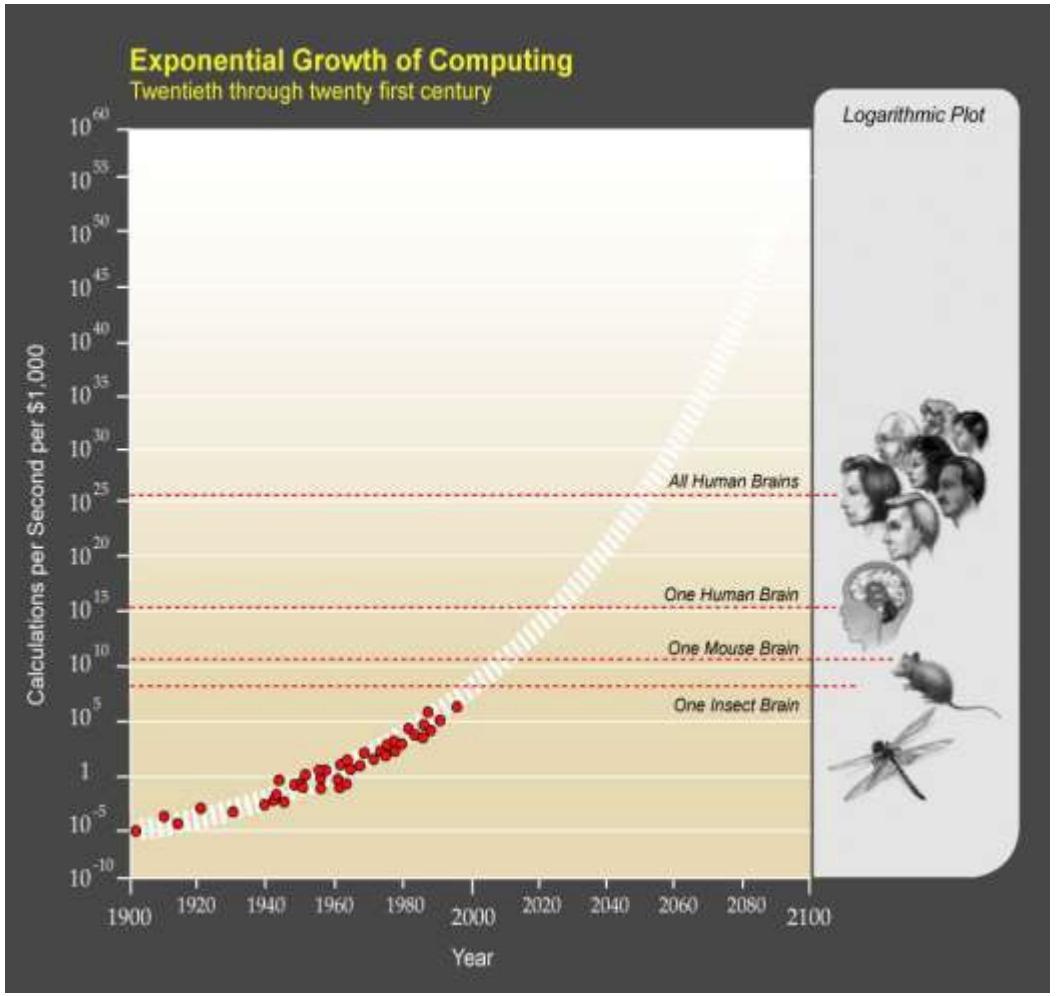
Вступ до машинного навчання

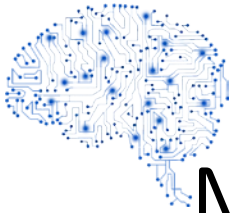
Професор, д.е.н. Ставицький А.В.

Щохвилини в Інтернеті

Закон Мура

- Закон Мура: максимальна обчислювальна потужність у світі подвоюється приблизно кожні два роки
- Тож світові комп’ютери в 1000 доларів зараз б’ють мишачий мозок, і вони перебувають приблизно на тисячній долі людського рівня





Мозок — наше все, але … Big Data

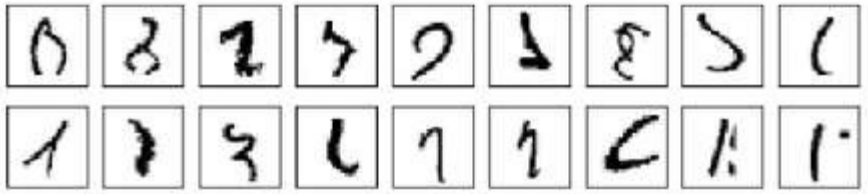
- Аналізувати дані вміють всі люди
- Виживання людини як біологічного виду обумовлено здатністю мозку:
  - бачити взаємозв’язки подій
  - робити висновки на основі фактів
  - вчитися на досвіді (своєму і чужому)
- Але даних дуже багато, тому доручаємо машині:
  - шукати зв’язки
  - виявляти закономірності
  - формувати відповіді на питання

DS — вже невід’ємна частина життя сучасної людини

- Розмови з голосовим асистентом у смартфоні (наприклад, Google Assistant)
- Надання рекомендацій відносно товару, що найкраще підходить (Amazon, Netflix…)
- Фільтрація спаму у вхідних повідомленнях електронної пошти
- Виявлення та діагностування внутрішніх захворювань

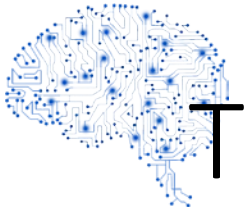
Наскільки добре програма розпізнає рукописні цифри?

- Набір даних зібраниq Національним інститутом стандартів і технологій США (NIST)
- У ньому 50000 зображень та 10 000 зображень для перевірки



73 рядки коду Python правильно розпізнали 9659 з 10000 (96,59%)

```
    network.py
import random

import numpy as np

class Network(object):

    def __init__(self, sizes):

        self.num_layers = len(sizes)

        self.sizes = sizes

        self.biases = [np.random.randn(y, 1) for y in sizes[1:]]

        self.weights = [np.random.randn(y, x)

                        for x, y in zip(sizes[:-1], sizes[1:])]

    def feedforward(self, a):

        for b, w in zip(self.biases, self.weights):

            a = sigmoid(np.dot(w, a)+b)

        return a

    def SGD(self, training_data, epochs, mini_batch_size, eta,

            test_data=None):

        if test_data: n_test = len(test_data)

        n = len(training_data)

        for j in xrange(epochs):

            random.shuffle(training_data)

            mini_batches = [

                training_data[k:k+mini_batch_size]

                for k in xrange(0, n, mini_batch_size)]

            for mini_batch in mini_batches:

                self.update_mini_batch(mini_batch, eta)

        if test_data:

                print "Epoch {0}: {1} / {2}".format(

                    j, self.evaluate(test_data), n_test)

            else:

                print "Epoch {0} complete".format(j)

    def update_mini_batch(self, mini_batch, eta):

        nabla_b = [np.zeros(b.shape) for b in self.biases]

        nabla_w = [np.zeros(w.shape) for w in self.weights]

        for x, y in mini_batch:

            delta_nabla_b, delta_nabla_w = self.backprop(x, y)

            nabla_b = [nb+dnb for nb, dnb in zip(nabla_b, delta_nabla_b)]

            nabla_w = [nw+dnw for nw, dnw in zip(nabla_w, delta_nabla_w)]

        self.weights = [w-(eta/len(mini_batch))*nw

                        for w, nw in zip(self.weights, nabla_w)]

        self.biases = [b-(eta/len(mini_batch))*nb

                       for b, nb in zip(self.biases, nabla_b)]

    def backprop(self, x, y):

        nabla_b = [np.zeros(b.shape) for b in self.biases]

        nabla_w = [np.zeros(w.shape) for w in self.weights]

        activation = x

        activations = [x]

        zs = []

        for b, w in zip(self.biases, self.weights):

            z = np.dot(w, activation)+b

            zs.append(z)

            activation = sigmoid(z)

            activations.append(activation)

        delta = self.cost_derivative(activations[-1], y) * \

            sigmoid_prime(zs[-1])

        nabla_b[-1] = delta

        nabla_w[-1] = np.dot(delta, activations[-2].transpose())

        for l in xrange(2, self.num_layers):

            z = zs[-l]

            sp = sigmoid_prime(z)

            delta = np.dot(self.weights[-l+1].transpose(), delta) * sp

            nabla_b[-l] = delta

            nabla_w[-l] = np.dot(delta, activations[-l-1].transpose())

        return (nabla_b, nabla_w)

    def evaluate(self, test_data):

        test_results = [(np.argmax(self.feedforward(x)), y)

                        for (x, y) in test_data]

        return sum(int(x == y) for (x, y) in test_results)

    def cost_derivative(self, output_activations, y):

        return (output_activations-y)

def sigmoid(z):

    return 1.0/(1.0+np.exp(-z))

def sigmoid_prime(z):

    return sigmoid(z)*(1-sigmoid(z))
```

