НЕТОЛОГИЯ
групп

ЗАНЯТИЕ 4.4

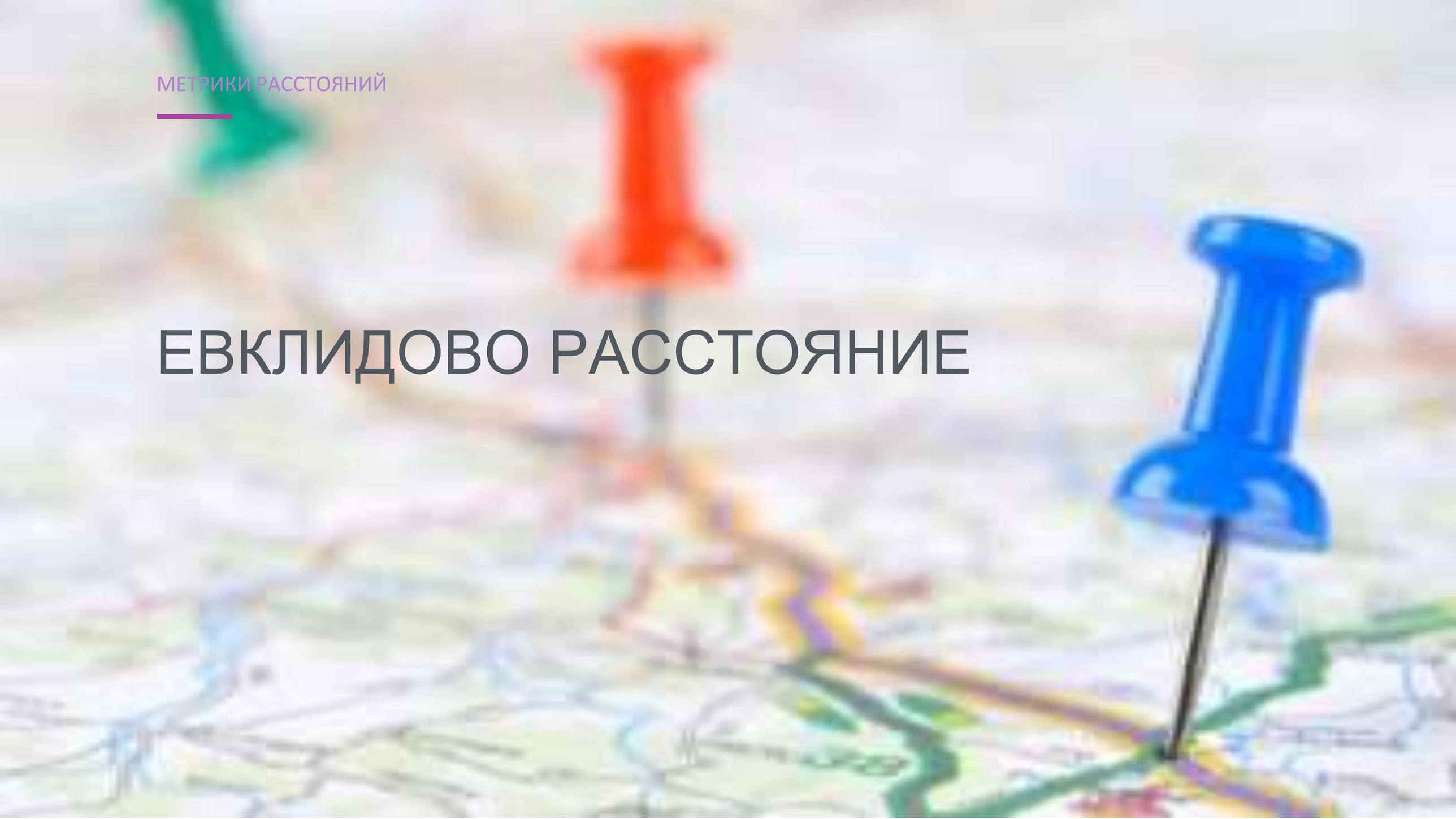
МЕТРИКИ РАССТОЯНИЙ И АЛГОРИТМ KNN

ЦЕЛИ ЗАНЯТИЯ

В КОНЦЕ ЗАНЯТИЯ ВЫ:

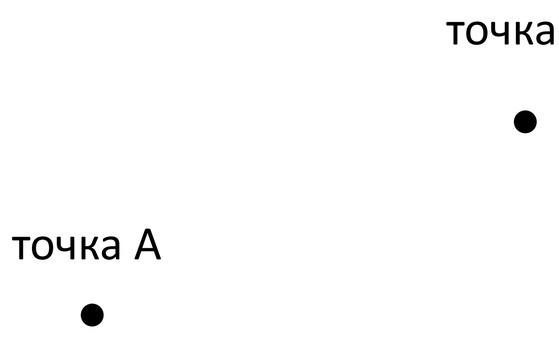
- будете знать как выбирать метрики близости;
- познакомитесь с алгоритмом KNN;
- потренируемся на различных метриках
- реализуете в коде задачу классификации и регрессии с помощью алгоритма KNN.

