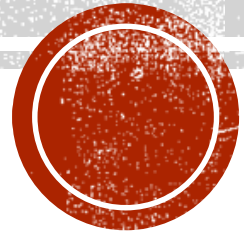
EVALUATION

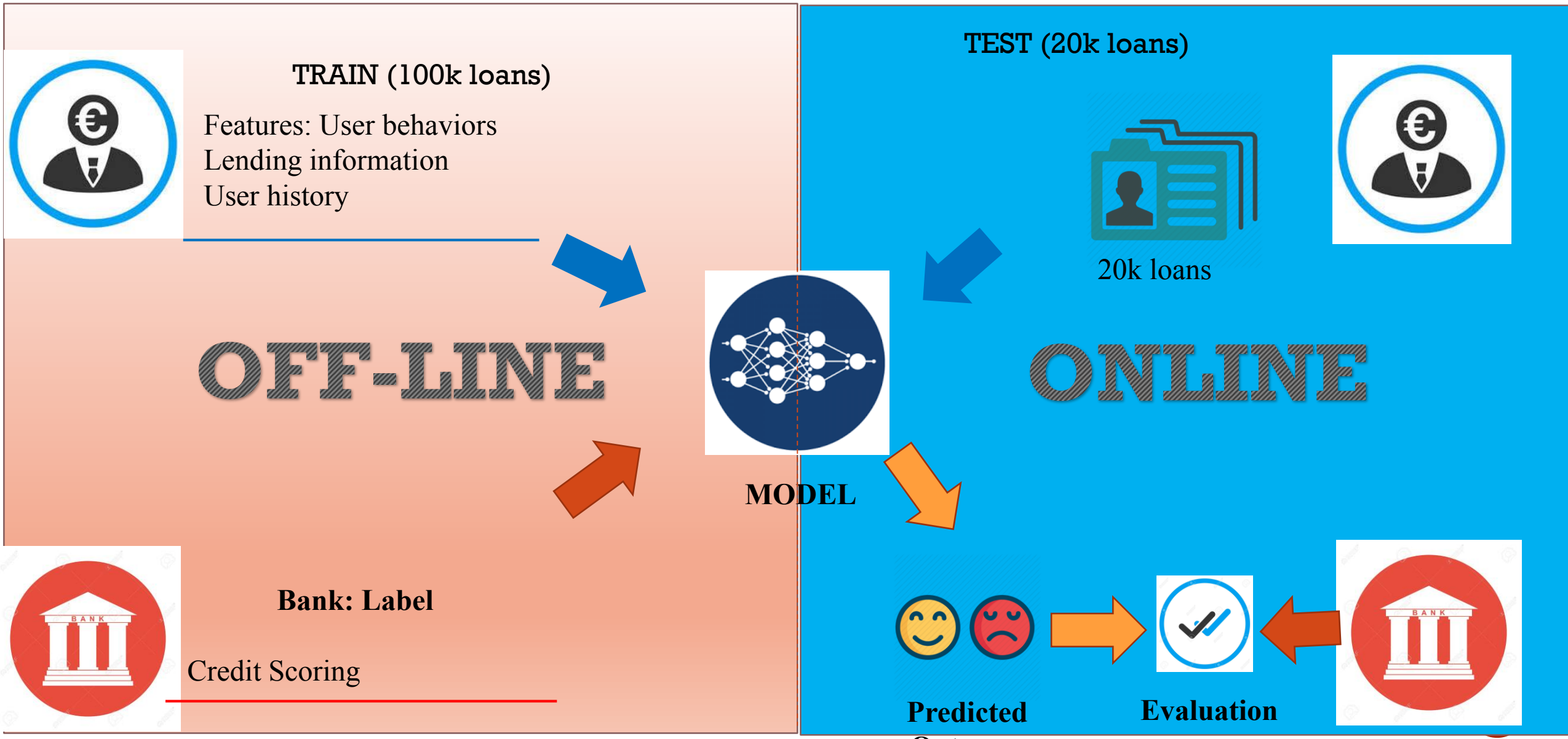
Sonpvh.2019.05.May



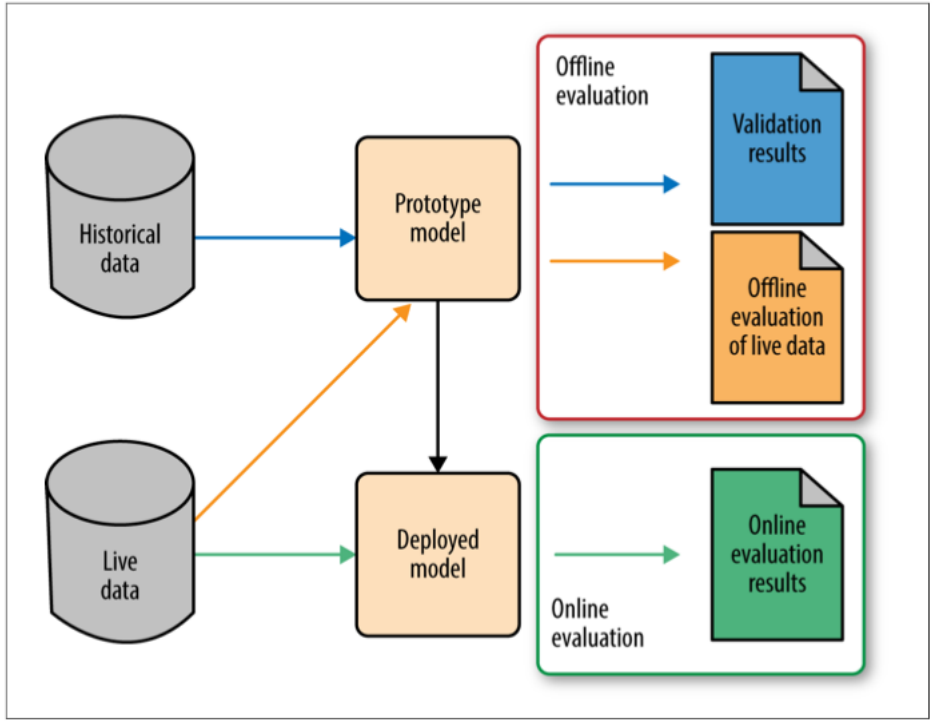
OUTLINE

1. The machine learning workflow
2. Evaluation Metrics
