Введение в машинное обучение

Сбертех, МФТИ

Зачем нужно машинное обучение?

Машинное обучение – это область обучения, которая дает компьютеру возможность учиться без явного программирования.

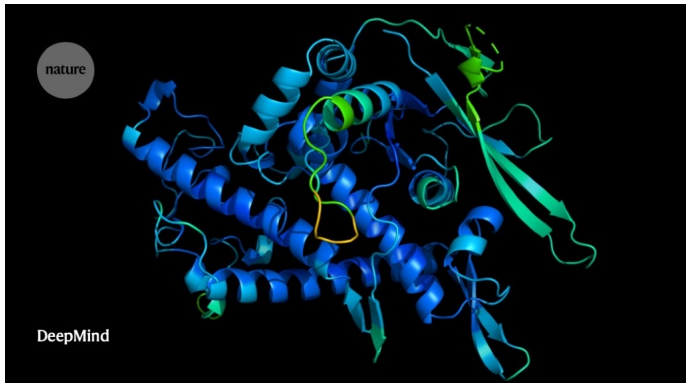
Обучаемая машина - это любое устройство, на действия которого влияет прошлый опыт.

AI vs ML

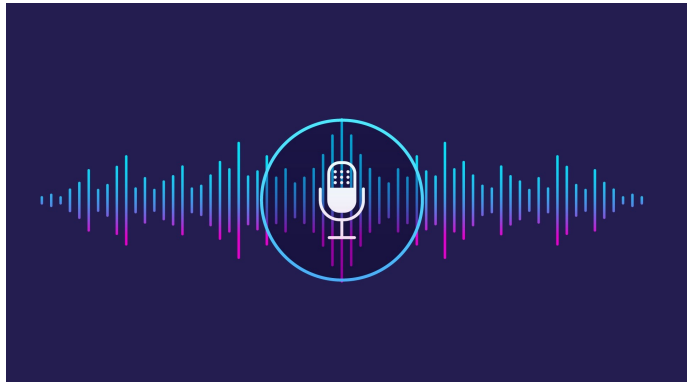

Искусственный Интеллект
Машинное Обучение
Нейросети
Глубокое Обучение
сотни других методов обучения

Почему ИИ это круто?

Движок рекомендаций Netflix предотвращает уход клиентов предлагая им новые серии



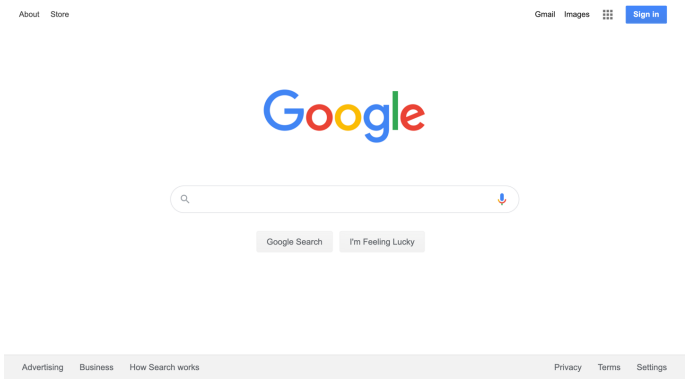
OpenAI создала модель предсказывающую структуру белка для ускорения исследования вирусов и создания лекарств



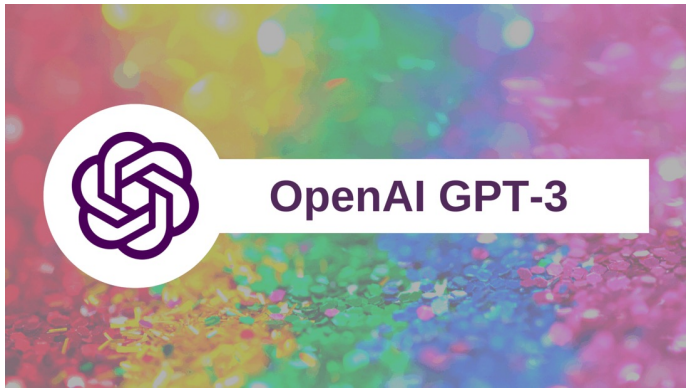
Голосовые ассистенты решают повседневные задачи пользователей без вмешательства человека

Yandex создает самоуправляемые машины

Почему ИИ это круто?



Поиск Google помогает найти любую информацию за считанные секунды



Система GPT-3 позволяет генерировать осознанный текст, код программ и запросов

Система Dall-E/MidJourney генерирует по текстовому запросу наиболее подходящую картинку

