Machine Learning data processing

Credit card fraud detection Case Demo

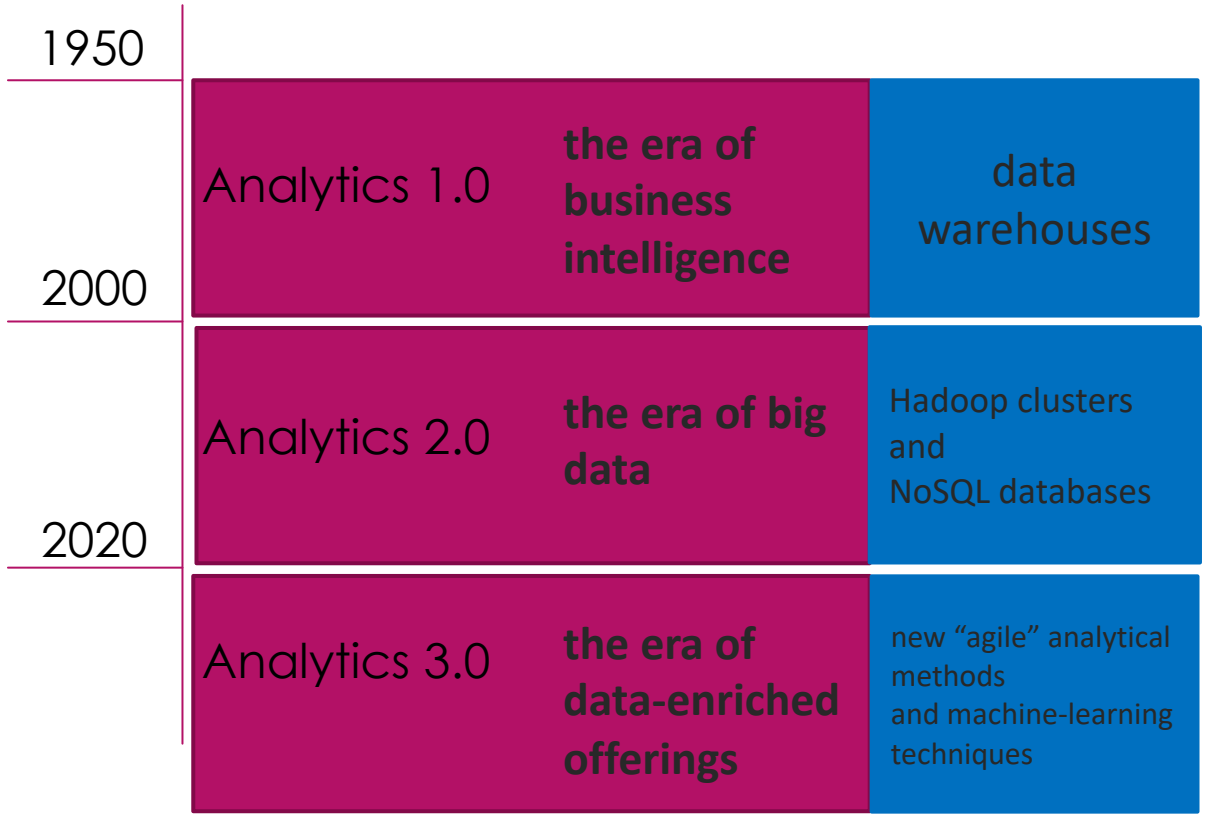
YAO ZHAO

Data everywhere!

