Lecture 2: Introduction to Feature Selection

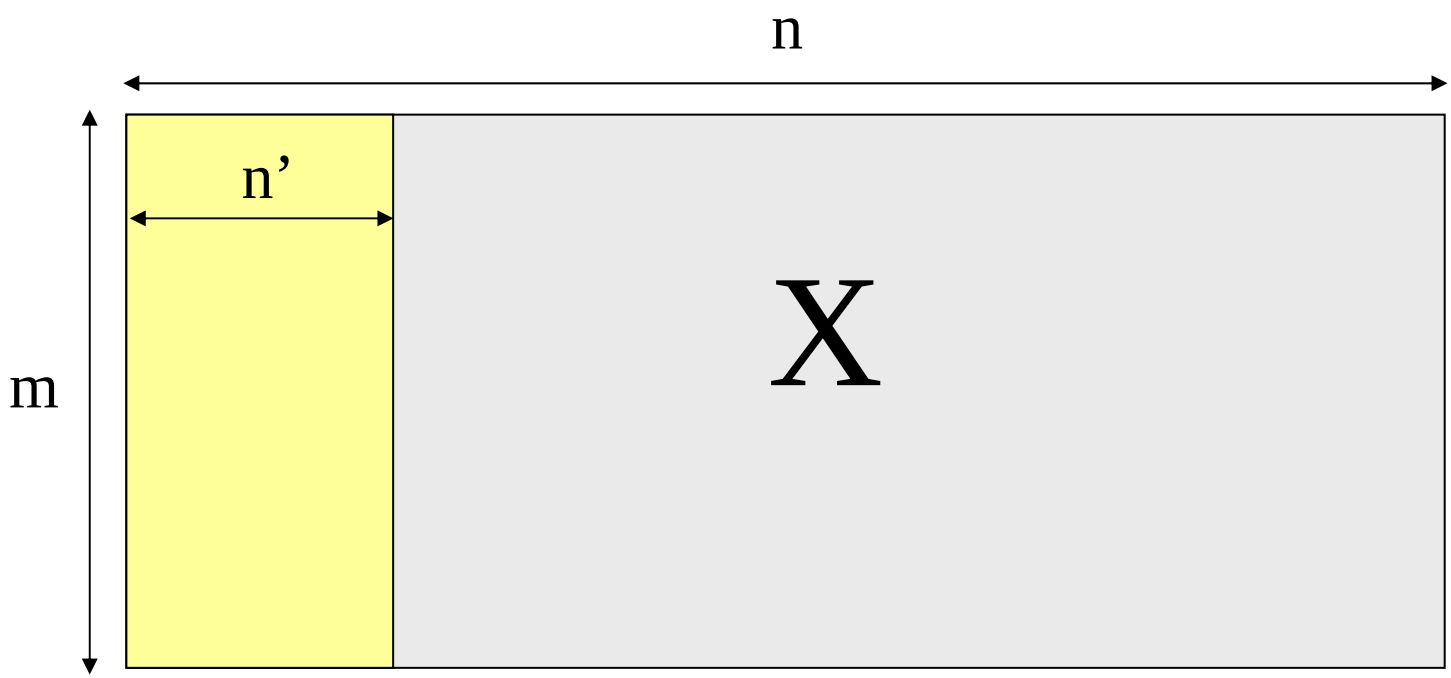
Isabelle Guyon

isabelle@clopinet.com

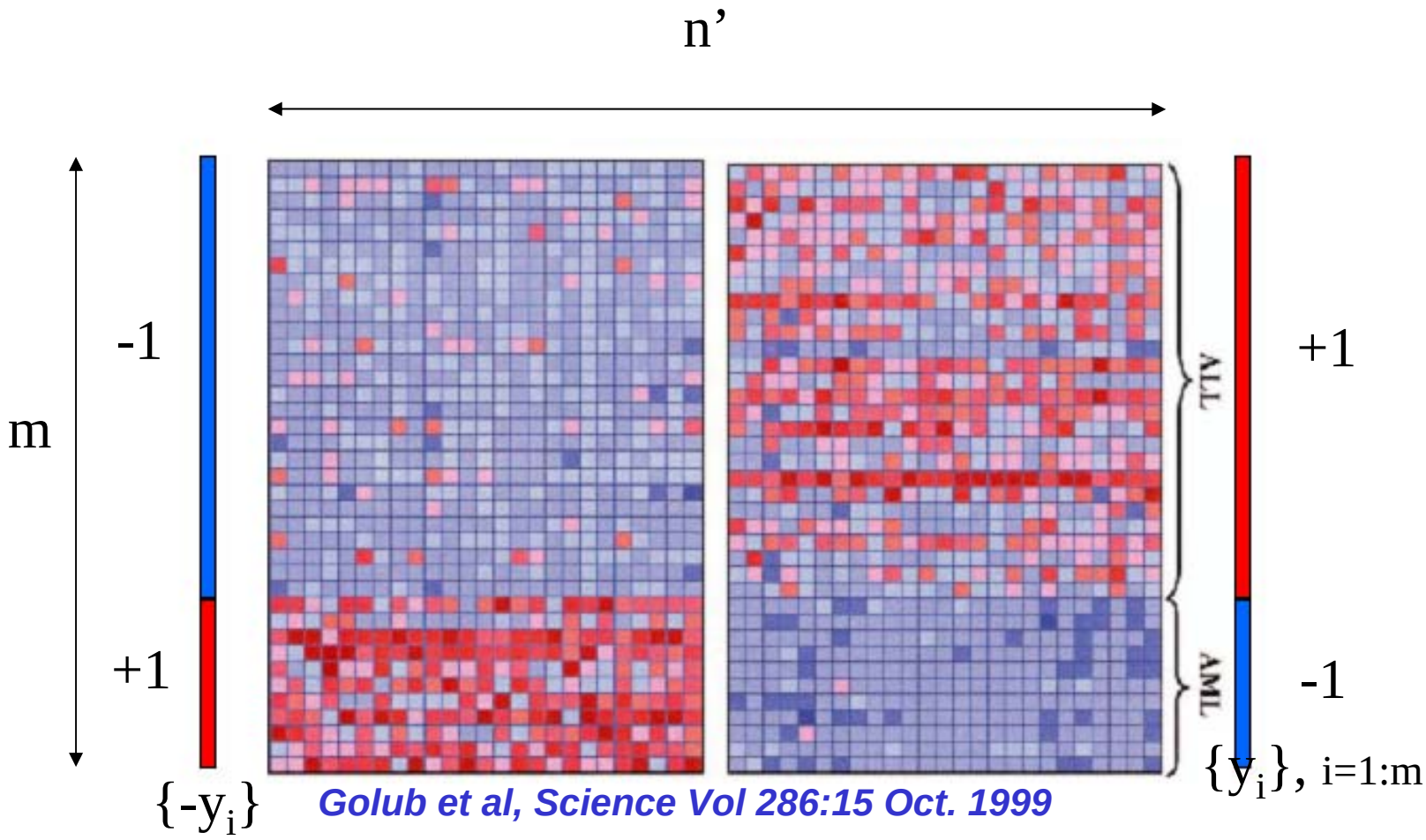
Notations and Examples

Feature Selection

- **Thousands to millions of low level features**: select the most relevant one to build **better, faster, and easier to understand** learning machines.

