Data Analysis

Somkiat Puisungnoen

Somkiat | Home 17

Somkiat Puisungnoen

Update Info | View Activity Log 10+ | ...

Timeline | About | Friends 2,941 | Photos | More

When did you attend Ubon Ratchathani University?

14 Pending Items

Intro

Software Craftsmanship

- Software Practitioner at สยามชำนาญกิจ พ.ศ. 2556
- Agile Practitioner and Technical at SPRINT3r
- Software analyst at TARAD.com
- Software Developer at True Corporation
- Former Software Engineer at Opendream

Status | Photo/Video | Live Video | Life Event

What's on your mind?

Public | Post

Somkiat Puisungnoen
1 hr · Twitter ·

พรุ่งนี้ไป share เรื่อง NoSQL ที่มหาวิทยาลัย

Like | Comment | Share

Nuttachot Dusitanont, Chitpong Few Wuttanan and 50 others

somkiat.cc

Somkiat | Home

Page | Messages | Notifications | Insights | Publishing Tools | Settings | Help

somkiat.cc

@somkiat.cc

- Home
- About
- Events
- Photos
- Likes
- Videos
- Posts
- Services

Liked | Following | Share | ...

+ Add a Button

Write something...

Personal Blog

Page Tips — See All

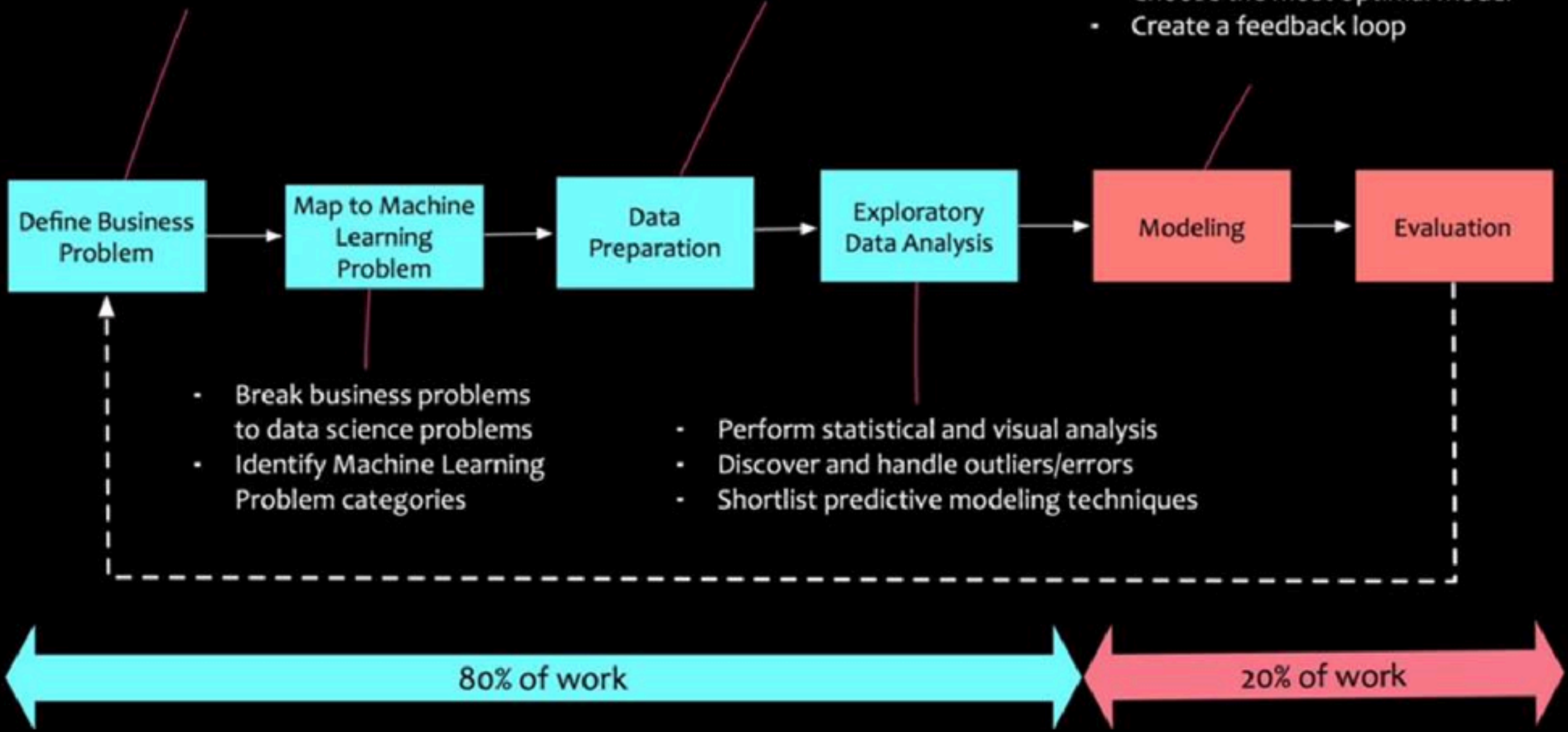
What's a Boosted Post?
A boosted post is the easiest way to reach more people on Facebook.

Data Science Process

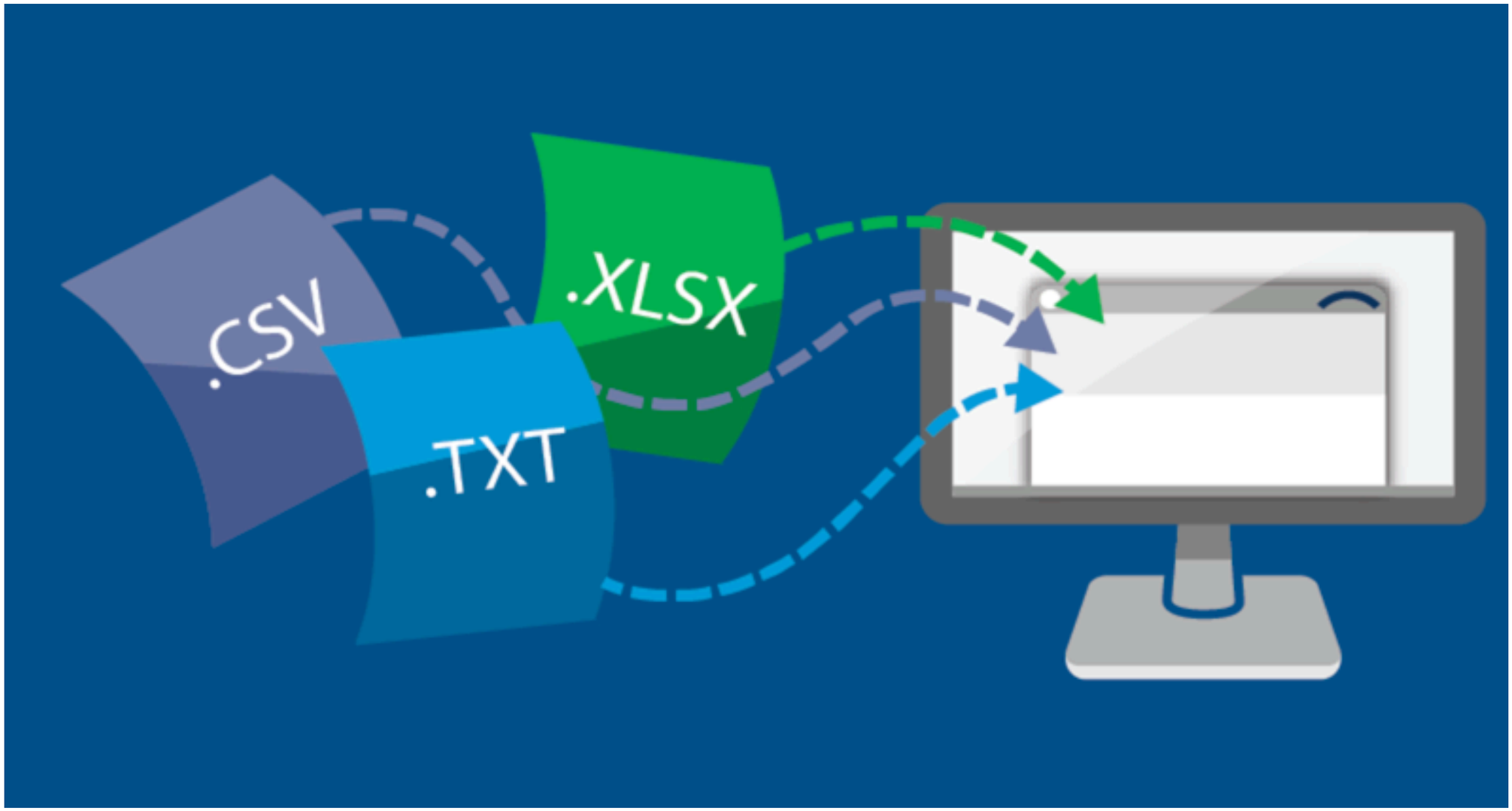
- Clearly defined business problem
- Set success criteria
- Define clear data science objectives

