# ROC Curves / Profit Curves

Clayton W. Schupp, Galvanize

Spring 2015

# Confusion Matrix

| Predicted class | Actual class: p | Actual class: n |
|---|---|---|
| Y | **True Positives** | **False Positives** |
| N | **False negatives** | **True negatives** |

TP rate = TP / P  recall (hit rate)

FP rate = FP / N  (false alarm rate)

Accuracy = (TP + TN) / (P + N)

Precision = TP / (TP + FP)

Other Common Terms

Sensitivity = recall

Specificity = TN / (FP + TN)
= 1 – FP rate

# Problems with Unbalanced Classes

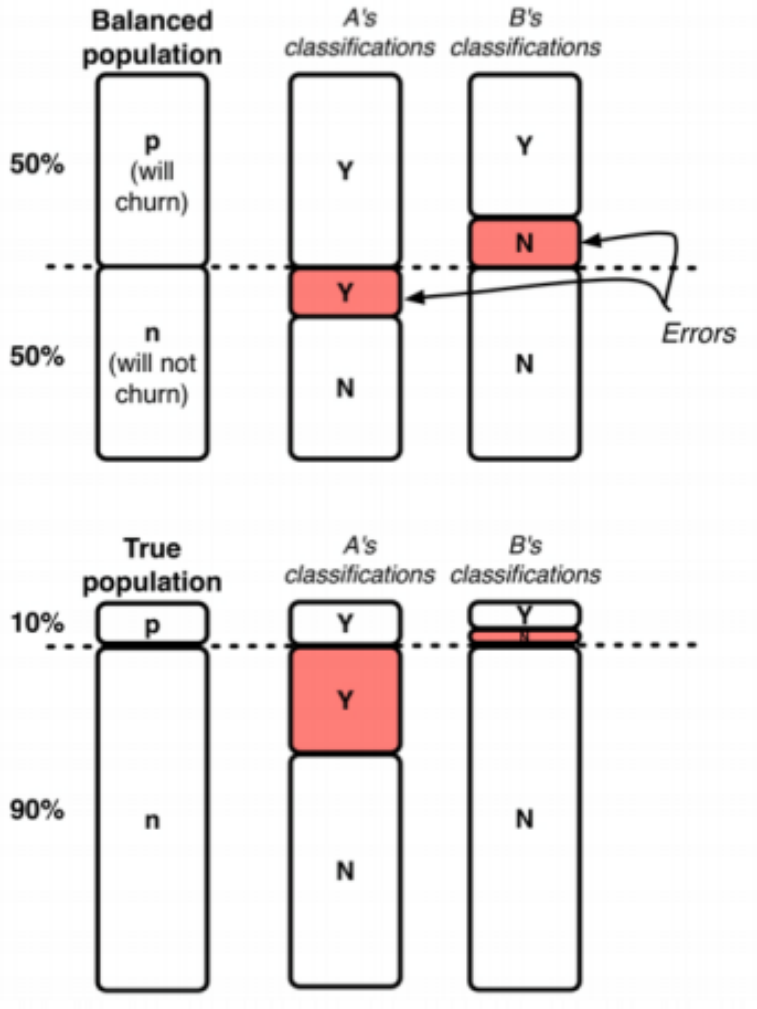
- Two different cases

*Table 7-2. Confusion matrix of A*

| | churn | not churn |
|---|---|---|
| Y | 500 | 200 |
| N | 0 | 300 |

*Table 7-3. Confusion matrix of B*

| | churn | not churn |
|---|---|---|
| Y | 300 | 0 |
| N | 200 | 500 |



# Building the ROC Curve

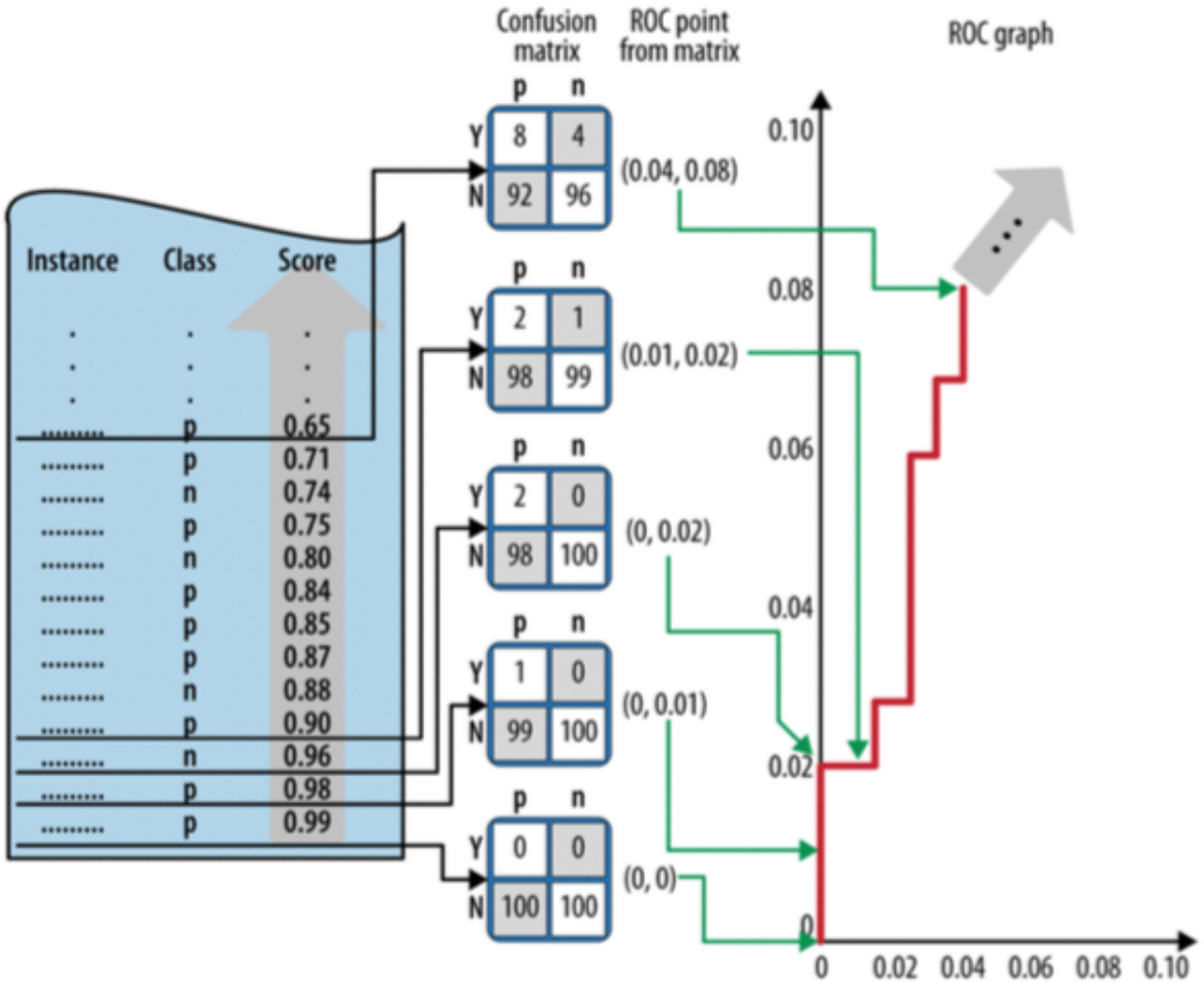
For a given model $f$, each threshold value T gives a point on the ROC Curve

