



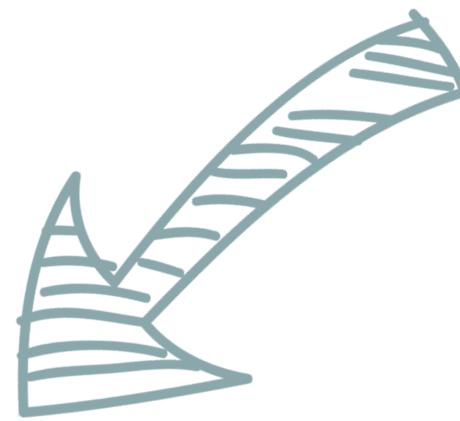
Data Science for Digital Humanities

*Что такое наука о данных (и почему мы
опять не можем использовать однозначное
определение этого направления...)*

Анна Сенина, Введение в DH 2023

+ Информатика

Статистика



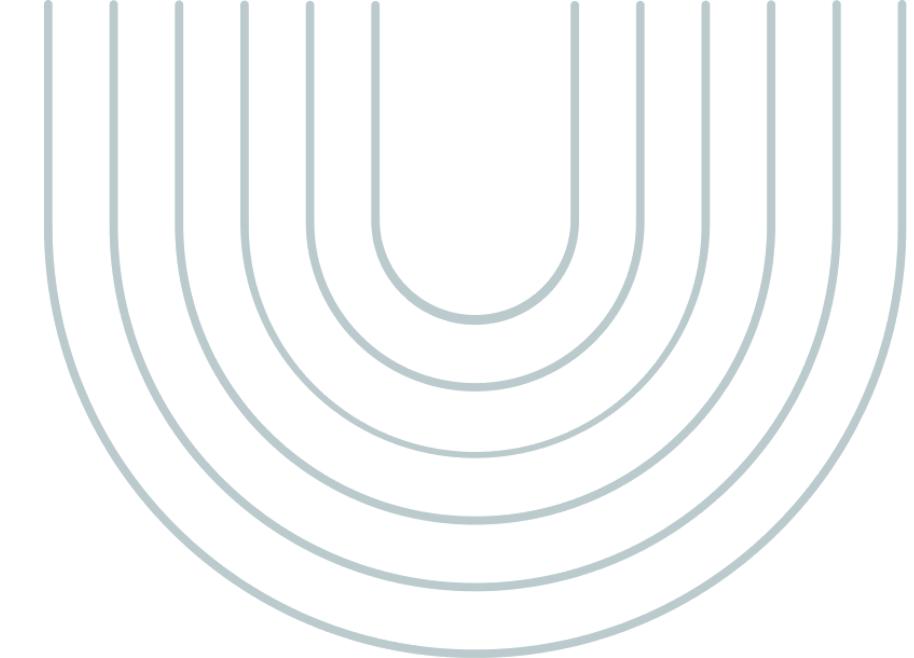
Анализ
данных

+ Программирование

Наука о
данных



Цифровые
гуманитарные
науки



01.

СБОР ДАННЫХ

Больших и малых!

02.

ОЧИСТКА

Подготовка данных к исследованию: чистые / грязные данные

03.

АНАЛИЗ + ПРОГНОЗЫ

Поговорим о разных видах

04.

ПРИНЯТИЕ РЕШЕНИЙ

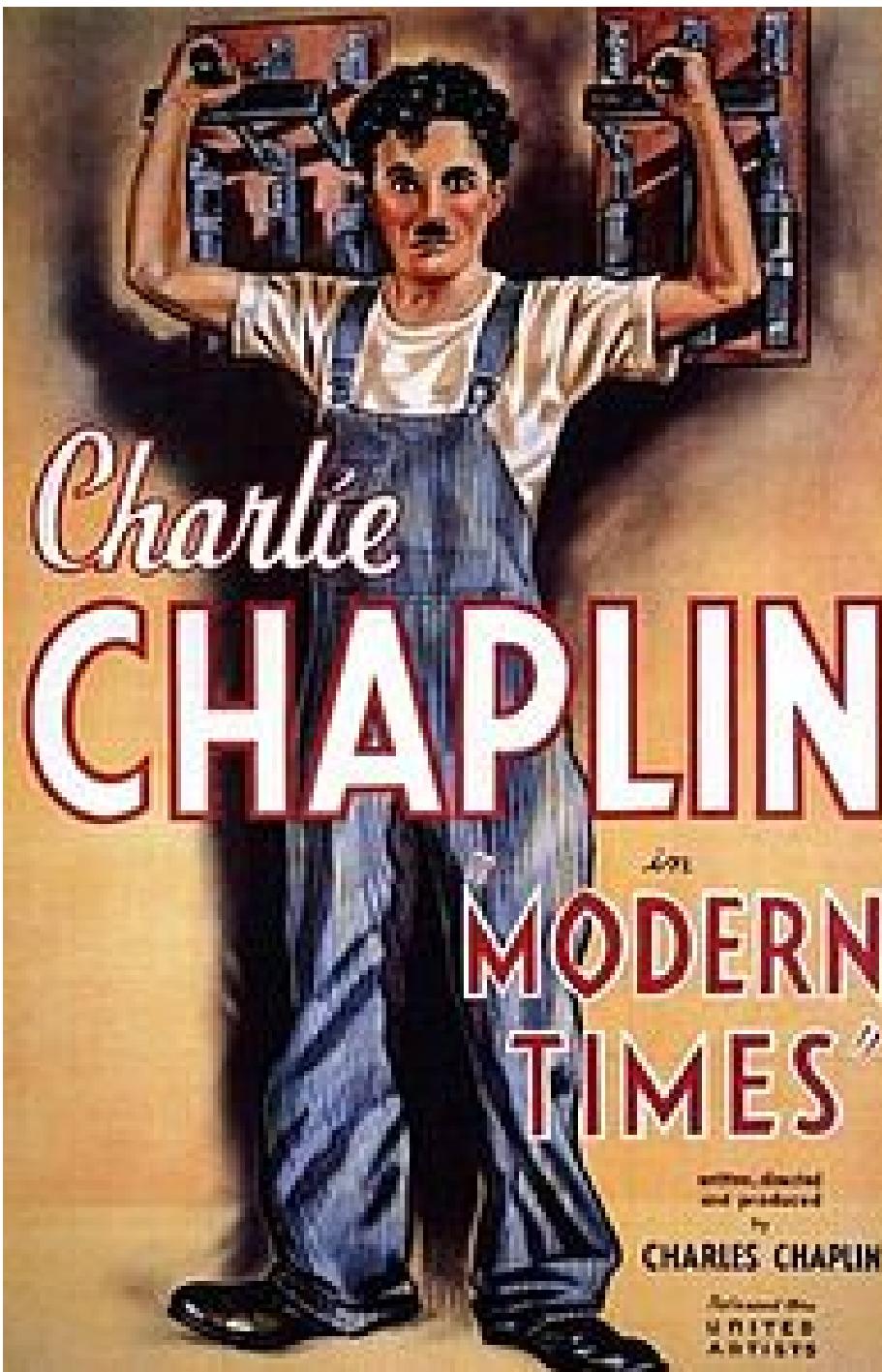
управленческих, маркетинговых, политических...

• • • • • • • •
• • • • • • • •
• • • • • • • •

Тогда чем все-таки наука
о данных занимается?..

1910-е годы

тейлоризм



В "1984" людей держат под контролем через причинение боли.



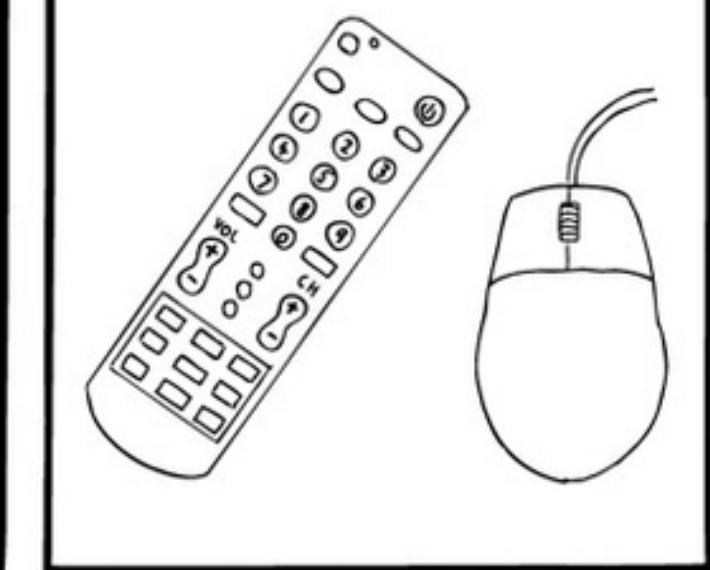
В "дивном новом мире" людей контролируют через доставление удовольствий.



Если вкратце, то Оруэлл боялся, что нас погубит то, что мы ненавидим.

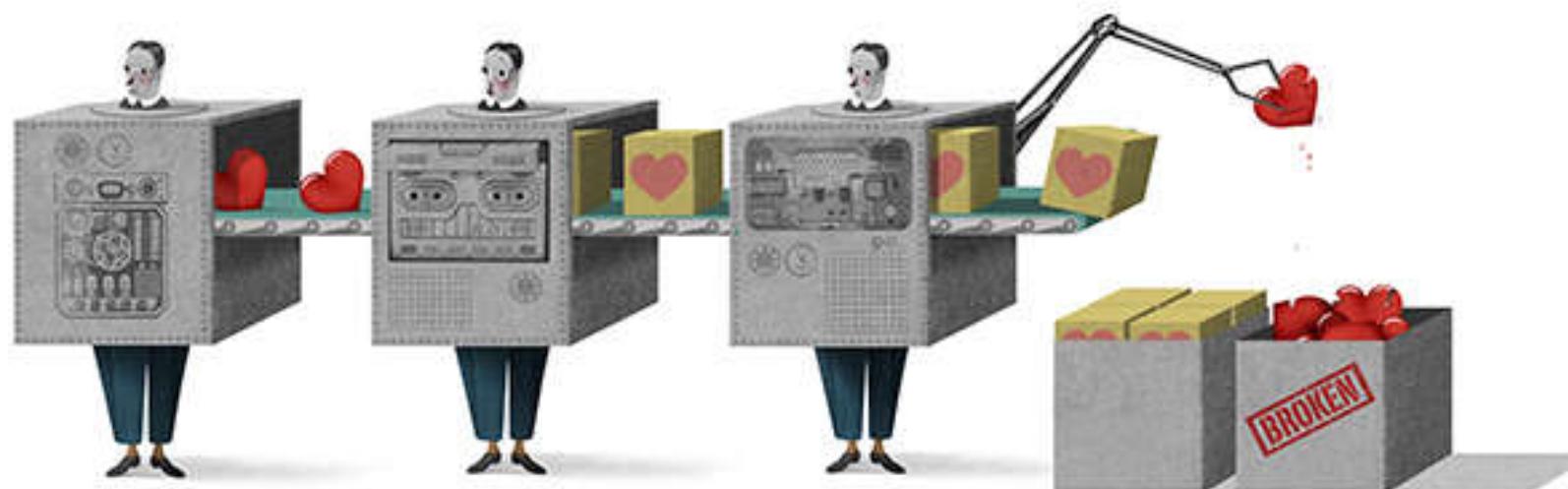


Хаксли боялся, что нас погубит то, что мы любим.



Взято из книги "AMUSING OURSELVES TO DEATH: PUBLIC DISCOURSE IN THE AGE OF SHOW BUSINESS" Нейла Постмана
Книги о том, что возможно был прав Хаксли, а не Оруэлл.

сегодня: цифровой тейлоризм



Shurick Agapitov 31 Jul



...

...
i...

Вы получили это письмо, потому что моя команда биг дата проанализировала ваши активности в жире, конфлюенс, гугл почте, чате, документах, дашбордах и пометила вас как невовлеченные и малопродуктивные сотрудники. Иными словами вы не всегда присутствовали на рабочем месте тогда, когда работали удаленно.

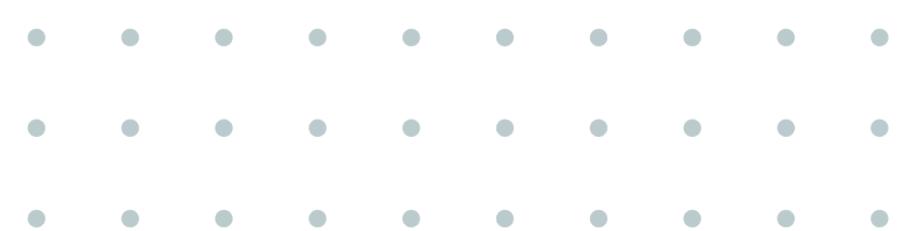
для многих из вас это может быть шоком, но я искренне верю, иксолла не для вас. так что все что не делается, все к лучшему. Надя и ее команда заботы орагнизовали партнерство с семью ведущими HR агенствами и мы поможем вам найти хорошее место, где вы будете получать еще больше, а работать еще меньше.

Саша вам поможет получить рекомендацию, включая от меня лично. А Наталья вам зачитает ваши права.

С 1960-х статистика
соединяется с
информатикой



Кто узнает этот
автомобиль?



В 1994 г. компании уже
обладают информацией
о людях, но обрабатывать
ее еще тяжело...



→ Данные, данные повсюду



В мире информации

Каждый год объем информации
увеличивается на 30 %

информационный взрыв!

