CATALOG DESCRIPTION

ECSE 6610 Advanced Character Recognition. 4 credits.
Principles and practice of the recognition of isolated or connected typeset, hand-printed, and cursive characters. Review of optical scanners, features, classifiers. Supervised and non-supervised estimation of classifier parameters. Expectation Maximization, the Curse of Dimensionality, language context. Advanced classification techniques including Classifier Combinations, Support Vector Machines, Hidden Markov Methods, Adaptation, Indirect Symbolic Correlation.
Prerequisites: ECSE 2610, Probability, Linear Algebra.
Spring term annually

ECSE-6610 FIRST DAY HANDOUT

Instructor:	Prof. George Nagy,
Office:	JEC 6020, RPI, Troy, NY
Office hours:	Hotel Bar, after class
Email:	nagy@ecse.rpi.edu

Text: ***Optical Character Recognition: An Illustrated Guide to the Frontier***
S. V. Rice, G. Nagy, T. A. Nartker, Kluwer Academic Publishers, 1999

Reference texts (on reserve at Folsom Library):

Duda, Hart, & Stork, 2001	[DHS 01]
Mitchell, McGraw-Hill 1997	[TM 97]
Nadler & Smith, Wiley 1993	[NS 93]
Schürmann, Wiley, 1996	[JS 96]
Theodoridis & Koutroumbas 1999	[TK 99]
Vapnik, Wiley 1998	[VV 98]

For additional sources, see the Text and the Bibliography.

Grading: Five programming assignments
Term Paper
Final Examination

Review: Intro to OCR (ECSE 2610)

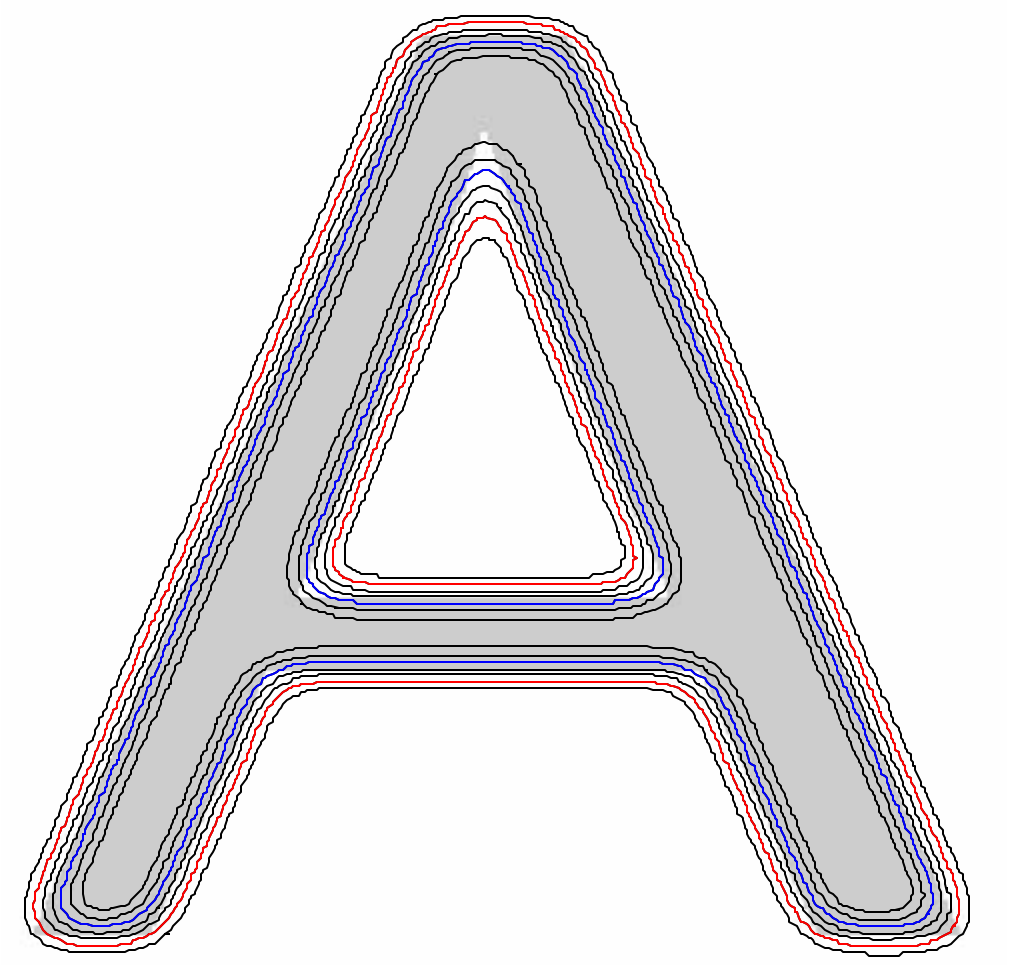
Digitization and preprocessing:

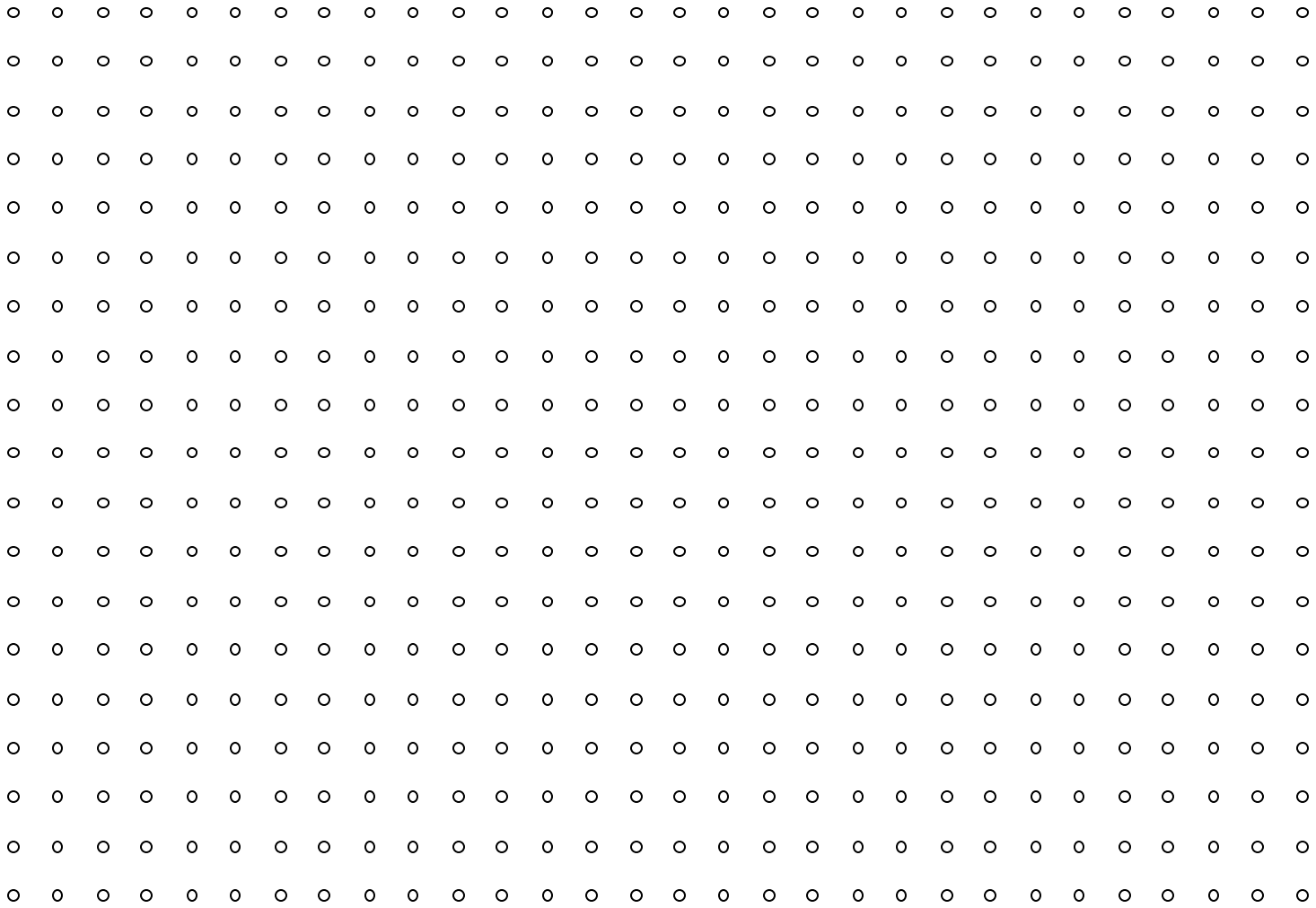
Scanner calibration; noise removal; recovery of scanner distortions [BE 01]

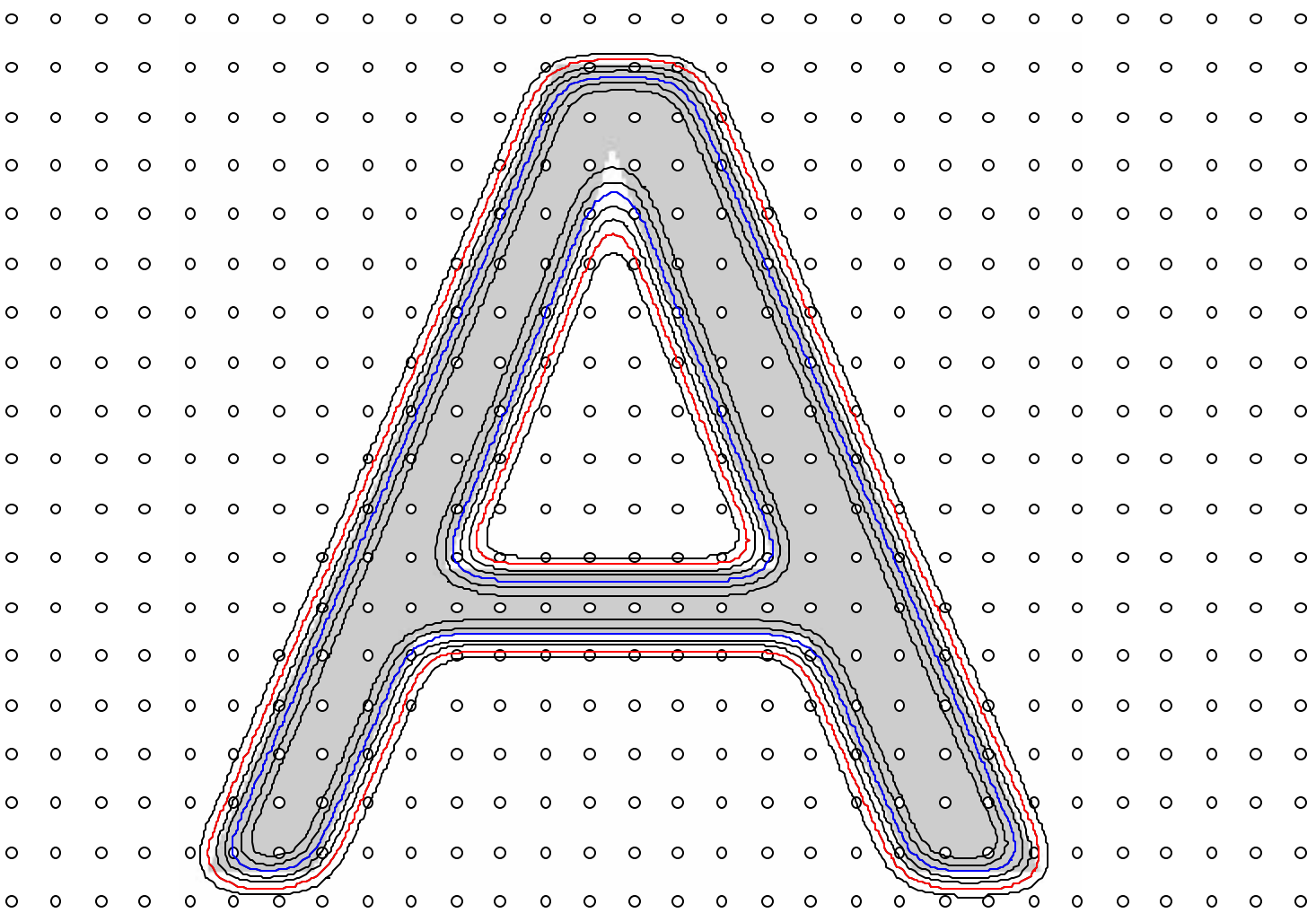
Character image defect models [KBH 94]

