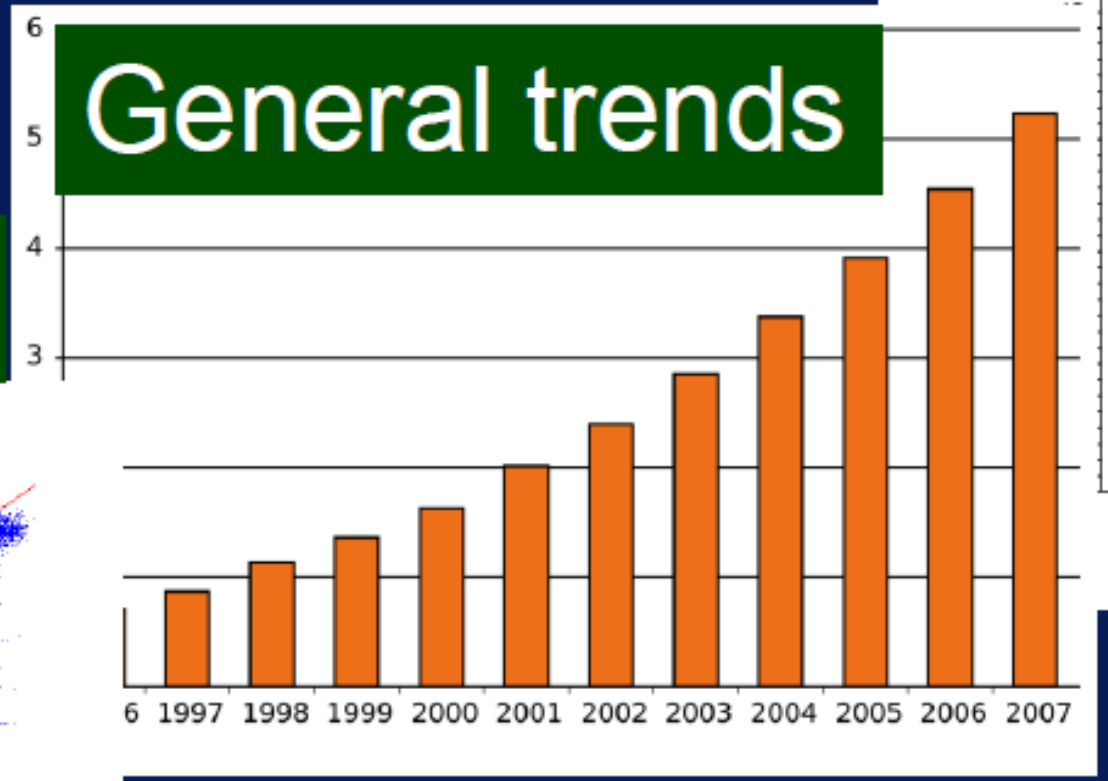
by University of California San Diego

# Why Explore?

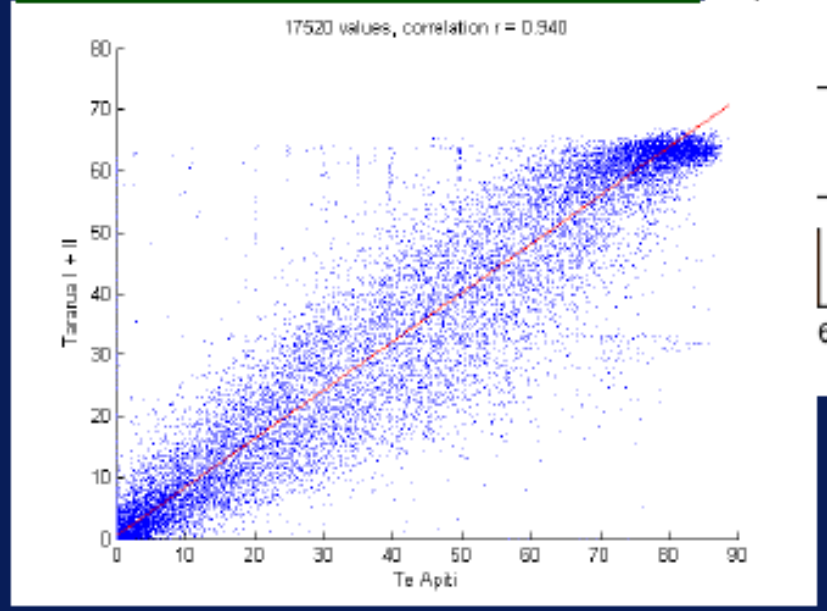
## Outliers




## General trends

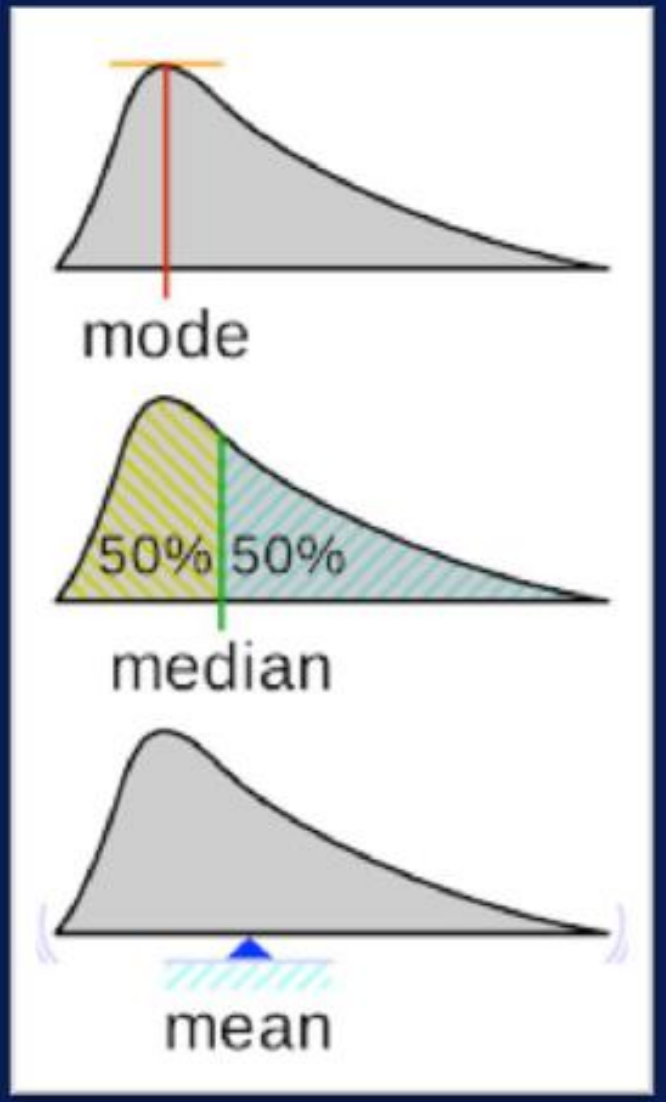


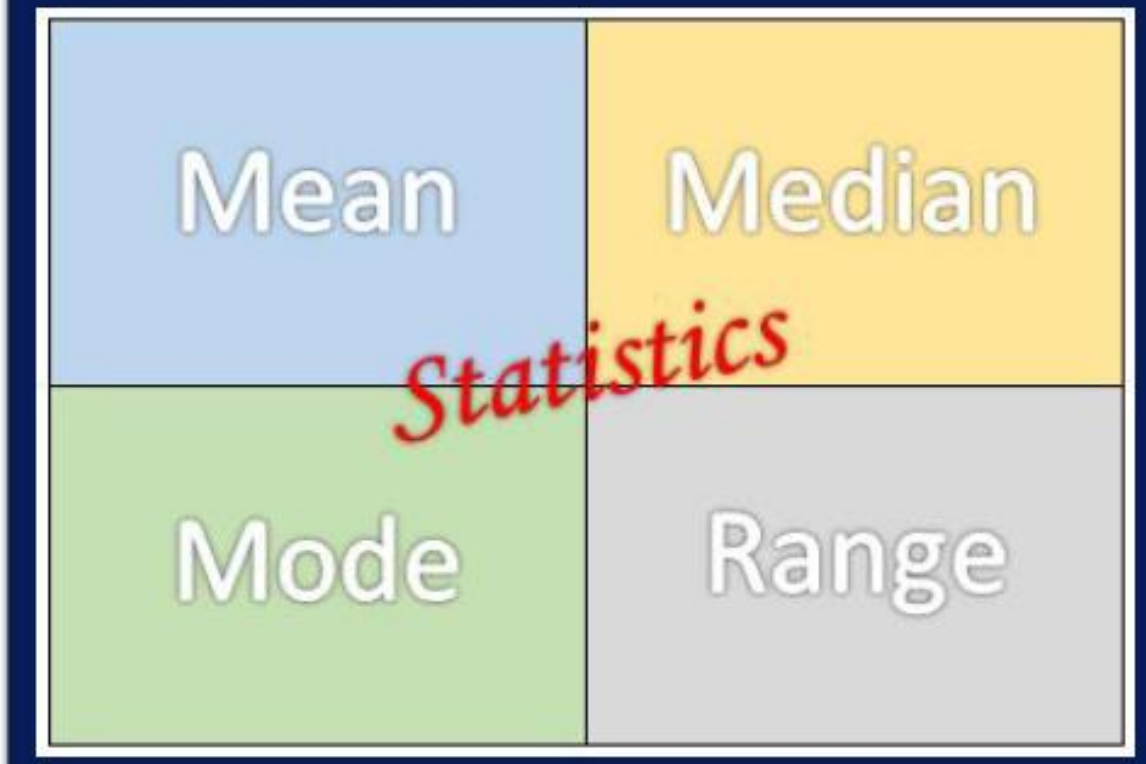

## Correlations




by University of California San Diego

