ЦЕНТР
АНАЛИЗА
ДАННЫХ

# Feature Engineering, Ensembles and Medicine Cases

19.11.16

ДЕПАРТАМЕНТ ЗДРАВООХРАНЕНИЯ ГОРОДА МОСКВЫ

Департамент информационных технологий города Москвы

ИАЦ ДЗМ

Data analysis center

https://analysiscenter.ru/

https://analysiscenter.github.io/about/

https://www.facebook.com/analysiscenter.ru

# Agenda

ЦЕНТР АНАЛИЗА ДАННЫХ


**Feature Engineering** – Алексей Ушаков

**Ensembles** – Кирилл Емельянов

**Case:** Прогнозирование спроса на терапевтическую помощь Дмитрий Подвязников

**Case:** Балансировка терапевтических участков Наталия Амелина

# **Feature Engineering** – Алексей Ушаков

# Feature Engineering

ЦЕНТР АНАЛИЗА ДАННЫХ

Чек-лист аналитика:

- Сформулировать корректную проблему
- Выбрать модель
- Выбрать алгоритм обучения
- Найти данные
- **Придумать/создать признаки, правильно представить, отобрать релевантные**

Feature (признак) – информация, потенциально полезная для предсказания

# Feature Engineering

ЦЕНТР АНАЛИЗА ДАННЫХ

Мы не можем формально определить, что такое Feature Engineering, но…
для успешной реализации нужно:

- Знать свои данные (область, проблематику, особенности)
- Понимать особенности задачи (регрессия/классификация, метрика и т.д.)
- Очень много экспериментировать


Но всему есть предел:

1. Рано или поздно скорость станет проблемой
2. Overfitting

**Поэтому нужно выбирать только «полезные», «релевантные» признаки**

# Feature Engineering

ЦЕНТР
АНАЛИЗА
ДАННЫХ

Как научиться:

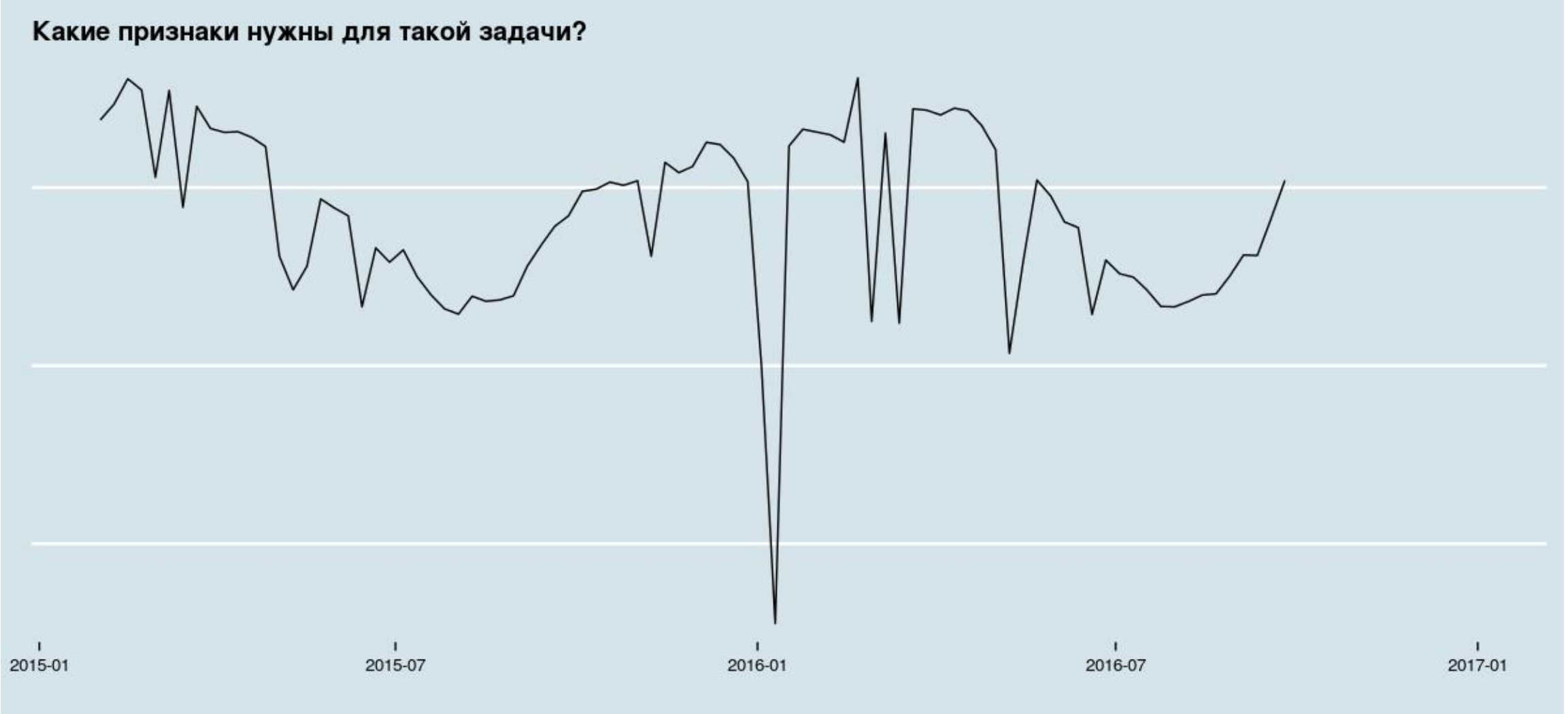
- Проводить больше экспериментов
- Читать статьи/скрипты под конкретные проблемы
- Придумывать задачу, «а как бы я сделал?»