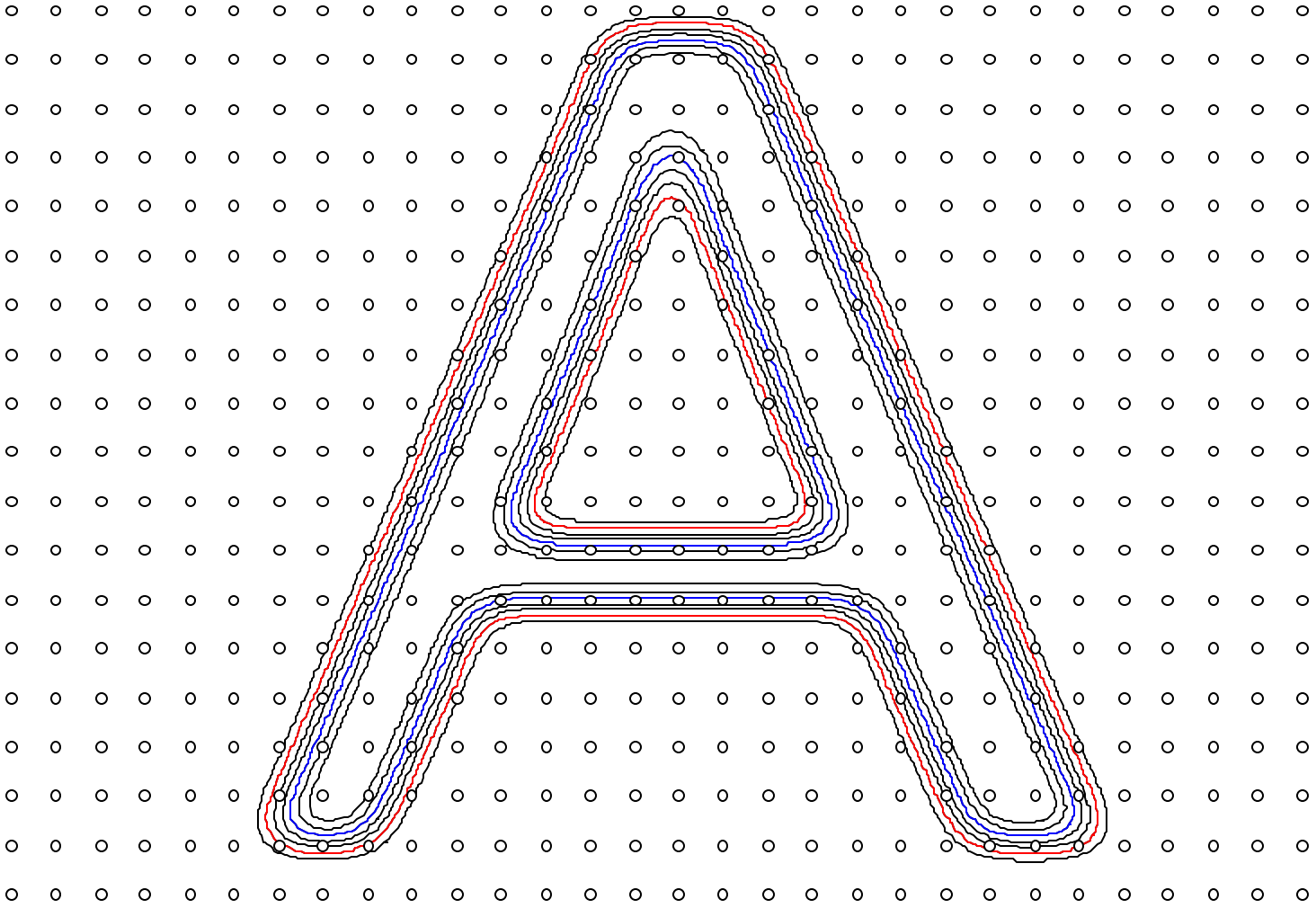
Help Session: Thursday 9:40 Prof. Elisa Barney Smith, BSU

MODEL OF DIGITIZATION









Review:

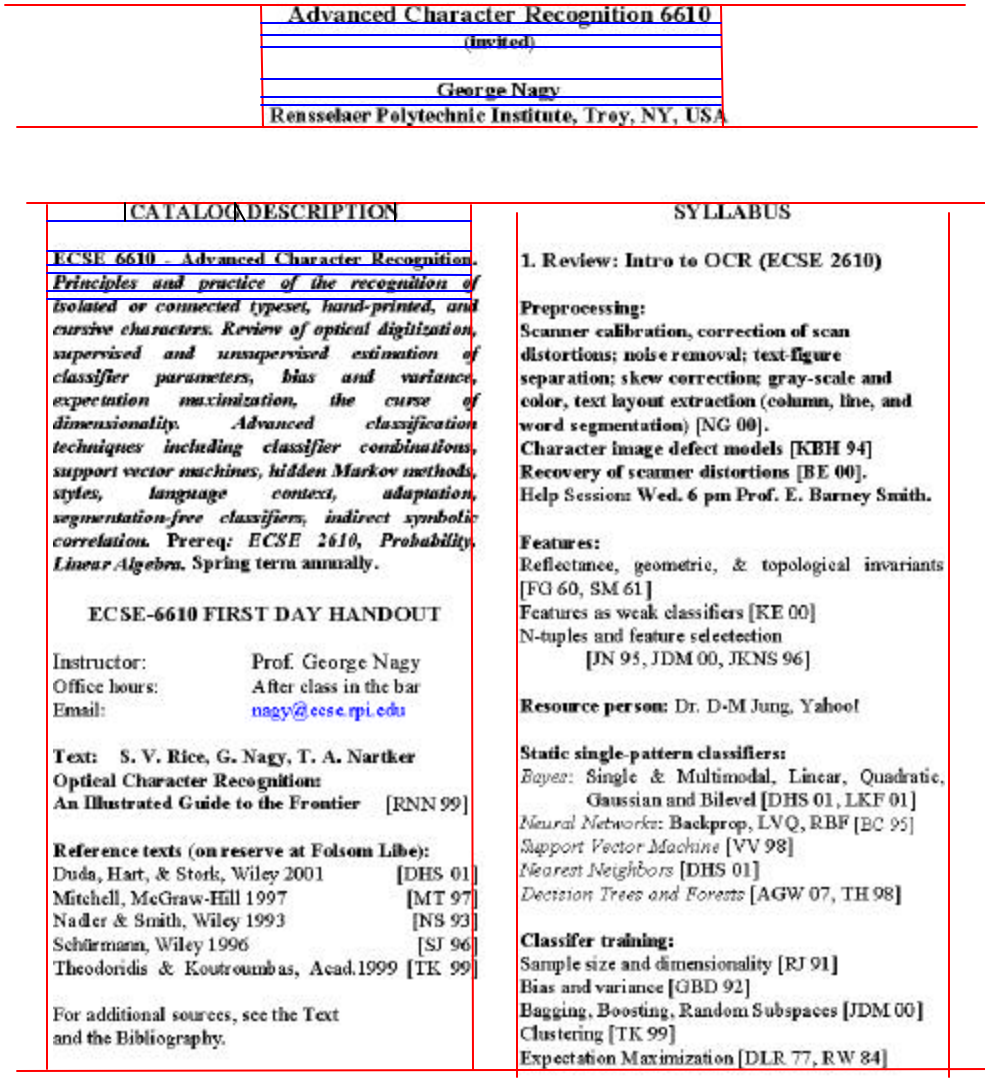
Text-figure separation; skew correction; text layout extraction (column, line, and word segmentation) [NG 00].

Help session: Wed. 11:00 Prof. Don Sylwester, Concordia College

Tables

Help session: Monday 13:15, 13:55 Dr. Dan Lopresti, Lucent Bell Labs

X-Y TREE



Advanced Character Recognition 6610

(Invited)

George Nagy

Rensselaer Polytechnic Institute, Troy, NY, USA

CATALOG DESCRIPTION

ECSE 6610 - Advanced Character Recognition. ***Principles and practice of the recognition of isolated or connected typeset, hand-printed, and cursive characters. Review of optical digitization, supervised and unsupervised estimation of classifier parameters, bias and variance, expectation maximization, the curse of dimensionality. Advanced classification techniques including classifier combinations, support vector machines, hidden Markov methods, styles, language context, adaptation, segmentation-free classifiers, indirect symbolic correlation.*** **Prereq:** ***ECSE 2610, Probability, Linear Algebra.*** **Spring term annually.**

ECSE-6610 FIRST DAY HANDOUT

Instructor:	Prof. George Nagy
Office hours:	After class in the bar
Email:	nagy@ecse.rpi.edu

Text: S. V. Rice, G. Nagy, T. A. Nartker
Optical Character Recognition:
An Illustrated Guide to the Frontier [RNN 99]

Reference texts (on reserve at Folsom Libe):

Duda, Hart, & Stork, Wiley 2001	[DHS 01]
Mitchell, McGraw-Hill 1997	[MT 97]
Nadler & Smith, Wiley 1993	[NS 93]
Schürmann, Wiley 1996	[SJ 96]
Theodoridis & Koutroumbas, Acad.1999	[TK 99]

For additional sources, see the Text and the Bibliography.

SYLLABUS

1. Review: Intro to OCR (ECSE 2610)

Preprocessing:
Scanner calibration, correction of scan distortions; noise removal; text-figure separation; skew correction; gray-scale and color, text layout extraction (column, line, and word segmentation) [NG 00].
Character image defect models [KBH 94]
Recovery of scanner distortions [BE 00].
Help Session: **Wed. 6 pm Prof. E. Barney Smith.**

Features:
Reflectance, geometric, & topological invariants [FG 60, SM 61]
Features as weak classifiers [KE 00]
N-tuples and feature selection [JN 95, JDM 00, JKNS 96]

Resource person: Dr. D-M Jung, Yahoo!

Static single-pattern classifiers:
Bayes: Single & Multimodal, Linear, Quadratic, Gaussian and Bilevel [DHS 01, LKF 01]
Neural Networks: **Backprop**, LVQ, RBF [BC 95]
Support Vector Machine [VV 98]
Nearest Neighbors [DHS 01]
Decision Trees and Forests [AGW 07, TH 98]

Classifer training:
Sample size and dimensionality [RJ 91]
Bias and variance [GBD 92]
Bagging, Boosting, Random Subspaces [JDM 00]
Clustering [TK 99]
Expectation Maximization [DLR 77, RW 84]

Review:

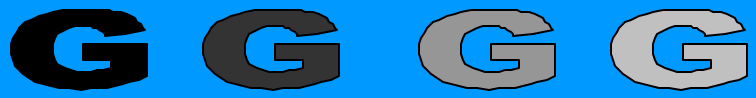
Features:

Reflectance, geometric, & topological *invariants*	[GF 60], [SM 61]
Features as weak classifiers	[KE 00]
N-tuples	[JN 95]
Feature selection	[JDM 00]

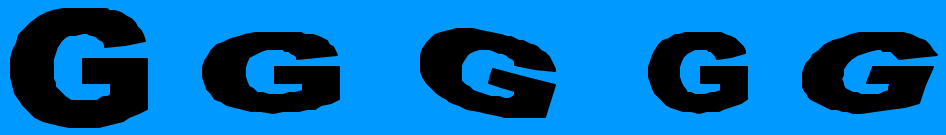
Resource person: Dr. D-M Jung, Yahoo!

