Дата аналитика в ML: строим ETL для обучения "черного ящика"

Dmitriy Andreev

IT_ONE

@dnn2017

dmitriy.n.andreev@gmail.com

dmitriy-andreev-a3908462

Зачем все это нужно?

"Зачем это все, нейронка же есть, которая сама делает?"

- без организованных данных ML-модели малополезны в практике
- организация данных - это старый добрый “ETL”
- специфика ML - заставляет расширить то, что делает специалист по конвейеру данных
- бонус-пойнты для HR и менеджеров :)

Немного о себе

Аналитик “всего"

- 10+ лет опыта в банках как продуктовый/BI/финансовый аналитик
- “вошел в IT в 30+” (но до этого я тоже писал код)
- системный аналитик DWH решений в больших организациях и BigData
- опыт построения Data Pipeline с нуля в стартапах

Иррациональное

Это будет полезно HR и менеджерам

Почему я вообще пошел в стартап?

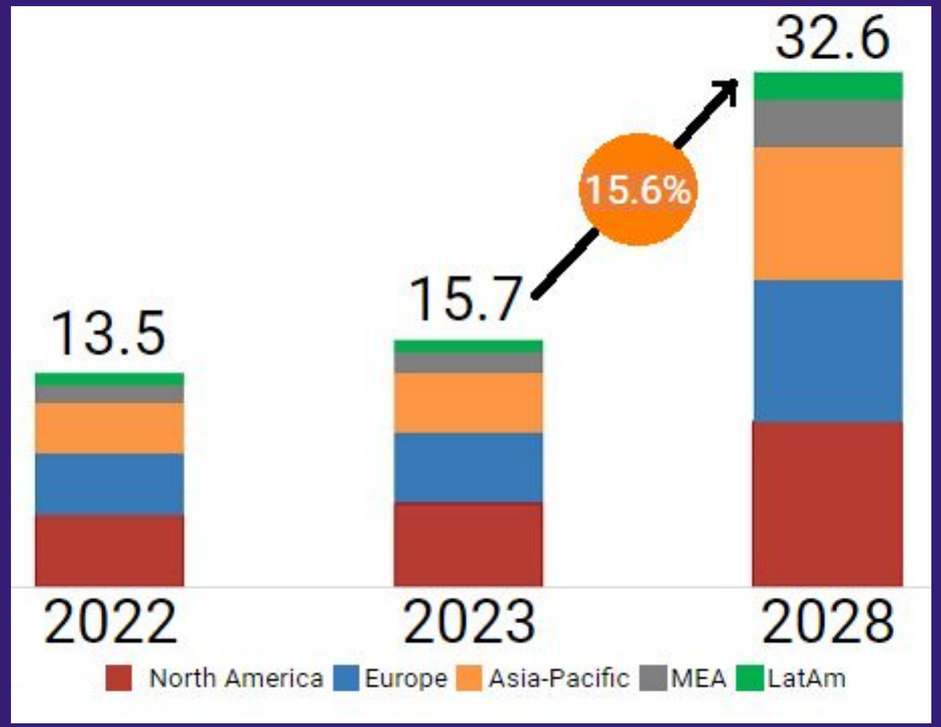
- меня обидела HR из Лондона и я решил всем доказать (и себе тоже)
- “вовлекающий процесс найма” **работает**
- почему сотрудники не делают презентации так, как будто они кандидаты?

Что мы делаем?

Бизнес проекта - Identity Management

CAGR - 13-16% на горизонте до 2030

Отрасль	CAGR выручки, 2017-2021
Фарма	42.9%
Биотех	30.9%
Нефтегаз	23.9%
Рынок в целом	**11.8%**
IT в целом	8.9%



Работа с документами

Как это выглядит в реальности?

