

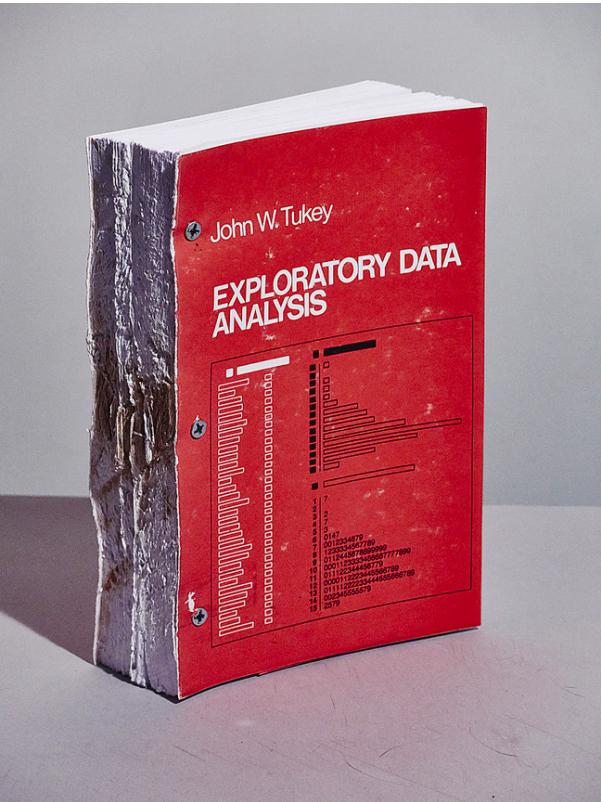
Введение в машинное обучение

МАКСИМОВСКАЯ
АНАСТАСИЯ

Что такое data science?

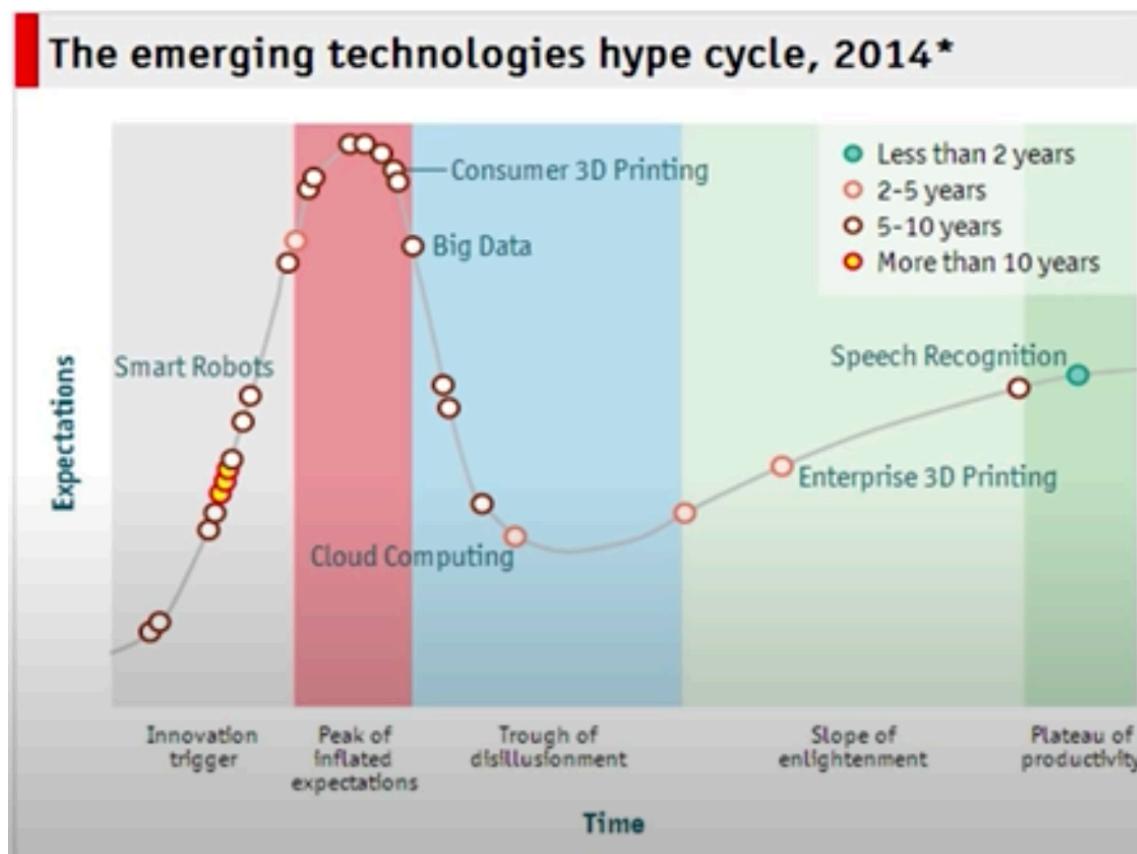
- Практическая деятельность по обработке, анализу и представлению данных
- Позволяет восстановить сложные зависимости по конкретному числу примеров

Когда все началось?