3. Offline Evaluation mechanisms
4. Hyperparameter turning
5. A/B Testing
6. Casual-Effect

1. MACHINE LEARNING WORKFLOW



1. MACHINE LEARNING WORKFLOW [1]



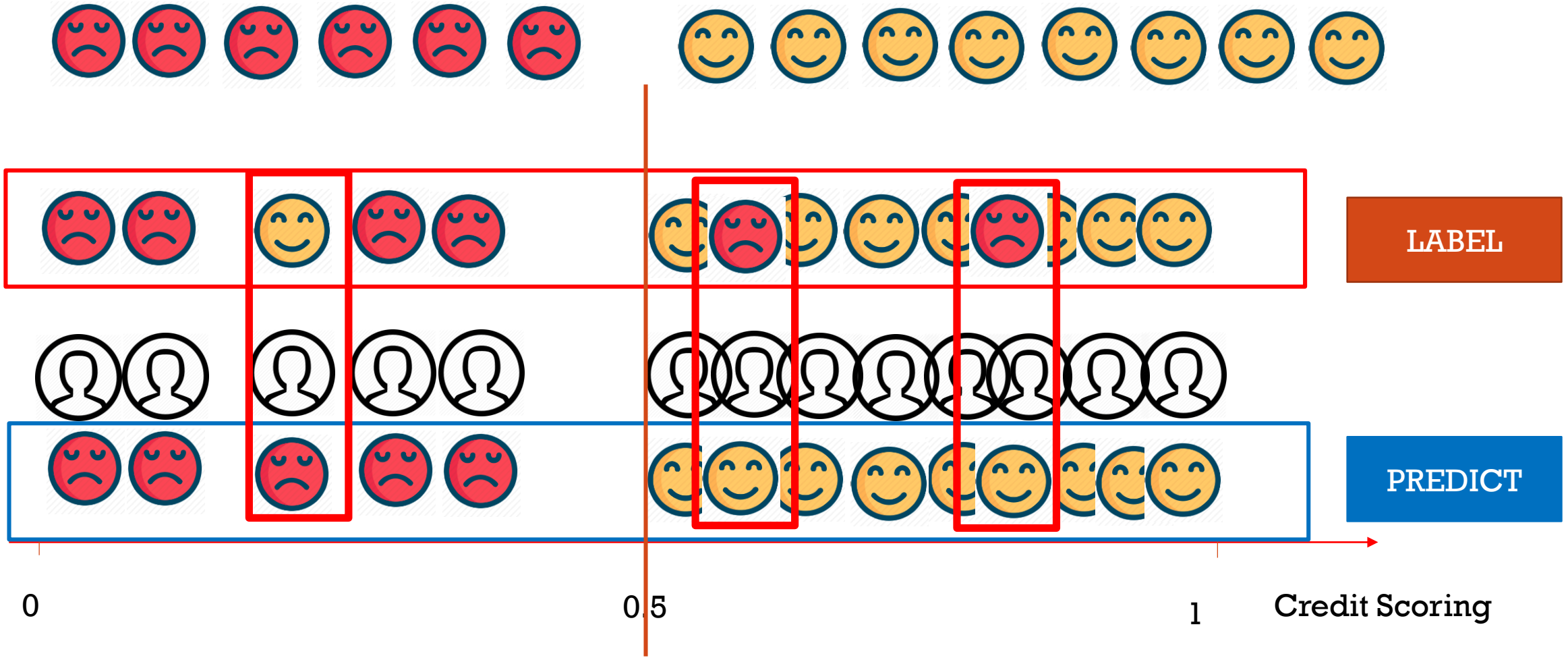
Why is it so complicated?

