



# **Прикладные задачи анализа данных**

## **ОЦЕНКИ СРЕДНЕГО, ВЕРОЯТНОСТИ И ПЛОТНОСТИ. ВЕСОВЫЕ СХЕМЫ**

**Дьяконов А.Г.**

**Московский государственный университет  
имени М.В. Ломоносова (Москва, Россия)**

## Что такое среднее?

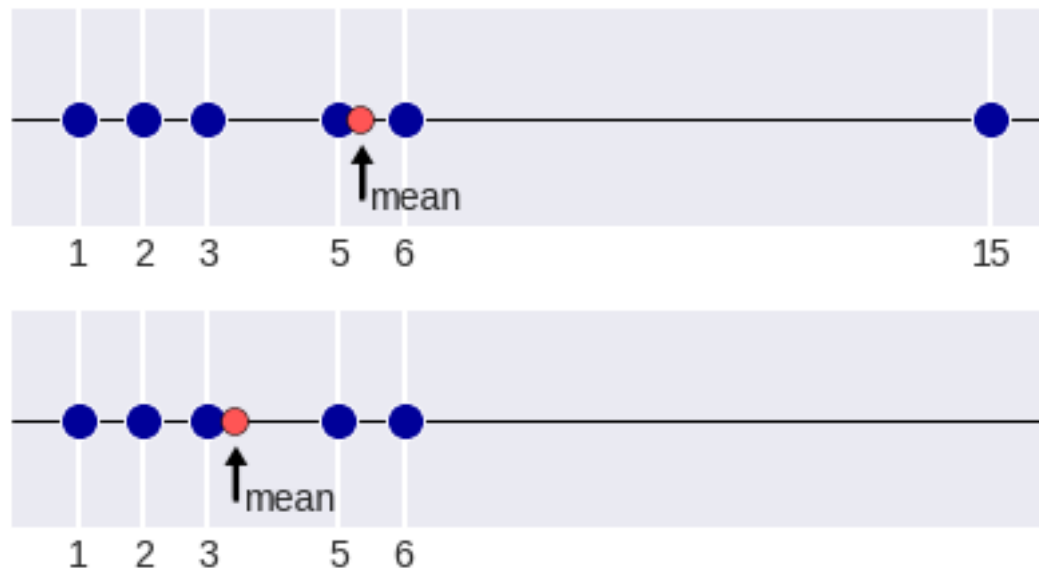
средний, типичный, среднестатистический

Естественная формализация – **среднее арифметическое**

$$\text{mean}(X) = \frac{x_1 + \dots + x_m}{m}$$

Большой плюс – среднее можно вычислять в  $\mathbb{R}^n$

### 1) Проблема выбросов



## Что такое среднее?

средний, типичный, среднестатистический

Естественная формализация – **среднее арифметическое**

### 2) Проблема «виртуальных точек»

**Признак «пол»:** [М, Ф, Ф, М, М, М, Ф, Ф, Ф, Ф]

– Какой у нас среднестатистический клиент?

– Он на 40% мужчина?

– Хочется конкретный пример!

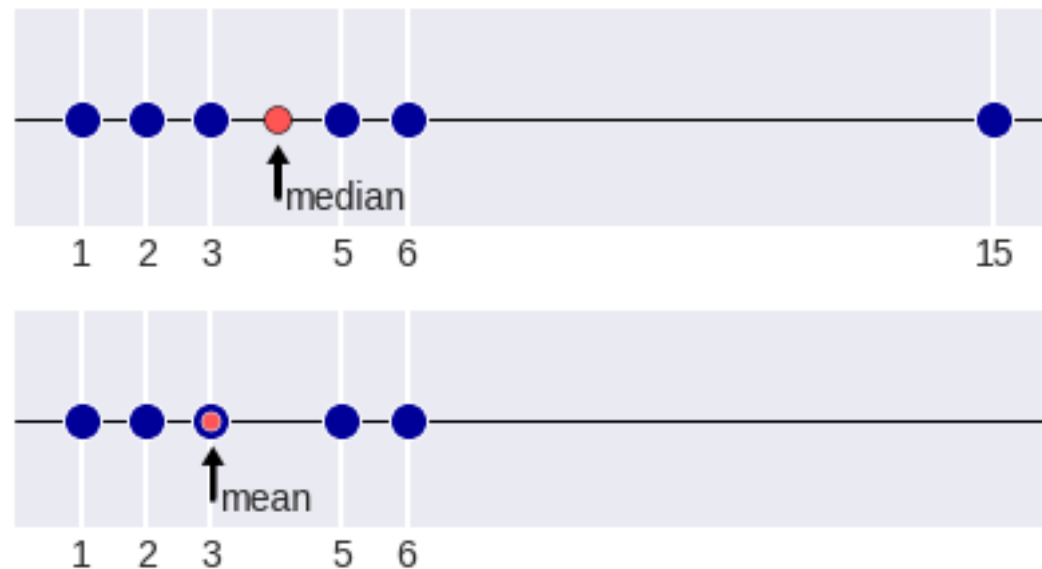
## Что такое среднее?

**Решение проблемы – медиана.**

$$\text{median}(X) = \frac{x_{\lfloor n/2 \rfloor} + x_{\lceil n/2 \rceil}}{2}$$

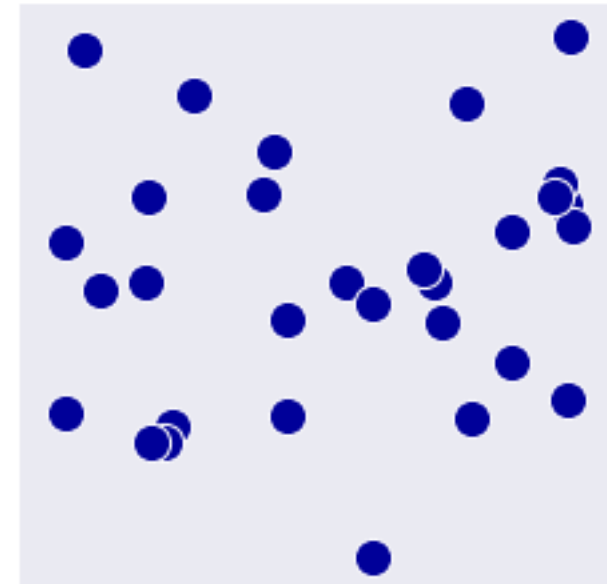
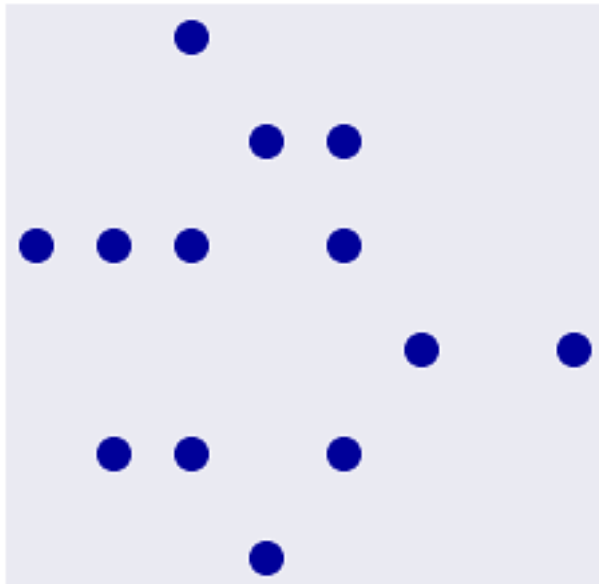
**1) устойчива к выбросам**

**2) является (можно сделать!) точкой выборки**



## Проблема медианы

**Что такое многомерная медиана?**



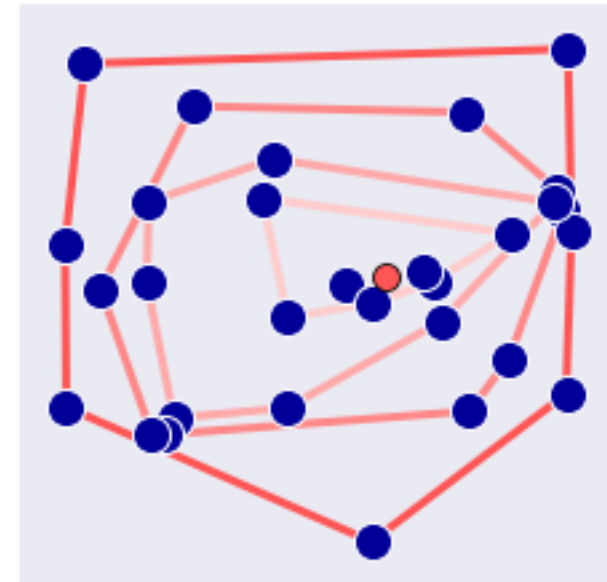
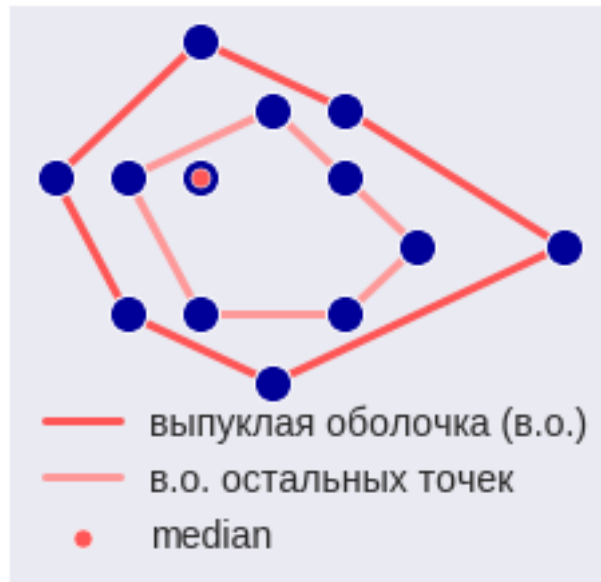
**Хочется инвариантность к**

- **движениям**
  - поворотам
  - сдвигам (параллельным переносам)
- **сжатиям**

**В одномерном случае должна совпадать с median!**

## Многомерная медиана

### Что такое многомерная медиана?



**Выход: сделать аналогичный процесс построения,  
как в одномерном случае  
удаление крайних элементов!**

## **Многомерная медиана**

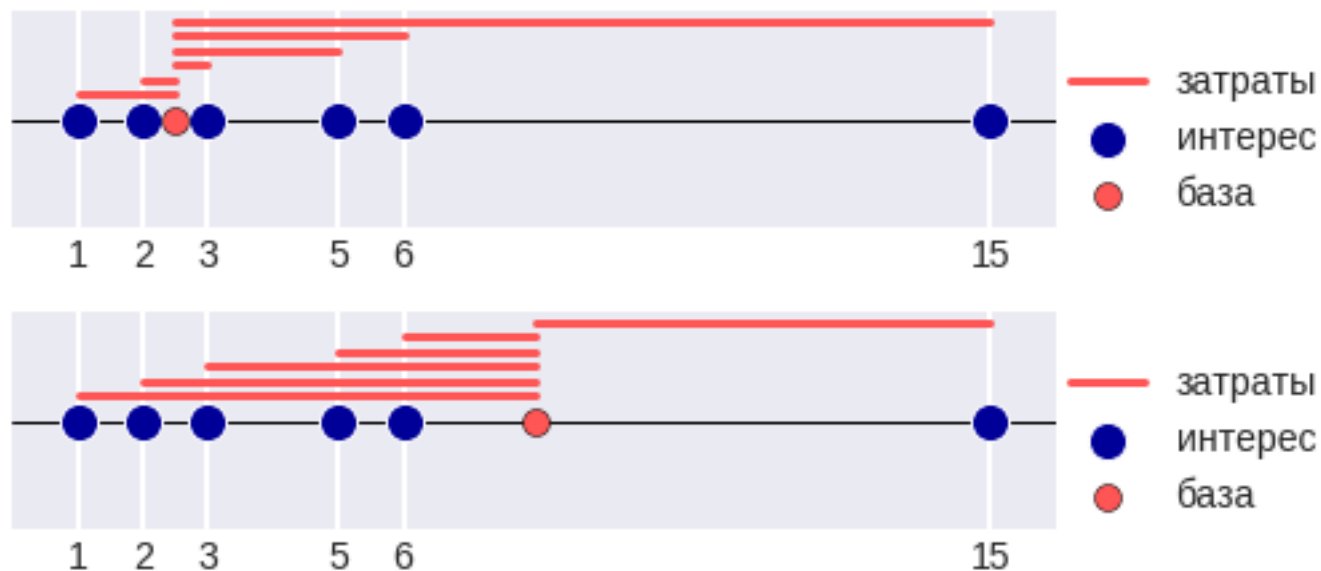
**Если признаки разнородны, неравноценны и т.п.  
(не нужно инвариантности к поворотам)**

**Всё равно можно применить подход  
«отбрасывания крайних элементов».**

**Вопрос: как, где?**

## Среднее как решение оптимизационной задачи

- Живём в одномерном мире «на базе»
  - Есть пункты интереса
  - Есть функция затрат
- Надо минимизировать суммарные затраты





## Среднее как решение оптимизационной задачи

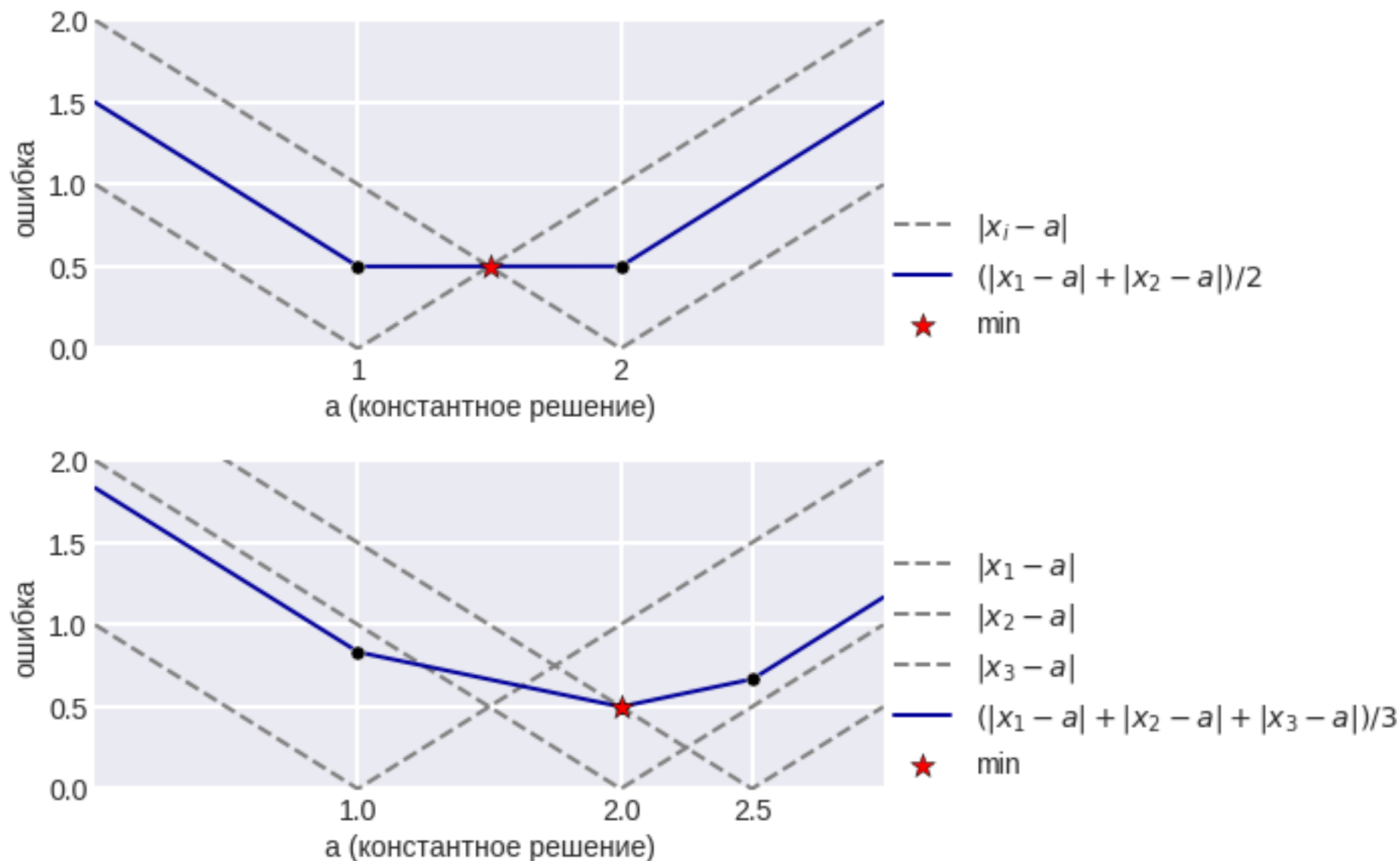
**Если суммарные затраты**

$$\sum_{i=1}^m |x_i - a| \rightarrow \min$$

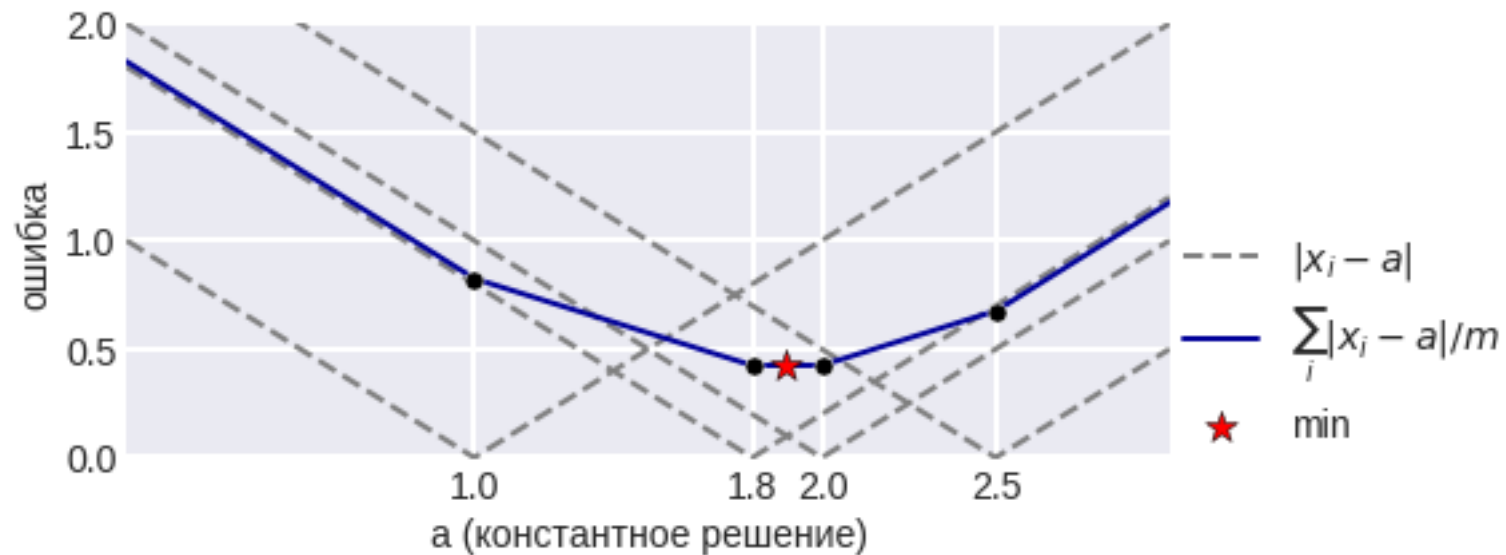
**то решение – медиана**



## Среднее как решение оптимизационной задачи



## Среднее как решение оптимизационной задачи



## Медиана в пространстве

### 2й способ формализации:

**аналогично минимизируем затраты**

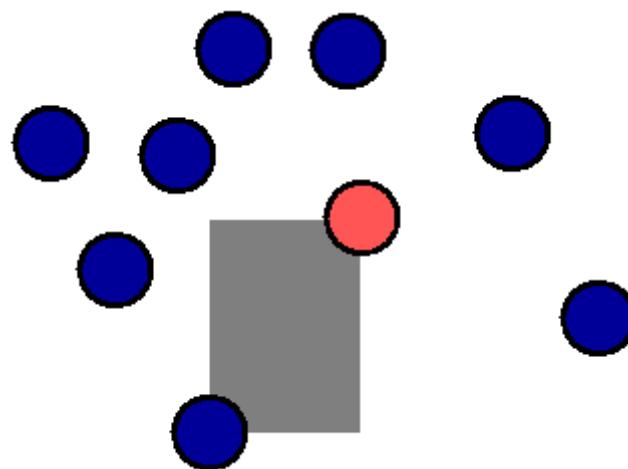
**но тут может быть зависимость от координат!**

$$\sum_{i=1}^m \left( |x_i - a_1|^d + |y_i - a_2|^d \right)^{1/d} \rightarrow \min$$

$$\sum_{i=1}^m |x_i - a_1| + \sum_{i=1}^m |y_i - a_2| \rightarrow \min$$

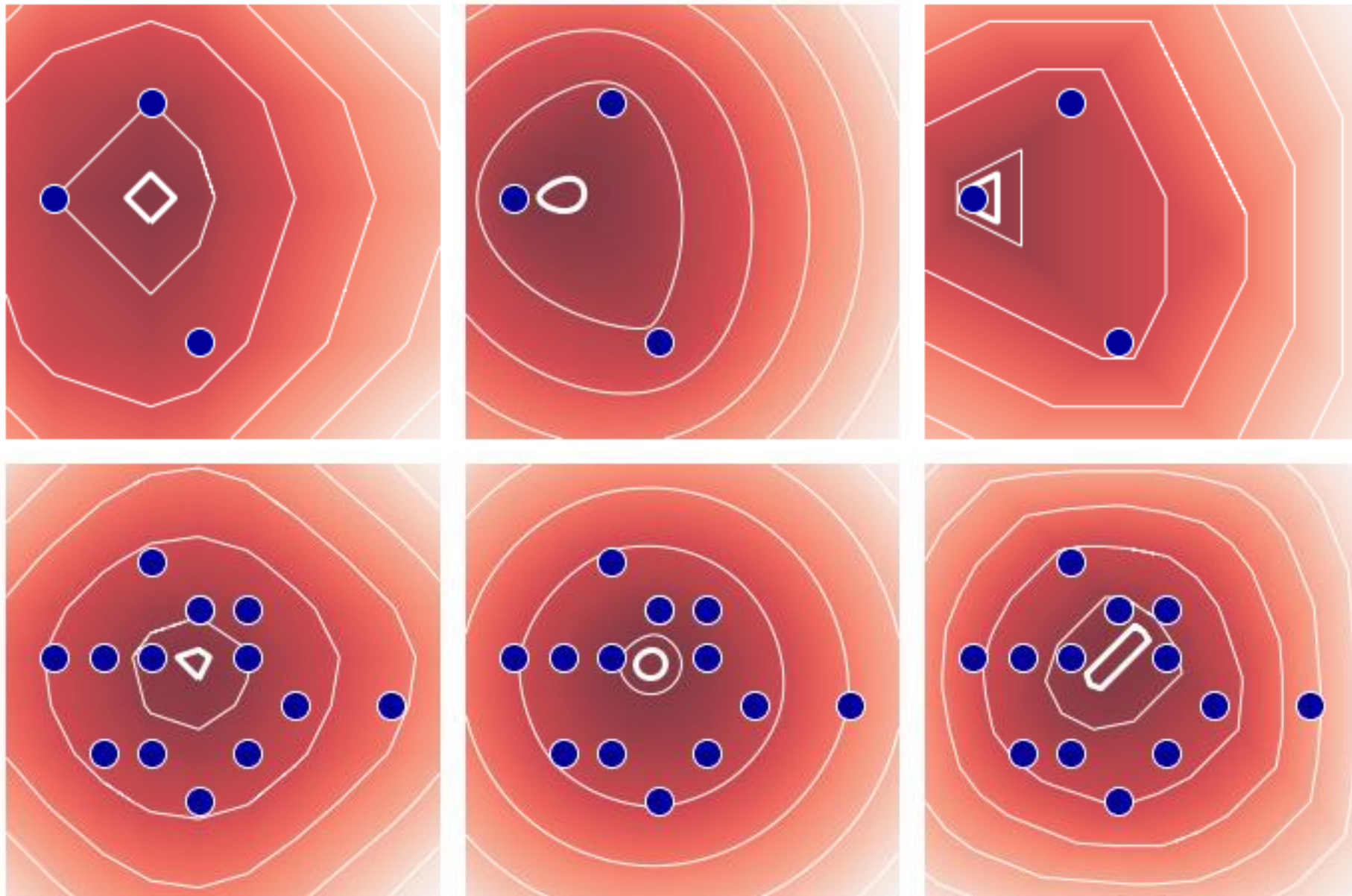
$$\sum_{i=1}^m \max[|x_i - \mu_1|, |y_i - \mu_2|] \rightarrow \min$$

