An Introduction to Boosting

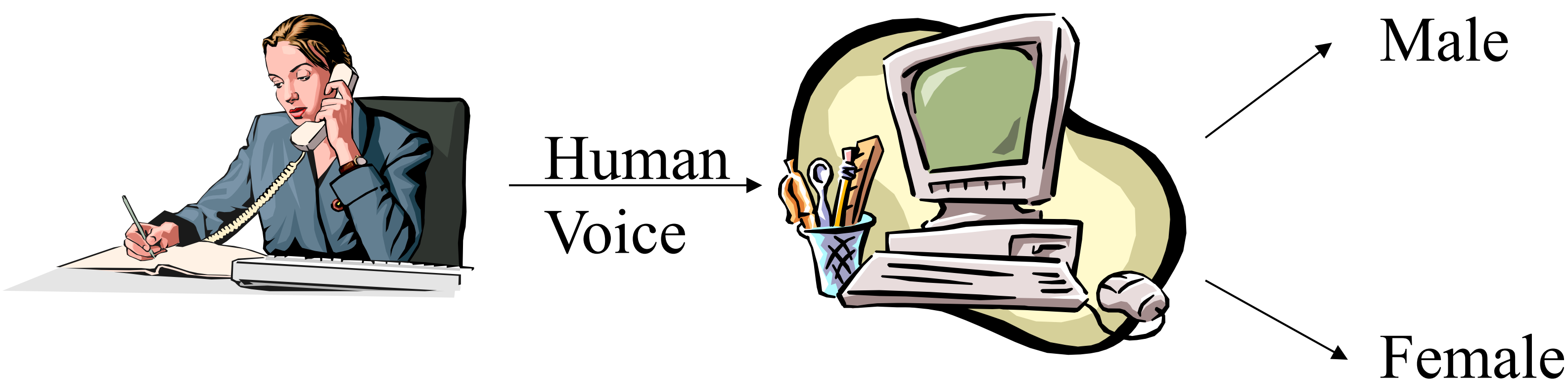
Yoav Freund
Banter Inc.

Plan of talk

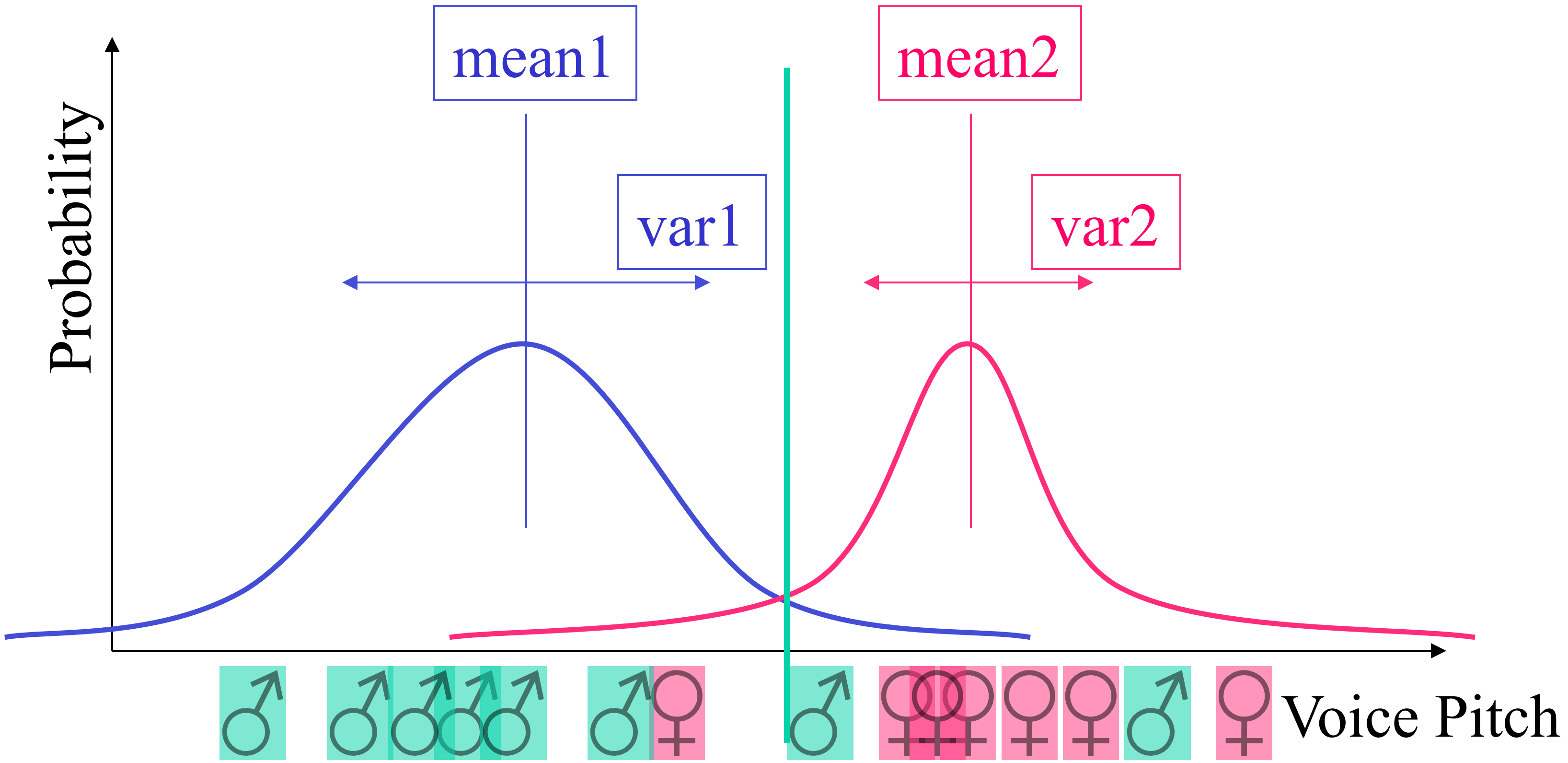
- Generative vs. non-generative modeling
- Boosting
- Alternating decision trees
- Boosting and over-fitting
- Applications

Toy Example

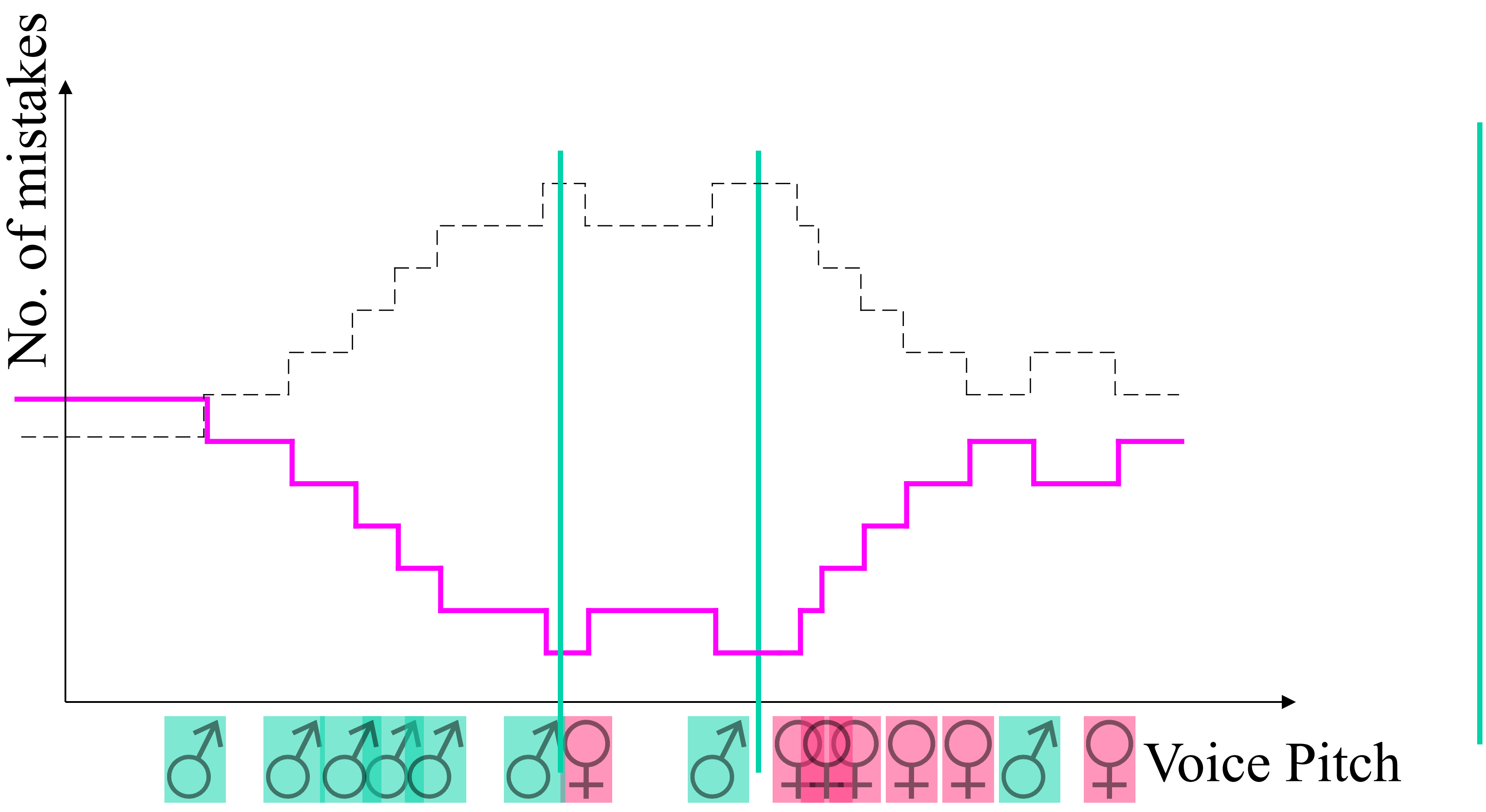
- Computer receives telephone call
- Measures Pitch of voice
- Decides gender of caller



