Методи передобробки та кластеризації часових рядів.

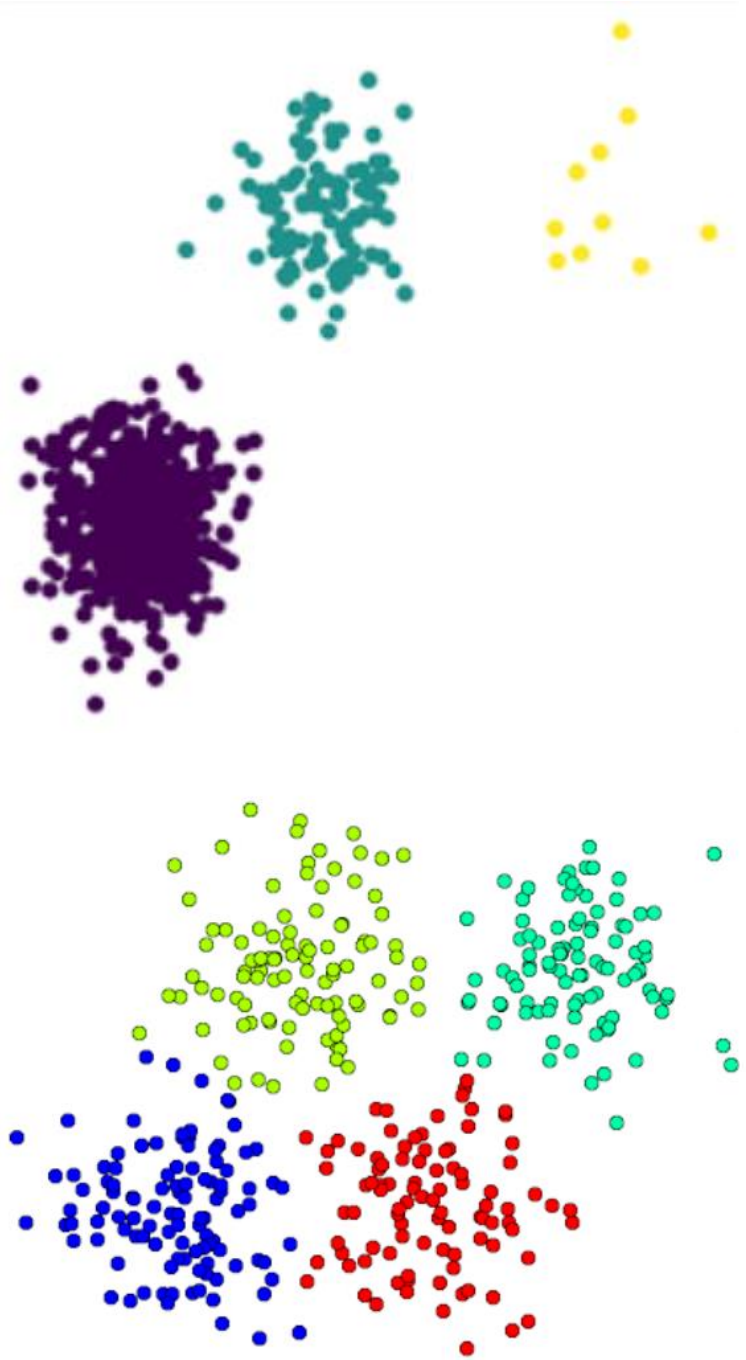
Купенко Ольга Петрівна

професор кафедри штучного інтелекту

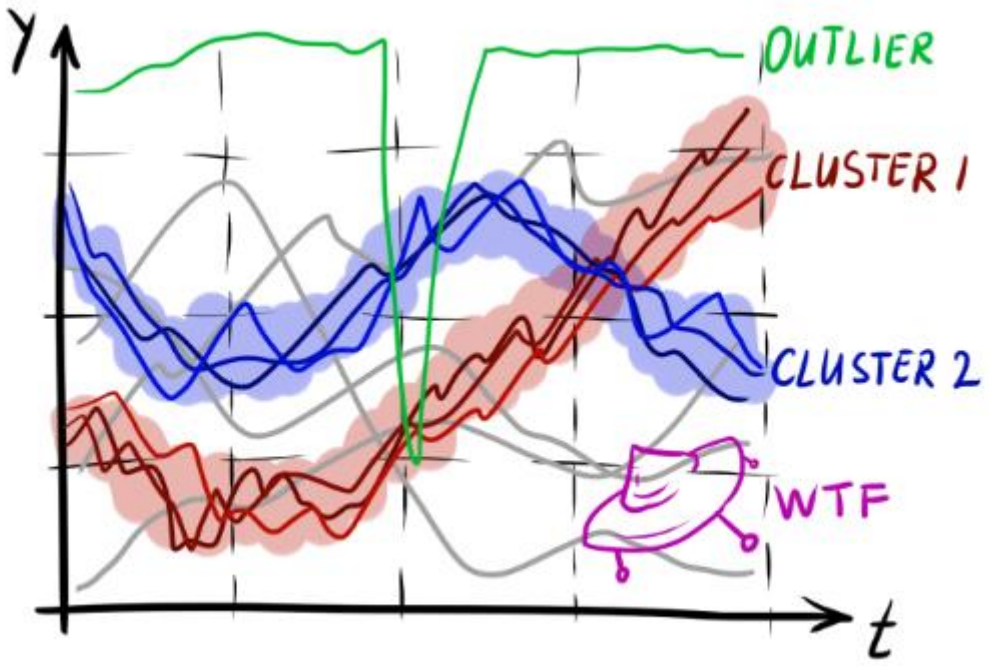
НТУУ «КПІ імені Ігоря Сікорського»

Що таке кластеризація?

- Кластеризація — це задача машинного навчання без учителя, що має за мету **розбити набір немаркованих об’єктів даних на однорідні** групи або кластери, у яких об’єкти мають максимальну схожість з іншими об’єктами в групі та мінімальну схожість з об’єктами в інших групах.
- Оскільки йдеться про навчання без учителя, ми **не маємо жодних глибоких знань** щодо визначень або класифікацій груп.
- Цей підхід можна використовувати для інтелектуального аналізу даних, **ідентифікації структур та паттернів у немаркованих наборах даних**, узагальнення даних або як етап попередньої обробки в більш складній системі моделювання.



- Кластеризація часових рядів отримала переваги від удосконалень за останні роки, зберігаючи таким чином свою репутацію корисного інструменту інтелектуального аналізу даних для **вилучення корисних шаблонів і знань із великих наборів даних**.
- Дані для багатьох програм та додатків зберігаються у форматі часових рядів, наприклад: дані про погоду, дані про продажі, біомедичні вимірювання, такі як артеріальний тиск і електрокардіограма, ціни на акції, біометричні дані тощо.
- Багато проектів, пов’язаних з аналізом часових рядів, було виконано в різних областях для різних цілей, таких як: зіставлення підпослідовностей, виявлення аномалій, візуалізація, сегментація, ідентифікація закономірностей, аналіз трендів, узагальнення та прогнозування.



Які основні виклики?

Кластеризація часових рядів у контексті великих наборів даних є складною проблемою з двох основних причин.

- По-перше, **дані часових рядів часто мають високу розмірність**, що робить обробку цих даних повільною та складною для багатьох алгоритмів кластеризації.
- Друга причина полягає у **вимірюванні подібності**, на чому власне і базується ідея утворення кластерів.