FEATURE INVARIANCES

Reflectance:



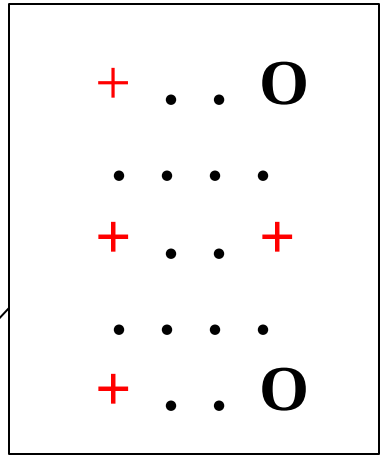
Geometry:



Topology:

N-TUPLE FEATURES (OR CLASSIFIERS?)

Bledsoe & Browning, 1959, ….., D-M Jung, 1995, ….



Review

Static Singleton Classifiers:

Bayes:
Single & Multimodal, Linear, Quadratic,
Gaussian and Bilevel [DHS 01]

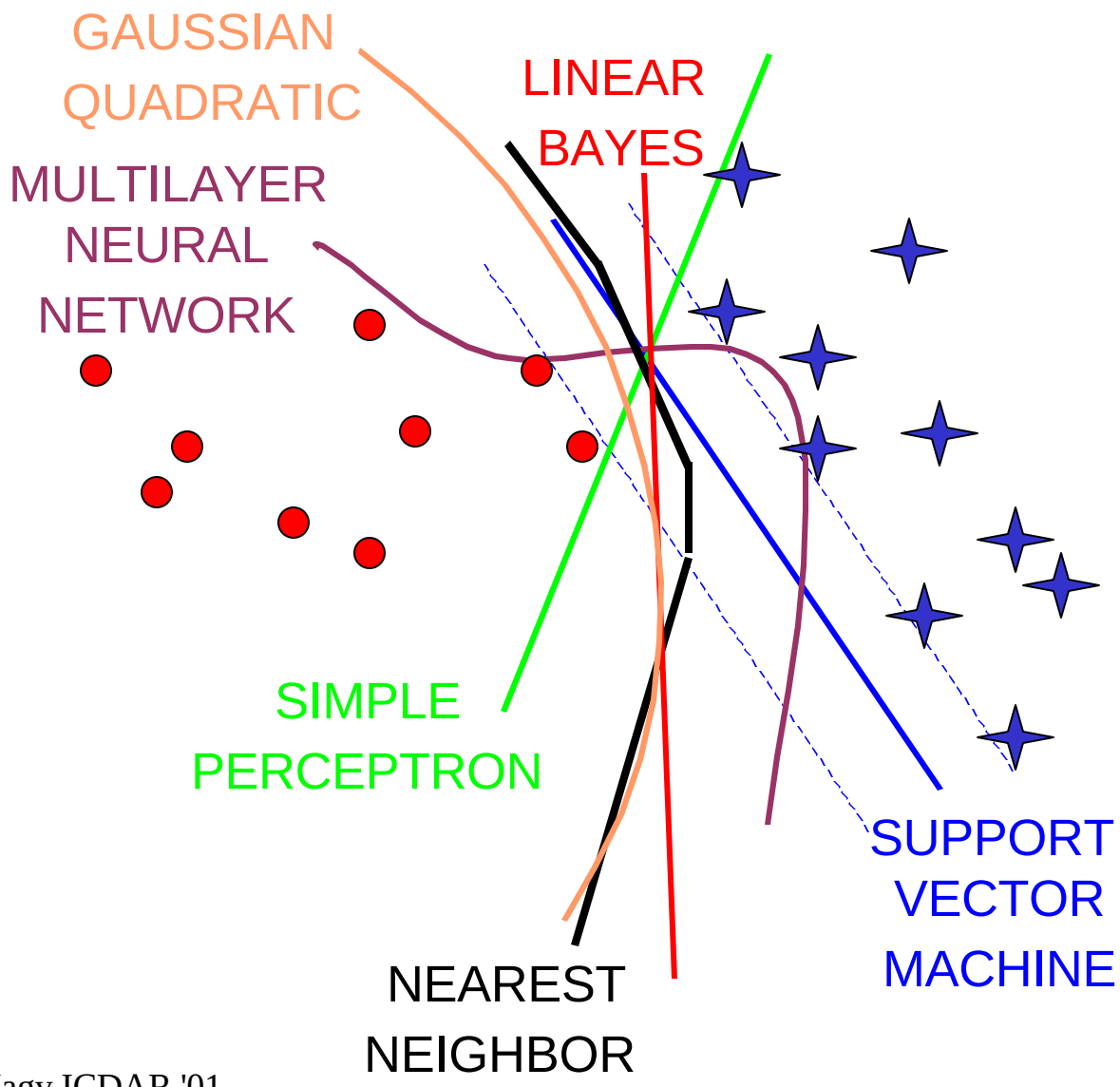
Neural Networks:, [BC 95]
Backpropagation, Learning Vector Quantization,
Radial Basis Functions

Support Vector Machines [VV 98]

Nearest Neighbors [DHS 01]

Decision Trees and Decision Forests [TH 98]

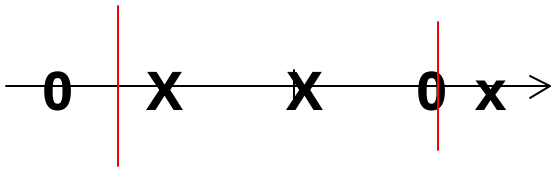
SOME CLASSIFIERS

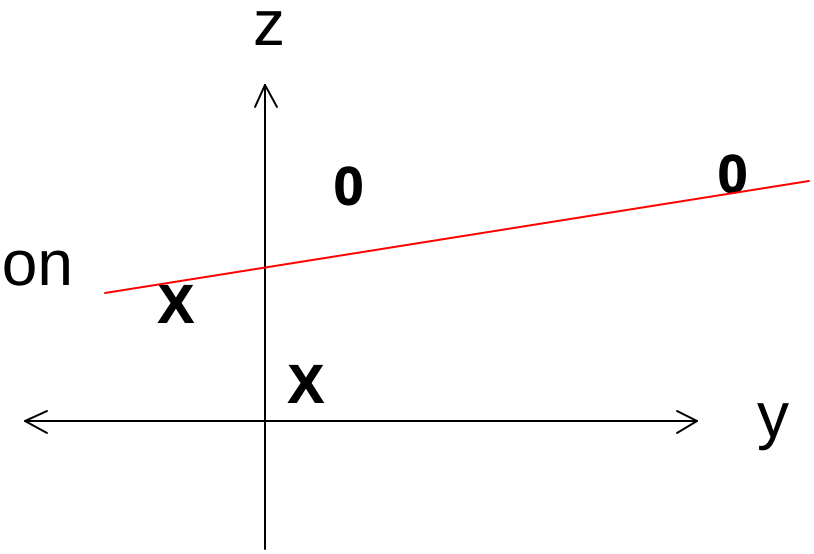


SUPPORT VECTOR MACHINE (V. Vapnik)

Kernel-induced transformation

$x \rightarrow \mathbf{v} = (y,z)$

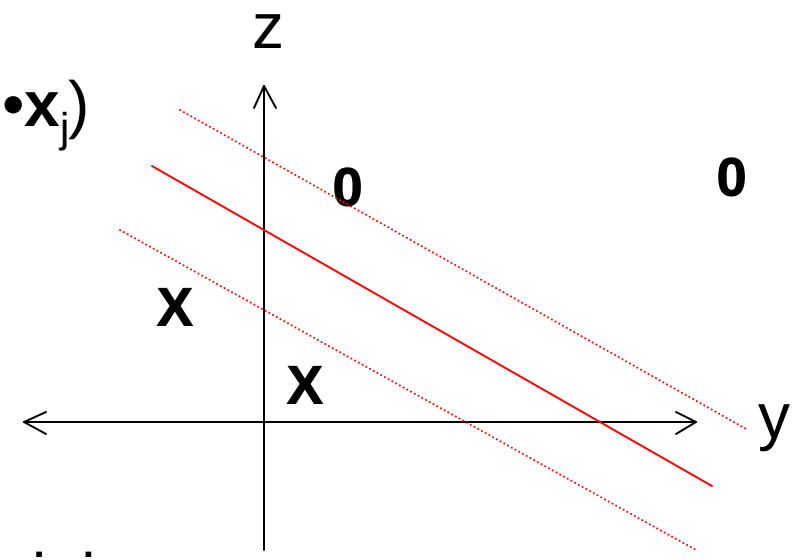




max min { $f(\mathbf{v}_i \bullet \mathbf{v}_j)$ } by QP

Mercer's theorem: $\mathbf{v}_i \bullet \mathbf{v}_j = K(\mathbf{x}_i \bullet \mathbf{x}_j)$

i.e., compute distances in high-dim space from distances in low-dim space



resists over-training

Review:

Static Singleton Classifiers:

Bayes: Single & Multimodal, Linear, Quadratic, Gaussian and Bilevel	[DHS 01]
Neural Networks: Backprop, LVQ, RBF,	[BC 95]
Support Vector Machine	[VV 98]
Nearest Neighbors	[DHS 01]
Decision Trees and Decision Forests	[TH 98]

Classifier training:

Dimensionality and sample size	[RJ 91, LSF 01]]
Bias and variance	[GBD 92],
Bagging, Boosting, Random Subspaces	[JDM 00]
Clustering & Expectation Maximization	[TK 99, DLR 77, RW 84]

Generalization, Validation, and Error Prediction:

Validation, Jackknife, Bootstrap	[DHS 01], [JDM 00]

SOME DEFINITIONS

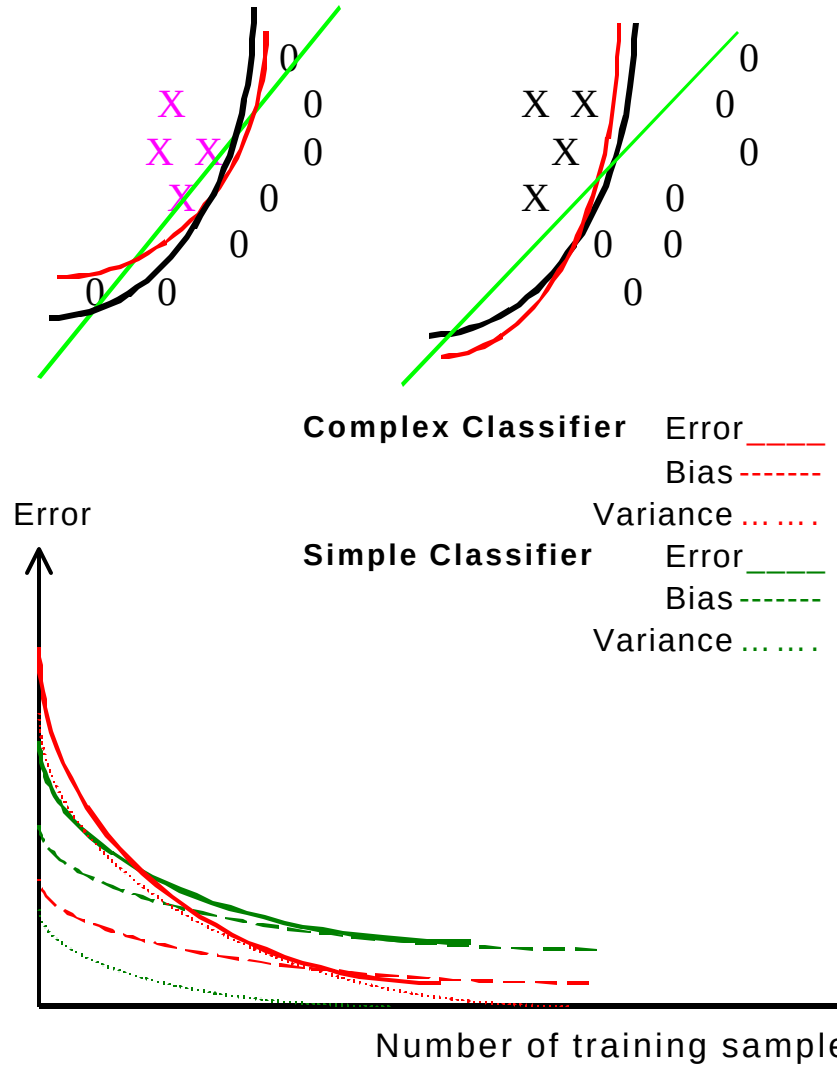
BOOTSTRAP PARAMETER ESTIMATE:
Mean of estimates of *m* sample sets obtained from the same data by *sampling with replacement*.
(Better than m-way partitioning for estimating variances.)

