

---

# Занятие № 4

Оценка точности  
модели,  
переобучение,  
регуляризация



---

# Содержание

---

- 1 Обучающая и тестовая выборка, кросс-валидация
- 2 Метрики качества: accuracy, precision, recall
- 3 Смещение и разброс (bias-variance tradeoff)
- 4 Признаки переобучения и регуляризация
- 5 Практика.



# Обучающая и тестовая выборка, кросс-валидация

**Обучающая выборка** содержит значения признаков и целевой переменной.

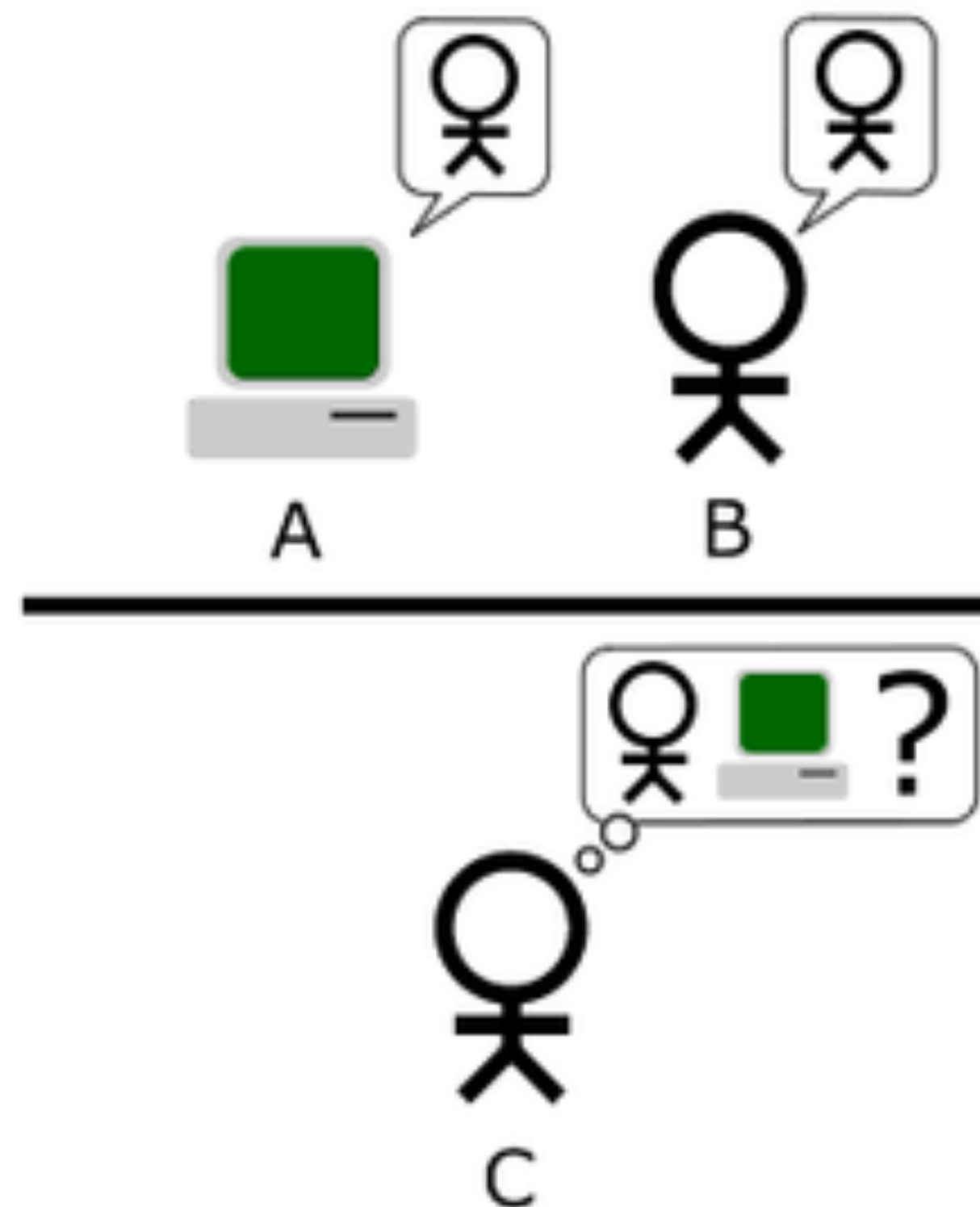
На обучающей выборке строим модель.