У літературі визначено чотири компоненти кластеризації часових рядів:

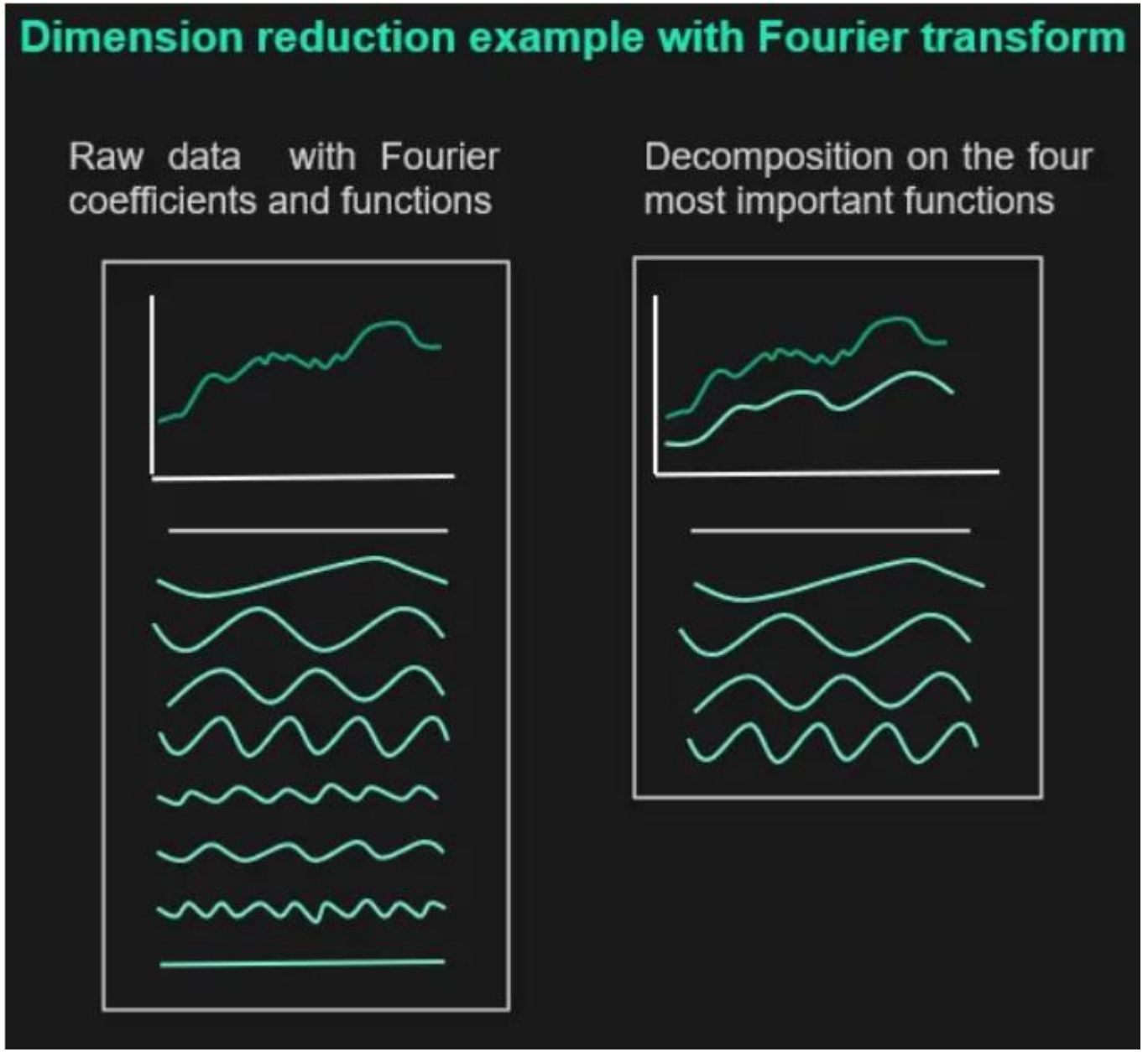
- зменшення розмірності або метод представлення
- вимірювання відстані між часовими рядами (вимірювання подібності)
- алгоритм кластеризації
- оцінка якості кластеризації та аналіз результатів.

Препроцесинг часових рядів та їх спрощення

- На практиці дуже рідко кластеризують сирі дані часових рядів, натомість виконують їх препроцесинг і певне спрощення.
- Необроблені дані можна обробити перед кластеризацією шляхом застосування перетворень: стандартизації, згладжування, інтерполяції, стаціонарізації тощо.
- Ці процеси можуть усунути шум і небажані тенденції в даних, і алгоритми кластеризації буде застосовано до результуючих рядів.
- Простим і добре відомим прикладом є використання швидкого перетворення Фур'є. Будь-яку «гарну» математичну функцію часу можна описати комбінацією функцій синуса та косинуса частоти:

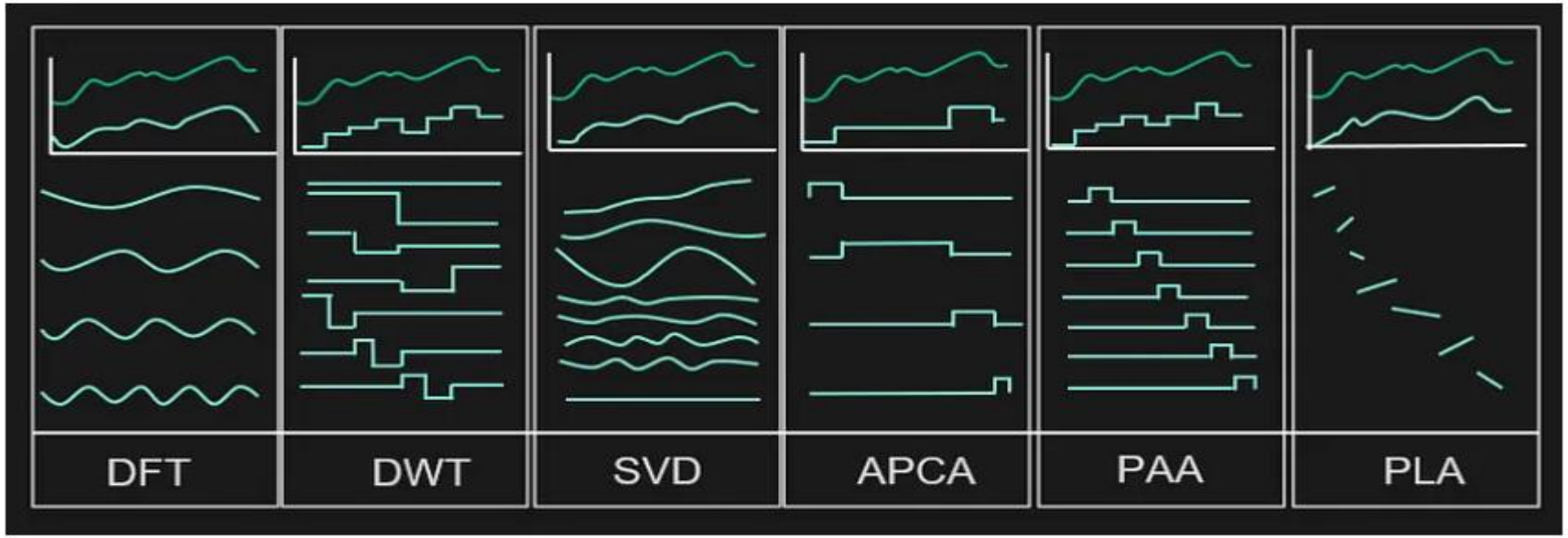
$$X(t) = \sum_{k=1}^{n} \left[a_k \cos(w_k t) + b_k \sin(w_k t) \right]$$

- У цьому розкладанні на синуси та косинуси коефіцієнти a і b представляють внесок коливань частоти ω в ряд X.
- **Нижні частоти відображатимуть тенденцію та цикл ряду**, зокрема, сезонні частоти.
- Таким чином, ми **можемо отримати спрощене представлення ряду, зберігаючи лише «найважливіші» коефіцієнти Фур'є.**



Методи декомпозиції рядів

- Дискретне перетворення Фур'є (DFT)
- Дискретне вейвлетне перетворення (DWT)
- SVD
- La SSA (аналіз сингулярного спектру)
- Кускове агрегатне наближення: PAA
- Адаптивна кусково-стала апроксимація: APCA
- Кусково-лінійна апроксимація: PLA
- **Empirical Mode Decomposition (EMD) -** це потужний алгоритм для декомпозиції часових рядів на набір компонент (IMFs) та залишковий член. EMD особливо підходить для нелінійних, нестаціонарних часових рядів. Він ітеративно вилучає компоненти IMF, які представляють коливальні компоненти в даних.

