Національний технічний університет
ДНІПРОВСЬКА ПОЛІТЕХНІКА
1899

Пістунов І.М.

Datamaning: навч. наоч. посіб. Дніпро : НТУ «ДП», 2024. 55 с.

В посібнику розглядається основні методи дейта манінгу, які включають в себе такі дисципліни як інформатика та ком’ютерна техніка, економетрика, теорія ймовірності, економічна кібернетика, кластерний аналіз, нейронні сітки, прогнозування часових рядів.
Наведено приклади розрахунків із застосування електронних таблиць Excel.
Призначено для студентів всіх спеціальностей.

Рецензенти:
Васильєва Н.К., завідувач каф. інформаційних систем ДДАЕУ, проф.
Алексеєв М.О., зав каф. програмних засобів комп’ютерних систем НТУ “ДП”, проф.

ЗМІСТ

Чи складно працювати
дата-аналітиком?

Спробуйте на безплатному
марафоні для новачків

Всього за 3 дні ви познайомитесь з основними робочими задачами дата-аналітика. Так зможете зрозуміти, чи підходить вам ця айтішна професія та які можливості відкриває. 😎

Марафон стартує регулярно, спробувати може будь-хто. Для участі потрібен лише ноутбук та 1 година вільного часу ввечері.

Якщо виконаєте всі домашки, берете участь у розіграші повного курсу Data Analyst в GoIT. Починайте хоч зараз!

Реєстрація:

https://i.goit.global/SrRuC

Що таке Data Mining

- **Data Mining** - це технологія виявлення неочевидних, об'єктивних та практично корисних закономірностей у великих обсягах даних.

- **Data Mining** - це апарат сучасної бізнес-аналітики та дослідження даних для виявлення прихованих закономірностей і побудови моделей прогнозування.

Застосування Датаманінгу:
Датаманінг необхідний для того, щоб перетворити сирі дані, які можуть бути складними для розуміння, на *корисну інформацію*, або *знання*. Він допомагає виявити закономірності, тенденції та взаємозв'язки в даних, що може допомогти приймати обґрунтовані рішення.
Звідки беруться дані в Датаманінгу:
Дані для датаманінгу можуть бути отримані з різних джерел: від баз даних компаній та організацій до соціальних мереж, сенсорних пристроїв та веб-сайтів. Ці дані можуть бути структурованими (наприклад, таблиці баз даних) або неструктурованими (тексти, зображення).
Цільова аудиторія Датаманінгу:
Датаманінг користуються різні фахівці та спеціалісти з різних галузей. Аналітики, дослідники, бізнес-аналітики, маркетологи, фінансисти та багато інших фахівців використовують датаманінг для здійснення аналізу даних, виявлення трендів, прогнозування та прийняття обґрунтованих рішень.
Отже, датаманінг допомагає робити дані більш зрозумілими та корисними для різних галузей та спеціалістів, а це робить його важливим інструментом в сучасному світі.

За допомогою Data Mining можна виявляти певні шаблони і тенденції, які можуть бути корисні для прийняття бізнес-рішень. Data Mining може використовуватися для різноманітних задач, таких як прогнозування продажів, виявлення шахрайства, управління ризиками, аналіз поведінки клієнтів та інших завдань, які вимагають обробки великих об'ємів даних.

В Data Mining використовуються різні методи, такі як класифікація, кластерний аналіз, асоціативні правила та інші. Для здійснення Data Mining можуть використовуватися спеціальні програмні засоби, що дозволяють здійснювати автоматичну обробку та аналіз великих об'ємів даних.

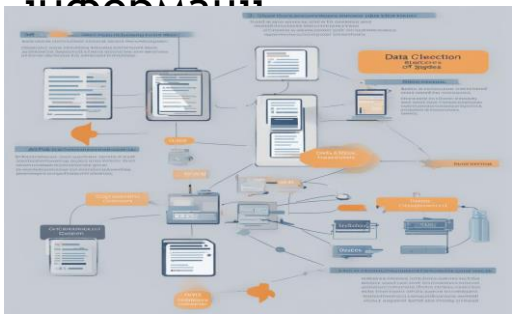
Рис. 1. Схема застосування Data Mining

Схема застосування Data Mining може включати наступні етапи:

Збір даних: на цьому етапі здійснюється збір потрібних даних з різних джерел, таких як бази даних, файли, веб-сторінки та інші джерела.
Підготовка даних: на цьому етапі відбувається очищення даних від помилок, дублікатів, непотрібних елементів та іншого, що може завадити подальшому аналізу.
Обробка даних: на цьому етапі виконується обробка даних, що містяться в базі, включаючи їх агрегацію, сортування, фільтрацію, інтеграцію та інші операції.
Моделювання: на цьому етапі виконується аналіз даних з використанням методів Data Mining, таких як класифікація, кластеризація, асоціативні правила та інші. В результаті моделювання формується математична модель, яка може використовуватись для передбачення результатів.
Оцінка та інтерпретація результатів: на цьому етапі проводиться оцінка та інтерпретація результатів моделювання. Оцінка включає оцінку точності моделі, що отримана в результаті моделювання. Інтерпретація результатів включає зрозуміння причинно-наслідкових зв'язків, що виявлені під час аналізу даних.
Використання результатів: на цьому етапі використовуються результати аналізу для вирішення конкретних завдань, наприклад, для передбачення попиту на товари, виявлення шахрайства, відшукання відмінностей в даних та інше.
Збереження та обробка: Отримані результати можуть бути збережені для подальшої обробки або використання, якщо ситуація вимагає подальшого моніторингу

Дані, Інформація та Знання:

Дані - це необроблені факти, цифри, символи, або спостереження, які ми отримуємо з навколишнього світу. Наприклад, це може бути набір чисел, список імен, або зображення. Дані самі по собі не мають особливого значення і не передають жодної інформації.



Інформація - це повідомлення або дані, які несуть у собі щось нове, раніше невідоме або розкривають зв'язки та закономірності, що роблять їх корисними та значущими для вироблення рішень, розуміння ситуацій та висновків.

Знання - це результат обробки та аналізу даних для отримання методів здійснення дій, розробки нового, пояснення подій або явищ та глибшого розуміння ситуацій. Воно включає в себе розуміння взаємозв'язків, тенденцій та закономірностей, що дозволяє використовувати інформацію в практичних цілях та приймати обґрунтовані рішення.

Датаманінг використовує різноманітні методи для обробки, аналізу та використання даних з метою виявлення корисної інформації. Ось деякі з найбільш поширених методів, які застосовуються в датаманінгу:

1. **Сортування та фільтрація :** Прості методи, які дозволяють відсортувати дані за певними змінними або вибрати підмножину даних, що задовольняють певні умови.
2. **Виявлення аномалій:** Виявлення відхилень або несподіваних зразків в даних, що може вказувати на потенційно важливі події або проблеми.

3. Агрегація: Об'єднання даних в групи з метою обчислення агрегованих статистик, таких як середнє, мінімум, максимум, сума тощо.

4. Кластеризація: Групування подібних об'єктів разом на основі спільних характеристик. Це дозволяє виявити приховані паттерни в даних.

5. Класифікація: Визначення приналежності об'єкта до певного класу або категорії на основі його характеристик. Цей метод допомагає робити прогнози на основі існуючих даних.

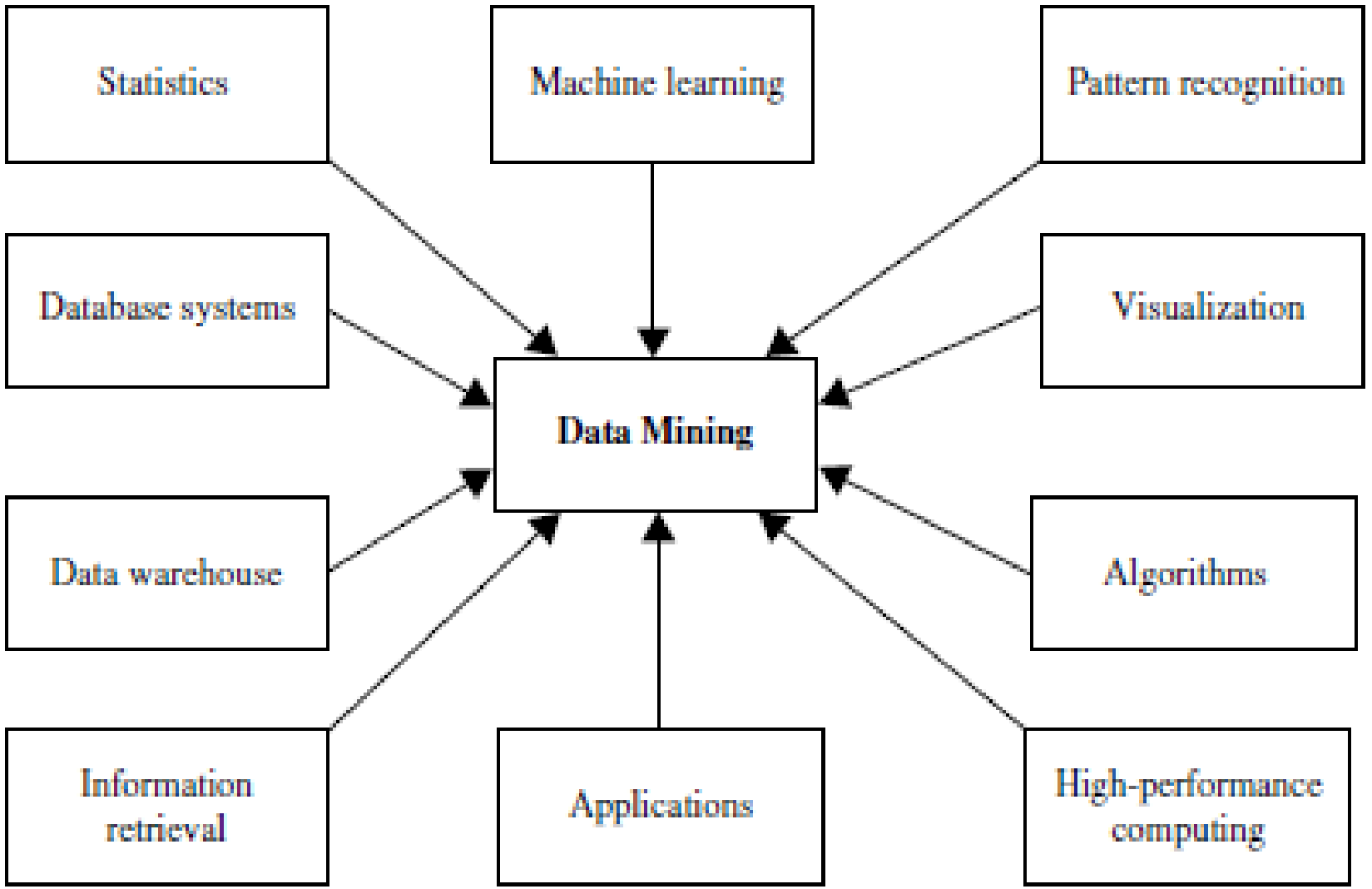
6. Асоціативний аналіз: Виявлення асоціацій та зв'язків між різними змінними в даних, зокрема, знаходження "часто спільних" об'єктів.

7. Регресійний аналіз: Визначення залежності між змінними та прогнозування значень однієї змінної на основі інших.

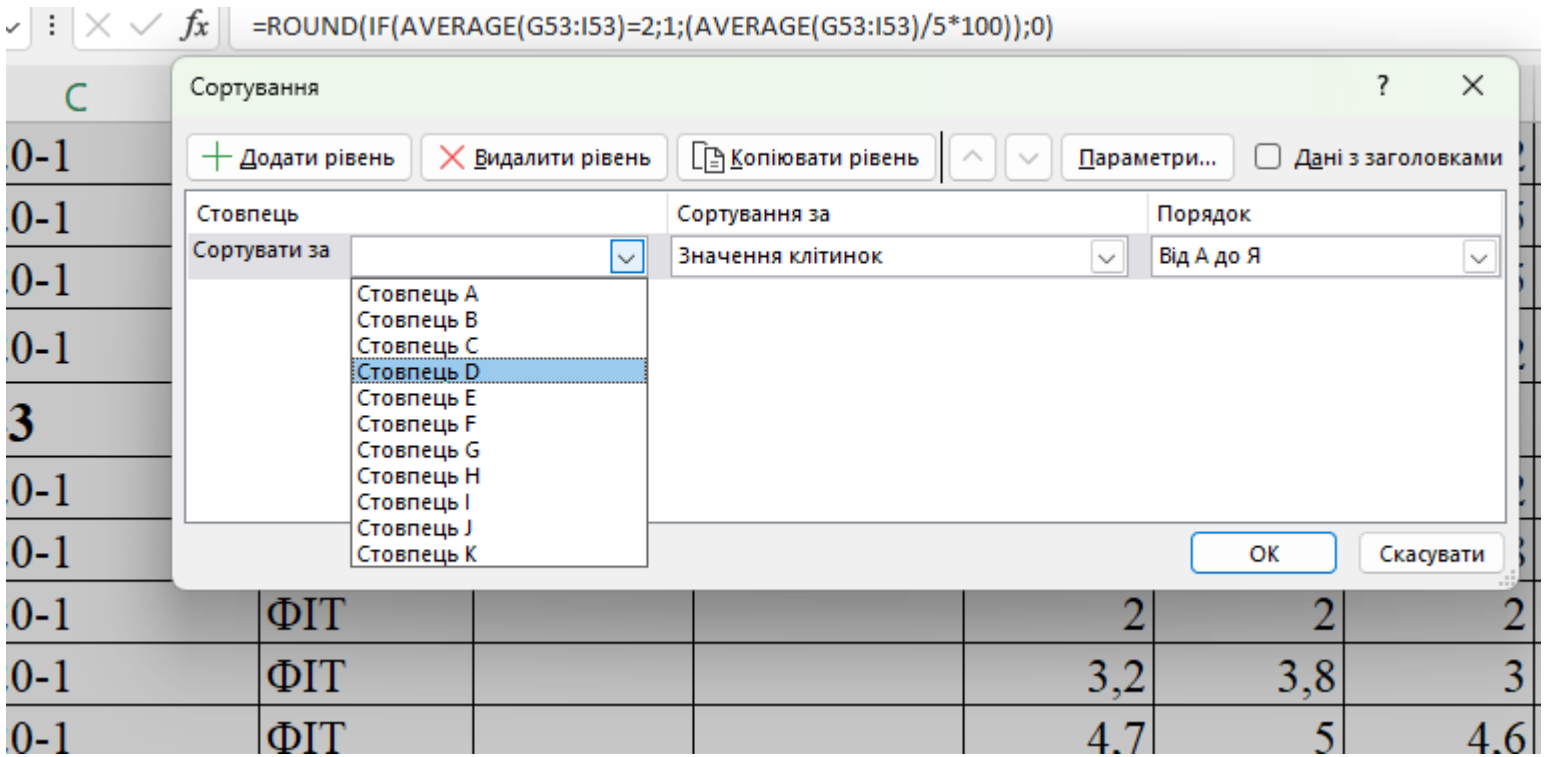
8. Прогнозування: Використання історичних даних для прогнозування майбутніх подій або трендів.