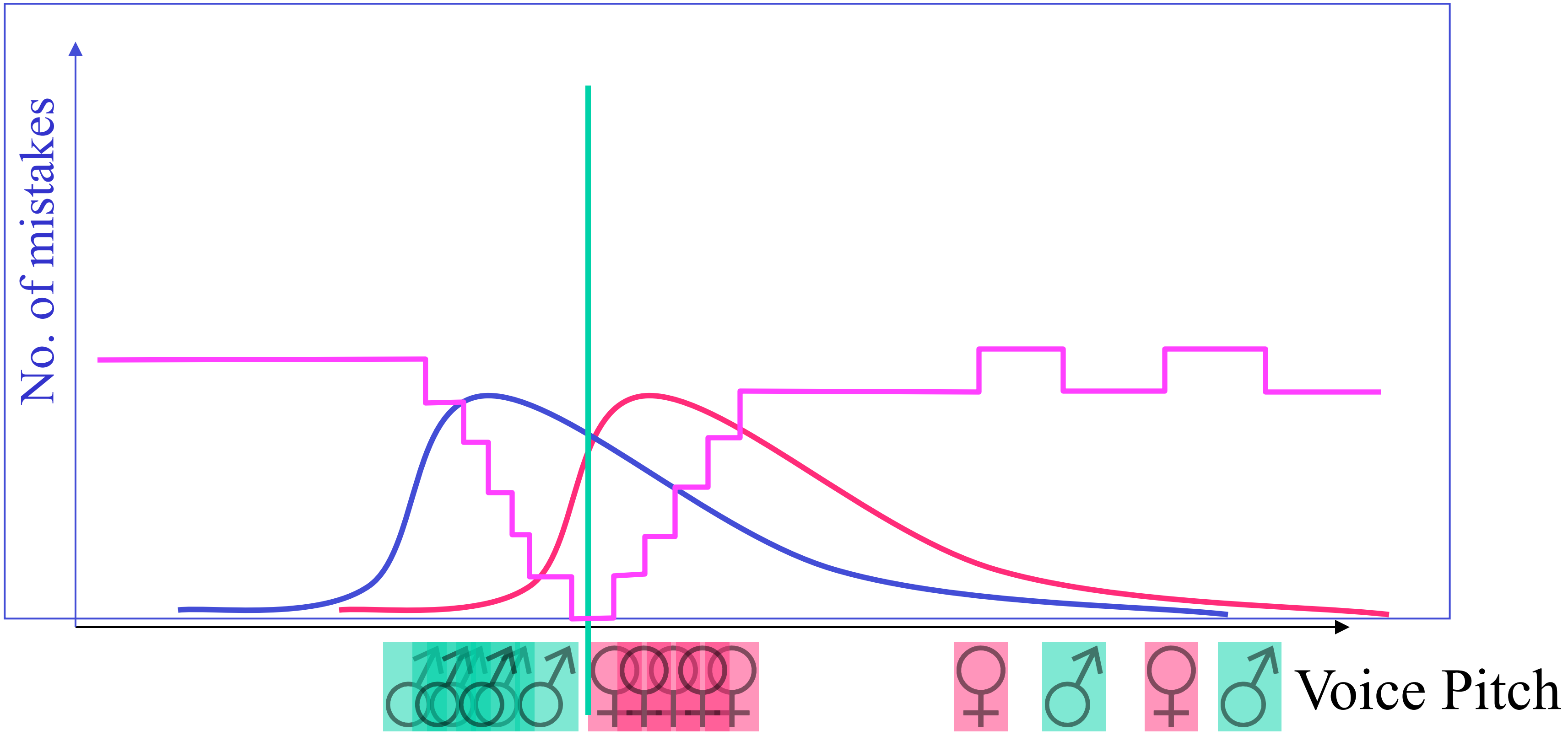
Generative modeling



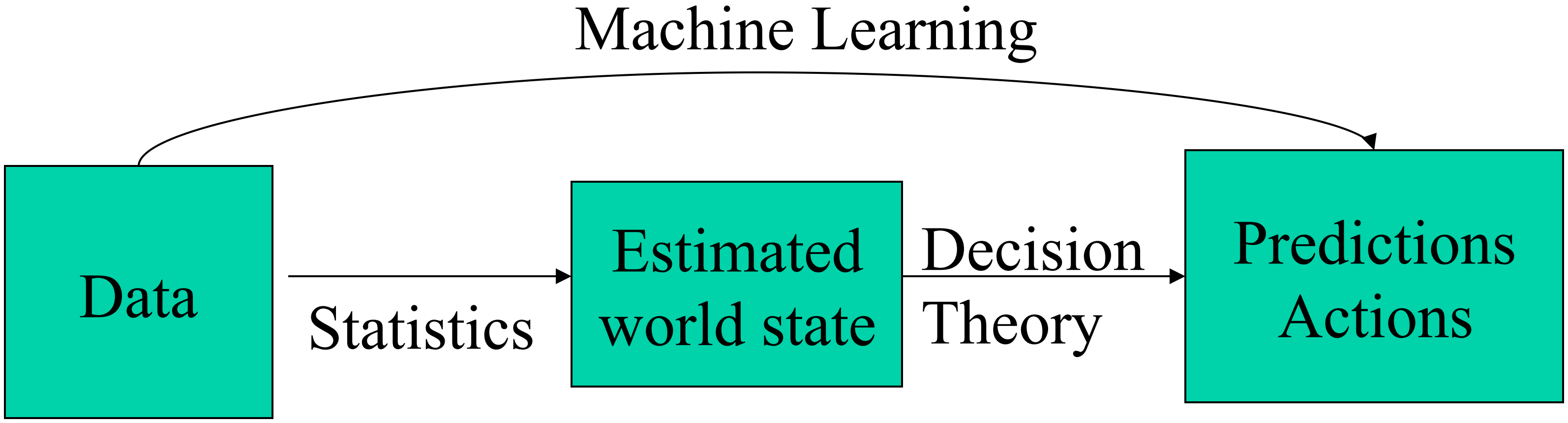
Discriminative approach



Ill-behaved data



Traditional Statistics vs. Machine Learning

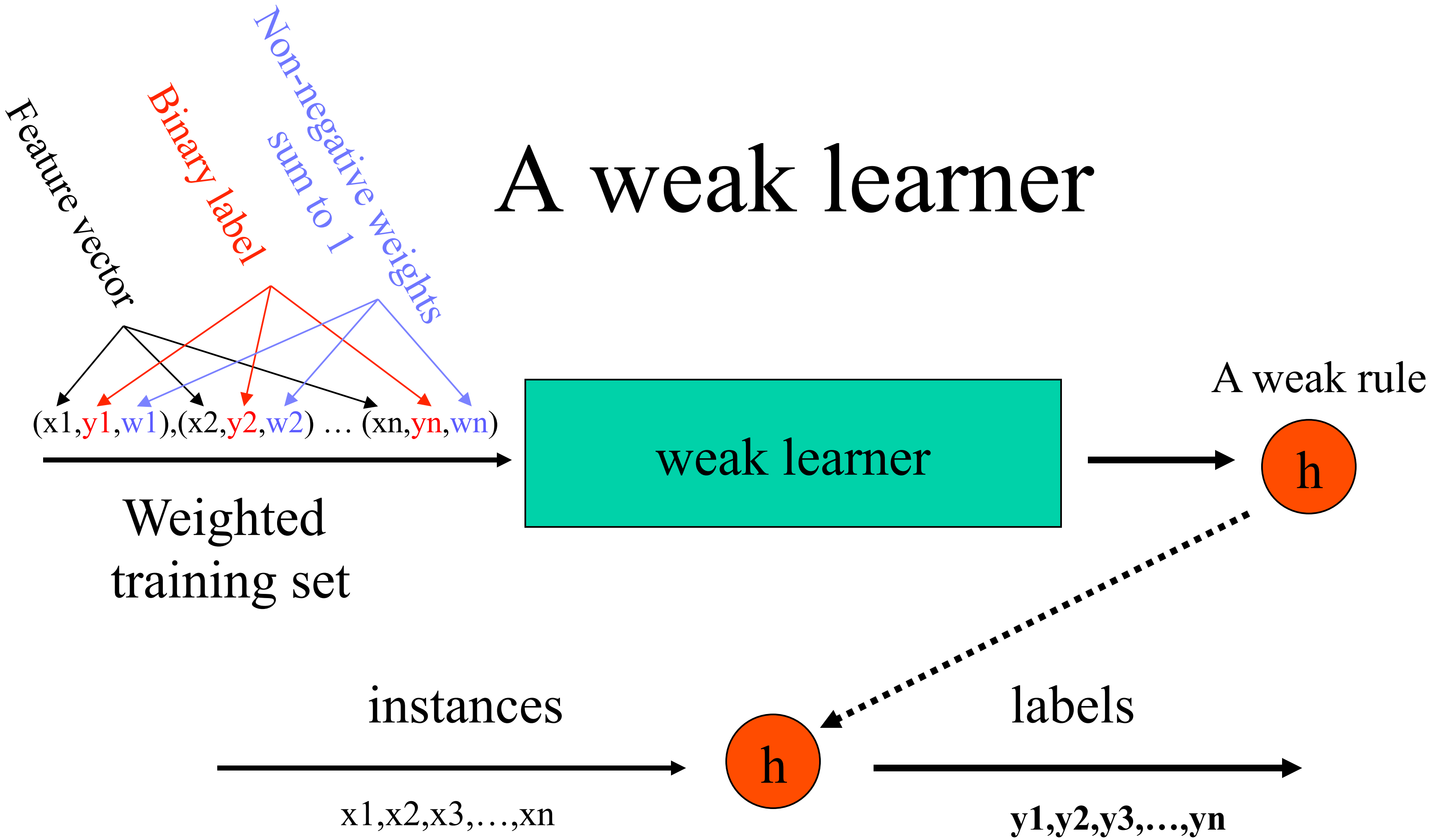


Comparison of methodologies

Model	Generative	Discriminative
Goal	Probability estimates	Classification rule
Performance measure	Likelihood	Misclassification rate
Mismatch problems	Outliers	Misclassifications

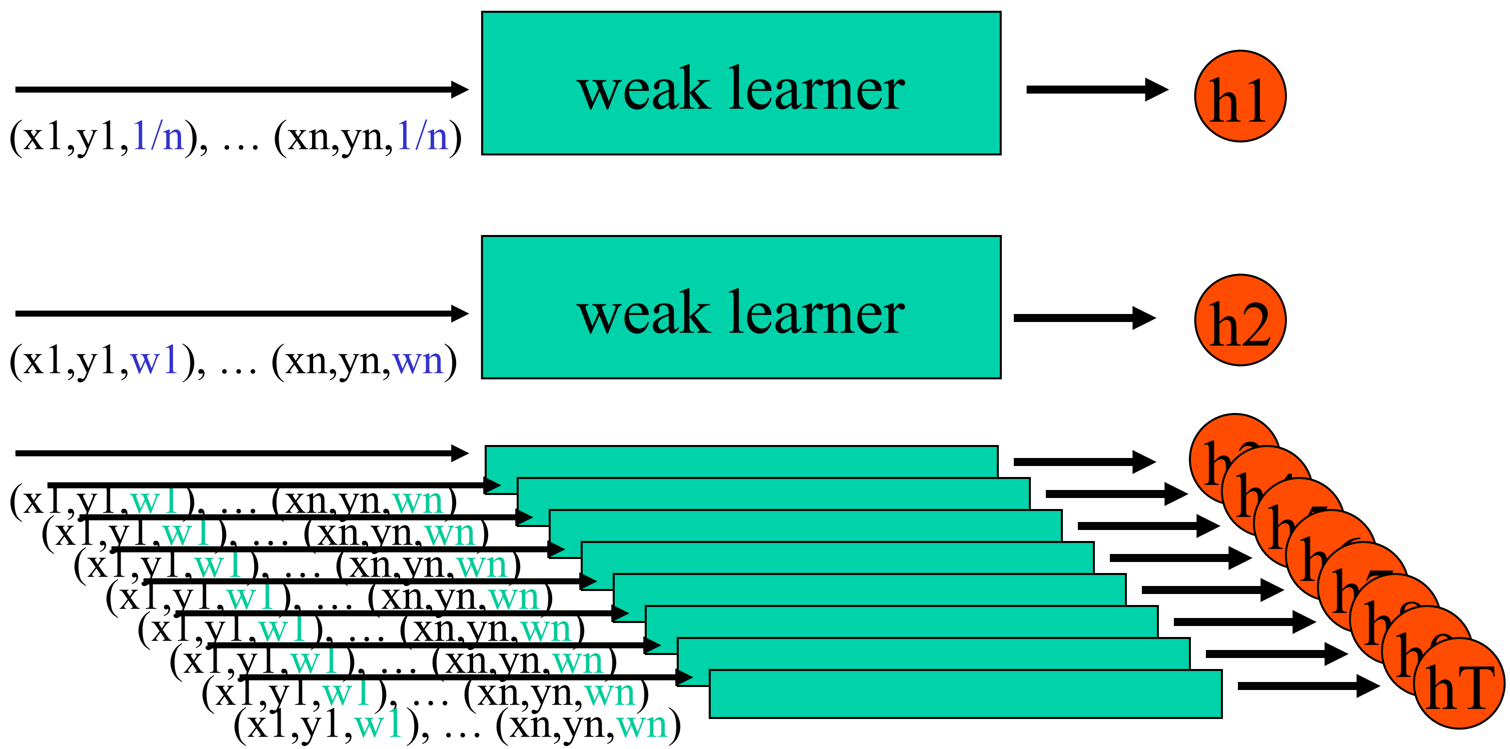
Boosting

A weak learner



The weak requirement: $\dfrac{\sum_{i:y_i \neq \mathbf{y}_i} w_i}{\sum_{i=1}^{n} w_i} < \dfrac{1}{2} - \gamma$

