

1
0
1
0
1
1 0 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1

ОТ ТЕЛЕФОННЫХ ОПРОСОВ В МИР СИНТЕТИЧЕСКИХ ЛЮДЕЙ

1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 1 0 0 1 0 0 1

МАРИЯ ЧАНИНА
РУКОВОДИТЕЛЬ ОТДЕЛА АНАЛИТИКИ ПАНЕЛЬНЫХ ИССЛЕДОВАНИЙ

1 0 1
0 0

БЕНИАМИН ЕРМОЛАЕВ
РУКОВОДИТЕЛЬ ОТДЕЛА ИССЛЕДОВАНИЙ И РАЗРАБОТОК ПО ИНТЕГРАЦИИ ДАННЫХ

1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 0 1 0 1 0 1
0 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1

1 1 1 0 0 1 1 1 1 1 1 0 1 1 1 0 1 1 0 1 0 0

0 0 0 1 1 0 0 0 0 1 1 0 1 0 1 0 1 0 1 0 1 1

1 1 0 1 1 1 1 0 0 0 1 1 1 0 0 1 0 0 0 0 0 0

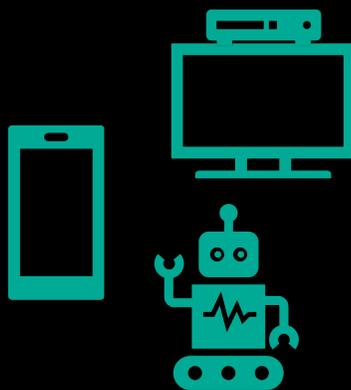
1 1 0 0 0 0 1 1 1 1 1 0 1 0 1 1 1 1 1 1 1 1

1 1 0 1 0 1 1 0 1 1 1 0 1 0 0 0 1 1 1 1 1 1

0 0 1 0 1 0 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1

RESEARCH

ЭВОЛЮЦИЯ ИССЛЕДОВАНИЙ / МИФ



```
1 1 1 1 1
0 0 0 0 0
1 1 1 1 1
0 0 0 0 0
1 1 1
1 1 1 1 1
1 0 1 1 0
```

BIG DATA

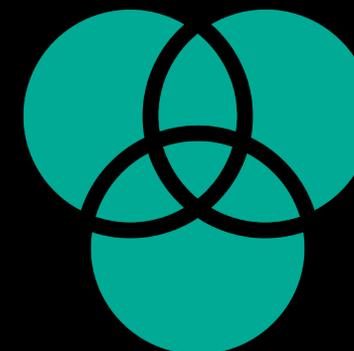
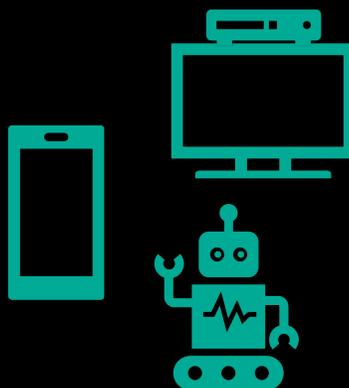


RESEARCH

ЭВОЛЮЦИЯ ИССЛЕДОВАНИЙ / РЕАЛЬНОСТЬ

СОЦИОЛОГИЯ

МАТЕМАТИКА

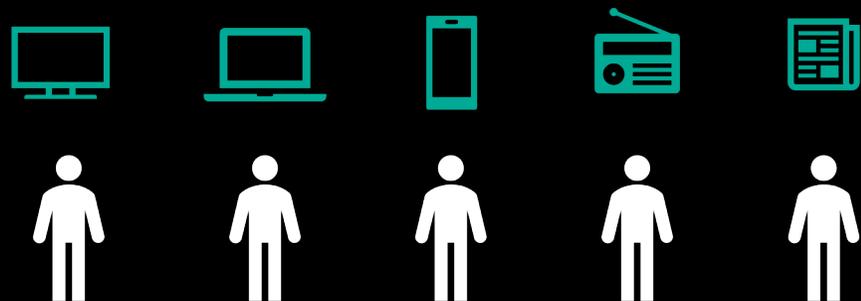


BIG DATA



RESEARCH

ЭВОЛЮЦИЯ МЕДИА ИССЛЕДОВАНИЙ



ИССЛЕДОВАНИЕ ОТДЕЛЬНЫХ МЕДИА



КРОСС-МЕДИА



SAME MEDIA, DIFFERENT SOURCES

ПАНЕЛИ КАК ОСНОВА ИЗМЕРЕНИЙ

ПАНЕЛЬ – ГРУППА,
РЕПРЕЗЕНТИРУЮЩАЯ
НАСЕЛЕНИЕ

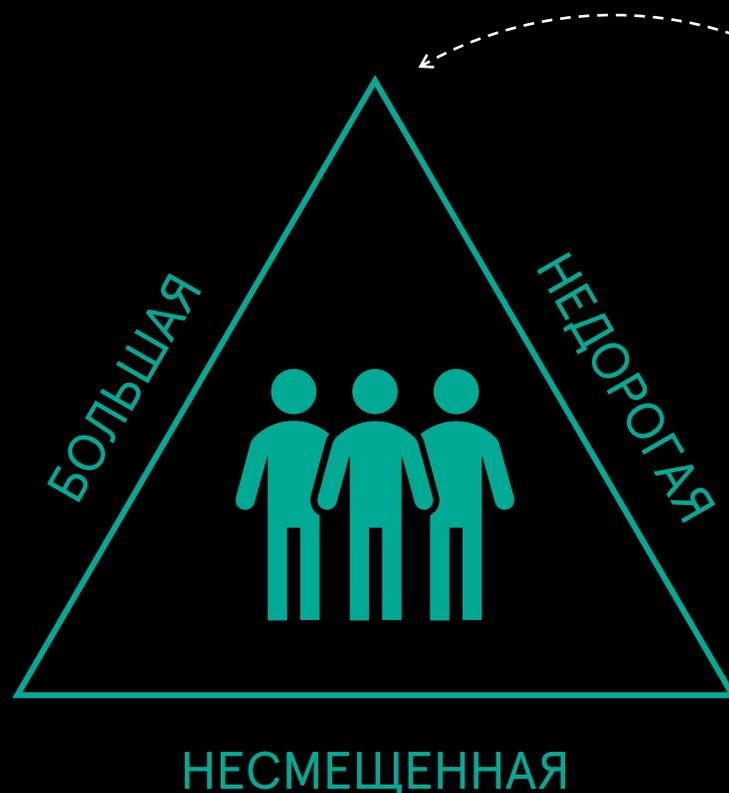


ИДЕАЛЬНАЯ ПАНЕЛЬ:

- БОЛЬШАЯ
- НЕСМЕЩЕННАЯ

ПОДХОДЫ К НАБОРУ ПАНЕЛИ

НАБРАТЬ БОЛЬШУЮ
НЕСМЕЩЕННУЮ ПАНЕЛЬ –
СЛОЖНО И ДОРОГО



НЕДОСТАТОК
DIGITAL РЕКРУТМЕНТА –
НЕИЗВЕСТНЫЕ СМЕЩЕНИЯ

СМЕЩЕНИЯ И ИХ КОРРЕКЦИЯ

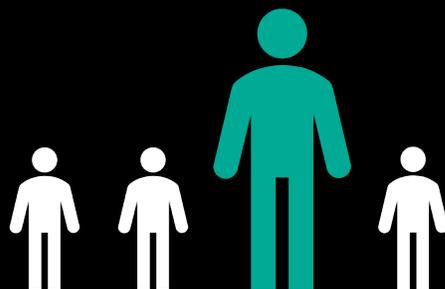
СМЕЩЕНИЯ:

01

ИЗВЕСТНЫЕ

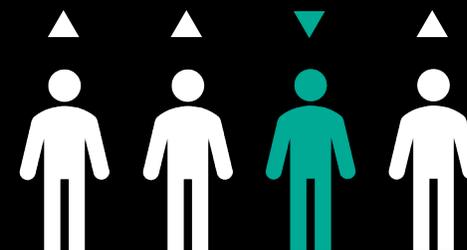
ПРИМЕР:

