Life Sciences and
Facility Management

Institute of
Computational Life Sciences

Active learning

Machine Learning

Norman Juchler

Learning objectives

- Know what active learning is and how it works.
- Understand how to use active learning to reduce the number of required labels?

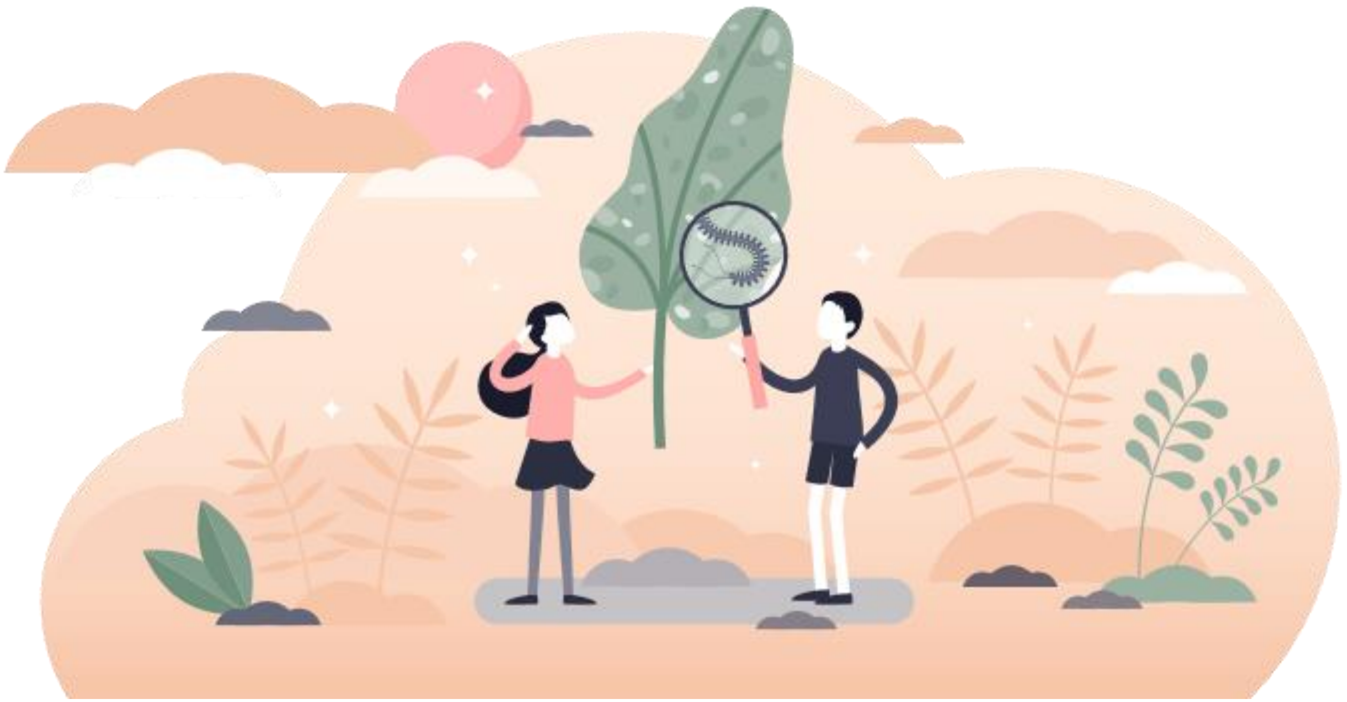
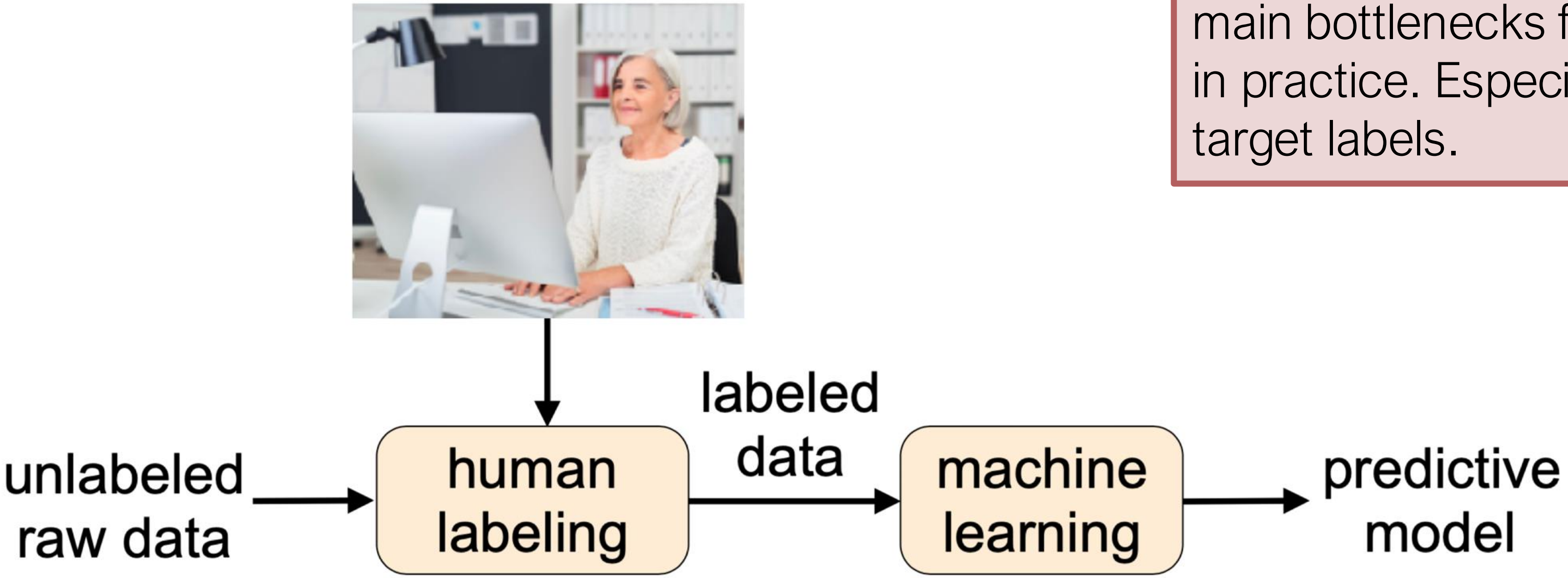