The boosting process



Final rule: Sign[a1 h1 + a2 h2 + + aT hT]

Adaboost

- Binary labels $y = -1,+1$
- $\mathrm{margin}(x,y) = y\,[\Sigma_t\, \alpha_t\, h_t(x)]$
- $P(x,y) = (1/Z)\exp(-\mathrm{margin}(x,y))$
- Given h_t, we choose α_t to minimize

$$\Sigma_{(x,y)} \exp(-\mathrm{margin}(x,y))$$

Main property of adaboost

- If advantages of weak rules over random guessing are: **$g_1, g_2, .., g_T$** then in-sample error of final rule is at most

$$\exp\left(-2\sum_{t=1}^{T}\gamma_t^2\right)$$

(w.r.t. the initial weights)

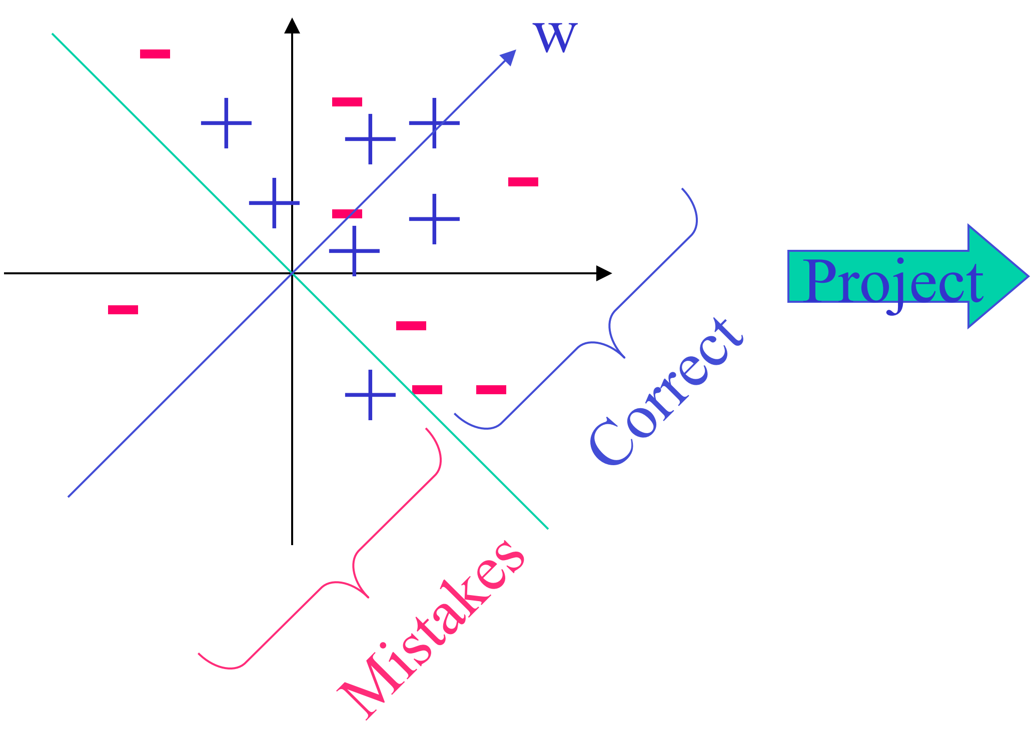
Adaboost as gradient descent

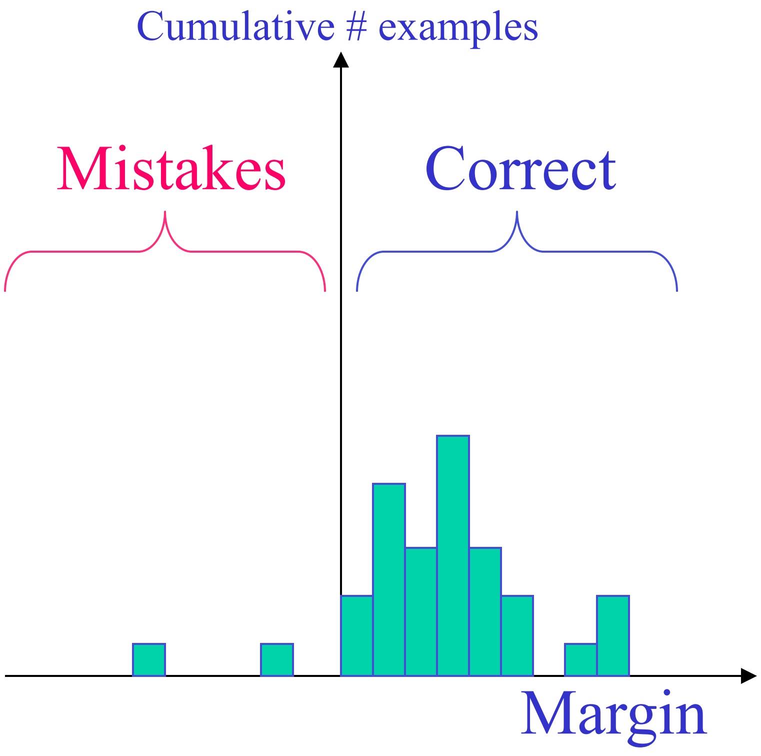
- Discriminator class: a linear discriminator in the space of "weak hypotheses"
- Original goal: find hyper plane with smallest number of mistakes
  - Known to be an NP-hard problem (no algorithm that runs in time polynomial in d, where d is the dimension of the space)
- Computational method: Use exponential loss as a surrogate, perform gradient descent.

Margins view

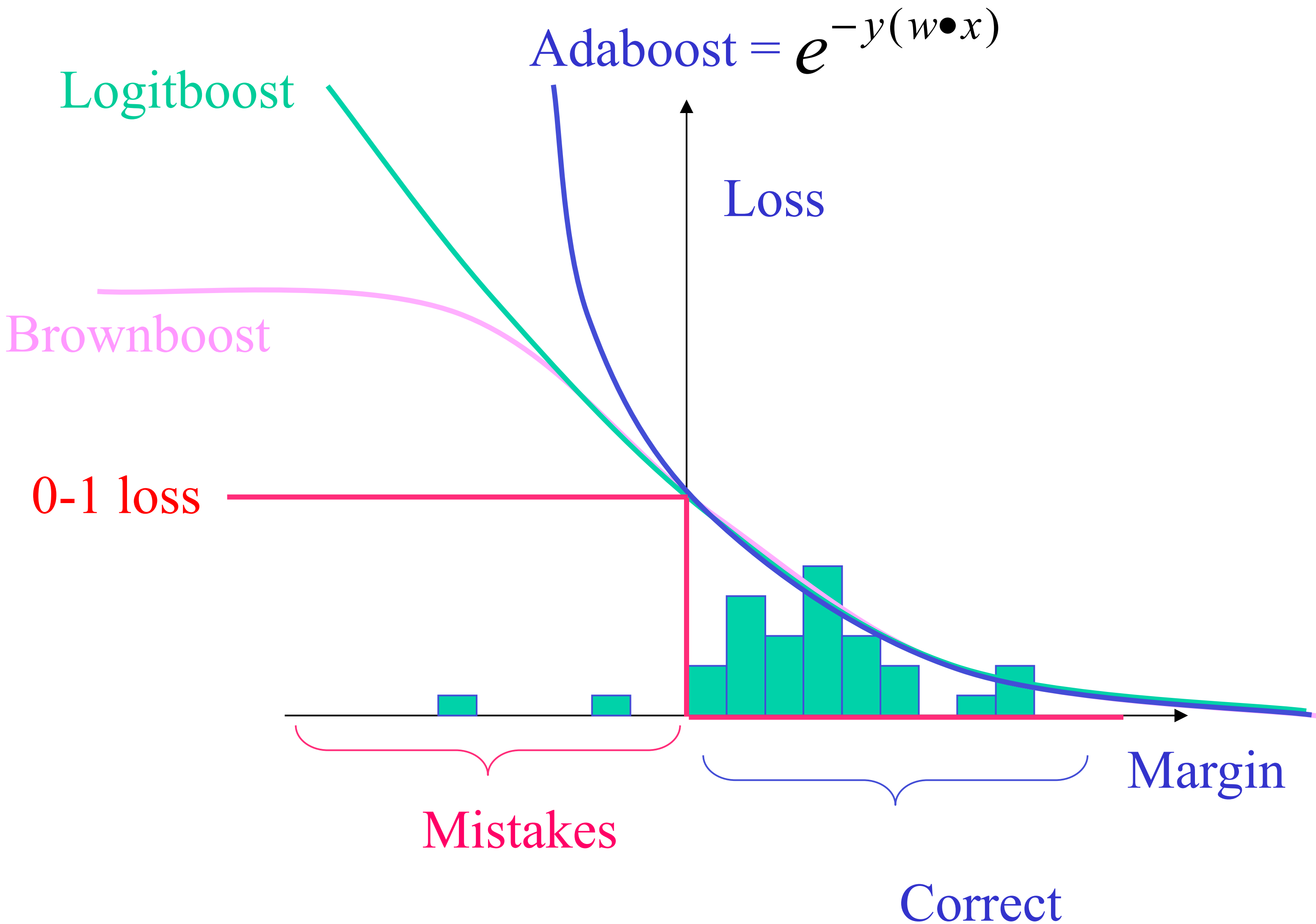
$x, w \in R^n; y \in \{-1,+1\}$ Prediction = $sign(w \bullet x)$