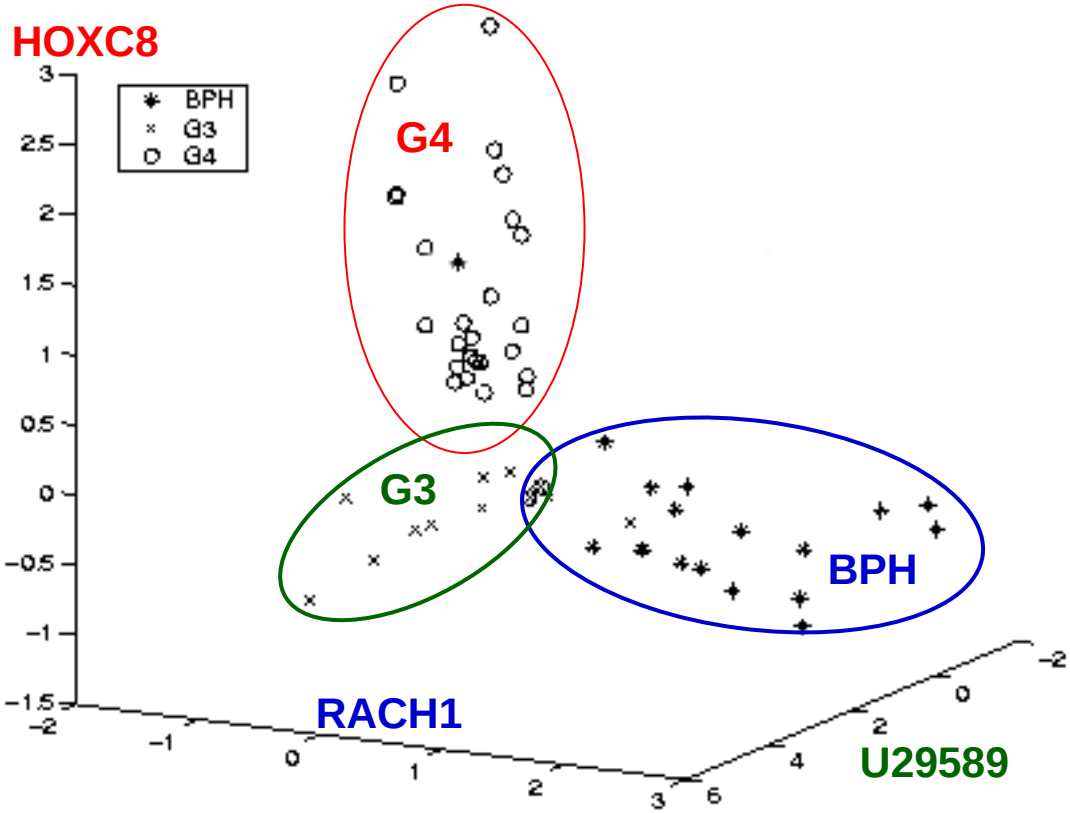


Leukemia Diagnosis



Golub et al, Science Vol 286:15 Oct. 1999

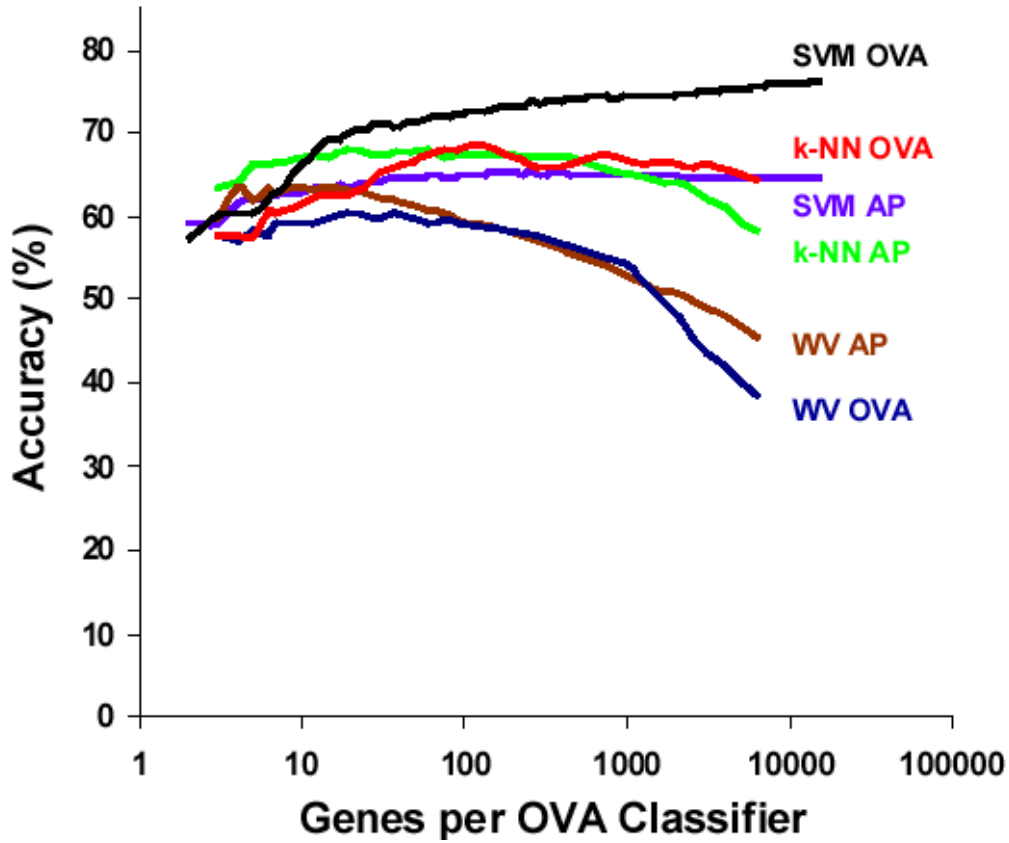
Prostate Cancer Genes



RFE SVM, ***Guyon-Weston, 2000. US patent 7,117,188***

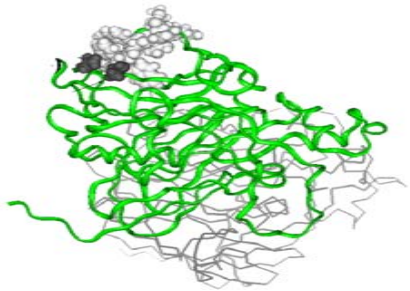
Application to prostate cancer. ***Elisseeff-Weston, 2001***

RFE SVM for cancer diagnosis



Differenciation of 14 tumors. *Ramaswamy et al, PNAS, 2001*

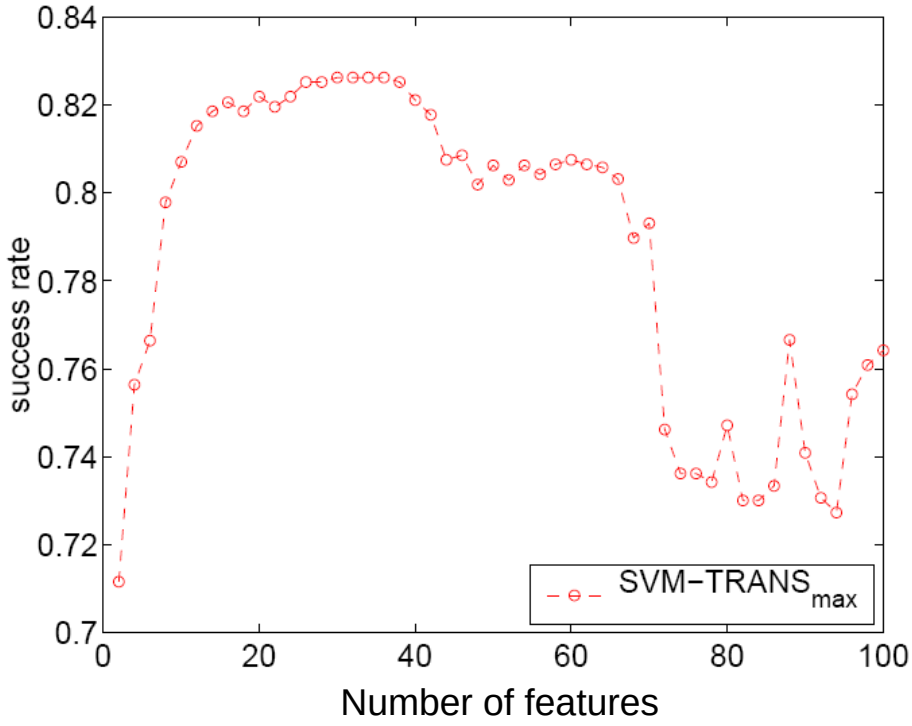
QSAR: Drug Screening



Binding to Thrombin (DuPont Pharmaceuticals)

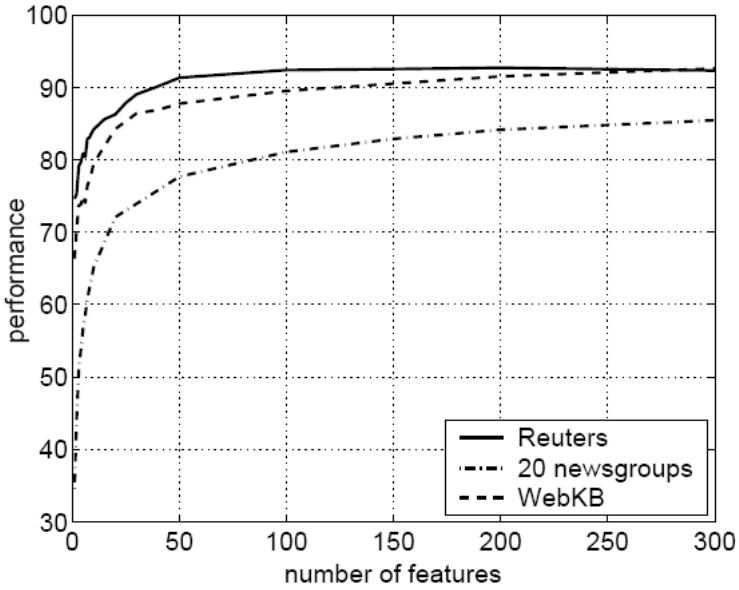
- 2543 compounds tested for their ability to bind to a target site on thrombin, a key receptor in blood clotting; 192 “active” (bind well); the rest “inactive”. Training set (1909 compounds) more depleted in active compounds.

