

February 16, Kyiv  
**Data Science & Mathematical  
Modeling Bachelor Program**

**Course “Basics of Machine Learning”**  
**Lecture 1: Introduction**



Oleg CHERTOV

Professor, Sc.D. (Doctor Habilitatus),  
Head of the Applied Mathematics Department



Applied Mathematics Department  
Igor Sikorsky Kyiv Polytechnic Institute  
Ukraine





О-о-о, які у нас класні студенти!



[Фото із  
[http://www.hopa.ro/poze/mare/08\\_podborka\\_33\\_1181636029.jpg](http://www.hopa.ro/poze/mare/08_podborka_33_1181636029.jpg)]

# Lecture 1. Introduction (Вступ). Зміст

1

Представлення  
лектора

2

Traditional  
Program-  
ming vs. ML

3

Середина  
2010-х – все  
змінилося.  
І назавжди!

4

Data Science &  
Machine  
Learning  
definition

5

Приклади  
конкретних  
задач  
класифікації  
та регресії

6

Machine  
learning types

# Summary of my CV

- 39 years of research work and 36 years of academic work:

- Textbook “Calculus for programmers, part I” (2005, 2017 II ed.)

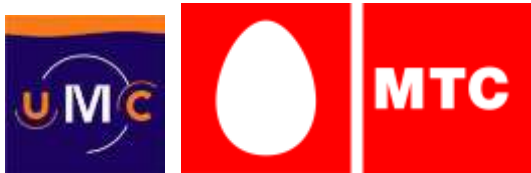
- 6 monographs



...

- > 150 research articles
- > 37 years of experience in real computer business
- **2014-present**      **Head of the App. Math Dep.**
- 1993–2013          Associate Professor
- 1990–1993          Assistant Professor

# Some big projects (with my contribution), > 4.5 bln UAH



THE WORLD BANK



Lecture 1. Introduction

## ❑ 2002-2004, project manager

Automated system “Census-2001” for processing data of the Ukrainian national census:

- ≈ 68 mln paper-based questionnaires
- the biggest OLAP system in Ukraine (at that time)

## ❑ 2005, project manager

Automated system “Moldova Census-2004”:

- ≈ 10 mln paper-based questionnaires

## ❑ October 2006 – September 2007, project manager

Implementation of the prepaid billing system Foris:

\$?? mln

## ❑ 2009-2013, IT-consultant, (World Bank, Ukraine)

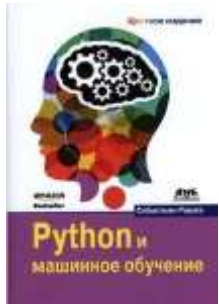
Development of State Statistics System for Monitoring Social & Economic Transformation Project: \$44 mln

## ❑ 2020-2021, Adviser & Business analyst

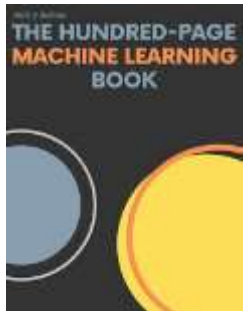
Information communication system “Information exchange” for processing data of American residents in Ukraine according to FATCA (Foreign Account Tax Compliance Act)



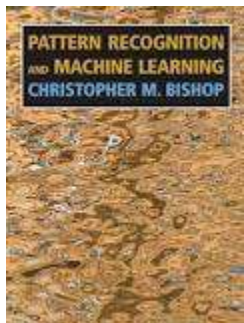
# Рекомендована література



1. Рашка С. Python и машинное обучение (2017)



2. Andriy Burkov. The Hundred-Page Machine Learning Book (2019)



3. Christopher M. Bishop. Pattern Recognition and Machine Learning. Springer (2006)

# Big data is like teenage sex



Photo from  
<https://www.facebook.com/TAUalumni/photos/t.704919/1179943592165007/?type=3&theater>

- Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

From (06.01.2013):

[https://www.facebook.com/dan.ariely/posts/904383595868?imm\\_mid=0a9701&cmp=em-strata-newsletters-strata-olc-20130529-elist](https://www.facebook.com/dan.ariely/posts/904383595868?imm_mid=0a9701&cmp=em-strata-newsletters-strata-olc-20130529-elist)

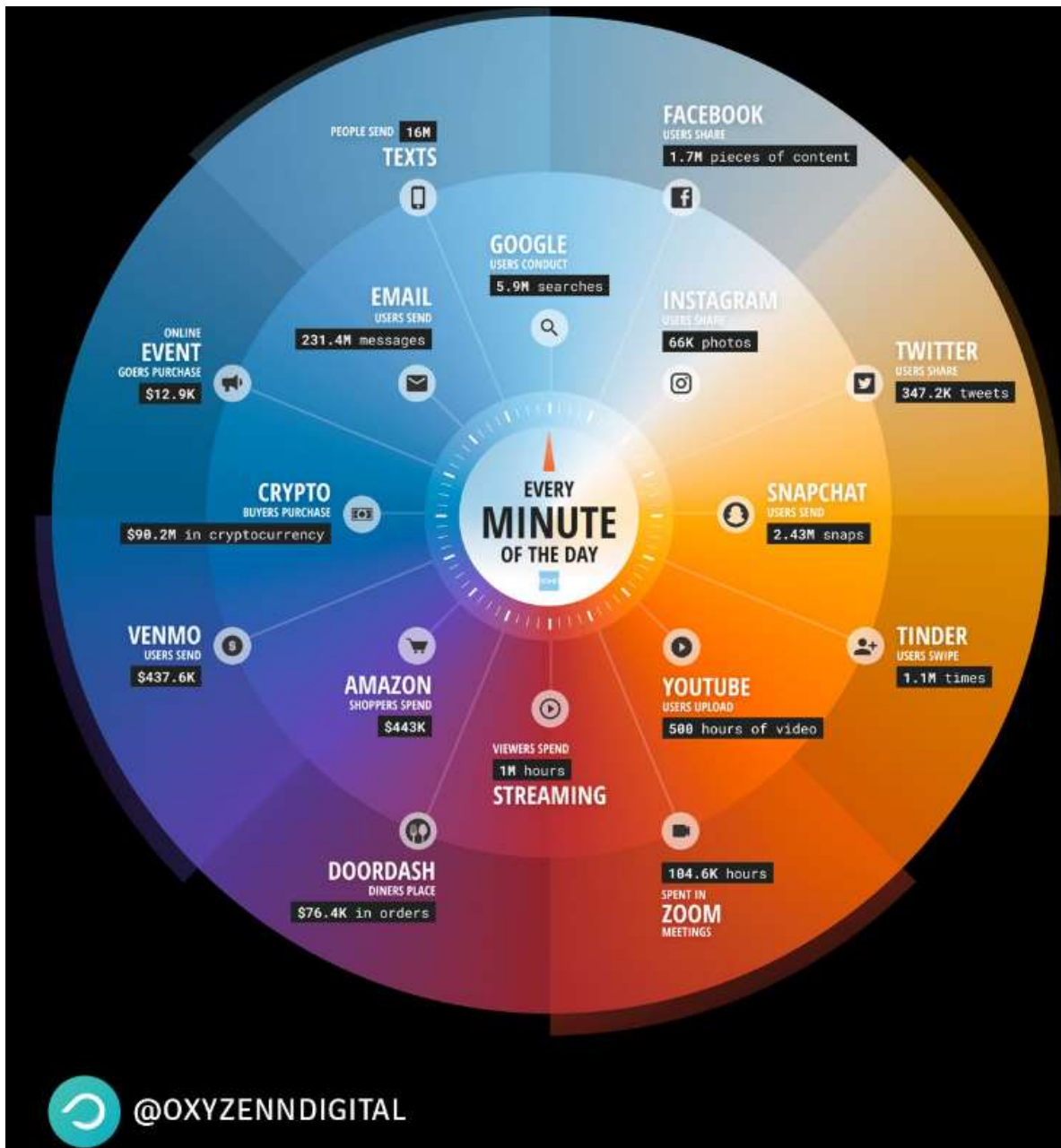
- **Dan Ariely** (29.04.1967-) is an Israeli-American professor and author
- He serves as a James B. Duke **Professor of psychology and behavioral economics** at Duke University
- Ariely's TED talks have been viewed over 15 million times



# 2022: This is what happens in an Internet minute

## How much data are generated every minute?

The incredible scale of e-commerce, social media, email, and other content creation that happens on the web



# This is what happens in an Internet minute

## 2021 This Is What Happens In An Internet Minute



eDiscovery Today  
@TodayEdiscovery

Hey @lorilewis, when are you going to publish the Internet Minute for 2022?

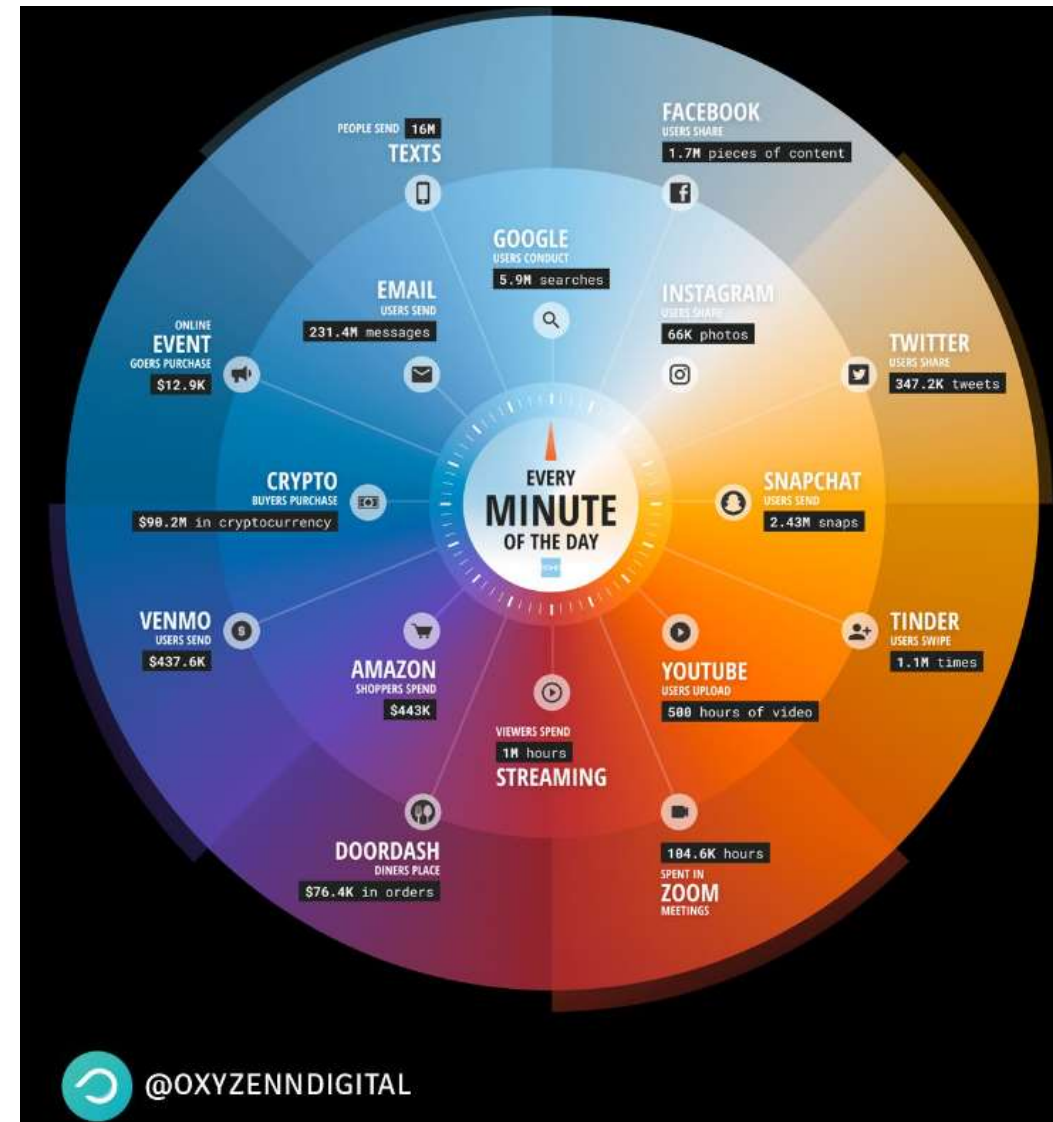
5:03 PM · May 11, 2022



Lori Lewis ✓ @lorilewis · May 11, 2022  
Replying to @TodayEdiscovery  
I will  
I'll start collecting data now.



# This is what happens in an Internet minute

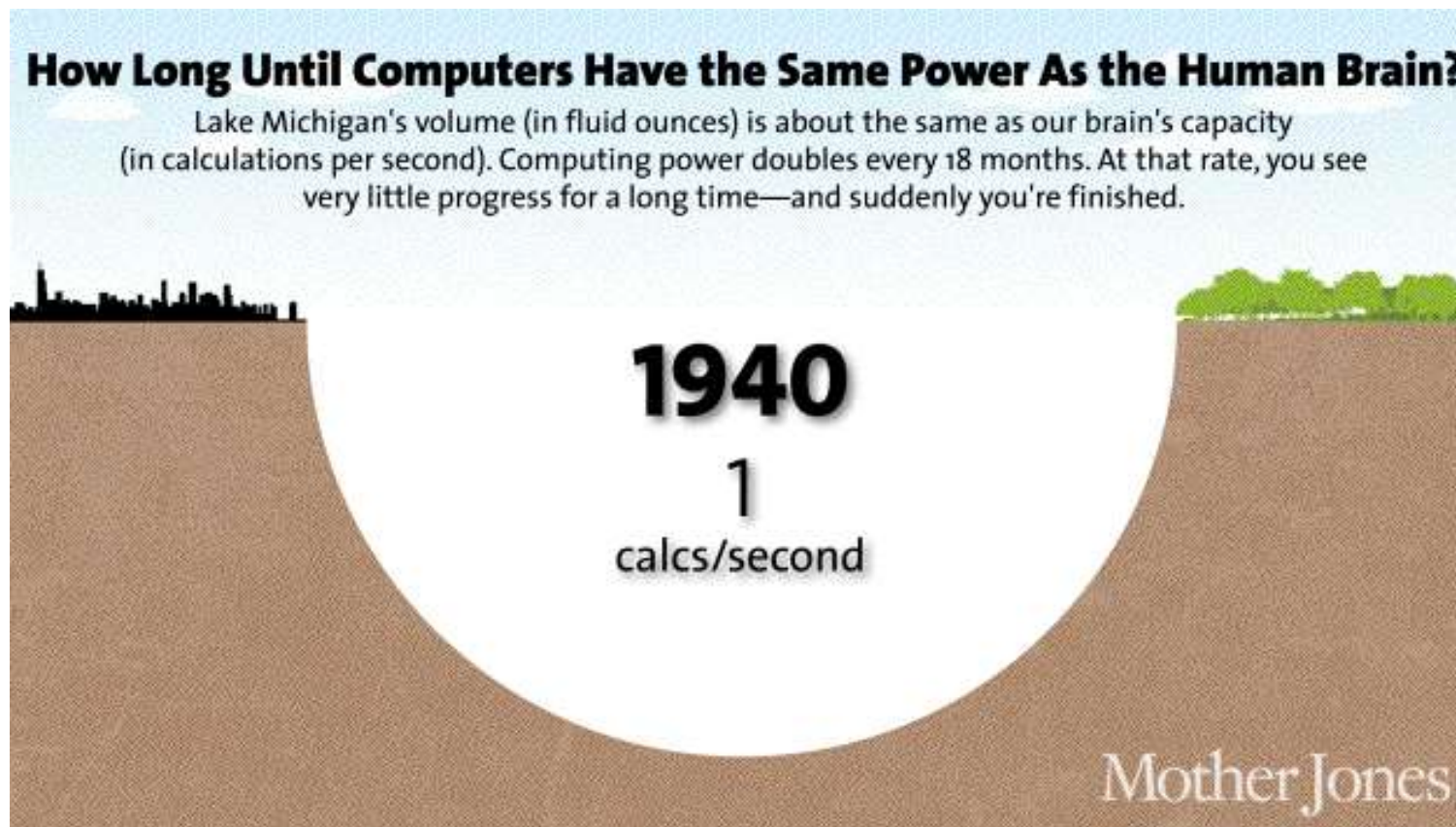


# А яке «залізо» потрібно? (оцінка 2015 р.)



- One way to express this capacity is in the total calculations per second (cps) the brain could manage, and you could come to this number by figuring out the maximum cps of each structure in the brain and then adding them all together
  - Raymond Kurzweil (Рэймонд Курцвейл) came up with a shortcut by taking someone's professional estimate for the cps of one structure and that structure's weight compared to that of the whole brain and then multiplying proportionally to get an estimate for the total
- Sounds a little iffy, but he did this a bunch of times with various professional estimates of different regions, and the total always arrived in the same ballpark—around  $10^{16}$ , or 10 quadrillion cps