Типы задач машинного обучения

Примеры задач машинного обучения

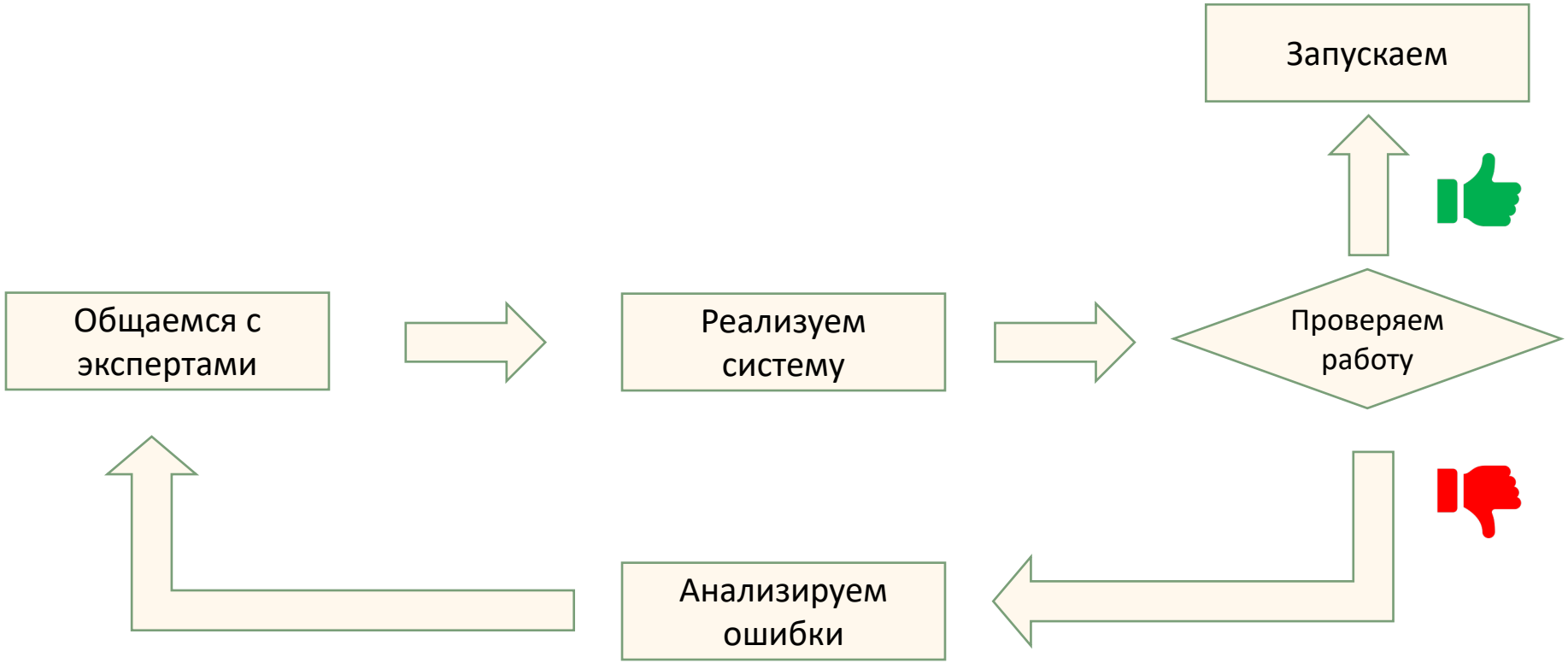
Тип задачи	Описание	Пример
Классификация	Определи 1 из классов	Собака, кот или вертолет на картинке
Регрессия	Предсказание действительного числа	Цена дома
Кластеризация	Группировка похожих примеров	Определение похожих документов
Правила ассоциации	Задача рекомендаций	Если пользователь купил матрас, купит ли он насос
Уменьшение размерности	Избавление от лишней информации объекта	Представление объекта не 10 х-ми , а 3
Структурированный вывод	Сложный вывод	Предсказание синтаксического дерева предложения, определение границ объекта на картинке
Ранжирование	Определение позиции объекта в выдаче	Google/Yandex

Rule-based подход

Рассмотрим задачу постановки мед. Диагноза

- Сначала общаемся с экспертами
- Получаем понимание установки диагноза
- Реализуем алгоритм предсказания

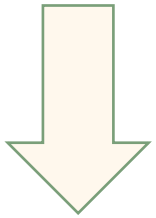
Процесс rule-based подхода может быть усложнен логическим выводом и использованием баз знаний.



Плюсы и минусы rule-based подхода

Плюсы

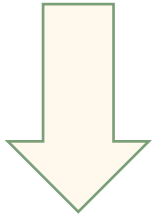
1. Можно быстро сделать хороший бейслайн решения
2. Не надо тратить время на сбор обучающей выборки и обучения
3. Хорошо справляются с ограниченными доменами



Нужен MVP – rule-based

Минусы

1. Большие системы сложно строить
2. Плохо справляются с неоднозначностями
3. Монолитны и плохо масштабируемы

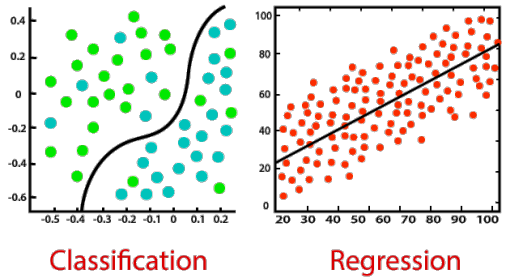


Нужно укрепить хороший бейслайн – модели МО

Современное машинное обучение

Рассмотрим задачу постановки медицинского диагноза:

1. Изучаем задачу, как ставить диагноз, какие данные нужны, разбираемся с целевой переменной
    1. Классификация – болен пациент или нет
    2. Регрессия – процент поражения легких
2. Собираем данные для обучения модели предсказания медицинского диагноза
3. Обучаем модель, которая сама устанавливает нужные зависимости в данных и учится делать правильный вывод



Ключевые понятия

- Типы задач машинного обучения
  - **Обучения с учителем** – есть целевая переменная Y
  - **Обучение без учителя** – нет целевой переменной Y
- **Модель предсказания** – предсказывает целевую переменную при входе данных X
- **Обучающая выборка** – набор пар (X, y)
- **Обучение с учителем** – модель предсказания учится на обучающей выборке

Процесс обучения моделей

- Понимание домена
- Оценка эффекта
- Оценка качества

- Понимание домена
- Hard skills – SQL, Python, DF Systems
- Crowd-source

- Понимание домена
- Статистические меры и свойства
- Линейная алгебра

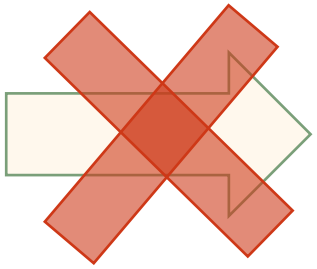
- Методы оценки параметров
- Оптимизация функций
- Алгоритмы машинного обучения

- Статистические методы оценки
- Статистика
- Оптимизация

Мышление в машинном обучении

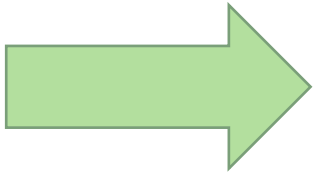
Для решения задачи машинного обучения в первую очередь надо знать данные задачи.