Model score is the probability of class membership

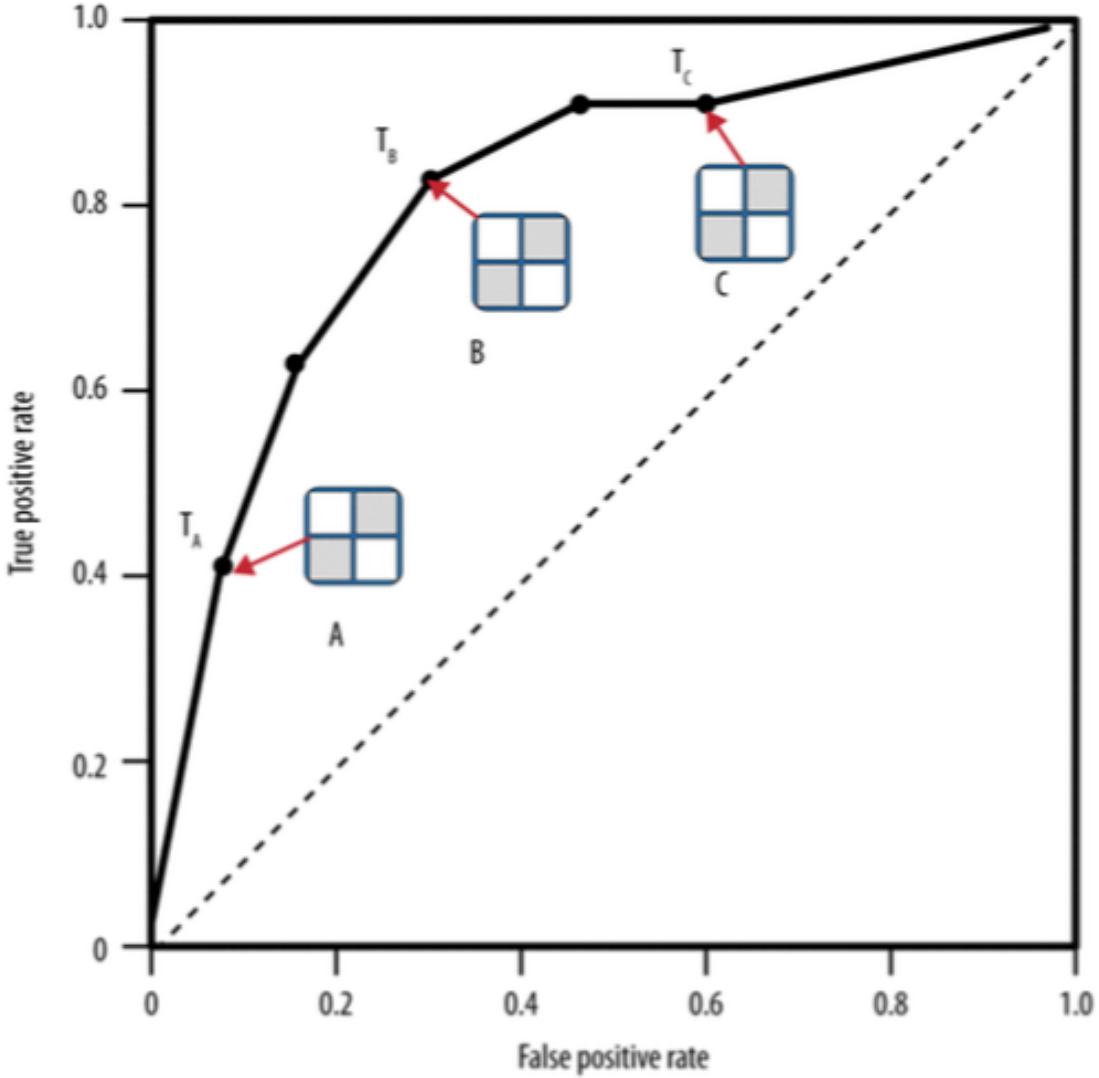
1. T = minimum score
2. TP=0, FP=0
3. For each observation, $i$:
   - If $i > T \longrightarrow$ increment TP
   - else $\longrightarrow$ increment FP
4. Add point (FP/N, TP/P) to ROC Graph

