

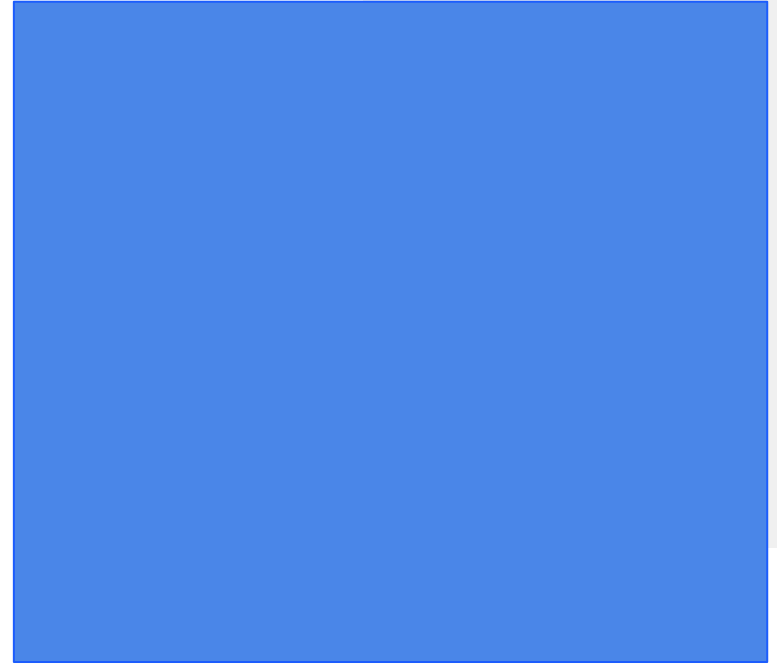


# Big Data Analytics

Dr Sirintra Vaiwsri | Email: [sirintra.v@itm.kmutnb.ac.th](mailto:sirintra.v@itm.kmutnb.ac.th)



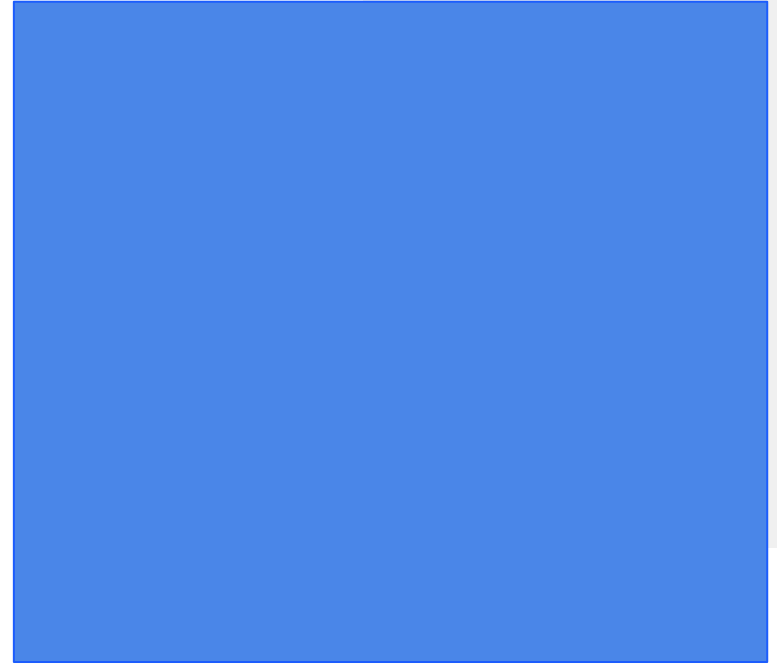
# Big Data Analytics Steps



# Big Data Analytics Steps (Dietrich et al., 2015)

- Discovery - Learn the business domain, identify a problem, and plan required resources.
- Data Collection - Consider types of data and data sources.
- Data Preparation and Storage - Pre-process data and add data into the data file/database.
- Data Processing and Analytics - Process data and analyse data.
- Visualisation - Design visualisation and visualise analysed data as tables, plots, or graphs.

