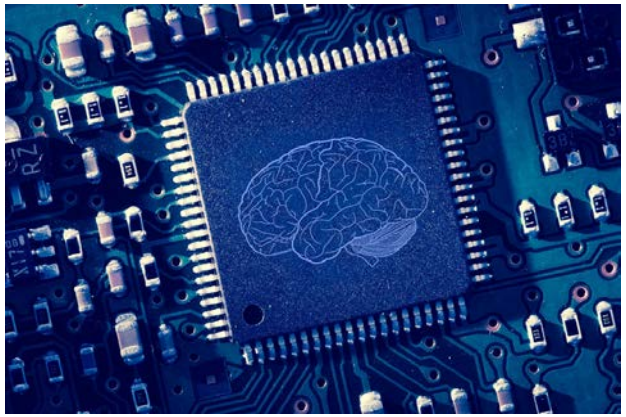
ΕΠΙΧΕΙΡΗΣΙΑΚΗ ΝΟΗΜΟΣΥΝΗ ΣΤΟΝ ΤΟΥΡΙΣΜΟ

Αριστείδης Γ. Βραχάτης, Dipl-Ing, M.Sc, PhD

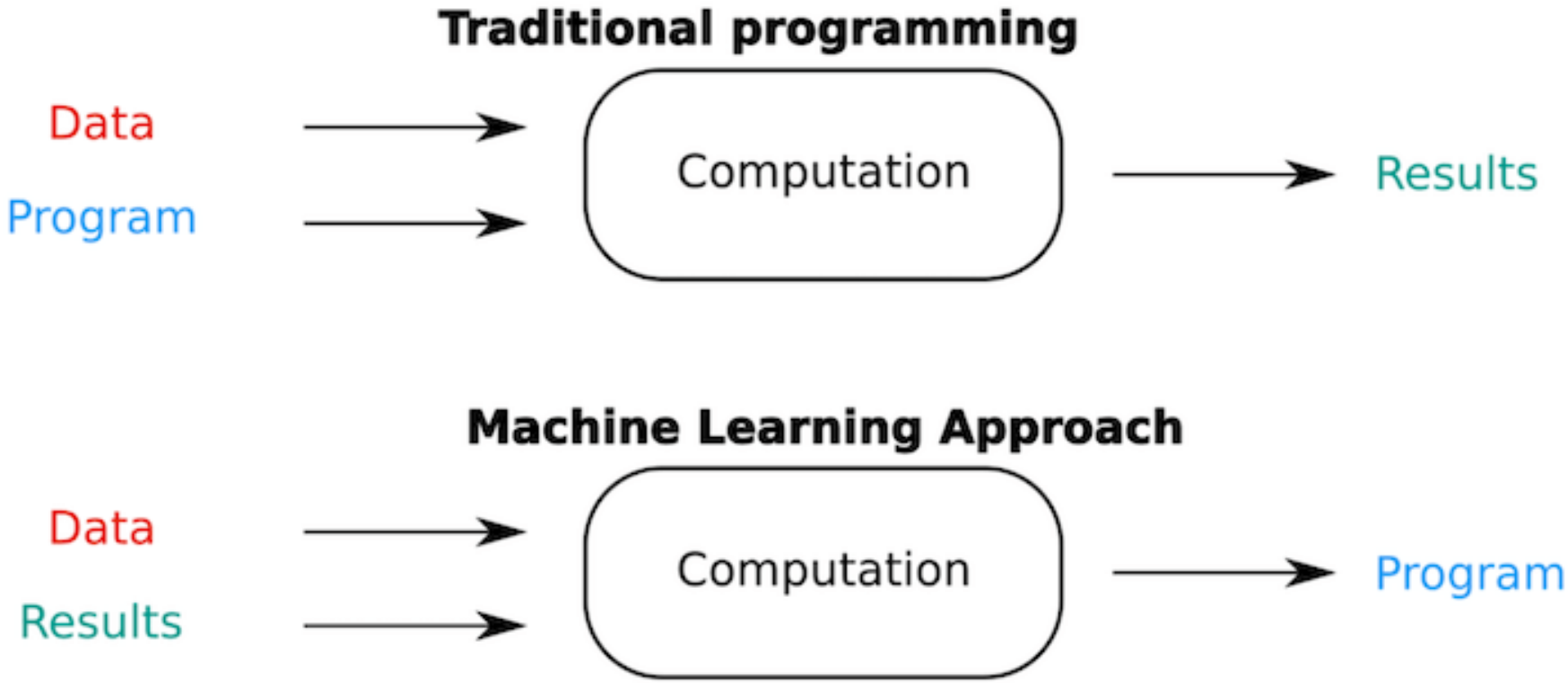
Επίκουρος Καθηγητής, Τμήμα Πληροφορικής, Ιόνιο Πανεπιστήμιο

