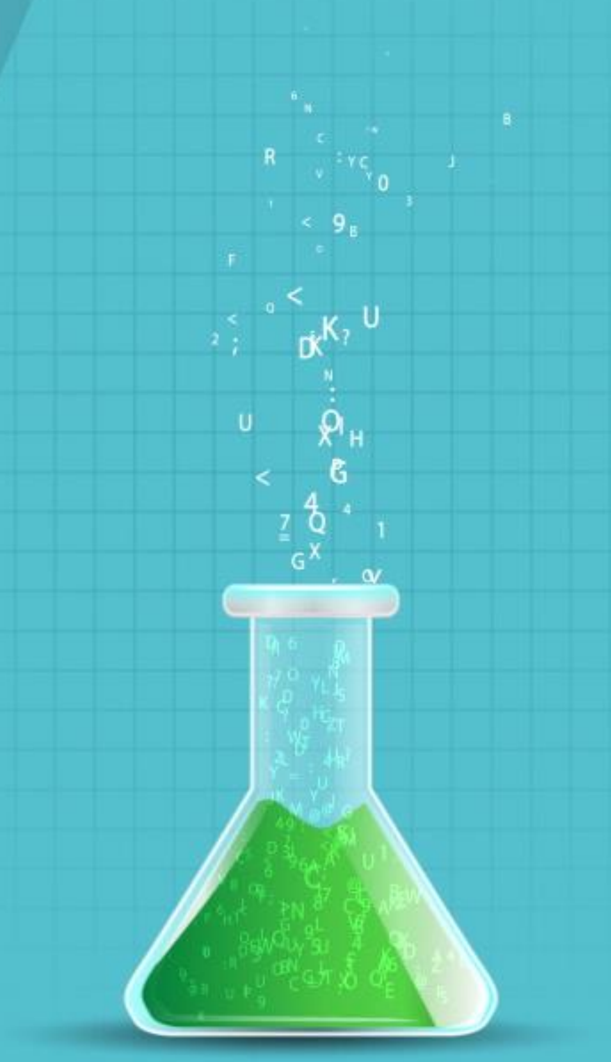
Data Science with Python

Lesson 8—Machine Learning with Scikit–Learn

DATA SCIENCE

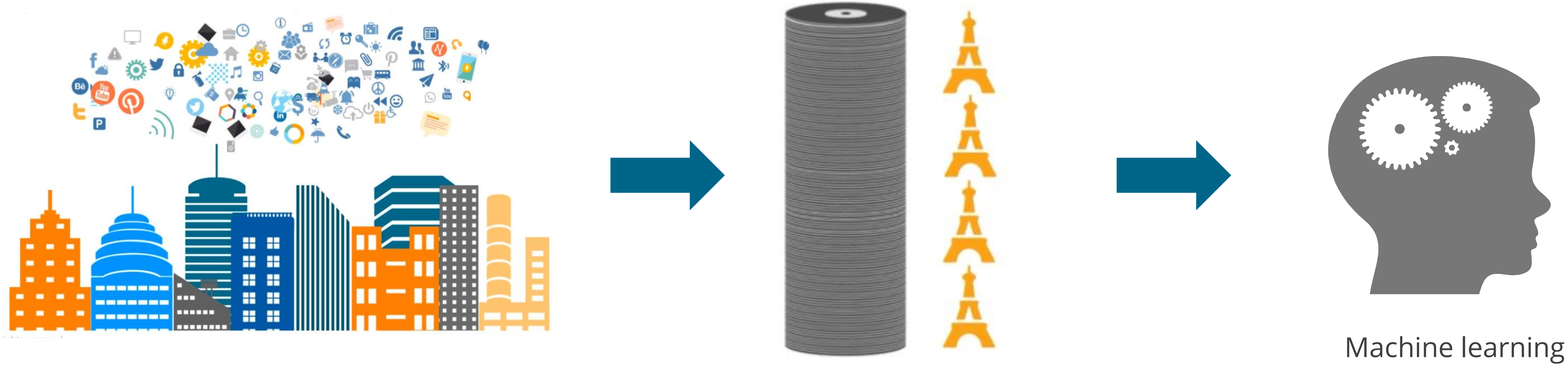
What You’ll Learn

- What machine learning is and why it is important
- The machine learning approach
- Relevant terminologies that help you understand a dataset
- Features of supervised and unsupervised learning models
- Algorithms such as regression, classification, clustering, and dimensionality reduction



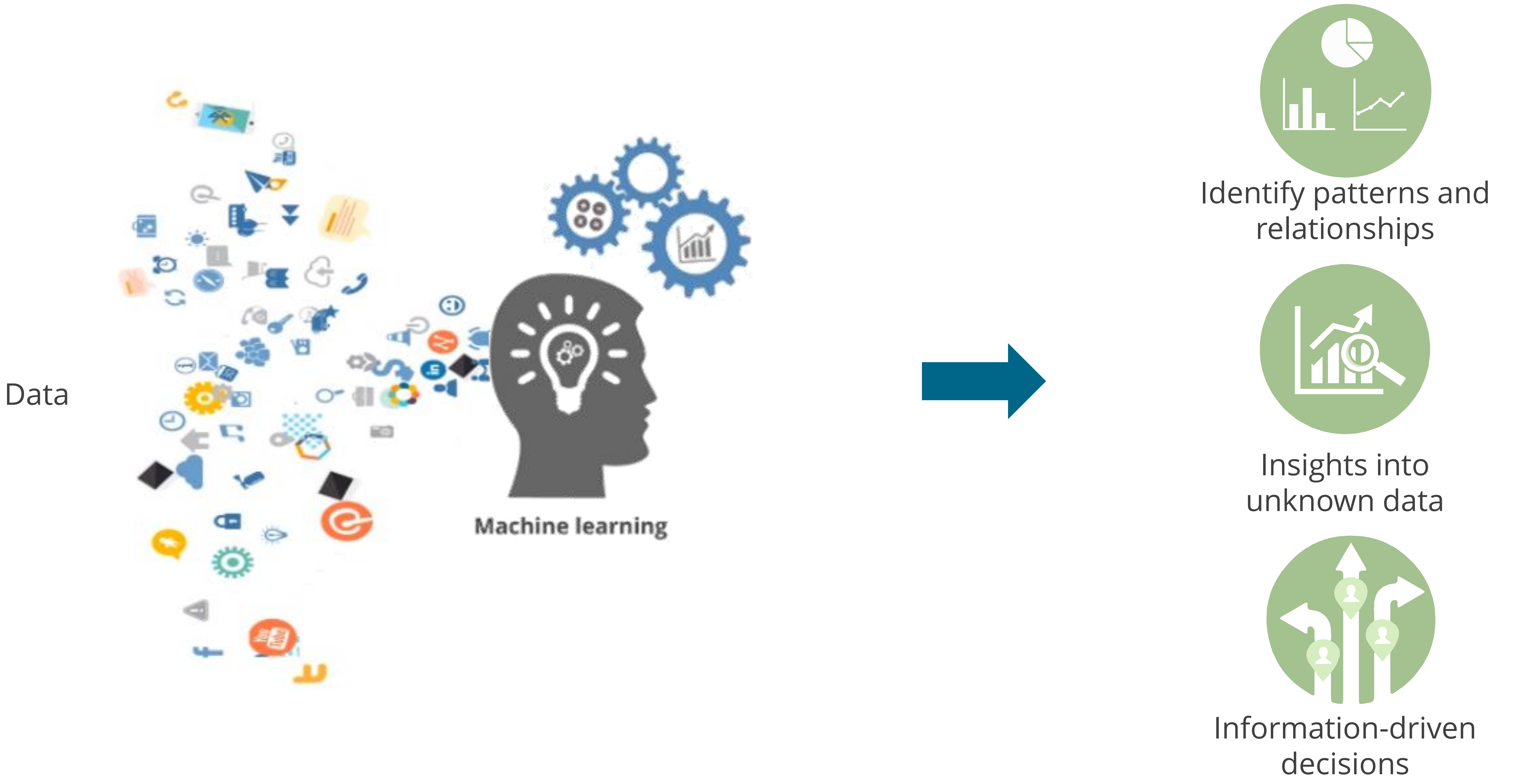
Why Machine Learning

If we stored the data generated in a day on Blu-ray disks and stacked them up, it would be equal to the height of four Eiffel towers! Machine learning helps analyze this data easily and quickly.



Purpose of Machine Learning

Machine learning is a great tool to analyze data, find hidden data patterns and relationships, and extract information to enable information-driven decisions and provide insights.



Data

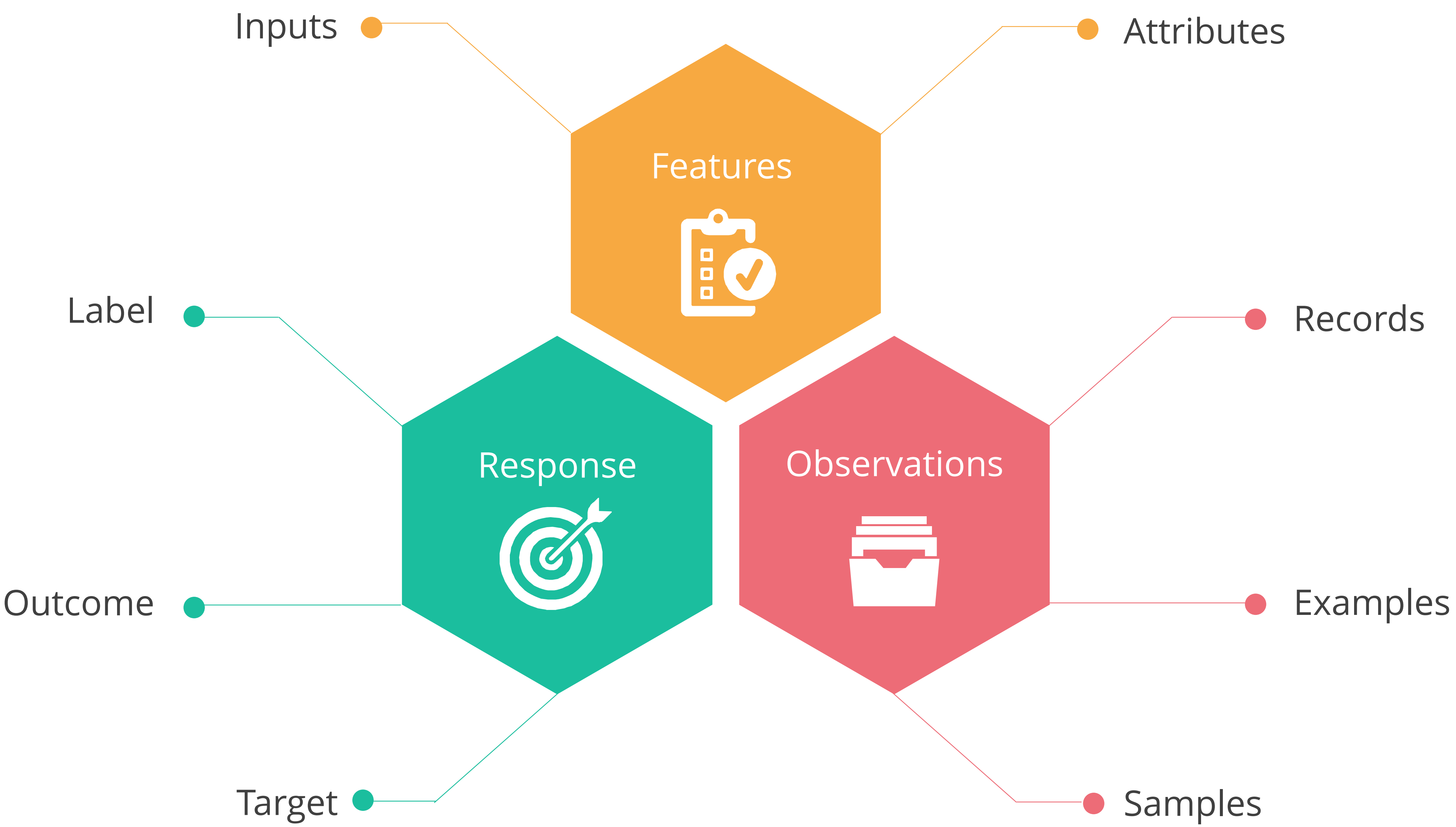
Identify patterns and relationships

Insights into unknown data

Information-driven decisions

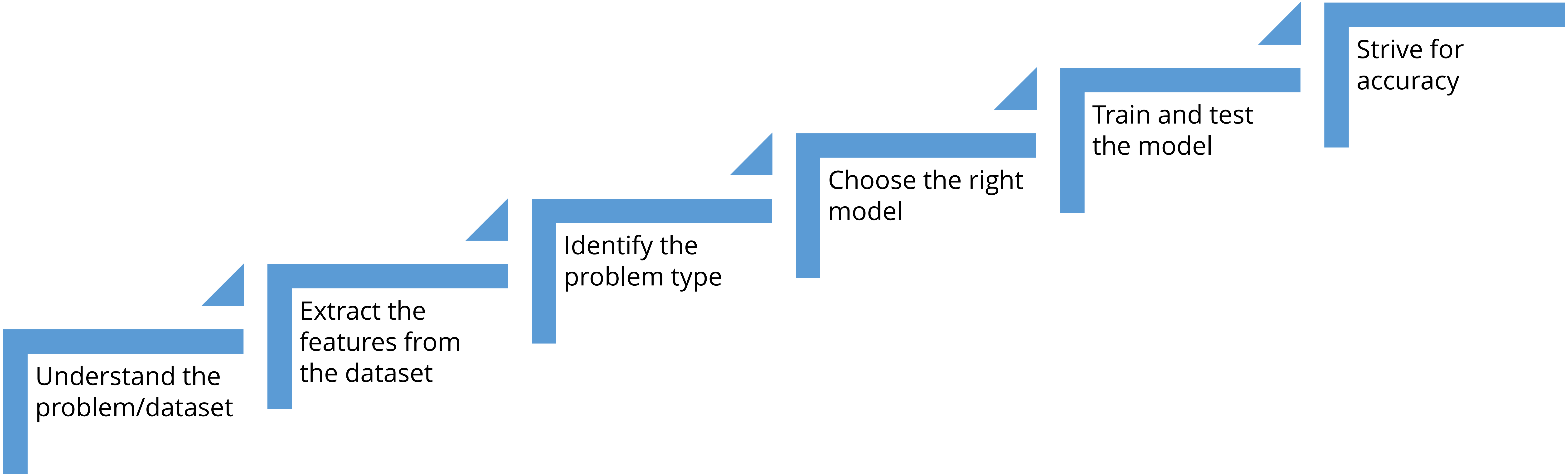
Machine Learning Terminology

These are some machine learning terminologies that you will come across in this lesson:



Machine Learning Approach

The machine learning approach starts with either a problem that you need to solve or a given dataset that you need to analyze.