# Data Analytics




# Data Analytics (Bahga and Madisetti, 2019)

- Raw data itself does not have a meaning until it is processed into information.
- Analytics is a process for creating information.
- Different goals need different technologies, algorithms, or frameworks for data analytics.
- For example,
  - To predict something from the data
  - To find patterns in data
  - To find relationships between data




# Data Analytics (Bahga and Madiseti, 2019)

- Data analytics can be categorised into 4 categories:
    - Descriptive analytics
    - Diagnostic analytics
    - Predictive analytics
    - Prescriptive analytics
- 



# Descriptive Analytics (Bahga and Madiseti, 2019)

- It analyses the past data to present the analysed data in a summarised form.
  - It answers the question “What has happened?”.
  - Example: number of reactions of all Lives on Facebook of the year 2023 and which Live gets the highest number of reactions.
  - Linear algebra, linear regression, and basic statistics functions such as counts, maximum, minimum, mean, top-N, percentage, etc. can be used for descriptive analytics.
- 

# Diagnostic Analytics (Bahga and Madiseti, 2019)

- It analyses past data to diagnose the reasons why certain events happened.
- Therefore, it answers the question of “Why did it happen?”.
- Example: From all Lives on Facebook, why does early morning or late night reach a higher number of reactions? Does work time affect the social used by users?
- Linear algebra, linear regression, graph analysis, clustering, etc. can be used for diagnostic analytics.



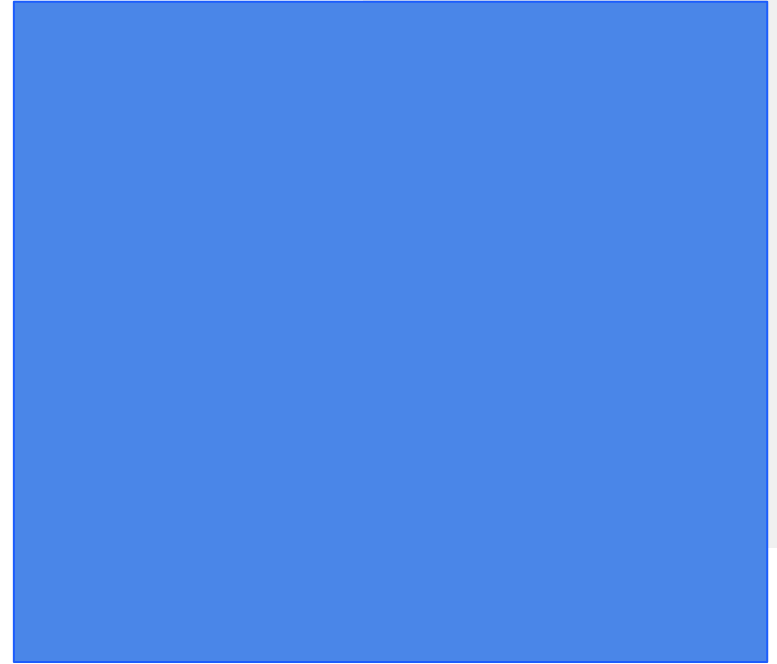
# Predictive Analytics (Bahga and Madisetti, 2019)

- It predicts the occurrence of an event or the likely outcome of an event.
- It answers the question of “What is likely to happen?”.
- The accuracy of prediction depends upon the quality of existing data.
- Example: What Facebook Live content will attract more views, thus, number of reactions? Will the number of reactions increase if you Live on Facebook in the early morning or late night?
- Linear algebra, linear regression, clustering, graph analysis, Bayesian inference, Markov Chain Monte Carlo, text analysis, Hidden Markov Model, etc. can be used for predictive analytics.

