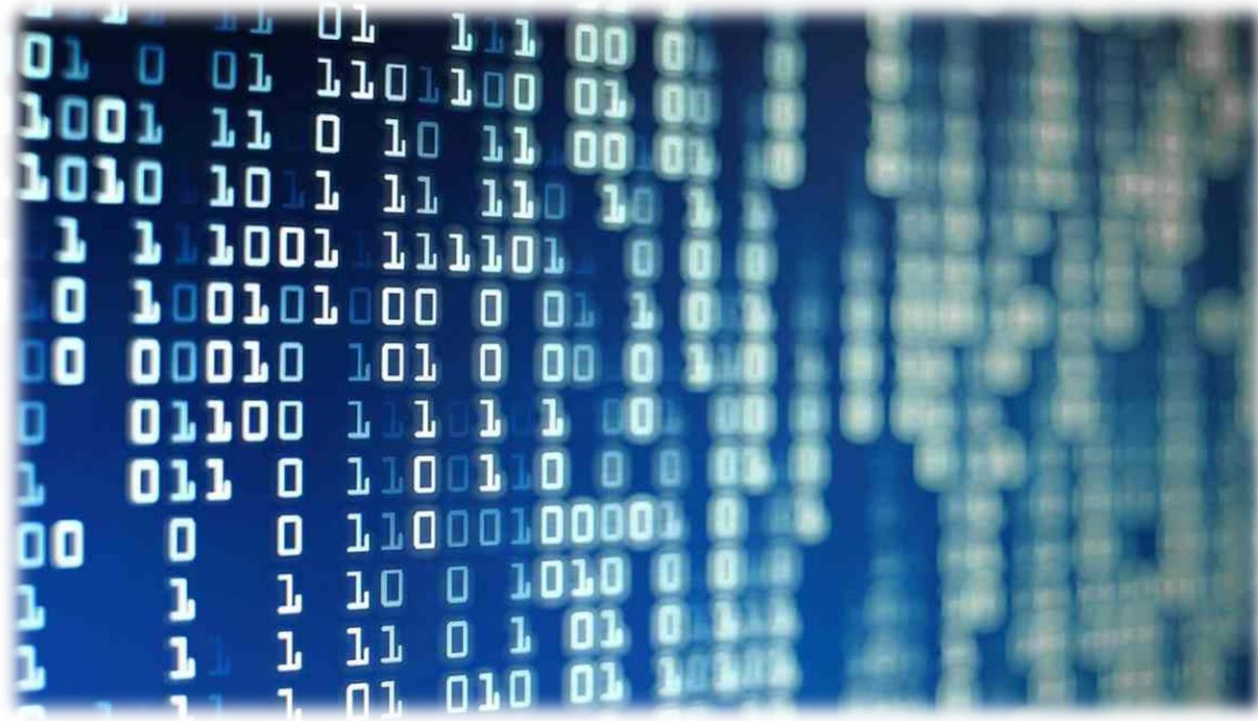
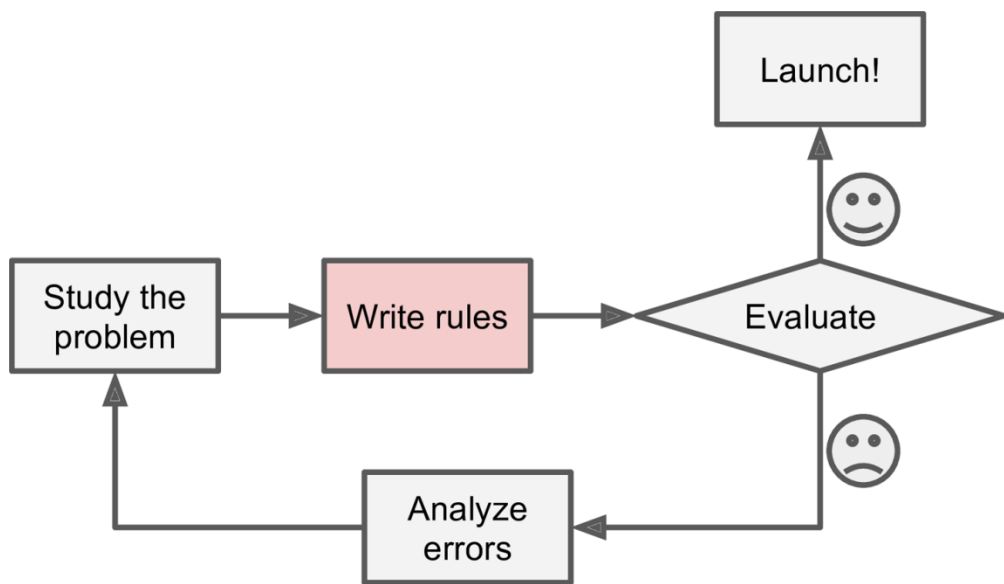


