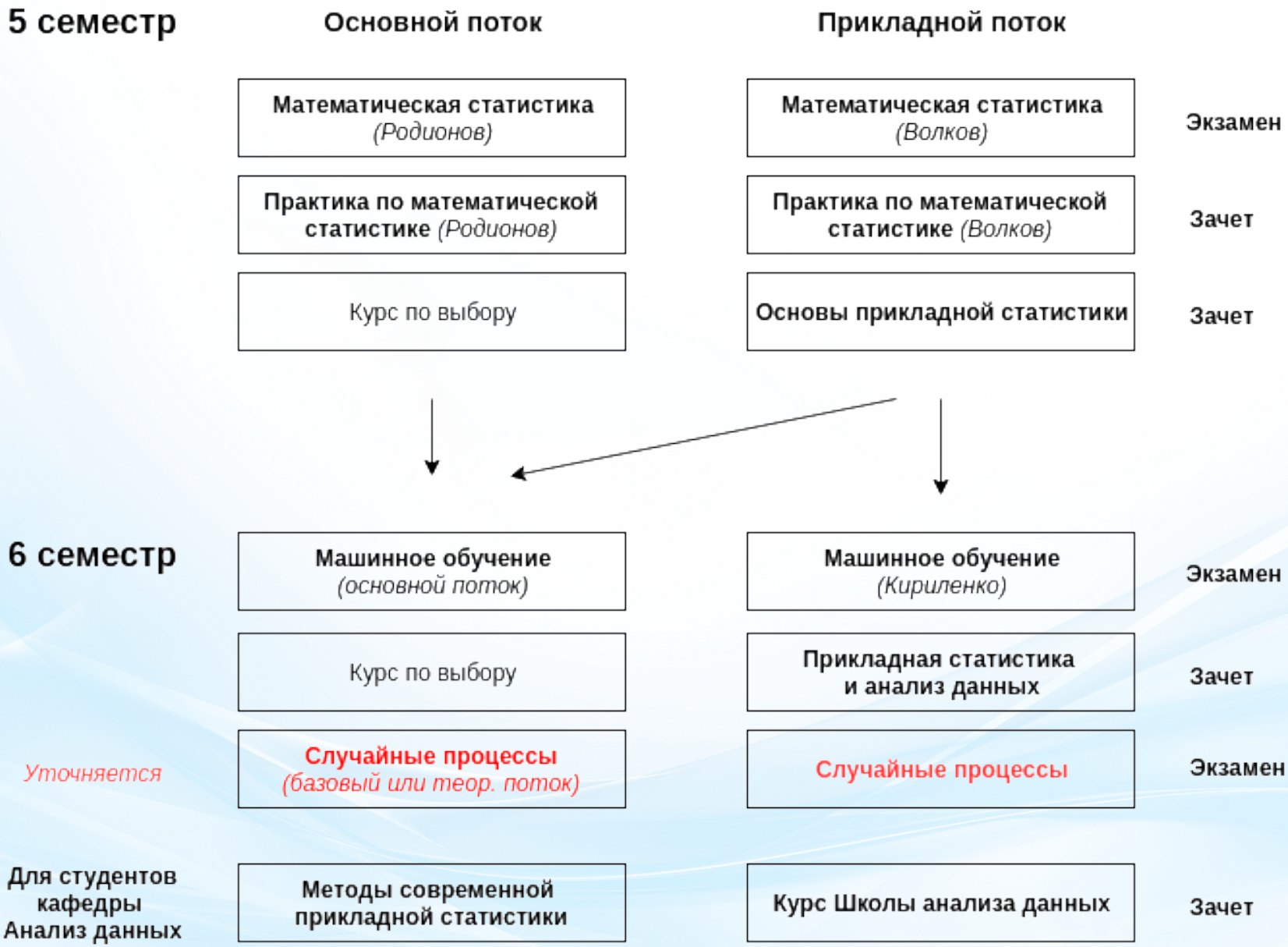
Статистика ФИВТ ПМИ

Прикладной поток


5 семестр
Основной поток
Прикладной поток
Математическая статистика (Родионов)
Математическая статистика (Волков)
Экзамен
Практика по математической статистике (Родионов)
Практика по математической статистике (Волков)
Зачет
Курс по выбору
Основы прикладной статистики
Зачет
6 семестр
Машинное обучение (основной поток)
Машинное обучение (Кириленко)
Экзамен
Курс по выбору
Прикладная статистика и анализ данных
Зачет
Уточняется
Случайные процессы (базовый или теор. поток)
Случайные процессы
Экзамен
Для студентов кафедры Анализ данных
Методы современной прикладной статистики
Курс Школы анализа данных
Зачет

О прикладном потоке

Лекции

Статистика

Никита Волков

Машинное обучение

Елена Кириленко

Семинары

Ольга Калиниченко

Роман Логинов

Дмитрий Лунин

Практические занятия

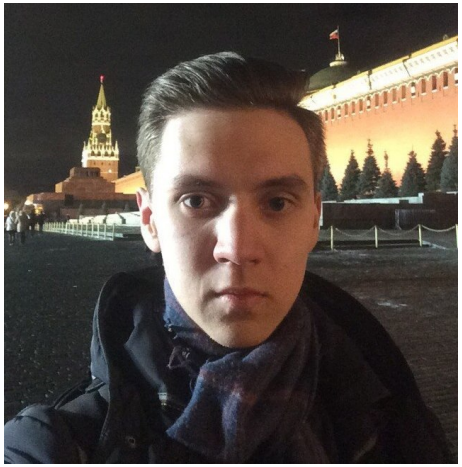
Анастасия Грачева

Елизавета Дахова

София Ожерельева

Другие члены команды

Евгений Иванин



Вячеслав Иванов

Егор Колодин

Мария Кривошапко

Артем Куприянов

Михаил Лепехин

Организационная информация

Сайт курса mipt-stats.gitlab.io

Почта mipt.stats@yandex.ru

Тема письма: "[S19] Имя Фамилия - задание 1"

Лекция — понедельник 17:50-20:00, Арктика, поточная аудитория

Семинары

- Дмитрий — четверг 11:30-13:45, 302 КПМ;
- Ольга — пятница 10:00-12:20, Цифра 516;
- Роман— пятница 13:55-16:10, 412 ГК.

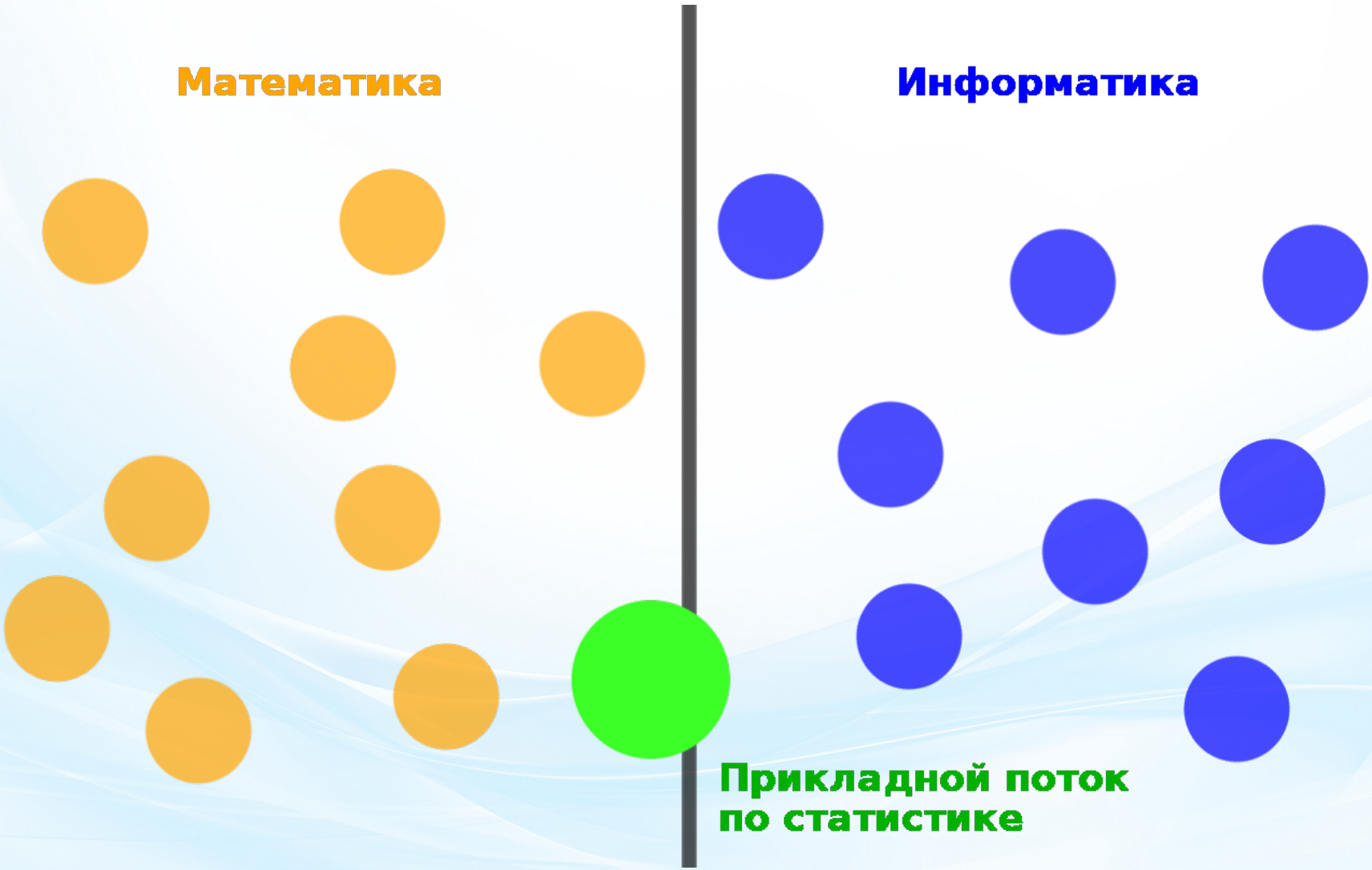
Практические занятия — информация появится на сайте

На первой недели семинаров и практических занятий нет.

Отбор на поток

- До 4 сентября 23:59 подать заявку на курс, ссылка на сайте;
- До 5 сентября 23:59 можно прислать вступительное задание;
- 6 сентября в первой половине дня будут результаты о том, кому необходимо пройти собеседование;
- Собеседования:
  6 сентября (вечер), 7 сентября (утро), 8 сентября (утро);
- 8 сентября (вечер) — итоговые результаты;
- 9-10 сентября — распределение по семинарским группам.

Курсы на ПМИ



Основные темы (осень)

1. Точечное оценивание;

   Свойства и методы поиска оценок; качество оценок

   и поиск наилучших оценок; оценки специального вида.

2. Доверительные интервалы;

3. Байесовский подход;

   Полный байесовский вывод, сопряженные распределения.

4. Непараметрический подход;

   Эмпирическое распределение; бутстреп; ядерная оценка плотности.

5. Проверка гипотез;

   Критерии; p-value; множественная проверка гипотез; критерии согласия.

6. Линейная регрессия.

   МНК; регуляризация; отбор признаков.

Основные темы (весна)

1. Регрессионный анализ;

   Анализ остатков; общая модель (в т.ч. логистическая регрессия).

2. Методы снижения размерности.

   PCA; t-SNE.

3. Корреляционный анализ;

   Коэффициенты корреляции; анализ категориальных признаков.

4. Дисперсионный анализ;

   Критерии для независимых и связных выборок; AB-тестирование; ANOVA.

5. Анализ причинности;

6. Детектирование аномалий;

   Методы скорейшего обнаружения разладок.

7. Последовательный анализ.

Реальная практика в курсе

1. Реальные примеры из практики;
2. Соревнования на Kaggle;
3. Гостевые лекторы, применяющие статистику на практике;
4. Разбор статей на тему анализа данных;
5. Бонусы за участие в хакатонах,
   соревнованиях по анализу данных и прочую активность;
6. Сотрудничество с компаниями, которые работают с data science,
   в том числе рекомендации на стажировку для лучших студентов.

