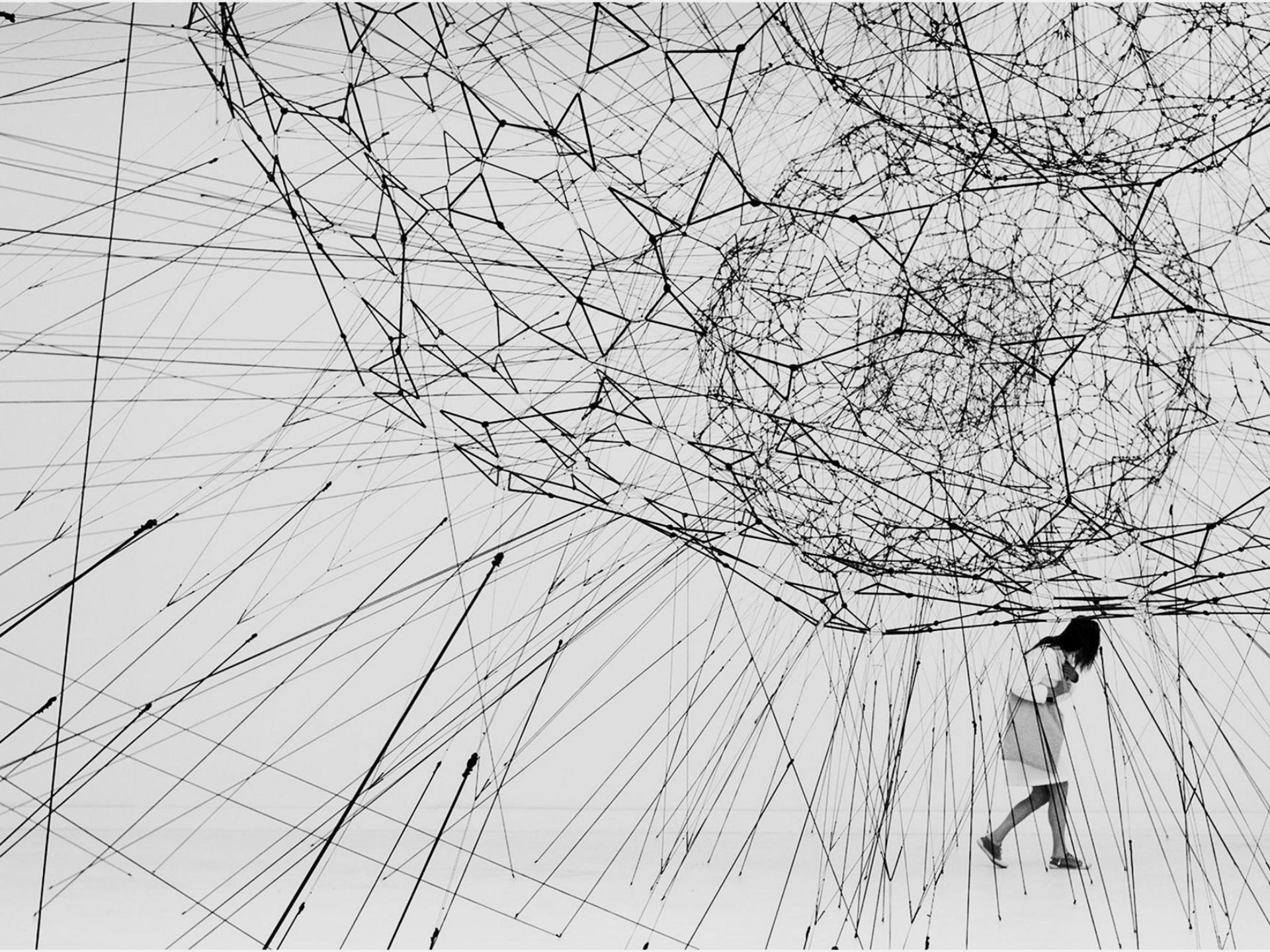


Data mining

Підготували:
Бортнік В.
Губенко М.
Кравчук О.
Кривонос А.
Пузир Д.
Серіков О.

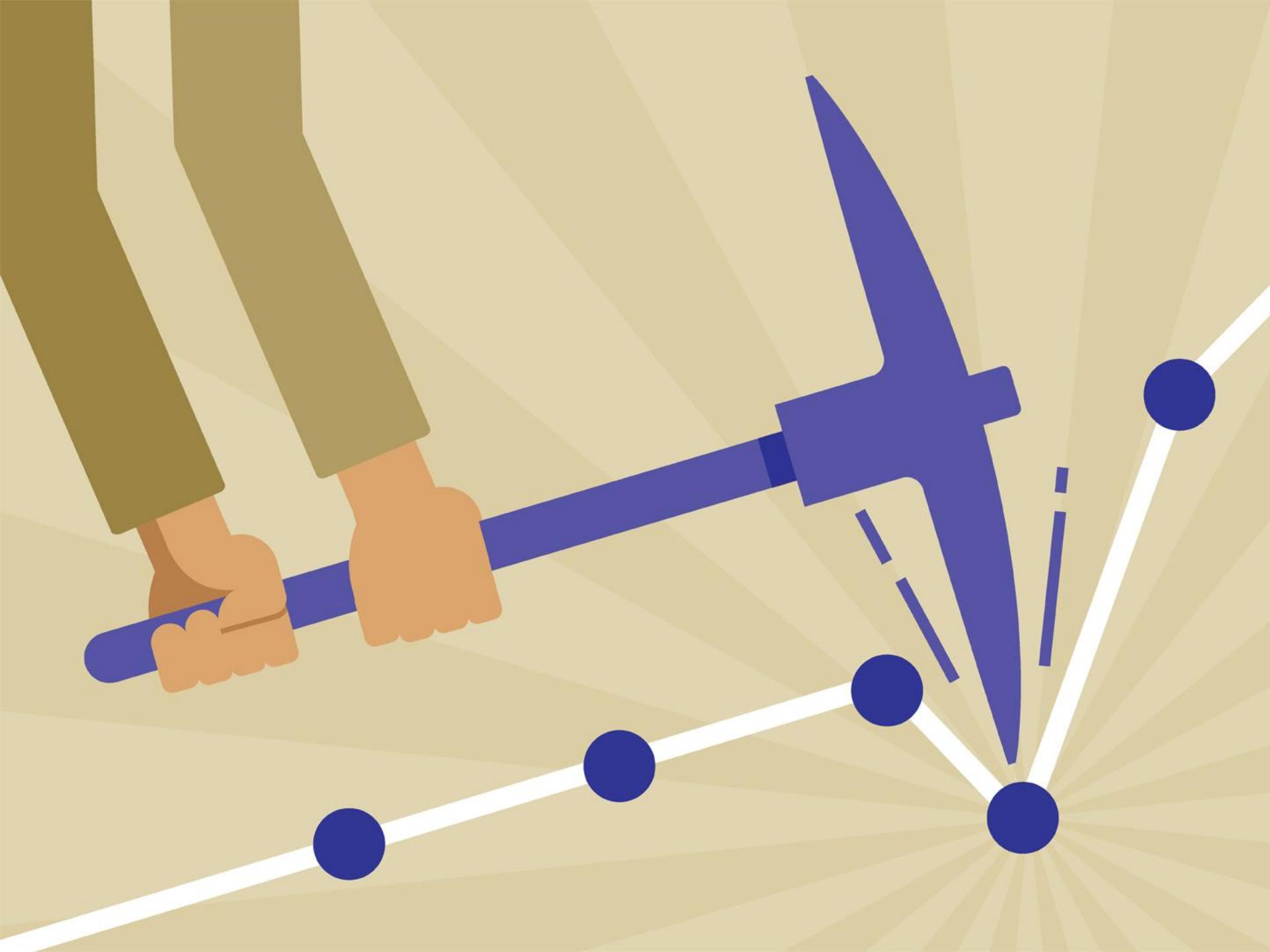


Data mining

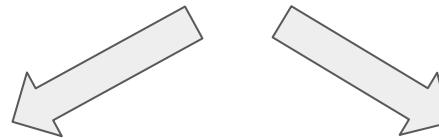
- Витяг, збір даних, видобуток даних (ще використовують Information Retrieval або IR);
- Витяг знань, інтелектуальний аналіз даних (Knowledge Data Discovery або KDD, Business Intelligence).

Завдання, які вирішуються Data Mining:

1. Класифікація
2. Кластеризація
3. Скорочення опису
4. Асоціація
5. Прогнозування
6. Аналіз відхилень
7. Візуалізація даних.



Типи даних, шкали



Просторові дані

Тимчасові ряди

| Вид: | Приклад: |
|--|--|
| Дані класифікації (номінальні) | Особи класифіковані за статтю, національністю |
| Ранжировані | Впорядкування регіонів за рейтингом |
| Дані вимірювання на інтервальній шкалі | Температура (шкала з довільною нульовою точкою і масштабом) |
| Дані вимірювання на відносній шкалі | Вимірювання ваги, висоти, об'єму (шкала з довільним масштабом, але фіксованою нульовою точкою) |

DATA ANALYSIS

STRUCTURE
QUALITY
WAYS
ONE
ANALYTICS
STATISTICS
FINDINGS
PEOPLE
ACCOUNT
CHECKED
DISCOVERY
ANALYSING
LEVEL
PERFORMED
TYPES
USING
LEVEL
PERFORMED
CHARACTERISTICS
TRANSFORMING
SUPPORTING
NORMALITY
PLOTS
USED
SPECIAL
SCATTER APPROACH
KURTOSIS
CONFIRMATORY STATISTICAL
BUSINESS INFORMATION

PHASE
MODELS
ASSESSED
APPLICATIONS
HYPOTHESES
INSTRUMENTS
CHECK
QUESTION
RESULTS
INTEGRATION
DESCRIPTIVE
DIFFERENT
ANALYSES
EITHER
RELIABLE
MAIN
STRUCTURE
QUALITY
WAYS
ONE
ANALYTICS
STATISTICS
FINDINGS
PEOPLE
ACCOUNT
CHECKED
DISCOVERY
ANALYSING
LEVEL
PERFORMED
TYPES
USING
LEVEL
PERFORMED
CHARACTERISTICS
TRANSFORMING
SUPPORTING
NORMALITY
PLOTS
USED
SPECIAL
SCATTER APPROACH
KURTOSIS
CONFIRMATORY STATISTICAL
BUSINESS INFORMATION

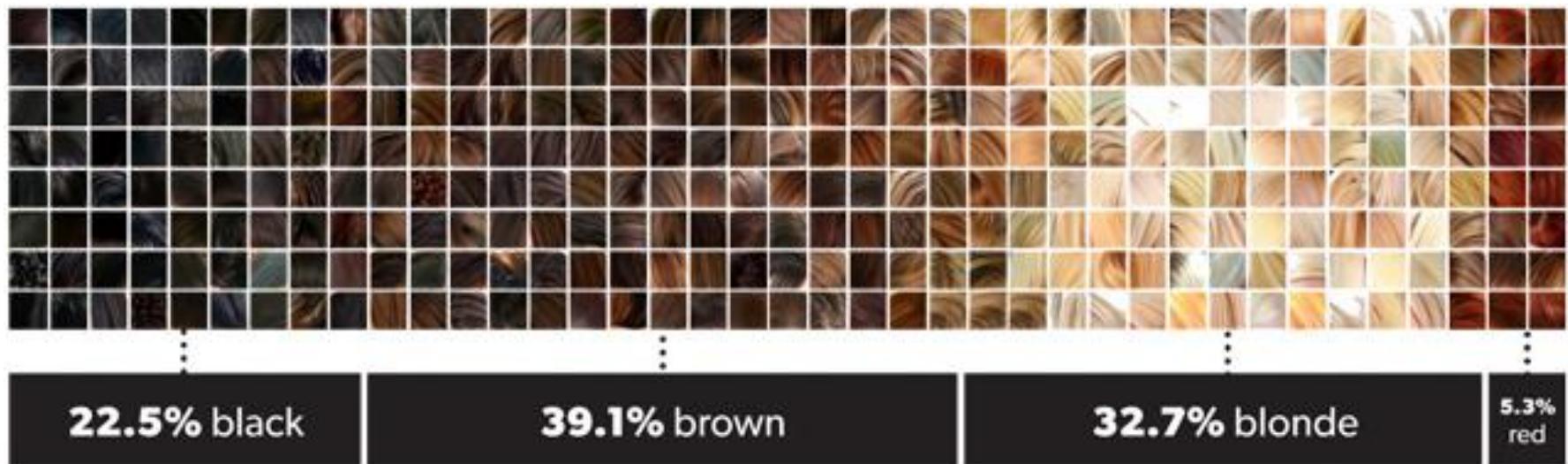
MEASUREMENT
FOCUSSES
CLOSELY
POSSIBLE
REPRODUCIBLE
RESEARCH
EXISTING
STRUCTURAL
TECHNIQUES
PLAN
TWO
EXPLORATORY
PLAN
EDA
CLEAR
HARD
LOOK
SKEWNESS
TAKE
NECESSARY
SUBGROUPS
USUALLY
ORIGINAL
FINAL
STAGE
DIVIDE
TEXTUAL
MODELING
PREDICTIVE
CDA
SAMPLE
MEDIAN
SEVERAL
SCIENCE
ADOPTED

