Python for Data Science

How does data science happen?

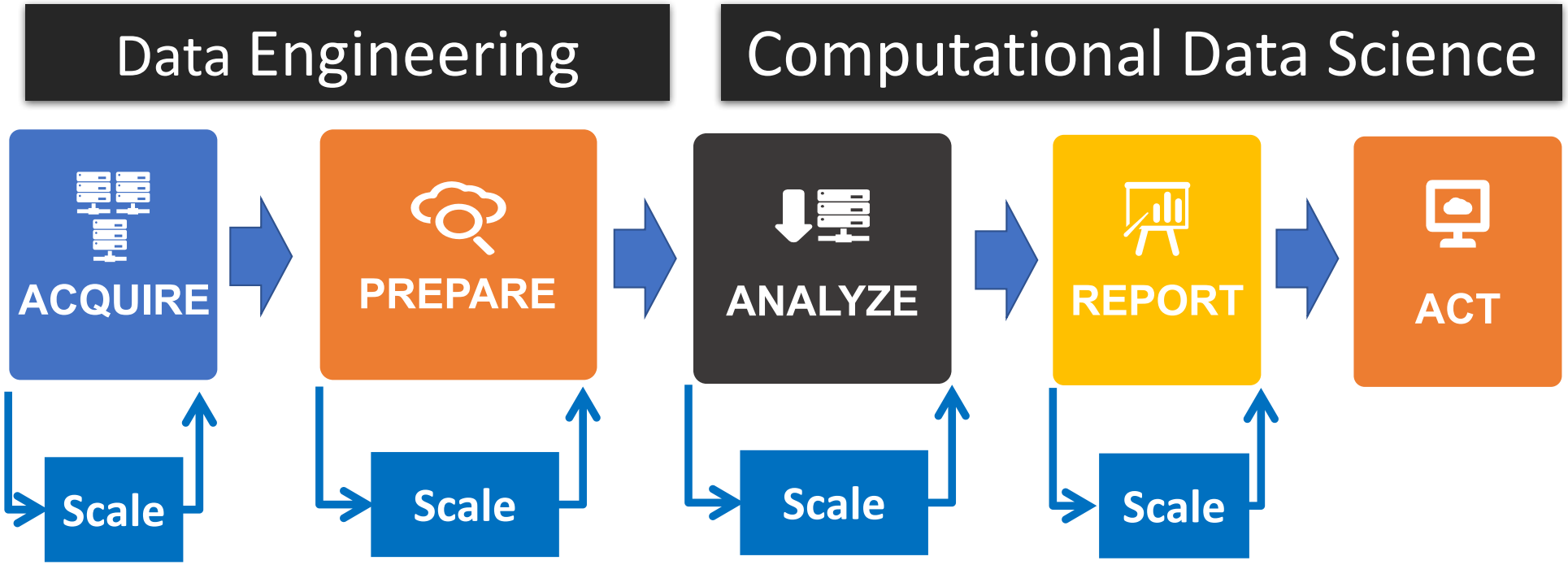
Dr. Ilkay Altintas and Dr. Leo Porter

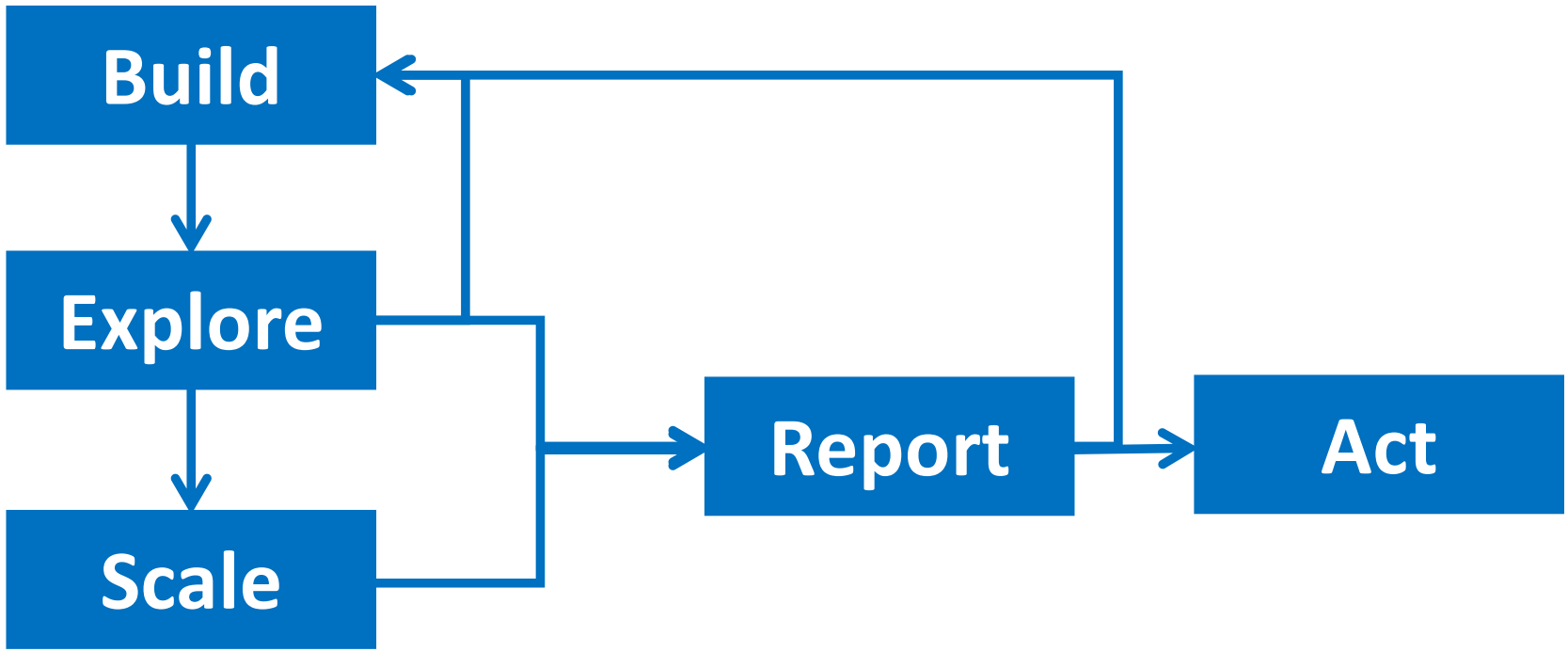
Twitter: #UCSDpython4DS

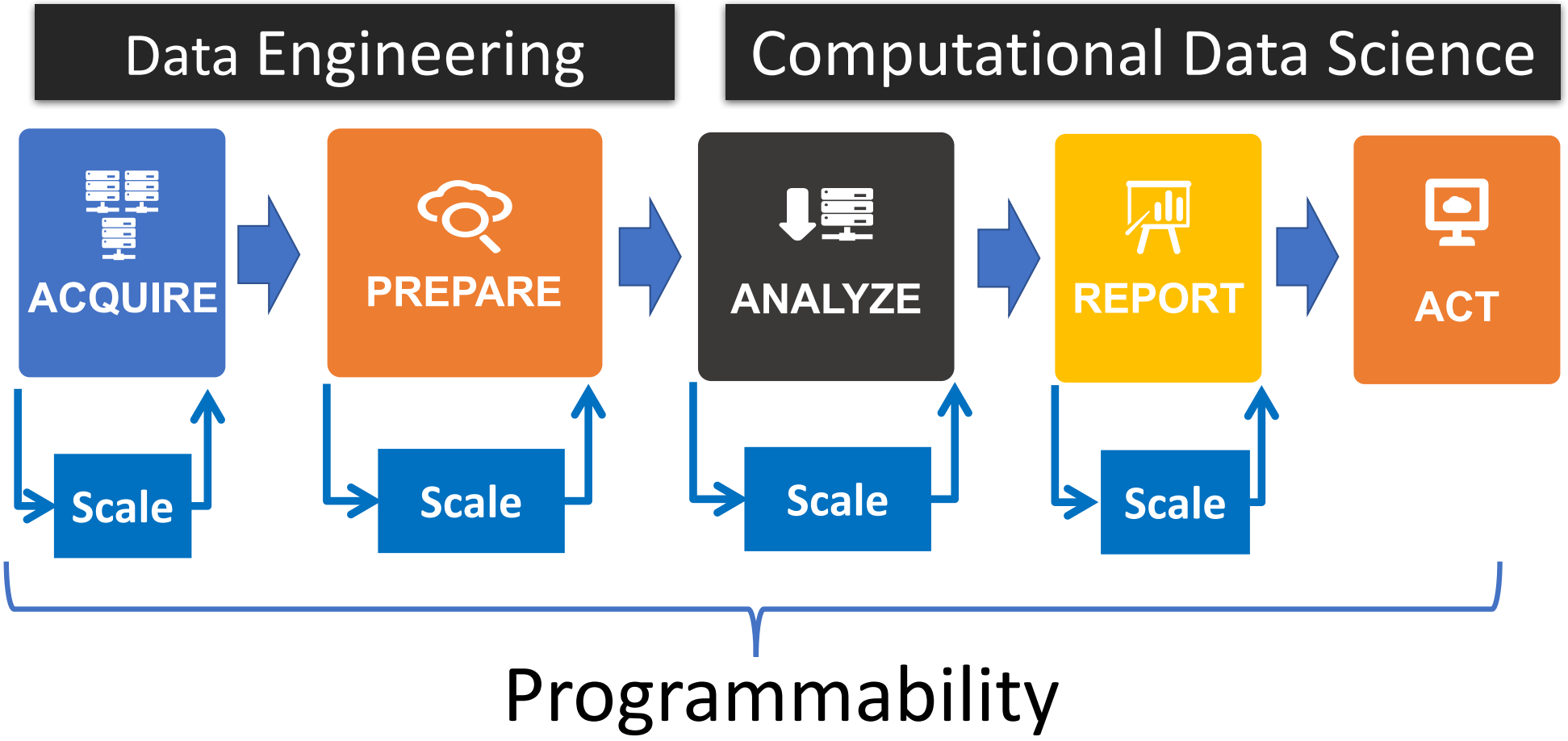
After this video you will be able to..

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

Data Science Process


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


Build
Explore
Scale
Report
Act


Data Engineering
Computational Data Science
ACQUIRE
PREPARE
ANALYZE
REPORT
ACT
Scale
Scale
Scale
Scale
Programmability

Python for Data Science

Asking the Right Question

Dr. Ilkay Altintas and Dr. Leo Porter

Twitter: #UCSDpython4DS

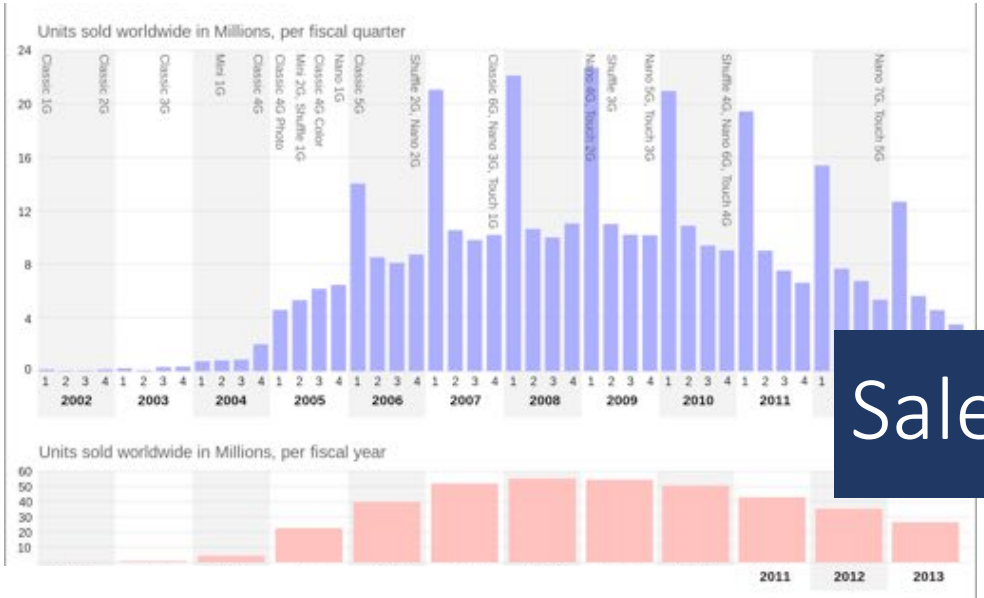
After this video you will be able to..

- 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

Sensor data

Sensor data

Sensor data

Detect equipment failure

Customer data

+

Marketing data

Better targeted marketing

1 **Market Size and Growth**

©The LPO Program 2012

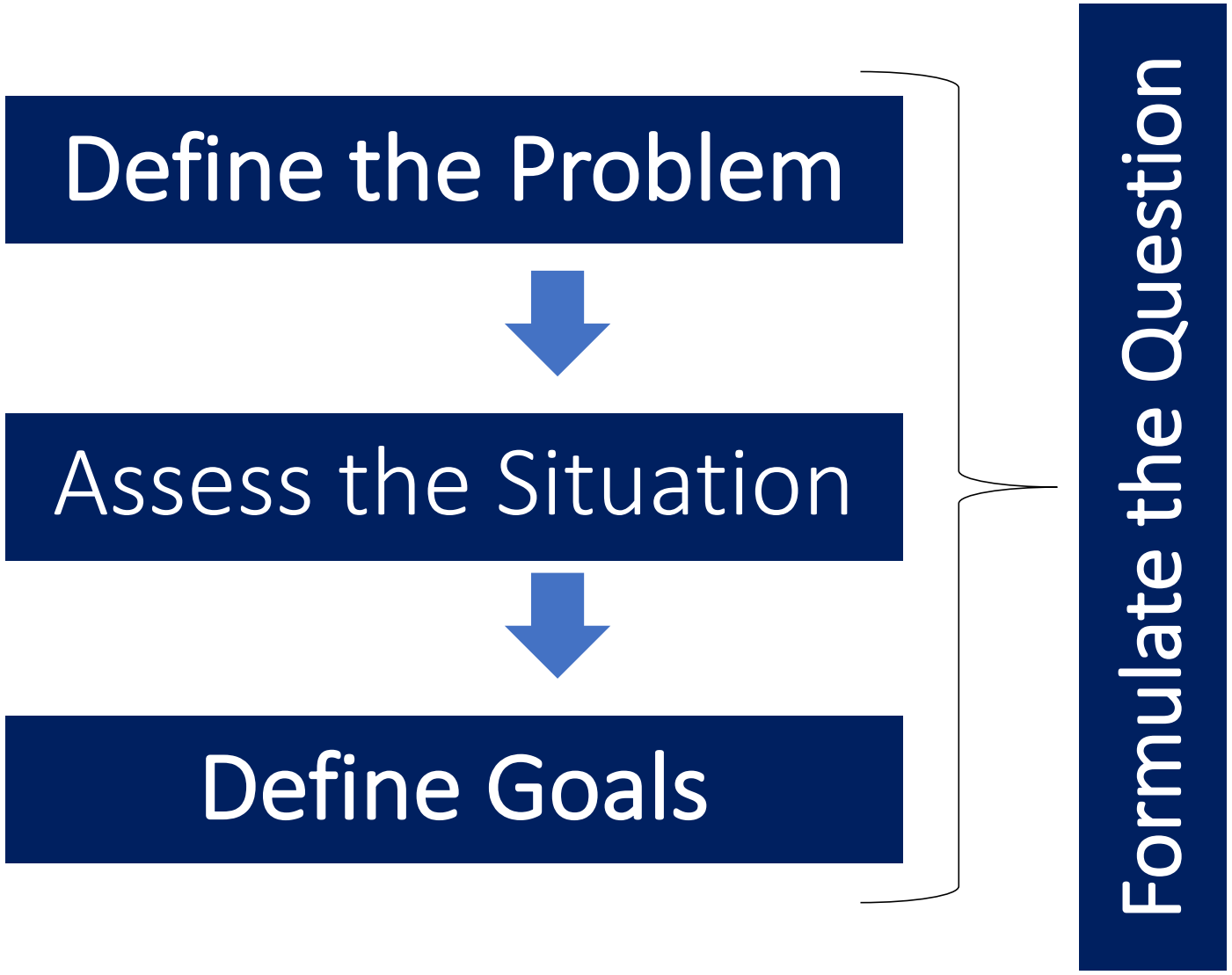
Assess the Situation

Assess the Situation

- Risks
- Benefits
- Contingencies
- Regulations
- Resources
- Requirements

Define Goals

- Objectives
- Criteria


Define the Problem
Assess the Situation
Define Goals
Formulate the Question

Python for Data Science

Steps in the Data Science Process

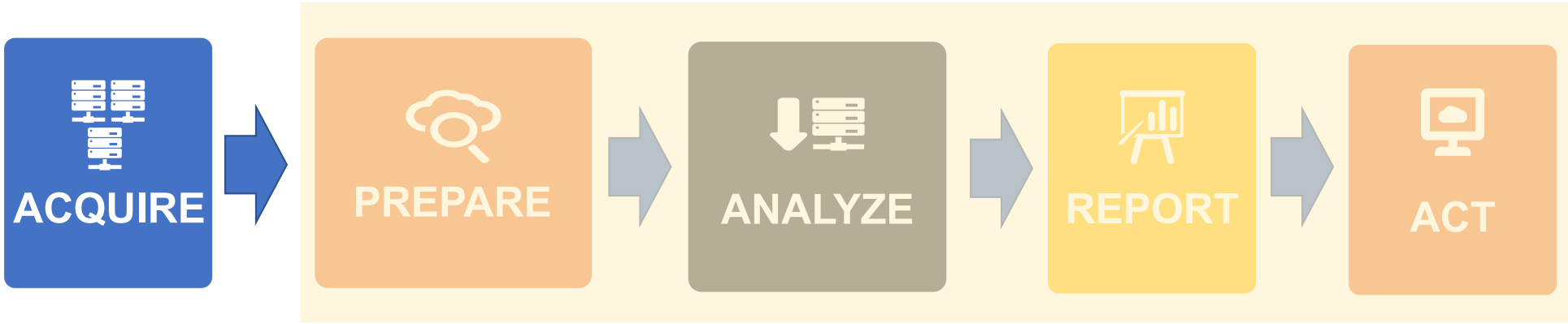
Dr. Ilkay Altintas and Dr. Leo Porter

Twitter: #UCSDpython4DS

After this video you will be able to..

