

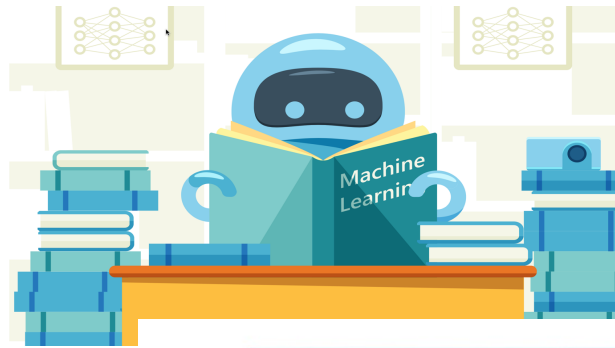
Building and Using Predictive Models

Machine Learning AI with Qlucore

The Qlucore logo, consisting of a stylized 'Q' made of two white circles and the word 'QLUCORE' in a bold, white, sans-serif font, is overlaid on a background image. The background image shows a person's hand typing on a laptop keyboard in a laboratory setting. The laptop screen displays a software interface with various data points and graphs. In the foreground, there are several small vials with black caps. The overall scene suggests a scientific or research environment where machine learning is being applied.

QLUCORE

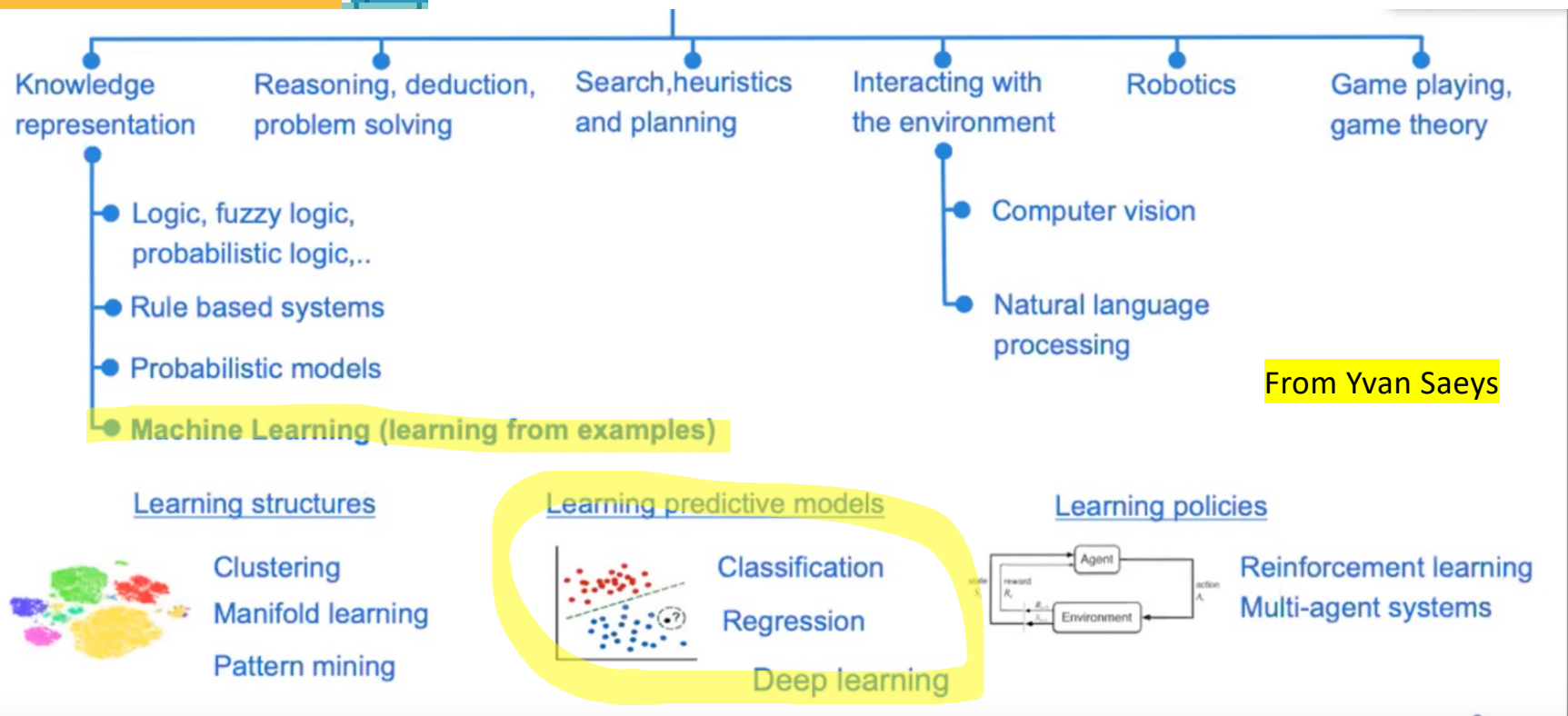
Yana Stackpole, PhD



Machine learning – a branch of AI

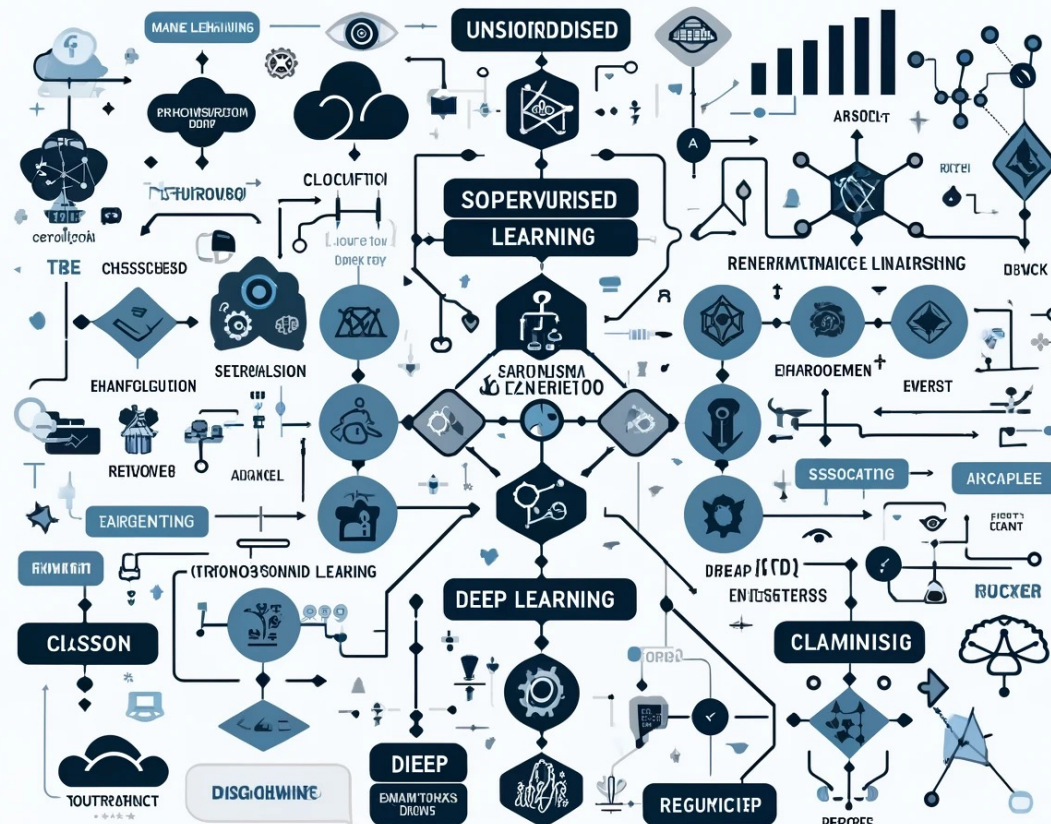
Computer program does not know the answer, but it knows how to learn from data we feed into it. It learns from examples/experience

Data Mining branch of AI – similar but does not make predictions



Machine learning flow chart, example of my prompt to ChatGPT

A lot of made up and nonsense



Good review article

[nature](#) > [nature medicine](#) > [news feature](#) > article

News Feature | [Published: 13 January 2021](#)

Eight ways machine learning is assisting medicine

[Mike May](#)

[Nature Medicine](#) 27, 2–3 (2021) | [Cite this article](#)

10k Accesses | 4 Citations | 92 Altmetric | [Metrics](#)

There has been a lot of hype around the applications of machine learning in medicine. But how is machine learning actually helping bench-to-bedside scientists and clinicians do their jobs?

