Machine Learning

The Data Science Process

1. Identify the question
2. Get the data
3. Clean the data
4. Explore the data
5. Model the data
6. Communicate the results

The Data Science Process

1. Identify the question
2. Get the data
3. Clean the data
4. Explore the data
5. Model the data
6. Communicate the results

The Data Science Process

1. Identify the question
2. Get the data
3. Clean the data
4. Explore the data
5. Model the data
6. Communicate the results

Identify the question

Identify the question

- Answerable
- Actionable
- Narrow
- Specific

Get the data

Data Sources

- kaggle.com/datasets
- https://registry.opendata.aws
- https://cloud.google.com/bigquery/public-data/
- data.gov
- archive.ics.uci.edu/ml/
- https://github.com/fivethirtyeight/data
- https://www.quandl.com/search
- public APIs (e.g., Twitter, Facebook, Spotify)
- web scraping
- your company

Data Sources

Google Dataset Search

toolbox.google.com/datasetsearch

ProPublica Data Store

propublica.org/datastore

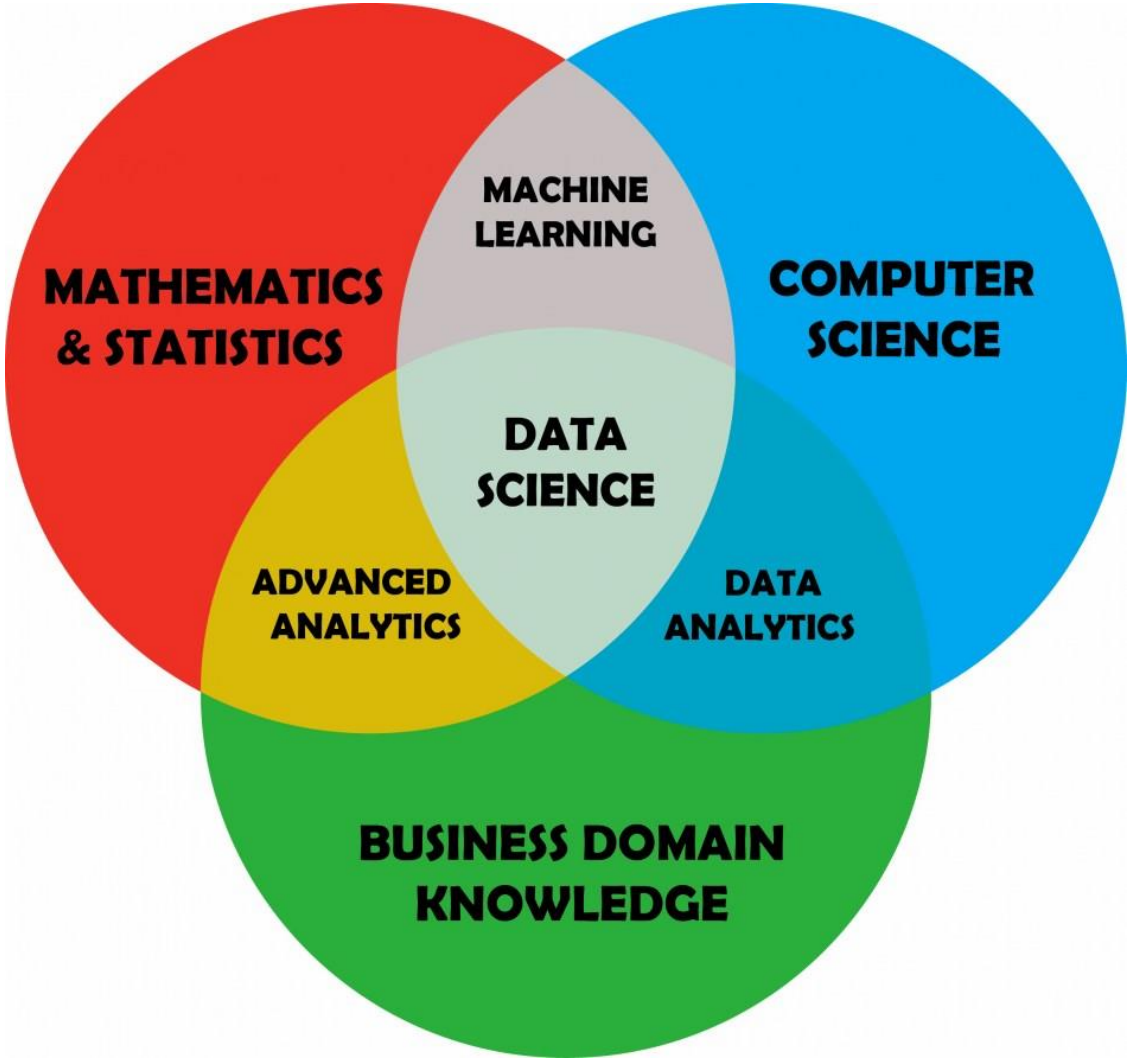
NASA's Open Data Portal

data.nasa.gov

World Bank Open Data

data.worldbank.org

Data Science



Python Libraries for Data Analysis

NumPy

Pandas

Matplotlib

SciPy

Scikit-learn

What is Machine Learning?

$$F = C * 1.8 + 32$$

Celsius	0	8	15	22	38
Fahrenheit	32	46.4	59	71.6	100.4

Machine Learning

Input: [**0, 8, 15 22**]

Machine Learning

Input: [**0, 8, 15 22**]

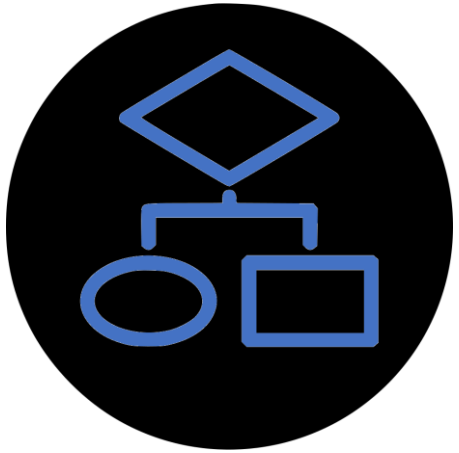
Output: [**32, 46.4, 59, 71.6**]

Machine Learning

Input: [**0, 8, 15 22**]

Relationship: **?**

Output: [**32, 46.4, 59, 71.6**]