Margin = $\dfrac{y(w \bullet x)}{\|w\| \cdot \|x\|}$



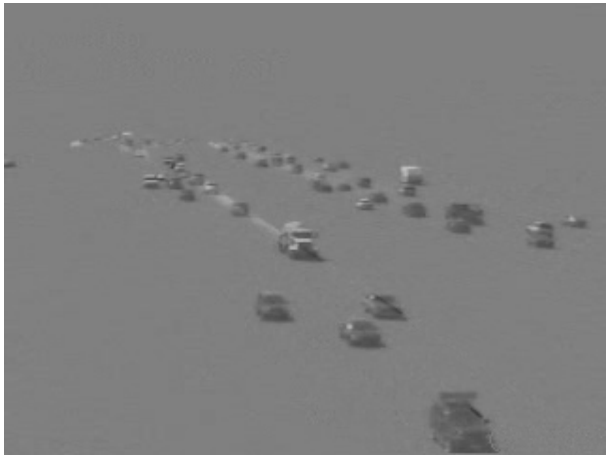


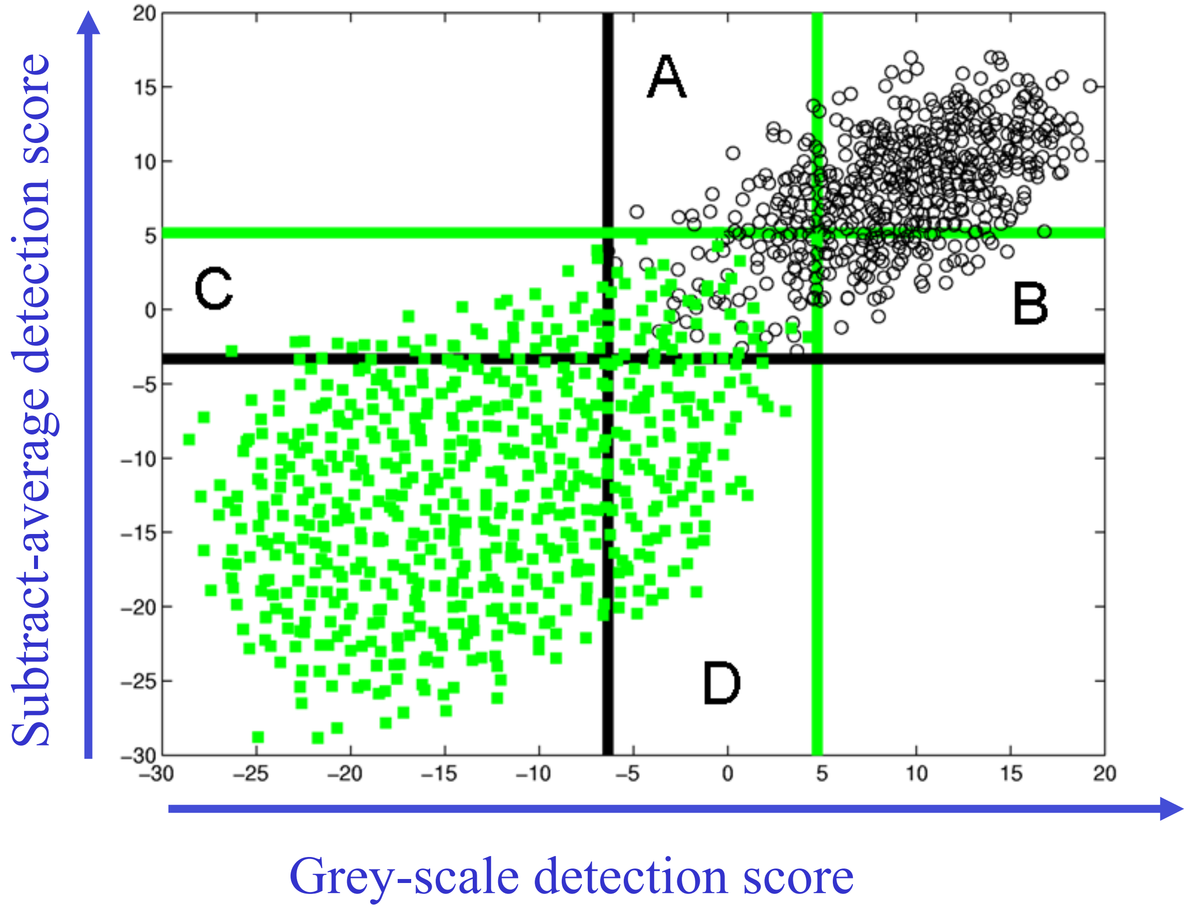
Adaboost et al.




I-405
NE 124TH ST
© WSDOT


SR 520
MIDSPAN
© WSDOT





A
B
C
D
20
15
10
5
0
−5
−10
−15
−20
−25
−30
−30
−25
−20
−15
−10
−5
0
5
10
15
20
Subtract-average detection score
Grey-scale detection score

One coordinate at a time

- Adaboost performs gradient descent on exponential loss
- Adds one coordinate ("weak learner") at each iteration.
- Weak learning in binary classification = slightly better than random guessing.
- Weak learning in regression – unclear.
- Uses example-weights to communicate the gradient direction to the weak learner
- Solves a computational problem

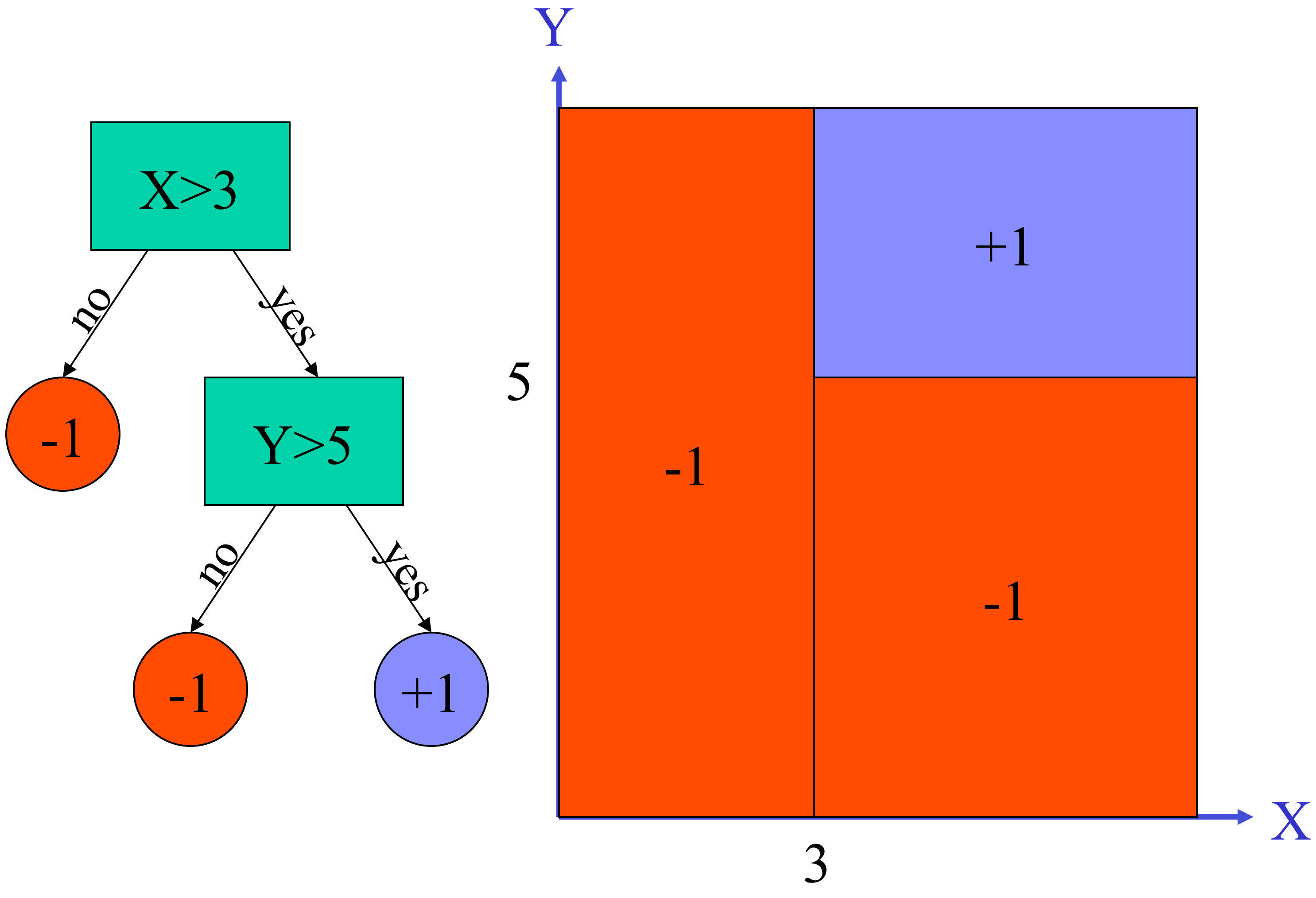
What is a good weak learner?

- The set of weak rules (features) should be flexible enough to be (weakly) correlated with most conceivable relations between feature vector and label.
- Small enough to allow exhaustive search for the minimal weighted training error.
- Small enough to avoid over-fitting.
- Should be able to calculate predicted label very efficiently.
- Rules can be "specialists" – predict only on a small subset of the input space and abstain from predicting on the rest (output 0).

