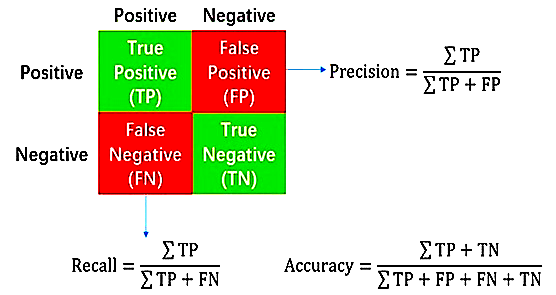
# Метрики оценки качества трекинга

Метрики для оценки качества работы нейросетевой классифицирующей модели сравнивают результат классификации с заранее известной информацией, такой как истинные метки классов (лейблы).

По существу – это общий вопрос о критериях проверки правильности статистических гипотез, основной и альтернативной. Ошибки классификации бывают двух видов: FP и FN. В статистике FP называют ошибкой I-го рода, а FN — ошибкой II-го рода (см. рисунок).



Где:

* 𝑇𝑃 (*True Positive*) - количество истинно положительных предсказаний;
* 𝑇𝑁 (*True Negative*) - количество истинно отрицательных предсказаний;
* 𝐹𝑃 (*False Positive*) - количество ложно положительных предсказаний;
* 𝐹𝑁 (*False Negative*) - количество ложно отрицательных предсказаний.

В задачах трекинга применяются следующие метрики:

***Accuracy (общая эффективность)*. *Recall (***полнота)**, *Precision*** (точность)**, *F1-score***

Первая наивная оценка эффективности трекинга (1) по процентному отношению числа найденных треков к общему числу треков-кандидатов (accuracy)– бесполезна и даже опасна, т.к. наша выборка очень сильно несбалансирована и среди треков-кандидатов может оказаться слишком много ложных, образованных из шумовых отсчетов и кусков разных треков, так что среди распознанных, кроме тех, что в самом деле являются реальными треками оказываются также и те, что ошибочно распознаны как треки, но они – ложные (ghosts- призраки).

Поэтому физики используют метрики recall (полнота) и precision (точность), а также их гармоническое среднее ***F1,***  которые не зависят, в отличие от accuracy, от соотношения классов и потому применимы в условиях несбалансированных выборок,***.*** Вместо precision чаще используется purity=1- precision (чистота), которая показывает процент ложных треков – ghosts, наличие которых в результирующих данных может серьезно исказить физические результаты эксперимента.

Принято результат трекинга считать приемлемым, если recall превышает 95%, а purity не выше 5-7%, т.е. F1~0,95