



Современные подходы для разметки данных в области информационной безопасности

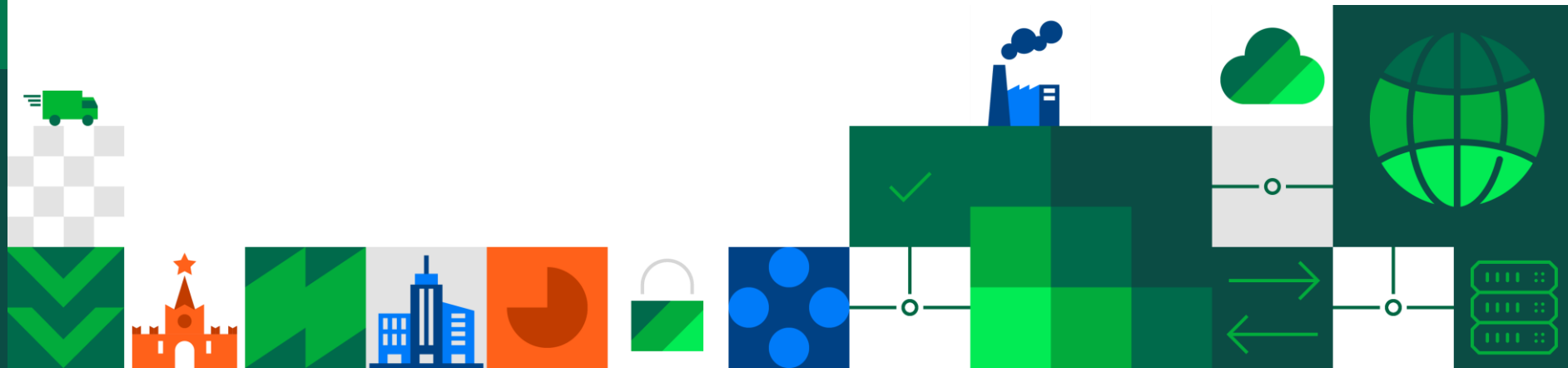
