True positive rate (Sensitivity)

$$\text{true positive rate} = \frac{\#\text{ of true positives}}{\#\text{ of known positives}}$$

(Proportion of actual positives that are correctly identified)

True negative rate (Specificity)

$$\text{true negative rate} = \frac{\text{\# of true negatives}}{\text{\# of known negatives}}$$

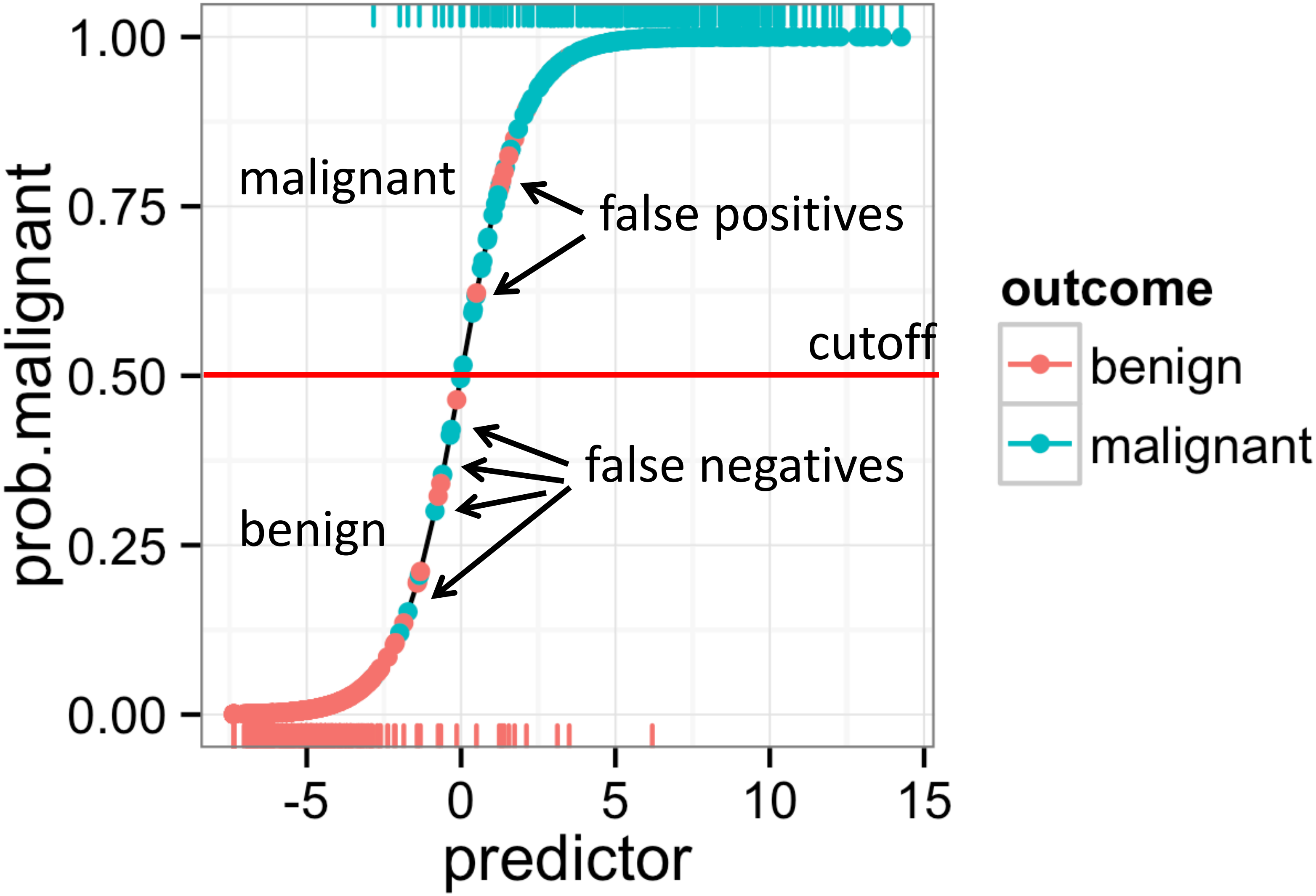
(Proportion of actual negatives that are correctly identified)

False positive rate (1 – Specificity)

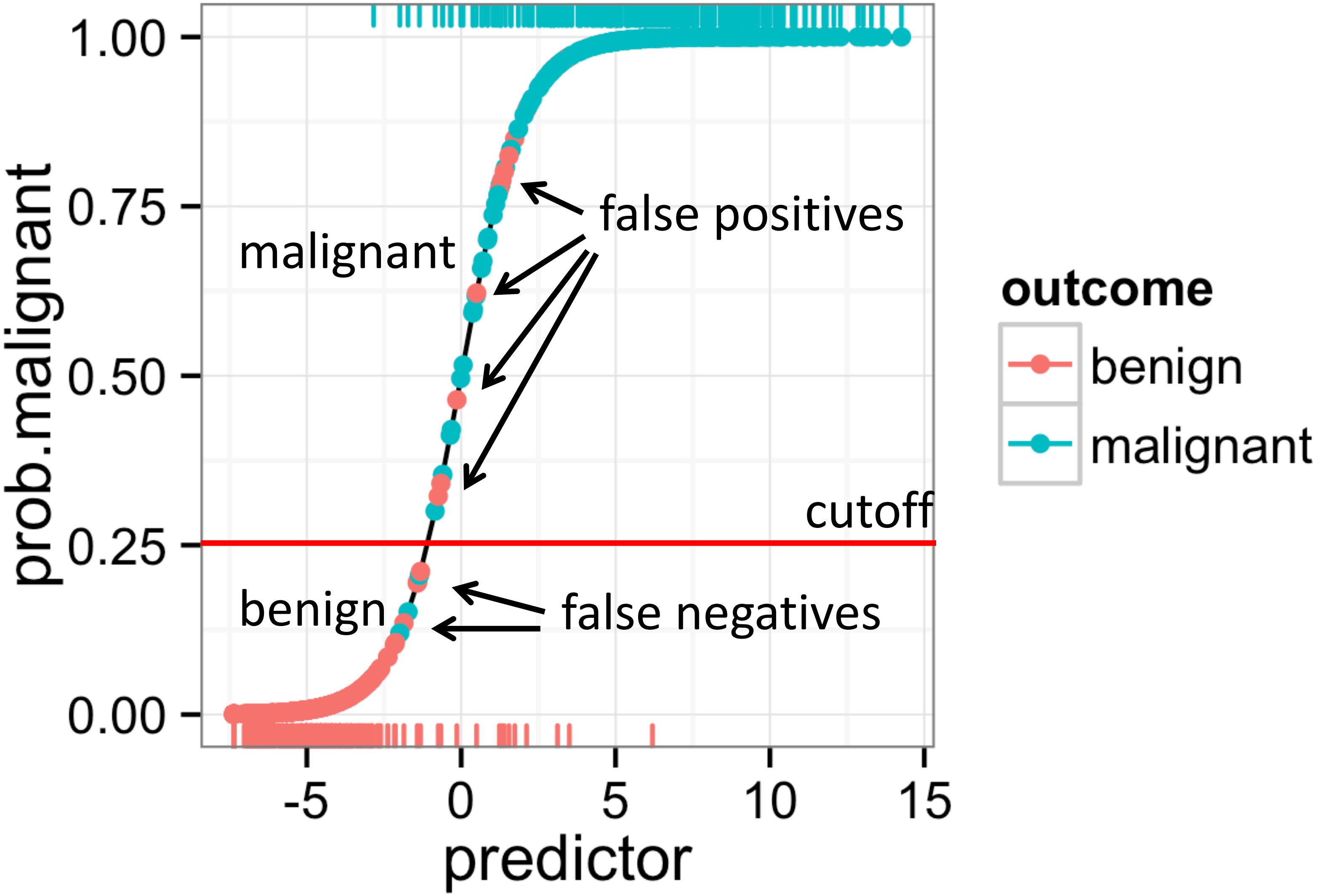
$$\text{false positive rate} = \frac{\text{\# of false positives}}{\text{\# of known negatives}}$$