Understand the problem/dataset

Extract the features from the dataset

Identify the problem type

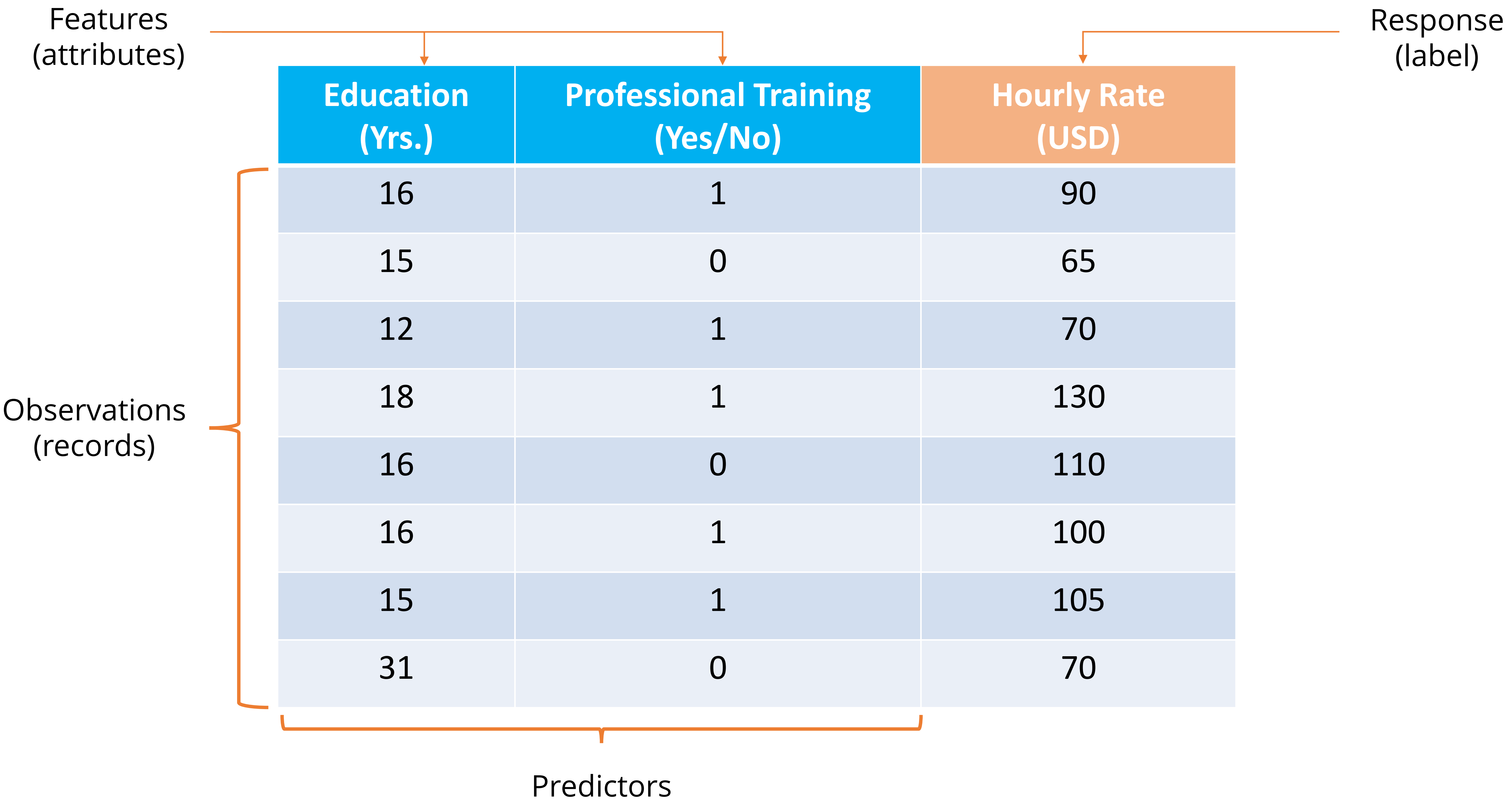
Choose the right model

Train and test the model

Strive for accuracy

Steps 1 and 2: Understand the Dataset and Extract its Features

Let us look at a dataset and understand its features in terms of machine learning.



Features (attributes)

Response (label)

Observations (records)

Predictors

Education (Yrs.)	Professional Training (Yes/No)	Hourly Rate (USD)
16	1	90
15	0	65
12	1	70
18	1	130
16	0	110
16	1	100
15	1	105
31	0	70

Steps 3 and 4: Identify the Problem Type and Learning Model

Machine learning can either be supervised or unsupervised. The problem type should be selected based on the type of learning model.

Concept | Problem Types | Example

Supervised Learning

- In supervised learning, the dataset used to train a model should have observations, features, and responses. The model is trained to predict the “right” response for a given set of data points.
- Supervised learning models are used to predict an outcome.
- The goal of this model is to “generalize” a dataset so that the “general rule” can be applied to new data as well.

Unsupervised Learning

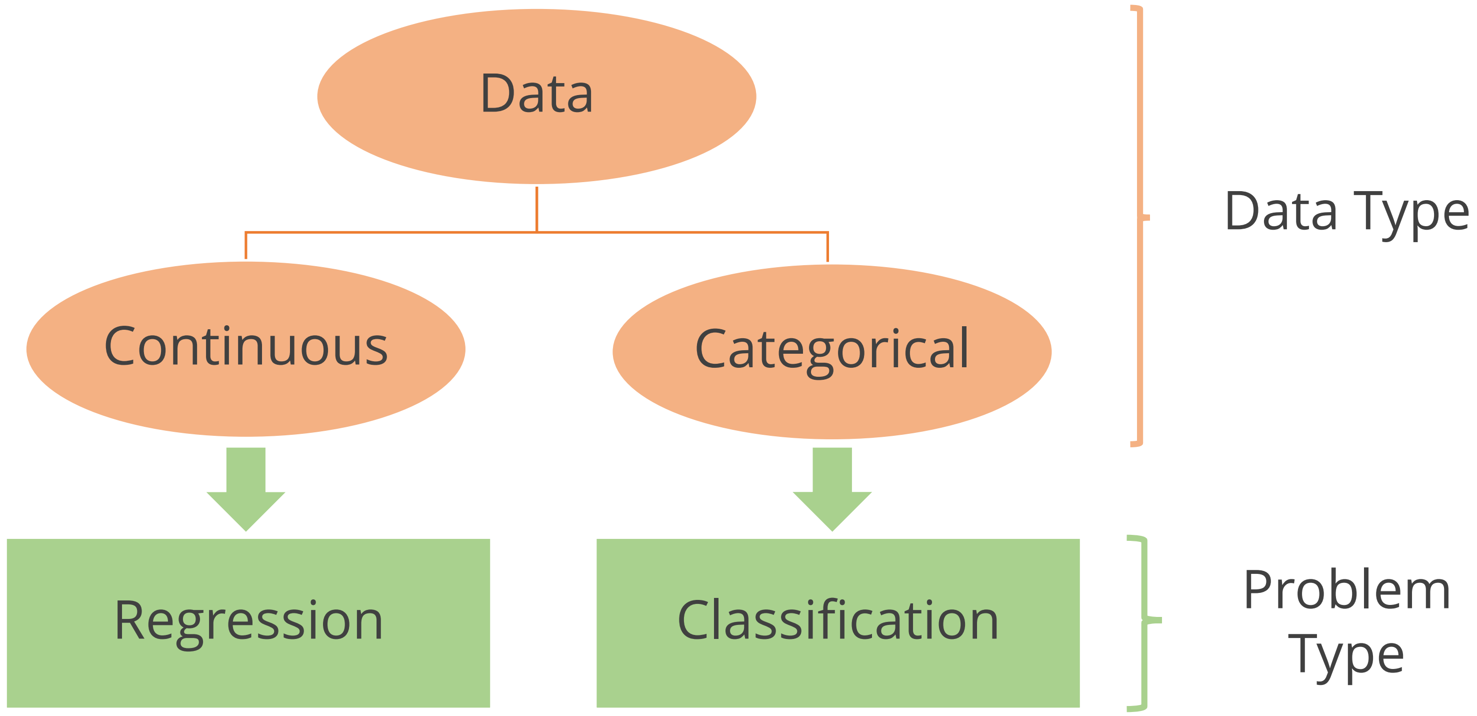
- In unsupervised learning, the response or the outcome of the data is not known.
- Supervised learning models are used to identify and visualize patterns in data by grouping similar types of data.
- The goal of this model is to “represent” data in a way that meaningful information can be extracted.

Steps 3 and 4: Identify the Problem Type and Learning Model (contd.)

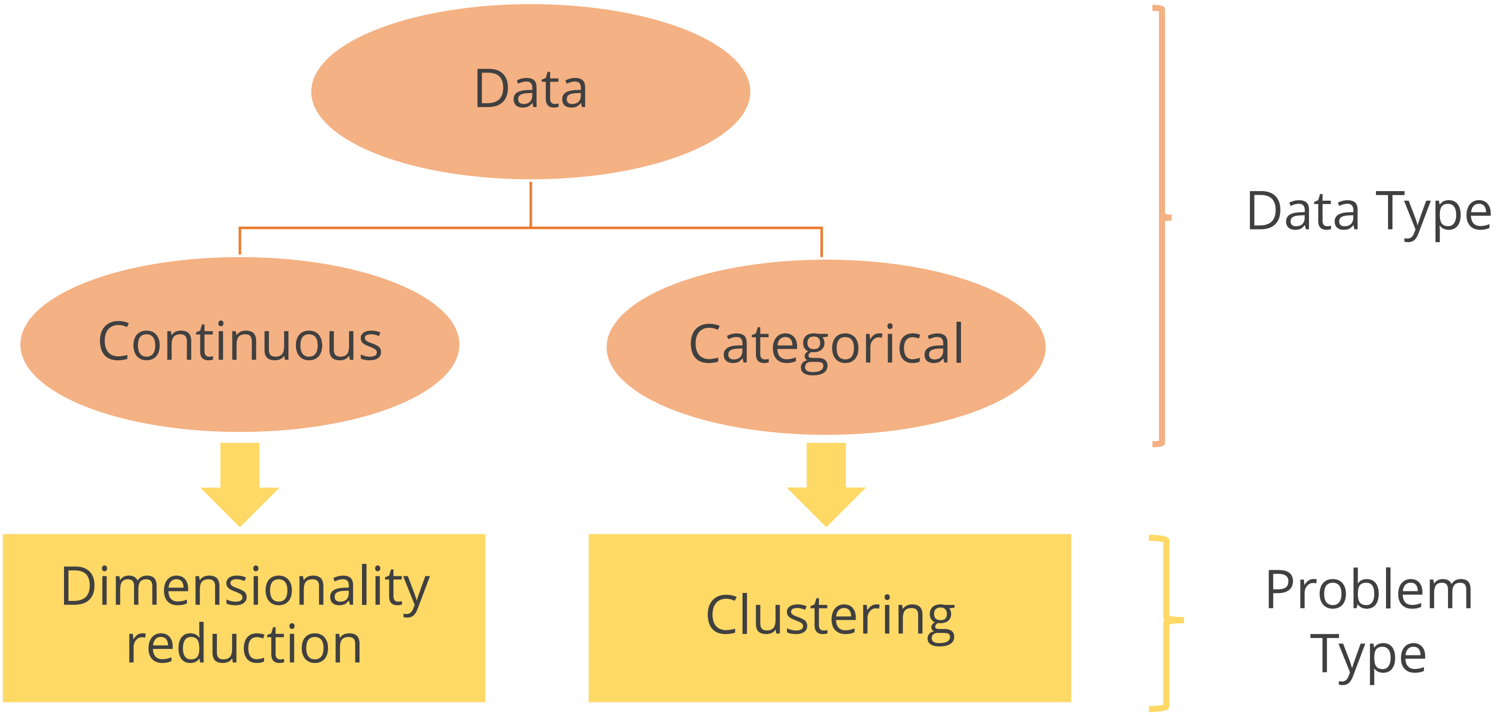
Data can either be continuous or categorical. Based on whether it is supervised or unsupervised learning, the problem type will differ.

Concept | Problem Types | Example

Supervised Learning



Unsupervised Learning

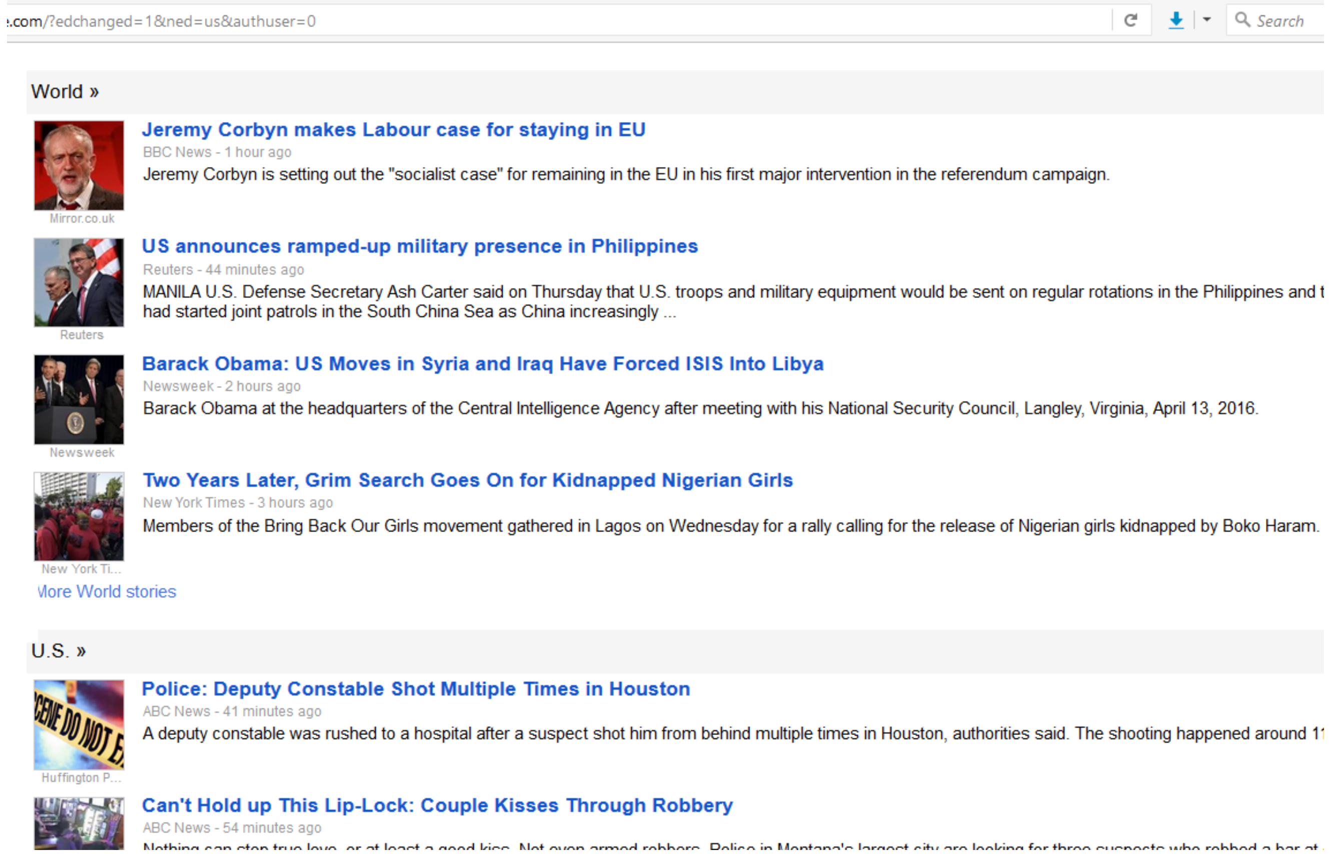


Steps 3 and 4: Identify the Problem Type and Learning Model (contd.)

Some examples of supervised and unsupervised learning models are shown here.