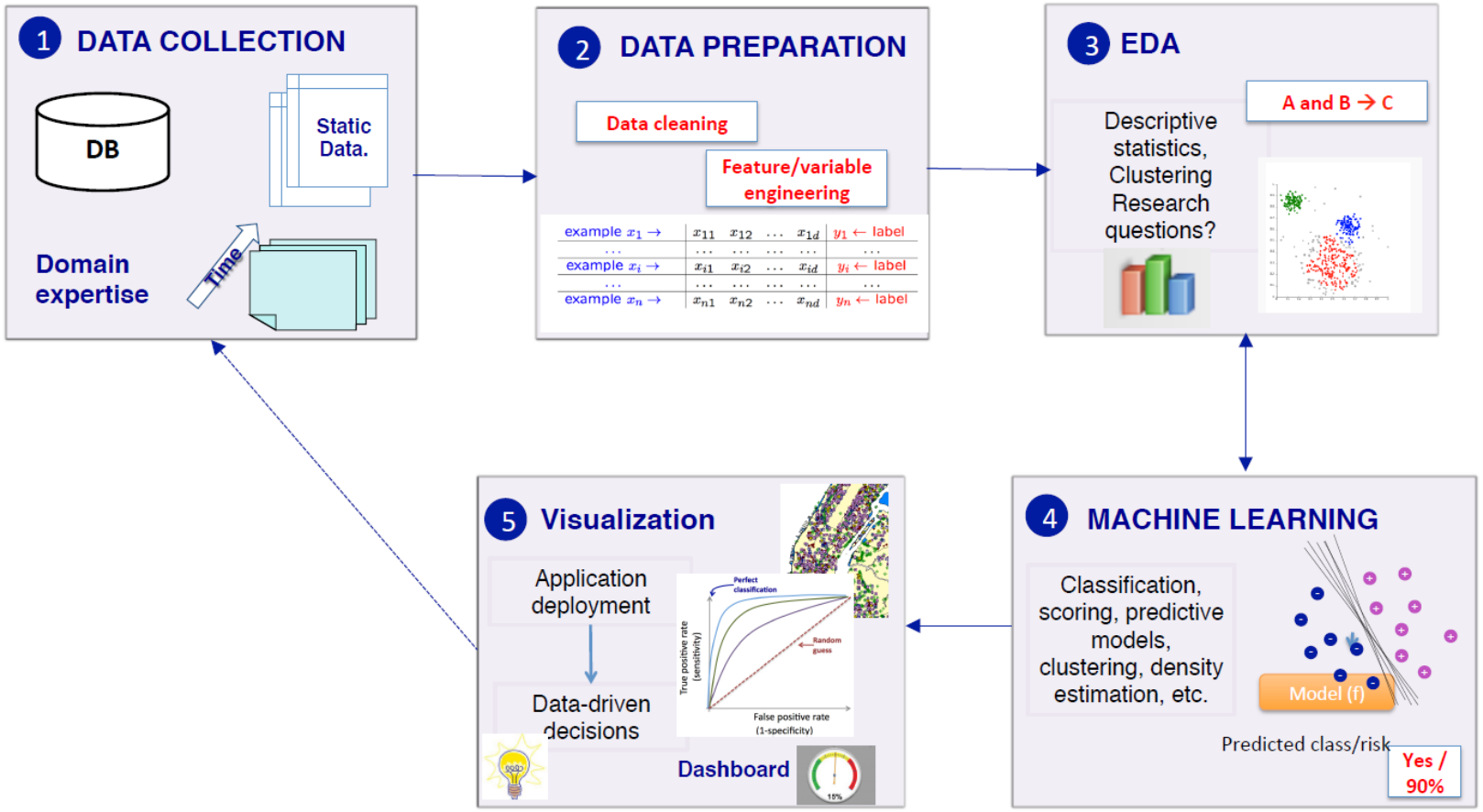
1. **Google**: processes 24 petabytes of data per day.
2. **Facebook**: 10 million photos uploaded every hour.
3. **YouTube**: 1 hour of video uploaded every second.
4. **Twitter**: 400 million tweets per day.
5. **Astronomy**: Satellite data is in hundreds of PB.
6. ...
7. **“By 2020 the digital universe will reach 44 zettabytes...“**

The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things, April 2014.

That's 44 trillion gigabytes!