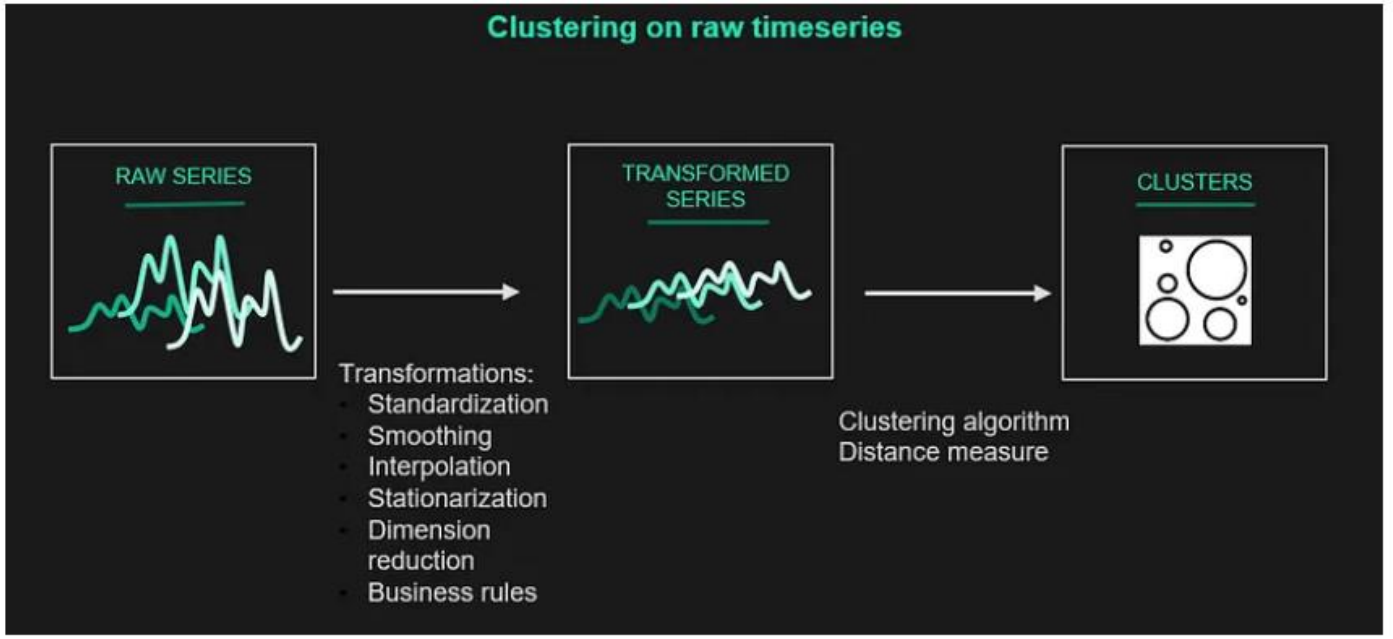


Застосування знань про досліджувану область

- Іншим можливим підходом до спрощення даних **є застосування «бізнес-правил», коли залежно від варіанту застосування**, **можливо зменшити розмірніть** без використання статистичних методів.
- Згадаємо, наприклад, часові ряди, що представляють споживання електроенергії за один рік із вимірюваннями, які проводяться з 30-хвилинними інтервалами.

Цілком можливо, залежно від мети:

- агрегувати ряди (наприклад, з погодинними інтервалами)
- ущільнити повторювані моделі та створити коротші ряди (середні значення зимових і літніх тижнів) тощо.



Кластеризація на основі моделі

- Ідея полягає в тому, щоб **змоделювати часові ряди**, з метою охопити й узагальнити їх динаміку. Тоді **два часові ряди можна вважати схожими, якщо їх встановлені моделі схожі.**
- Наприклад можна використовувати модель авторегресії або навіть ARIMA, які підганяються під часовий ряд. Потім ці моделі порівнюються за певними показниками.
- Однак на практиці показано, що зазвичай підходи, засновані на моделі, мають проблеми з масштабованістю і можуть призвести до зниження продуктивності, коли кластери розташовані близько один до одного. Тому ці методи використовуються рідко.

Кластеризація на основі вилучених ознак часового ряду (Feature-based approach)

Тут логіка кластеризації на основі моделі просувається далі:

- ми використовуємо перелік показників, які дозволяють кваліфікувати часові ряди (**тенденція/тренд, сезонність, автокореляція, нелінійність, асиметрія, ексцес тощо**). Це ознаки, які можна виділити для опису кожного часового ряду.
- Потім переходимо **від необробленого часового ряду до набору його статистичних та аналітичних характеристик**.
- Для кожного часового ряду таких характеристик можна обрахувати сотні. Тому може бути цікаво налагодити **процес вибору найкращих/релевантних ознак та/або застосувати алгоритм зменшення розмірності**.
- Нарешті, ми можемо обрати спосіб обчислення відстані/міри подібності та алгоритму, залежно від обмежень реалізації.

Міри схожості/несхожості в кластеризації часових рядів

- У кластеризації часових рядів показники подібності/несхожості відіграють важливу роль і мають великий вплив на результати кластеризації.
- Однак вибрати правильну міру не є очевидним завданням.
- Деякі міри подібності пропонуються на основі конкретних трансформованих часових рядів, а деякі з них працюють незалежно від застосованого перетворення або сумісні з необробленими часовими рядами.

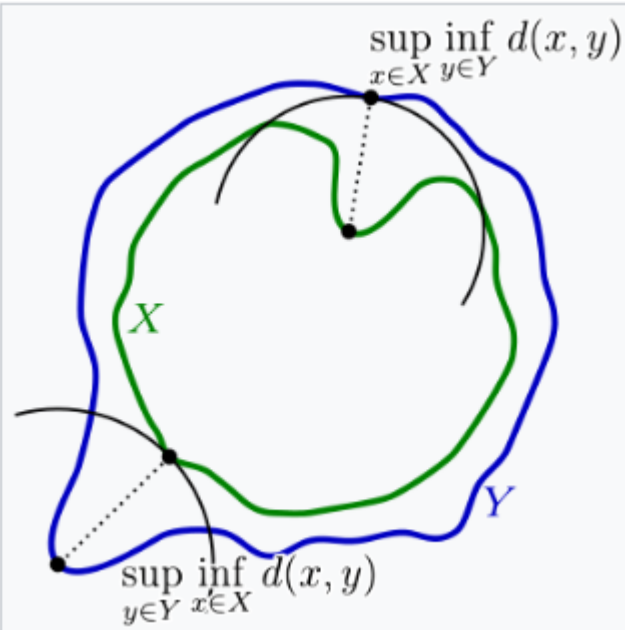
- Зазвичай **при виконанні «класичної» кластеризації** відстані між індивідами «ґрунтуються на відповідності», тобто **кожна конкретна ознака поєднується з відповідною такою ж ознакою** іншого індивіда.
- Однак **для часових рядів відстань кластеризації можна розрахувати приблизно**. Це дає змогу розрахувати відстань між рядами з нерегулярними інтервалами семплювання та неоднаковою довжиною.
- Існують популярні та ефективні засоби вимірювання відстані для часових рядів:

- відстань Хаусдорфа (DH)
- модифікована відстань Хаусдорфа (MODH)
- динамічне викривлення часу (DTW)
- евклідова відстань,
- евклідова відстань у підпросторі головних компонент
- найдовша спільна підпослідовність (LCSS)

Hausdorff Distance (DH)/ Modified Hausdorff Distance (MODH)