МЕТРИКИ РАССТОЯНИЙ

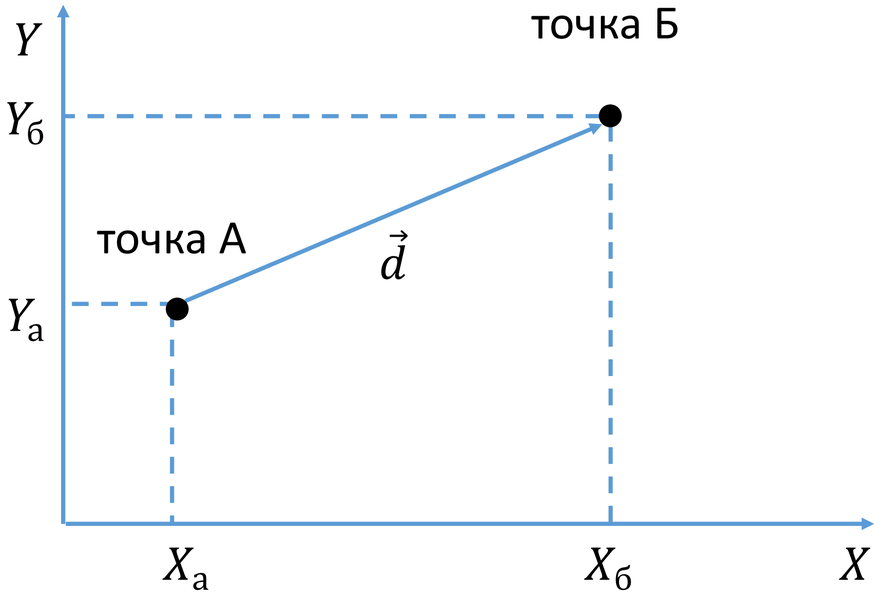


ЕВКЛИДОВО РАССТОЯНИЕ

ТОЧКИ НА ПЛОСКОСТИ

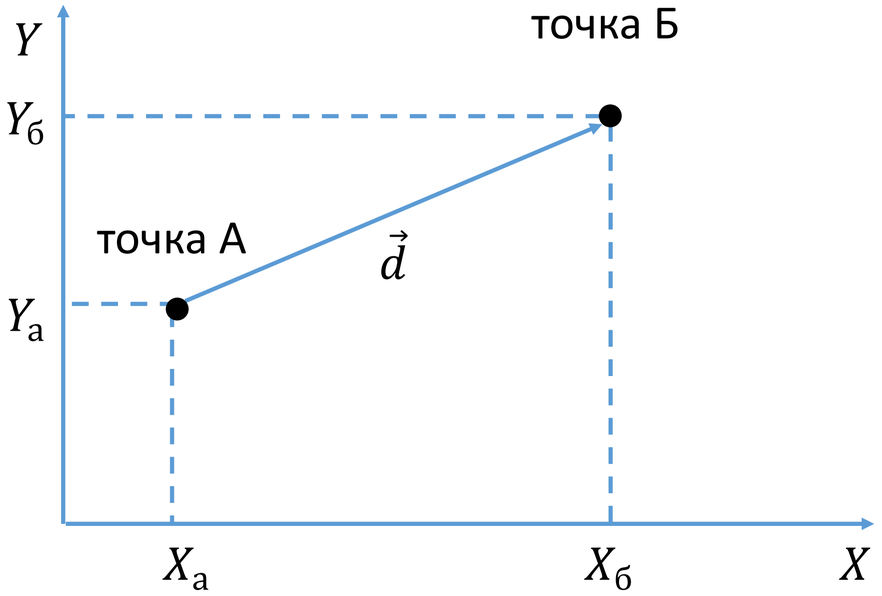


ТОЧКИ НА ПЛОСКОСТИ



$$d = \sqrt{(X_б - X_а)^2 + (Y_б - Y_а)^2}$$

ТОЧКИ НА ПЛОСКОСТИ



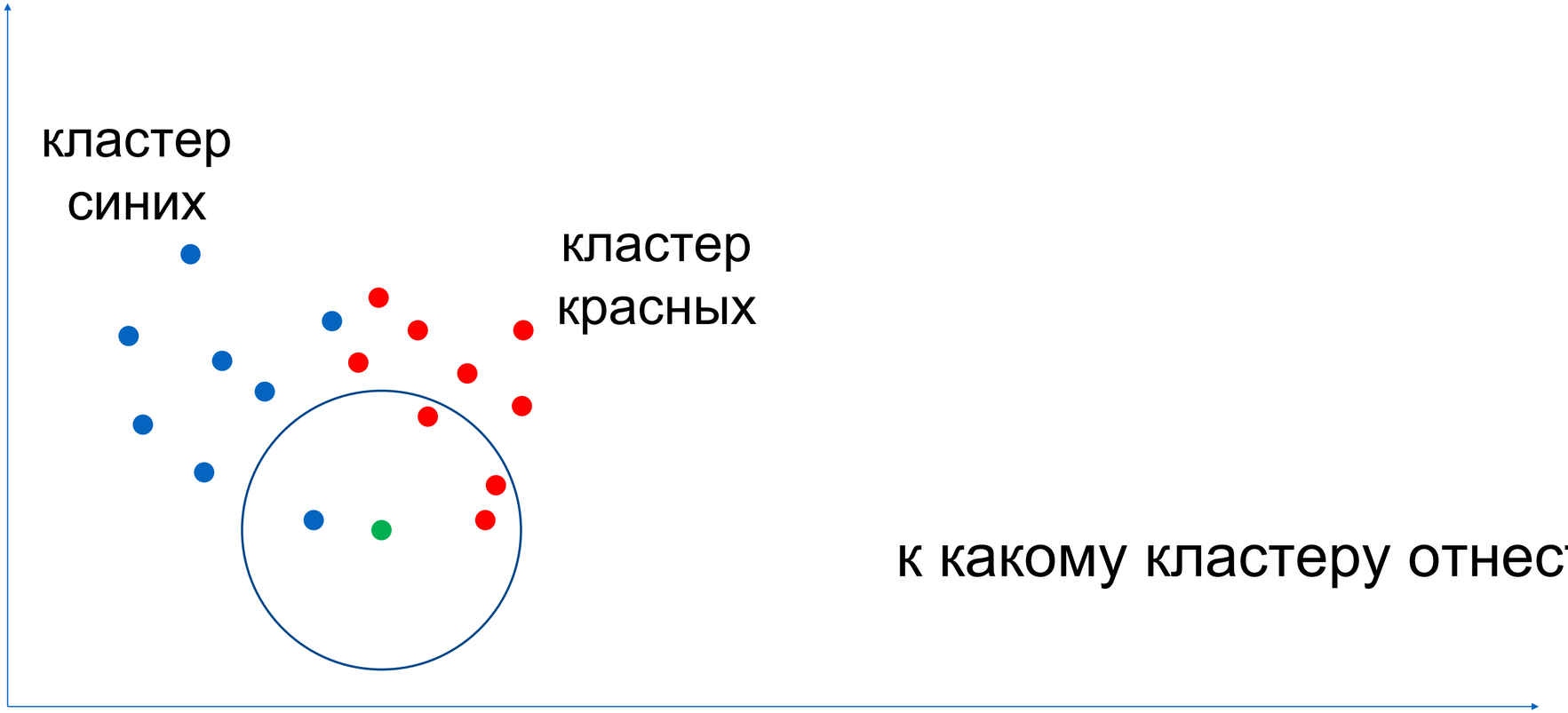
$$d = \sqrt{(X_б - X_а)^2 + (Y_б - Y_а)^2}$$

$$d = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$$

K NEAREST NEIGHBOR

ИДЕЯ АЛГОРИТМА

ИДЕЯ АЛГОРИТМА



к какому кластеру отнести зеленую точку?

ИДЕЯ АЛГОРИТМА

