Logistic Regression

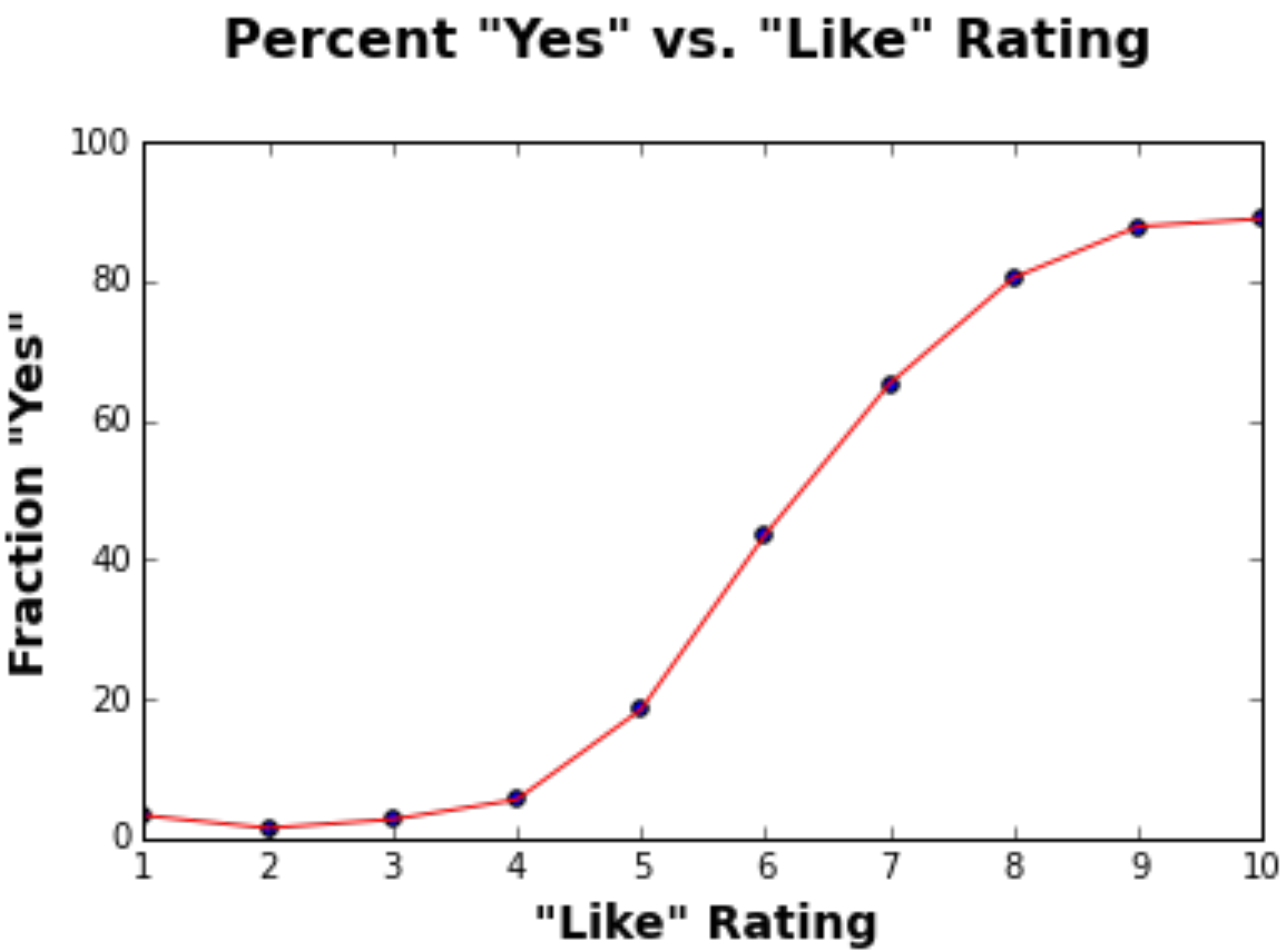
gSchool Data Science Spring 2015

Regression vs. Classification

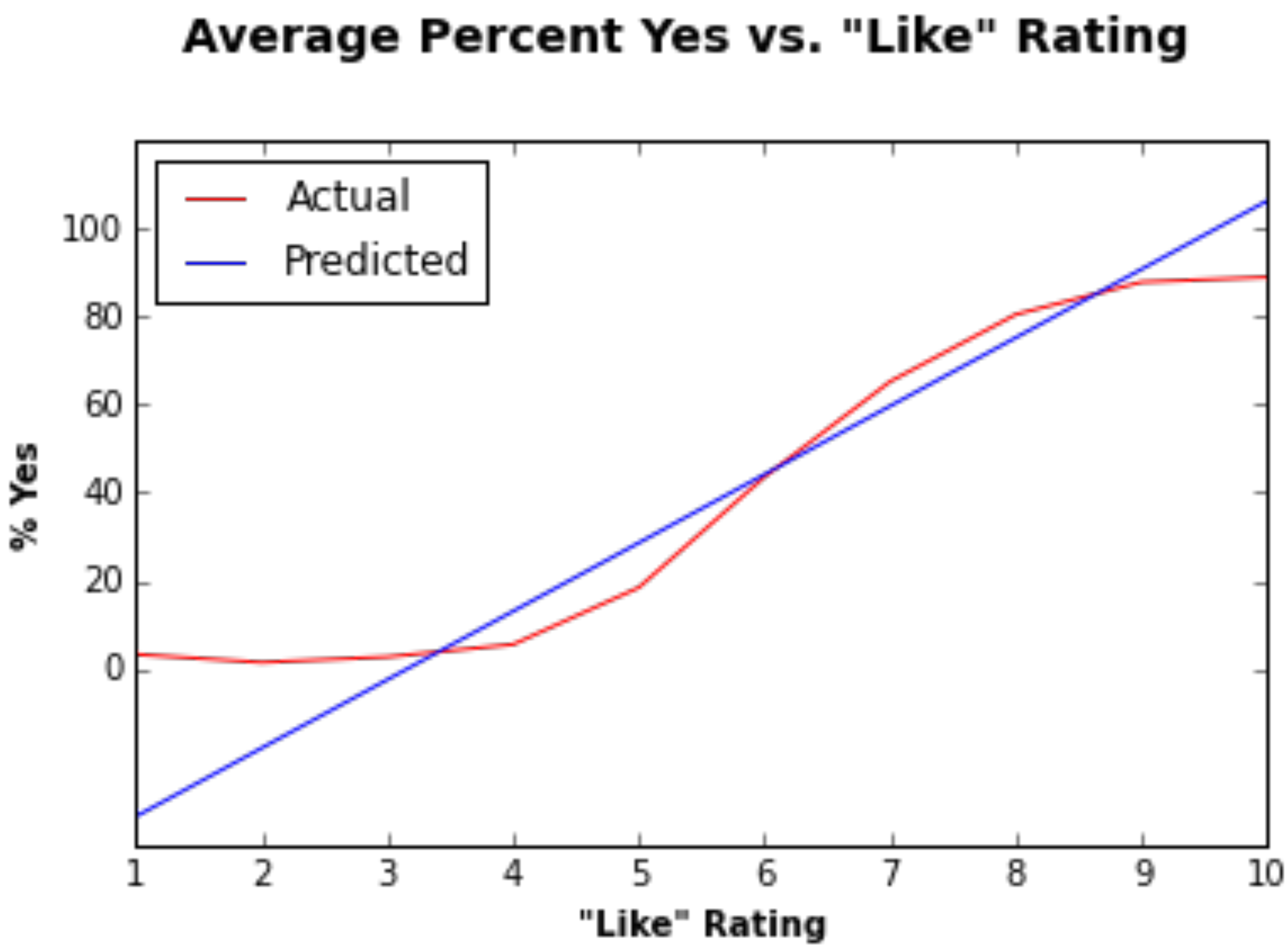
- So far we've looked at **regression** problems: predict a real number in terms of other real numbers.
- Today we study binary **classification** problems: predict whether or not example is in a *category*.
- Convention: construct a numeric variable which is 1 if example is in the category else 0.