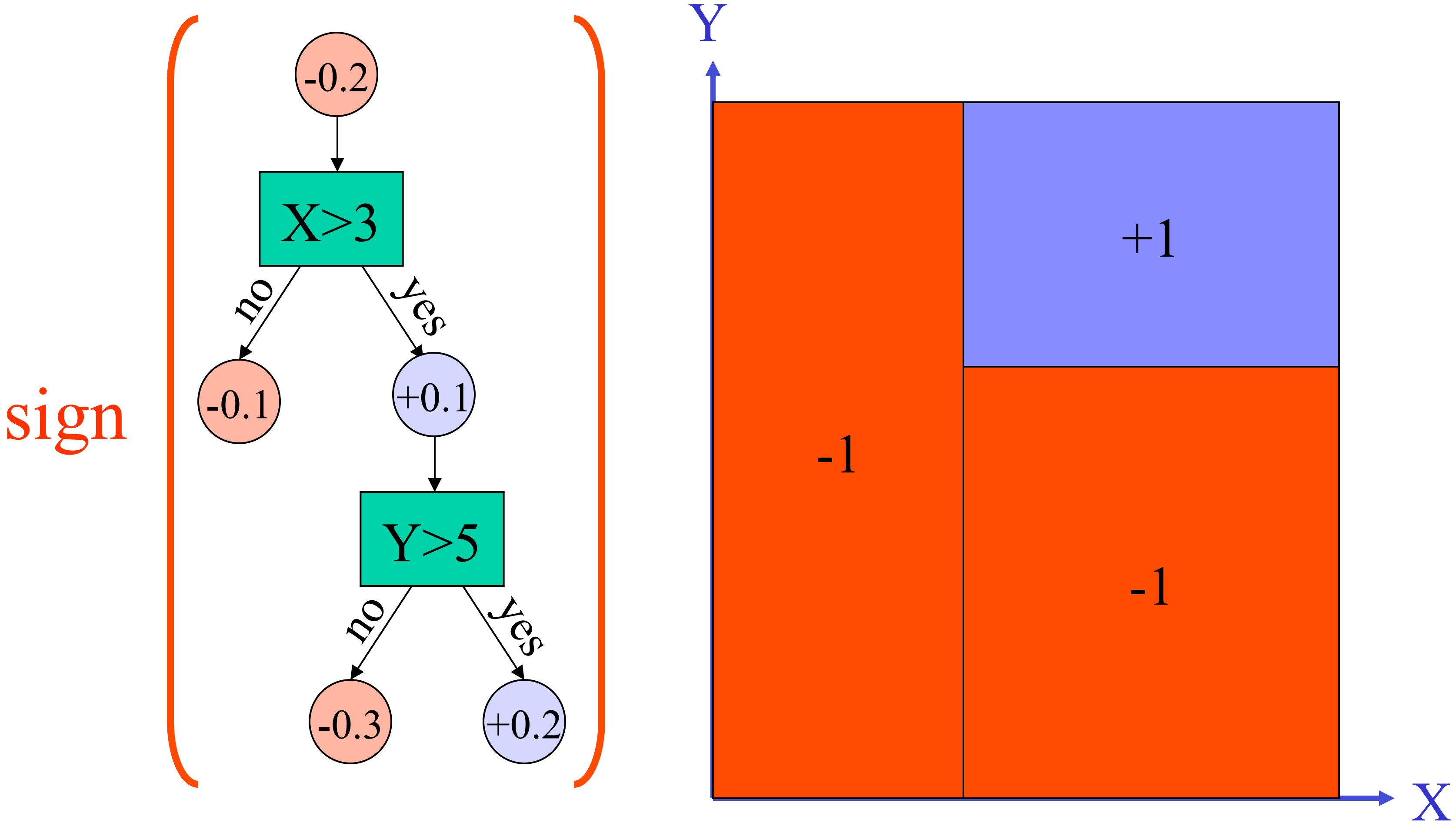
Alternating Trees

Joint work with Llew Mason

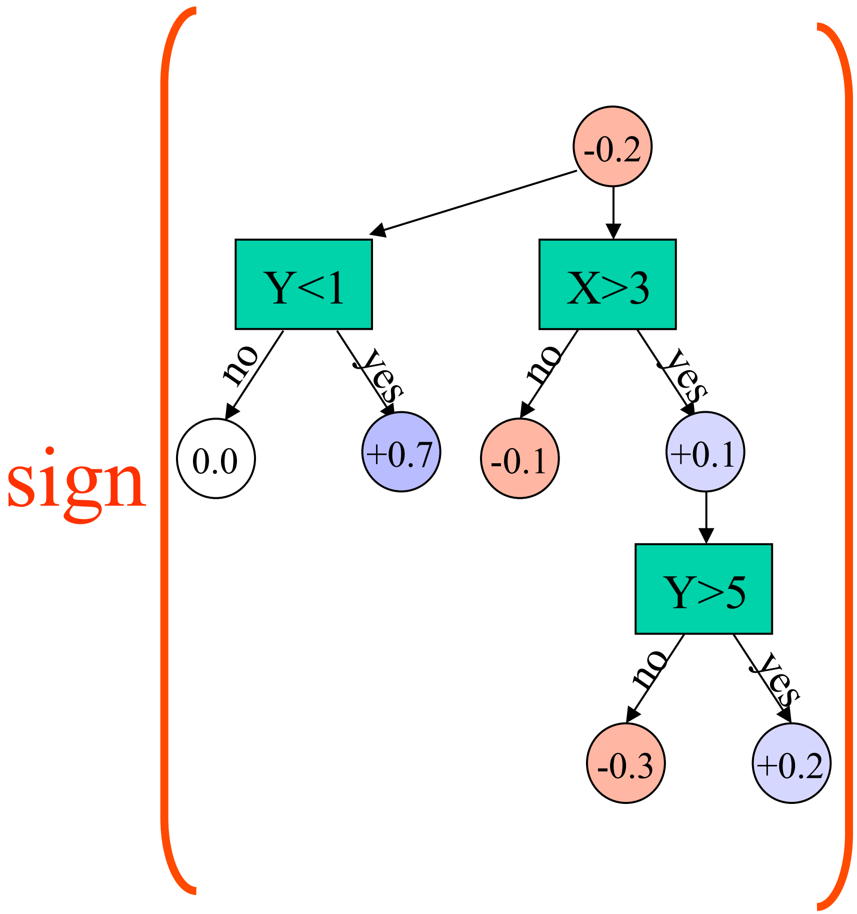
Decision Trees

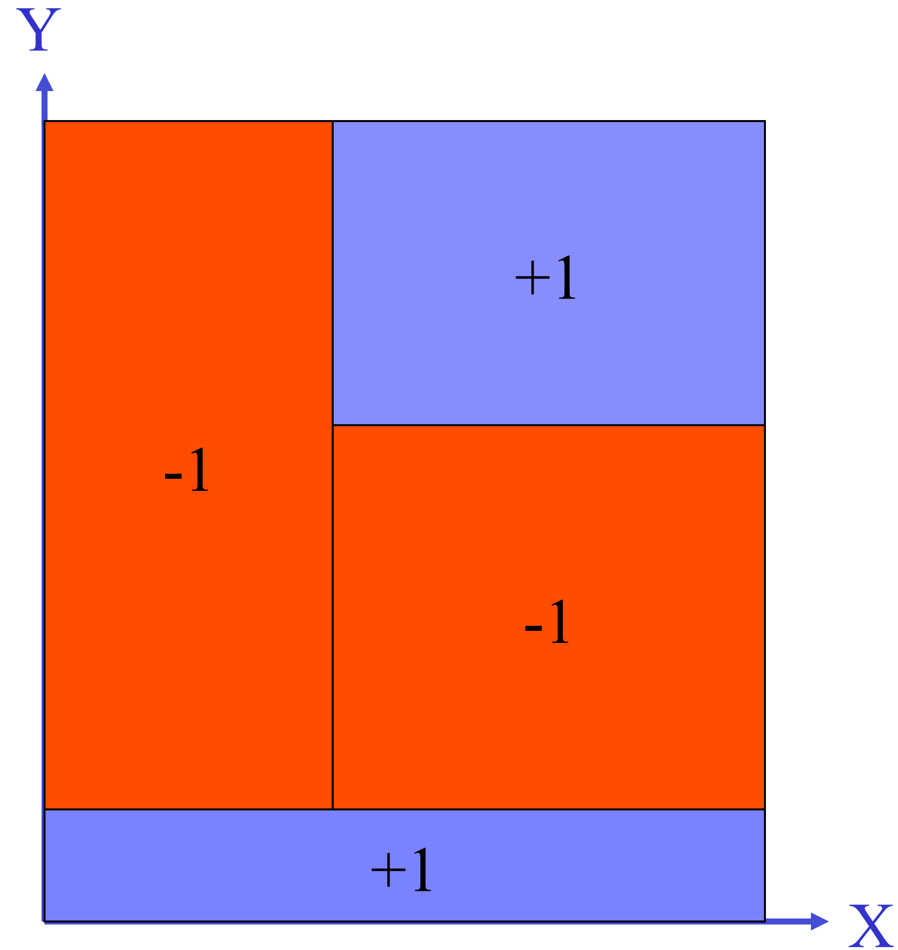


Decision tree as a sum



An alternating decision tree

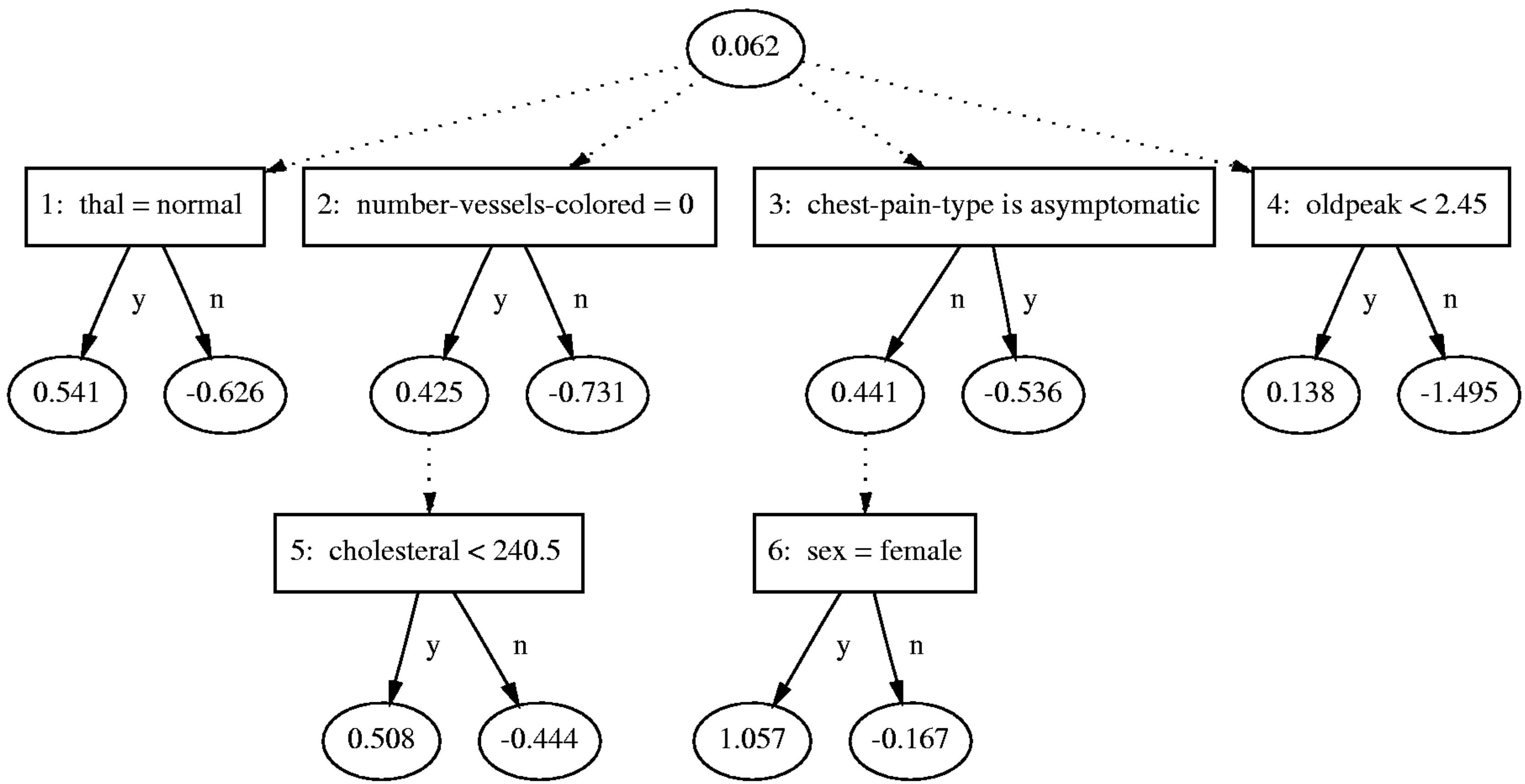




Example: Medical Diagnostics

- `Cleve` dataset from UC Irvine database.
- Heart disease diagnostics (+1=healthy,-1=sick)
- 13 features from tests (real valued and discrete).
- 303 instances.

Adtree for Cleveland heart-disease diagnostics problem



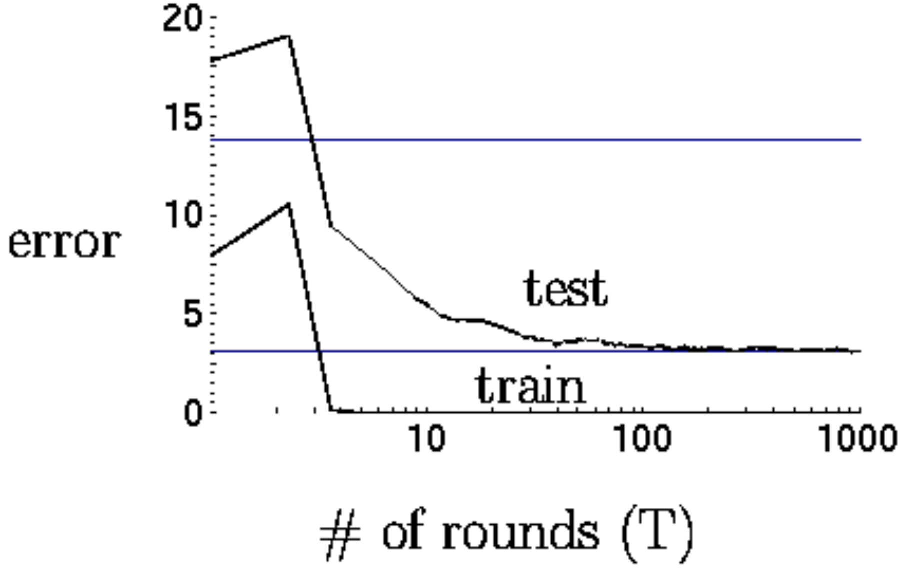
Cross-validated accuracy

Learning algorithm	Number of splits	Average test error	Test error variance
ADtree	6	17.0%	0.6%
C5.0	27	27.2%	0.5%
C5.0 + boosting	446	20.2%	0.5%
Boost Stumps	16	16.5%	0.8%

