Matching

Lecture 2: Production realities

Anton Ryabtsev

Moscow Institute of Physics and Technology

Autumn 2023

A day in the life of a marketplace

TODAY

1. Velomobile

76 925 ₽

Целевая аудитория	Детская
Материал рамы	Сталь
Форма поставки	В разобранном виде
Макс. нагрузка, кг	50
Вес в собранном состоянии, кг	18

2. Spark plug 4T A7TC

183 ₽

Фото временно отсутствует

Тип	Свеча зажигания
Бренд	TMMP
Вид техники	Мотоциклы, Мопеды, Скутеры

A day in the life of a marketplace

TODAY

1. Velomobile



76 925 ₽

Целевая аудитория	Детская
Материал рамы	Сталь
Форма поставки	В разобранном виде
Макс. нагрузка, кг	50
Вес в собранном состоянии, кг	18

2. Spark plug 4T A7TC

183 ₽

Фото временно отсутствует

Тип	Свеча зажигания
Бренд	TMMP
Вид техники	Мотоциклы, Мопеды, Скутеры

TOMORROW

1. Velomobile Berg Buddy Fendt



76 925 ₽

Целевая аудитория	Детская
Материал рамы	Сталь
Форма поставки	В разобранном виде
Макс. нагрузка, кг	50
Вес в собранном состоянии, кг	18

2. Spark plug 4T A7TC

196 ₽

Партномер (артикул производителя)	16105775
Артикул	16106062
Тип	Свеча зажигания
Бренд	TMMP
Вид техники	Мотоциклы, Мопеды, Скутеры

3. Greenfield Green Ginseng Tea

98 ₽

Бренд	Greenfield
Вид чая	Зеленый

Problem Statement In Practice

TODAY

Given:

$I^t = \{i_1, i_2, \cdots, i_N\}, N > 10^8 -$ a set of items for today

$D^t = \{d_1^t, d_2^t, \cdots, d_N^t\} -$ a set of data describing the product for today

Find:

$M^t = \|m_{i,j}\|^t -$ matches matrix of shape $N \times N,\ m_{i,j} \in \{0, 1\}$

Result:

$\tilde{M}^t = \|\tilde{m}_{i,j}\|^t -$ matches matrix of shape $K^t \times K^t,\ \ K^t \ll N$

Problem Statement In Practice

TODAY

Given:

$I^t = \{i_1, i_2, \cdots, i_N\}, N > 10^8 -$ a set of items for today

$D^t = \{d_1^t, d_2^t, \cdots, d_N^t\} -$ a set of data describing the product for today

Find:

$M^t = \|m_{i,j}\|^t -$ matches matrix of shape $N \times N,\ m_{i,j} \in \{0, 1\}$

Result:

$\tilde{M}^t = \|\tilde{m}_{i,j}\|^t -$ matches matrix of shape $K^t \times K^t,\ \ K^t \ll N$

TOMORROW

Given:

$I^{t+1} = I^t + \{i_{N+1}, \cdots, i_L\} -$ a set of items for tomorrow

$D^{t+1} = \{d_1^{t+1}, d_2^{t+1}, \cdots, d_N^{t+1}\} -$ a set of data describing the product for tomorrow

Find:

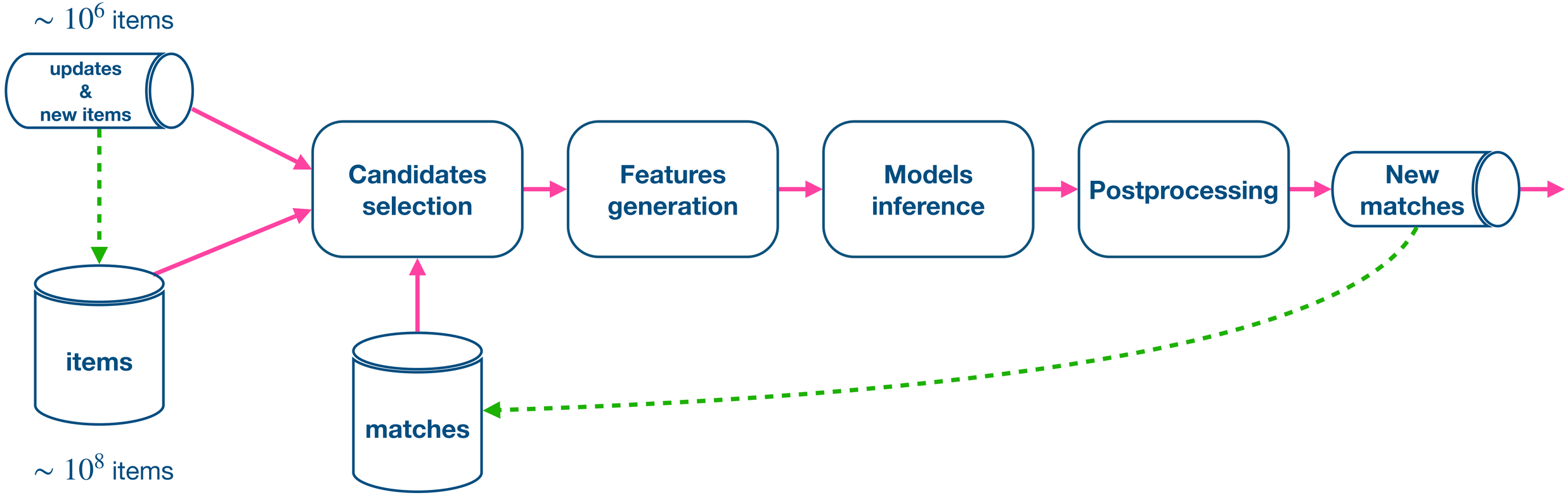
$M^{t+1} = \|m_{i,j}\|^{t+1} -$ matches matrix of shape $L \times L,\ m_{i,j} \in \{0, 1\}$

