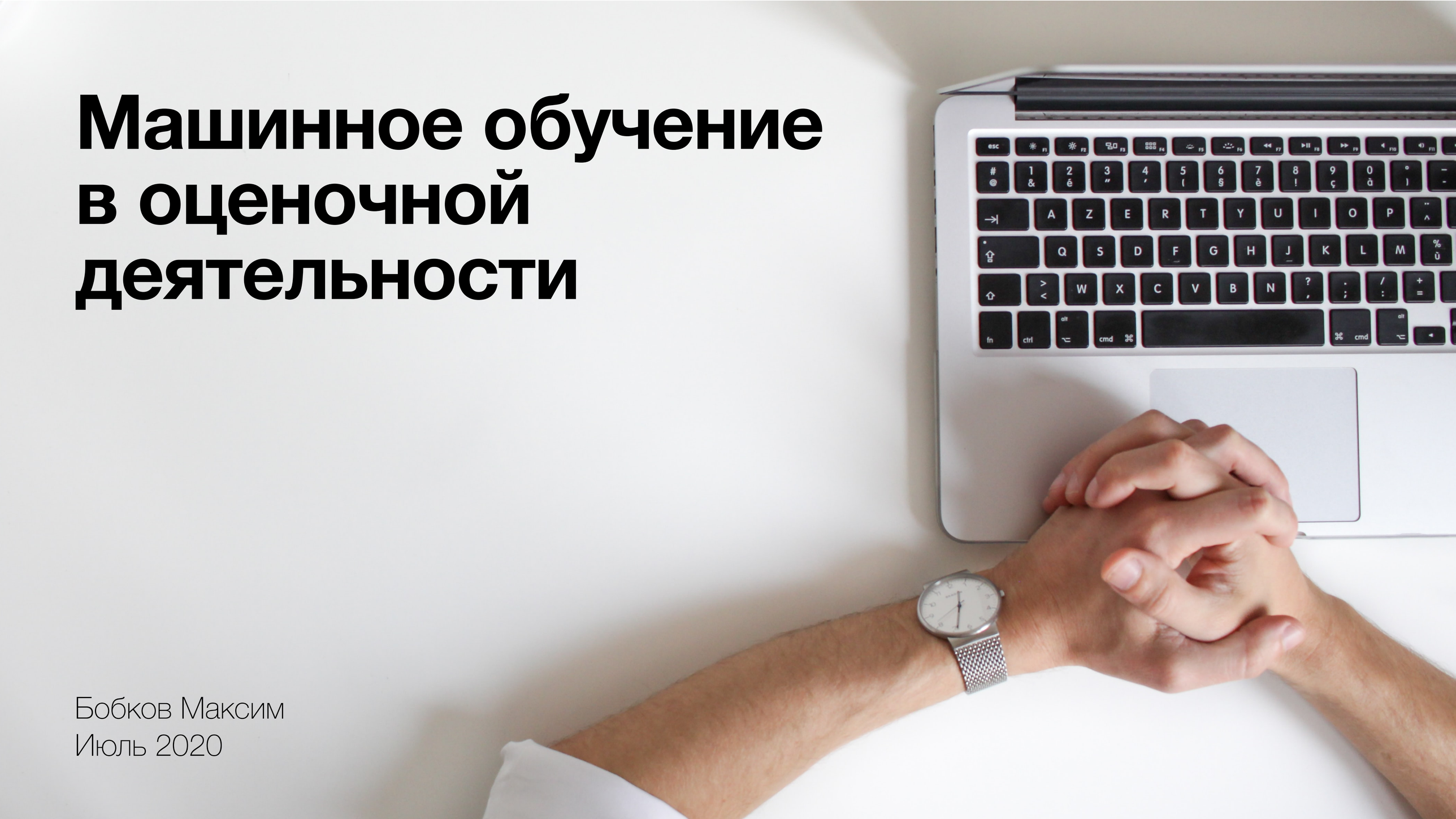
Машинное обучение в оценочной деятельности



Бобков Максим
Июль 2020

Немного о себе

- Занимаюсь оценкой 17 лет, из них последние 10 в банковских залогах
- Программирую на Питоне 2 года, интересуюсь машинным обучением и экспериментирую с ним в оценке стоимости
- Написал и внедрил в рабочий процесс автоматизированную оценку автомобилей, сократив время выпуска заключения с 30 до 5 минут
- Автор бота @flattabot в Телеграме, бот вставляет фотографии объекта в Ворд и высылает готовый файл на почту
- Написал бота @date2picbot, проставляющего дату на фотографии с телефона
- Создал канал в @gosocenka, в котором автоматически публикуются заказы на оценку с сайта https://zakupku.gov.ru
- Почта: max.bobkoff@gmail.com
- Телеграм: @maxximax