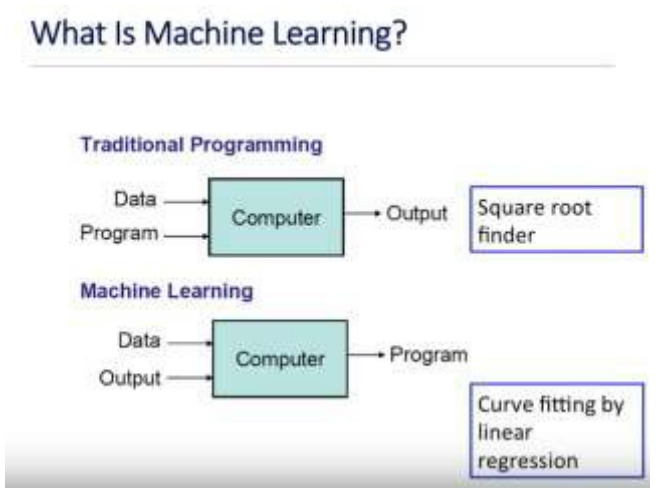


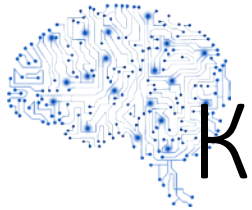
Традиційне програмування проти машинного навчання

- «За останні десятиліття комп'ютери автоматизували багато процесів, які програмісти могли описати через точні правила і алгоритми. Сучасні техніки машинного навчання дозволяють нам робити те ж саме з завданнями, для яких набагато складніше поставити чіткі правила» (С) Дж.Безос

Традиційне програмування проти машинного навчання – 2

- А що таке взагалі навчання?
- Важко сформулювати?! Давайте розберемося на прикладі: як Ви навчилися запрограмовувати розв’язання квадратних рівнянь.
- Шляхом тренування здобули необхідний досвід – що є різні випадки, що коренів може не бути тощо.
- Схема навчання – як на верхній схемі.
- Але ж буває й інший вид навчання.
- Навіть щура можна навчити, що якщо натиснеш цю кнопку, то отримаєш їжу, а якщо цю – то вдарить струмом. З часом щур фактично придумає алгоритм: якщо йому хочеться їсти, то натискує одну кнопку, а якщо хочеться адреналіну – то іншу.
- Це – інше, так зване «машинне» навчання, коли на вході дані та правильні відповіді, а на виході – програма (алгоритм), яку комп’ютер сам написав! Це – навчання з учителем, бо він знає правильні відповіді. Але є ще крутішій варіант машинного навчання – без учителя!





Концептуальна зміна у середині 2010-х років

- Невизначеність – це не баг, це фіча!

Розпізнавання зображень – 1

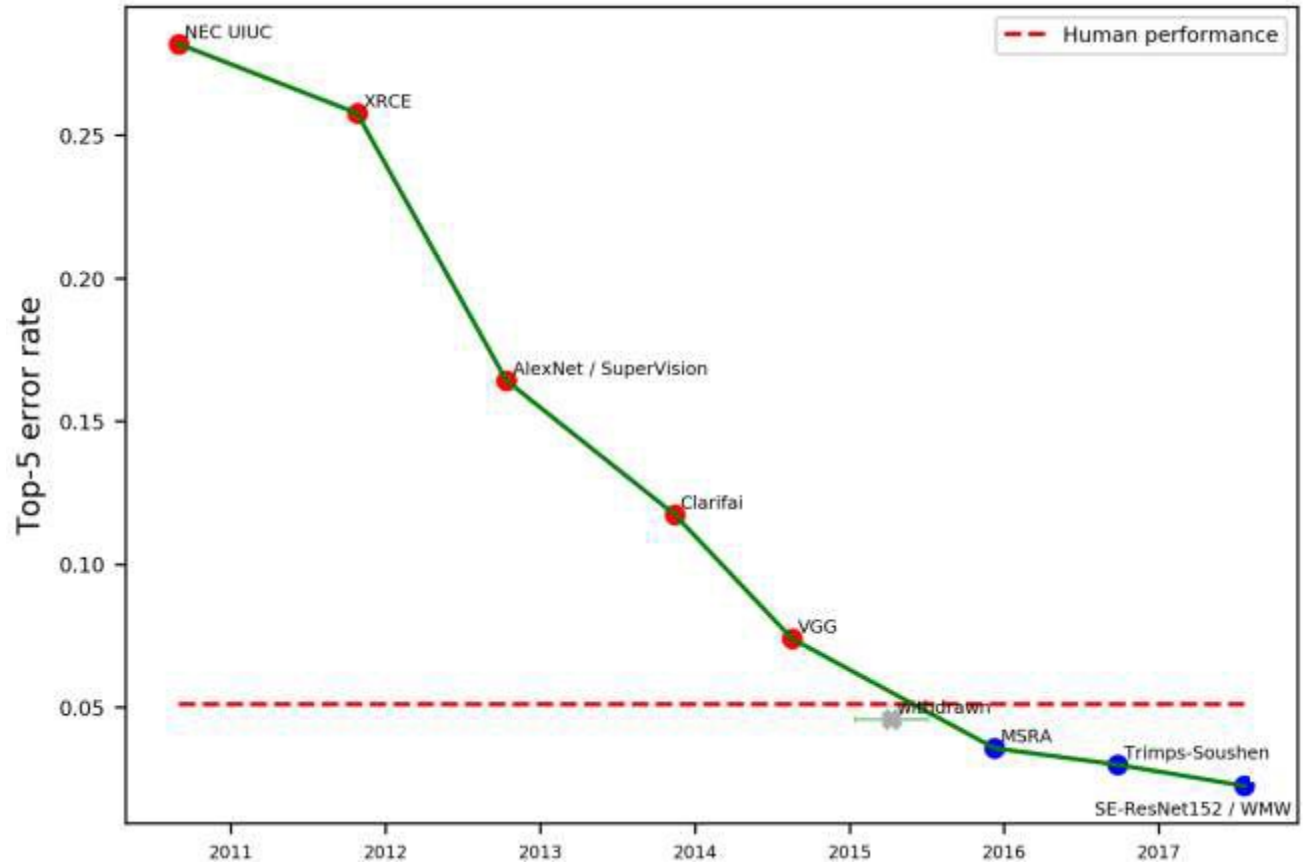
- Понад 14 мільйонів зображень було вручну анотовано проектом ImageNet, щоб вказати, які об'єкти зображені у принаймні в одному мільйоні картинок
- ImageNet містить понад 20000 категорій, кожна з яких складається з декількох сотень зображень