Τεχνητή Νοημοσύνη

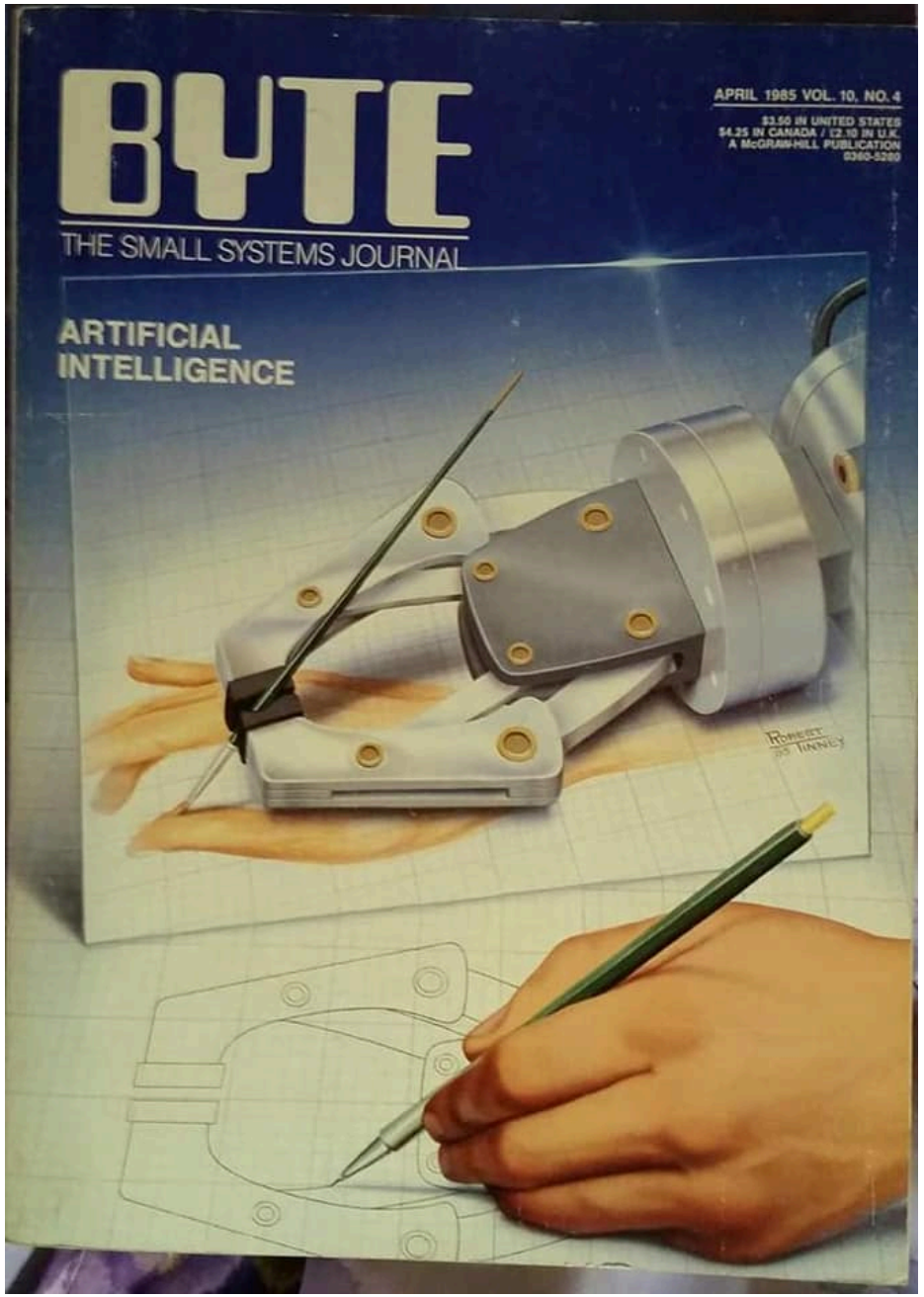
– Οι υπολογιστές μπορούν να επιδείξουν αληθινή ευφυΐα ???



ΤΙ ΕΊΝΑΙ Η ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ ???


Traditional programming
Data
Program
Computation
Results
Machine Learning Approach
Data
Results
Computation
Program

Τι έχει αλλάξει τα τελευταία χρόνια ?



Δεδομένα….το νέο «πετρέλαιο»

Γιουβάλ Νόα Χαράρι στην «Κ»: Οι πόλεμοι θα γίνονται με data

The Fourth Industrial Revolution

1st Industrial Revolution
WATER & STEAM

Steam and water power replace human and animal power with machines.