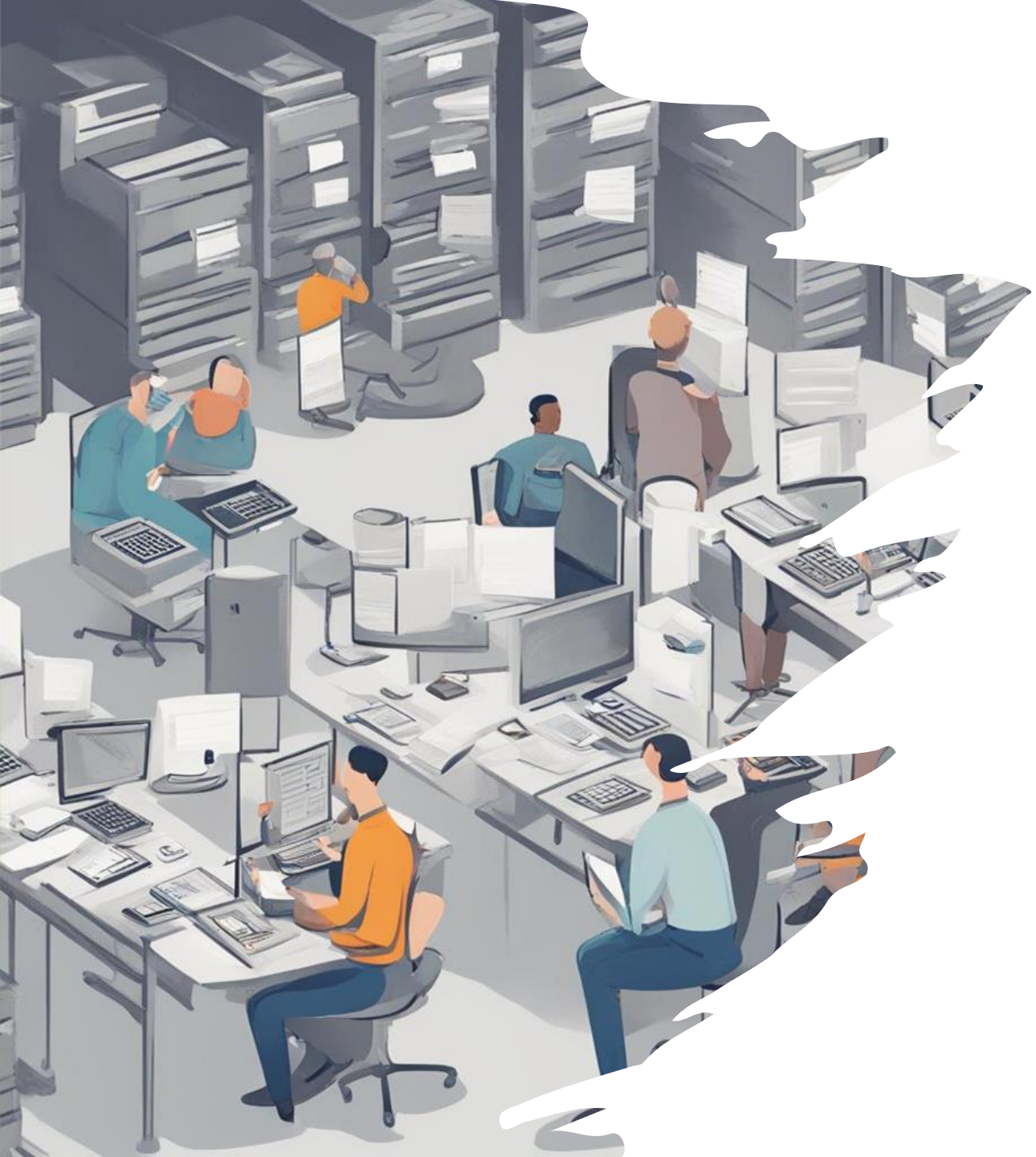
9. Текстовий аналіз: Обробка та аналіз текстових даних для виявлення тенденцій, сентименту, ключових слів тощо.

10. Візуалізація даних: Використання графіків, діаграм, карт та інших візуальних представлень для кращого розуміння структури та характеристик даних.


Statistics
Machine learning
Pattern recognition
Database systems
Visualization
Data Mining
Data warehouse
Algorithms
Information retrieval
Applications
High-performance computing

Сортування та фільтрація





• Фільтрація в Датаманінгу:

• Фільтрація - це інструмент, який допомагає зменшити обсяг даних, вибираючи лише ті частини, які відповідають певним умовам, і дозволяє нам зосередитися на конкретних аспектах даних, які є важливими для наших цілей аналізу та рішень.

• У датаманінгу, коли ми працюємо з великою кількістю даних, фільтрація допомагає зосередитися на тих даних, які є важливими для нашої конкретної задачі. Ми встановлюємо певні умови, і тільки дані, які задовольняють ці умови, залишаються видимими, а інші виключаються.

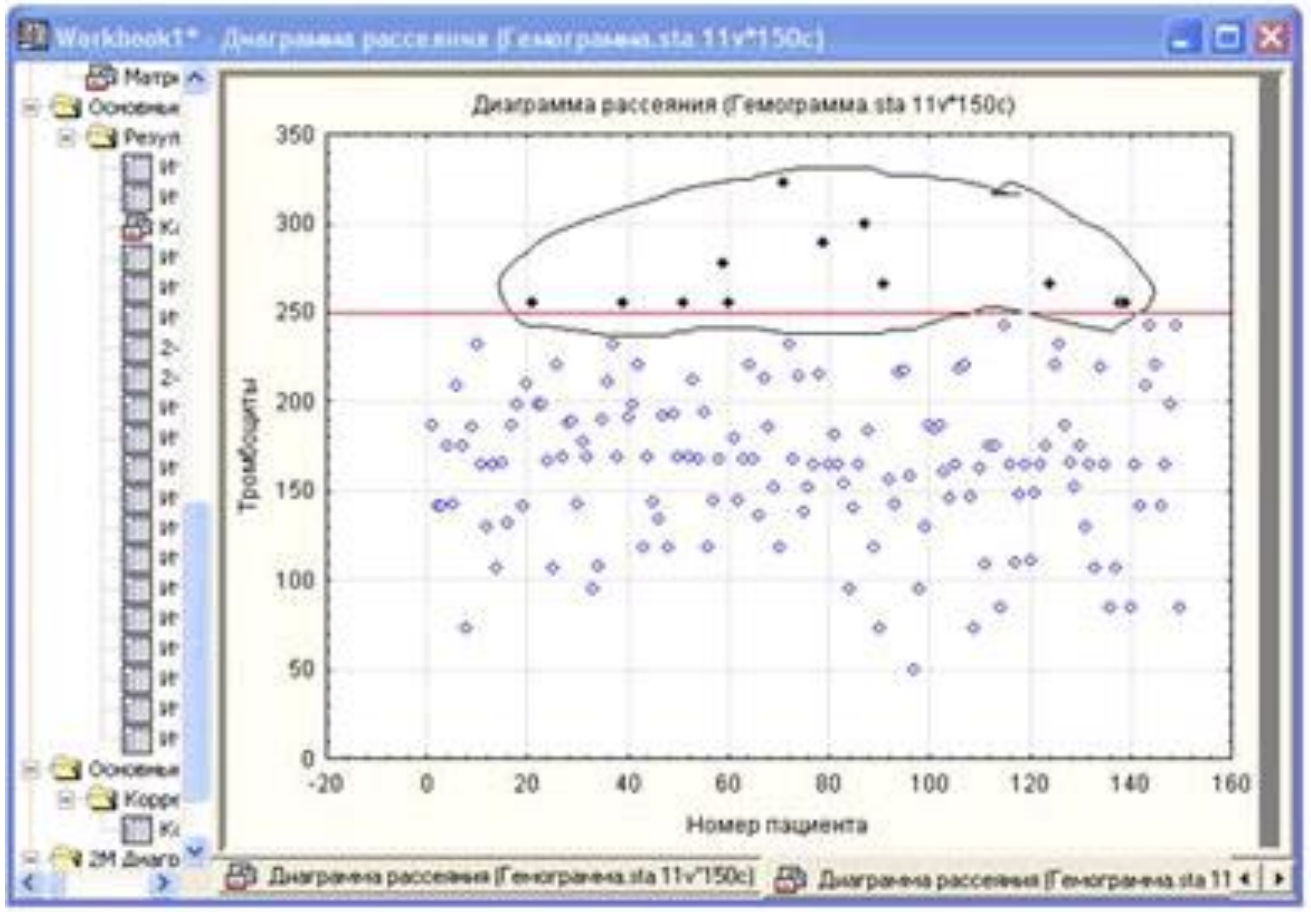
• Наприклад, якщо ми аналізуємо дані про продажі, ми можемо використовувати фільтрацію, щоб побачити лише продажі, які відбулися в певний період часу, або продукти, що коштують більше певної суми. Це дозволяє нам зосередитися на конкретних аспектах даних, які мають значення для нашого аналізу чи прийняття рішень.

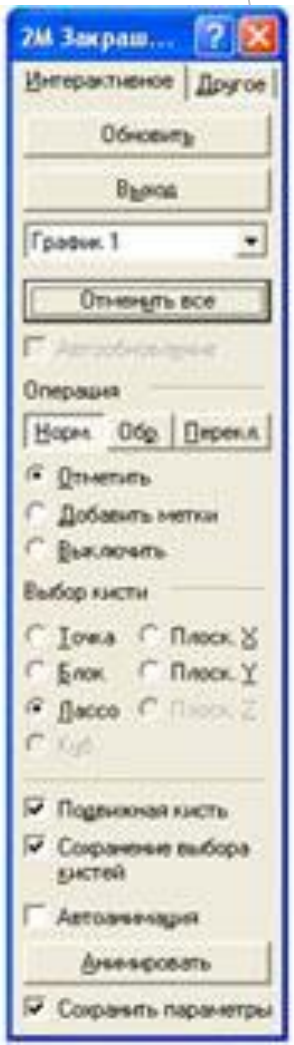
приклад.xlsx

Виявлення аномалій

- Аналіз викидів

Викидами є спостереження, що різко виділяються.
Візуальний аналіз діаграми розсіювання даних





Агрегація:

Бази даних

База даних, складається із набору взаємопов’язаних даних і набору програмних засобів для управління даними та доступу до них.

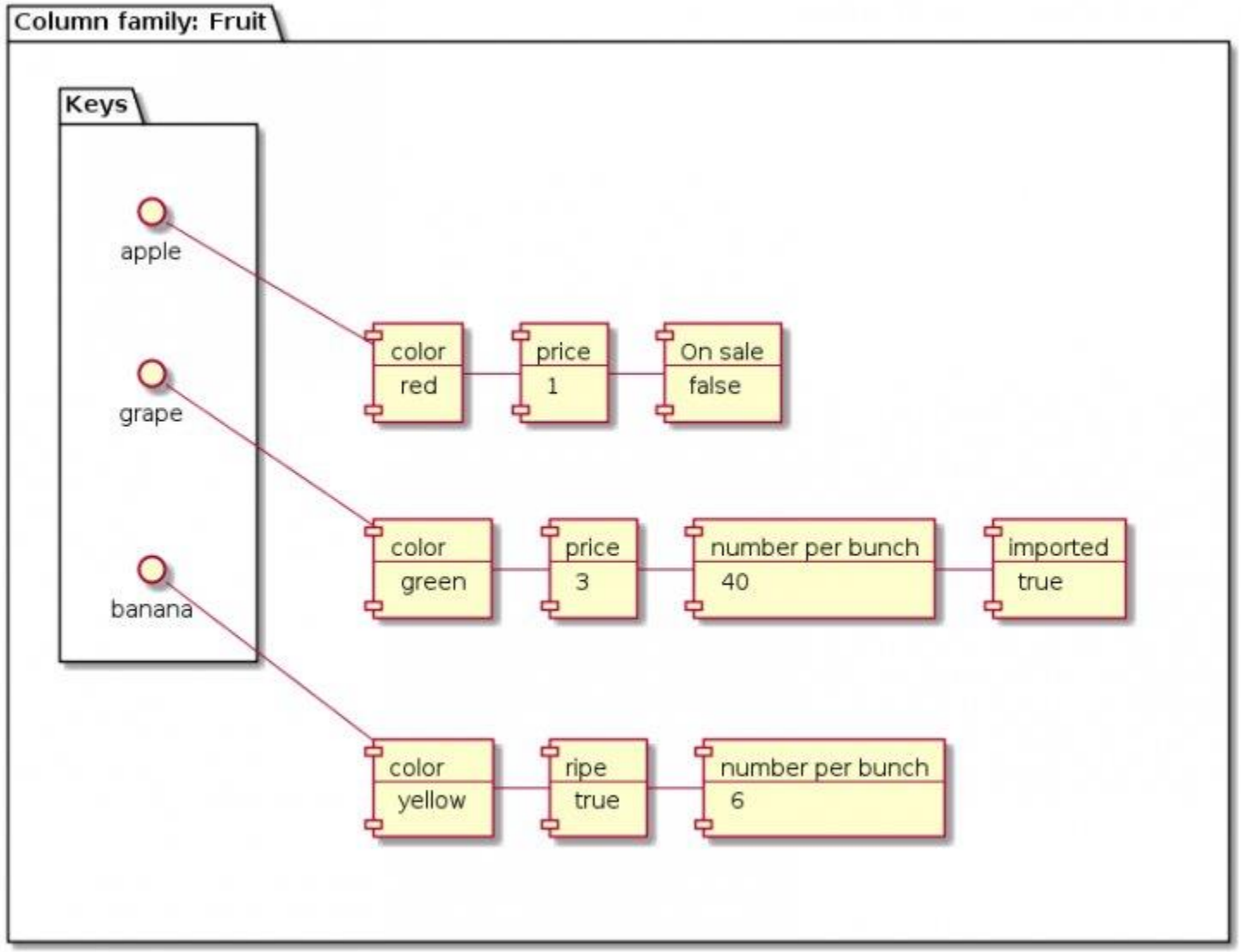
Програми забезпечують:

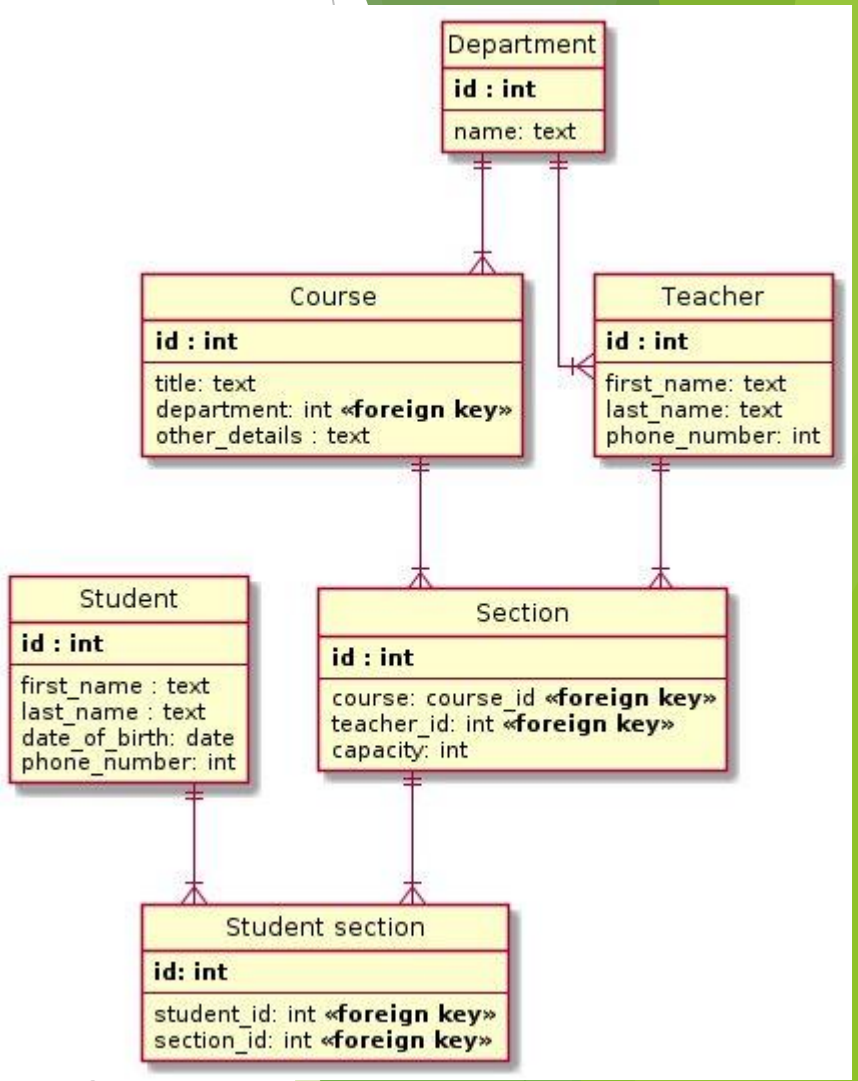
- механізми взаємодії елементів баз даних та зберігання даних;
- управління одночасним, спільним або розподіленим доступом до даних;
- безпеку інформації, що зберігається, незважаючи на збої системи або спроби несанкціонованого доступу.

