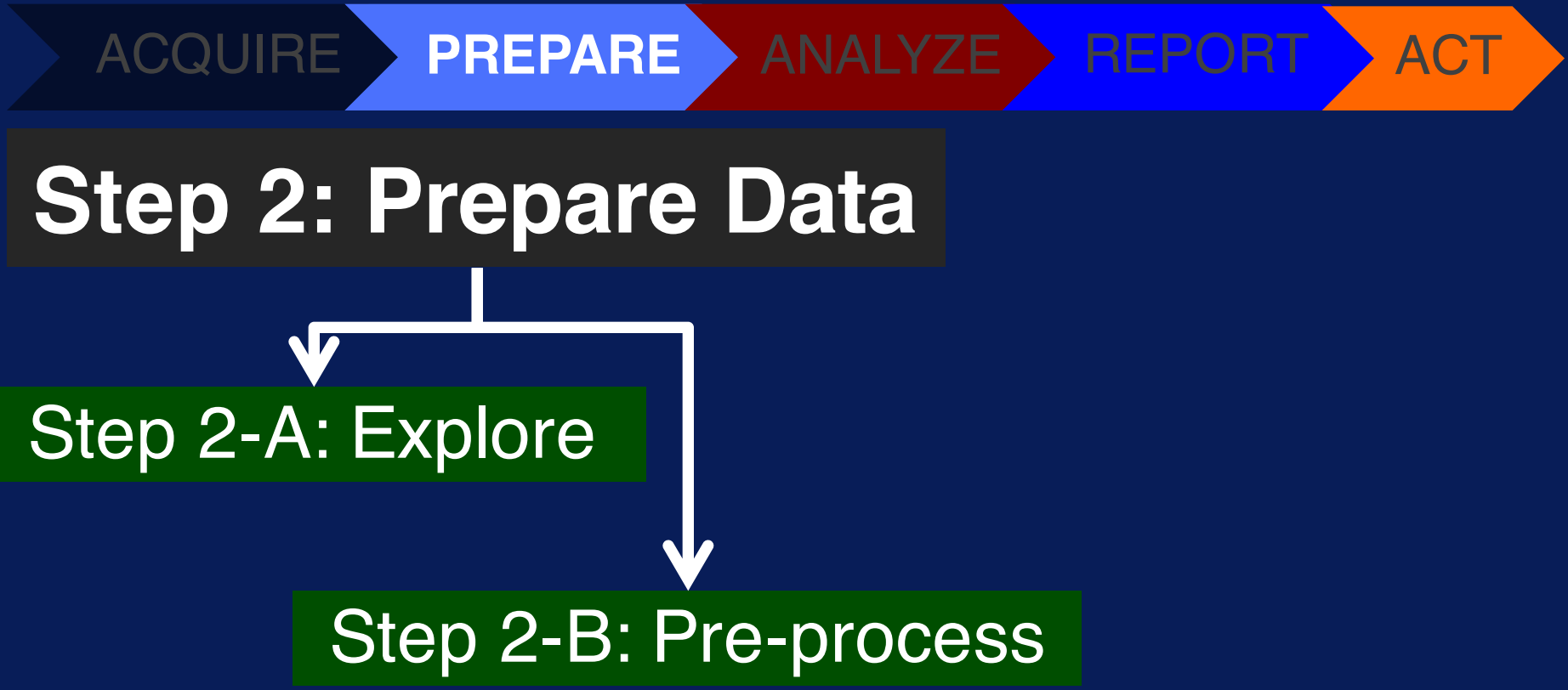
Steps in the Data Science Process


ACQUIRE
PREPARE
ANALYZE
REPORT
ACT

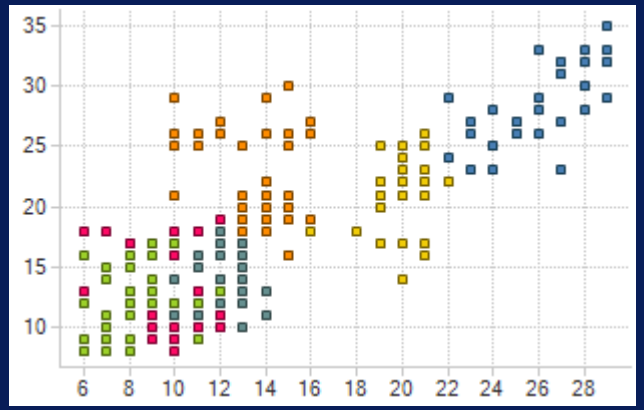
ACQUIRE PREPARE ANALYZE REPORT ACT

Step 1: Acquire Data

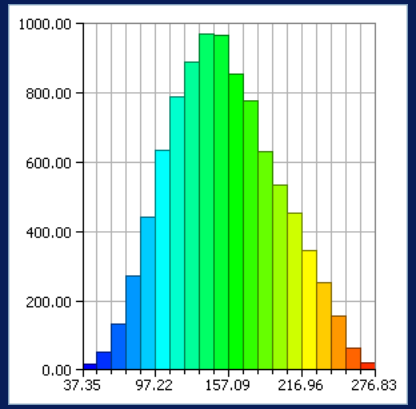
- Identify data sets
- Retrieve data
- Query data


ACQUIRE
PREPARE
ANALYZE
REPORT
ACT
Step 2: Prepare Data
Step 2-A: Explore
Step 2-B: Pre-process

Step 2-A: Explore Data



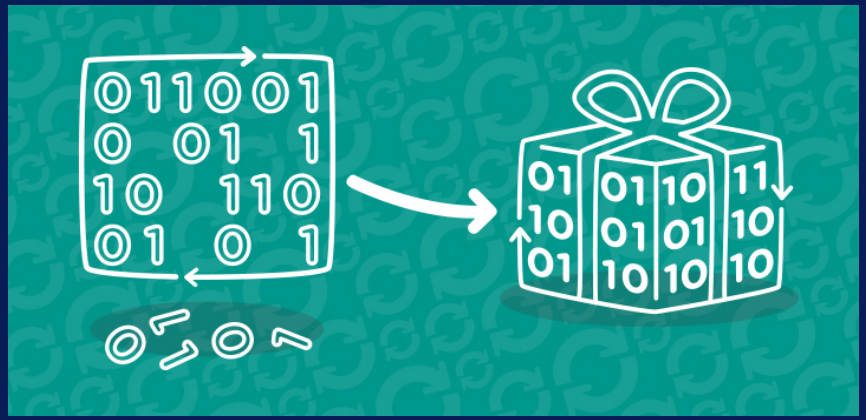
Understand nature of data



Preliminary analysis

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

Step 2-B: Pre-process Data



Clean | Integrate | Package

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

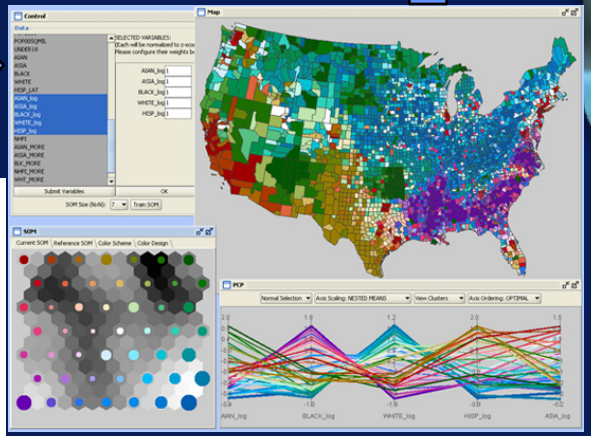
Step 3: Analyze Data

Select analytical techniques

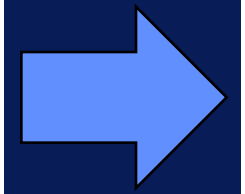
Build models

ACQUIRE PREPARE ANALYZE **REPORT** ACT

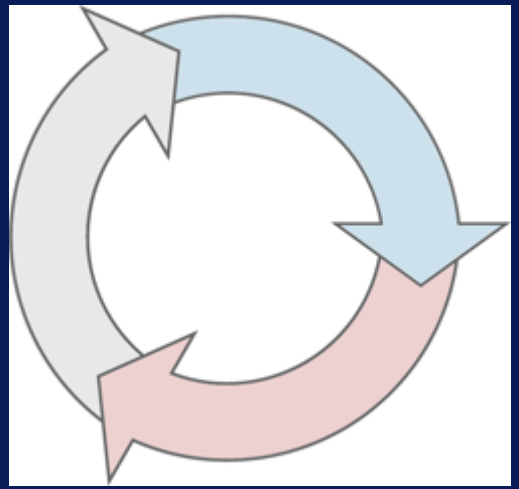
Step 4: Communicate Results



Step 5: Apply Results



ACQUIRE PREPARE ANALYZE REPORT ACT



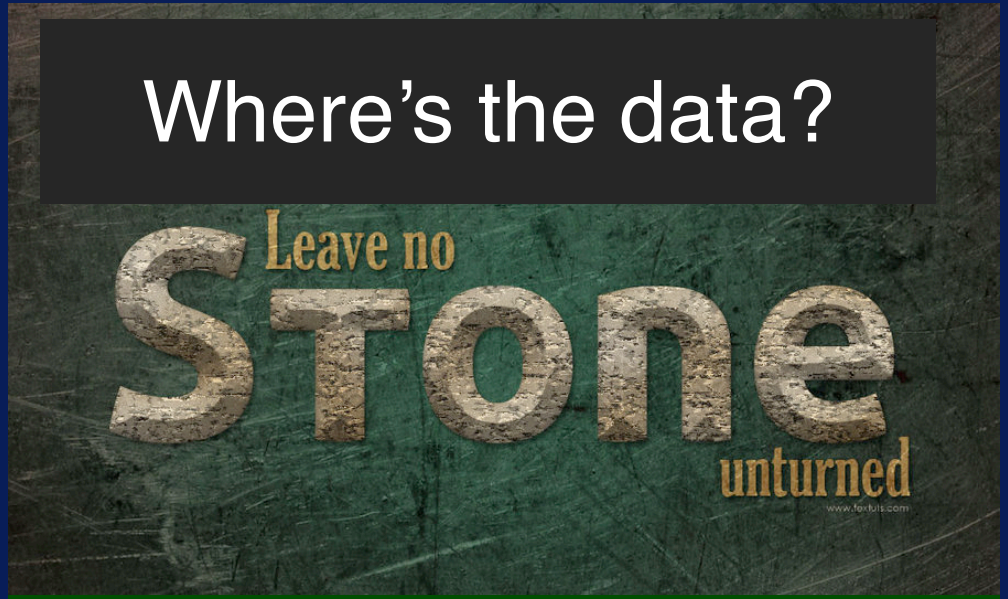
Iterative process

Step 1:

Acquiring Data


Big Data Engineering
Computational Big Data Science
ACQUIRE
PREPARE
ANALYZE
REPORT
ACT

Where’s the data?



