

Как оценить качество данных

ГЕОРГИЙ КАСПАРЬЯНЦ



Образование

Механико-математический факультет МГУ



Опыт



Teleport app – Middle data scientist



Picsart – Senior data scientist



Labelme – Founder & CEO



Gradient – Head of AI

LabelMe



СЫРОСТЬ ДАННЫХ

Неоднородность, ошибки на уровне чтения этих данных, ошибочные пропуски и другие технические проблемы



Битые картинки, аудио, видео



Json-файлы с ошибками



Неудобная структура данных



Отсутствие ключей в словарях

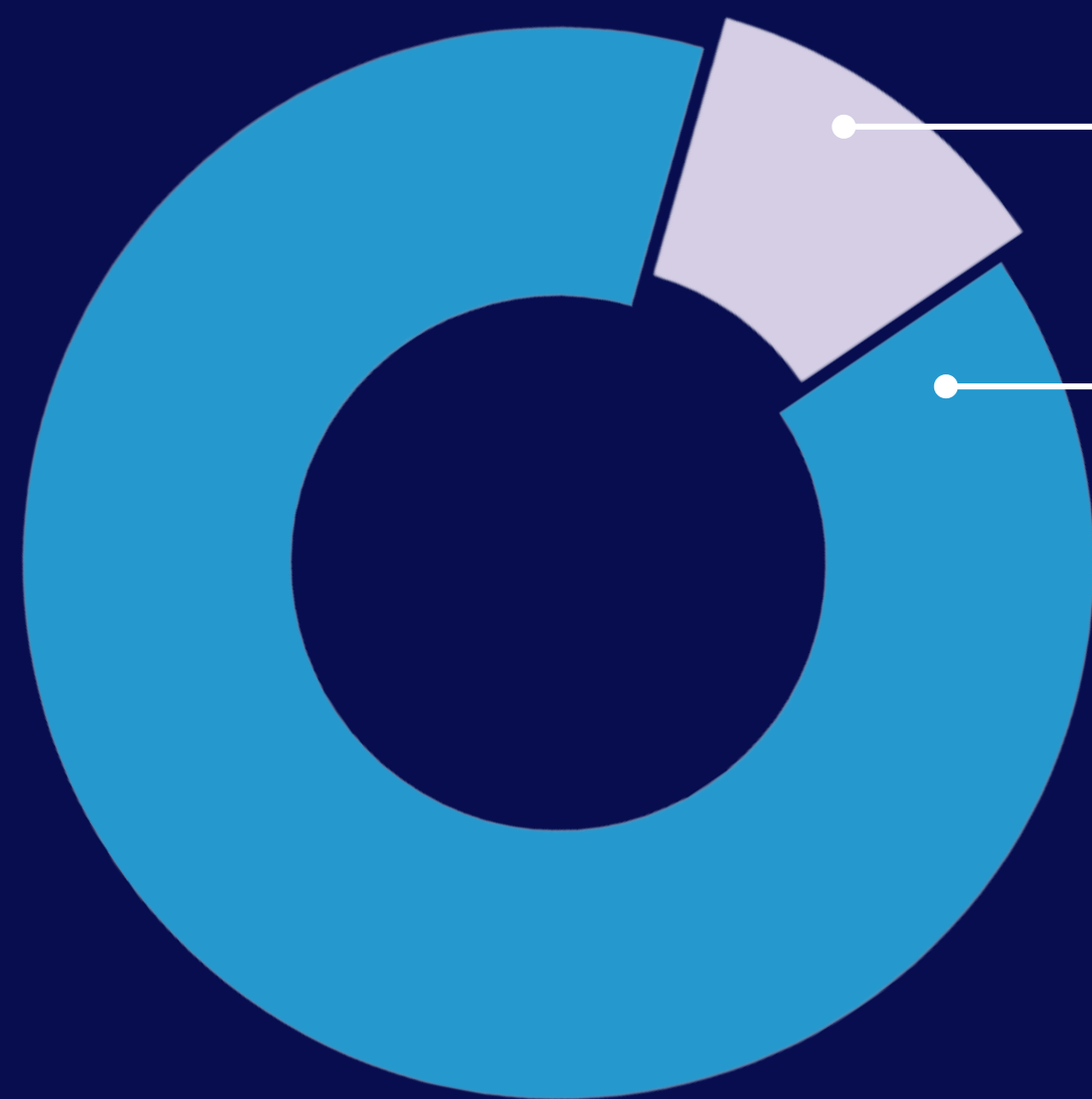


Неудобное место хранения данных

Откуда берется «сырость»?