- Understand data points and constraints
- Formulate data analytics strategy
- Perform required transformation

- Experiment with multiple models
- Choose the most optimal model
- Create a feedback loop

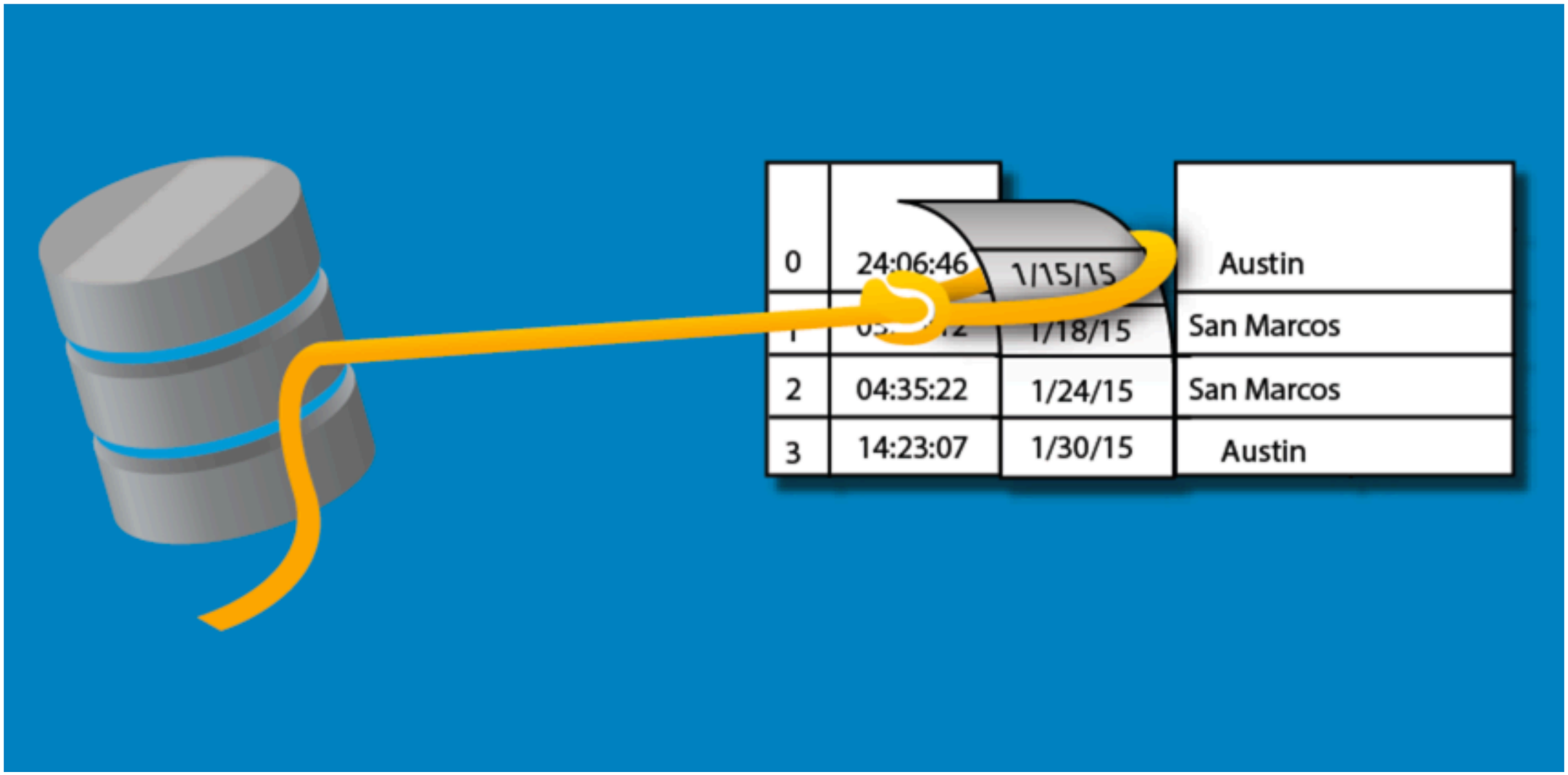


Access data from multiple source

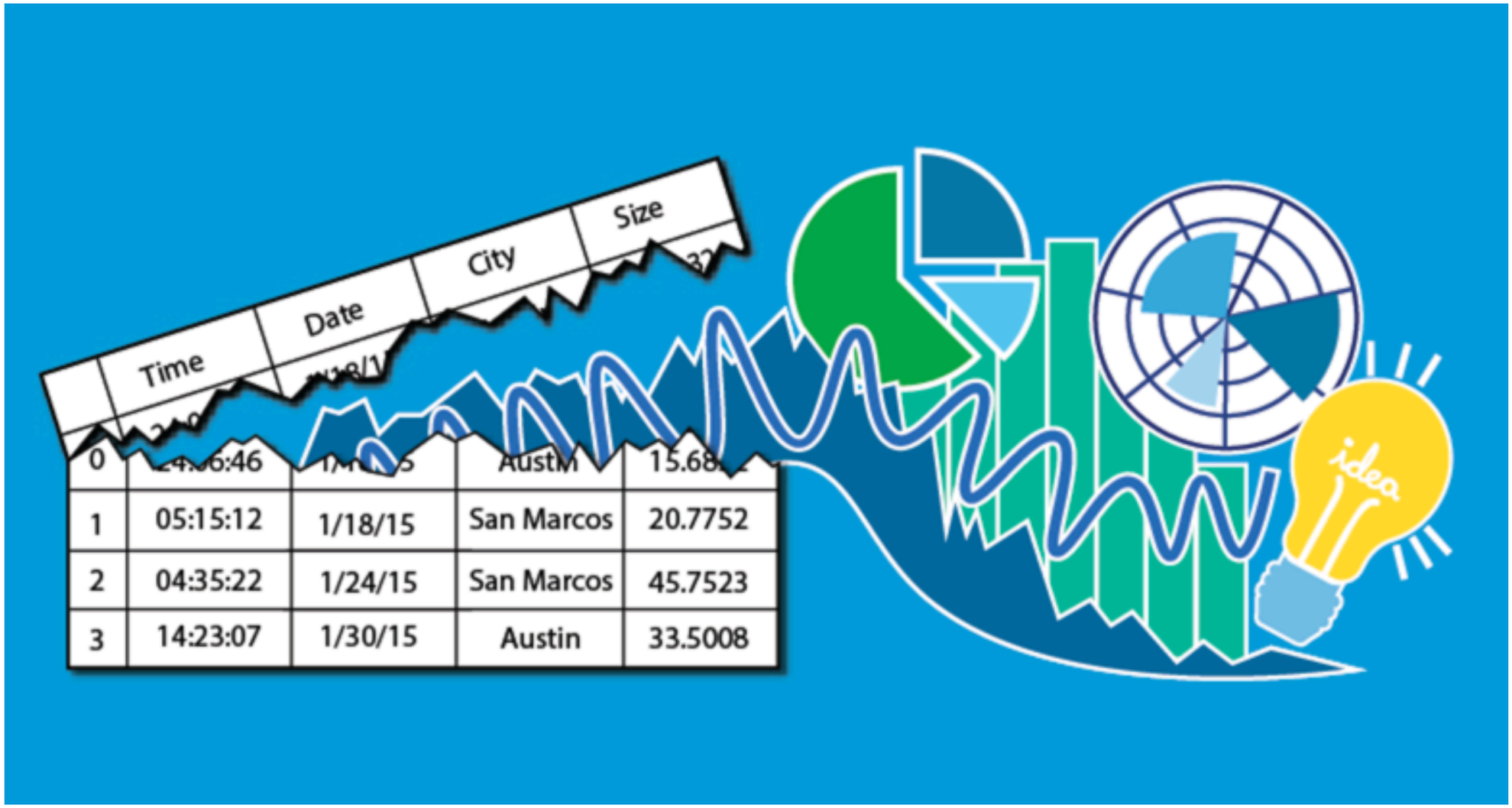


Cleaning and preparing data

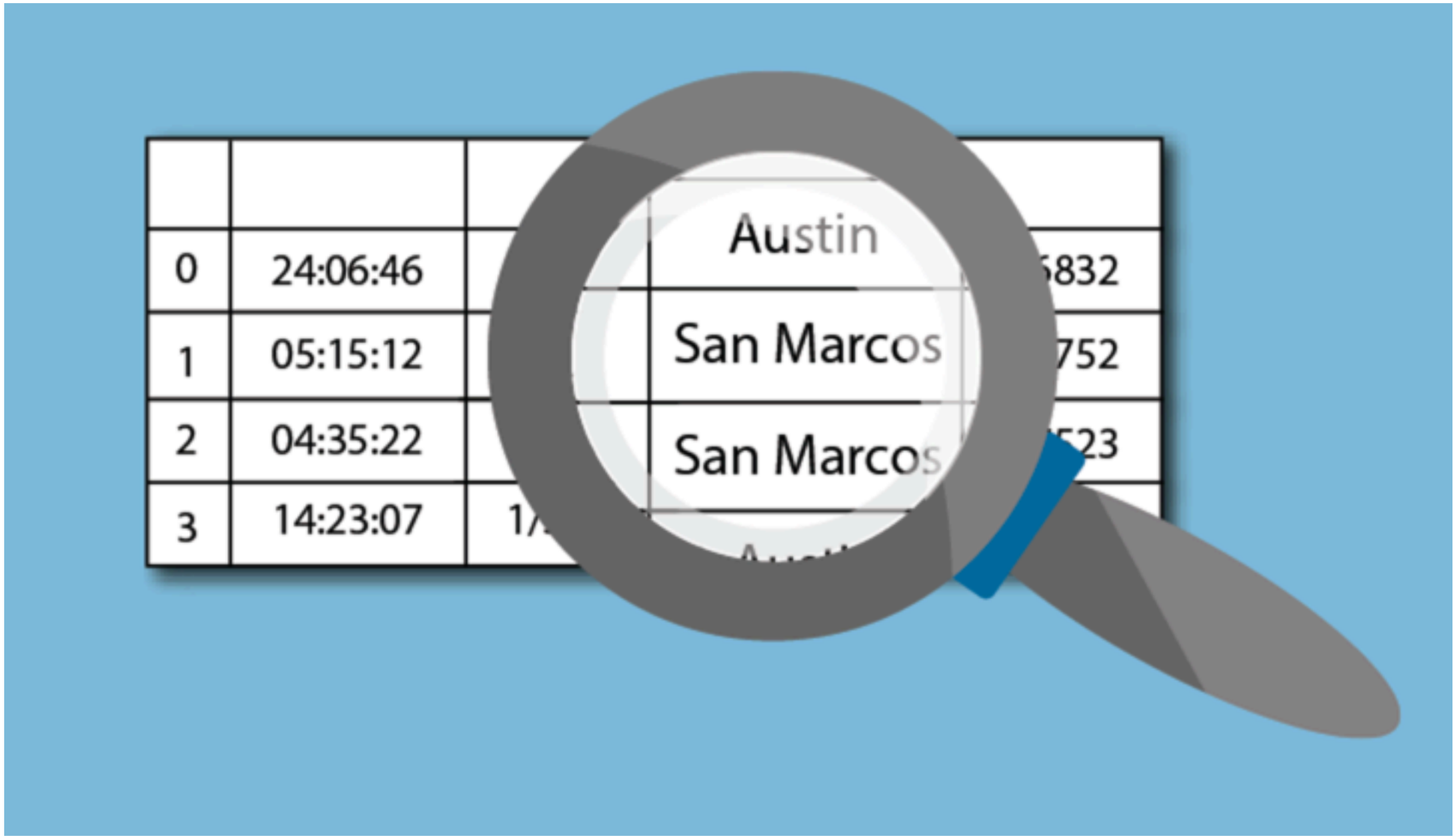
Data Wrangling