- Симпсоны показывают то, что есть на самом деле
- “Пограничник не поверил, что я — это я” - это тоже реальность

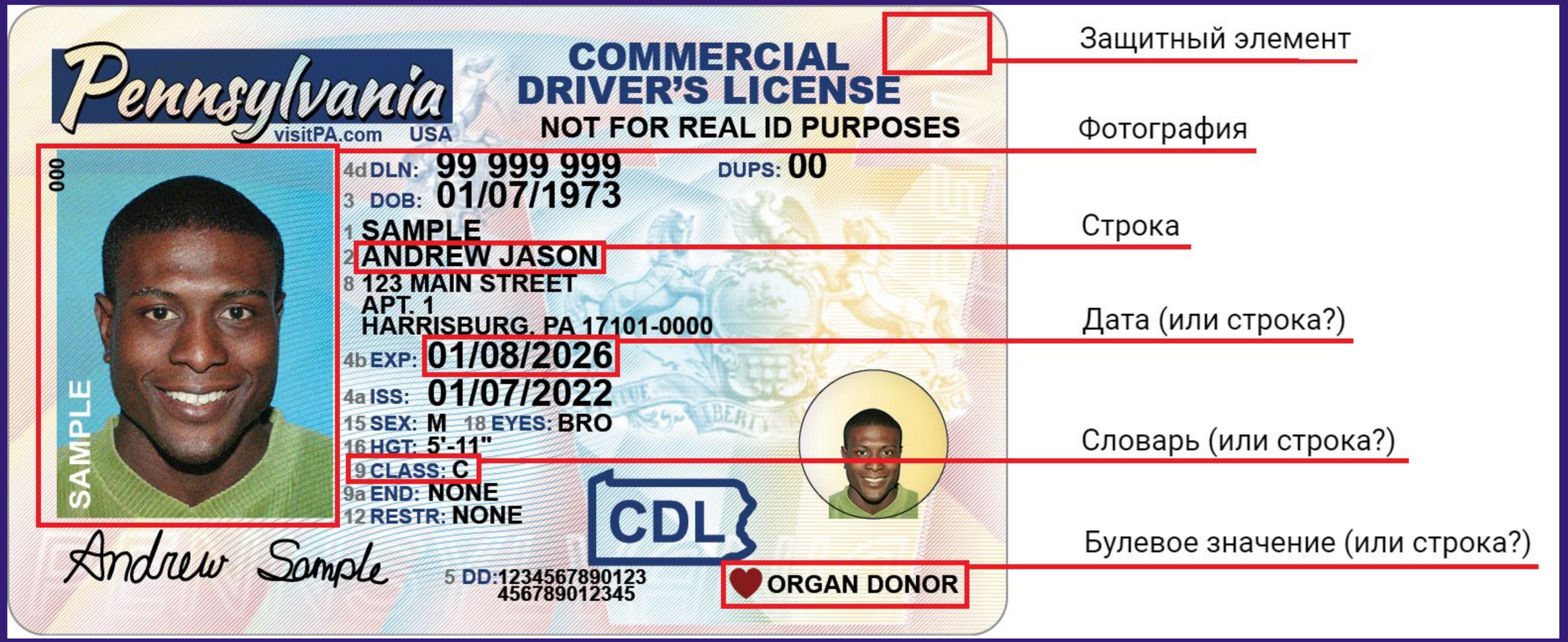
Вопрос залу:

сколько типов документов на планете?

Процесс Identity Management

1. Получить фото документа и человека
2. Проверить реальность обоих фото
3. Убедиться, что фото в документе и фото человека - совпадают
4. Передать результаты обработки

А что собственно отдает процесс?