2nd Industrial Revolution
ELECTRICITY

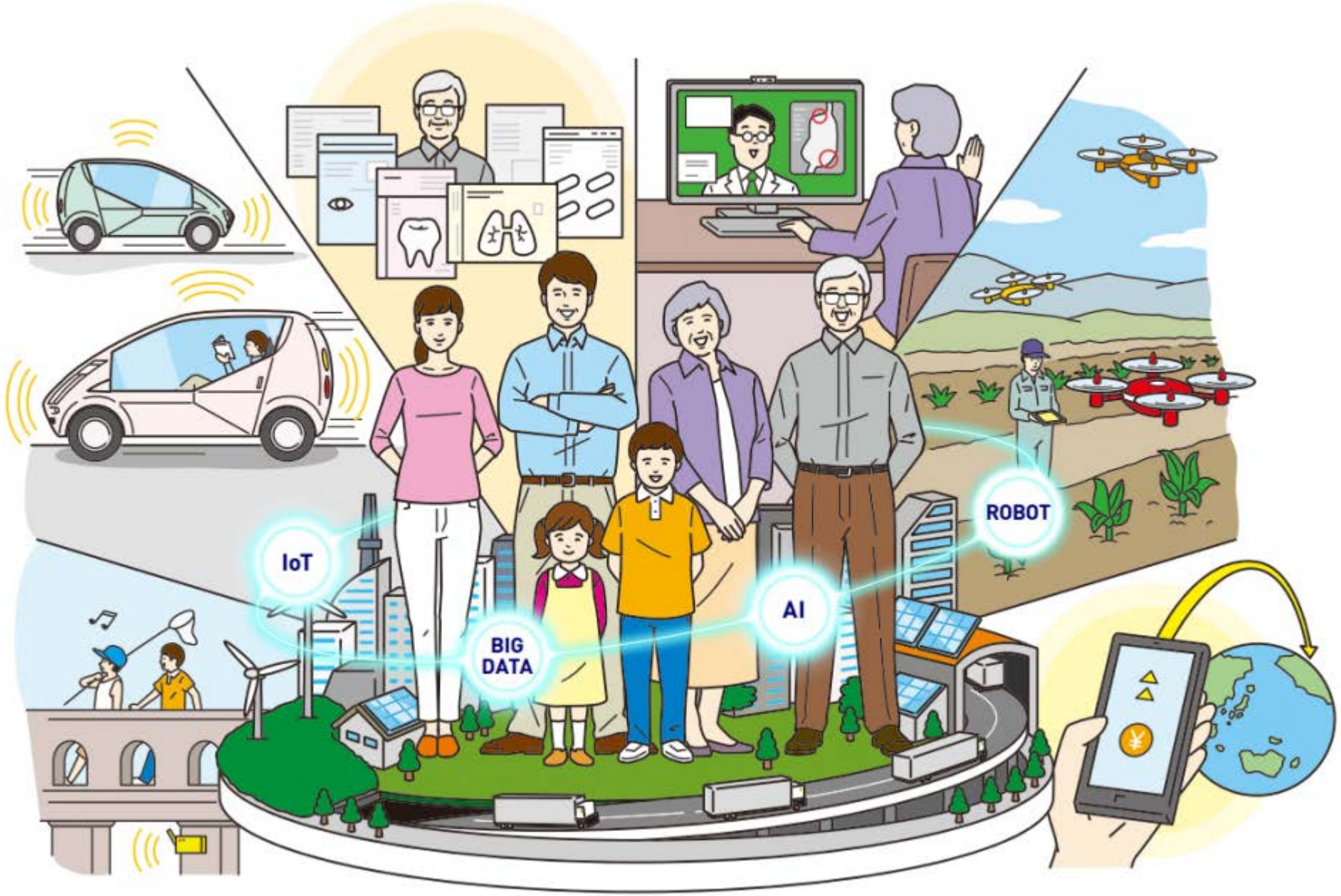
Electricity, internal combustion engines, airplanes, telephones, cars, radio, and mass production.

3rd Industrial Revolution
AUTOMATION

Electronics, the internet and IT used to further the automation of mass production.

4th Industrial Revolution
CYBER-PHYSICAL SYSTEMS

Driverless cars, smart robotics, materials that are lighter and tougher, and a manufacturing process built around 3D printing.


IoT
BIG DATA
AI
ROBOT

An Example: Self Driving Cars


amazon
fulfillment

Τεχνητή Νοημοσύνη - Αμφισβήτηση της ορολογίας

Χαρακτηριστικό	Τι Κάνει η ΤΝ	Τι Δεν Κάνει η ΤΝ
Εκπαίδευση	Εκπαιδεύεται σε συγκεκριμένα δεδομένα	Δεν "σκέφτεται" όπως ο άνθρωπος
Προσαρμογή	Προσαρμόζεται σε νέα δεδομένα	Δεν "καταλαβαίνει" το περιεχόμενο των δεδομένων
Δεδομένα	Χρειάζεται μεγάλες ποσότητες δεδομένων	Δεν έχει βαθιά κατανόηση των θεμάτων
Αποτελεσματικότητα	Είναι αποτελεσματική σε συγκεκριμένες εργασίες	Δεν έχει συνείδηση, συναισθήματα ή προθέσεις
Εξειδίκευση	Εξειδικευμένη για συγκεκριμένες εργασίες	Δεν μπορεί να εφαρμοστεί ευρέως χωρίς νέα εκπαίδευση

