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

Machine Learning

What is Machine Learning?

Traditional Software Development

Traditional Software Development

Convert **inches** to **cm**

Traditional Software Development

Convert **inches** to **cm**

Input:

Output:

Traditional Software Development

Convert to **cm**

Input: **inches**

Output:

Traditional Software Development

Input: **inches**

Relationship: **cm** =

Output:

Traditional Software Development

Input: **inches**

Relationship: **cm** = **inches** * 2.54

Output:

Traditional Software Development

Input: **inches**

Relationship: **cm** = **inches** * 2.54

Output: **cm**

Traditional Software Development

Traditional Software Development

Convert a **number** to its **absolute value**

Traditional Software Development

Convert a **number** to its **absolute value**

Input:

Output:

Traditional Software Development

Convert a to its **absolute value**

Input: **number**

Output:

Traditional Software Development

Convert a to its **absolute value**

Input: **number**

Rules:

Output:

Traditional Software Development

Input: **number**

Rules:

abs. value =

Output:

Traditional Software Development

```
Input: number

Rules:
    if number >= 0:  abs. value = number


Output:
```

Traditional Software Development

Input: **number**

Rules:

```
    if number >= 0:  abs. value = number
    else:            abs. value = number * -1
```

Output:

Traditional Software Development

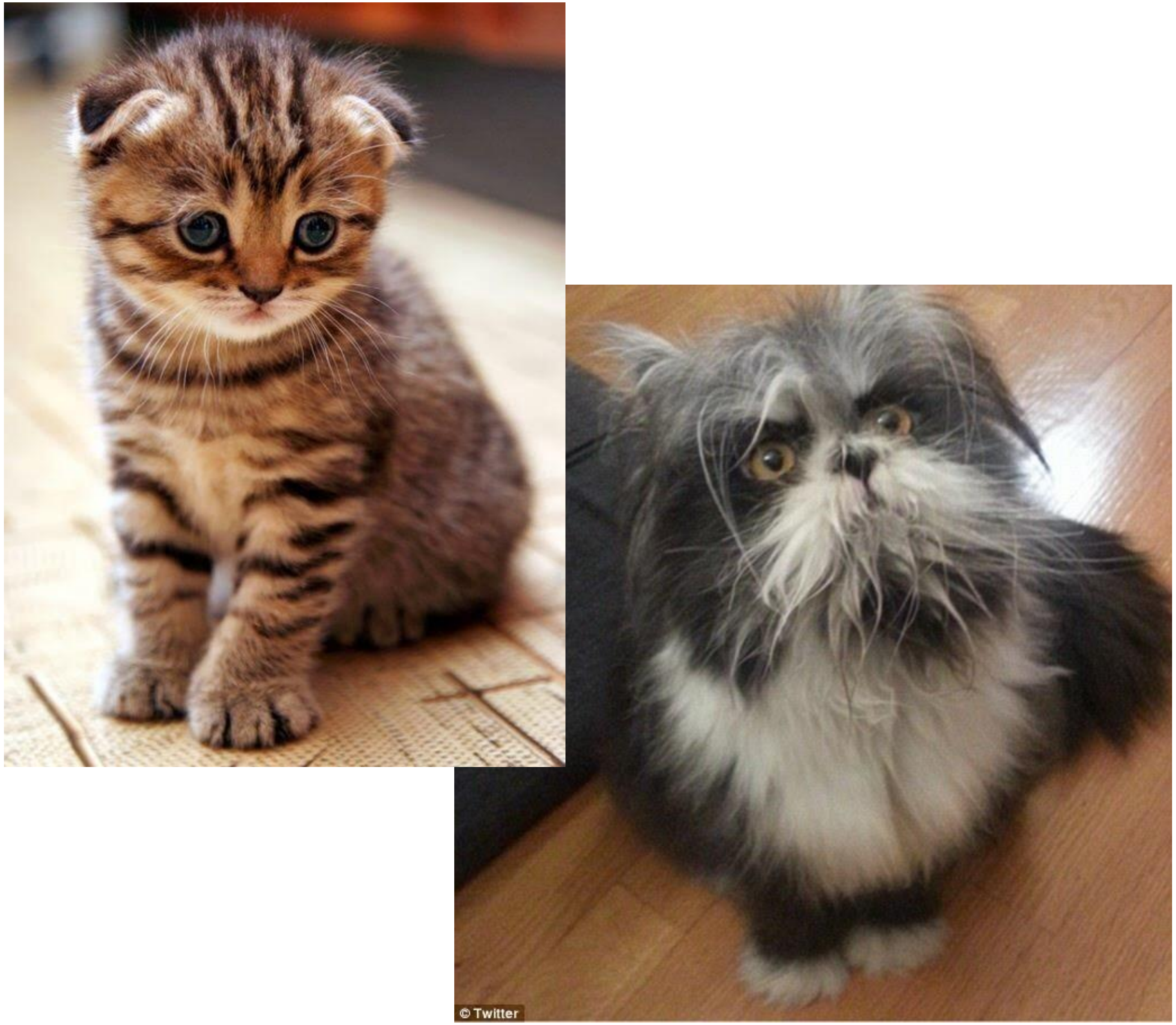
```
Input: number

Rules:
    if number >= 0:  abs. value = number
    else:            abs. value = number * -1

Output: abs. value
```

Traditional Software Development

Traditional Software Development



© Twitter

© Twitter

Traditional Software Development

Input:

Rules:

Output:

Traditional Software Development

Input:

Rules:

rule 1
rule 2
rule 3 …

Output:

Traditional Software Development

Input:

Rules: rule 1
rule 2
rule 3 …

Output: “cat”

Machine Learning

Machine Learning

Input	0	8	15	22	38
Output	32	46.4	59	71.6	?

Machine Learning

Input	**0**	**8**	**15**	**22**	**38**
Output	32	46.4	59	71.6	100.4

$$F = C * 1.8 + 32$$

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**]

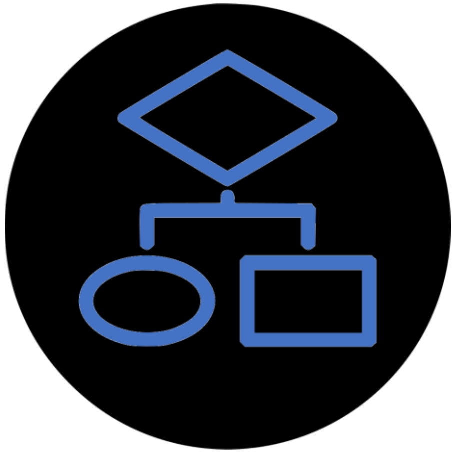
Output: [**32, 46.4, 59, 71.6**]

Machine Learning

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

Relationship: **?**

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



=

Common ML Algorithms

Linear Regression

Logistic Regression

Naïve Bayes

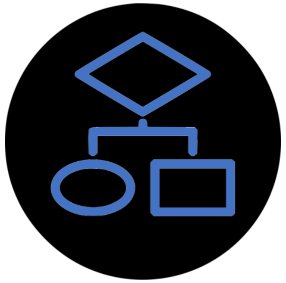
Support Vector Machine

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**

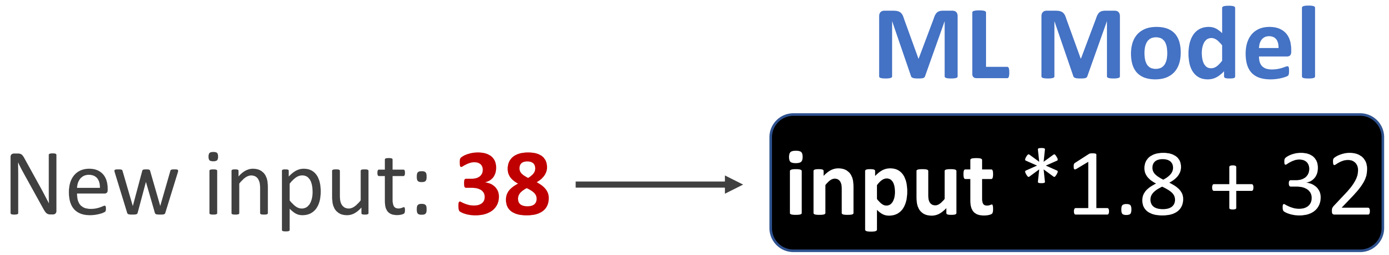
Output: [**32, 46.4, 59, 71.6**]