до 15 тыс.
рекламных
сообщений в
день (vs 500
в советское
время)

в среднем
7 часов в
Интернете
каждый день

от 300 до 1 тыс.
сообщений в
ленте в день



Да, мы все еще не в научном
треке, но скоро туда попадем...

ГЕНДЕРНЫЕ ДАННЫЕ

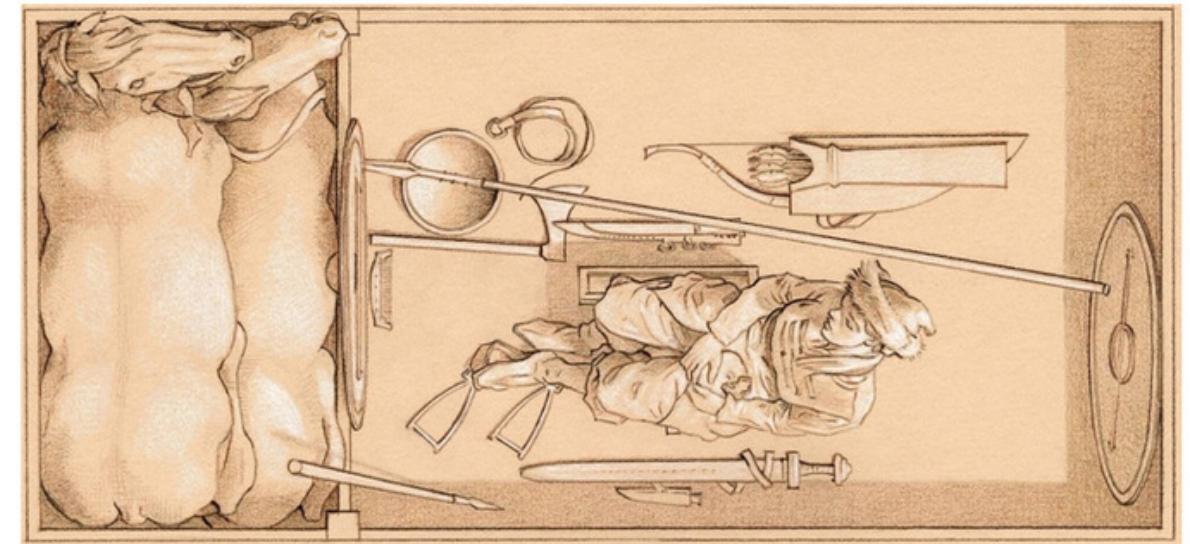
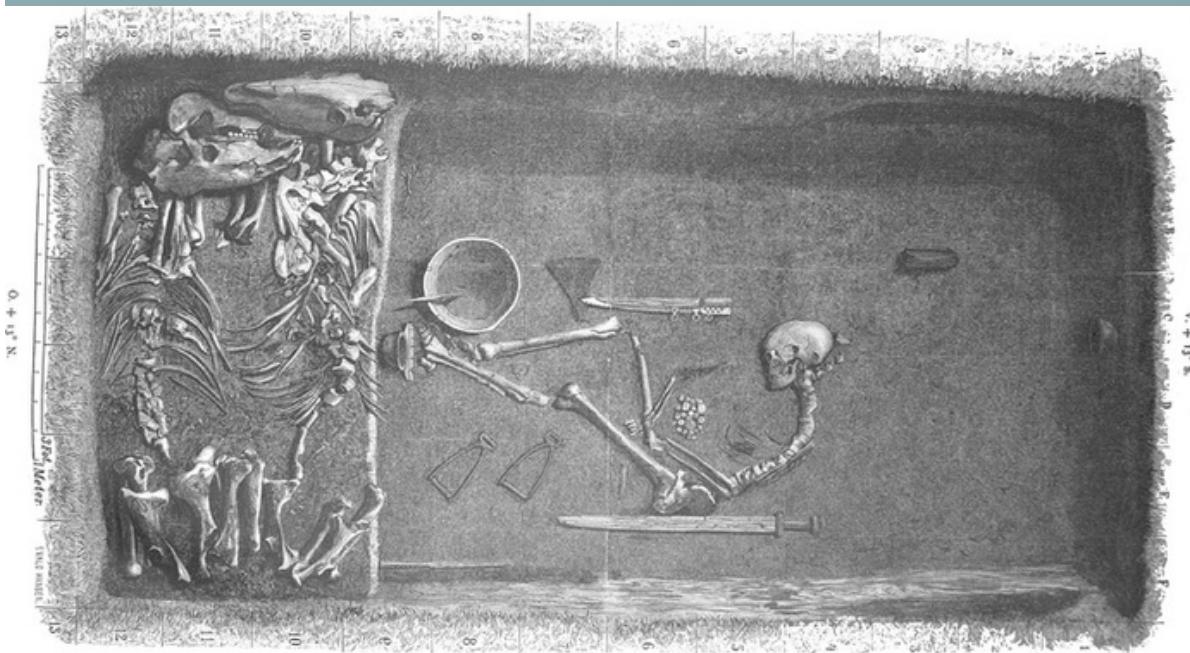
просто примеры:

- уборка снега (в скандинавских странах)
- испытание лекарств
- автомобильные ремни: краш-тесты
- униформа и многое другое...

• • • • • • •
• • • • • • •
• • • • • • •



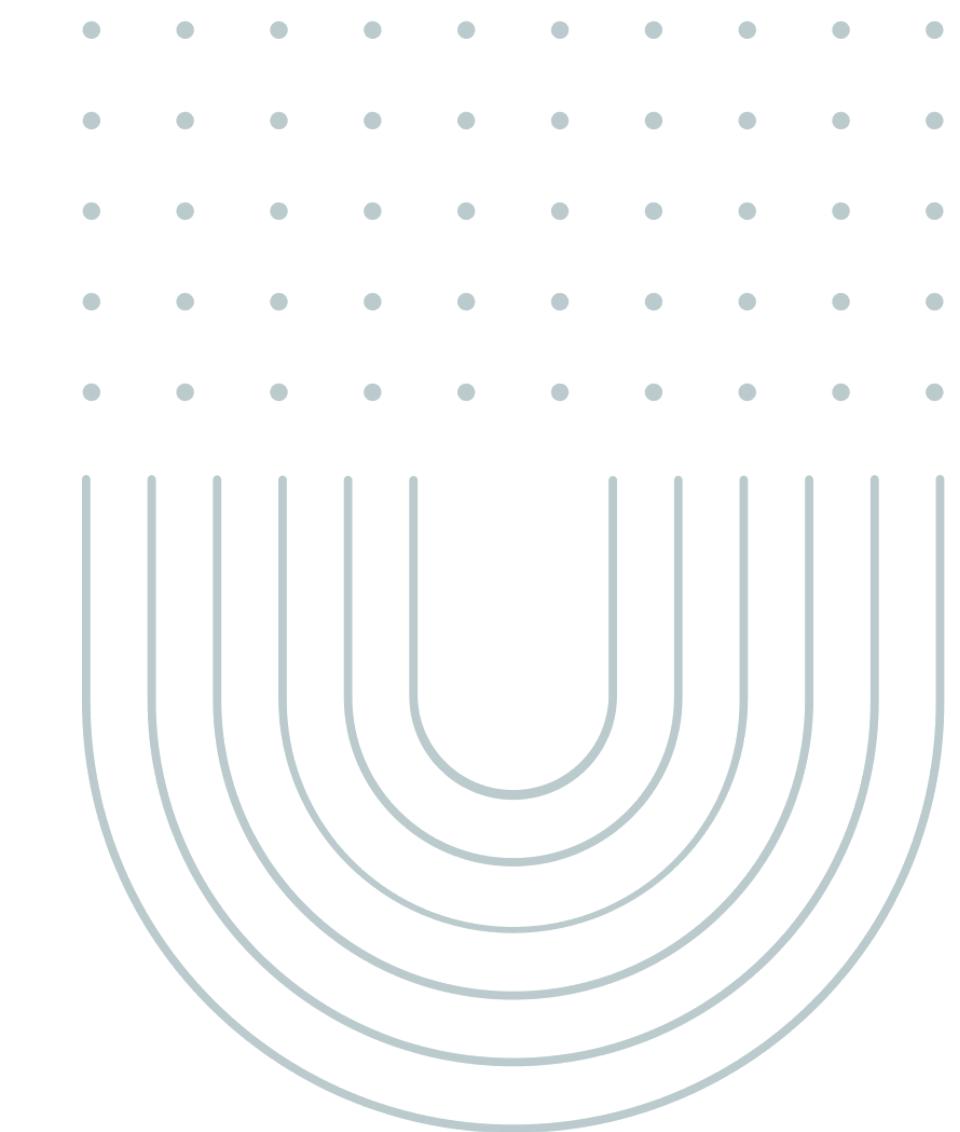
02.



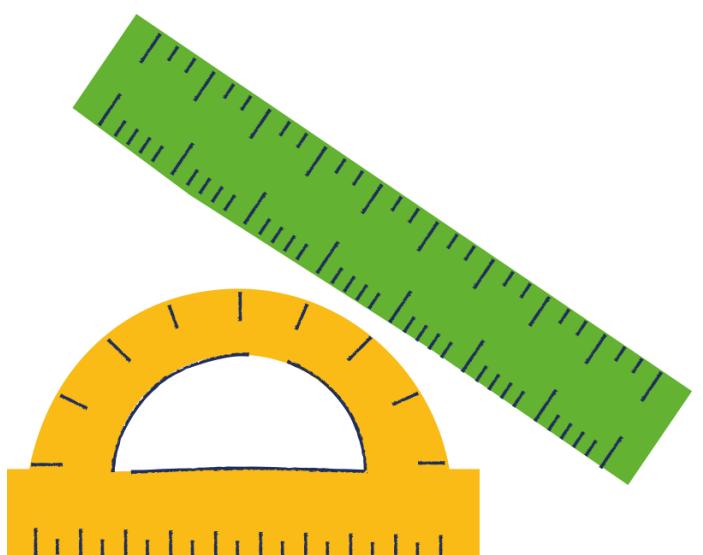
А где здесь история? (или хотя бы DH)

*Обзор методов математики,
статистики и Data Science*





Неметрические шкалы
шкала наименований
порядковая шкала



Метрические шкалы
шкала интервалов
шкала отношений

**ИЗМЕРЕНИЯ
И ШКАЛЫ**

• • • • •

*“Жизнь общества и поведение
людей измеримы и моделируемы,
иначе не работали бы
политические и рекламные
технологии”*

(Д.А. Гагарина)

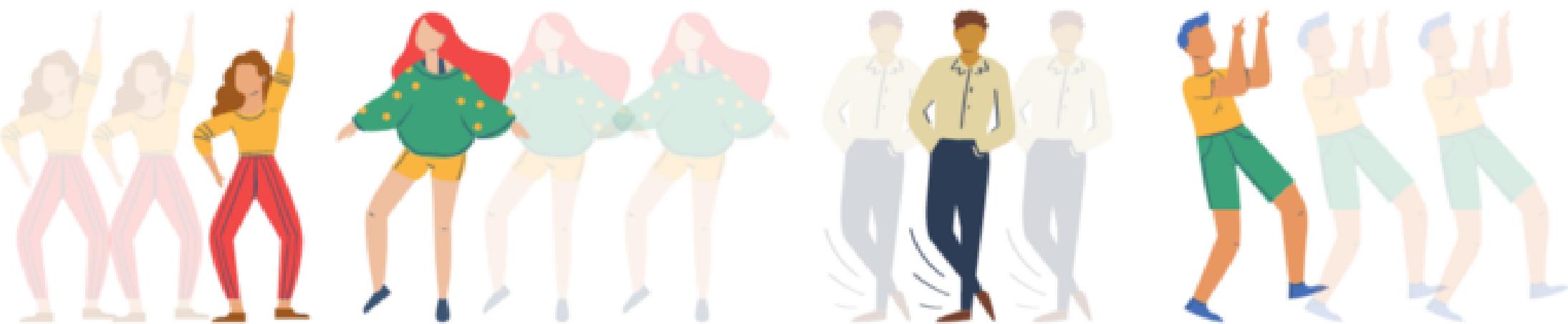




Неслучайные выборки

Вынужденный вариант:

