

ВСТУП ДО ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ (DATA MINING)

ЛЕКЦІЯ 1



МОВИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

План

- План курсу
- Данні
- Аналіз даних & інтелектуальний аналіз даних
- Моделі
- Трактуювання даних
- Помилки при роботі з даними.
- Типи даних.
- Чому саме Python?



ПЛАН КУРСУ

32 години лекцій
32 години практичні роботи
залік

Практичні роботи 8 ($100 * 0,60 = 60$)
Тести тематичні – 4 ($4 * 5 = 20$)

Додаткові 20 балів

СЕРТИФІКАТ

ПРОЕКТ

<https://www.kaggle.com/>

Портал - <https://learn.ztu.edu.ua/course/view.php?id=7720#section-1>

ДАННІ



*0,5 TB даних під час польоту Boeing 787, за
сутки 100 000 перельотів в середньому*

За 1 хв.

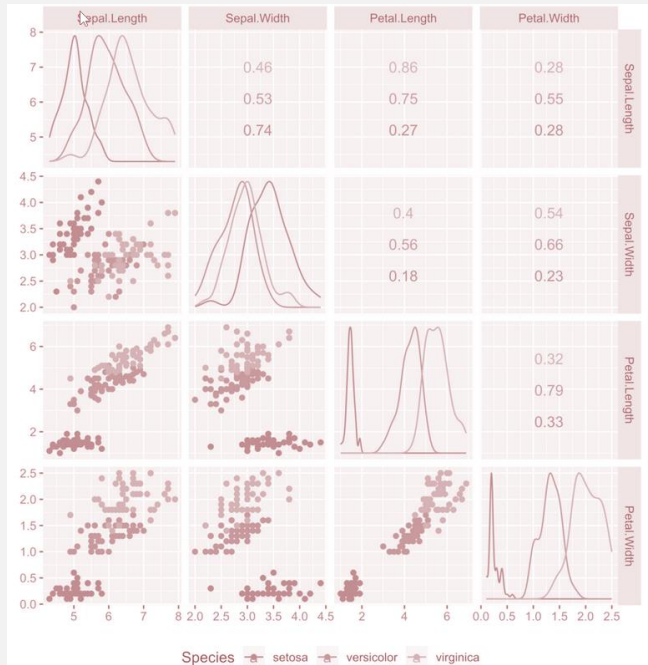
- Google - приблизно 2.5 млн. запитів
- Facebook – приблизно 700 000 тисяч користувачів заходить

надають інформацію

- Uber - найбільша в світі служба таксі, не володіє жодним авто
- Alibaba - retailer, нічого не виробляє, діяльність – торгові операції
- Airbnb - сервіс короткострокової оренди житла, не володіє жодним помешканням

ДАННІ

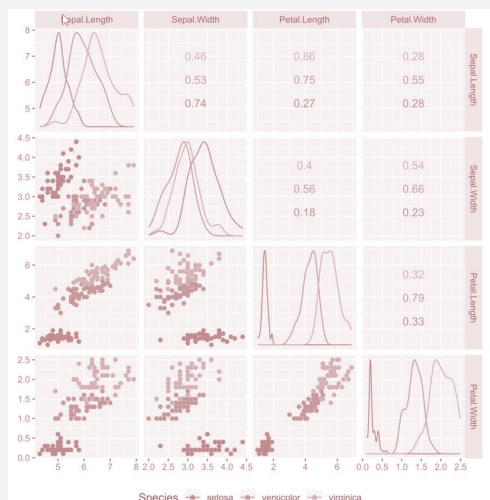
Коли дані мають сенс



- оцінка ризику трансгенних продуктів;
- оцінка ризику нових вакцин;
- передбачення кількості захворювань на грип або інші захворювання по країнах, регіонах тощо;
- передбачення результатів виборів;
- голосові асистенти;
- самокеровані автомобілі тощо.

ДАННІ

Коли дані мають сенс



КН

Коли ви можете виявити приховані закономірності (patterns) у великих масивах, які неможливо побачити "неозброєним оком".

Приклад. Аналіз логів користувачів для оптимізації архітектури хмарного сервісу. Сенс з'являється, коли модель каже: «Завтра о 14:00 навантаження зросте на 40%, розгортайте додаткові вузли».

