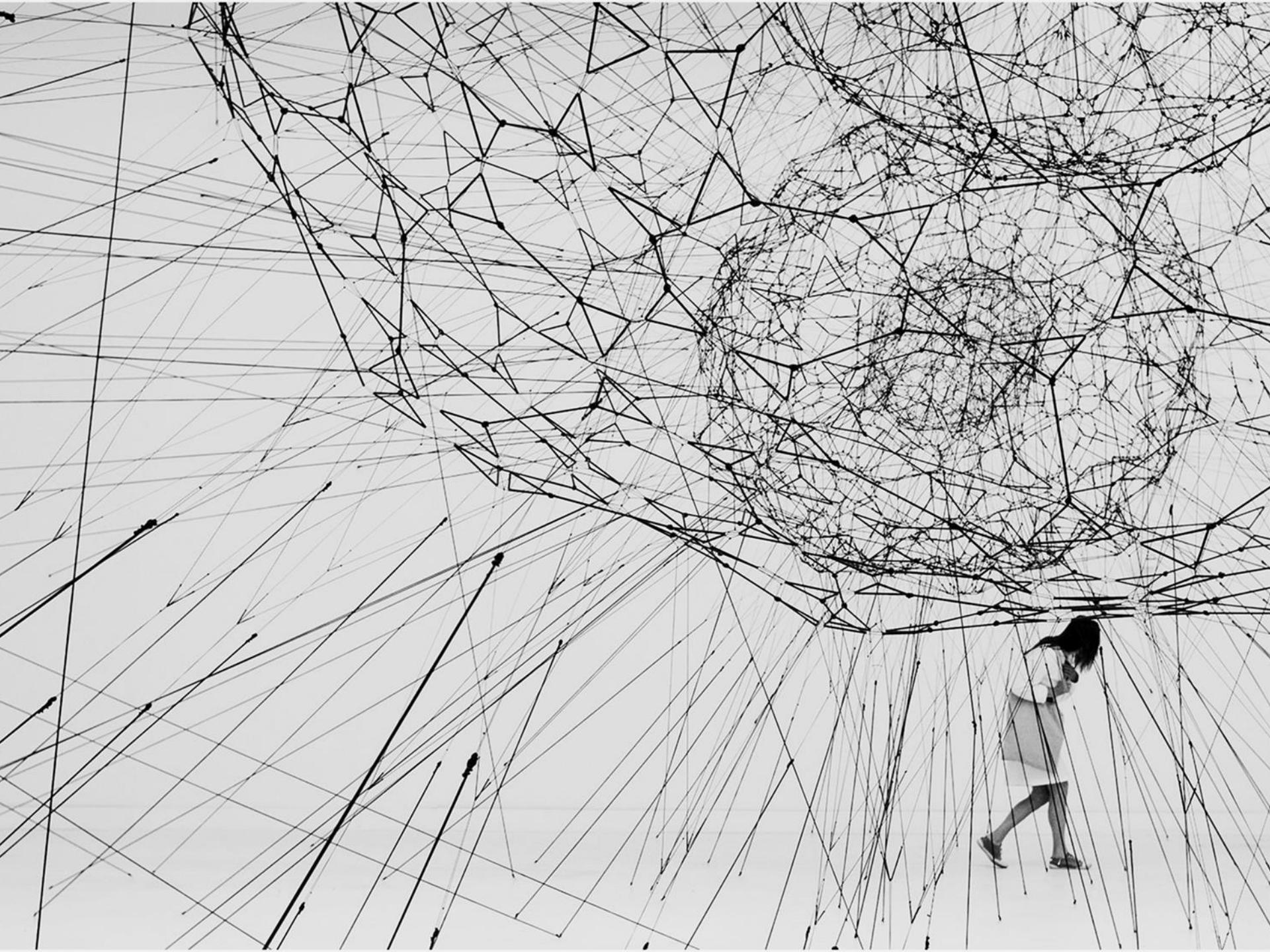


Data mining

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

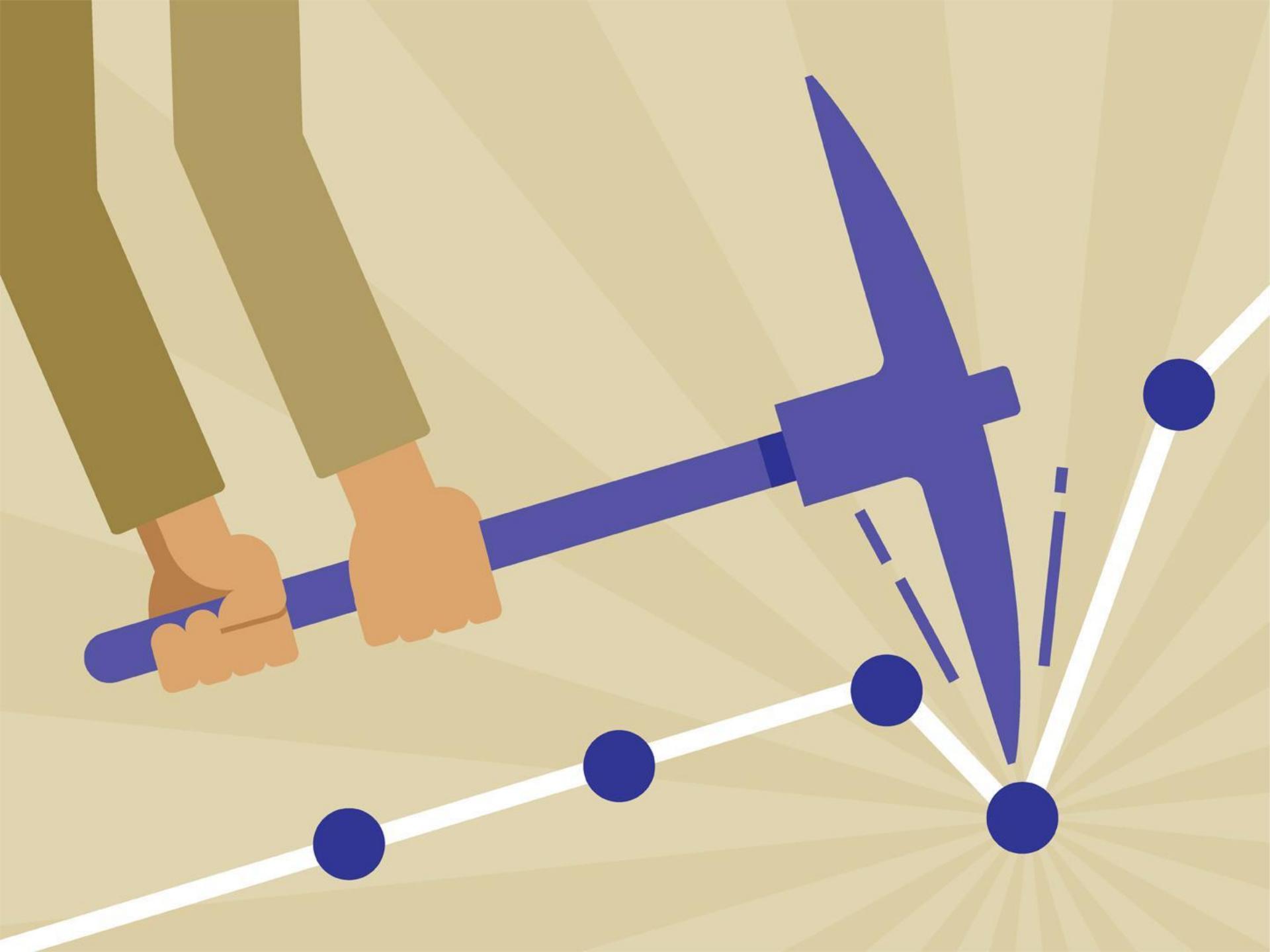


Data mining

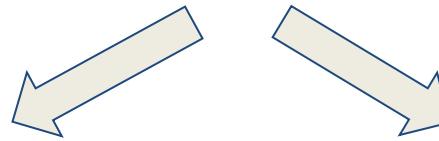
- Витяг, збір даних, видобуток даних (ще використовують Information Retrieval або IR);
- Витяг знань, інтелектуальний аналіз даних (Knowledge Data Discovery або KDD, Business Intelligence).

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 інтервальний 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)

Statistics

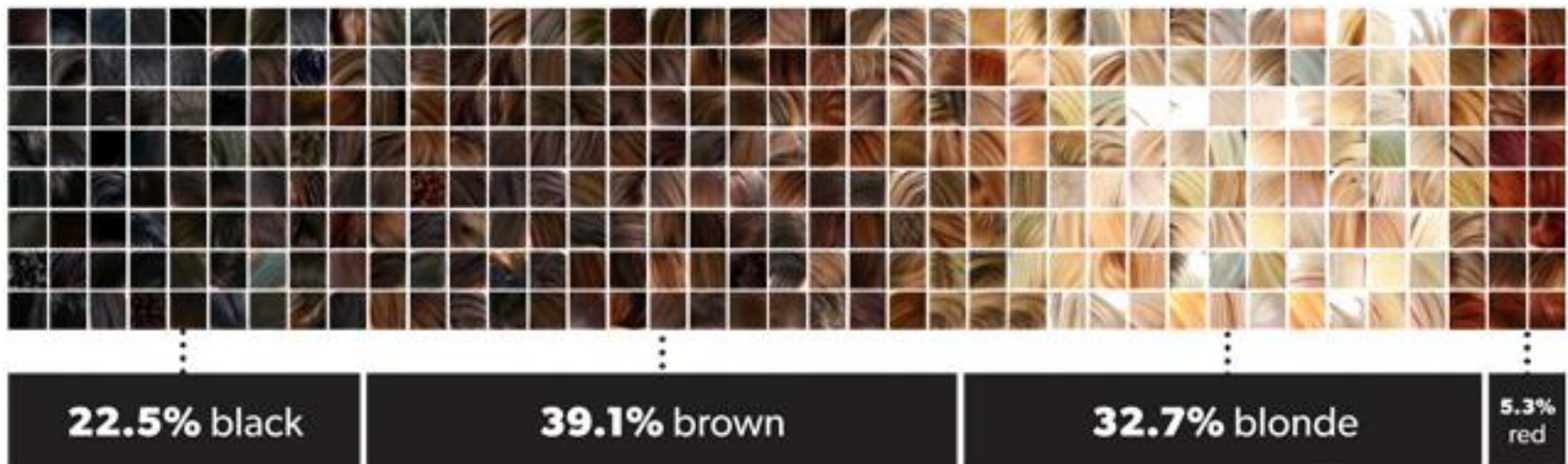
application experience
deduce part considered problems social working useful ways improve process
accounts accounts soundness terms
organization empirical summarize successful element logically companies usually
given samples since collecting larger roots wrong analysis
grouped inductive disciplines observations provide one confused survey
total mean statistic particular sampling set hypothesis well prove
forecasting particularly increase statistic disciplines
related applicable business access median particular sampling
a.k.a. available without advance leads
tools quantity median
prediction addition relevant predictive population statisticians
probability describe tested surveys science
experiments subject explanation used referring holds
prediction basis predictions referring discipline
mathematical singular true results government direction
branch method together communicating uncertainty
Inf calculated plural rather