# 딥러닝 데이터 준비

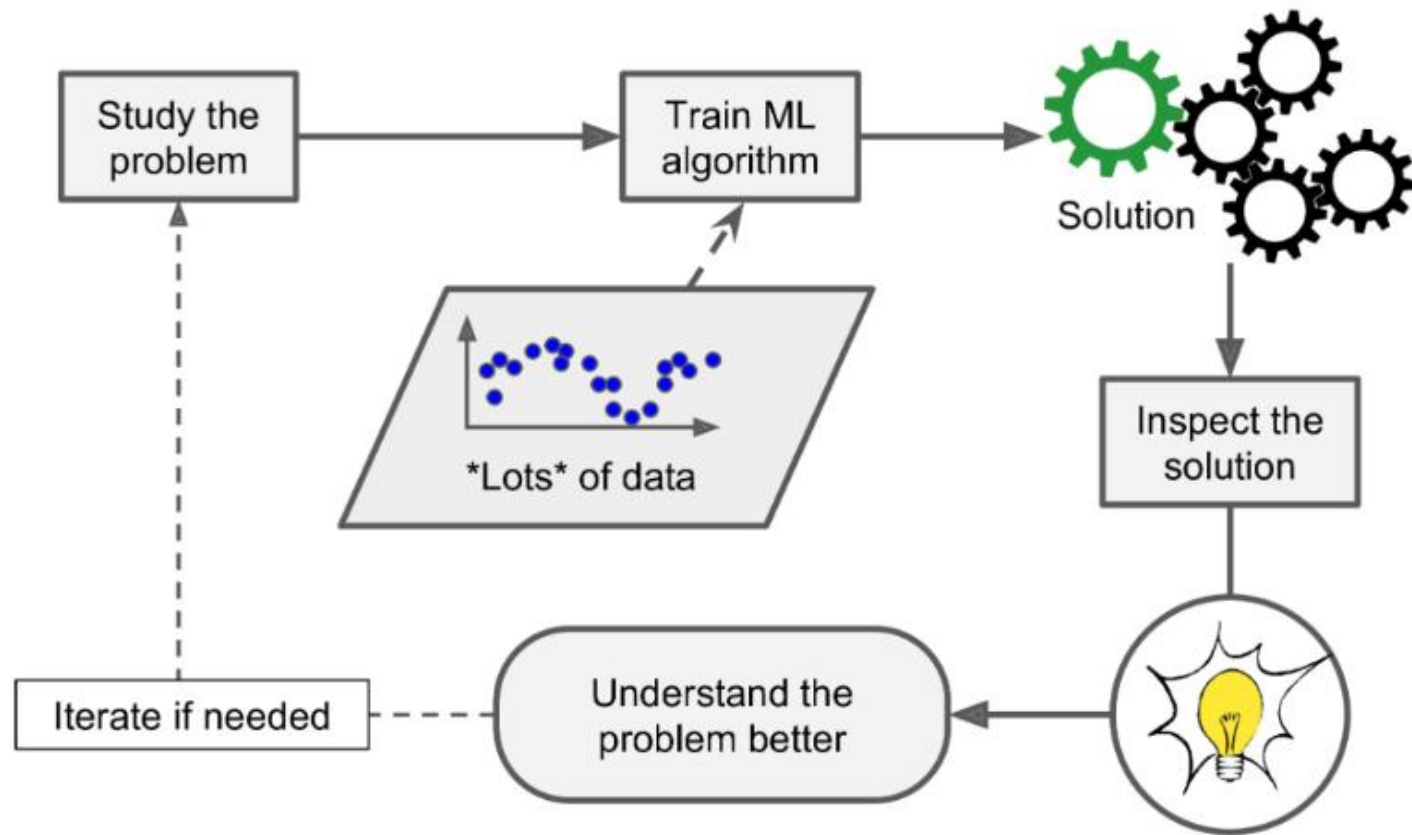


# 데이터

## ■ 기존 방식

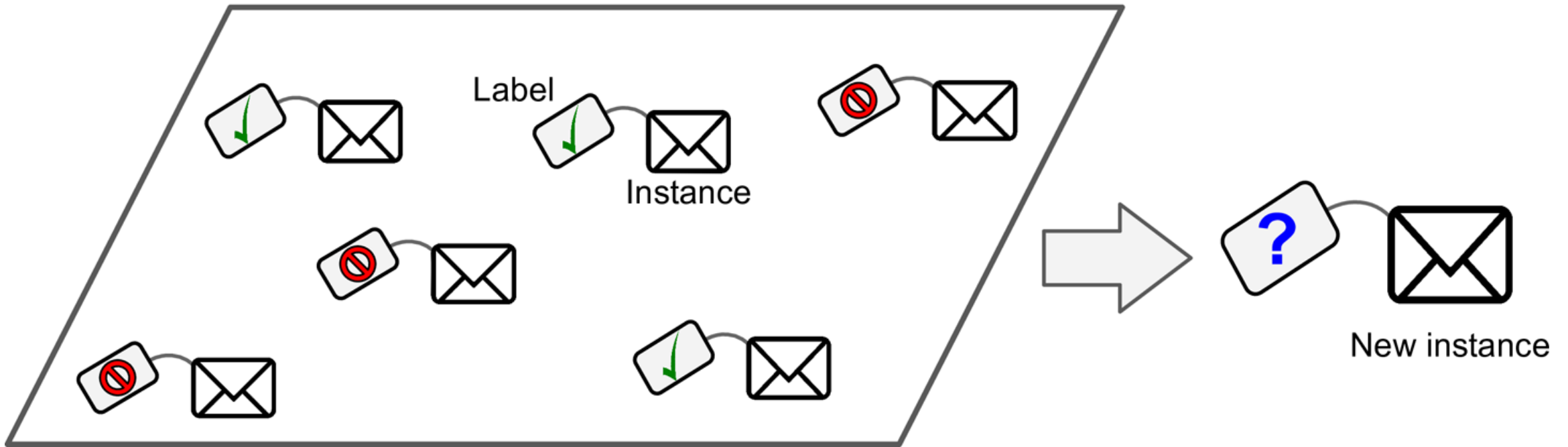


## ■ 머신러닝 방식

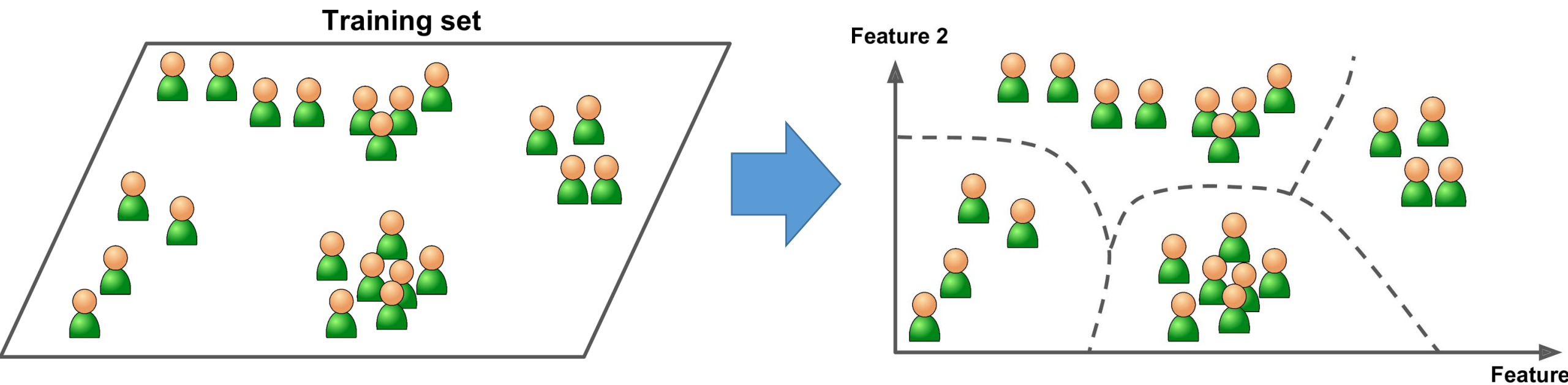


# 지도학습(Supervised Learning)

Training set



# 비지도 학습(Unsupervised Learning)



# 데이터

**kaggle Datasets**

<https://www.kaggle.com/datasets>



<http://archive.ics.uci.edu/ml/index.php>

**Registry of Open Data on AWS**

<https://registry.opendata.aws/>



[https://en.wikipedia.org/wiki/List\\_of\\_datasets\\_for\\_machine-learning\\_research](https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research)

