Support vector machine evaluation of peptide identification via mass spectrometry

William Stafford Noble

Department of Genome Sciences

Department of Computer Science and Engineering

University of Washington

SVMs in computational biology

- Splice site recognition
- Protein sequence similarity detection
- Protein functional classification
- Regulatory module search
- Protein-protein interaction prediction
- Gene functional classification from microarray data
- Cancer classification from microarray data

- Dave Anderson

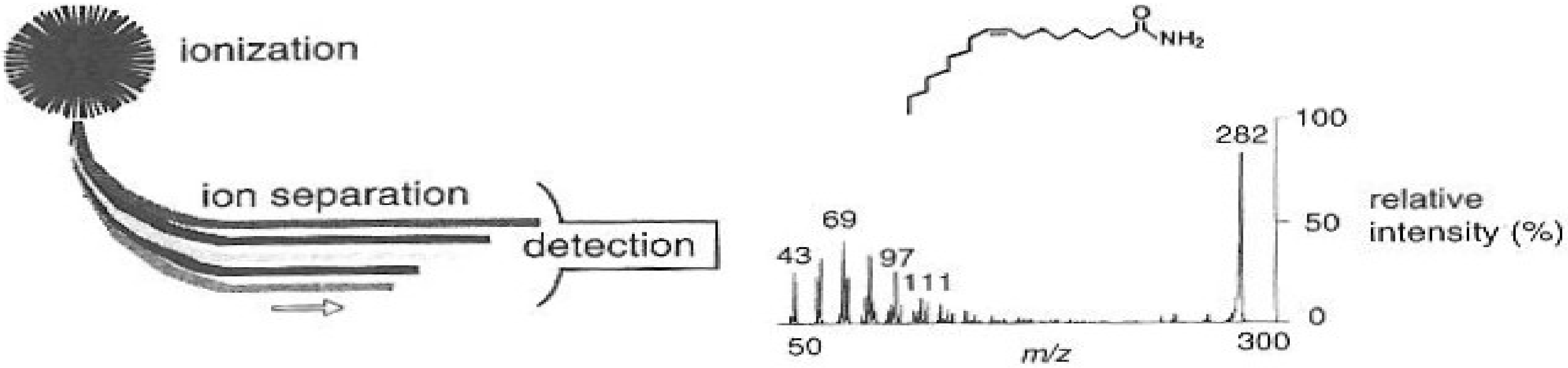
- Don Payan
- Weiqun Li

Rigel, Inc.

Outline

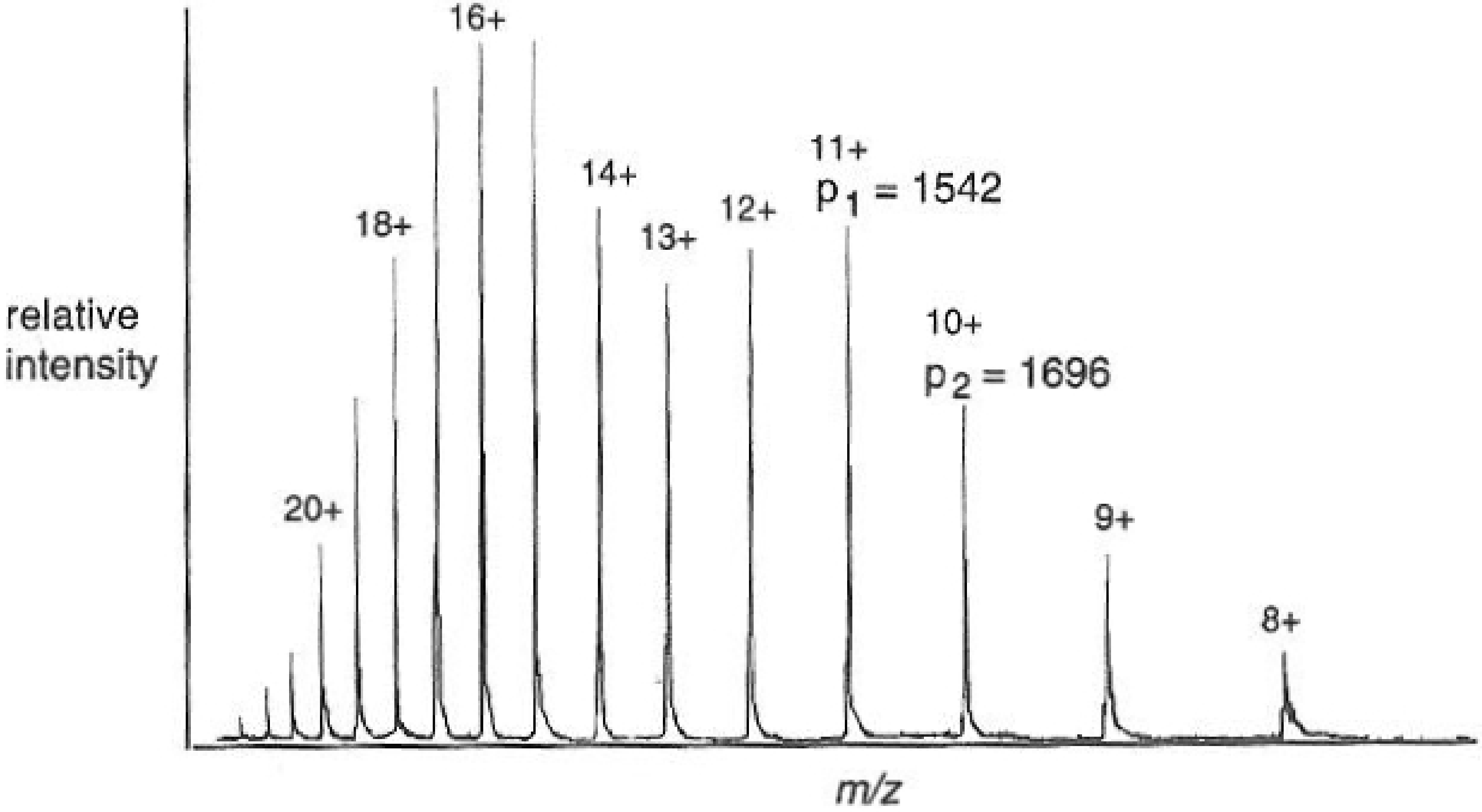
- The task: Recognizing correct peptide matches.
- The data: Thirteen informative features.
- The algorithm: The support vector machine.
- The results: Multiple data sets, and comparison to other approaches.

Protein mass spectrometry



(Siuzdak 1996)

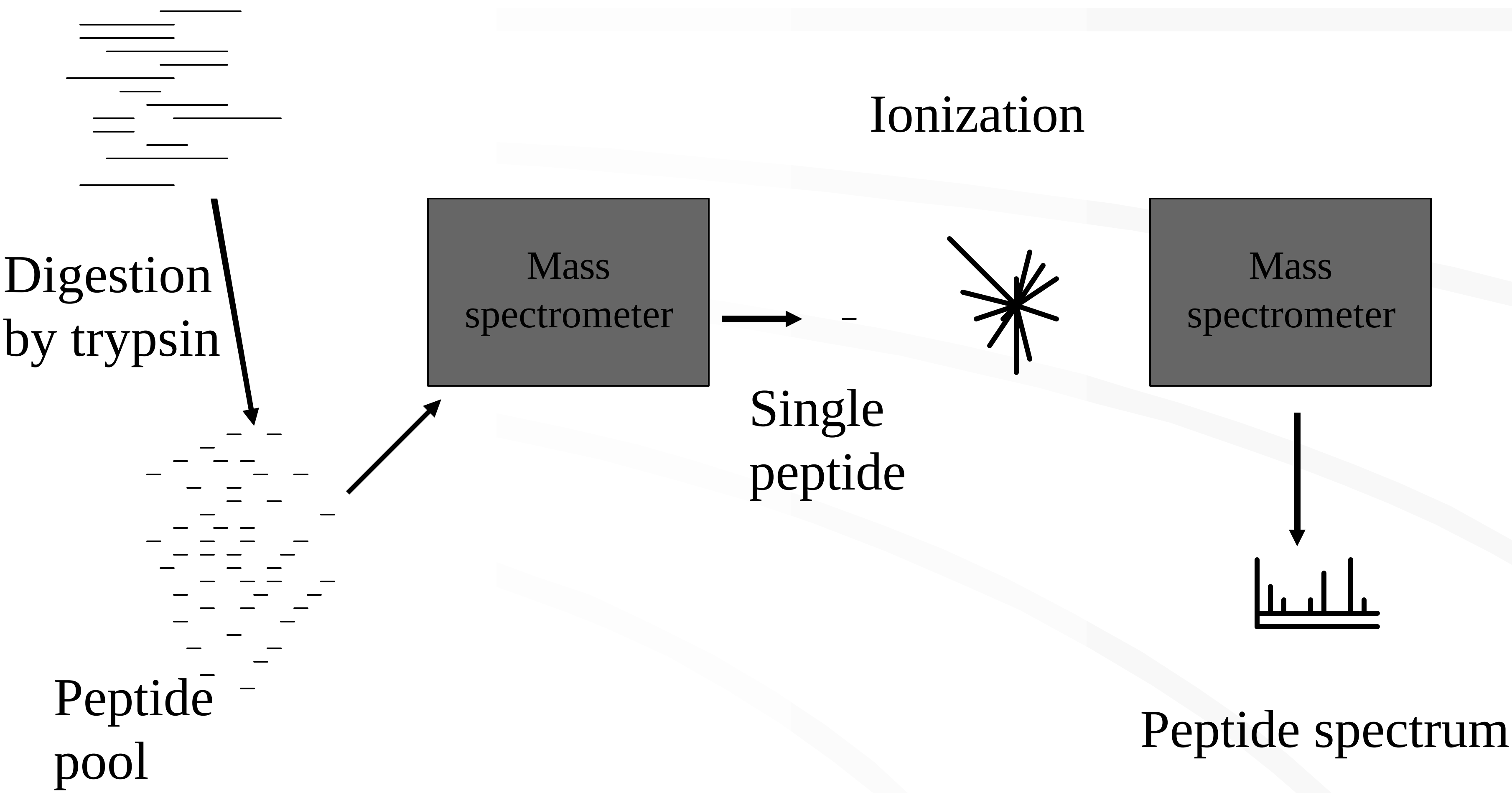
Mass spectrum



(Siuzdak 1996)

