


Яндекс



# Данные для метеохакатона

Иван Бушмаринов, аналитик



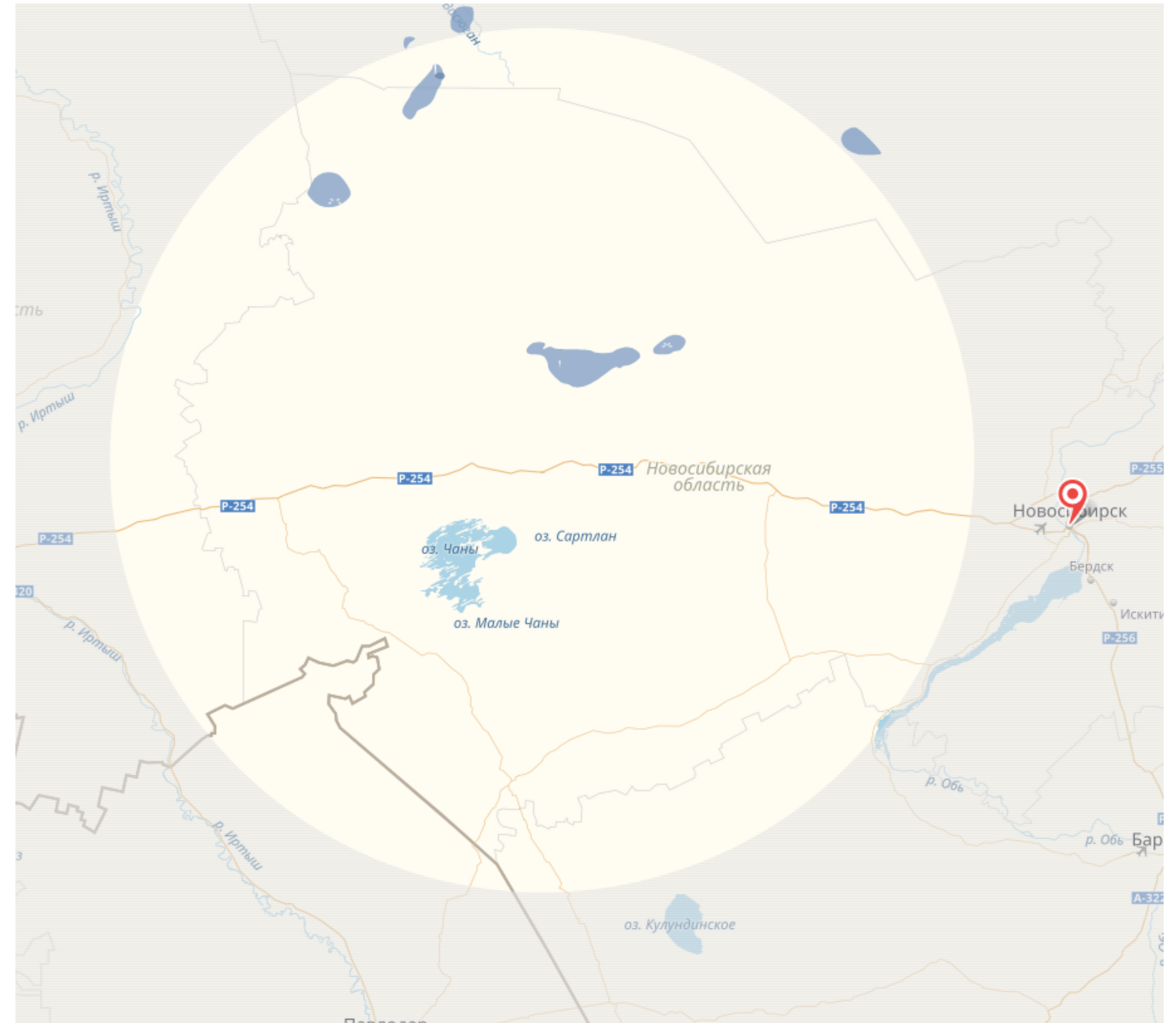
Толпа многоголова,  
как Гидра и Цербер,  
но она не делает  
погоду, как  
Гидрометцентр

Охххуmiron, «Слово мэра»

# Зачем

Радарные данные для дождей — это замечательно, но они доступны не во всех городах. Например, в Новосибирске положение радара не совпадает с положением города.

