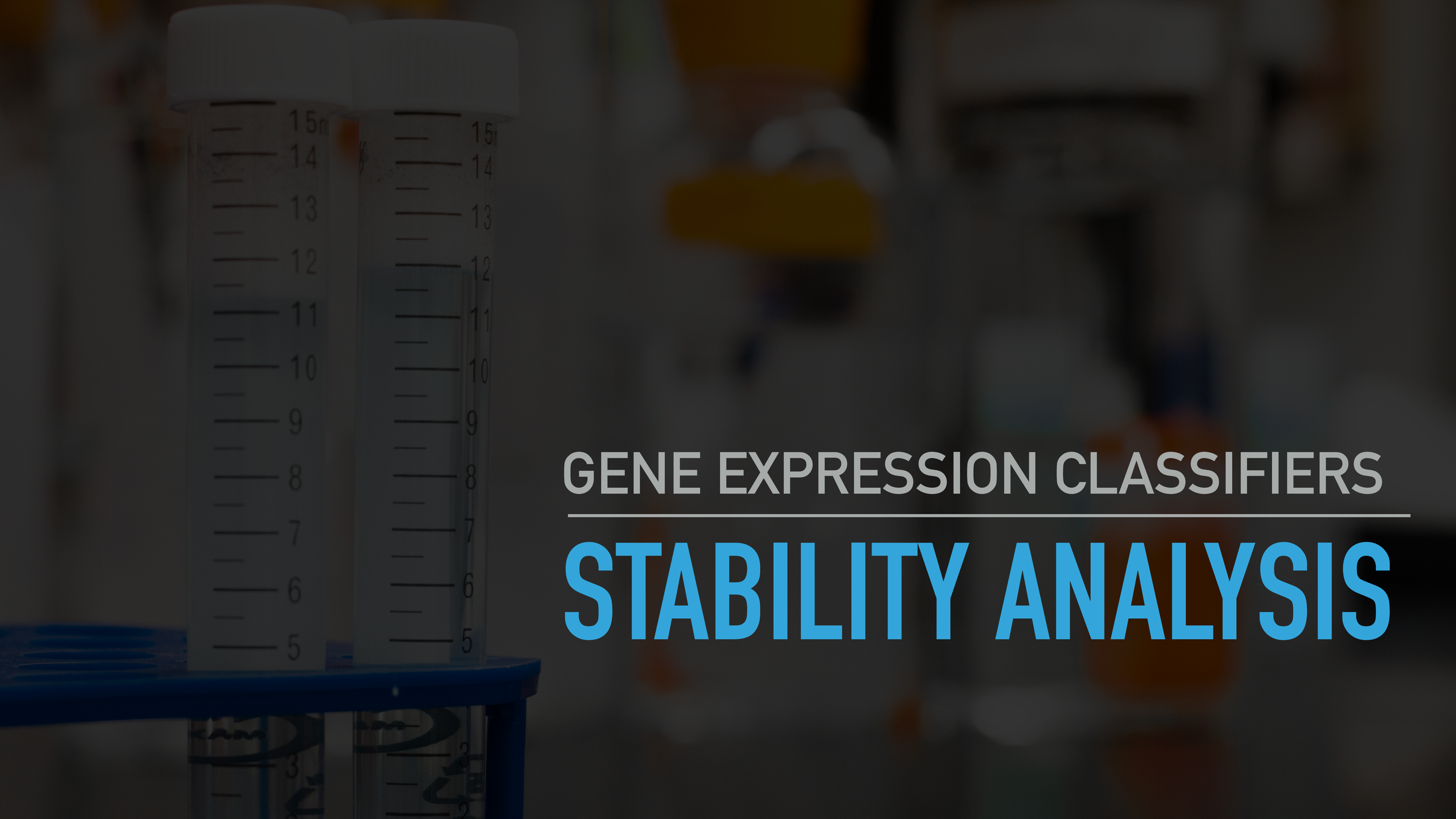
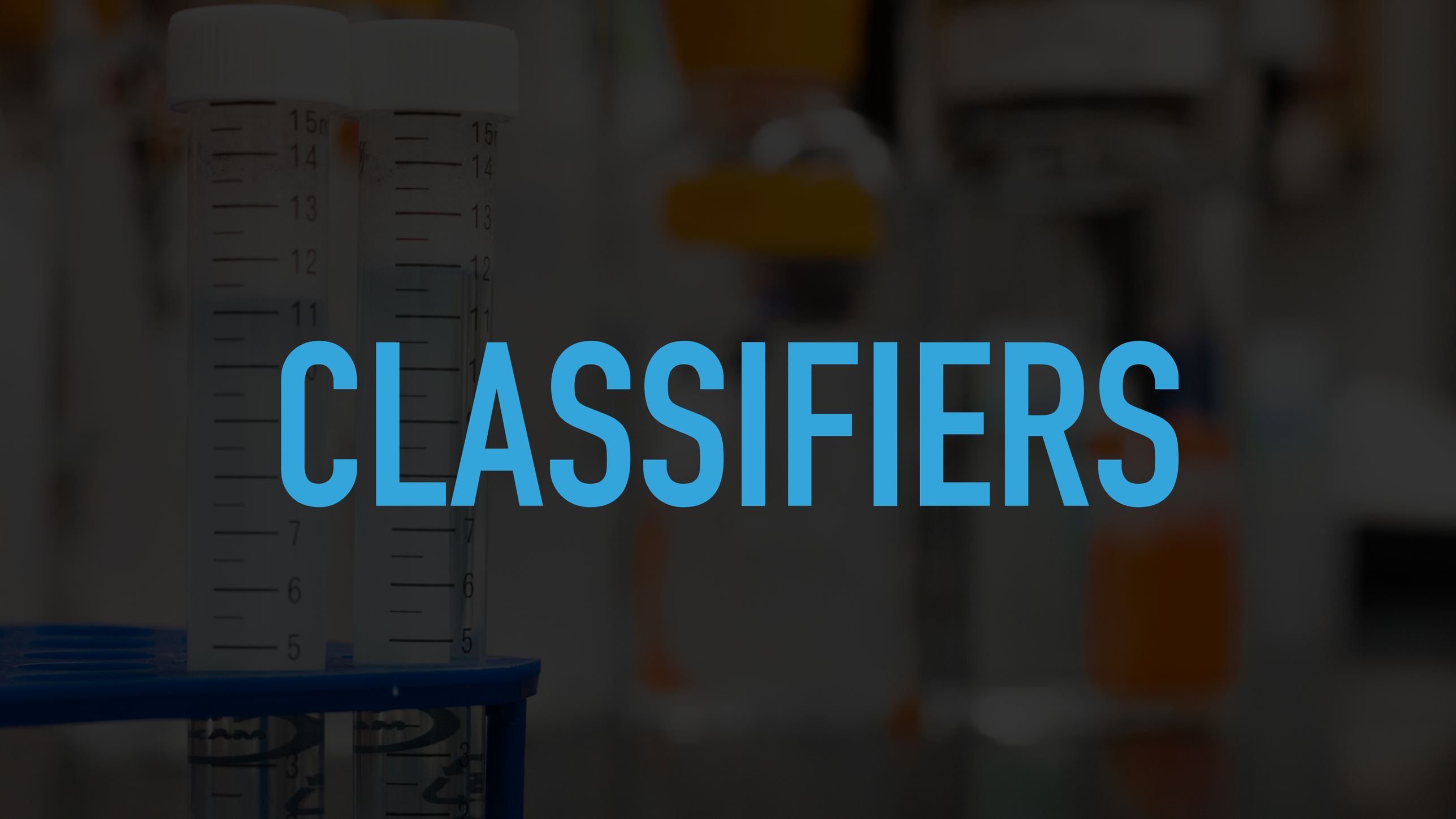