Concept | Problem Types | Example

Supervised Learning



.com/?edchanged=1&ned=us&authuser=0

World »

Jeremy Corbyn makes Labour case for staying in EU
BBC News - 1 hour ago
Jeremy Corbyn is setting out the "socialist case" for remaining in the EU in his first major intervention in the referendum campaign.
Mirror.co.uk

US announces ramped-up military presence in Philippines
Reuters - 44 minutes ago
MANILA U.S. Defense Secretary Ash Carter said on Thursday that U.S. troops and military equipment would be sent on regular rotations in the Philippines and t had started joint patrols in the South China Sea as China increasingly ...
Reuters

Barack Obama: US Moves in Syria and Iraq Have Forced ISIS Into Libya
Newsweek - 2 hours ago
Barack Obama at the headquarters of the Central Intelligence Agency after meeting with his National Security Council, Langley, Virginia, April 13, 2016.
Newsweek

Two Years Later, Grim Search Goes On for Kidnapped Nigerian Girls
New York Times - 3 hours ago
Members of the Bring Back Our Girls movement gathered in Lagos on Wednesday for a rally calling for the release of Nigerian girls kidnapped by Boko Haram.
New York Ti...

lore World stories

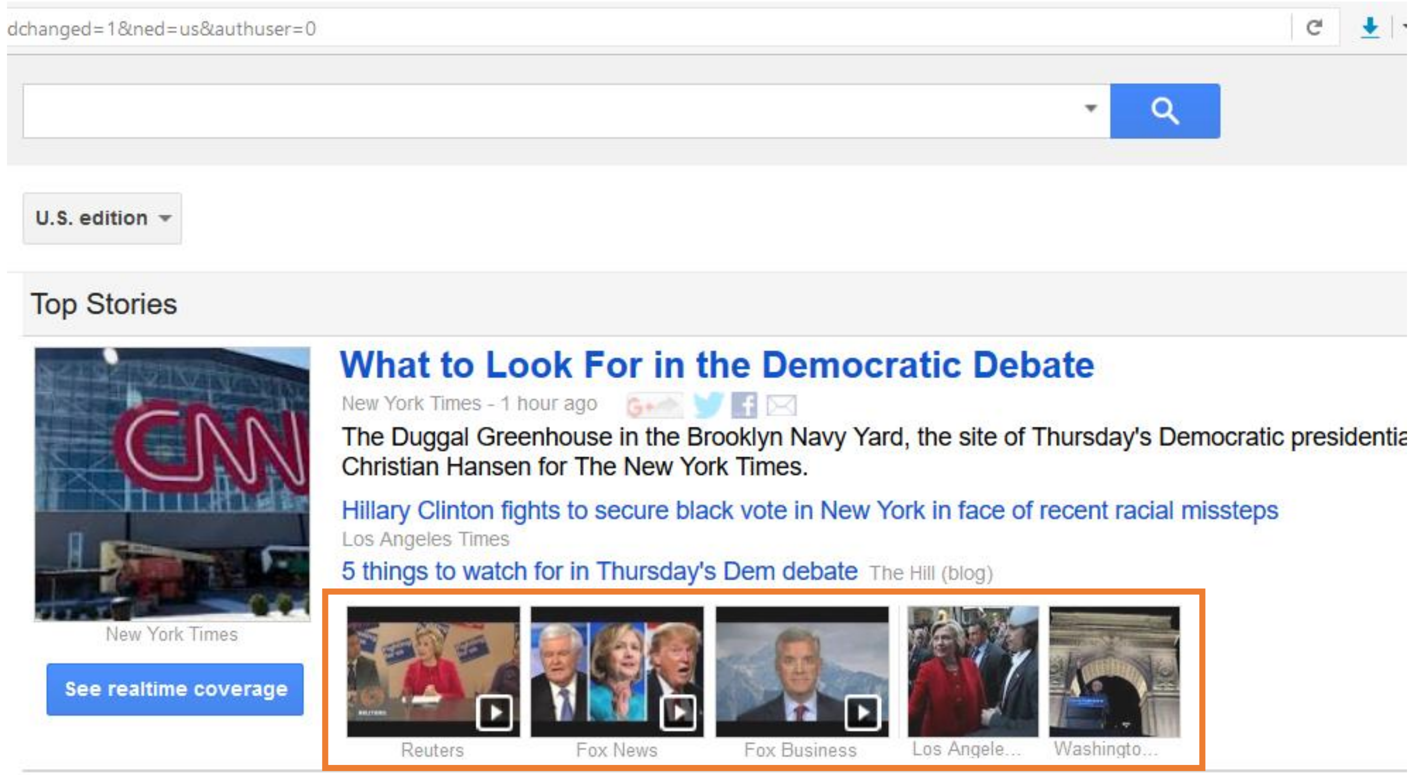
U.S. »

Police: Deputy Constable Shot Multiple Times in Houston
ABC News - 41 minutes ago
A deputy constable was rushed to a hospital after a suspect shot him from behind multiple times in Houston, authorities said. The shooting happened around 1
Huffington P...

Can't Hold up This Lip-Lock: Couple Kisses Through Robbery
ABC News - 54 minutes ago
Nothing can stop true love, or at least a good kiss. Not even armed robbers. Police in Montana's largest city are looking for three suspects who robbed a bar at

Categories of news based on the topics

Unsupervised Learning



dchanged=1&ned=us&authuser=0

U.S. edition

Top Stories

What to Look For in the Democratic Debate
New York Times - 1 hour ago
The Duggal Greenhouse in the Brooklyn Navy Yard, the site of Thursday's Democratic presidentia Christian Hansen for The New York Times.

Hillary Clinton fights to secure black vote in New York in face of recent racial missteps
Los Angeles Times

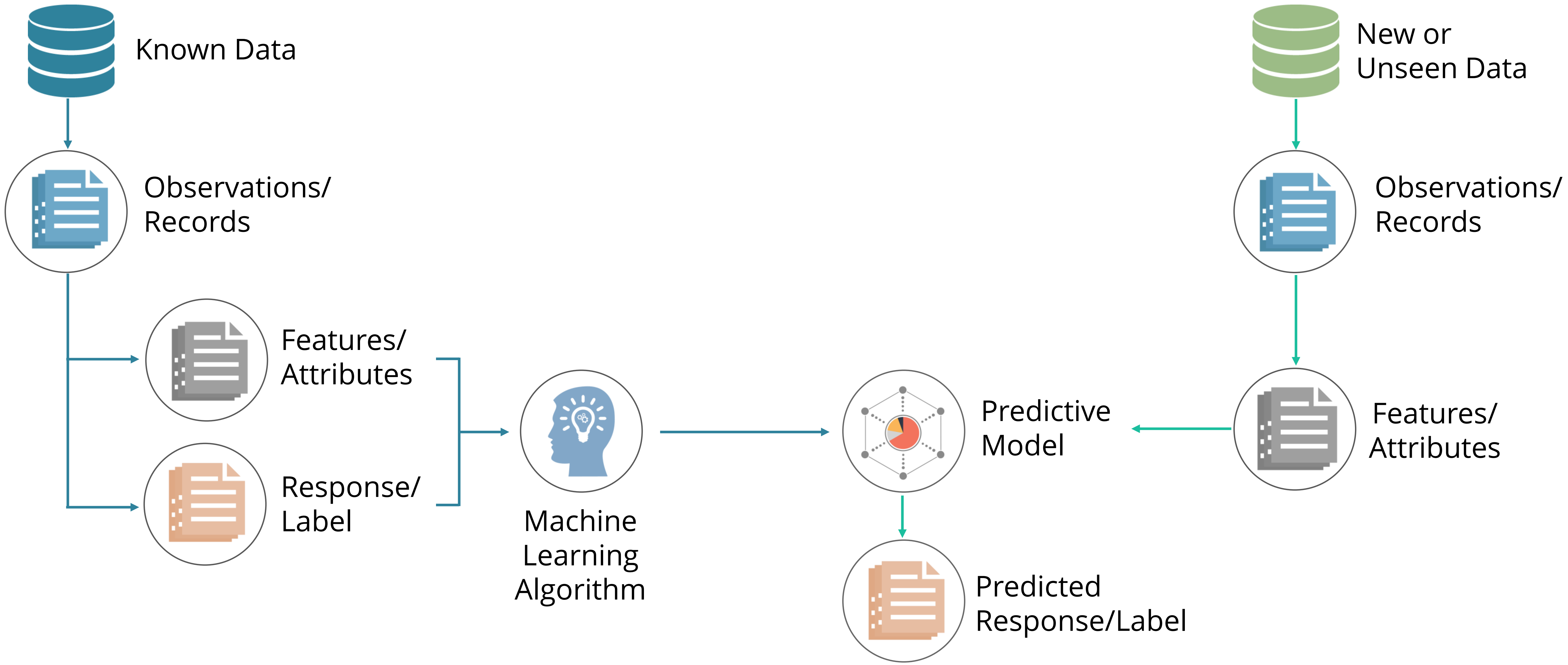
5 things to watch for in Thursday's Dem debate The Hill (blog)

New York Times

Grouping of similar stories on different news networks

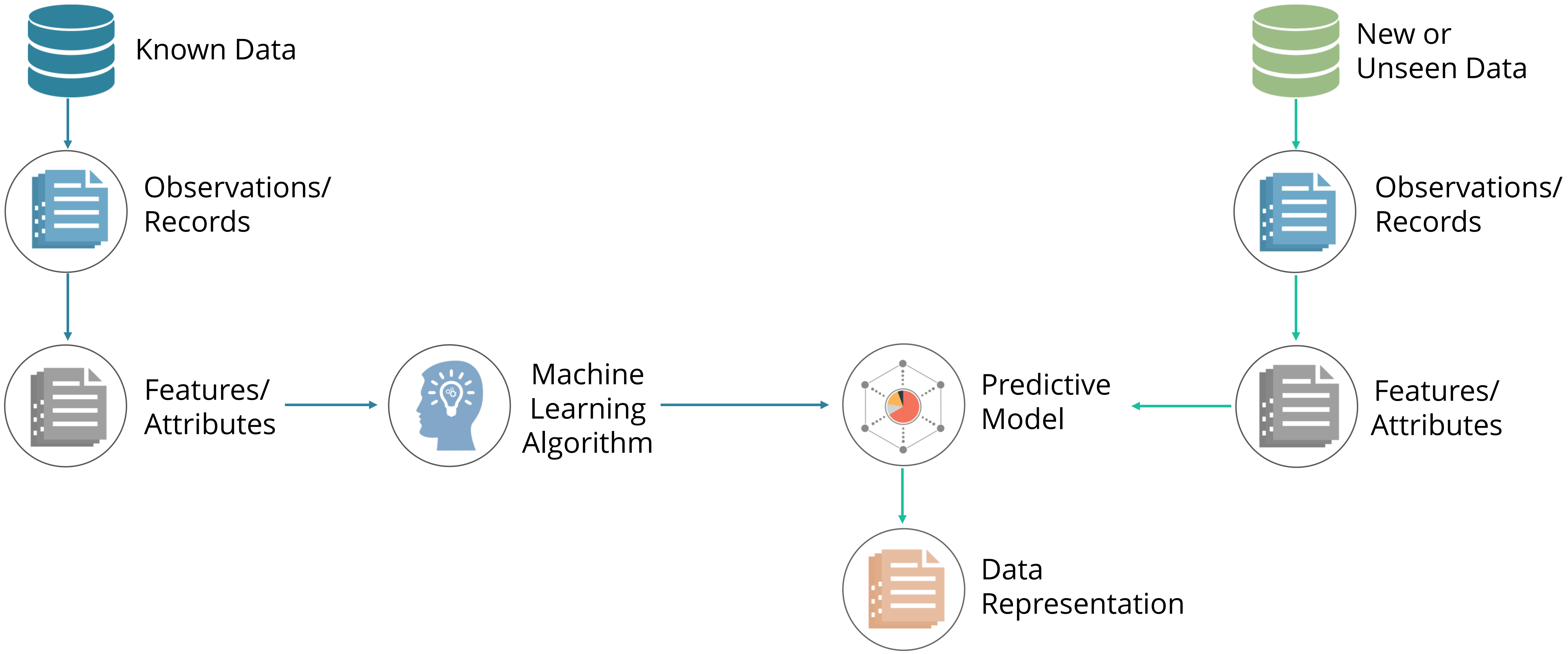
How it Works—Supervised Learning Model

In supervised learning, a known dataset with observations, features, and response is used to create and train a machine learning algorithm. A predictive model, built on top of this algorithm, is then used to predict the response for a new dataset that has the same features.



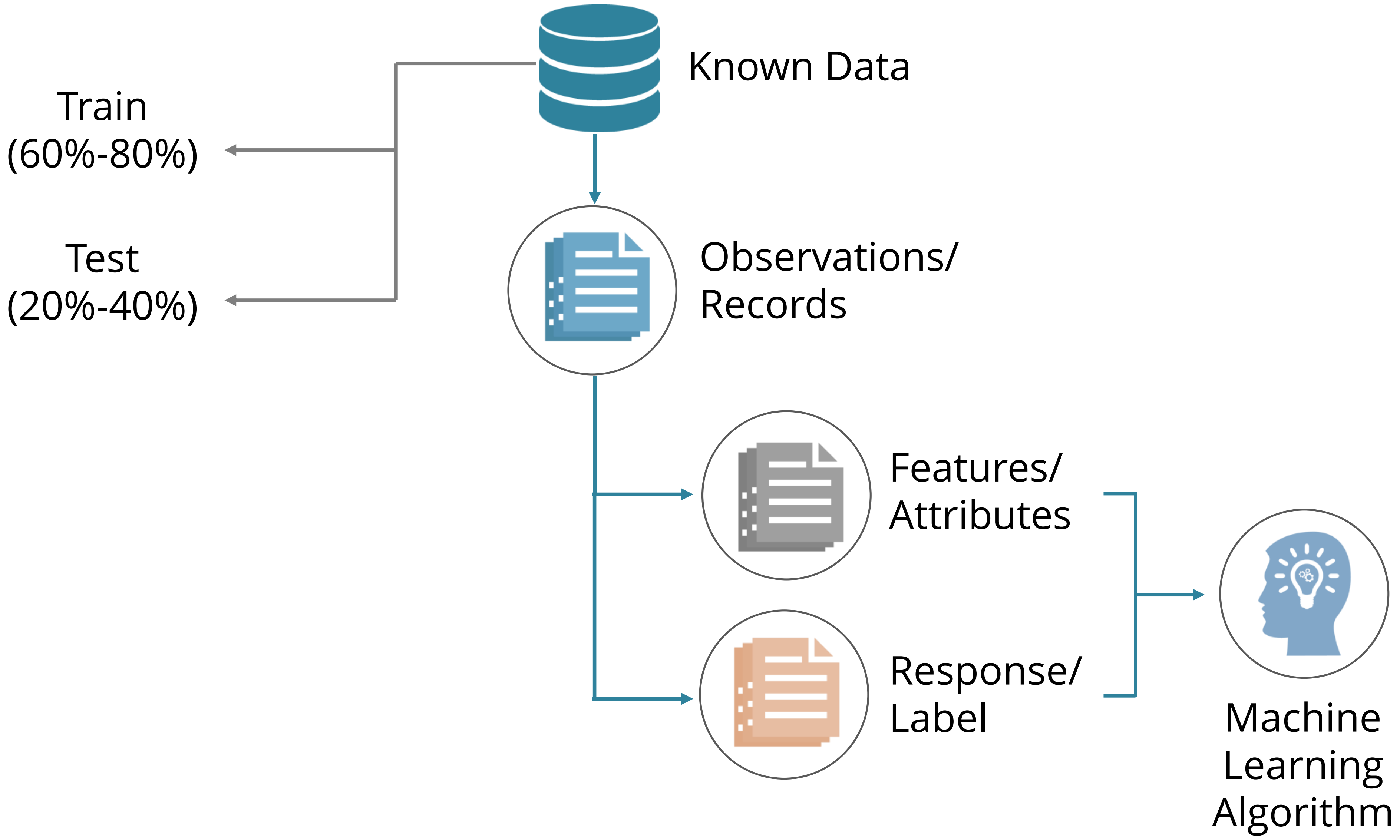
How it Works—Unsupervised Learning Model

In unsupervised learning, a known dataset has a set of observations with features. But the response is not known. The predictive model uses these features to identify how to classify and represent the data points of new or unseen data.



Steps 5 and 6: Train, Test, and Optimize the Model

To train supervised learning models, data analysts usually divide a known dataset into training and testing sets.



Steps 5 and 6: Train, Test, and Optimize the Model (contd.)

Let us look at an example to see how the split approach works.

Model Training

Observation

🟧 Train set

🟨 Test set

ID	Education	Professional training	Hourly rate
10	16	1	90
45	15	0	65
83	12	1	70
45	18	1	130
54	16	0	110
67	16	1	100
71	15	1	105
31	15	0	70

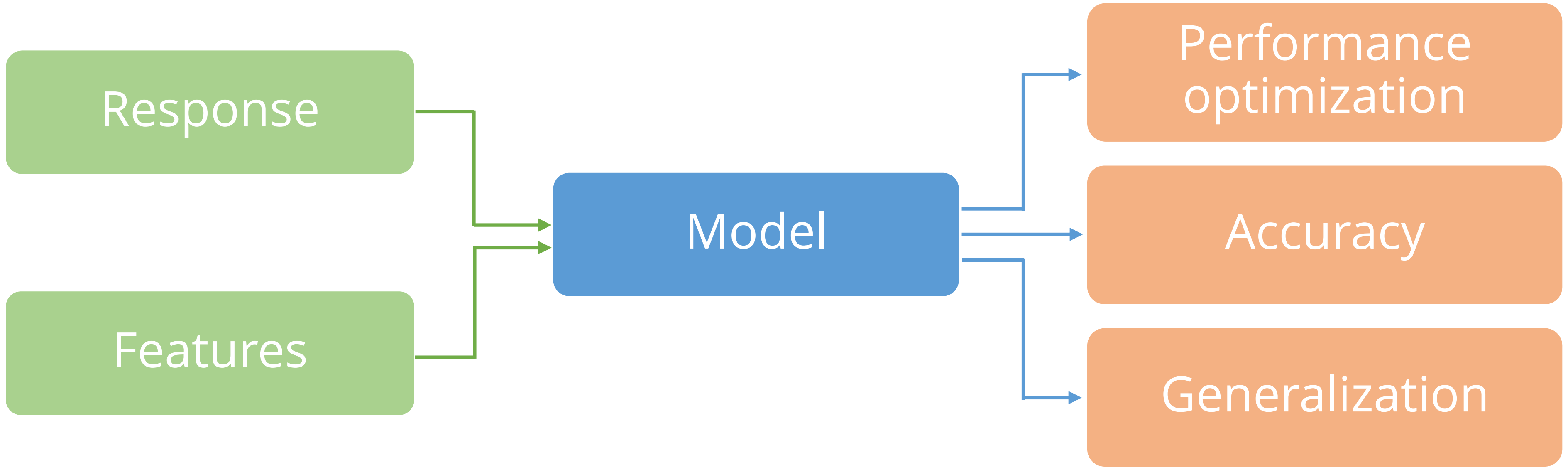
Response

🟦 Train set

🟩 Test set

Supervised Learning Model Considerations

Some considerations of supervised and unsupervised learning models are shown here.