Protein sample

Tandem mass spectrometry



Digestion by trypsin

Peptide pool

Mass spectrometer

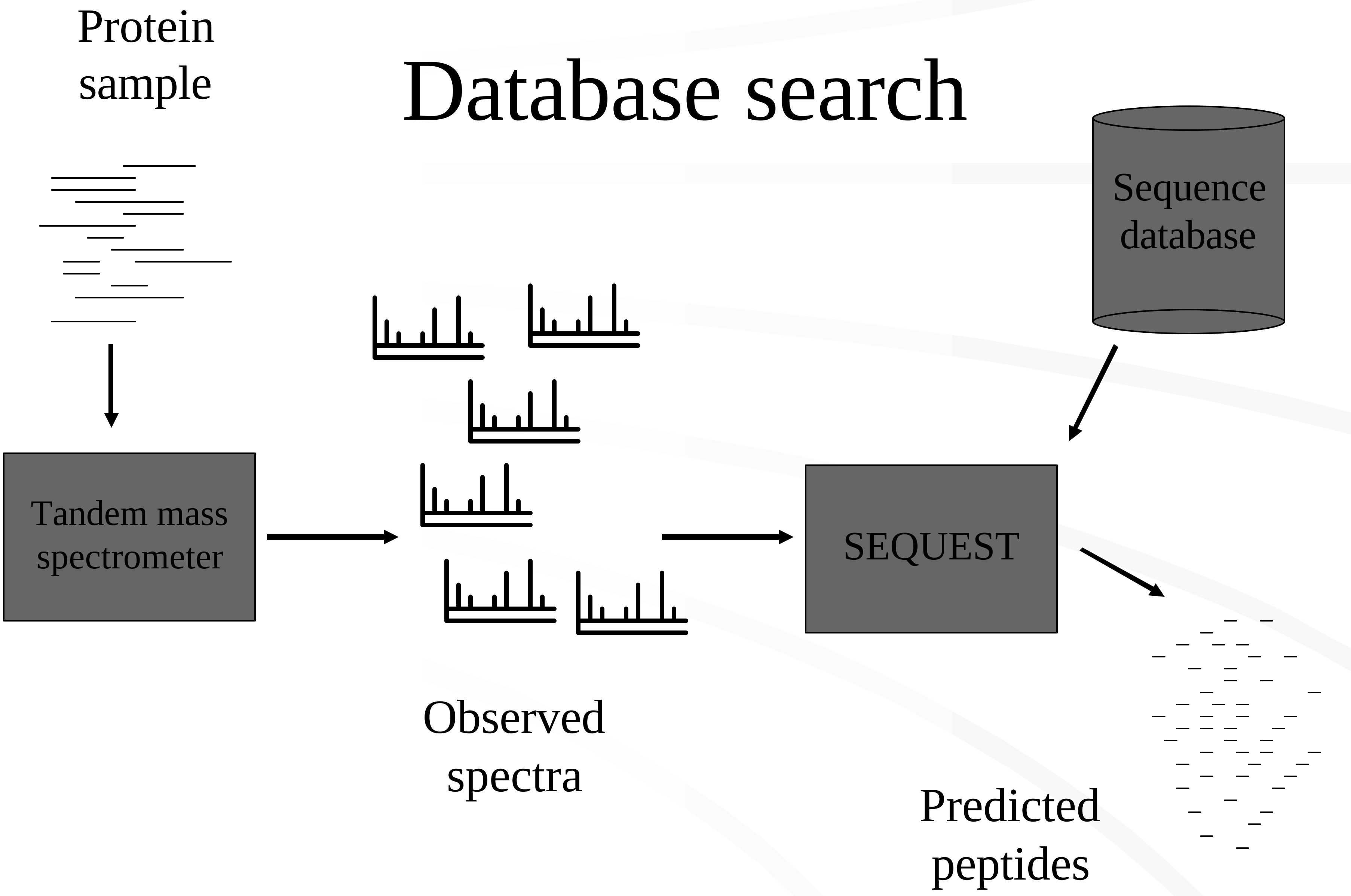
Single peptide

Ionization

Mass spectrometer

Peptide spectrum

Database search



Protein sample

Tandem mass spectrometer

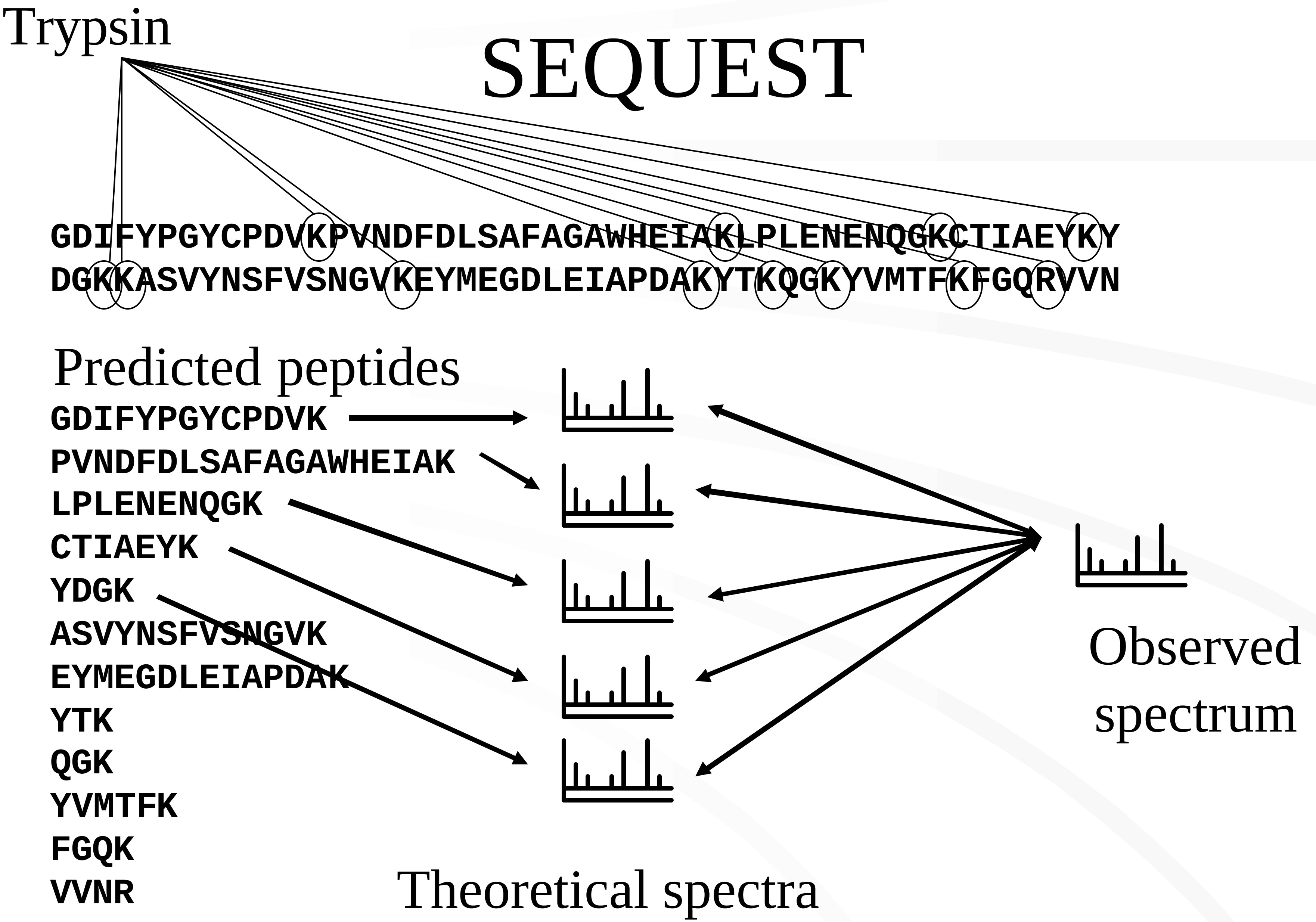
Observed spectra

Sequence database

SEQUEST

Predicted peptides

SEQUEST



Trypsin

GDIFYPGYCPDVKPVNDFDLSAFAGAWHEIAKLPLENENQGKCTIAEYKY
DGKKASVYNSFVSNGVKEYMEGDLEIAPDAKYTKQGKYVMTFKFGQRVVN

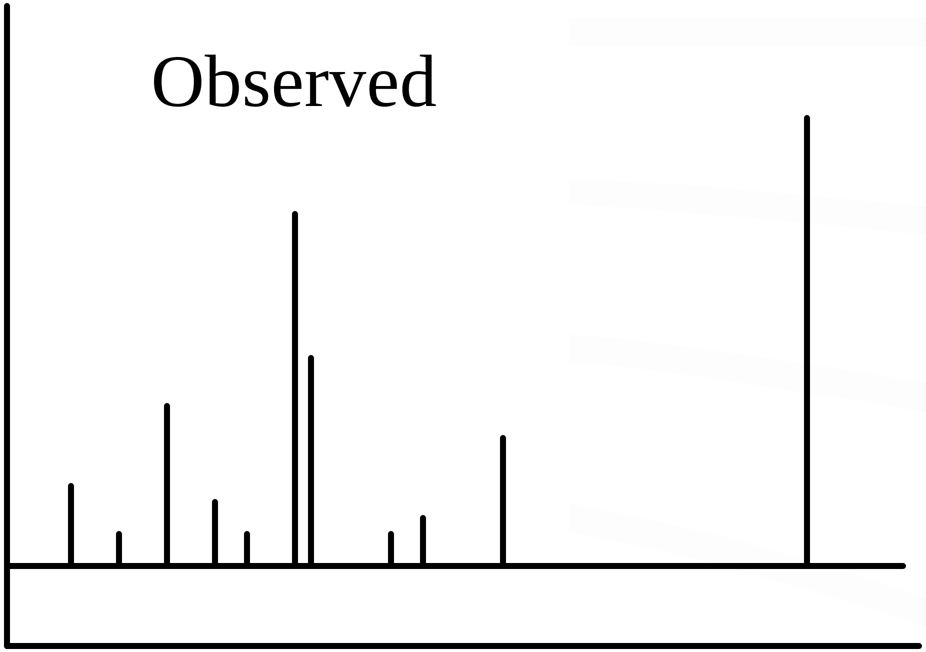
Predicted peptides

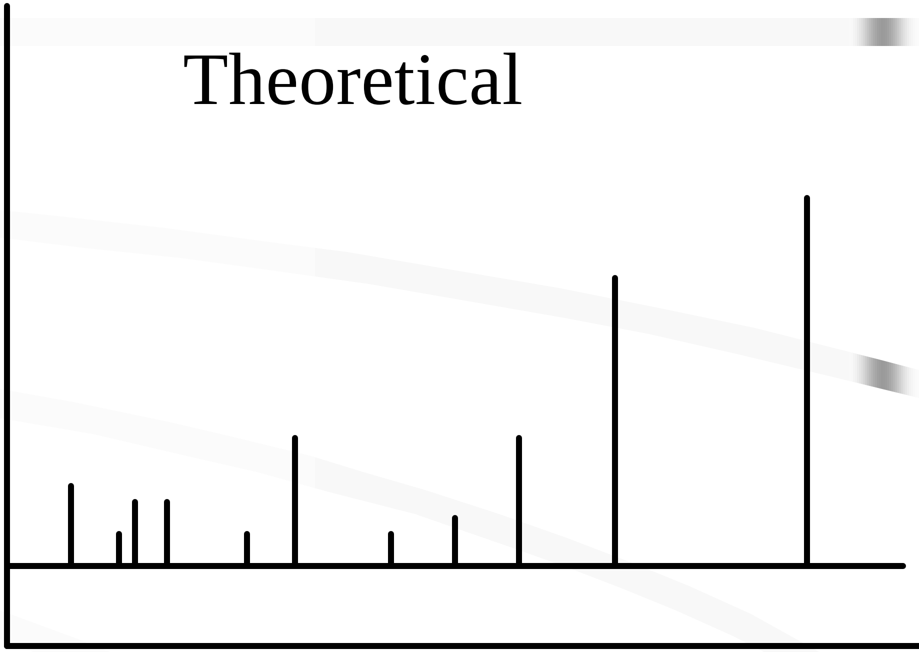
GDIFYPGYCPDVK
PVNDFDLSAFAGAWHEIAK
LPLENENQGK
CTIAEYK
YDGK
ASVYNSFVSNGVK
EYMEGDLEIAPDAK
YTK
QGK
YVMTFK
FGQK
VVNR

Theoretical spectra

Observed spectrum

The learning task





- We are given SEQUEST output: paired observed and theoretical spectra.
- Question: Is the pairing correct?

We need to choose

- the feature set and
- the learning algorithm.

Properties of the observed spectrum

- Total peptide mass. Too small yields little information; too large (>25 amino acids) yields uneven fragmentation.
- Charge (+1, +2 or +3). Provides some evidence about amino acid composition.
- Total ion current. Proportional to the amount of peptide present.
- Peak count. Small indicates poor fragmentation; large indicates noise.

Observed vs. theoretical spectra

- Mass difference.
- Percent of ions matched.
- Percent of peaks matched.
- Percent of peptide fragment ion current matched.
- Preliminary SEQUEST score.
- Cross-correlation.
- Cross-correlation rank.