О чем подумать, работая с данными:

- Есть ли N/A, Missing values?
- Какое распределение у признака? Насколько сильно оно зависит от подвыборки? От времени/типа сегментации?
- Есть ли дубликаты или пересекающиеся данные?
- Все ли данные дискретны/непрерывны?

Хорошая статья:
“A Few Useful Things to Know about Machine Learning” Pedro Domingos

## Спрос на услуги терапевтов в Москве



**Ensembles** – Кирилл Емельянов

# План

ЦЕНТР
АНАЛИЗА
ДАННЫХ

- Что такое ансамбли и зачем их применять
- Примеры ансамблей
- Random forest
- Boosting
- Stacking

# Основные понятия

ЦЕНТР АНАЛИЗА ДАННЫХ

- Обучающая выборка $X^l = (x_i, y_i)_{i=1}^l$
- Вероятностная природа данных $(x_i, y_i) \sim p(x, y)$
- Метод обучения $\mu : X^l \to A$ ,

  где $A$ --- пространство алгоритмов.
- Функция потерь $L(z, y)$
- Функционал потерь $Q(\mu) = E_{X^l} E_{(x,y)} \left(\mu(X^l)(x) - y\right)^2$

# Типы композиций

ЦЕНТР АНАЛИЗА ДАННЫХ

**Основные подходы к построению композиции алгоритмов:**

1) Простое голосование:

$$\hat{b}(x) = F\left(b_1(x), \ldots, b_T(x)\right) = \frac{1}{T}\sum_{t=1}^{T} b_t(x)$$

2) Взвешенное голосование:

$$\hat{b}(x) = F\left(b_1(x), \ldots, b_T(x)\right) = \sum_{t=1}^{T} \alpha_t \cdot b_t(x), \quad x \in X, \alpha_t \in \mathbf{R}$$

3) Смесь алгоритмов:

$$\hat{b}(x) = F\left(b_1(x), \ldots, b_T(x)\right) = \sum_{t=1}^{T} g_t(x) \cdot b_t(x), \quad g_t(x) : X \to \mathbf{R}$$