JACKKNIFE: leave-one-out estimate (of classifier accuracy).

BAGGING (Bootstrap AGgregation):
Increasing the nominal size of the training set by sampling with replacement.

BOOSTING: reintroducing misclassified training samples into the classifier construction process.

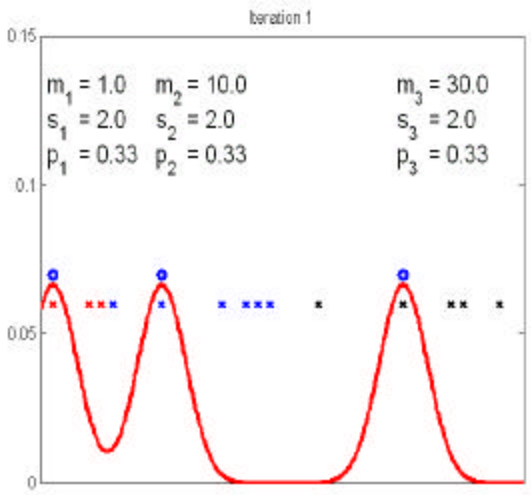
CLASSIFIER BIAS AND VARIANCE

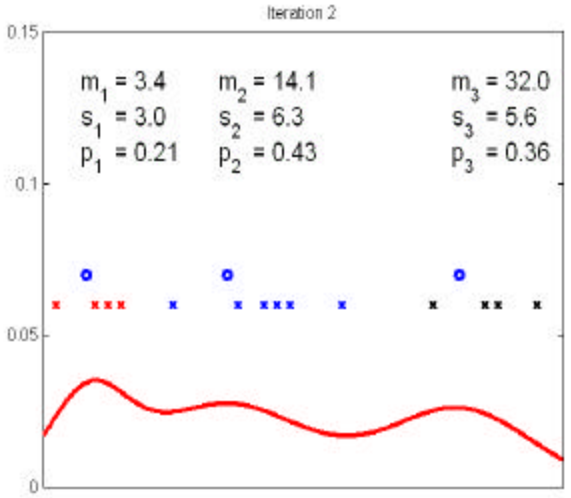


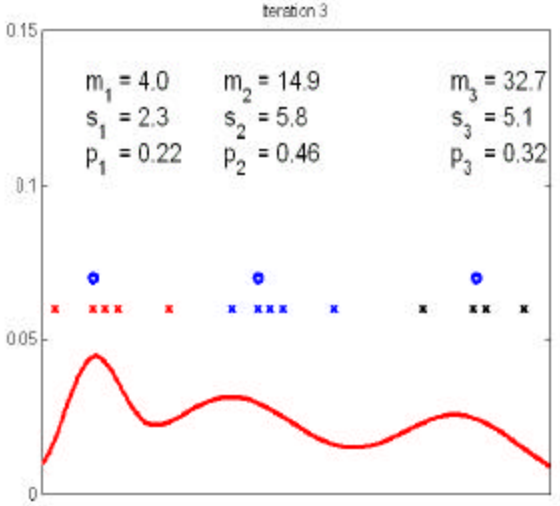
CLASSIFIER BIAS AND VARIANCE DON'T ADD!

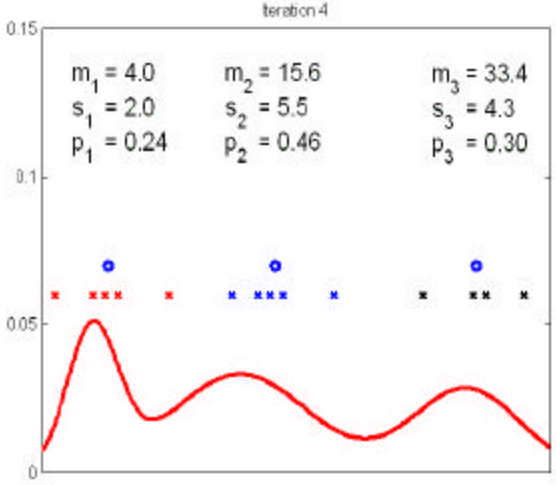
Any classifier can be shown to be better than any other.

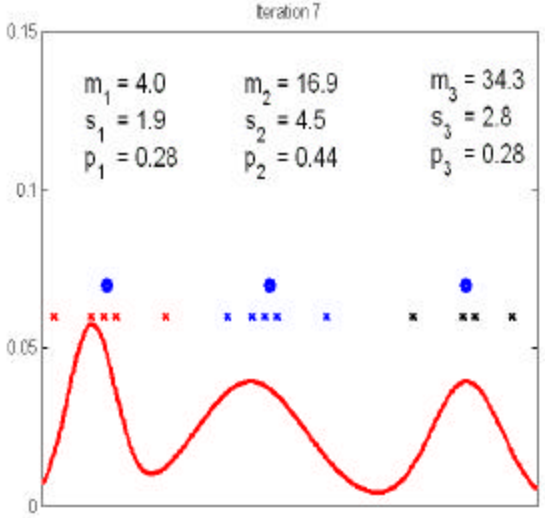
K-MEANS AND EXPECTATION MAXIMIZATION

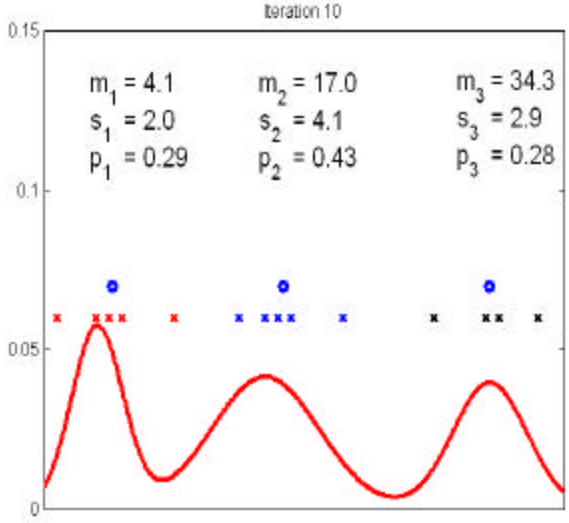












OCR 6610 TOPICS

CLASSIFIER COMBINATION (Dr. Tin Kam Ho, Bell Labs)

CONTEXT (Drs. H. Fujisawa, J.J. Hull, Profs. S. Srihari, S. Seth)

STYLES (Dr. P. Sarkar, Harsha Veeramachaneni)

UNSUPERVISED ADAPTATION (Dr. H.S. Baird)

UNSEGMENTED TEXT

- DOCUMENT-SPECIFIC CLASSIFIERS (Dr. Yihong Xu, EMC)
- N-GRAM-BASED CLASSIFICATION (Adnan El-Nasan)
- INDIRECT SYMBOLIC CORRELATION (Harsha V.)

CLASSIFIER COMBINATION [Tin Kam Ho '01]

DECISION OPTIMIZATION:
Combine a set of complete, fully trained classifiers.

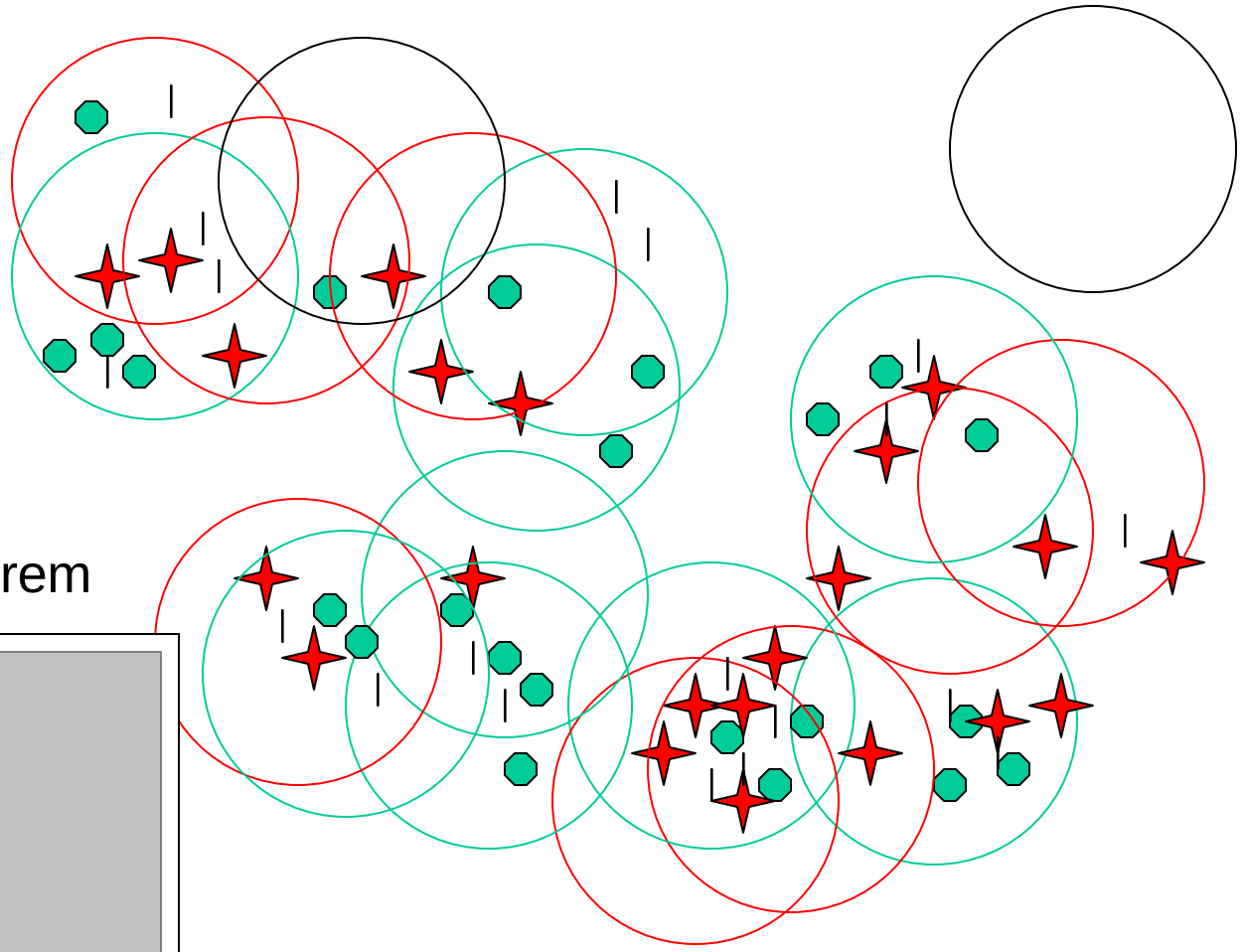
COVERAGE OPTIMIZATION:
Tune each classifier to a different aspect of training set
(Different subspaces or training samples).