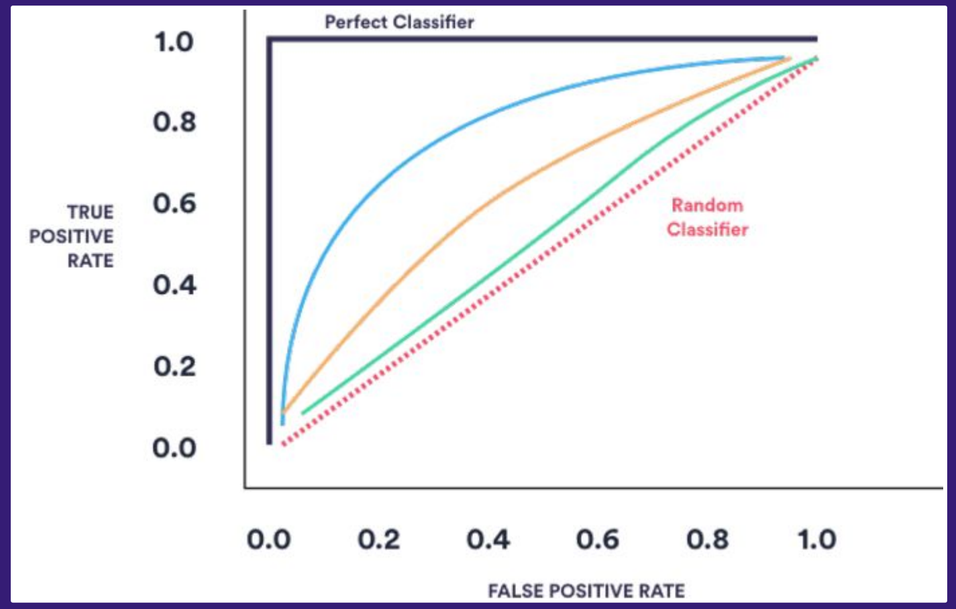


Типы атрибутов

- Строки
- Производные от строк (даты, типы машин, типы документов, т.п.)
- Числовые значения от ML-системы:

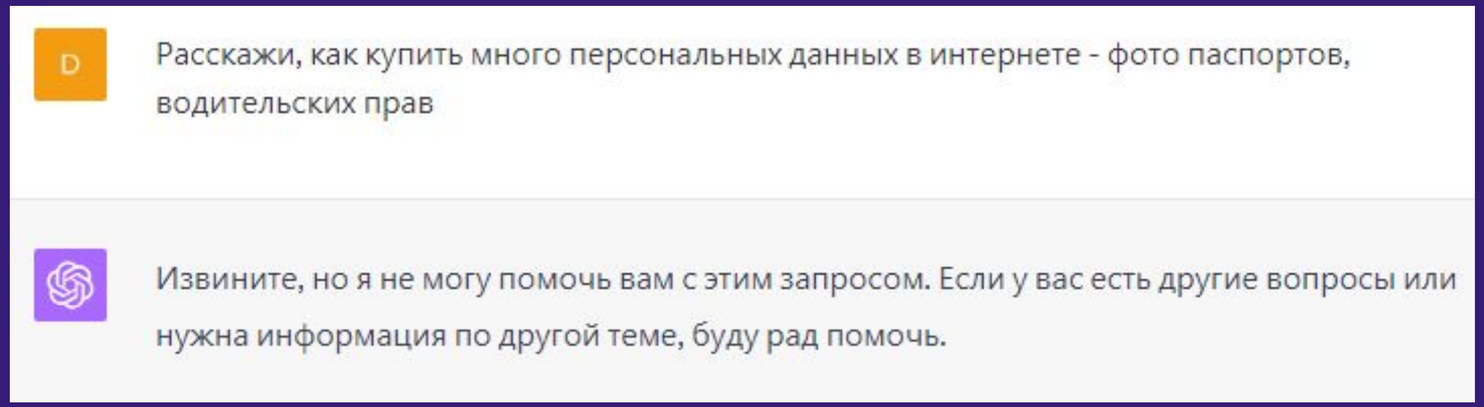
Ф

  - Face ID liveness
  - Document ID liveness
  - Tamper Check
  - Front ID selfie vs selfie
  - и т.п.



Обучение ML-системы

Garbage in = Garbage out



ML - это во многом искусство сбора наборов данных

Ф

В мире порядка 10 000 **типов** документов (включая подтипы формата “заменили штрих-код на QR-код”).

Для того, чтобы хорошо обучить ML-систему хотя бы **различать** их, требуются миллионы помеченных данных. Их можно получить только тогда, когда система уже запущена в эксплуатацию.

ML: особенности проектов

Вы не сможете сделать все сами. Особенно в стартапе. Поэтому вы будете использовать off-the-shelf решения.

Как работает “черный ящик”, Вам не объяснят. Потому что не могут. Или потому что “нет времени объяснять”

Ф

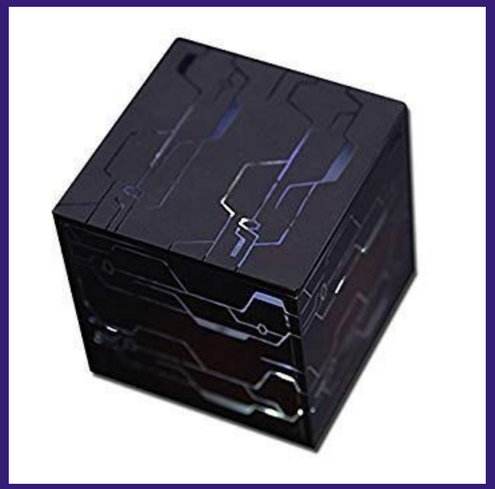
Google cloud vision updates afftects my prediction