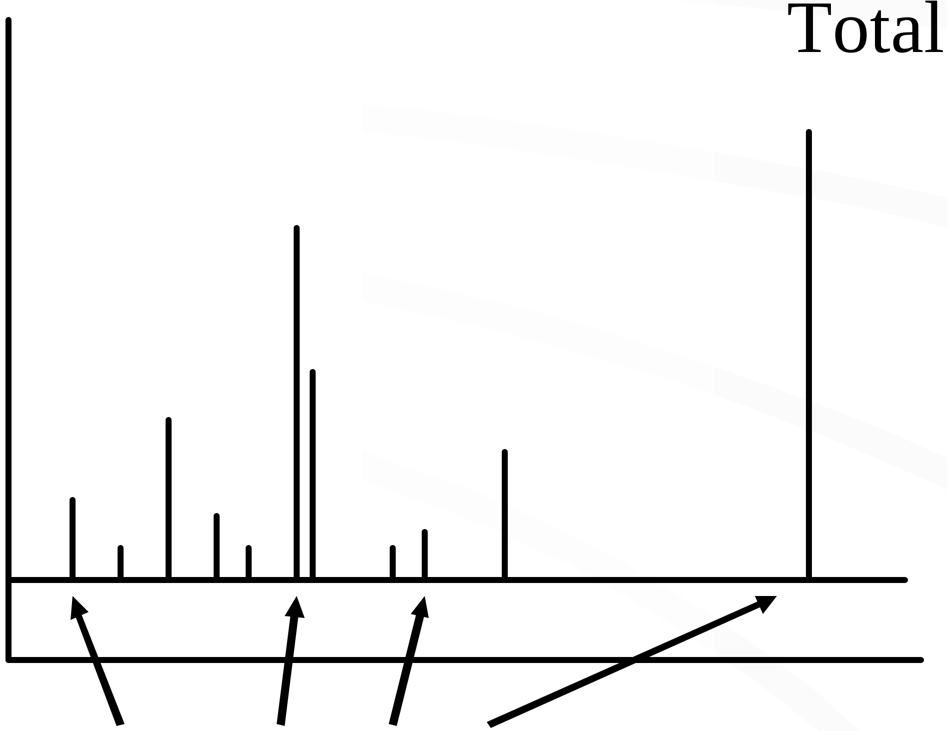
Percent of ions matched

YCPDVK	*
YCPDV	*
YCPD	
YCP	*
YC	
Y	*
CPDVK	*
PDVK	
DVK	*
VK	*
K	

$$\frac{\text{7 matched ions}}{\text{11 possible ions}} = 63.6\%$$

Percent of peaks matched

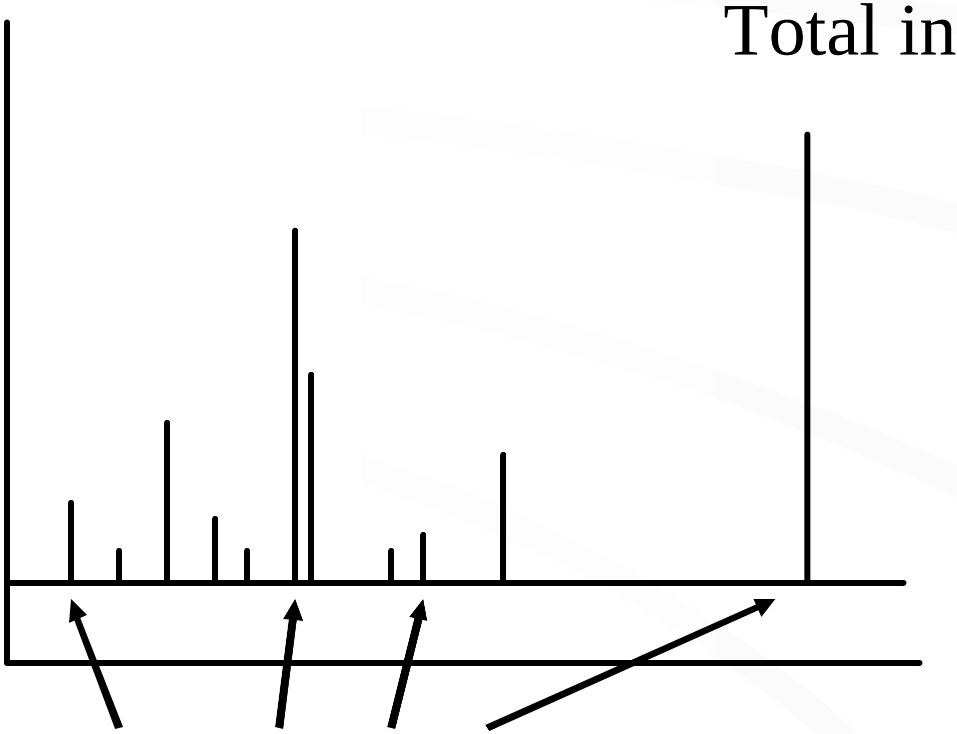
$$\frac{\text{Number of matched peaks}}{\text{Total number of peaks}}$$



Observed peaks

Percent of peptide fragment ion current matched.

$$\frac{\text{Total intensity of matched peaks}}{\text{Total intensity of all peaks}}$$



Observed peaks

This metric weights for matching large peaks

Preliminary SEQUEST score

The score S_p

- is only computed for pairs within a defined mass tolerance,
- accounts for percent of ion matches, continuity, and other factors, and
- can be computed efficiently.

Cross-correlation

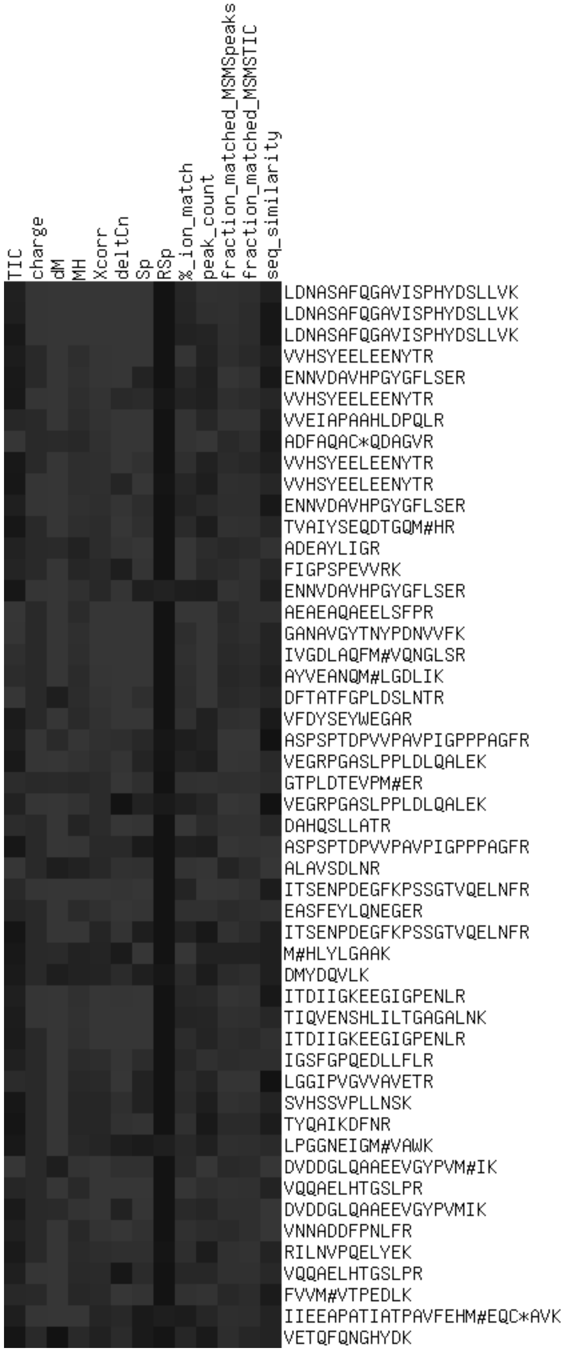
$$R_t(x,y)=\sum_{i=0}^{n-1} x[i]y[i+t]$$

- Theoretical and observed spectra are x and y, and t is the offset between them.
- The correlation is computed via FFT.
- C_n (a.k.a. Xcorr) is the maximal R_t divided by the mean R_t for $-75 < t < 75$, normalized to 1.0.
- Cross-correlation is only computed for the top-scoring 500 peptides.
- Correlation rank is the location of the theoretical spectrum in a list ranked by cross-correlation.

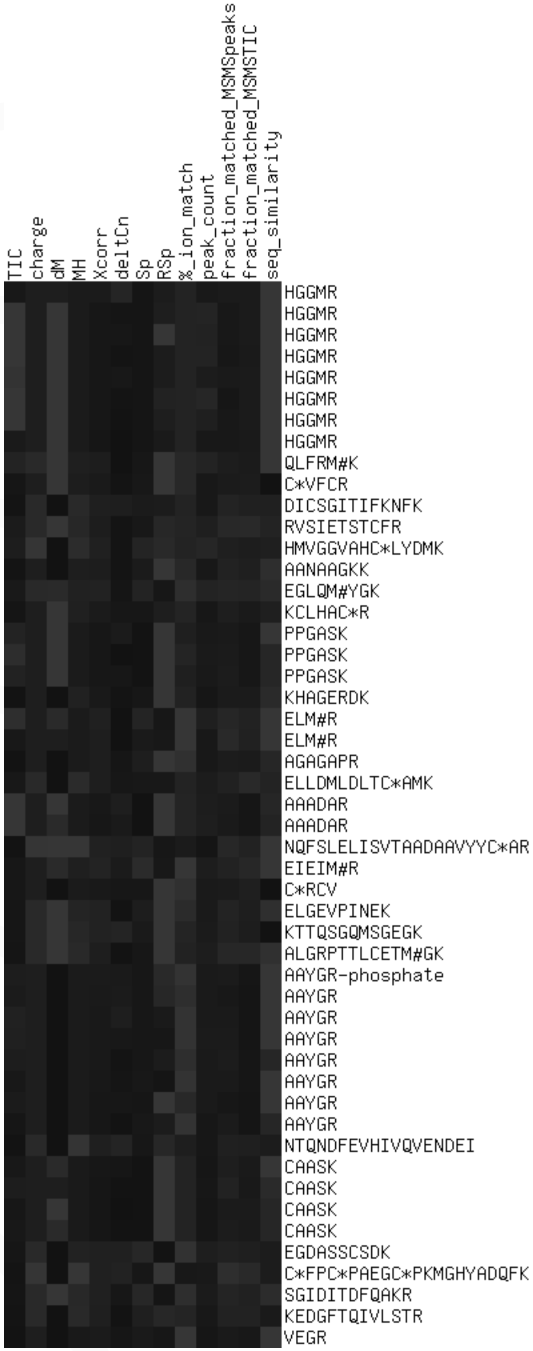
Top-ranked vs. second-ranked peptides

- Change in cross-correlation. Compute the difference in C_n for the top-ranked and second-ranked peptide. 0.1 or greater indicates a significant difference between the first two choices.
- Percent sequence identity. Usually anti-correlated with change in cross-correlation.