УВЕЛИЧЕННАЯ МИНИ-ПАНЕЛЬ
ВНУТРИ ОСНОВНОЙ ПАНЕЛИ
(ЖИТЕЛИ ГОРОДА)



КОРРЕКЦИЯ:

РОТАЦИЯ
ВЗВЕШИВАНИЕ



СМЕЩЕНИЯ И ИХ КОРРЕКЦИЯ

СМЕЩЕНИЯ:

ПРИМЕР:

КОРРЕКЦИЯ:

01

ИЗВЕСТНЫЕ

УВЕЛИЧЕННАЯ МИНИ-ПАНЕЛЬ
ВНУТРИ ОСНОВНОЙ ПАНЕЛИ
(ЖИТЕЛИ ГОРОДА)

РОТАЦИЯ
ВЗВЕШИВАНИЕ

02

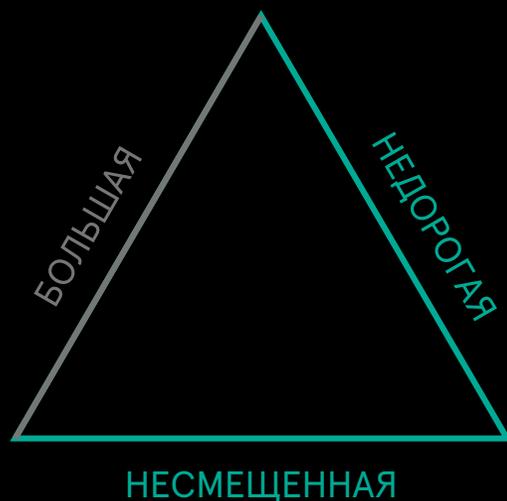
НЕИЗВЕСТНЫЕ

ПОСЛЕДСТВИЯ НЕСЛУЧАЙНОГО
РЕКРУТМЕНТА (ПОЛЬЗОВАТЕЛИ
САЙТОВ С БАННЕРАМИ)

PROPENSITY
SCORE
ADJUSTMENT

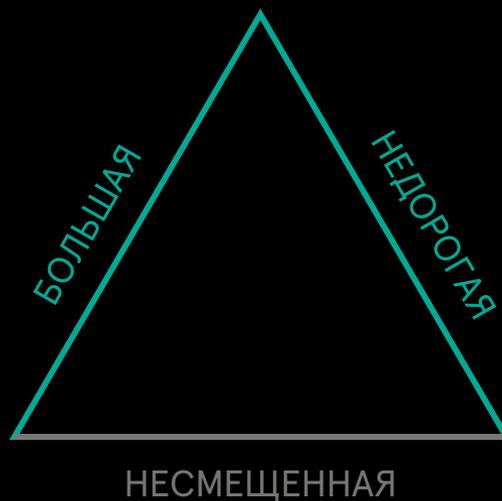
PROPENSITY SCORE ADJUSTMENT

ПАНЕЛЬ № 1
КАЧЕСТВЕННАЯ,
НО МАЛЕНЬКАЯ

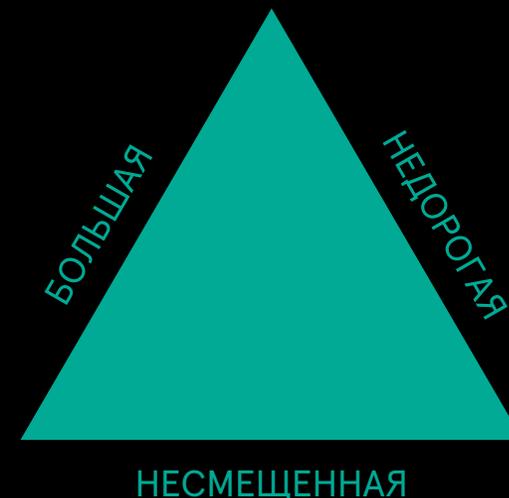


+

ПАНЕЛЬ № 2
БОЛЬШАЯ,
НО СМЕЩЕННАЯ



=



PROPENSITY SCORE ADJUSTMENT

ЭТАЛОННАЯ (1) И
КОРРЕКТИРУЕМАЯ (0)
ПАНЕЛИ

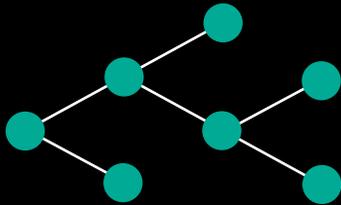


КАЧЕСТВЕННАЯ,
НО МАЛЕНЬКАЯ



БОЛЬШАЯ,
НО СМЕЩЕННАЯ

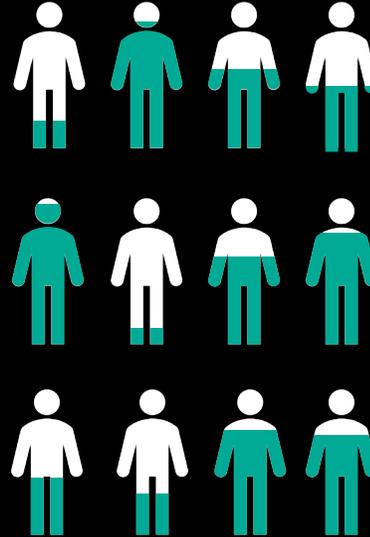
АЛГОРИТМ
КЛАССИФИКАЦИИ
(XGBOOST)



КЛАССИФИКАЦИЯ
ПО ВСЕМ
ПРИЗНАКАМ:
ДЕМОГРАФИЯ,
ГЕОГРАФИЯ,
АКТИВНОСТЬ

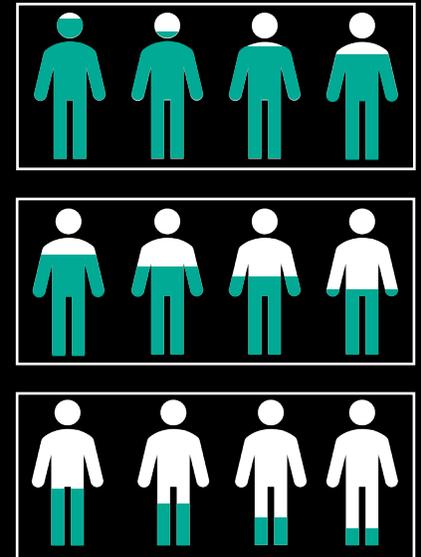
XGBoost

ВЕРОЯТНОСТИ
ПРИНАДЛЕЖНОСТИ
К ЭТАЛОННОЙ ПАНЕЛИ



ВЕРОЯТНОСТИ
ОТ 0 ДО 1

СОРТИРОВКА И
РАСЧЕТ ПОПРАВОЧНЫХ
КОЭФФИЦИЕНТОВ

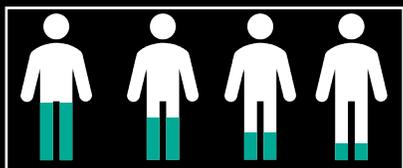
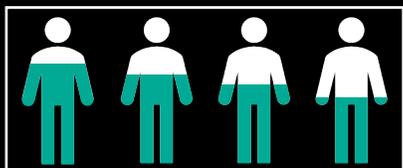
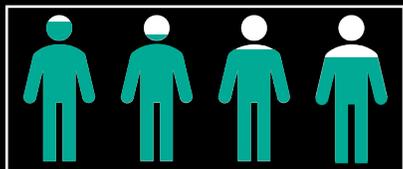