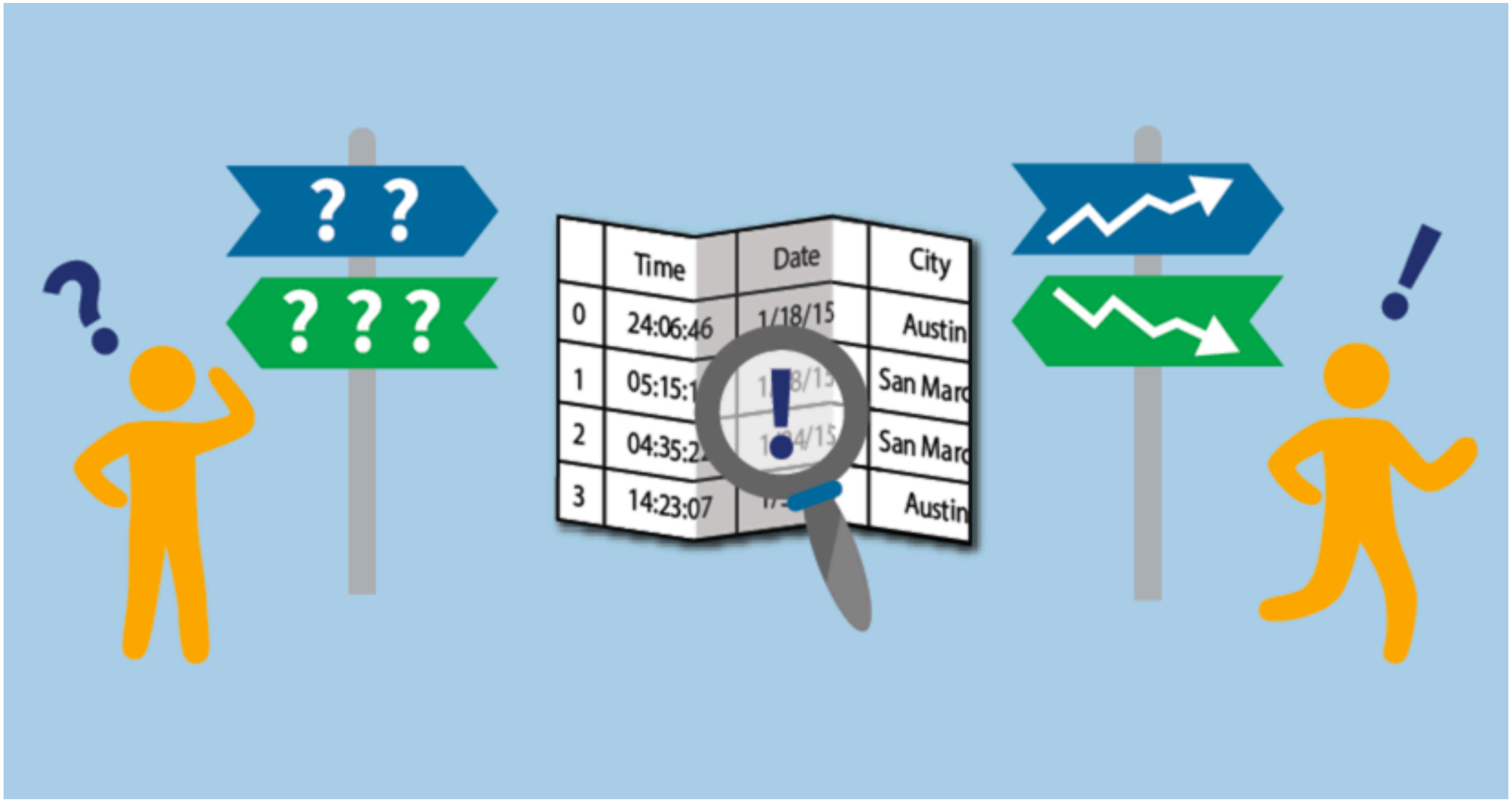
Data Visualization



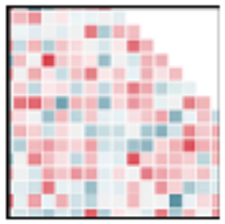
Data Analysis



Modeling and Problem Solving



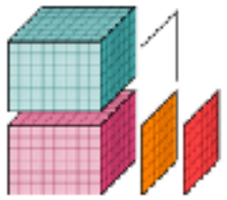
Learning path



Seaborn

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



xarray

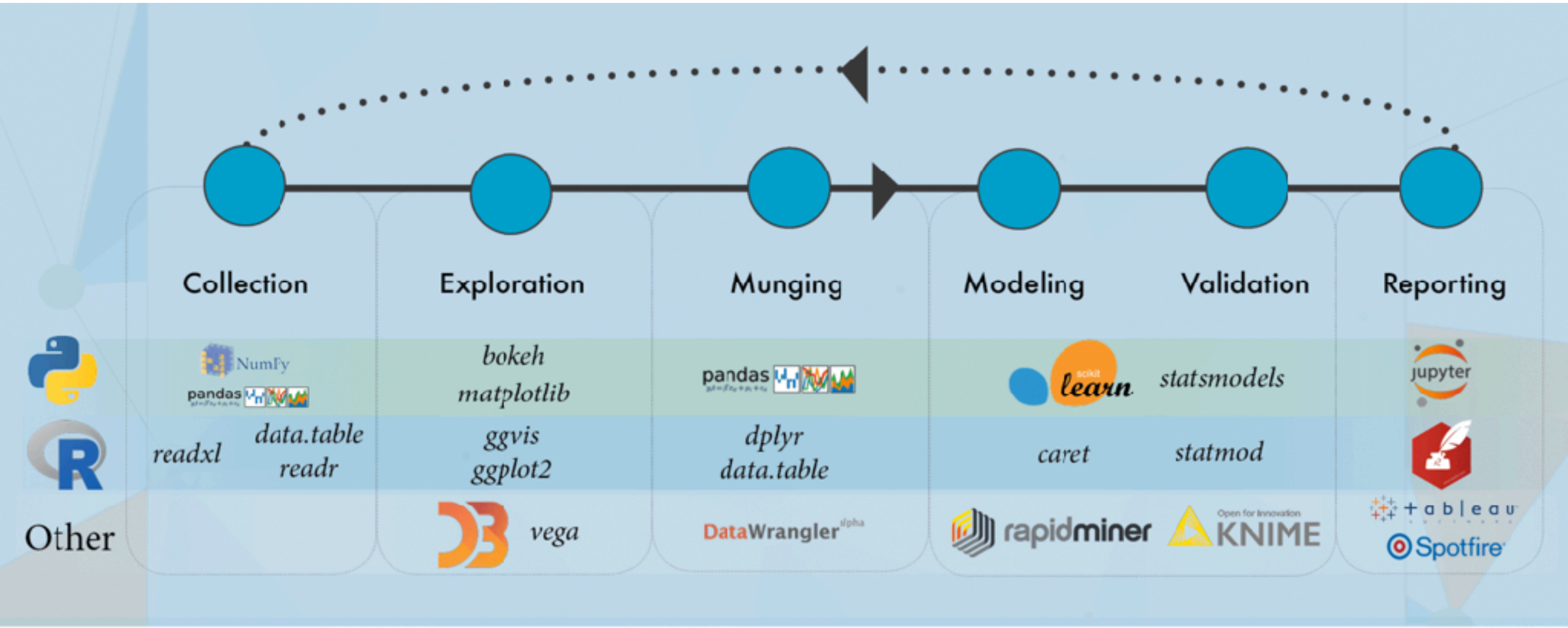
scikit-image
image processing in python

NumPy

matplotlib

IP[y]:
IPython

Data Science Workflow



Import libraries

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
import sklearn as sk
import statsmodels as sm
```

NumPy

(Numerical Python)

NumPy

Efficiency Multi-dimensional array
Fast operations on array without loop
Read and write data
Linear algebra
Random number generator

Operation for data analysis

Data munging/wrangling
Data cleaning
Data filtering
Data transformation
Data aggregation/summarize
Data sorting

Create array with 1M