Positive examples

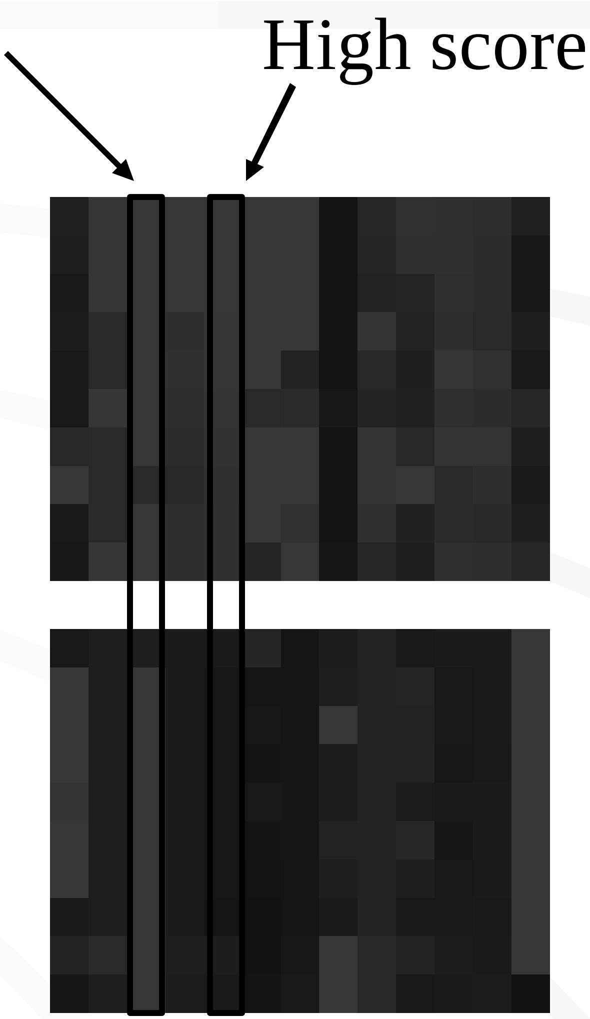


Negative examples



Fisher criterion score

$$\frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}$$



Feature ranking

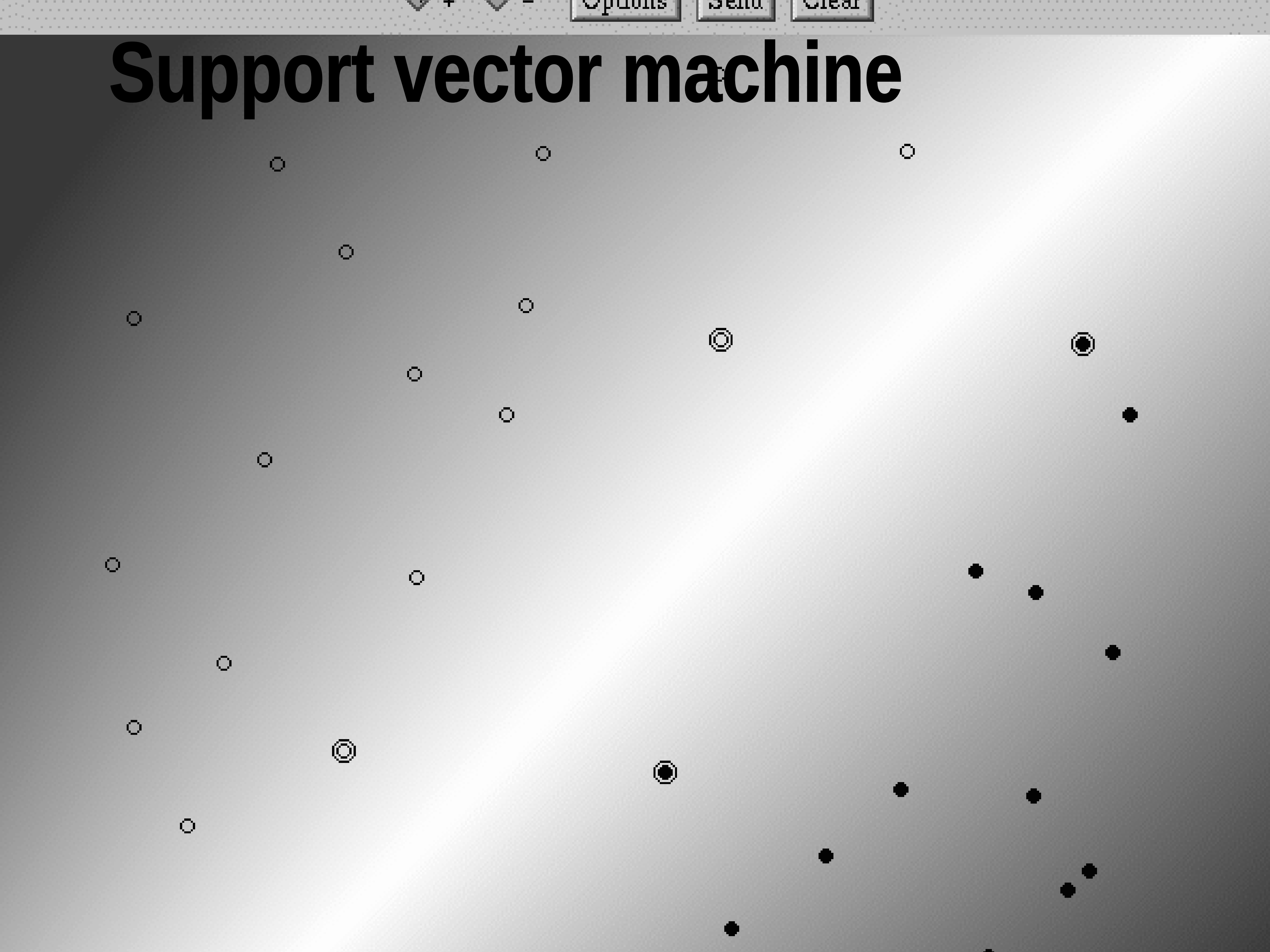
delta C_n	2.861
% match total ion current	2.804
C_n	2.444
% match peaks	2.314
S_p	1.158
mass	0.704
charge	0.488
rank S_p	0.313
peak count	0.209
sequence similarity	0.115
% ion match	0.079
total ion current	0.026
delta mass	0.024

Pairwise feature ranking

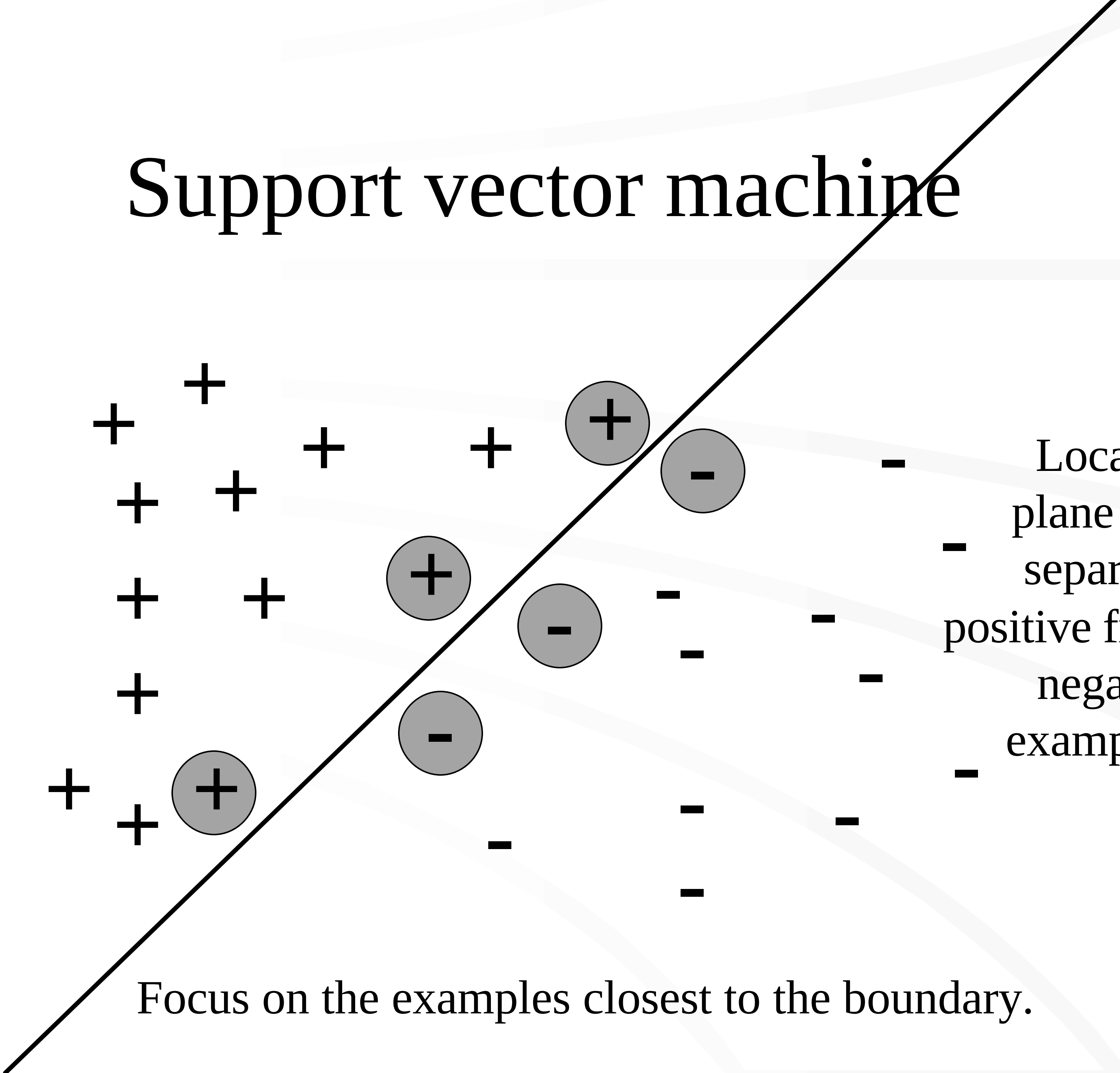
% match TIC-delta Cn	4.741
% match peaks-delta Cn	4.233
% match TIC-Cn	3.819
delta Cn-Cn	3.597
delta Cn-charge	3.563
% match peaks-Cn	3.377
delta Cn-mass	3.119
% match TIC-% match peaks	2.823
% ion match-delta Cn	2.812
Sp-delta Cn	2.799
% match TIC-Sp	2.579
% match peaks-Sp	2.383

% match TIC-mass	2.097
% ion match-mass	2.091
Cn-charge	1.943
Sp-mass	1.922
% match TIC-charge	1.898
Cn-mass	1.884
Sp-Cn	1.881
% ion match-Cn	1.827
Sp-charge	1.770
% match peaks-mass	1.668
% match peaks-charge	1.528
% match TIC-% ion match	1.473

Support vector machine



Support vector machine



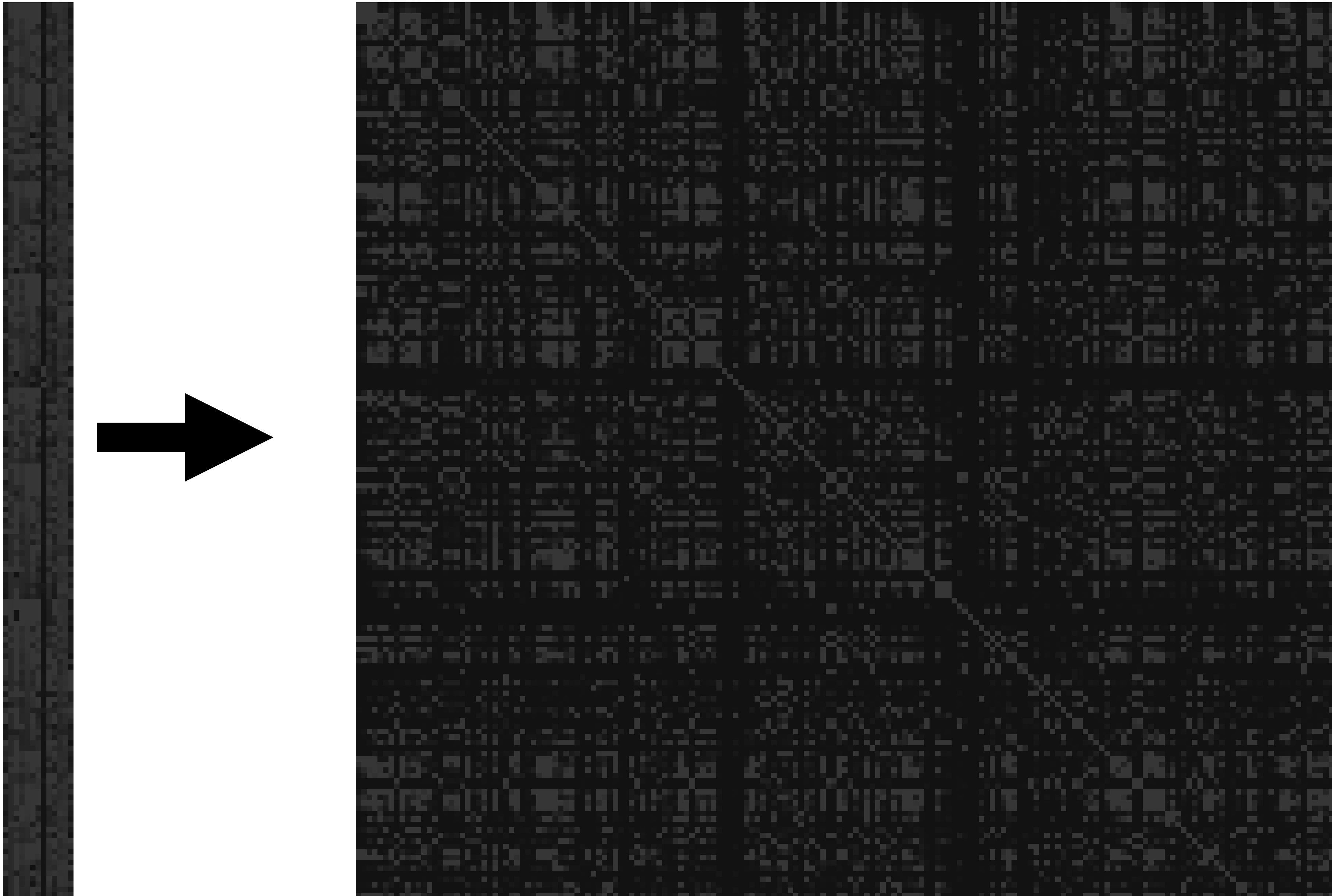
Locate a plane that separates positive from negative examples.

Focus on the examples closest to the boundary.