https://ua.puma.com/ru/size-guide.html

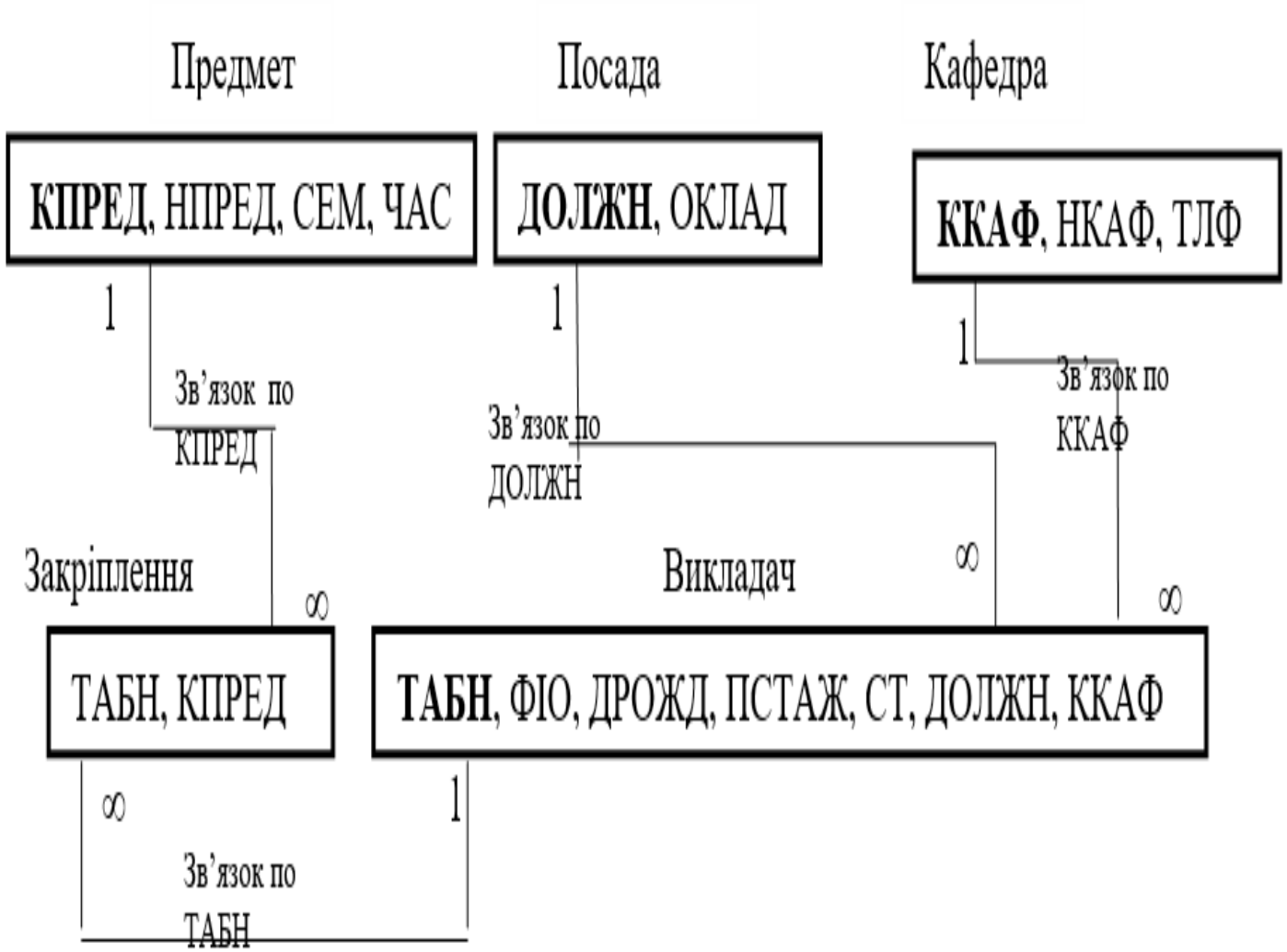
ТИПИ БАЗ ДАНИХ

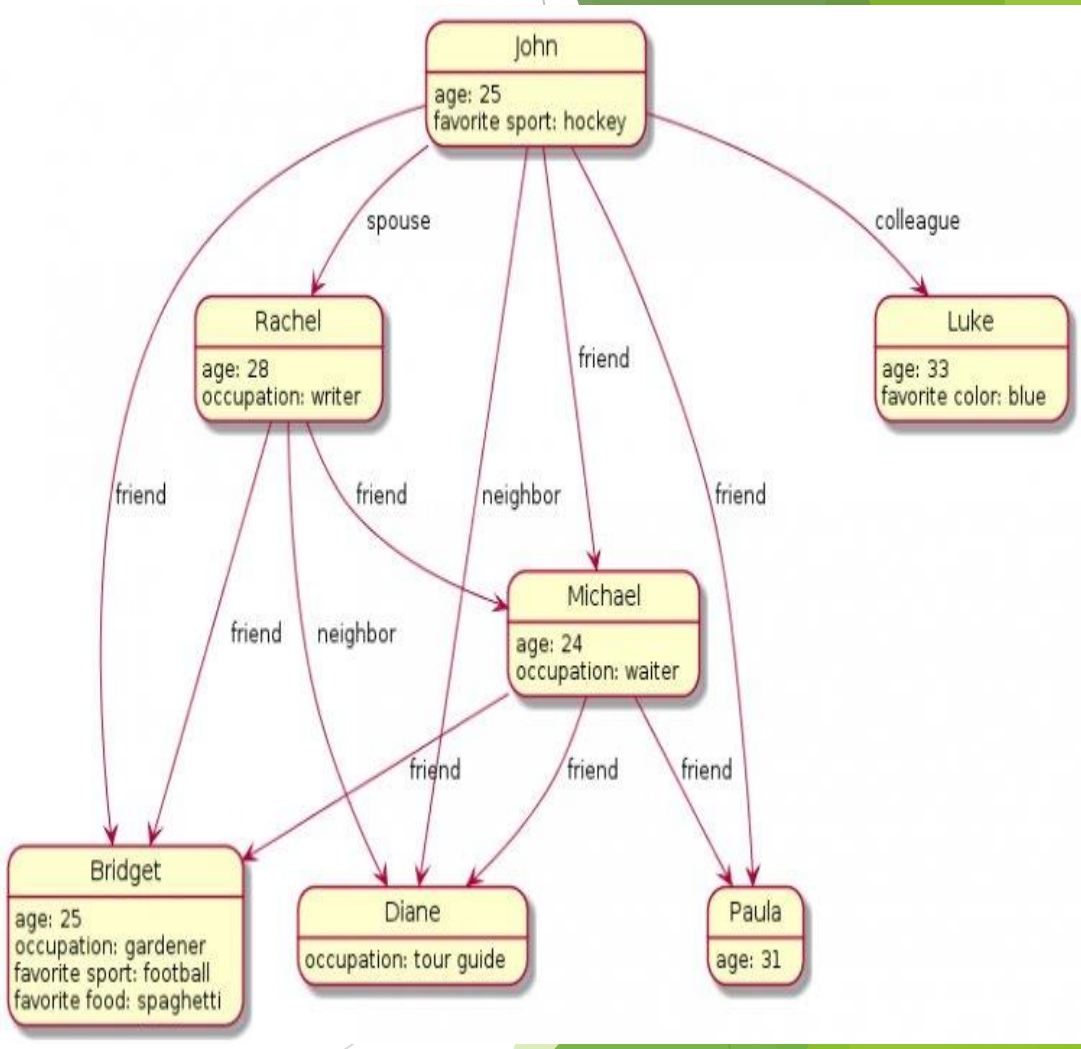
- **Ієрархічна модель бази даних** — структура даних, які впорядковані за підляганням від загального до конкретного.
- **Мережева модель бази даних** — подібна до ієрархічної, але між елементами довільний, не обмежений кількістю елементів — зв’язок.
- **Реляційна модель бази даних** — являє собою одну таблицю або сукупність взаємопов'язаних двовимірних таблиць.

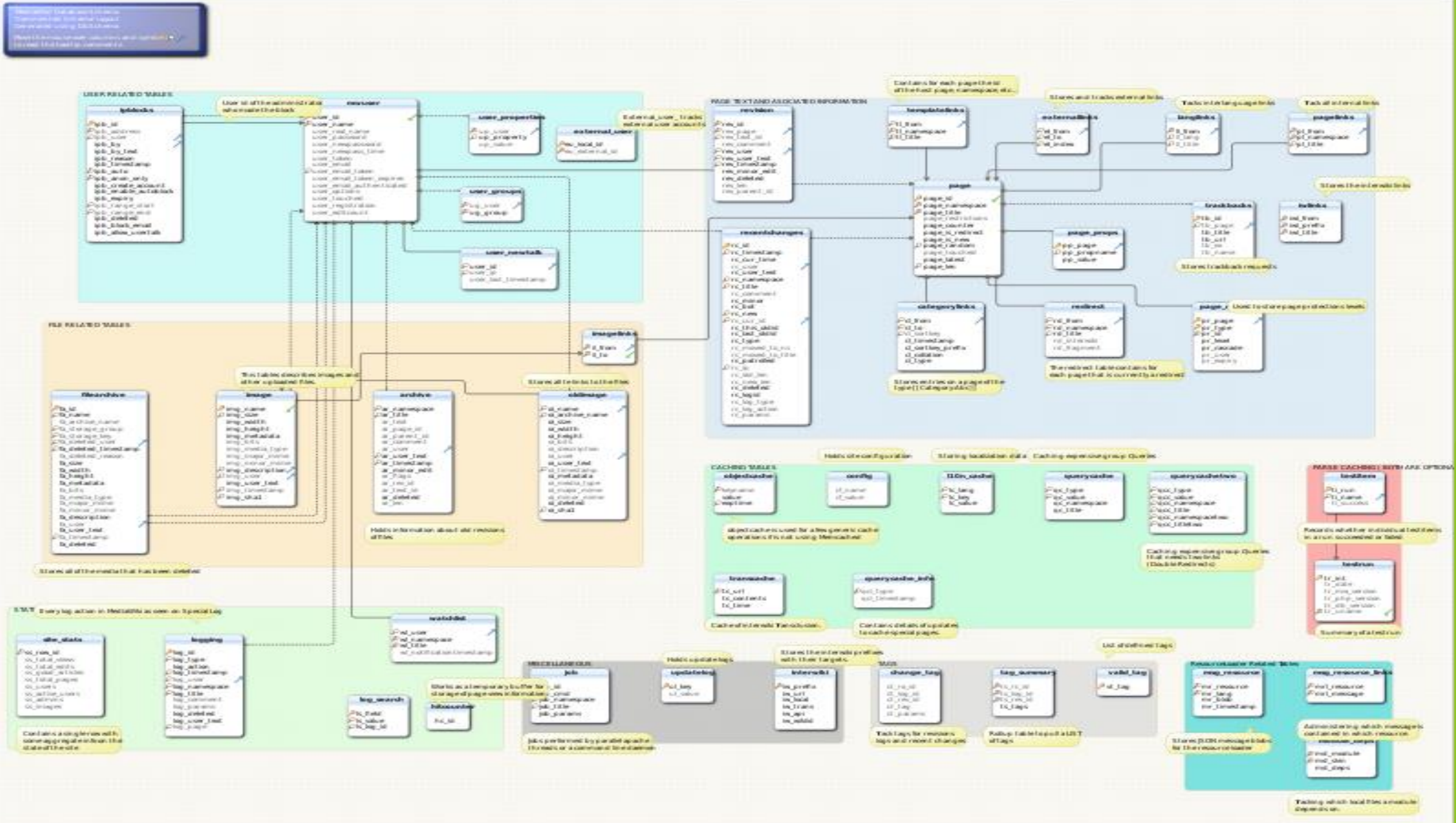

Column family: Fruit
Keys
apple
grape
banana
color
red
price
1
On sale
false
color
green
price
3
number per bunch
40
imported
true
color
yellow
ripe
true
number per bunch
6


Department
id : int
name: text
Course
id : int
title: text
department: int «foreign key»
other_details : text
Teacher
id : int
first_name: text
last_name: text
phone_number: int
Student
id : int
first_name : text
last_name : text
date_of_birth: date
phone_number: int
Section
id : int
course: course_id «foreign key»
teacher_id: int «foreign key»
capacity: int
Student section
id: int
student_id: int «foreign key»
section_id: int «foreign key»

7.7.5. Логічна структура бази даних *Кафедри університету*





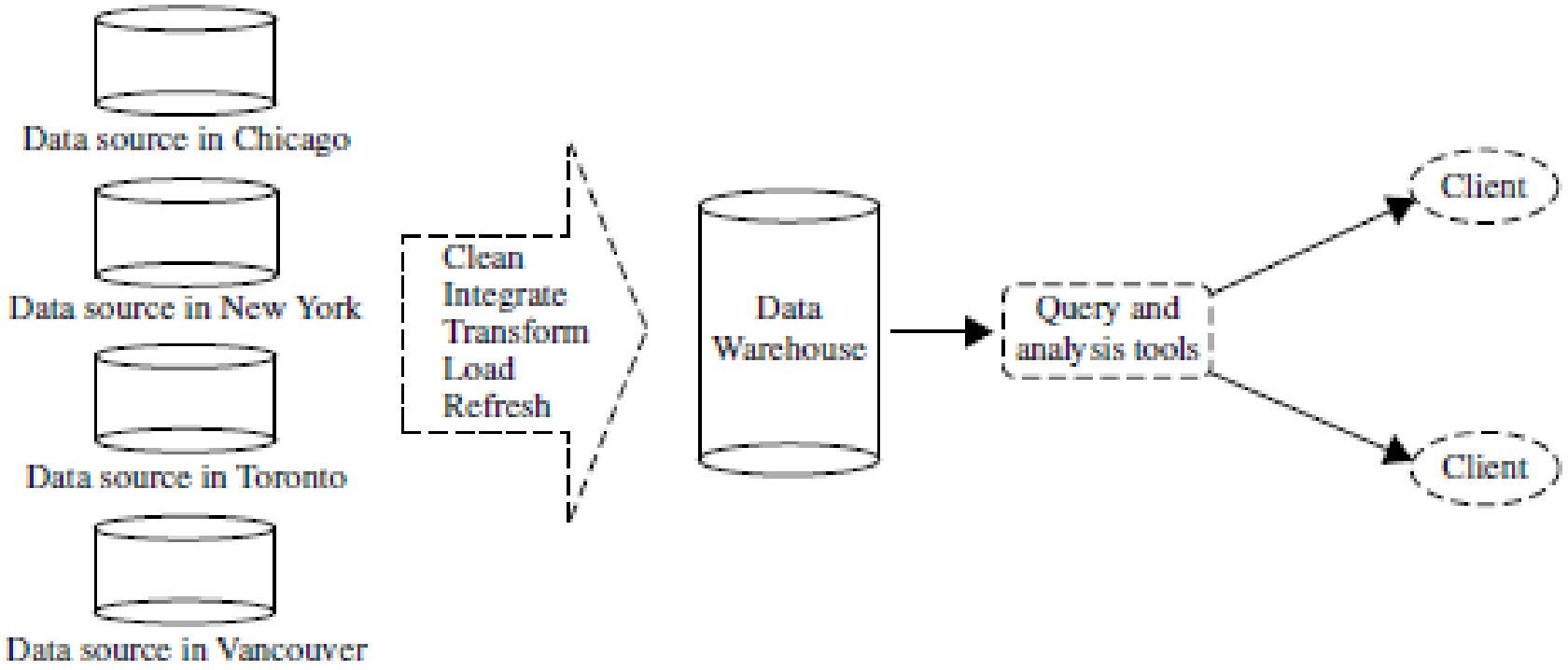

USER RELATED TABLES
ipblocks
user
user_properties
external_user
user_groups
user_newtalk
PAGE TEXT AND ASSOCIATED INFORMATION
revision
templatelinks
externallinks
langlinks
pagelinks
page
trackbacks
iwlinks
recentchanges
categorylinks
redirect
FILE RELATED TABLES
imagelinks
filearchive
image
oldimage
CACHING TABLES
objectcache
querycache
querycachetwo
transcache
querycache_info
PARSE CACHING WITH AVE OPTIONS
linkstore
text
searchindex
site_stats
logging
watchlist
log_search
job
updatelog
interwiki
change_tag
tag_summary
valid_tag
recentchanges
user_former_groups
archive

Сховище даних - це спеціальна база даних, яка зберігає інформацію з різних джерел для аналітичних цілей.