The idea of improving medicine with computation is almost as old as digital computers. In the early 1960s, scientists [used](#) a computer in diagnosing blood diseases, and that was just one pioneering example in this field. In the branch of artificial intelligence (AI) called ‘machine learning’, computer software learns from experience. The results teach medical researchers and clinicians new ways of studying diseases, making medicines and treating patients.

In the Lab

- To simulate disease course and drug response - at the individual patient level
- To ID most appropriate patient group for enrolling into clinical trials

In the Clinic

- Developing diagnostics models (most success with images so far):

April 2018 - FDA approved the first AI-based diagnostic, IDx-DR, to detect diabetic retinopathy by analyzing retinal images. Qlucore Dx for blood cancers is used in a clinical setting (limited so far, in works)

- Improving disease prognosis, personalized approach
- Predicting response to treatment before it is given
- Predicting survival time / Risk group
- Patient monitoring – Apple Watch alerting you of arrhythmia
- Qlucore is building now custom RNAseq data-based models for cancer diagnostics – **Qlucore Dx**



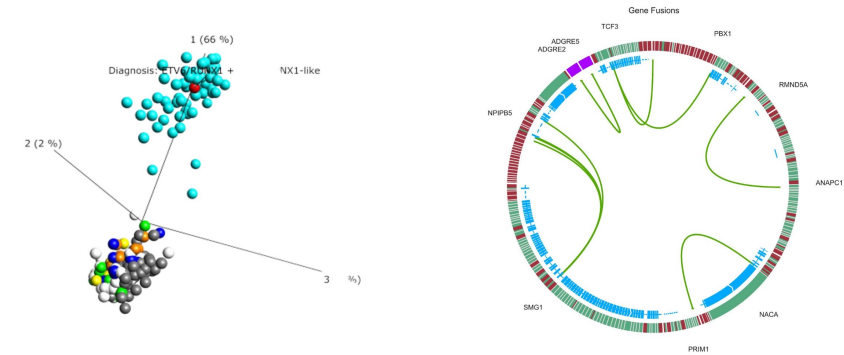
Analysis result

Patient name	Ellen Ripley	Analysis date	2022-05-05 14:43:07
Patient ID	19790525-1234	Model	Acute Lymphoblastic Leukemia
Sample file	case_189.bam	Model version	1.0.904
Reference Genome	hg19	SW version	0.9.5
Gene fusion file(s)	star-fusion.fusion_predictions.abridged.tsv	Technology	RNA Seq.

Method description RNA-seq based analysis for detection of gene fusions and gene expression classification for Acute Lymphoblastic Leukemia.

Conclusion The subtype based on gene expression signature for patient Ellen Ripley is ETV6::RUNX1 or ETV6::RUNX1-like with confidence level 1.0. The following gene fusions have been detected: .

Example of Qlucore Insights Report



How it works

1. **Define exactly what you want to predict.** Formulate precisely and carefully "Find a model based on specific OMICs (single or multi omics) to correctly predict X class for Y samples"
2. **Collect Training dataset** – with already known samples, dataset relatable to the real life applications you have in mind
3. **Collect Validation dataset** (desirable but optional)
4. **Samples you want to predict status/class for** - were you do not know it
5. **Train your model**

Sample ID	Output (Diagnostic)	Output (Predictive)
Patients – Training set	Stage 1	Relapse in 24 months
Patients – Training set	Stage 2	No relapse in 24 months
Patient – Validation set	Stage 1	Relapse in 24 months
Patients – Validation set	Stage 2	No relapse in 24 months
New Patients to classify	Stage X	Predicted Relapse in 24 months – Yes or No

Live Demo

Important Considerations

Very easy execution in Qlucore GUI 😊

... If you have right data to build your particular model ☹

“Machine learning **fueled by the right data** has the power to transform the development of breakthrough, new medicines and optimize their use in patient care.” Hill

What makes your data “right” for the model you want to build?