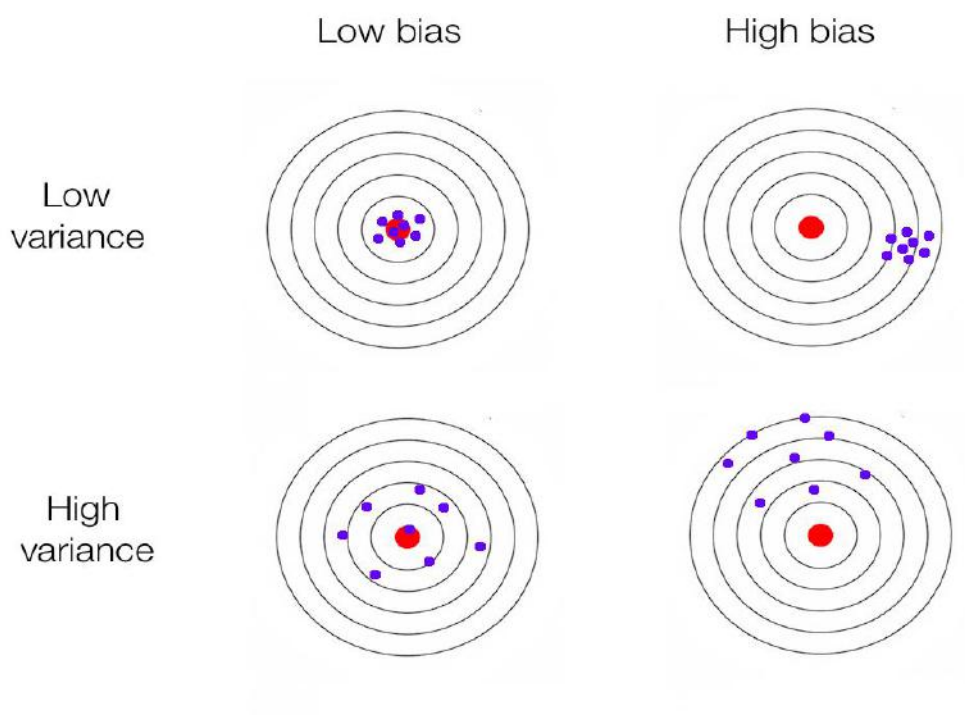
# Bias–Variance trade–off

ЦЕНТР АНАЛИЗА ДАННЫХ

**Теорема:**

$$Q(\mu) = \underbrace{E_{(x,y)}\left(a^*(x) - y\right)^2}_{noise} + \overbrace{E_{(x,y)}\left(\bar{a}(x) - a^*(x)\right)^2}^{bias} + \underbrace{E_{(x,y)}E_{X^l}\left(\mu(X^l)(x) - \bar{a}(x)\right)^2}_{variance}$$

где $\bar{a}(x) = E_{X^l}\left(\mu(X^l)(x)\right), \quad a^*(x) = E(y|x) = \int_Y y\, p(y|x)\, dx$



# Стохастические алгоритмы построения композиций



- Построение такой композиции уменьшает разброс (variance).
- Качество увеличивается, если алгоритмы композиции преимущественно независимы.
- Bagging (Bootstrap aggregation):

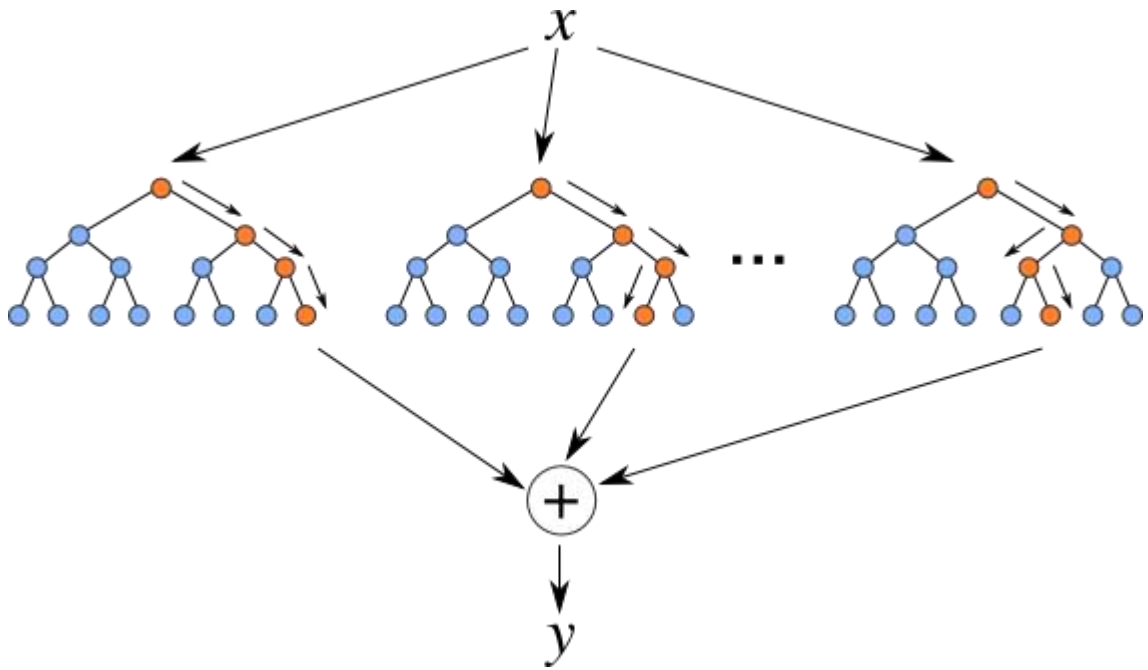
  из обучающей выборки формируются различные обучающие подвыборки меньшего размера с помощью бутстрепа (выбор с возвращением).
- RSM (Random Subspace Method):