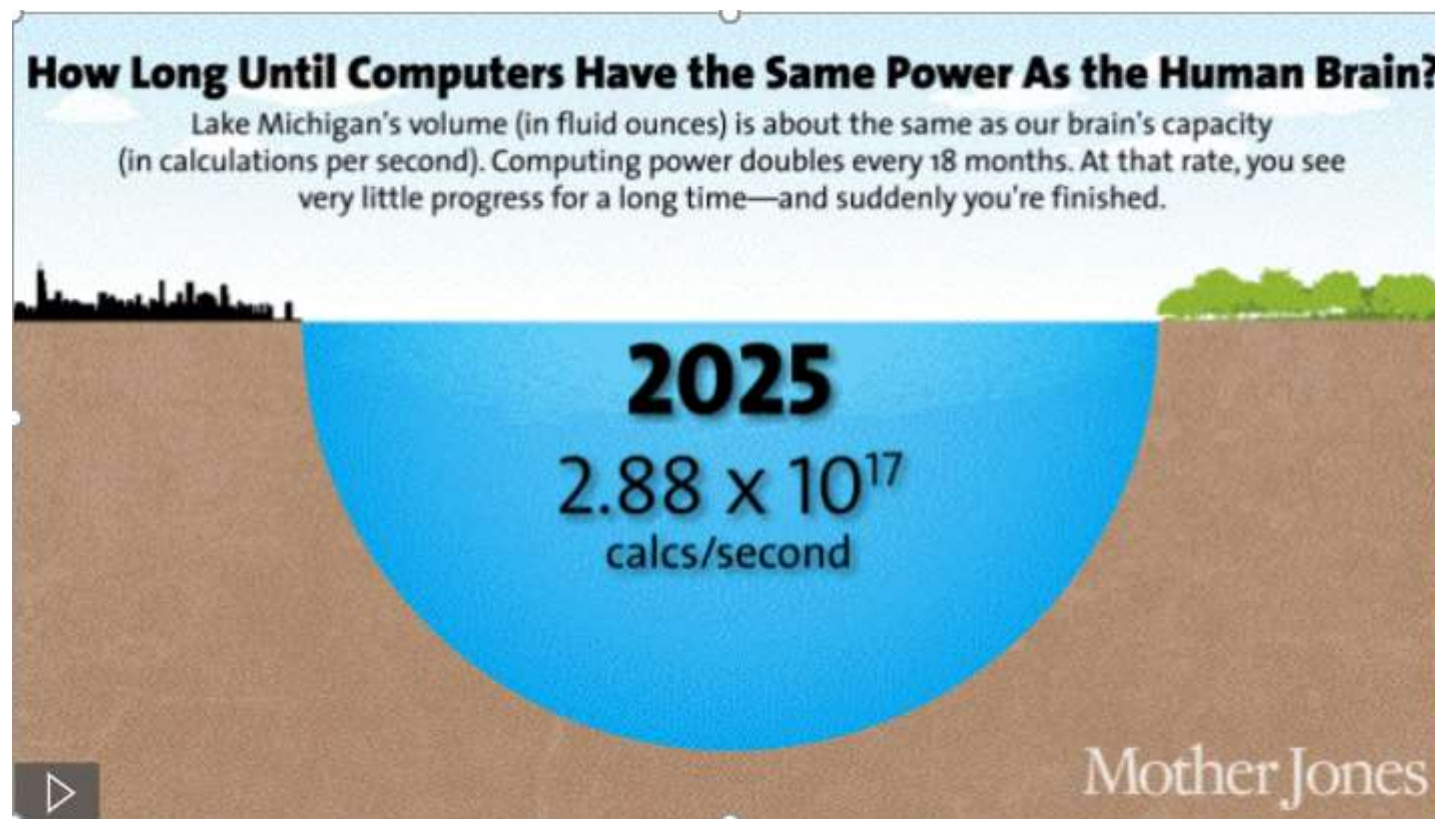
# А коли стане достатньо?



<https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html>



# А коли стане достатньо?



<https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html>



# Чи є у нас достатні обчислювальні потужності?



- It currently has achieved 1.102 Exaflop/s (=1102 quadrillion floating point operations per second) using 8,730,112 cores. The new HPE Cray EX architecture combines 3rd Gen AMD EPYC™ CPUs optimized for HPC and AI with AMD Instinct™ 250X accelerators and Slingshot-11 interconnect

□ **Frontier** [frɒn'tɪə] is the new No. 1 system in the TOP500. This HPE Cray EX system is the first US system with a peak performance exceeding one ExaFlop/s. It is currently being integrated and tested at the **Oak Ridge National Laboratory in Tennessee, USA**, where it will be operated by the Department of Energy (DOE).

<https://www.top500.org/lists/top500/2022/11/>

# Чи є у нас достатні обчислювальні потужності?



- Info about #5 (**Summit**, an IBM-built system at the Oak Ridge National Laboratory in Tennessee, USA)
- At over 340 tons, Summit's cabinets, file system, and overhead infrastructure **weigh more than a large commercial aircraft**
- Occupying 5,600 sq. ft. of floor space, Summit **could fill two tennis courts**

<https://www.olcf.ornl.gov/2018/06/08/summit-by-the-numbers/>

Photo from <https://www.timesfreepress.com/news/opinion/times/story/2018/jun/11/pams-points-lets-elect-summit-supercomputer-p/472848/>

- More than 4,000 gallons of water pump through Summit's cooling system every minute, carrying away about 10 megawatts of heat
- Just 20 watts - **the human brain's energy requirement**

(<https://www.popsoci.com/technology/article/2009-11/neuron-computer-chips-could-overcome-power-limitations-digital>)

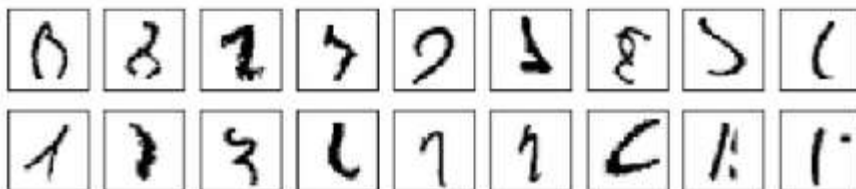
# Мозок – наше все, але ... Big Data

- 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 (Alexa від Amazon, Google Assistant від Google, Siri від Apple, Cortana від Microsoft)
- Надання рекомендацій відносно товару, що найкраще підходить (Amazon, Netflix, ...)
- Фільтрація спаму у вхідних повідомленнях електронної пошти
- Виявлення та діагностування внутрішніх захворювань

# How well does the program recognize handwritten digits?



- **MNIST data set** is based on two data sets collected by the United States' National Institute of Standards and Technology (NIST)
- 50,000 image data set + 10,000 image validation set