- когда есть общие сведения о генеральной совокупности
- эксперты
- снежный ком

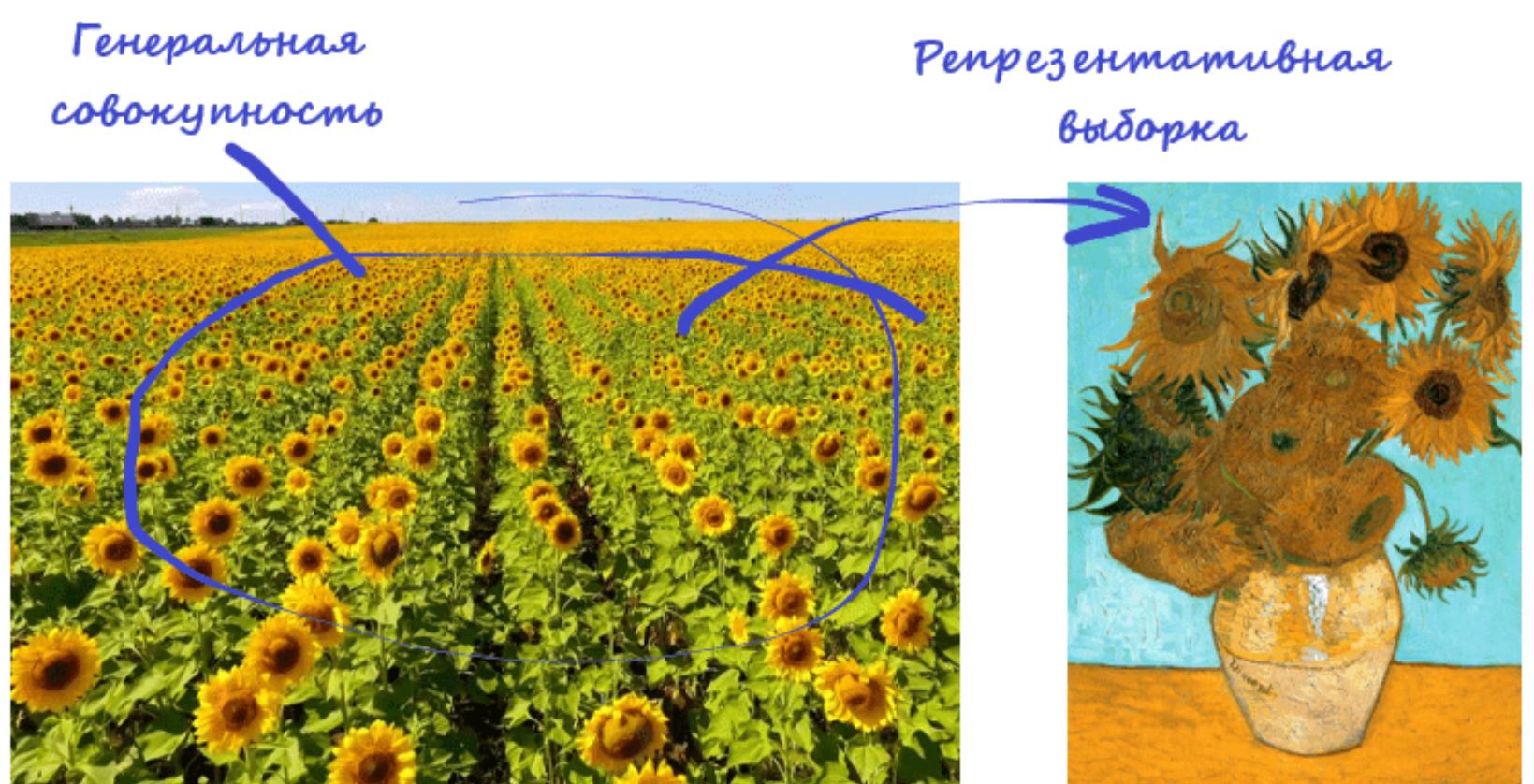


Случайные выборки

Всегда лучше, потому что статистика гарантирует достоверность выводов, а случайную ошибку можно измерить

Репрезентативность выборки

Если суп хорошо перемешать, то достаточно одной ложки, чтобы сделать вывод о вкусе всей кастрюли (почти Д.Гэллоп)



Генеральная
совокупность

НЕрепрезентативная
выборка

Предвыборный опрос «Литрери Дайджест», 1936 г.

Почтовый опрос общественного мнения о вероятных результатах грядущих президентских выборов в США. До 1936 года опрос всегда правильно предсказывал победителя.

Цель – предсказать результаты президентских выборов в США.

Разосланы 10 млн анкет:

- подписчикам журнала;
- людям, выбранным по телефонным книгам;
- по спискам регистрации автомобилей.

Около 2,5 млн анкет заполнены:

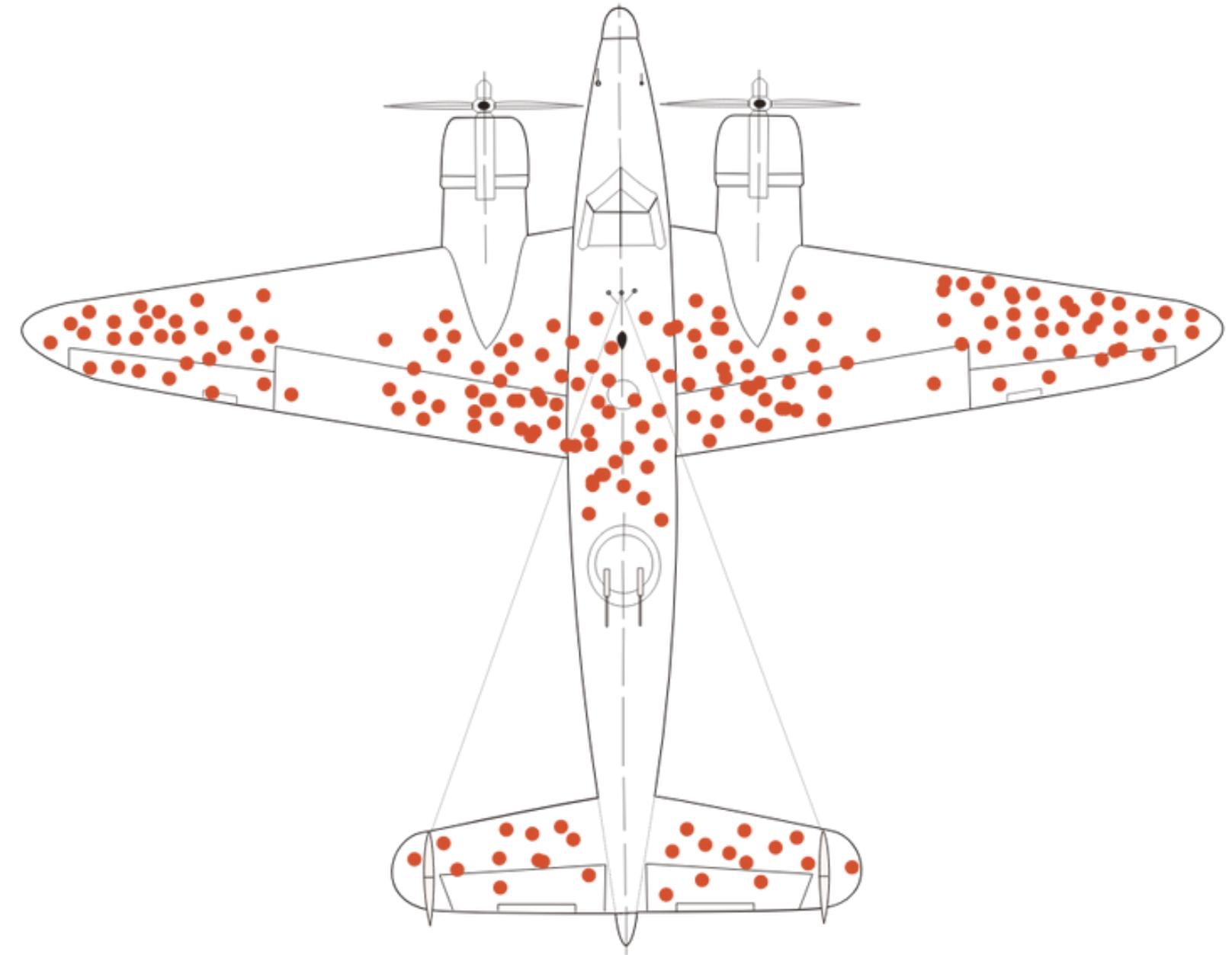
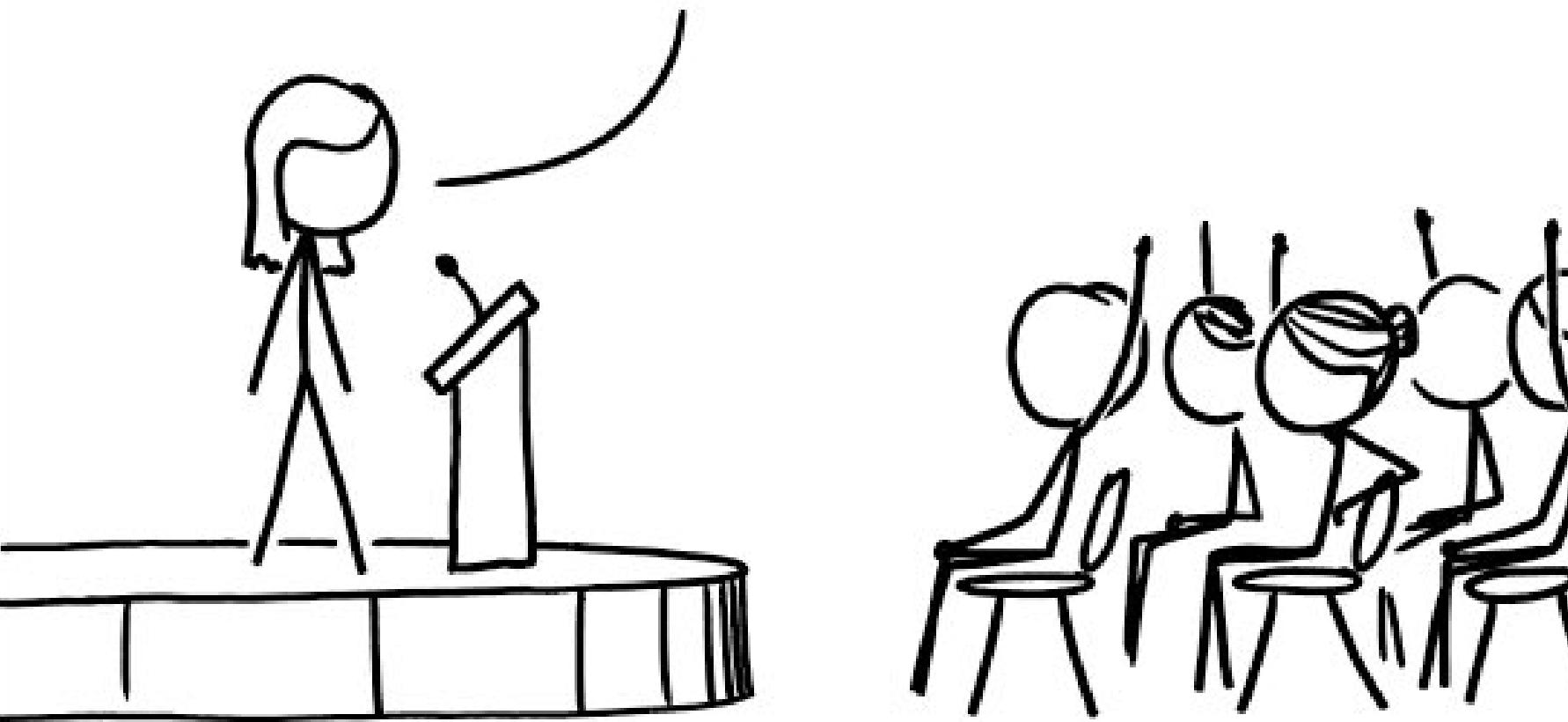
- 57 % – за республиканца Альфа Лэндона,
- 40 % – за демократа Франклина Рузвельта.

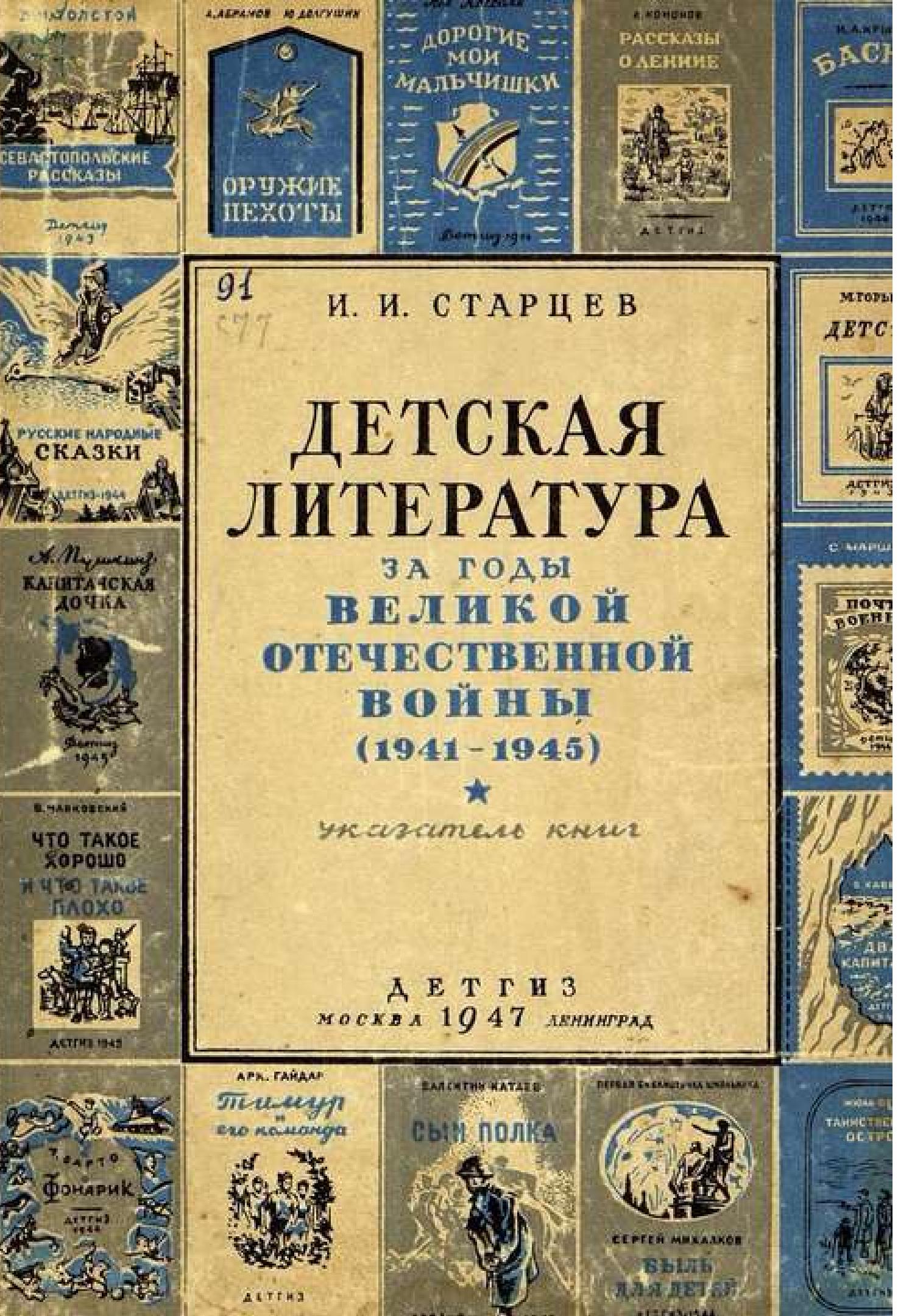
Что могло
пойти не так?