Розпізнавання зображень – 2

Леопард

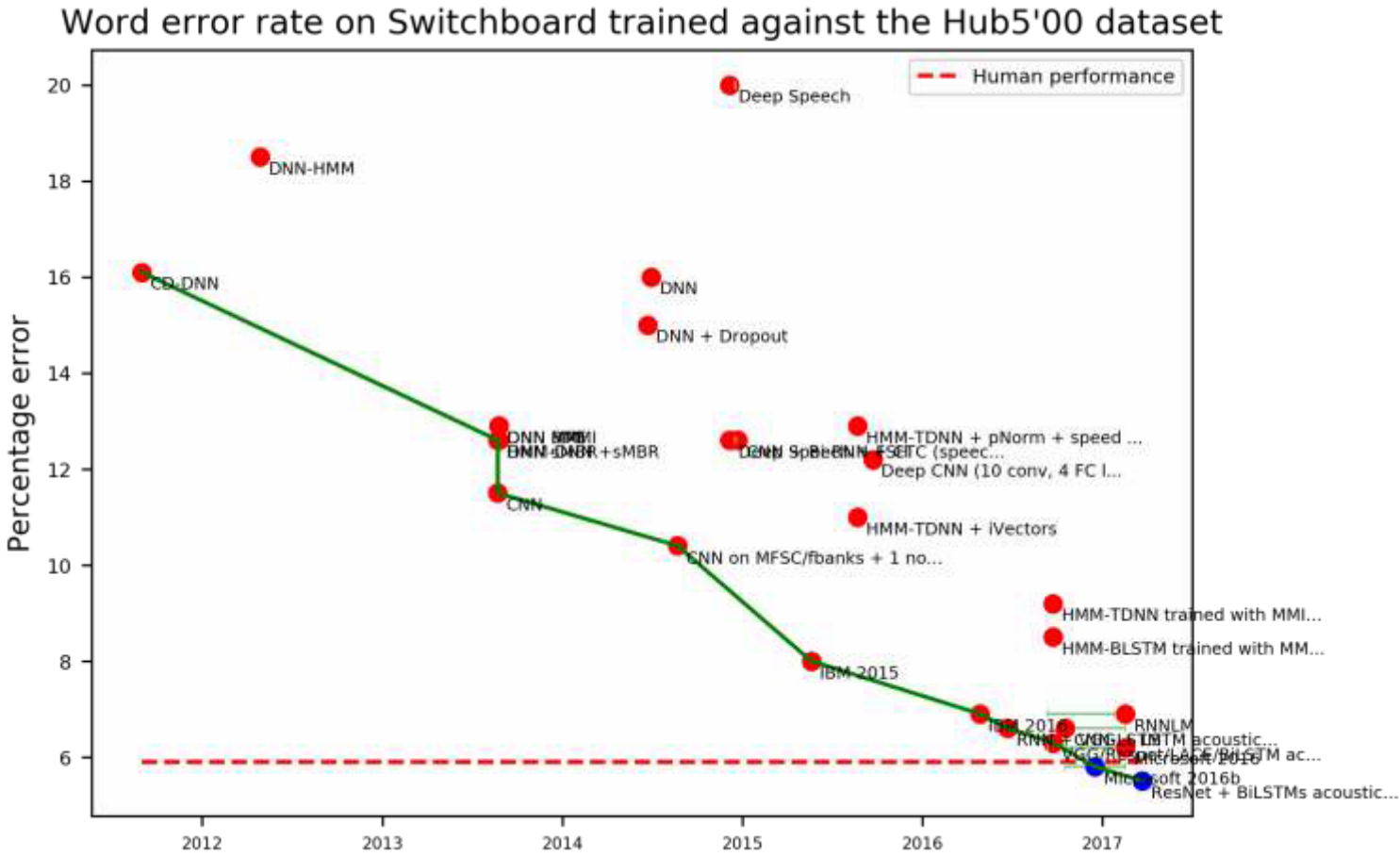
Собака, далматинець, вишня?

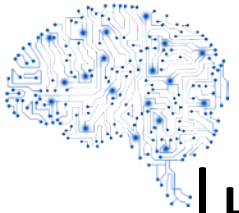
Розпізнавання зображень – 3



Розпізнавання мови

- Дані були взяті з набору неформальних телефонних розмов на задані теми:
- 40515 слів
- 4583 речення
- 30 людей

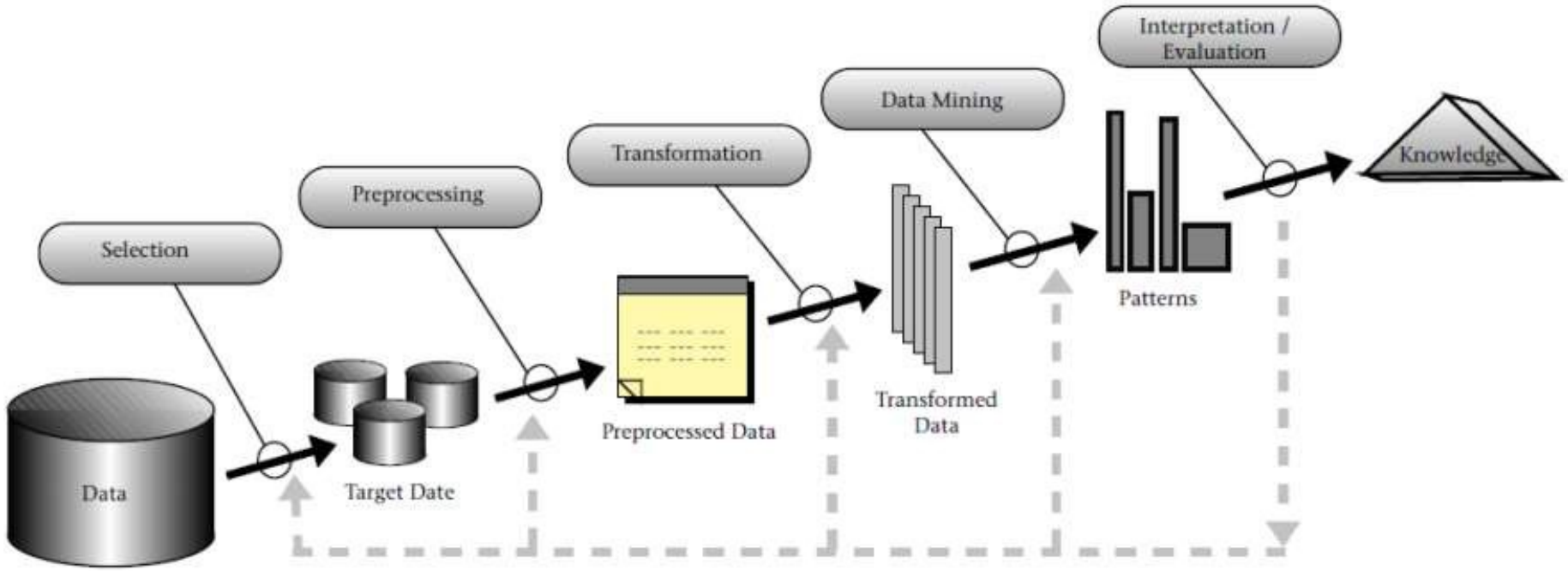




Інвестиції в ШІ зросли під час пандемії

- Чверть IT-спеціалістів збільшила рівень інвестицій у ШІ завдяки COVID-19, 42% утримали його на колишньому рівні, але 75% планують продовжити або розпочати нові проекти ШІ протягом наступних 6-9 місяців.

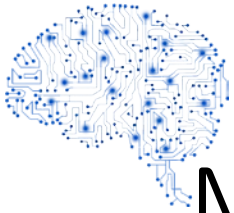
Виявлення знань у базах даних



Data mining - це процес виявлення цікавих закономірностей та знань із великих обсягів даних

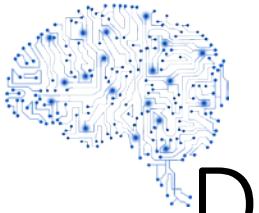
Аналіз даних проти математичної статистики

Різниця	Математична статистика	Аналіз даних
Очевидна (різні обсяги даних, що обробляються)	Не Big Data	Big Data
Неочевидні (різні типи основних завдань)	намагається вивести властивості навколишнього світу на основі спеціально зібраних даних. Приклад: методи перевірки статистичних гіпотез відіграють дуже важливу роль, коли агроном хоче зрозуміти, який сорт насіння, за інших рівних умов, принесе найкращий урожай, або лікар намагається визначити, чи новий метод обробки дає помітно кращий результат, ніж існуюча техніка. Щоб відповісти на такі запитання, потрібно ретельно провести експеримент, отримати порівнянні дані та точно порівняти результати з урахуванням їх випадкового розарування.	зосереджена на пошуку будь-яких закономірностей, структурі наявних даних. Приклад: дані в результаті чийогось спостереження. Це можуть бути дані про соціально-економічний стан деяких країн за один рік. Або це може бути колекція повідомлень, надісланих учасниками соціальної мережі протягом певного періоду. У таких ситуаціях типові запитання: Що означають ці дані? Чи існує якась структура даних для набору об'єктів, про які йдеться? Чи можуть ці фактори допомогти передбачити їх?



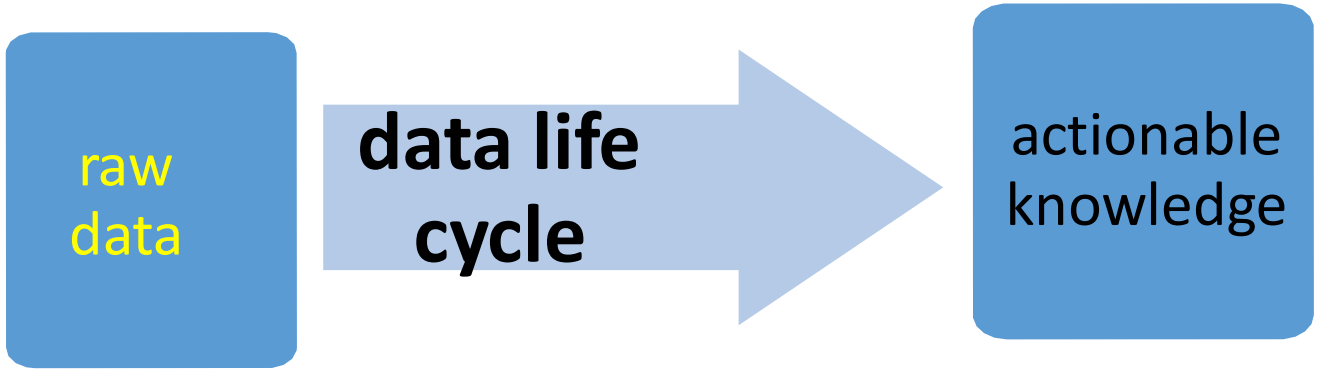
Малі дані

- У деяких областях малі дані дуже поширені, особливо в медицині, клінічних випробуваннях тощо, коли є лише 20 або 30 зразків.
- (Надзвичайно) мала кількість спостережень викликана наступним фактом, що важко знайти достатню кількість бажаючих взяти участь у дослідженні. Ось чому найважливішою частиною кожного клінічного випробування є визначення мінімального обсягу вибірки, необхідної для отримання передбачуваної точності.