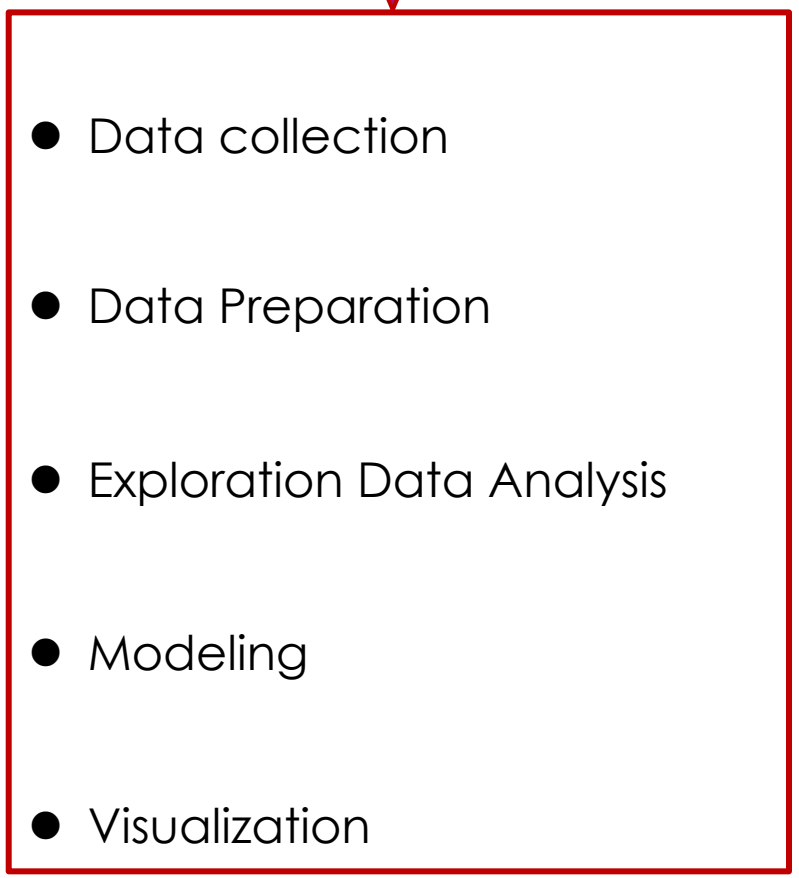

1950
2000
2020
Analytics 1.0
the era of business intelligence
data warehouses
Analytics 2.0
the era of big data
Hadoop clusters and NoSQL databases
Analytics 3.0
the era of data-enriched offerings
new “agile” analytical methods and machine-learning techniques

The Data Science process



Machine Learning

- Spam filtering
- Credit card fraud detection
- Digit recognition on checks, zip codes
- Detecting faces in images
- MRI image analysis
- Recommendation system
- Search engines
- Handwriting recognition
- Scene classification
- etc...



How to find datasets?

Big data competition platform

- Kaggle: https://www.kaggle.com/datasets

Colleges and Universities

- UCI: https://archive.ics.uci.edu/ml/datasets.html

Enterprises or public welfare organizations

- Google dataset: https://cloud.google.com/bigquery/public-data/

Government's open data

- The U.S. Government's open data: https://data.gov/

Searching datasets:

- https://datasetsearch.research.google.com/

Credit card fraud detection

- **Data collection**
- Data Preparation
- Exploration Data Analysis
- Modeling
- Visualization

www.kaggle.com/mlg-ulb/creditcardfraud

Download the dataset : creditcard.csv

- Data collection
- **Data Preparation**
- Exploration Data Analysis
- Modeling
- Visualization

Data Preparation

Eliminates duplicate and null values, corrupt data, inconsistent data types, invalid entries, missing data, and improper formatting.

Python libraries for Data Preparation

NumPy

- NumPy is the foundation for many other packages that hold the data science ecosystem like Pandas, Matplotlib and Scikit-learn.

Pandas

- Pandas offer developers fast, efficient and optimized objects for data manipulation in various academic and industrial fields.