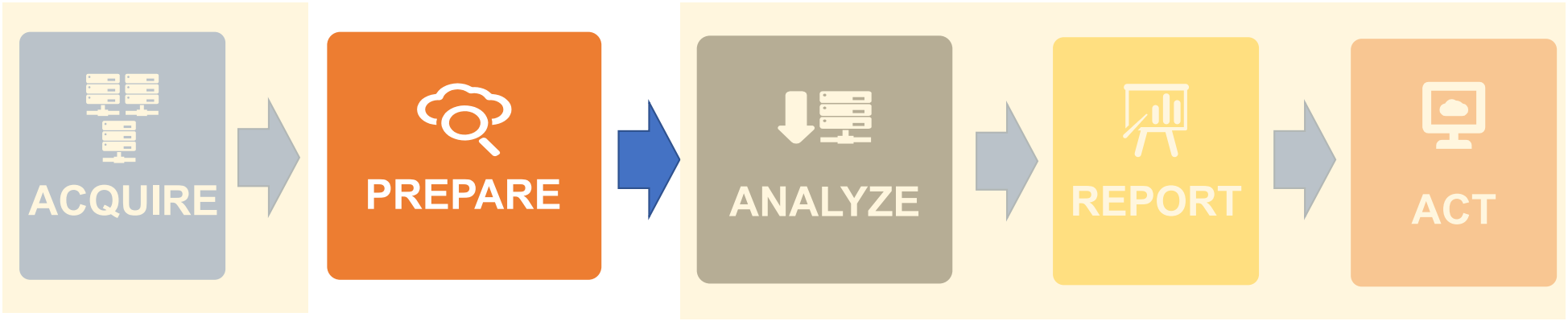
- Identify the steps in the data science process
- Understand what each step involves


ACQUIRE
PREPARE
ANALYZE
REPORT
ACT



Step 1: Acquire Data

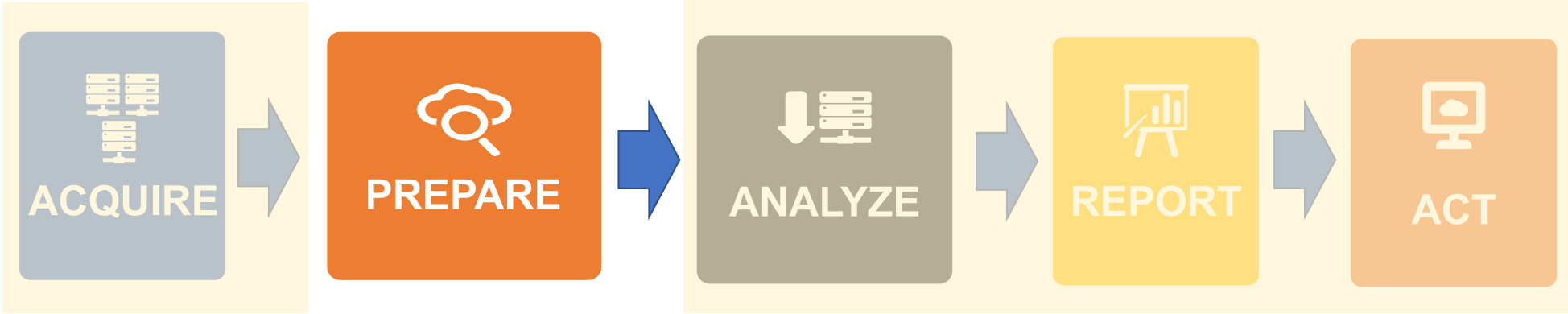
- Identify data sets
- Retrieve data
- Query data



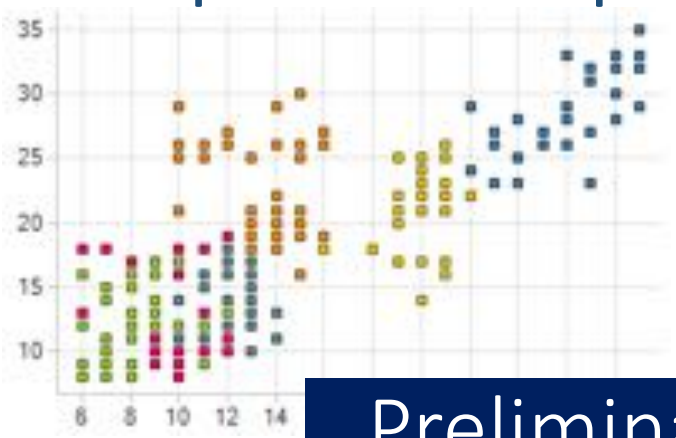
Step 2: Prepare Data

Step 2-A: Explore

Step 2-B: Pre-process

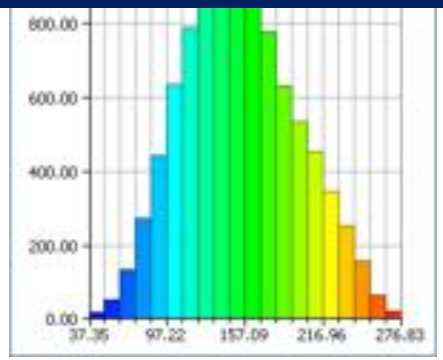


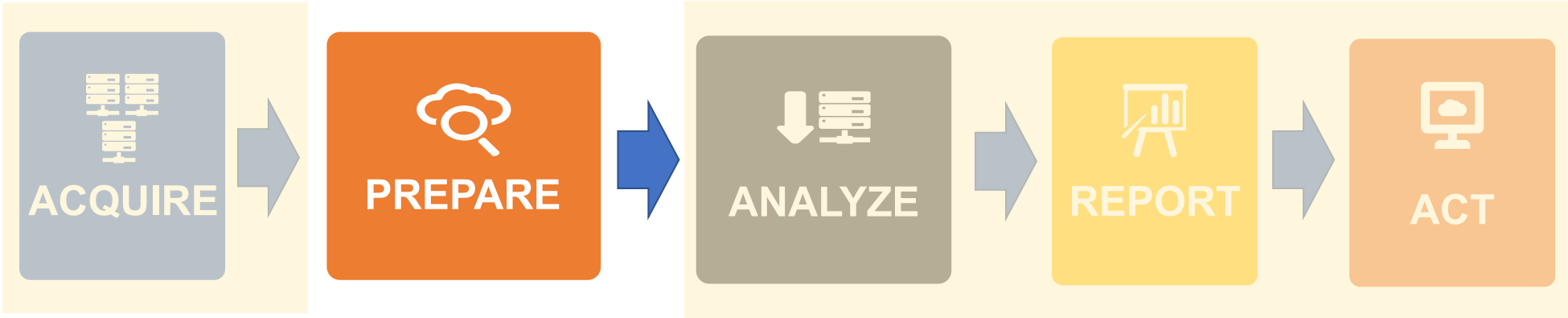
Step 2-A: Explore Data



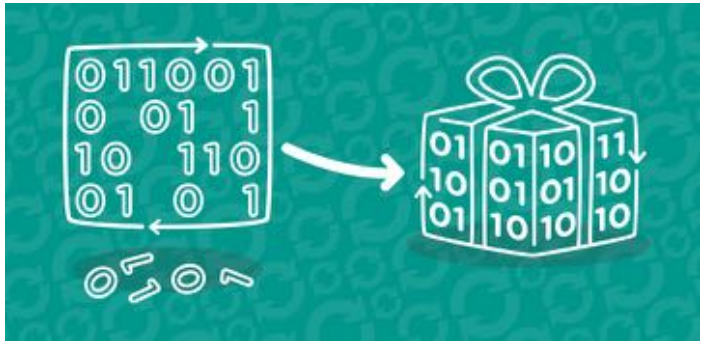
Preliminary analysis

Understand nature of data

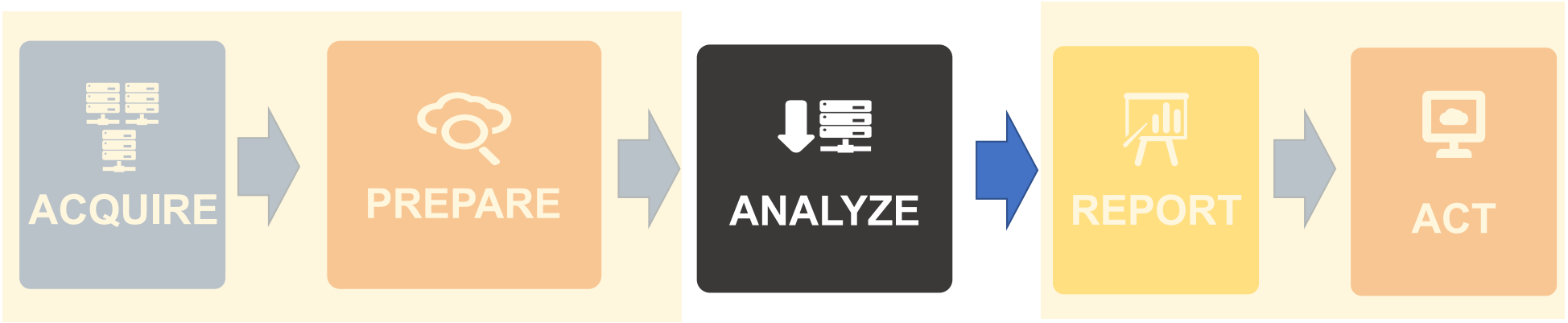




Step 2-B: Pre-process Data



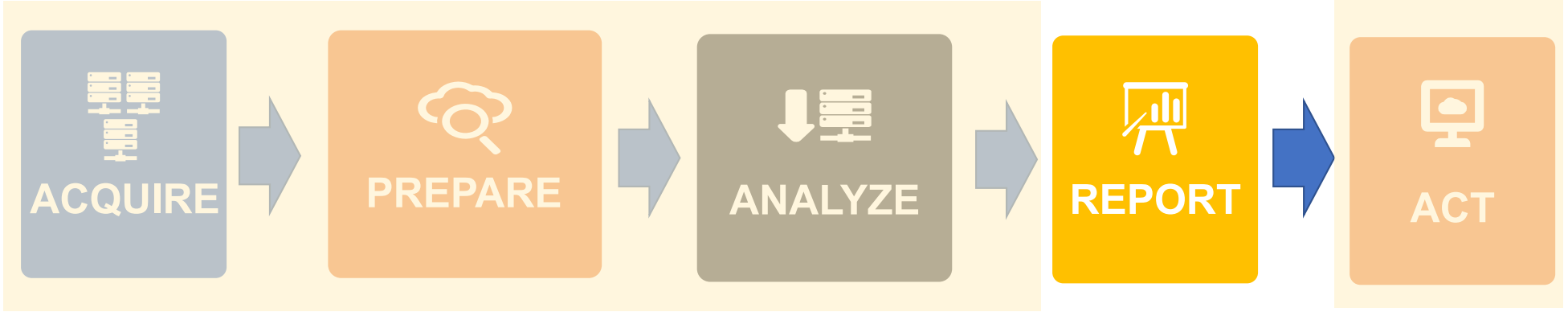
Clean | Integrate | Package



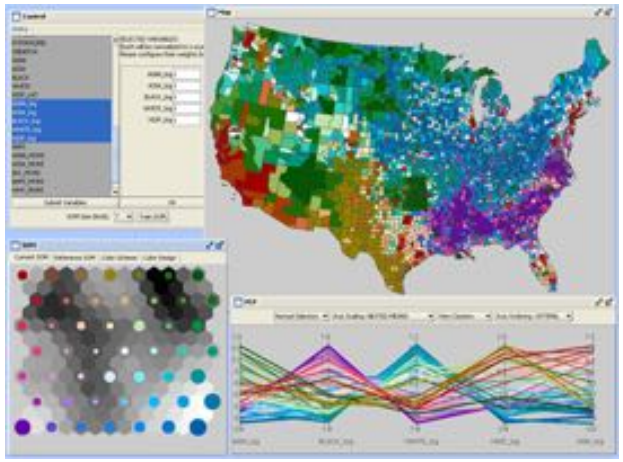
Step 3: Analyze Data

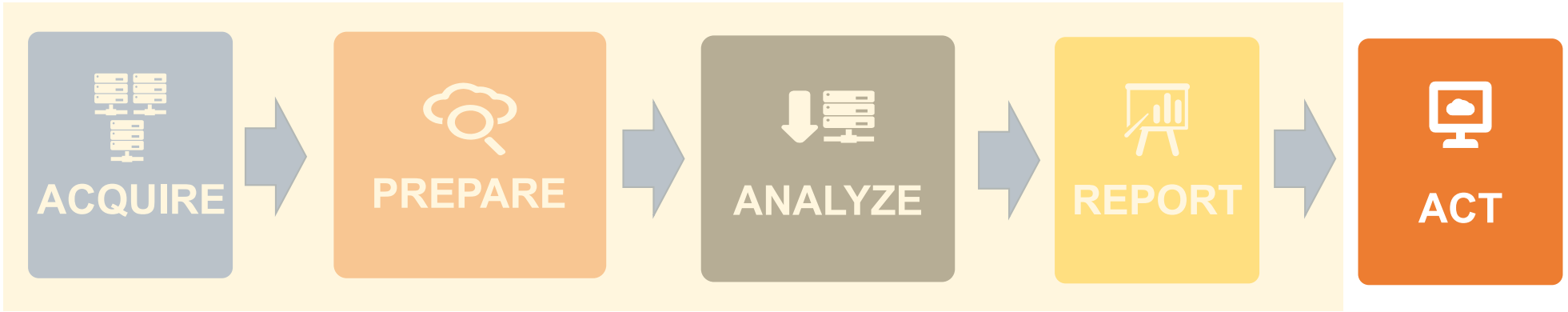
Select analytical techniques

Build models

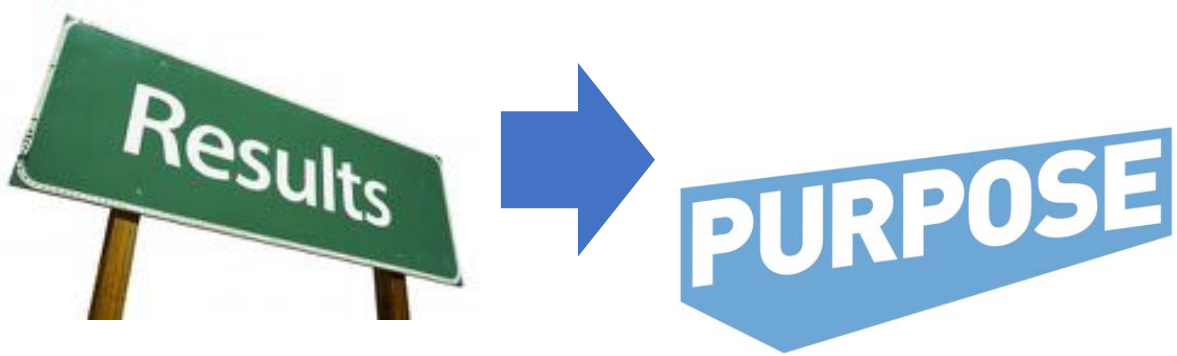


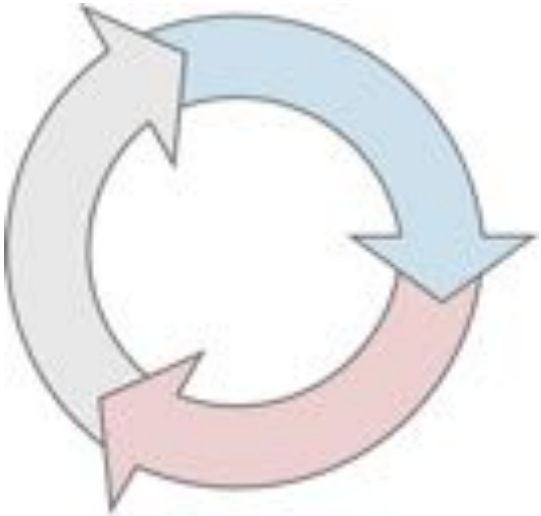
Step 4: Communicate Results





Step 5: Apply Results





Iterative process

Python for Data Science

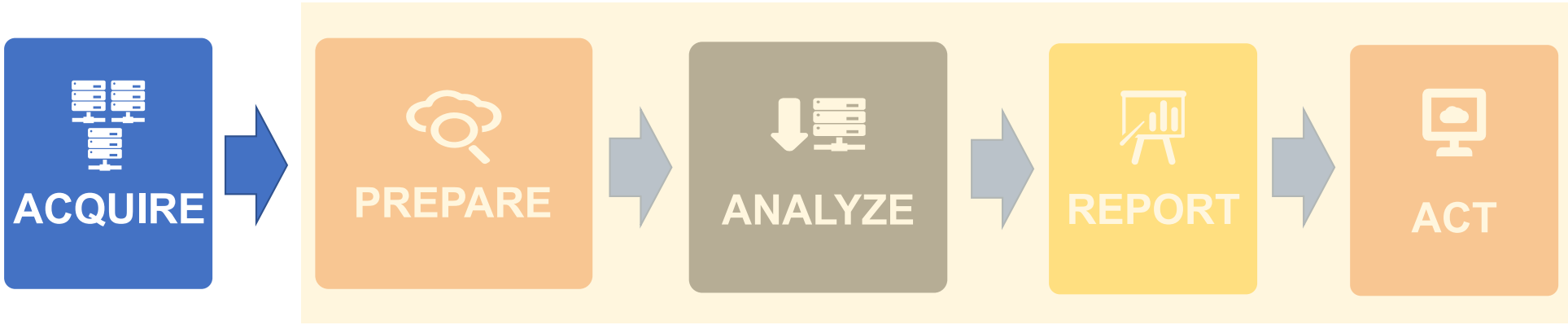
Step 1: Acquiring Data

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

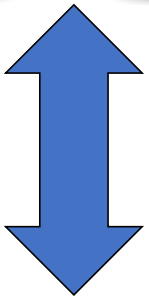
Where's the data?