1. **Offline evaluation**: accuracy, precision, recall, MSE …

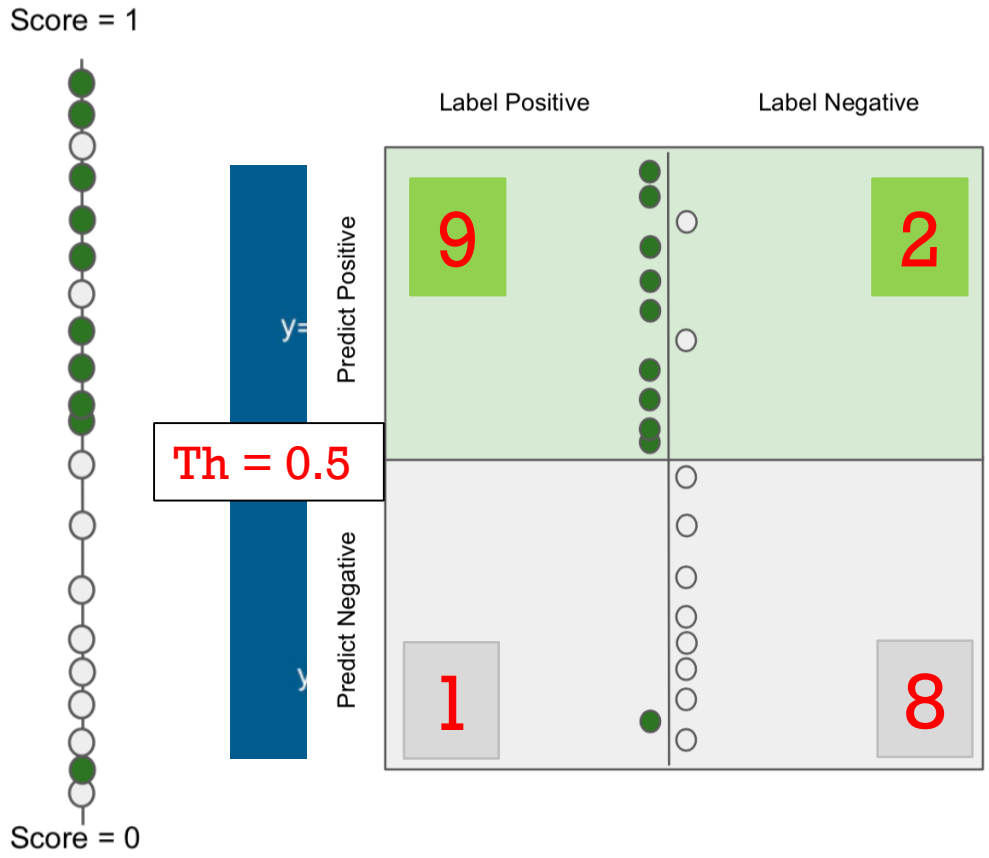
   Online evaluation: business metrics

2. Distribution drift: the distribution of data changes overtime, so keep track the models performance on the validation metrics of live data.

2. EVALUATION METRICS: BINARY CLASSIFICATION[2]



2. EVALUATION METRICS: BINARY CLASSIFICATION [2]



Confusion Matrix

Th	TP	TN	FP	FN	Acc	Pre	Recall	F
0.5	9	8	2	1	0.85	0.81	0.9	0.85

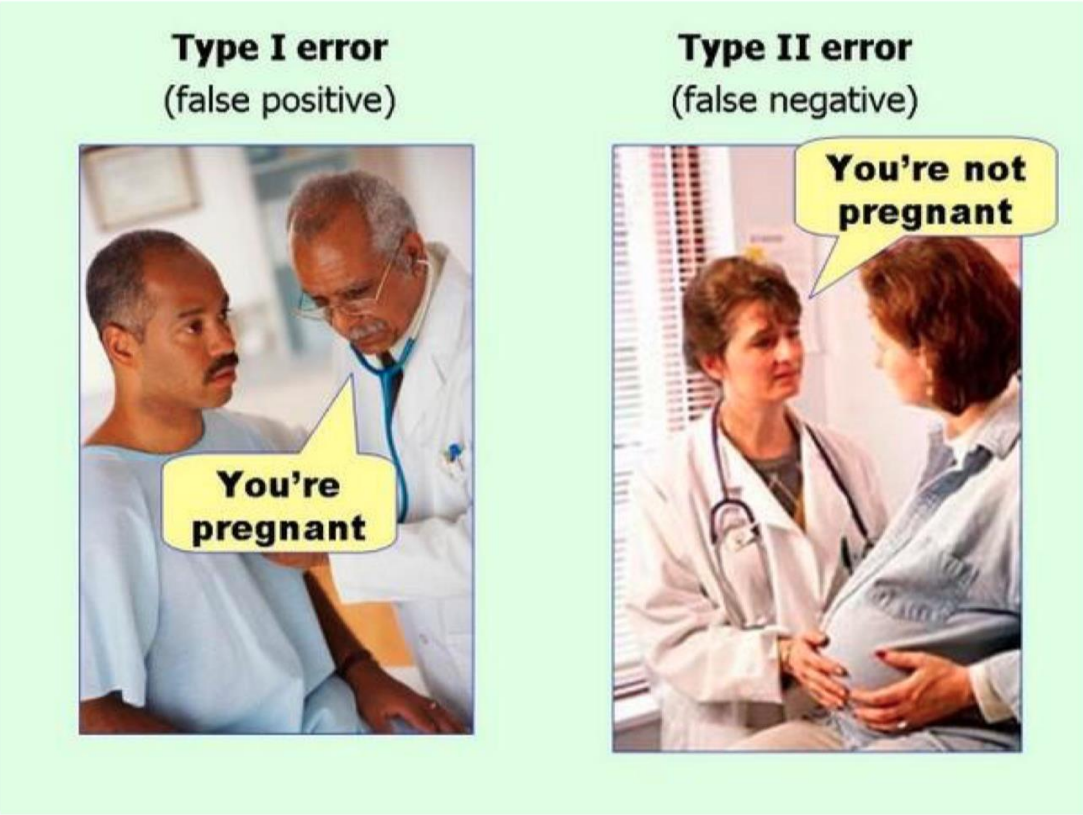
$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