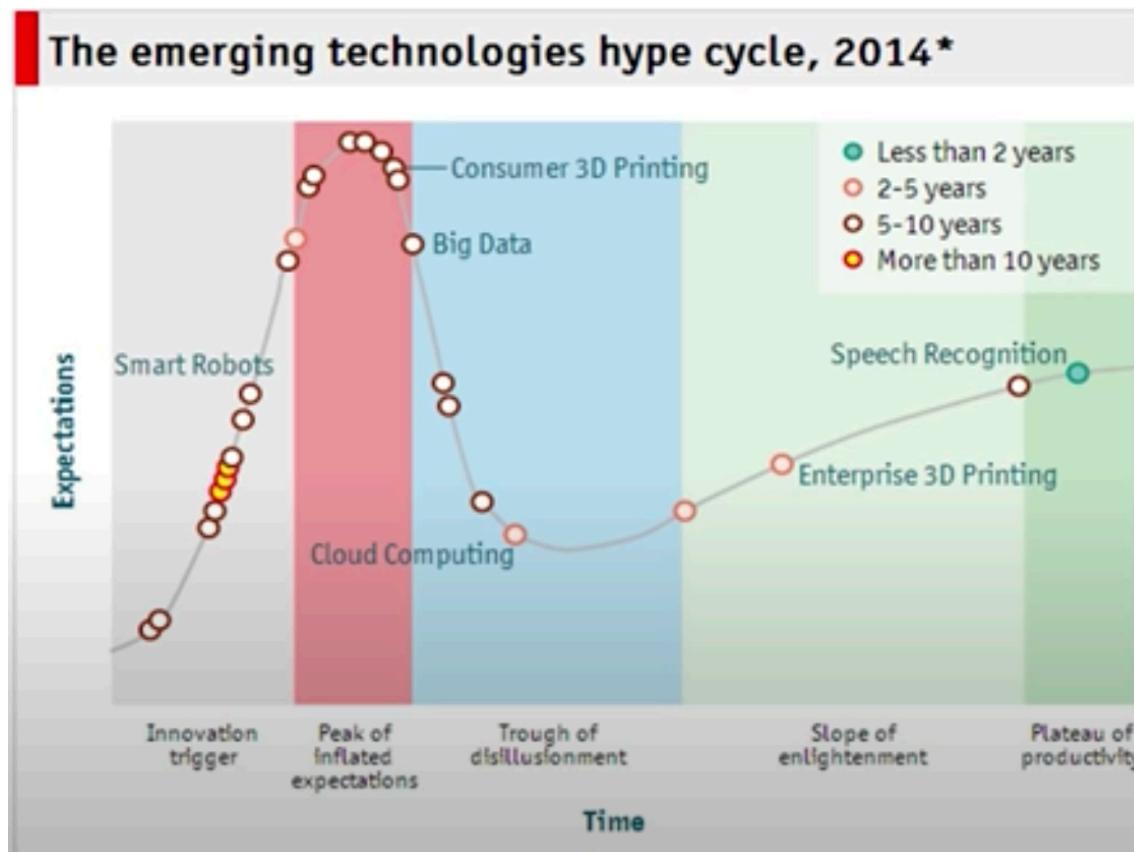


В 1960-1970х американский математик John W. Tukey публикует ряд работ, где предлагает ставить гипотезы, основываясь на данных.

Gartner Hype Cycle 2014



Gartner Hype Cycle 2014



В 2015 Gartner исключили область Big Data. Почему?

Основные термины

Пример

- Клиент – сеть ресторанов, которая хочет открыть еще один
- Несколько вариантов размещения
- Необходимо выбрать тот, который принесет наибольшую прибыль

Терминология

- x – объект, то, для чего делаем предсказание (расположение ресторана)
- \mathbb{X} – пространство объектов (все возможные расположения ресторана)
- y – целевая переменная, то, что предсказываем (прибыль ресторана)
- \mathbb{Y} – пространство решений, все значения ответа (все вещественные числа)

Обучающая выборка

- Собираем много объектов с известным значением целевой переменной
- Уже имеющиеся рестораны, их признаковое описание и прибыль

Признаковое описание

- Объект – абстрактная сущность, а компьютеры работают с числовыми
- Признак (feature) – числовая характеристика объекта

Признаки бывают:

- Числовые
- Бинарные (0/1)
- Категориальные
- Признаки со сложной внутренней структурой

Признаковое описание

- Объект – абстрактная сущность, а компьютеры работают с числовыми
 - Признак (feature) – числовая характеристика объекта
-
- Какие признаки будем собирать для данной задачи?

Возможные ответы

- Локация
- Количество проезжающих мимо машин за день
- Расстояние до ближайшего конкурента
- Средняя стоимость квадратного метра жилья поблизости
- Средний возраст жителей ближайших кварталов
- и так далее

Локация – как учесть?

- У нас есть координаты – долгота и широта
- Не очень вероятно, что есть взаимосвязь между этими двумя числами и целевой переменной – прибылью
- Что делать?

Локация – как учесть?

- У нас есть координаты – долгота и широта
 - Генерируем признаки на основе этих данных (feature generation)
-
- Какие можно придумать?

Локация – как учесть?

- У нас есть координаты – долгота и широта
- Генерируем признаки на основе этих данных (feature generation)

Примеры:

- удаленность от центра, от ближайшей станции метро
- если рестораны в разных странах – часовые пояса
- и так далее

Алгоритм

- $a(x)$ – алгоритм, модель (функция, предсказывающая ответ для любого объекта)
- Отображает \mathbb{X} в \mathbb{Y} (пространство объектов в пространство решений)

Функционал качества

- После предыдущих шагов мы получили модель, которая выдает предсказания прибыли для ресторанов
- Теперь необходимо оценить качество полученных предсказаний
- Для этого используется метрики – меры качества работы алгоритма на выборке
- Подбирается исходя из бизнес-требований

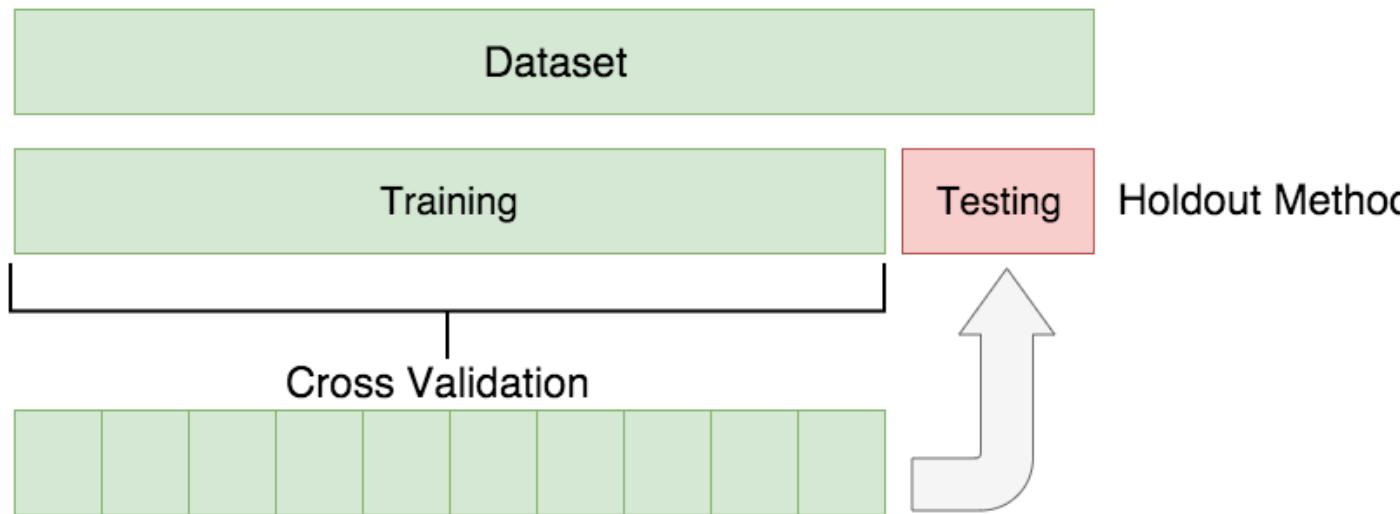
Пример метрики для задачи регрессии

Mean Squared Error (среднеквадратичное отклонение):

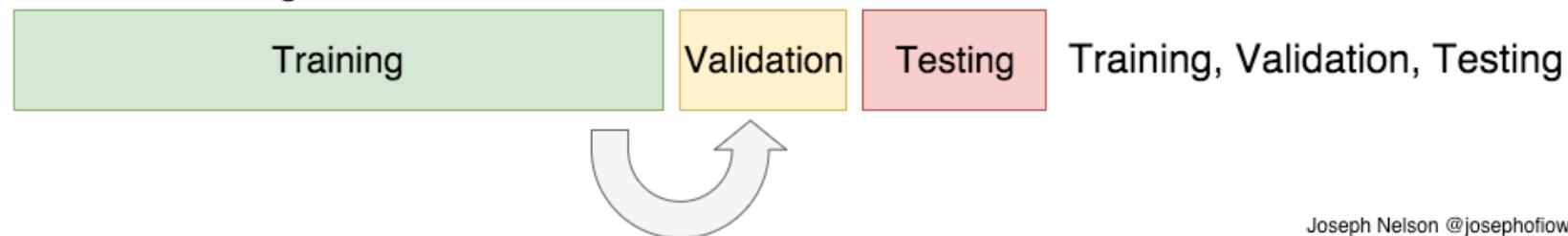
$$\text{MSE}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