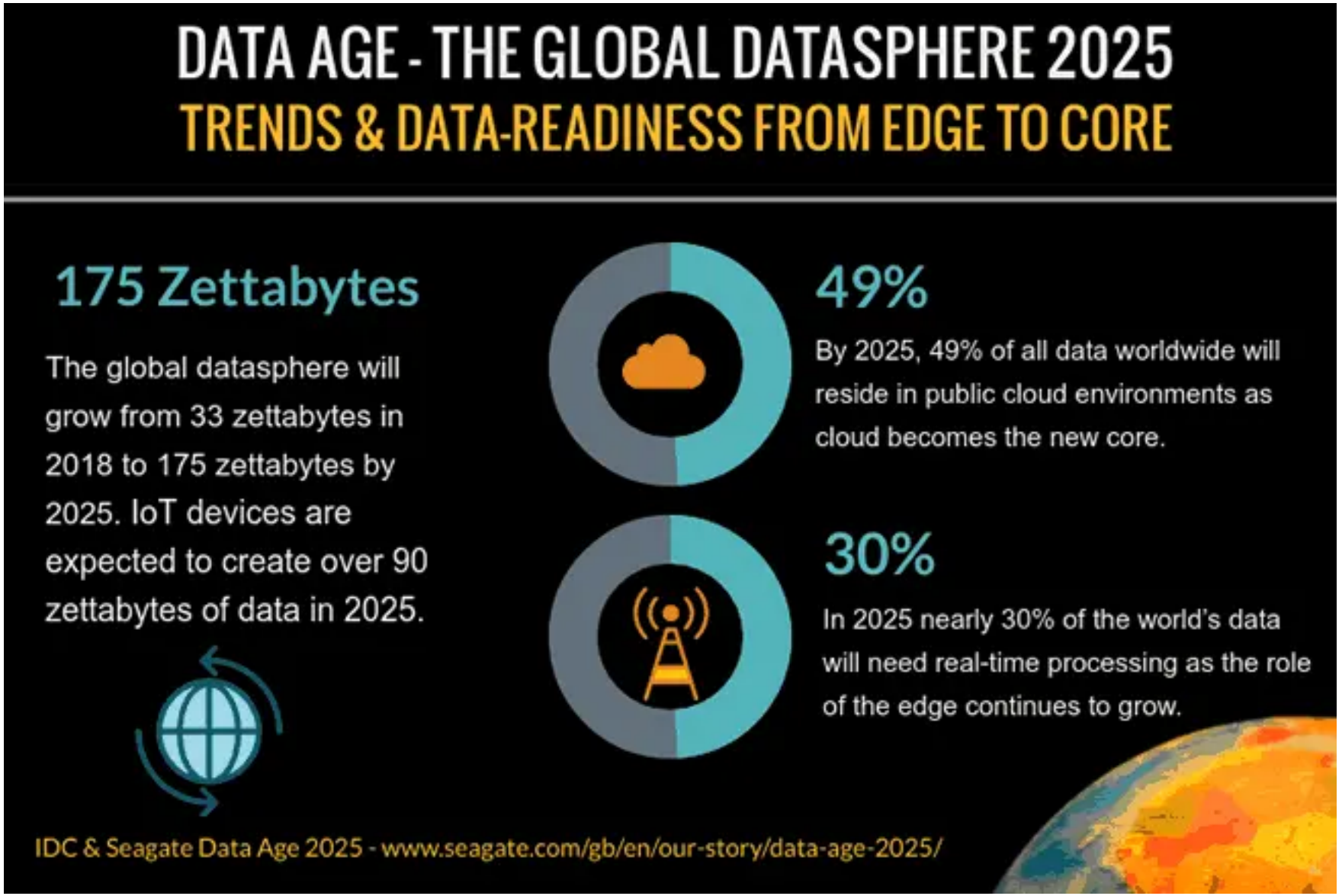

DATA AGE - THE GLOBAL DATASPHERE 2025
TRENDS & DATA-READINESS FROM EDGE TO CORE
175 Zettabytes
The global datasphere will grow from 33 zettabytes in 2018 to 175 zettabytes by 2025. IoT devices are expected to create over 90 zettabytes of data in 2025.
49%
By 2025, 49% of all data worldwide will reside in public cloud environments as cloud becomes the new core.
30%
In 2025 nearly 30% of the world's data will need real-time processing as the role of the edge continues to grow.
IDC & Seagate Data Age 2025 - www.seagate.com/gb/en/our-story/data-age-2025/