(Proportion of actual negatives that are incorrectly identified)

Sensitivity and specificity depend on a chosen cutoff

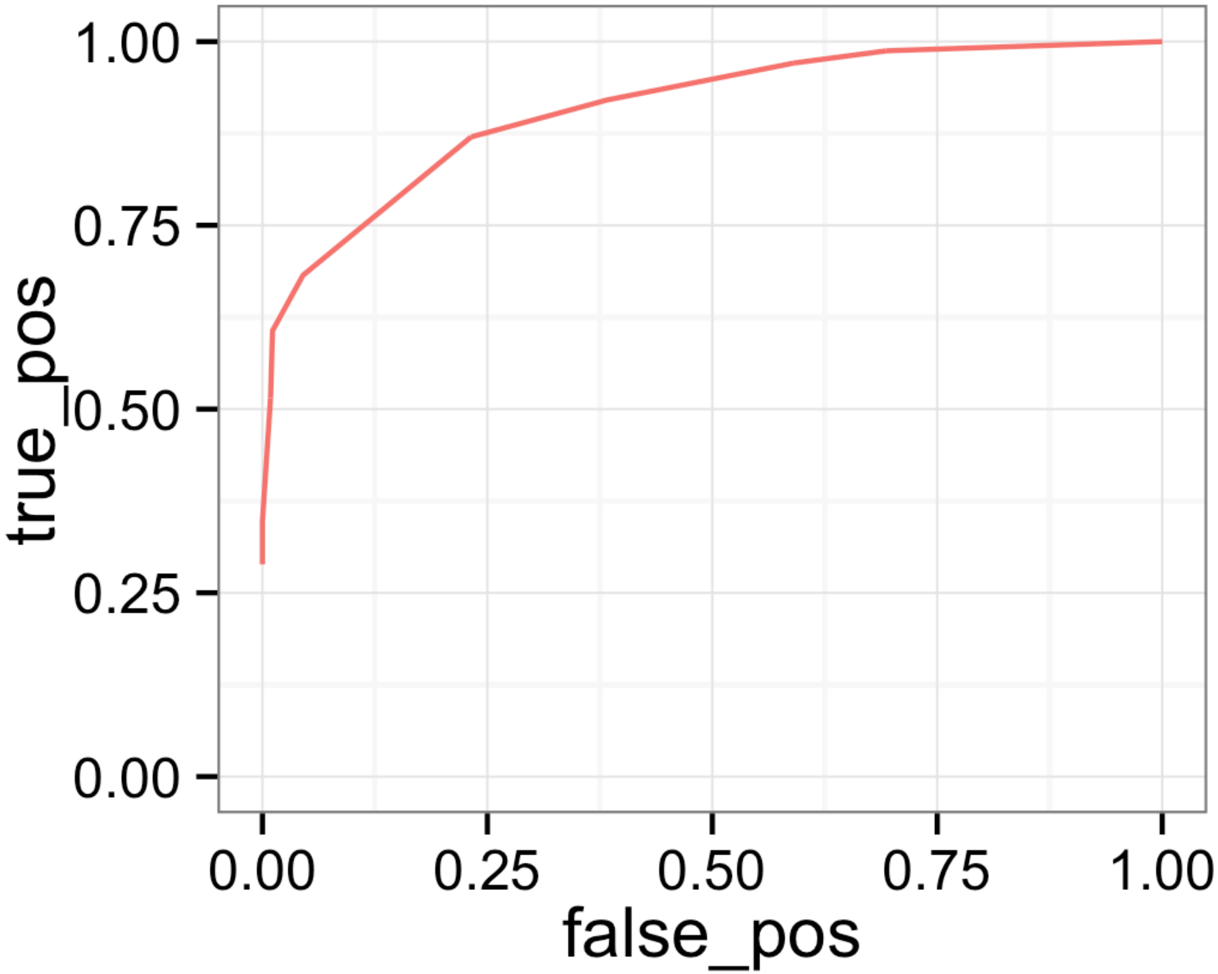


Sensitivity and specificity depend on a chosen cutoff



Do Part 1 of the worksheet now

We usually plot the true pos. rate vs. the false pos. rate for all possible cutoffs