- Identify suitable data
- Acquire all available data

Data comes from many places…

…with many ways to access it

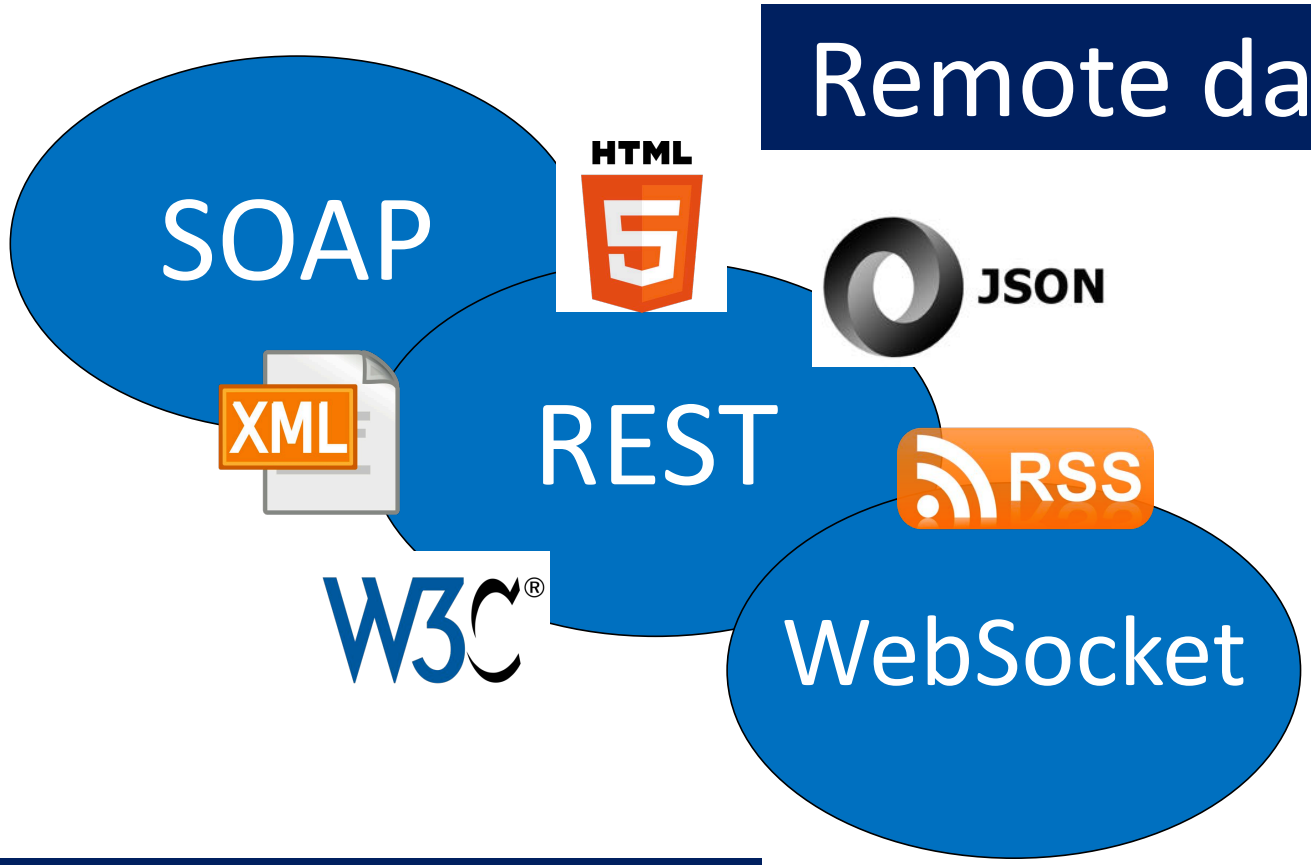
Traditional databases

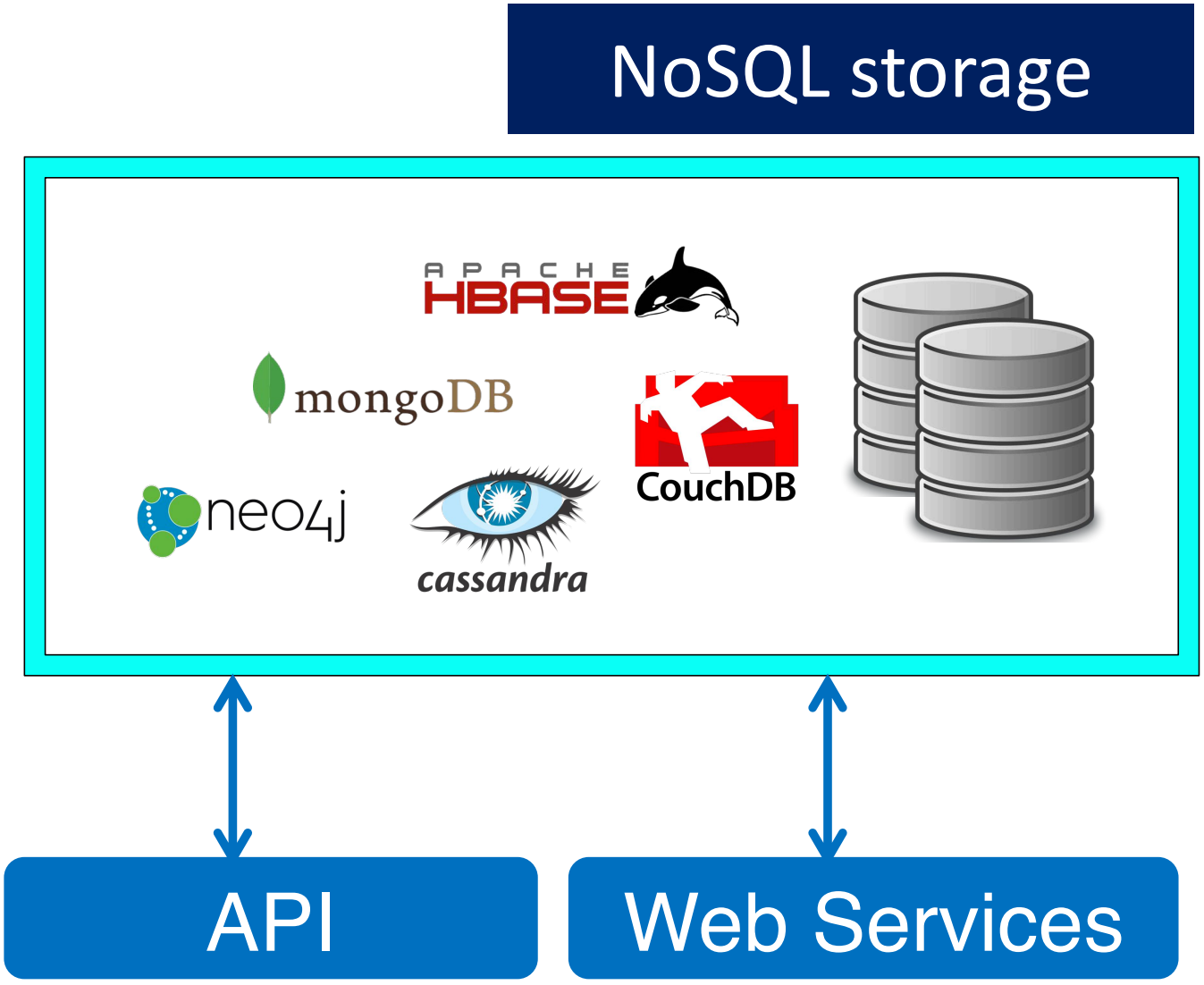


SQL and query browsers

Text files

Scripting languages


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


NoSQL storage
APACHE HBASE
mongoDB
CouchDB
neo4j
cassandra
API
Web Services

Acquiring data related to wildfires…

Historical weather

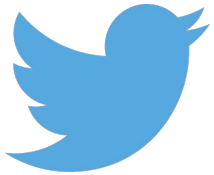
SQL

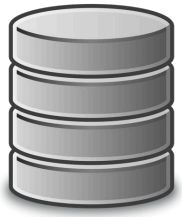
Current weather

WebSocket

Real-time tweets near fires

REST





Traditional databases

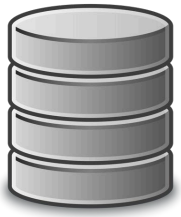
SQL and query browsers

Remote data

Web Services

Text files

Scripting languages



NoSQL storage

Web Services

Programming Interfaces

Python for Data Science

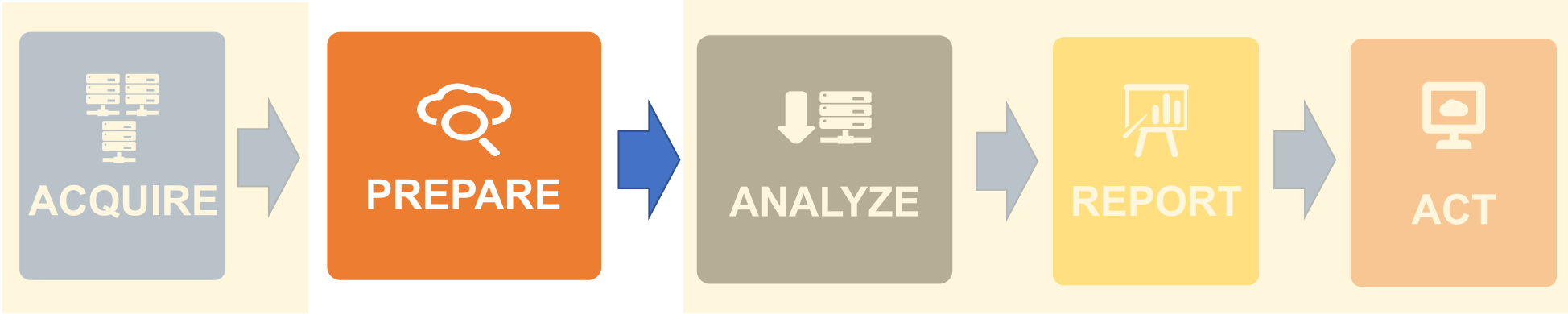
Step 2-A: Exploring Data

Dr. Ilkay Altintas and Dr. Leo Porter

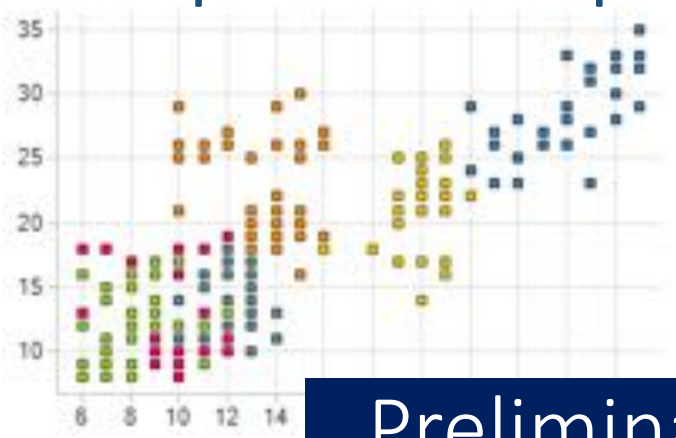
Twitter: #UCSDpython4DS

After this video you will be able to..

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

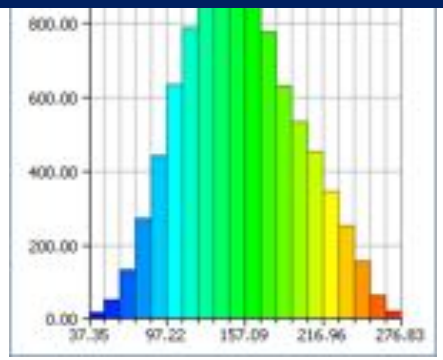


Step 2-A: Explore Data



Preliminary analysis

Understand nature of data

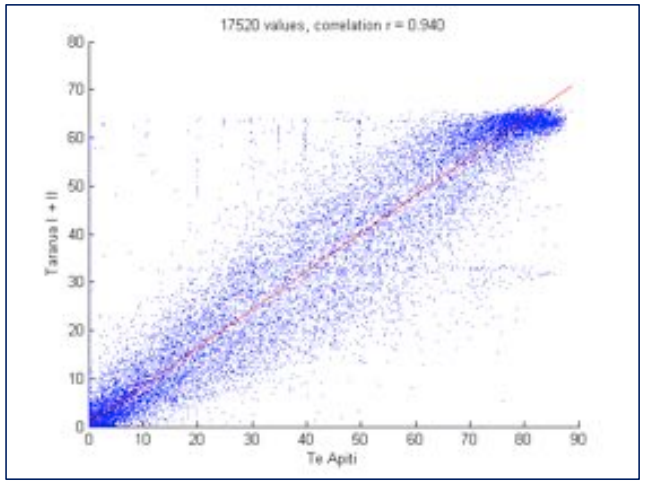


Why explore?

Goal: Understand your data

Why explore?

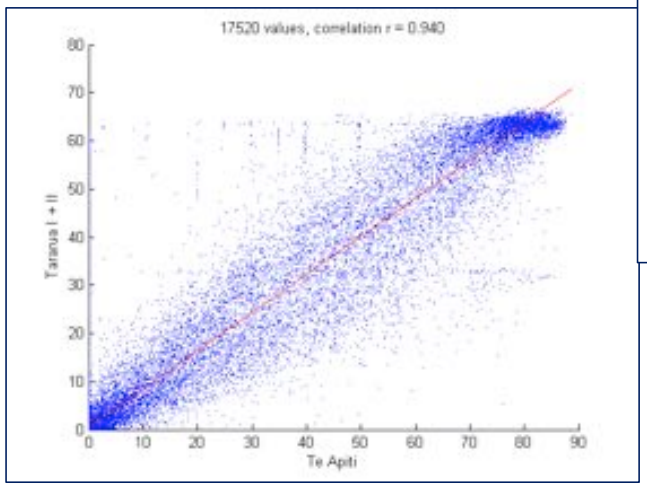
Correlations



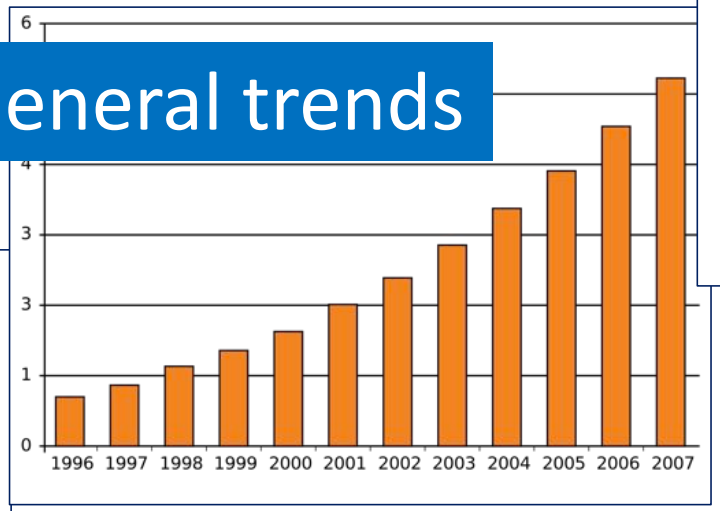
Python for Data Science

Why explore?