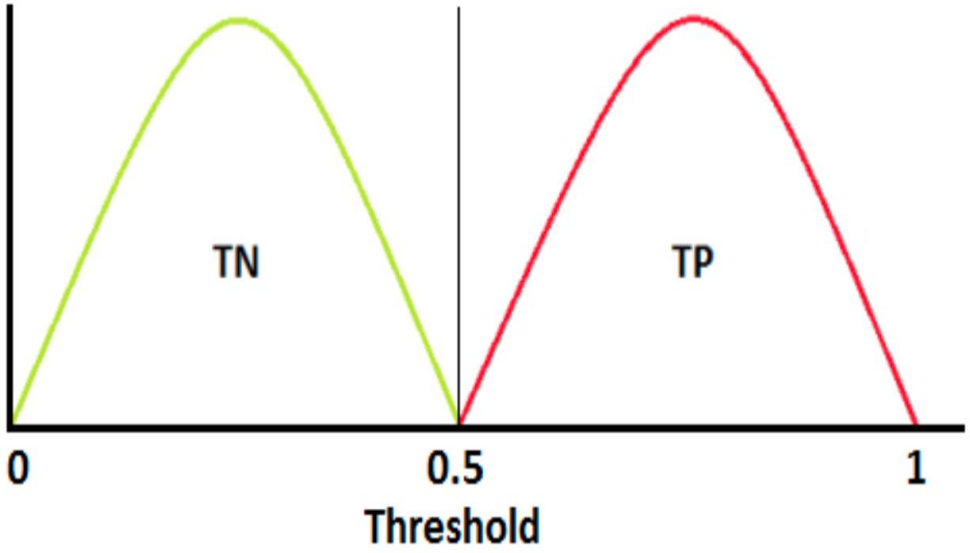
2. EVALUATION METRICS [2]

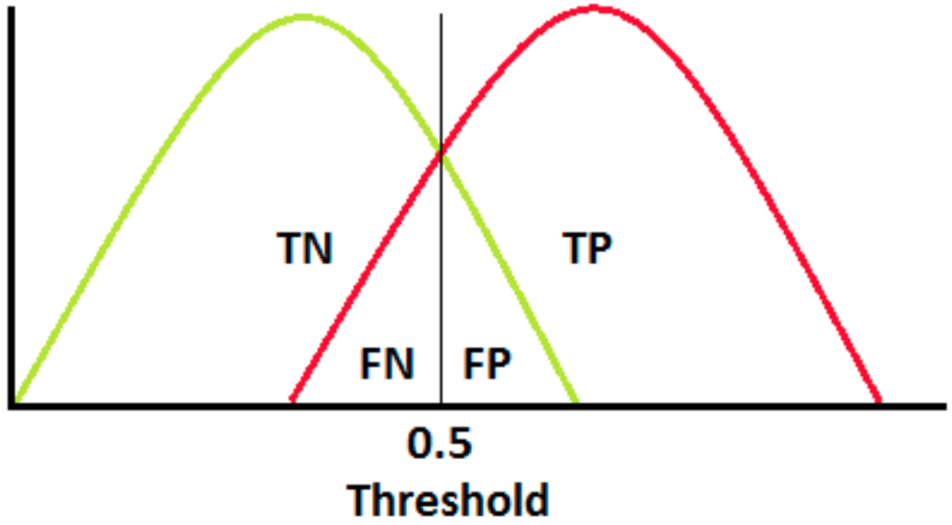


2. EVALUATION METRICS: CHANGING THRESHOLD [2]

Threshold	TP	TN	FP	FN	Accuracy	Precision	Recall	Specificity	F1
1.00	0	10	0	10	0.50	1	0	1	0
0.95	1	10	0	9	0.55	1	0.1	1	0.182
0.90	2	10	0	8	0.60	1	0.2	1	0.333
0.85	2	9	1	8	0.55	0.667	0.2	0.9	0.308
0.80	3	9	1	7	0.60	0.750	0.3	0.9	0.429
0.75	4	9	1	6	0.65	0.800	0.4	0.9	0.533
0.70	5	9	1	5	0.70	0.833	0.5	0.9	0.625
0.65	5	8	2	5	0.65	0.714	0.5	0.8	0.588
0.60	6	8	2	4	0.70	0.750	0.6	0.8	0.667
0.55	7	8	2	3	0.75	0.778	0.7	0.8	0.737
0.50	8	8	2	2	0.80	0.800	0.8	0.8	0.800
0.45	9	8	2	1	0.85	0.818	0.9	0.8	0.857
0.40	9	7	3	1	0.80	0.750	0.9	0.7	0.818
0.35	9	6	4	1	0.75	0.692	0.9	0.6	0.783
0.30	9	5	5	1	0.70	0.643	0.9	0.5	0.750
0.25	9	4	6	1	0.65	0.600	0.9	0.4	0.720
0.20	9	3	7	1	0.60	0.562	0.9	0.3	0.692
0.15	9	2	8	1	0.55	0.529	0.9	0.2	0.667
0.10	9	1	9	1	0.50	0.500	0.9	0.1	0.643
0.05	10	1	9	0	0.55	0.526	1	0.1	0.690
0.00	10	0	10	0	0.50	0.500	1	0	0.667

2. RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE, AREA UNDER THE ROC (AUC) [3]

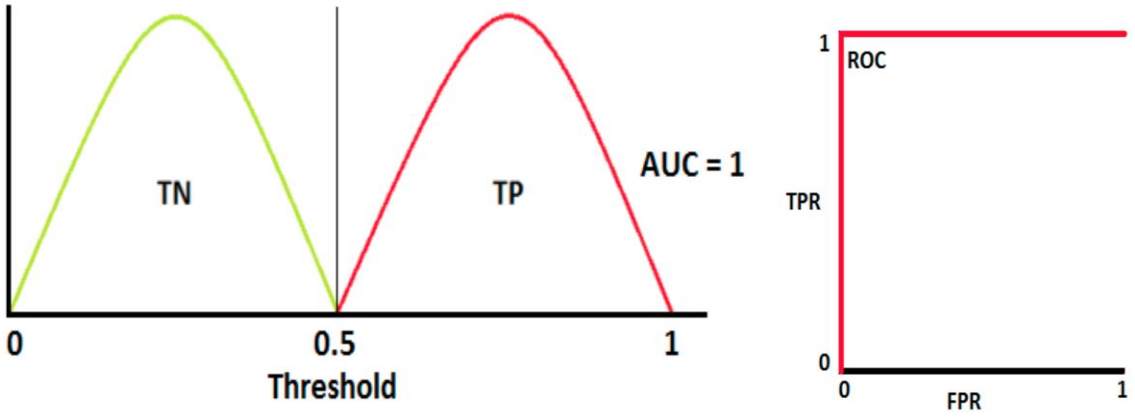


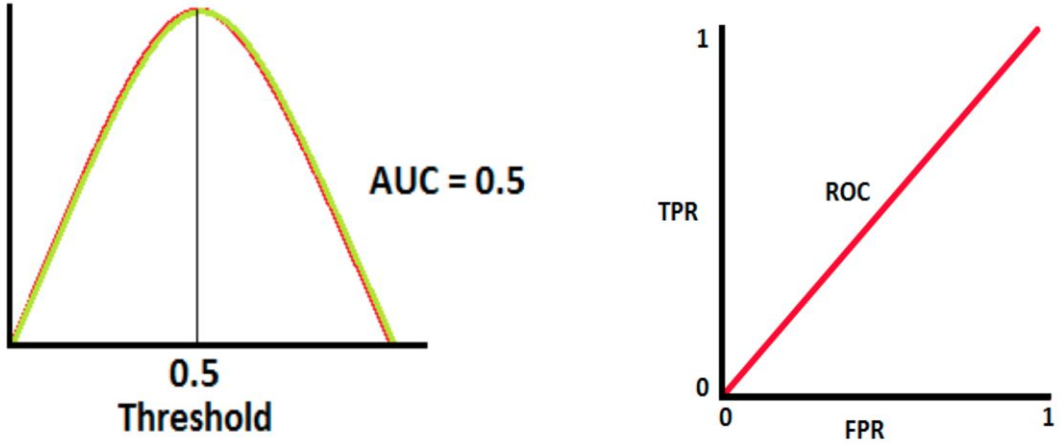


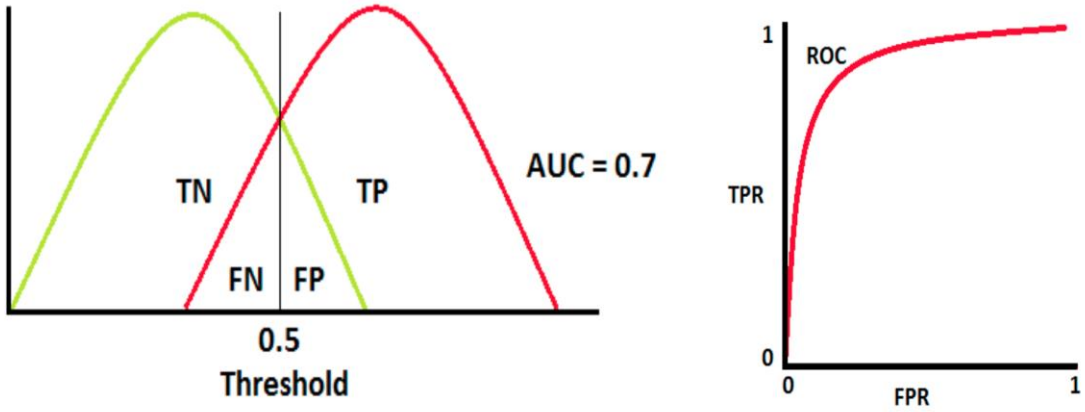
$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \text{RECALL (SENSITIVITY)}$$

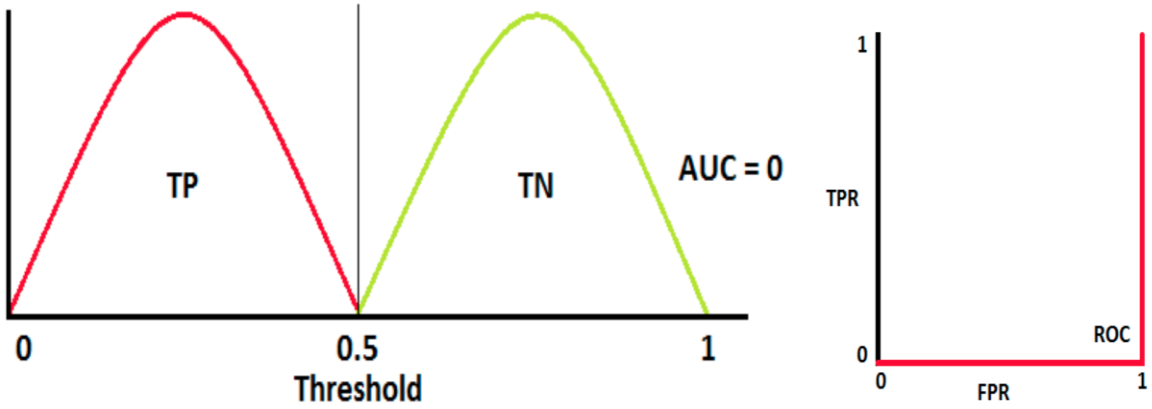
$$\textit{F}\text{PR} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