Result:

$\tilde{M}^{t+1} = \|\tilde{m}_{i,j}\|^{t+1} -$ matches matrix of shape $K^{t+1} \times K^{t+1},\ \ K^{t+1} \ll N + L$

Matching Pipeline

High level design



Problem Solving In Practice

Curse of dimensionality

Full matrix cartesian product:

10^8 items $\Rightarrow$ $10^8 \times 10^8 = 10^{16}$ (ten quadrillion) pairs

New and updated items (up to 10^6 per day):

$10^6 \times 10^8 = 10^{14}$ (one hundred trillion)

100 most similar for each of new and updated items:

$10^6 \times 10^2 = 10^8$

Narrowing the Search Space

Negatives that are easy to recognize

≠

≠

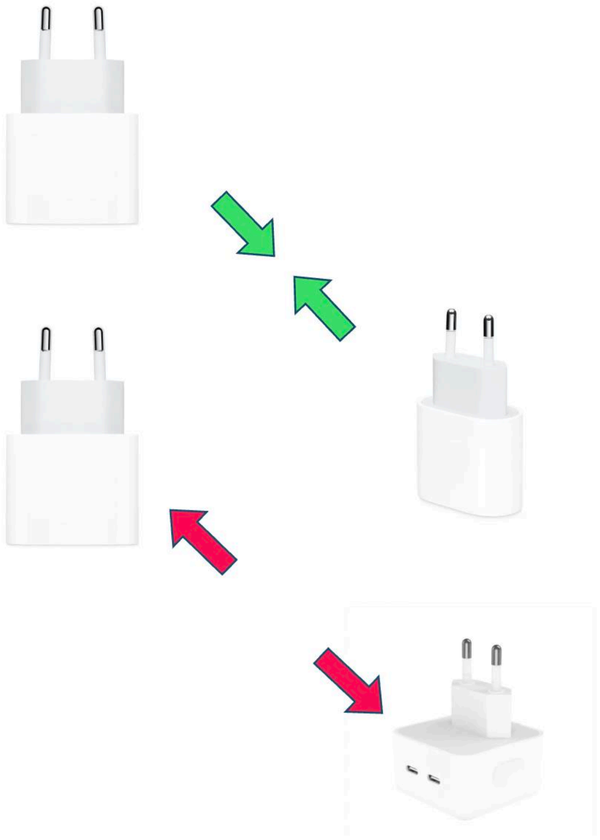
USB Adapter ≠ **Sweatshirt with print**

Banana chips ≠ **Cappuccinatore Kitfort**

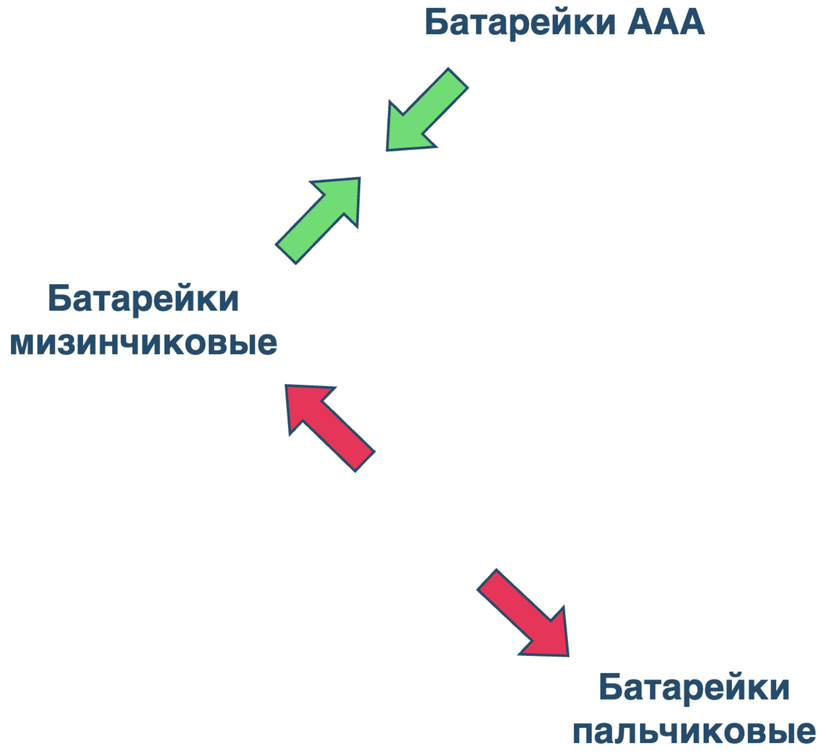
Outdoor thermometer ≠ **Dried mango «King»**