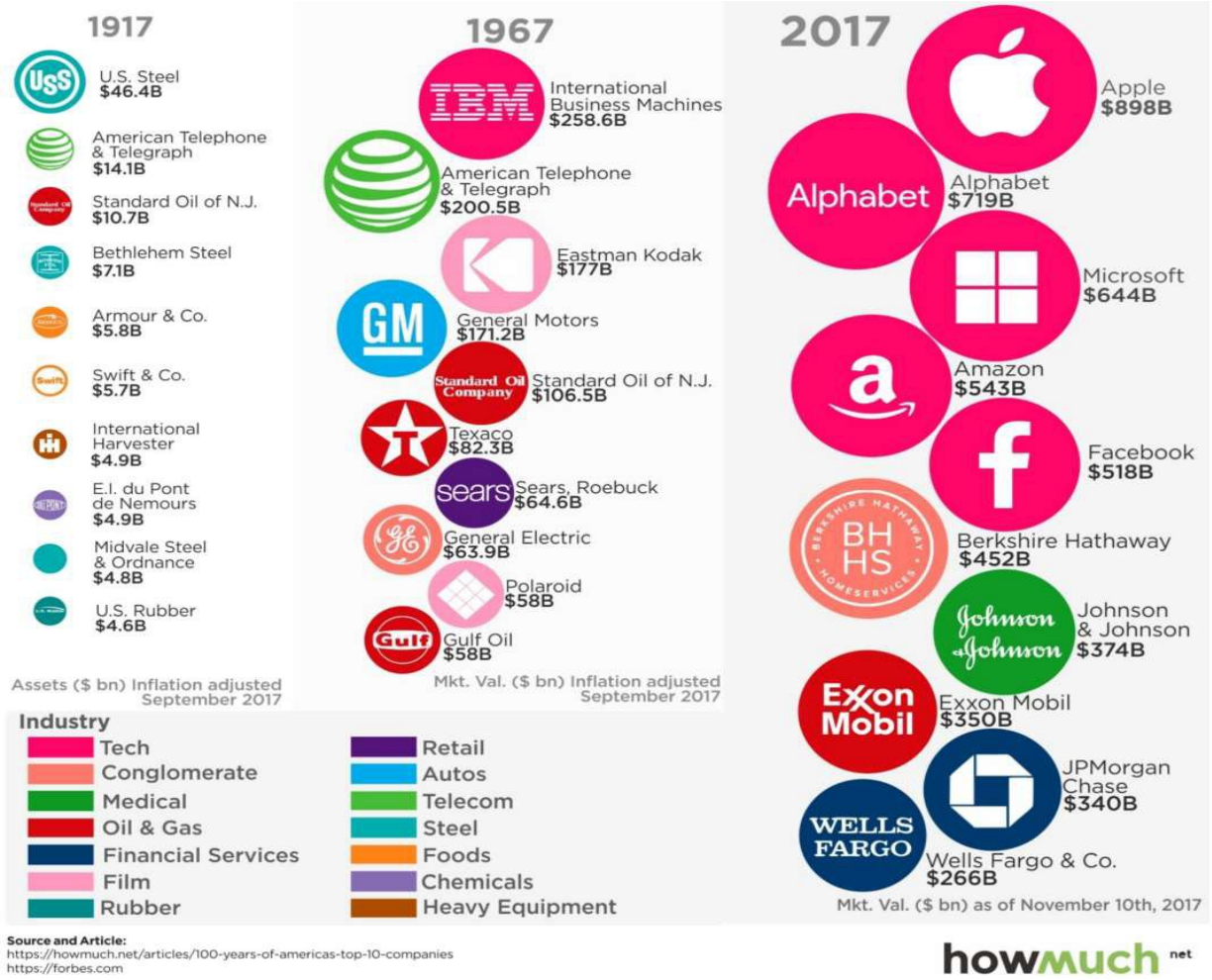


Data Science

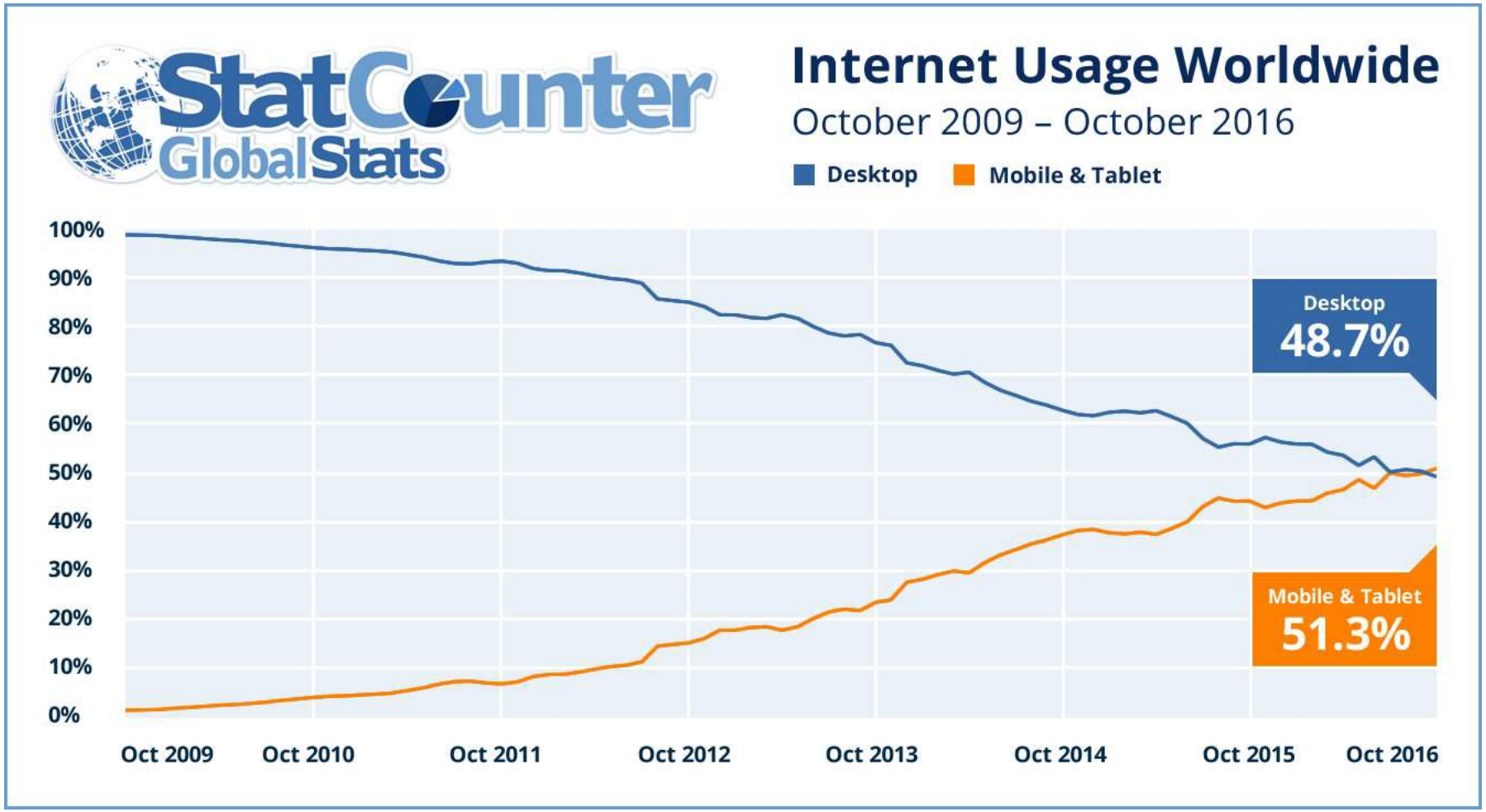
- Отримання знань безпосередньо з даних шляхом процесу відкриття, або формулювання гіпотези та перевірки гіпотез
- Практик, який володіє достатніми знаннями в предметній області, аналітичними навичками та навчиками розробки програмного забезпечення та систем для управління наскрізними процесами даних у життєвому циклі даних

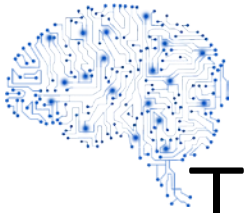


Яка компанія є найбільш креативною?



