Guide to evaluating the application of machine learning methods in genetics literature

Laura Saba, PhD
Associate Professor
Co-Director of the NIDA Center of Excellence in Omics, Systems Genetics, and the Addictome
Department of Pharmaceutical Sciences
Skaggs School of Pharmacy and Pharmaceutical Sciences
University of Colorado Anschutz Medical Campus
Laura.Saba@cuanschutz.edu

Rationale for machine learning in genetics and omics studies

- **Millions of measurable molecular phenotypes –** Multiple omics technologies can measure millions of molecular phenotypes (e.g., DNA variants, RNA transcripts, metabolites, proteins, methylation marks)
- **Polygenicity of complex traits and pleiotropy of individual genes –** Not only are almost all complex traits influenced by multiple genes, but genes also have multiple functions in the cell.
- **Abundance of complex gene-by-gene interactions and gene-by-environment interaction –** GXG and GXE interactions are the norm rather than the exception.

How do we harness this volume of information for biological discovery?

Rationale for This Webinar

- The utilization machine learning methods in genetics studies and multi-omics research is on the rise and will likely continue to grow.
- Learning the basics of machine learning allows researchers to critically review current literature and to identify areas of their own research that may benefit from these types of approaches.
- To evaluate literature, we need to be cognizant of important features of a machine learning study that indicate the rigor of the study and the relevance to research and clinical practice.

Learning Objectives

By the end of this lecture, you should have a basic understanding of:

- The relationship between artificial intelligence (AI), machine learning (ML), and deep learning (DL).
- General scenarios when ML is appropriate.
- Methods for comparing the performance of ML models
- General criteria to examine when evaluating medical literature that includes machine learning algorithms

Outline

- General overview of big data methods
- Artificial intelligence and machine learning
- Evaluation of ML model performance
- Criteria for evaluating ML applications in medical literature

Introduction to Big Data Methods

predictors/causes **outcomes/effects**

Prediction. To be able to predict what the response is going to be based on input variables

Information. To extract some information about how nature is associating the response variable to the input variables.

Breiman L. Statistical modeling: Two cultures. Statistical Science. 2001 16(3):199-231

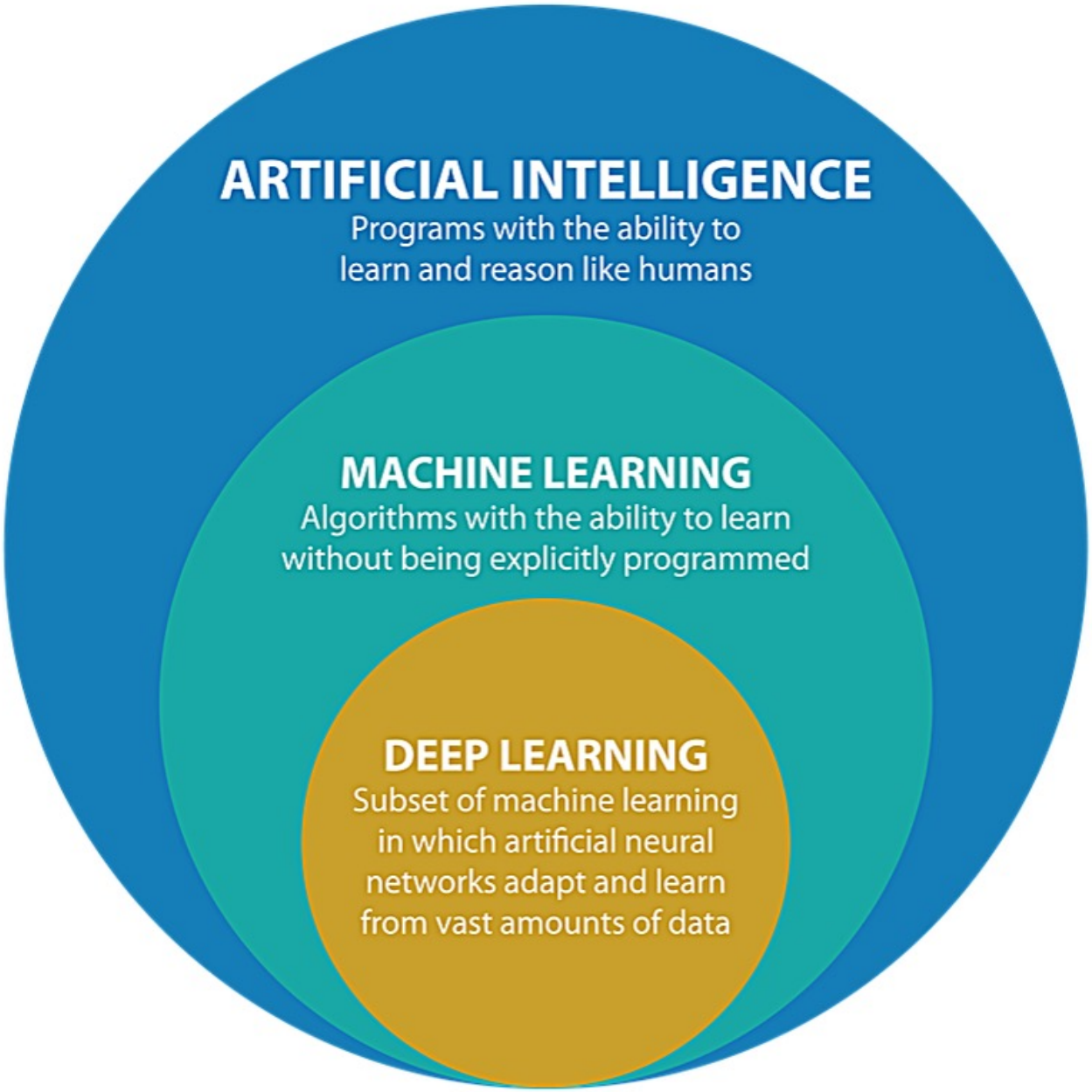
Machine Learning
(prediction analysis)

- Definition: algorithm that identifies factors than can predict an outcome (often computationally intensive)
- How/why a factor influences an outcome is often not directly modeled
- Benchmark for success: accuracy of predictions
- Ideal for identifying biomarkers of disease or drug response

Network Analysis
(functional analysis)

- Definition: mathematical model for describing the relationship among predictors and outcomes (often not a clear distinction between the two)
- Interpretability of the model (and model parameters) is often favored over prediction accuracy
- Benchmark for success: ability to mimic nature
- Ideal for identifying novel drug targets

Artificial Intelligence and Machine Learning



https://blog.digitalogy.co/the-difference-between-artificial-intelligence-machine-learning-and-deep-learning/

Machine can

Forecast

Memorize

Reproduce

Choose best item

Machine cannot

Create something new

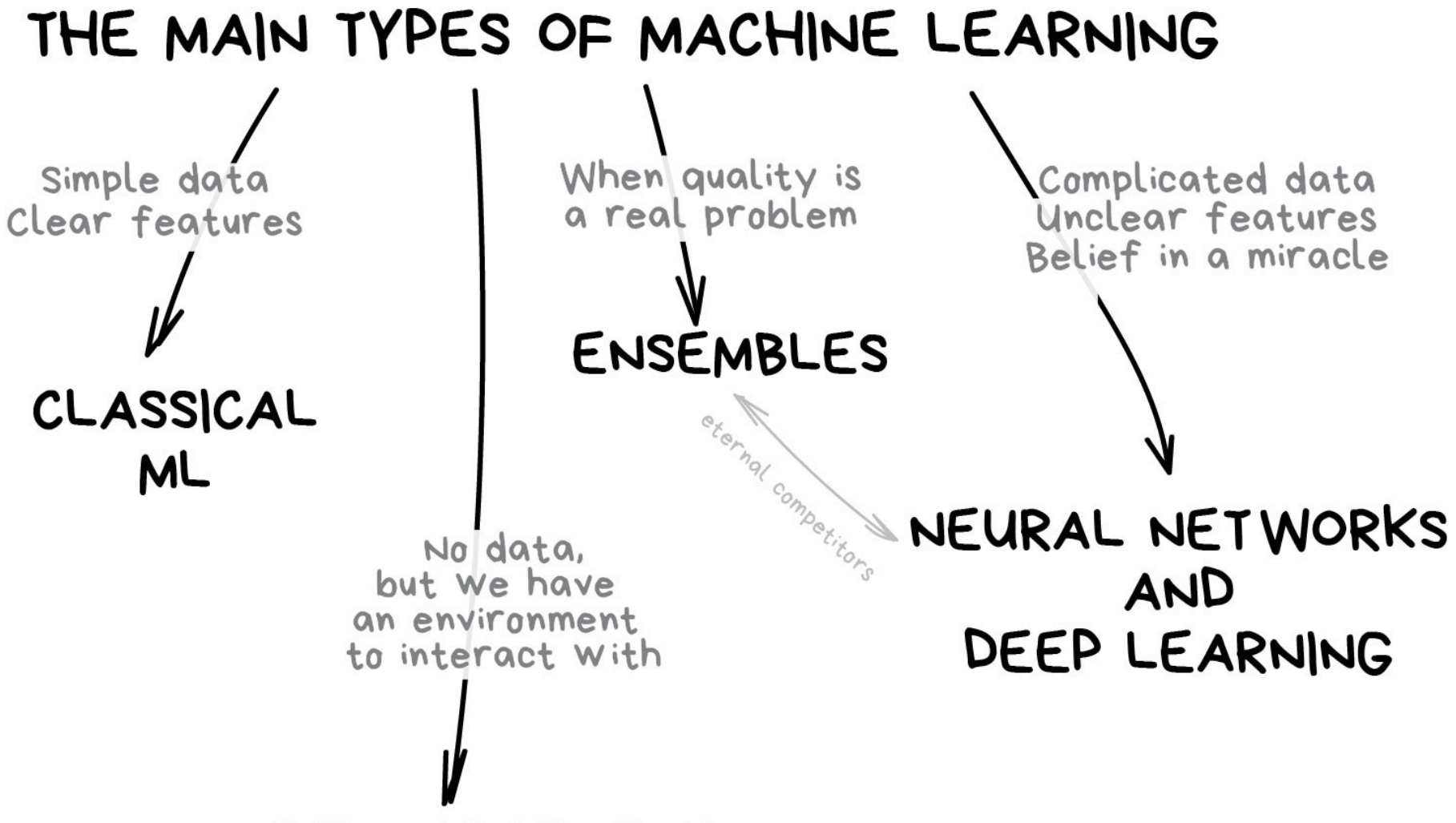
Get smart really fast

Go beyond their task

Kill all humans

https://vas3k.com/blog/machine_learning/

Terminology



Features = Characteristics used to predict the outcome

https://vas3k.com/blog/machine_learning/

CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

SUPERVISED

Predict a category

CLASSIFICATION

«Divide the socks by color»