Identify suitable data

Acquire all available data

Data comes from many places…







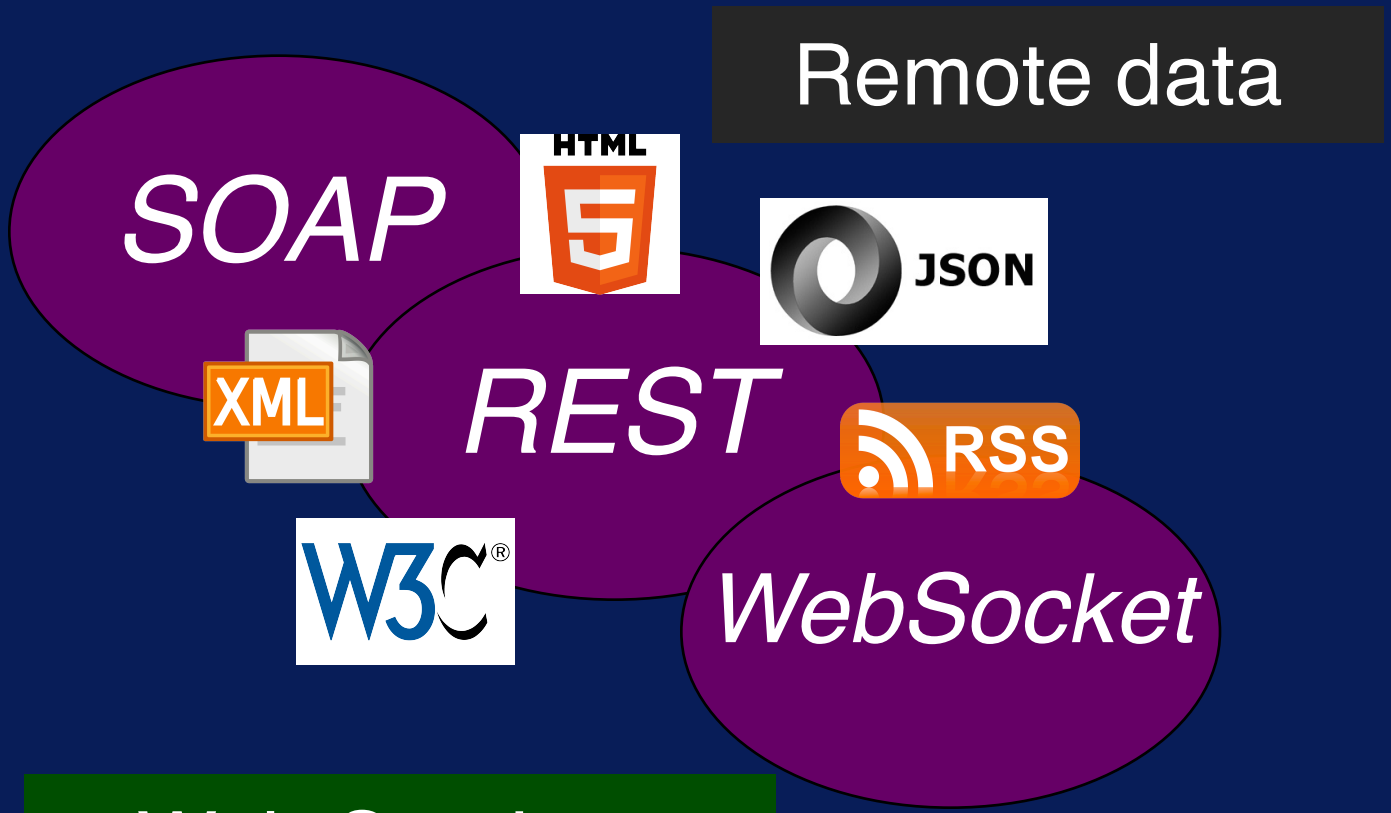
…with many ways to access it

Traditional databases

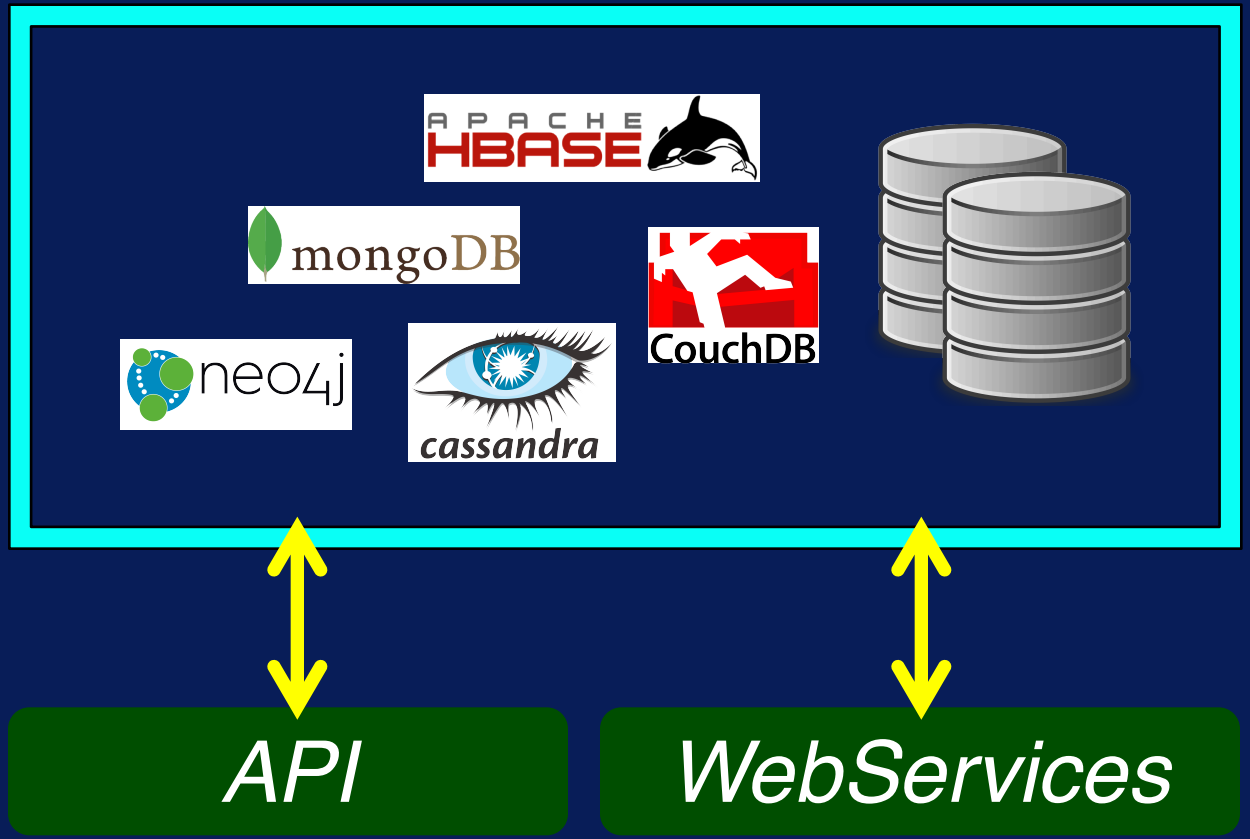
SQL and query browsers

Text files

Scripting languages


Remote data
SOAP
HTML
JSON
XML
REST
RSS
W3C
WebSocket
Web Services

NoSQL storage



Acquiring Data From WIFIRE

Historical weather

SQL

Current weather

WebSocket

Real-time tweets near fires

REST

Traditional databases

SQL and query browsers

Remote data

Web Services

Text files

Scripting languages

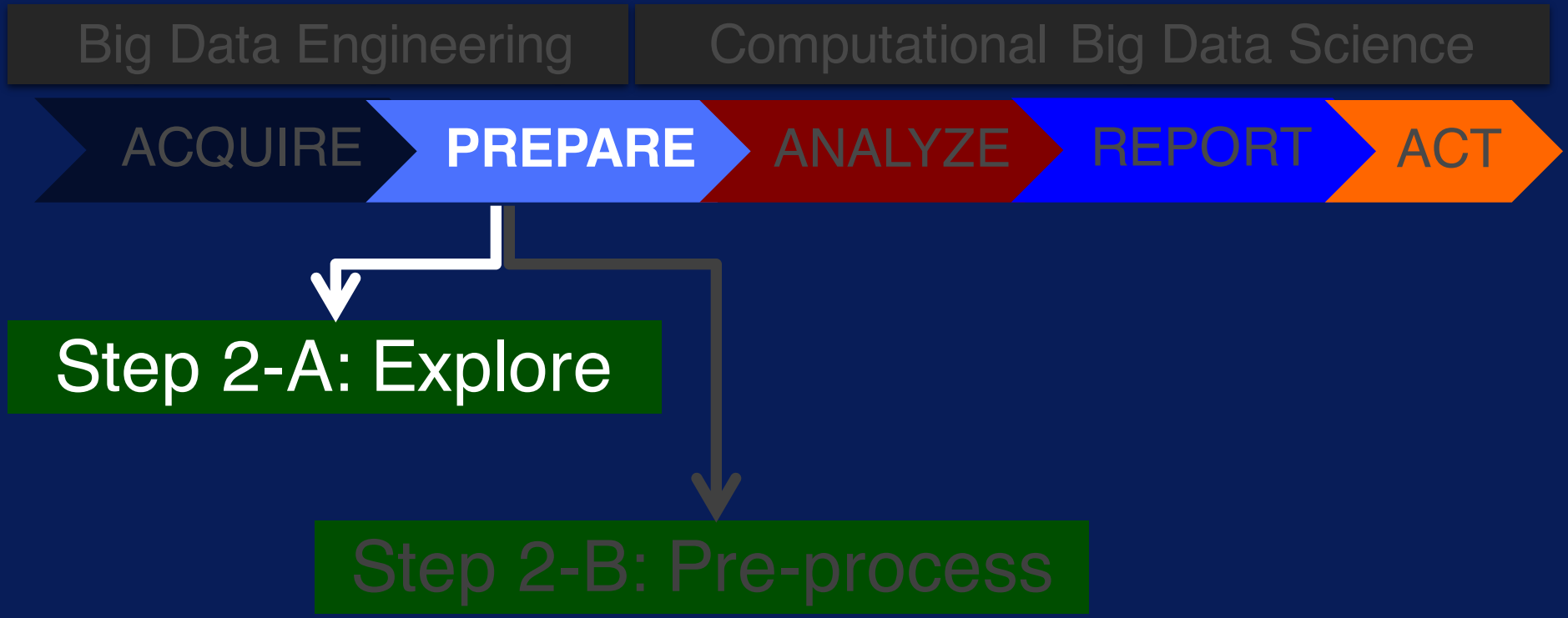
NoSQL storage

Web Services

Programming Interfaces

Step 2-A:

Exploring Data

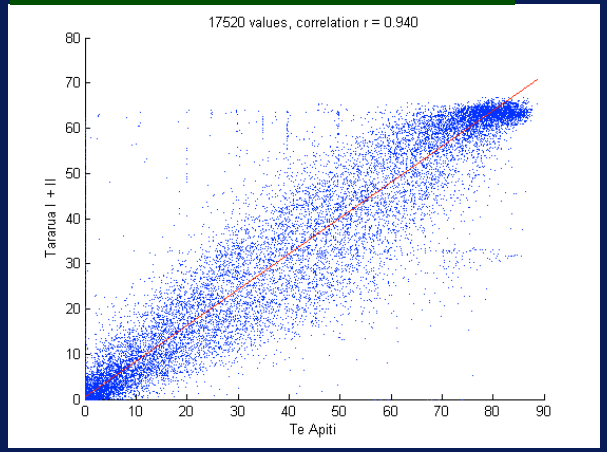

Big Data Engineering
Computational Big Data Science
ACQUIRE
PREPARE
ANALYZE
REPORT
ACT
Step 2-A: Explore
Step 2-B: Pre-process

Why Explore?

Goal: Understand your data

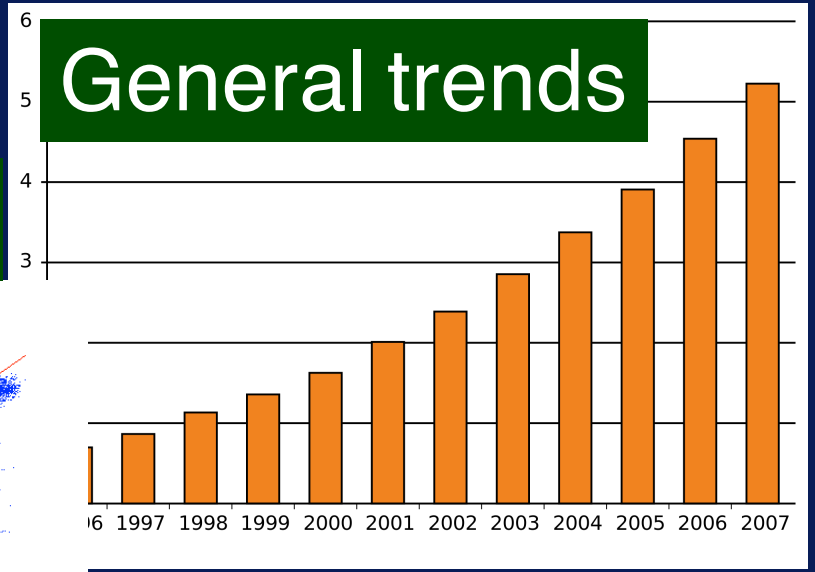
Why Explore?