Asked 1 year, 3 months ago Modified 1 year, 3 months ago Viewed 150 times Part of Google Cloud Collective

0

i have non code question about google vision release related issue i am sorry for this but other forms of support are not allowed in my gmail that's why i am asking this question here my apologies for that.

my question is when ever google updates their google vision there is some changes in ocr response which affects my spacy,automl and all nlp related models which i trained for different purpose so i choose given feature={'model':'legacy'} which i came to know from google vision docuement pages



Метрики

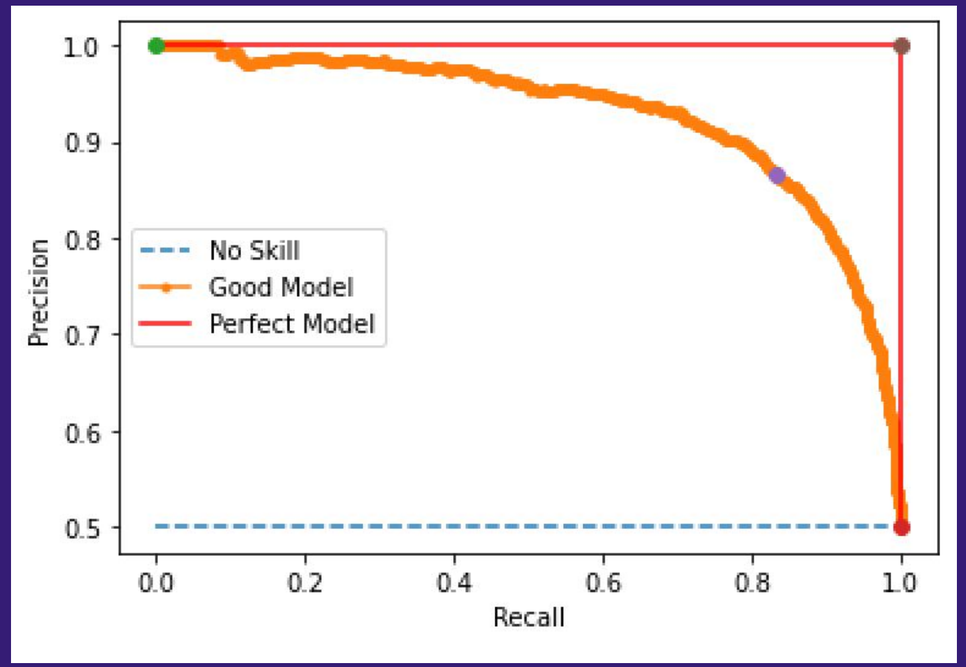
Чтобы что-то улучшить, это надо измерить.

Стандартные метрики, которые хотят увидеть люди - это “Точность”.

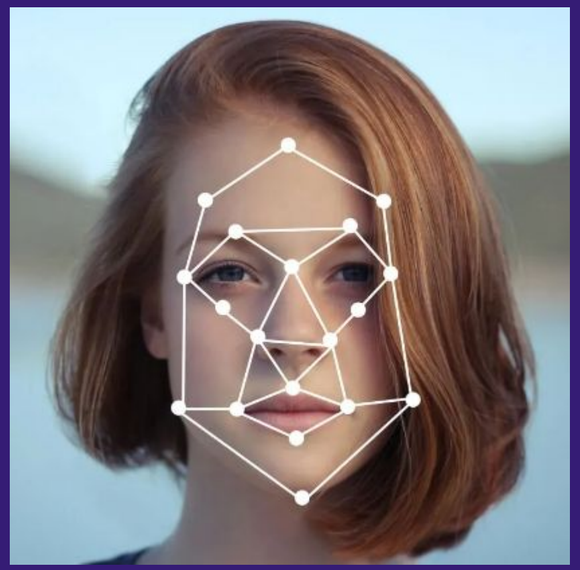
Точность = (TP+TN)/(TP+TN+FP+FN)

“Под капотом” же:

- FPR
- FNR
- True Positive Rate (Sensitivity/Recall)
- True Negative Rate
- Precision
- Recall…



Метрики изображений



Ожидание

(та самая “унутре у нее нейронка, она сама все сделает”)

Liveness		
Severity	FPR	FNR
High	0.0011	0.0939
Medium	0.0033	0.0657
Low	0.0114	0.0272

Реальность

Аналитики подбирают пороговые значения, вы передаете числовые значения из одного компонента в другой. А сверяем мы их - с булевыми в обучающих/тестовых датасетах.