Аналогично Пробкам хотим собирать карту по пользовательским данным



# Источники данных



Вышки сотовой связи



Любительские метеостанции  
Netatmo



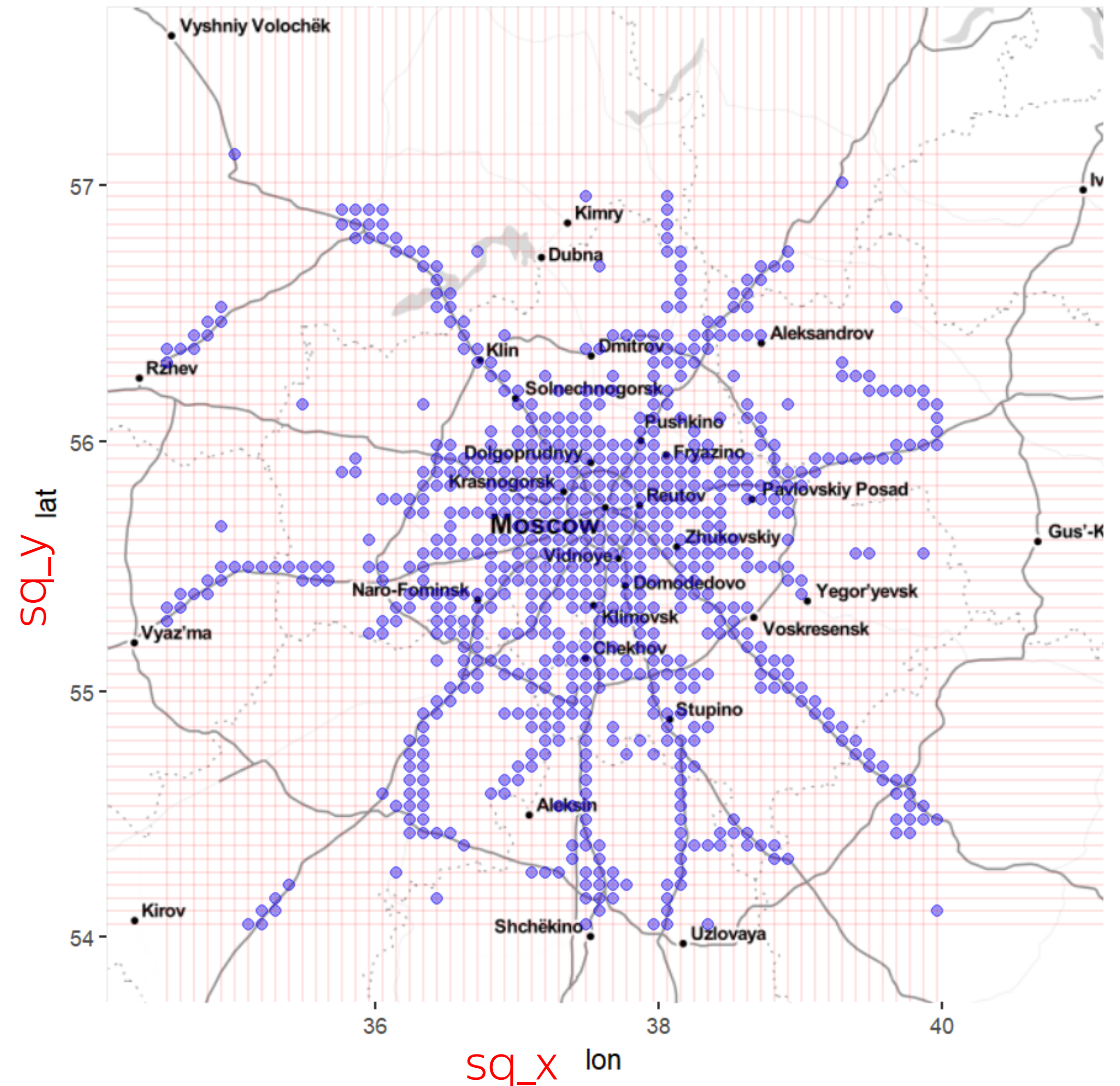
# Агрегированные радарные данные (таргет)

Дождь на момент конца часа,  
усредненный по квадрату 6\*6 км.  
 $\text{rain} = \text{precipitation} > 0.25$

Все данные изначально разбиты по  
квадраточасам. Предлагается  
предсказывать для квадраточасов с  
пользовательскими данными.

