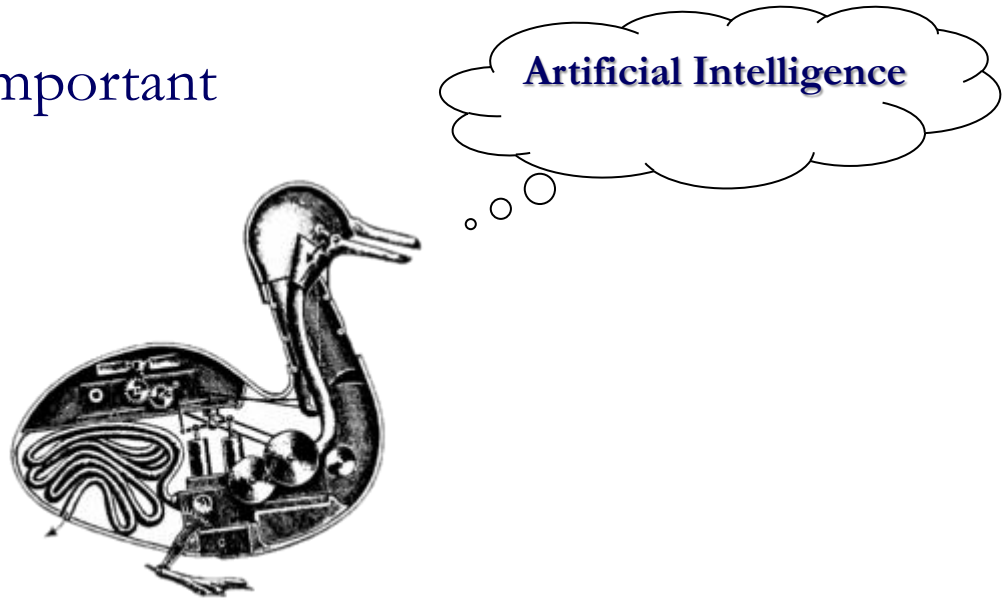
Machine learning applications in biotechnology research

UC Davis Biotechnology Program 2012
Davis, CA

Tobias Kind
UC Davis Genome Center
FiehnLab - Metabolomics

Machine learning

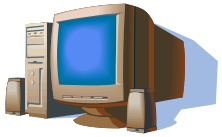
- Machine learning commonly used for prediction of future values
- Complex models (black box) do not always provide causal insight
- Predictive power is most important



Picture Source: Wikipedia

Why Machine Learning?

People cost money, are slow, don't have time



(1984)

Let the machine (computer) do it...

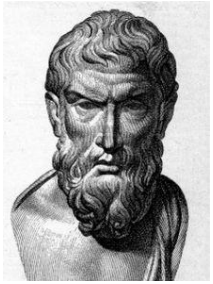
???: 42

Replace people with computers...

The Terminator (2029)

Picture Source: Wikipedia; Office; RottenTomatoes; Wikimedia-Commons

Machine Learning Personalities



Epicurus
(341 BC)

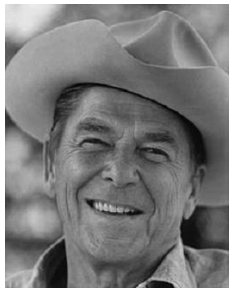
Epicurus:
Principle of multiple explanations
"All consistent models should be retained".

Lenin*

Trust is good, control is better

Ockham
(1288)

Occam's Razor:
Of two equivalent theories or explanations, all other things being equal, the simpler one is to be preferred.



Ronald Reagan*

Trust, but verify

Alan Turing
(1912)

Turing Test
Can machines think?

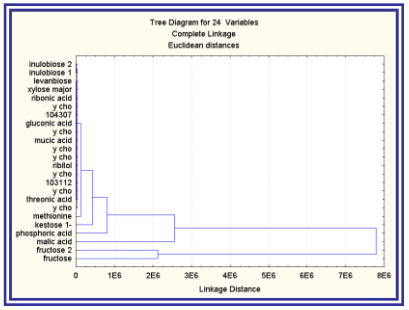
Marvin Minsky
(1927)

Artificial intelligence, Neural networks

Sources: Wikipedia;

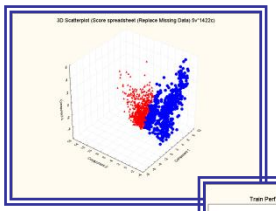
(*) In principle

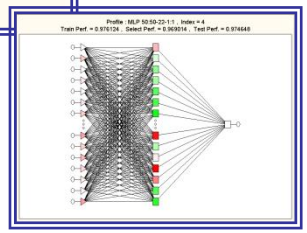
Machine Learning Algorithms

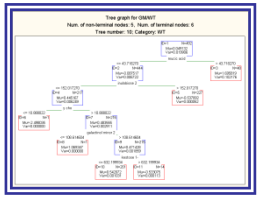


Unsupervised learning: Clustering methods

Supervised learning: Support vector machines
MARS (multivariate adaptive regression splines)
Neural networks
Naive Bayes classifier





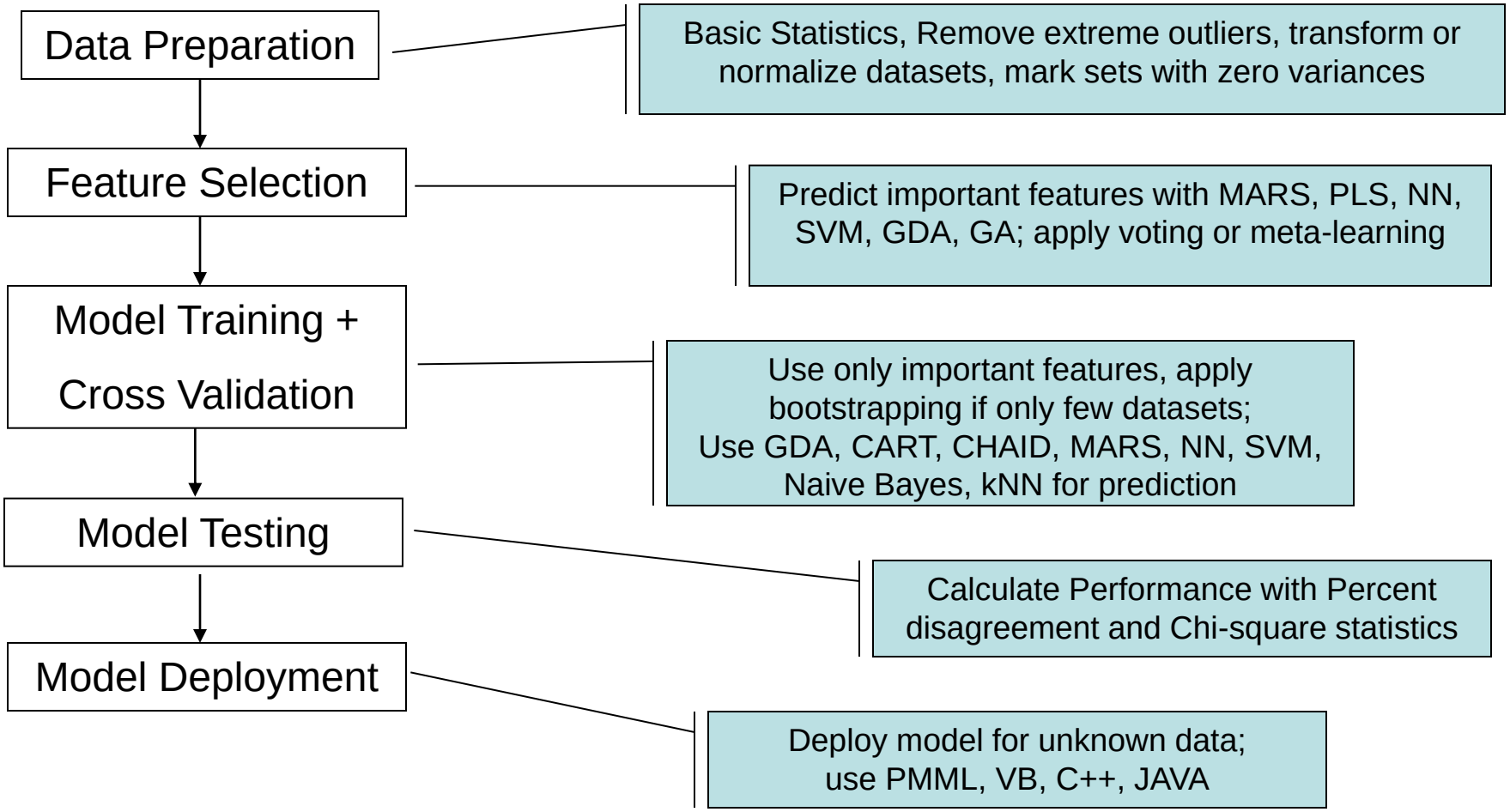


Random Forest, Boosting trees, Honest trees, Decision trees
CART (Classification and regression trees)
Genetic programming