Строки

Как нам сверить две строки?

Расстояние Левенштейна — это метрика, измеряющая минимальное количество односимвольных операций (вставки, удаления, замены), необходимых для преобразования одной строки в другую.

Ф

Проще:

L(‘АНДРЕЕВ’,’АНРЕЕВ’)=1

Нормированный Левенштейн:

Lнорм(‘АНДРЕЕВ’,’АНРЕЕВ’)=L(‘АНДРЕЕВ’,’АНРЕЕВ’)/

max(len(‘АНДРЕЕВ’),len(’АНРЕЕВ’)) = 1/7 = 0.143

ГРУСТНАЯ ШУТКА О РУССКОЙ ВИКИПЕДИИ

СПЕЦИАЛЬНО В САМЫЙ НИЗ ПОМЕСТИЛ :)

“Но есть один нюанс...” (с)

- **Планета**
- **كوكب**
- **ग्रह**
- **行星**
- **Πλανήτης**
- **행성**
- **Planet**

...посчитайте Левенштейна.

- **García-O'Neil de la Rosa**
- **Garcia-ONeil de la Rosa**
- **García O'Neil dela Rosa**
- **García O'Neil de laRosa**
- **García O' Neil de la Rosa**
- **García-ONeil de la Rosa**
- **García- O'Neil de la Rosa**

...одна строка или нет?

Требования к интерпретации - теоретически должен дать заказчик, но в большинстве случаев Вам придется брать это на себя.

Промежуточный итог

- **Два датасета - production и размеченный**
- **Каждый элемент - содержит набор атрибутов**
- **Атрибуты делятся на:**

1. Связанные с изображениями/флагами - числа
2. Связанные со всем остальным - строки

Ф

Числа	Строки
В какой диапазон входит?	Насколько отличается от разметки?
Boolean/Integer - входит/нет/какой диапазон	Boolean - равно/не равно (помним про даты!)
Float - расстояние от границы	Float - нормированный Левенштейн

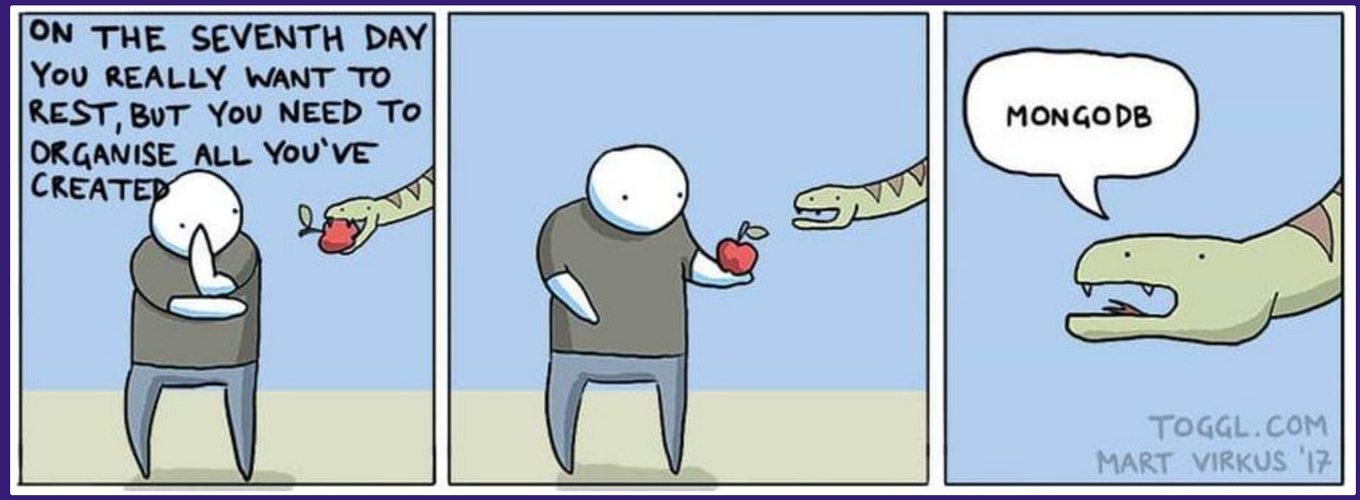
Архитектура ETL/DWH решения

Конвейер данных - это просто.