КБ

Коли дані перетворюються на «сигнал тривоги». Величезний обсяг трафіку — це шум. Але якщо Data Mining виявляє мікроскопічне відхилення в протоколі передачі, яке характерне для Zero-day атаки, — ці дані безцінні.

Приклад. Побудова моделей виявлення вторгнень (IDS), де сенс полягає в мінімізації "false positives" (хибних спрацювань).

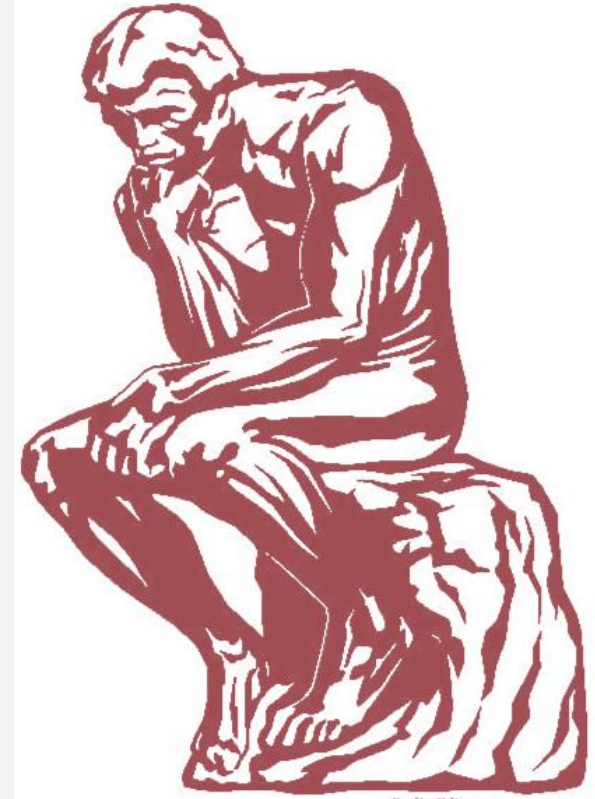
БІ

Коли ви можете трансформувати складні сигнали (ЕКГ, ЕЕГ, МРТ) у конкретний показник стану здоров'я.

Приклад. Аналіз мінливості серцевого ритму. Окремі удари серця — це просто цифри. Сенс з'являється, коли алгоритм аналізує ці інтервали й попереджає про ризик ішемії за годину до її прояву.

АНАЛІЗ ДАНИХ - ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ

- **Аналіз даних** - розділ математики, що займається розробкою методів обробки даних незалежно від їх природи.
- **Інтелектуальний аналіз даних** це сучасна концепція аналізу даних. Інтелектуальний аналіз даних - це обробка інформації та виявлення в ній моделей і тенденцій, які допомагають приймати рішення



Поняття аналізу даних з'явилося у 1978 році. Найбільшої популярності термін «аналіз даних» набув у першій половині 90-х років