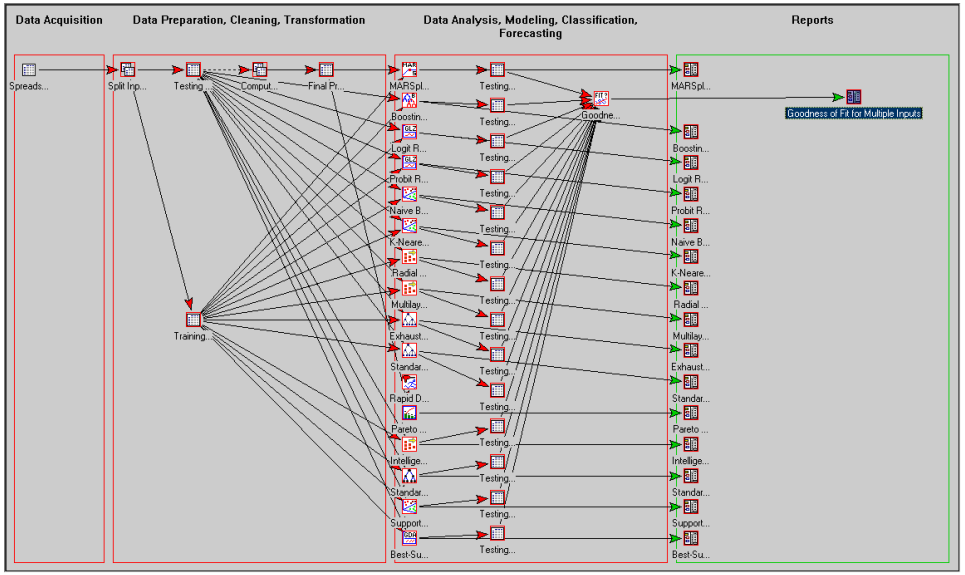
Transduction: Bayesian Committee Machine
Transductive Support Vector Machine

...thanks to WIKI and Stuart Gansky

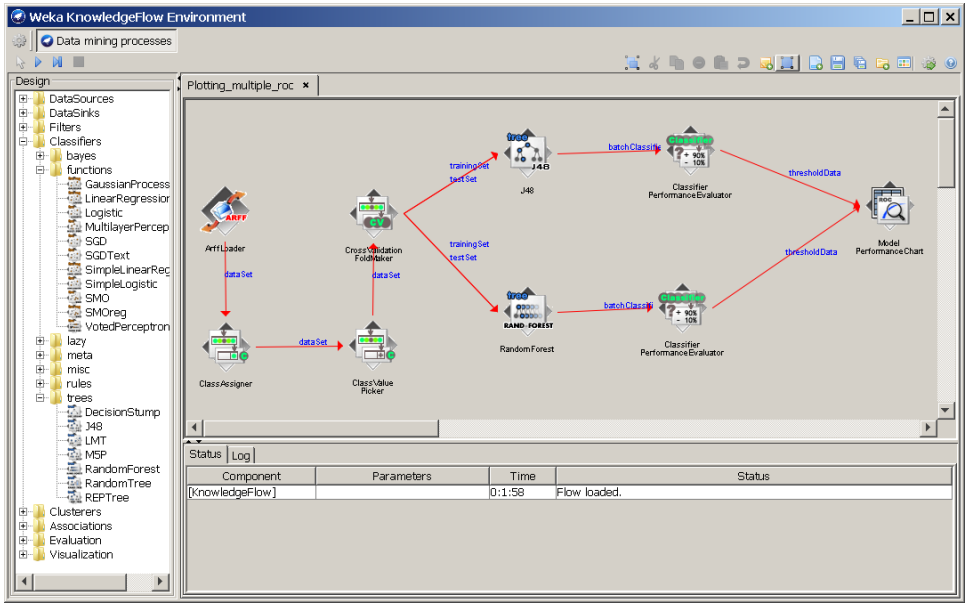
Concept of predictive data mining for classification



Automated machine learning workflows – tools of the trade



Statistica Dataminer workflow

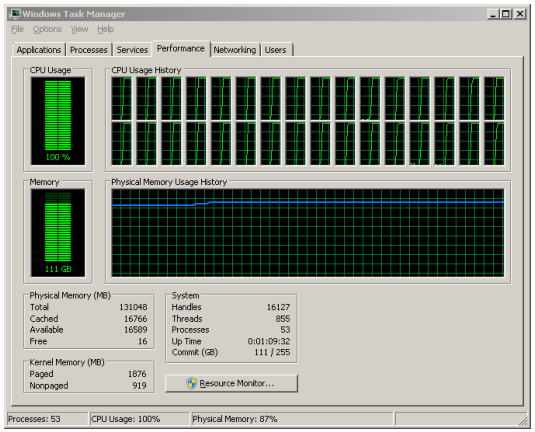


WEKA KnowledgeFlow workflow

Statistica – academic license (www.onthehub.com/statsoft/); WEKA – open source (www.cs.waikato.ac.nz/ml/**weka**/)

Massive parallel computing

Free lunch is over – concurrency in machine learning



Modern workstation
(with 4-64 CPUs)

GPU computing
(with 1000 stream processors)

Cloud computing
(with 10,000 CPUs/GPUs)

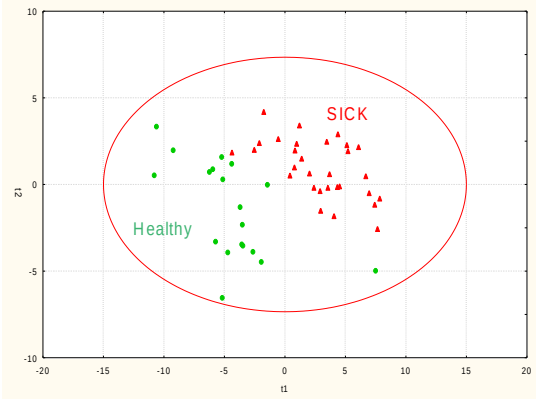
Google Prediction API

Amazon Elastic MapReduce

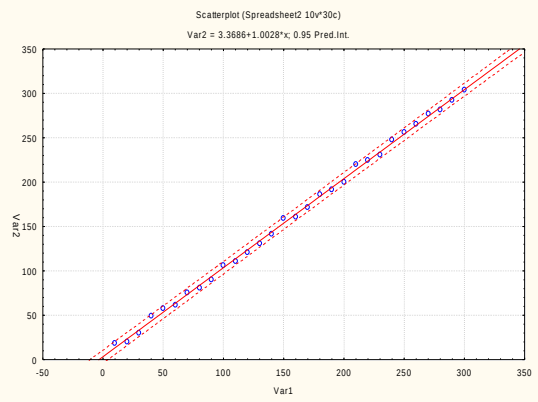
Picture Source: ThinkMate Workstation; Wikipedia; Office; Google; Amazon

Common ML applications in biotechnology

Classification - genotype/wildtype, sick/healthy, cancer grading, toxicity prediction and evaluations (FDA, EPA)

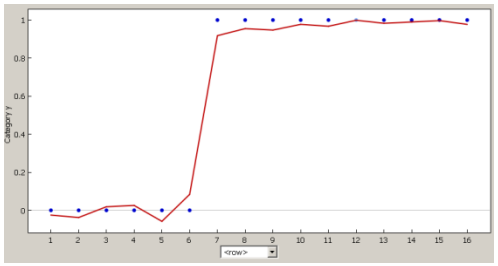


Regression - predicting biological activities, toxicity evaluations, prediction of molecular properties of unknown substances (QSAR and QSPR)



