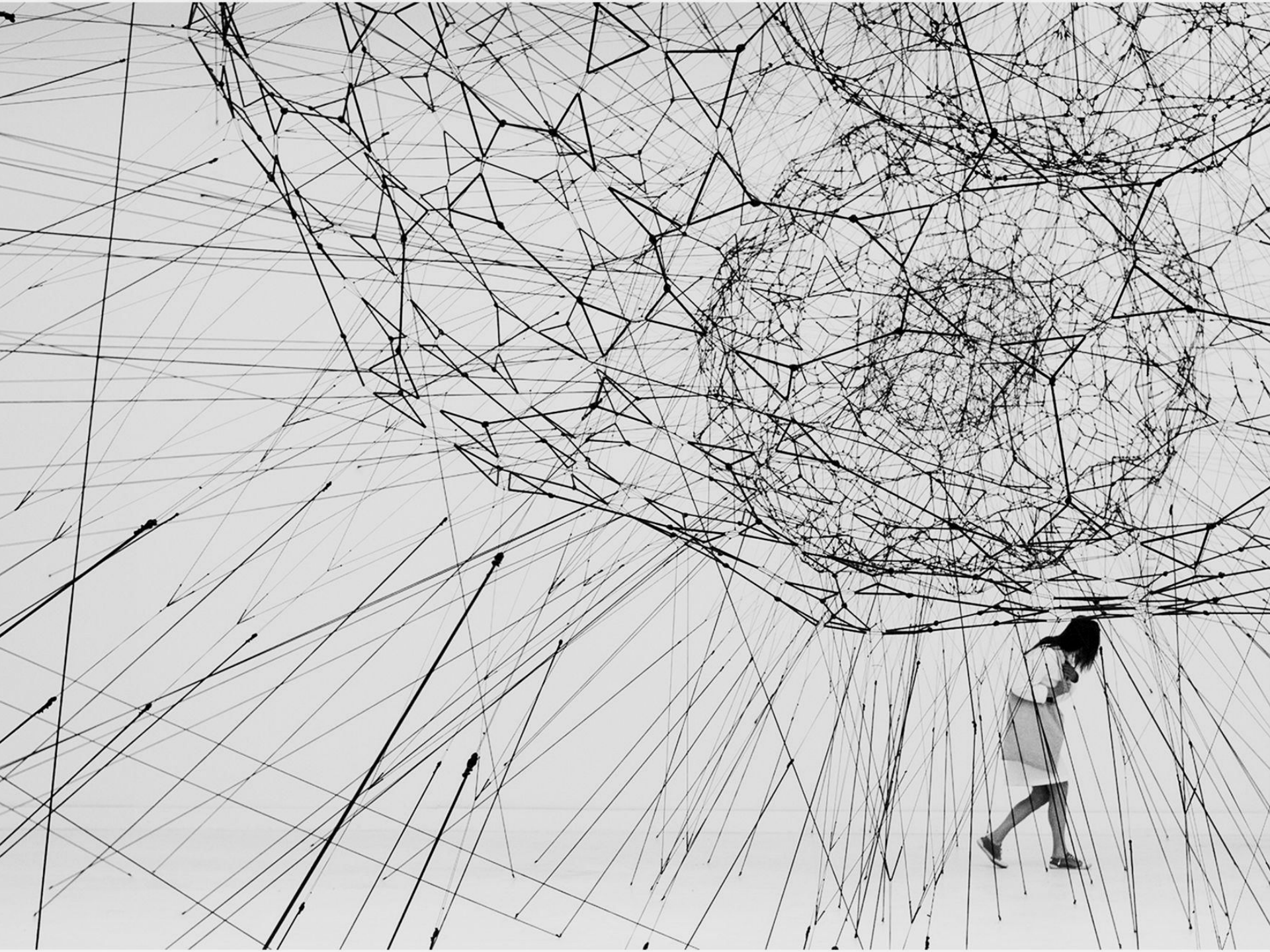




Data mining

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

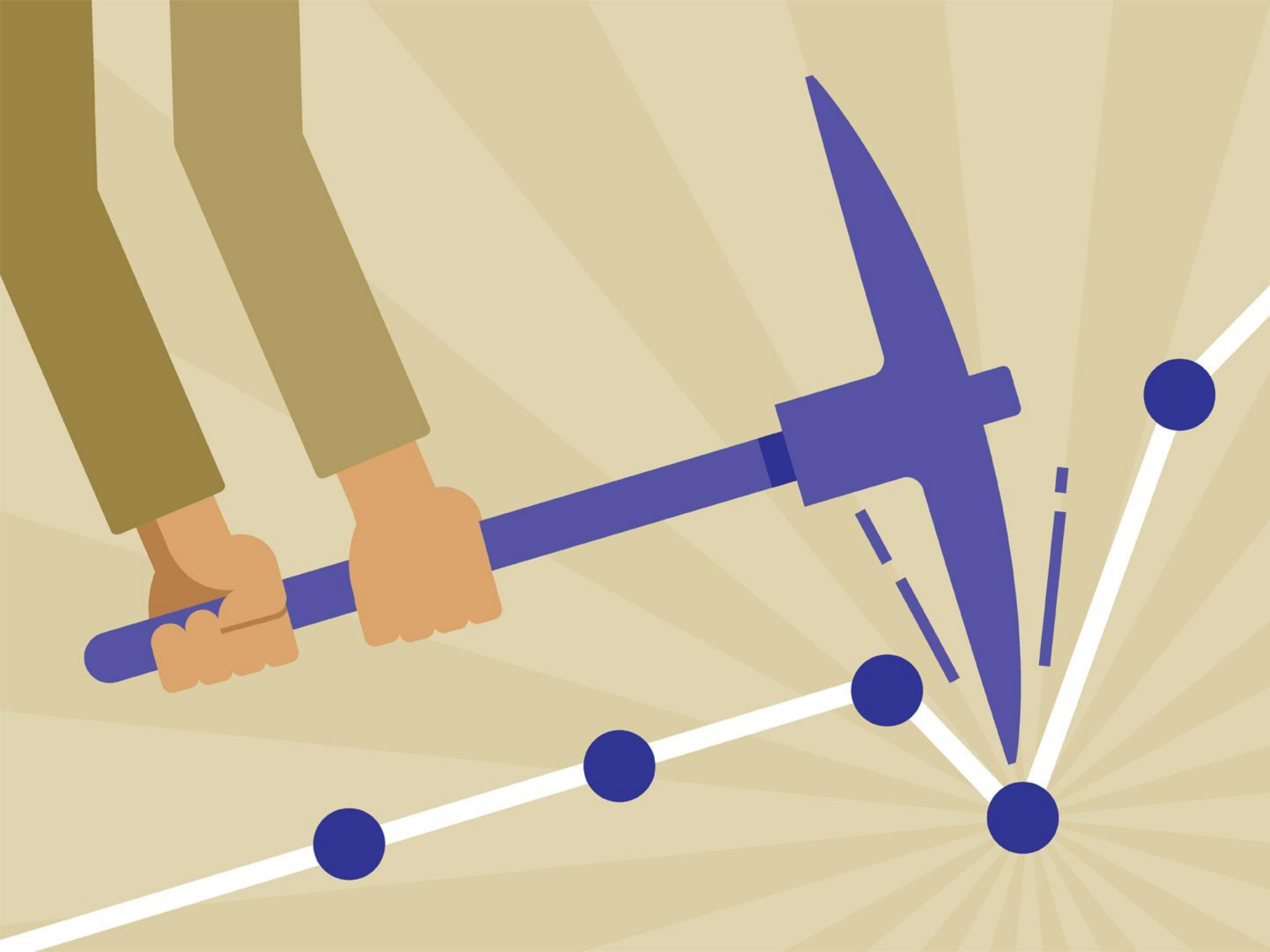


Data mining

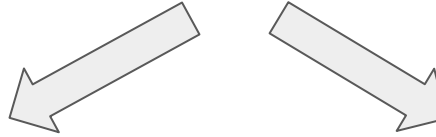
- Витяг, збір даних, видобуток даних (ще використовують Information Retrieval або IR);
- Витяг знань, інтелектуальний аналіз даних (Knowledge Data Discovery або KDD, Business Intelligence).