**Michael Nielsen.**  
**Neural Networks**  
**and Deep**  
**Learning**

# 73 rows of Python code reached 9,659 out of 10,000 (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))
```



# Traditional Programming vs. ML

- Jeff Bezos, Amazon CEO, у своєму листі інвесторам:
- «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»
- In his latest letter to Amazon shareholders: «Over the past decades, computers have broadly automated tasks that programmers could describe with clear rules and algorithms. Modern machine learning techniques now allow us to do the same for tasks where describing the precise rules is much harder»

<https://www.businessinsider.com/read-amazon-ceo-jeff-bezos-2016-letter-to-shareholders-2017-4>



**Jeff Bezos. Mario  
Tama/Getty Images**

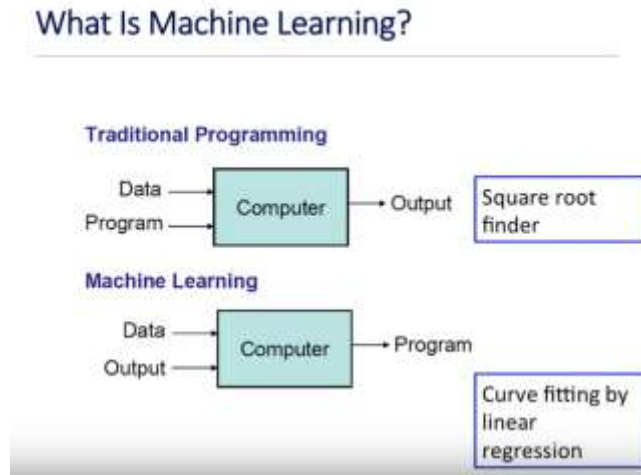
# Штучний інтелект небезпечніший від ядерної зброї



**Yuval Noah Harari** (photo: REUTERS / Denis Balibouse)  
Professor, Department of History, The Hebrew University of Jerusalem  
Historian, philosopher and the author of the bestsellers “Sapiens: A Brief History of Humankind”, “Homo Deus: A Brief History of Tomorrow”, and “21 Lessons for the 21st Century”  
*World Economic Forum, Davos, January 21, 2020,*  
<https://www.weforum.org/events/world-economic-forum-annual-meeting-2020/sessions/realizing-chinas-technology-potential>

- Проблема зі штучним інтелектом в порівнянні з ядерною зброєю полягає в тому, що небезпека не така очевидна і деякі гравці бачать величезну вигоду від використання ШІ
- Що стосується атомних бомб, то всі знали, що їх застосування призведе до кінця світу. Неможливо виграти повномасштабну ядерну війну. У випадку з ШІ багато хто думає - і на це є вагомі причини - що гонку озброєнь в сфері ШІ можна виграти
- І це дуже небезпечно, тому що в такому випадку набагато більше спокуси виграти цю гонку, щоб почати домінувати над світом

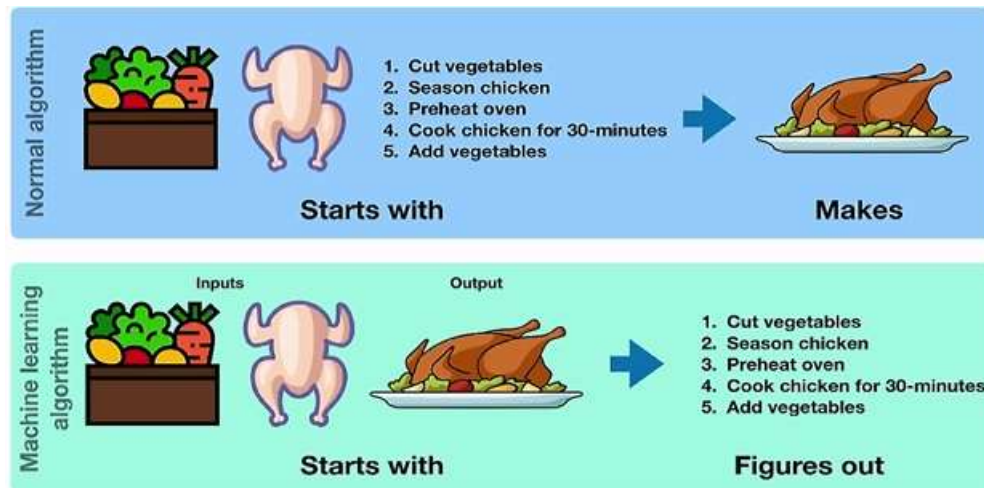
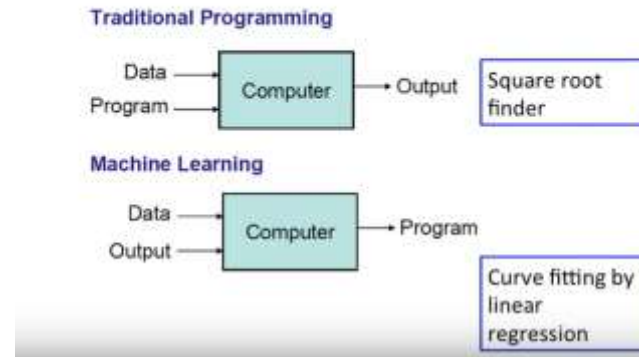
# Traditional Programming vs. ML



- **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 придумав 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 крутіший варіант машинного навчання – без учителя!

# Traditional Programming vs. ML

## What Is Machine Learning?

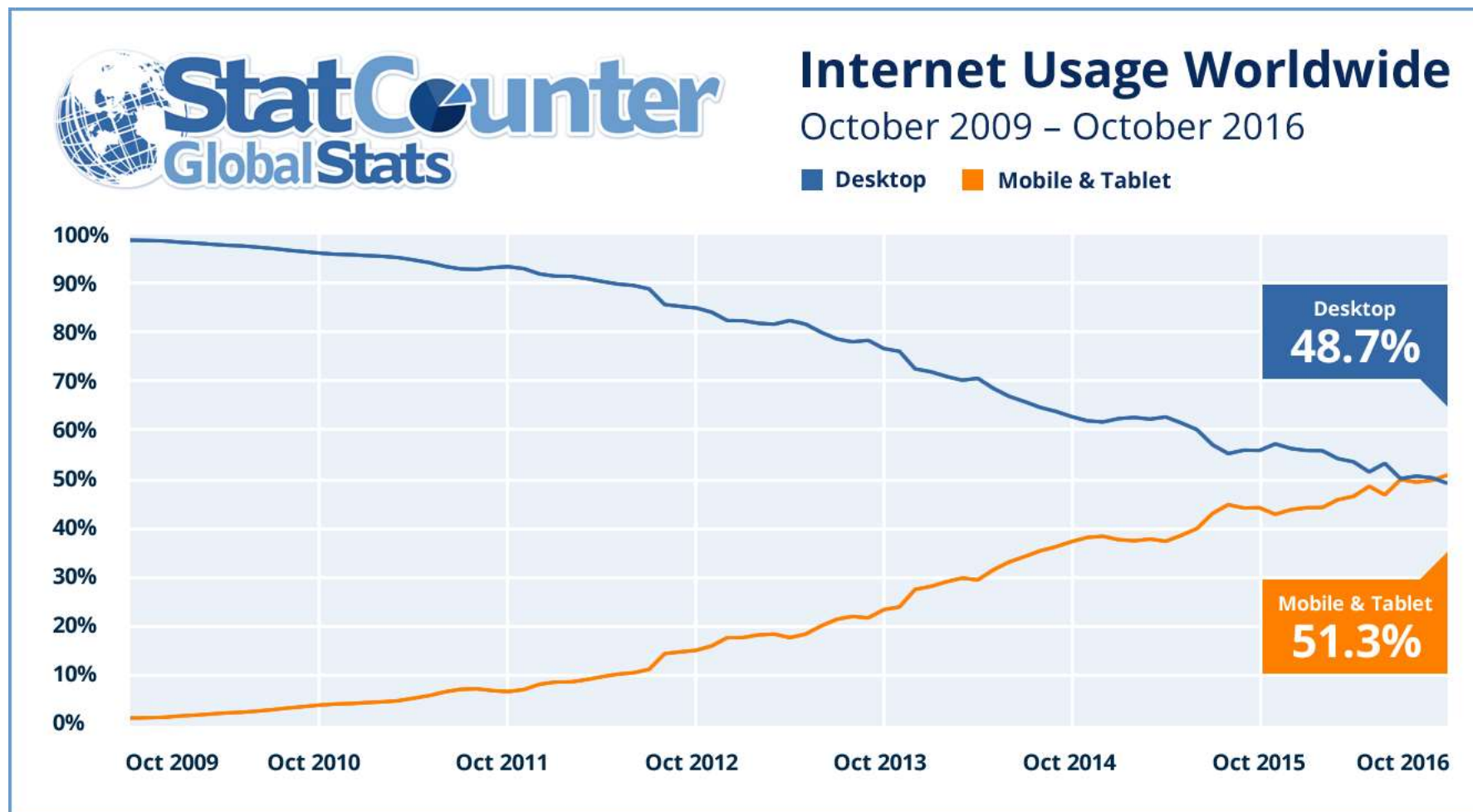


<https://virgool.io/@hamidreza.firooze/%D9%85%D9%87%D9%85%D8%AA%D8%B1%DB%8C%D9%86-%D8%A7%D9%84%DA%AF%D9%88%D8%B1%DB%8C%D8%AA%D9%85-%D9%87%D8%A7%DB%8C-%DA%A9%D8%A7%D8%B1%D8%A8%D8%B1%D8%AF%DB%8C-%DB%8C%D8%A7%D8%AF%DA%AF%DB%8C%D8%B1%DB%8C-%D9%85%D8%A7%D8%B4%DB%8C%D9%86-%DA%86%DB%8C%D8%B3%D8%AA-cxhxyt8lev4n>

Середині 2010-х — все змінилося.  
**І назавжди!**

Невизначеність — це не баг (bug),  
а «фіча» (feature), особливість

# Жовтень 2016 р.: користувачів Інтернет з моб. пристроями > ніж з настільними



Mobile  
first!



# Alan Turing. Father of the Modern Computer



Alan Turing (1912-1954)

Source: Beryl Turing and King's College Library, Cambridge

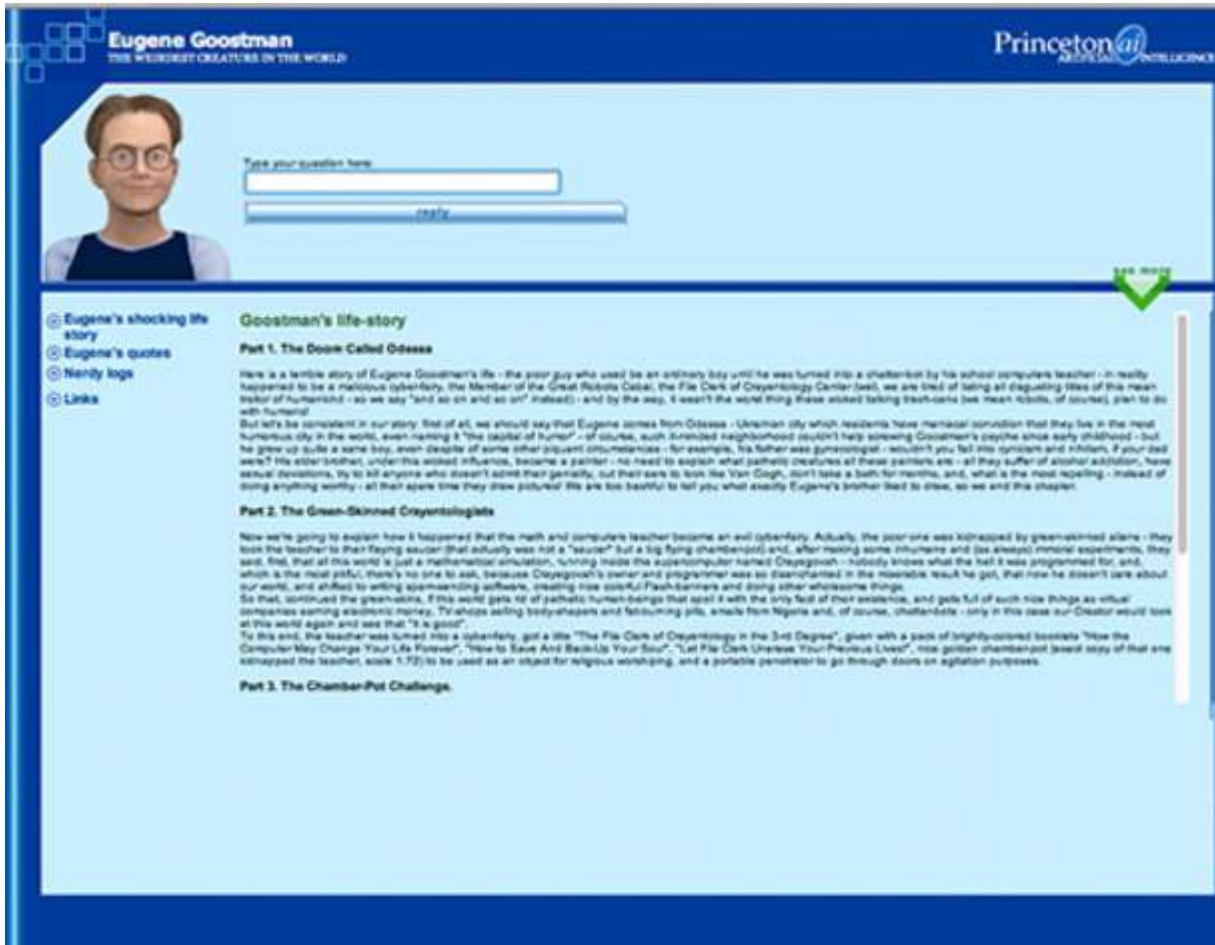
- The earliest large-scale electronic digital computers, the British Colossus (1944) and the American ENIAC (1945), did not store programs in memory.
- To set up these computers for a fresh task, it was necessary to modify some of the machine's wiring, re-routing cables by hand and setting switches.
- The basic principle of the modern computer—the idea of controlling the machine's operations by means of a program of coded instructions stored in the computer's memory—was conceived by Alan Turing.

From:

B. Jack Copeland, Diane Proudfoot. Alan Turing. Father of the Modern Computer (web-book)

<http://www.rutherfordjournal.org/article040101.html#sdfootnotei3anc>

# Чатбот «Євген Густман» - єврейський хлопчик із Одеси?



- Turing, Alan (October 1950), "Computing Machinery and Intelligence", Mind, LIX (236): 433–460,

<https://academic.oup.com/mind/article-pdf/LIX/236/433/9866119/433.pdf>

- This paper was the first to introduce his concept of what is now known as the Turing test to the general public
- Тест Тюрінга пройдено в 2014 р.!
- Не стільки рівень комп'ютерних програм, виріс, скільки рівень людей знизився ☺

# ImageNet



**Leopard**

- More than 14 million images have been hand-annotated by the project to indicate what objects are pictured and in at least one million of the images, bounding boxes are also provided
- ImageNet contains more than 20,000 categories with a typical category, consisting of several hundred images

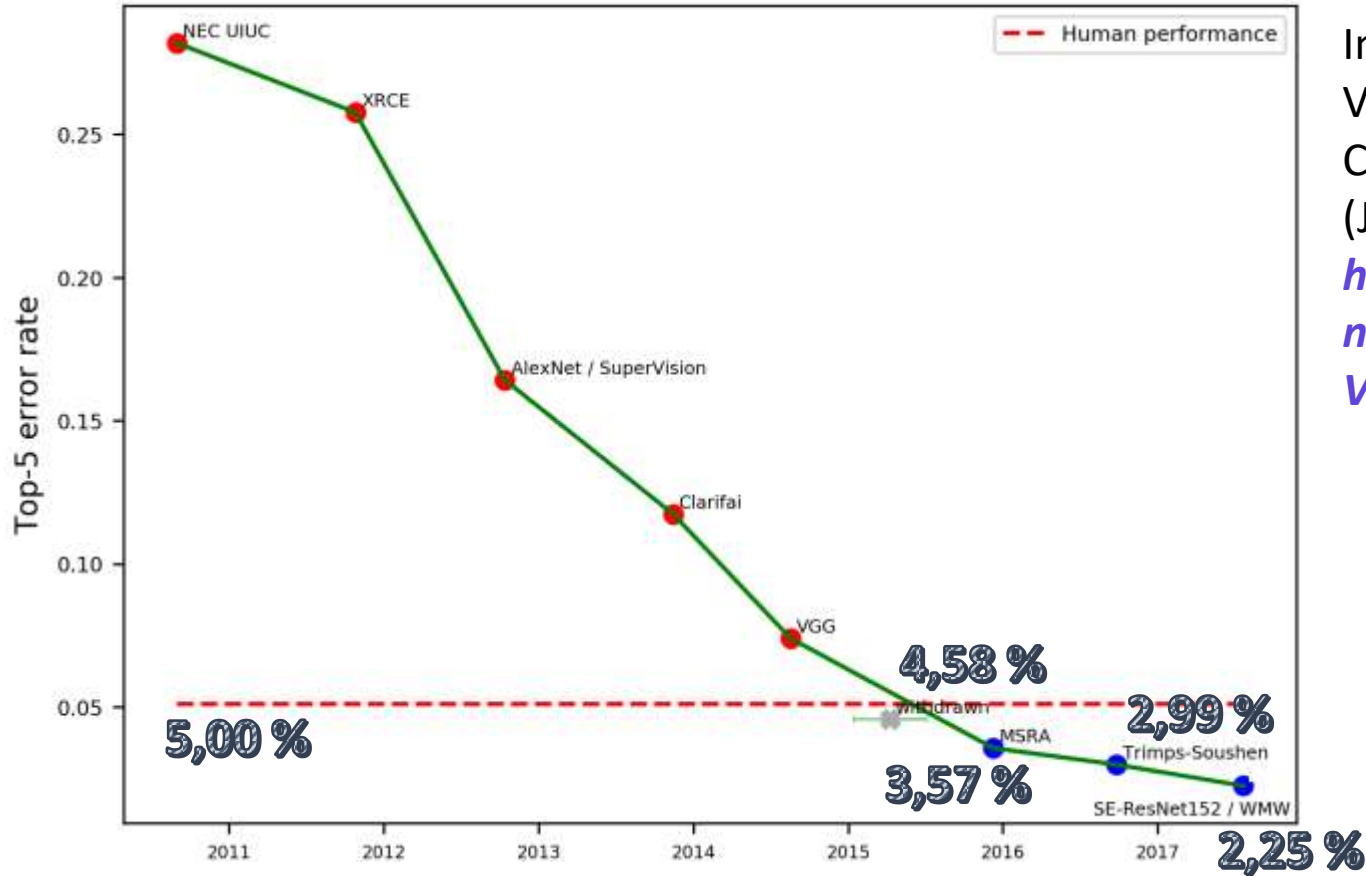
# ImageNet



**Dalmatian?**  
**Cherry?**

- More than 14 million images have been hand-annotated by the project to indicate what objects are pictured and in at least one million of the images, bounding boxes are also provided
- ImageNet contains more than 20,000 categories with a typical category, consisting of several hundred images

# Image Recognition



<https://www.eff.org/ai/metrics>

ImageNet Large Scale Visual Recognition Competition 2017: (July 2017)  
<http://image-net.org/challenges/LSVRC/2017/results>

WMW  
*Jie Hu (Momenta)*  
*Li Shen*  
(University of Oxford)  
*Gang Sun*  
(Momenta)

**Momenta** was founded in September 2016

- In 2017, 29 of 38 competing teams had greater than 95% accuracy.
- In 2017 ImageNet stated it would roll out a new, much more difficult, challenge in 2018 that involves classifying 3D objects using natural language

