

Machine Learning in R to Predict Outcomes from Biomarker Data

Prakash Dev, M.B.B.S., M.B.A.
Benno Kurch, B.A.

labcorp
Drug Development

Agenda

- What is the problem we are trying to solve?
 - Vast amount of biomarker data can't be effectively analyzed by statistical modeling techniques
- What are biomarkers?
 - Unique characteristics
- Can Machine Learning solve the problem?
 - Gain new insights post-hoc from data
 - Transform data into actionable knowledge
 - Create value for organization – biomarker discovery and validation
- Different Machine Learning techniques
 - Supervised and Unsupervised
- Performance metrics for the prediction
 - Lift, Cost function, Receiver Operating Characteristic (ROC)

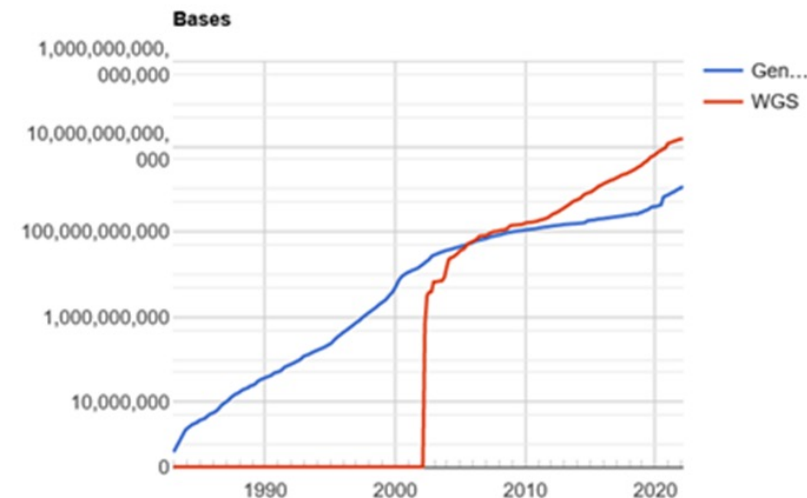
What problem are we trying to solve?

- Data production outpaces human ability to analyze the data
- Decision Support Systems generally utilize static algorithms and do not seek out new information
- Massive amount of biomarker data from drug discovery and Real-World Evidence (RWE)
 - Omics data explosion since completion of Genome Project
 - Platforms that measure thousands of genes at once
 - Duration for testing has become shorter
 - More data produced from smaller amount of samples
 - Significant investments in newer technologies have added to data volume to analyze
 - Focus has shifted from gene identification to its quantitation in cellular microenvironment
 - e.g., If TNFa associated with Inflammation → How much TNFa is suppressed or increased with treatment
- Statistical modeling is usually not suited for biomarker data analysis
 - No statistical distribution can be assumed
 - Predictor variables may be highly correlated

What problem are we trying to solve?

- Omics data explosion
 - Extensive adoption of new technologies
 - Genome-wide abundance of transcripts
 - Intensity-based proteomics data
- Technologies over past decades
 - Next Generation Sequencing (NGS)
 - RNASeq
 - Multiplex assays
 - MSD, Luminex
 - Microarrays
 - cDNA, SNP
 - Proteomics
 - MS-based, e.g., MALDI-TOF, LC, SomaScan

- GenBank Data Repository*
 - 2022/04: >1.25 trillion base pairs(x100), >237 million nucleotide sequences(
 - 2000/12: ~11 billion base pairs, ~10 million nucleotide sequences



* Source: <https://www.ncbi.nlm.nih.gov/genbank/statistics/>

Biomarkers

- FDA/NIH definition:

“A defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions”
- Can be utilized as surrogate endpoints
- Categories
 - Predictive Biomarkers: e.g., TMB for efficacy of immune checkpoint inhibitors
 - Identify individuals who are more likely to experience a favorable or unfavorable effect from exposure to a medical product
 - Prognostic Biomarkers – PSA (prostate cancer)
 - Identify likelihood of a clinical event in patients who have the medical condition
 - Response/Pharmacodynamic Biomarkers – INR (Warfarin therapy)
 - Show that a biological response has occurred in an individual who has been exposed to a medical product