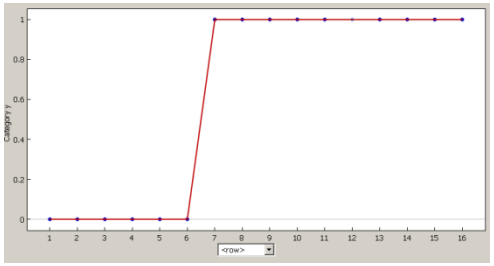
Supervised learning with categorical data
Classification

	Category y	Value x1	Value x2	Value x3
Sample1	blue	615.4603	3.363	0.0561
Sample2	blue	371.3181	3.491	0.0582
Sample3	blue	285.2924	3.636	0.0606
Sample4	blue	571.4323	3.785	0.0631
Sample5	blue	419.3184	3.933	0.0656
Sample6	blue	659.4875	4.091	0.0682
Sample7	green	832.6272	4.255	0.0709
Sample8	green	681.4981	4.418	0.0736
Sample9	green	549.4212	4.575	0.0763
Sample10	green	527.4065	4.736	0.0789
Sample11	green	458.3863	4.893	0.0816
Sample12	green	628.5179	5.501	0.0917
Sample13	green	304.3019	5.565	0.0928
Sample14	green	796.5588	5.62	0.0937
Sample15	green	774.5773	5.686	0.0948
Sample16	green	650.4938	5.76	0.0960



Good solution

Category y = sin((40.579*Value x3 + cos(sin(4.2372*Value x1)) - 3.25702)/(1.43018 + sin(Value x2)))



Perfect solution

Category y = round(0.1219*Value x2)

y = function (x values)
where y are discrete categories such as text
multiple categories (here colors) are possible

Picture Source: Nutonian Formulize

Figures of merit for classifications

A) Calculate prediction and true/false values

	Category y	Value x1	Value x2	Value x3		predicted	true/false
Sample1	blue	615.4603	3.363	0.0561		blue	TRUE
Sample2	blue	371.3181	3.491	0.0582		blue	TRUE
Sample3	blue	285.2924	3.636	0.0606		blue	TRUE
Sample4	blue	571.4323	3.785	0.0631		blue	TRUE
Sample5	blue	419.3184	3.933	0.0656		blue	TRUE
Sample6	blue	659.4875	4.091	0.0682		blue	TRUE
Sample7	green	832.6272	4.255	0.0709		blue	FALSE
Sample8	green	681.4981	4.418	0.0736		green	TRUE
Sample9	green	549.4212	4.575	0.0763		green	TRUE
Sample10	green	527.4065	4.736	0.0789		green	TRUE
Sample11	green	458.3863	4.893	0.0816		green	TRUE
Sample12	green	628.5179	5.501	0.0917		green	TRUE
Sample13	green	304.3019	5.565	0.0928		green	TRUE
Sample14	green	796.5588	5.62	0.0937		green	TRUE
Sample15	green	774.5773	5.686	0.0948		green	TRUE
Sample16	green	650.4938	5.76	0.0960		green	TRUE

Example is special case of binary classification
multiple categories are possible

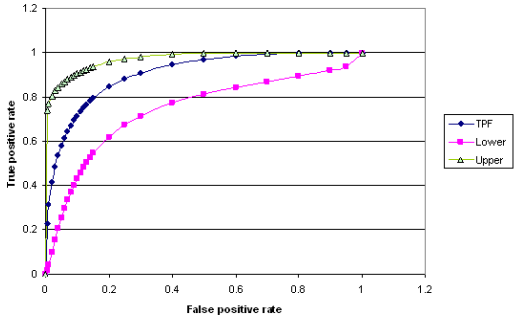
B) Confusion matrix

true positives	true negative
false positives	false negatives

C) Figures of merit

True positive rate or
sensitivity or recall = TP/(TP+FN)
False positive rate = FP/FP+TN)
Accuracy = (TP+TN)/(TP+TN+FP+FN)
Specificity = (TN/(FP+TN)
Precision = TP/(TP+FN)
Negative predictive value = TN/(TN+FN)
False discovery rate = FP/(FP+TP)

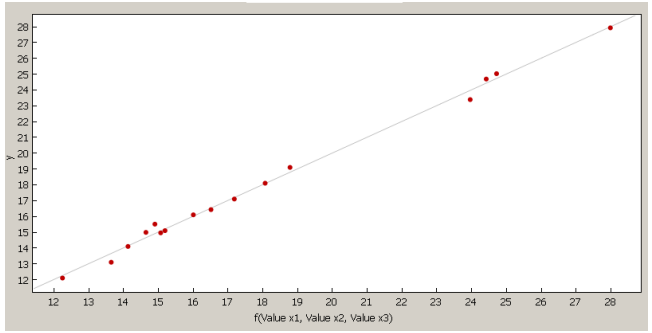
D) ROC curves



Source: http://en.wikipedia.org/wiki/Sensitivity_and_specificity

Supervised learning with continuous data
Regression

	y	Value x1	Value x2	Value x3
Sample1	12.10	615.4603	3.363	0.05605
Sample2	14.96	371.3181	3.491	0.058183
Sample3	13.10	285.2924	3.636	0.0606
Sample4	15.51	571.4323	3.785	0.063083
Sample5	14.10	419.3184	3.933	0.06555
Sample6	14.99	659.4875	4.091	0.068183
Sample7	15.10	832.6272	4.255	0.070917
Sample8	25.03	681.4981	4.418	0.073633
Sample9	16.10	549.4212	4.575	0.07625
Sample10	16.43	527.4065	4.736	0.078933
Sample11	17.10	458.3863	4.893	0.08155
Sample12	24.69	628.5179	5.501	0.091683
Sample13	18.10	304.3019	5.565	0.09275
Sample14	27.93	796.5588	5.62	0.093667
Sample15	19.10	774.5773	5.686	0.094767
Sample16	23.39	650.4938	5.76	0.096



Good solution

y = mod(Value x2*Value x2, 6.61)*tan(0.2688*sin(5.225*Value x2*Value x2))*max(Value x2 + -25.59/(Value x2*Value x2) + cos(42.32*Value x2), floor(mod(Value x2*Value x2, 6.61))) + round(4.918*Value x2) - 3.945 - 3.672*ceil(sin(5.225*Value x2*Value x2))

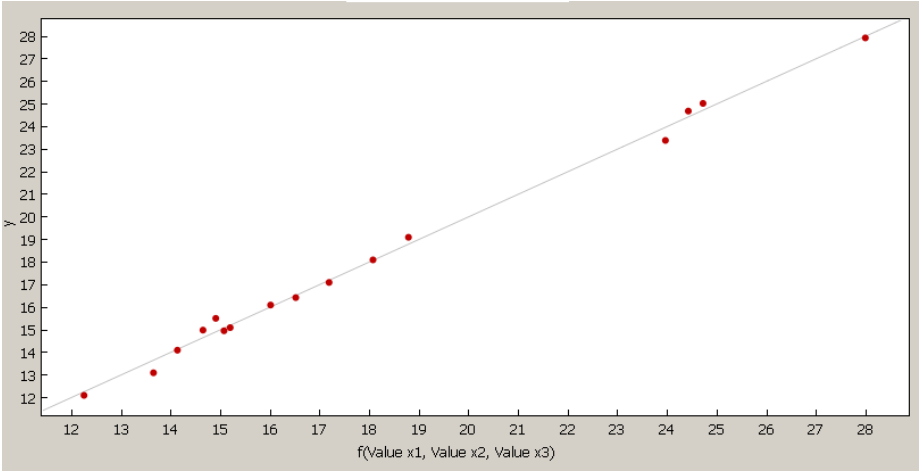
R^2 Goodness of Fit	**0.99522414**
Correlation Coefficient	**0.99760921**
Maximum Error	**0.60953998**
Mean Squared Error	**0.092117594**
Mean Absolute Error	**0.22821247**

y = function (x values)

