

Яндекс

ШАД

# Применение методов анализа данных: «Антифрод»

Даниил Тарарухин

dante@yandex-team.ru

# Fraud в вебе: виды накруток

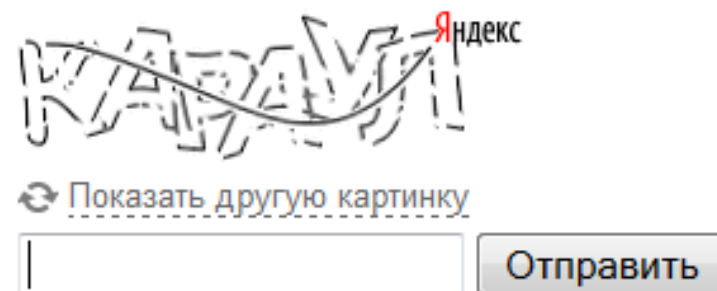
## 1. Баннерные сети

модель оплаты: CPM – cost per mille

## 2. Контекстная реклама в Yandex/Google/Begun

модель оплаты: CPC – cost per click

## 3. Спам



## 4. Эмулирование поведения пользователей

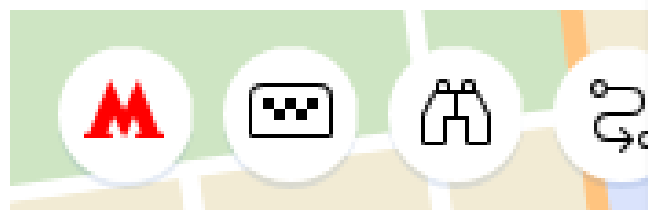
# Эмулирование пользовательского поведения

Яндекс

Найдётся всё



Карта Москвы



Авиабилеты

Поиск Карты Маркет Новости Словари Картинки Видео

накрутка лайков

накрутка лайков вконтакте

накрутка лайков

накрутка подписчиков вконтакте

накрутка лайков в инстаграме

накрутка подписчиков в инстаграме

накрутка лайков в вк

накрутка подписчиков

накрутка

накрутка лайков вконтакте likest

накрутка друзей в контакте

# Усложнение схем монетизации

1. Поставили баннер -> накликали
2. Создали сайт/приложение -> накликали лайки -> продвинулись -> собрали трафик -> открутили рекламу
3. Создали «фабрику накрутки» - специализация на накрутке
4. «Биржи накруток» -> значительное снижение порога входа на накруточный рынок