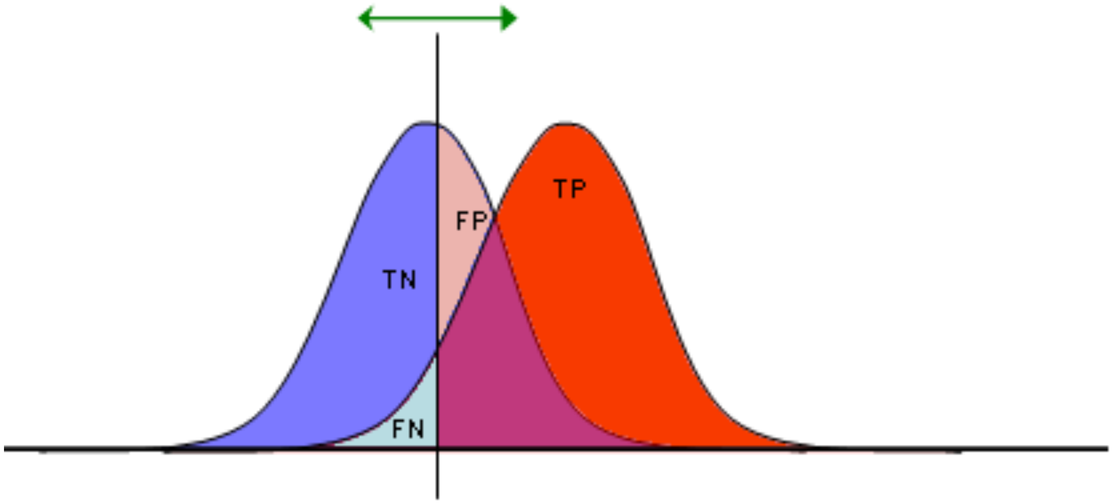
ROC curve

Receiver

Operating

Characteristic

curve



TP	FP
FN	TN
1	1