where y are continuous values such as numbers

Picture Source: Nutonian Formulize

Figures of merit for regressions

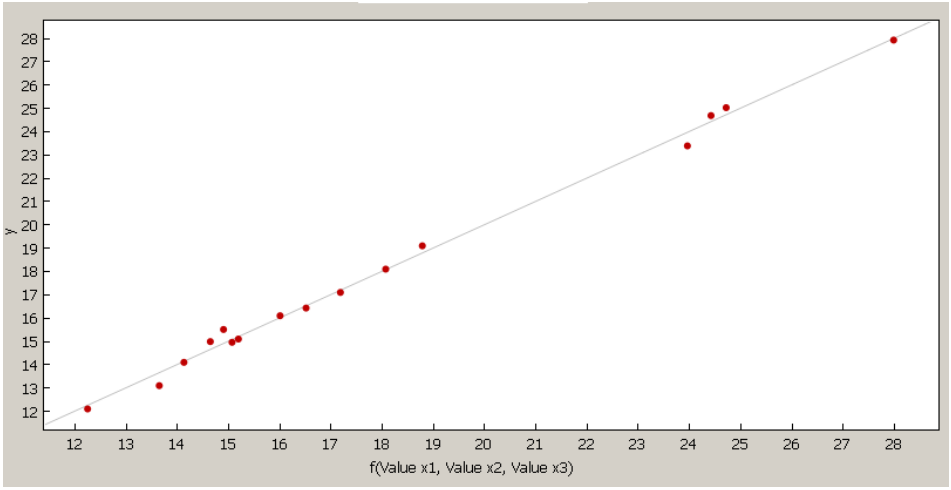


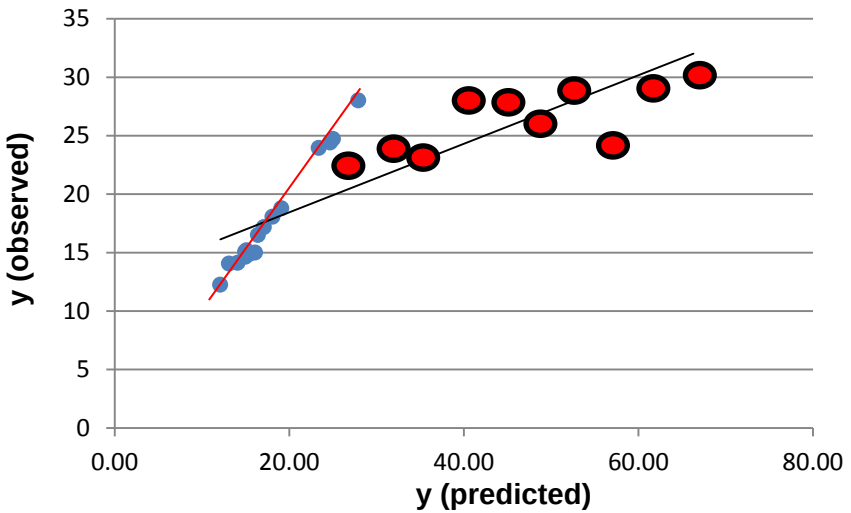
R^2 Goodness of Fit	0.99522414
Correlation Coefficient	0.99760921
Maximum Error	0.60953998
Mean Squared Error	0.092117594
Mean Absolute Error	0.22821247

Figures of merit are also calculated for external test and validation sets such as the **predictive squared correlation coefficient Q^2**

Picture Source: Nutonian Formulize

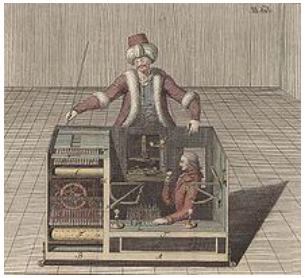
Overfitting – trust but verify





Old model applied to new data
$R^2 = 0.995 \rightarrow Q^2 = 0.7227$

External validation failed
Prediction power is most important



Mechanical turk

The problem of overfitting.; DM Hawkins
J. Chem. Inf. Comput. Sci.; 2004; 44(1) pp 1 - 12;

Picture Source: Wikipedia;

Overfitting – avoid the unexpected

Dogs

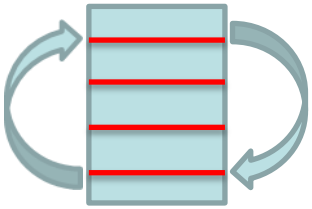
Cats

New Kid on the block

Picture Source: Wikipedia; Icanhzcheesburger.com

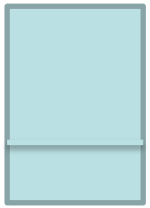
Avoid overfitting

Internal cross-validation (weak)



n-fold CV

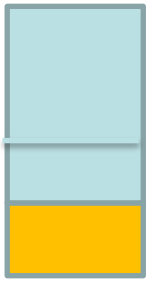
70/30 split development/test set (good)



Training (70%)

Test (30%)

External validation set or blind hold-out (best)

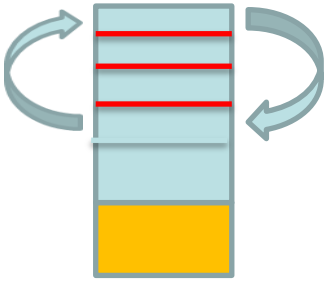


Training (70%)

Test (30%)

External validation set (+30%)

TOP:
Combine all three methods

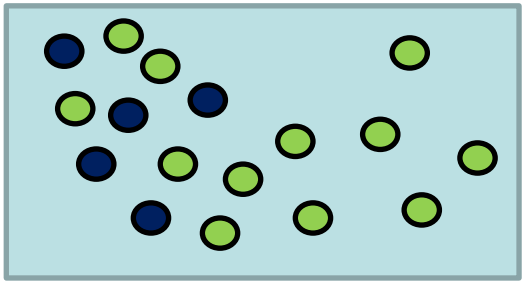


Training (70%) with CV

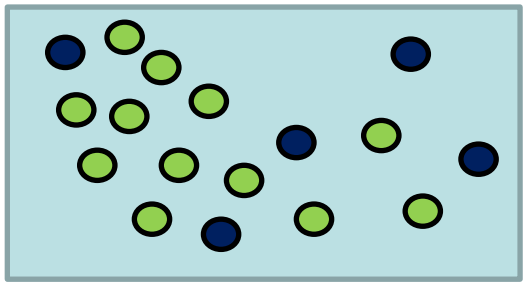
Test (30%)

External validation set (+30%)

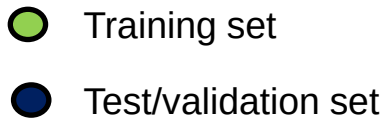
Sample selection for testing and validation



Bad selection



Good selection



Sample selection for test and validation set split should be truly randomized

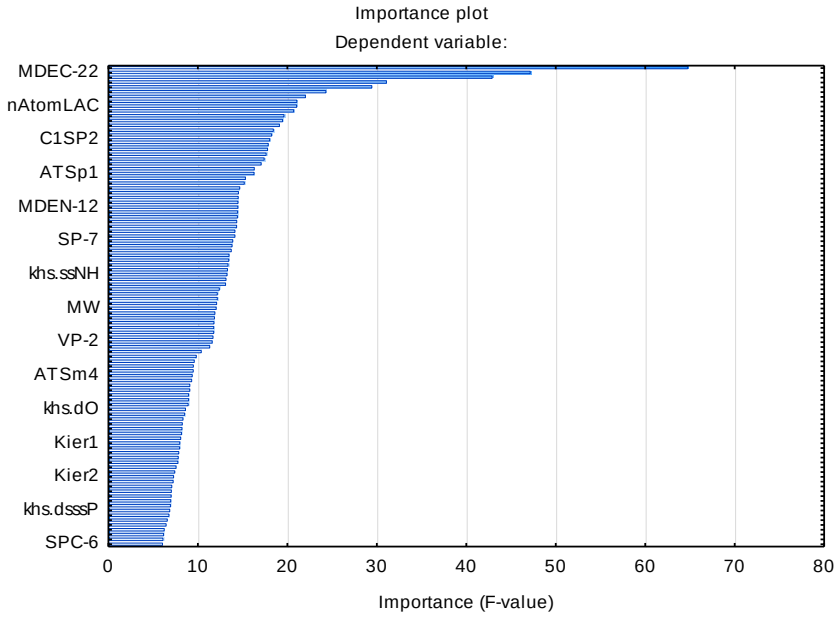
Range of the y-coordinate (activity or response) should be completely covered

Training and test set variables should not overlap

Picture Source: Nutonian Formulize

Why do we need feature selection?