2. ROC, AUC [3,4]


TN
TP
AUC = 1
0
0.5
1
Threshold
1
ROC
TPR
0
0
FPR
1


AUC = 0.5
0.5
Threshold
1
TPR
ROC
0
0
FPR
1


TN
TP
FN
FP
AUC = 0.7
0.5
Threshold
1
ROC
TPR
0
0
FPR
1


TP
TN
AUC = 0
0
0.5
1
Threshold
1
TPR
ROC
0
0
FPR
1

2. EVALUATION METRICS: REGRESSION

- Ex: prediction…
-

$$\text{RMSE} = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}}$$

$$\text{MAPE} = \text{median}(|(y_i - \hat{y}_i)/y_i|)$$

2. EVALUATION METRICS: CLASSIFICATION

- Ex: spam detection, prostitute detection…

$$\text{accuracy} = \frac{\#\ \text{correct predictions}}{\#\ \text{total data points}}$$

	Predicted as positive	Predicted as negative
Labeled as positive	80	20
Labeled as negative	5	195

Confusion matrix

Per-class accuracy

$$\text{log-loss} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p_i + (1 - y_i) \log (1 - p_i)$$

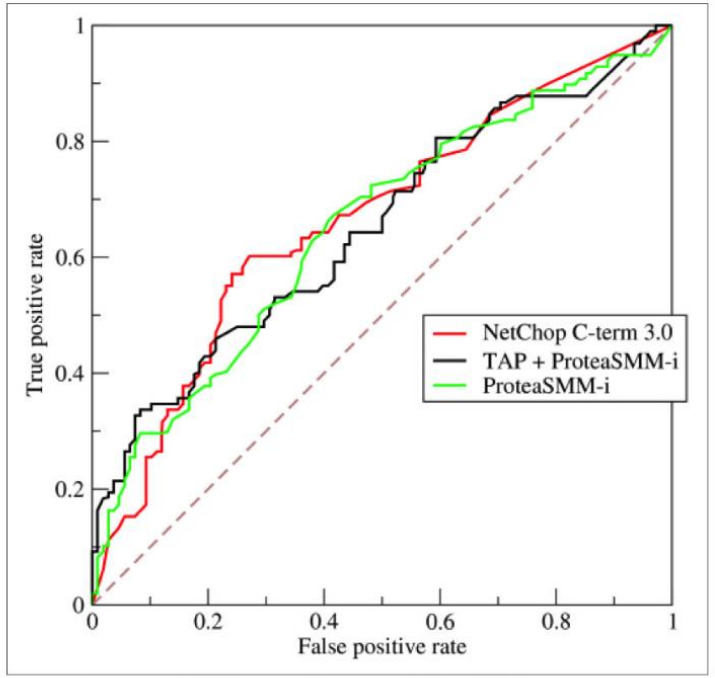


Figure 2-2. Sample ROC curve (source: Wikipedia)

ROC: receiver operating characteristic
AUC: area under the curve

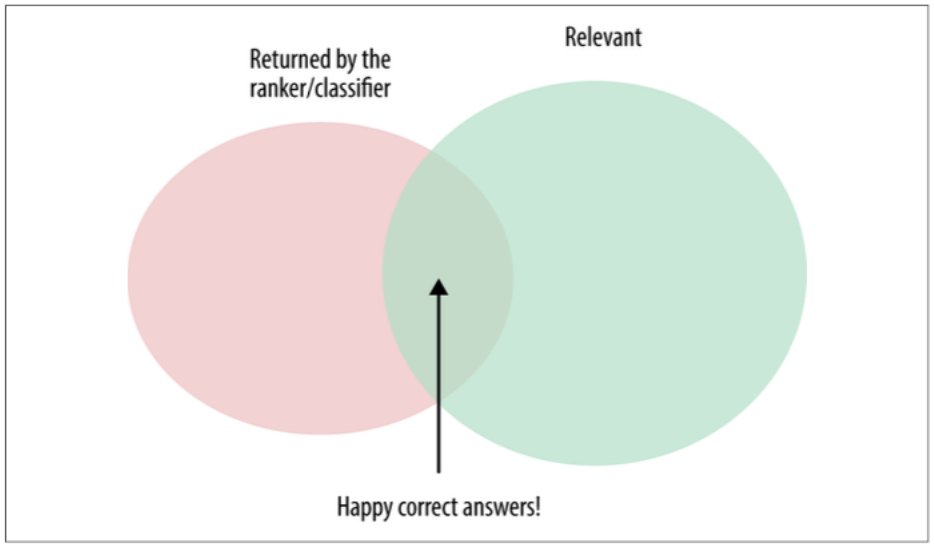
2. EVALUATION METRICS: RANKING

- Ex: search ranker, personalized recommendation ..

		Reality	
		Actually Good	Actually Bad
Prediction	Rated Good	True Positive (tp)	False Positive (fp)
	Rated Bad	False Negative (fn)	True Negative (tn)

All recommended

All good items



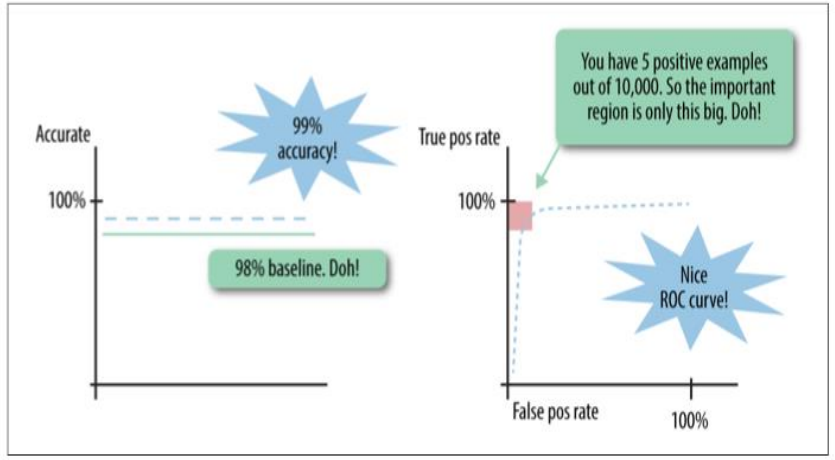
$$\text{precision} = \frac{\#\text{ happy correct answers}}{\#\text{ total items returned by ranker}}$$

$$\text{recall} = \frac{\#\text{ happy correct answers}}{\#\text{ total relevant items}}$$

