

Введение в журналистику данных

Алексей Смагин | 2022

Давайте знакомиться!



Алексей Смагин

Дата-журналист. Работал в РБК, «Новой газете», Студии инфографики РИА Новости, Тинькофф Журнале, сейчас — в исследованиях Яндекса.

Веду занятия по работе с данными, автор двух курсов в Высшей Школе Экономики.

Для связи:

Telegram – @BlackPineapple
Facebook – facebook.com/blackpn

журналистика данных

Данные – представление фактов, понятий или инструкций в форме, приемлемой для общения, интерпретации, или обработки человеком или с помощью автоматических средств

Данные – представление фактов
в форме,
приемлемой для
обработки
человеком или с помощью
автоматических средств

В общем, **данные** – это любая
информация, имеющая хоть какой-то
смысл

а не 'цифры' или 'статистика'

Когда в данных можно выделить
объекты, обладающие общими
свойствами, информацию будет нести
не только каждый объект в
отдельности, но и их совокупность

Объекты – новости на сайте



ПОЛИТИКА

Великобритания эвакуирует дипломатов из Киева

Европа

11 февраля 2022, 23:10

0



ЭКОНОМИКА

В ПФР напомнили, как получить выплату по беременности и родам

Москва

11 февраля 2022, 22:33

0

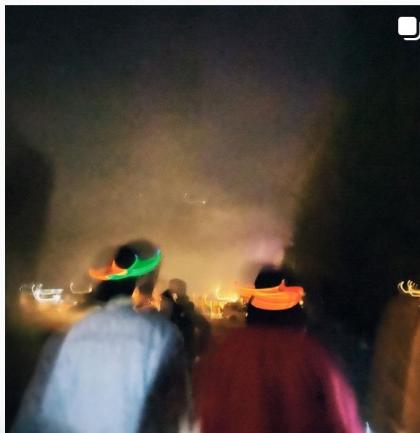
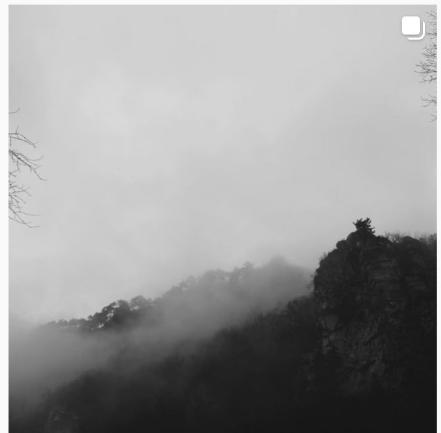
О каких темах агентство пишет чаще? В какой день было больше всего новостей?

Объекты – файлы в папке “Загрузки”

Имя	Размер	Тип	Дата добавления
2. Сведения о доходах.docx	190 КБ	Micros... (.docx)	15 февр. 2020 г., 03:48
2.jpg	3,2 МБ	JPEG	18 марта 2020 г., 23:23
3 (1).jpg	4,9 МБ	JPEG	19 марта 2020 г., 01:32
3-2-3_2019 (1).doc	100 КБ	Micros...t (.doc)	10 июля 2020 г., 12:29
3-2-3_2019.doc	100 КБ	Micros...t (.doc)	24 июня 2020 г., 14:04
3-2-3-2017 (1).doc	87 КБ	Micros...t (.doc)	10 июля 2020 г., 12:28
3-2-3-2017.doc	87 КБ	Micros...t (.doc)	24 июня 2020 г., 14:10
3-4 (1).doc	439 КБ	Micros...t (.doc)	28 апр. 2020 г., 14:47
3-4.doc	439 КБ	Micros...t (.doc)	28 апр. 2020 г., 12:39
3.jpg	4,9 МБ	JPEG	18 марта 2020 г., 23:23
04-04.doc	49 КБ	Micros...t (.doc)	19 мая 2020 г., 17:44
05-05_2017-2018 (1).xls	34 КБ	Micros...ok (.xls)	10 июля 2020 г., 12:29

Файлов какого типа больше всего? Сколько места занимают все файлы в папке?

Объекты – посты в Instagram



Как часто добавляются
фото?
Какой цвет на них
преобладает?

Объекты – предметы на моём столе



Сколько места это всё
займёт, если я решу
переехать?

Мы можем анализировать совокупность объектов с одинаковыми свойствами, чтобы получить новое знание

Это знание может быть полезным, важным, интересным

Журналистика данных — это способ создания материалов в медиа, при котором мы анализируем данные и получаем некоторое **новое знание**, интересное широкой аудитории

большие данные

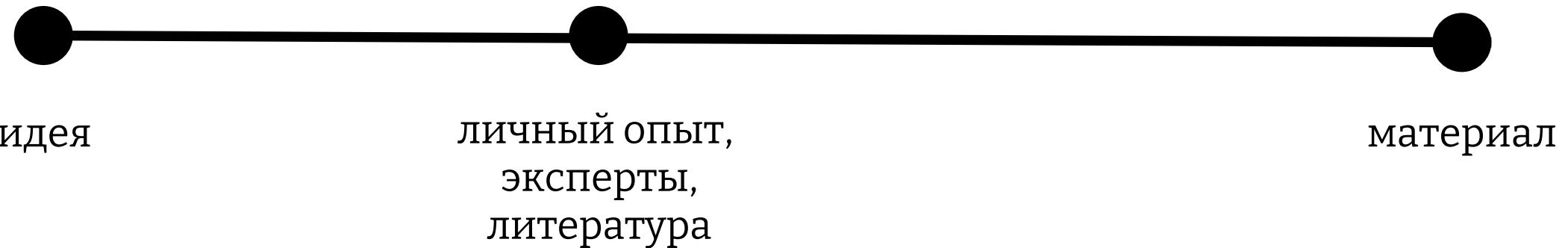


большие данные

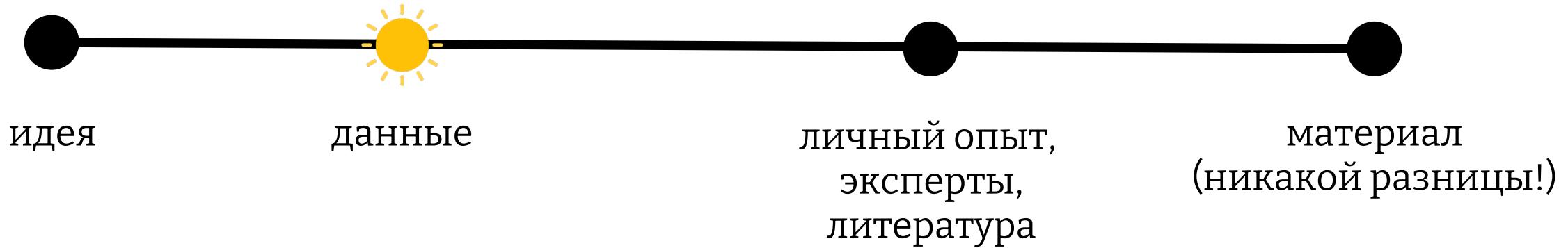
Большие данные – огромные потоки информации, поступающие регулярно (например, информация обо всех абонентах мобильной связи в стране, совершивших звонки за час)

В **журналистике данных** мы работаем чаще всего с сотнями или тысячами объектов. Реже – с миллионами (хотя это всё ещё не Big Data в строгом смысле)

Как работает журналист



Как работает **дата-журналист**



**журналистика
данных
в примерах**

**Методы работы дата-
журналистов существовали
задолго до развития
компьютерной техники**

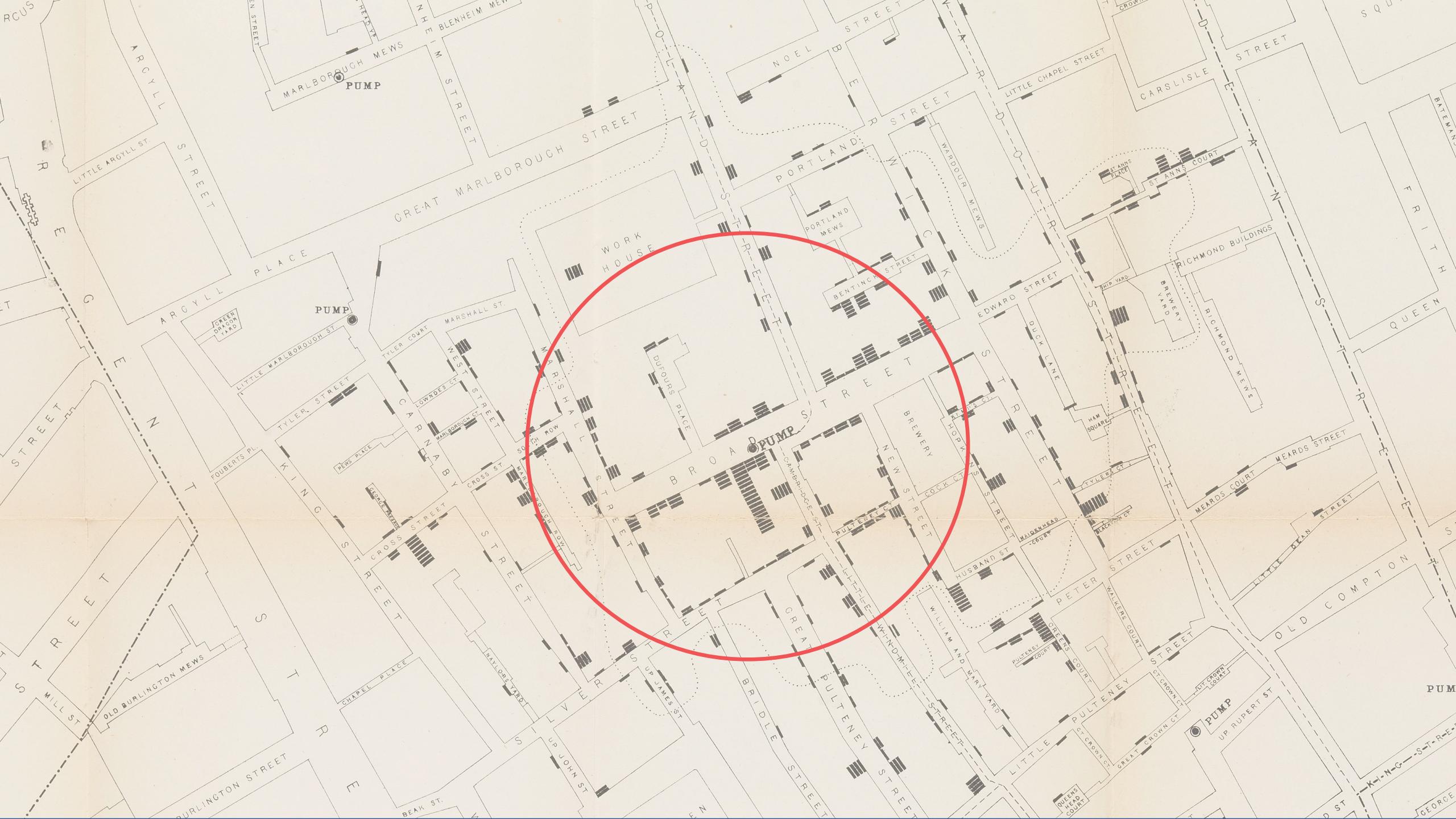
1854 год. в мире — вспышки холеры. умирают
сотни тысяч людей в разных странах

популярна теория «миазмов», которая гласит,
что люди заражаются, вдыхая нездоровый
воздух