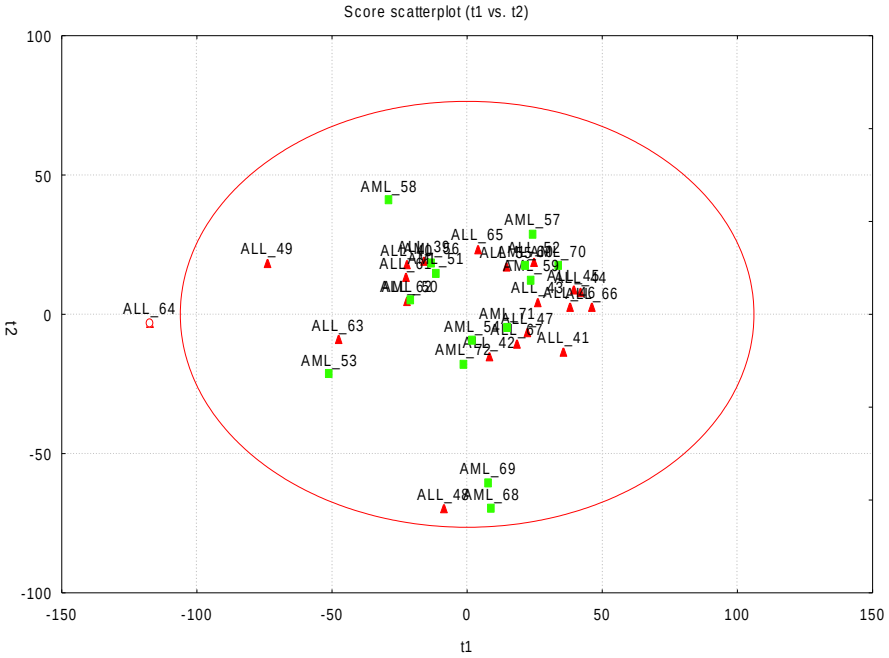
- Reduces computational complexity
- Curse of dimensionality is avoided
- Improves accuracy
- The selected features can provide insights about the nature of the problem*

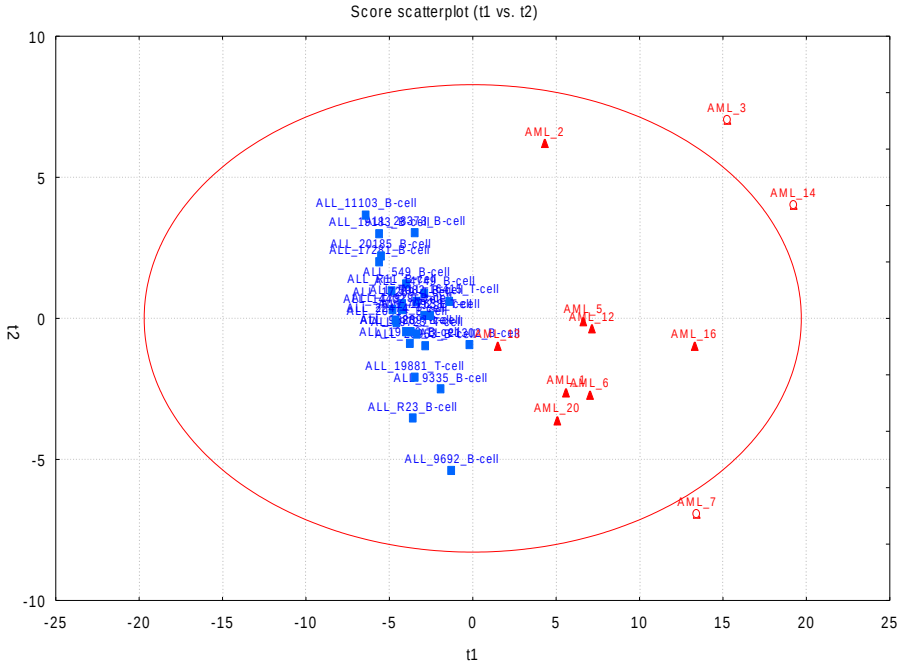


* Margin Based Feature Selection Theory and Algorithms; Amir Navot

Feature selection example

Principal component analysis (PCA) microarray data





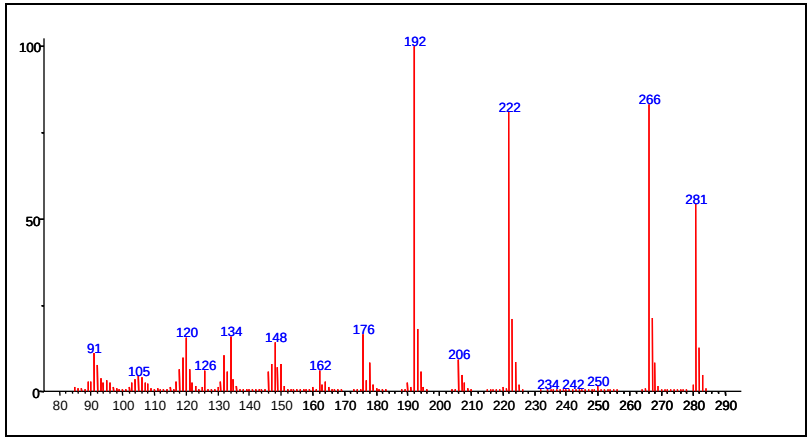
NO feature selection → no separation

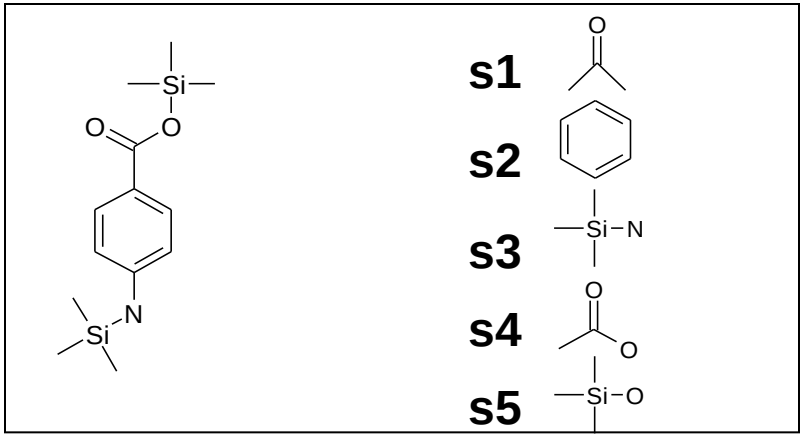
With feature selection → separation

Data: Golub, 1999 Science Mag

Approach: Automated substructure detection

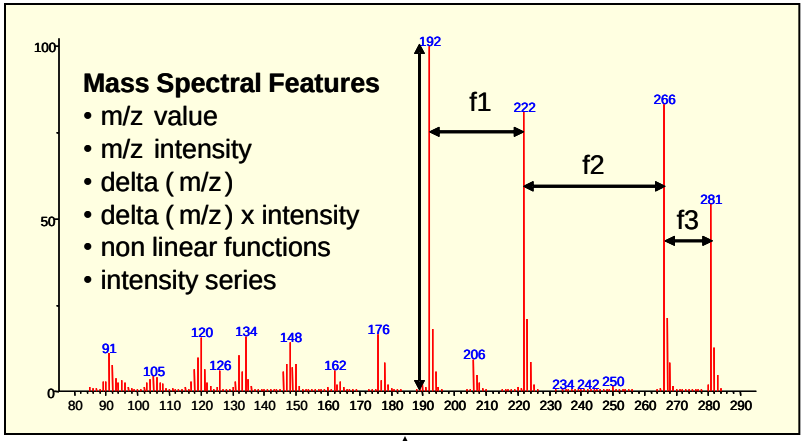
Aim1: take unknown mass spectrum – predict all substructures
Aim2: classification into common compound classes (sugar, amino acid, sterol)





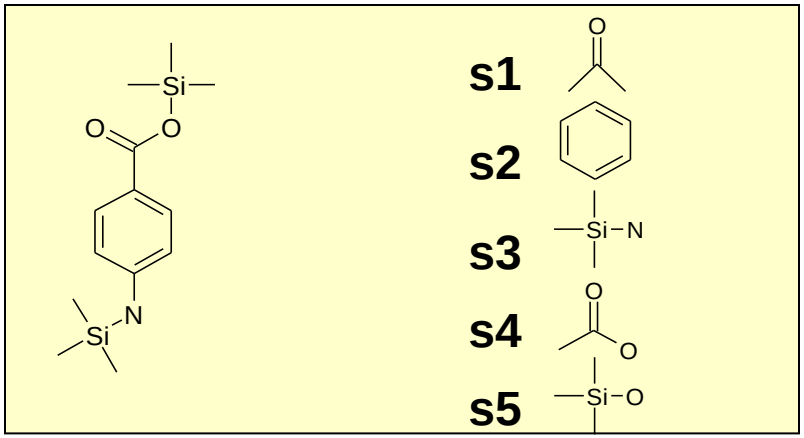
Pioneers: Dendral project at Stanford University in the 1970s
Varmuza at University of Vienna
Steve Stein at NIST
MOLGEN-MS team at University Bayreuth