PROP_SCORE =
 $P(1)/(1 - P(1)) = P(1)/P(0)$

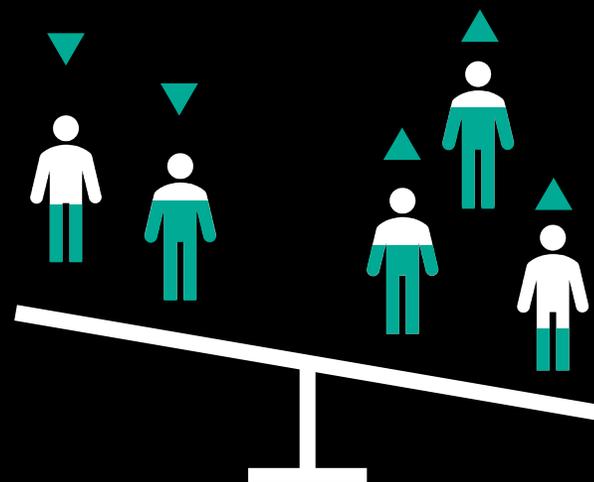
SAME MEDIA, DIFFERENT SOURCES

PROPENSITY SCORE ADJUSTMENT

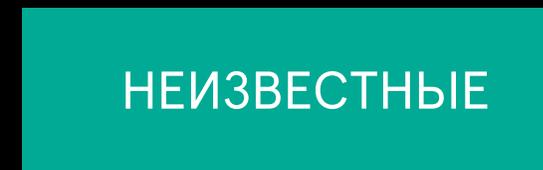
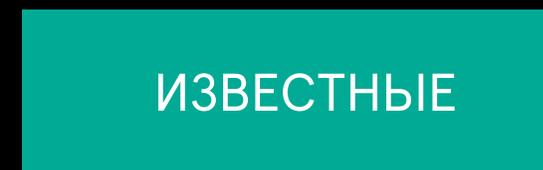
ПОПРАВочНЫЕ
КОЭФФИЦИЕНТЫ –
СТАРТОВЫЕ ВЕСА



КЛАССИЧЕСКАЯ
ПРОЦЕДУРА
ВЗВЕШИВАНИЯ

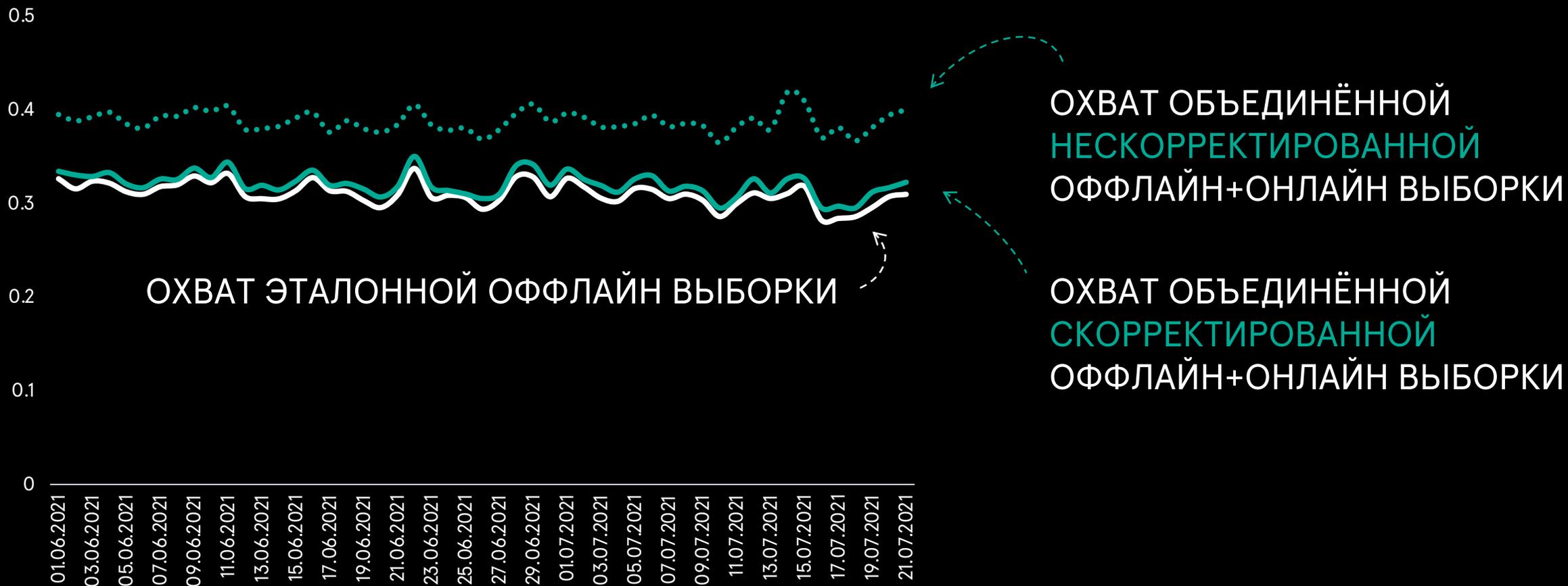


СКОРРЕКТИРОВАНЫ ВСЕ
СМЕЩЕНИЯ – ИЗВЕСТНЫЕ
И НЕИЗВЕСТНЫЕ



SAME MEDIA, DIFFERENT SOURCES

PROPENSITY SCORE ADJUSTMENT / ИТОГ



PROPENSITY SCORE ADJUSTMENT | ПРИМЕРЫ

PEW RESEARCH

2016

[EVALUATING ONLINE NONPROBABILITY SURVEYS](#)

Courtney Kennedy, Andrew Mercer, Scott Keeter, Nick Hatley, Kyle Mcgeeney, Alejandra Gimenez

2017

[THEORY AND PRACTICE IN NONPROBABILITY SURVEYS: PARALLELS BETWEEN CAUSAL INFERENCE AND SURVEY INFERENCE](#)

Andrew Mercer, Frauke Kreuter, Scott Keeter, Elizabeth Stuart

2018

[FOR WEIGHTING ONLINE OPT-IN SAMPLES, WHAT MATTERS MOST?](#)

Andrew Mercer, Arnold Lau, Courtney Kennedy

HORVITZ THOMPSON ESTIMATOR

2016

[A CASE STUDY OF THE IMPACT OF DATA-ADAPTIVE VERSUS MODEL-BASED ESTIMATION OF THE PROPENSITY SCORES ON CAUSAL INFERENCES FROM THREE INVERSE PROBABILITY WEIGHTING ESTIMATORS](#)

Romain Neugebauer, Julie A. Schmittiel, Mark J. Van Der Laan

2020

[PROPENSITY SCORE ADJUSTMENT USING MACHINE LEARNING CLASSIFICATION ALGORITHMS TO CONTROL SELECTION BIAS IN ONLINE SURVEYS](#)

Ramón Ferri-garcía, María Del Mar Rueda

GOOGLE RESEARCH

2016

[A METHOD FOR MEASURING ONLINE AUDIENCES](#)

Jim Koehler, Evgeny Skvortsov, Wiesner Vos

AAPOR

2010

[AAPOR REPORT ON ONLINE PANELS](#)

Reg Baker, Stephen Blumberg and others

MEDICAL RESEARCH

2018

[PROPENSITY SCORE METHODS: THEORY AND PRACTICE FOR ANESTHESIA RESEARCH](#)

Phillip Schulte, Edward Mascha

2020

[PROPENSITY SCORE WEIGHTING FOR COVARIATE ADJUSTMENT IN RANDOMIZED CLINICAL TRIALS](#)

Shuxi Zeng, Fan Li, Rui Wang

MEDIASCOPE RESEARCH – СКОПО НА HABR.COM

PART 2

DIFFERENT MEDIA, DIFFERENT SOURCES

1
0
1
0
1
1 0 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 0 1 1 0 1 1 1 1 0 1 1 0 0
0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 1 0 0 0 1
0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1
0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1
1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1
0
0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 0 0
0 0 1 1 1 1 1 1 1 1 1 0 0 0 1 1 0 1 1 1
1 1 1 1 1 1 0 1 1 1 1 0 1 1 1 0 1 1 0 0
0 0 0 1 1 1 1 1 0 0 0 0 0 0 1 1 0 1 0 0
1 1 0 0 0 0 1 0 0 0 1 1 1 0 0 1 0 0 0 0
1 1 0 0 0 0 0 1 1 1 1 0 1 0 1 1 1 1 1 1
1 1 0 1 0 1 1 1 0 0 1 1 0 1 0 0 0 1 1 1
0 0 1 0 1 0 1 0 1 1 0 0 1 1 1 1 1 1 1 1

