Big Analytics. Підходи та методи глибокого аналізу (великих) даних.

Олександр С. Балабанов
д.ф.-м.н.,
Інститут програмних систем НАН України, Київ

Big Data Analysis

Olexandr S. Balabanov
Institute of Software Systems of NAS of Ukraine, Kyiv.

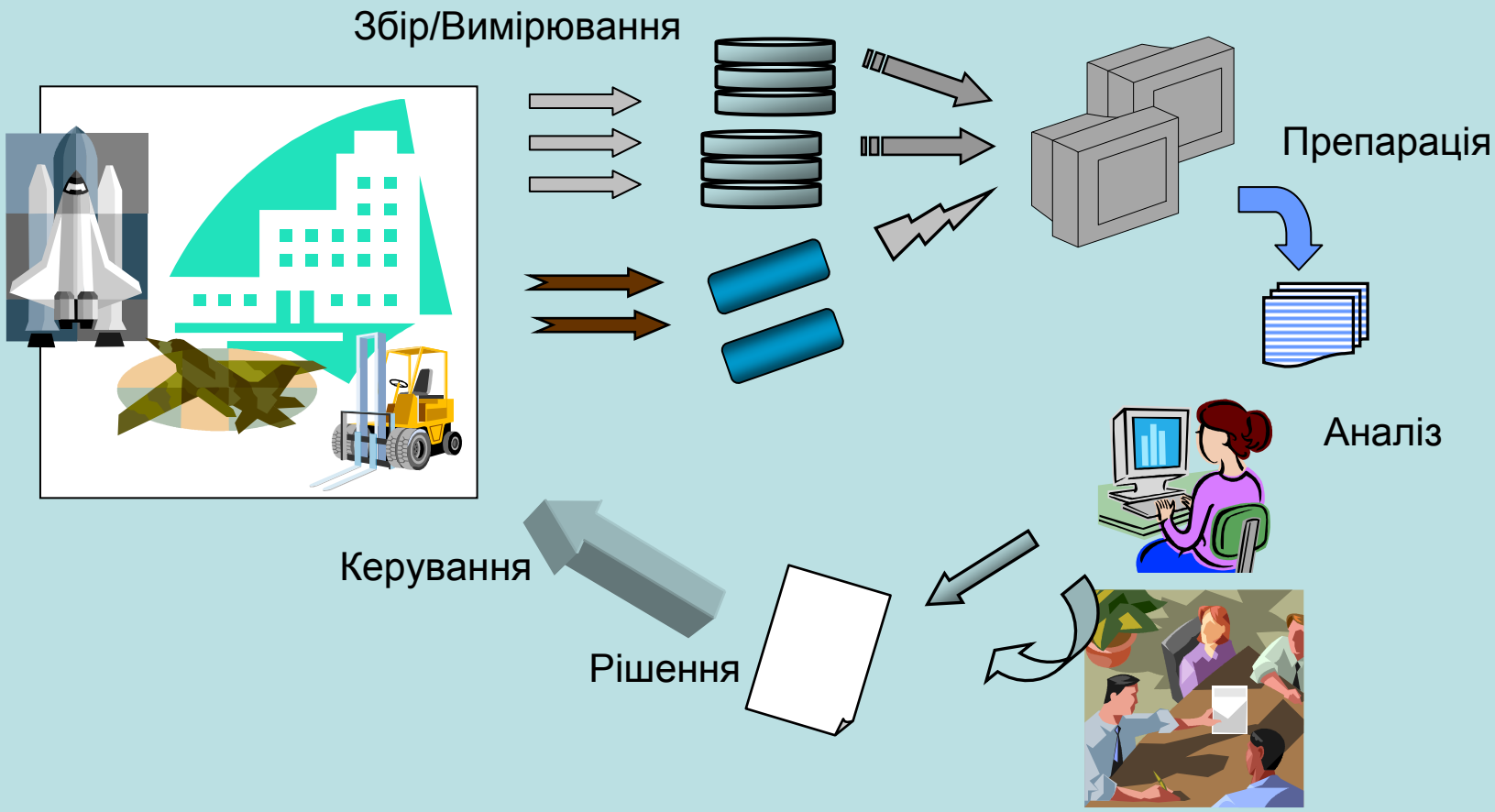
2018

Потреби і потенційні сфери впровадження

Традиційно масиви емпіричних даних вибірково та неповно залучалися до підготовки планів та прогнозування наслідків управлінських рішень, а також аналітичних досліджень. Вибір і обґрунтування рішень робляться на основі експертних суджень і оцінок, адекватність й актуальність яких важко контролювати. Необхідно позбуватися консерватизму й суб'єктивізму в процесах підготовки й обґрунтування важливих рішень.

Вихід на ринок Великих Даних дозволяє кардинально оновити технологію й практику вироблення рішень (і аналітику досліджень).

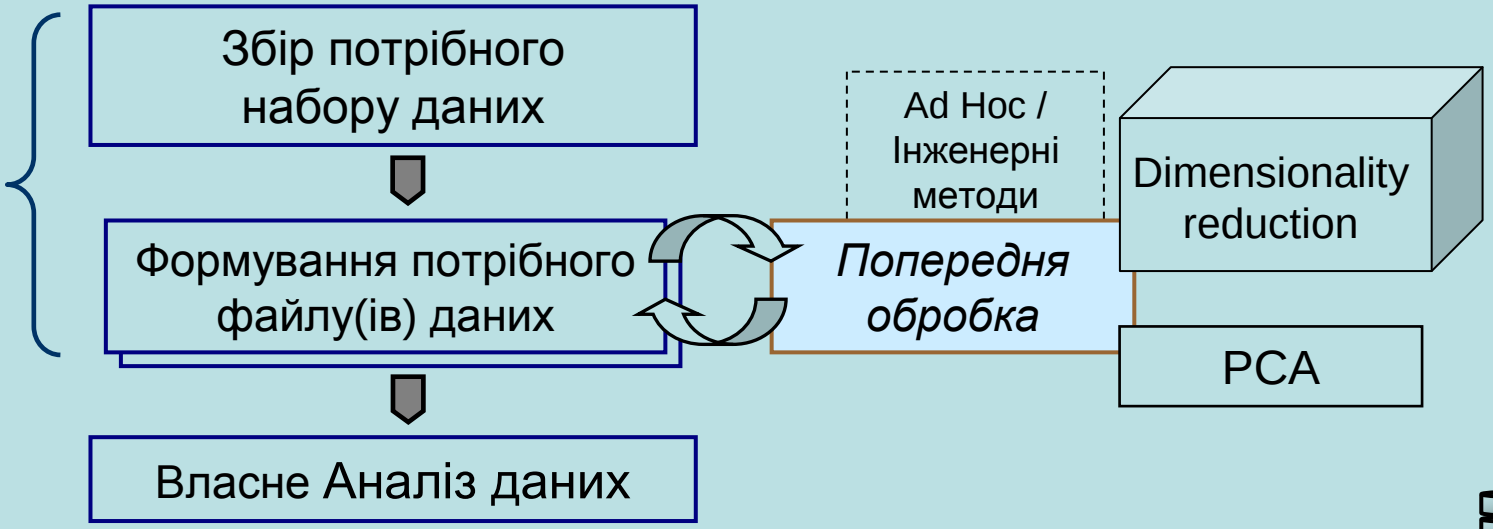
Доступність Великих Даних дозволяє отримати широкий спектр інформації про об'єкт та середовище
можливість побудувати замкнений комп'ютеризований цикл керування (з переважаючою роллю комп'ютерних технологій)

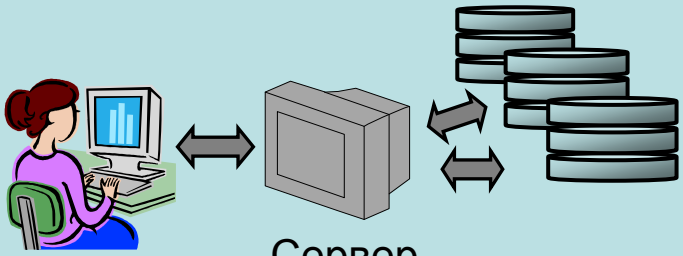


Оскільки збираються переважно “сирі”, різнорідні, неузгоджені та невпорядковані дані,
то для отримання з них корисного “сенсу” необхідні два етапи:

1) компіляція та інтеграція даних (добір, фільтрація, зменшення розмірності, агрегація, комплектування, синтез, синхронізація, переформатування);

2) глибокий аналіз підготовлених даних (Велика аналітика).



