

## ЗАНЯТИЕ 4.4

# МЕТРИКИ РАССТОЯНИЙ И АЛГОРИТМ KNN



**Артур Сапрыкин**

---

# 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.

---

О ЧЁМ ПОГОВОРИМ И ЧТО  
СДЕЛАЕМ

- 
1. Что такое близость объектов и в каких задачах это нужно: примеры
  2. Идея и особенности алгоритма KNN: теория;
  3. Еще немного практики на различных метриках
  4. Пример решения задачи классификации KNN: практика;
  5. Пример решения задачи регрессии через KNN: тоже практика.

---

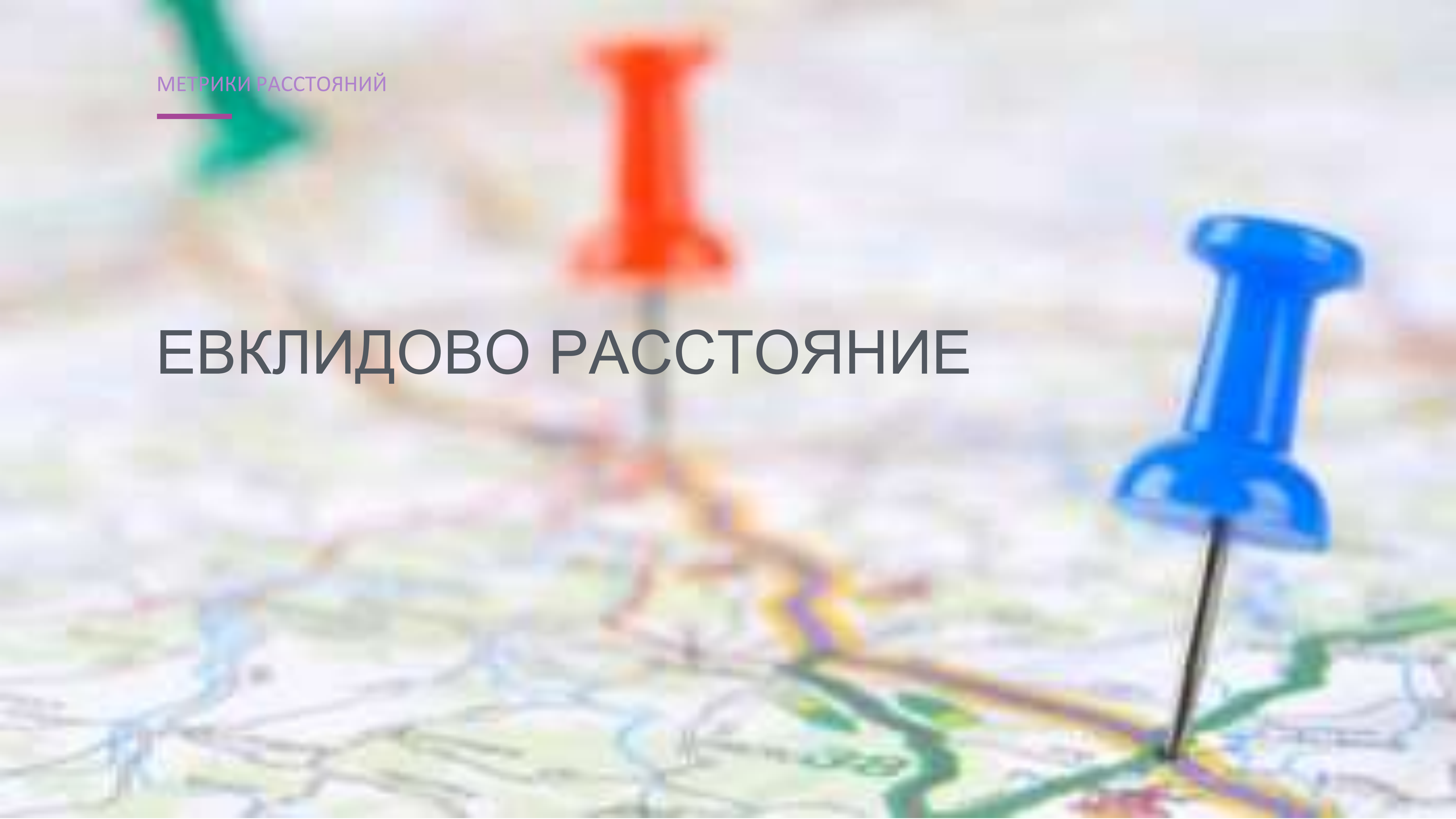
# МЕТРИКИ РАССТОЯНИЙ



МЕТРИКИ РАССТОЯНИЙ

---

# ЕВКЛИДОВО РАССТОЯНИЕ





ЕВКЛИДОВО РАССТОЯНИЕ



# ТОЧКИ НА ПЛОСКОСТИ

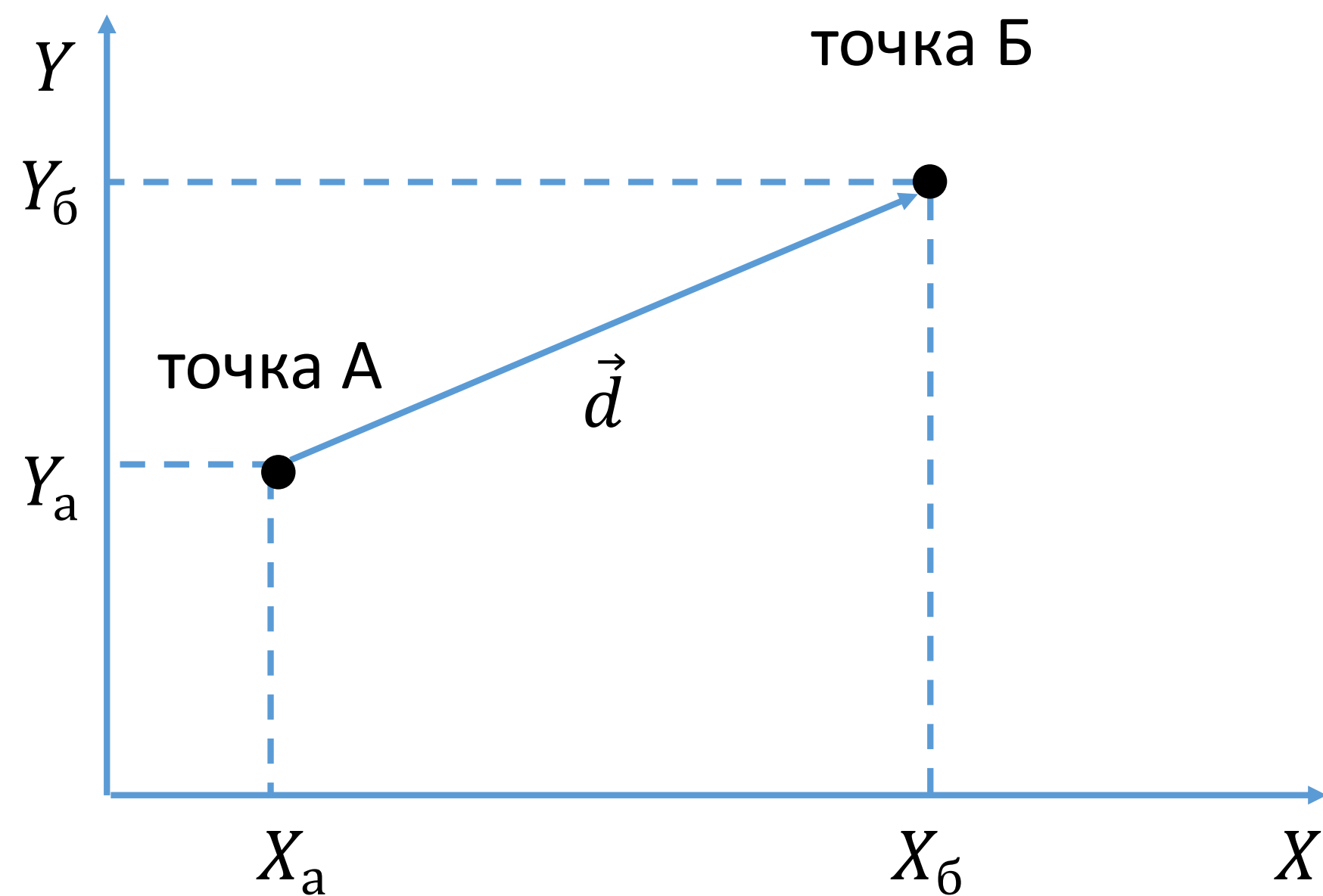
точка Б



точка А

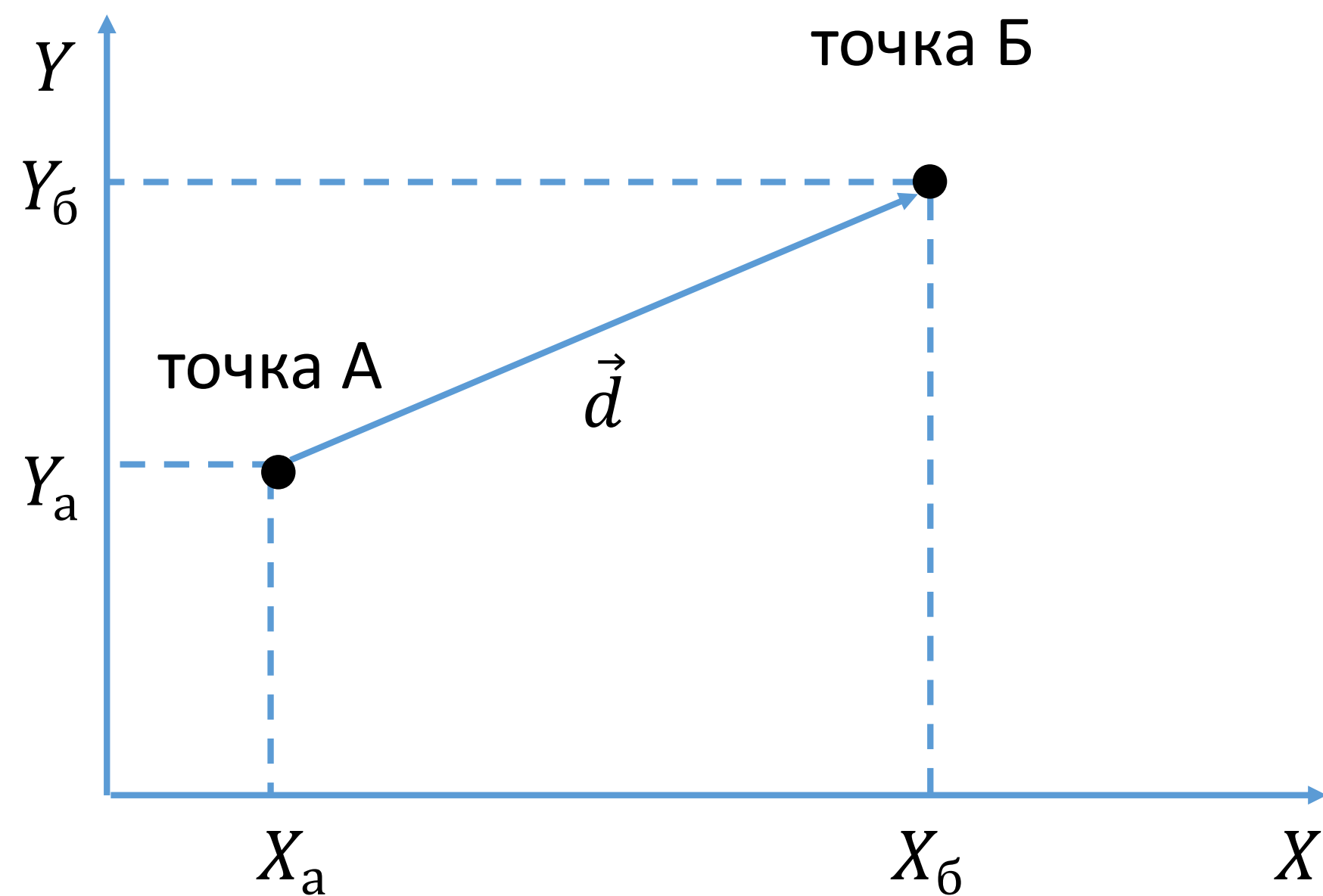