Image source: Adobe stock (577250168)

The conventional ML process



Producing training data is one of the main bottlenecks for ML applications in practice. Especially valuable are target labels.

The value of training data

- Depending on the task and ML method, thousands, if not millions, of samples are required to train a model.
- Company experts (e.g., lawyers, medical doctors) are expensive
- ...and often unavailable

Goal: Can we train models with less training data and less human supervision?

The collection of data has a value. A selection of companies offering data collection services. Source: [Link]

The active learning process

Idea: Use a machine ("active learner") to automatically and adaptively select most informative data for labeling

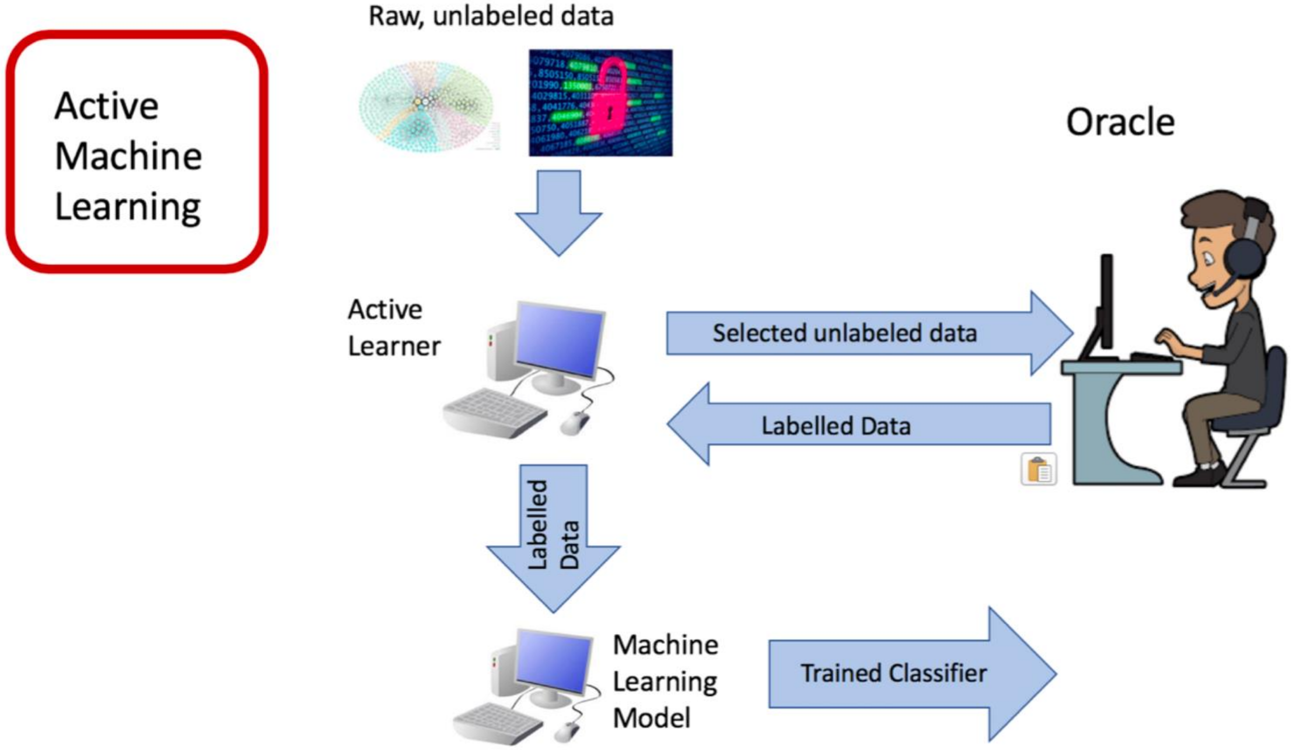


Illustration of active learning. The use of machine learning during a data labeling process. Image source: Link

Core ideas of active learning

- **Iterative process**:
  The active learner is trained on an initial labeled dataset, then iteratively selects and requests labels for the most uncertain or impactful samples.

- **Uncertainty sampling**:
  The model identifies data points where it has the least confidence in its predictions (e.g., probabilities close to 0.5 in binary classification).

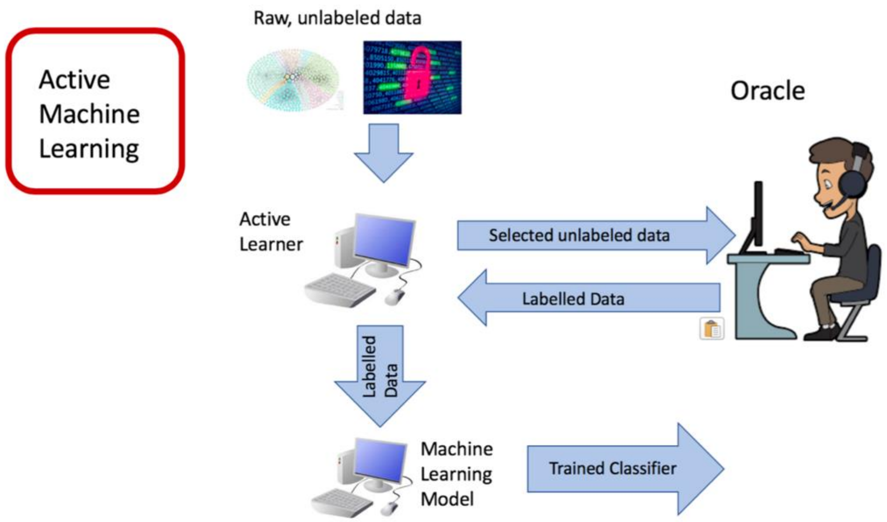
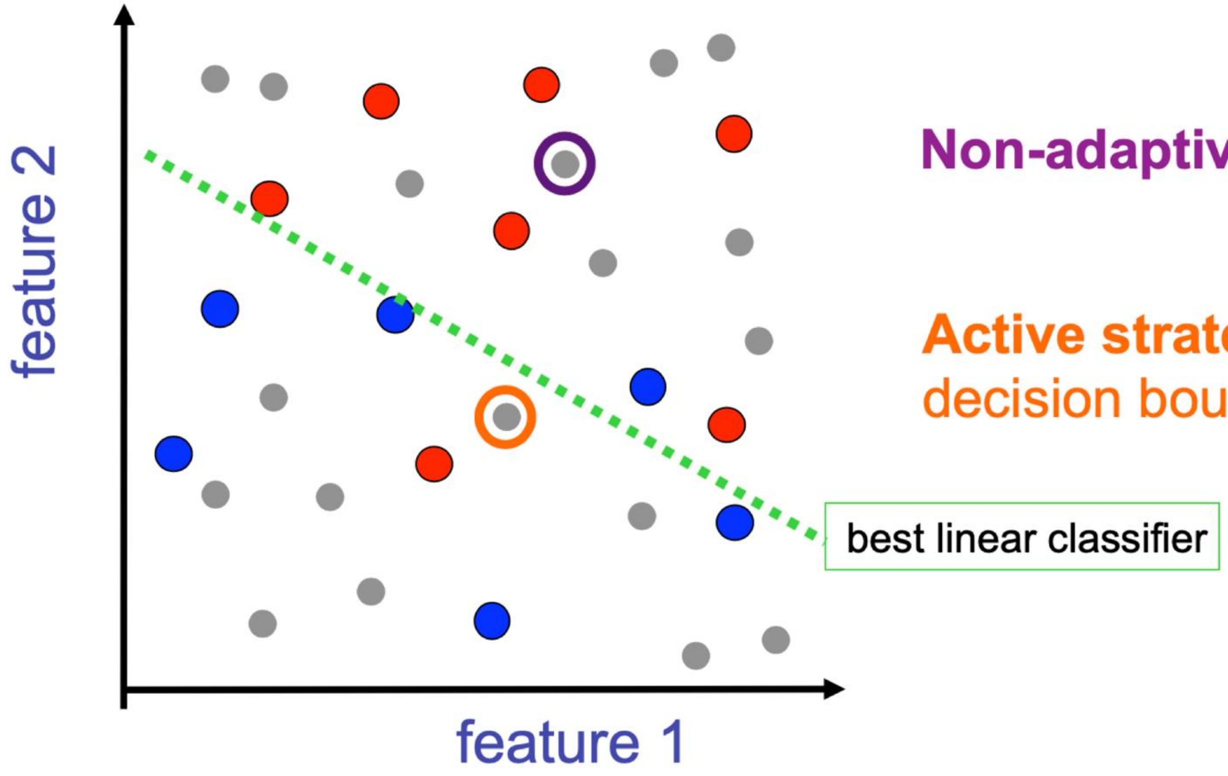