Требования



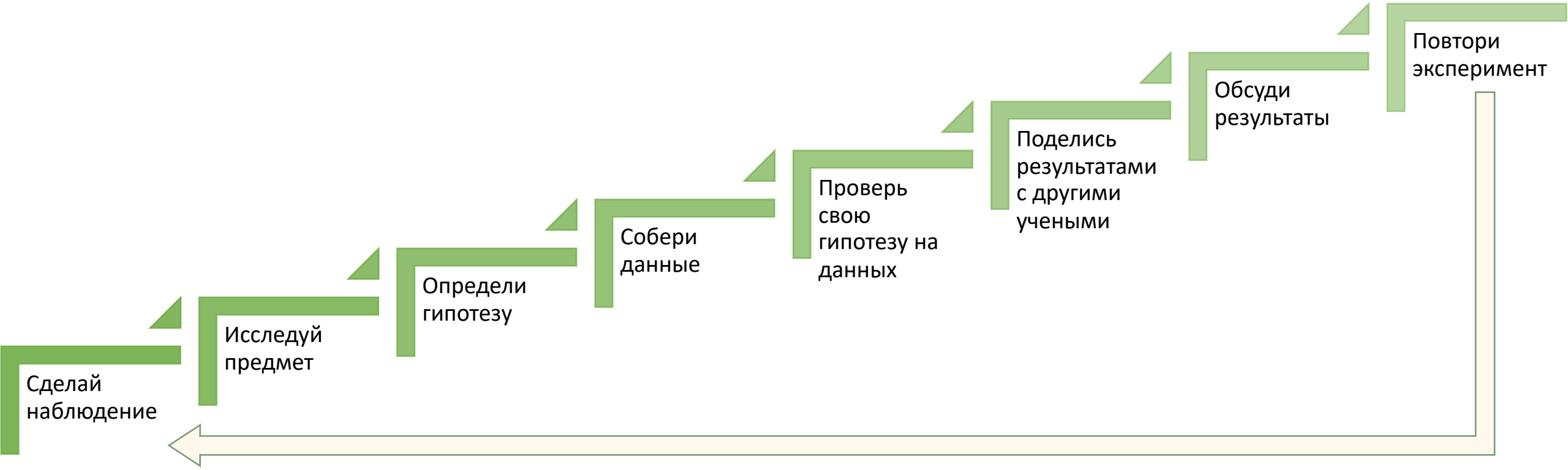
Продукт

Данные



Продукт

Научный метод



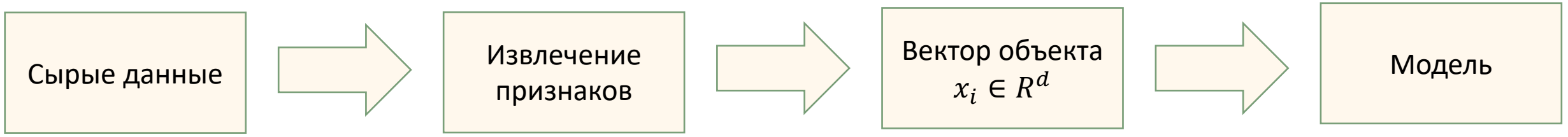
Научный метод в машинном обучении

Данные в машинном обучении

Модели обучаются на числовых признаках – потому что ничего другое компьютер не может понять.

Что может использоваться для машинного обучения?

- Текстовые документы
- Картинки
- Видео
- Последовательности ДНК
- Звук
- **Все, откуда можно извлечь признаки!**



- Чем лучше признаки – тем качественнее можно обучить модель
- Иногда сам разработчик генерирует признаки, иногда это делает модель
- Можно обучать модель генерировать хорошие признаки

Пример извлечения признаков

Задача – предсказать, является ли данная последовательность букв, электронным адресом

example@yandex.ru

Название признака	Значение признака
Присутствует @	1
Заканчивается на ru	1
Количество букв до @	7
Домен yandex	1
Заканчивается на com	0
Заканчивается на ru	1
…	…

exampleyandexru

Название признака	Значение признака
Присутствует @	0
Заканчивается на ru	1
Количество букв до @	0
Домен yandex	1
Заканчивается на com	0
Заканчивается на ru	1
…	…

Размеченные данные

Признак 1 (категориальный)	Признак 2 (вещественный)	Признак 3 (целочисленный)	Признак 4 (порядковый)	Целевая переменная
0	0.72	12000	1	1
1	0.3	32111	2	0
1	0.54	43442	3	1
0	0.99	9800	1	1

Объект обучающей выборки

Пространство признаков

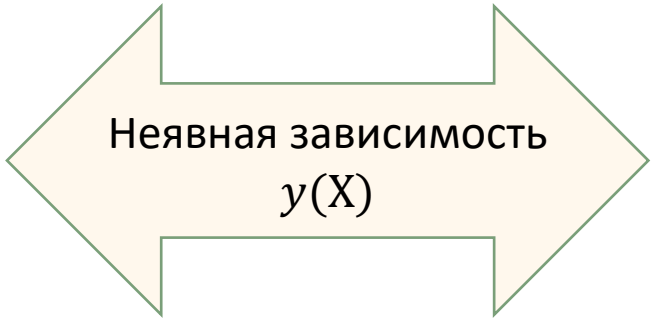
Целевое значение

Обучающая выборка

Задача машинного обучения

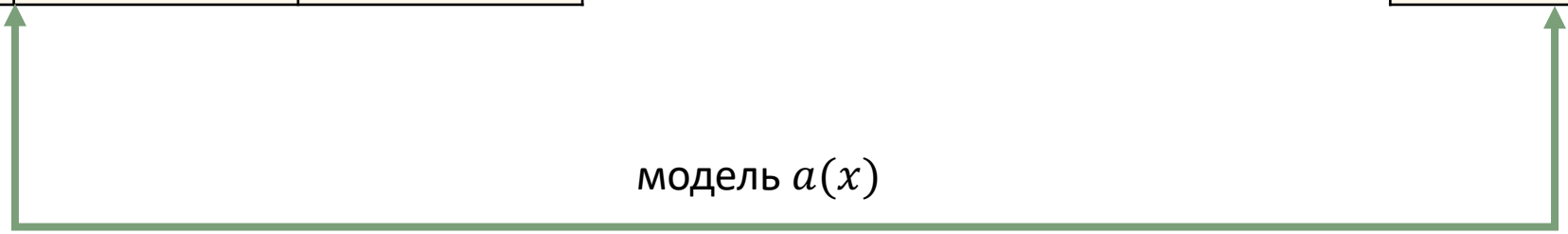
X

Признак 1	Признак 2	Признак 3	Признак 4
0	0.72	12000	1
1	0.3	32111	2
1	0.54	43442	3
0	0.99	9800	1