Increment T from min-score to max-score, repeating steps 1-4

# Building the ROC Curve



# Sample ROC Curve

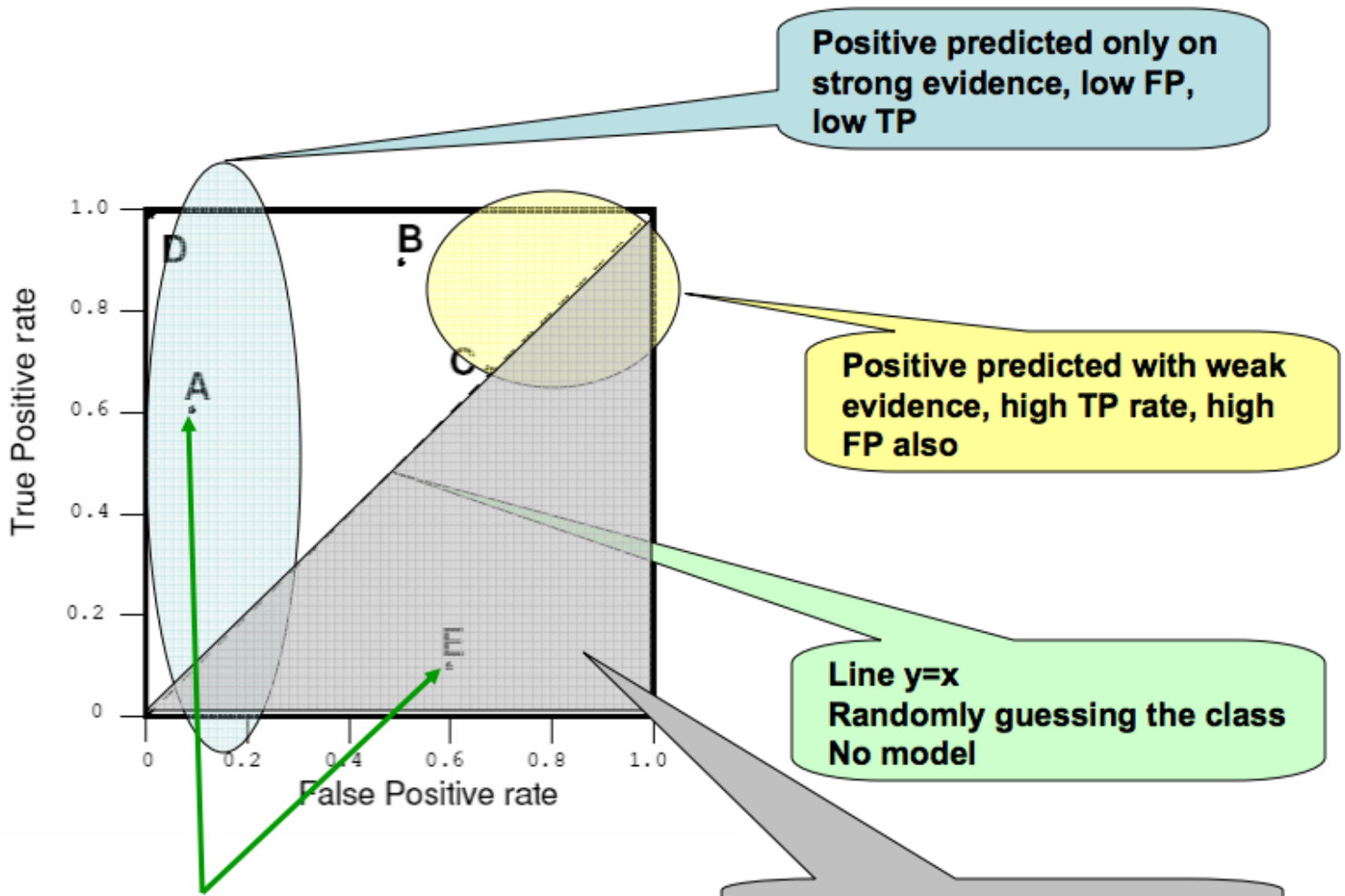


# Choosing Between Models

How do we go about choosing a model based on the ROC curve?