Matplotlib

- The core package used for data visualization. Matplotlib offers various plots and figures developers can use to create different visualizations

Pandas

DataFrame : A Pandas DataFrame is a 2-dimensional data structure, like a 2-dimensional array, or a table with rows and columns.

pandas.read_csv(): Load a CSV into a DataFrame

pandas.DataFrame.head(): returns the headers and a specified number of rows, starting from the top.

pandas.DataFrame.tail(): viewing the *last* rows of the DataFrame

pandas.DataFrame.info():viewing more information about the data set.

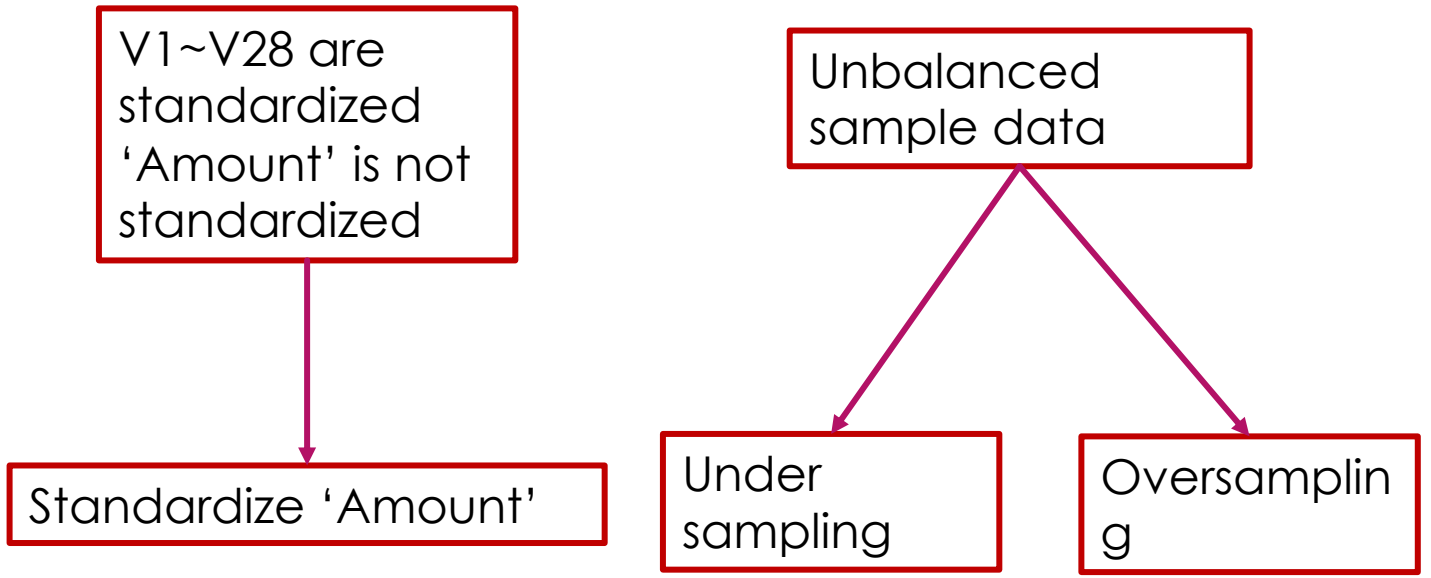
pandas.DataFrame.loc(): Access a group of rows and columns by label(s) or a boolean array.

DataFrame.describe(): Generate descriptive statistics.Descriptive statistics include those that summarize the central tendency, dispersion and shape of a dataset's distribution, excluding NaN values.