Narrowing the Search Space

Metric spaces



Contrastive Loss

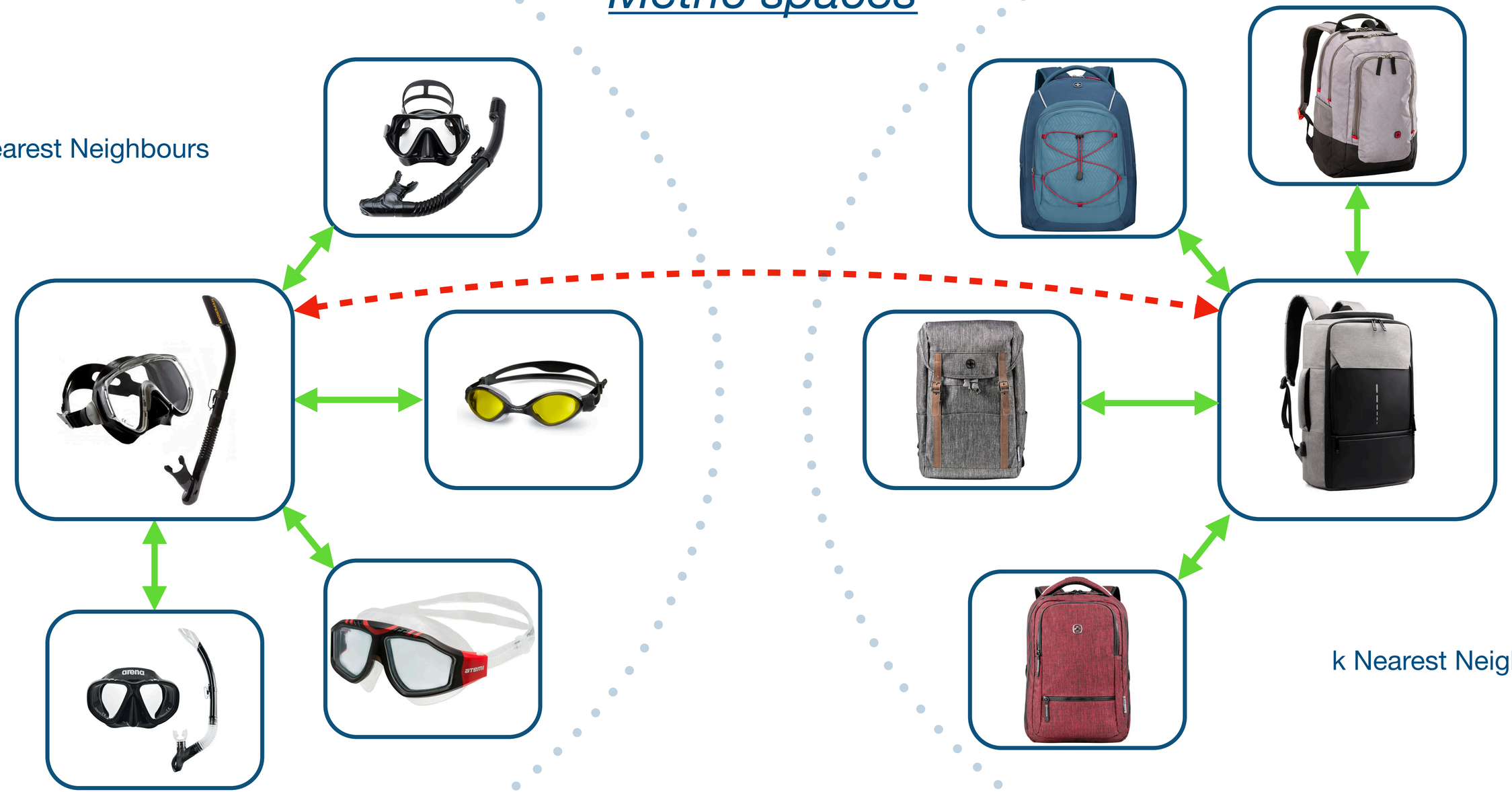


Triplet Loss

Narrowing the Search Space

Metric spaces

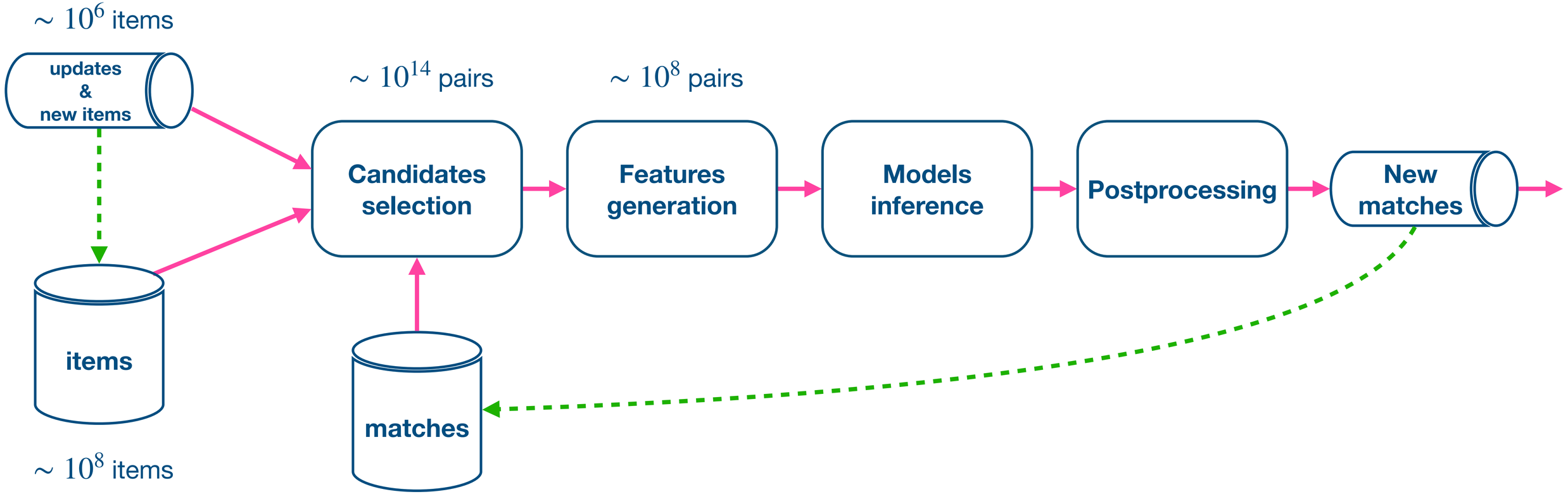
k Nearest Neighbours



k Nearest Neighbours