STATISTICS
CONFERENCE
~2022~

RAISE YOUR HAND
IF YOU'RE FAMILIAR
WITH SELECTION BIAS.

AS YOU CAN SEE,
IT'S A TERM MOST
PEOPLE KNOW...





Закон нормального распределения (Гаусс)

1

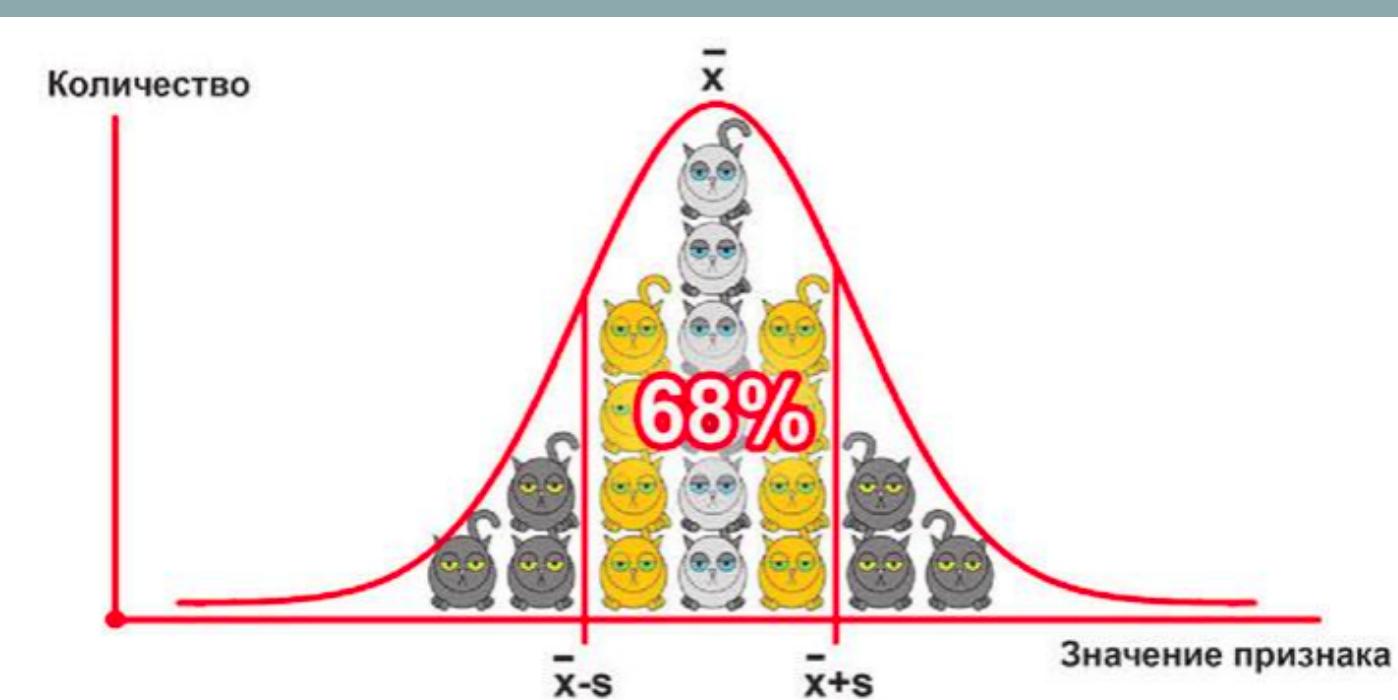
Большая часть данных в мире нормально распределена вокруг среднего

2

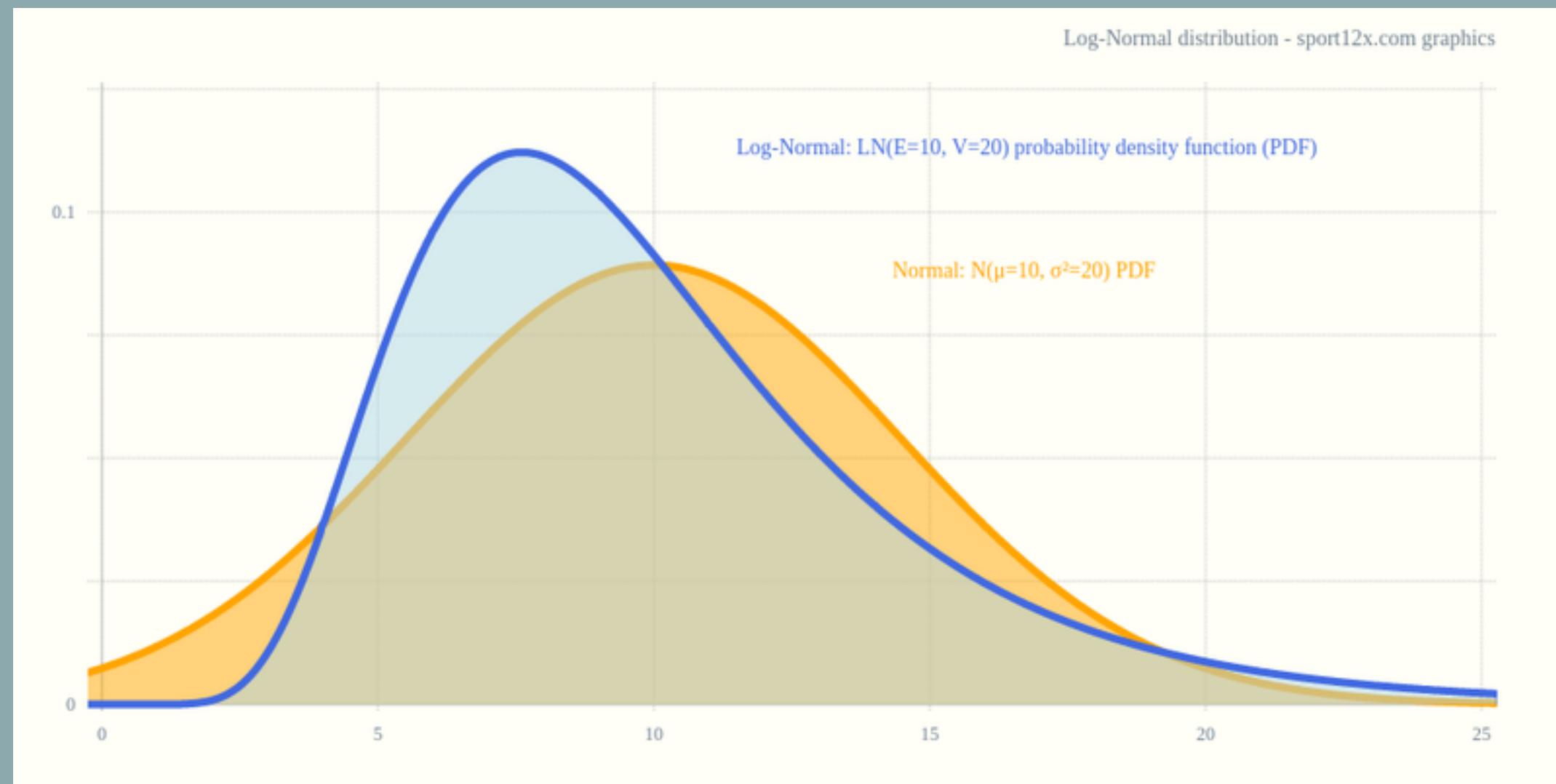
Аномалии – должны иметь объяснение

3

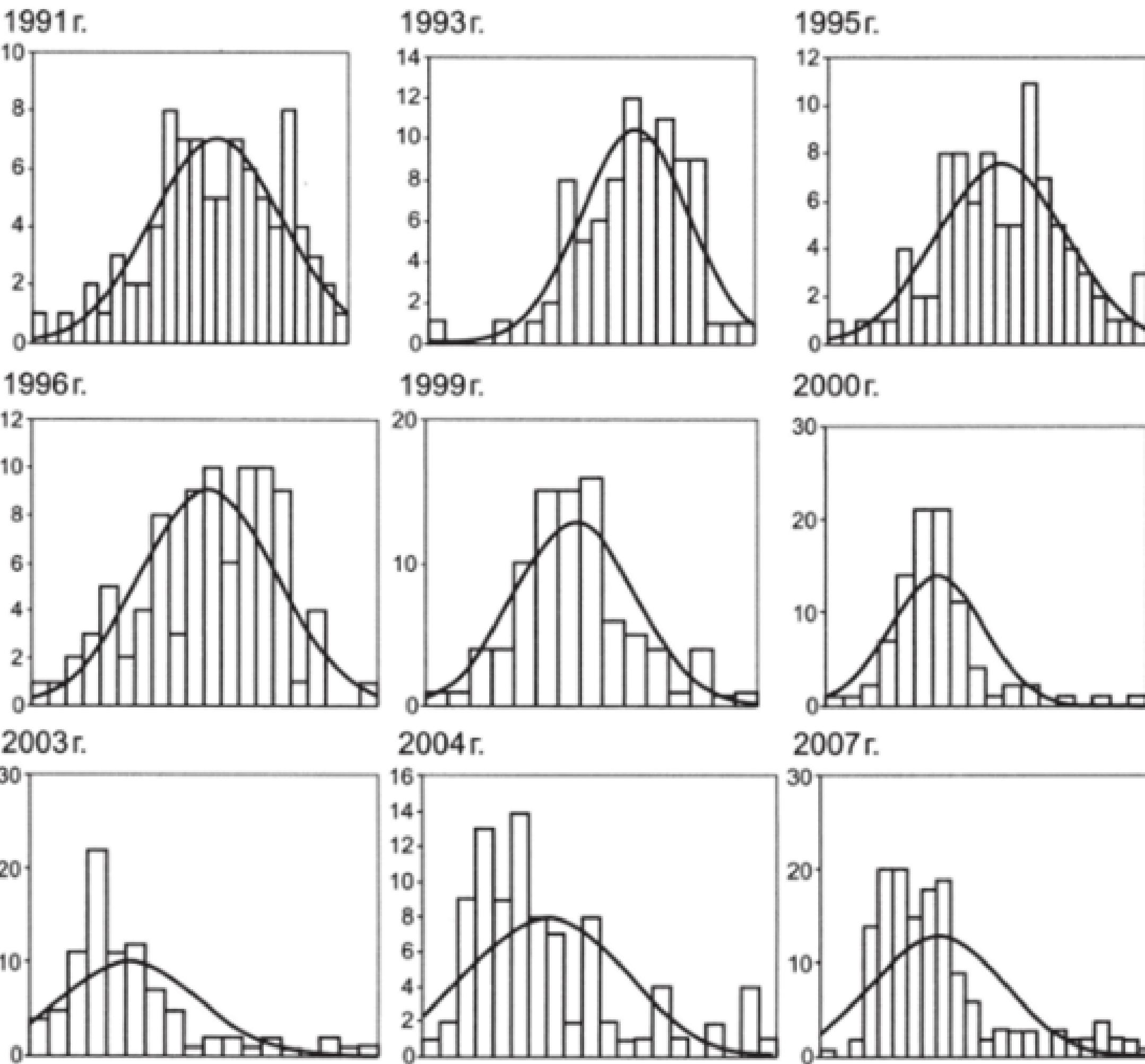
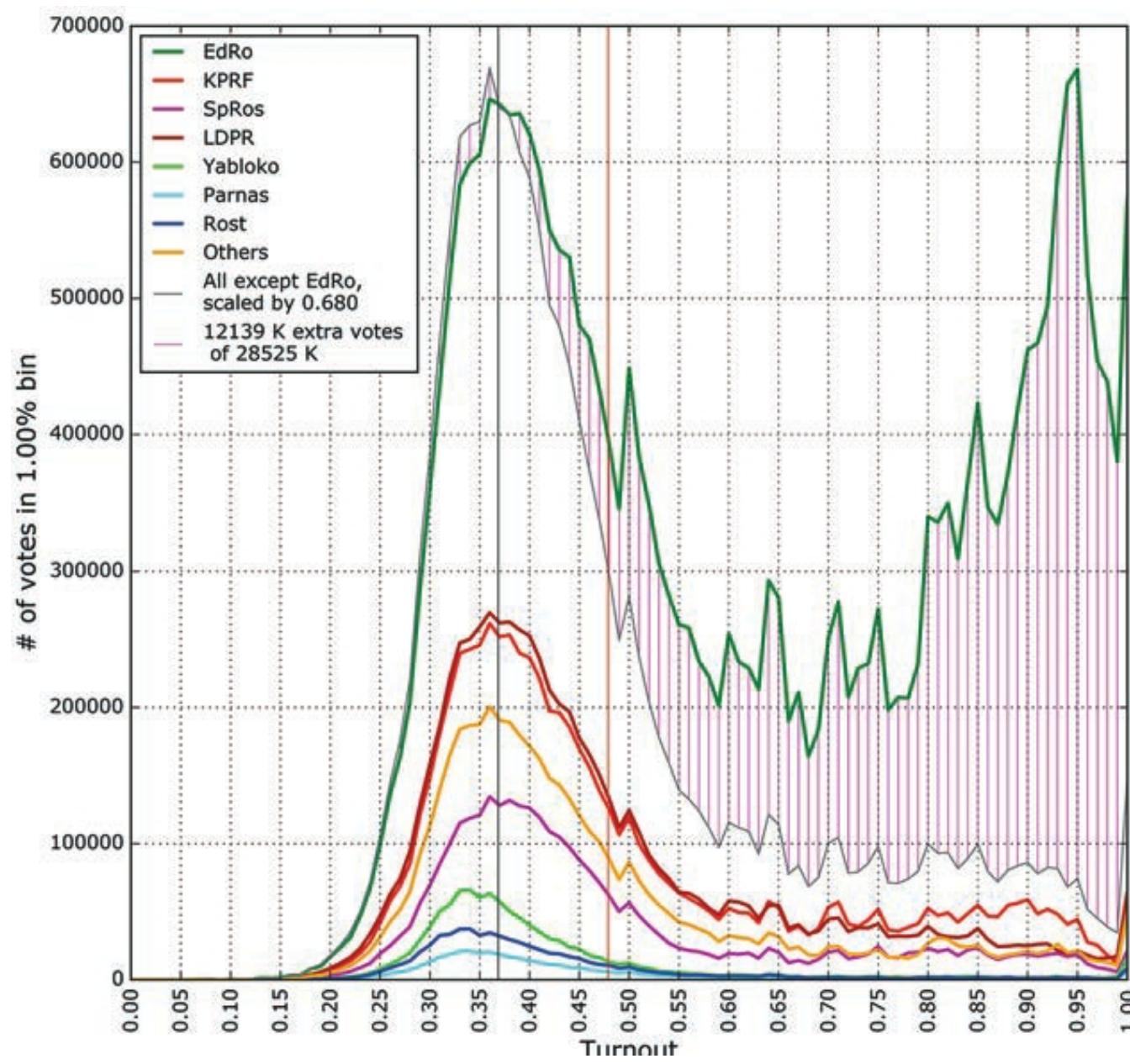
У данных есть интервалы доверия (и вероятность ошибки в выводах)