Data Mining

...аналіз 10 000 акторів
фільмів для дорослих



Розподіл акторів за кольором волосся



Розподіл акторів за кольором шкіри

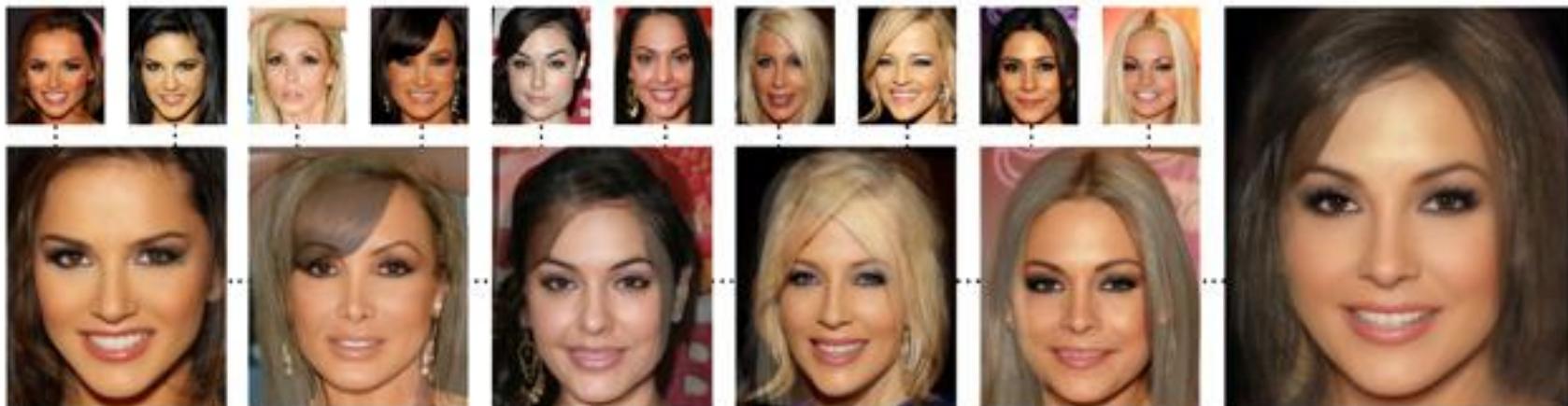


Розподіл акторів за наявністю татуювань



Морфінг 10 облич топ-10 актор_ecc

Facial morphs of 10 of the most popular adult performers



Tori Black &
Sunny Leone

Lisa Ann &
Nikki Benz

Sasha Grey &
Nina Mercedez

Alexis Texas &
Puma Swede

Raylene &
Jesse Jane

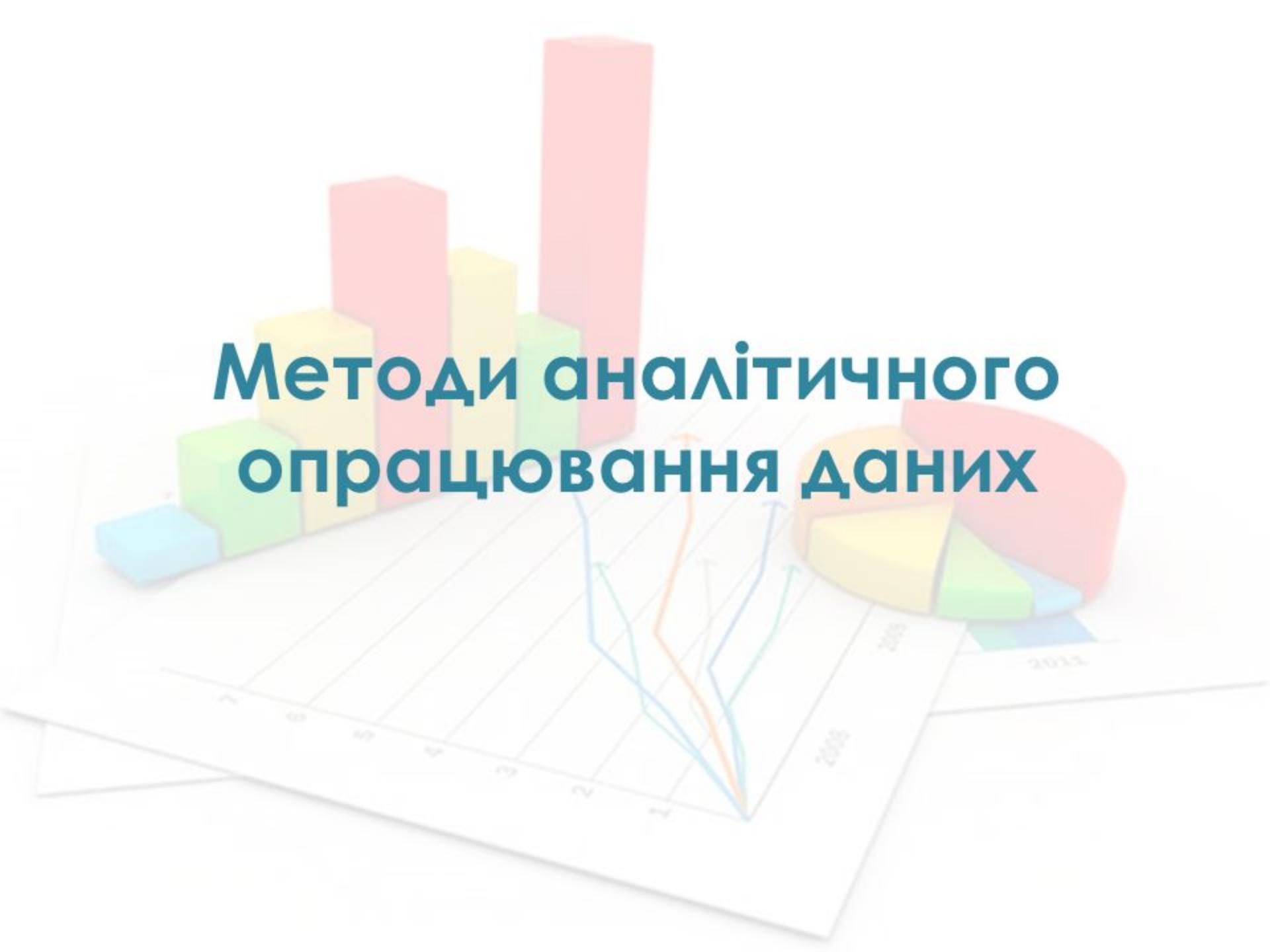
The Average Face of Ten
Top Female Pornstars



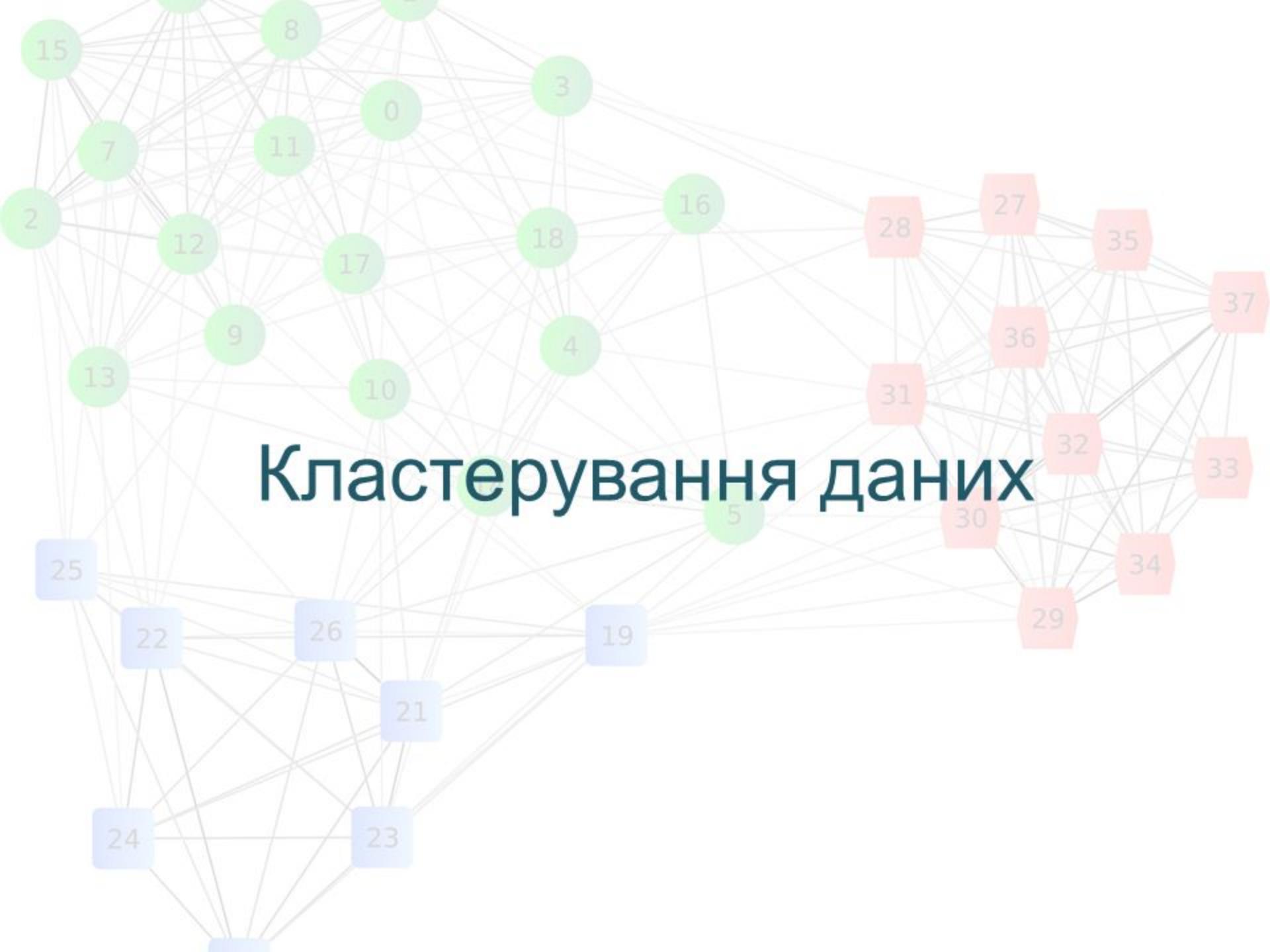
Посилання: jonmillward.com



Методи аналітичного опрацювання даних



Кластерування даних



Застосування кластерного аналізу в загальному вигляді зводиться до наступних етапів:

- Відбір вибірки об'єктів для кластеризації.
- Визначення безлічі змінних, за якими будуть оцінюватися об'єкти у вибірці. При необхідності - нормалізація значень змінних.
- Обчислення значень міри схожості між об'єктами.
- Застосування методу кластерного аналізу для створення груп схожих об'єктів (кластерів).
- Представлення результатів аналізу

Нормалізація

- Перед використанням алгоритмів кластеризації часто виклостовують нормалізацію, щоб всі компоненти давали одинаковий вклад при розрахунку «відстані».
- У процесі нормалізації всі значення приводяться до деякого діапазону, наприклад, [-1, -1] або [0, 1]
- Наприклад міні-макс нормалізація:

$$x' = (x - \text{MIN}[X]) / (\text{MAX}[X] - \text{MIN}[X])$$

Вимірювання відстані

- Евклідова відстань

$$\rho(x, x') = \sqrt{\sum_i^n (x_i - x'_i)^2}$$

- Квадрат евклідової відстані

$$\rho(x, x') = \sum_i^n (x_i - x'_i)^2$$

- Відстань між міськими кварталами (Манхеттенська відстань)

$$\rho(x, x') = \sum_i^n |x_i - x'_i|$$

- Відстань Чебишева

$$\rho(x, x') = \max(|x_i - x'_i|)$$

- Степеннева відстань

$$\rho(x, x') = \sqrt[p]{\sum_i^n (x_i - x'_i)^p}$$

Алгоритми кластеризації умовно можна розділити на ієрархічні та плоскі.

- Ієрархічні алгоритми (також називають алгоритмами таксономії) будують систему вкладених розбиттів. Тобто на виході ми отримуємо дерево кластерів, коренем якого є вся вибірка, а листям - найбільш дрібні кластери.
- Плоскі алгоритми будують одне розбиття об'єктів на кластери.

Метод к-середніх

Метод к-середніх створює k-груп з набору об'єктів таким чином, щоб члени групи були найбільш однорідними. Це популярна техніка кластерного аналізу для дослідження набору даних.

Вхідні данні: число кластерів.

Як працює метод к-середніх?

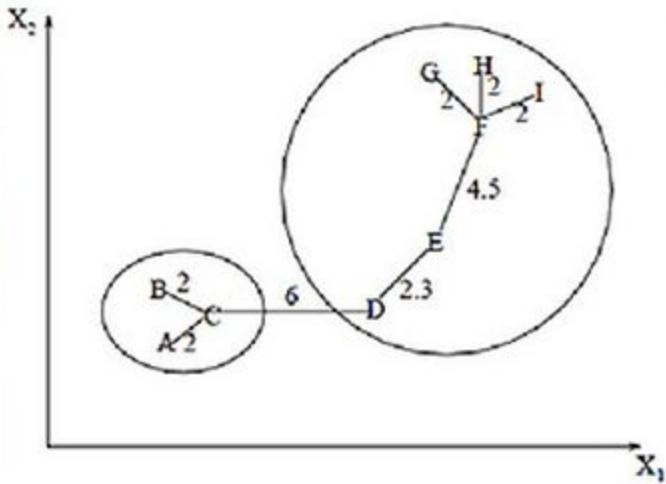
- Метод к-середніх вибирає точки багатовимірного простору, які будуть представляти к-кластери. Ці точки називаються центрами тяжіння. Перший раз, за відсутності припущень, центри тяжіння можна вибирати випадково
- Кожен пацієнт буде розташовуватися найближче до однієї з точок.
- Тепер у нас є к-кластерів, і кожна точка - це член якогось з них.
- Метод к-середніх, враховуючи положення членів кластера, знаходить центр кожного з k-кластерів. Обчислений центр стає новим центром тяжіння кластера.
- Оскільки центр ваги перемістився, точки могли виявитися більше до інших центрів тяжіння. Іншими словами, вони могли змінити членство.
- Кроки 2-6 повторюються до тих пір, поки центр ваги не перестануть змінюватися і членство не стабілізується. Це називається збіжністю.

Реалізації методу к-середніх

- Apache Mahout
- Julia
- R
- SciPy
- Weka
- MATLAB
- SAS

Алгоритм мінімального покривачого дерева

Алгоритм мінімального покриває дерево спочатку буде на графі мінімальне покриває дерево, а потім послідовно видаляє ребра з найбільшою вагою. На малюнку зображено мінімальне покриває дерево, отримане для дев'яти об'єктів.



