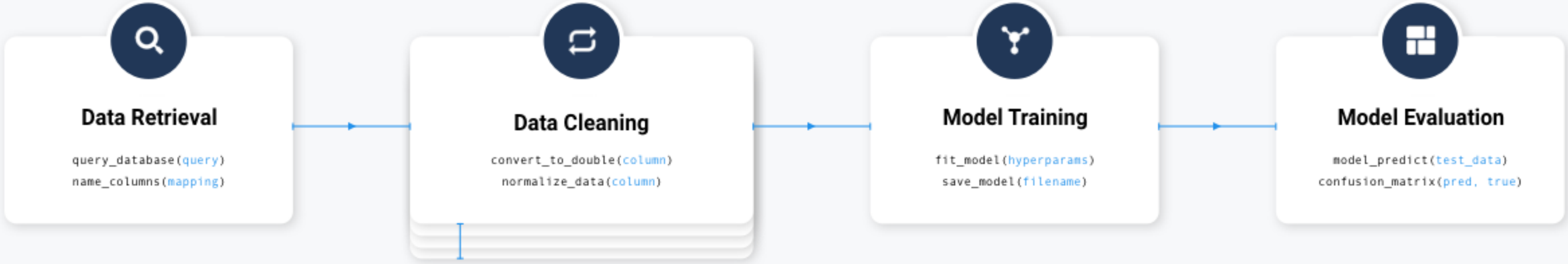
Telkomsel
by Telkom Indonesia

SPILL THE TECH WORKSHOP

Data Science Pipeline

Randy Galawana
Data Engineer

Data Science Pipeline in General



Data Retrieval

```
query_database(query)
name_columns(mapping)
```

Data Cleaning

```
convert_to_double(column)
normalize_data(column)
```

Model Training

```
fit_model(hyperparams)
save_model(filename)
```

Model Evaluation

```
model_predict(test_data)
confusion_matrix(pred, true)
```

Data Engineer Domain

- Big Data
- Different Tech in all use case
- Data Warehouse and Data Lake
- Advance Data Transformation
- Data Quality Checks

Data Scientist Domain

- Sample Data
- Using most common tech / playground
- Statistic and Visualization
- Simple Data Cleaning
- Simple Data Transformation

Prerequisites

What need to be installed before the workshop

-Anaconda Jupyter Notebook

-Python v3.6 +

Or use google collabs

https://colab.research.google.com/

Telkomsel
by Telkom Indonesia

Get Sample Code

Download sample code from

https://github.com/galawana/tselworkshop1

(Download Zip and extract or git clone repository)
If using google collabs, upload to drive.google.com

01

Data Retrieval / Ingestion

How to Ingest data from multiple source

Data Retrieval using Pandas

Retrieve data from known data sources

Database (JDBC / ODBC)

Using pandas read_sql function with sqlalchemy connection, example in notebook. Support most of ODBC and JDBC database

IO File

Using pandas io read https://pandas.pydata.org/docs/user_guide/io.html

Support almost file extension and format

API's

Using pandas io read https://pandas.pydata.org/docs/user_guide/io.html

Support almost file extension and format

Open Sample code in

Data_Retrieval.ipnyb

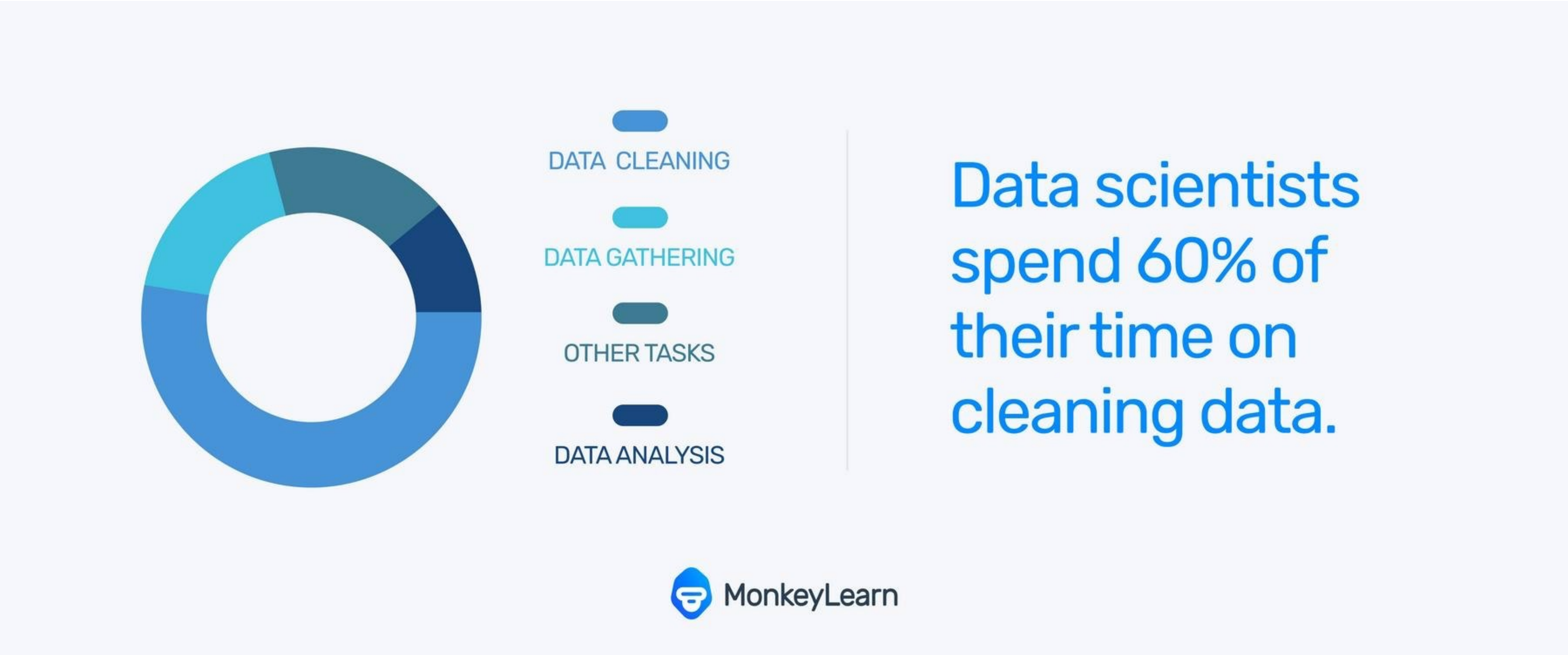
Demo and Hands on

02

Data Cleaning and Transformation

Clean the data, and do simple to advance transformation

Why Data Cleaning need to include in Pipeline



Data scientists spend 60% of their time on cleaning data.