- There must be a **signal** in the data capturing difference between classes you want to predict.
- **Sample pool** for Training must correspond to the sample pool you will be classifying. Consider all kinds of variation – gender, age, race, location, comorbidities... anything potentially contributing to the variance in data.
- Signal vs Noise – problem of **overfitting** (trained to see noise) and **underfitting** (too simple, not enough).
- Nature of the data noise in your **Training, Validation** and **To classify** datasets must be similar or accounted for (sequencing platforms, type of gene expression data, scale if cross-OMICs)

Statistical Analysis vs ML

Different approaches, purposes, and techniques (ChatGPT)

1. Purpose and Goals

Statistical Analysis: The primary goal is to understand relationships within data, test hypotheses, and make inferences about a population from a sample. It focuses on explaining data and providing insights into how variables are related.

Machine Learning: The main goal is to make predictions or decisions based on data. It focuses on developing models that can generalize well to new, unseen data, often prioritizing predictive accuracy over interpretability.

2. Approach

Statistical Analysis: It often involves predefined models and hypotheses. Statistical methods rely on probability theory and require assumptions about the data distribution (e.g., normality, independence).

Machine Learning: It is more flexible and data-driven, often without strong assumptions about the underlying data distribution. ML can automatically learn patterns from data, even when the relationships are complex and not easily modeled by traditional statistical methods.

3. Techniques

Statistical Analysis: Common techniques include regression analysis, ANOVA, hypothesis testing, chi-square tests, and descriptive statistics. These methods are usually designed to be interpretable and provide clear statistical significance.

Machine Learning: Techniques include supervised learning (e.g., decision trees, support vector machines, neural networks), unsupervised learning (e.g., clustering, dimensionality reduction). They involve complex algorithms that may not be as easily interpretable as traditional statistical models.

Statistical Analysis vs ML

4. Data Requirements

Statistical Analysis: Typically requires smaller, well-structured data sets where assumptions about the data can be verified. It often involves cleaning and preprocessing data to meet these assumptions.

Machine Learning: Can handle large, high-dimensional, and unstructured data sets. It is designed to work with vast amounts of data, including big data, and often involves more extensive data preprocessing and feature engineering.

5. Outcome Interpretation

Statistical Analysis: Results are often presented with confidence intervals, p-values, and other measures of statistical significance. The focus is on understanding the data and the relationships between variables. One variable at a time.

Machine Learning: Results are usually evaluated based on metrics such as accuracy, precision, recall, F1 score, and ROC-AUC. The focus is on the performance of the model in making accurate predictions. Variables are treated as groups

6. Domain Knowledge

Statistical Analysis: Requires strong biology/pathology knowledge to choose appropriate models and correctly interpret results. The analyst needs to understand the underlying biology.

Machine Learning: Also benefits from domain knowledge, especially for feature selection, but it can often uncover patterns and relationships in the data without explicit domain knowledge.

Statistical analysis is more focused on inference and understanding relationships within data, whereas machine learning prioritizes prediction and pattern recognition.

More Considerations

- The model must be useful for new samples we will be classifying. **Know exactly what you want** to classify before you collect training data, and before you train your model.
- We want to choose the **smallest possible marker set** (50 is a good reference number). Even when we start with 30k genes, even if 10k will be differentiating your groups!
- **Markers** we are selecting from the model must be stable enough in Training, Validation and To Classify data. Keep in mind known and possible variability factors for your measurements.
- **Accuracy**. Trade off Sensitivity / Specificity ☹ (ideally we would like both to be perfect...)
Model Sensitivity = proportion of observed positives predicted to be positive
Model Specificity = proportion of observed negatives predicted to be negative
- **Sample size (n) for training a model**: “big-p, little-n” or **predictors/markers >> n**. Most machine learning algorithms assume $p \ll n$ --- not easy to fulfil with OMICs (20k genes, 100 samples for example, $p \gg n$). To really benefit of using deep learning algos you need more samples (n) than predictors (p) 20k genes – like 30k samples...
- Deep learning algos now are way ahead of sample number requirement for them. Not many can really benefit of deep learning so far. **You may be better off with simpler, even linear models** like SVM etc.