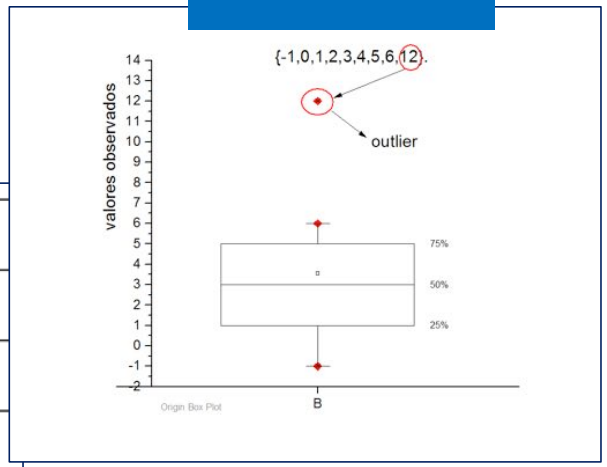
Correlations



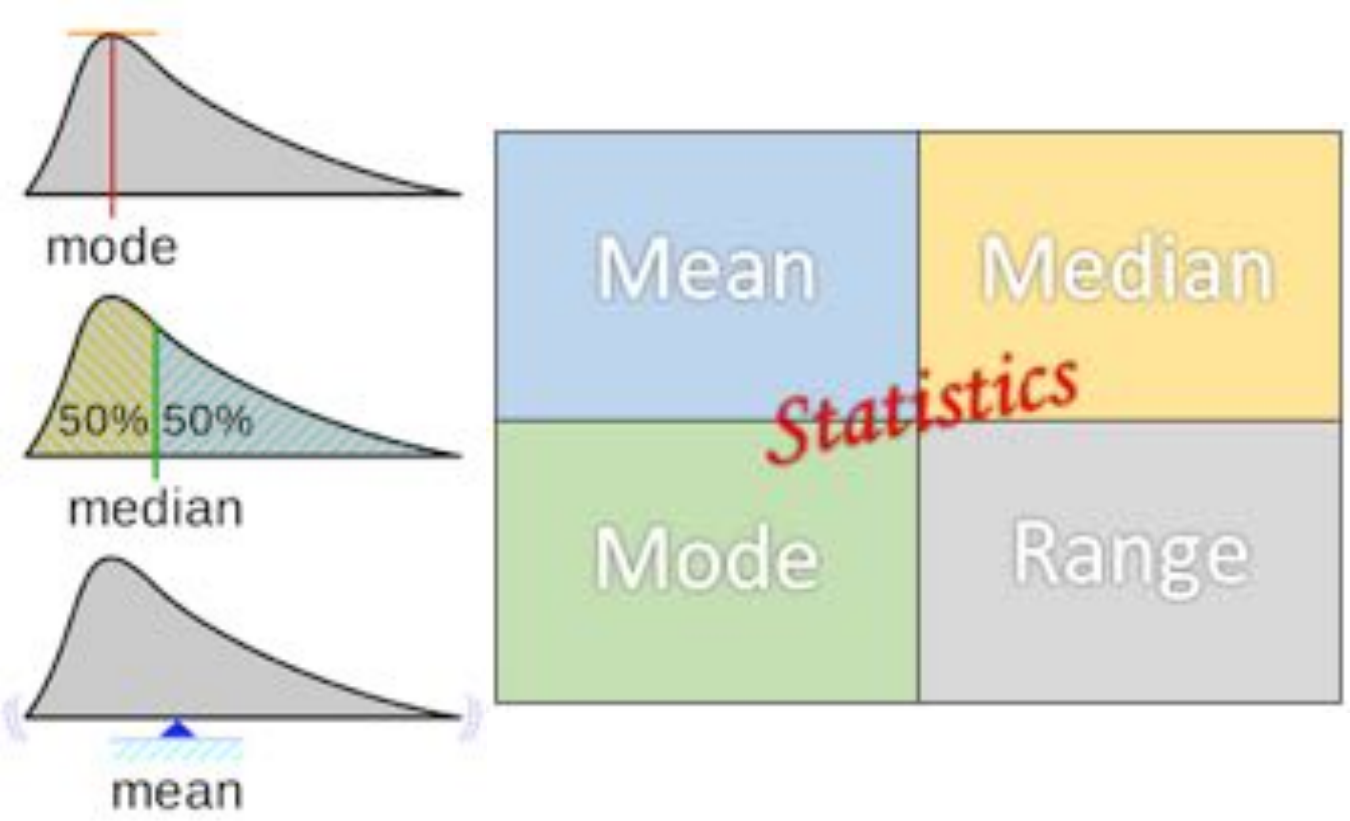
General trends



Outliers



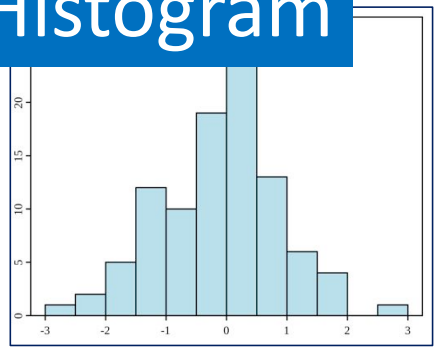
Describe Your Data



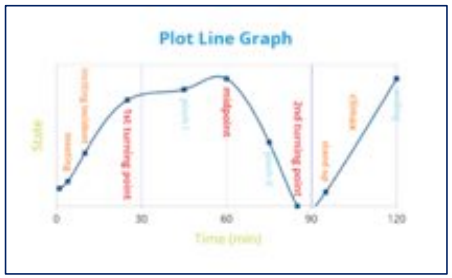
Python for Data Science

Visualize Your Data

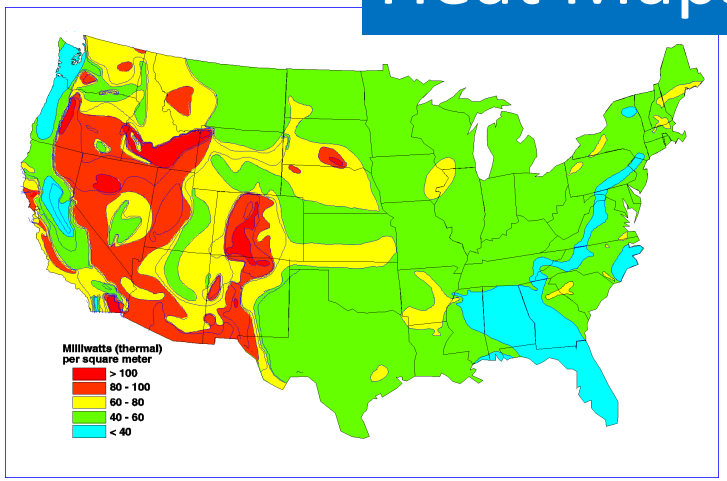
Histogram



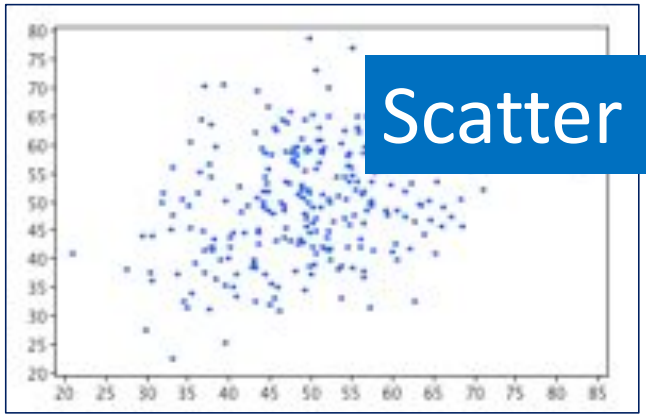
Line graphs



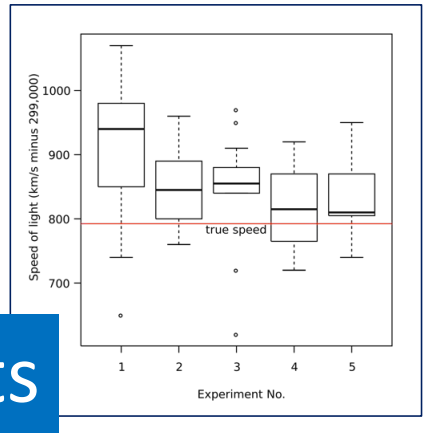
Heat Maps

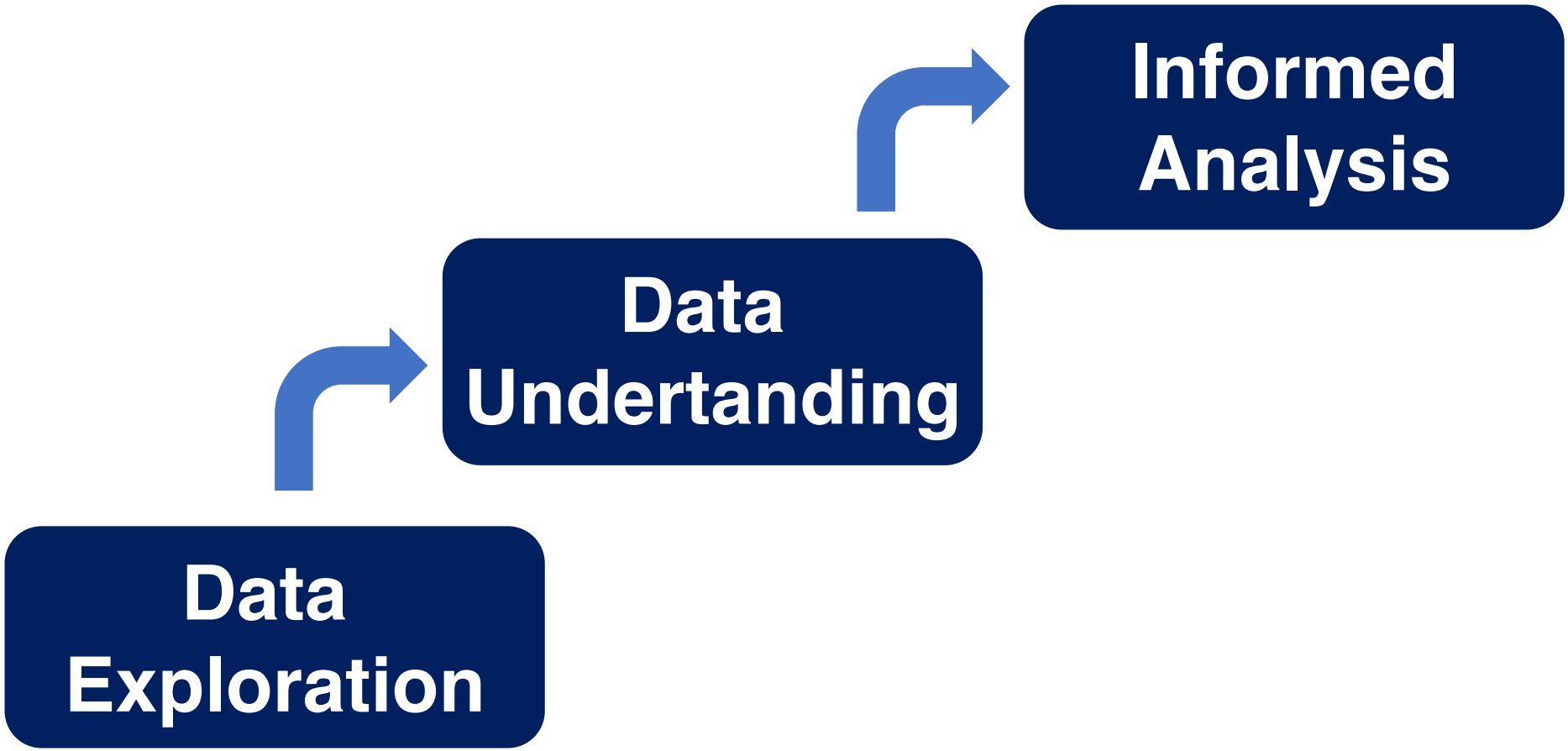


Scatter plots



Boxplots




Data Exploration
Data Undertanding
Informed Analysis

Python for Data Science

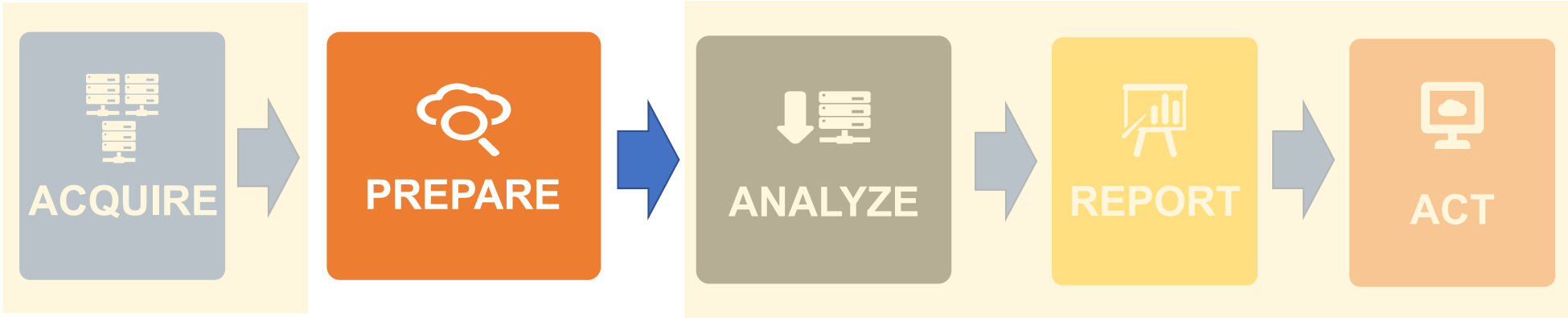
Step 2-B: Pre-processing Data

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