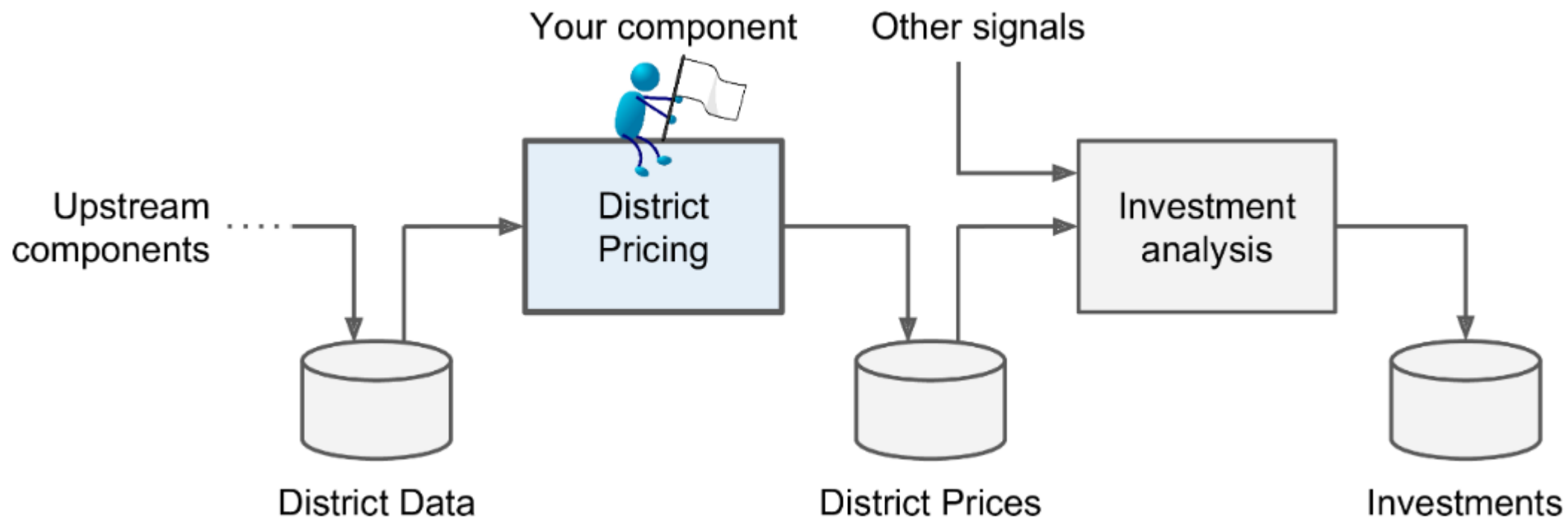


<https://www.data.go.kr/>



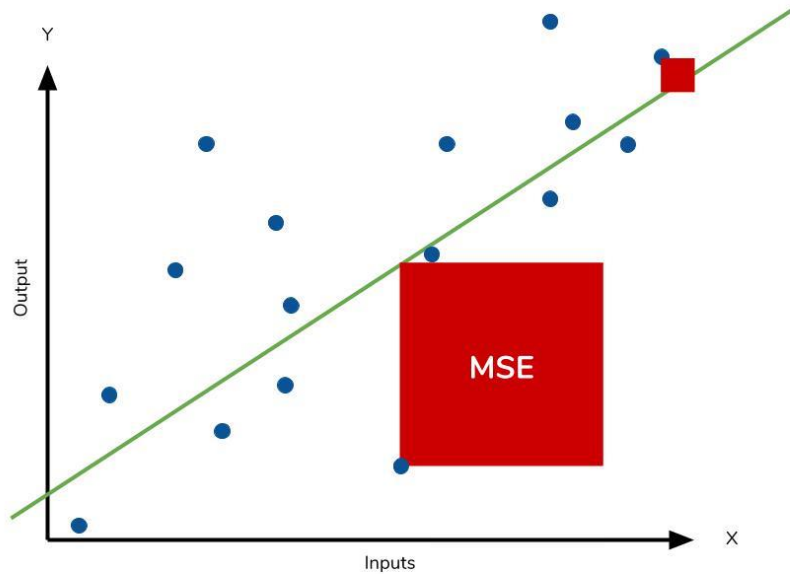
<https://aihub.or.kr/>

# 문제 정의



# 성능지표 선택 (회귀모델)

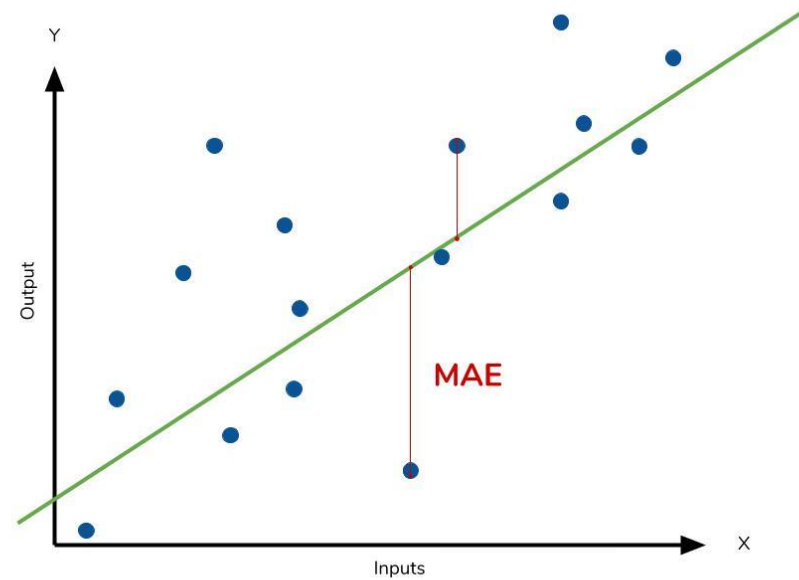
## ■ MSE(Mean Squared Error)



$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

레이블 값 (실제값)       $\hat{y}$  (모델이 예측한 값)

## ■ MAE(Mean Absolute Error)



$$\frac{1}{n} \sum_{i=1}^n (|y_i - \hat{y}_i|)$$

참고자료 : <https://mizykk.tistory.com/102>

# 데이터 다운로드

```
import os
import tarfile
from six.moves import urllib
```

```
DOWNLOAD_ROOT = "https://raw.githubusercontent.com/ageron/handson-ml/master/"
HOUSING_PATH = os.path.join("datasets", "housing")
HOUSING_URL = DOWNLOAD_ROOT + "datasets/housing/housing.tgz"
```

```
def fetch_housing_data(housing_url=HOUSING_URL, housing_path=HOUSING_PATH):
    if not os.path.isdir(housing_path):
        os.makedirs(housing_path)
    tgz_path = os.path.join(housing_path, "housing.tgz")
    urllib.request.urlretrieve(housing_url, tgz_path)
    housing_tgz = tarfile.open(tgz_path)
    housing_tgz.extractall(path=housing_path)
    housing_tgz.close()
```



