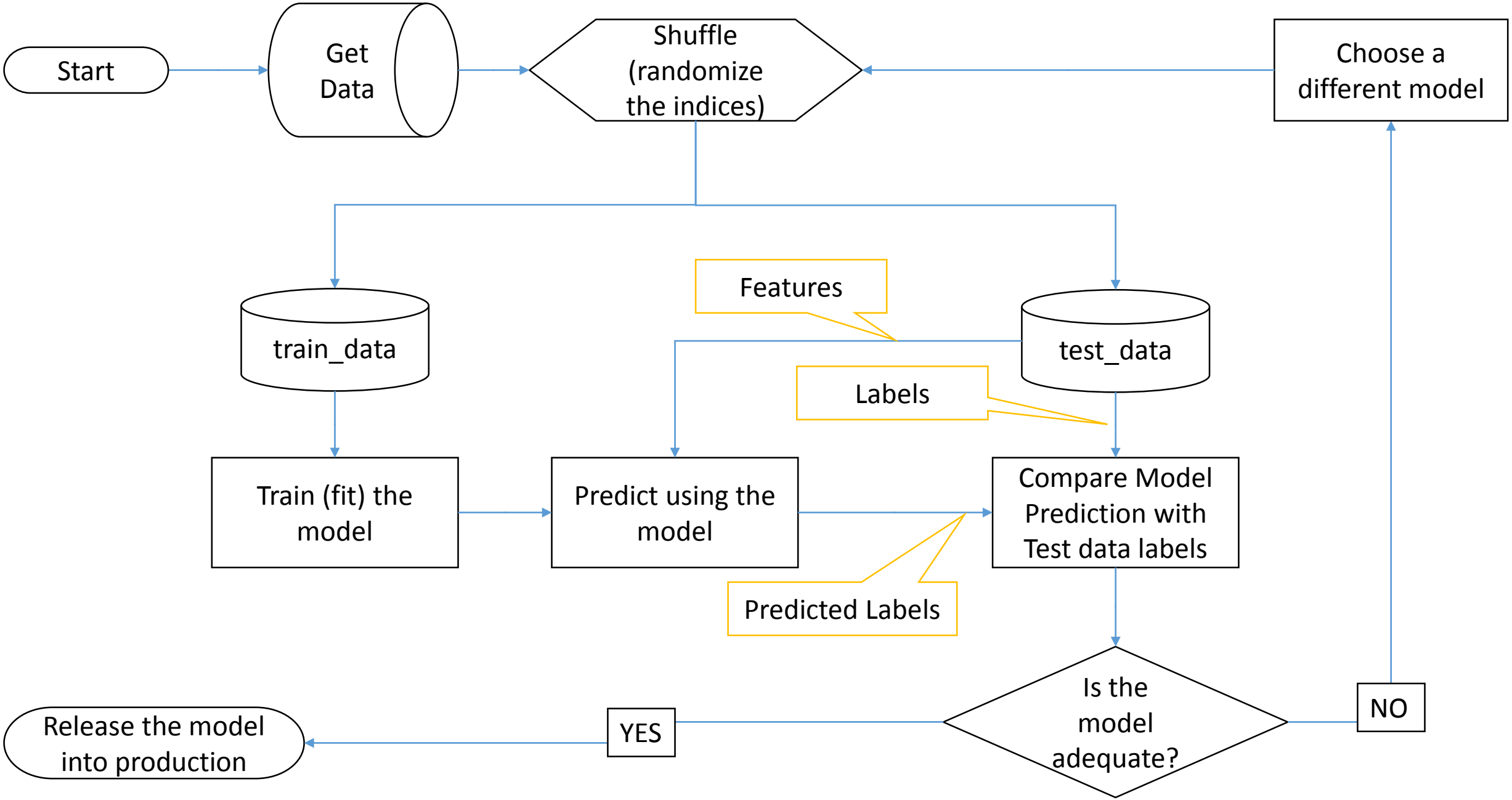
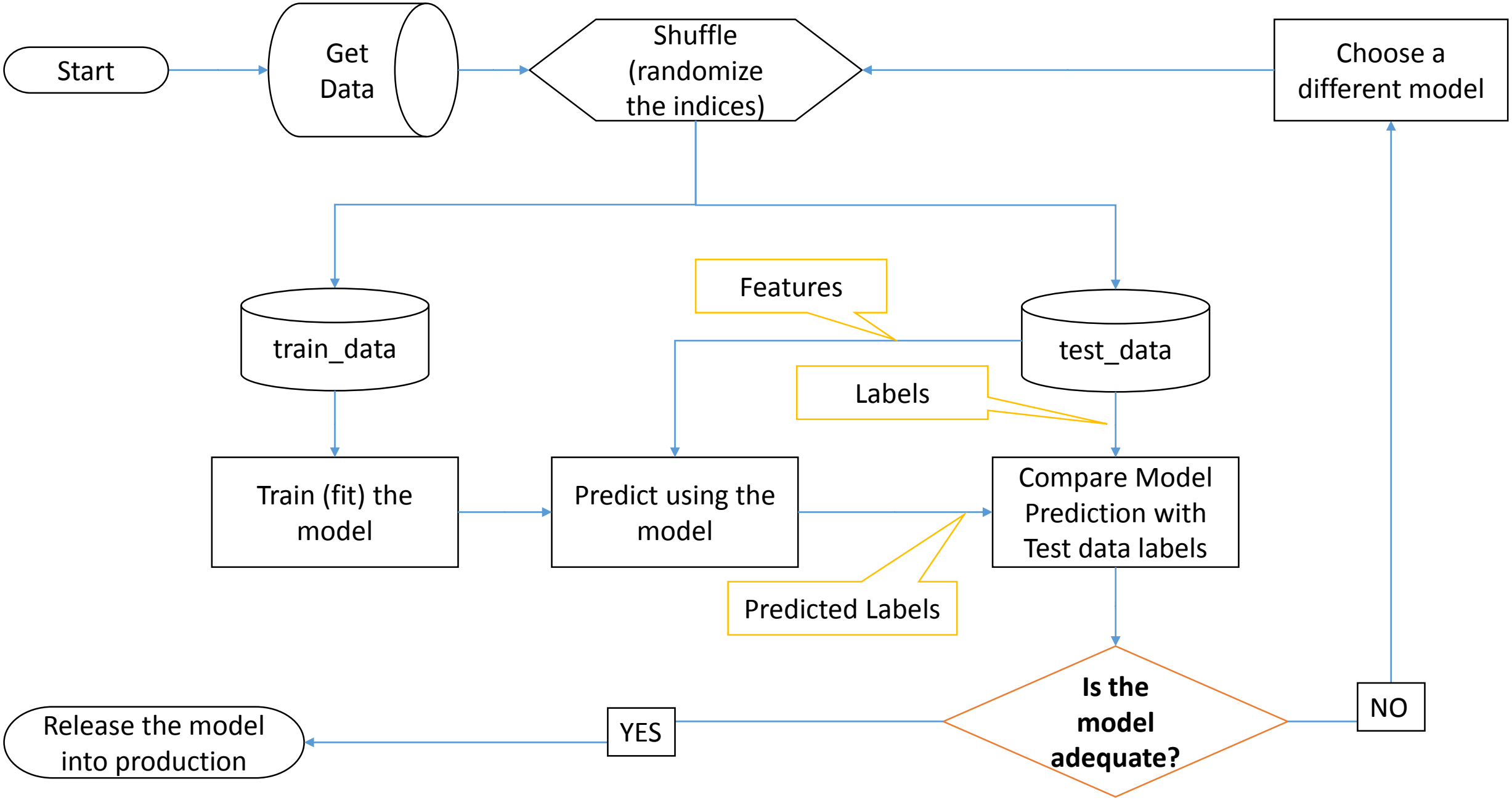