Жовтень 2016 року





Технології МН в Google

- Технології МН суттєво застосовуються в сервісах пошуку, перекладання та обробки зображень
- Із листопада 2015 р. розроблена бібліотека Tensor Flow, що прискорює процес побудови глибинних нейронних мереж, має відкритий вихідний код
- У липні 2018 р. представлено окремий сопроцесор спеціального призначення Edge TPU (Tensor Processing Unit) для апаратного прискорення вже навчених нейроалгоритмів на пристроях типу датчики чи камери

Визначення машинного навчання

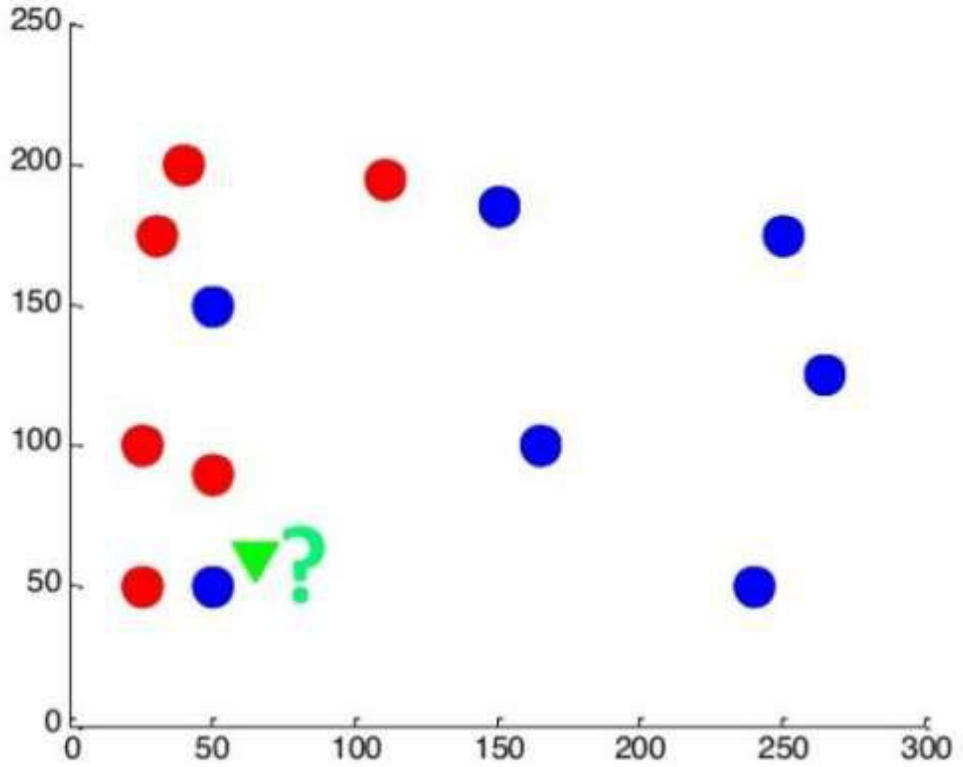
- Машинне навчання — це процес, в результаті якого машина (комп'ютер) здатна показувати поведінку, яка в неї не була явно закладена (запрограмована).
- Якщо бути точнішим, ми говоримо, що **машина навчається** стосовно певного завдання T, якщо її продуктивність відносно метрики P поліпшується з набуттям досвіду E.

Приклади класів задач Т в МН

- **Класифікація** — віднесення об'єкта до однієї з категорій на підставі його ознак
- **Регресія** — прогнозування кількісної ознаки об'єкта на підставі інших його ознак
- **Кластеризація** — розбиття множини об'єктів на групи на підставі ознак цих об'єктів так, щоб усередині груп об'єкти були схожі між собою, а поза однієї групи — менш схожі
- **Пошук аномалій (=викидів)** — пошук об'єктів, «сильно несхожих» на всі інші у вибірці або на якусь групу об'єктів

Задача класифікації

- Є кружечки тільки двох кольорів: червоні та сині
- Зеленим трикутником позначено невідомий нам кружечок
- До якого класу його віднести?

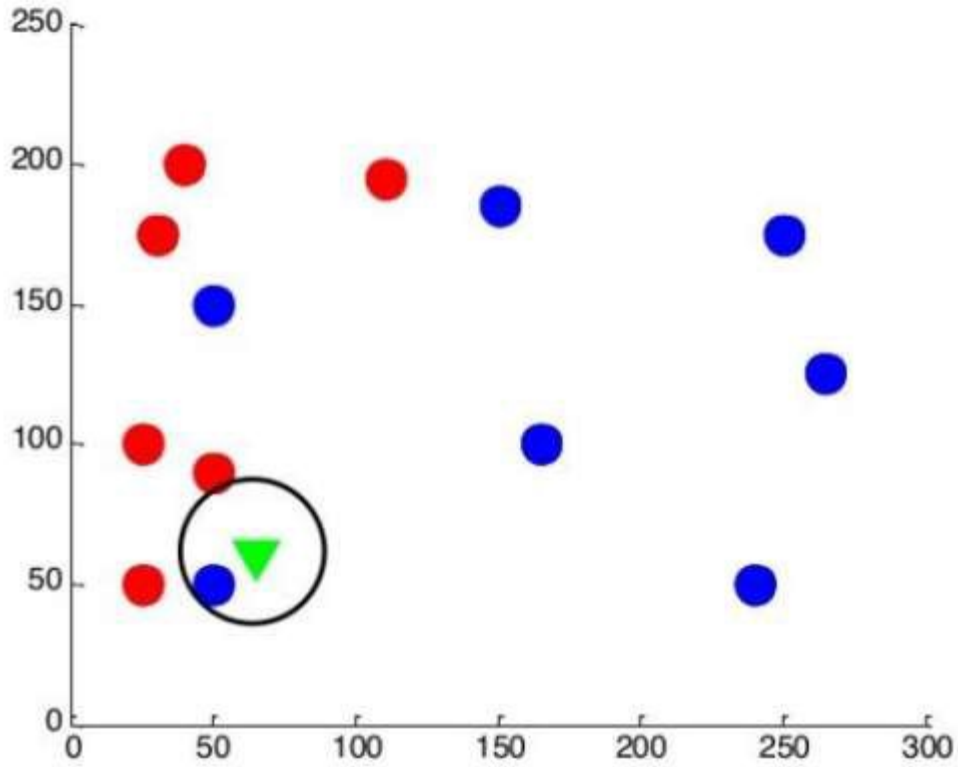


https://habrahabr.ru/company/yandex/blog/206058/

Метод найближчого сусіда

- Відстань між кружечками рахуємо у звичайній евклідовій метриці:

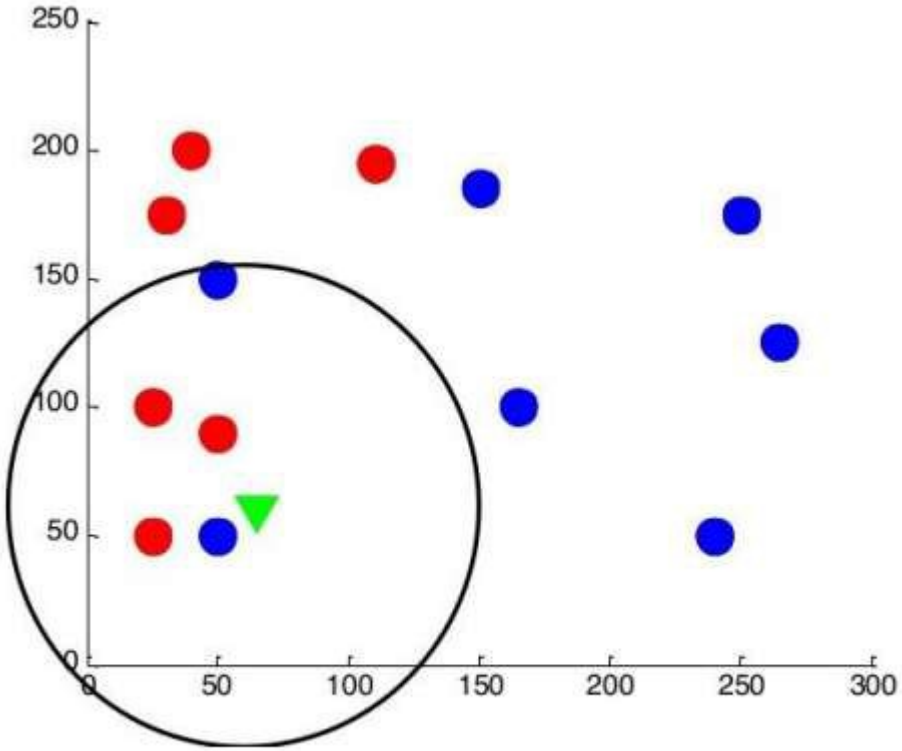
$$d_{ij} = \sqrt{\sum_{k=1}^{m} \left(x_{ik} - x_{jk} \right)^2}$$



https://habrahabr.ru/company/yandex/blog/206058/

Метод 5 найближчих сусідів

- А скільки взяти найближчих сусідів до уваги?
- На «нескінченних» вибірках метод k найближчих сусідів дає оптимальний метод класифікації