Correlations

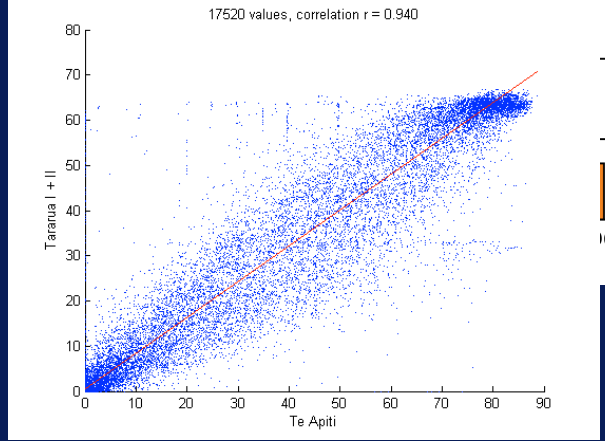


Why Explore?

General trends

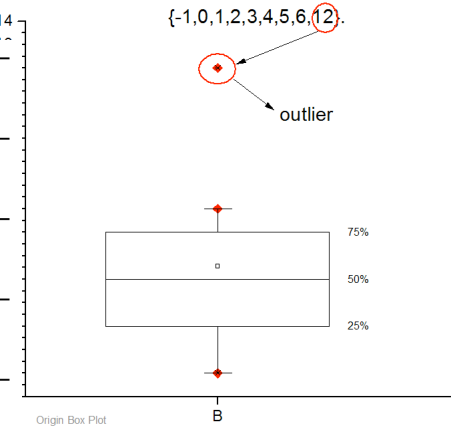


Correlations

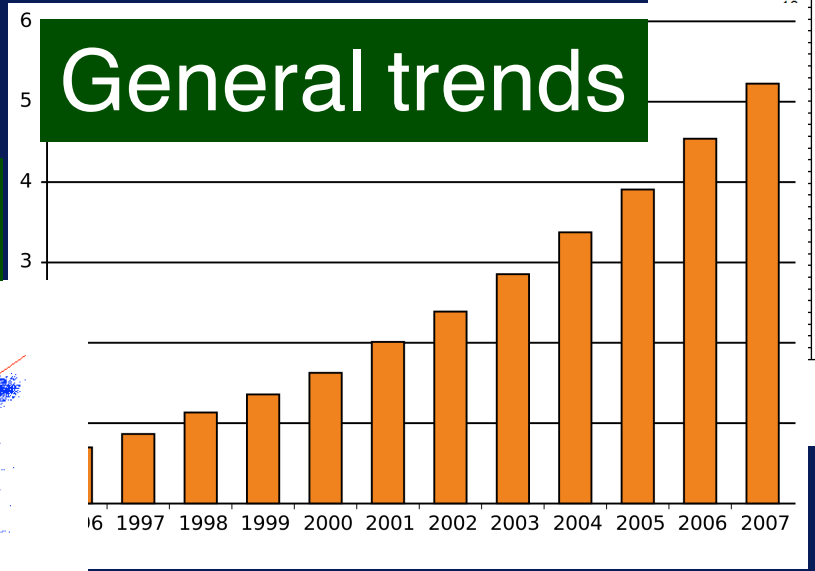


Why Explore?