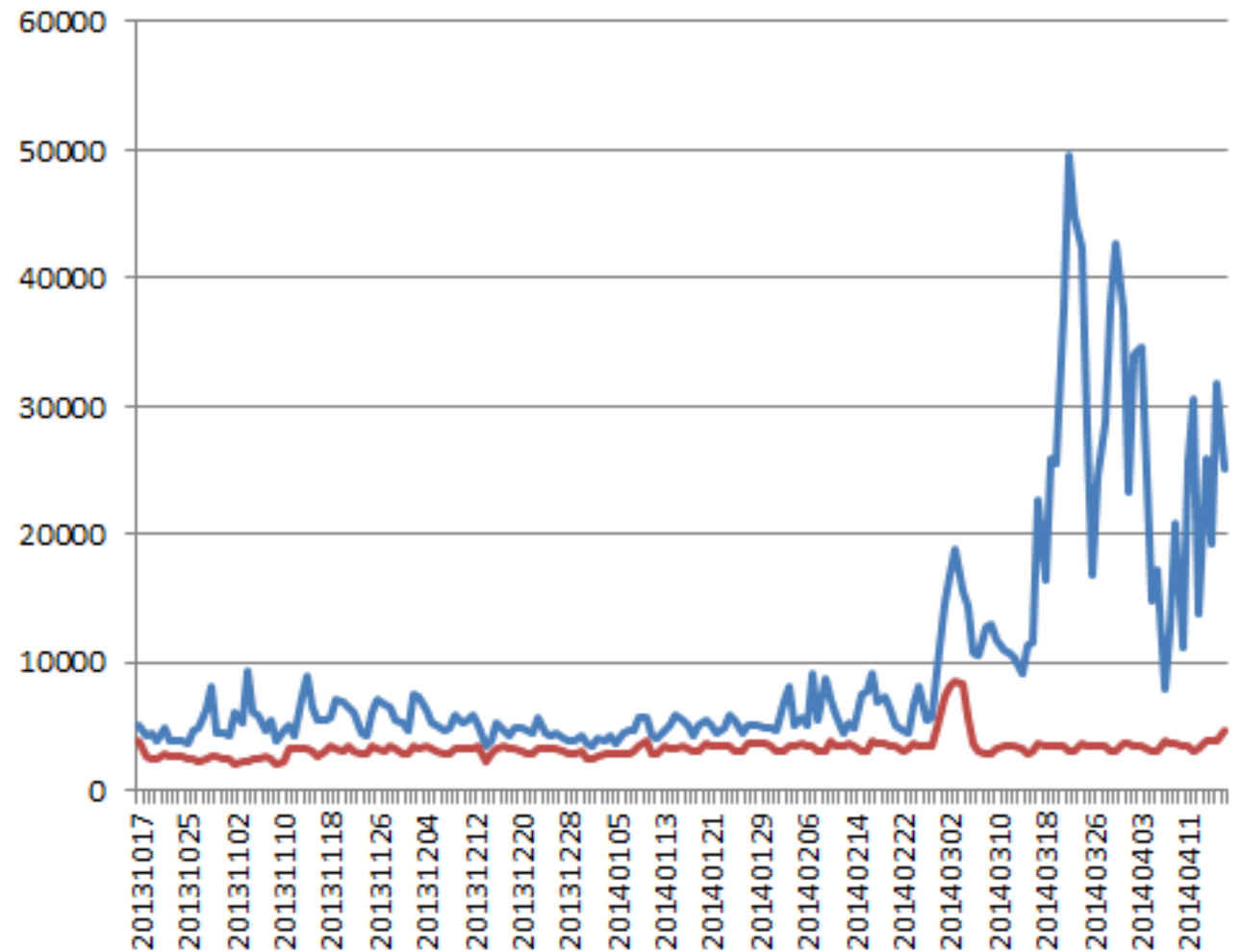
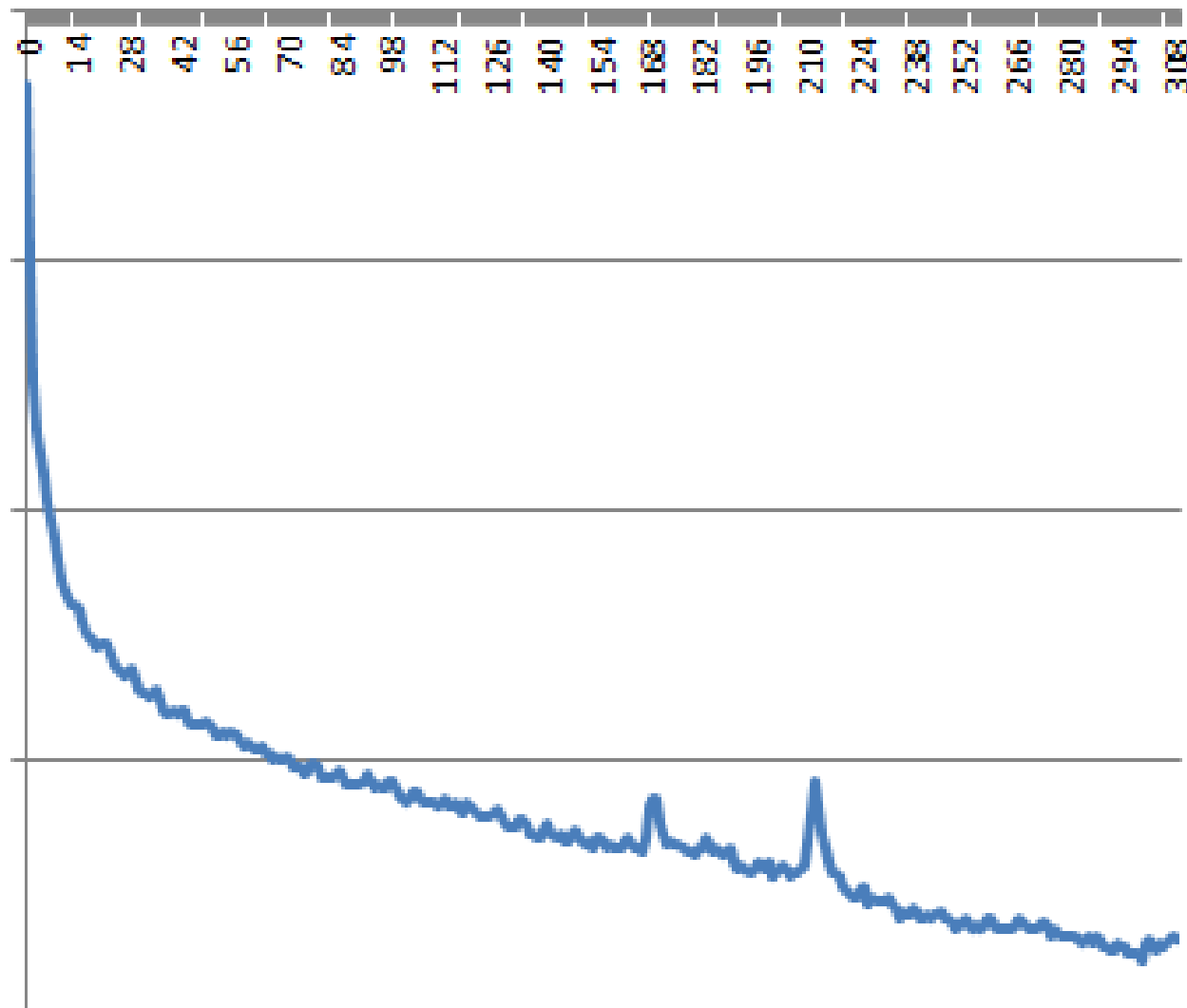
# Анекдот про хакера в столовой