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ

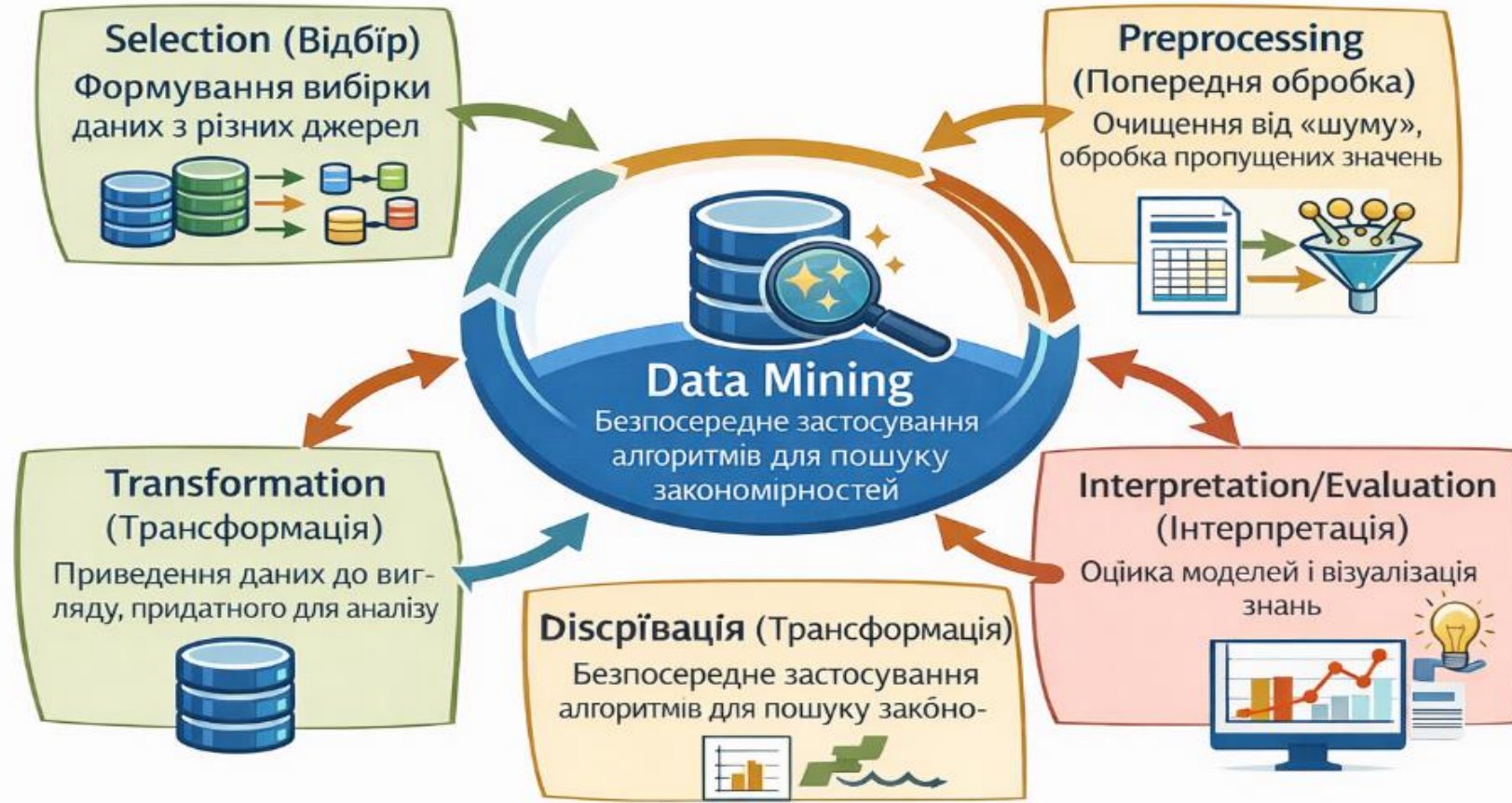


Моделі для знаходження нового знання з інформаційного сховища

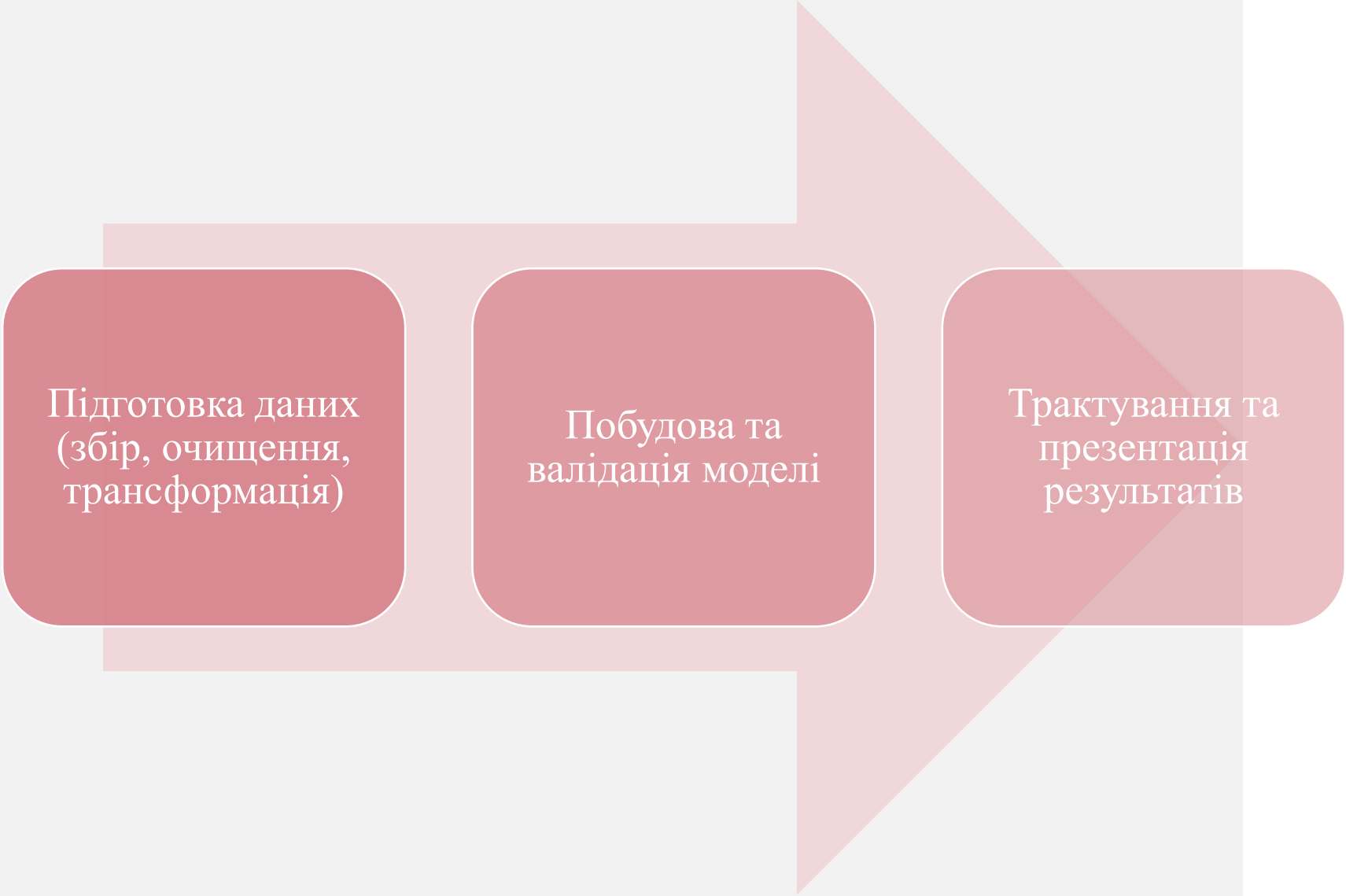
Виявлення відносин в даних



Терміни Data Mining та KDD часто вживають як синоніми, але Data Mining — лише один етап глобального процесу KDD.



ТРАКТУВАННЯ ДАНИХ



Підготовка даних
(збір, очищення,
трансформація)

Побудова та
валідація моделі

Трактування та
презентація
результатів

ТРАКТУВАННЯ ДАНИХ

Парадокс Сімпсона