# ТОЧКИ НА ПЛОСКОСТИ



$$d = \sqrt{(X_b - X_a)^2 + (Y_b - Y_a)^2}$$

# ТОЧКИ НА ПЛОСКОСТИ



$$d = \sqrt{(X_б - X_a)^2 + (Y_б - Y_a)^2}$$

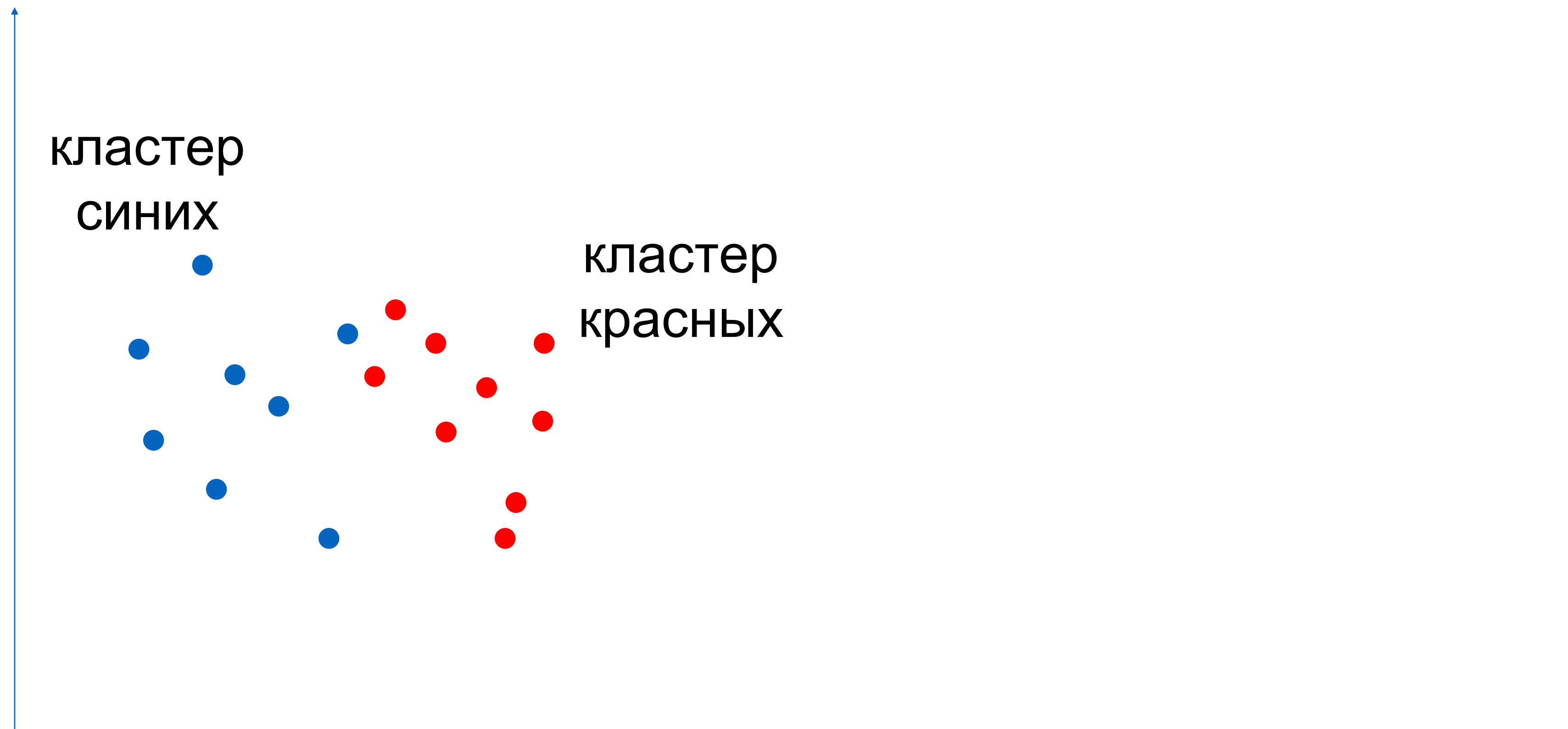
$$d = \sqrt{\sum_{i=1}^n (X_i - Y_i)^2}$$

---

# K NEAREST NEIGHBOR

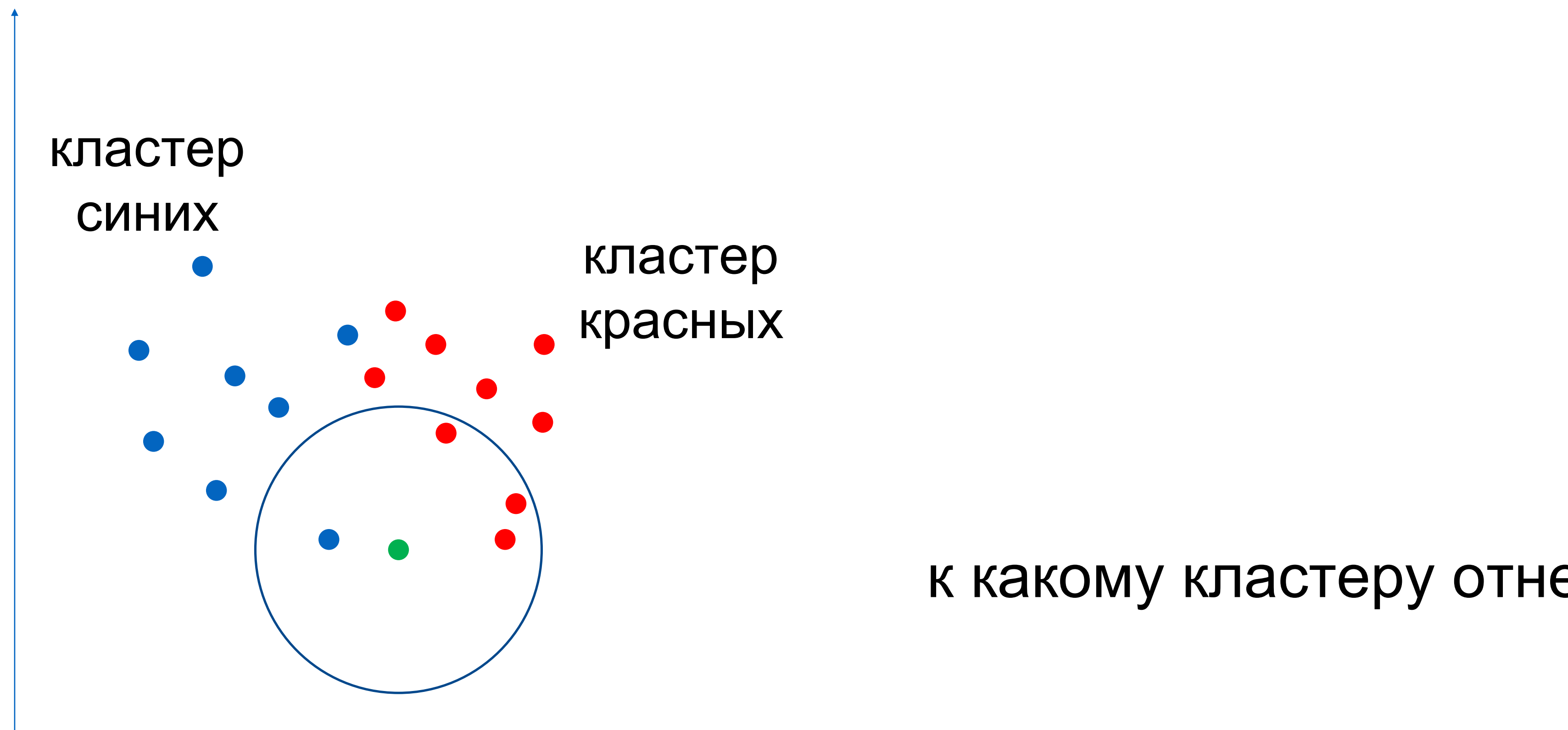
К БЛИЖАЙШИХ СОСЕДЕЙ

# ИДЕЯ АЛГОРИТМА



К БЛИЖАЙШИХ СОСЕДЕЙ

# ИДЕЯ АЛГОРИТМА



к какому кластеру отнести зеленую точку?