Data Mining

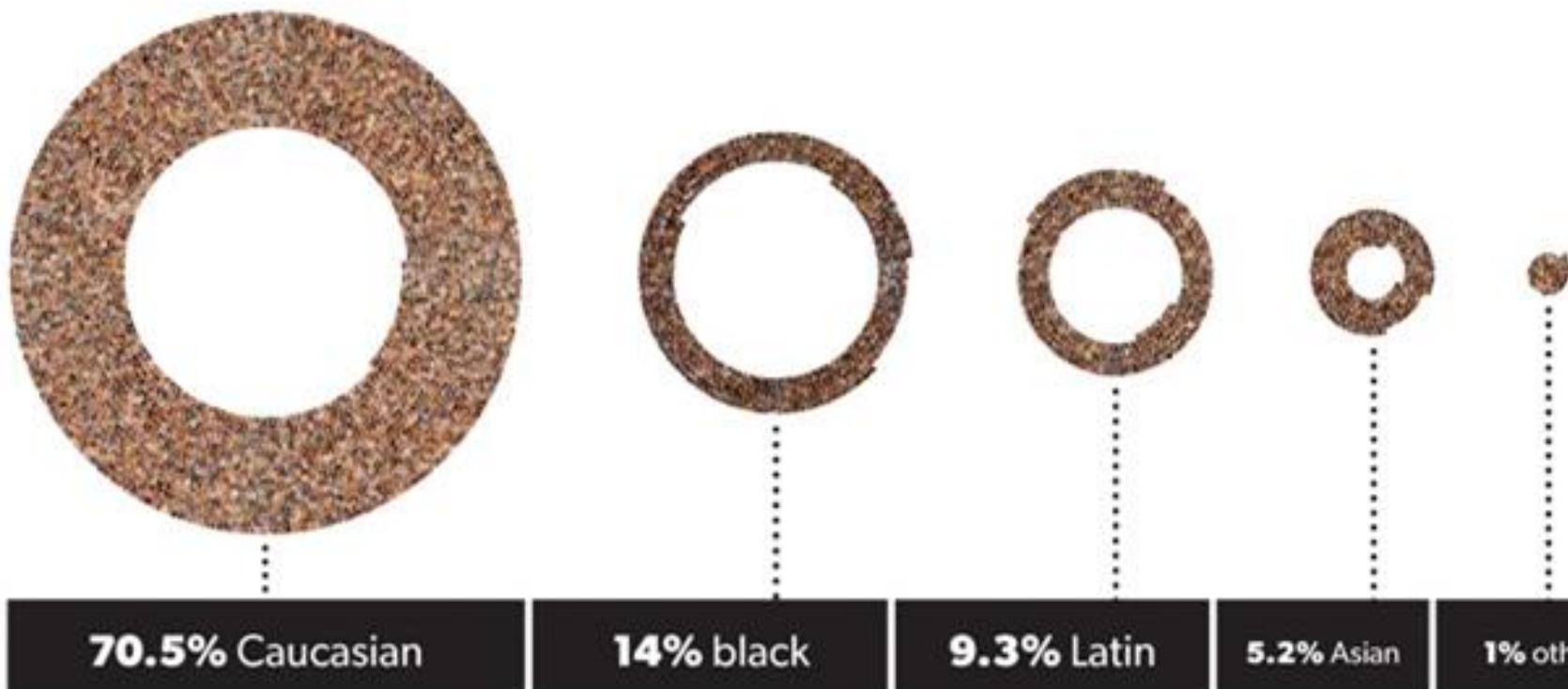


...аналіз 10 000 акторів
фільмів для дорослих

Розподіл акторів за кольором волосся



Розподіл акторів за кольором шкіри



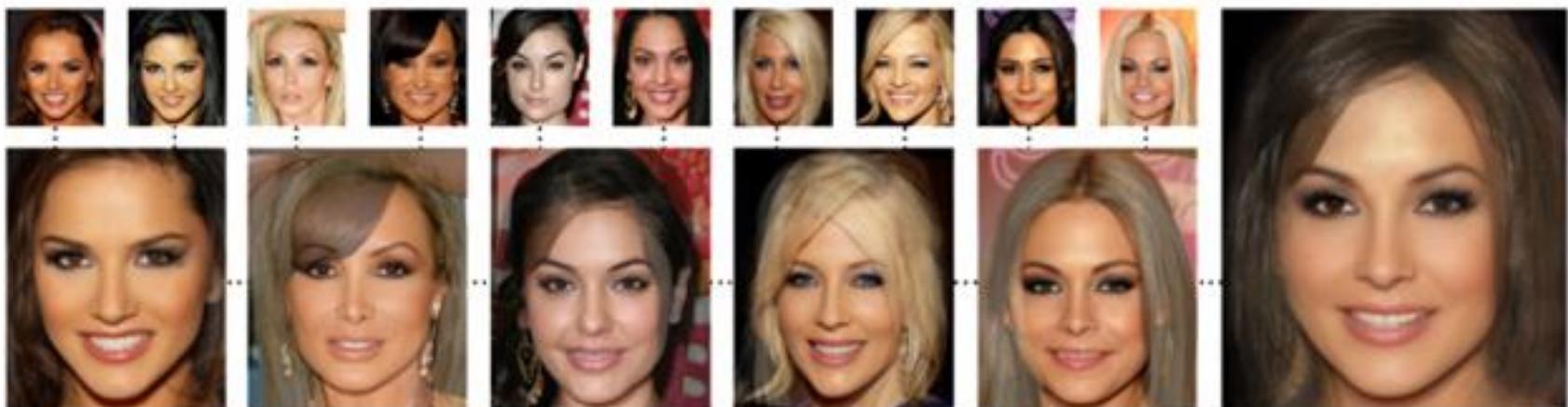
Розподіл акторів за наявністю татуювань

The % of porn stars who have a tattoo



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

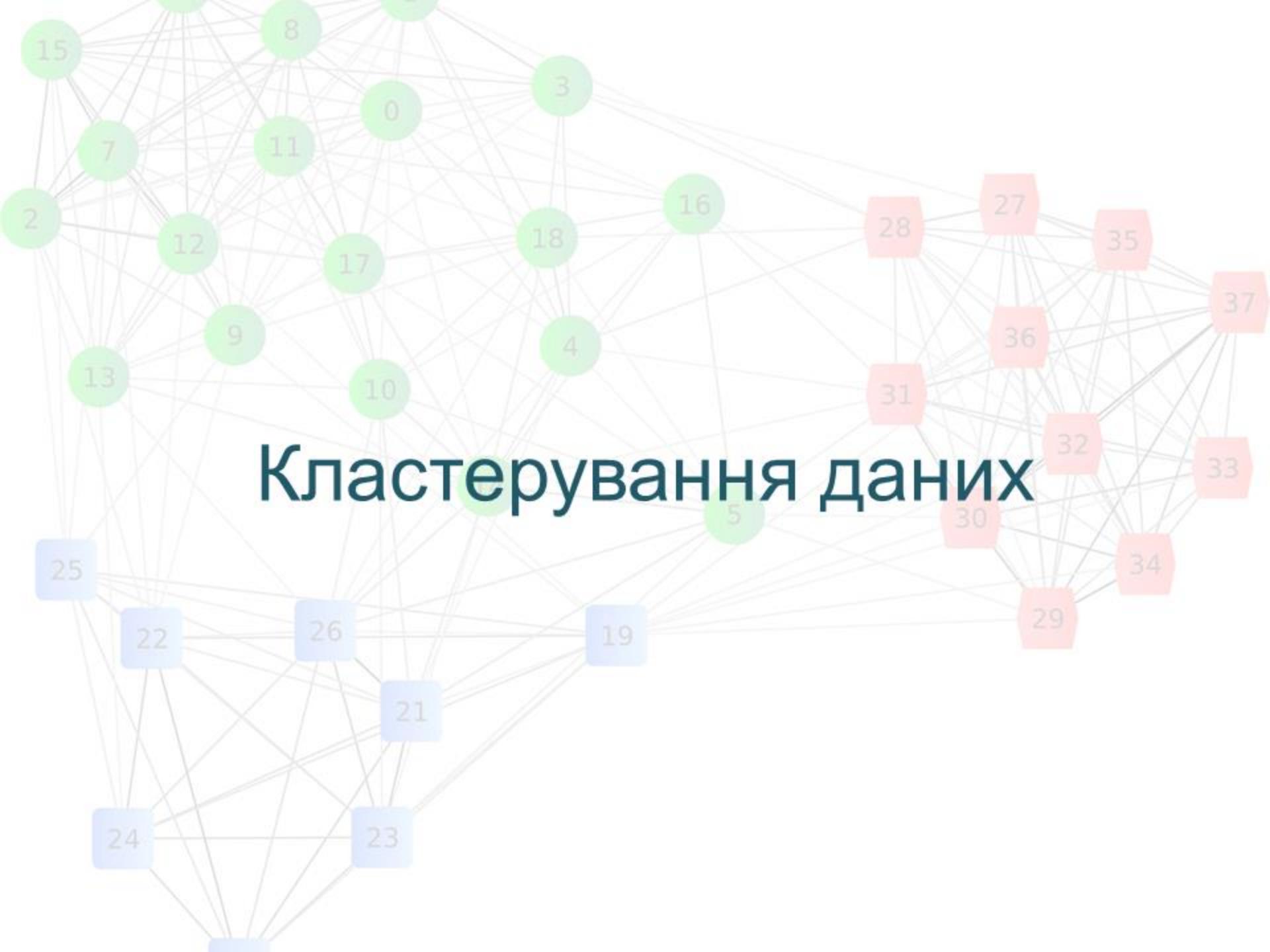
The Average Face of Ten
Top Female Pornstars



Посилання: jonmillward.com



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

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

Як працює метод к-середніх?

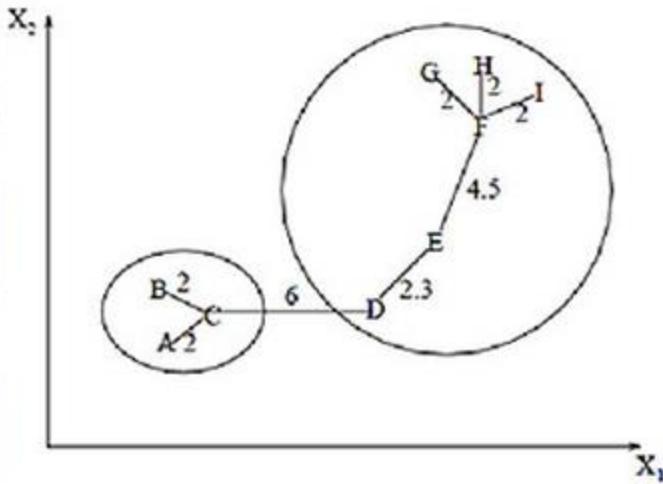
- Метод к-середніх вибирає точки багатовимірного простору, які будуть представляти к-кластери. Ці точки називаються центрами тяжіння. Перший раз, за відсутності припущень, центри тяжіння можна вибирати випадково
- Кожен пацієнт буде розташовуватися найближче до однієї з точок.
- Тепер у нас є к-кластерів, 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 зображене 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

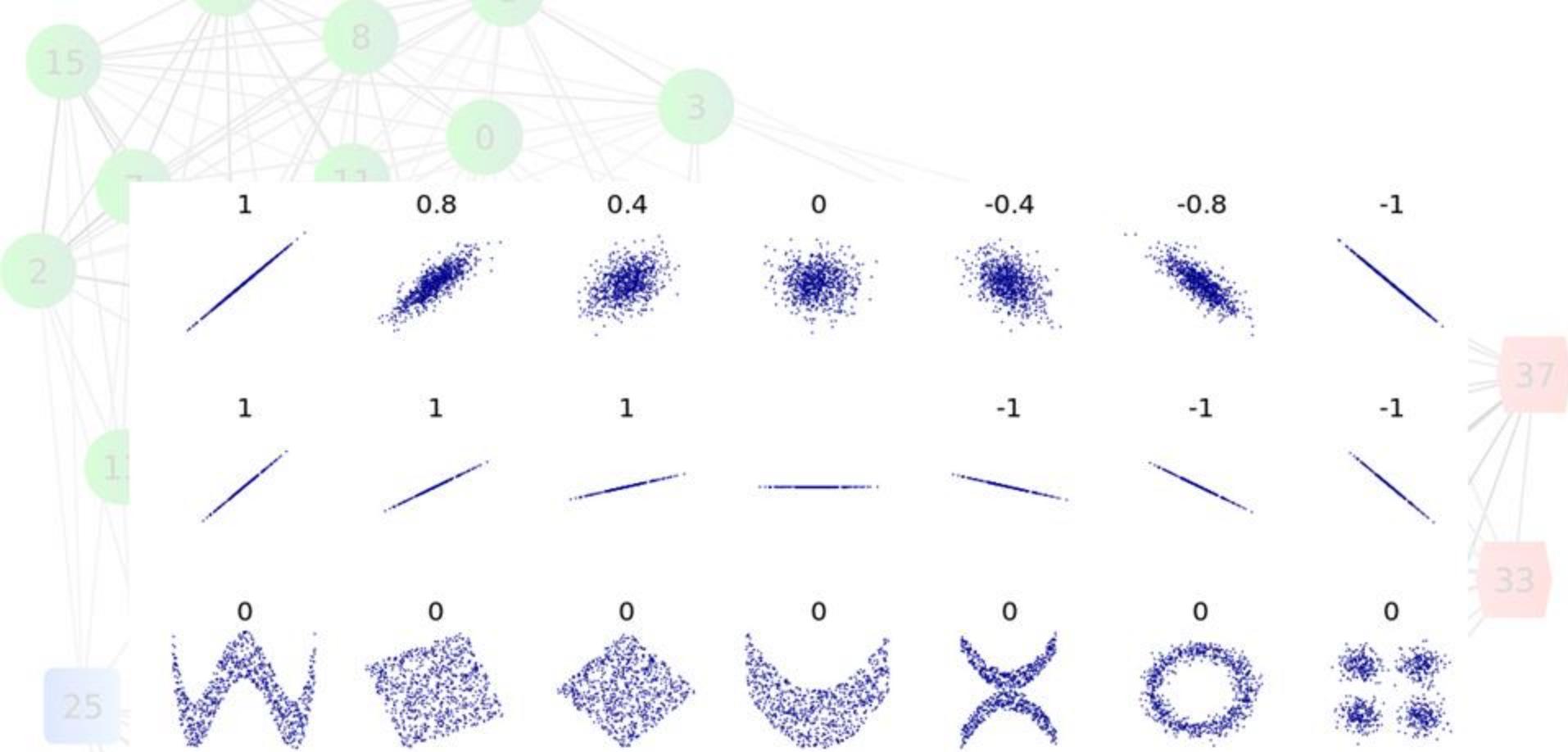
Порівняння деяких алгоритмів кластеризації

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 належності
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:

$$K_{xy}^{(21)} = \frac{1}{n} \sum_{i=0}^n (x_i y_i - m_x m_y)$$

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?