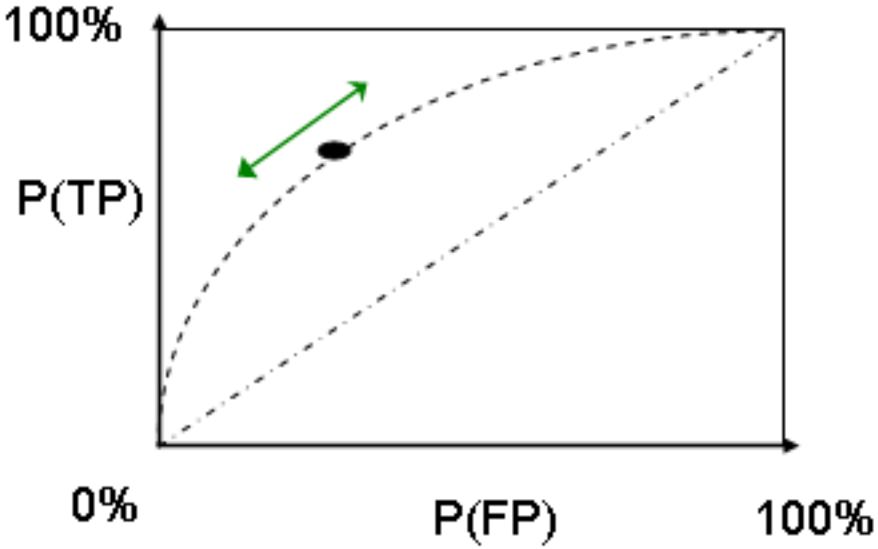
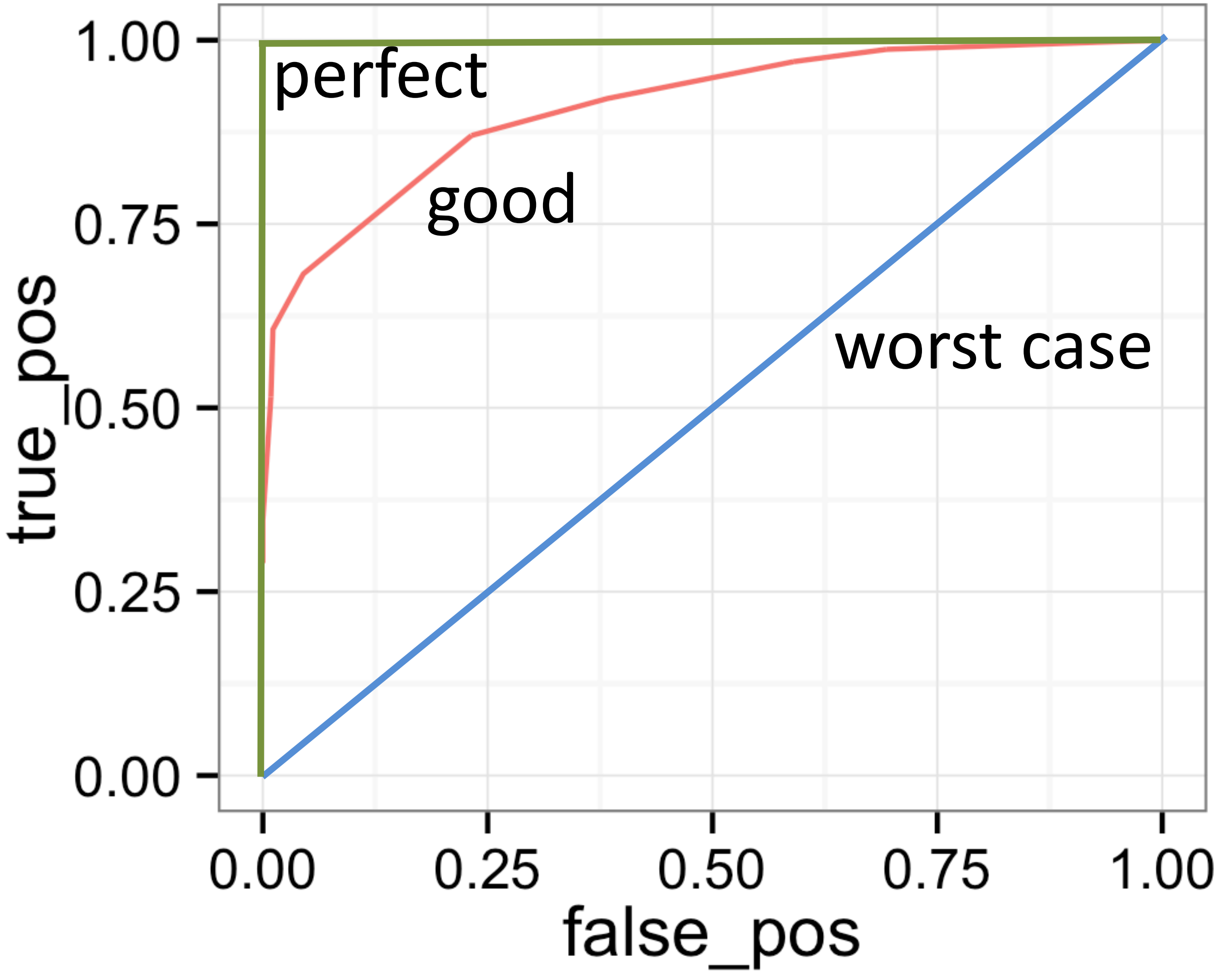
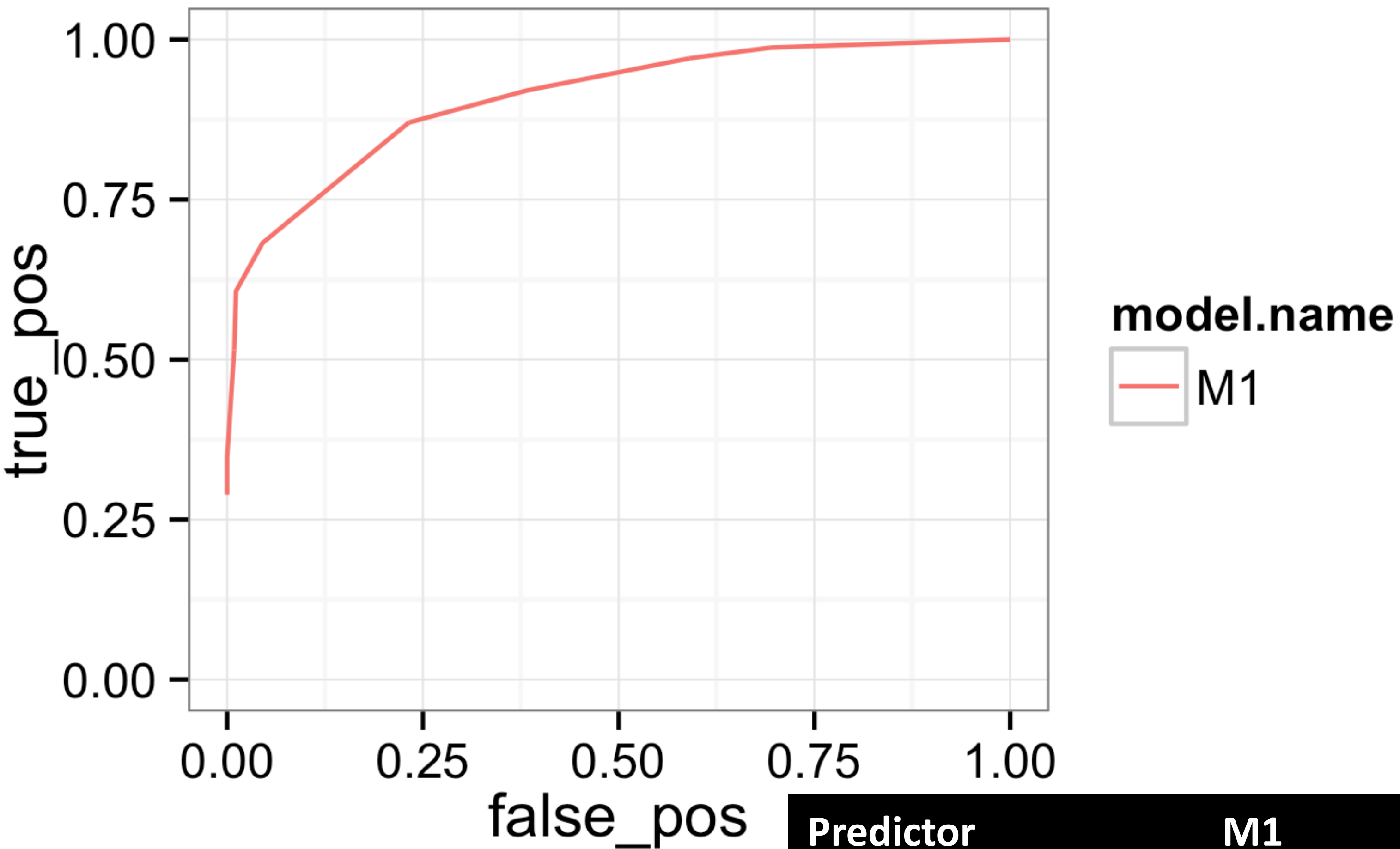