Knowledge Check

KNOWLEDGE CHECK

In machine learning, which one of the following is an observation?

a. Features

b. Attributes

c. Records

d. Labels

KNOWLEDGE CHECK

In machine learning, which one of the following is an observation?

a. Features

b. Attributes

c. Records

d. Labels

The correct answer is c .

Explanation: An observation is a set of examples, records, or samples.

KNOWLEDGE CHECK

If data is continuous and has labels (response), then it fits which of the following problem types?

a. Supervised learning: classification

b. Unsupervised learning: clustering

c. Unsupervised learning: dimensionality reduction

d. Supervised learning: regression

KNOWLEDGE CHECK

If data is continuous and has labels (response), then it fits which of the following problem types?

a. Supervised learning: classification

b. Unsupervised learning: clustering

c. Unsupervised learning: dimensionality reduction

d. Supervised learning: regression

The correct answer is d.

Explanation: The regression algorithm belonging to the supervised learning model is best suited to analyze continuous data.

KNOWLEDGE CHECK

Identify the goal of unsupervised learning. *Select all that apply*.

a. To predict the outcome

b. To understand the structure of the data

c. To generalize the dataset

d. To represent the data

KNOWLEDGE CHECK

Identify the goal of unsupervised learning. *Select all that apply*.

a. To predict the outcome

b. To understand the structure of the data

c. To generalize the dataset

d. To represent the data

The correct answer is **b, d**.

Explanation: The goal of unsupervised learning is to understand the structure of the data and represent it. There is no right or certain answer in unsupervised learning.

Scikit-Learn

Scikit is a powerful and modern machine learning Python library for fully and semi-automated data analysis and information extraction.

Efficient tools to identify and organize problems (Supervised/ Unsupervised)

Free and open datasets

Rich set of libraries for learning and predicting

Model support for every problem type

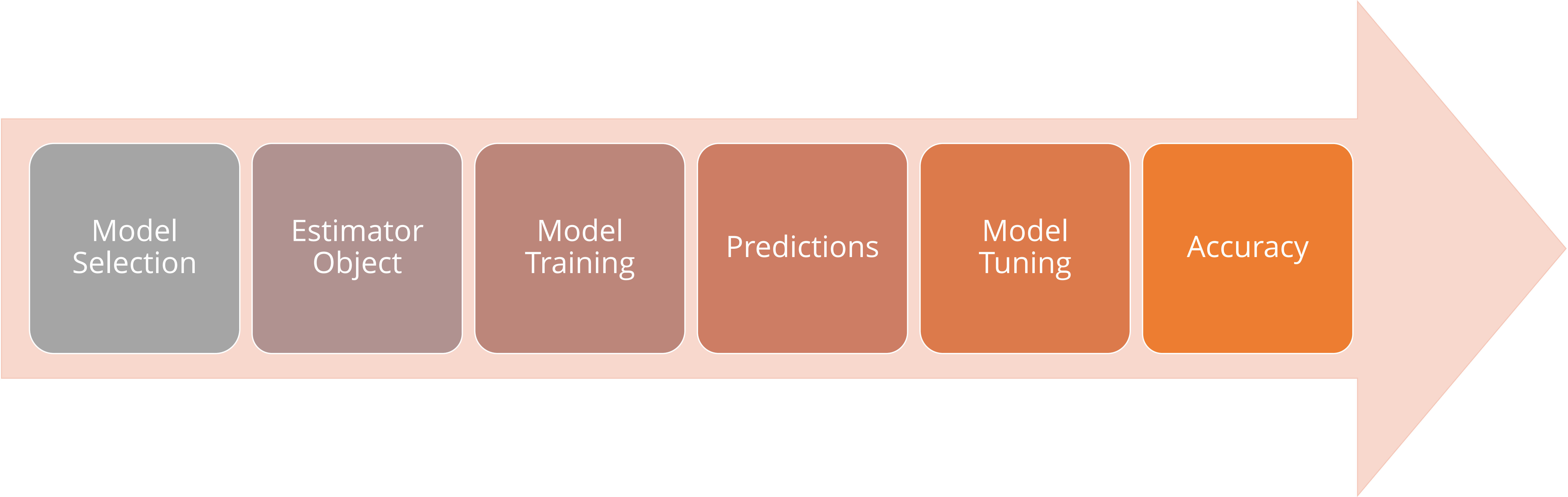
Model persistence

open source initiative®

Open source community and vendor support

Scikit-Learn—Problem-Solution Approach

Scikit-learn helps Data Scientists organize their work through its problem-solution approach.



Scikit-Learn—Problem-Solution Considerations

While working with a Scikit-Learn dataset or loading your own data to Scikit -Learn, always consider these four points:

- ✓ Create separate objects for feature and response.
- ✓ Ensure that features and response have only numeric values.
- ✓ Features and response should be in the form of a NumPy ndarray.
- ✓ Since features and response would be in the form of arrays, they would have shapes and sizes.
- ✓ Features are always mapped as *x*, and response is mapped as *y*.

Knowledge Check

KNOWLEDGE CHECK

The estimator instance in Scikit-learn is a _____.

a. model

b. feature

c. dataset

d. response

KNOWLEDGE CHECK

The estimator instance in Scikit-learn is a _____.

a. model

b. feature

c. dataset

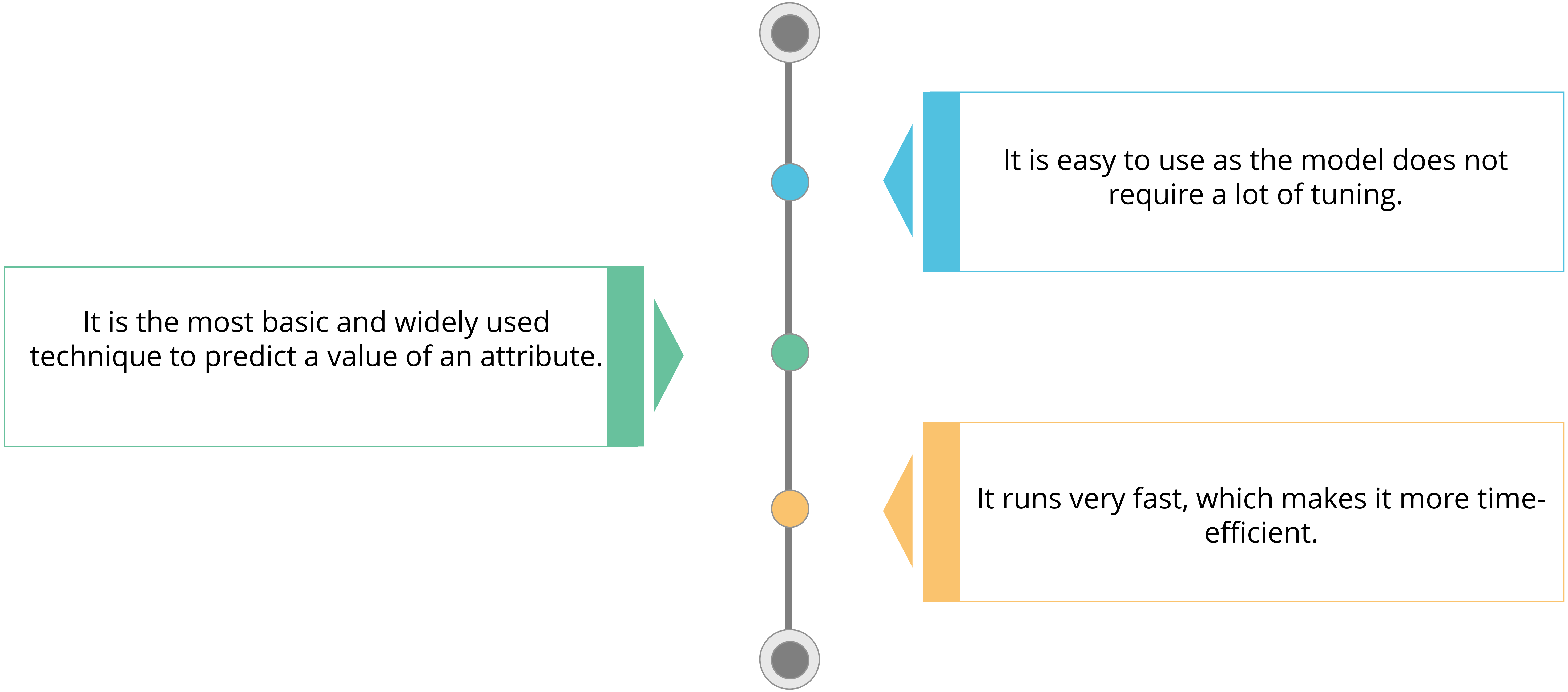
d. response

The correct answer is **a** .

Explanation: The estimator instance or object is a model.

Supervised Learning Models: Linear Regression

Linear regression is a supervised learning model used to analyze continuous data.



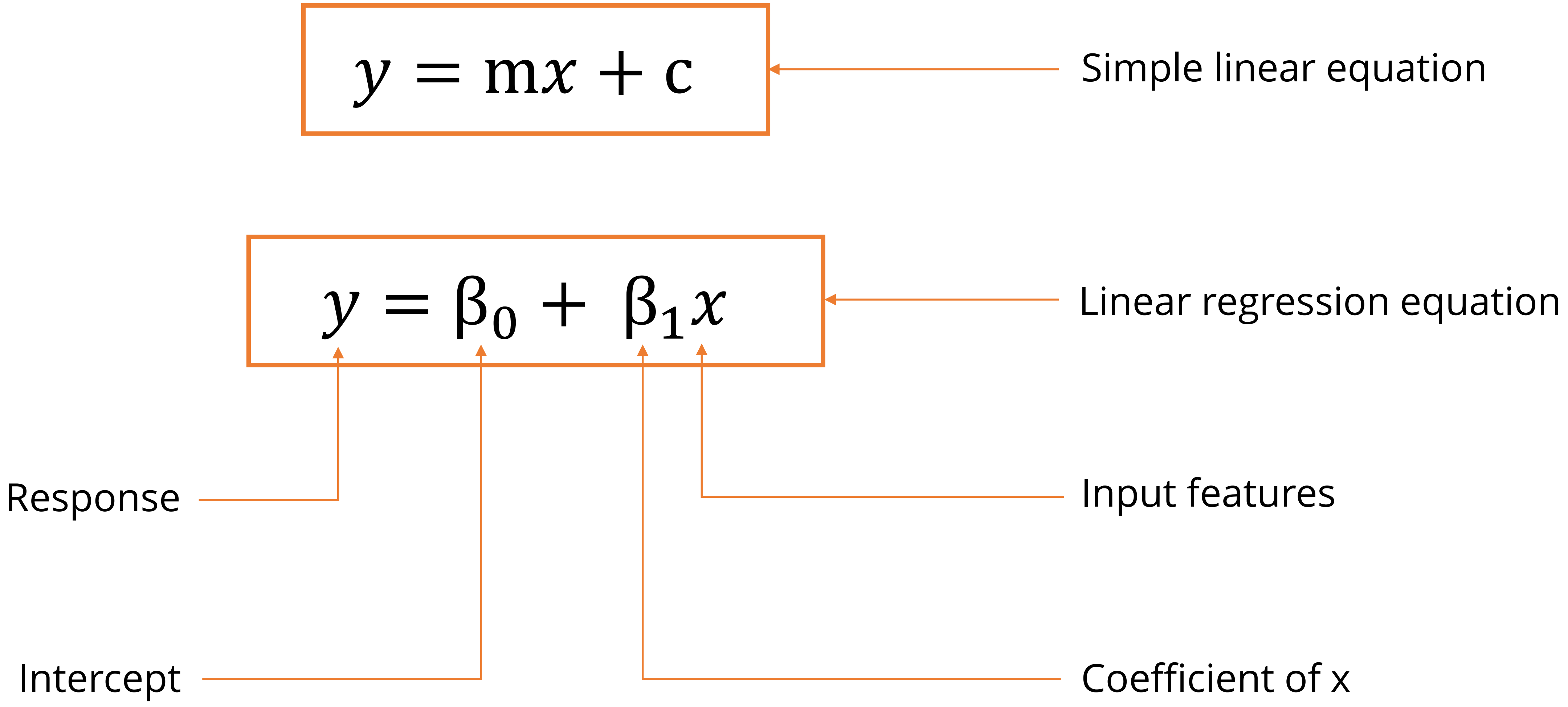
It is easy to use as the model does not require a lot of tuning.

It is the most basic and widely used technique to predict a value of an attribute.

It runs very fast, which makes it more time-efficient.

Supervised Learning Models: Linear Regression (contd.)

The linear regression equation is based on the formula for a simple linear equation.



$$y = \mathrm{m}x + \mathrm{c}$$

Simple linear equation

$$y = \beta_0 + \ \beta_1 x$$

Linear regression equation

Response

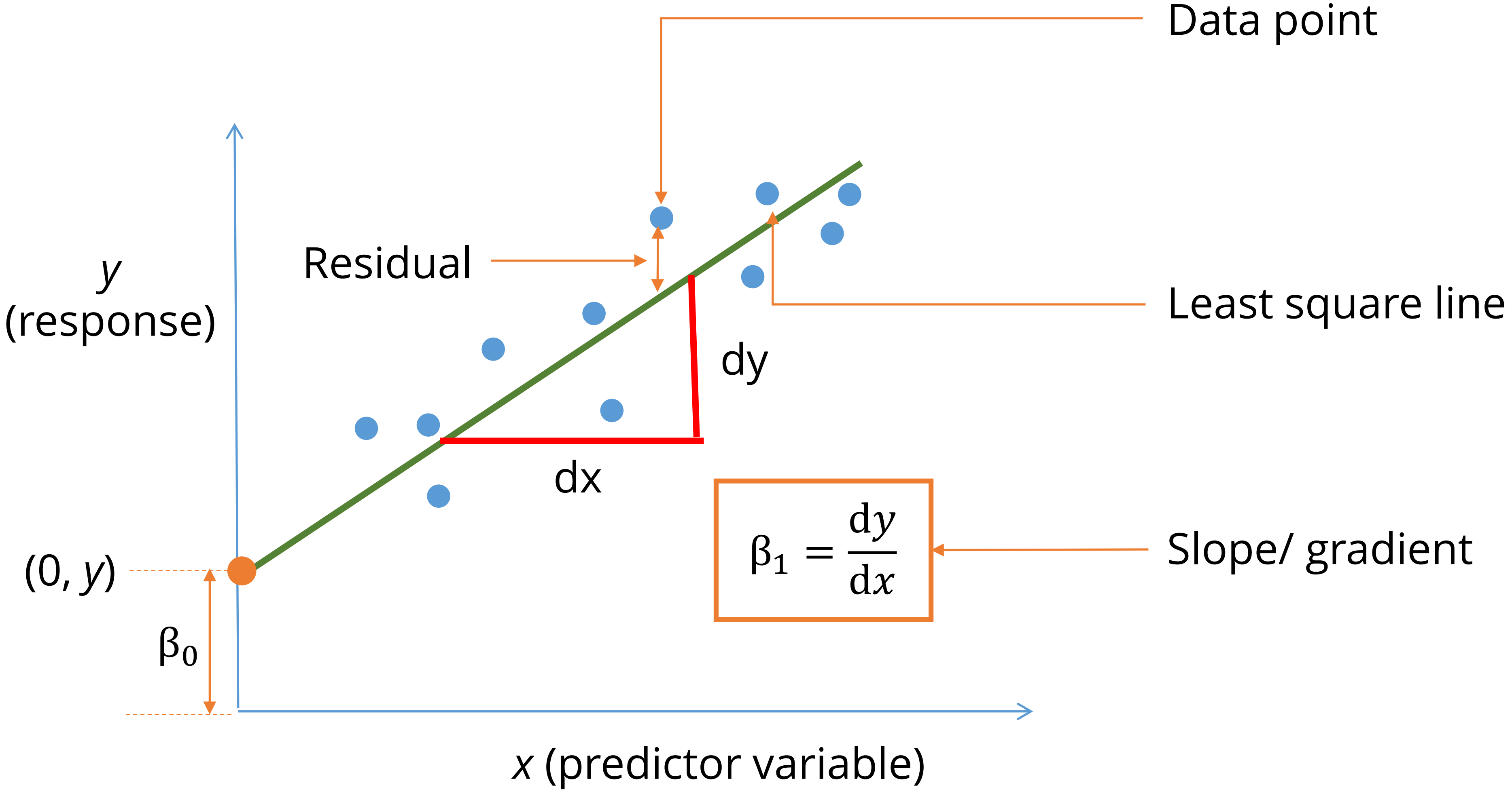
Intercept

Input features

Coefficient of x

Supervised Learning Models: Linear Regression (contd.)