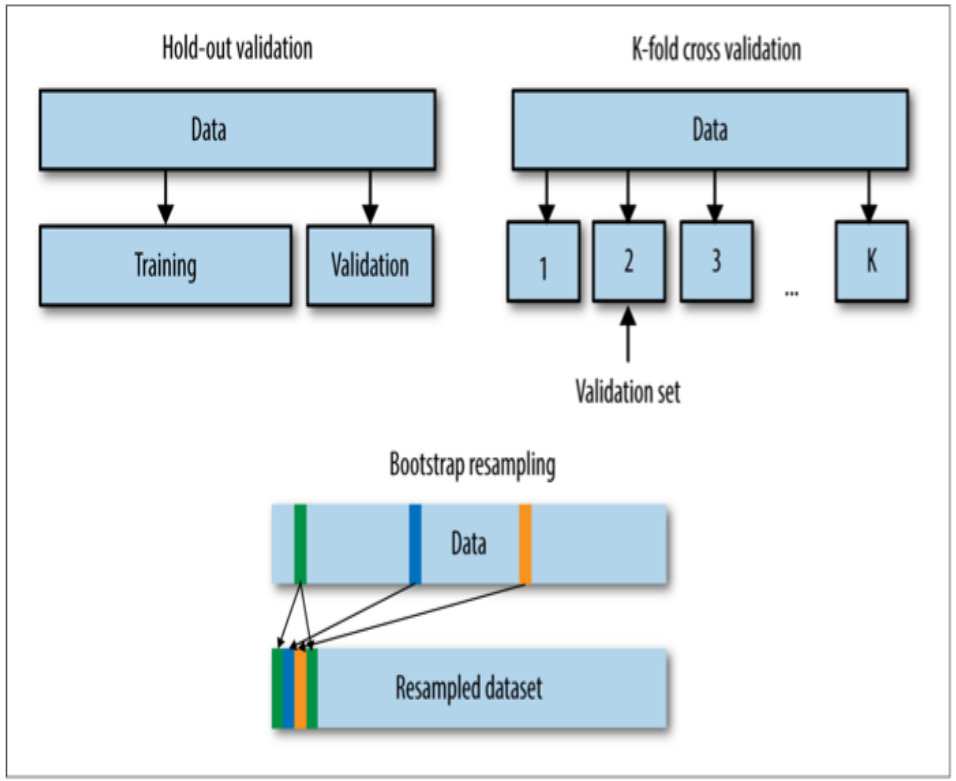
"The precision is the proportion of recommendations that are good recommendations, and recall is the proportion of good recommendations that appear in top recommendations."

2. EVALUATION METRICS: BEST PRACTICE

- Evaluation metrics # model log loss function: Train a personalized recommender by minimizing the loss between its predictions and observed ratings, and then use this recommender to produce a ranked list of recommendations. AVOID
- Skewed data, imbalanced, classes, outliers, rare data: analysis carefully before doing anything else



3. OFFLINE EVALUATION MECHANISM



Cross validation:
Independently and Identically distributed

4. HYPER-PARAMETER TURNING

- Model parameter: $y = W^T x$
- Hyper-parameter (nuisance parameters): optimization state.
- Ex:
- Linear regression: regularization parameter,
- Decision trees: desired depth and number of leaves.
- SVMs: misclassification penalty term
- ..

5. A/B TESTING AND ITS PITFALLS

1. Split into randomized control/experimentation groups.
2. Observe behavior of both groups on the proposed methods.
3. Compute test statistics.
4. Output decision.