- 139,351 binary features, which describe three-dimensional properties of the molecule.



Weston et al, Bioinformatics, 2002

Text Filtering



Reuters: 21578 news wire, 114 semantic categories.

20 newsgroups: 19997 articles, 20 categories.

WebKB: 8282 web pages, 7 categories.

Bag-of-words: >100000 features.

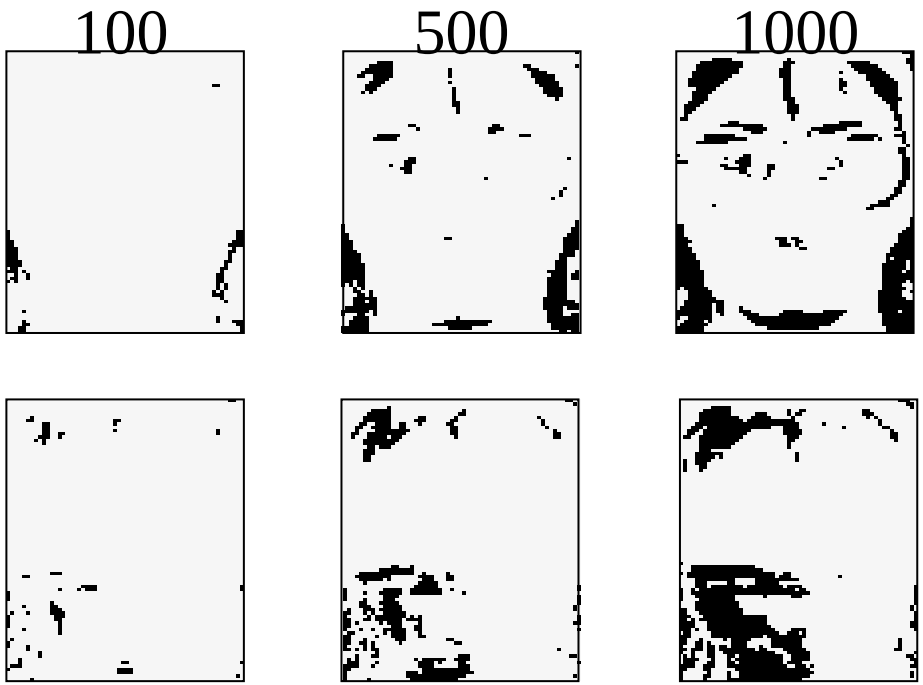
Top 3 words of some categories:

- **Alt.atheism**: atheism, atheists, morality
- **Comp.graphics**: image, jpeg, graphics
- **Sci.space**: space, nasa, orbit
- **Soc.religion.christian**: god, church, sin
- **Talk.politics.mideast**: israel, armenian, turkish
- **Talk.religion.misc**: jesus, god, jehovah

Bekkerman et al, JMLR, 2003

Face Recognition

- Male/female classification
- 1450 images (1000 train, 450 test), 5100 features (images 60x85 pixels)



Navot-Bachrach-Tishby, ICML 2004

Nomenclature

- **Univariate method**: considers one variable (feature) at a time.
- **Multivariate method:** considers subsets of variables (features) together.
- **Filter method:** ranks features or feature subsets independently of the predictor (classifier).
- **Wrapper method:** uses a classifier to assess features or feature subsets.