A Classification Problem

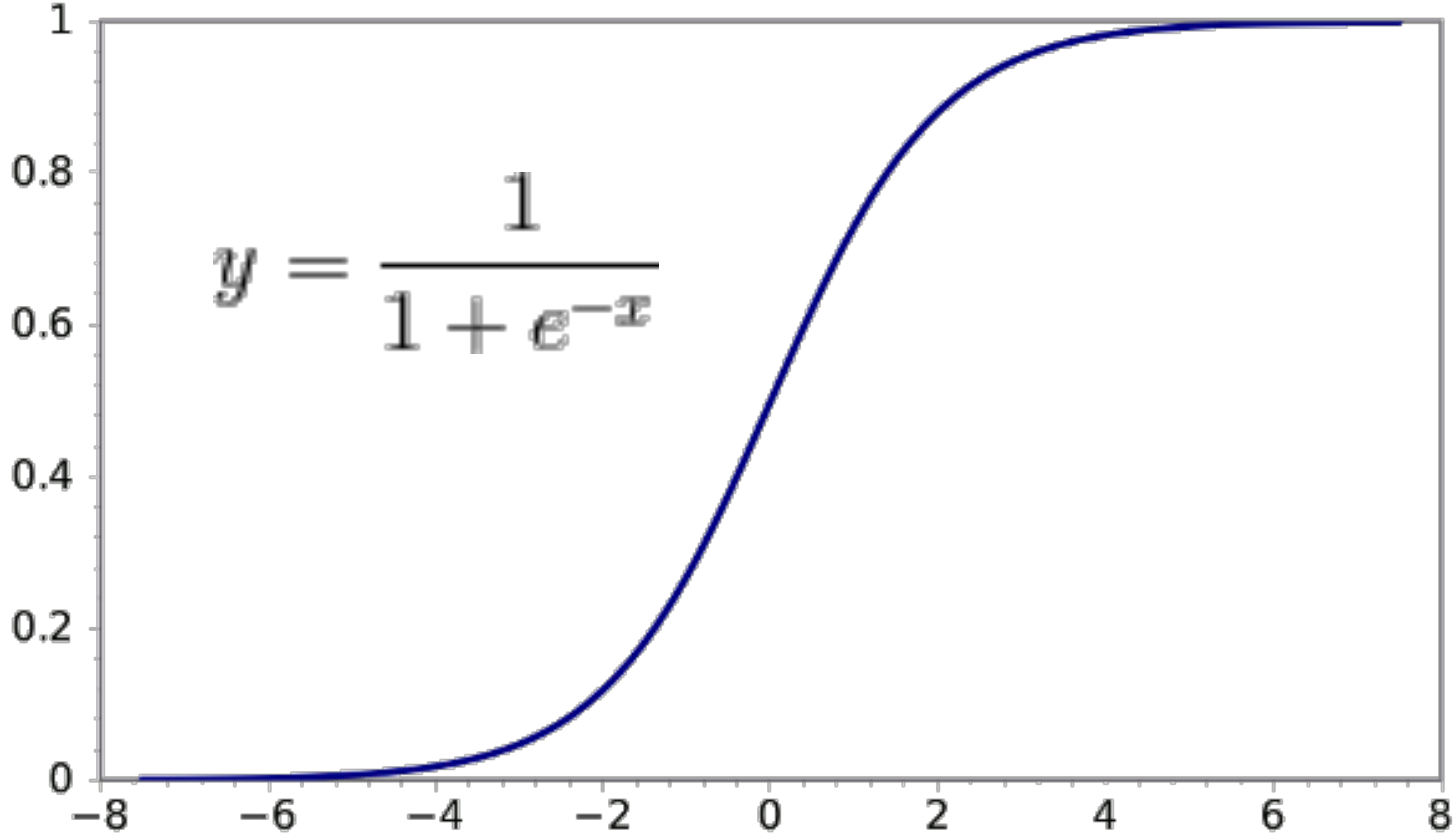
- Predict decisions at speed dating events.
- Predictor: “Like” rating from 1 to 10.
- Prediction —> probability.