```
from timeit import default_timer as timer

start = timer()

my_list = list(range(1000000))
for _ in range(10): my_list2 = [x * 2 for x in my_list]

end = timer()
print(end - start)
```

Create array with 1M

```
import numpy as np
from timeit import default_timer as timer

start = timer()

my_arr = np.arange(1000000)
for _ in range(10): my_arr2 = my_arr * 2

end = timer()
print(end - start)
```

Numpy Performance

10-100 times faster than pure Python

Multi-dimensional array object

N-dimensional array object (ndarray)
Fast and flexible for large data set

Create array with random data

```python
import numpy as np

data = np.random.randn(2, 3)
print(data)

# Operation on array
print(data * 2)
print(data + data)

# Properties of araay
print(data.dtype)
print(data.shape)
```

Create ndarray

```
import numpy as np

data1 = [1, 2, 3.5, 4, 5]
arr1 = np.array(data1)
print(arr1)
```

```
data2 = [[1, 2, 3,], [4, 5, 6]]
arr2 = np.array(data2)
print(arr2)
```

Create ndarray

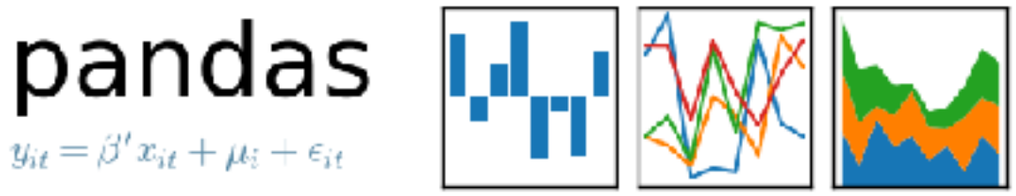
np.zeros(5)
np.ones(5)
np.empty(5)
np.arange(5)

Operation on array and scalar

```python
import numpy as np

arr = np.array([[1,2,3], [4,5,6]])
print(arr)
print(arr + arr)
print(arr - arr)
print(arr * arr)