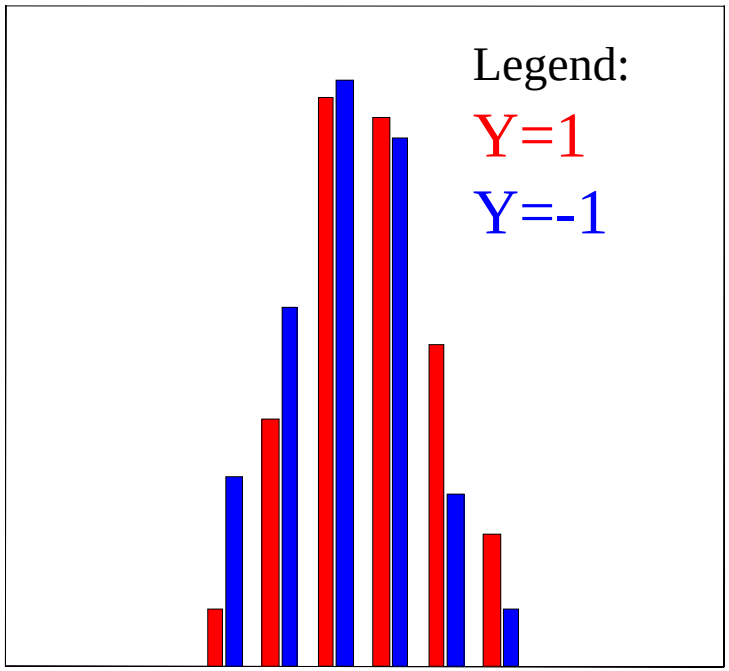
Univariate Filter Methods

Individual Feature Irrelevance

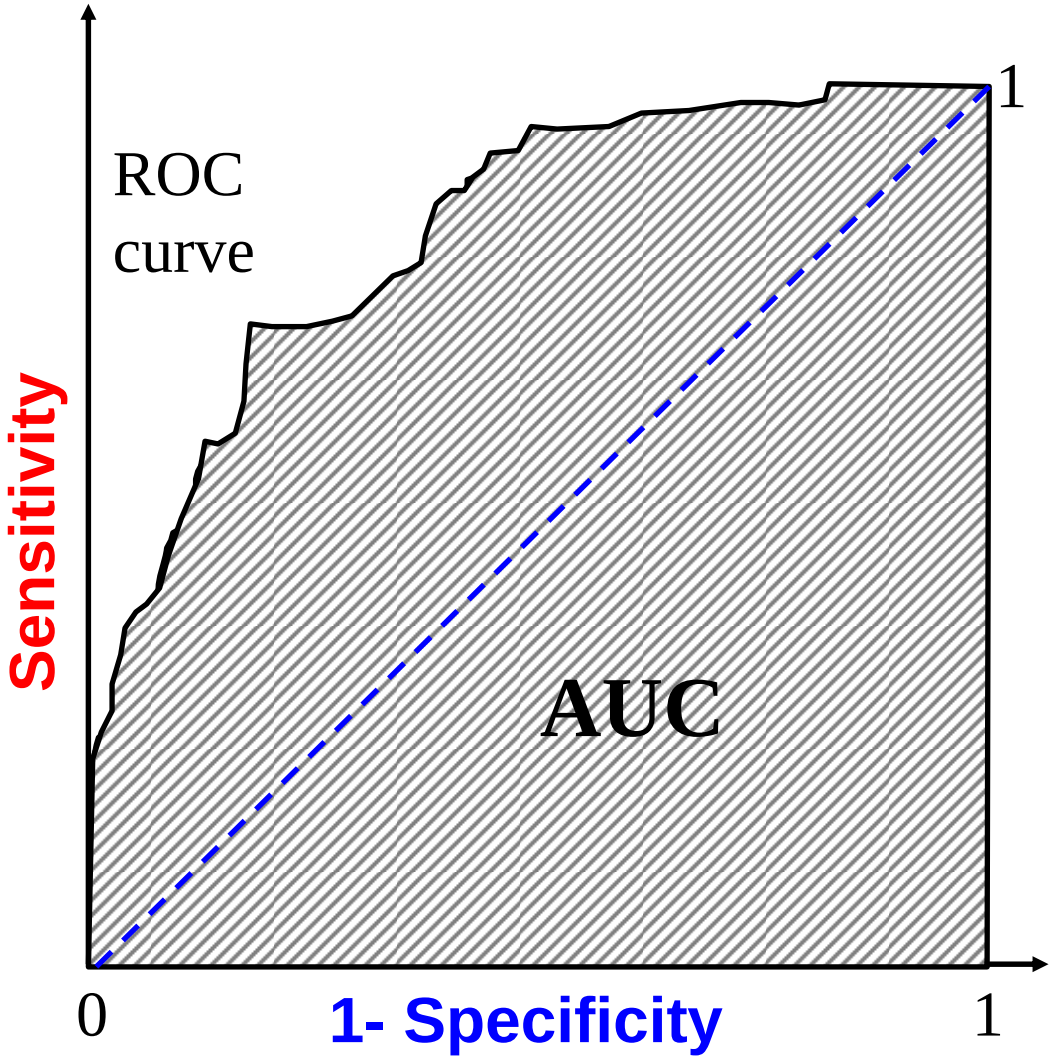
$P(X_i, Y) = P(X_i)\, P(Y)$

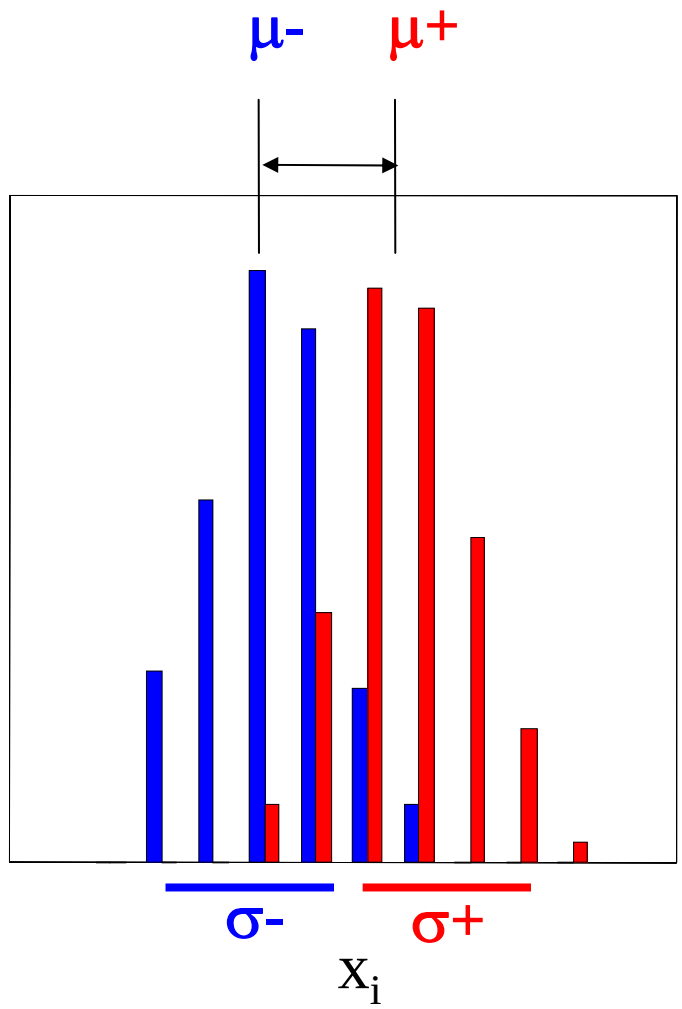
$P(X_i | Y) = P(X_i)$

$P(X_i | Y=1) = P(X_i | Y=-1)$

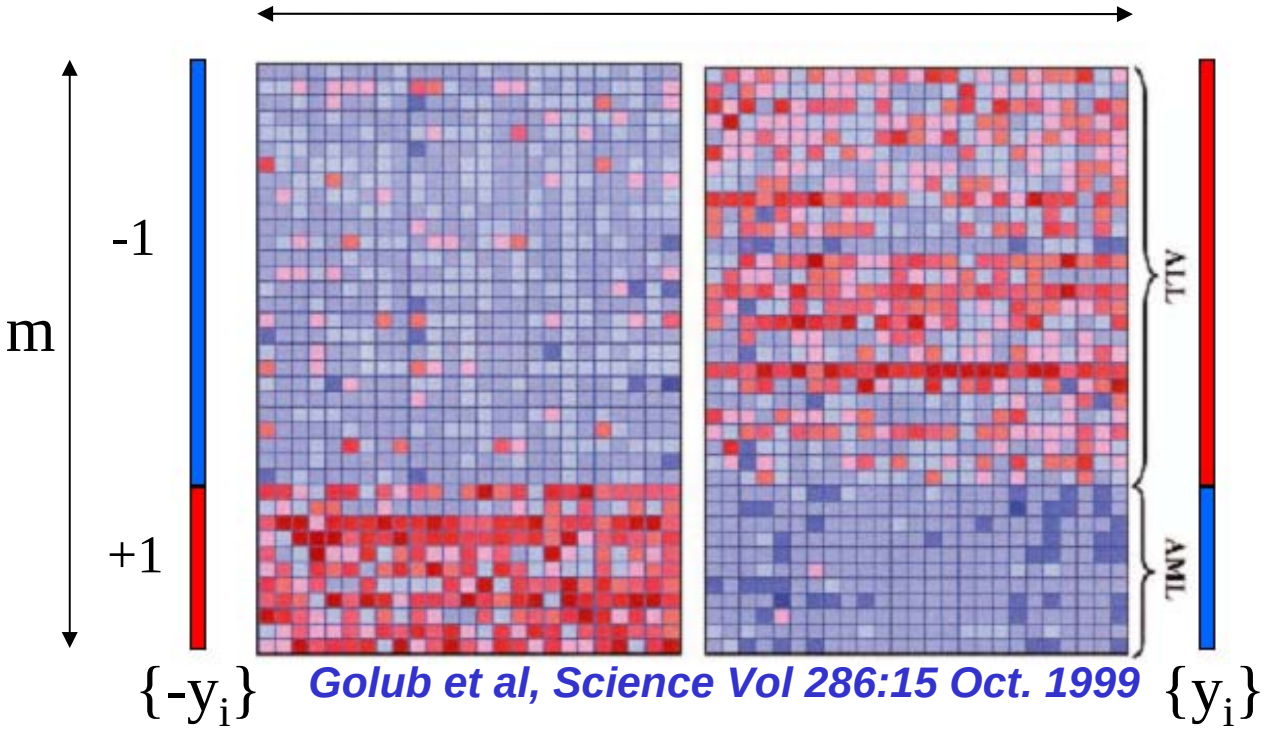


Individual Feature Relevance





S2N

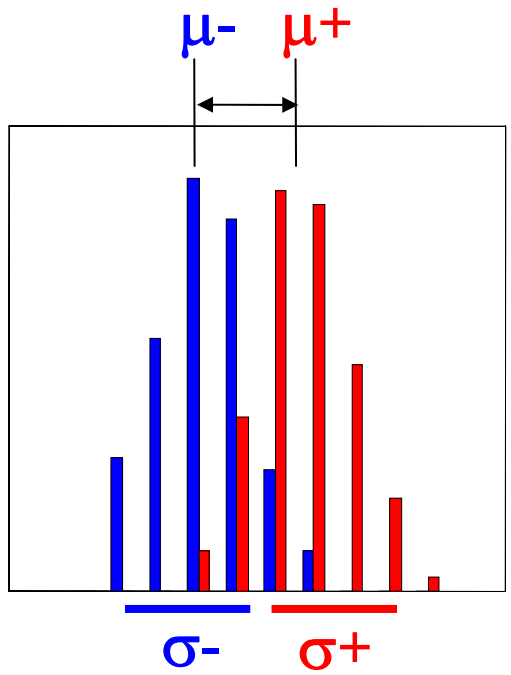


Golub et al, Science Vol 286:15 Oct. 1999

$$S2N = \frac{|\mu_+ - \mu_-|}{\sigma_+ + \sigma_-}$$

$S2N \cong R \sim \mathbf{x} \bullet \mathbf{y}$

after "standardization" $\mathbf{x} \leftarrow (\mathbf{x}-\mu_x)/\sigma_x$



Univariate Dependence

- Independence:

$$P(X, Y) = P(X)\,P(Y)$$

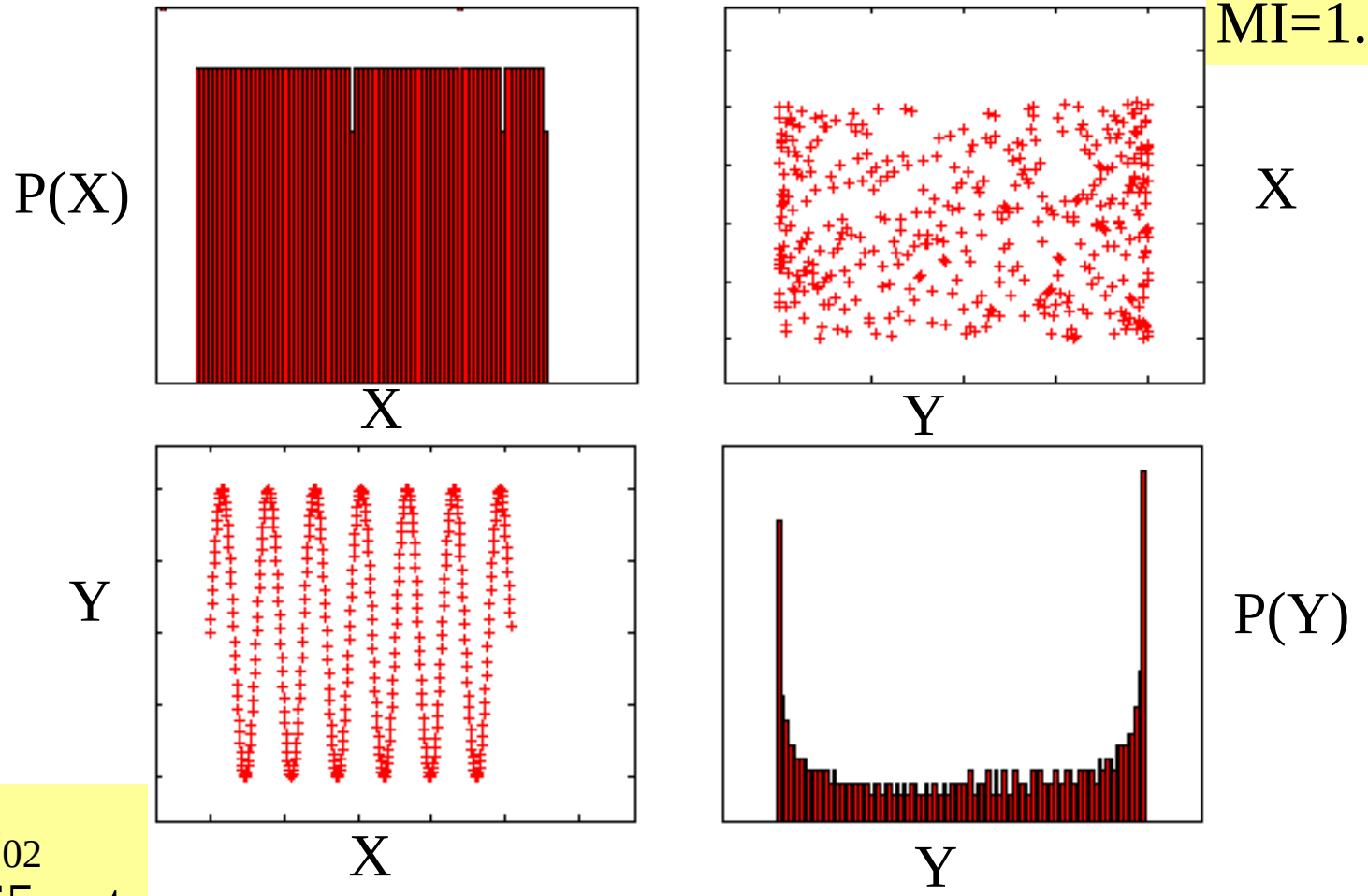
- Measure of dependence:

$$MI(X, Y) = \int P(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)}\, dX\, dY$$

$$= KL(P(X,Y) \,||\, P(X)P(Y))$$

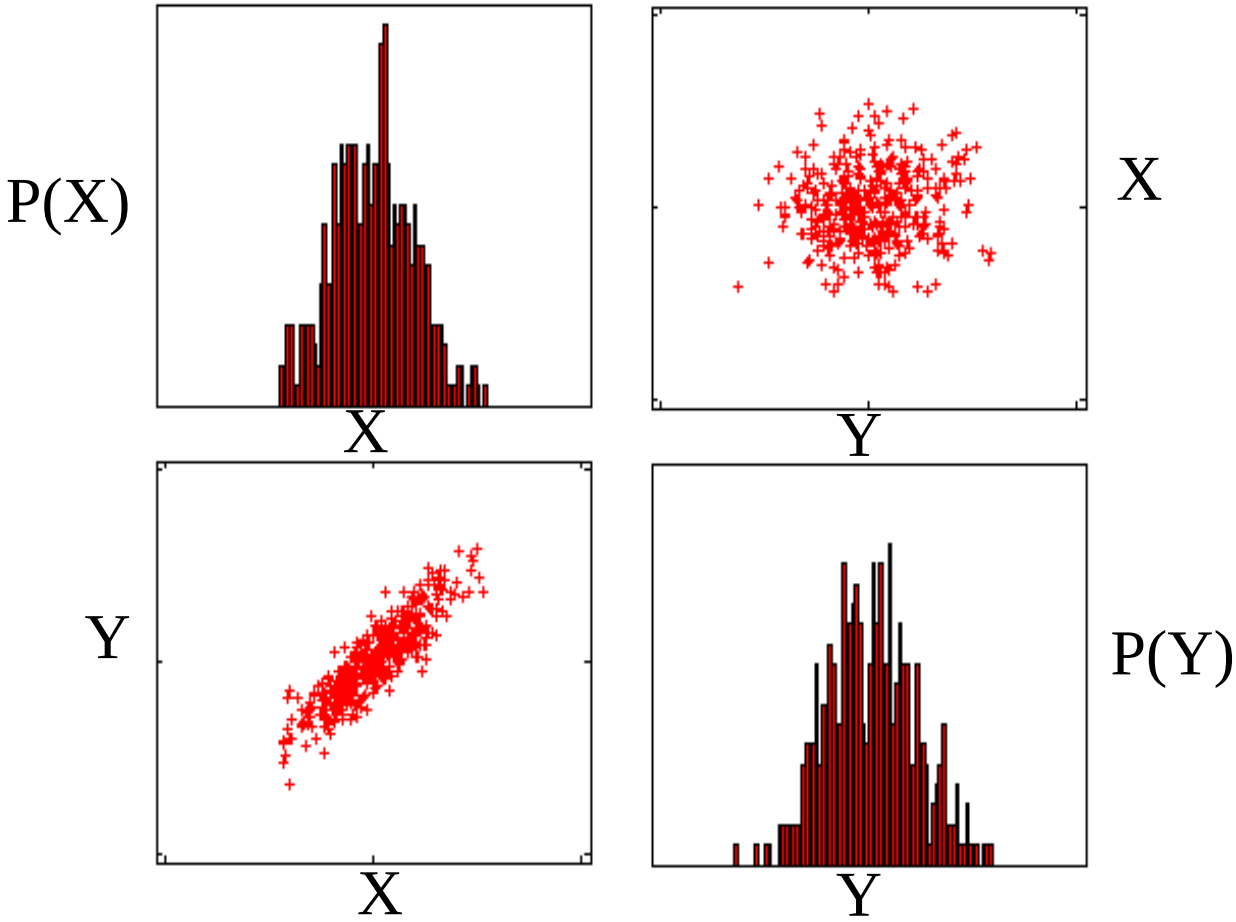
Correlation and MI

R=0.02
MI=1.03 nat



R=0.00$_{02}$
MI=1.65 nat

Gaussian Distribution

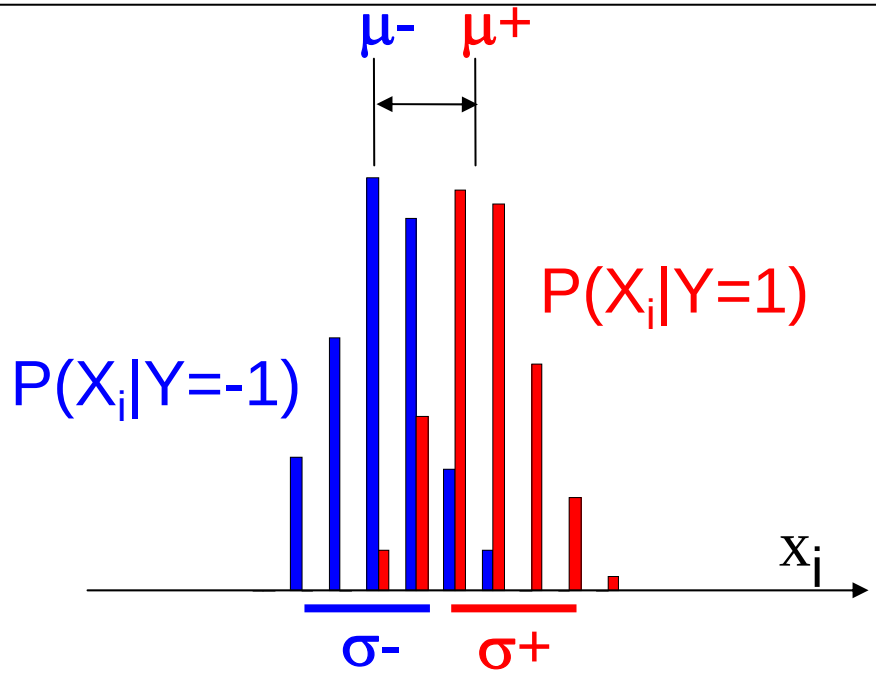


$$MI(X, Y) = -(1/2)\log(1-R^2)$$

Other criteria ( chap. 3)