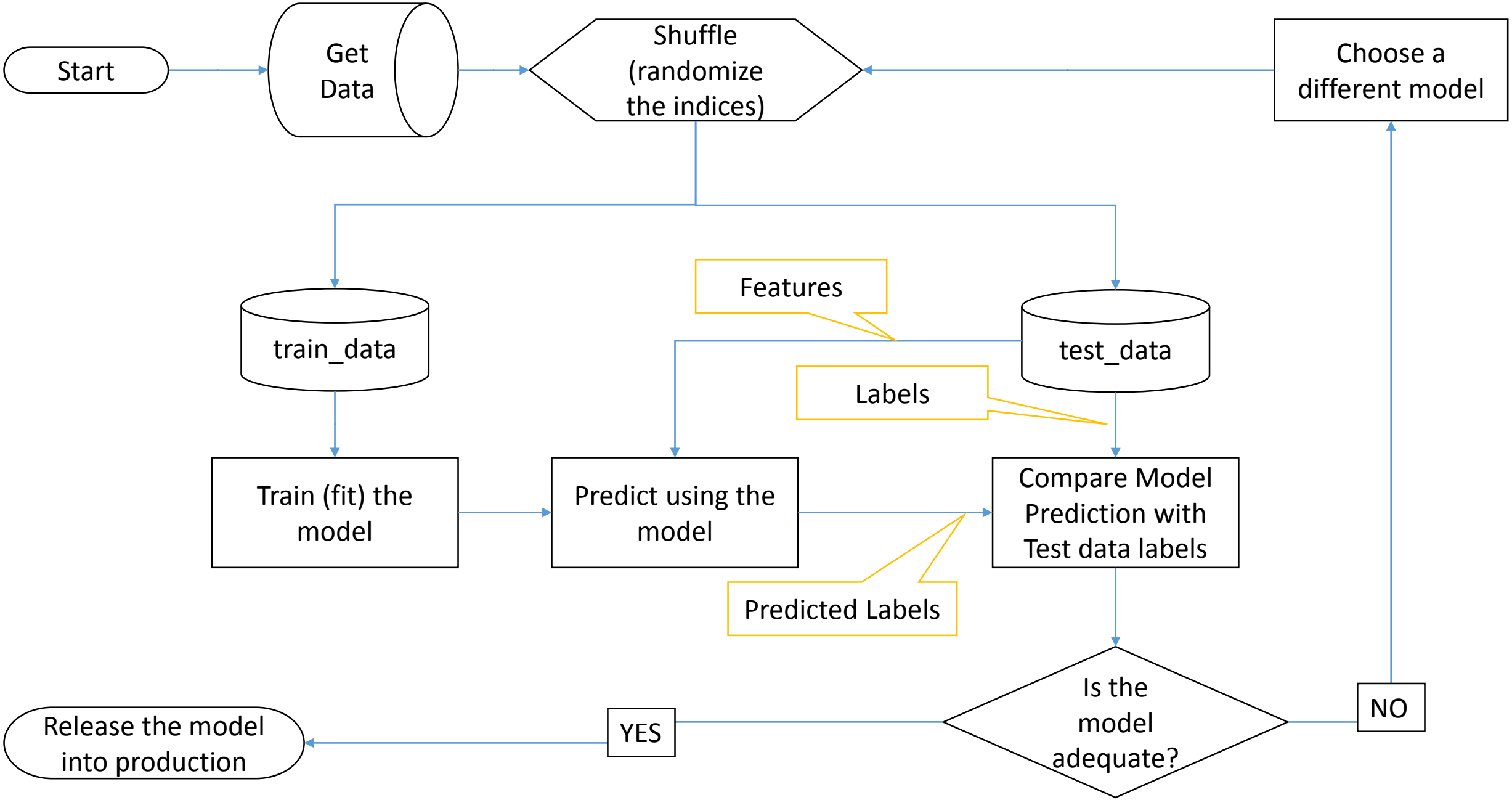
Supervising the Unsupervised