$$\sum_{i=1}^m |x_i - a_1| \cdot |y_i - a_2| \rightarrow \min$$

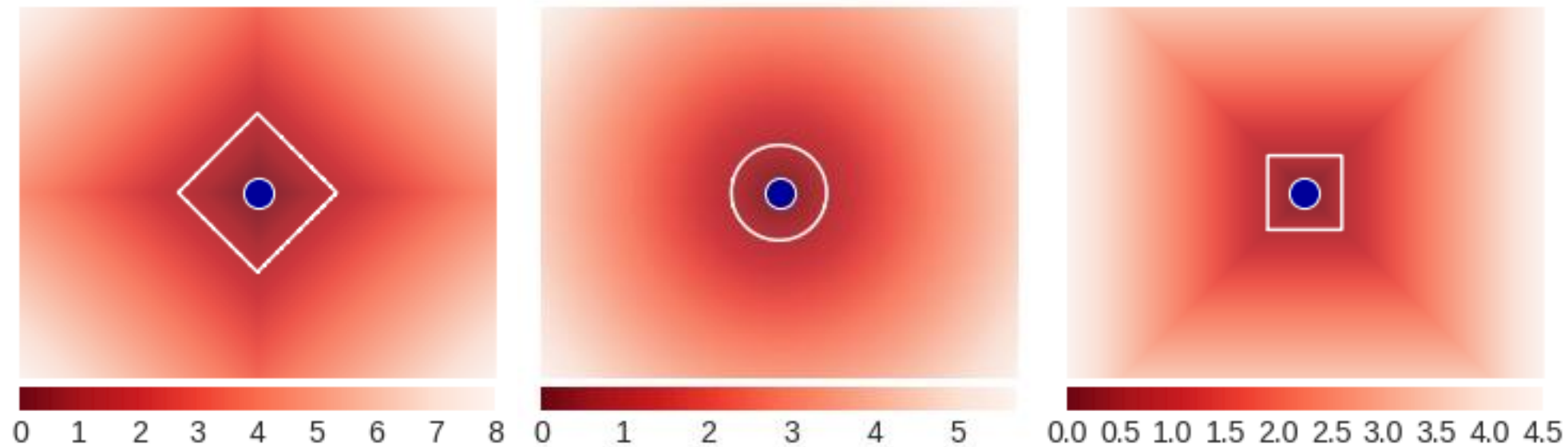


**Решаем перебором по точкам выборки!!!**

## «Степень медианности» – какие функции представлены?



## «Степень медианности»



$$\sum_{i=1}^m |x_i - a_1| + \sum_{i=1}^m |y_i - a_2| \rightarrow \min$$

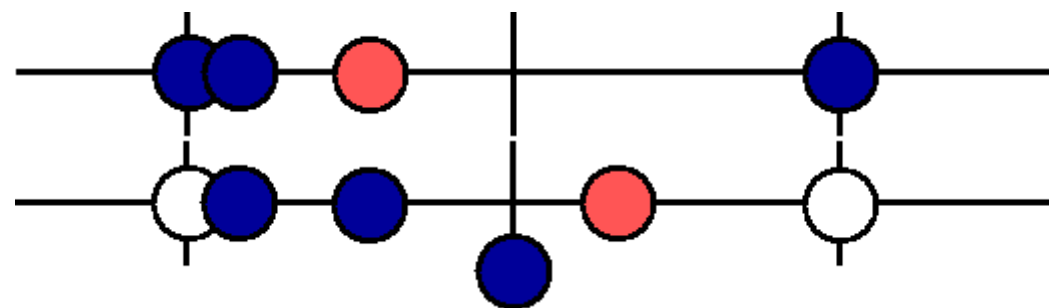
$$\sum_{i=1}^m (|x_i - a_1|^2 + |y_i - a_2|^2)^{1/2} \rightarrow \min$$

$$\sum_{i=1}^m \max[|x_i - a_1|, |y_i - a_2|] \rightarrow \min$$

**ДЗ: есть ли тут желанные свойства медианы?**

## Эвристический способ борьбы с выбросами

$$a = \frac{1}{m} \sum_{i=1}^m x_i$$



### Алгоритм Шурыгина

- 1. Если  $m \leq 2$ , то пользуемся формулой (\*). Выход.**
- 2. Пусть  $x_1 \leq \dots \leq x_m$  (без ограничения общности).**
- 3. Если  $\frac{x_1 + x_m}{2} \leq x_2$ , то удаляем из выборки  $x^1$ . Переходим к п.1 (с соответствующей перенумерацией объектов).**
- 4. Если  $\frac{x_1 + x_m}{2} \geq x_{m-1}$ , то удаляем из выборки  $x_m$ . Переходим к п.1 (с соответствующей перенумерацией объектов).**
- 5. Исключаем из выборки  $x_1, x_m$ , но добавляем в неё  $\frac{x_1 + x_m}{2}$ .**

## **Борьба с выбросами**

**В чём недостаток алгоритма Шурыгина?**

**Практика:** часто забываем о выбросах



## Что минимизирует «среднее»

$$\text{median}(X) = \arg \min \sum_{i=1}^m |x_i - a|$$

$$\text{mean}(X) = \arg \min \sum_{i=1}^m |x_i - a|^2$$

**Для минимизации можно выбрать «что угодно»**

$$\text{mid}(X) = \arg \min \sum_{i=1}^m f(x_i, a)$$

**– оценка минимального контраста**

**... другие формализации понятия «среднее»**

## Оценка минимального контраста

**Если после дифференцирования  
(здесь рассматриваем одномерный случай)**

$$\sum_{i=1}^m \psi(x_i - a) = \sum_{i=1}^m (x_i - a) \xi(x_i - a) = 0,$$

**для некоторых функций  $\psi$  (оценочная функция) и  $\xi$  (весовая функция), то часто успешно применяется итеративный способ вычисления параметра  $a$  по формуле**

$$a = \frac{\sum_{i=1}^m x_i \xi(x_i - a)}{\sum_{i=1}^m \xi(x_i - a)}.$$

**Д/З Проверить применимость формулы**

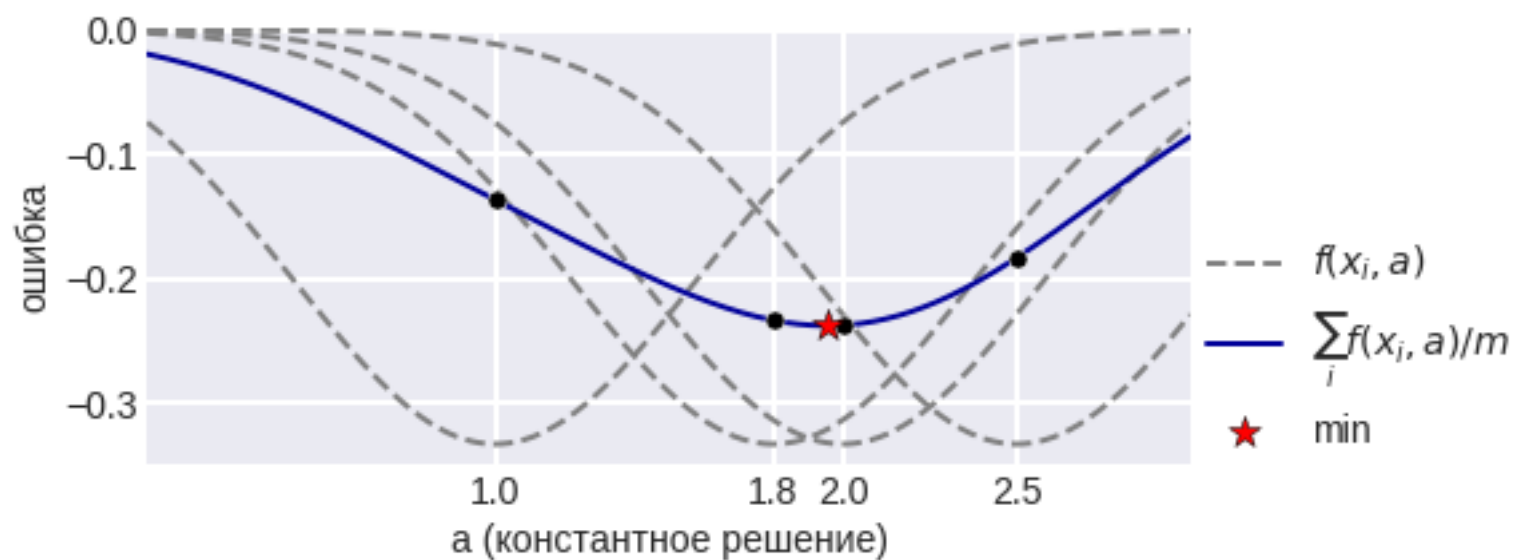
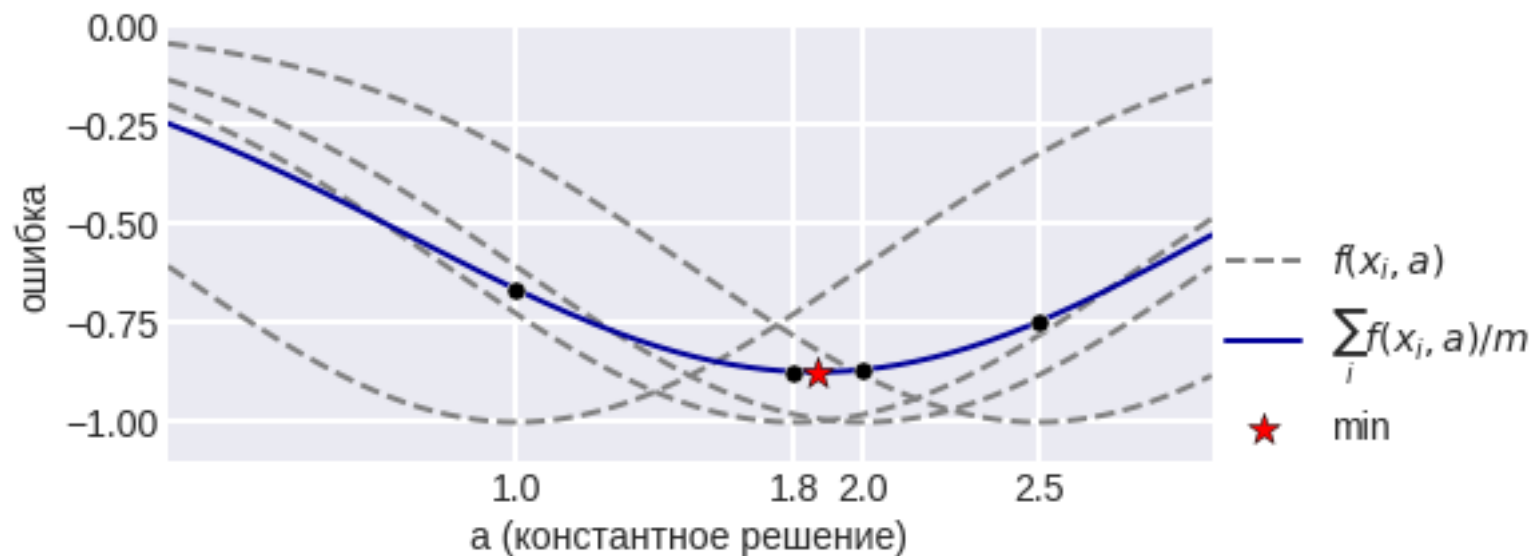
## Принстонский эксперимент 1972 года подбор различных функций

**Мешалкин Л.Д. (1977) предлагал**

$$f(x, a) = -\frac{1}{\lambda} e^{-\frac{\lambda(x-a)^2}{2}}$$

$$\psi(z) = ze^{-\lambda z^2/2}, \quad \xi(z) = e^{-\lambda z^2/2}.$$

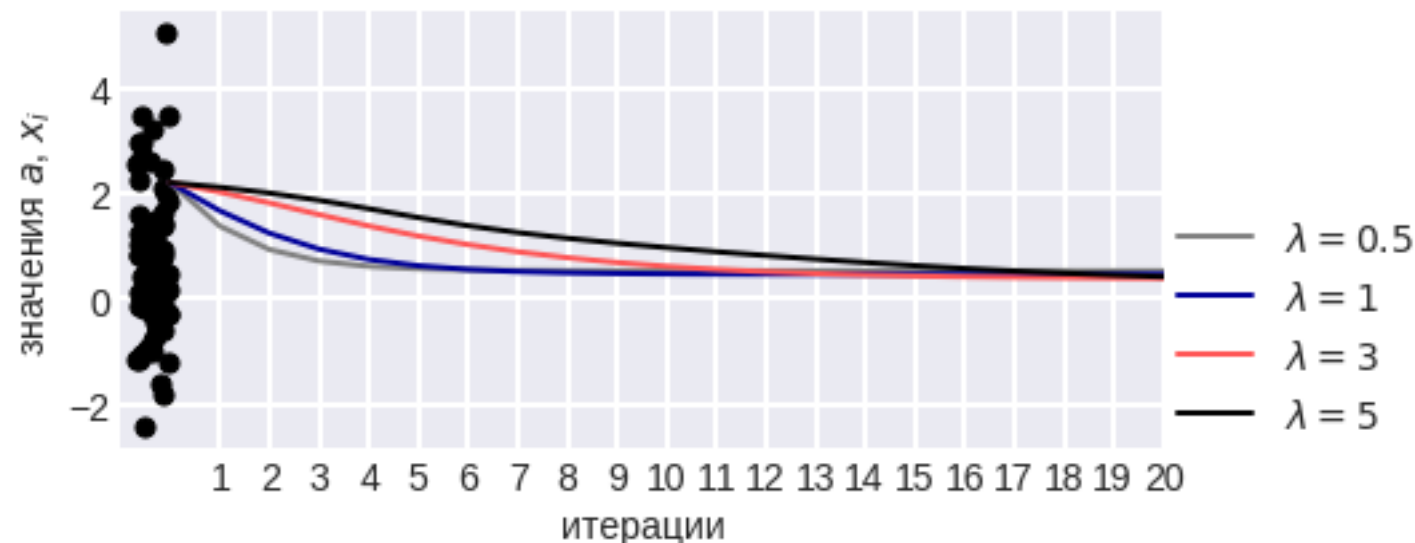
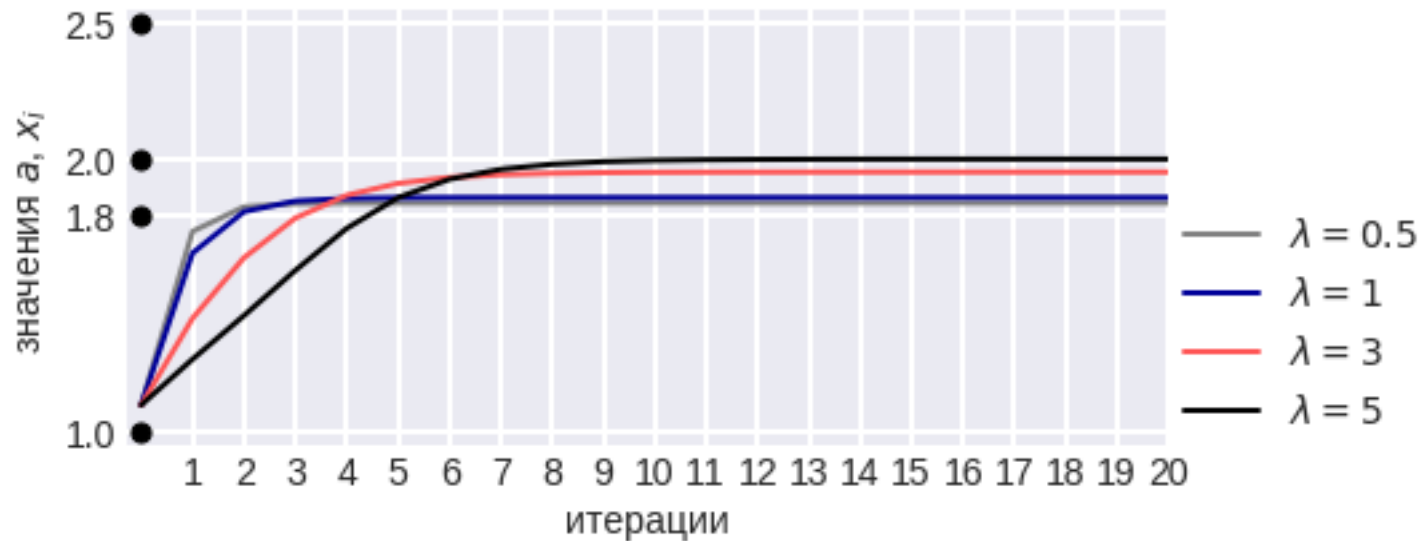
## Чем отличаются рисунки?



## Чем отличаются рисунки?

$$\lambda = 1 \quad \lambda = 3$$

**Результаты пересчёта: что важно, как в любой задаче оптимизации?**



## **Что важно?**

**Начальное приближение  
Масштаб**

**Для справки**

**Система уравнений для их поиска оценок среднего и матрицы ковариации для многомерного распределения:**

$$\begin{cases} \sum_{i=1}^m (x^i - \mu) e^{-\lambda \cdot q_i / 2} = 0, \\ \sum_{i=1}^m \left( (x^i - \mu)(x^i - \mu)^T - \frac{1}{1 + \lambda} C \right) \cdot e^{-\lambda \cdot q_i / 2} = 0, \end{cases}$$

$$q_i = (x^i - \mu)^T C^{-1} (x^i - \mu)$$

**Обобщение медианы на многомерный случай**

$$\mu = \frac{\sum_{i=1}^m \frac{x^i}{\sqrt{q_i}}}{\sum_{i=1}^m \frac{1}{\sqrt{q_i}}}.$$