# Describe Your Data

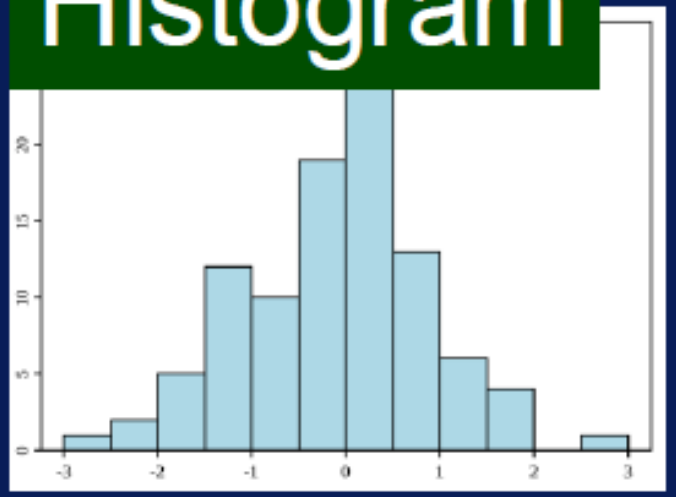


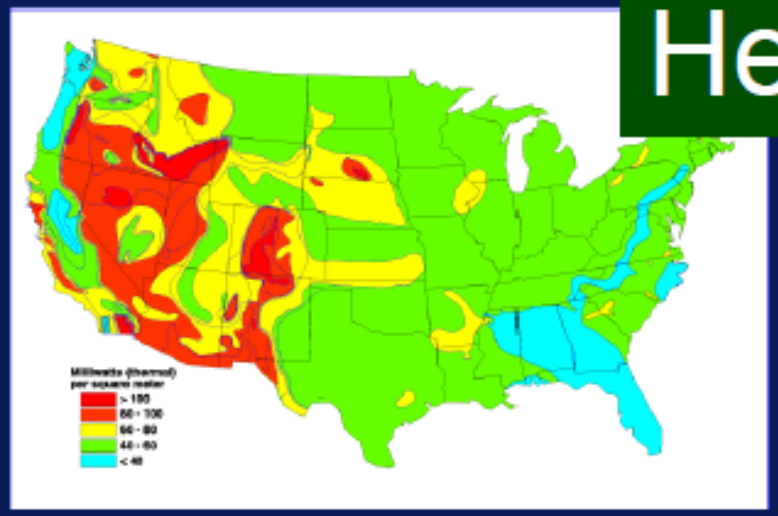

by University of California San Diego

# Visualize Your Data

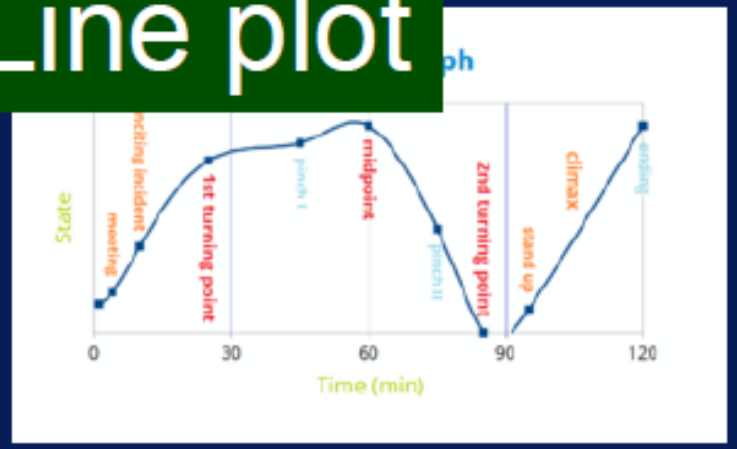
## Histogram



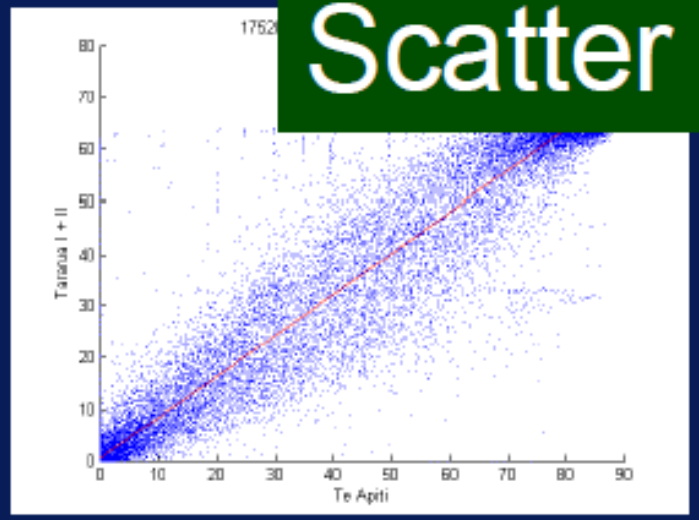

## Heat map



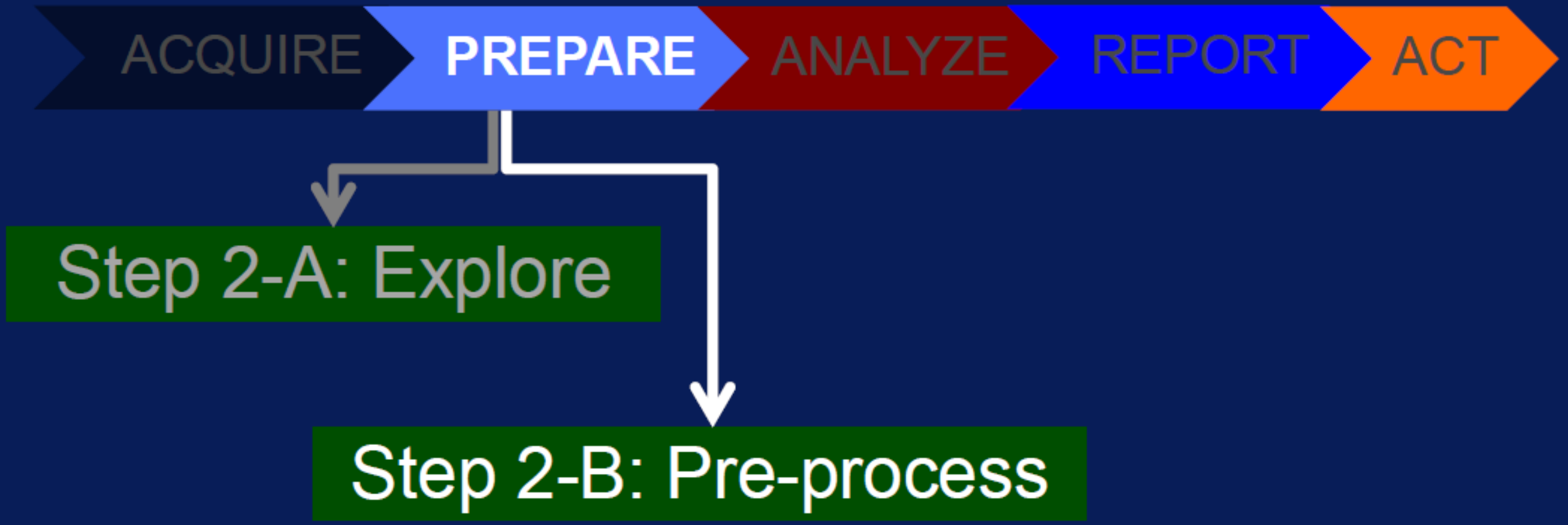