(не)нормальное распределение



(не)нормальное распределение

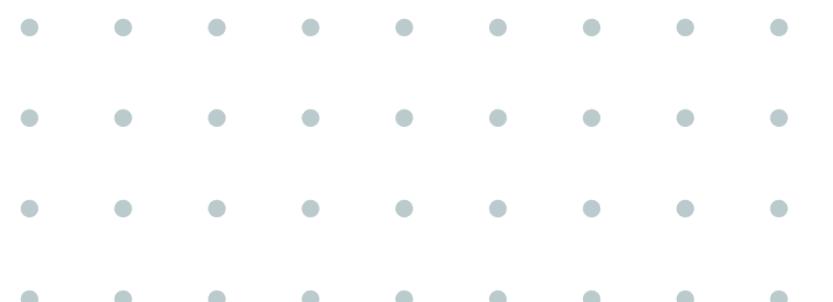


ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Ключевая идея: наиболее точно описать типичный портрет, город, страну...
...и не получить среднюю температуру по больнице

Среднее арифметическое – **не всегда** надежная мера

Есть показатель точнее: медиана
(среднее по порядку значение)



доходы в обществе 1

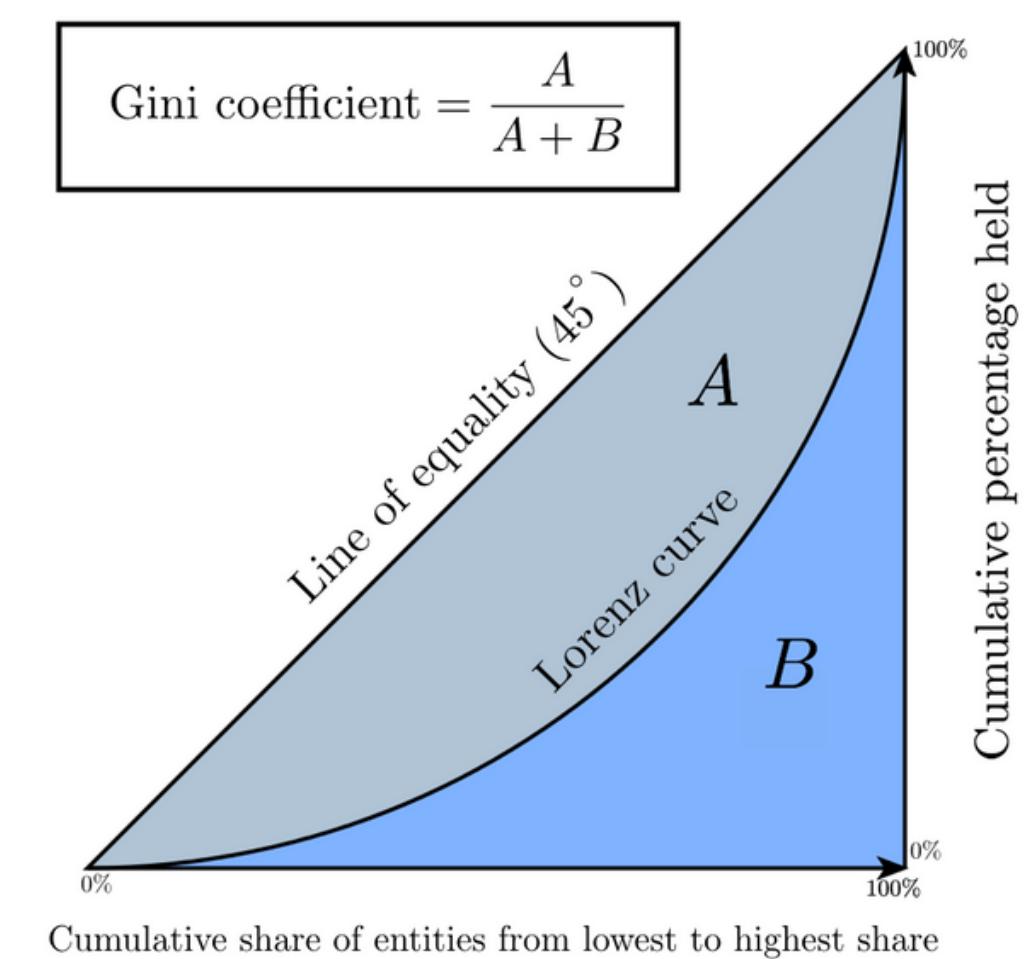
49, 49, 49, 49, 50, 50, 50, 50, 51, 51, 52 ...
Среднее = 50

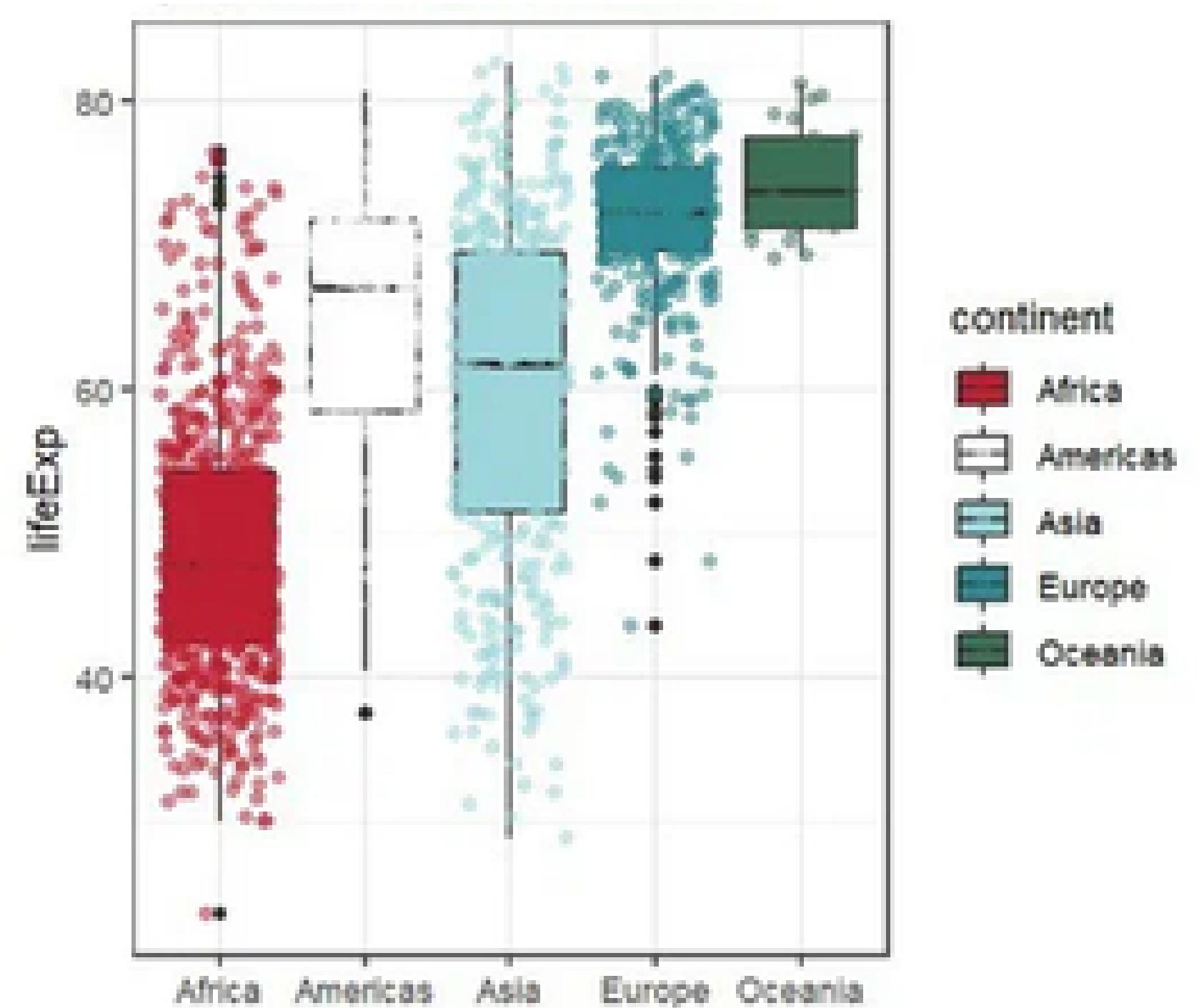
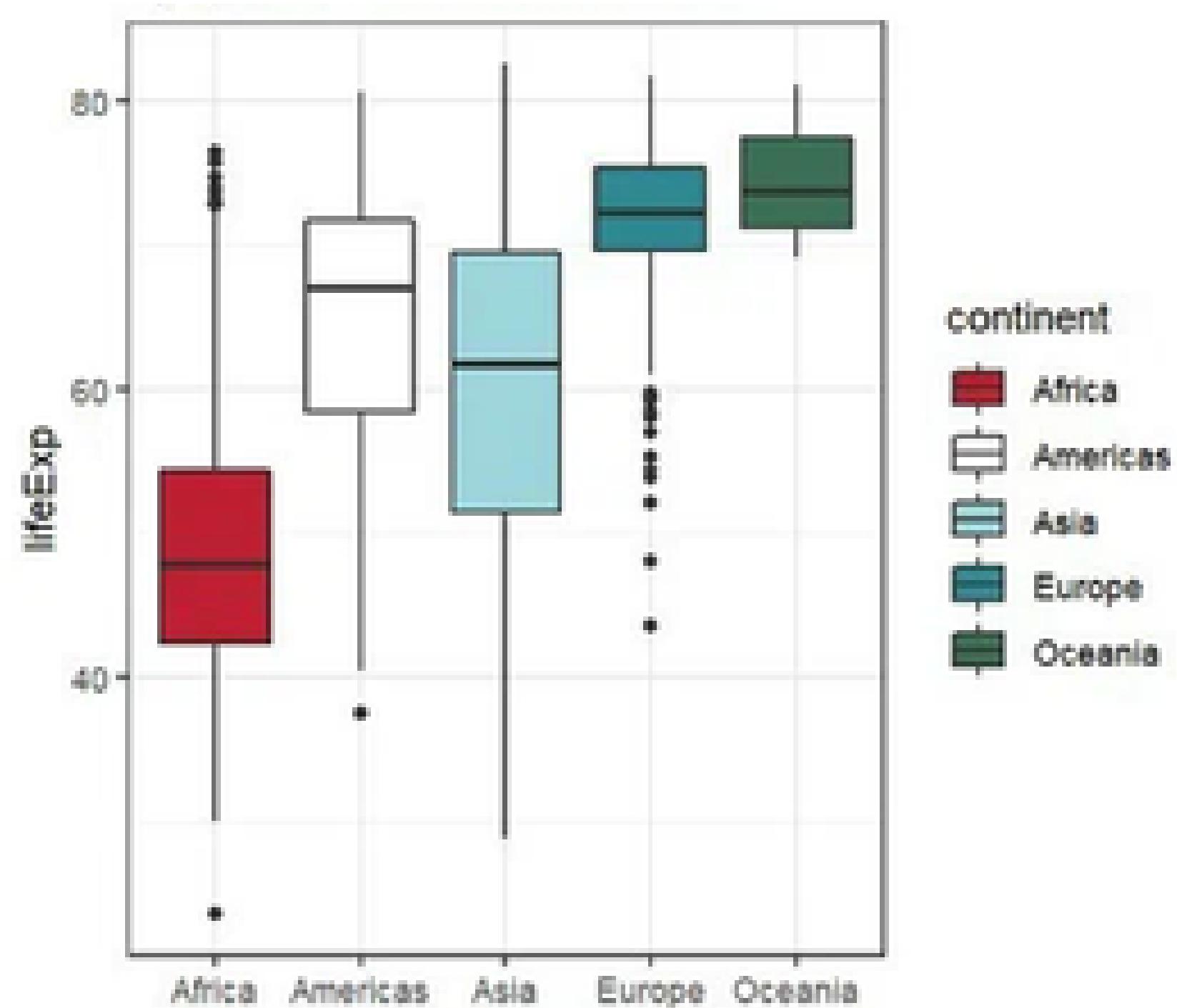
доходы в обществе 2