Сховище даних має такі особливості:

1. Воно орієнтоване на певну предметну область, наприклад, продажі, фінанси, маркетинг тощо.
2. Воно інтегрує дані з різних систем і форматів у єдину структуру, що спрощує доступ і запити.
3. Воно зберігає дані в історичній перспективі, тобто враховує зміни даних за часом і дозволяє проводити порівняльний аналіз.
4. Воно незмінне, тобто дані, які потрапили в сховище, не можуть бути модифіковані або видалені.

Схема роботи сховища даних



Сховища даних будуються за допомогою процесу очищення даних, інтеграції даних, перетворення даних, завантаження даних та періодичного оновлення даних.

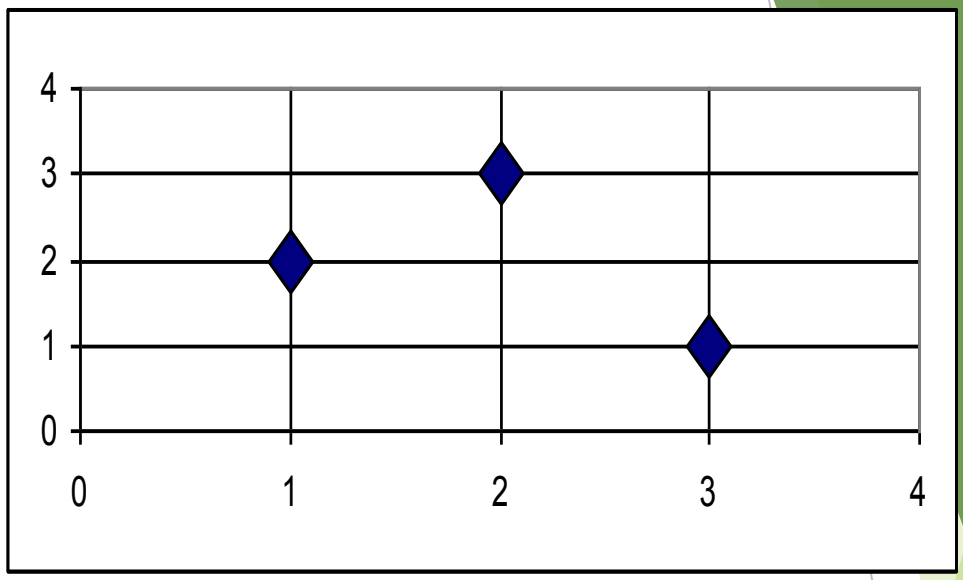
Кластеризація

Кластерізація - це метод аналізу даних у сфері машинного навчання та датамайнінгу, який допомагає групувати схожі об'єкти разом на основі їх характеристик чи спільних властивостей. У контексті студентів університету, це означає розділення студентів на групи або "кластери" на підставі схожості їхніх атрибутів, таких як результати екзаменів, спеціалізація, інтереси тощо.
Кластерізація дозволяє ідентифікувати приховані закономірності та групи серед даних, що можуть бути важливими для подальшого аналізу або прийняття рішень..

Кластерний аналіз

Відстані між двома об'єктами

$$d_S(x_i;y_i)=\left(\sum_{i=1}^{Nf}\left|x_i-y_i\right|^{p}\right)^{\frac{1}{r}}$$



Представлення трьох об’єктів,
як точок на площині

$$d_M(x_i;y_i)=\sqrt{\sum_{i=1}^{Nf}(\sqrt{x_i}-\sqrt{y_i})^2}$$

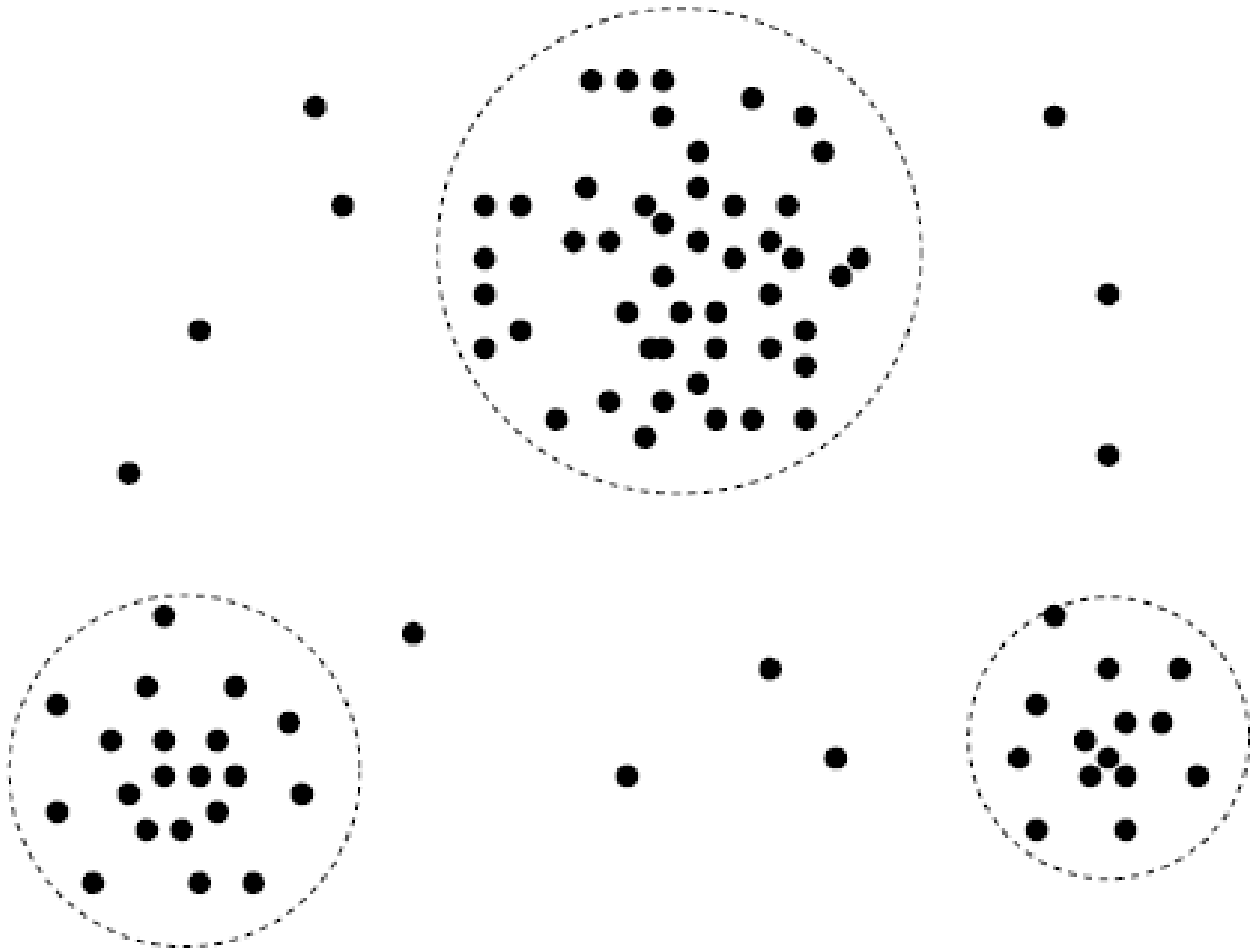
$$d_E(x_i;y_i)=\sqrt{\sum_{i=1}^{Nf}(x_i-y_i)^2}$$

$$d_{XEM}(x_i;y_i)=\sum_{i=1}^{Nf}(x_i-y_i)$$

$$d_{SUP}(x_i;y_i)=SUP\left|x_i-y_i\right|$$

$$d_L(x_i;y_i)=\sum_{i=1}^{Nf}\left|x_i-y_i\right|$$

Кластеризація



Кластеризація повним перебором об'єктів

$$Z = \sum_{i=1}^{N_0}\sum_{j=1}^{N_0} q_{ij} d_{ij} \rightarrow \max \qquad \sum_{i=1}^{N_0} q_{ij} \le N_0 \qquad \sum_{j=1}^{N_0} q_{ij} = 1$$

$$\sum_{i=1}^{N_0}\sum_{j=1}^{N_0} q_{ij} = N_0$$

	1	2	3	4	5
1	0	0,89	0,41	0,74	0,46
2	0,89	0	0,86	0,87	0,61
3	0,41	0,86	0	0,96	0,66
4	0,74	0,87	0,96	0	0,62
5	0,46	0,61	0,66	0,62	0

		Об'єкти					Сума по кластерам
		1	2	3	4	5	
Кластери	1	0	1	0	0	0	1
	2	1	0	0	0	0	1
	3	0	0	0	1	1	2
	4	0	0	1	0	0	1
	5	0	0	0	0	0	0
Сума по постовпцям		1	1	1	1	1	5

Класифікація

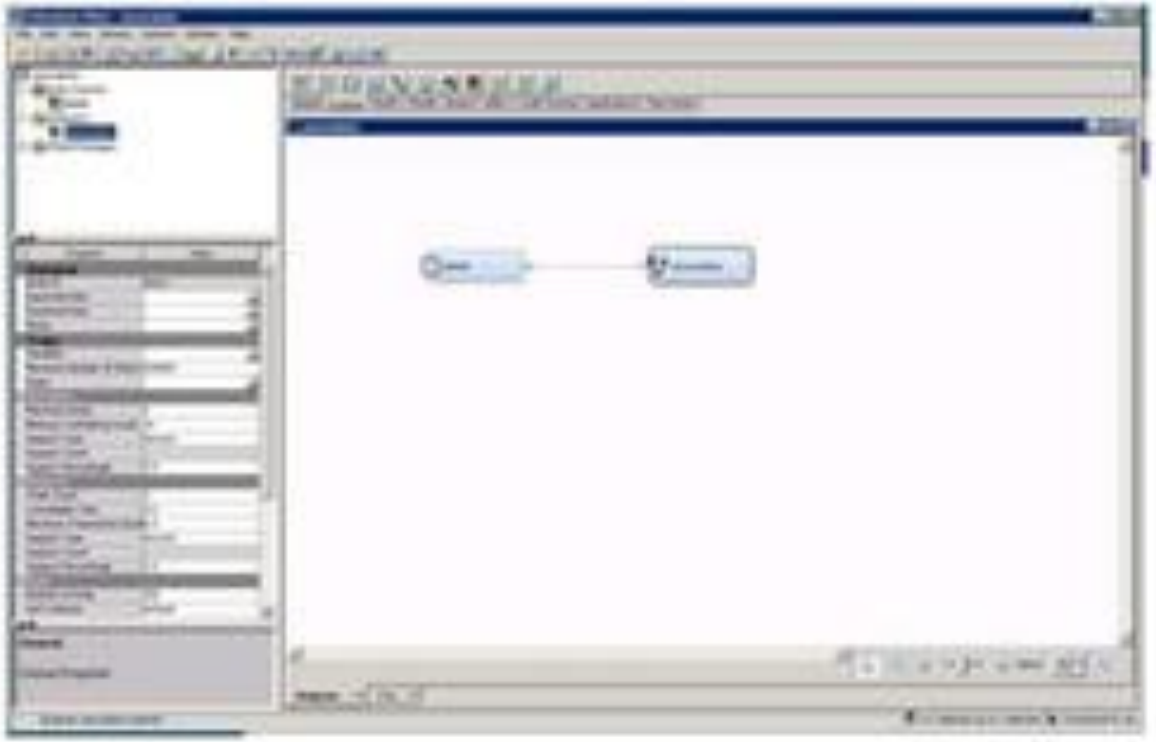
Класифікація — система розподілення об'єктів по групах відповідно до наперед визначених ознак. В деяких випадках, вживають термін категоризація у значенні розподілення об'єктів на категорії.

Метод К-середніх – це частковий випадок GMM, коли коваріація кожного кластера за всіма параметрами прямує до 0. По-друге, оскільки GMM використовує ймовірності при розрахунку, то точки можуть одночасно належати кільком кластерам.

Асоціативний аналіз:

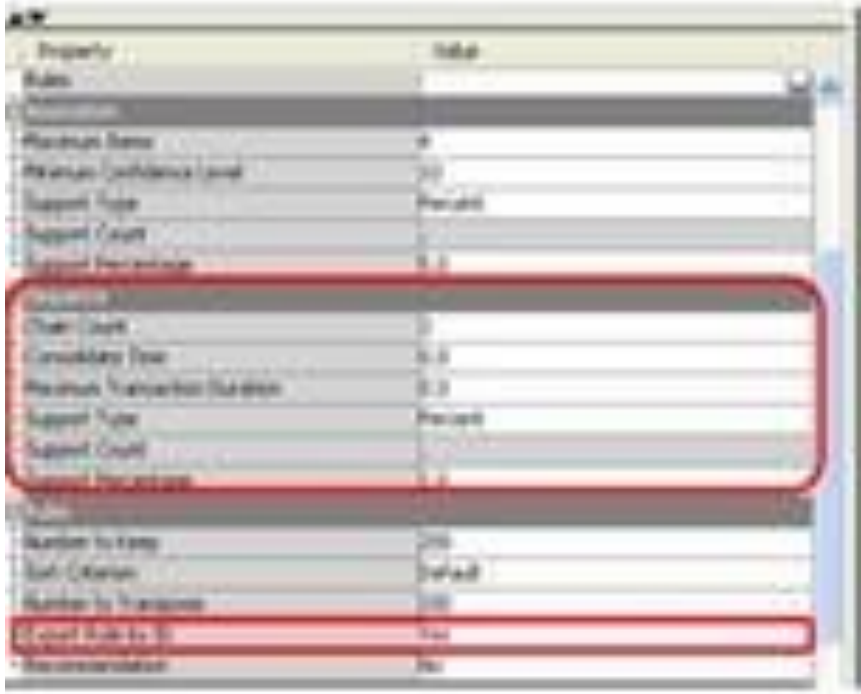
Додання інструмента асоціативного аналізу на діаграму

- Меню ***Explore***, інструмент ***Association***
- З'єднання інструмента асоціативного аналізу з джерелом



Додання ще одного інструмента асоціативного аналізу на діаграму

- Меню ***Explore***, інструмент ***Association***
- З'єднання інструмента з джерелом та перейменування в **Sequence Analysis**



Регресійний аналіз:

Синтез статистичних лінійних та квазілінійних моделей

Можливі перетворення

$$y = a_0 + \sum_{i=1}^{K} a_i x_i \qquad y = a_0 \ell^{a_i x_i} \qquad y = a_0 \log_n x$$

$$y = a_0 + \sum_{i=-n}^{-1} a_i x^i \qquad x_1\ x_2,\ x_1/\ x_2,\ x_1\text{-}x_2,\ \log\ x_1\ x_2,$$