Image from: http://en.wikipedia.org/wiki/Receiver_operating_characteristic

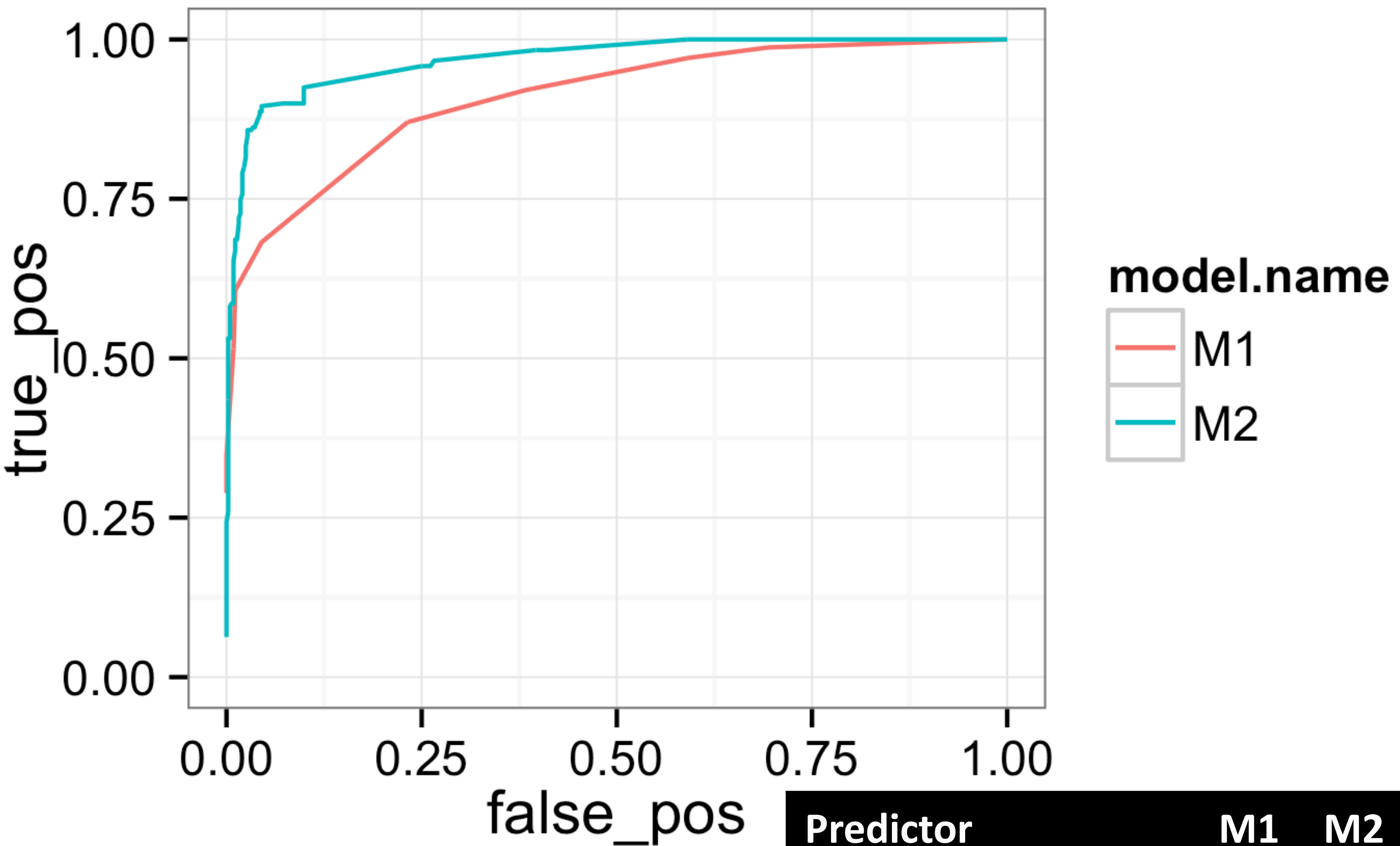
The area under the curve tells us how good a model's predictions are



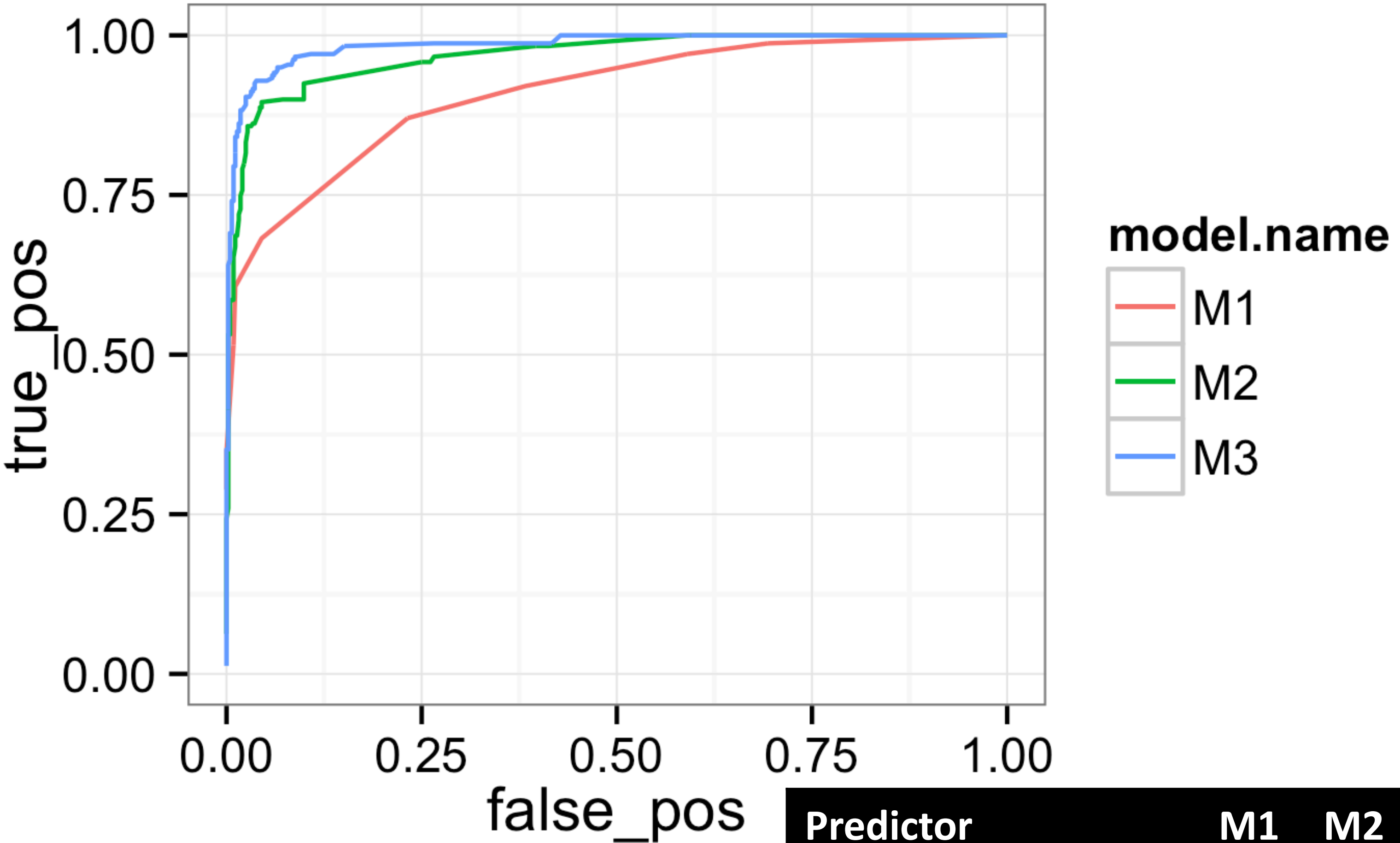
Let’s look at the performance of several different models for the biopsy data set