Predict a number

REGRESSION

«Divide the ties by length»

Data is not labeled in any way

UNSUPERVISED

Divide by similarity

CLUSTERING

«Split up similar clothing into stacks»

Find hidden dependencies

DIMENSION REDUCTION (generalization)

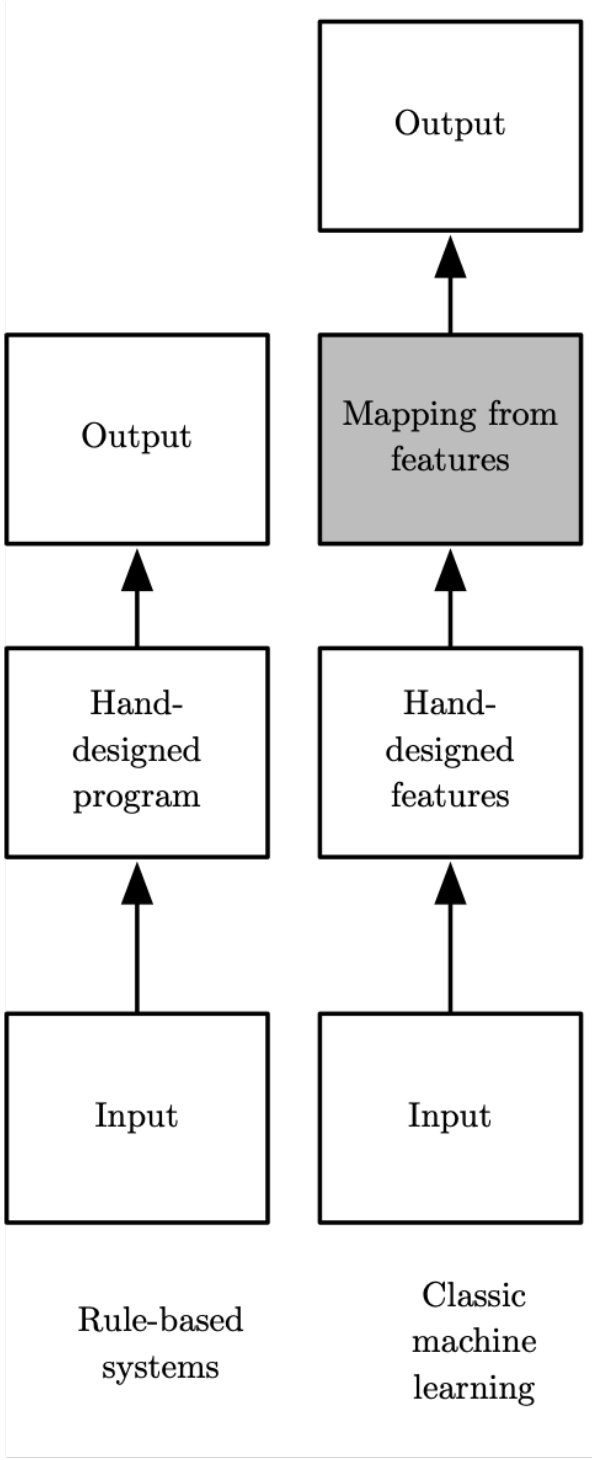
«Make the best outfits from the given clothes»

Identify sequences

ASSOCIATION

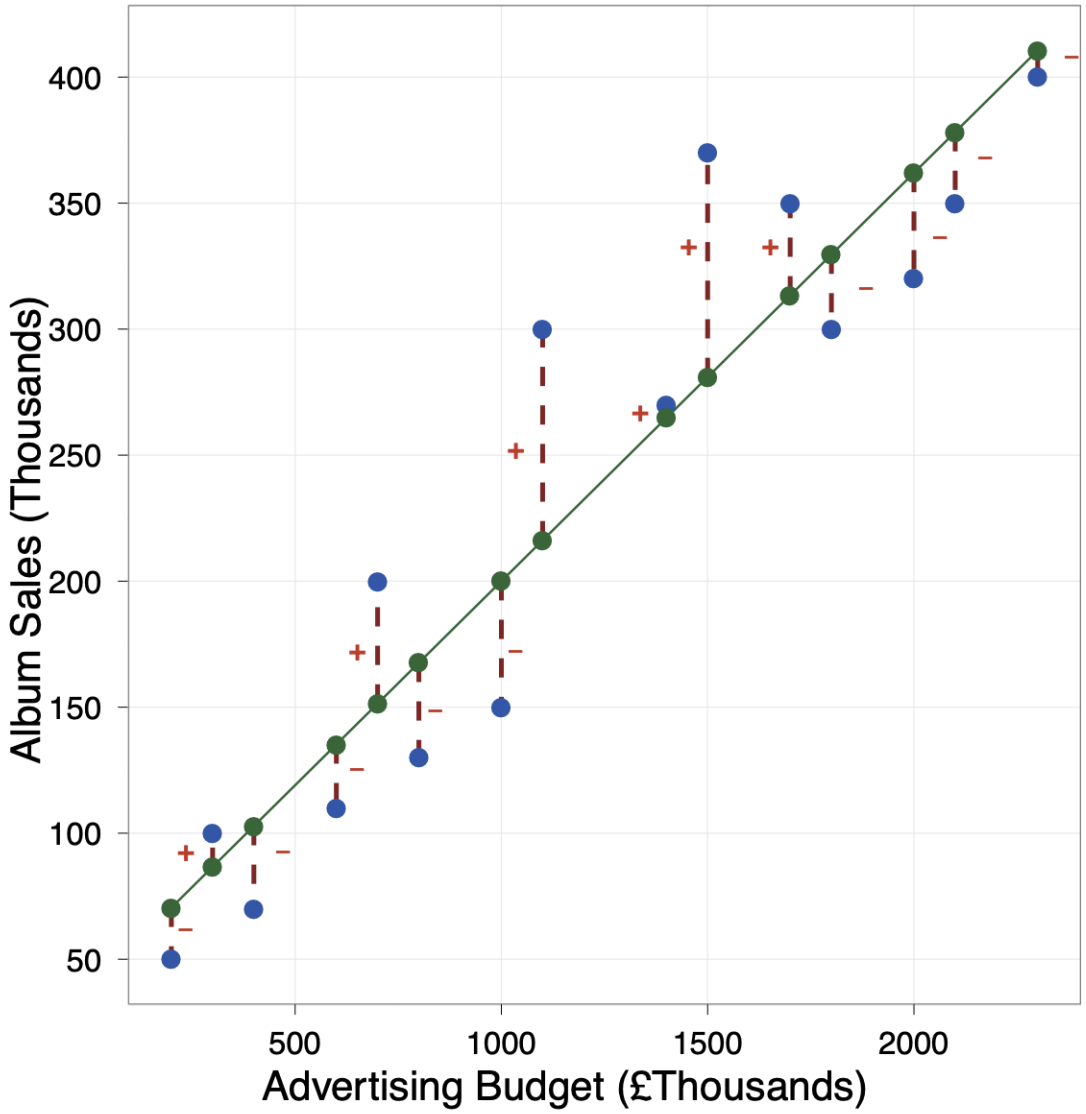
«Find what clothes I often wear together»