# Speech recognition. Switchboard corpus

Data were drawn from the Switchboard corpus of informal telephone conversations on prescribed topics:

- 40,515 words
- 4,583 sentences
- 30 speakers

*Godfrey, John, and Edward Holliman. Switchboard-1 Release 2 LDC97S62. Web Download. Philadelphia: Linguistic Data Consortium, 1993.*

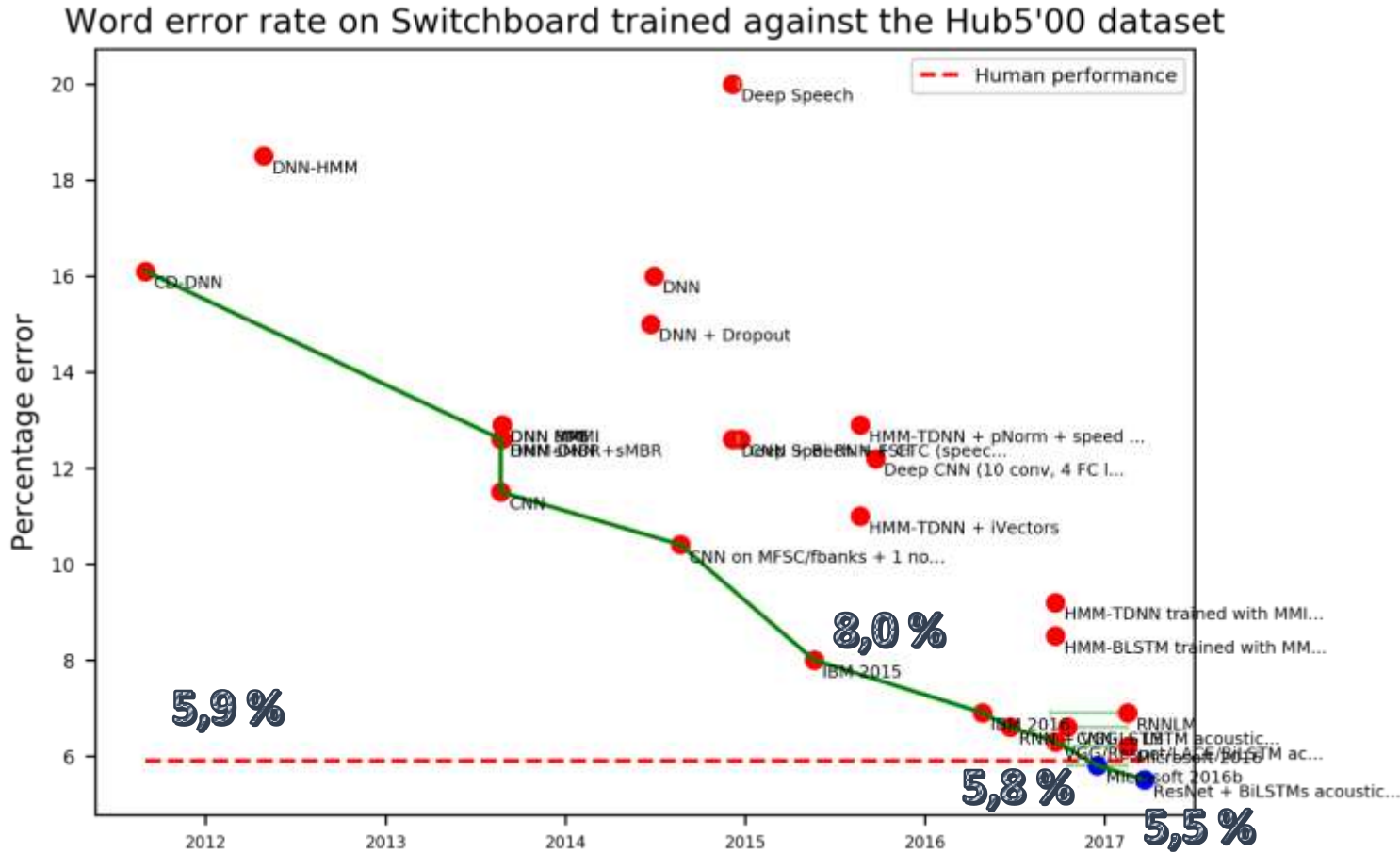
Audio sample from

<https://catalog.ldc.upenn.edu/LDC97S62>





# Speech recognition



<https://www.eff.org/ai/metrics>

W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu and G. Zweig. Achieving Human Parity in Conversational Speech Recognition (Microsoft Research, 17.12.2016)

<https://arxiv.org/pdf/1610.05256.pdf>


George Saon et al. English  
Conversational Telephone  
Speech Recognition by  
Humans and Machines  
(06.03.2017)

<https://arxiv.org/abs/1703.02136>

# Data scientist: the sexiest job of the 21st century

- Data scientist – the sexiest job of the 21st century

Thomas Davenport & DJ Patil, Harvard Business Review (2012)



**50 Best Jobs in America**

This report ranks jobs according to each job's Glassdoor Job Score, determined by combining three factors: number of job openings, salary, and overall job satisfaction rating.

United States | 2018

0 Shares | Facebook | Twitter | LinkedIn | Print

**1 Data Scientist**



**4.8 / 5**  
Job Score

**\$110,000**  
Median Base Salary

**4.2 / 5**  
Job Satisfaction

**4,524**  
Job Openings

[View Jobs](#)

[https://www.glassdoor.com/List/Best-Jobs-in-America-LST\\_KQ0,20.htm](https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm)

- 
- Data scientist – the best job of 2016-2019 in USA
  - 2020 & 2022 – the 3<sup>rd</sup> place
  - 2021 – the 2<sup>nd</sup> place

This report ranks jobs according to each job's Glassdoor Job Score, determined by combining three factors: number of job openings, salary, and overall job satisfaction rating

# AI investment increased during the pandemic, and many business plan to do more, Gartner found

STAMFORD, Conn., October 1, 2020

## Gartner Survey Reveals 66% of Organizations Increased or Did Not Change AI Investments Since the Onset of COVID-19

Cost Optimization, Customer Experience, and Revenue Growth are Top Focus Areas for AI Initiatives

A Gartner, Inc. poll of roughly 200 business and IT professionals on September 24, 2020 revealed that 24% of respondents' organizations increased their artificial intelligence (AI) investments and 42% kept them unchanged since the onset of [COVID-19](#). Growth – namely customer experience and retention, and revenue growth – along with cost optimization were the top focus areas for their current AI initiatives.

Over the course of the next six-nine months, 75% of respondents will continue or start new AI initiatives as they move into the Renew phase of their organization's post-pandemic [Reset](#).

### Contacts

**Katie Costello**  
Gartner  
[katie.costello@gartner.com](mailto:katie.costello@gartner.com)

**Meghan Rimol**  
Gartner  
[meghan.rimol@gartner.com](mailto:meghan.rimol@gartner.com)

### Share this:

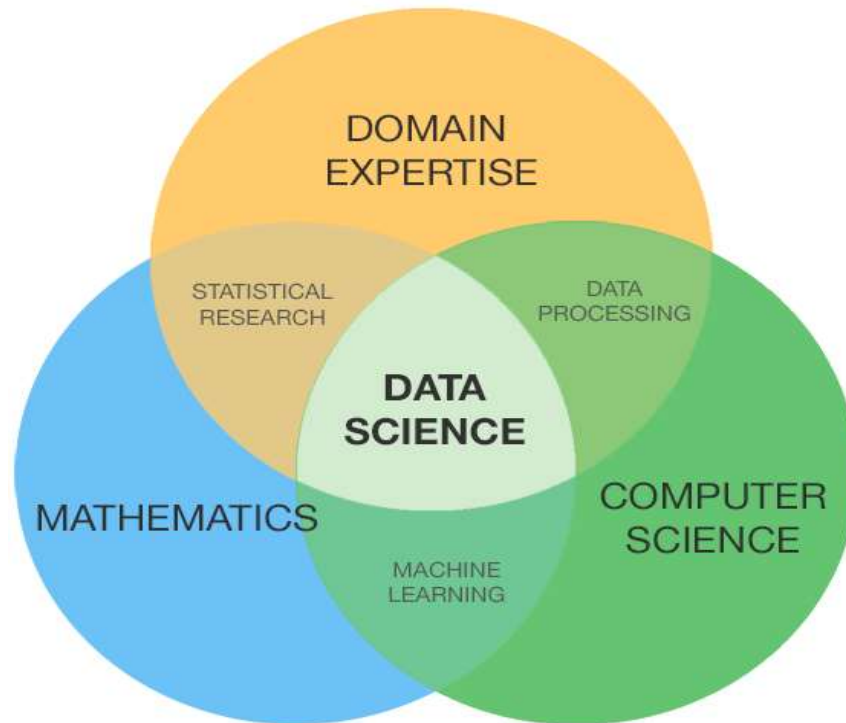


[Newsroom](#)

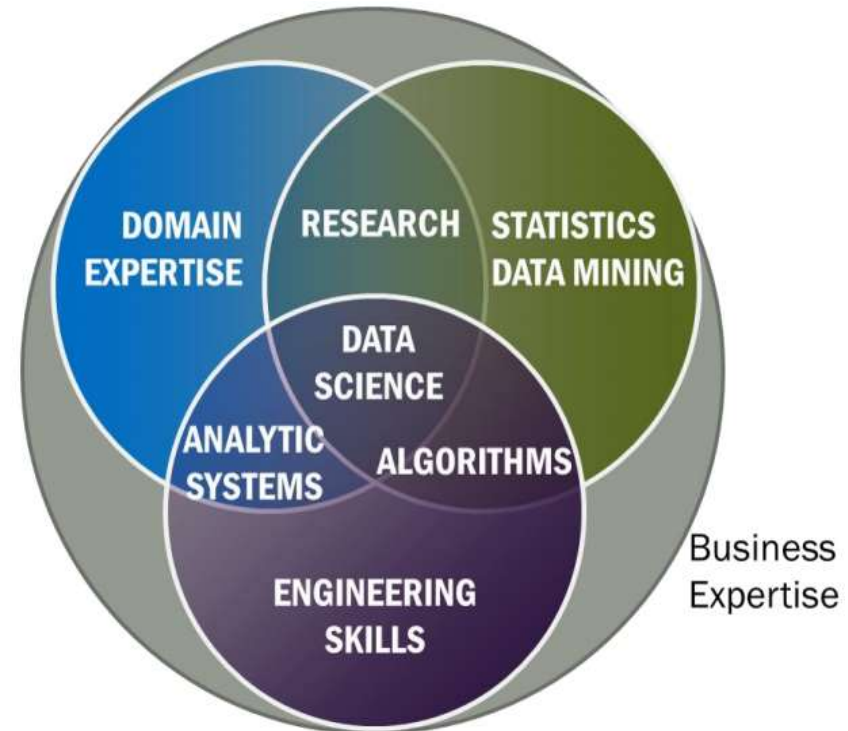
- A quarter of IT professionals increased AI investment levels due to COVID-19, 42% kept it at the same level, but 75% plan to continue or begin new AI projects in the next 6-9 months

# Data Science definition (in figures)

Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015



NIST SP 1500-1 NIST Big Data interoperability Framework: Volume 1: Definitions, September 2015

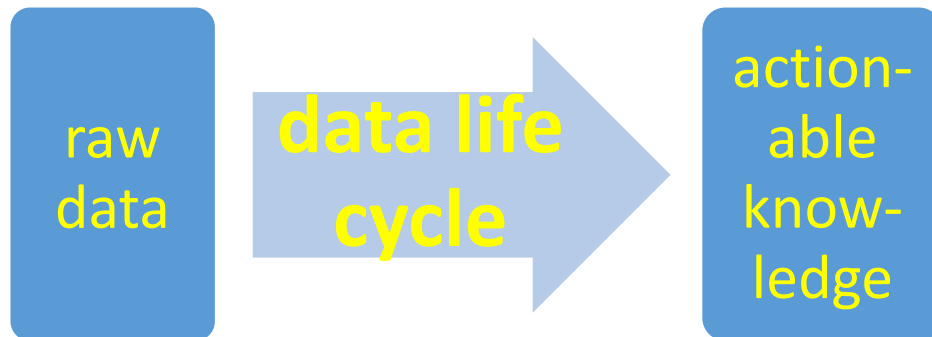


# Data Science definition (in text)

*from NIST Big Data Working Group, 2015*

## Data science

- Extraction of actionable knowledge directly from data through a process of discovery, or hypothesis formulation and hypothesis testing



## Data scientist

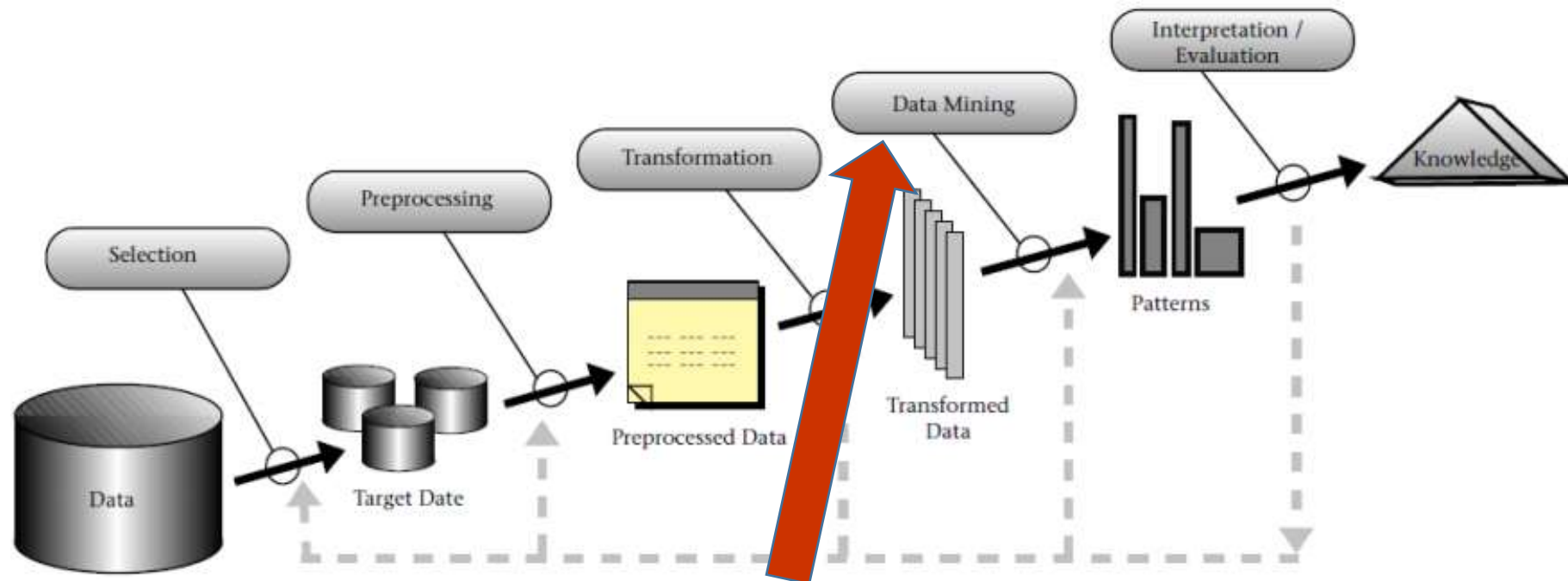
- Practitioner who has sufficient knowledge in the overlapping regimes of business needs, domain knowledge, analytical skills, and software and systems engineering to manage the end-to-end data processes in the data life cycle



# Knowledge discovery in Databases

*from Fayyad, Piatetsky-Shapiro, and Smyth, 1996*

**Knowledge discovery in Databases** is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data



**Data mining** is the process of discovering interesting patterns and knowledge from large amounts of data

*from Han et al., Data Mining. Concepts and Techniques, 3rd Ed., 2012*

# Paulo Villegas: DM vs ML vs AI vs DS vs ...

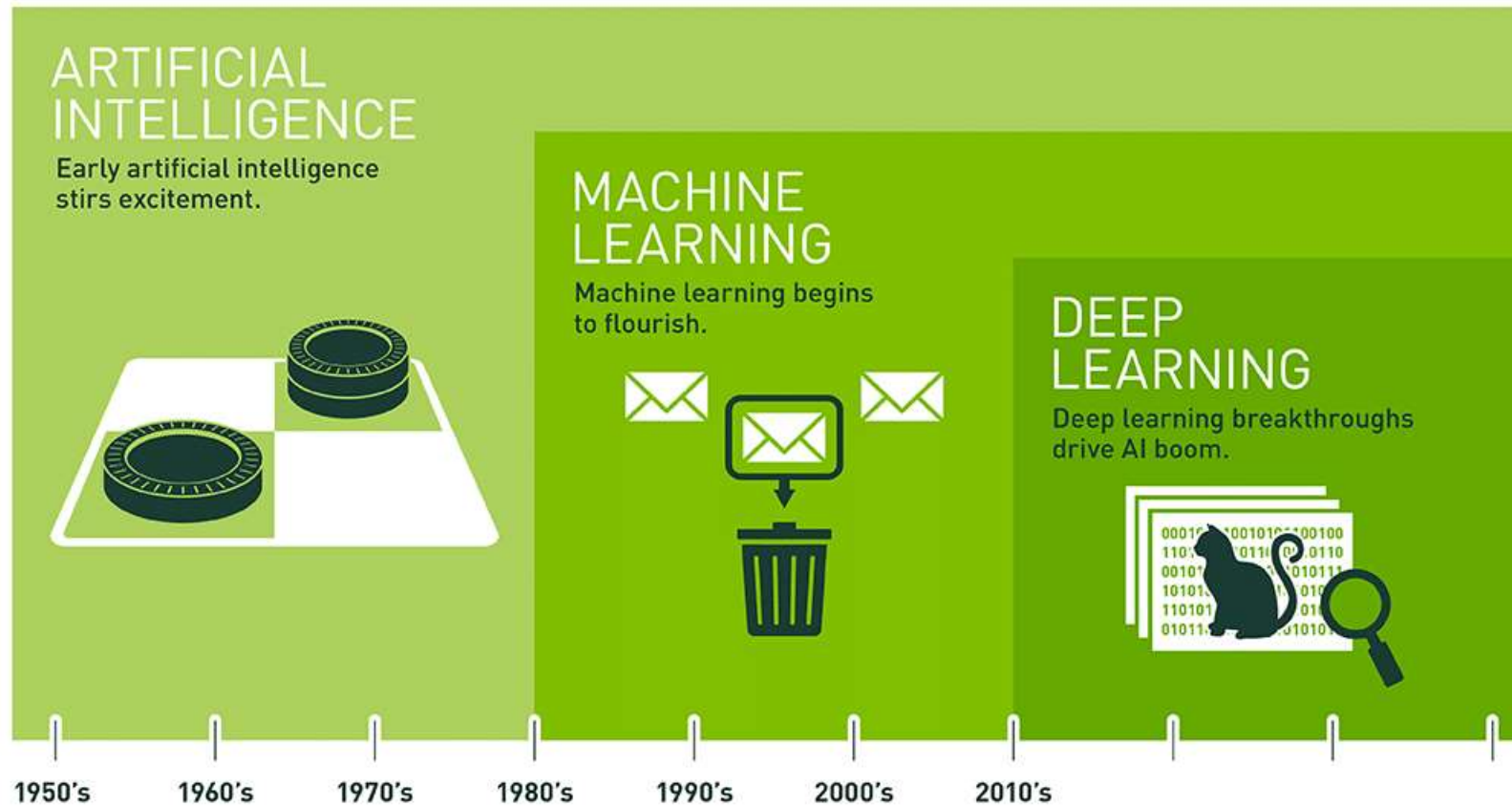
- Data Analysis, Data Mining, Machine Learning and Mathematical Modeling are **tools**: means towards an end
- Analytics, BI, Econometrics and AI are **application areas**: domains that use the tools above (and others) to produce results within its subject
- Statistics is a **branch** of Mathematics providing theoretical and practical support to the above tools
- Data Science describes using those all tools to provide answers in those all areas (and in others), specially when dealing with Big Data, which is nothing more than a label meaning doing any of the above but when the datasets are huge

# What's the Difference Between Artificial Intelligence, and Machine Learning?

*If it's written in Python, it's machine learning;  