$(\text{sq\_time}, \text{city\_code}) \rightarrow \text{hour\_hash}$

$(\text{sq\_x}, \text{sq\_y}, \text{hour\_hash}) \rightarrow \text{rain}$



# По каким данным предсказываем?

Семплированные (по Москве 0.1% пользователей), обфусцированные данные из Аппметрики + практически сырые данные любительских метеостанций в Москве, Санкт-Петербурге и Казани.

hackathon\_tosubmit.tsv, test\_kazan\_netatmo.tsv,  
test\_spb\_features.tsv, train\_kazan\_netatmo.tsv, train\_msk.tsv,  
test\_msk\_features.tsv, test\_spb\_netatmo.tsv, train\_kazan.tsv,  
train\_spb\_netatmo.tsv, test\_kazan\_features.tsv, test\_msk\_netatmo.tsv,  
train\_msk\_netatmo.tsv, train\_spb.tsv

# hackathon\_tosubmit.tsv

**id** — уникальный номер квадрато-часа (1 точка таргета)

**hour\_hash** — уникальный id часа конкретного дня в конкретном городе

**sq\_x** — положение квадрата по горизонтали

**sq\_y** — положение квадрата по вертикали

Сабмитим вероятность дождя, проверка по ROC AUC.



# Постановка задачи

Цель – для квадрато часа с пользовательскими данными восстановить, шел дождь или нет.

Предположительно, дождь ослабляет сигнал от телефона до сотовой вышки (основные файлы train и test). Также дождь регистрируется любительскими метеостанциями (файлы netatmo).

Тренировочные данные – начало июля 2017. Тестовые – июль-август 2017.

# Данные о сигнале (обучающий набор)

Каждая строка — измерение 1 сигнала 1 пользователя в 1 момент времени до 1 вышки.

hours\_since, sq\_time, precipitation, rain — *только в тренировочном наборе*

precipitation — осадки, мм/ч

rain = precipitation > 0.25 — **таргет**

sq\_time — время измерения (timestamp, UTC)

hours\_since — число часов с полуночи 1 июля по Москве.

# Данные о сигнале (квадраточас)

`sq_lat` — широта центра квадрата, градусы

`sq_lon` — долгота центра квадрата, градусы

`sq_x` — абсцисса квадрата на сетке агрегации, целая

`sq_y` — ордината квадрата на сетке агрегации, целая

`day_hour` = `hours_since` % 24

`city_code` — город (16, 77 или 78)

`hour_hash` — хэш часа измерения в конкретном городе

# Данные о сигнале (пользователь)

`u_hashed` — обфусцированный DeviceID (в тесте постоянен в пределах часа)

`ver_hash` — обфусцированная версия Андроида

`device_model_hash` — обфусцированная модель телефона

`ulat` — широта пользователя

`ulon` — долгота пользователя



# Данные о сигнале (сигнал)

EventTimestampDelta — время до конца часа, с

SignalStrength — сила сигнала, дБ

cell\_hash — обфусцированный ID ячейки

LAC — Location Area Code ячейки

OperatorID — оператор (1 — МТС, 99 — Билайн...)

eventid — идентификатор измерения

radio — тип станции (1 — GSM, 2 — LTE, 3 — UMTS)

# Данные о сигнале (локация)

■ Все локации по GPS.

**LocationTimestampDelta** — время от определения положения до конца часа, с

**LocationAltitude** — высота, м

**LocationDirection** — направление телефона

**LocationPrecision** — точность определения положения, м

**LocationSpeed** — скорость движения

# Данные о сигнале (OpenCellID)

■ Для части вышек доступны открытые координаты.

`cell_lat` — широта вышки

`cell_lon` — долгота вышки

`range` — "дальнобойность" вышки

# Netatmo

Метеоданные с шагом в 20 минут, также за квадраточас. Названия полей говорят сам за себя.

`city_code, day_hour, hour_hash, netatmo_humidity_percent,  
netatmo_latitude, netatmo_longitude, netatmo_pressure_mbar,  
netatmo_sum_rain_1h, netatmo_sum_rain_24h,  
netatmo_temperature_c, netatmo_timestamp_delta, netatmo_uid,  
netatmo_wind_direction_deg, netatmo_wind_gust_direction_deg,  
netatmo_wind_gust_speed_kmh, netatmo_wind_speed_kmh,  
point_latitude = sq_lat, point_longitude = sq_lon, sq_x, sq_y`



# Читерство

**Явное читерство** — использование внешних данных без согласования с организаторами. Если мы одобрим использование какого-то источника, об этом будет объявлено публично и ссылка появится в хакатонном чатике. Данные традиционного метеопрогнозирования заведомо не будут одобрены.

**Неявное читерство** — использование фичей из будущего. Обфускация данных и привязка к конкретному часу это частично предотвращают, но нужно быть готовым ответить на вопросы по решению, если попадете в топ3 и будете претендовать на призы.