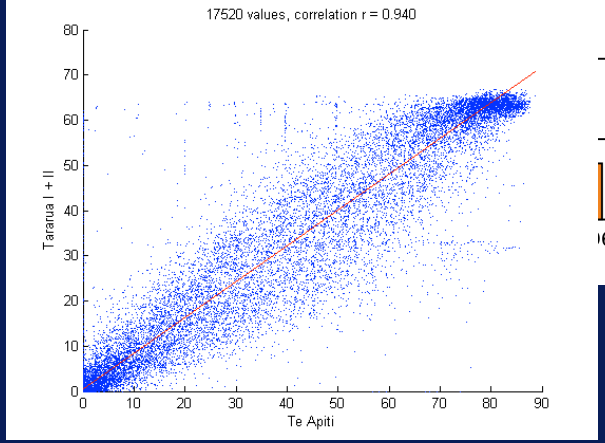
Outliers



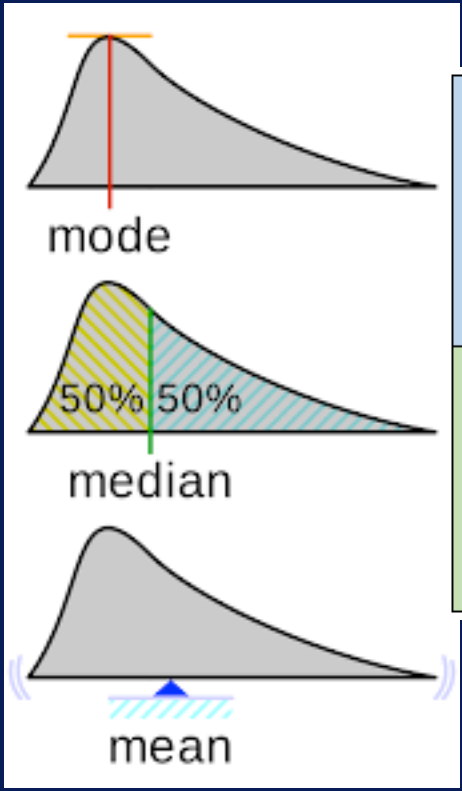
General trends



Correlations



Describe Your Data



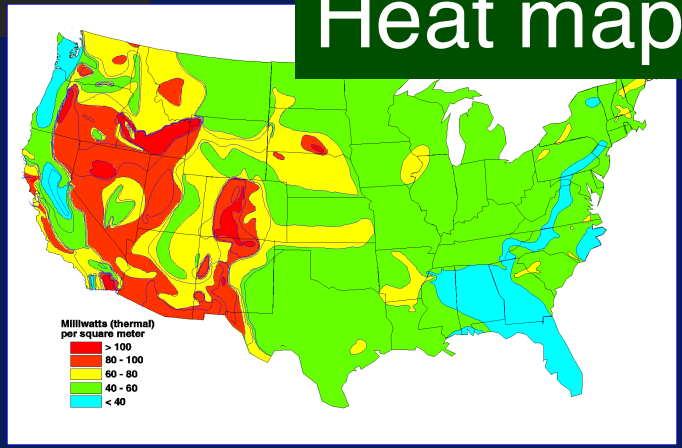
Mean	Median
Mode	Range

Statistics

Visualize Your Data

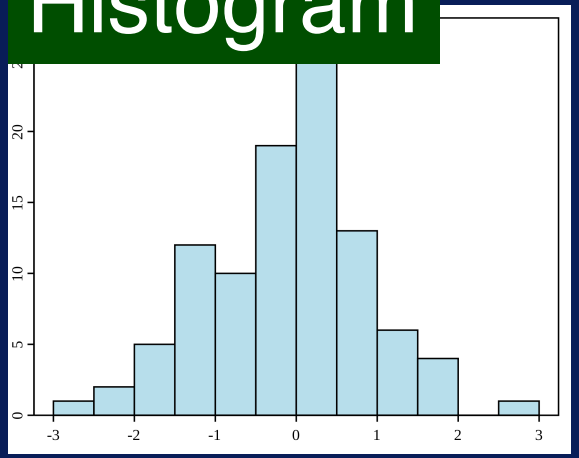
Visualize Your Data

Heat maps

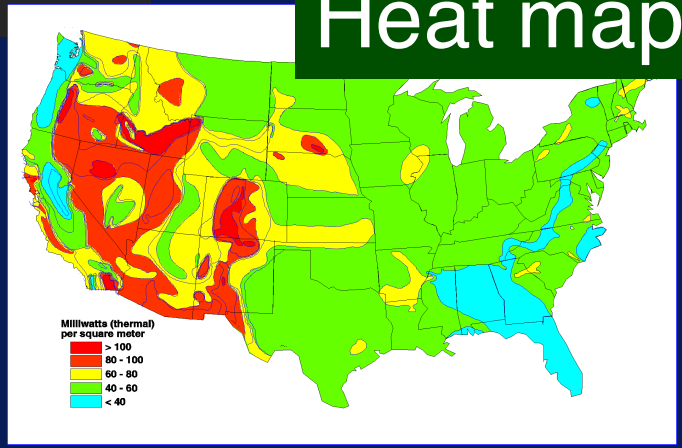


Visualize Your Data

Histogram

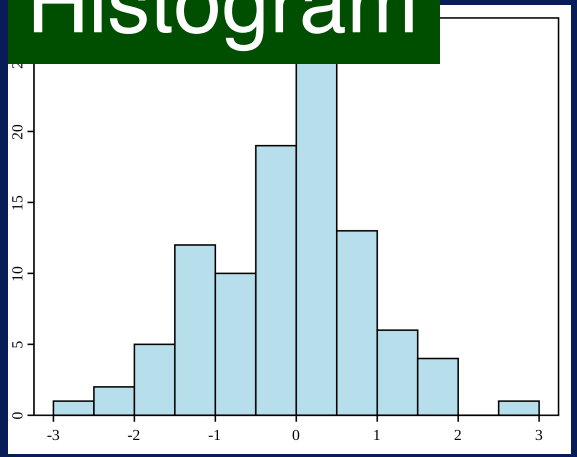


Heat maps

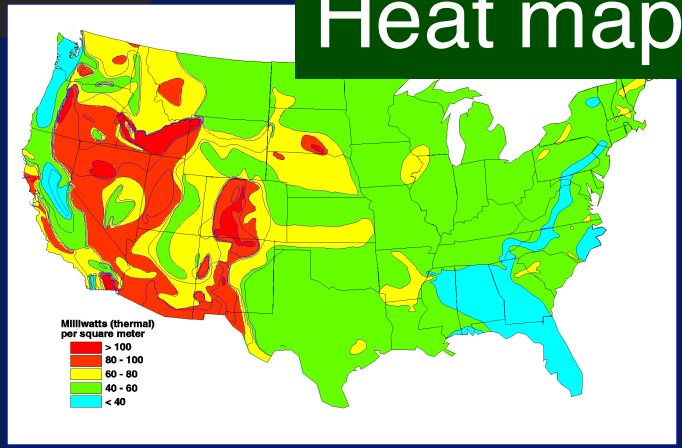


Visualize Your Data

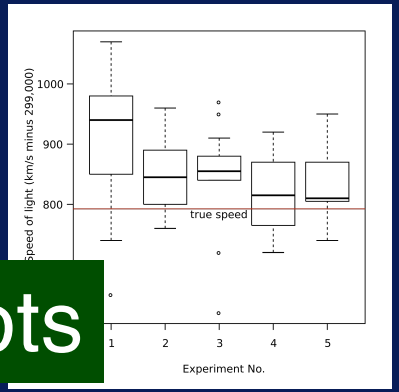
Histogram



Heat maps

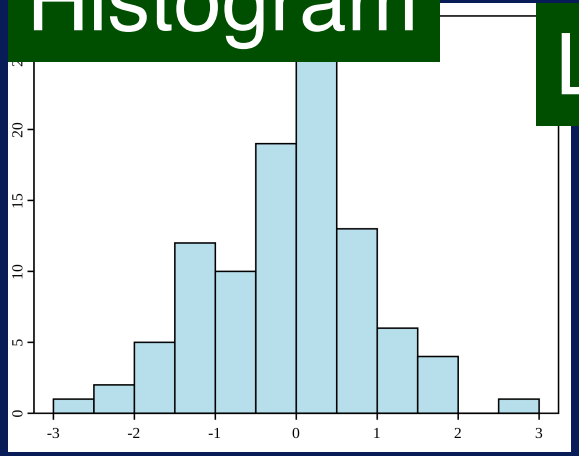


Boxplots



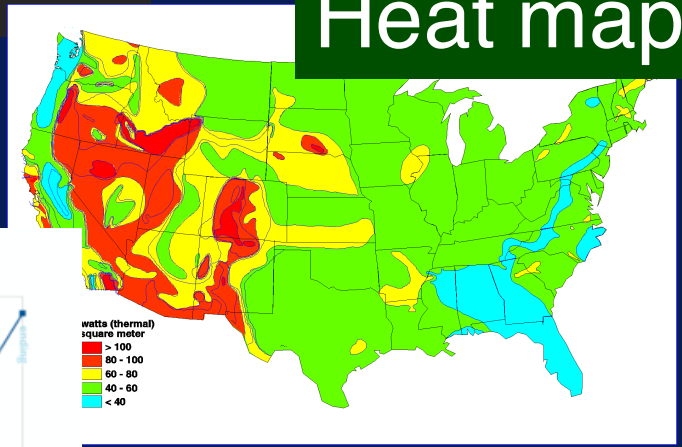
Visualize Your Data

Histogram

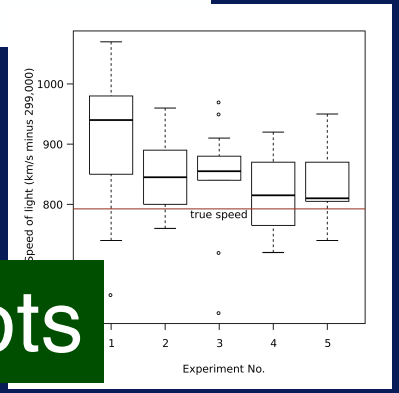


Line graphs

Heat maps

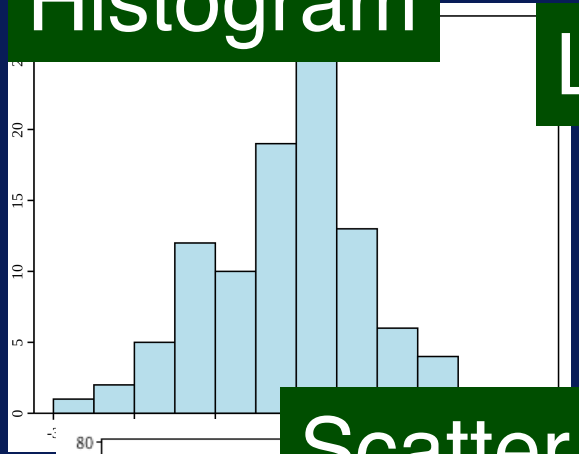


Boxplots

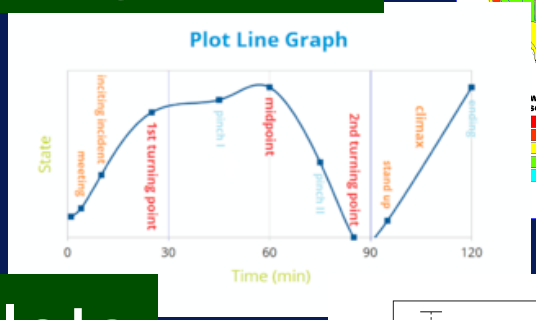


Visualize Your Data

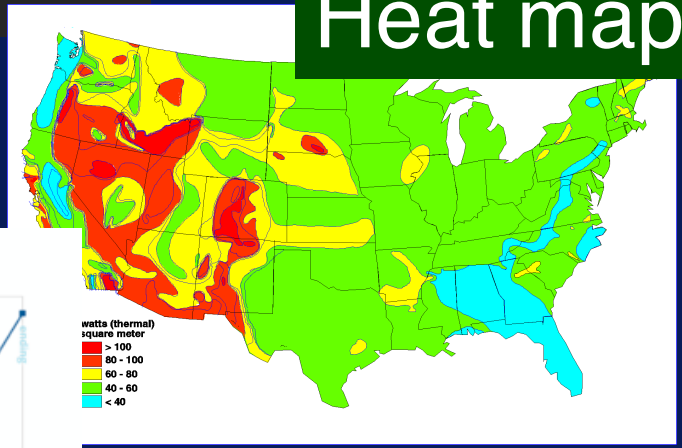
Histogram



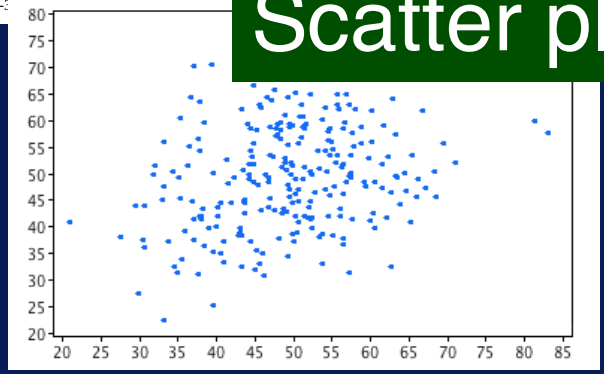
Line graphs



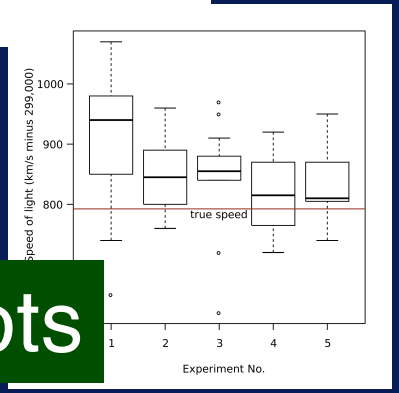
Heat maps

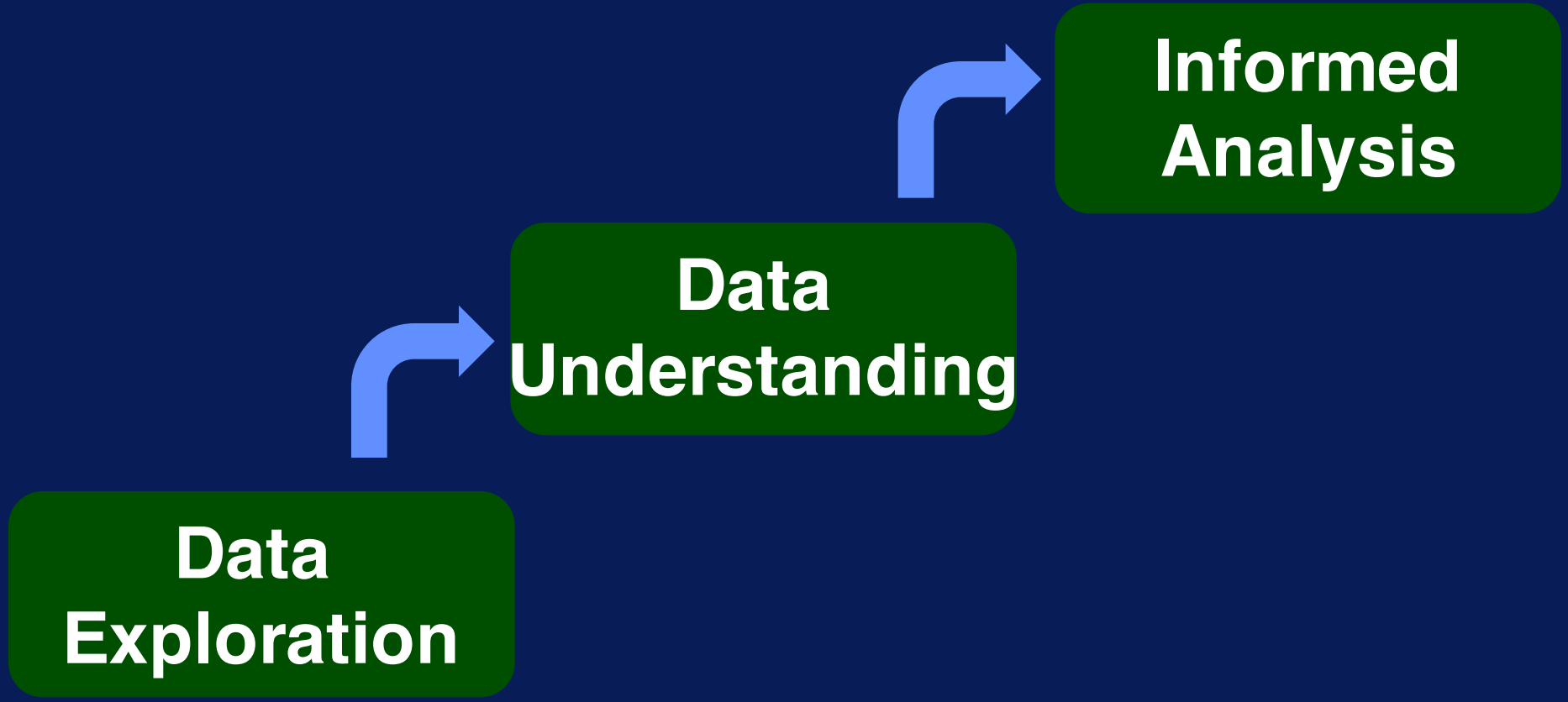


Scatter plots



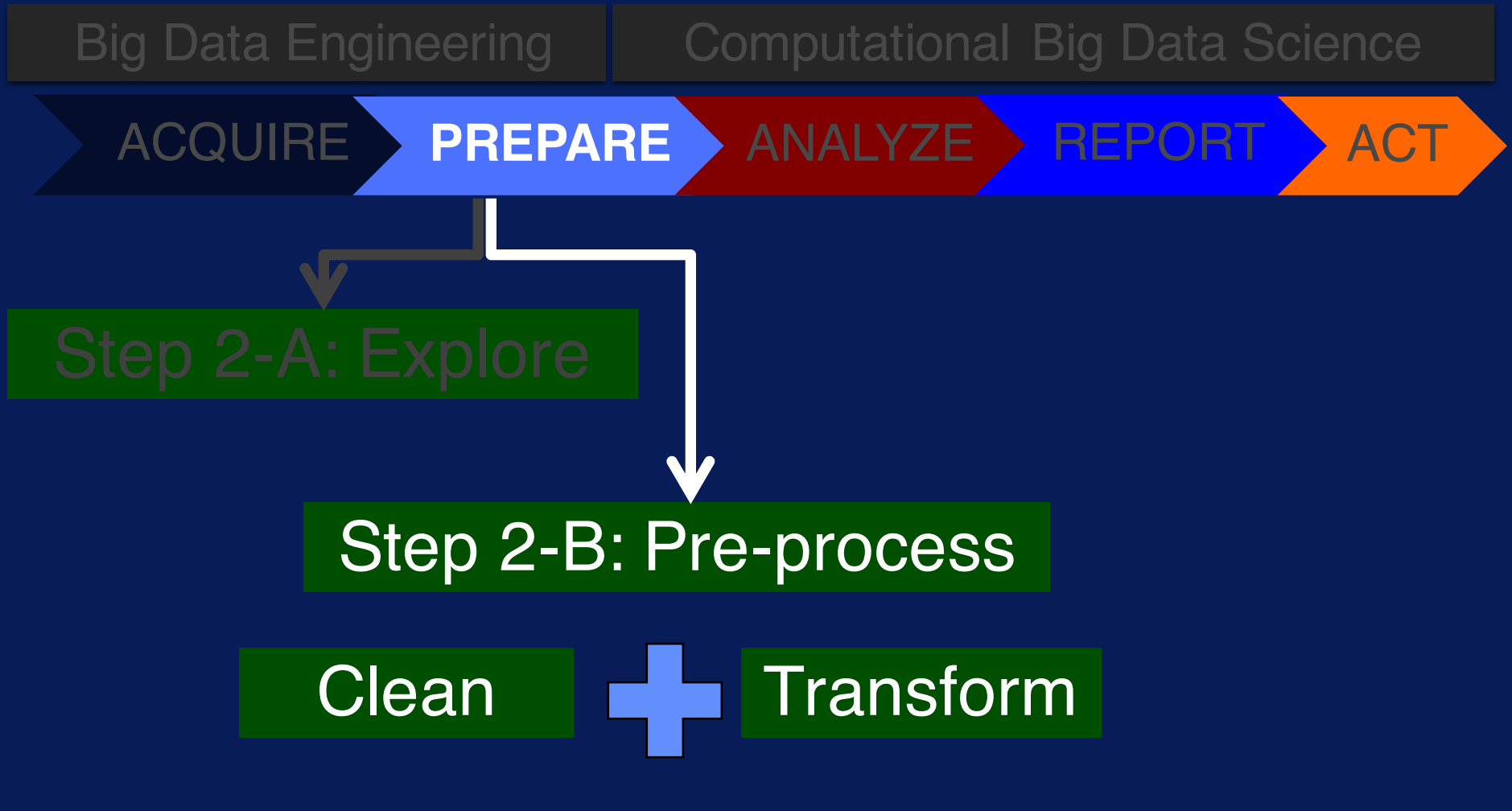
Boxplots




Informed Analysis
Data Understanding
Data Exploration

Step 2-B:

Pre-processing Data


Big Data Engineering
Computational Big Data Science
ACQUIRE
PREPARE
ANALYZE
REPORT
ACT
Step 2-A: Explore
Step 2-B: Pre-process
Clean
Transform

Real-world data is messy!