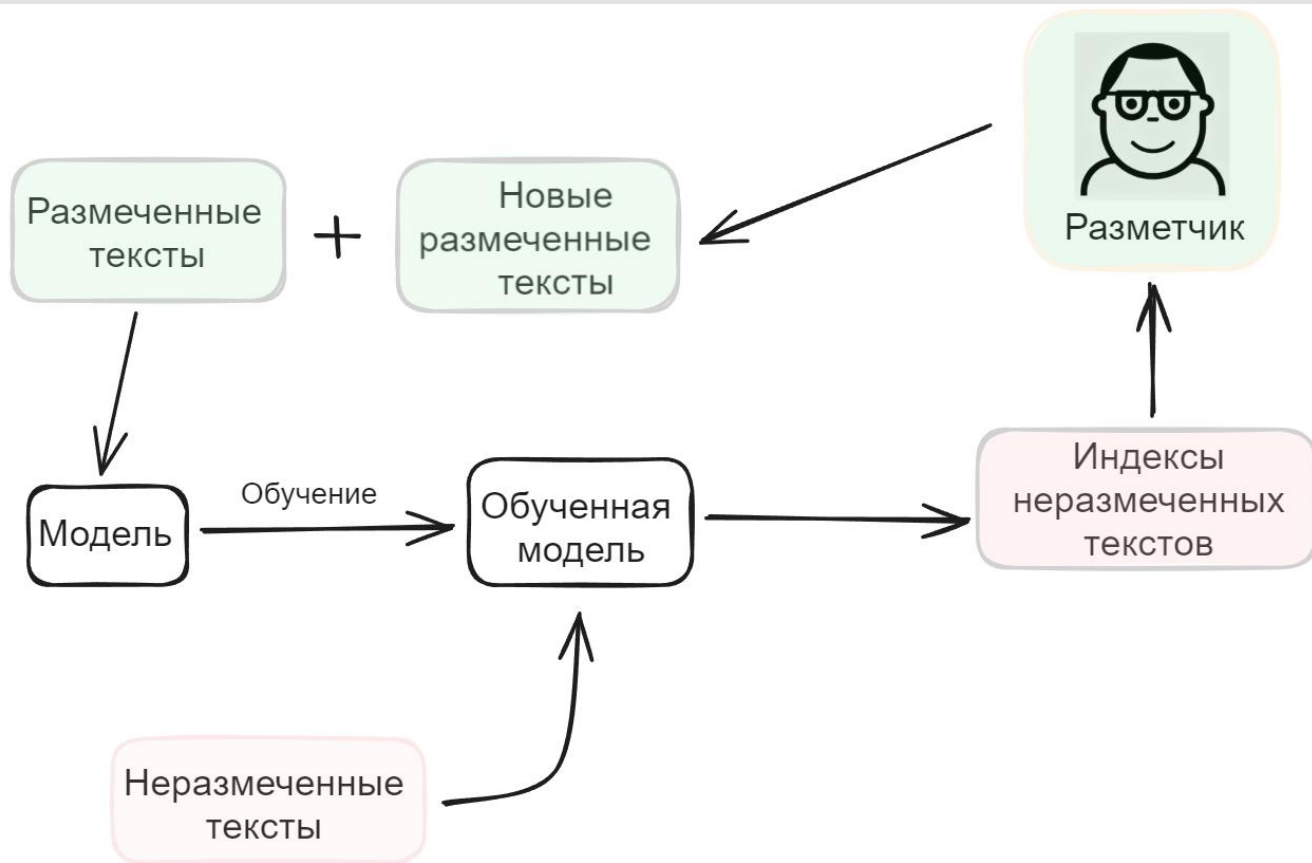
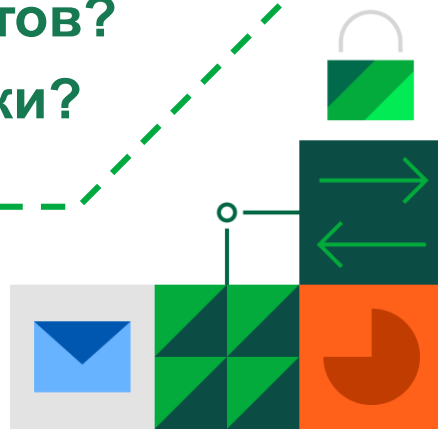


