Abstract
We investigate the application of machine learning methods to find statistical arbitrage opportunities in the stock market using pair trading strategy. Pairs are recognized using clustering methods, while trading signals are predicted by multiple supervised learning algorithms.

Motivation
The key to successful pairs trading is the ability to detect patterns in spreads and correctly identify when a spread is likely to converge back to its mean. Sophisticated machine learning techniques can be used at every step of the pairs trading process.

Methods
PCA: Feature dimension reduction and component generation, preparing for clustering.
DBSCAN: Creates clusters and identify points that are not part of any cluster.
t-SNE: Method to visualize clusters from high dimension to 2-D space.
Gradient Boosting: A sequential ensemble model to capture complex patterns.
Random Forest: A bagging decision tree to reduce bias.
LSTM: Apply more weight to recent observations in time series prediction. Comparing with the standard RNN, LSTM diminishes the problems of long-term dependencies.

Clustering: Our approach to pairs trading is to apply PCA for dimension reduction on a large set of features in order to ease computation of DBSCAN filter for clustering stocks. Afterwards, we use cointegration test to extract all possible combinations of stocks in each cluster that are within 5% significance level.

Results
We gained experience on training neural networks, dimensionality reduction, and supervised learning on time-series data.

What we have learned
We gained experience on training neural networks, dimensionality reduction, and supervised learning on time-series data.

Conclusion
Pair trading is still a feasible trading strategy and machine learning can improve its performance.