# Prescriptive Analytics (Bahga and Madisetti, 2019)

- It uses multiple prediction models to predict various outcomes and the best action for each outcome.
- It answers the question of “What can we do to make it happen?”.
- It predicts the outcomes based on the current actions.
- Example: Should you live on TikTok or YouTube to get more number of views? Does a photo or video attract more views in the early morning? What content type should you post in the early morning to get more views from users with the age above 20?
- Clustering, graph analysis, optimisation, Bayesian inference, Markov Chain Monte Carlo, text analysis, Hidden Markov Model, etc. can be used for prescriptive analytics.

# Machine Learning



# Big Data Analytics and Machine Learning

(Guller, 2015; Russell, 2018)


- As Big Data works with a vast amount of data (volume) where the accuracy (variety and veracity) and speed (velocity) are of concern for creating information (value), machine learning can be applied to Big Data Analytics.
- Machine Learning (ML) is a trained program to learn from data and then does some actions.
- ML is used when:
  - Complex problems need to be solved
  - Non-stable system where new data can always be ingested into the system
  - Improving performance for solving problems that require various rules to find the solution.

# Machine Learning Terminology (Guller, 2015)

- **Feature** (independent variable) represents a property of an observation. For example, a row represents an observation while their columns represent different features (different independent variables).
  - A category feature is a descriptive feature (qualitative), for example, a name.
  - A numerical feature is a numerical feature (quantitative), for example, a number of reactions on FBLiveTH.
- **Label** (dependent variable) is a variable that an ML system learns to predict.
  - A categorical label represents a category, for example, an ML application learns to predict a category of news articles such as business and technology categories. These categories are categorical labels.
  - A numerical label represents a numerical, for example, an ML application learns to predict the price of a house. The price is a numerical label.



# Machine Learning Terminology (Guller, 2015)

- **Model** estimates the relationship between the dependent variable (label) and independent variable (feature) in a dataset.
    - A model can predict the value of a dependent variable (label) using independent variables (features). For example, using the number of reactions, status published, and status type in FBLiveTH to predict a future number of comments.
    - In other words, we can say that a model is a mathematical function that uses features as input and outputs a value (label).
- 

# Machine Learning Terminology (Guller, 2015)

- **Training Data**, also called “training set”, is the data that is used for training a model.
- Training data can be categorised into:
  - Labelled dataset - a dataset has at least one of the columns that contains a label.
  - Unlabeled dataset - a dataset does not have a column that contains a label.
- **Test Data**, also called “test set”, is a dataset that is used for testing a model after the model has been trained for evaluating the performance of the model.

# Supervised and Unsupervised Learning Algorithms

(Chambers and Zaharia, 2018; Guller, 2015)

- A supervised learning algorithm uses a labelled dataset to train a model, where these labels in a training dataset may be generated manually or sourced from other datasets.
- A supervised learning algorithm aims to predict a label for a data.
- An unsupervised learning algorithm uses an unlabeled dataset to train a model.
- An unsupervised learning algorithm aims to discover a structure in data.