Linear regression is the most basic technique to predict a value of an attribute.



Data point

Residual

Least square line

Slope/ gradient

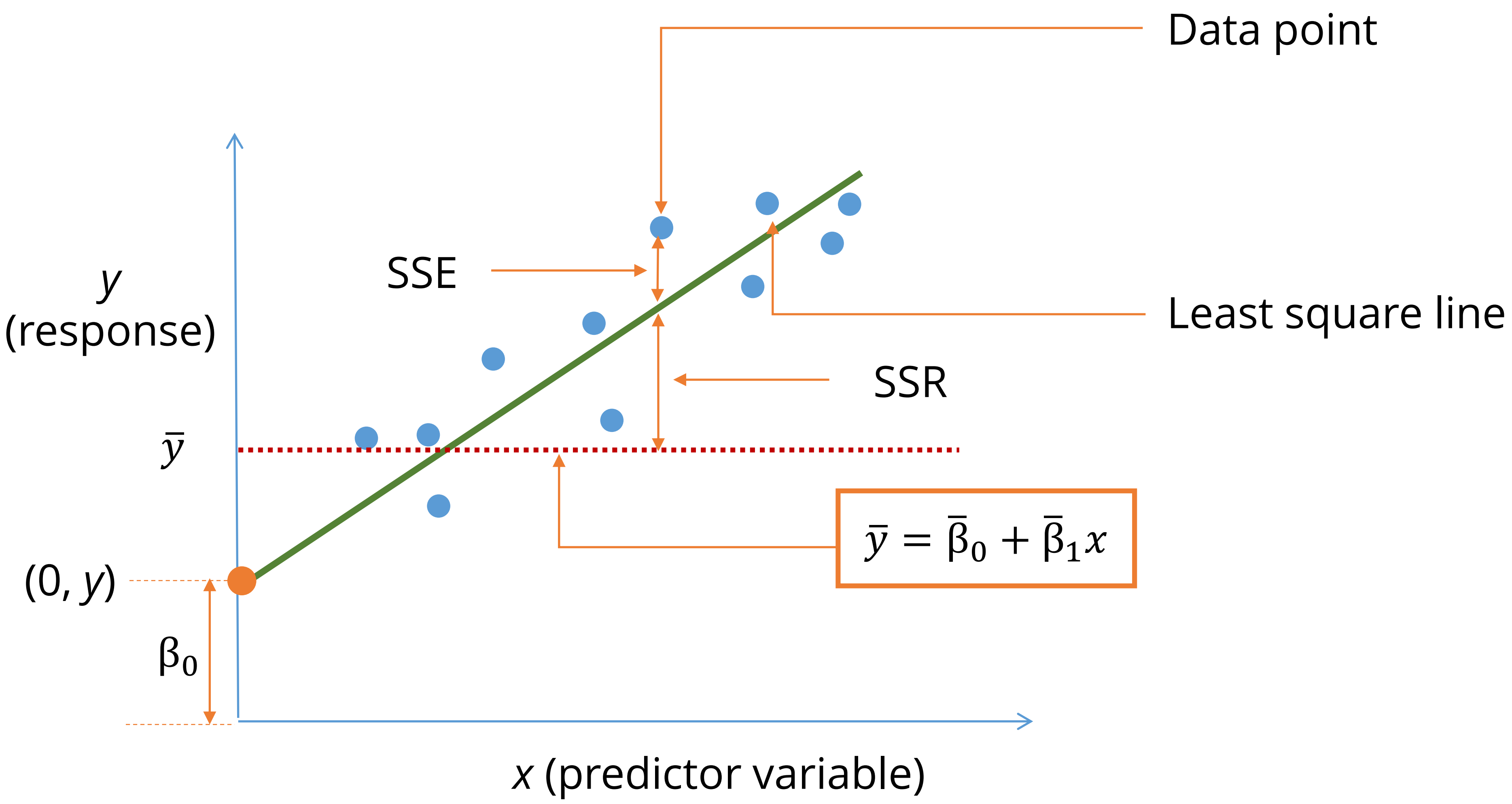
$$\beta_1 = \frac{dy}{dx}$$

$$\boxed{\underset{\text{Actual value}}{y} = \underbrace{\beta_0 + \beta_1 x}_{\text{Predicted value}} + \overset{\text{Residual}}{u}}$$

The attributes are usually fitted using the “least square” approach.

Supervised Learning Models: Linear Regression (contd.)

Smaller the value of SSR or SSE, the more accurate the prediction will be, which would make the model the best fit.



$$y = \beta_0 + \beta_1 x + \mathrm{u}$$

$$\mathrm{SSR} = \sum (\hat{y}_i - \bar{y})^2$$

Regression of sum of squares

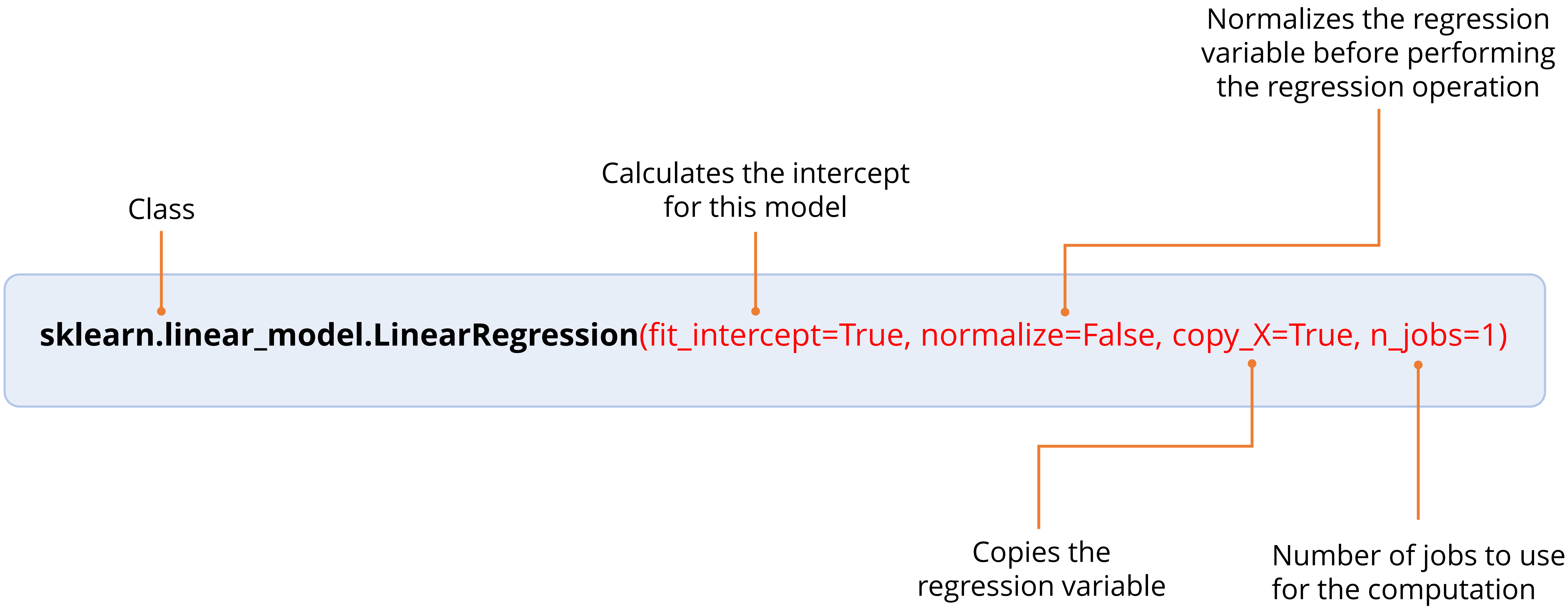
$$\mathrm{SSE} = \sum (y_i - \hat{y}_i)^2$$

Error of sum of squares

! The attributes are usually fitted using the “least square” approach.

Supervised Learning Models: Linear Regression (contd.)

Let us see how linear regression works in Scikit-Learn.



```
sklearn.linear_model.LinearRegression(fit_intercept=True, normalize=False, copy_X=True, n_jobs=1)
```

Class

Calculates the intercept for this model

Normalizes the regression variable before performing the regression operation

Copies the regression variable

Number of jobs to use for the computation

Demo 01—Loading a Dataset

Demonstrate how to load a built-in scikit-learn dataset

Demo 02—Linear Regression Model

Demonstrate how to create and train a linear regression model

Supervised Learning Models: Logistic Regression

Logistic regression is a generalization of the linear regression model used for classification problems.

$$\pi = \Pr(y = 1|x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

Probability of *y* = 1, given *x*

Change in the log-odds
for a unit change in x

The purpose of K-NN is to predict the class for each observation.

Supervised Learning Models: Logistic Regression (contd.)

Logistic regression is a generalization of the linear regression model used for classification problems.

$$\text{Odds} = \frac{\pi}{1-\pi}$$

Probability

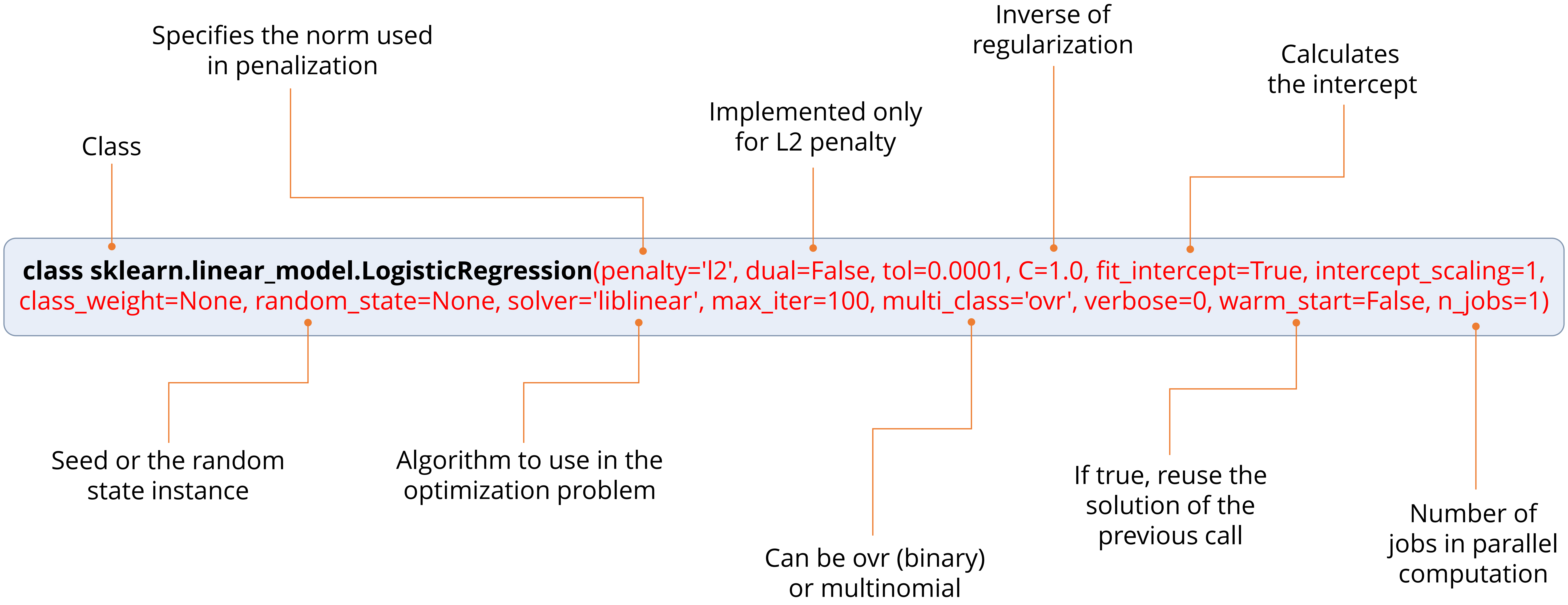
$$\log\left(\frac{\pi}{1-\pi}\right) = \log\left(e^{\beta_0+\beta_1 x}\right) = \beta_0 + \beta_1 x$$

Logarithm of odds

Linear regression

Supervised Learning Models: Logistic Regression (contd.)

Logistic regression is a generalization of the linear regression model used for classification problems.



```
class sklearn.linear_model.LogisticRegression(penalty='l2', dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1,
class_weight=None, random_state=None, solver='liblinear', max_iter=100, multi_class='ovr', verbose=0, warm_start=False, n_jobs=1)
```

Class

Specifies the norm used in penalization

Implemented only for L2 penalty

Inverse of regularization

Calculates the intercept

Seed or the random state instance

Algorithm to use in the optimization problem

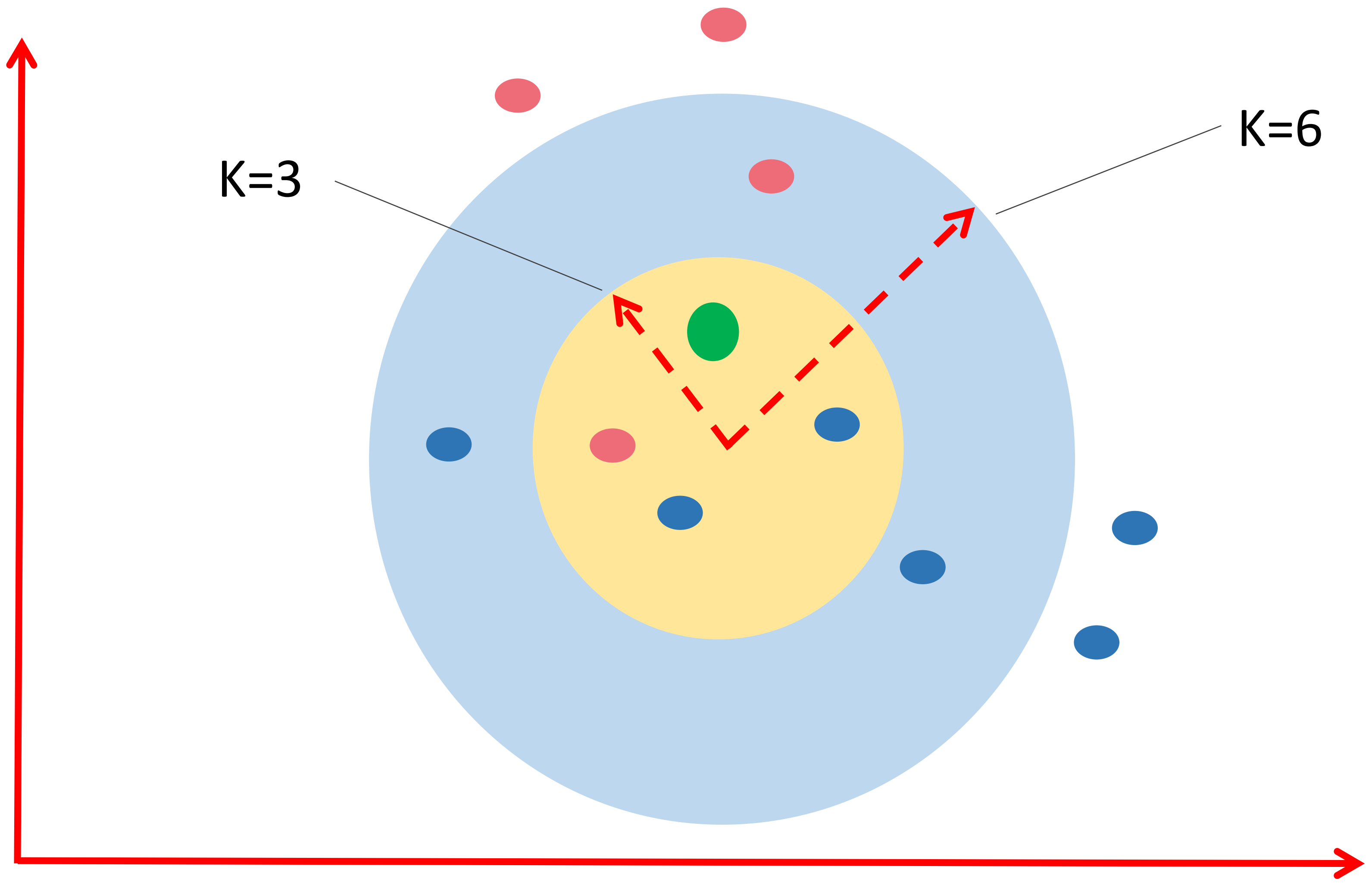
Can be ovr (binary) or multinomial

If true, reuse the solution of the previous call

Number of jobs in parallel computation

Supervised Learning Models: K Nearest Neighbors (K-NN)

K-nearest neighbors, or K-NN, is one of the simplest machine learning algorithms used for both classification and regression problem types.