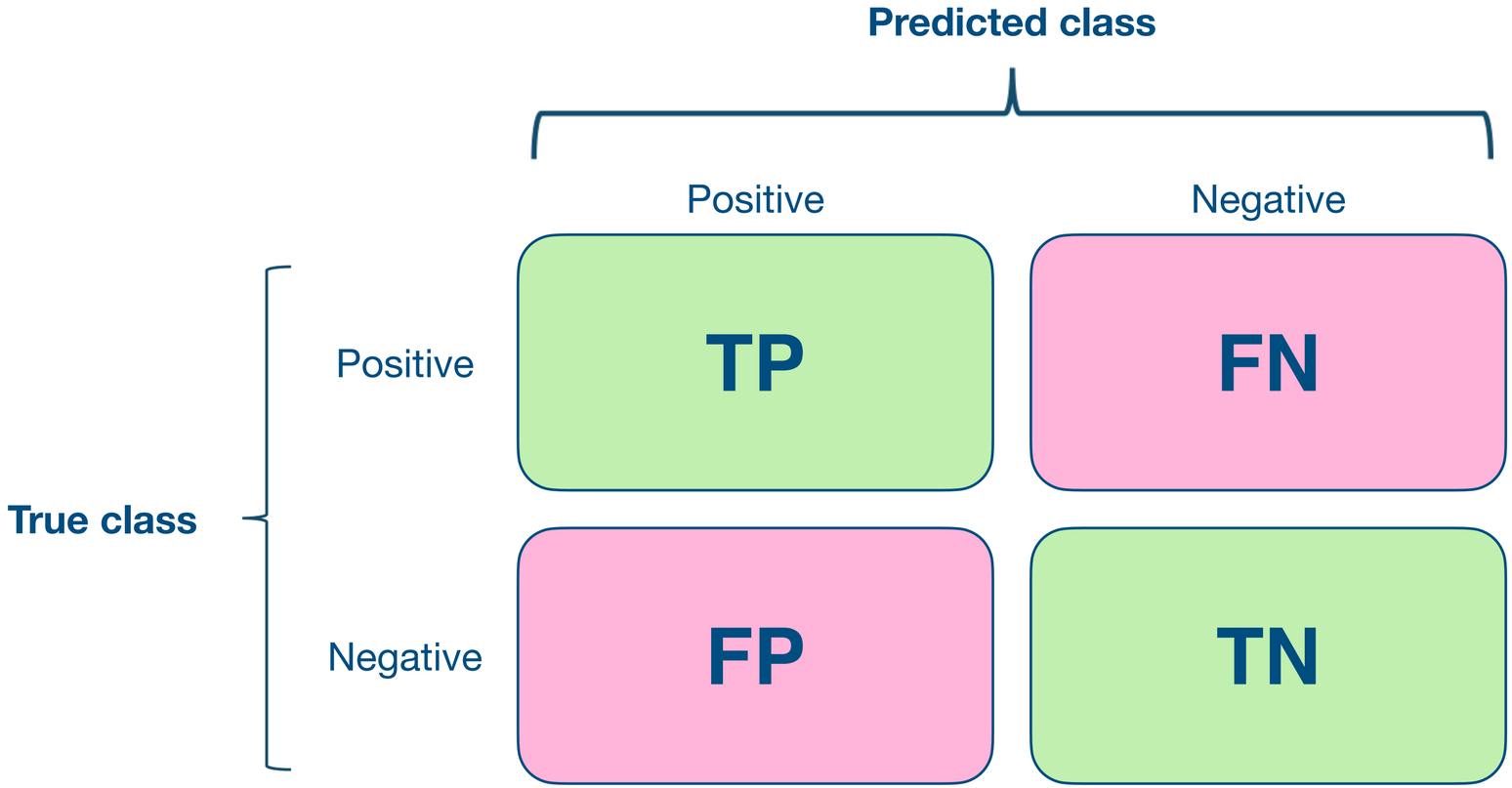
Matching Pipeline

High level design



Solution Quality Evaluation

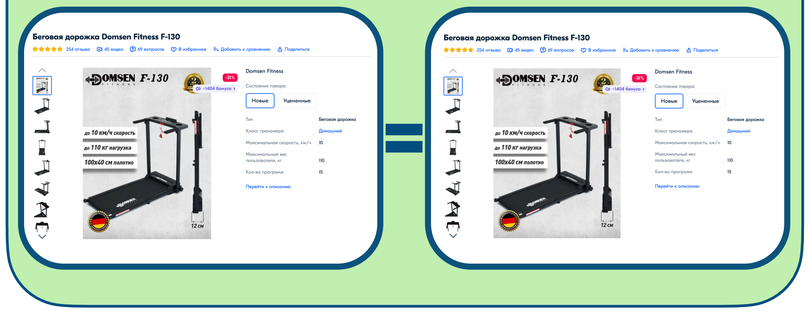
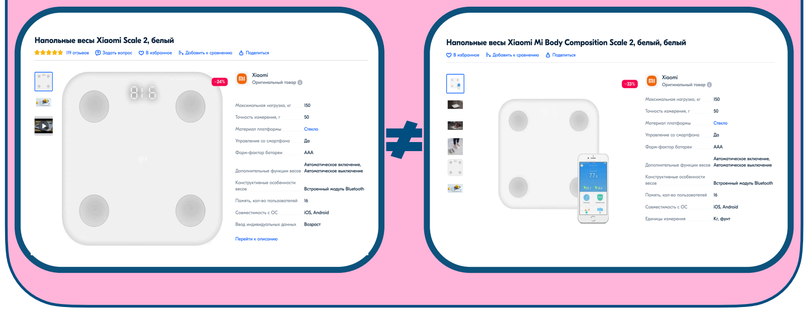
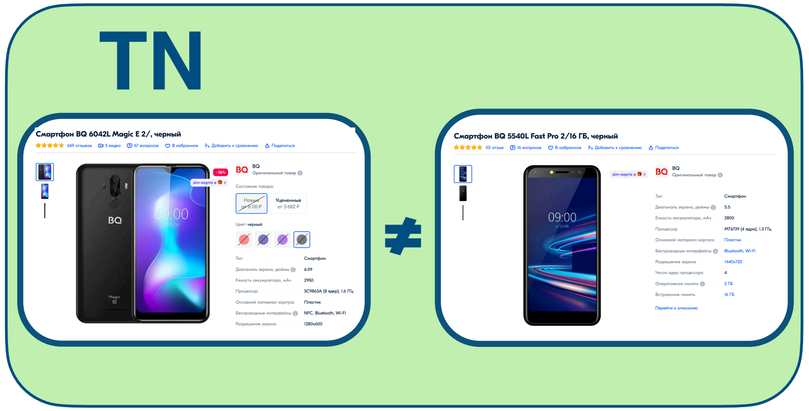
Correct predictions and mistakes