- **Означення DH** : Вимірює максимальну відстань між двома наборами точок. Визначає наскільки далеко множина точок лежить від найближчої точки іншої множини.

$$DH(A,B) = \max(\sup_{a\in A}\inf_{b\in B}\|a-b\|, \sup_{b\in B}\inf_{a\in A}\|b-a\|)$$



- **Обмеження**: Чутлива до аутлаєрів.
- **Означення MODH**: Узагальнення відстані Хаусдорфа, що усереднює відстані між точками множин. Менш чутлива до аутлаєрів та шумів.

$$MODH(A,B) = \frac{1}{|A|}\sum_{a\in A}\min_{b\in B}\|a-b\|$$

Dynamic Time Warping (DTW)

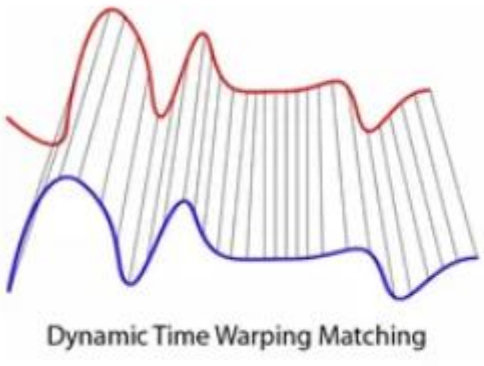
- **Означення**: вирівнює два часові ряди, деформуючи часову вісь, щоб мінімізувати відстань. Враховує зсуви та викривлення часових рядів.
- **Використання**: порівняння часових рядів різної довжини або з рівними часовими періодами (наприклад, мова, жести, курс акцій, тощо).
- **Перевага**: стійкість до нелінійних перетворень.
- **Обмеження**: інтенсивні обчислення для довгих послідовностей.

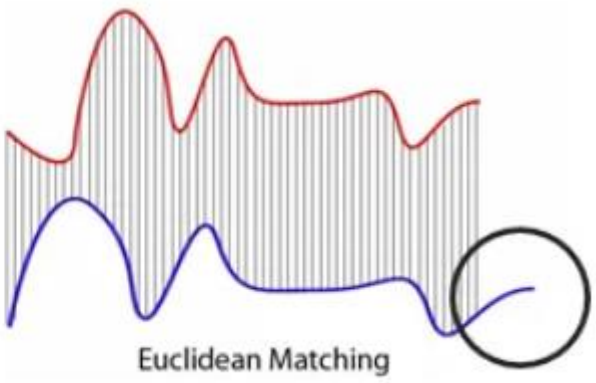
$$DTW(x, y) = \min_{\pi} \sqrt{\sum_{(i,j)\in\pi} d(x_i, y_j)^2}$$

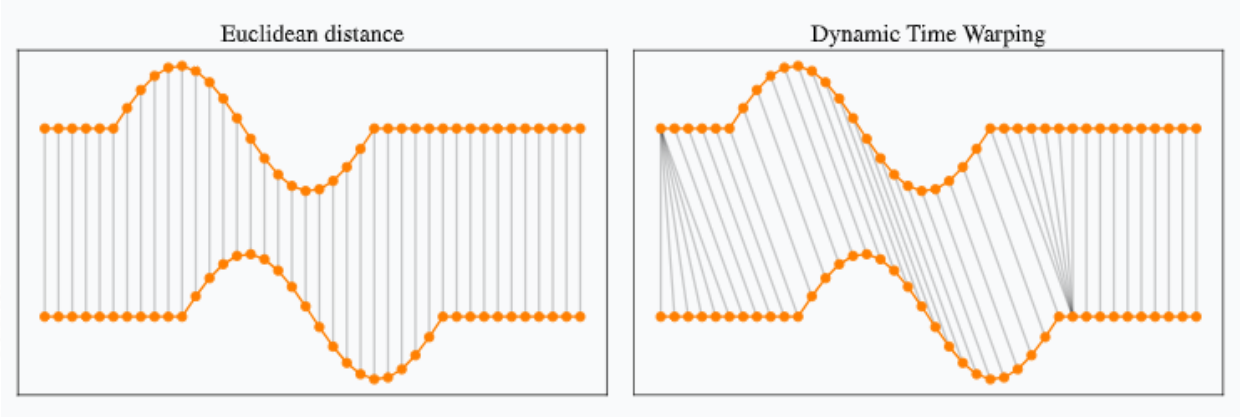
where $\pi = [\pi_0, \ldots, \pi_K]$ is a path that satisfies the following properties:

- it is a list of index pairs $\pi_k = (i_k, j_k)$ with $0 \leq i_k < n$ and $0 \leq j_k < m$
- $\pi_0 = (0, 0)$ and $\pi_K = (n-1, m-1)$
- for all $k > 0$, $\pi_k = (i_k, j_k)$ is related to $\pi_{k-1} = (i_{k-1}, j_{k-1})$ as follows:
  - $i_{k-1} \leq i_k \leq i_{k-1} + 1$
  - $j_{k-1} \leq j_k \leq j_{k-1} + 1$

- Підсумовуючи формулу DTW: DTW обчислюється як квадратний корінь із суми квадратів відстаней **між кожним елементом в X та його найближчою точкою в Y.**
- Наприклад, ми маємо дві різні криві — червону та синю різної довжини.
- Дві криві **мають однаковий патерн**, однак **синя крива дов**ша за червону.
- Якщо ми застосовуємо евклідову відповідність один-до-одного (показано праворуч), відображення не буде ідеально синхронізовано, і хвіст синьої кривої залишиться поза увагою. DTW вирішує цю проблему, розробляючи відповідність «один-до-багатьох», щоб аналогічний шаблон ідеально збігався, і не залишалося поза увагою точок обидвох кривих (показано ліворуч).







Евклідова відстань/Евклідова відстань у підпросторі головних компонент

- **Означення ЕВ**: вимірює відстань по прямій лінії між двома векторами. Для часових рядів передбачається, що обидва ряди мають однакову довжину.

$$d(A, B) = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$$

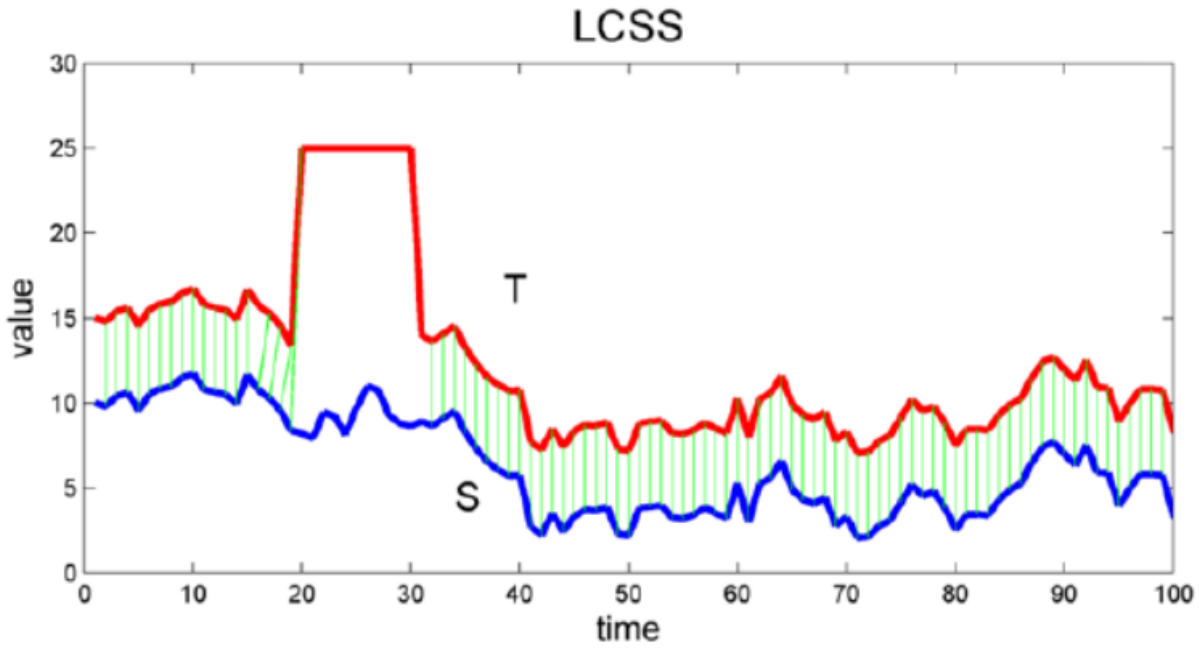
- **Використання**: прості порівняння часових рядів однакової довжини.
- **Означення ЕВ у ПГК**: проектує багатовимірні часові ряди в низьковимірний підпростір за допомогою аналізу головних компонентів (PCA) і обчислює евклідову відстань у цьому підпросторі.
- **Кроки**:
  - Залучаємо PCA для зменшення розмірності.
  - Обчислюємо евклідову відстань у стисненому просторі.
- **Використання**: ефективне порівняння багатовимірних часових рядів.
- **Перевага**: Зменшує шум і фокусується на основних патернах.