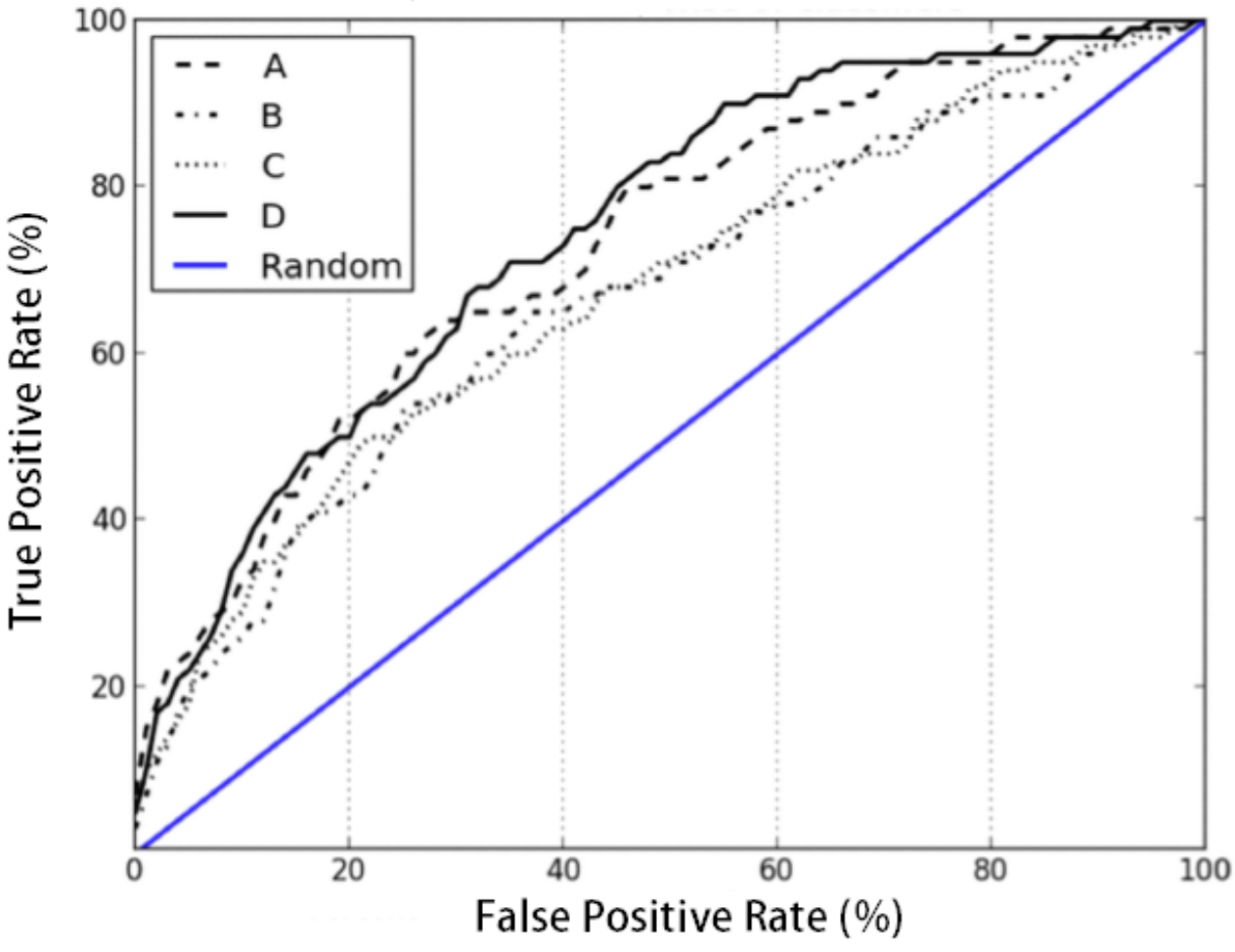
- Depends on the goal of the model
  - Screening Test: it is more important to have a high True Positive Rate regardless of the False Positive Rate $\longrightarrow$ Want to identify potential people with a disease than miss them
  - Diagnostic Test: It is more important to have a low False Positive Rate $\longrightarrow$ Unnecessary treatment for healthy patients
- We can examine the regions of the ROC curve based on desired result