КАК ЭТО БЫЛО УСТРОЕНО РАНЬШЕ



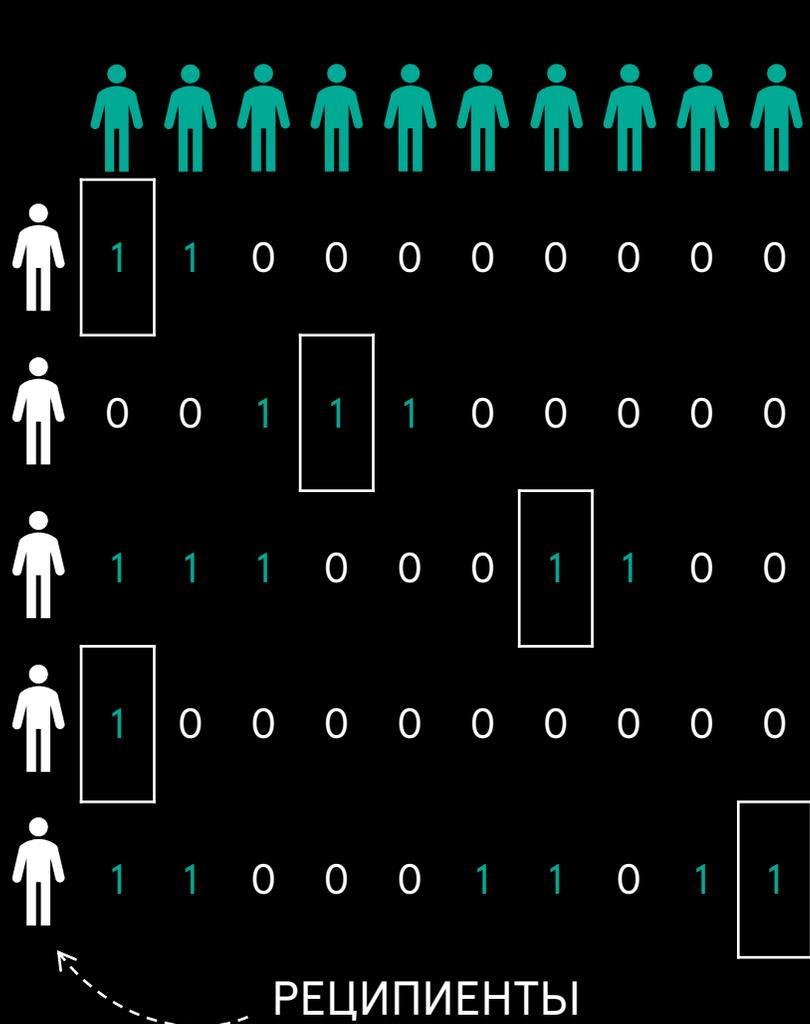
ДОНОРЫ

ЗАДАЧА ПОИСКА ПОХОЖИХ РЕСПОНДЕНТОВ
РЕШАЕТСЯ ВНУТРИ **СТРАТ**

СТРАТЫ – ЭТО МНОЖЕСТВО РЕСПОНДЕНТОВ
С ОДИНАКОВЫМ СОЦ-ДЕМ ПРОФИЛЕМ

РЕЦИПИЕНТЫ

КАК ЭТО БЫЛО УСТРОЕНО РАНЬШЕ



ДОНОРЫ

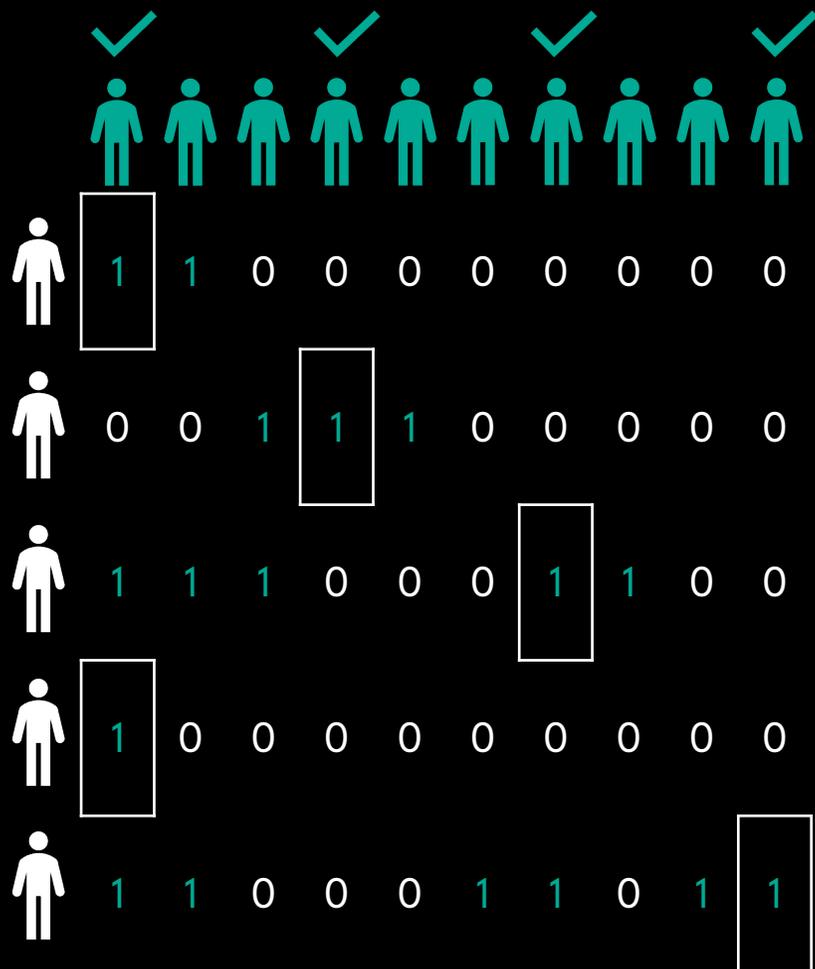
ЗАДАЧА ПОИСКА ПОХОЖИХ РЕСПОНДЕНТОВ
РЕШАЕТСЯ ВНУТРИ **СТРАТ**

СТРАТЫ – ЭТО МНОЖЕСТВО РЕСПОНДЕНТОВ
С ОДИНАКОВЫМ СОЦ-ДЕМ ПРОФИЛЕМ

ИЗ ОПТИМАЛЬНОГО РЕШЕНИЯ ВЫДЕЛЯЕТСЯ
ПОДМНОЖЕСТВО

ИТОГОВОЕ РЕШЕНИЕ ПРЕДСТАВЛЯЕТ СОБОЙ
МАТРИЦУ «**ОДИН КО МНОГИМ**»