If you are using this method for binary classification, choose an odd number for k to avoid the case of a "tied" distance between two classes.

Demo 03—K-NN and Logistic Regression Models

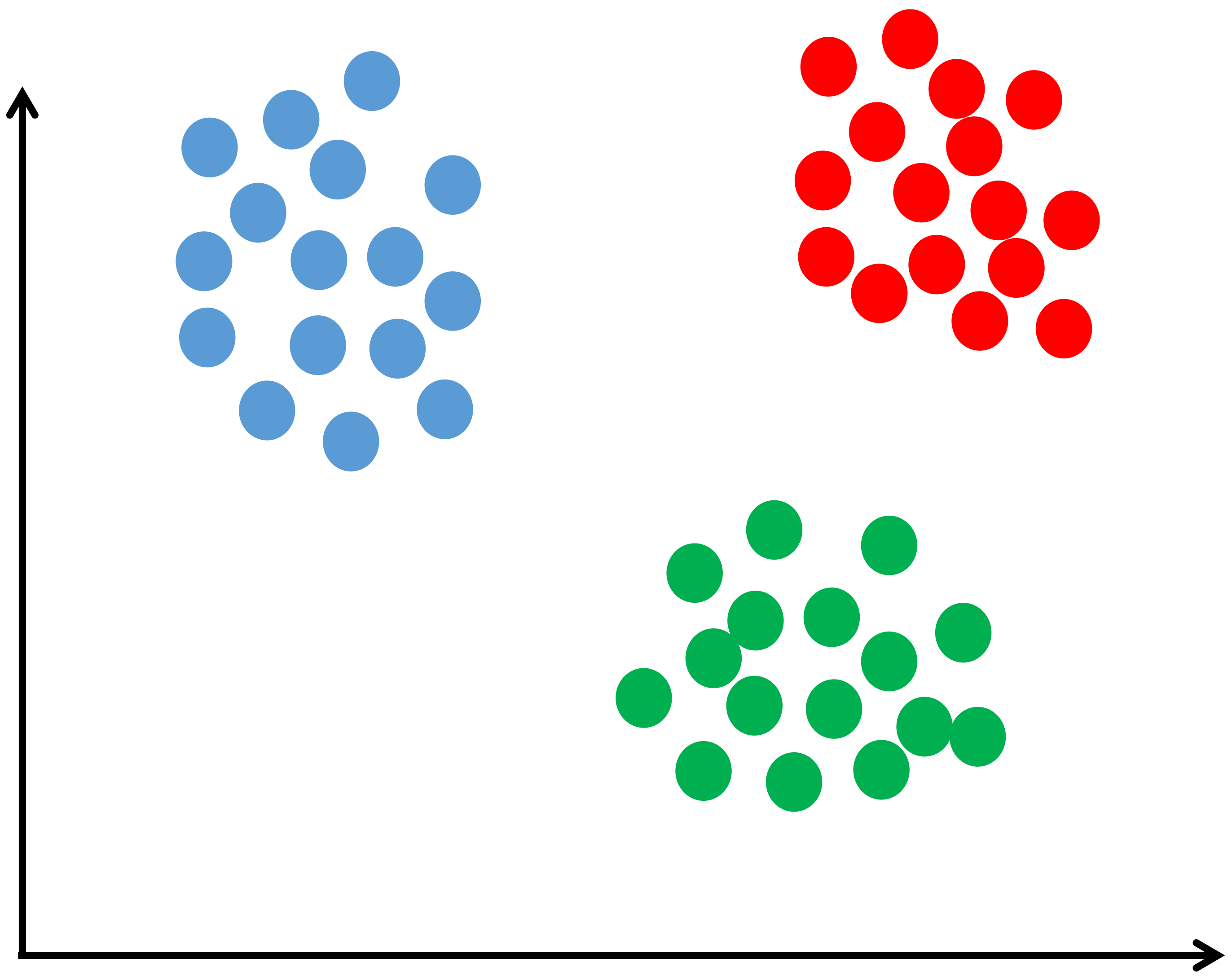
Demonstrate the use of K-NN and logistic regression models

Unsupervised Learning Models: Clustering

A cluster is a group of similar data points.

It is used:

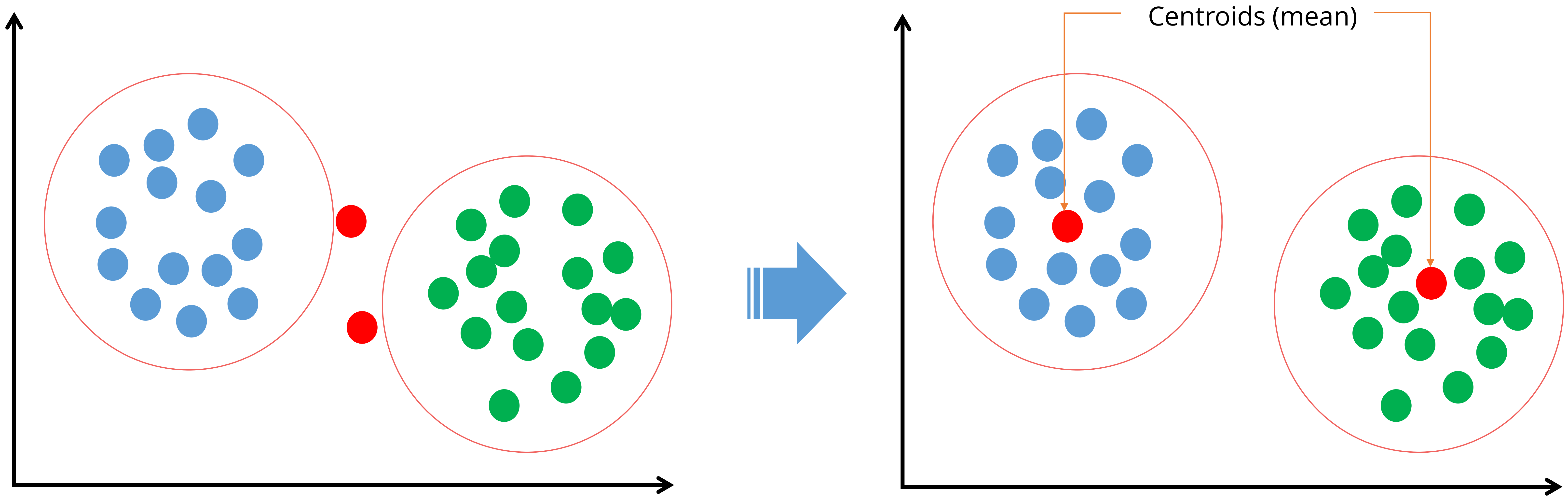
- To extract the structure of the data
- To identify groups in the data



Greater similarity between data points results in better clustering.

Unsupervised Learning Models: K-means Clustering

K-means finds the best centroids by alternatively assigning random centroids to a dataset and selecting mean data points from the resulting clusters to form new centroids. It continues this process iteratively until the model is optimized.



Assign

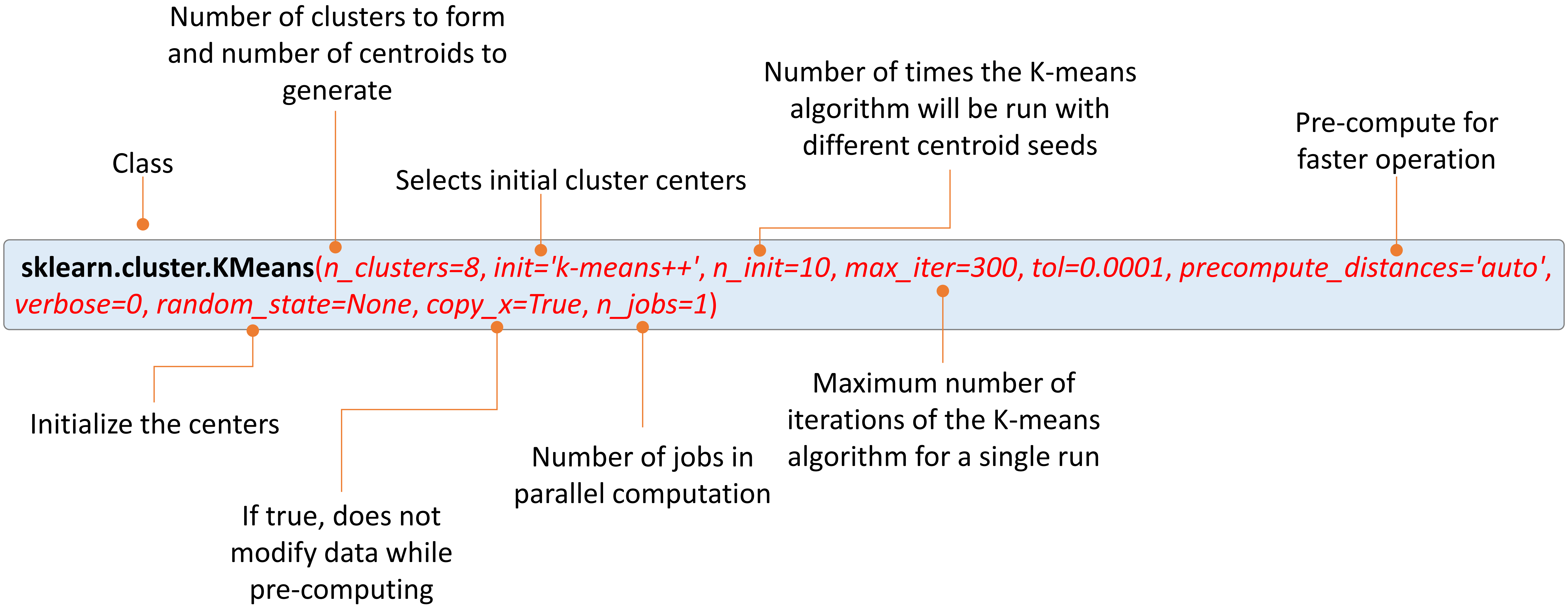
Find the number of clusters and assign mean

Optimize

Iterate and optimize the mean for each cluster for its respective data points

Unsupervised Learning Models: K-means Clustering (contd.)

Let us see how the k-means algorithm works in Scikit-Learn.



Number of clusters to form and number of centroids to generate

Number of times the K-means algorithm will be run with different centroid seeds

Pre-compute for faster operation

Class

Selects initial cluster centers

```
sklearn.cluster.KMeans(n_clusters=8, init='k-means++', n_init=10, max_iter=300, tol=0.0001, precompute_distances='auto',
verbose=0, random_state=None, copy_x=True, n_jobs=1)
```

Maximum number of iterations of the K-means algorithm for a single run

Initialize the centers

Number of jobs in parallel computation

If true, does not modify data while pre-computing

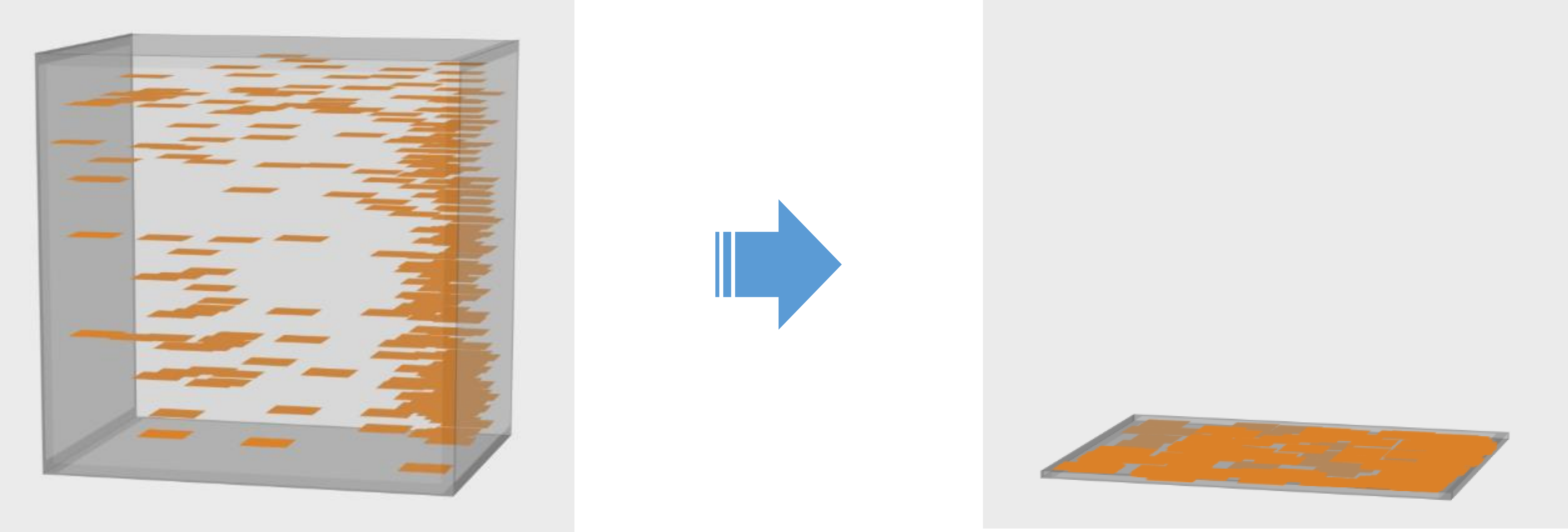
Demo 04—K-means Clustering

Demonstrate how to use k-means clustering to classify data points

DATA SCIENCE

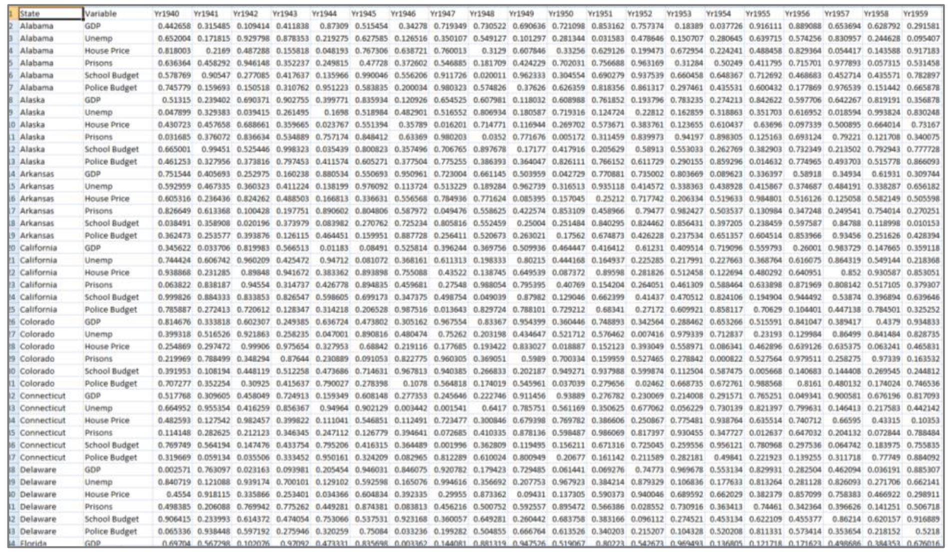
Unsupervised Learning Models: Dimensionality Reduction

It reduces a high-dimensional dataset into a dataset with fewer dimensions. This makes it easier and faster for the algorithm to analyze the data.