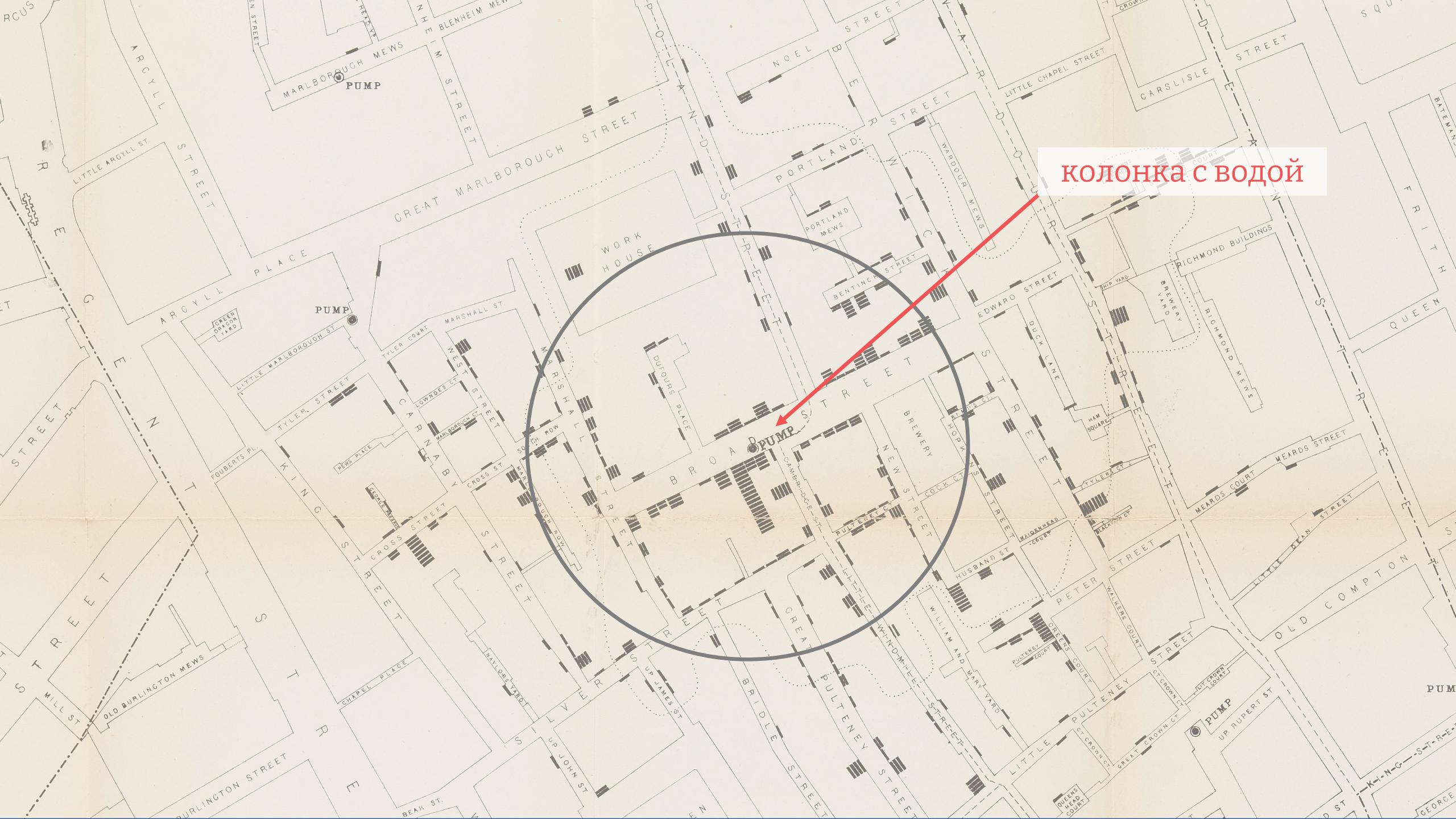


своё исследование проводит доктор Джон Сноу. он опрашивает местных жителей в Лондоне, отмечая на карте, в каких домах живут больные.





КОЛОНКА С ВОДОЙ

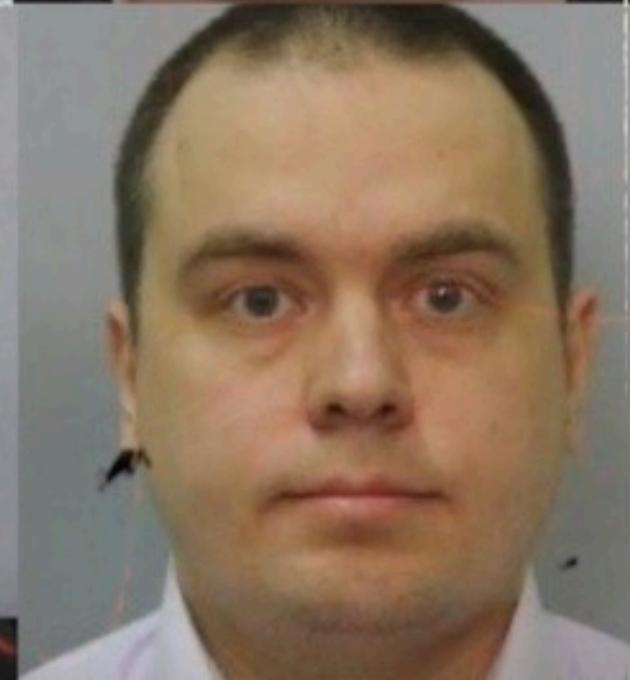
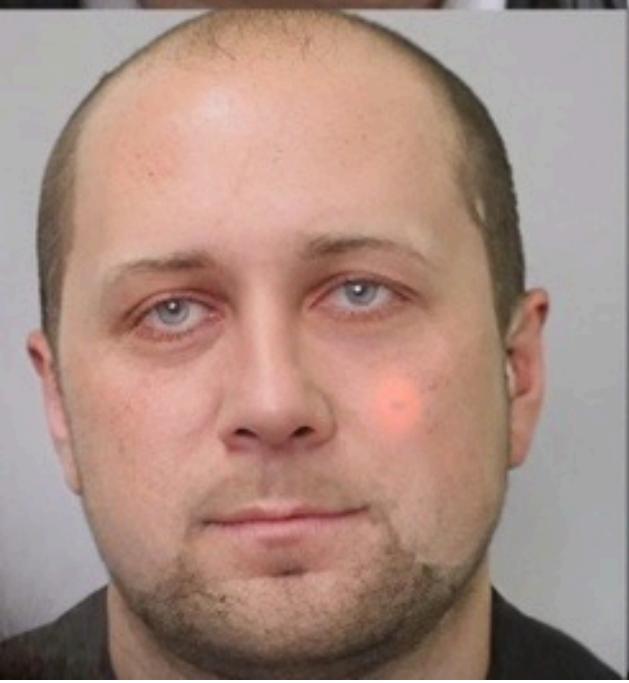
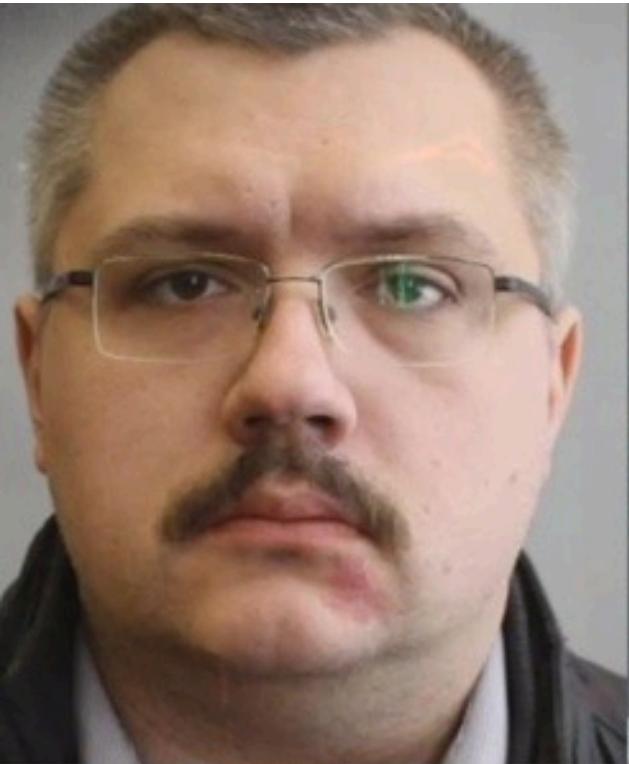




В монастыре рядом с
пивоварней нет больных:
монахи пили только пиво



**Работа с данными помогает
узнавать вещи, которые не
смогли бы найти обычные
журналисты**



Методы журналистики данных в расследовании Bellingcat

В расследовании Bellingcat использовались как традиционные методы работы с данными: ручная проверка сведений о сотрудниках ФСБ, чтение документов «глазами», так и методы **журналистики данных**

<https://ru.bellingcat.com/novosti/russia/2020/12/28/navalny-fsb/>

Методы журналистики данных в расследовании Bellingcat



Биллинги
телефонных разговоров



Список пассажиров
в самолётах

Биллинги телефонных разговоров

Расследователи «пробили» номера руководителей НЦ «Сигнал».

В биллингах содержится информация обо всех входящих и исходящих вызовах по данному номеру с указанием даты и времени.

Биллинги телефонных разговоров

- Но вы же просто видите, что люди созваниваются, — как вы при таких ограниченных данных понимаете, что это был особенно важный рабочий разговор? Они же могли просто, например, задержки зарплат обсуждать?
- Здесь важно искать исключения. **Всплеск звонков**, в котором вдруг появляются **новые номера**, — это уже что-то. Например, в одном таком «пике» мы увидели номер [главы Центра специальной техники ФСБ генерал-майора Владимира] Богданова, их большого шефа, который вообще редко появлялся в звонках, — и мы проверили, кому он позвонил сразу после. И оказалось, что он звонил в «Сигнал». А потом еще двое участников этих переговоров звонят в «Сигнал». И мы понимаем, что этот всплеск должен быть как-то связан с «Новичком».

<https://meduza.io/feature/2020/12/16/ne-tak-to-prosto-otravit-cheloveka-novichkom>

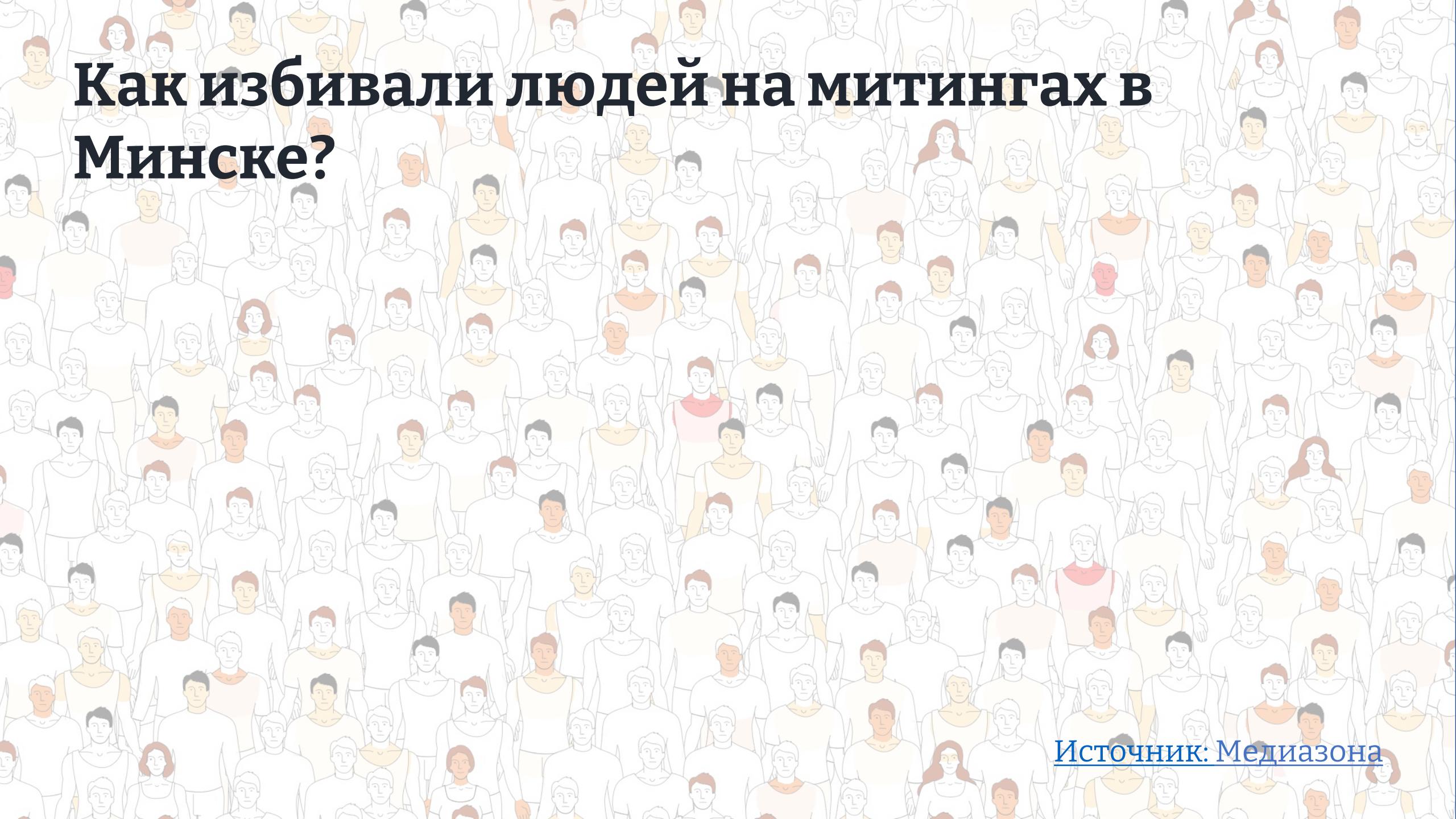
Списки пассажиров в самолётах

Гипотеза: сотрудники ФСБ летают вслед за Навальным в тот же день или с разницей в один день.