Illustration of active learning. The use of machine learning during a data labeling process. Image source: [Link]

Core ideas of active learning



Non-adaptive strategy: Label a random sample

Active strategy: Label a sample near best decision boundary based on labels seen so far

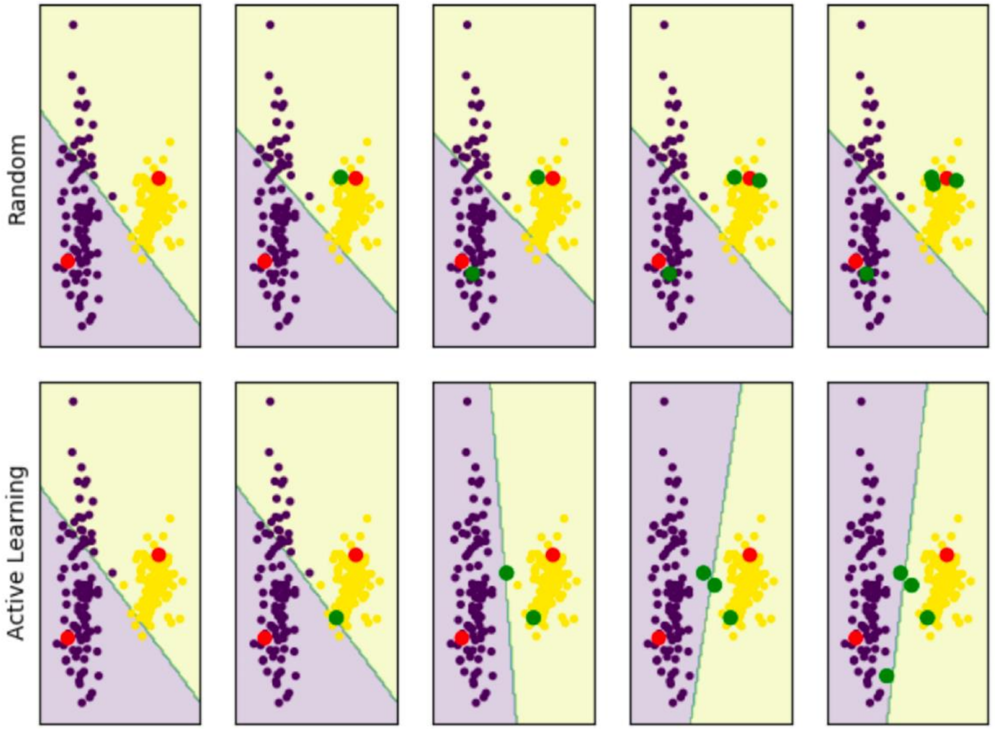
Uncertainty sampling algorithm

- Goal: Select the data points for which a (classification) model is most uncertain
- Example for a simple measure of uncertainty:

> The very simple measure of the uncertainty of the classification can be define as follows:
>
> $$U(x) = 1 - P(\hat{x}|x)$$
>
> where x is the instance to be predicted and $\hat{x}$ is the most likely prediction. For example, if you have classes `[0,1,2]` and classification probabilities `[0.1, 0.2, 0.7]`, the most likely class according to the classifier is 2 with uncertainty 0.3.
> ⇒ **The sample with the highest uncertainty is chosen for labeling.**