Data Quality Issues

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

Addressing Data Quality Issues

Addressing Data Quality Issues

Remove data with missing values

Addressing Data Quality Issues

Remove data with missing values

Merge duplicate records

Addressing Data Quality Issues

Remove data with missing values

Generate best estimate for invalid values

Merge duplicate records

Addressing Data Quality Issues

Remove data with missing values

Generate best estimate for invalid values

Merge duplicate records

Remove outliers

Addressing Data Quality Issues

Remove data with missing values

Generate best estimate for invalid values

Merge duplicate records

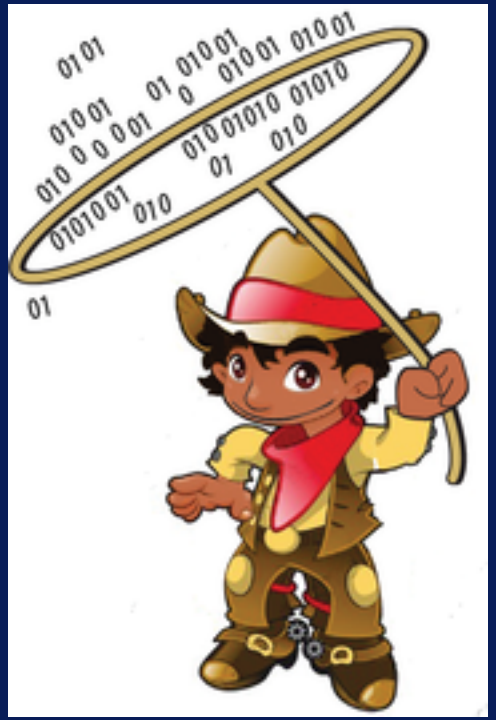
Remove outliers

Domain Knowledge

Getting Data in Shape

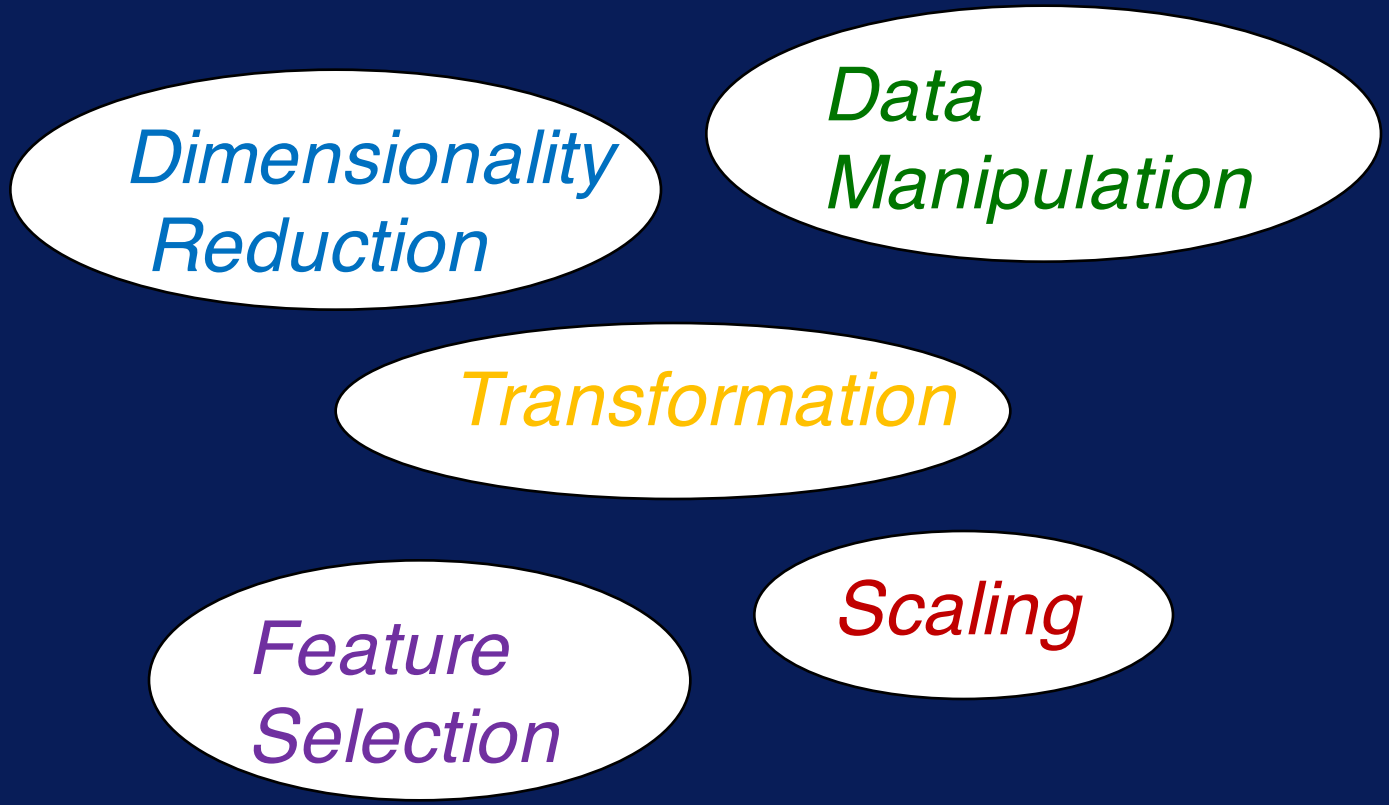
Data Munging