Парадокс Сімпсона названо на честь дослідника Едварда Сімпсона, який у 1951 описав цей феномен. Хорошою ілюстрацією буде ситуація, що склалася в університеті Берклі в 1973. Тоді університет звинуватили в гендерній нерівності. Для ілюстрації трохи спростимо умови. Нехай в університеті є всього два факультети: А та В.

Факультет А			
	Подано заяв	прийнято	Відсоток прийнятих
чоловіки	900	450	50
жінки	100	80	80
Факультет В			
чоловіки	100	10	10
жінки	900	180	20

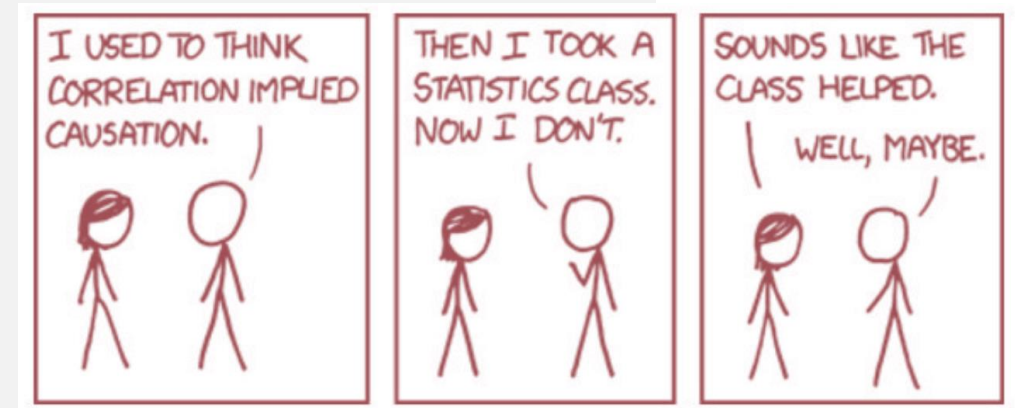
ТРАКТУВАННЯ ДАНИХ

Парадокс Сімпсона

Разом факультет А і В			
	Подано заяв	прийнято	Відсоток прийнятих
чоловіки	1000	460	46
жінки	1000	260	26