Незаметные технологии

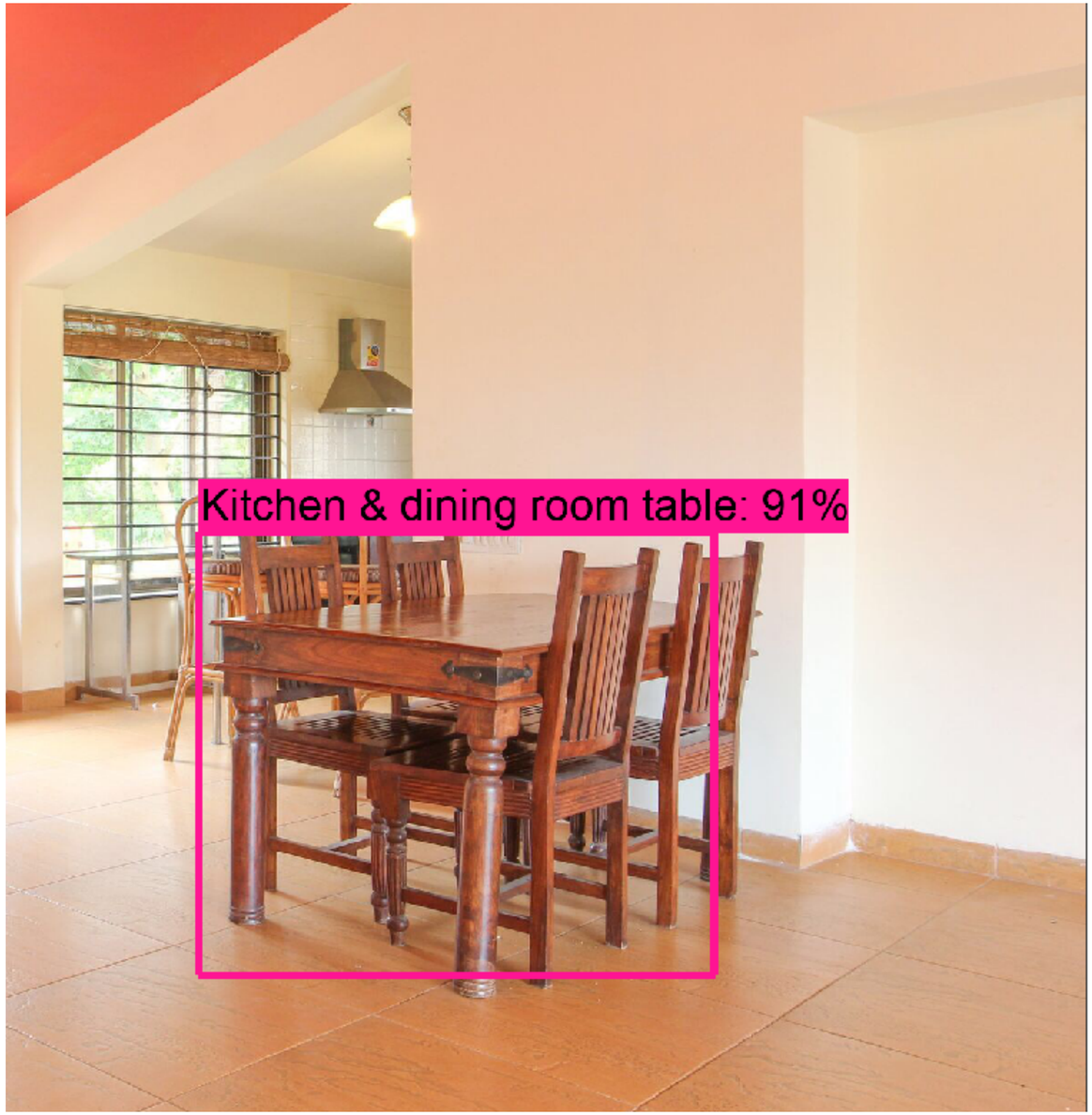
Яндекс.такси
проверяет автомобили

- Водитель фотографирует автомобиль перед выходом на смену
- Ручной контроль был медленный, приходилось жертвовать качеством
- После внедрения машинного зрения согласование ускорилось
- Проверяется цвет, номер, модель, чистота автомобиля и отсутствие повреждений
- https://habr.com/ru/company/yandex/blog/433386/

airbnb.com

контролирует объявления

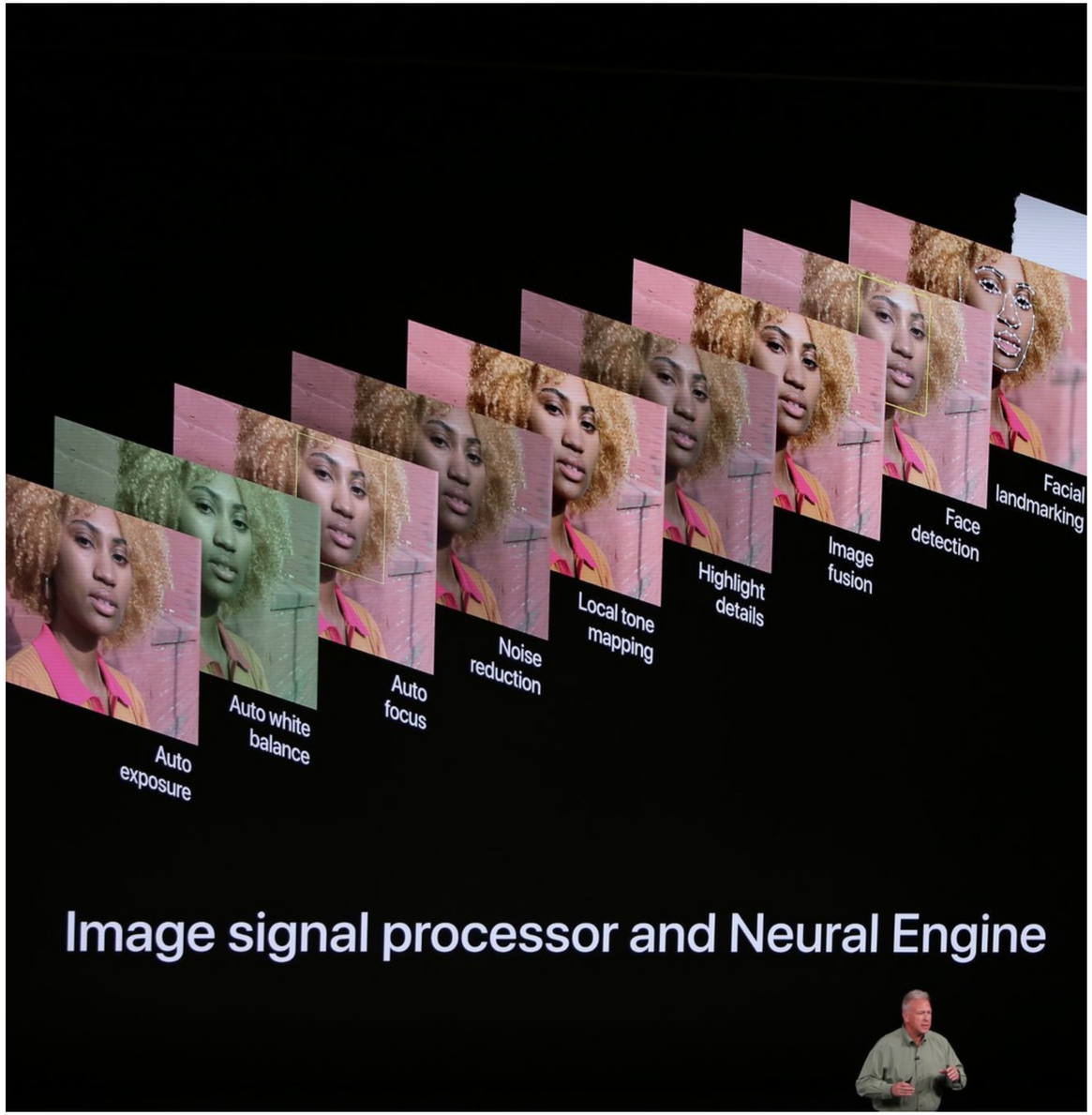
- Миллионы пользователей загружают фотографии своих домов, платформе приходится бороться с фейками
- Машинное обучение помогло определять не только качество фотографий, но и тип помещений/ удобств, сверяя их с заявленными в объявлении
- https://medium.com/airbnb-engineering/categorizing-listing-photos-at-airbnb-f9483f3ab7e3