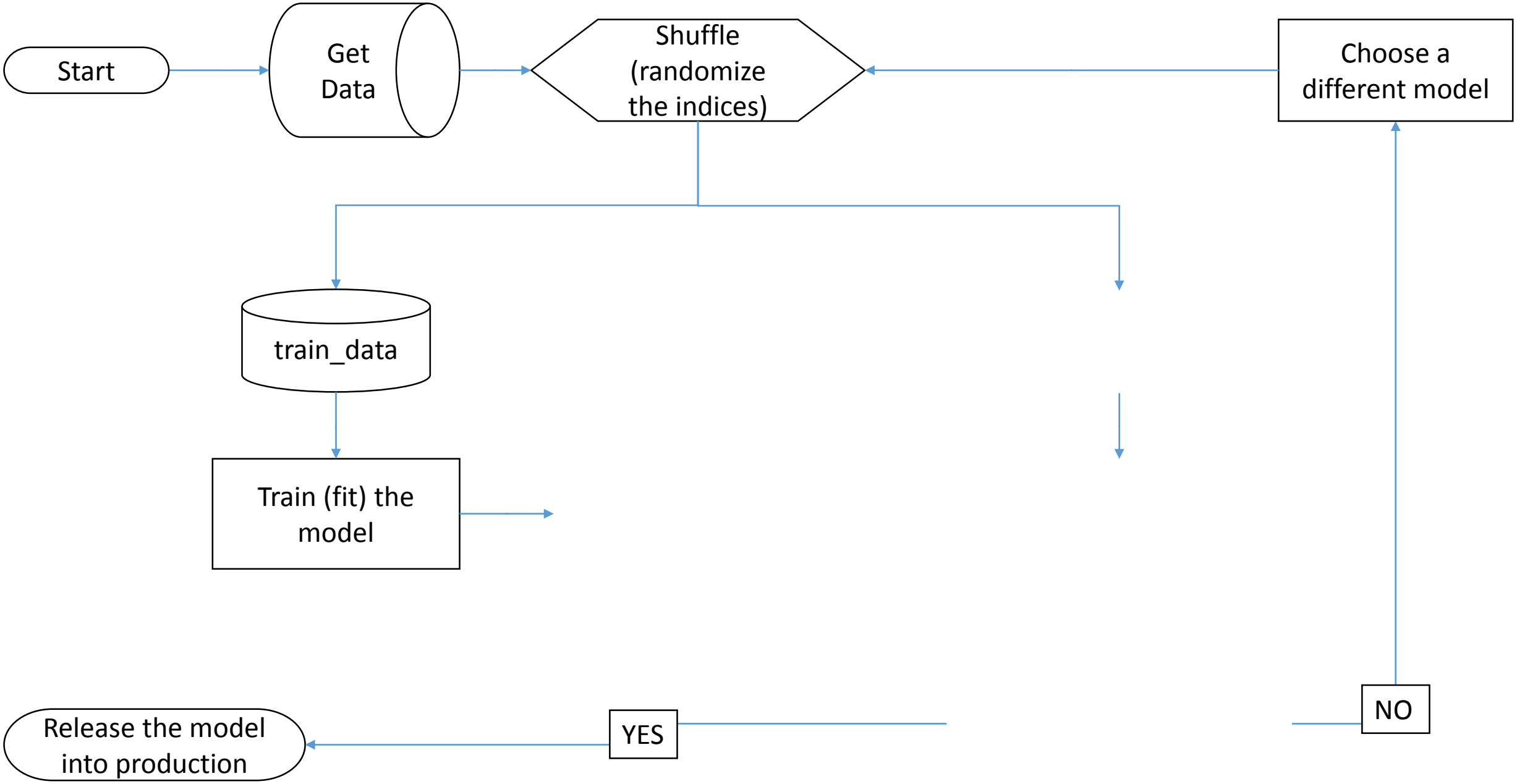
(NOT herding the kittens)