y

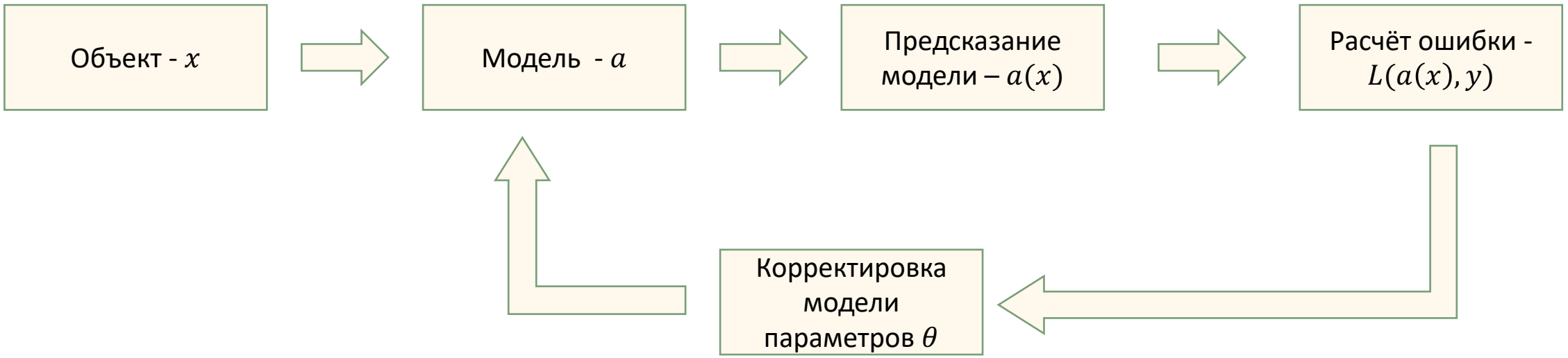
Целевая переменная
1
0
1
1



Модель a должна наилучшим образом приближена к y

Сведение задачи обучения к задаче оптимизации

- Модель – параметрическое семейство функций $A = \{g(x,\theta)|\theta \in \Theta\}$, где $g: X \times \theta \to Y$- функция, Θ – множество допустимых параметров θ
- $L(a,x)$ – величина ошибки модели a на объекте x
  - Функция ошибки на задаче классификации – $L(a,x) = [a(x) \neq y(x)]$
  - Функция ошибки на задаче регрессии – $L(a,x) = \big(a(x) - y(x)\big)^2$ - ***mean squared error(MSE)***
- Эмпирический риск – значение функции $L(a,x)$ на X – $Q(a_\theta, X) = \frac{1}{n}\sum_{i=1}^{n} L(a_\theta, x_i)$
- Задача оптимизации – $\underset{a \in A}{\operatorname{argmin}}\, Q(a,X)$



Оценка данных

Данные задачи

1) ***Обучающая выборка(train)*** – часть данных, на которых происходит обучение модели машинного обучения.
   - Чем больше обучающая выборка, тем лучше обучится алгоритм.
   - Распределение данных(как признаков, так и целевых) в обучающей выборке должно быть такое же, какое вы ожидаете увидеть при деплое модели
   - Обычно для обучающей выборки берут 60-80%.

2) ***Валидационная выборка(val)*** – часть данных, которая используется для тестирования во время обучения для подбора гиперпараметров (*внешних эффектов, которые влияют на обучение модели*) модели.
   - После подбора оптимальных параметров обычно добавляется в трейн(так как чем больше данных, тем лучше качество)

3) ***Тестовая выборка(test)*** – часть данных, на которых проходит тестирование итоговой модели.
   - Тестовая выборка не используется ни в train, ни в val. Она используется, как отложенный контроль для проверки модели.
   - Если данных мало, то тестовую выборку стоит сделать небольшой, но репрезентативной.

Проблемы с данными

- **Разные данные/разное распределение данных** в тренировочной и тестовой выборке

- **Утечка данных** – информация в данных прямым образом влияет на целевую переменную.
  - Название картинки совпадает с классом картинки
  - В тренировочной выборке в задаче рекомендации фильмов используются фильмы которые пользователь посмотрел после определенной даты(в тестовой выборке)
  - Использование рейтинга отзыва при классификации тональности

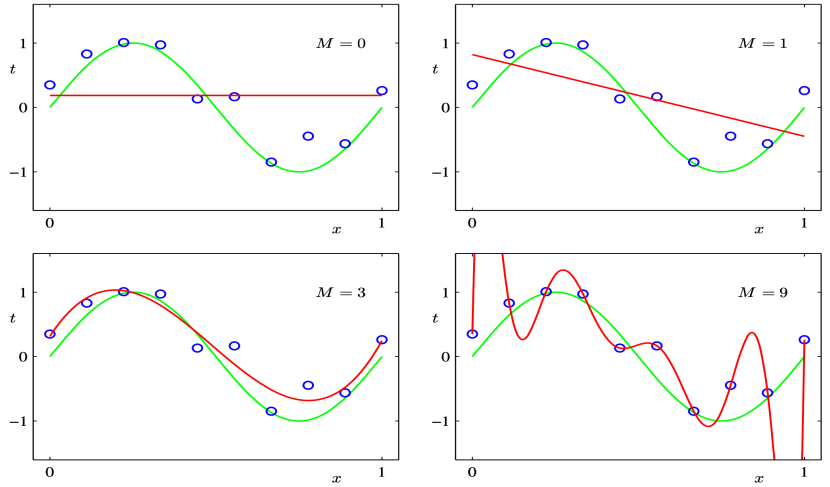
- **Нестационарность в данных** – данные для моделирования меняются со временем
  - Ковариационный сдвиг – изменения распределения данных на этапе обучения и тестирования
    - Популярность запросов в поисковике – старые запросы перестают быть популярными, новые становятся популярными
  - Концептуальный сдвиг – изменения истинного значения при одних и тех же входных данных во времени
    - *Летом люди обычно не покупают зимние куртки*
    - *На прошлой неделе я искал квартиру, на этой не ищу*