# Обучающая и тестовая выборка, кросс-валидация

**Тестовая выборка** содержит значения признаков, по которым необходимо предсказать значение целевой переменной.

Оцениваем качество различных вариантов модели.



# Обучающая и тестовая выборка, кросс-валидация

## **Проблемы:**

Модель может хорошо работать на обучающей выборке, однако сильно терять в качестве на тестовой (один из вариантов - переобучение).

Преобразования данных на обучающей выборке должны быть повторены и иметь смысл для тестовой.



# Обучающая и тестовая выборка, кросс-валидация

Разбиваем обучающую выборку на 2 части.

На одной будем тренировать модель, на другой – проверять (т. е. использовать в качестве тестовой, только с известной целевой переменной)

```
from sklearn.model_selection import train_test_split  
X_train, X_test, y_train, y_test = train_test_split( X, y, test_size = 0.3, random_state = 0 )
```

ОБУЧАЮЩАЯ ВЫБОРКА

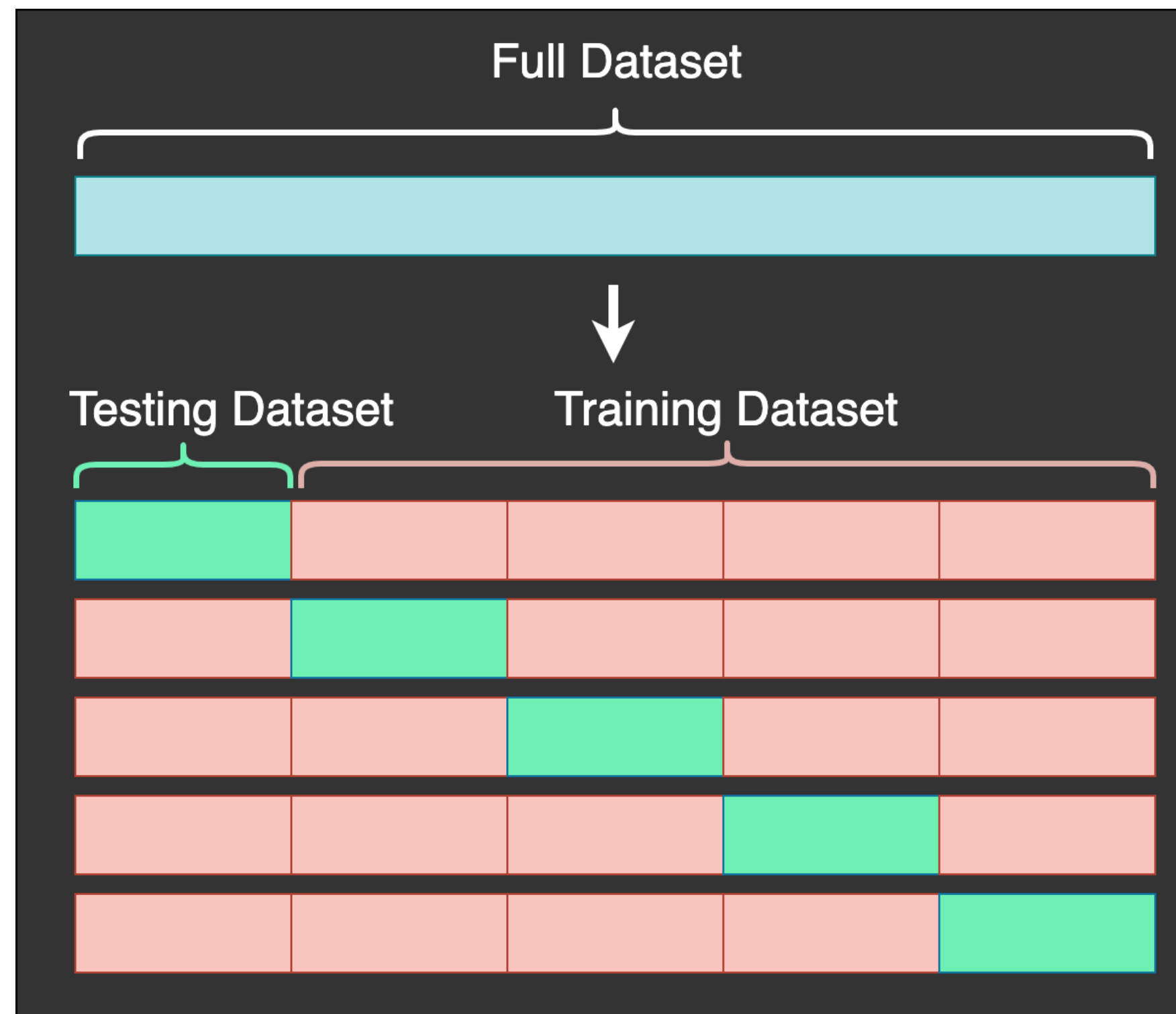


The diagram illustrates the process of splitting a dataset. At the top, a solid blue horizontal bar is labeled 'ОБУЧАЮЩАЯ ВЫБОРКА' (Training Sample). A large blue arrow points downwards from this bar to a second horizontal bar below it. This second bar is divided into two segments: a larger blue segment on the left labeled 'TRAINING' and a smaller yellow segment on the right labeled 'TEST'.

TRAINING

TEST

# Обучающая и тестовая выборка, кросс-валидация





# Метрики качества: accuracy, precision, recall

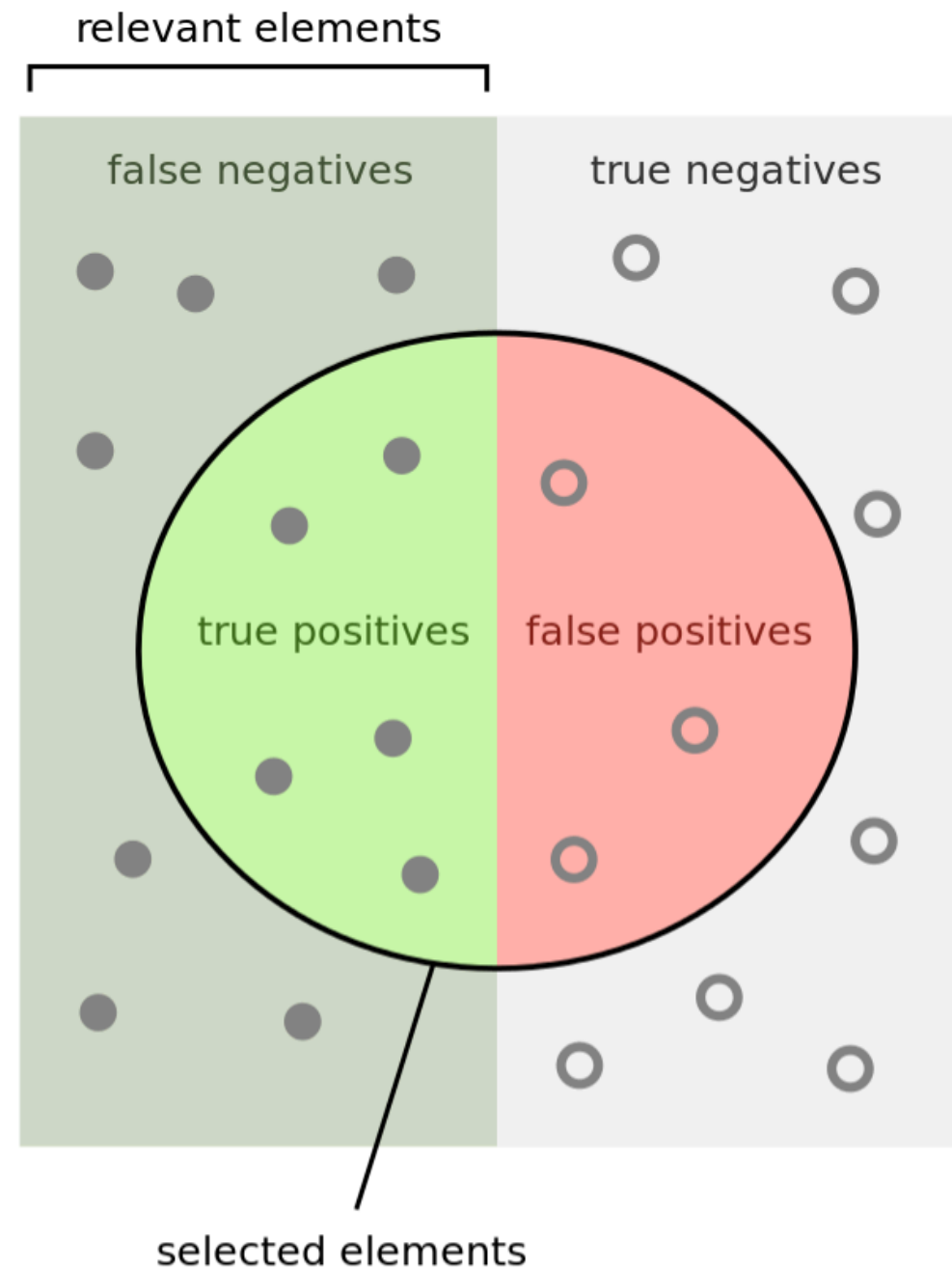
n=165		Predicted: NO	Predicted: YES	
Actual: NO		TN = 50	FP = 10	60
Actual: YES		FN = 5	TP = 100	105
		55	110	