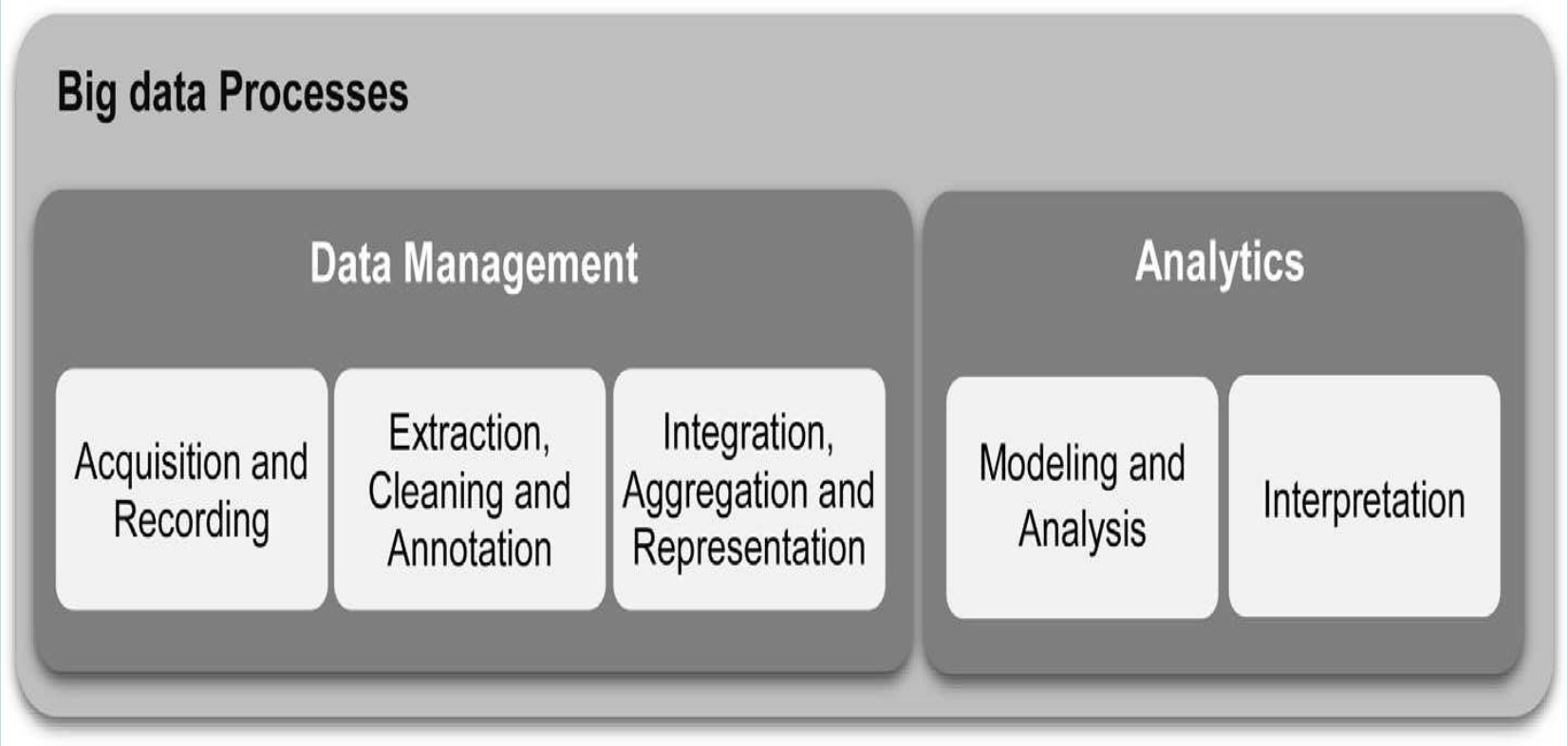


Аналіз з “поточними” запитами (in-memory operations?)

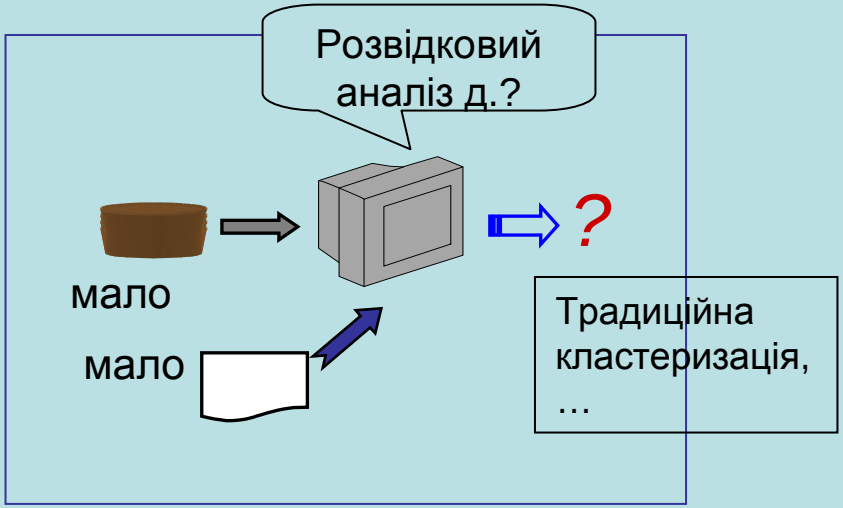
Сервер

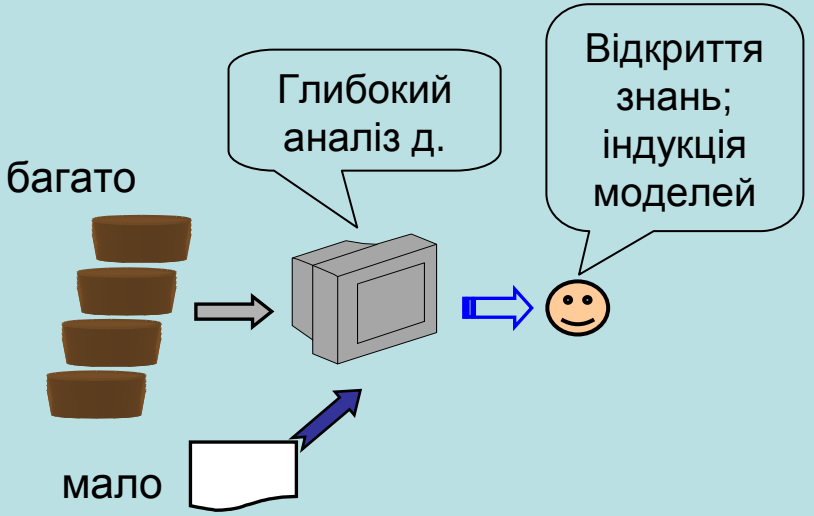
Схема з огляду:

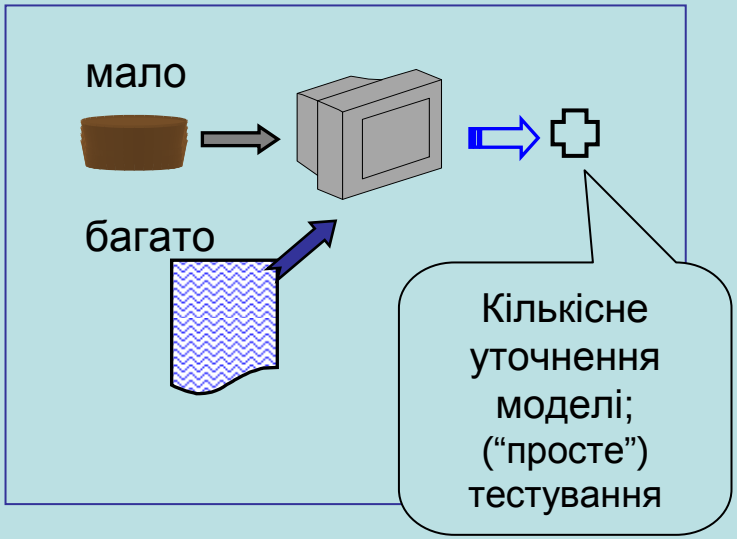
[Gandomi A., Haider M. (2015). Beyond the hype: Big data concepts, methods, and analytics / International Journal of Information Management.]