Реальная практика в курсе

Спойлеры:

1. Поймем, почему на лабах по физике
   вам могли говорить неправду;
2. На примере схемы Бернулли рассмотрим случай, когда заказчик
   формулирует задачу не так, как ему хотелось бы на самом деле;
3. Разберем задачу, которую спрашивают
   на большинстве собеседований на аналитика;
4. Изучим, какие факторы влияют на объем времени,
   проведенный во внебрачных отношениях.

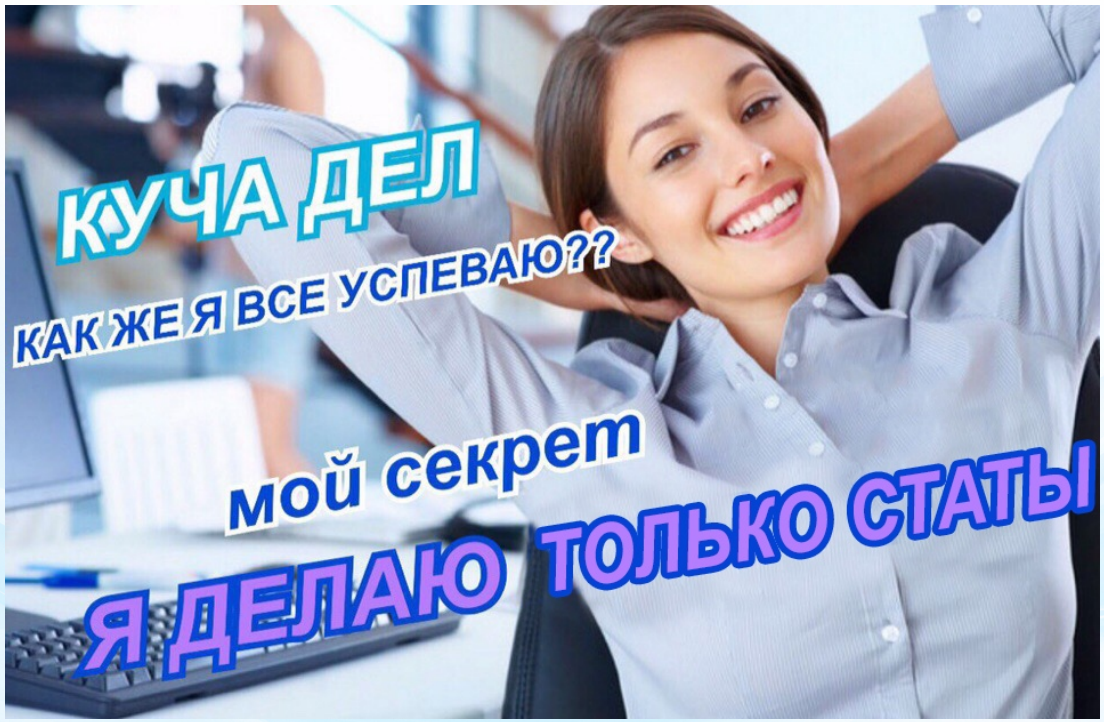
Основные книги по курсу

- Лагутин М.Б., Наглядная математическая статистика;
- L. Wasserman, All of Statistics;
- Russell B. Millar, Maximum Likelihood Estimation and Inference;
- Боровков А.А., Математическая статистика.