Percent "Yes" vs. "Like" Rating
Fraction "Yes"
"Like" Rating
100
80
60
40
20
0
1
2
3
4
5
6
7
8
9
10

Linear NOT OK



Sigmoid (our friend)



Logistic regression

$$F(x) = \frac{1}{1 + e^{-(\beta_o + \beta_1 x)}}$$

- **x:** predictor
- **F(x):** target variable
- **Beta_1**: Change in Log Odds Ratio (LOR)

Odds Ratios

Motivation: p = 0.99 and p = 0.9999 are very different!

$$OR = \frac{p}{1-p}$$

$$p - - - -| - (1-p)$$

$$p = 0.8 \Rightarrow OR = 4$$
$$p = 0.99 \Rightarrow OR = 99$$

Log Odds Ratio

- Fixes issue of odds ratio blowing up!

- Symmetry between p & 1 - p.

$$LOR = \log(OR) = \log\left(\frac{p}{1-p}\right)$$

p	0.0001	0.01	0	0.99	0.9999
LOR	-9.2	-4.6	0	4.6	9.2

LOR & Sigmoid

$$LOR = \log\left(\frac{p}{1-p}\right)$$

$$p = \frac{e^{LOR}}{e^{LOR}+1} = \frac{1}{1+e^{-LOR}}$$

$$LOR = \beta_0 + \beta_1 x$$

Interpretation of Betas

$$LOR_{\text{decision}} = -5.72 + 0.89 \cdot \text{like}$$

- Coefficient on “like”: change in LOR with each additional “like” point.

Like	**4**	**5**	**6**	**7**	**8**
LOR	-2.2	-1.3	-0.4	0.5	1.4
Probability	0.1	0.2	0.4	0.6	0.8

Likelihood function

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p_i^{y_i} \cdot (1 - p_i)^{1-y_i}$$

- **y_i:** Target variable.
- **p_i**: Probability model assigns to y_i.
- Choose betas to maximize this.
- Log is easier to work with.

