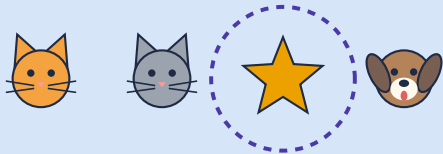
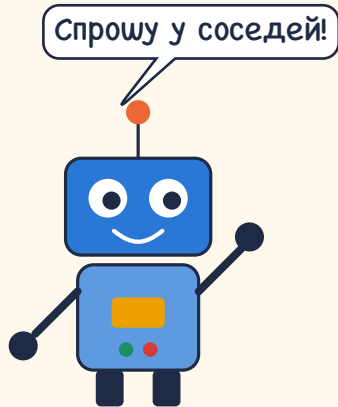
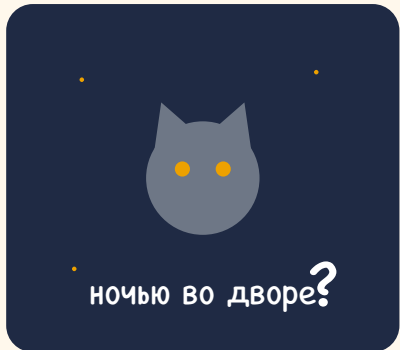


КТО ТВОИ СОСЕДИ?

Алгоритм **К ближайших соседей** (KNN)
— простыми словами



Кто это прибежал?



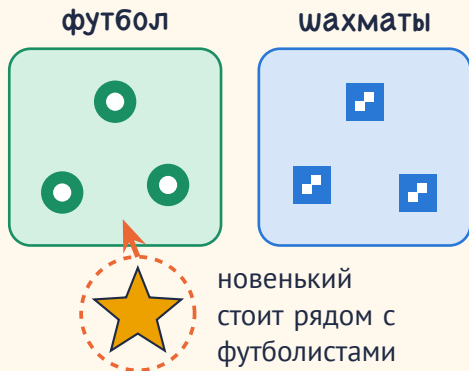
Ночью во двор прибежал зверёк. Бип видит только:

- он **небольшой**,
- у него **длинные усы**.

Котик это или собачка?

Бип вспоминает зверей, которые больше всего похожи на этого. Если почти все похожие — котики, то и это, скорее всего, котик!

Скажи мне, кто твой сосед...



В класс пришёл новенький. Он встал рядом с ребятами из **футбольной** секции.

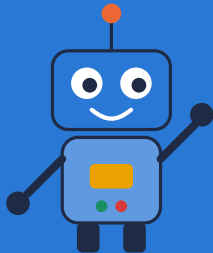
Скорее всего, ему тоже нравится футбол!

Идея KNN: похожие вещи лежат **рядом**. Чтобы узнать, кто ты, посмотри на своих **ближайших соседей**.

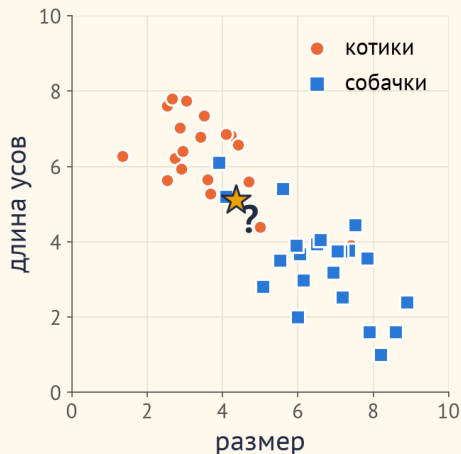
KNN — K-Nearest Neighbors, «К ближайших соседей».

Часть 1

Как работает KNN



Каждый зверь — точка на карте



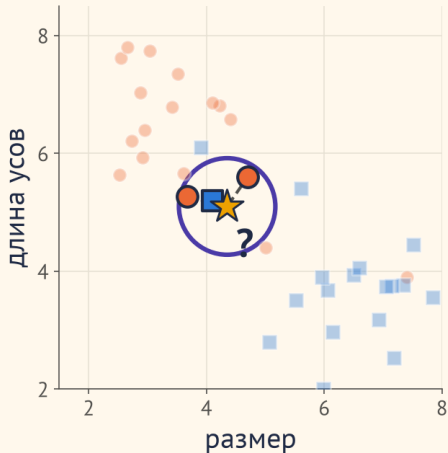
- Для каждого зверя Бип записал два числа: **размер** и **длину усов**. Это его место на карте.
- **Котики** собрались слева вверху: маленькие, усатые.
- **Собачки** — справа внизу: крупные, усы короткие.
- **Звёздочка** — наш ночной гость.