# 데이터 다운로드

```
fetch_housing_data()
```

```
import pandas as pd
```

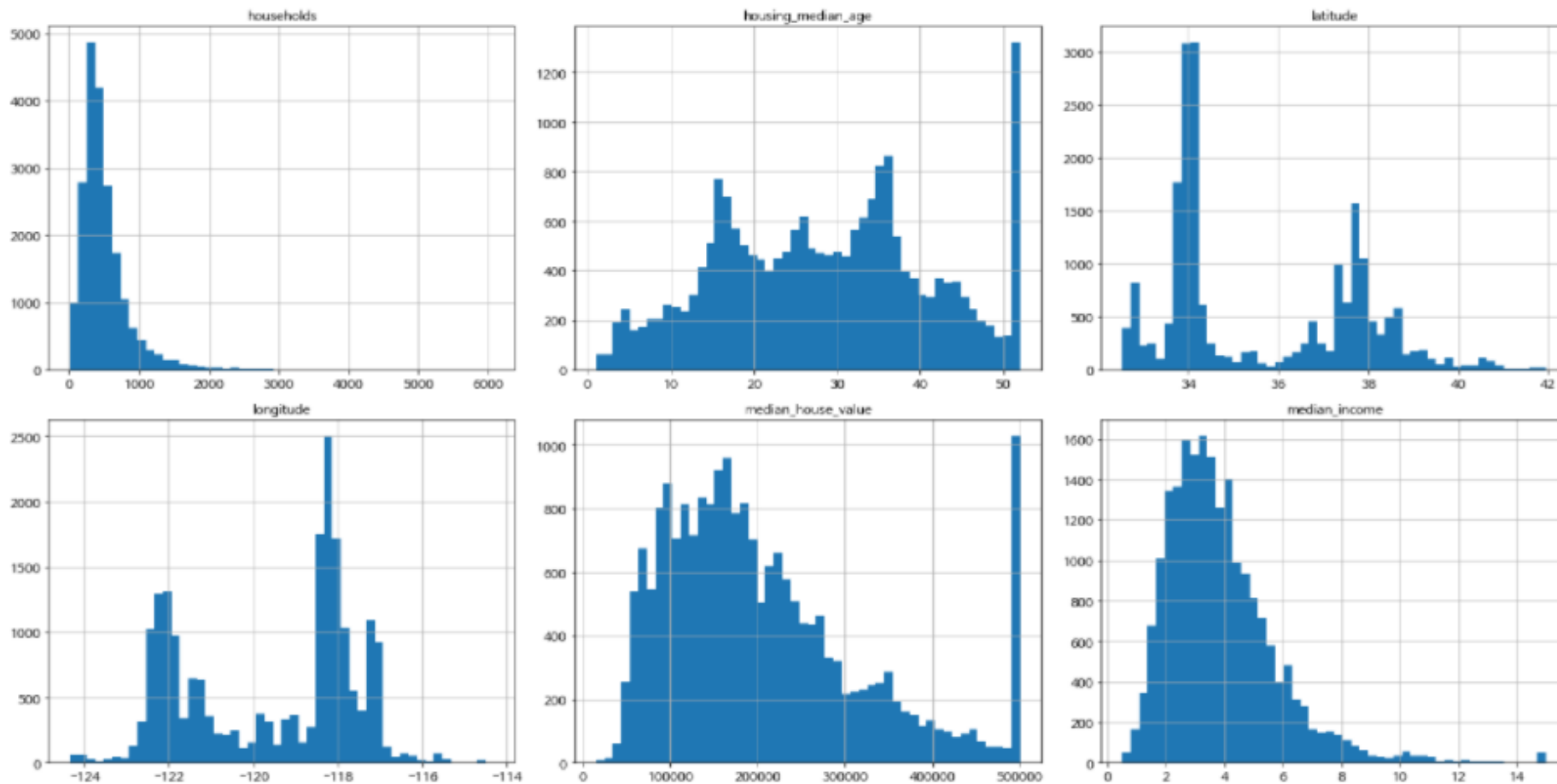
```
def load_housing_data(housing_path=HOUSING_PATH):  
    csv_path = os.path.join(housing_path, "housing.csv")  
