University of Roehampton London

Machine Learning

Dr Changjiang He, Dr Kuo-Ming Chao
Computer Science| School of Art
University of Roehampton

University of Roehampton London

Lesson 8.2

Self-Training & Co-Training

Self-Training

University of Roehampton London

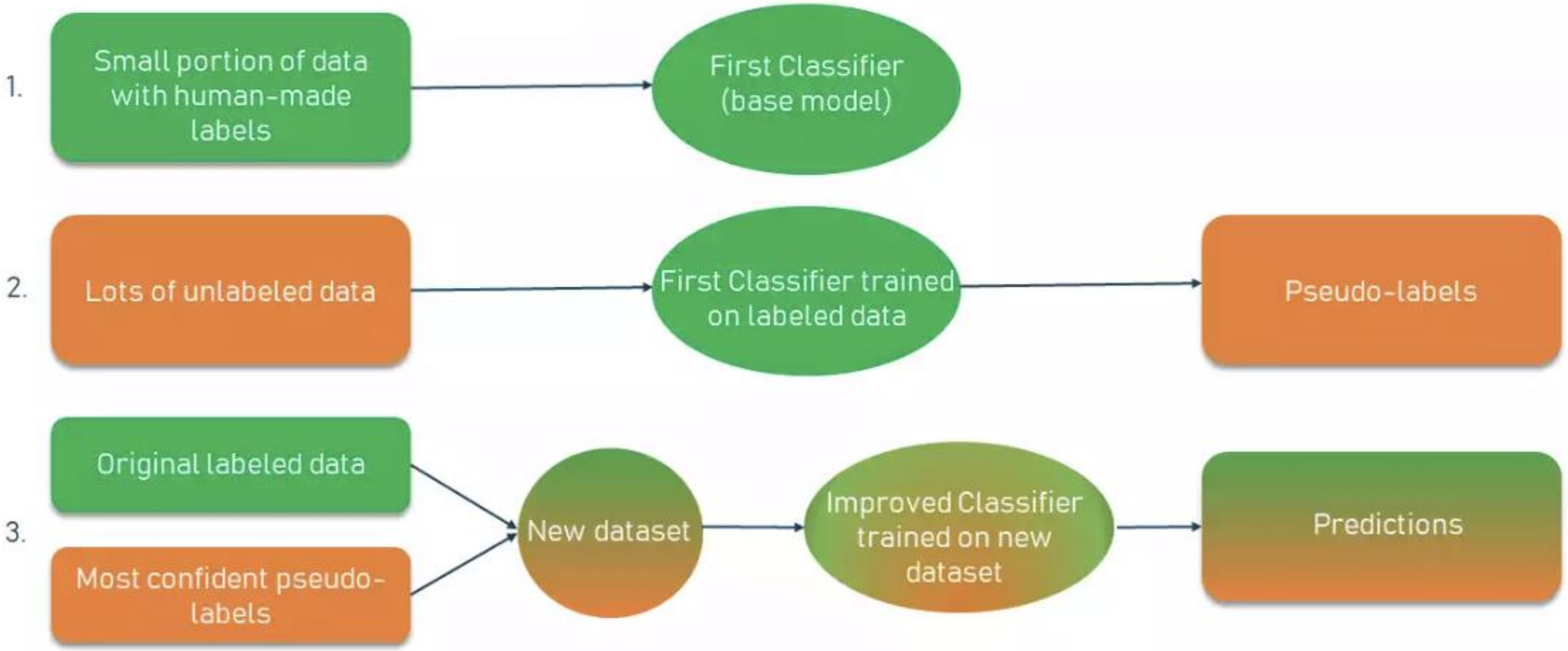
- Instead of adding tags to the entire dataset, we go through and hand-label just a small part of the data and use it to train a model, which then is applied to the ocean of unlabelled data.

- One of the simplest examples of semi-supervised learning, in general, is self-training.

- **Self-training** is the procedure in which you can take any supervised method for classification or regression and modify it to work in a semi-supervised manner, taking advantage of labelled and unlabelled data.

Self-training

University of Roehampton London



Self-training Process

University of Roehampton London

- Pick a small amount of labelled data.
- Apply the process known as *pseudo-labelling*.
- From this point, we take the most confident predictions made with the model.
- The process can go through several iterations (10 is often a standard amount) with more and more pseudo-labels being added every time.