Method		X			Y			Comments
Name	Formula	B	M	C	B	M	C	
Bayesian accuracy	Eq. 3.1	+	s		+	s		Theoretically the golden standard, rescaled Bayesian relevance Eq. 3.2.
Balanced accuracy	Eq. 3.4	+	s		+	s		Average of sensitivity and specificity; used for unbalanced dataset, same as AUC for binary targets.
Bi-normal separation	Eq. 3.5	+	s		+	s		Used in information retrieval.
F-measure	Eq. 3.7	+	s		+	s		Harmonic of recall and precision, popular in information retrieval.
Odds ratio	Eq. 3.6	+	s		+	s		Popular in information retrieval.
Means separation	Eq. 3.10	+	i	+	+			Based on two class means, related to Fisher's criterion.
T-statistics	Eq. 3.11	+	i	+	+			Based also on the means separation.
Pearson correlation	Eq. 3.9	+	i	+	+	i	+	Linear correlation, significance test Eq. 3.12, or a permutation test.
Group correlation	Eq. 3.13	+	i	+	+	i	+	Pearson's coefficient for subset of features.
χ^2	Eq. 3.8	+	s		+	s		Results depend on the number of samples m.
Relief	Eq. 3.15	+	s	+	+	s	+	Family of methods, the formula is for a simplified version ReliefX, captures local correlations and feature interactions.
Separability Split Value	Eq. 3.41	+	s	+	+	s		Decision tree index.
Kolmogorov distance	Eq. 3.16	+	s	+	+	s	+	Difference between joint and product probabilities.
Bayesian measure	Eq. 3.16	+	s	+	+	s	+	Same as Vajda entropy Eq. 3.23 and Gini Eq. 3.39.
Kullback-Leibler divergence	Eq. 3.20	+	s	+	+	s	+	Equivalent to mutual information.
Jeffreys-Matusita distance	Eq. 3.22	+	s	+	+	s	+	Rarely used but worth trying.
Value Difference Metric	Eq. 3.22	+	s		+	s		Used for symbolic data in similarity-based methods, and symbolic feature-feature correlations.
Mutual Information	Eq. 3.29	+	s	+	+	s	+	Equivalent to information gain Eq. 3.30.
Information Gain Ratio	Eq. 3.32	+	s	+	+	s	+	Information gain divided by feature entropy, stable evaluation.
Symmetrical Uncertainty	Eq. 3.35	+	s	+	+	s	+	Low bias for multivalued features.
J-measure	Eq. 3.36	+	s	+	+	s	+	Measures information provided by a logical rule.
Weight of evidence	Eq. 3.37	+	s	+	+	s	+	So far rarely used.
MDL	Eq. 3.38	+	s		+	s		Low bias for multivalued features.

T-test



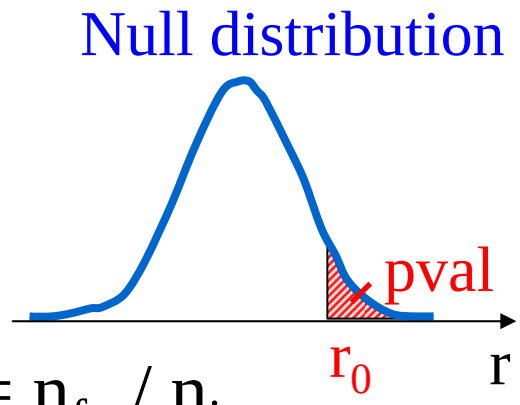
• Normally distributed classes, equal variance σ^2 unknown; estimated from data as σ^2_{within}.

• Null hypothesis H_0: $\mu+ = \mu-$

• T statistic: If H_0 is true,

$$t = (\mu+ - \mu-)/(\sigma_{within}\sqrt{1/m^+ + 1/m^-}) \sim \text{Student}(m^+ + m^- - 2 \text{ d.f.})$$

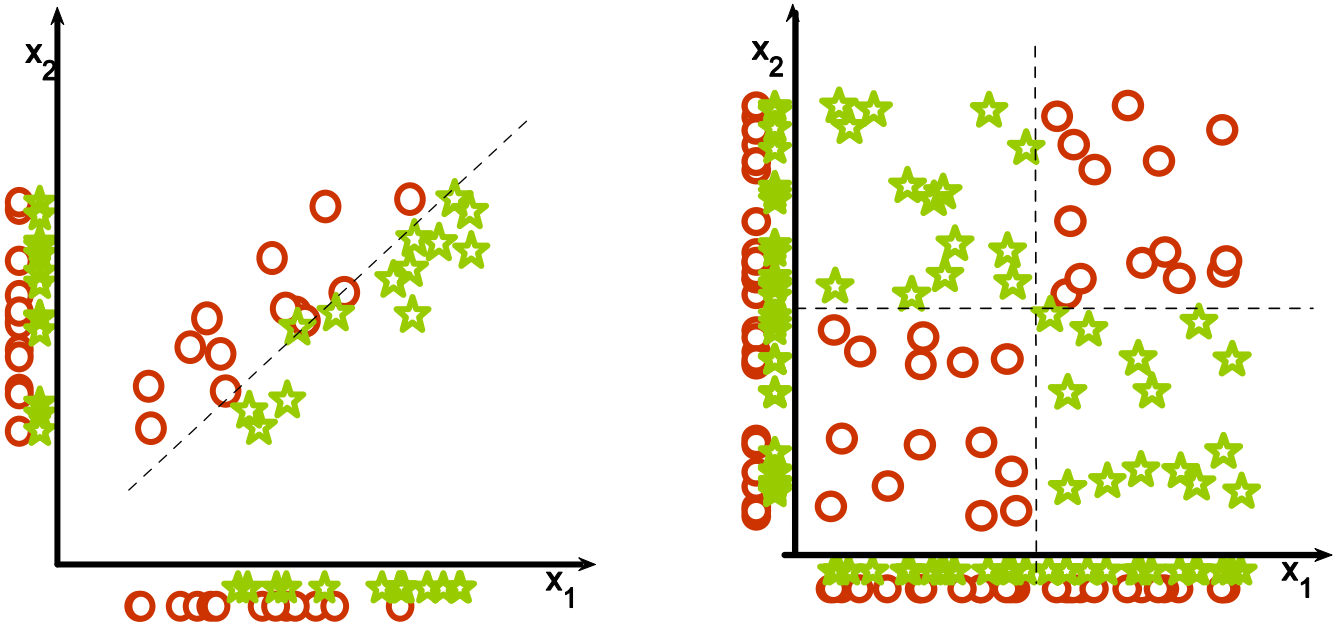
Statistical tests ( *chap. 2)*



- H_0: X and Y are independent.
- Relevance index ⇔ test statistic.
- Pvalue ⇔ false positive rate FPR = n_{fp} / n_{irr}
- Multiple testing problem: use Bonferroni correction pval ← n pval
- False discovery rate: FDR = $n_{fp} / n_{sc} \leq$ FPR n/n_{sc}
- Probe method: FPR ≅ n_{sp}/n_p