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)



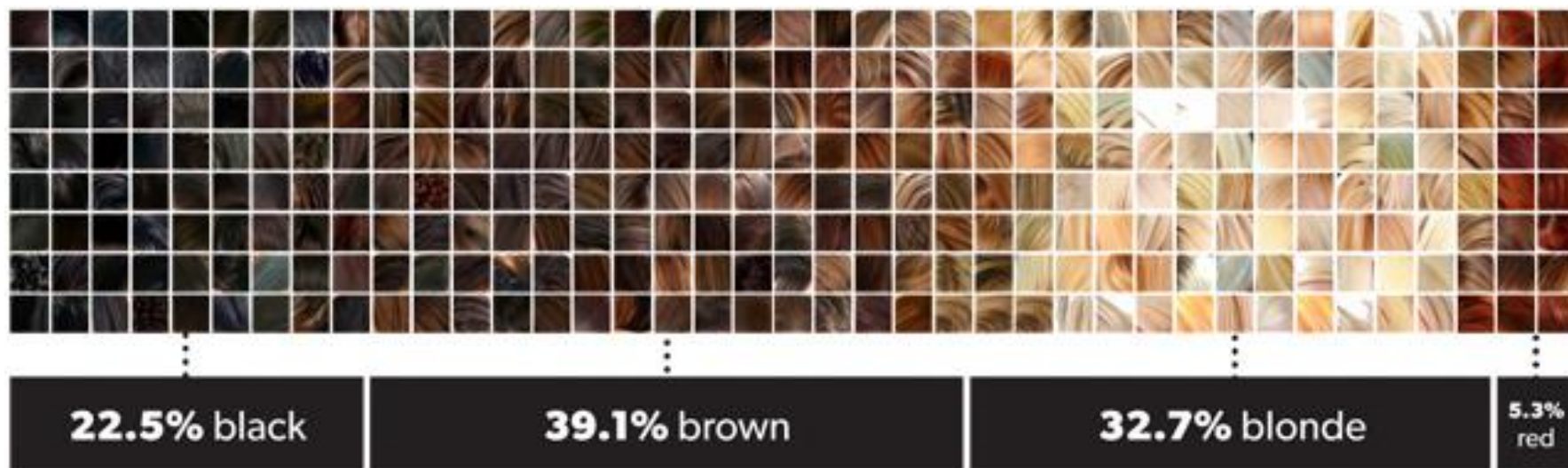
STRUCTURE
QUALITY
PHASE
DATA
ANALYSIS
MEASUREMENT
FOCUSES
TECHNIQUES
EXPLORATORY
PREDICTIVE
STATISTICAL
BUSINESS
INFORMATION
MODELS
ASSESSED
HYPOTHESES
INSTRUMENTS
CHECKED
DISCOVERY
ACCOUNT
PEOPLE
WAYS
FINDINGS
ONE
STATISTICS
APPROACHES
IMPORTANT
FREQUENCY
CLOSING
MUST
TYPES
USING
LEVEL
PERFORMED
SUPPORTING
TRANSFORMING
CHARACTERISTICS
PLOTS
NORMALITY
USED
SPECIAL
SCATTER
APPROACH
CONFIRMATORY
KURTOSIS
EITHER
RELIABLE
MAKING
MEDIAN
SEVERAL
FACETS
SCIENCE
ADOPTED
DIVIDE
TEXTUAL
ORIGINAL
USUALLY
SUBGROUPS
NECESSARY
FINAL
STAGE
LOOK
HARD
TAKE
REPRODUCIBLE
CLOSELY
POSSIBLE
EXISTING
STRUCTURAL
ED
PLAN
TWO
MODELING
ORIGINAL
TEXTUAL
DIVIDE
FACETS
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_есс