- Чем меньше, тем лучше

Оценка результатов



Data Permitting:

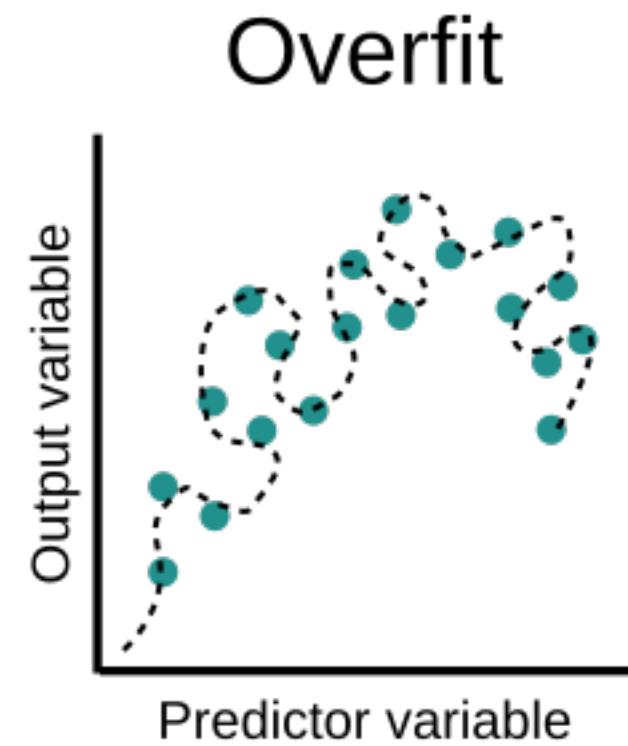
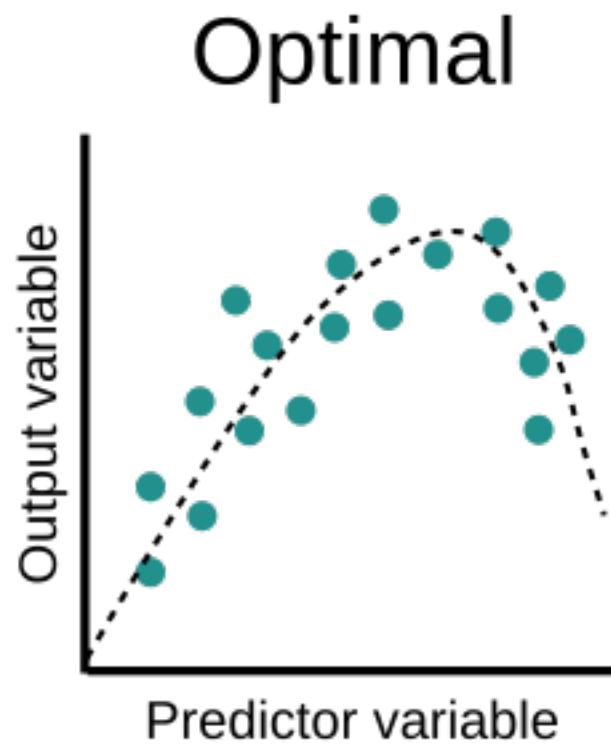
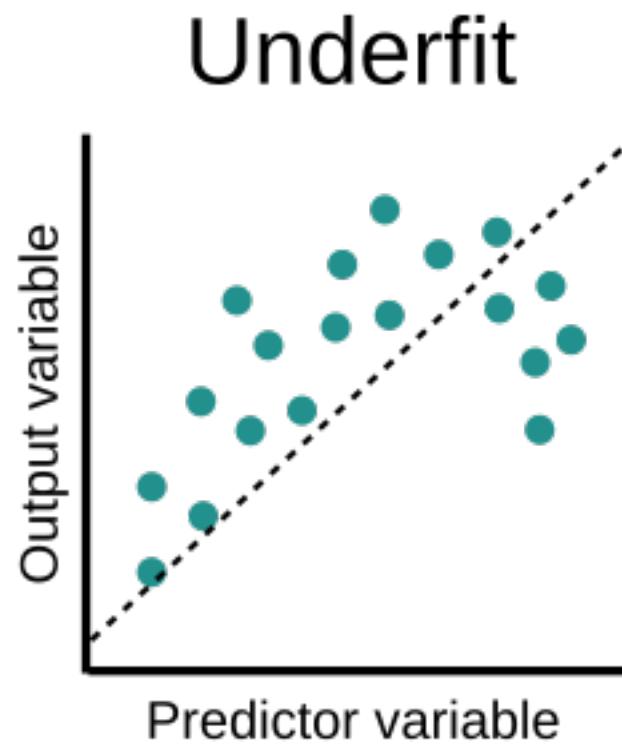


Joseph Nelson @josephofiowa

Обучение алгоритма

- Обучение – поиск оптимального алгоритма с точки зрения качества
- Имея обучающую выборку и функционал качества (метрику) подбираем подходящий алгоритм из некого семейства алгоритмов
- Важные критерии: качество, отвечающее бизнес-требованиям, устойчивость результатов, в некоторых случаях – интерпретируемость