По мотивам <https://xakep.ru/2006/12/16/35784/>

Мораль: атака всегда имеет преимущество над защитой.

Дополнение: вообще-то нет, не всегда.

# Фрод как аномальное поведение



Злоумышленник не знает некоторых «эталонных» распределений, они известны только на стороне сервиса. Иногда можно подобрать параметры так, что злоумышленник будет аномальным по этим параметрам.

# Realtime- и offline-системы



Realtime



Offline

Realtime фильтрует на входе, offline добанивает остальных.

Часто в realtime важна точность в ущерб полноте (нельзя забанить невиновного).

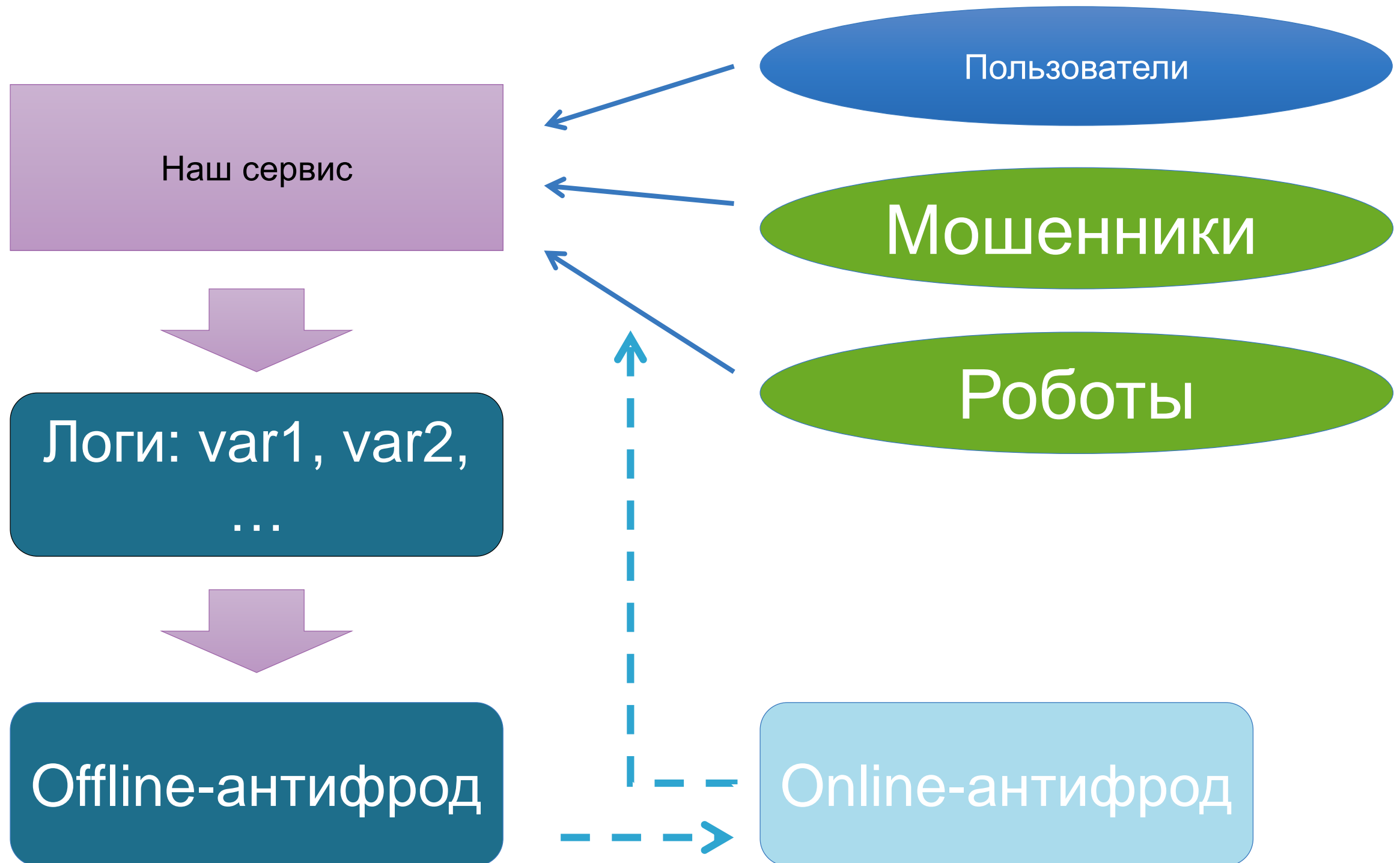
Часто realtime – какая-то подсистема offline, способная работать быстро и точно.