Boosting and over-fitting

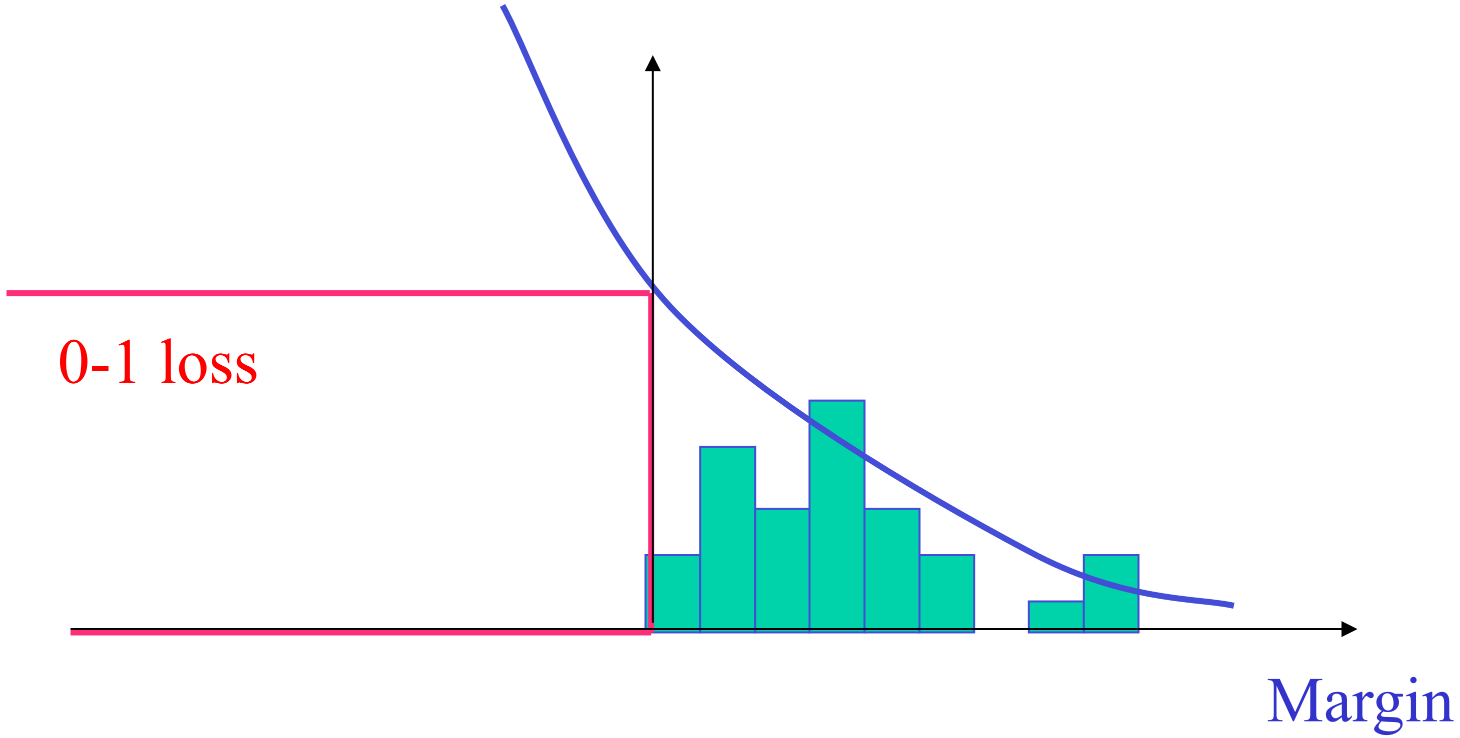
Curious phenomenon

Boosting decision trees

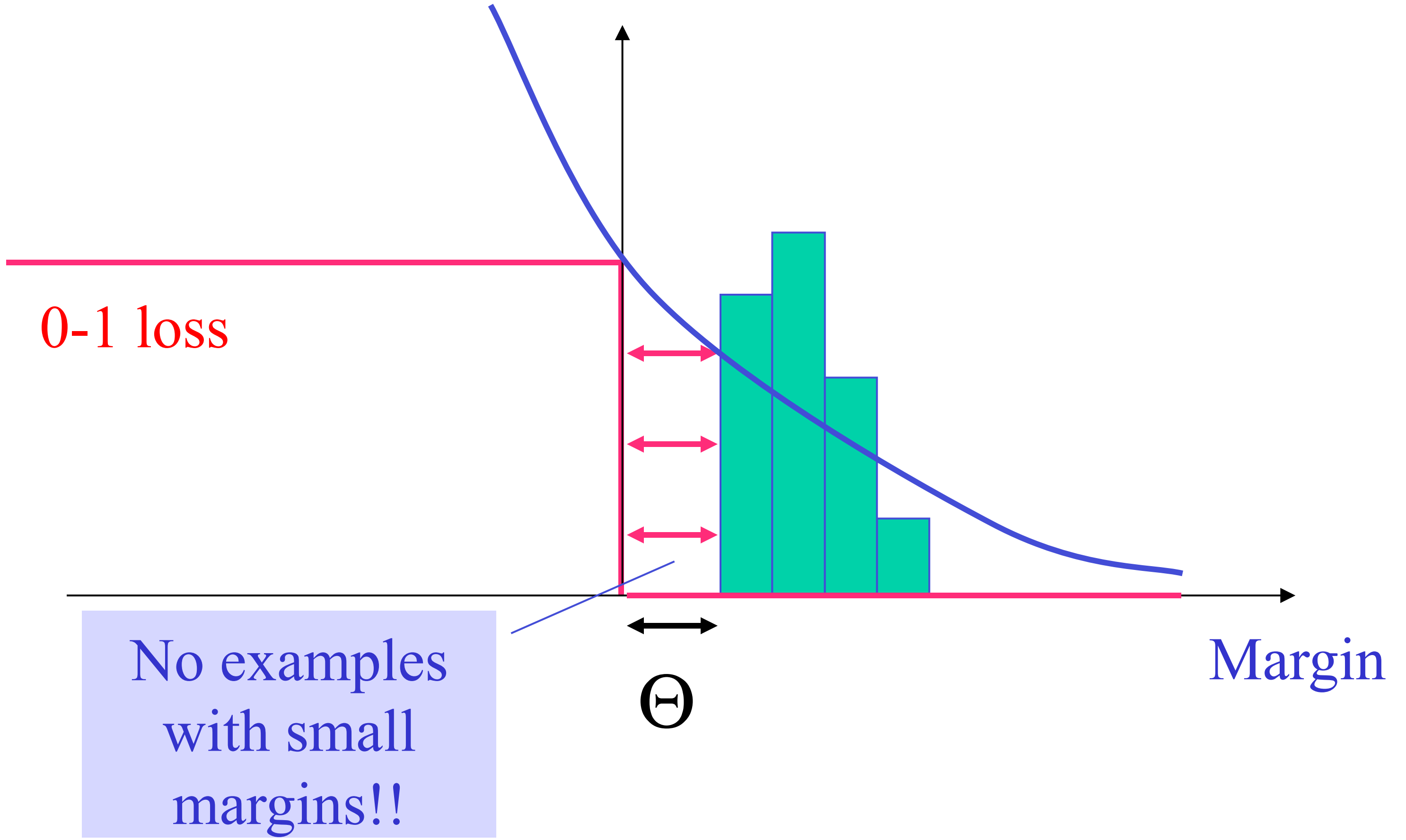


Using <10,000 training examples we fit >2,000,000 parameters

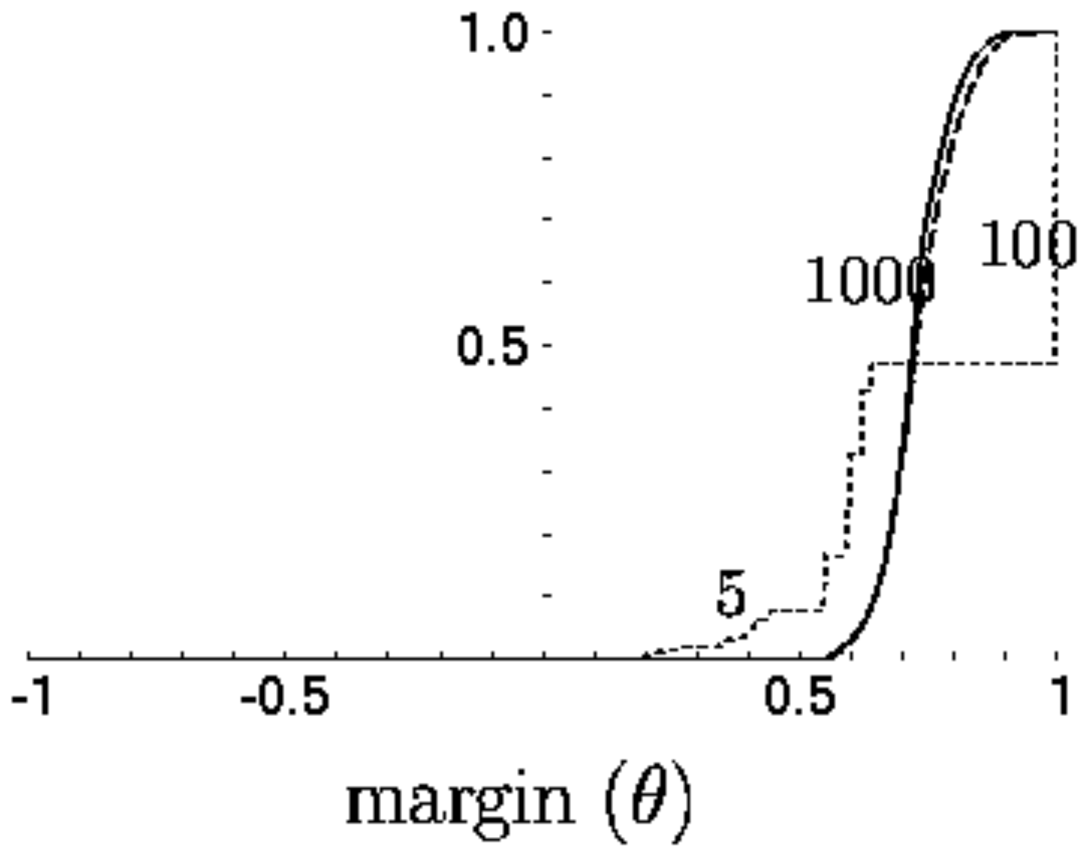
Explanation using margins



Explanation using margins



Experimental Evidence



Theorem

Schapire, Freund, Bartlett & Lee
Annals of stat. 98

For any convex combination and any threshold $\forall f \in C, \forall \theta > 0:$

$$\underbrace{P_{(x,y)\sim D}\left[\mathrm{sign}(f(x)) \neq y\right]}_{\text{Probability of mistake}} \leq \underbrace{P_{(x,y)\sim S}\left[margin_f(x,y) \leq \theta\right]}_{\text{Fraction of training example with small margin}}$$

