DLI Accelerated Data Science Teaching Kit

# Lecture 18.3 - Learning Process

# Learning on Streaming Data

Learning from streaming data requires real-time (or near real-time) updates to a model with these challenges:

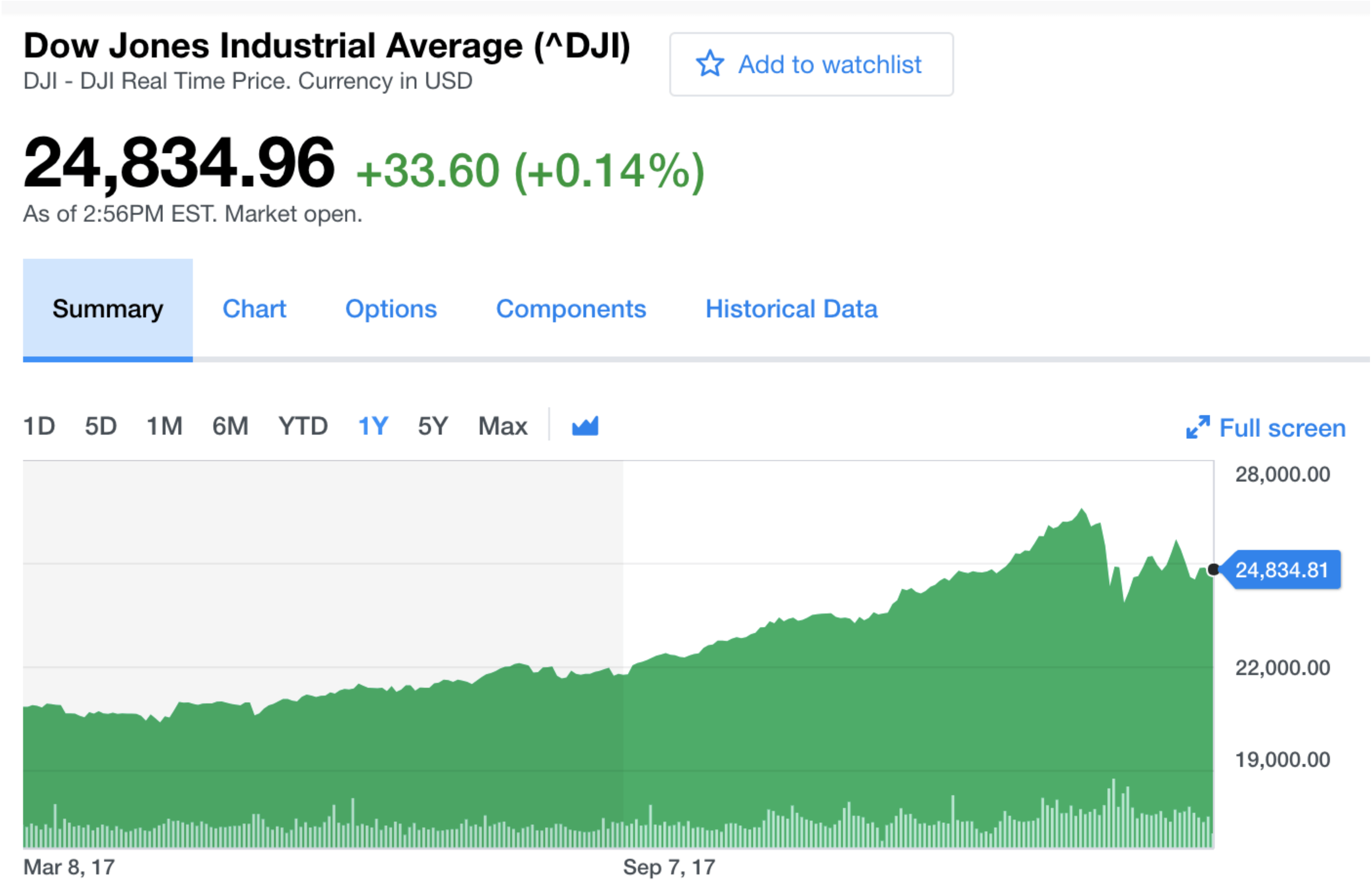
- Building the relationship between data streams and time series;
- Addressing the problem of dealing with partially and delayed labels;
- Learning on imbalanced data streams;
- Detecting anomalies from streaming data

# Time Series

Time series data may commonly arrive in the real-time manner, and it thus can be treated as a data stream.

Data streams may often involve temporal dependence and thus be considered as time series.

Unlike a regular data stream, where instances are assumed to be independently and identically distributed (i.i.d.), data points in a time series are expected to exhibit strong temporal dependence.