КАК ЭТО БЫЛО УСТРОЕНО РАНЬШЕ



ИСПОЛЬЗУЕМ ДАННЫЕ **ТОЛЬКО ЧАСТИ** РЕСПОНДЕНТОВ ДОНОРСКОЙ ПАНЕЛИ



СИЛЬНАЯ **ЗАВИСИМОСТЬ РЕЗУЛЬТАТОВ** ОТ РОТАЦИИ ДОНОРСКОЙ ПАНЕЛИ

КАК ЭТО УСТРОЕНО ТЕПЕРЬ



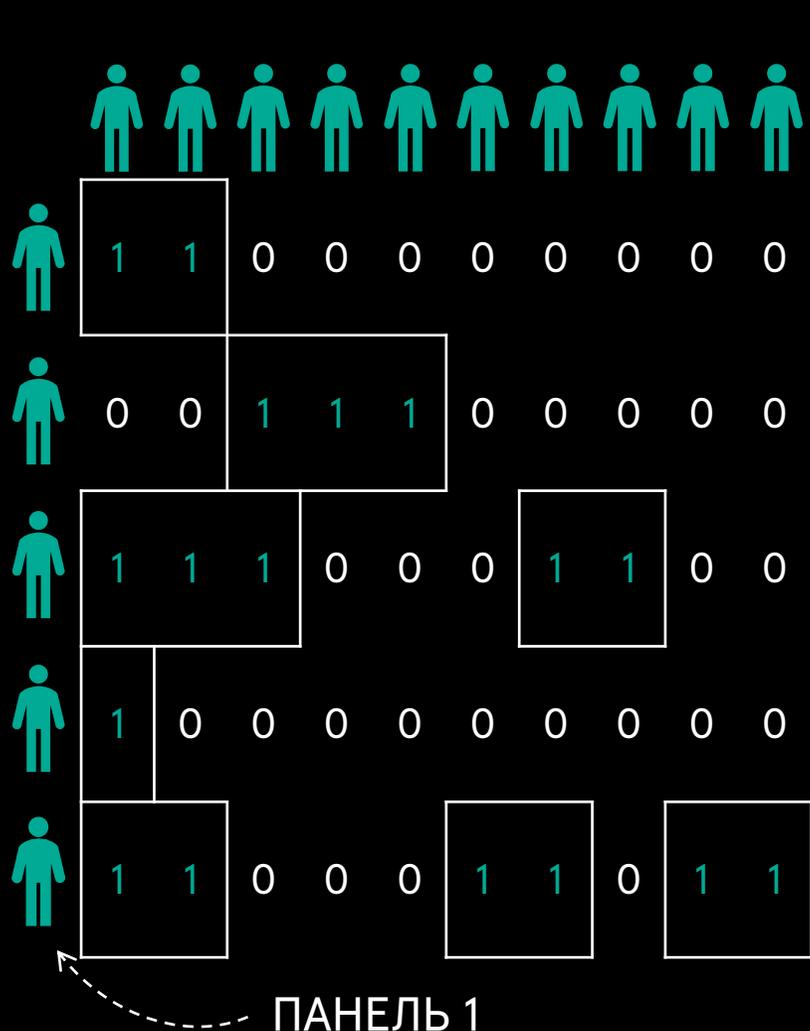
ПАНЕЛЬ 2

ЗАДАЧА ПОИСКА ПОХОЖИХ РЕСПОНДЕНТОВ
РЕШАЕТСЯ ВНУТРИ **СТРАТ**

СТРАТЫ – ЭТО МНОЖЕСТВО РЕСПОНДЕНТОВ
С ОДИНАКОВЫМ СОЦ-ДЕМ ПРОФИЛЕМ

ПАНЕЛЬ 1

КАК ЭТО УСТРОЕНО ТЕПЕРЬ



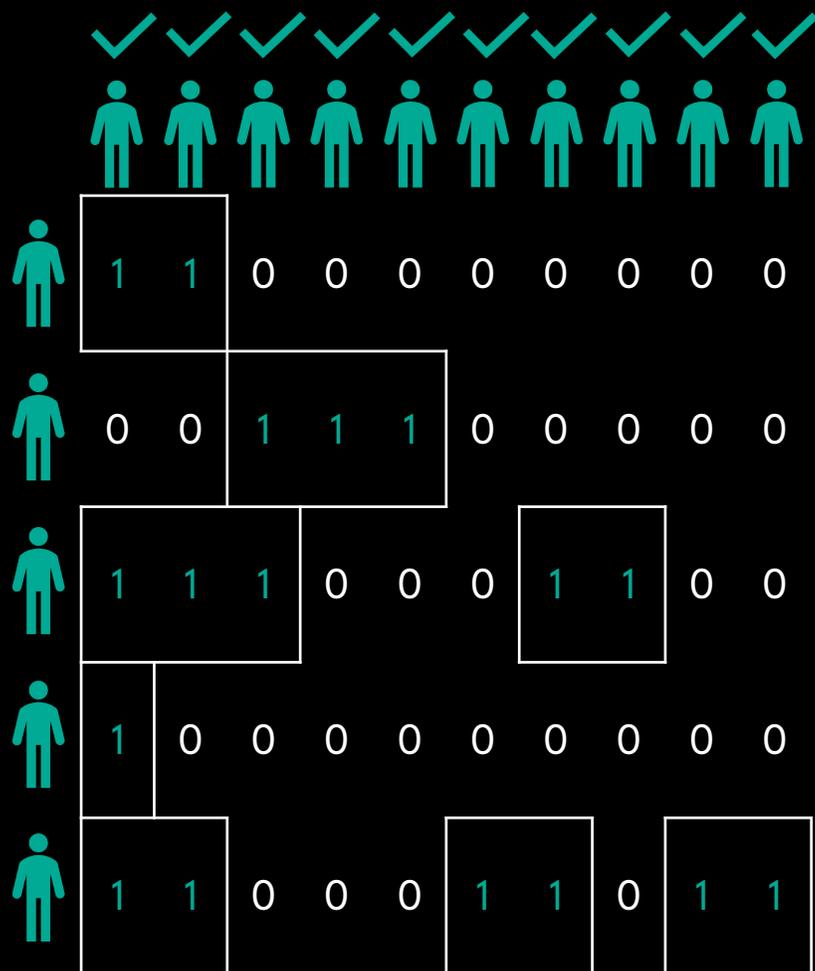
ПАНЕЛЬ 2

ЗАДАЧА ПОИСКА ПОХОЖИХ РЕСПОНДЕНТОВ
РЕШАЕТСЯ ВНУТРИ **СТРАТ**

СТРАТЫ – ЭТО МНОЖЕСТВО РЕСПОНДЕНТОВ
С ОДИНАКОВЫМ СОЦ-ДЕМ ПРОФИЛЕМ

ОПТИМАЛЬНОЕ РЕШЕНИЕ ПРЕДСТАВЛЯЕТ
СОБОЙ МАТРИЦУ ПАР «**МНОГИЕ КО МНОГИМ**»