Найдовша загальна підпослідовність (LCSS)

- **Означення**: вимірює подібність на основі найдовшої підпослідовності, спільної для обох часових рядів, дозволяючи деяким елементам не збігатися.

$$LCSS(A, B) = \max(|S|) \quad \text{where } S \subseteq A \cap B$$

- **Використання**: порівняння часових рядів із шумом або відсутніми даними (наприклад, аналіз траєкторії).
- **Перевага**: Стійка до шуму та пропусків у даних часових рядів.



- Вибір міри відстані, очевидно, залежить від
  - мети використання,
  - характеристик часового ряду (таких як довжина, семплювання та ін.),
  - методу його представлення.

Дослідження показують, що загальні властивості часових рядів, такі як:

- шум
- масштаб амплітуди
- зсуви
- поздовжнє масштабування,
- пропуски в часовому ряді
- часовий дрейф

можуть створити проблеми при виборі та інтерпретації вимірювань відстані.

Недоліки відстані Евкліда

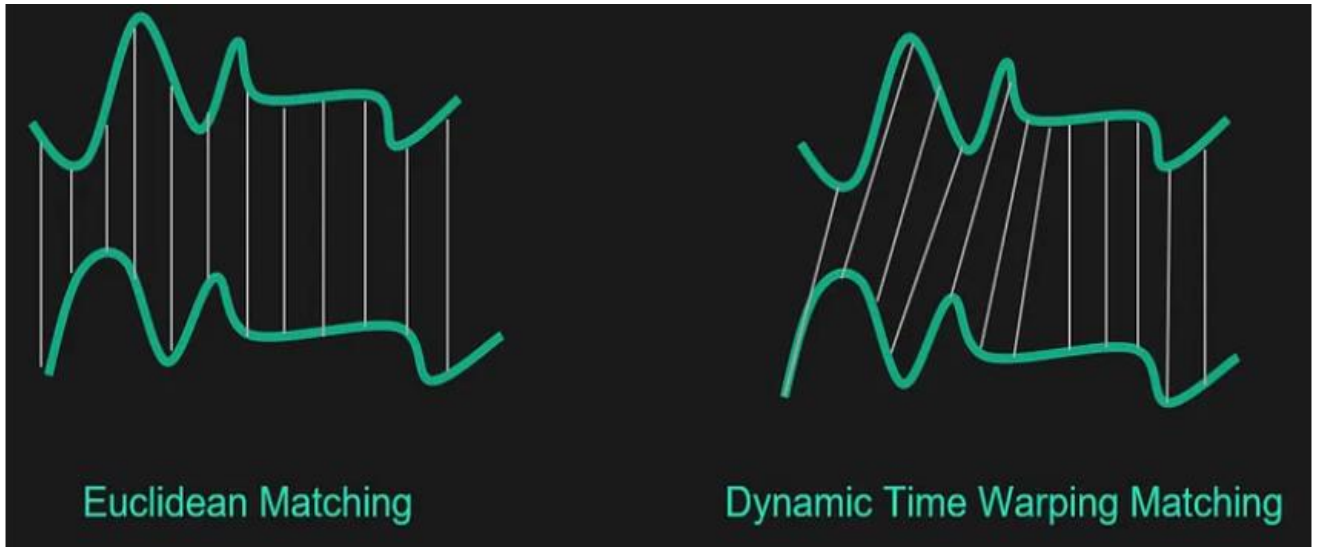
- Для класифікації даних було запропоновано сотні варіантів вимірювання відстані, **найпопулярнішою серед яких є евклідова відстань**.
- Однак, коли йдеться про **кластеризацію часових рядів**, використання евклідової відстані або будь-якої іншої метрики Мінковського на необроблених даних **може призвести до неінтуїтивних результатів**.
- Зокрема, **ця відстань дуже чутлива до ефектів масштабу та до нетипових або відсутніх точок даних**.
- Евклідова відстань **не дозволяє врахувати часові зміщення**. Дійсно, дуже подібні сигнали можуть бути зміщені в часі і не будуть вимірюватися як «близькі» на стандартній евклідовій відстані.

Для вирішення цієї проблеми існують адаптовані показники, такі як **відстань DTW** (Dynamic Time Warping).

Ця відстань розглядає час як нелінійний та еластичний, а, отже, пояснює відносні часові зсуви двох часових рядів через деформацію осі часу.

Таким чином**, евклідова відстань враховує лише одночасні події**, тоді як відстань **DTW враховує події поза фазою**.

Звичайно, це не магічний підхід, який працює у всіх випадках. Немає метрики чи алгоритму кластеризації, який би перевершив усі інші - все залежить від контексту.



Залежно від обраної міри відстані, отримані кластери можуть бути «щільнішими» за часом або формою.

- На ці відмінності в параметрах кластера також **можуть впливати методи обробки даних, розглянуті вище**.
- Кластеризацію можна застосувати до необроблених часових рядів, до спрощеного представлення рядів, до моделей часових рядів або до вилучених ознак. Потім можна застосувати заходи, що впливають на часову або геометричну подібність кластерів:
- **Міра подібності, заснована на часовій близькості**, полягає в тому, щоб знайти подібні часові ряди за відліком часу і формою – евклідова відстань, DTW, LCSS тощо.
- **Міра подібності на основі стиснення**: підходить для коротких і довгих часових рядів - коефіцієнт кореляції Пірсона та відповідні відстані.
- **Міра подібності на основі вилучених ознак**: підходить для довгих часових рядів.
- **Засіб вимірювання подібності на основі моделі**, наприклад ARMA

Алгоритми кластеризації часових рядів

Загалом методи кластеризації можна розділити на п’ять груп:

- ієрархічні
- ітеративні
- на основі моделі
- на основі щільності
- багатокрокові або гібридні алгоритми кластеризації.

Далі обговорюється застосування кожної групи до кластеризації часових рядів.