Недообучение, переобучение и сложность

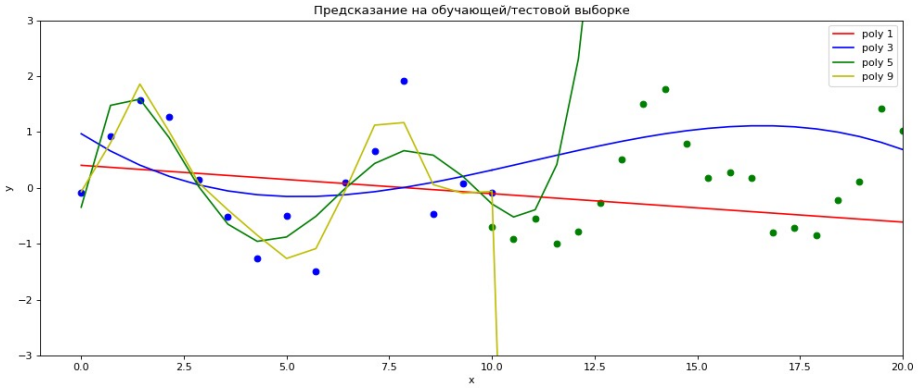
- **Переобучение (overfitting)** – явление, когда ошибка на тестовой выборке заметно выше чем на обучающей.
  - Обычно происходит из-за небольшого количества данных

- **Недообучение (underfitting)** – явление, когда ошибка на обучающей выборке слишком большая.
  - В таком случае говорят – модель не смогла настроится на обучающую выборку(например, модель очень простая для моделирования скрытой зависимости в данных)

- **Сложность модели** – оценка, насколько разнообразно семейство алгоритмов в модели с точки их функциональных свойств(например, способности настраиваться на выборку).
  - Высокая сложность модели решает проблему недообучения и вызывает переобучение.



Переобучение на тренировочной выборке



Работа модели на тренировочной и тестовой выборках

Из чего состоит ошибка модели?

Разложим функцию ошибки $(a - y)^2$ на компоненты

Целевая(скрытая зависимость) – $y = f(x) = \sin(x) + \epsilon, \epsilon \sim N(0, \sigma^2)$

1. Посмотрим на мат. ожидание квадрата отклонения предсказания от истинного значения:

$$E(a-y)^2 = E(a^2 - 2ay + y^2)$$

2. Воспользуемся свойством линейности E, выделим полные квадраты в мат. ожидании

$$E(a^2 - 2ay + y^2) = E(a^2) - \big(E(a)\big)^2 + \big(E(a)\big)^2 - 2E(ay) + E(y^2) - \big(E(y)\big)^2 + \big(E(y)\big)^2$$

3. По формуле дисперсии $\boldsymbol{D(x) = E(x^2) - \big(E(x)\big)^2}$

$$E(a^2) - \big(E(a)\big)^2 + \big(E(a)\big)^2 - 2E(ay) + E(y^2) - \big(E(y)\big)^2 + \big(E(y)\big)^2 = D(a) + \big(E(a)\big)^2 - 2E(ay) + D(y) + \big(E(y)\big)^2$$

4. Обратное преобразование

$$D(a) + \big(E(a)\big)^2 - 2E(ay) + D(y) + \big(E(y)\big)^2 = D(a) + D(y) + \big(E(a-y)\big)^2 = variance(a) + \sigma^2 + bias^2(a, y)$$

Bias & Variance

$$variance(a) + \sigma^2 + bias^2(a, y)$$

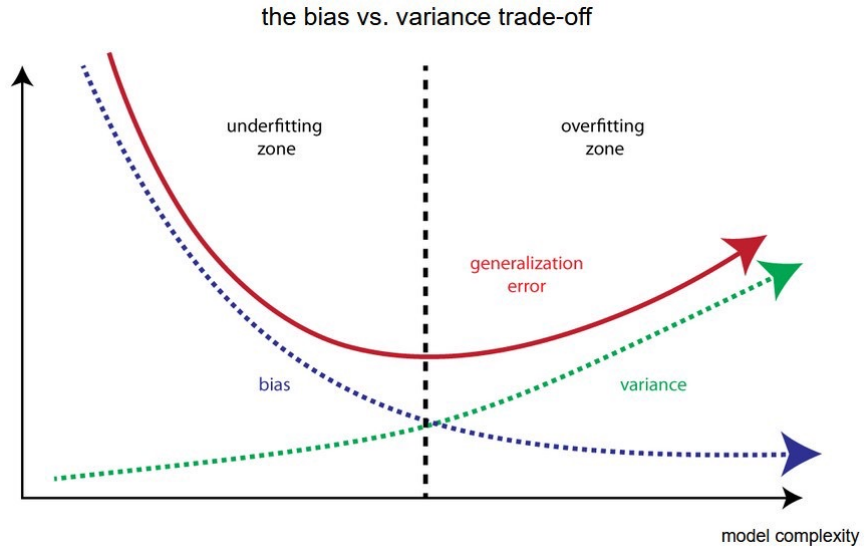
- Смещение(**bias**) – разность между предсказанием и истинным ответом - $(a - y)^2$
- Разброс(**variance**) – чувствительность алгоритма a к изменениям в обучающей выборке, дисперсия предсказания относительно корректного значения

- Когда модель сложная(**переобучилась – high variance, low bias**) – у нас высокий разброс предсказаний
- Когда модель слишком простая(**недообучилась, low variance, high bias**) – у нас высокое смещение относительно целевой переменной

- Ошибка предсказания состоит из 3 компонент:
  - Неустранимая ошибка в данных(шум)
  - Дисперсия модели предсказания
  - Разница предсказания и истинного ответа

- **Как найти оптимальную модель для наших данных?**

Bias and Variance trade-off



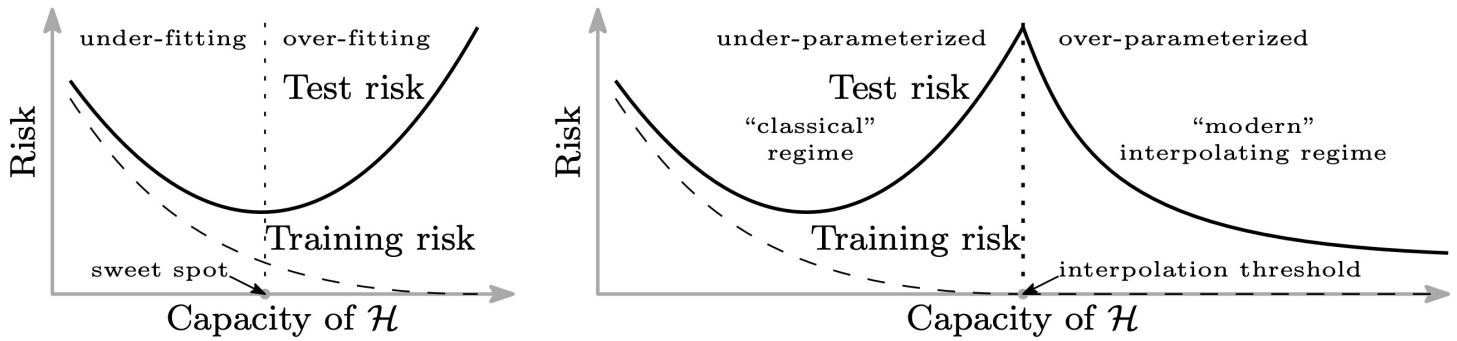
- Чем сложнее модель, тем ближе ее предсказание к тренировочной выборке.
  - Чем ближе модель к тренировочной выборке, тем меньше смещение.