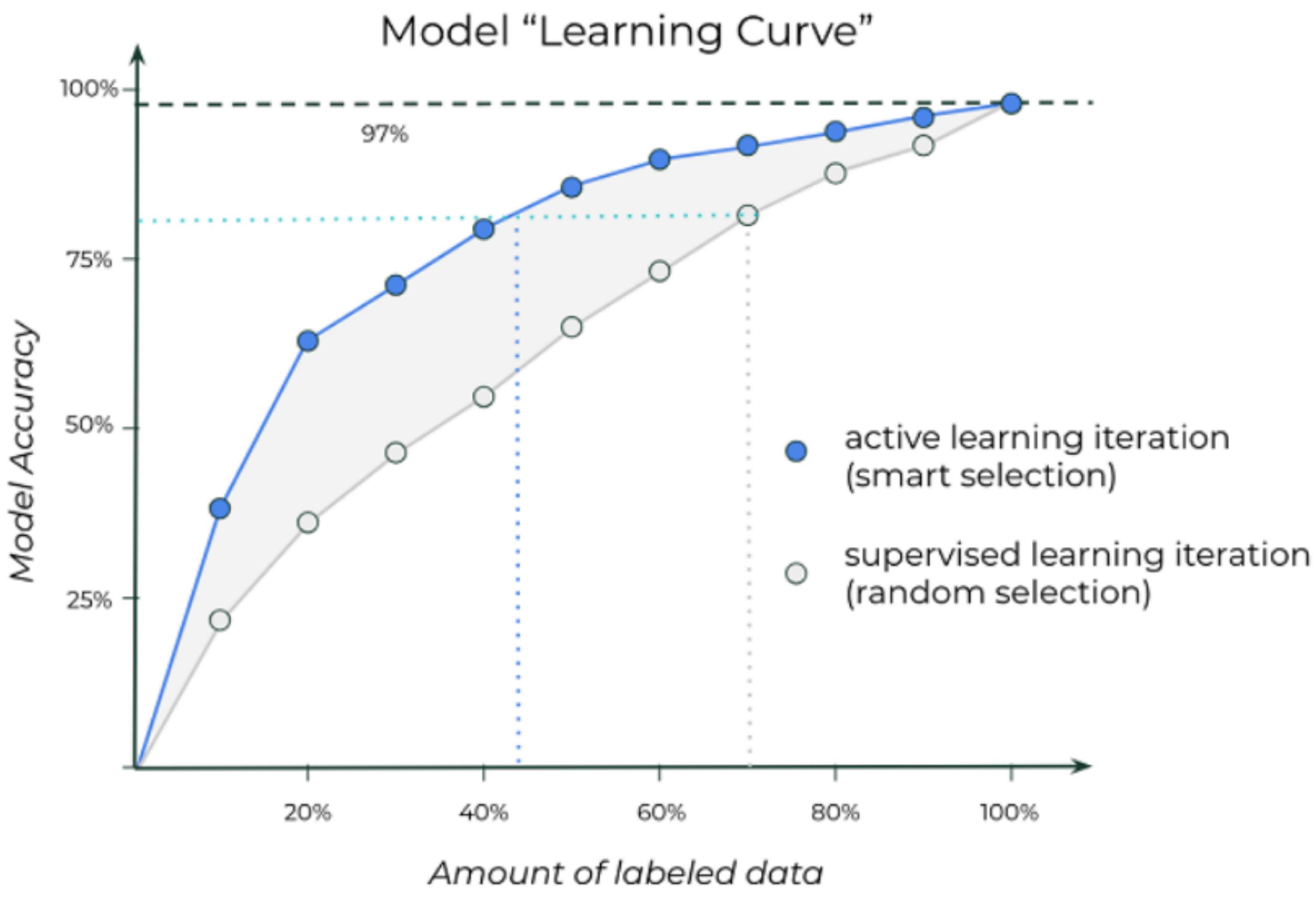
- Alternative measures of classification uncertainty: classification margin, classification entropy

Example: Effect of Active Learning



Comparison of random sampling and active learning for a binary classification problem (yellow and purple classes). Both algorithms get one initial labeled point from each blob (red dots) and sequentially ask for the label of additional points (green dots). Uncertainty sampling selects evaluation points near the decision boundary. Image source: Link

Example: Effect of Active Learning



Takeaways

- Training data generation is one of the biggest bottlenecks for ML applications
- Training data is very valuable
- Active learning can find an optimal classifier with less human supervision
- AL achieves higher performance (for the same number of samples to label)