Machine Learning

ML Model

input *1.8 + 32

Machine Learning

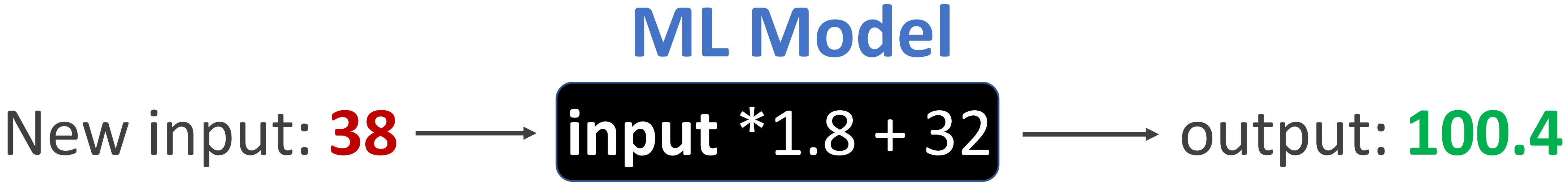


ML Model

New input: **38**

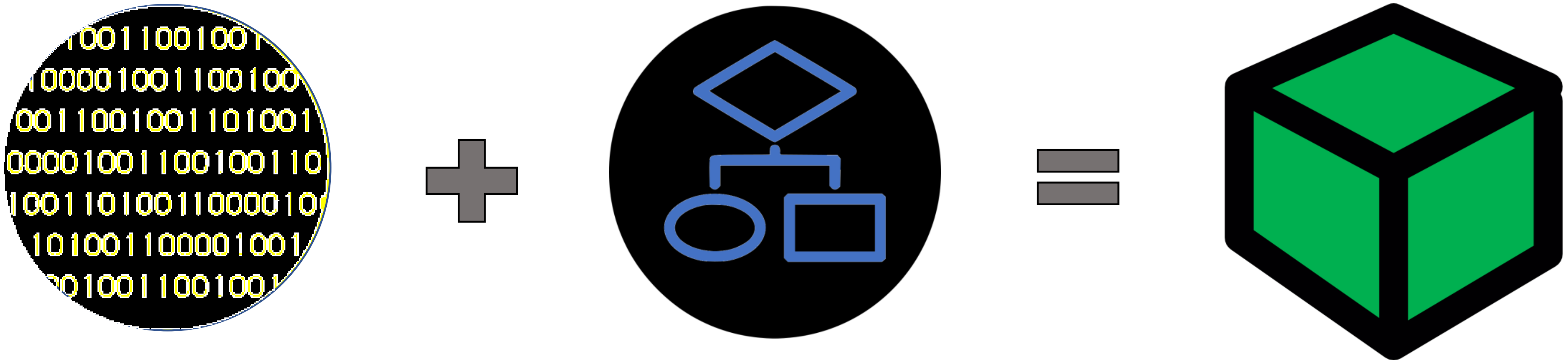
input *1.8 + 32

Machine Learning

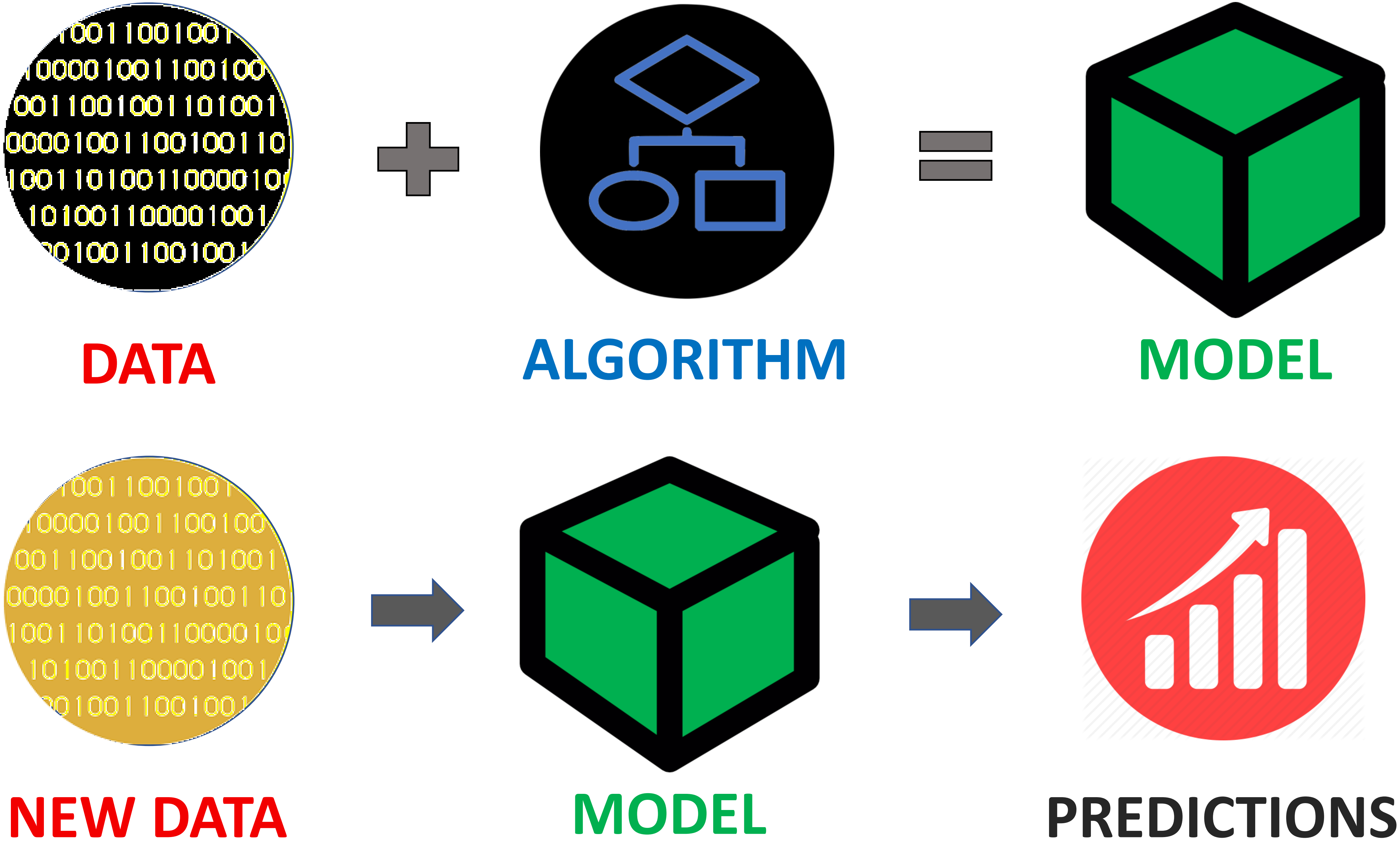


ML Model

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



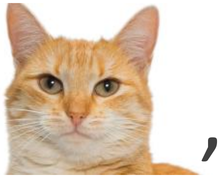
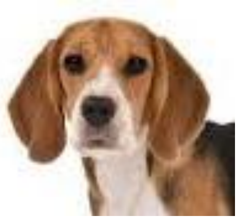
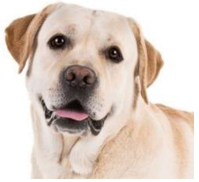
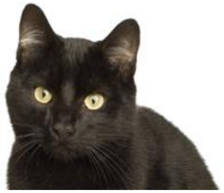
DATA + ALGORITHM = MODEL



DATA + ALGORITHM = MODEL

NEW DATA → MODEL → PREDICTIONS

Machine Learning

Input:[ ,  ,  , ]

Relationship:

Output: [“cat”, “dog”, “dog”, “cat]

Machine Learning

Input:[, , ,]

Relationship:

Output: [“cat”, “dog”, “dog”, “cat]

Pareidolia

Machine Learning

Instead of programming a computer, you give a computer examples and it **learns** what you want.

Why ML Now?

Why ML Now?

- Increasing availability of data

Why ML Now?

- Increasing availability of data
- Sophistication of ML algorithms

Why ML Now?