MonkeyLearn

Data Cleaning using Pandas

Most common Data cleaning using pandas and python

Handling Null Values / Missing

Imputer

- Simple Imputer
- Mode, Mean Values
- Hot Deck

Drop Na

- Thresholding

Standardization and Normalization

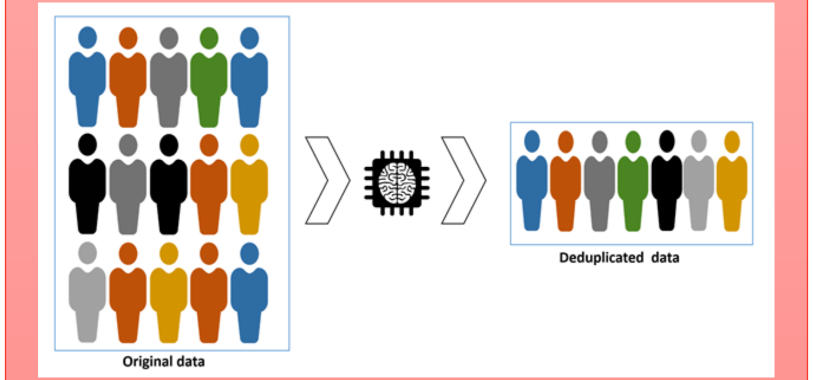
Category Encoder

- One hot encoding
- Label encoder, etc

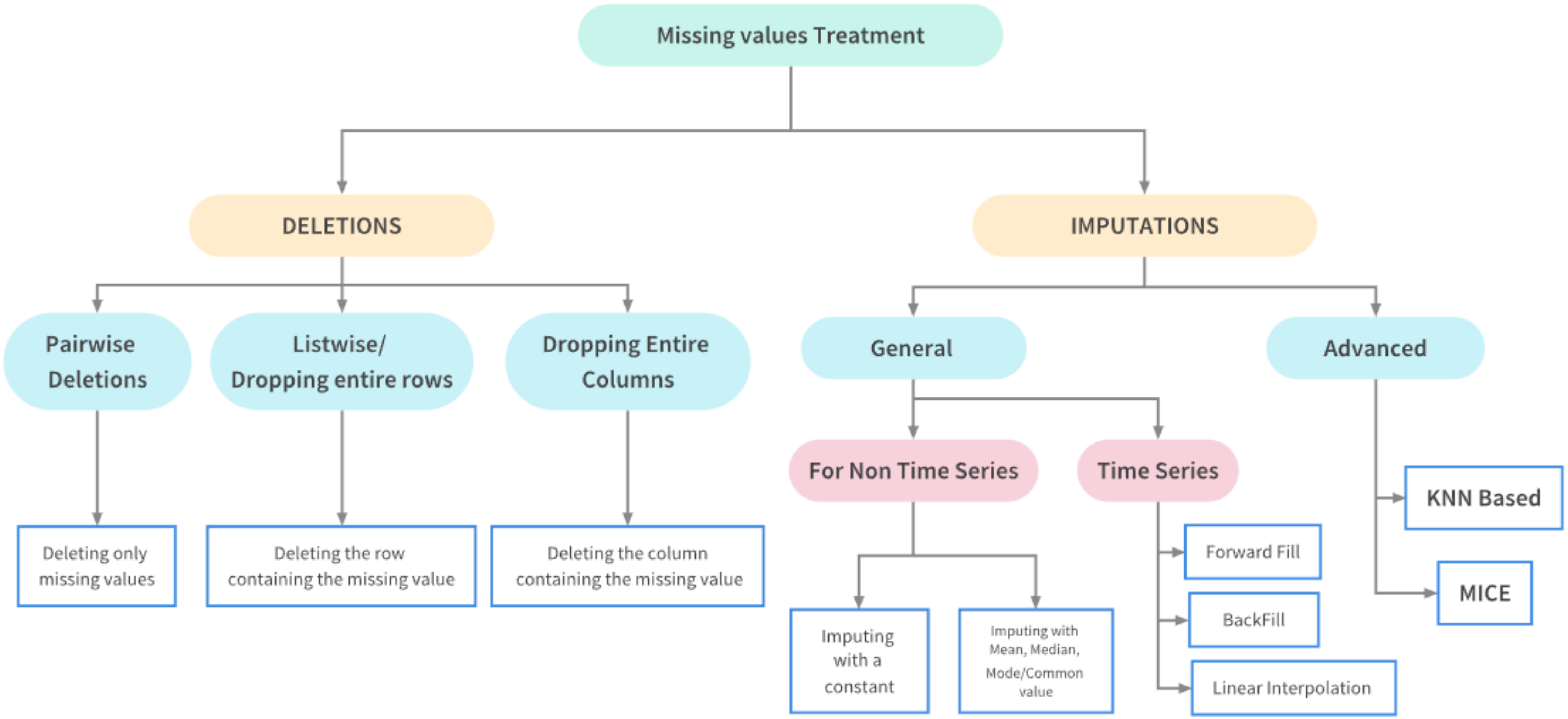
Scaler

- MinMaxScaler
- StandartScaler, etc

De-Duplication



Missing Value Treatment



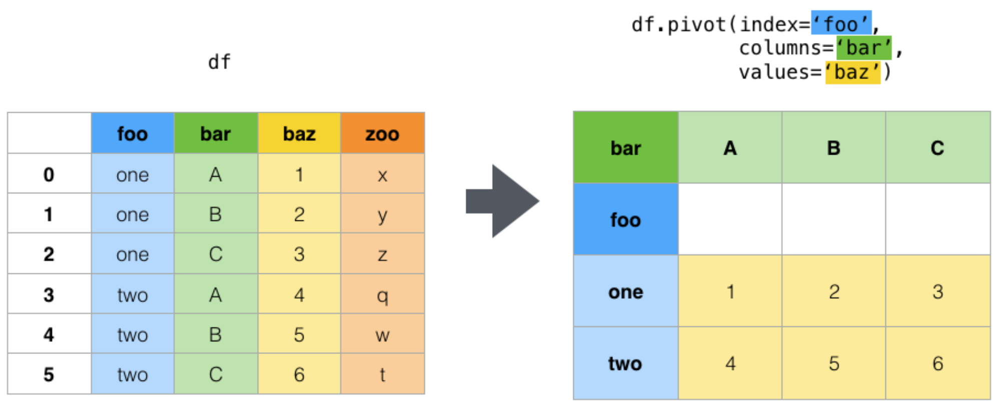
Open Sample code in

Data_Cleaning.ipnyb

Demo and Hands on

