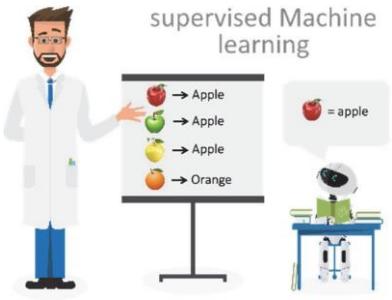
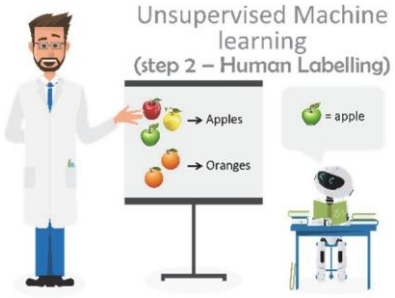
Machine Learning (ML) uses techniques that learn to make predictions or decisions from data without explicit programming. The ability for "*ML software to learn from real-world feedback (training) and improve its performance (adaptation) makes these technologies uniquely situated among software*"[11]. There are different types of ML including supervised and unsupervised learning.

- **Supervised learning** tools are initially fed with training data (i.e. the input / question) and their interpretation (i.e. the output / answer). The algorithm can learn from human experience through "training data" and make future decisions and predictions (i.e. answer similar questions from new data). Once the system is trained, its accuracy can be verified using test data (i.e. data without the answer to the problem). Training data can be adjusted as necessary until the desired learning is achieved. Overall, the easier the problem is to solve, the less training data you need. As an example, consider decisions on when to raise a query. Most cases are simple such as when data is out of range, then raise a query. So, you can train the system to do so with limited training data. From our CDM experience most scenarios are not as simple. If data is out of range at a visit, you may want to check if the same issue was queried at a previous visit. If it was, did the site confirm the data "as is" due to a patient's medical condition? If yes, you would not query again. You would likely close the query and enter a comment. This behavior could be trained too but would require more training data and time.



- **Unsupervised learning** tools are fed with data without their interpretation. In that case, the algorithm can only determine which data are more similar to another. Unsupervised learning is trying to understand how things work. The system can learn through observations (not through experience). In the scenario of query generation, CDM could feed the system with query history without expressing the ideal behaviors. The system would process all the information to determine the different query handling scenarios. The system would require enough cases (i.e. data) to correctly learn by itself. Scenarios would then be confirmed and labelled by humans (e.g. one scenario being about out of range data).



In simplistic terms, a human can train a system to recognize apples and oranges using supervised learning. When unsupervised, a machine can learn by itself to differentiate apples from oranges without knowing what an apple or an orange is. From here, a human has to "label" what classified group defines oranges and which one defines apples to complete the learning process. The unsupervised process is ultimately easier to implement as it does not require the pre-definition of training data which could be complex in Clinical Research. Today, a robust form of unsupervised learning, called "deep learning" can successfully classify patterns. As an example, deep learning can robustly analyze images to make medical diagnosis (e.g. Classify images with vs. without disease characteristics such as nodules).

Without a doubt, unsupervised learning systems especially those using deep learning, have proven their ability to make sound predictions. Unfortunately, they cannot yet provide the reasoning behind their predictions. This is making their use for regulated activities challenging. These techniques need careful review of the classification generated by the algorithm. It is however expected that learning algorithms will be able to explain reasoning back to humans in an understandable way in the very near future. At that stage, CDM will then be able to consider more regulated scenarios where humans would only verify the prediction(s) determined by machines.

Lastly, it is very important to understand that ML is designed to learn and adapt. Like humans, ML Tools will make decisions tomorrow that are better than the one they made today. So, CDM needs to carefully anticipate the implication of evolving systems in our regulated environment when reproducibility of results is expected. This has prompted FDA "*to reimagine an approach to premarket review for AI/ML-driven software modifications"* [11]. While focused on Software as Medical Device (SaMD), this FDA discussion paper highlights the need for new approaches with AI. We also need to realize that ML learning could be biased by the training datasets or if human supervisors are 'just' following the decisions suggested by the algorithm which would re-enforce the biased behavior. CDM will need to leverage its risk-based data management strategy when implementing such disruptive technology to ensure reliable and ethical (i.e. unbiased) decision making.