Фотоаппарат

делает красиво

- Камеры в телефонах снимают лучше зеркалок несмотря на крошечные объективы
- Современная фотография больше вычисляется, чем просто фиксирует свет на матрице
- https://vas3k.ru/blog/computational_photography/



Ближе к делу

Искусственный интеллект

понятие модное, но переоцененное

- Искусственный интеллект - маркетинговое словечко, помогающее продавать, которое вставляют в рекламу по поводу и без
- Нейронные сети не повторяют структуру мозга разумных существ, несмотря на свое название
- Машинное обучение - ряд технологий, позволяющих натренировать алгоритмы на имеющихся данных, чтобы они могли с определенной долей ошибок делать выводы на основе новых данных того же типа
- Несмотря на то, что такие алгоритмы работают гибче и точнее, чем набор заранее написанных инструкций для компьютера, здесь нет никакой магии, вместо нее сложные функции, которые оптимизируют или аппроксимируют

Границы возможностей

Алгоритмы могут

- Предсказывать
- Запоминать
- Воспроизводить
- Выбирать лучшее

Алгоритмы не могут

- Создавать новое
- Резко поумнеть
- Выйти за рамки задачи
- Убить всех людей

Виды обучений

С учителем

- Классификация
- Регрессия

Нужно много (десятки тысяч) примеров с заранее известными метками классов/значениями

Машина ищет закономерности в данных и старается применить из на незнакомых примерах