Data Transformation using pandas

Pivot

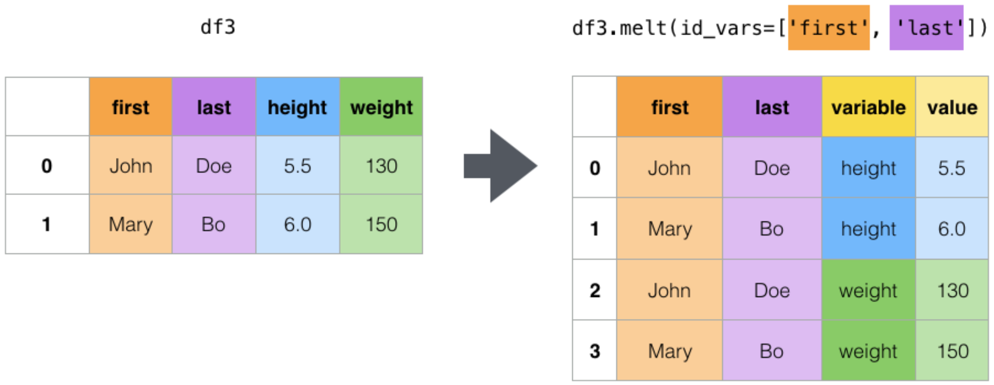


	foo	bar	baz	zoo
0	one	A	1	x
1	one	B	2	y
2	one	C	3	z
3	two	A	4	q
4	two	B	5	w
5	two	C	6	t

```
df.pivot(index=‘foo’,
         columns=‘bar’,
         values=‘baz’)
```

bar	A	B	C
foo			
one	1	2	3
two	4	5	6

Melt

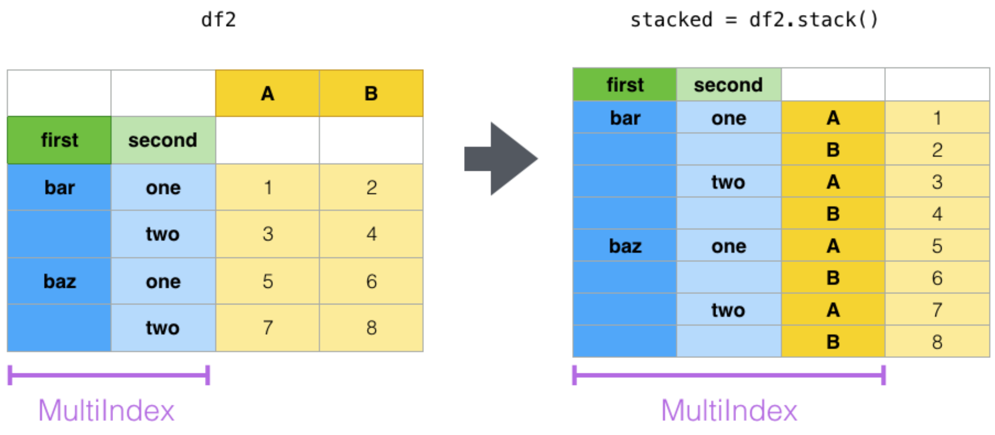


	first	last	height	weight
0	John	Doe	5.5	130
1	Mary	Bo	6.0	150

```
df3.melt(id_vars=['first', 'last'])
```

	first	last	variable	value
0	John	Doe	height	5.5
1	Mary	Bo	height	6.0
2	John	Doe	weight	130
3	Mary	Bo	weight	150