Берем К ближайших соседей к зеленой точке. Берем класс, наиболее часто встречающийся среди соседей.

Варианты:

- **Берем ближайшую точку (k = 1) – группа синих**
- **Учитываем несколько соседей (k = 4) – группа красных**
- **Учитываем вес, обратно пропорциональный расстоянию до точки**

ПРЕИМУЩЕСТВА И НЕДОСТАТКИ

+ Простая реализация и интерпретация

+ Применим ко многим задачам классификации и регрессии

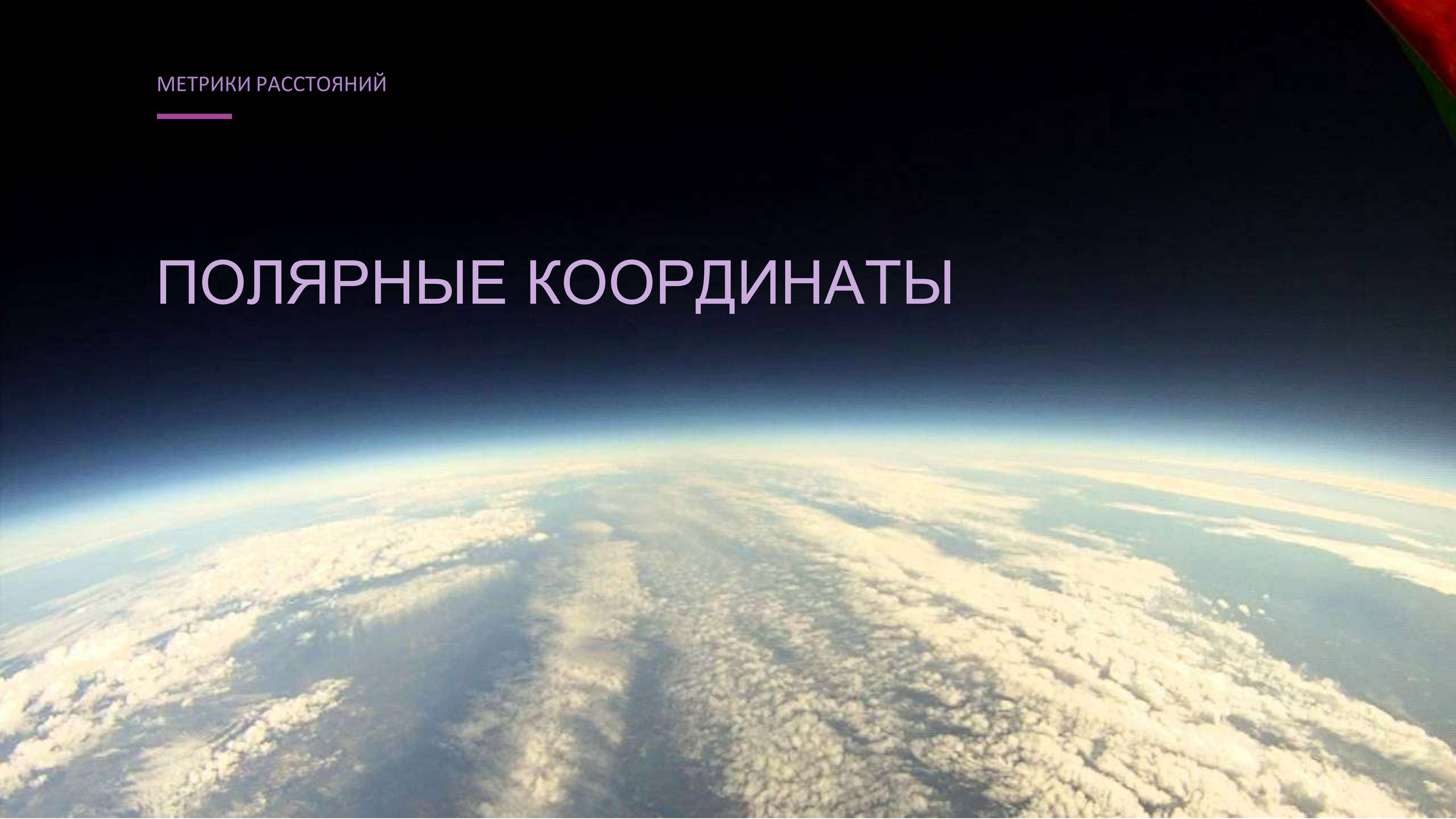
- Число соседей нужно задавать заранее, что иногда определяет результат