=

Common ML Algorithms

Linear Regression

Logistic Regression

Naïve Bayes

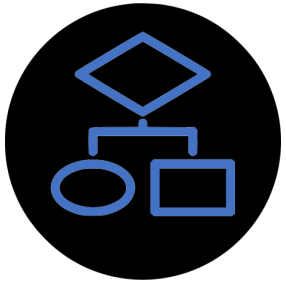
Support Vector Machine

Decision Tree

K-Nearest Neighbor

Machine Learning

Input: [**0, 8, 15 22**]

Relationship: 

Output: [**32, 46.4, 59, 71.6**]

Machine Learning

Input: [**0, 8, 15 22**]

Relationship: **input** *1.8 + 32

Output: [**32, 46.4, 59, 71.6**]

Machine Learning

Input: [**0, 8, 15 22**]

Relationship: **input** *1.8 + 32 ← **Model**

Output: [**32, 46.4, 59, 71.6**]

Machine Learning

ML Model

input *1.8 + 32

Machine Learning

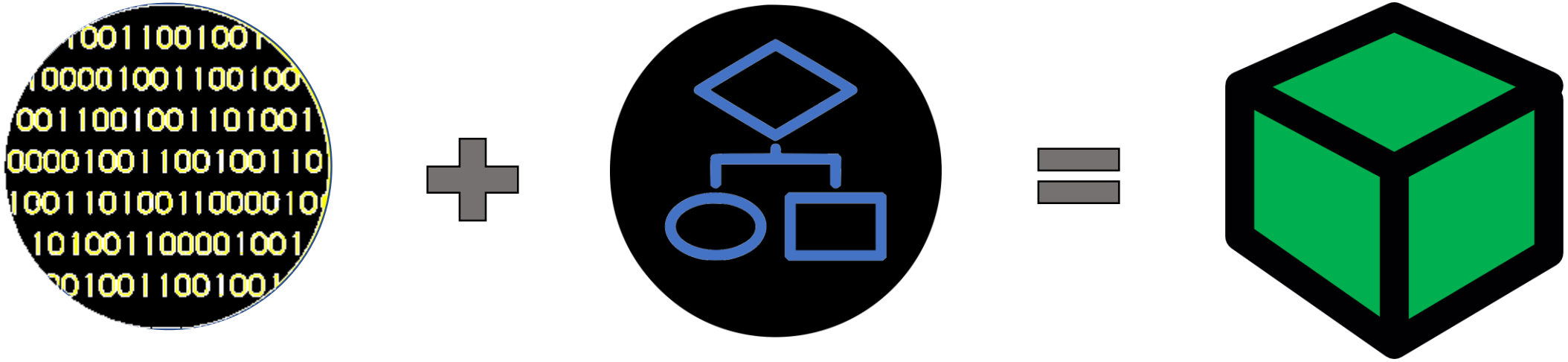
ML Model

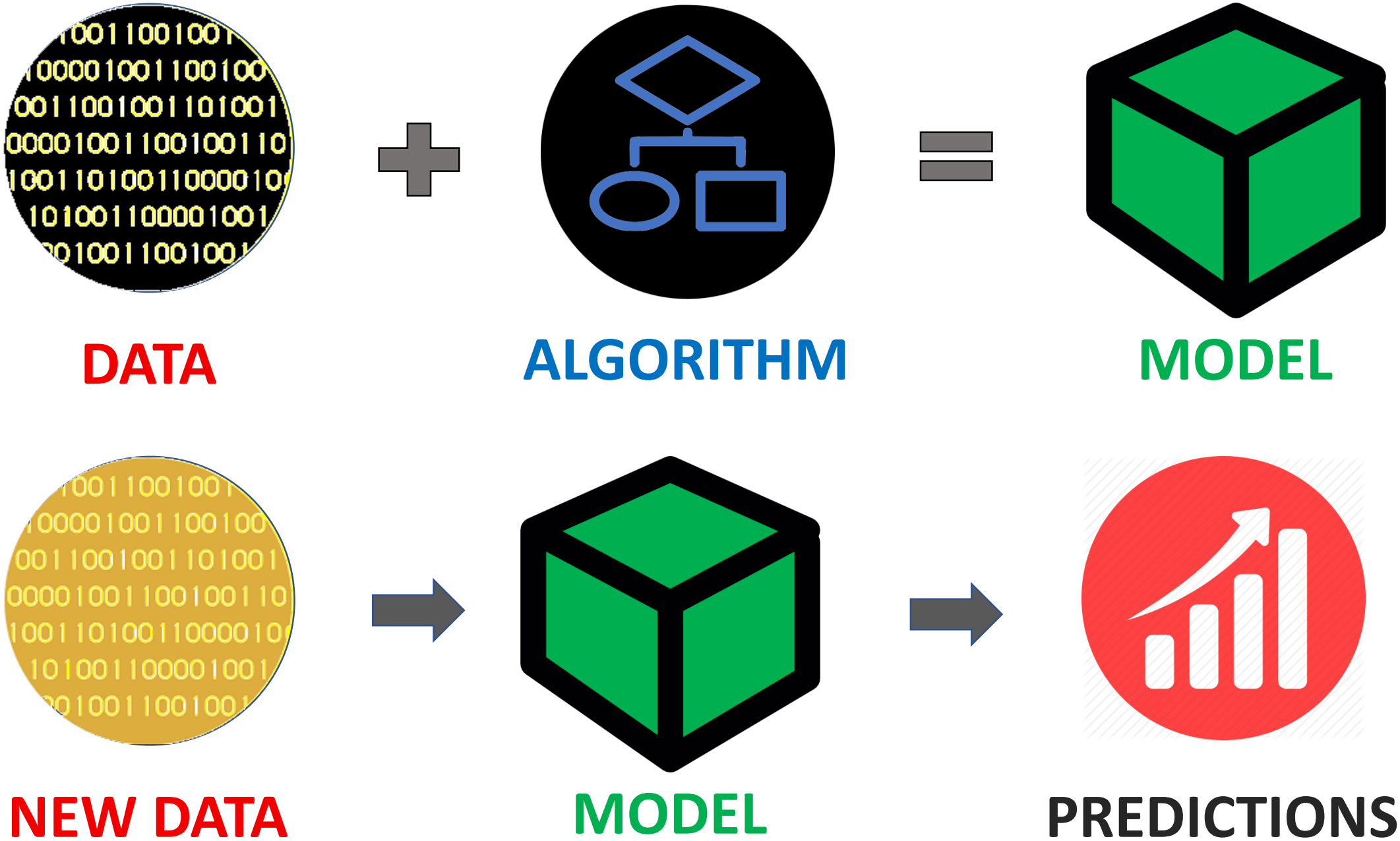
New input: **38** → **input** *1.8 + 32

Machine Learning

ML Model

New input: **38** → **input** *1.8 + 32 → output: **100.4**


DATA
ALGORITHM
MODEL


DATA
ALGORITHM
MODEL
NEW DATA
MODEL
PREDICTIONS

Types of Machine Learning

Supervised

Unsupervised

Music

Song	Artist	Genre	Liked
Breathing Light	Frameworks	Alternative Rock	Yes
Superior	Silver Maple	Pop	No
Icicle	AK	Pop	No
Jazzin	Flap Jack	R&B	Yes
The Way You Do	Schlomo	R&B	Yes
Mirror Maru	Cashmere	Rock	Yes
Never Too Far	Sorrow	Pop	No