Расследователи покупают списки пассажиров на все рейсы в даты полетов Навального с разницей \pm день

Они проверяют совпадения в две даты: дату прилёта и дату отлёта. Так они выходят на группу из трёх сотрудников ФСБ, которые летали за Навальным в его роковую поездку.

**Данные помогают понять
масштаб проблем**

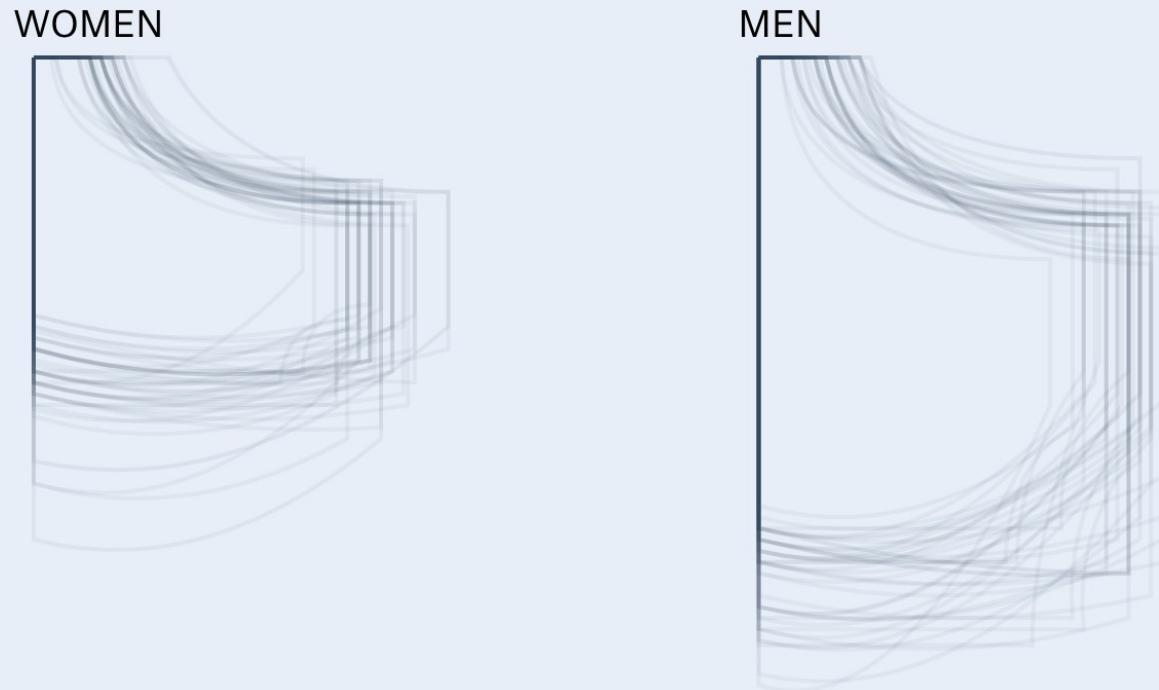


Как избивали людей на митингах в Минске?

[Источник: Медиазона](#)

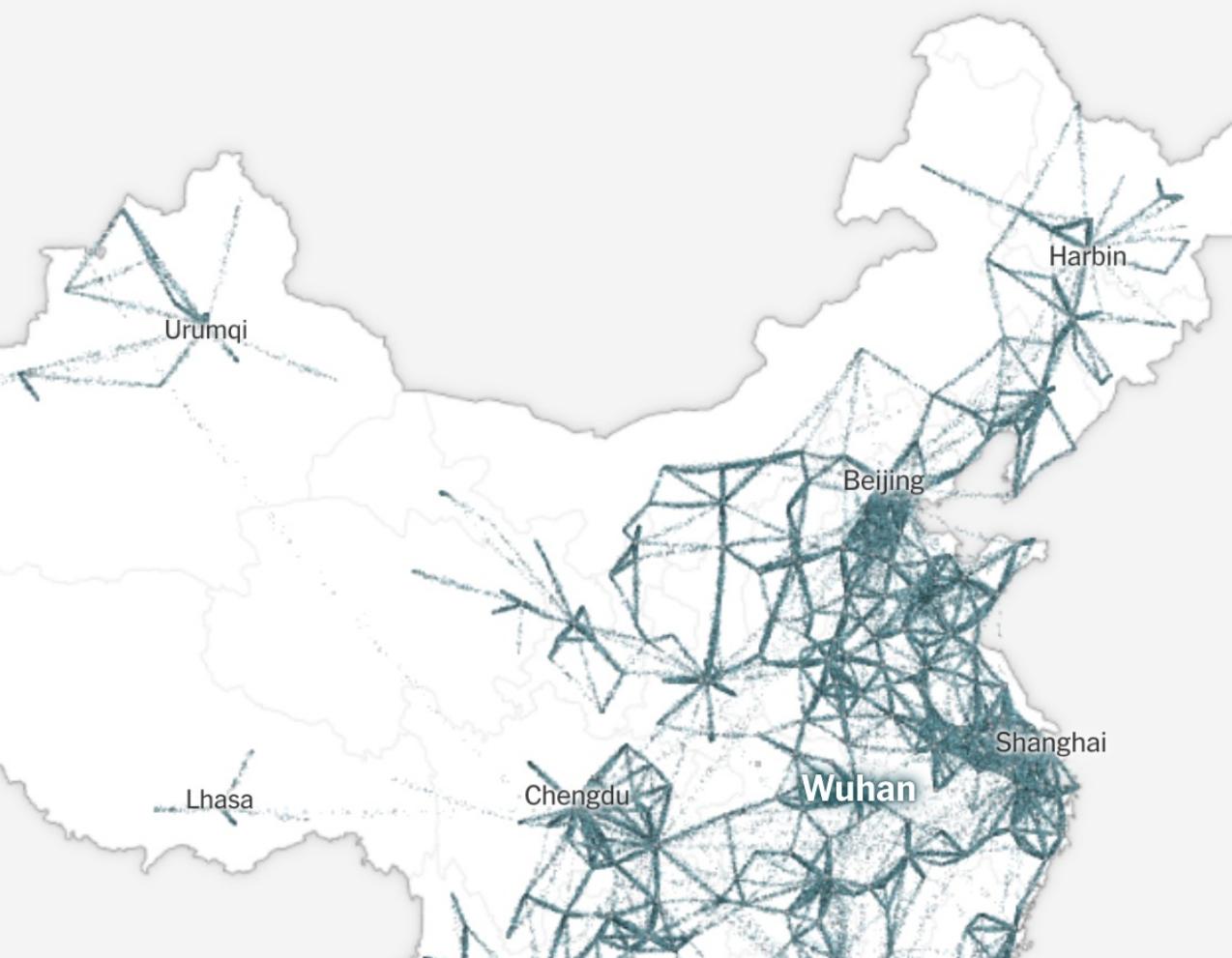
... или наглядно их отразить

Карманы в мужских и женских джинсах



[Источник: The Pudding](#)

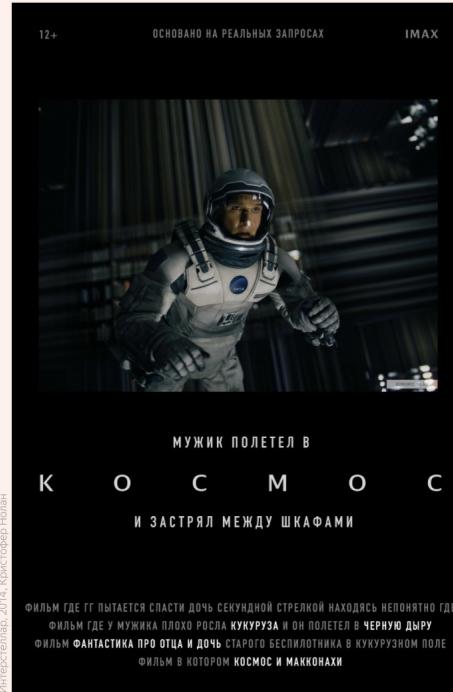
Как коронавирус распространился по всему миру



[New York Times](#)

Данные – это не скучно

Как найти фильм, если не помнишь название?



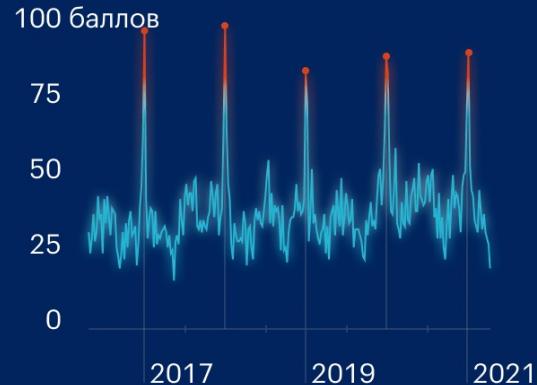
[Исследования Яндекса](#)

И не обязательно – сложно

Как меняется интерес россиян к вопросам здоровья

🔍 Уснуть

Чаще всего россияне беспокоятся о сне
в неделю после новогодних праздников



🔍 Давление

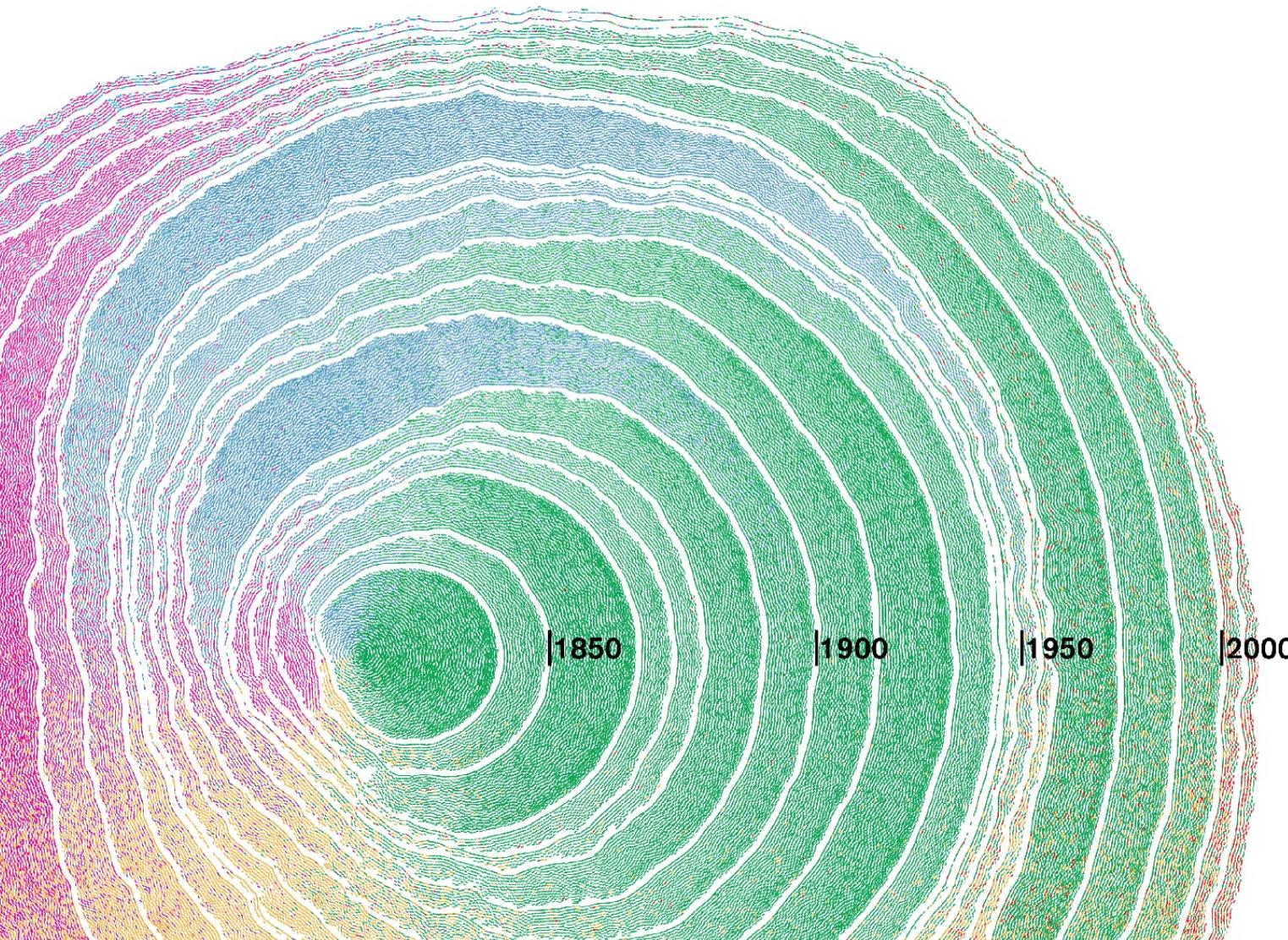
Запросы про давление популярны в январе
и наименее популярны в начале сентября