Data Preprocessing



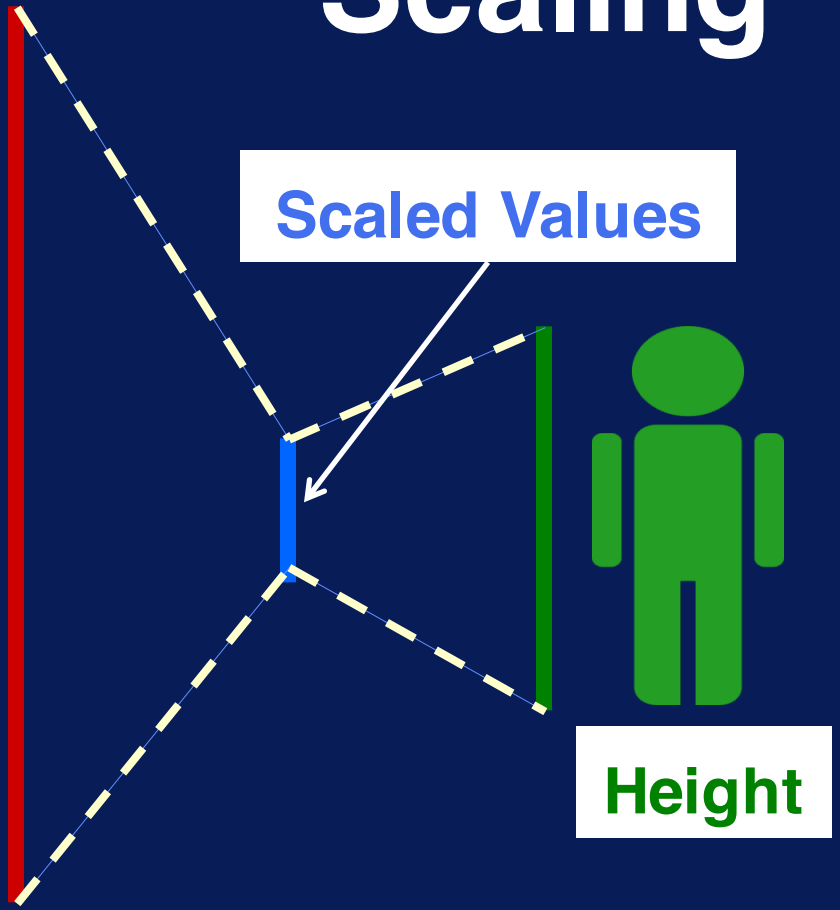
Data Wrangling

Data Munging



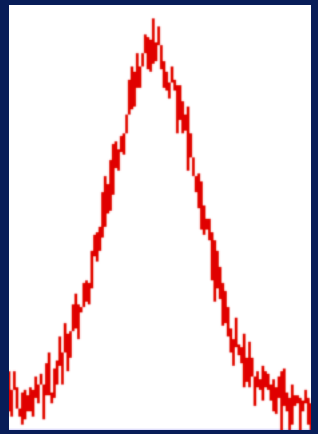
Scaling

Weight

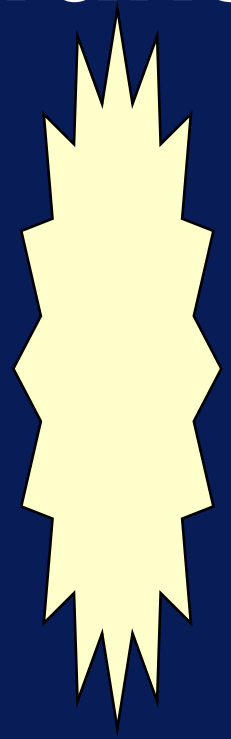


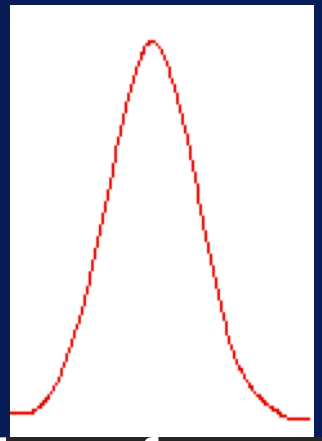
Height

Transformation



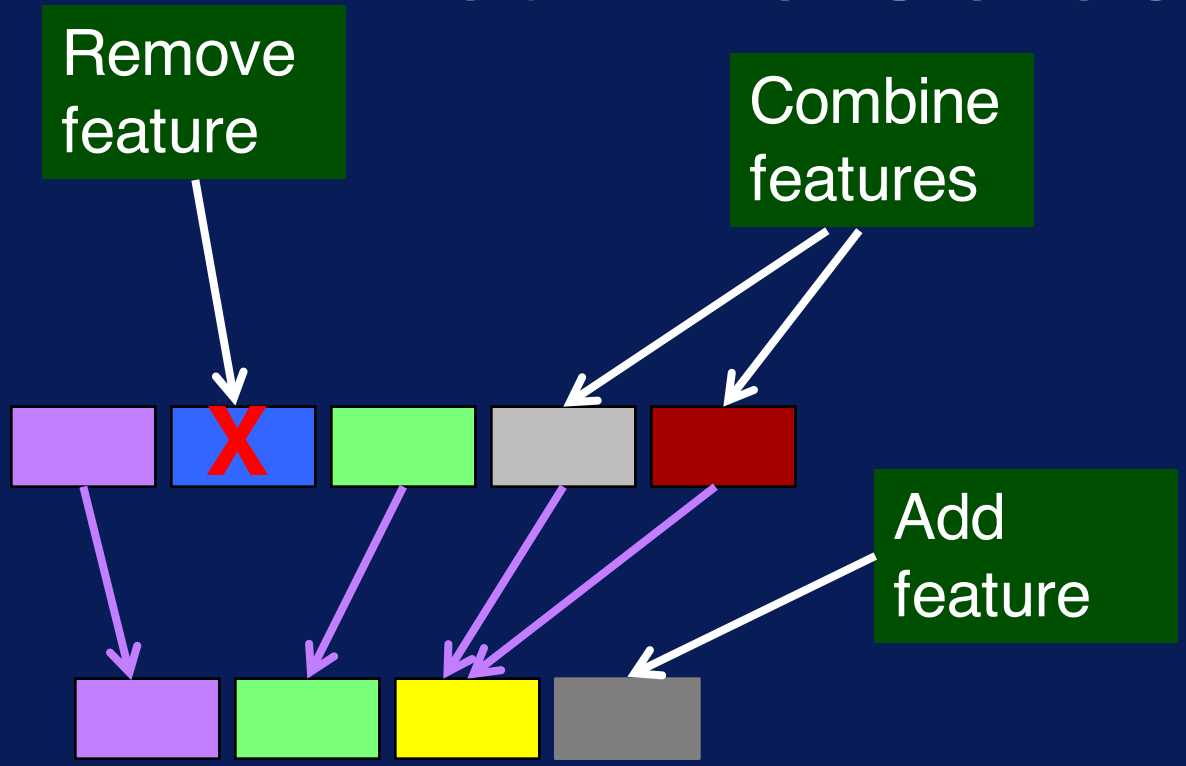
Original Data



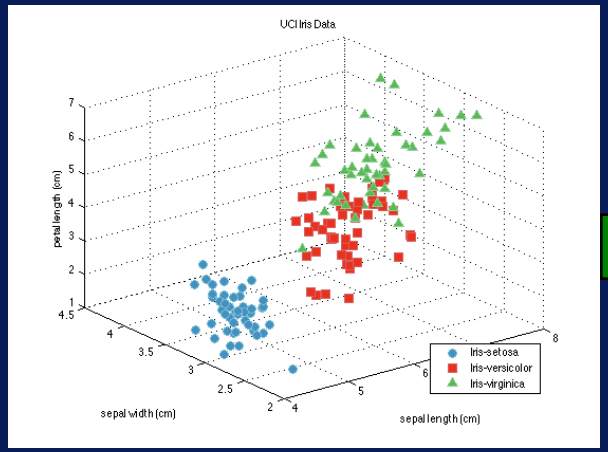


Transformed Data

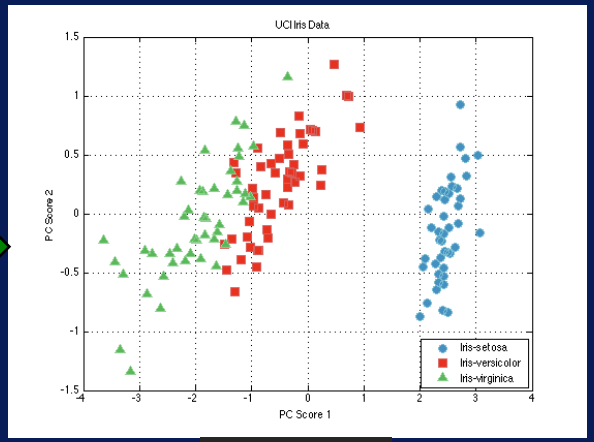
Feature Selection



Dimensionality Reduction

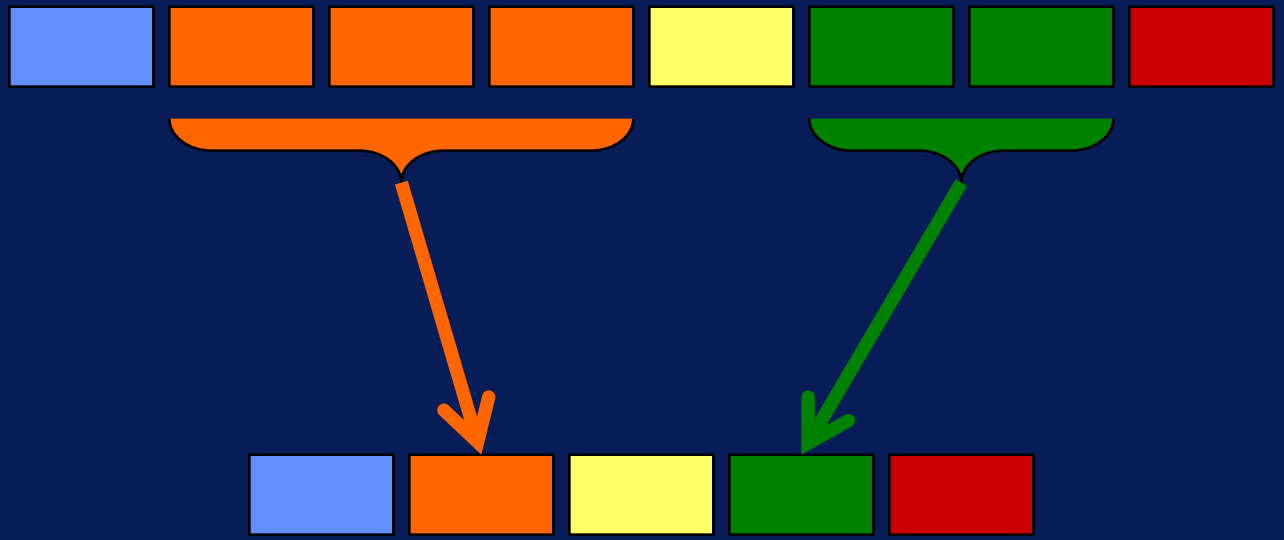


3D



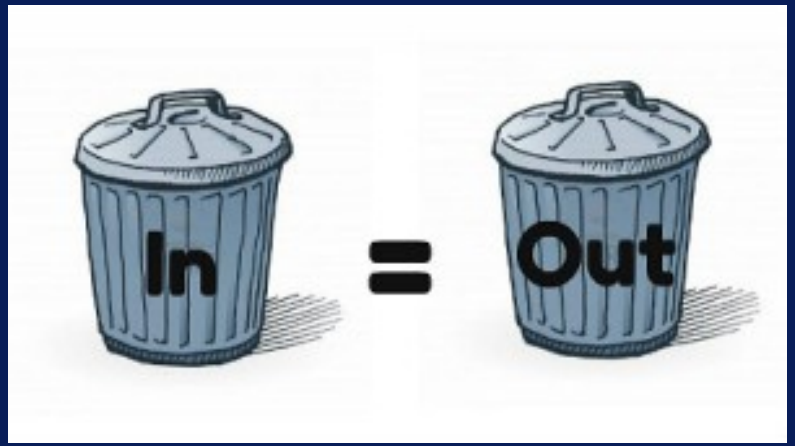
2D

Data Manipulation



Always Remember!

Garbage in = Garbage out



Data preparation is very important for meaningful analysis!

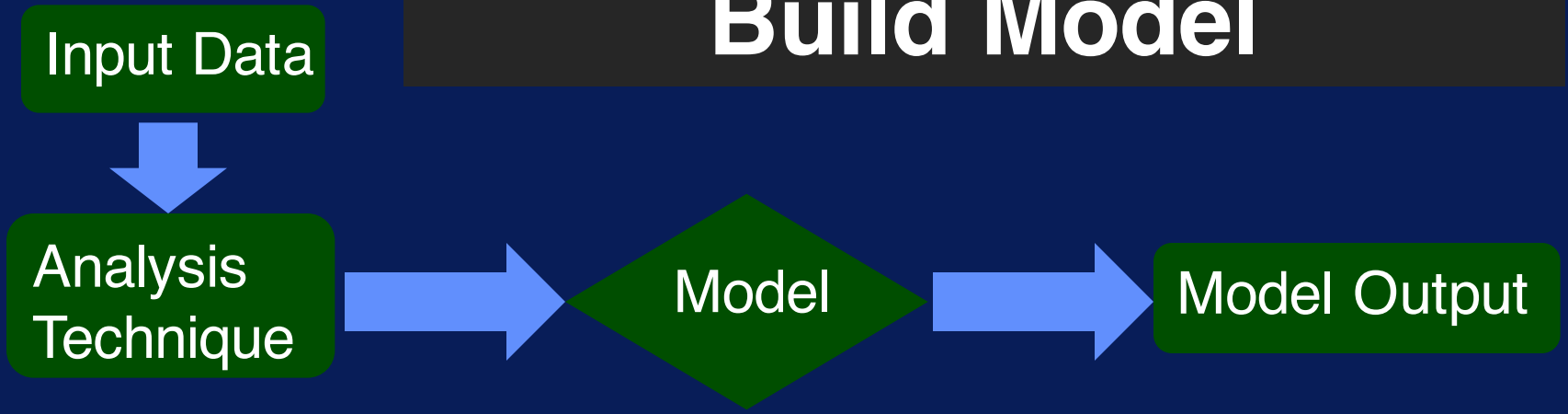
Step 3:

Analyze Data

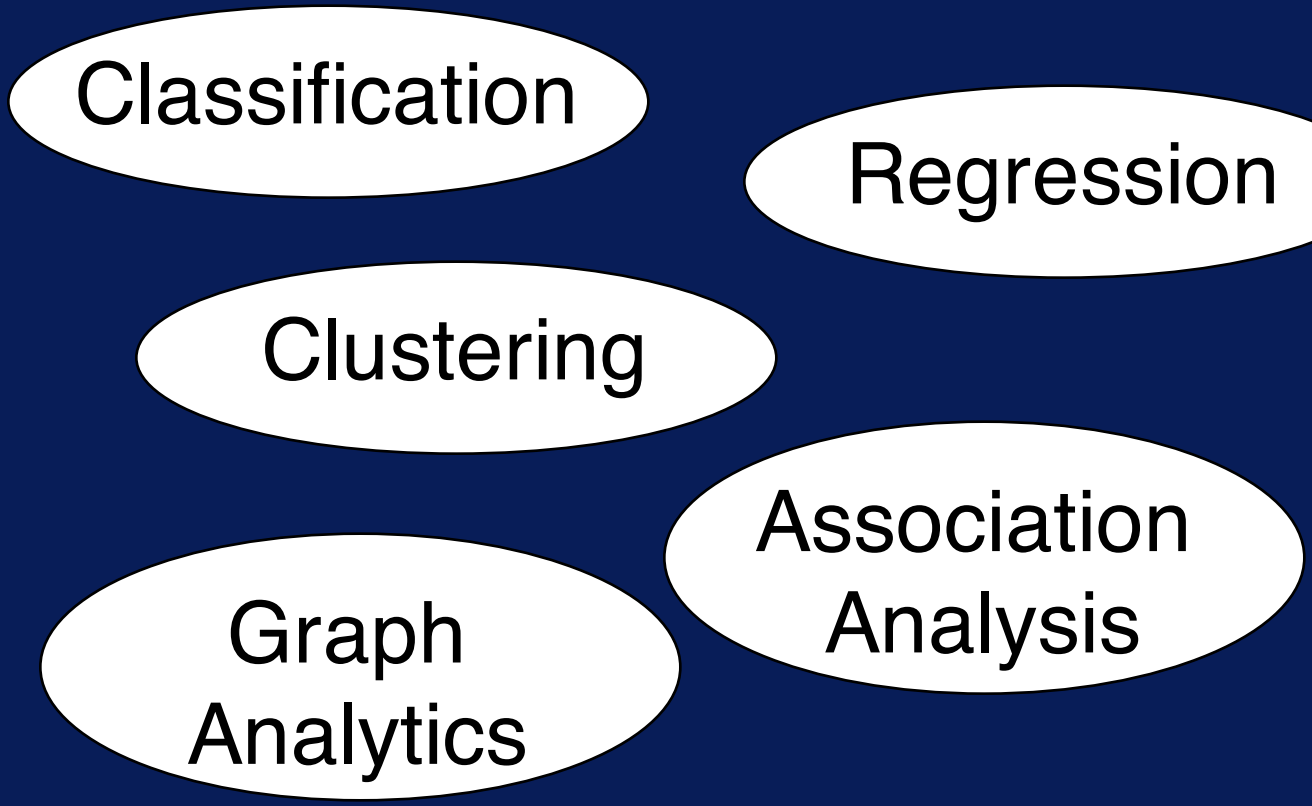
Big Data Engineering | Computational Big Data Science