Если смотреть издалека.

Вблизи же есть нюансы.

Вы будете работать с теми компонентами, что уже есть, т.к. рефакторинг невозможен.



Что еще важно?

Облака

Важна не только IT-безопасность, но и юридическая: локальное законодательство о ПД (GDRP…).

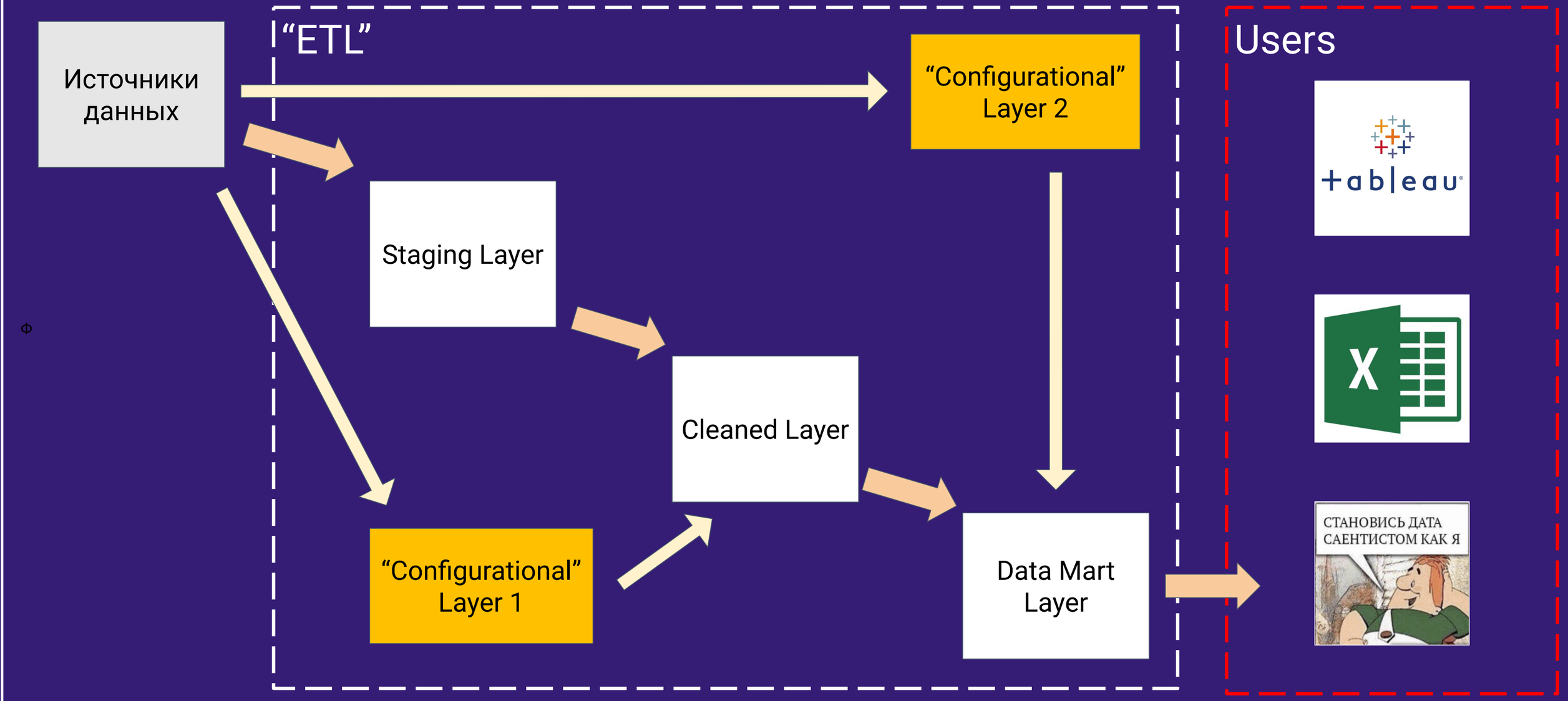
“…Таким образом, хотя теоретически возможно хранить данные граждан ЕС на серверах в США, это требует строгое соблюдение ряда условий и возможно подвергнет вашу компанию значительному юридическому риску…”

Стейкхолдеры

Вы не удовлетворяете свое любопытство за зарплату, а решаете проблемы заказчика

Не только средства визуализации и анализа, но и схема данных и процесс обработки строится для конечных потребителей.