STOCHASTIC DISCRIMINATION [KE 00]:

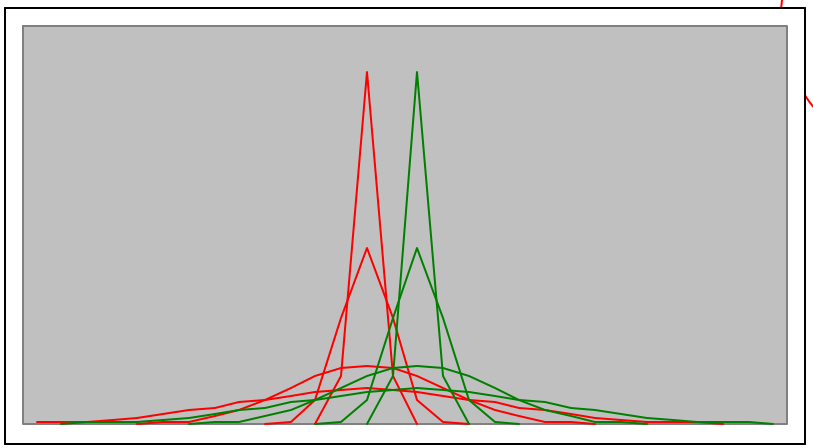
1. Enrichment: each of *many* weak "models" must favor some class;
2. Uniform coverage of populated region;
3. Characteristics of the sample space of weak models with respect to any point same whether that point is a training point or a test point.

STOCHASTIC DISCRIMINATION [Eugene Kleinberg ‘00]

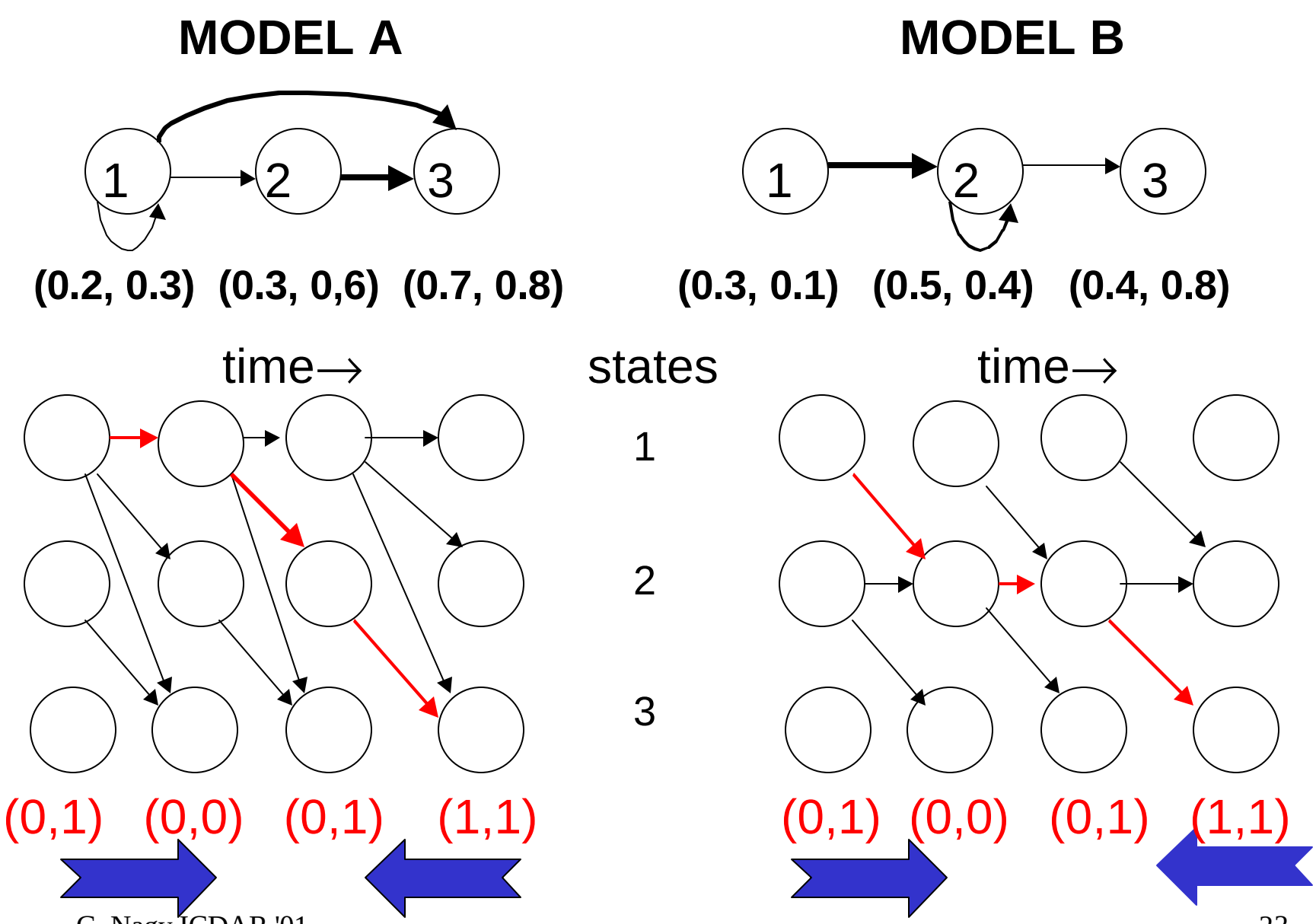
Duality between sample space of weak models and feature space



The Central Limit Theorem

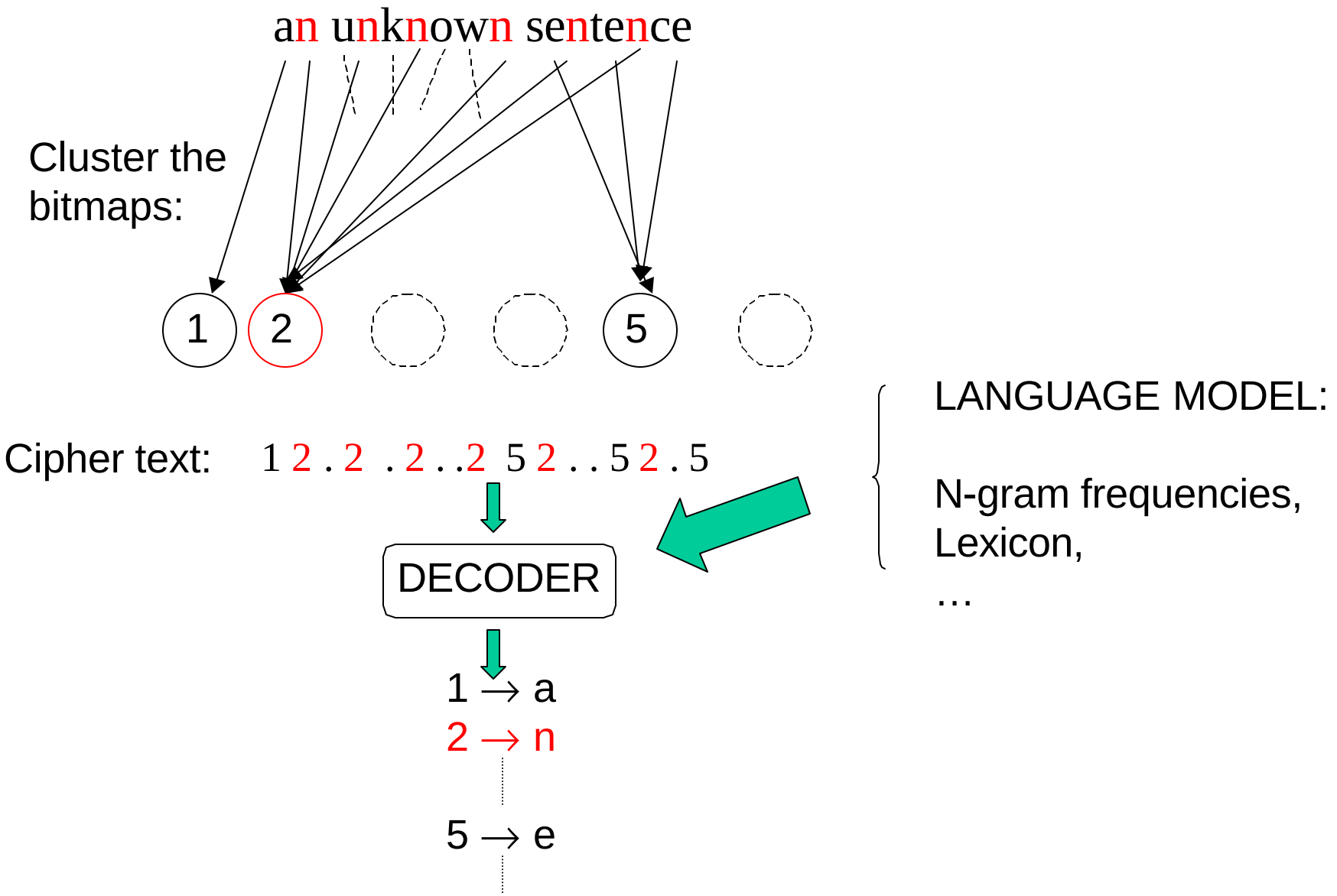


CONTEXT: HIDDEN MARKOV MODELS



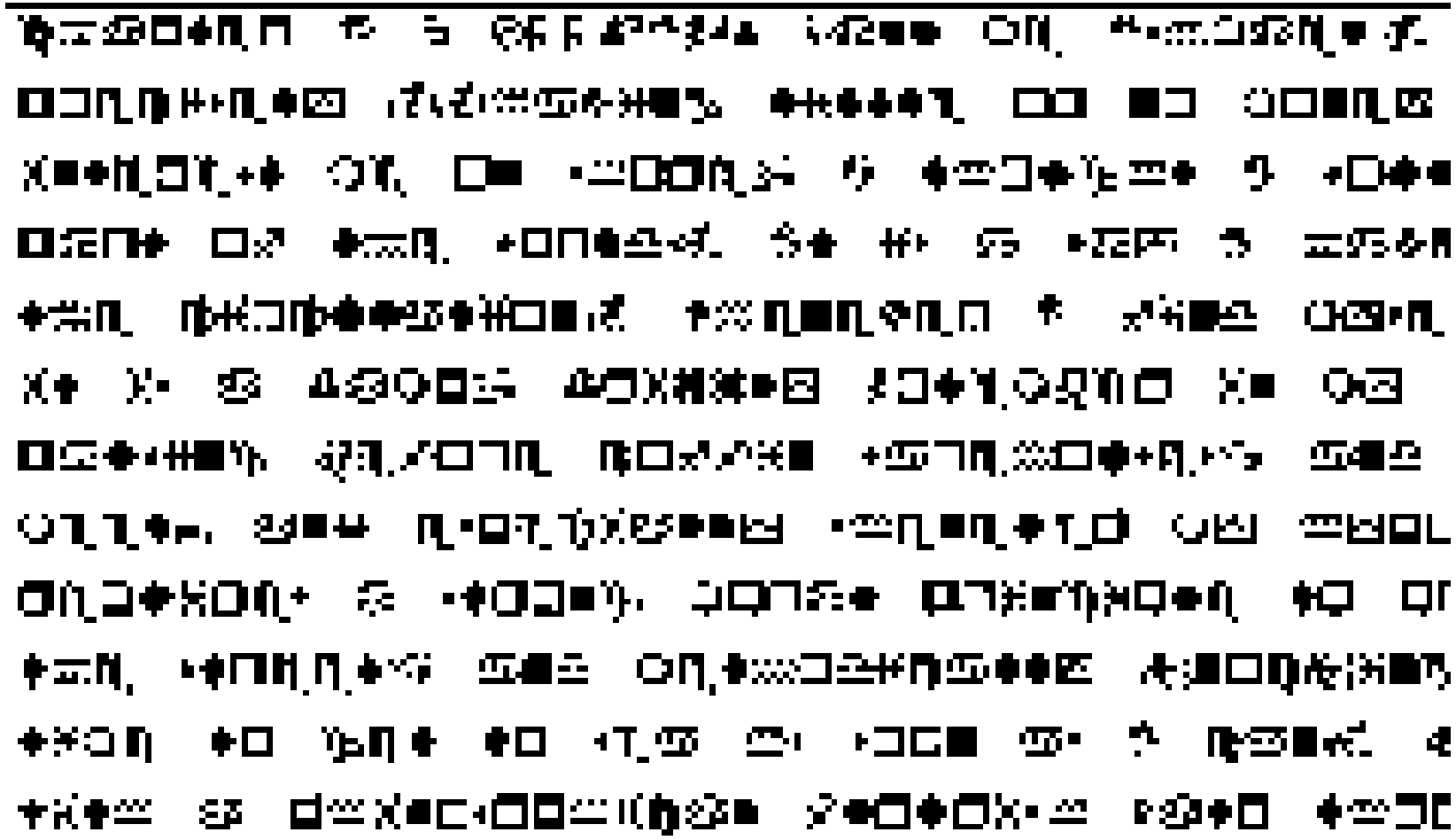
TRAINING: joint probs via Baum-Welch Forward-Backward (EM)

CONTEXT BY DECODING A SUBSTITUTION CIPHER



LINGUISTIC CONTEXT [HN 00]

TEXT PRINTED WITH SPITZ GLYPHS



DECODED TEXT

chapter i _ bee_inds

_all me ishmaels some years ago__never mind how long precisely __having little or no money in my purses and nothing particular to interest me on shores i thought i would sail about a little and see the watery part of the worlds it is a way i have ...