- Плохо работает при сильно пересекающихся данных

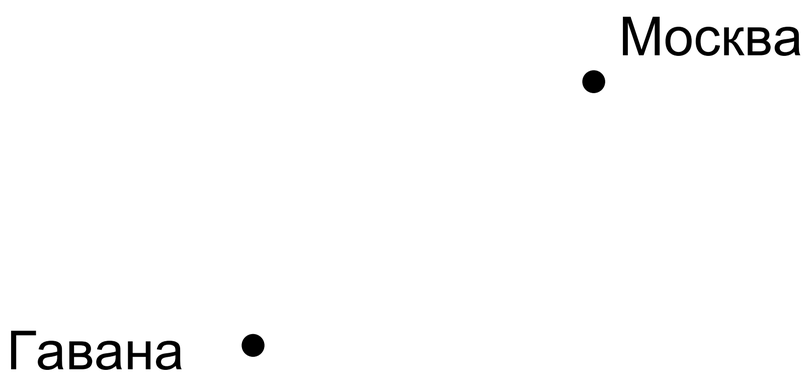
ВРЕМЯ ПРАКТИКИ

KNN.IPYNB

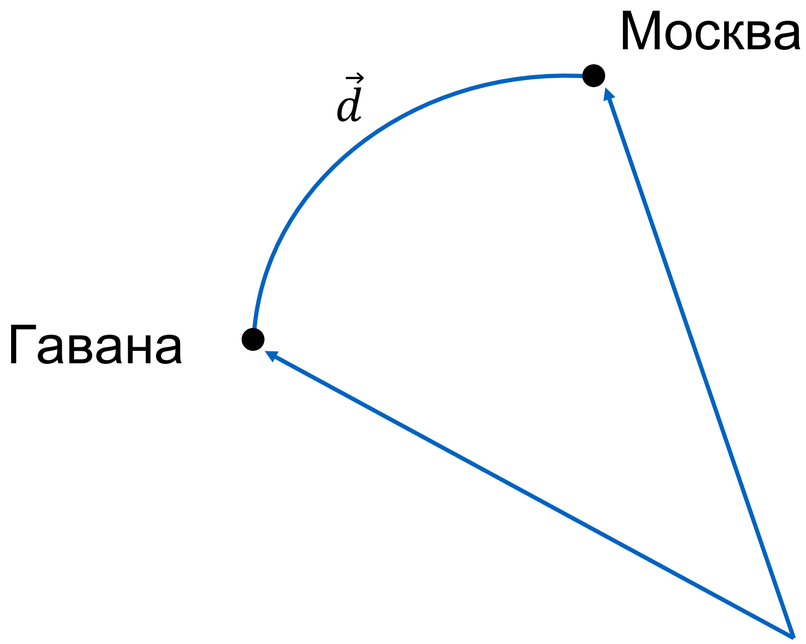
ПОЛЯРНЫЕ КООРДИНАТЫ



УЧЕТ КРИВИЗНЫ ПОВЕРХНОСТИ

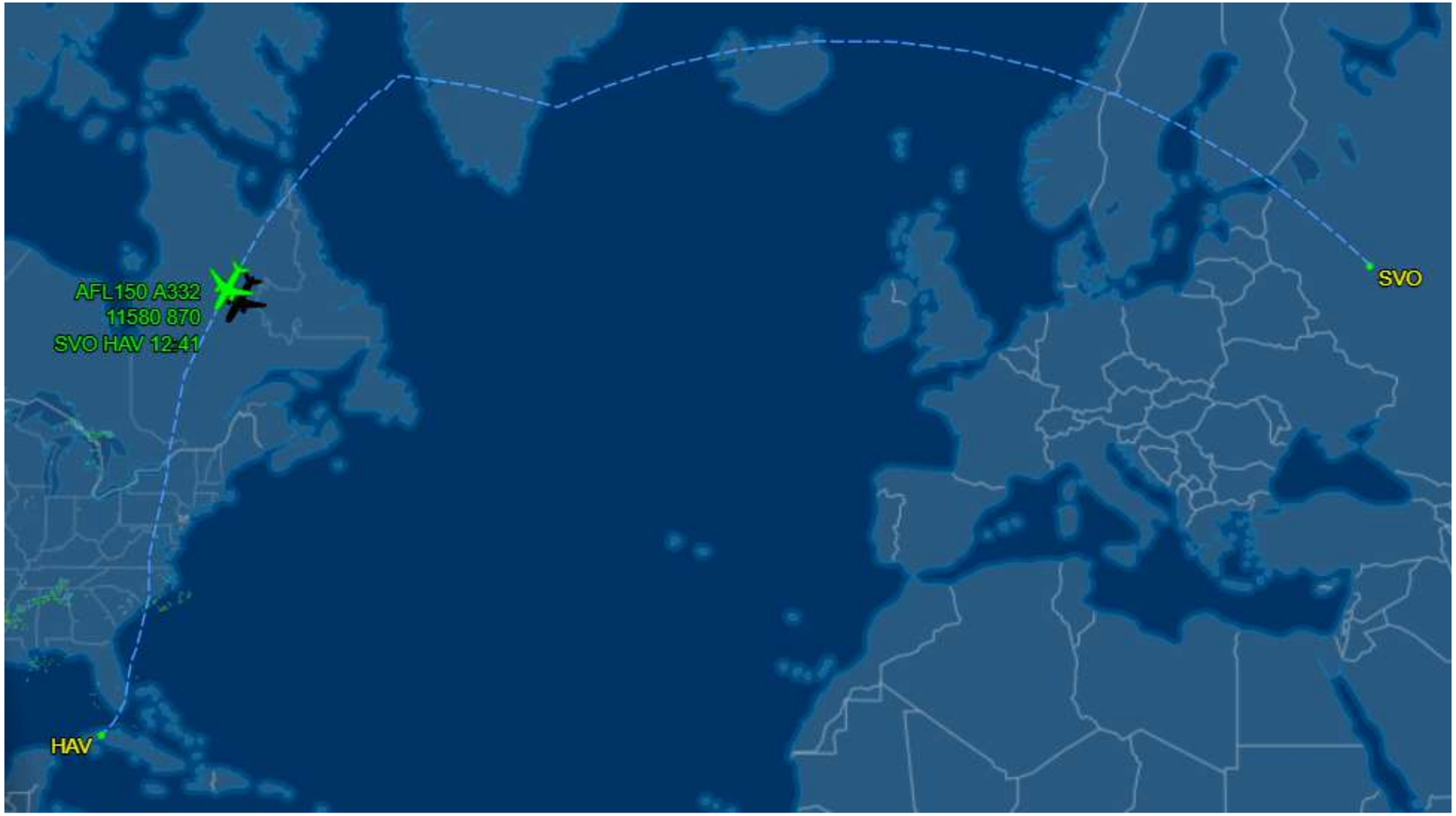


УЧЕТ КРИВИЗНЫ ПОВЕРХНОСТИ



d – длина дуги в полярных координатах

КАК НА САМОМ ДЕЛЕ

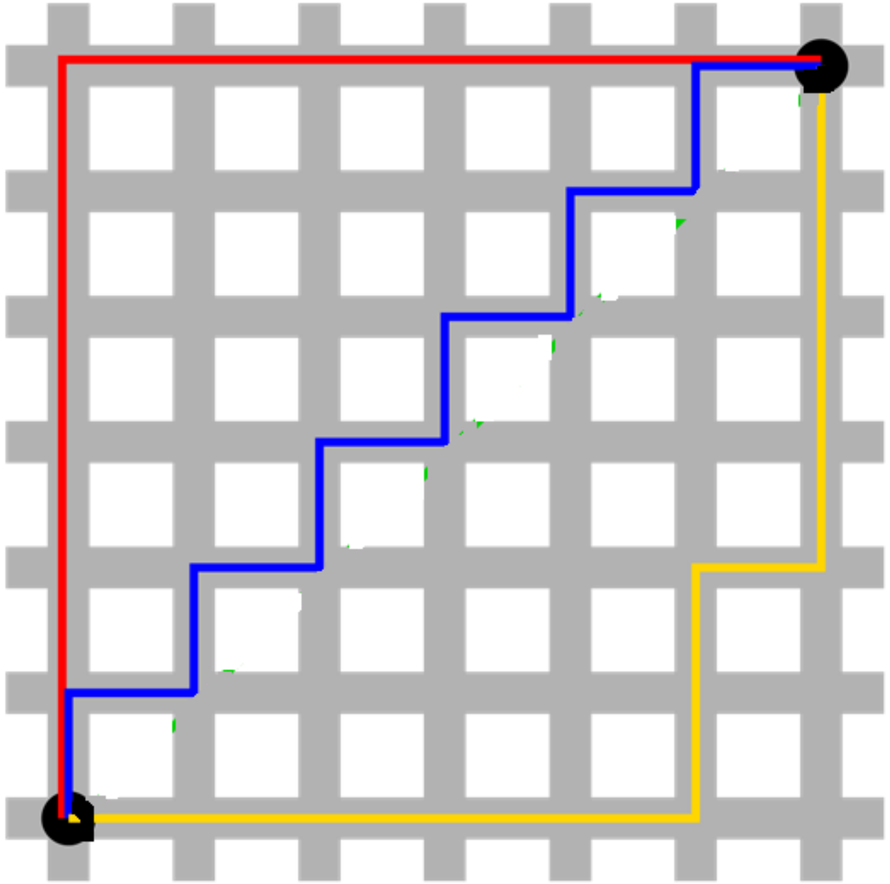


РАССТОЯНИЕ И ПУТЬ

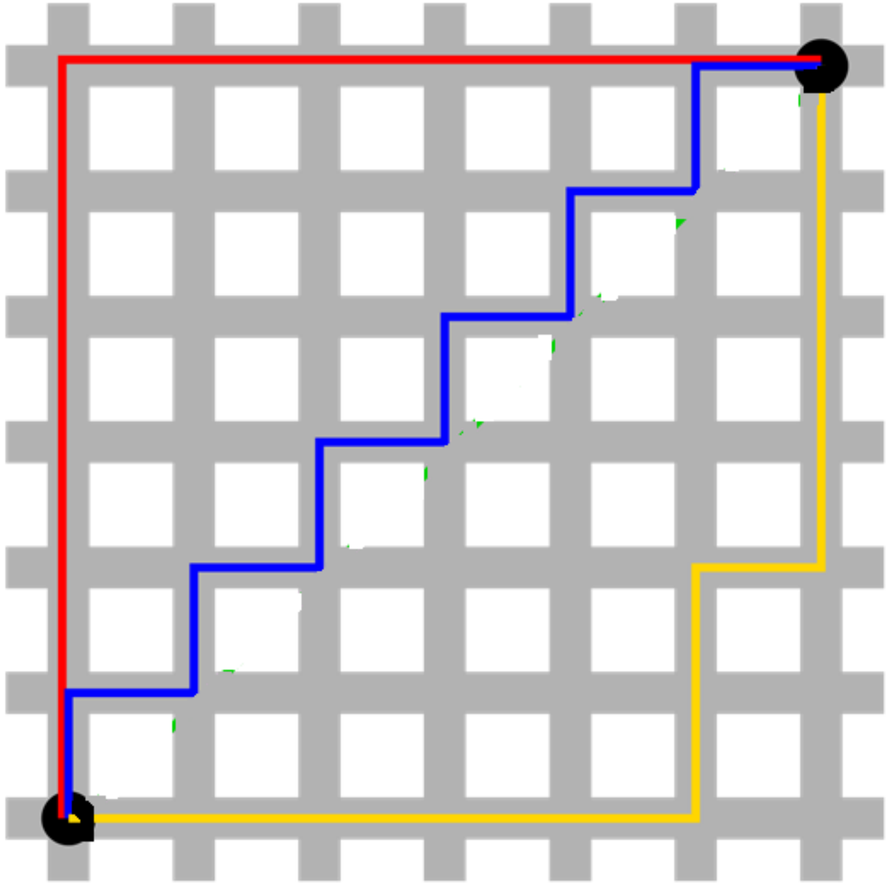
МАНХЭТТЕНСКОЕ РАССТОЯНИЕ

Улицы Манхэттена перпендикулярны друг другу

ДЛИНЫ ВСЕХ ПУТЕЙ РАВНЫ



ДЛИНЫ ВСЕХ ПУТЕЙ РАВНЫ