- ➔ Когда разметчики используют разный софт
- ➔ Когда подготовкой датасета занимаются люди без опыта и навыков
- ➔ Когда данные не прошли финальную проверку

Время на исправление некачественной разметки



11% – исправление данных

+

89% – вся разработка

=

Издержки на зарплату сотрудников

Отклонения от бизнес-плана

Дополнительная нагрузка на штат

КАЧЕСТВО ДАННЫХ

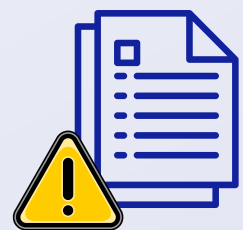
Качество данных — это целевая метрика от того, что разметили, и то, что требовалось на самом деле



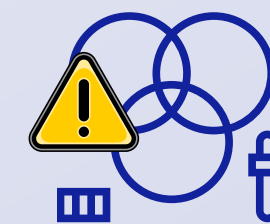
Ошибочные классы в классификации



Неточные границы в сегментации



Извлечены не все классы сущностей



Цвета масок у разных разметчиков отличаются



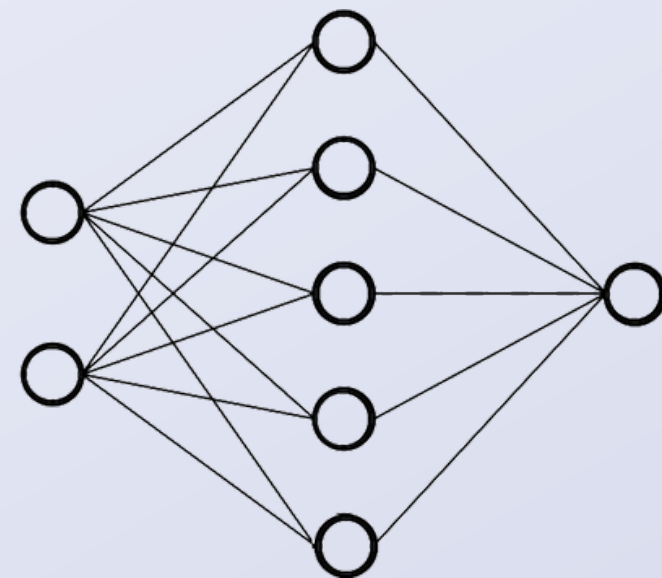
Неверные слова при транскрибации

Почему возникают проблемы с качеством?

- Данные не проходят проверку на этапе разметки
- Разметкой занимаются необученные люди
- Отсутствие подробного ТЗ
- Когда нет перекрестной проверки (сам разметчик проверяет свою работу)

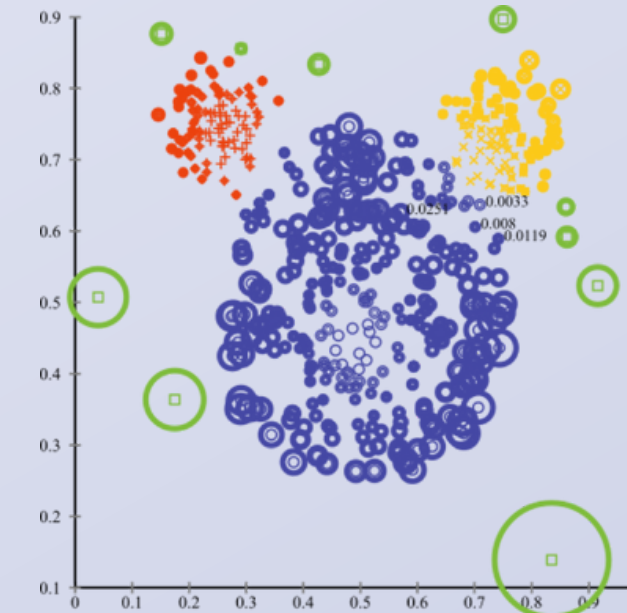
КАК ИСПРАВЛЯТЬ НЕКАЧЕСТВЕННУЮ РАЗМЕТКУ?