Схема Active Learning

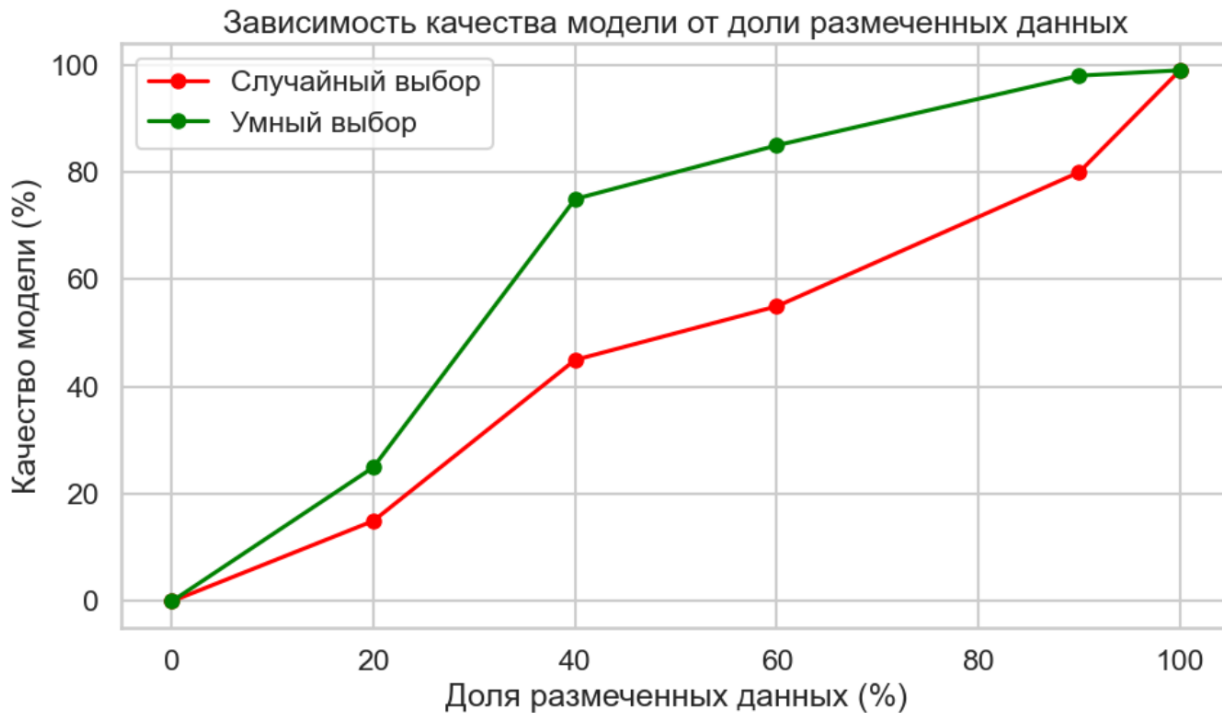


Вопросы по реализации AL

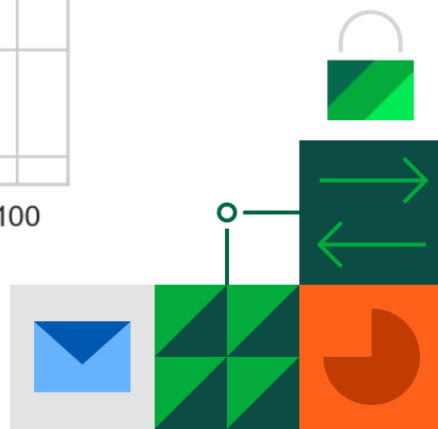
1. Как выбрать **тексты** для разметки?
2. Какие тексты разметить на **старте**?
3. Когда **останавливать** AL?
4. Как оценить **стоимость** разметки текстов?
5. Как **ускорить** цикл обучения и разметки?



Как выбрать тексты для разметки?

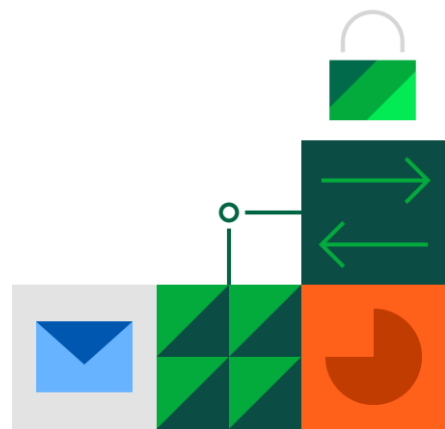


**Важно руководствоваться тем, что модель знает
и что она не знает!**



Квадрат знаний модели

Такие данные нет смысла подавать,
модель не узнает ничего нового



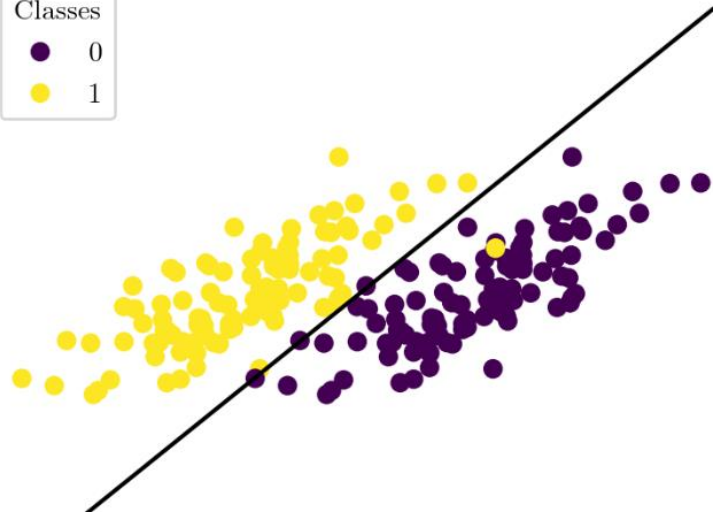
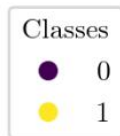
Uncertainty Sampling

Известное незнание обнаруживается в ходе анализа предсказаний модели.

Смысл Uncertainty Sampling – важно обучить модель на тех объектах, в которых она **не уверена**.