Stack

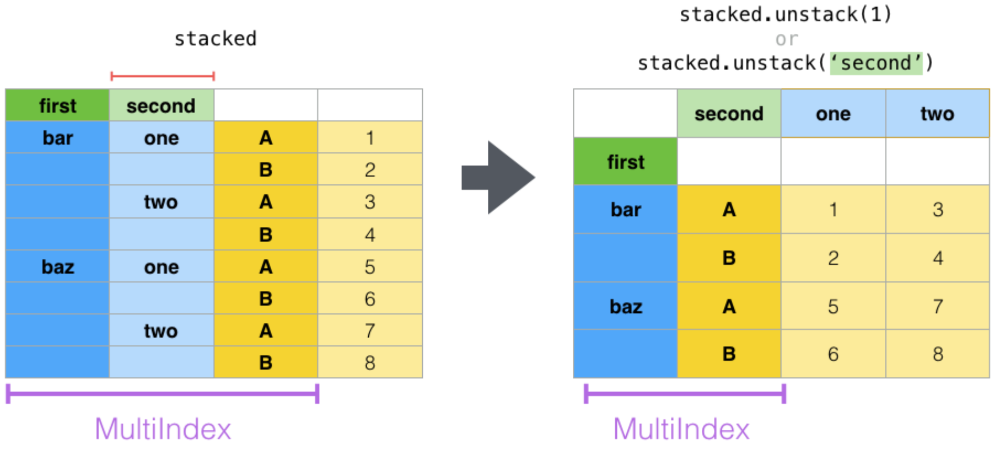


		A	B
first	second		
bar	one	1	2
	two	3	4
baz	one	5	6
	two	7	8

```
stacked = df2.stack()
```

first	second		
bar	one	A	1
		B	2
	two	A	3
		B	4
baz	one	A	5
		B	6
	two	A	7
		B	8

Unstack(1)



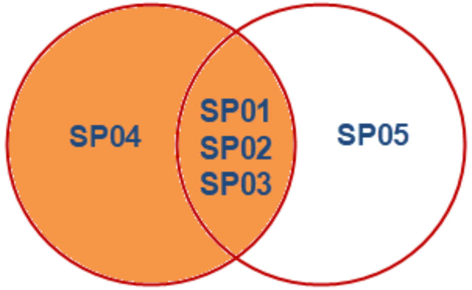
first	second		
bar	one	A	1
		B	2
	two	A	3
		B	4
baz	one	A	5
		B	6
	two	A	7
		B	8

```
stacked.unstack(1)
      or
stacked.unstack(‘second’)
```

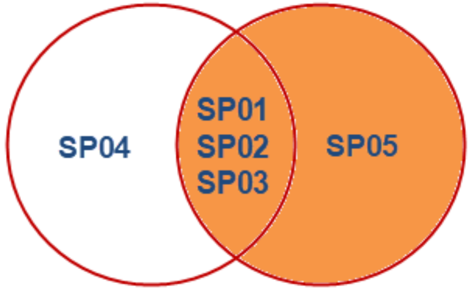
	second	one	two
first			
bar	A	1	3
	B	2	4
baz	A	5	7
	B	6	8