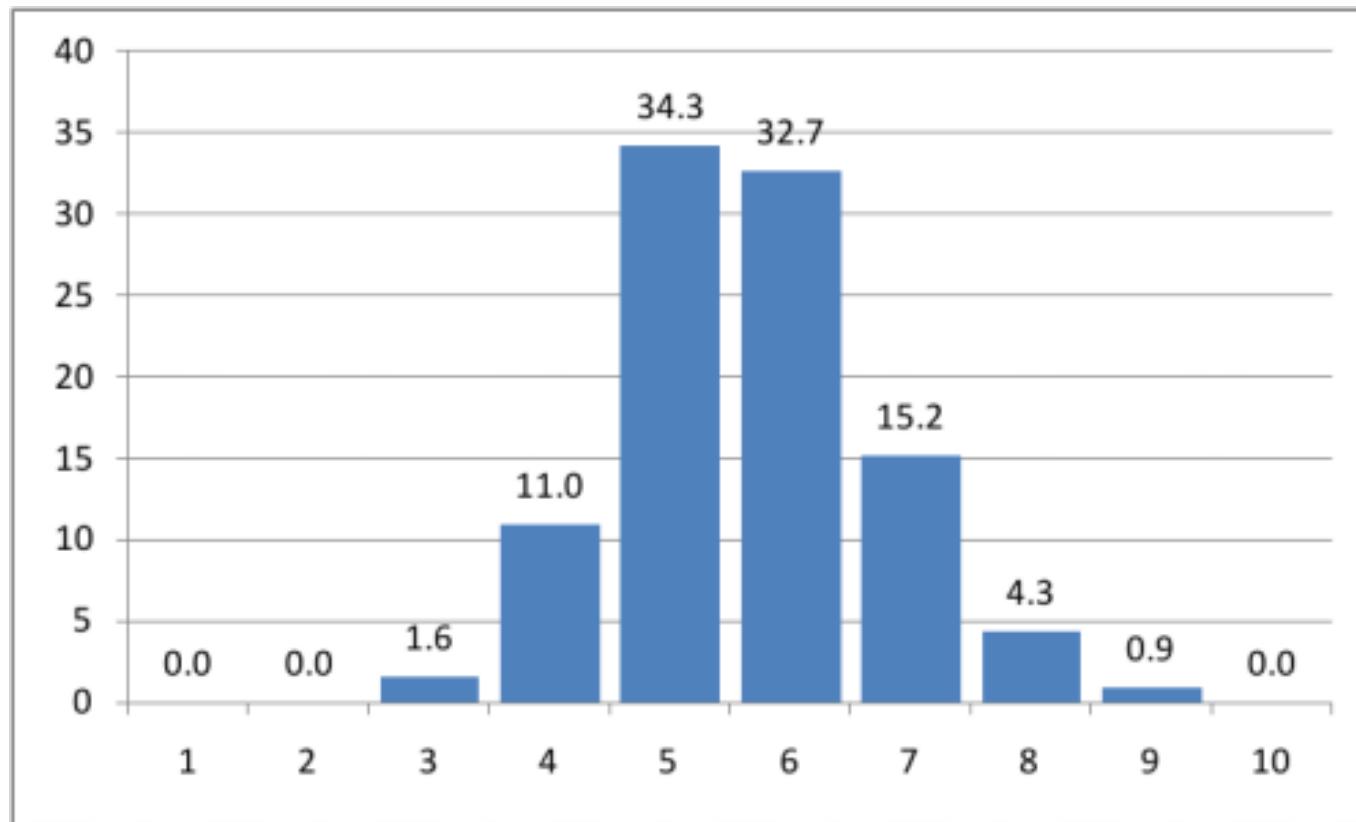
Кожна людина на Землі знайома з будь-якою іншою через ланцюжок з п'яти друзів, тобто, через шість рукостискань



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”



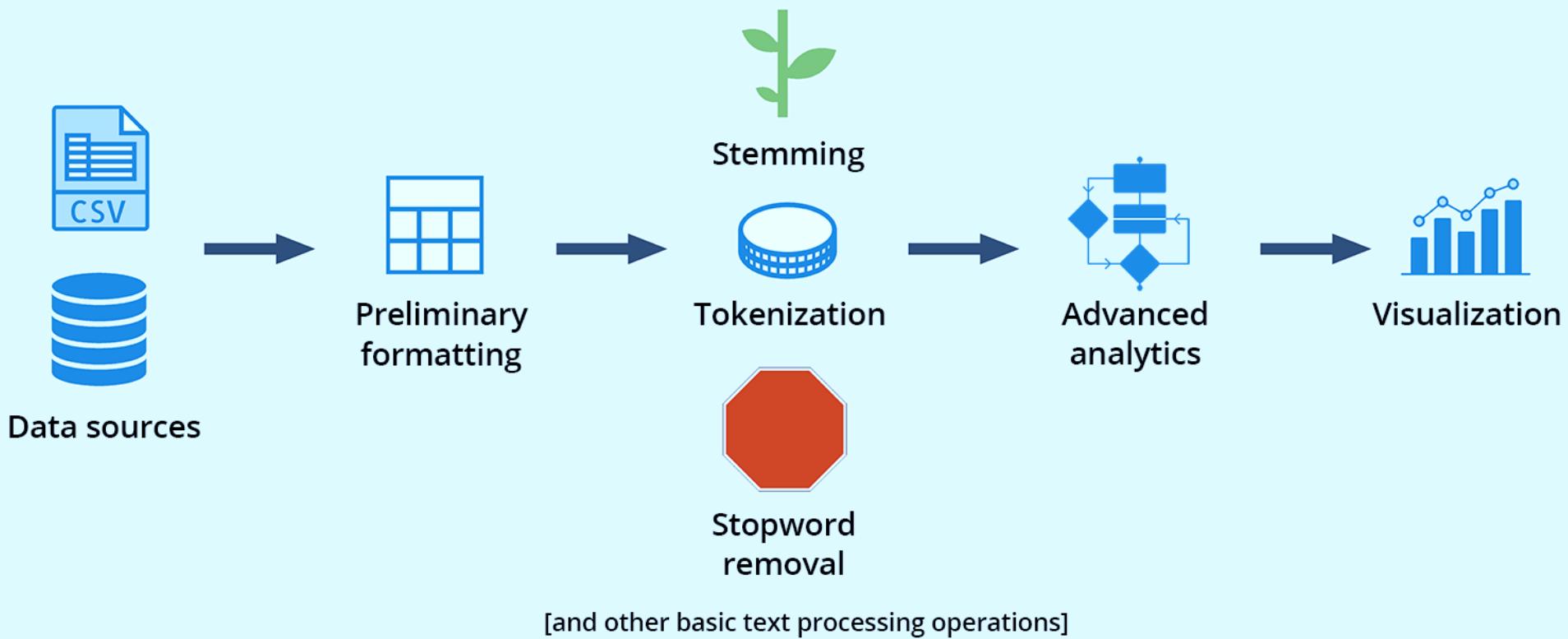
Основні бібліотеки - “спорядження”

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



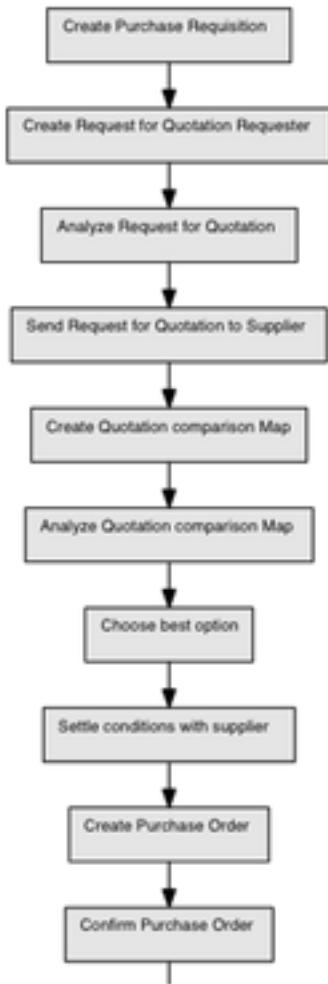
Виды интеллектуального анализа. Визуальный анализ

