Pattern Classification in the Presence of Class Imbalance

James R. Green

Systems and Computer Engineering

Carleton University

Recall… Case study: PIPE II

- The challenge:
  - Yeast has 6200 proteins in its proteome.
  - Every possible pair of yeast proteins could potentially interact.
  - Based on biological evidence, it is believed that approx 50K interactions exist in yeast.
  - Would like to computationally predict from sequence alone whether a given pair will interact.
  - It is very expensive to verify a prediction experimentally.
- The solution:
  - We have developed a classifier which tests a given pair of protein sequences and predict whether they will interact *in vivo*.
  - We have reduced the computational complexity to the point where we can run it on all 18million pairs.
  - Through parameter tuning, we can achieve either:
    - 1) High specificity of 99% with medium sensitivity (%50)
    - 2) Very high specificity of 99.9% at the cost of a low sensitivity (25%)
- The $1M questions:
    - **Which parameter set is preferred?**
    - **How many of the predicted interactions are likely to be true interactions?**

The Effect of Class Imbalance

Case 1

		Actual Class	
		A (+)	B (-)
Predicted Class	A (+)	25K	180K
	B (-)	25K	17.8M

Sn=50%
Sp=99%

The Effect of Class Imbalance

Case 1

		Actual Class	
		A (+)	B (-)
Predicted Class	A (+)	25K	180K
	B (-)	25K	17.8M

Sn=50%
Sp=99%

Case 2

		Actual Class	
		A (+)	B (-)
Predicted Class	A (+)	12.5K	18K
	B (-)	37.5K	18M

Sn=25%
Sp=99.9%

The Effect of Class Imbalance

Case 1

		Actual Class	
		A (+)	B (-)
Predicted Class	A (+)	25K	180K
	B (-)	25K	17.8M

Sn=50%
Sp=99%
Prec=25K/205K=12%

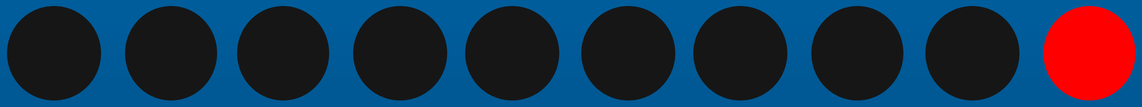
Case 2

		Actual Class	
		A (+)	B (-)
Predicted Class	A (+)	12.5K	18K
	B (-)	37.5K	18M

Sn=25%
Sp=99.9%
Prec=12.5K/30.5K=42%

Let’s Learn a Rule!

- You receive a big bag of coloured balls.
- You draw 10 balls:



- You now must guess the colour of the next 10 balls, one ball at a time.
- What colour should you guess for each?

Thought Experiment

Point #1: Don’t blame the classifier

Thought Experiment

Point #1: Don’t blame the classifier

Point #2: What if the red balls are patients with cancer?

Class Imbalance

- Many events of interest are rare:
  - ~500K interactions among ~250M human protein pairs (→ 1:500 ratio)
  - 11M pseudo-miRNA RNA hairpins; only ~2600 known miRNA in MiRBase (→ <1:4000 ratio)
  - Most biopsies are ‘normal’ (→ 1:1000? ratio)
  - Most financial transactions are legitimate (ratio?)
- A dataset is **imbalanced** if the classification categories are not approximately equally represented

Class Imbalance

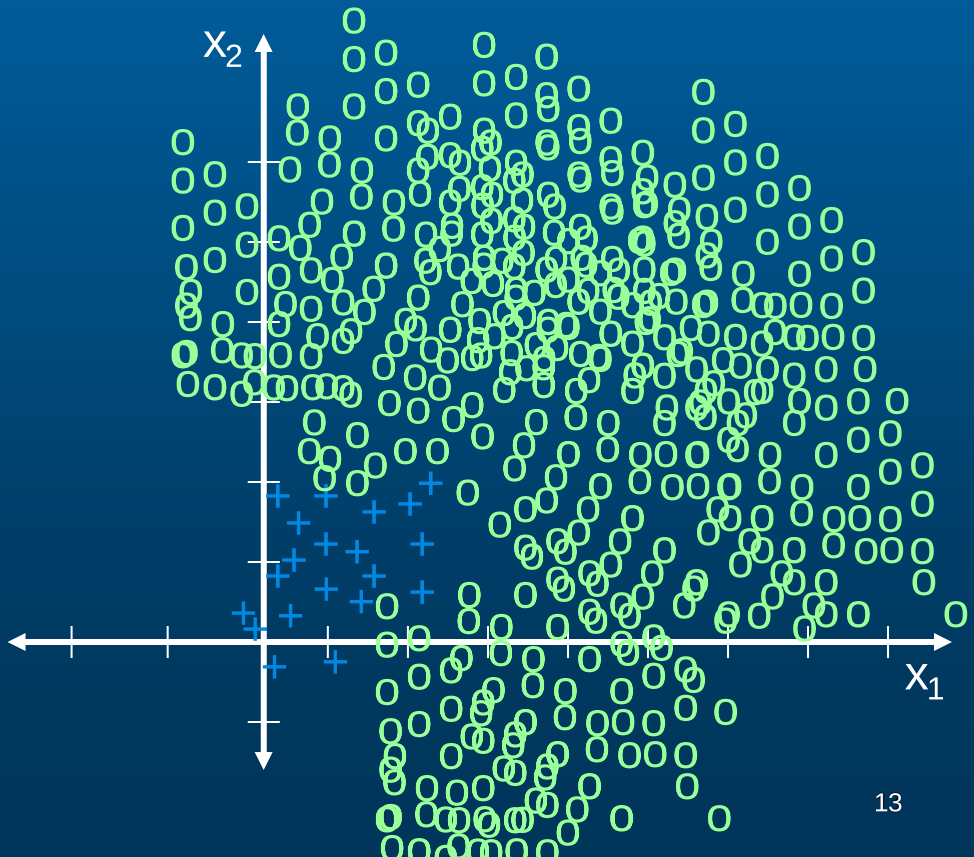
- 2 Problems:
  - 1) Classifiers tend to always predict dominant class & ignore rare class
    - Often we are most interested in the rare class!
  - 2) The rare class should only be predicted rarely!
    - Over-predicting rare class can lead to a useless classifier
- Solution:
  - Must consider TPR-FPR tradeoff…
  - Must be addressed during both <u>training</u> & <u>evaluation</u> of predictor

Class Imbalance During Training

- Avoiding problem 1 *(ignoring rare class)*:
  - Undersample dominant class
  - Oversample rare class
  - Weight errors on each class
  - Collect more data!
    - Active learning… later…
- Avoiding problem 2 *(over-predicting rare class)*:
  - Bayesian approach
  - Train secondary classifier

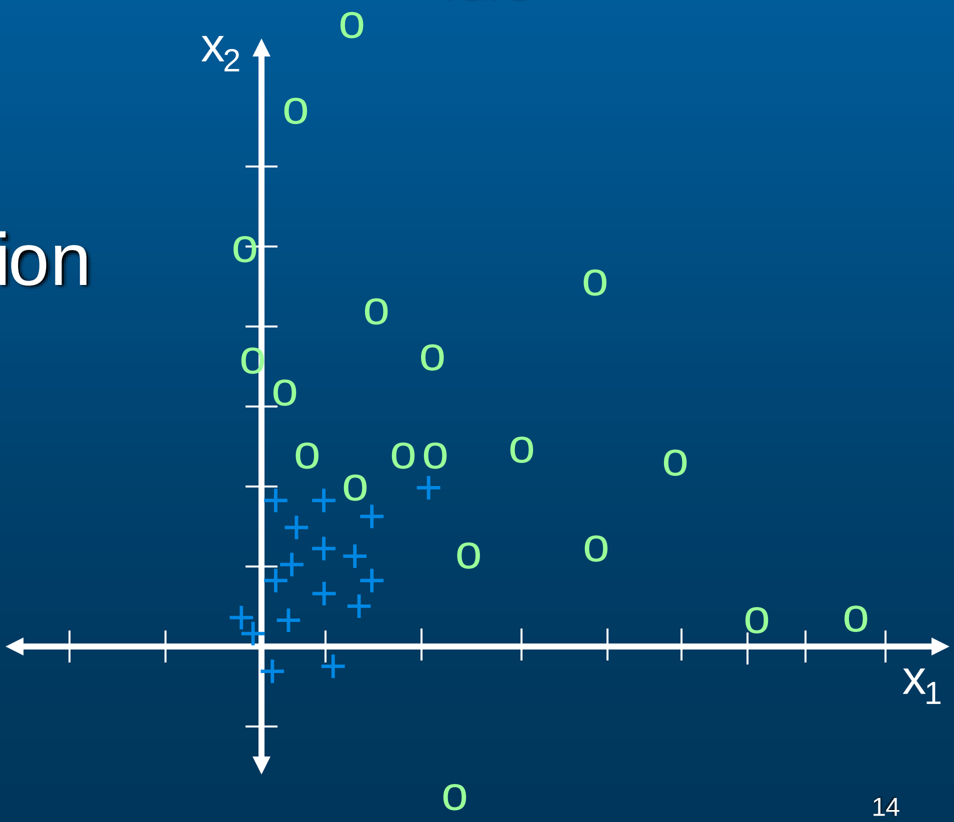
Undersample Majority Class

- Goal: achieve class balance in training data
- Options:
  - Randomly select subset of size N_{rare} from majority class



Undersample Majority Class

- Goal: achieve class balance in training data
- Options:
  - Randomly select subset of size N_{rare} from majority class
  - Pro: simple
  - Con: losing information