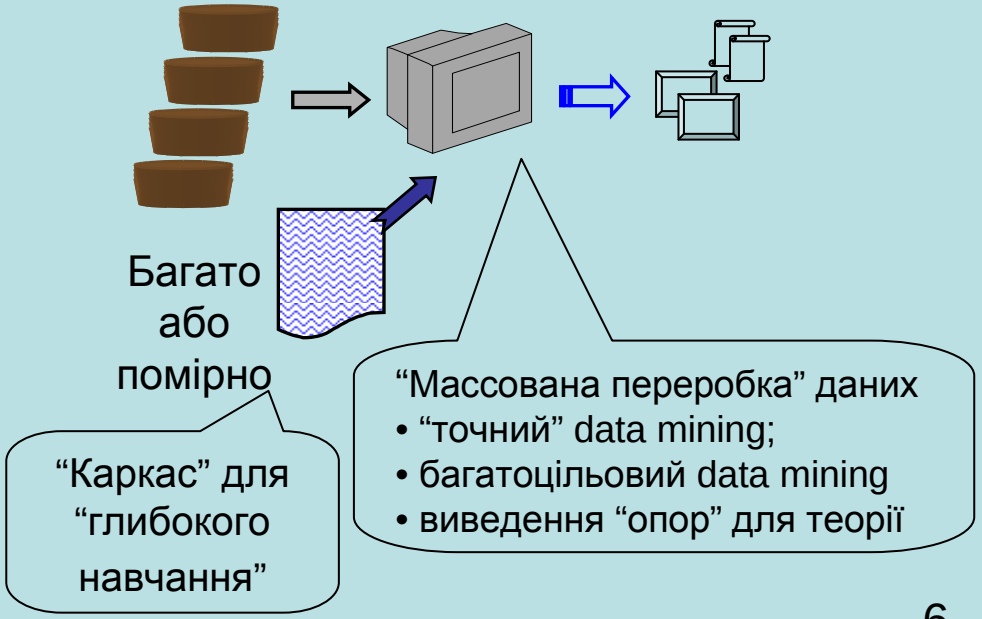


Аналіз даних ; проблемні (когнітивні) ситуації

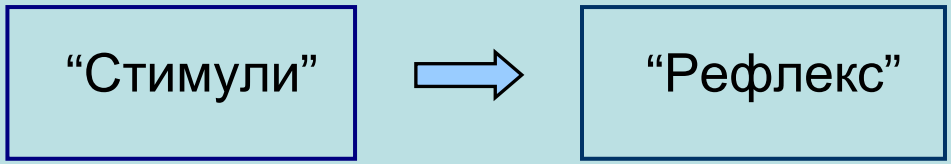








(Ідеальна) повністю замкнена комп'ютерна технологія (схема)



Дані

На виході – кінцевий результат для практики

Така схема (“чорна скриня”) працює для спеціальних задач типу розпізнавання. (де ціль вказана і відношення (ролі) в принципі відомі).

Але

• для задач керування складним об'єктом

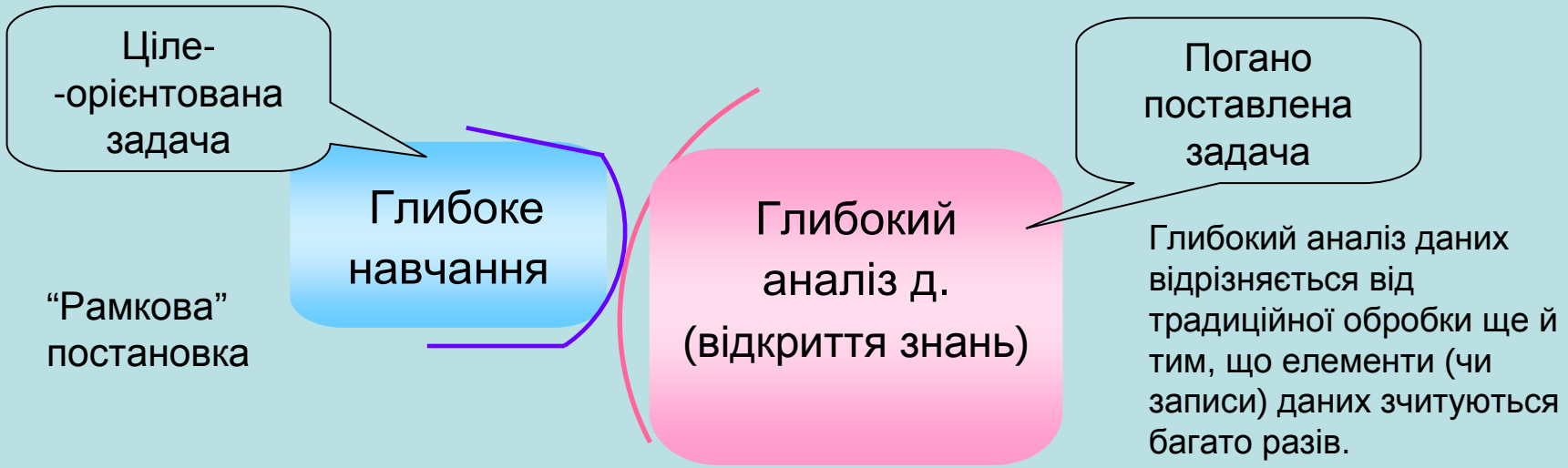
• для дослідницьких задач, коли апріорі майже нічого не відомо..

така схема – нереалістична й недоцільна.

