

"Програмування систем штучного інтелекту"

Методи та інструменти покращення якості даних



Чому якість даних критично важлива для AI?

Основні виклики

- Неточні дані призводять до помилкових рішень
- Неповні дані знижують ефективність моделей
- Непослідовність даних ускладнює навчання
- Застарілі дані дають неактуальні результати

Згідно з дослідженнями, до 80% часу data scientists витрачають на очищення та підготовку даних, а не на розробку моделей.



П'ять ключових способів впливу якості даних на AI-рішення

Точність прогнозів

Якісні дані забезпечують високу точність моделей машинного навчання

Швидкість розробки

Чисті дані скорочують час на preprocessing і налаштування

Довіра до результатів

Надійні дані підвищують впевненість у прийнятті рішень

Масштабованість

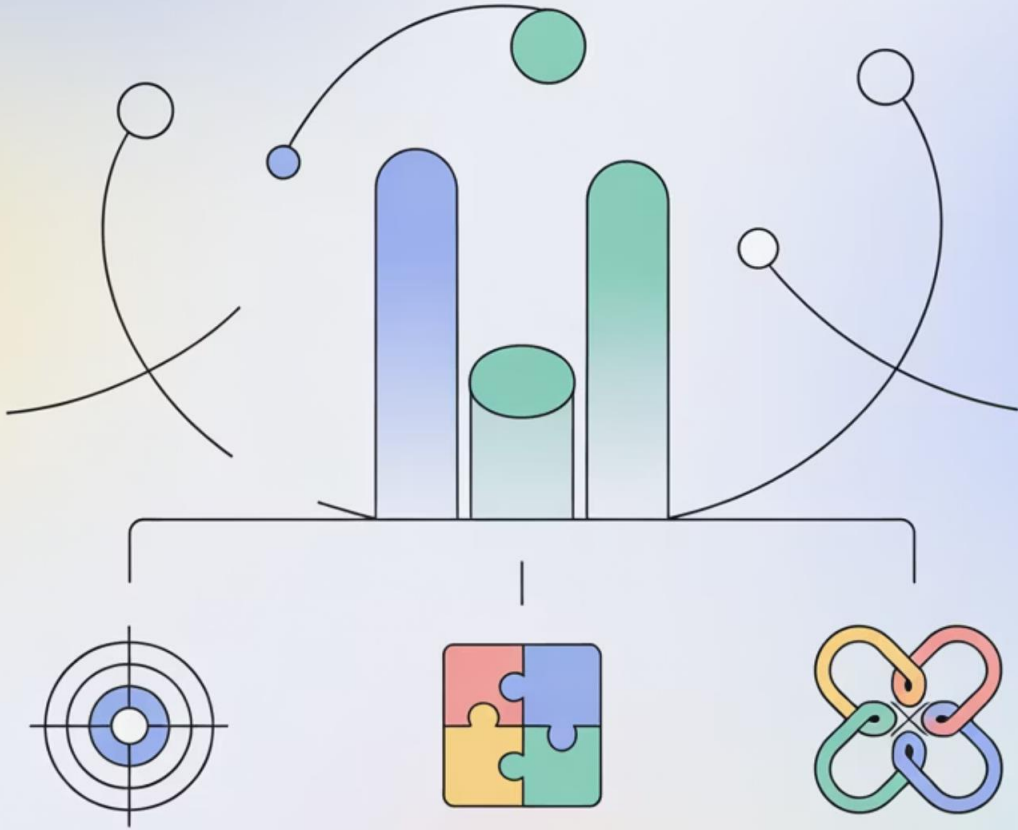
Структуровані дані полегшують розширення систем

Економічна ефективність

Якісні дані знижують витрати на виправлення помилок

Основні виміри якості даних

Data Quality Dimensions



Accuracy

Locustesc ocultig linuuct
contetgij dirmoheoodt
ingtimoting iirstukcent
cemuotigysiliit cluppes
closenotacodm doeilita
botentibietoeclierdt
dint eeing coluiss.

Completeness

Licoustees ocultig linuuct
bortetgij dirmoheoodt
ingtimoting iirstukcent
cemuotigysiliit cluppes
closenotacodm doeilita
botentibietoeclierdt
dint eeing coluiss.

Consistency

Licoustees ocultig linuuct
bortetgij dirmoheoodt
ingtimoting iirstukcent
cemuotigysiliit cluppes
closenotacodm doeilita
botentibietoeclierdt
dint eeing coluiss.