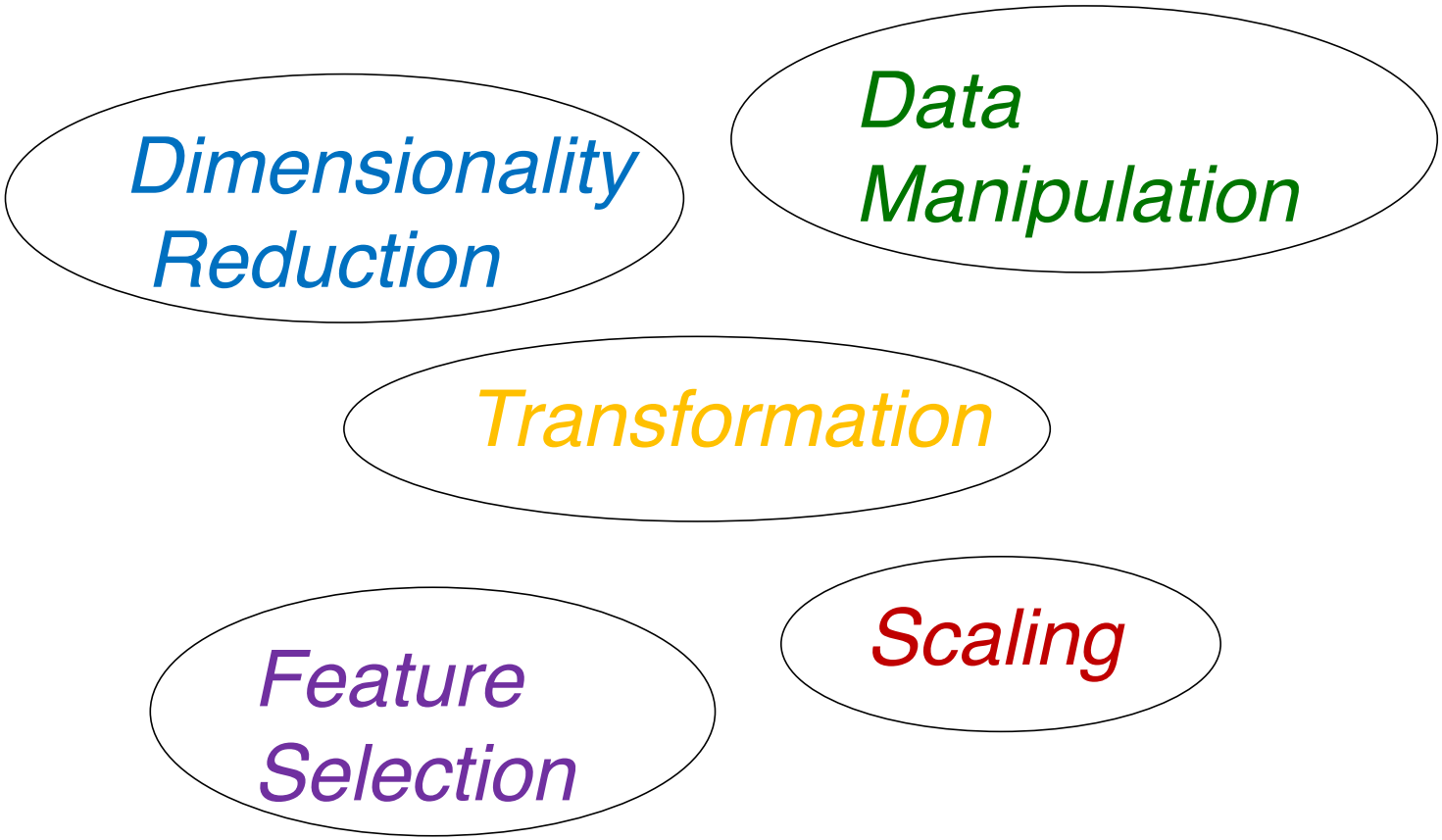
Getting Data in Shape

Data Munging

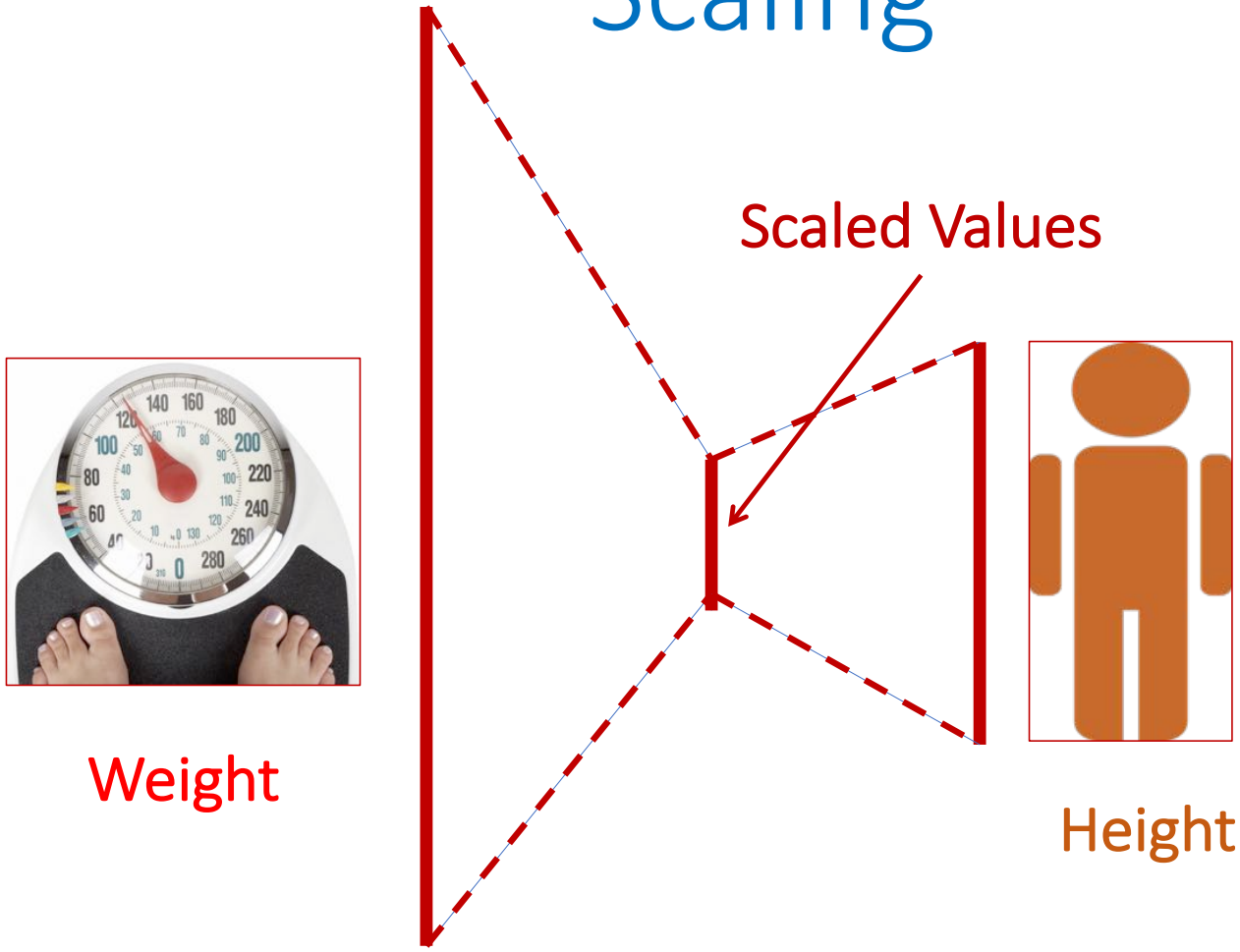
Data Preprocessing

Data Wrangling

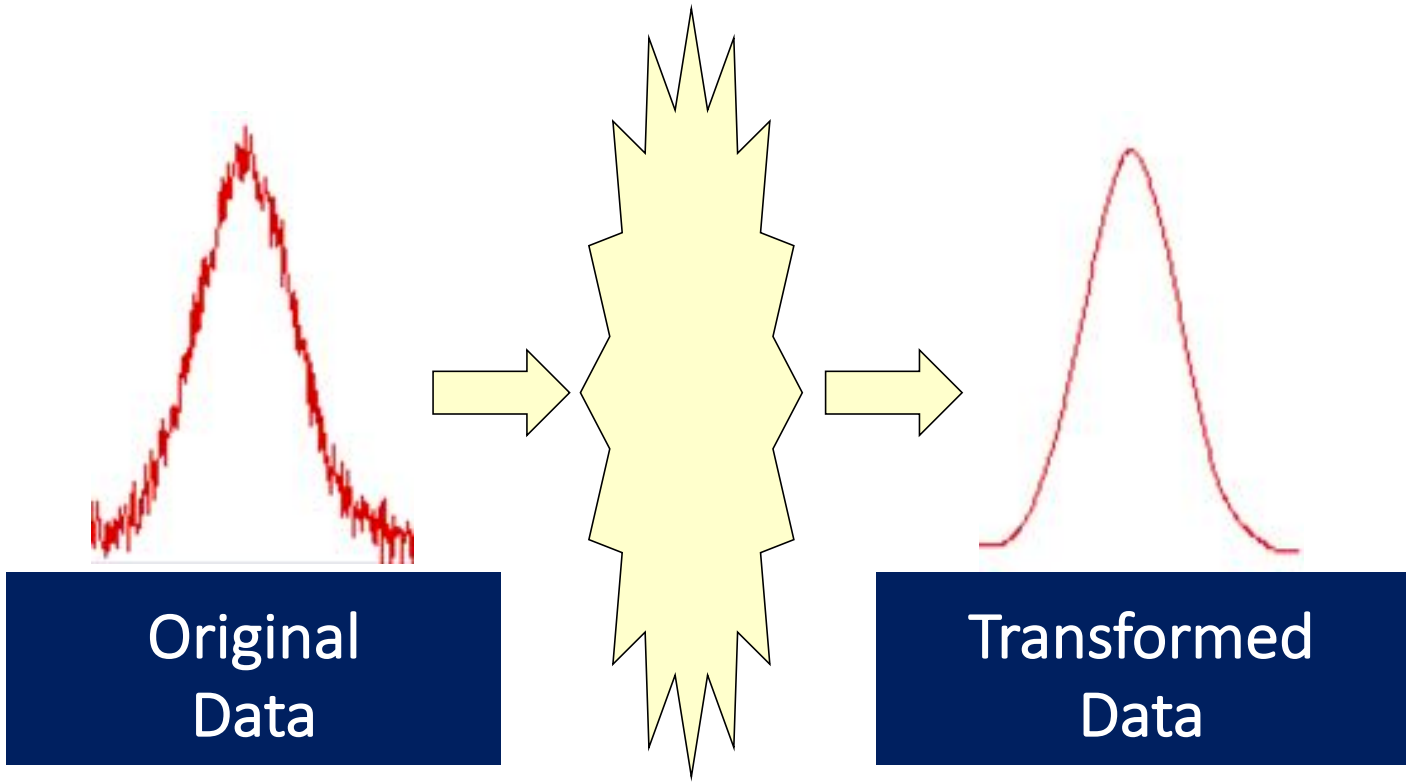
Data Munging



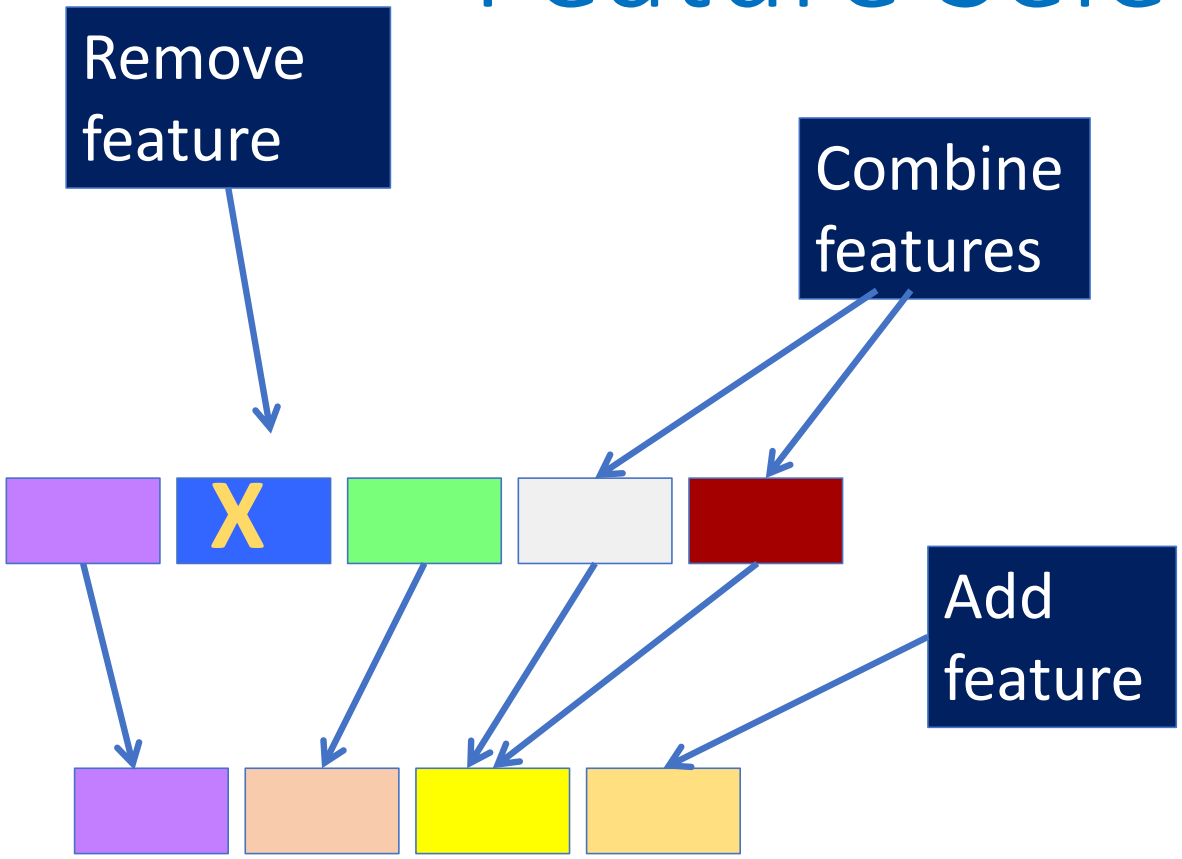
Scaling



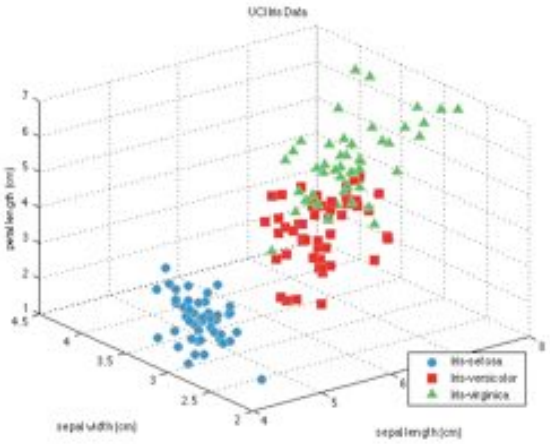
Transformation



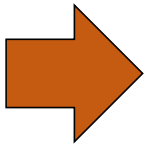
Feature Selection

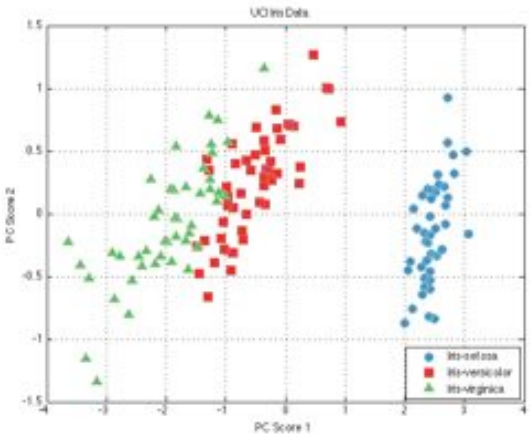


Dimensionality Reduction



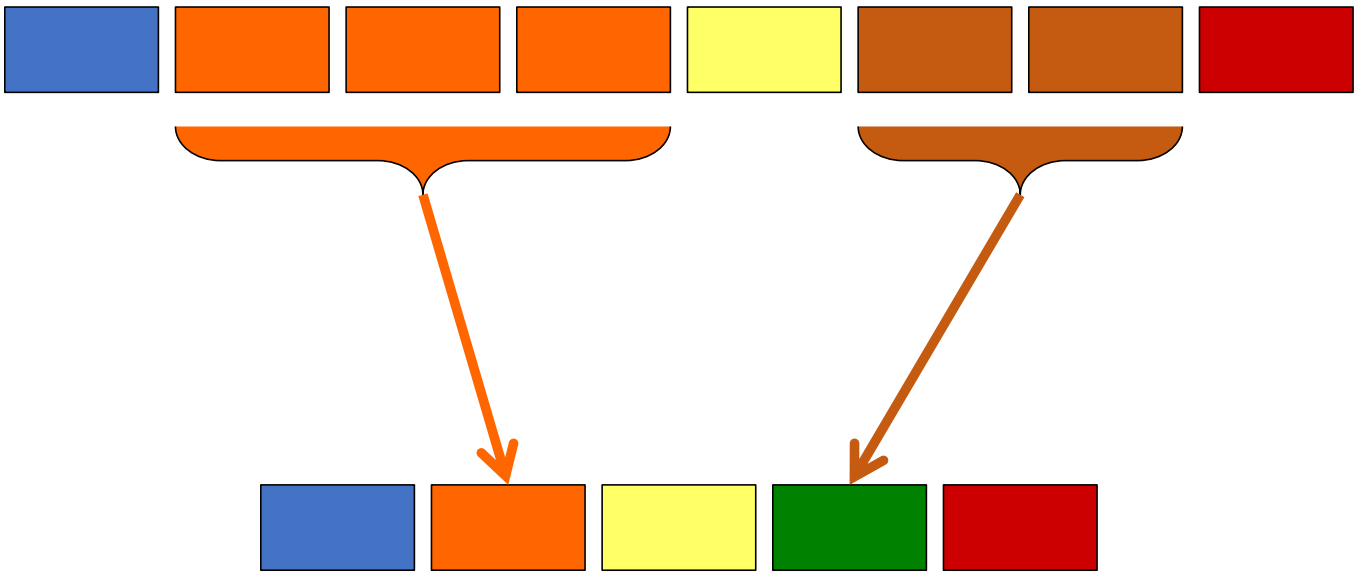
3D





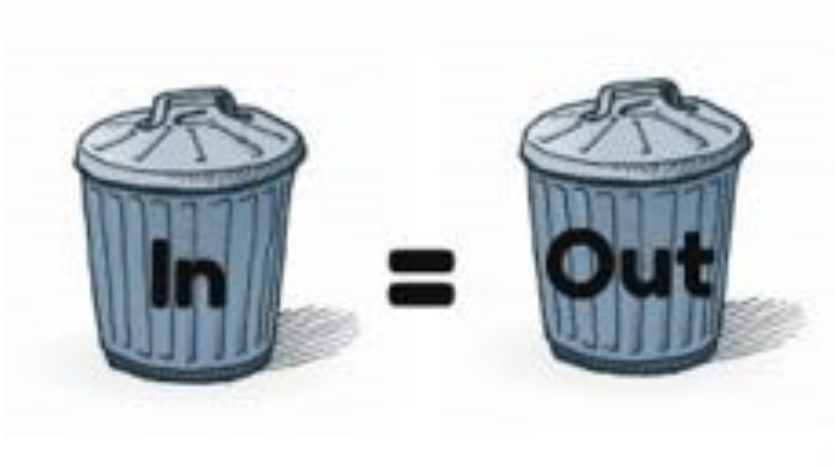
2D

Data Manipulation



Always Remember!

Garbage in = Garbage out



Data preparation is very important for meaningful analysis!