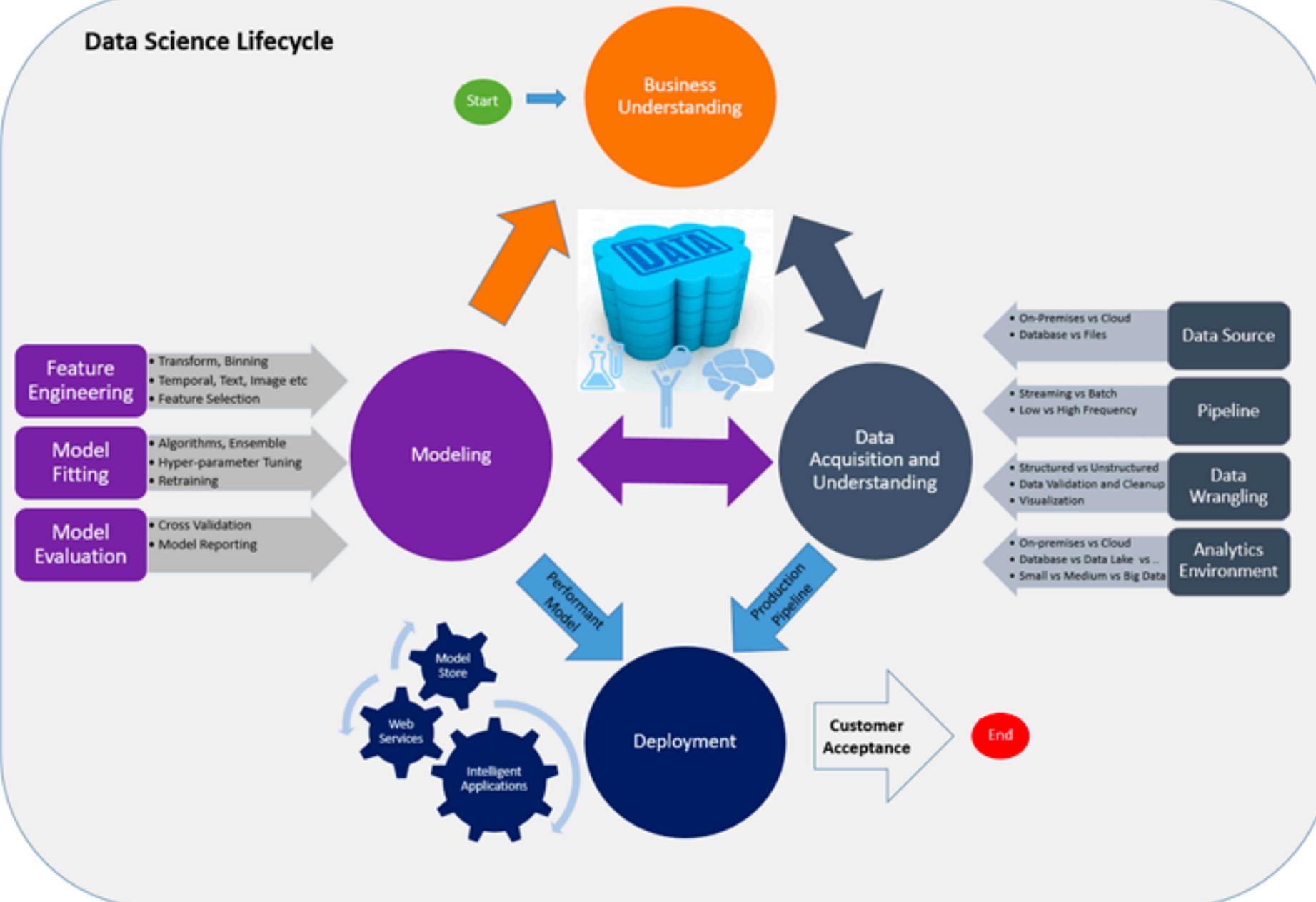
Переобучение и недообучение

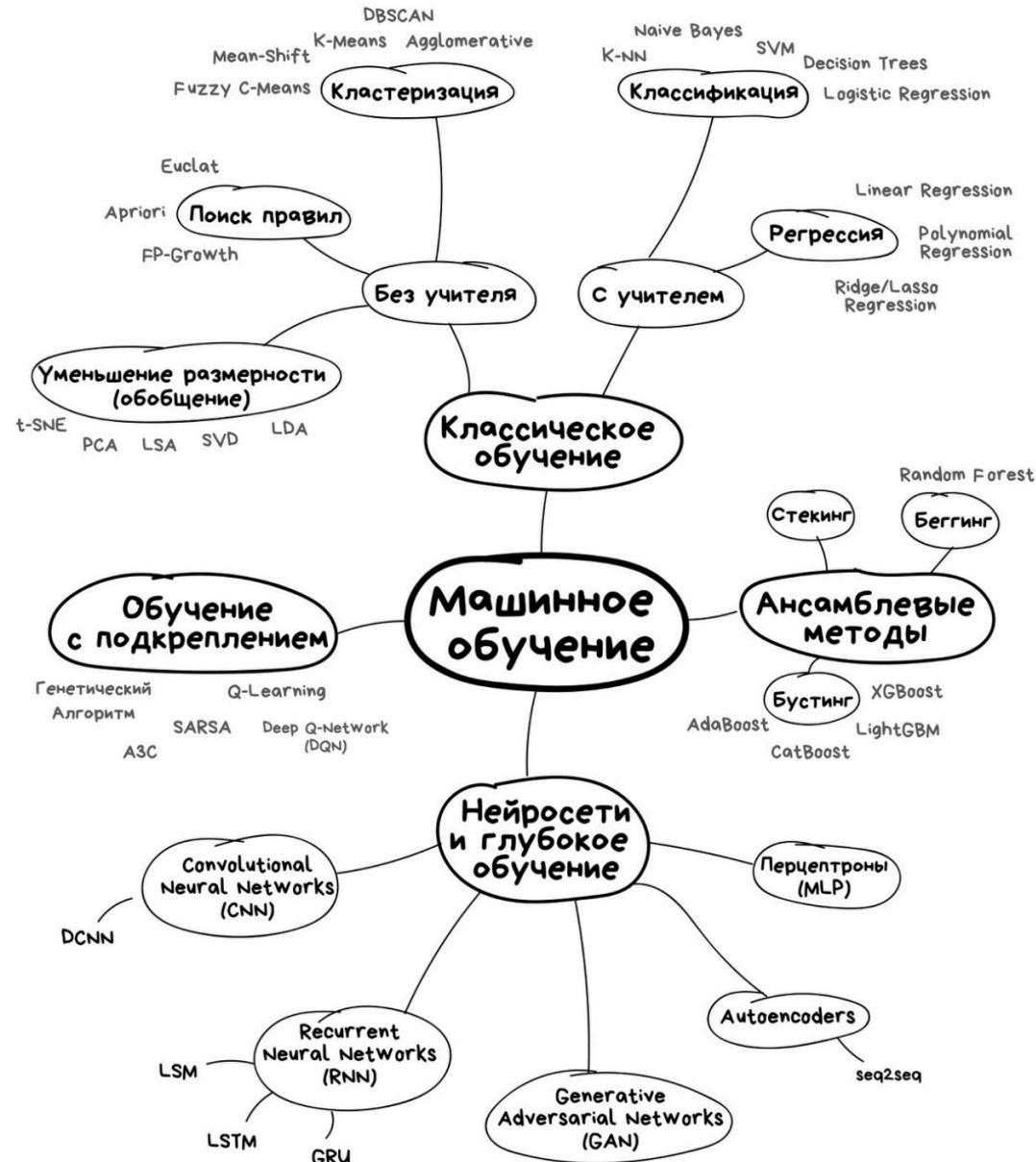


Цикл data science проекта

1. Определение задачи и целевой метрики
2. Сбор данных
3. Предобработка данных и EDA (exploratory data analysis, разведочный анализ данных)
4. Генерация дополнительных признаков
5. Подбор и валидация моделей
6. Как правило, возвращение к пункту 2 или 3

Data Science Lifecycle





Источник: [URL](#)

Задача регрессии

- **Регрессия** — класс задач обучения с учителем, когда по определённому набору признаков объекта нужно предсказать целевую переменную.
- **Задача регрессии** — нахождение зависимостей между определяющими переменными и определяемой переменной, если она является непрерывным числом. Например, определить стоимость дома по его площади.
- **Целевое значение** — любое действительное число.
- **Задача линейной регрессии** — нахождение такой зависимости, если она линейная.