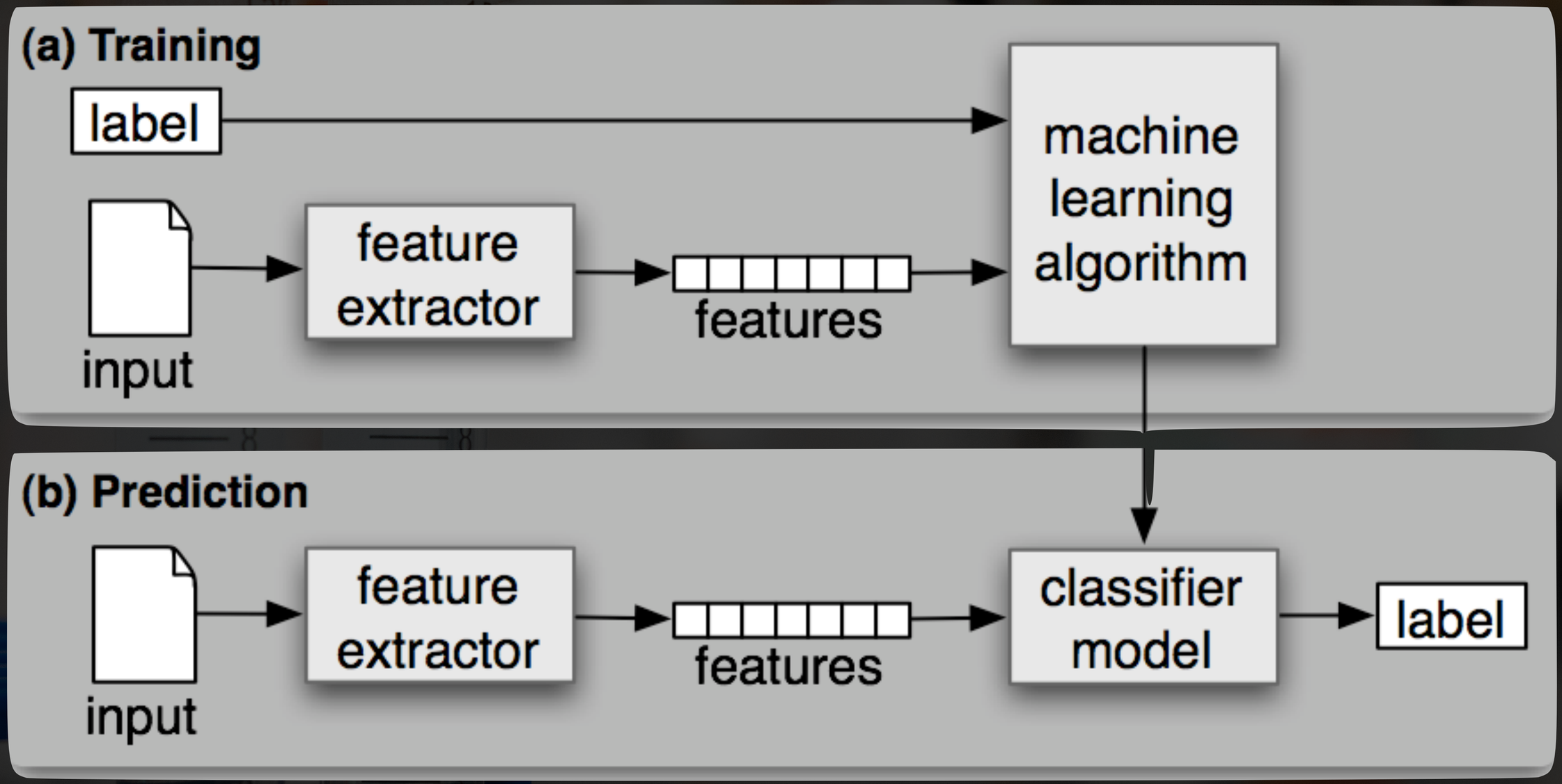