Приклад нормування-денормування

$y = a_0 + a_1x + a_2x^2$

$$y = \left[M_y + \sigma_y \left(a_0 - \frac{a_1 M_x}{\sigma_x} - \frac{a_2 M_{x2}}{\sigma_{x2}} \right) \right] + \frac{a_1 \sigma_y}{\sigma_x} x + \frac{a_2 \sigma_y}{\sigma_{x2}} x^2$$

$\mathrm{Ln}y = a_0 + a_1 * \mathrm{Ln}x_1 + a_2 * \mathrm{Ln}x_2$

$$\mathrm{Ln}y = \left[M_y + \sigma_y \left(a_0 - \frac{a_1 M_x}{\sigma_x} - \frac{a_2 M_{x2}}{\sigma_{x2}} \right) \right] + \frac{a_1 \sigma_y}{\sigma_x} \mathrm{Ln}x_1 + \frac{a_2 \sigma_y}{\sigma_{x2}} \mathrm{Ln}x_2$$

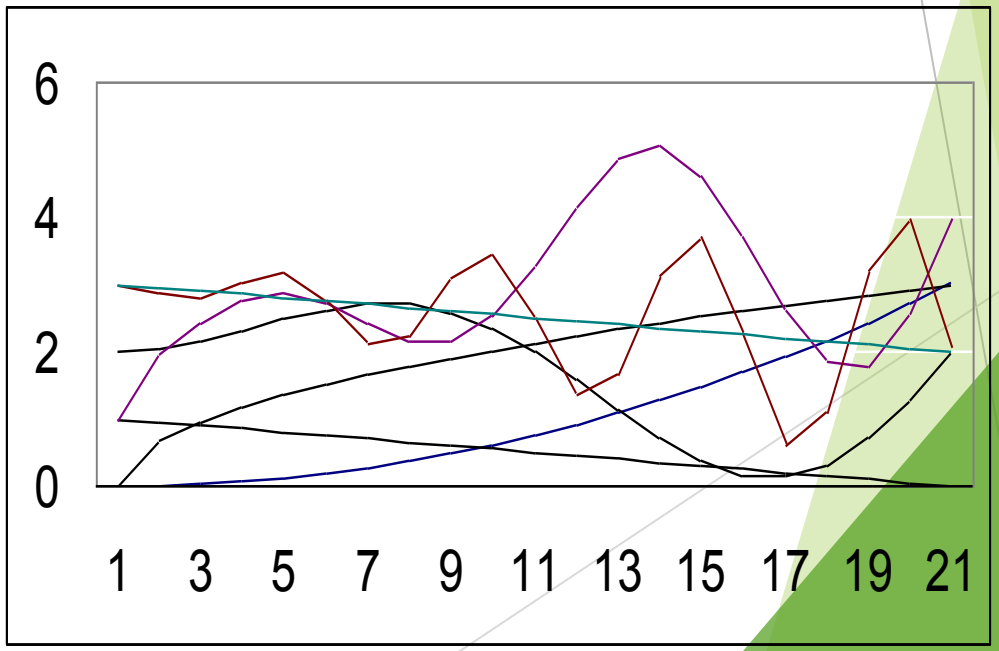
Синтез авторегресійних моделей

Рентабельність	Попереднє значення рентабельності
5%	-
4%	5%
1%	4%
12%	1%

Синтез періодичних моделей

$$y = Ax^B + C(1 - e^{Dx})$$

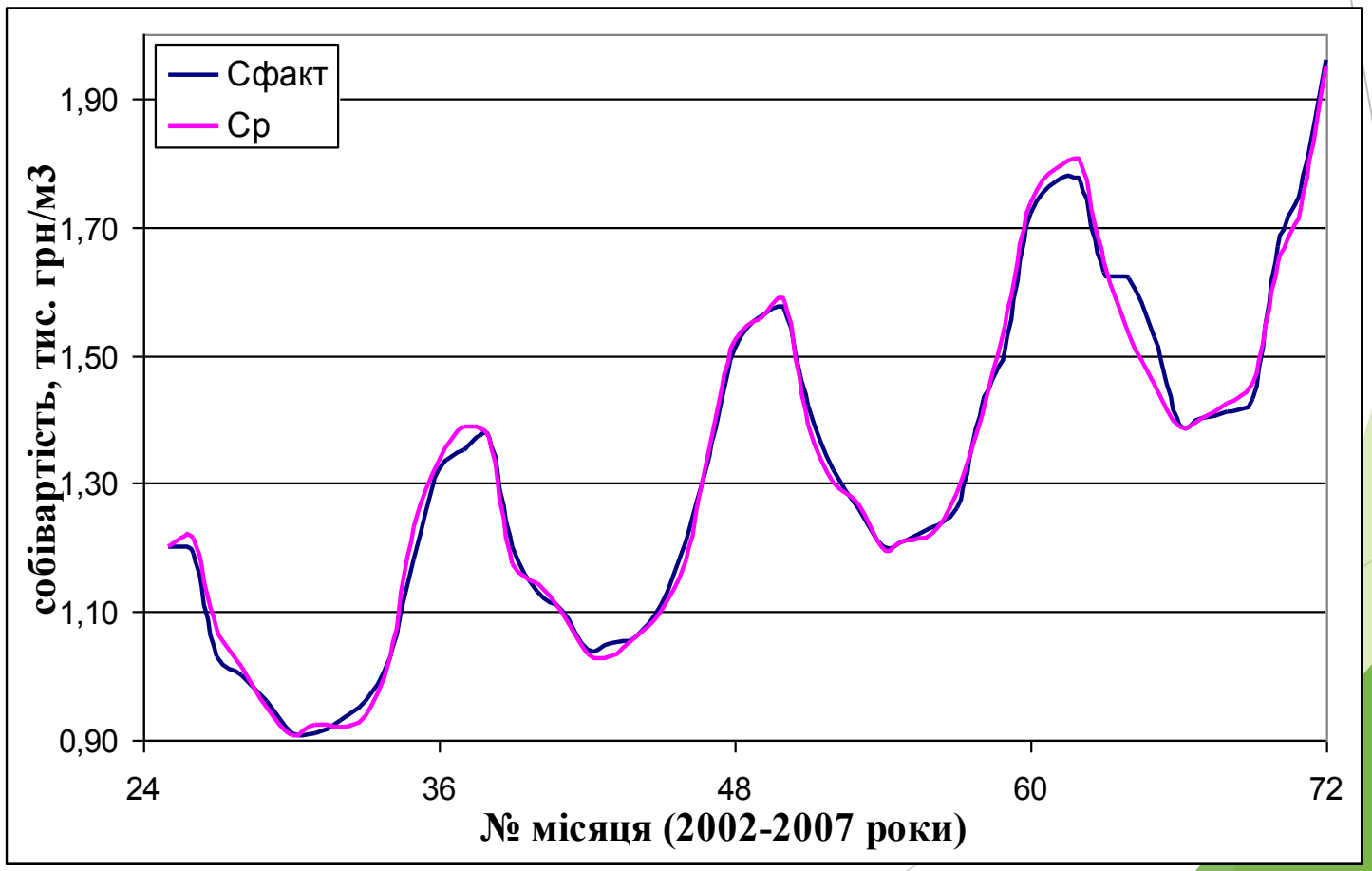
$$Sin(Ex^F + G) + H$$



Приклад періодичної моделі

$$C_{t,1}=1{,}342c_{t-12,1}^{\,0{,}887}+0{,}525(1-e^{-0{,}976c_{t-12,1}})\sin(0{,}524c_{t-12,1}^{\,1{,}187}+0{,}664)-$$

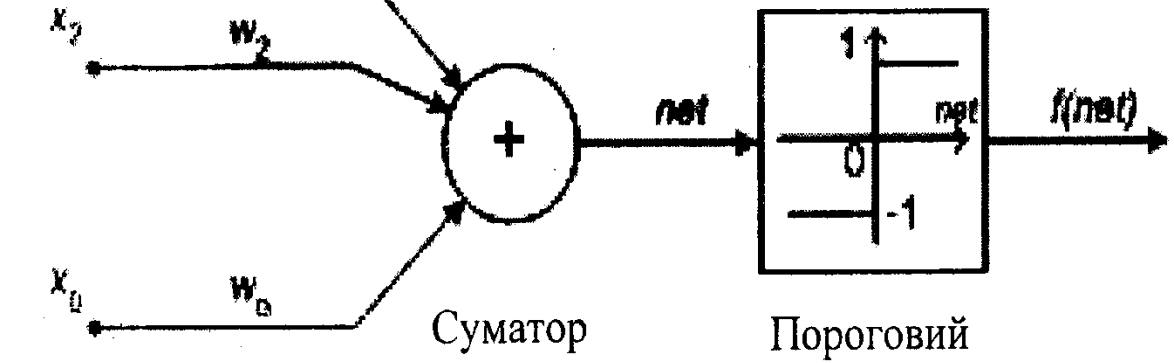
$$-0{,}402c_{t-24,1}^{\,0{,}686}+0{,}288(1-e^{-0{,}106c_{t-24,1}})\sin(0{,}524c_{t-24,1}^{\,0{,}808}+0{,}195)-0{,}146$$



Нейронні сітки

$$net = \sum_{i=1}^{n} x_i w_i$$

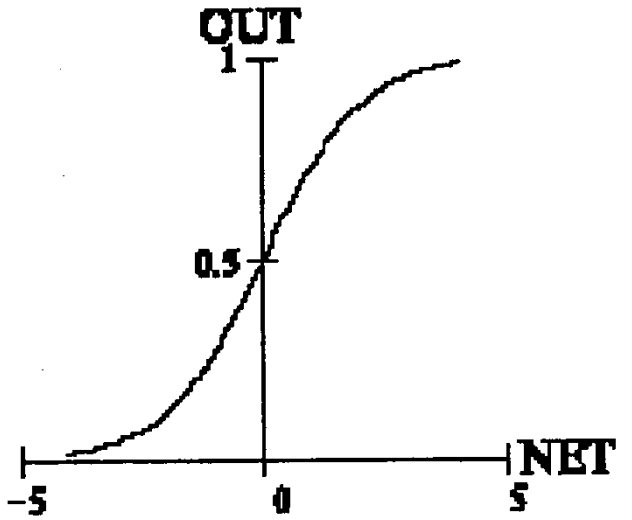
$$f(net) = \mathrm{sgn}(net) = \begin{cases} +1, & \mathrm{net} > 0 \\ -1, & \mathrm{net} < 0 \end{cases}$$



Вхід Вага Суматор Пороговий логічний елемент

Модель граничного нейрона МакКаллоха-Піттса

$$f(net) = \frac{1}{1 + e^{-\lambda \cdot net}}$$



Вид логістичної функції

Алгоритм навчання перцептрону

1. Початкові ваги можуть бути будь-якими. Корекція провадиться пропорційно величині похідної по даній координаті. Похідна береться від функції активації. Підстроювання *j* ваги для *i* нейрона здійснюється за формулою

, $\Delta w_{ij} = \eta \cdot [d_i - f(net_i)] \cdot f'(net_i) \cdot x_j$ де *j=1,2,...,n* -

коефіцієнт навчання, підбирається евристично

2. Помилка при навчанні на *k* кроці: $E_k = \frac{1}{2}[d_i - f(net_i)]^2$

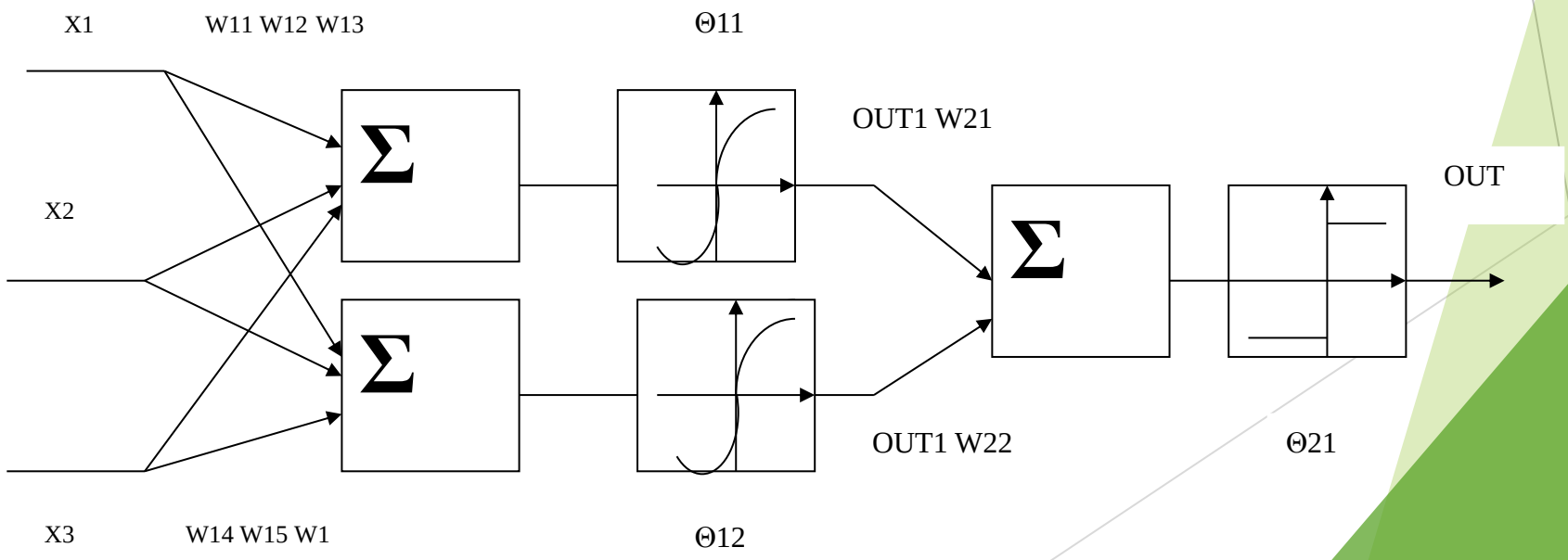
де *di* - очікуваний вихід

3. Загальна помилка при навчанні: $E = \frac{1}{2 \cdot p} \sum_{k=1}^{p} [d_i - f(net_i)]^2$

де *p* - число прикладів у навчальній вибірці

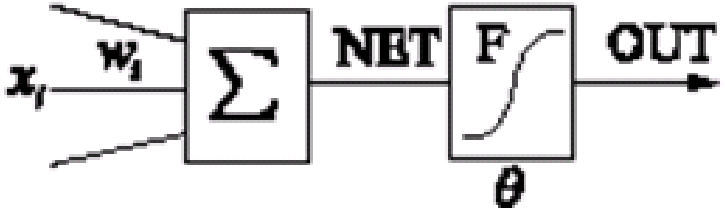
4. Похідна від сигмоїди $f'(net) = \lambda \cdot f(net) \cdot [1 - f(net)]$

де *p* - число прикладів *у* навчальній вибірці.