Unsupervised Learning Models: Dimensionality Reduction (contd.)

These are some techniques used for dimensionality reduction:



1	State	Variable	Yr1940	Yr1941	Yr1942	Yr1943	Yr1944	Yr1945	Yr1946	Yr1947	Yr1948	Yr1949	Yr1950	Yr1951	Yr1952	Yr1953	Yr1954	Yr1955	Yr1956	Yr1957	Yr1958	Yr1959
2	Alabama	GDP	0.442658	0.315485	0.109414	0.411838	0.87309	0.515454	0.34278	0.719349	0.730522	0.690636	0.721098	0.853162	0.757374	0.18389	0.037726	0.916111	0.889088	0.653694	0.628792	0.291581
3	Alabama	Unemp	0.652004	0.171815	0.929798	0.878353	0.219275	0.627585	0.126516	0.350107	0.549127	0.101297	0.281344	0.031583	0.478646	0.150707	0.280645	0.639715	0.574256	0.830957	0.244628	0.095407
4	Alabama	House Price	0.818003	0.2169	0.487288	0.155818	0.048193	0.767306	0.638721	0.760013	0.3129	0.607846	0.33256	0.629126	0.199473	0.672954	0.224241	0.488458	0.829364	0.054417	0.143588	0.917183
5	Alabama	Prisons	0.636364	0.458292	0.946148	0.352237	0.249815	0.47728	0.372602	0.546885	0.181709	0.424229	0.702031	0.756688	0.963169	0.31284	0.50249	0.411795	0.715701	0.977893	0.057315	0.531458
6	Alabama	School Budget	0.578769	0.90547	0.277085	0.417637	0.135966	0.990046	0.556206	0.911726	0.020011	0.962333	0.304554	0.690279	0.937539	0.660458	0.648367	0.712692	0.468683	0.452714	0.435571	0.782897
7	Alabama	Police Budget	0.745779	0.159693	0.150518	0.310762	0.951223	0.583835	0.200034	0.980323	0.574826	0.37626	0.626359	0.818356	0.861317	0.297461	0.435531	0.600432	0.177869	0.976539	0.151442	0.665878
8	Alaska	GDP	0.51315	0.239402	0.690371	0.902755	0.399771	0.835934	0.120926	0.654525	0.607981	0.118032	0.608988	0.761852	0.193796	0.783235	0.274213	0.842622	0.597706	0.642267	0.819191	0.356878
9	Alaska	Unemp	0.047899	0.329383	0.039415	0.261495	0.1698	0.518984	0.482901	0.516552	0.806934	0.180587	0.719316	0.124724	0.22812	0.162859	0.318863	0.351703	0.616952	0.018594	0.993824	0.830248
10	Alaska	House Price	0.430723	0.457658	0.688661	0.359665	0.023767	0.551394	0.35789	0.016201	0.714771	0.116944	0.269702	0.573671	0.383761	0.123655	0.610437	0.63696	0.097339	0.500895	0.664014	0.73167
11	Alaska	Prisons	0.031685	0.376072	0.836634	0.534889	0.757174	0.848412	0.63369	0.980203	0.0352	0.771676	0.005172	0.311459	0.839973	0.94197	0.898305	0.125163	0.693124	0.79221	0.121708	0.340075
12	Alaska	School Budget	0.665001	0.99451	0.525446	0.998323	0.035439	0.800823	0.357496	0.706765	0.897678	0.17177	0.417916	0.205629	0.58913	0.553033	0.262769	0.382903	0.732349	0.213502	0.792943	0.777728
13	Alaska	Police Budget	0.461253	0.327956	0.373816	0.797453	0.411574	0.605271	0.377504	0.775255	0.386393	0.364047	0.826111	0.766152	0.611729	0.290155	0.859296	0.014632	0.774965	0.493703	0.515778	0.866093
14	Arkansas	GDP	0.751544	0.405693	0.252975	0.160238	0.880534	0.550693	0.950961	0.723004	0.661145	0.503959	0.042729	0.770881	0.735002	0.803669	0.089623	0.336397	0.58918	0.34934	0.61931	0.309744
15	Arkansas	Unemp	0.592959	0.467335	0.360323	0.411224	0.138199	0.976092	0.113724	0.513229	0.189284	0.962739	0.316513	0.935118	0.414572	0.338363	0.438928	0.415867	0.374687	0.484191	0.338287	0.656182
16	Arkansas	House Price	0.605316	0.236436	0.824262	0.488503	0.166813	0.336631	0.556568	0.784936	0.771624	0.085395	0.157045	0.25212	0.717742	0.206334	0.519633	0.984801	0.516126	0.125058	0.582149	0.505598
17	Arkansas	Prisons	0.826649	0.613368	0.100428	0.197751	0.890602	0.804806	0.587972	0.049476	0.558625	0.422574	0.853109	0.458966	0.79477	0.982427	0.503537	0.130984	0.347248	0.249541	0.754014	0.270251
18	Arkansas	School Budget	0.038491	0.358908	0.020196	0.373979	0.083982	0.270762	0.725234	0.805816	0.552459	0.25004	0.251484	0.840295	0.824462	0.856431	0.397205	0.238459	0.597587	0.84788	0.118998	0.010153
19	Arkansas	Police Budget	0.362473	0.253577	0.393876	0.126115	0.464451	0.159951	0.887728	0.256411	0.520673	0.263021	0.17562	0.674873	0.426228	0.237534	0.651357	0.604514	0.853966	0.93456	0.251626	0.428394
20	California	GDP	0.345622	0.033706	0.819983	0.566513	0.01183	0.08491	0.525814	0.396244	0.369756	0.509936	0.464447	0.416412	0.61231	0.409514	0.719096	0.559793	0.26001	0.983729	0.147665	0.359118
21	California	Unemp	0.744424	0.606742	0.960209	0.425472	0.94712	0.081072	0.368161	0.611313	0.198333	0.80215	0.444168	0.164937	0.225285	0.217991	0.227663	0.368764	0.616075	0.864319	0.549144	0.218368
22	California	House Price	0.938868	0.231285	0.89848	0.941672	0.383362	0.893898	0.755088	0.43522	0.138745	0.649539	0.087372	0.89598	0.281826	0.512458	0.122694	0.480292	0.640951	0.852	0.930587	0.853051
23	California	Prisons	0.063822	0.838187	0.94554	0.314737	0.426778	0.894835	0.459681	0.27548	0.988054	0.795395	0.40769	0.154204	0.264051	0.461309	0.588464	0.633898	0.871969	0.808142	0.517105	0.379307
24	California	School Budget	0.999826	0.884333	0.833853	0.826547	0.598605	0.699173	0.347375	0.498754	0.049039	0.87982	0.129046	0.662399	0.41437	0.470512	0.824106	0.194904	0.944492	0.53874	0.396894	0.639646
25	California	Police Budget	0.785887	0.272413	0.720612	0.128347	0.314218	0.206528	0.987516	0.013643	0.829724	0.788101	0.729212	0.68341	0.27172	0.609921	0.858117	0.70629	0.104401	0.447138	0.784501	0.325252
26	Colorado	GDP	0.814676	0.333818	0.602307	0.249385	0.636724	0.473802	0.305162	0.967554	0.83367	0.954399	0.360446	0.748893	0.342564	0.288462	0.653266	0.515591	0.841047	0.389417	0.4379	0.934833
27	Colorado	Unemp	0.399318	0.516526	0.921863	0.258235	0.047001	0.890816	0.480474	0.75262	0.203198	0.434647	0.521712	0.576462	0.007416	0.979339	0.712837	0.23193	0.129984	0.86499	0.841484	0.828735
28	Colorado	House Price	0.254869	0.297472	0.99906	0.975654	0.327953	0.68842	0.219116	0.177685	0.193422	0.833027	0.018887	0.152123	0.393049	0.558971	0.086341	0.462896	0.639126	0.635375	0.063241	0.465831
29	Colorado	Prisons	0.219969	0.788499	0.348294	0.87644	0.230889	0.091053	0.822775	0.960305	0.369051	0.5989	0.700334	0.159959	0.527465	0.278842	0.000822	0.527564	0.979511	0.258275	0.97339	0.163532
30	Colorado	School Budget	0.391953	0.108194	0.448119	0.512258	0.473686	0.714631	0.967813	0.940385	0.266833	0.202187	0.949271	0.937988	0.599874	0.112504	0.587475	0.005668	0.140683	0.144408	0.269545	0.244812
31	Colorado	Police Budget	0.707277	0.352254	0.30925	0.415637	0.790027	0.278398	0.1078	0.564818	0.174019	0.545961	0.037039	0.279656	0.02462	0.668735	0.672761	0.988568	0.8161	0.480132	0.174024	0.746536
32	Connecticut	GDP	0.517768	0.309605	0.458049	0.724913	0.159349	0.608148	0.277353	0.245646	0.222746	0.911456	0.93889	0.276782	0.230069	0.214008	0.291571	0.765251	0.049341	0.900581	0.676196	0.817093
33	Connecticut	Unemp	0.664952	0.955354	0.416259	0.856367	0.94964	0.902129	0.003442	0.001541	0.6417	0.785751	0.561169	0.350625	0.677062	0.056229	0.730139	0.821397	0.799631	0.146413	0.217583	0.442142
34	Connecticut	House Price	0.482593	0.127542	0.982457	0.399822	0.111041	0.546851	0.112491	0.723477	0.300846	0.679398	0.769782	0.386606	0.250867	0.775481	0.938764	0.635514	0.740712	0.66595	0.43315	0.10353
35	Connecticut	Prisons	0.114148	0.282625	0.212123	0.346345	0.247112	0.126779	0.394641	0.072685	0.410335	0.878136	0.598487	0.986069	0.817397	0.930455	0.347727	0.012637	0.647032	0.204132	0.072844	0.788484
36	Connecticut	School Budget	0.769749	0.564194	0.147476	0.433754	0.795206	0.416315	0.364489	0.001996	0.362809	0.119495	0.156211	0.671316	0.725045	0.259556	0.956121	0.780968	0.297536	0.064742	0.183975	0.755835
37	Connecticut	Police Budget	0.319669	0.059134	0.035506	0.333452	0.950161	0.324209	0.082965	0.812289	0.610024	0.800949	0.20677	0.161142	0.211589	0.282181	0.49841	0.221923	0.139255	0.311718	0.77749	0.884092
38	Delaware	GDP	0.002571	0.763097	0.023163	0.093981	0.205454	0.946031	0.846075	0.920782	0.179423	0.729485	0.061441	0.069276	0.74773	0.969678	0.553134	0.829931	0.282504	0.462094	0.036191	0.885307
39	Delaware	Unemp	0.840719	0.121088	0.939174	0.700101	0.129102	0.592598	0.165076	0.994616	0.356692	0.207753	0.967923	0.384214	0.879329	0.106836	0.177633	0.813264	0.281128	0.826093	0.271706	0.662141
40	Delaware	House Price	0.4554	0.918115	0.335866	0.253401	0.034366	0.604834	0.392335	0.29955	0.873362	0.09431	0.137305	0.590373	0.940046	0.689592	0.662029	0.382379	0.857099	0.758383	0.466922	0.298911
41	Delaware	Prisons	0.498385	0.206088	0.769942	0.775262	0.449281	0.874381	0.083813	0.456216	0.500752	0.592557	0.895472	0.566386	0.028552	0.730916	0.363413	0.74461	0.342364	0.396626	0.141251	0.506718
42	Delaware	School Budget	0.906415	0.233993	0.614372	0.474054	0.753066	0.537531	0.923168	0.360057	0.649281	0.260442	0.683758	0.383166	0.096112	0.274521	0.453134	0.622109	0.455377	0.86214	0.620157	0.916889
43	Delaware	Police Budget	0.065336	0.938448	0.597192	0.275946	0.320259	0.75084	0.033236	0.199282	0.504855	0.666764	0.613526	0.340203	0.215207	0.104328	0.520208	0.811331	0.573414	0.353654	0.218152	0.5218
44	Florida	GDP	0.69704	0.567298	0.102076	0.97092	0.473331	0.835698	0.003362	0.144081	0.881319	0.947526	0.519067	0.80223	0.542673	0.969493	0.136805	0.121718	0.171623	0.498686	0.384353	0.676016

Large dataset
(a few thousand columns and rows)

Drop data columns with missing values

Drop data columns with low variance

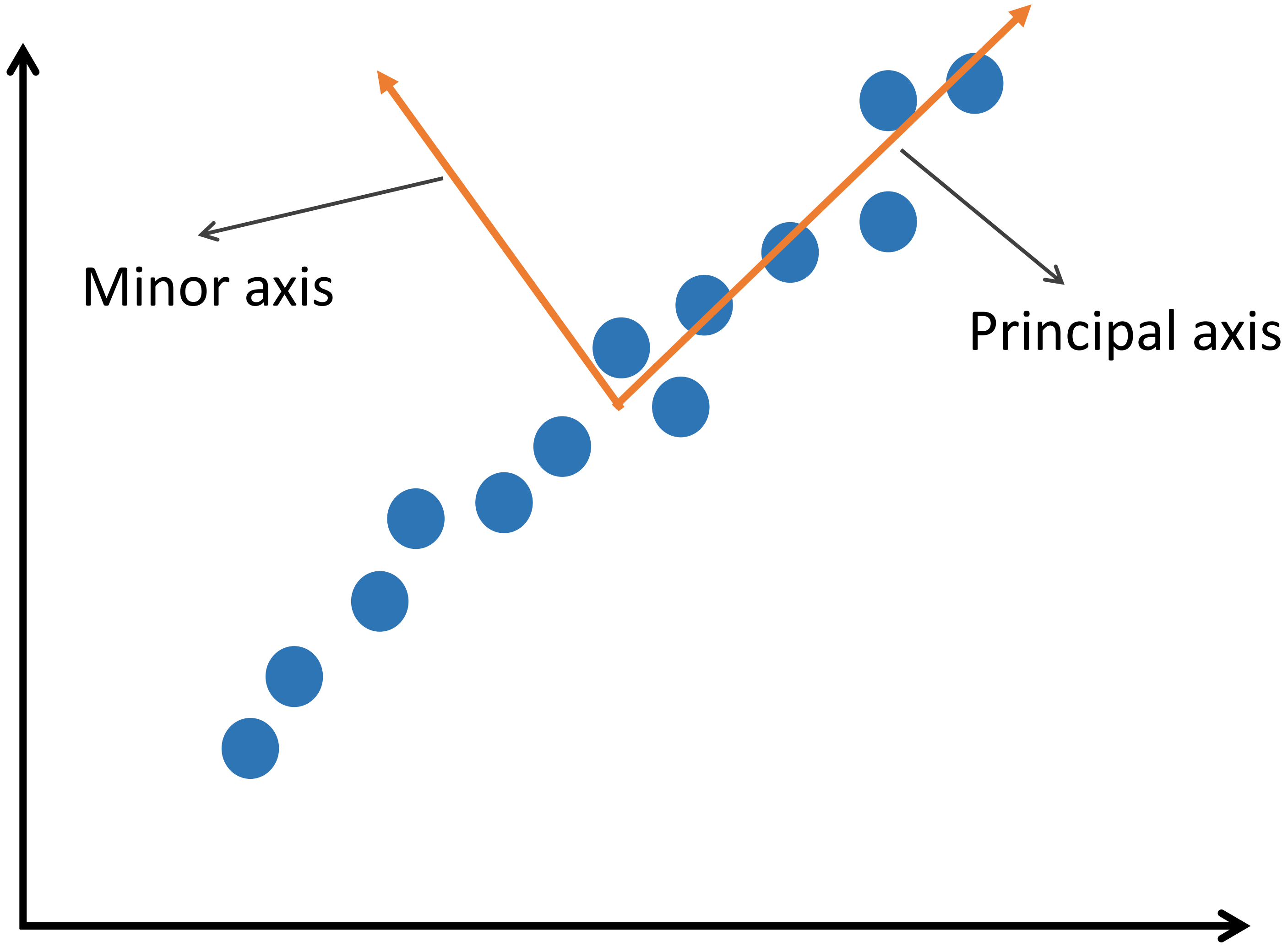
Drop data columns with high correlations

Apply statistical functions - PCA

Unsupervised Learning Models: Principal Component Analysis (PCA)

It is a linear dimensionality reduction method which uses singular value decomposition of the data and keeps only the most significant singular vectors to project the data to a lower dimensional space.