GENE EXPRESSION CLASSIFIERS

STABILITY ANALYSIS


CLASSIFIERS

GENE EXPRESSION CLASSIFIER

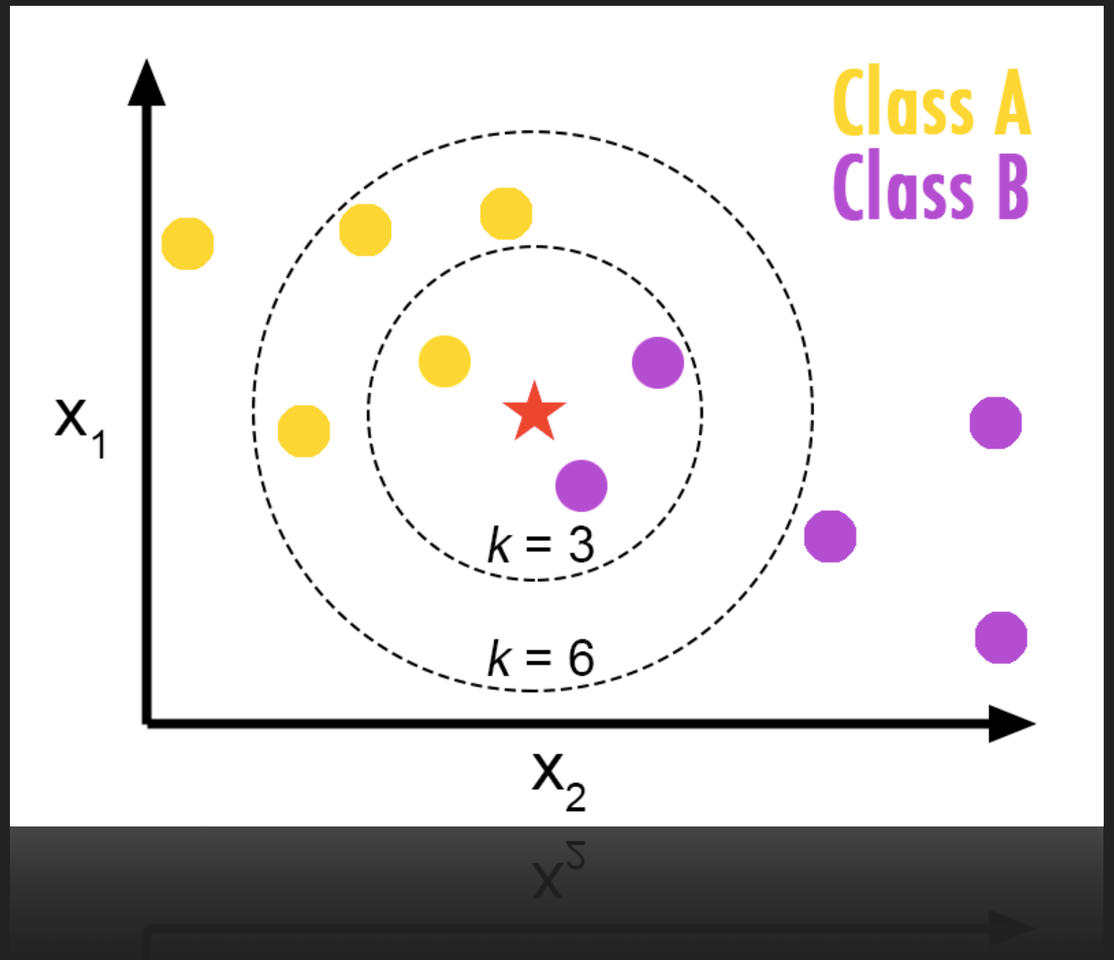


CHI-SQUARED

- select a subset of relevant features (genes) for use in model construction
  - simplification of model = easier to interpret
  - shorter run time