Multivariate Methods

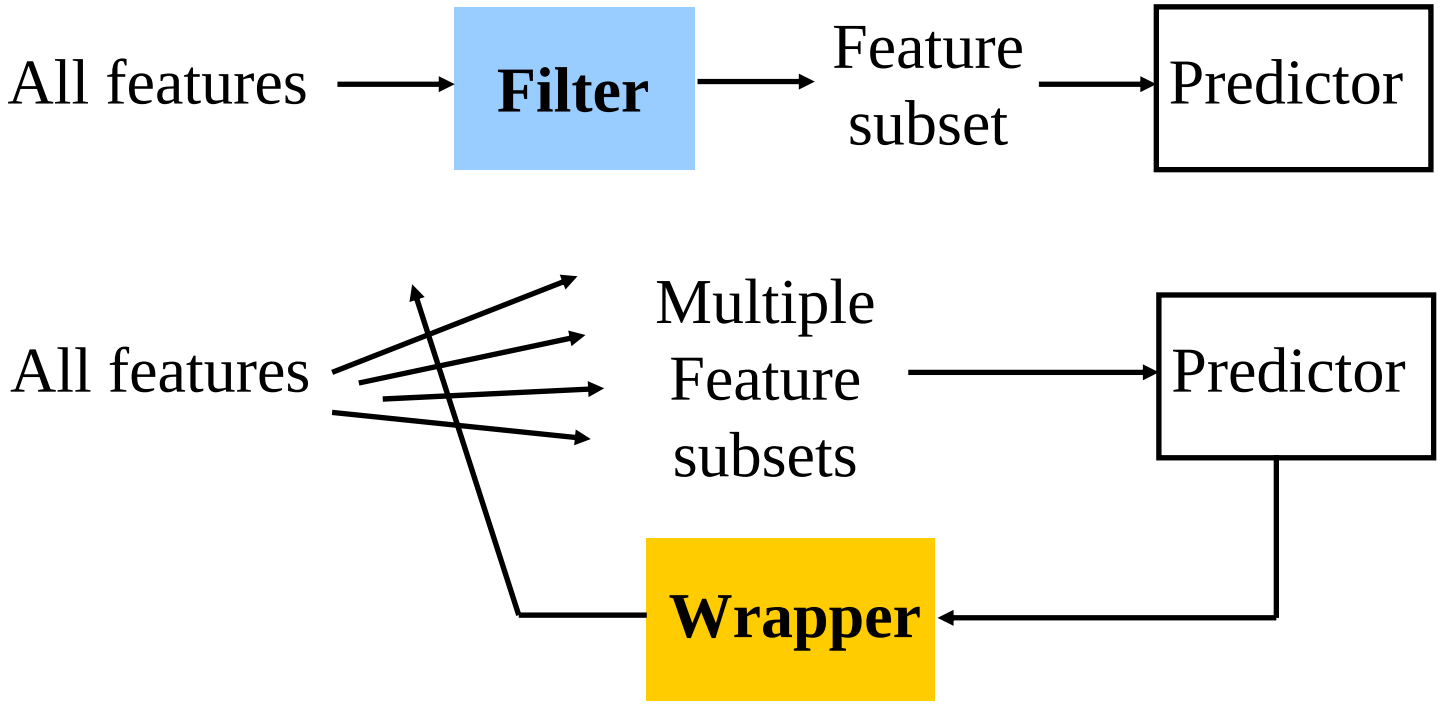
Univariate selection may fail



Guyon-Elisseeff, JMLR 2004; Springer 2006

Filters vs. Wrappers

- **Main goal:** rank subsets of useful features.

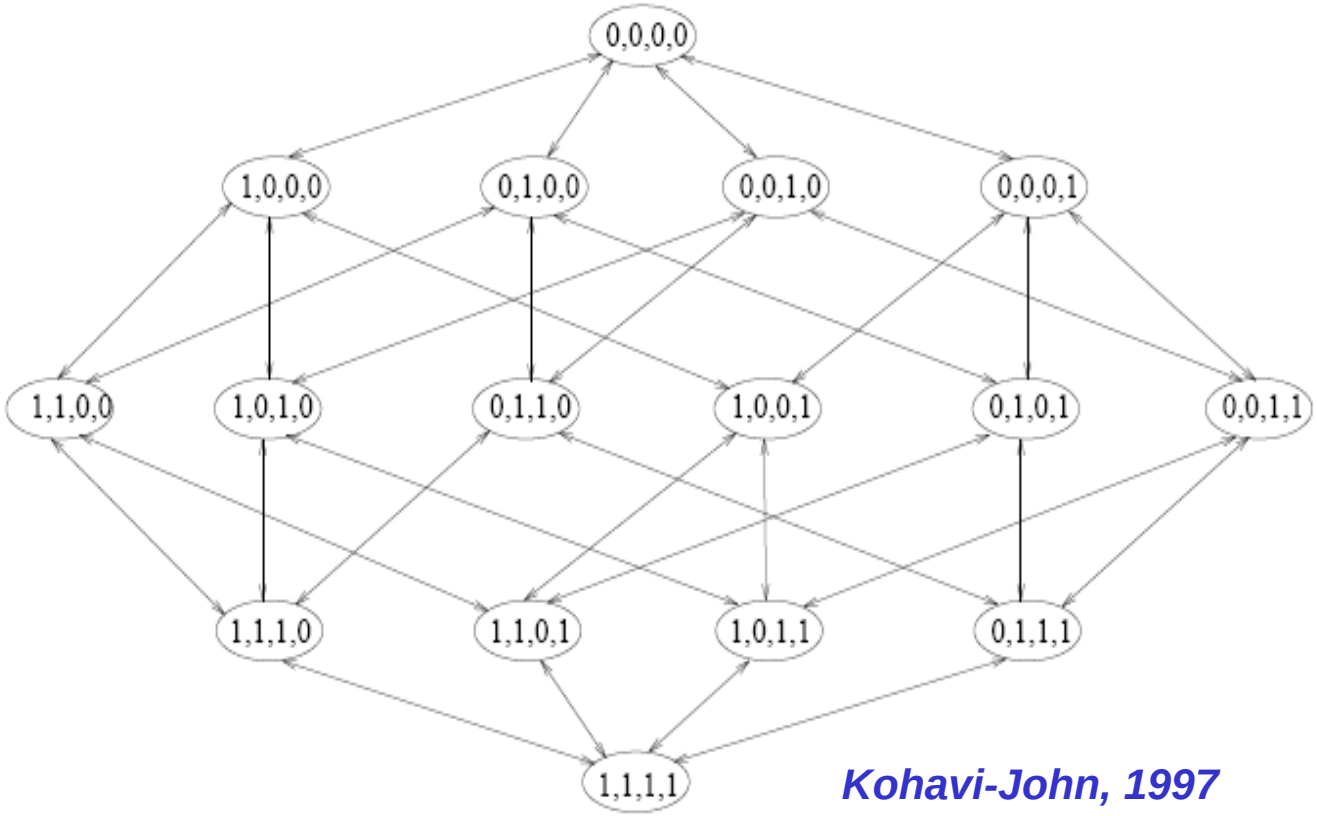


- **Danger of over-fitting** with intensive search!

Search Strategies ( chap. 4)

- **Forward selection** or **backward elimination.**
- **Beam search:** keep k best path at each step.
- **GSFS:** generalized sequential forward selection – when (n-k) features are left try all subsets of g features i.e. $\binom{n-k}{g}$ trainings. More trainings at each step, but fewer steps.
- **PTA(l,r):** plus l , take away r – at each step, run SFS l times then SBS r times.
- **Floating search** (SFFS and SBFS): One step of SFS (resp. SBS), then SBS (resp. SFS) as long as we find better subsets than those of the same size obtained so far. Any time, if a better subset of the same size was already found, switch abruptly.