Dow Jones Industrial Average (^DJI)

DJI - DJI Real Time Price. Currency in USD

24,834.96 +33.60 (+0.14%)

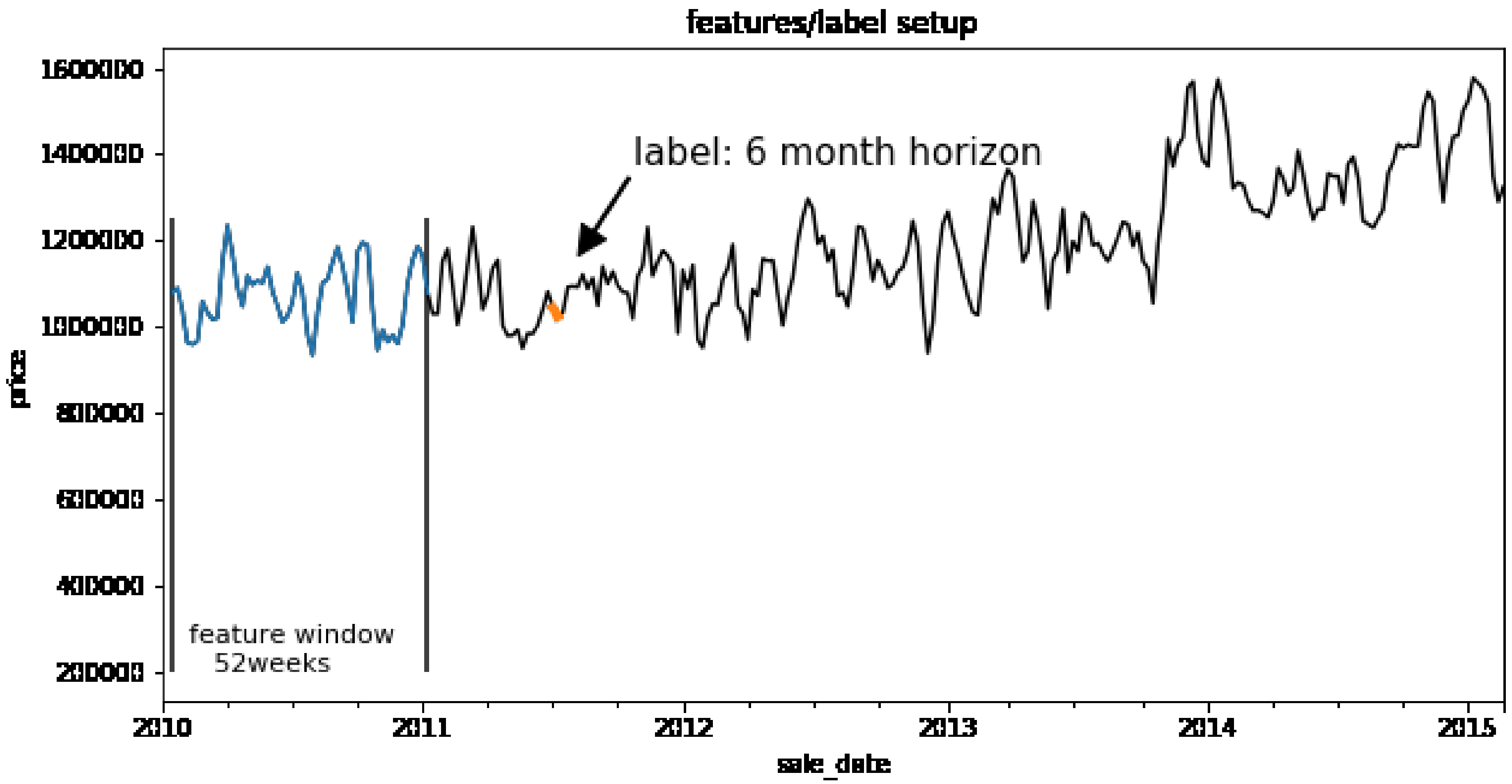
As of 2:56PM EST. Market open.

Source: https://www.influxdata.com/what-is-time-series-data/

# Time Series

If $P(y_t \,|x_t \,, x_{t-1}\,) \; = \; P(y_t \,|x_t)$ does not hold, it indicates temporal dependence in the data stream.

The idea is to produce a new stream of instances $x_t' \coloneqq [x_t, x_{t-1}, \dots, x_{t-w}]$ over a window of size ***w***, sufficient such that $P(y_t \,|x_{\prime t}) \; = P(y_t \,|x'_t \,, x'_{t-1}\,)$; thus, producing a temporally-independent (i.e., ‘regular’) data stream.