  - reduce overfitting and under-fitting of training data
- how likely it is that any observed difference between the sets arose by chance
  - calculate chi2 statistic for each sample, selecting top k features
  - Typical k ranges between 50 and 300, for this study k = 50

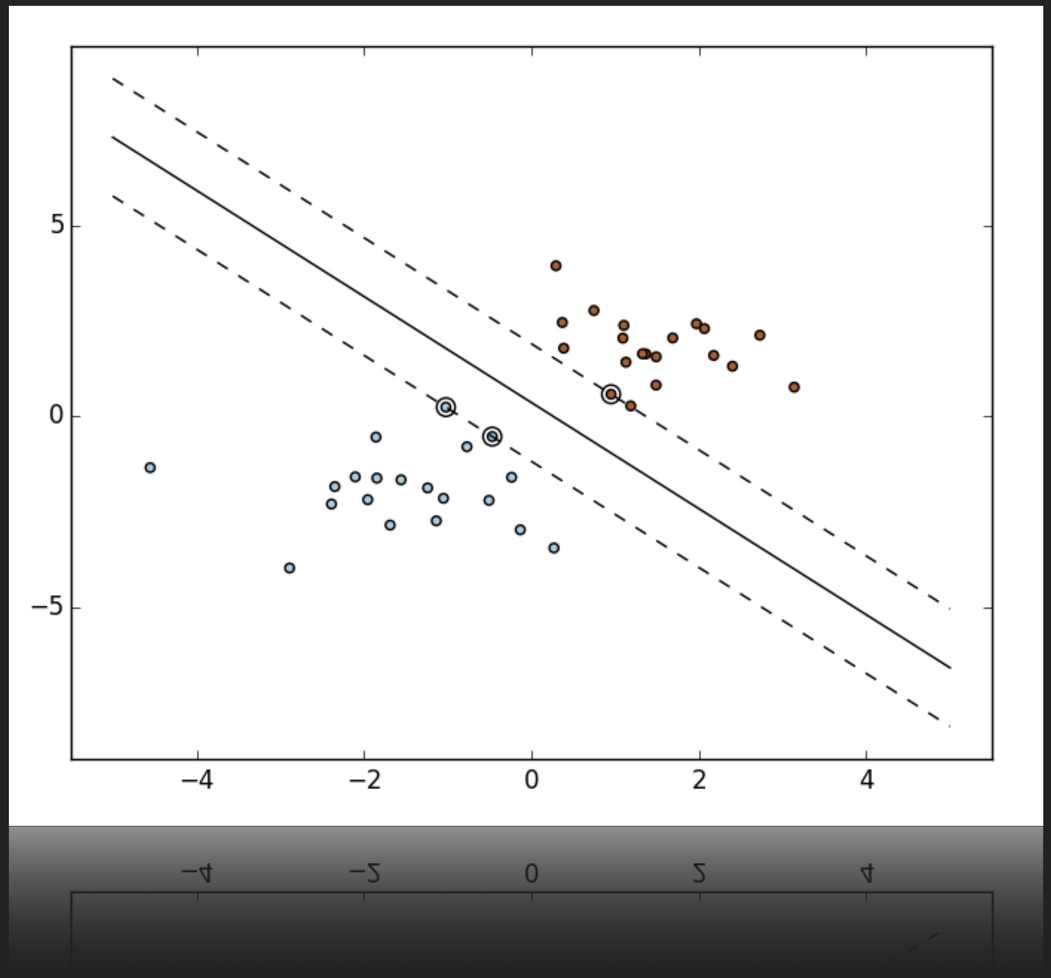
K-NEAREST NEIGHBORS (KNN)