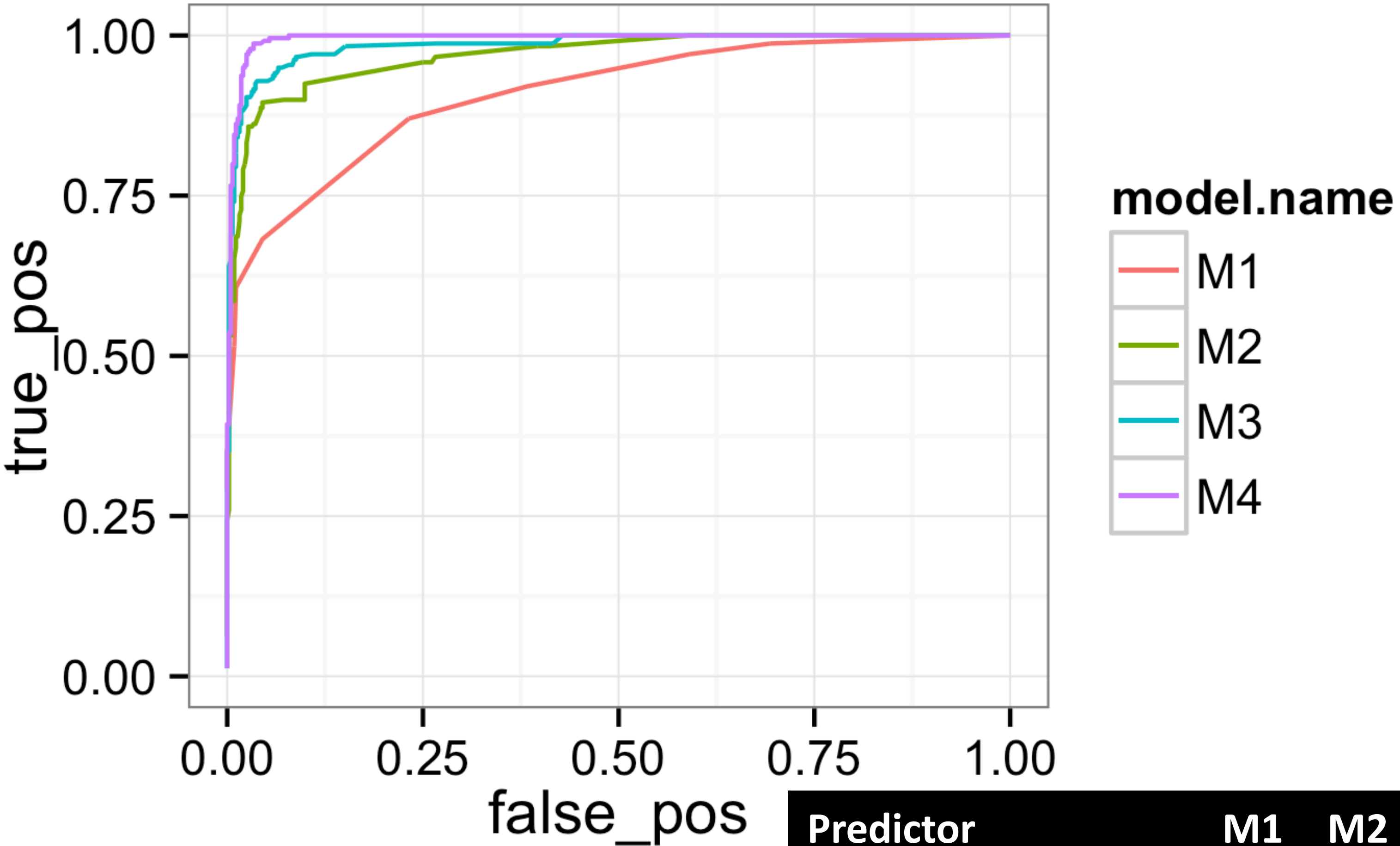
Predictor	M1
clump_thickness	✔
normal_nucleoli	
marg_adhesion	
bare_nuclei	
uniform_cell_shape	
bland_chromatin	