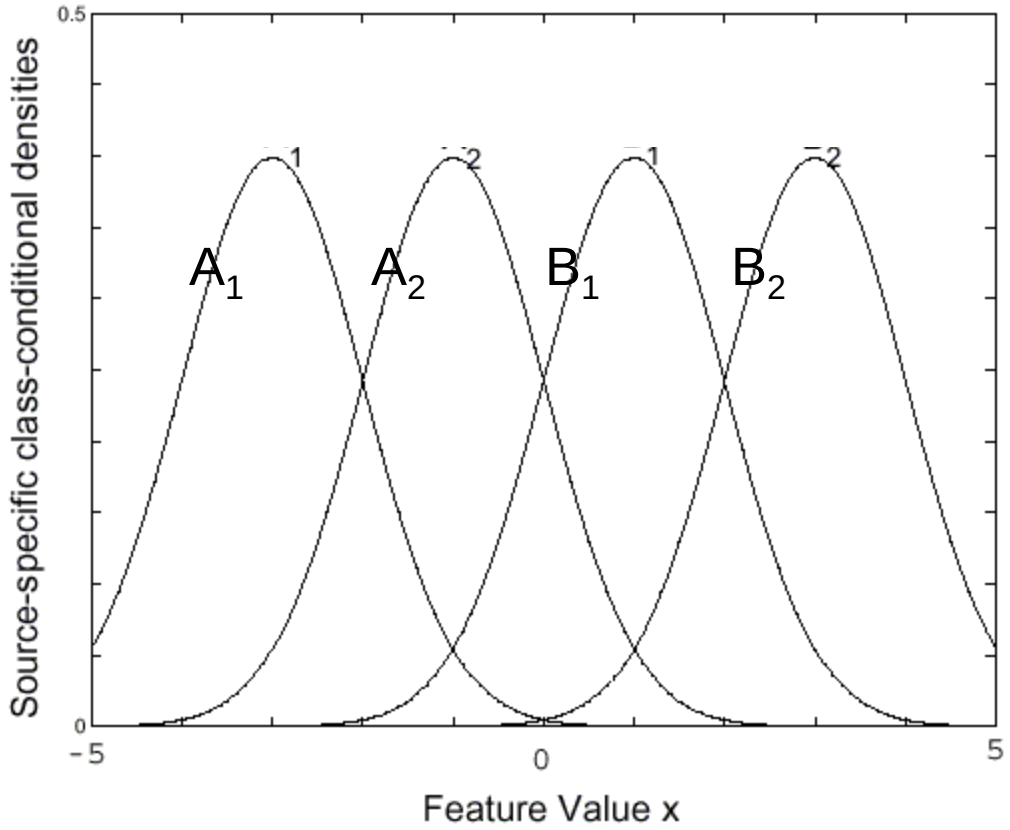
chapter I 2 LOOMINGS

Call me Ishmael. Some years ago – never mind how long precisely – having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have ...

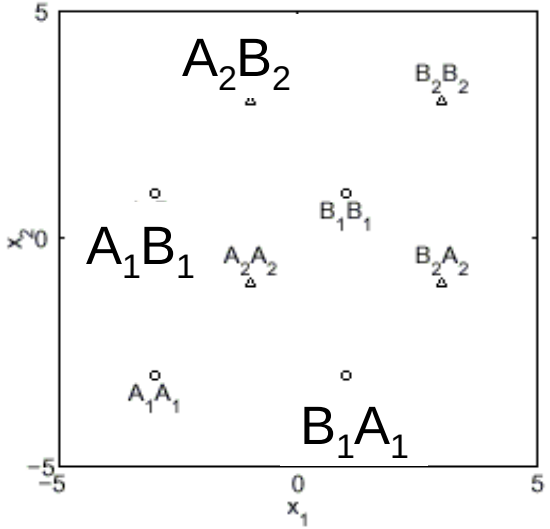
STYLES

(Prateek Sarkar
Thursday 10:00,
Harsha Veeramachaneni)

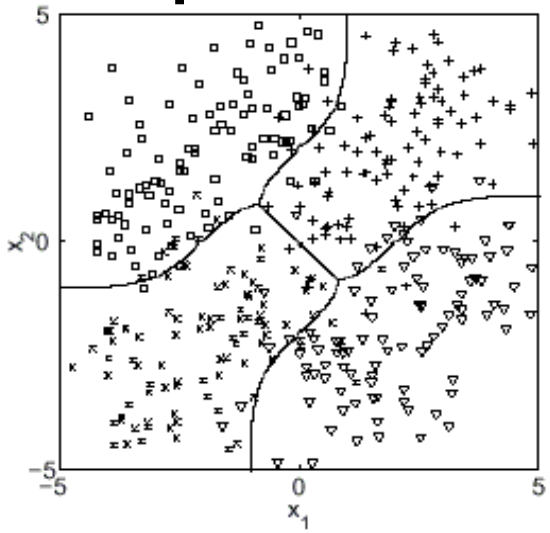
Writer 1 Writer 2



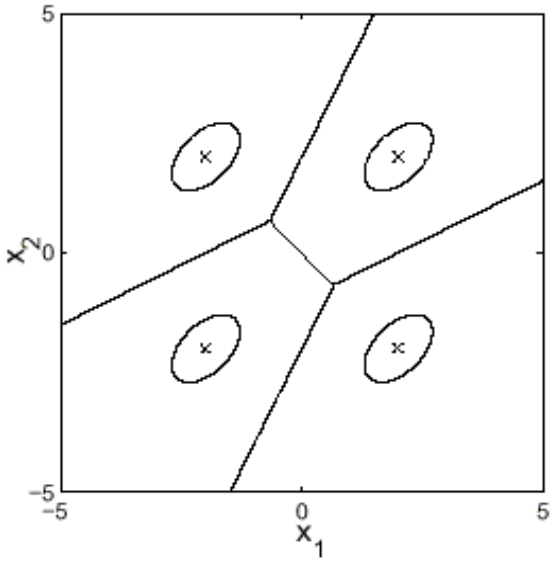
STYLE-CONSCIOUS CLASSIFIERS in field-feature space