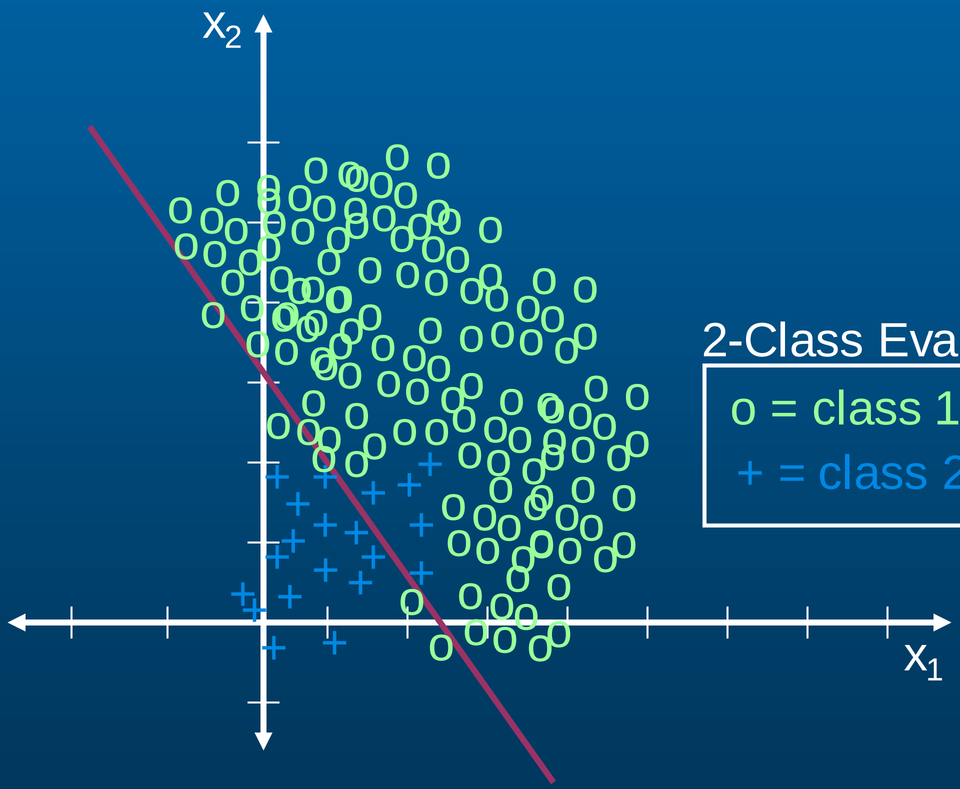


Oversample Rare Class

- Goal: achieve class balance in training data
- Options:
  - Include repeated copies of rare instances
    - Effect?
  - Generate synthetic data
    - Interpolate or add noise to rare samples
    - E.g. Synthetic Minority Oversampling TEchnique (SMOTE)

Weight Errors Differentially

- Goal: force classifier to pay more attention to one class



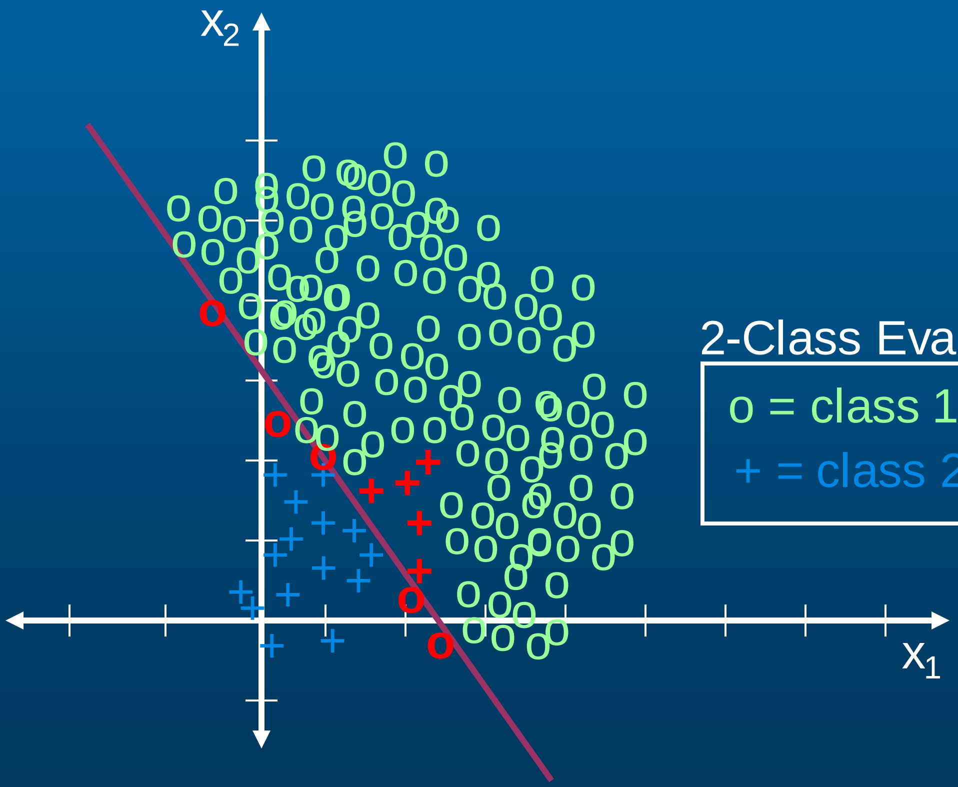
2-Class Evaluation Data Set

o = class 1 (80 samples)

+ = class 2 (20 samples)

Weight Errors Differentially

- Goal: force classifier to pay more attention to one class



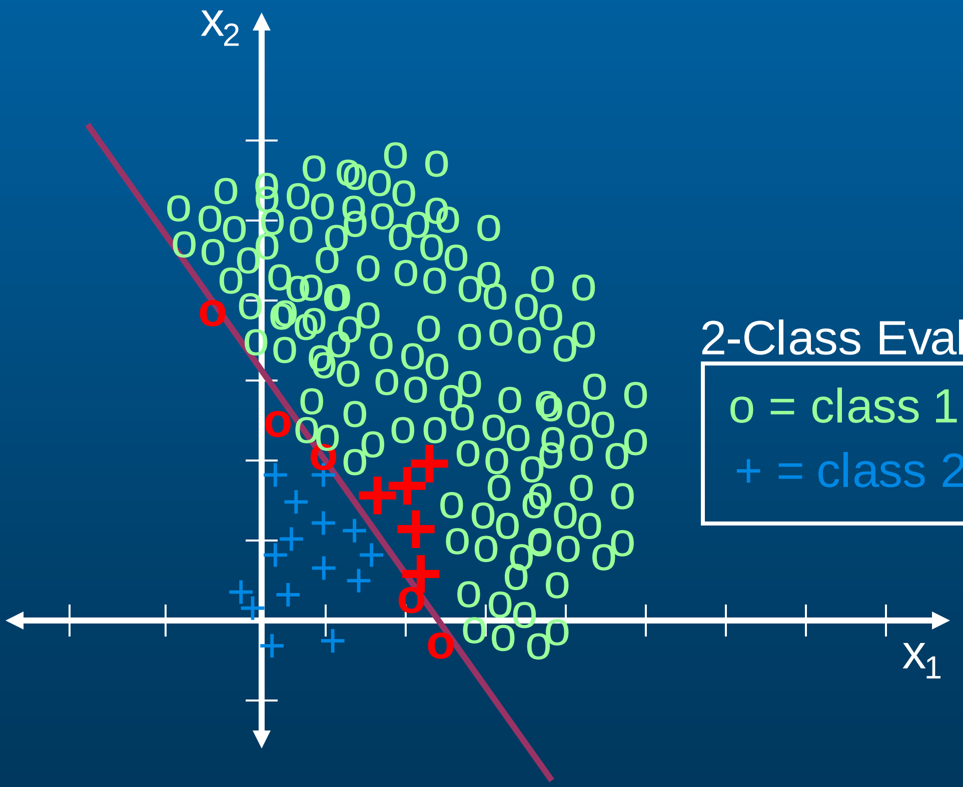
2-Class Evaluation Data Set

o = class 1 (80 samples)
+ = class 2 (20 samples)

Unweighted fitness = correct classification rate = 90/100 = 0.9

Weight Errors Differentially

- Goal: force classifier to pay more attention to one class



2-Class Evaluation Data Set

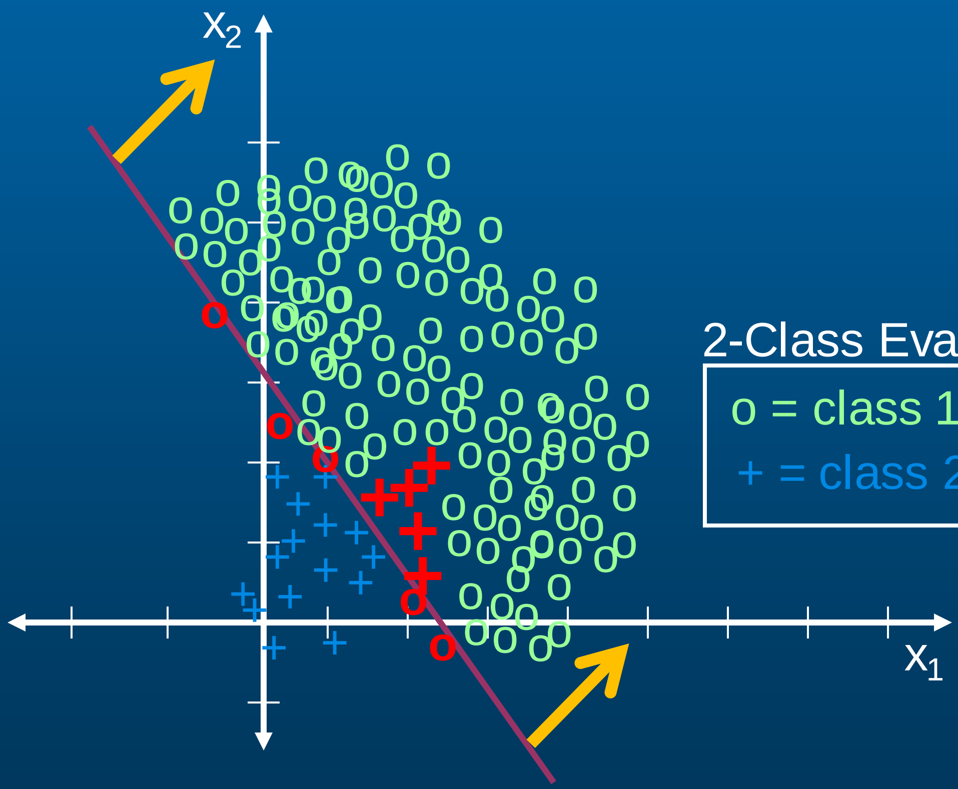
o = class 1 (80 samples)