if it's written in PowerPoint, it's AI*

Joke from <https://mc.ai/what-is-machine-learning%E2%80%8A-%E2%80%8AA-visual-explanation/>

# What's the Difference Between Artificial Intelligence, Machine Learning and Deep Learning?



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Fig. from <https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/>

# Our kids are using AI technology every day. It's time to learn about it



- Machine Learning

- Video from <https://www.youtube.com/watch?v=vT5QDtX4mp8>

“Дамир учится самостоятельно сползать с кровати в 10 месяцев”



Our kids are using AI technology every day.  
It's time to learn about it



- Data Mining

- Video from <https://www.dailymotion.com/video/x7vomke>

“My Son The Genius”



# Data Science competence groups

- **Data Analytics**
- Use appropriate data analytics and statistical techniques on available data to discover new relations and deliver insights into research problem or organizational processes and support decision-making
- **Data Science Engineering**
- Use engineering principles and modern computer technologies to research, design, implement new data analytics applications; develop experiments, processes, instruments, systems, infrastructures to support data handling during the whole data lifecycle
- **Data Management**
- Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing
- **Research Methods and Project Management**
- Create new understandings and capabilities by using the scientific method (hypothesis, test/artefact, evaluation) or similar engineering methods to discover new approaches to create new knowledge and achieve research or organisational goals
- **Domain related Competences: Applied to BA**
- Use domain knowledge to develop relevant data analytics applications; adopt general DS methods to domain specific data types and presentations, data and process models, organisational roles and relations

# Data Analysis versus Mathematical Statistics

| Difference  | Mathematical Statistics  | Data Analysis  |
|---|--|--|
| <b>Obvious</b> (different amounts of data being processed)  | not Big Data   | Big Data   |
| <b>Unobvious</b> (different types of tasks that are fundamental)<br><br>Examples from:<br>Миркин Б. Г.<br>Введение в анализ данных : учебник и практикум для бакалавриата и магистратуры. — М. : Издательство Юрайт, 2019. — 174 с. | <p>tries to deduce the properties of the surrounding world based on specially collected data.</p> <p>Example: methods for testing statistical hypotheses play a very important role when, an agronomist wants to understand which variety of seeds, other things being equal, will bring the best yield, or a doctor is trying to determine whether a new treatment method gives a noticeably better result than existing technique.</p> <p>To answer such questions, one must carefully set up an experiment, obtain comparable data and accurately compare the results, taking into account their random spread.</p> | <p>is focused on finding any patterns, regularities, structure in the available data.</p> <p>Example: data as a result of someone's observation. This can be data on the socio-economic state of some countries in one year. Or it could be a collection of messages sent by members of a social network over a period of time.</p> <p>In situations like this, typical questions are. What is the meaning of this data? Is there some structure in the data for the set of objects in question? Can these factors help predict those?</p> |

# “Small data” — it's another world

- In some domains, small data is very common especially in medicine, clinical trials etc. So, it you end up with 20 or 30 samples only
- Even lower number is not unlikely (Phase 0)
- And 1000-10 000 of subjects in II or 3rd phase trial are not Big Data
- The (extremely) small number of observations is caused by the following facts:
  - you treat people. You may not find a sufficient number of them willing to participate in your study (dangerous or “controversial” therapy, etc.). A certain disease may be rare. A drop-out may be significant due to a very serious condition the patients are under. This is why the crucial part of every clinical trial is to determine the minimum sample size required to obtain assumed precision
  - it costs real money. Assuming there are thousands of volunteers willing to participate in your trial, you have to organize everything and pay for it

<https://www.quora.com/Why-do-so-many-statisticians-not-want-to-become-data-scientists-Why-are-they-not-interested-in-Big-Data/answer/Adrian-Olszewski-1>

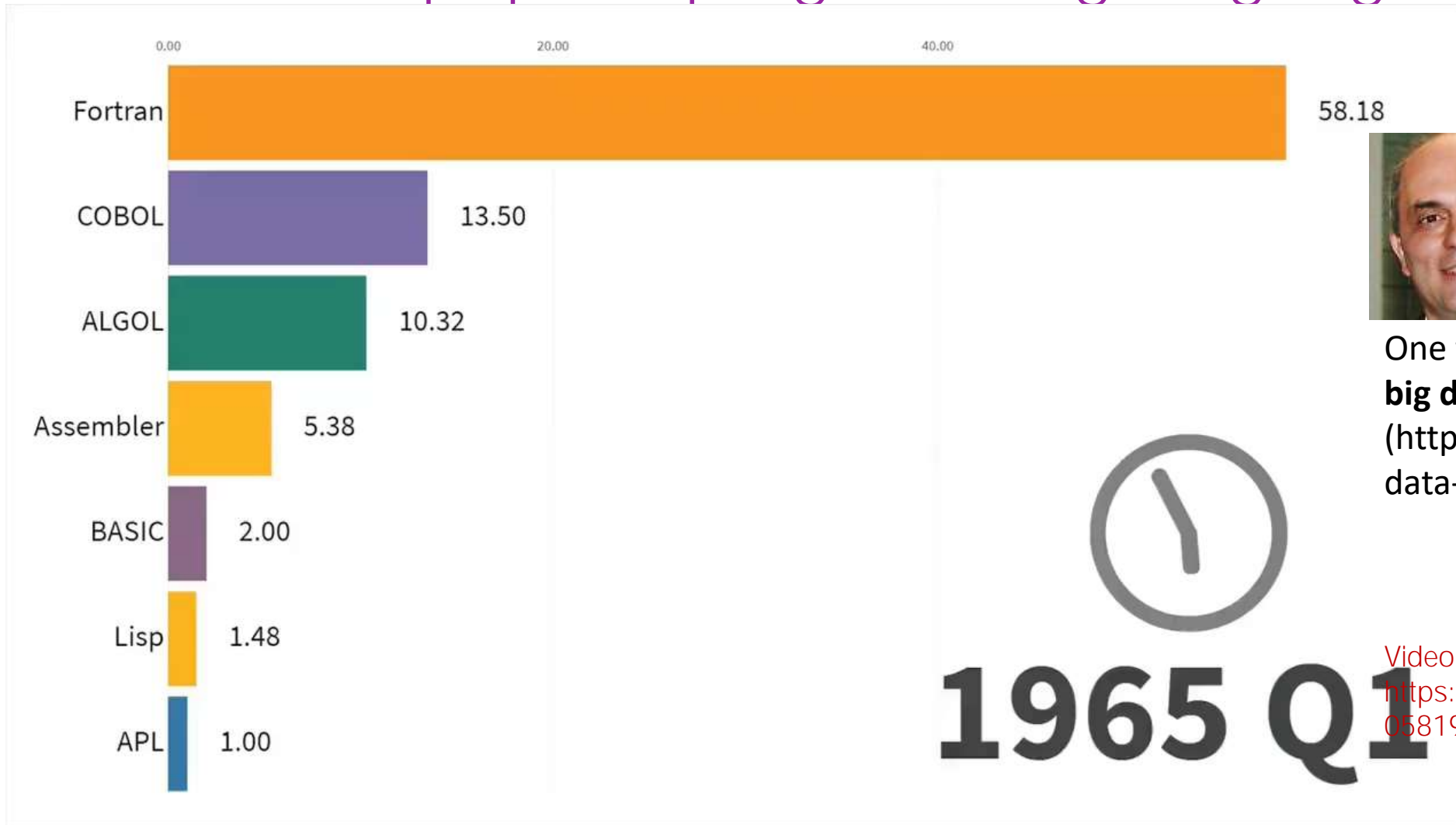
# Inferential statistics are no worse than DS

- **Inferential statistics** use a random sample of data taken from a population to describe and make inferences about the population
- Inferential statistics are valuable when examination of each member of an entire population is not convenient or possible
- For example, to measure the diameter of each nail that is manufactured in a mill is impractical
- You can measure the diameters of a representative random sample of nails. You can use the information from the sample to make generalizations about the diameters of all of the nails

<https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/basic-statistics/inference/supporting-topics/basics/what-are-inferential-statistics/>