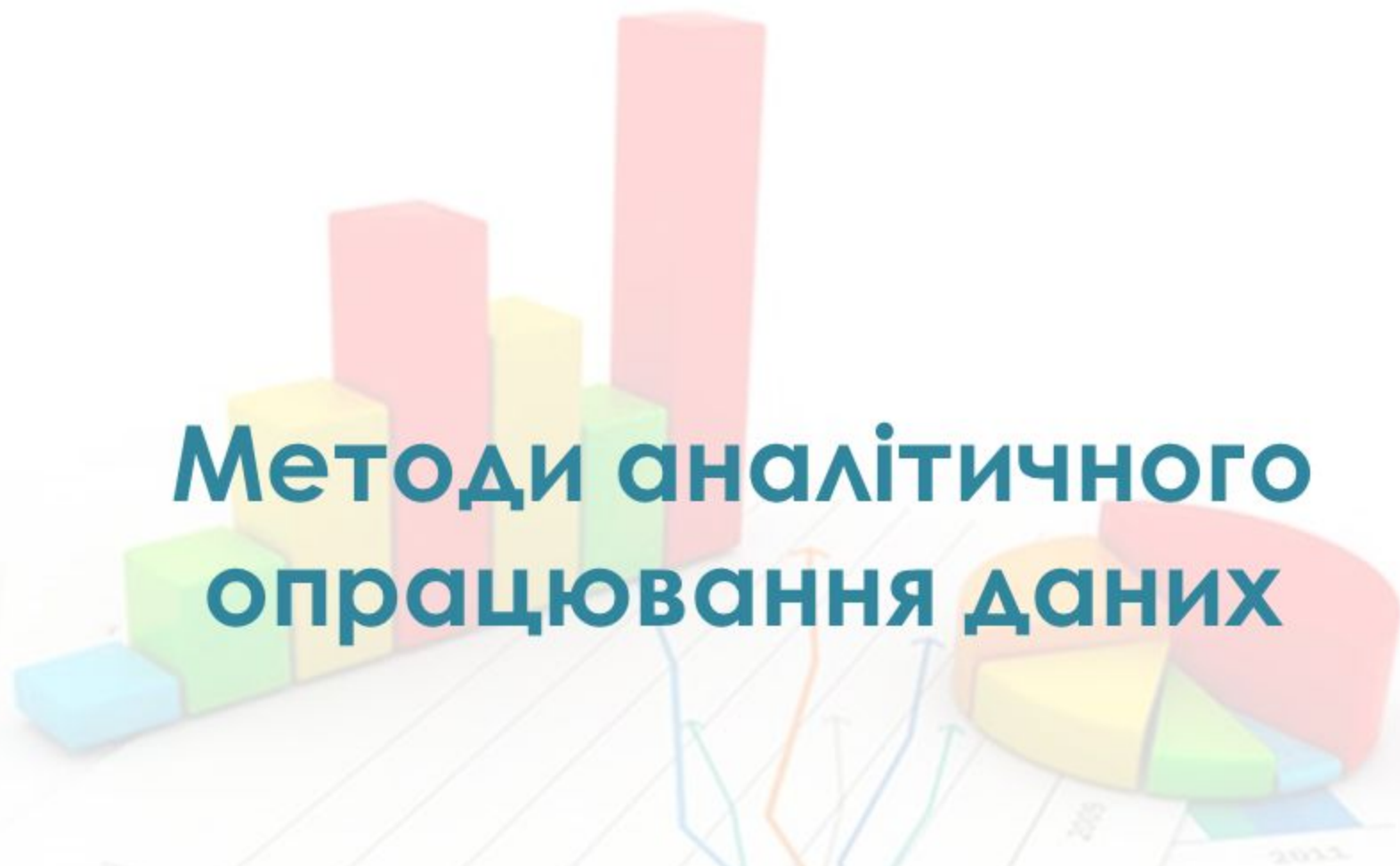
Facial morphs of 10 of the most popular adult performers