Також для кластеризації використовують наступні алгоритми:

- с-средніх
- Мінімальне покриваюче дерево
- Пошарова кластеризація
- C4.5
- Метод опорних векторів
- Apriori
- EM-алгоритм
- PageRank
- AdaBoost
- k-найближчих сусідів

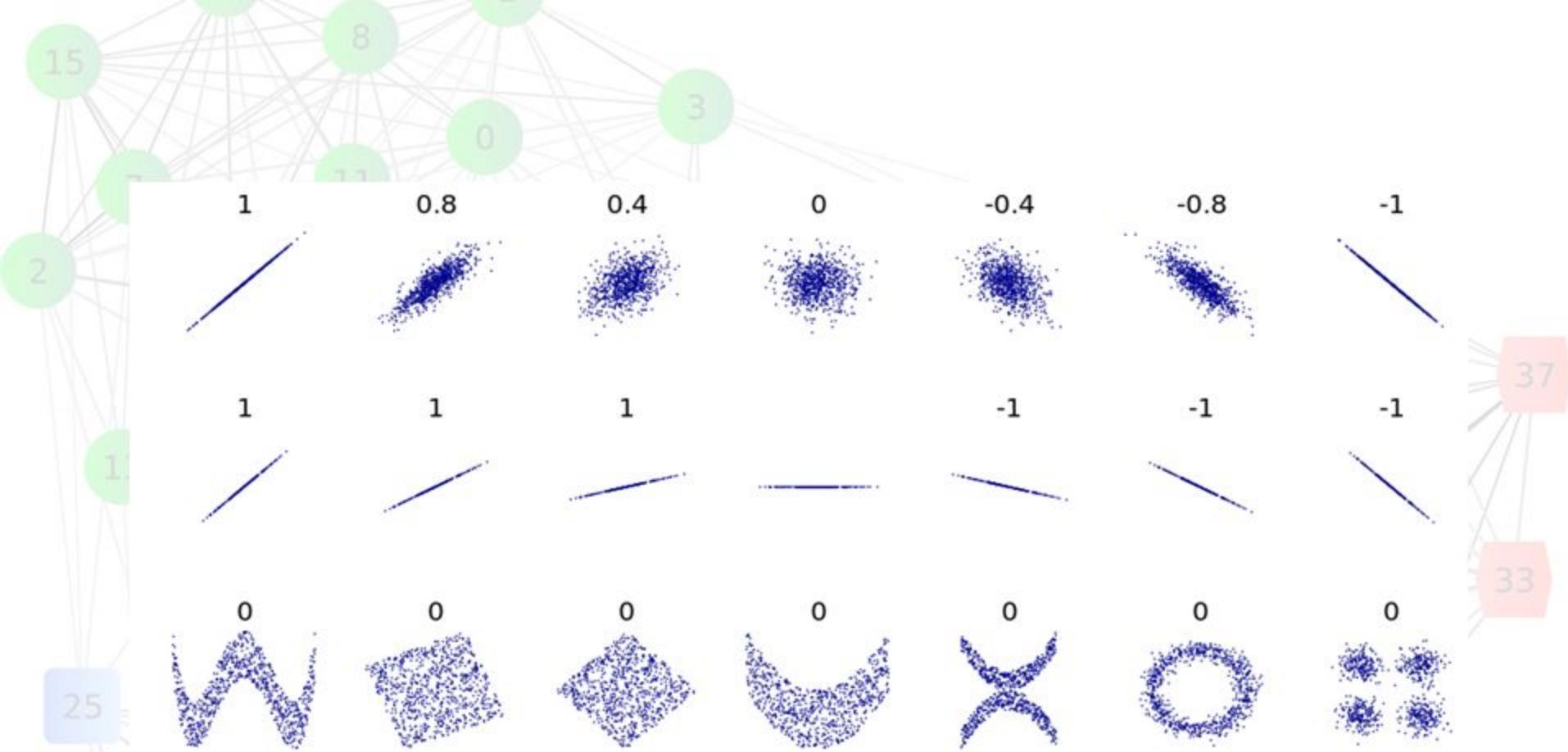
Порівняння деяких алгоритмів кластеризації

| Алгоритм кластеризації | Обчислювальна складність |
|------------------------------|--|
| Ієрархічний | $O(n^2)$ |
| k-средніх | $O(nkl)$, где k – число кластерів, l – число ітерацій |
| c-средніх | |
| Мінімальне покриваюче дерево | $O(n^2 \log n)$ |
| Пошарова кластеризація | $O(\max(n, m))$, где $m < n(n-1)/2$ |

| Алгоритм кластеризації | Форма кластерів | Вхідні дані | Результати |
|------------------------------|-----------------|---|--|
| Ієрархічний | Довільна | Число кластерів или порог відстані для усічення ієархії | Бінарне дерево кластерів |
| k-средніх | Гіперсфера | Число кластерів | Центри кластерів |
| c-средніх | Гіперсфера | Число кластерів, степень нечіткості | Центри кластерів, матриця приналежності |
| Виділення зв'них компонент | Довільна | Порог відстані R | Древоподібна структура кластерів |
| Мінімальне покриваюче дерево | Довільна | Число кластерів ичи порог відстані для видалення ребер | Древоподібна структура кластерів |
| Пошарова кластеризація | Довільна | Полідовність границь відстані | Древоподібна структура кластерів з різними рівнями |

Статистичні методи аналізу даних. Кореляційний аналіз

- Кореляційний аналіз - метод обробки статистичних даних, що полягає у вивченні коефіцієнтів (кореляції).
- При цьому порівнюються коефіцієнти кореляції між однією парою або великою кількістю пар ознак, для встановлення між ними статистичних взаємозв'язків.



Декілька наборів точок (x, y), над кожним з яких вказано коефіцієнт кореляції Пірсона величин x і y

З теорії ймовіості:

Для системи з двох неперервних випадкових величин (X, Y) єснує поняння коваріації або кореряційного моменту):

$$K_{xy} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - m_x)(y - m_y) f(x, y) dx dy.$$

Де $f(x,y)$ -функція густини розподулі вірогідності

Для характеристики зв'язку між величинами(X, Y) вводять наступну величну:

$$r_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y},$$

Для дискретних величин кореляційний момент можна знайти наступним чином:

$$K_{xy} = \sum_{i=1}^n \sum_{j=1}^n (x_i - m_x)(y_j - m_y)p_{ij}$$
$$= \frac{1}{n} \sum_i^n \sum_j^n (x_i - m_x)(y_j - m_y)$$

Також на практиці зазвичай використовують іншу формалу, яка дає менш точні результати, але потребує менше обчислень:

$$K_{xy} = \frac{1}{n} \sum_{i=0}^n (x_i y_i - m_x m_y)$$

Мат. очікування та дисперсія обчислюються за наступними формулами:

$$m_x = \frac{1}{n} \sum_{i=0}^n x_i$$

$$D_x = \frac{1}{n} \sum_{i=0}^n x_i^2 - m_x^2, \sigma_x = \sqrt{D_x}$$

$$r_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y}$$

Якщо даний коефіцієнт рівний нулю, то величини незалежні між собою.

1 - абсолютно залежні

-1 також залежні, але збільшення X призводить до зменшення Y і навпаки.

Data Mining

...практична перевірка
теорії шести рукостискань



В чому полягає теорія шести рукостискань?

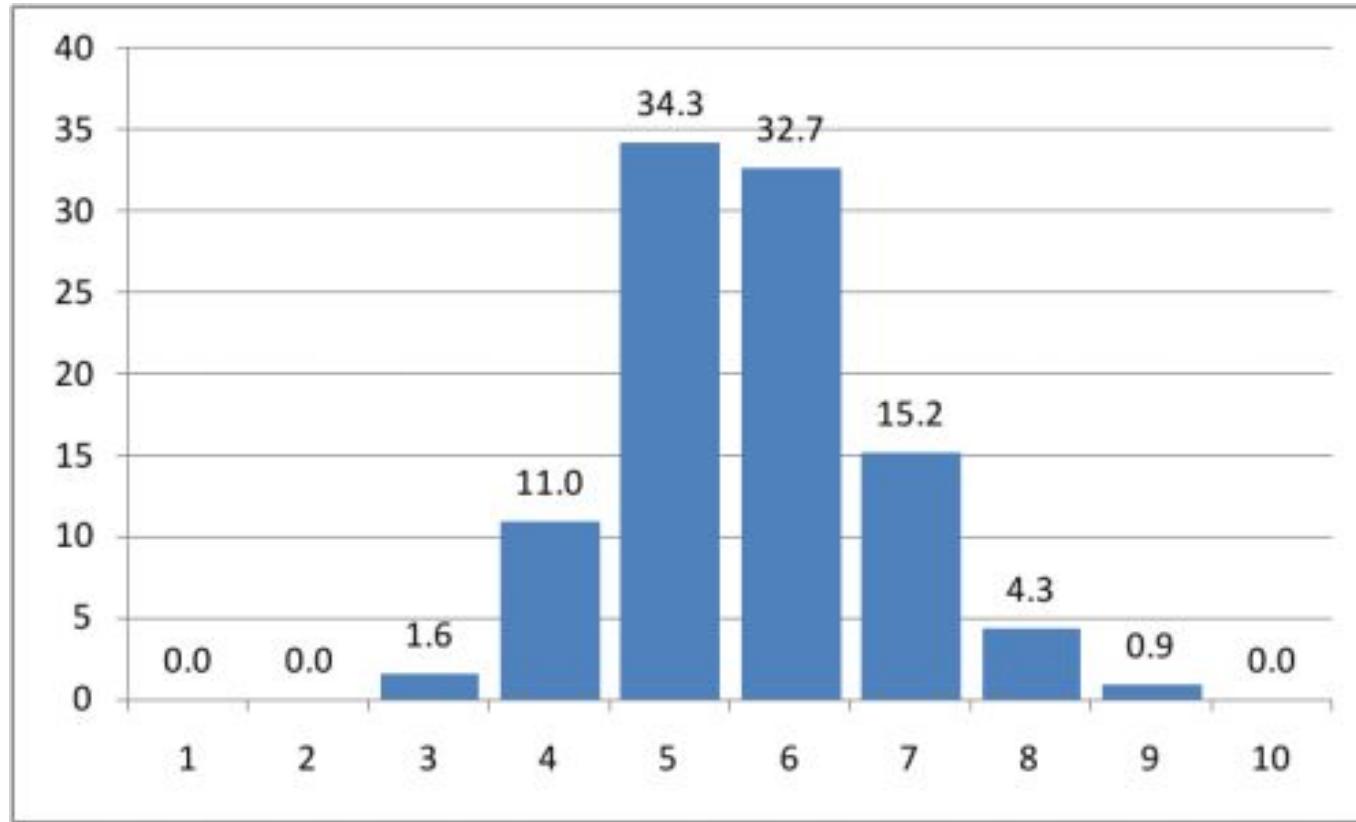
Кожна людина на Землі знайома з будь-якою іншою через ланцюжок з п'яти друзів, тобто, через шість рукостискань



Теория шести рукопожатий

sakson.lit-dety.ru

Результати проведеного дослідження



По осі X - довжина найкоротшого ланцюжка друзів,
по осі Y - ймовірність її знайти

Посилання: habr.com/post/132558/



Data Mining

Інструменти



Python як основний “шахтарський” інструмент

- опенсорсний
- простий у використанні
- велика спільнота
- легко освоїти нові бібліотеки
- код, зрозумілий навіть “непосвящонним”

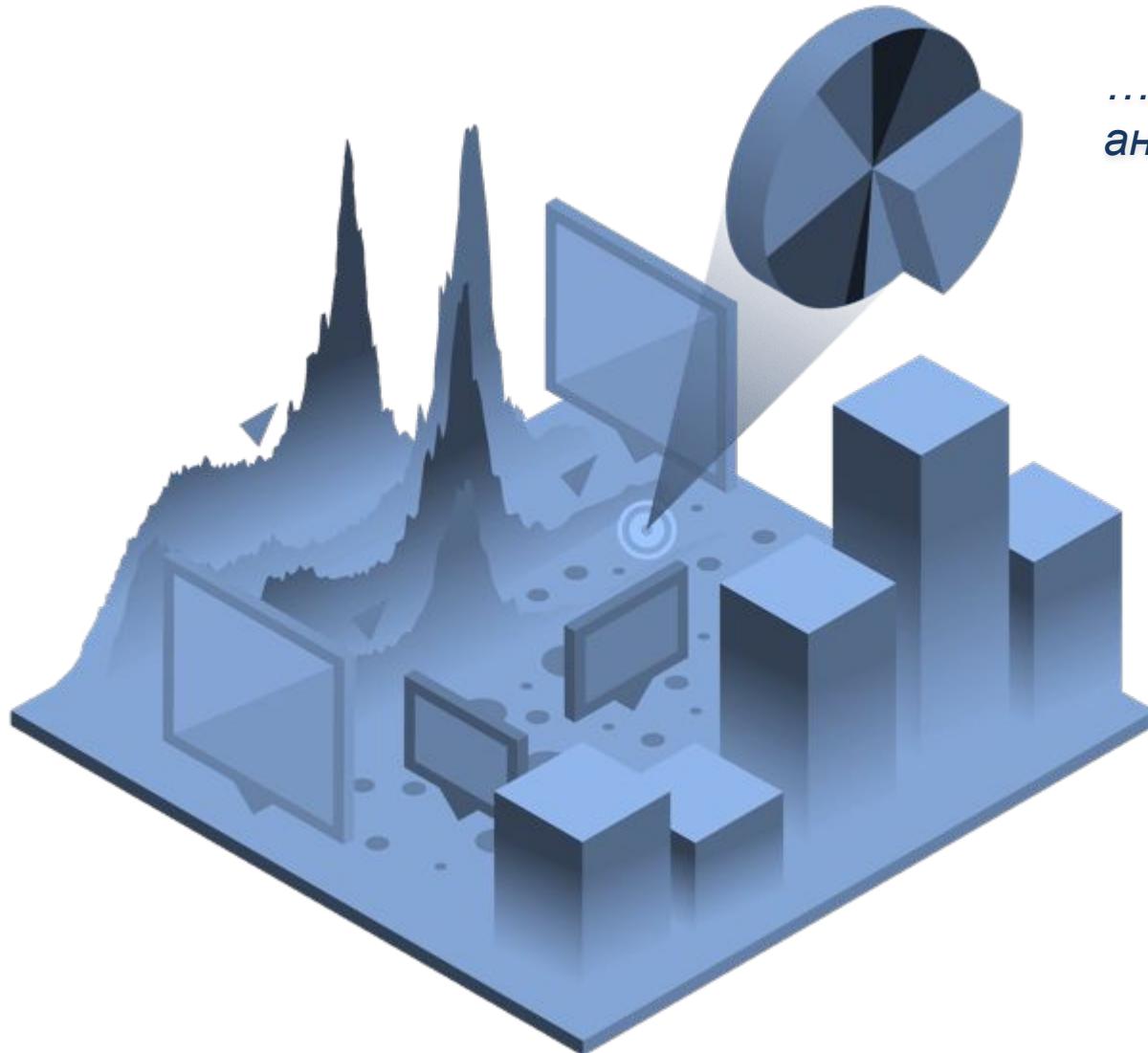


Основні бібліотеки - “спорядження”