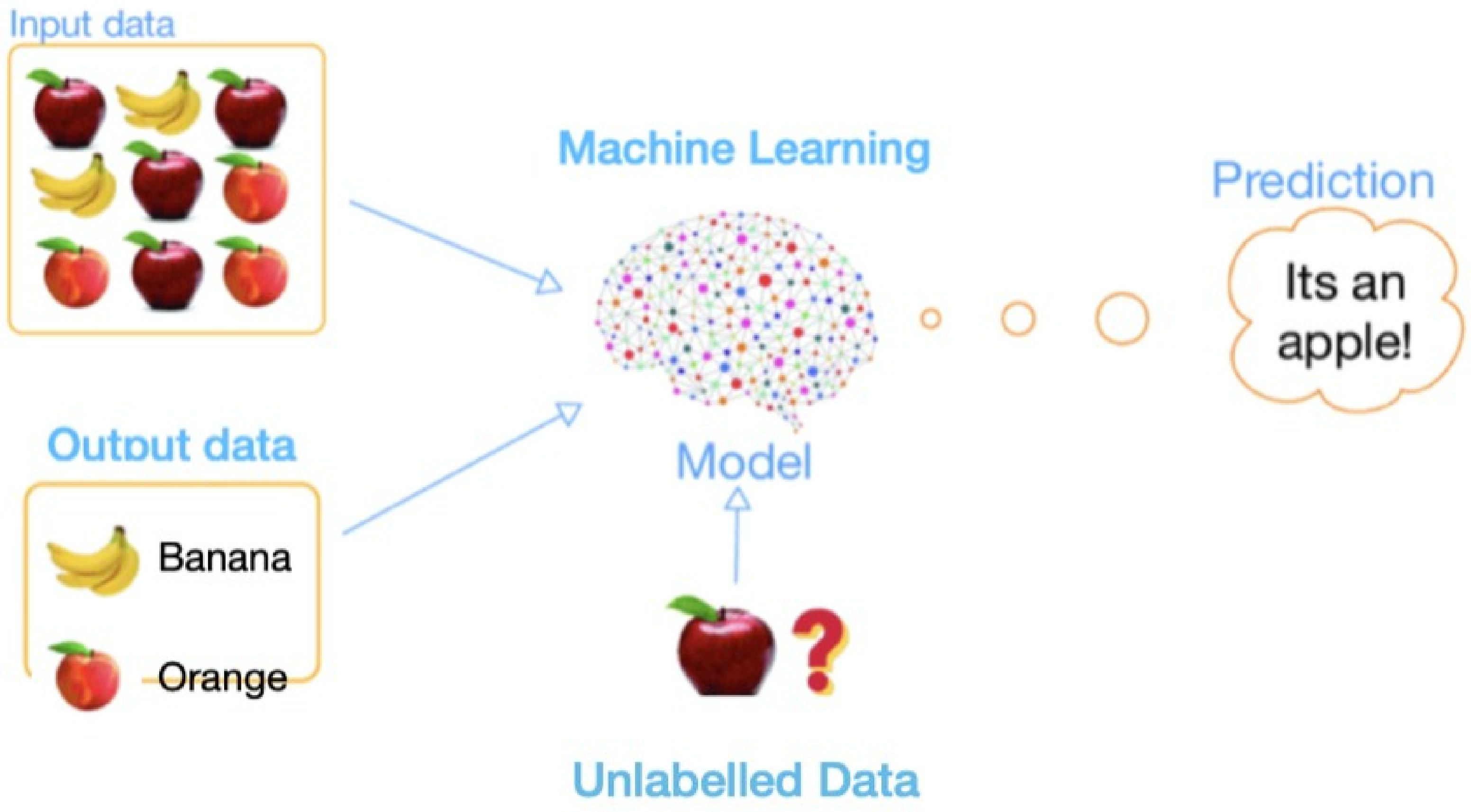
Source: https://cloud.google.com/blog/products/ai-machine-learning/how-t quickly-solve-machine-learning-forecasting-problems-using-pandas-and-bigquery

# Semi-supervised Learning

Semi-supervised learning (SSL) is particularly relevant to streaming applications where data are abundant but labeled data may be rare.

The SSL techniques for streaming data includes unsupervised learning combined with supervised learning; It is hybrid approach.

- Each of these approaches makes assumptions about the problem and the data distribution, but not very often these assumptions are explicitly discussed.