Алгоритм в четыре шага



Никаких сложных правил — только
«кто рядом?» и честное голосование

Голосуем! Спросим 3 соседей



Берём $K = 3$. Рисуем круг вокруг звёздочки, пока внутрь не попадут 3 зверя.



2 : 1

Котиков больше. Ответ: КОТИК!

Ближе всех стоит маленькая собачка (чихуахуа), но двое других соседей — котики.

Бип не учит правила — он помнит альбом

Обычный робот

Сначала долго учится и придумывает **правило**, например линию между котиками и собачками.
Потом отвечает быстро.

✓ «Выучиться» можно мгновенно

Робот-KNN

Ничего не придумывает заранее: просто **хранит весь альбом**.

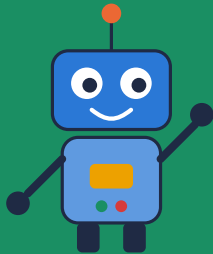
На каждый вопрос заново листает альбом и ищет соседей.

✗ Если зверей миллион — отвечать долго

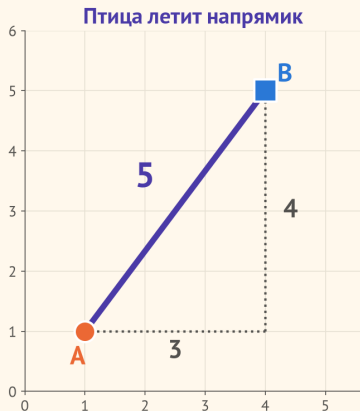
Учёные называют это «ленивым обучением» (lazy learning).

часть 2

Как измерить
расстояние?

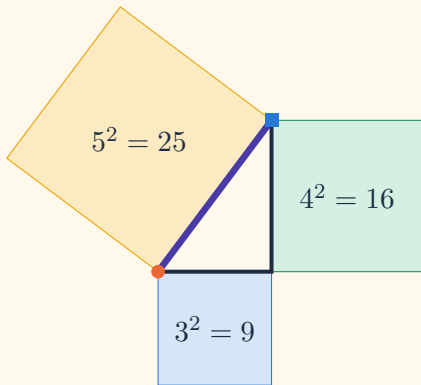


Два способа дойти от А до В



Птица летит по прямой — **5 шагов**. Такси объезжает дома — **7 шагов**.
Какое расстояние «правильное»? Оба! Просто они меряют по-разному.

Евклидово расстояние: как летит птица



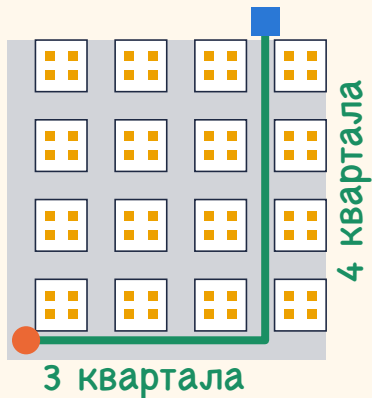
Это теорема **Пифагора**! Два маленьких квадрата вместе равны большому: $9 + 16 = 25$.

Правило птицы:

1. Найди разницу по каждой стороне: 3 и 4.
2. Возведи в квадрат и сложи: $9 + 16 = 25$.
3. Извлеки корень: $\sqrt{25} = 5$.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Манхэттенское расстояние: такси



Манхэттен — район Нью-Йорка, где улицы идут ровными клеточками. Такси не может проехать сквозь дома!

Правило такси:

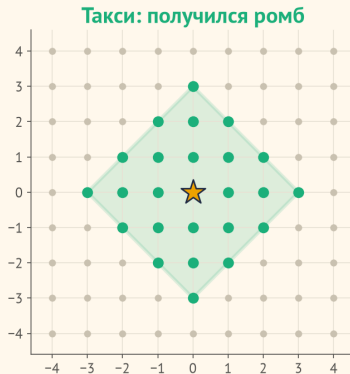
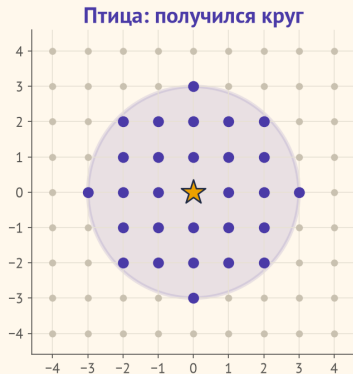
Просто сложи шаги по каждой стороне:

$$3 + 4 = 7$$

$$d = |x_1 - x_2| + |y_1 - y_2|$$

Палочки $|$ значат «без минуса»: $|-3| = 3$. Шаги назад тоже считаются шагами.