## Line plot




## Scatter plot




by University of California San Diego




by University of California San Diego

ACQUIRE | **PREPARE** | ANALYZE | REPORT | ACT

Step 2-A: Explore

Goal: Create data for analysis

Step 2-B: Pre-process

Clean

Select

Transform

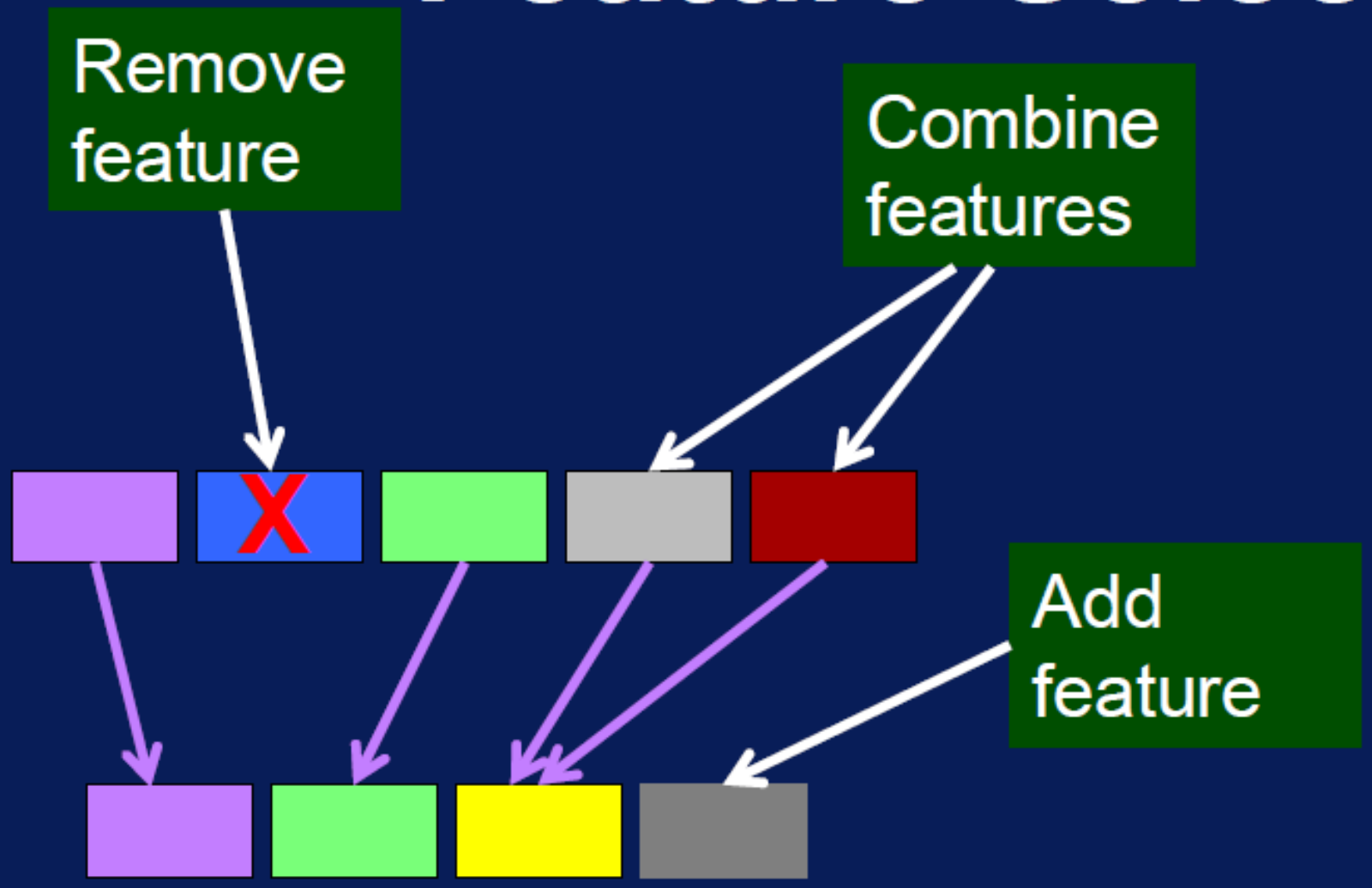
by University of California San Diego