# Метрики и KPI

1. False-negative / полнота / recall
2. False-positive / точность / precision
3. Метрики денег
4. Метрики времени
5. Метрики производительности

# Модельная задача



# Модельная задача

Роботы – «тупые», мошенники – более умные, более похожие на настоящих пользователей.

Надо создать какой-то метод поимки мошенников или роботов, в предположении, что те и другие демонстрируют аномальное поведение, в чем-то отличающееся от поведения живых пользователей.

# Первичная обработка данных

1. Проверка соответствия типов, чистка артефактов
2. Шкалирование, построение гистограмм, анализ распределений
3. Исключение низковариативных переменных
4. Исключение сильно скоррелированных переменных

# Поиск закономерностей

1. Уменьшение размерности
2. Кластеризация
3. Дисперсионный анализ (ANOVA)
4. Ассоциативные правила
5. Машинное обучение
6. И др.

# Примеры распределений в реальной жизни

Распределение Бернулли (0 или 1)

Равномерное на отрезке  $[a, b]$

Нормальное  $N(a, s^2)$ : случайные независимые отклонения в обе стороны от среднего, от «нормы».

Экспоненциальное  $\text{Exp}(a)$ : время жизни объекта, не обладающего свойством отсутствия памяти. «Тонкий хвост».

Лог-нормальное  $\text{exp}(N(a, s^2))$ : интенсивность затухания луча, количество денег у домохозяйств