**Source: https://www.enjoyalgorithms.com/blogs/supervised-unsupervise- semisupervised-learning**

# Ensemble Learning

Ensemble learning receives much attention for data stream learning as ensembles can be integrated with drift detection algorithms and incorporate dynamic updates.

Ensemble models can rely on the reactive strategy to cope with concept drift that continuously updates the ensemble, often assigning different weights to base models according to their prediction performance.

The challenge is how to maintain the characteristics of the ensemble methods and efficiently train them over several machines.

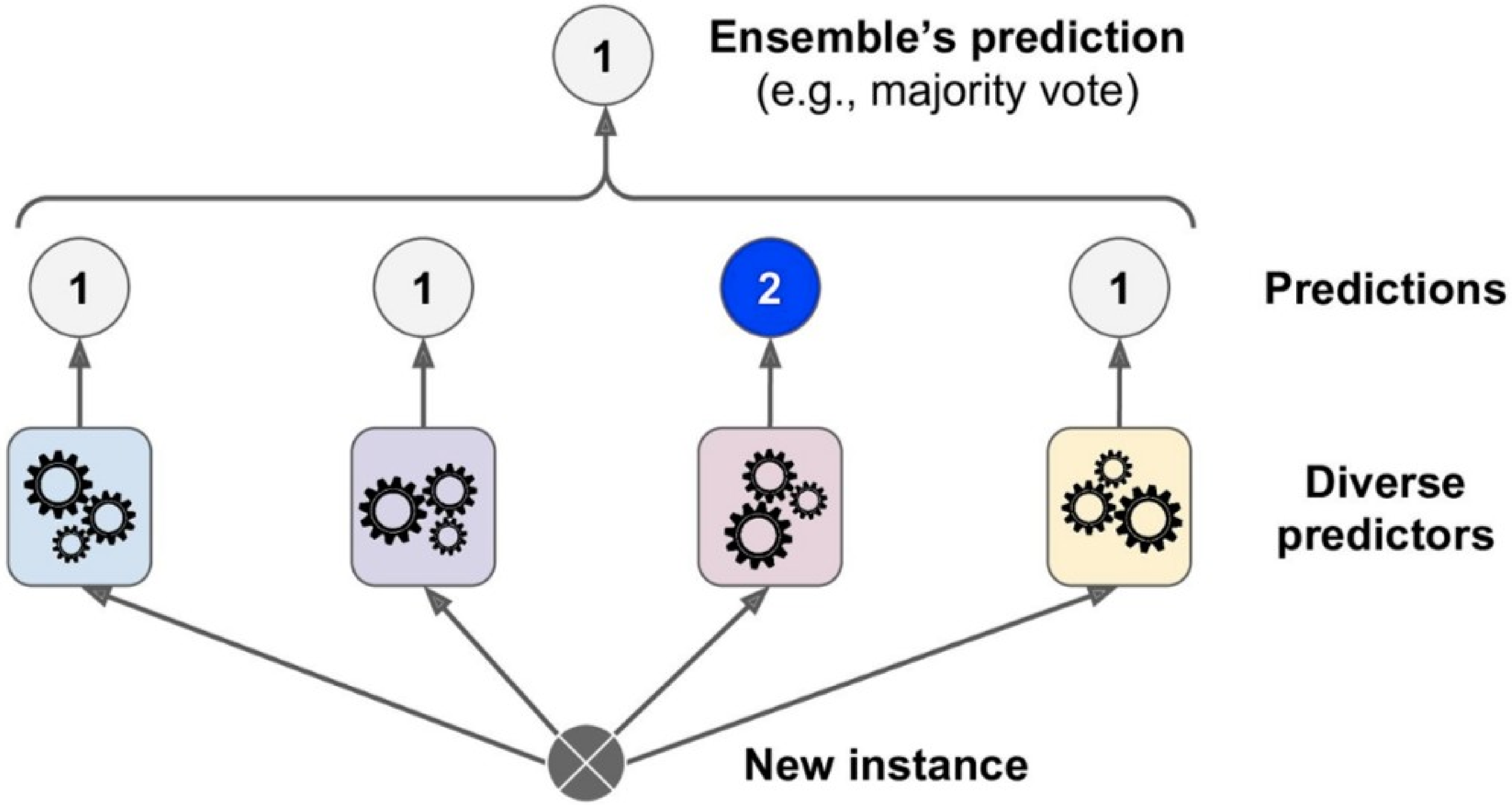


*Figure 7-2. Hard voting classifier predictions*

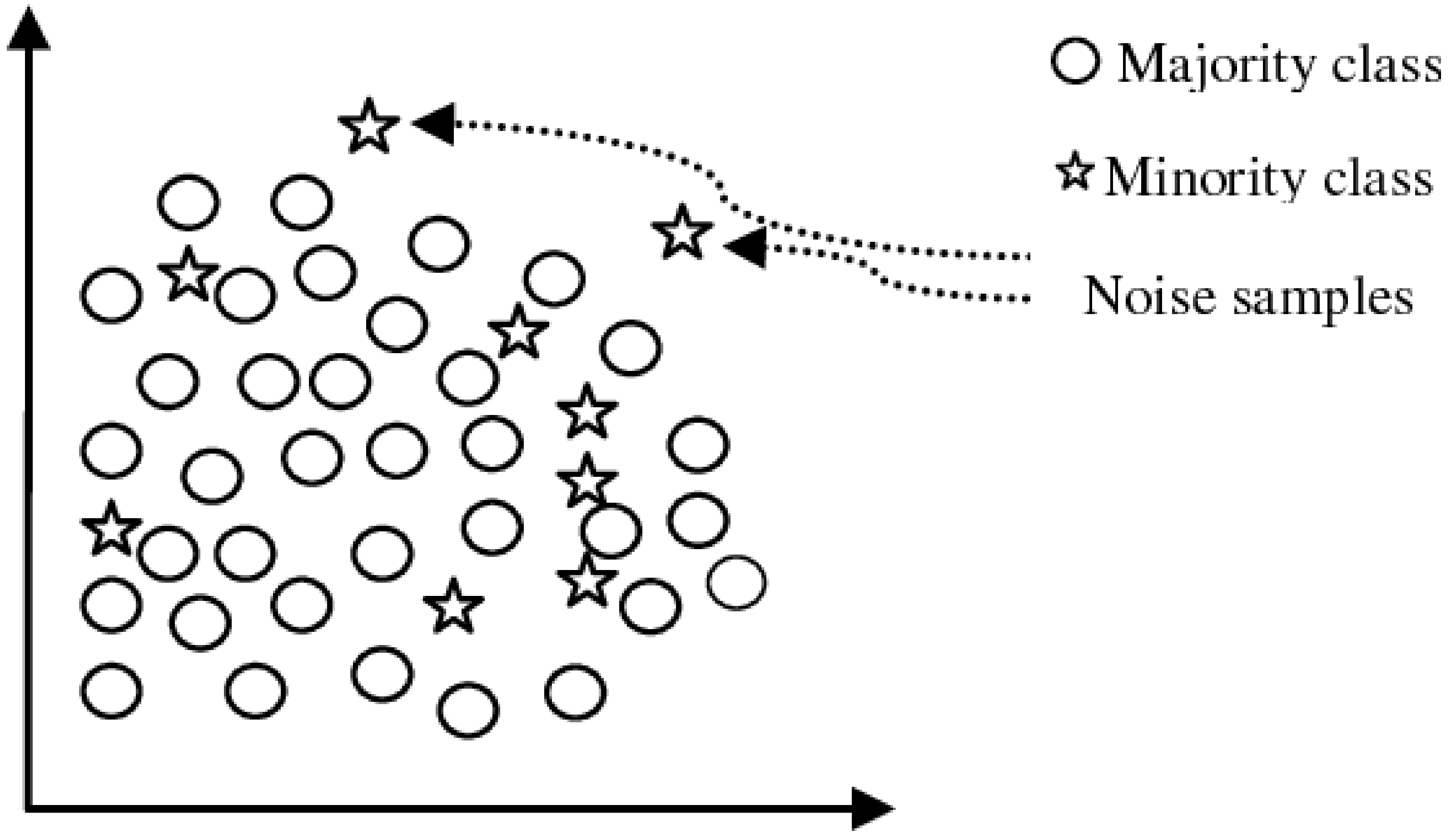
**Source: https://www.kdnuggets.com/2019/01/ensemble-learning-5 approaches.html**

# Imbalanced Learning

Imbalanced datasets are characterized by one class outnumbering the instances of the other one. The latter is referred to as the minority class, while the former is identified as the majority class.

The imbalance may be inherent to the problem (intrinsic) or caused by some fault in the data acquisition (extrinsic).

Learning from imbalanced datasets is challenging as most learning algorithms are designed to optimize for generalization, and as a consequence, the minority class may be completely ignored.



**Source: https://www.semanticscholar.org/paper/A-Review-on-Class-Imba Problem%3A-Analysis-and-Maheshwari-Jain/ 70b83c369dce441b765bcf5a8f16550c7e87b53f**

DLI Accelerated Data Science Teaching Kit

# Thank You