- Increasing availability of data
- Sophistication of ML algorithms
- Increasing power and availability of computing hardware and software

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

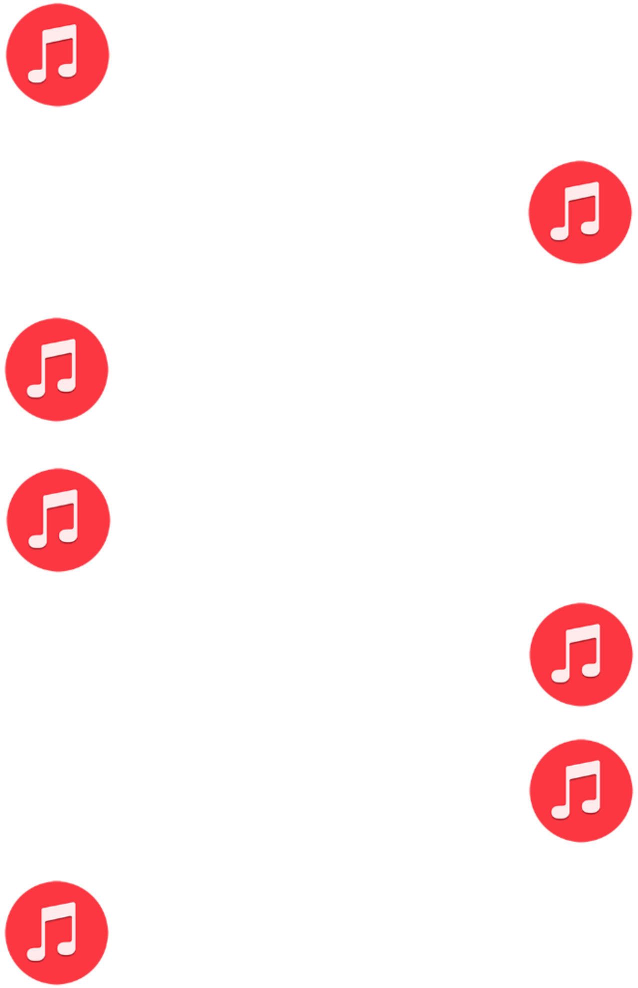
Features	Label
♫	Yes
♫	No
♫	No
♫	Yes
♫	Yes
♫	Yes
♫	No