Решить проблему можно только тяжелым математическим аппаратом или полуручными методами разметки



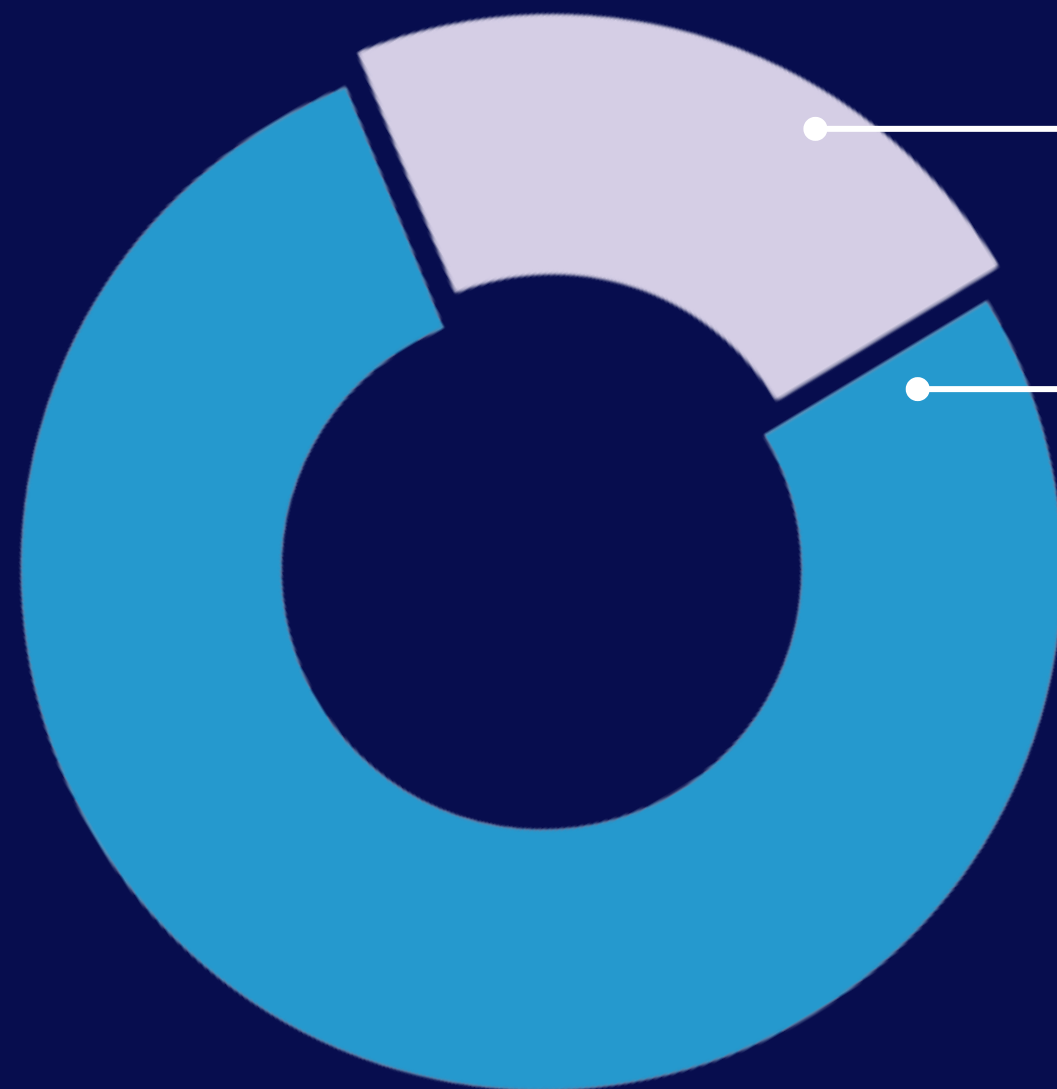
Отсевивание классификатором

LabelMe



Детекции аномалий в данных

Время на исправление некачественной разметки



27% - исправление данных

+

73% - вся разработка

=

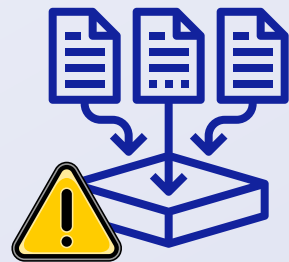
Издержки на зарплату сотрудников
Существенные отклонения от бизнес-плана
Издержки на переразметку
Дополнительная нагрузка на штат

ПОЛНОТА ДАННЫХ

Данные могут и качественными, и не сырыми, а технология не работает. Почему?



Ошибки дата-инженеров в работе с данными



Датасет не покрывает все кейсы используемой технологии

Как возникает проблема полноты данных?

- Изначальная инструкция по сбору или разметке данных не включает все кейсы работы технологии
- Не проводились тестирования технологии
- За основу берутся не кастомные решения, а ограниченные open source датасеты

Как добиться исчерпывающей полноты данных?