Без учителя

- Уменьшение размерности (обобщение)
- Кластеризация

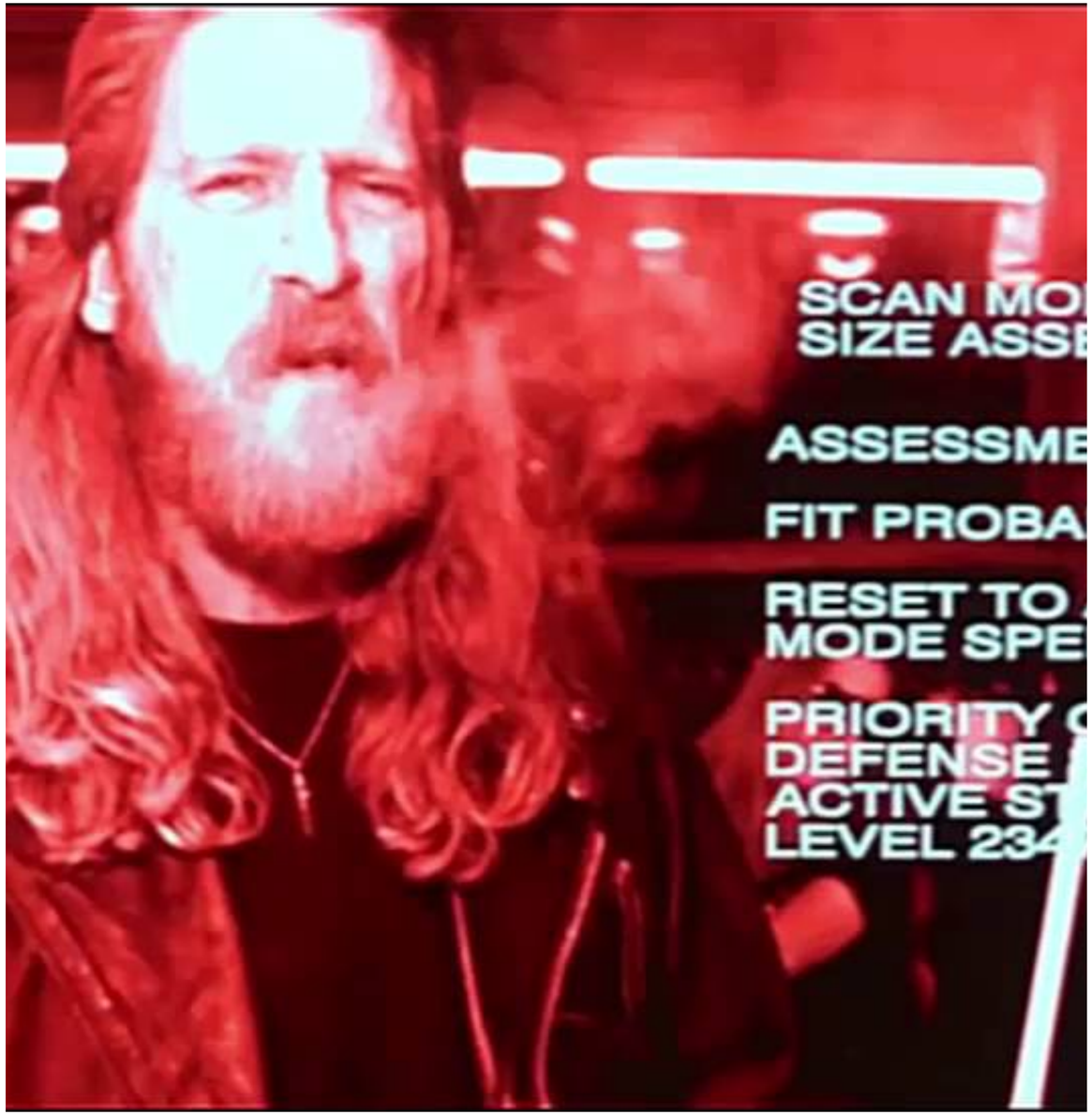
Датасет заранее не размечен

Алгоритм сам ищет общие признаки, по которым можно разбивать данные на группы

Как применять

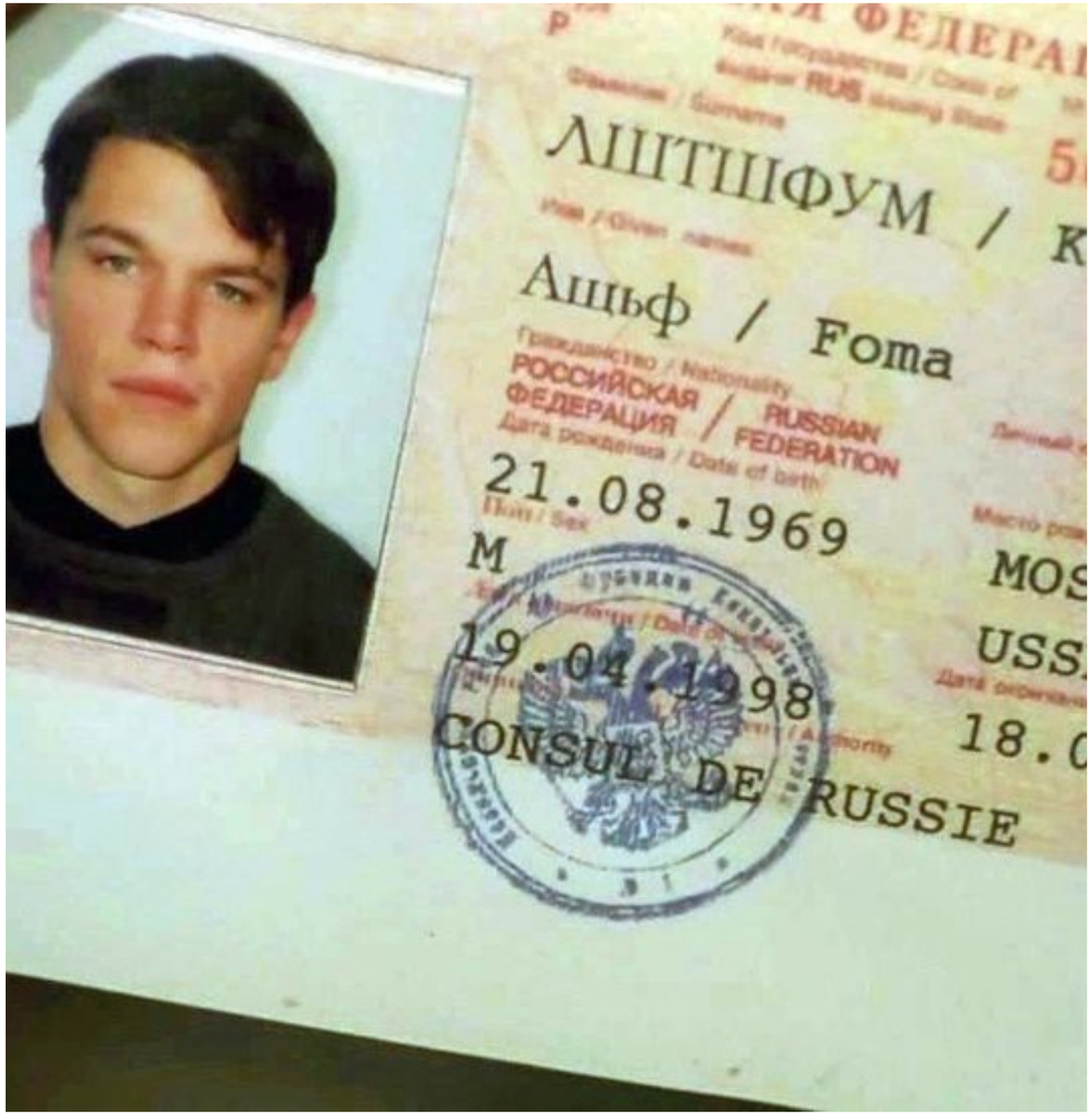
Компьютерное зрение

- Детектирование объектов - автоматический подсчет пешеходного/автомобильного трафика
- Классификация изображений - работа с большими объемами фотографий объектов/ аналогов
- https://habr.com/ru/post/461365/