# ИДЕЯ АЛГОРИТМА

Берем  $K$  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

К БЛИЖАЙШИХ СОСЕДЕЙ

ВРЕМЯ ПРАКТИКИ

KNN.IPYNB



МЕТРИКИ РАССТОЯНИЙ

---

# ПОЛЯРНЫЕ КООРДИНАТЫ



ПОЛЯРНЫЕ КООРДИНАТЫ

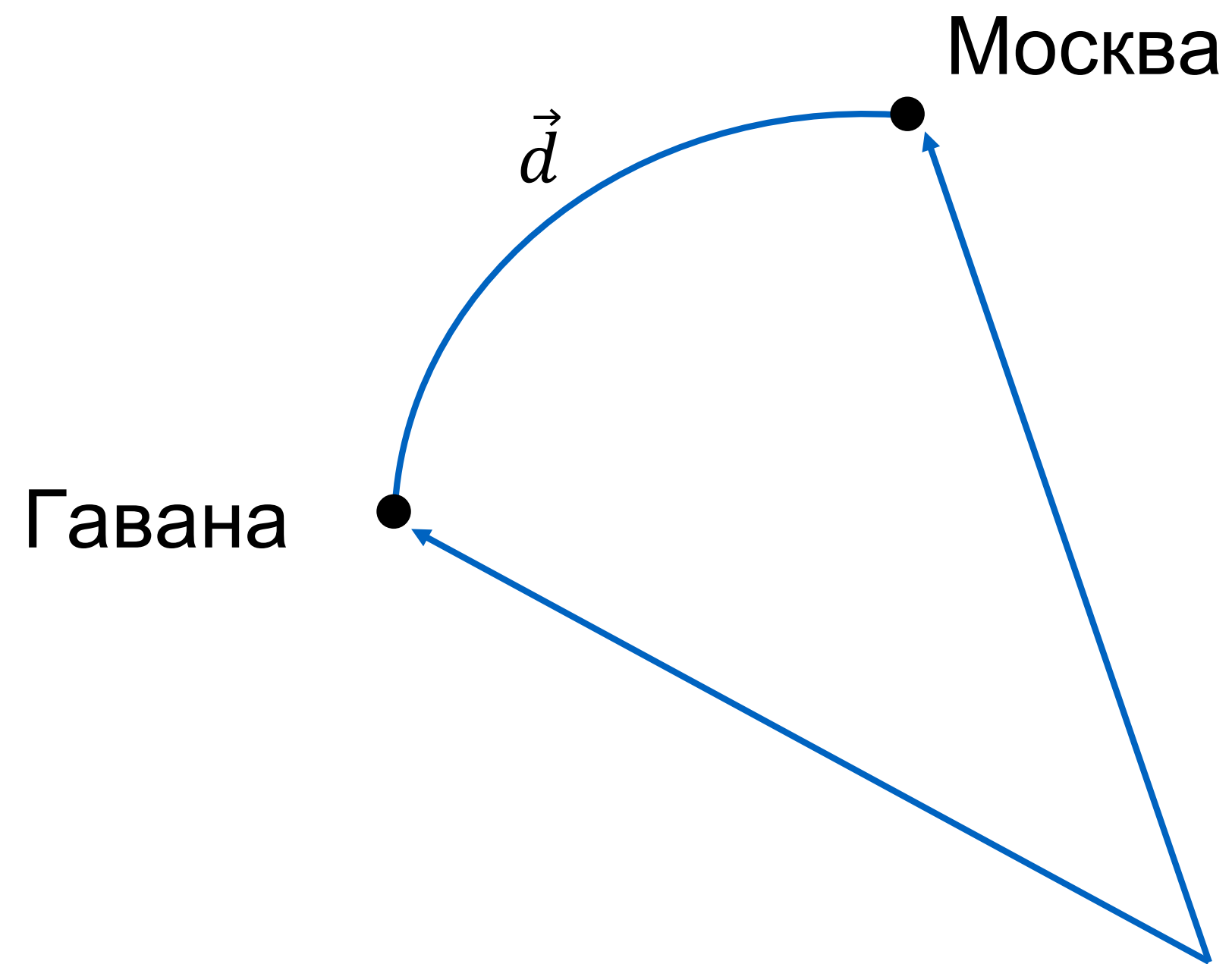


# УЧЕТ КРИВИЗНЫ ПОВЕРХНОСТИ

● Москва

Гавана ●

# УЧЕТ КРИВИЗНЫ ПОВЕРХНОСТИ



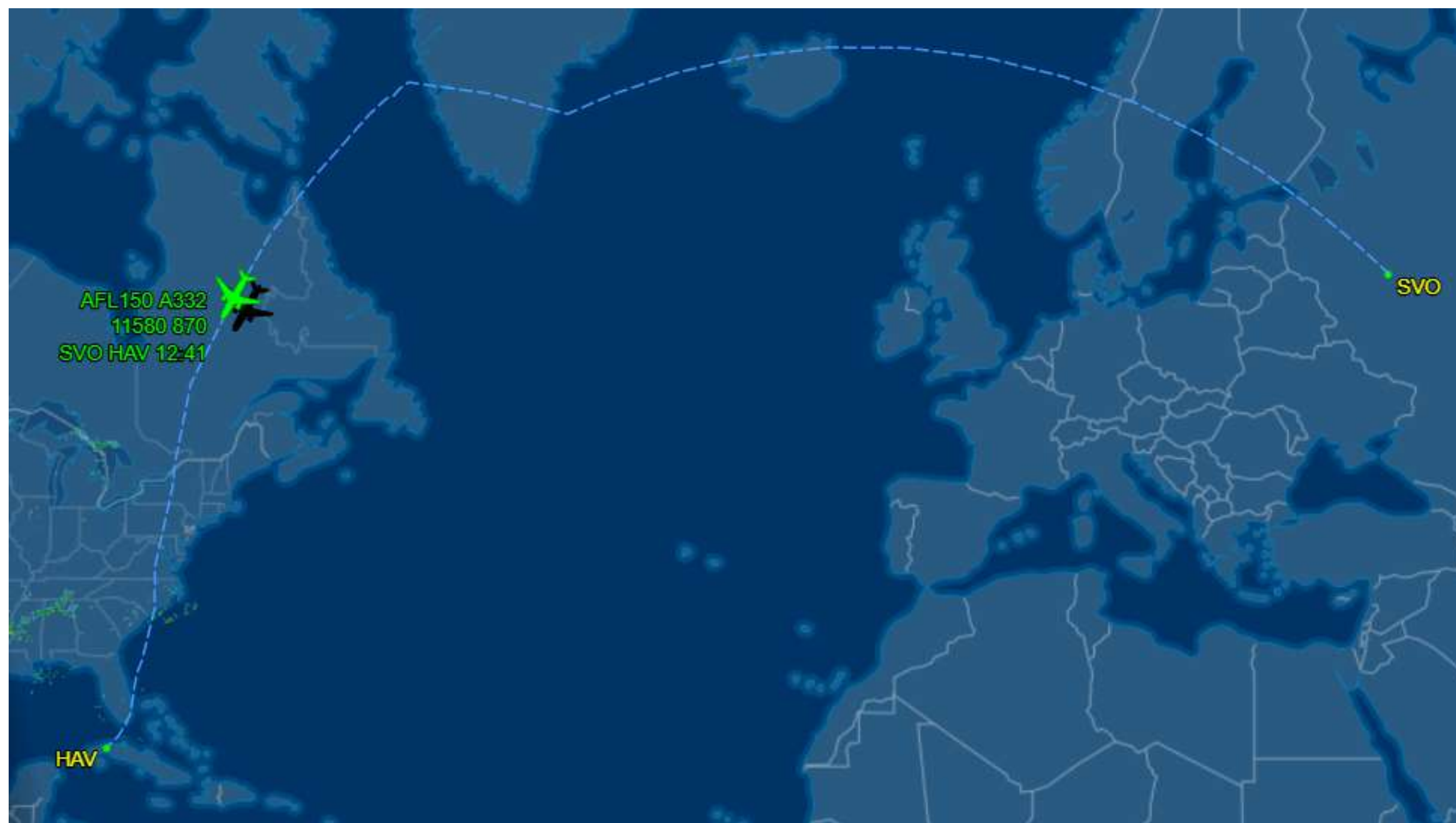
$d$  – длина дуги в полярных координатах



