**Industry problem recap**

Financial compliance regulations necessitate organizations to monitor all communication channels used by their regulated employees. Demonstrating awareness of any unlawful actions by traders requires thorough examination of all sent and received communications, promptly identifying any signs of malicious behaviour. The sheer volume of communication data generates numerous alerts, many of which turn out to be false alarms. By implementing machine learning (ML), the number of false positives (innocent communications causing unnecessary work for reviewers) can be reduced, and false negatives (potential regulatory breaches leading to substantial fines and sanctions) can be minimized. Communications are scored using complex lexicon search criteria and this score is passed as an input to another system to raise the appropriate alerts.

**What information would you need in order to make the most accurate predictions?**

To make the most accurate predictions in the context of financial compliance monitoring using machine learning, the following information would be crucial:

* Frequency of alerts
* Frequency of communication
* Surrounding communications
* Parties involved in the conversation
* The means of communication, i.e. the channel through which they were communicating
* HR records
* Location of the sender and recipient
* The sentiment and emotion of the message body
* The time and day of the message
* The baseline emotion of the sender
* The context / topic of conversation
* Position / role of the sender and recipient within the organisation
* Market news and alerts
* Trade and market data

**How would you aggregate the information?**

The data listed above would have to be aggregated differently depending on how it is gathered and measured. Although it is tough and possibly incorrect to state how the data *would* be aggregated in a functional model. I will list here some suggested techniques by which the data *could* be aggregated.

1. Personnel aggregation – The people involved in a given communication and alert would be the most obvious, basic separation to make.
2. Modal aggregation – Categorisation of communications and alerts based on the mode (most popular) value for some of the data gathered above may be more useful than including all the details in the dataset. For example, emotion and sentiment can change in a single message; noting only the most common emotion and sentiment throughout and categorising on this may reduce noice and variance.
3. Spatial Aggregation – The location of the sender and recipient could be used to define different regions.
4. Time-based aggregation – The time of day/day of week of the communication could be used to separate the communications into time-based, ordinal, categories.
5. Text data aggregation – A technique such as TD-IDF (term frequency-inverse document frequency) could be used to analyse the content of the messages. This measures the importance of words in a document relative to their frequency in a corpus.
6. Hierarchical aggregation – As mentioned, HR data (including corporate hierarchies) could be gathered and used to categorise the communications by team, department, or level.

**Of the machine learning techniques, you have learnt so far, which approaches would you try?**

Given the nature of how the alerts are raised given a communication (see the scoring discussed in the first paragraph), a regression model could yield good results. This would hopefully give an interpretable model and help highlight feature importance. The predictive power of regression models is generally high, which is very important when trying to predict future alerts. Some examples of regression models could include:

* Linear regression
* Decision trees
* Random forests

KNN (K-Nearest Neighbors) could give a model with higher predictive accuracy with the downside of having less interpretability.

Time series analysis would be applicable in this case due to the communications being gathered over a period. The model should be able to identify patterns, trends and irregularities within the data which may not be obvious. Time series models, like ARIMA (autoregressive integrated moving average) and LSTM (long short-term memory), could be employed to forecast future scores and alerts.

Deep learning models, like neural networks, could work well due to the complex patterns contained in the body of the communications. These models excel in learning complex patterns allowing for high predictive accuracy, exactly meeting our goals.