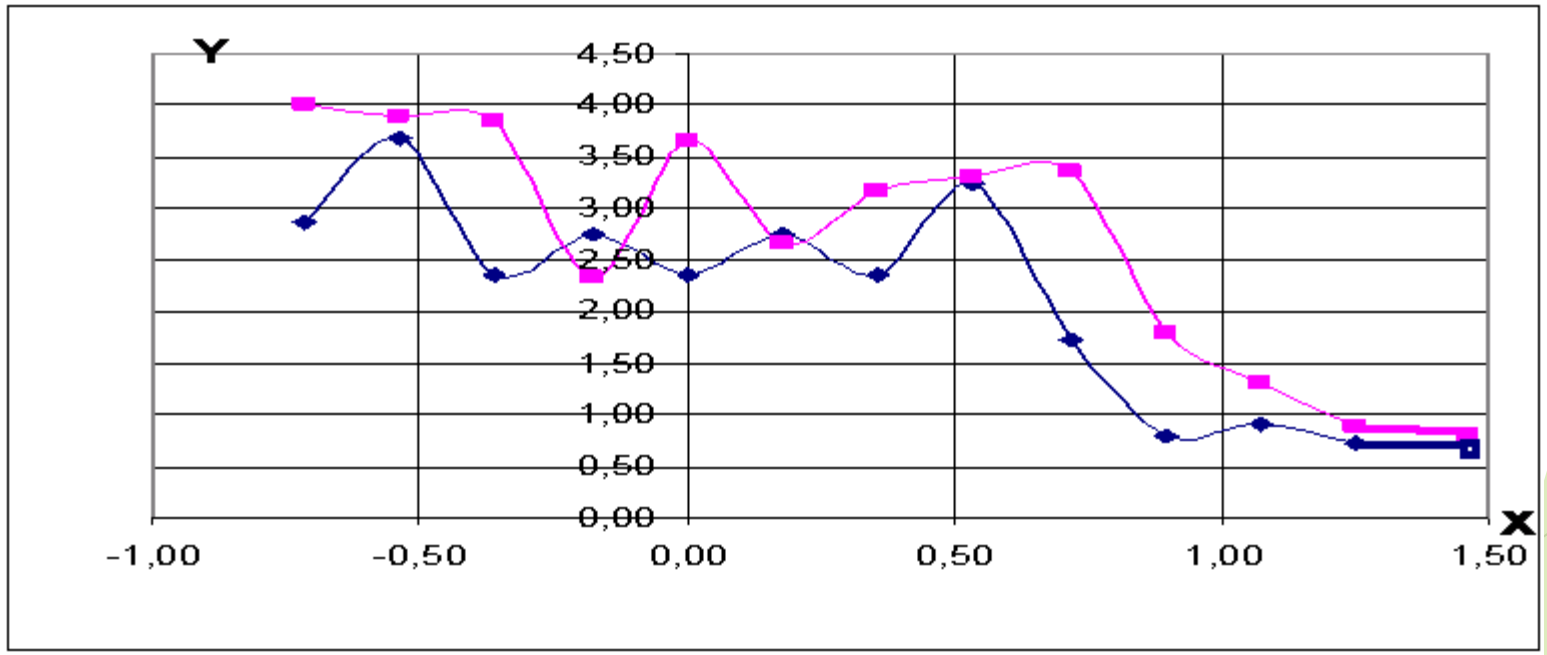


Приклади нейронних сіток

Загальний вигляд схеми перцептрона з одним нейроном та суматором на вході

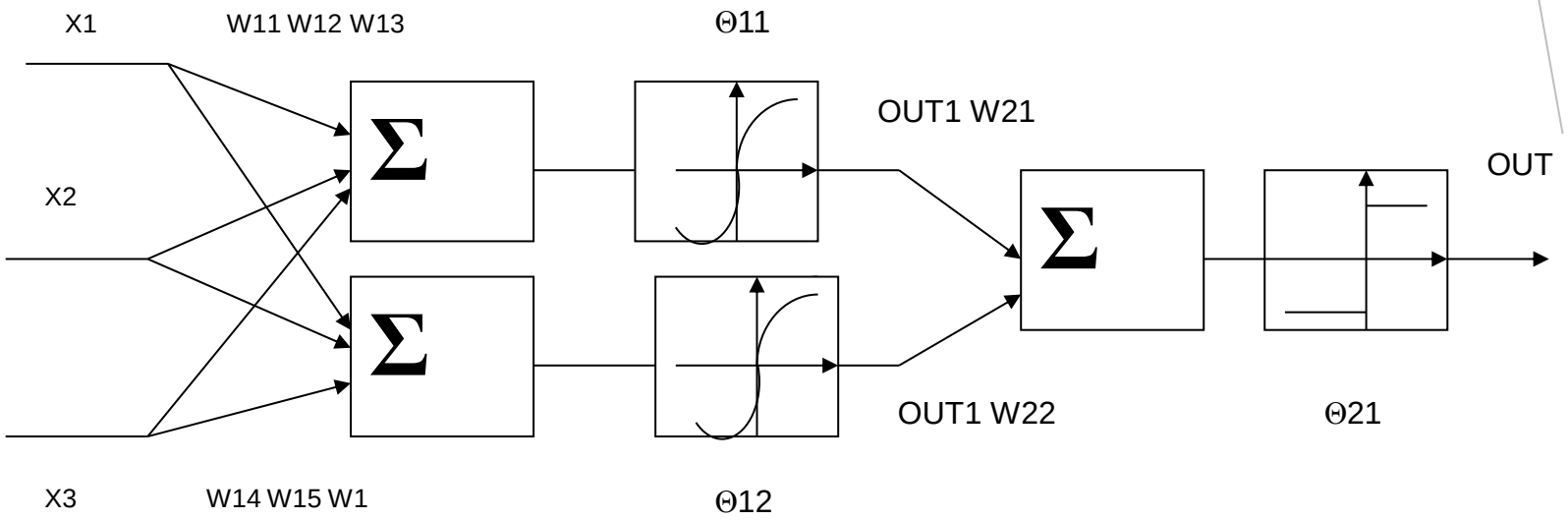


$$OUT = \frac{1}{1 + \ell^{-(w_1 x_i + w_2 x_{i+1} + w_3 x_{i+2})}}$$



Графік кількості викликів Y
по годинам робочого дня X.
(♦ – експериментальна крива,
■ – розрахована крива)

Схема двошарового перцептрона з трьома входами на кожному нейроні



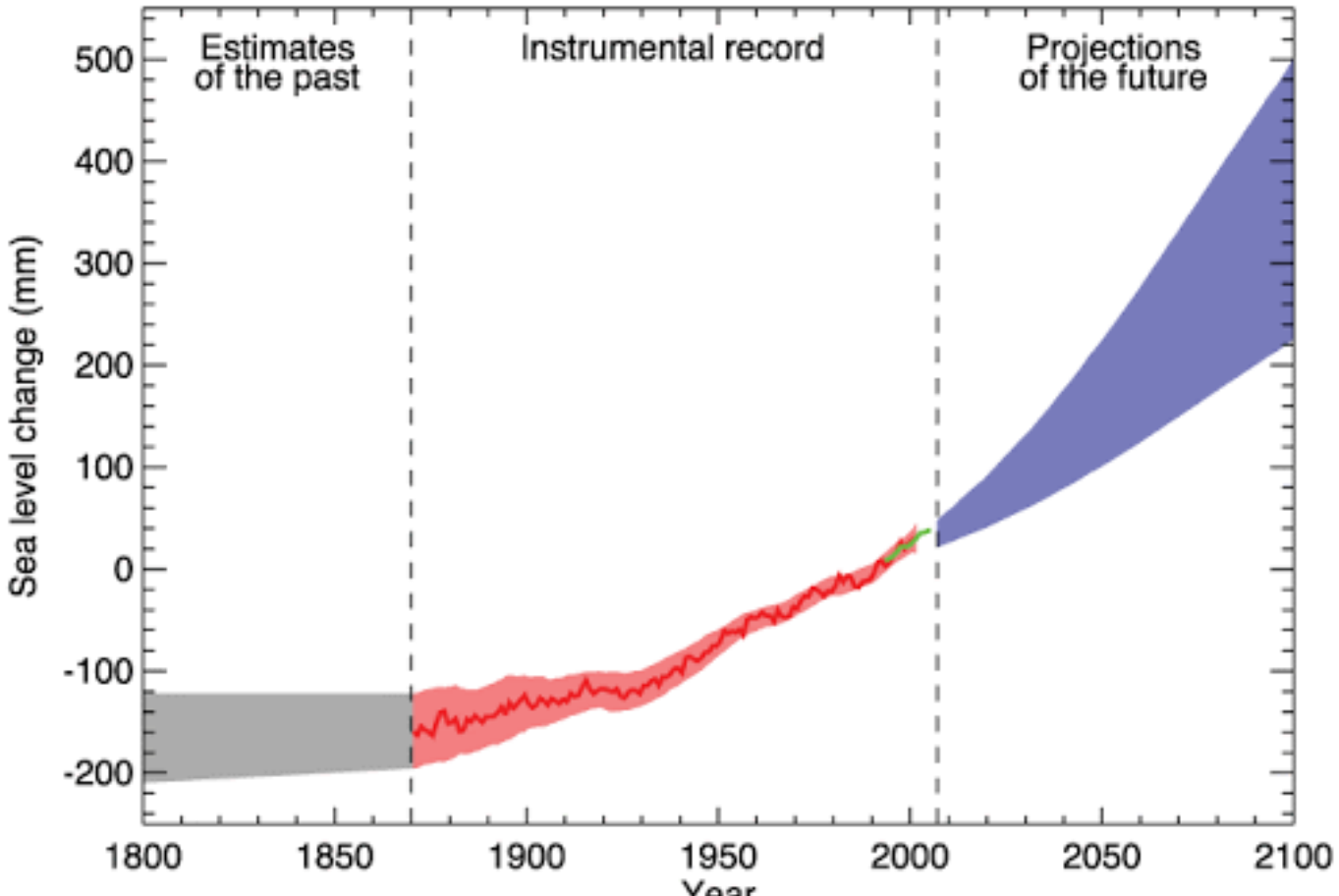
$$OUT = Sign\left[\frac{10{,}37669298}{\left[1+\ell^{\,-(-1{,}443525248X_1+1{,}561436345X_2+1{,}582503396X_3+4{,}683681367)}\right]}+\frac{7{,}940045065}{\left[1+\ell^{\,-(1{,}556474752X_1+4{,}561436345X_2+4{,}582503396X_4+-3{,}683681367)}\right]}+2{,}883681367\right]$$

Модель прогнозування інвестицій на вугільній шахті О.Ю. Чурікановоі

$$OUT = sigm(NET) = \left[\frac{1}{1+\exp\left(\lambda_3 \left\{ \frac{1}{1+\exp\left(\lambda_1 \left(\sum_{i=1}^{26}\sum_{j=1}^{3} x_{ij}^{1} w_{ij}^{1}\right)\right)} + \frac{1}{1+\exp\left(\lambda_2 \left(\sum_{i=1}^{3}\sum_{j=1}^{1} x_{ij}^{2} w_{ij}^{2}\right)\right)} \right\}\right)} \right],$$

Прогнозування

Прогнозування — процес передбачення майбутнього стану предмета чи явища на основі аналізу його минулого і сучасного, систематично оцінювана інформація про якісні й кількісні характеристики розвитку обраного предмета чи явища в перспективі. Результатом прогнозування є знання про майбутнє і про ймовірний розвиток сьогочасних тенденцій конкретного явища-об'єкту в подальшому існуванні.



Прогнозування часового ряду методом ковзної середньої

для лінійної інтерполяції

нійної інтерполяції

$$y_t^* = \frac{\sum_{t=t-p}^{t+p} y_t}{m}, \quad t > p.$$

$$y_t^* = \frac{\sum_{t=t-p}^{t+p} p_t y_t}{\sum_{t=t-p}^{t+p} p_t},$$

причому вага p_t визначається за допомогою методу найменших квадратів. Ця вага обчислюється для різних степенів поліному апроксимації і різних інтервалів згладжування. Так, для поліномів другого та третього степенів чисельна послідовність ваги при інтервалі згладжування $m = 5$ має значення [-3, 12, 17, 12, -3], а при $m = 7$ — [-2, 3, 6, 7, 6, 3, -2]. Для поліномів четвертої та п'ятої степенів при $m = 7$ послідовність ваги буде [5, -30, 75, 131, 75, -30, 5].

Експоненційне згладжування

основні формули

значення коефіцієнту *a* обирається в межах 0,1 - 0.3

В окремих випадках Браун запропонував визначати величину параметра згладжування залежно від довжини згладжування ряду за формулою

$$y_t^* = \alpha y_t + (1-\alpha) y_{t-1}^*, \qquad \alpha = \frac{2}{n+1}.$$

Відома також для визначення величини α формула Мейєра вигляду

$$y_t^* = \alpha \sum_{i=0}^{t-1} (1-\alpha)^i y_{t-i} + (1-\alpha)^t y_0, \qquad \alpha \approx \frac{\sigma_n}{\sigma_\varepsilon},$$

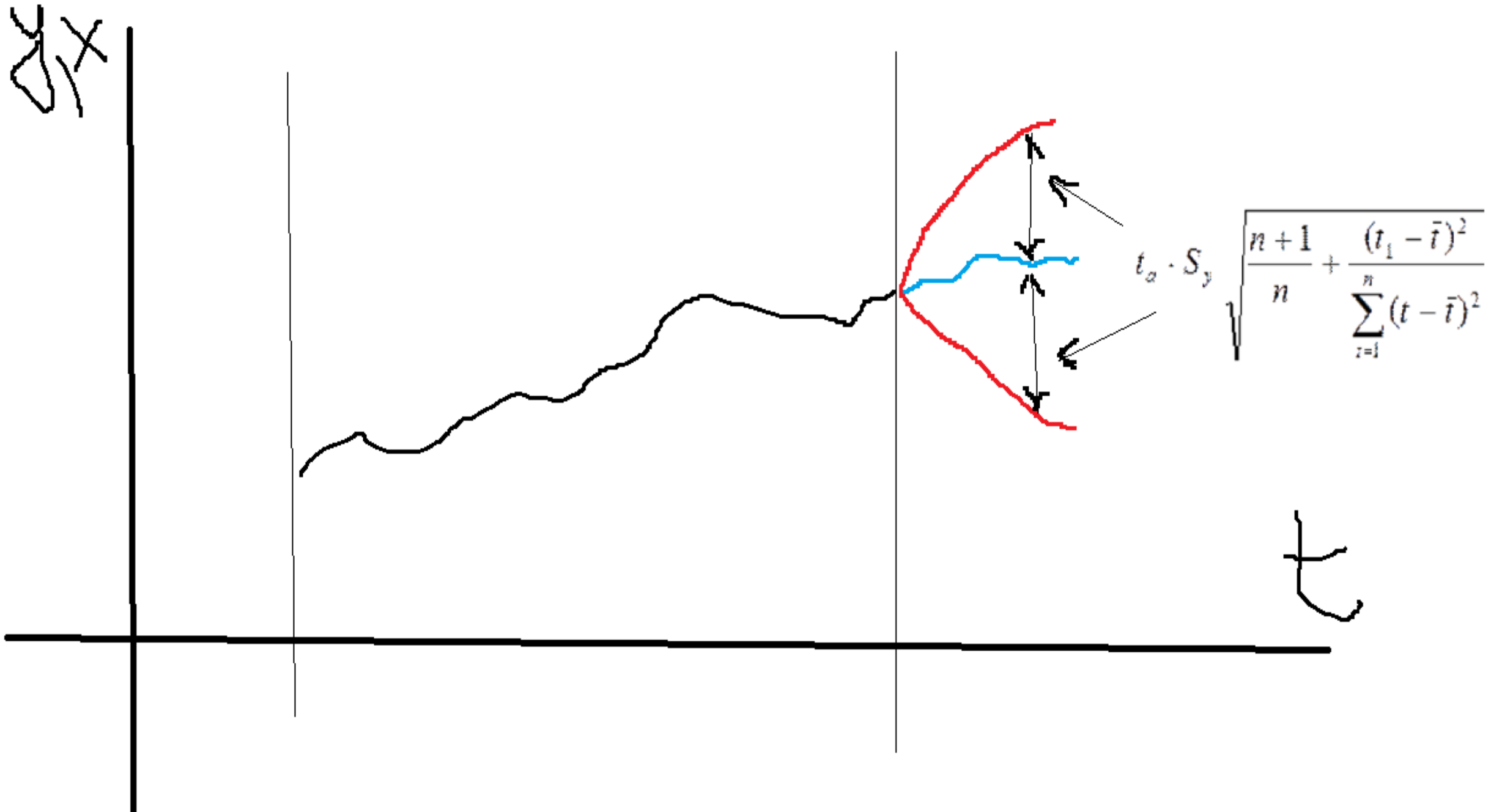
$y_0 = \dfrac{y_1 + y_2 + y_3}{3}$ де $\sigma_n, \sigma_\varepsilon$ – середньоквадратичне відхилення моделі та вихідного ряду відповідно.