Метрики качества: accuracy,  
precision, recall

$$accuracy = \frac{correct}{correct + incorrect}$$

# Метрики качества: accuracy, precision, recall



# Метрики качества: accuracy, precision, recall

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

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

$$F1 = 2 \cdot \frac{\textit{precision} \cdot \textit{recall}}{\textit{precision} + \textit{recall}}$$

$TP$  = True positive

$TN$  = True negative

$FP$  = False positive

$FN$  = False negative

# Метрики качества: accuracy, precision, recall

		PREDICTIVE VALUES		
		POSITIVE (1)	NEGATIVE (0)	
ACTUAL VALUES	POSITIVE (1)	TP = 3	FN = 1	4
	NEGATIVE (0)	FP = 2	TN = 4	6
		5	5	

Diagram illustrating a Confusion Matrix for classification metrics: accuracy, precision, and recall.

The matrix is structured as follows:

- ACTUAL VALUES (Rows):**
  - POSITIVE (1)
  - NEGATIVE (0)
- PREDICTIVE VALUES (Columns):**
  - POSITIVE (1)
  - NEGATIVE (0)

Key components and metrics highlighted:

- TP (True Positive) = 3:** Correctly classified positive instances.
- FN (False Negative) = 1:** Positive instances incorrectly classified as negative.
- FP (False Positive) = 2:** Negative instances incorrectly classified as positive.
- TN (True Negative) = 4:** Correctly classified negative instances.