# From 1965 to 2019, the most popular programming languages



Marcus Borba  
@marcusborba  
14/12/2019

One from the **Biggest influencers in big data in Q2 2020**  
(<https://www.verdict.co.uk//big-data-3/>)

Video from  
<https://twitter.com/marcusborba/status/1205819678842609664>

# PYPL PopularitY of Programming Language Index

## February, 2023

Worldwide, Feb 2023 compared to a year ago:

| Rank | Change | Language    | Share   | Trend  |
|------|--------|-------------|---------|--------|
| 1    |        | Python      | 27.7 %  | -0.7 % |
| 2    |        | Java        | 16.79 % | -1.3 % |
| 3    |        | JavaScript  | 9.65 %  | +0.6 % |
| 4    | ↑      | C#          | 6.97 %  | -0.5 % |
| 5    | ↓      | C/C++       | 6.87 %  | -0.6 % |
| 6    |        | PHP         | 5.23 %  | -0.8 % |
| 7    |        | R           | 4.11 %  | -0.1 % |
| 8    | ↑↑     | TypeScript  | 2.83 %  | +0.8 % |
| 9    |        | Swift       | 2.27 %  | +0.3 % |
| 10   | ↓↓     | Objective-C | 2.25 %  | -0.1 % |

- The PYPL PopularitY of Programming Language Index is created by analyzing how often language tutorials are searched on Google
- The more a language tutorial is searched, the more popular the language is assumed to be. It is a leading indicator. The raw data comes from Google Trends
- The index is updated once a month
- <http://pypl.github.io/PYPL.html>

# TIOBE programming community index

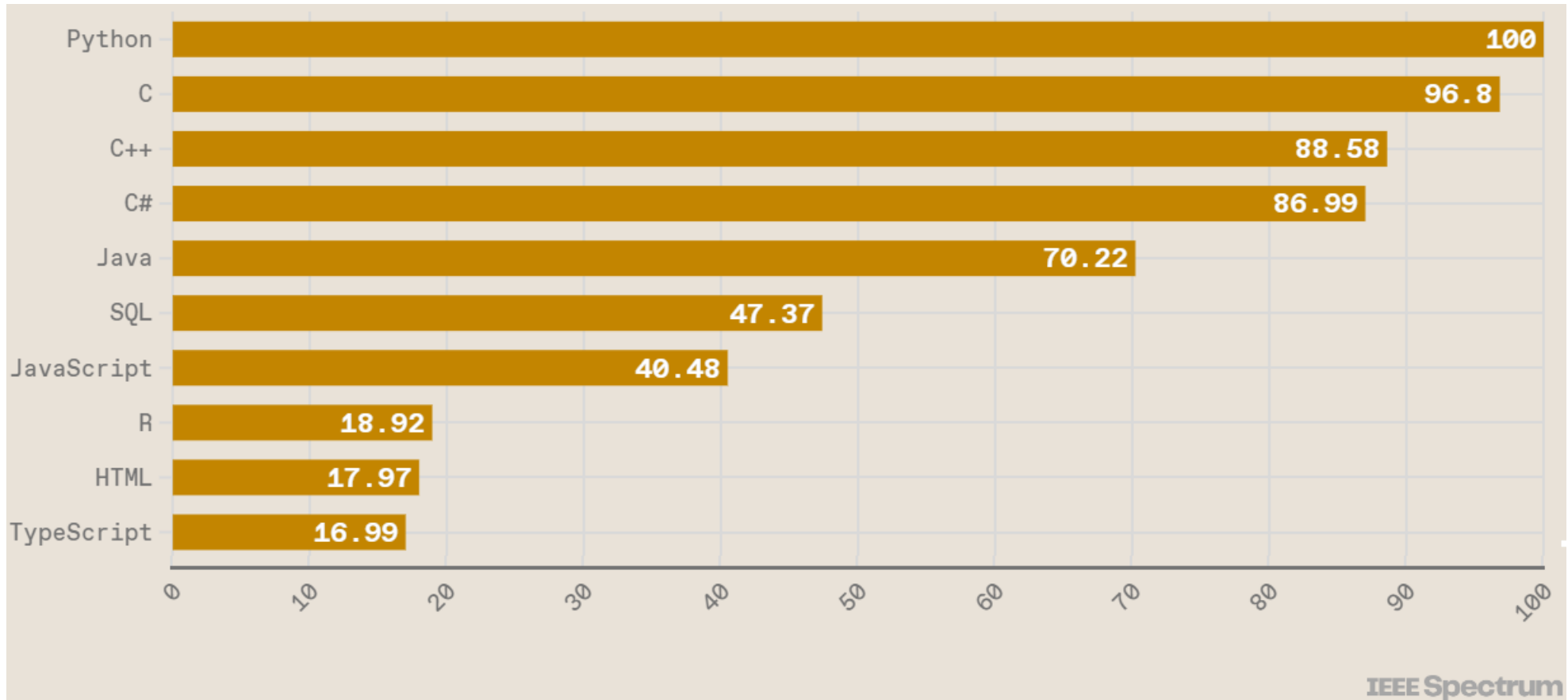
## February, 2023

| Feb 2023 | Feb 2022 | Change | Programming Language  | Ratings | Change |
|----------|----------|--------|---|---------|--------|
| 1        | 1        |        |  Python              | 15.49%  | +0.16% |
| 2        | 2        |        |  C                   | 15.39%  | +1.31% |
| 3        | 4        | ▲      |  C++                 | 13.94%  | +5.93% |
| 4        | 3        | ▼      |  Java                | 13.21%  | +1.07% |
| 5        | 5        |        |  C#                  | 6.38%   | +1.01% |
| 6        | 6        |        |  Visual Basic        | 4.14%   | -1.09% |
| 7        | 7        |        |  JavaScript          | 2.52%   | +0.70% |
| 8        | 10       | ▲      |  SQL               | 2.12%   | +0.58% |
| 9        | 9        |        |  Assembly language | 1.38%   | -0.21% |
| 10       | 8        | ▼      |  PHP               | 1.29%   | -0.49% |

- TIOBE programming community index is a measure of popularity of programming languages, created and maintained by the TIOBE Company based in Eindhoven, the Netherlands
- The index is calculated from the number of search engine results for queries containing the name of the language
- The index covers searches in Google, Google Blogs, MSN, Yahoo!, Baidu, Wikipedia and YouTube
- The index is updated once a month
- <https://www.tiobe.com/tiobe-index/>

# IEEE Spectrum Top Programming Languages

2022



- Rankings are created by weighting and combining 11 metrics from eight sources — CareerBuilder, GitHub, Google, Hacker News, the IEEE, Reddit, Stack Overflow, and Twitter
- The sources cover contexts that include social chatter, open-source code production, and job postings
- 2021: The 3rd consecutive year the top 3 remains unchanged: Python, **Java** and C

# Reason for Python's Popularity

- The major reason for Python's popularity is — it's easy to learn. Its syntax is simple compared to other languages and anyone can learn the basics of Python in a few hours or days
- Even after learning other languages such as C++ or Java, developers often prefer to stick to Python. That's because there is a python library for almost everything one can ask for
- Libraries and simple syntax made developing software in Python, simple and productive. These advantages made Python the language #1 for DS



# Машинне навчання — це ключ до НОВИХ МОЖЛИВОСТЕЙ



- Машинне навчання — це ключ. Це те, що здатно перетворити все, що ми робимо. Зараз ми намагаємося так чи інакше застосовувати МН у всіх наших продуктах, будь то Google Search, YouTube або Google Play. Корпорація тільки починає активну інтеграцію, але зовсім скоро ми будемо застосовувати МН у кожній області, причому на систематичній основі
- "[#AI](#) will be more profound for humanity than fire and electricity"

*Сундар Пічай (Sundar Pichai), генеральний директор Google, 2016*

# The Global Unicorn Club (Sept. 2020) founded by Ukrainians

- **Revolut, \$ 5.5 billion.** A financial technology company headquartered in London. One of the co-founders is **Vlad Yatsenko** from Odesa. Revolut's flagship product is a super-low commission worldwide payment card
- **Grammarly, \$ 1 billion.** Online service to help you write texts in English. Grammarly was founded in 2009 by three Ukrainians: **Oleksiy Shevchenko, Maxym Lytvyn and Dmytro Leader**
- **Bitfury, a \$ 1 billion.** Industrial miner and software developer for the Bitcoin blockchain. The company was founded by Ukrainian **Valeriy Nebesniy** and Latvian **Valeriy Vavilov**
- **GitLab, \$ 2.7 billion.** This is a cloud service for programmers, analogous to the less convenient and outdated GitHub. It was founded by Ukrainian **Dmytro Zaporozhets** with a Dutch partner

A **unicorn** is a private company with a valuation over \$1 billion

As of September 2020, there are more than **400** unicorns around the world



<https://www.cbinsights.com/research-unicorn-companies>