Частное двух величин

# Корреляционный анализ

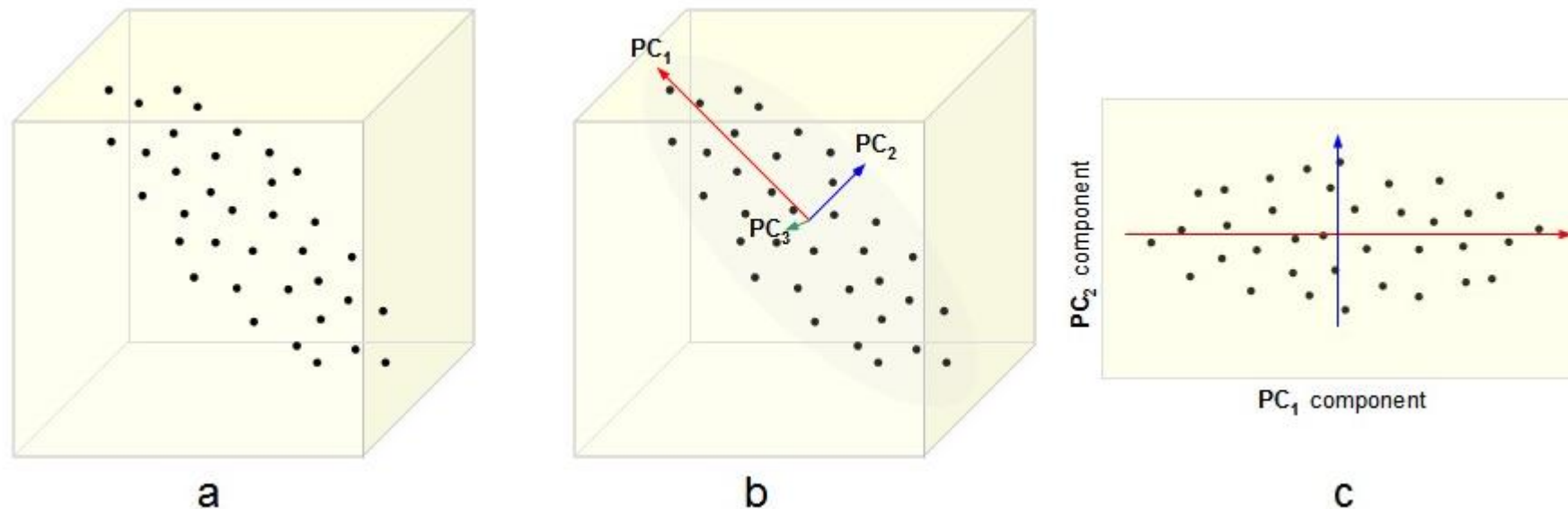
$$r_{XY} = \frac{\text{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}$$

где  $\bar{X} = \frac{1}{n} \sum_{t=1}^n X_t$ ,  $\bar{Y} = \frac{1}{n} \sum_{t=1}^n Y_t$  — среднее значение выборок.

Суть анализа: вычисляем матрицу попарных корреляций  
и пристально смотрим на нее.

# Метод главных компонент

## Principal component analysis (PCA)





Оно же – приведение эллипсоида к главным осям


- 1) Отбираем самые крупные компоненты (объясняющие много дисперсии)
- 2) Оставляем переменные, которые значимо участвуют в отобранных компонентах




# Метод главных компонент

$D =$   данные

$A =$   координаты главных компонент

$A^T =$   координаты старого базиса в главных компонентах

$AA^T = A^TA =$  

$DA$  – матрица данных в базисе главных компонент

# Кластерный анализ

Кластеризация:

- Иерархическая
- K-means
- Естественная (в качестве кластеров берем значения какой-то категориальной переменной. Или переменной, которая может быть воспринята как категориальная.)

Разложение совокупной дисперсии на внутригрупповую и межгрупповую

# Дисперсионный анализ (ANOVA)

$$\sum_{i=1}^{n_j} (x_{i,j} - M)^2 = \sum_{i=1}^{n_j} (M_j - M)^2 + \sum_{i=1}^{n_j} (x_{i,j} - M_j)^2,$$