Распознавание данных

- Автоматическая обработка сканированных документов
- Выявление именованных сущностей и фактов из текстов объявлений
- https://demo.deeppavlov.ai/#/ru/ner
- https://alexeykalina.github.io/technologies/tomita-parser.html



Анализ текстов

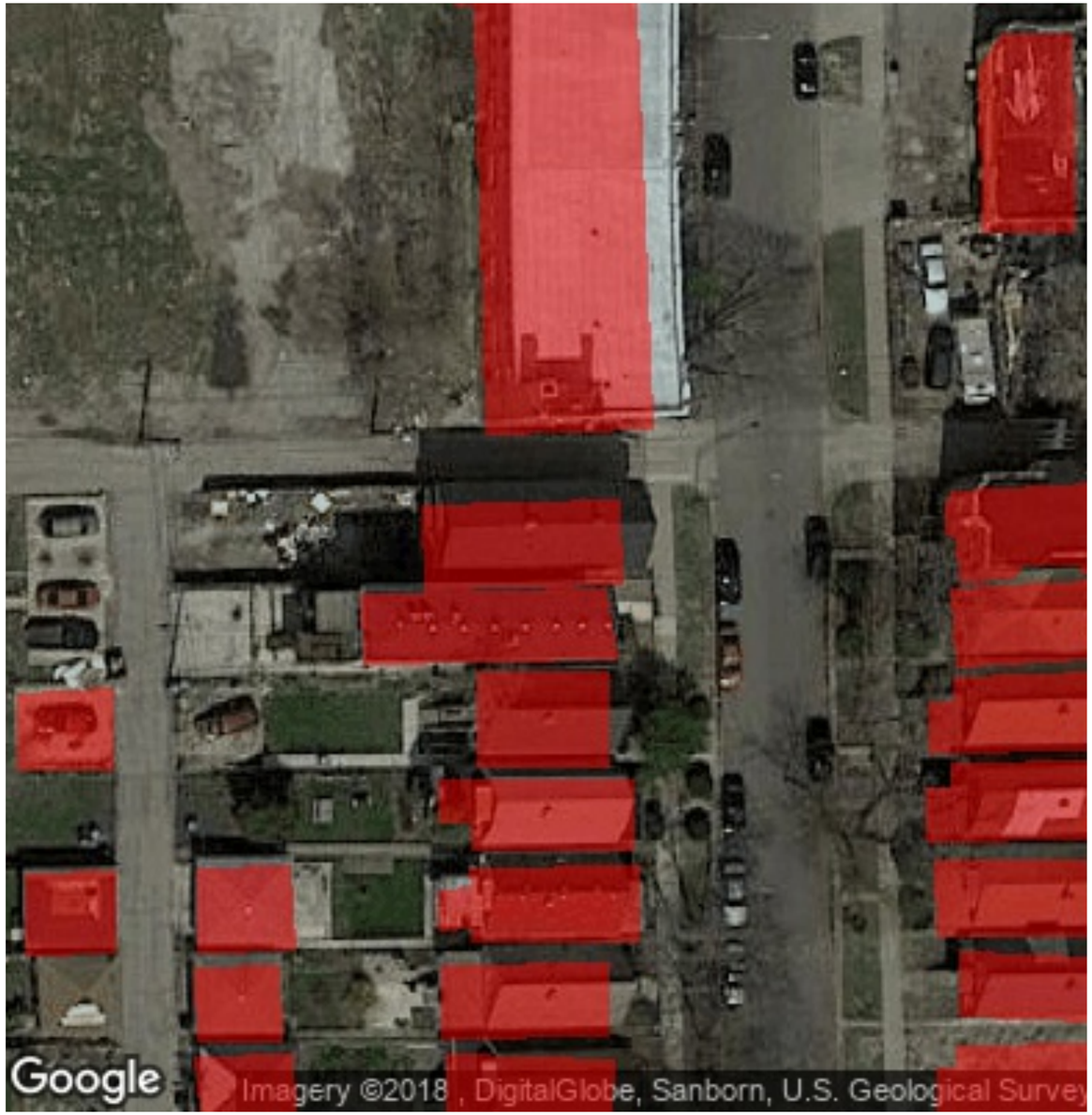
- Анализ объявлений из открытых источников
- Наработки команды Алексея Зумберга по анализу объявлений о продаже объектов-аналогов
- https://srosovet.ru/press/news/250619/
- https://demo.deeppavlov.ai/#/ru/textqa

Исследование городских территорий

- Моделирование трафика на основе данных о плотности населения, расположении транспорта, активности абонентов сотовой связи, пешеходных маршрутов в микрорайонах
- Выявление перспективных локаций
- https://bestplace.ai/ru
- https://habr.com/ru/post/270513/
- https://geophy.com/neighborhoods/