- sample is classified by majority vote of its k nearest neighbors
- if k=1, sample is assigned to nearest neighbor
- among simplest of all machine learning algorithms
- for all stability tests, k=1



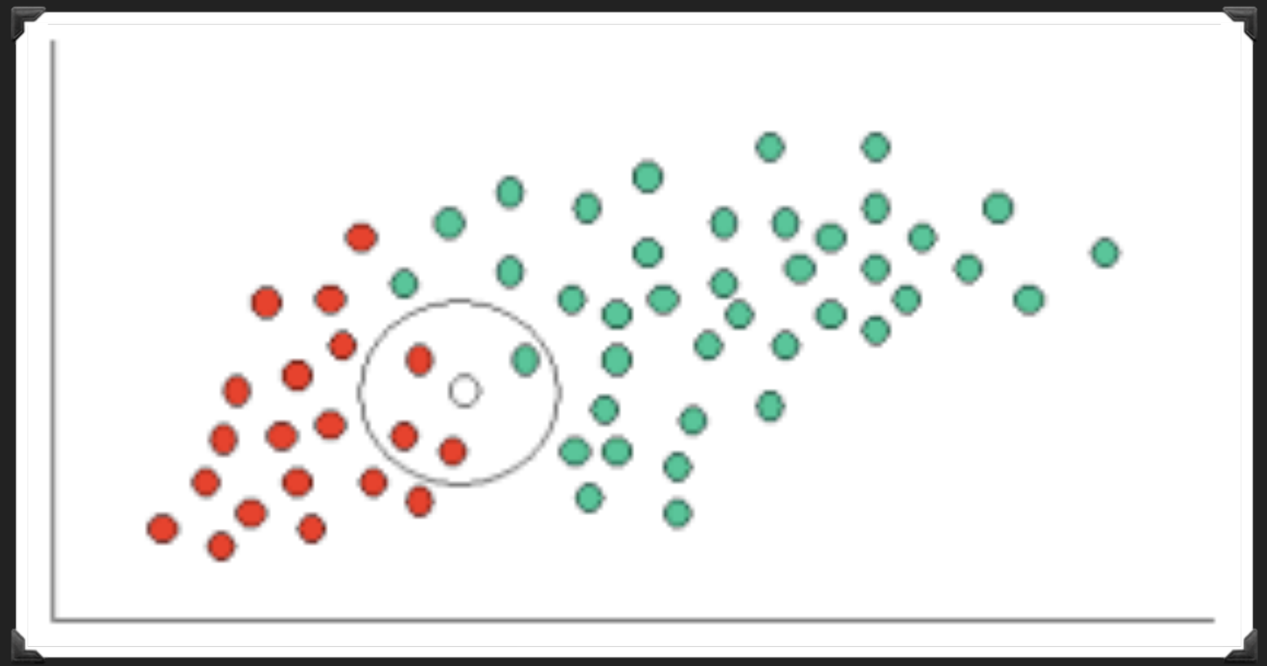
SUPPORT VECTOR MACHINE (SVM)

- constructs a set of hyper-planes used for classification
- good separation is achieved by the hyper-plane that has the largest distance to the nearest training data points of any class
- points that lie closest to this max-margin hyperplane are called the support vectors



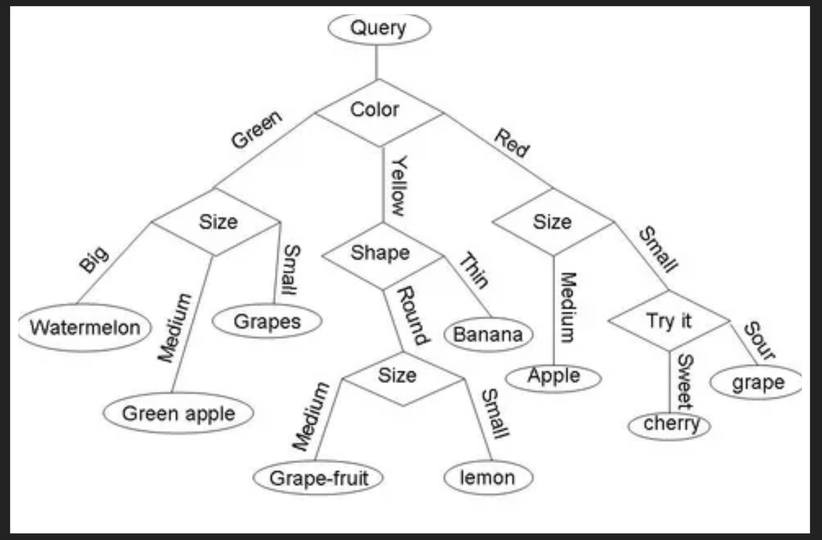
NAIVE BAYES (NB)