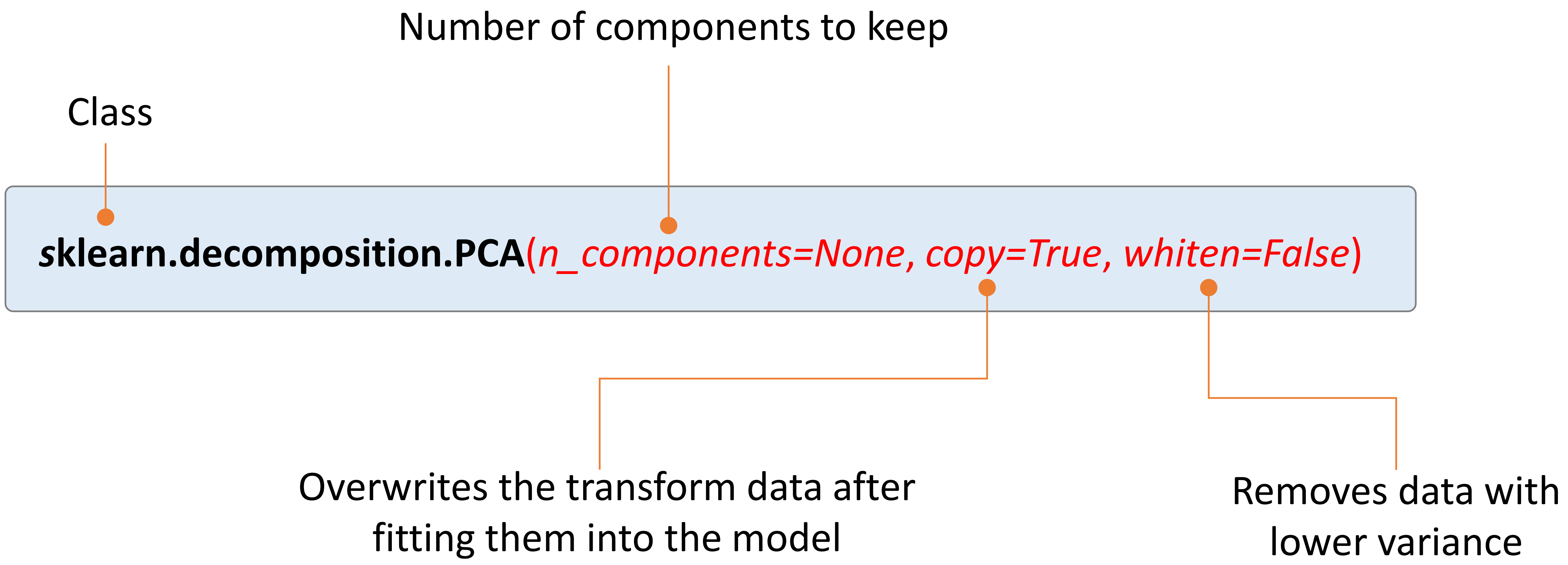
- It is primarily used to compress or reduce the data.
- PCA tries to capture the variance, which helps it pick up interesting features.
- PCA is used to reduce dimensionality in the dataset and to build our feature vector.
- Here, the principal axes in the feature space represents the direction of maximum variance in the data.



This method is used to capture variance.

Unsupervised Learning Models: Principal Component Analysis (PCA)

Let us look at how the PCA algorithm works in Scikit-Learn.



Number of components to keep

Class

```
sklearn.decomposition.PCA(n_components=None, copy=True, whiten=False)
```

Overwrites the transform data after fitting them into the model

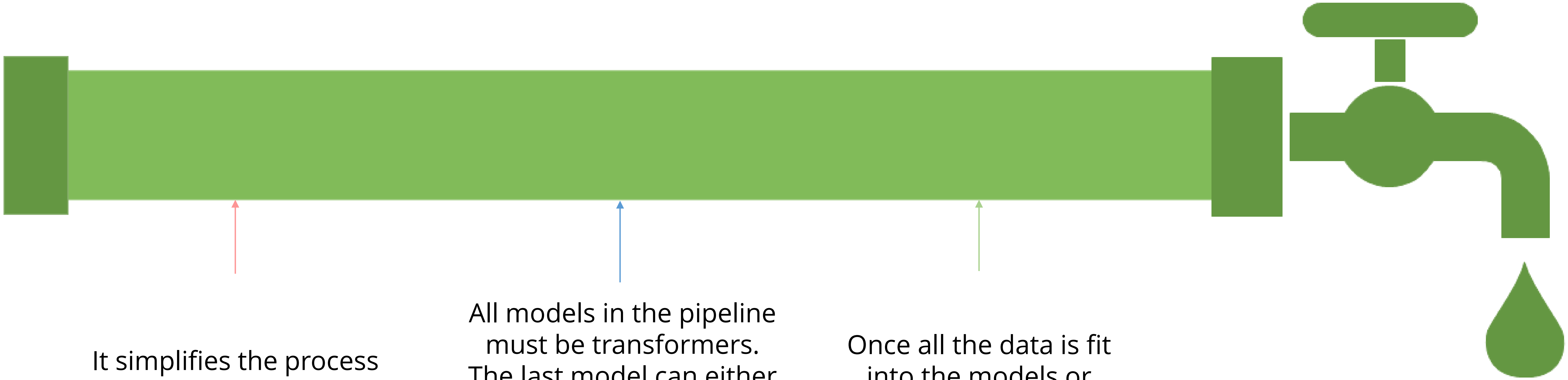
Removes data with lower variance

Demo 05—Principal Component Analysis (PCA)

Demonstrate how to use the PCA model to reduce the dimensions of a dataset

Pipeline

Pipeline is mainly used to combine multiple models or estimators. Its characteristics are as follows:



It simplifies the process where more than one model is required or used.

All models in the pipeline must be transformers. The last model can either be a transformer or a classifier, regressor, or other such objects.

Once all the data is fit into the models or estimators, the predict method can be called.

Estimators are known as ‘model instance’.

Demo 06—Pipeline

Demonstrate how to build a pipeline

DATA SCIENCE

Model Persistence

Save model for the future use. No need to retrain your model every time when you need them.

It is possible to save a model by using Python's Pickle method.

Scikit-learn has a special replacement for pickle called joblib.

You can use joblib.dump and joblib.load methods.

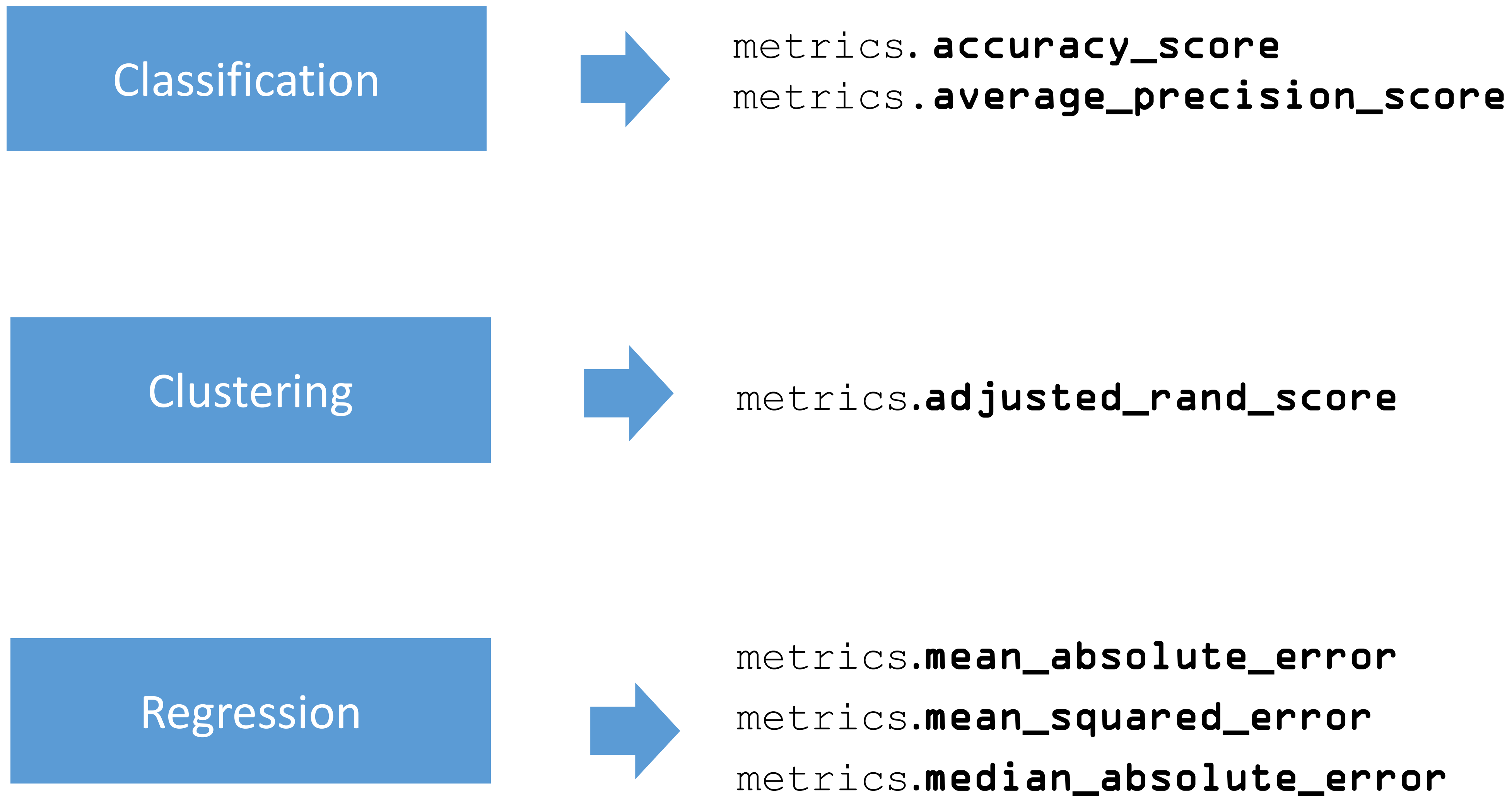
These are more efficient for Big Data.

Demo 06—Model Persistence

Demonstrate how to persist a model for future use

Model Evaluation: Metric Functions

You can use the “Metrics” function to evaluate the accuracy of your model’s predictions.



Classification

```
metrics. accuracy_score
metrics.average_precision_score
```

Clustering

```
metrics.adjusted_rand_score
```

Regression

```
metrics.mean_absolute_error
metrics.mean_squared_error
metrics.median_absolute_error
```

Knowledge Check

KNOWLEDGE CHECK

What is the best way to train a model?

a. Use the entire dataset as a training and testing set

b. Split the known dataset into separate training and testing sets

c. Ask the source to provide continuous data

d. Ask the source to provide categorical data

KNOWLEDGE CHECK

What is the best way to train a model?

a. Use the entire dataset as a training and testing set both

b. Split the known dataset into separate training and testing sets

c. Ask the source to provide continuous data

d. Ask the source to provide categorical data

The correct answer is **b**.

Explanation: The best way to train a model is to split the known dataset into training and testing sets. The testing set varies from 20% to 40%.

Assignment

Assignment 01

Problem

Instructions

The given dataset contains ad budgets for different media channels and the corresponding ad sales of XYZ firm. Evaluate the dataset to:

- Find the features or media channels used by the firm
- Find the sales figures for each channel
- Create a model to predict the sales outcome
- Split as training and testing datasets for the model
- Calculate the Mean Square Error (MSE)

Assignment 01

Problem Instructions

Instructions on performing the assignment:

- Download the “Advertising Budget and Sales.csv” file from the “Resource” tab. You can load the saved file to the Jupyter notebook that you would be using to complete the assignment.

Common instructions:

- If you are new to Python, download the “Anaconda Installation Instructions” document from the “Resources” tab to view the steps for installing Anaconda and the Jupyter notebook.
- Download the “Assignment 01” notebook and upload it on the Jupyter notebook to access it.
- Follow the provided cues to complete the assignment.

Assignment

Assignment 02

Problem

Instructions

The given dataset lists the glucose level readings of several pregnant women taken either during a survey examination or routine medical care. It specifies if the 2 hour post-load plasma glucose was at least 200 mg/dl. Analyze the dataset to:

1. Find the features of the dataset,
2. Find the response label of the dataset,
3. Create a model to predict the diabetes outcome,
4. Use training and testing datasets to train the model, and
5. Check the accuracy of the model.

Assignment 02

Problem | Instructions

Instructions on performing the assignment:

- Download the “pima-indians-diabetes.DATA” and “pima-indians-diabetes.NAMES” files from the “Resources” tab. Load the .DATA file to the Jupyter notebook to work on it.
- Open the .NAMES file with a notepad application to view its text. Use this file to view the features of the dataset and add them manually in your code.

Common instructions:

- If you are new to Python, download the “Anaconda Installation Instructions” document from the “Resources” tab to view the steps for installing Anaconda and the Jupyter notebook.
- Download the “Assignment 01” notebook and upload it on the Jupyter notebook to access it.
- Follow the provided cues to complete the assignment.

Quiz

QUIZ

1

Which of the following is true with a greater value of SSR or SSE? *Select all that apply*.

a. The prediction will be more accurate, making it the best fit model.

b. The prediction will start becoming less accurate.

c. The outcome remains unaffected.

d. The model will not be the best fit for the attributes.

QUIZ

1

Which of the following is true with a greater value of SSR or SSE? Select all that apply.

a. The prediction will be more accurate, making it the best fit model.

b. The prediction will start becoming less accurate.

c. The outcome remains unaffected.

d. The model will not be the best fit for the attributes.

The correct answer is **b, d**.

Explanation: With higher SSR or SSE, the prediction will be less accurate and the model will not be the best fit for the attributes.

QUIZ

2

Class sklearn.linear_model.LogisticRegression, random_state _____.

a. indicates the seed of the pseudo random number generator used to shuffle data

b. defines the features state

c. represents the number of random iterations

d. specifies a random constant to be added to the decision function

QUIZ

2

Class sklearn.linear_model.LogisticRegression, random_state ____.

a. indicates the seed of the pseudo random number generator used to shuffle data

b. defines the features state

c. represents the number of random iterations

d. specifies a random constant to be added to the decision function

The correct answer is a.

Explanation: The class “sklearn.linear_model.LogisticRegression, random_state” indicates the seed of the pseudo random number generator used to shuffle data.

QUIZ

3

What are the requirements of the K-means algorithm? ***Select all that apply*****.**

a. Number of clusters should be specified

b. More than one iteration should meet requisite criteria

c. Centroids should minimize inertia

d. Features should be labeled

QUIZ

3

What are the requirements of the K-means algorithm? *Select all that apply*.

a. Number of clusters should be specified

b. More than one iteration should meet requisite criteria

c. Centroids should minimize inertia

d. Features should be labeled

The correct answer is **a, b, c**.

Explanation: The K-means algorithm requires that the number of clusters be specified and that centroids that minimize inertia be selected. It requires several iterations to fine tune itself and meet the required criteria to become the best fit model.

QUIZ

4

In Class sklearn.decomposition.PCA, the transform(X) method , where X is multi-dimensional ______.

a. fits the model with X and applies the dimensionality reduction on X

b. transforms the data back to its original space

c. applies the dimensionality reduction on X

d. computes data co-variance with the generative model

QUIZ

3

In Class sklearn.decomposition.PCA, the transform(X) method , where X is multi-dimensional _____.

a. fits the model with X and applies the dimensionality reduction on X

b. transforms the data back to its original space

c. applies the dimensionality reduction on X

d. computes data co-variance with the generative model

The correct answer is **c.**

Explanation: In Class “sklearn.decomposition.PCA,” the transform(X) method applies the dimensionality reduction on X.

Key Takeaways

- Scikit-learn has many built-in functions and algorithms which make it easy for Data Scientists to build machine learning models.
- Machine learning can easily and quickly extract information from large sources of data.
- Supervised and unsupervised machine learning models are two of the most widely used learning models.
- Supervised learning models are used to predict the outcome of a dataset.
- Unsupervised learning models are used to find the structure of a dataset.
- For continuous data, you can use regression algorithms if it is a supervised learning model. However, use dimensionality reduction for unsupervised learning.
- For categorical data, use classification algorithms if it is a supervised learning model and clustering if it is an unsupervised learning model.

This concludes “Machine Learning with Scikit-Learn.”

The next lesson is “Natural Language Processing (NLP) with Scikit-Learn.”