Music

Features (X) →

Song	Artist	Genre	Liked
Breathing Light	Frameworks	Alternative Rock	Yes
Superior	Silver Maple	Pop	No
Icicle	AK	Pop	No
Jazzin	Flap Jack	R&B	Yes
The Way You Do	Schlomo	R&B	Yes
Mirror Maru	Cashmere	Rock	Yes
Never Too Far	Sorrow	Pop	No

Music

Song	Artist	Genre	Liked
Breathing Light	Frameworks	Alternative Rock	Yes
Superior	Silver Maple	Pop	No
Icicle	AK	Pop	No
Jazzin	Flap Jack	R&B	Yes
The Way You Do	Schlomo	R&B	Yes
Mirror Maru	Cashmere	Rock	Yes
Never Too Far	Sorrow	Pop	No

← Target (y)

Music

Song	Artist	Genre	Liked
Breathing Light	Frameworks	Alternative Rock	Yes
Superior	Silver Maple	Pop	No
Icicle	AK	Pop	No
Jazzin	Flap Jack	R&B	Yes
The Way You Do	Schlomo	R&B	Yes
Mirror Maru	Cashmere	Rock	Yes
Never Too Far	Sorrow	Pop	No

← Labels

Supervised

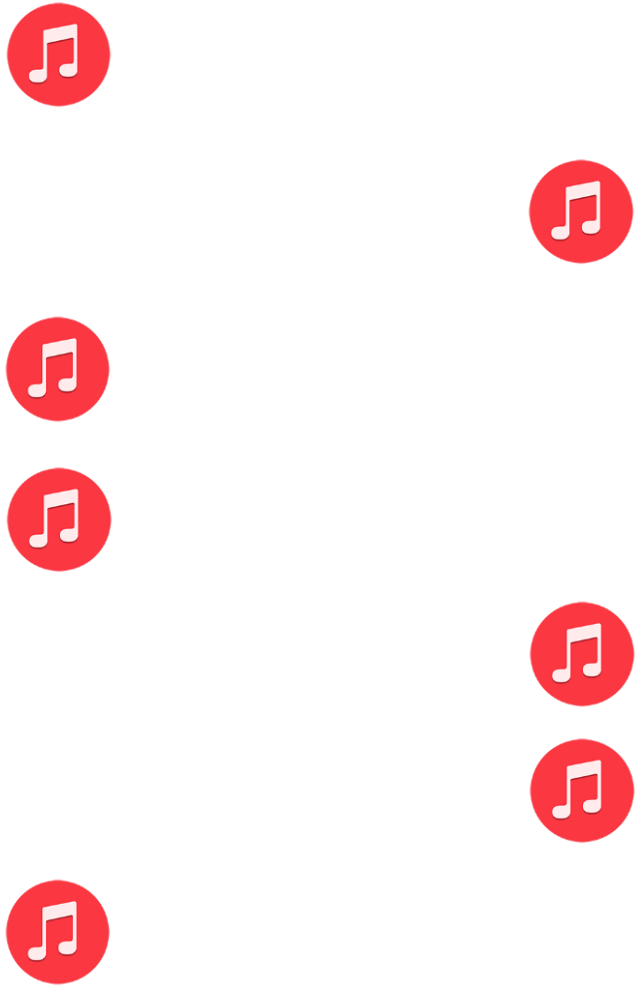
Features	Label
	Yes
	No
	No
	Yes
	Yes
	Yes
	No