    return pd.read_csv(csv_path)
```

```
housing = load_housing_data()  
housing.head()
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY

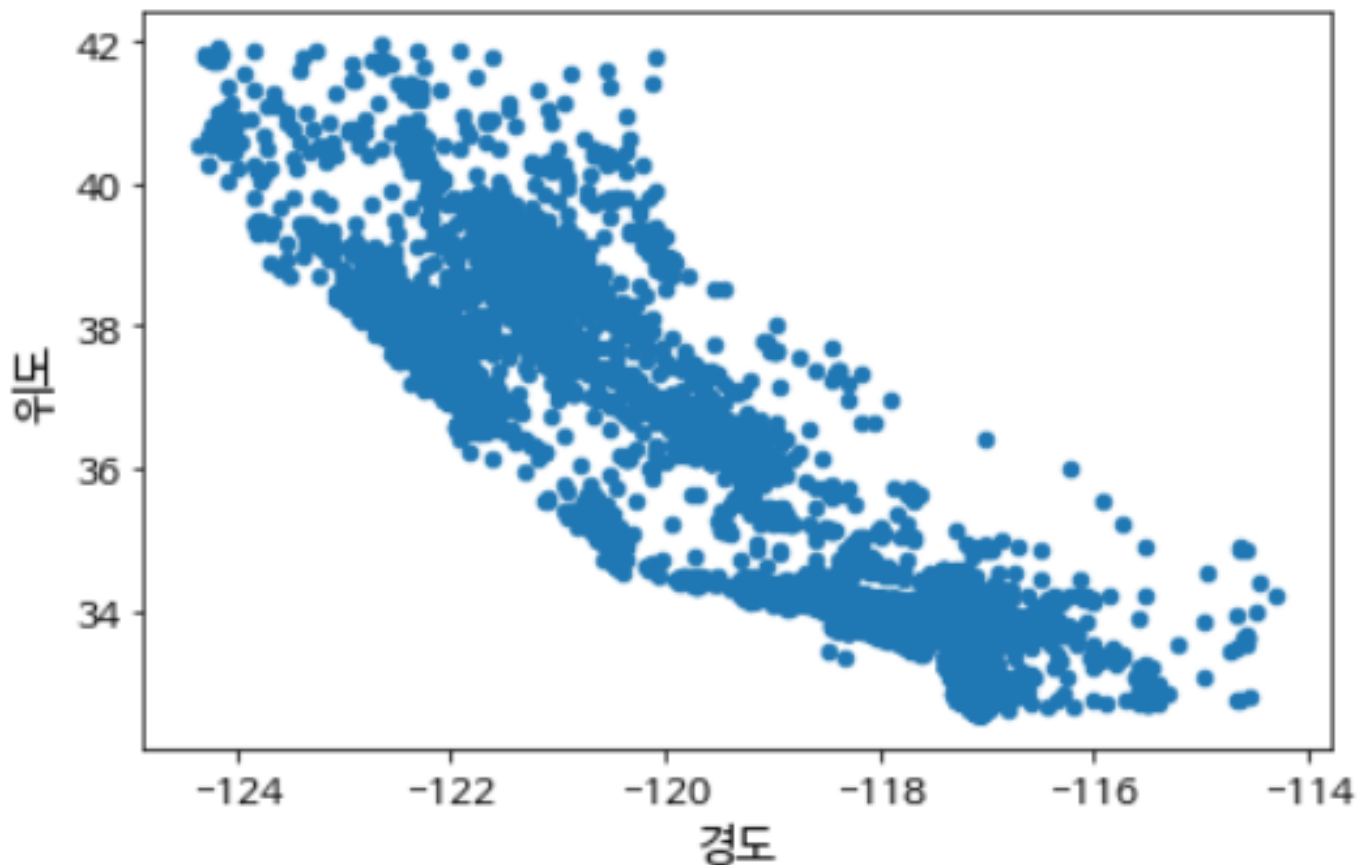
# 데이터 탐색과 시각화