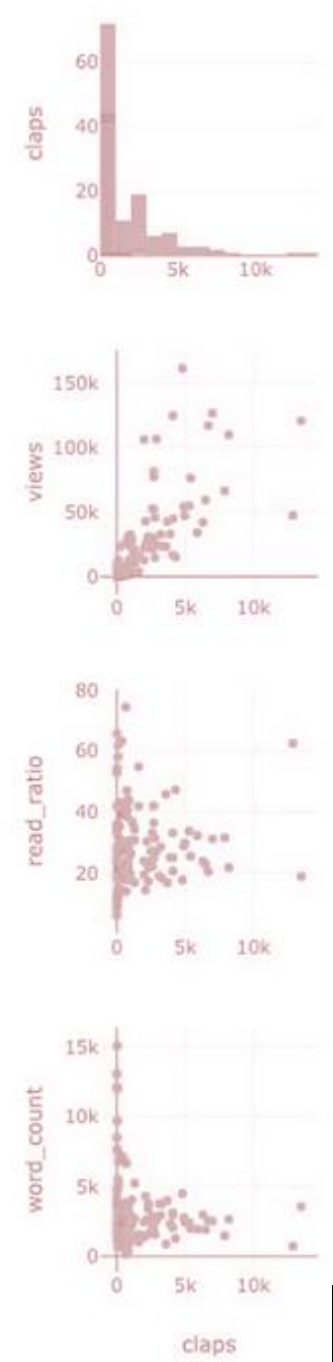
ПОМИЛКИ ПРИ РОБОТІ З ДАНИМИ

- помилки підготовки даних;
- обмежений доступ до даних у середині компанії;
- використання нерелевантних даних (неструктуровані, полу-структуровані);
- помилки вибірки;
- помилки кореляцій;
- уявна кореляція між непов'язаними факторами;
- втрата даних;
- помилки моделювання;
- неправильна інтерпретація даних;
- екстраполяція окремого випадку на загальну ситуацію;
- зайва віра в дані;
- отримати некоректні висновки через велику кількість викидів.

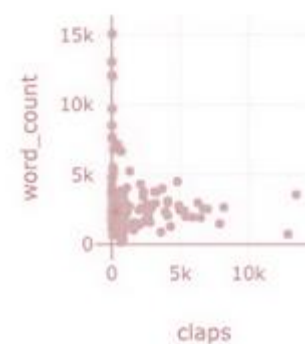
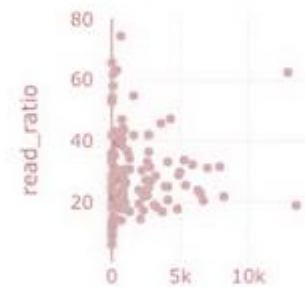
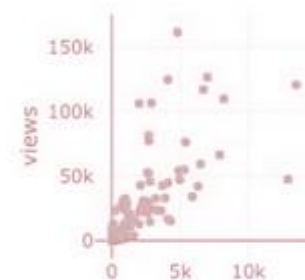
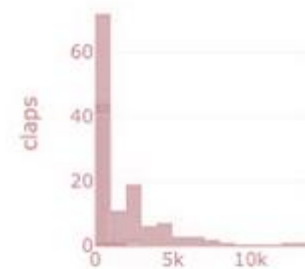


ТИПИ ДАНИХ

Кількісні	дискретні	кількість дітей у сім'ї
		кількість медалей олімпійської збірної
		кількість учнів
		кількість хворих
	неперервні	ріст
		вага
		заробітна плата
		швидкість
		площа



Категоріальні	невпорядковані	імена
		стать
		область
		назви міст
		групи крові
	впорядковані	бакалавр, магістр, доктор філософії
		погоджуюсь, не погоджуюсь, важко відповісти
	бінарні	так, ні
		0, 1
		є, немає



ПРИКЛАД. ЩО ВІДОМО ПРО ПСА?