print(1 / arr)
print(arr ** 2)
```

Pandas

(**Pan**el **Da**ta **S**tructure)



Pandas

Python data analysis library
Build on top of Numpy

Key features

DataFrame object for data manipulation
Read and write data
Data alignment and **missing data**
Reshaping and pivoting of data
Merging, Joining and grouping data
Time series functionality

Data Structure

Data Structure

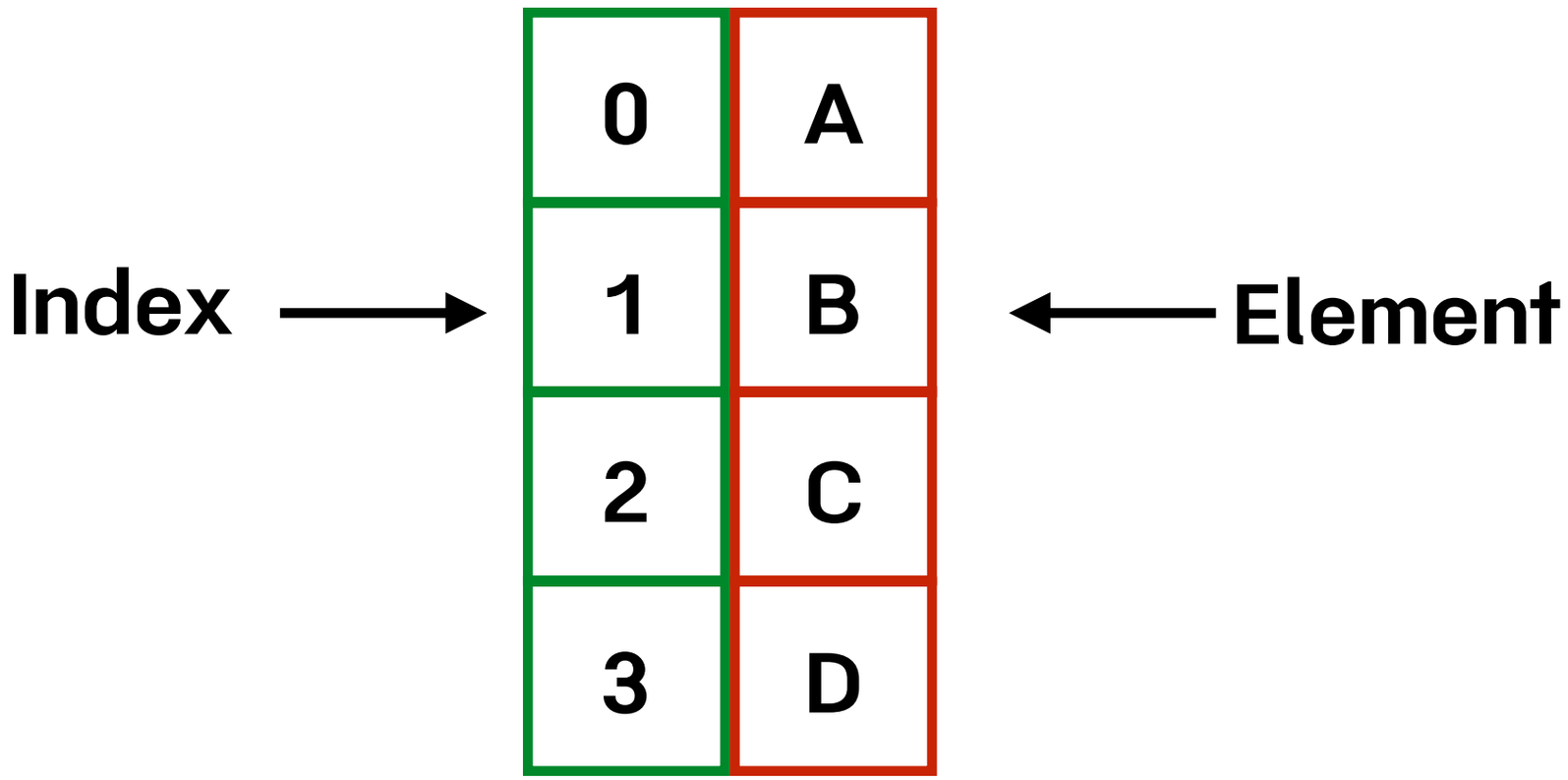
Series (1D)

DataFrame (2D)

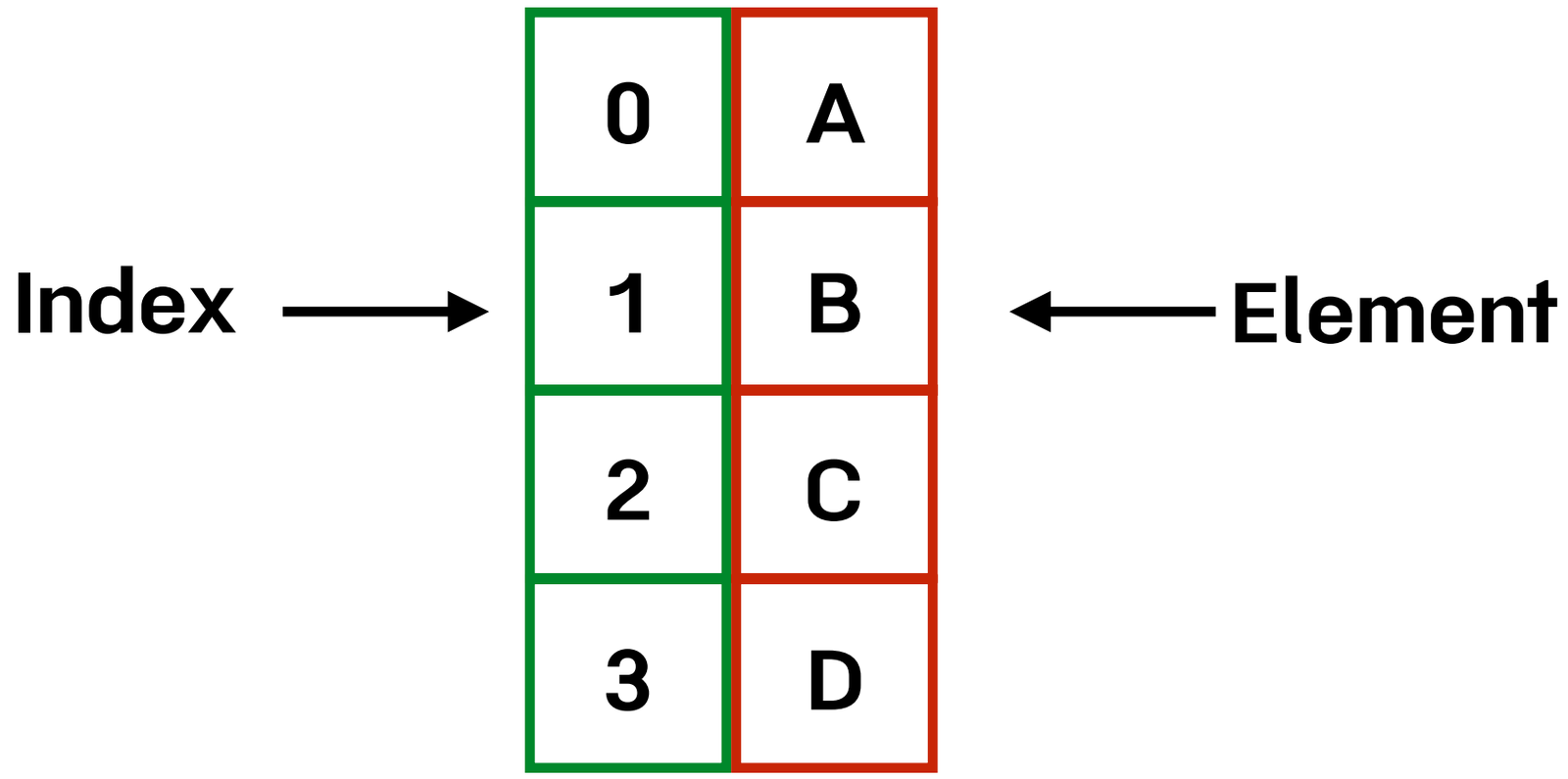
Panel (3D)

Series

A one-dimensional **labeled array** capable of holding any data type

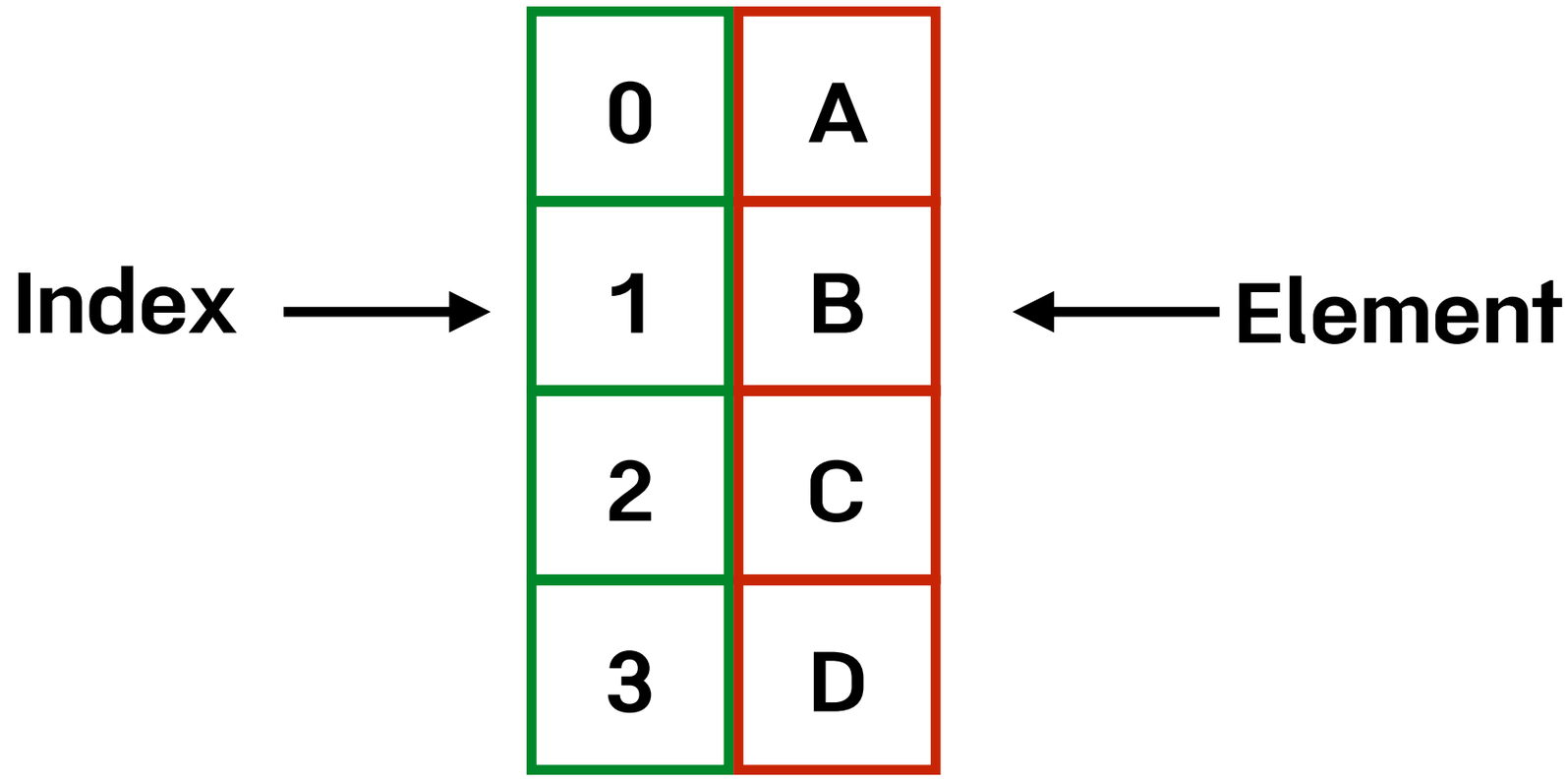


Series