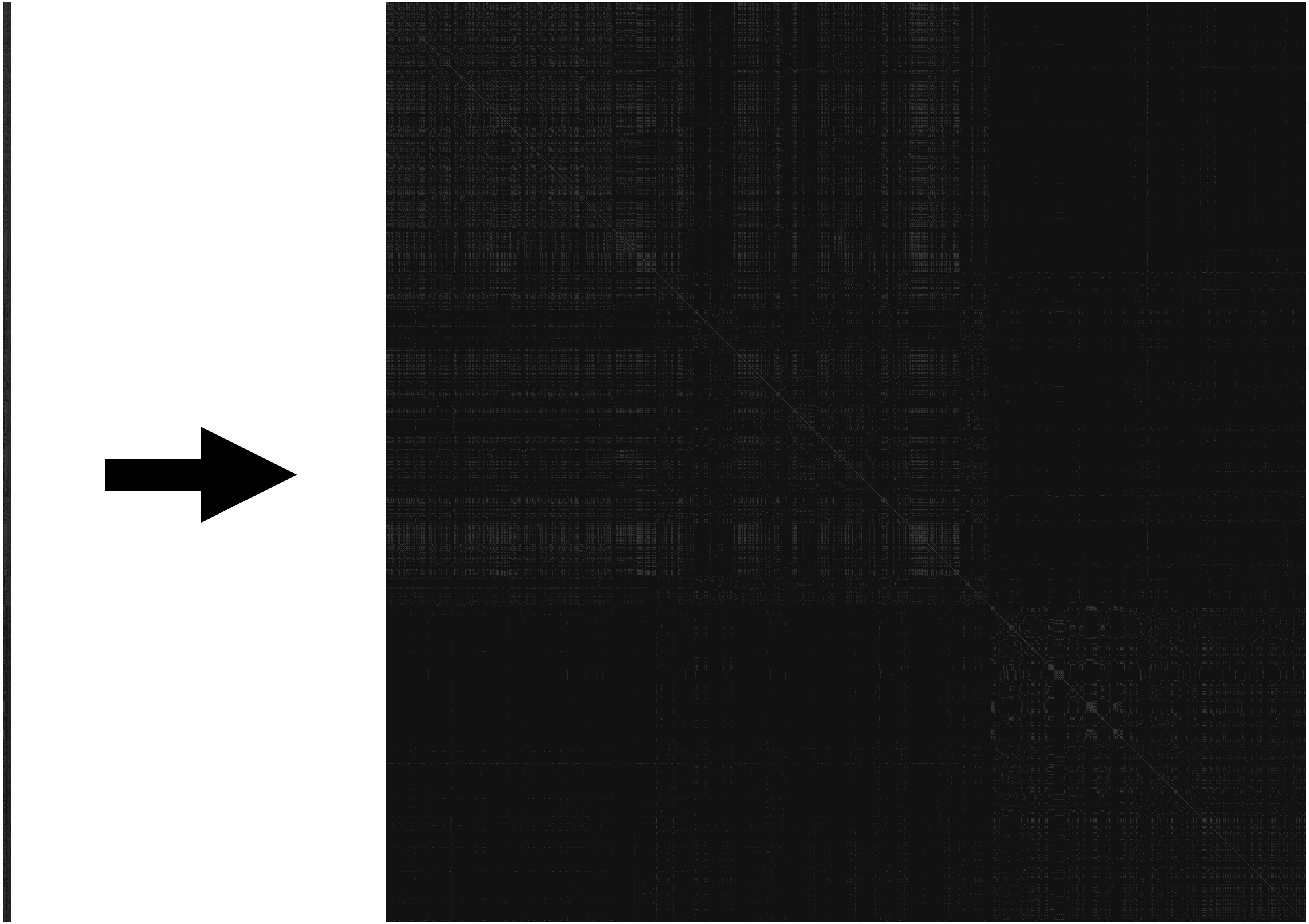
Support vector machine learning

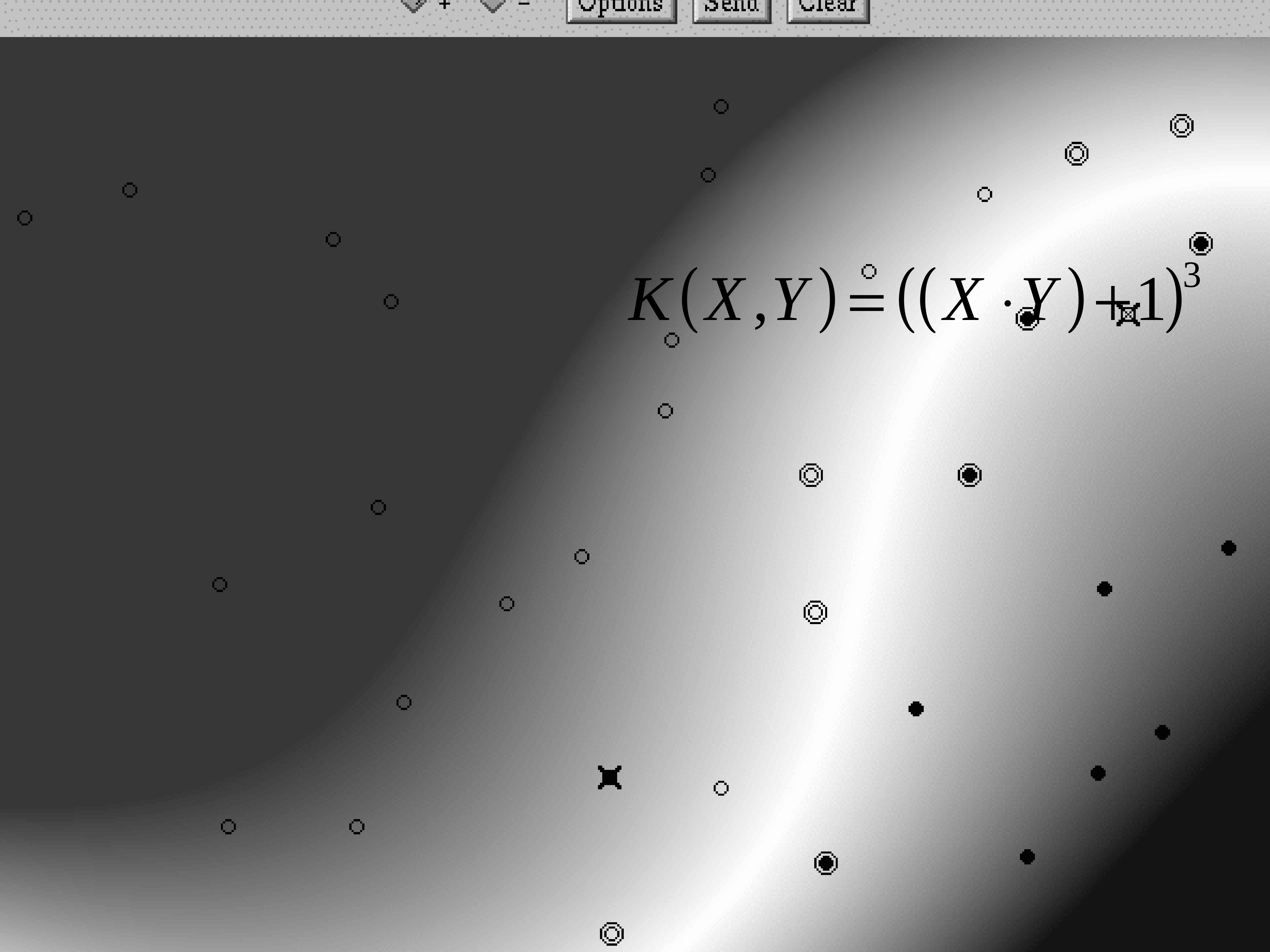
- The SVM learning algorithm finds a linear decision boundary.
- The hyperplane maximizes the margin; i.e., the distance from any training example.
- The optimization is convex; the solution is sparse.
- A soft margin allows for noise in the training set.
- A complex decision surface can be learned by using a non-linear kernel function.

Kernel matrix representation

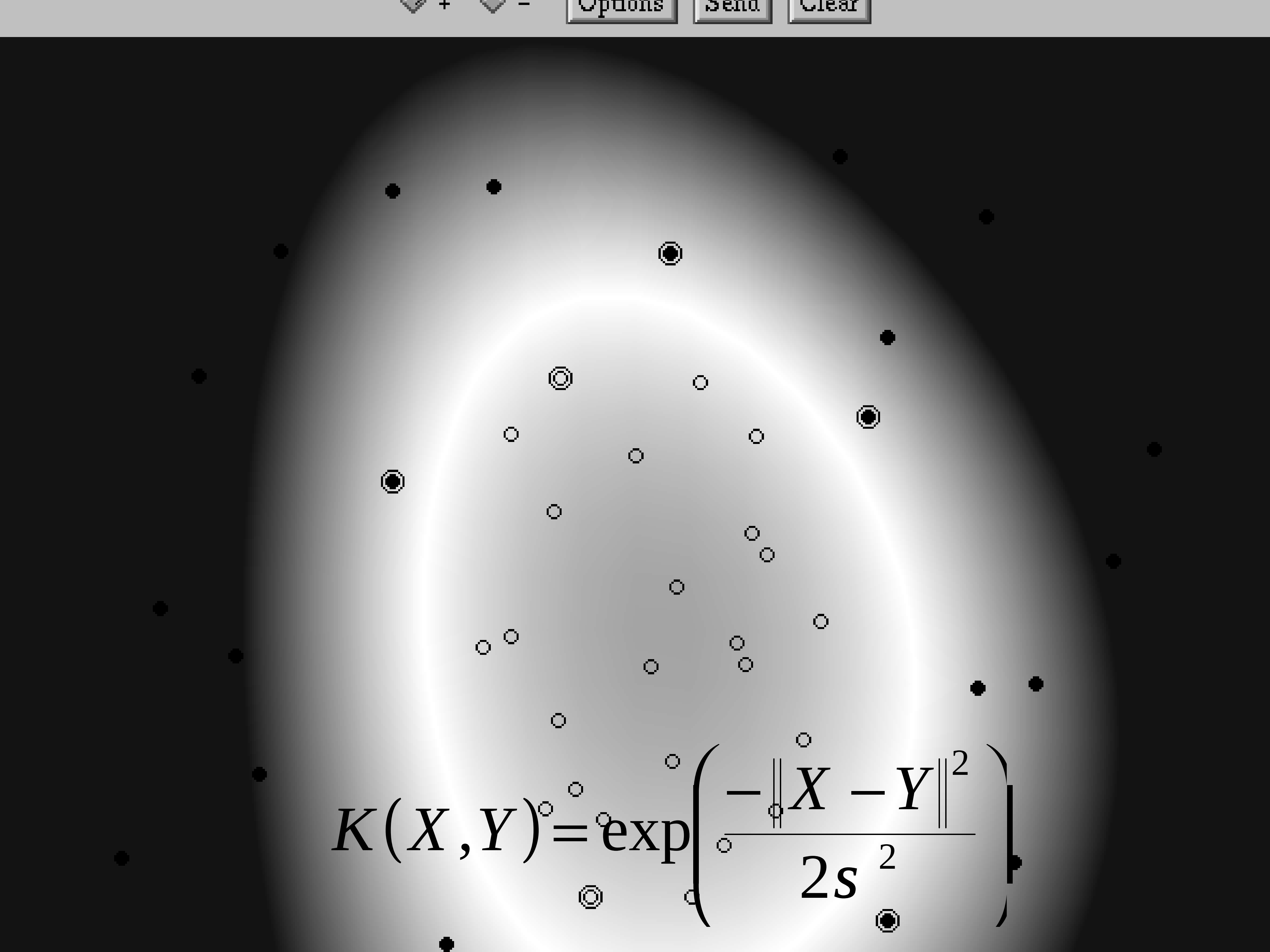


Kernel matrix representation





$$K(X,Y) = ((X \cdot Y) + 1)^3$$



$$K(X,Y)=\exp\left(\frac{-\left\|X-Y\right\|^{2}}{2s^{2}}\right)$$

Kernel function

- Let p(x,y) be the function that computes a 13-element vector of parameters for a pair of spectra, x and y.
- The kernel function K operates on pairs of observed and theoretical spectra:

$$K\left(S_o^A : S_t^A, S_o^B : S_t^B\right) = K\left(p\left(S_o^A, S_t^A\right), p\left(S_o^B, S_t^B\right)\right)$$

$$= \left(p\left(S_o^A, S_t^A\right) \cdot p\left(S_o^B, S_t^B\right) + 1\right)^2$$