- Чем сложнее модель, тем больше она обучается под тренировочную выборку
- Чем больше модель обучается под тренировочную выборку, тем больше разброс на тестовой выборке.

- Низкое смещение на тренировочной выборке и высокий разброс на тестовой – переобучение.
- Высокое смещение на тренировочной выборке и низкий разброс на тестовой – недообучение.

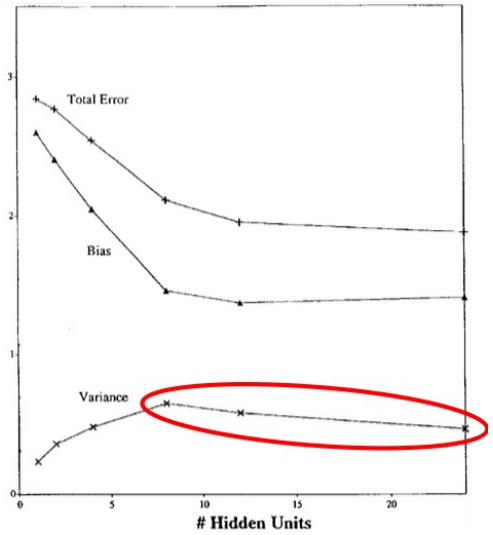
- Существует оптимальная точка выбора смещения и разброса.

Double descent



С увеличение ширины первого слоя, у модели начинает снижаться разброс.

https://arxiv.org/pdf/1812.11118.pdf



Geman et al. on Bias & Variance dilemma

https://arxiv.org/pdf/1810.08591.pdf

- Bias-variance trade off не всегда выполняется, существуют случаи(модели/данные) когда этот подход не работает
- В общем, не всегда существует trade-off между смещением и разбросом – стоит всегда пытаться его искать при обучении моделей(вместо этого можно смотреть на другие метрики)

Кейс. Обучение с учителем.

Определение оттока клиентов. Описание задачи.

- **Постановка и описание задачи**
  - Телеком оператору нужно своевременно определить клиентов, которые хотят уйти от нее/переключиться на другого оператора.
  - Обычно привлечение новых клиентов дороже, чем удержание старых(маркетинговые компании достаточно дорогие)
  - Компания может заблаговременно удержать клиента, дав ему скидку или предложив новые условия.
  - Необходимо предсказать отток клиента, причем:
    - Если клиенту предложили скидку и он собирался уйти – компания сохранила деньги
    - Если клиенту предложили скидку, а он не собирался уйти – компания потеряла деньги

- **Данные**
  - Информация о всех клиентах за два года
  - Информация об времени уходе клиента

- **Бизнес задача – научится удерживать клиентов**

Давайте разберемся

1. Что мы будем предсказывать?
2. Какие данные мы будем использовать на этапе обучения / оценки ?
3. Какие метрики нам нужны для оценки?

Кейс. Обучение без учителя.

Определение места склада для быстрой доставки

Компания **Доставка Джо** по быстрой доставке продуктов хорошо развивается в Москве и теперь готова выходить в новый город Воронеж. Работа **Доставка Джо** простая – за 30 минут доставить со склада продукты пользователю. Ваша задача определить – где в самом начале должны стоять в новом городе склады с продуктами.

На какие вопросы мы должны ответить перед исследованием?

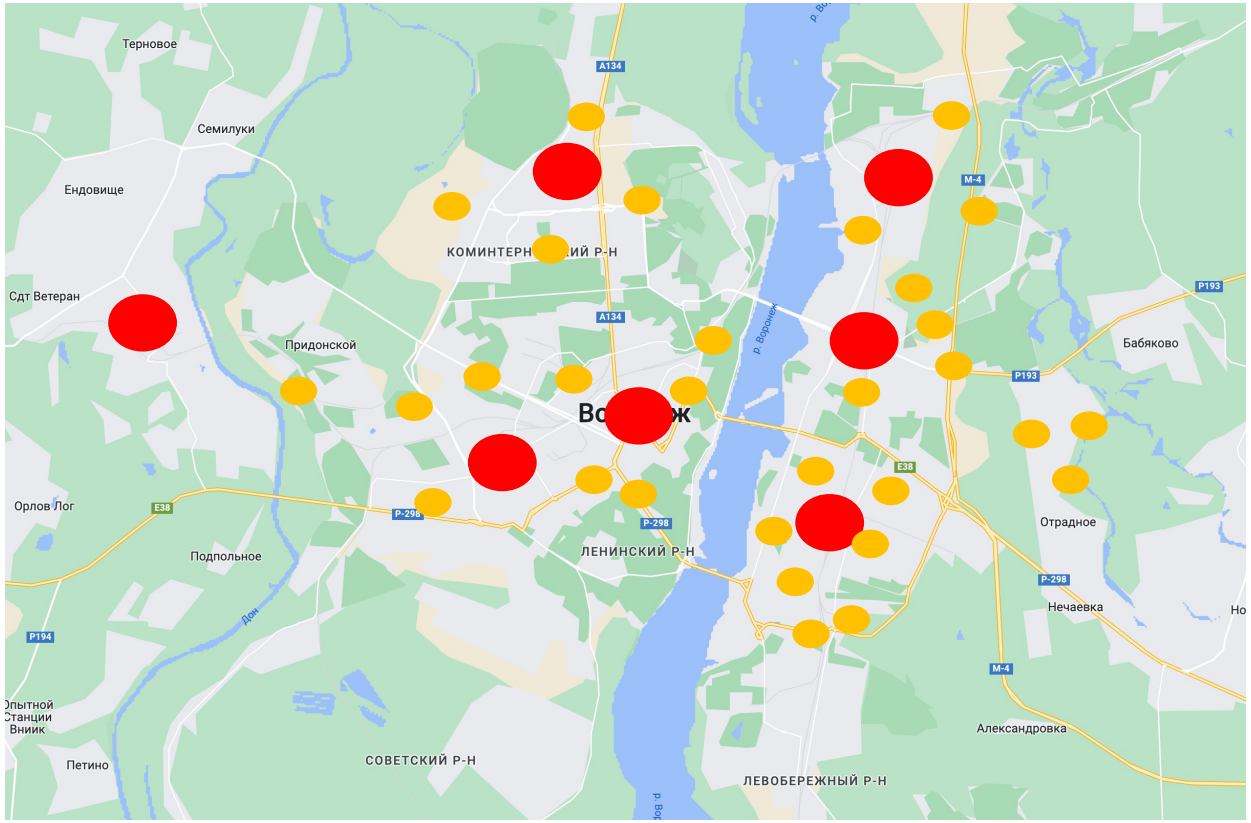
Определение места склада для быстрой доставки