Источник: Т–Ж

**В конце-концов,
это прекрасно**

История американской иммиграции

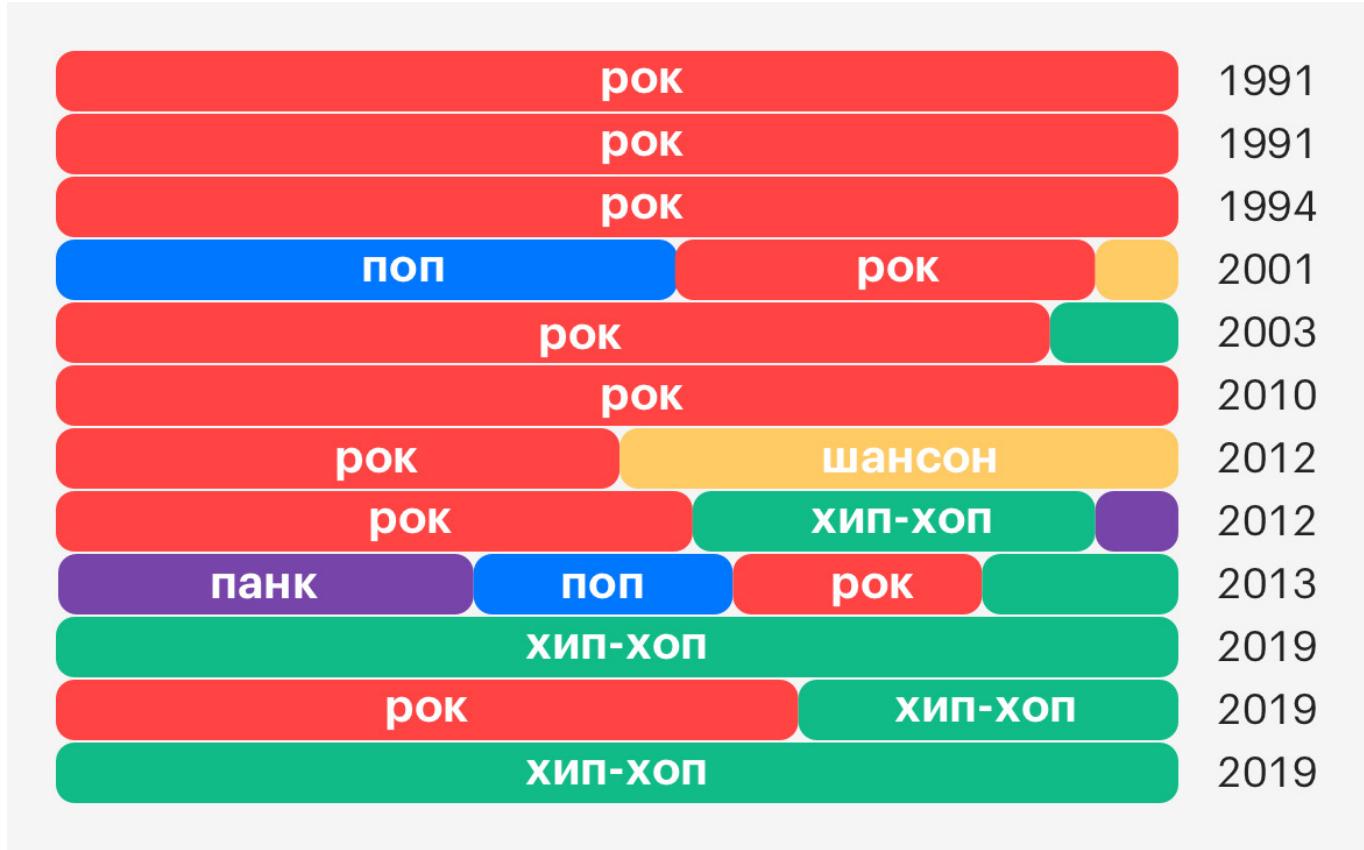


[Источник: Northeastern University](#)

Как сделаны эти материалы?

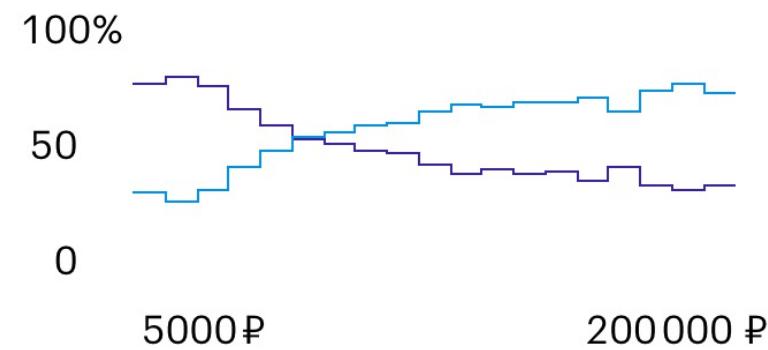
Попробуйте ответить на вопросы:

- О чём рассказывает этот материал?
- Какие данные использовались?
- Получены ли эти данные из открытых источников?
- Нравится ли вам этот материал? Если да, то чем он вас зацепил?

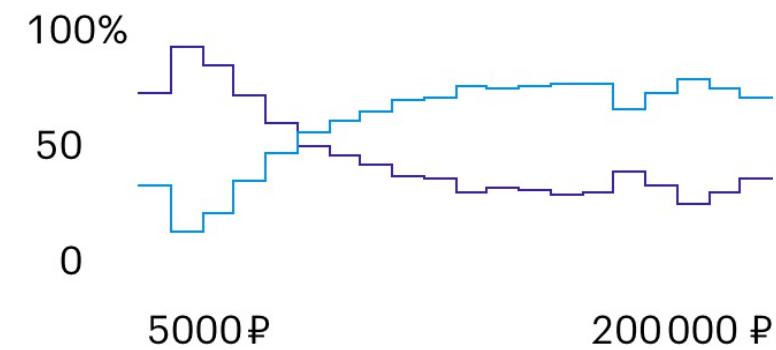


Источник: РБК

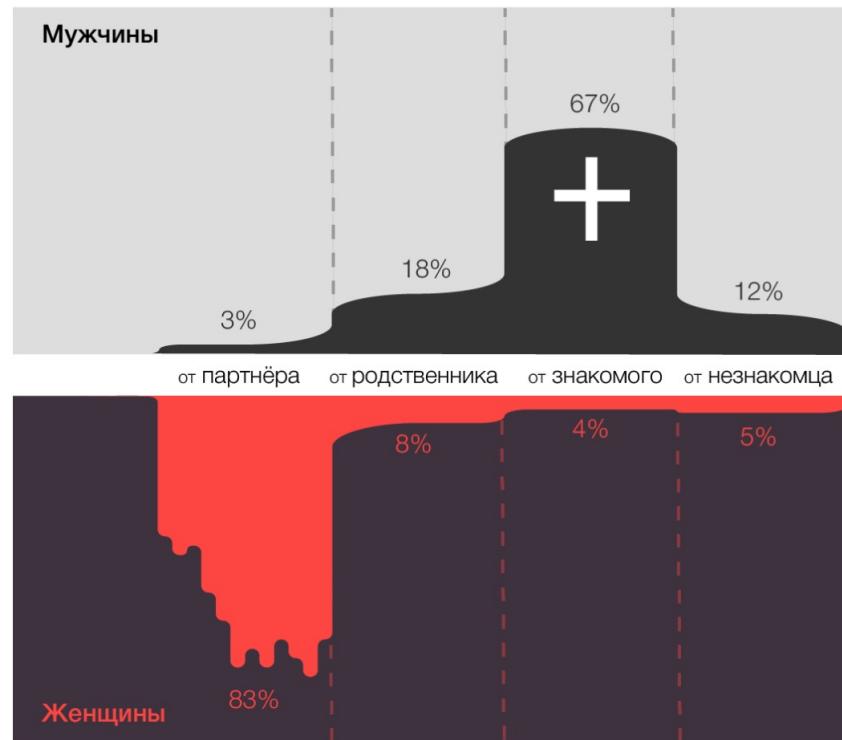
Маркетинг, пиар

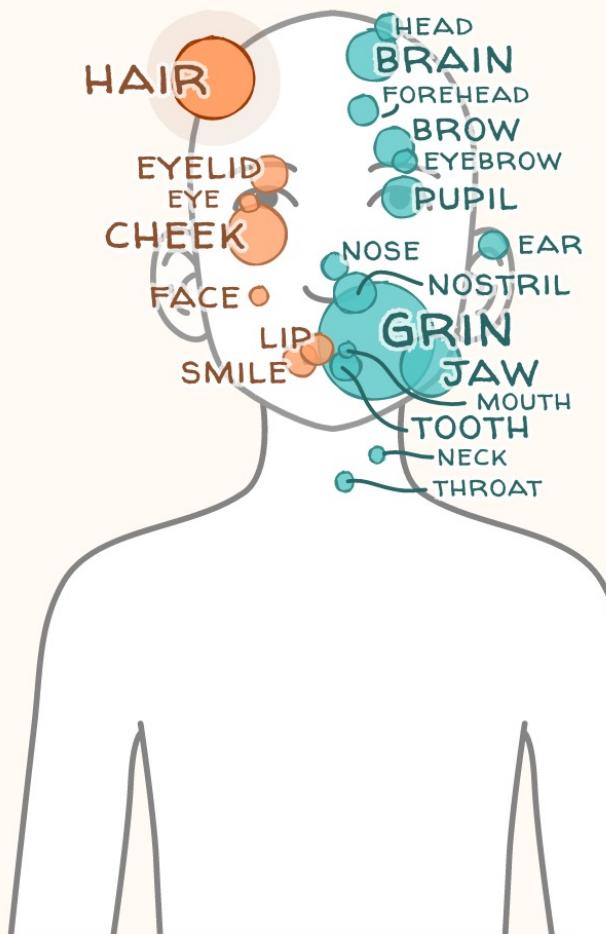


Продажи



[Тинькофф Журнал](#)





The Pudding

Процесс

Формулирование вопросов

Формулирование вопросов

Поиск данных

Формулирование вопросов
Поиск данных
Получение данных

Формулирование вопросов
Поиск данных
Получение данных
Подготовка данных

Формулирование вопросов
Поиск данных
Получение данных
Подготовка данных
Анализ данных

Формулирование вопросов
Поиск данных
Получение данных
Подготовка данных
Анализ данных
Интерпретация данных

Формулирование вопросов
Поиск данных
Получение данных
Подготовка данных
Анализ данных
Интерпретация данных
Упаковка

Какие навыки нужны?

Журналиста

придумывать истории,
писать текст, искать
информацию

Аналитика

работать с данными

Дизайнера

делать инфографику и
макеты сложных
проектов

Программиста

для сложной работы с
данными и
интерактивных проектов

Глава 1. Собянин



АКТИВНЫЙ
ГРАЖДАНИН



23 мая 2018



Виктория 8:57

Лёш, увидела тут плакатик, что платформе «активный гражданин» 26 мая 4 года

Мб вам актуально будет сделать материал в духе «какие решения на самом деле приняли граждане за это время»
Насчет этой платформы оч много вопросов: относительно важных проблем мнения на самом деле не учитываются, а люди могут повлиять на что то совсем мизерное



Результаты голосований в проекте «Активный гражданин»

Таблица

Паспорт

Скачать



Не нашли объект?



Фильтры

#	▲ Название голосования	⚙	Ссылка на результаты голосования
1	Результаты голосования: Благоустройство Проспекта Мира и Ярославского шоссе: оценка москвичей		ag.mos.ru/check?poll_i...
2	Результаты голосования: Благоустройство Ленинградки: оценка москвичей		ag.mos.ru/check?poll_i...
3	Результаты голосования: Экскурсии в природных парках		ag.mos.ru/check?poll_i...
4	Результаты голосования: Благоустройство улиц у Новодевичьего монастыря: оценка москвичей		ag.mos.ru/check?poll_i...
5	Результаты голосования: Благоустройство Неглинной улицы: оценка москвичей		ag.mos.ru/check?poll_i...