Text Mining

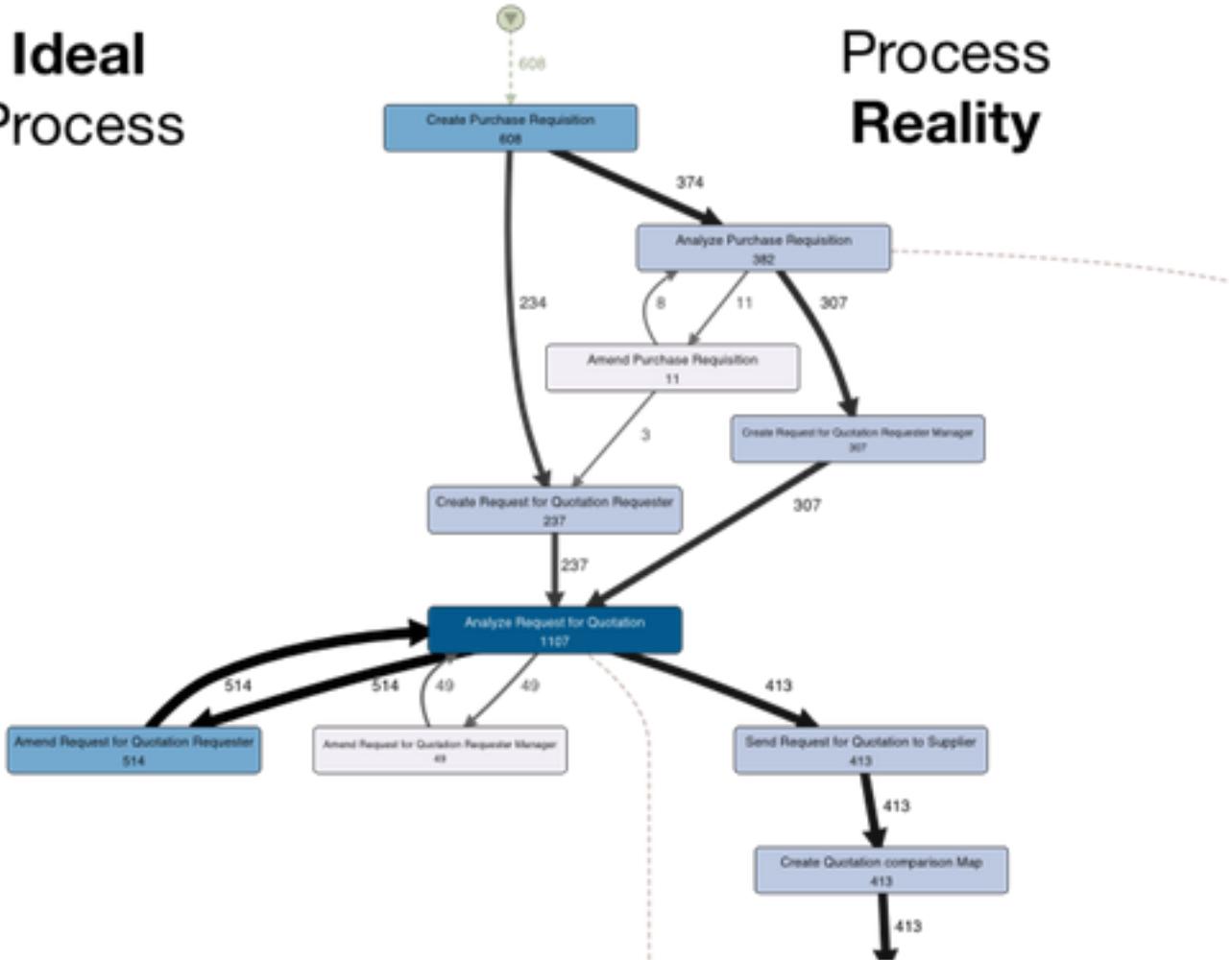




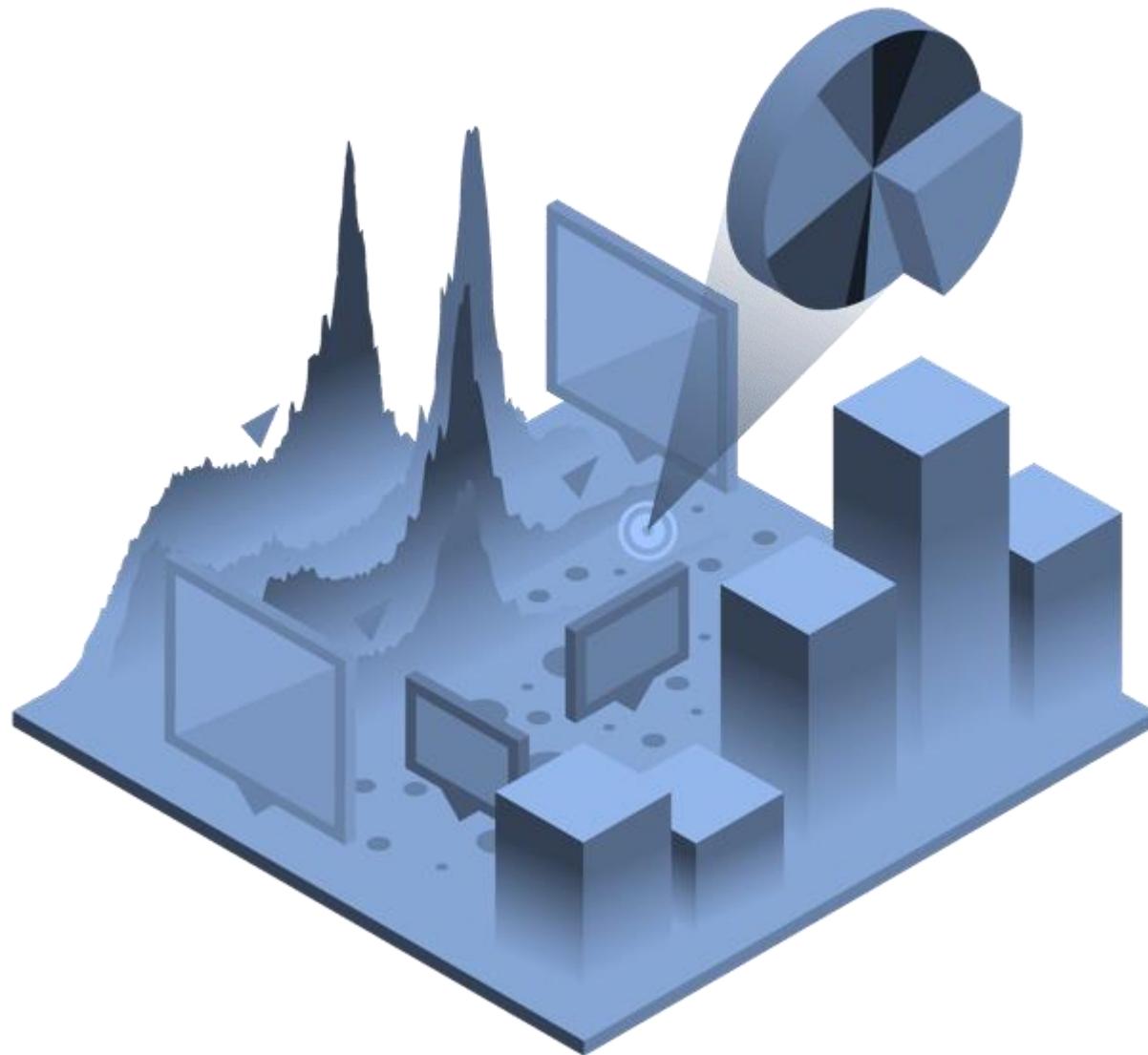
Process Mining



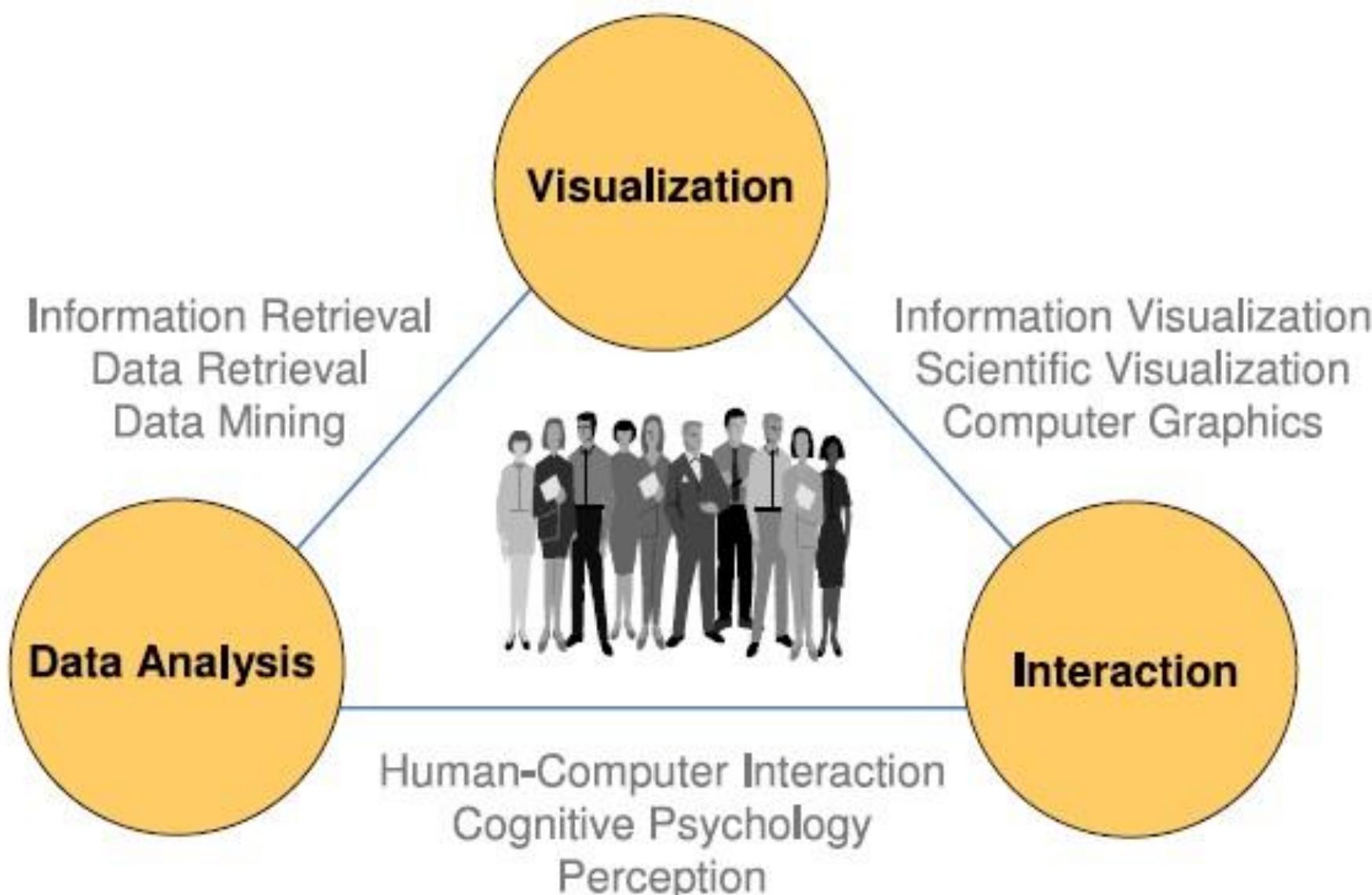
Process Reality



Visual Mining



Визуальный анализ



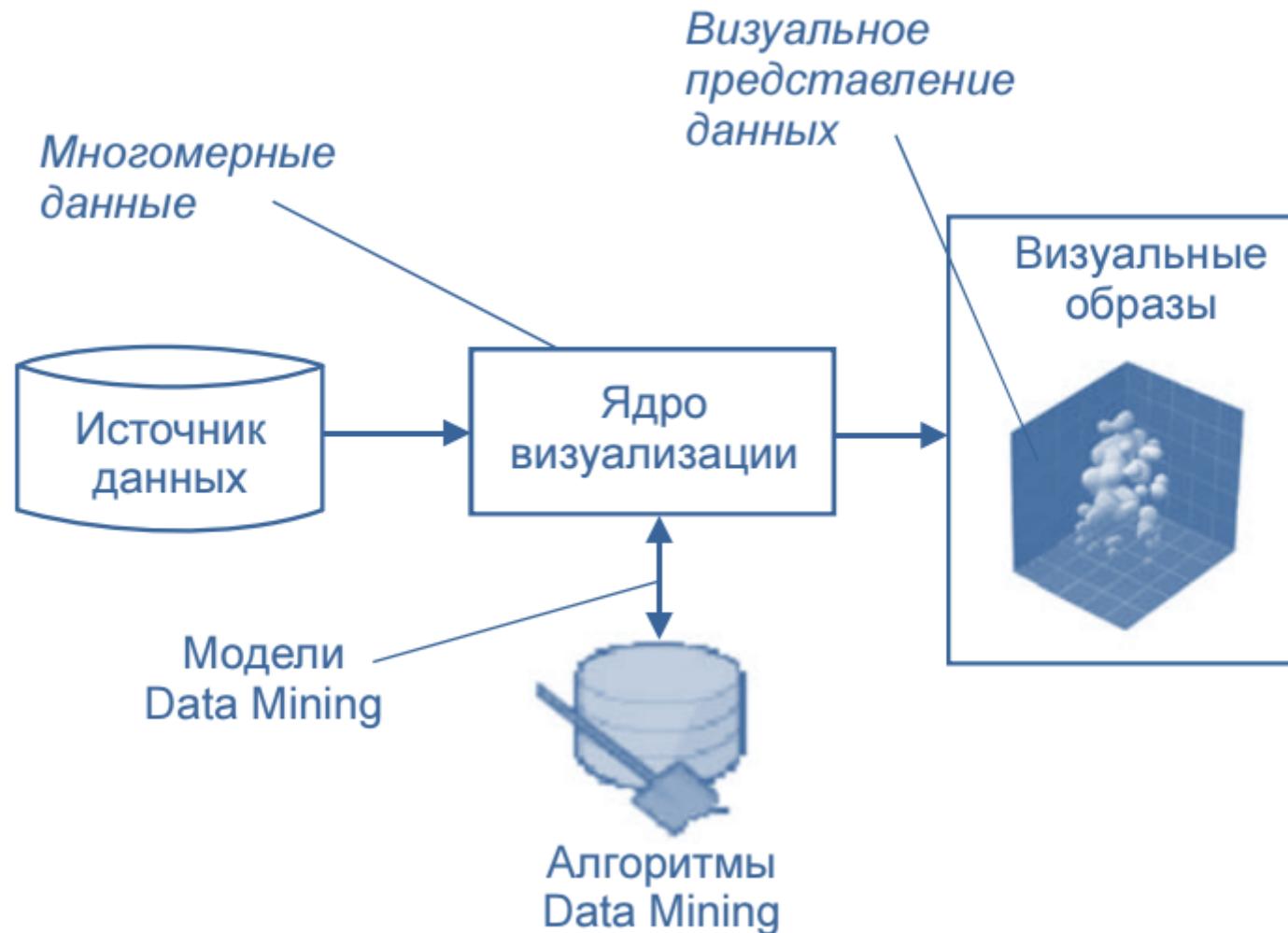
Почему же всё-таки Visual Mining?

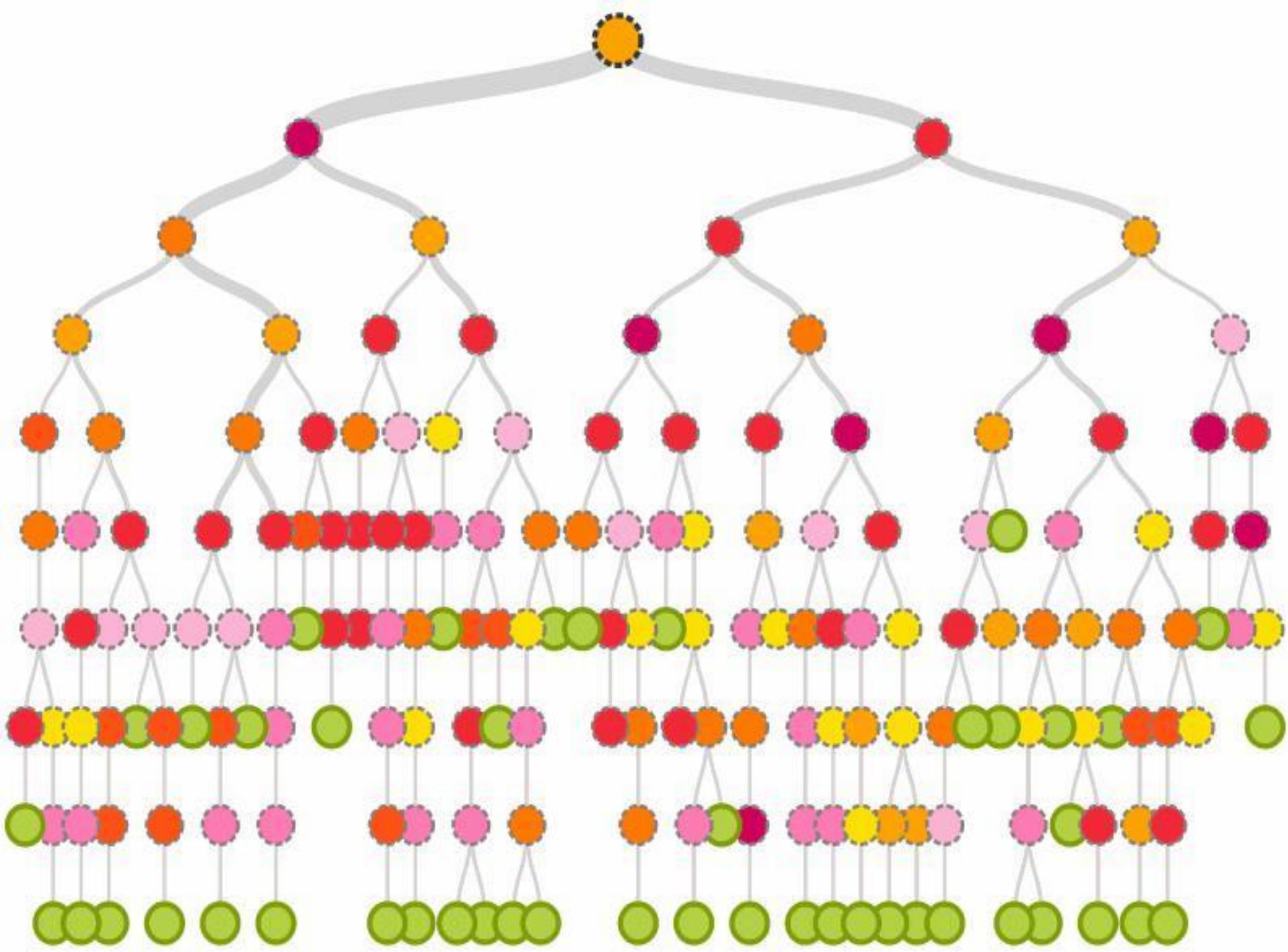
Преимущества:

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

	Data Mining	Visual Mining
<i>Действенность</i>	+	-
<i>Качественная оценка</i>	+	-
<i>Гибкость</i>	-	+
<i>Вовлечение пользователя</i>	-	+