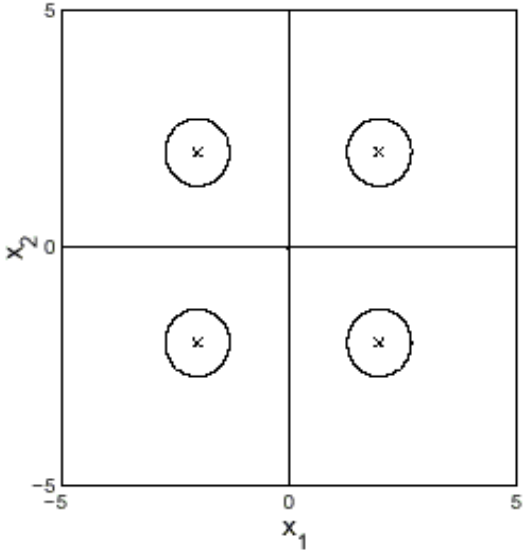
Position of means



Optimal field classifier



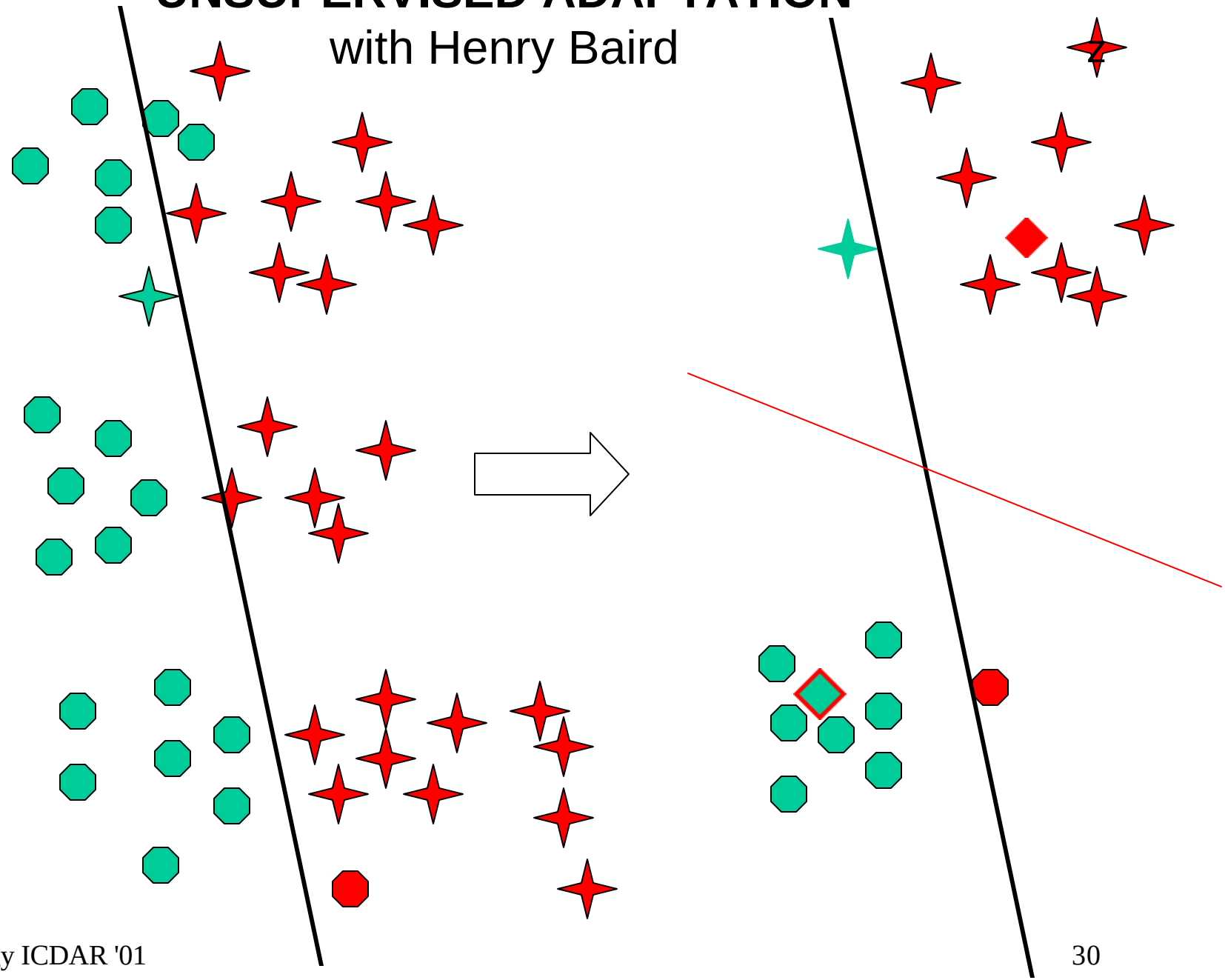
Quadratic field classifier



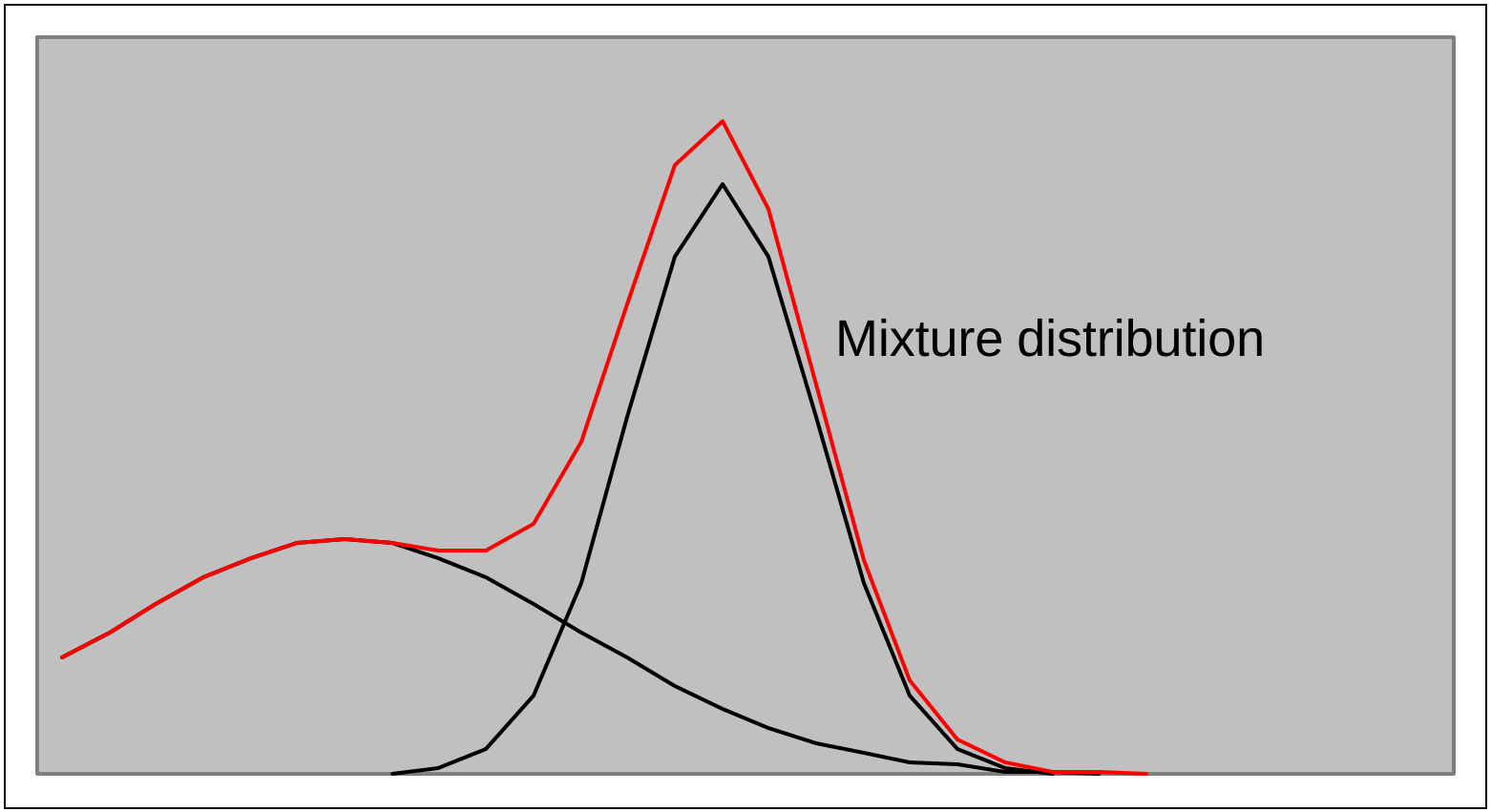
Singlet classifier

UNSUPERVISED ADAPTATION

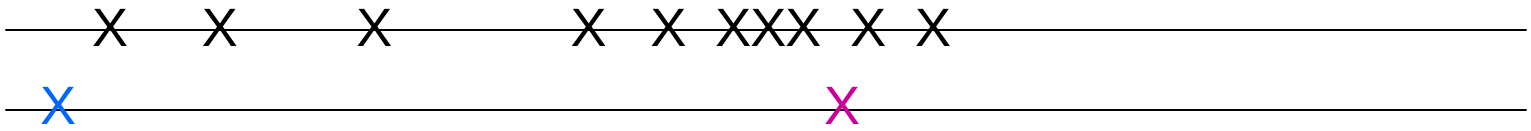
with Henry Baird



THE EXPONENTIAL VALUE OF LABELED SAMPLES (Tom COVER)

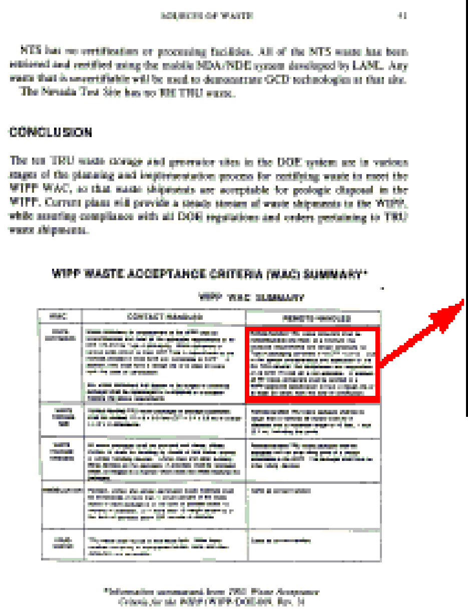


To estimate **m, s**

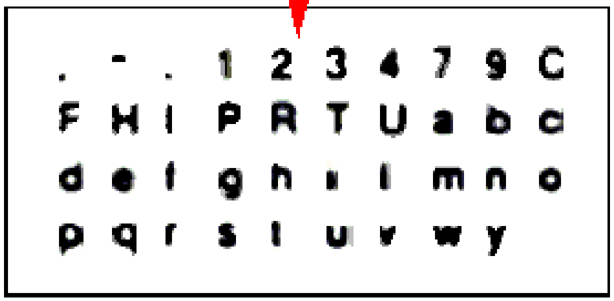


Which is which?

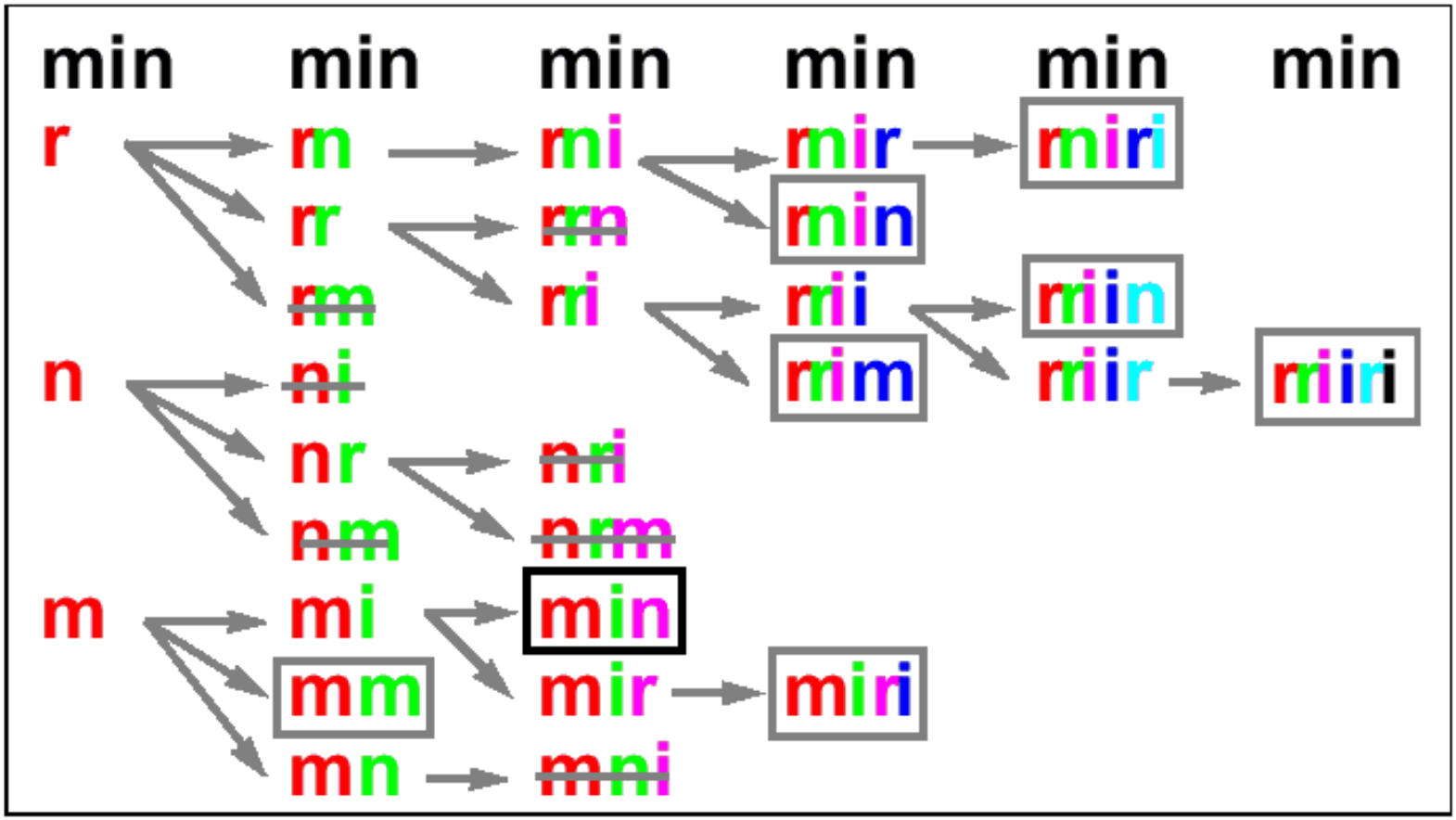
Document-specific Prototype



Remote-handled TRU waste containers shall be noncombustible and meet, as a minimum, the structural requirements and design conditions for Type A packaging contained in 49 CFR 173.412 . Due to the special characteristics and application of the RH TRU canister, the compression test requirement of 49 CFR 173.465 (d) is not applicable. In addition, all RH waste containers shall be certified to a WIPP-approved specification to have a design life of at least 20 years from the date of certification.