Confusion Matrix

	True (Predicted)	False (Predicted)
True (Actual)	TP	FN
False (Actaual)	FP	TN

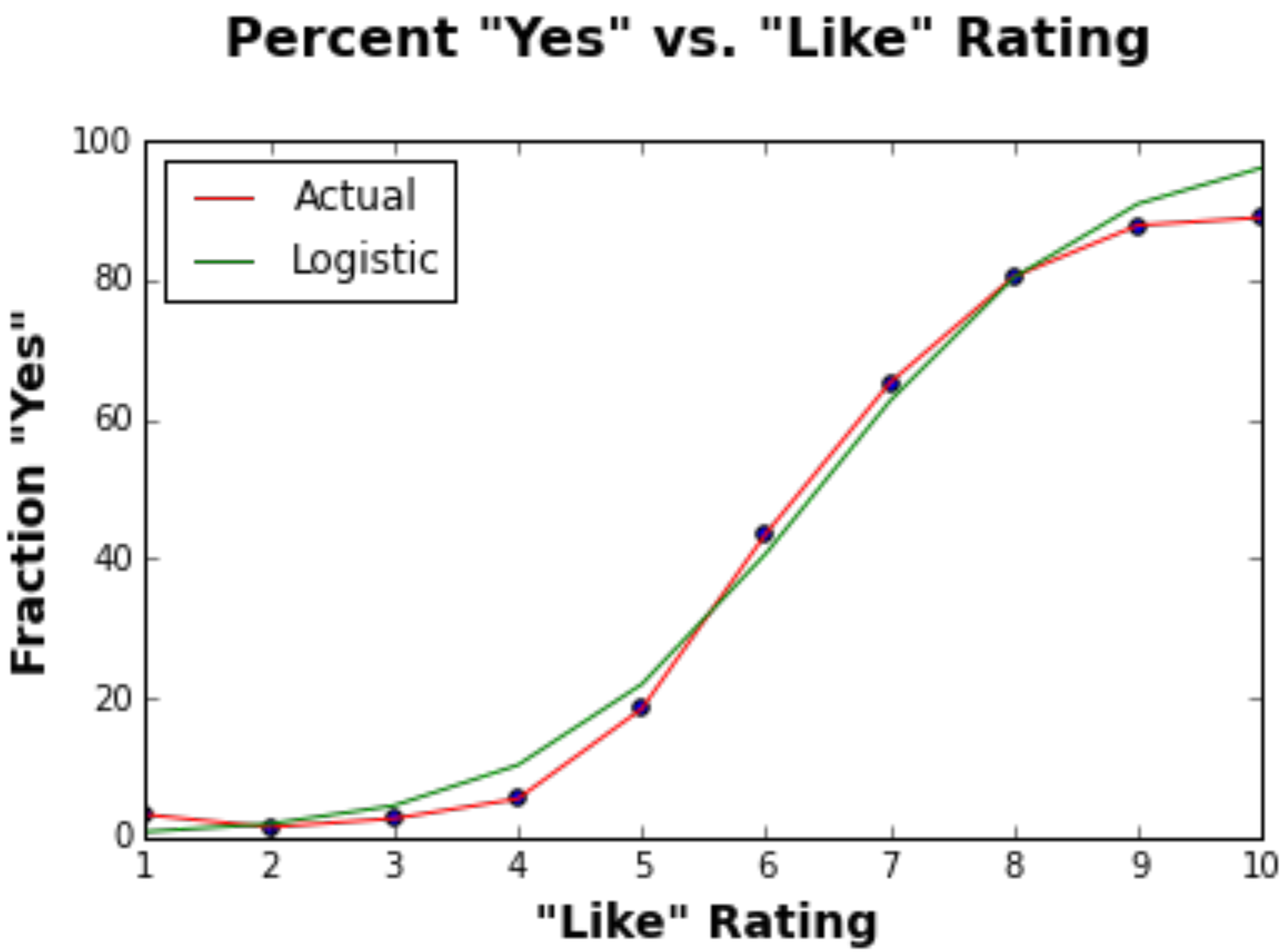
Log Loss

$$-\log(\mathrm{L}) = -\sum_{i=1}^{n} y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)$$

$$-\log(L) = \sum_{i=1}^{n} \mathrm{Cost}_i$$

$$\mathrm{Cost}_i = \begin{cases} -\log(p_i) & \text{if } y_i = 1 \\ -\log(1 - p_i) & \text{if } y_i = 0. \end{cases}$$

Much Better!