Co-training

University of Roehampton London

- Derived from the self-training approach and being its improved version, co-training is another semi-supervised learning technique used when only a small portion of labelled data is available.

- Unlike the typical process, co-training trains two individual classifiers based on two *views* of data.

Co-training

University of Roehampton London

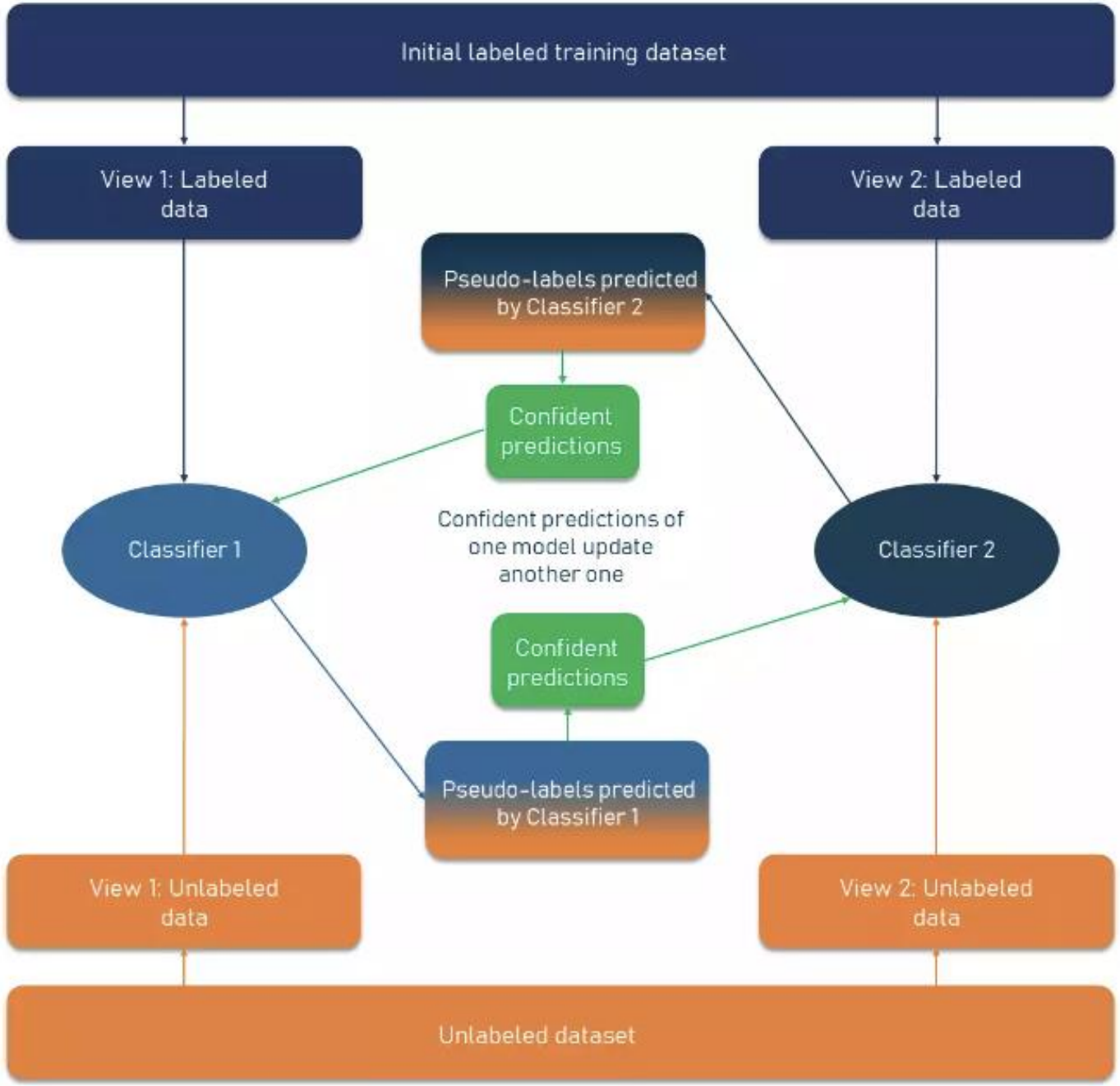
- The views are basically different sets of features that provide additional information about each instance, meaning they are independent given the class.

- Also, each view is sufficient — the class of sample data can be accurately predicted from each set of features alone.