Неофициальный фан-клуб

Мемы про Матстаты для Маленьких Мальчиков

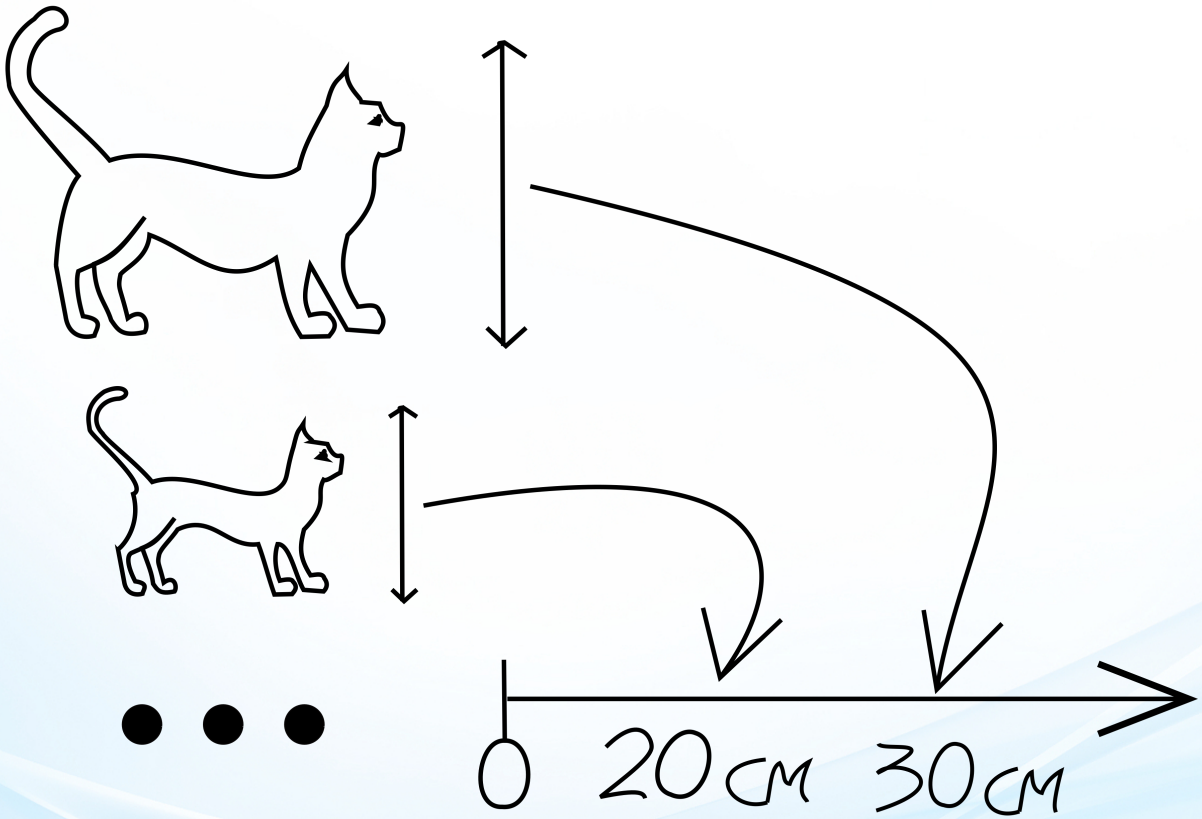
https://vk.com/meme_stats



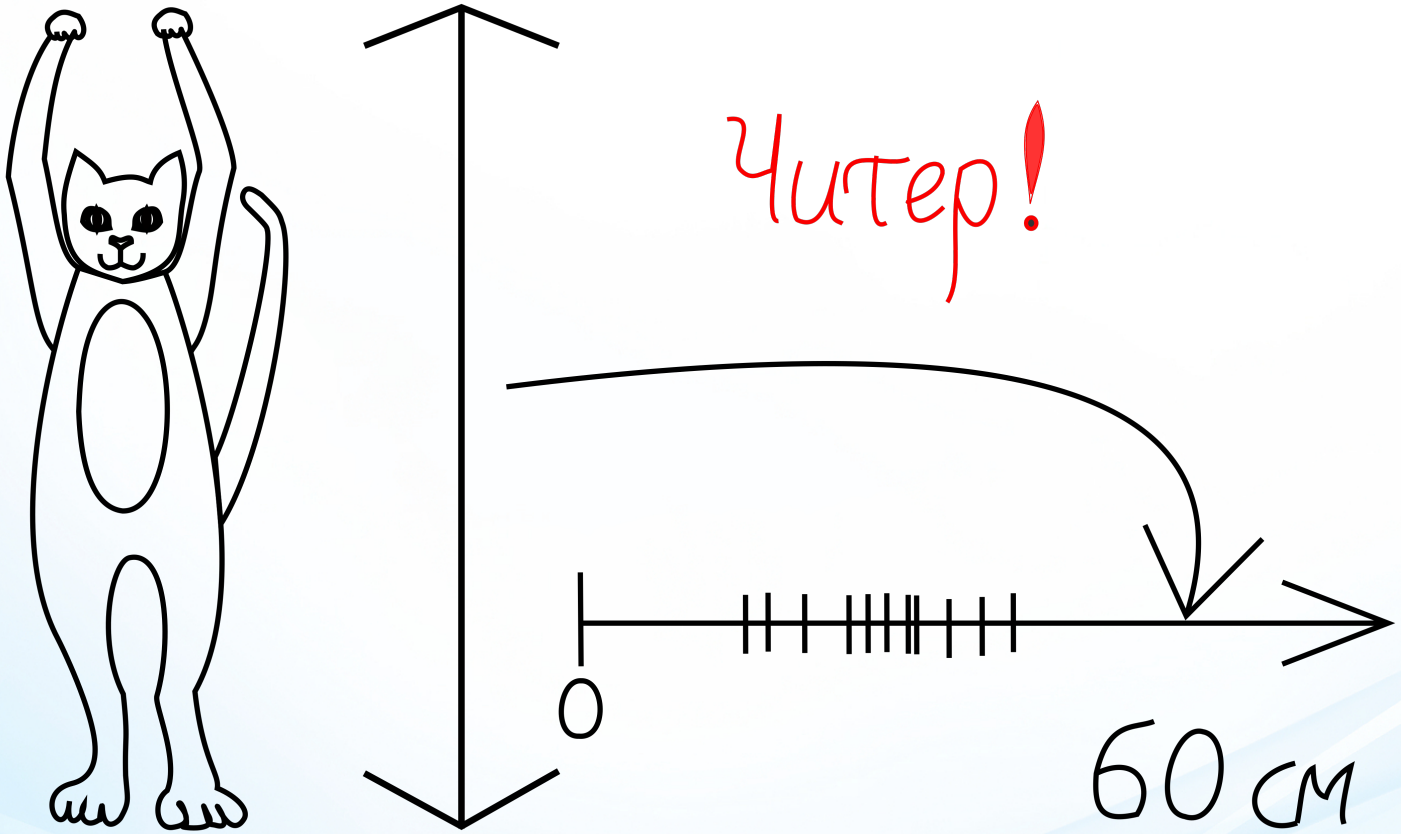
Обзор статистики

Мурмурландия

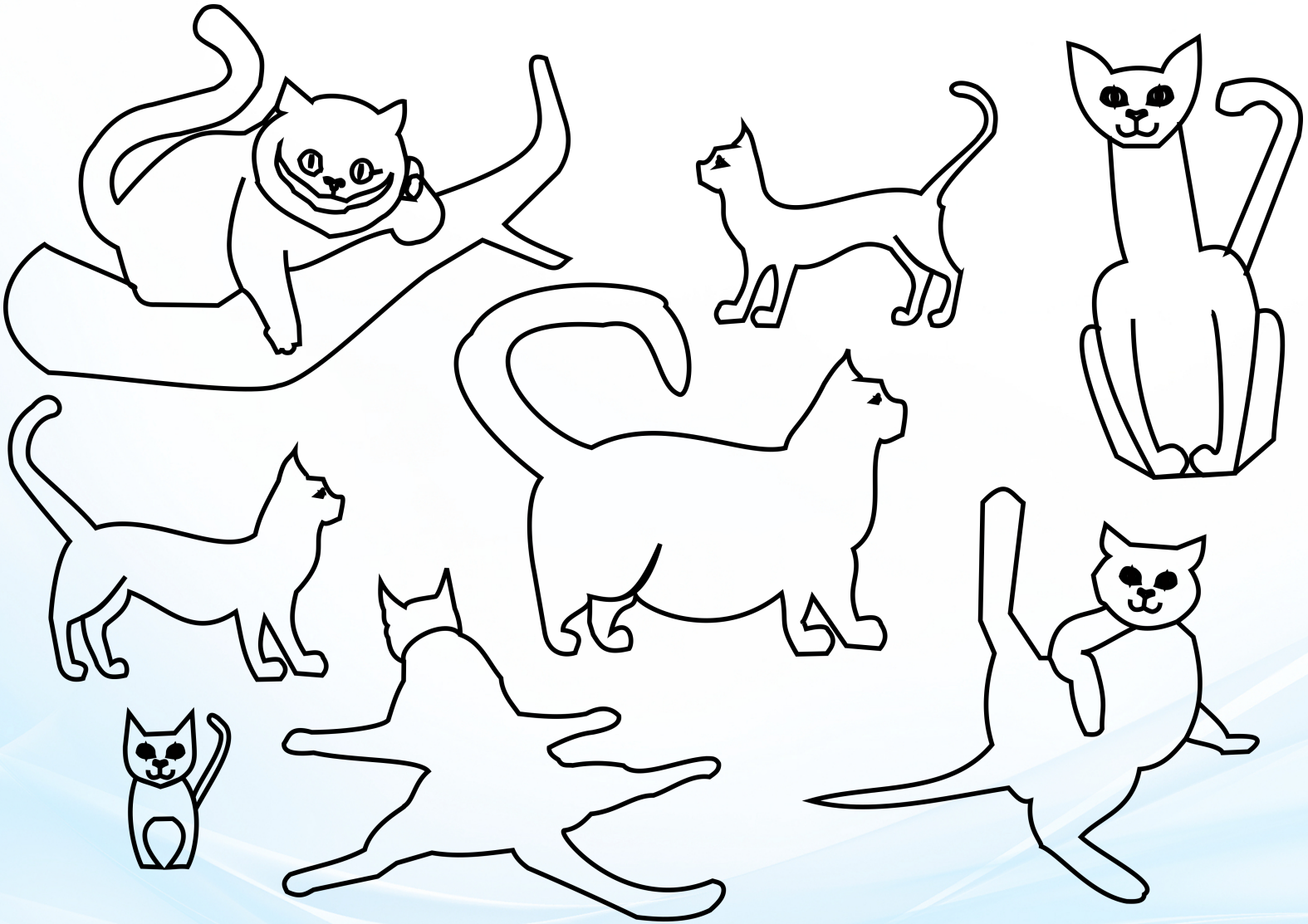
Каков средний рост котиков?