```
import matplotlib.pyplot as plt
housing.hist(bins=50, figsize=(20,15))
save_fig("attribute_histogram_plots")
plt.show()
```



# 데이터 이해를 위한 탐색과 시각화

```
ax = housing.plot(kind="scatter", x="longitude", y="latitude")  
ax.set(xlabel='경도', ylabel='위도')  
save_fig("bad_visualization_plot")
```



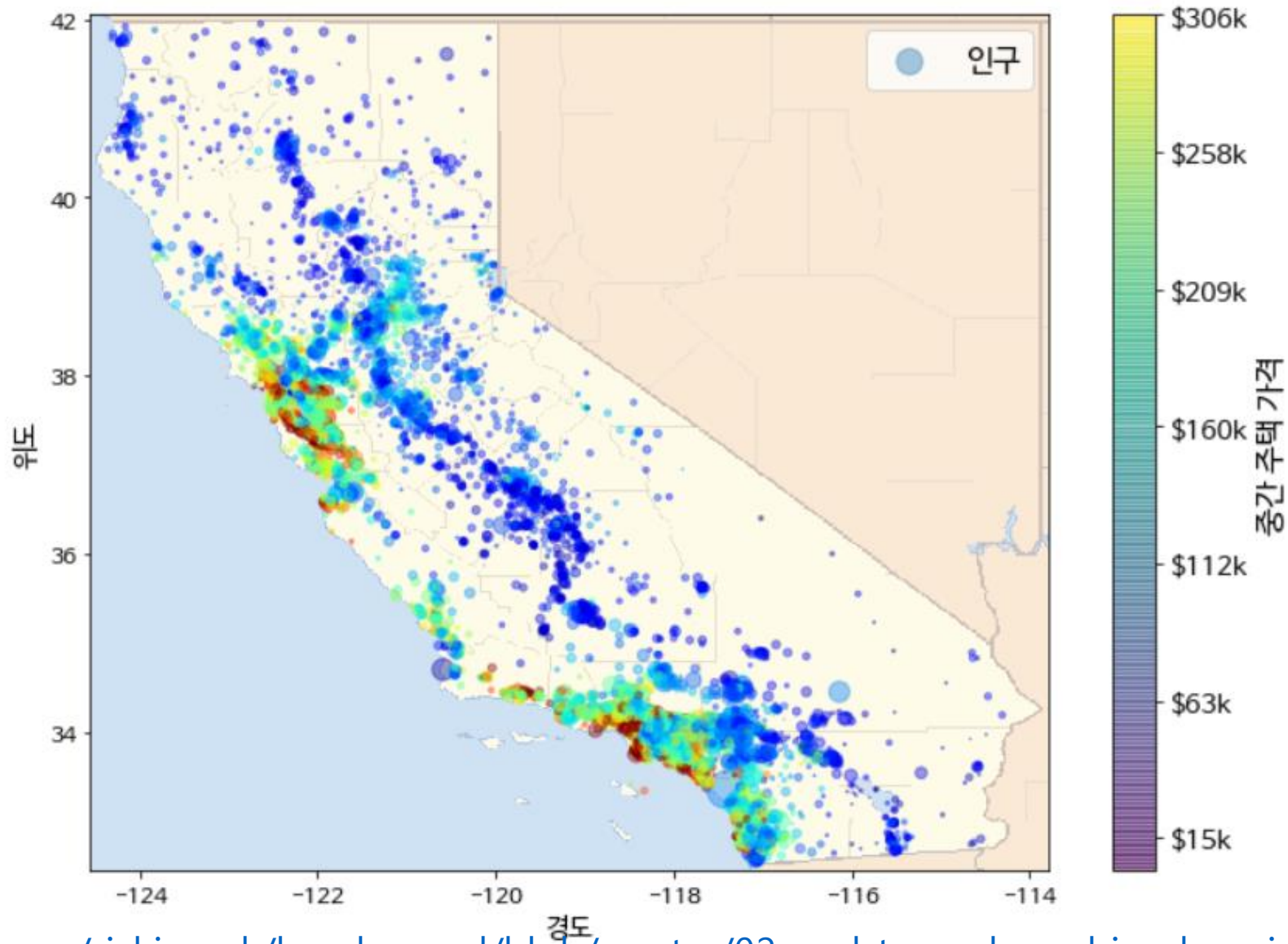
# 데이터 이해를 위한 탐색과 시각화

```
import matplotlib.image as mpimg
california_img=mpimg.imread(PROJECT_ROOT_DIR + '/images/end_to_end_project/california.png')
ax = housing.plot(kind="scatter", x="longitude", y="latitude", figsize=(10,7),
                  s=housing['population']/100, label="인구",
                  c="median_house_value", cmap=plt.get_cmap("jet"),
                  colorbar=False, alpha=0.4,
                  )
plt.imshow(california_img, extent=[-124.55, -113.80, 32.45, 42.05], alpha=0.5)
plt.ylabel("위도", fontsize=14)
plt.xlabel("경도", fontsize=14)

prices = housing["median_house_value"]
tick_values = np.linspace(prices.min(), prices.max(), 11)
cbar = plt.colorbar()
cbar.ax.set_yticklabels(["$%dk"%(round(v/1000)) for v in tick_values], fontsize=14)
cbar.set_label('중간 주택 가격', fontsize=16)

plt.legend(fontsize=16)
save_fig("california_housing_prices_plot")
plt.show()
```

# 데이터 이해를 위한 탐색과 시각화



[https://github.com/rickiepark/handson-ml/blob/master/02\\_end\\_to\\_end\\_machine\\_learning\\_project.ipynb](https://github.com/rickiepark/handson-ml/blob/master/02_end_to_end_machine_learning_project.ipynb)

# 딥러닝 데이터 준비 실습

[https://github.com/rickiepark/handson-ml/blob/master/02\\_end\\_to\\_end\\_machine\\_learning\\_project.ipynb](https://github.com/rickiepark/handson-ml/blob/master/02_end_to_end_machine_learning_project.ipynb)

<https://github.com/>



<https://colab.research.google.com/github/>

[https://colab.research.google.com/github/rickiepark/handson-ml/blob/master/02\\_end\\_to\\_end\\_machine\\_learning\\_project.ipynb](https://colab.research.google.com/github/rickiepark/handson-ml/blob/master/02_end_to_end_machine_learning_project.ipynb)



<https://youtu.be/KMa5z3amwv4>

<https://youtu.be/NK-poleR9JY>

<https://youtu.be/8-milNfxCm4>