где

$$SS_{total} = \sum_{i=1}^{n_j} (x_{i,j} - M)^2$$

$$SS_{BG} = \sum_{i=1}^{n_j} (M_j - M)^2$$

$$SS_{WG} = \sum_{i=1}^{n_j} (x_{i,j} - M_j)^2$$

Следовательно

$$SS_{total} = SS_{BG} + SS_{WG}.$$

# ANOVA – «плохие» кластеры

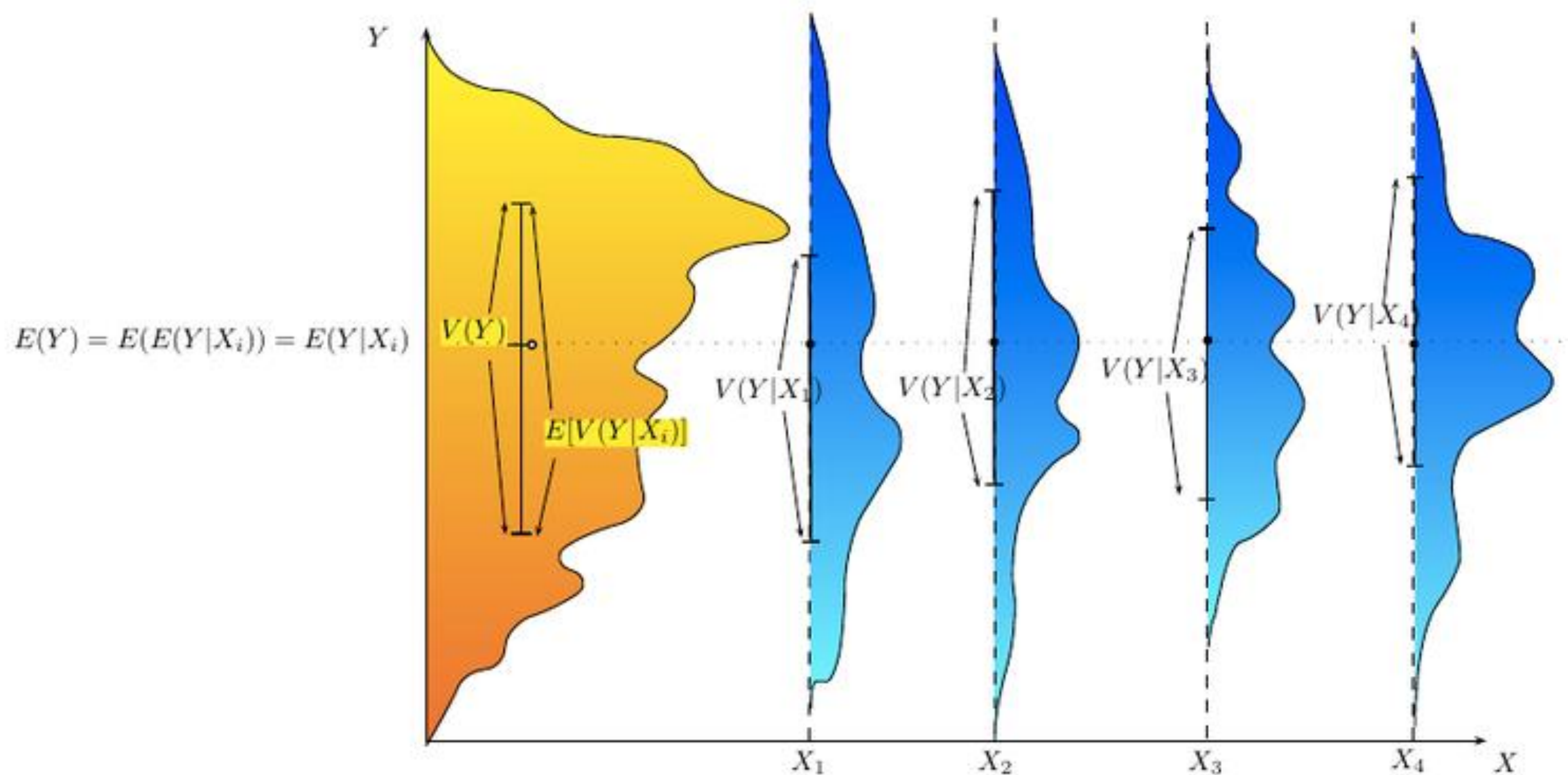


Figure 2: ANOVA : No fit

# ANOVA

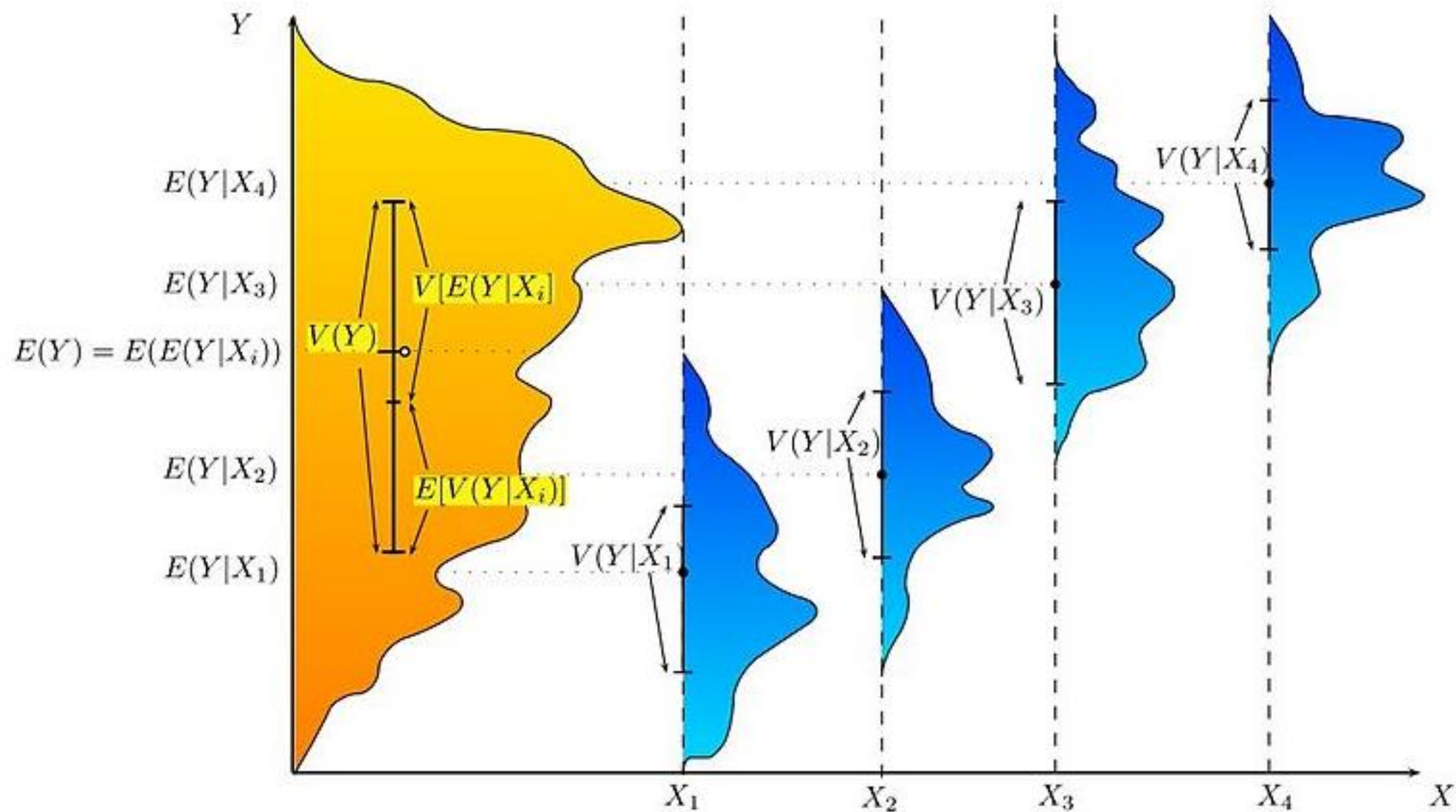


Figure 1: ANOVA : Fair fit

# ANOVA – «хорошие» кластеры

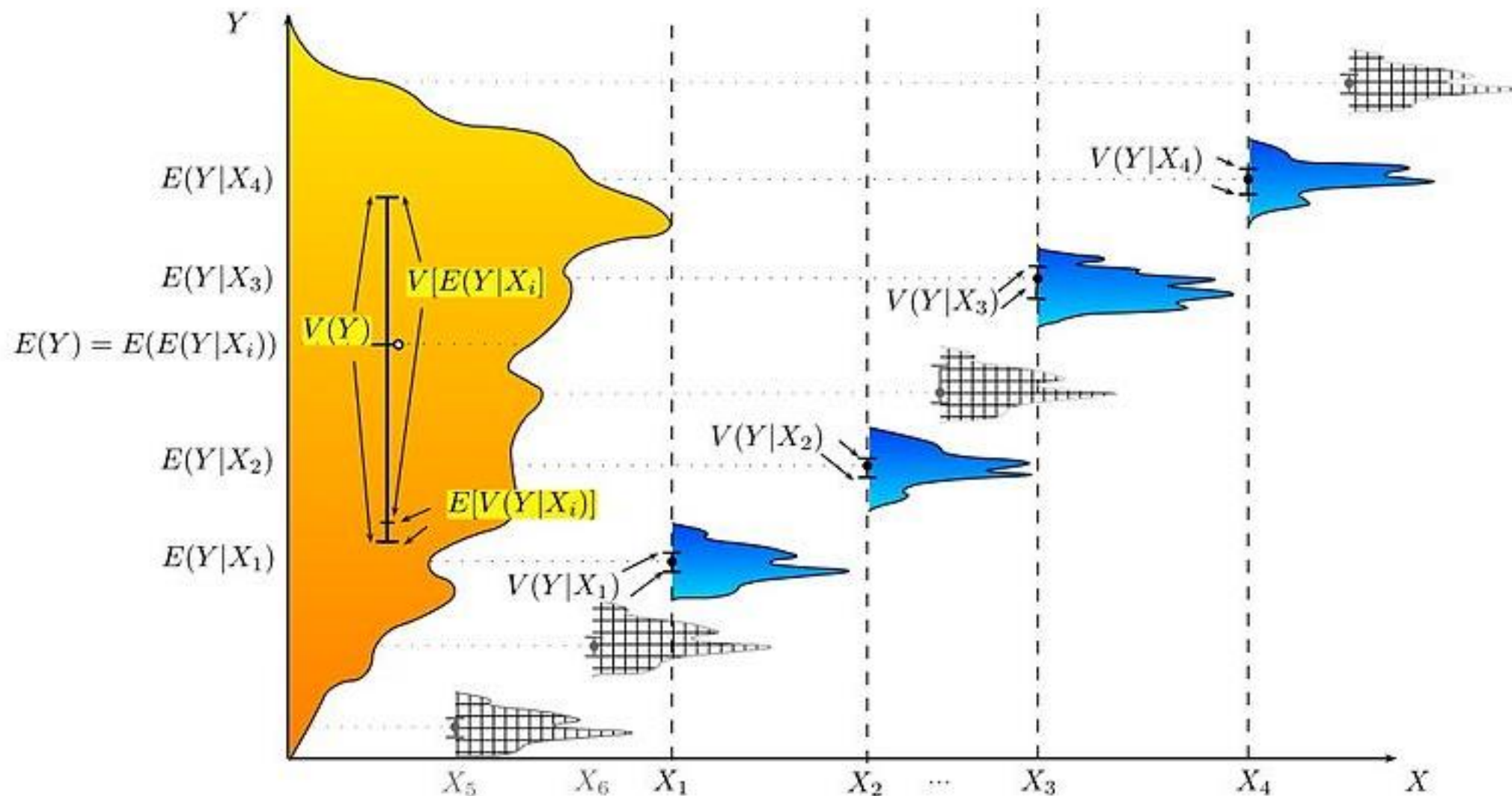


Figure 3: ANOVA : very good fit

Конечно, самые «хорошие» кластеры, с максимальной межкластерной и минимальной внутрикластерной дисперсией получатся, если каждую точку выделять в свой кластер. Но это не имеет смысла. Надо