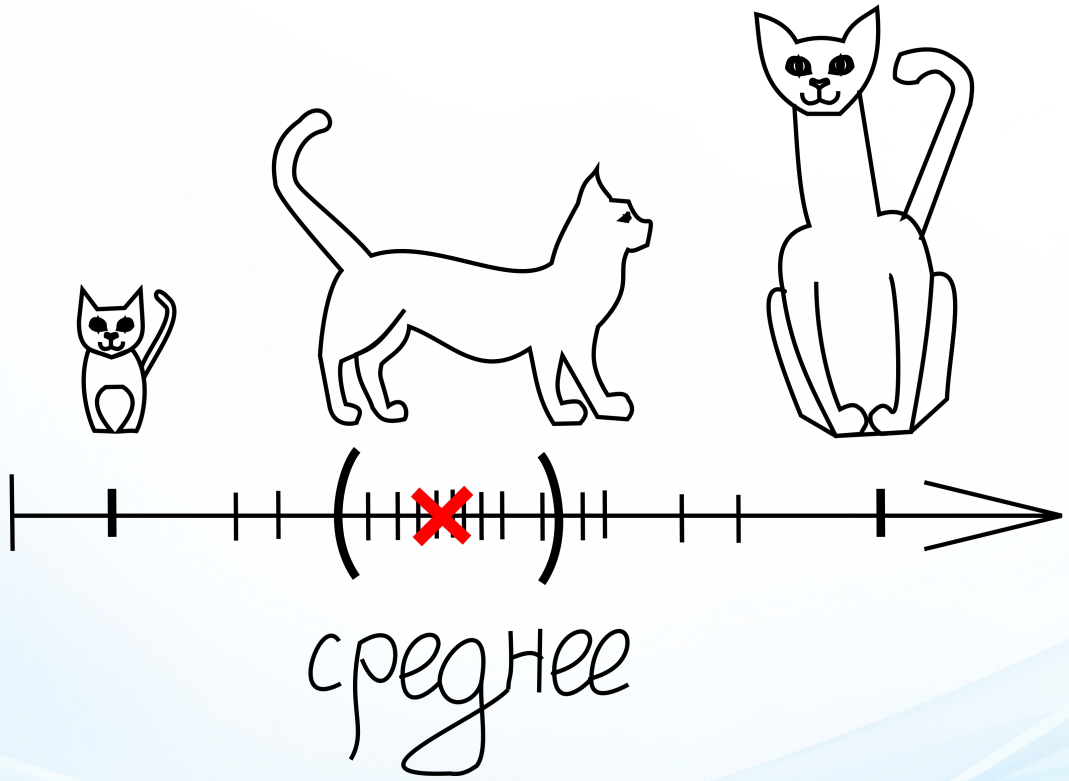
Точечное оценивание



Выбросы

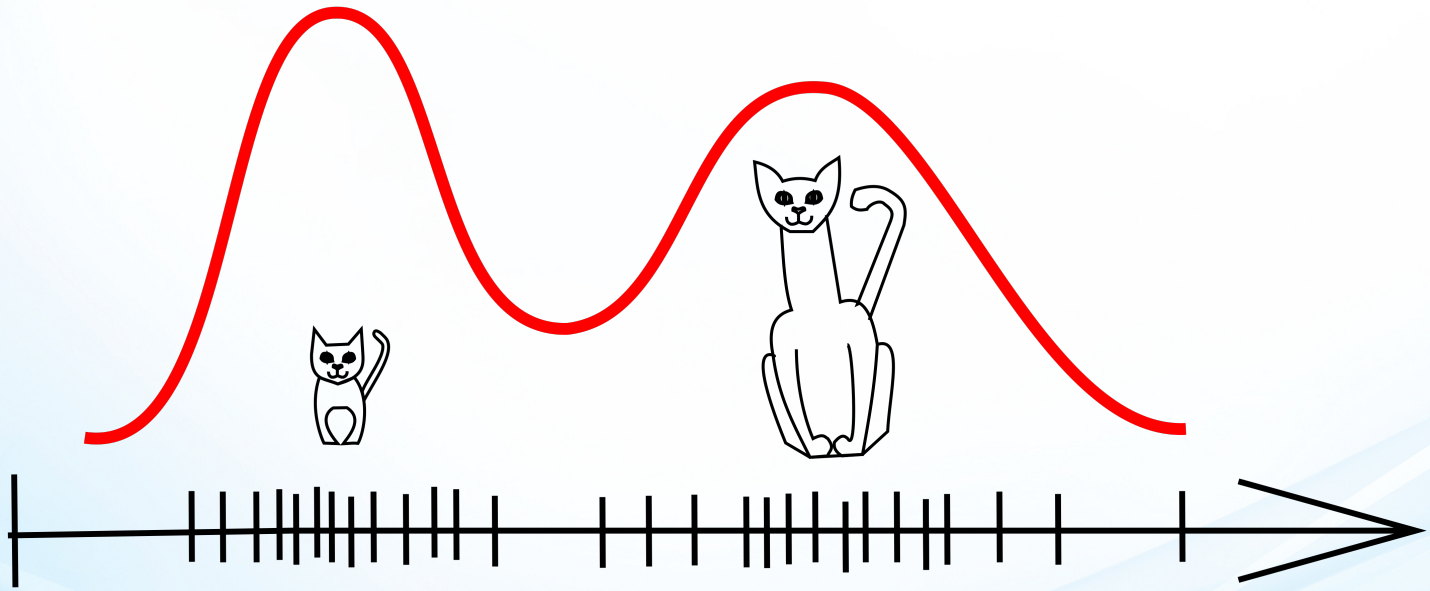


Среднее определяется неточно



Интервальное оценивание

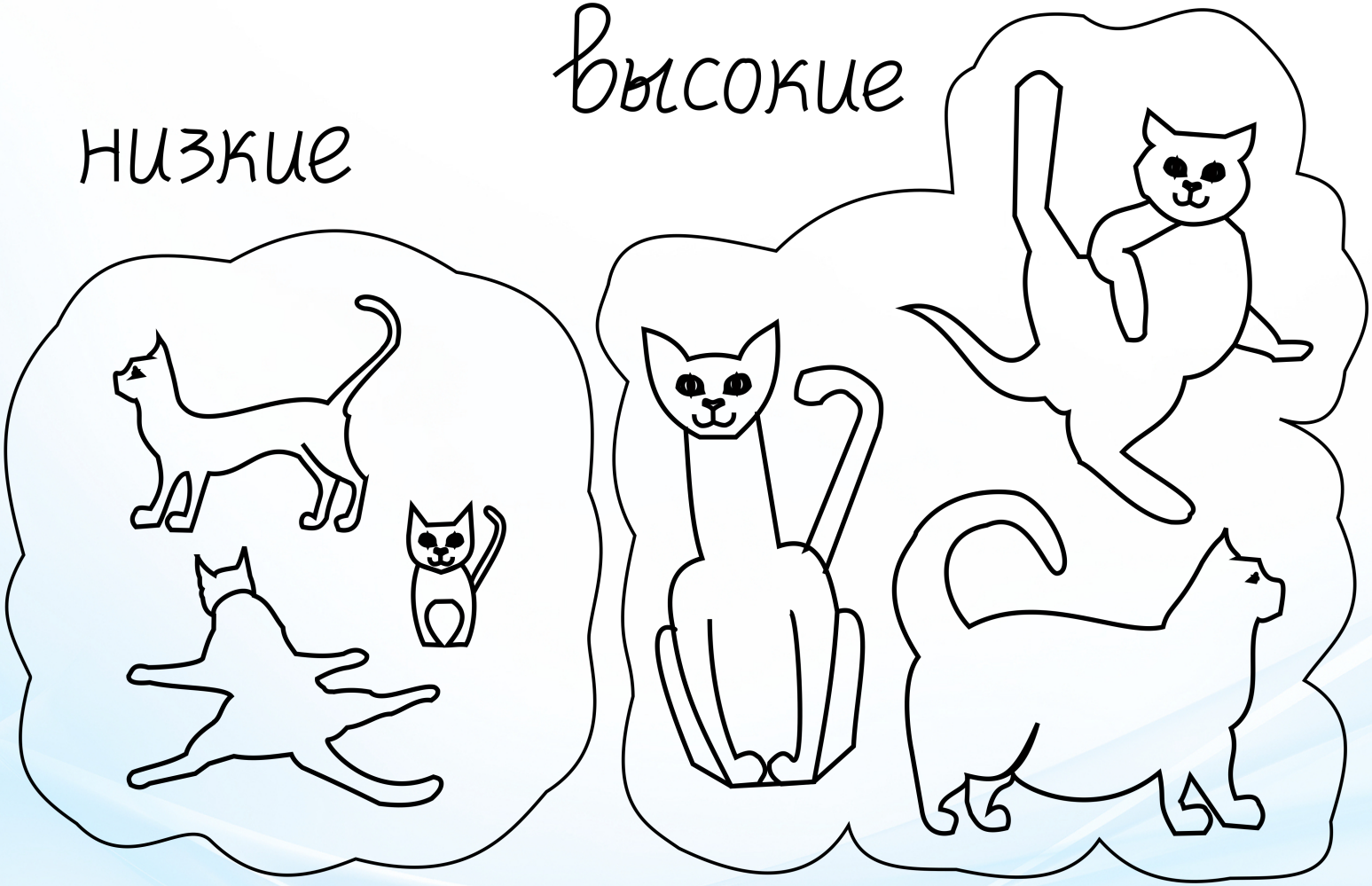
Характер распределения



Непараметрическое оценивание

низкие

высокие

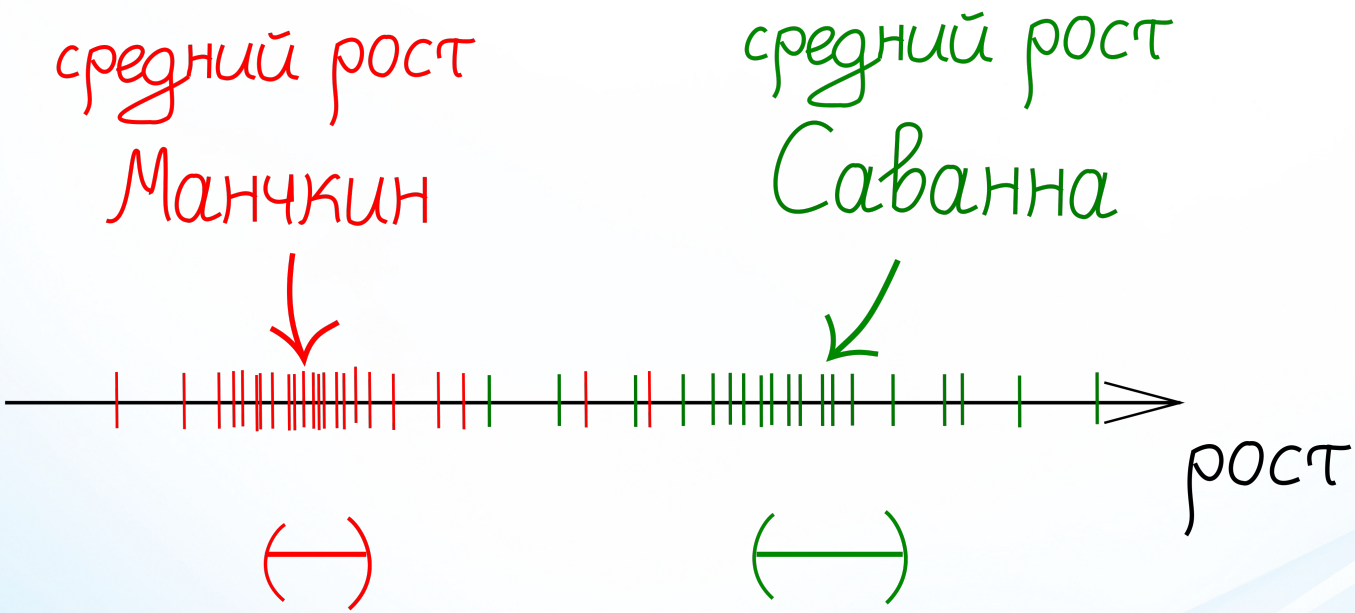


Саванна

Манчкин

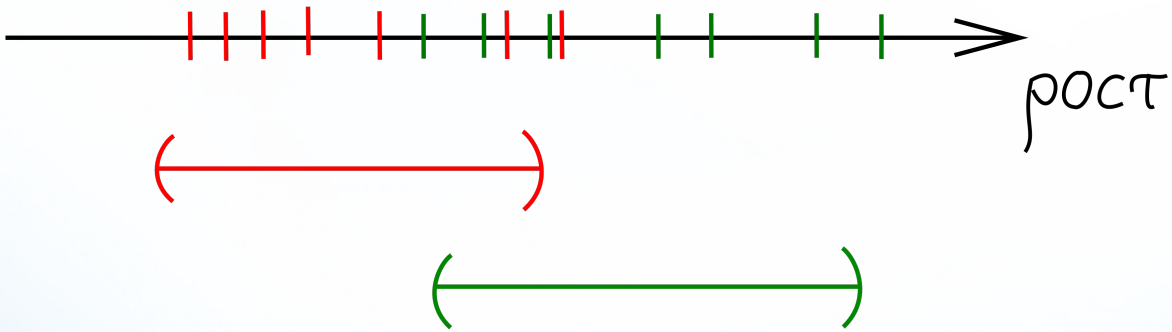
Отличается ли их средний рост?