Data Transformation using pandas

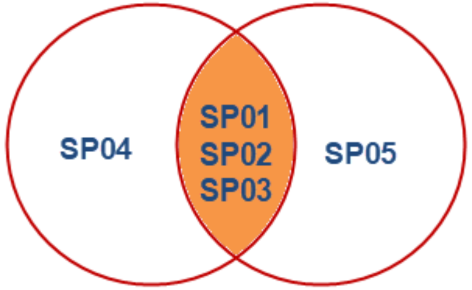
Join / Merge



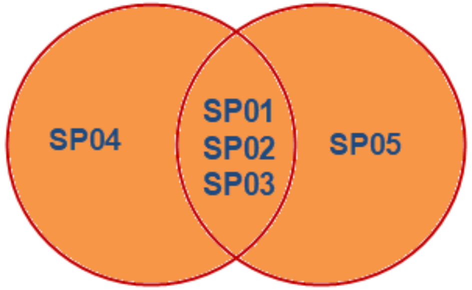
LEFT



RIGHT

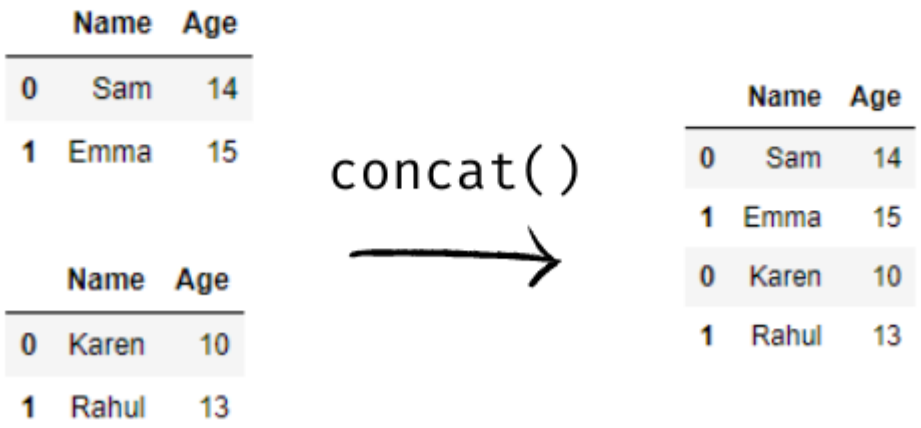


INNER



OUTER

Concat



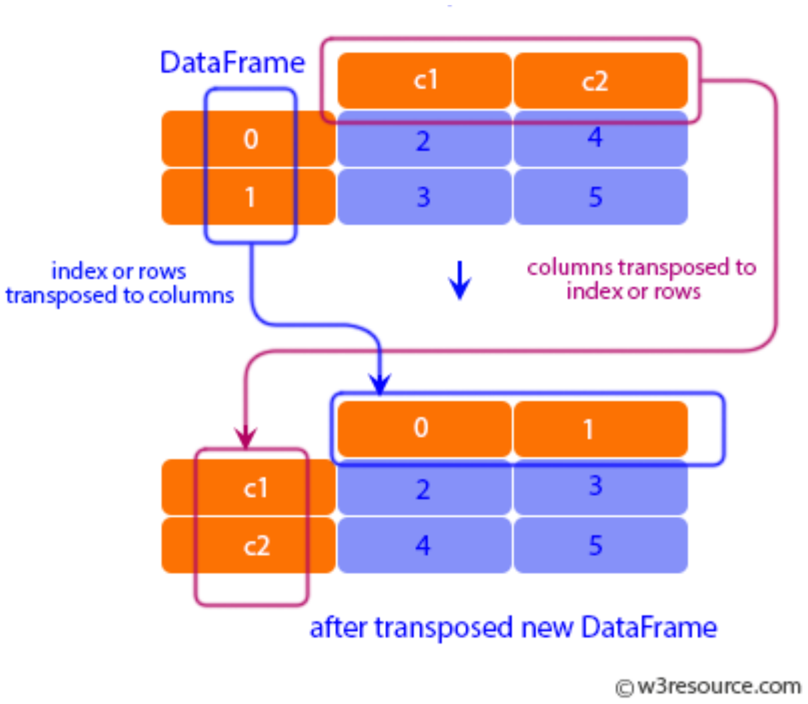
	Name	Age
0	Sam	14
1	Emma	15

	Name	Age
0	Karen	10
1	Rahul	13

	Name	Age
0	Sam	14
1	Emma	15
0	Karen	10
1	Rahul	13

Data Transformation using pandas

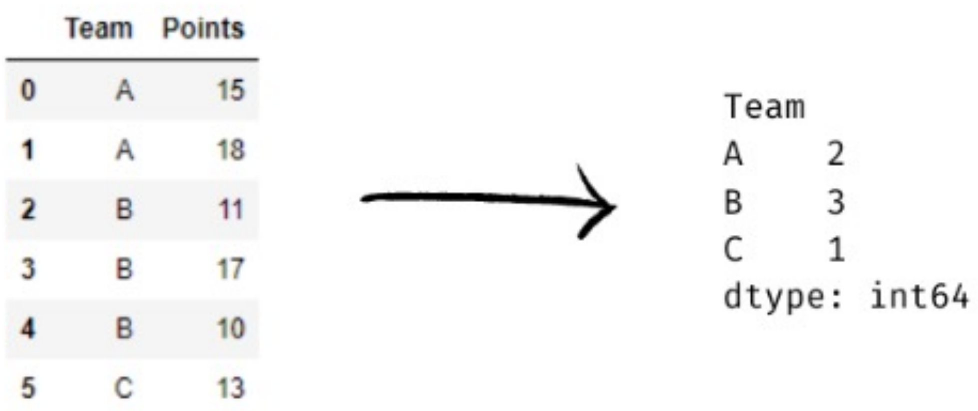
Transpose


DataFrame
index or rows transposed to columns
columns transposed to index or rows
after transposed new DataFrame
©w3resource.com

	c1	c2
0	2	4
1	3	5

	0	1
c1	2	3
c2	4	5

Group by / Aggregate



	Team	Points
0	A	15
1	A	18
2	B	11
3	B	17
4	B	10
5	C	13

```
Team
A      2
B      3
C      1
dtype: int64
```