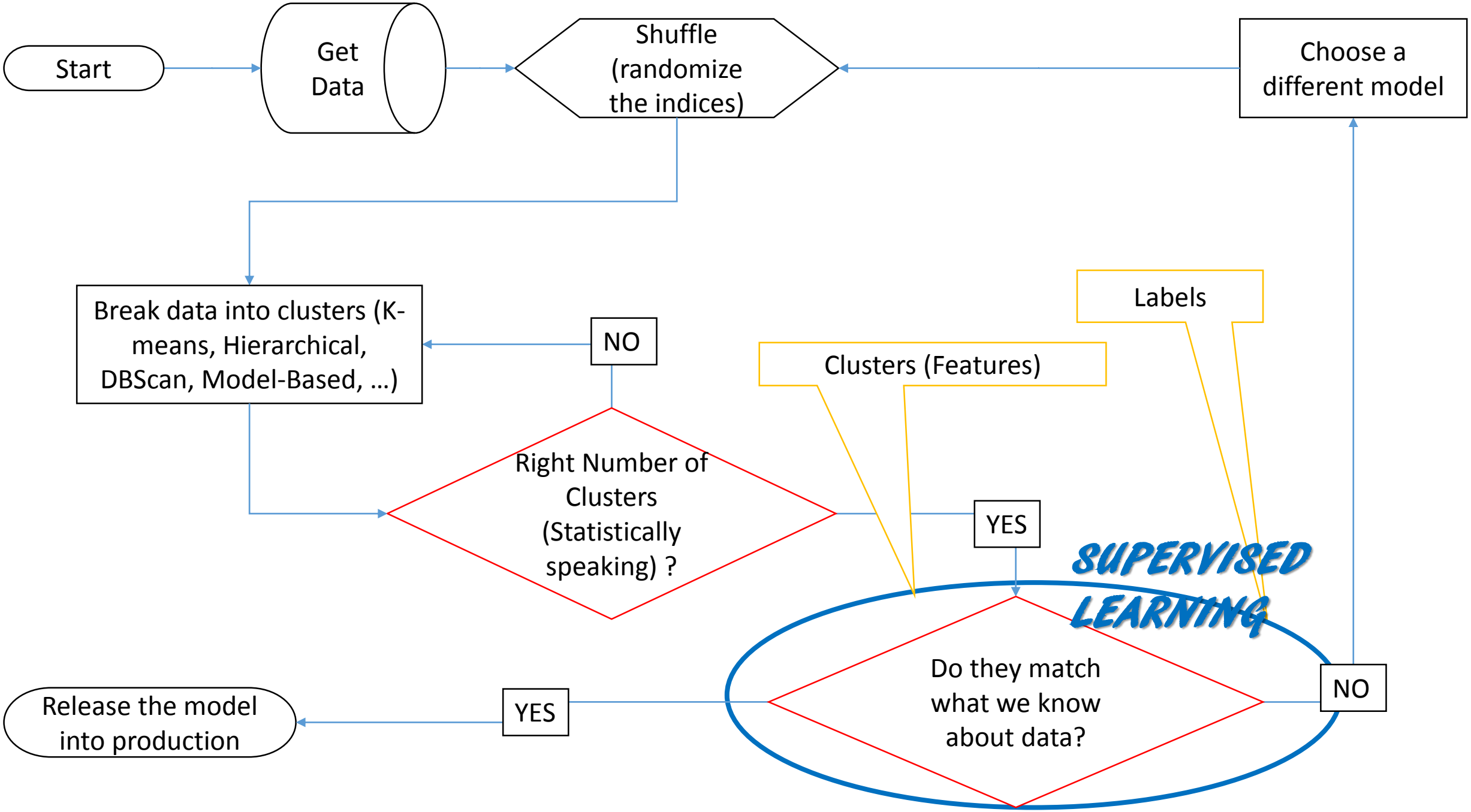

Start
Get Data
Shuffle (randomize the indices)
Choose a different model
train_data
Features
test_data
Labels
Train (fit) the model
Predict using the model
Compare Model Prediction with Test data labels
Predicted Labels
Is the model adequate?
NO
YES
Release the model into production