**итерационный алгоритм**

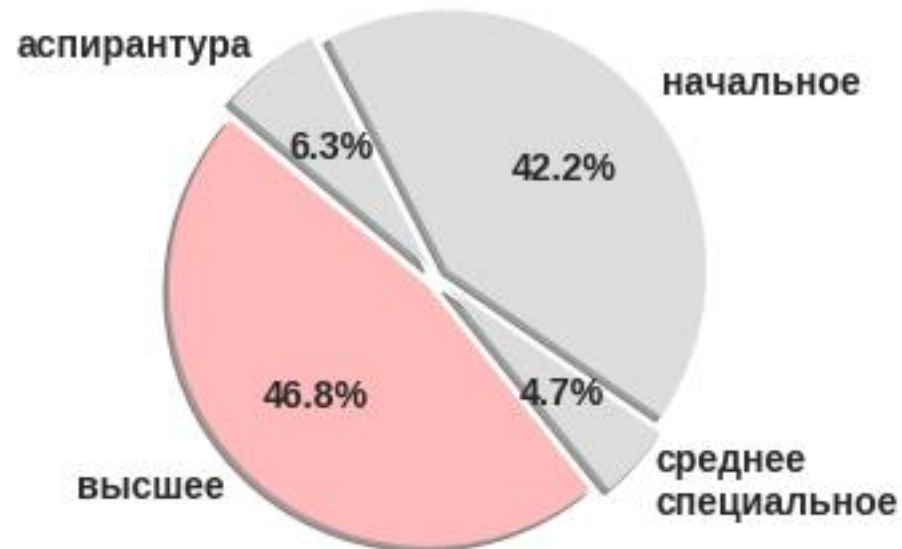
**[см. Шурыгин]**

## Что такое среднее для номинальных признаков?



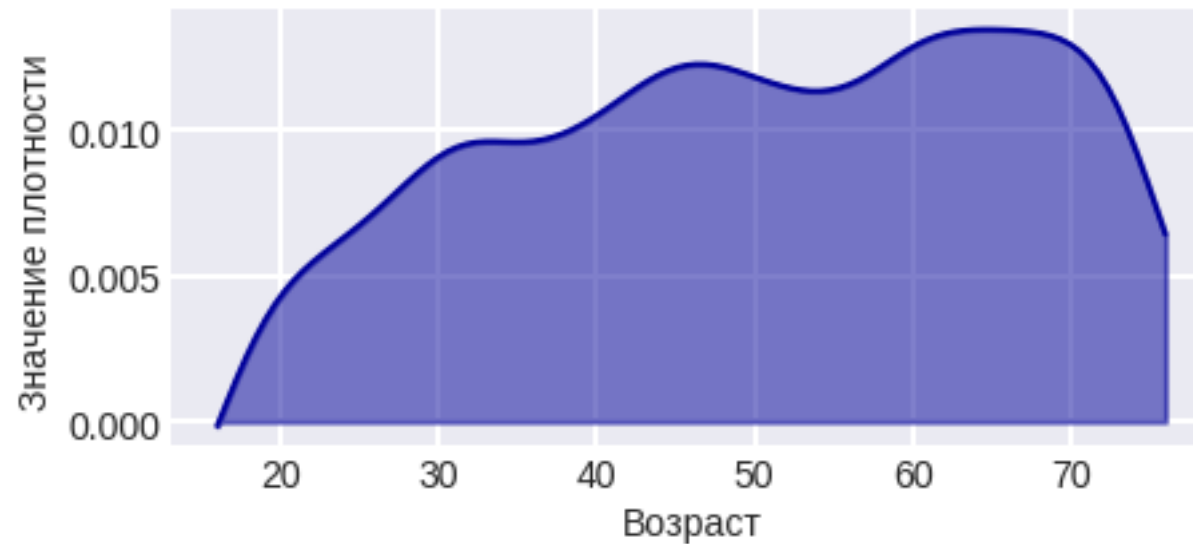


## Что такое среднее для номинальных признаков?

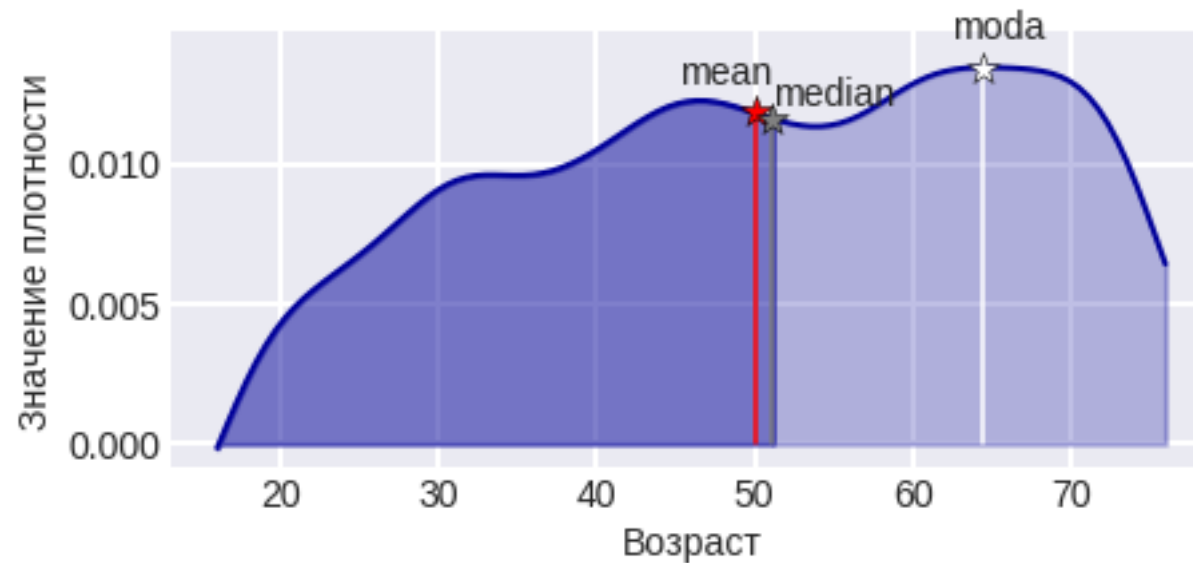


**Мода – самое популярное значение**  
– самое вероятное значение

## Где матожидание, медиана, мода?



## Где матожидание, медиана, мода?



## **Практика: придумывать не функционал, а среднее**

### **Среднее по А.Н.Колмогорову**

$$\varphi^{-1}\left(\frac{\varphi(x_1) + \dots + \varphi(x_n)}{n}\right)$$

**среднее арифметическое**  $\varphi(x) = x$

**среднее геометрическое**  $\varphi(x) = \log x$

**среднее гармоническое**  $\varphi(x) = x^{-1}$

**среднее квадратическое**  $\varphi(x) = x^2$

**где медиана и мода?**

**что такое среднее по Коши?**

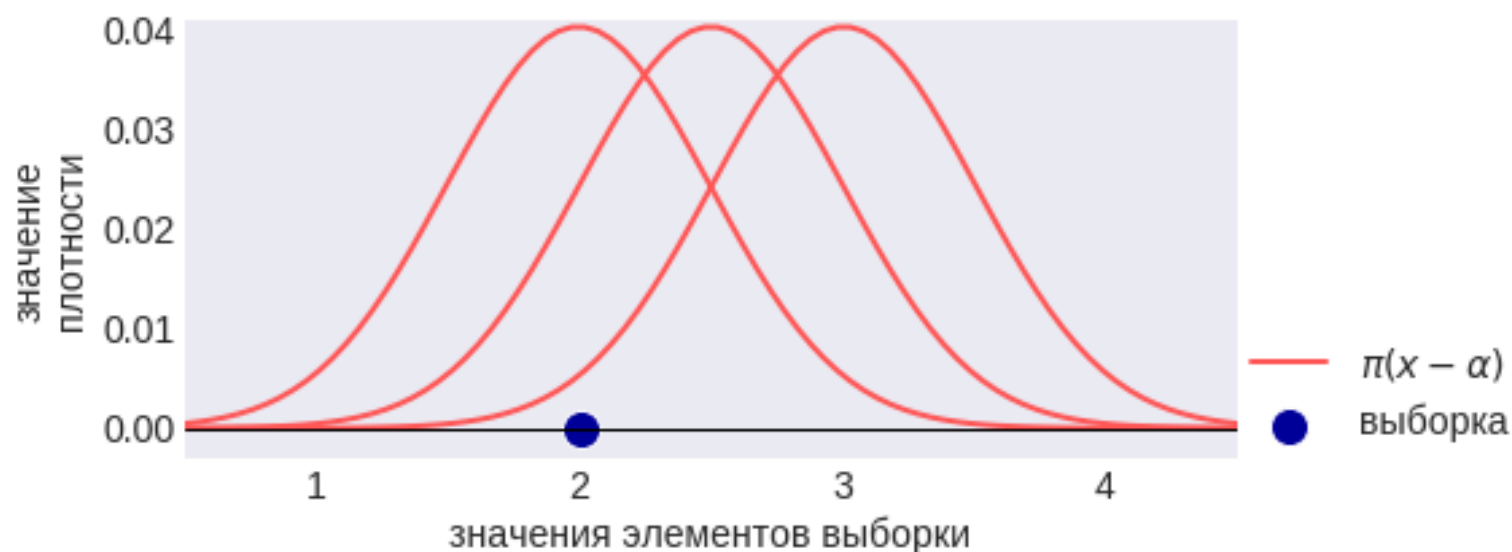
## Оценивание вероятности

тоже, в некотором смысле, усреднение... сейчас объясним

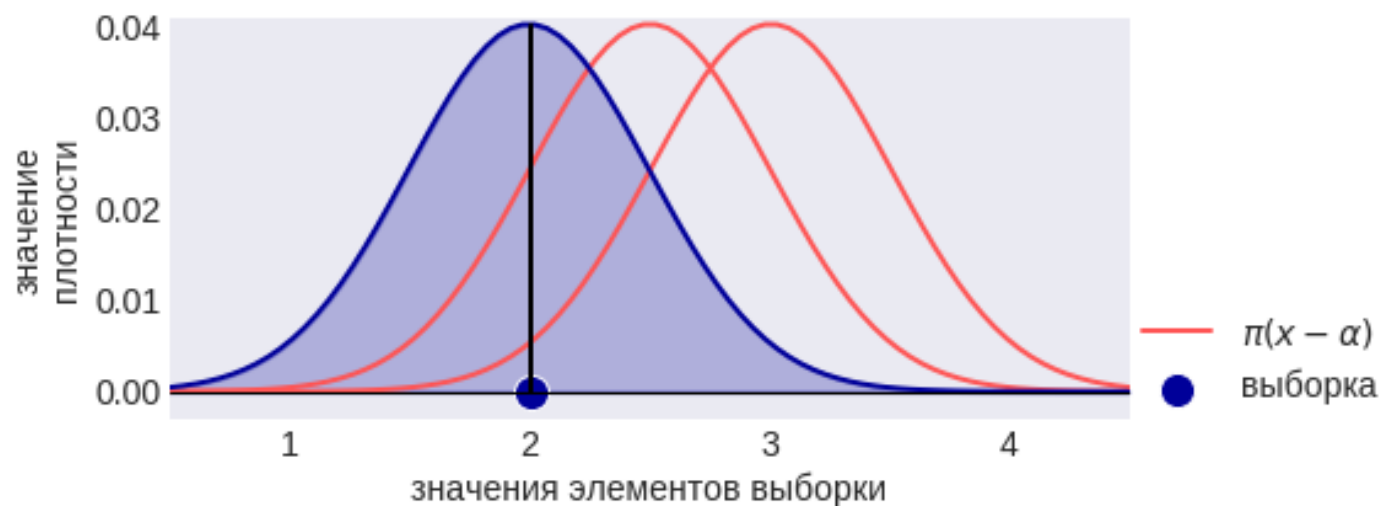
### Метод максимального правдоподобия

Есть выборка  $x_1, \dots, x_n$  какое распределение  $\pi_\alpha(x)$  ?

Пусть  $m = 1$ ,  $\pi_\alpha(x) = \pi(x - \alpha)$  какое распределение выбрать?



## Метод максимального правдоподобия



$$\pi_\alpha(x_1) \rightarrow \max_\alpha$$

Пусть  $m = 2$



## Метод максимального правдоподобия

Пусть  $m = 2$



$$\pi_{\alpha}(x_1) \cdot \pi_{\alpha}(x_2) \rightarrow \max_{\alpha}$$

**Общий случай:**

$$\prod_{i=1}^m \pi_{\alpha}(x_i) \rightarrow \max_{\alpha}$$

**Как максимизируют?**

## Случай биномиального распределения

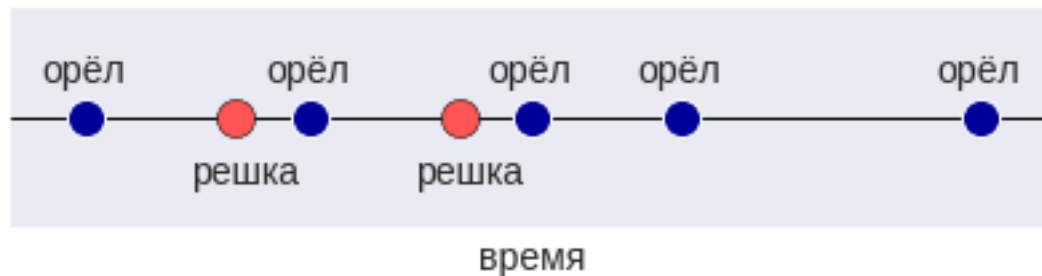
$$\pi_p(x) = \begin{cases} p, & x = 1, \\ 1 - p, & x = 0. \end{cases}$$

$$\Pi = \prod_{i=1}^n \pi_p(x_i) = p^m (1-p)^{n-m} \sim m \log p + (n-m) \log(1-p)$$

$$(\log \Pi)' = \frac{m}{p} - \frac{(n-m)}{1-p} = 0$$

$$p = \frac{m}{n}$$

**Самый очевидный ответ для оценки вероятности!**

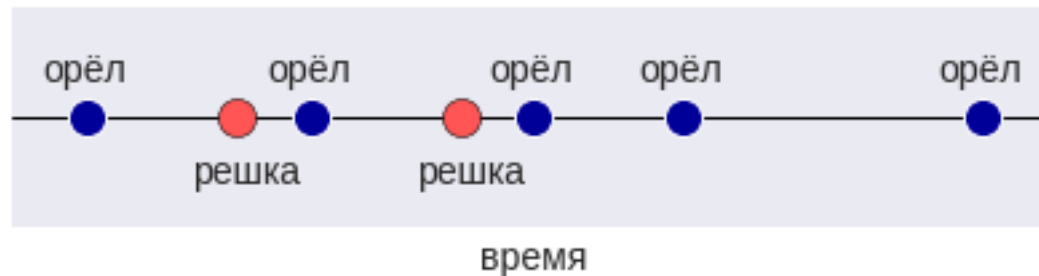


$$p = \frac{5}{5+2} = \frac{5}{7} \approx 0.71$$

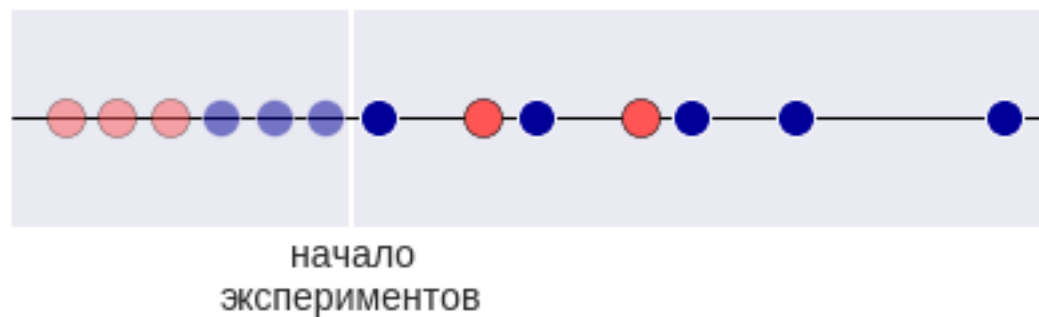


## Оценивание вероятности – сглаживание Лапласа

**тоже, в некотором смысле, усреднение**



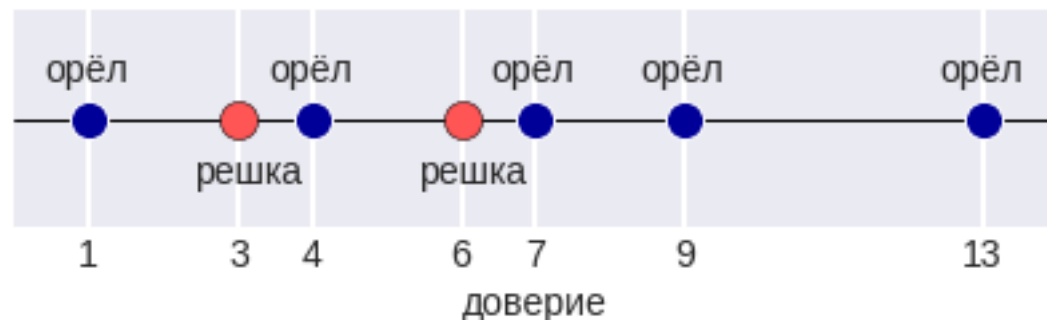
**на практике есть априорная вероятность**



$$\frac{m + \lambda \cdot p}{n + \lambda} = \frac{5 + 3 \cdot 0.5}{5 + 2 + 3} = 0.65$$

## Вторая особенность практики

**Не все эксперименты равнозначны!**



$$\frac{1 + 4 + 7 + 9 + 13}{1 + 3 + 4 + 6 + 7 + 9 + 13} = 0.79$$




## Весовая схема

$$\frac{w_{i_1} + \dots + w_{i_m}}{w_1 + \dots + w_n}$$

**Веса (доверие) возникают даже там, где нет эксперта**

- есть временная ось
- есть «такие же условия»
- есть кластеры (и схожесть вообще)

## Зодиакальный скоринг

Знак зодиака		Сколько представителей знака допускают хотя бы одну просрочку
Овен		35.3%
Дева		35%
Рыбы		34.2%

**где ошибка?**

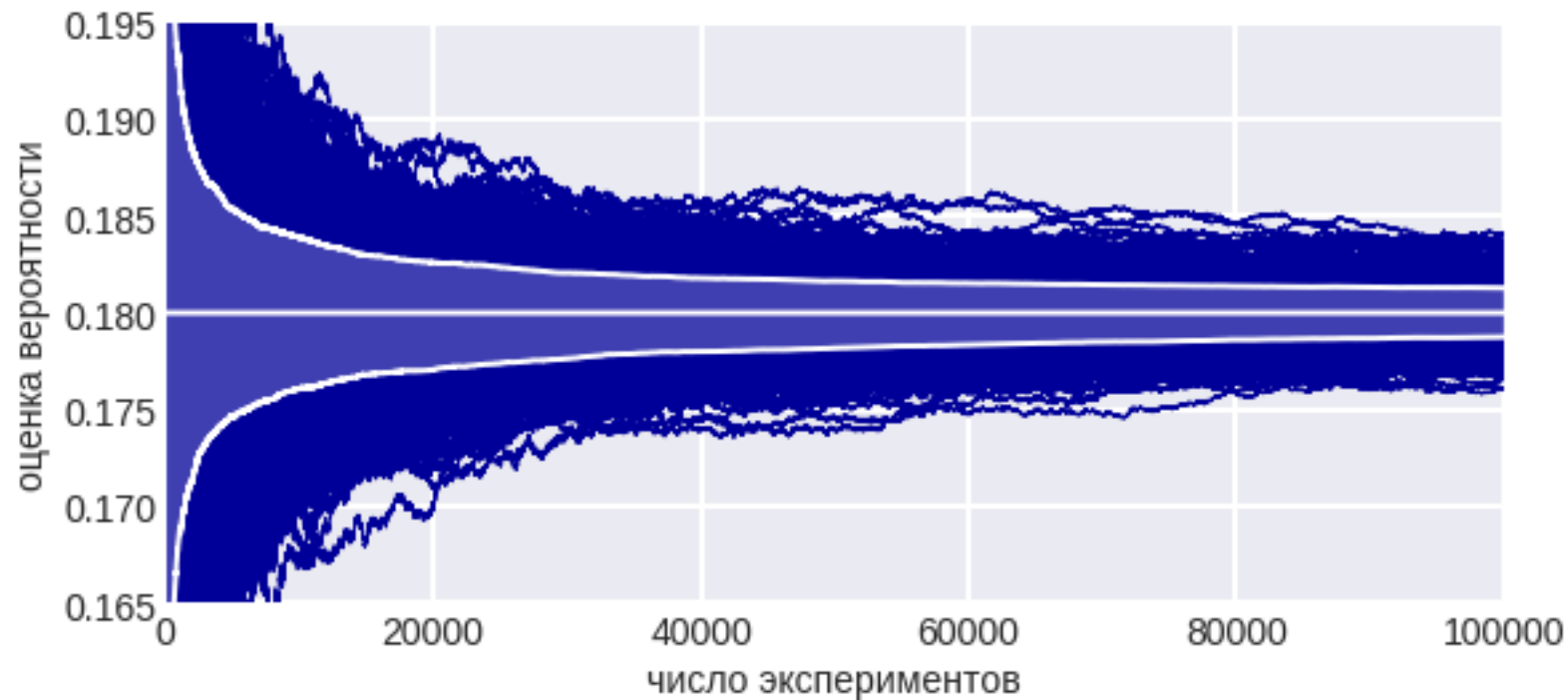
<http://www.banki.ru/news/daytheme/?id=7408493>

<http://moneyman.ru/articles/goroskop-moneyman>

## Что ещё нужно знать про вероятности

### Объёмы выборок

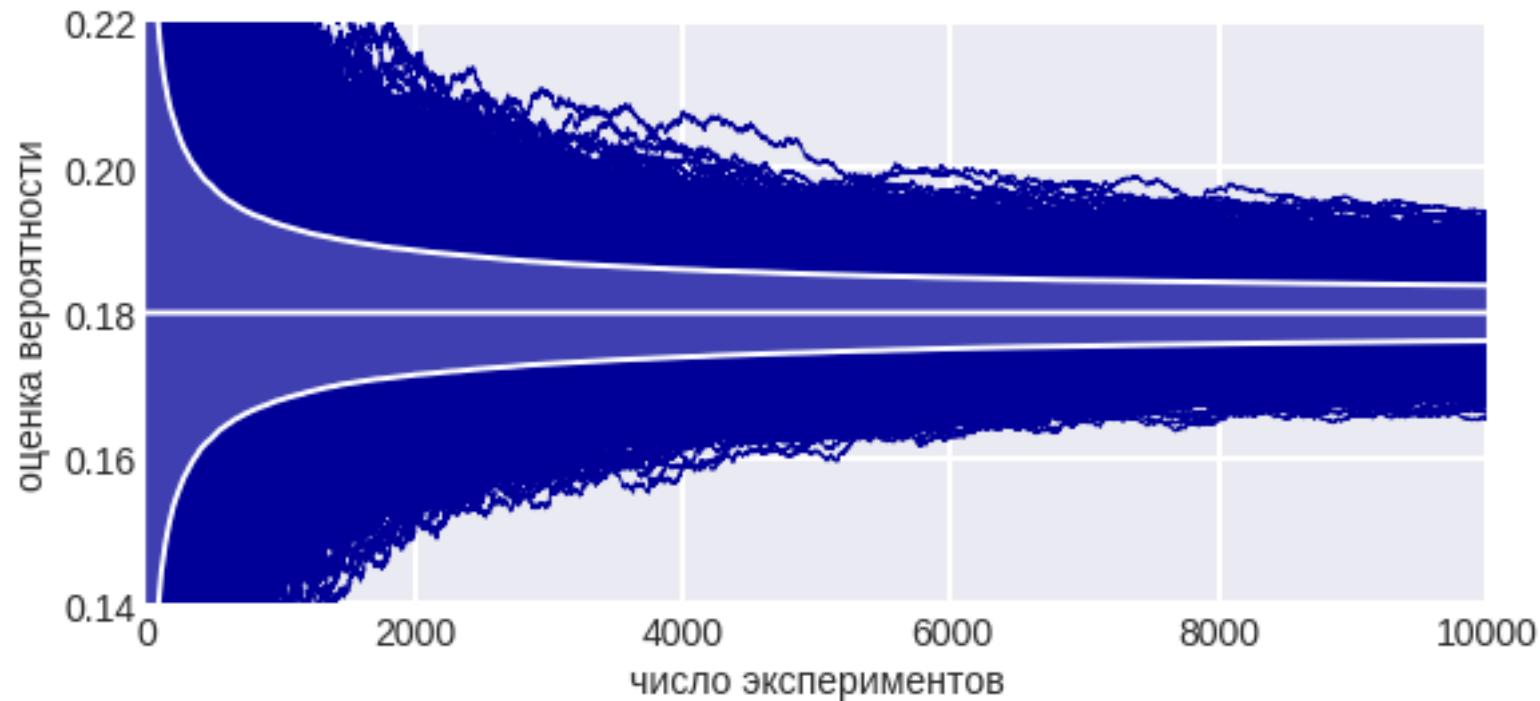
**Оцениваем вероятность в схеме Бернулли (неизвестная  $p=0.18$ )**



**1000 экспериментов**

## Что ещё нужно знать про вероятности

### Объёмы выборок



**Выборки 10000 достаточно, но это чтобы оценить с точность  $\pm 0.01$   
с точностью 99%**

**Д/З так ли это?**

## Что ещё нужно знать про вероятности

**Классика статистики: есть точность,  
а есть вероятность того, что мы оценили с этой точностью**

**Д/З сколько нужно опросить перед выборами людей, чтобы получить  
достоверную оценку общественного мнения?  
что здесь такое «достоверная»?**