# Regions of the ROC Curve




Note: E (bad) is negation of A

# ROC Curve for Multiple Classifiers



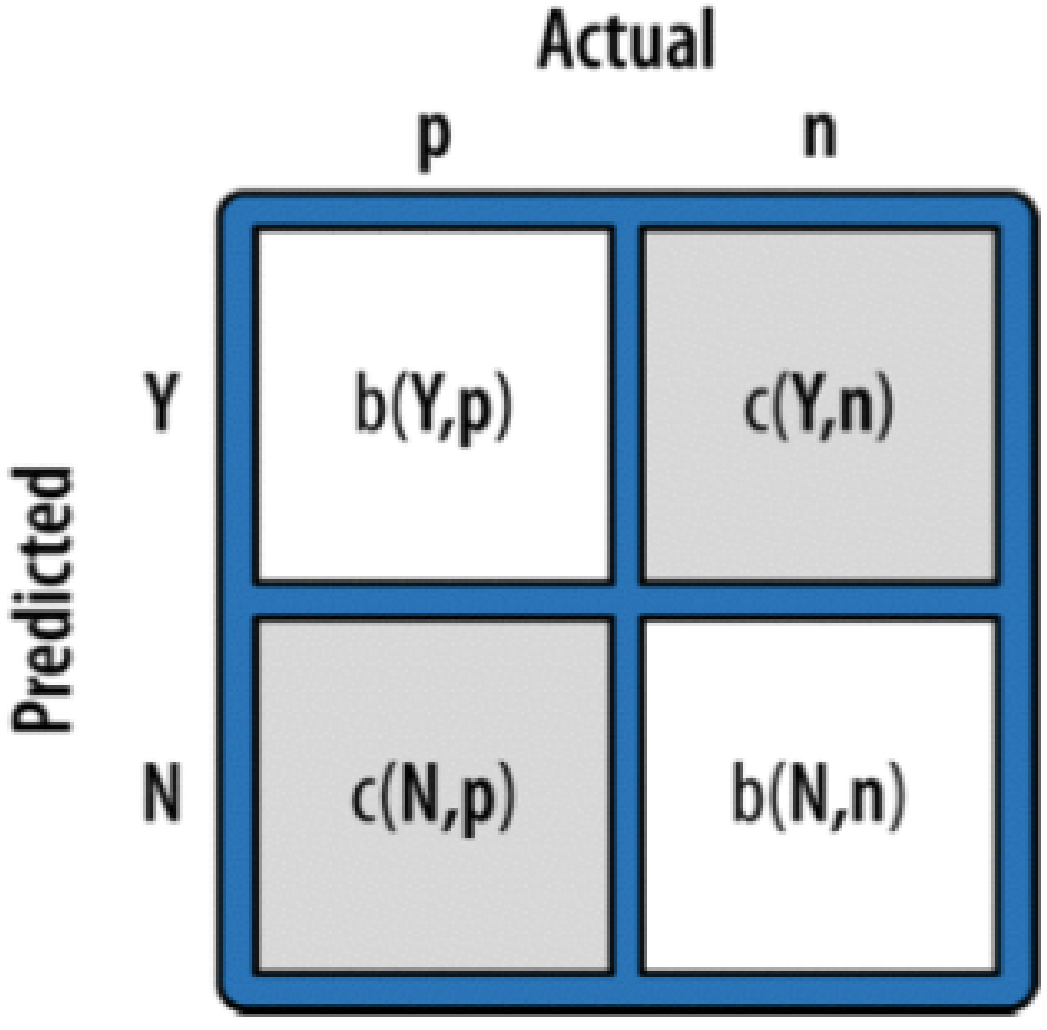
# Cost-Benefit Information

What if instead there is an inherent cost to losing a customer or benefit to keep them?