Solution Quality Evaluation

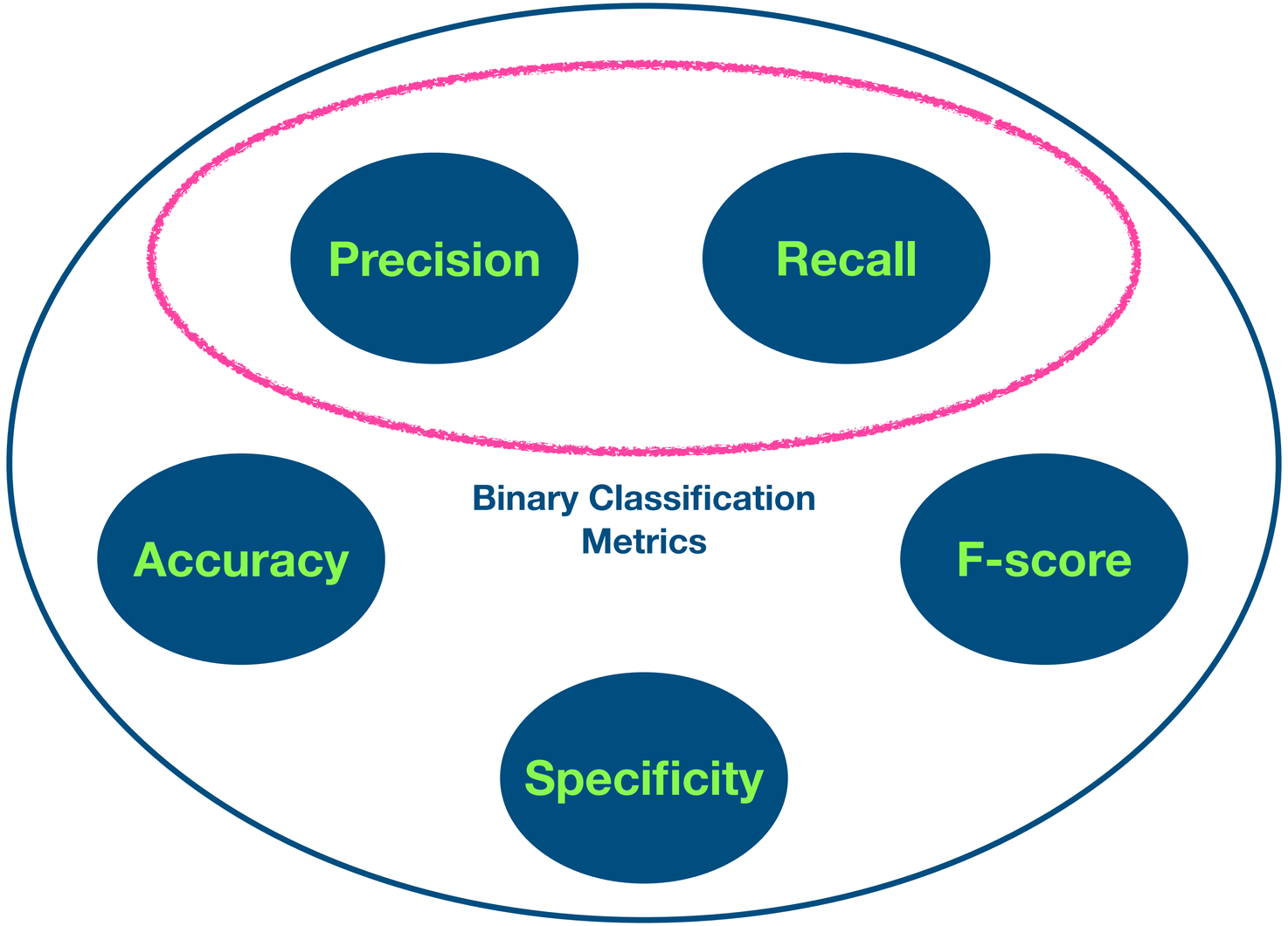
Correct predictions and mistakes

Predicted class

True class	Positive	Negative
Positive	**TP** 	**FN** 
Negative	**FP** 	**TN** 

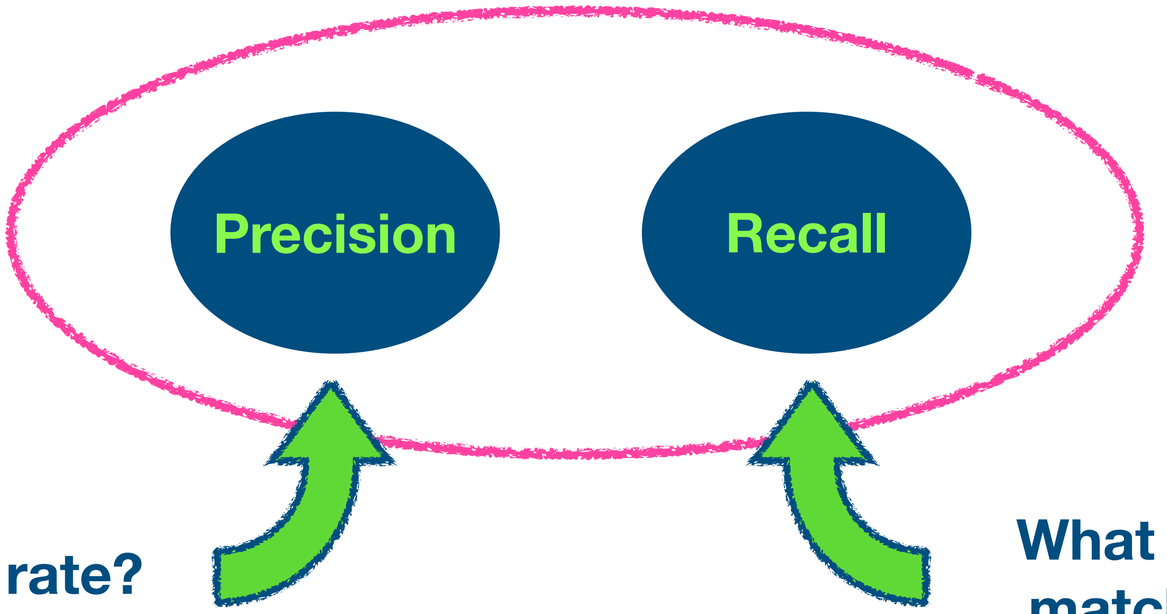
Solution Quality Evaluation

Metrics Zoo



Solution Quality Evaluation

Matching Business Metrics



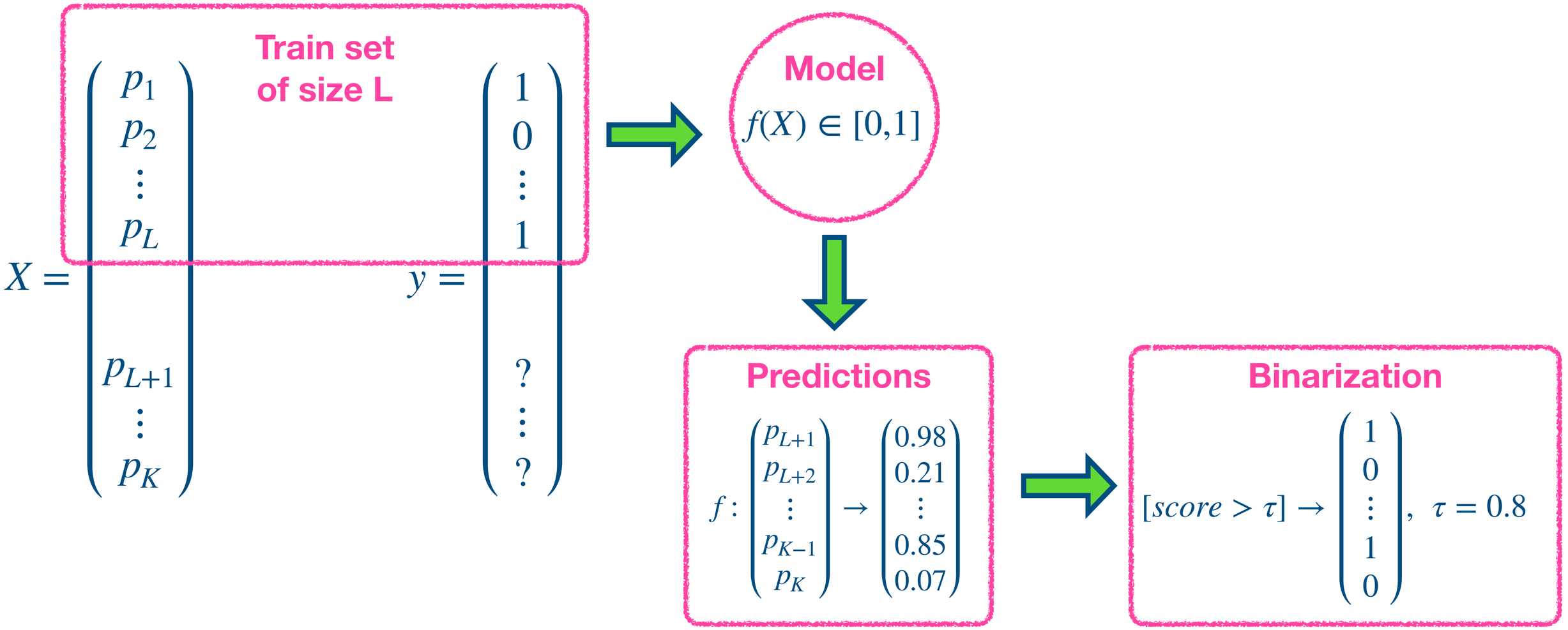
What is the error rate?