ПОЛЯРНЫЕ КООРДИНАТЫ

---

# КАК НА САМОМ ДЕЛЕ



---

# РАССТОЯНИЕ И ПУТЬ



МЕТРИКИ НА ПЛОСКОСТИ

---

# МАНХЭТТЕНСКОЕ РАССТОЯНИЕ

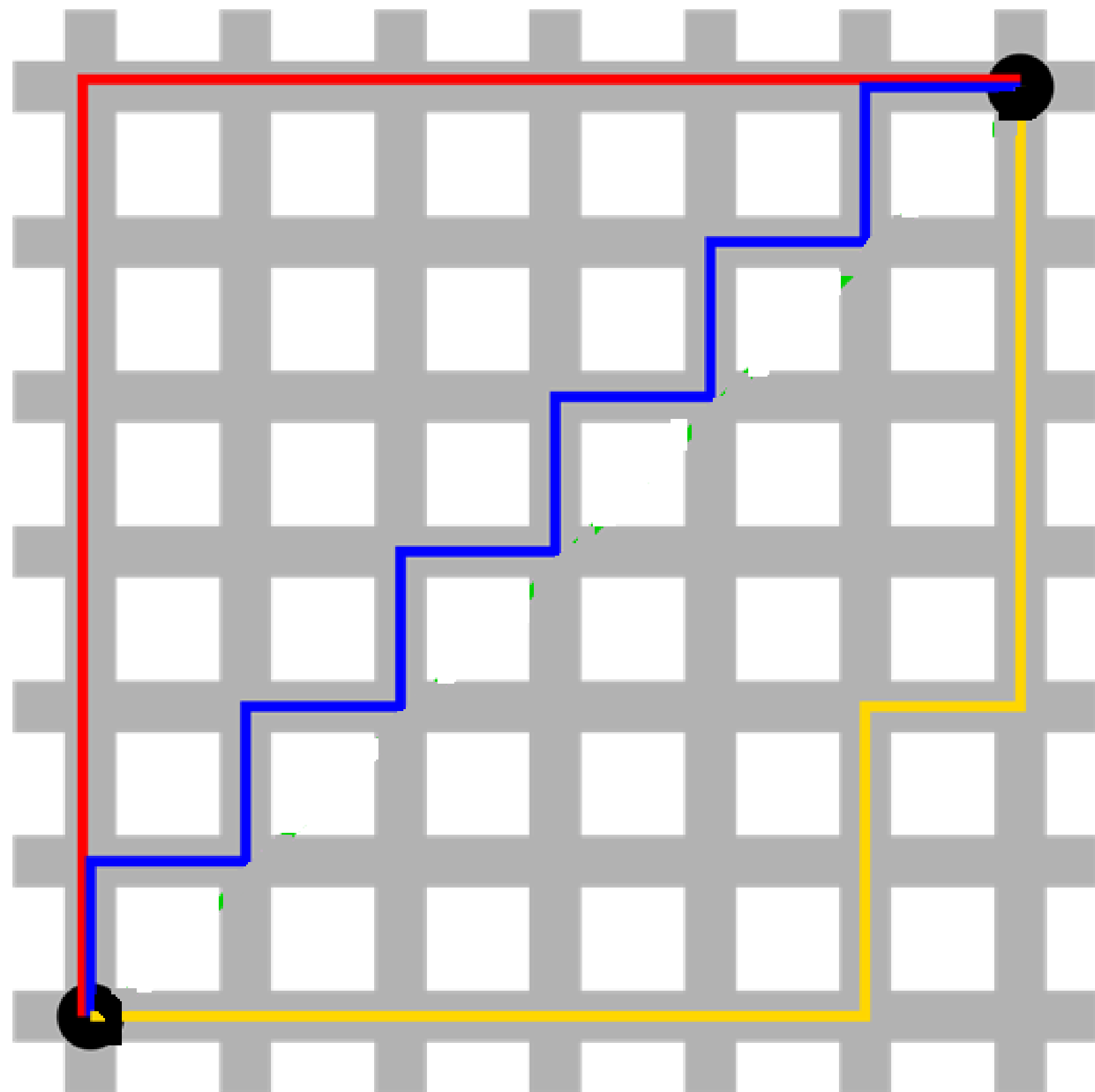
Улицы Манхэттена перпендикулярны друг другу