- Data collection
- Data Preparation
- **Exploration Data Analysis**
- Modeling
- Visualization

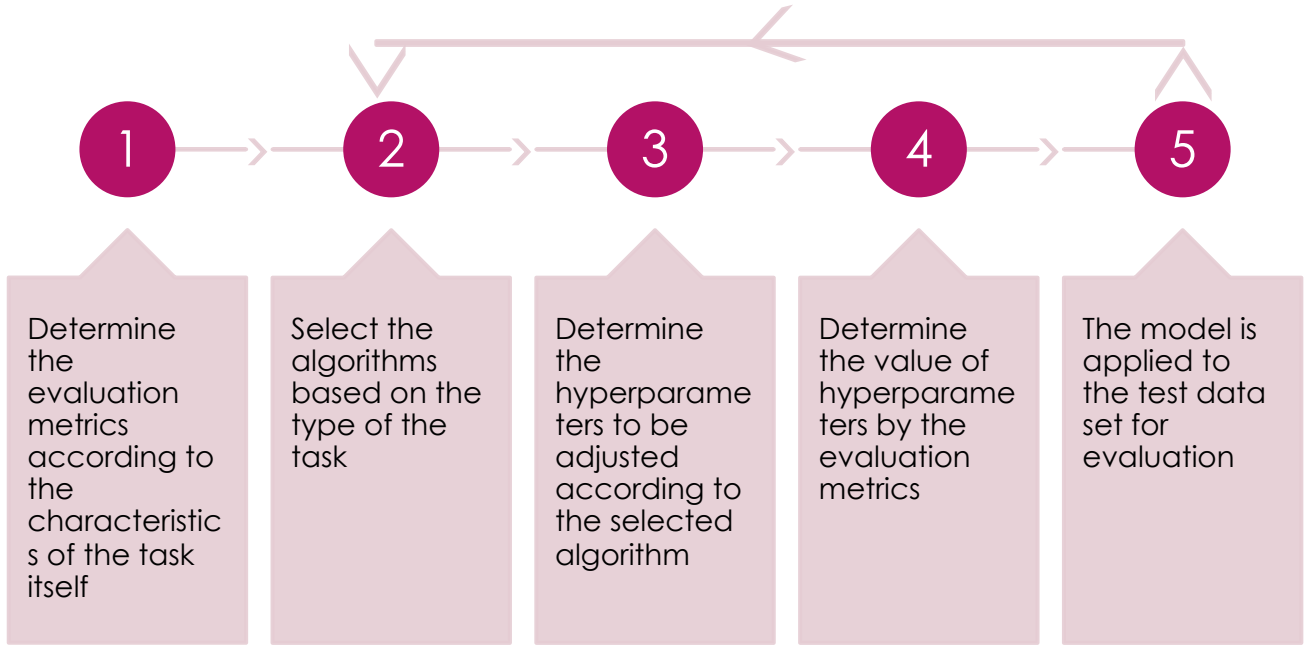
EDA

Analyze and investigate the data set and summarize its key characteristics



- Data collection
- Data Preparation
- Exploration Data Analysis
- **Modeling**
- Visualization