Кількісні

дискретні

4 лапи

1 господар

неперервні

Вага: 25,5 кг

Зріст 565 мм

Категоріальні

Окрас: чорний

Характер: спокійний

Шерсть: довга



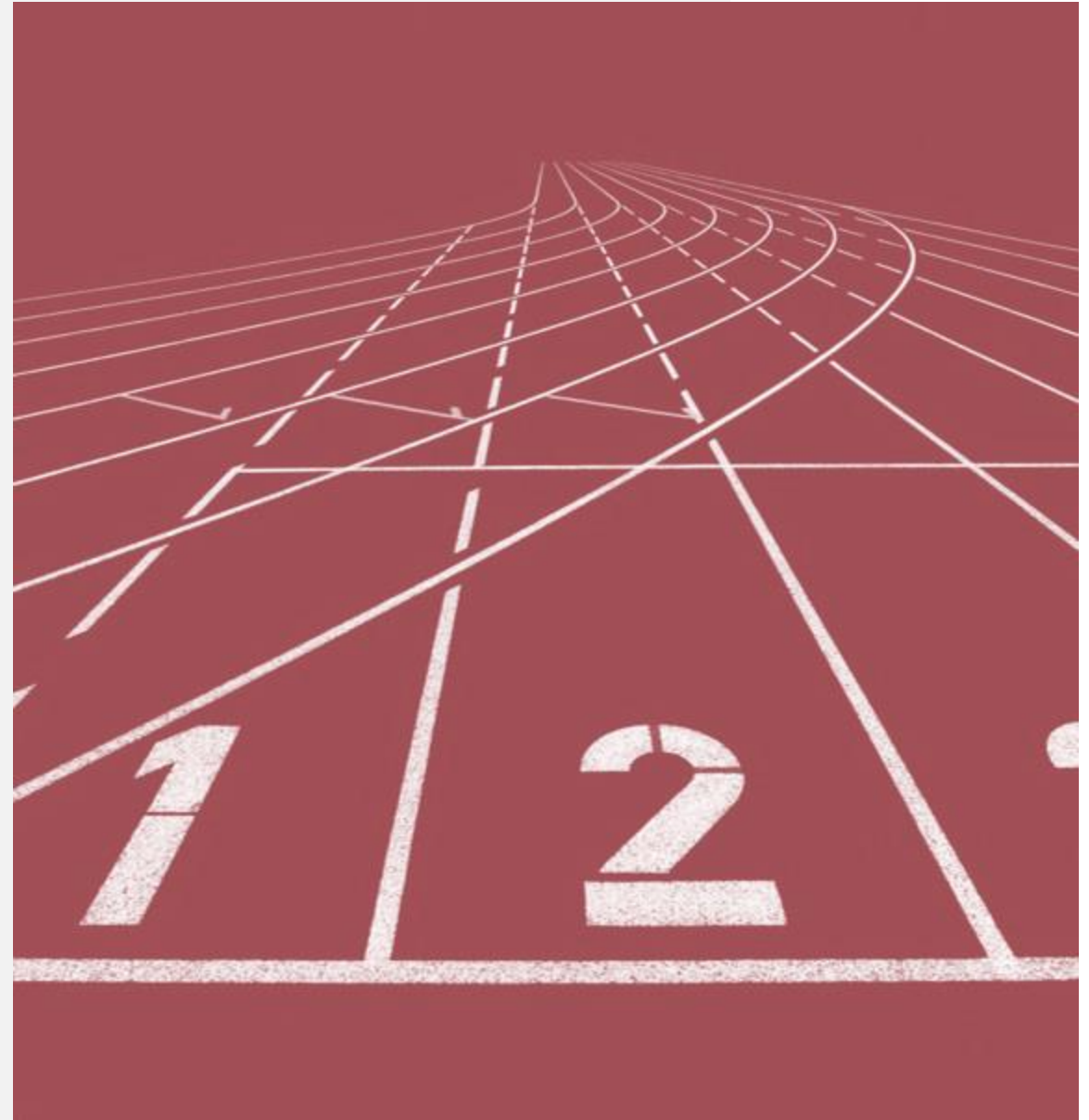
РІВНІ ВИМІРУ

Рівень виміру	Суть	Чи є порядок	Чи має сенс різниця	Абсолютний нуль	Приклади	Допустимі статистики
Номінальний	Категорії без порядку	☐ Ні	☐ Ні	☐ Ні	Стать, колір очей, штат, група крові	Частота, мода
Порядковий	Категорії з порядком	☒ Так	☐ Ні	☐ Ні	Освіта, рівень задоволеності, оцінки	Медіана, ранжування
Інтервальний	Числові дані з рівними інтервалами	☒ Так	☒ Так	☐ Ні	Температура (°C), роки, IQ	Середнє, дисперсія, кореляція
Відношень	Повноцінні числові дані	☒ Так	☒ Так	☒ Так	Вік, дохід, населення, кількість випадків	Усі статистичні методи