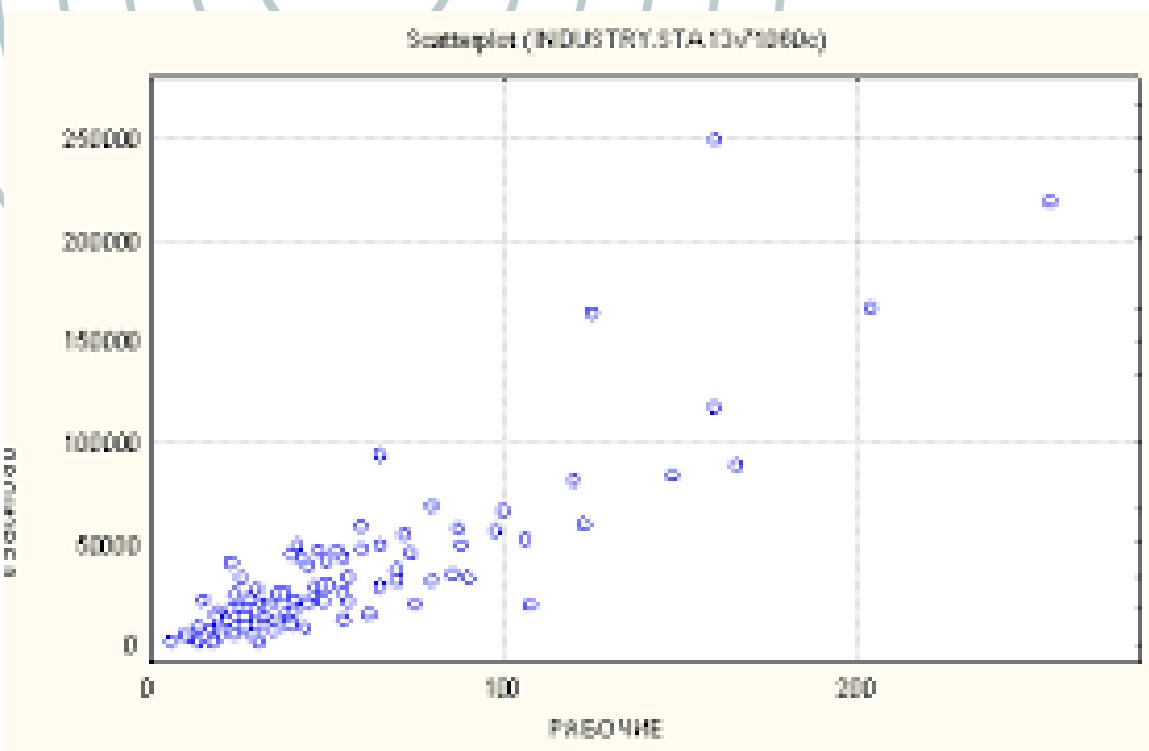
20, 20, 20, 20, 20 ... 20, 1000, 1000, 1000
 $(95 * 20 + 3 * 1000) / 98$
Среднее = 50

=> социальное расслоение

=> социальное равенство





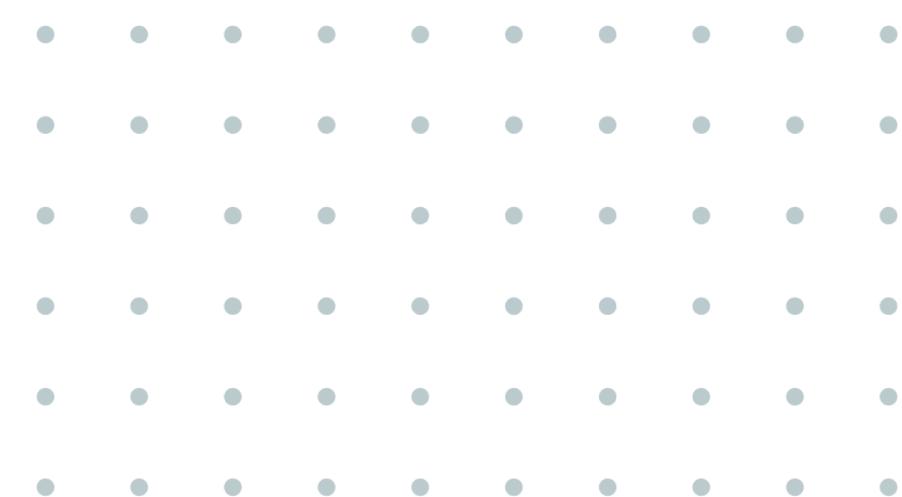


Коэффициент корреляции

отображает связь между
случайными величинами

Корреляционный анализ

метод обработки
статистических данных, с
помощью которого
измеряется теснота связи
между двумя или более
переменными



ЕЩЕ ВЗАИМОСВЯЗИ



ГИПОТЕЗЫ

Гипотеза о независимости
переменных (хи-квадрат
независимости Пирсона)



МОДЕЛИРОВАНИЕ

В исторической информатике:
математические модели в
истории



СЕТИ

Связи объектов в сети



ФАКТОРНЫЙ АНАЛИЗ

Структура зависимостей

МОДЕЛИРОВАНИЕ В ИСТОРИИ

Модель

- объект, замещающий в процессе исследования объект-оригинал
- некоторый образ реального объекта, отражающий существенные свойства объекта и заменяющий его в процессе решения задачи

Моделирование

- процесс создания модели

Объекты моделирования

- материальные объекты
- явления и процессы (клиодинамика)
- исторические источники
- знания



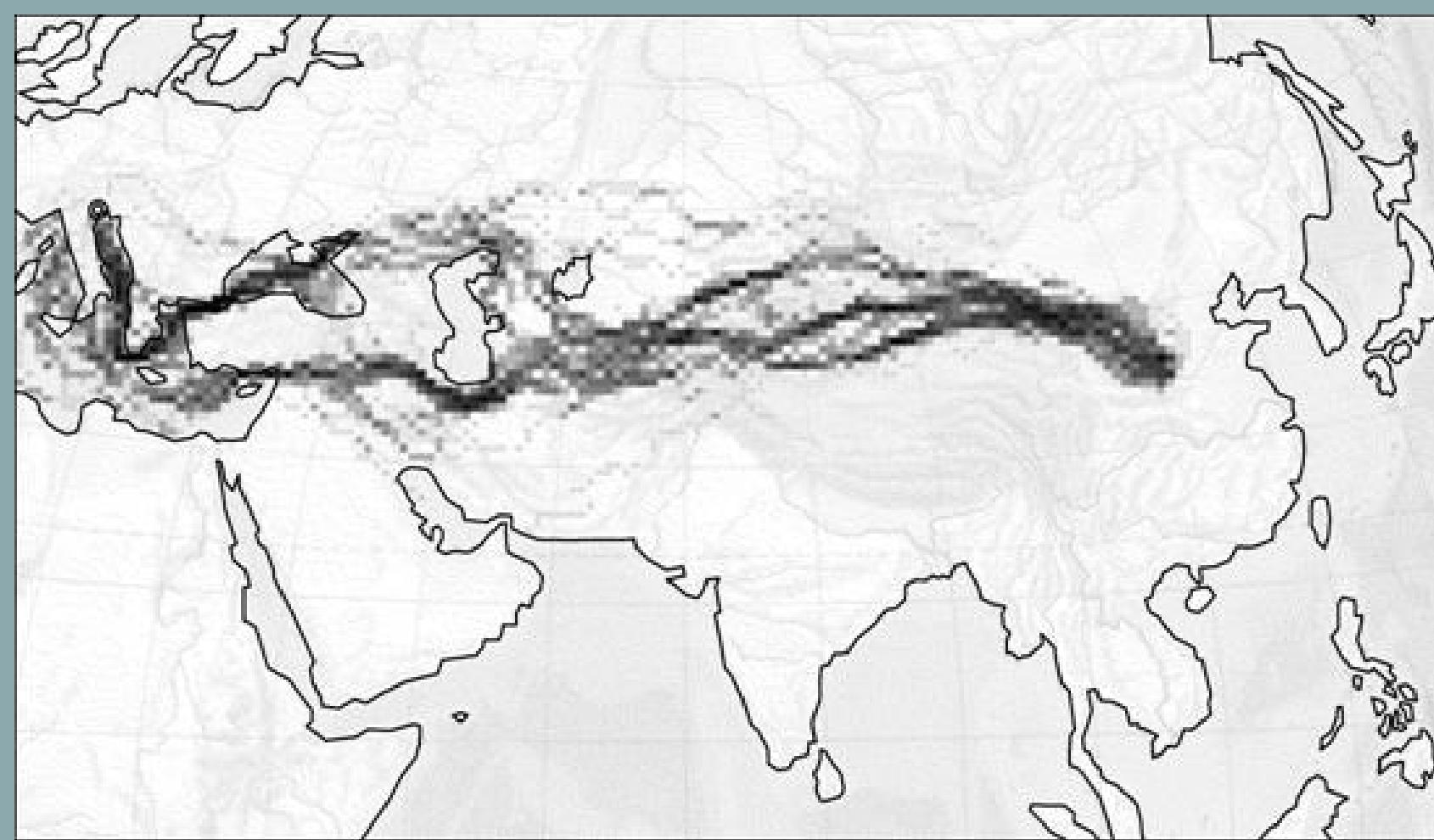
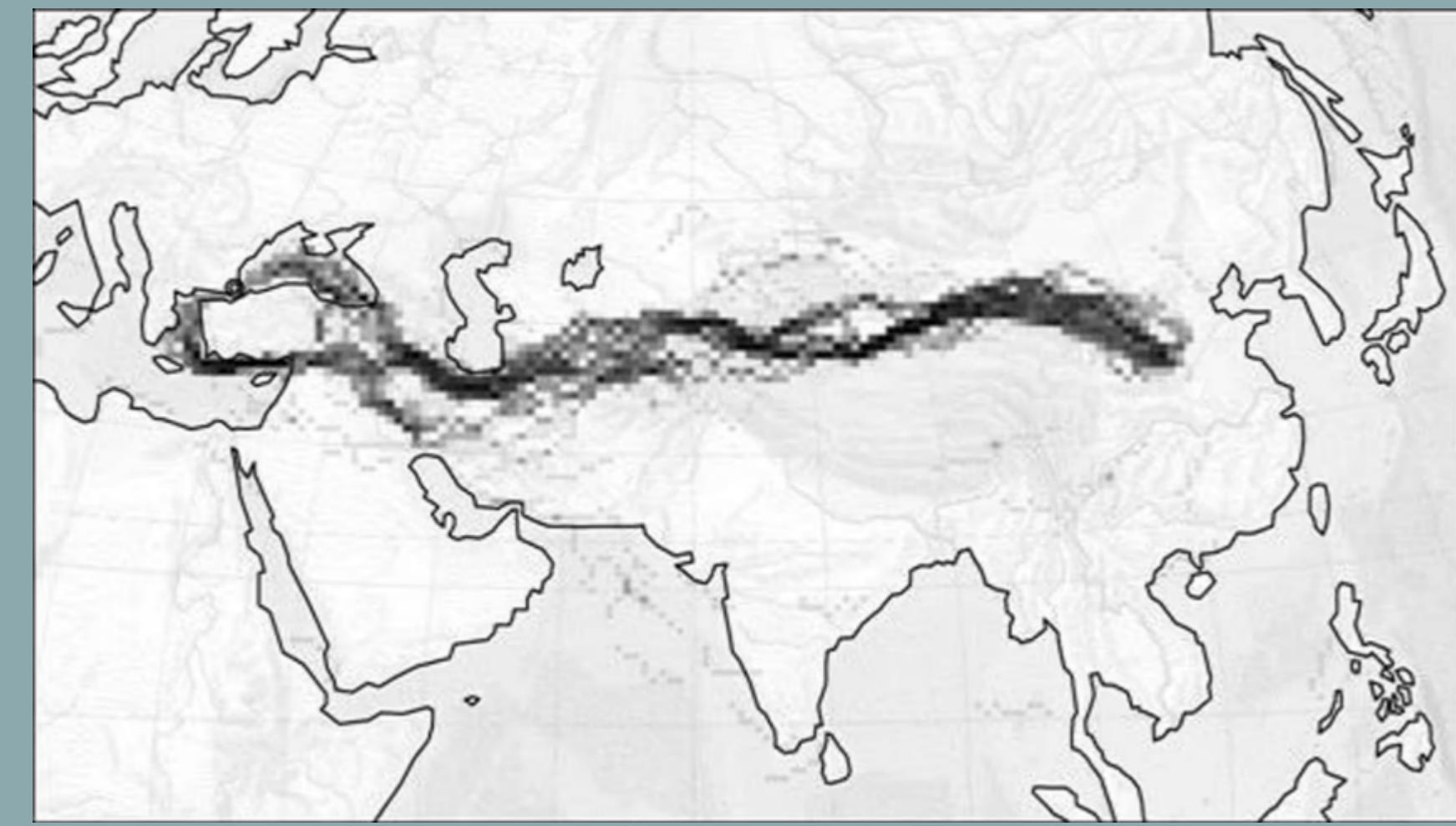
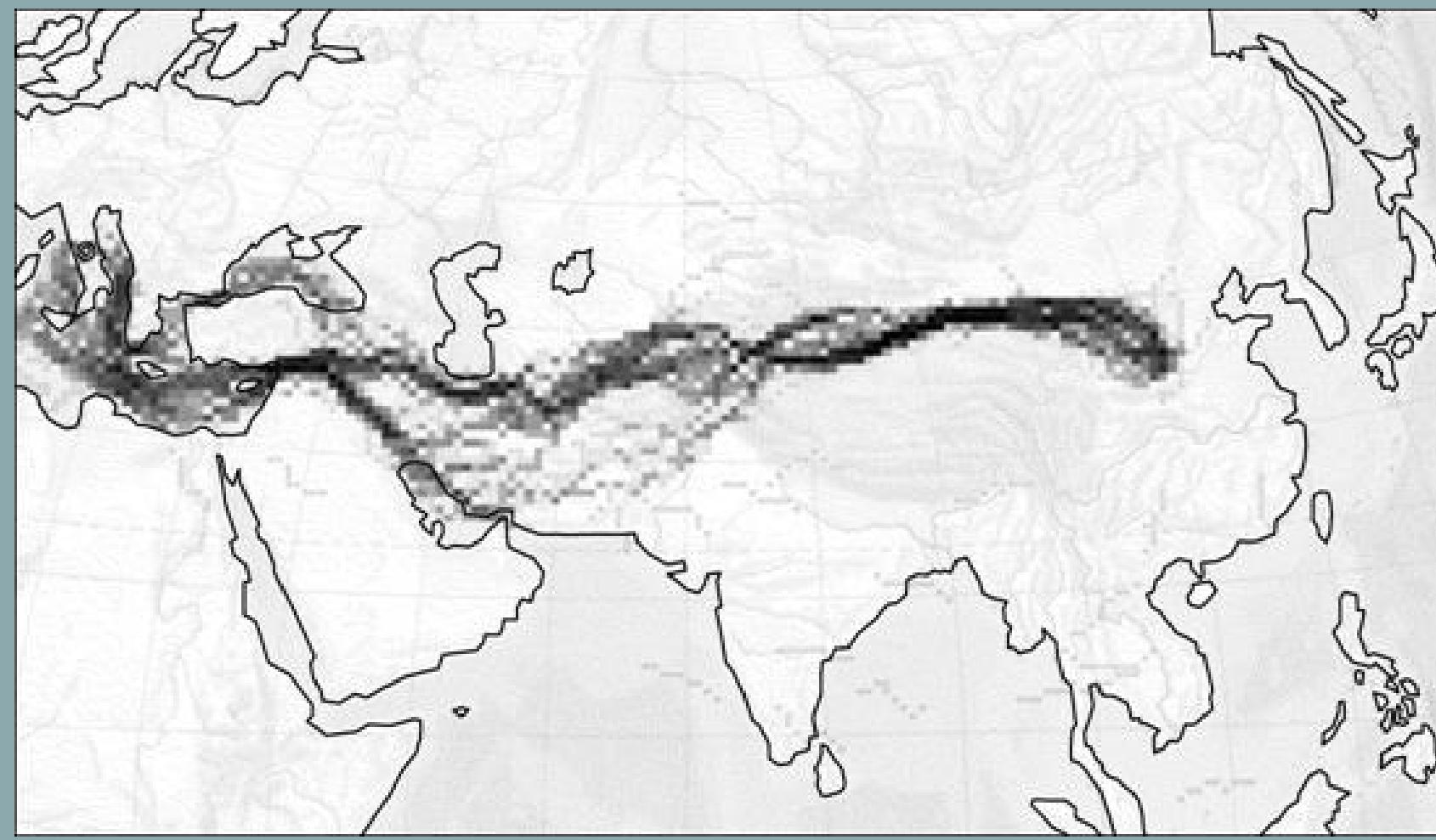
ВИДЫ МОДЕЛЕЙ