What proportion of all matches do we find?

Both works with labels {0, 1}.

How to get a solution?

Train a binary classification model and use it!

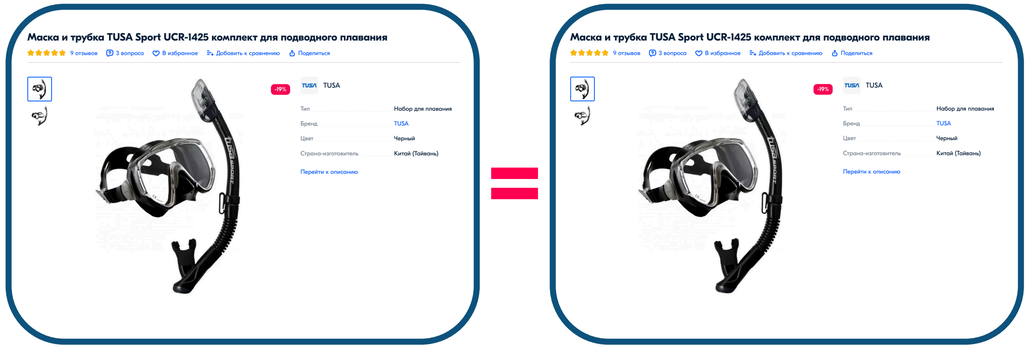


Solution Quality Evaluation

Practical approach

1. Recall @ (Precision = X%)

- Precision is 1 priority metric.
- Precision should be high enough (e.g. 95%).
- Is useful for search for identical products.