Unsupervised

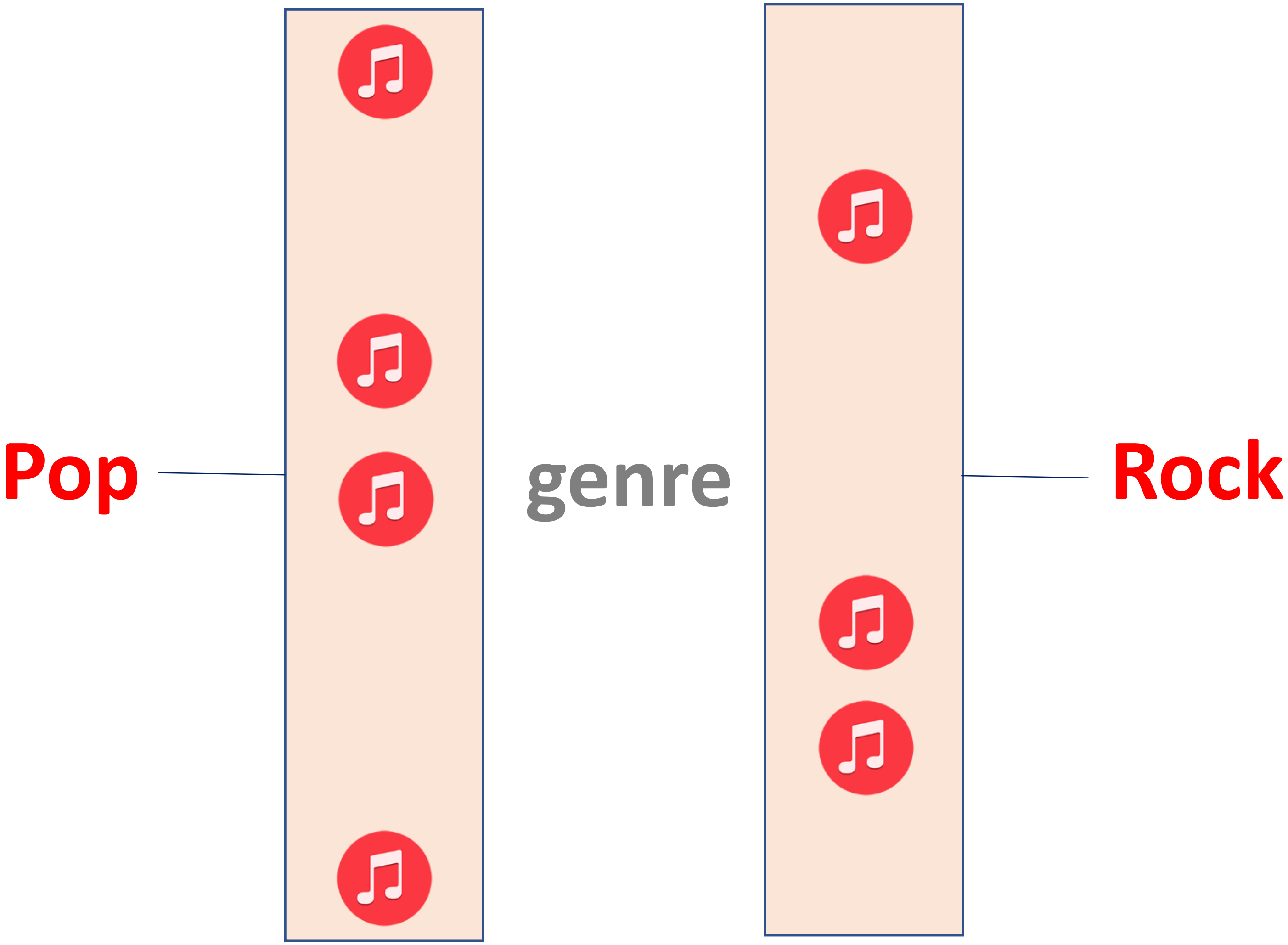
Features	Label
♫	
♫	
♫	
♫	
♫	
♫	
♫	

Clustering

Clustering



Clustering

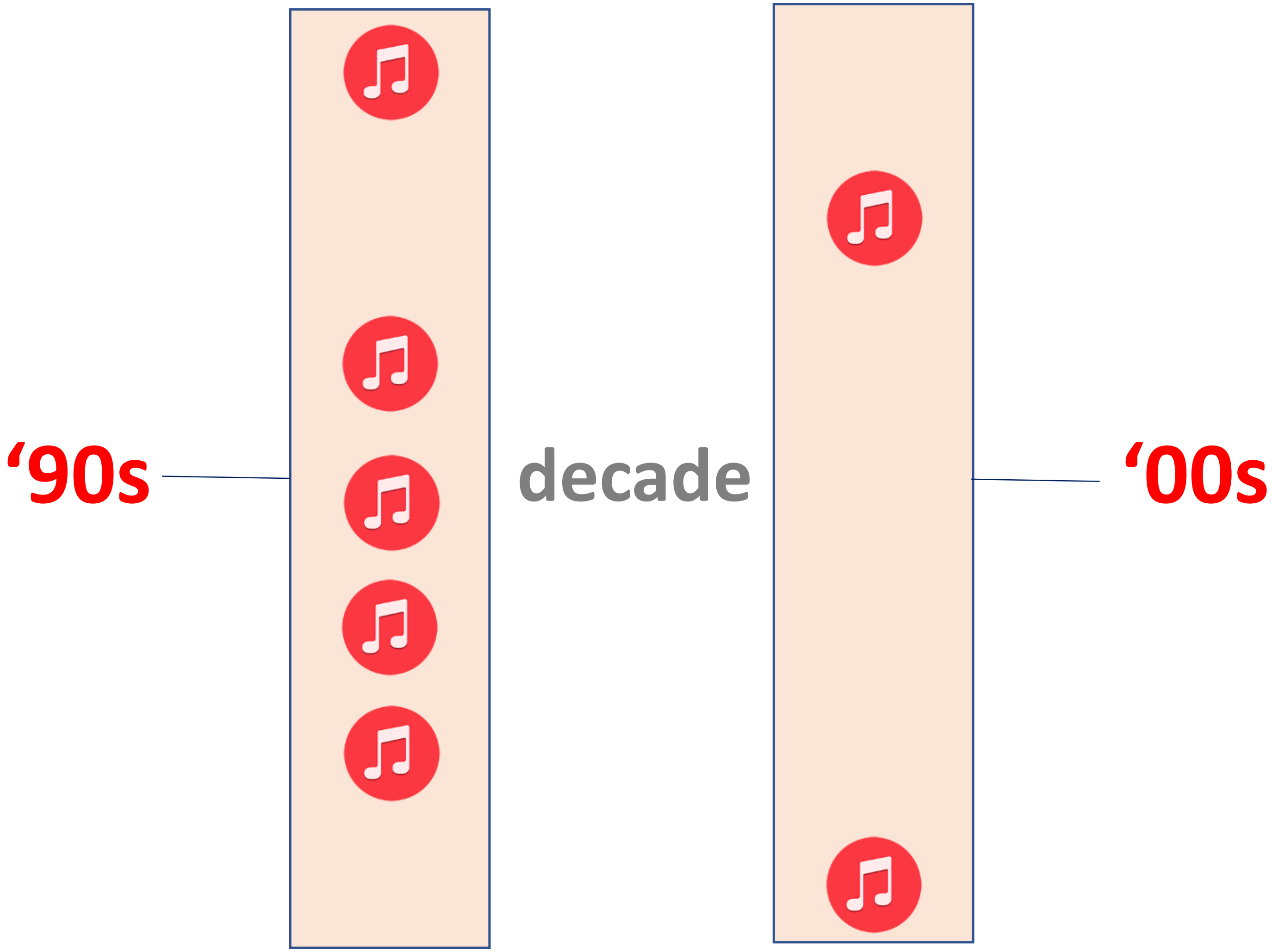


Pop

genre

Rock

Clustering



Supervised

Regression

Classification

Unsupervised

Clustering

Data

The best data has 3 qualities:

- Clean
- Coverage
- Complete

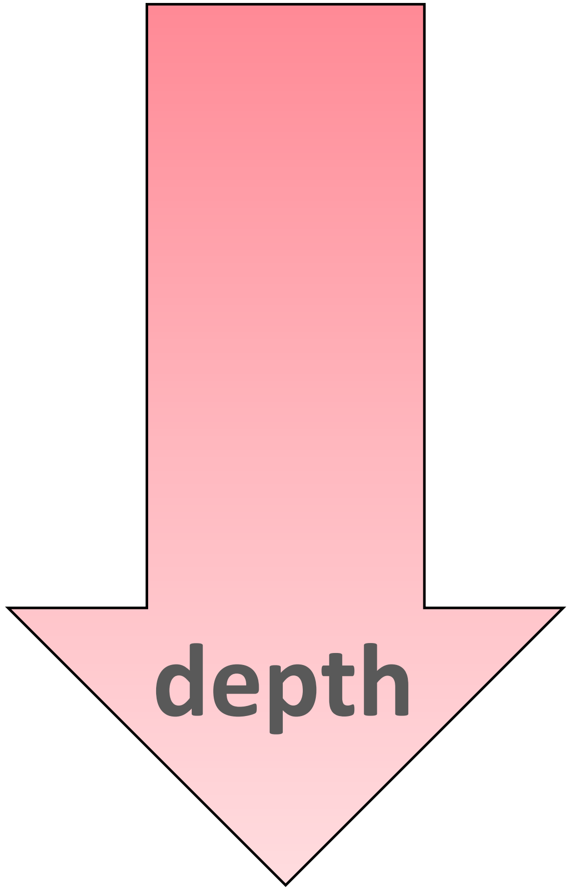
The best data has 3 qualities:

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

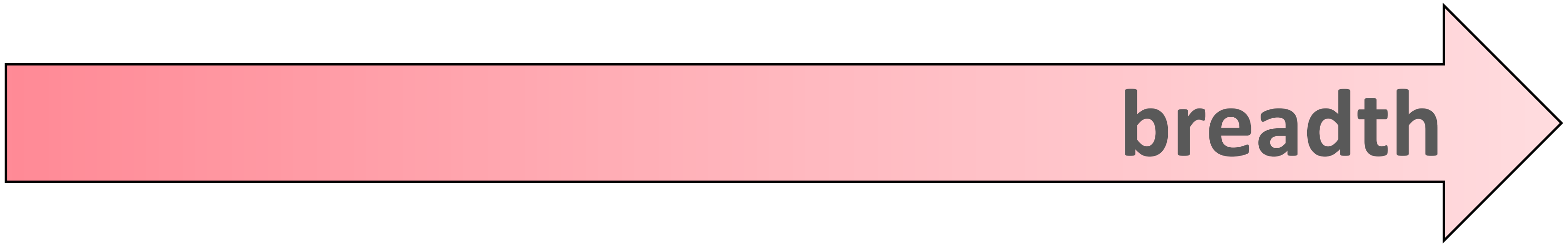
Coverage



depth

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



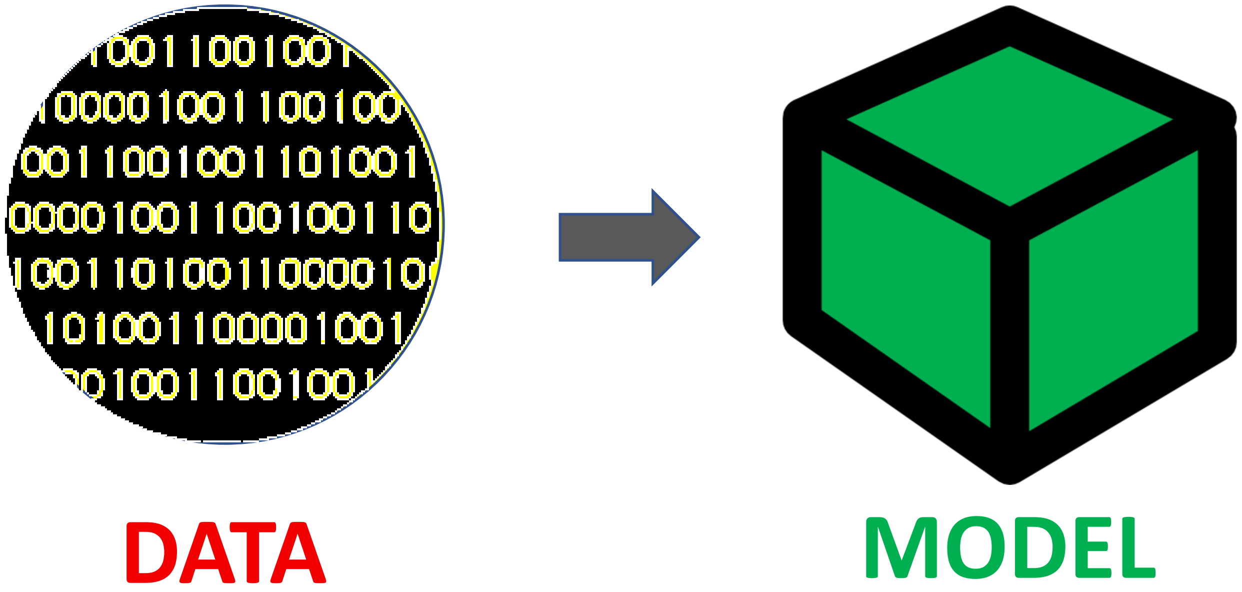
breadth

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

If ML is a rocket engine, data is the fuel

Model Training

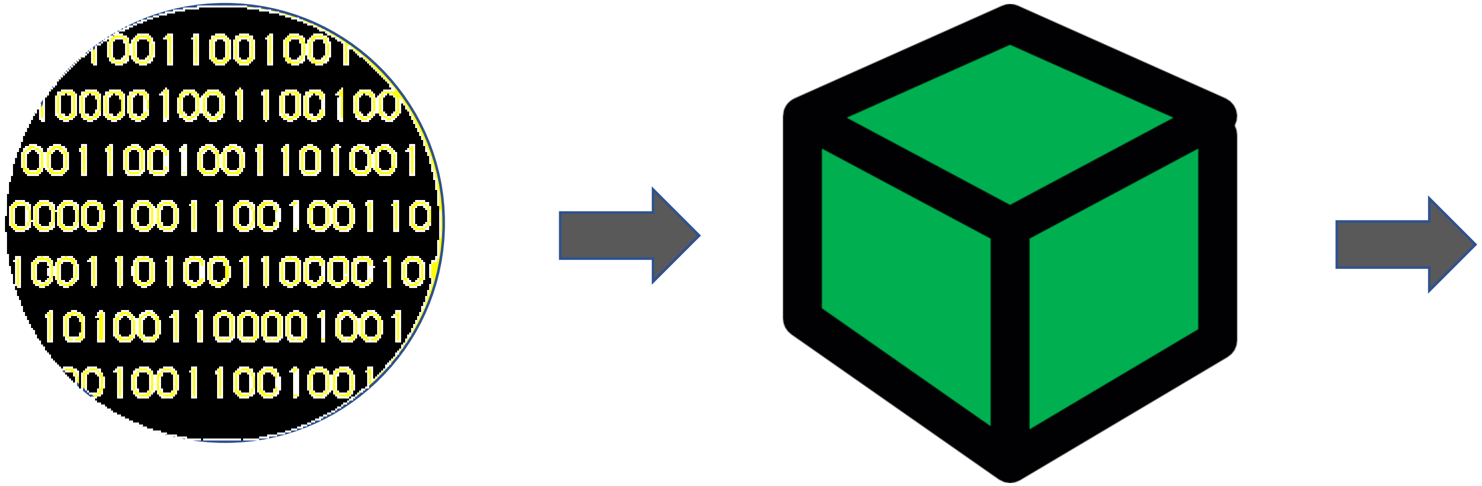
Model Training



DATA

MODEL

Model Training