- Отражательно-измерительные
- Имитационно-прогностические
- Контрфактические

Реконструкция Тамбовской крепости (Кончаков Р.Б., Жеребятьев Д.И.)



Схема комплексной системы артефактной реконструкции



БУКВАЛЬНО НЕСКОЛЬКО СЛОВ О МАШИННОМ ОБУЧЕНИИ

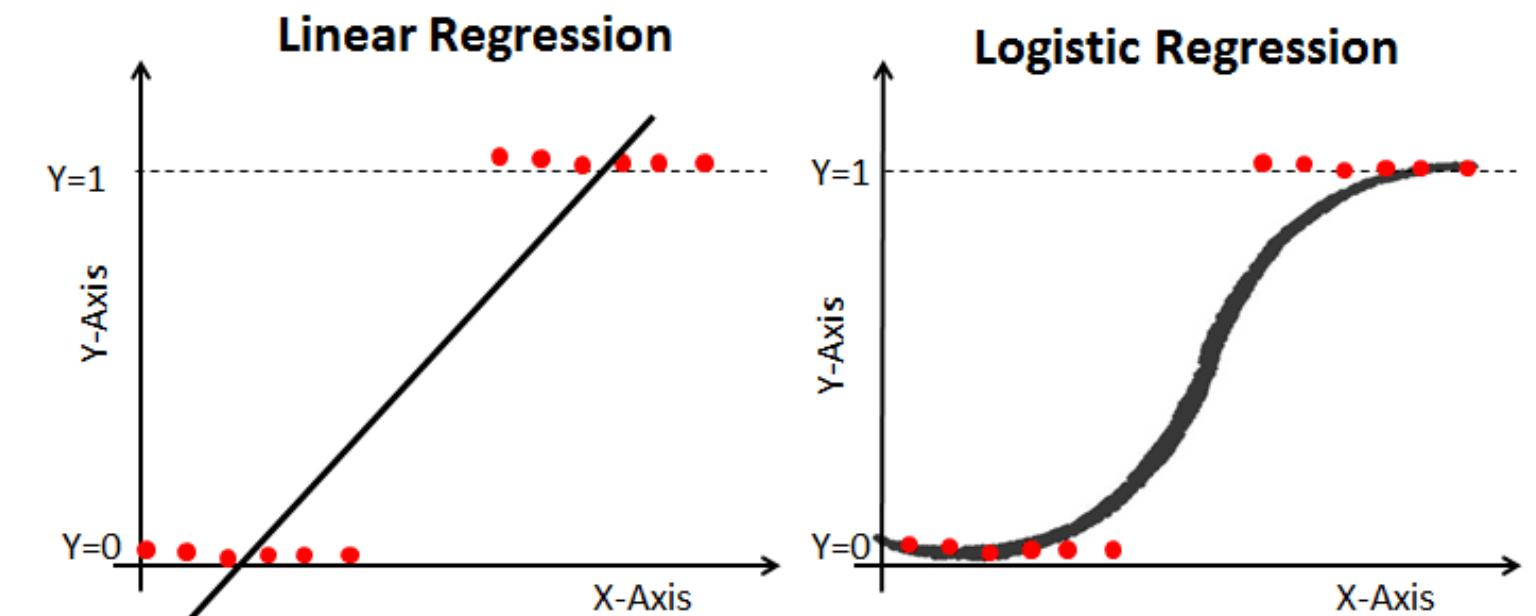
Классификация

Предсказание, умрет
пассажир Титаника или
нет в зависимости от пола
и пассажирского класса



Кластеризация

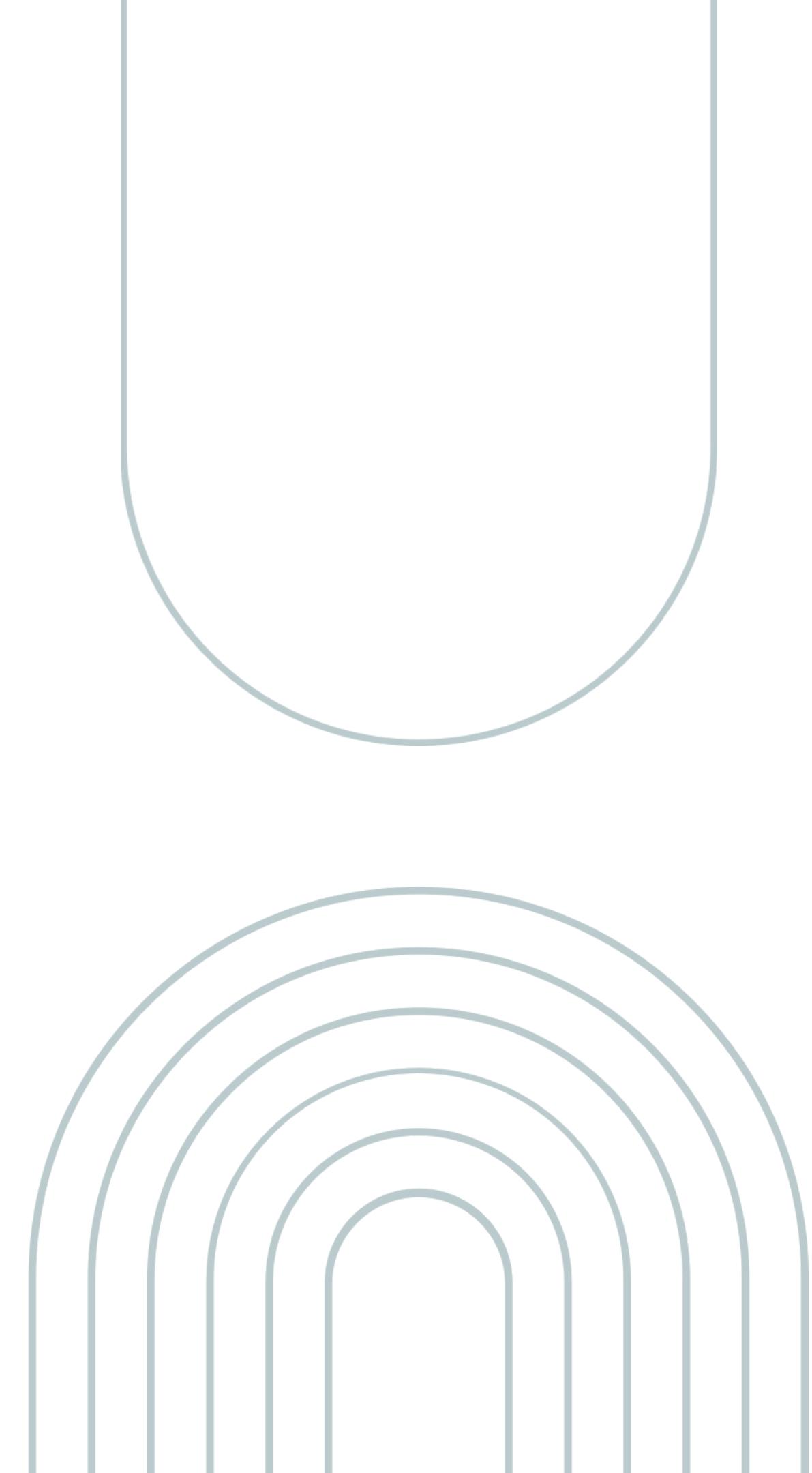
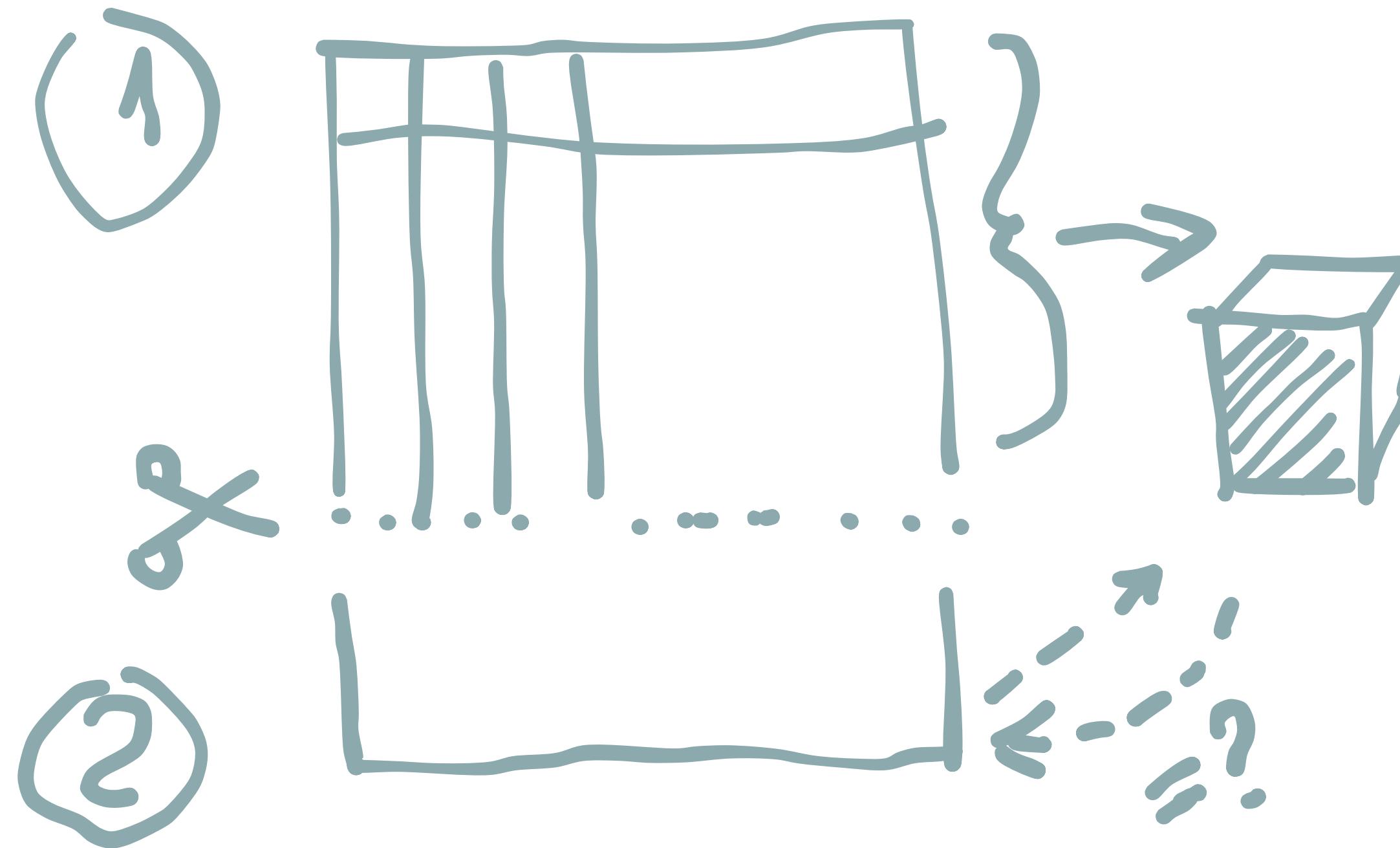
Разделим наши данные на
группы похожих объектов