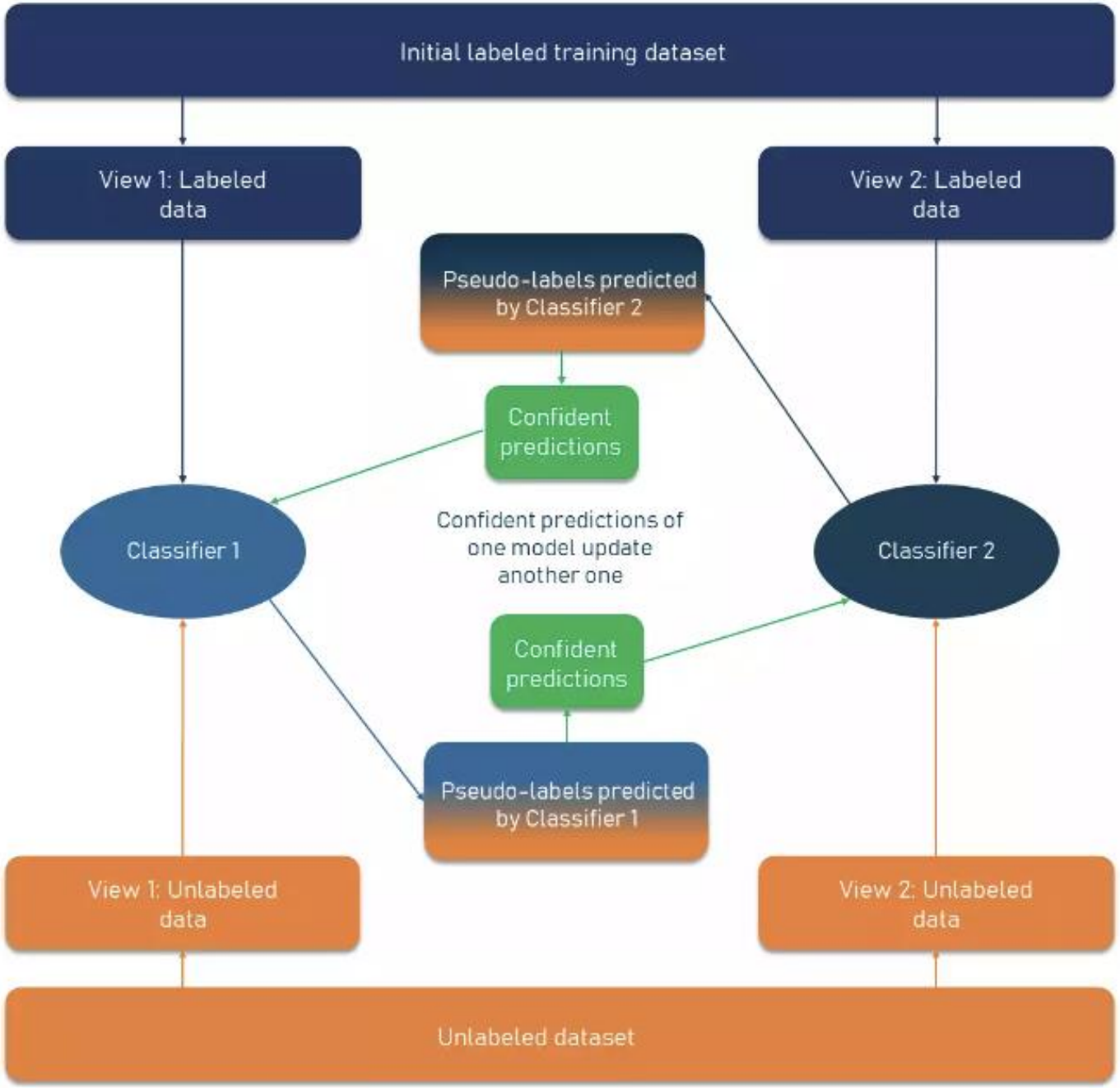
• Train a separate classifier (model) for each view with the help of a small amount of labelled data.

•Then the bigger pool of unlabelled data is added to receive pseudo-labels.

•Classifiers co-train one another using pseudo-labels with the highest confidence level.



•If the first classifier confidently predicts the genuine label for a data sample while the other one makes a prediction error, then the data with the confident pseudo-labels assigned by the first classifier updates the second classifier and vice-versa.



- The final step involves the combining of the predictions from the two updated classifiers to get one classification result.
- As with self-training, co-training goes through many iterations to construct an additional training labelled dataset from the vast amounts of unlabelled data.