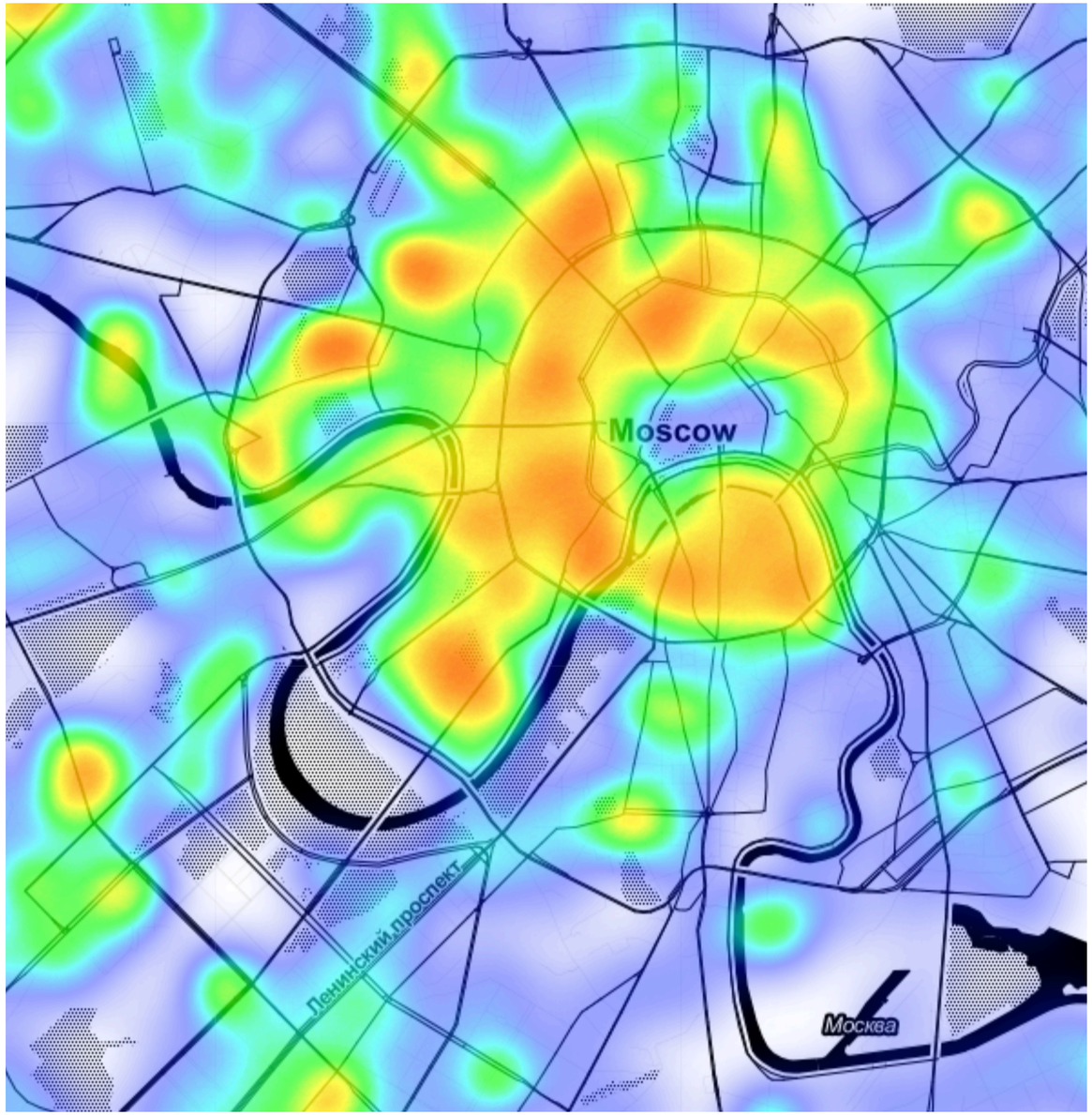
Классификация строений

- Анализ геоснимков городских территорий, выявление жилого сектора, промышленных территорий и т.п.
- Изучение структуры населенного пункта, его точек роста, зонирования
- https://github.com/Sardhendu/PropertyClassification



Предсказание цен на недвижимость

- Наиболее известный сервис автоматической оценки жилья - zillow.com
- В 2017 году Zillow проводил конкурс по машинному обучению на kaggle.com с целью улучшения своей расчетной модели: https://www.kaggle.com/c/zillow-prize-1
- Конкурс Сбербанка по предсказанию цен на квартиры: https://youtu.be/Eo4WMIcT7uo
- geophy.com - голландский сервис оценки недвижимости на основе машинного обучения



С чего начать

План действий

Оценка квартир в Москве

- Попробовать как работает машинное обучение можно на примере датасета из 63 000 объявлений о продаже, собранного мной за период ноябрь 2019 - январь 2020 г. Ссылка на гитхаб: https://github.com/maxbobkov/ml_moscow_flats

- В статье есть подробная инструкция как запустить код в облачном сервисе Google Colab: https://medium.com/@max.bobkov/machine-learning-moscow-flats-appraising-25a1e9f171db

- С кодом можно экспериментировать и брать за основу своих исследований

Где искать датасеты

- https://datasetsearch.research.google.com/
- https://opengovdata.ru/it/dataset
- https://www.kaggle.com/datasets
- https://data.gov.ru/
- https://data.mos.ru/
- https://www.gks.ru/opendata/

Парсинг данных

- Парсинг - автоматический сбор информации с веб-сайтов, программа-парсер запрашивает у сервера все интересующие нас страницы, анализирует html-код, вытаскивая оттуда сведения, по заранее описанным правилам и сохраняет данные в табличном виде для дальнейшего использования
- Существуют как готовые сервисы для парсинга: https://habr.com/ru/post/340038/
- Так и библиотеки для самостоятельного написания парсеров (для Python, к примеру - requests, beautiful soup, selenium): https://nuancesprog.ru/p/2715/
- Сервисы противодействуют парсингу, вычисляют и банят роботов, поэтому приветствуется творческий подход и аккуратность (делаем задержки между запросами, чтобы не попасть под блокировку)