+ = class 2 (20 samples)

Weighted fitness << 0.9

Weight Errors Differentially

- Goal: force classifier to pay more attention to one class



2-Class Evaluation Data Set

o = class 1 (80 samples)
+ = class 2 (20 samples)

Weighted fitness increases

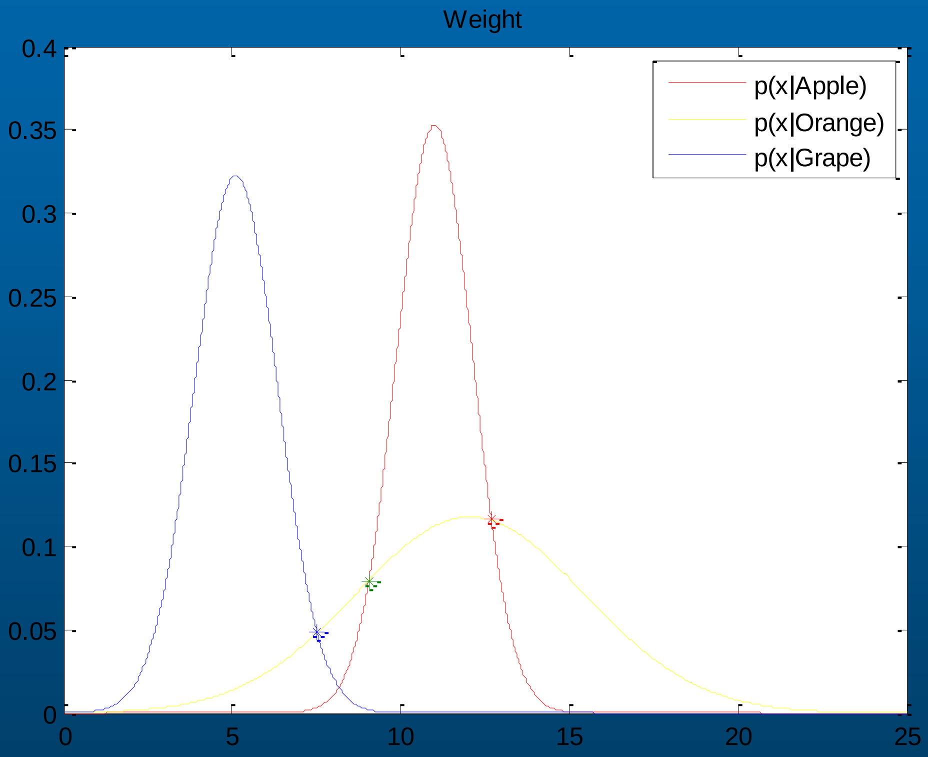
Bayesian Approach

- Methods above emphasize rare class
  - Pro: rare class is not ignored
  - Con: rare class may be over-predicted (FP problem)

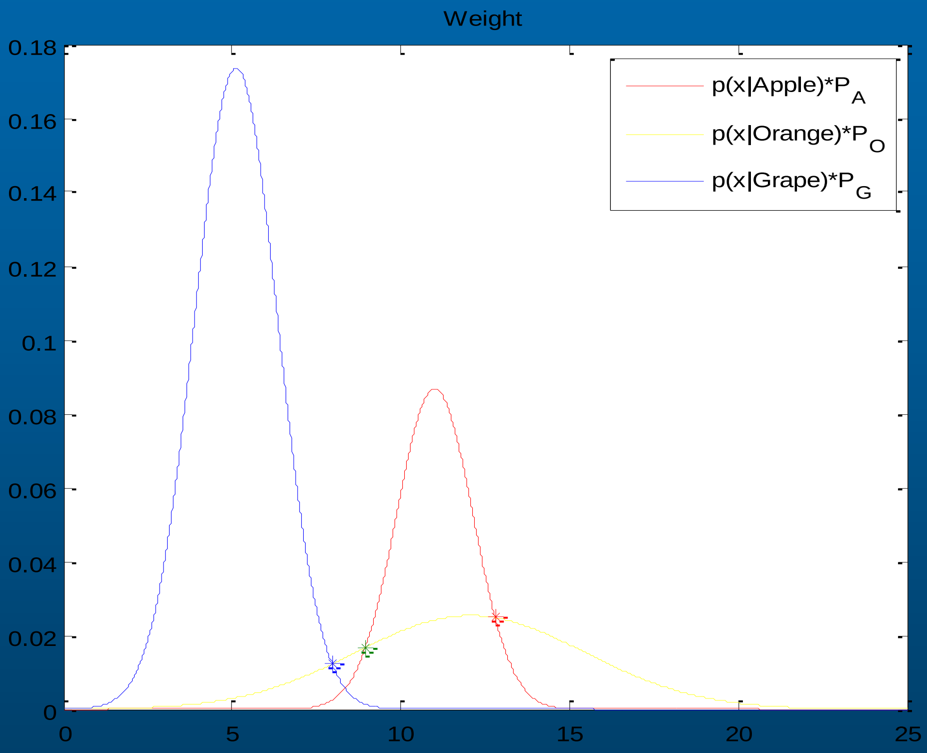
- If ‘balanced’ dataset used to get classifier score, consider these to be p(score|ω_1) & p(score|ω_2)
- Can now apply “prior” belief based on prevalence of each class to get posterior probability

$$P(\omega_j \mid x) = \alpha \left[p(score \mid \omega_j) \cdot P(\omega_j) \right]$$

Bayesian Approach

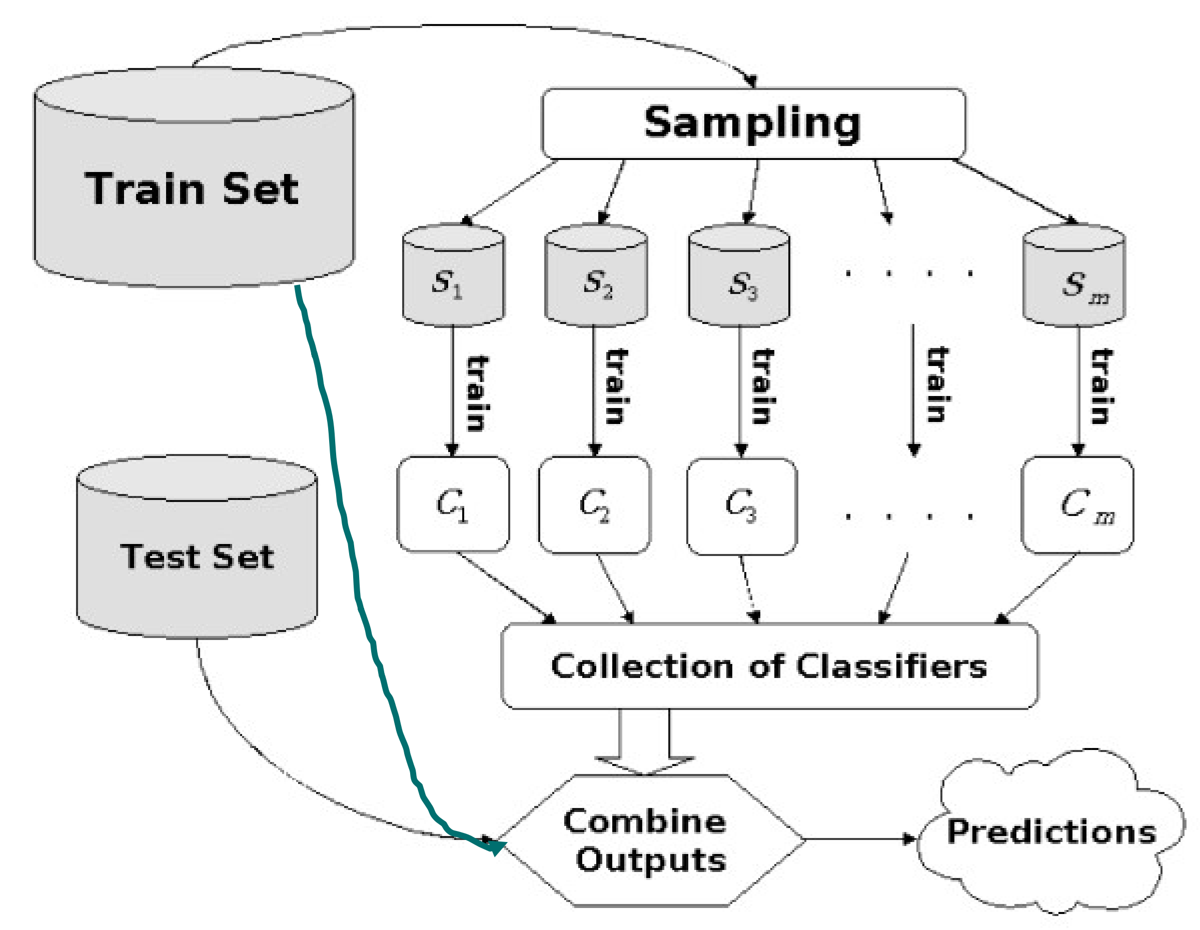


No Priors



With Priors

Train Secondary Classifier



Caragea *et al. BMC Bioinformatics* 2007 **8**:438

“Balanced” datasets used for training component classifiers

“Natural” distribution datasets used for training ensemble logic and for evaluation

Avoid over-prediction of rare class

Class Imbalance During Testing

- Under-prediction problem now solved.