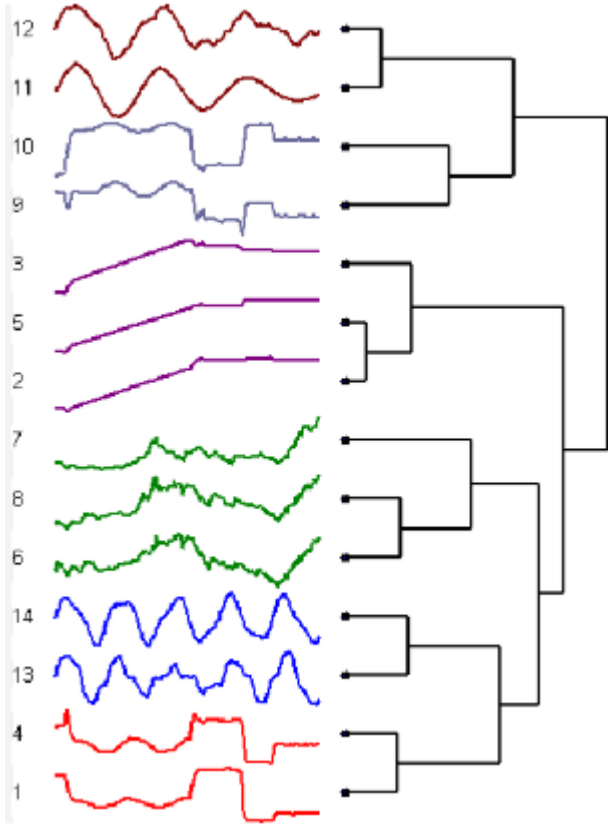
Ієрархічна кластеризація

Ієрархічний кластерний аналіз — один із найпопулярніших і простих для розуміння методів кластеризації. Цей клас кластеризації має два типи алгоритмів:

- Агломеративна ієрархічна кластеризація: розглядає кожен елемент як кластер, а потім поступово об’єднує кластери (знизу вгору)
- Роздільна ієрархічна кластеризація: починається з усіх об’єктів як єдиного кластера, а потім розділяє кластер, щоб досягти кластерів з одним об’єктом (зверху вниз).

Загалом, **ієрархічні алгоритми є слабкими з точки зору якості**. Як наслідок, зазвичай **ієрархічні алгоритми кластеризації поєднуються з іншим алгоритмом** в межах гібридного підходу до кластеризації для вирішення цієї проблеми.

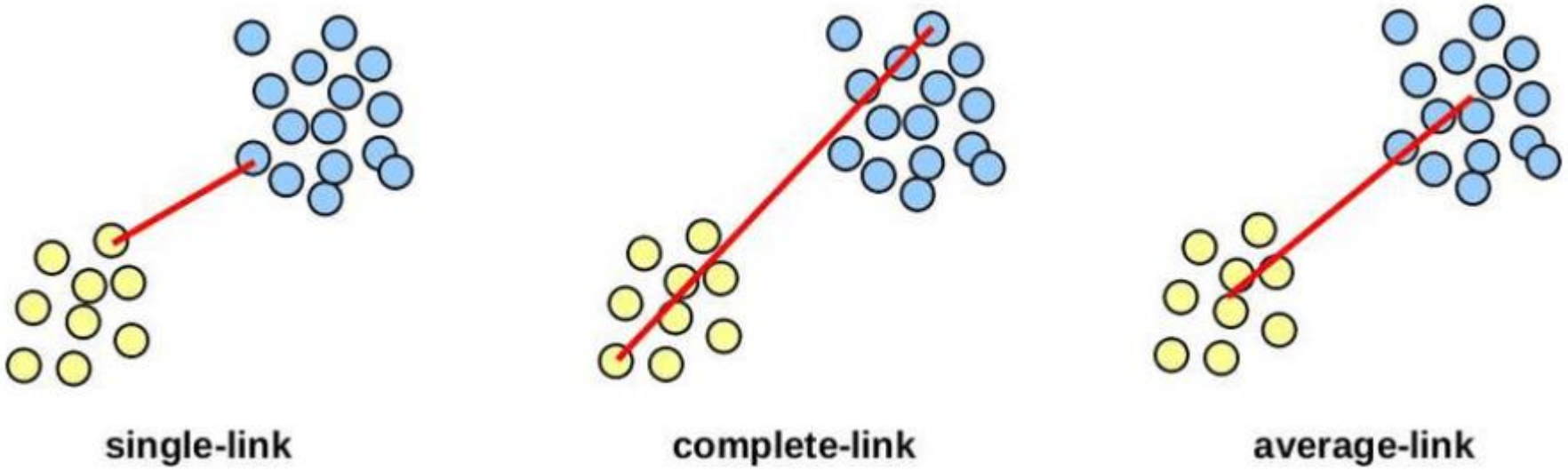
- Алгоритм **генерує попарну матрицю відстаней часових рядів**. Для створення цієї матриці **необхідно вибрати міру відстані/подібності**, що очевидно, вплине на результати алгоритму.
- Однією з **сильних сторін** цього алгоритму є те, що **не потрібно вказувати кількість кластерів як вхідний параметр**. Це чудово, оскільки визначити цей параметр може бути дуже важко.



- Крім того, **ієрархічні алгоритми кластеризації мають здатність кластеризувати часові ряди різної довжини**. За допомогою цього алгоритму **можна кластеризувати часові ряди різної довжини**, якщо для обчислення відмінності/подібності часових рядів використовується динамічне викривлення часу (DTW) або найдовша загальна підпослідовність (LCSS).
- Однак **ієрархічна кластеризація фактично нездатна мати справу з великими часовими рядами через її квадратичну обчислювальну складність**, і така погана масштабованість призводить до її обмеженого використання на невеликих наборах даних.
- **Окрім вибору алгоритму та міри відстані, необхідно вибрати критерій агрегування**. З кількох ітерацій ієрархічного алгоритму доведеться розглянути відстань між групами окремих часових рядів і виникне наступне питання: як обчислити відстань за наявності груп? Існують різні можливості.

Перш за все, три критерії, які працюють незалежно від міри відстані/схожості, є (серед багатьох інших):

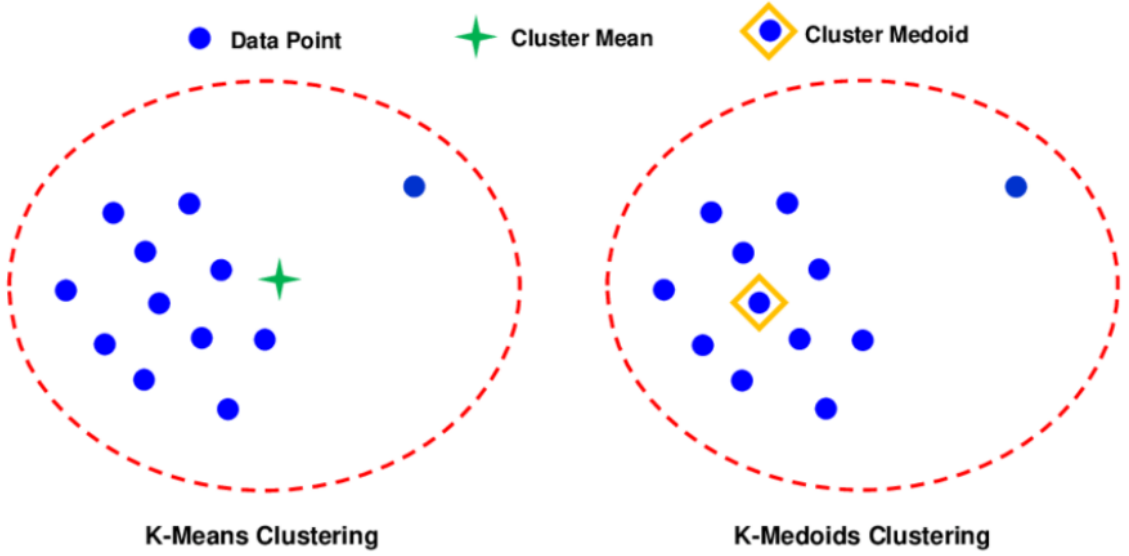
- **Критерій одиночного зв’язку**, який базується на двох найближчих часових рядах;
- **Критерій максимального зв’язку** (повного зв’язку), який базується на двох найбільш віддалених елементах;
- **Критерій середнього зв’язку**, який використовуватиме середнє значення відстаней між елементами кожного класу для виконання групувань.
- Залежно від обраного критерію ми можемо отримати деревовидні структури дуже різної форми.
- Зрештою, важливо пам’ятати, що не існує єдиного найкращого методу, так само як немає єдиного методу кластеризації. Рекомендується випробувати кілька підходів і проаналізувати результати.