Посилання: jonmillward.com

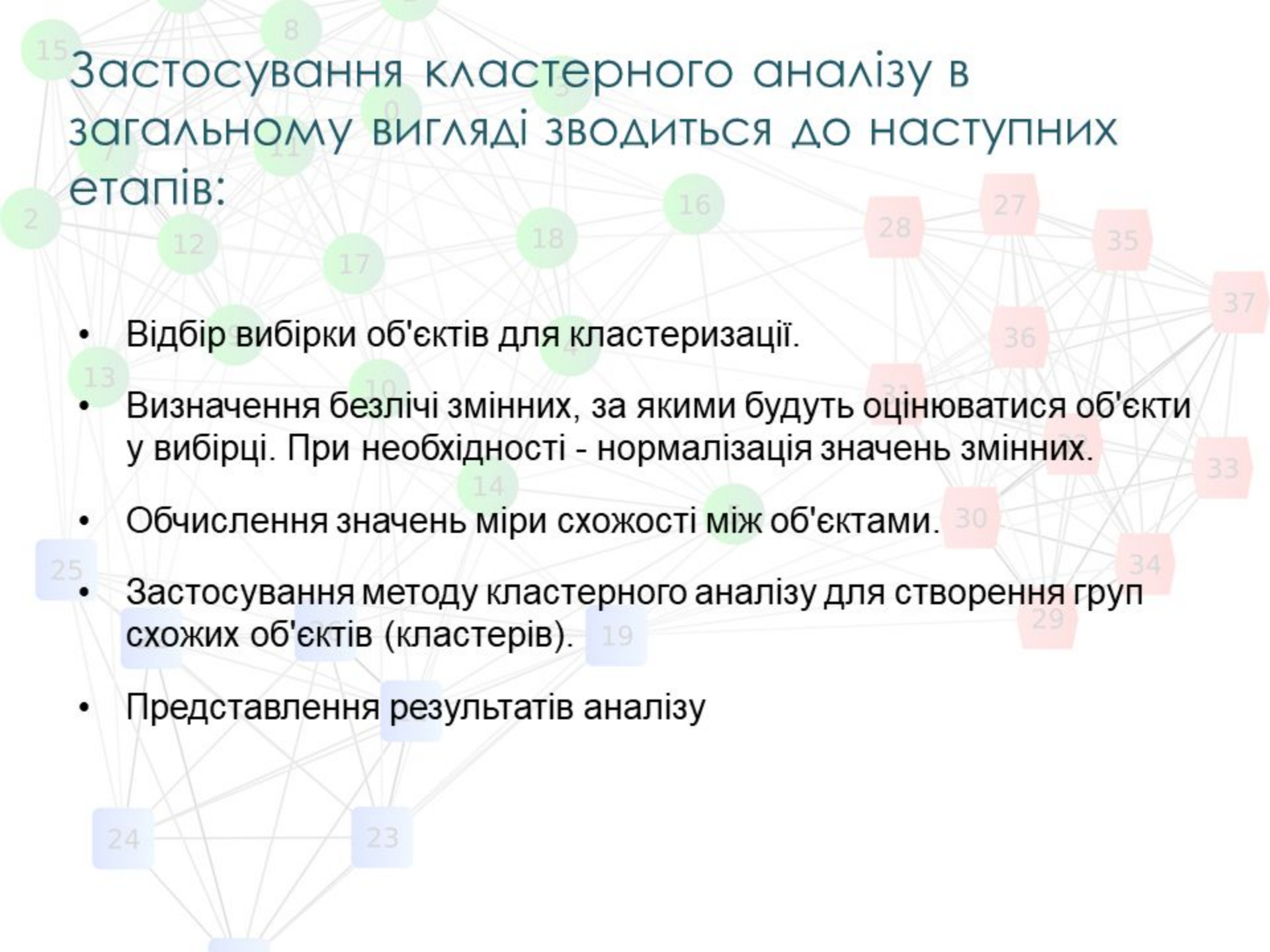


Методи аналітичного опрацювання даних





Кластерування даних

A background network diagram featuring a complex web of nodes and edges. The nodes are represented by various shapes and colors: green circles (e.g., 15, 8, 12, 17, 18, 16, 13, 14, 10, 7, 0, 2), red pentagons (e.g., 28, 27, 35, 36, 37, 33, 34, 29, 30), and blue squares (e.g., 25, 19, 24, 23). The edges are thin, light gray lines connecting the nodes, creating a dense, interconnected pattern.

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 виконують нормалізацію, щоб всі компоненти давали однаковий 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 (Мангеттенська відстань)