Обработка и анализ данных

- Сырые данные содержат пропуски, некорректные или аномальные значения, ошибки. Перед обучением модели датасет необходимо изучить, очистить, подготовить, сгенерировать дополнительные признаки (Feature engineering)
- Исследование датасета в десятки тысяч строк - нетривиальная задача, решаемая с помощью визуализаций, изучения статистических характеристик
- Подробный пример такого анализа и обработки датасета с фрагментами кода на Python можно увидеть по ссылке: https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python

Построение предсказательных моделей

- Дерево решений (Decision tree) — один из самых известных алгоритмов машинного обучения, хорошо работающий с табличными данными. Его суть — последовательно разбивать (сплитить) выборку на все более мелкие части так, чтобы в конце максимально точно разложить классифицировать примеры
- Для регрессии деревья тоже хорошо работают, в отличие от дискретной классификации, предсказываются точные значения (в нашем случае - цена предложения квартиры
- Случайный лес (Random forest). Идея случайного леса в том, что строится сразу много деревьев решений, слабых по отдельности, но сильных своим общим вердиктом (при регрессии предсказанные значения деревьев усредняются, при классификации финальное решение выбирается голосованием). При этом, каждое из слабых деревьев строится только на части датасета.

Дерево Решений

Построение предсказательных моделей

- Следующим шагом в борьбе за качество моделей стал градиентный бустинг, позволяющий делать много циклических заходов с построением деревьев, обращая особое внимание на неточности, совершенные на предыдущем круге, делая “работу над ошибками” и доучивая предыдущую версию модели.
- Последние годы в соревнованиях по ML почти всегда лидируют решения, построенные на алгоритмах бустинга XGBoost и LightGBM.
- Яндекс разрабатывает открытую библиотеку машинного обучения CatBoost, работающую “из коробки” с категориальными признаками (в отличии от остальных алгоритмов, которые не могут переварить понятия “панель”, “кирпич” и “монолит” в чистом виде, требуя кодировать слова цифрами, CatBoost берет датасет в работу даже в таком виде)
- https://medium.com/nuances-of-programming/алгоритм-xgboost
- https://towardsdatascience.com/predicting-airbnb-prices-with-machine-learning-and-deep-learning-f46d44afb8a6

Метрики качества

- Для оценки качества модели датасет случайным образом делят на обучающую и тестовую выборки, как правило, в соотношении 70/30 процентов, чтобы сохранить часть примеров в тайне от алгоритма.
- Модель учится на первой выборке, сопоставляя признаки и итоговый результат. Затем получает характеристики объектов из тестовой выборки и пытается предсказать значения по ним.
- Последним действием вычисляется ошибка - насколько предсказанные значения отклонились от фактических.
- Метрики качества позволяют сравнивать модели между собой и понимать, насколько можно доверять результатам работы алгоритма.
- В своих исследованиях оценки квартир в Москве достигал абсолютной медианой ошибки в пределах 5,2%.
- Подробнее о метриках для моделей: https://habr.com/ru/company/ods/blog/328372/

Черный ящик

объяснимость работы моделей

- Почему алгоритм выдал именно такой результат? Как он к нему пришел?
- Недоверие к моделям из-за непрозрачности их работы
- Одиночные деревья решений - наиболее объяснимые модели, ансамбль деревьев (случайный лес) уже нет
- Нейронные сети - максимально черный ящик
- Нам доступна оценка качества моделей - проверка работы на новых данных, для которых мы знаем правильный ответ и можем посчитать величину ошибки

Особенности и недостатки

- Предсказательная способность модели тем выше, чем полнее данные и чем сильнее прослеживается зависимость между параметрами и предсказываемым значением
- В текущей экономической ситуации значительная часть ретроспективных данных для обучения моделей стала бессмысленной, в частности, это касается планов Сбербанка предсказывать выручку предприятий: https://habr.com/ru/news/t/460293/
- Зато от этих факторов не страдает сфера компьютерного зрения, работающая с изображениями

Инструменты

Python

дистрибутив Anaconda

- Удобный пакет для одновременной установки Python 3.7, Jupyter-ноутбука и десятков библиотек для вычислений и машинного обучения