Шість основних вимірів якості даних



Точність (Accuracy)

Наскільки дані відповідають реальному стану речей. Неточні дані можуть призвести до хибних висновків та неправильних бізнес-рішень.



Повнота (Completeness)

Відсутність пропущених значень у критично важливих полях. Неповні дані обмежують можливості аналізу та знижують якість моделей.



Послідовність (Consistency)

Узгодженість формату та значень даних у різних джерелах. Непослідовні дані ускладнюють інтеграцію та аналіз.



Актуальність (Timeliness)

Своєчасність отримання та оновлення інформації. Застарілі дані можуть призвести до неправильних прогнозів у динамічних середовищах.



Валідність (Validity)

Відповідність даних встановленим правилам та обмеженням. Невалідні дані порушують цілісність бази даних.



Унікальність (Uniqueness)

Відсутність дублікатів записів. Дублікати спотворюють статистичні показники та результати машинного навчання.

Процес управління якістю даних



Оцінка поточного стану

Аналіз існуючих даних, виявлення проблем якості та встановлення базових метрик



Планування покращень

Розробка стратегії підвищення якості, встановлення цілей та пріоритетів



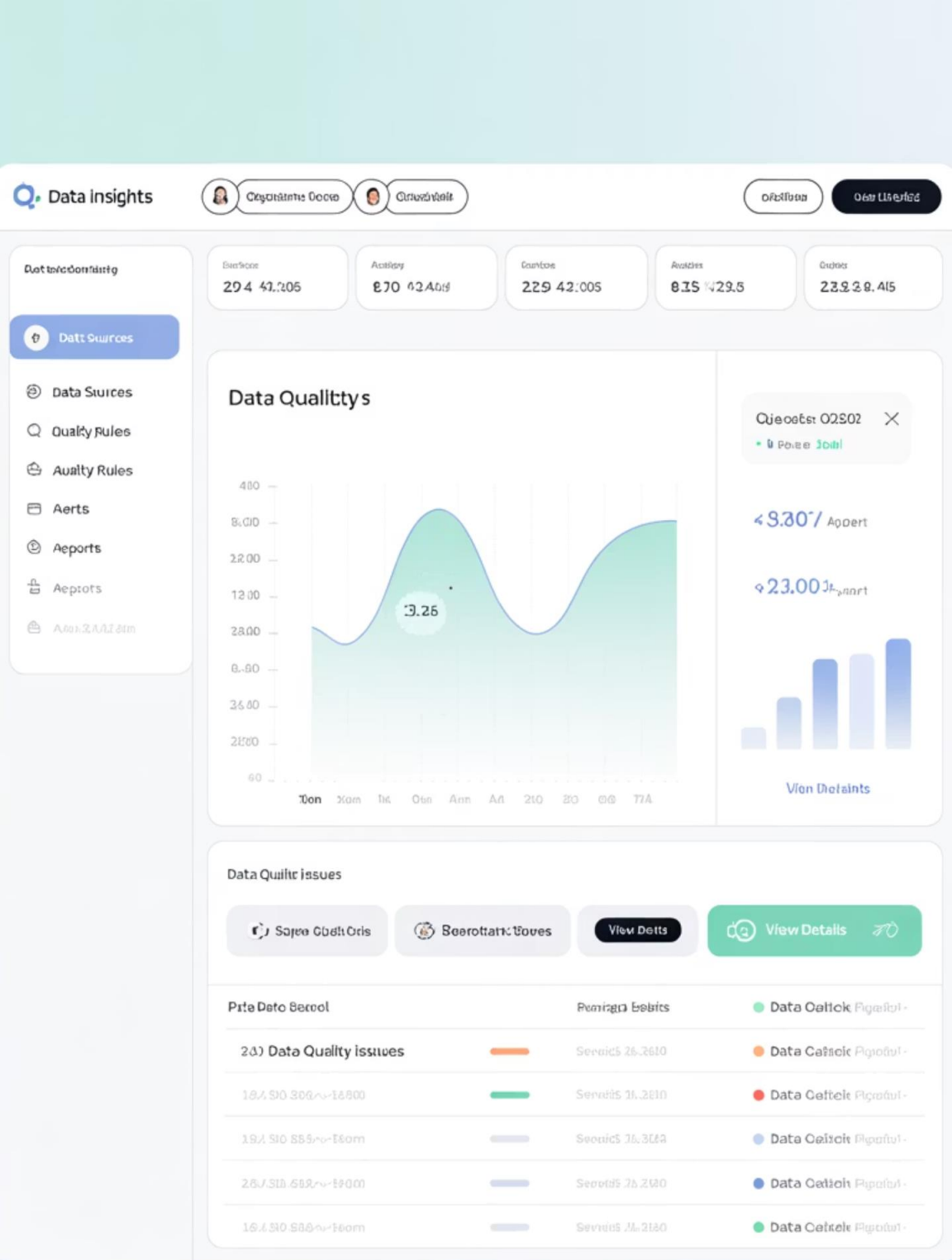
Впровадження рішень

Застосування інструментів очищення, валідації та стандартизації даних



Моніторинг і контроль

Постійне відстеження якості та своєчасне усунення нових проблем



Основні інструменти для покращення якості даних

Data Profiling Tools

- Аналіз структури та змісту даних
- Виявлення аномалій та відхилень
- Статистичний аналіз розподілу значень

ETL інструменти

- Витяг, трансформація та завантаження
- Стандартизація форматів даних
- Автоматизація процесів очищення

Системи валідації

- Перевірка бізнес-правил
- Контроль цілісності даних
- Автоматичне виявлення помилок

Методи очищення даних

Технічні методи

01 Видалення дублікатів

Алгоритми нечіткого зіставлення для виявлення схожих записів

02 Заповнення пропусків

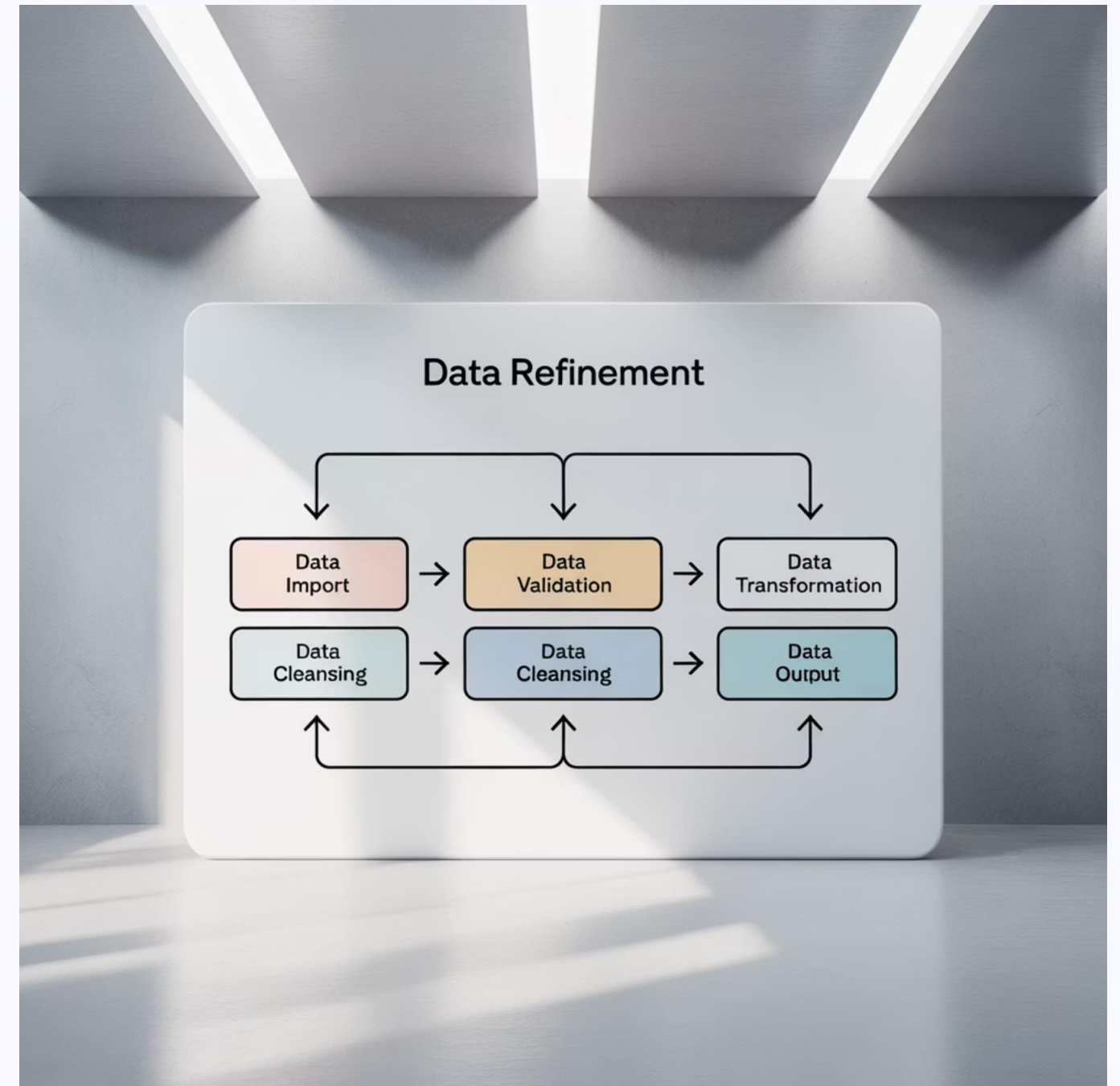
Використання статистичних методів або ML для відновлення даних