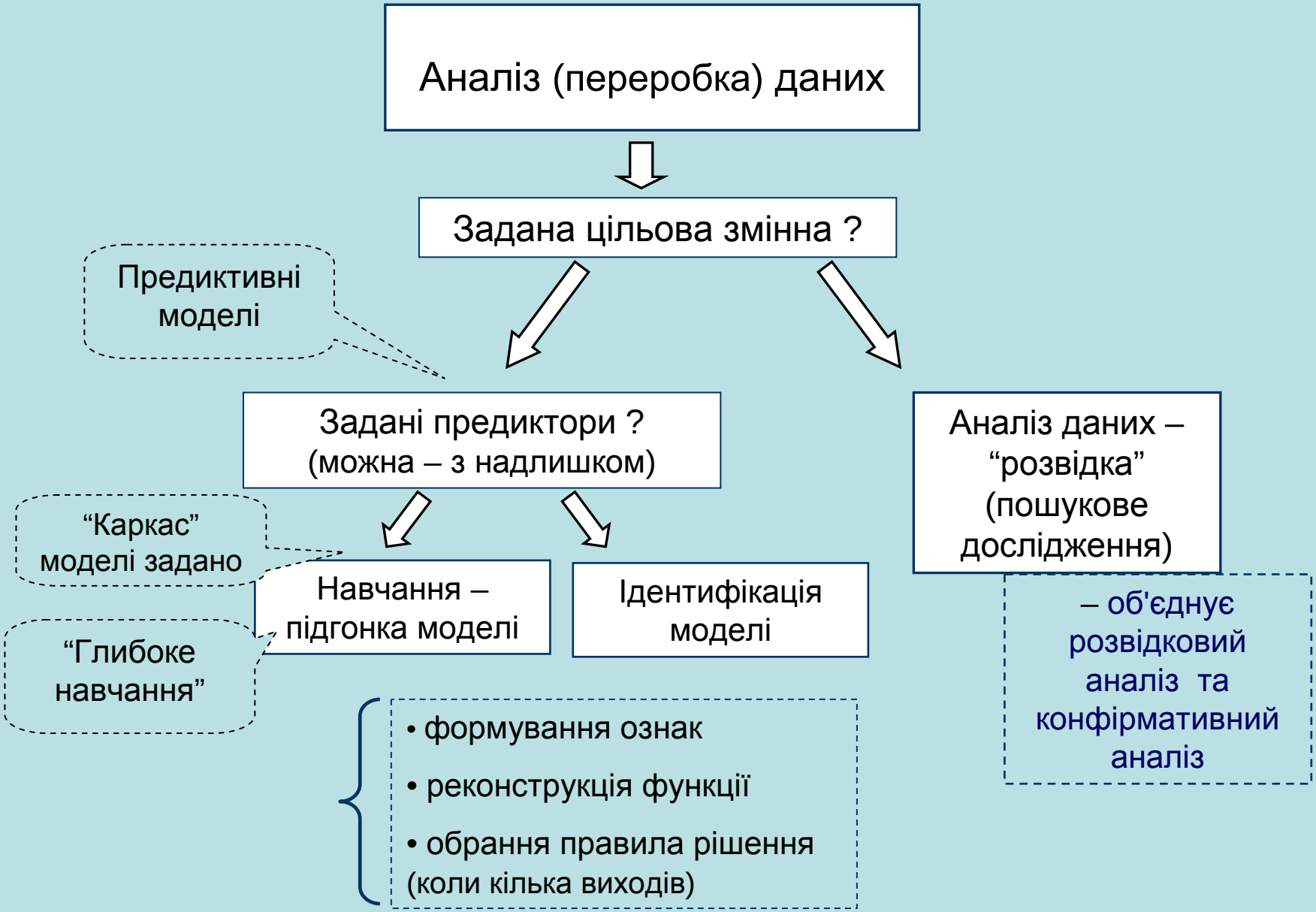
Реалістично – на основі даних спостережень вивести модель, яка відтворює «портрет» об’єкту, відтворює систему зв’язків та впливів між характеристиками (показує, «як воно розгортається»)

⎩ відтворити генеративну модель.

Мета – ідентифікувати систему зв’язків та впливів між характеристиками об’єкту у середовищі (бажано – не використовуючи обмежуючих апріорних знань);
Відтворити «портрет» об’єкту у середовищі, що надасть “інсайт” аналітику й замовнику



Спочатку вивести модель. А прогноз кінцевого ефекту або цільового показника (наприклад, прибутку) можна оцінити як функцію змінних моделі, з можливістю врахувати ризики, невизначеність та фактори, що залишилася поза вхідними даними.


Аналіз (переробка) даних
Задана цільова змінна ?
Предиктивні моделі
Задані предиктори ?
(можна – з надлишком)
Аналіз даних – “розвідка” (пошукове дослідження)
“Каркас” моделі задано
Навчання – підгонка моделі
Ідентифікація моделі
“Глибоке навчання”
– об'єднує розвідковий аналіз та конфірматичний аналіз
• формування ознак
• реконструкція функції
• обрання правила рішення
(коли кілька виходів)

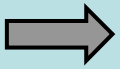
“Рамкова” постановка –

• Задано “каркас” моделі:

• Задана одна цільова характеристика

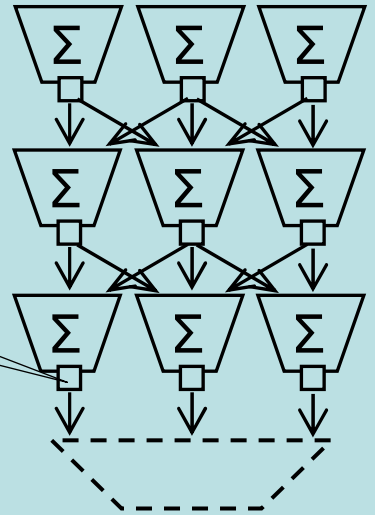
• Всі інші задані на вході характеристики є кандидатами у предиктори (фактори, аргументи) або їх компоненти (елементи).

• Задано форми перетворення (параметричні родини моделей) та арсенал “цеглин” (форм, будівельних блоків), з яких можна конструювати модель (функцію, процедуру)



“Глибоке навчання”

Нейронні мережі та “Глибоке навчання”



Сігмоїдні ф-ції від лінійних комбінацій входів

Порогова функція

формування ознак

(формування ознак 2-го порядку)

реконструкція функції

(Вибір) правило рішення (коли кілька виходів)

Нелінійна ф-ція лінійної функції ознак

$$z_i = \sigma(\alpha_{i,0} + \alpha_{i,1} \cdot x_1 + \ldots)$$

$$y_k = g_k(\beta_{k,0} + \beta_{k,1} \cdot z_1 + \ldots)$$

Зазвичай g_k подібна до σ, але має 3 вільні параметри

“Навчання” – підбір параметрів з метою мінімізувати с.-кв. відхилення прогнозу від фактичних значень y_i

Навчання – підгонка

Аналіз – дослідження

- ♣ Тренування процедур і моделей (розпізнавання, класифікації)
- ♣ Глибокий аналіз (емпіричних) даних.

Типи задач та результатів:

- Групування випадків (записів, об'єктів..)
(кластеризація (– *значущі кластери*),)
- Виявлення регулярних патернів (типових повторювань)

  а) структурних

  послідовних (motifs), 3-вимірних, графових, ...

  б) наборів (itemsets, market baskets, асоціацій)..
- Виявлення послідовностей у часі (лінки,.. значущі зв'язки дій)
- Виявлення трендів та “періодичності” (в даних із темпоральною прив'язкою)
- Виявлення структур залежностей
- Відтворення каузальної моделі

виведення формул

“Глибоке навчання”

Глибина – ієрархічність, багаторівневість моделі, складність формул.

Адекватність такої складної моделі можлива завдяки тому, що задано “каркас” моделі і що модель високоспеціалізована.

Глибокий аналіз (емпіричних) даних

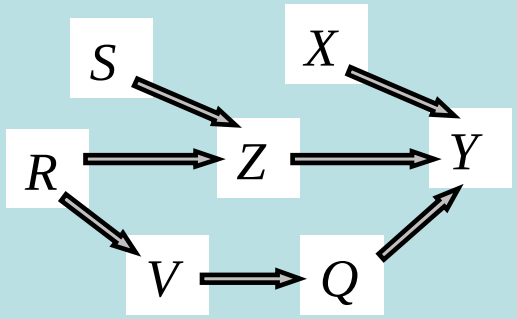
Мета – ідентифікувати систему зв’язків та впливів між характеристиками об’єкту у середовищі (з мінімальним використанням обмежень та апріорних знань);
Відтворити «портрет» об’єкту у середовищі, що надасть “інсайт” аналітику й замовнику.
Виявити реальні фактори й причини.
Надати інструмент рішень (керування).
(“Прескриптивна” аналітика)

Глибина – підйом від “сирих” даних до змістовної “картини” (знань); пошуково-дослідницький характер аналізу.

Потрібна модель, яка допомагає зрозуміти зв'язки та взаємозалежності у реальному середовищі, де функціонує об'єкту інтересу.
Бажано, аби виведена модель була придатна для прогнозу наслідків виконання рішень (дій) користувача (менеджера), навіть якщо дані було зібрано як пасивні спостереження.

Такі властивості мають каузальні мережі.

Прогноз наслідків втручання (керування) виконується на каузальній мережі процедурами виведення умовного розподілення ймовірностей цільового показника, після застосування правил do-calculus, відповідно до планового втручання.