  базовые алгоритмы обучаются на различных подмножествах признаков, выделяемых случайным образом.

# Пример: случайный лес



- Бэггинг над решающими деревьями.
- RSM. Признак в каждой вершине дерева выбирается из случайного подмножества признаков (k из n).
- Оптимальные значения для параметра k (эвристики):
  - $k = \lfloor \frac{n}{3} \rfloor$ для задачи регрессии
  - $k = \lfloor \sqrt{n} \rfloor$ для задачи классификации



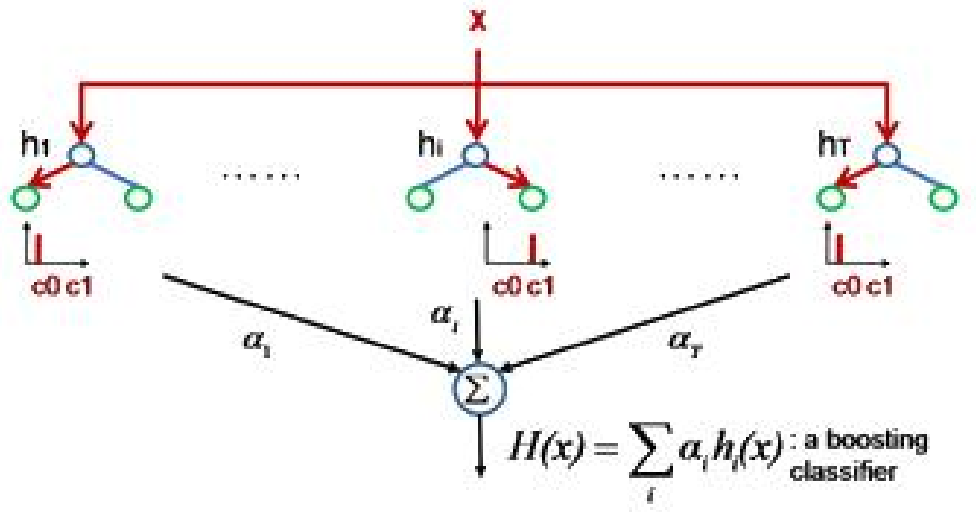
# Бустинг



- Комбинация слабых классификаторов, которая сама является сильным классификатором.
- Уменьшает смещение.
- Обычно в качестве слабых классификаторов выступают решающие деревья небольшой глубины (~5).

Примеры функций потерь:

- AdaBoost: $L(z) = e^{-z}$
- LogitBoost: $L(z) = \log_2(1 + e^{-z})$
- GentleBoost: $L(z) = (1 - z)^2$



# Пример: градиентный бустинг



- Произвольная дифференцируемая функция потерь.
- На каждой итерации добавляется алгоритм, который сильнее всех остальных уменьшает ошибку композиции.