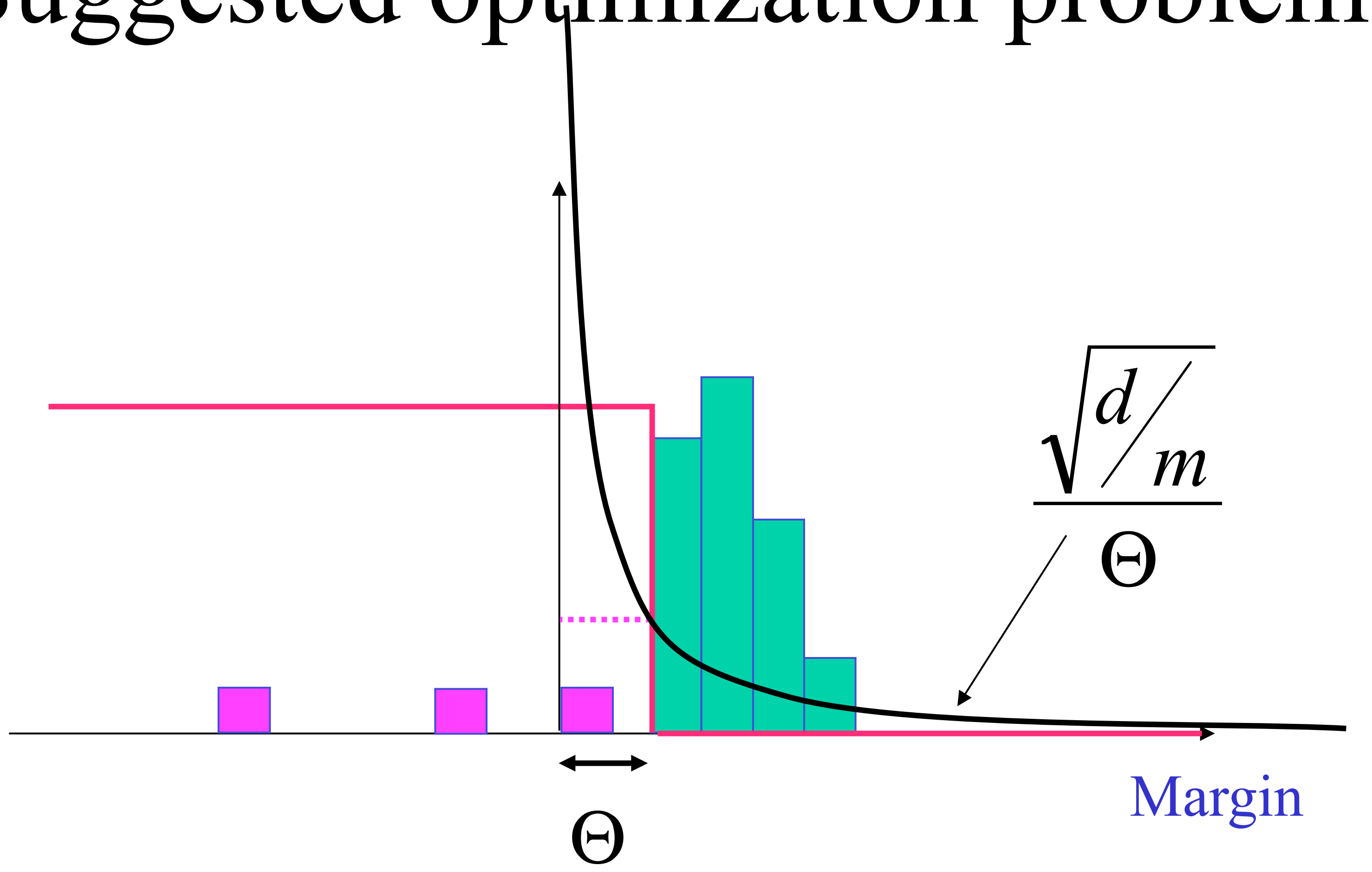
Size of training sample

$$+ \; \tilde{O}\left(\frac{\sqrt{d/m}}{\theta}\right)$$

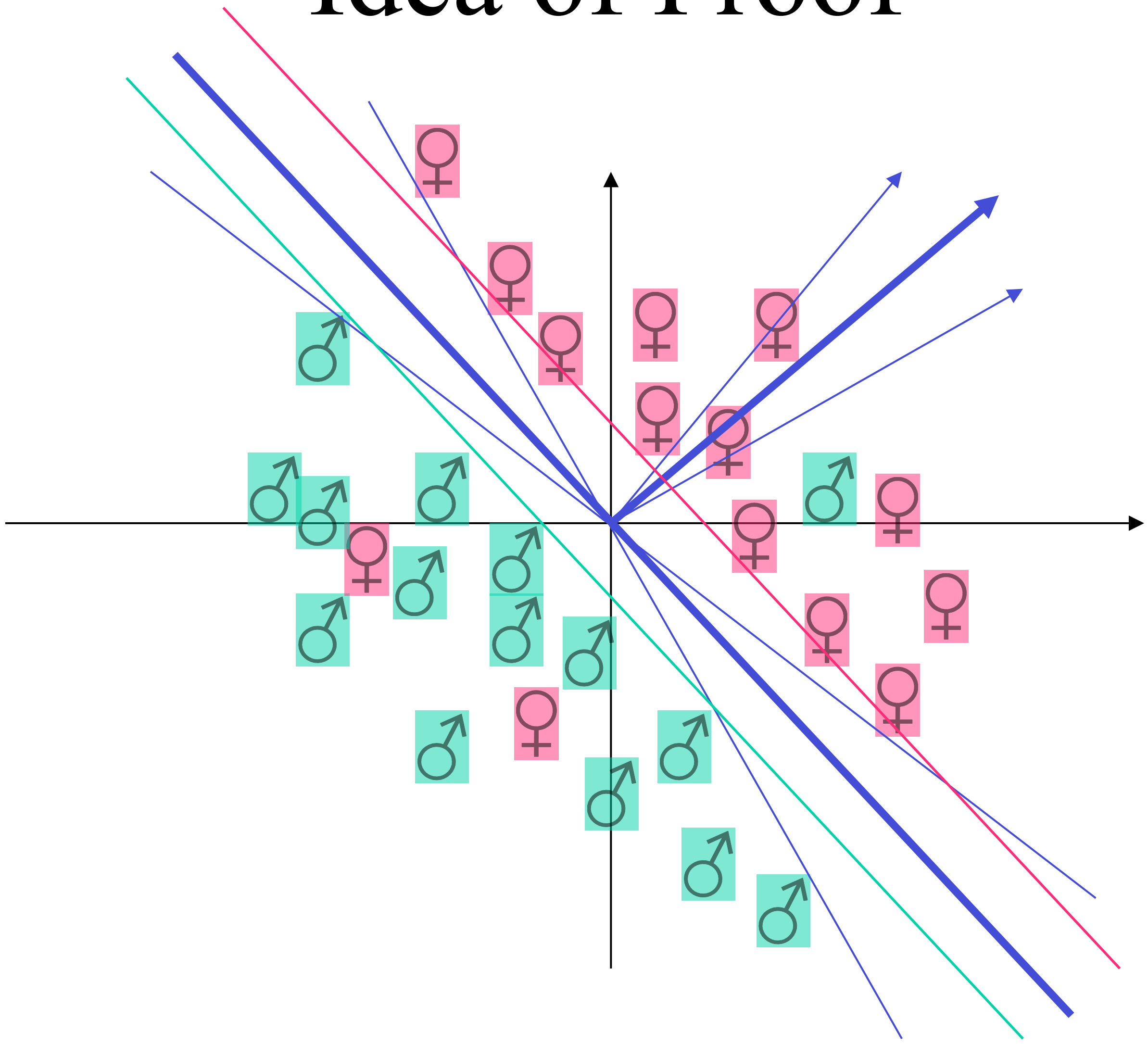
VC dimension of weak rules

No dependence on number of weak rules that are combined!!!

Suggested optimization problem



Idea of Proof



Applications

Applications of Boosting

- Academic research
- Applied research
- Commercial deployment

Academic research

% test error rates

Database	Other	Boosting	Error reduction
Cleveland	27.2 (DT)	16.5	39%
Promoters	22.0 (DT)	11.8	46%
Letter	13.8 (DT)	3.5	74%
Reuters 4	5.8, 6.0, 9.8	2.95	~60%
Reuters 8	11.3, 12.1, 13.4	7.4	~40%

Schapire, Singer, Gorin 98

Applied research

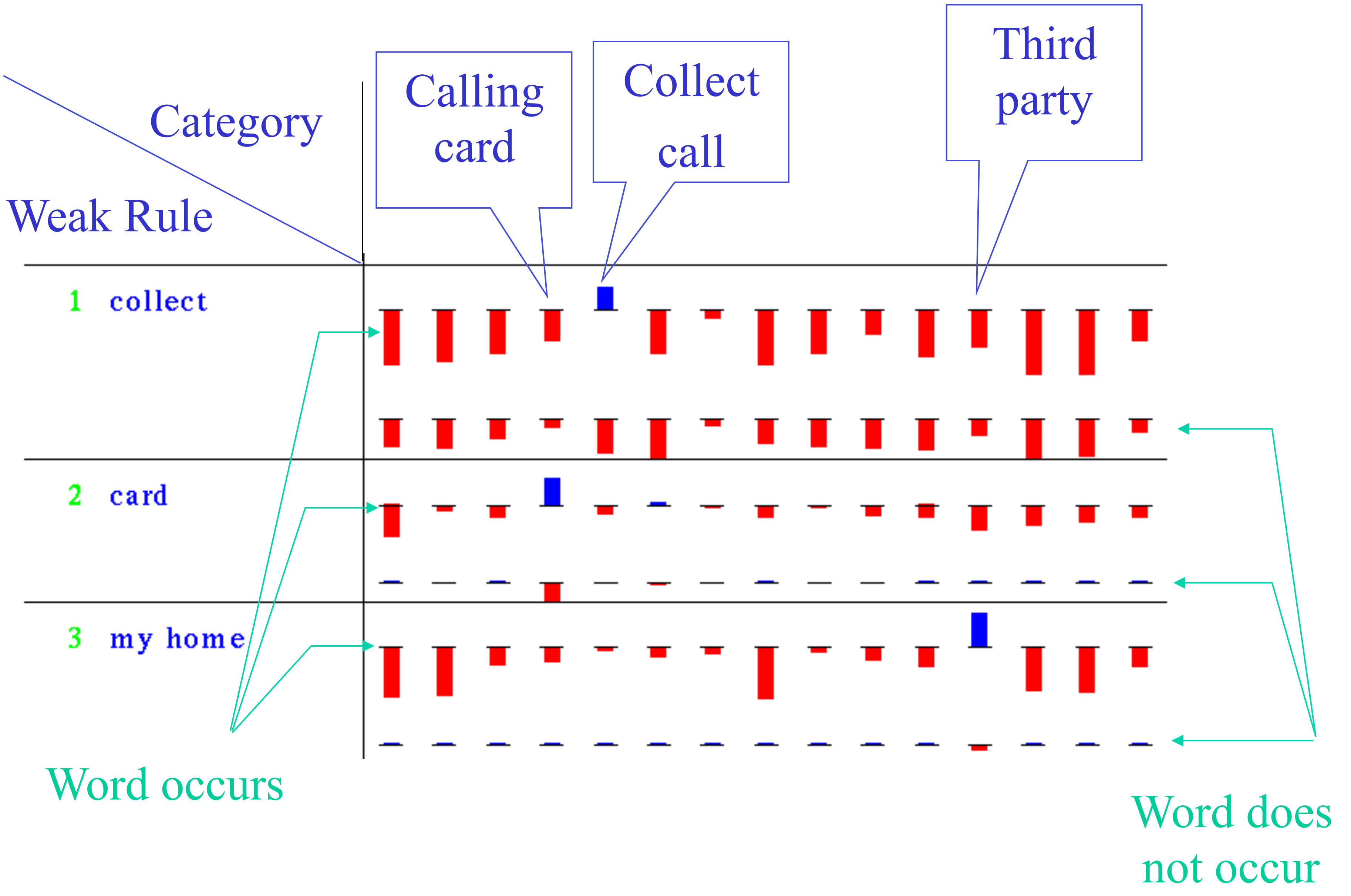
- *"AT&T, How may I help you?"*
- Classify voice requests
- Voice -> text -> category
- Fourteen categories

  Area code, AT&T service, billing credit, calling card, collect, competitor, dial assistance, directory, how to dial, person to person, rate, third party, time charge ,time

Examples

- Yes I'd like to place a collect call long distance please
  - collect
- Operator I need to make a call but I need to bill it to my office
  - third party
- Yes I'd like to place a call on my master card please
  - calling card
- I just called a number in Sioux city and I musta rang the wrong number because I got the wrong party and I would like to have that taken off my bill
  - billing credit

Weak rules generated by “boostexter”



Results

- 7844 training examples
  - hand transcribed
- 1000 test examples
  - hand / machine transcribed
- Accuracy with 20% rejected
  - Machine transcribed: 75%
  - Hand transcribed: 90%