Unsupervised

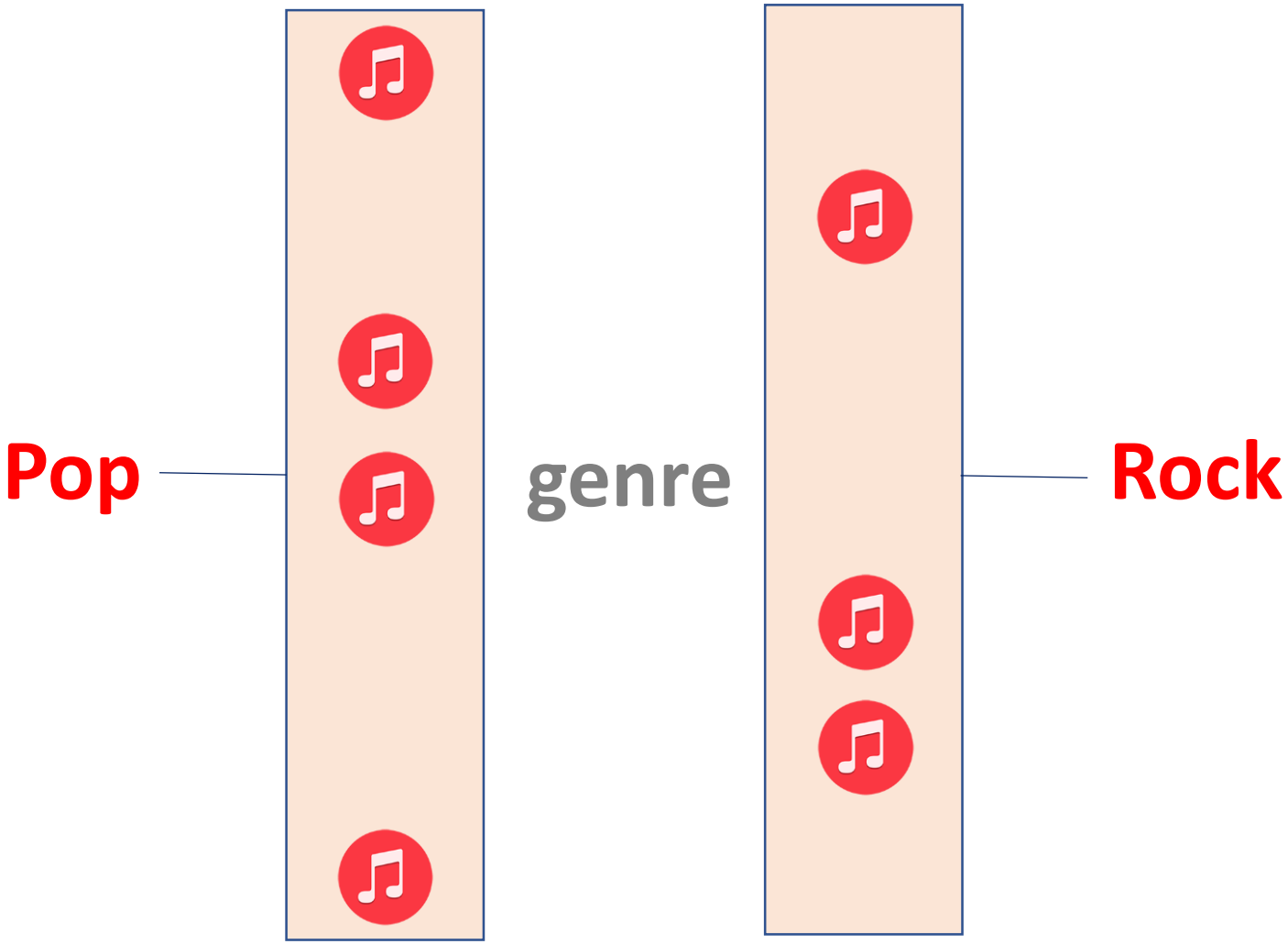
Features	Label

Clustering

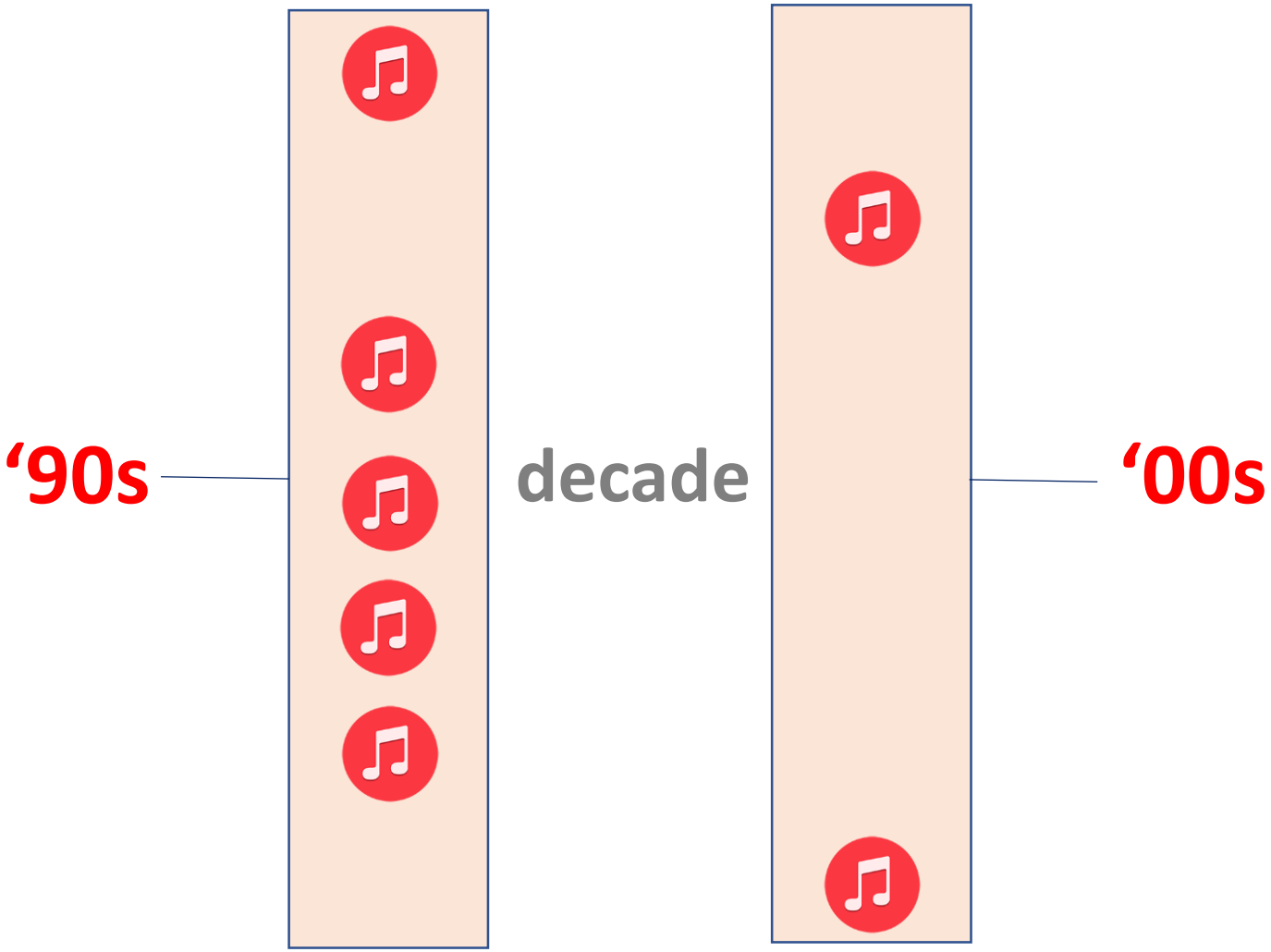
Clustering



Clustering



Clustering



Supervised	Unsupervised
Regression	Clustering
Classification	

Data

The best data has 3 qualities:

- Clean
- Coverage
- Complete

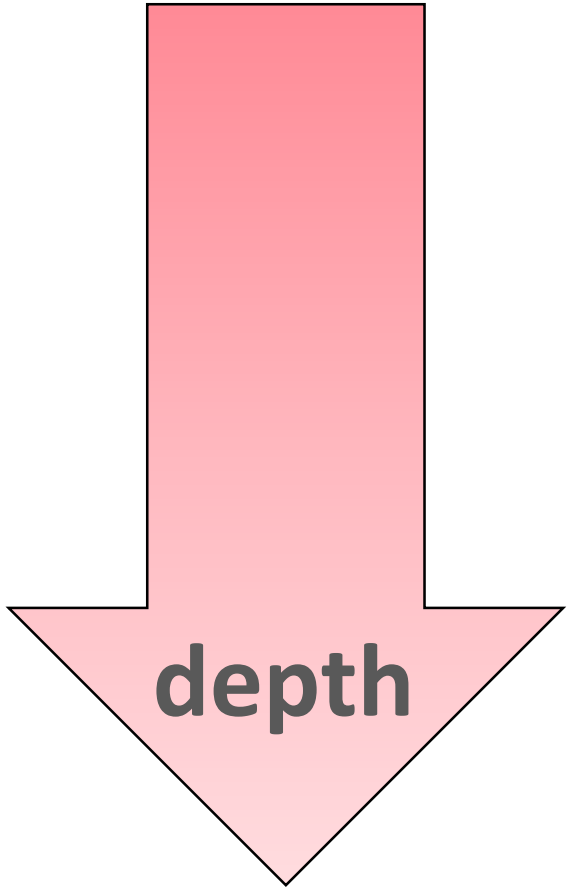
The best data has 3 qualities:

Feature 1	Feature 2	Feature 3	Feature 4
Male	200	1	Yes
Female	316	3	No
F	190	1	No
Male	244		Yes
Male	128	2	Yes
Male		3	Yes
Female	302	2	No

Clean

Feature 1	Feature 2	Feature 3	Feature 4
Male	200	1	Yes
Female	316	3	No
F	190	1	No
Male	244	13	Yes
Male	128	2	Yes
Male		3	Yes
Female	302	2	No

Coverage



Feature 1	Feature 2	Feature 3	Feature 4
Male	200	1	Yes
Female	316	3	No
F	190	1	No
Male	244		Yes
Male	128	2	Yes
Male		3	Yes
Female	302	2	No

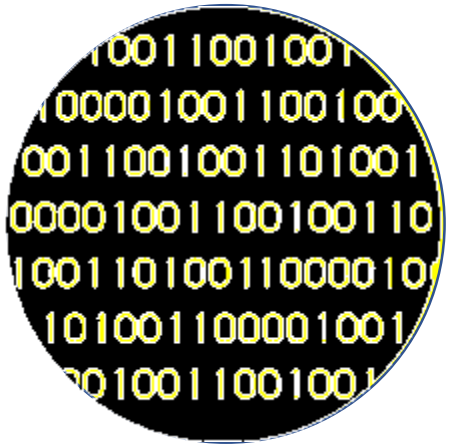
Complete

breadth

Feature 1	Feature 2	Feature 3	Feature 4
Male	200	1	Yes
Female	316	3	No
F	190	1	No
Male	244		Yes
Male	128	2	Yes
Male		3	Yes
Female	302	2	No