2. Precision @ (Recall = Y%)

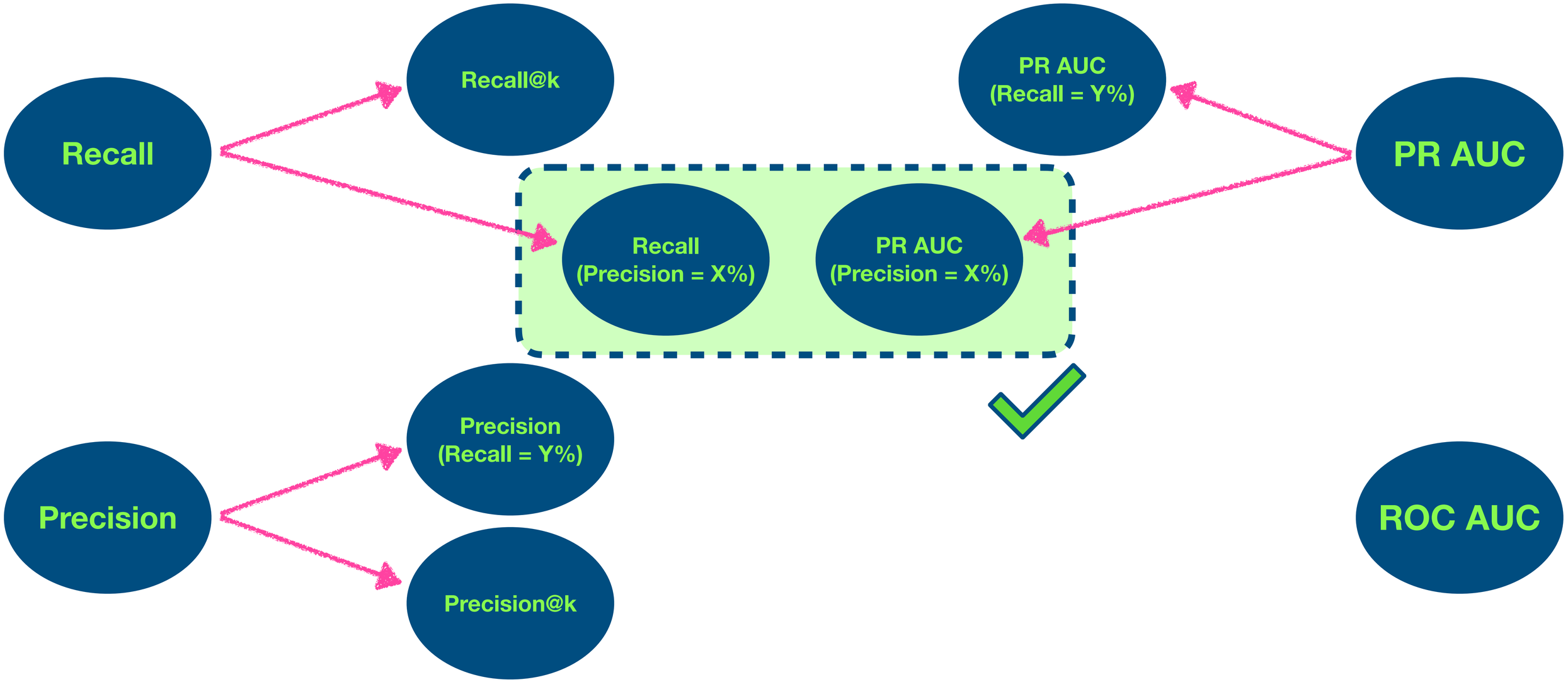
- Recall is 1 priority metric.
- Recall should be high enough (e.g. 75%).
- Is useful for search for substitute products or to select candidates.