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

- Відстань Чебишева

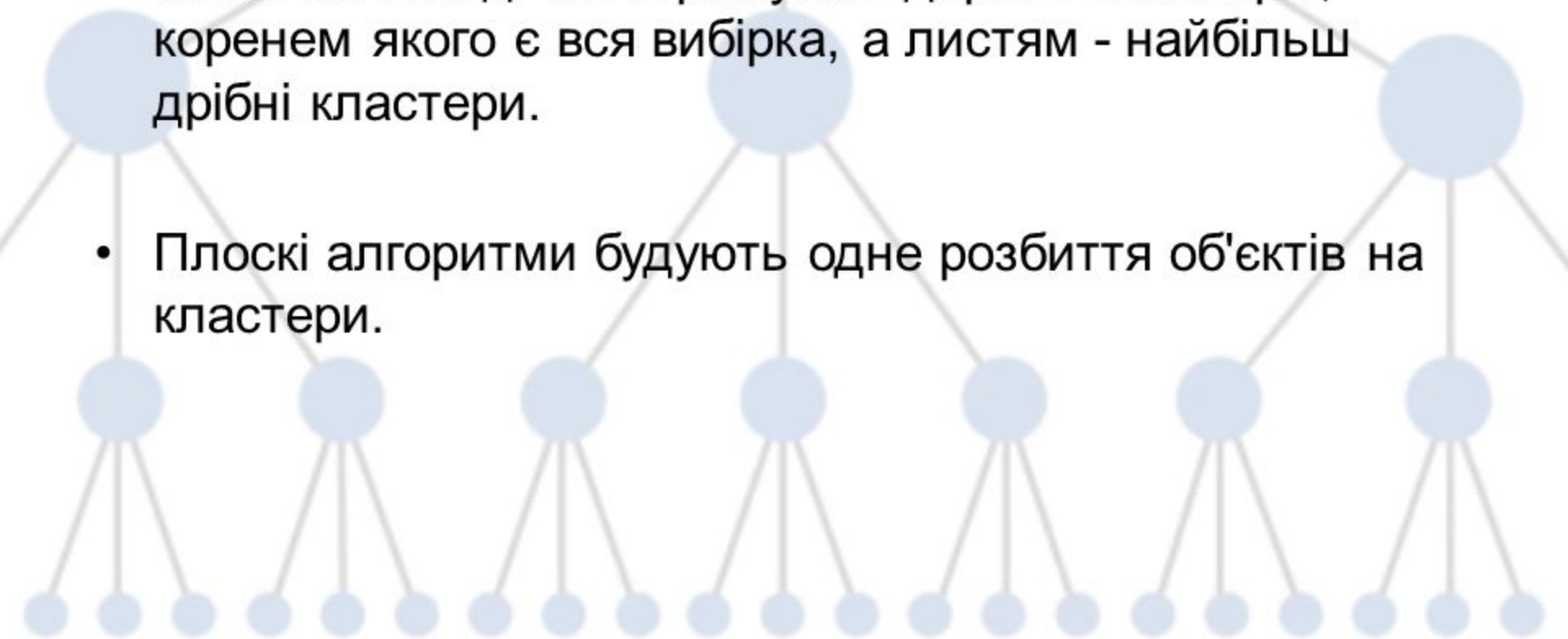
$$\rho(x, x') = \max(|x_i - x'_i|)$$

- 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 к-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?

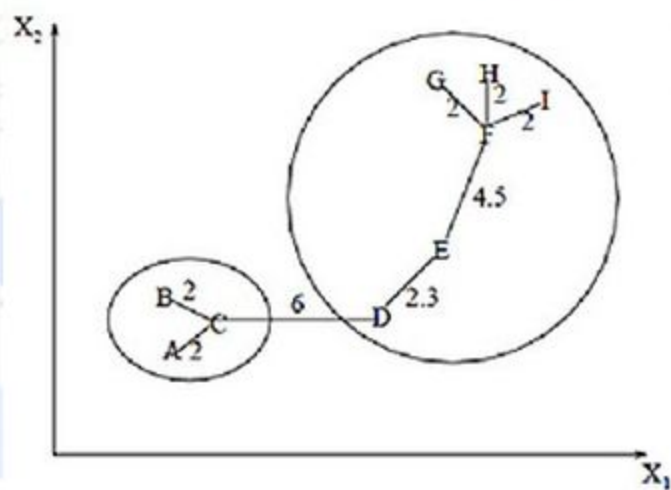
- Метод к-середніх вибирає точки багатовимірного простору, які будуть представляти к-кластери. Ці точки називаються центрами тяжіння. Перший раз, за відсутності припущень, центри тяжіння можна вибирати випадково
- Кожен пацієнт буде розташовуватися найближче до однієї з точок.
- Тепер у нас є к-кластерів, і кожна точка- це член якогось з них.
- Метод к-середніх, враховуючи положення членів кластера, знаходить центр кожного з к-кластерів. Обчислений центр стає новим центром тяжіння кластера.
- Оскільки центр ваги перемістився, точкою могли виявитися ближче 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 дерева спочатку будує 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