Model Training

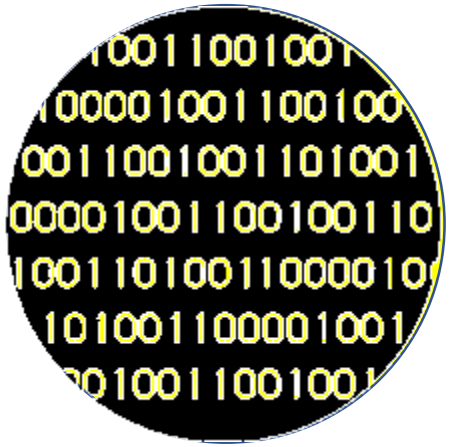
Model Training



DATA

MODEL

Model Training

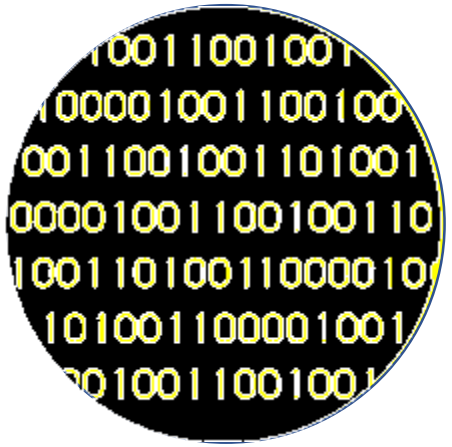


DATA

MODEL

Prediction
0
1
0
0
1
0

Model Training

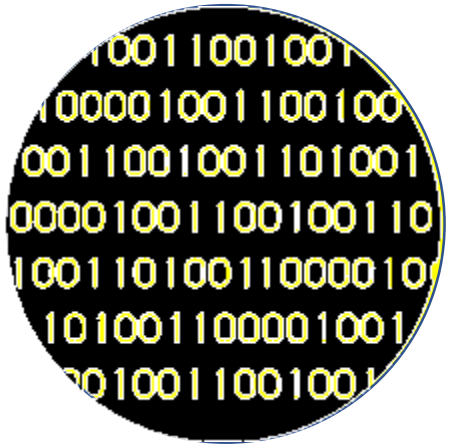


DATA

MODEL

Prediction	Label
0	1
1	1
0	0
0	1
1	0
0	0

Model Training

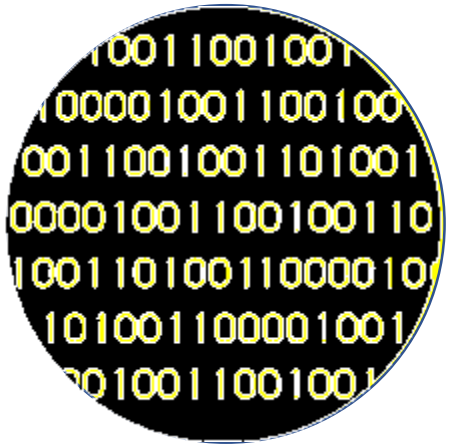


DATA

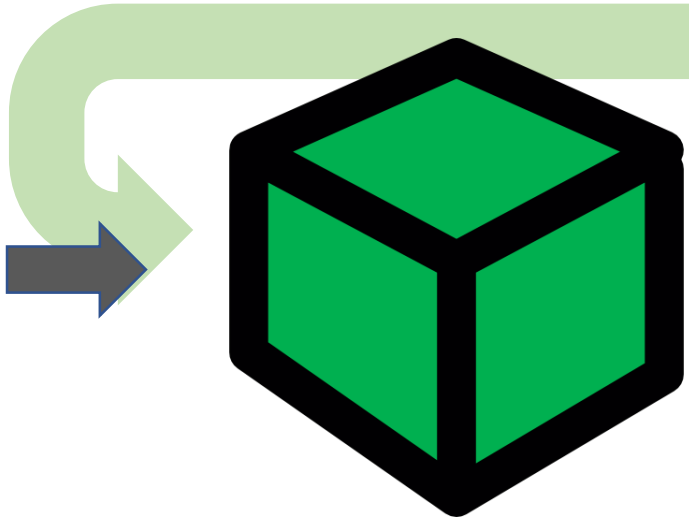
MODEL

Prediction	Label
0	1
1	1
0	0
0	1
1	0
0	0

Model Training



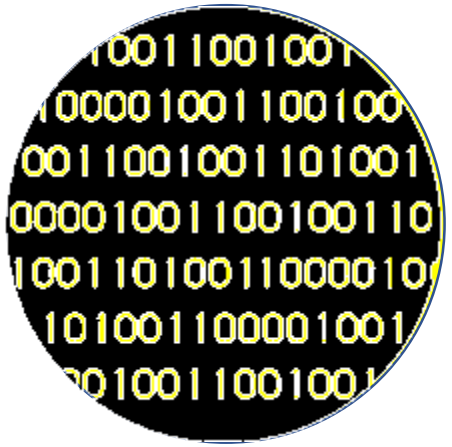
DATA



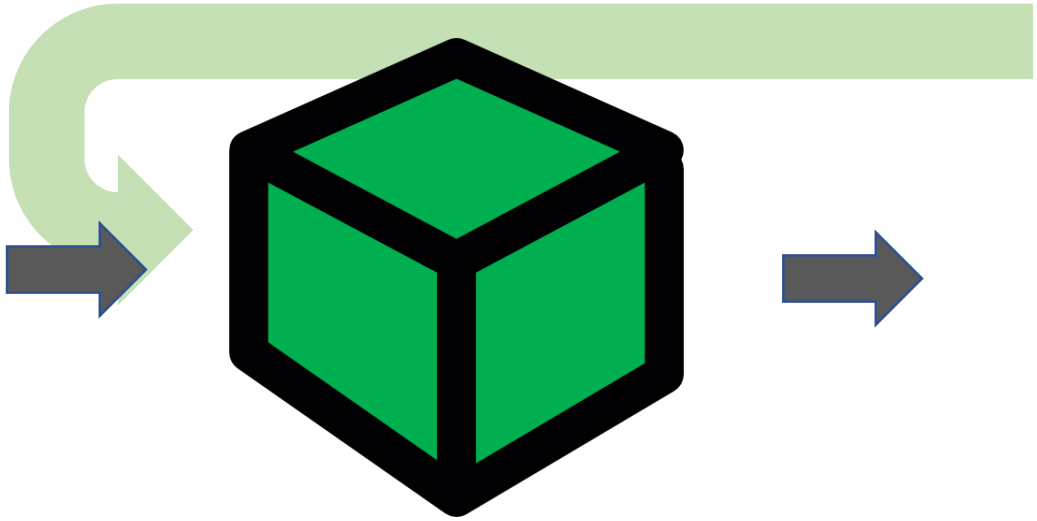
MODEL