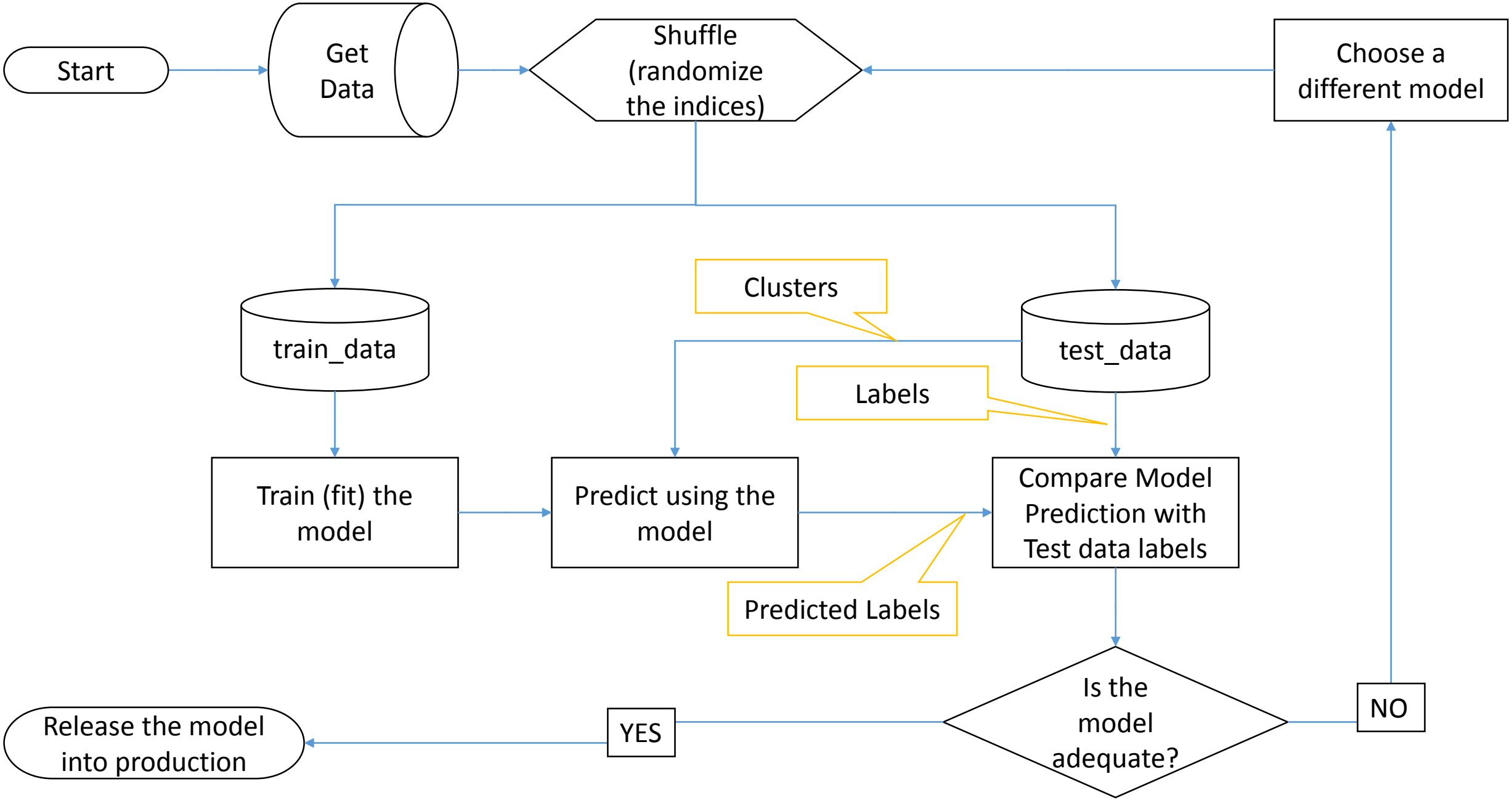

Start
Get Data
Shuffle (randomize the indices)
Choose a different model
train_data
Features
test_data
Labels
Train (fit) the model
Predict using the model
Compare Model Prediction with Test data labels
Predicted Labels
Is the model adequate?
NO
YES
Release the model into production

Unsupervised Learning: Clustering


Start
Get Data
Shuffle (randomize the indices)
Choose a different model
train_data
Features
test_data
Labels
Train (fit) the model
Predict using the model
Compare Model Prediction with Test data labels
Predicted Labels
Is the model adequate?
NO
YES
Release the model into production


Start
Get Data
Shuffle (randomize the indices)
Choose a different model
train_data
Train (fit) the model
NO
YES
Release the model into production


Start
Get Data
Shuffle (randomize the indices)
Choose a different model
Break data into clusters (K-means, Hierarchical, DBScan, Model-Based, …)
NO
Labels
Clusters (Features)
Right Number of Clusters (Statistically speaking) ?
YES
SUPERVISED LEARNING
Do they match what we know about data?
NO
YES
Release the model into production


Start
Get Data
Shuffle (randomize the indices)
Choose a different model
train_data
Clusters
test_data
Labels
Train (fit) the model
Predict using the model
Compare Model Prediction with Test data labels
Predicted Labels
Is the model adequate?
NO
YES
Release the model into production