Figure 1 – Annual Size of the Global Datasphere

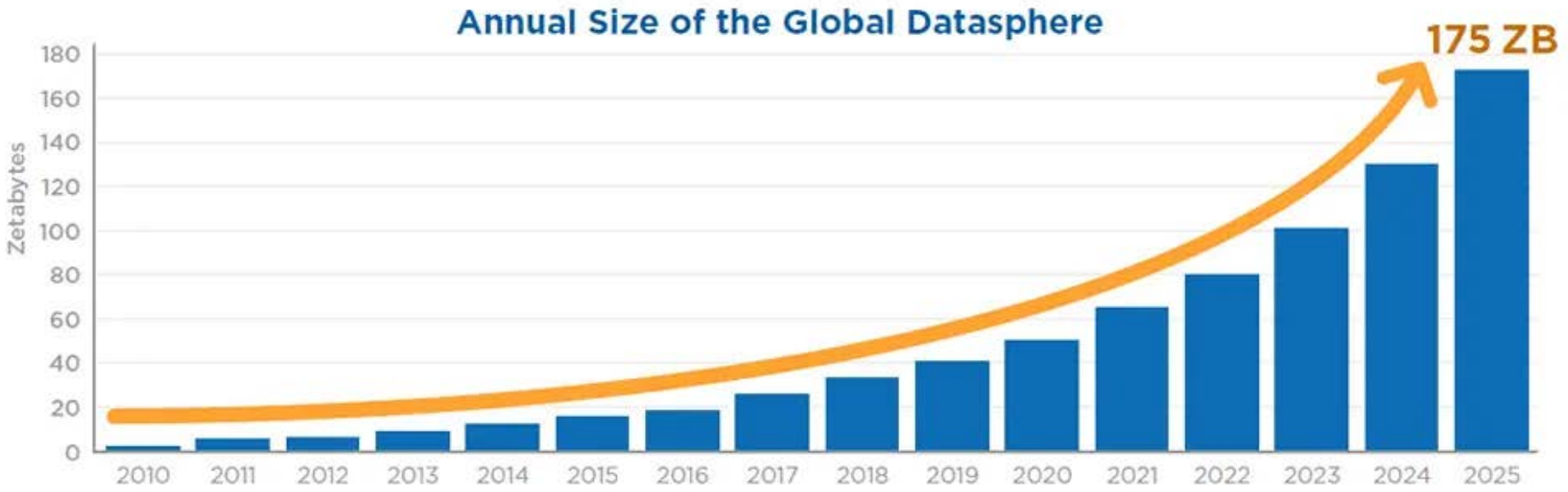
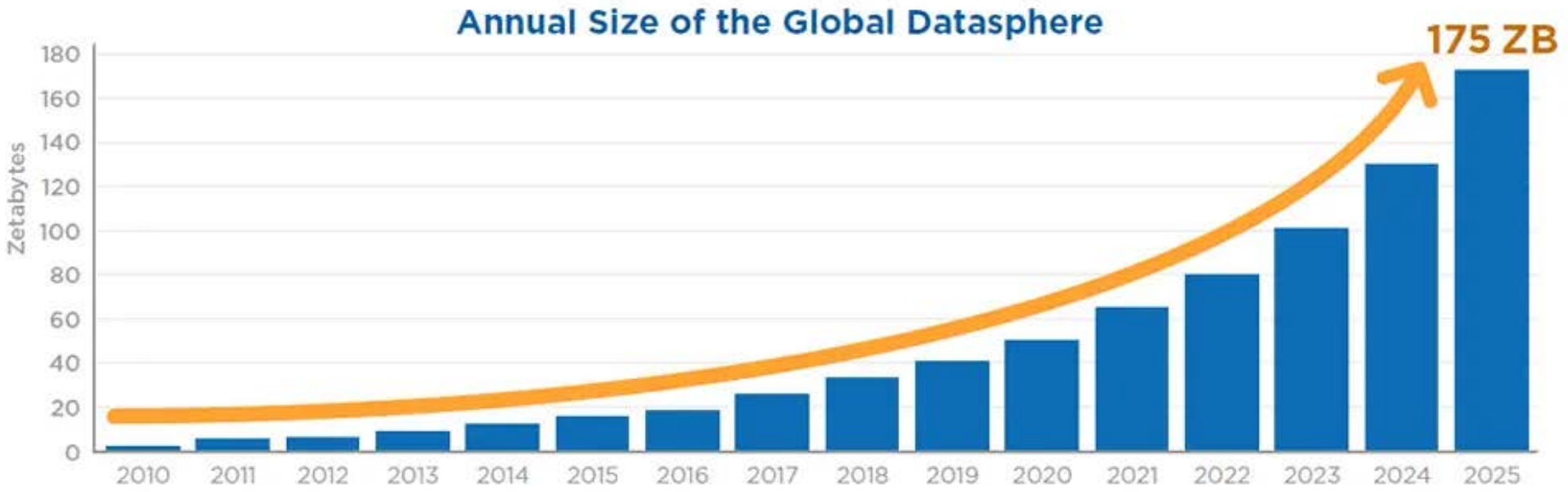


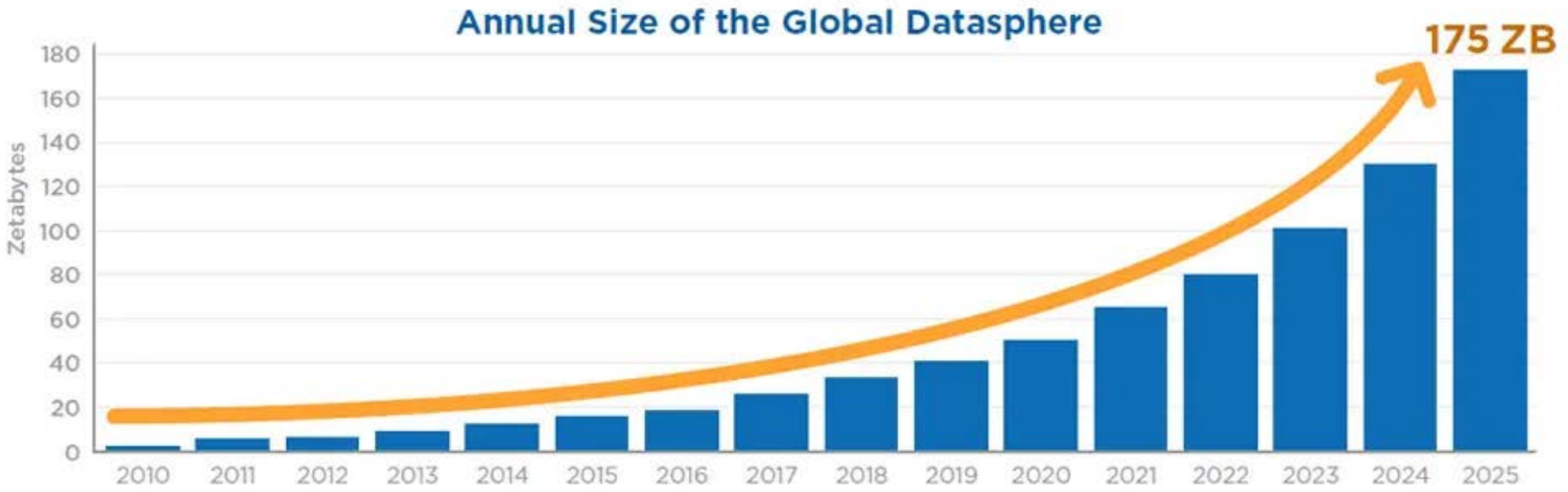
Figure 1 – Annual Size of the Global Datasphere



1 zettabyte =

1.0×10^{21} bytes

Figure 1 – Annual Size of the Global Datasphere



1 zettabyte =

1.0×10^{21} bytes


EXAMPLES

Παραδείγματα της Τεχνητής Νοημοσύνης στον Τουρισμό

1. **Εξατομικευμένη Πρόταση Ταξιδιού**: Πλατφόρμες κρατήσεων και ταξιδιωτικές εφαρμογές χρησιμοποιούν αλγόριθμους μηχανικής μάθησης για να προτείνουν εξατομικευμένες προτάσεις ταξιδιού βάσει των προτιμήσεων του χρήστη, ιστορικού ταξιδιών και άλλων δεδομένων.