Types of Biomarker Data in Basic Research

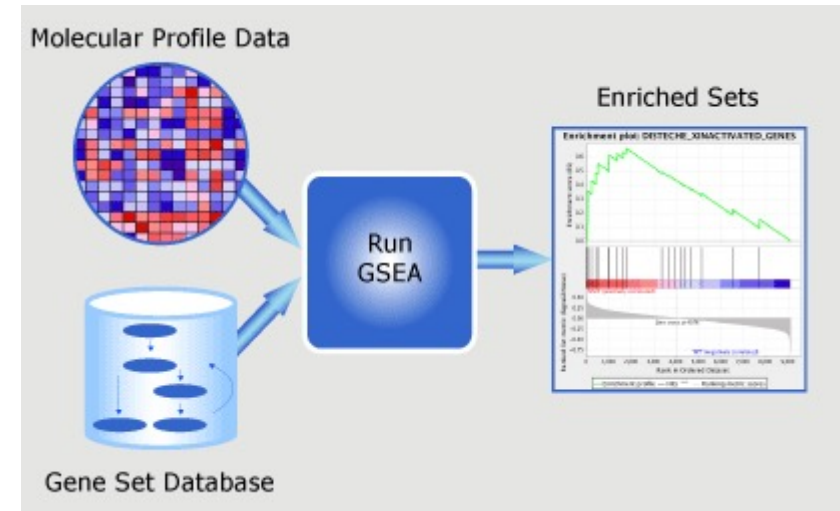
- Gene products – nucleotides, transcripts, peptides
- Circulating small molecules
 - Cell leakage products
- Sequence Variations/Genotypes/Haplotypes
- Gene expression: Transcripts/Transcriptome
- Metabolomics: Peptides/Proteins/Proteome

Complexities of Biomarker Data

- High dimensionality/too many predictor variables
- Background noise/house keeping genes
- Collinearity/interactions
- Data non-normality
- Numerous data formats – sequence (FASTA, FASTQ), gene expression (Illumina-TXI), peptides (MS, LC, SomaScan)
- Lack of standards – across source data from vendors, systems
- Need for specialized software
 - Vendor-specific LIMS

Current Approaches to Analyze Biomarker Data

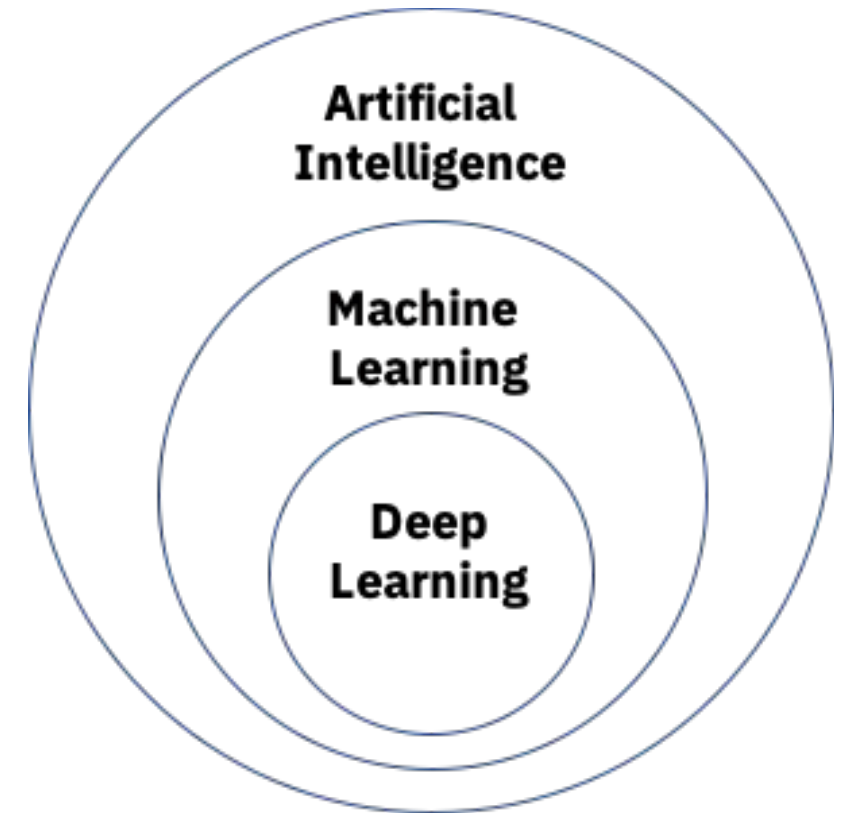
- Analysis of gene sets
 - Enrichment analysis
- Individual gene effects
 - Linear models – LIMMA methodology
- Often depends upon visual detection of effects
 - Heat maps/Volcano plots/box plots
- Few genes analyzed – loss of information
- Focused on exploratory end points
 - Association between treatment and gene(s) or pathway(s)