```
pd.Series(['A', 'B', 'C', 'D'])
```

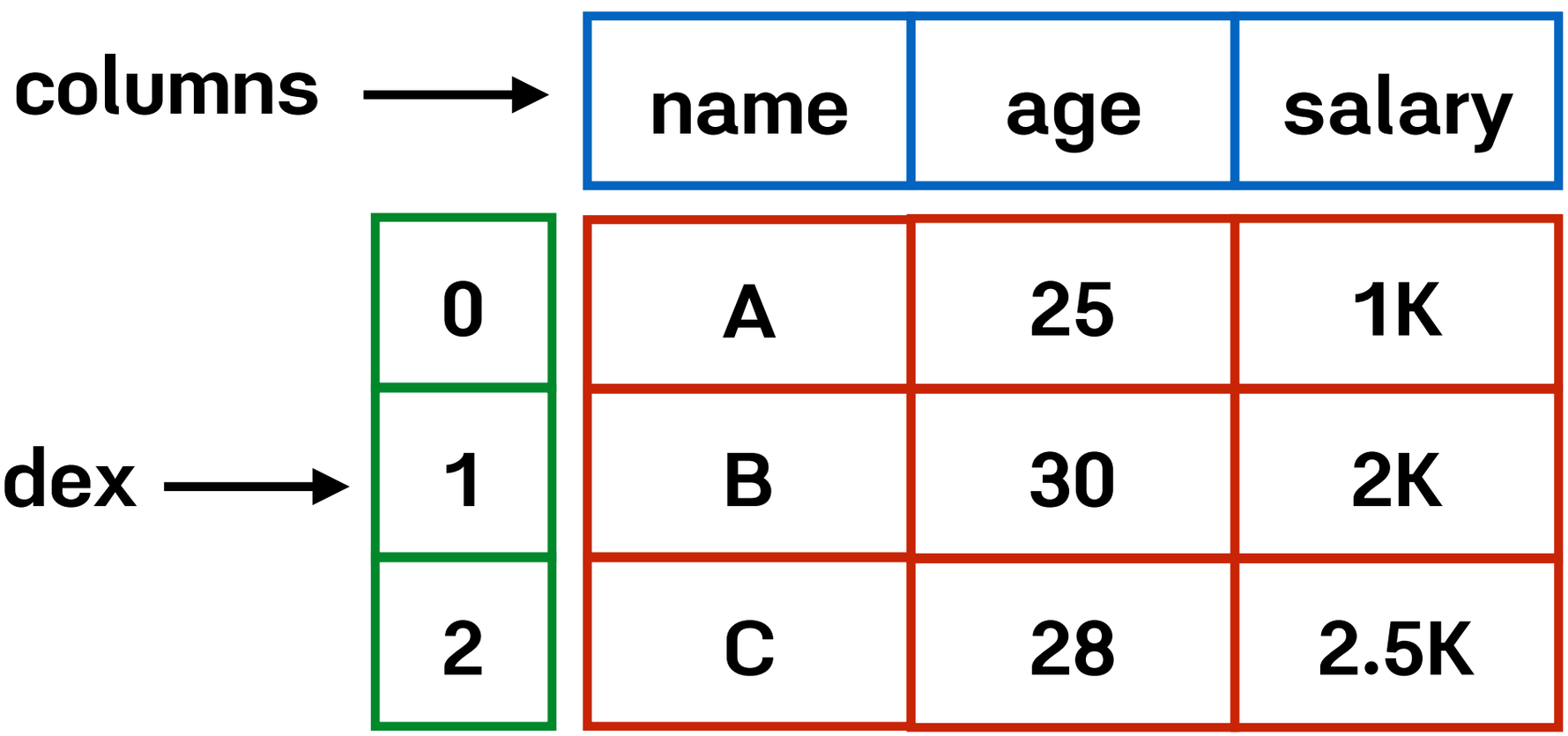
Series



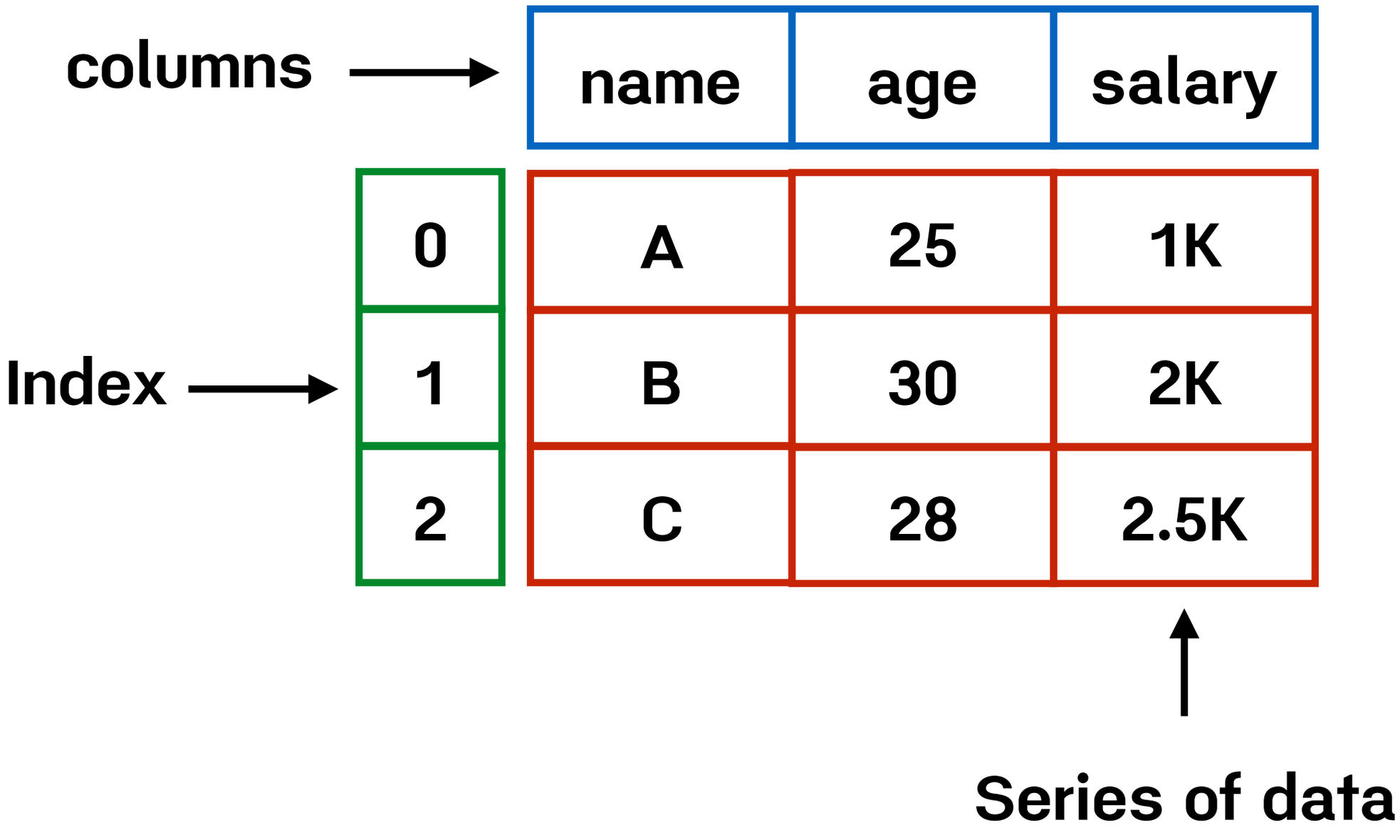

```
pd.Series(['A', 'B', 'C', 'D'], index=[0, 1, 2, 3])
```

DataFrame

A two-dimensional labeled data structure with columns of potentially different types



DataFrame


columns
name
age
salary
Index
0
A
25
1K
1
B
30
2K
2
C
28
2.5K
Series of data

DataFrame

Create dataframe from Python's dictionary