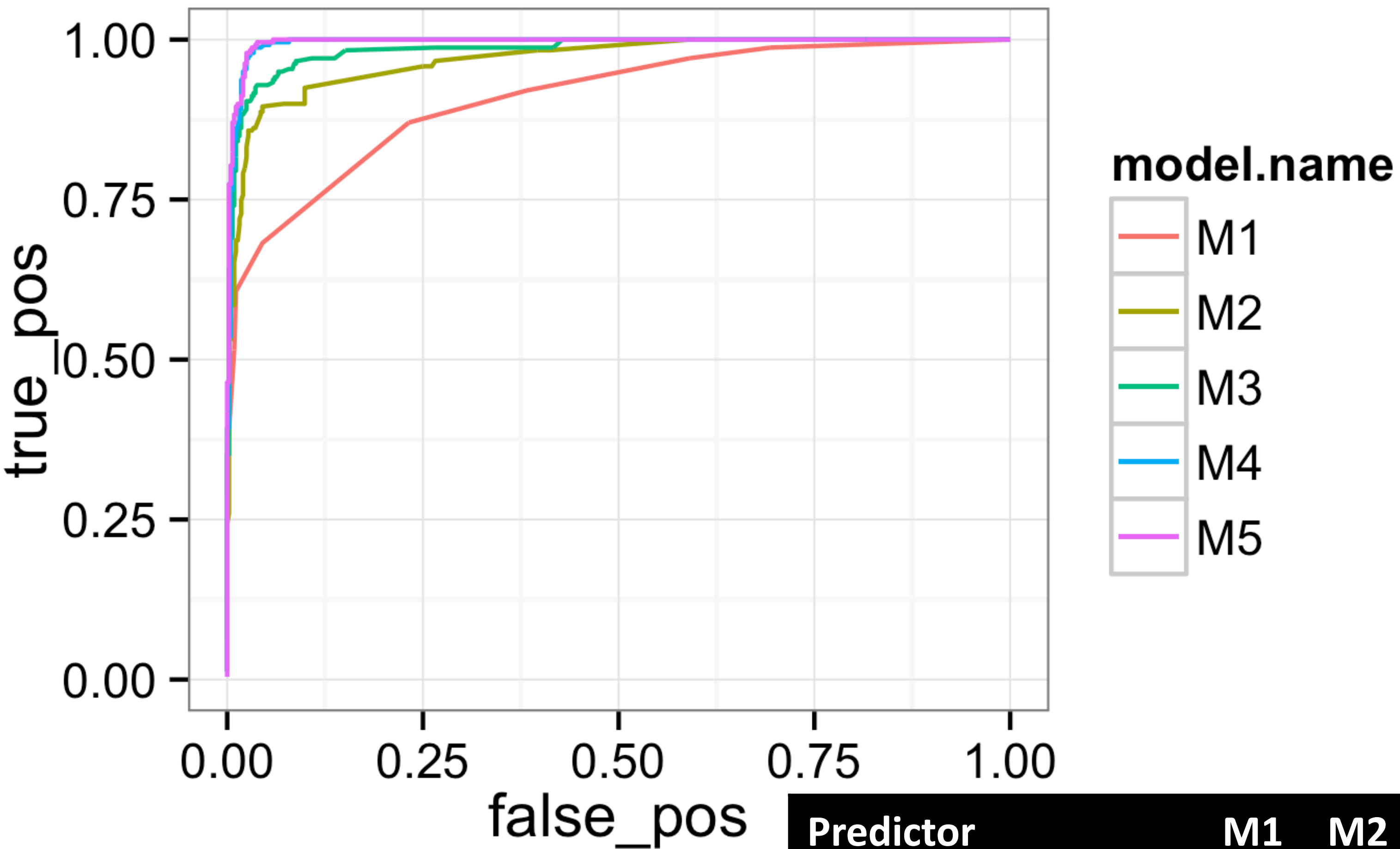
Predictor	M1	M2
clump_thickness	✔	✔
normal_nucleoli		✔
marg_adhesion		
bare_nuclei		
uniform_cell_shape		
bland_chromatin		



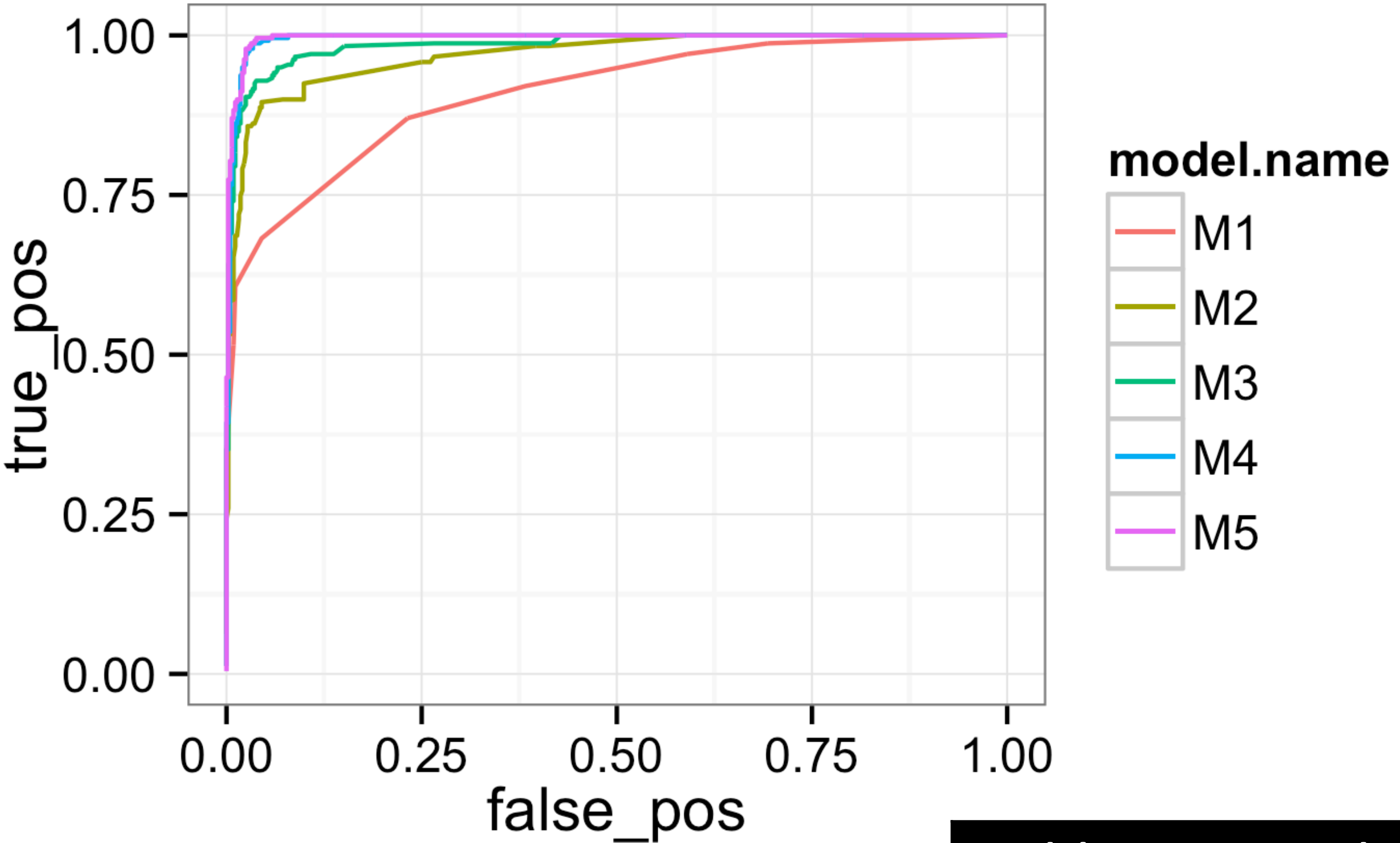
Predictor	M1	M2	M3
clump_thickness	✔	✔	✔
normal_nucleoli		✔	✔
marg_adhesion			✔
bare_nuclei			
uniform_cell_shape			
bland_chromatin			