[Broad Institute \(gsea-msigdb.org\)](https://www.broadinstitute.org/gsea-msigdb.org/gsea/)

Machine Learning (ML)

- Computer Program (machine) that can Learn from data without explicit instructions
- Performance of the machine improves as it is exposed to more data
- Focuses on entirety of the data
- No need to design experiments
- Minimal assumptions about data distribution
- ML Categories
 - Supervised Learning – Task driven
 - Unsupervised Learning – Data driven
 - Reinforcement Learning – Learning from mistake



Supervised Machine Learning

- Machine Learning using labelled dataset
- Both Inputs (predictors) and output (response/target) variables are present
- Classification and Regression Problem
- Build statistical model for estimating (predicting) an outcome (Y) based on one or more inputs (X)

$$Y = f(X) + e$$

[f = systematic information, e = random error]

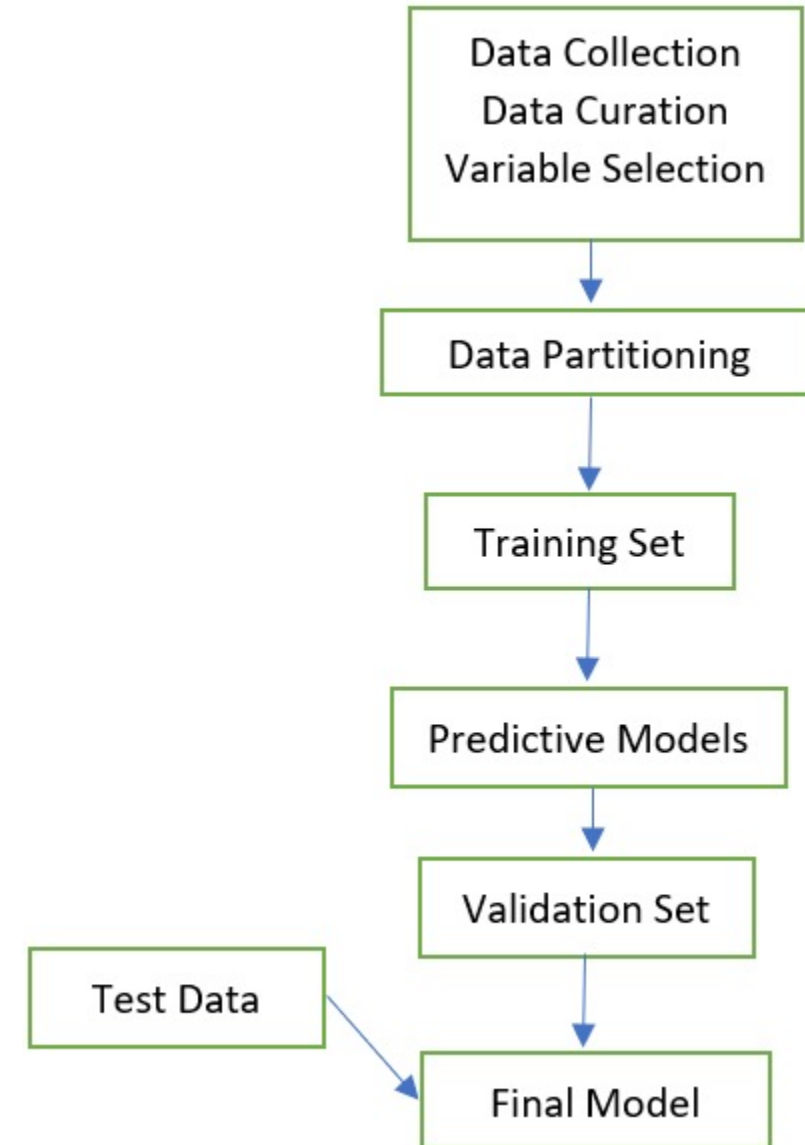
- Examples:
 - Decision Tree
 - Regression – linear/logistic
 - SVM
 - Neural Network