Modeling

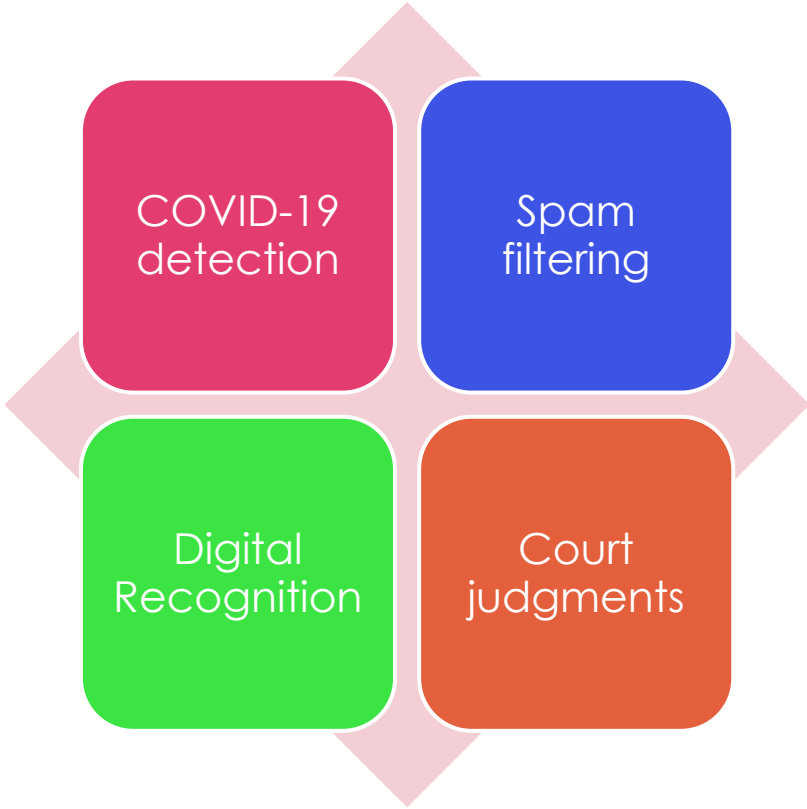


Modeling 1

		Actual Label	
		Positive	Negative
Predicted Label	Positive	**True Positive (TP)**	**False Positive (FP)**
	Negative	**False Negative (FN)**	**True Negative (TN)**

		Actual Label	
		Positive	Negative
Predicted Label	Positive	**True Positive (TP)**	**False Positive (FP)**
	Negative	**False Negative (FN)**	**True Negative (TN)**

Accuracy	(TP + TN) / (TP + TN + FP + FN)	The percentage of predictions that are correct
Precision	TP / (TP + FP)	The percentage of positive predictions that are correct
Sensitivity (Recall)	TP / (TP + FN)	The percentage of positive cases that were predicted as positive
Specificity	TN / (TN + FP)	The percentage of negative cases that were predicted as negative



Modeling 1

1

Determine the evaluation metrics according to the characteristics of the task itself

Credit Card Fraud Detection

		Actual Label	
		Positive	Negative
Predicted Label	Positive	**True Positive (TP)**	**False Positive (FP)**
	Negative	**False Negative (FN)**	**True Negative (TN)**

Accuracy	(TP + TN) / (TP + TN + FP + FN)	The percentage of predictions that are correct
Precision	TP / (TP + FP)	The percentage of positive predictions that are correct
Sensitivity (Recall)	TP / (TP + FN)	The percentage of positive cases that were predicted as positive
Specificity	TN / (TN + FP)	The percentage of negative cases that were predicted as negative

Modeling 2

2

Select the algorithms based on the type of the task

Credit Card Fraud Detection

Supervised Learning

Training data: "examples" x with "labels" y.

$$(x_1, y_1), \ldots, (x_n, y_n),\ x_i \in \mathbb{R}^d$$

- **Classification**: y is discrete. To simplify, $y \in \{-1, +1\}$

$$f\colon \mathbb{R}^d \to \{-1, +1\}$$ (f is called a **binary classifier**)

Example: Approve credit yes/no, spam/ham, banana/orange.

- Methods:
  - Logistic Regression
  - SVM
  - Neural network
  - decision tree
  - …

Modeling 3

3

Determine the hyperparameters to be adjusted according to the selected algorithm

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html?highlight=logisticregression#examples-using-sklearn-linear-model-logisticregression

Hyperparameters:

- test_size
- random_state
- penalty
- C
- max_iter
- threshold
- ...