Алгоритм:

1. Считаем уверенность для всех неразмеченных объектов
2. Оставляем n самых неуверенных на разметку
3. Обучаем модель заново с этими объектами

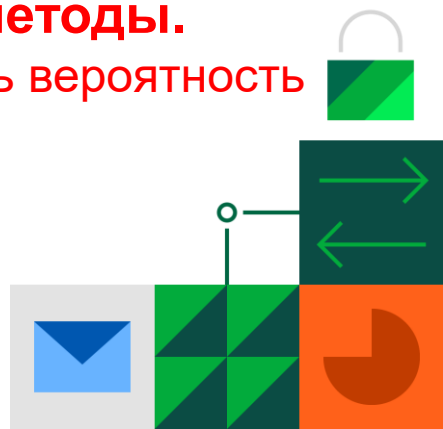


Примеры алгоритмов Uncertainty Sampling

Фокусируемся на объектах, в которых модель сомневается. Для оценки уверенности модели применяют:

1. Least Confident
2. Margin of Confidence
3. Ratio of Confidence
4. Label Entropy
5. Query by committee

Вероятностные методы.
Могут переоценивать вероятность

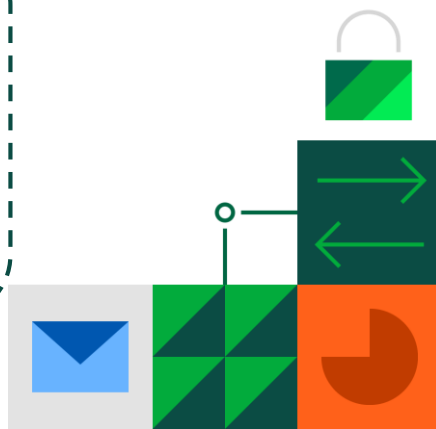


Diversity Sampling

Подход ориентирован на те знания, про которые модель **вообще ничего не знает**. Это выявляется по распределению данных.

В реальных задачах ИБ данные часто **смещены**:

1. Отрицательных классов больше, чем положительных
2. Разные источники атак
3. Разный стиль написания кода/атак
4. Разные языки

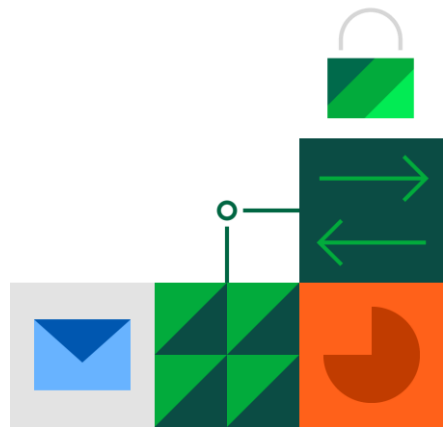


Diversity Sampling

Мы хотим, чтобы модель могла работать со всеми данными. То есть нам хотелось бы находить такие примеры, которые **отличаются** от остального корпуса!

То есть, методы Diversity Sampling **увеличивают разнообразие**:

1. Model-based Sampling
2. Cluster-based Sampling
3. Representative Sampling



Комбинирование методов сэмплирования

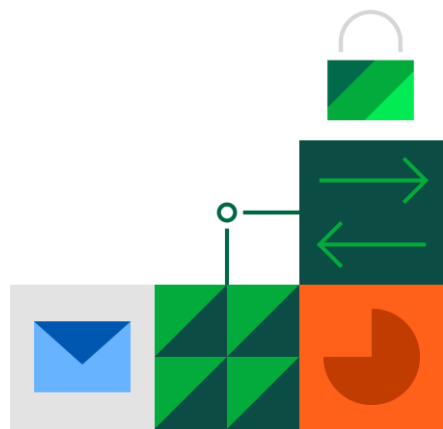
Uncertainty Sampling

- Похожесть данных
- Слабая обобщаемость

Diversity Sampling

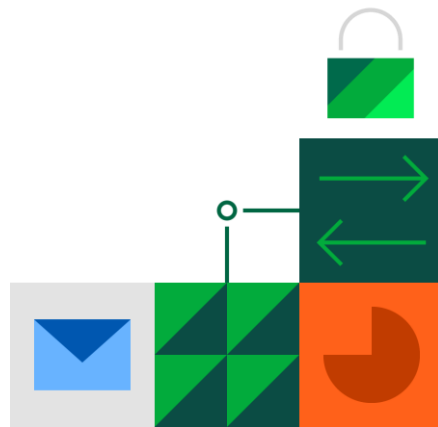
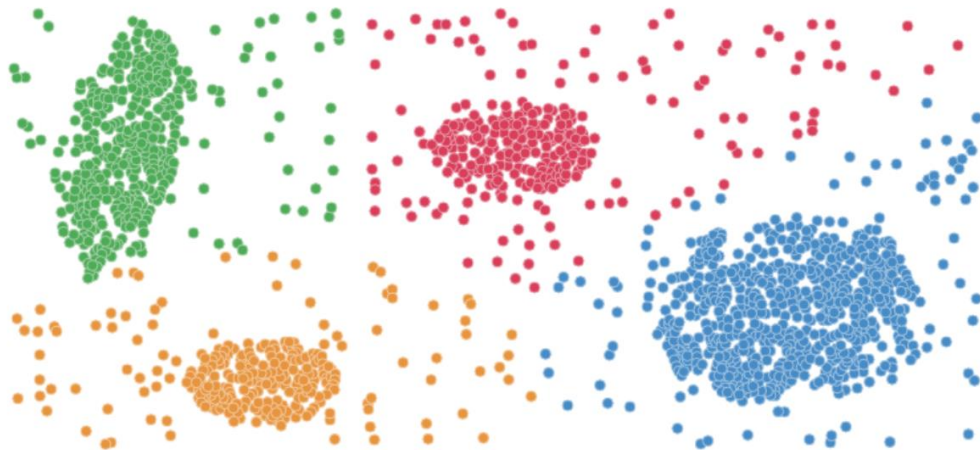
- Простые примеры
- Фокусируется на выбросах

Комбинируем их!



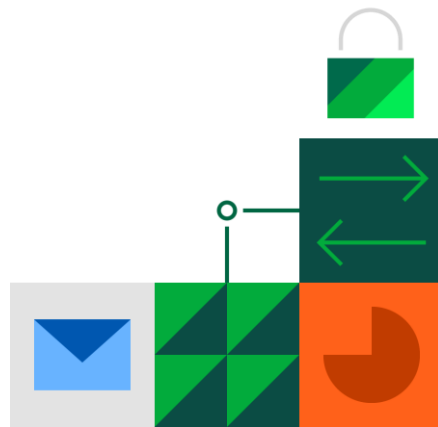
Какие данные размечать на старте?

Не стоит выбирать данные случайно. Лучше **разбить всю выборку на кластеры** и сэмплировать центры кластеров.



Когда останавливать AL?

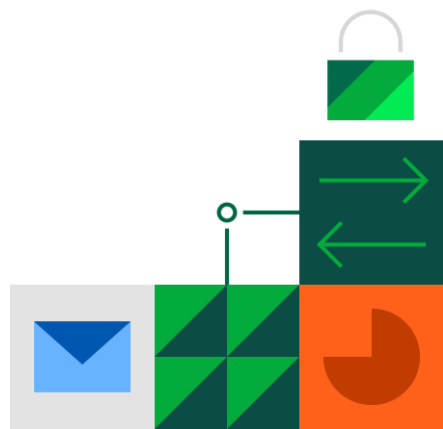
- 
- При израсходовании бюджета
 - При достижении
определенного качества
 - Модель становится уверенной в
предсказаниях
 - Незначительная разница между
итерациями AL



Оценка стоимости разметки данных

Нельзя считать стоимость разметки всех атак одинаковой, так как они имеют разную длину и сложность!

Верный подход: обучить простую модель для предсказания сложности на основе разных текстовых статистик: длина, количество уникальных слов, язык + учесть компетенции разметчиков!



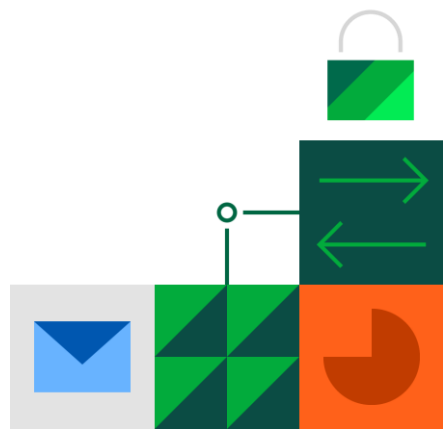
Как использовать стоимость разметки?