Principle of mass spectral features



MS Feature matrix

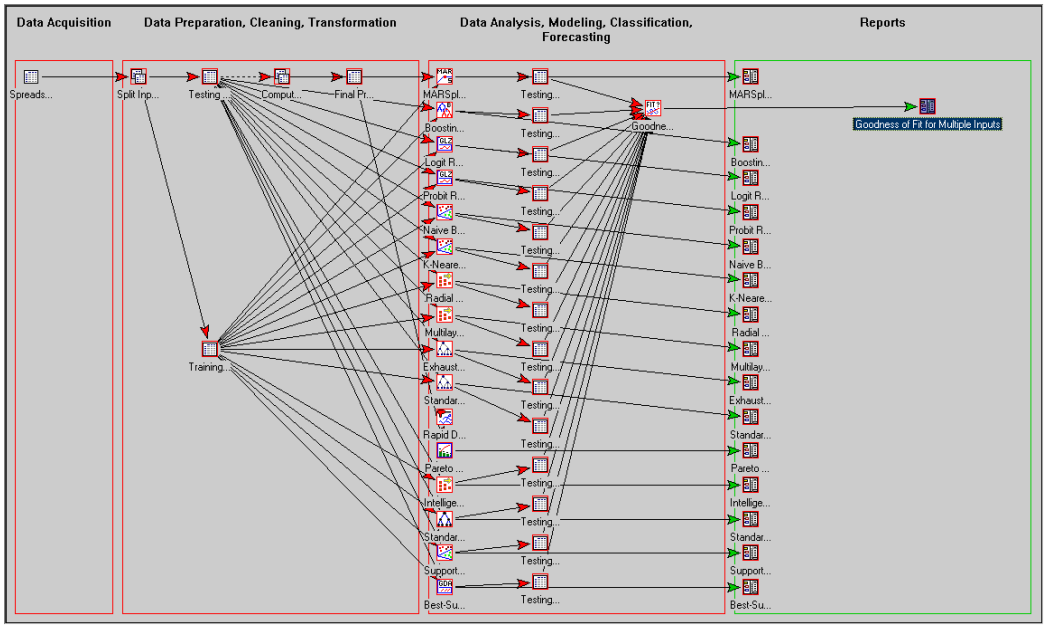
MS Spectrum	f1	f2	f3	f4	f5	fn
MS1	100	20	50	60	0	0
MS2	100	20	50	60	0	20
MS3	100	20	60	50	0	0
MS4	0	40	20	50	0	40
MS5	0	40	20	50	0	40



Substructure matrix

Substructure	s1	s2	s3	s4	s5	sn
Molecule1	Y	Y	N	Y	Y	N
Molecule2	Y	Y	N	Y	Y	N
Molecule3	Y	Y	N	Y	Y	N
Molecule4	N	N	N	Y	Y	Y
Molecule5	N	N	N	Y	Y	Y

Application - Substructure detection and prediction



Generalized Linear Models (GLM)
- General Discriminant Analysis
- Binary logit (logistic) regression
- Binary probit regression

Nonlinear models
- Multivariate adaptive regression splines (MARS)

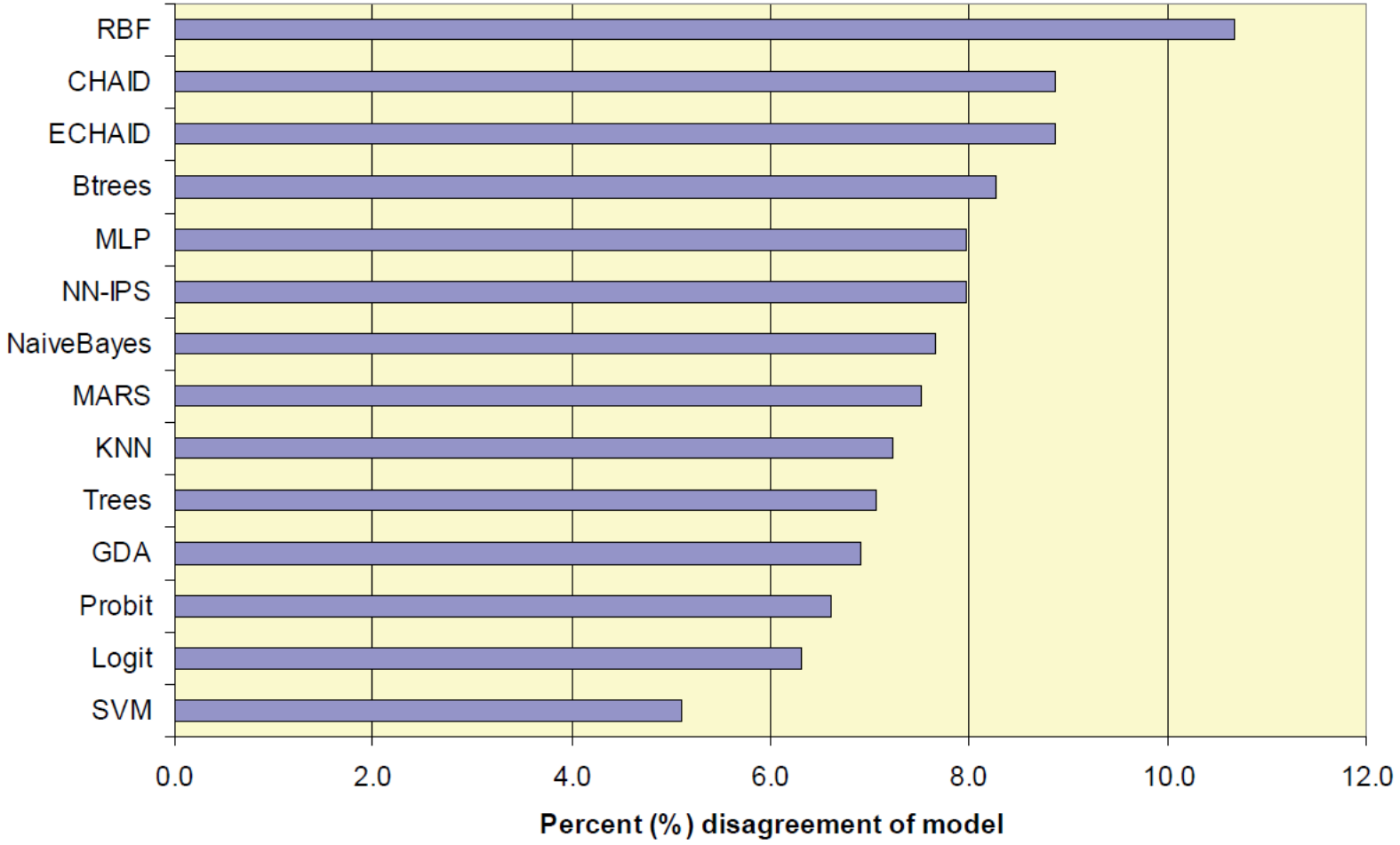
Tree models
- Standard Classification Trees (CART)
- Standard General Chi-square Automatic Interaction Detector (CHAID)
- Exhaustive CHAID
- Boosting classification trees
- M5 regression trees

Meta Learning

Neural Networks
- Multilayer Perceptron
- Neural network (MLP)
- Radial Basis Function neural network (RBF)

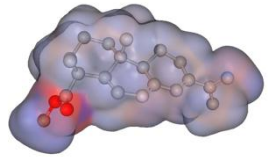
Machine Learning
- Support Vector Machines (SVM)
- Naive Bayes classifier
- k-Nearest Neighbors (KNN)