https://vas3k.com/blog/machine_learning/



Classic Machine Learning

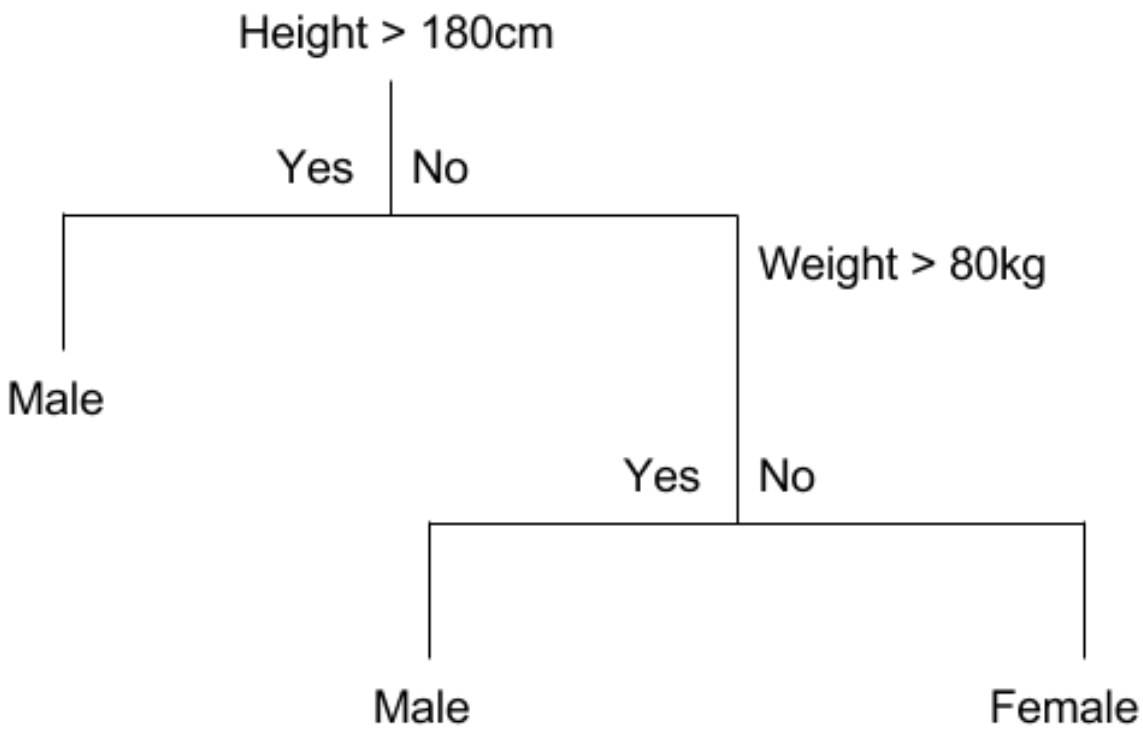
Linear/Logistic Regression



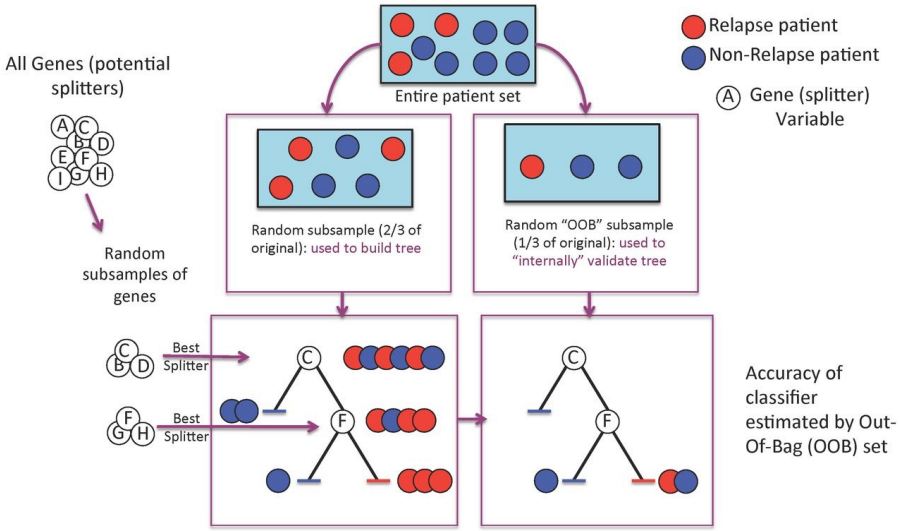
Sparse modeling

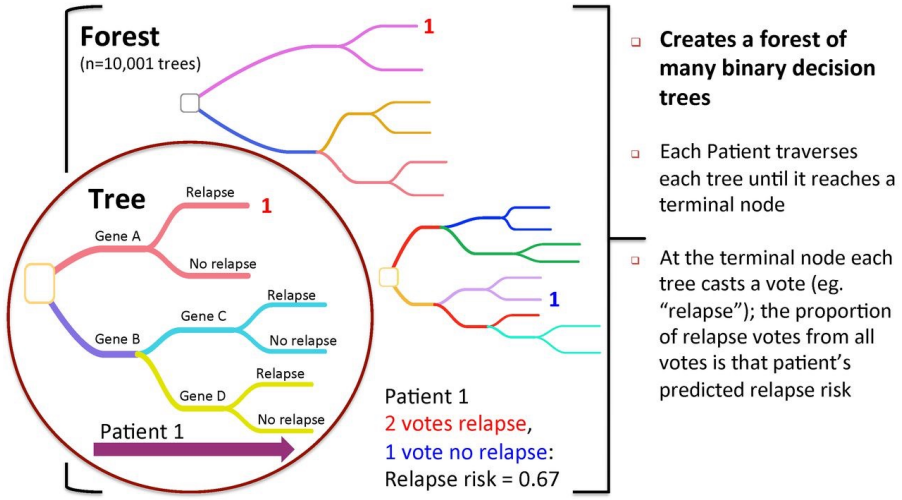
- Modeling the outcome using multiple predictors
  - When the number of possible predictors is large, traditional model building approaches tend to be computationally expensive.
- From LASSO to Ridge Regression
  - **Lasso regularization**
    - Identifies important predictors and eliminates redundant predictors
    - ‘Shrinks’ coefficients to avoid over fitting (many shrink to zero)
  - **Ridge regression**
    - ‘Shrinks’ coefficients to avoid over fitting, but not necessarily to zero
  - **Elastic net**
    - Hybrid of the two approaches
    - Selects important predictors, but doesn’t eliminate all redundant predictors