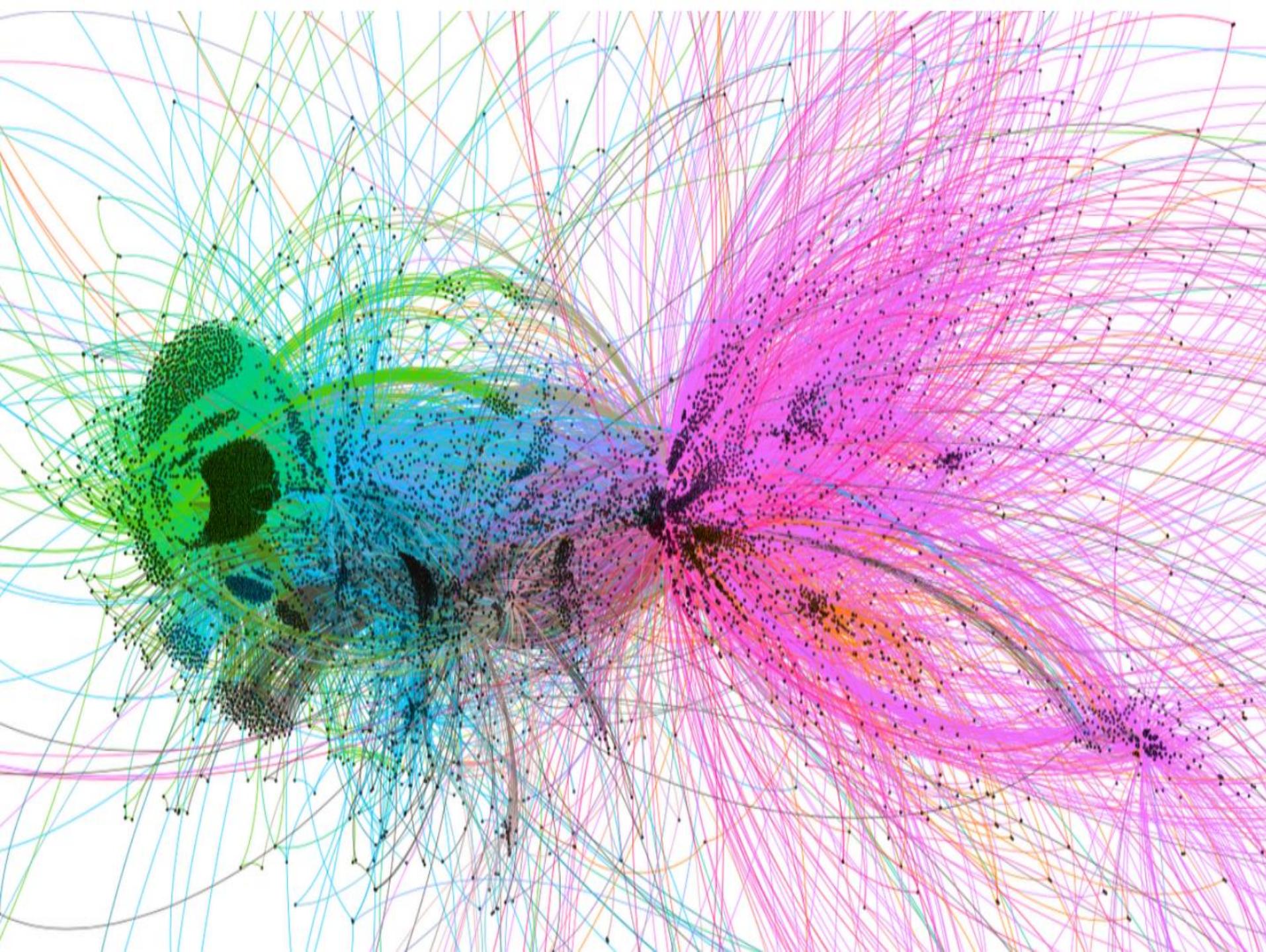
Методы визуального анализа

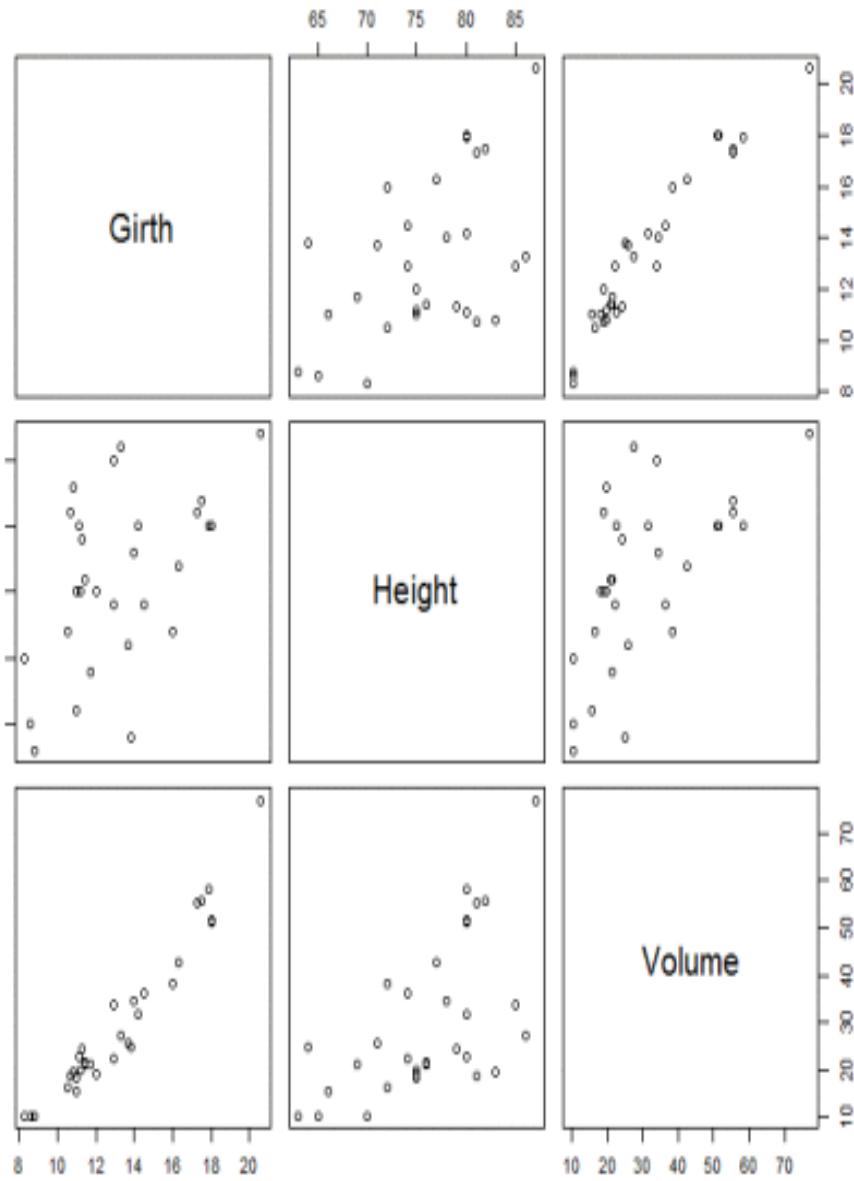
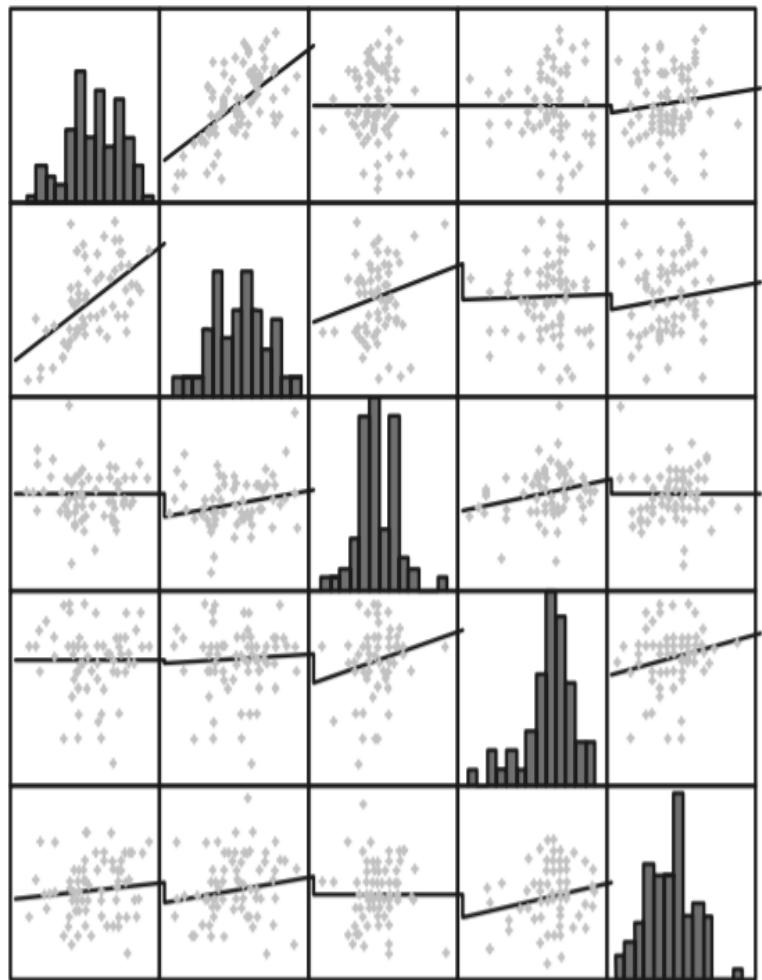


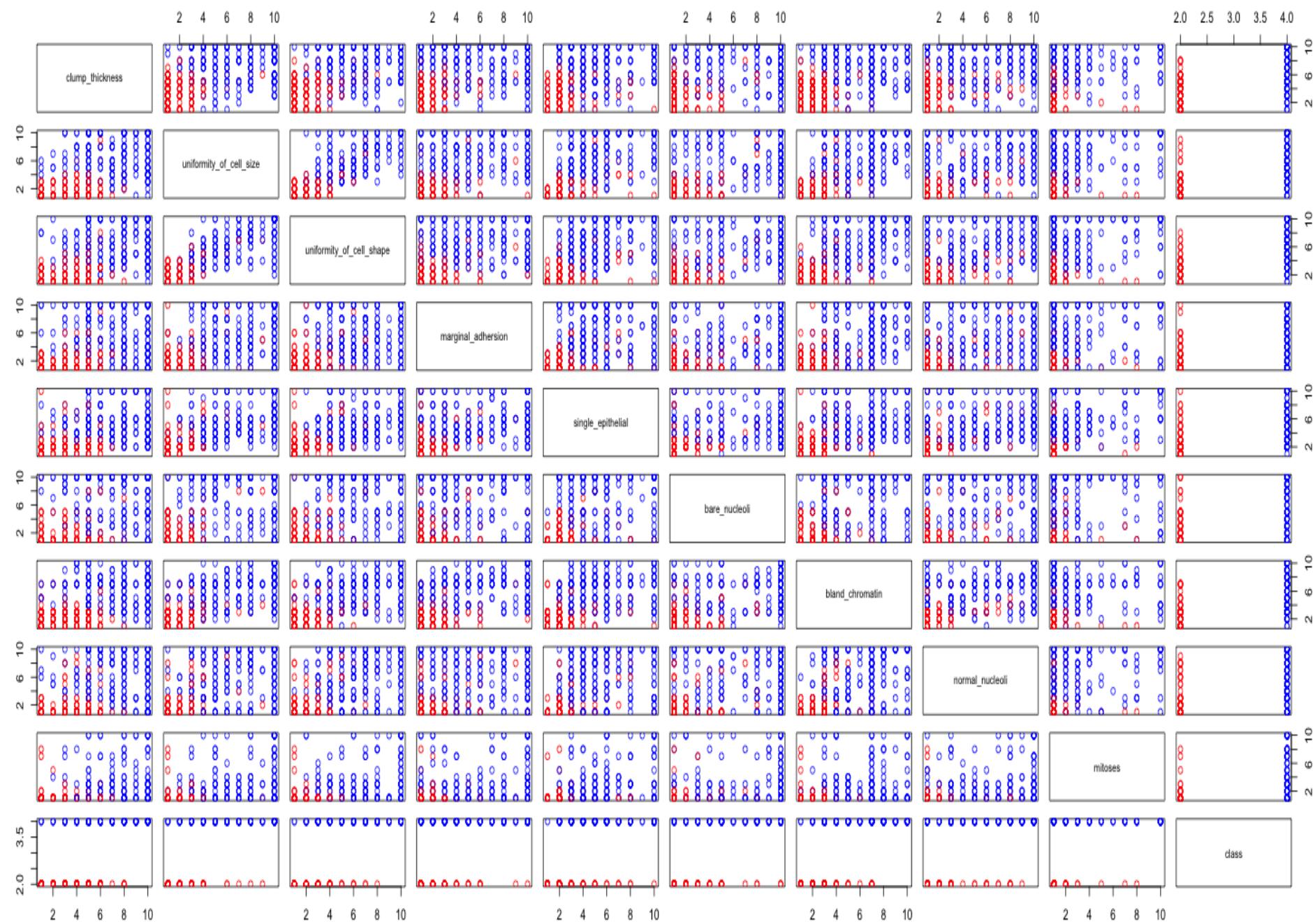


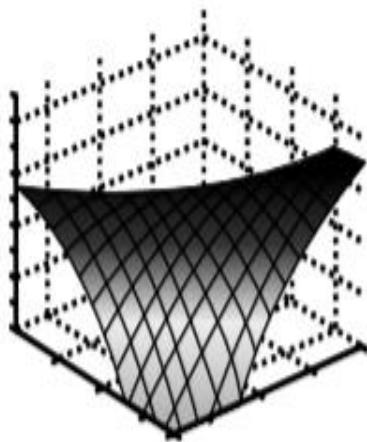
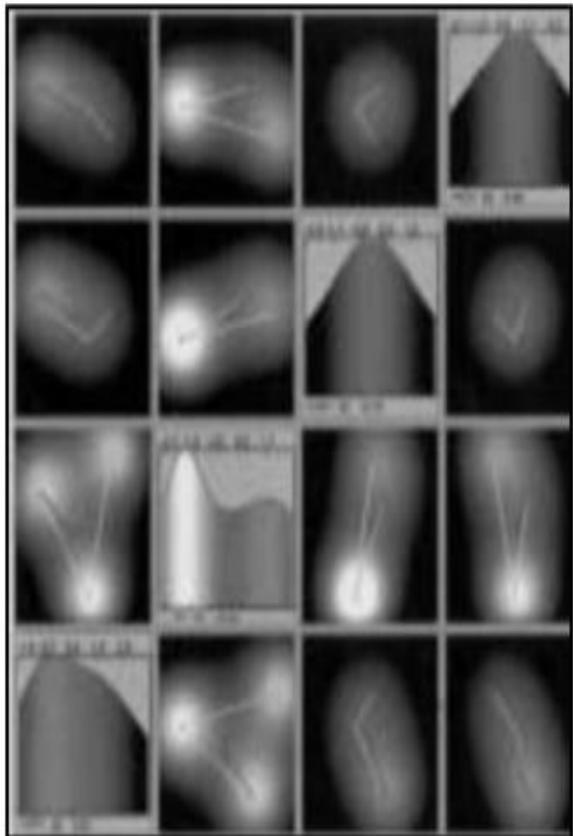
0	4	8	16	19	24	32
Версия	Длина	Тип сервиса		Общая длина		
		Идентификация	Флаги	Смещение фрагмента		
Время жизни		Протокол		Контрольная сумма заголовка		
		IPv4-адрес отправителя				
		IPv4-адрес получателя				
		Опции IPv4		Заполнение		
				Данные		