Задача классификации

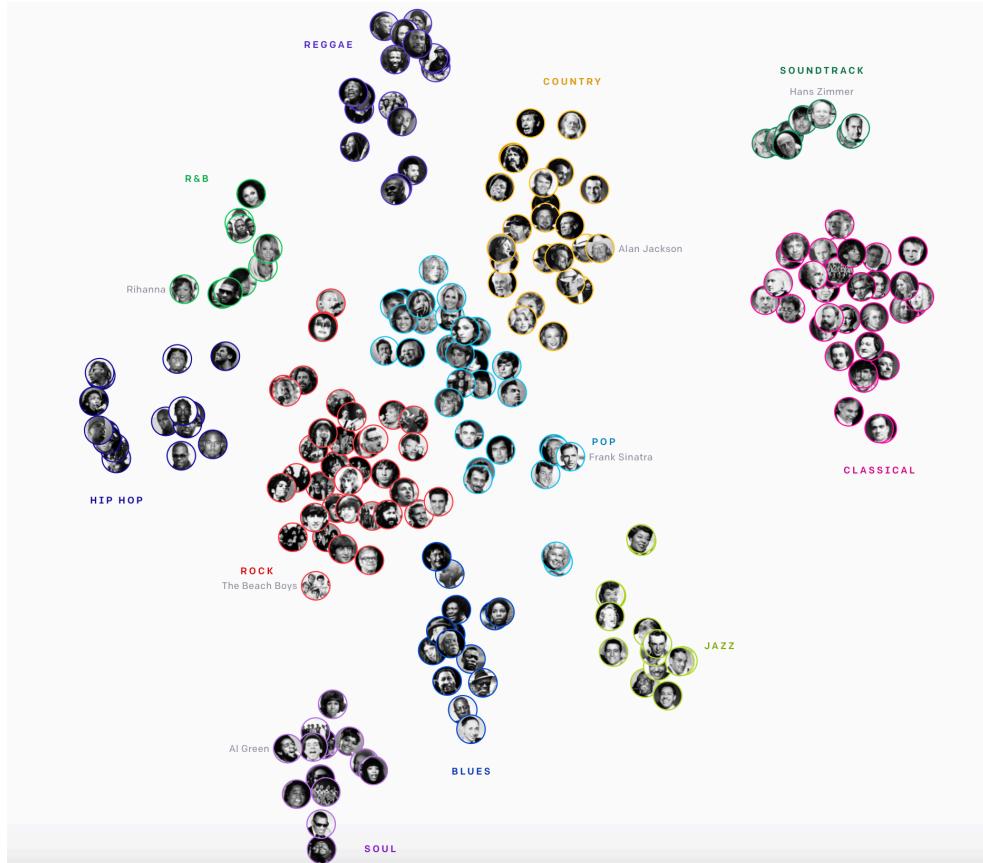
- **Классификация** — класс задач обучения с учителем, когда по определённому набору признаков объекта нужно предсказать его принадлежность к определенному классу.
- **Задача регрессии** — нахождение зависимостей между определяющими переменными и определяемой переменной, если она принадлежит к фиксированному набору значений. Например, определить вернет человек кредит или нет (бинарная классификация) или определение научной области статьи (многоклассовая классификация).
- **Целевое значение** – класс (один из фиксированного набора).
- Существует также многоклассовая классификация с пересекающимися классами.

Задача кластеризации

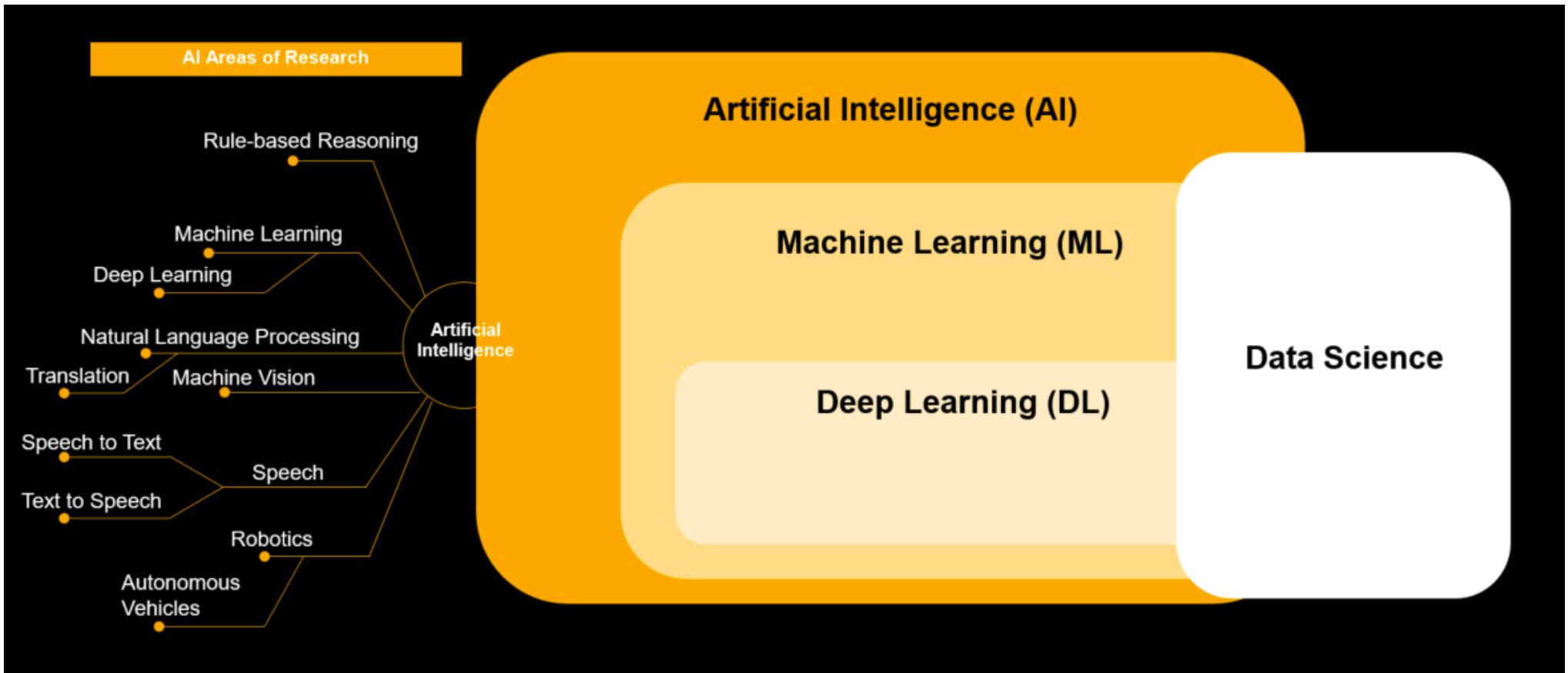
- Кластеризация — класс задач обучения без учителя, объекты разделяются на группы, обладающими некоторыми свойствами. Внутри одного кластера объекты должны быть похожи
- Похожесть определяется мерой расстояния
- Пример: евклидово расстояние – геометрическое расстояние в многомерном пространстве