Experimental design

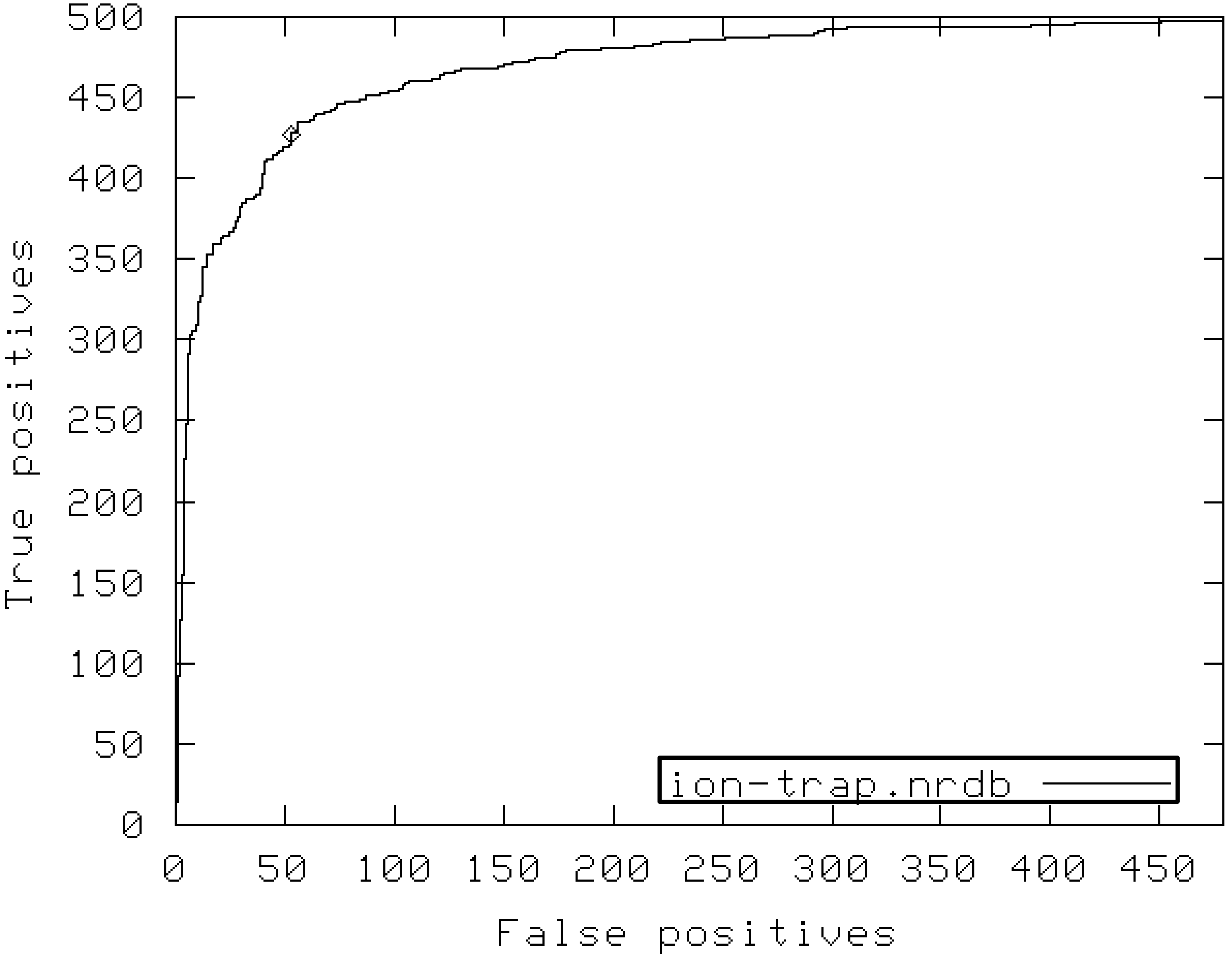
- Data consists of one 13-element vector per predicted peptide.
- Each feature is normalized to sum to 1.0 across all examples.
- The SVM is tested using leave-one-out cross-validation.
- The SVM uses a second-degree polynomial, normalized kernel with a 2-norm asymmetric soft margin.

Three data sets

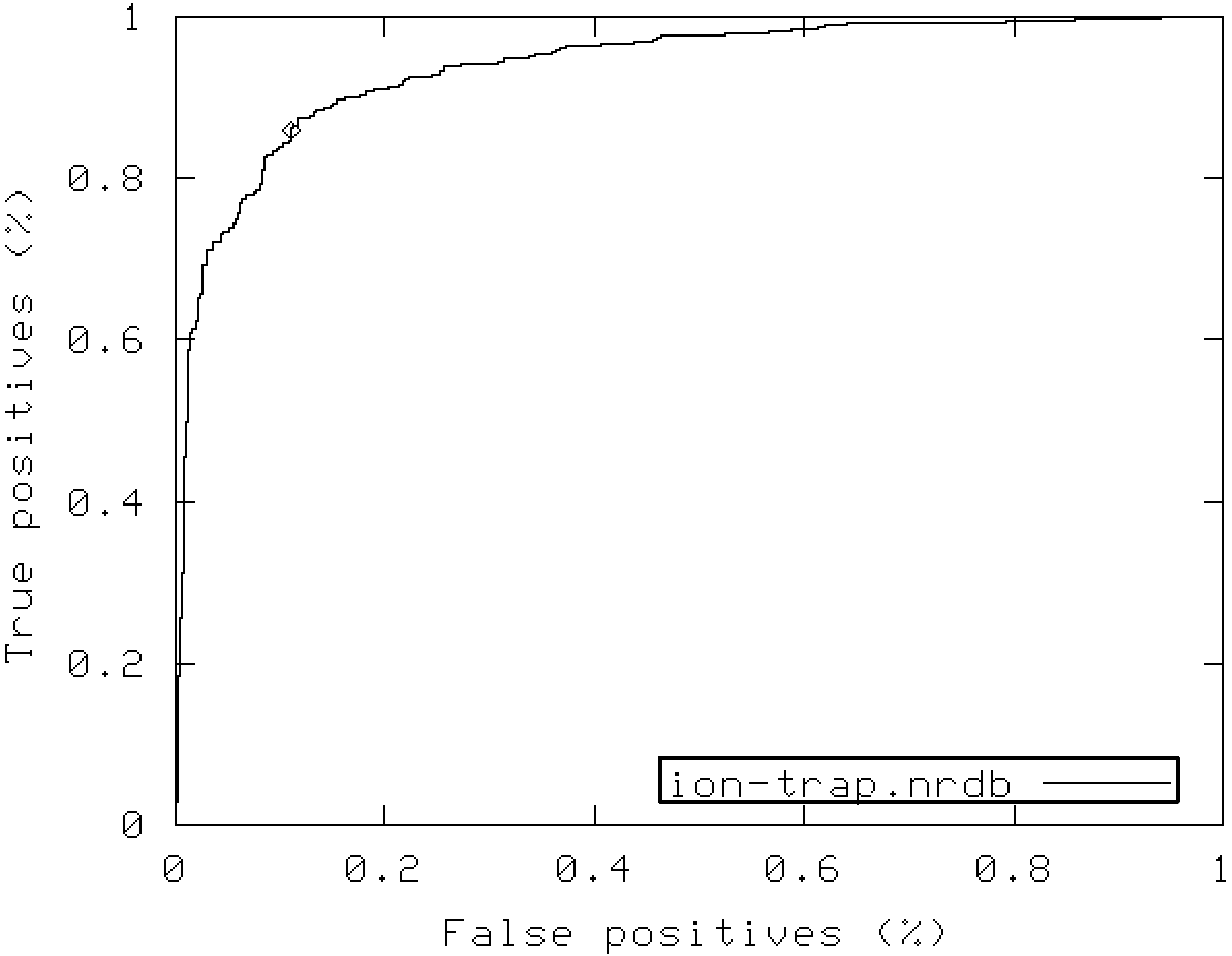
- Set 1: Ion trap mass spectrometer. Sequest search on the full non-redundant database.
- Set 2: Ion trap mass spectrometer. Sequest search on human NRDB.
- Set 3: Quadrupole time-of-flight mass spectrometer. Sequence search on human NRDB.