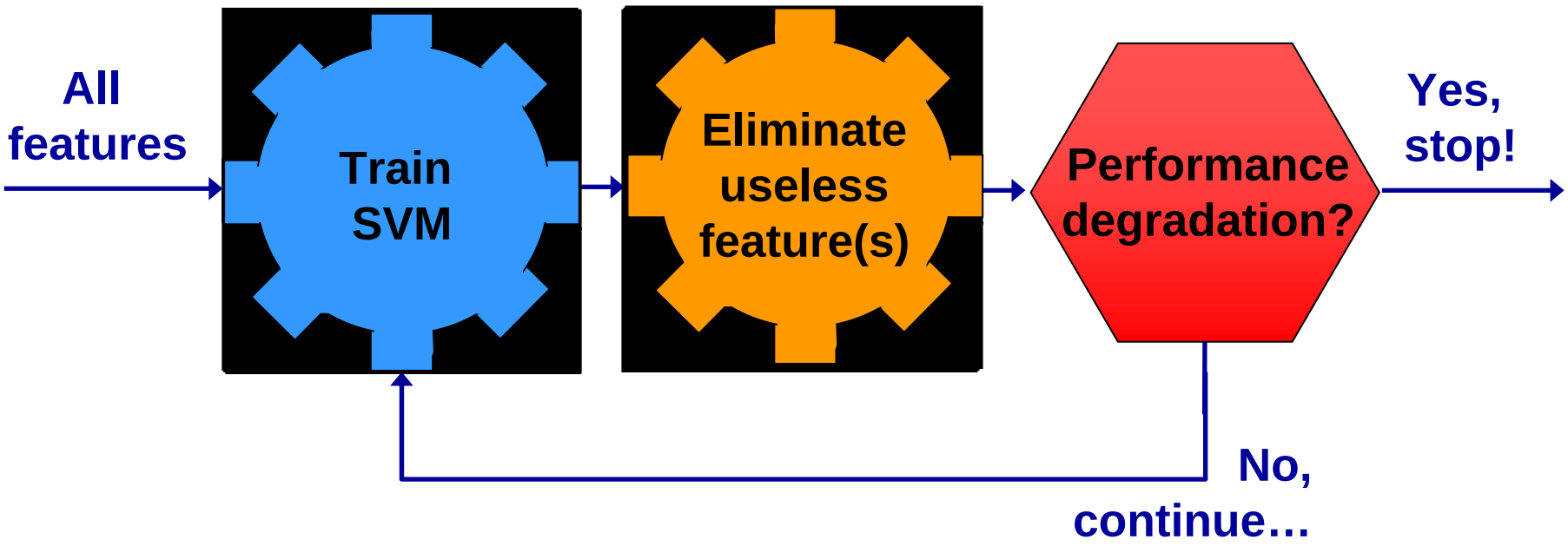
Multivariate FS is complex



Kohavi-John, 1997

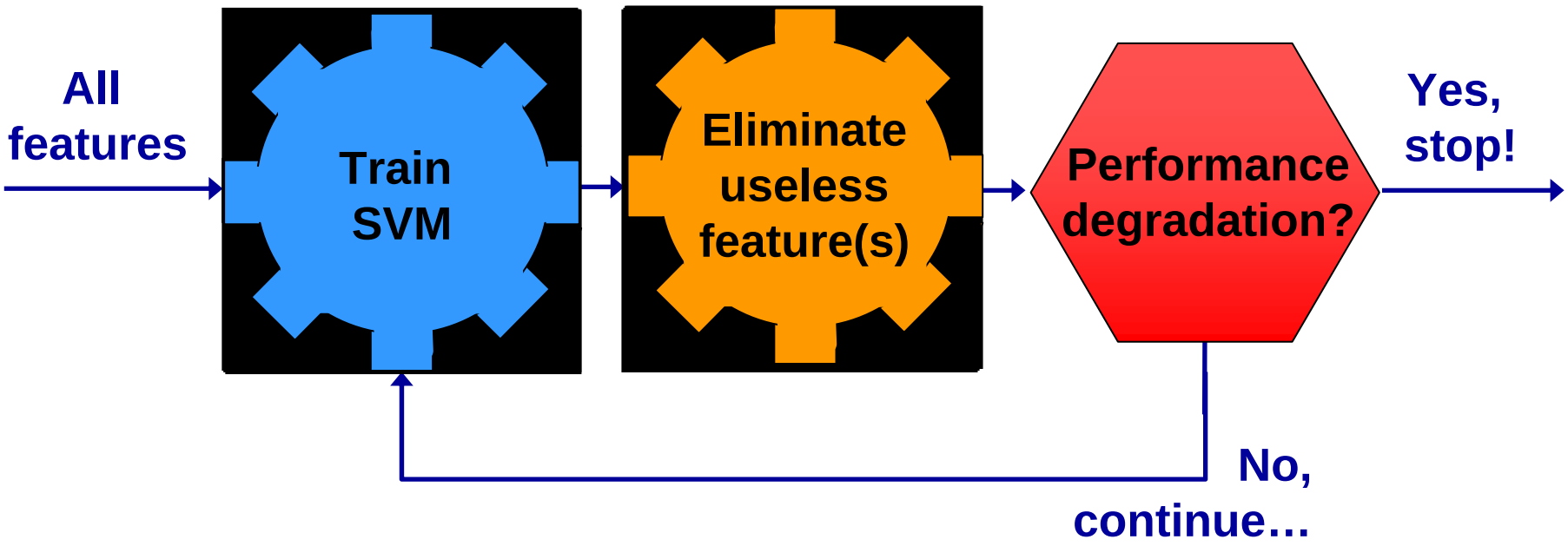
N features, 2^N possible feature subsets!

Embedded methods



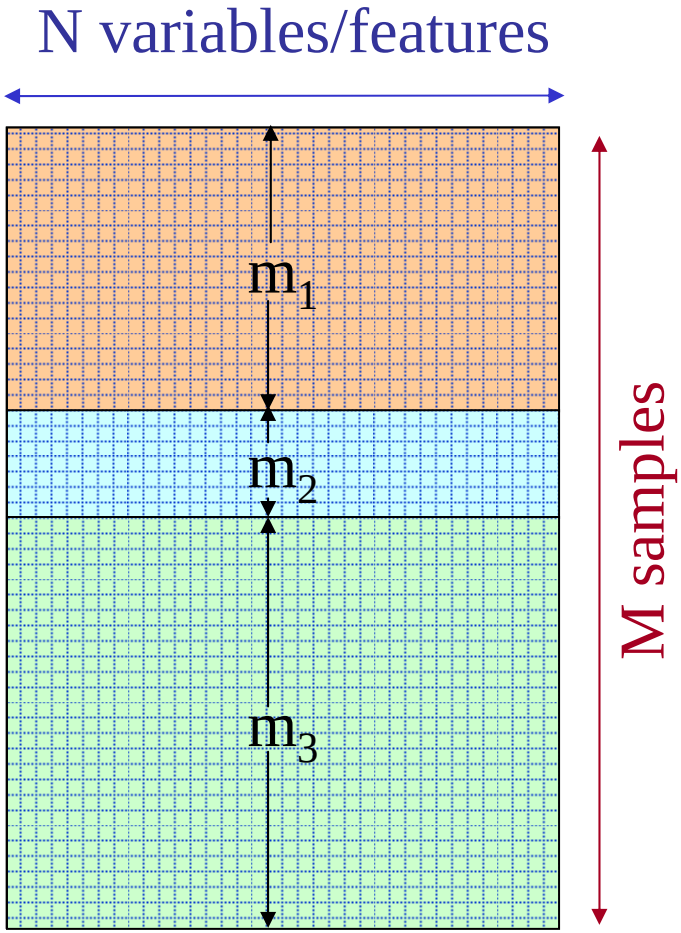
Recursive Feature Elimination (RFE) SVM. *Guyon-Weston, 2000. US patent 7,117,188*

Embedded methods



Recursive Feature Elimination (RFE) SVM. *Guyon-Weston, 2000. US patent 7,117,188*

Feature subset assessment



Split data into 3 sets:

training, validation, and test set.

1) For each feature subset, train predictor on training data.

2) Select the feature subset, which performs best on validation data.

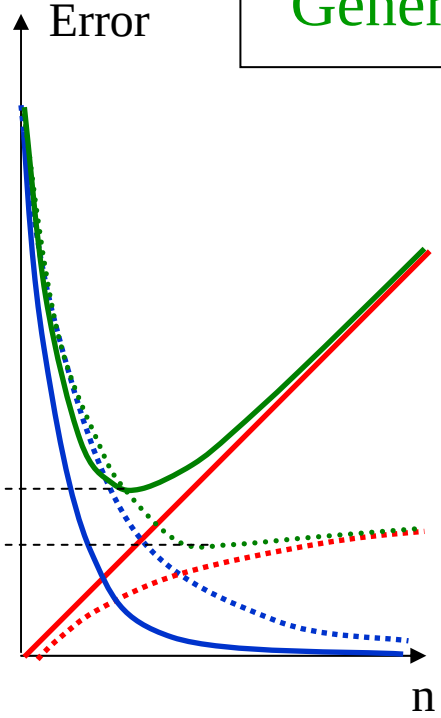
   – Repeat and average if you want to reduce variance (cross-validation).

3) Test on test data.

Complexity of Feature Selection

With high probability:

$$\text{Generalization\_error} \leq \text{Validation\_error} + \varepsilon(C/m_2)$$



Method	**Number of subsets tried**	**Complexity C**
Exhaustive search wrapper	2^N	N
Nested subsets Feature ranking	N(N+1)/2 or N	log N

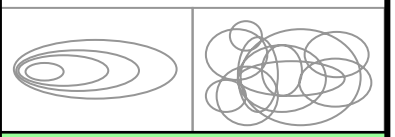
m_2: number of *validation* examples,
N: total number of features,
n: feature subset size.

Try to keep C of the order of m_2.

Examples of FS algorithms

keep C = O(m_2)

keep C = O(m_1)

	Univariate	Multivariate
		
Linear	T-test, AUC, feature ranking	RFE with linear SVM or LDA
Non-linear 	Mutual information feature ranking	Nearest Neighbors Neural Nets Trees, SVM

In practice...

- **No method is universally better:**
  - **wide variety of types of variables, data distributions, learning machines, and objectives.**
- **Match the method complexity to the ratio M/N:**
  - univariate feature selection may work better than multivariate feature selection; non-linear classifiers are not always better.
- **Feature selection is not always necessary to achieve good performance.**

NIPS 2003 and WCCI 2006 challenges : **http://clopinet.com/challenges**

Book of the NIPS 2003 challenge

Feature Extraction, Foundations and Applications
I. Guyon et al, Eds.
Springer, 2006.
http://clopinet.com/fextract-book