Unsupervised ML: Pattern Recognition

- No explicit target labels - no outcome variables
- No a priori hypothesis or model
- Clustering and Association problem
- Hypothesis formulated post-hoc from the results
- Examples
 - K-means Clustering
 - Principal Component Analysis
 - Nearest Neighbor Mapping

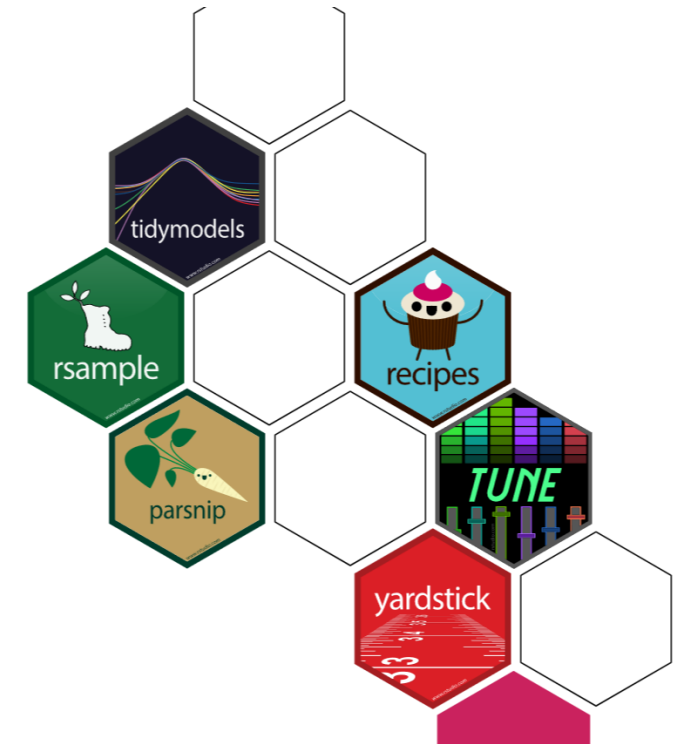
Machine Learning Workflow

- Prepare data
 - Variable selection
 - Transformation
- Split the data into partitions
 - Train (60%)
 - Validation (20-30%)
 - Test (10-20%)
- Fit a model and evaluate
- Fine tune hyperparameters
- Evaluate on validation set
- Select best model and evaluate the final model on test set



Machine Learning – Considerations

- Tradeoffs: prediction accuracy and model interpretability
 - Restrictive approach – easier interpretation, e.g., linear model
 - Flexible approach – hard to ascertain relationships, e.g., spline
- Performance
 - Training – data/computing resources
 - Predicting – accuracy
- Measuring the Quality of Fit
 - Good fit vs. Over-fit/Under-fit
- Parsimony – Fewer important variables
- Black-box – All available predictor variables