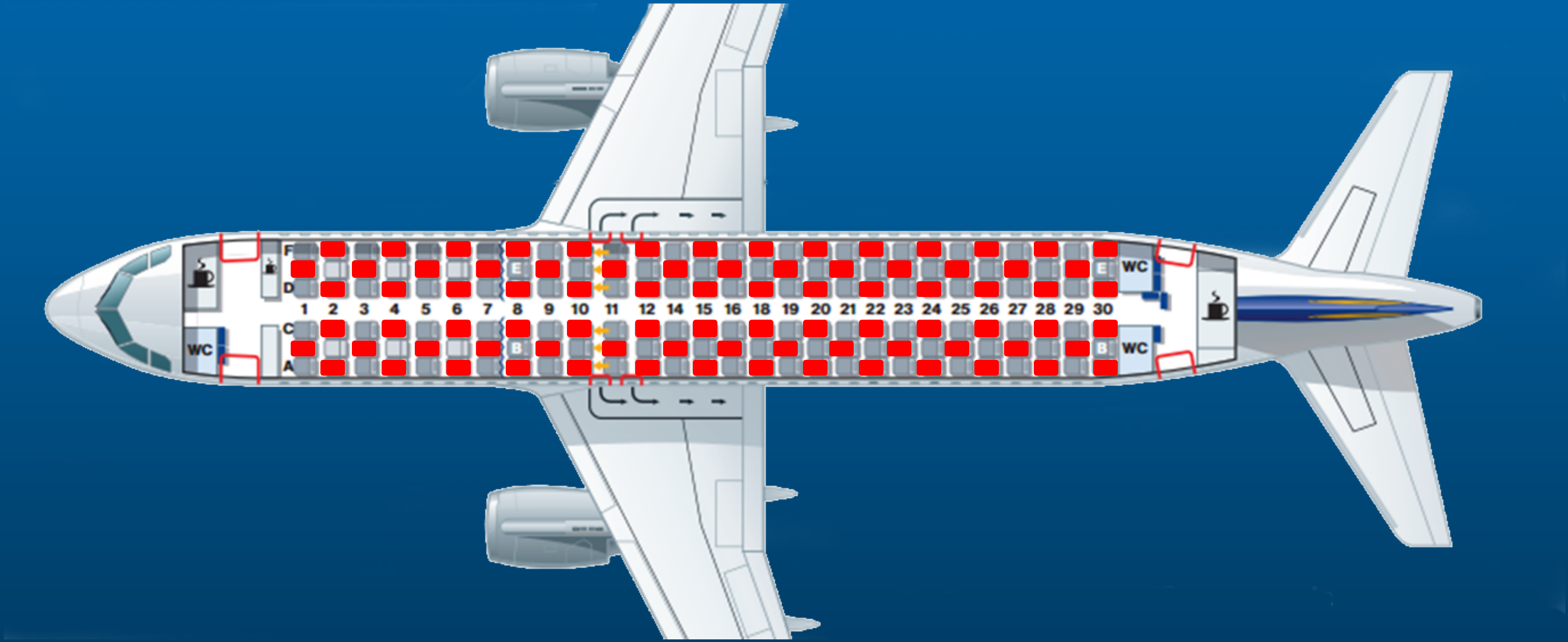
- What about testing?
  - 2 pitfalls to avoid:

    1) Using inappropriate test data

    2) Using inappropriate performance metrics

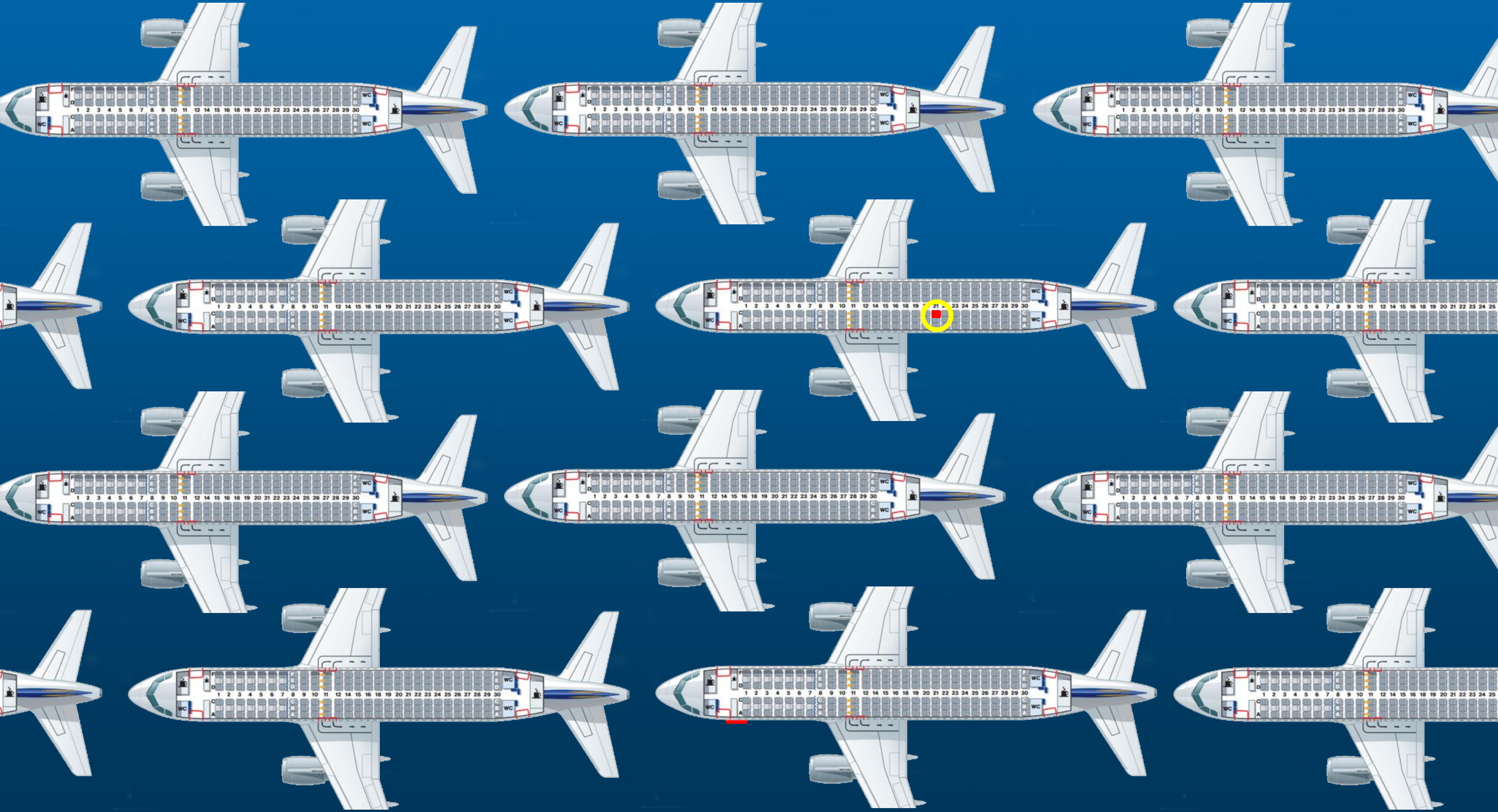
Screening Travellers for Quarantine

Given test data:



Screening Travellers for Quarantine

Actual deployment:

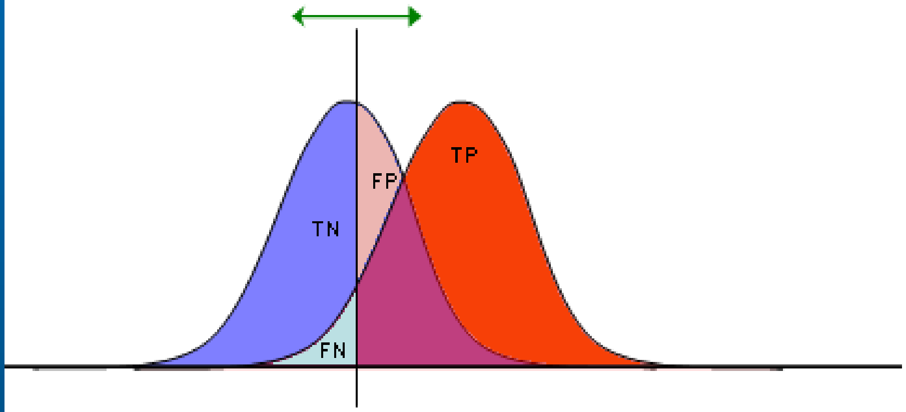


Class Imbalance During Testing

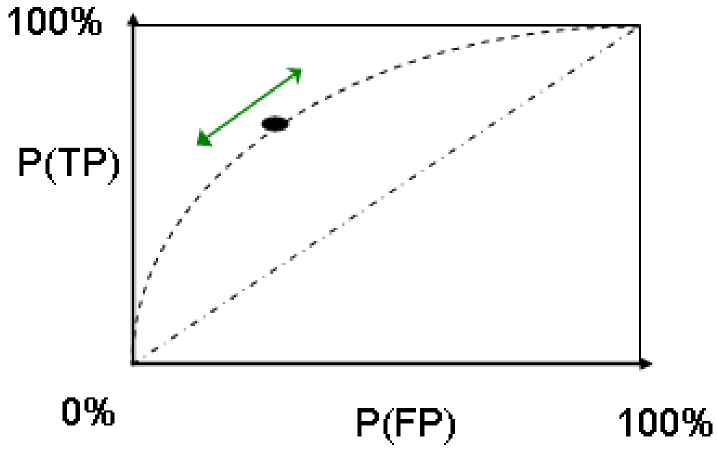
- It’s ok to create “balanced” training sets
  - But test data should reflect future data
    - (*natural* prevalence/ratio)
- Use appropriate performance metrics
  - Not CCR, ROC, AUC
- Precision
  - Precision-recall curves
  - Effect of prevalence