2. **Συστήματα Οδήγησης και Πλοήγησης**: Εφαρμογές όπως το Google Maps χρησιμοποιούν τεχνητή νοημοσύνη για να προτείνουν βέλτιστες διαδρομές, να αναγνωρίζουν κίνηση και να παρέχουν προβλέψεις κατά τη διάρκεια του ταξιδιού.

3. **Εικονικοί Βοηθοί και Chatbots**: Τα chatbots και οι εικονικοί βοηθοί μπορούν να χρησιμοποιηθούν για την παροχή πληροφοριών σχετικά με ταξίδια, κρατήσεις, αξιοθέατα και πολλά άλλα, προσφέροντας έτσι εξατομικευμένη εξυπηρέτηση σε πραγματικό χρόνο.

4. **Εκτίμηση Κόστους και Αξιολόγηση**: Πολλές πλατφόρμες ταξιδίων χρησιμοποιούν αλγόριθμους για να εκτιμήσουν το κόστος του ταξιδιού, να ανιχνεύσουν τις καλύτερες τιμές και να παρέχουν αξιολογήσεις και κριτικές για ξενοδοχεία, εστιατόρια και άλλα καταλύματα.

5. **Αναγνώριση Εικόνας και Προορισμών**: Τεχνολογίες όπως η αναγνώριση εικόνας μπορούν να χρησιμοποιηθούν για να αναγνωρίζουν αξιοθέατα, αναγνώριση τοποθεσίας και να παρέχουν πληροφορίες για το περιβάλλον ενός ταξιδιού.

EXPERTS

Πώς η Τεχνητή Νοημοσύνη οδηγεί τον Αεροπορικό Κλάδο στο Μέλλον

Η τεχνητή νοημοσύνη και η μηχανική μάθηση παίζουν ολοένα και πιο καθοριστικό ρόλο στη διαμόρφωση του μέλλοντος του κλάδου των αερομεταφορών

EXPERTS

Πώς η Τεχνητή Νοημοσύνη οδηγεί τον Αεροπορικό Κλάδο στο Μέλλον

Η τεχνητή νοημοσύνη και η μηχανική μάθηση παίζουν ολοένα και πιο καθοριστικό ρόλο στη διαμόρφωση του μέλλοντος του κλάδου των αερομεταφορών

Συμβολή στον τομέα της ασφάλειας

Η τεχνητή νοημοσύνη και η μηχανική μάθηση χρησιμοποιούνται σε μεγάλο βαθμό για να ενισχύσουν την ασφάλεια των αεροσκαφών και των πτήσεων. Οι αλγόριθμοι μηχανικής μάθησης αναλύουν τεράστιες ποσότητες δεδομένων (μεγάλα δεδομένα – big data) που προέρχονται από τις πτήσεις για να εντοπίσουν πιθανούς κινδύνους, δυσλειτουργίες καθώς και να προβλέψουν τις ανάγκες συντήρησης των αεροσκαφών έτσι ώστε τελικώς να βοηθήσουν στην πρόληψη ατυχημάτων.

EXPERTS

Πώς η Τεχνητή Νοημοσύνη οδηγεί τον Αεροπορικό Κλάδο στο Μέλλον

Η τεχνητή νοημοσύνη και η μηχανική μάθηση παίζουν ολοένα και πιο καθοριστικό ρόλο στη διαμόρφωση του μέλλοντος του κλάδου των αερομεταφορών

Βελτιώσεις στην συντήρηση

Τα αεροσκάφη είναι εξοπλισμένα με πολυάριθμους αισθητήρες που παράγουν μεγάλα δεδομένα κατά τη διάρκεια των πτήσεων, όπως αναφέρθηκε και παραπάνω. Τα συστήματα προληπτικής συντήρησης με βάση την τεχνητή νοημοσύνη και την μηχανική μάθηση, μπορούν να αναλύουν αυτά τα δεδομένα σε πραγματικό χρόνο, εντοπίζοντας πιθανά προβλήματα δυσλειτουργίες και αστοχίες, προτού γίνουν αυτά γίνουν κρίσιμα. Αυτό μειώνει την συντήρηση εκτός προγράμματος και ενισχύει την αξιοπιστία και την αξιοπλοΐα των αεροσκαφών.