$$\rho(x, x') = \sqrt{\sum_i^N (x_i - x'_i)^2}$$

Задача кластеризации



<https://openai.com/blog/jukebox/>

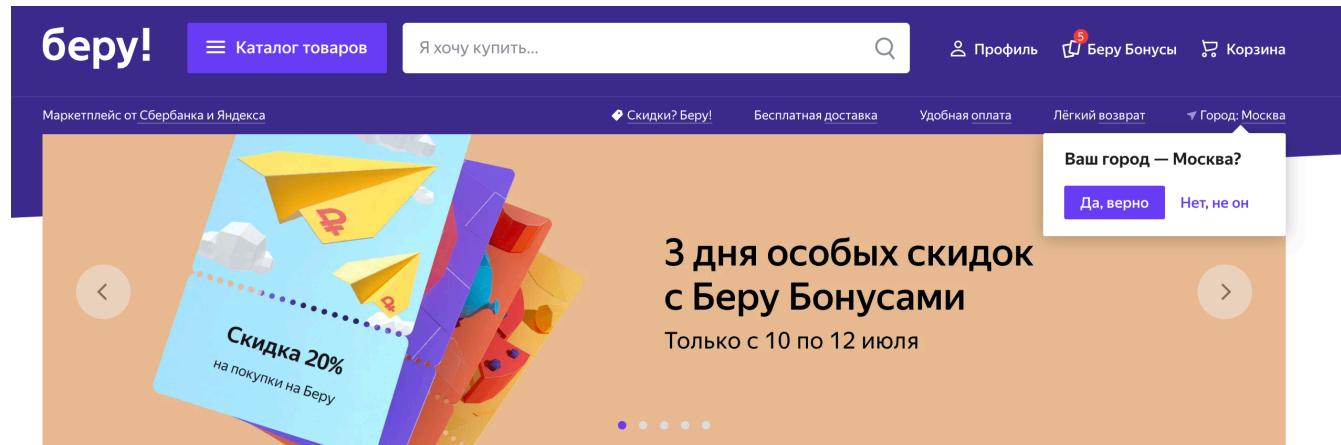


Направления работы

Рекомендательные системы

- Предсказываем, что будет интересно пользователю на основе информации о его профиле
- Примеры: Netflix, Кинопоиск, ЯндексМузыка, Youtube

Рекомендательные системы



Стоит приглядеться



4 038 ₽
от 146 ₽ / мес.



10 990 ₽
от 397 ₽ / мес.



1 200 ₽
4.5 ⭐⭐⭐⭐⭐ 1 548 отзывов



1 790 ₽
4.5 ⭐⭐⭐⭐⭐ 921 отзывов



10 690 ₽ ~~13 990 ₽~~
от 386 ₽ / мес.



889 ₽
4.5 ⭐⭐⭐⭐⭐ 35 отзывов

NLP – Natural Language Processing

➤ Работа с естественным языком

NLP + ранжирование

Google

how does netflix recommendation system work

Все Картинки Видео Новости Покупки Ещё Настройки Инструменты

Результатов: примерно 8 460 000 (0,47 сек.)



The recommendation system works putting together data collected from different places. ... Every time you press play and spend some time watching a TV show or a movie, Netflix is collecting data that informs the algorithm and refreshes it. The more you watch the more up to date the algorithm is. 2 авр. 2018 г.

uxplanet.org › netflix-binging-on-the-algorithm-a3a74... ▾

[Netflix: Binging on the Algorithm | by Josefina Blattmann | UX ...](#)

Подробнее о выделенных описаниях... Оставить отзыв

Похожие запросы

- How do I fix my Netflix recommendations?
- What information system does Netflix use?
- How does Netflix know what I want to watch?

Оставить отзыв

help.netflix.com › node ▾ Перевести эту страницу

[How Netflix's Recommendation System Works](#)

Яндекс

how does netflix recommendation system work

Найти

Поиск Картинки Видео Карты Маркет Новости Переводчик Эфир Коллекции Кью Услуги Ещё

Нашлось 19 млн результатов

N How Netflix's Recommendations System Works
help.netflix.com > en/node/100639 ▾

The recommendations system does not include demographic information (such as age or gender) as part of the decision making process. ... Below is a description of how the system works over time, and how these pieces of information influence what we present... Читать ещё >

Q How does the Netflix movie recommendation system work?
quora.com > How...Netflix...recommendation-system-work- ▾

More than 80 per cent of the TV shows people watch on Netflix are discovered through the platform's recommendation system. That means the majority of what you decide to watch on Netflix is the result of decisions made by a mysterious, black box of an algorithm. Intrigued? Here's how it works. Netflix uses machine learning and... Читать ещё >

M How Netflix's Recommendation Engine Works? - Medium
medium.com > ...ind/how-netfixs-recommendation...works... ▾

How does Netflix come up with such precise genres for its 100 million-plus subscriber ... Netflix's recommendation systems have been developed by hundreds of engineers ... Whenever a user accesses Netflix services, the recommendations system estimates... Читать ещё >

RT How does Netflix's recommendation system work?
radiotimes.com > news...netflix...how-does-it-work/ ▾

The streaming service is working hard to match viewers to new shows they'll like - but if you feel like Netflix doesn't "get" you, you're not alone.

W This is how Netflix's secret recommendation system works
wired.co.uk > article/how-do-netflixs...work-machine... ▾

This is how Netflix's top-secret recommendation system works. ... Netflix uses machine learning and algorithms to help break viewers' preconceived notions and find ... To do this, it looks at nuanced threads within the content, rather than relying on broad genres to... Читать ещё >

NLP

АНГЛИЙСКИЙ (ОПРЕДЕЛЕН АВТОМАТИЧЕСКИ)	РУССКИЙ	АНГЛИЙСКИЙ	РУССКИЙ	АНГЛИЙСКИЙ	УКРАИНСКИЙ
<p>The recommendation system works putting together data collected from different places. Recommended rows are tailored to your viewing habits. That's why you can tell when your little cousins have been using your account to watch a billion hours of Peppa Pig. In this case, algorithms are often used to facilitate machine learning. Systems like Netflix based on machine learning rewrite themselves as they learn from their own users. Every time you press play and spend some time watching a TV show or a movie, Netflix is collecting data that informs the algorithm and refreshes it. The more you watch the more up to date the algorithm is.</p>	<p>Система рекомендаций работает, собирая данные, собранные из разных мест. Рекомендуемые строки с учетом ваших привычек просмотра. Вот почему вы можете сказать, когда ваши маленькие двоюродные братья использовали вашу учетную запись, чтобы посмотреть миллиард часов Peppa Pig. В этом случае алгоритмы часто используются для облегчения машинного обучения. Такие системы, как Netflix, основанные на машинном обучении, переписывают себя, учась у своих пользователей. Каждый раз, когда вы нажимаете кнопку воспроизведения и проводите некоторое время за просмотром телепередачи или фильма, Netflix собирает данные, которые информируют алгоритм и обновляют его. Чем больше вы смотрите, тем более современным является алгоритм.</p>	<p>Sistema rekomendatsiy rabotayet, sobiraya dannyye, sobrannyye iz raznykh mest. Rekomenduyemye stroki s uchetom vashikh privychech prosmotra. Vot pochemu vy mozhete skazat', kogda vashi malen'kiye dvoyurodnyye brat'ya ispol'zovali vashu uchetnuyu</p>	<p>Развернуть</p>		