ROC Curves

Tunable decision threshold



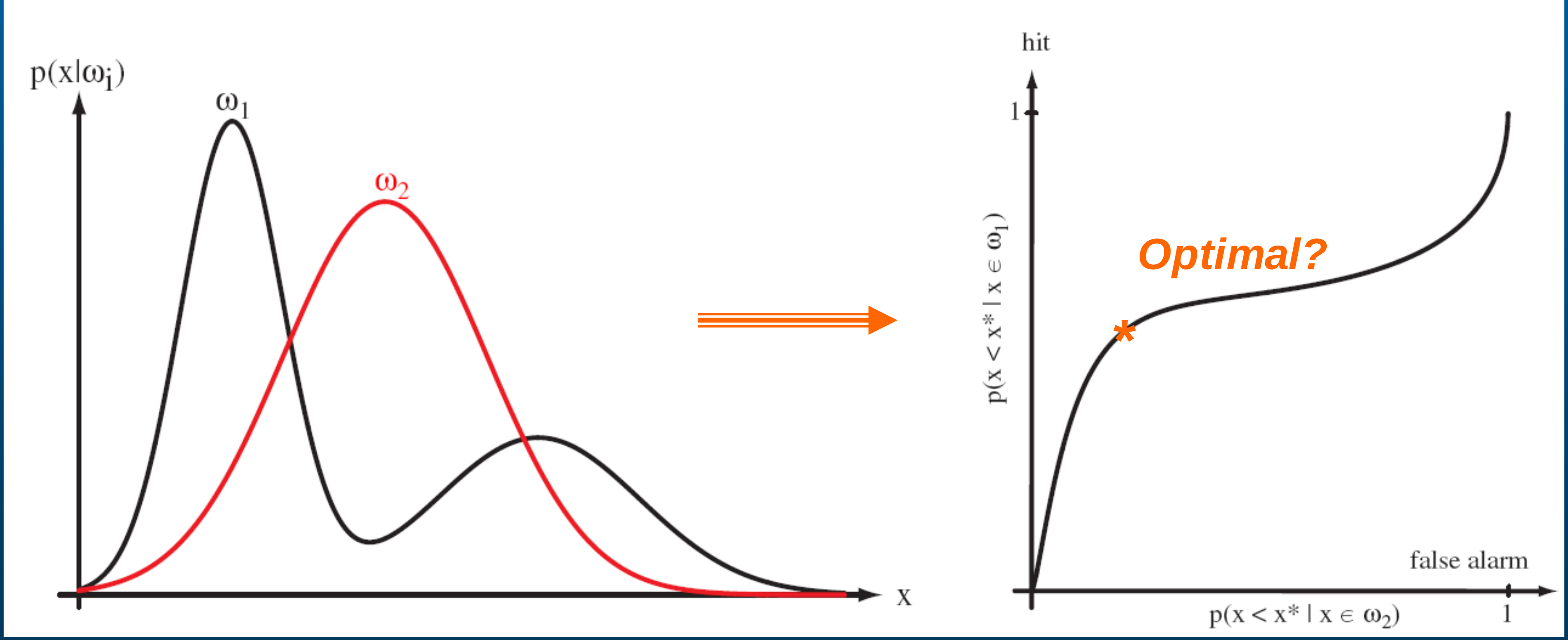
TP	FP
FN	TN
1	1



See http://www.anaesthetist.com/mnm/stats/roc/Findex.htm for a great ROC demo

ROC Curve

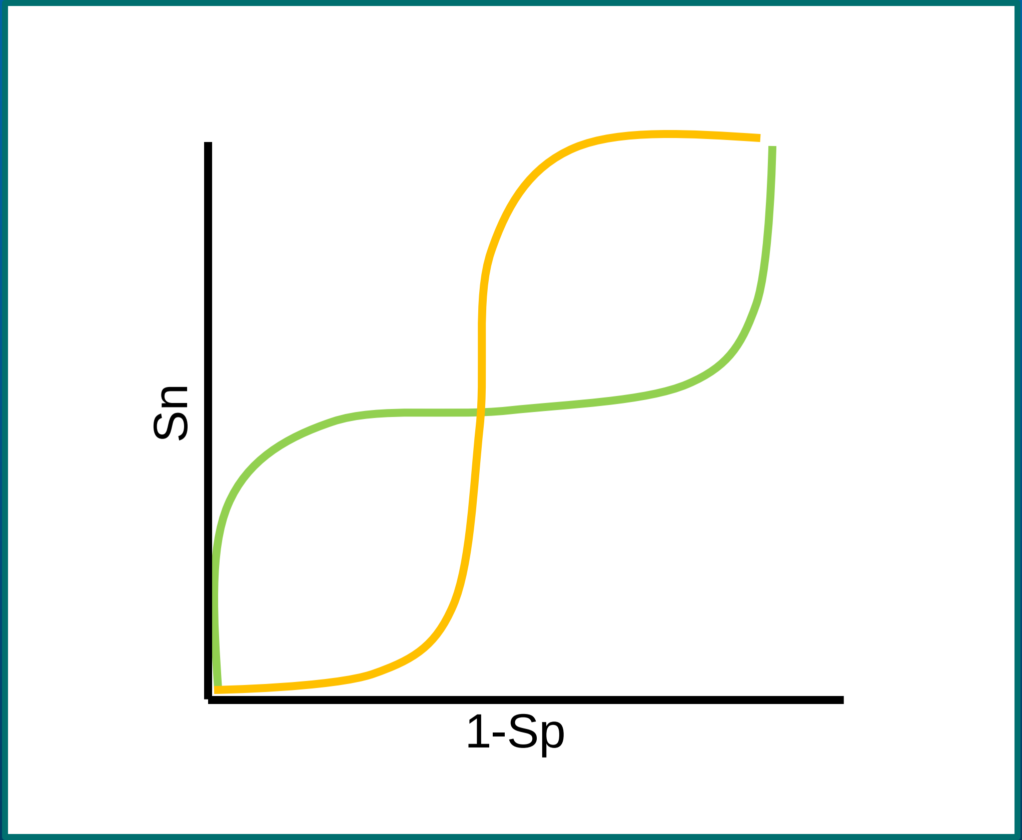
- Curve is not necessarily symmetric
- Can be informative in setting threshold to balance benefit of TP against cost of FP