Scientific Method

STEP 1: Ask a question

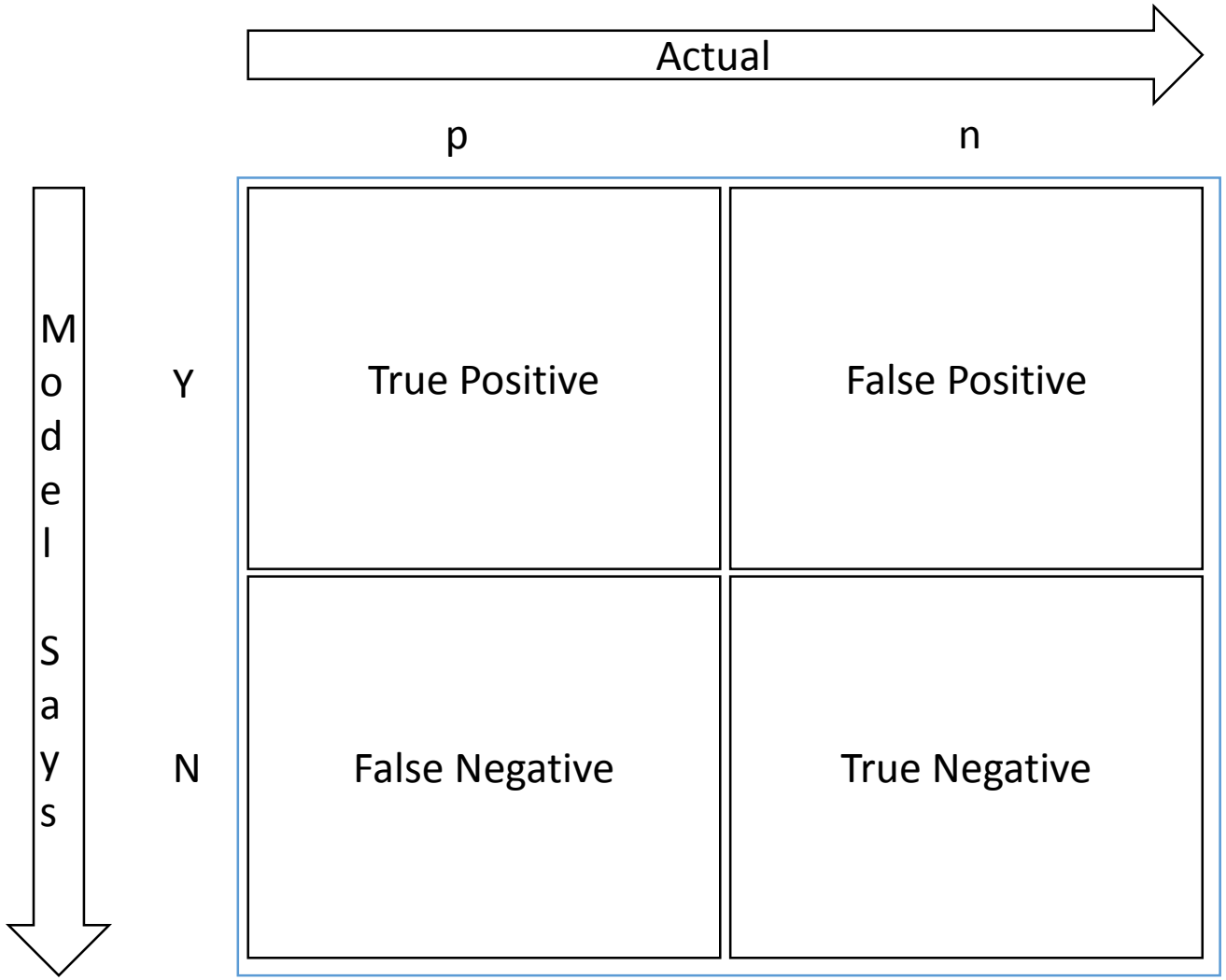
STEP 2: State a hypothesis

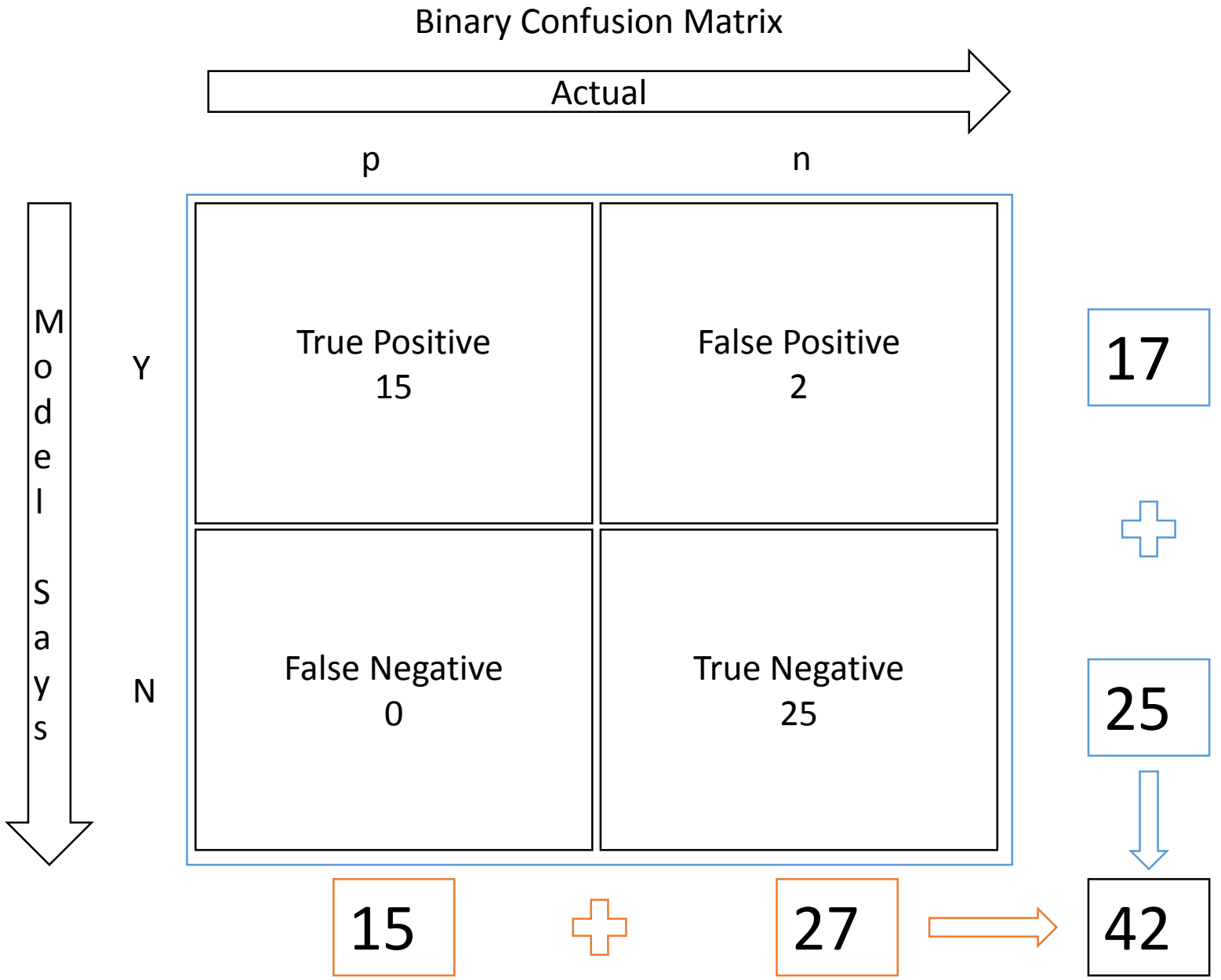
STEP 3: Conduct an experiment

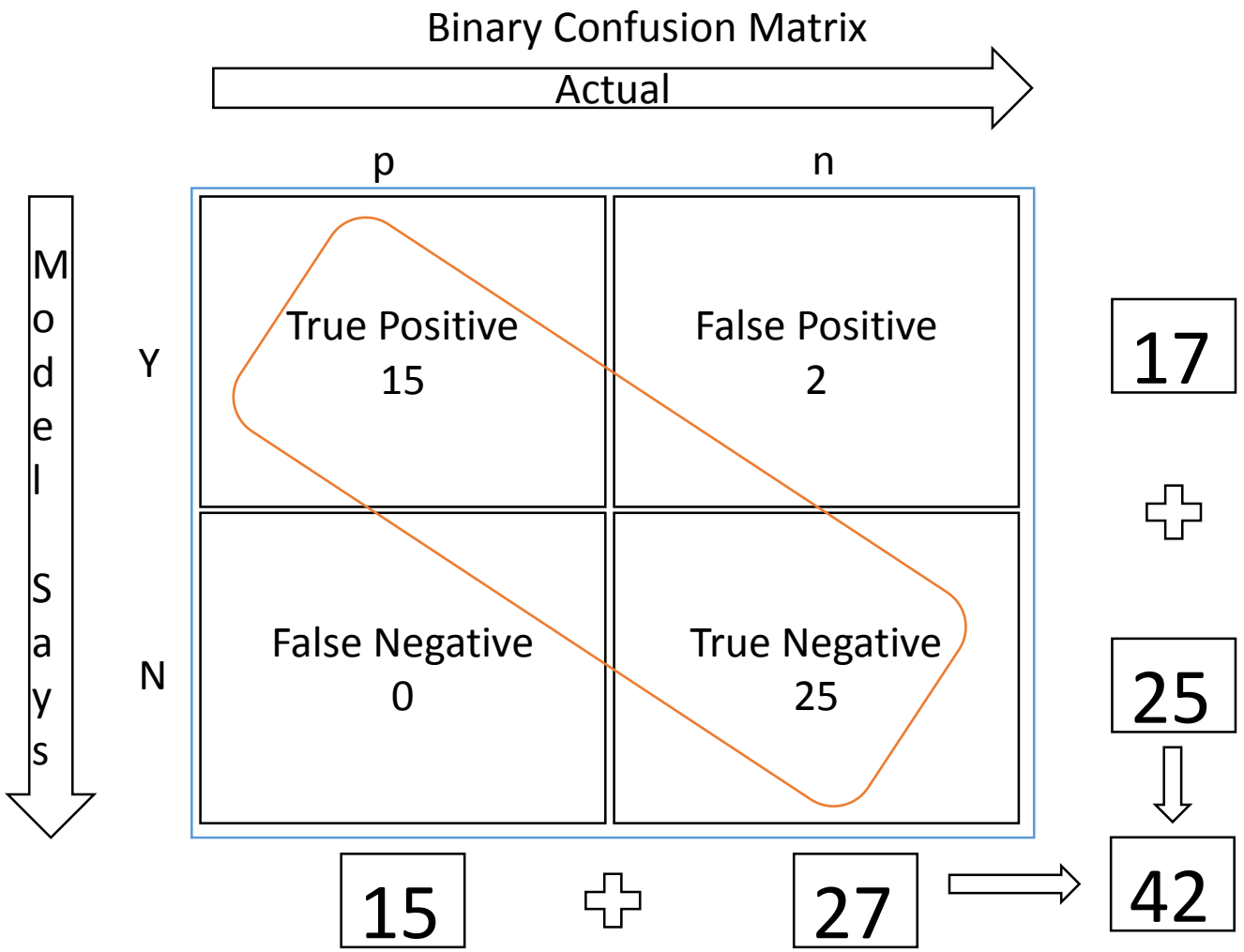
STEP 4: Analyze the results