Regression trees for classification



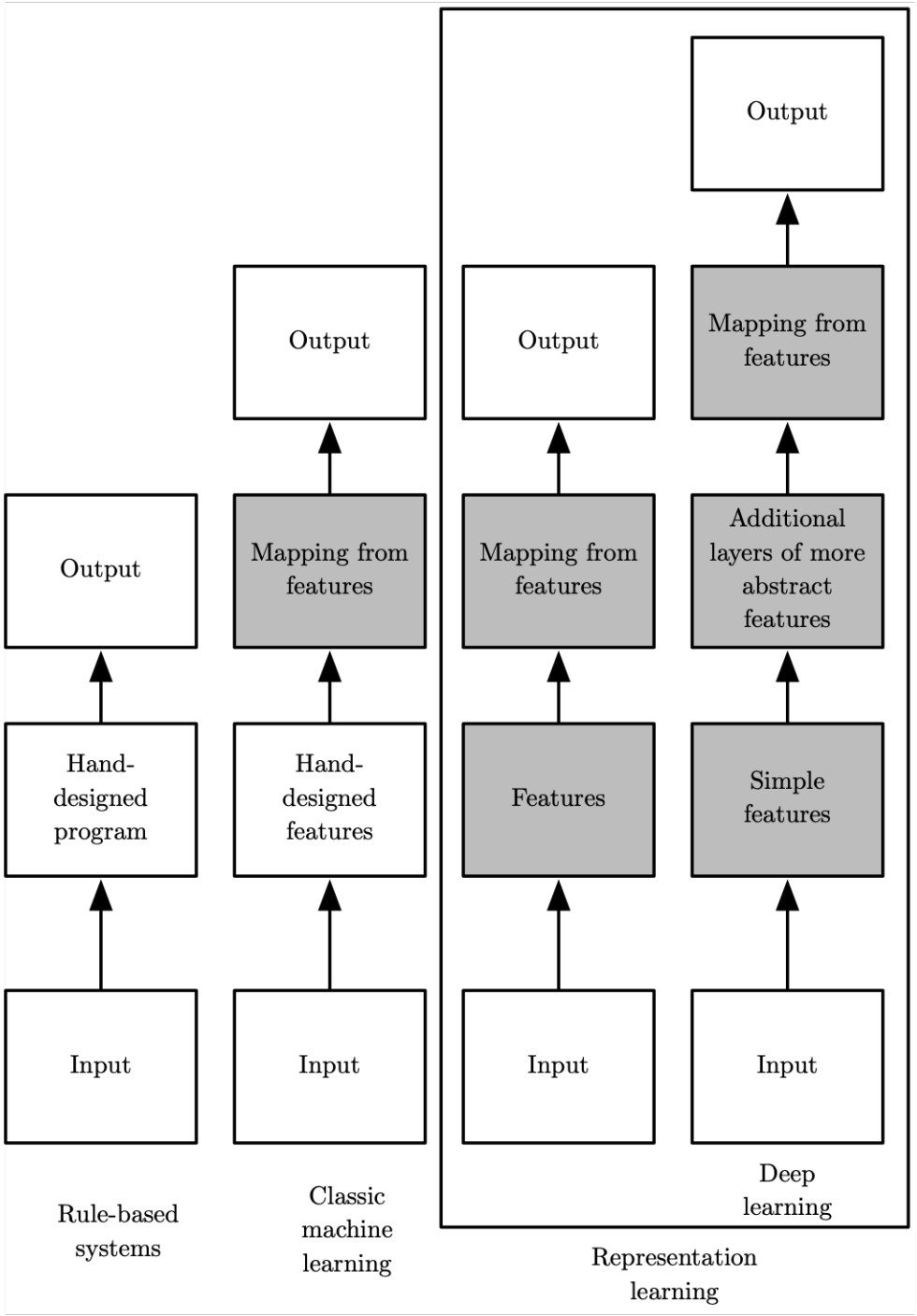
Random forest



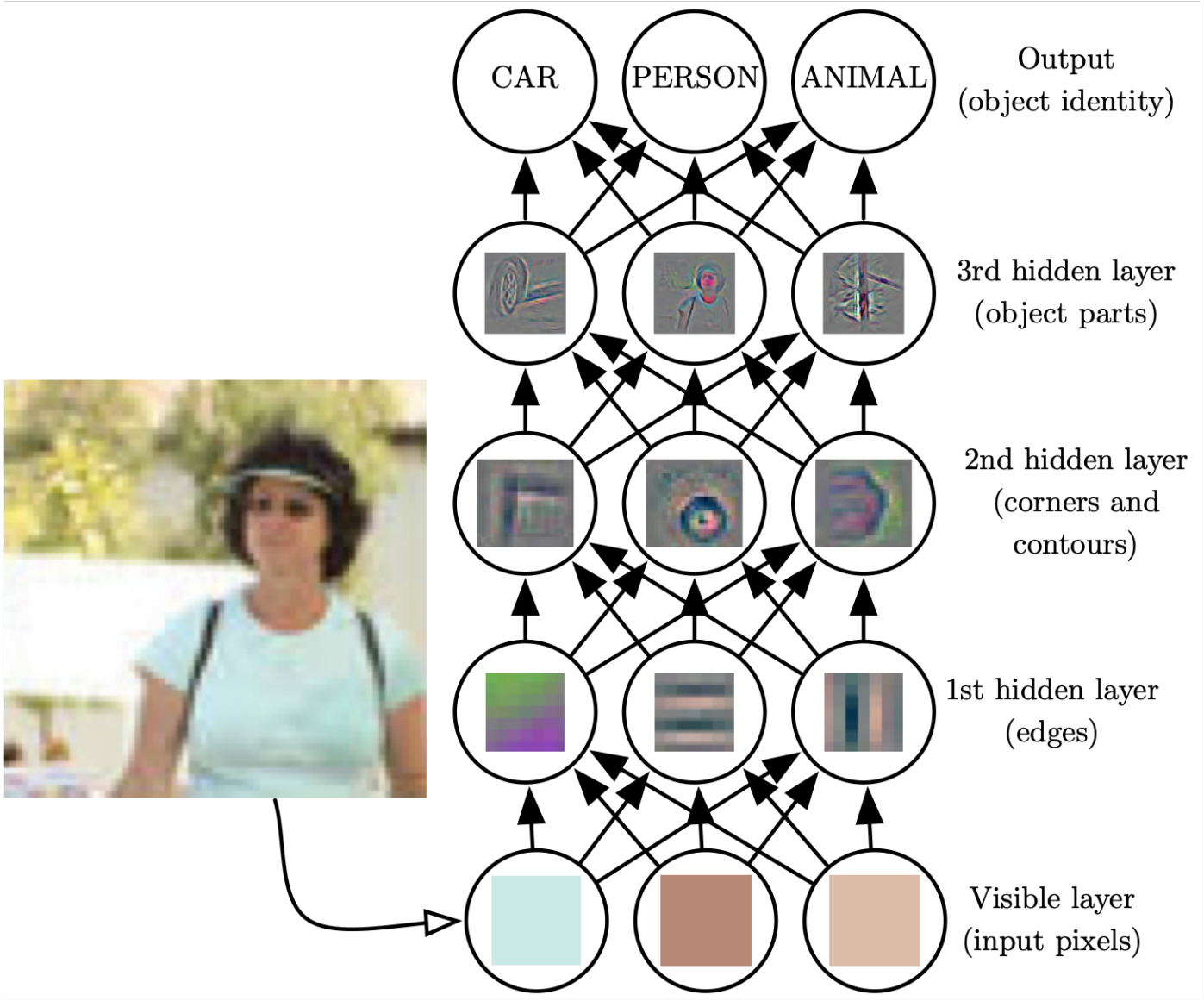


https://www.biostars.org/p/86981/

Deep learning



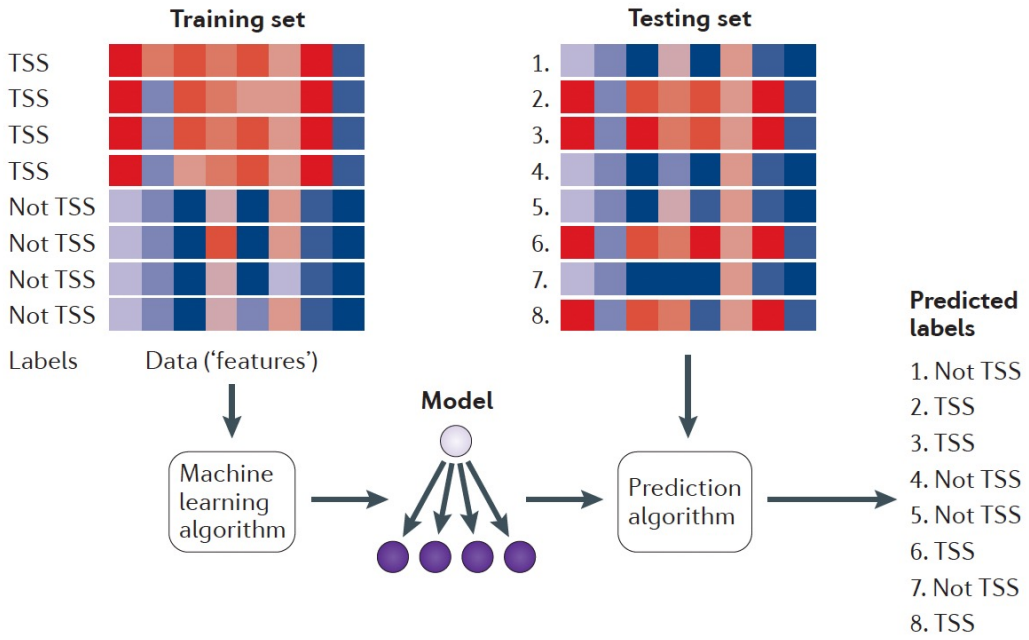
Deep learning



Evaluation of ML Model Performance

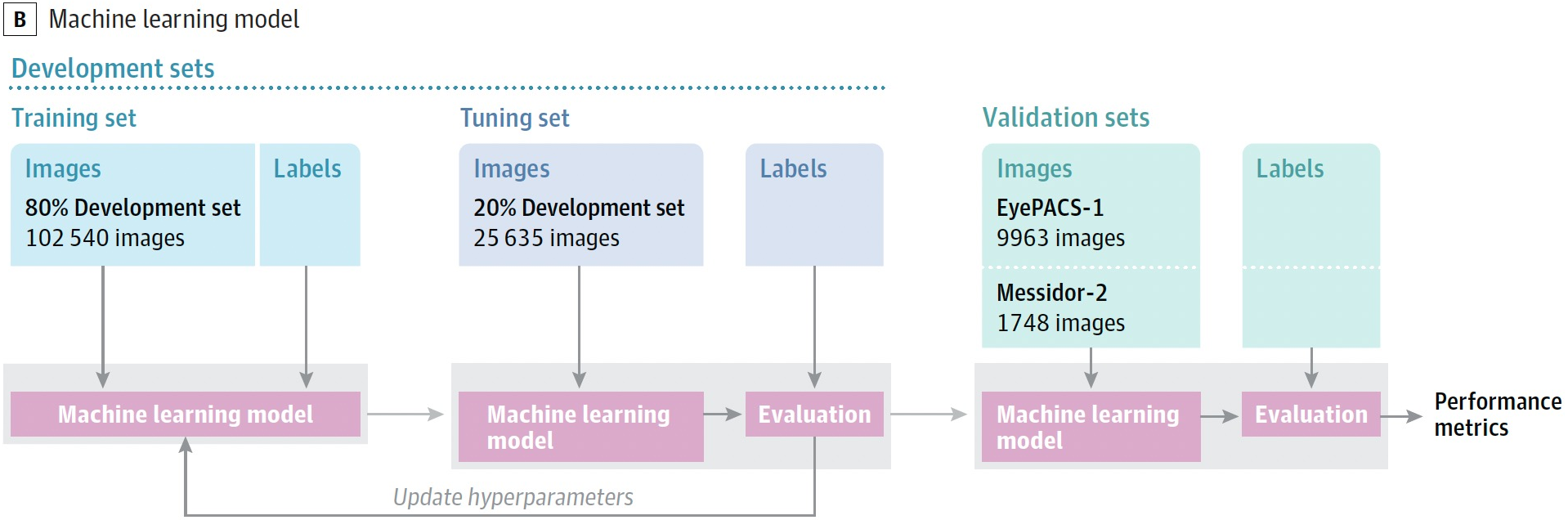
General Supervised ML Process

1. Preparing to Build a Model
   - Task definition
   - Data collection
   - Data preparation
2. Training Model
3. Evaluating Model
4. Implementation



Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015 Jun;16(6):321-32.

Evaluating Data Sets and Data Elements



Liu Y, Chen PC, Krause J, Peng L. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. *JAMA.* 2019;322(18):1806–1816.

a **hyperparameter** is a parameter whose value is used to control the learning process and must be set by the analyst

Data sets involved

- **Development data set** - used to build the predictive model
  - **Training data** - used to estimate the model parameters
  - **Tuning data** - used to tune the hyperparameters in more complex ML methods.

- **Test data (validation data)** - used to evaluate algorithm performance

Both development and test data sets need to have measurements for predictors/features AND the outcome/label

Data Set Questions

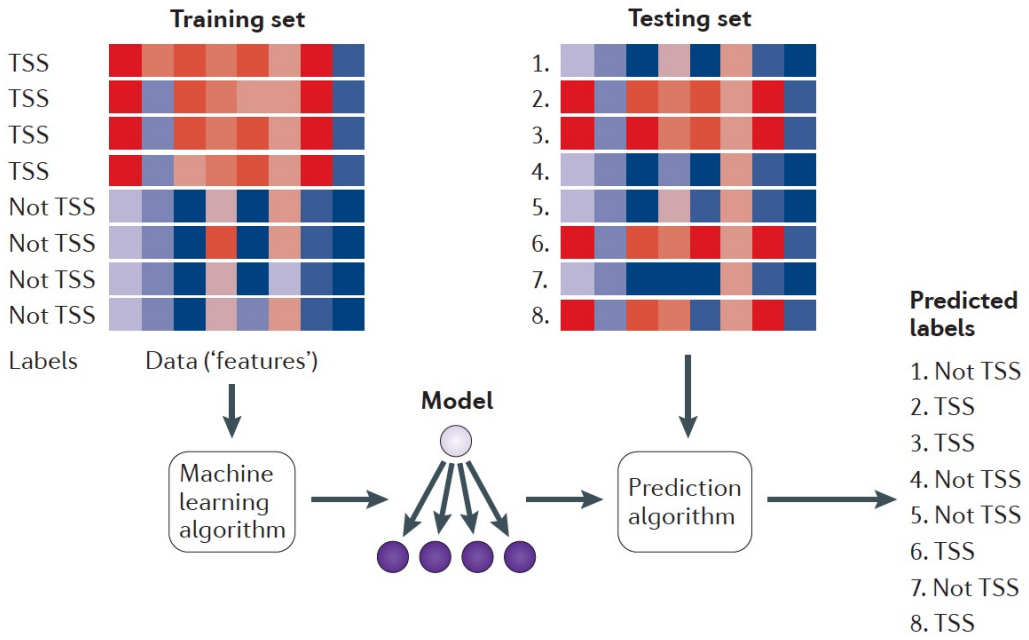
- Do the authors clearly define the training/development data vs. the test/validation data?
- Do the data in the training/development data set include the types of observations/patients that the algorithm will be eventually applied to?
- Is the validation data set completely independent from the development data?
- What are potential confounders and sources of bias in the data set?