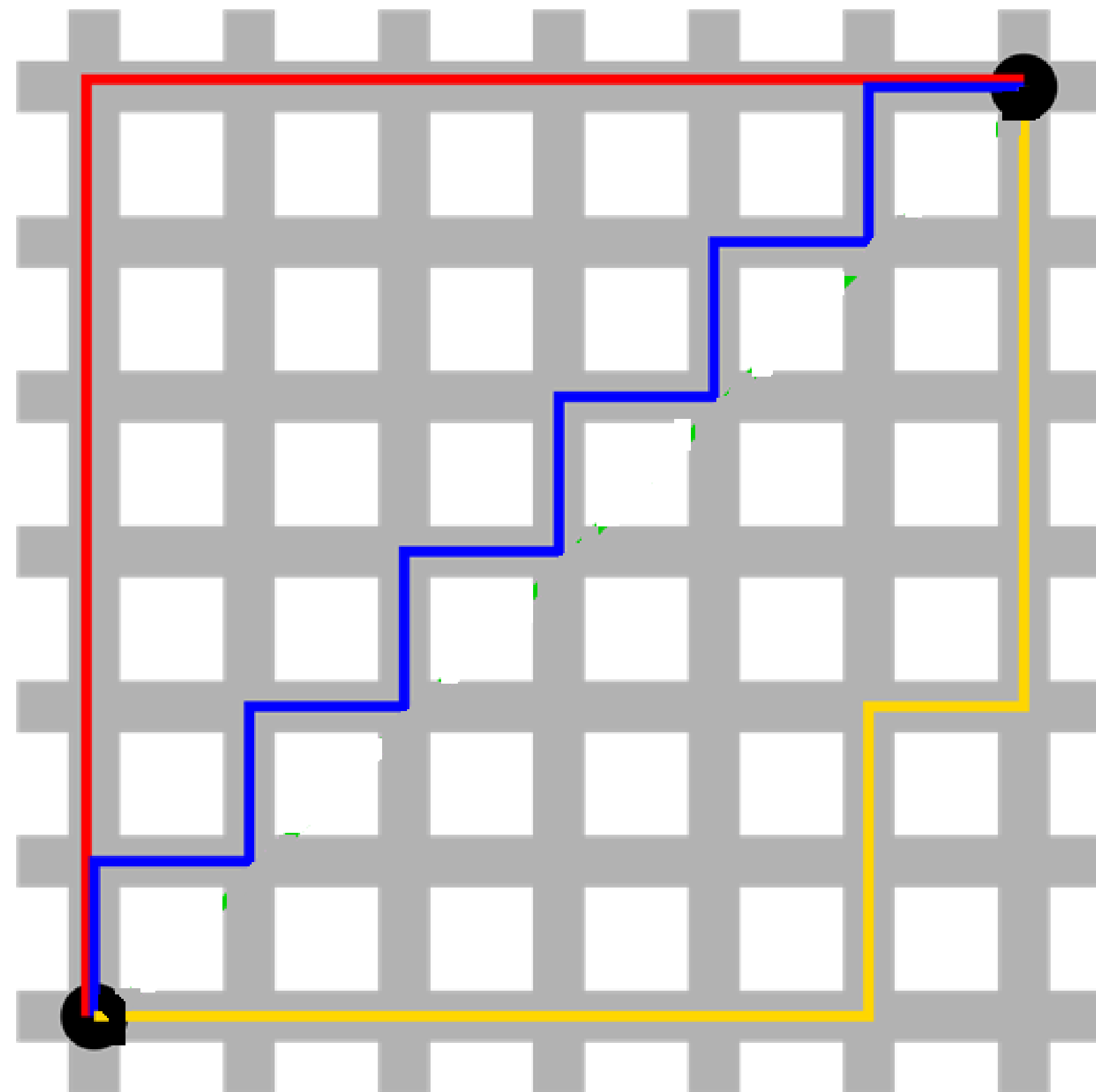
МАНХЭТТЕНСКОЕ РАССТОЯНИЕ



# ДЛИНЫ ВСЕХ ПУТЕЙ РАВНЫ



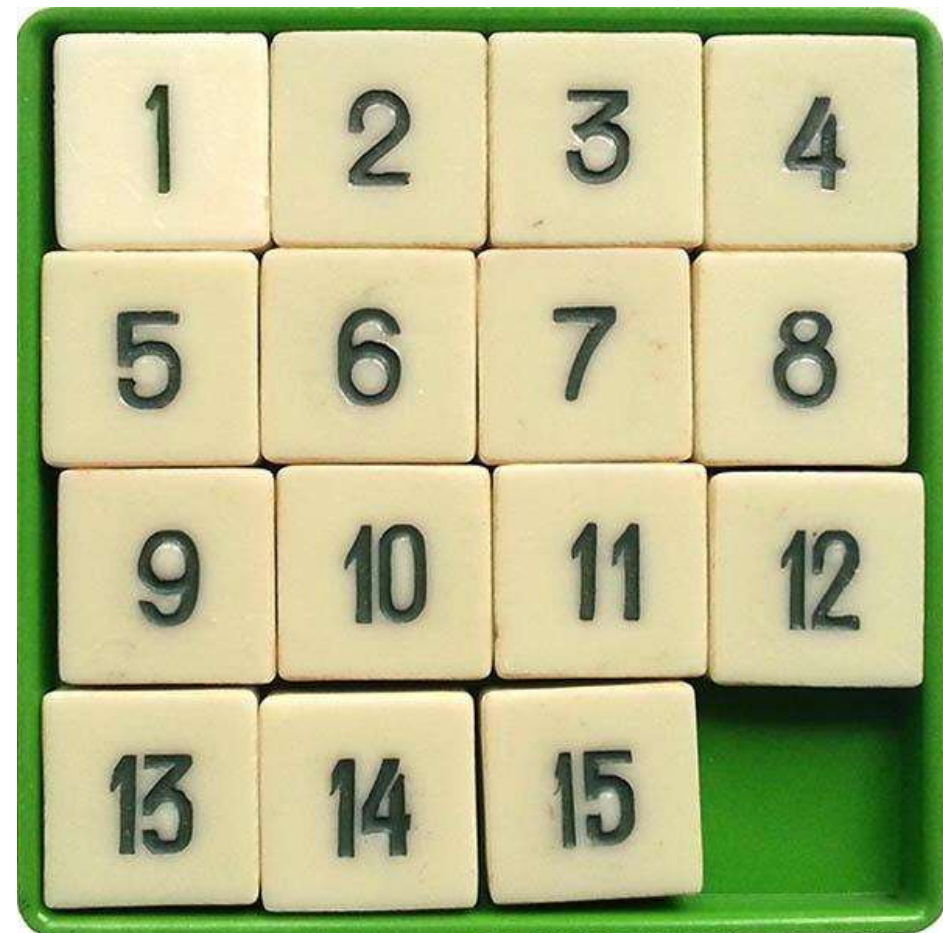
# ДЛИНЫ ВСЕХ ПУТЕЙ РАВНЫ



Расстояние городских кварталов

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

# ДЛЯ ПОИСКА ОПТИМАЛЬНОГО РЕШЕНИЯ



Сумма манхэттенских расстояний между костяшками и позициями, в которых они находятся в решённой головоломке «Пятнашки», используется в качестве эвристической функции для поиска оптимального решения