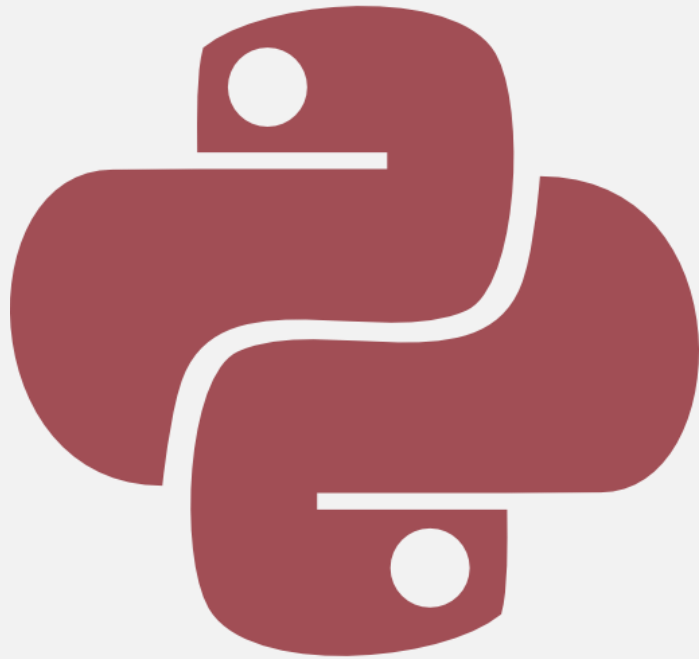
PYTHON

Чому саме Python?

- Простий, але виразний синтаксис.
- Багатий вибір бібліотек. І йдеться не лише про бібліотеки алгоритмів машинного навчання — на Python розробляють хмарні сховища, стрімінгові сервіси, ігри тощо.
- Висока культура документації. Python чудово документований і зазвичай бібліотеки на ньому продовжують цю традицію.



PYTHON



- Python - інтерпретована мова програмування
- Динамічна типізація
- Гарна підтримка модульності
- Підтримка об'єктно-орієнтованого програмування
- Вбудована підтримка Unicode в рядках
- Автоматичне прибирання сміття
- Інтеграція з C/C++, якщо можливостей Python недостатньо
- Зрозумілий та лаконічний синтаксис, що сприяє ясному відображенню коду
- Величезна кількість модулів, які входять в стандартну поставку Python 3
- Кросплатформеність.

ПРИКЛАД



Є таблиця з даними студентів (відвідування, оцінки, активність).
Які задачі Data Mining тут можливі?

Тип задачі	приклад	Алгоритми (Мови аналізу)
Класифікація	Прогнозування «групи ризику». Визначити, чи отримає студент незадовільну оцінку в кінці семестру на основі перших 4 тижнів навчання.	Logistic Regression, Decision Trees, Random Forest
Кластеризація	Сегментація студентів. Поділ на групи: «теоретики», «практики», «активні лідери», «пасивні слухачі» для персоналізації підходу.	K-means, DBSCAN
Регресія	Прогноз фінального балу. Розрахунок конкретної цифри балу (0–100) на основі поточної динаміки.	Linear Regression, Gradient Boosting (XGBoost/LGBM)
Пошук асоціативних правил	Аналіз контенту. Виявлення зв'язків типу: «Студенти, які дивилися відео А, зазвичай успішніше здають тест Б».	Apriori, FP-Growth

САМОСТІЙНА РОБОТА

Colaboratory (<https://colab.research.google.com/>)

дозволяє писати та виконувати код Python у браузері. При цьому:

- не потрібно ніякого налаштування;
- безкоштовний доступ до графічних процесорів;
- надавати доступ до документів іншим людям дуже просто.

Це відмінне рішення для фахівців з обробки даних та дослідників у галузі штучного інтелекту.