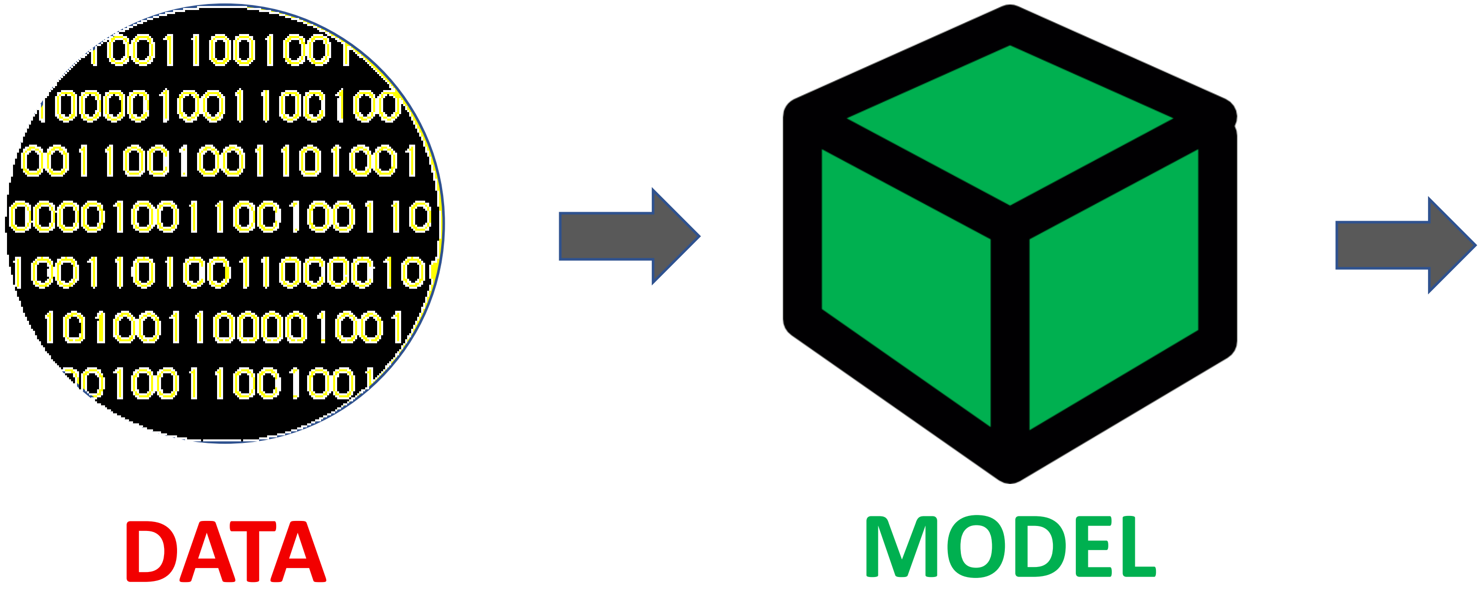


DATA

MODEL

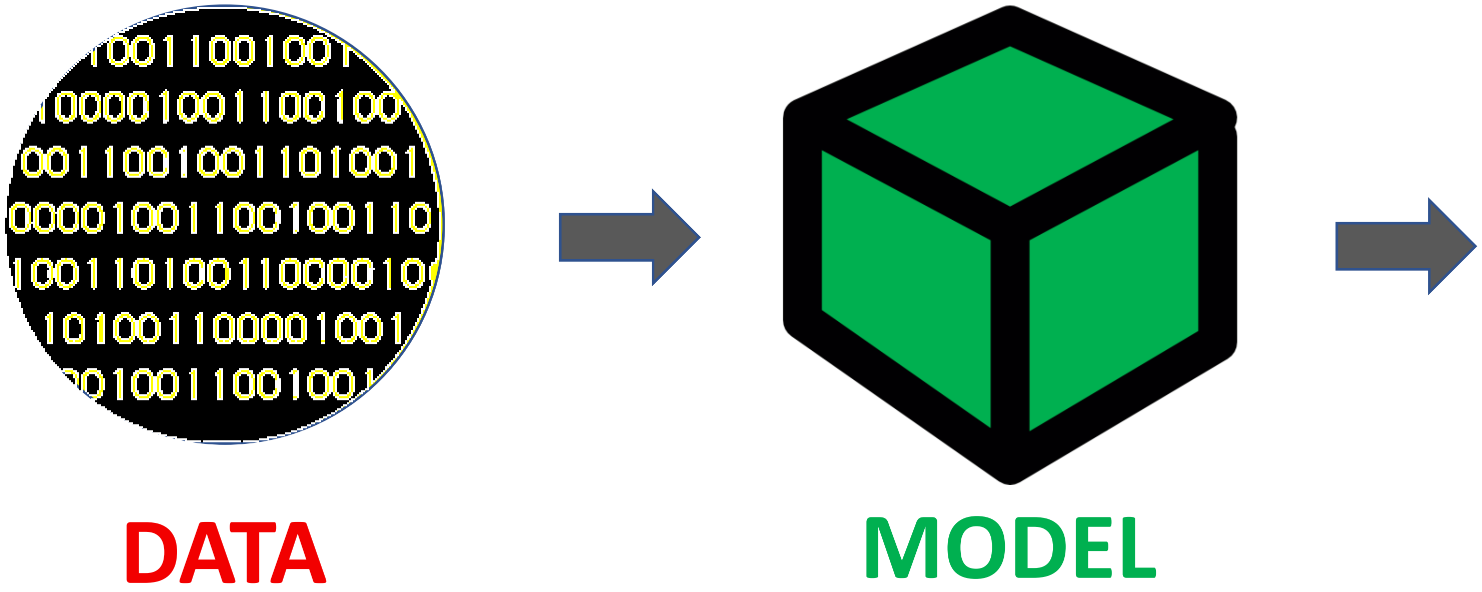
Prediction
0
1
0
0
1
0

Model Training



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

Model Training

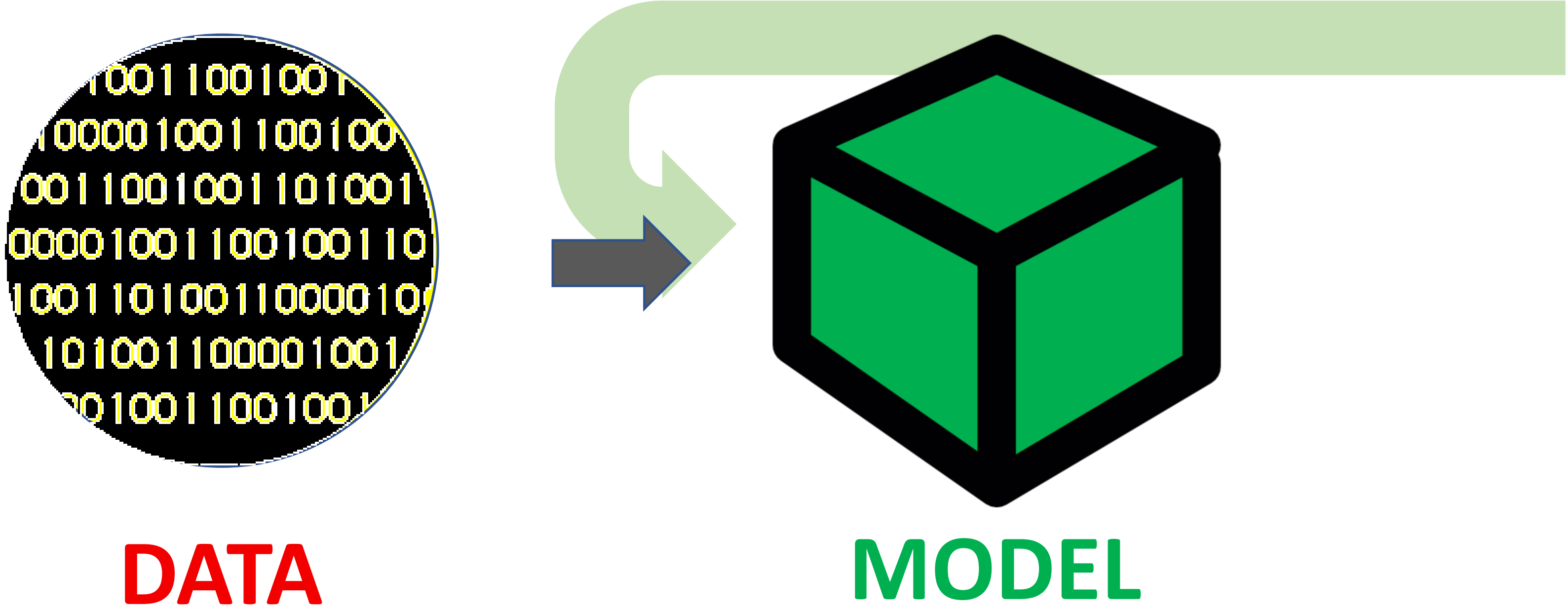


DATA

MODEL

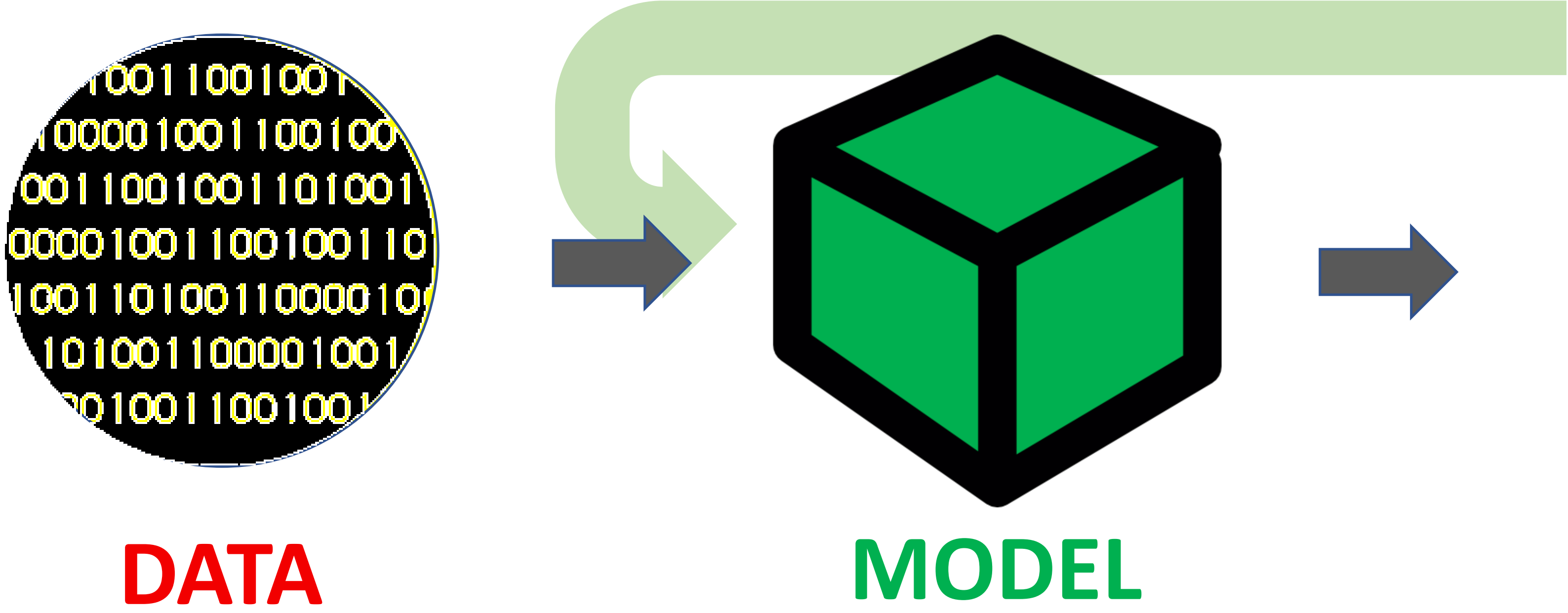
Prediction	Label
0	1
1	1
0	0
0	1
1	0
0	0

Model Training



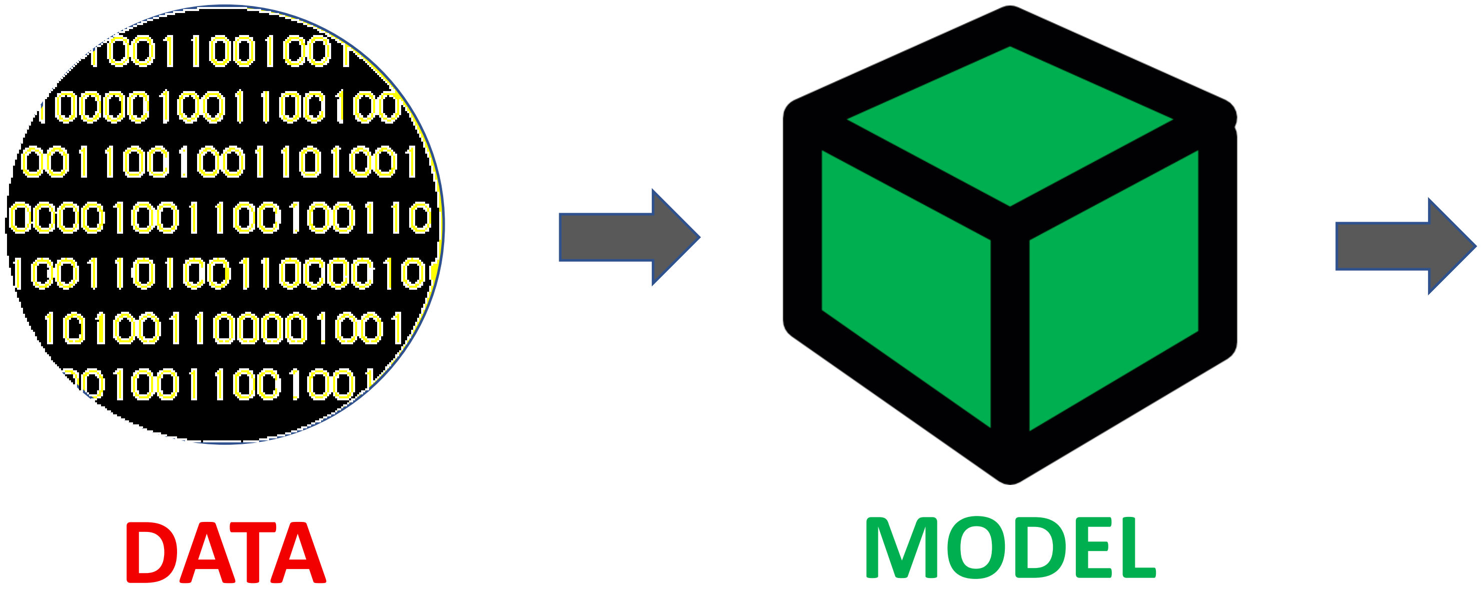
Prediction	Label
0	1
1	1
0	0
0	1
1	0
0	0