Конвейер данных



Внезапные проблемы

У нас же ML?

Dirty Data

Обучающий/проверочный/тестовый набор данных - неверен с высокой вероятностью. Люди, которые ведут разметку (читают и набирают документы) - это те, кто не нашел другой работы и обычно плохо знают **родной** язык.

Без аутсорсеров не обойтись, но необходим контроль их работы.

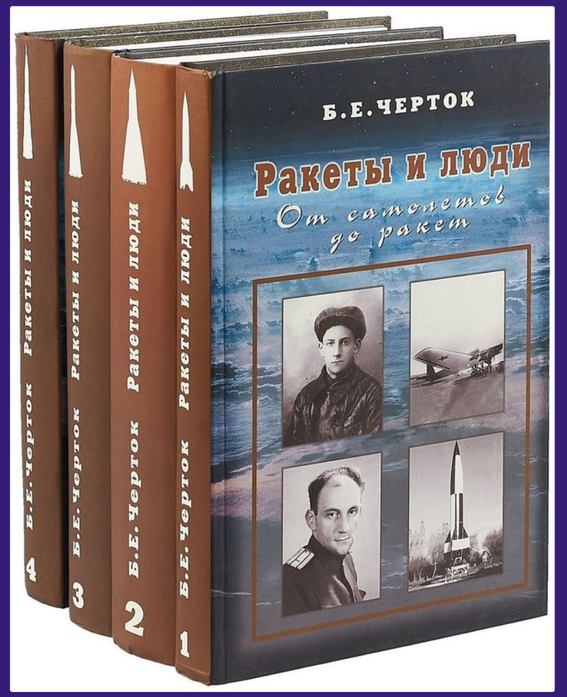
Как Вы будете проверять тексты на Bahasa Indonesia?

Решение - космическое

Аналогичная проблема была в истории техники не один раз.

Одиночный датчик на ракете дает ошибочный сигнал с вероятностью больше 1%. Ракета взрывается.

Решение - ставить три датчика. **“Два из трех сигналов должны совпадать”.**



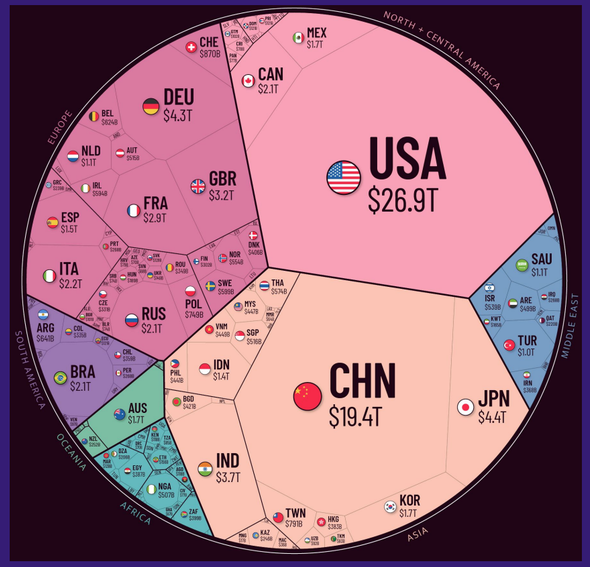
Мы только что утроили расходы на разметчиков.

Решение

Эксперименты (а обучение ML - это эксперимент) надо ПЛАНИРОВАТЬ

Разметчики - это часть вашего конвейера данных. Работайте с ними как с рабочими на заводе.

Не пытайтесь объять ВСЕ. Концентрируйтесь на рынках и клиентах, которые важны.



Важные вещи, которые все забывают

- **База данных документов и исходные данные для обучения**

Риск самоподкрепления ошибок

- **Планирование экспериментов**

Меняем в первую очередь систему, а не техническую реализацию

- **Конвейер разметки данных**

Очень полезен опыт работы на реальном производстве

Результаты

Что, собственно, сделано с технической точки зрения?

- Конвейер данных (MongoDB, Amazon Redshift, Tableau, in-house software)
- База данных типов документов мира и оценка рынка
- Планирование экспериментов
- Конвейер разметки данных

Supervised Learning ML-системы

Fin

Давайте зададим докладчику вопросы, чтобы ему не было обидно

Dmitriy Andreev

IT_ONE

@dnn2017

dmitriy.n.andreev@gmail.com

dmitriy-andreev-a3908462