- Формально, на каждой итерации алгоритма решается следующая задача:

$$\sum_{i=1}^{l} L(y_i, a_{N-1}(x_i) + \gamma_N b_N(x_i)) \to \min_{b_N \in B}$$

- Ответы алгоритма должны наилучшим образом приближать градиент функции потерь:

$$b_N(x) = \arg\min_{b \in B} \sum_{i=1}^{l} (b(x_i) + \frac{\partial L}{\partial z}|_{z=a_{N-1}(x_i)})^2$$

- Вес алгоритма выбирается по аналогии с наискорейшим спуском:

$$\sum_{i=1}^{l} L(y_i, a_{N-1}(x_i) + \gamma_N b_N(x_i)) \to \min_{b_N \in B}$$

$$\gamma_N = \arg\min_{\gamma \in \mathbf{R}} \sum_{i=1}^{l} L(y_i, a_{N-1}(x_i) + \gamma b_N(x_i))$$

# Random Forest vs Gradient Boosting




- Бустинг дает лучшие результаты на обучающих выборках большого размера.
- Бэггинг и RSM более предпочтительны для коротких обучающих выборок.
- RSM позволяет сократить число признаков в случае, когда из больше чем объектов, или среди них есть много неинформативных.
- Для бэггинга возможна параллельная реализация.

# Стэкинг



При наличии построенных $t$ различных базовых алгоритмов, вектор их предсказаний $(b_1(x), \ldots, b_t(x))$ кно принять за новое признаковое описание объекта, после чего построить по ним новый мета-алгоритм $a$:

$$\sum_{i=1}^{l} L(y_i, a(b_1(x_i, \ldots, b_N(x_i)))) \to \min_a$$

- Обучение всех базовых алгоритмов и мета-алгоритма на одной обучающей выборке приводит к переобучению.
- Исходную выборку разбивают на K блоков

  $X = X_1 \sqcup \cdots \sqcup X_k$
- Базовые алгоритмы обучают на исходной выборке, из которой исключен один из блоков.

# Взвешенное голосование (blending)



- Частный случай стэкинга
- Мета-алгоритм является линейной комбинацией базовых:

$$a(x) = \sum_{i=1}^{k} w_i b_i(x), \quad \sum_{i=1}^{k} w_i = 1$$

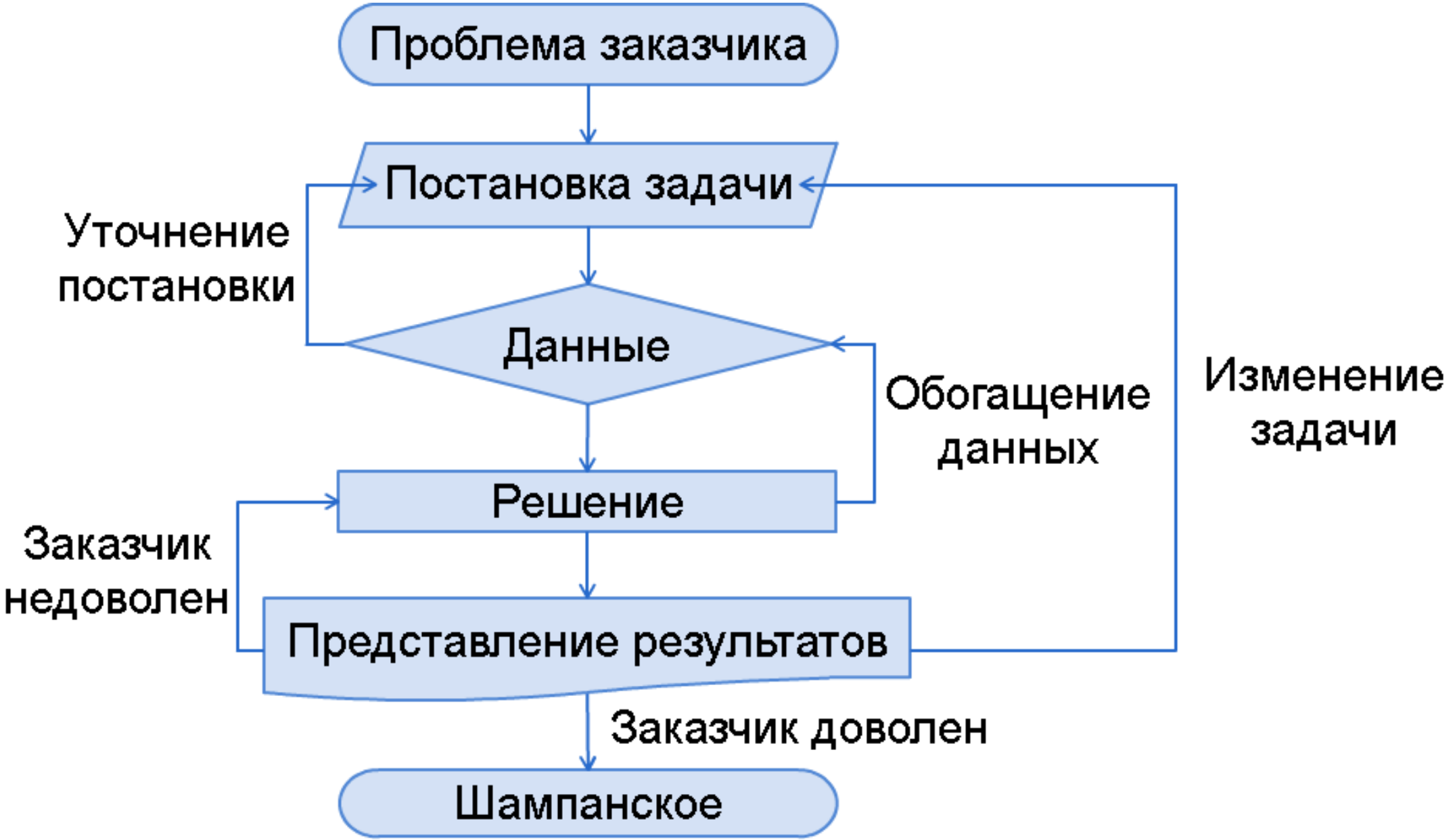
- Для обучения весов можно использовать различные линейные методы машинного обучения:
  - Логистическая регрессия
  - Линейный дискриминант Фишера
  - Метод опорных векторов (SVM)

- Если базовый входит в мета-алгоритм с отрицательным весом, то его нужно исключить из композиции

**Case:** Прогнозирование спроса на терапевтическую помощь Дмитрий Подвязников

# Процесс

ЦЕНТР АНАЛИЗА ДАННЫХ



# Проблема заказчика

ЦЕНТР
АНАЛИЗА
ДАННЫХ

Когда отпускать работников в отпуск?

Когда нанимать новых работников?

Когда отправлять работников на обучение?

Сколько заказать вакцин для прививок?

Как составить расписание работы?

Удостовериться, что всех пациентов приняли?

Как сделать так, чтобы не было очередей и все могли записаться?

# Постановка задачи

ЦЕНТР
АНАЛИЗА
ДАННЫХ

Что нужно?

Прогноз числа посещений в неделю, на 12 недель вперед

Когда нужно?

Через 2 недели

…

# Данные

ЦЕНТР АНАЛИЗА ДАННЫХ

| | ID | CLINIC_ID | SUNDAY_DATE | RECEIPTS | CHRONICS | PATIENTS | VISITS | DISTRICT |
|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 22 | 2015-02-01 | 1376.0 | 2561.0 | 52307.0 | 1491.0 | 1 |
| **1** | 1 | 22 | 2015-02-08 | 1376.0 | 2561.0 | 52307.0 | 1713.0 | 1 |
| **2** | 2 | 22 | 2015-02-15 | 1376.0 | 2561.0 | 52307.0 | 1725.0 | 1 |
| **3** | 3 | 22 | 2015-02-22 | 1376.0 | 2561.0 | 52307.0 | 1637.0 | 1 |
| **4** | 4 | 22 | 2015-03-01 | 1384.0 | 2570.0 | 48437.0 | 1431.0 | 1 |
| **5** | 5 | 22 | 2015-03-08 | 1384.0 | 2570.0 | 48437.0 | 1661.0 | 1 |
| **6** | 6 | 22 | 2015-03-15 | 1384.0 | 2570.0 | 48437.0 | 1355.0 | 1 |
| **7** | 7 | 22 | 2015-03-22 | 1384.0 | 2570.0 | 48437.0 | 1615.0 | 1 |
| **8** | 8 | 22 | 2015-03-29 | 1384.0 | 2570.0 | 48437.0 | 1567.0 | 1 |
| **9** | 9 | 22 | 2015-04-05 | 1443.0 | 2563.0 | 48845.0 | 1485.0 | 1 |