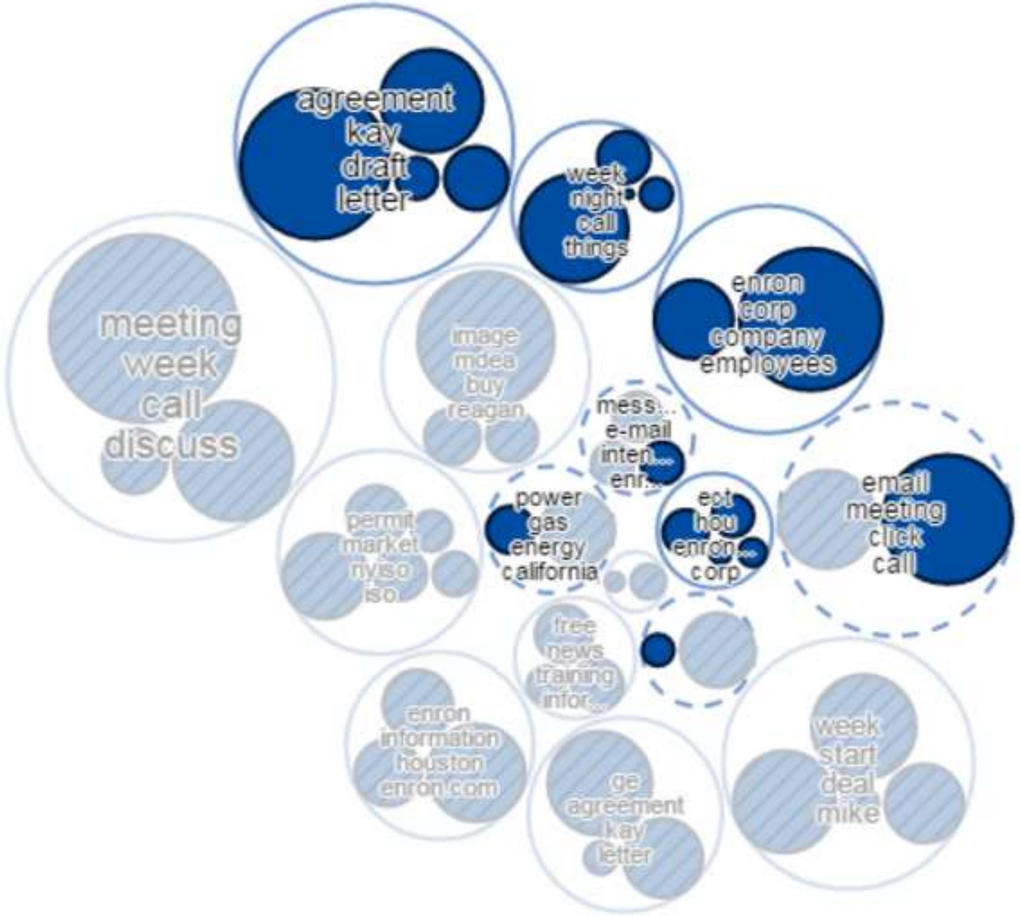


Расстояние городских кварталов

$$d = \sum_{i=1}^{n} |X_i - Y_i|$$

МЕТРИКИ БЛИЗОСТИ ОБЪЕКТОВ

СРАВНЕНИЕ ТЕКСТОВ



ПРОБЛЕМА

СТАРТОВЫЙ ЛИСТ

1	Шехавцова Анна	Ж	1998	РГАУ-МСХА
2	Гречихина Наталья	Ж	1994	МГУ
3	Козлова Алена	Ж	1994	МГУ
4	Груздева Алина	Ж	1998	РГУНГ
5	Кущенко Анна	Ж	1997	МГУ
6	Чистякова Анастасия	Ж	1998	РГАУ-МСХА

РАСПОЗНАВАНИЕ РЕЧИ

```
# результат расшифровки речи диктора

speech_recognition = [
    'кучменко она',
    'кущенко оксана',
    'груздь алина',
    'рычихина наталья',
    'шиховцева на',
    'чистова анастасия'
]
```

РАССТОЯНИЕ ХЭММИНГА

Число позиций, в которых соответствующие символы двух слов одинаковой длины различны

кареты
ракеты
} 2

РАССТОЯНИЕ ХЭММИНГА

В телекоме для отслеживания ошибок

В биоинформатике для оценки стабильности цепи

https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.spatial.distance.hamming.html

РАССТОЯНИЕ ЛЕВЕНШТЕЙНА

Минимальное количество операций вставки одного символа, удаления одного символа и замены одного символа на другой, необходимых для превращения одной строки в другую.