- ScraPy - власне, сама кирка, приціл до неї та перемикач на режим “автомат”
- Pandas - вагонетка
- NumPy - все ще вагонетка
- Matplotlib - каменерізний станок

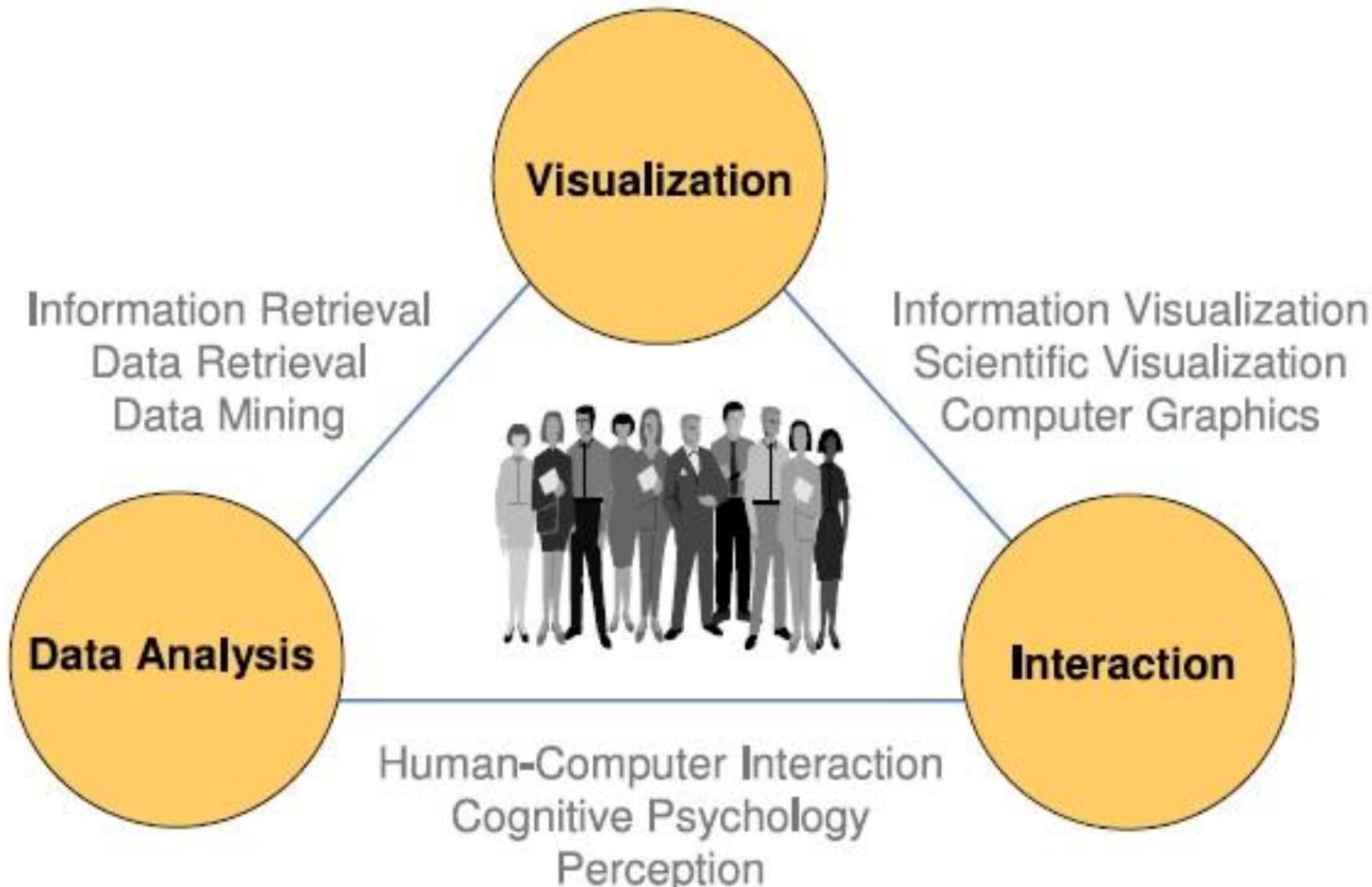


Visual Mining



*...а также тех.
анализ*

Визуальный анализ



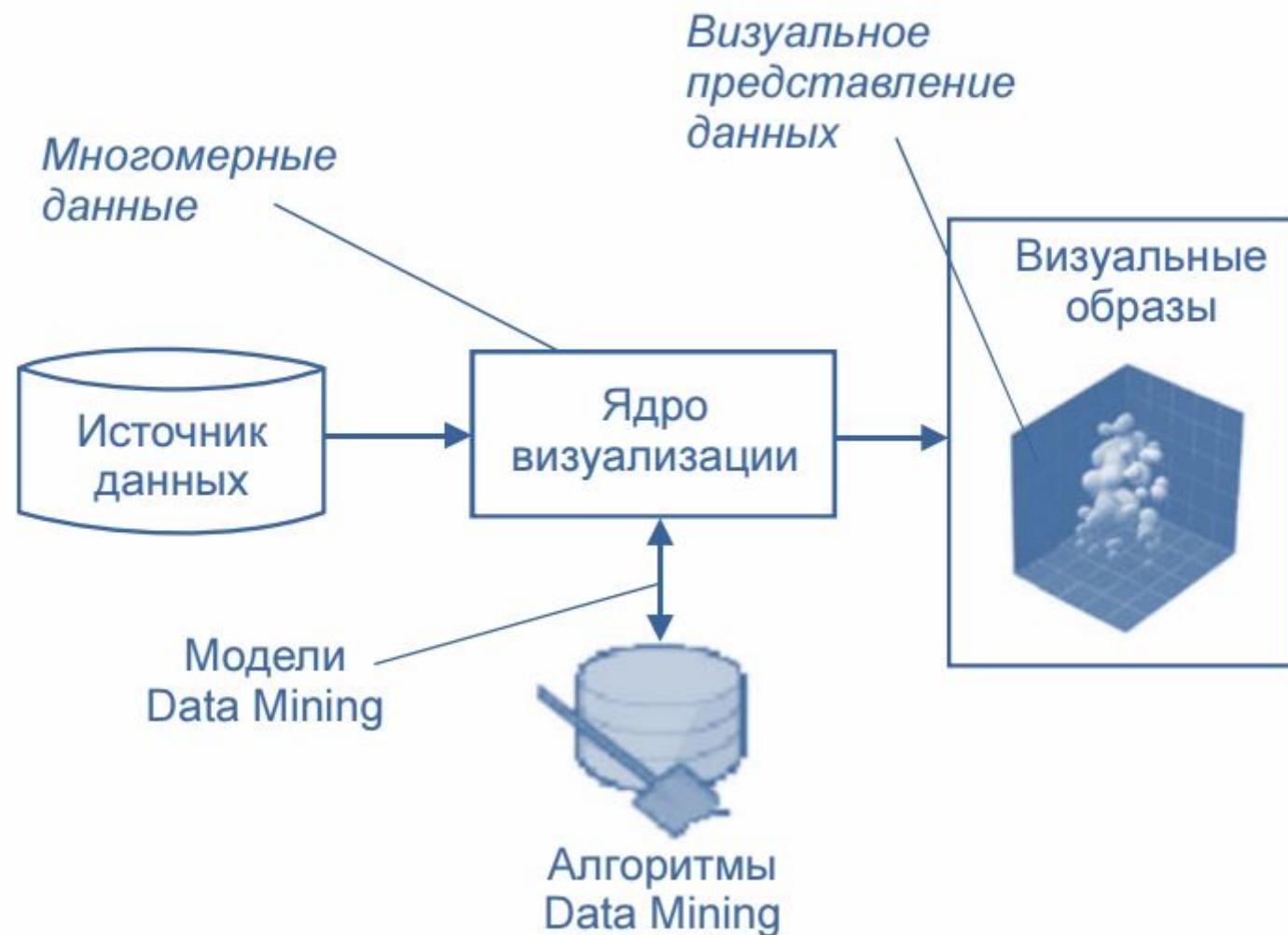
Почему же всё-таки Visual Mining?

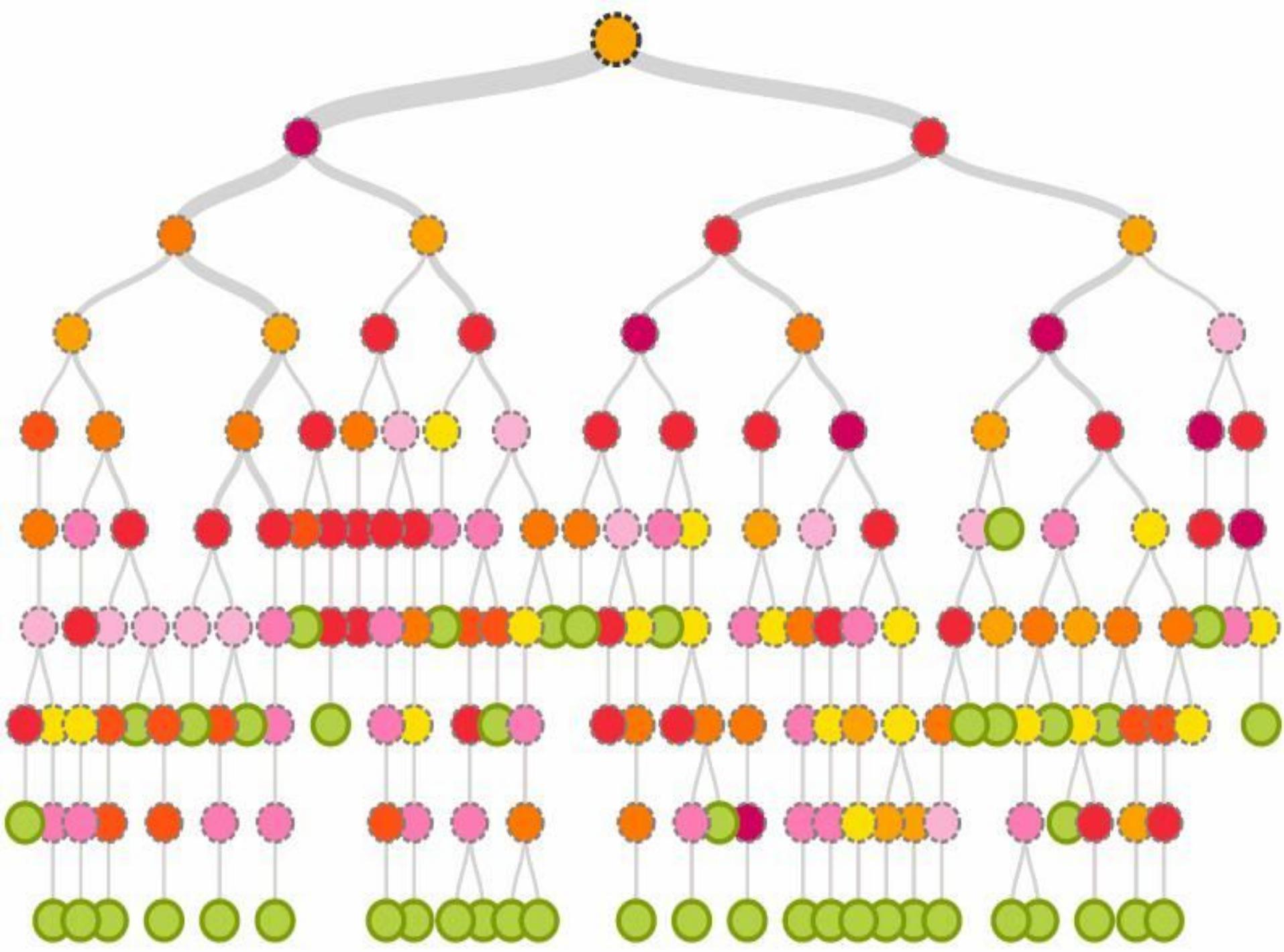
Преимущества:

- Гибкость человеческого мышления;
- Более развитый причинно-следственный анализ;
- Обширная база знаний.

| | Data Mining | Visual Mining |
|-------------------------|-------------|---------------|
| Действенность | + | - |
| Качественная оценка | + | - |
| Гибкость | - | + |
| Вовлечение пользователя | - | + |