Python for Data Science

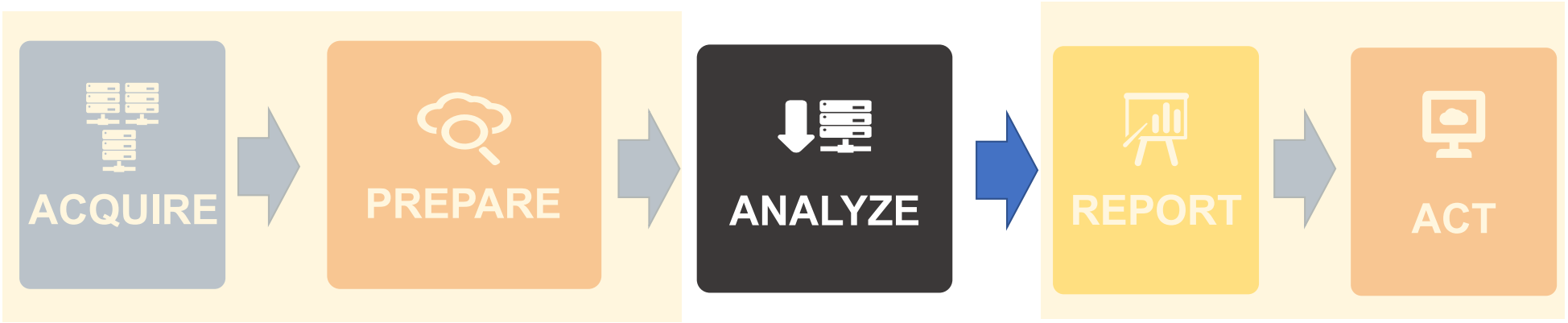
Step 3: Analyze Data

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

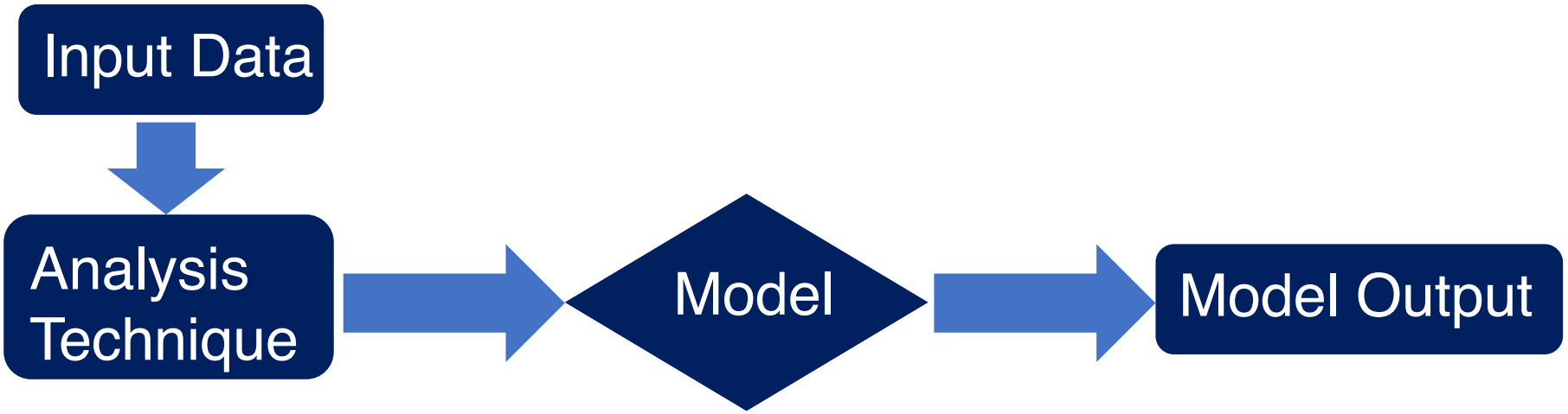


Step 3: Analyze Data

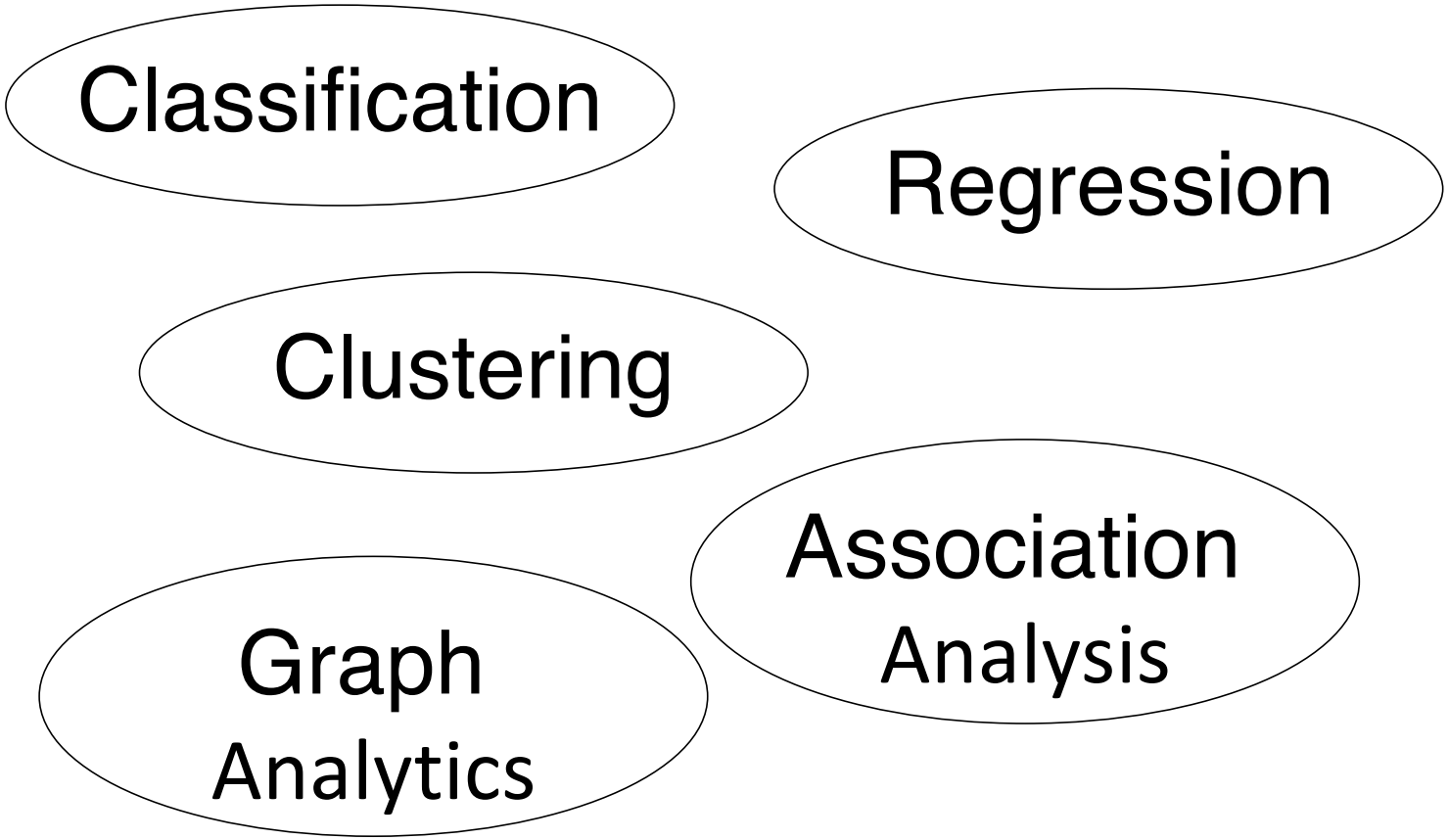
Select analytical techniques

Build models

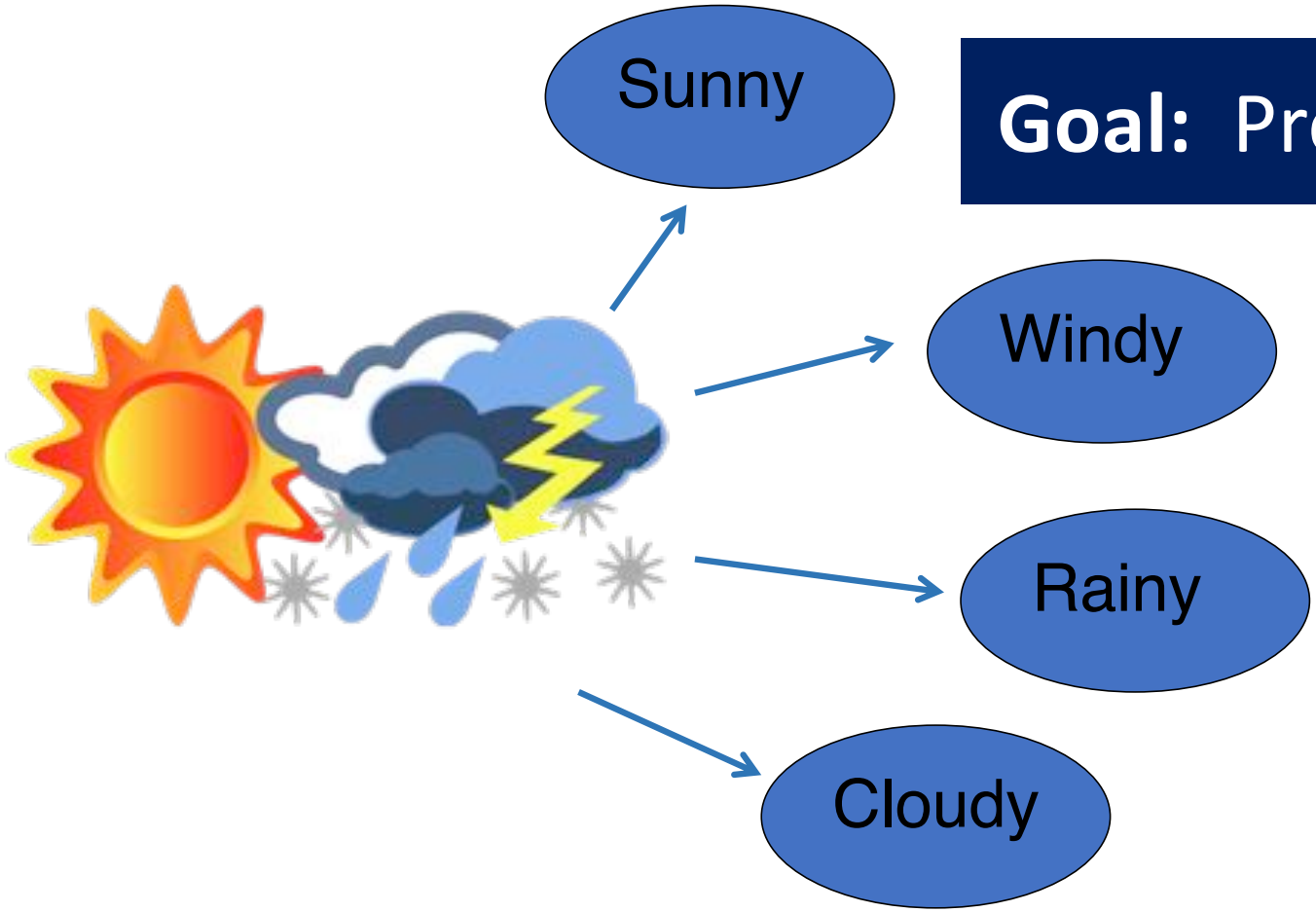
Build Model



Categories of Analysis Techniques



Classification



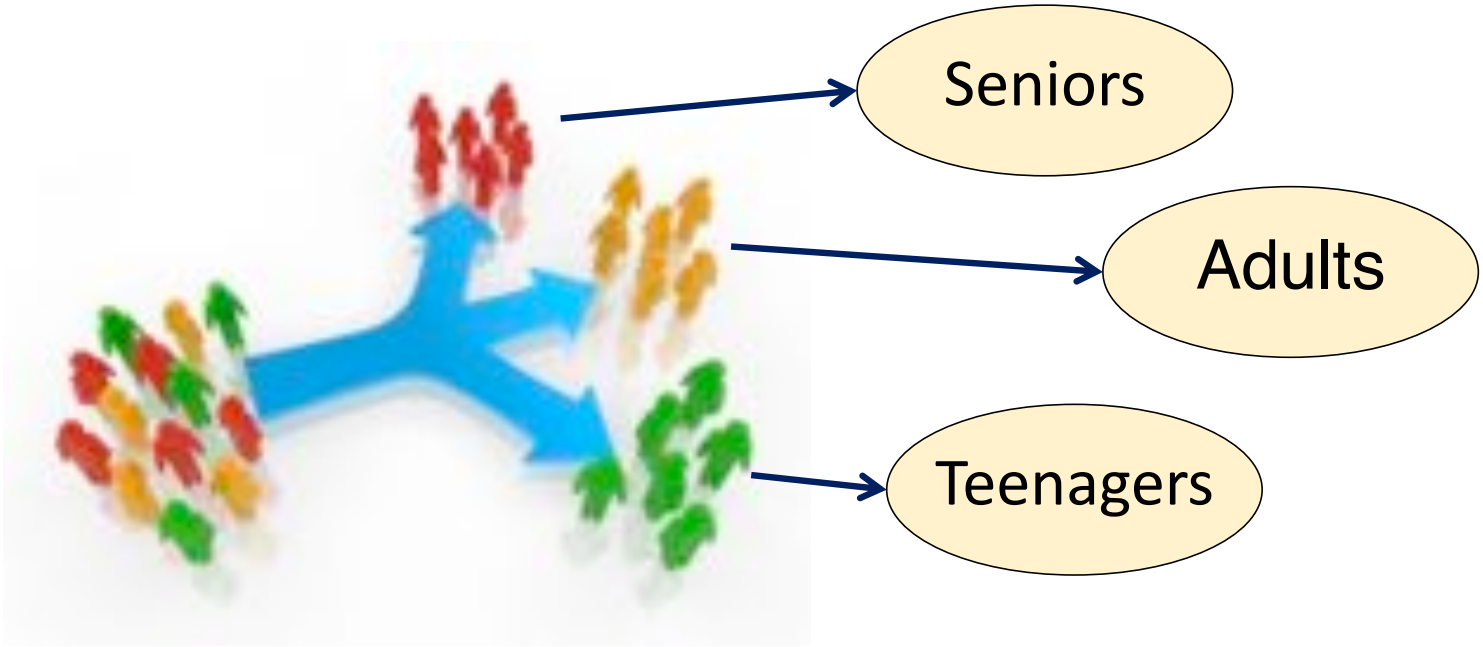
Goal: Predict category

Regression

Goal: Predict numeric value

Clustering

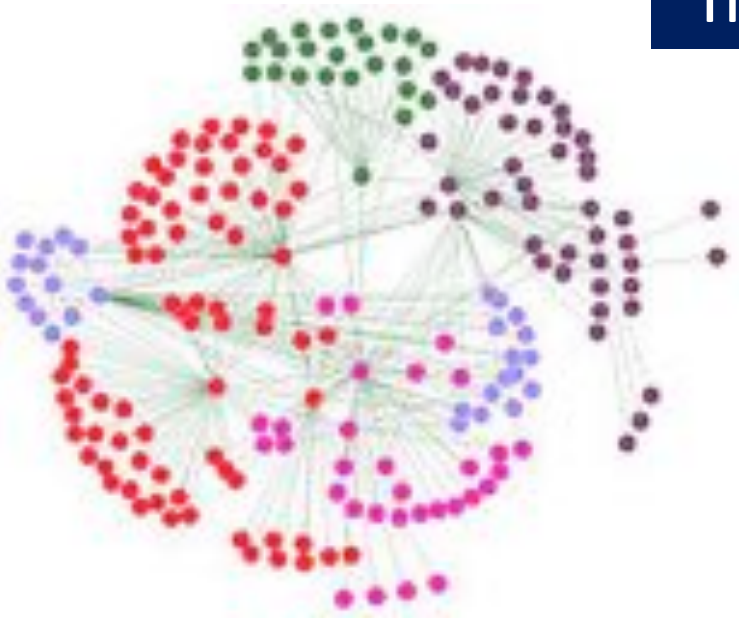
Goal: Organize similar items into groups



Association Analysis

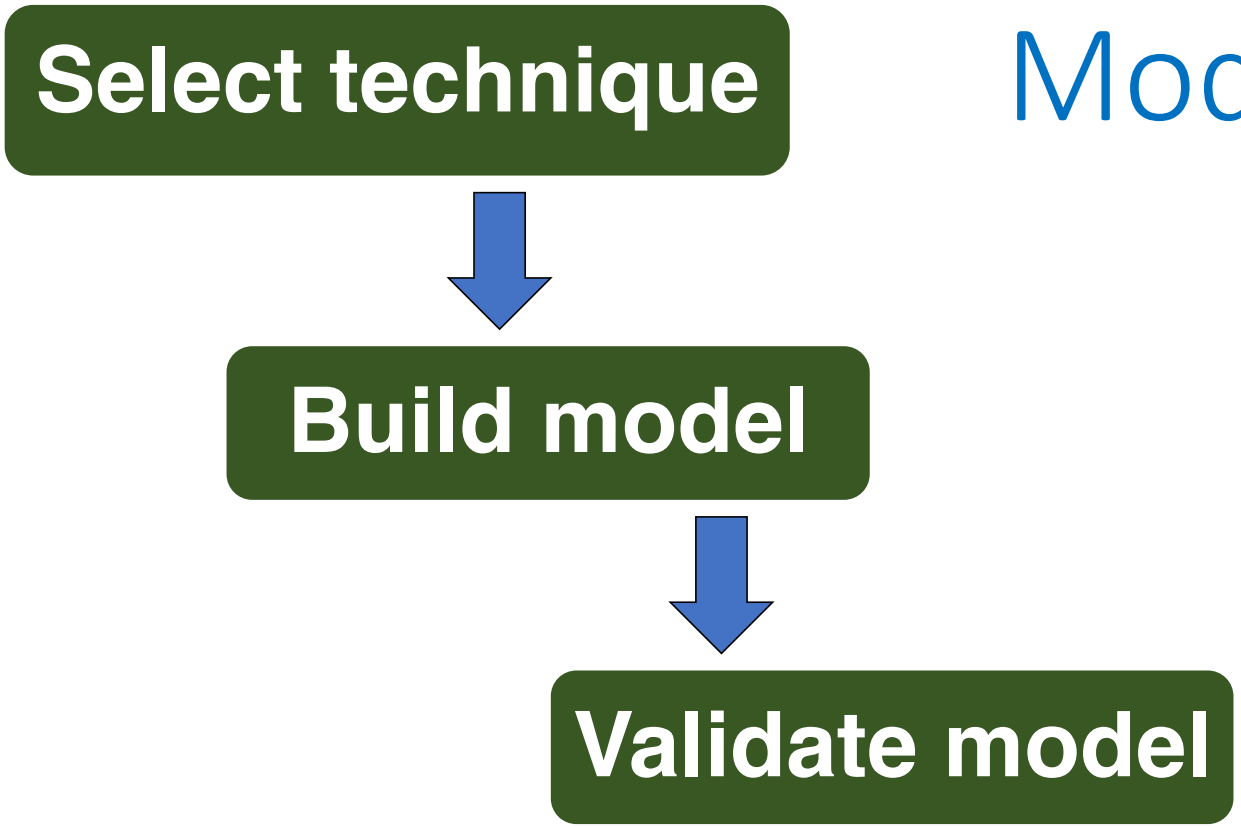
Goal: Find rules to capture associations between items