искать кластеризации, когда кластеров относительно немного, а межкластерная дисперсия высокая.

# Машинное обучение

Обучающее множество: матрица

$$X_{11}, X_{12}, \dots, X_{1m}, \Rightarrow Y_1$$

...

$$X_{n1}, X_{n2}, \dots, X_{nm}, \Rightarrow Y_n$$

Генеральная совокупность: хотим подобрать  $Y$

Классификация:  $Y_i = 0$  или  $1$

Точность: частота правильного определения объекта

Полнота: % правильно определенных объектов

F1-мера: среднее гармоническое

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

# Bonus track

Можно отойти от представления данных в виде матрицы факторов для объектов. Либо переходить к анализу макро-объектов.

- разладки: данные как временные ряды

Допустим, мы можем наблюдать за каким-то объектом во времени. Изменение его поведения – возможно, признак того, что он стал аномальным.

- кластеризация для случая неразличимых объектов.

Настоящий пользователь, производящий мало действий, неотличим по логам от робота, производящего мало действий, т.к. никакие факторы, которые можно придумать, не различают две строки в логе, состоящие из нулей.

Однако можно определить правильную долю таких «нулевых» пользователей, и затем искать кластера, в т.ч. «естественные кластеризации», в которых доля нулевых пользователей будет значительно выше ожидаемых значений.



**Спасибо за внимание**