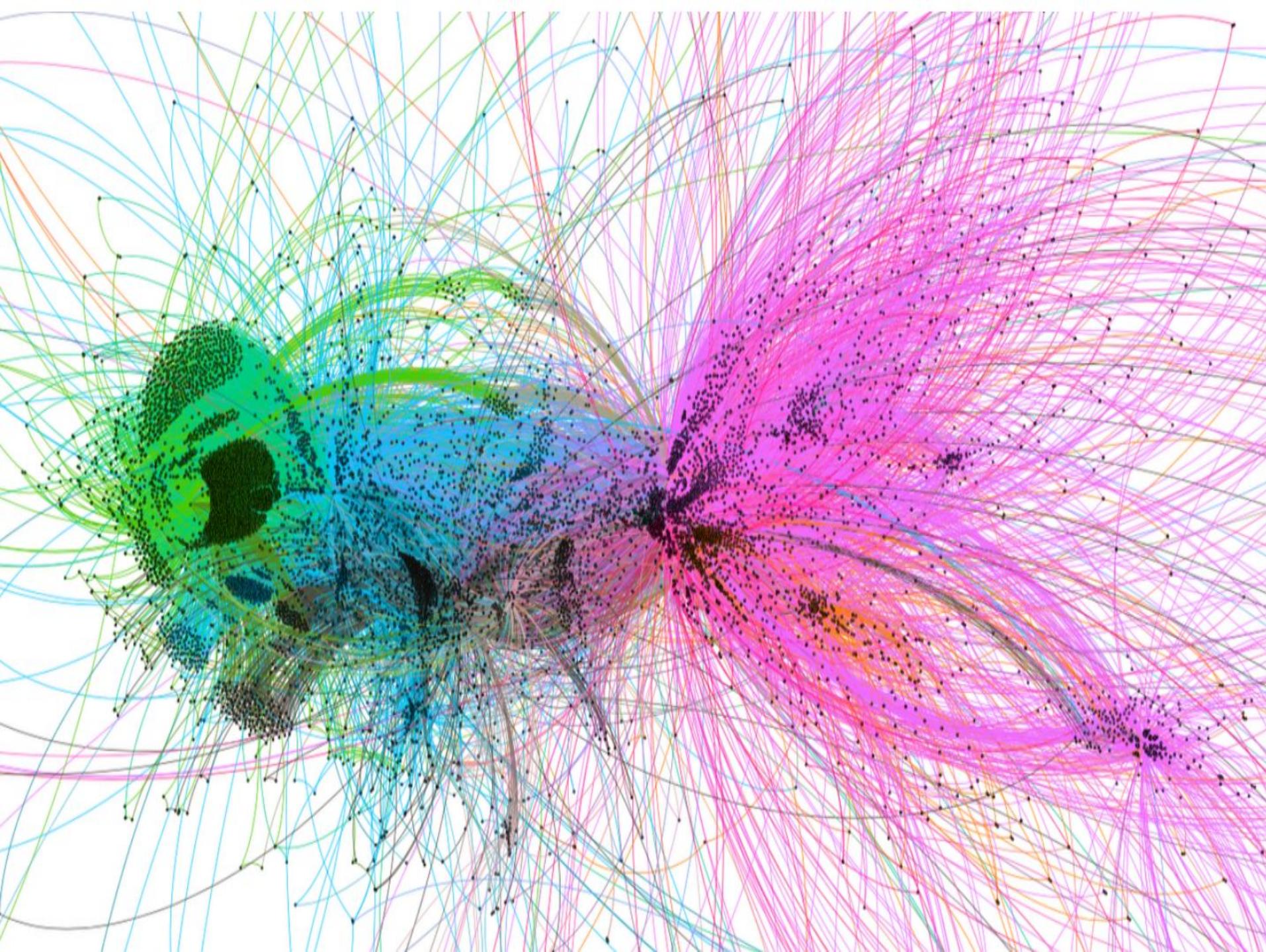
Методы визуального анализа

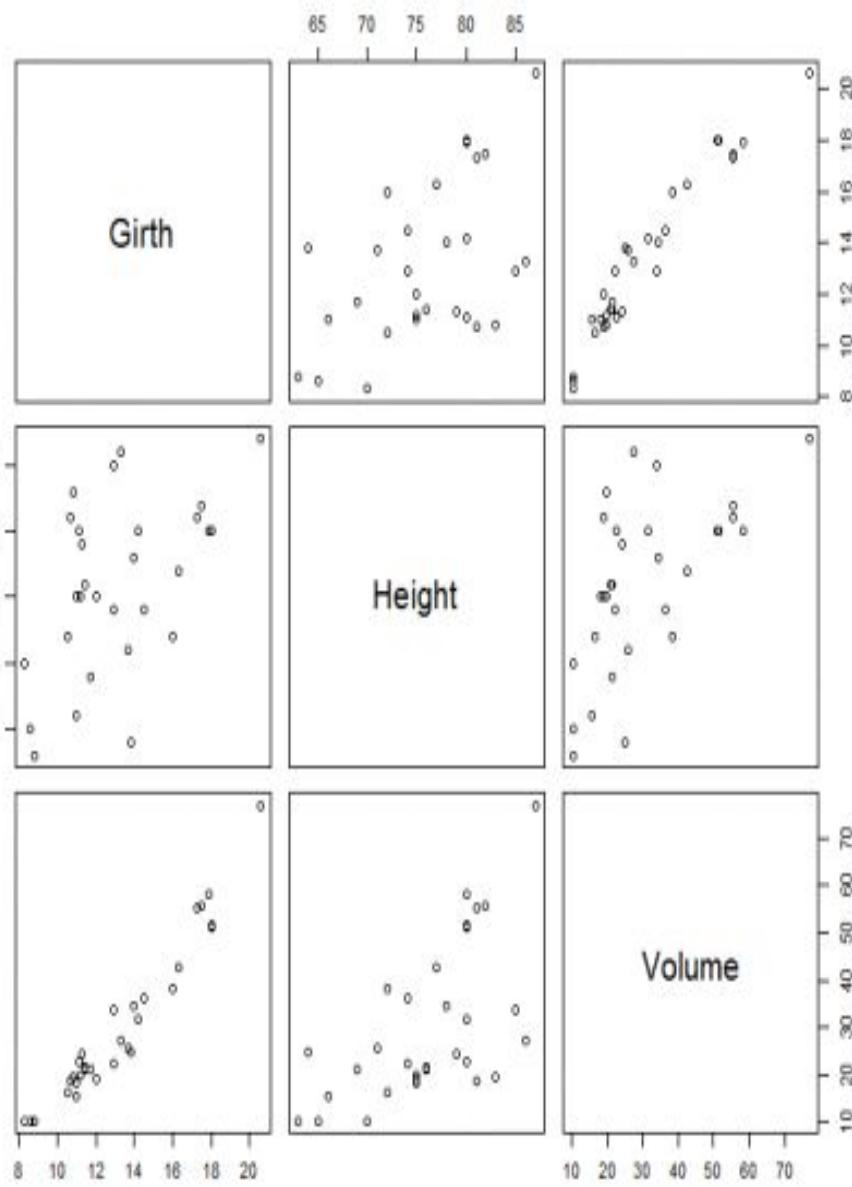
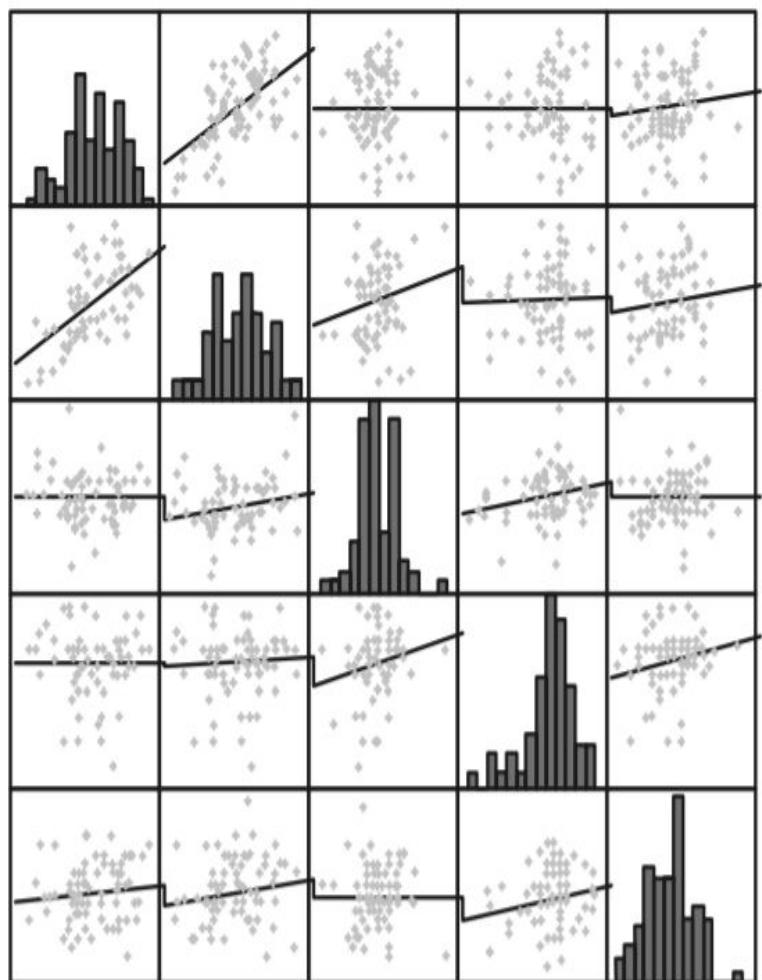


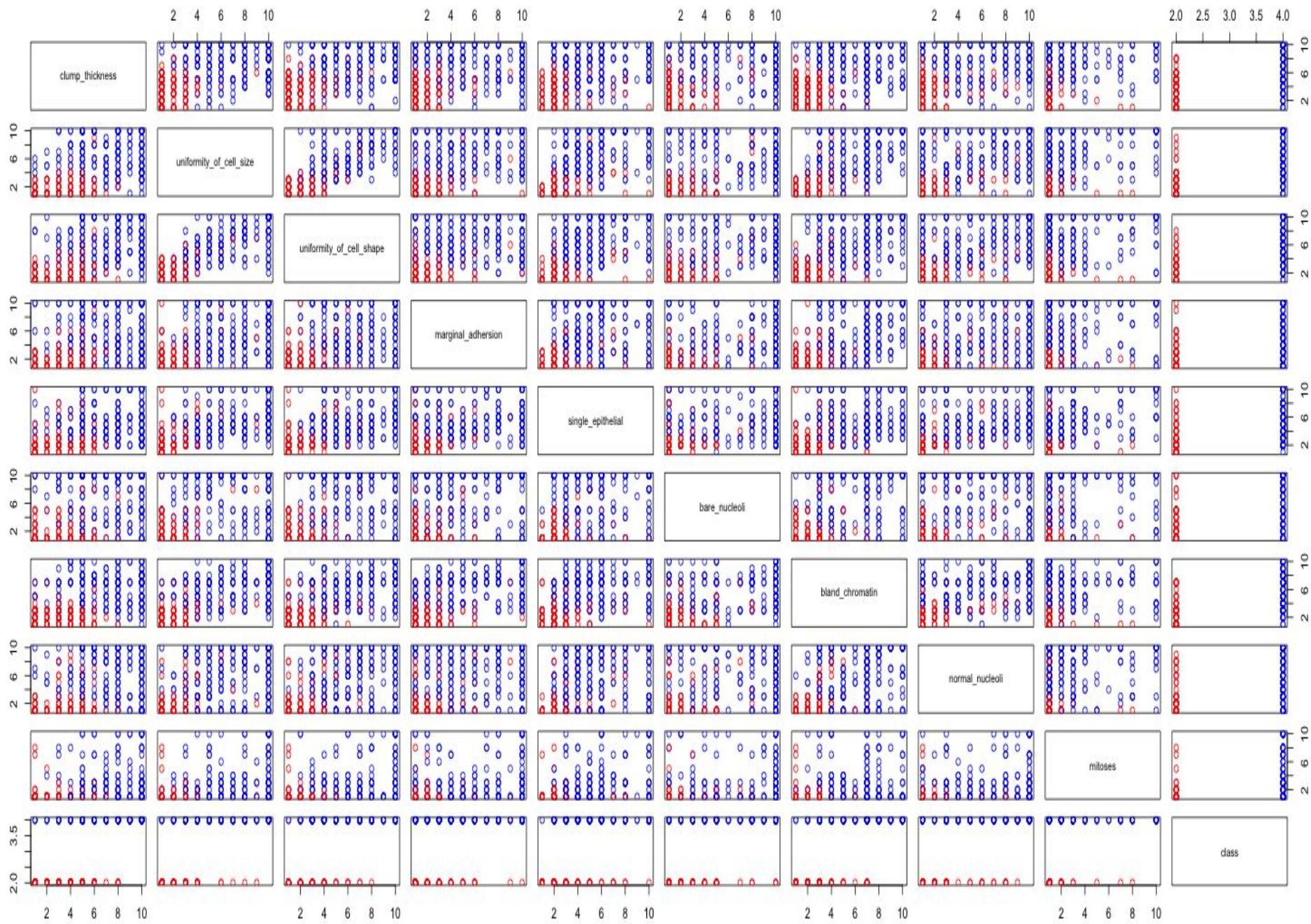


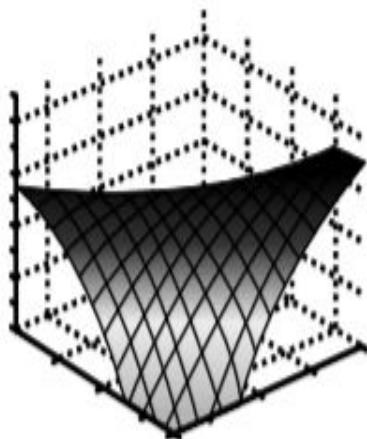
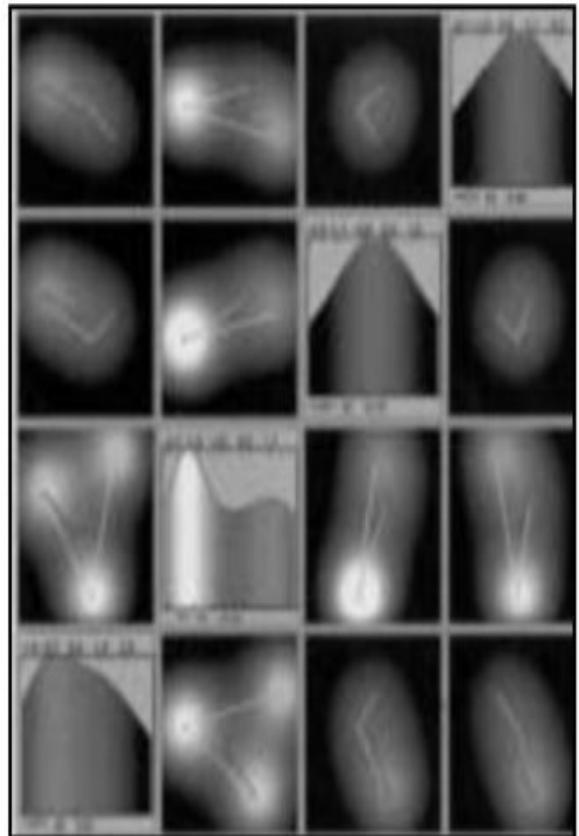
| 0 | 4 | 8 | 16 | 19 | 24 | 32 |
|-------------|-------|----------------------|--------|-----------------------------|----|----|
| Версия | Длина | Тип сервиса | | Общая длина | | |
| | | Идентификация | Флаги | Смещение фрагмента | | |
| Время жизни | | Протокол | | Контрольная сумма заголовка | | |
| | | IP-адрес отправителя | | | | |
| | | IP-адрес получателя | | | | |
| | | Опции IP | | Заполнение | | |
| | | | Данные | | | |