Tidymodels R packages:
<https://www.tidymodels.org/>

Predictive Model: Logistic Regression

- Type of regression when the target variable is binary
- Link function
 - Transformation of target variable
 - Logit function
- Model parameters and optimization
 - Link function: logit, probit, loglog
- Fit statistics:
 - ASE, Misclassification rate

R function: glm

```
fit1 <- glm(formula= EP ~ feature1 + feature2+..+featureN,  
data=bmk, family=binomial)
```

```
Summary(fit1)
```

Model: Decision Tree

- Hierarchical segmentation of the data by applying rules
- Consists of parent node and child (leaf) nodes
- Leaf nodes have posterior probability
- The set of rules do not have any equations or coefficients

```
library(tree)
```

```
dt1 <- tree(EP ~ feature1+feature2+...+featureN,  
data=bmk)
```

```
summary(dt1)
```

```
plot(dt1)
```

```
text(dt1)
```

Model: Random Forest

- Merges a collection of independent decision trees
- Aggregates the predictions from individual trees
- Uses majority voting for final prediction
- Fast and high performing

```
library(randomForest)
```

```
rf1 <- randomForest(Status ~ AgeAtStart + Height +  
Weight + Diastolic + Systolic + MRW + Smoking +  
Cholesterol, data=heart1)
```

```
rf1
```

```
pred <- predict(rf1, newdata=testing)
```

```
cm <- table(testing$Status,pred)
```

```
Testing1 <- testing  
testing1$Status[1] <- "Dead"
```

```
cm1 <- table(testing1$Status,pred)
```

Boosting – Using R Programming

- Combine several weak models to create one strong model
 - Weak = random guessing
 - Strong = true classification
- Learn from each iteration
- Weighted Voting

```
library(gbm)
```

```
boost1 = gbm(Status ~ AgeAtStart + Height + Weight + Diastolic +  
Systolic + MRW + Smoking + Cholesterol, data=heart1, distribution =  
"gaussian", n.trees = 10000, shrinkage = 0.01, interaction.depth = 4)
```

```
summary(boost1)
```

```
# Relative significance of AgeAtStart, Cholesterol
```

```
plot(boost1, i="AgeAtStart")
```

```
plot(boost1, i="Cholesterol")
```

Ensemble Model

- Combines several machine learning models into single model
- Retains results from best performing models
- Decreases variance across different models
- Utilizes several techniques to improve performance
 - Boosting
 - Bagging
 - Stacking

Scoring

- Process of predicting values based on a trained machine learning model
 - Uses test data
 - Generates metrics to measure performance
- Performed on a test (new) dataset – not previously exposed to model
- Performance measures from confusion matrix
 - Accuracy = $(TP+TN)/(TP+FP+FN+TN)$
 - Precision = $(TP/TP+FP)$
 - Recall (Sensitivity) = $(TP/TP+FN)$

		TRUE CLASS	
		Positive	Negative
PREDICTED CLASS	Positive	TP	FP
	Negative	FN	TN

TP: True Positive, TN: True Negative
FP: False Positive, FN: False Negative

Conclusions

- Massive growth in biomarkers data from basic and early clinical research outpaces capacity for exploratory analysis due to time and resource limitation
- Complex Omics data can take advantage of Machine Learning techniques
- Numerous tools are available to optimize Machine Learning models and improve performance of models
- Machine Learning provides an efficient method of predicting associations, trends and patterns in biomarker data

References

- Tidymodels - <https://www.tidymodels.org/>
- Biomarker Collaborative - <https://biomarkercollaborative.org/>
- Machine Learning | JAMA Network - <https://jamanetwork.com/channels/machine-learning>
- Lantz, Brett. Machine Learning with R. PACKT Publishing, 2013