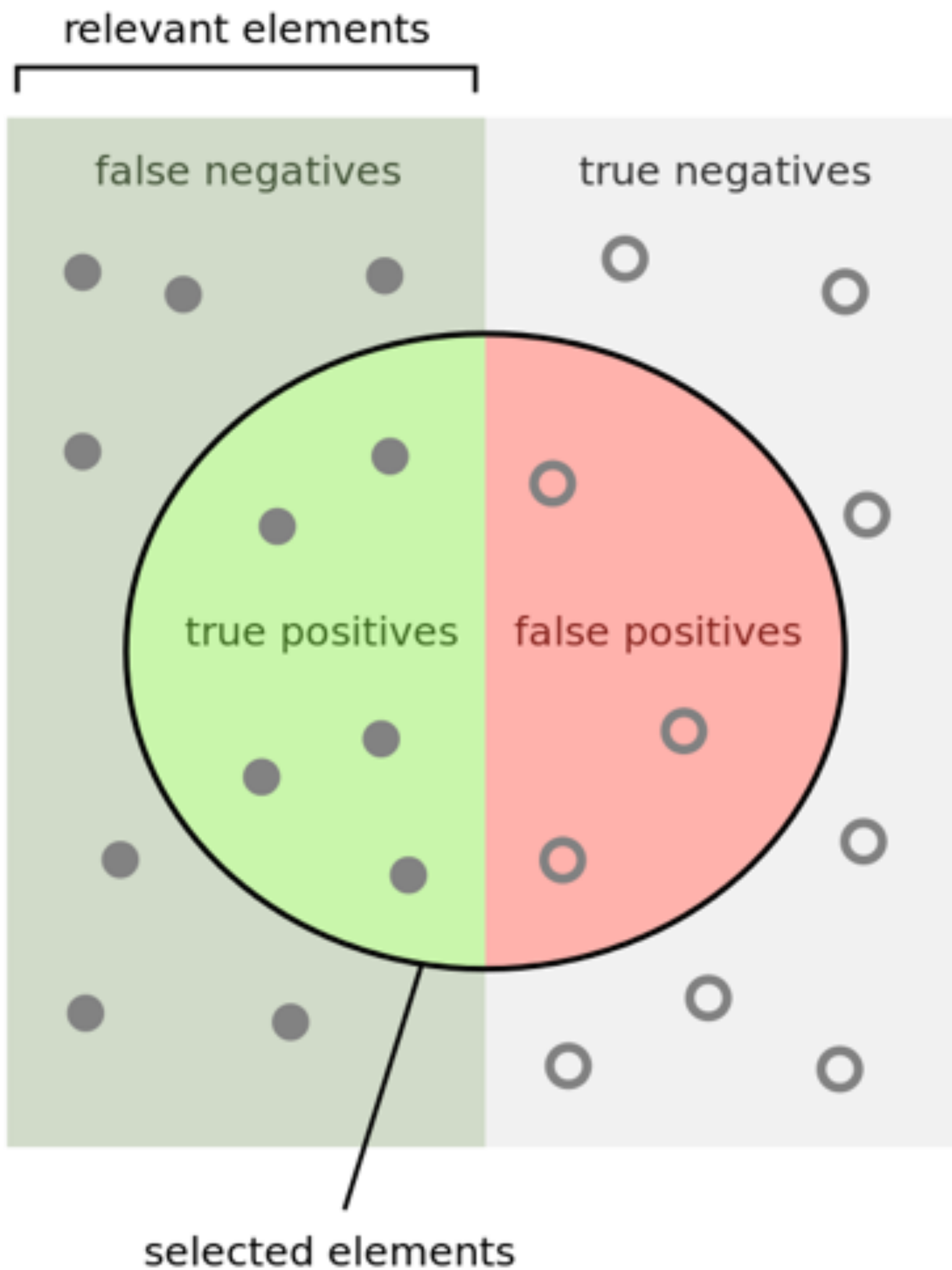
Percent "Yes" vs. "Like" Rating
Actual
Logistic
Fraction "Yes"
"Like" Rating
0
20
40
60
80
100
1
2
3
4
5
6
7
8
9
10

Classification

- Logistic regression ——> probabilities.
- Probabilities —-> predictions.
- Convention: predict “yes” iff $p > 0.5$.
- Measure quality of model.