Рис. 5.16. Формат дейтаграммы

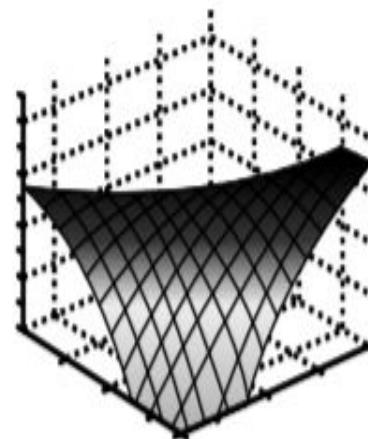




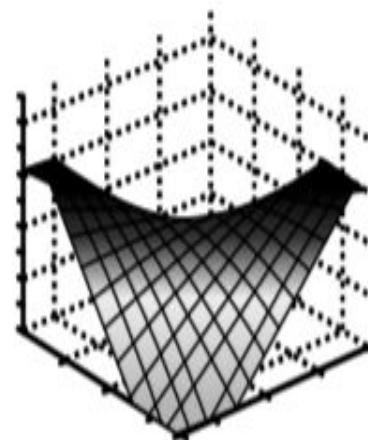
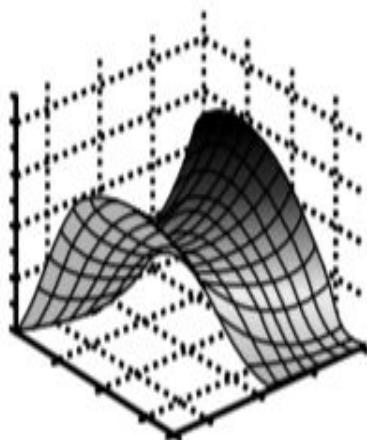