Data Element Questions

- Are the outcomes (i.e., labels) for both the development and test data sets accurate?
- How many potential predictors were included? Did the algorithm ‘choose’ which predictors to include and how they would be represented?

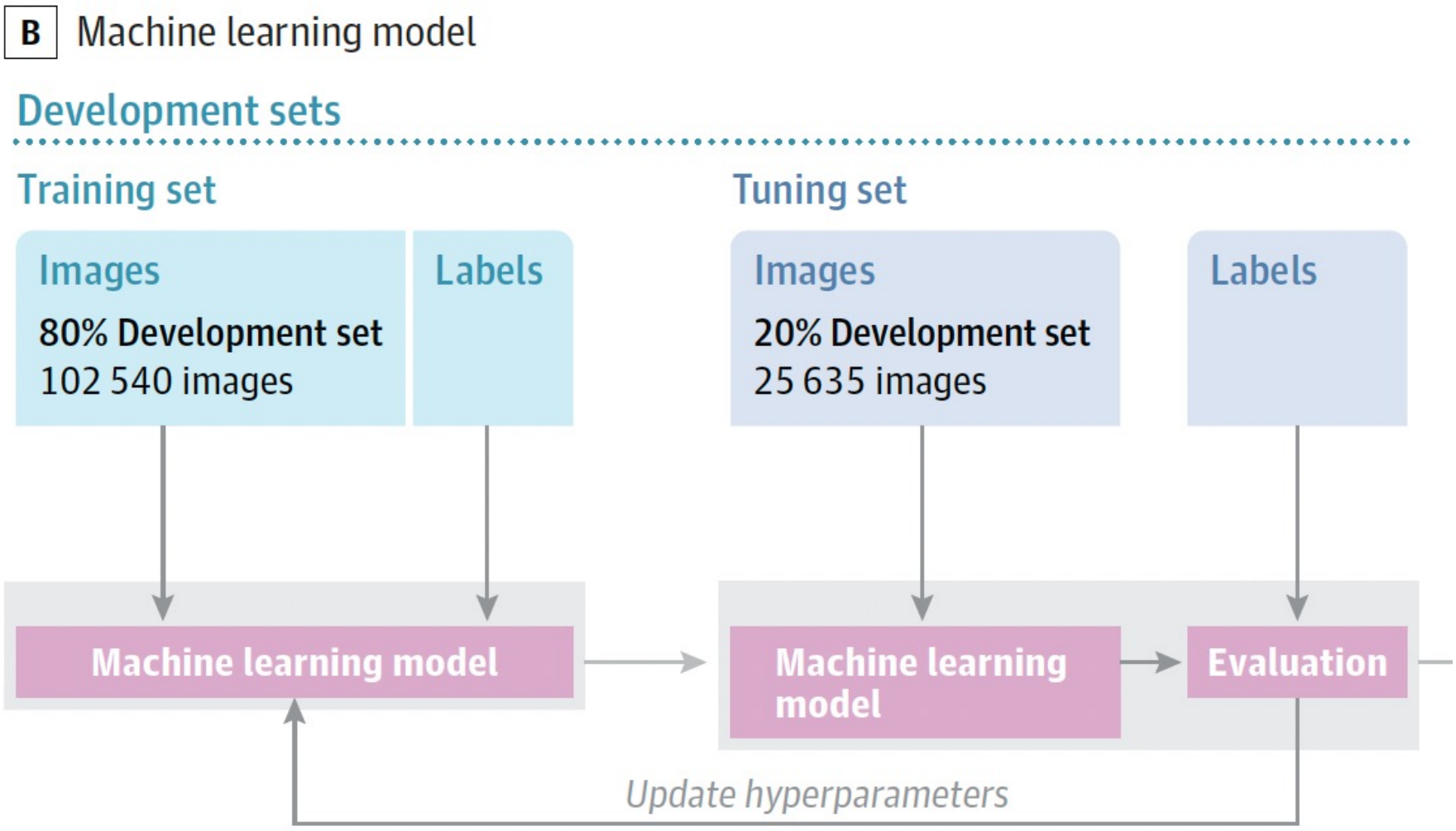
General Supervised ML Process

1. Preparing to Build a Model
   - Task definition
   - Data collection
   - Data preparation
2. **Training Model**
3. Evaluating Model
4. Implementation



Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015 Jun;16(6):321-32.

Training Model



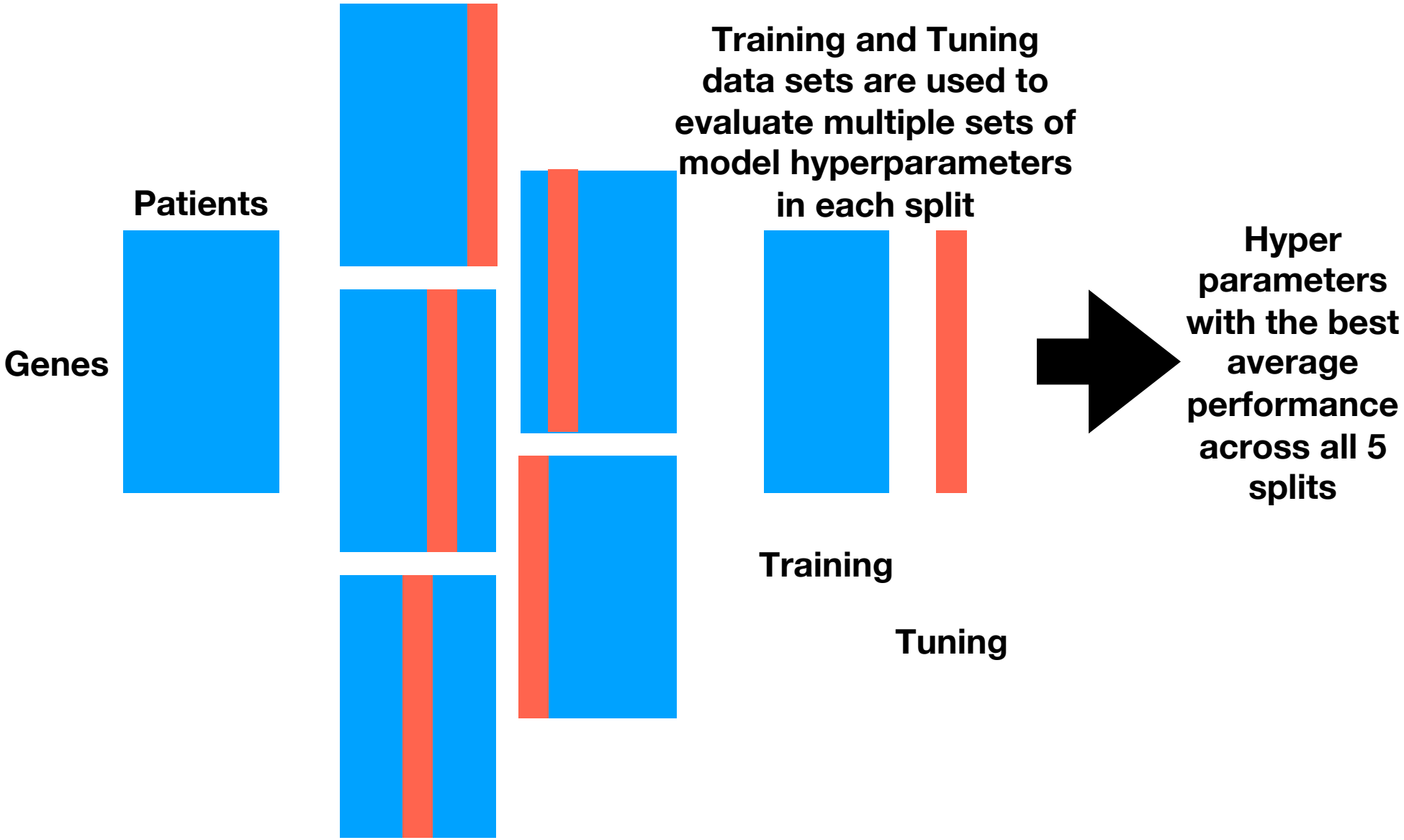
Role of Hyperparameters and Tuning Data

Hyperparameters: parameters that are established before a model is trained and remain fixed throughout the training process