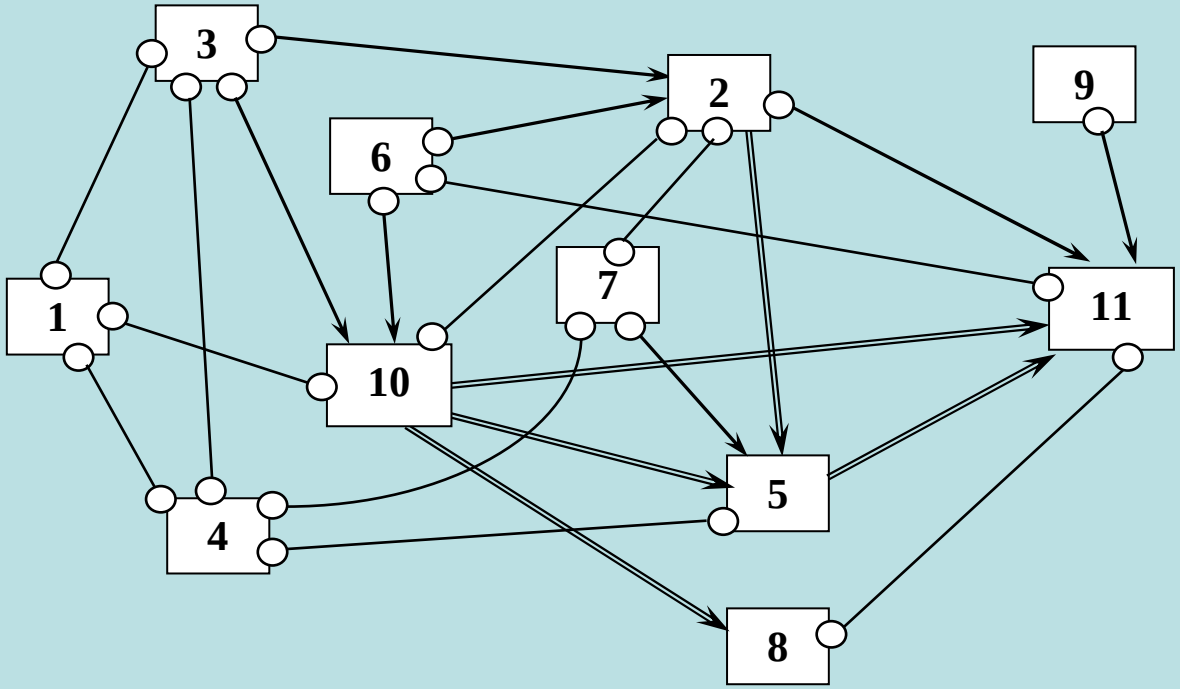
$$z = f(r,s) + \varepsilon_Z$$

$$y = g(x,z,q) + \varepsilon_Y$$

“Глибокий” аналіз – тому що на виході нові знання.

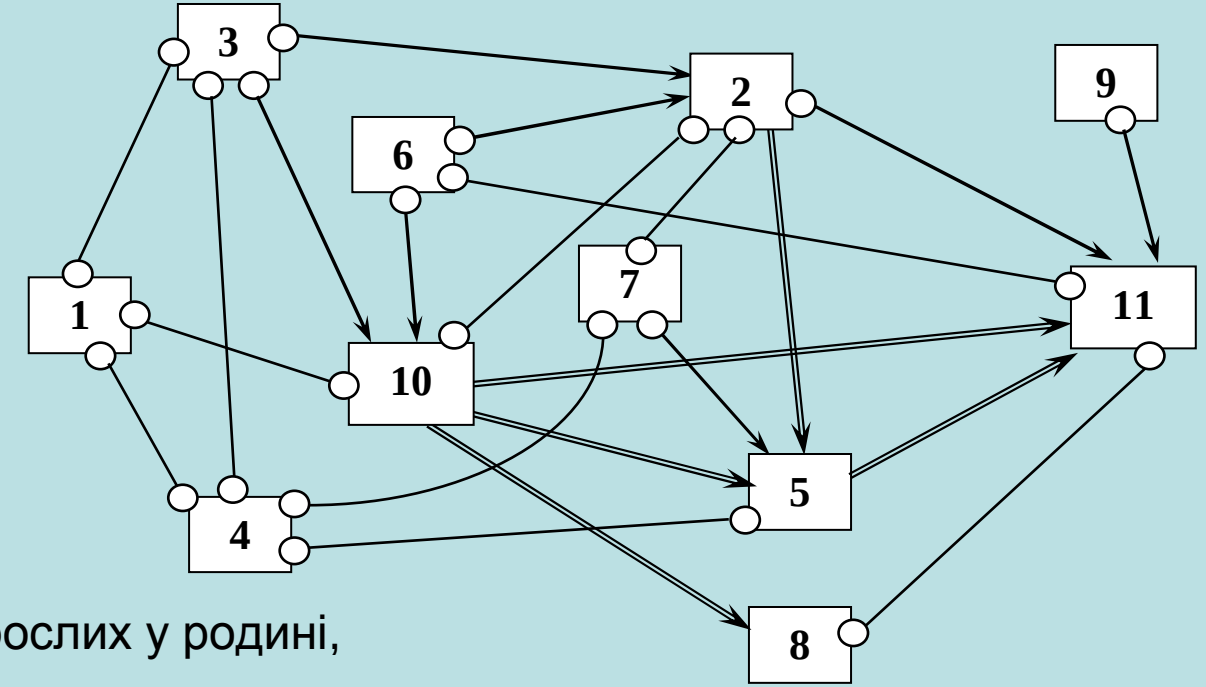
Приклад аналізу реальних даних. Дані охоплюють фактори, пов'язані з віком матері при народженні першої дитини в США.

Результат виведення структури моделі нашим алгоритмом Razor-1.3.
(*Метод, оснований на незалежності.*)



1) професія батьків;
2) раса;
3) відсутність братів (сестер);
4) матір жила на фермі;
5) регіон США; 6) наявність двох дорослих у родині, де росла матір;
7) релігія; 8) паління сигарет; 9) був чи ні викидень;
10) освітній рівень матері (на час виходу заміж);
11) вік матері при народженні першої дитини.

1) професія батьків;
2) раса;
3) відсутність братів (сестер);
4) матір жила на фермі;
5) регіон США;



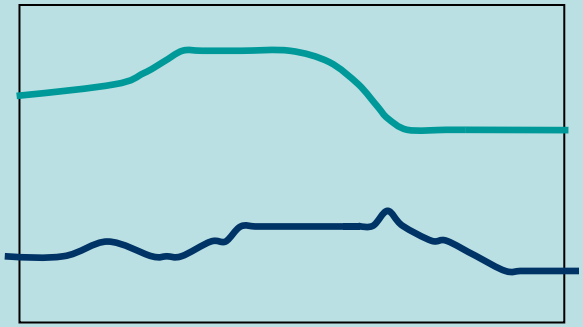
6) наявність двох дорослих у родині, де росла матір;
7) релігія; 8) паління сигарет; 9) був чи ні викидень;
10) освітній рівень матері (на час виходу заміж);
11) вік матері при народженні першої дитини.

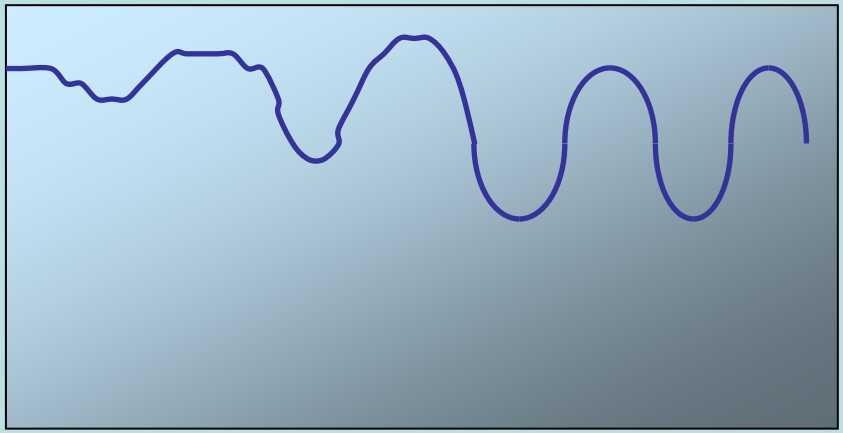
Виявлено п’ять каузальних зв’язків. На вік матері при народженні першої дитини впливають освітній рівень матері та регіон проживання (‘10→11’ та ‘5→11’).