Summary

Machine learning **really works** in Biomedicine

However, **we don't fully understand** why it works (in any field). We cannot prove it theoretically (Pierre Dupont)

It is very useful, but **it does not replace human intelligence as** it does not simulate human mind.

You – (your human mind) are responsible for:

- Studying a model knowledge – explore selected markers, reason with them;
- Making biological interpretation in that context – think why those markers are/could be of interest;.
- Working on clinical validation.

Were We Want To Be

“Precision medicine requires interpretable, sparse and stable models, estimated from a few samples”

Pierre Dupont

“Should Biologists use Machine Learning?”

ChatGPT

Yes, biologists can benefit greatly from using machine learning (ML) in their research. Here are several reasons why:

- 1. Handling Large Data Sets:** Modern biological research often involves large and complex data sets, such as genomic sequences, proteomic data, and ecological data. Machine learning algorithms can handle, analyze, and interpret these large data sets more efficiently than traditional methods.
- 2. Pattern Recognition:** ML algorithms excel at recognizing patterns in data, which can be extremely useful in biology for tasks such as identifying gene expression patterns, predicting protein structures, and detecting anomalies in ecological data.
- 3. Predictive Modeling:** Machine learning can be used to build predictive models that can forecast biological phenomena, such as the spread of diseases, the impact of climate change on ecosystems, or the likelihood of a patient responding to a particular treatment.
- 4. Automation:** Many biological experiments involve repetitive tasks that can be automated using machine learning, saving researchers time and reducing the potential for human error.
- 5. Insights and Discoveries:** Machine learning can help uncover new insights and discoveries that might be missed by human researchers. For example, ML algorithms can identify subtle correlations and causal relationships in data that are not immediately apparent.
- 6. Personalized Medicine:** In the field of medicine, machine learning is crucial for developing personalized treatment plans based on an individual's genetic makeup, lifestyle, and other factors.

However, successful application of machine learning in biology requires interdisciplinary collaboration. Biologists need to work closely with data scientists and machine learning experts to ensure that the algorithms are appropriately applied and interpreted within the context of biological research.

Example – Improve a Writing Sample

My Input

for the prompt “Improve the webinar invite below”:

Useful options for visuals suitable for presenting and publishing results of OMICS data analysis In Qlucore Omics Explorer we interact with data via constant and instant visuals, hence seamlessly navigating between visuals, or working simultaneously with multiple synchronized plots.

In this webinar we will go over the following options.

- Customize the view of your plots, colors, titles, fonts etc.;
- Creating plots for publication vs presentation or analysis records;
- Handle labels, color legends;
- Export created plots;
- Saving preferences for complex plot configurations to pick up the work at a later time. Come join us!

Result, 1st iteration:

Webinar Invitation: Enhance Your Omics Data Visualization Skills

- **Title:** Useful Visuals for Presenting and Publishing Omics Data Analysis
- **Description:** In Qlucore Omics Explorer, data interaction is brought to life through constant and instant visuals. This allows for seamless navigation between visuals or simultaneous work with multiple synchronized plots.
- Join us for an informative webinar where we will cover the following key topics:
- **Customizing Plot Views:** Learn how to personalize your plots with various colors, titles, fonts, and more.
- **Creating Plots for Different Purposes:** Understand the distinctions between plots for publication, presentation, or analysis records.
- **Managing Labels and Color Legends:** Discover efficient ways to handle plot labels and color legends for clarity.
- **Exporting Plots:** Gain insights on exporting your created plots for various uses.
- **Saving Plot Preferences:** Learn how to save complex plot configurations to easily continue your work later.
- **Why Attend?** This webinar is designed to enhance your skills in visualizing and presenting Omics data. Whether you're preparing for a publication, a presentation, or simply need to maintain detailed analysis records, this session will provide you with valuable tools and techniques.
- **Join Us!** Don't miss this opportunity to elevate your data visualization skills with Qlucore Omics Explorer. We look forward to seeing you there!
- **[Register Now]**
- For any queries, please contact [Contact Information].