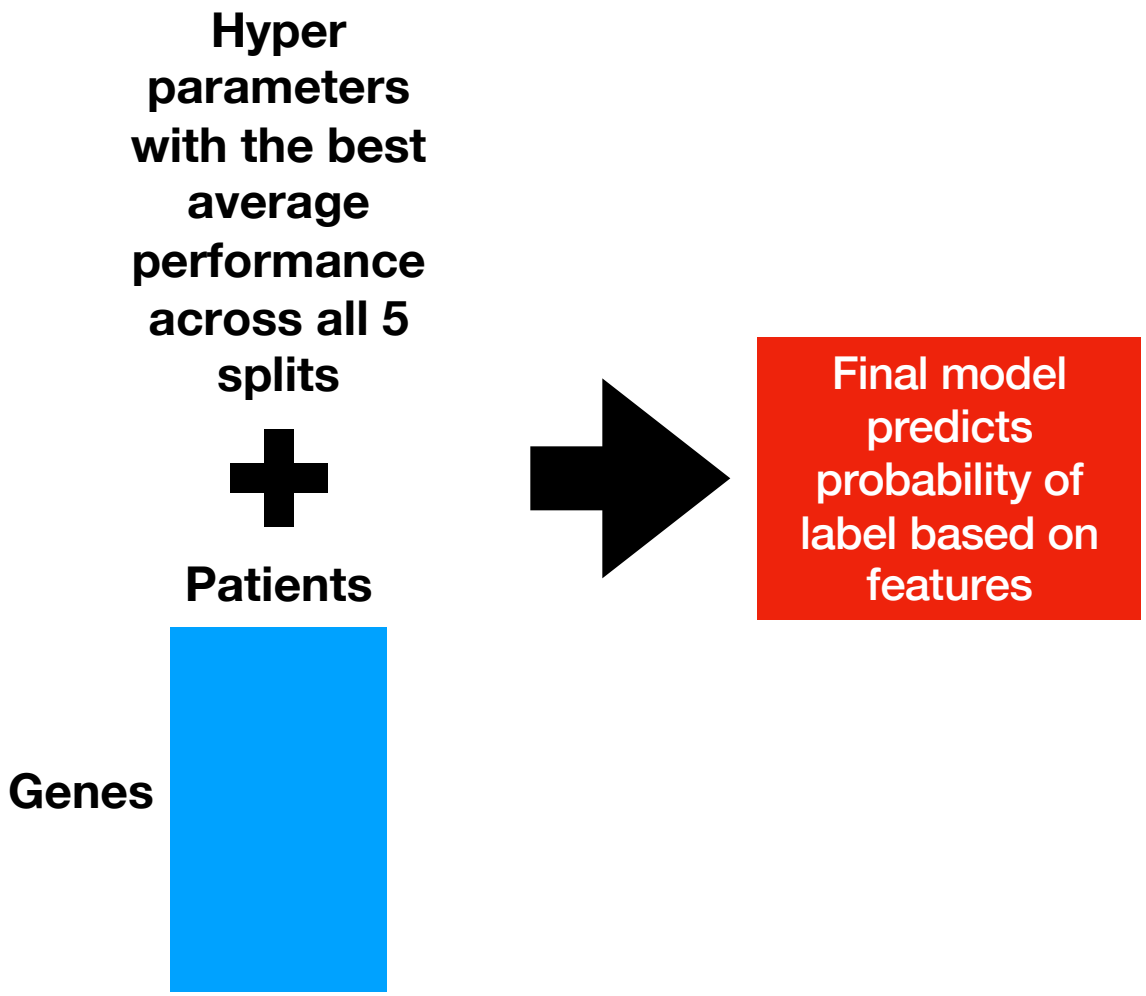
- Affect the parameters that are learned during training and can have a large influence on the final accuracy.
- One of the difficulties in machine learning is in determining sets of hyperparameters that optimize the model fit.

Cross Validation

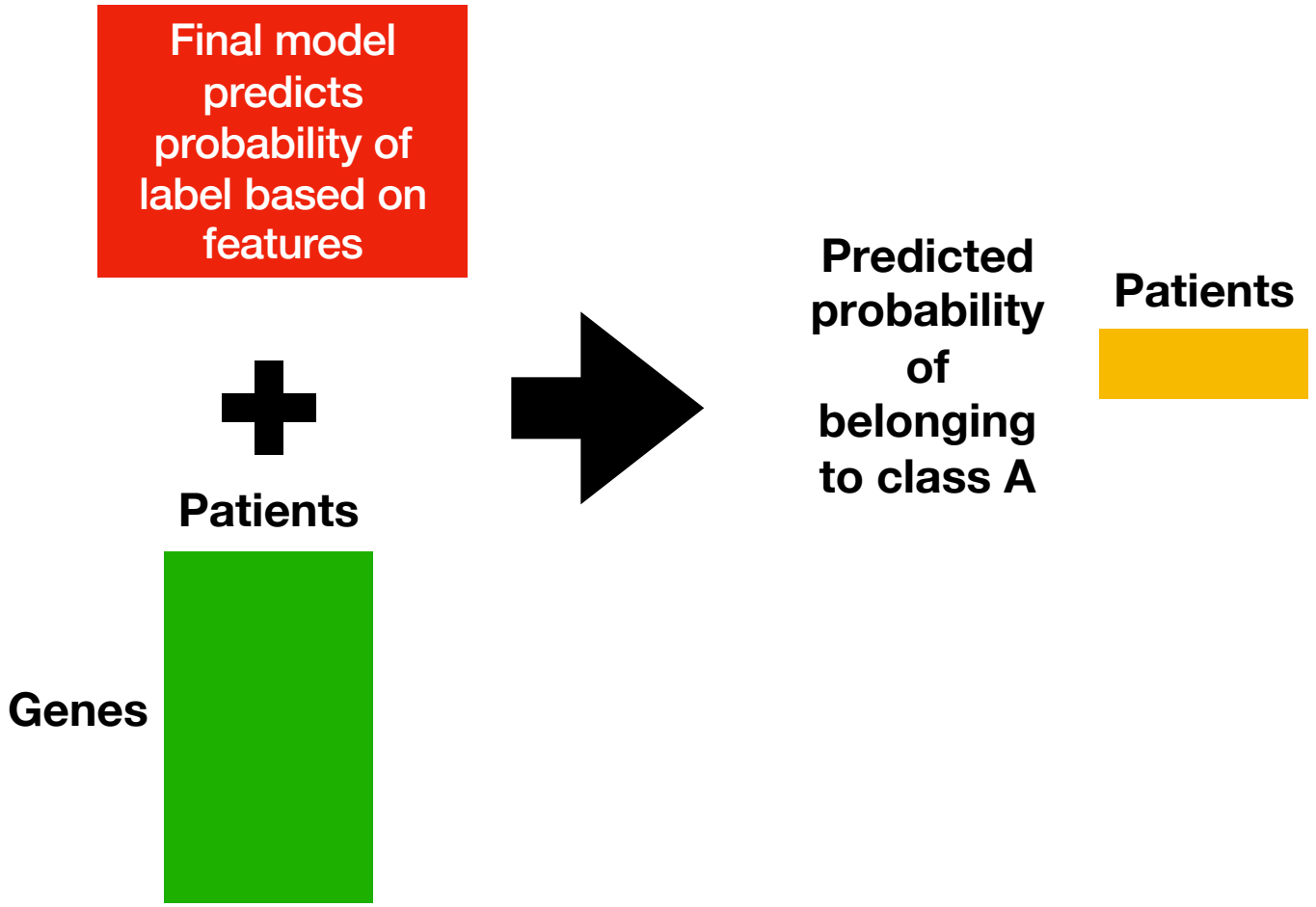
Often cross-validation is used as the ‘tuning data’



Using Hyperparameters to "Learn" the final model

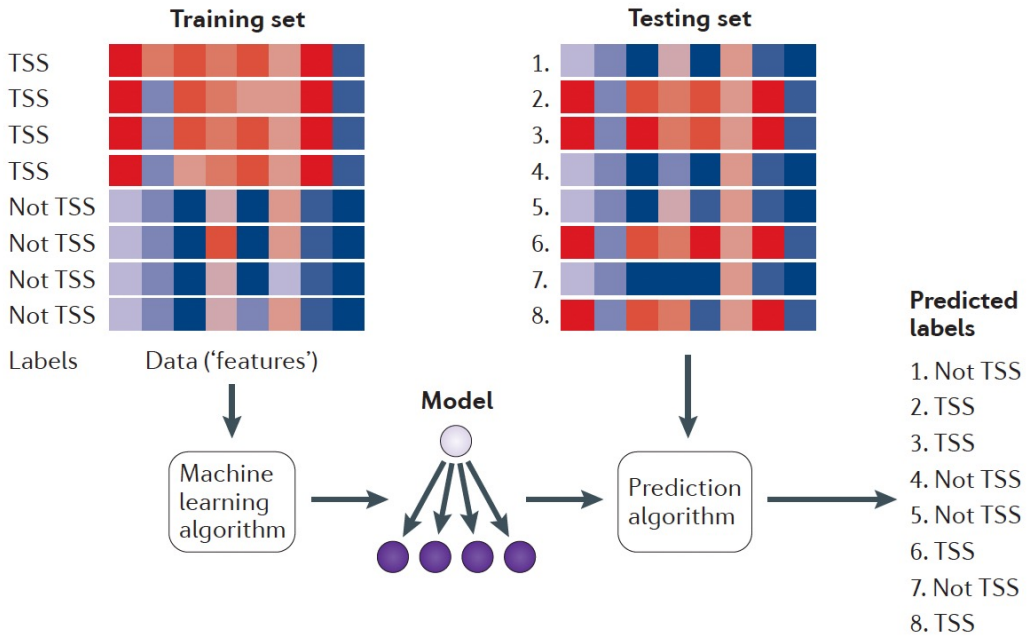


Using ML to predict classification of individuals in validation data set



General Supervised ML Process

1. Preparing to Build a Model
   - Task definition
   - Data collection
   - Data preparation
2. Training Model
3. **Evaluating Model**
4. Implementation



Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015 Jun;16(6):321-32.

Outcome of ML model

Each individual is assigned a probability of a particular label (i.e., outcome).

- e.g., individual i has a 90% probability of having the disease (if labels were disease/no disease)

For a given probability threshold, we can calculate a **confusion matrix**.