M	M	M	R	I	M	R	R
C	O	N	N		E	C	T
C	O	N	E	H	E	A	D

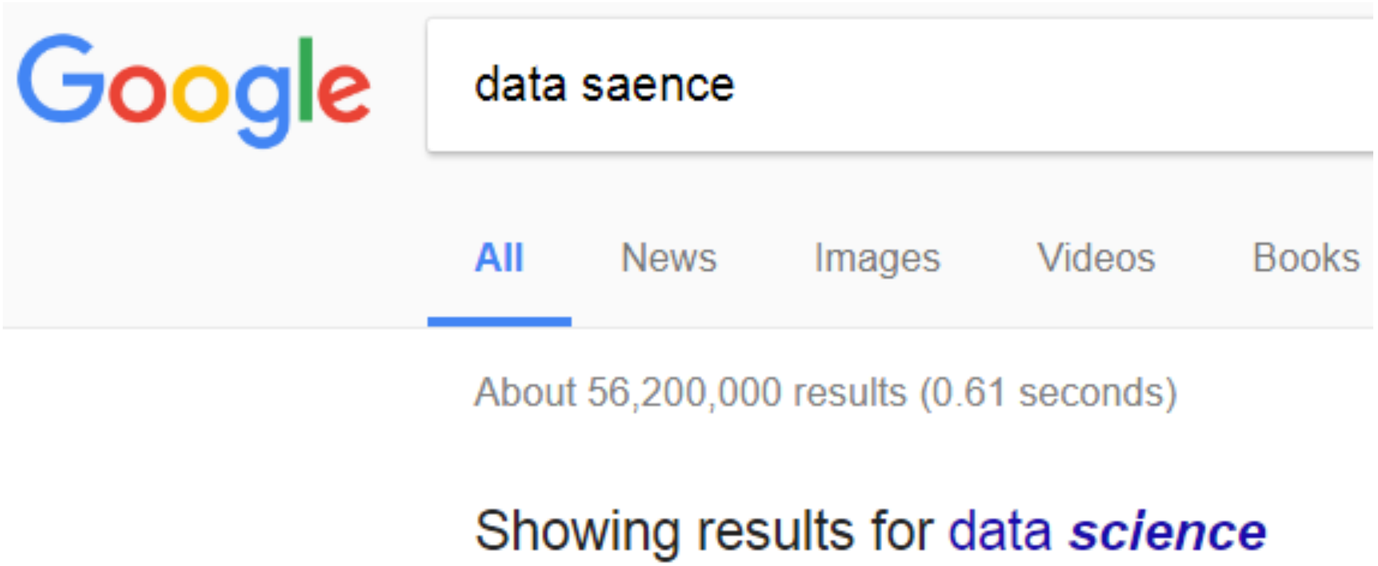
D – удалить,

I – вставить,

R – заменить,

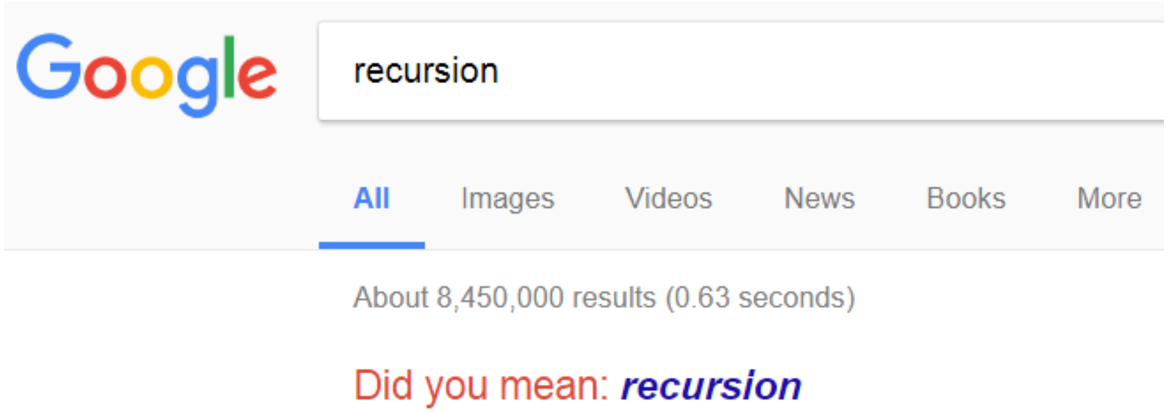
M – совпадение

РАССТОЯНИЕ ЛЕВЕНШТЕЙНА



РАССТОЯНИЕ ДАМЕРАУ-ЛЕВЕНШТЕЙНА

То же самое, но с добавлением операции транспозиции (перестановки символов)



юмор Гугла

СЛОВА И ВЕКТОРЫ

МЕТРИКА TF-IDF

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j

df_i = number of documents containing i

N = total number of documents

МЕРА ВАЖНОСТИ ДОКУМЕНТА

TF (term frequency — частота слова) — отношение числа вхождений некоторого слова к общему числу слов документа

IDF (inverse document frequency — обратная частота документа) — инверсия частоты, с которой некоторое слово

встречается в документах коллекции

МЕРА ВАЖНОСТИ ДОКУМЕНТА

TF-IDF имеет много модификаций под разные задачи

Вариант определения для поисковых систем (т. н. BM25)

Пусть дан запрос Q, содержащий слова $q_1, \ldots, q_n$, тогда функция BM25 даёт следующую оценку релевантности документа D запросу Q:

$$\mathrm{score}(D,Q) = \sum_{i=1}^{n} \mathrm{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\mathrm{avgdl}})},$$