КАК ЭТО УСТРОЕНО ТЕПЕРЬ



ИСПОЛЬЗУЕМ ДАННЫЕ
ПО ВСЕМ ПАНЕЛИСТАМ



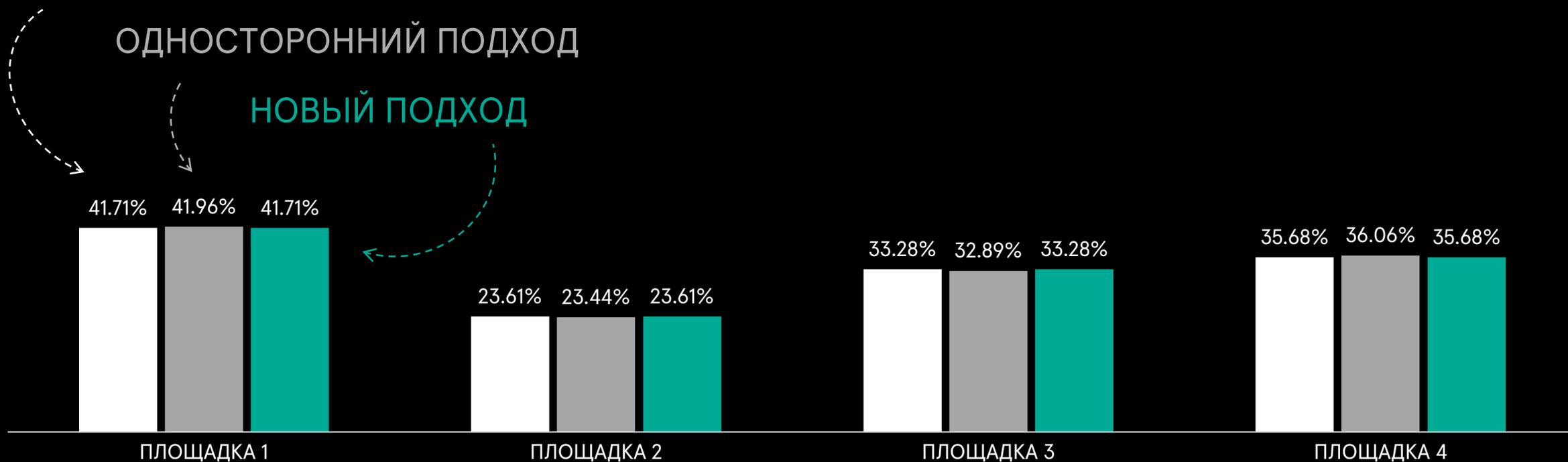
ТОЧНО ВОСПРОИЗВОДИМ
ДНЕВНЫЕ ПОКАЗАТЕЛИ



ВОСПРОИЗВОДИМ КАРТИНУ
НАКОПЛЕННЫХ ПОКАЗАТЕЛЕЙ
СО ЗНАЧИТЕЛЬНО МЕНЬШЕЙ
ПОГРЕШНОСТЬЮ

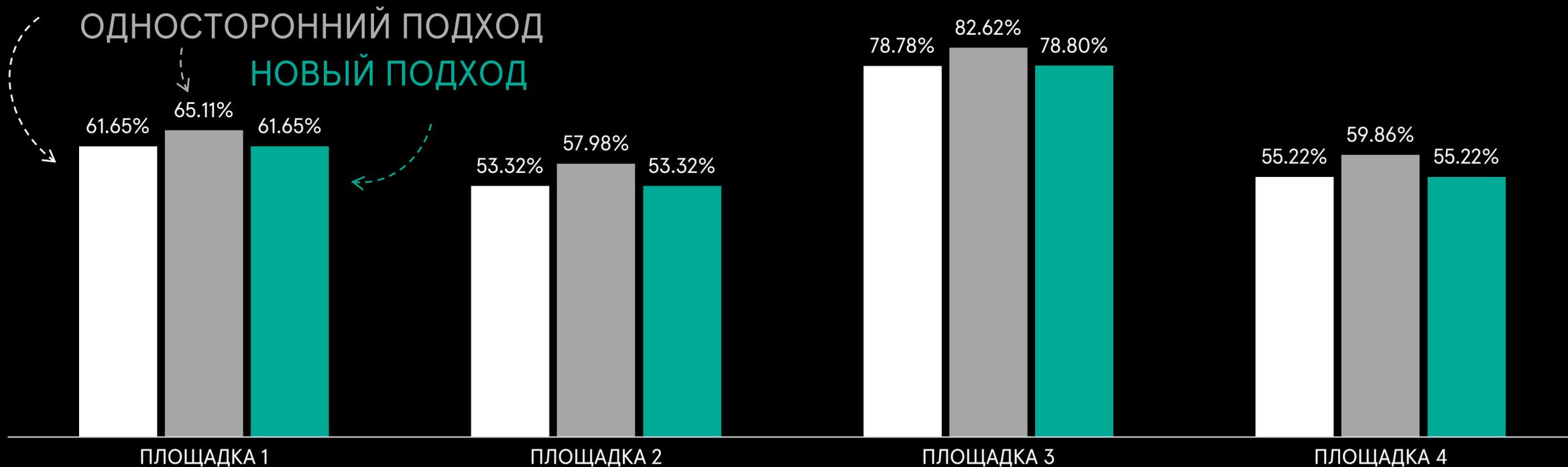
ОХВАТ ЗА ДЕНЬ

ДОНОРСКАЯ ПАНЕЛЬ



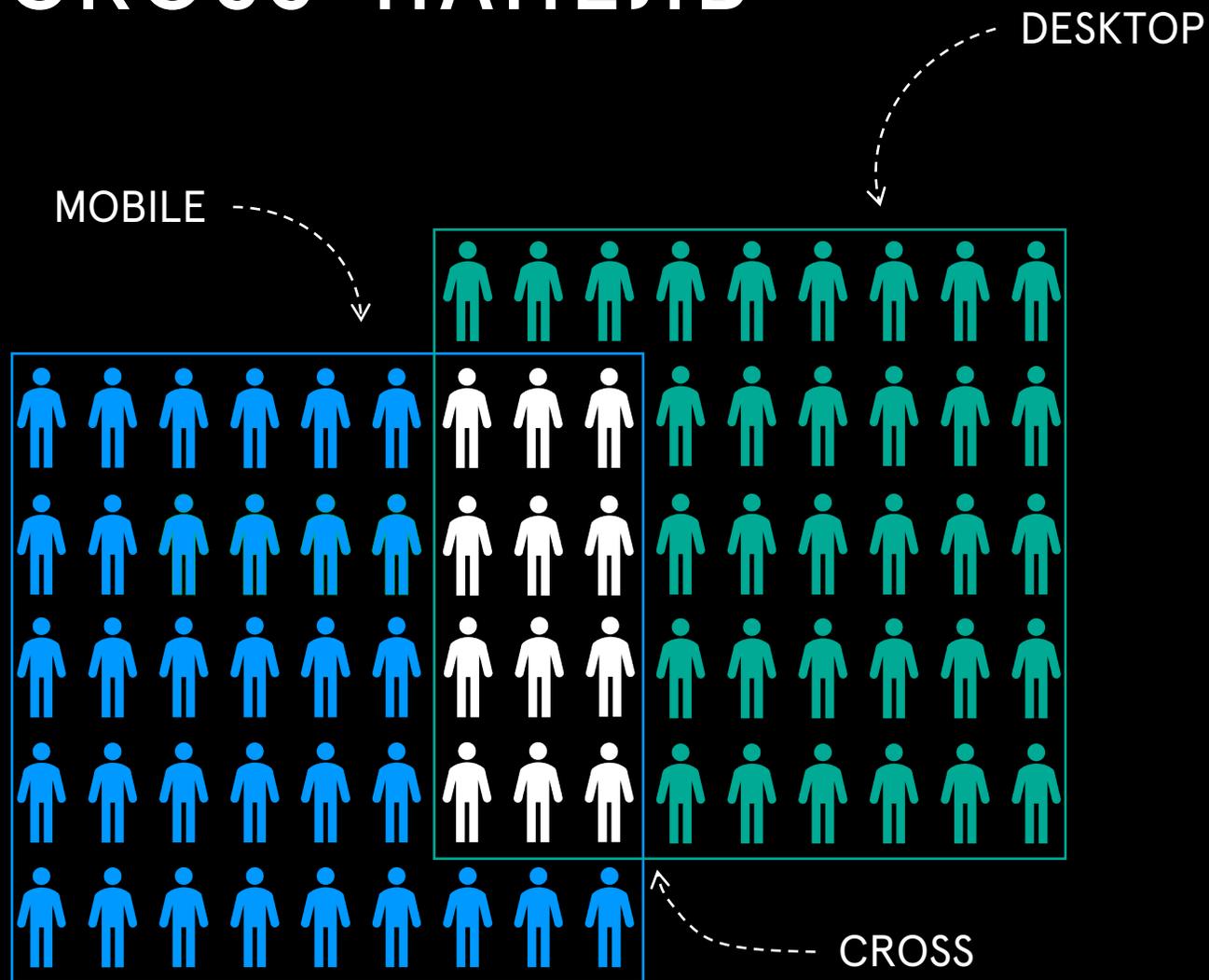
ОХВАТ ЗА МЕСЯЦ

ДОНОРСКАЯ ПАНЕЛЬ



DIFFERENT MEDIA, DIFFERENT SOURCES

CROSS-ПАНЕЛЬ

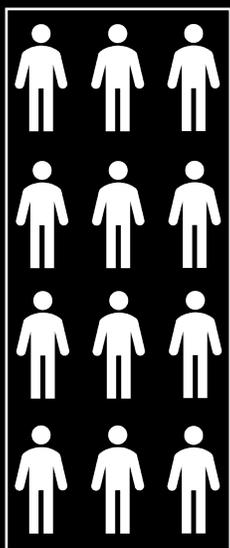


~4 500

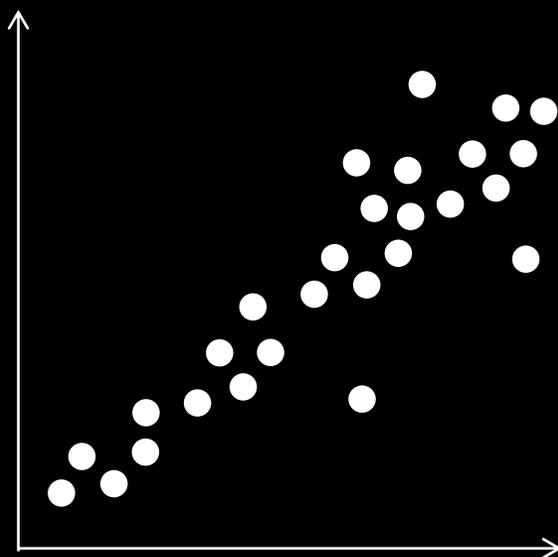
ЧЕЛОВЕК В CROSS-ПАНЕЛИ ЕЖЕДНЕВНО

ИСПОЛЬЗОВАНИЕ CROSS-ПАНЕЛИ

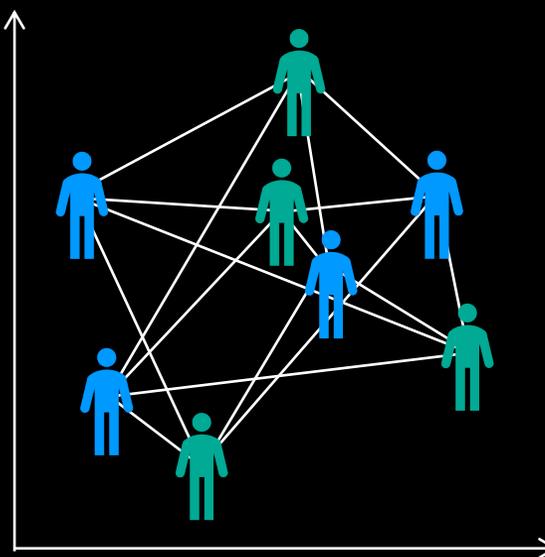
ДАННЫЕ
CROSS
ПАНЕЛИ



ПОИСК ЗНАЧИМЫХ
КОРРЕЛЯЦИЙ (CANONICAL
CORRELATION ANALYSIS)