Graph Analytics



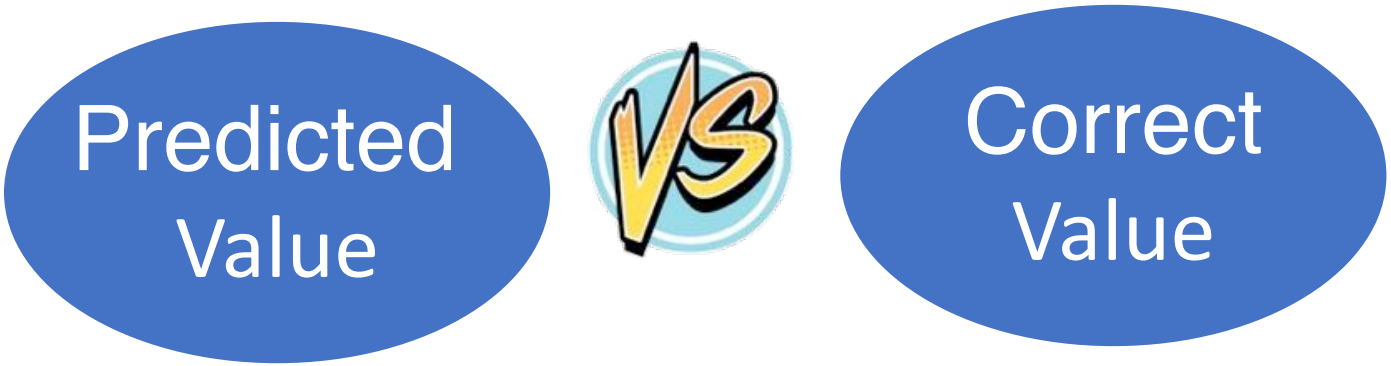
Goal: Use graph structures to find connections between entities

Modeling



Evaluation of Results

Classification and Regression



Clustering

Association Analysis and 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

Python for Data Science

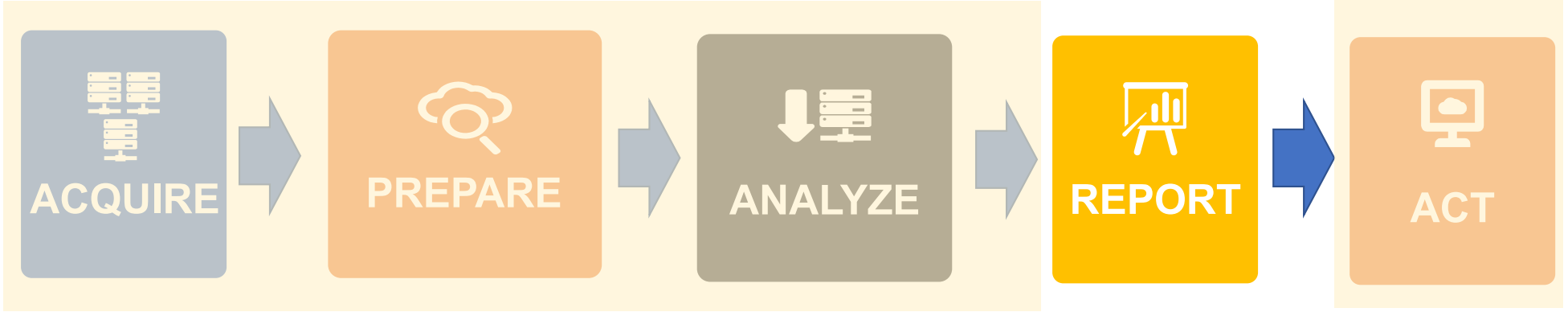
Step 4: Reporting Insights

Dr. Ilkay Altintas and Dr. Leo Porter

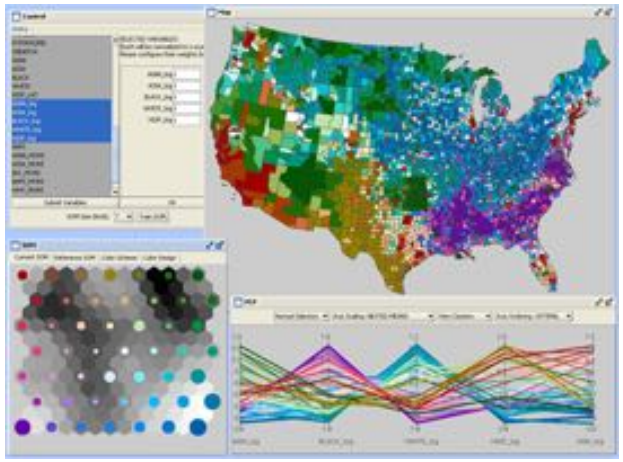
Twitter: #UCSDpython4DS

After this video you will be able to..

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

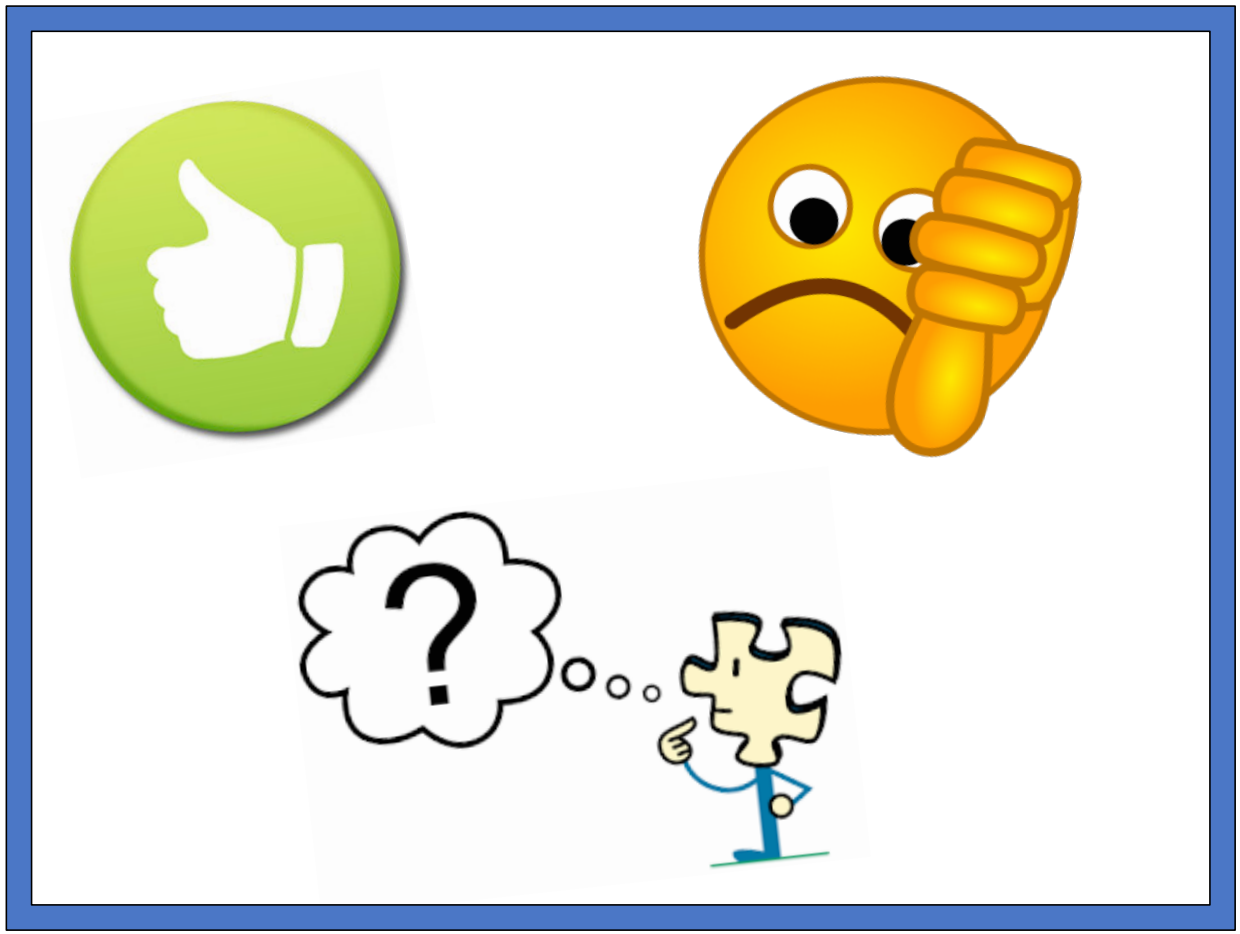


Step 4: Communicate Results



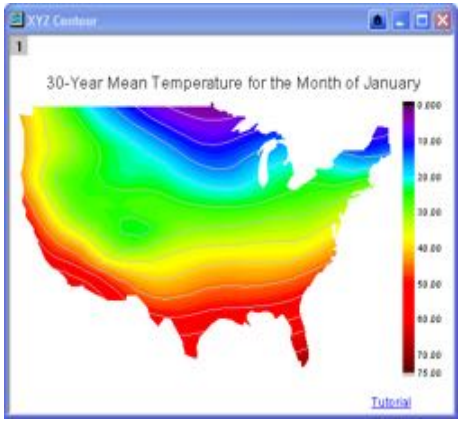
What to Present

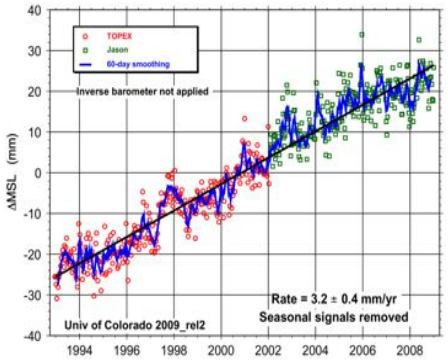
What to Present



Python for Data Science

How to Present

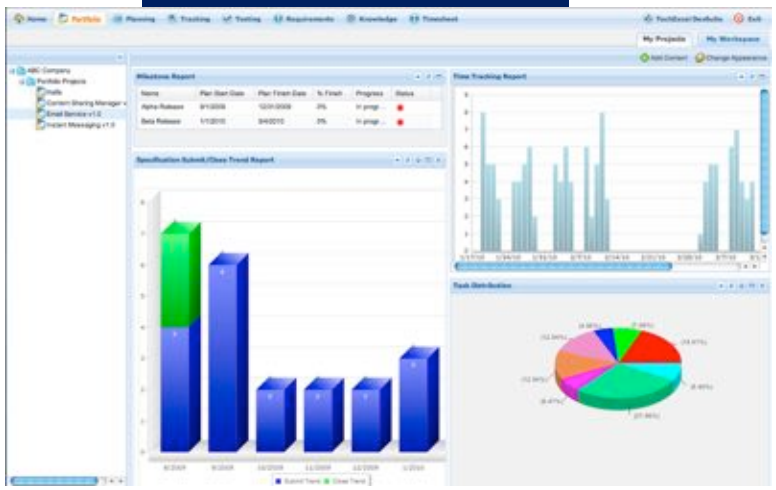




Visualization Tools

Present

with



using

tableau

Google Developers CHARTS

Python for Data Science

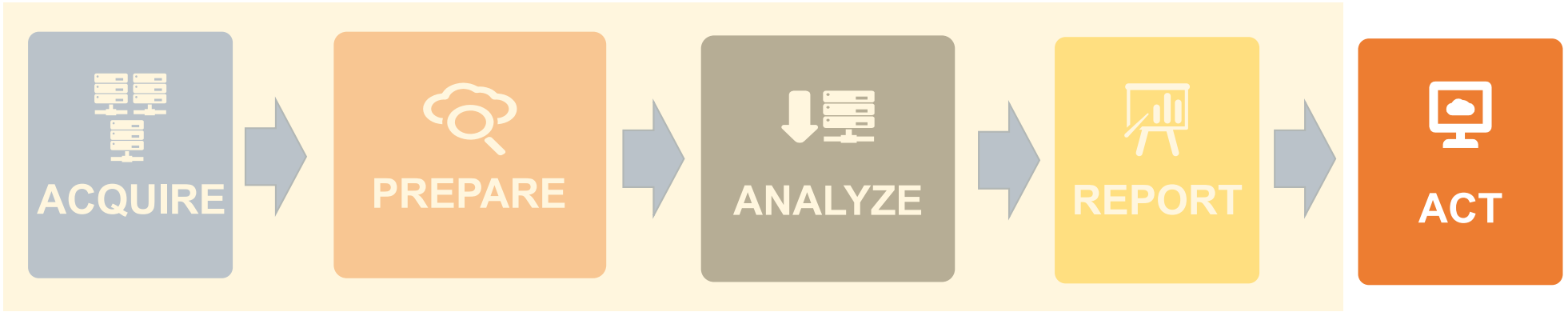
Step 5: Turning Insights into Action

Dr. Ilkay Altintas and Dr. Leo Porter

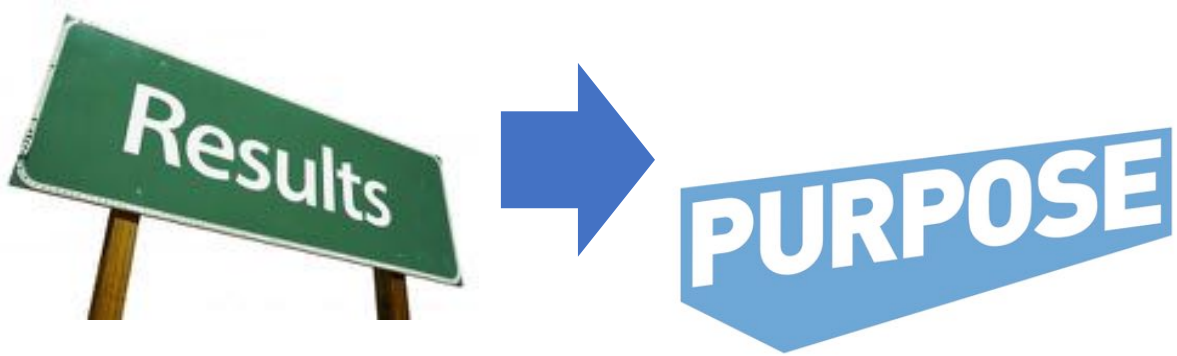
Twitter: #UCSDpython4DS

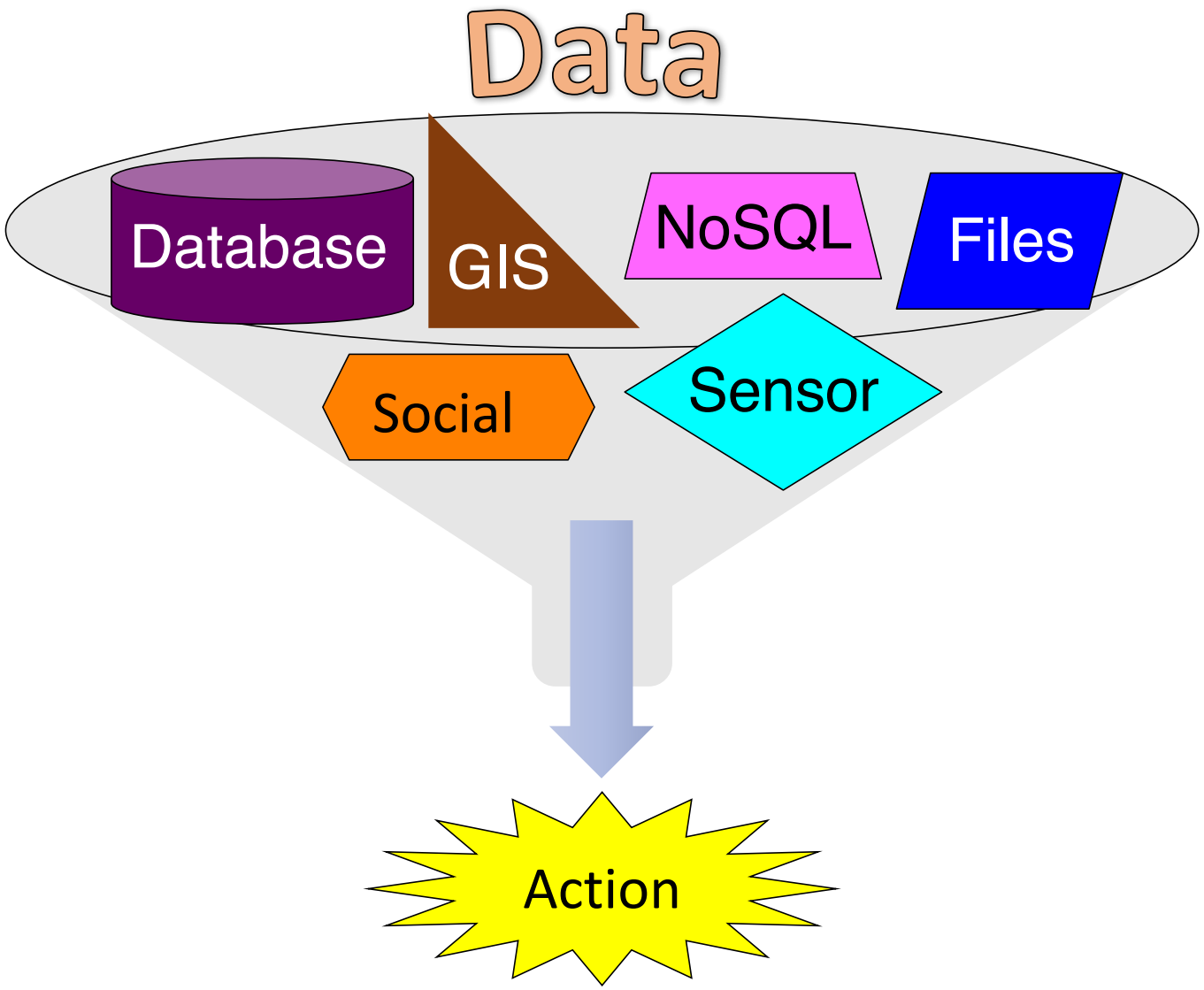
After this video you will be able to..