Model Training



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

Model Training



DATA

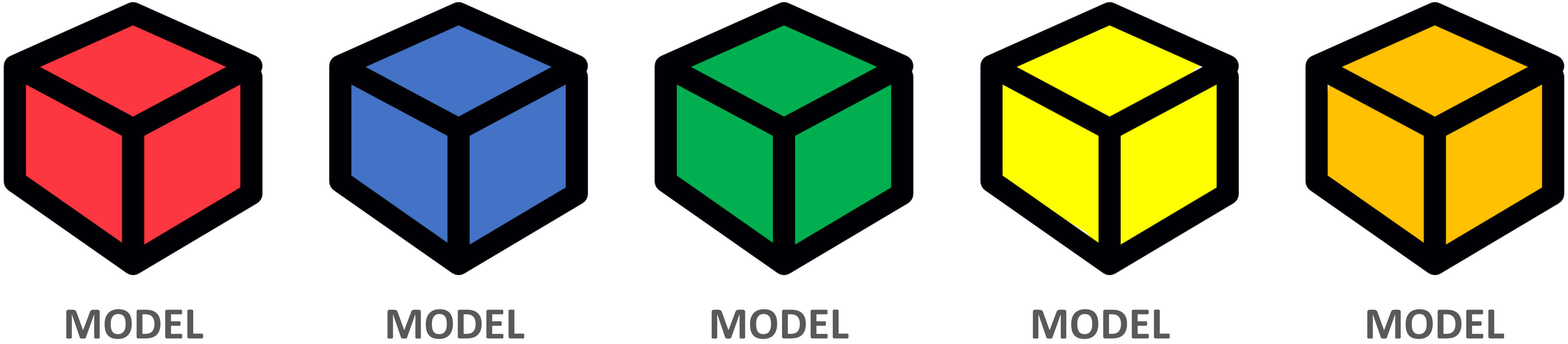
MODEL

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

Model Training

TRAINED MODEL

Model Training



MODEL

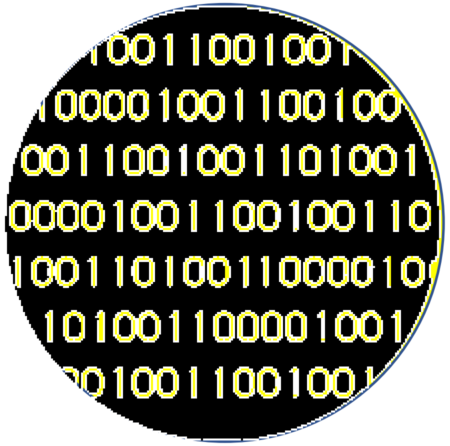
MODEL

MODEL

MODEL

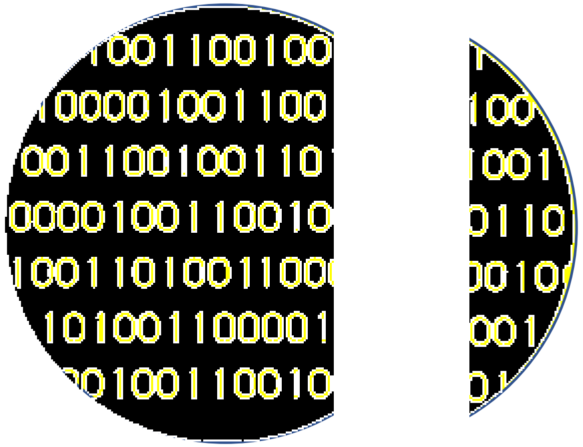
MODEL

Evaluate the Model

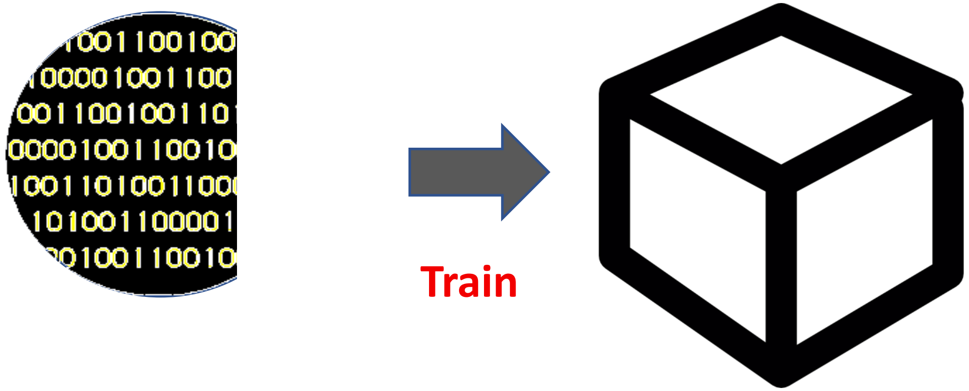


Evaluate the Model

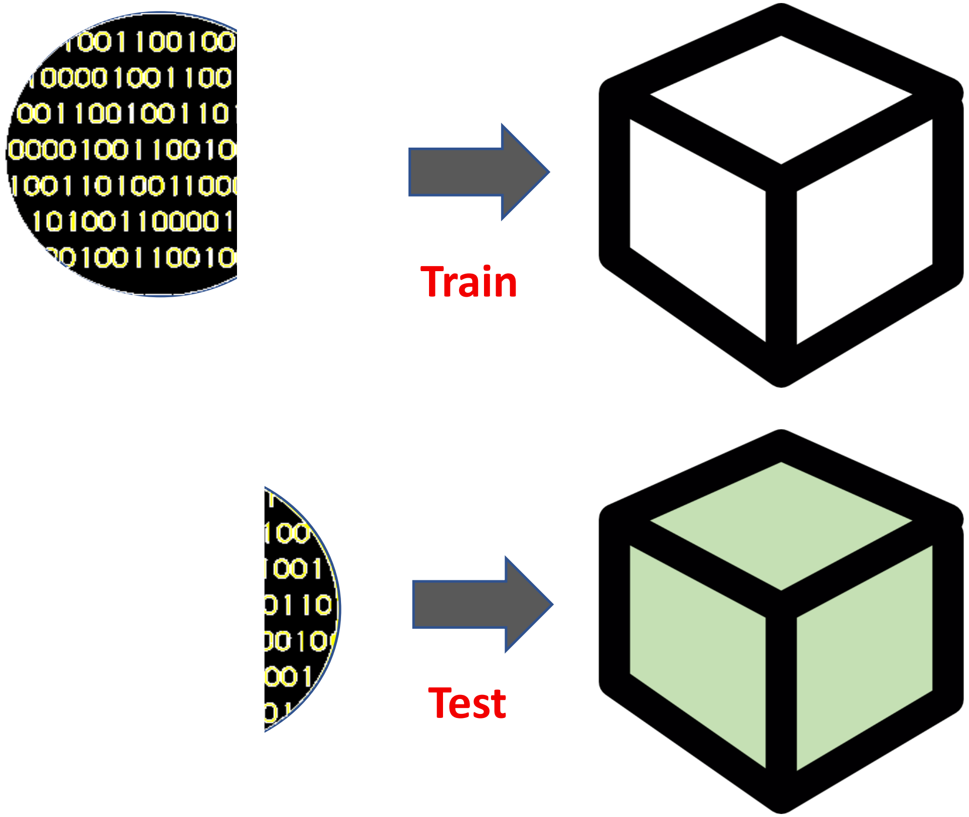
Training Data



Test Data

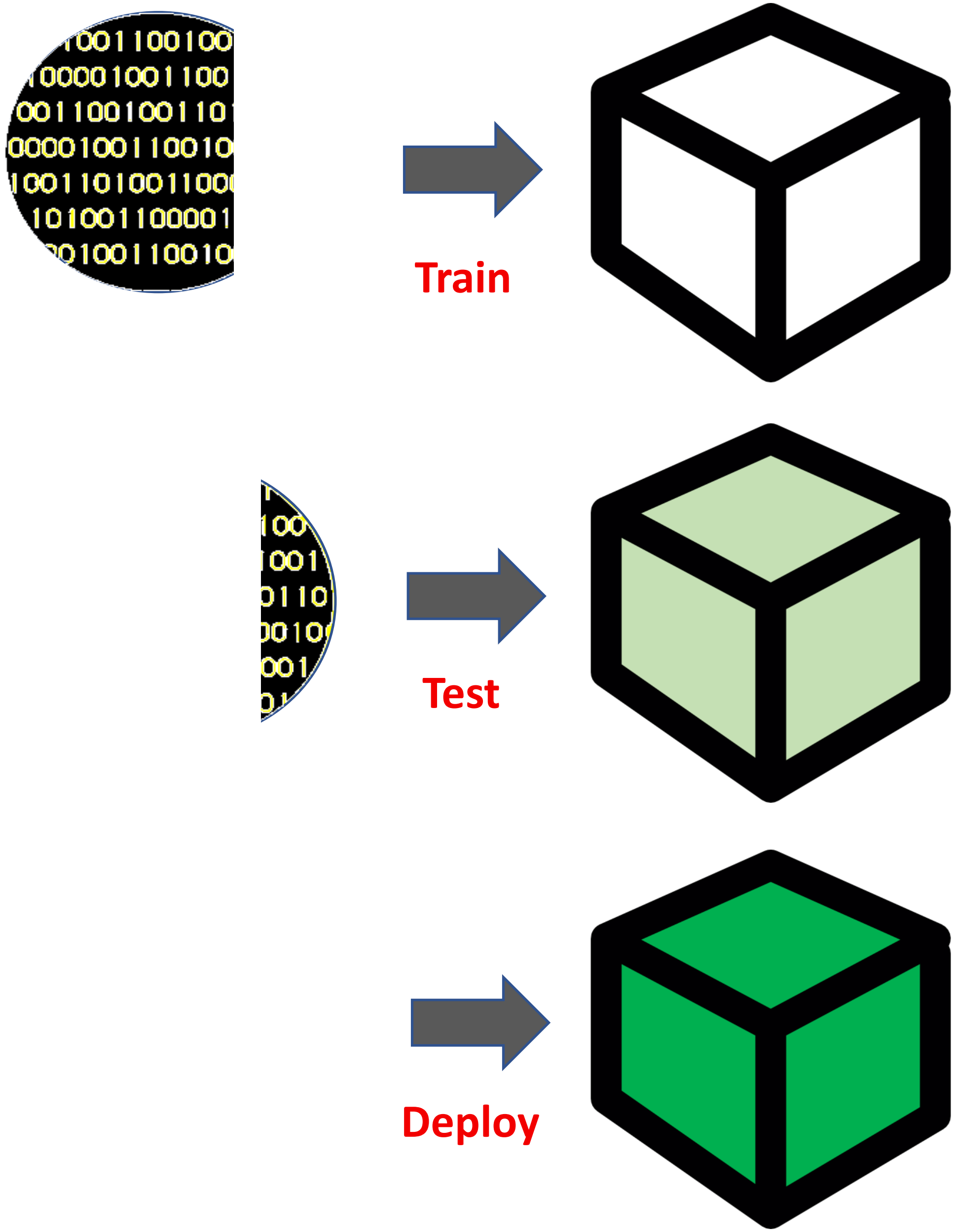


Train



Train

Test

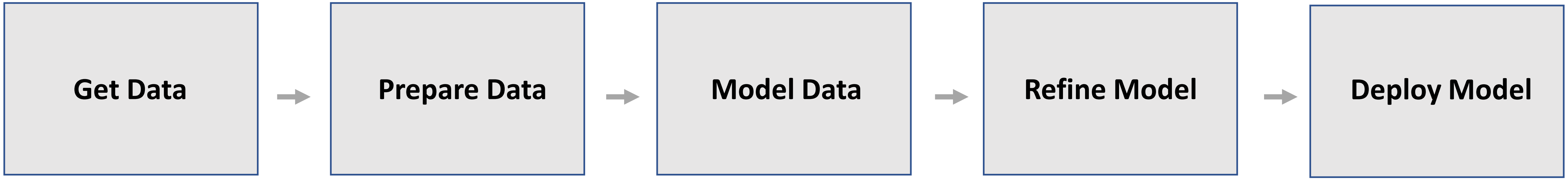


Train

Test

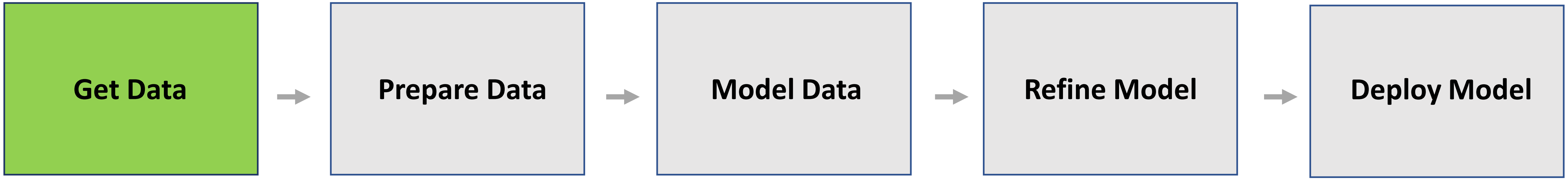
Deploy

ML Process



Get Data → Prepare Data → Model Data → Refine Model → Deploy Model