Неможливо обчислити безпосередню кореляцію між ‘1’ та ‘3’ , тому що невідомо, чи залежні вони через ‘4’.

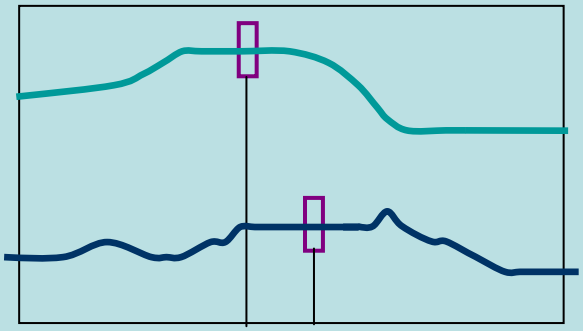
Головна підстава сумніватися у точності цих результатів – припущення про лінійність залежностей (тим паче, що маємо дискретні змінні).

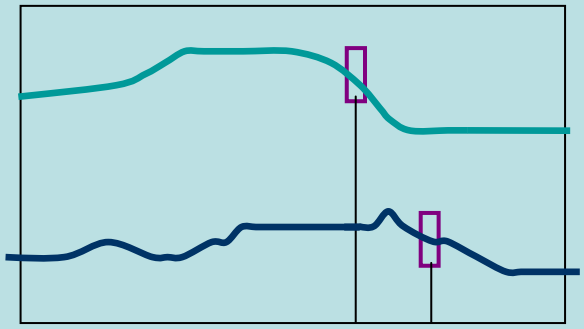
Дані типу “процеси (поведінка)”

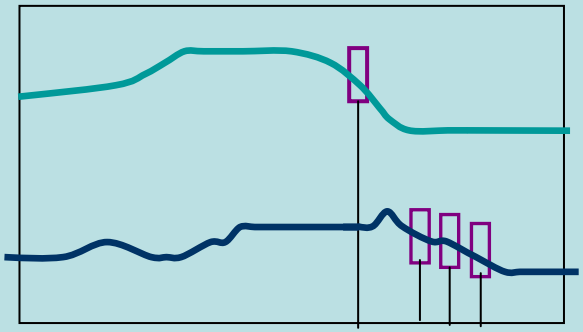




Вимірювання даних. Схема вимірювання

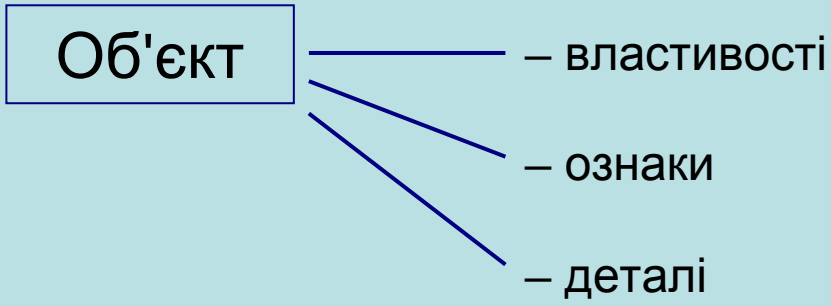






Дані за схемою “об'єкт – ознаки (властивості)”

Зв'язки (відношення) :



(Зв'язок “частина – ціле”)

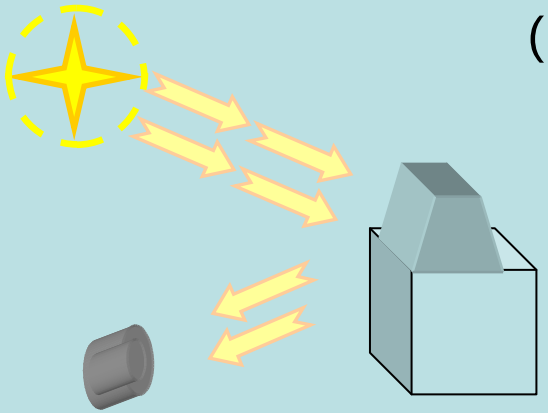
Такі відношення відображаються неорієнтованими статистичними зв'язками.

(Взагалі) Типи Зв'язків: суміжність (близкість), приналежність, атрибут (характеристика), залежність (асоціація), слідування у часі, вплив, …

Але навіть у задачах, де розглядаються статичні відношення між характеристиками частин та характеристиками цілого (об'єкту), може виникати модель типу процес та каузальність.

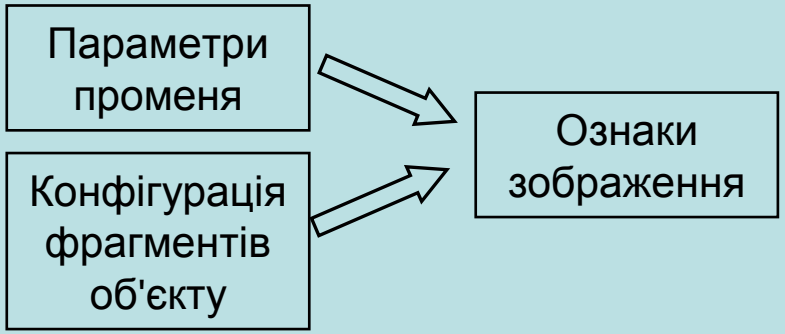
Зазвичай дані несуть не безпосередньо самі властивості об'єкту, а результат його сприйняття через певний механізм.

Замість атрибуту – його репрезентант.



(*Технічний зір*)

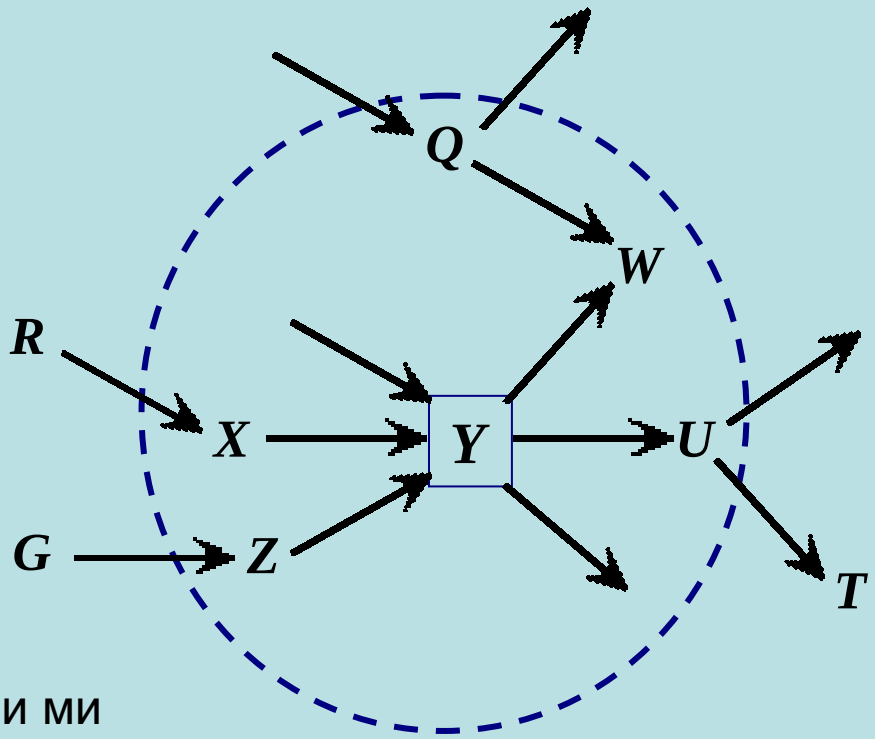
Ознака – світла пляма, яка відображає грань. Але освітленість і форма плями можуть змінюватись як результат взаємодії променя з фактурою грані й прилеглих ділянок.



Інформативність змінних
(регресорів, предикторів, ознак, атрибутів) –
– через каузальну структуру

Пасивна предикції (класифікації,…)
– оцінити Y , знаючи пов'язані з нею характеристики.

Наприклад, якщо задано значення $X\ G\ Q\ W\ T$, то всі вони корисні для цього.