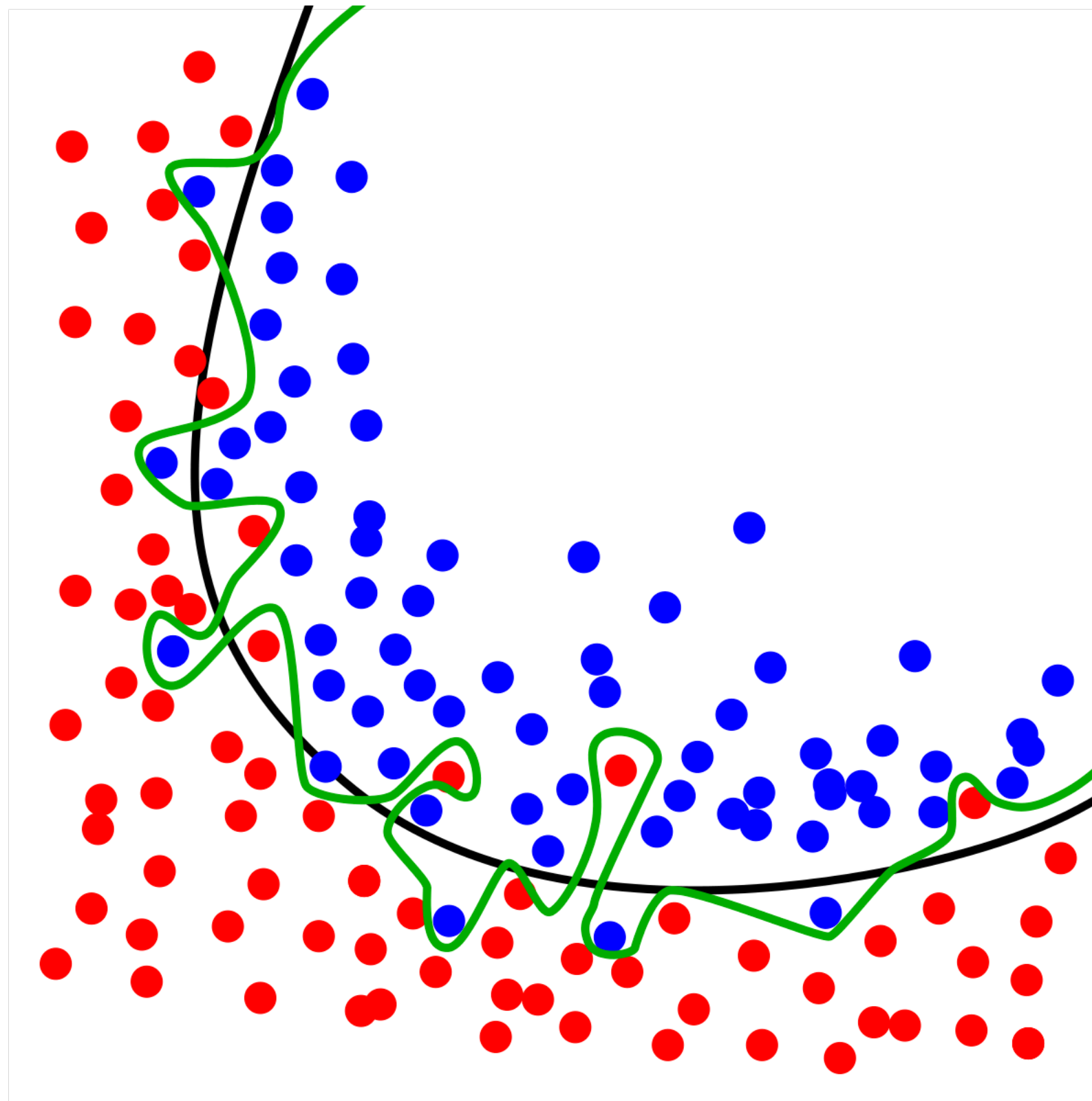
Row and Column Totals:

- Row 1 (Actual Positive): 4
- Row 2 (Actual Negative): 6
- Column 1 (Predicted Positive): 5
- Column 2 (Predicted Negative): 5