# Ой, щось відбувається прямо зараз 😊

- **ChatGPT** – чат-бот зі штучним інтелектом, розроблений компанією OpenAI і здатний працювати в діалоговому режимі, що підтримує запити природними мовами.
- 30 листопада 2022 р.
- **DALL-E** – нейронна мережа OpenAI, створена за фінансової підтримки Microsoft, здатна генерувати високоякісні зображення, виходячи з текстових описів англійською мовою
- 5 січня 2021 р. (перша версія)
- 6 квітня 2022 р. (друга версія)



# Arthur Samuel's Checkers program



- Arthur Samuel (1901 – 1990)
- On February 24, 1956, Arthur Samuel's Checkers program, which was developed for play on the IBM 701, was demonstrated to the public on television
- The program was a sensational demonstration of the advances in both hardware and skilled programming and caused IBM's stock to increase 15 points overnight

# Перше визначення машинного навчання

- **Machine learning** gives computers the ability to learn without being explicitly programmed

Arthur Samuel. Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development, 1959, 3 (3): 210–229 Arthur Samuel (1901 – 1990)

- Іншими словами, **машинне навчання** — це процес, в результаті якого машина (комп'ютер) здатна показувати поведінку, яка в неї не було **явно** закладена (запрограмована)



# Класичне визначення машинного навчання

- To be more precise, we say that a machine **learns** with respect to a particular task  $T$ , performance metric  $P$ , and type of experience  $E$ , if the system reliably improves its performance  $P$  at task  $T$ , following experience  $E$
- Кажуть, що комп'ютерна програма **НАВЧАЄТЬСЯ** при розв'язанні деякої задачі із класу  $T$ , якщо її продуктивність згідно з метрикою  $P$  поліпшується з набуттям досвіду  $E$

T.M. Mitchell. Machine Learning. McGraw-Hill, 1997

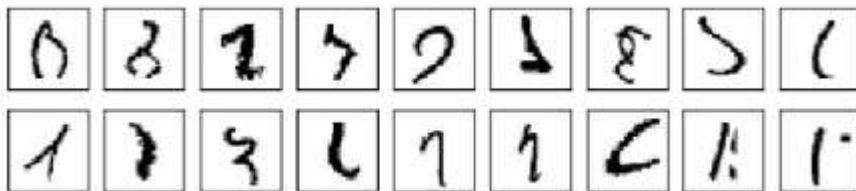
# Класичне визначення машинного навчання

- To be more precise, we say that a **machine learns** with respect to a particular task  $T$ , performance metric  $P$ , and type of experience  $E$ , if the system reliably improves its performance  $P$  at task  $T$ , following experience  $E$
  - Кажуть, що **комп'ютерна програма *навчається*** при розв'язанні деякої задачі із класу  $T$ , якщо її продуктивність згідно з метрикою  $P$  поліпшується з набуттям досвіду  $E$
- T.M. Mitchell. Machine Learning. McGraw-Hill, 1997

# Приклади класів задач $T$ в машинному навчанні

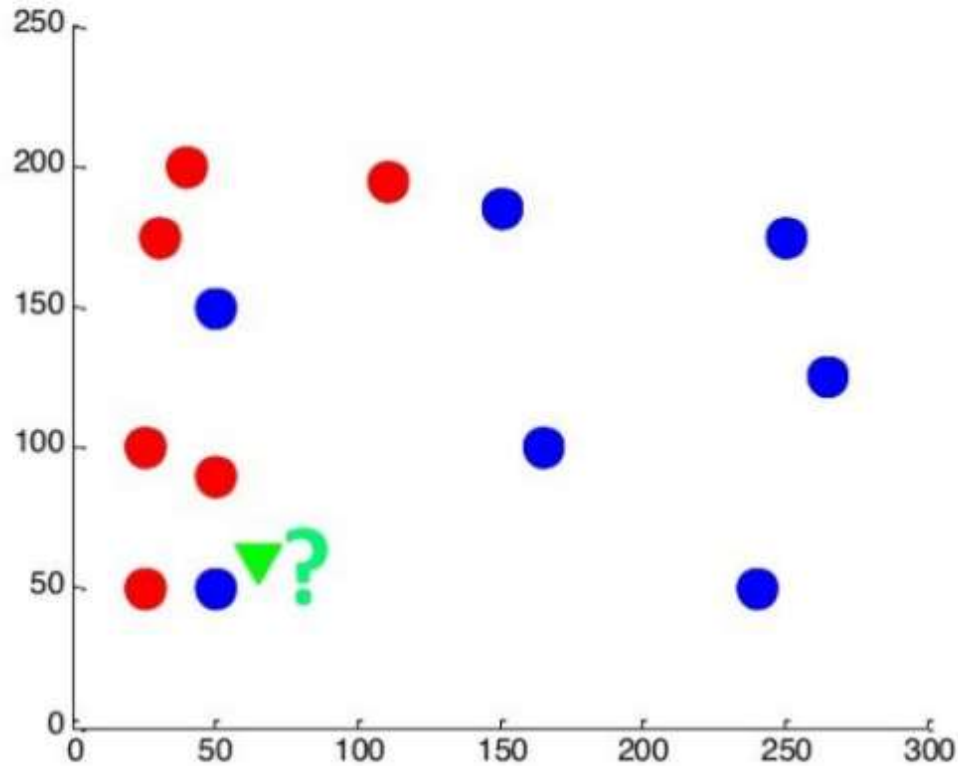
- **Класифікація** – віднесення об'єкта до однієї з категорій на підставі його ознак
- **Регресія** – прогнозування кількісної ознаки об'єкта на підставі інших його ознак
- **Кластеризація** – розбиття множини об'єктів на групи на підставі ознак цих об'єктів так, щоб усередині груп об'єкти були схожі 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

# How well does the program recognize handwritten digits?



- **MNIST data set** is based on two data sets collected by the United States' National Institute of Standards and Technology (NIST)
- 50,000 image data set + 10,000 image validation set

# Віднесення об'єкта до однієї з категорій на підставі його ознак



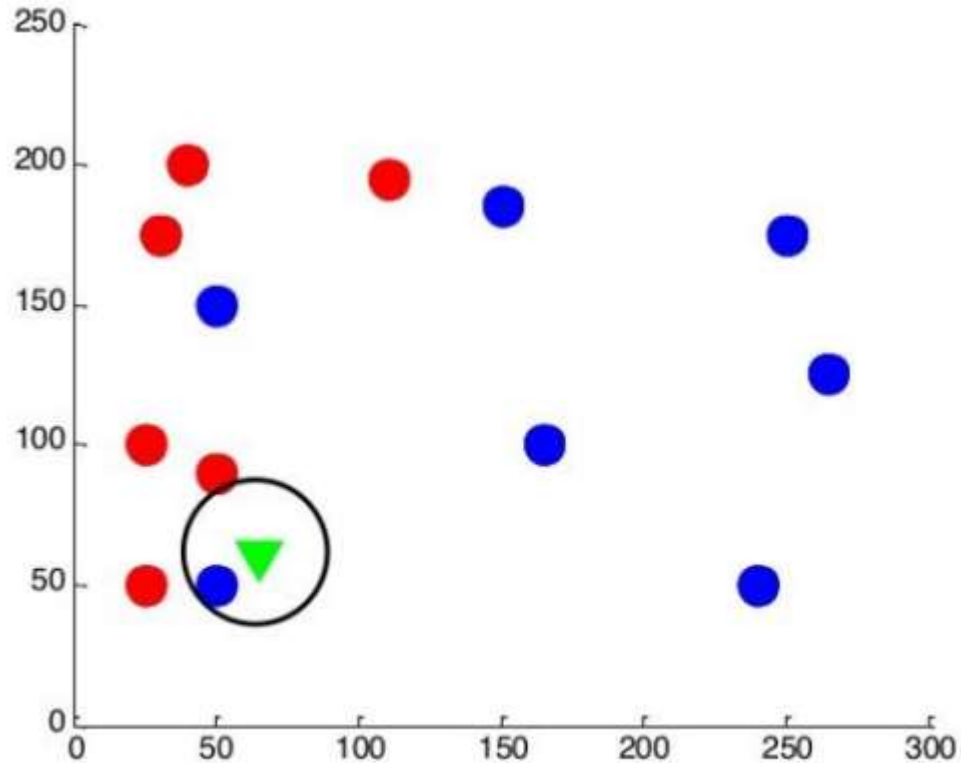
<https://habrahabr.ru/company/yandex/blog/206058/>

## Задача класифікації

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



# Метод найближчого сусіда



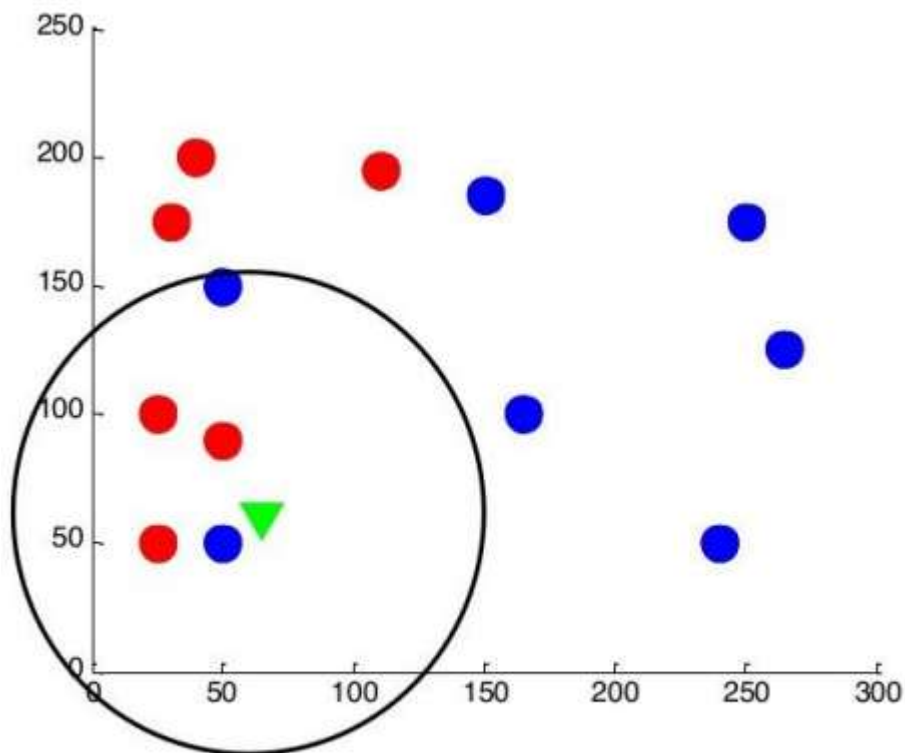
<https://habrahabr.ru/company/yandex/blog/206058/>

- Відстань між кружечками рахуємо у звичайній евклідовій метриці:

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

## Задача класифікації

# Метод 5 найближчих сусідів



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

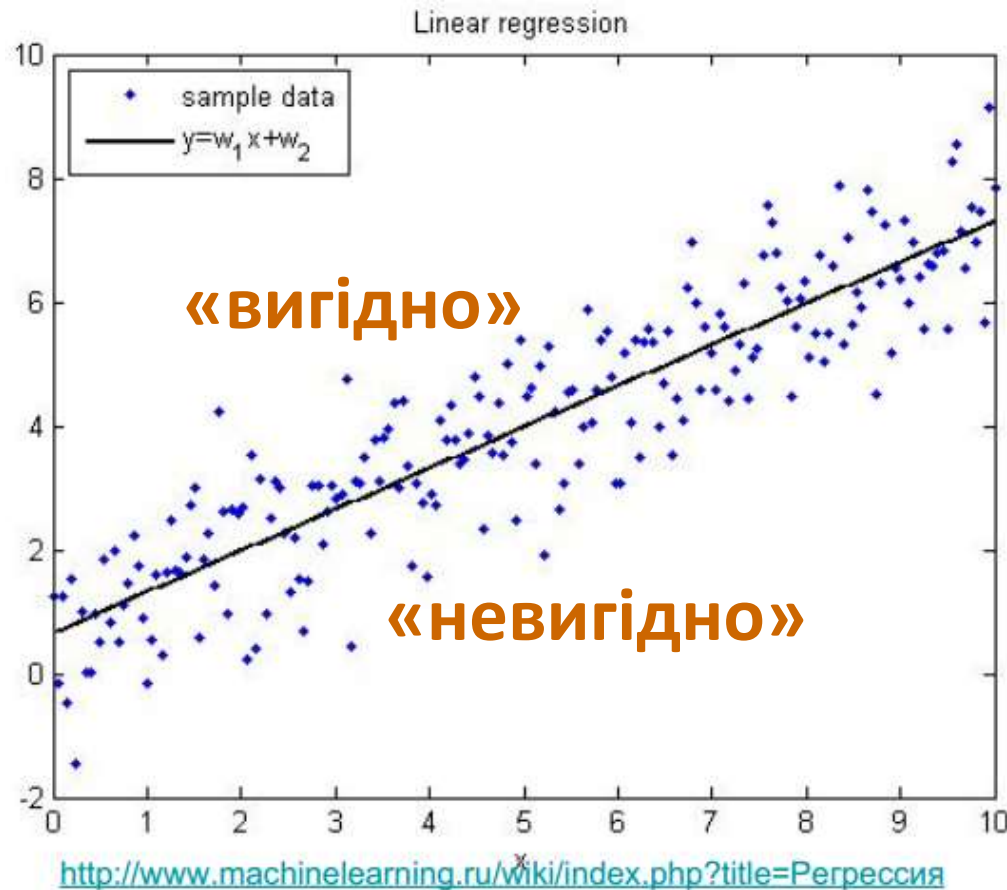
- А скільки взяти найближчих сусідів до уваги?
- На «нескінченних» вибірках метод  $k$  найближчих сусідів дає оптимальний метод класифікації

## Задача класифікації

# 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

# Прогнозування кількісної ознаки об'єкта на підставі інших його ознак



## Задача регресії

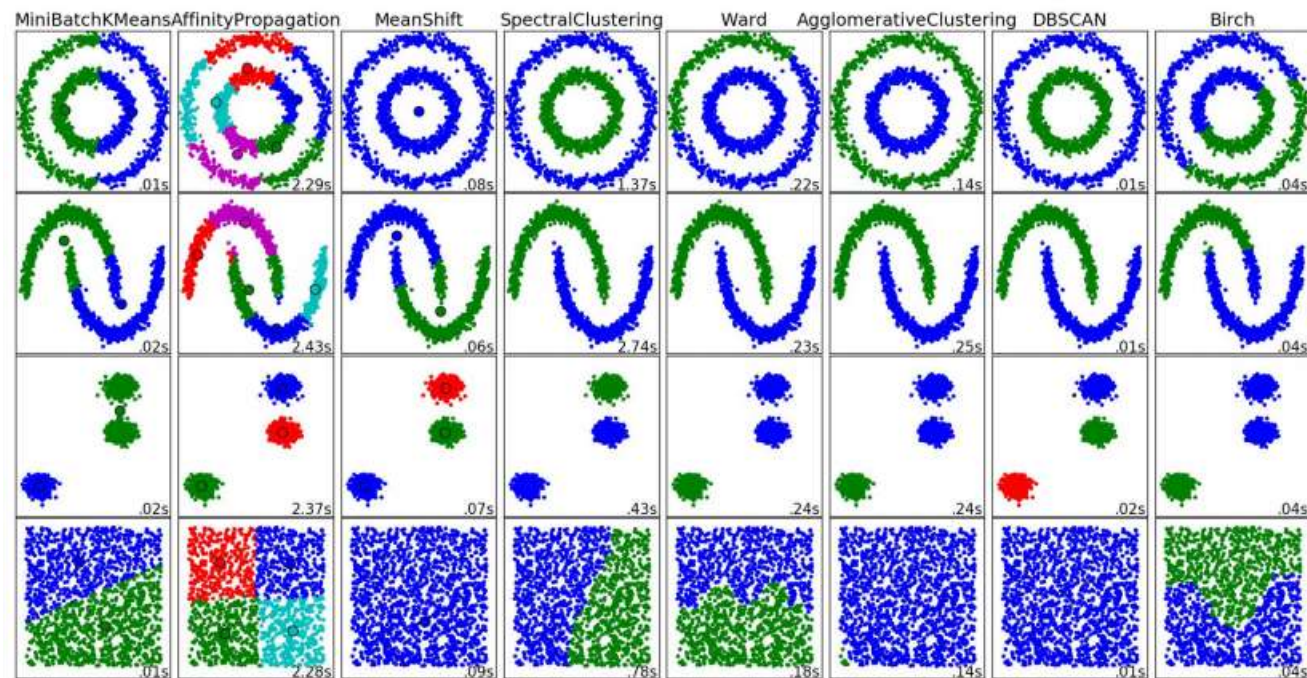
- Вартість будинку як функція від його площі
- Задачі класифікації та регресії часто взаємопов'язані:
  - ❖ «вигідно» і «невигідно»
  - ❖ кредитний скорінг

# 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**



# Розбиття множини об'єктів на групи



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

- **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

# 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)

# Класичне визначення машинного навчання

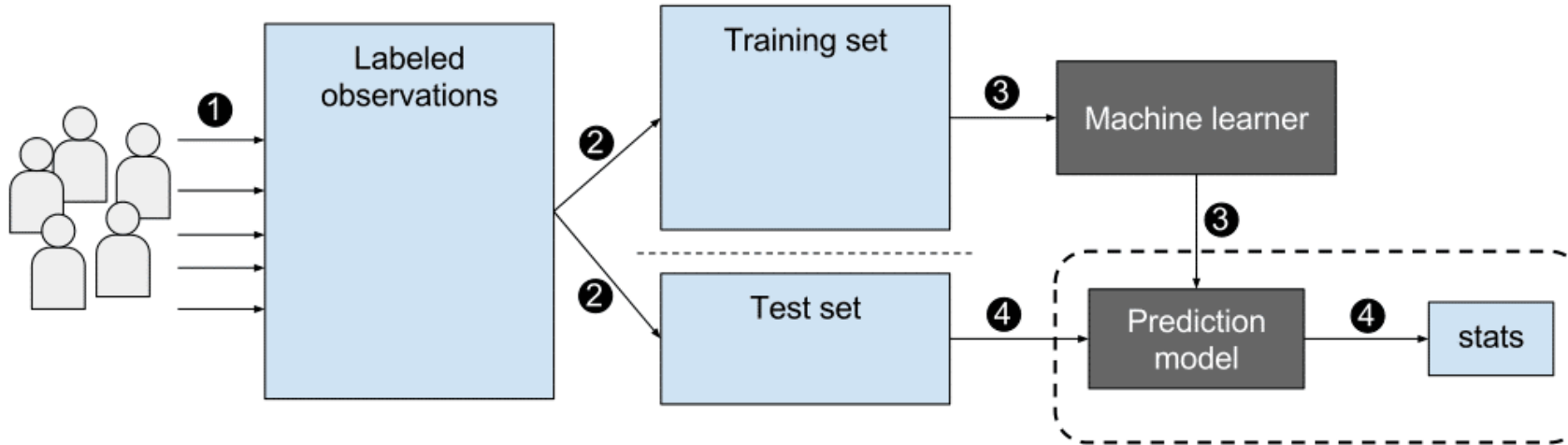
- To be more precise, we say that a machine learns with respect to a particular task  $T$ , performance metric  $P$ , and type of **experience  $E$** , if the system reliably improves its performance  $P$  at task  $T$ , following **experience  $E$**
- Кажуть, що комп'ютерна програма **навчається** при розв'язанні деякої задачі із класу  $T$ , якщо її продуктивність згідно з метрикою  $P$  поліпшується з набуттям **досвіду  $E$**

T.M. Mitchell. Machine Learning. McGraw-Hill, 1997

# Досвід у машинному навчанні = прецеденти

- 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)

# What Is Supervised Learning?



- If you're learning a task under supervision, someone is present judging whether you're getting the right answer. Similarly, in supervised learning, that means having a full set of labeled data while training an algorithm.