Каузальний прогноз – оцінити ***Y*** , коли ми будемо маніпулювати змінними *X G Q W T* , на об'єкті . Керування змінними *Q W T* **не** дасть ефекту.

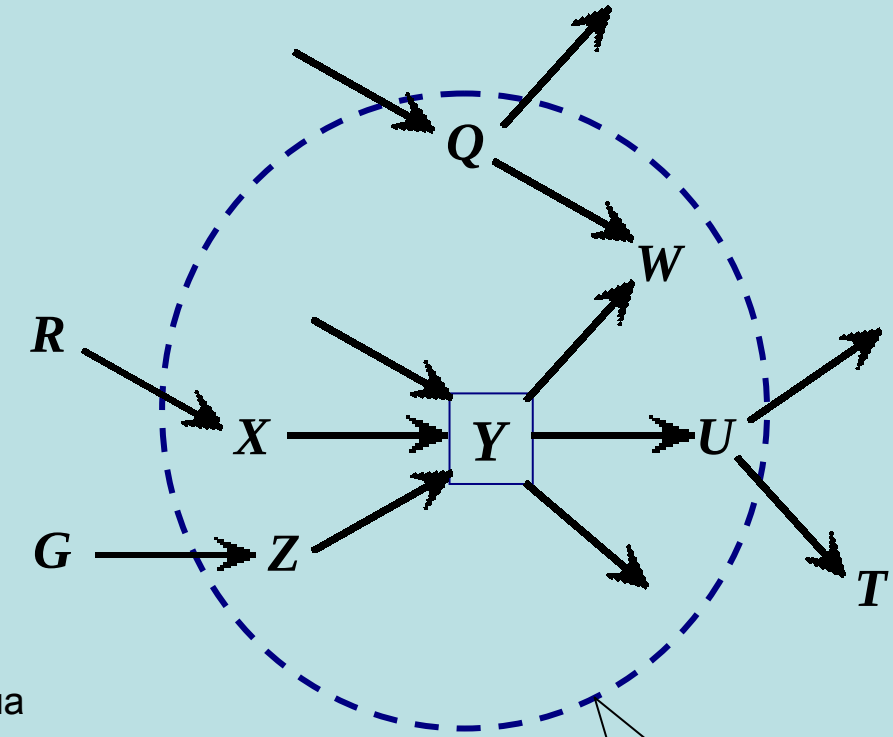
Інформативність змінних

Принципова відмінність **впливу** і **асоціації**

Для пасивної предикції (класифікації, ...) безумовно інформативні всі суміжні змінні (безпосередні причини й наслідки).

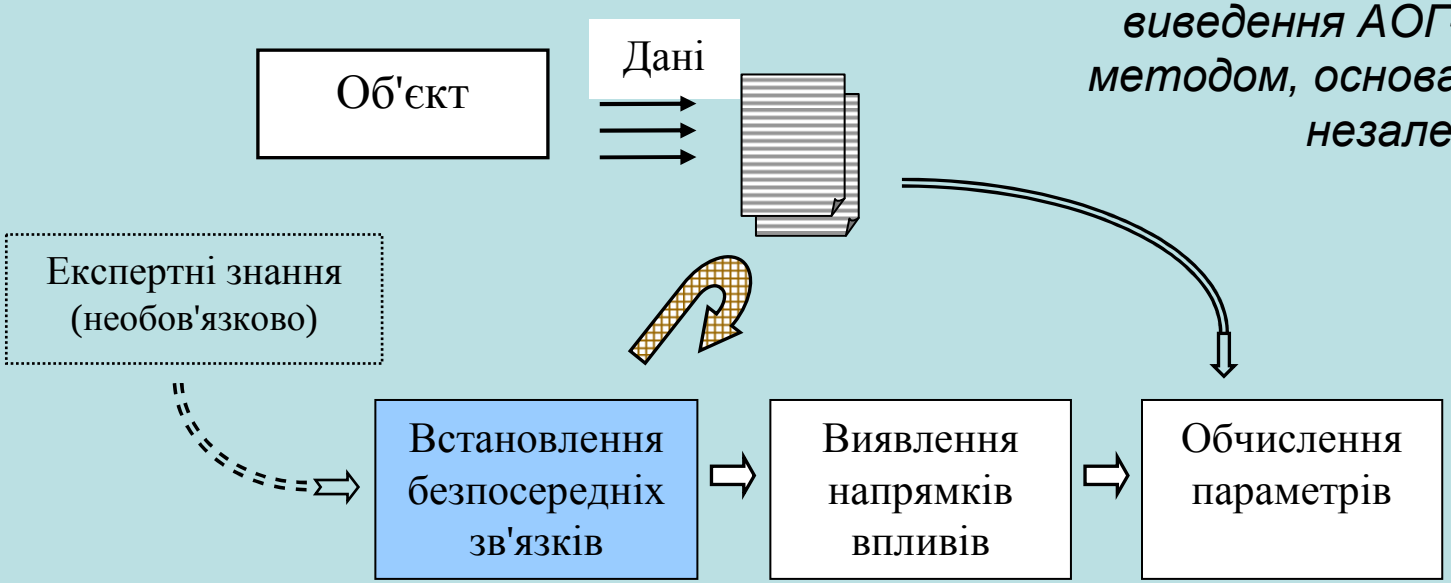
Всі інші змінні можуть стати інформативними за умови присутності/відсутності інших змінних.

Якщо задано всі змінні марківського бланкету, решта змінних – неінформативна

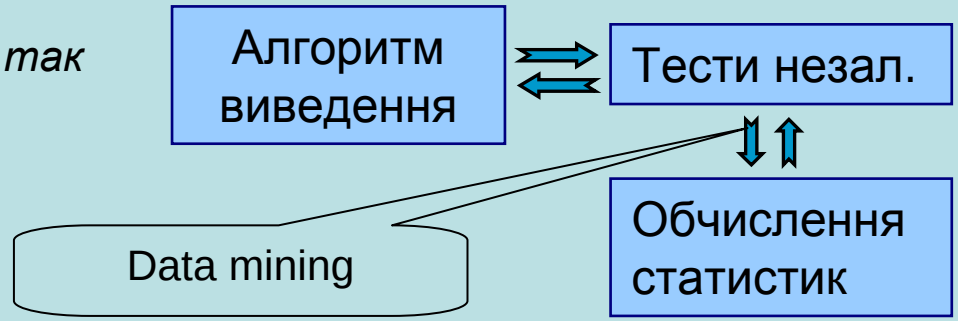


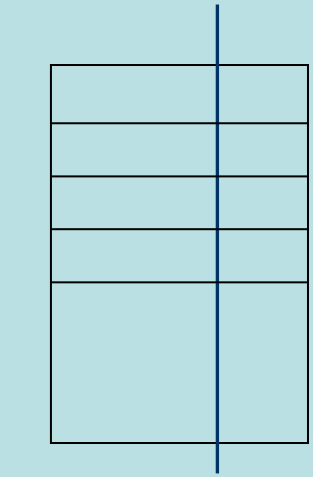
Для *каузального прогнозу* інформативні змінні розділяються на два типи й враховуються по-різному.

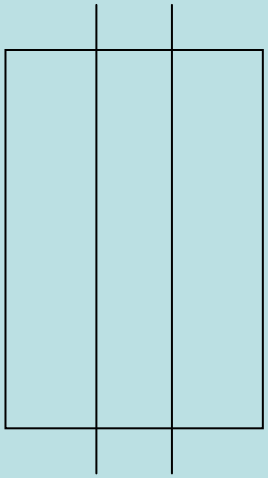
(Змістовна) блок-схема виведення АОГ-моделі методом, основаним на незалежності



З обчислювальної точки зору відтворення моделі виглядає так







Фундаментальне вимога більшості статистичних методів аналізу (виведення моделі) – дані є статистичною вибіркою записів-випадків (прецедентів, екземплярів, транзакцій, актів) популяції однорідних об'єктів.

Кожний запис – набір значень “атрибутів” з єдиної номенклатури , що характеризує конкретний випадок “реалізації” акту.