Рис. 5.16. Форматдейтаграммы

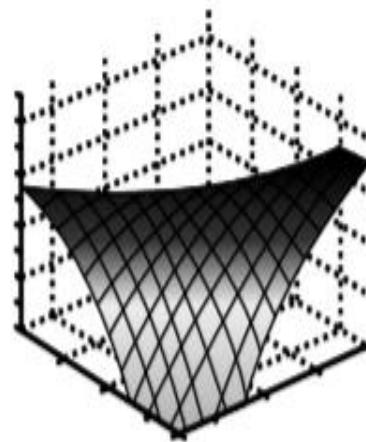




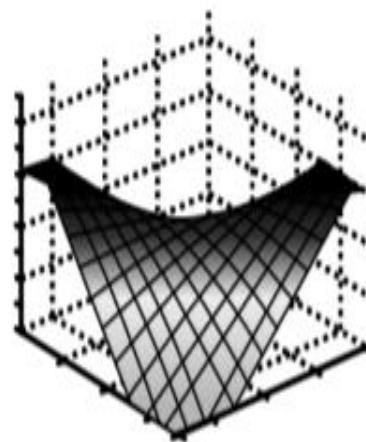
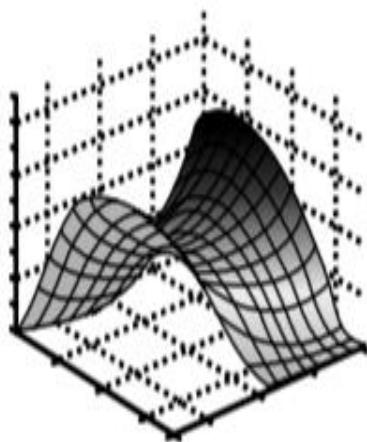