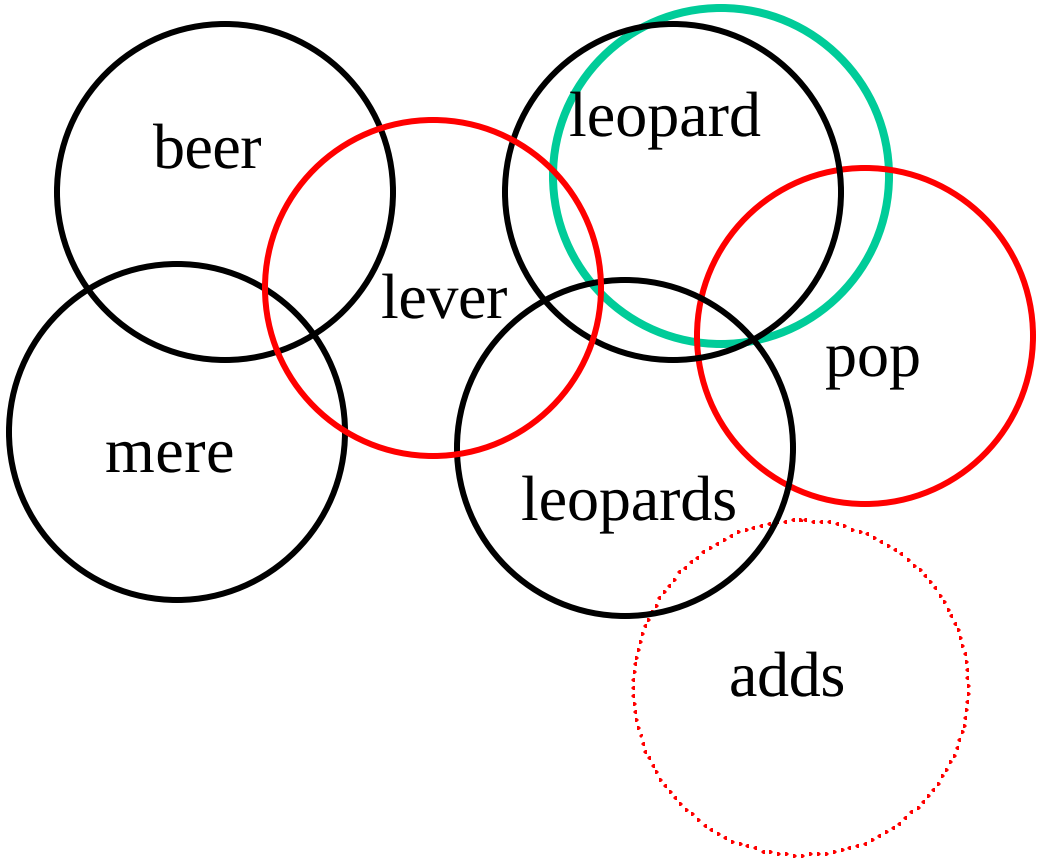


Level-Building Algorithm



Word Discrimination using N-grams

Adnan El-Nasan, Harsha Veermachaneni, Monday 13:35



LEGEND

Unknown word

Lexicon word

Reference word that has a common bigram with the unknown

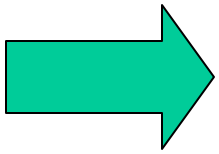
Reference word that has no common bigram with the unknown

Word Discrimination using n-grams

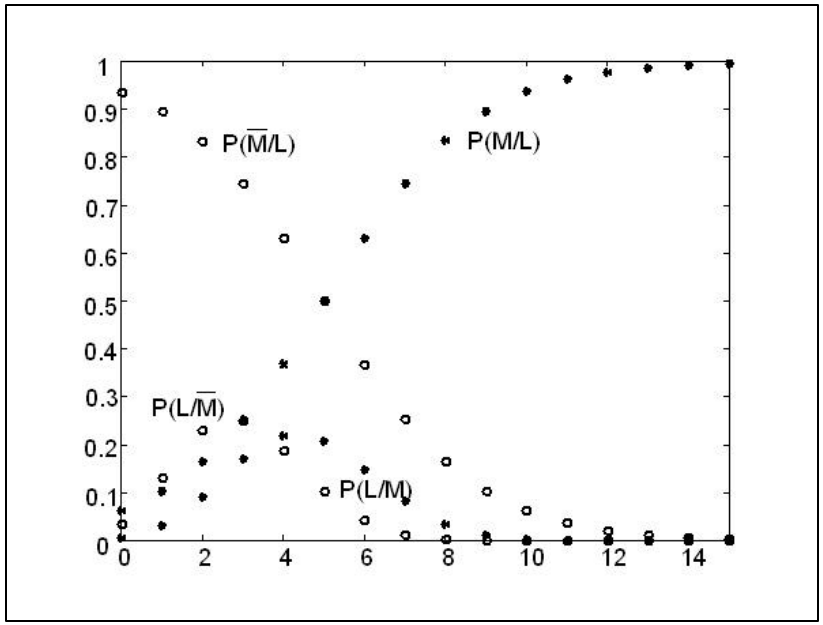
Adnan El-Nasan, Harsha Veermachaneni, Monday 13:35

Lex / Ref	people	herd	open	mile	hazard	ever	period	tripod
have	0	0	0	0	1	1	0	0
lever	1	1	0	1	0	1	1	0
people	1	0	1	1	0	0	1	0
position	0	0	0	0	0	0	1	1

Reference Words	Unknown
	period
have have	4
lever lever	8
people people	10
position position	11



Unknown
4
8
10
11



L	P(Match/L)
4	0.3
8	0.8
10	0.9
11	0.95

S(people) = .7×.8×.9×.05 = .0252

S(herd) = .7×.8×.1×.05 = .0028

S(open) = .7×.2×.9×.05 = .0063

S(mile) = .3×.8×.9×.05 = .0108

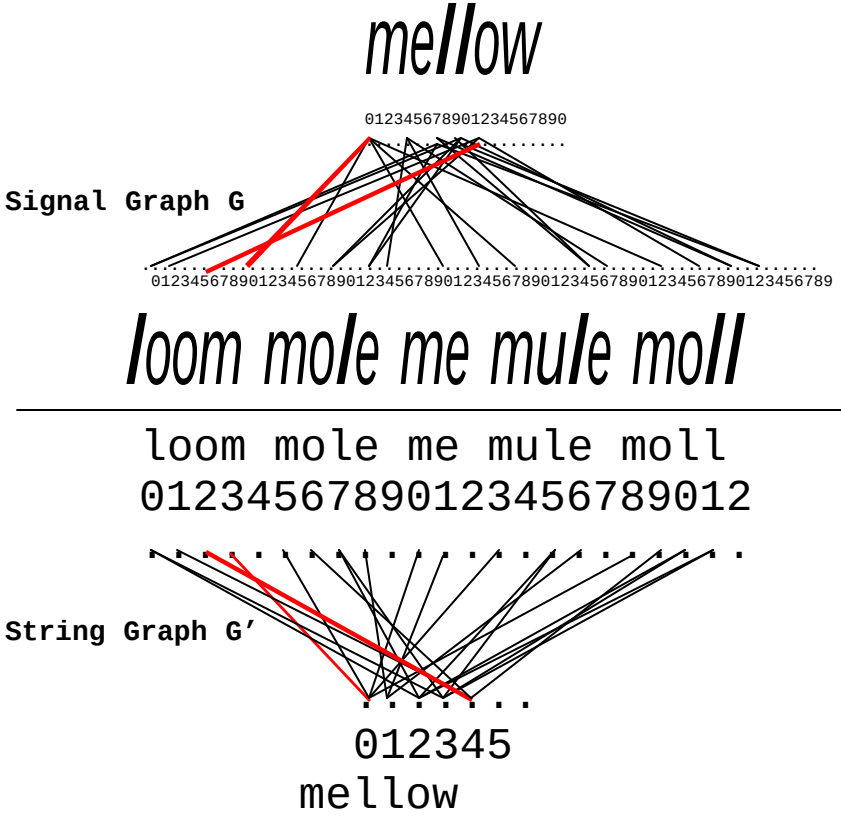
S(hazard) = .3×.2×.1×.05 = .0003

S(ever) = .3×.8×.1×.05 = .0012

S(period) =.7´.8´.9´.95 = .4788

S(tripod) =.7×.2×.1×.95 = .0133

SYMBOLIC INDIRECT CORRELATION



Lexicon L	{low, me, mole, mule, moll, mellow, wool, loom, we}
Unkown q(t)	mellow
Ref Signal r(t)	low me mole mule moll
Ref String r_s	low me mole mule moll
X_{mellow}	{(0,2), (0,3), (1,4), **(2,4), (3,0),** (5,0), (6,4), (7,2), …}
X_{we}	{(8,1), (11,1), (16,1)}
X'(q)	{(0,7), (0,9), (2,11), **(6,11), (10,0),** (15,0), (19,11)…}
f(q)	mellow

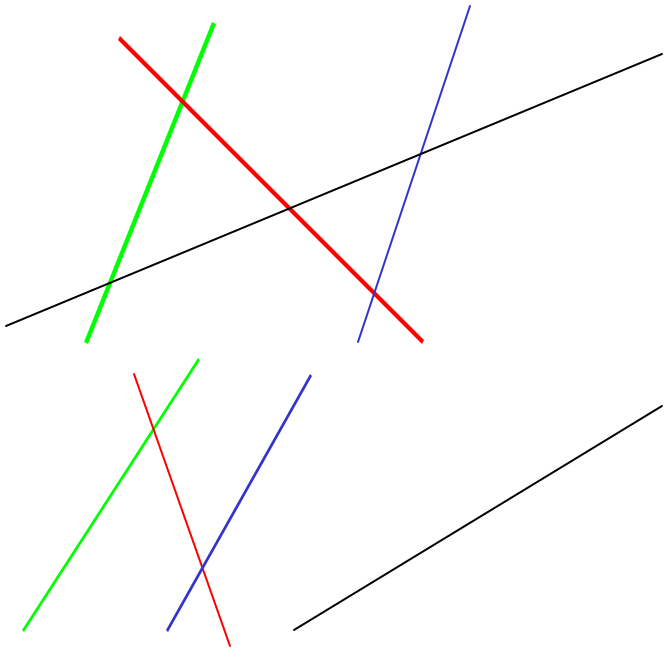
MAXUMUM CLIQUE FORMULATION

(Harsha Veeramachaneni, Prof. M. Krishnamoorthy (RPI))

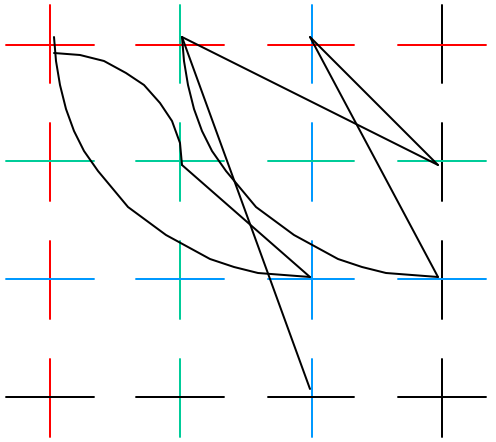
Consider an *Association Graph* where each vertex corresponds to a pair of edges, one in the signal graph and one in the in match graph.

Signal Graph:

String Graph:



Association Graph:



Edges in the Association Graph indicate order compatibility between *pairs of pairs* in the match graphs.

Here the maximum clique is of size 3, the largest *subset order isomorphism*.

SNAP TEST

1. Does Bagging reduce classifier bias, variance, or neither?
   Justify your answer.
2. Can feature correlation reduce classification error over
   uncorrelated features with the same class-conditional means?
3. If two separate sets of features are available, is it better to combine
   feature information before or after classification? Why?
4. What are the three essential conditions for Stochastic Discrimination?
5. What kind of confusions encourage using style-conscious classification?
6. Where and why is EM used in HMM classification?
7. How can unlabeled samples improve classifier accuracy?
   Give an example where the classification after adaptation is worse.
8. Find two English words of at least seven letters that share exactly
   the same letter bigrams.
9. What role do Maximal Cliques play in Indirect Symbolic Correlation?

10. More accurate OCR will result most likely from:

a. Better digitization

b. Improved preprocessing (e.g. layout analysis)

c. More discriminating features

d. More advanced classifiers for isolated patterns

d. Further exploitation of linguistic context

f. Style context

g. Unsupervised adaptation

h. Whole-word, line, or page classification

i. None of the above.

Please justify your choice.

THANK YOU!