### **Зодиакальный скоринг**

- **достаточно ли велика выборка**

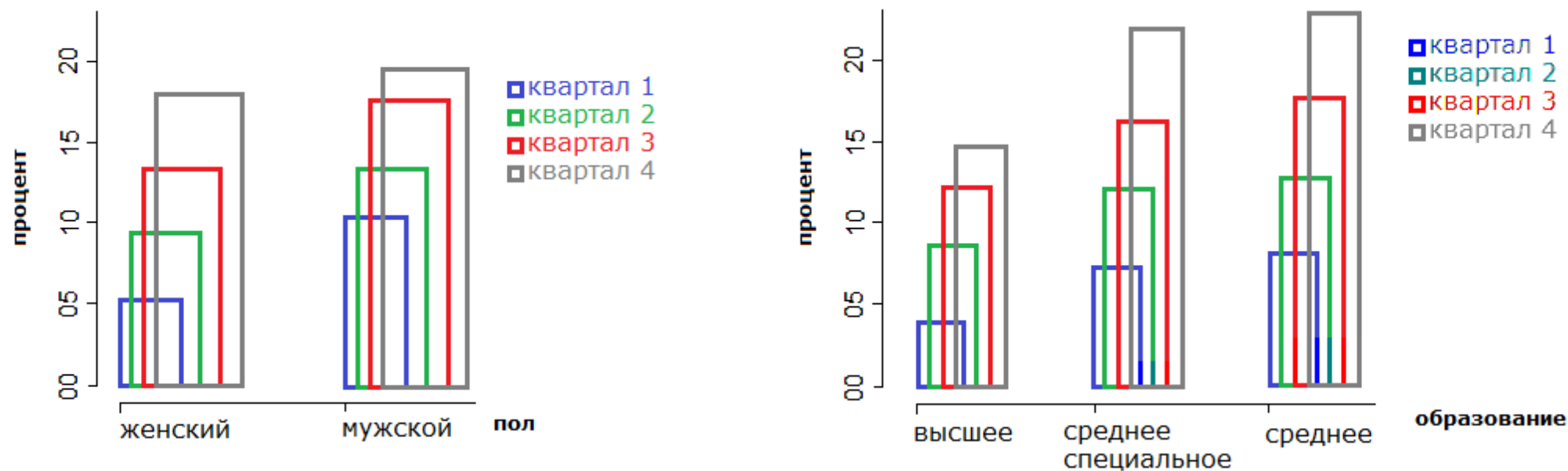
**более 250000 +  $<10\%$  каждого знака + 10% получили микрозаймы**

- **значимы ли отклонения в процентах**
- **насколько закономерности устойчивы  
(ex: не зависят от времени)**

## Эксперименты с банковскими данными

**300000 клиентов**

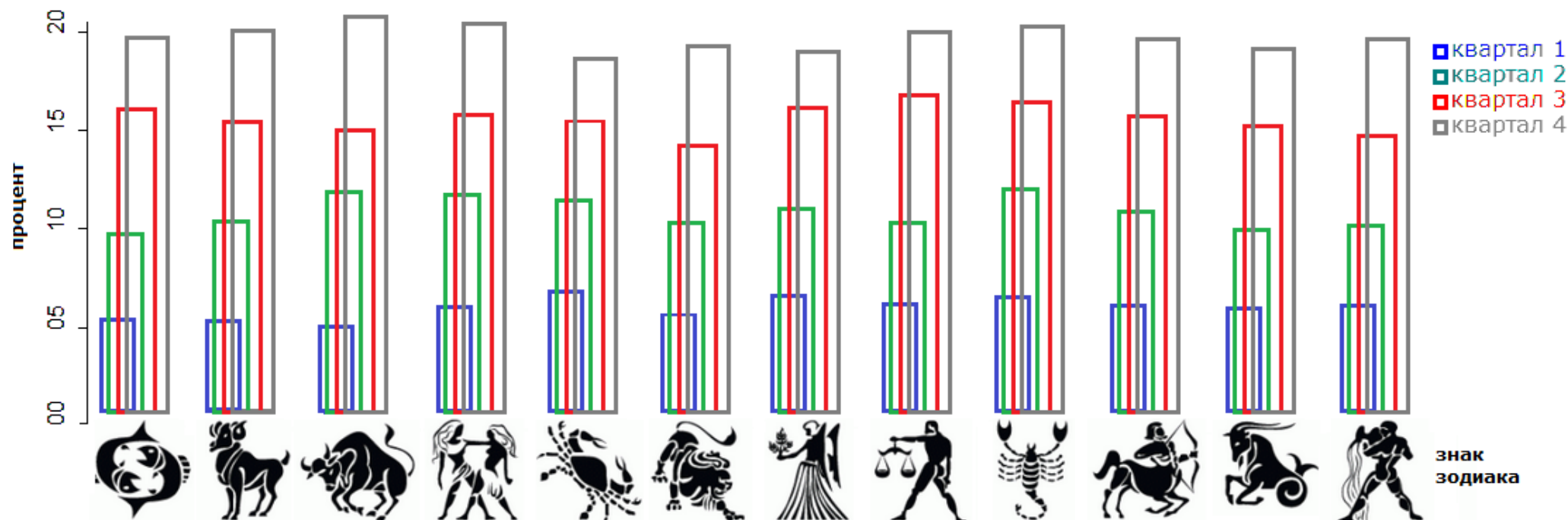
**классические скоринговые признаки**



**Есть устойчивость по кварталам!**

## Эксперименты с банковскими данными

### Неклассические скоринговые признаки



**Нет устойчивости по кварталам!**

**Логическая закономерность тогда является таковой, когда с её помощью можно что-то предсказать!**



## **Эксперименты с банковскими данными**

**Д/З в чём слабость наших аргументов?**

## Итог

**формализаций средних много  
(по Колмогорову + медиана, мода, ...)**

**среднее**

- **формула**
- **решение задачи оптимизации**
- **ответ некоторого алгоритма**

**важны априорные знания (сглаживание Лапласа)!**

**Не все объекты равноценны (весовые схемы)**

**Объём выборки для правильных выводов**

**Д/З другие способы обобщения медианы...**

## Задача

# Прогнозирование визитов покупателей супермаркетов и сумм их покупок

<http://www.kaggle.com/c/dunnhumbychallenge/>

## Международное соревнование «dunnhumby's Shopper Challenge»

Опишем лучший алгоритм из 287

#	Team Name	\$10,000 • 279 teams	Score ?	Entries
1	D'yakonov Alexander (MSU, Moscow, Russia) *	18.83	68	
2	NSchneider *	18.67	20	
3	Ben Hamner *	18.57	19	
4	William Cukierski	18.44	75	



**Дано:** статистика визитов

**Предсказать:** день **первого** визита + сумму покупки  
с точностью до 10 \$

покупатель, дата визита, сумма

56, 2011-06-30, 35.01

56, 2011-06-08, 35.17

56, 2011-07-10, 24.12

56, 2011-07-12, 7.73

57, 2011-05-13, 29.38

57, 2011-05-19, 41.00

...

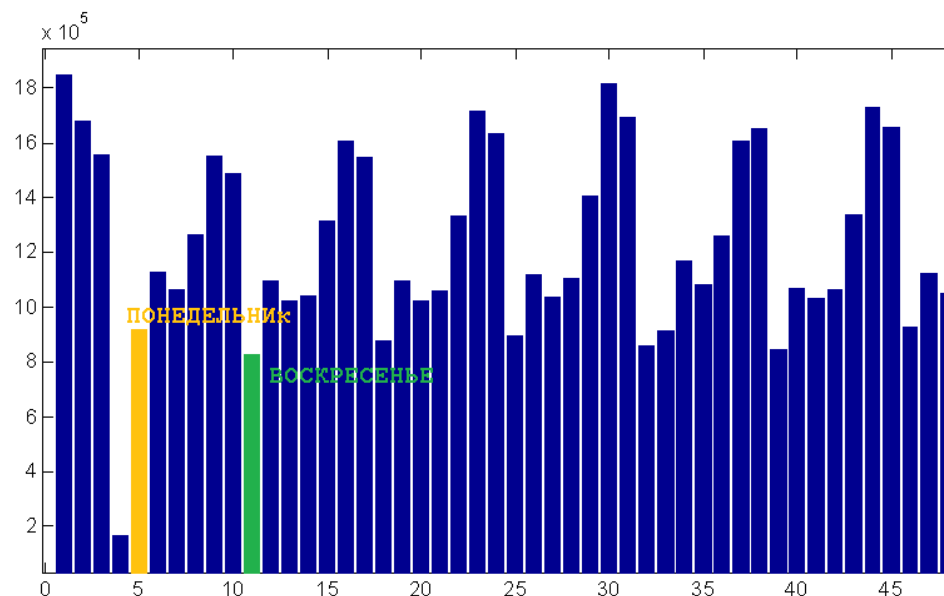
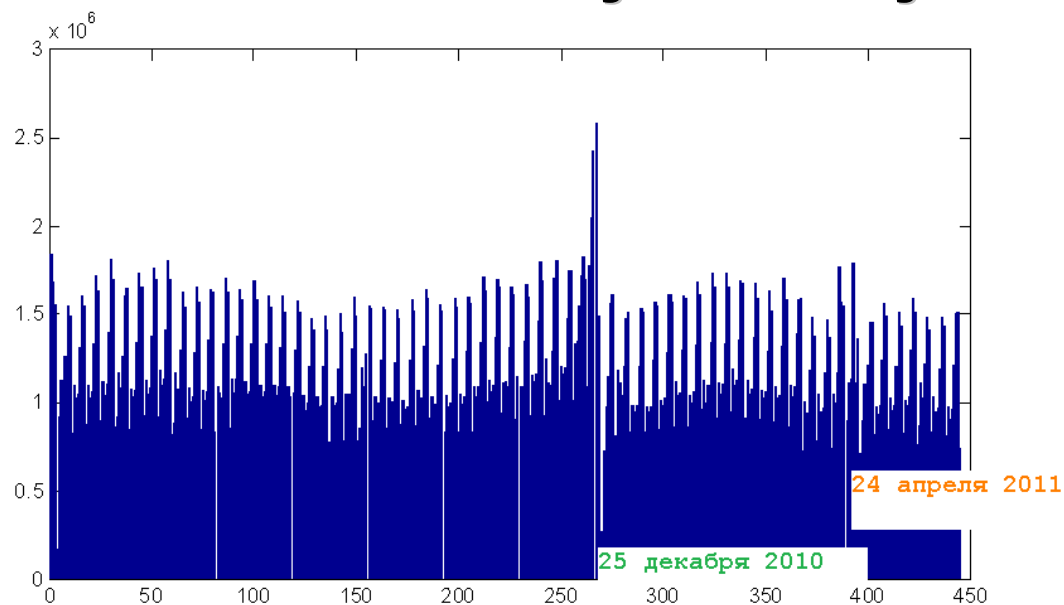
>100000 клиентов **customers**

T = 1 год

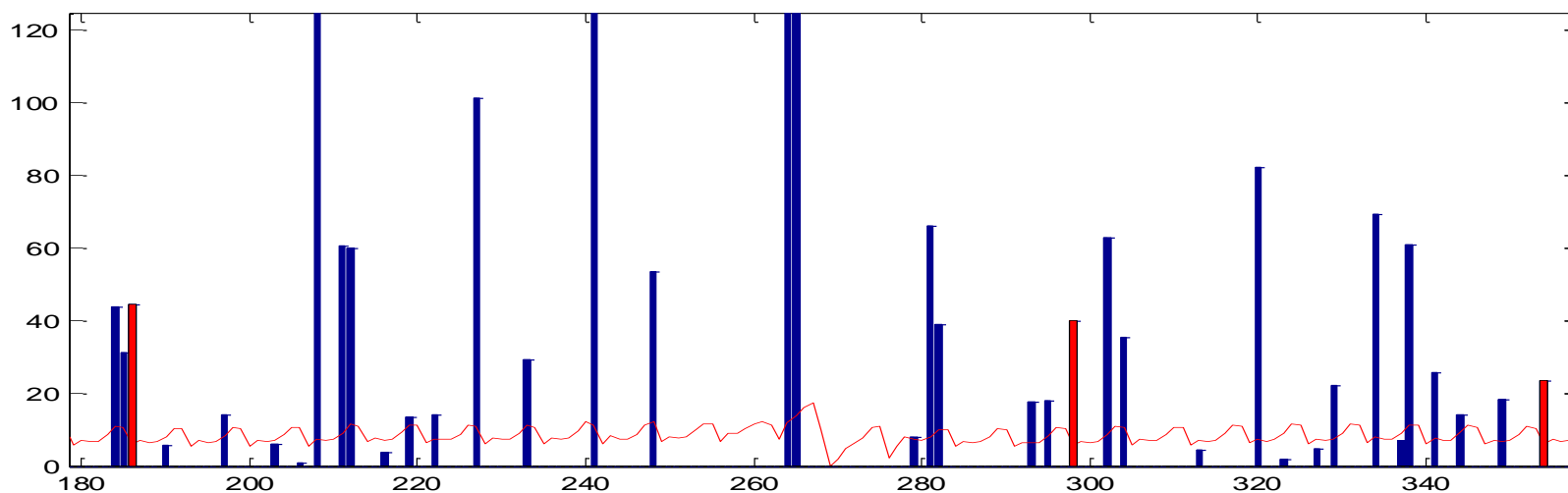
**Статистика визитов одного клиента:**

Март	Март	Март	Март	Март	Март	Март	Март	Март	Март	Март	Апрель	Апрель	Апр
21	22	23	24	25	26	27	28	29	30	31	1	2	3
5\$		45\$	5\$				35\$		60\$		?	?	?

## Суммы покупок всех клиентов



## Покупки одного клиента



## Предположение:

**Все клиенты независимы**

**Будем анализировать каждого клиента отдельно**

## Разбиение на недели:

Март 21	Март 22	Март 23	Март 24	Март 25	Март 26	Март 27	Март 28	Март 29	Март 30	Март 31	Апрель 1	Апрель 2	Апрель 3
5\$		45\$	5\$				35\$		60\$		?	?	?
неделя				неделя									

				Март 22	Март 23	Март 24
				5\$	45\$	5\$
Март 25	Март 26	Март 27	Март 28	Март 29	Март 30	Март 31
				35\$	60\$	
Апрель 1	Апрель 2	Апр				
?	?	?				



200			42		50	
10						
62			40		45	5
			35		60	

### Матрица разбивки по неделям:

The diagram illustrates the transformation of a 7x7 grid from a sparse state to a dense state. The left grid shows a sparse state with yellow cells at (1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (1,7), (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), (2,7), (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), (3,7), (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), (4,7), (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), (5,7), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6), (6,7), (7,1), (7,2), (7,3), (7,4), (7,5), (7,6), (7,7). The right grid shows a dense state with red cells at (1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (1,7), (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), (2,7), (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), (3,7), (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), (4,7), (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), (5,7), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6), (6,7), (7,1), (7,2), (7,3), (7,4), (7,5), (7,6), (7,7).

## Сработало устранение пустых недель...

## Вероятностная модель поведения клиента

**Матрица затрат:**  $S = \| s_{ij} \|_{d \times 7}$

**Матрица визитов:**  $V = \|v_{ij}\|_{d \times 7}$ ,  $v_{ij} = 1 \Leftrightarrow s_{ij} > 0$ .

## Вероятности визитов

оценки вероятностей...

100						10
					18	
52			50			
200			42		50	
10						
62			40		45	5
			35		60	

$5/N$     $0$     $0$     $4/N$     $0$     $4/N$     $2/N$   
 ▲   ▲  
**вероятности визитов**

$$5/N \quad ((N-5)/N) \cdot 0 = 0$$

$$((N-5)/N) \cdot 1 \cdot 0 = 0$$

$$((N-5)/N) \cdot 1 \cdot 1 \cdot (4/N) \quad \dots$$

**вероятности первых визитов**

первых визитов

$$p_1$$

$$\tilde{p}_1 = p_1$$

$$p_2$$

$$\tilde{p}_2 = (1 - p_1) p_2$$

...

...

$$p_7$$

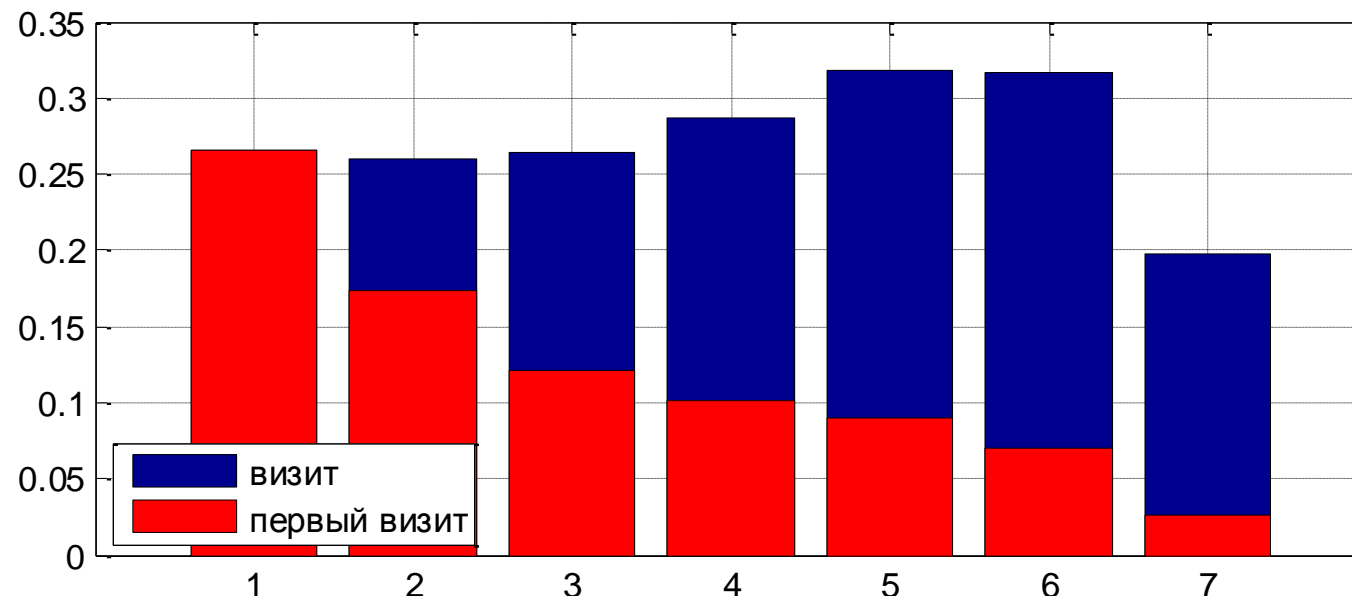
$$\tilde{p}_7 = \prod_{i=1}^6 (1 - p_i) p_7$$

**Находим максимум вероятностей!**

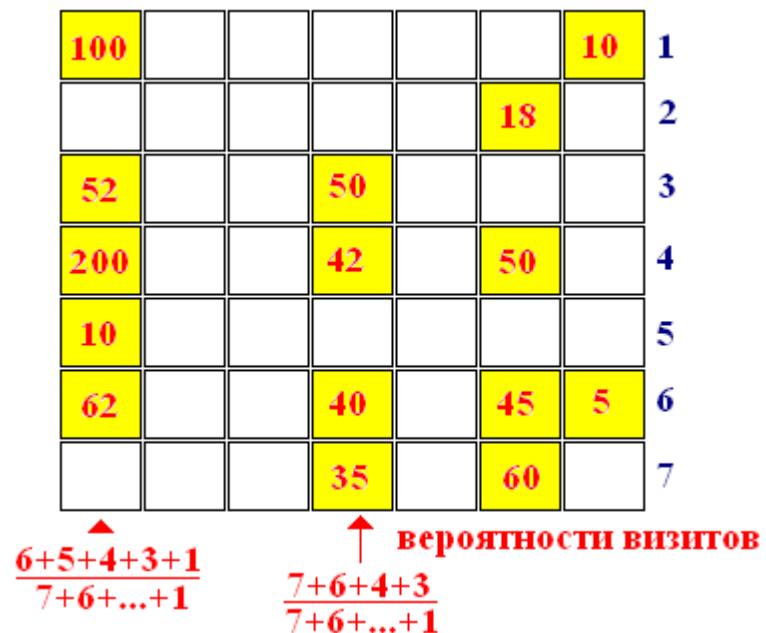
**Предположение: Каждый клиент обязательно посетит магазин в течение следующей недели.**



## Процент визитов и первых визитов на неделе



## «Более свежие» данные о клиенте важнее устаревших!



## Весовые схемы!

**Взвешенная схема оценки вероятности:**

$$p_j = \sum_{i=1}^d w_i v_{ij},$$

$$w_1 \geq w_2 \geq \dots \geq w_d \geq 0, \sum_{i=1}^d w_i = 1.$$

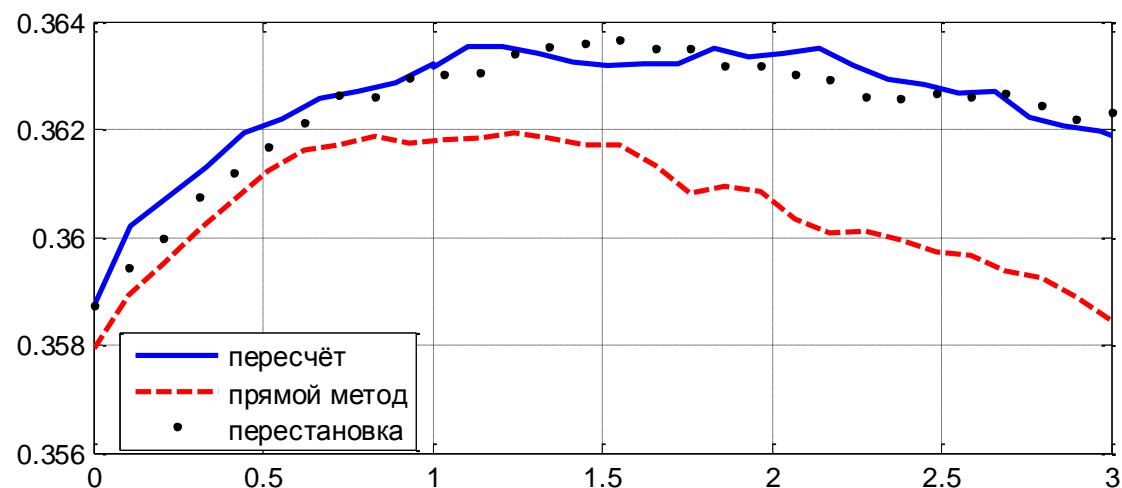
**Способы**

$$w_i^N = \left( \frac{d-i+1}{d} \right)^\delta, \quad i \in \{1, 2, \dots, d\},$$

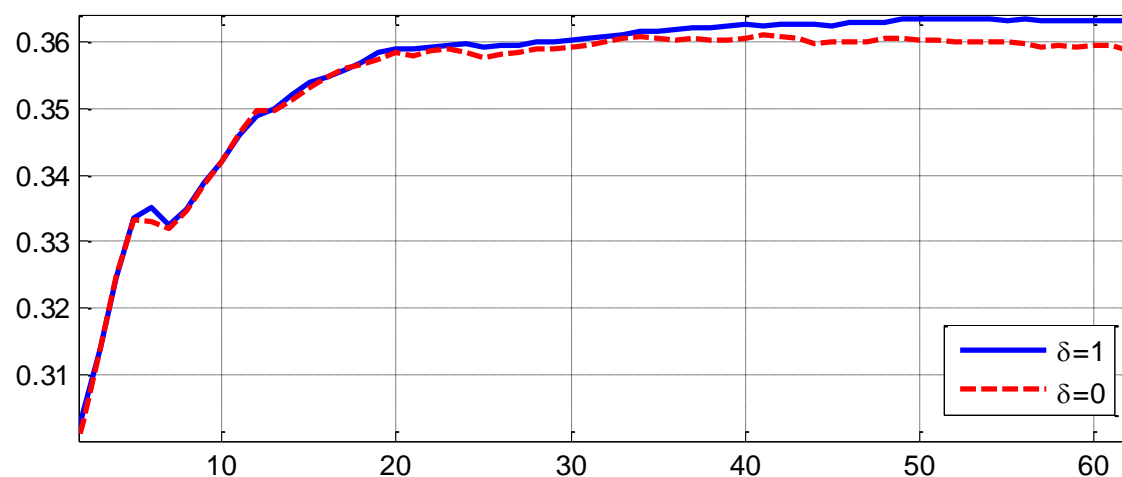
$$w_i = \frac{w_i^N}{\sum_{i=1}^d w_i^N}, \quad i \in \{1, 2, \dots, d\}. \text{ [просто нормировка]}$$