STEP 5: Make a conclusion

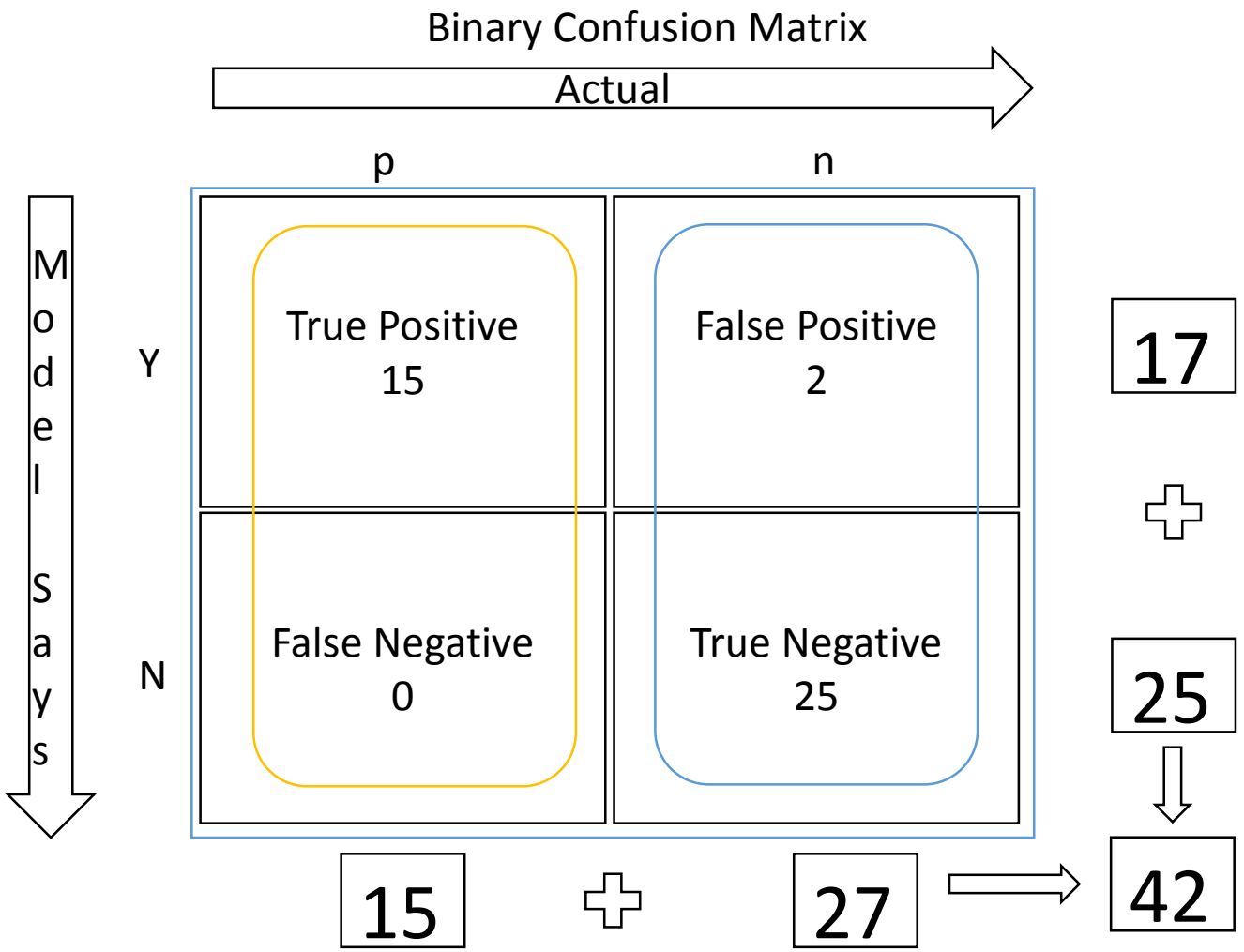
Binary Confusion Matrix




Binary Confusion Matrix
Actual
p
n
Model Says
Y
N
True Positive
15
False Positive
2
False Negative
0
True Negative
25
17
25
15
27
42