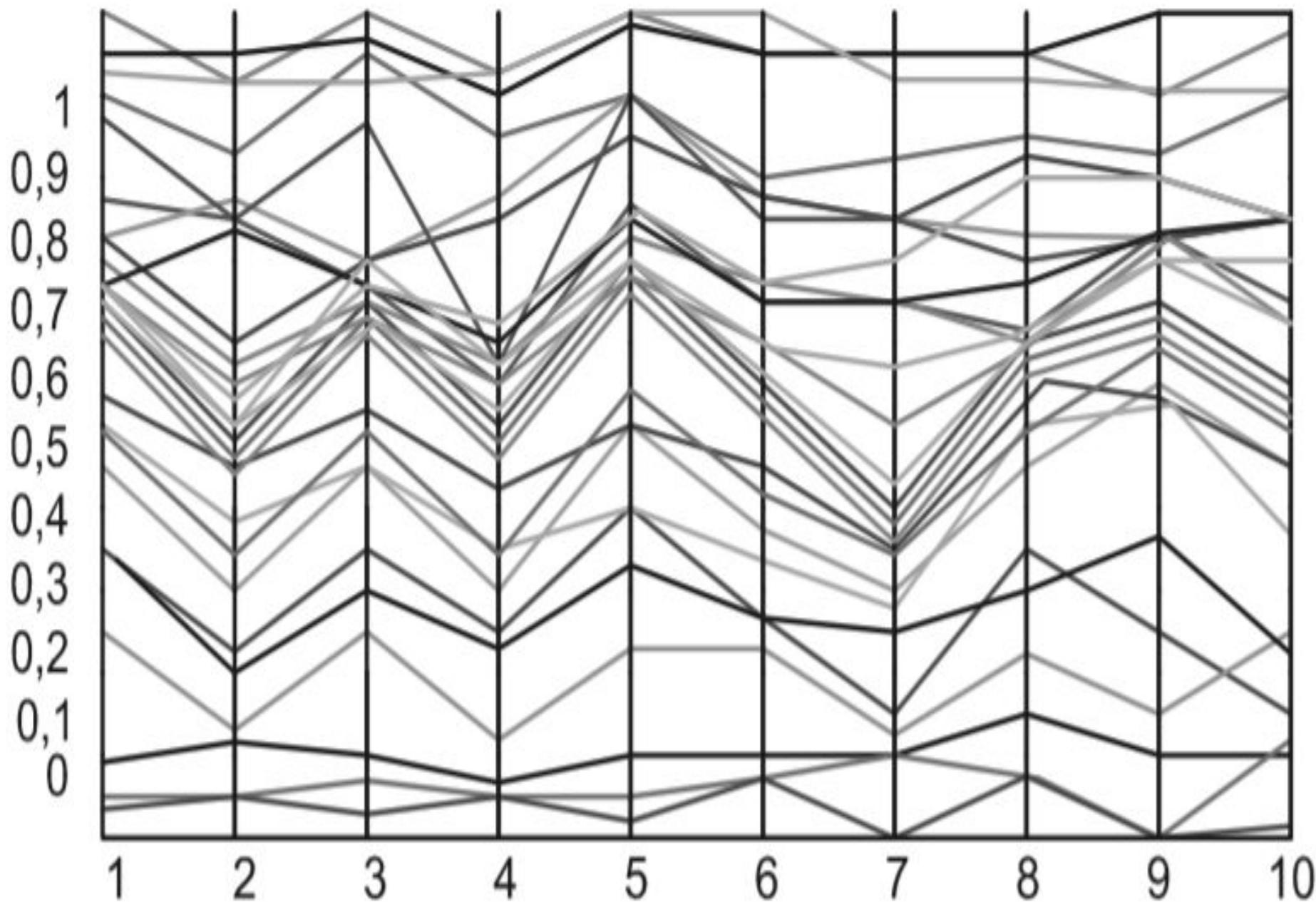


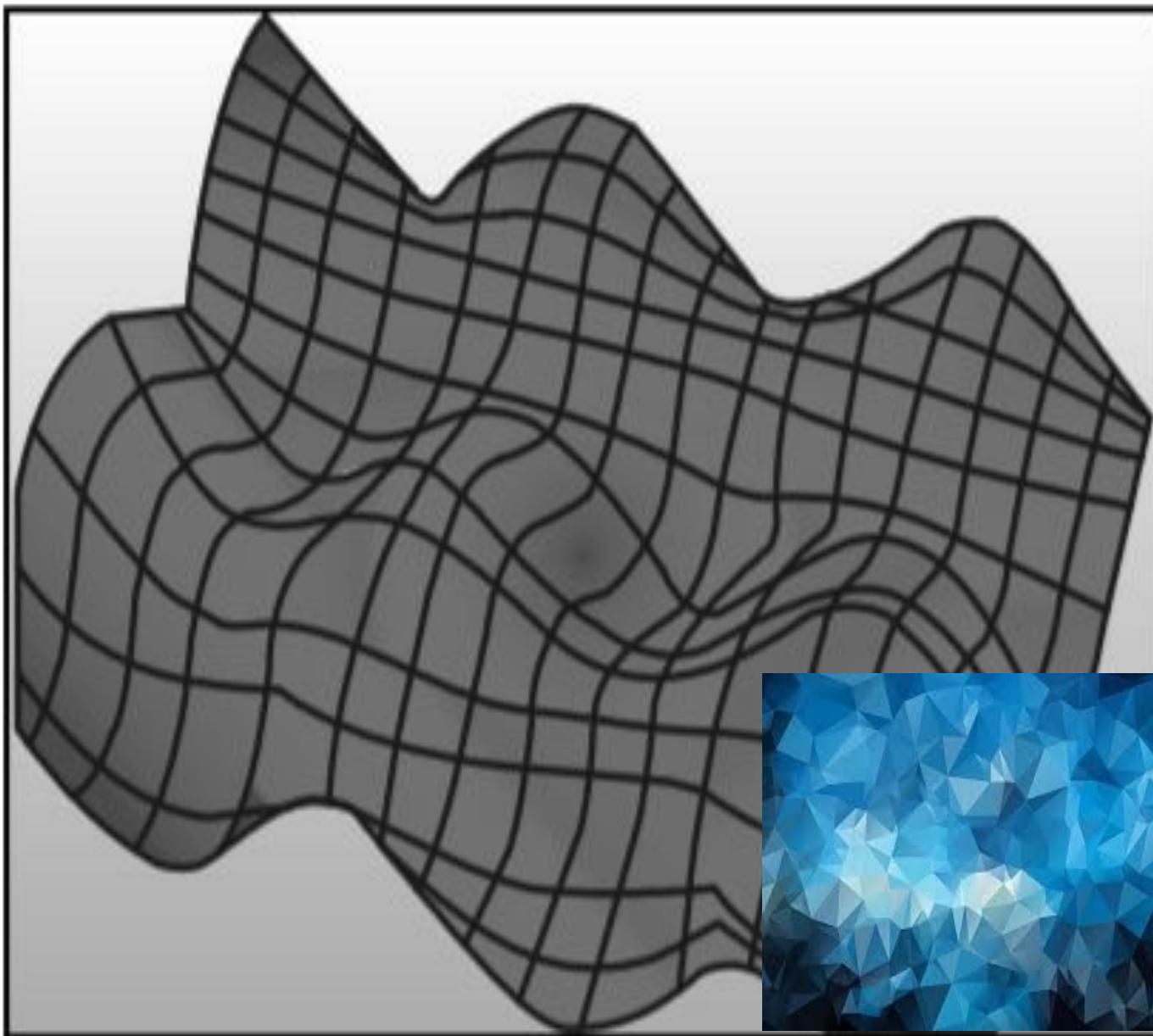
Replication 1

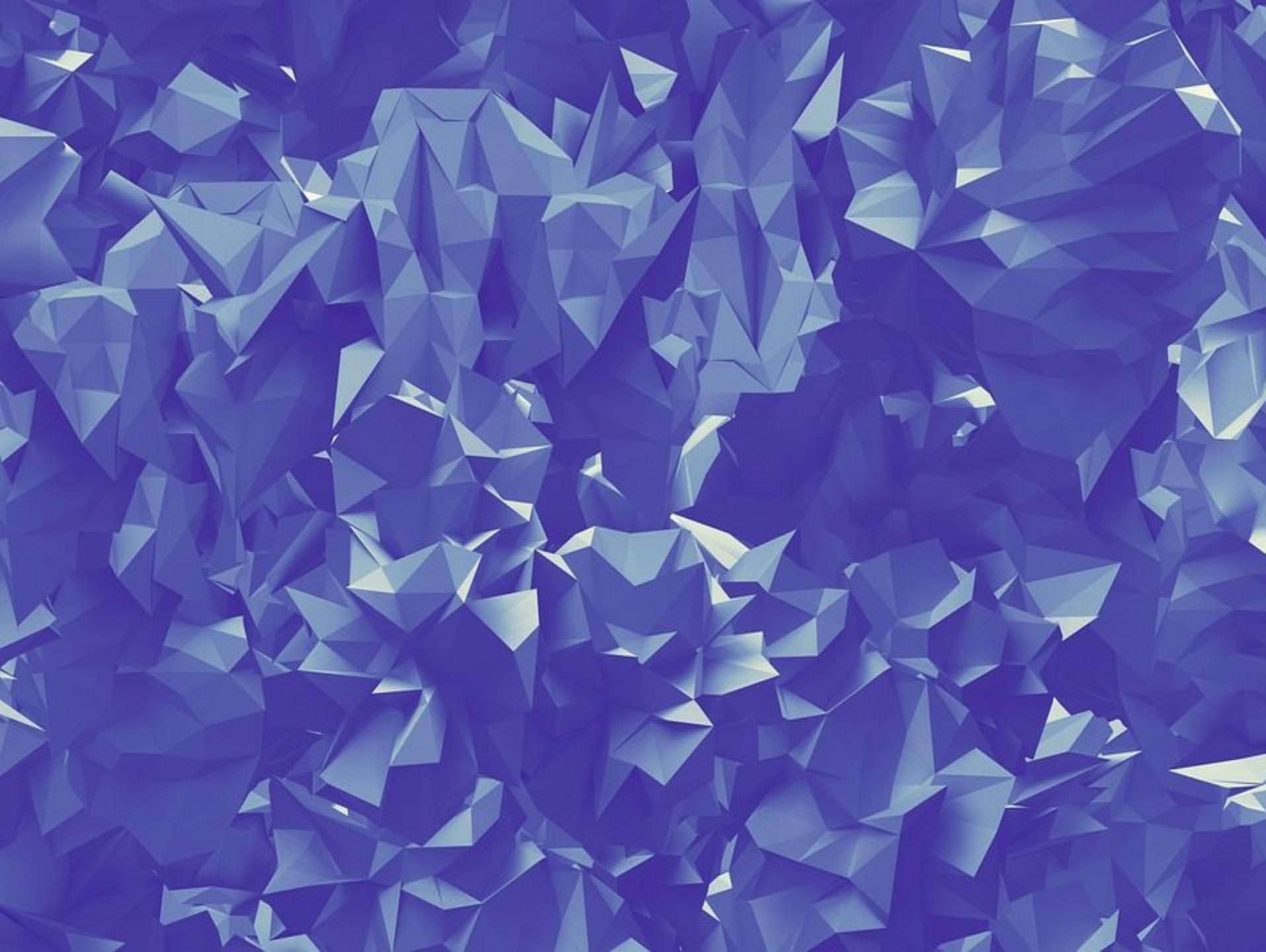


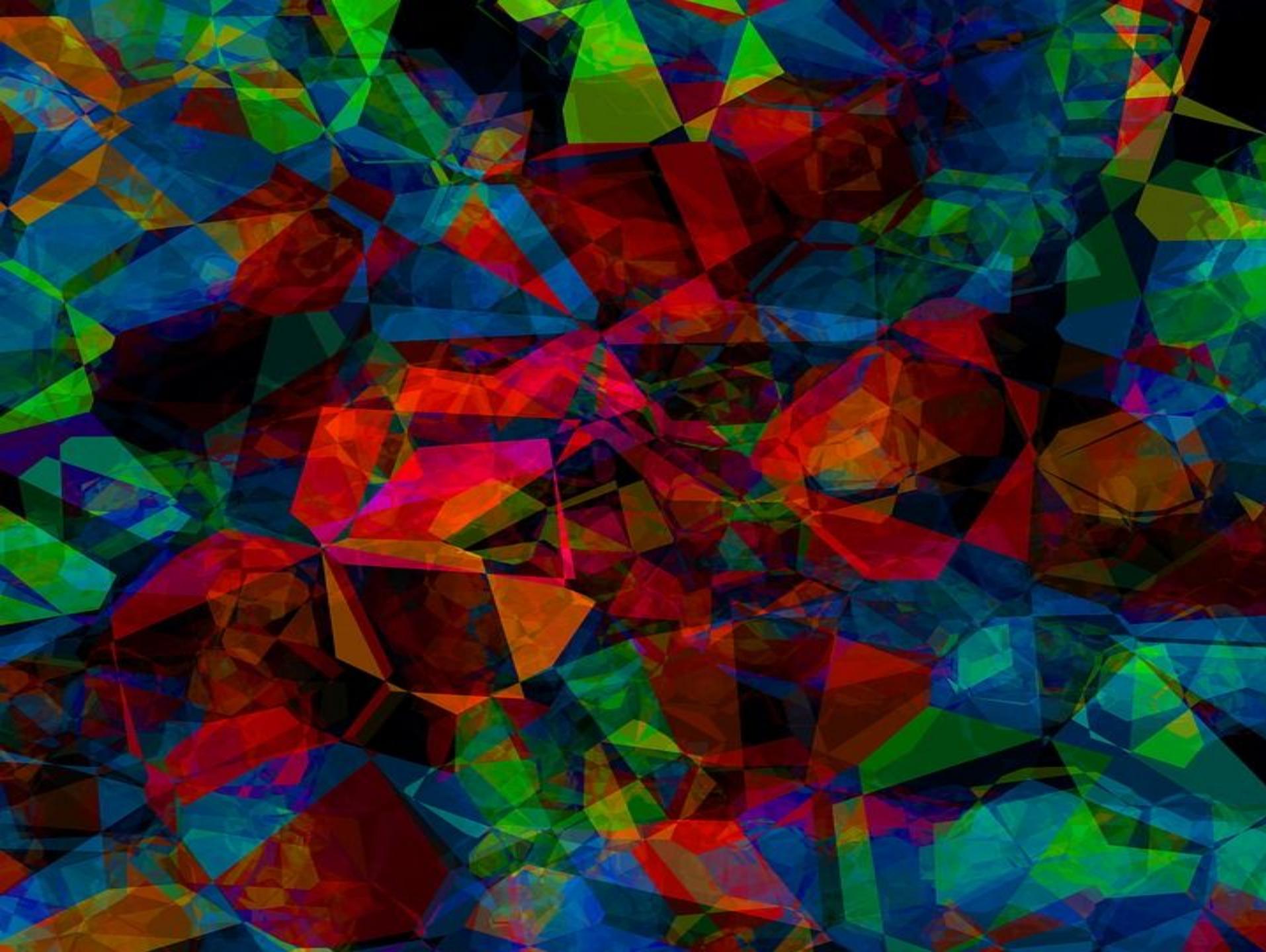
Replication 2

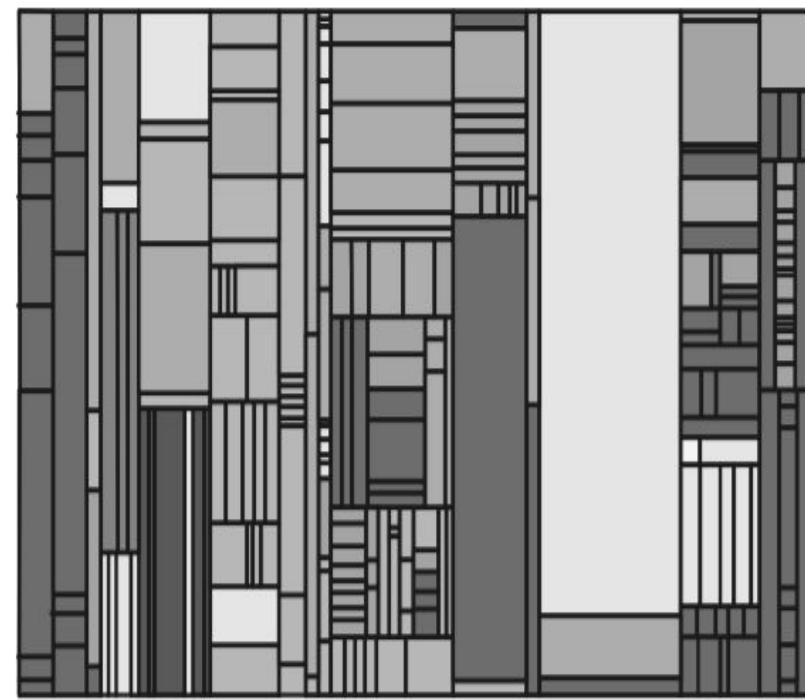
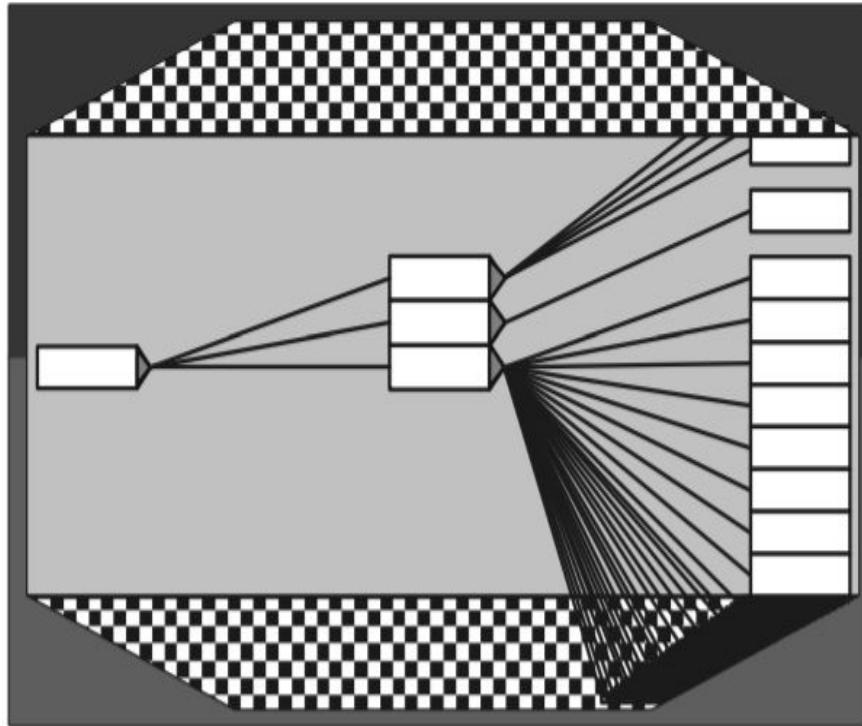
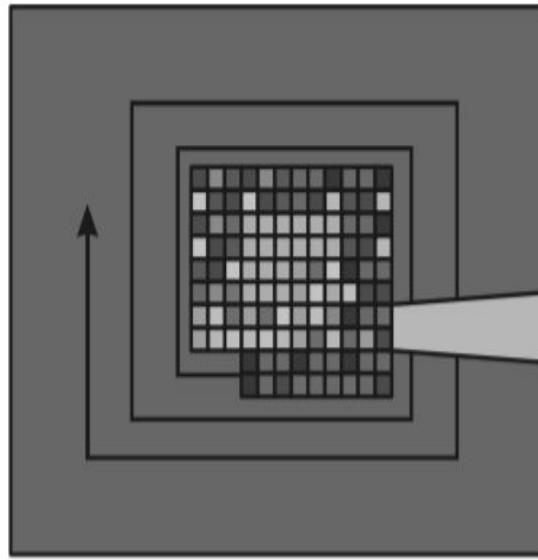




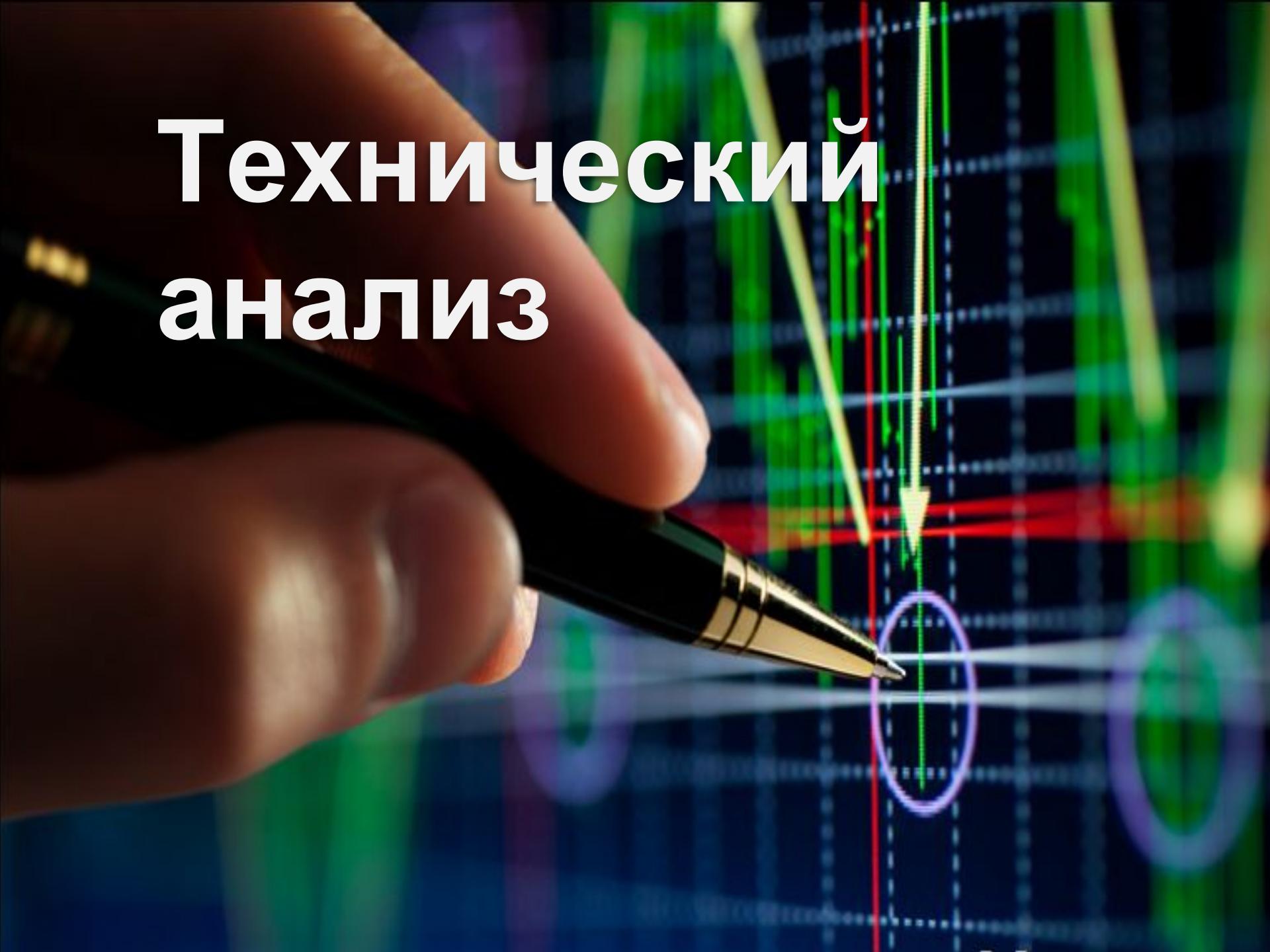








Технический анализ



**Хомма Мунэхиса
(1724 - 1803)**

**Создатель японских
свечей и первого
хедж-фонда**



JuckFava

EURUSD

1

30

1h

15



FACEBOOK INC, D, BATS

O 81.41 H 81.52 L 80.18 C 80.42

● loading data

Vol (20) loading... n/a - n/a



JuckFava

British Pound/Japanese Yen, D, FX

O 187.882 H 188.601 L 187.643 C 187.647

● closed

Vol (20)