# Решение



KEEP
CALM
AND
DO YOUR
WORK

# Представление результатов

ЦЕНТР АНАЛИЗА ДАННЫХ

Рассказать про данные

- Какие важные закономерности удалось найти?
- Какое новое знание удалось выявить?

Объяснить вашу модель

- Почему такой подход?
- Почему выбрали / добавили эти фичи?

Обосновать решения

- Чем обусловлен выбор метрики?
- Почему такое деление на обучающую / тестовую выборки?

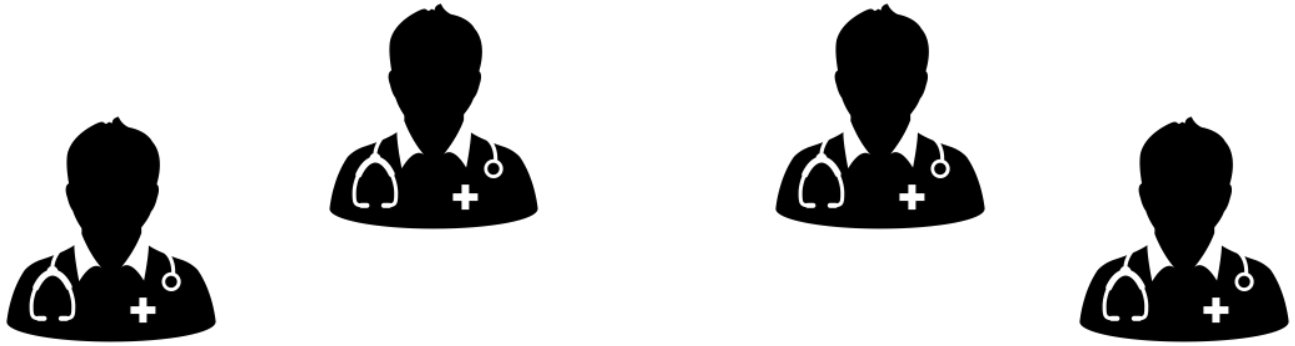
# Case: Балансировка терапевтических участков

Наталия Амелина

# Балансировка участков. Что сейчас?

ЦЕНТР АНАЛИЗА ДАННЫХ

Для уменьшения очередей к терапевтам, отменили принцип «участкового» врача. Пациенты могут записаться к любому доступному в поликлинике врачу



**КТО** теперь отвечает за пациента?

**КАК** оценить работу конкретного врача?

Участки не несут сейчас смысла и создают **РАЗНУЮ** нагрузку

# Постановка задачи заказчиком

ЦЕНТР АНАЛИЗА ДАННЫХ

Задача **восстановить участковую службу**:

- ✓ Сопоставимая нагрузка (количество приемов) с одного участка в период
- ✓ Нагрузка остается стабильной через N лет

- ✓ Минимальное количество людей меняет участок
- ✓ Не трогать тех, кто хорошо знает своего врача (untouchable)
- ✓ * Восстановить семьи
- ✓ * Выделить хроников в отдельные участки