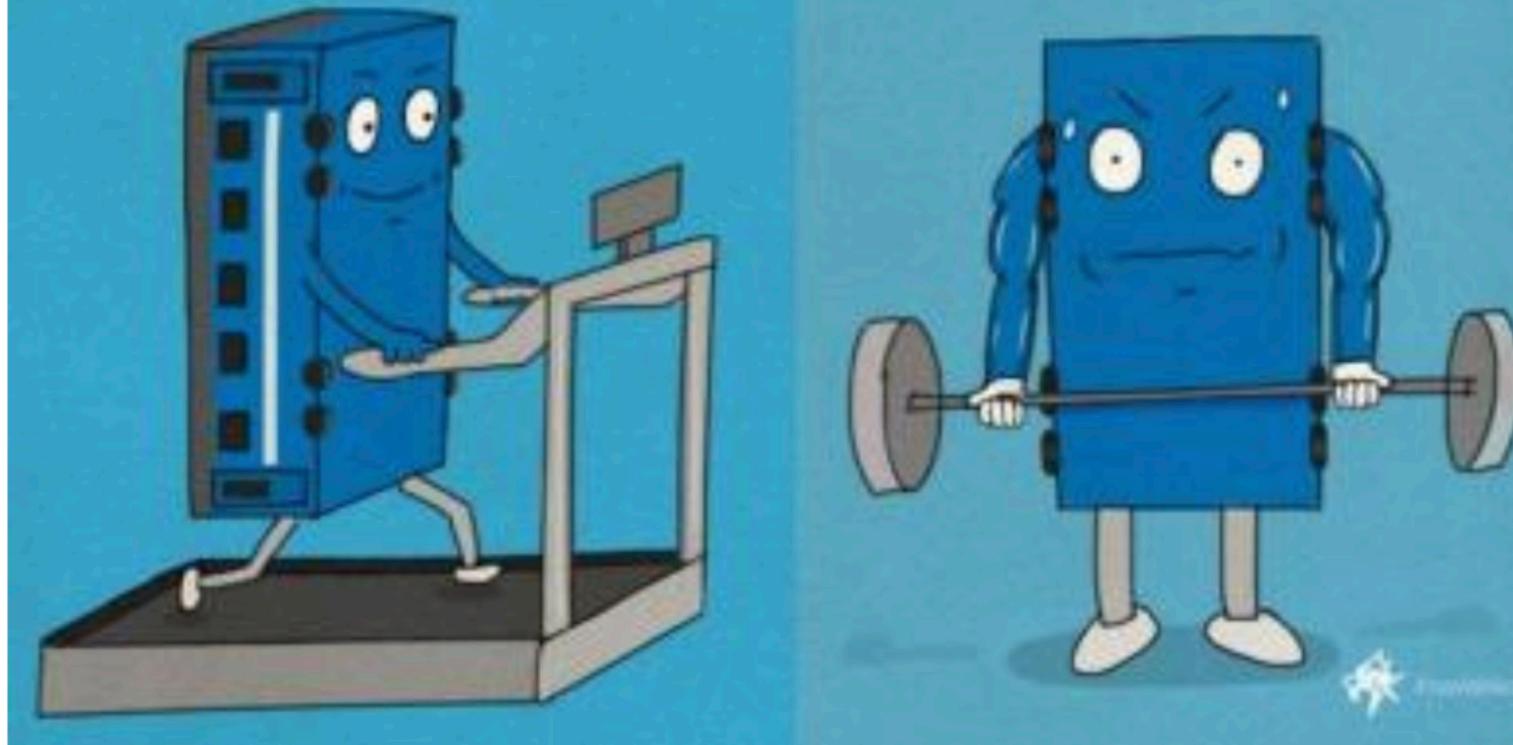
  637/5000 

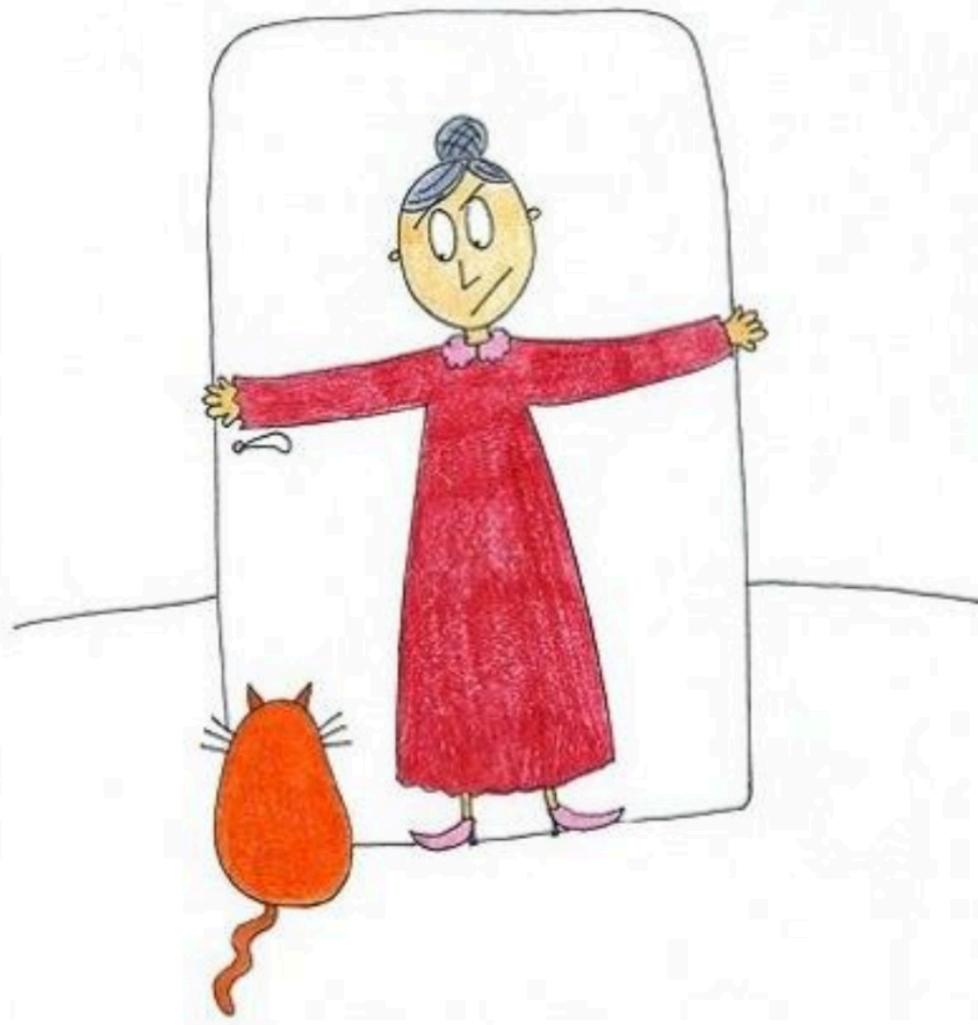
СТЕПАН ПЕТРОВИЧ
ПРОЦВЕТАЕТ



Голубой вагон
бежит — качается



МАЛЬВИНА ГЕНРИХОВНА ЗАЩИЩАЕТ ДОКТОРСКУЮ



Кластеризация текстов

Сейчас в СМИ в Москве Интересное 11.07, 11:56

-  В Хабаровске прошел несанкционированный митинг в поддержку Фургала
 -  WP: Трамп подтвердил кибератаку против российской компании в 2018 году
 -  В России предложили провести деноминацию рубля
 -  Эрдоган подписал указ о превращении собора Святой Софии в мечеть
 -  Голикова предложила возобновить международное авиасообщение с 15 июля
- USD 70,73 -0,19 EUR 79,88 -0,12 НЕФТЬ 43,32 +2,24% ...