Собираем данные

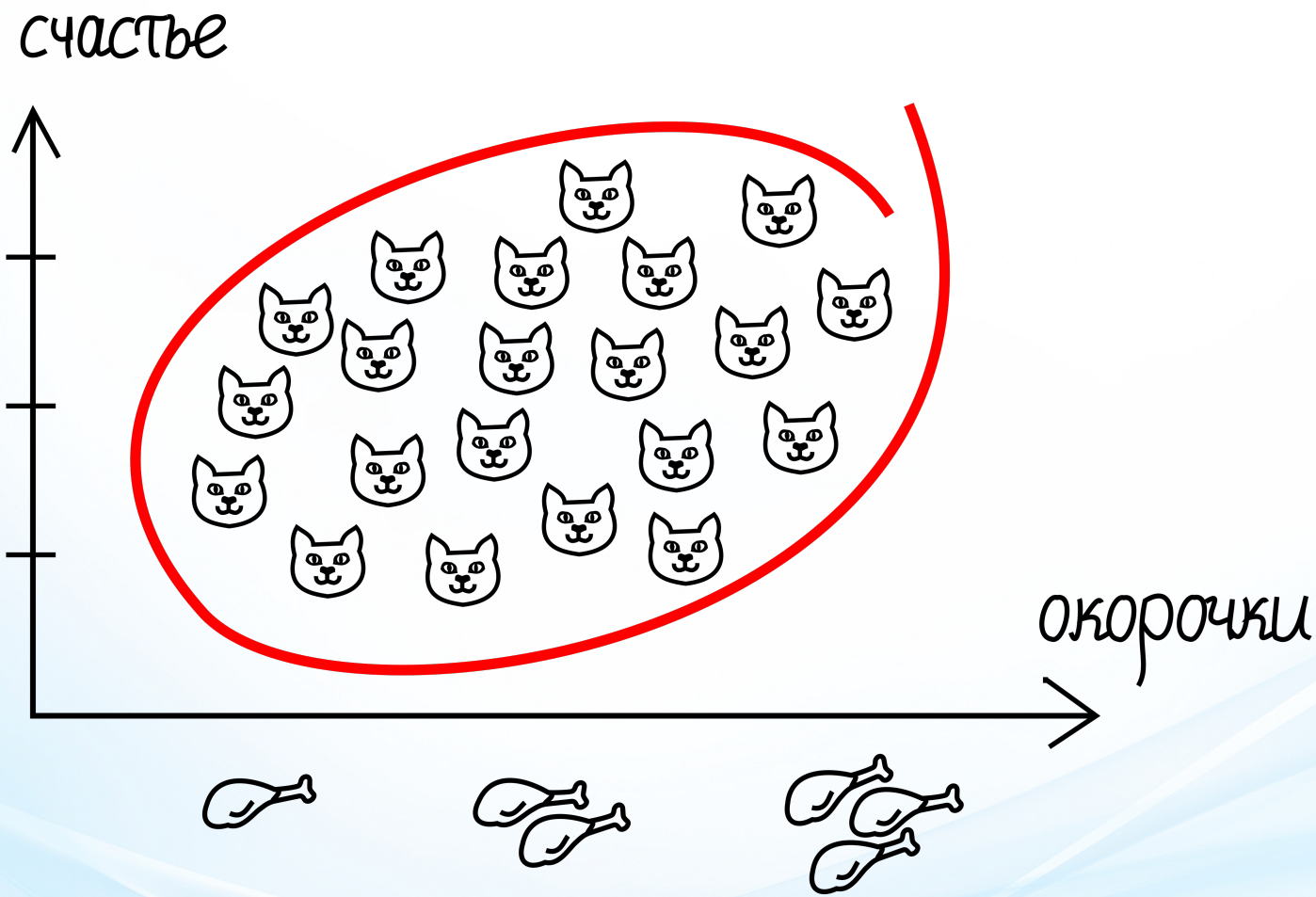


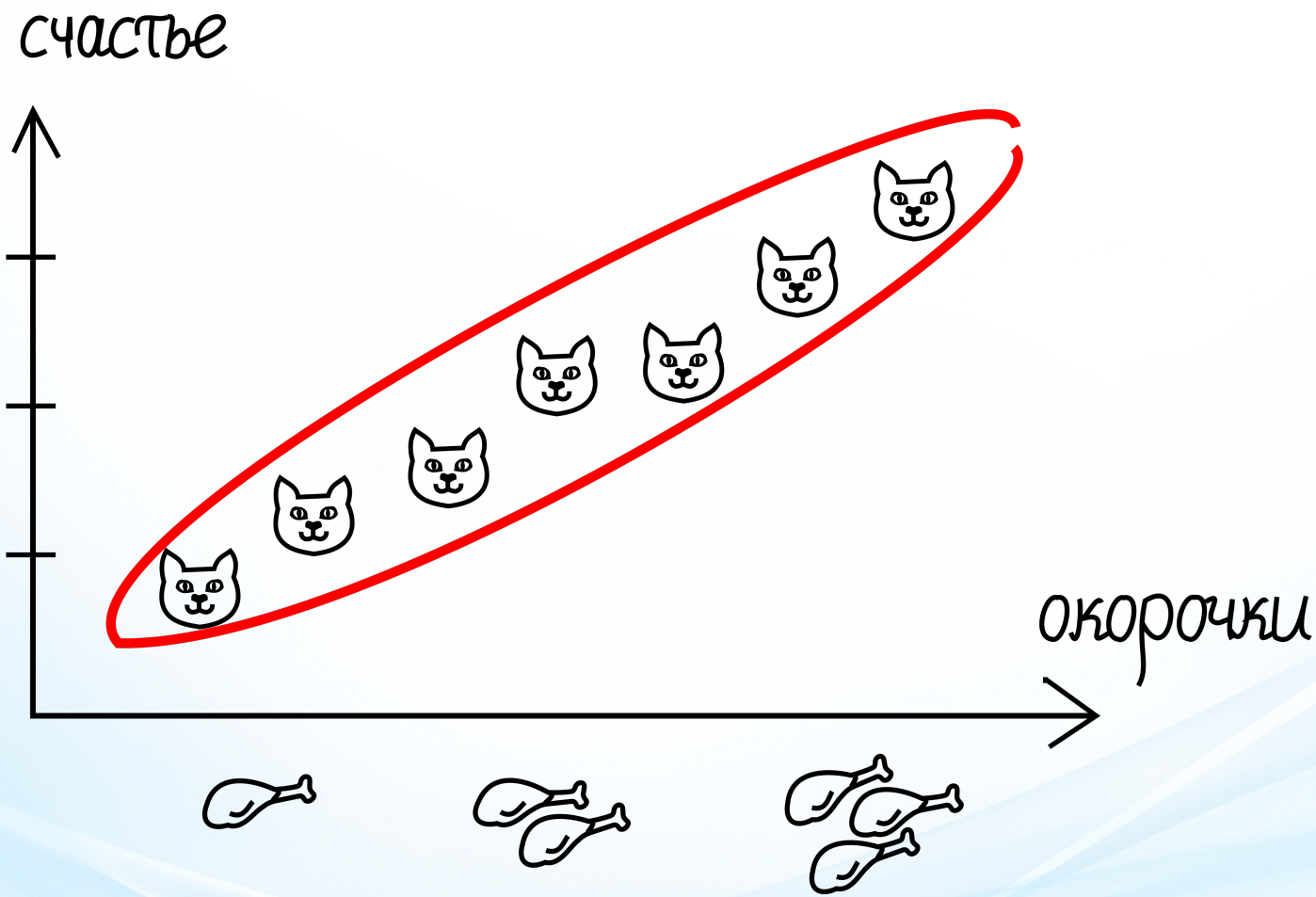
отличается

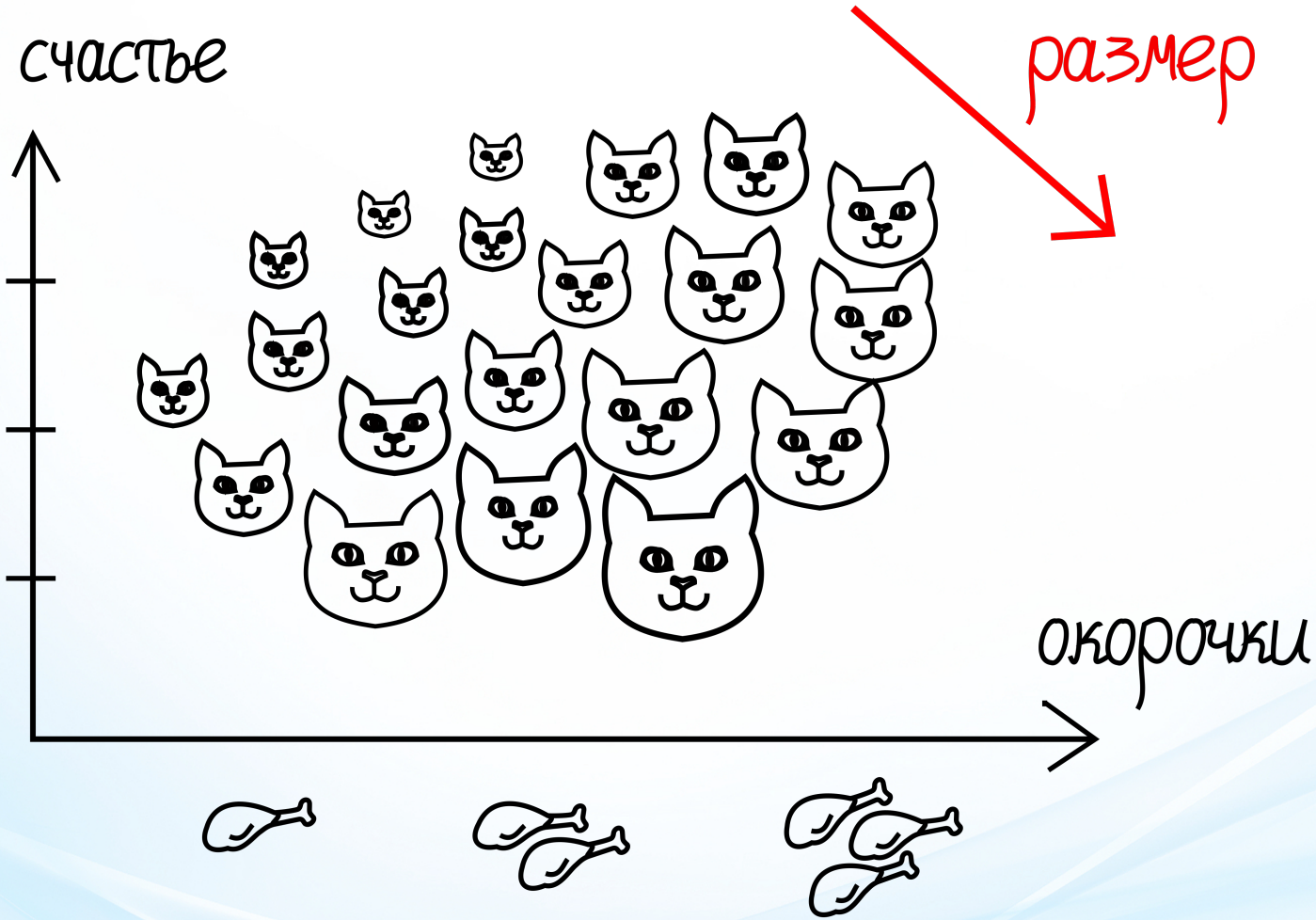
Если данных мало



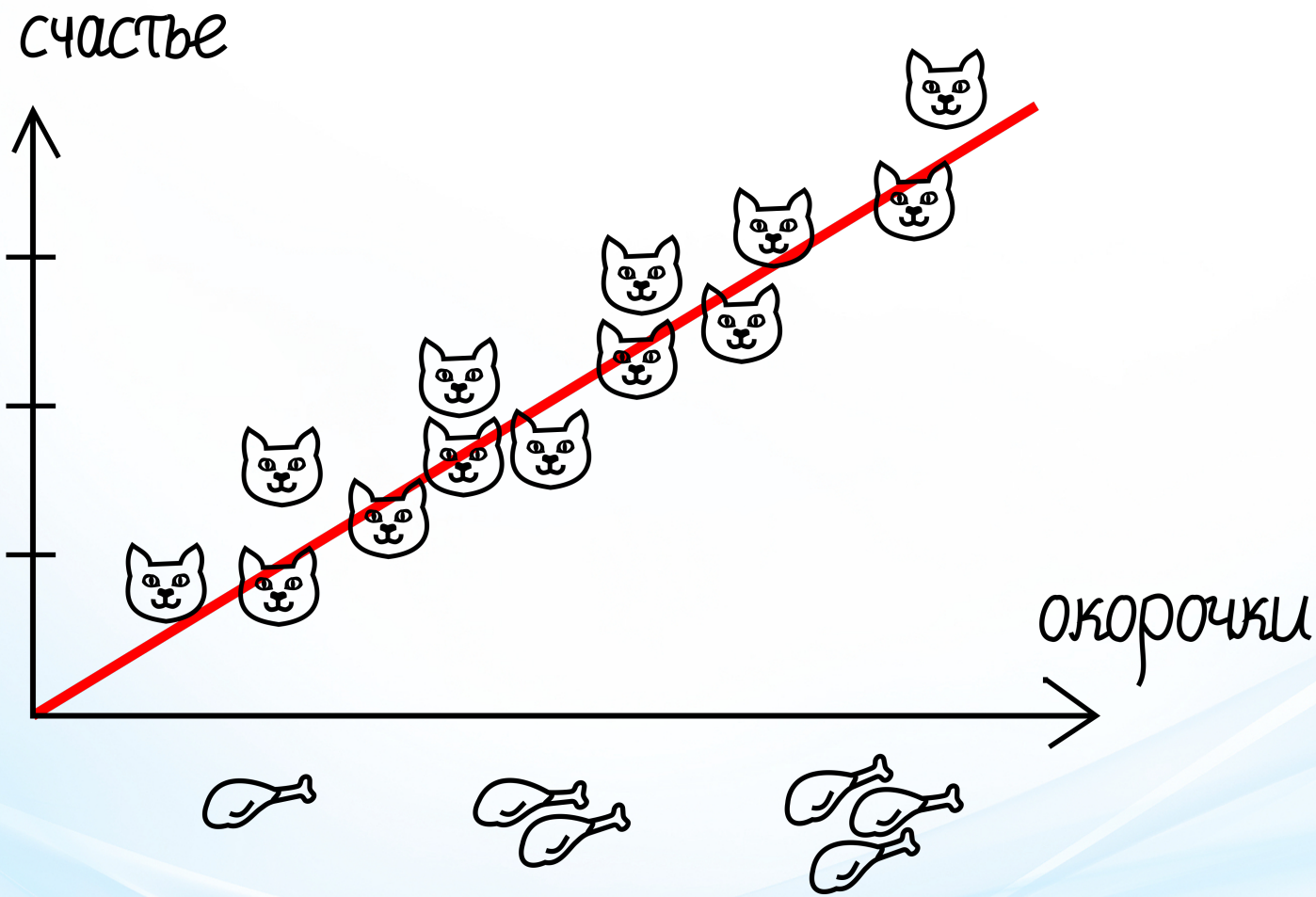
непонятно


счастье
окорочки

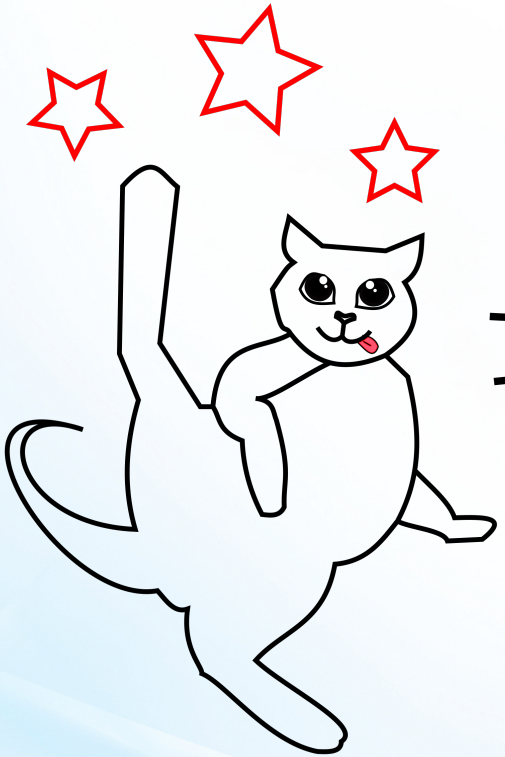

счастье
окорочки



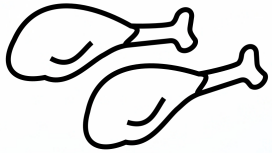
Корреляционный анализ


счастье
окорочки

Формула счастья

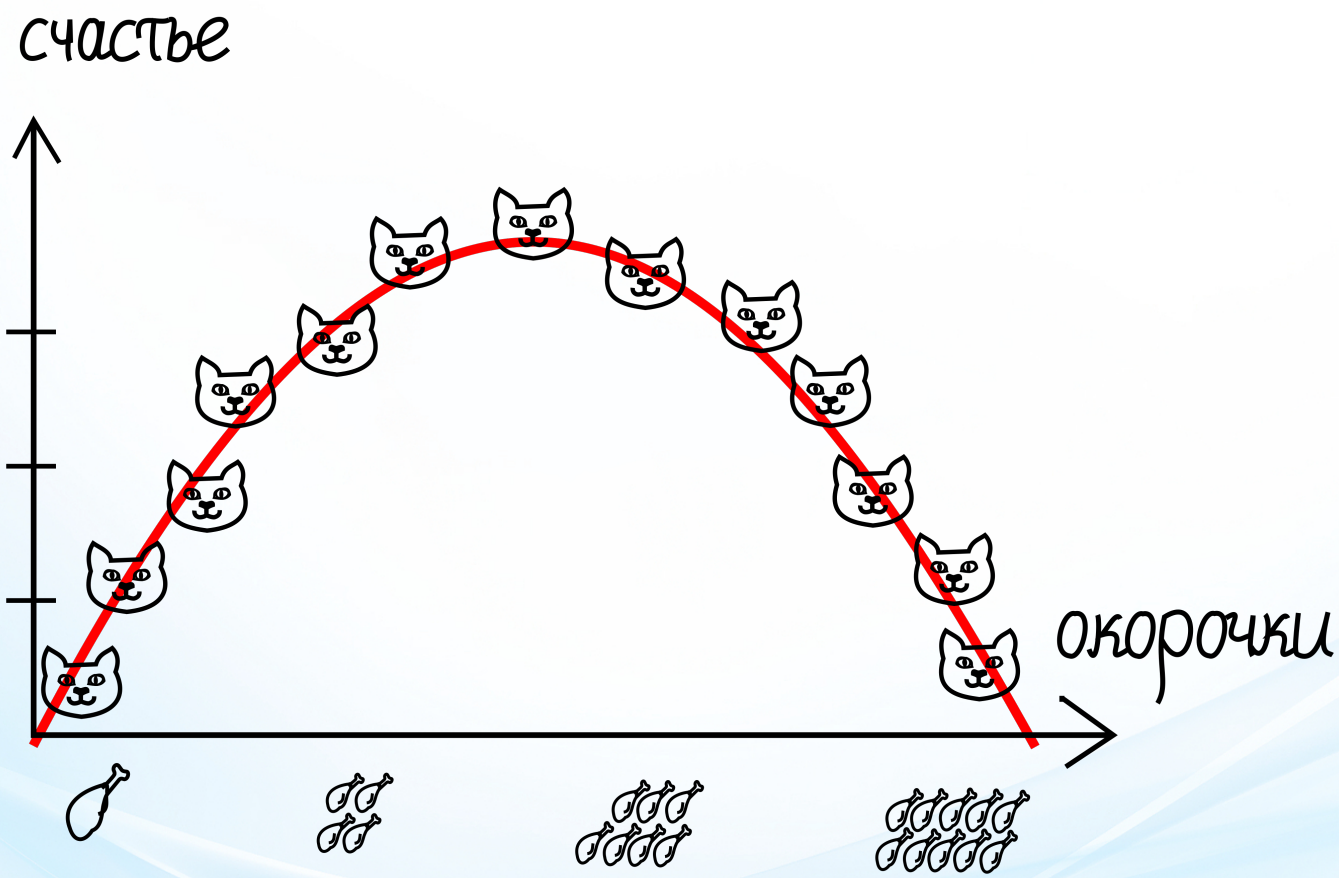


$$= \theta_0 + \theta_1 \times \text{кол-во}$$

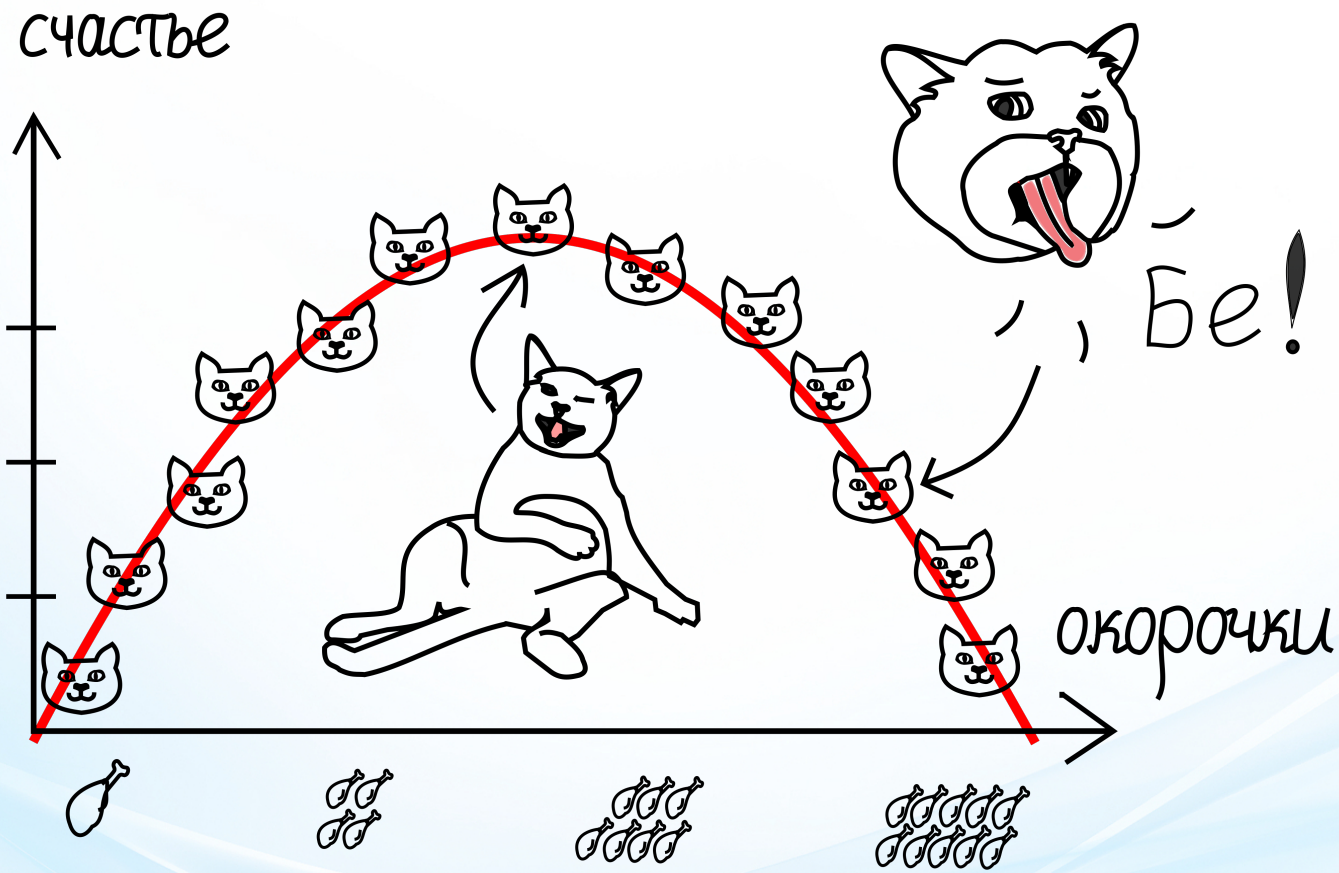


$$+ \text{погрешности}$$

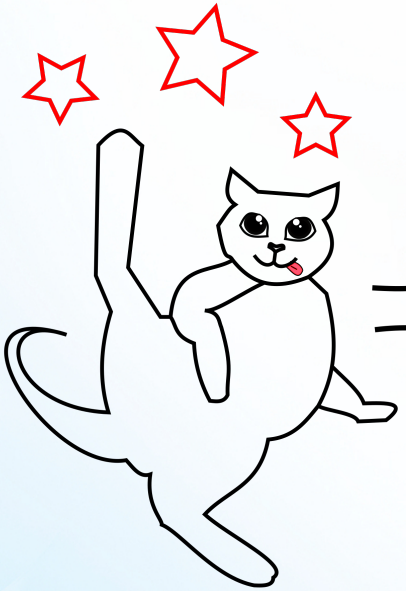
Больше окорочков



Больше окорочков