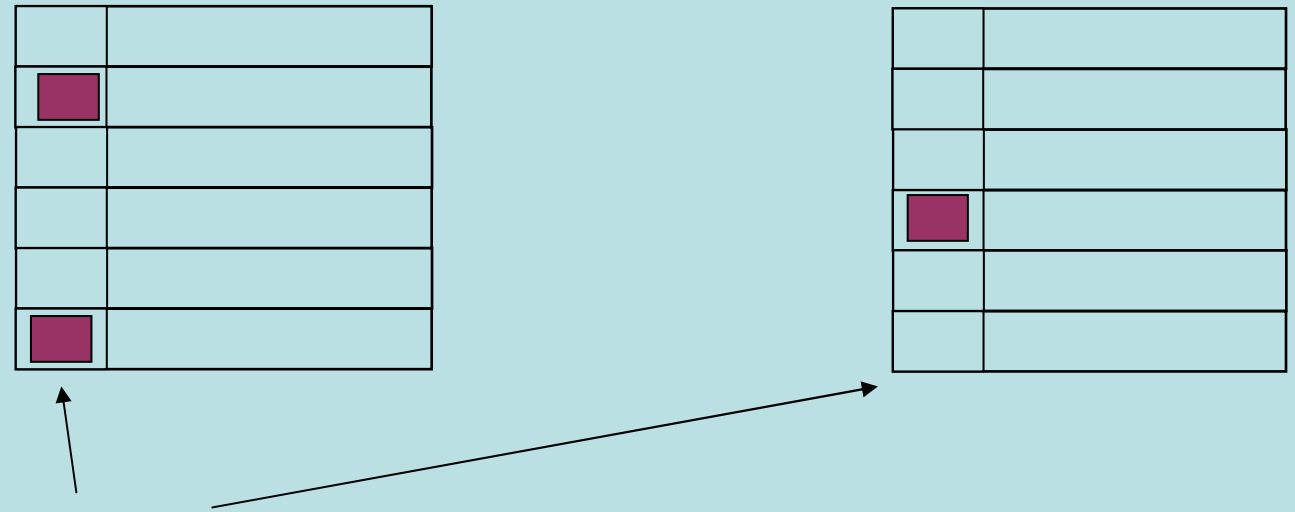
Якщо розрізати таблицю вертикально (роз'єднати набори стовпчиків), переставити рядки і забути номери рядків, то багатовимірний аналіз стане неможливим.

Хоча є спец. методи непрямого відтворення залежностей між “рознесеними атрибутами”, коли збережено достатній набір спільних атрибутів.

На поточний момент складається враження, що для глибокого аналізу доступних нам даних головною проблемою стане

не добір, фільтрація, зменшення розмірності,

а агрегація, комплектування, синтез
(тобто формування “випадків”)



IP-address

Якщо замість конкретних випадків аналізувати “середні” характеристики, то зв'язки втрачаються

References

Statistical inference, learning and models in Big Data / B. Franke, J.-F. Plante, R. Roscher, et.al. (2016). – Intern. Statistical Review, 84(3): P.371–389.

H.J. Watson (2014). Tutorial: Big Data Analytics: Concepts, Technologies, and Applications // Communications of the Association for Information Systems. Volume 34 , Article 65, P.1247–1268.

Gandomi A., Haider M. (2015). Beyond the hype: Big data concepts, methods, and analytics / International Journal of Information Management, 35(2), P. 137–144.

Critical analysis of Big Data challenges and analytical methods / U. Sivarajah, M.M. Kamal, Z. Irani, V. Weerakkody (2017) / Journal of Business Research. V.70, P. 263–286.

Hastie T., R. Tibshirani, J. Friedman (2009). The Elements of Statistical Learning. – 2nd ed., Springer. –745p.

References (2)

Cukier K. (2010). Data, data everywhere: A special report on managing information / The Economist, 2010, February 25.

Jiang, H., Chen, Y., Qiao, Z., Weng, T. H., Li, K. C. (2015). Scaling up MapReduce-based big data processing on multi-GPU systems. / Cluster Computing, 18(1), 369–383.

E. Bareinboim and J. Pearl (2016): Causal inference and the data-fusion problem / Proc of Nat. Acad. Sciences of USA, 113(27): 7345–7352.

Андон Ф.И., Балабанов А.С. Выявление знаний и изыскания в базах данных: подходы, модели, методы и системы (обзор) // Проблемы программирования. – 2000. – № 1–2. – С. 513–526.

Балабанов А. С. Выделение знаний из баз данных – передовые компьютерные технологии интеллектуального анализа данных / А.С. Балабанов // Математичні машини і системи. – 2001. – № 1–2. – С. 40–54.

References (3)

О.С. Балабанов. Відкриття знань в даних та каузальні моделі в аналітичних інформаційних технологіях / Проблеми програмування. – 2017. № 3: С.76–92.

The anatomy of big data computing / R. Kune, P. K. Konugurthi, A. Agarwal, R.R. Chillarige, and R. Buyya / Software: Practice and Experience, 2016. – V.46: P. 79–105.

Labrinidis A., Jagadish H. V. (2012). Challenges and opportunities with big data / Proceedings of the VLDB Endowment, 5(12), P. 2032–2033.

Peter Bühlmann, Sara van de Geer (2018). Statistics for big data: A perspective / Statistics and Probability Letters. (in press).

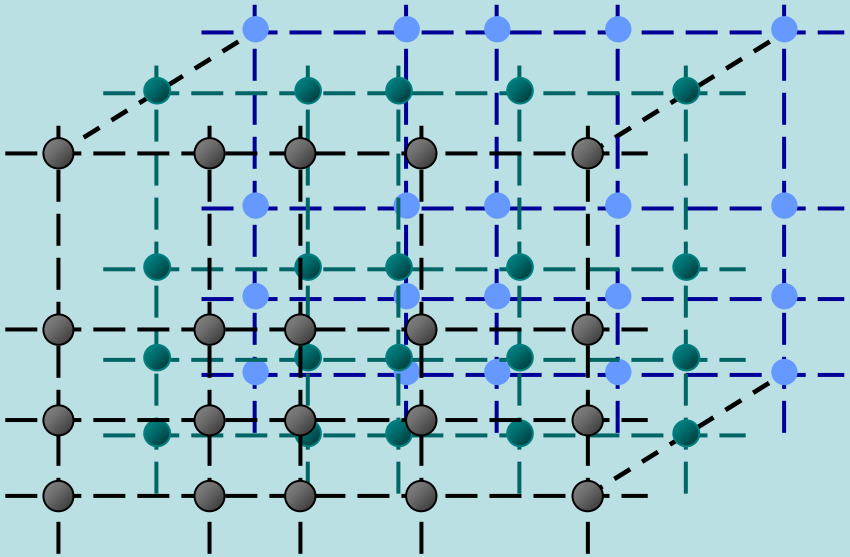
Y. LeCun, Y. Bengio, G. Hinton (2015). Deep learning / Nature, V. 521, P.436-444.

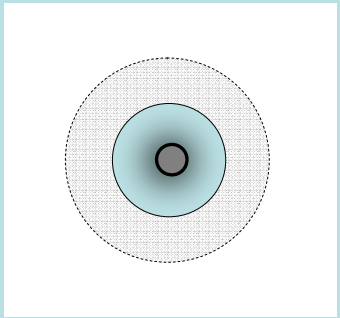
J. Pearl, and E. Bareinboim (2014). External Validity: From Do-Calculus to Transportability Across Populations / Statistical Science, Vol. 29, N. 4, P. 579–595.

Дякую за увагу !

Тестування умовної незалежності за способом хі-квадрат через «згладжену дискретизацію»

Статистика для тесту обчислюється в кожному вузлі «решітки».
Решітка утворюється на основі «опорних» (маркерних) значень





(ці моделі погано інтерпретуються); 2) призначені для отримання знань і придатні для керування у варіабельних умовах.

Моделі можна розділити на ті, що: 1) призначені для “предикції” значення певної характеристики; (ці моделі погано інтерпретуються); 2) призначені для отримання знань і придатні для керування у варіабельних умовах.