**Параметр**  $\delta \in [0, +\infty)$ .

## Веса – от равномерных к «агрессивным»

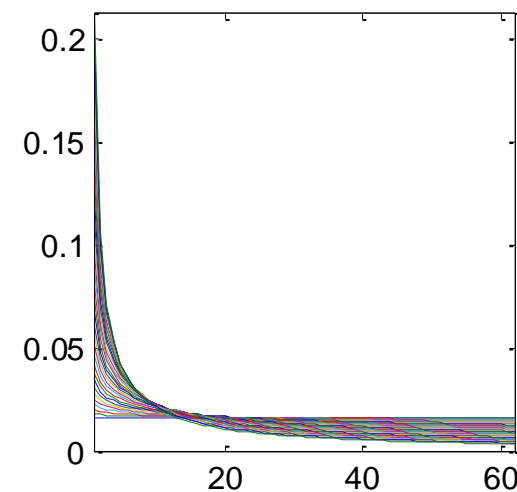
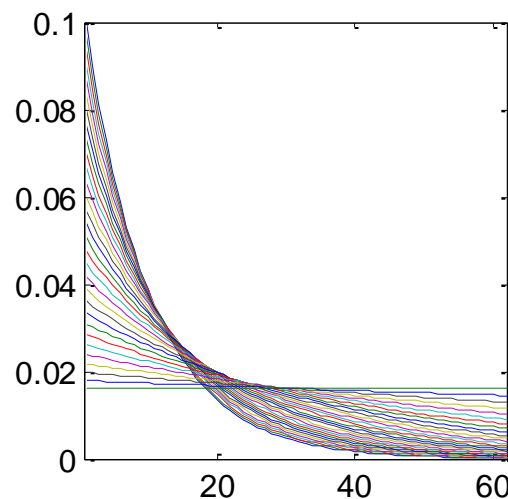
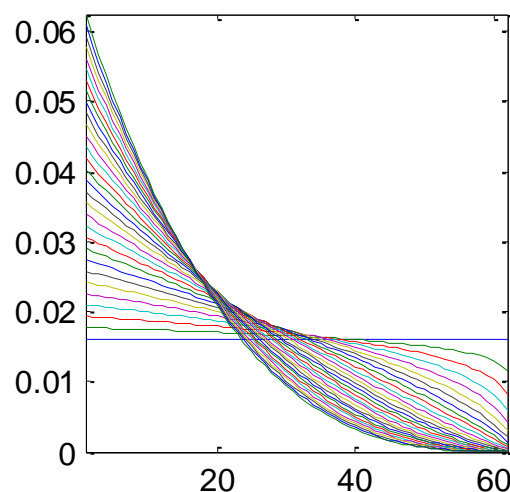


## Зависимость качества прогноза от степени $\delta$



## Зависимость качества прогноза от числа учитываемых недель

## Три разные весовые схемы



**вес недели в зависимости от её номера**

$$w_i^N = \left( \frac{d-i+1}{d} \right)^\delta$$

$$\delta \in [0, +\infty)$$

$$w_i^N = \lambda^i$$

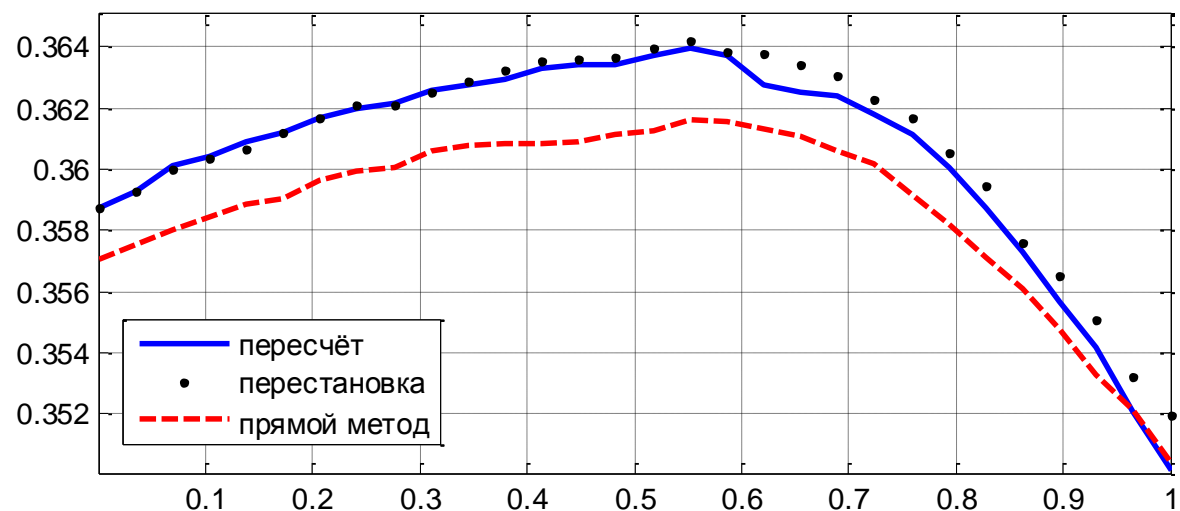
$$\lambda \in (0, 1]$$

$$w_i^N = \frac{1}{i^\gamma},$$

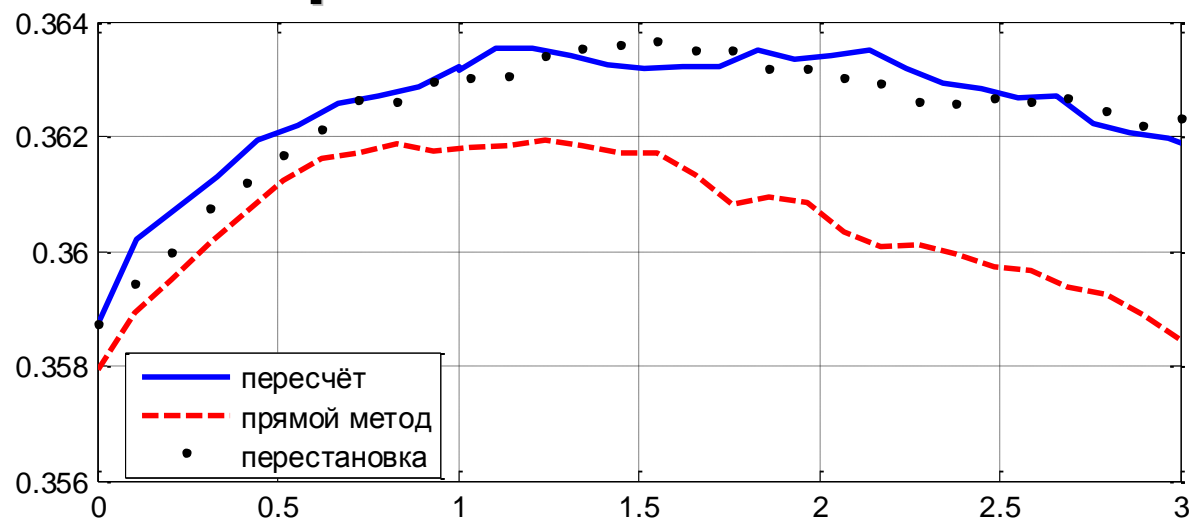
$$\gamma \in [0, +\infty)$$

**Вопрос: какие ещё?**

## Принципиально всё одинаково...



## Третья весовая схема



## Первая весовая схема

## Два способа оценки вероятности первого визита

### Прямой метод

$$\tilde{p}_j^2 = \frac{1}{d} |\{i \in \{1, 2, \dots, d\} : v_{i1} = \dots = v_{i,j-1} = 0, v_{ij} = 1\}|$$

Более естественный, **но хуже!**

### матрица первых визитов

$$V' = \|v'_{ij}\|_{d \times 7}$$

$$\tilde{p}_j^2 = \sum_{i=1}^d w_i v'_{ij}$$



$$\begin{array}{cccccc} \frac{1}{6} & \frac{2}{6} & \frac{1}{6} & \frac{2}{6} & & & \\ \frac{1}{6} & \frac{3}{6} & \frac{1}{6} & \frac{5}{6} & & \frac{1}{6} & \frac{2}{6} \end{array}$$

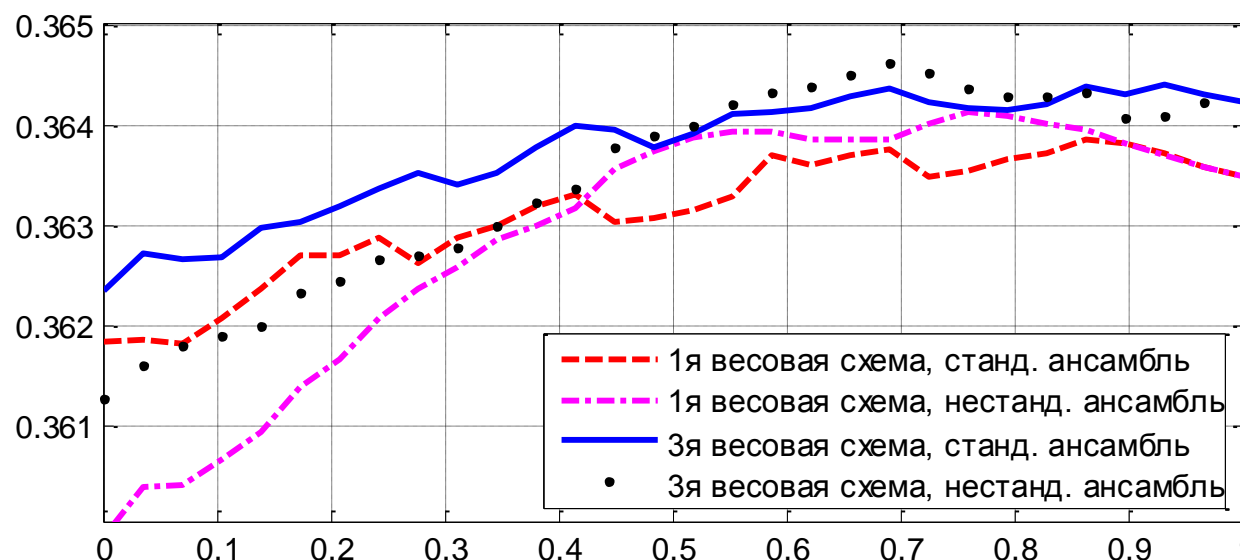
## Ансамблирование

**«Стандартный ансамбль» – взять выпуклую комбинацию:**

$$\tilde{p}_j = \alpha \tilde{p}_j^1 + (1 - \alpha) \tilde{p}_j^2, \quad \alpha \in [0, 1].$$

**«Нестандартный ансамбль»**

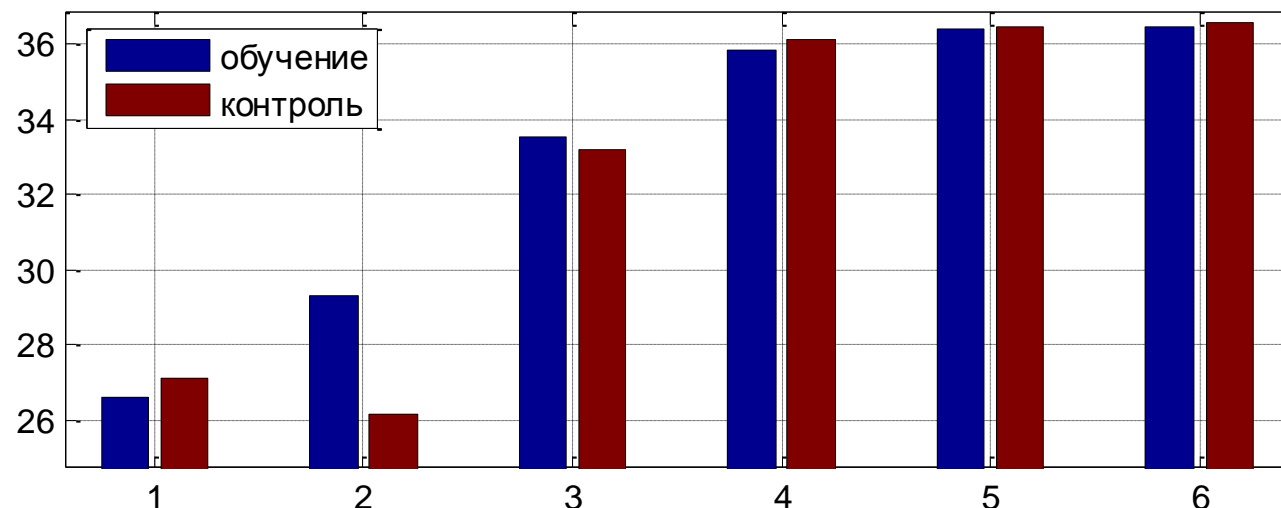
$$\alpha p_j + (1 - \alpha) \tilde{p}_j^2 = \alpha \sum_{i=1}^d w_i v_{ij} + (1 - \alpha) \sum_{i=1}^d w_i v'_{ij} = \sum_{i=1}^d w_i (\alpha v_{ij} + (1 - \alpha) v'_{ij})$$



**Качество ансамблирования от параметра  $\alpha \in [0, 1]$**



## Про переобучение

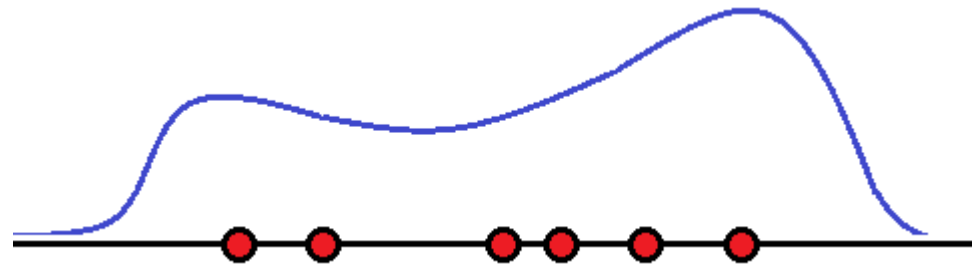


### Качество на обучении и отложенном контроле для шести алгоритмов

1. Константный («клиент придёт на следующий день»),
2. Визит клиента как на прошлой неделе,
3. Вероятности (\*) оценены по последним 5 неделям,
4. Вероятности оценены по всем неделям,
5. Оптимальные значения весов,
6. Оптимальное нестандартное ансамблирование.

**Не усложнение, а сглаживание!**

## Восстановление плотности



**Какие методы знаете?**

## **Восстановление плотности**

### **1. Параметрические**

**Плотность известна с точностью до параметров**

### **2. Непараметрические**

**Вид плотности не известен**

### **3. Восстановление смесей**

**Плотность = сумме плотностей**

# Непараметрические методы восстановления плотности

## Метод окон Парзена:

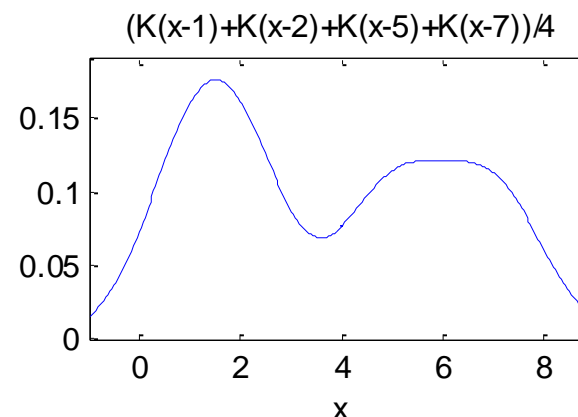
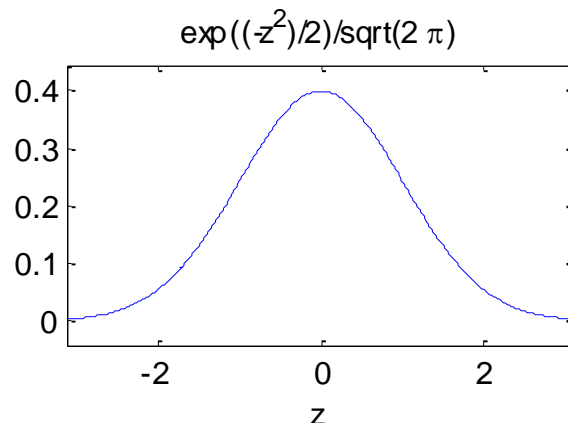
**Выборка**  $x^1, \dots, x^m$   
**в пространстве**  $\mathbf{R}^d$

$$\frac{1}{mh} \sum_{i=1}^m K\left(\frac{x - x^i}{h}\right),$$

**где**  $K(x)$  – **функция окна.**

$$K((z_1, \dots, z_d)) = \begin{cases} 1, & \forall j \in \{1, 2, \dots, d\} \mid |z_j| \leq 0.5 \\ 0, & \text{иначе.} \end{cases}$$

$$K(\tilde{z}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\tilde{z}^T \tilde{z}}{2}\right)$$



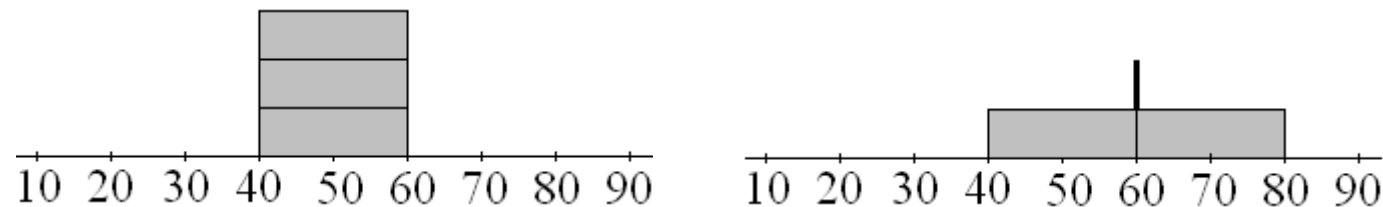
## Предсказание суммы покупки

= непараметрическое восстановление плотности по Парзену

«Суммы ступенек» при покупках

50, 50, 50

50, 70

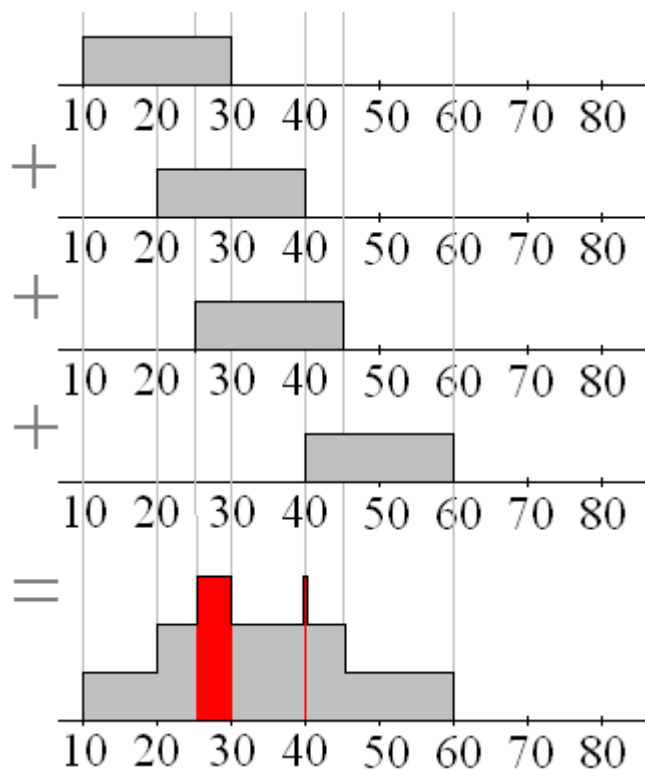


**Наилучшая стратегия предсказания суммы  
при условии, что пользователь  
ведёт себя как раньше**

**т.е. это оценка среднего**

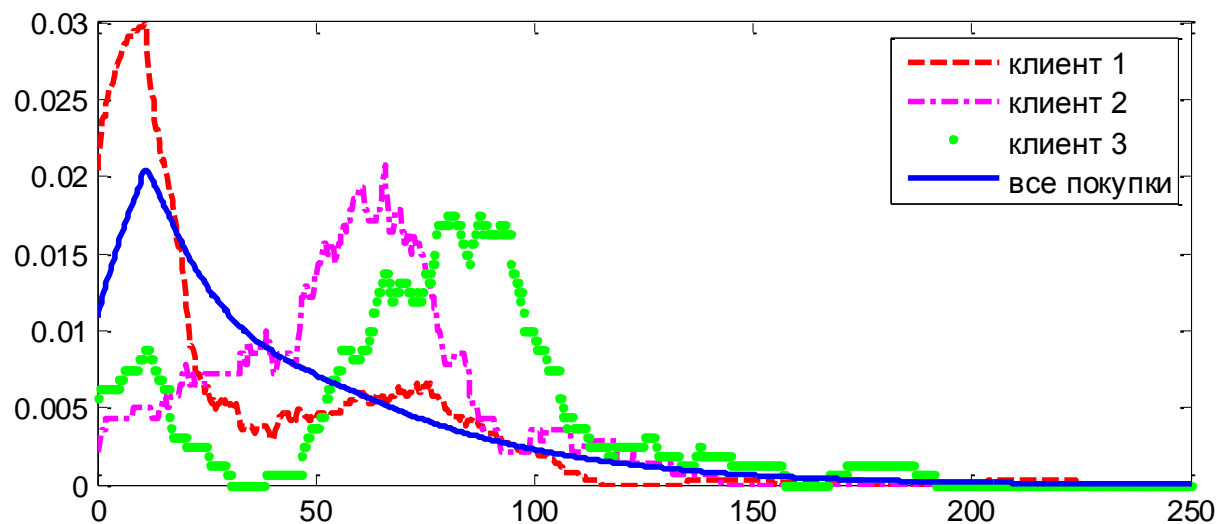
## Прогноз с помощью моды

«Суммы ступенек» при покупках **20, 30, 35, 50** –

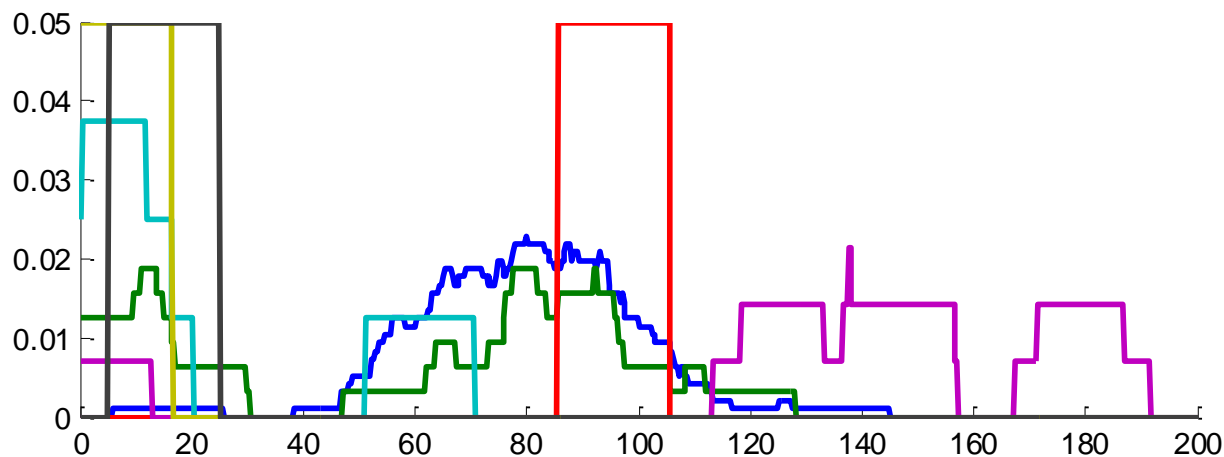


максимум достигается на отрезке **[25, 30]** и в точке **40**.