Prediction	Label
0	1
1	1
0	0
0	1
1	0
0	0

Model Training



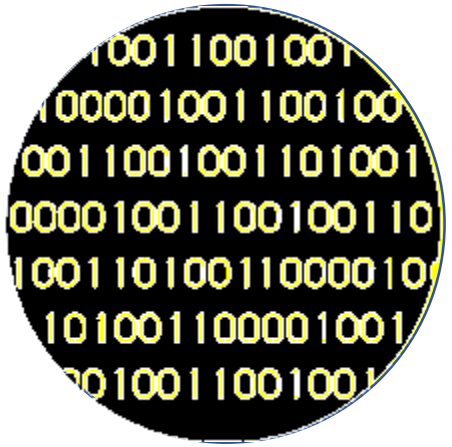
DATA



MODEL

Prediction	Label
1	1
1	1
0	0
1	1
0	0
0	0

Model Training



DATA

MODEL

Prediction	Label
1	1
1	1
0	0
1	1
0	0
0	0

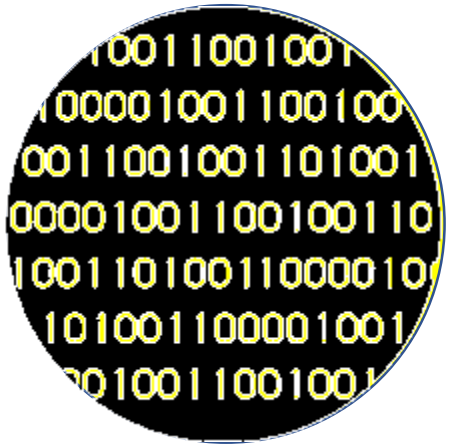
Model Training

TRAINED MODEL

Model Training

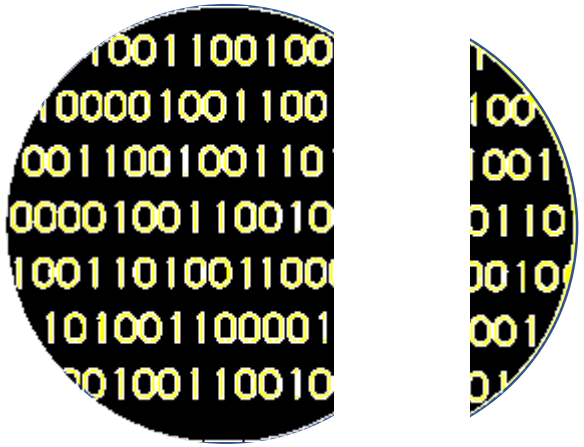
MODEL MODEL MODEL MODEL MODEL

Evaluate the Model

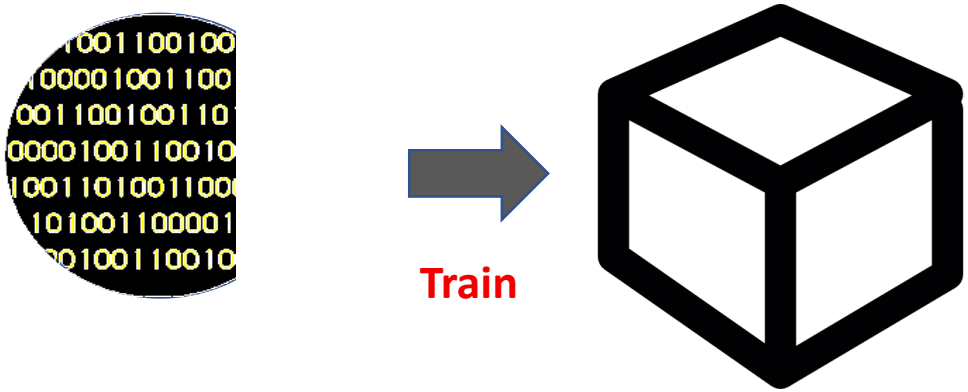


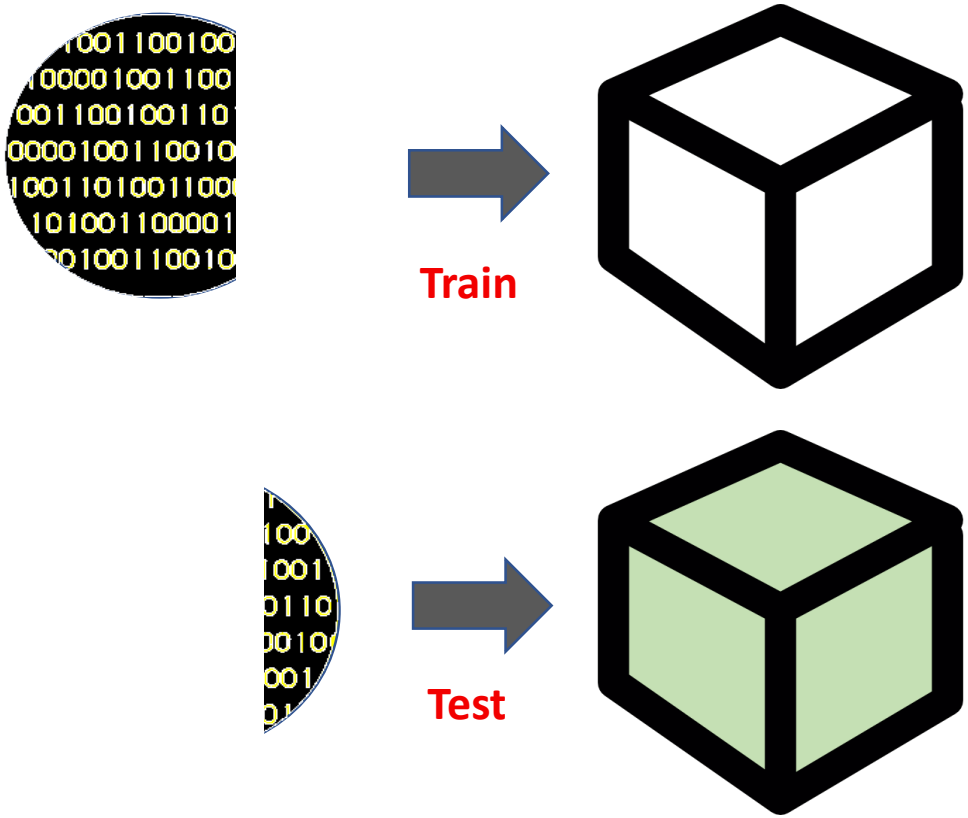
Evaluate the Model

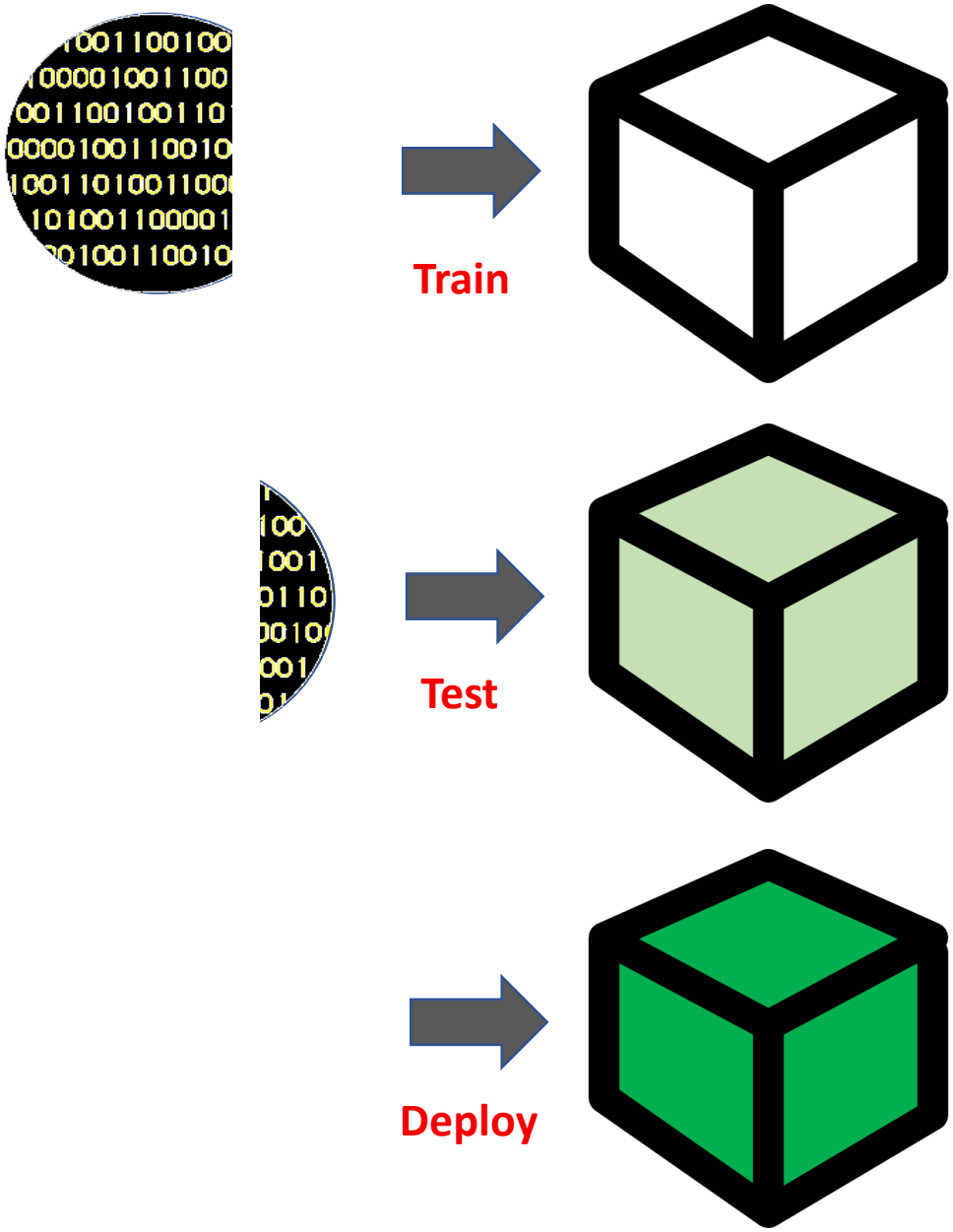
Training Data