Розрахунок довірчих інтервалів прогнозу

n - довжина часового ряду;
L- період випередження;
y_{n+L} - точковий прогноз на момент n + L;
ta - значення t-статистики Стьюдента;
S_y - середня квадратична помилка ряду;
t- порядковий номер рівнів ряду, t = 1,2, ..., n.
Тоді, для поліному першого порядку, довірчий інтервал:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

$$\hat{y}_{n+L} \pm t_\alpha \cdot S_y \sqrt{\frac{n+1}{n} + \frac{(t_i - \bar{t})^2}{\sum_{i=1}^{n}(t - \bar{t})^2}}$$

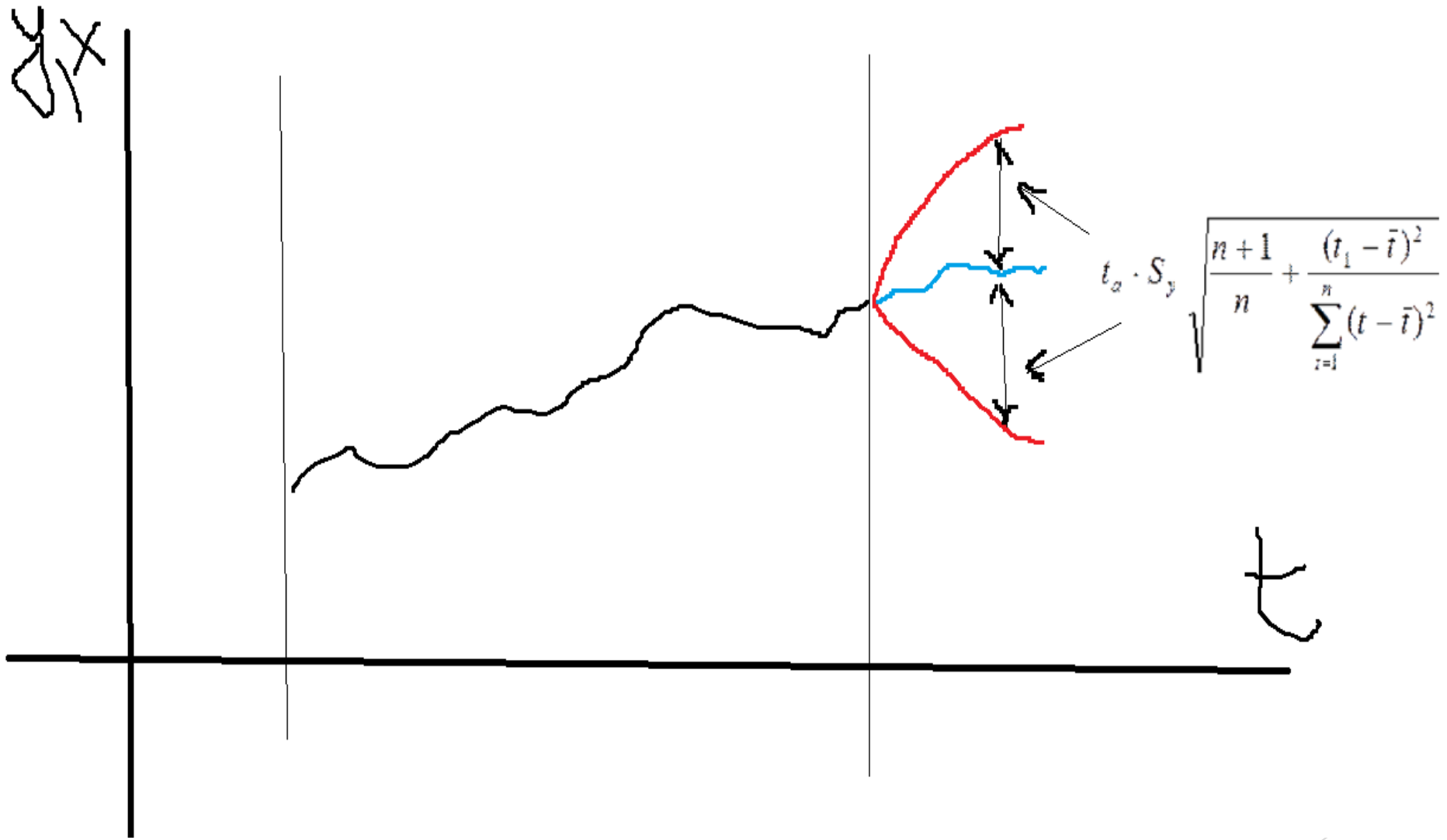

x
t
$t_\alpha \cdot S_y \sqrt{\frac{n+1}{n} + \frac{(t_1 - \bar{t})^2}{\sum_{i=1}^{n} (t - \bar{t})^2}}$

Розрахунок довірчих інтервалів прогнозу

n - довжина часового ряду;
L- період випередження;
y_{n+L} - точковий прогноз на момент n + L;
ta - значення t-статистики Стьюдента;
S_y - середня квадратична помилка ряду;
t- порядковий номер рівнів ряду, t = 1,2, ..., n.
Тоді, для поліному першого порядку, довірчий інтервал:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

$$\hat{y}_{n+L} \pm t_\alpha \cdot S_y \sqrt{\frac{n+1}{n} + \frac{(t_i - \bar{t})^2}{\sum_{i=1}^{n}(t - \bar{t})^2}}$$


y/x
t
$t_\alpha \cdot S_y \sqrt{\frac{n+1}{n} + \frac{(t_1 - \bar{t})^2}{\sum_{i=1}^{n} (t - \bar{t})^2}}$

Методи Бокса-Дженкінса

AR(p) -авторегресивна модель порядку p

Модель має вигляд:

$$Y(t) = f_0 + f_1 \cdot Y(t-1) + f_2 \cdot Y(t-2) + \ldots + f_p \cdot Y(t-p) + E(t)$$

де Y(t) –залежна змінна у момент часу t. $f_0, f_1, f_2 \ldots, f_p$ - оцінювані параметри. E(t) - помилка від впливу змінних, які не враховуються в даній моделі. Завдання полягає в тому, щоб визначити $f_0, f_1, f_2 \ldots, f_p$. Їх можна оцінити різними способами. Найправильніше шукати їх через систему рівнянь Юла-Уолкера, для складання цієї системи буде потрібно розрахунок значень автокореляційної функції. Можна поступити простішим способом - порахувати їх методом найменших квадратів.

MA(q) -модель з ковзаючим середнім порядку q

Модель має вигляд:

$$Y(t) = m + e(t) - w_1 \cdot e(t-1) - w_2 \cdot e(t-2) - \ldots - w_p \cdot e(t-p)$$

Де Y(t) -залежна змінна у момент часу t. $w_0, w_1, w_2 \ldots, w_p$ - оцінювані параметри.

Авторегресійне ковзне середнє ARMA(p,q)

Під позначенням ARMA(p,q) [3] розуміється модель, p авторегресійних складових, що містить q, ковзаючих середніх.

Точніше модель ARMA(p,q) включає моделі AR(p) і MA(q):

$$X_t = c + e_t + \sum_{i=1}^{q} \theta_i e_{t-i} + \sum_{i=1}^{p} \phi_i X_{t-i},$$

Зазвичай значення помилки e_t вважають незалежними однаково розподіленими випадковими величинами, узятими з нормального розподілу з нульовим середнім: $e_t \sim N\left(0, \sigma^2\right)$, де σ^2 — дисперсія. Припущення можна ослабити, але це може привести до зміни властивостей моделі. Наприклад, якщо не припускати незалежності і однакового розподілу помилок, поведінка моделі суттєво міняється.

ARIMA (p,d,q)

У завданні аналізу тимчасового ряду з складною структурою часто використовуються моделі класу ARIMA(p,d,q)[2] (авторегресійне інтегрування ковзаючого середнього - Autoregressive Integrated Moving Average) порядку (p,d,q), які моделюють різні ситуації, що зустрічаються при аналізі стаціонарних і нестаціонарних рядів. Залежно від аналізованого ряду модель ARIMA (p,d,q) може трансформуватися до авторегресійної моделі AR(p), моделі ковзного середнього MA(q) або змішаній моделі ARMA (p,q). При переході від нестаціонарного ряду до стаціонарного значення параметра d, що визначає порядок різниці, приймається рівним 0 або 1, тобто цей параметр має тільки цілочисельні значення. Зазвичай обмежуються вибором між d = 0 і d = 1. Проте з поля зору дослідників випадає ситуація, коли параметр d може приймати дробові значення.

ARFIMA(p,d,q)

Для ситуації розгляду дробових значень порядку різниці, в роботах зарубіжних учених, в першу чергу, C.W.Granger, J.R.Hosking, P.M.Robinson, R. Beran, був запропонований новий клас моделей ARFIMA(p,d,q)[2] (F: fractional - дріб), що допускає можливість нецілого параметра d і авторегресійний дріб інтегрований процес ковзного середнього. Такі ряди володіють своєю специфікою: самоподібністю, дробовою розмірністю, поволі спадаючою кореляцією. Прогнозування тимчасових рядів за допомогою моделі ARFIMA(p,d,q) відкриває ширші перспективи для підвищення точності прогнозу.

Модель вигляду ARIMA (p,d,q)(P,D,Q)S

ARIMA (p,d,q)(P,D,Q)S [1],

де: p - авторегресійні доданки;

D - різниці;

q - доданки ковзаючого середнього;

P – сезонні авторегресійні доданки;

D – сезонні різниці на інтервалі S;

Q – доданки сезонного ковзаючого середнього

Текстовий аналіз:

Нечіткі моделі

Функція приналежності $A = \{x/mA(x)>0\}$.

Типи функцій приналежності

Трикутна

$$MF(x)=\begin{cases}1-\dfrac{b-x}{b-a}, & a\le x\le b\\ 1-\dfrac{x-c}{c-b}, & b\le x\le c\\ 0, \text{ в інших випадках}\end{cases}$$

Трапецевидна

$$MF(x)=\begin{cases}1-\dfrac{b-x}{b-a}, & a\le x\le b\\ 1, & b\le x\le c\\ 1-\dfrac{x-c}{d-c}, & c\le x\le d\\ 0, \text{ в інших випадках}\end{cases}$$

Гаусіана

$$MF(x)=\exp\left[-\left(\frac{x-c}{\sigma}\right)^2\right]$$

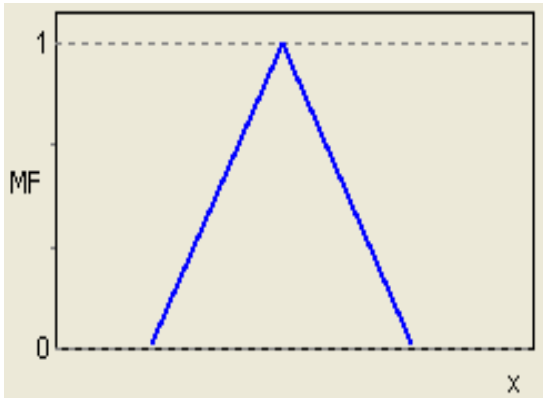
Кругова

$$\mu(x)=\frac{1}{1+\dfrac{(x-a)^2}{b}}/x$$

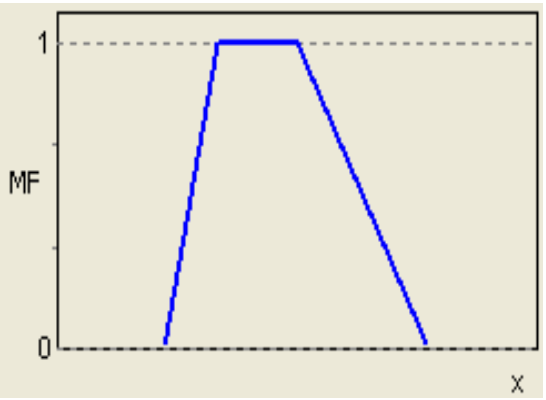
$$\mu(x)=a\sqrt{\frac{x}{x_{max}}\left(1-\frac{x}{x_{max}}\right)}$$

Приклад вигляду функцій приналежності

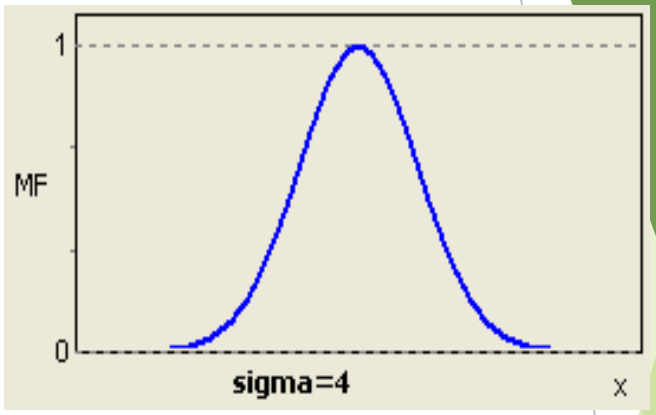
Трикутна



Трапецевидна



Гаусіана



Структура нечіткої системи управління

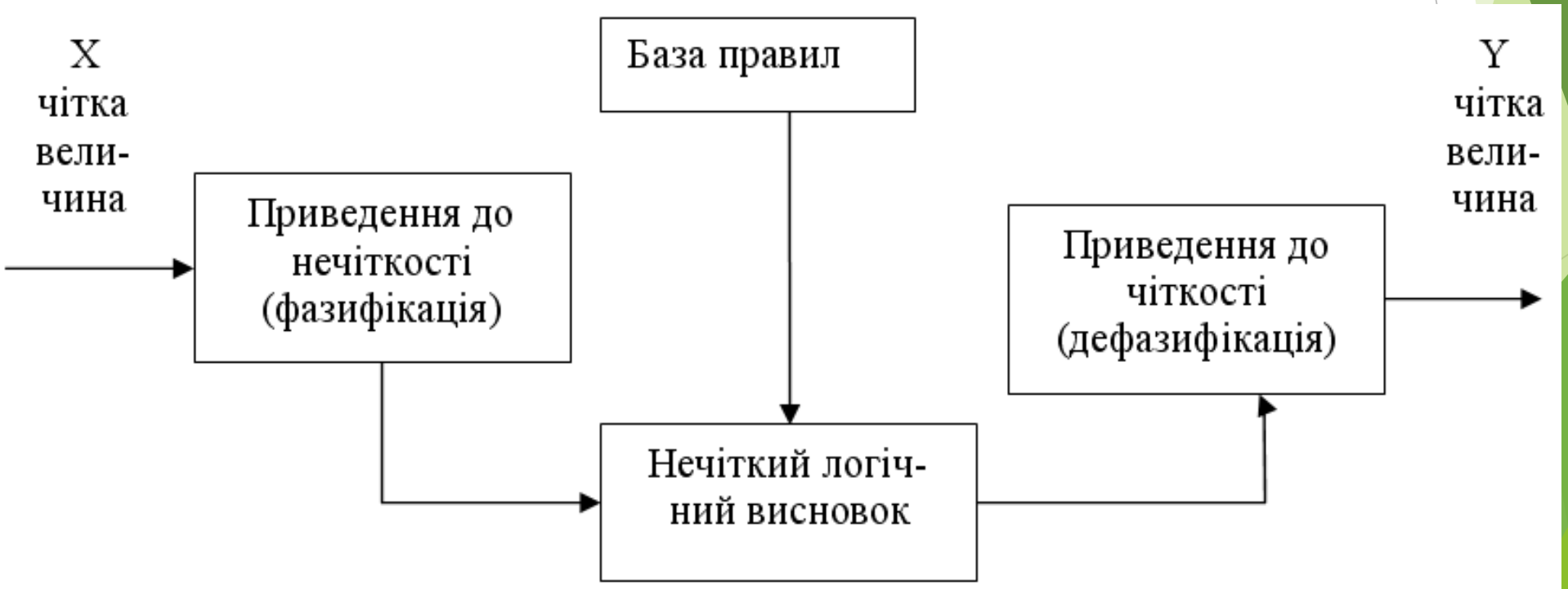
ЯКЩО x_1 це A_1 . І . x_2 це A_2, ТО y це B.