# Data Cleaning

- Missing values
- Duplicate data
- Inconsistent data
- Noise
- Outliers

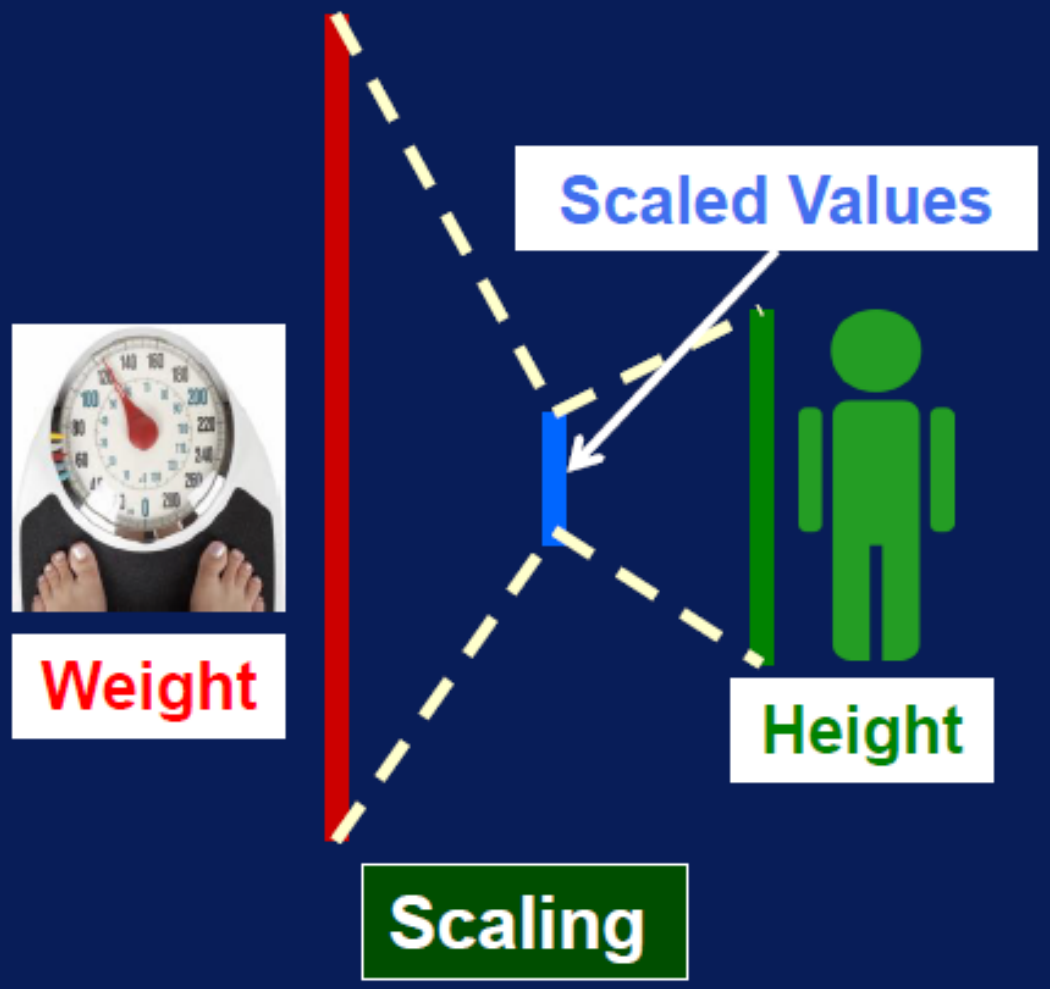

by University of California San Diego

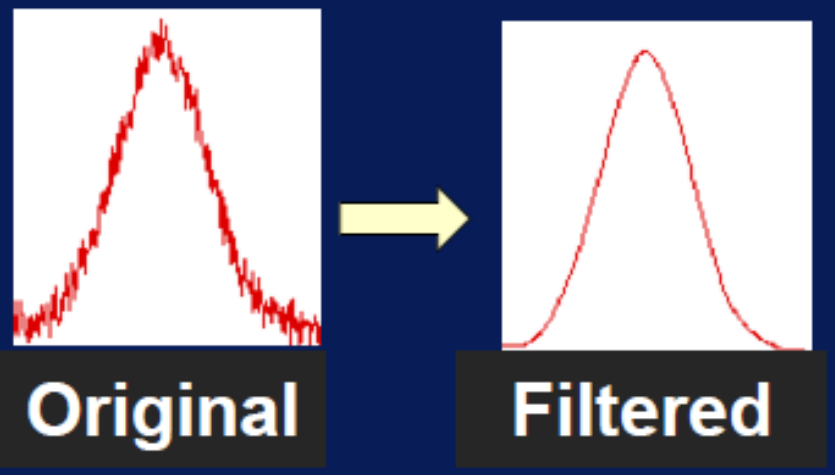
# Feature Selection




by University of California San Diego

# Feature Transformation




3D

2D