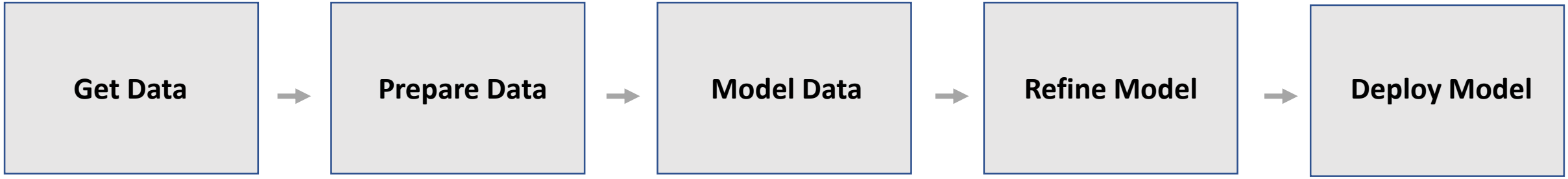
Test Data


Train

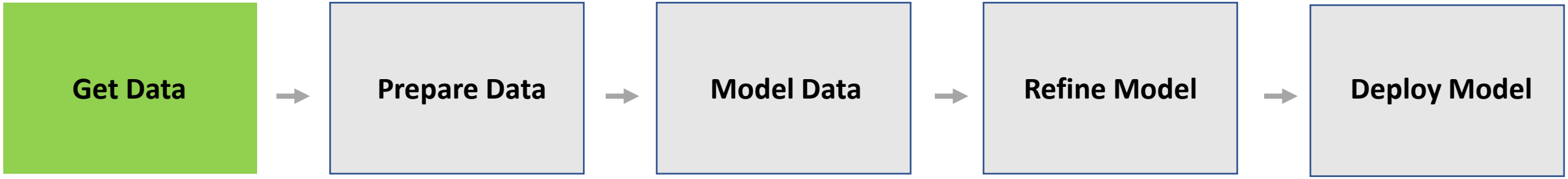

Train
Test


Train
Test
Deploy

ML Process

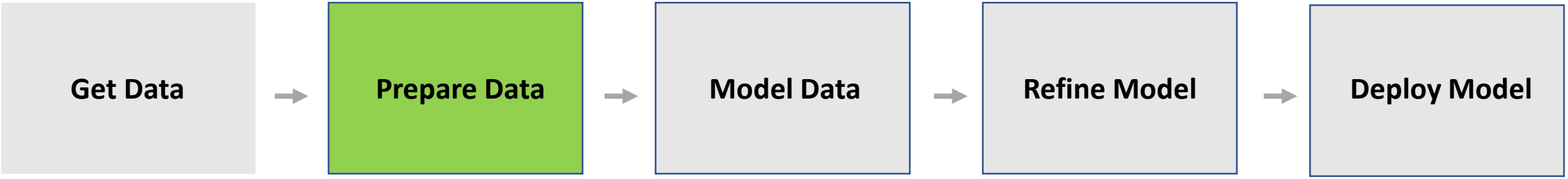


ML Process



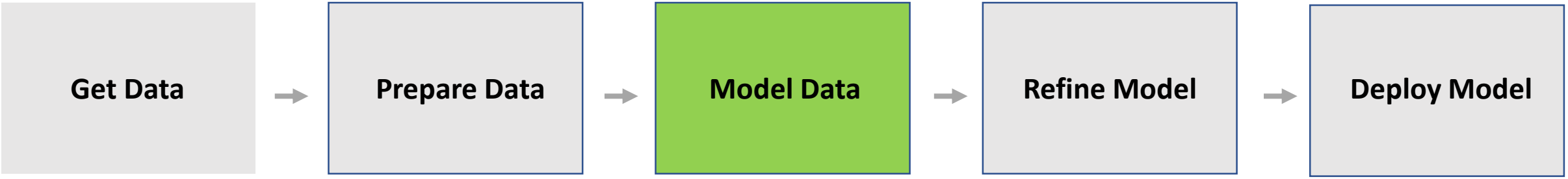
What data should you use?
Is it labeled?

ML Process



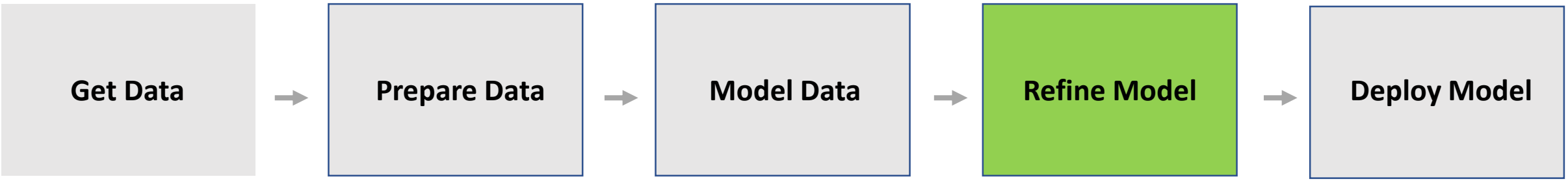
Is your data **complete**, **clean**, does it have **coverage**?

ML Process



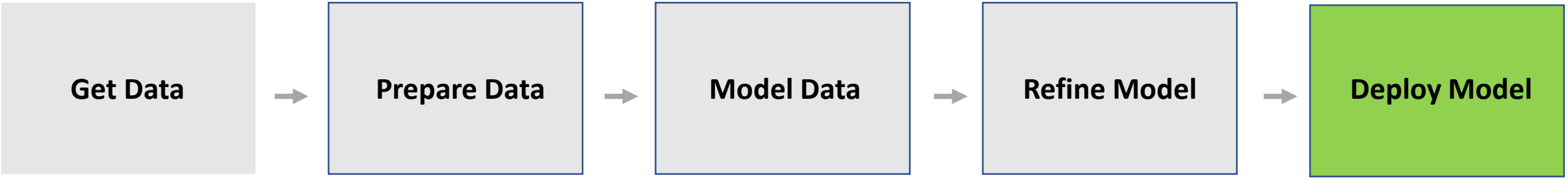
Which algorithms should you use?

ML Process



What level of performance is sufficient?

ML Process



Make predictions.