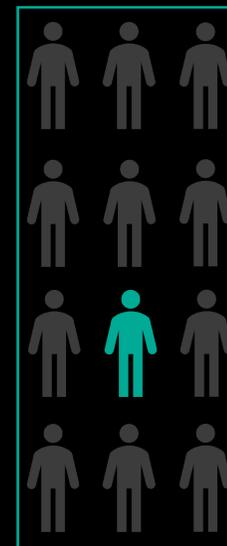


РАСЧЕТ РАССТОЯНИЙ
МЕЖДУ РЕСПОНДЕНТАМИ
ОСНОВНЫХ ПАНЕЛЕЙ

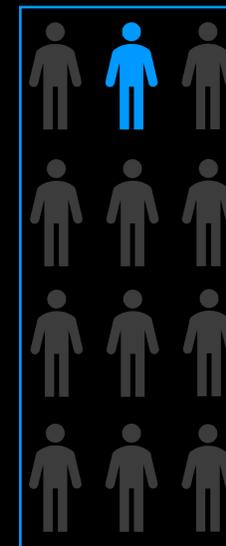


ПОИСК В ПАНЕЛЯХ
НАИБОЛЕЕ ПОХОЖИХ
РЕСПОНДЕНТОВ

DESKTOP



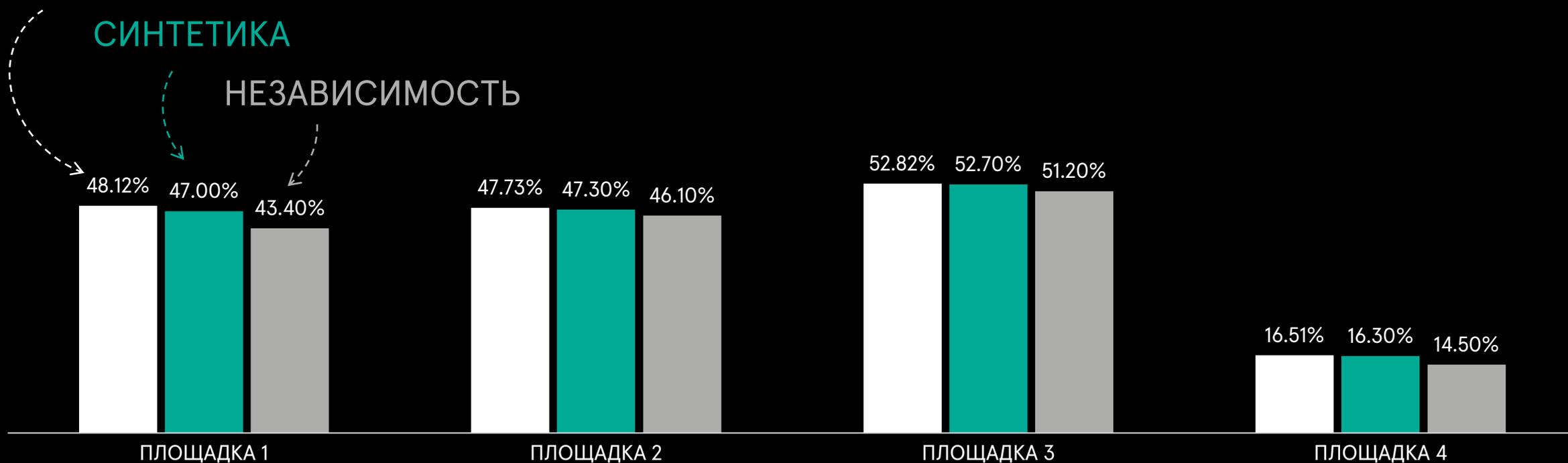
MOBILE



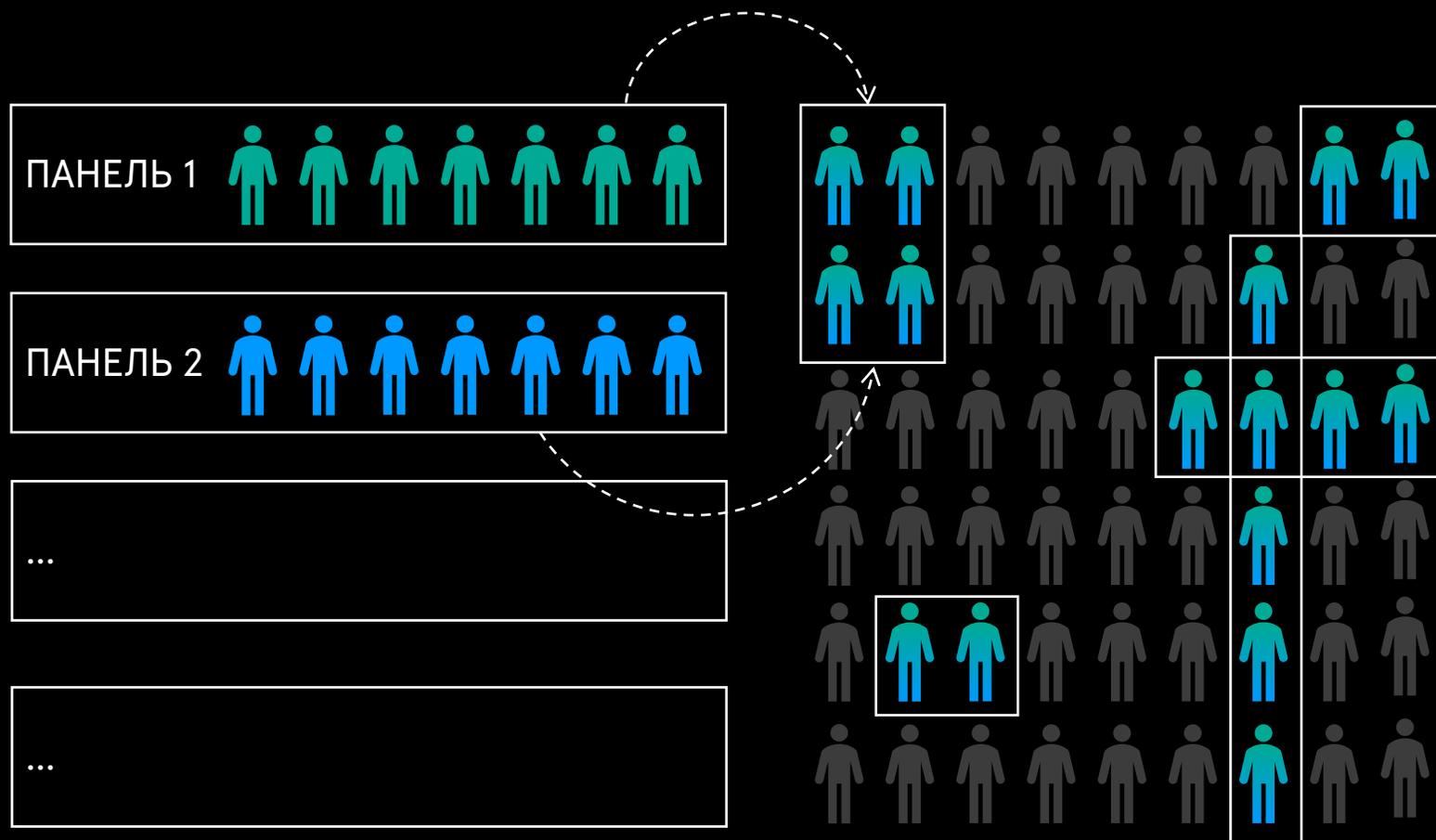
DIFFERENT MEDIA, DIFFERENT SOURCES

CROSS-ОХВАТ

CROSS-ПАНЕЛЬ



СИНТЕТИЧЕСКАЯ ПОПУЛЯЦИЯ



РЕПОРТИНГ ДАННЫХ
ПО **ЮНИТАМ**

ЮНИТ - СИНТЕТИЧЕСКИЙ
РЕСПОНДЕНТ, ПОЛУЧЕННЫЙ
В РЕЗУЛЬТАТЕ ИНТЕГРАЦИИ
ДАННЫХ ИЗ РАЗНЫХ
ИСТОЧНИКОВ

КАЖДЫЙ ЮНИТ
РЕПРЕЗЕНТИРУЕТ
100 РЕАЛЬНЫХ ЛЮДЕЙ
В НАСЕЛЕНИИ

DIFFERENT MEDIA, DIFFERENT SOURCES

BIG DATA PLATFORM

Audience

Задача посчитать аудиторию проектов:

- Avito.ru – аудитория одного проекта
- Avito.ru, lvi.ru – аудиторию 2-х проектов, т.е. количество людей, которые посещают оба сайта
- Total – Total Internet

```
facility = 'mobile' # возможные значения: 'desktop', 'mobile', 'desktop-mobile'
rcats = rc.ResponsumCats(facility)
rtask = rt.ResponsumTask(facility)
# задаем период расчета
date_from = '2020-05-01'
date_to = '2020-05-31'
# задаем Типы пользования Интернетом
usetypes = rcats.get_usetype('all')
# задаем Население, нас интересует города с населением 100тыс. и больше
population = rcats.get_population('Russia100+')
# задаем Возрастные группы, нас интересует 12+
ages = rcats.get_age_groups('12+')
```

Расчет аудитории по проекту Avito.ru

```
# Задаем название проекта для отображения в DataFrame
project_name = 'avito.ru'
# Задаем медиа фильтр, в нашем случае это ID проекта Avito.ru
media_filter = "site = 15828"
# Задаем список статистик для расчета
statistics=["UnwReach", "Reach", "ADR", "AWR", "Uni", "ReachPer", "UnwReachPer", "ReachPer", "ADR", "AWR", "Uni", "ReachPer"]
# Указываем порядок группировки, в нашем случае ее нет
```

```
"slices": [
    "sex",
    "ageGroup",
    "useTypeId",
    "useTypeName",
    "crossMediaThemeId"
],
"filter": {
    "dateFilter": {
        "operand": "AND",
        "elements": [
            {
                "unit": "researchDate",
                "relation": "GTE",
                "value": "2021-04-01"
            },