Dimensionality Reduction

by University of California San Diego

ACQUIRE PREPARE **ANALYZE** REPORT ACT

Goals:

- Build model
- Evaluate results

by University of California San Diego

# Analyze

Select technique

Build model

Evaluate

Classification
Regression
Cluster Analysis
Association
Analysis


by University of California San Diego

# Analyze

Select technique

↓

Build model

↓

Evaluate results

ACQUIRE → PREPARE → ANALYZE → **REPORT** → ACT

# Goal: Communicate results and recommend actions

# Present Report


with

using

tableau

Google Developers CHARTS

by University of California San Diego

ACQUIRE → PREPARE → ANALYZE → REPORT → **ACT**

Goal: Determine actions based on insights

by University of California San Diego

# Act

Determine action

↓

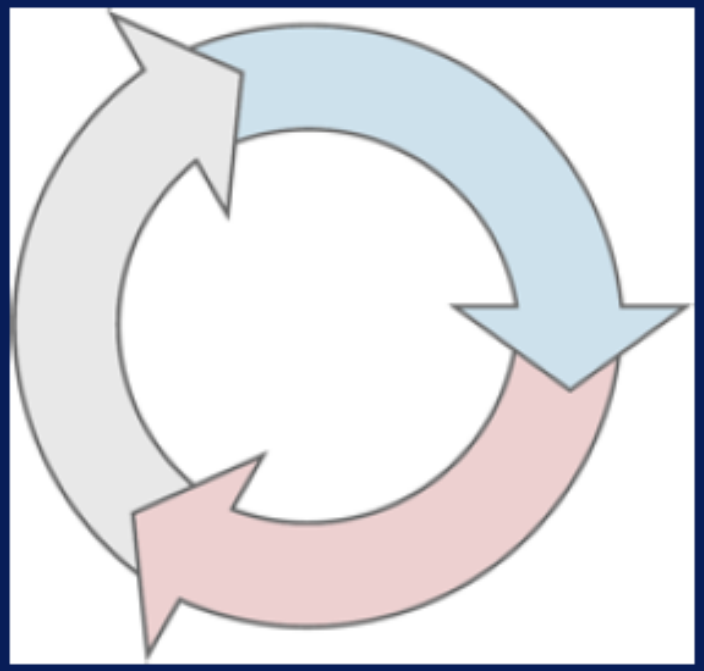
Implement

↓

Assess impact

by University of California San Diego

ACQUIRE → PREPARE → ANALYZE → REPORT → ACT



# Iterative process

by University of California San Diego