Freund, Mason, Rogers, Pregibon, Cortes 2000

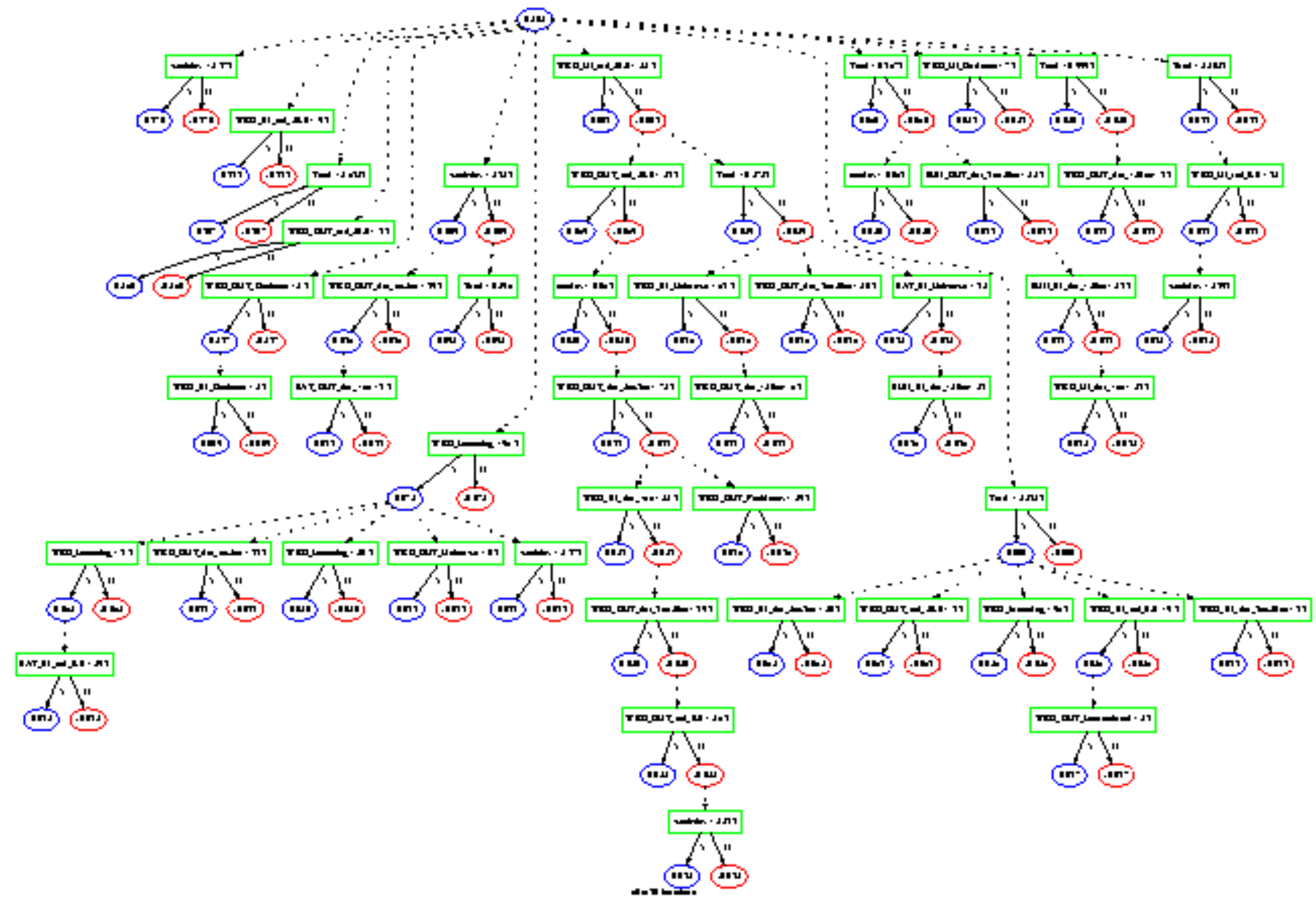
Commercial deployment

- Distinguish business/residence customers
- Using statistics from call-detail records
- Alternating decision trees
  - Similar to boosting decision trees, more flexible
  - Combines very simple rules
  - Can over-fit, cross validation used to stop

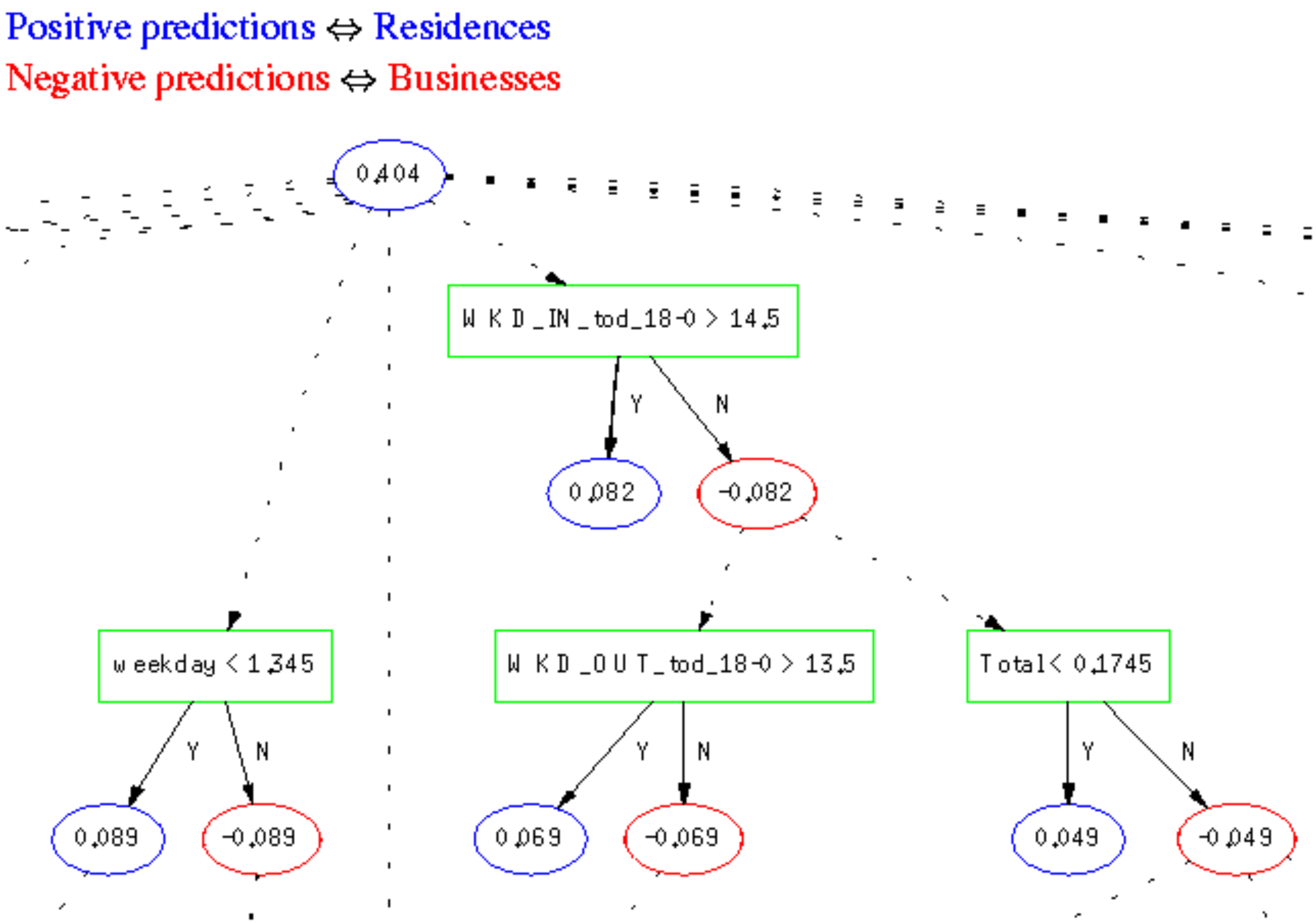
Massive datasets

- 260M calls / day
- 230M telephone numbers
- Label unknown for ~30%
- Hancock: software for computing statistical signatures.
- 100K randomly selected training examples,
- ~10K is enough
- Training takes about 2 hours.
- Generated classifier has to be both accurate and efficient

Alternating tree for “buizocity”

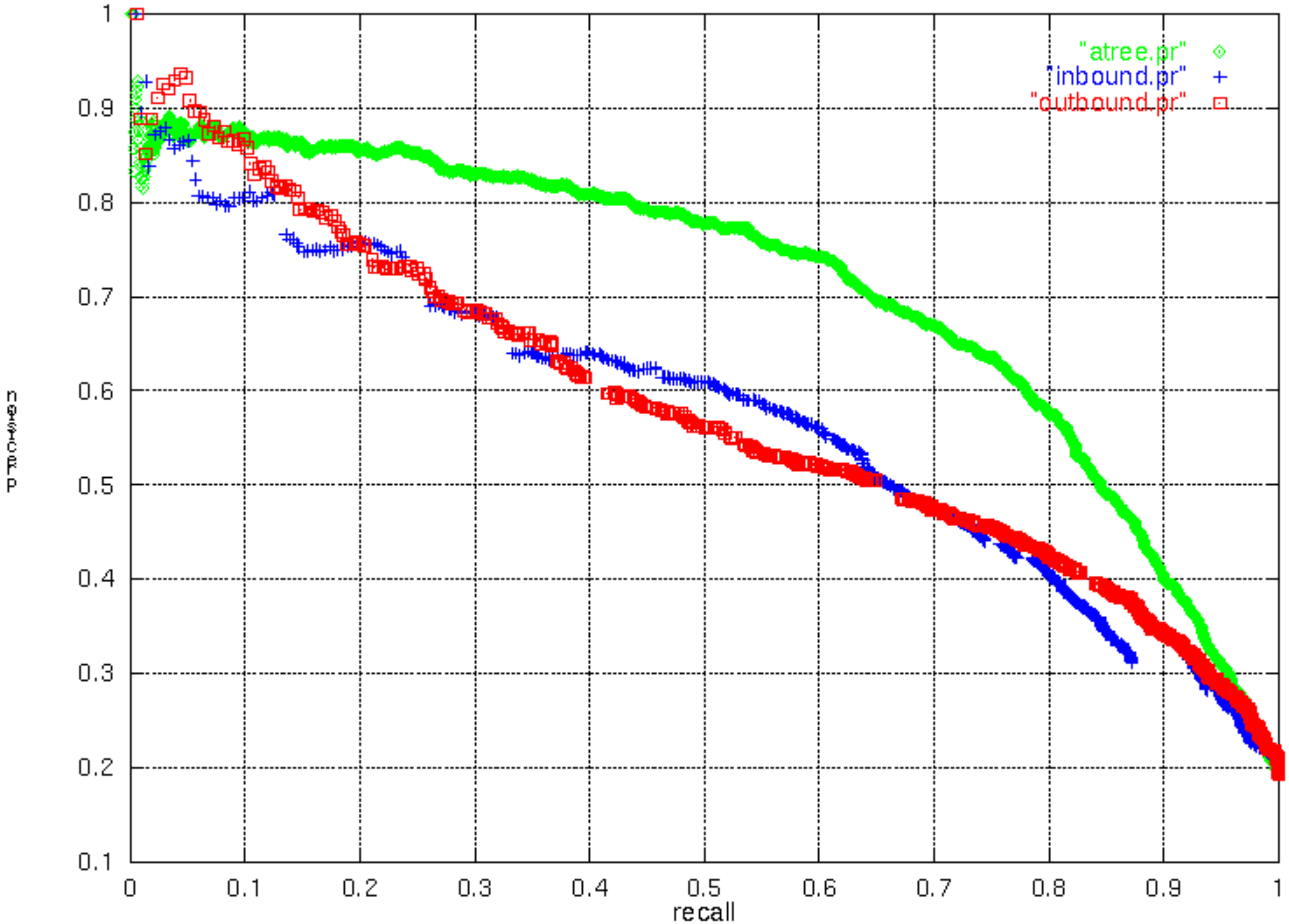


Alternating Tree (Detail)



Precision/recall graphs

Accuracy



Score

Business impact

- Increased coverage from 44% to 56%
- Accuracy ~94%
- Saved AT&T 15M$ in the year 2000 in operations costs and missed opportunities.

Summary

- Boosting is a computational method for learning accurate classifiers
- Resistance to over-fit explained by margins
- Underlying explanation –
  large "neighborhoods" of good classifiers
- Boosting has been applied successfully to a variety of classification problems

Come talk with me!

- Yfreund@banter.com
- http://www.cs.huji.ac.il/~yoavf