 скачать все результаты

№	Уникальный идентификатор пользователя	Вопрос	Ответ
1	9d853ba01d78e89f55ecb993b0b36bf1 Голос принят в: 08.01.2016 12:37	<p>Развитию спорта в Москве уделяется особое внимание. Каждый год в городе открываются новые спорткомплексы, стадионы, развивается спортивная инфраструктура. Так, в 2015 для москвичей распахнули свои двери ледовый дворец «Арена легенд» и летний аквакомплекс «Лужники». Начал работу уникальный спортивный центр по подготовке сборных команд. А московские спортсмены стали призерами многих международных соревнований. Как вы считаете, какое событие уходящего года в сфере спорта можно назвать «Главным делом Москвы»?</p> <p>Не все же время работать, надо и отдыхать. Москва предлагает массу вариантов проведения досуга. Концерты, праздники, фестивали давно вошли в привычную жизнь горожан. Например, в этом году миллионы москвичей посетили пятый юбилейный фестиваль «Круг света», покатались на коньках на самом большом катке Европы на ВДНХ, послушали Aerosmith на Дне города и многое другое. Как вы считаете, какое событие уходящего года в сфере культуры и развлечений можно назвать «Главным делом Москвы»?</p>	<p>Затрудняюсь ответить</p> <p></p> <p>Ничего из перечисленного</p>

Какие места оценивали в «Активном гражданине»

Всюду голосуют одни и те же люди



■ % голосовавших в обеих точках от числа участников в меньшем опросе

Источник: ag.mos.ru

Новая газета, 2018

Вот особенно прекрасное:

Работы выполнены хорошо, но есть на что обратить внимание (укажите): Я думаю, что вы пидаrasы

Работы выполнены хорошо, но есть на что обратить внимание (укажите): Я не знаю что, я голосую ради баллов

21:02

[https://novayagazeta.ru/articles/2018/09/21/77914-
prodam-svoy-golos-za-svitshot](https://novayagazeta.ru/articles/2018/09/21/77914-prodam-svoy-golos-za-svitshot)

Глава 2. Оскар

номинанты

2019







Алина 22:29

ну, что точно можно сказать: сейчас есть тенденция давать Оскар фильмам, затрагивающих меньшинства в любом проявлении (женщины, темнокожие, люди с нетрадиционной ориентацией)



Алексей 22:29

я вот чего боюсь – можно ли сказать что-то новое на эту тему?



Алина

16 сен 2019

ну, что точно можно сказать: сейчас есть тенденция давать Оскар фильмам, затрагивающих меньшинства в любом проявлении (женщины, темнокожие, люди с нетрадиционной ориентацией)



Алина 22:29

про Oscar bait знаешь?



Алексей 22:30

Нет

Россман и Шилк разработали алгоритм, рассчитывающий соотношение вложенных инвестиций/платы киноакадемикам с шансами получить ту или иную награду. При этом количество полученных премий прямо пропорционально размерам вероятных кассовых сборов. Это сопряжено с большими рисками, если картина не получит никаких премий или даже не пройдёт номинацию, она обречена на провал. Поэтому подобную маркетинговую компанию можно сравнивать с лотереей^[21].

**«Узнайте, насколько
ваш любимый фильм
снят под Оскар»**

Close, But No Cigar: The Bimodal Rewards to Prize-Seeking

American Sociological Review
2014, Vol. 79(1) 86–108
© American Sociological
Association 2014
DOI: 10.1177/0003122413516342
<http://asr.sagepub.com>



Gabriel Rossman^a and Oliver Schilke^a

Abstract

This article examines the economic effects of prizes with implications for the diversity of market positions, especially in cultural fields. Many prizes have three notable features that together yield an emergent reward structure: (1) consumers treat prizes as judgment devices when making purchase decisions, (2) prizes introduce sharp discontinuities between winners and also-rans, and (3) appealing to prize juries requires costly sacrifices of mass audience appeal. When all three conditions obtain, winning a prize is valuable, but seeking it is costly, so trying and failing yields the worst outcome—a logic we characterize as a Tullock lottery. We test the model with analyses of Oscar nominations and Hollywood films from 1985 through 2009. We create an innovative measure of prize-seeking, or “Oscar appeal,” on the basis of similarity to recent nominees in terms of such things as genre, plot keywords, and release date. We then show that Oscar appeal has no effect on profitability. However, this zero-order relationship conceals that returns to strong Oscar appeals are bimodal, with super-normal returns for nominees and large losses for snubs. We then argue that the effect of judgment devices on fields depends on how they structure and refract information.

for films associated with a particular theme and divide this sum by the frequency of the theme so as to give an indication of the Oscar-ness for the theme's average film. The result is λ_{it} , which can be understood conceptually as how tightly associated with the Oscars a particular theme is at a particular time.

$$\lambda_{it} := \frac{\sum_{j=1}^J \zeta_{ij} \frac{\ln(\kappa_j + \sqrt{\kappa_j^2 + 1})}{\sqrt{n_{j(i)}}}}{n_{i(j)}} \quad (1)$$

Although there is some change from year to year, on a fairly consistent basis the highest λ_{it} genres are drama, war, history, and biography, whereas the lowest are usually horror, science fiction, action, and family. The λ_{it} of keywords oscillates more than that of genre; for convenience we draw examples with reference to the year 2009, when among the high

$$\tau_j := \sum_{i=1} \lambda_{it} \zeta_{ij} \quad (2)$$

Estimation of Oscar Appeal

Because films are not explicitly rated by their Oscar appeal, we must measure Oscar appeal operationally as how closely a film conforms to the model of recent films that garnered Oscar nominations. Specifically, in Table 1 we model the number of major category Oscar nominations each film garnered using negative binomial regression as a function of observable traits. We use negative binomial regression because the dependent variable (number of above-the-line nominations per film) is almost perfectly described by a negative binomial distribution (with a mean of .3, over-dispersion of 13.2, and no zero-inflation).¹⁰ To allow for nonlinear effects, we break our most important continuous vari-

```
799  
800 *-----  
801 *lambda & tau  
802 *-----  
803 use $bigstuff/imdb19752010_eligible_oscars, clear  
804 merge 1:m imdbtitle using $bigstuff/keywords.dta  
805 keep if _merge==3  
806 drop _merge  
807 save $bigstuff/keywords_oscars, replace  
808  
809 sort imdbtitle  
810 by imdbtitle: gen nkeywords=[_N]  
811 gen pl_totwins_scaled=pl_totwins/(nkeywords^0.5)  
812 gen pl_totnoms_scaled=pl_totnoms/(nkeywords^0.5)  
813 gen weightedfreq=1/(nkeywords^0.5)  
814 gen keywordfreq=1  
815 collapse (sum) pl_totwins_scaled pl_totnoms_scaled weight  
816 lab var weightedfreq "keyword, down-weighted by how many times it appears in a movie"  
817 * notation for keywordfreq: n_i(j)  
818 gen keywordratio=pl_totnoms_scaled/keywordfreq  
819 gen keywordratio_w=pl_totwins_scaled/keywordfreq  
820 save keyword_weightedfreq.dta, replace  
821 sort keywordratio  
822 tail  
823  
824 drop pl_totwins_scaled pl_totnoms_scaled  
825 save keyword_scores, replace /*illustrative only, time  
826
```



Алексей Смагин <asblackpn@gmail.com>

to rossman ▾

Sep 25, 2019, 2:53 PM



Hello!

I'm a data-journalist from Russia.

We are going to make data-based story about Oscar soon and I think that your work about Oscar Appeal is very interesting. Thank you for this work!

We want to explain the reasons of "Oscar Bait" and use your formula to predict nominees for the next Oscar.

The main problem is that in your work you mentioned algorithm to get the Oscar Appeal variable, but don't describe that algorithm. So I have two questions:

- 1) Could you please share final formula of Oscar Appeal variable?
- 2) Could you explain how Actors, Writers and Directors are used in this formula?

I'll be grateful you for any help.

With appreciation,

Alexey



Алексей Смагин <asblackpn@gmail.com>

to Gabriel ▾

Oct 10, 2019, 6:45 PM



Hello again!

I am trying to reproduce your Oscar appeal data, but my NB regression gives very bad results.

	coef
<hr/>	
genres_tau	-3.2391
keywords_tau	0.4298
actors_nomineed	0.4097
directors_nomineed	0.8146
writers_nomineed	0.3810
major	-0.0290
indimajor	0.7212
days_from_year_start	-0.0087

Have you got dataset with computed tau, prior nomination contributors and other regressors? If so, **can you please send your data to me?** I want to check all the variables and find where I was wrong.

If you are interested in my data, csv in the attachment. I have no MPAA data yet, because it is downloading right now, but I don't think that it is the root of the problem.

With gratitude,
Alexey



Gabriel Rossman <rossman@soc.ucla.edu>

to me ▾

Oct 15, 2019, 1:26 AM



Yeah, I doubt MPAA matters much as that had only a small effect in my work.

Note that several of the variables had linear splines in my version whereas I don't see splines in your output.

I attached the file necessary to generate table 1 in both Stata and tab-separated values. If you are using R you should be able to read either one, though you'll need library(foreign) to open the Stata file.

Note that the numbers are slightly off from those in the paper, not sure why, but only slightly. Here is the output from when I checked it.

```
. nbreg totnoms gr_lowmed gr_high kr_lowmed kr_high mpaa_r actornomsum_dummy write  
> rnomsum_dummy directornomsum_dummy major indymajor dy_*
```

Fitting Poisson model:

Iteration 0: log likelihood = -2563.6017
Iteration 1: log likelihood = -1902.1679
Iteration 2: log likelihood = -1549.5828
Iteration 3: log likelihood = -1538.4922
Iteration 4: log likelihood = -1538.3797
Iteration 5: log likelihood = -1538.3797