EXPERTS

Πώς η Τεχνητή Νοημοσύνη οδηγεί τον Αεροπορικό Κλάδο στο Μέλλον

Η τεχνητή νοημοσύνη και η μηχανική μάθηση παίζουν ολοένα και πιο καθοριστικό ρόλο στη διαμόρφωση του μέλλοντος του κλάδου των αερομεταφορών

Συμβολή στην διαχείριση εναέριας κυκλοφορίας

Η τεχνητή νοημοσύνη διαδραματίζει κρίσιμο ρόλο στη βελτιστοποίηση της διαχείρισης της εναέριας κυκλοφορίας. Οι προηγμένοι αλγόριθμοι μπορούν να προβλέψουν και να μετριάσουν τη συμφόρηση, μειώνοντας τις καθυστερήσεις και την κατανάλωση καυσίμων. Τα συστήματα που λειτουργούν με τεχνητή νοημοσύνη και μηχανική μάθησης, υποστηρίζουν επίσης πιο αποτελεσματική δρομολόγηση και προγραμματισμό πτήσεων.

EXPERTS

Πώς η Τεχνητή Νοημοσύνη οδηγεί τον Αεροπορικό Κλάδο στο Μέλλον

Η τεχνητή νοημοσύνη και η μηχανική μάθηση παίζουν ολοένα και πιο καθοριστικό ρόλο στη διαμόρφωση του μέλλοντος του κλάδου των αερομεταφορών

Βελτιώσεις στον σχεδιασμό και την αποδοτικότητα αεροσκαφών

Η αξιοποίηση της τεχνητής νοημοσύνης στον σχεδιασμό των αεροσκαφών είναι καθοριστική καθώς θεωρείται εργαλείο για μεγαλύτερη αεροδυναμική και αποδοτικότερη κατανάλωση καυσίμων. Η μηχανική μάθηση βοηθά επίσης στην βελτιστοποίηση του σχήματος, των υλικών και των συστημάτων των αεροσκαφών για την ελαχιστοποίηση της αντίστασης, τη μείωση των εκπομπών και την αύξηση της συνολικής απόδοσης.

EXPERTS

Πώς η Τεχνητή Νοημοσύνη οδηγεί τον Αεροπορικό Κλάδο στο Μέλλον

Η τεχνητή νοημοσύνη και η μηχανική μάθηση παίζουν ολοένα και πιο καθοριστικό ρόλο στη διαμόρφωση του μέλλοντος του κλάδου των αερομεταφορών

Συνεισφορά στην εμπειρία των επιβατών

Η τεχνητή νοημοσύνη δεν ωφελεί μόνο την αεροπορική βιομηχανία στο επιχειρησιακό και λειτουργικό της κομμάτι. Αντιθέτως, βελτιώνει και την ταξιδιωτική εμπειρία των επιβατών. Από τις εξατομικευμένες προτάσεις και λύσεις ψυχαγωγίας κατά τη διάρκεια της πτήσης έως τα chatbots με τεχνητή νοημοσύνη για την εξυπηρέτηση αυτών, οι επιβάτες ήδη απολαμβάνουν τα οφέλη των υπηρεσιών με τεχνητή νοημοσύνη. Ενδεικτικό είναι το παράδειγμα της Emirates, που τα τελευταία χρόνια έχει εφαρμόσει check – in ανέπαφα και οπτικά με την χρήση βιομετρικών στοιχείων.

EXPERTS

Πώς η Τεχνητή Νοημοσύνη οδηγεί τον Αεροπορικό Κλάδο στο Μέλλον

Η τεχνητή νοημοσύνη και η μηχανική μάθηση παίζουν ολοένα και πιο καθοριστικό ρόλο στη διαμόρφωση του μέλλοντος του κλάδου των αερομεταφορών

Πρόγνωση καιρού και πρόβλεψη αναταράξεων

Ο καιρός είναι ένας σημαντικός παράγοντας που επηρεάζει τις αερομεταφορές. Η τεχνητή νοημοσύνη και η μηχανική μάθηση βελτιώνουν την ακρίβεια της πρόβλεψης του καιρού, επιτρέποντας στους πιλότους να λαμβάνουν τεκμηριωμένες αποφάσεις. Μπορεί επίσης, να προβλέψει και να μετριάσει τις αναταράξεις, ενισχύοντας την άνεση και την ασφάλεια των επιβατών.

EXPERTS

Πώς η Τεχνητή Νοημοσύνη οδηγεί τον Αεροπορικό Κλάδο στο Μέλλον

Η τεχνητή νοημοσύνη και η μηχανική μάθηση παίζουν ολοένα και πιο καθοριστικό ρόλο στη διαμόρφωση του μέλλοντος του κλάδου των αερομεταφορών