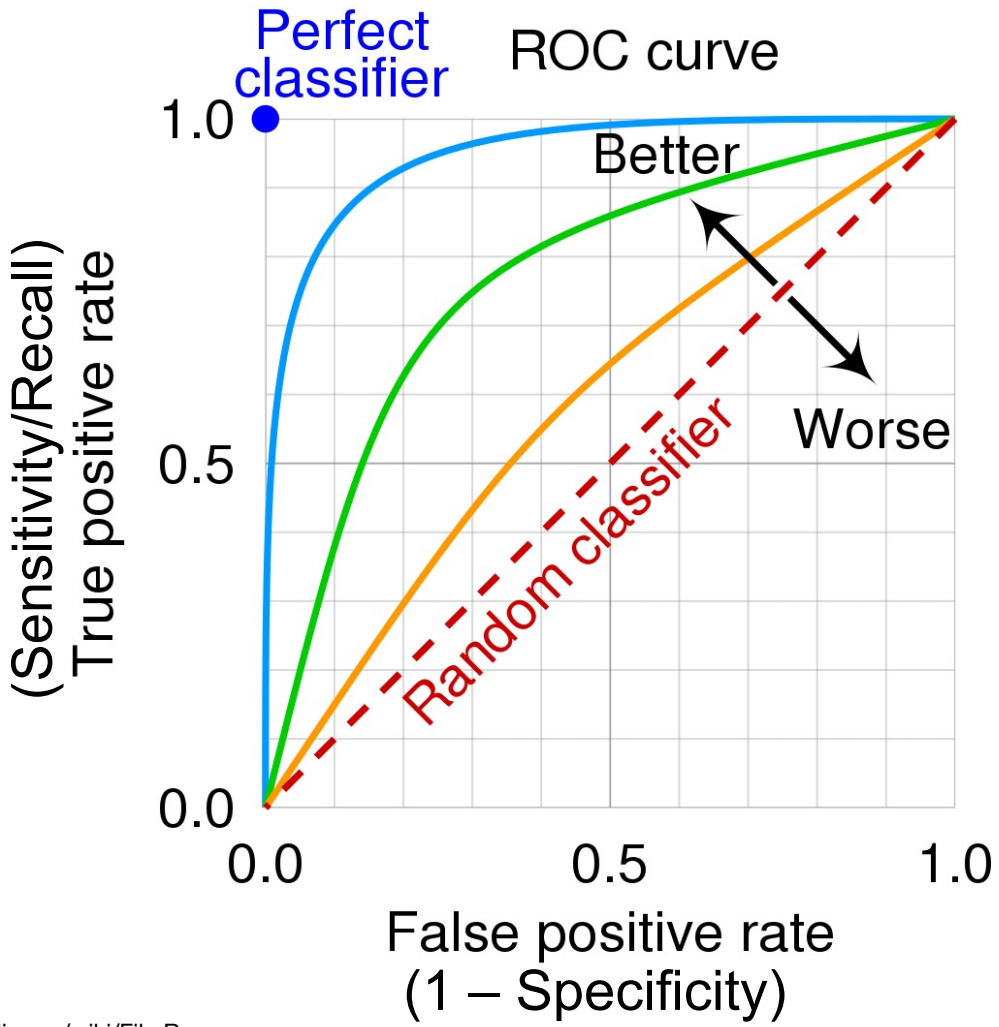
		Predicted Class	
		Predicted to be from Class A (positive)	Predicted to NOT be from Class A (negative)
Actual Class	From Class A	True Positive (TP)	False Negative (FN)
	NOT From Class A	False Positive (FP)	True Negative (TN)

Typical Summaries of a Confusion Matrix

- **Sensitivity/Recall** – TP / (TP + FN); e.g., if the patient does have the disease what is the probability the model detects it
- **Specificity (or 1 – Fallout)** – TN / (TN + FP); e.g., if the patient does NOT have the disease what is the probability the model indicates that they do not have the disease
- **Precision** – TP / (TP + FP); e.g., what is the probability that the patient does have the disease if the model indicates that they have the disease
- **Accuracy** - (TP + TN) / (TP + TN + FP + FN); what is the probability that the model correctly classified the patient

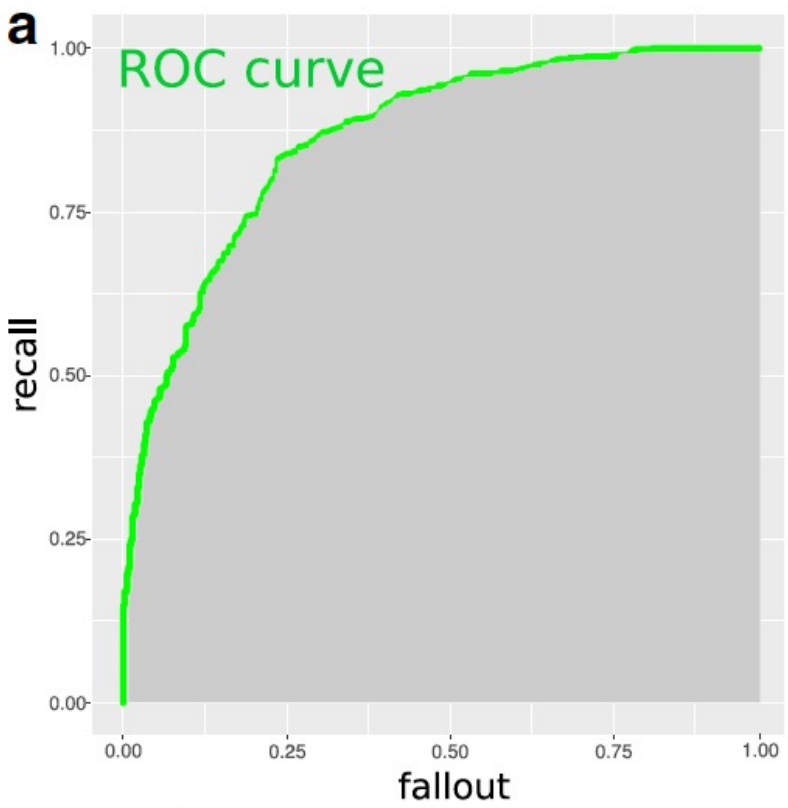
Curves to summarize results across various thresholds for a positive test result

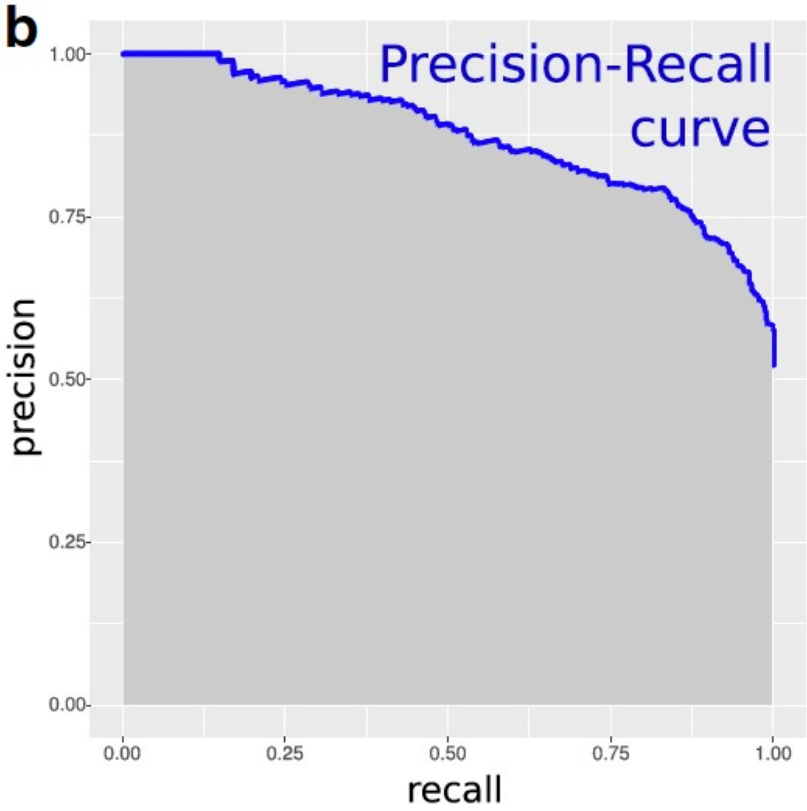
Probability Threshold	Sensitivity / Recall	1 - Specificity
0	100%	0%
0.25	92%	14%
0.5	53%	43%
0.75	12%	88%
1	0%	100%



https://commons.wikimedia.org/wiki/File:Roc_curve.svg

Curves to summarize results across various thresholds for a positive test result





$$recall = \frac{TP}{TP + FN} \quad fallout = \frac{FP}{FP + TN}$$

$$precision = \frac{TP}{TP + FP} \quad recall = \frac{TP}{TP + FN}$$

Precision-Recall curve is often preferred for <u>rare</u> outcome because true negatives (TN) are not included in the measures.

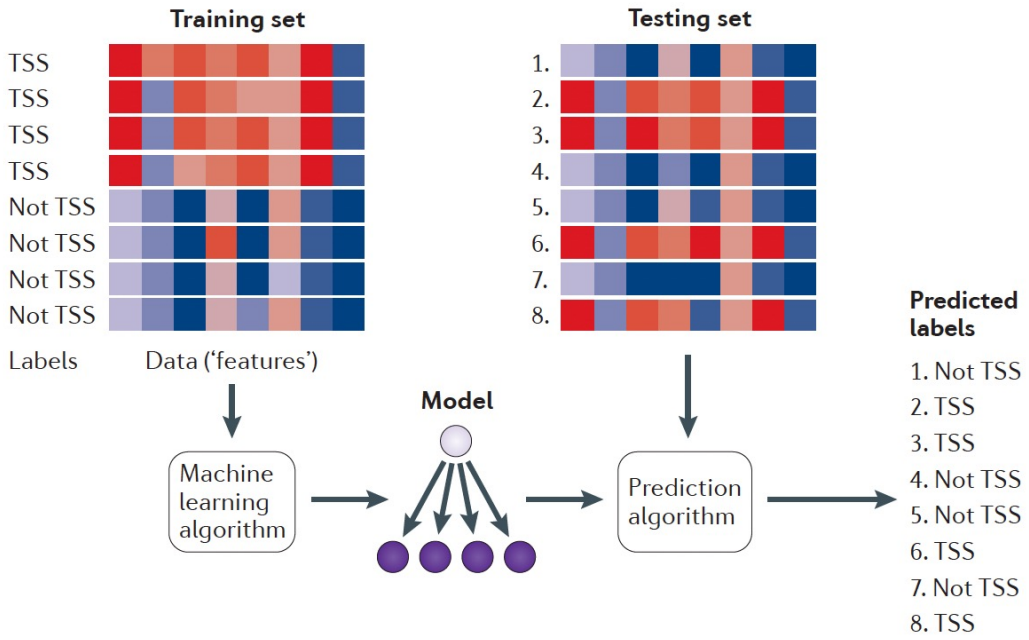
Overfitting

Overfitting – A scenario in which a machine learning model is trained to predict the training data too well, such that it does not generalize to new data sets.

- Often caused by including too many features/parameters that make the model specific to the training data
- One route of detection is to compare the performance of the model between the tuning and the validation data sets
- Overfitting leads to lack of repeatability and reproducibility.

General Supervised ML Process

1. Preparing to Build a Model
   - Task definition
   - Data collection
   - Data preparation
2. Training Model
3. Evaluating Model
4. **Implementation**



Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015 Jun;16(6):321-32.