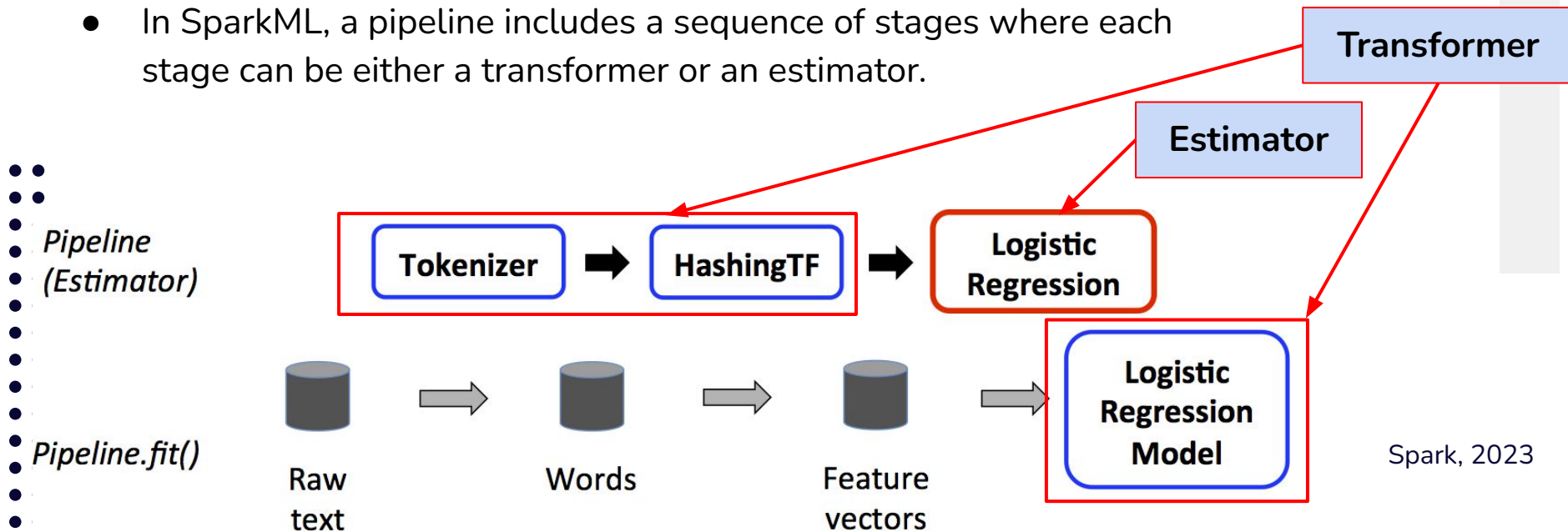


# SparkML (Chambers and Zaharia, 2018; Spark, 2023)

- DataFrame from SparkSQL is used as an ML dataset.
- The transformer includes feature transformers and learned models.
  - It inputs a DataFrame, then transforms (*transform()*) the input to output which is another DataFrame.
  - Transformer examples are such as cutting some features, adding more features, manipulating the current features, or formatting data.
- The estimator is an algorithm or learning algorithm that is used to fit (*fit()*) data into the machine learning model (transformer).

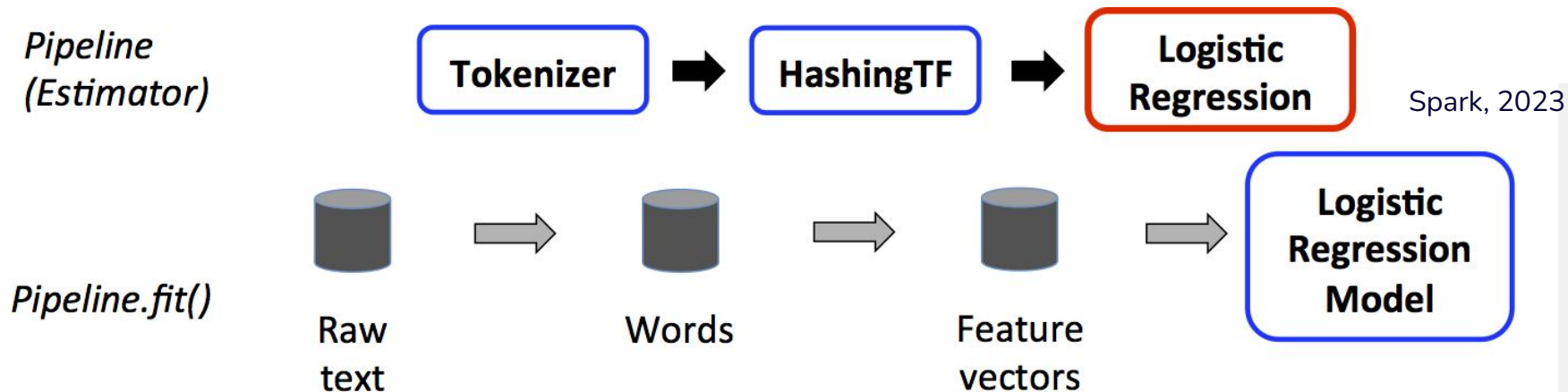
# Pipeline (Chambers and Zaharia, 2018; Spark, 2023)

- A pipeline in machine learning is a sequence of algorithms for processing and learning from data.
- In SparkML, a pipeline includes a sequence of stages where each stage can be either a transformer or an estimator.