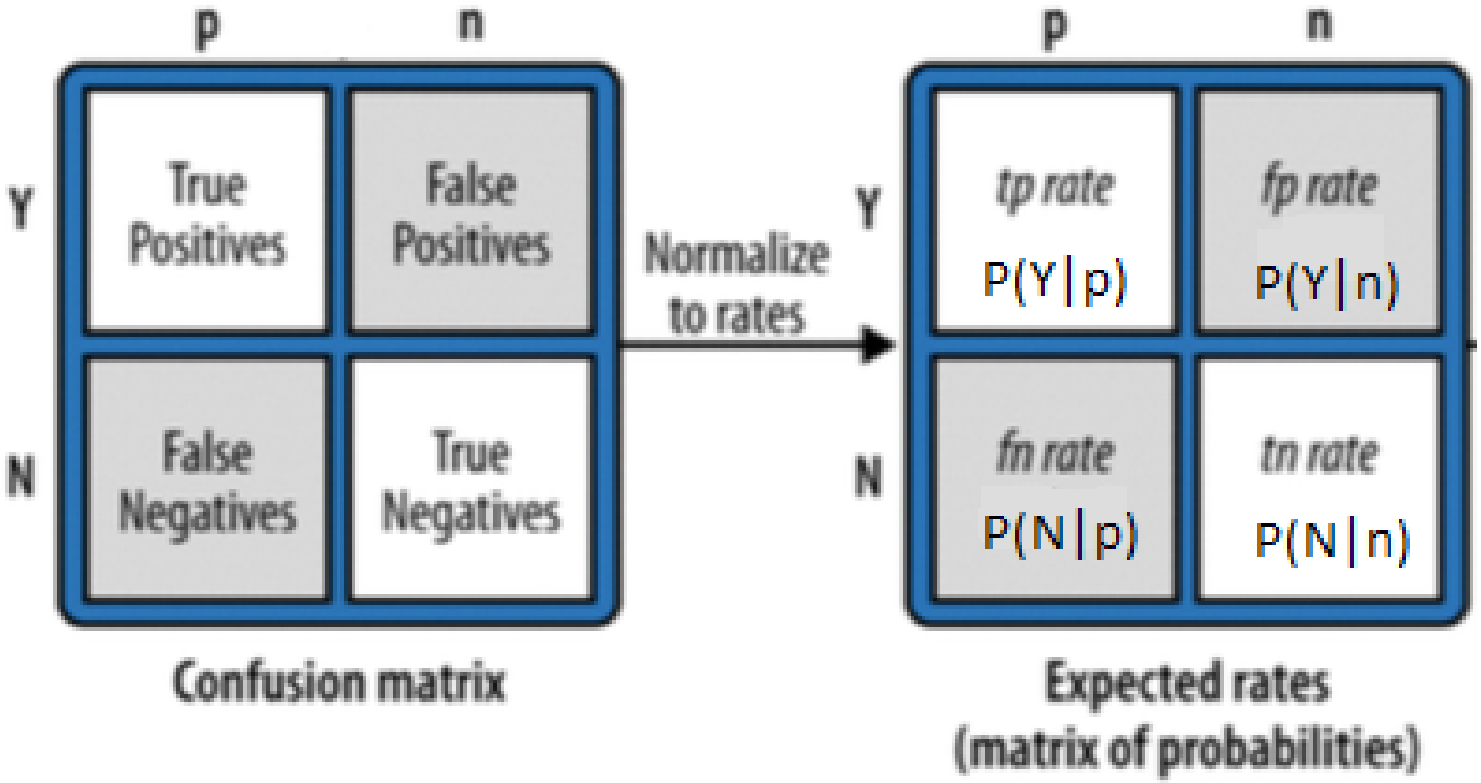
- ROC Curves alone assume equal cost of misclassification
- In reality, the cost of a misclassification or benefit of correct classification may be higher for one class than another

Profit Curves allow us to compare models and select the one that will maximize profit for a specified cost-benefit

# Cost-Benefit Matrix



# Normalize Confusion Matrix to Rates



# Expected Profit

By combining information from the Confusion Matrix and the Cost-Benefit Matrix, we can calculate the Expected Profit:

$$
\begin{aligned}
E[Profit] &= P(Y,p)\cdot b(Y,p) + P(Y,n)\cdot c(Y,n) + \\
&= P(N,p)\cdot c(N,p) + P(N,n)\cdot b(N,n) \\
&= P(Y|p)\cdot P(p)\cdot b(Y,p) + P(Y|n)\cdot P(n)\cdot c(Y,n) + \\
&= P(N|p)\cdot P(p)\cdot c(N,p) + P(N|n)\cdot P(n)\cdot b(N,n) \\
&= P(p)\cdot [P(Y|p)\cdot b(Y,p) + P(N|p)\cdot c(N,p)] + \\
&= P(n)\cdot [P(Y|n)\cdot c(Y,n) + P(N|n)\cdot b(N,n)]
\end{aligned}
$$