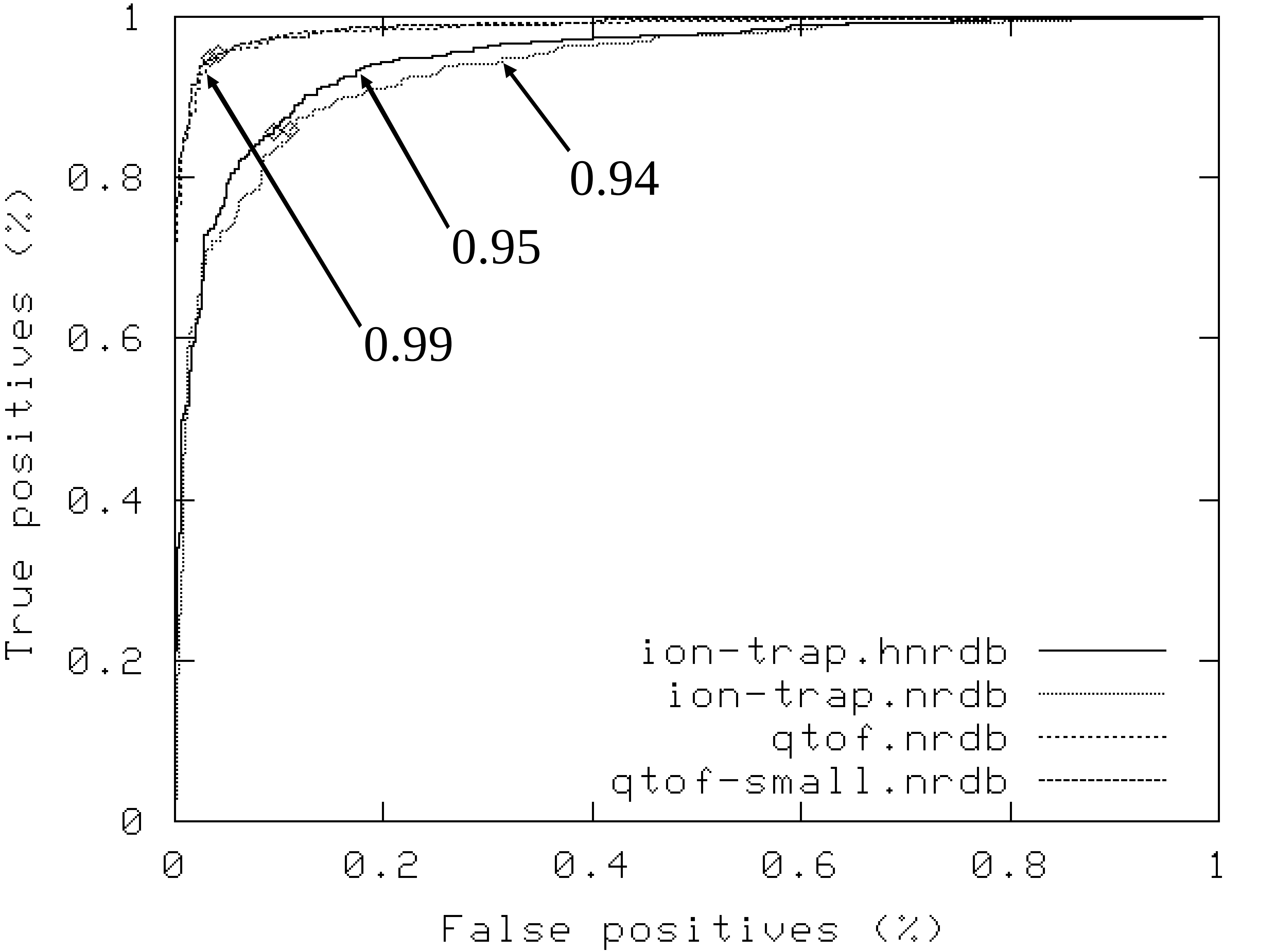
Data set sizes

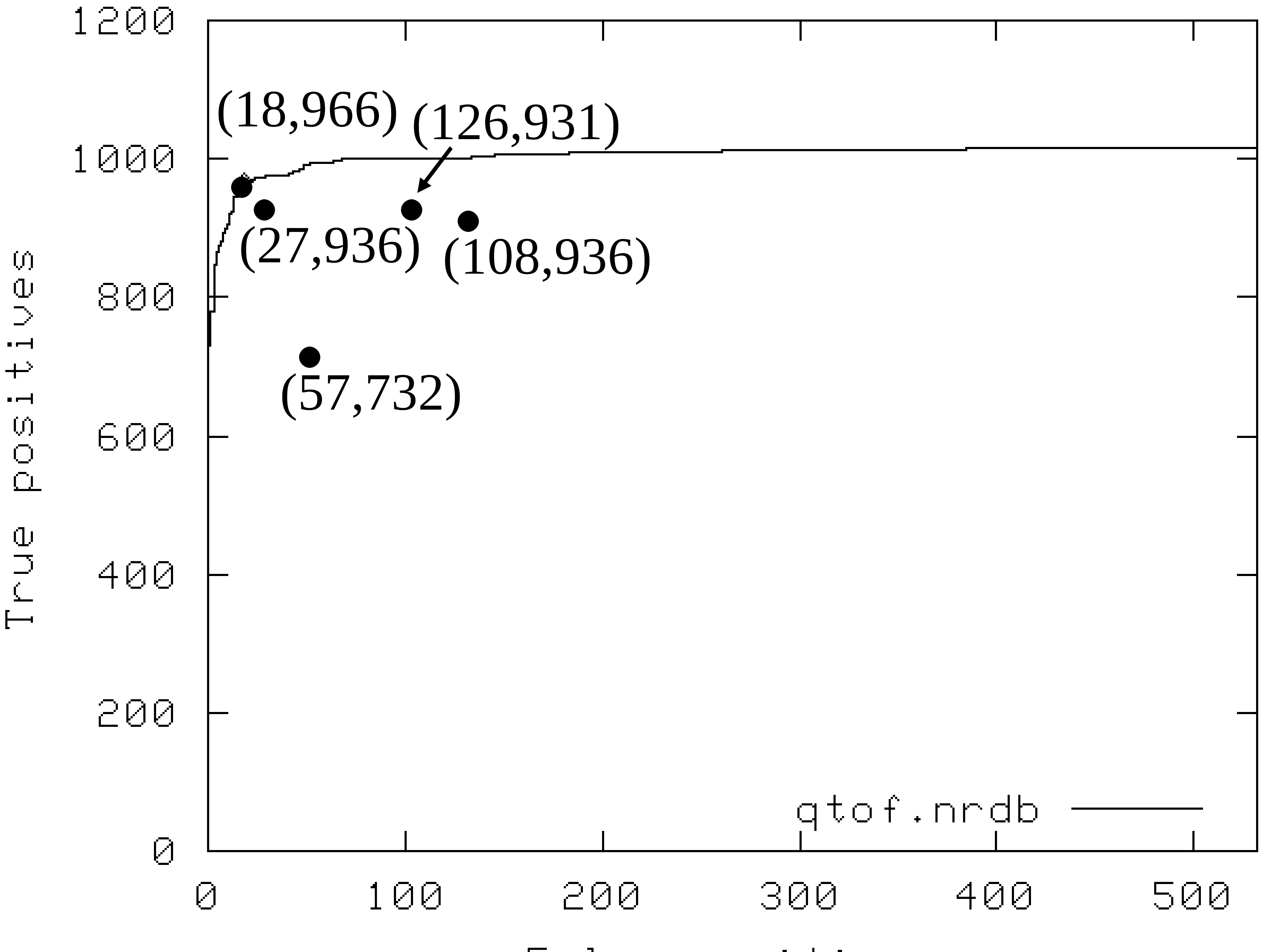
	Positive	Negative	Total
Ion-trap NRDB	497	479	976
Ion-trap HNRDB	696	465	1161
QTOF HNRDB	1017	523	1540

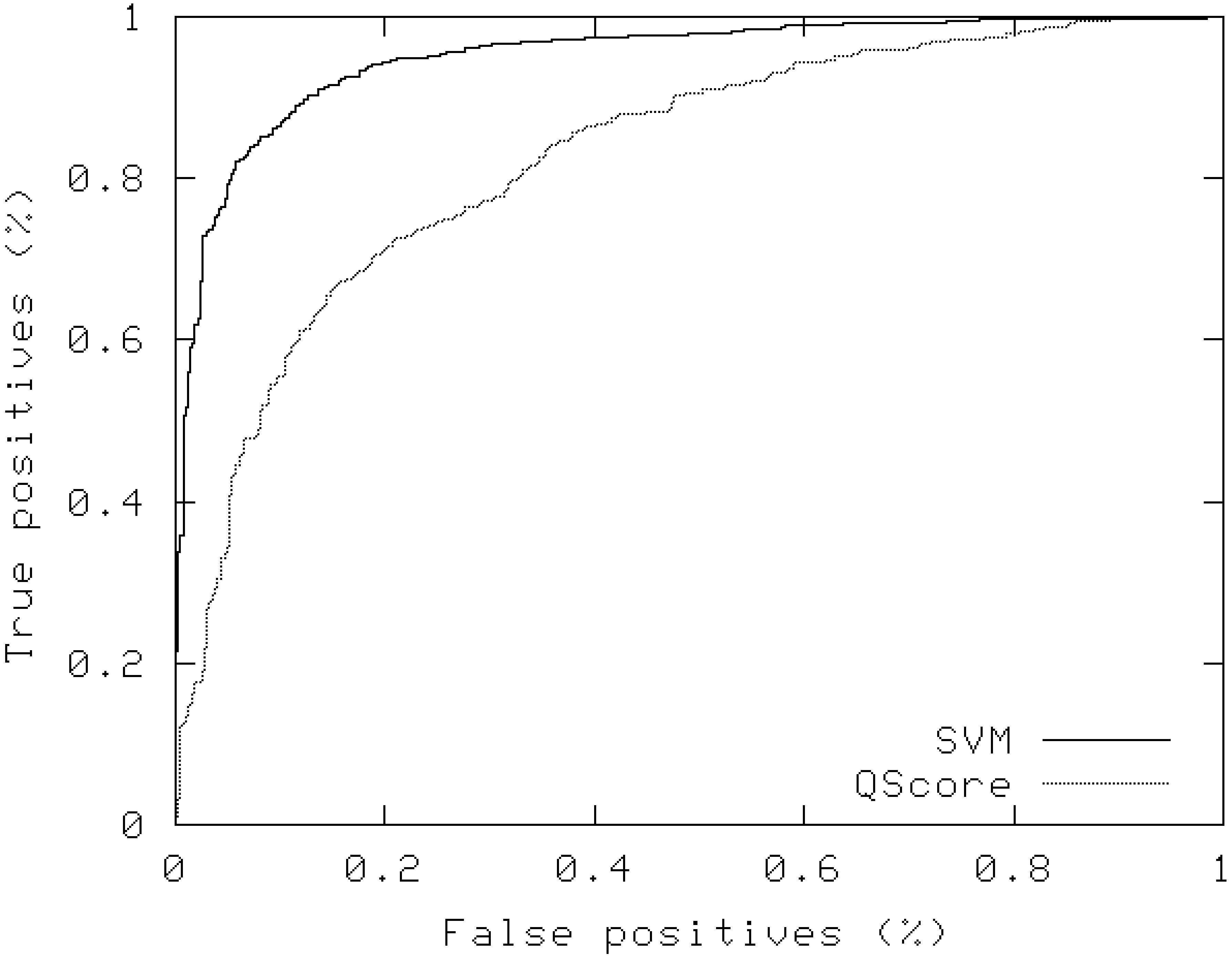


True positives
False positives
0
50
100
150
200
250
300
350
400
450
500
0
50
100
150
200
250
300
350
400
450
ion-trap.nrdb


1
0.8
0.6
0.4
0.2
0
True positives (%)
0
0.2
0.4
0.6
0.8
1
False positives (%)
ion-trap.nrdb


1
0.8
0.6
0.4
0.2
0
True positives (%)
0
0.2
0.4
0.6
0.8
1
False positives (%)
0.94
0.95
0.99
ion-trap.hnrdb
ion-trap.nrdb
qtof.nrdb
qtof-small.nrdb

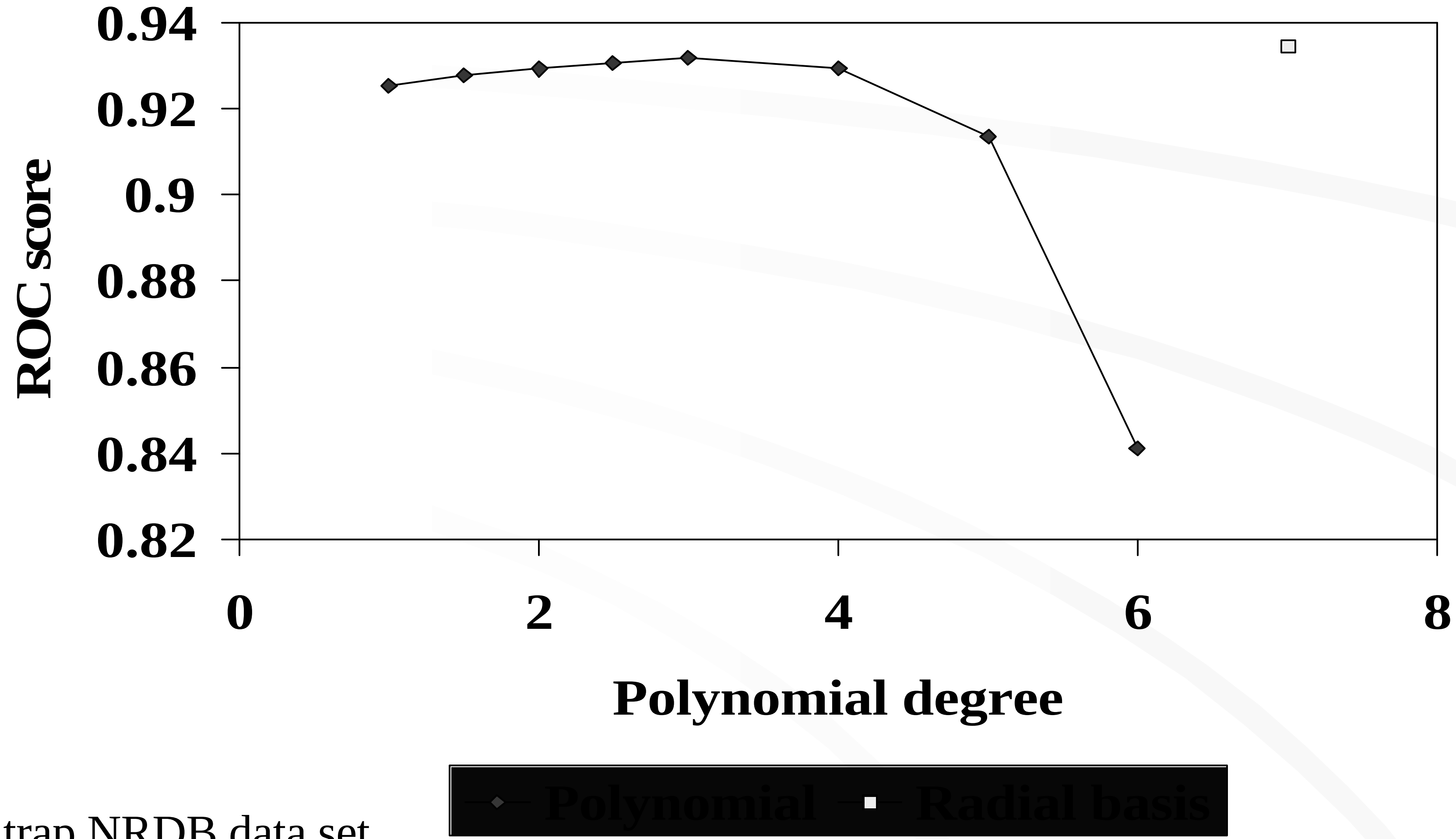

1200
1000
800
600
400
200
0
0
100
200
300
400
500
True positives
False positives
qtof.nrdb
(18,966)
(126,931)
(27,936)
(108,936)
(57,732)


True positives (%)
False positives (%)
0
0.2
0.4
0.6
0.8
1
0
0.2
0.4
0.6
0.8
1
SVM
QScore