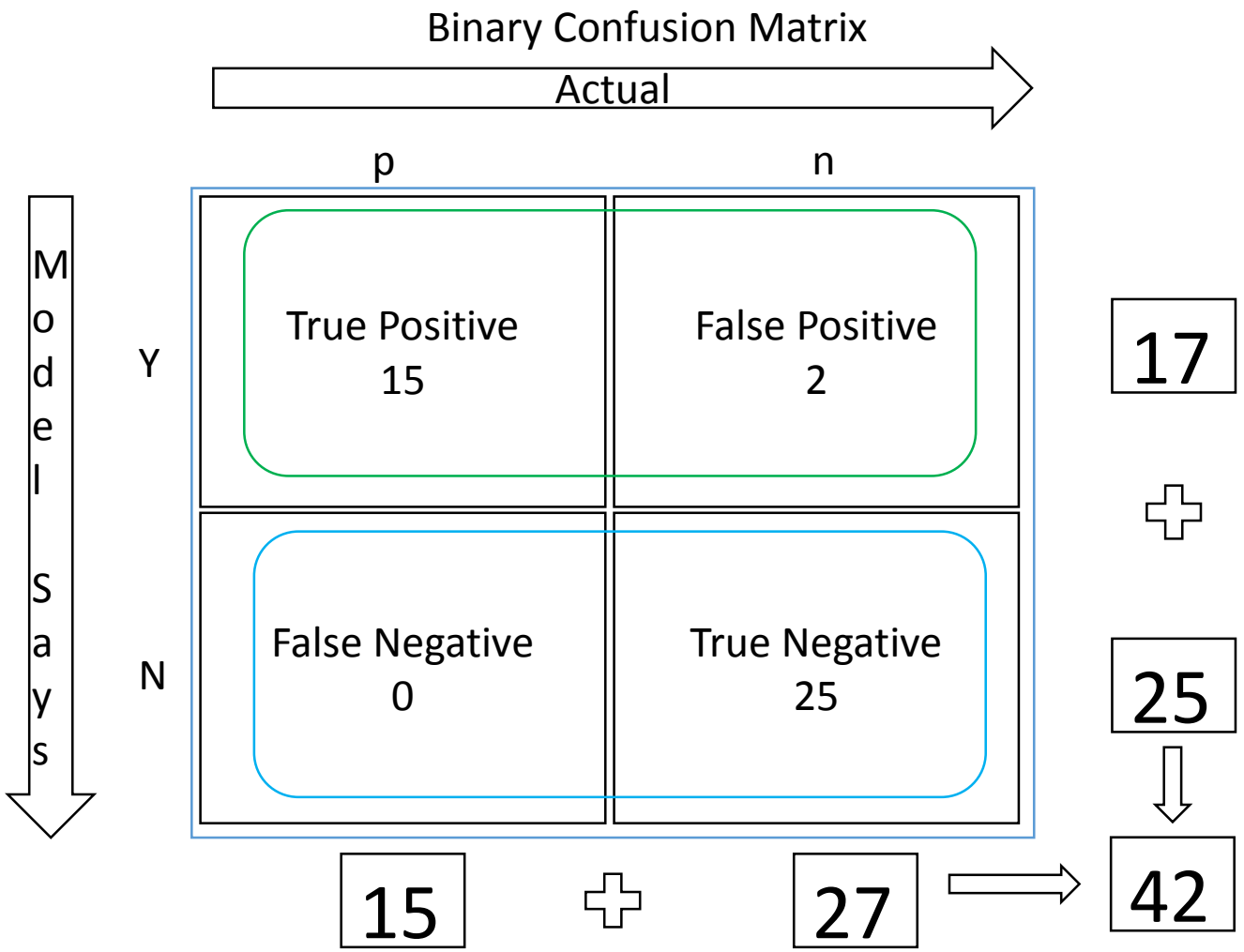
$$Accuracy = \frac{TP + TN}{P + N}$$



$$Accuracy = \frac{TP + TN}{P + N}$$

$$Sensitivity = TPR = \frac{TP}{TP + FN}$$

$$Specificity = TNR = \frac{TN}{TN + FP}$$



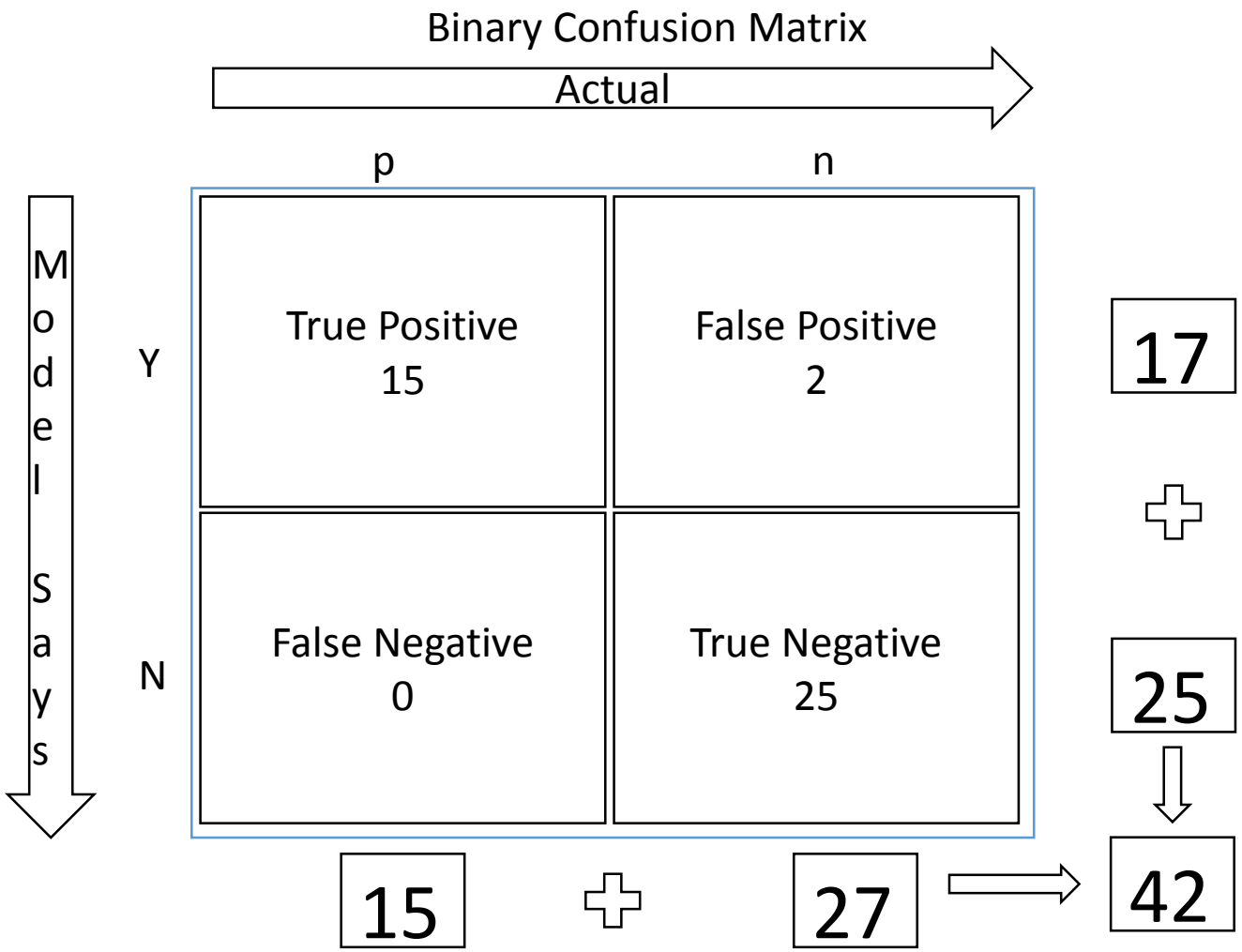
$$Accuracy = \frac{TP + TN}{P + N}$$

$$Sensitivity = TPR = \frac{TP}{TP + FN}$$

$$Specificity = TNR = \frac{TN}{TN + FP}$$

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

$$Recall = \frac{TP}{TP + FN}$$



$$Accuracy = \frac{TP + TN}{P + N}$$

$$Sensitivity = TPR = \frac{TP}{TP + FN}$$

$$Specificity = TNR = \frac{TN}{TN + FP}$$

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

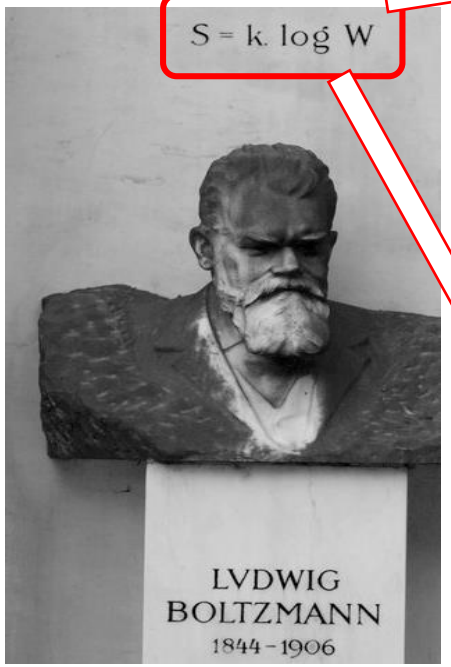
$$Recall = \frac{TN}{TN + FN}$$

$$\mathrm{F_{measure}} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Entropy

$$W = \{w : w \in T(v)\}$$

$$H = -\sum_{w \in W} p(w) \log_2 p(w)$$



Confusion and Diffusion

Cl Shannon

Confusion

The relationship between the key and the ciphertext as complex and as involved as possible.

e.g. Enigma & complex substitution (S-boxes)

011011

Diffusion

Statistics of the plaintext is "dissipated" in the statistics of the ciphertext. If we change a character of the plaintext, then several characters of the ciphertext should change.

http://en.wikipedia.org/wiki/Permutation_box

P-Box

$$H_t = -\sum_{i=1}^{n} \pi_{it} \ln \pi_{it}$$

Progress imposes not only new possibilities for the future but new restrictions. It seems almost as if progress itself and our fight against the increase of entropy intrinsically must end in the downhill path from which we are trying to escape.

— Norbert Wiener —

$$\Delta S = S(set) - S(set|\theta) \Leftrightarrow \Delta H = H(set) - H(set|\theta)$$

$$= -\left[\sum_{i=1}^{N} p_i * \log_{10} p_i\right] - \left\{-\left[\sum_{i=1}^{N} p_{i|\theta} * \log_{10} p_{i|\theta}\right]\right\}$$

$$\sim -\left[\sum_{i=1}^{N} p_i * \log_2 p_i\right] - \left\{-\left[\sum_{i=1}^{N} p_{i|\theta} * \log_2 p_{i|\theta}\right]\right\}$$

$$\sim -\left[\sum_{i=1}^{N} p_i * ln\, p_i\right] - \left\{-\left[\sum_{i=1}^{N} p_{i|\theta} * ln\, p_{i|\theta}\right]\right\}$$

As long as you stay consistent with the base of the logs, you are doing the right thing.