Гарриет



2019 г. · Драма · 2 ч 5 мин



Посмотреть трейлер на YouTube

5,5/10

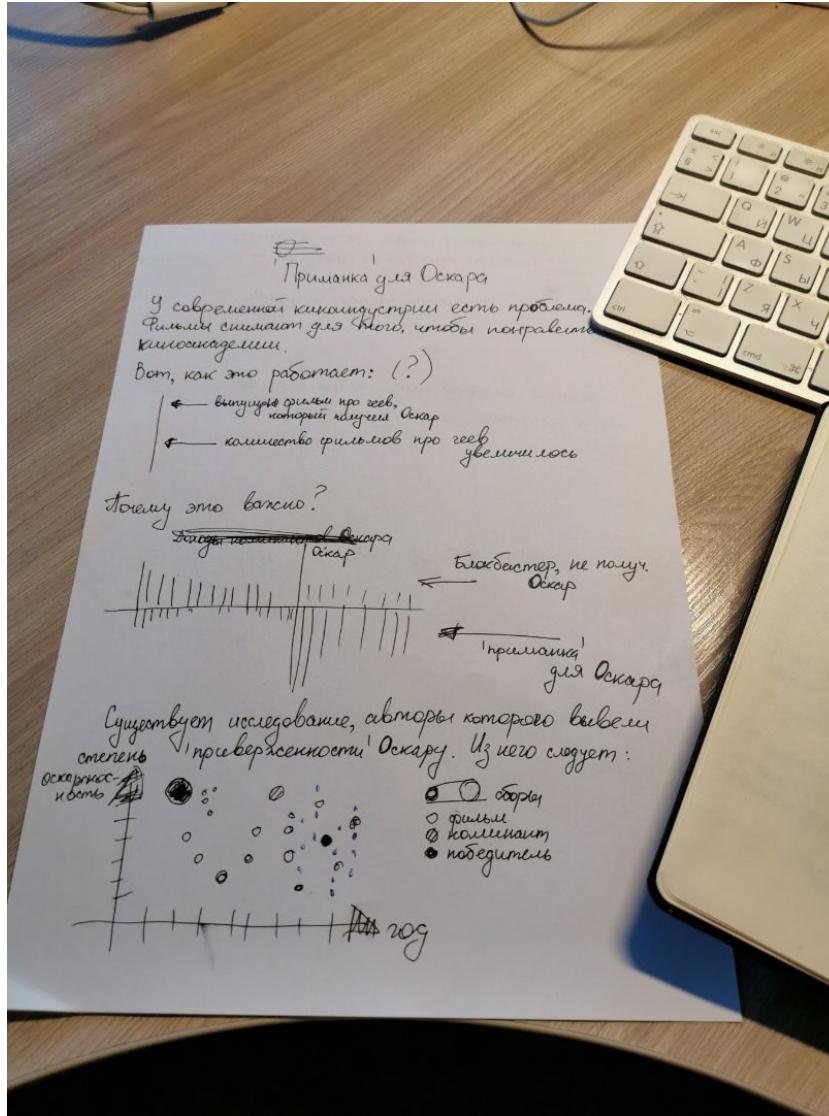
IMDb

63 %

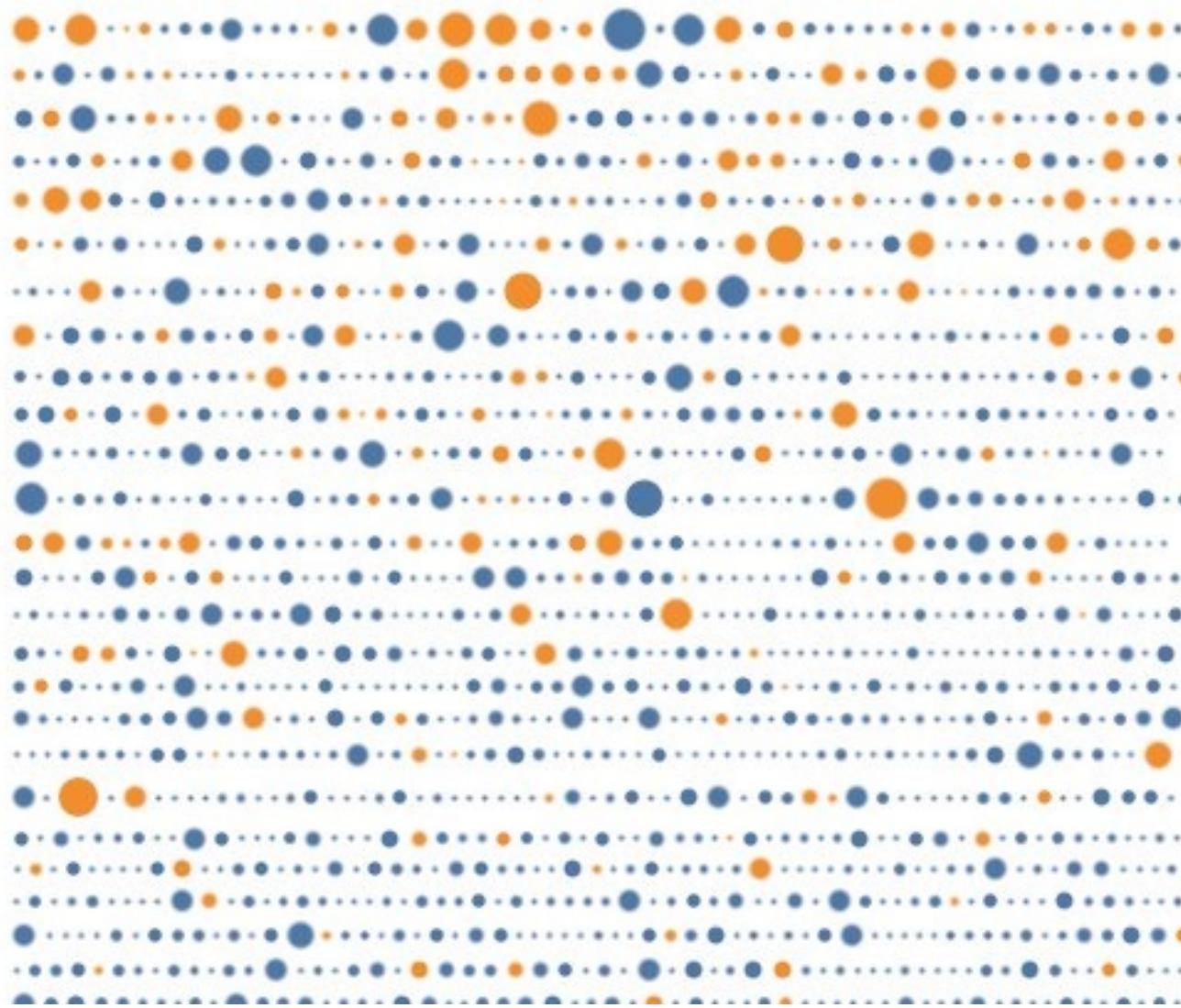
Metacritic

История Гарриет Табмен, бывшей рабыни и аболиционистки, которая посвятила свою жизнь борьбе против рабства. Гарриет регулярно совершала поездки на юг, откуда вывозила рабов, и приняла активное участие в деятельности подземной железной дороги, переправлявшей беглых рабов из южных штатов на Север или в Канаду.

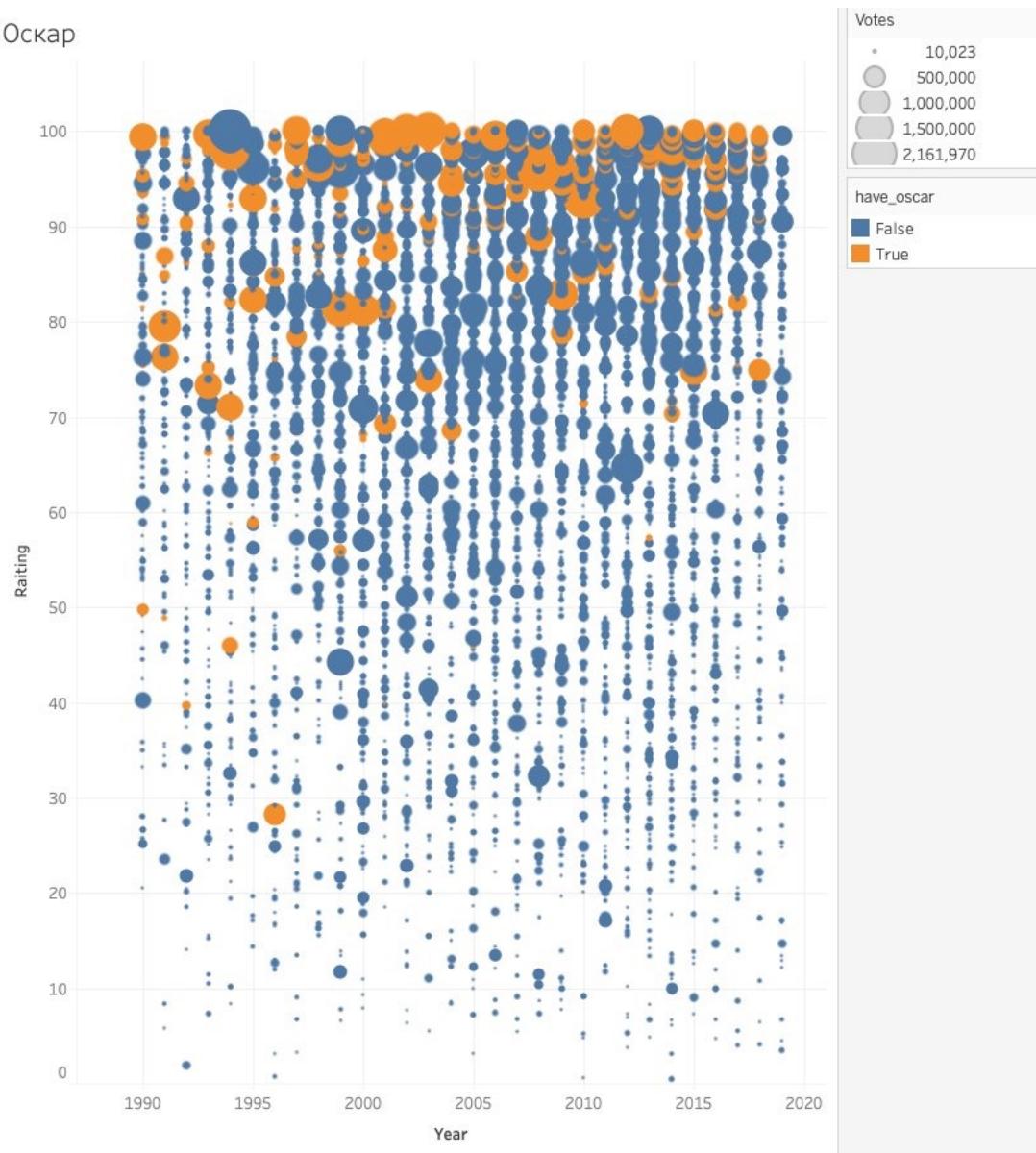
Дата премьеры: 23 января 2020 г. (Россия)