Tuning SVM parameters

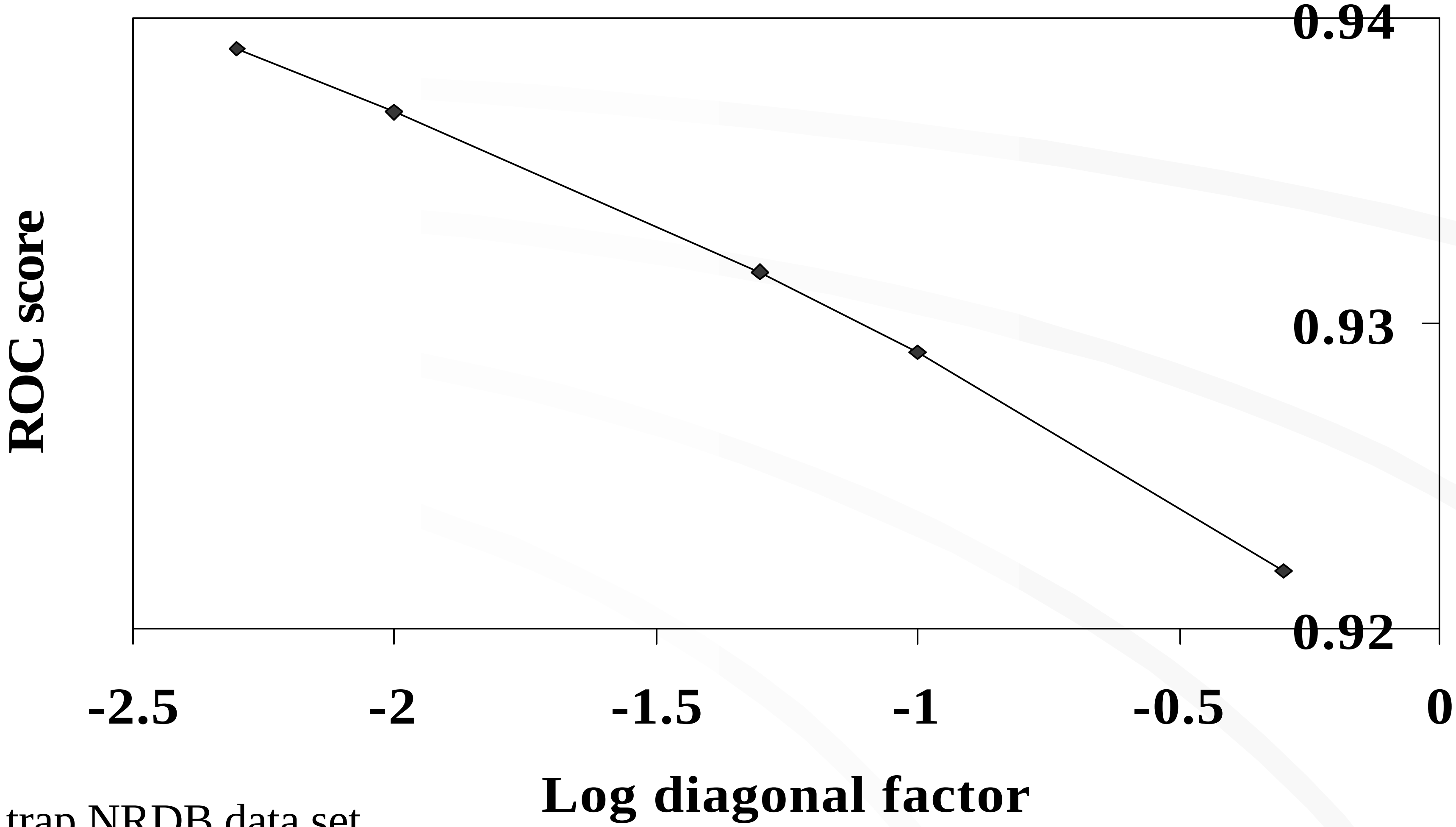
- The polynomial degree controls the dimensionality of the feature space.
- The “diagonal factor” controls the softness of the margin.

Varying the polynomial

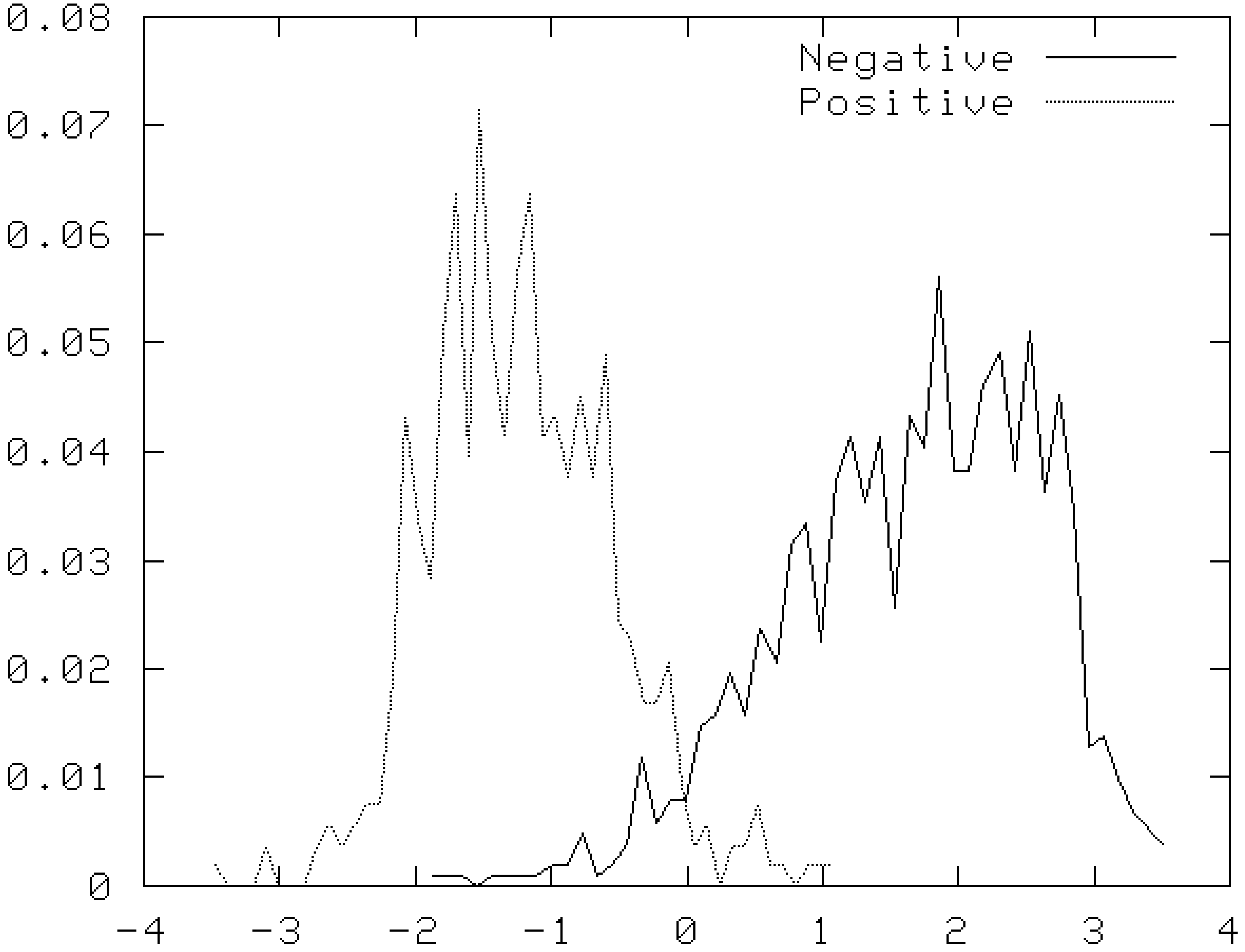


Ion trap NRDB data set.

Adjusting the soft margin



Ion trap NRDB data set.


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
-4
-3
-2
-1
0
1
2
3
4
Negative
Positive

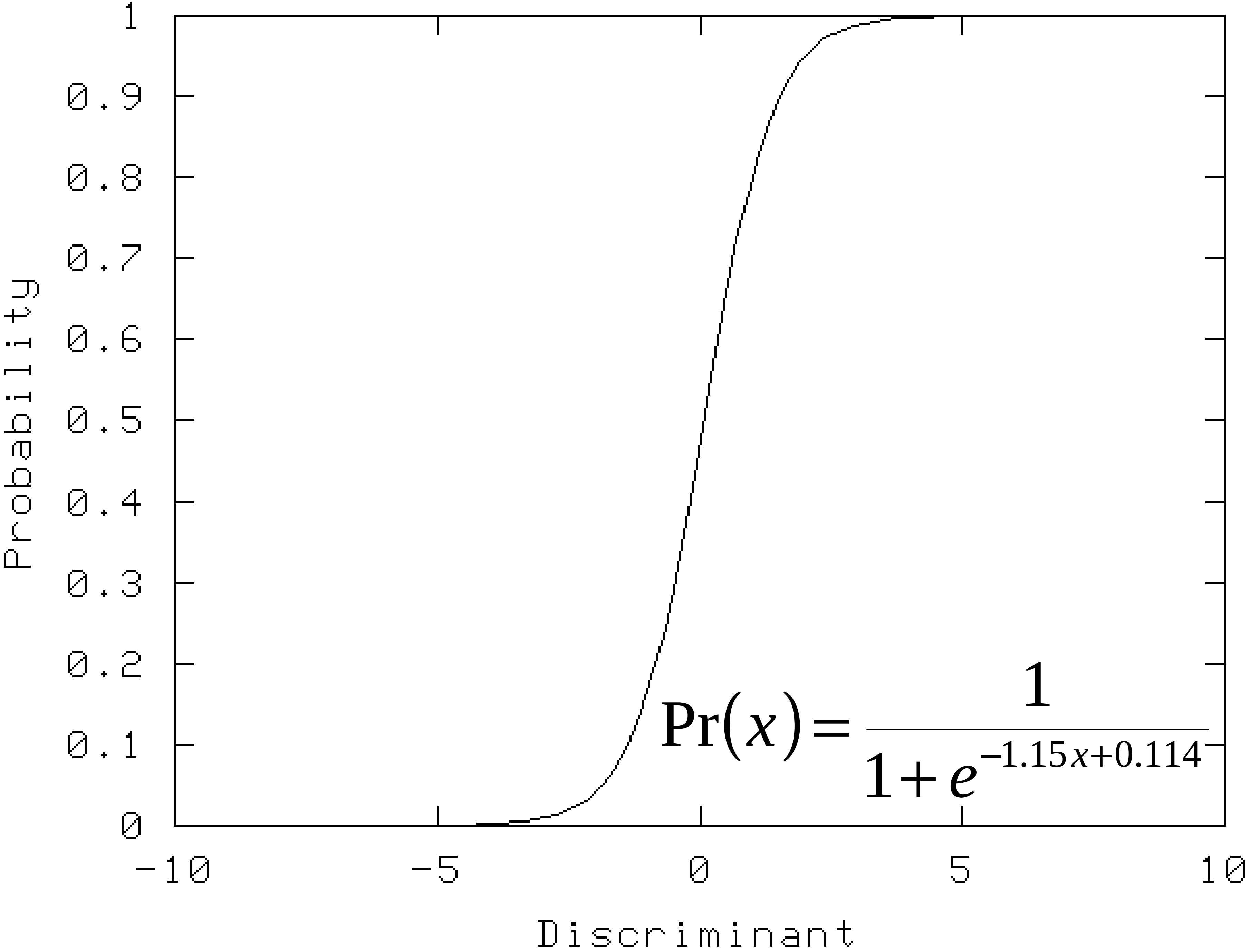
Conversion to probabilities

- Hold out a subset of the training examples.
- Use the hold-out set to fit a sigmoid.

$$\Pr\left(y=1 \middle| f\right)=\frac{1}{1+e^{Af+B}}$$

y = label
f = discriminant

- This is equivalent to assuming that the SVM output is proportional to the log-odds of a positive example.


Probability
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-10
-5
0
5
10
Discriminant
$\Pr(x) = \frac{1}{1+e^{-1.15x+0.114}}$

Analysis of cellular ubiquitinated proteins

- Ubiquitin is a 76 amino acid protein that, when covalently attached to other cell proteins, targets them to the proteasome for degradation.
- Ubiquitin attachment can also be used to regulate cellular processes by mechanisms other than degradation.
- Proteins are labeled with ubiquitin in lysed cells using cellular enzyme systems specialized for attaching ubiquitin to proteins.
- Peptides of proteins affinity extracted with an anti-epitope tag-ubiquitin antibody are analyzed by mass spectrometry.
- Proteins identified by only 1 or 2 peptides are analyzed by the SVM and the appropriate training set to calculate the probability that the peptide sequence match is correct.

Results

Peptide	Prob	Protein	Comment
VTIAQGGVLPNIQAVLLPK	0.971	Histone H2A	known to be ubiquitinated (positive control)
AENYDIPSADR	0.929	E1 ubiquitin activating enzyme	important component of the ubiquitin-proteasome system
NKLDFLRPYTVPNK	0.858	26S proteasome beta 7 subunit	combined with other data, indicates affinity extraction of the proteasome
VLVALYEEPEKPNSALDFLK	0.922	c-myc binding protein	binds c-myc, which when disregulated is oncogenic in a variety of cancers

Future work

- Test the SVM’s generalization to other data sets.
- Develop a more complete feature set.
- Design algorithms for other mass spec instruments.
- Combine peptide-level predictions into protein-level predictions.

- Anderson, DC, W Li, DG Payan and WS Noble. “A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores.” *Journal of Proteome Research.*
- http://www.gs.washington.edu/~noble