- **Количество складов для начала**
- **Распределение свободных помещений в городе**
- **Первые области для старта**
- **Цены на аренду**
- **Минимальное количество человек на склад**
- **Что делают конкуренты?**
- **Предпочтения по количеству/качеству у стейкхолдеров**

Определение места склада для быстрой доставки

Как нам оптимизировать расположение складов?

Откуда получить ответ на вопрос?



Определение места склада для быстрой доставки

Информация о точках доставки

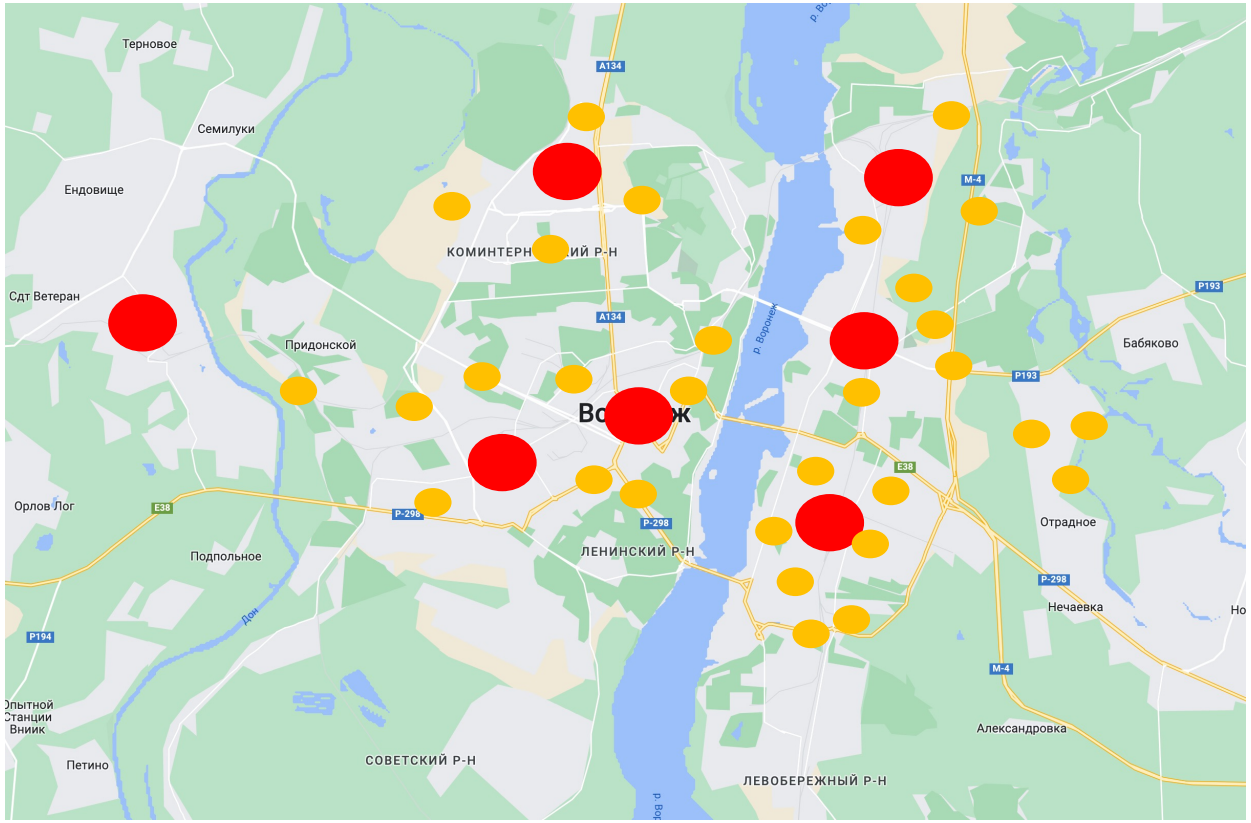
$$\sum_{i=1}^{S} \sum_{u \in U_i} |u - \mu_i| \rightarrow min$$

u – координаты пользователя

μ_i - координаты центра склада

Ищем оптимальные центры среди наших пользователей.