Modeling 4

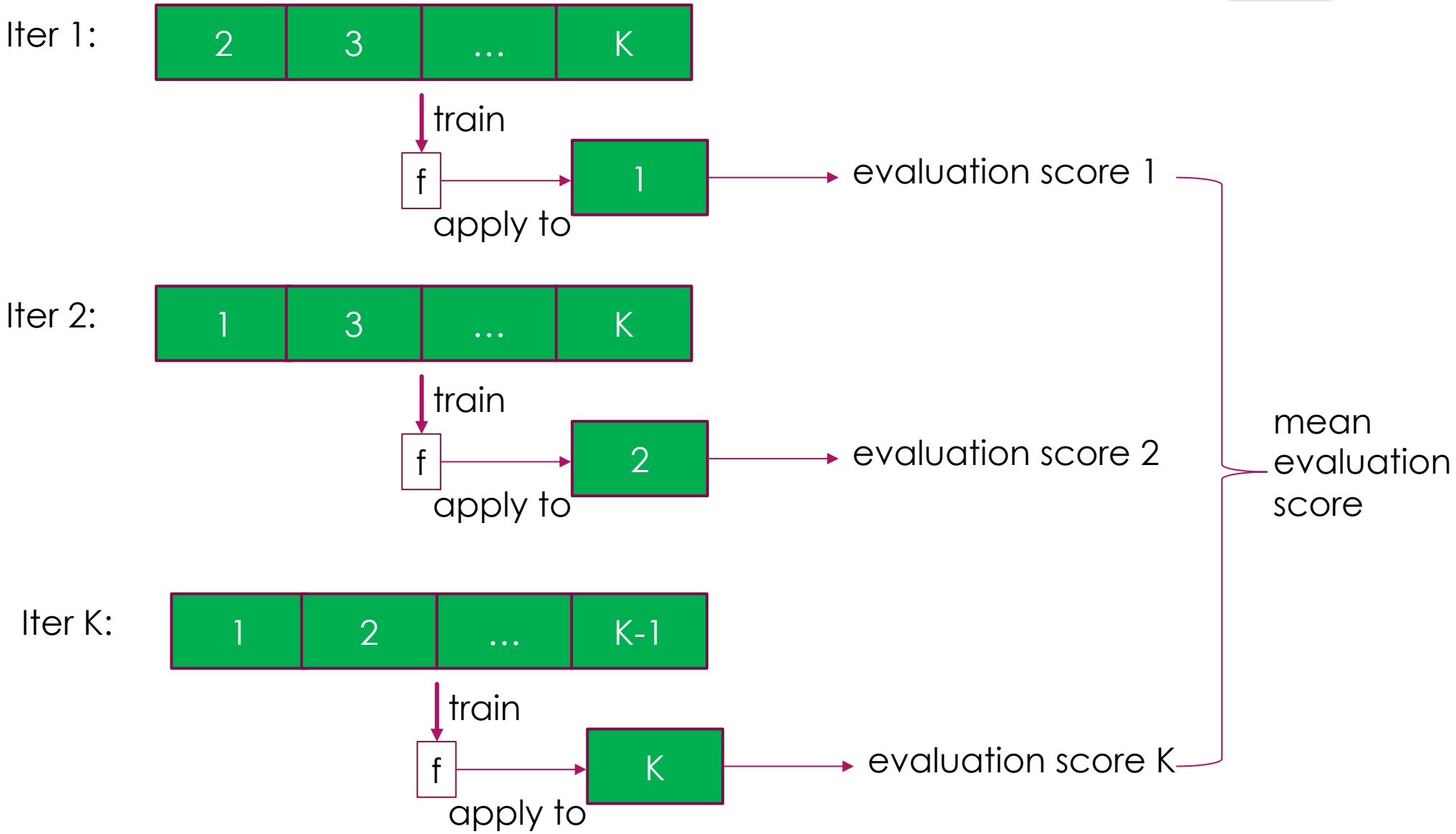
K-fold Cross Validation

4

Determine the value of hyperparameters by the evaluation metrics

Modeling 4

K-fold Cross Validation



Modeling 5

5

The model is applied to the test data set (further apply to the original test data)for evaluation

Accept
or try other method of sampling
or other ML algorithms

Visualization

- Data collection
- Data Preparation
- Exploration Data Analysis
- Modeling
- **Visualization**

Visualization methods refer to the use of visual representation to display complex resource content after the original data is converted into visual elements and to deepen the user‘s understanding. Some important visualization techniques are histograms, Scatter Plots, timelines, Box and Whisker Plots, and treemaps.