Ітеративні алгоритми кластеризації

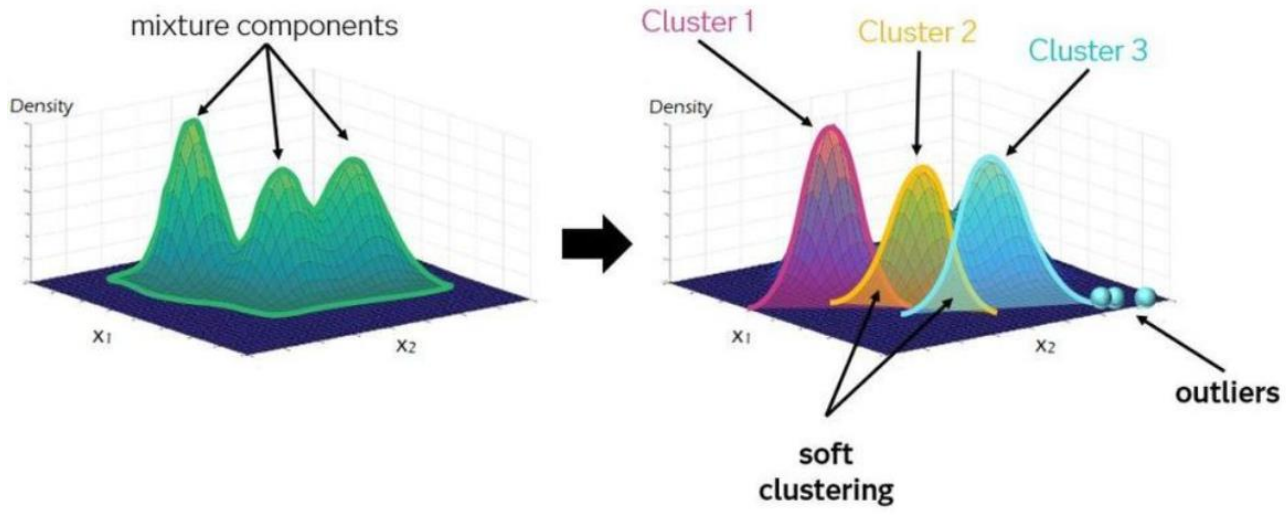
- **У цих алгоритмах кластеризації кількість кластерів, k, повинна бути попередньо призначена**, що є дуже складним завданням навіть для кластеризації даних, які не є часовими рядами.
- **Це ще складніше з даними часових рядів, оскільки набори даних дуже великі**, а діагностичні перевірки для визначення кількості кластерів непрості – тож тут має сенс відштовхуватись від предметної області.
- Однак **ці алгоритми дуже швидкі** порівняно з ієрархічною кластеризацією, і це робить їх вельми придатними для кластеризації часових рядів, як було показано в багатьох роботах.
- **Кластери можна утворювати «жорстким» або «м'яким» способом**, що означає, що або об'єкт чітко призначається кластеру (жорсткий), або об'єкту призначається ймовірність перебування в кластері (м'який).

- **Для заданої міри відстані та відомої кількості класів k** легко уявити просте й оптимальне рішення кластеризації: **перерахувати всі можливі можливості кластеризації та залишити найкращу**. Однак це рішення не застосовне на практиці, оскільки кількість можливих комбінацій угруповань швидко стає величезною. Наближені рішення можна отримати завдяки евристикам, і існує кілька таких алгоритмів. Їх основний принцип заснований на методі рухомих центрів.
- **Принцип роботи рухомих центрів полягає в мінімізації загальної відстані** (як правило, евклідової відстані) **між усіма об'єктами в кластері від центру кластера.**
- Алгоритм не обов'язково призводить до глобального оптимуму: **кінцевий результат залежить від початкових центрів**. Найвідоміші алгоритми, які використовують мобільні центри:
- Метод K-середніх
- Методи динамічної кластеризації
- Методи K-medoids



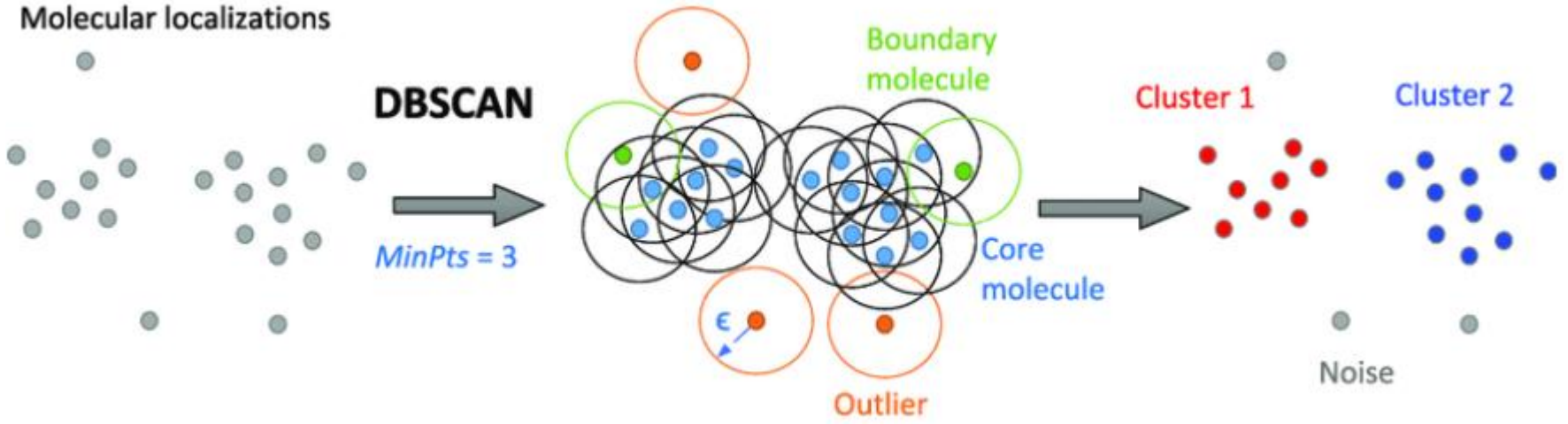
Кластеризація на основі моделювання

- **Кластеризація на основі моделі припускає, що дані були згенеровані моделлю**, і намагається відновити вихідну модель з даних.
- У літературі досить мало наукових публікацій, які демонструють цінність цих методів. **У більшості випадків вони будують кластеризацію за допомогою поліноміальних моделей, змішаних моделей Гауса, ARIMA або нейронної мережі**.
- Загалом **кластеризація на основі моделі** має два недоліки: по-перше, вона **має повільний час обробки великих наборів даних**.
- По-друге, і що важливіше, **кластеризація на основі моделі вимагає вхідних параметрів і припущень**, які можуть бути неправильними для варіанта використання та призвести до неточних кластерів.



Кластеризація на основі щільності

- Кластеризація на основі щільності відноситься до неконтрольованих методів навчання, які **ідентифікують відмінні групи/кластери в даних на основі ідеї, що кластер у просторі даних є неперервною областю високої щільності, відокремленої від інших таких кластерів суміжними областями низької щільності**.
- Одним із найвідоміших алгоритмів, заснованих на цій концепції на основі щільності, є DBSCAN.
- Однак огляд літератури показує, що кластеризація на основі щільності не широко застосовувалася для кластеризації даних часових рядів через високу складність цього завдання.



Гібридний підхід (багатоступенева кластеризація)

На попередніх слайдах алгоритми кластеризації були представлені окремо. Однак на практиці може бути корисно використовувати їх разом. Комбінуючи підходи, можна використовувати основні переваги різних методів, а саме:

- **Можливість аналізу великої кількості окремих часових рядів, що є сильною стороною неієрархічних методів** (для певної кількості спостережень стає важко застосовувати ієрархічні методи безпосередньо);
- **Вибір оптимальної кількості класів, можливий завдяки ієрархічній класифікації.**
- На практиці, як тільки набір даних стає великим, рекомендується застосувати гібридний підхід, щоб отримати найкращий компроміс між продуктивністю та швидкістю виконання.