Как меняется с **возрастом** вероятность визита к врачу**?**

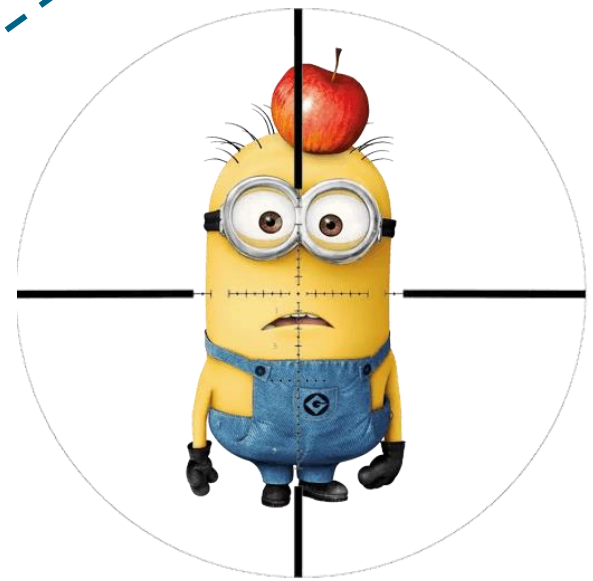
Лишь 34% застрахованных пришли к врачу за исследуемый период. Как работать с **неполными данными?**

В неидеальном мире нужно неидеальное решение: **когда** остановить балансировку**?**

# Наша интерпретация задачи

ЦЕНТР АНАЛИЗА ДАННЫХ

1. Скоринг человека



2. Балансировка участков

# Данные

ЦЕНТР АНАЛИЗА ДАННЫХ

| Признак | Описание |
|---|---|
| untouchable | признак пациента, которого нельзя переносить с участка |
| **pat_id** | идентификатор пациента |
| lpu_district | район ЛПУ (поликлиники) |
| pers_district | район проживания пациента |
| MOSCOW_REGISTRATION | признак наличия московской регистрации |
| invalid | признак наличия инвалидности |
| nadomnik | признак неходячего пациента |
| age_calculate_y | возраст |
| sovmest_deti | количество совместно проживающих детей до 14 лет |
| sovmest_lgotnik | количество совместно проживающих льготников |
| sovmest_pens | количество совместно проживающих пенсионеров |
| sphere_new_Учреждения образования | работник образования |
| sphere_new_Учреждения здравоохранения | работник зравоохранения |
| sphere_new_Учреждения соцзащиты | работник соц·защиты |
| sphere_ZHKH_ЖКХ | работник ЖКХ |
| bezrab_безработный сейчас | безработный на 01.07.2016 |
| bezrab_был безработный 2015/16 | был безработным в период 01.01.2016 – 30.06.2016 |
| bezrab_дело на рассмотрении | потенциально станет безработным |
| COMMON_GENDER | пол |
| lux1 | признак материального благосостояния № 1 |
| lux2 | признак материального благосостояния № 2 |
| FLAG_BESSR | признак бессрочного льготника |
| HAD_APPOINTMENTS | появлялся ли пациент хотя бы один раз за последние 2 года |
| from_region | региональный полис |
| **lpu_id** | идентификатор ЛПУ (поликлиники) |
| **district_id** | идентификатор участка |
| **target** | целевая переменная (количество приемов) - преобразованные |

# Представление результатов

ЦЕНТР АНАЛИЗА ДАННЫХ

Презентация «**как заказчику = врачу**» + рабочий ноутбук до 1 декабря

Представление результатов и обсуждение 3 декабря

1. **Pre-research**

Что интересного есть в данных?

2. **Скоринговая модель человека**

Обоснование выбора модели, метрики точности

Интерпретируемая модель! Не черный ящик, а набор правил понятный врачу

3. **Балансировка участков на скоринговой модели на 3 ЛПУ (13, 17, 19)**

Не трогать untouchable

Обосновать принцип / функционал балансировки, метрика точности

Почему с этим решением участки не разбалансируются в будущем?

В каком моменте алгоритм останавливается – квазиоптимальное решение?

# Решение кейсов до 1 декабря на n.amelina@analysiscenter.ru

ЦЕНТР АНАЛИЗА ДАННЫХ


**Feature Engineering** – Алексей Ушаков
a.ushakov@analysiscenter.ru

**Ensembles** – Кирилл Емельянов
k.emelyanov@analysiscenter.ru

**Case:** Прогнозирование спроса на терапевтическую помощь Дмитрий Подвязников, d.podvyaznikov@analysiscenter.ru

**Case:** Балансировка терапевтических участков
Наталия Амелина, n.amelina@analysiscenter.ru
amelinans@gmail.com