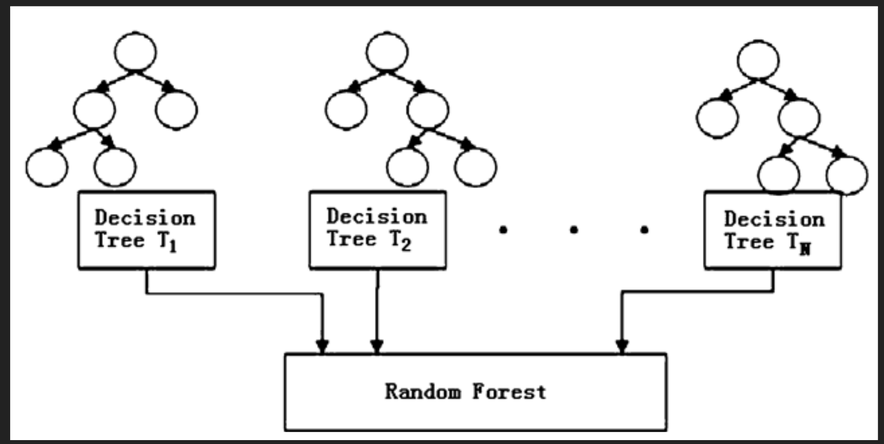
- creates rules based on Bayes' theorem
- uses probabilistic induction to assign class labels to test samples, assuming independence among the features
- probability and likelihood
- simplified assumptions



RANDOM FOREST (RF)

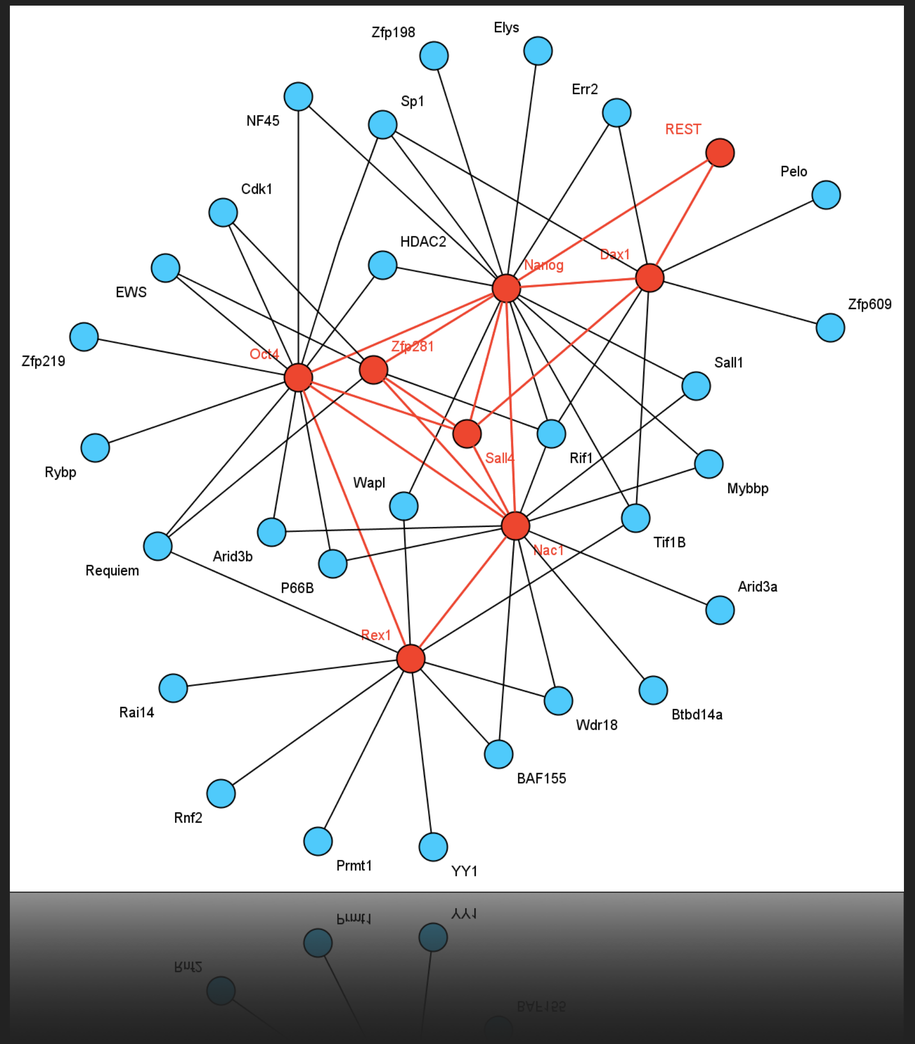
- builds multiple decision trees
  - internal nodes = splitting criterion
  - leaf nodes = class label
- test samples classified by assigning to class that takes majority vote over all decision trees