Оцінювання якості кластеризації часових рядів

- Правильна оцінка ефективності процедури кластеризації завжди є відкритою проблемою, як загалом, так і конкретно для кластеризації часових рядів.
- **У неконтрольованому навчанні ми не маємо міток для встановлення показників ефективності**. Визначення «хорошої» кластеризації залежить від проблеми і **часто є суб'єктивним**. Кількість кластерів, розмір кластера, визначення викидів і визначення подібності між часовими рядами в задачі – це поняття, які залежать від поставленого завдання.
- Оцінка кластеризації часових рядів повинна відповідати деяким рекомендаціям :
- **Треба уникати упередженості реалізації шляхом ретельного планування експериментів**
- Нові методи вимірювання подібності слід порівнювати з простими та стабільними метриками, такими як евклідова відстань.

- Результати можна оцінити за допомогою різних заходів. Візуалізація та скалярні вимірювання є основними методами оцінки якості кластеризації, які також відомі як оцінка валідності кластеризації.
- **Типові цільові функції** в кластеризації формалізують мету **досягнення високої внутрішньокластерної подібності** (об’єкти в кластері подібні) і **низької міжкластерної подібності** (об’єкти з різних кластерів і несхожі).
- Внутрішні індекси валідності оцінюють результати кластеризації, використовуючи лише характеристики елементів та інформацію, властиву набору даних. Зазвичай вони використовуються в тому випадку, якщо істинні рішення невідомі.

- Кластери можуть бути охарактеризовані за допомогою, наприклад, загальних статистичних інструментів, таких як середнє значення, стандартне відхилення, мінімальне/максимальне значення тощо.
- Сума квадратичних помилок (SSE) — цільова функція, яка описує когерентність даного кластера. Для кожного часового ряду помилкою є відстань до найближчого кластера. Очікується, що «кращі» кластери дадуть нижчі значення SSE. Фактично, SSE – це ніщо інше, як внутрішньокластерна дисперсія. Ця сума буде приймати менші значення, якщо часові ряди у кластері будуть корельовані.

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

Також існує багато внутрішніх індексів:

- індекс силуету,
- індекс Девіса-Болдіна,
- Калінскі-Харабаж,
- індекс Кржановського-Лаї,
- індекс Хартігана,
- Індекс середньоквадратичного стандартного відхилення (RMSSTD),
- індекс відстані між двома кластерами (CD),
- зважений внутрішній індекс,
- індекс однорідності та індекс поділу.

Індекс силуету

- **Означення**: оцінює якість кластерів шляхом вимірювання згуртованості (відстань між кластерами) і розділення (відстань між кластерами).

де:

$$s(x) = \frac{b(x) - a(x)}{\max(a(x), b(x))}$$

a(x): Середня відстань x до точок у тому самому кластері.

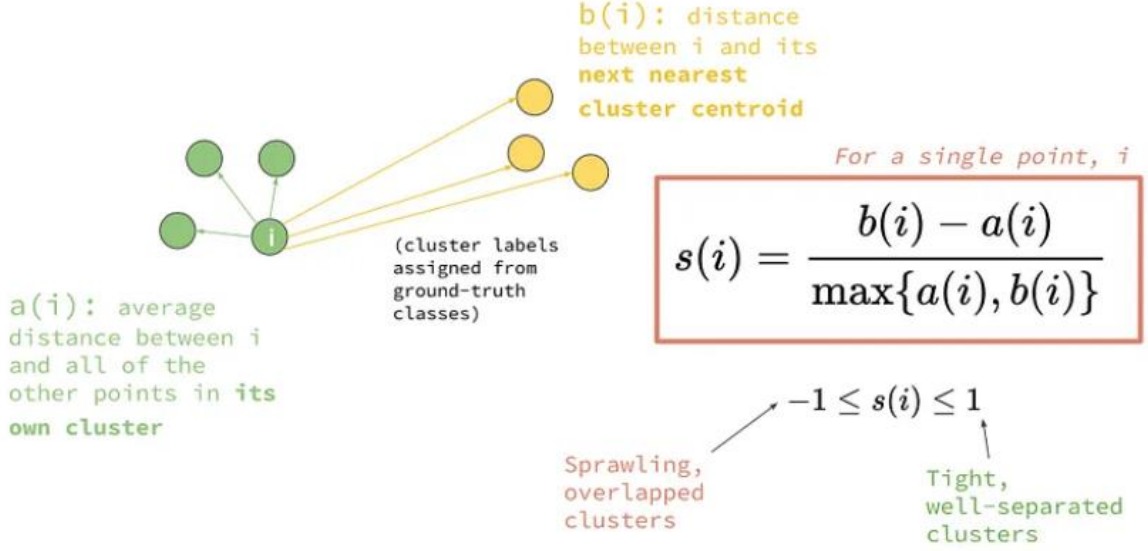
b(x): Мінімальна середня відстань x до точок в іншому кластері.

- **Інтерпретація**:

s(x) близьке до 1: добре згруповані кластери.

s(x) близьке до 0: неоднозначна кластеризація.

s(x) близьке до -1: неправильно згруповане.



- Однак у контексті кластеризації саме бізнес-перспектива часто є визначальним фактором при розгляді ефективності кластеризації. Можуть бути нереалістичні результати, яких ми не хочемо, конкретні очікування, з тих чи інших причин наперед відомакількість кластерів, що слід поважати, тощо.

Висновки

- Внутрішні характеристики часових рядів, роблять кластеризацію складною задачею.
- Дійсно, важко сліпо застосовувати звичайні методи кластеризації, які в більшості випадків не працюють. **Основними перешкодами є висока розмірність, дуже висока кореляція ознак і (як правило, велика) кількість шуму, що характеризує дані часових рядів**.

Виходячи з цих проблем, кластеризацію часових рядів, а також пов’язані з нею дослідження можна розділити на дві тенденції:

- Літературні джерела показують, що були докладені зусилля для розгляду високовимірних характеристик даних часових рядів і **розробки методів представлення часових рядів у менших вимірах**, сумісних із звичайними алгоритмами кластеризації.
- Багато досліджень зосереджено на **представленні вимірювання відстані** на основі необроблених часових.

- Тому **першим кроком часто є зменшення розмірності**. Спектр методів широкий. Ми можемо застосувати прості моделі, як ковзні середні, бізнес-правила, або більш складні моделі, як перетворення Фур’є, PCA тощо.
- Далі, важливо **визначити найкращу міру подібності**. Найпопулярнішими мірами подібності в кластеризації часових рядів є евклідова відстань і DTW.
- Нарешті, **постає питання про алгоритм**. З цього приводу практично немає різниці між даними часових рядів і контекстом «класичної» кластеризації.
- Порівняно з іншими алгоритмами, **ітеративні алгоритми** широко використовуються через швидкість. Однак, оскільки кількість кластерів потрібно призначати заздалегідь, ці алгоритми важко застосувати в більшості додатків в реальному часі.
- **Ієрархічна кластеризація**, з іншого боку, **не потребує попереднього визначення кількості кластерів**, а також **має велику потужність візуалізації** в кластеризації часових рядів.
- Але ієрархічна кластеризація **обмежена невеликими наборами даних через її квадратичну обчислювальну складність.**

- Хоча існують можливості для вдосконалення всіх чотирьох аспектів кластеризації часових рядів, можна зробити висновок, що **головною можливістю для майбутніх робіт у цій галузі є робота над новими гібридними алгоритмами з використанням існуючих підходів до кластеризації, щоб збалансувати якість і витрати на кластеризацію часових рядів.**