# Training Example (Spark, 2023)

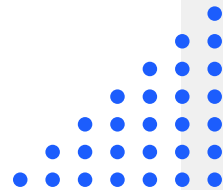
The pipeline calls `LogisticRegression.fit()` to create a Logistic Regression Model which is a transformer.



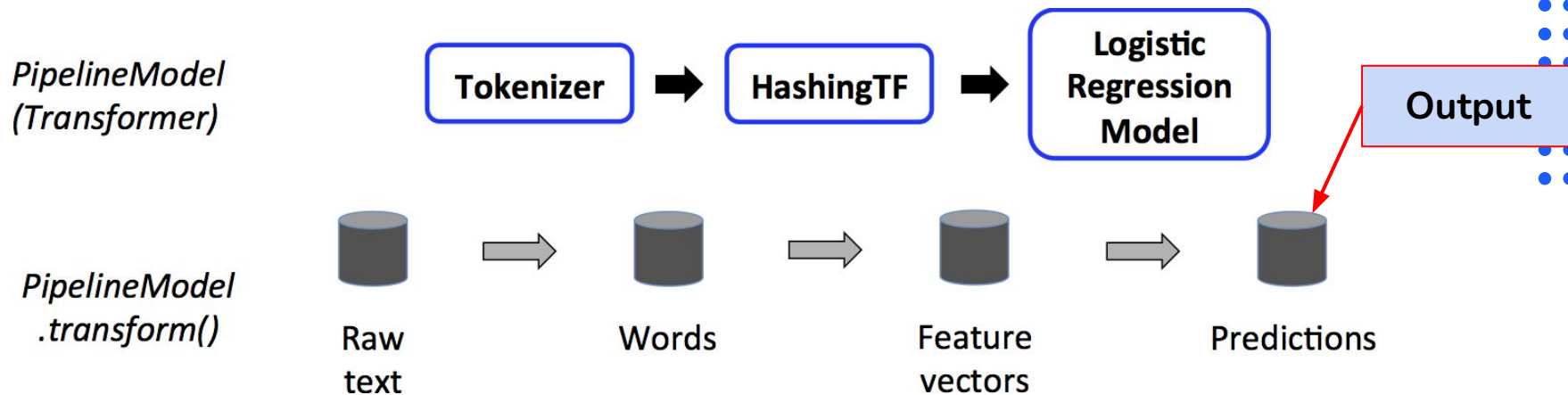
`Pipeline.fit()` is called to create a `PipelineModel`

Raw text is tokenized to words using `Tokenizer.transform()` and adding a new column for storing words.

Words in the column are converted to feature vectors using `HashingTF.transform()` and add a new column for storing feature vectors.



# Testing Example (Spark, 2023)



Spark, 2023

PipelineModel is used for testing the model

All Estimators become Transformers



# References

- Dietrich, D., Heller, B., & Yang, B. (2015). *Data science & big data analytics: discovering, analyzing, visualizing and presenting data*. Wiley.
  - Bahga, A., & Madisetti, V. (2019). *Big Data analytics: A hands-on approach*.
  - Guller, M. (2015). *Big data analytics with spark*.
  - Russell, R. (2018). *Machine Learning: Step-by-step guide to implement machine learning algorithms with python*. (Knxb).
  - Chambers, B., & Zaharia, M. (2018). *Spark: The definitive guide: Big data processing made simple*. "O'Reilly Media, Inc."
  - Spark. <https://spark.apache.org>. Accessed: 2023-09-05.
- 