            {
                "unit": "researchDate",
                "relation": "LTE",
                "value": "2021-04-07"
            }
        ]
    },
    "useTypeFilter": {
        "operand": "OR",
        "elements": [
```

```
(rtask.get_result(task_audience), project_name=project_name)
```

38f22028d7d2> поступила в обработку
=====] время расчета: 0:00:55.706044

reach	unreach	adr_per	awr_per	awr	adr
46592.158751	10385.0	8.157812	21.737146	26600.286646	9982.918132

Таблицы

Итоговые результаты в один DataFrame

```
df = pd.concat([df_avito_ru, df_avito_lvi, df_total])
```

reach	unreach	adr_per	awr_per	awr	adr
39969.333508	9007.0	7.307620	19.022255	23278.007107	8942.516823
46592.158751	10385.0	8.157812	21.737146	26600.286646	9982.918132
95336.209469	19812.0	69.766854	76.068995	93087.523456	85375.436012

```
df.to_excel(filename('audience-avito-ivi'))
df.info()
df.to_excel('report', index=False)
df.to_excel('report', index=False)
```

МАСШТАБИРУЕМОСТЬ

ОБЪЕДИНЕНИЕ ДАННЫХ ИЗ ДВУХ И БОЛЕЕ ИСТОЧНИКОВ

ТРЕБОВАНИЯ К ИСТОЧНИКАМ:



НАЛИЧИЕ ОБЩИХ АТТРИБУТОВ,
ХАРАКТЕРИЗУЮЩИХ МЕДИА-АКТИВНОСТЬ РЕСПОНДЕНТОВ



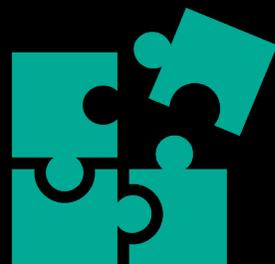
НАЛИЧИЕ CROSS-ПАНЕЛИ, НА КОТОРОЙ МОЖНО ИЗМЕРИТЬ
ВЗАИМОСВЯЗИ В ПОТРЕБЛЕНИИ МЕДИА

DIFFERENT MEDIA, DIFFERENT SOURCES

СИНТЕТИЧЕСКАЯ ПОПУЛЯЦИЯ – НУВ ДЛЯ:



ПОСТРОЕНИЯ СЛОЖНЫХ
ЦЕЛЕВЫХ ГРУПП



ОБОГАЩЕНИЯ СТОРОННИМИ
БОЛЬШИМИ ДАННЫМИ

СПАСИБО!

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 0 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1
1 1 1 1 0 1 1 0 1 1 0 1 1 1 1 0 1 1 0
0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 1 1
0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1
1 0 0 1 0 1 0 0 0 1 0 0 0 1 0 0 1 0 0
0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1
0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 0 0 1 1 0 1 1 1 1 1 1 1 0 1 0 0
0 0 1 1 1 1 1 1 1 1 0 0 0 1 1 0 1 1 1
1 1 1 0 1 1 1 0 1 1 1 1 0 1 1 1 0 1 0
0 0 0 1 1 0 0 0 0 0 1 1 0 1 0 1 0 1 0
1 1 0 1 1 1 1 0 0 0 1 1 1 0 0 1 0 0 0
1 1 0 0 0 0 1 1 1 1 1 0 1 0 1 1 1 1 1
1 1 0 1 0 1 1 0 1 1 0 1 0 0 0 1 1 1 1
0 0 1 0 1 0 1 1 1 0 0 1 1 1 1 1 1 1 1