Open Sample code in
Data_Transformation.ipnyb

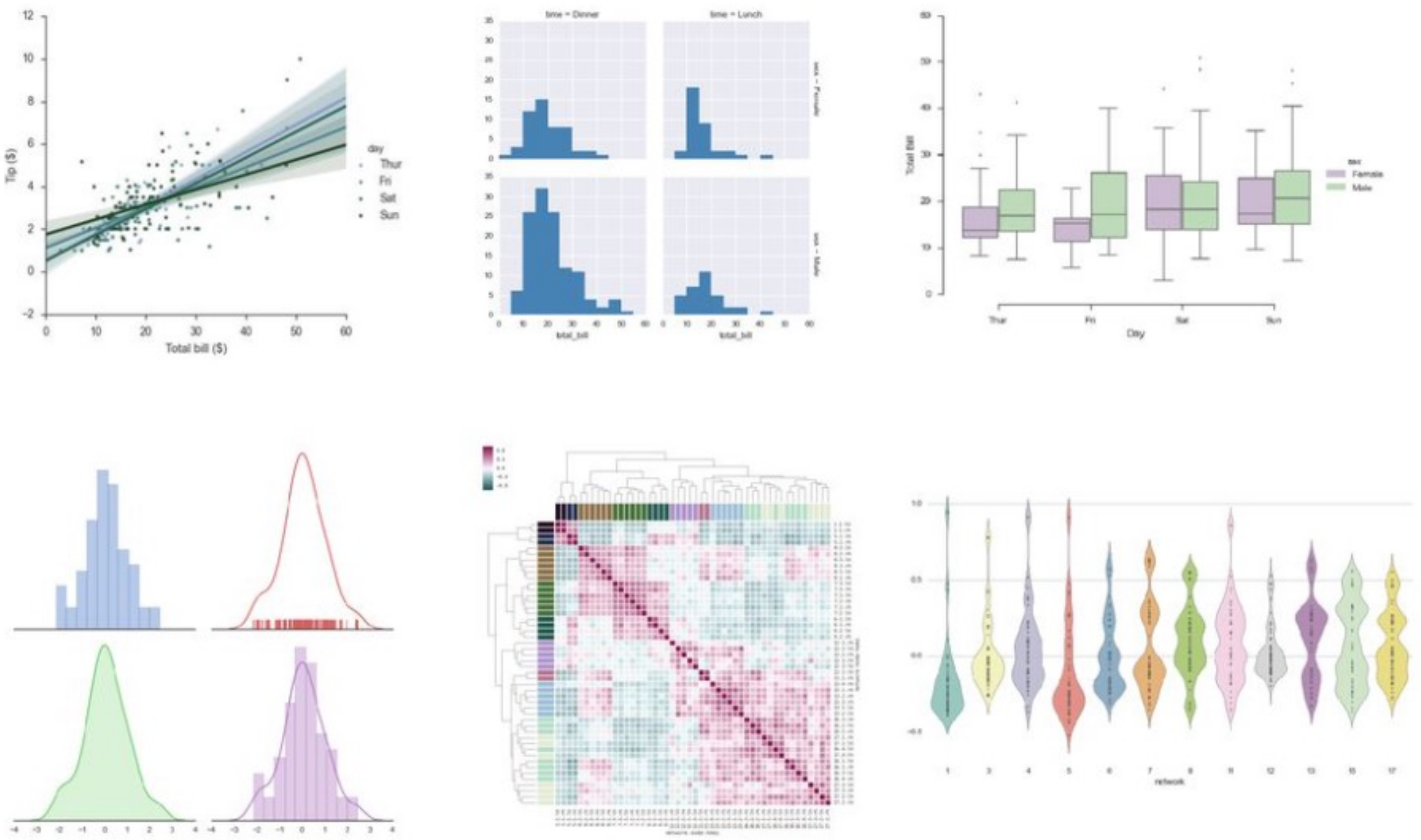
Demo and Hands on

03

EDA and Data Visualization with Seaborn

Explore and Tell the story from the data using
seaborn visualization library

Seaborn chart library



Telkomsel
by Telkom Indonesia

Open Sample code in
EDA.ipnyb

Demo and Hands on

04

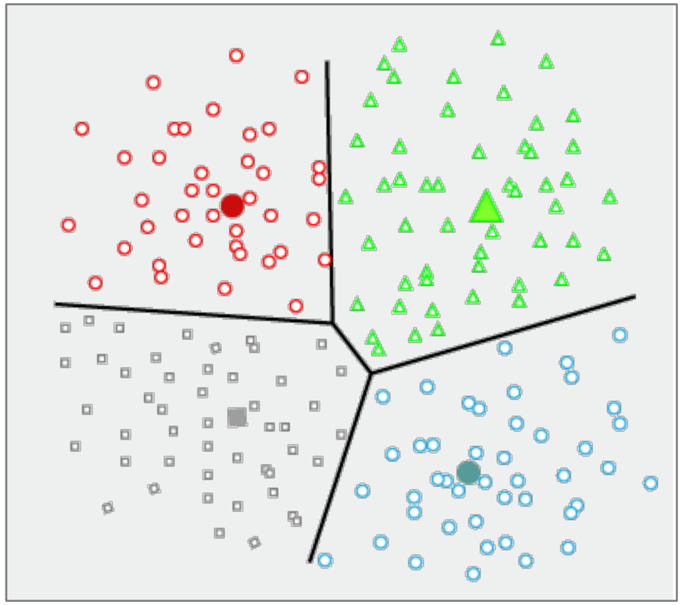
Model Training dan Evaluation with scikit learn