03 Стандартизація форматів

Приведення даних до єдиних стандартів

04 Виявлення викидів

Статистичні методи для знаходження аномальних значень



Практичний приклад: Очищення клієнтської бази

Вихідні дані (проблеми):

Ім'я	Email	Телефон	Місто
Іван Петров	ivan@gmail.com	+380501234567	Київ
І. Петров	ivan@gmail.com	050-123-45-67	Kiev
Марія	maria@invalid		Харків
Петро Сидоров	petro@test.ua	0671234567	Львів

Результат після очищення:

Ім'я	Email	Телефон	Місто
Іван Петров	ivan@gmail.com	+380501234567	Київ
Петро Сидоров	petro@test.ua	+380671234567	Львів

Застосовані методи: видалення дублікатів, стандартизація номерів телефонів, виправлення назв міст, видалення записів з некоректними email.

Метрики якості даних

95%

Цільова точність

Рекомендований
рівень точності для
критично
важливих систем

99%

Повнота даних

Мінімальний
відсоток
заповнених
обов'язкових полів

<5%

Рівень дублікатів

Максимально
допустимий
відсоток
дублювання
записів

24h

Час оновлення

Максимальний час
між отриманням та
обробкою даних

Регулярний моніторинг цих метрик дозволяє підтримувати високу якість даних та своєчасно виявляти проблеми.

Найкращі практики управління якістю даних

01

Встановлення стандартів

Розробка чітких правил введення, зберігання та валідації даних для всіх систем організації

02

Автоматизація перевірок

Впровадження автоматичних систем контролю якості на всіх етапах життєвого циклу даних

03

Навчання персоналу

Регулярні тренінги для співробітників щодо важливості якості даних та правильних практик

04

Постійний моніторинг

Налаштування системи алертів та регулярних звітів про стан якості даних

05

Культура якості

Формування організаційної культури, де якість даних є пріоритетом для всіх підрозділів

Практичний приклад: Підготовка даних для ML-моделі

Задача: Прогнозування цін на нерухомість

Проблеми в початкових даних:

- 15% відсутніх значень площі
- Різні формати адрес
- Викиди в цінах (помилки введення)
- Неконсистентні категорії районів

Методи очищення:

- Заповнення пропусків медіанними значеннями
- Стандартизація адрес через геокодування
- Видалення викидів ($>3 \sigma$)
- Створення єдиного довідника районів

Результат: Точність моделі підвищилась з 73% до 89% після очищення даних



Інструменти та технології

Python (Pandas, NumPy)

Популярні бібліотеки для очищення та аналізу даних з широкими можливостями для роботи з різними типами даних

Apache Spark

Платформа для обробки великих обсягів даних з вбудованими функціями очищення та валідації

SQL та NoSQL бази

Використання вбудованих функцій баз даних для валідації, очищення та забезпечення цілісності інформації

Спеціалізовані DQ інструменти

Talend, Informatica, Great Expectations – професійні рішення для комплексного управління якістю даних

Контрольні питання

1. Назвіть та поясніть шість основних вимірів якості даних.
2. Які основні етапи процесу управління якістю даних?
3. Як неякісні дані впливають на ефективність AI-моделей?
4. Які методи можна використати для заповнення пропущених значень?
5. Наведіть приклад практичного застосування метрик якості даних.
6. Які інструменти ви б рекомендували для очищення великих обсягів даних?
7. Чому важливо автоматизувати процеси контролю якості даних?
8. Як організаційна культура впливає на якість даних?



Дякую за увагу!

Питання та обговорення