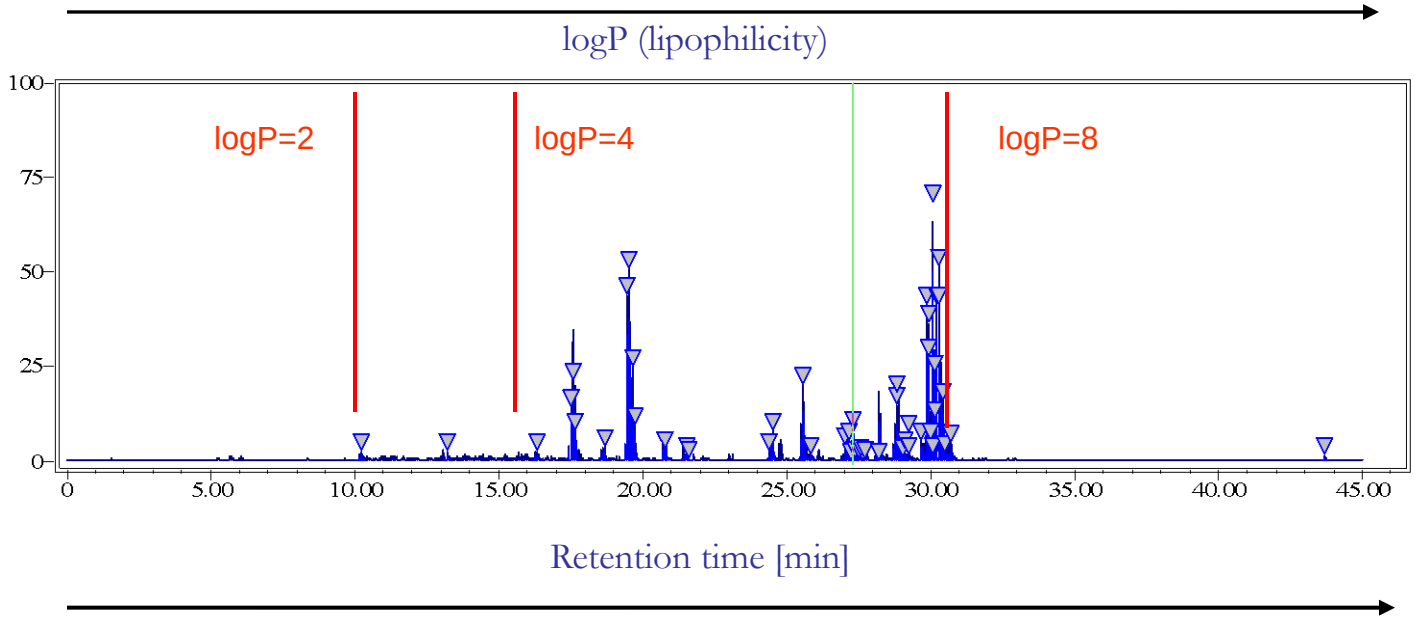
Strategy - let all machine learning algorithms compete



Lower is better

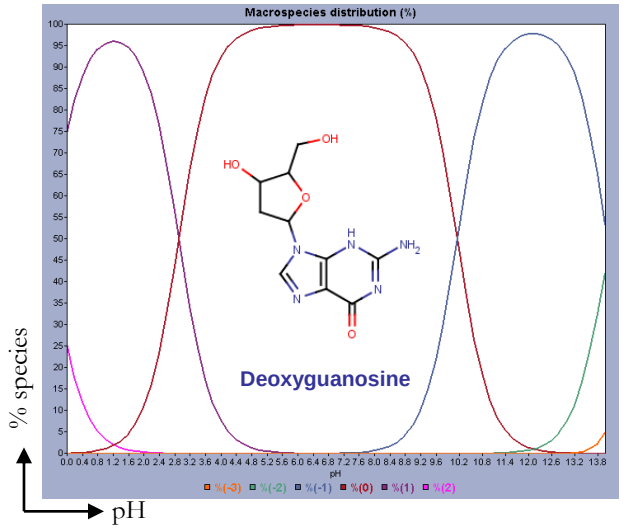
Application: Retention time prediction for liquid chromatography



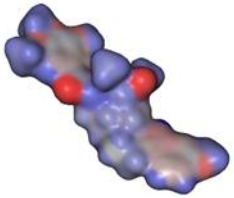


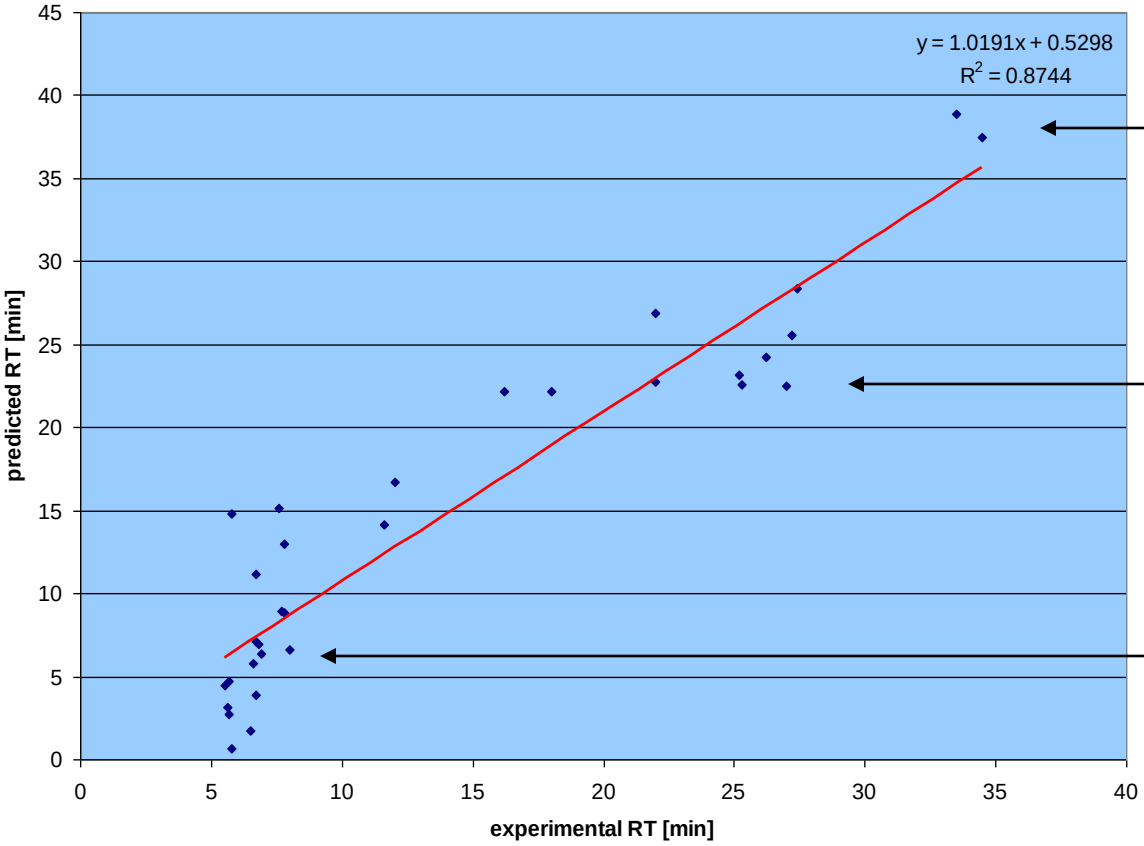
Calibration using logP concept for reversed phase liquid chromatography data

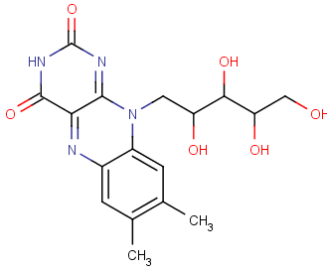
- very simplistic and coarse filter for RP only
- problematic with multi ionizable compounds
- logD (includes pKa) better than logP
- possible use as time segment filter



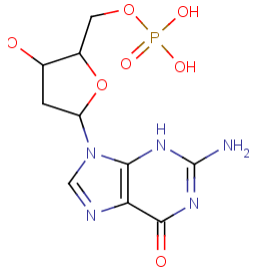
Application: Retention time prediction for liquid chromatography



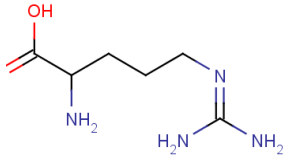




Riboflavin



Deoxyguanosine monophosphate (dGMP)