1. Не размечать тексты дороже определенного порога. Лучше несколько, но дешевле.
2. Использовать ROI (Return of investment).

ROI: предположим, что значение ROI **растет** при **убывании уверенности** модели. А затем поделим вклад от разметки на стоимость разметки атаки:

$$\text{ROI}(x) = \frac{\text{uncertainty}(x)}{\text{cost}(x)}$$

Берем для разметки атаки с наибольшим ROI!



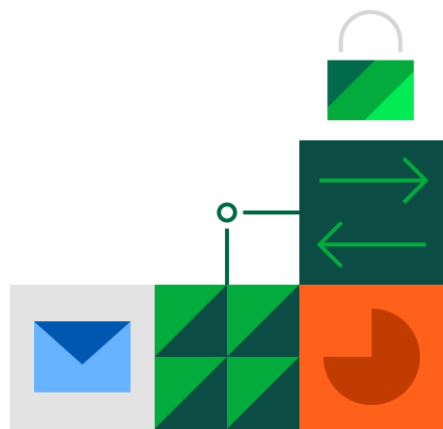
Ускорение цикла обучения и разметки



Дообучать модель на каждой итерации, а не с нуля

Обучать маленькую модель

Обучать модель, пока разметчики заняты разметкой



Что делать, если модель не улучшается?

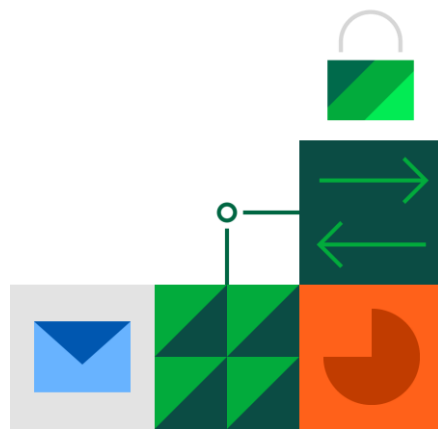


Размечаете слишком мало новых данных

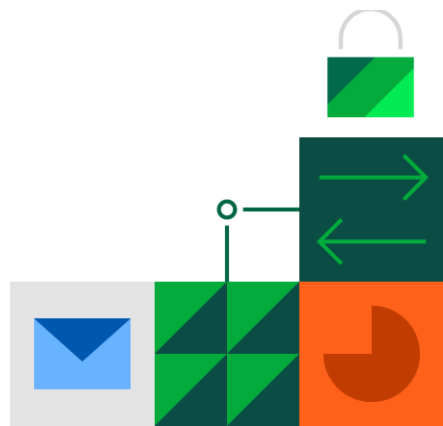
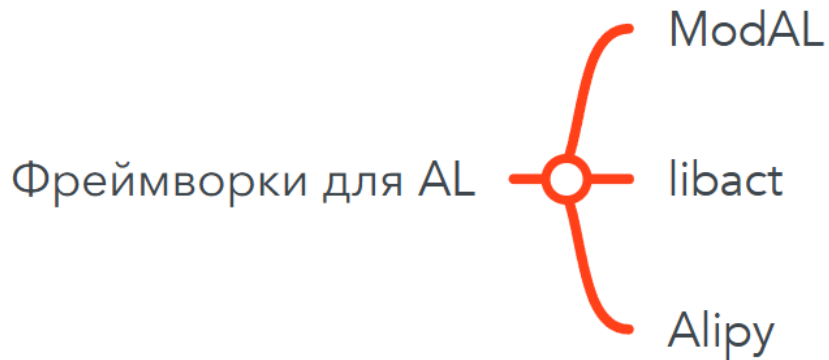
Стратегия отбора уже не оптимальна и ее надо менять

Баг в коде

Датасет уже слишком большой



- Старайтесь автоматизировать все, что можно!
- Убедитесь, что ваш сервис быстро догружает данные для разметки
- Маленький размер батча лишь потратит время на циклы загрузки и выгрузки данных
- Не стоит брать сразу огромный батч данных





Спасибо за внимание!

info@securitycode.ru
www.securitycode.ru