- https://www.anaconda.com/products/individual

Рабочая среда

Jupyter notebook

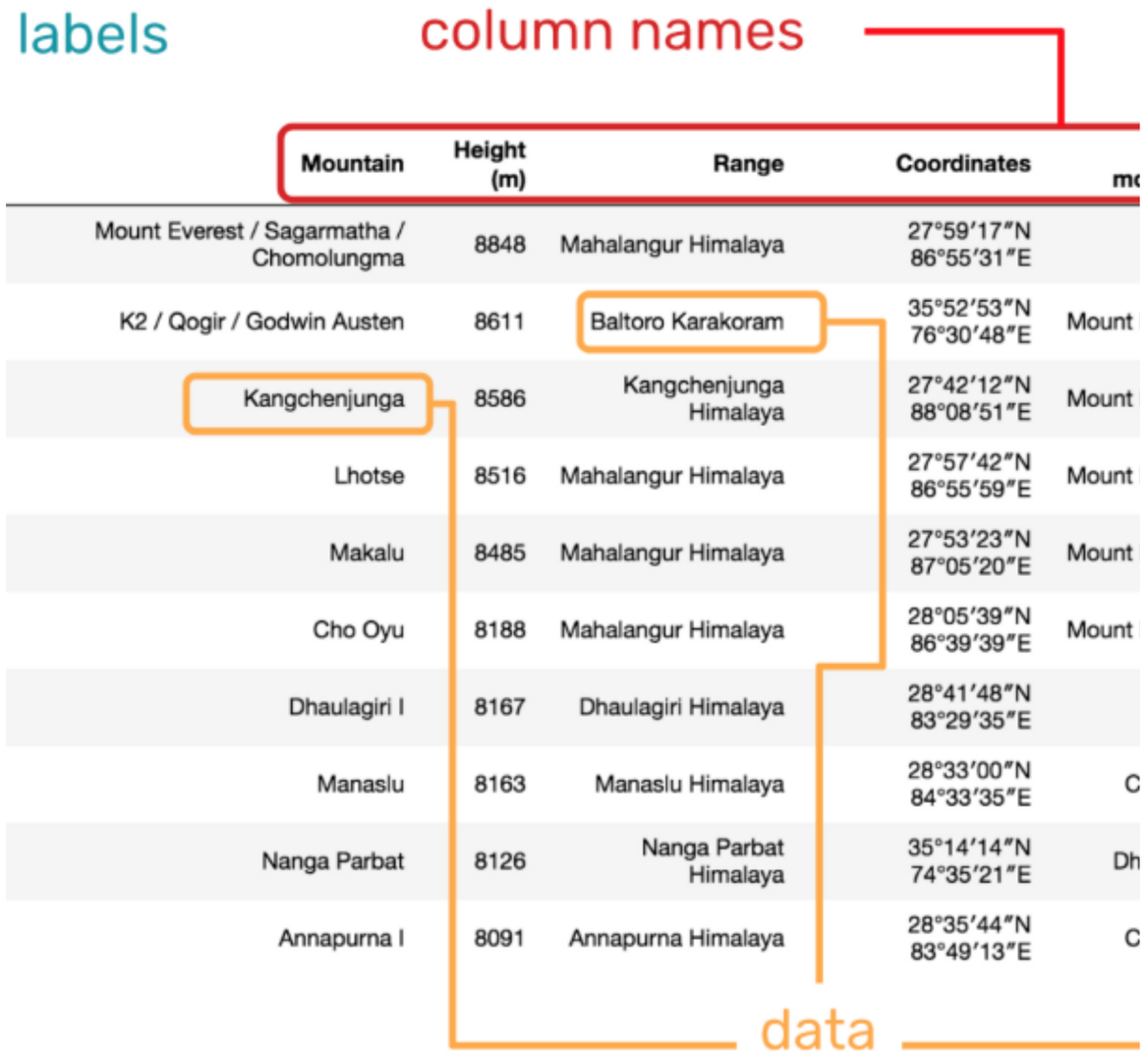
- Работает в браузере, удобен для научных исследований
- Код можно запускать отдельными блоками и сразу видеть результат
- Построение графиков и таблиц в едином рабочем пространстве с кодом
- https://jupyter.org/

.

Pandas

эксель для программистов

- Исключительно удобная и мощная библиотека для работы с табличными данными
- Фильтрация, трансформация датасетов, сводные таблицы, любые манипуляции с данными, в том числе со специфическими картографическими форматами
- https://smysl.io/blog/pandas/



Google Colab

бесплатные мощности

- Colab - это Jupyter-ноутбуки в облаке
- Бесплатно доступен мощный графический ускоритель вычислений NVIDIA T4
- Есть 12-часовое ограничение на срок непрерывной работы, не забывайте сохранять промежуточные результаты
- https://colab.research.google.com/

Куда копать дальше

Курсы

- https://stepik.org/course/58852/promo - Python для начинающих
- https://stepik.org/course/4852/promo - Введение в Data Science и машинное обучение - увлекательный и доходчивый вводный курс
- https://www.coursera.org/learn/machine-learning - Классический курс от Эндрю Ына
- https://www.kaggle.com/learn/overview - Ряд неплохих мини-курсов на Kaggle
- https://praktikum.yandex.ru/ - онлайн-школа Яндекса

Книги

- https://www.litres.ru/el-sveygart/avtomatizaciya-rutinnyh-zadach-s-pomoschu-python-prakticheskoe-rukovodstvo-dlya-nachinauschih-38272545/ - идеальная для новичков книга по Python, помогающая быстро стартовать, автоматизируя рутинные операции на компьютере
- https://www.litres.ru/pol-berri/izuchaem-programmirovanie-na-python-25562287/ - отличный самоучитель по языку Python с комиксами
- https://www.litres.ru/zed-shou/legkiy-sposob-vyuchit-python-25206565/ - автор со своеобразным стилем и методом обучения, все четко и по полочкам
- https://buildmedia.readthedocs.org/media/pdf/pandasguide/latest/pandasguide.pdf - бесплатная книга по основам библиотеки Pandas

Статьи

- https://vas3k.ru/blog/machine_learning/ - о машинном обучении простыми словами - самый доходчивый обзор технологий машинного обучения на русском языке
- https://habr.com/ru/hub/machine_learning/ - подборка статей по машинному обучению на Хабре
- https://dyakonov.org/ - блог Александра Дьяконова - специалиста в DS и преподавателя
- https://ods.ai/ - русскоязычное сообщество датасайентистов
- https://smysl.io/blog/pandas/ - ликбез по библиотеке Pandas
- https://pythonru.com/baza-znanij/jupyter-notebook-dlja-nachinajushhih - все о Jupyter notebook
- https://medium.com/@max.bobkov/machine-learning-moscow-flats-appraising-25a1e9f171db - статья о моих экспериментах с машинным обучением
- https://habr.com/ru/company/mailru/blog/462769/ - применение машинного обучения в промышленности - мегасборник ссылок на примеры кода

Соревнования по ML

- https://www.kaggle.com/competitions - актуальные соревнования на самой известной площадке Kaggle.com
- https://mltrainings.ru/ - разбор задачек с соревнований и ссылки на открытые соревнования
- https://mlcontests.com/ - агрегатор соревнований
- https://www.drivendata.org/competitions/ - еще один агрегатор

Спасибо за внимание!