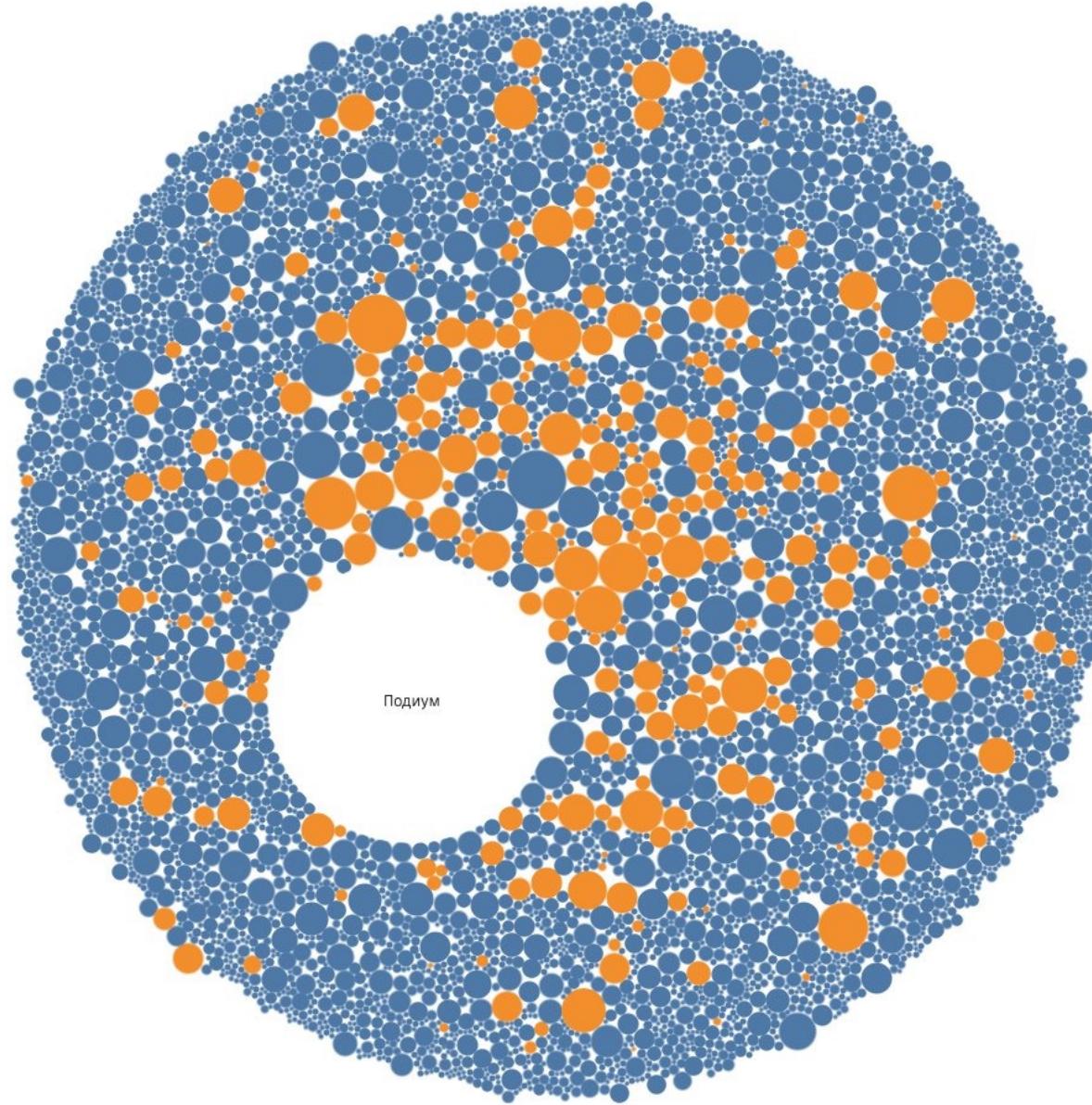


Оскар

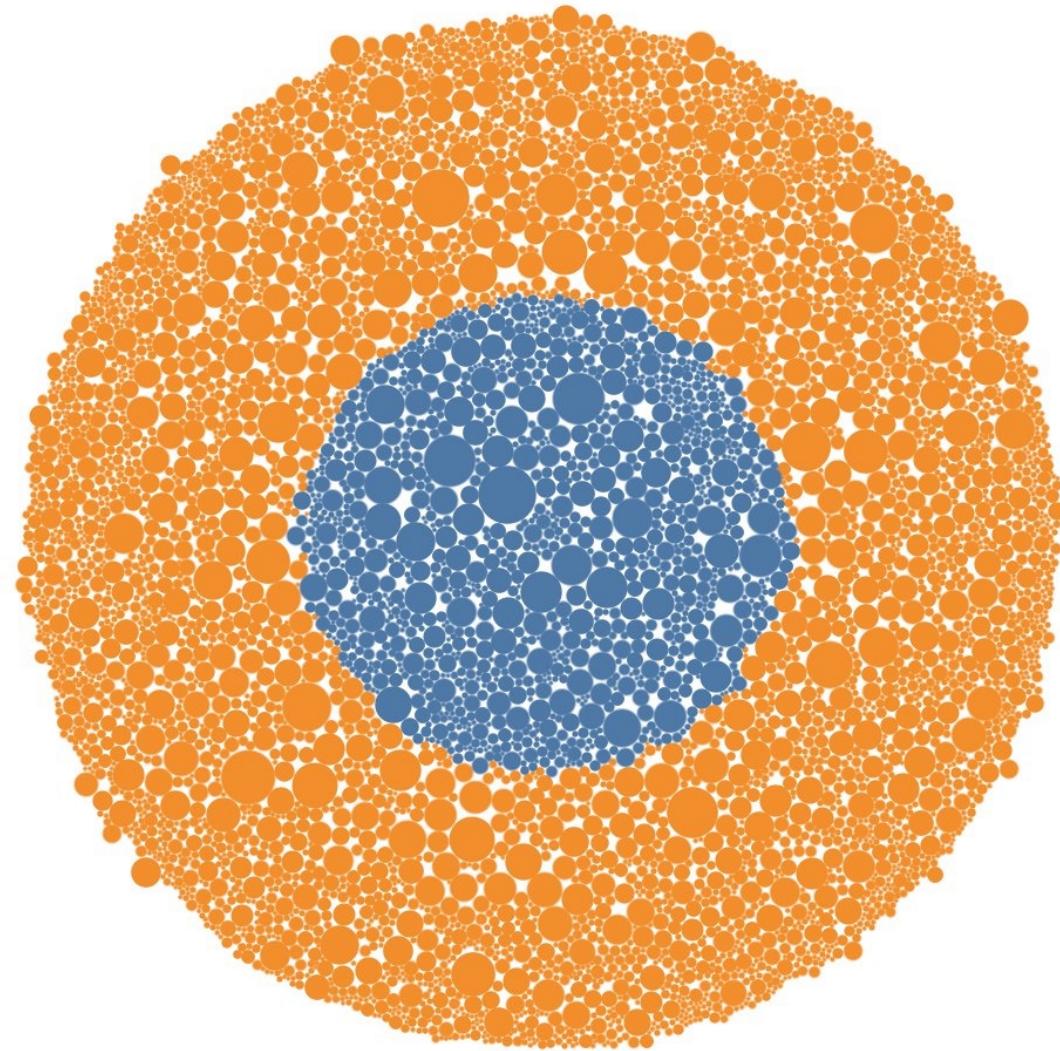


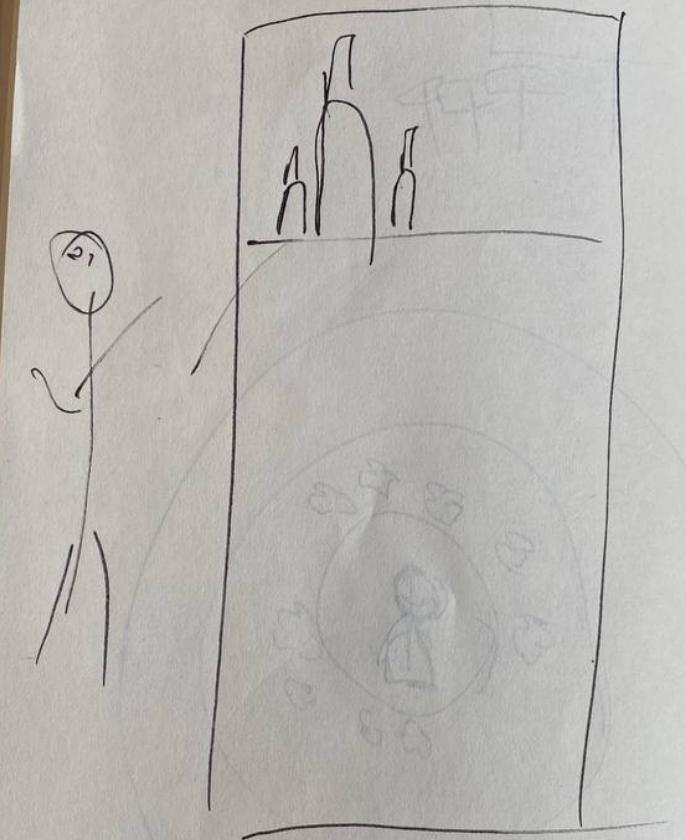
Оскар

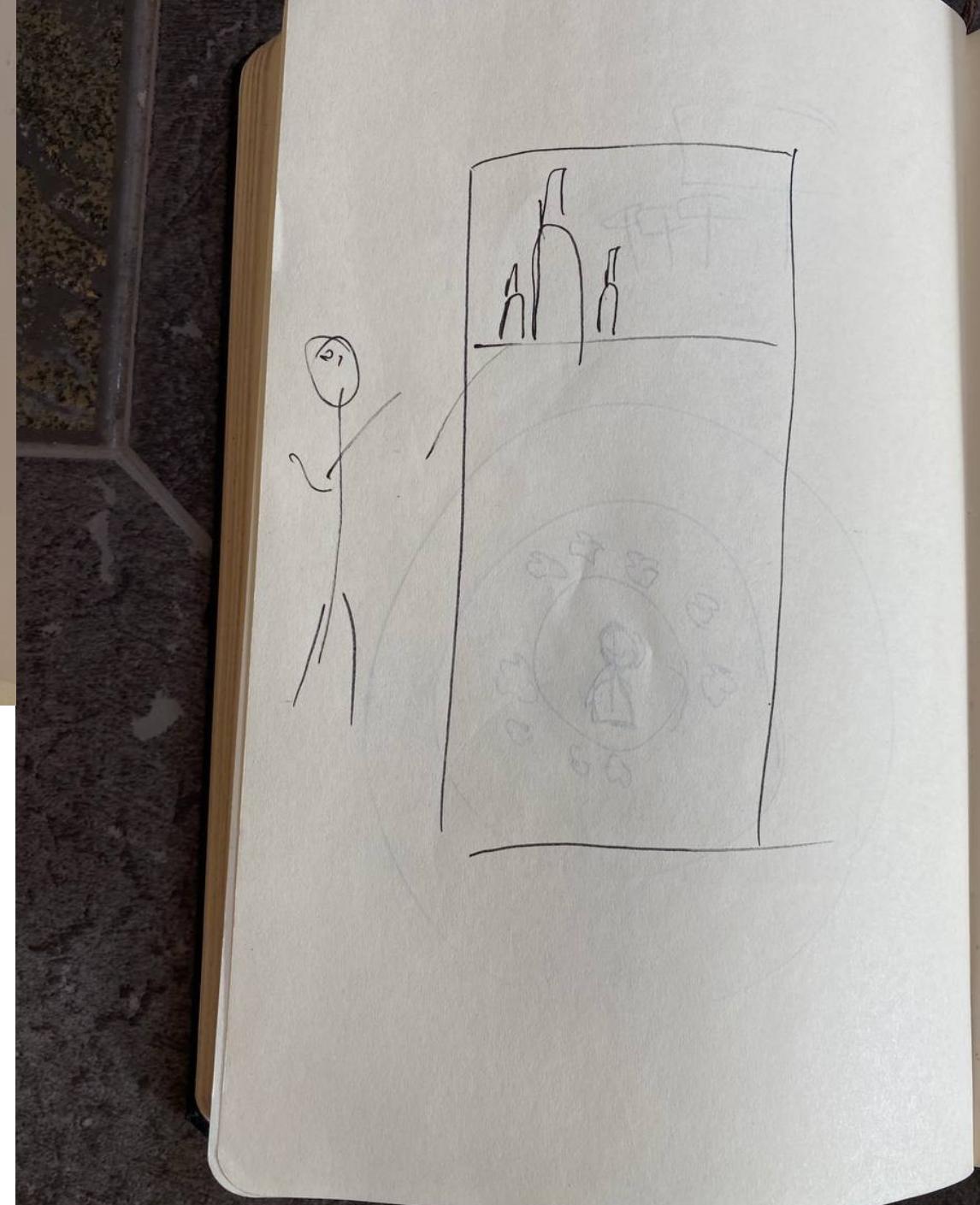
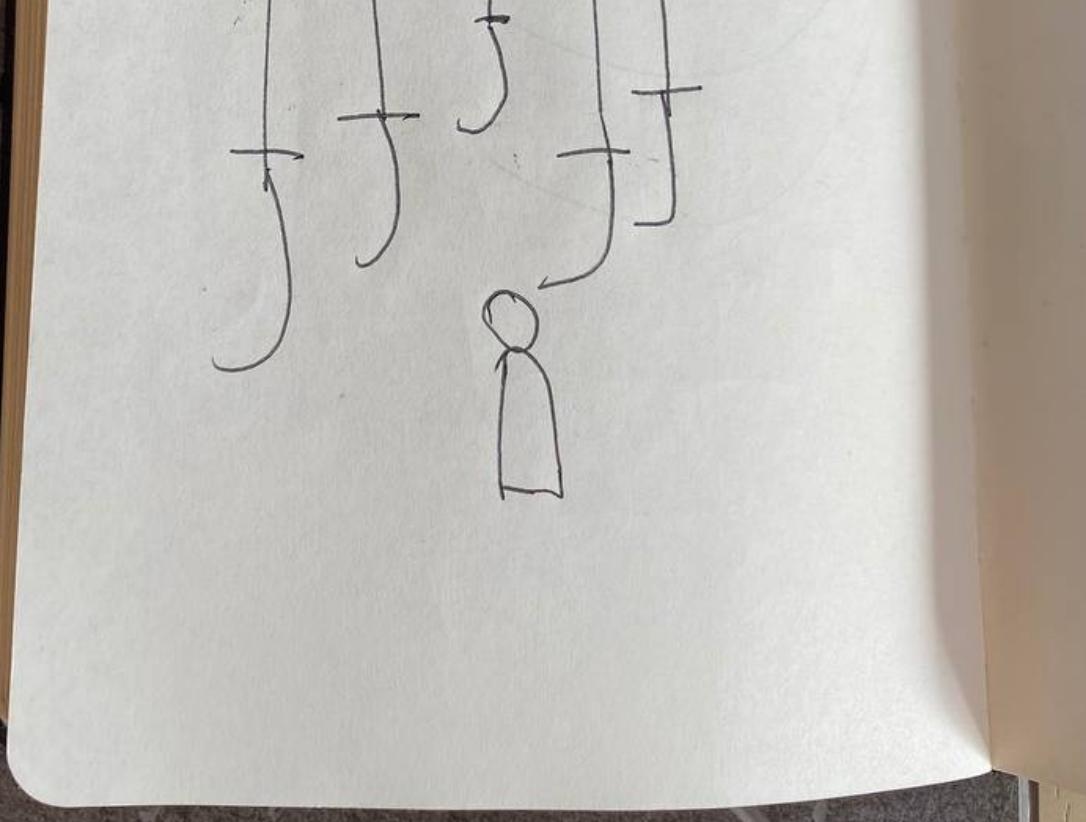