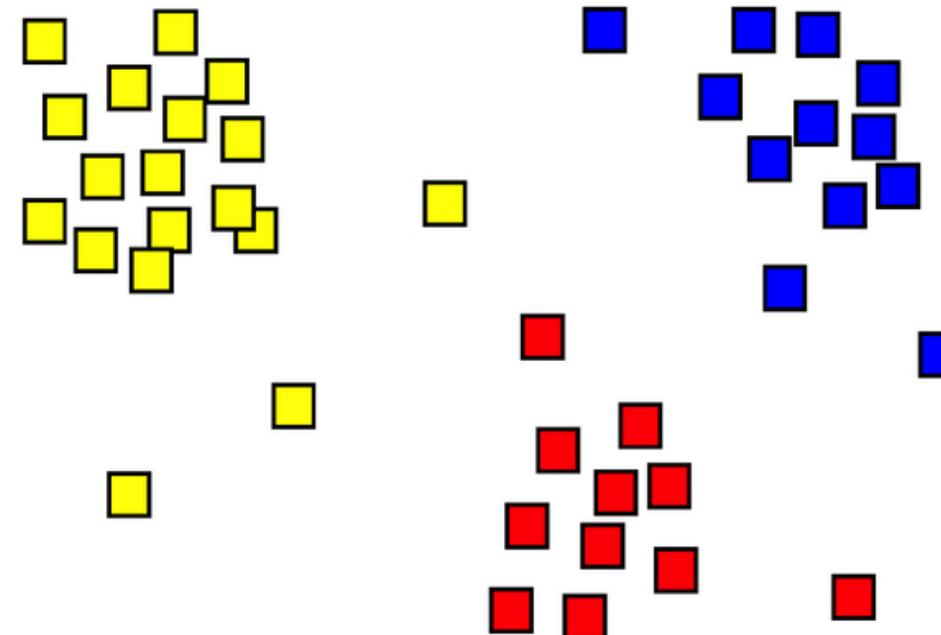
Регрессия

Предскажем точное значение
показателя в моменте
(похожая постановка
проблемы у Р. Фогеля –
контрфактическая модель
экономики и железных дорог,
экономики до Гражданской
войны в США)

ОБЩАЯ ИДЕЯ МАШИННОГО ОБУЧЕНИЯ



Кластеризация



	COKE	D_COKE	D_PEPSI	D_7UP	PEPSI	Sprite	TAB	SEVENUP
--	------	--------	---------	-------	-------	--------	-----	---------

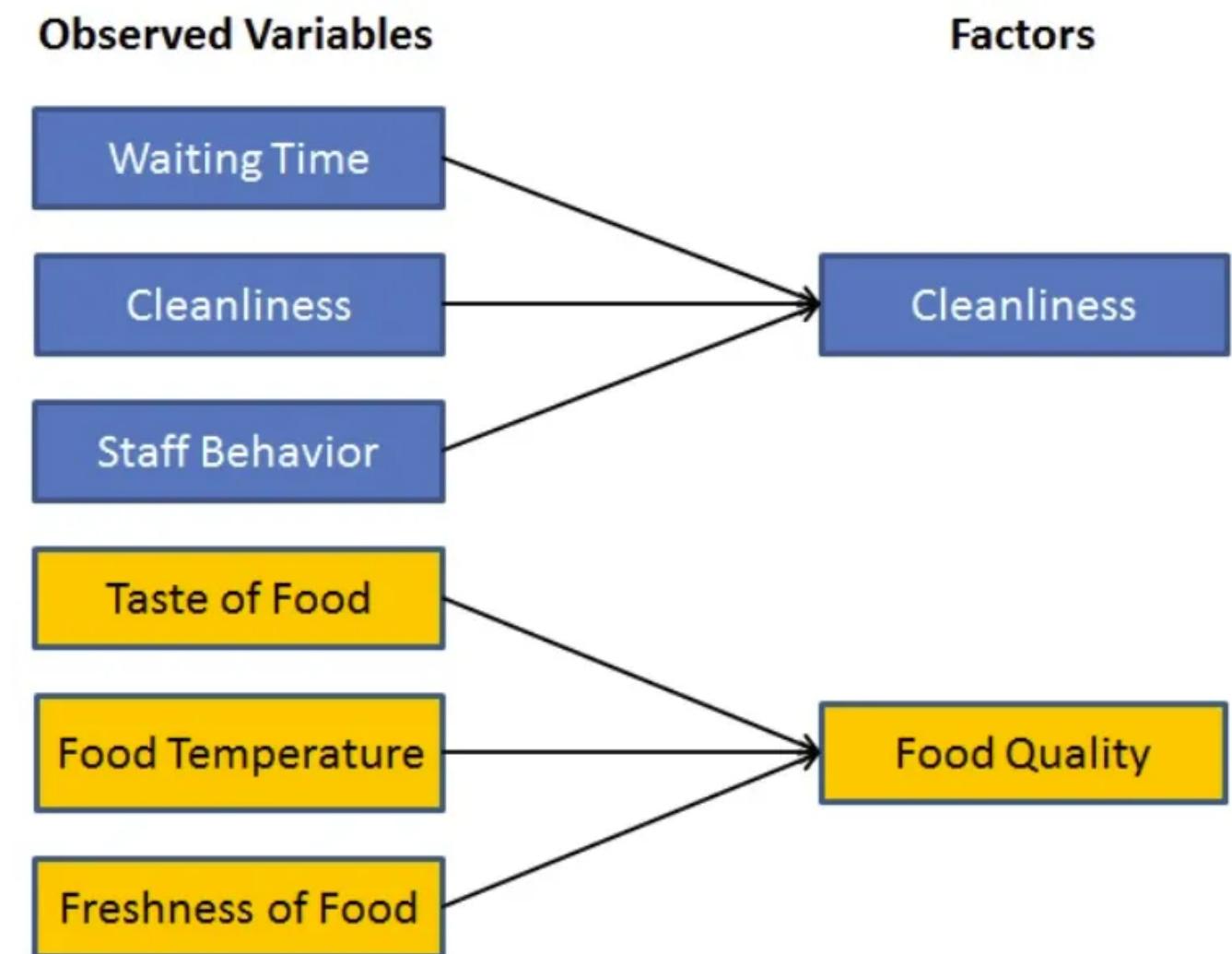
cluster

1	0.00	1.000000	0.545455	0.545455	0.000000	0.000000	0.909091	0.000000
2	1.00	0.272727	0.090909	0.000000	1.000000	0.000000	0.000000	0.272727
3	0.75	0.250000	0.083333	0.083333	0.416667	0.916667	0.083333	0.500000

ФАКТОРНЫЙ АНАЛИЗ

Суть: вместо разных переменных, описывающих наши данные, получить меньшее количество факторов (несуществующих факторов! факторы = представители наших данных)

Интересное: эти факторы отсутствуют в исходных данных, но порой можно найти новое и удивительное



Измерение неизмеримого?

- измерить силу любви
- измерить отношение пациентов к доктору
- удовлетворенность сортом кофе
- приверженность к курению
- лояльность торговой марке
- вероятность разорения фирмы...



	LUNGES	BITES	ZIGZAGS	NEST	SPINES	DNEST	BOUT
0	0.715035	0.956805	-0.086423	-0.241141	0.358615	-0.282219	0.166892
1	-0.019296	0.098533	0.367721	0.848657	0.073328	0.513705	-0.329849

- LUNGES Количество нападений (удары) на модель самца
- BITES Количество нападений (укусы) на модель самца
- ZIGZAGS Плавание зигзагом, которое является частью поведения, направленного на привлечение самок
- NEST Действия, связанные с построением гнезда
- SPINES Число раз, когда топорщился колючий верхний плавник
- DNEST Суммарная длительность времени, проведенного за построением гнезда
- BOUT Количество элементов поведения, характеризующих готовность к схватке

	LUNGES	BITES	ZIGZAGS	NEST	SPINES	DNEST	BOUT
0	0.715035	0.956805	-0.086423	-0.241141	0.358615	-0.282219	0.166892
1	-0.019296	0.098533	0.367721	0.848657	0.073328	0.513705	-0.329849

А можно измерить и для каждой рыбки

factor1 factor2

0 -0.938788 -0.776068

1 -0.159453 -0.591389

2 -0.818776 -0.458188

3 1.444012 0.094633

4 -0.638620 3.662073

5 0.397292 -0.037810

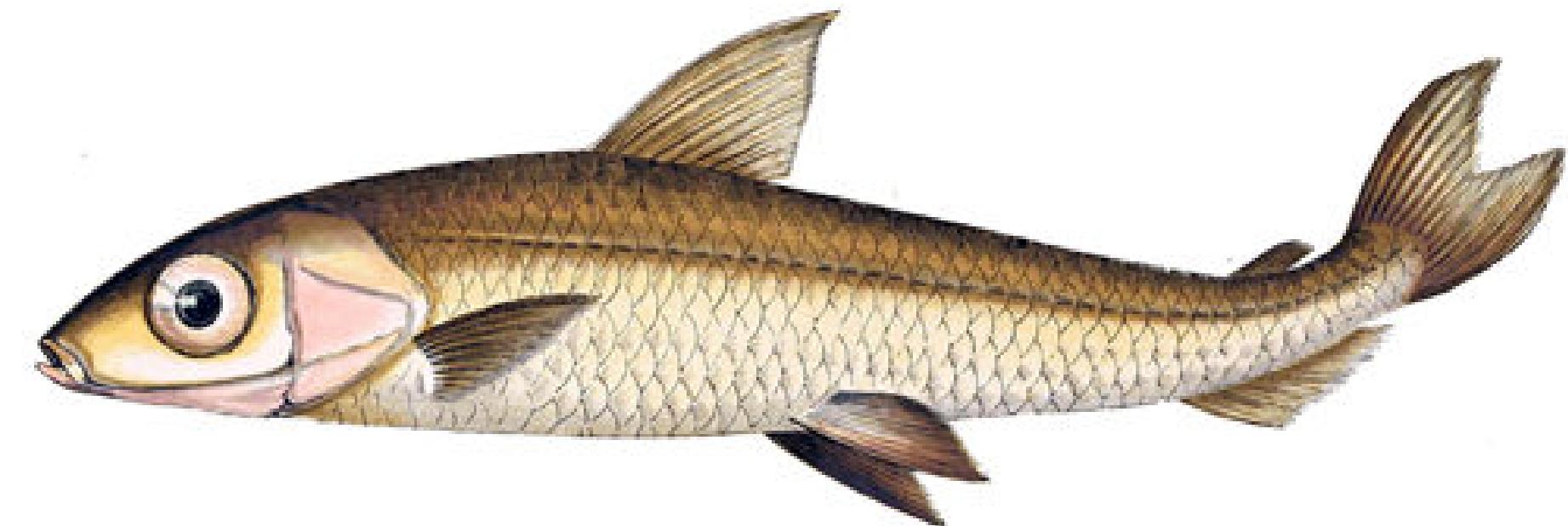
6 2.342290 0.179116

7 -0.151254 0.045436

8 1.072923 -0.406284

9 -0.908872 0.227997

10 -0.420181 -0.232168



LUNGES

BITES

ZIGZAGS

NEST

SPINES

DNEST

BOUT

0	0.715035	0.956805	-0.086423	-0.241141	0.356015	-0.282219	0.166692
1	-0.019296	0.098533	0.367721	0.848657	0.073328	0.513705	-0.329849