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

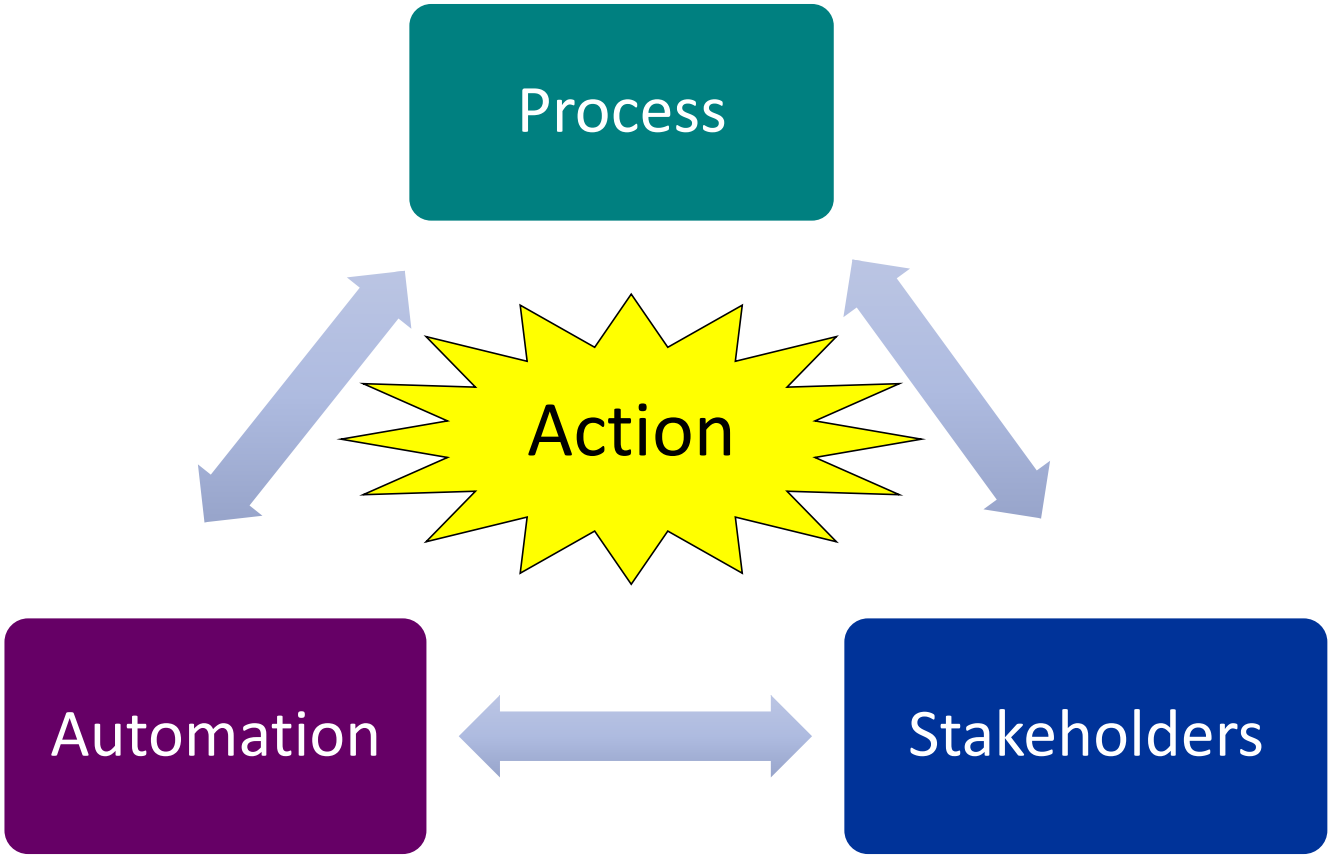


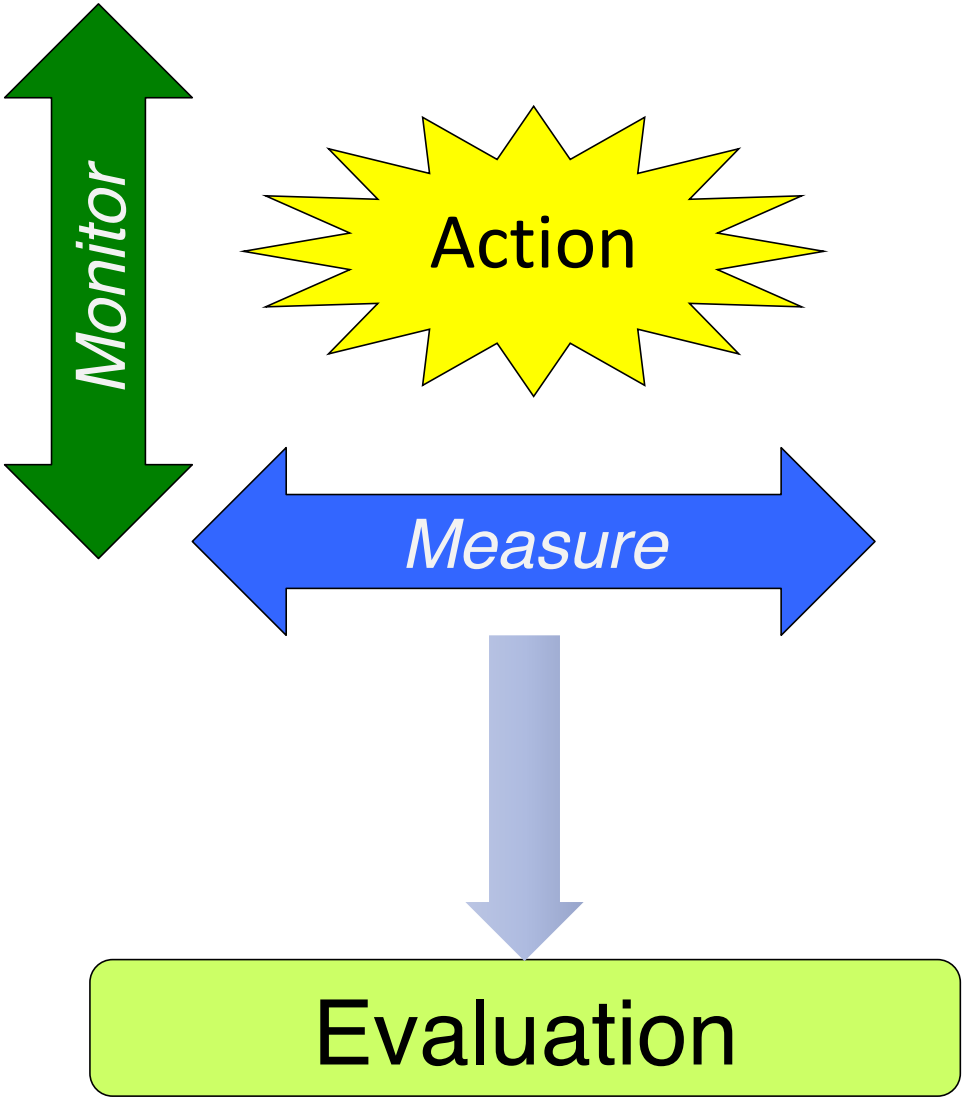
Step 5: Apply Results




Data
Database
GIS
NoSQL
Files
Social
Sensor
Action

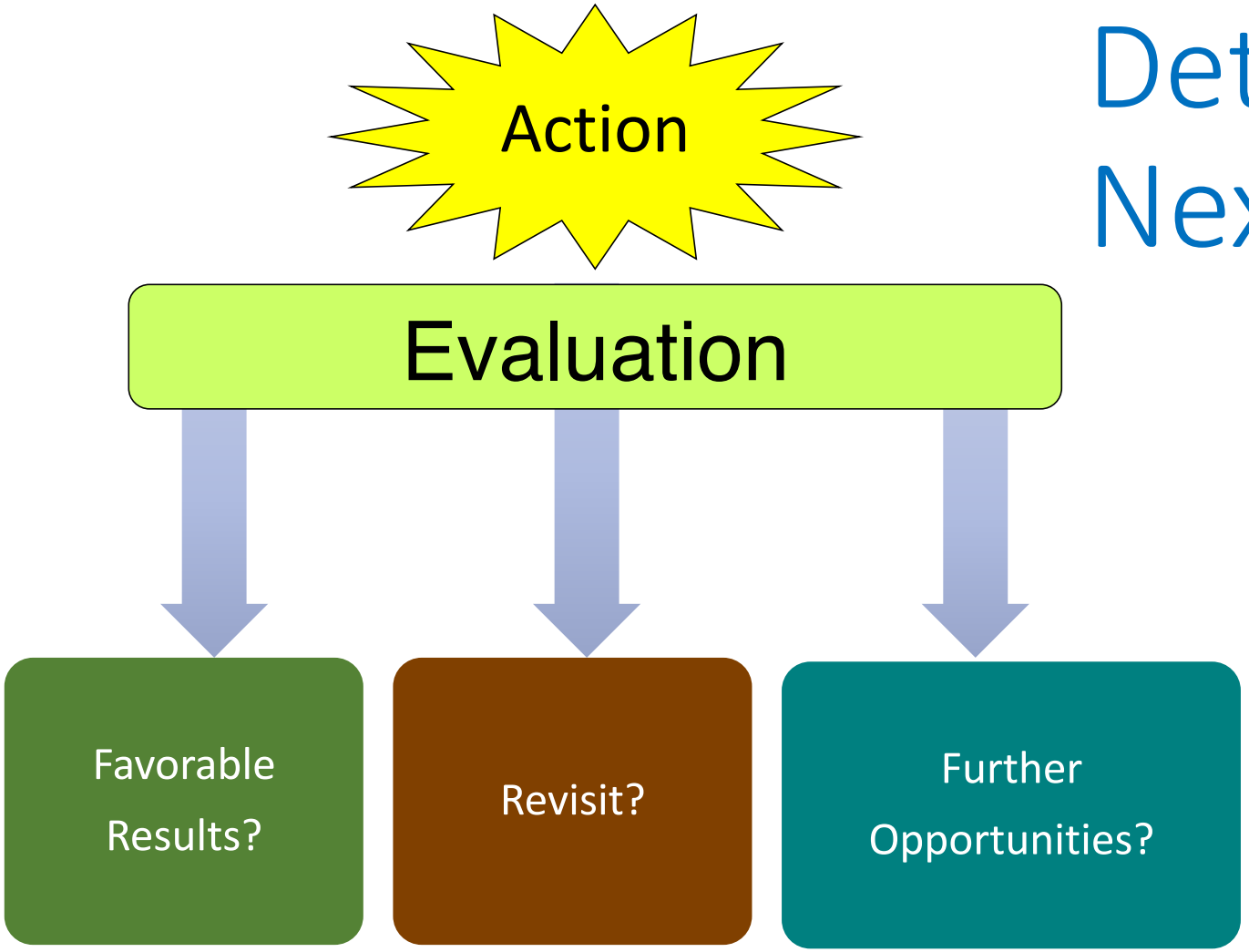
Implementation

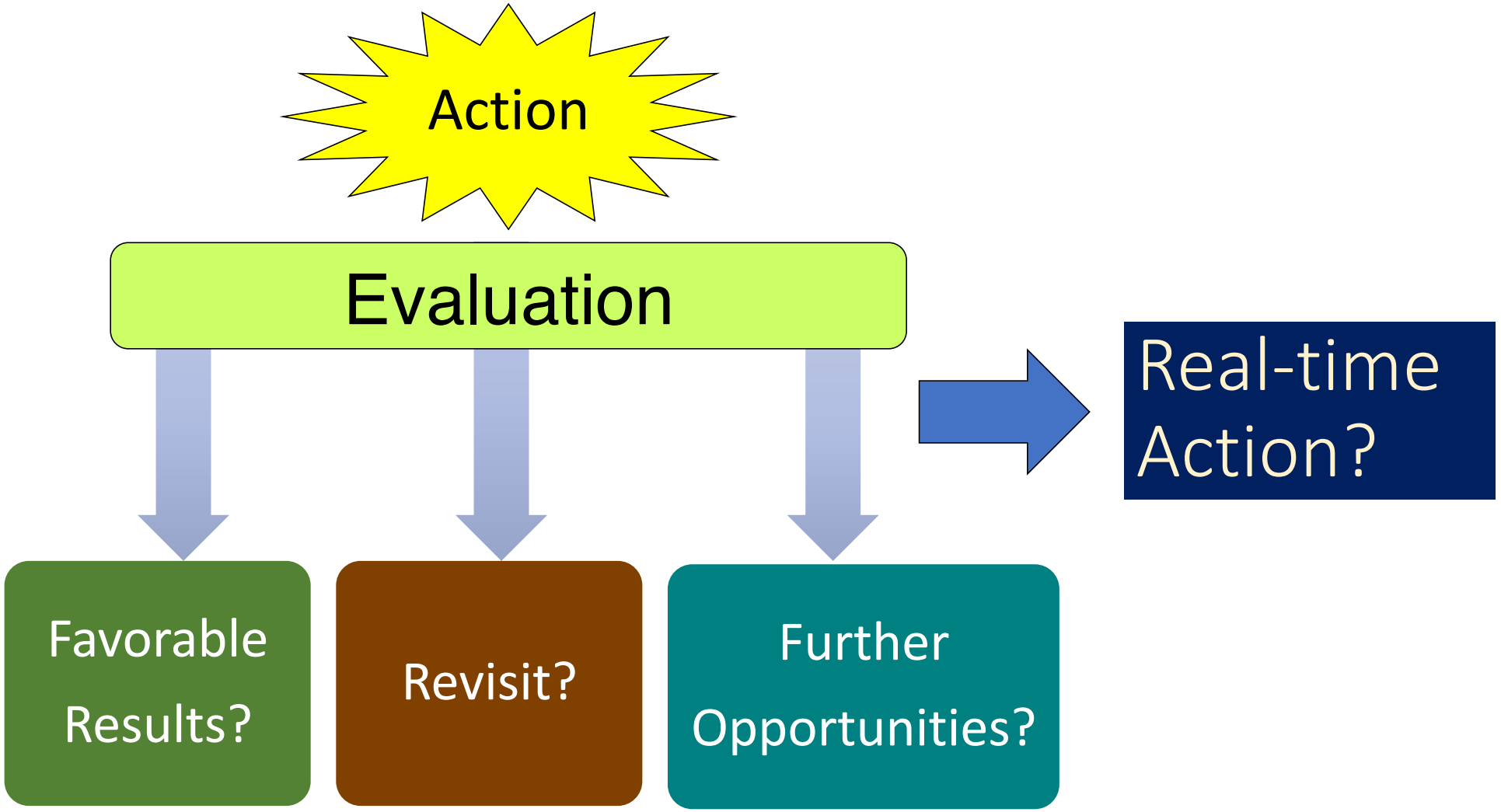




Assess Impact

Determine Next Steps




Action
Evaluation
Real-time Action?
Favorable Results?
Revisit?
Further Opportunities?