Predictor	M1	M2	M3	M4
clump_thickness	✔	✔	✔	✔
normal_nucleoli		✔	✔	✔
marg_adhesion			✔	✔
bare_nuclei				✔
uniform_cell_shape				
bland_chromatin				

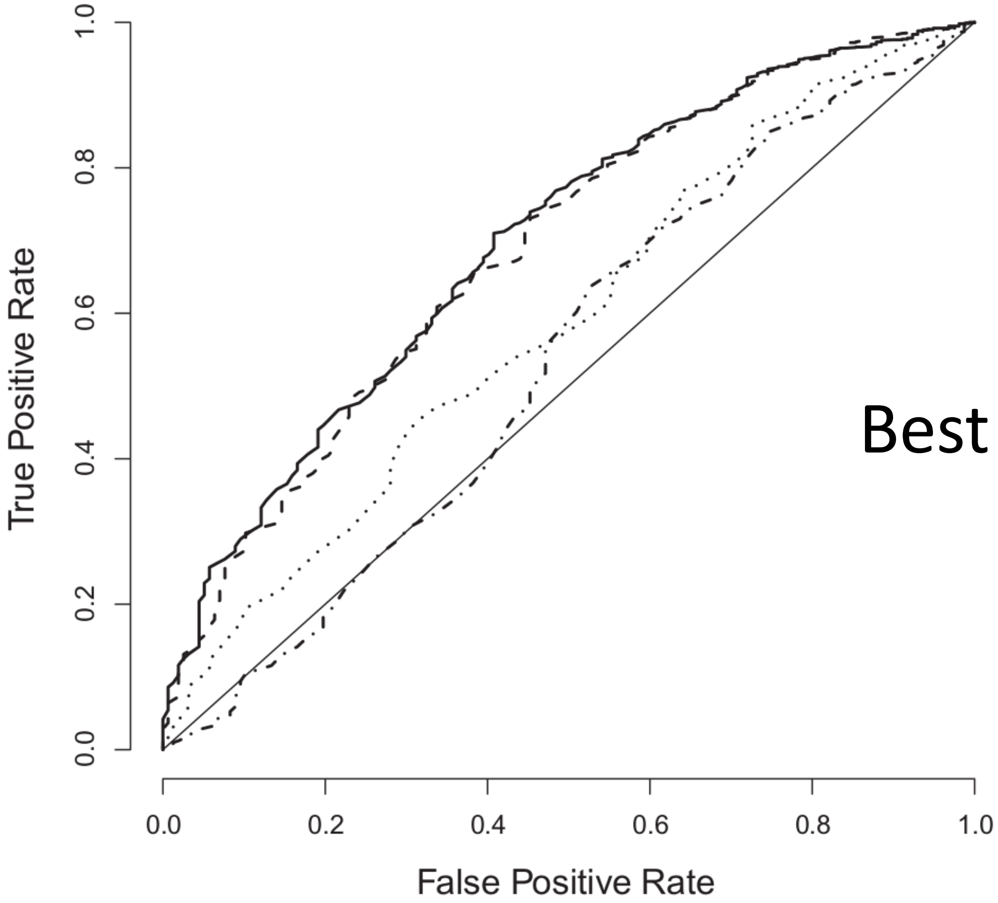


Predictor	M1	M2	M3	M4	M5
clump_thickness	✔	✔	✔	✔	✔
normal_nucleoli		✔	✔	✔	✔
marg_adhesion			✔	✔	✔
bare_nuclei				✔	✔
uniform_cell_shape					✔
bland_chromatin					✔



Model	Area Under Curve (AUC)
M1	0.940
M2	0.974
M3	0.985
M4	0.995
M5	0.996

Things usually look much worse in real life



Best AUC (solid line): 0.70

Keller, Mis, Jia, Wilke. Genome Biol. Evol. 4:80-88, 2012

Calculating ROC curves in R

We need a custom-built function: calc_ROC()

```
calc_ROC <- function(probabilities, known_truth, model.name=NULL)
{
  outcome <- as.numeric(factor(known_truth))-1
  pos <- sum(outcome)   # total known positives
  neg <- sum(1-outcome) # total known negatives
  pos_probs <- outcome*probabilities     # probabilities for known positives
  neg_probs <- (1-outcome)*probabilities # probabilities for known negatives
  true_pos <- sapply(probabilities,
                     function(x) sum(pos_probs>=x)/pos) # true pos. rate
  false_pos <- sapply(probabilities,
                     function(x) sum(neg_probs>=x)/neg) # false pos. rate
  if (is.null(model.name))
    result <- data.frame(true_pos, false_pos)
  else
    result <- data.frame(true_pos, false_pos, model.name)
  result %>% arrange(false_pos, true_pos)
}
```

(We do not cover here how this function works, we just use it as a black box.)

Using the function `calc_ROC()`

Using the function `calc_ROC()`

```
# fit a logistic regression model
glm.out <- glm(outcome ~ clump_thickness,
               data=biopsy, family=binomial)
```

Using the function `calc_ROC()`

```
# fit a logistic regression model
glm.out <- glm(outcome ~ clump_thickness,
               data=biopsy, family=binomial)

# calculate ROC curve
ROC1 <- calc_ROC(probabilities=glm.out$fitted.values,
                 known_truth=biopsy$outcome,
                 model.name='M1')
```

Using the function calc_ROC()

```
# fit a logistic regression model
glm.out <- glm(outcome ~ clump_thickness,
               data=biopsy, family=binomial)

# calculate ROC curve
ROC1 <- calc_ROC(probabilities=glm.out$fitted.values,
                 known_truth=biopsy$outcome,
                 model.name='M1')

# Result
> ROC1
     true_pos   false_pos model.name
1   0.2887029 0.000000000         M1
2   0.2887029 0.000000000         M1
3   0.2887029 0.000000000         M1
4   0.2887029 0.000000000         M1
5   0.2887029 0.000000000         M1
6   0.2887029 0.000000000         M1
```

Do Part 2 of the worksheet now