## Как выглядят плотности



## Плотности распределения покупок



## Плотности покупок одного пользователя в разные дни недели

**И здесь сделаем весовую схему!**

$$f(x) = \frac{1}{m} \sum_{i=1}^m K(|s_i - x|)$$

$$2 \int_0^{+\infty} K(x) dx = 1.$$

$$K(|s - x|) = \begin{cases} 1/2\varepsilon, & |s - x| \leq \varepsilon, \\ 0, & |s - x| > \varepsilon. \end{cases}$$

**Весовая схема:**

$$f(x) = \sum_{i=1}^m w_i K(|s_i - x|)$$



## Весовая схема учёт времени, дня недели

Пусть  $s_1, \dots, s_m$  – все упорядоченные покупки пользователя,  
 $s'_1, \dots, s'_{m'}$  – покупки, сделанные в этот день недели.

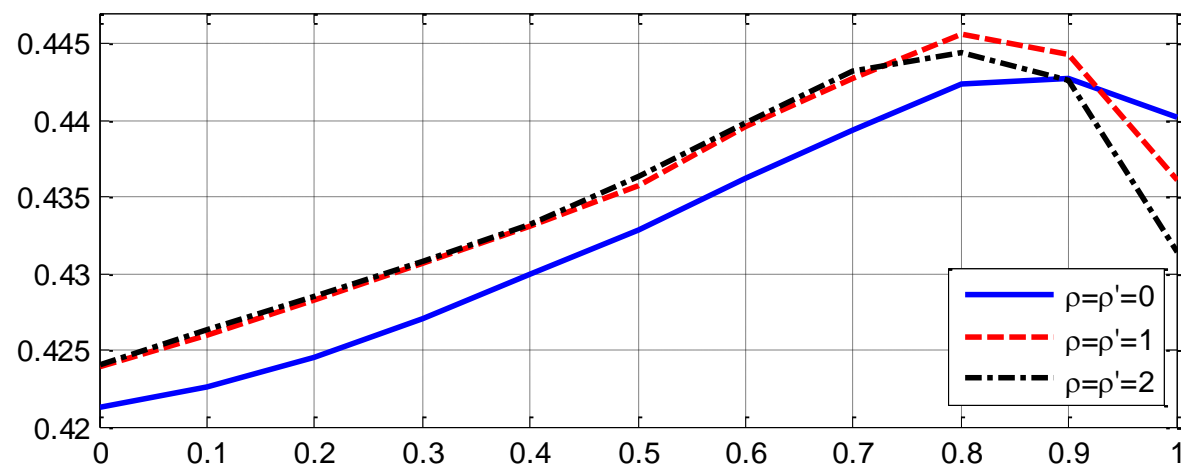
Плотность будем восстанавливать для расширенного набора  
 $s'_1, \dots, s'_{m'}, s_1, \dots, s_m$ .

**Веса:**

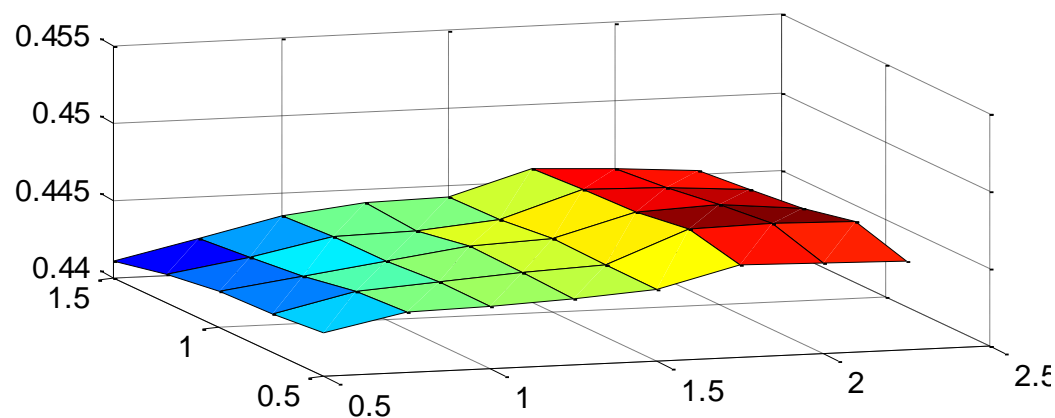
$$s'_i \leftrightarrow \beta \frac{(m' - i + 1)^{\rho'}}{\sum_{j=1}^{m'} j^{\rho'}}$$

$$s_i \leftrightarrow (1 - \beta) \frac{(m - i + 1)^{\rho}}{\sum_{j=1}^m j^{\rho}}$$

## Весовая схема



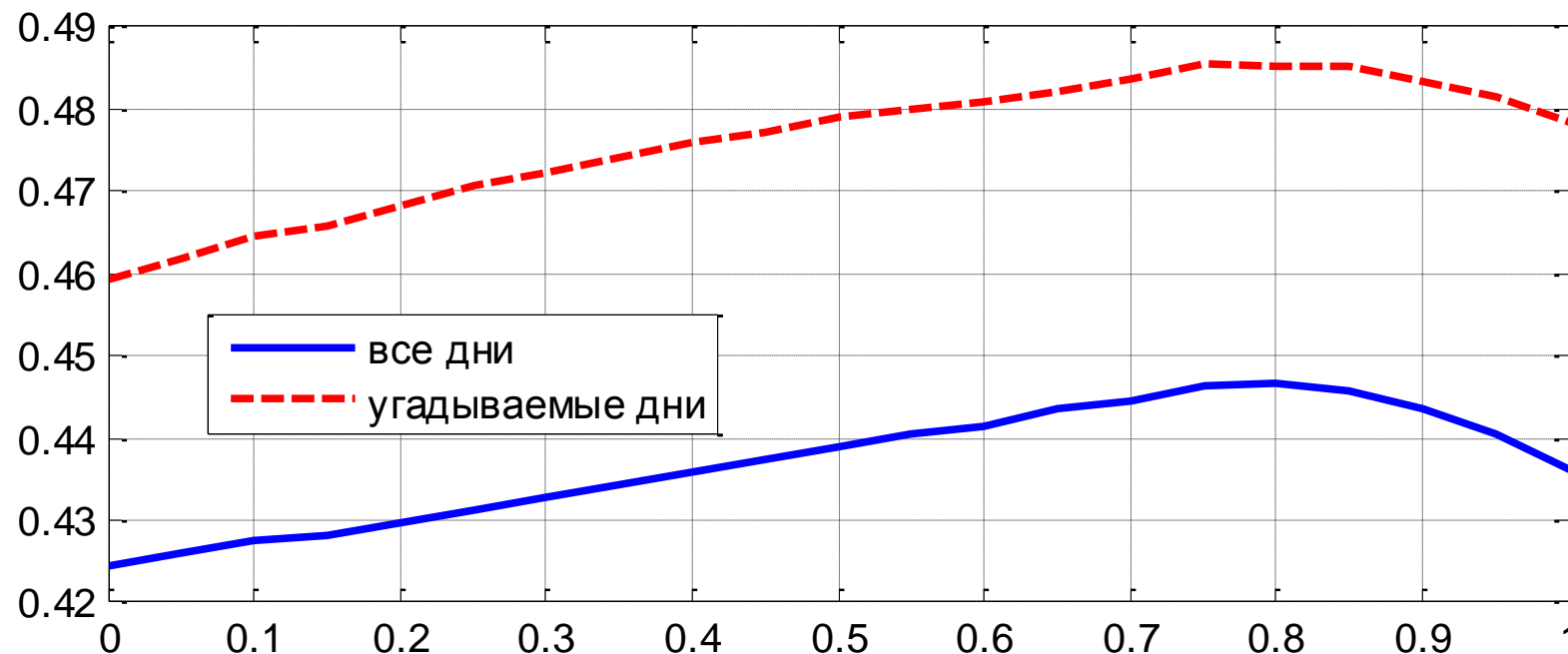
**Качество прогноза суммы покупок от параметра  $\beta$**



**Качество прогноза в зависимости от степеней при  $\beta = 0.8$**

## Как настраивать, точнее где...

- на всей выборке
- на угадываемых днях (на остальных – бесполезно для функционала)



**Качество прогноза суммы покупок  
от параметра  $\beta$  при  $\rho = 0.7$ ,  $\rho' = 1.6$ .**

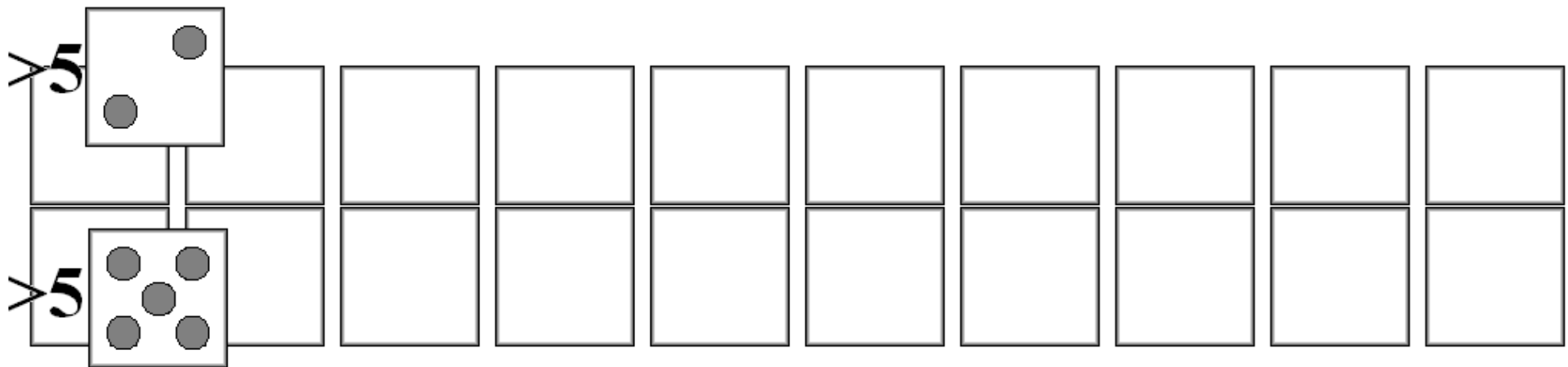
## Улучшение алгоритма

Есть:

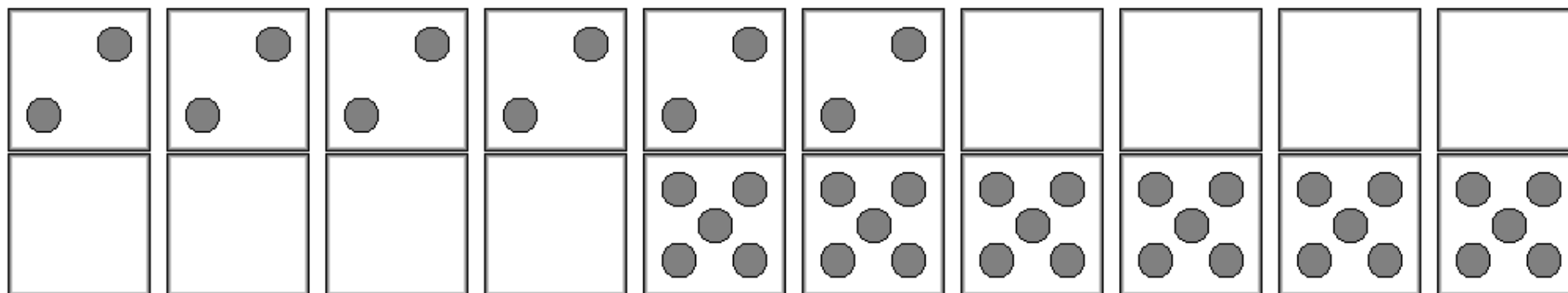
- метод предсказания даты визита (вероятностный пересчёт)
- метод предсказания суммы покупки (непараметрическое восстановление)

**Можно ли так осуществить прогноз?**

**Все прогнозировали так...**



## Почему метод работает не очень хорошо...



«И» в условии не означает «И» в решении

Найти день **И** сумму.

**Понедельник:** 10\$, 50\$, 220\$, 100\$, 310\$, 5\$, 250\$, 75\$, 500\$

**Вторник:** 40\$, 42\$, 40\$

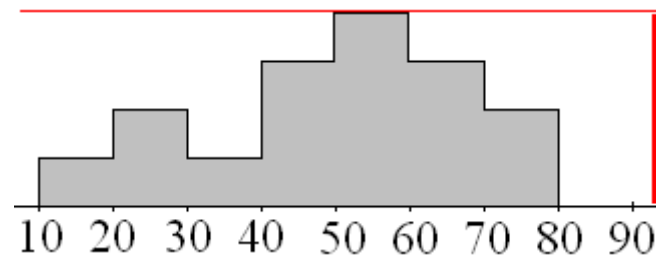
(вероятность угадать день) \* (вероятность угадать сумму)

$$0.9 * 0.1 = 0.09$$

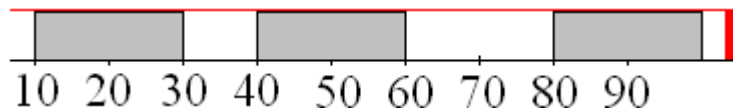
$$0.1 * 1 = 0.1 \text{ выгоднее ставить на вторник}$$

**Надо: вычислить вероятность угадывания дня и суммы**

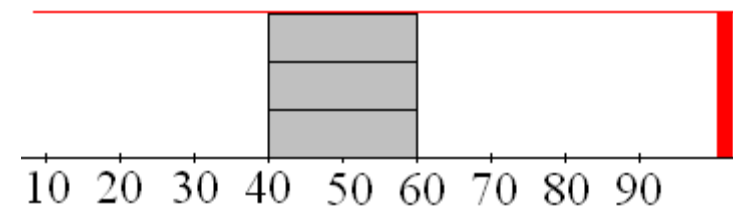
## Как вычислить стабильность поведения клиента?



**высота графика плотности**



**низкая стабильность**



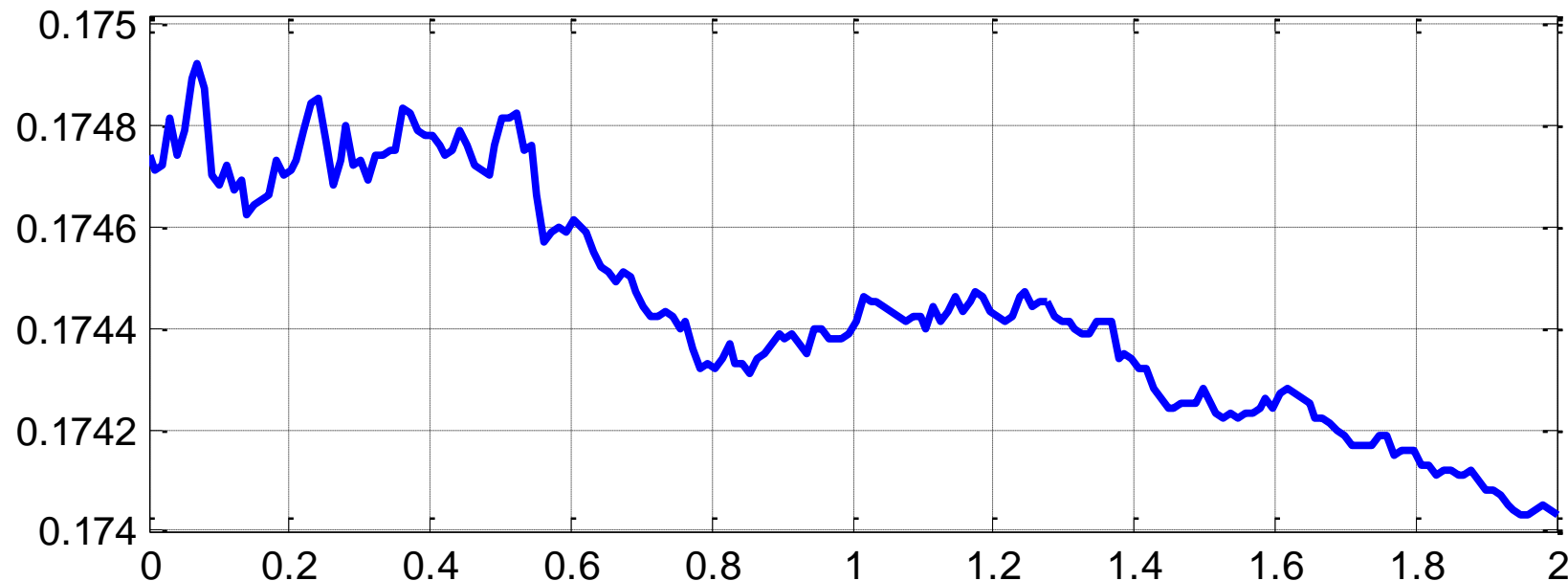
**высокая стабильность**

**учёт стабильности = улучшение результата**

## Неполный учёт стабильности

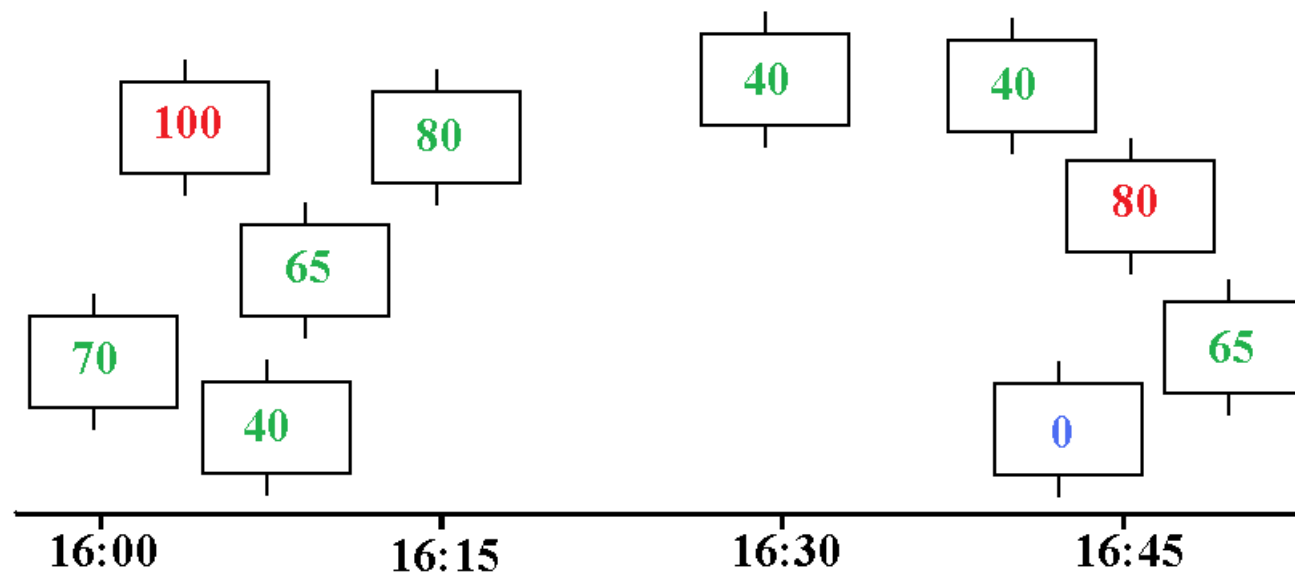
$$\tilde{p}_j(q_j + h) \rightarrow \max_j$$