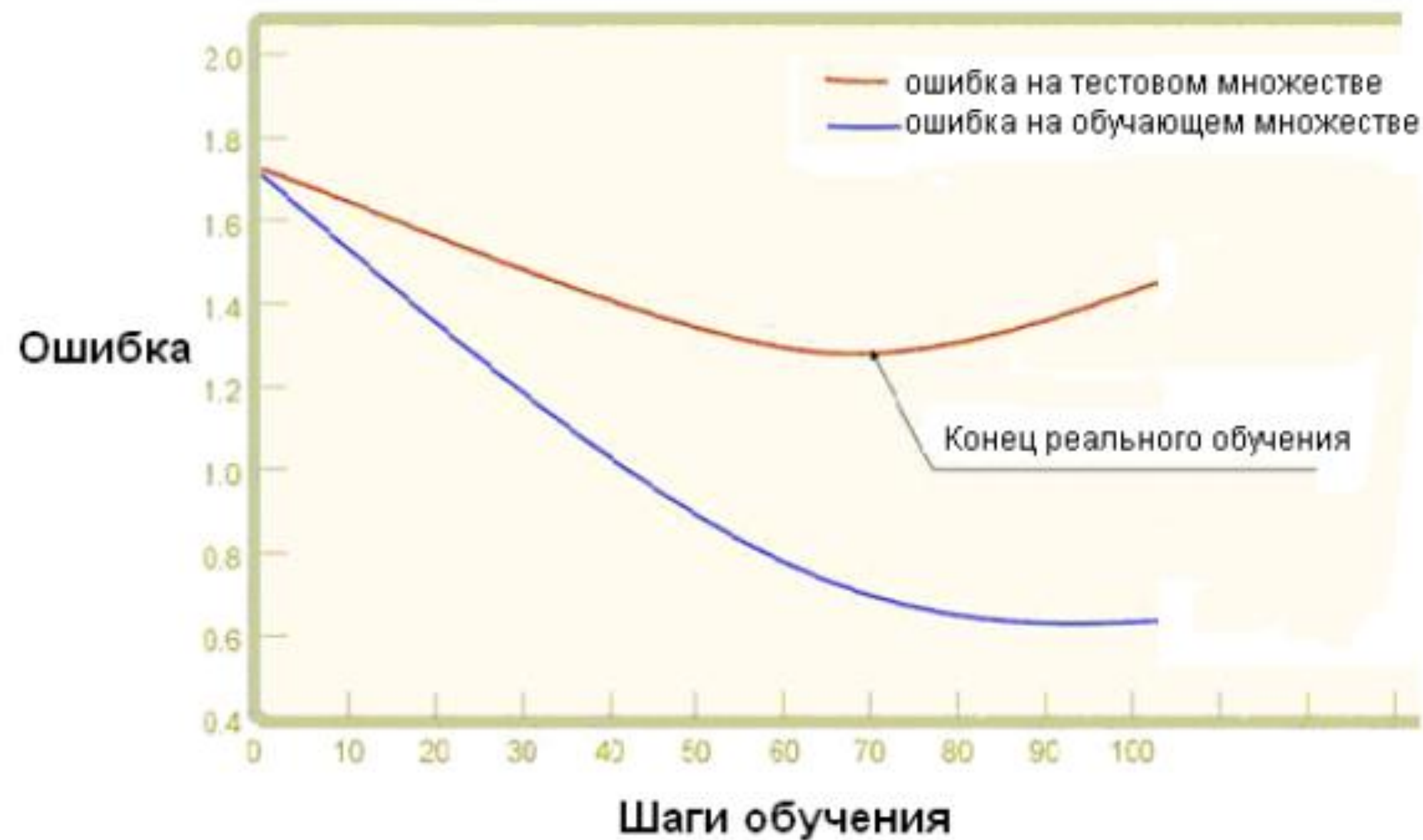
Metrics highlighted with callouts:

- PRECISION:** Indicated by a green box around the TP and FP cells.
- RECALL:** Indicated by a red box around the TP and FN cells.

# Признаки переобучения и регуляризация



# Признаки переобучения и регуляризация



# Признаки переобучения и регуляризация

L1 Regularization

$$\text{Cost} = \sum_{i=0}^N (y_i - \sum_{j=0}^M x_{ij} W_j)^2 + \lambda \sum_{j=0}^M |W_j|$$

L2 Regularization

$$\text{Cost} = \sum_{i=0}^N (y_i - \sum_{j=0}^M x_{ij} W_j)^2 + \lambda \sum_{j=0}^M W_j^2$$

Loss function

Regularization  
Term



**ПРАКТИКА**



---

# Спасибо за внимание!

---

**Сапрыкин Артур**  
Data Scientist



[fb.com/asaprykin92](https://fb.com/asaprykin92)



[asaprykin92@gmail.com](mailto:asaprykin92@gmail.com)