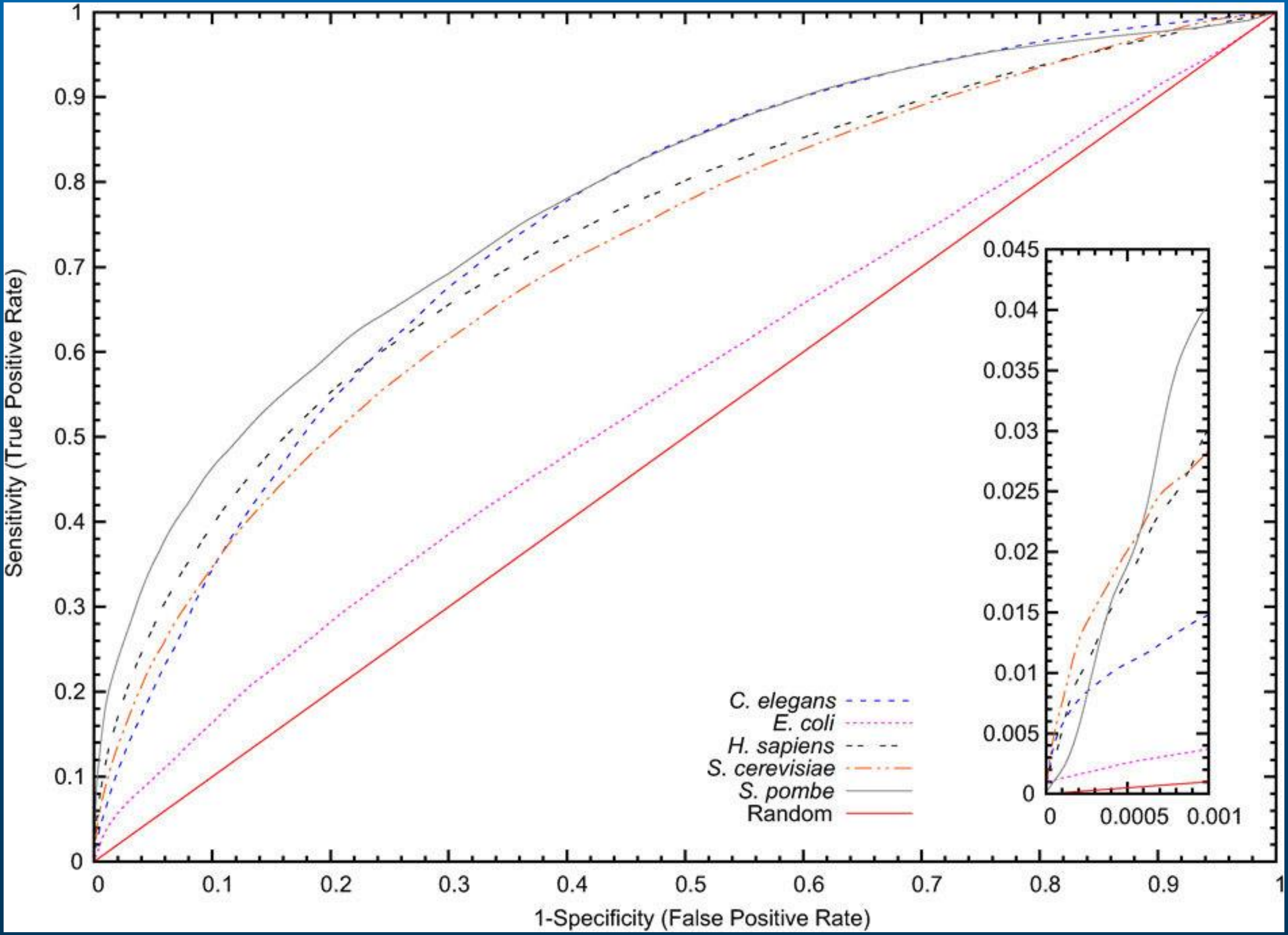
Area under the ROC Curve

- Area under an ROC curve (AUC) summarizes performance of a classifier
  - Independent of particular cost function which might influence threshold placement
  - Ranges from 1 (perfect) to 0 (worst)
    - Random = 0.5
  - BUT, AUC is just one facet of classifier performance. May not be the most important one
    - E.g. PIPE must perform at one extreme end of the curve…

2 ROC Curves with Same AUC

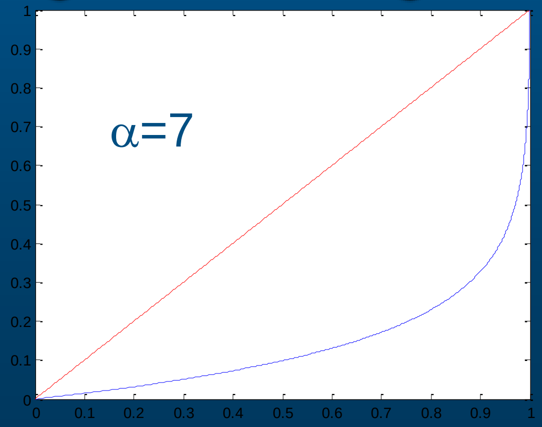


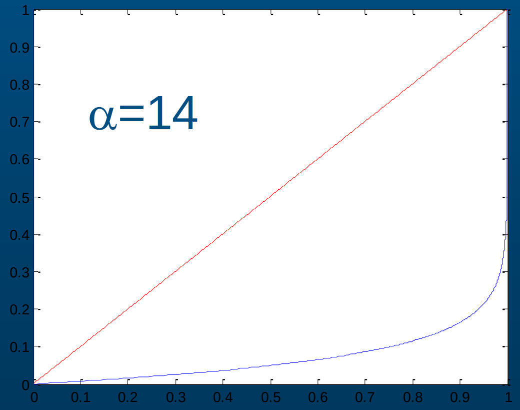
PIPE ROC Curve



The CROC Curve

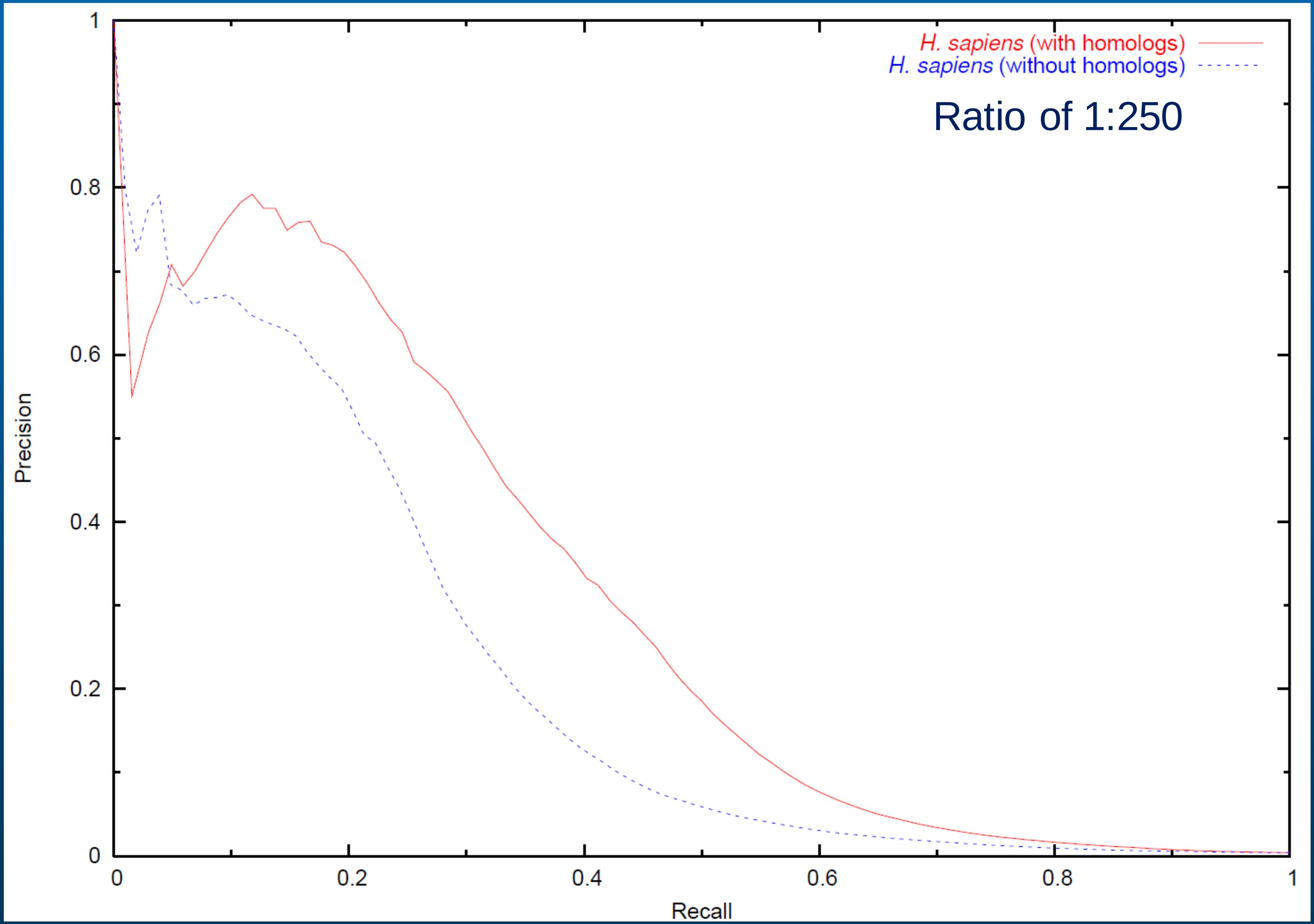
- Compress x-axis (FPR)
  - Accentuate performance in high Sp region
  - AUC more meaningful
  - Plot Sn vs. f(FPR): $(f(x) = (1 - e^{(-\alpha x)})/(1 - e^{(-\alpha)}))$
    - Adjust α (α=7 → Sp=90% → f[0.1]=0.5)
    - Analogous to using a log scale on FPR axis