**это и регуляризация  
и ансамблирование**  $(\underbrace{\tilde{p}_j q_j}_{\max} + h \underbrace{\tilde{p}_j}_{\max})$



**Качество предсказания поведения в зависимости от параметра  $h$ .**

## Пример: задача о пробках

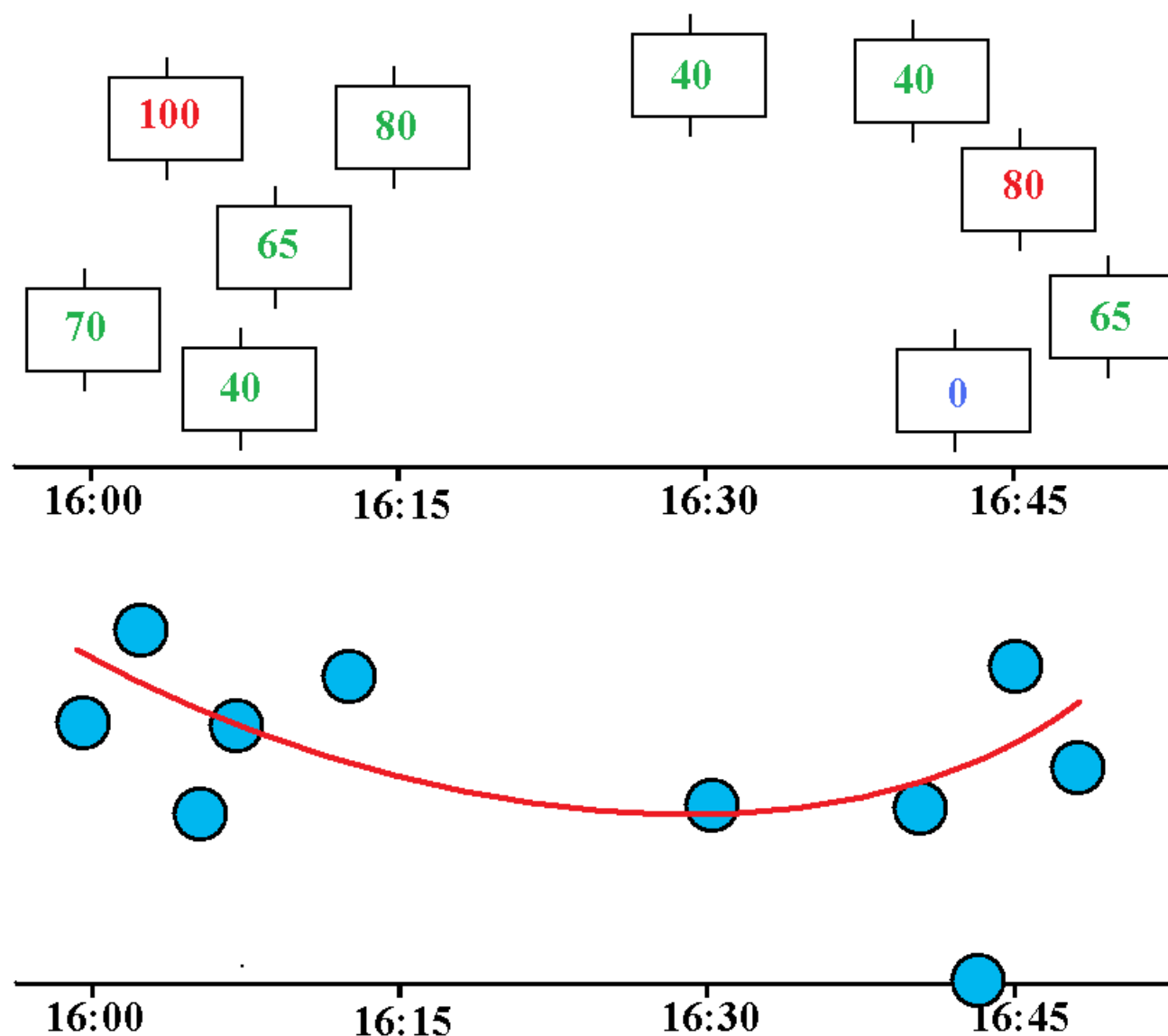


**Нужно знать «среднюю» скорость на дороге  
в каждый момент времени**

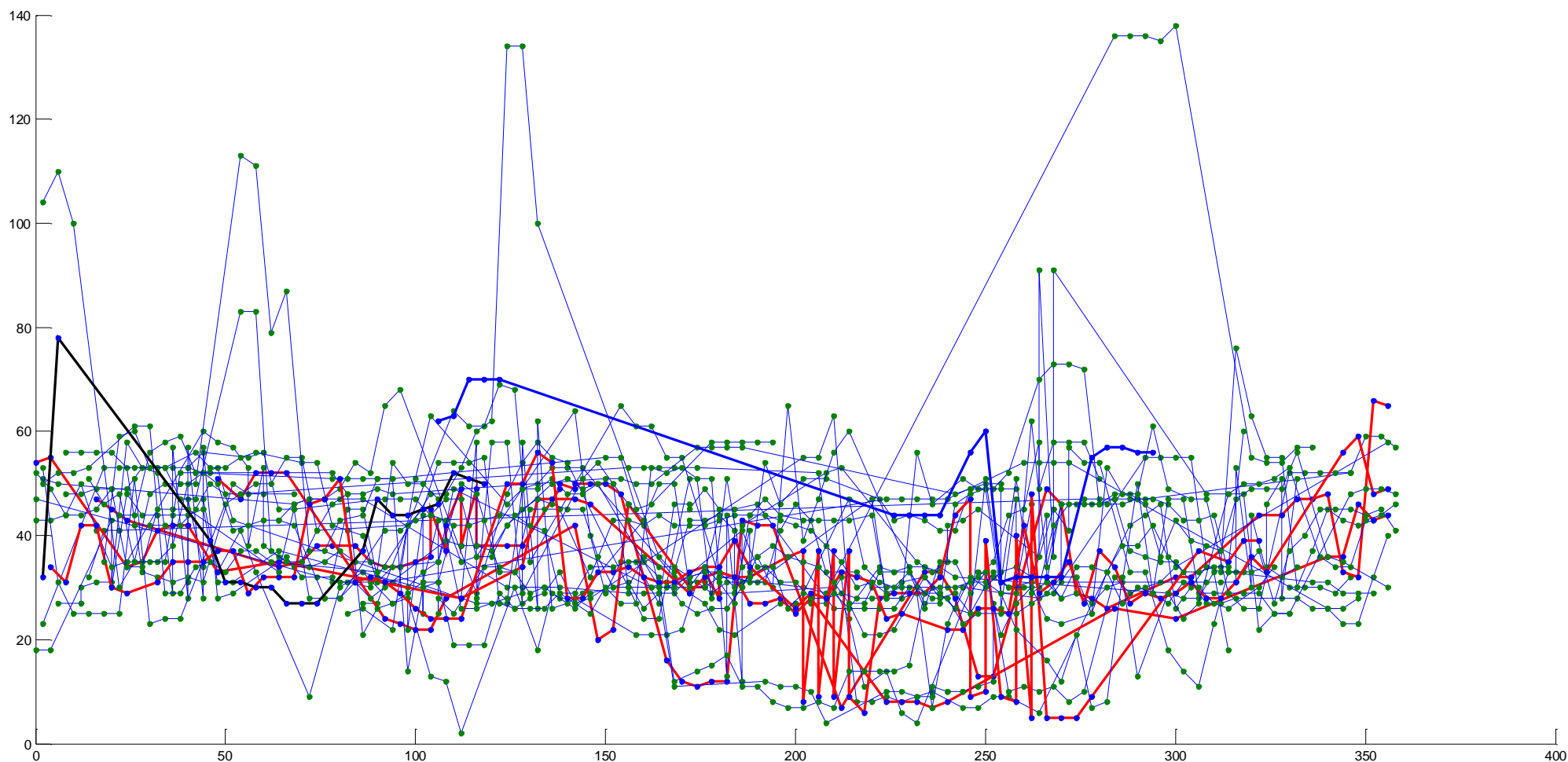
**т.е. + требование непрерывности**



## «Существенно двухмерное» усреднение

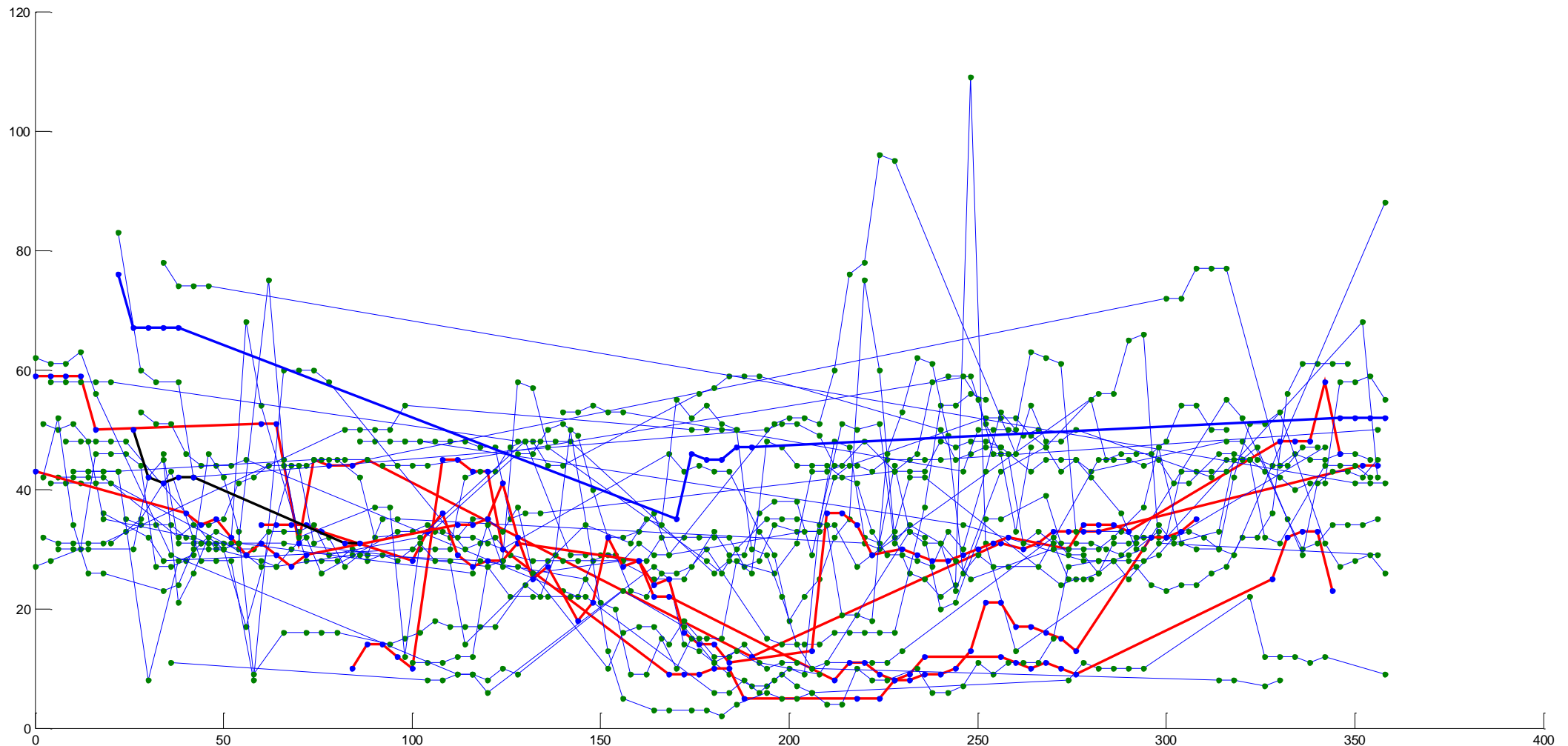


## Как выглядят данные:



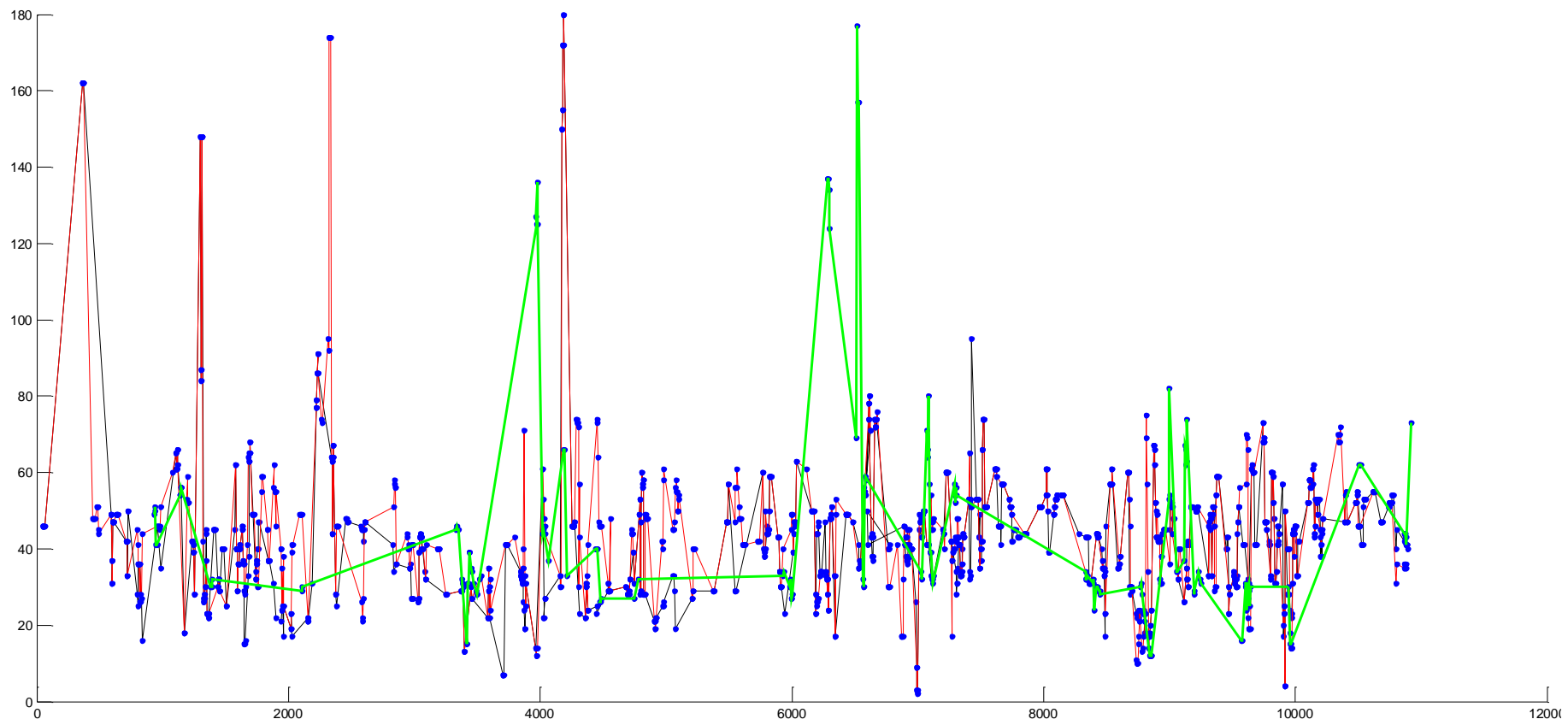
**Чёрный – наш день,**  
**Красный – этот день недели,**  
**Синий – предыдущий день.**

## Другая дуга (граф ориентированный):



## Замечаем странности:

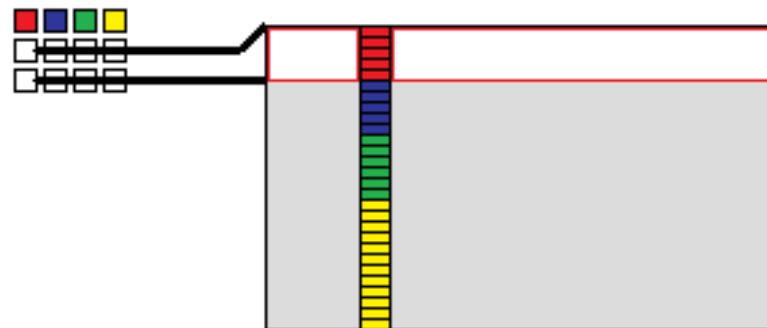
1. По некоторым дугам статистика совпадает
2. Или почти совпадает.
3. Скорость «теряется» при переходе на другую дугу.



**Разные дороги: чёрный, красный, зелёный.**

## В процессе обработки данных открыл для себя приём: Выборка по факторам...



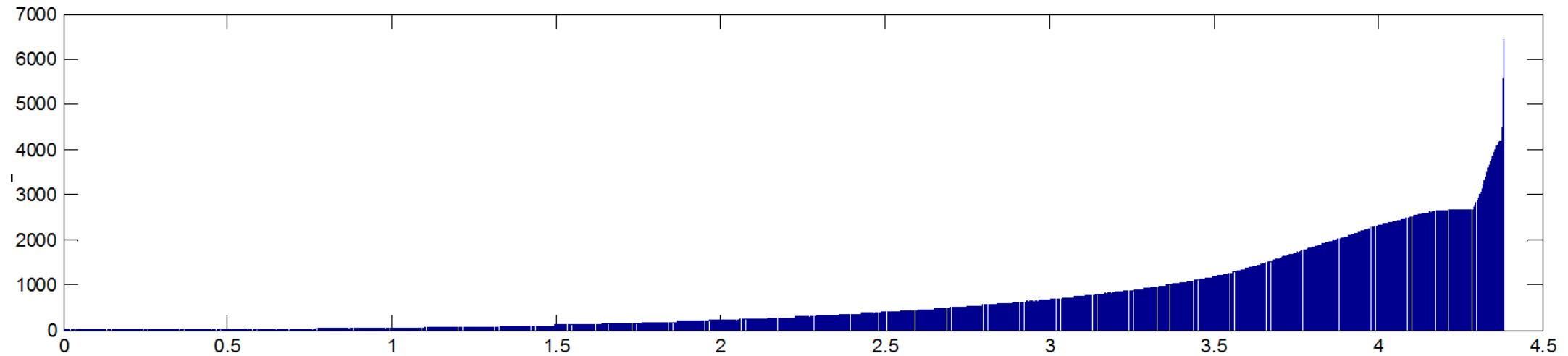
$$M[M[:,i]==a,:]$$


	<div style="display: inline-block; width: 15px; height: 15px; background-color: red; border: 1px solid black;"></div>	<div style="display: inline-block; width: 15px; height: 15px; background-color: blue; border: 1px solid black;"></div>	<div style="display: inline-block; width: 15px; height: 15px; background-color: green; border: 1px solid black;"></div>	<div style="display: inline-block; width: 15px; height: 15px; background-color: yellow; border: 1px solid black;"></div>
begin	<div style="display: inline-block; width: 15px; height: 15px; background-color: white; border: 1px solid black;"></div>	<div style="display: inline-block; width: 15px; height: 15px; background-color: white; border: 1px solid black;"></div>	<div style="display: inline-block; width: 15px; height: 15px; background-color: white; border: 1px solid black;"></div>	<div style="display: inline-block; width: 15px; height: 15px; background-color: white; border: 1px solid black;"></div>
end	<div style="display: inline-block; width: 15px; height: 15px; background-color: white; border: 1px solid black;"></div>	<div style="display: inline-block; width: 15px; height: 15px; background-color: white; border: 1px solid black;"></div>	<div style="display: inline-block; width: 15px; height: 15px; background-color: white; border: 1px solid black;"></div>	<div style="display: inline-block; width: 15px; height: 15px; background-color: white; border: 1px solid black;"></div>

### Сортировка

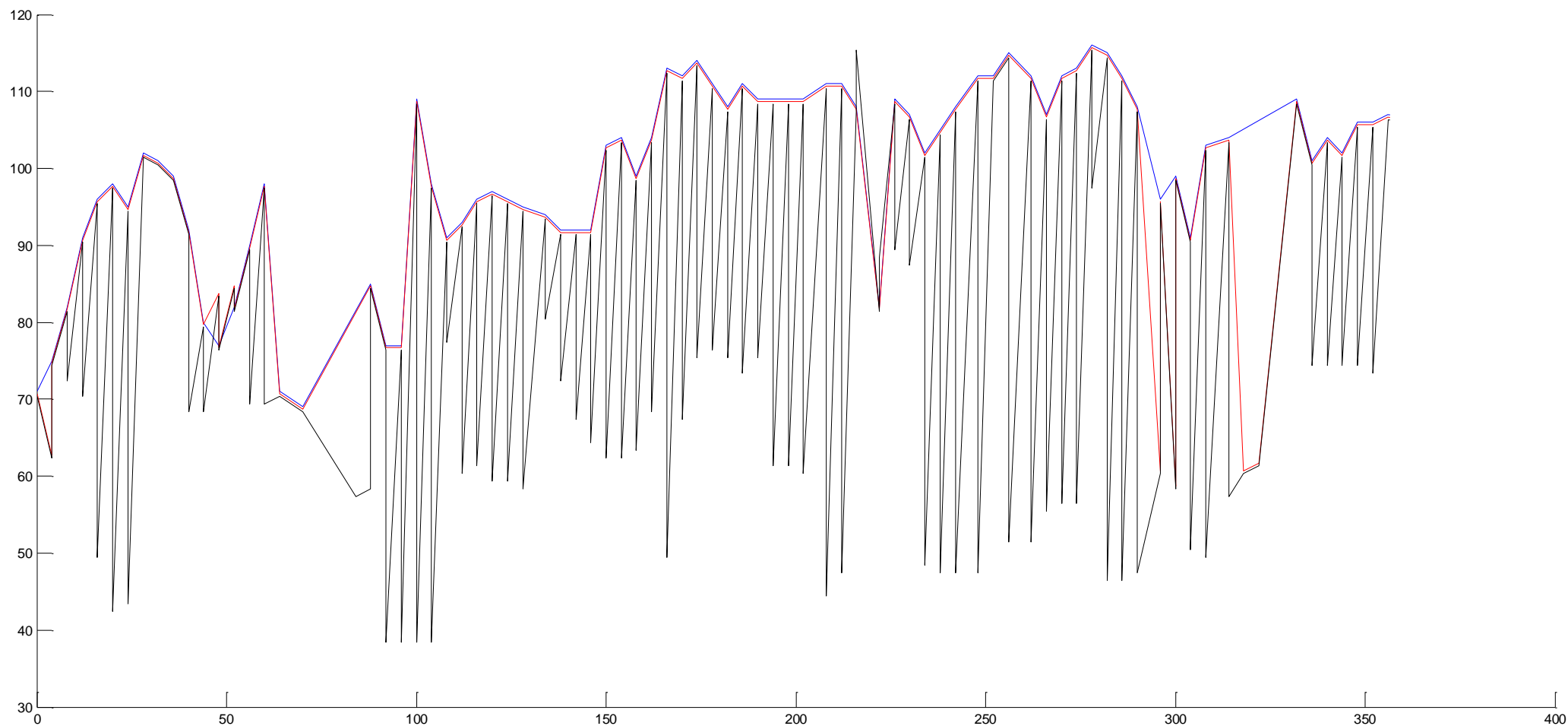
**Хранить  
начала и концы разных  
факторов.**

## Распределение длин дорог



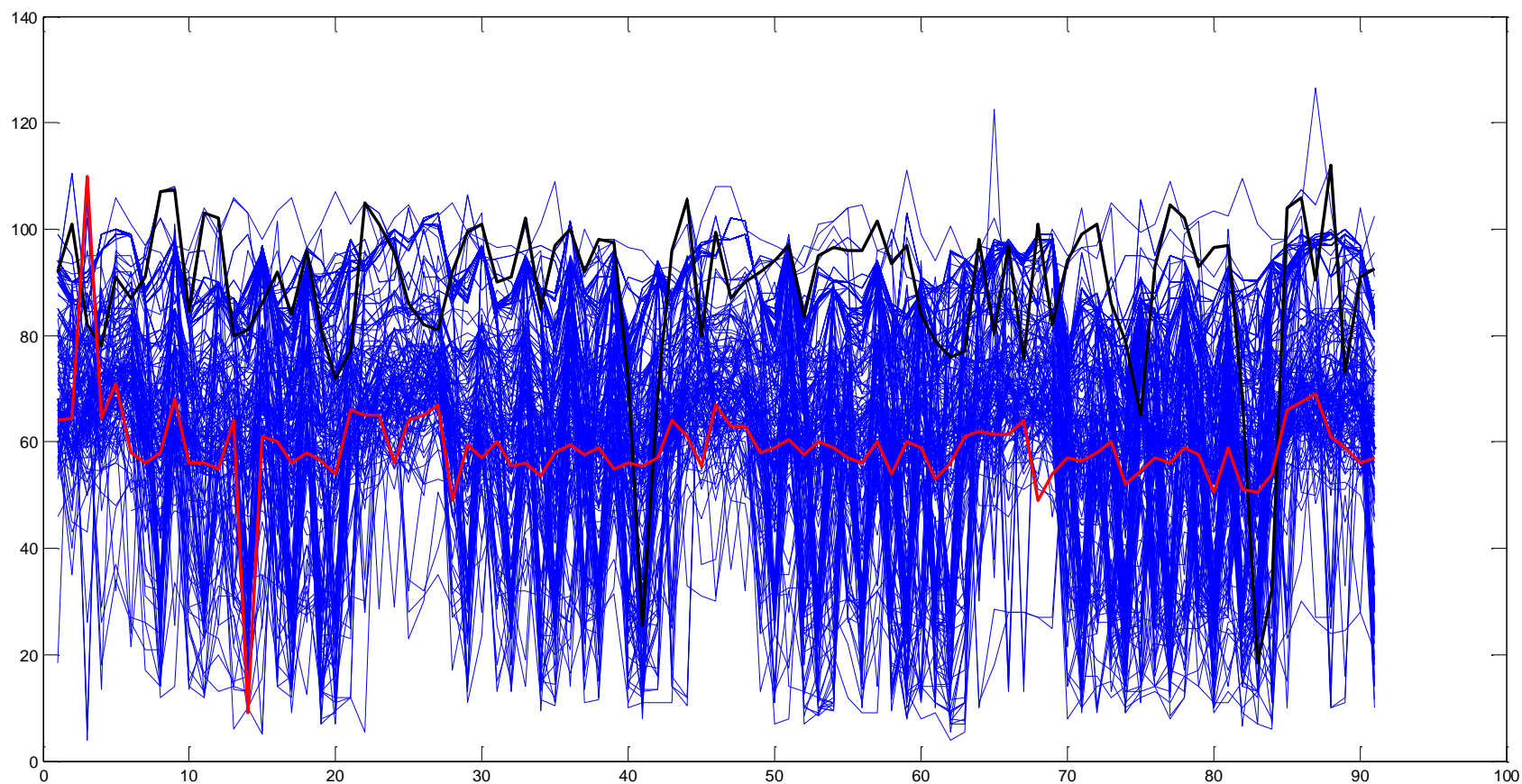
**опять нет нормального распределения...**