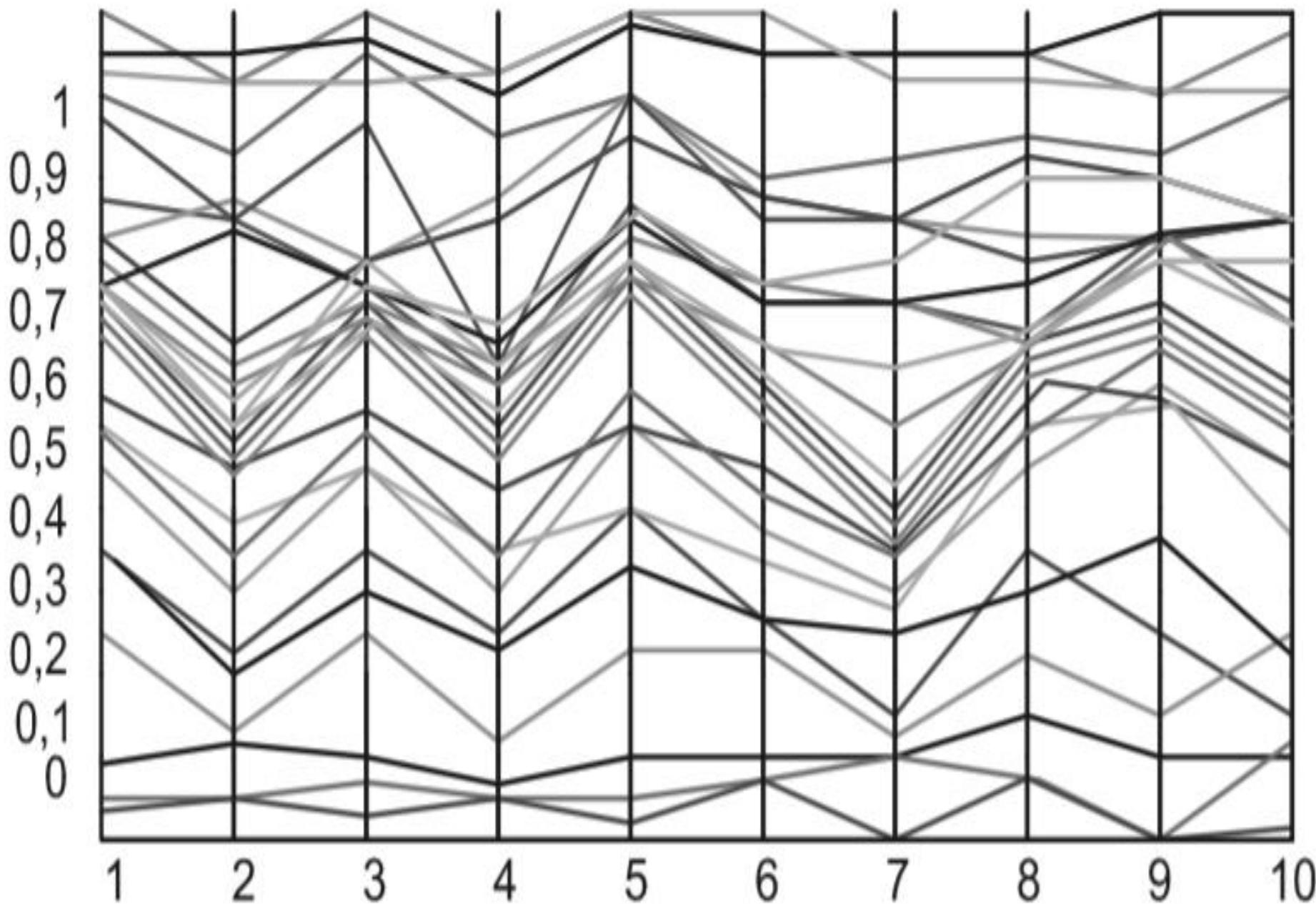


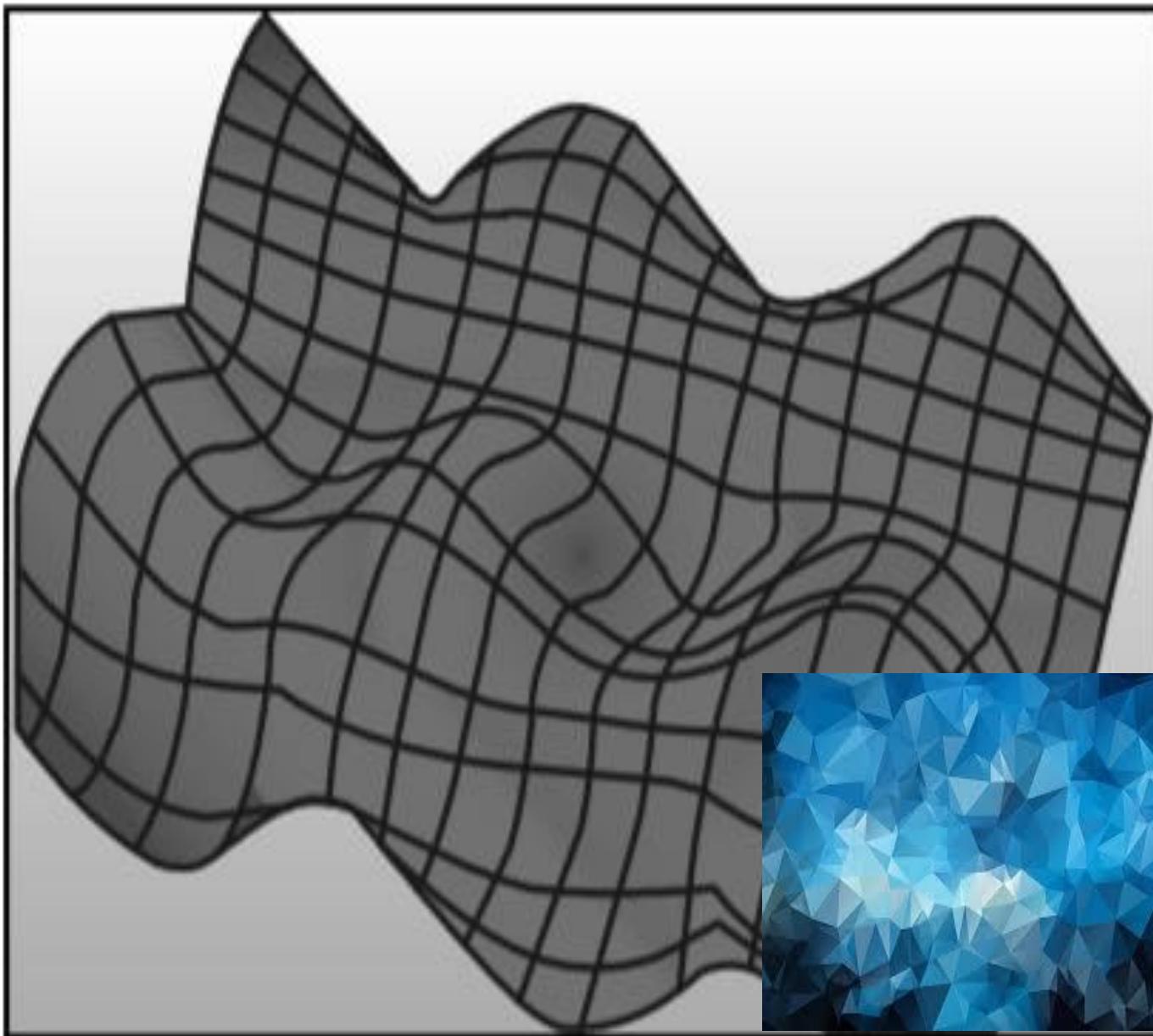
Replication 1

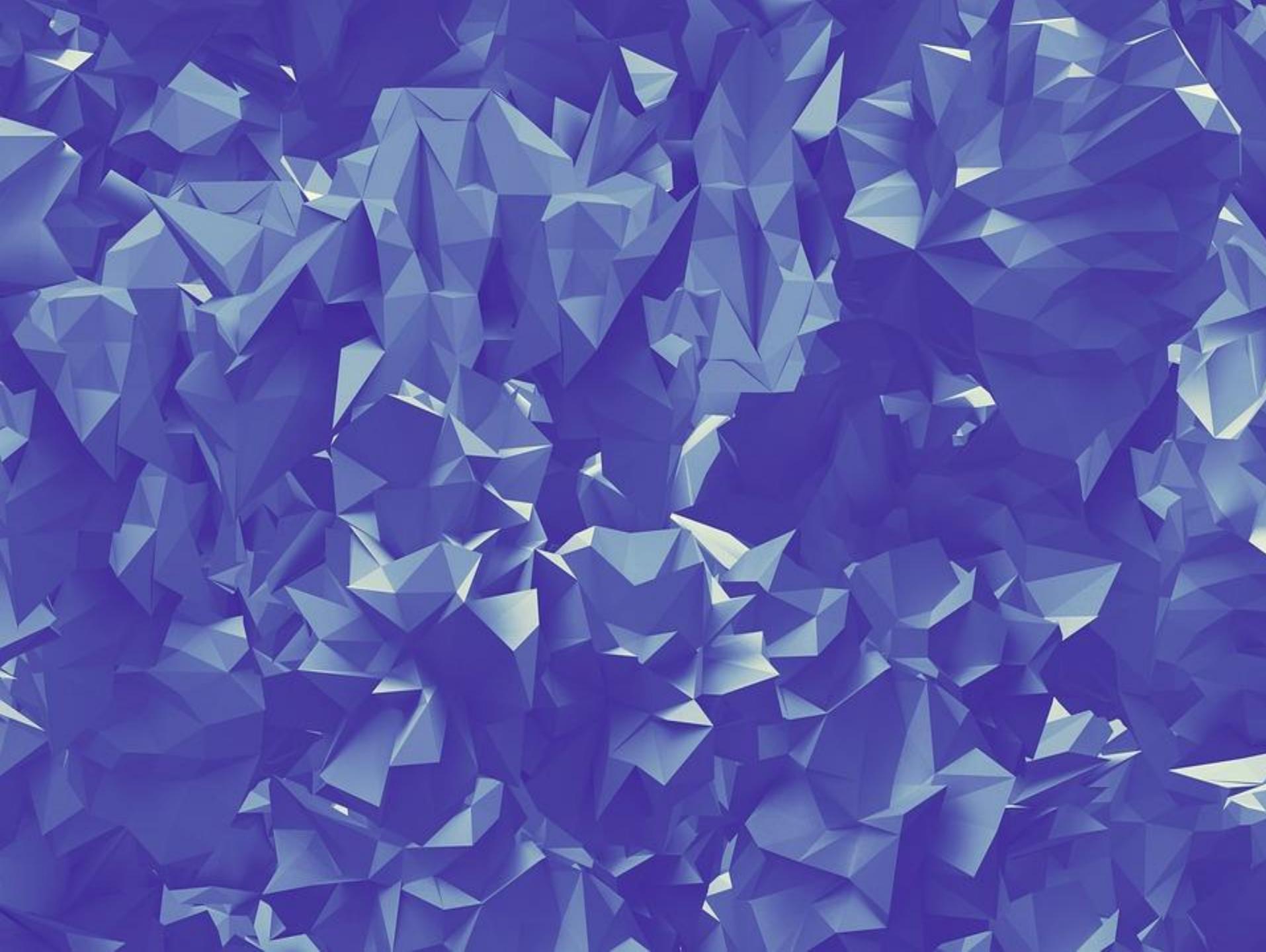


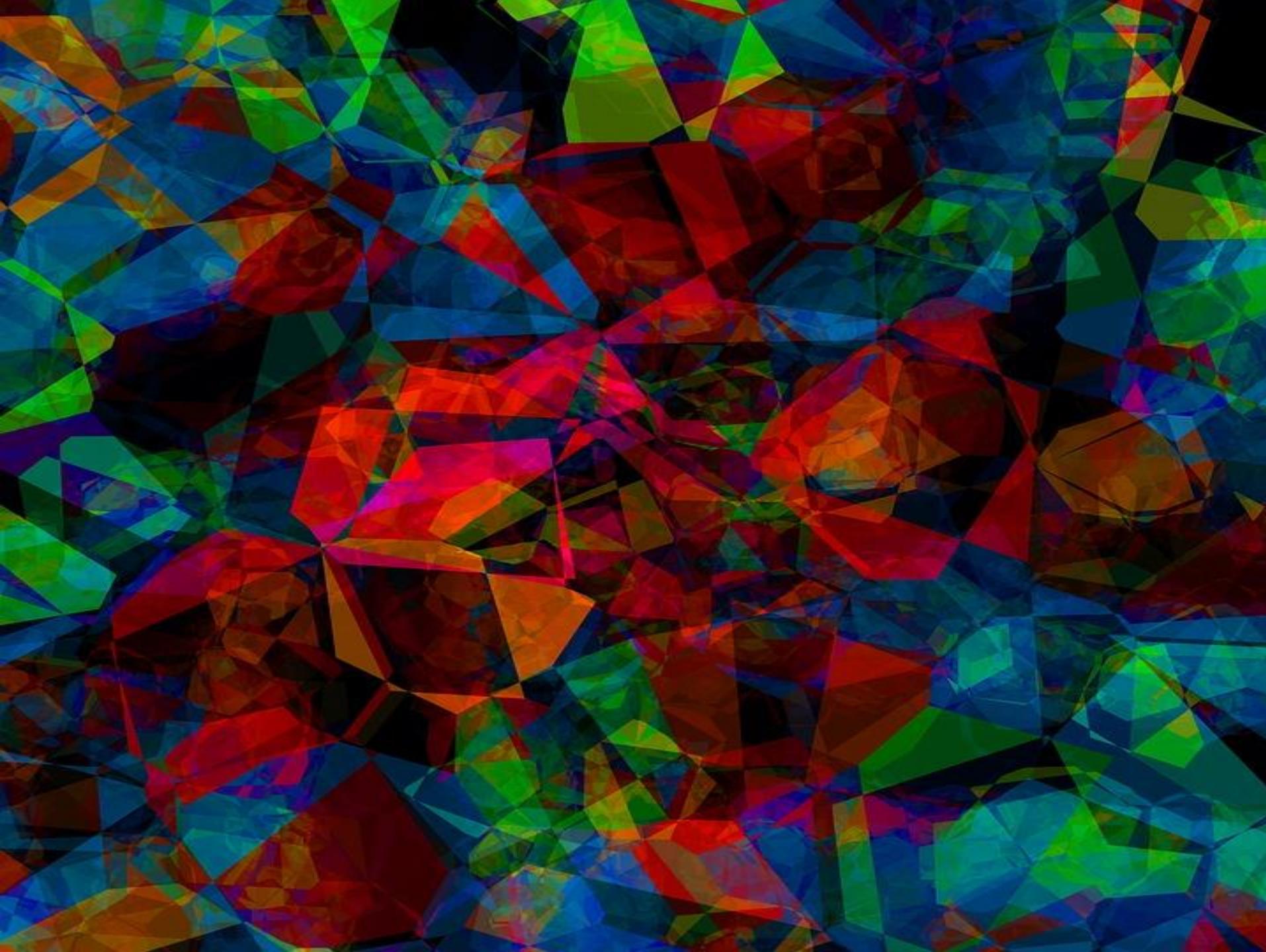
Replication 2

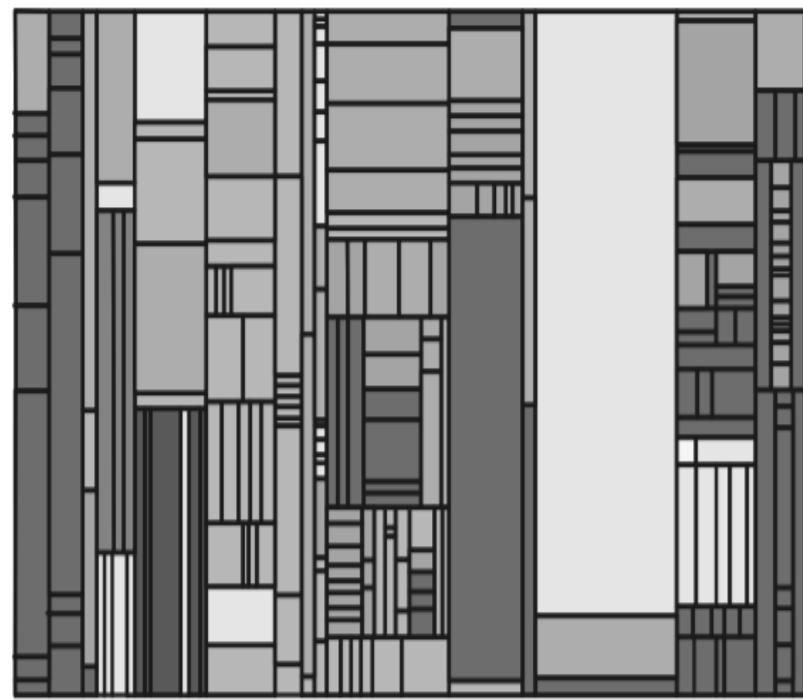
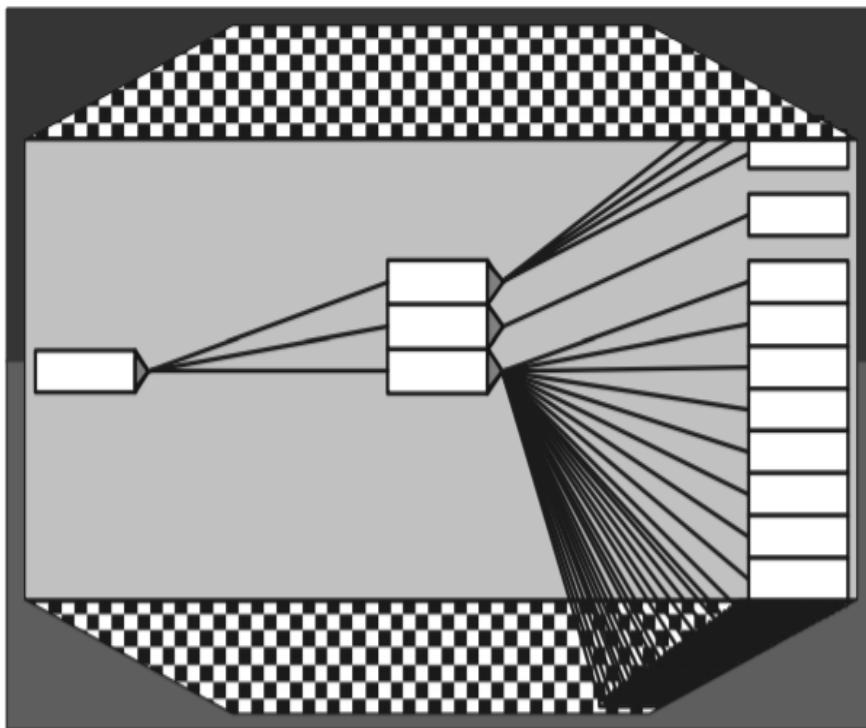
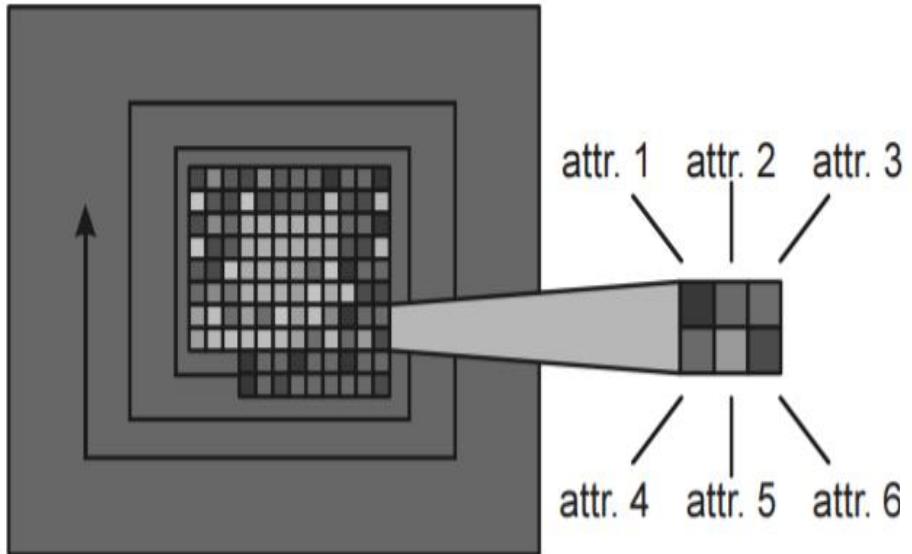




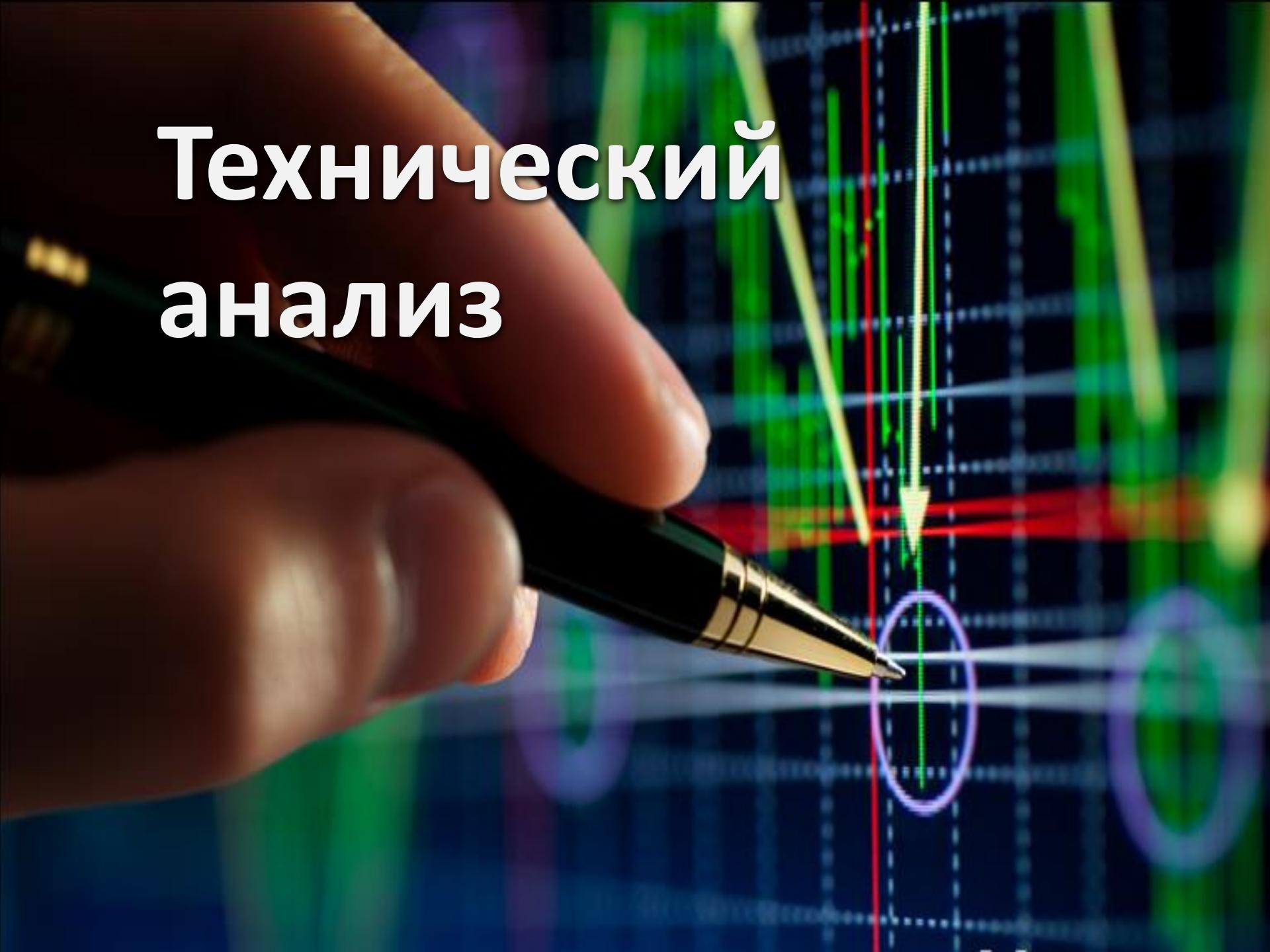








Технический анализ



Хомма Мунэхиса (1724 - 1803)

Создатель японских
свечей и первого
хедж-фонда



EURUSD

1

30

1h

15



FACEBOOK INC, D, BATS

O 81.41 H 81.52 L 80.18 C 80.42

loading...