где $f(q_i, D)$ есть частота слова (англ. *term frequency, TF*) q_i в документе D, $|D|$ есть длина документа (количество слов в нём), а $avgdl$ — средняя длина документа в коллекции. k_1 и b — свободные коэффициенты, обычно их выбирают как $k_1 = 2.0$ и $b = 0.75$.

СХОЖЕСТЬ ПОЛЬЗОВАТЕЛЕЙ

КОЭФФИЦИЕНТ ЖАККАРА

$$K = \frac{n(A \cap B)}{n(A \cup B)}$$

Отношение количества элементов, общих для множеств A и B, к общему количеству элементов в этих множествах

КОЭФФИЦИЕНТ ЖАККАРА

Удобно использовать в рекомендательных системах

Товары

Признак	Телефон 1 vs 2
Память	совпадает
Экран	разный
Процессор	совпадает

Предпочтения пользователей

Фильм	Пользователь 1	Пользователь 2
Гадкий Я	★★★★	★
Мумия	★★	★★★
Пираты	★★★★★	★★★★★

KNN И РЕГРЕССИЯ

ПРОСТО ПОСМОТРИМ КОД

KNN REGRESSION.IPYNB

ЧТО МЫ СЕГОДНЯ УЗНАЛИ

ЧТО МЫ СЕГОДНЯ УЗНАЛИ

1. Метрики расстояний и близости объектов в применении к различным задачам.
2. Рассмотрели идею алгоритма KNN.
3. Реализовали на практике алгоритм KNN в задачах классификации и регрессии.

ПОЛЕЗНЫЕ МАТЕРИАЛЫ

1. Примеры различных корреляций
   http://www.tylervigen.com/spurious-correlations
2. Блог Open Data Science
   https://habrahabr.ru/company/ods/blog/322534/#metod-blizhayshih-sosedey
3. Еще примеры метрик https://ru.coursera.org/learn/supervised-learning/lecture/gqbPI/mietriki-v-knn

НЕТОЛОГИЯ
групп

Спасибо за внимание!

АРТУР САПРЫКИН