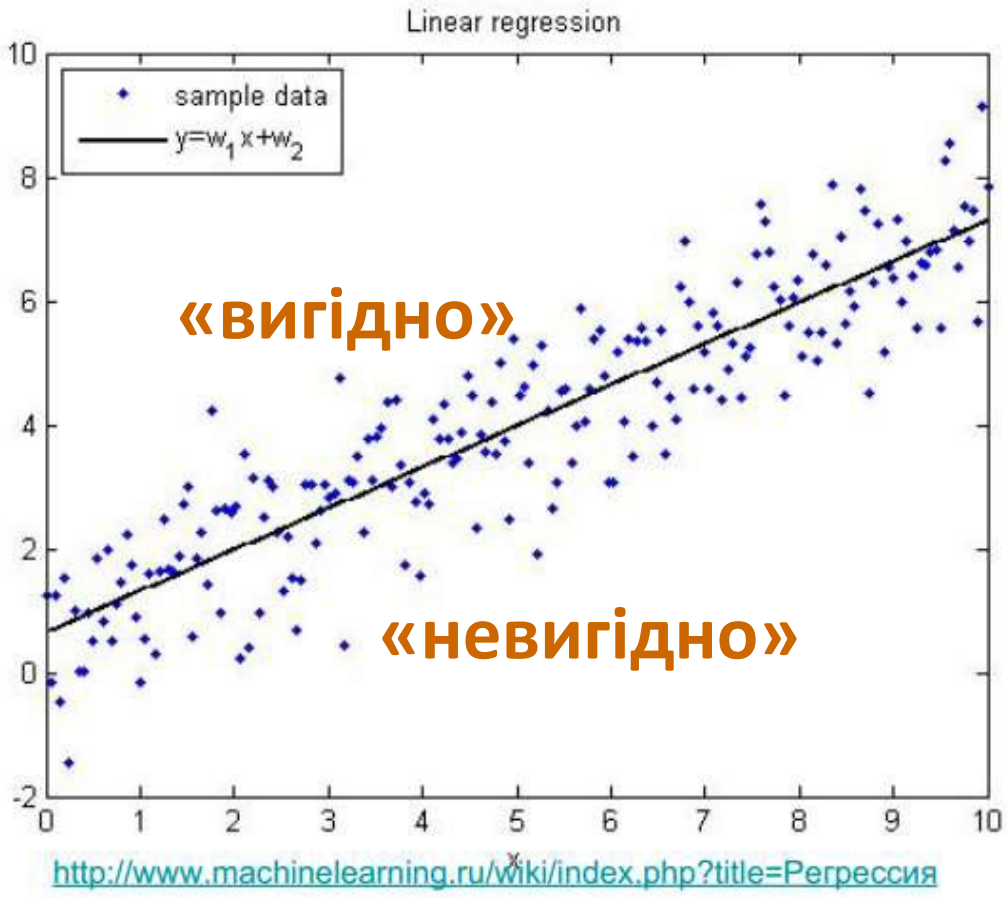
https://habrahabr.ru/company/yandex/blog/206058/

Приклади конкретних задач класифікації

- Кредитний скоринг
- Ідентифікація вигідних клієнтів
- Пошук нафтових чи газових родовищ, золотих рудників тощо на основі даних про відомі місця
- Пошук імен людей чи назв географічних місць у тексті
- Ідентифікація людей по фотографіям чи за записом голосу
- Ідентифікація захворювань
- Передбачення команди, що виграє Лігу чемпіонів

Задача регресії

- Вартість будинку як функція від його площі
- Задачі класифікації та регресії часто взаємопов’язані:
  - «вигідно» і «невигідно»
  - кредитний скоринг

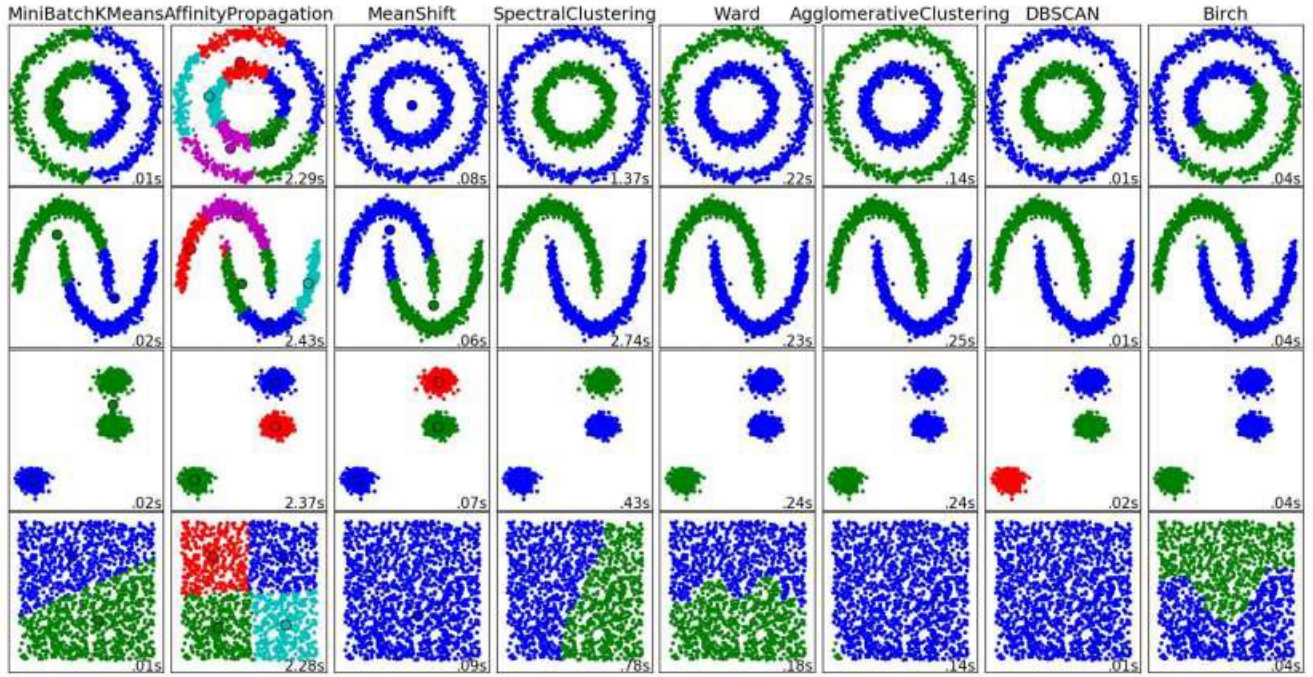


Приклади конкретних задач регресії

- Кредитний скоринг
- Ідентифікація вигідних клієнтів
- Пошук нафтових чи газових родовищ, золотих рудників тощо на основі даних про відомі місця
- Прогнозування суми, що людина витратить на певний продукт
- Прогнозування річного доходу компанії

Задача кластеризації

- Розбиття множини об'єктів на групи на підставі ознак цих об'єктів так, щоб усередині груп об'єкти були схожі між собою, а поза однієї групи — менш схожі.