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



British Pound/Japanese Yen, D, FX

O 187.882 H 188.601 L 187.643 C 187.647

clos

Vol.(20)

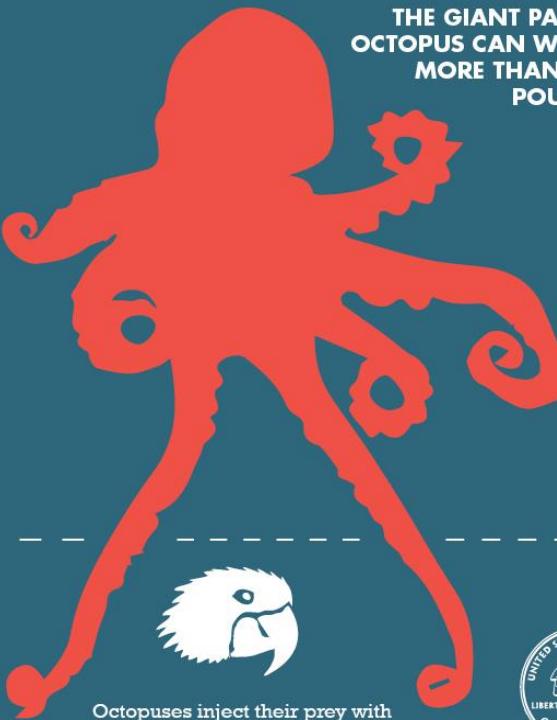


n/a n/a





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

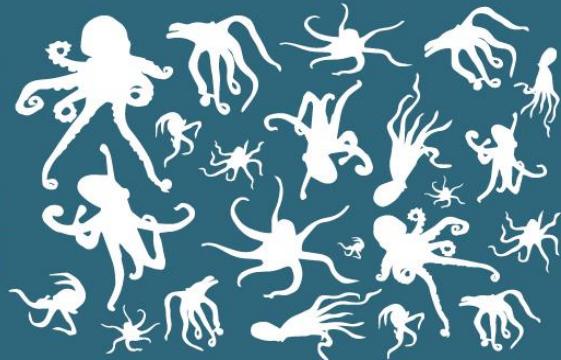


A mature female octopus can have up to 280 suckers on each arm! Each sucker contains thousands of chemical receptors, with sensitivities to both touch and taste.

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



Menu File New Pidchart Save Preview Download Share

Shapes & Lines

Icons

Search icon here Technology All Color Mono

Unselect

Background

Text

Tools

Photos

Photo Frame

New Pidchart

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.

The screenshot shows a user interface for creating infographics. On the left, there's a sidebar with categories like 'Shapes & Lines', 'Icons' (with a search bar and a 'Technology' filter), 'Background', 'Text', 'Tools', 'Photos', and 'Photo Frame'. The main workspace displays a template with a dark background. At the top, the title 'HOW TO START' is in large white letters, followed by a subtitle 'HOW TO START AND DESIGN AN INFOGRAPHIC' in a pink box. Below this, there are two sections: 'INSPIRATION FROM AROUND THE WEB' and 'PICKING THE RIGHT COLOR SCHEME'. The 'INSPIRATION' section contains text about learning from experts. The 'COLOR SCHEME' section contains text about picking colors and includes a bar chart and a donut chart. The bar chart has values 90, 70, 50, and 20. The donut chart has segments labeled 60%, 15%, and 25%. A 'need help?' button is in the bottom right corner.

Add chart



Search by chart type

Sort by Chart type ▾

Line



Lines

Bar



Bar



Stacked



Grouped



Radial

Column



Column

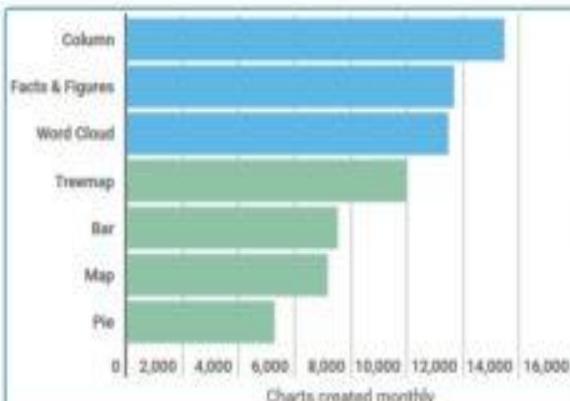


Stacked

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

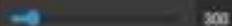
Chart type



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?



Main Title



LOREM
IPSUM

LOREM
IPSUM

Lorem ipsum dolor
sit amet, possit
scit quis presentis

LOREM
IPSUM

Lorem ipsum dolor
sit amet, possit
scit quis presentis

LOREM IPSUM
DOLOR

Lorem ipsum dolor sit amet, possit
scit quis presentis
nihil debet possit
scit quis presentis

LOREM
IPSUM

Lorem ipsum dolor
sit amet, possit
scit quis presentis

LOREM
IPSUM

Lorem ipsum dolor
sit amet, possit
scit quis presentis

LOREM
IPSUM

LOREM
IPSUM

LOREM
IPSUM

LOREM
IPSUM

SHOW MORE

- 100% +

?



Поиск среди 1 000 000 изображений...

Logo & Icons

21

Интервалы

Копировать

Сортировка

X

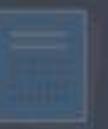
P



Бесплатные

Сети

Дизайн



Фигуры

Линии

Упаковка



Знаки

Диаграммы

КМС-дизайн



Поиск

Basic Shapes

Arrows

Block Shapes

Stars And Bullets

Social Media Icons

Infographics



Column Chart



Bar Chart



Line Chart



Area Chart



Pie Chart



Donut Chart



Bubble Chart

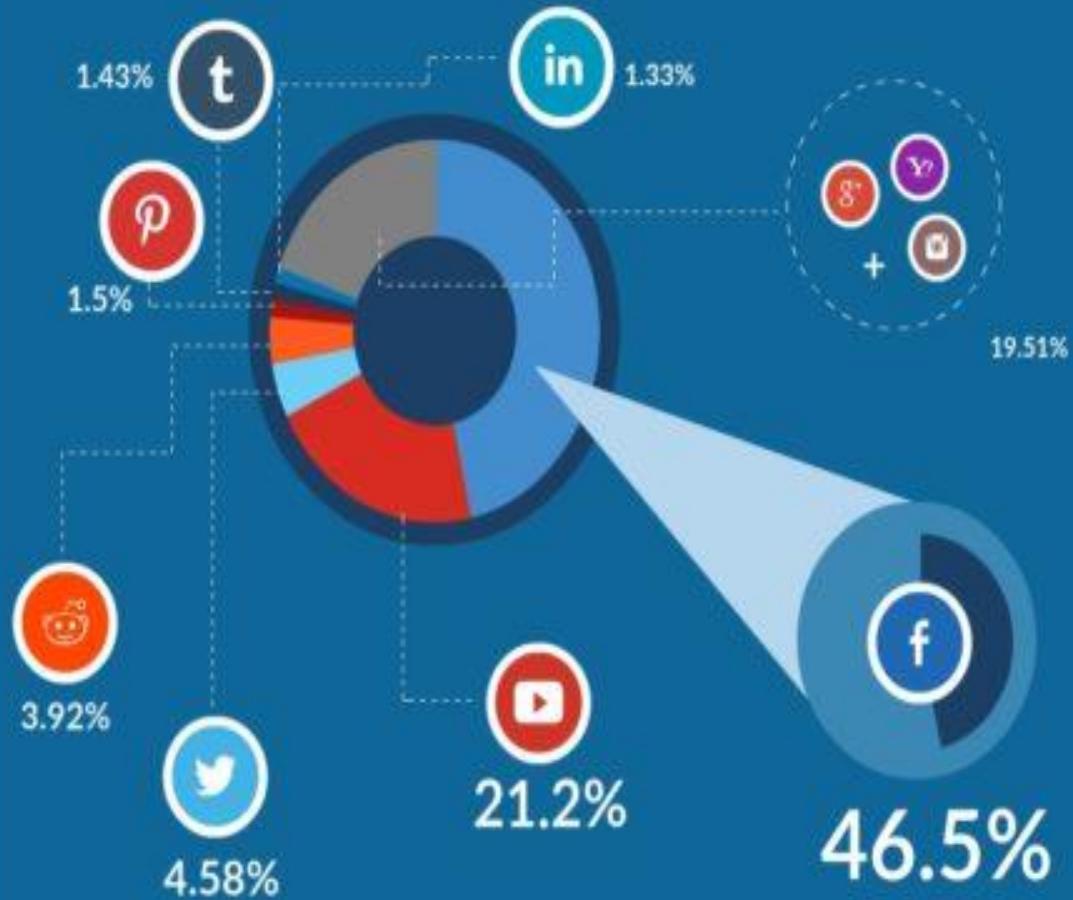
Percentage Bar



Demographic Repeater

+ Получить более обще...

MOST POPULAR SOCIAL MEDIA WEBSITES IN THE UNITED STATES IN 2015



Импорт

Экспорт

Публикация

100%

Сообщество Задачи Комментарии Публикации Справка

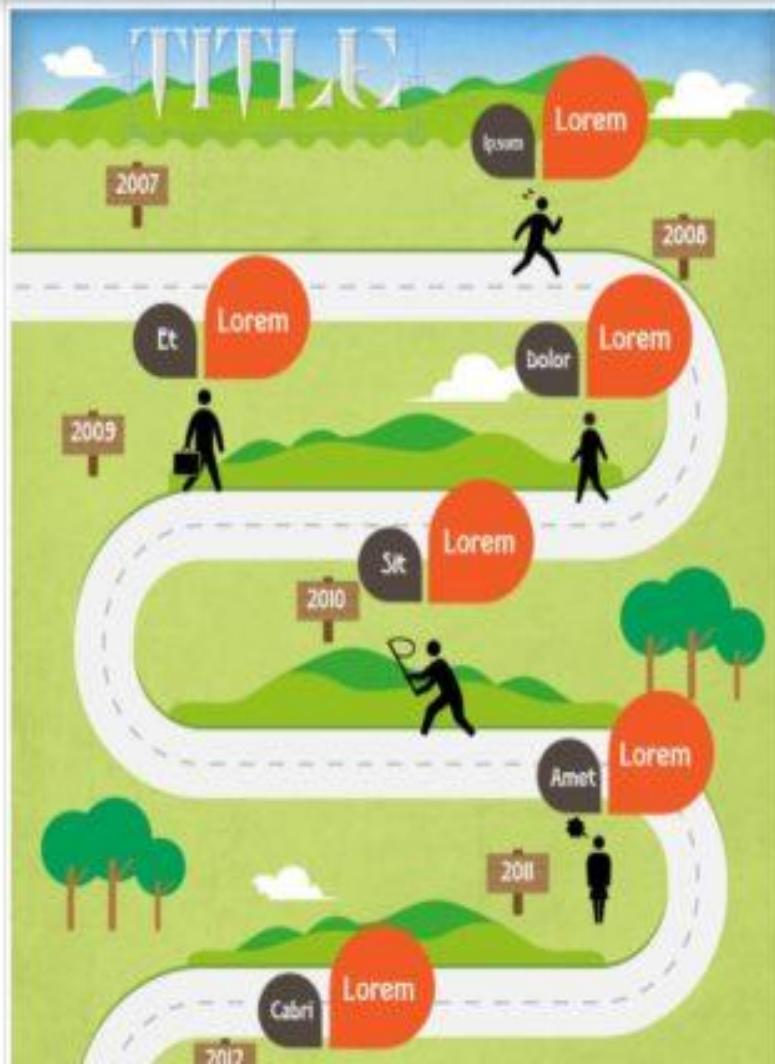
[Templates](#)[Objects](#)[Media](#)[Backgrounds](#)[Lines](#)[Shapes](#)[Text](#)[Image](#)[Upload](#)

Zoom 100%

grid

undo

redo

[Present](#)My Area
Area, colors...[Bar](#)[Column](#)[Line](#)[Radar](#)Want more charts? [Go Pro!](#)[Chat with us](#)

Untitled Project Save As... Preview Publish Upgrade

Objects (65) Trebuchet MS 48 + B I U A A

All Shapes Recently Used Lines

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

W: 482 A: 275 C: 81 H: 116 Y: 48

STARTING A DESIGN PROJECT

Lore ipsum

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

Lore ipsum

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

AI EPS EPS PSD

PPT CGM PNG JPG

Lore ipsum Lore ipsum

Lore ipsum Lore ipsum

SLIDES

- 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
- RADAR



Источники

- <http://infographer.ru/tag/parallelnye-koordinaty/>
- <http://kek.ksu.ru/eos/WM/AnalizDannihProcessov.pdf>
- <https://habr.com/ru/company/ods/blog/323210/>
- https://ru.wikipedia.org/wiki/Интеллектуальный_анализ_текста
- <http://datareview.info/article/osnovnyie-tehnologii-text-mining/>
- <https://binguru.net/>