Формула счастья

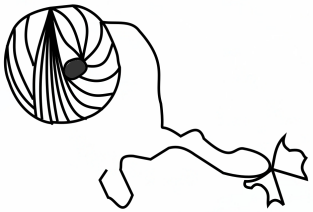


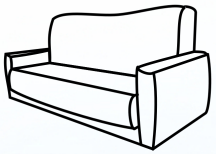
$$= \theta_0 + \theta_1 \times (\text{кол-во}) - \theta_2 \times (\text{кол-во})^2$$

$$+ \text{погрешности}$$

Другие факторы

$$= \theta_0 + \theta_1 \times \text{  } - \theta_2 \times \left(\text{  }\right)^2$$

$$+ \theta_3 \times \text{  }$$

$$+ \theta_4 \times \text{  }$$

$$+ \text{погрешности}$$

Регрессионный анализ

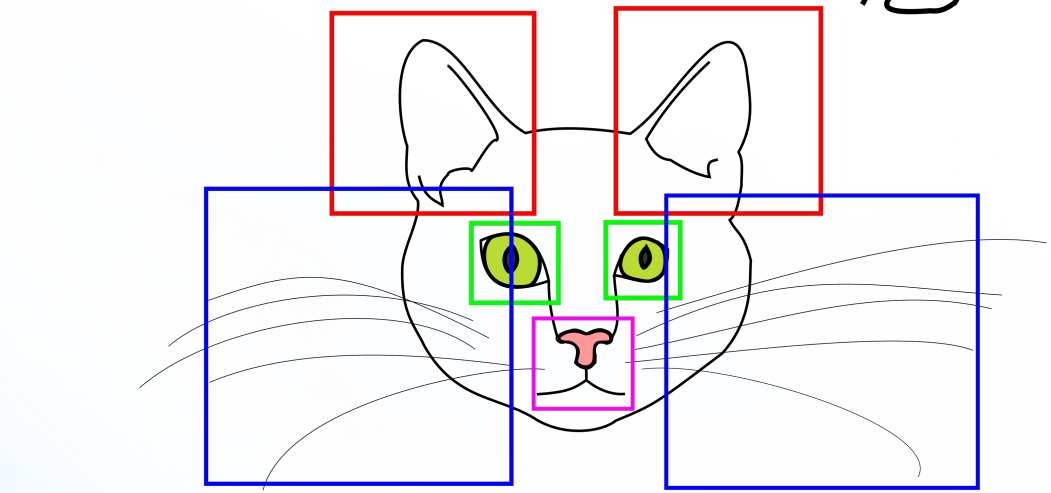
Классификация котиков

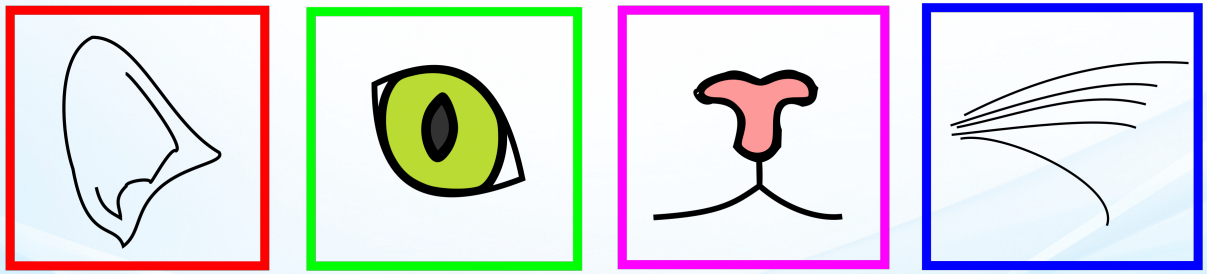
Классификация

Собираем данные

котик	порода	рост	шерсть
	Саванна	50 см	да
	Сфинкс	30 см	нет
	Манчкин	15 см	да
	Саванна	40 см	да