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

Build Model



Categories of Analysis Techniques



Classification

Sunny

Windy

Rainy

Cloudy

Goal: Predict category

Regression

Goal: Predict numeric value

Clustering

Goal: Organize similar items into groups

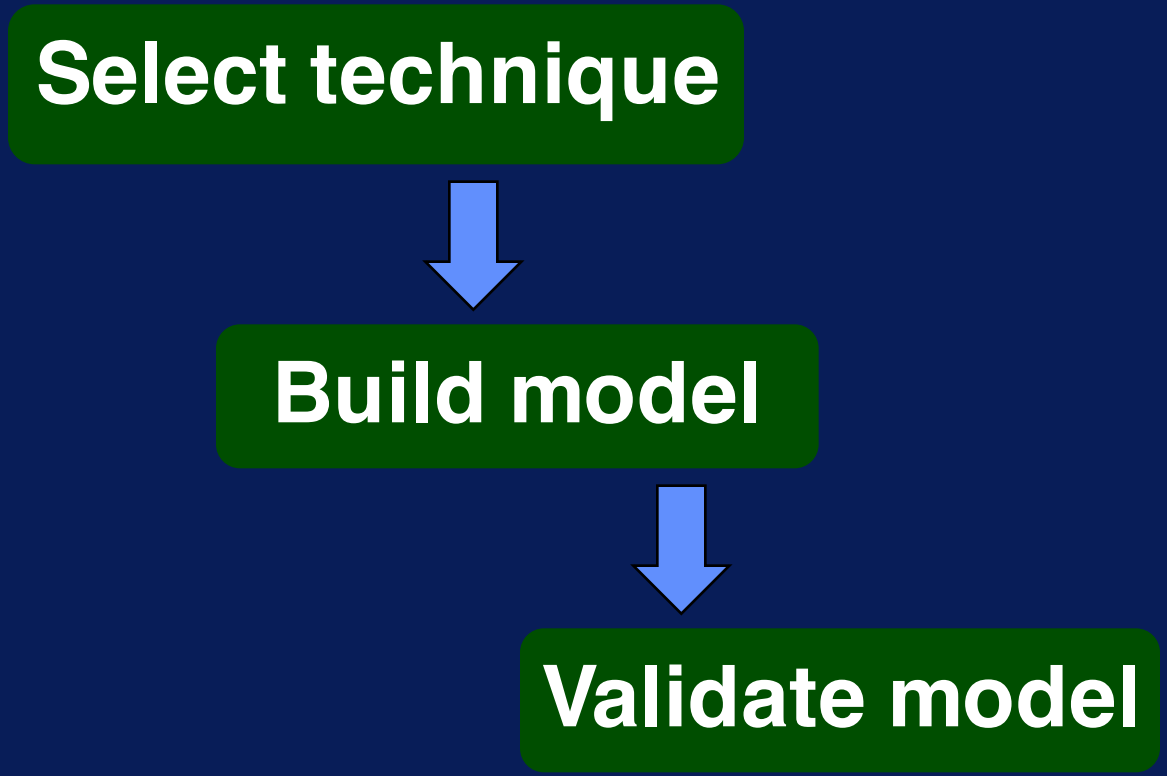
Association Analysis

Goal: Find rules to capture associations between items

Graph Analytics

Goal: Use graph structures to find connections between entities

Modeling



Evaluation of Results

Classification & Regression

Predicted Value

Correct Value

Clustering

Association Analysis & Graph Analytics

Investigate

Validate

Determine Next Steps

Repeat analysis?

Take deeper dive?

Act on results?

Select technique

Build model

Evaluate

Classification
Regression
Clustering
Association
Analysis
Graph Analytics

Step 4:

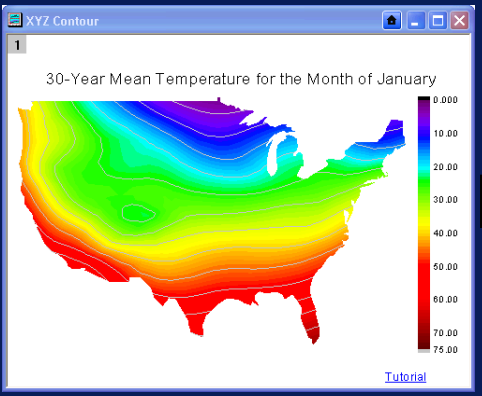
Reporting Insights

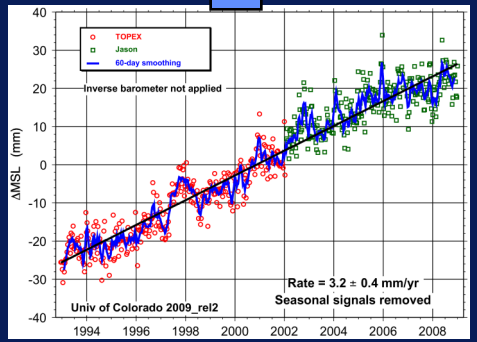

Big Data Engineering
Computational Big Data Science
ACQUIRE
PREPARE
ANALYZE
REPORT
ACT

What to Present

What to Present

How to Present

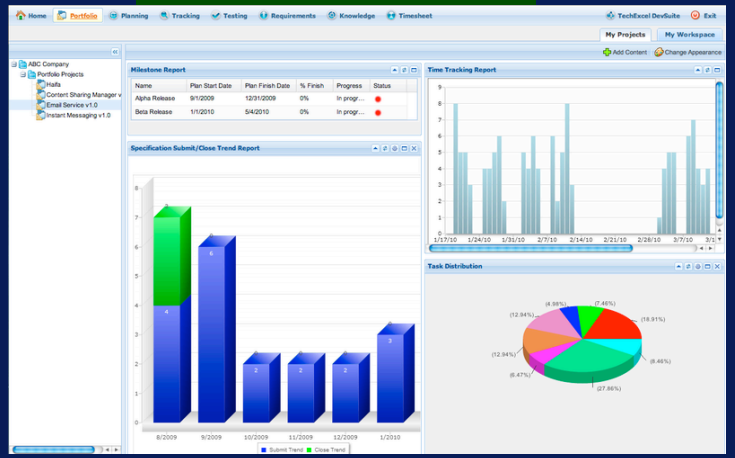




Visualization Tools

Present

with



using

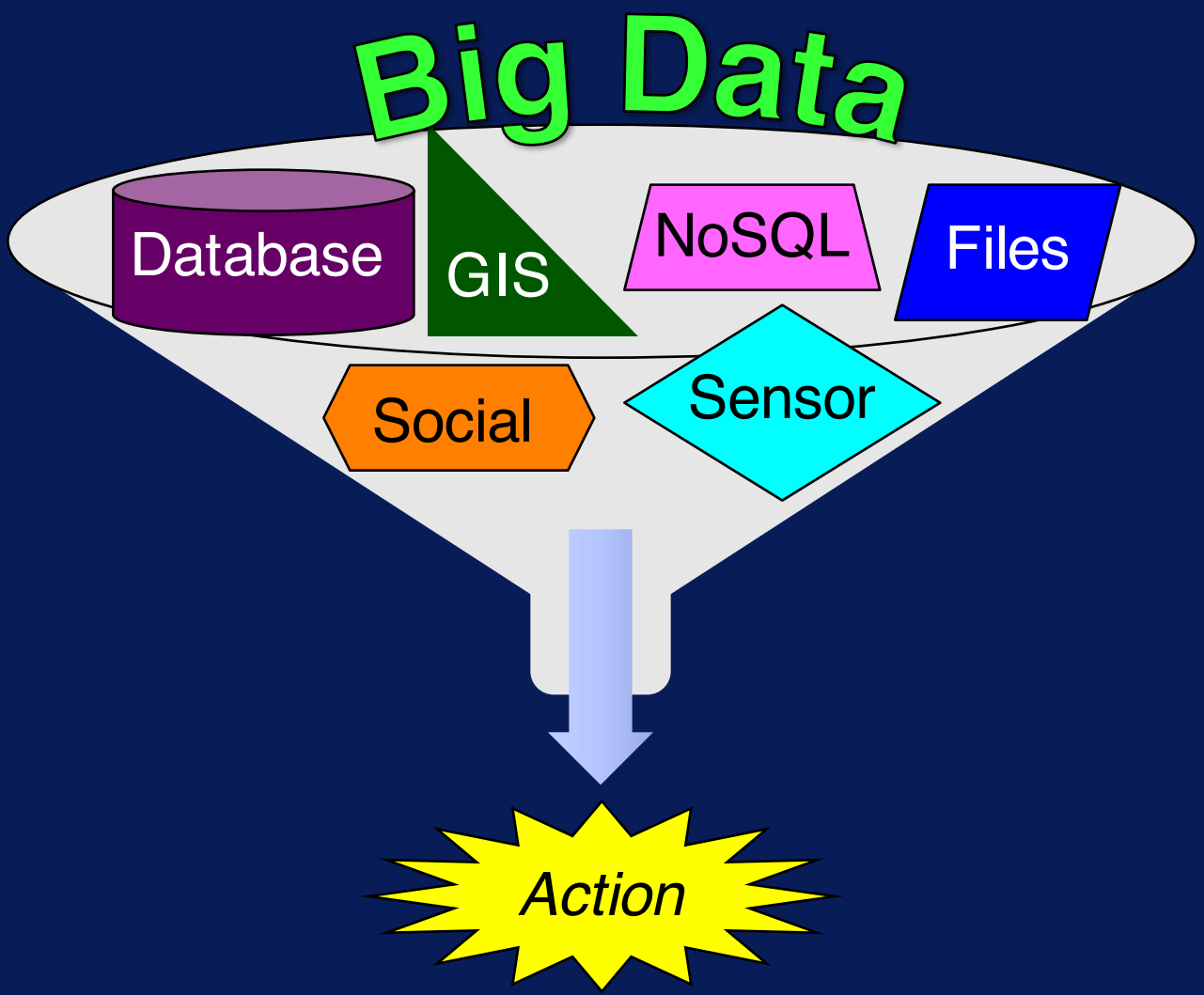
tableau

Google Developers CHARTS

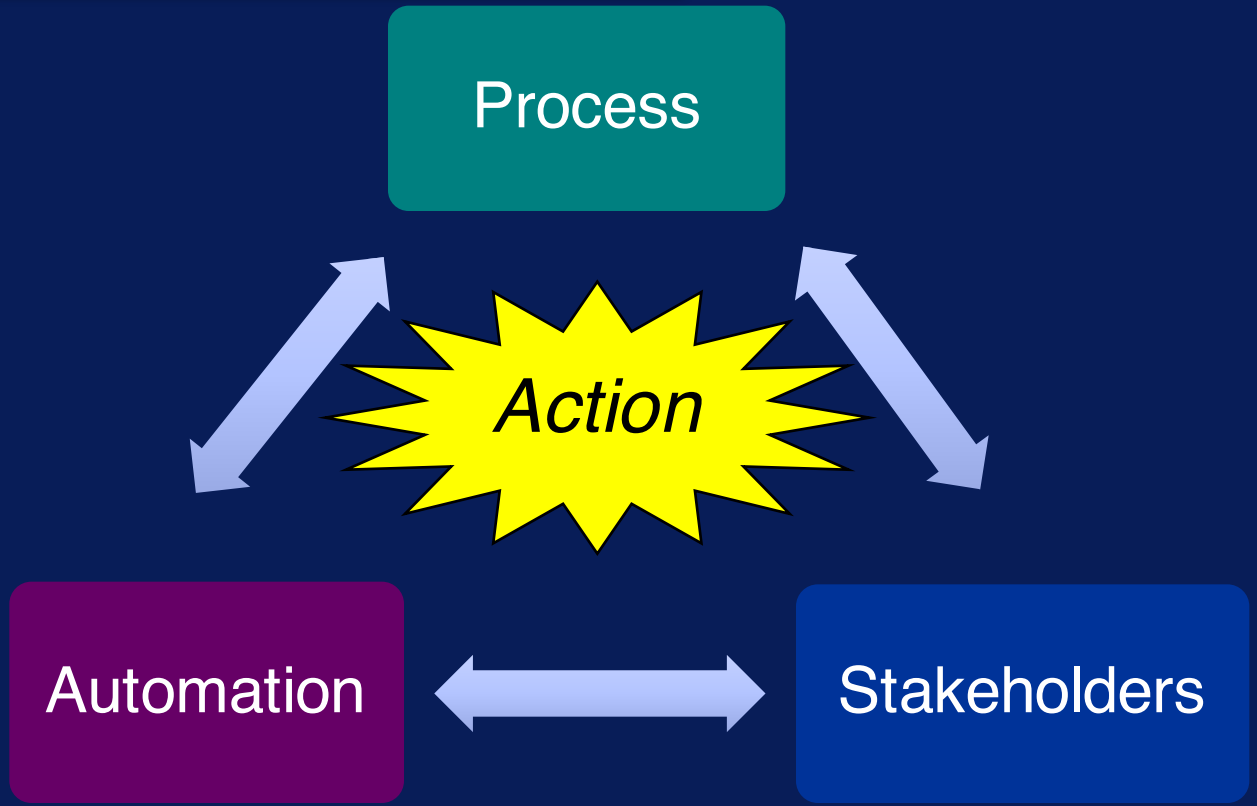
Step 5:

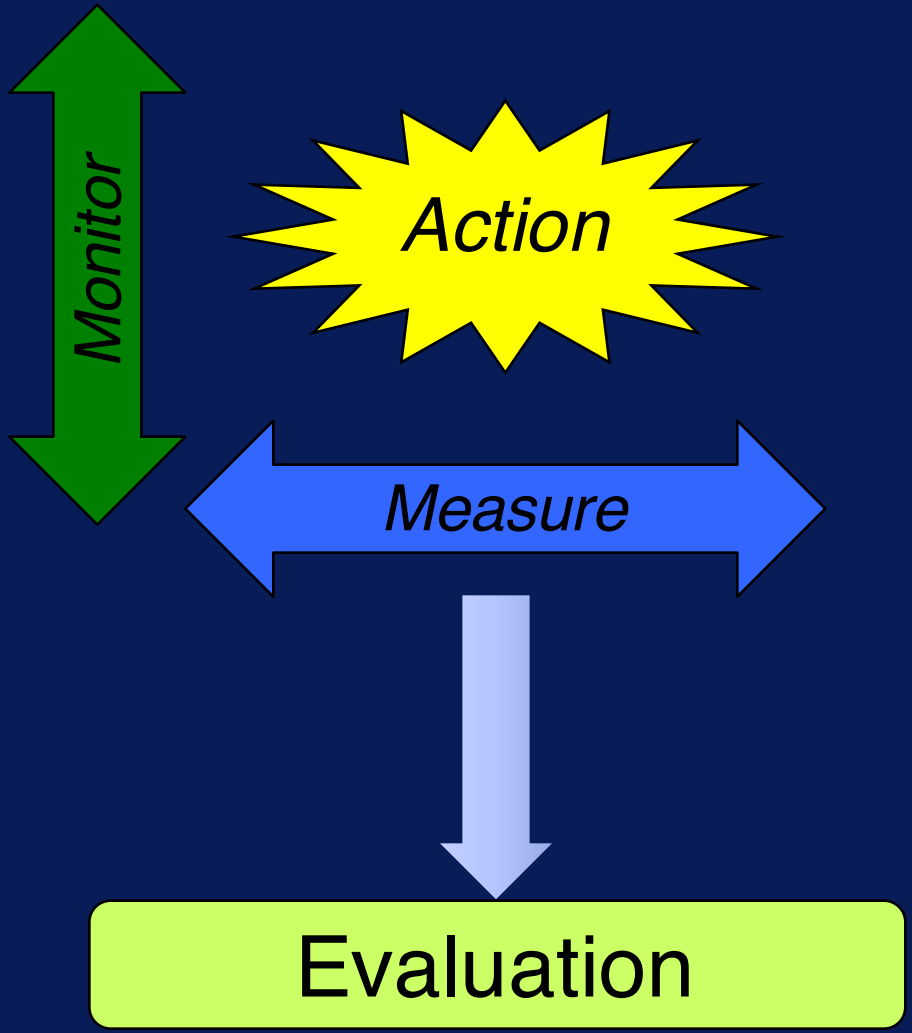
Insights into Action


Big Data Engineering
Computational Big Data Science
ACQUIRE
PREPARE
ANALYZE
REPORT
ACT

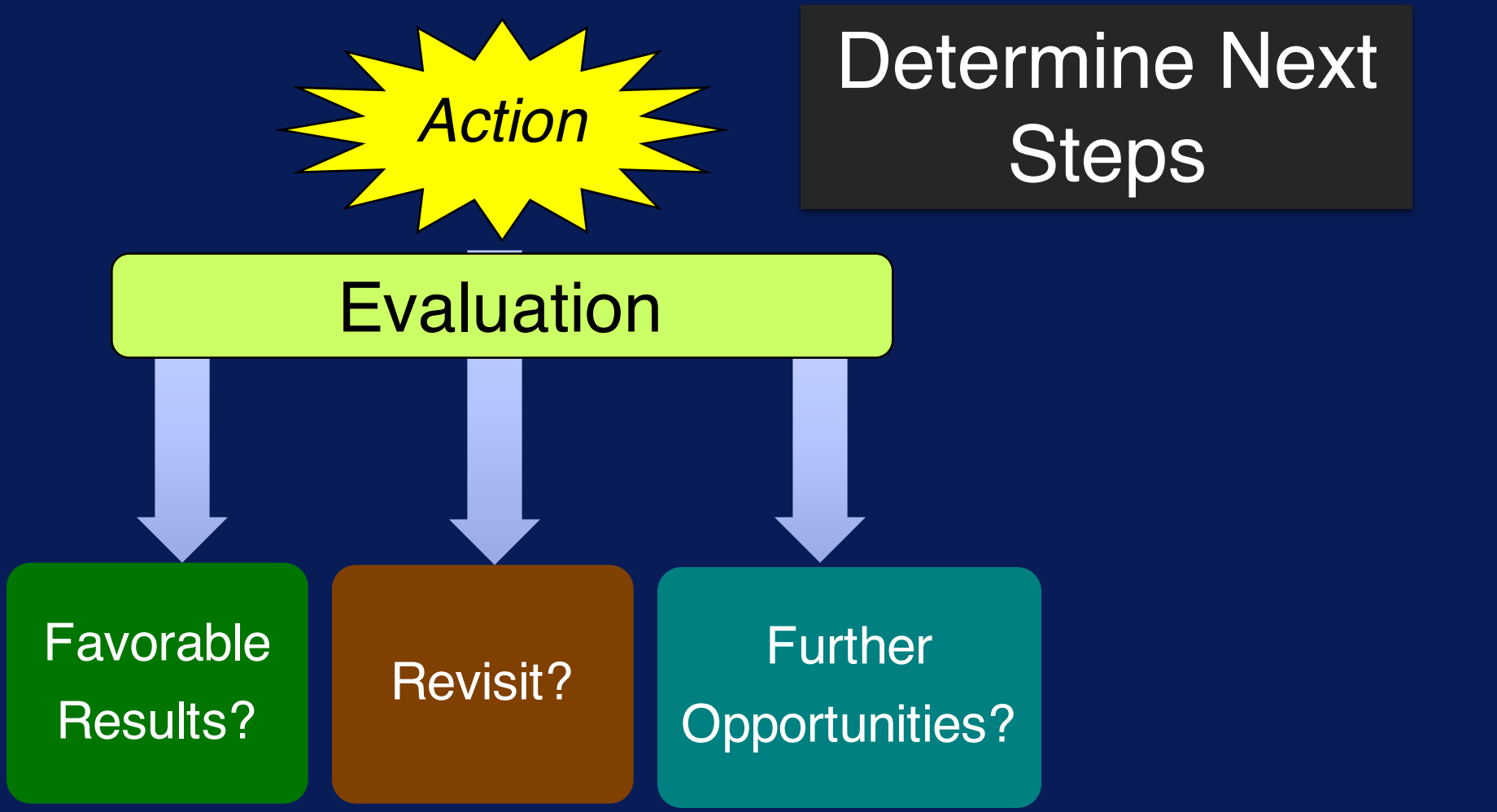

Big Data
Database
GIS
NoSQL
Files
Social
Sensor
Action

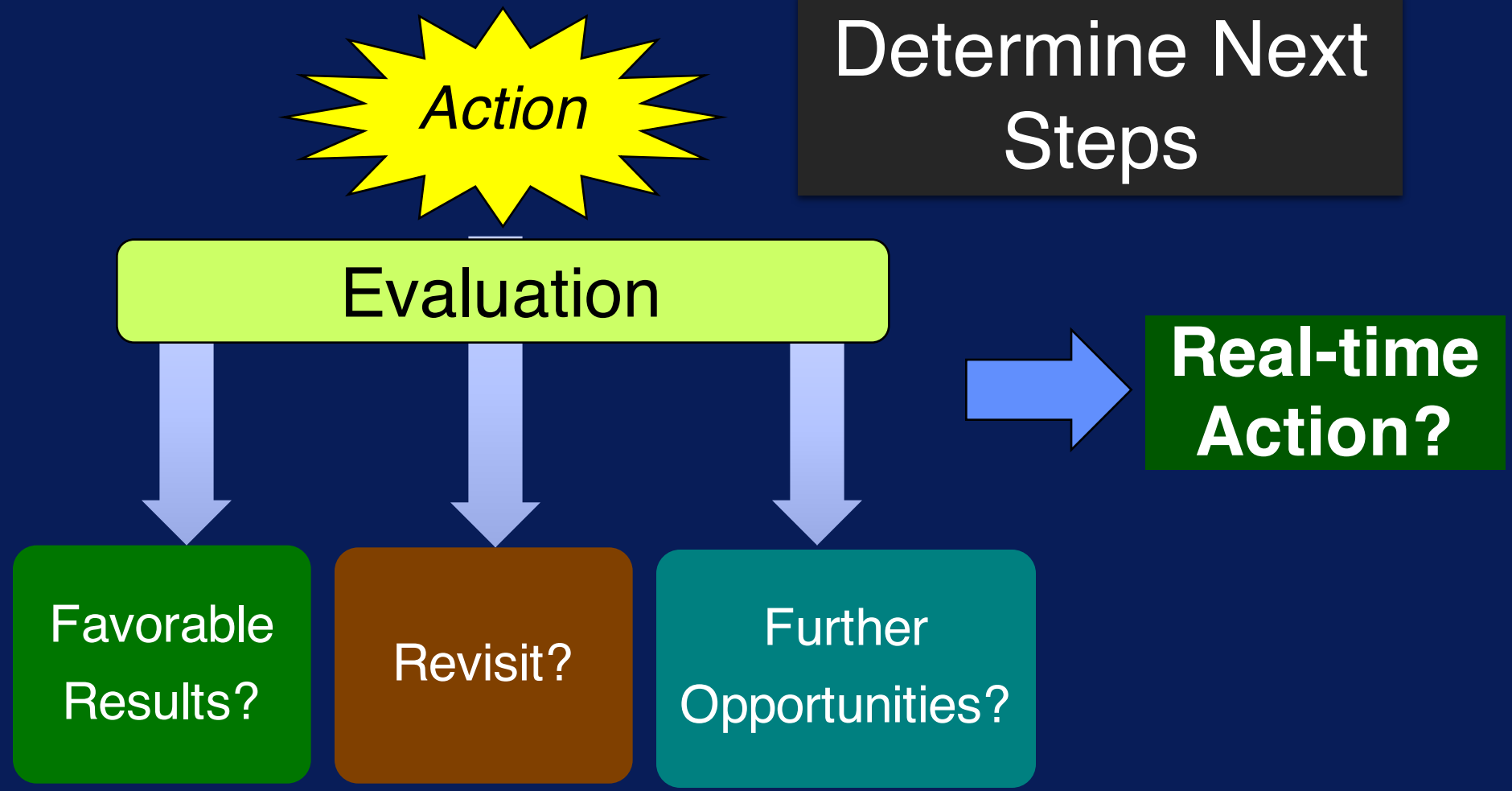
Implementation





Assess Impact


Action
Determine Next Steps
Evaluation
Favorable Results?
Revisit?
Further Opportunities?


Action
Determine Next Steps
Evaluation
Real-time Action?
Favorable Results?
Revisit?
Further Opportunities?