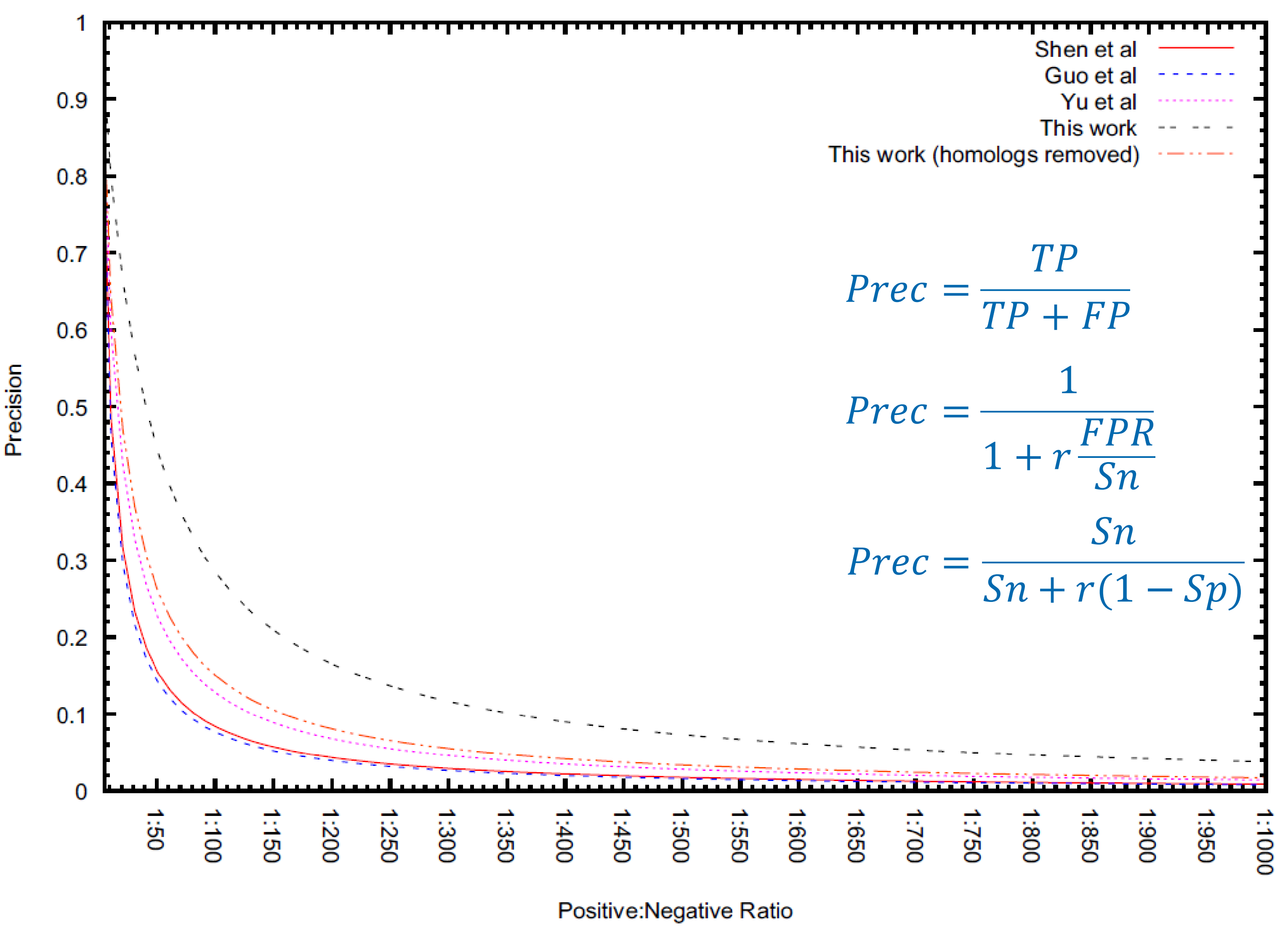


Swamidass SJ, Azencott C-A, Daily K, Baldi P: **A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval.** *Bioinformatics* 2010, **26**:1348–56.

Precision-Recall Curves



Precision vs. Prevalence

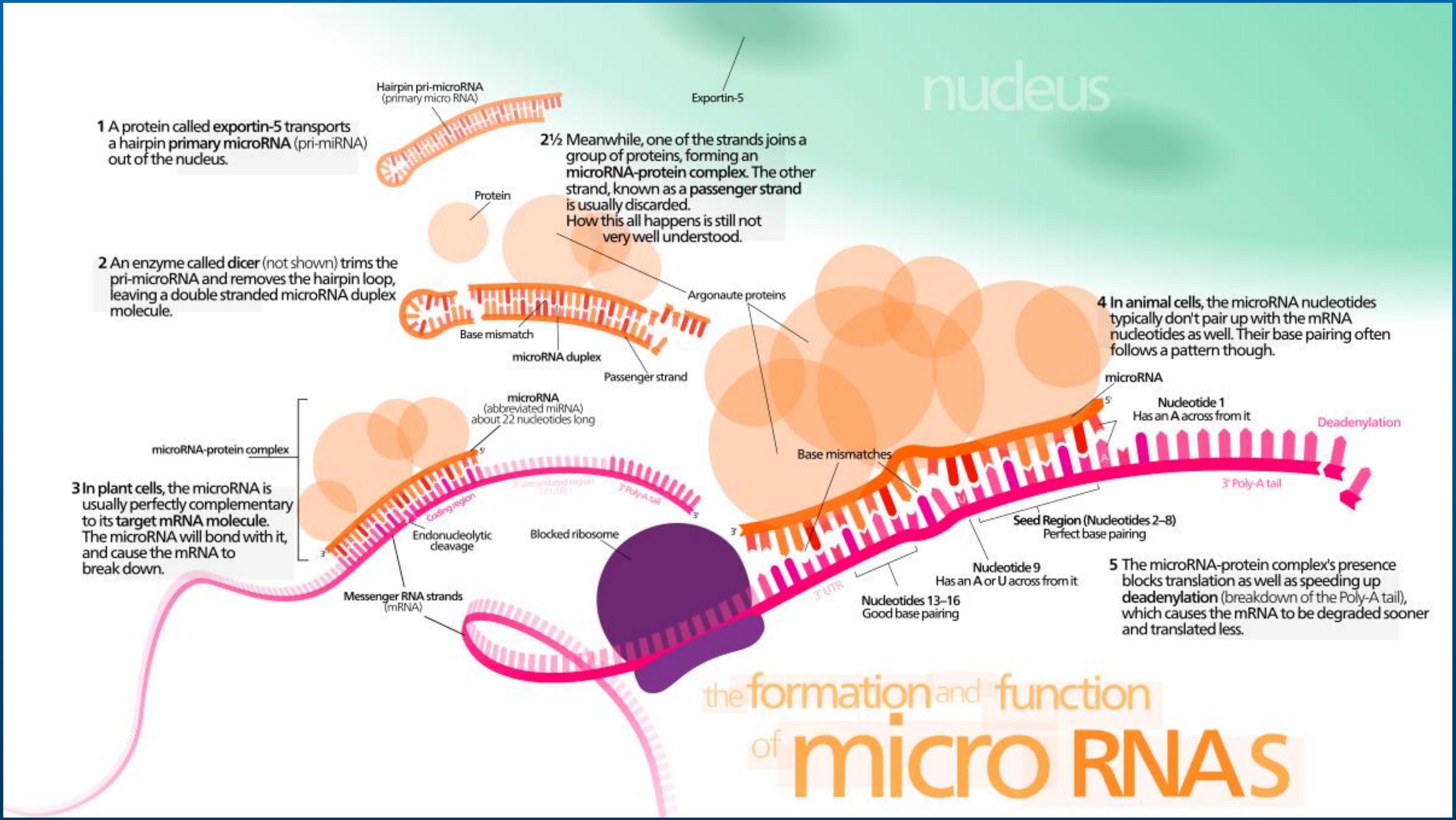


$$Prec = \frac{TP}{TP + FP}$$

$$Prec = \frac{1}{1 + r\frac{FPR}{Sn}}$$

$$Prec = \frac{Sn}{Sn + r(1 - Sp)}$$

miRNA



1 A protein called **exportin-5** transports a hairpin **primary microRNA** (pri-miRNA) out of the nucleus.

2 An enzyme called **dicer** (not shown) trims the pri-microRNA and removes the hairpin loop, leaving a double stranded microRNA duplex molecule.

2½ Meanwhile, one of the strands joins a group of proteins, forming an **microRNA-protein complex.** The other strand, known as a **passenger strand** is usually discarded. How this all happens is still not very well understood.

3 **In plant cells,** the microRNA is usually perfectly complementary to its **target mRNA molecule.** The microRNA will bond with it, and cause the mRNA to break down.

4 **In animal cells,** the microRNA nucleotides typically don't pair up with the mRNA nucleotides as well. Their base pairing often follows a pattern though.

5 The microRNA-protein complex's presence blocks translation as well as speeding up **deadenylation** (breakdown of the Poly-A tail), which causes the mRNA to be degraded sooner and translated less.

the formation and function of micro RNAs

Wikipedia Commons

miRNA Prediction

- microPred most widely used miRNA prediction tool
  - Trained on human known miRNAs
  - Uses 21 features, 5 of which relate to secondary structure free energy
    - Problem?
  - Accuracy evaluated using geometric mean
    - What are they failing to account for?
  - Tested on other species, sensitivity maintained
    - What is missing?

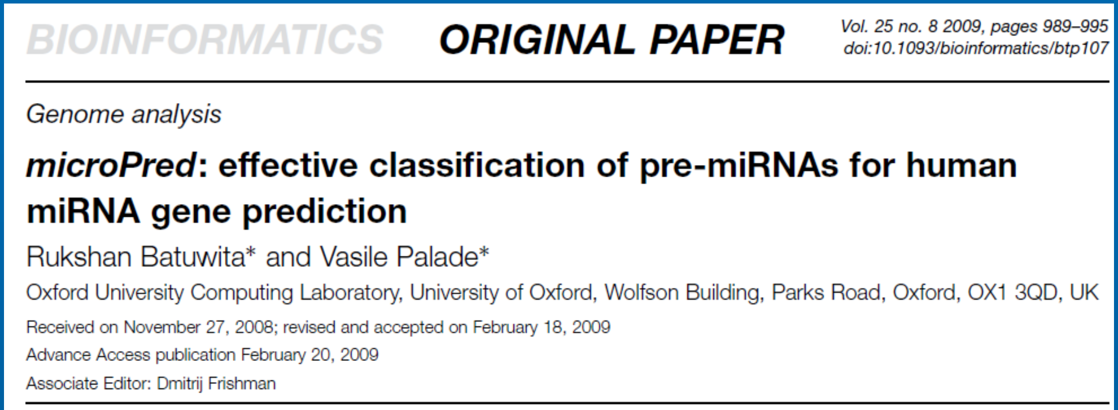
Effect of Class Imbalance

- Batuwida & Palade could achieve either:

	Sn	Sp	G-mean
Approach A	83.36%	99.0%	90.84%
Approach B	90.02%	97.28%	**93.58%**

- However considering class imbalance of 1000 negatives per positive:

	Sn	Sp	G-mean	Precision
Approach A	83.36%	99.0%	90.84%	7.6%
Approach B	90.02%	97.28%	**93.58%**	**3.2%**



BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 8 2009, pages 989–995
doi:10.1093/bioinformatics/btp107