МАНХЭТТЕНСКОЕ РАССТОЯНИЕ



# ДЛЯ ПОИСКА ОПТИМАЛЬНОГО РЕШЕНИЯ

			2

В примере манхэттенское расстояние равно 4

---

# ПРАКТИЧЕСКОЕ ЗАДАНИЕ 1

МАНХЭТТЕНСКОЕ РАССТОЯНИЕ

---

МАНХЭТТЕНСКОЕ РАССТОЯНИЕ

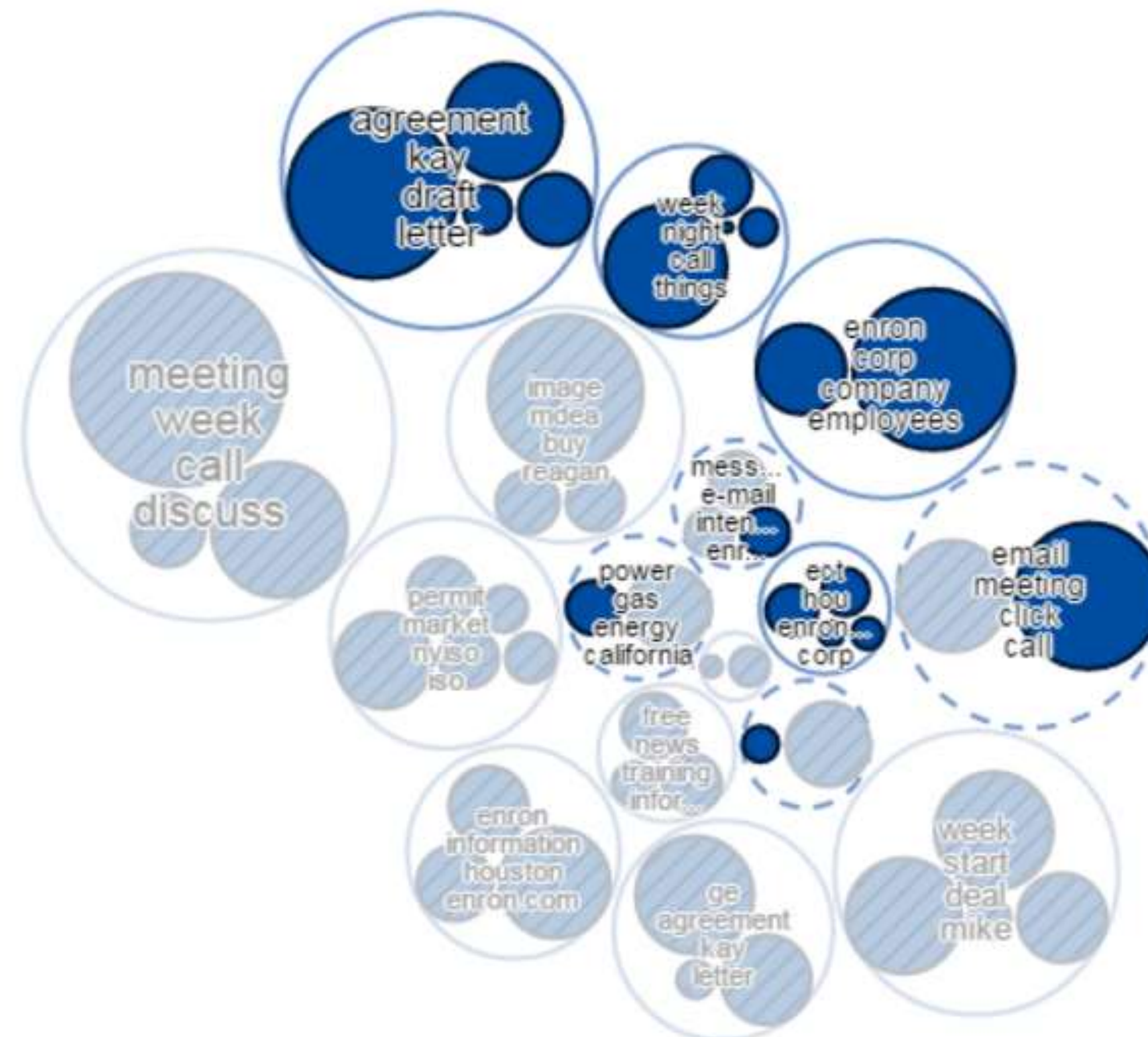
BARLEY-BREAK.IPYNB

---

# МЕТРИКИ БЛИЗОСТИ ОБЪЕКТОВ

БЛИЗОСТЬ ОБЪЕКТОВ

# СРАВНЕНИЕ ТЕКСТОВ



# 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
7				



# РАСПОЗНАВАНИЕ РЕЧИ

```
# результат расшифровки речи диктора  
  
speech_recognition = [  
    'кучменко она',  
    'кущенко оксана',  
    'груздь алина',  
    'рычихина натаалья',  
    '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>

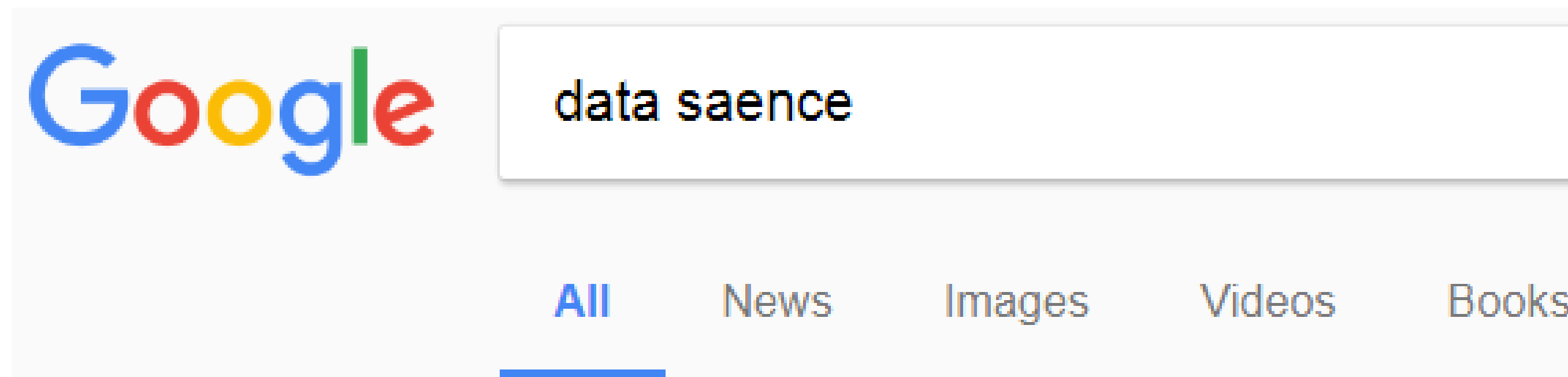
# РАССТОЯНИЕ ЛЕВЕНШТЕЙНА

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

М	М	М	Р	І	М	Р	Р
С	О	Н	Н		Е	С	Т
С	О	Н	Е	Н	Е	А	Д

D — удалить,  
I — вставить,  
R — заменить,  
M — совпадение

# РАССТОЯНИЕ ЛЕВЕНШТЕЙНА

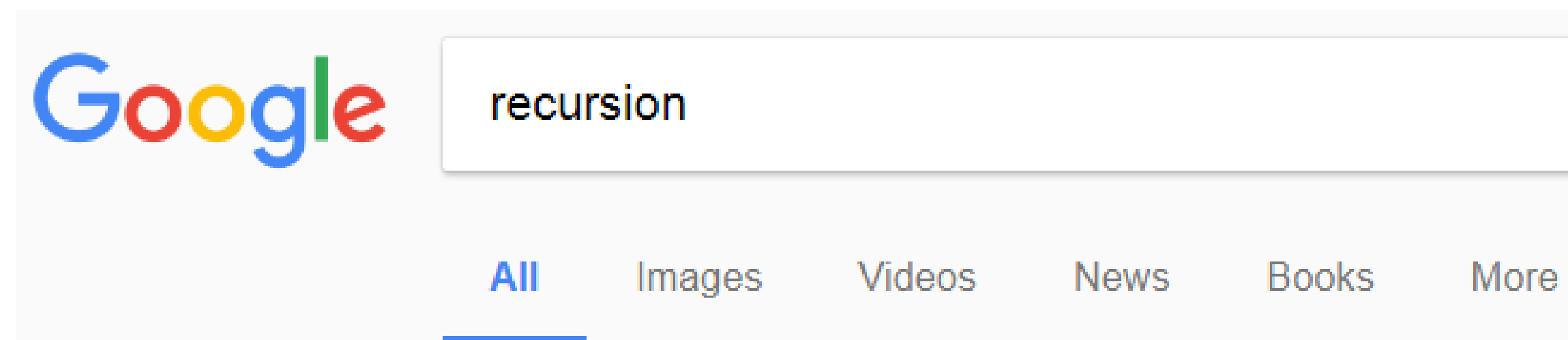


About 56,200,000 results (0.61 seconds)

Showing results for **data science**

# РАССТОЯНИЕ ДАМЕРАУ-ЛЕВЕНШТЕЙНА

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



About 8,450,000 results (0.63 seconds)

Did you mean: *recursion*

юмор Гугла

---

# ПРАКТИЧЕСКОЕ ЗАДАНИЕ 2

## (ДОМАШНЕЕ)



СРАВНЕНИЕ ТЕКСТОВ

---

ВРЕМЯ КОДА

LEVENSHTEIN DISTANCE. IPYNB

---

# 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)

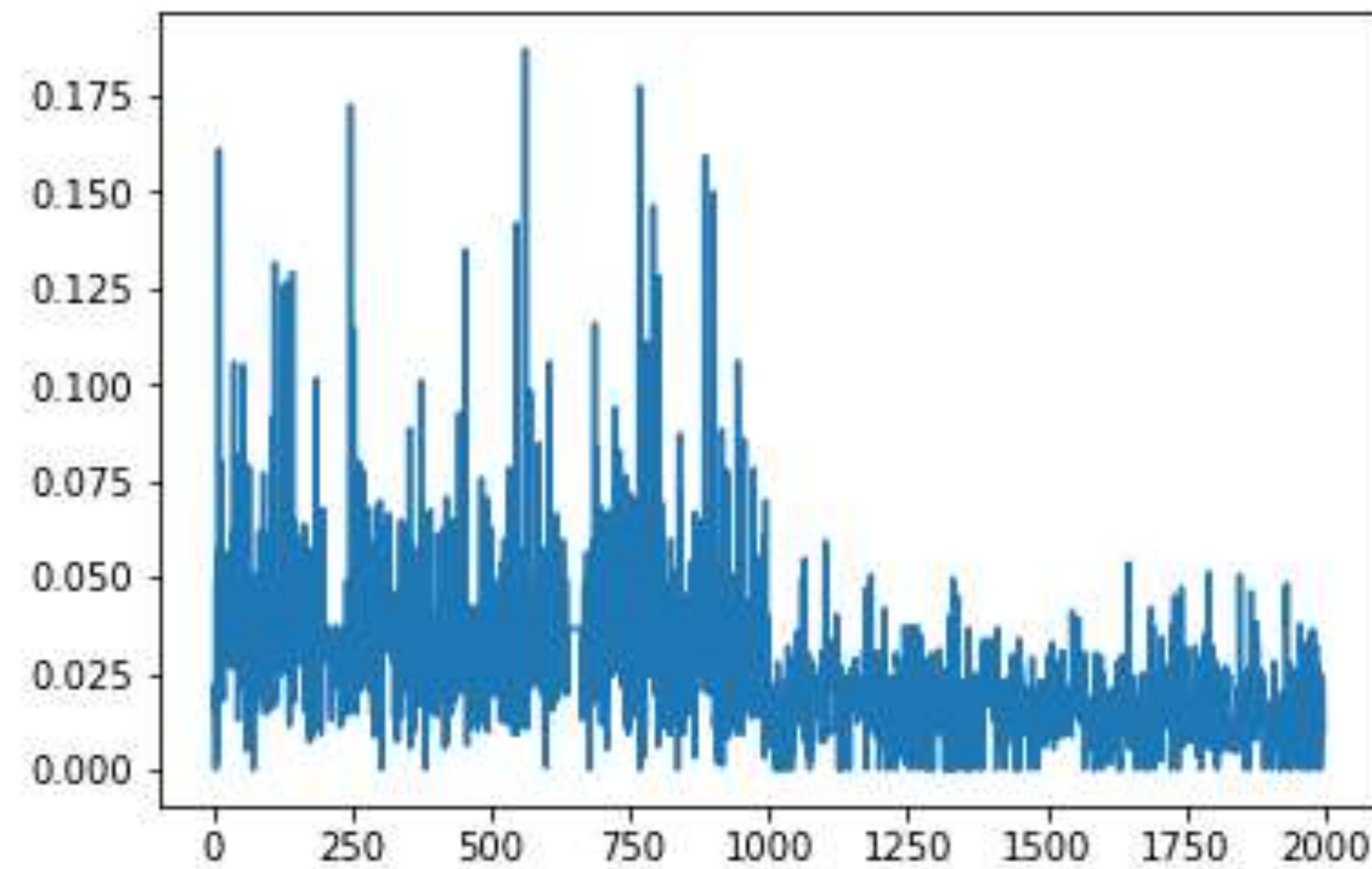
Пусть дан запрос  $Q$ , содержащий слова  $q_1, \dots, q_n$ . тогда функция BM25 даёт следующую оценку релевантности документа  $D$  запросу  $Q$ :

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

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  $b = 0.75$ .

TF-IDF

# ПРИМЕР РАЗЛИЧИЯ ТЕКСТОВ ВАКАНСИЙ



Тексты бухгалтерских вакансий (1-1000)  
«ближе» по косинусной мере к данному  
тексту, чем вакансии разработчиков  
(тексты 1001-2000)

---

# МОДИФИКАЦИИ TF-IDF



# КЛАССИФИКАЦИЯ ЗАПРОСОВ

Имеется набор запросов поисковика. Необходимо отнести каждый поисковый запрос к одному из наших проектов или пометить его как undef

Уникальных фраз в месяц – 30 млн. штук

Проектов для классификации – 20 штук

# КЛАССИФИКАЦИЯ ЗАПРОСОВ

1. В качестве обучающей выборки из Яндекс.Метрики берем ключевые слова, по которым переходили на проект. Количество переходов для каждой фразы тоже записываем, это наш аналог частоты