http://scikit-learn.org/stable/modules/clustering.html

Приклади конкретних задач кластеризації

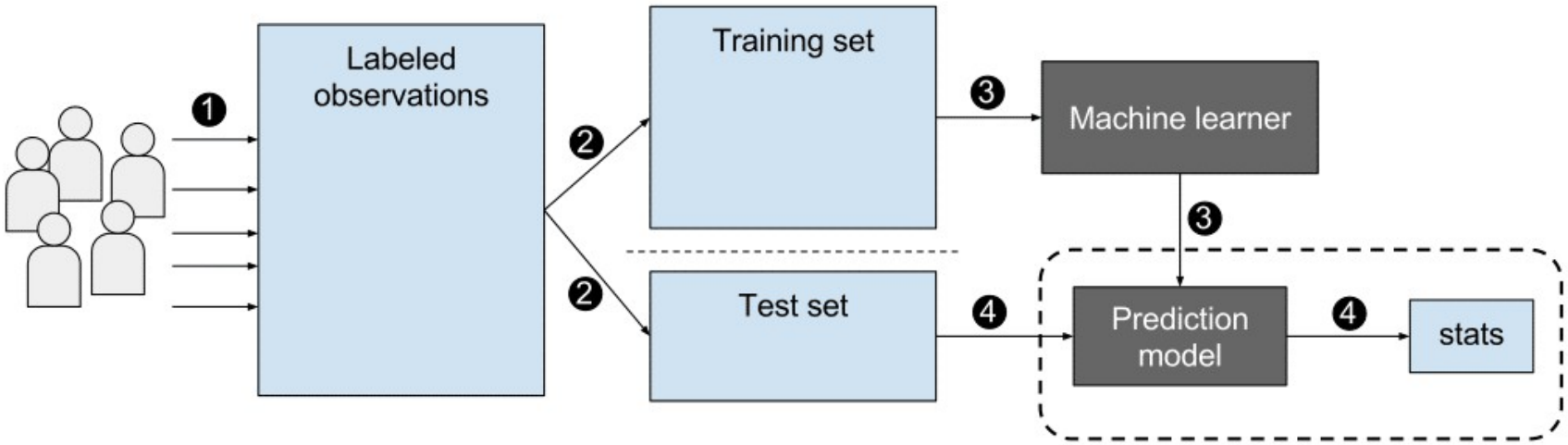
- Сегментація цільової аудиторії сайту
- Ідентифікація груп сімей — споживачів певного товару для розробки стратегії позиціонування бренду
- Тематичне моделювання електронних листів
- Кластеризація символів в незалежності від їх шрифту, розміру тощо (для подальшого розпізнавання)

Поняття досвіду у МН

- Під досвідом E розуміються дані
- Алгоритми машинного навчання діляться на ті, що навчаються з учителем і без учителя (контрольоване і неконтрольоване навчання, supervised & unsupervised learning)
- У завданнях навчання без учителя є вибірка, що складається з об'єктів, які описуються набором ознак
- У завданнях навчання з учителем на додачу до цієї навчальної вибірки для кожного об’єкта відома цільова ознака — те, що хотілося б спрогнозувати для інших об'єктів (=об'єктів не з навчальної вибірки)

Навчання під наглядом

- Якщо ви вивчаєте завдання під наглядом, хтось присутній судить, чи отримуєте ви правильну відповідь. Подібним чином, під контролем навчання, це означає наявність повного набору маркованих даних під час навчання алгоритму.
- Це означає, що кожен приклад у навчальному наборі даних позначений відповіддю, яку алгоритм повинен придумати самостійно. Отже, модель для маркованого набору даних квіткових зображень покаже картинки, де були троянди, маргаритки і нарциси. Коли показується нове зображення, модель порівнює його із навчальними прикладами, щоб передбачити правильну мітку



Навчання без нагляду

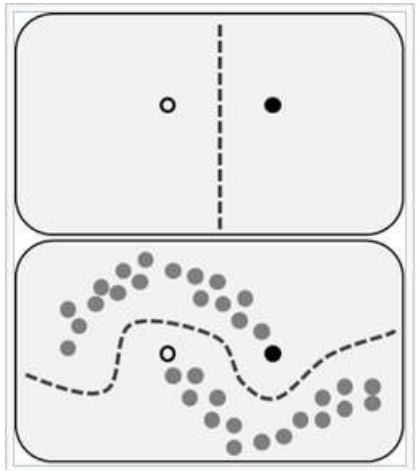
- Залежно від розглянутої проблеми, модель навчання без нагляду може організувати дані по-різному.
- **Кластеризація:** не будучи експертом-орнітологом, можна переглянути колекцію фотографій птахів та розділити їх приблизно за видами, спираючись на такі сигнали, як колір пір’я, розмір або форма дзьоба. Модель глибокого навчання шукає навчальні дані, подібні між собою, і групує їх разом.
- **Виявлення аномалій:** банки виявляють шахрайські транзакції, шукаючи незвичні закономірності в купівельній поведінці клієнта. Наприклад, якщо одна і та ж кредитна картка використовується в двох містах протягом одного дня, це викликає підозру. Подібним чином, неконтрольоване навчання може бути використано для позначення відхилень у наборі даних.
- **Асоціація:** Наповніть інтернет-кошик памперсами, чашками з яблуками та молоком, і веб-сайт просто може порекомендувати додати до вашого замовлення нагрудник та няню. Це приклад асоціації, коли певні ознаки вибірки даних корелюють з іншими ознаками. Переглядаючи пару ключових атрибутів точки даних, некерована модель навчання може передбачити інші атрибути, з якими вони зазвичай пов’язані.

Приклади задач різних типів навчання

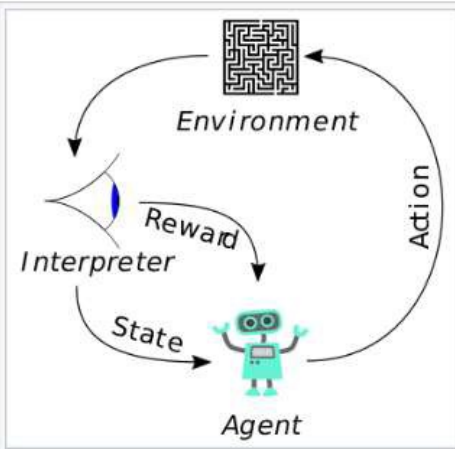
- Класифікація ==> навчання з учителем
- Регресія ==> навчання з учителем
- Кластеризація = = >навчання без учителя
- Пошук аномалій ==> навчання без учителя

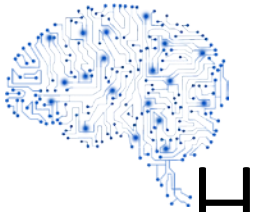
Приклади інших типів навчання

- напівконтрольоване навчання (semi-supervised learning)



- навчання з підкріпленням (reinforcement learning)