Только анализом кейсов применения технологии, досбором и доразметкой данных

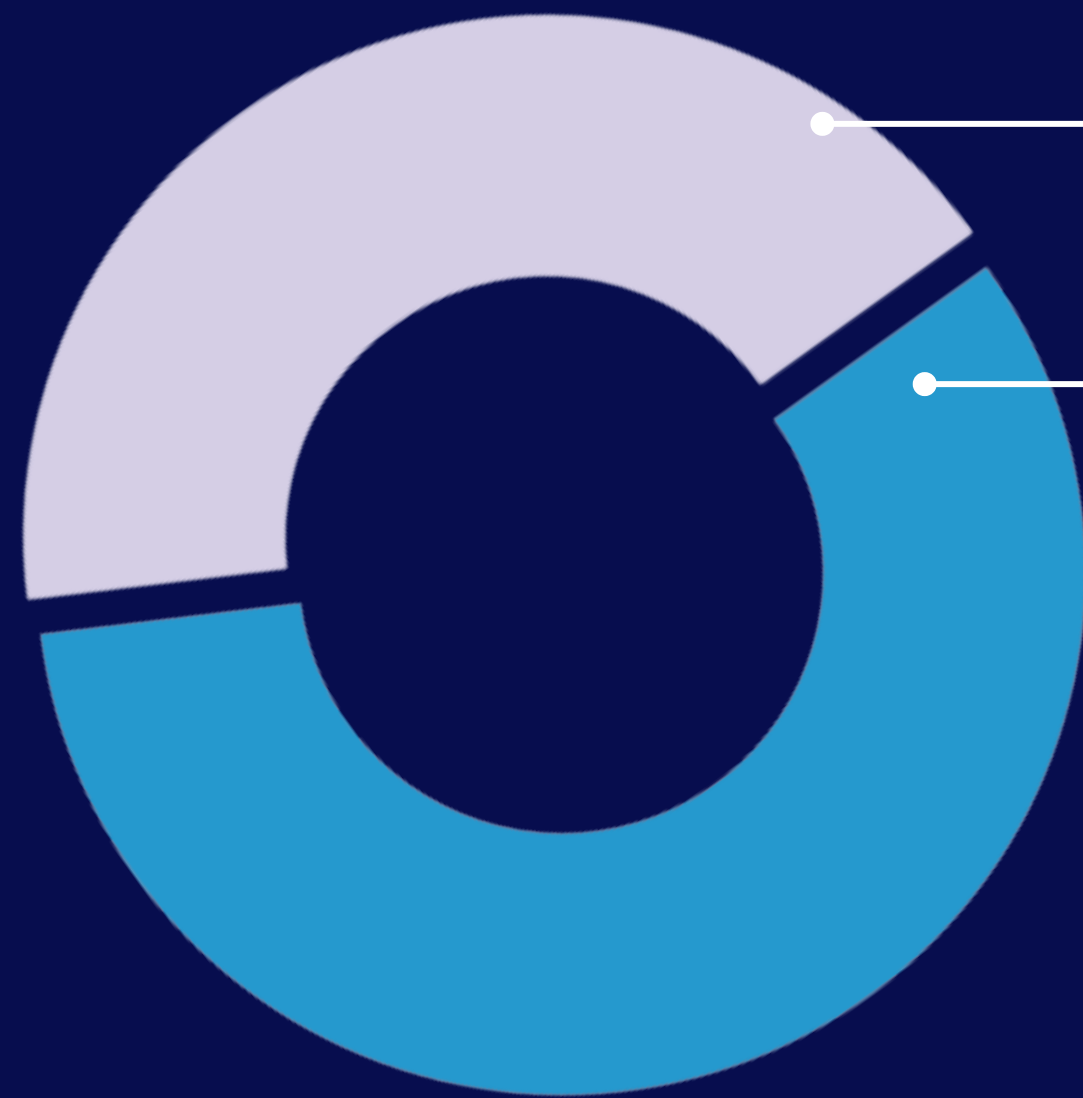
Изучить кейсы смежных проектов

Продумать ТЗ

Запустить досбор и доразметку

Адаптировать технологию под новые данные

Время на исправление некачественной разметки



42% - исправление данных

+

58% - вся разработка

=

Издержки на зарплату сотрудников
Критические отклонения от бизнес-плана
Издержки на досбор и доразметку данных
Дополнительная нагрузка на штат

Как LabelMe добивается лучшего качества разметки

- Собираем бесплатный тестовый датасет, чтобы предотвратить проблемы и согласовать все нюансы
- Внимательно изучаем и дополняем ваше ТЗ
- Допускаем к разметке только исполнителей с опытом
- Проверяем разметку на каждом этапе работы
- Выстраиваем работу в единой системе для получения однородных результатов



- 1 Включите камеру на телефоне
- 2 Сканируйте qr-код
- 3 Давайте решать проблемы с данными вместе