Кто «рядом» со звёздочкой?



Закрасили все точки, до которых **не больше 3 шагов**.

Для птицы «близко» — это круг, для такси — ромб. Поэтому соседи иногда разные!

Осторожно: разные линейки!

Пусть Бип записал **вес в граммах**, а **усы в сантиметрах**:

	вес, г	усы, см
Котик Мурзик	4000	7
Котик Барсик	4500	7
Собачка Тузик	4100	2

Мурзик — Барсик: разница 500 г $\Rightarrow$ **500**

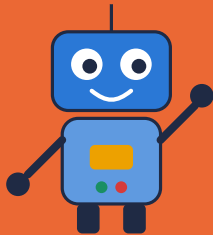
Мурзик — Тузик: 100 г и 5 см $\Rightarrow$ **около 100**

✗ Ближайший сосед котика — **собачка!** Граммы «кричат» громче, а усы никто не слышит.

✓ **Решение:** перевести все признаки в одинаковую шкалу, например от 0 до 10. Это называется **масштабированием**.

Часть 3

Сколько соседей
спросить?



Ответ зависит от К

соседи
по порядку



1-й



2-й



3-й



4-й



5-й



6-й



7-й

котики : собачки

К = 1



0 : 1 $\Rightarrow$ **собачка**

К = 3



2 : 1 $\Rightarrow$ **КОТИК**

К = 5



4 : 1 $\Rightarrow$ **КОТИК**

К = 7

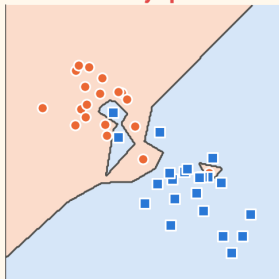


4 : 3 $\Rightarrow$ **КОТИК**

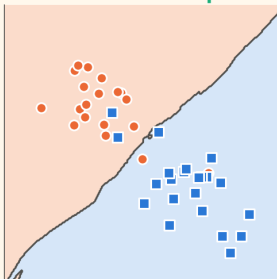
С **К = 1** Бип слушает только самого близкого — маленькую собачку. Спросив больше соседей, он передумал.

Карта ответов Бипа

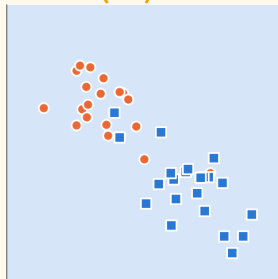
K = 1: зубрила



K = 7: в самый раз



K = 41 (все): ленивый



Цвет фона — что ответит Бип, если звёздочка окажется в этом месте.

Маленький K

верит каждому, даже
ошибке: островки и
зигзаги

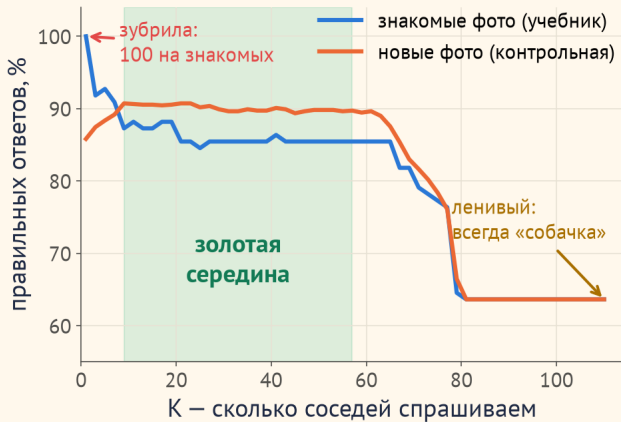
Средний K

ровная граница, случайные
звери не мешают

Огромный K

спрашивает всех —
собачек больше, значит,
всегда «собачка»

Ищем золотую середину



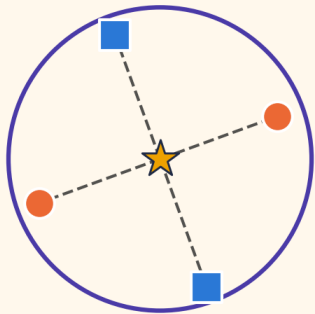
K = 1 — зубрила. На знакомых 100 %: каждый зверь — сам себе сосед. На новых — хуже.