Basics of Good Implementation of ML Models

- Do the authors provide a software implementation or webservice implementation of the model where users can input feature and receive outcome/label probabilities?
- What kind of preprocessing was used for the labels? e.g., do features need to be transformed or normalized
- Will the model be updated as new data become available?
- Were the validation data sets similar to the data sets I would like to apply the model to?

Criteria for evaluating ML applications in medical literature for clinical implementation

Clinical Review & Education

JAMA | Users' Guides to the Medical Literature

How to Read Articles That Use Machine Learning
Users' Guides to the Medical Literature

Yun Liu, PhD; Po-Hsuan Cameron Chen, PhD; Jonathan Krause, PhD; Lily Peng, MD, PhD

In recent years, many new clinical diagnostic tools have been developed using complicated machine learning methods. Irrespective of how a diagnostic tool is derived, it must be evaluated using a 3-step process of deriving, validating, and establishing the clinical effectiveness of the tool. Machine learning–based tools should also be assessed for the type of machine learning model used and its appropriateness for the input data type and data set size. Machine learning models also generally have additional prespecified settings called hyperparameters, which must be tuned on a data set independent of the validation set. On the validation set, the outcome against which the model is evaluated is termed the reference standard. The rigor of the reference standard must be assessed, such as against a universally accepted gold standard or expert grading.

JAMA. 2019;322(18):1806-1816. doi:10.1001/jama.2019.16489

Viewpoint page 1765 and **Editorial** page 1777

Supplemental content

CME Quiz at jamanetwork.com/learning

Author Affiliations: Google Health, Palo Alto, California.

Corresponding Author: Yun Liu, PhD, Google Health, 3400 Hillview Ave, Palo Alto, CA 94304 (liuyun@google.com).

What is the "Truth" that they are training the model with?

Did they test the model on a independent and realistic patient population

Was the population of subjects that got measured for the "truth" not dependent on the outcome the algorithm or their underlying disease state?

Box 1. Evaluating and Applying the Results of Studies of Diagnostic Tests[a]

Are the results of the study valid?

- Primary guides
  - Was there an independent, blind comparison with a reference standard?
  - Did the patient sample include an appropriate spectrum of patients to whom the diagnostic test will be applied in clinical practice?
  - Was there a completely independent validation set?
- Secondary guides
  - Did the results of the test being evaluated influence the decision to perform the reference standard?
  - Were the methods for performing the test described in sufficient detail to permit replication?

What were the results?

- Are likelihood ratios, sensitivity, and specificity for the test results presented or data necessary for their calculation provided?

Will the results help me in caring for my patients?

- Will the reproducibility of the test result and its interpretation be satisfactory in my setting?
- Are the results applicable to my patient?
- Will the results change my management?
- Will patients be better off as a result of the test?

[a] Information in this box is based on Jaeschke et al.[6,7]

Summary

- Machine learning methods are typically used for creating prediction models where **accuracy and precision** outweigh interpretation.
    - Models can be as simple as linear regression or as complex as a multi-layer deep learning model and the right choice of method depends on the eventual implementation.
- When evaluating ML applications for clinical implementation is important to understand 1) the **quality of the reference** standard, 2) the **independence** of and **generalizability** of the testing set, and 3) whether the ML model can be **realistically implemented** in a research setting or in clinical practice.

Acknowledgements

Saba Lab:

- Current: Cheyret Wood, Samuel Rosean, Keenan Manpearl, Jack Pattee and Angela Yoder
- Former: Lauren Vanderlinden, Harry Smith, Ryan Lusk, and Sean Hickey

Boris Tabakoff, Paula Hoffman, and their lab:

- Spencer Mahaffey and Jenny Mahaffey

Financial Support:

- NIDA Core “Center of Excellence" in Omics, Systems Genetics and the Addictome (NIDA - P30DA044223; MPIs - Williams, Saba)
- The heritable transcriptome and alcoholism (NIAAA - R24AA013162; MPIs - Tabakoff, Hoffman, Saba
- Identification of genes and genetic networks contributing to opioid use disorder traits in the Hybrid Rat Diversity Panel (NIDA - U01DA051937; MPIs – Ehringer, Bachtell, Saba)
- NIEHS, NEI, NIAMS, NIBIB, NHLBI, NIDA, NIAAA, and Skaggs Scholars

Looking for Computational Bioscience/Statistical Genetics Post Docs!

Clinical Review & Education

JAMA | Users' Guides to the Medical Literature

How to Read Articles That Use Machine Learning
Users' Guides to the Medical Literature

Yun Liu, PhD; Po-Hsuan Cameron Chen, PhD; Jonathan Krause, PhD; Lily Peng, MD, PhD

In recent years, many new clinical diagnostic tools have been developed using complicated machine learning methods. Irrespective of how a diagnostic tool is derived, it must be evaluated using a 3-step process of deriving, validating, and establishing the clinical effectiveness of the tool. Machine learning–based tools should also be assessed for the type of machine learning model used and its appropriateness for the input data type and data set size. Machine learning models also generally have additional prespecified settings called hyperparameters, which must be tuned on a data set independent of the validation set. On the validation set, the outcome against which the model is evaluated is termed the reference standard. The rigor of the reference standard must be assessed, such as against a universally accepted gold standard or expert grading.

JAMA. 2019;322(18):1806-1816. doi:10.1001/jama.2019.16489

Viewpoint page 1765 and **Editorial** page 1777

Supplemental content

CME Quiz at jamanetwork.com/learning

Author Affiliations: Google Health, Palo Alto, California.

Corresponding Author: Yun Liu, PhD, Google Health, 3400 Hillview Ave, Palo Alto, CA 94304 (liuyun@google.com).

The NEW ENGLAND JOURNAL of MEDICINE

REVIEW ARTICLE

FRONTIERS IN MEDICINE

Machine Learning in Medicine

Alvin Rajkomar, M.D., Jeffrey Dean, Ph.D., and Isaac Kohane, M.D., Ph.D.

A 49-year-old patient notices a painless rash on his shoulder but does not seek care. Months later, his wife asks him to see a doctor, who diagnoses a seborrheic keratosis. Later, when the patient undergoes a screening colonoscopy, a nurse notices a dark macule on his shoulder and advises him to have it evaluated. One month later, the patient sees a dermatologist, who obtains a biopsy specimen of the lesion. The findings reveal a noncancerous pigmented lesion. Still concerned, the dermatologist requests a second reading of the biopsy specimen, and invasive melanoma is diagnosed. An oncologist initiates treatment with systemic chemotherapy. A physician friend asks the patient why he is not receiving immunotherapy.

Machine learning applications in genetics and genomics

Maxwell W. Libbrecht[1] *and William Stafford Noble*[1,2]

Abstract | The field of machine learning, which aims to develop computer algorithms that improve with experience, holds promise to enable computers to assist humans in the analysis of large, complex data sets. Here, we provide an overview of machine learning applications for the analysis of genome sequencing data sets, including the annotation of sequence elements and epigenetic, proteomic or metabolomic data. We present considerations and recurrent challenges in the application of supervised, semi-supervised and unsupervised machine learning methods, as well as of generative and discriminative modelling approaches. We provide general guidelines to assist in the selection of these machine learning methods and their practical application for the analysis of genetic and genomic data sets.

Machine learning
A field concerned with the development and application of computer algorithms that improve with experience.

[1]*Department of Computer Science and Engineering, University of Washington, 185 Stevens Way, Seattle, Washington 98195–2350, USA.*
[2]*Department of Genome Sciences, University of Washington, 3720 15th Ave NE Seattle, Washington 98195–5065, USA.*
Correspondence to W.S.N.
e-mail: william-noble@uw.edu
doi:10.1038/nrg3920
Published online 7 May 2015

The field of machine learning is concerned with the development and application of computer algorithms that improve with experience[1]. Machine learning methods have been applied to a broad range of areas within genetics and genomics. Machine learning is perhaps most useful for the interpretation of large genomic data sets and has been used to annotate a wide variety of genomic sequence elements. For example, machine learning methods can be used to 'learn' how to recognize the locations of transcription start sites (TSSs) in a genome sequence[2]. Algorithms can similarly be trained to identify splice sites[3], promoters[4], enhancers[5] or positioned nucleosomes[6]. In general, if one can compile a list of sequence elements of a given type, then a machine learning method can probably be trained to recognize those elements. Furthermore, models that each recognize an individual type of genomic element can be combined, along with (learned) logic about their relative locations, to build machine learning systems that are capable of annotating genes — including their untranslated regions (UTRs), introns and exons — along entire eukaryotic chromosomes[7].

As well as learning to recognize patterns in DNA sequences, machine learning algorithms can use input data generated by other genomic assays — for example, microarray or RNA sequencing (RNA-seq) expression data; data from chromatin accessibility assays such as DNase I hypersensitive site sequencing (DNase-seq), micrococcal nuclease digestion followed by sequencing (MNase–seq) and formaldehyde-assisted isolation of regulatory elements followed by sequencing (FAIRE–seq); or chromatin immunoprecipitation followed by sequencing (ChIP–seq) data of histone modification or transcription factor binding. Gene expression data can be used to learn to distinguish between different disease phenotypes and, in the process, to identify potentially valuable disease biomarkers. Chromatin data can be used, for example, to annotate the genome in an unsupervised manner, thereby potentially enabling the identification of new classes of functional elements.

Machine learning applications have also been extensively used to assign functional annotations to genes. Such annotations most frequently take the form of Gene Ontology term assignments[8]. Input of predictive algorithms can be any one or more of a wide variety of data types, including the genomic sequence; gene expression profiles across various experimental conditions or phenotypes; protein–protein interaction data; synthetic lethality data; open chromatin data; and ChIP–seq data of histone modification or transcription factor binding. As an alternative to Gene Ontology term prediction, some predictors instead identify co-functional relationships, in which the machine learning method outputs a network in which genes are represented as nodes and an edge between two genes indicates that they have a common function[9].

Finally, a wide variety of machine learning methods have been developed to help to understand the mechanisms underlying gene expression. Some techniques aim to predict the expression of a gene on the basis of