Распознавание мордочек





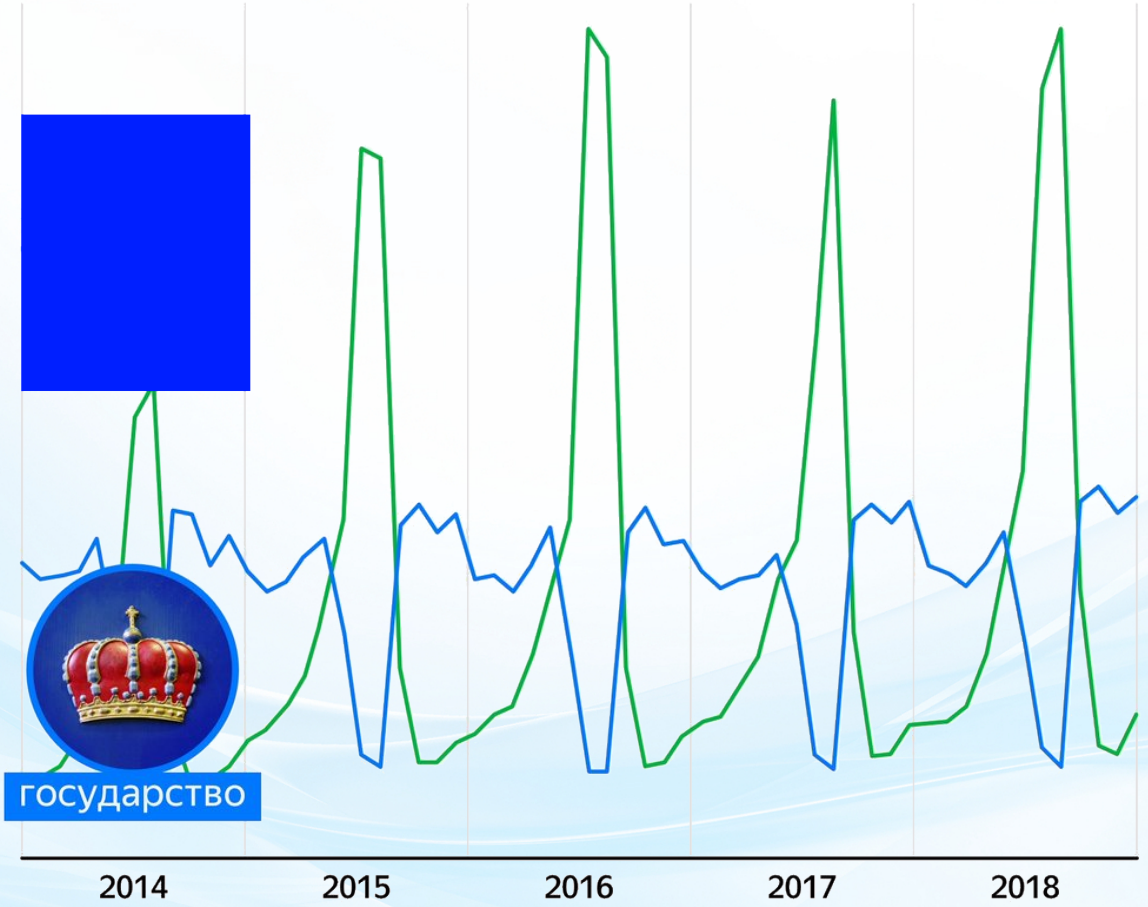
Нейронные сети

Художник курса

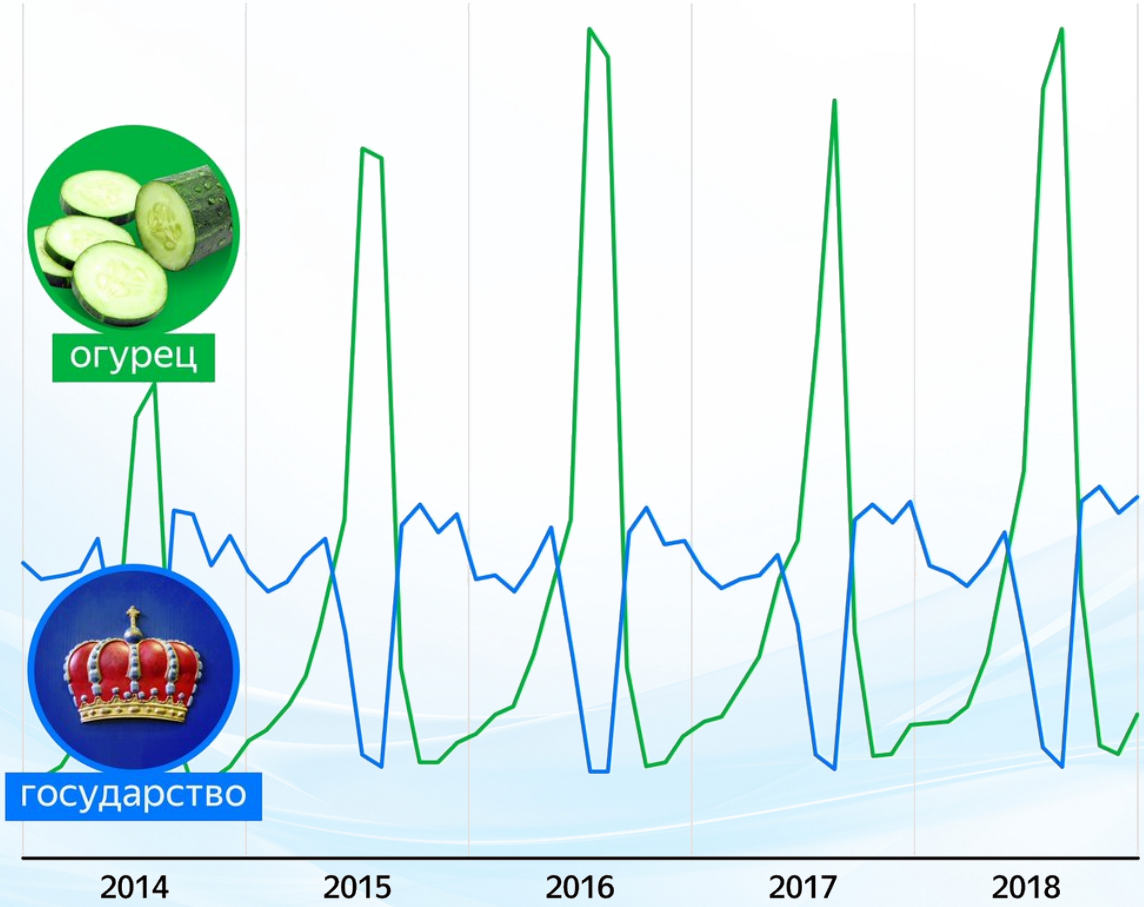
Книга с похожим содержанием

Яндекс

Когда в Поиске растёт интерес к [redacted] снижается доля запросов со словом государство

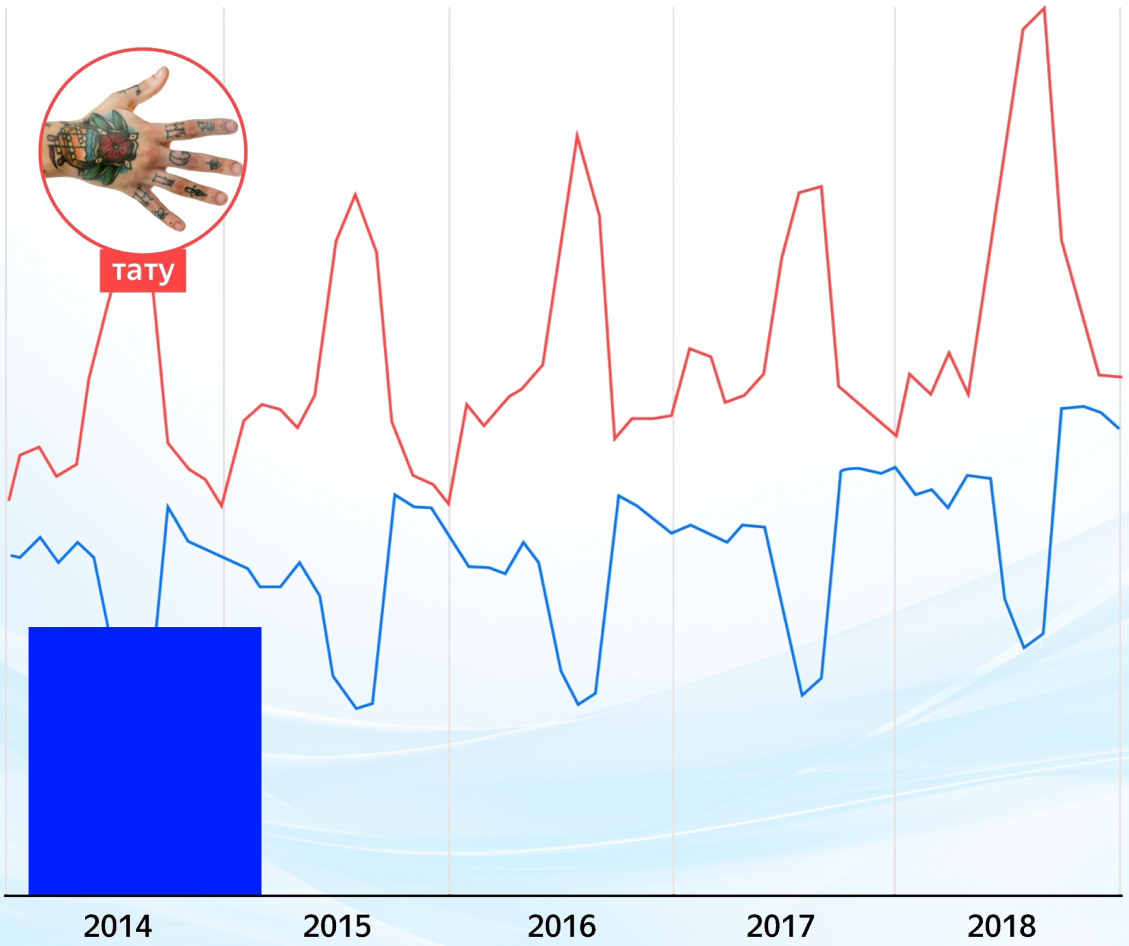


Когда в Поиске растёт интерес к огурцам, снижается доля запросов со словом государство



Яндекс

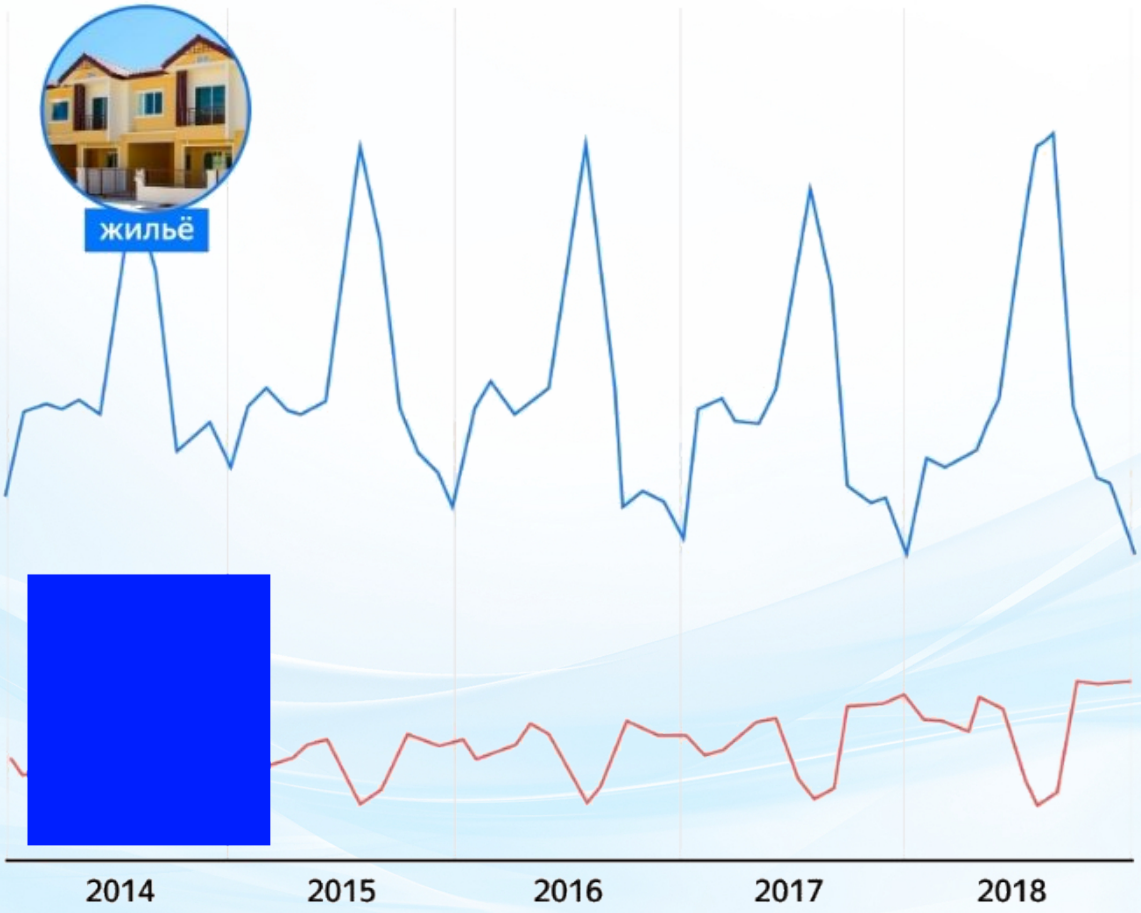
Когда в Поиске растёт интерес к тату,
снижается доля запросов со словом [redacted]



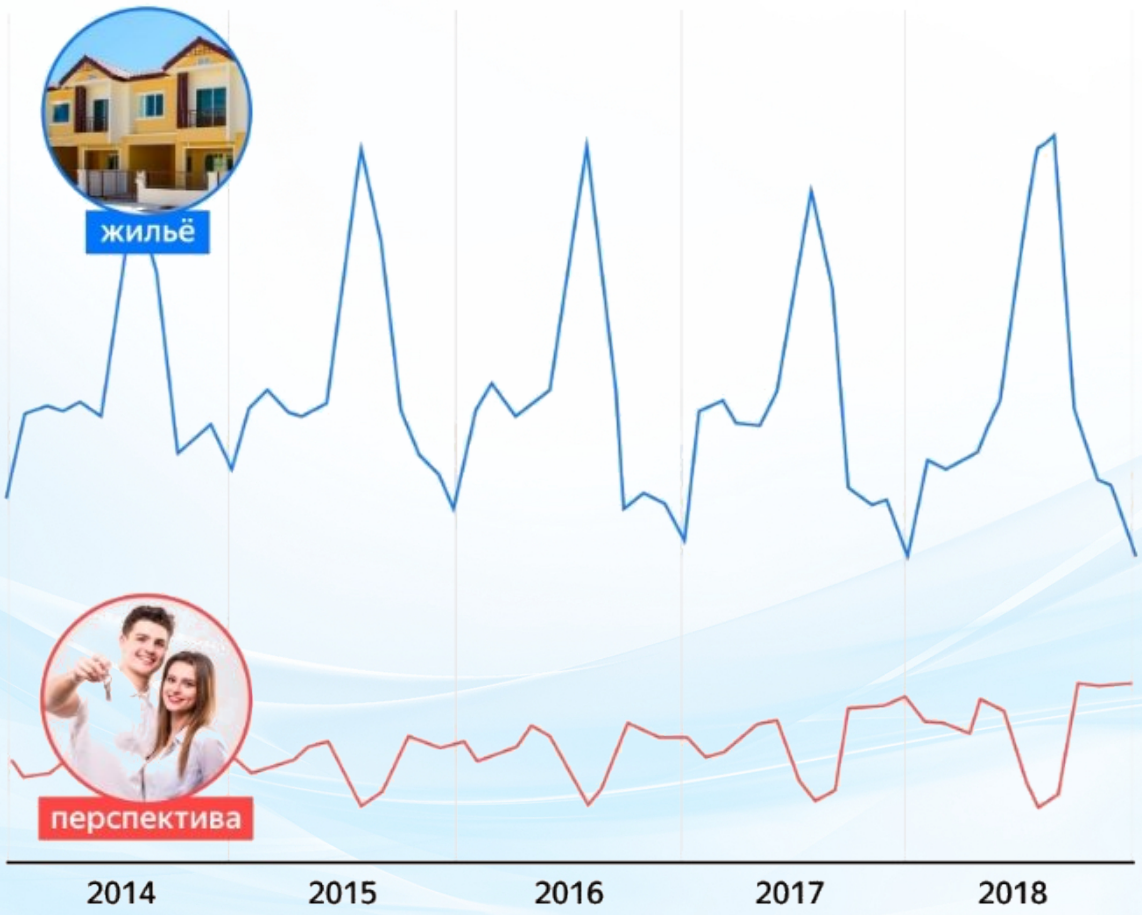
Когда в Поиске растёт интерес к тату,
снижается доля запросов со словом смысл

Яндекс

Когда в Поиске растёт интерес к жилью, снижается доля запросов со словом [redacted]



Когда в Поиске растёт интерес к жилью, снижается доля запросов со словом перспектива



Президентские выборы в США в 1948 г.