Genome analysis

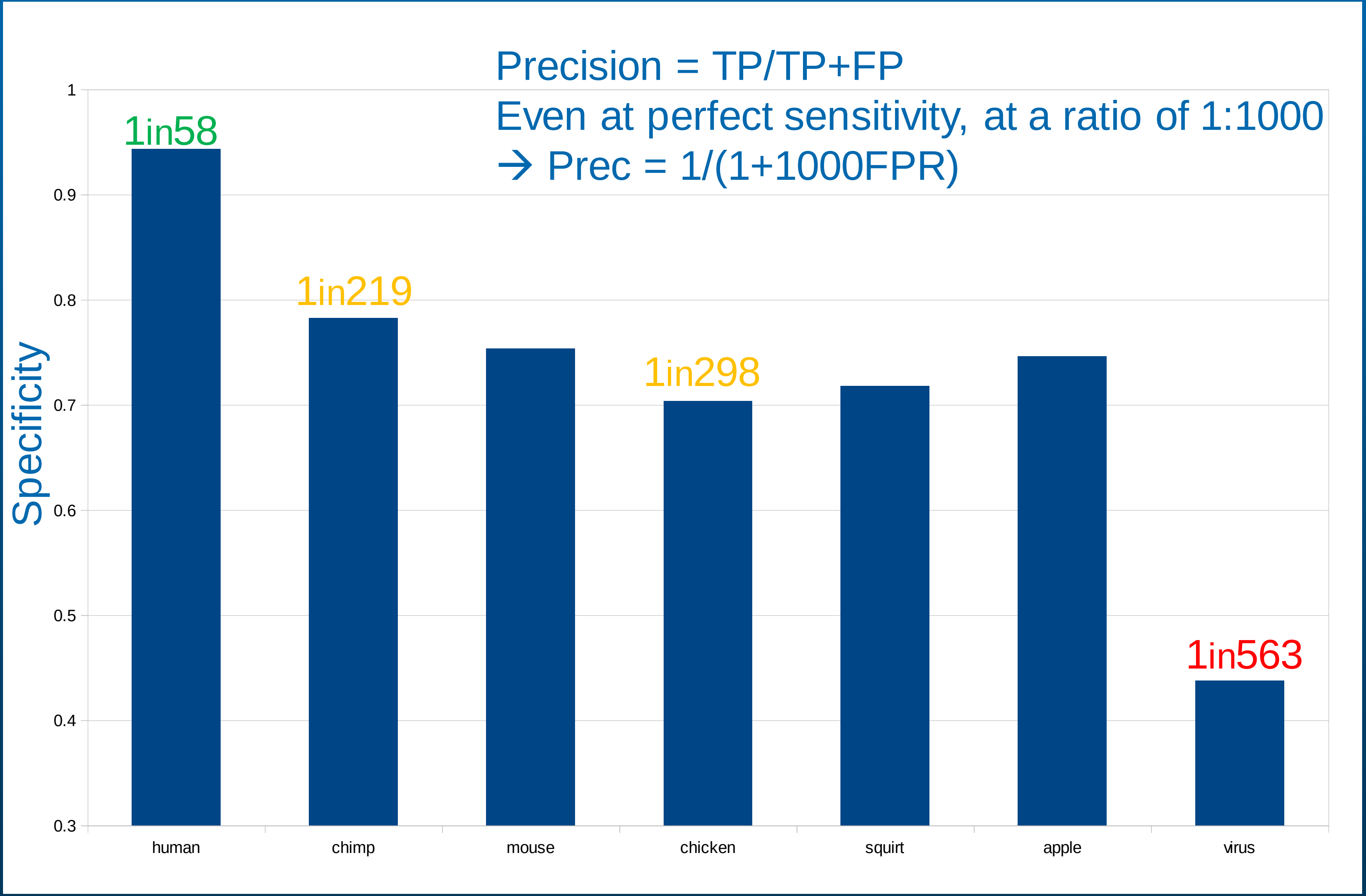
microPred: effective classification of pre-miRNAs for human miRNA gene prediction

Rukshan Batuwita* and Vasile Palade*
Oxford University Computing Laboratory, University of Oxford, Wolfson Building, Parks Road, Oxford, OX1 3QD, UK

Received on November 27, 2008; revised and accepted on February 18, 2009
Advance Access publication February 20, 2009
Associate Editor: Dmitrij Frishman

- “We validated the *microPred* predictions on the other animal (non-human) and viral pre-miRNAs published in the *miRBase12*, and obtained a high sensitivity. Out of 6095 other animal pre-miRNAs across 49 species, *microPred* identified 5651 correctly with 92.71% of recognition rate. Out of 139 viral pre-miRNAs across 12 species, 131 were predicted correctly with 92.24% of recognition rate.”

Specificity for non-human species



Precision = TP/TP+FP

Even at perfect sensitivity, at a ratio of 1:1000

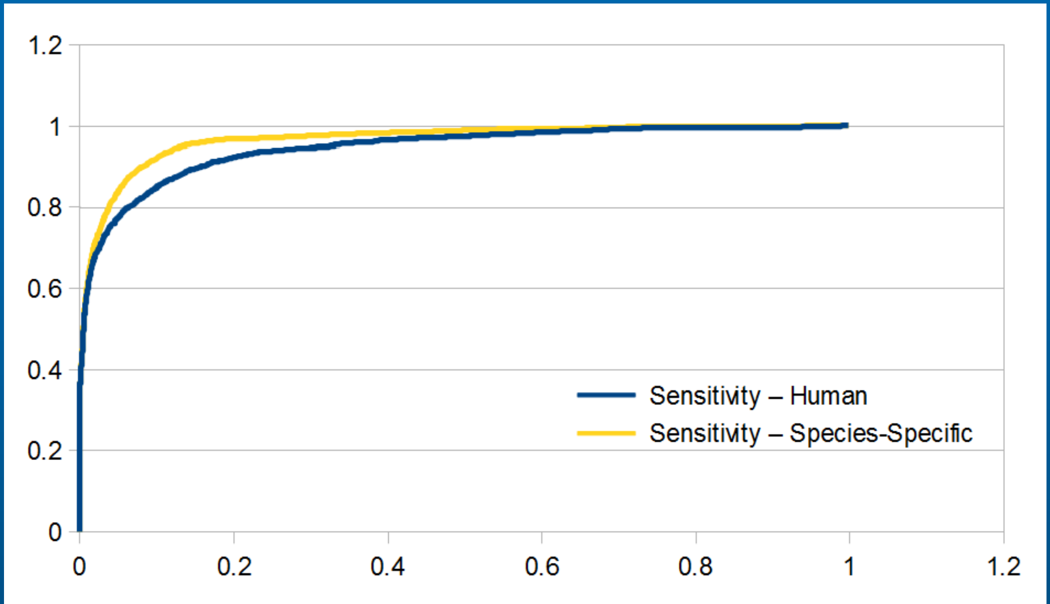
→ Prec = 1/(1+1000FPR)

Our miRNA Prediction Approach

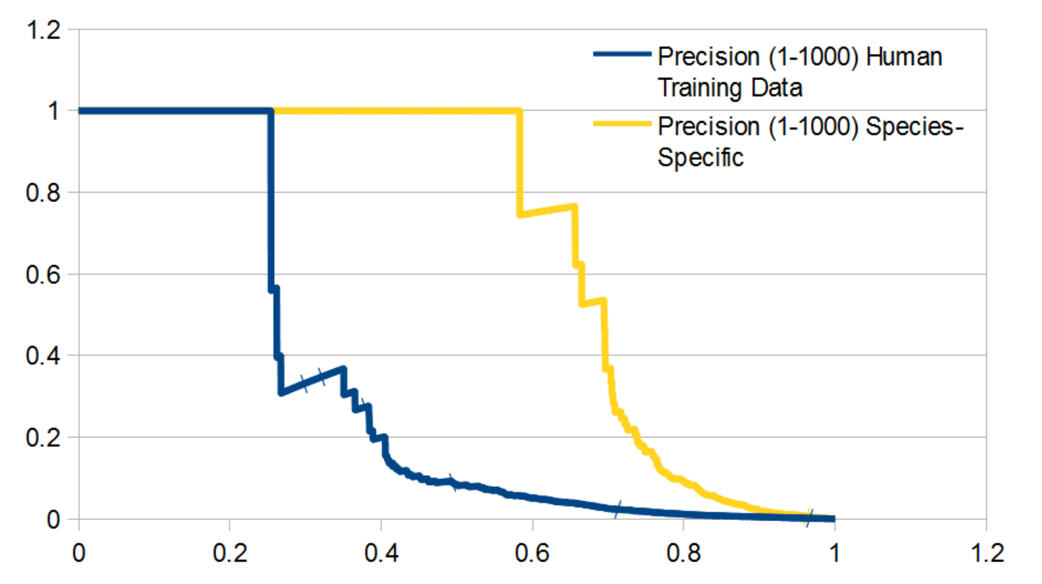
1. Cluster known miRNA from all species
2. Select largest N clusters
3. From each cluster, select representative closest to the target species → +ve training
   - Use SMOTE to generate synthetic minority data
   - Avoid redundant features
4. Get -ve training data from “related” species
   - Hairpin regions of known coding regions
5. Apply *leave-one-species-out* testing
6. Measure performance using precision-recall
   - Prevalence-corrected precision (1000:1 ratio)

Train: Xenopus Tropicalis
Test: Anolis Carolinensis

ROC



Prec-Recall



Summary

- Many problems of interest have class imbalance
- Must consider prevalence during both training and testing to avoid the pitfalls:

1) Completely ignoring minority class

2) Over-predicting minority class

3) Testing using unrealistic data

4) Using inappropriate performance metrics

Acknowledgments

1. Funding: NSERC, CFI, ORF, MITACS, Carleton U
2. Graduate students
    1. microRNA prediction: Robert Peace
    2. PIPE: Sylvain Pitre, Catalin Patulea, Andrew Schoenrock, Adam Amos-Binks, Allen Amos-Binks, Brad Barnes, Kevin Dick, Chris North, several biologists!
3. Collaborators
    1. microRNA prediction: Kyle Biggar, Ken Storey
    2. PIPE: Frank Dehne, Ashkan Golshani, Alex Wong, Michel Dumontier, Kyle Biggar, several biologists!

Pattern Classification in the Presence of Class Imbalance

James R. Green

Systems and Computer Engineering

Carleton University