пользователи склады



Определение места склада для быстрой доставки

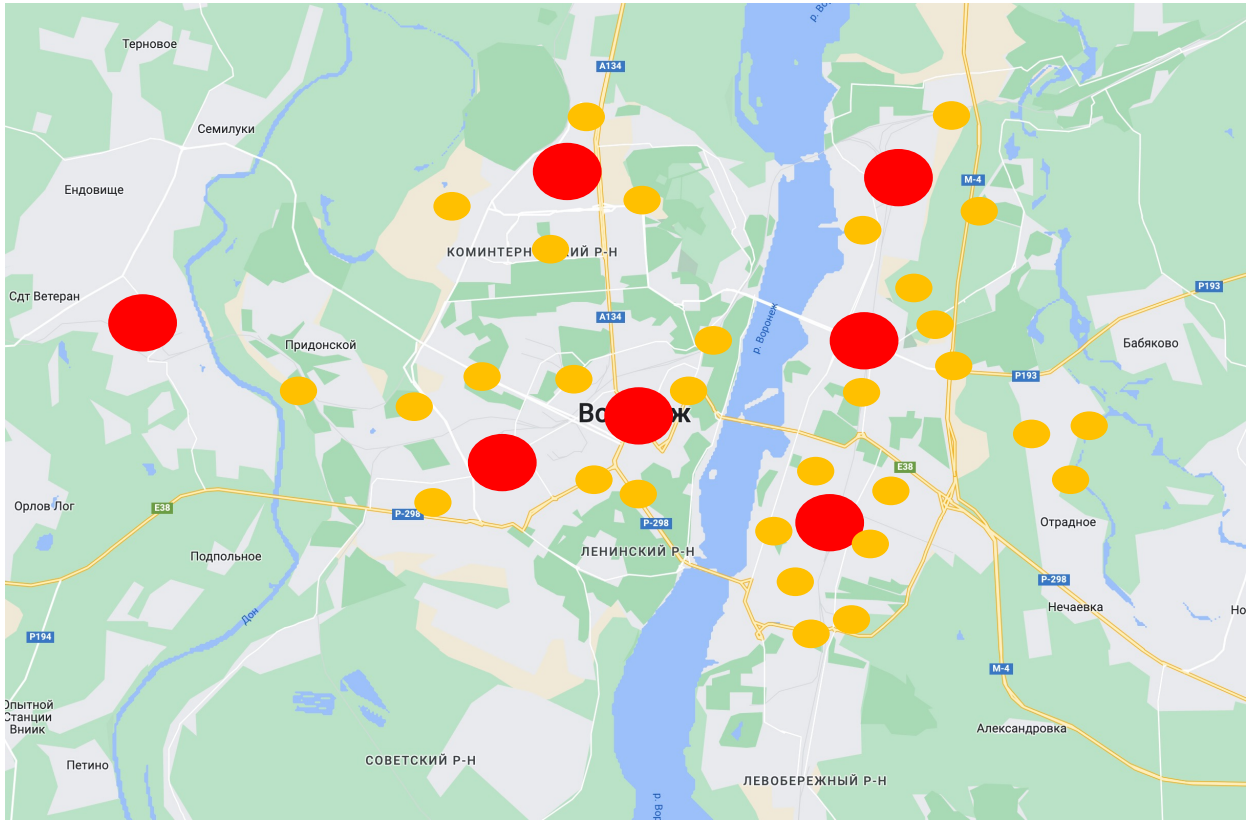
Перерасчёт позиций склада дал результаты – теперь склады в новой позиции уменьшают время доставки. Но возможно ли сделать еще лучше?

Информация о точках доставки

$$\sum_{i=1}^{S} \sum_{u \in U_i} |u - \mu_i| \rightarrow min$$

u – координаты пользователя

μ_i - координаты центра склада



Определение места склада для быстрой доставки

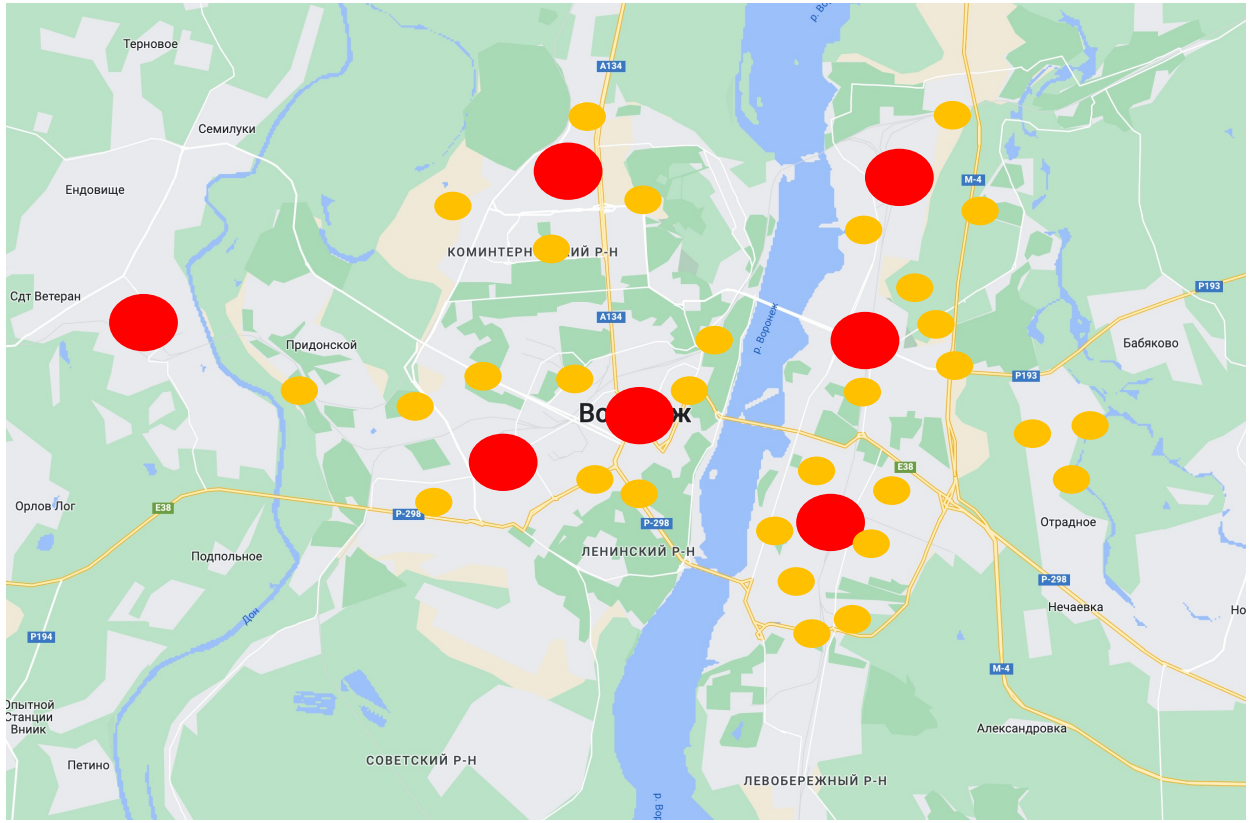
$$\sum_{i=1}^{S} \sum_{u \in U_i} |u - \mu_i| \rightarrow min$$

Но у нас не Евклидово расстояние – у нас есть пробки, ремонт дорог, мост.

u – координаты пользователя

μ_i - координаты центра склада

Матрица истинного расстояния доставки от склада пользователю



Поиск наиболее критичного склада и замена на новый