Confusion Matrix

	Yes (Predicted)	**No (Predicted)**
Yes (Actual)	TP	FN
No (Actaual)	FP	TN

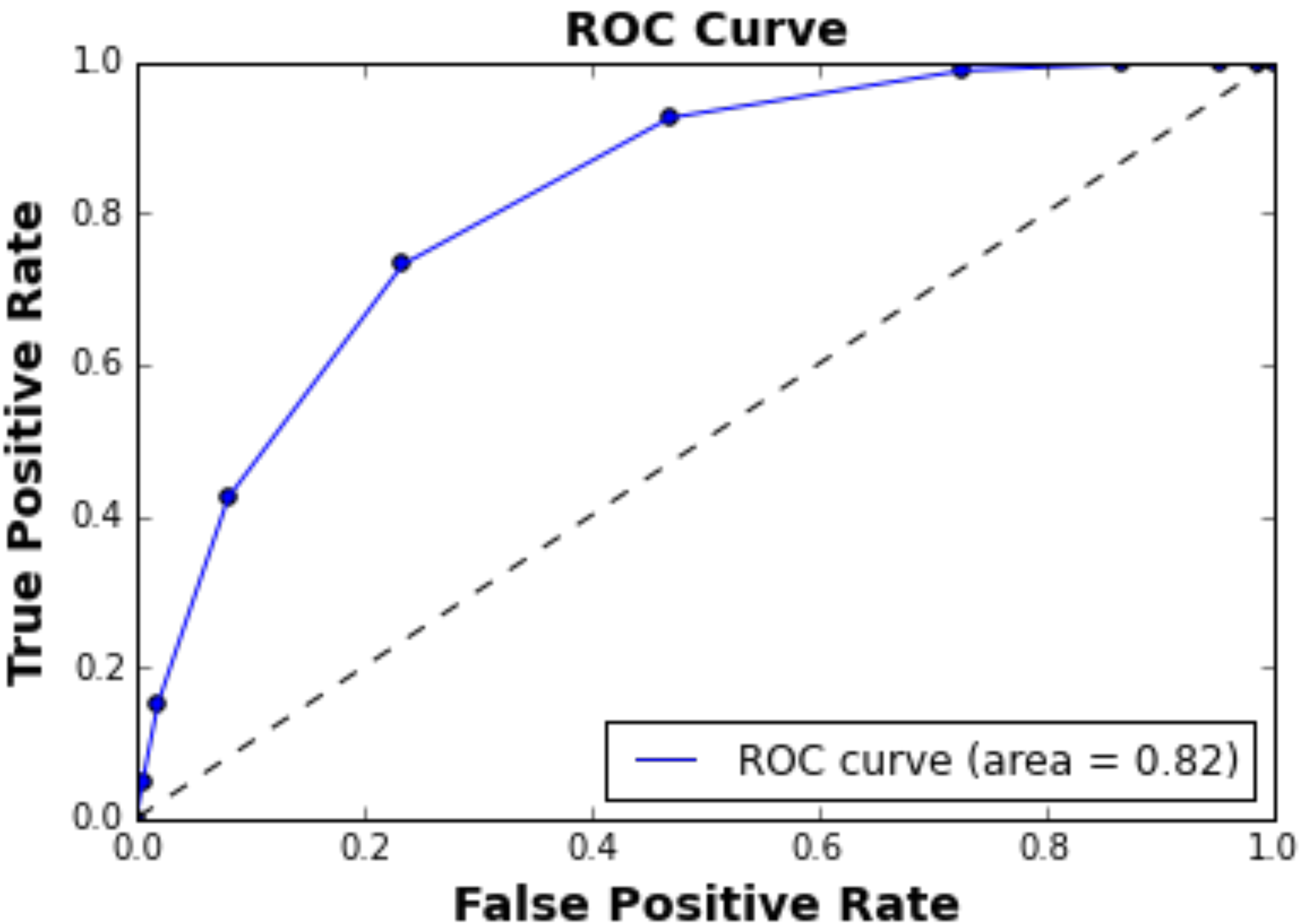


relevant elements
false negatives
true negatives
true positives
false positives
selected elements


How many selected items are relevant?
How many relevant items are selected?
Precision =
Recall =

Model performance

- Accuracy = (TP + TN)/N
- Precision = TP/(TP + FP)
- Recall = TP/(TP + FN)
- ROC Curve


ROC Curve
True Positive Rate
False Positive Rate
ROC curve (area = 0.82)
0.0
0.2
0.4
0.6
0.8
1.0