n/a

n/a

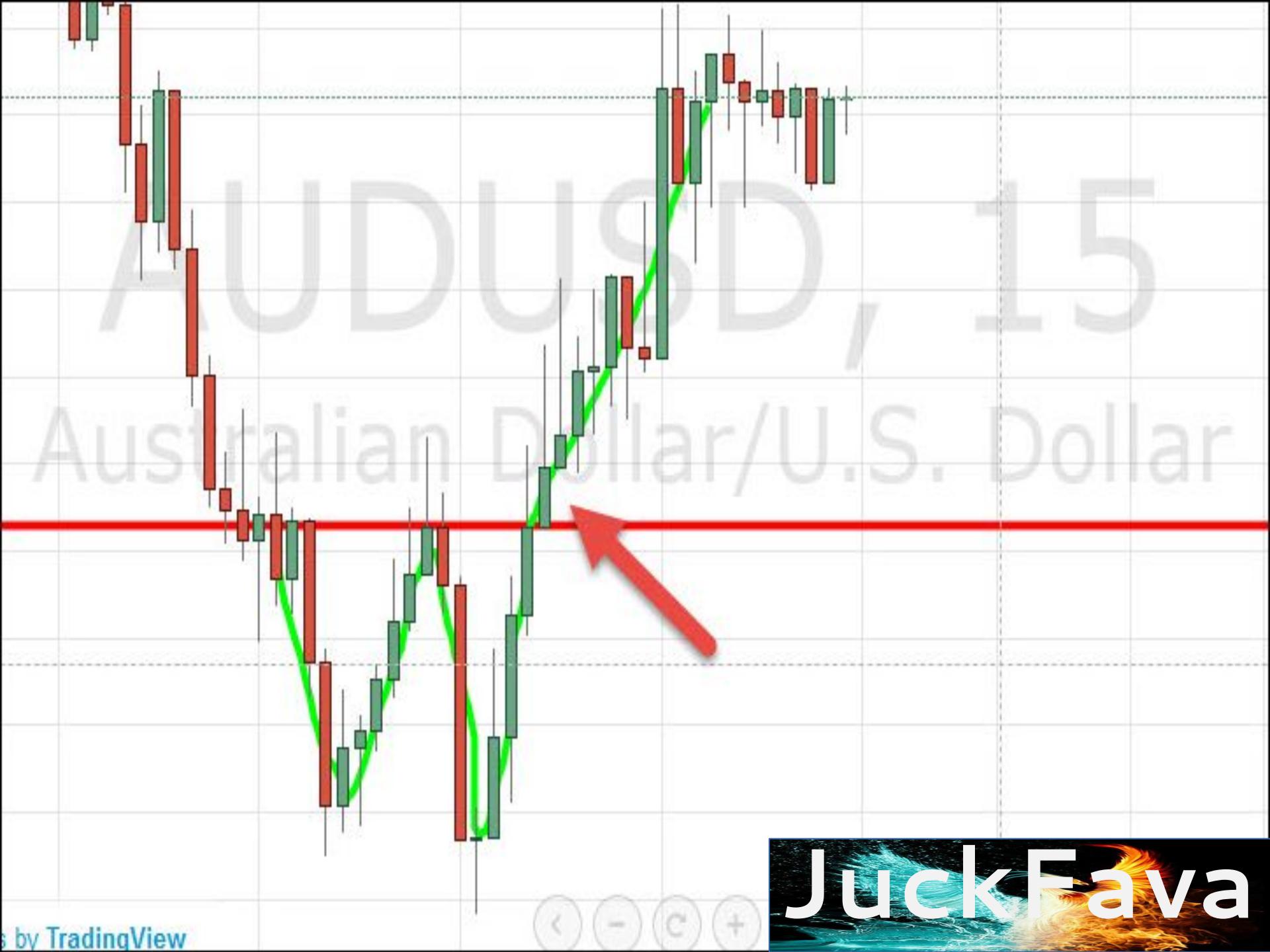
GBPJPY D
British Pound/Japanese Yen



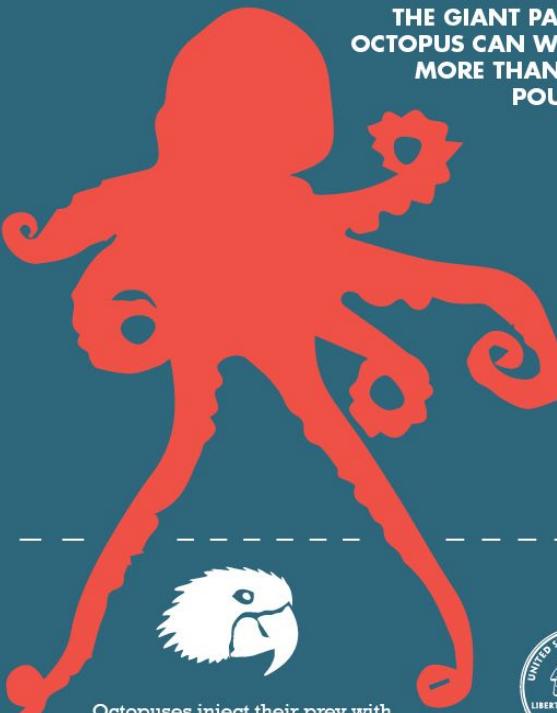
GBPJPY chart by [TradingView](#)



JuckFava



WORLD OCTOPUS DAY



Octopuses inject their prey with venom using a beak similar to a bird's made from the same tough material as a lobster shell.



THE GIANT PACIFIC OCTOPUS CAN WEIGH MORE THAN 600 POUNDS



ALL SPECIES ARE VENOMOUS, BUT THE BLUE-RINGED OCTOPUS IS THE ONLY ONE DANGEROUS TO HUMANS, RESPONSIBLE FOR AT LEAST TWO DEATHS.

one hundred thousand

IS THE MAXIMUM NUMBER OF EGGS THAT A FEMALE OCTOPUS CAN LAY, BUT THE AVERAGE LITTER SIZE IS ONLY 80.

OCTOPUSES VS. OCTOPI

THE PLURAL IN ENGLISH IS "OCTOPUSES," BUT THE GREEK PLURAL FORM "OCTOPODES" IS SOMETIMES USED. "OCTOPI," WHILE COMMONLY USED, IS CONSIDERED INCORRECT.



AN OCTOPUS HAS 3 HEARTS



OCTOPUSES ARE ABOUT
90%
MUSCLE



THE GIANT PACIFIC OCTOPUS CAN INHABIT DEPTHS OF UP TO 5,000 FEET

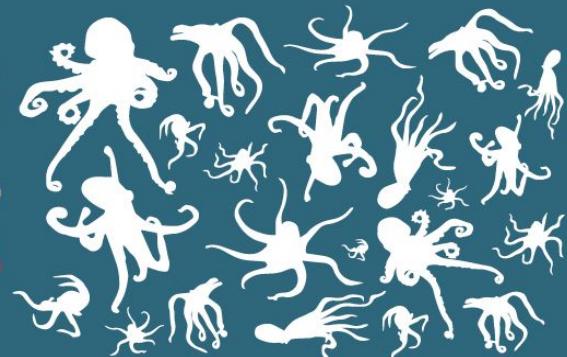


A mature female octopus can have up to 280 suckers on each arm! Each sucker contains thousands of chemical receptors, with sensitivities to both touch and taste.

OCTOPUSES CAN QUICKLY CHANGE THE COLOR AND TEXTURE OF THEIR SKIN

BECAUSE THEY DON'T HAVE BONES, EVEN LARGE OCTOPUSES CAN FIT THROUGH AN OPENING THE SIZE OF A QUARTER

300
RECOGNIZED
SPECIES
OF OCTOPUS



NATIONAL AQUARIUM.



aqua.org

JuckFava

Save Preview Download Share

New Pidchart

Shapes & Lines

Icons

Search icon here

Technology

All Color Mono

Photos

Photo Frame

Shapes & Lines

Icons

Search icon here

Technology

All Color Mono

Photos

Photo Frame

HOW TO START

HOW TO START AND DESIGN AN INFOGRAPHIC

INSPIRATION FROM AROUND THE WEB

When you are new to something, taking a look at what experts did, and how they did it is a fantastic way to begin. For starters, it will help you avoid most rookie mistakes without doing it.

PICKING THE RIGHT COLOR SCHEME

Picking the right colours palette for your infographic is a key step. You will need to take into account your audience, your imagery and your desired goal.

A screenshot of a graphic design software interface showing an infographic template. The template has a dark background with white and red text. It features a large title 'HOW TO START' and a subtitle 'HOW TO START AND DESIGN AN INFOGRAPHIC'. Below these are two main sections: 'INSPIRATION FROM AROUND THE WEB' and 'PICKING THE RIGHT COLOR SCHEME'. Under 'INSPIRATION', there is a paragraph of text. Under 'COLOR SCHEME', there is another paragraph. At the bottom left is a bar chart with four bars of increasing height in purple, red, orange, and yellow. At the bottom right is a donut chart divided into three segments with percentages: 60% (purple), 15% (orange), and 25% (red). The sidebar on the left contains various tools and categories like 'Shapes & Lines', 'Icons', 'Photos', and 'Photo Frame'.

JuckFava

Add chart



Search by chart type

Sort by: Chart type ▾

Line



Lines

Bar



Bar

Stacked



Grouped

Radial

Column



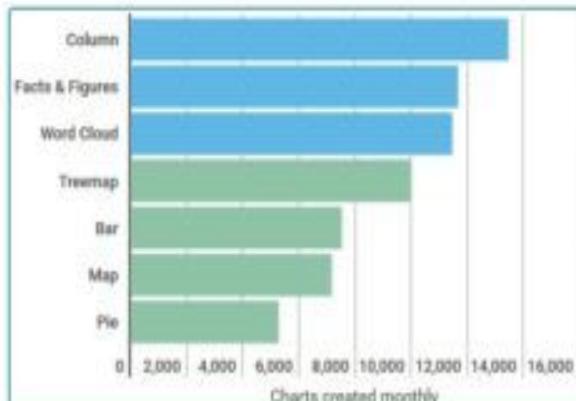
Column

Stacked

How to choose the right chart?

Your first project title

Here are our most popular chart types. To start [double click](#) to edit any text, chart or map.



Over 1 Billion

Views on content created with Infogram



5,124,059

Total charts and infographics created



Easily publish your content

Add your data visualizations to your website with our responsive embeds. You can also download them as PDF, PNG or as an animated GIF!

Chart

Chart type



Edit data

 Chart category height (px) Chart height (px)

300

 Colors Axis & Grid Number format Switch rows and columns Play controls Values Show values outside

How can we help you?

JuckFava



Main Title



LOREM
IPSUM

LOREM
IPSUM

Lorem ipsum dolor sit amet, poset scitia quis premevit.

LOREM
IPSUM

Lorem ipsum dolor sit amet, poset scitia quis premevit.

LOREM IPSUM
DOLOR

Lorem ipsum dolor sit amet, poset scitia quis premevit. Actus nolo presenti premevit. vestes. Lorem ipsum dolor sit amet, poset scitia quis premevit. confidemus vestes.

LOREM
IPSUM

Lorem ipsum dolor sit amet, poset scitia quis premevit.

LOREM
IPSUM

Lorem ipsum dolor sit amet, poset scitia quis premevit.

LOREM
IPSUM

LOREM
IPSUM

LOREM
IPSUM

LOREM
IPSUM

SHOW MORE

- 100% +



JuckFava



Поиск среди 1 000 000 изображений...

Logo Gothic

21

:

B

I

≡

AA

≡

Интервалы

Копировать

Сортировка



Бесплатно



Скрытие



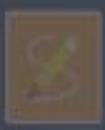
Редакт.



Фигуры



Линии



Уникальные



Значки



Диаграммы



Карта



- 75% +

JuckFava



JuckFava

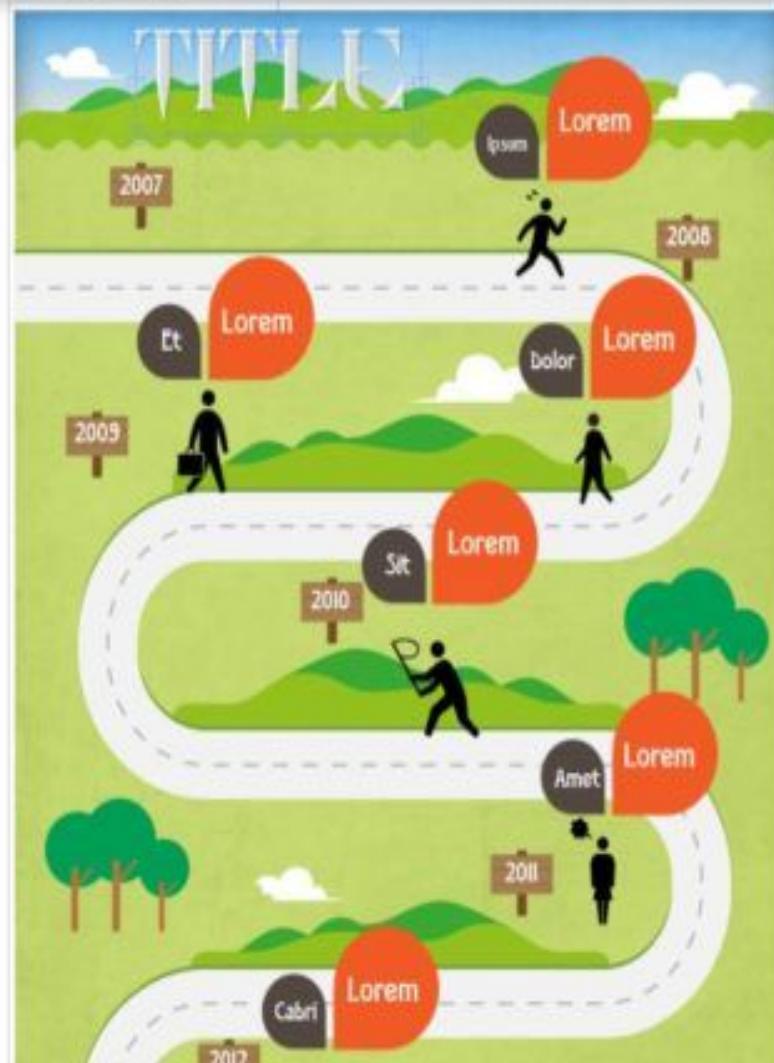
[Templates](#)[Objects](#)[Media](#)[Backgrounds](#)[Lines](#)[Shapes](#)[Text](#)[Chart](#)[Upload](#)

Zoom @ 100%

grid

undo

redo

[Present](#)pre-area
line, shapes...[Bar](#)[Column](#)[Line](#)[Radar](#)Want more charts? [Go Pro!](#)[Chat with us](#)

JuckFava

Untitled Project Save Adobe XD Preview Publish Upgrade

Objects (66) Trebuchet MS 48 + B I U A A

All Shapes Recently Used Lines Q

Shapes Lines Animals Arrows Banners Buildings & Landmarks Business Buttons Celebration Clothing & Shoes Decorative Elements Education Emotions Entertainment Food Geography Gestures

W: 480 H: 375 C: 116 Y: 48

STARTING A DESIGN PROJECT

Lore ipsum

Lore ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

Lore ipsum

Lore ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh .

AI EPSB EPS10 PSD

PPT CDR PNG JPG

Lore ipsum Lore ipsum

Lore ipsum Lore ipsum

Slides

JuckFava

- HORIZONTAL BAR
- VERTICAL BAR
- STACKED HORIZONTAL BAR
- STACKED VERTICAL BAR
- GROUPED HORIZONTAL BAR
- GROUPED VERTICAL BAR
- DONUT
- PIE
- LINE
- AREA
- STACKED AREA
- SCATTER PLOT
- BUBBLE
- RADER



JuckFava

Источники

- <http://infographer.ru/tag/parallelnye-koordinaty/>
- <http://kek.ksu.ru/eos/WM/AnalizDannyhProcessov.pdf>
- <https://habr.com/ru/company/ods/blog/323210/>