Оптимізація управління соціально-економічної системи,
заданої не

R_1: ЯКЩО x_1 це A_{11} . І . x_n це A_{1n}, ТО y це B_1

R_i: ЯКЩО x_1 це A_{i1} . І . x_n це A_{in}, ТО y це B_i.

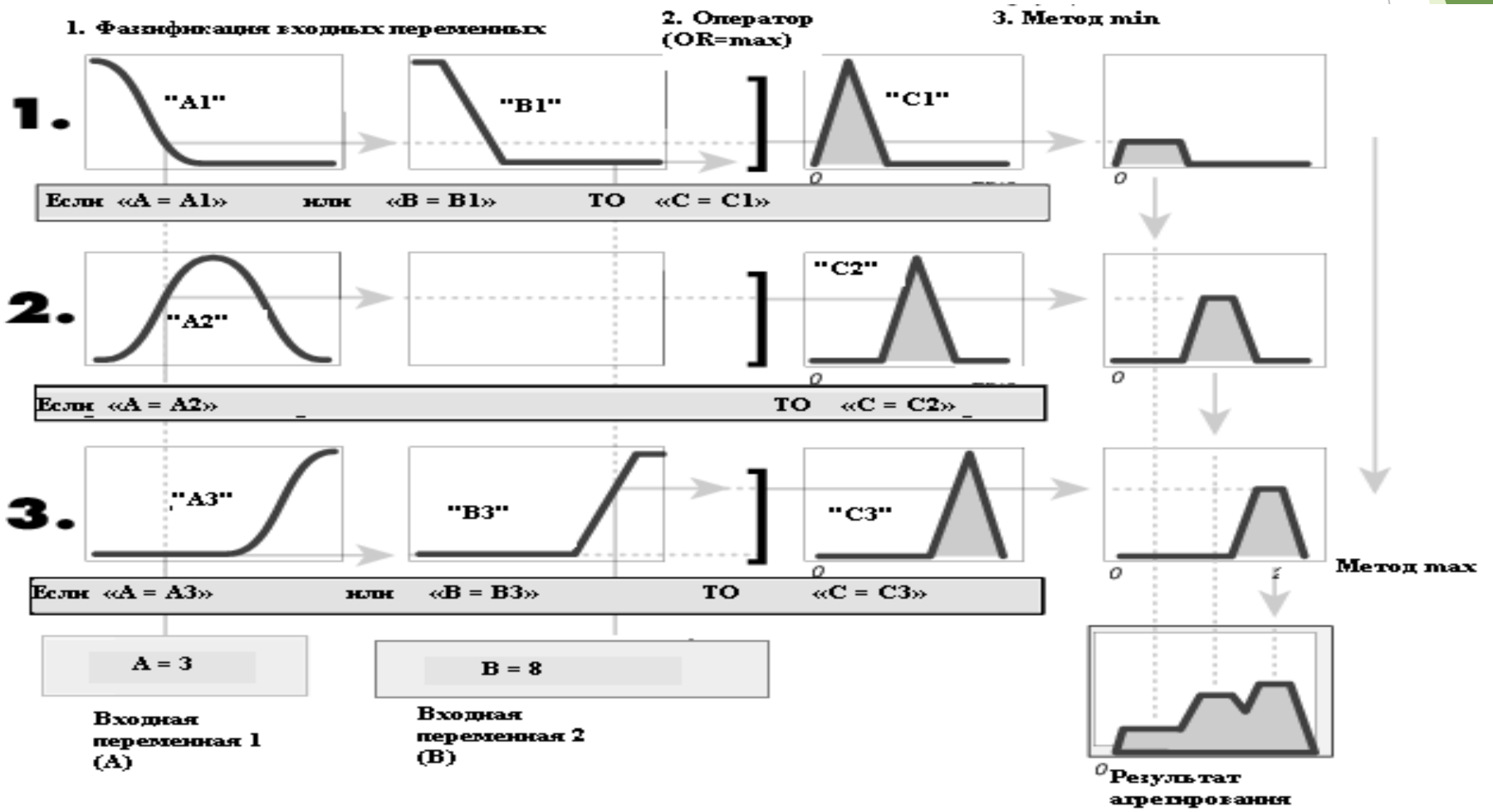
R_m: ЯКЩО x_1 це A_{m1} . І . x_n це A_{mn}, ТО y це B_m.



Нечіткий вивід за Мамдані

$$\alpha_i = \min_i(A_{ik}(x_k)) \quad B_i^*(y) = \min_i(\alpha_i, B_i(y)) \quad MF(y) = \max_i(B_i^*(y))$$

$$y = \frac{1}{y_{\max} - y_{\min}} \int_{y_{\min}}^{y_{\max}} MF(y)dy = \frac{1}{y_{\max} - y_{\min}} \int_{y_{\min}}^{y_{\max}} \max(B_i^*(y))dy$$



Нечіткий вивід за Такаги-Сугено

База знань Сугено аналогічна базі знань Мамдані за винятком висновків правил dj, які задаються не нечіткими термами, а лінійною функцією від входів:

$$d_j = b_{j,0} + \sum_{i=\overline{1,n}} b_{j,i} \cdot x_i$$

Ступені приналежності вхідного вектора

$$X^* = \left(x_1^*, x_2^*, \ldots x_n^*\right)$$

до значень dj розраховуються як

$$\mu_{d_j}\left(X^*\right) = \bigvee_{p=\overline{1,k_j}} w_{jp} \cdot \bigwedge_{i=\overline{1,n}} \left[\mu_{jp}\left(x_i^*\right)\right], \; j = \overline{1,m}$$

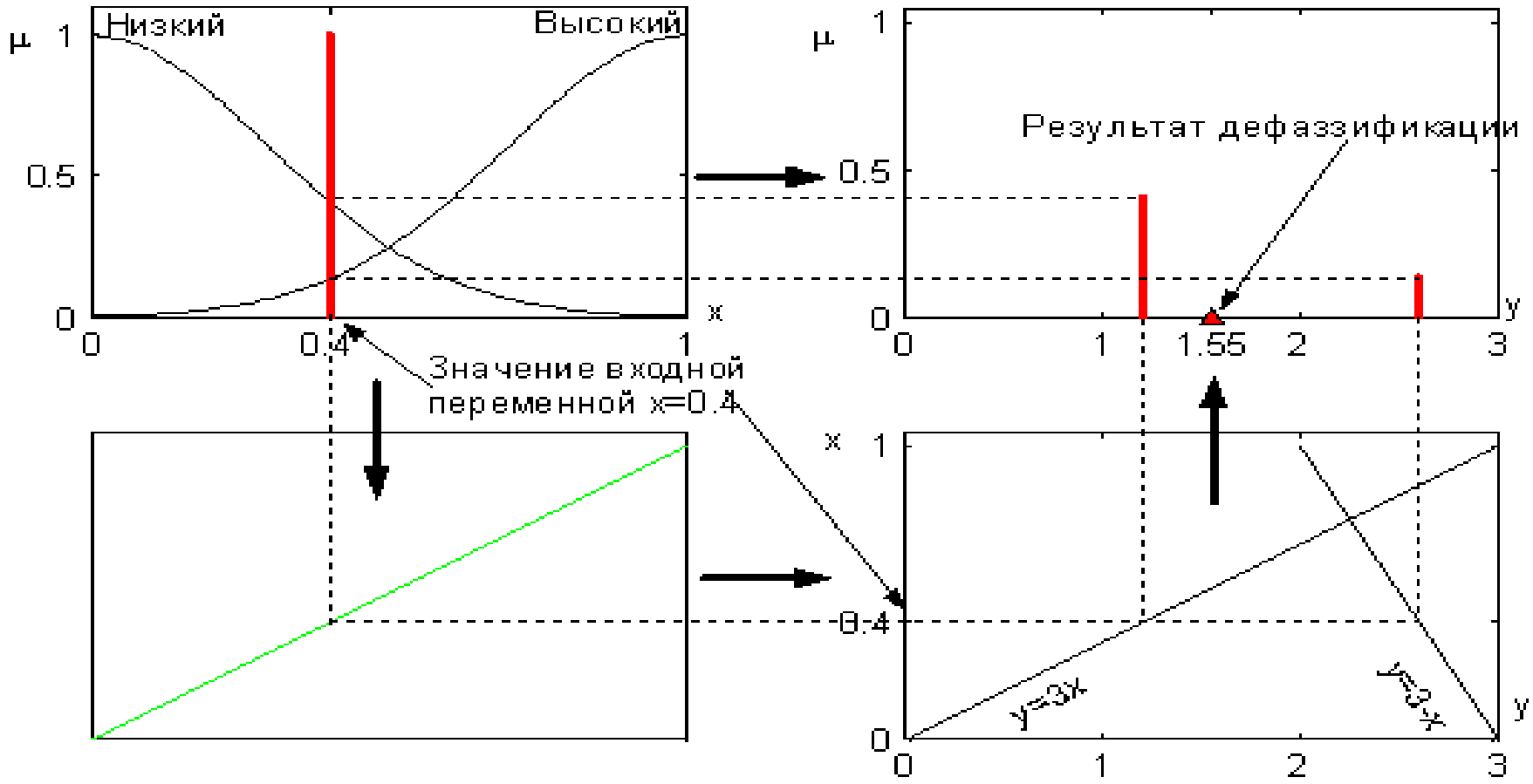
Тут знаки логічного «І» та «АБО» використовуються згідно наперед заданого правила. Тоді отримується набір значень функції приналежності виду

$$\tilde{y} = \frac{\mu_{d_1}(x^*)}{d_1} + \frac{\mu_{d_2}(x^*)}{d_2} + \ldots + \frac{\mu_{d_m}(x^*)}{d_m}$$

А дефазицікація здійснюється знову за правилом середньої звіаженої суми

$$y = \frac{\sum_{j=\overline{1,m}} \mu_{d_j}(X^*) \cdot d_j}{\sum_{j=\overline{1,m}} \mu_{d_j}(X^*)}$$

Приклад дефазифікіції за методом Такаги-Сугено

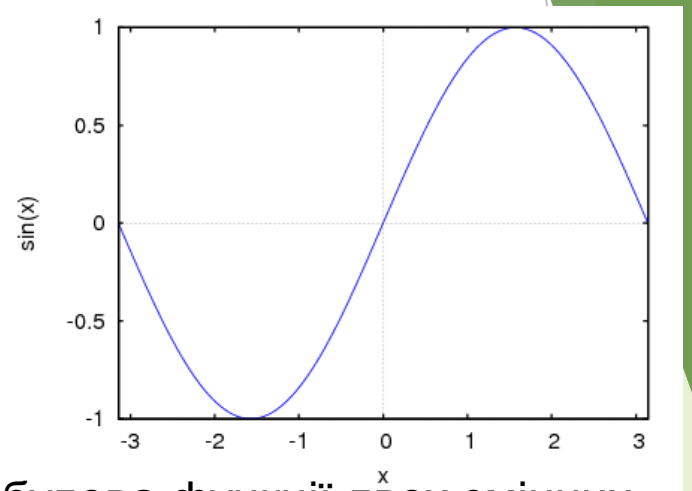


Візуалізація даних:

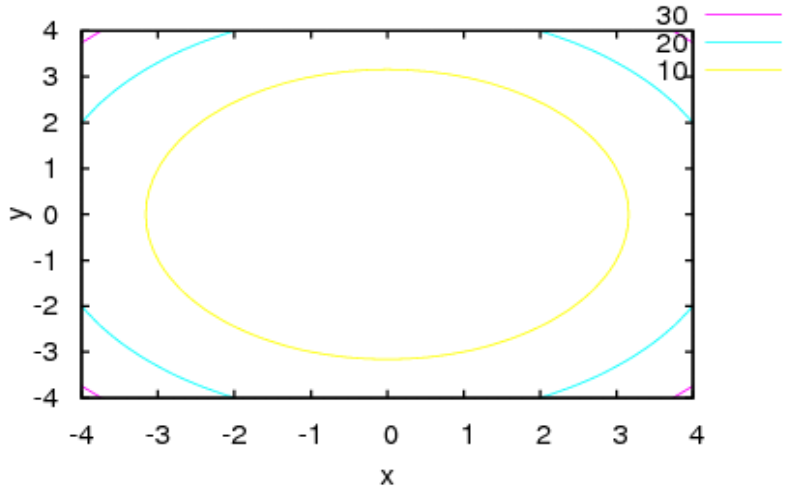
ВИКОРИСТАННЯ КОМП'ЮТЕРА (ПАКЕТ MAXIMA)
ГРАФІКІВ

Побудова функції однієї змінної:
(%i1) plot2d (sin(x), [x, -%pi, %pi])$

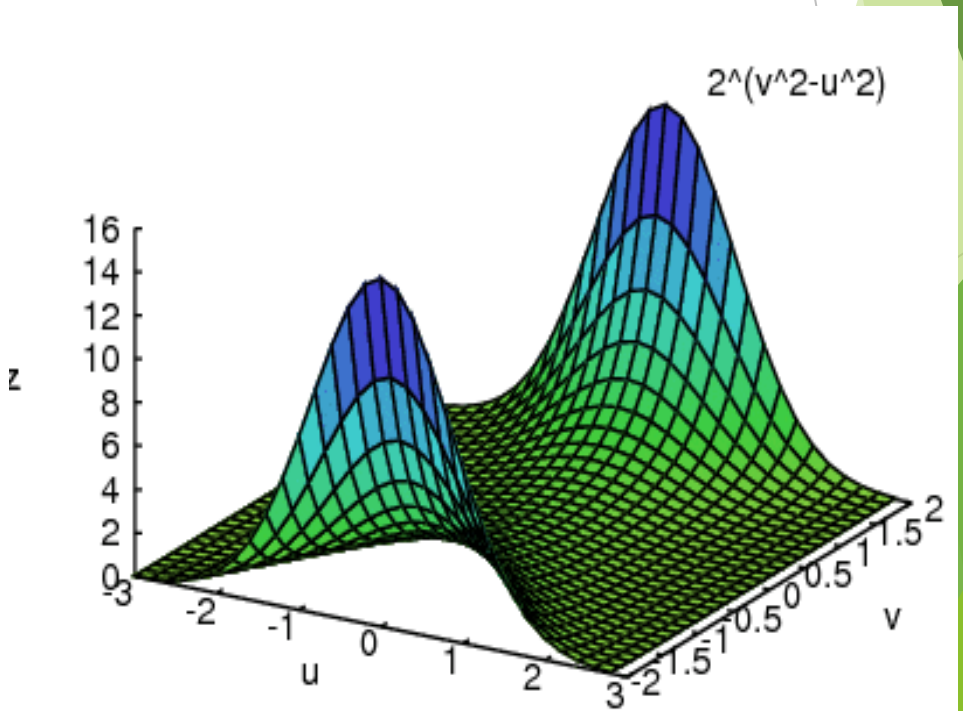
або
(%i1) plot2d (sec(x), [x, -2, 2], [y, -20, 20])$



Побудова функції двох змінних
(%i1) contour_plot (x^2 + y^2, [x, -4, 4], [y, -4, 4])$



Побудова тривимірного зображення
(%i1) plot3d (2^(-u^2 + v^2), [u, -3, 3], [v, -2, 2])$



```
plt.text(1550, 71, 'India')

plt.text(5700, 80, 'China')
```