Training ML model and evaluate the model

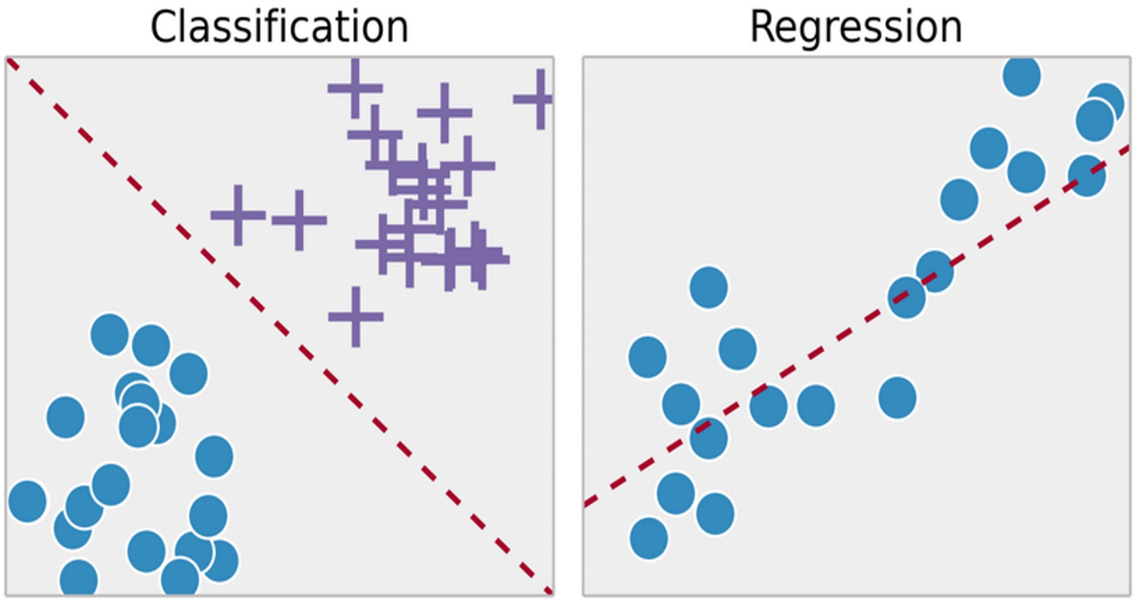
Type of Machine Learning

Unsupervised Learning

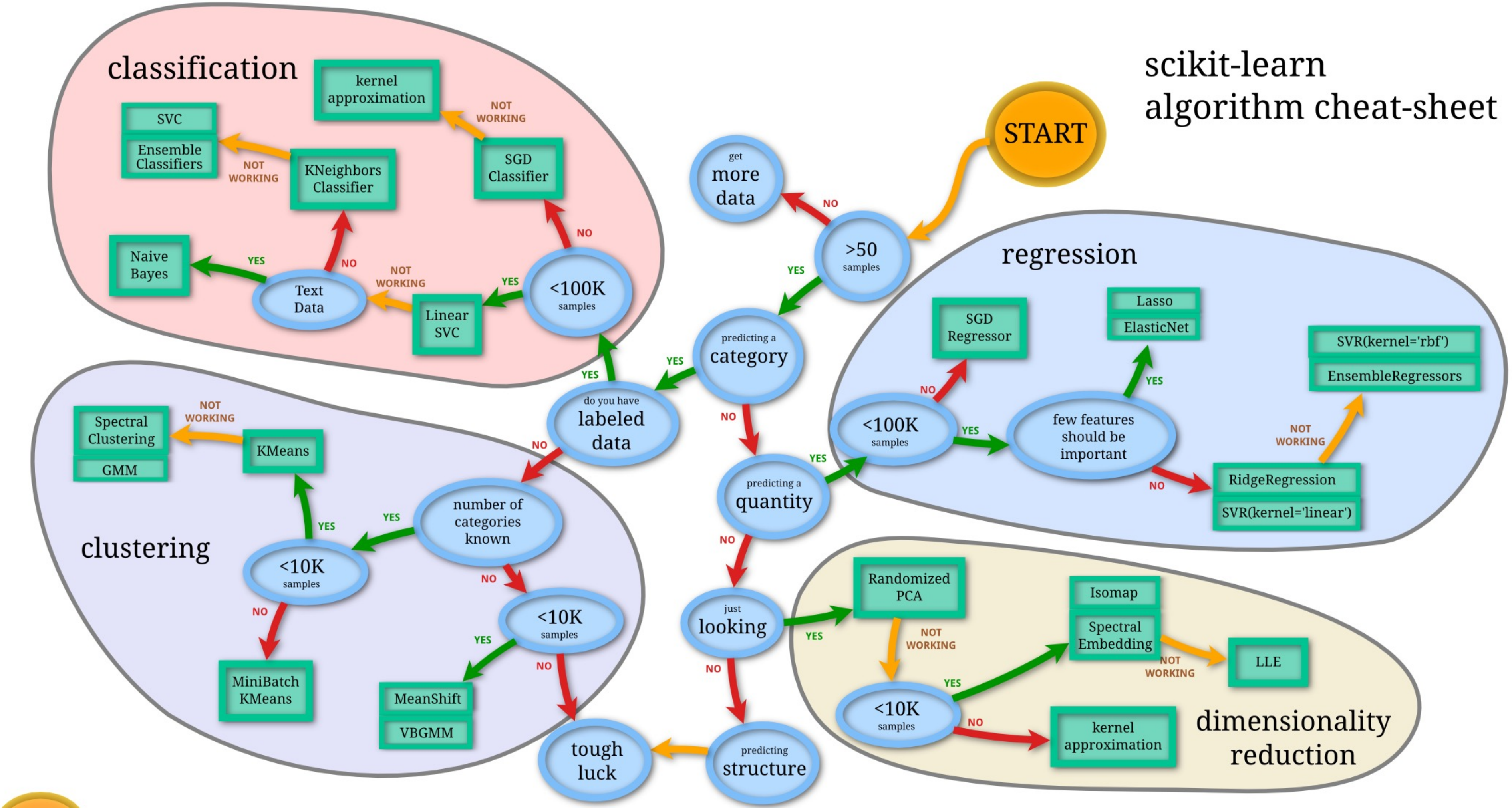
Clustering



Supervised Learning

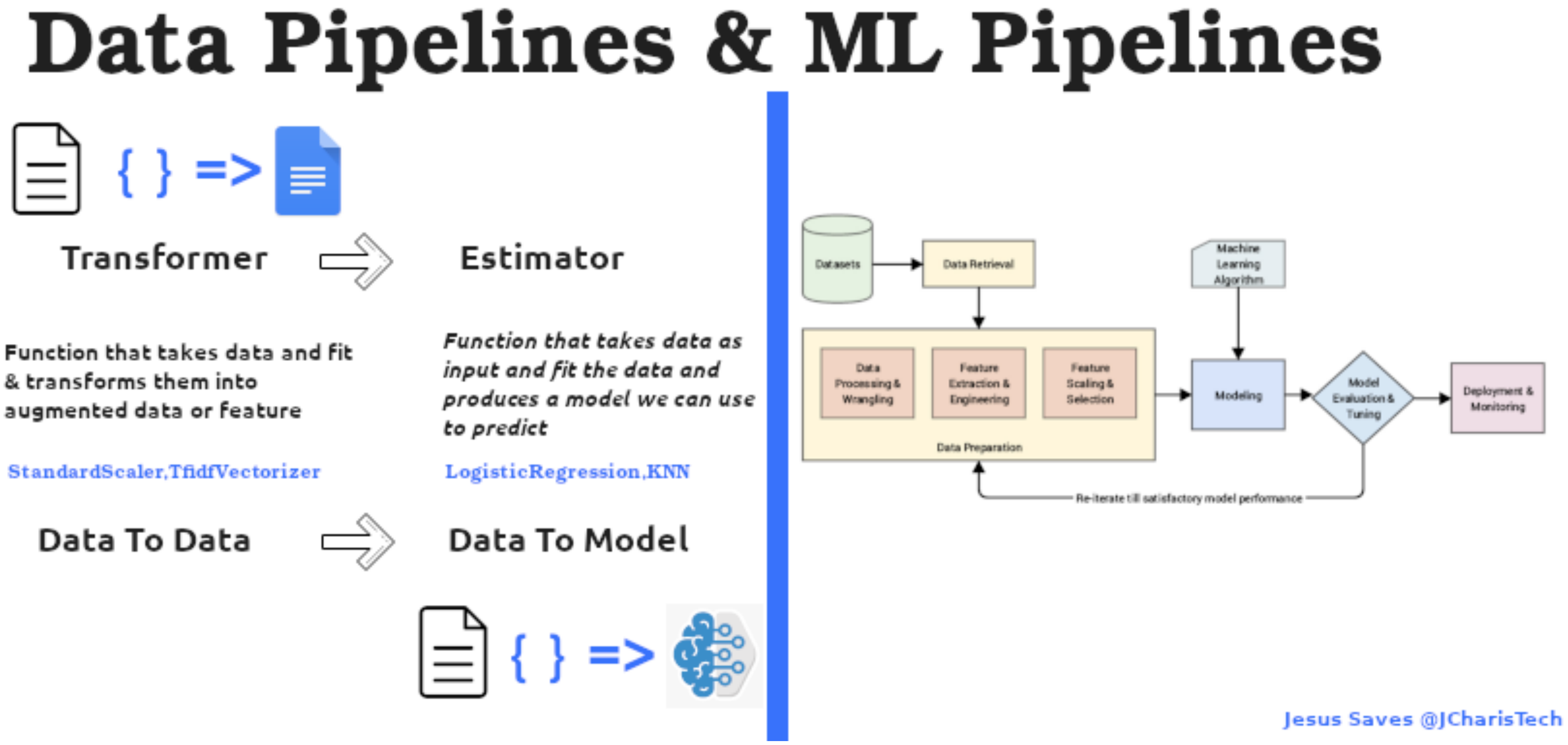


Scikit Learn ML Library



Back

Scikit Learn Pipelines



Data Pipelines & ML Pipelines

{ } =>

Transformer → Estimator

Function that takes data and fit & transforms them into augmented data or feature

StandardScaler,TfidfVectorizer

Function that takes data as input and fit the data and produces a model we can use to predict

LogisticRegression,KNN

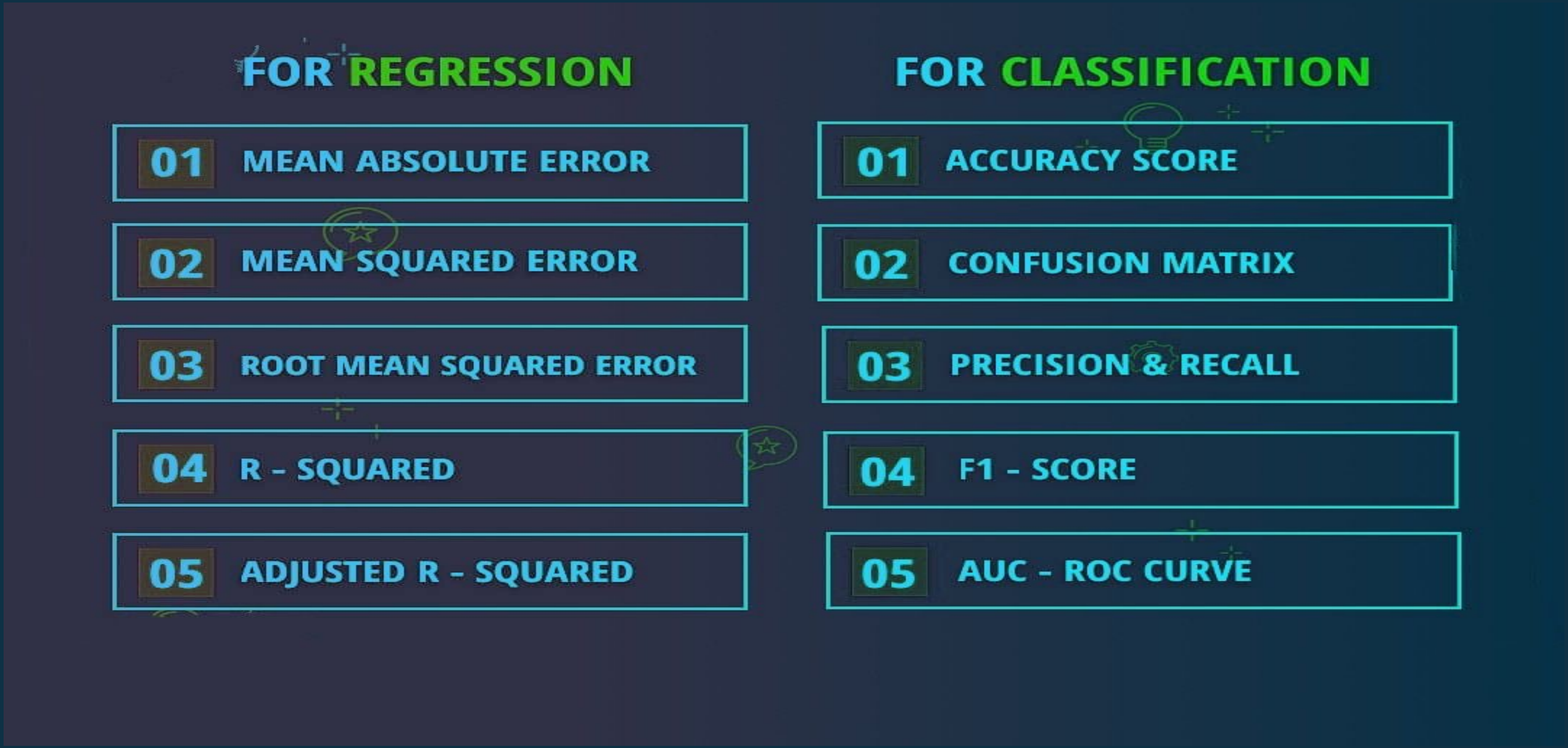
Data To Data → Data To Model

{ } =>

Datasets
Data Retrieval
Machine Learning Algorithm
Data Processing & Wrangling
Feature Extraction & Engineering
Feature Scaling & Selection
Data Preparation
Modeling
Model Evaluation & Tuning
Deployment & Monitoring
Re-iterate till satisfactory model performance

Jesus Saves @JCharisTech

Model Evaluation



FOR REGRESSION

01 MEAN ABSOLUTE ERROR

02 MEAN SQUARED ERROR

03 ROOT MEAN SQUARED ERROR

04 R – SQUARED

05 ADJUSTED R – SQUARED

FOR CLASSIFICATION

01 ACCURACY SCORE

02 CONFUSION MATRIX

03 PRECISION & RECALL

04 F1 – SCORE

05 AUC – ROC CURVE

Open Sample code in

Machine_Learning.ipnyb

Demo and Hands on

Telkomsel
by Telkom Indonesia

Thank You