## Данные с трёх дуг



**Данные двух дуг совпадают,  
+ с половиной данных третьей дуги.**

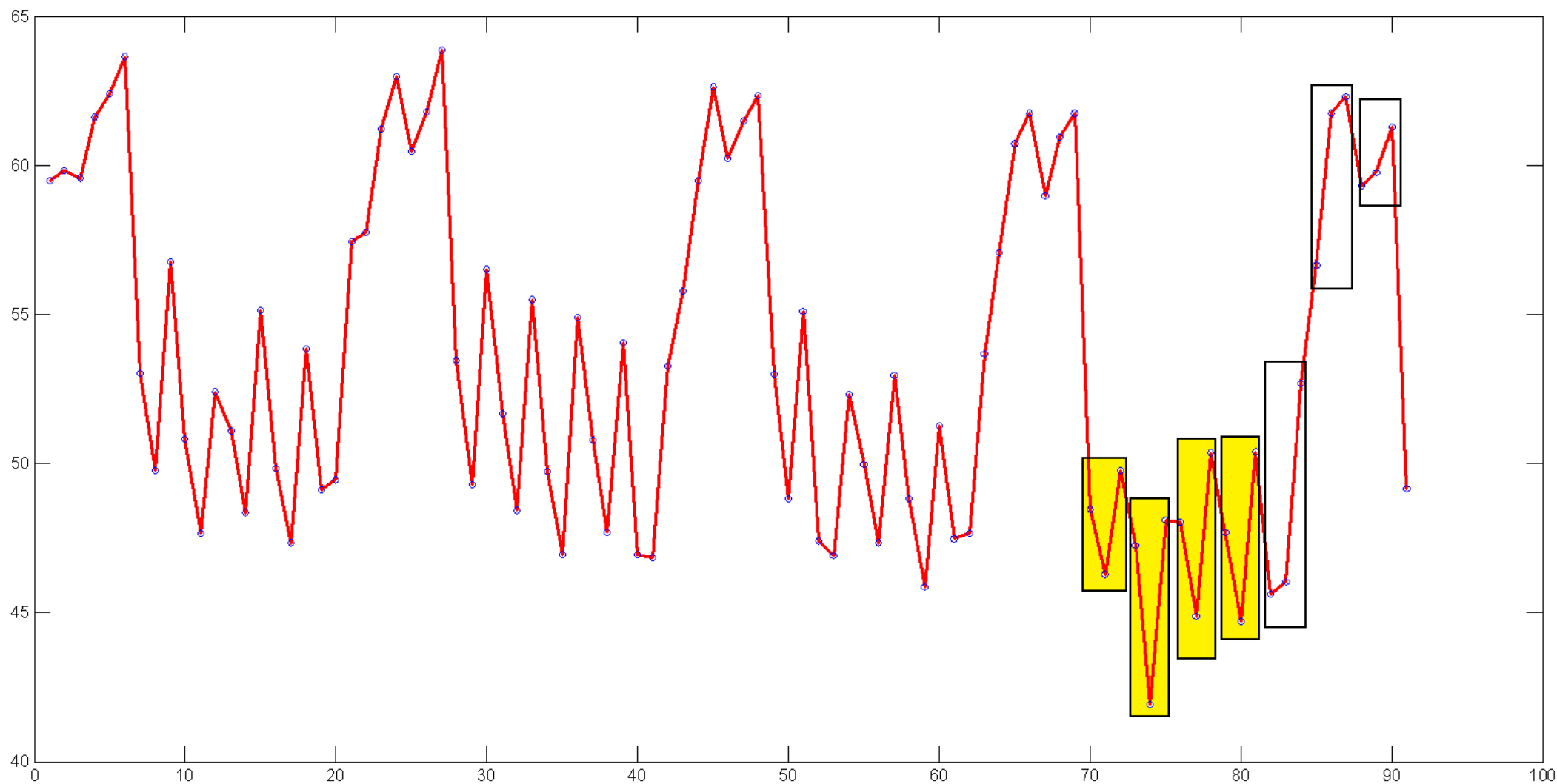
## Медианные данные по всем дням



**Что можно сказать?**

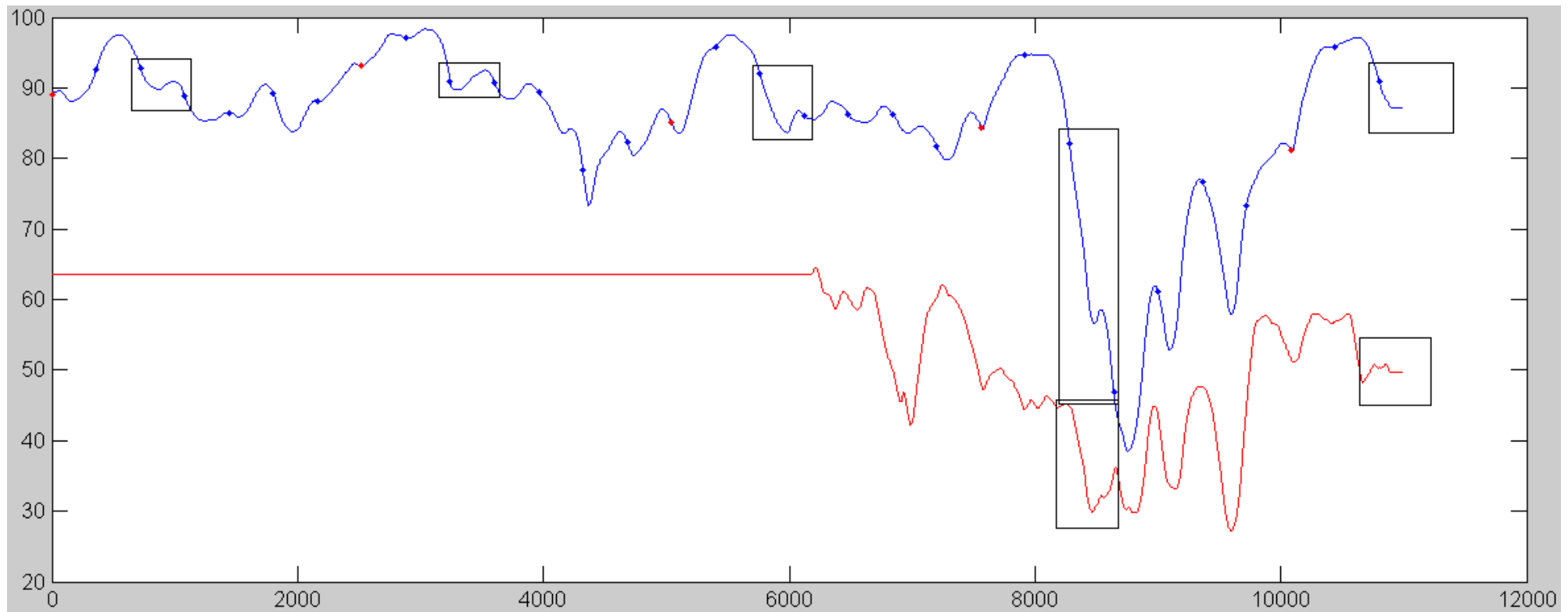


**Ответ: Идентифицировать дни недели.**



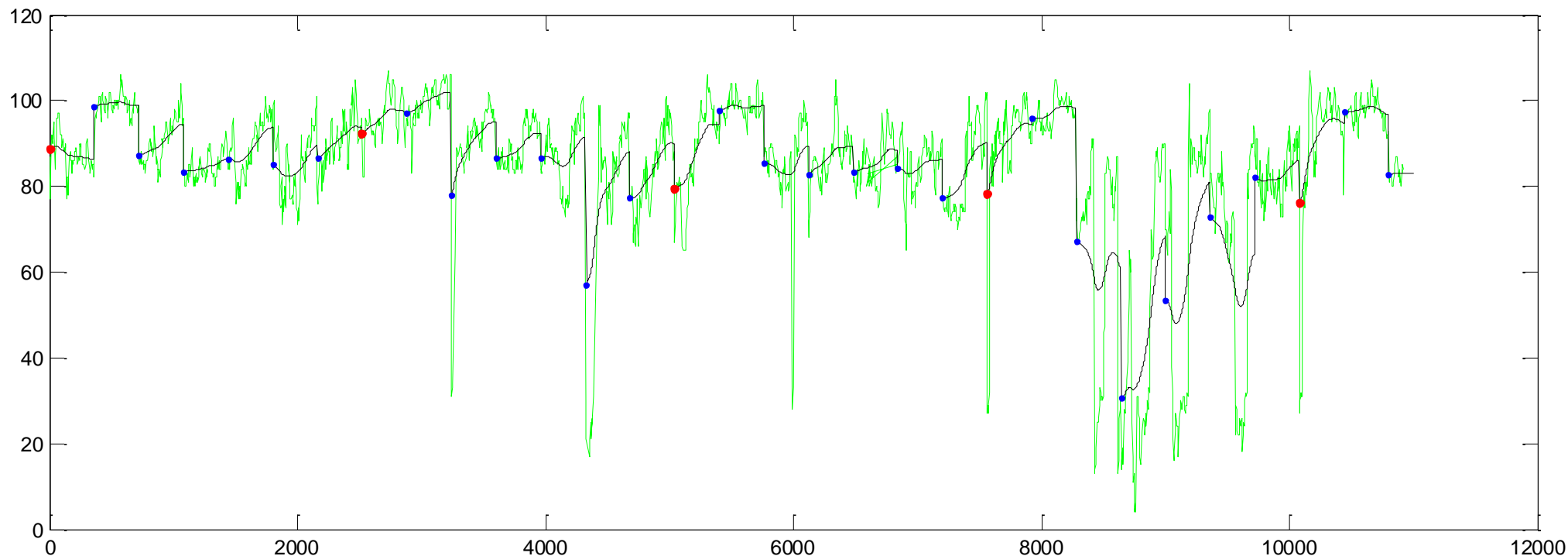
**и даже видна идея решения!**

## Иллюстрация сглаживания



**Данные по двум конкретным дорогам.  
Выделены участки одного дня недели.  
По **красной** нет достаточно статистики,  
но она коррелирует с **синей**, по которой есть!**

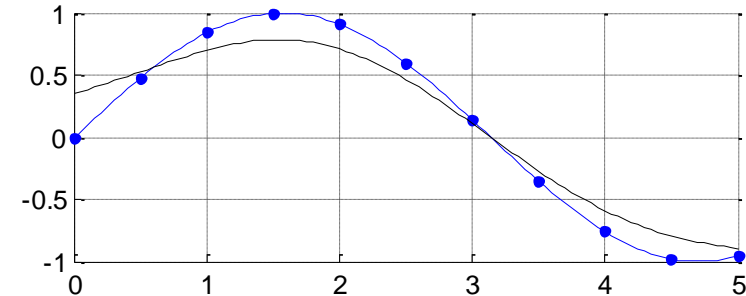
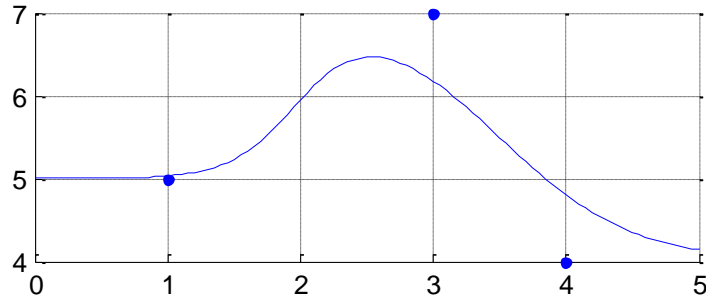
## Пример сглаживания



$$y(x) = \sum_{i=1}^n w_i y(x_i)$$
$$w_i = K(x, x_i) \approx^N e^{-\rho(x, x_i)}$$

**Формула Надарая-Ватсона – а ведь это тоже весовая схема!**

## «Регрессия» по формуле Надарая-Ватсона

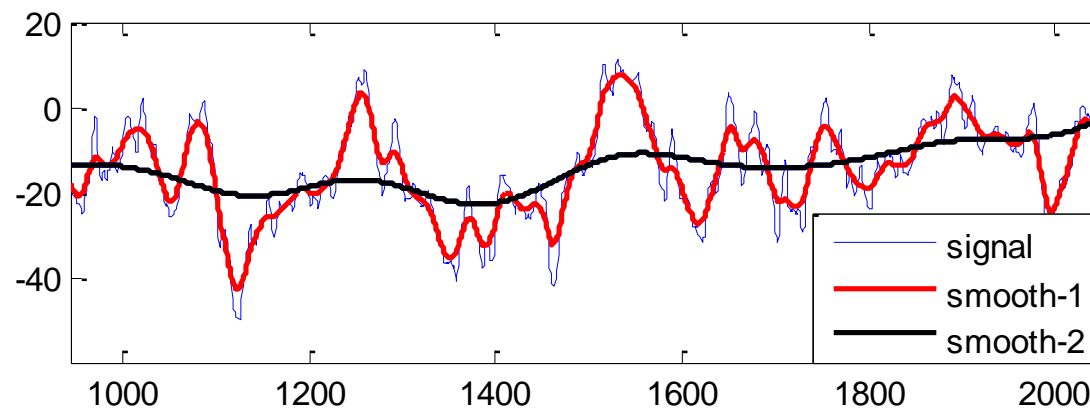


```

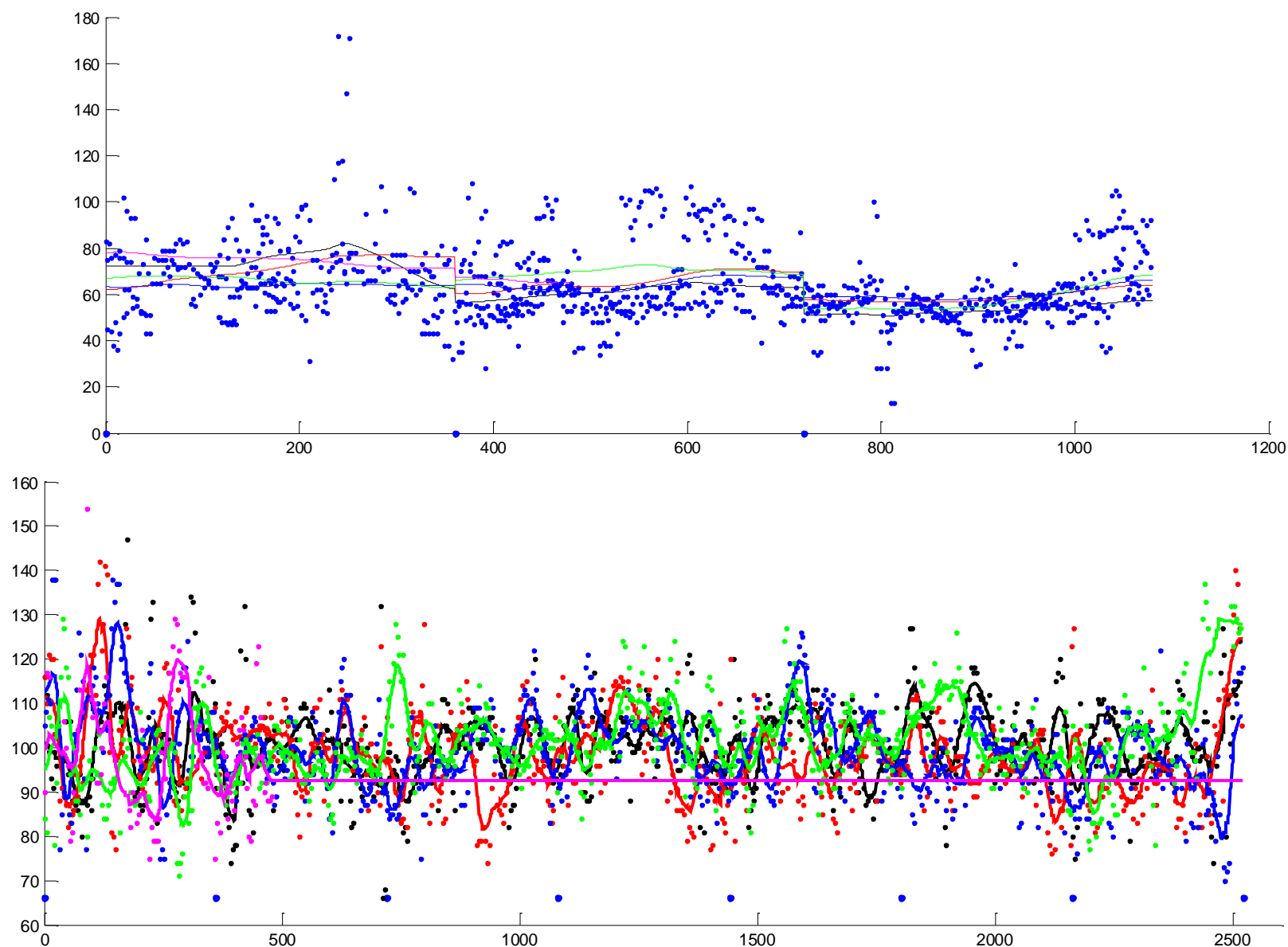
K = @(x) exp(-x.^2); % функция K
x = 0:0.05:5; % отрезок
f = sin(x); % истинные значения функции
X = x(1:10:end); % обучающая выборка - объекты
Y = sin(X); % - их метки
t = repmat(x',1,length(X)) - repmat(X,length(x),1);
t = arrayfun(K, t);
sumt = sum(t, 2);
t = sum(t.*repmat(Y,length(x),1), 2)./sumt; % значения
% ф-лы Н-В
clf; hold on; grid on; % Графика
plot(x, f, 'b'); % как должно быть
scatter(X, Y, 20, 'filled'); % обучение
plot(x, t, 'k'); % что получилось

```

## Сглаженная электрокортикограмма при различных $h$ .

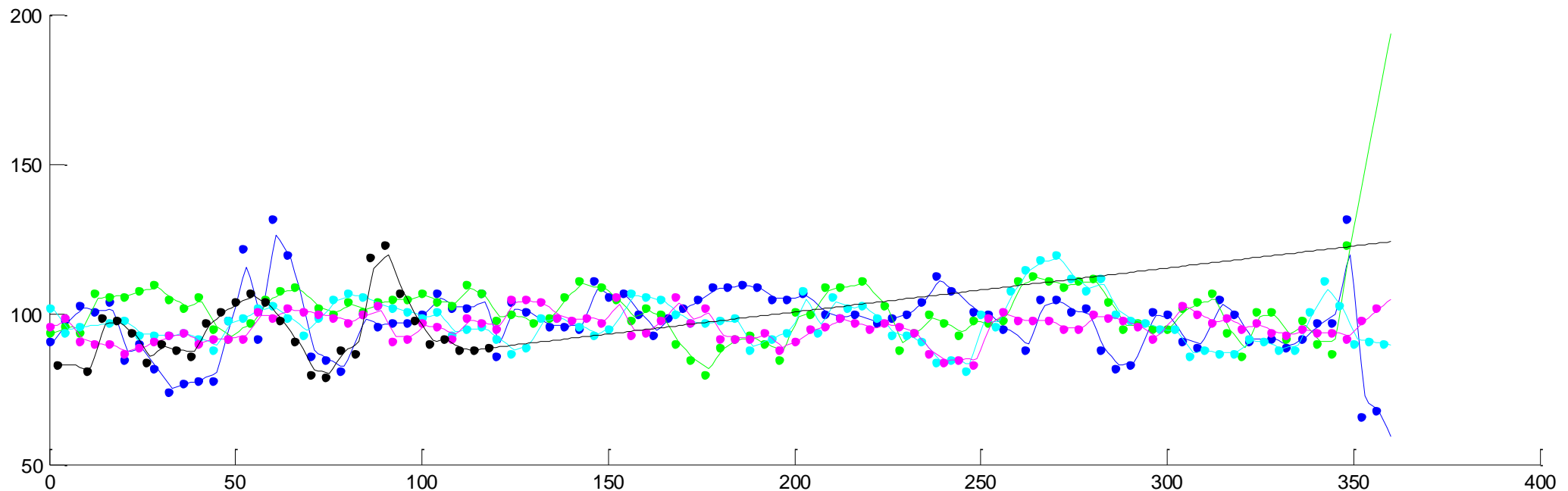


## Зачем нужно сглаживать... скорость на одной дороге в разные дни



## ЛИНЕЙНЫЙ Надарая-Ватсон

**достаточно опасный:**



### В обычном

- не проходит через точки
- почти всё считает выбросом
- не экстраполирует
- проблема подбора ширины окна (ядра)

## **Рецепт по усреднению:**

### **Что усреднять:**

- 1. Данные этого дня**
- 2. Данные вчерашнего дня (тек. день - пн)**
- 3. Данные этого дня недели**

**Как – эксперименты!**

## Литература

- **Шурыгин А.М. Математические методы прогнозирования // М., Горячая линия — Телеком, 2009, 180 с.**  
нужные фрагменты есть в <http://www.machinelearning.ru/wiki/images/7/7e/Dj2010up.pdf>
- **Дьяконов А.Г. Прогноз поведения клиентов супермаркетов с помощью весовых схем оценок вероятностей и плотностей // Бизнес-информатика. 2014. № 1 (27). С. 68–77.**  
<https://bijournal.hse.ru/data/2014/04/15/1320713004/8.pdf>
- **Неправильные интерпретации и ложные закономерности в анализе данных**  
<https://alexanderdyakonov.files.wordpress.com/2015/07/dyakonovfunnydm.pdf>
- **Оценка вероятности: когда к нам придёт клиент? //**  
<https://vimeo.com/119925869> (не действительна)