2. Считаем модифицированную метрику TF-IDF для каждого слова в запросе, беря вместо частоты количество переходов из пункта 1
3. Относим поисковый запрос к проекту с максимальным значением TF-IDF

---

# НЕМНОГО О ВЕКТОРАХ WORD2VEC

# ОПЕЧАТКИ И ПОДСКАЗКИ

Представление слов в виде векторов позволяет оценить их близость

## “нятвуч”

яндех 0.733752369881  
янекс 0.711848974228  
янлекс 0.698952019215  
гмайл 0.694112420082  
яндес 0.676702141762  
майлру 0.671700835228  
яедекс 0.665594637394  
янднкс 0.649016737938  
меил 0.644513964653  
яндек 0.640007019043

## “асаd + сирия”

игил 0.720687747002  
ливии 0.673956394196  
сирии 0.641393125057  
воюет 0.627352774143  
мусульмане 0.621693730354  
асада 0.616143345833  
ирак 0.597613811493  
сирию 0.585588812828  
ввела 0.583445549011  
госдолг 0.579698681831

# СЕМАНТИЧЕСКИ БЛИЗКИЕ СЛОВА

## “замок+дверь”

заклинила 0.5750108361  
дверной 0.542515099049  
замка 0.541278779507  
металлическую 0.533510  
личинку 0.525701284409  
заднюю 0.521266341209  
багажника 0.5204059481  
врезать 0.516109466553  
врезной 0.508815765381  
обшивку 0.508574485779

## “замок + ваз”

инжектор 0.655442953  
2110 0.639303267002  
2107 0.638078570366  
2115 0.629198908806  
2114 0.612880945206  
2106 0.608877778053  
замка 0.605489730835  
ручника 0.5999284982  
2109 0.595452368259  
стеклоподъемник 0.59

## “замок + швейцария”

швеция 0.452736794  
финляндия 0.451543  
германия 0.4477606  
австрия 0.43222194  
шв 0.431186437607  
щвеция 0.428391844  
лихте 0.4273314476  
футбольный 0.41752  
лихтенштейн 0.41578  
бронза 0.411839365

# АВТОМАТИЧЕСКИЙ ПОИСК СЮЖЕТОВ

Выборка текстов новостей

## спасение летчика

cluster 8

парашют

охотиться

приземление

спецназ

ведение

вблизи

кратковременный

пытаться

## гены и допинг

cluster 27

препарат

здоровье

среднее

организм

пациент

эффект

вирус

сопровождать

## санкции Турции и IT

cluster 48

постановление

ограничение

ответный

запрет

продление

расширение

договор

оборонный

предусматривать

---

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

СХОЖЕСТЬ ПОЛЬЗОВАТЕЛЕЙ

# 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	★★★★★	★★★★★

СРАВНЕНИЕ ТЕКСТОВ

---

СНОВА ПРИМЕР

JACCARD.ІРҮНВ

# ПРО ПРОИЗВОЛЬНЫЕ РАССТОЯНИЯ

KNN NBA.IPYNB

---

# KNN И РЕГРЕССИЯ



СРАВНЕНИЕ ТЕКСТОВ

---

ПРОСТО ПОСМОТРИМ КОД

KNN REGRESSION.IPYNB

---

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