Гарри Труман (демократы) vs. Томас Дьюи (республиканцы)

В ночь на оглашение результатов газета Chicago Tribune опубликовала заголовок: **DEWEY DEFEATS TRUMAN**

Chicago Daily Tribune

THE WORLD'S GREATEST NEWSPAPER

HOME

DEWEY DEFEATS TRUMAN

G.O.P. Sweep Indicated in State; Boyle Leads in City

REPUBLICAN TICKET AHEAD OF 1944 VOTE

Town Balloting Gives Trend

Tops Coghlan in Hot Race for Attorney

Probable Winners

RECORD CITY VOTE SEEN IN LATE TALLIES

Suburbs Ballot Near 575,000

BULLETINS ON ELECTIONS

Early Count Gives G.O.P. Senate Edge

PUTS G.O.P. BACK IN THE WHITE HOUSE

Sizable Electoral Margin Seen

После закрытия участков газета провела опрос, обзвонив большое число избирателей, все предвещало оглушительную победу Дьюи.

Президентские выборы в США в 1948 г.

Смеющийся Труман, победитель выборов 1948 года.

Что же пошло не так?

В 1948 году телефон был доступен только людям определенного достатка и редко встречался у людей с небольшим заработком.
Выборка не учитывала достаточно широкий пласт избирателей Трумана, т.к. как правило демократы имеют большую долю голосов среди бедного населения, которым телефон в свою очередь был недоступен.

1. Введение

1.1. Основная задача математической статистики

1.2. Вероятностно-статистическая модель

1.3. Виды подходов к статистике

Введение

Теория вероятностей

Зная природу случайного явления,

посчитать характеристику этого явления.

Математическая статистика

По результатам экспериментальных данных

высказать суждение о том, какова была природа этого явления.

Классический пример

На третьем курсе N студентов; из них M выбирает прикладной поток.

Задача в теории вероятностей

P(среди n чел. ровно m слушателей прикладного потока)$-$?

Предполагается, что M известно.

Задача в математической статистике

Среди случайных n чел. есть m слушателей прикладного потока.

Оценить M.

Предполагается, что M не известно.

Еще пример

$\xi \sim \mathcal{N}(a, \sigma^2)$ — случайная величина

Задача в теории вероятностей

Известно, что $a = 2.3, \sigma = 7.1$

$$\mathrm{P}(\xi \in [0, 1]) - ?$$

$$\mathrm{E}\xi - ?$$

Задача в математической статистике

$x_1, ..., x_n$ — независимые реализации случайной величины ξ.

Оценить a и σ.

Задача математической статистики

Пусть $x_1, ..., x_n$ — численные характеристики n-кратного повторения некоторого явления.

Будем их воспринимать как независимые реализации $\xi \sim \mathrm{P}$.

Задача: по значениям $x_1, ..., x_n$ высказать некоторое суждение о распределении P.

Решение: *статистический вывод* или *обучение*.

1. Введение

1.1. Основная задача математической статистики

1.2. Вероятностно-статистическая модель

1.3. Виды подходов к статистике

Однократный эксперимент

$\mathscr{X}$ — *выборочное пространство* = множество всех возможных значений эксперимента;

$\mathscr{B}_{\mathscr{X}}$ — некоторая σ-алгебра на $\mathscr{X}$;

P — некоторое неизвестное распределение на $(\mathscr{X}, \mathscr{B}_{\mathscr{X}})$;

Предполагается $\mathrm{P} \in \mathscr{P}$ — некоторое семейство распределений.

Вероятностно-статистическая модель

$$(\mathscr{X}, \mathscr{B}_{\mathscr{X}}, \mathscr{P}).$$

Для оперирования с результатами эксперимента как со случайными величинами, определим случайную величину $X : \mathscr{X} \to \mathscr{X}$ по правилу $X(x) = x \ \forall x \in \mathscr{X}$, которую будем называть *наблюдением*.

Многократный эксперимент

Вероятностно-статистическая модель

$$(\mathscr{X}^n, \mathscr{B}_{\mathscr{X}^n}, \mathscr{P}^n),$$

- $\mathscr{X}^n = \mathscr{X} \times ... \times \mathscr{X}$;
- $\mathscr{B}_{\mathscr{X}^n} = \sigma(B_1 \times ... \times B_n, B_i \in \mathscr{B}_{\mathscr{X}})$;
- $\mathscr{P}^n = \{\mathrm{P}^n, \mathrm{P} \in \mathscr{P}\}$, причем

  $\mathrm{P}^n(B_1 \times ... \times B_n) = \mathrm{P}(B_1)...\mathrm{P}(B_n)\ \forall B_i \in \mathscr{B}_{\mathscr{X}}$.

  Для любой $\mathrm{P} \in \mathscr{P}$ определенная таким образом P^n существует и единственна (теорема о продолжении вероятностной меры).

Многократный эксперимент

Наблюдение, соответствующее i-му эксперименту:

Сл. вел. $X_i : \mathscr{X}^n \to \mathscr{X}$, т.ч. $X_i(x) = x_i \ \forall x \in \mathscr{X}^n$.

Случайный вектор $X = (X_1, ..., X_n)$ — *выборка* размера n.

Выборка является вектором независимых

одинаково распределенных случайных величин,

каждая компонента которого имеет распределение P.

Бесконечный эксперимент

Вероятностно-статистическая модель

$$(\mathscr{X}^\infty, \mathscr{B}_{\mathscr{X}^\infty}, \mathscr{P}^\infty),$$

- $\mathscr{X}^\infty = \mathscr{X} \times \mathscr{X} \times ...$;
- $\mathscr{B}_{\mathscr{X}^\infty} = \sigma(B_1 \times ... \times B_n \times \mathscr{X}^\infty,\ B_i \in \mathscr{B}_{\mathscr{X}}, n \in \mathbb{N})$;
- $\mathscr{P}^\infty = \{\mathsf{P}^\infty, \mathsf{P} \in \mathscr{P}\}$, причем

  $\mathsf{P}^\infty(B_1 \times ... \times B_n \times \mathscr{X} \times ...) = \mathsf{P}^n(B_1 \times ... \times B_n)\ \forall B_i \in \mathscr{B}_{\mathscr{X}}$.

  Для любой $\mathsf{P} \in \mathscr{P}$ определенная таким образом P^n существует и единственна (теорема о продолжении вероятностной меры).

Бесконечный эксперимент

Наблюдение, соответствующее i-му эксперименту:

Сл. вел. $X_i : \mathscr{X}^\infty \to \mathscr{X}$, т.ч. $X_i(x) = x_i \ \forall x \in \mathscr{X}^\infty$.

Случайная последовательность $X = (X_1, X_2, ...)$ —
выборка неограниченного размера.

Выборка неограниченного размера является последовательностью независимых одинаково распределенных случайных величин, каждая компонента которого имеет распределение P.

Далее

- Для простоты будем опускать индексы n и ∞;
- Будем считать, что в качестве сигма-алгебры используется борелевская, если не сказано обратное;
- Да и вообще забудем про ее существование :)

1. Введение

1.1. Основная задача математической статистики

1.2. Вероятностно-статистическая модель

1.3. Виды подходов к статистике

Классификация по типу методов вывода

1. *Параметрический*

   Предполагается, что истинное распределение P принадлежит некоторому классу распределений $\mathscr{P}$,

   которое параметризовано параметром $\theta \in \Theta$.

   $\mathsf{P} \in \{\mathsf{P}_\theta \mid \theta \in \Theta\}$

2. *Непараметрический*

   Предполагается, что истинное распределение P принадлежит некоторому классу распределений $\mathscr{P}$,

   на котором не введен параметр.

Параметрический подход

1. $X_1, ..., X_n \sim Exp(\theta)$, где $\theta > 0$, подразумевает:

   $\mathscr{X} = (0, +\infty), \mathscr{P} = \{Exp(\theta) \mid \theta > 0\}, \Theta = (0, +\infty).$

   Статистический вывод: указание числа из множества Θ.

2. Схема испытаний Бернулли $X_1, ..., X_n$ подразумевает:

   $\mathscr{X} = \{0, 1\}, \mathscr{P} = \{Bern(\theta) \mid 0 \leqslant \theta \leqslant 1\}, \Theta = [0, 1].$

   Статистический вывод: указание числа из множества Θ.

3. $X_1, ..., X_n \sim \mathcal{N}(a, \sigma^2)$, где оба параметра неизвестны:

   $\mathscr{X} = \mathbb{R}, \mathscr{P} = \{\mathcal{N}(a, \sigma^2) \mid a \in \mathbb{R}, \sigma > 0\},$

   $\theta = (a, \sigma), \Theta = \mathbb{R} \times (0, +\infty).$

   Статистический вывод: указание пары чисел из множества Θ.

Пример

$\mathscr{X} = \mathbb{R}$;

$\mathscr{P} = \{U(0, \theta) \mid \theta > 0\}$;

Дана выборка $(X_1, X_2, X_3) = (1, 2, 3)$.

Может ли истинное значение θ быть равным 100 , 3 , 1.5 , -1?

- 100 и 3 — да;
- -1 — нет, поскольку $\theta > 0$;
- 1.5 — да, поскольку $\mathscr{X} = \mathbb{R}$

  $\Rightarrow$ возможны любые вещественные числа, правда вероятность получения хотя бы одного числа вне отрезка $[0, \theta]$ равна нулю.

Непараметрический подход

1. В отсутствии предположений:

   $\mathscr{X} = \mathbb{R}$;

   $\mathscr{P}$ — все распределения на $\mathbb{R}$.

   В качестве статистического вывода можно

   некоторым образом оценить функцию распределения;

2. Предполагается, что выборка взята из непрерывного распред.:

   $\mathscr{X} = \mathbb{R}$;

   $\mathscr{P}$ — все непрерывные распределения на $\mathbb{R}$.

   В качестве статистического вывода можно оценить плотность.

- Любое семейство распределений может рассматриваться как в параметрическом подходе, так и в непараметрическом;
- Смысл деления на два типа — принципиально разные методы к оценке неизвестного распределения:
  - Методы параметрического подхода как-либо оценивают параметр, соответствующий неизвестному распределению.
    На практике обычно $\Theta \subset \mathbb{R}^d$, размерность d фиксирована;
  - Непараметрические методы пытаются некоторым способом напрямую оценить неизвестное распределение.
    На практике обычно содержат нефиксированное количество параметров.

Классификация по способу вывода

Два доминирующих подхода:

1. *Частотный*

   Построение суждения о распределении P происходит только на основе выборки $X_1, ..., X_n$.

   Все распределения из класса $\mathscr{P}$ равноправны.

2. *Байесовский*

   На распределениях из $\mathscr{P}$ задано некоторое распределение ("априорное знание"), которое учитывается при построении суждения, как правило, с помощью формулы Байеса.

   Пример: $\mathscr{P} = \{\mathcal{N}(\theta, 1) \mid \theta \in \mathbb{R}\}$,

   и предполагается, что истинное значение θ_0 выбрано из $\mathcal{N}(0, 1)$.


ВСЁ!