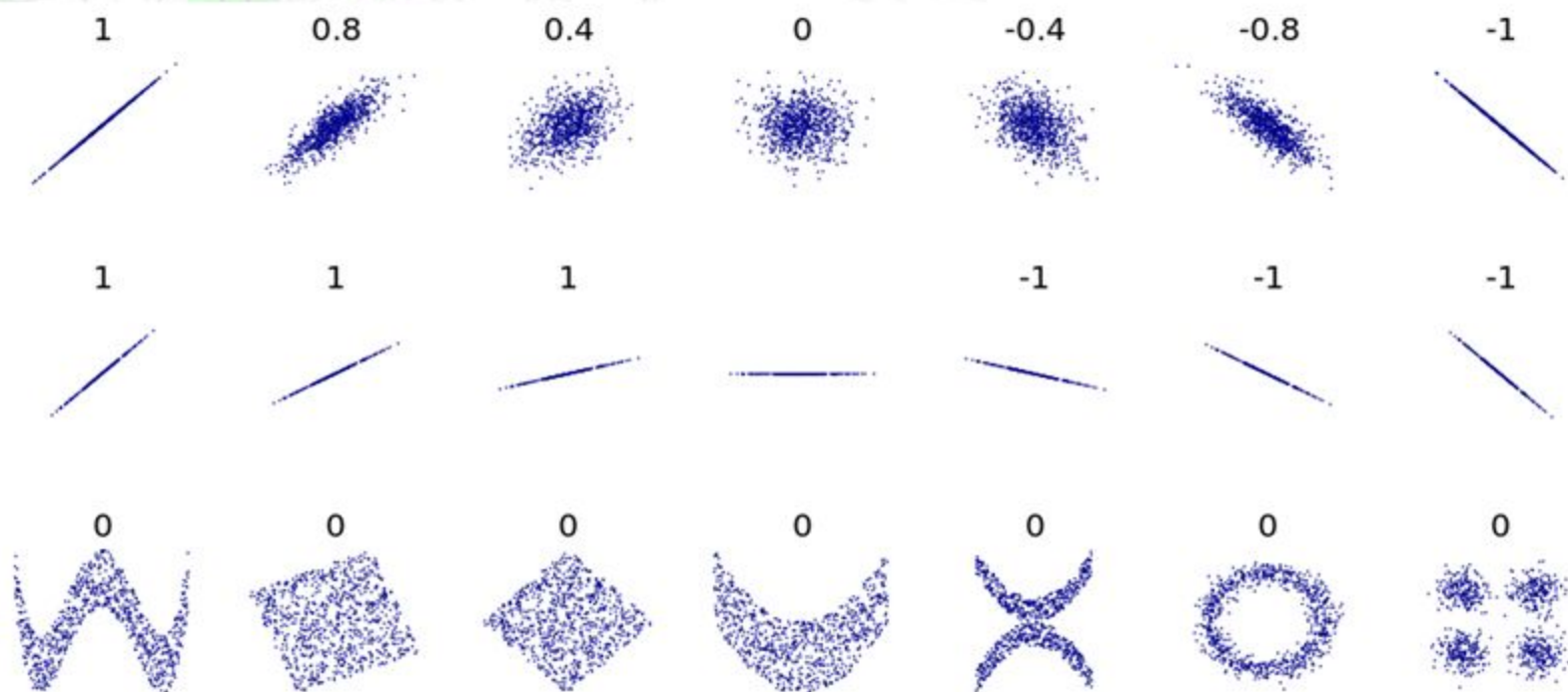
Порівняння деяких алгоритмів кластеризації

Алгоритм кластеризації	Обчислювальна складність
Ієрархічний	$O(n^2)$
к-середніх	$O(nkl)$, где k – число кластерів, l – число ітерацій
с-середніх	
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 ієрархії	Бінарне дерево кластерів
k-середніх	Гіперсфера	Число кластерів	Центри кластерів
c-середніх	Гіперсфера	Число кластерів, ступень нечіткості	Центри кластерів, матриця приналежності
Виділення зв'язних компонент	Довільна	Порог відстані R	Деревоподібна структура кластерів
Мінімальне покриваюче дерево	Довільна	Число кластерів и порог відстані для видалення ребер	Деревоподібна структура кластерів
Пошарова кластеризація	Довільна	Полідовність границь відстані	Деревоподібна 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) існує поняття коваріації або кореляційного 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 величину:

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

Для дискретних величин кореляційний момент можна знайти наступним чином:

$$\begin{aligned} K_{xy} &= \sum_i^n \sum_j^n (x_i - m_x) (y_j - m_y) p_{ij} \\ &= \frac{1}{n} \sum_i^n \sum_j^n (x_i - m_x) (y_j - m_y) \end{aligned}$$

Також на практиці зазвичай використовують іншу формулу, 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 рукопискань

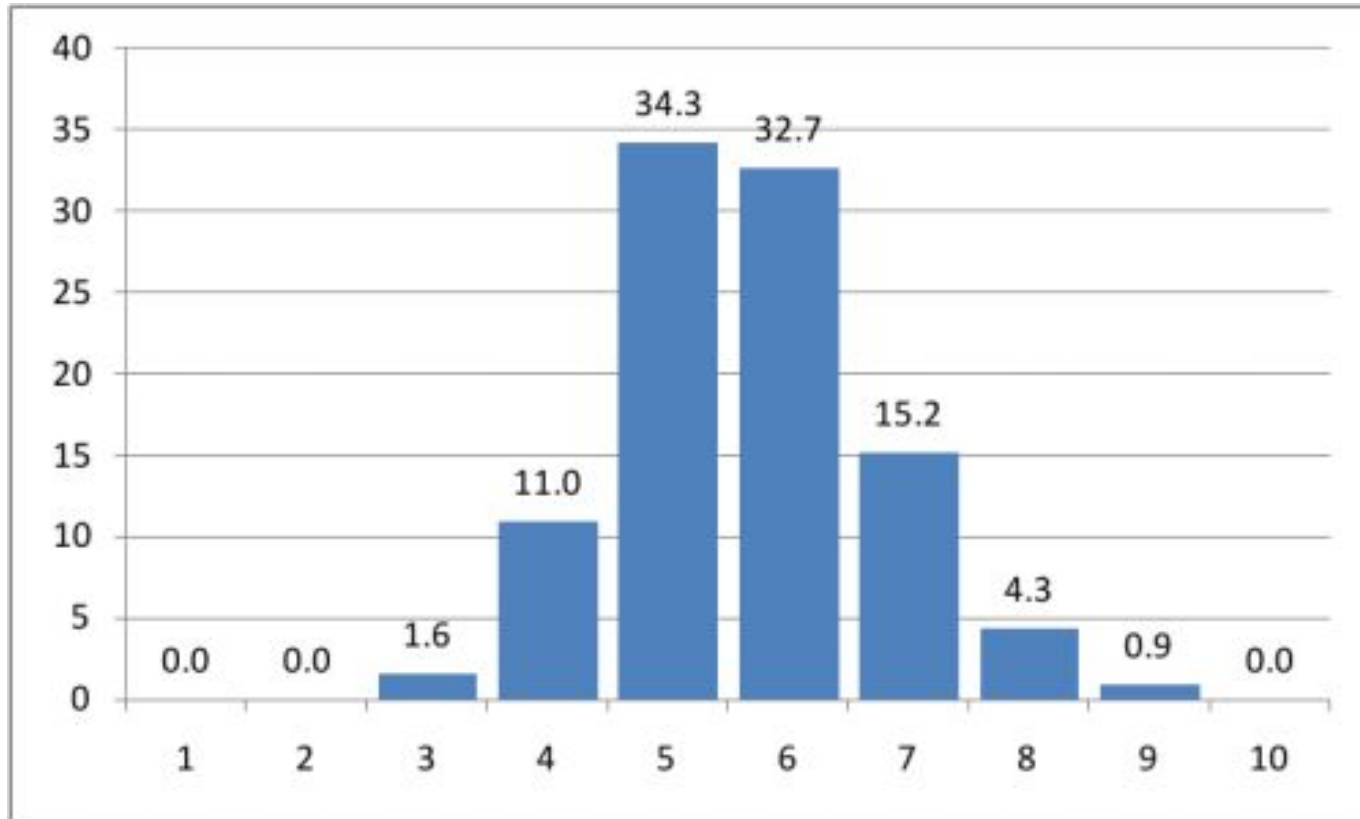


В чому полягає теорія шести рукошестикань?

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



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