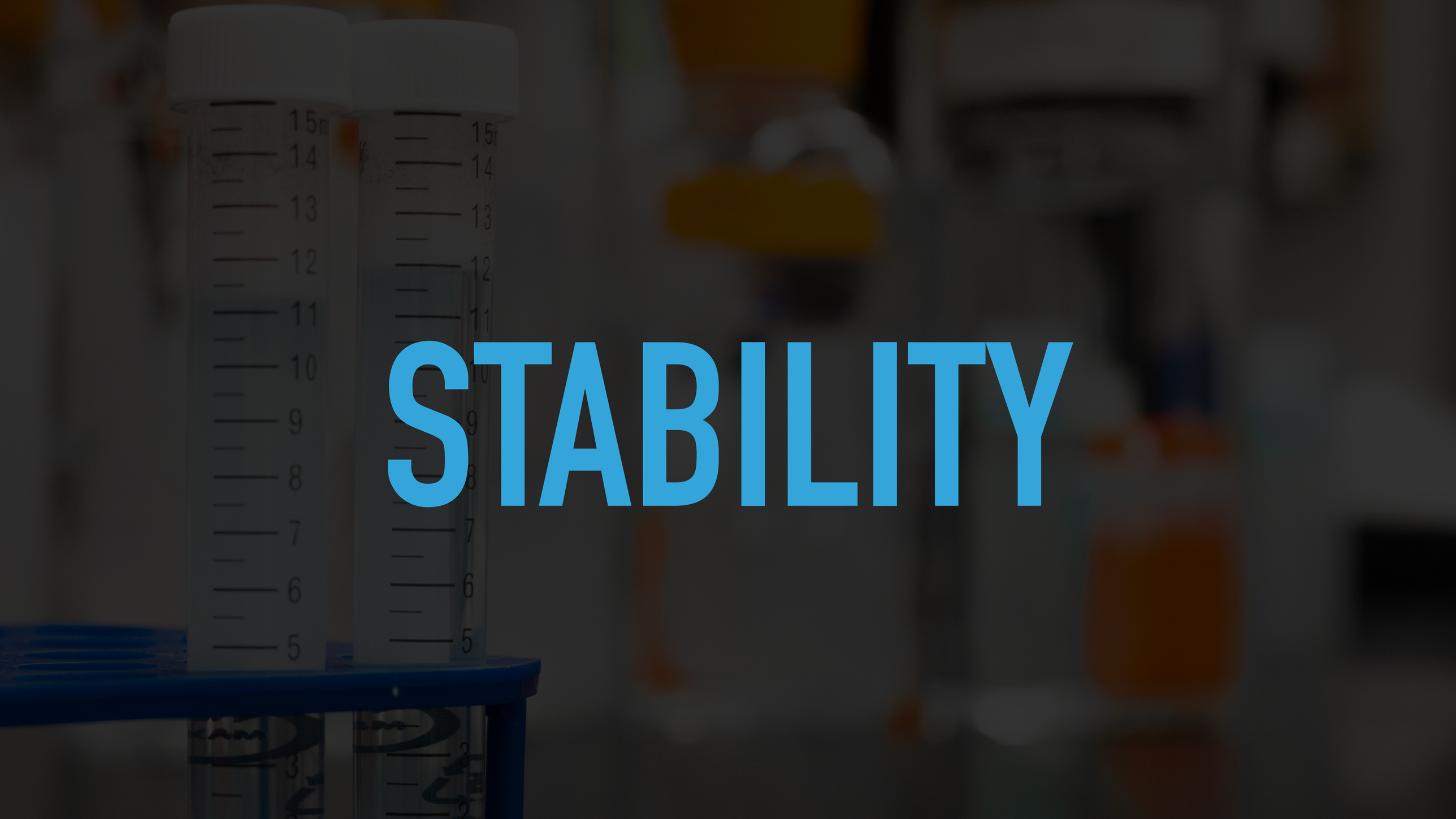




NETWORK-BASED CLASSIFIER (NBC)

- leverages the underlying gene network
- constructs hypothetical expression from neighbors
  - $f(g) = x_1*g_1 + \ldots + x_k*g_k$
- lowest root mean square error is class
  - predicted - actual




STABILITY

TESTING STABILITY

- how a machine learning algorithm performs if data is perturbed
  - looking for robust algorithms - does not produce a wildly different result for very small change in the input data
- types of perturbation considered here
  - alter every feature (gene) by some amount (test I)
  - alter a subset of features by some amount (test II-IV)
    - best way to select the features?