# Building the Profit Curve

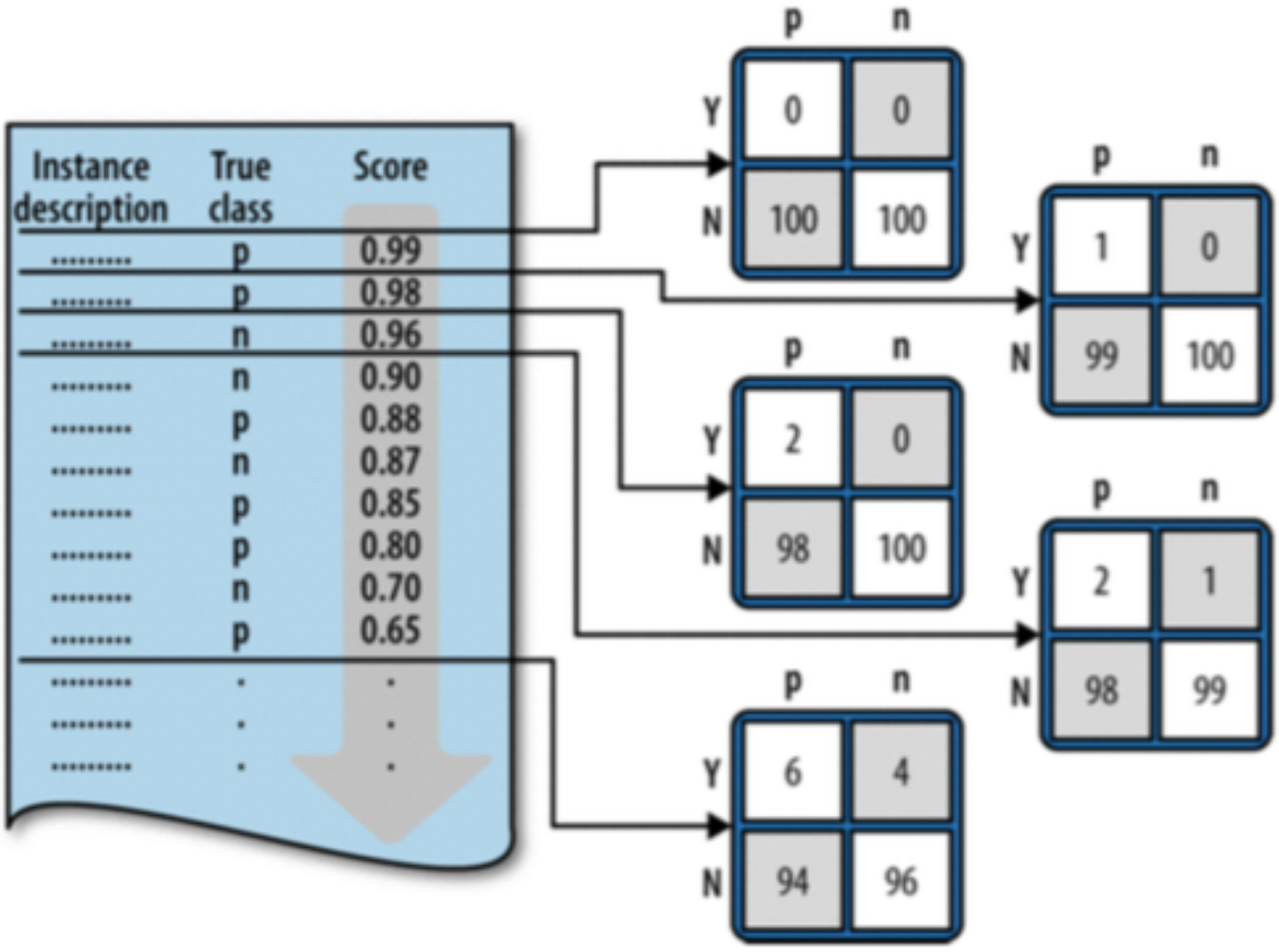
Similar to building ROC Curve. For a given model $f$, each threshold value T gives a point on the Profit Curve

Model score is the probability of class membership

1. T = maximum score
2. Using the confusion matrix and cost-benefit matrix, calculate $E[Profit]$
3. For each observation, $i$:
   - If $i > T \longrightarrow$ increment TP
   - else $\longrightarrow$ increment FP
4. Add point ($E[Profit]$, % Test Instances) to Profit Graph

Increment T from max-score to min-score, repeating steps 1-4

# Building the Profit Curve



# Profit Curves for Multiple Classifiers