- Fully labeled means that each example in the training dataset is tagged with the answer the algorithm should come up with on its own. So, a labeled dataset of flower images would tell the model which photos were of roses, daisies and daffodils. When shown a new image, the model compares it to the training examples to predict the correct label



# What Is Unsupervised Learning?

Depending on the problem at hand, the unsupervised learning model can organize the data in different ways.

- **Clustering:** Without being an expert ornithologist, it's possible to look at a collection of bird photos and separate them roughly by species, relying on cues like feather color, size or beak shape. That's how the most common application for unsupervised learning, clustering, works: the deep learning model looks for training data that are similar to each other and groups them together.
- **Anomaly detection:** Banks detect fraudulent transactions by looking for unusual patterns in customer's purchasing behavior. For instance, if the same credit card is used in California and Denmark within the same day, that's cause for suspicion. Similarly, unsupervised learning can be used to flag outliers in a dataset.
- **Association:** Fill an online shopping cart with diapers, applesauce and sippy cups and the site just may recommend that you add a bib and a baby monitor to your order. This is an example of association, where certain features of a data sample correlate with other features. By looking at a couple key attributes of a data point, an unsupervised learning model can predict the other attributes with which they're commonly associated.

# What Is Unsupervised Learning?

- **Autoencoders:** Autoencoders take input data, compress it into a code, then try to recreate the input data from that summarized code. It's like starting with Moby Dick, creating a SparkNotes version and then trying to rewrite the original story using only SparkNotes for reference. While a neat deep learning trick, there are fewer real-world cases where a simple autocoder is useful. But add a layer of complexity and the possibilities multiply: by using both noisy and clean versions of an image during training, autoencoders can remove noise from visual data like images, video or medical scans to improve picture quality.
- Because there is no “ground truth” element to the data, it's difficult to measure the accuracy of an algorithm trained with unsupervised learning. But there are many research areas where labeled data is elusive, or too expensive, to get. In these cases, giving the deep learning model free rein to find patterns of its own can produce high-quality results.

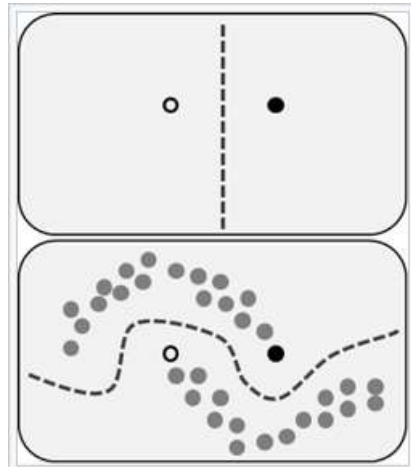
# Приклади задач різних типів навчання

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

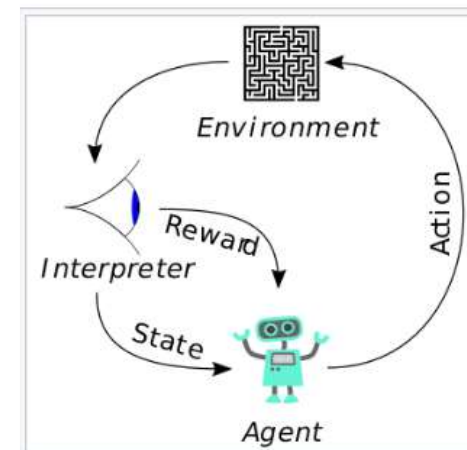
---

## Приклади інших типів навчання:

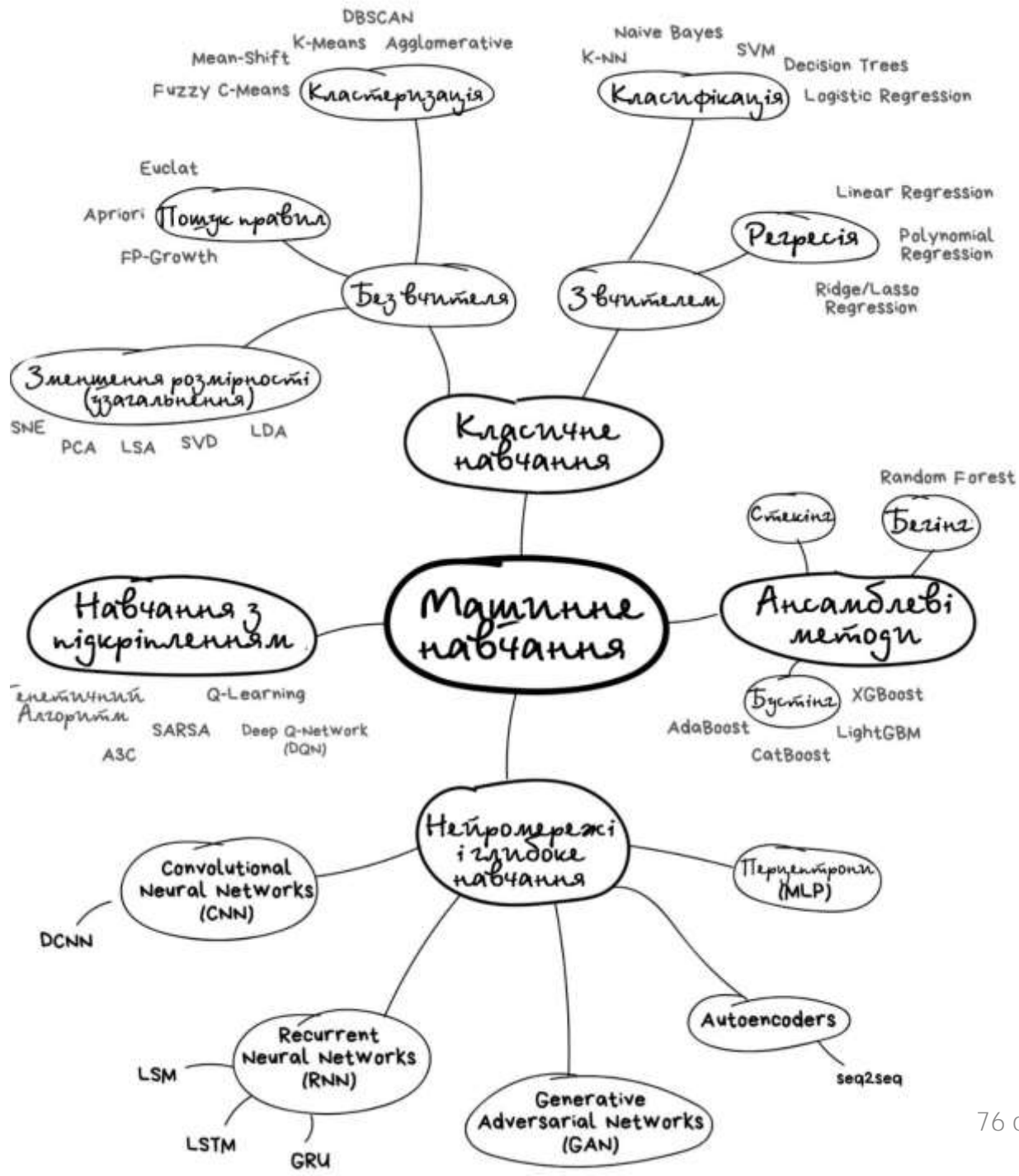
❖ напівконтрольоване навчання  
(semi-supervised learning)



❖ навчання з підкріпленням  
(reinforcement learning)



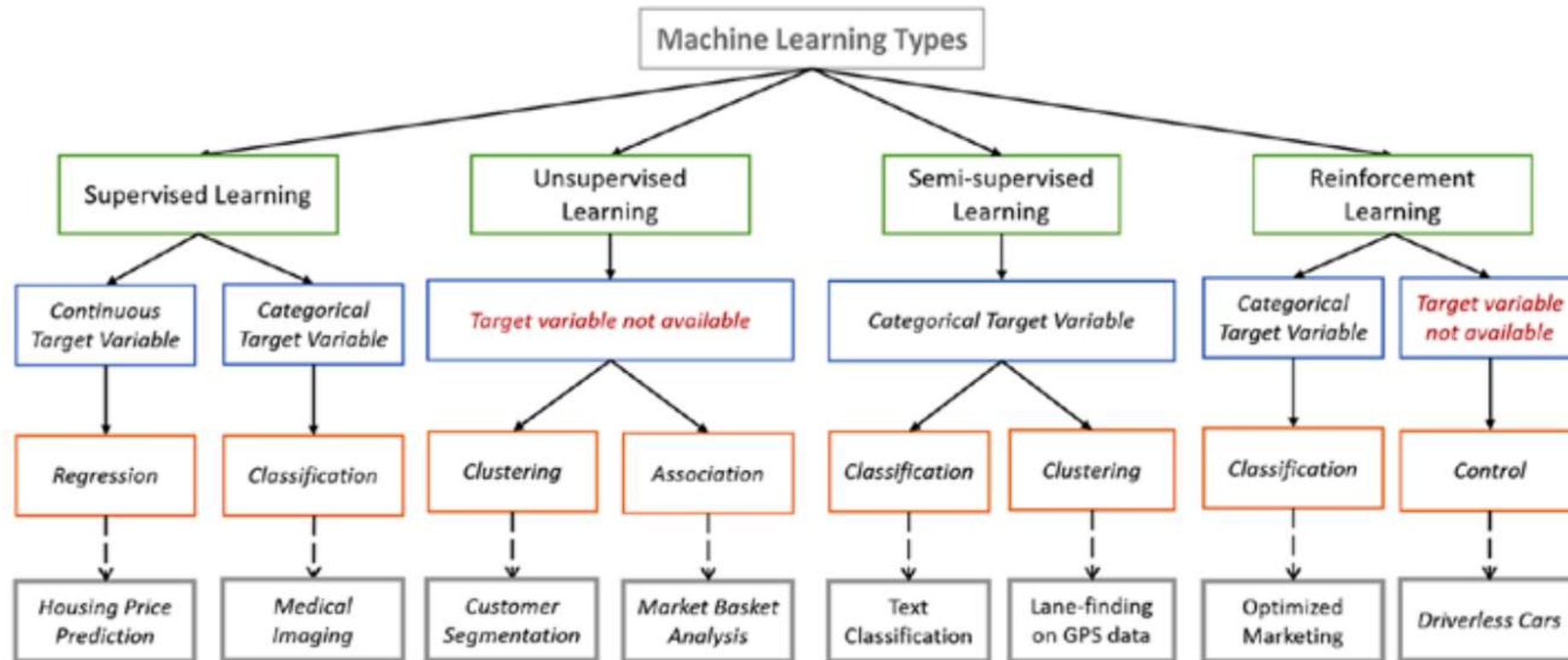
# Machine learning types



source: <http://www.mmf.lnu.edu.ua/ar/1739>

в дійсності, поцуплено з:  
[https://vas3k.ru/blog/machine\\_learning/](https://vas3k.ru/blog/machine_learning/)

# Machine learning types



[https://cdn-images-1.medium.com/max/1400/1\\*ZCeOEBhvEVLmwCh7vr2RVA.png](https://cdn-images-1.medium.com/max/1400/1*ZCeOEBhvEVLmwCh7vr2RVA.png)

# Класичне визначення машинного навчання

- To be more precise, we say that a machine learns with respect to a particular task  $T$ , **performance metric  $P$** , and type of experience  $E$ , if the system reliably improves its performance  $P$  at task  $T$ , following experience  $E$
- Кажуть, що комп'ютерна програма **навчається** при розв'язанні деякої задачі із класу  $T$ , якщо її продуктивність згідно з **метрикою  $P$**  поліпшується з набуттям досвіду  $E$

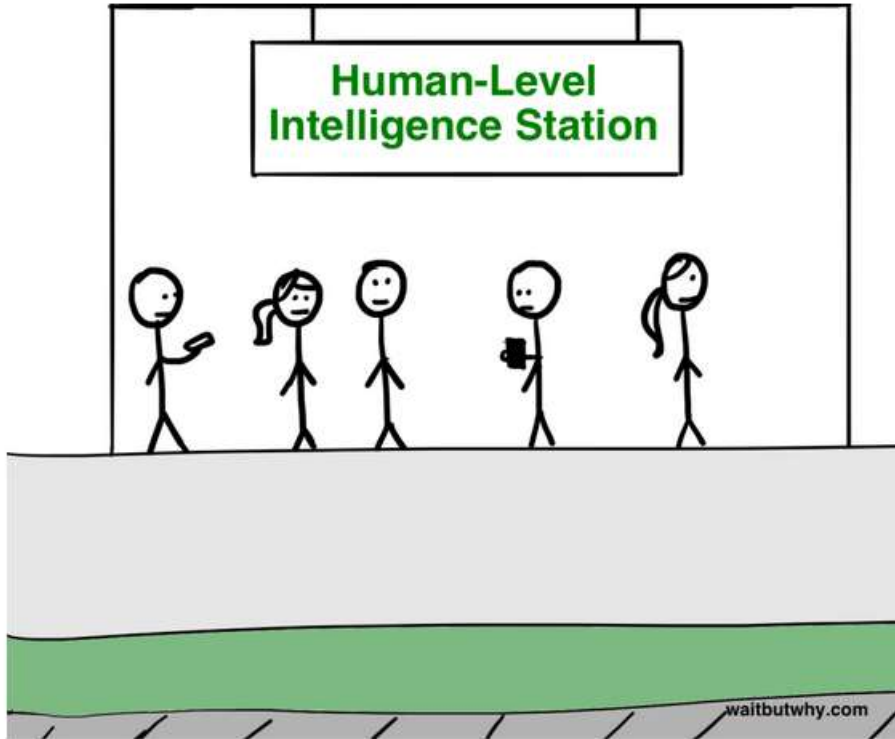
T.M. Mitchell. Machine Learning. McGraw-Hill, 1997



# Метрика $P$ оцінки продуктивності алгоритму

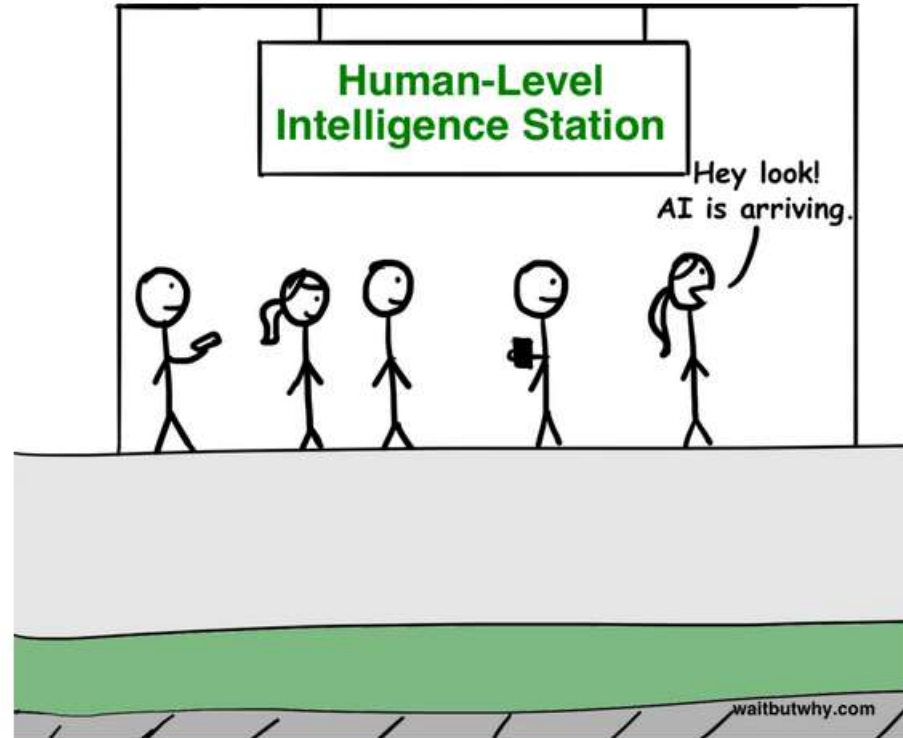
- Розрізняються для різних задач і алгоритмів
- Розглянемо декілька ключових метрик оцінки продуктивності алгоритму, що розв'язує задачу класифікації
- На прикладі задачі діагностики хворих людей

# Чи замінить штучний інтелект людину?



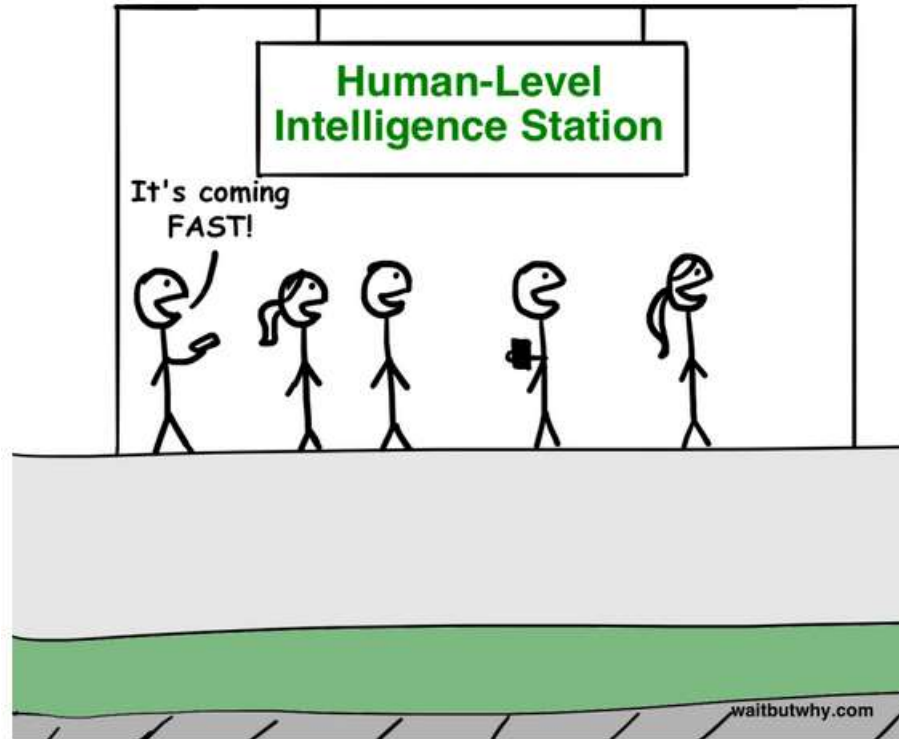
- AI won't replace managers, but managers who use AI will replace those who don't [HARVARD BUSINESS REVIEW, JULY, 2017]

# Чи замінить штучний інтелект людину?



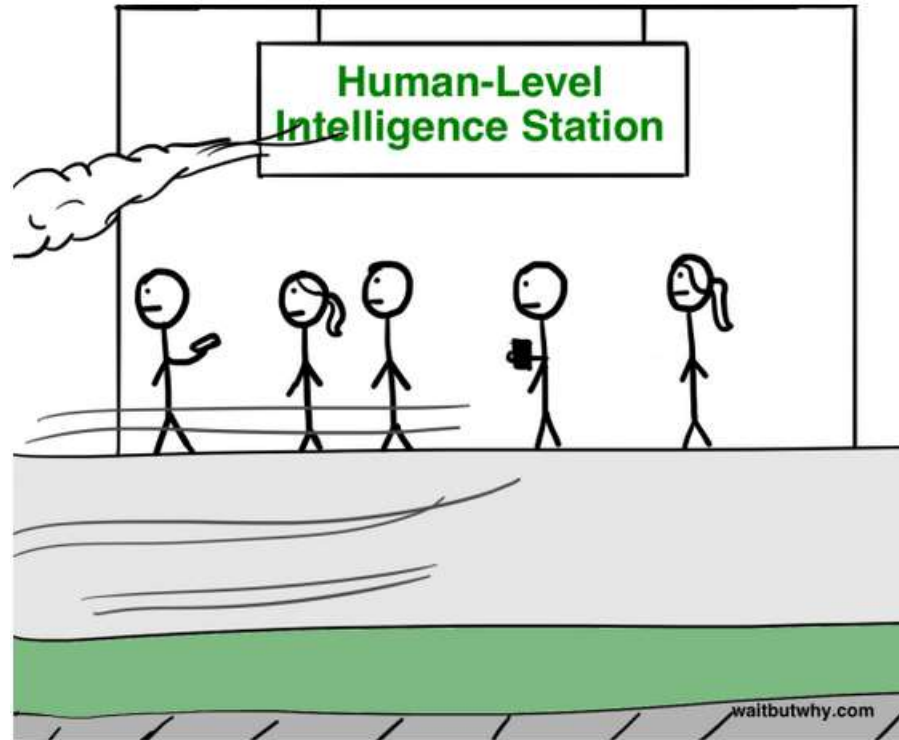
- AI won't replace managers, but managers who use AI will replace those who don't [HARVARD BUSINESS REVIEW, JULY, 2017]

# Чи замінить штучний інтелект людину?



- AI won't replace managers, but managers who use AI will replace those who don't [HARVARD BUSINESS REVIEW, JULY, 2017]

# Чи замінить штучний інтелект людину?



- AI won't replace managers, but managers who use AI will replace those who don't [HARVARD BUSINESS REVIEW, JULY, 2017]

# Thanks for Your Attention!

- Please do not hesitate to contact me by e-mail if you have any questions on my class:
- [chertov@i.ua](mailto:chertov@i.ua)
- But preferable: Slack, #machine-learning-course