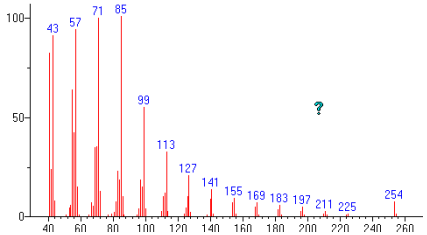


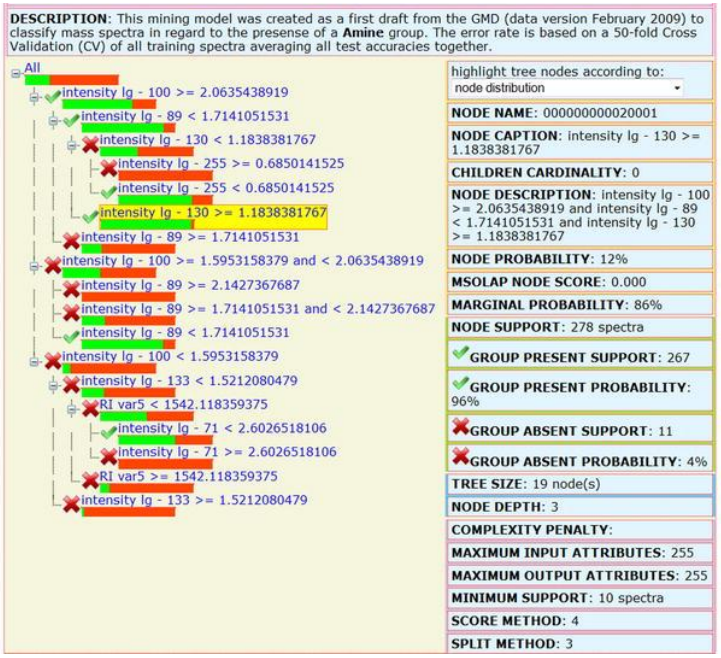
Arginine

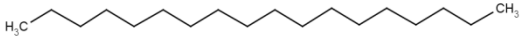
- Based on logD, pKa, logP and Kier & Hall atomic descriptors;
- 90 compounds; (n_{dev}= 48, n_{test} = 32); Std error 3.7 min
- Good models need development set n>500, needs to be highly diverse
- **Prediction power is most important**

QSRR Model: Tobias Kind (FiehnLab) using ChamAxon Marvin and WEKA
Data Source: Lu W, Kimball E, Rabinowitz JD. J Am Soc Mass Spectrom. 2006 Jan;17(1):37-50; LC method using 90 nitrogen metabolites on RP-18

Application: Decision tree supported substructure prediction of metabolites from GC-MS profiles



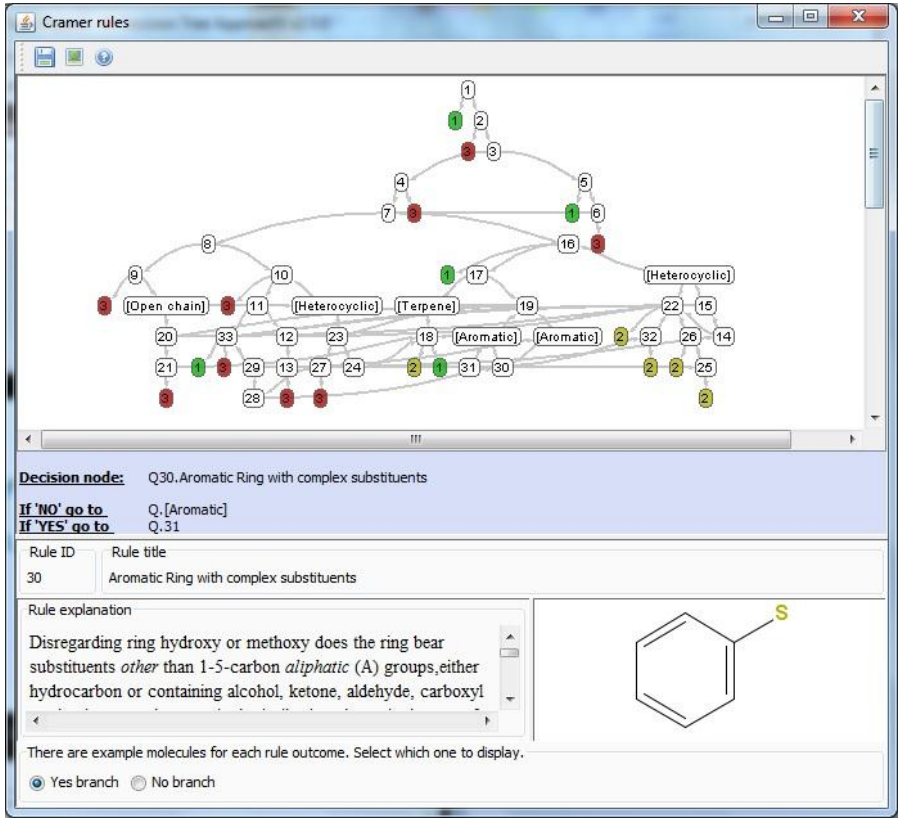


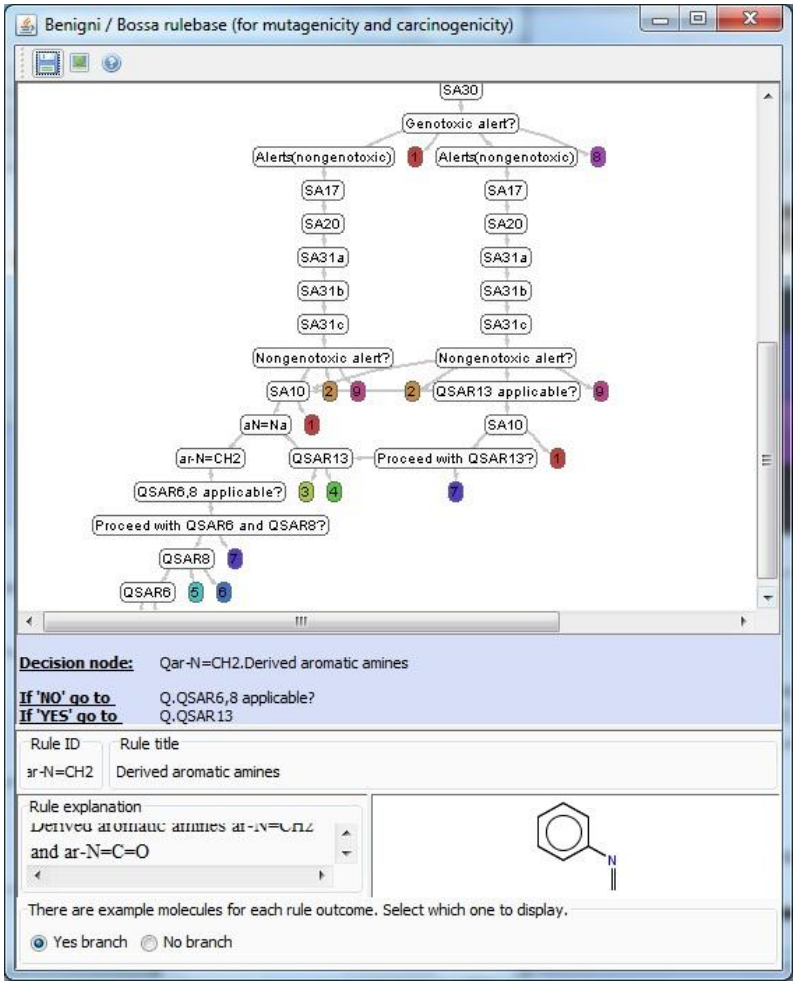


Spectrum → Decision tree → Compound structure

Source: Metabolomics. 2010 Jun;6(2):322-333. Epub 2010 Feb 16.
Decision tree supported substructure prediction of metabolites from GC-MS profiles.
Hummel J, Strehmel N, Selbig J, Walther D, Kopka J.

Toxicity and carcinogenicity predictions with ToxTree





Source: Patlewicz G, Jeliazkova N, Safford RJ, Worth AP, Aleksiev B. (2008) An evaluation of the implementation of the Cramer classification scheme in the Toxtree software. SAR QSAR Environ Res. ; http://toxtree.sourceforge.net

Conclusions – Machine Learning

Classification (categorical data) and **regression** (continuous data) for prediction of future values

Let algorithms compete for best solution (voting, boosting, bagging)

Validation (trust but verify) **is the cornerstone of machine learning** to avoid false results and wishful thinking

Modern algorithms do not necessarily provide direct causal insight they rather provide the best statistical solution or equation

Domain knowledge of the learning problem is important and **helpful for artifact removal** and final interpretation

Prediction power is most important

Thank you!