TEST I – CHANGE ALL GENES

- change all genes by some amount
  - p = {0.05, 0.10, … 0.95, 1.00}
  - For each gene, g
    - g’ = rand(g - g*p , g + g*p)
- represents an error in calibration of equipment
- subject to bias

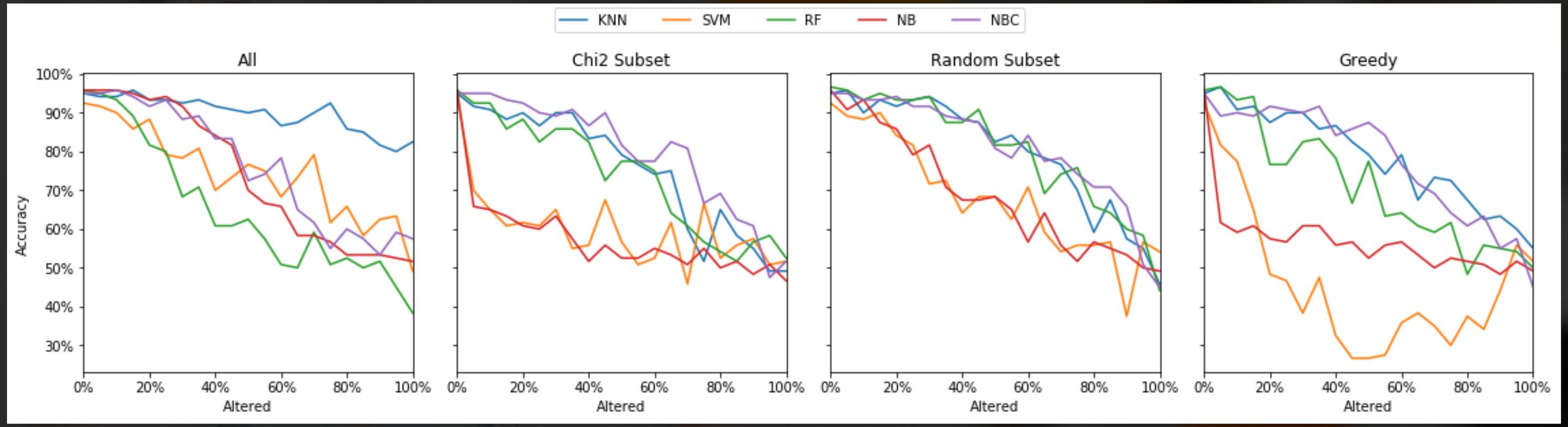
TEST II AND III – RANDOM/CHI2 SUBSET

- test II: pick a subset of genes randomly
- test III: pick a subset of genes according to chi2 score
- for both tests
  - p = {0.05, 0.10, … 0.95, 1.00}
  - for subset of genes of percent size p
    - g’ = rand(min, max)
- represents an error in reading data

TEST IV – GREEDY

- choose subset greedily
  - select next gene that produces minimum accuracy
- allows determination of minimum number of genes to make algorithm unstable
- represents a “smart” adversary
- not necessarily only solution for selected genes to make algorithm unstable
- could determine which genes are most important in disease

RESULTS


KNN
SVM
RF
NB
NBC
All
Chi2 Subset
Random Subset
Greedy
Accuracy
100%
90%
80%
70%
60%
50%
40%
30%
0%
20%
40%
60%
80%
100%
Altered

TEST I – CHANGE ALL GENES

- kNN is most stable under this condition
  - could indicate that data is well separated
  - could also be byproduct of testing method - relationships maintained
- RF is most unstable under this condition
  - could be due to interdependence of data
  - may take into consideration factors that are not as relevant
    - not many genes are correlated for a disease
    - probability doesn’t matter for gene expression of disease

TEST II AND III – RANDOM/CHI2 SUBSET

- kNN, RF, NBC are most stable under this condition
  - chi2 **does not** perform significantly better than random choice
- NB and SVM are most unstable under this condition
  - could be reflection of instability of single gene changes
  - chi2 **does** perform significantly better than random choice

TEST IV – GREEDY

- kNN, RF, NBC are most stable under this condition
  - NBC also performs well under sub-stable testing.
- NB and SVM are most unstable under this condition
  - could be reflection of instability of single gene changes
- Tests demonstrate that kNN performs the best under all conditions.