Solution Quality Evaluation

Offline metrics

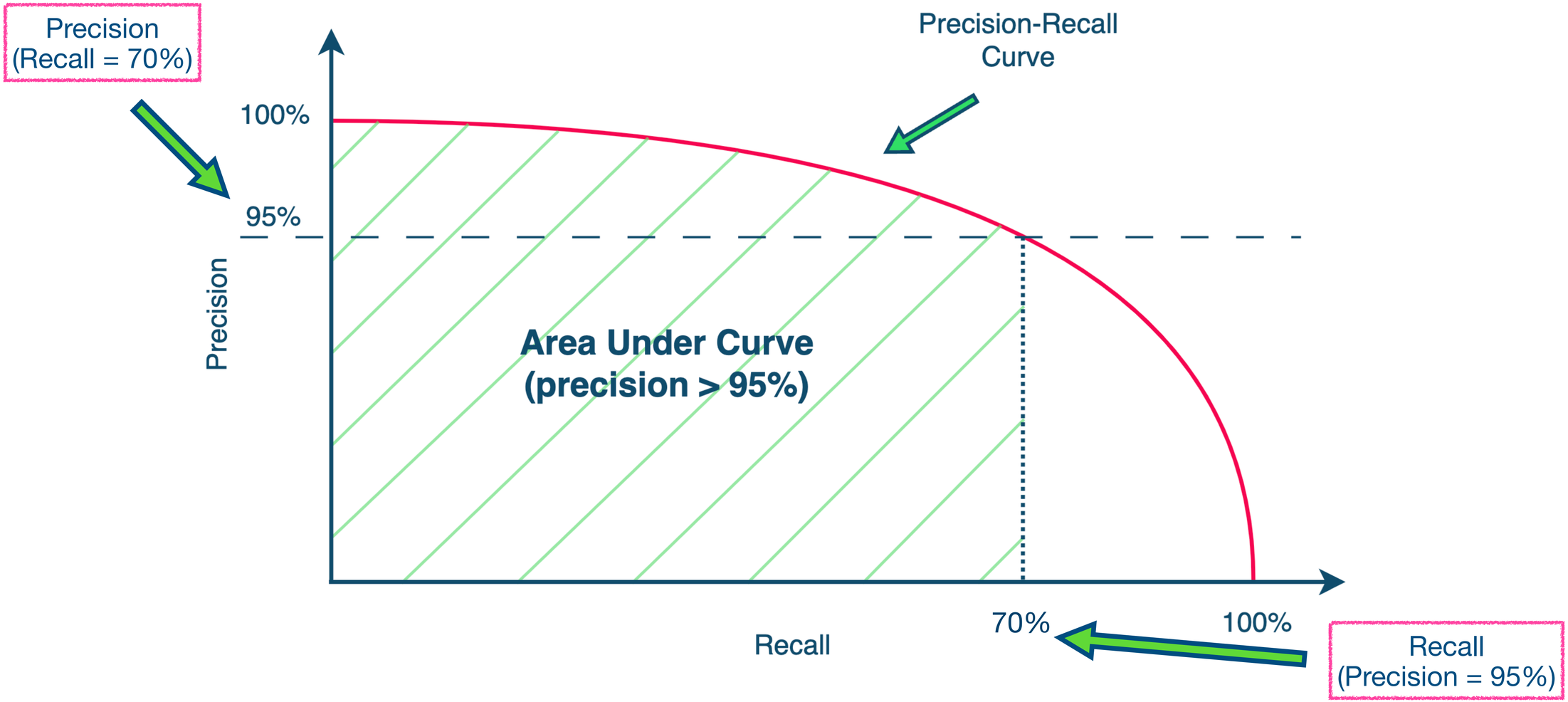
Solution Quality Evaluation

Offline metrics



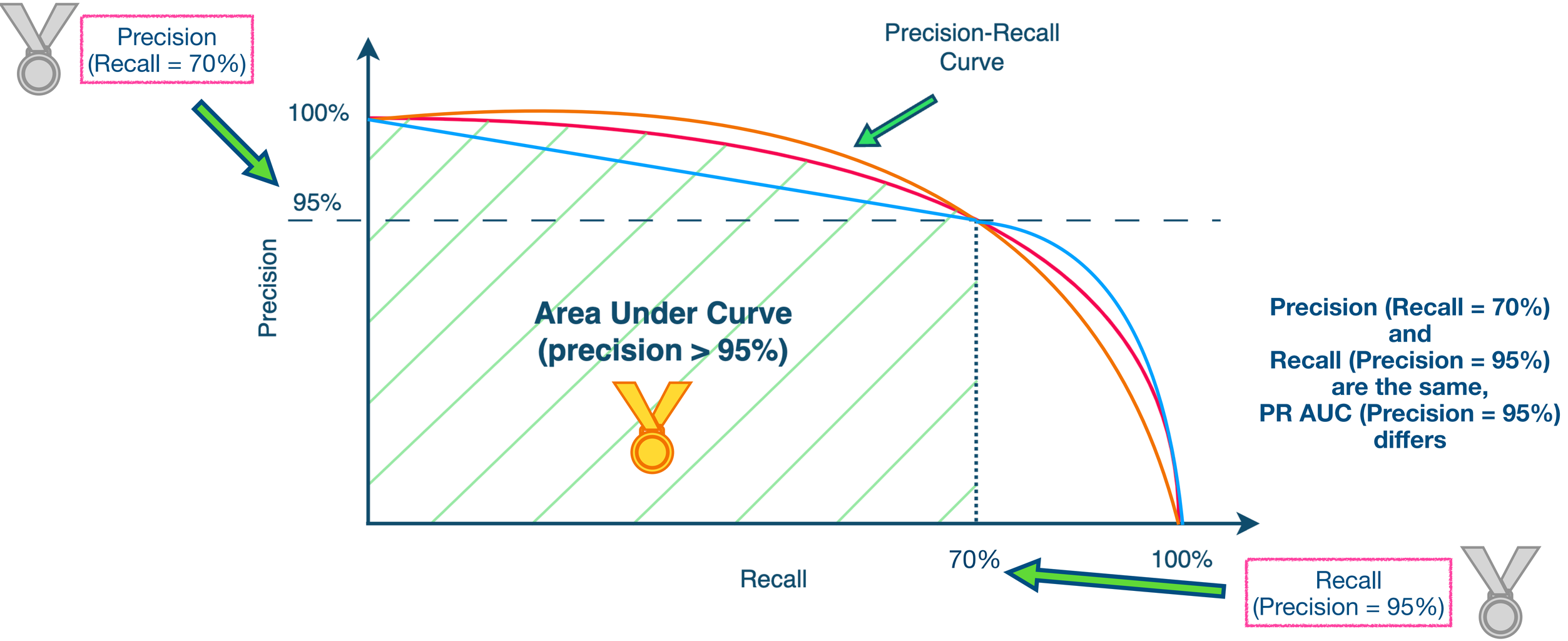
Solution Quality Evaluation

Offline metrics EXPLAINED



Solution Quality Evaluation

Offline metrics EXPLAINED



Important to keep in mind

Secrets, tricks and folk wisdom

- Threshold chosen by train set could lead to unexpected results since data distribution in production almost surely differs from train set.
- Offline metrics could be used to choose one model of many but, again, compare models on data with production-like distribution.
- Use bootstrap to exclude the random factor.
- Make sure you don't have any bugs in the metric code…