Извлечение информации



Фрэнк Розенблatt



Психолог

Фрэнк Розенблatt — известный американский учёный в области психологии, нейрофизиологии и искусственного интеллекта.

[Википедия](#)

Родился: 11 июля 1928 г., Нью-Рошелл, Нью-Йорк, США

Умер: 11 июля 1971 г., Чесапикский залив, США

Известность: Перцептрон

Образование: Старшая школа (Бронкс) с углубленным изучением наук, Корнелльский университет

В запросе я написала имя, в ответ мне вылетела карточка с фотографией и краткой информацией

NLP – Natural Language Processing

- Работа с естественным языком
- Машинный перевод, чат боты, фильтр спама, подсказки в поисковой системе
- А в совокупности с speech recognition – Siri, Alexa, Алиса

CV – computer vision

- Хотим понять, что находится на изображении
- Тест Тьюринга для задач компьютерного зрения: ответить на любой вопрос про изображение, на который может ответить человек

Выделение объектов



- outdoor
- city
- Beijing, China
- Tiananmen Square

Классификация и поиск похожих

Sky is blue

Wind is slow

Male, Mao Zedong

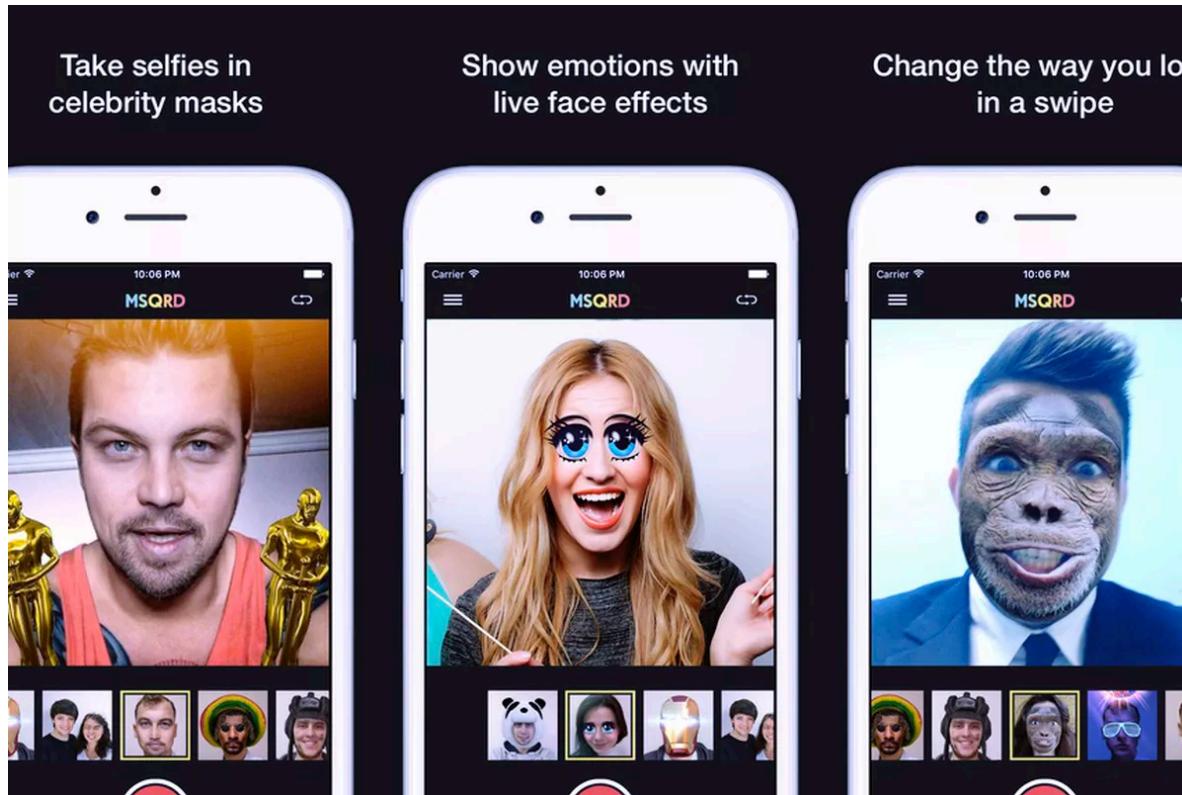


Profile, Female,
Unknown

Frontal, Male,
Unknown



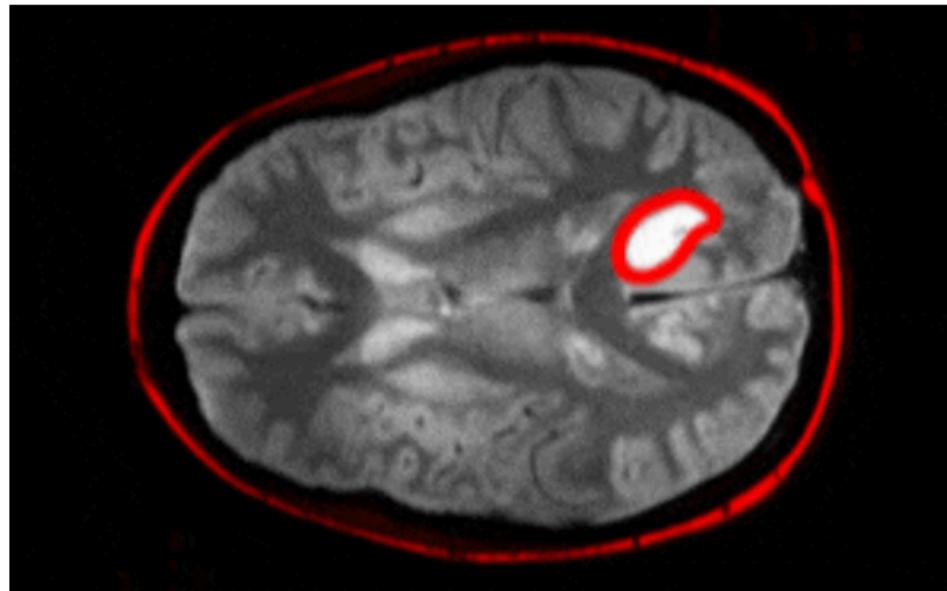
Где используем?



Где используем?



Где используем?



Детекция опухоли головного мозга