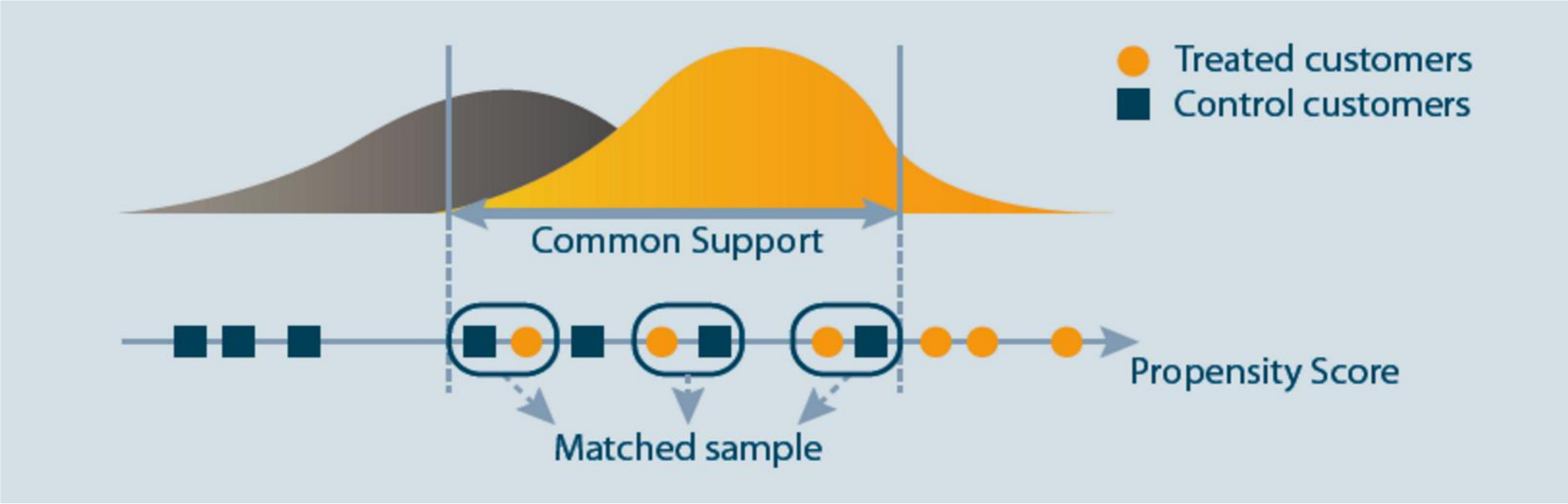
5. A/B TESTING AND ITS PITFALLS

1. Baggage of the old: should do A/A testing first
2. Choose metrics, indexes (business design)
3. Did you count right?
4. How many observations do you need?
5. Is the distribution of the metric Gaussian?
6. Variances equal?
7. Multiple models, multiple hypotheses: A/A1/A2/…/B testing
8. How long to run the test?
9. Catching distribution drift: stationarity assumption

6. PSM - PROPENSITY SCORE MATCHING METHOD

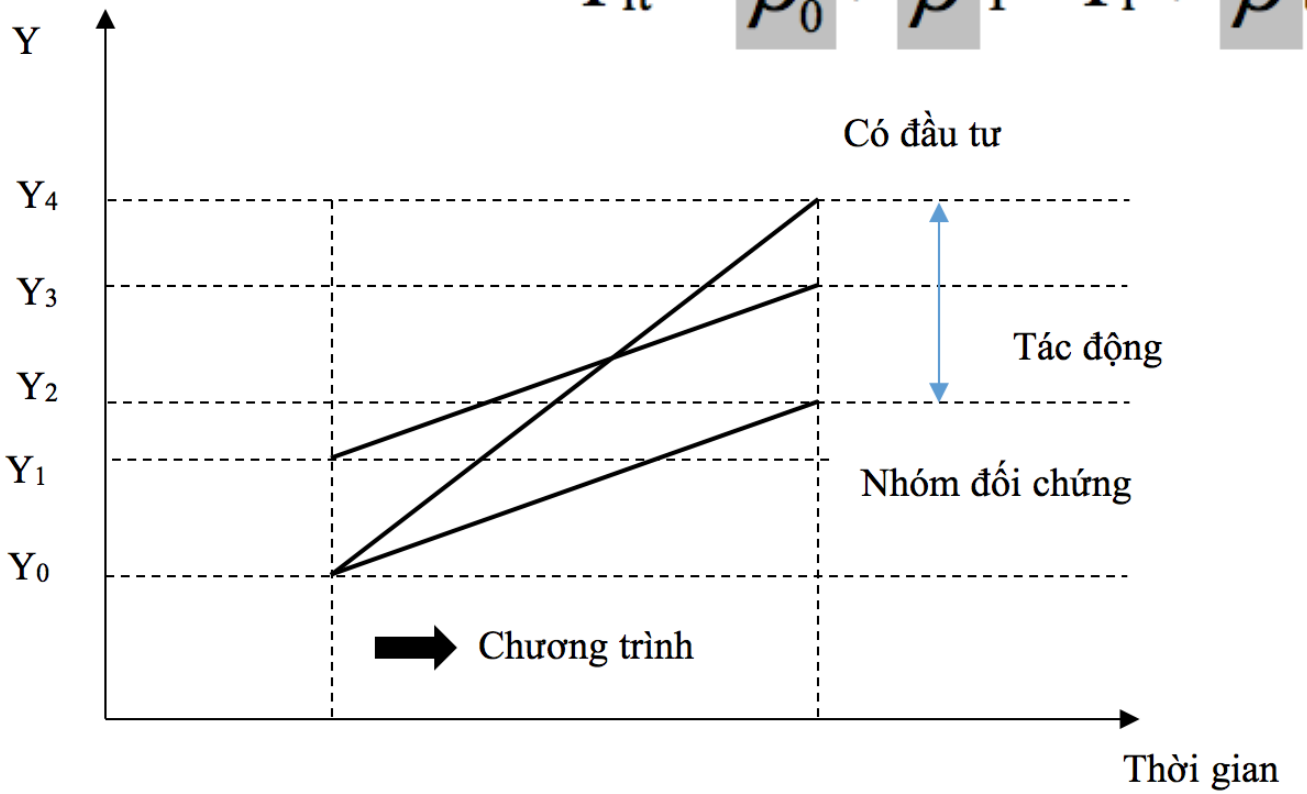
1. (Conditional) independence
2. Common support

3. TOT = $E_{P(X)|T=1}$ {$E[Y^{(1)}|T=1, P(X)] - E[Y^{(0)}|T=0, P(X)]$}



6. DID - DIFFERENCE IN DIFFERENCE

$$Y_{it} = \beta_0 + \beta_T * T_i + \beta_t * t + \beta_{iT} * t * T_i + \sigma_i \quad (6)$$



REFERENCES:

1. Alice Zheng - Evaluating Machine Learning Models - O'Reilly Media, Inc. 2015
2. http://cs229.stanford.edu/section/evaluation_metrics.pdf
3. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
4. http://www.navan.name/roc/