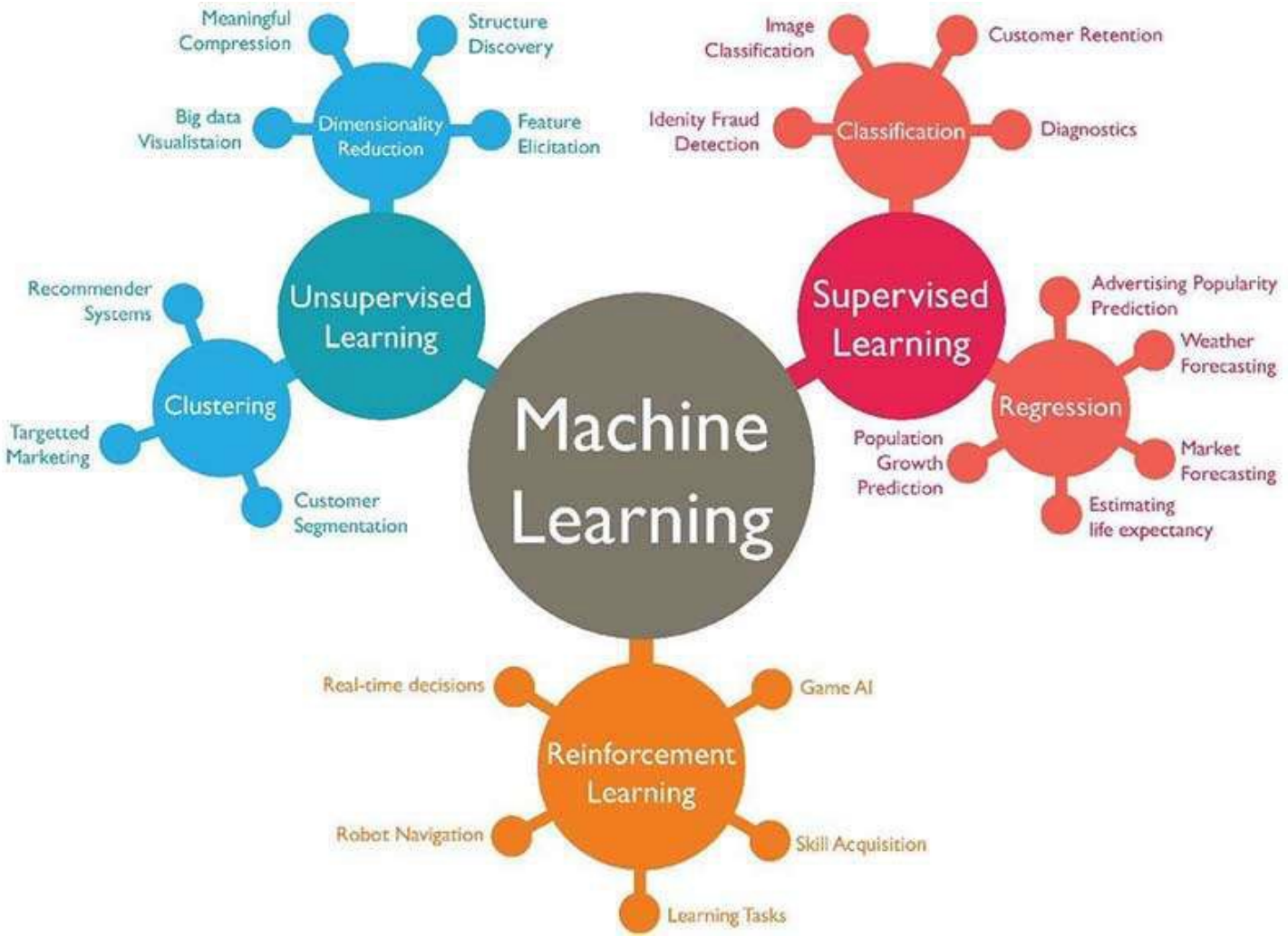




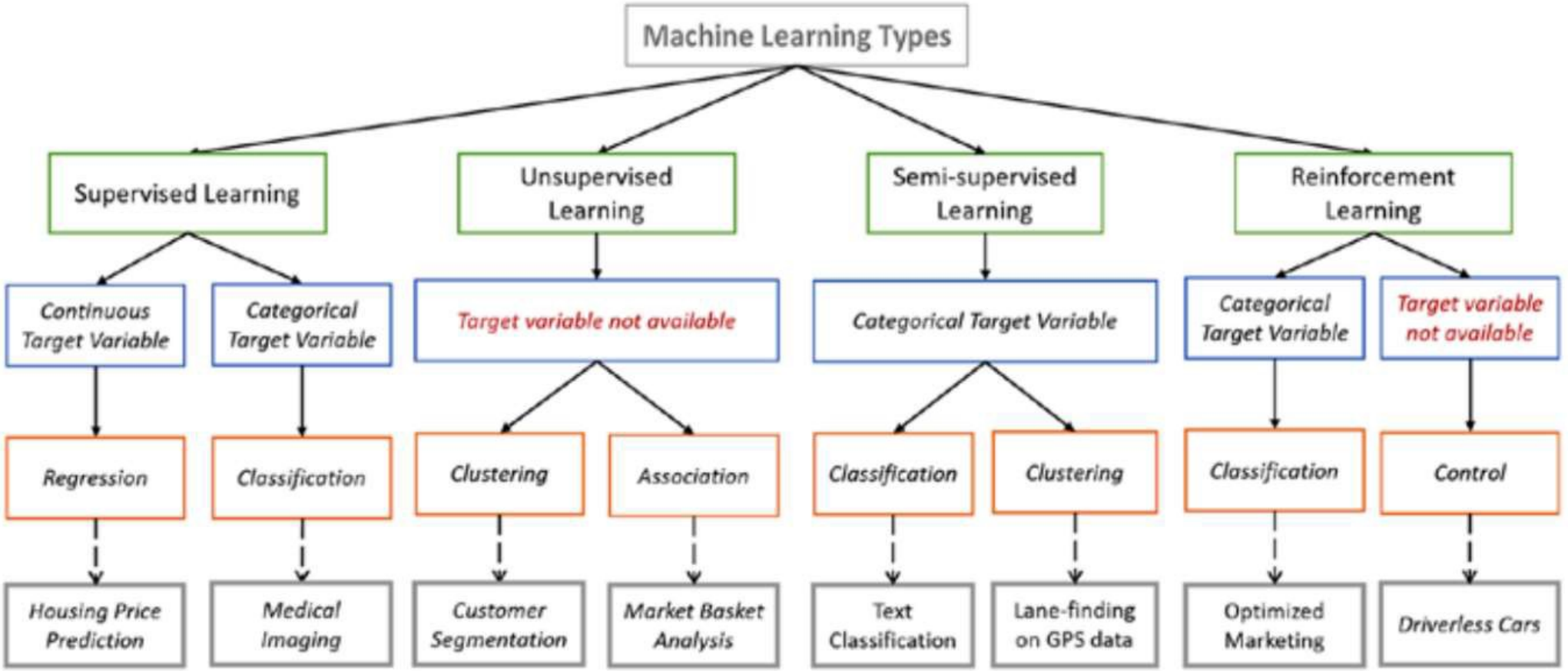
Напівконтрольоване навчання

- Напівконтрольоване навчання - набір навчальних даних містить як марковані, так і немарковані дані. Цей метод особливо корисний, коли приклади маркування - це трудомістке завдання для експертів.
- Навчання під наглядом особливо корисне для медичних зображень, де невелика кількість маркованих даних може призвести до значного підвищення точності.
- Поширеною ситуацією для такого роду навчання є медичні зображення, такі як КТ або МРТ. Кваліфікований рентгенолог може пройти і позначити невелику частину сканувань на наявність пухлин або захворювань. Було б занадто трудомістким і дорогим маркування всіх сканів вручну, але мережа глибокого навчання все одно може отримати вигоду від невеликої частки маркованих даних та підвищити їх точність порівняно з повністю некерованою моделлю.

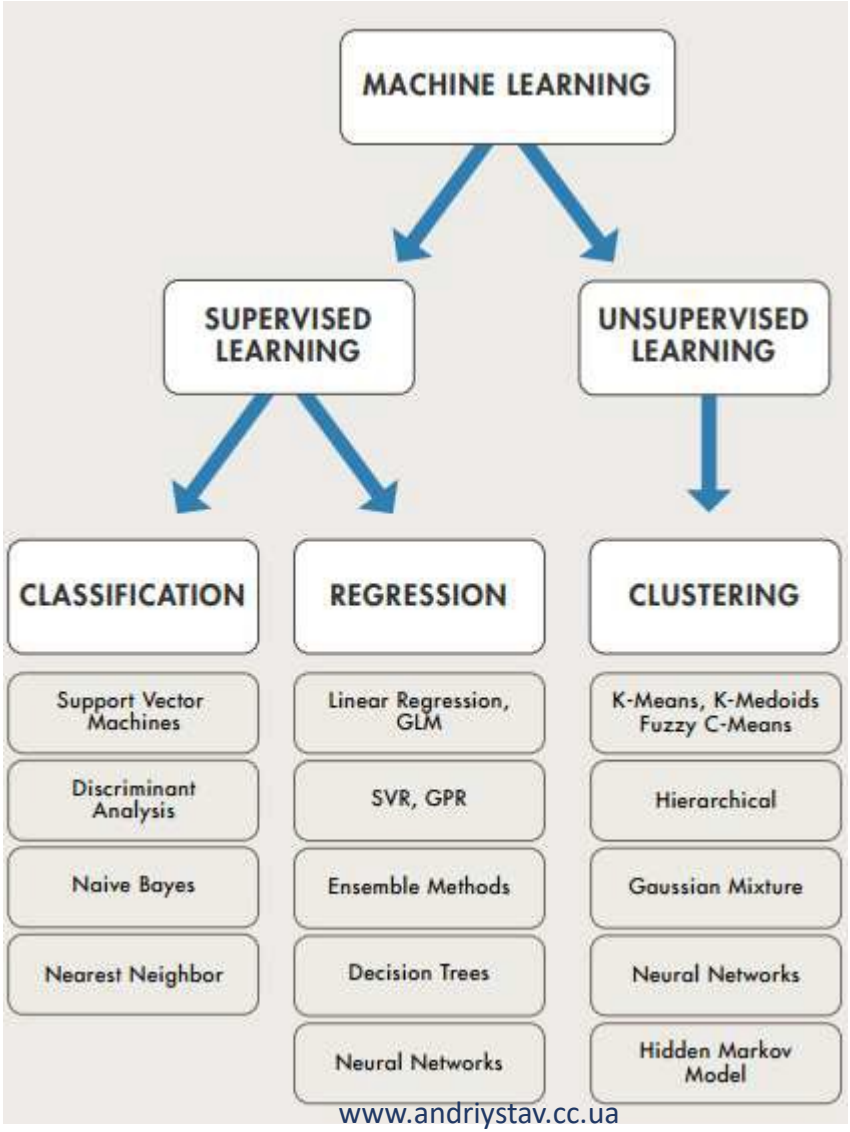
Типи машинного навчання – 1



Типи машинного навчання – 2



Типи машинного навчання – 3



Питання?