ML Process

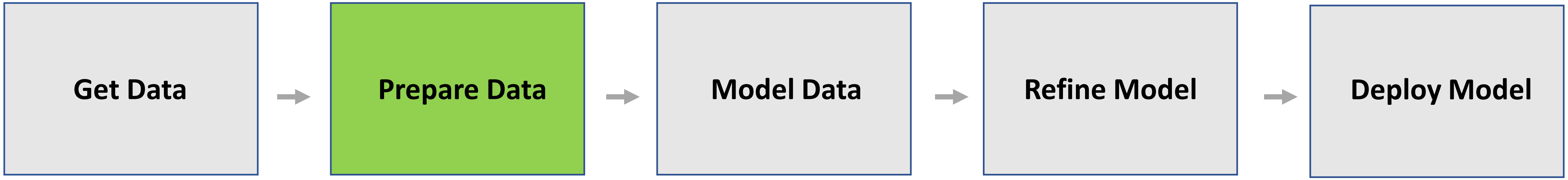


Get Data → Prepare Data → Model Data → Refine Model → Deploy Model

What data should you use?

Is it labeled?

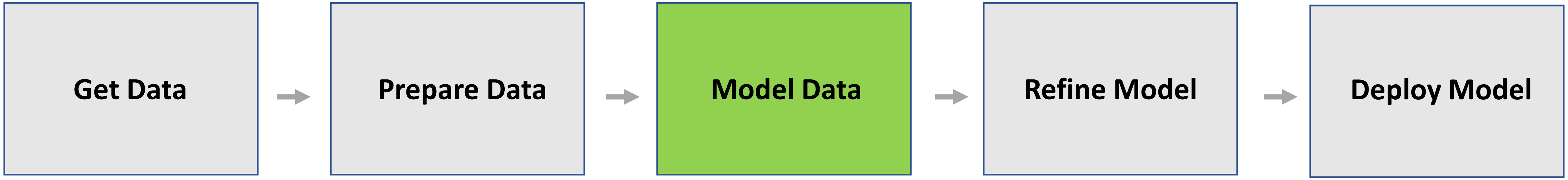
ML Process



Get Data → Prepare Data → Model Data → Refine Model → Deploy Model

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

ML Process



Get Data

Prepare Data

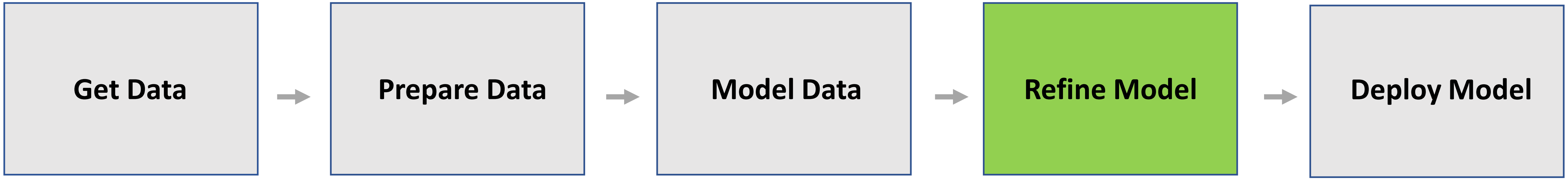
Model Data

Refine Model

Deploy Model

Which algorithms should you use?

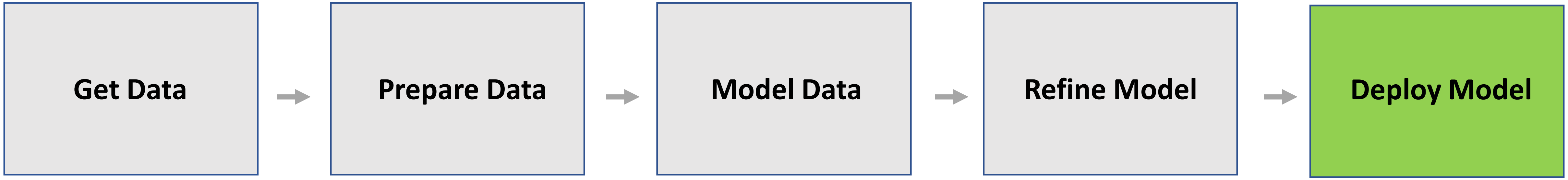
ML Process



Get Data → Prepare Data → Model Data → Refine Model → Deploy Model

What level of performance is sufficient?

ML Process



Get Data → Prepare Data → Model Data → Refine Model → Deploy Model

Make predictions.