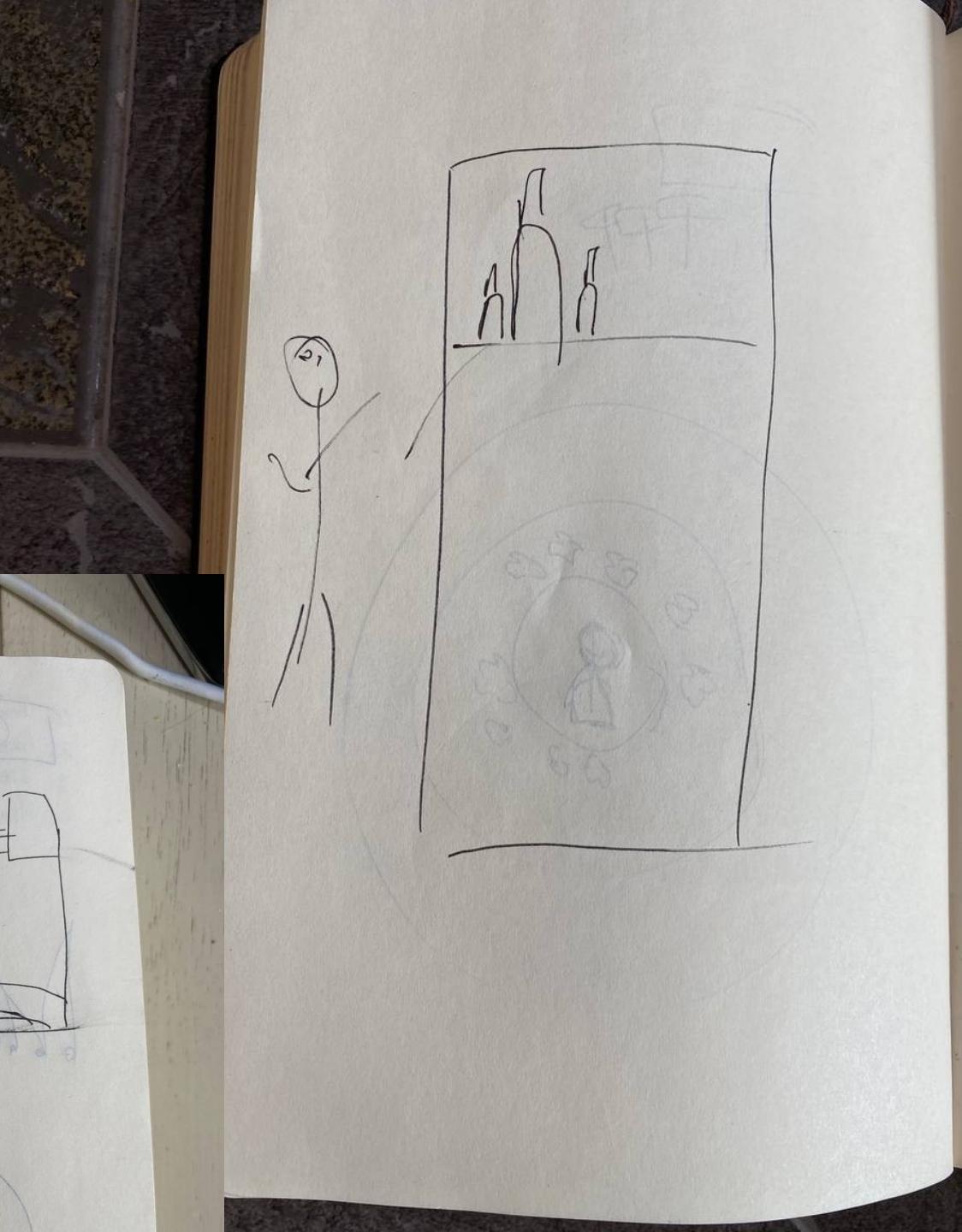
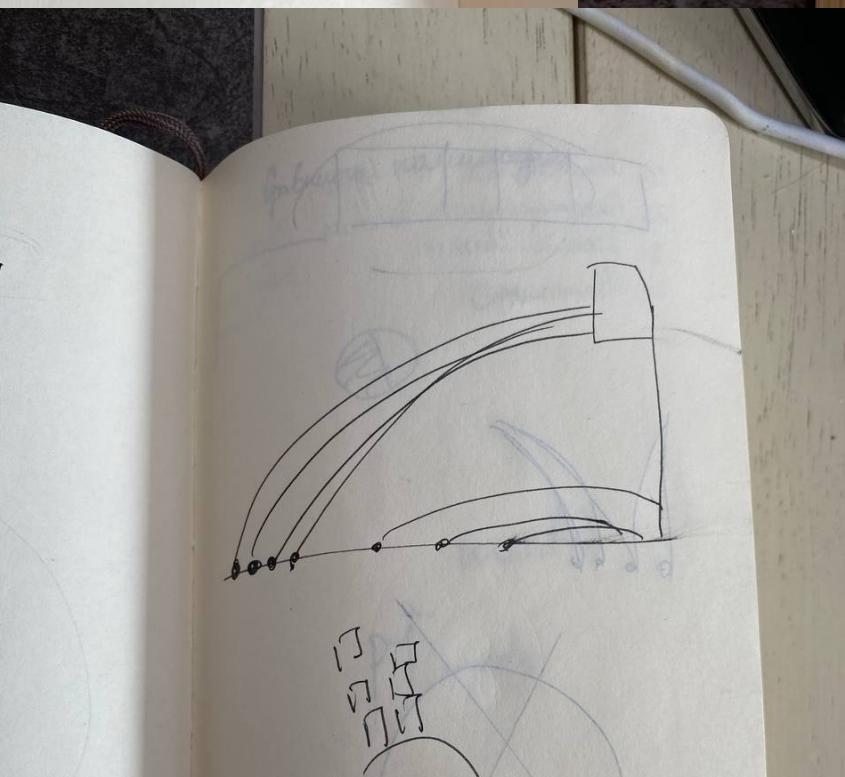
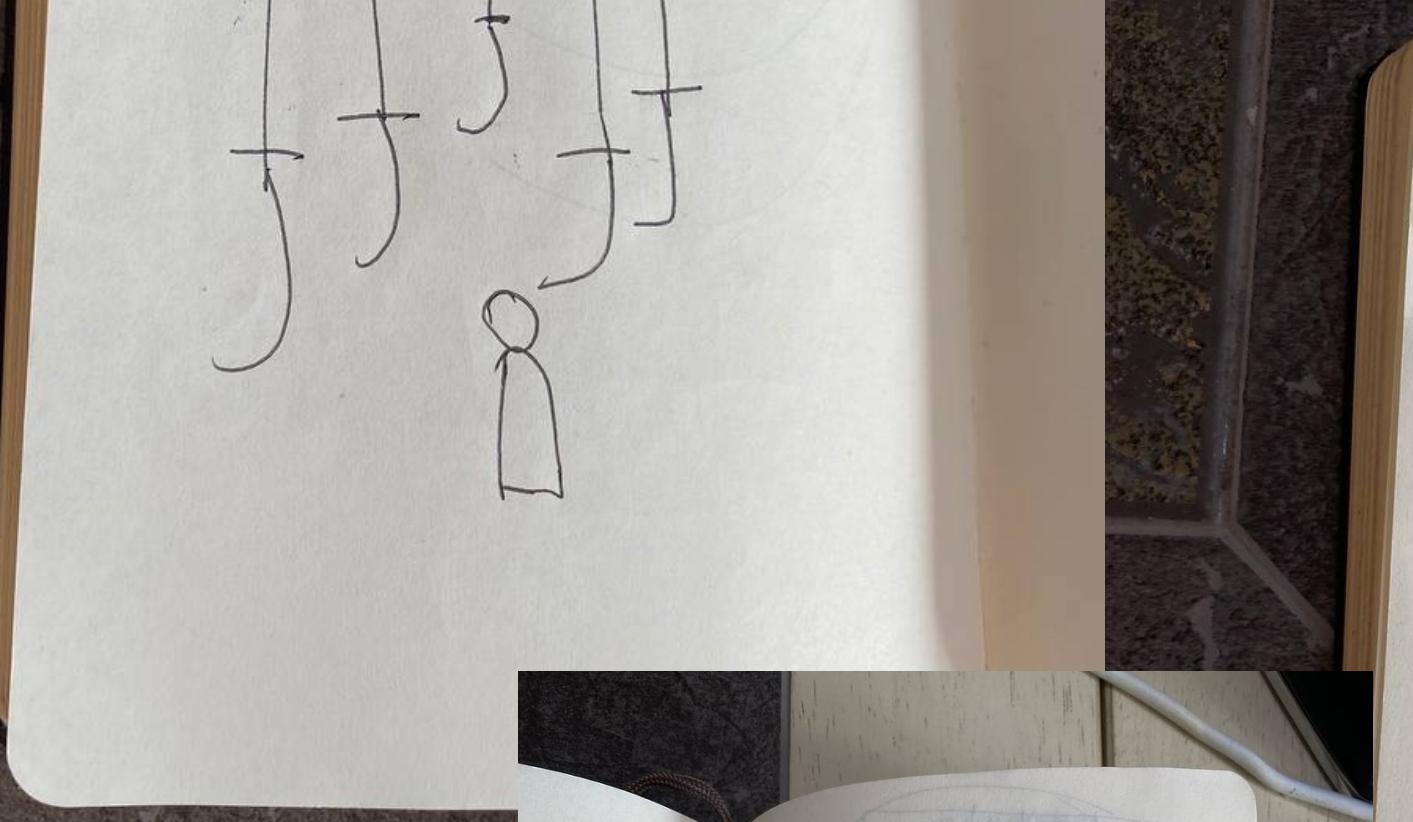


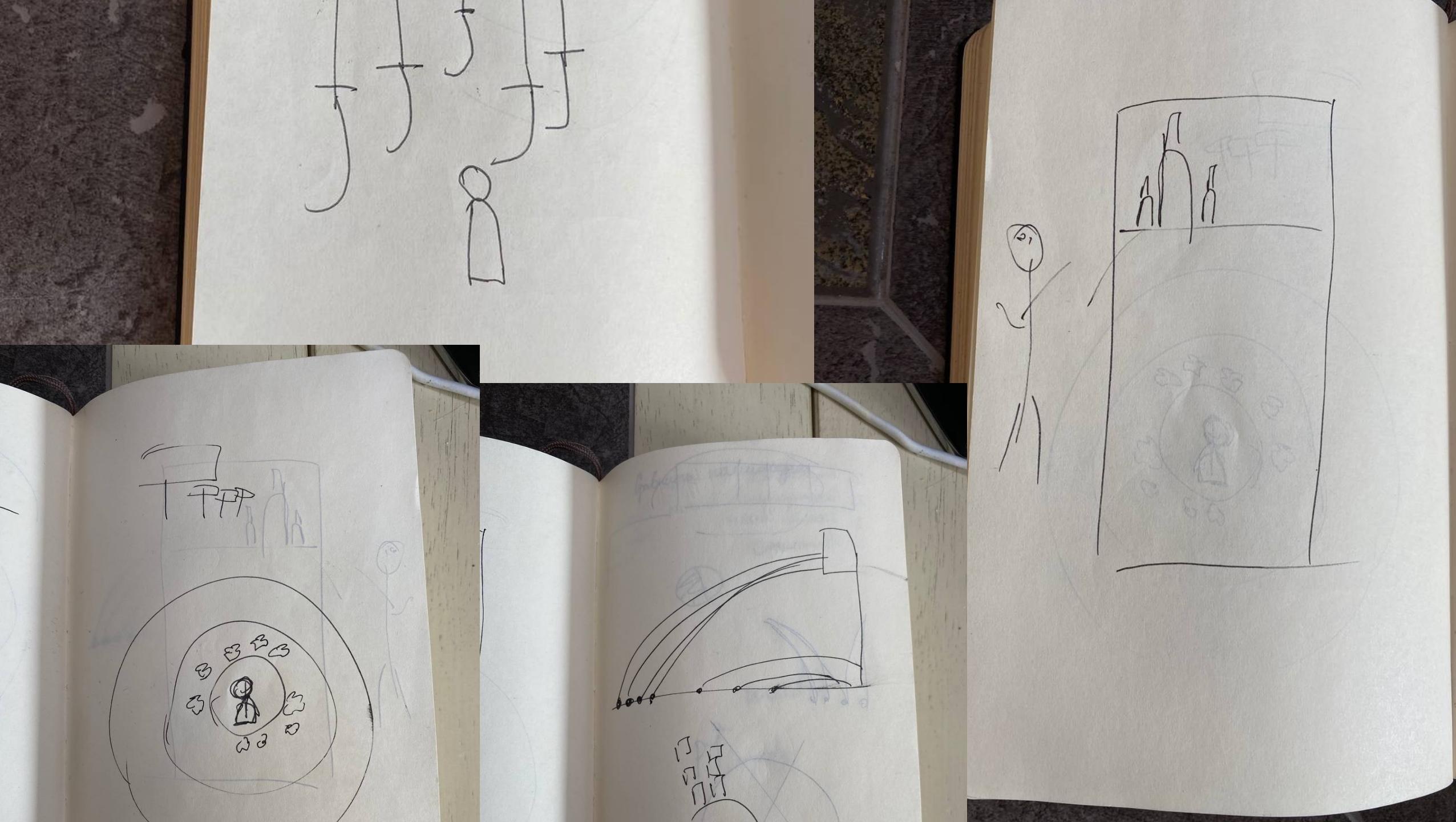
Подиум











Формула «приманки»

Далеко не всякий фильм-«приманка» в итоге получает премию, как и не всякий победитель «Оскара» непременно снимался под вкусы киноакадемии. Однако фильмы, которые номинируются на «Оскар», часто обладают некоторыми характерными особенностями

- Получили оscar
- Не получили оscar

«Приманка» с Оскаром

Рейтинг приманки 100%



Список Шиндлера

12 номинаций
7 наград

Историческая драма про Холокост, которая пробивает на эмоции не только зрителей, но и жюри

«Приманка» без наград

Рейтинг приманки 100%



Волк с Уолл-стрит

5 номинаций
Без наград

Ди Каприо опять остались без статуэтки :/

Молчание ягнят
71% приманка

Список Шиндлера
100% приманка

Волк с Уолл-стрит
100% приманка

Наведите на кружочек, чтобы
получить описание фильма



Gabriel Rossman 21:33

KOMy: MHe ▾



Wow, that's a very impressive visualization.
Congratulations on working your way through my
messy code and adapting it so well.

<https://ria.ru/20200207/1564342281.html>