```
data = { 'name': ['A', 'B', 'C'],
         'age': [25, 30, 28],
         'salary': [10000, 20000, 25000]
}

df = pd.DataFrame(data)
```

DataFrame

Improve sequence of columns

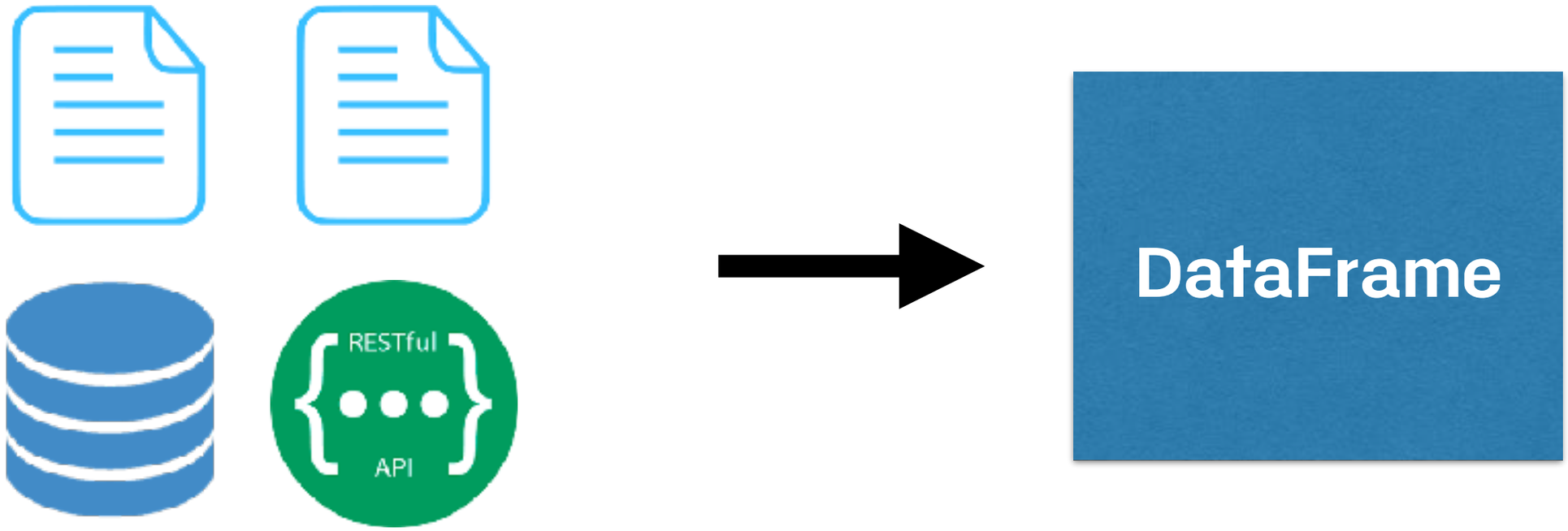
```
data = { 'name': ['A', 'B', 'C'],
         'age': [25, 30, 28],
         'salary': [10000, 20000, 25000]
}

df = pd.DataFrame(data,
                  columns=['name', 'age', 'salary'])
```

Data loading

Read data from text format

Function	Description
read_csv	อ่านข้อมูลในรูปแบบ CSV โดยแยกด้วย comma
read_table	อ่านข้อมูลในรูปแบบ table โดยแยกด้วย TAB
read_fwf	อ่านข้อมูลในรูปแบบ fixed-length ของ column
read_excel	อ่านข้อมูลในรูปแบบของ MS Excel
read_html	อ่านข้อมูลในรูปแบบ HTML
read_json	อ่านข้อมูลในรูปแบบ JSON (JavaScript Object Notation)
read_sql	อ่านข้อมูลจาก SQL Query ใช้งานผ่าน SQLAlchemy

Read data from datasource



Data Sources

Text file
Binary
Web API
Database

Data cleaning and preparation

Data wrangling

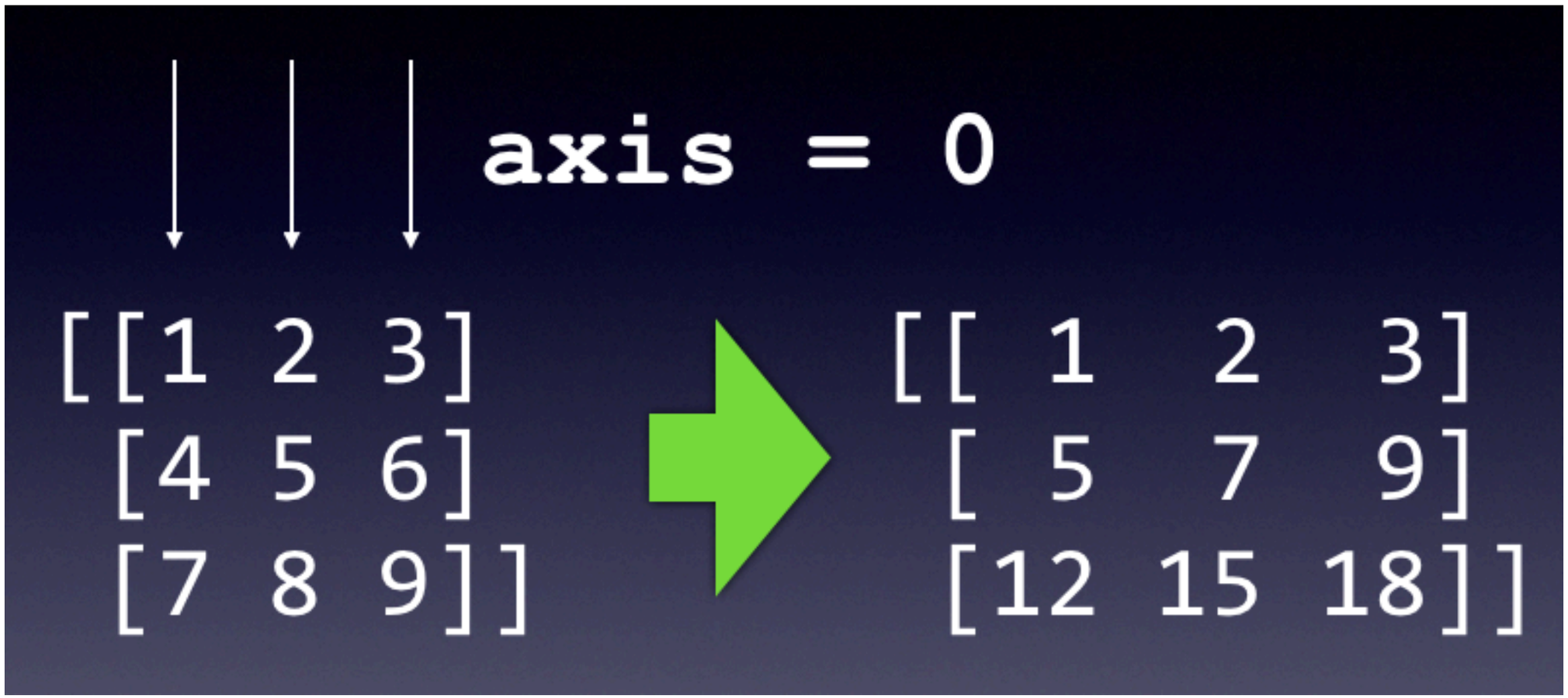
Data aggregation

Plotting and Visualization

Time Series

workshop

axis = 0



axis = 1