K = все — ленивый. Всегда «собачка»: собачек в альбоме больше.

В середине — лучше всего на новых зверях. Такой K и выбираем.

K подбирают на **тренировочной** кучке фото — как мы делали на прошлом уроке.

Почему К берут нечётным?



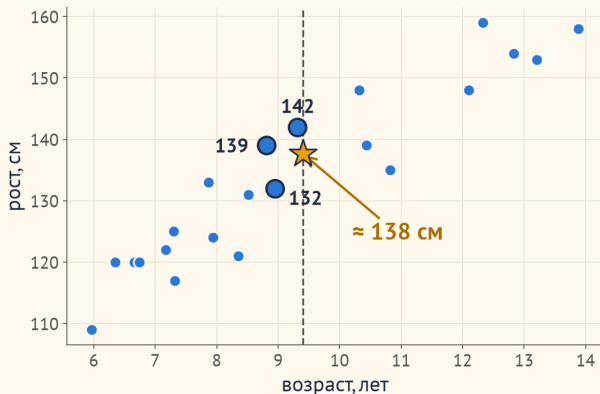
2 : 2 — ничья!

При $K = 4$ может выйти ничья: 2 котика и 2 собачки. Кто победил? Непонятно!

Когда выбираем из **двух** ответов, берём нечётное K : 3, 5, 7... Тогда ничьей не бывает.

Ещё одна хитрость: дать ближним соседям больше голосов, чем дальним.

Соседи угадывают и числа



Какой рост у мальчика, которому **9 лет и 5 месяцев**?

Найдём 3 ребят ближе всего по возрасту и возьмём их **средний** рост:

$$\frac{142 + 132 + 139}{3} \approx 138 \text{ см}$$

Для «котик или собачка» соседи голосуют, а для чисел — считаем среднее.

Чем хорош и чем плох KNN

✓ Хорошо

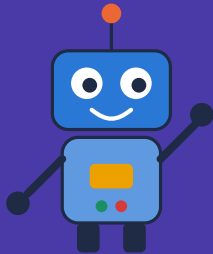
- очень простая идея — можно объяснить другу
- ничего не надо учить заранее
- новый зверь в альбоме — сразу помогает
- умеет и голосовать, и считать числа

✗ Плохо

- медленно, если альбом огромный
- нужно хранить все примеры
- обязательно масштабировать признаки
- надо подобрать хорошее K

Часть 4

Повторим и проверим
себя



Шпаргалка



Слово	Детское объяснение
KNN	Спроси К соседей и сделай как большинство
К	Сколько соседей спрашиваем
Евклидово	Как летит птица: $\sqrt{3^2 + 4^2} = 5$
Манхэттенское	Как едет такси по кварталам: $3 + 4 = 7$
Масштабирование	Все признаки — на одинаковой линейке
Маленький К	Зубрила: верит каждой случайности
Большой К	Ленивый: всегда отвечает одно и то же

Проверь себя!



1. Точки $A(1; 2)$ и $B(4; 6)$. Найди **евклидово** и **манхэттенское** расстояния.
2. Соседи по порядку: котик, собачка, котик, собачка, котик. Что ответит Бип при $K = 1$, $K = 2$ и $K = 5$?
3. Бип с $K = 1$ угадал 100 % знакомых зверей. Значит, он отличный робот?
4. Рост записали в миллиметрах, а длину хвоста — в метрах. Что пойдёт не так?

Ответы



1. Разницы 3 и 4. Евклидово: $\sqrt{9 + 16} = 5$. Манхэттенское: $3 + 4 = 7$.
2. $K = 1$: **КОТИК**. $K = 2$: **НИЧЬЯ** $1 : 1$. $K = 5$: **КОТИК** $(3 : 2)$.
3. Нет! Каждый знакомый зверь — сам себе ближайший сосед. Проверять надо на **НОВЫХ** зверях.
4. Миллиметры — большие числа, они «перекричат» хвост. Нужно **масштабирование**.

Спасибо!

Вопросы?