Ασφάλεια και ανίχνευση απάτης

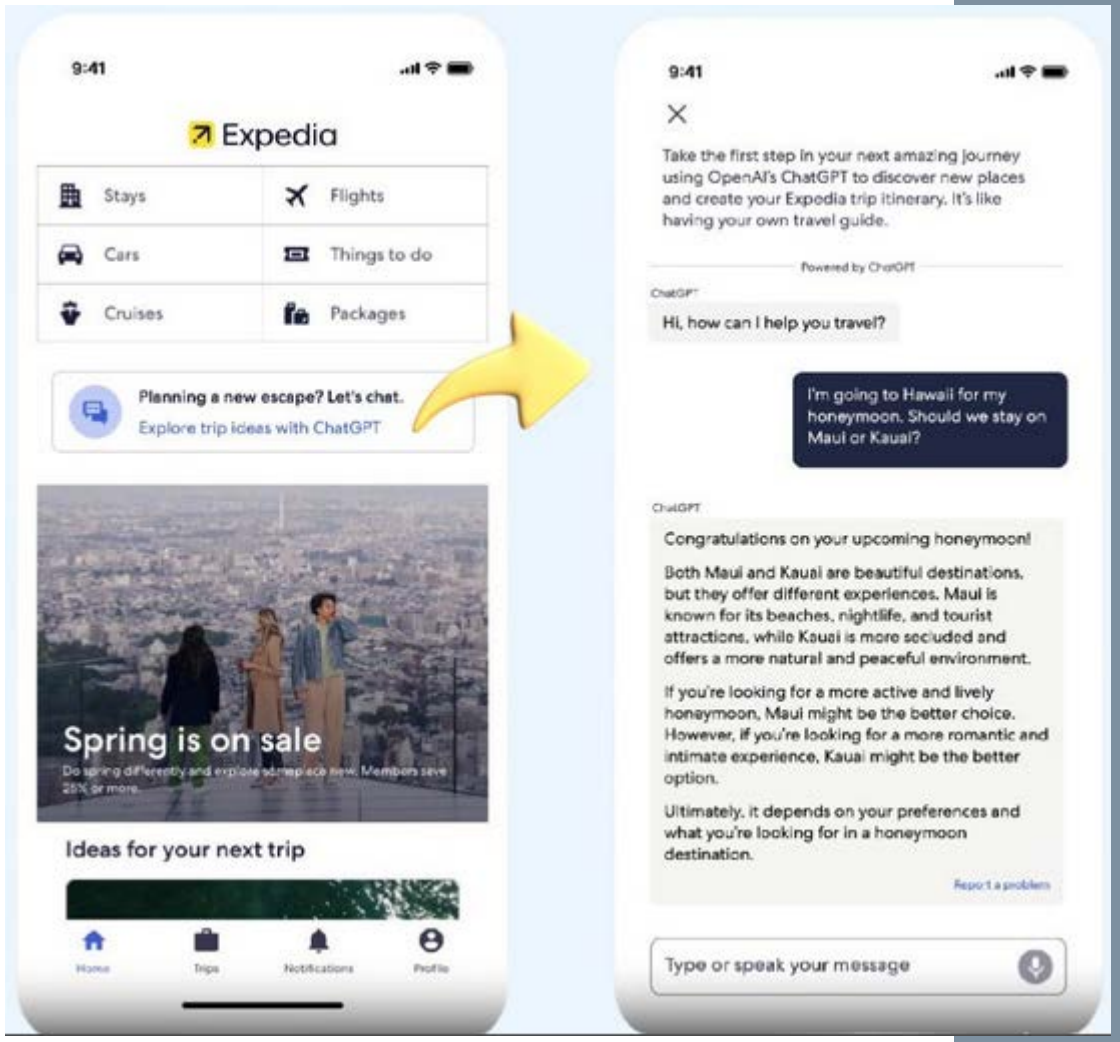
Η τεχνητή νοημοσύνη χρησιμοποιείται για την ενίσχυση των μέτρων ασφαλείας στους αερολιμένες, τον εντοπισμό πιθανών απειλών και τη βελτίωση των διαδικασιών ελέγχου των επιβατών. Επιπλέον, χρησιμοποιείται στην ανίχνευση απάτης στις συναλλαγές για τις αεροπορικές εταιρείες, μειώνοντας τον αντίκτυπο των υποκλοπών στοιχείων και ενεργειών με δόλο. Αρκετές αεροπορικές εταιρείες σε όλο τον κόσμο έχουν ενσωματώσει ενεργά την τεχνητή νοημοσύνη (AI) στις δραστηριότητές τους για να βελτιώσουν την αποτελεσματικότητα, την εξυπηρέτηση των πελατών, την ασφάλεια και άλλες πτυχές του οργανισμού.

Chatbots για διαδικτυακή εξυπηρέτηση πελατών

Οι μεγαλύτερες εταιρείες στον κλάδο των ταξιδιωτικών κρατήσεων όπως

- η Booking,
- η Skyscanner και
- η Expedia

έχουν υιοθετήσει τέτοια chatbots, προκειμένου να βελτιώσουν την εμπειρία των πελατών τους.



Εφαρμογές AI για πρόβλεψη πτήσεων

- Η τεχνητή νοημοσύνη στην τουριστική βιομηχανία, έχει σημαντικές εφαρμογές και στην πρόβλεψη πτήσεων.

- Μέσα από την πρόσβαση σε μεγάλο όγκο δεδομένων και ιστορικά δεδομένα πτήσης, η τεχνητή νοημοσύνη μπορεί να είναι ανεκτίμητη για την πραγματοποίηση προβλέψεων και τον εντοπισμό τάσεων που σχετίζονται για παράδειγμα με την ταχύτητα και διάρκεια πτήσης.

- Έτσι, μπορεί να προβλέψει με ακρίβεια πότε θα φτάσουν οι πτήσεις και πόσο καιρό είναι πιθανό να διαρκέσουν οι καθυστερήσεις.

- Ακόμη, προγνώσεις μπορούν να πραγματοποιηθούν και για πρόβλεψη του προς τα που κατευθύνονται οι τιμές των αεροπορικών εταιρειών, ώστε οι χρήστες να γνωρίζουν πότε τους συμφέρει καλύτερα να προβούν σε αγορά εισιτηρίου.

Χρήση Voice Assistants

- Οι voice assistants (φωνητικοί βοηθοί) χρησιμοποιούνται ήδη σε μεγάλο βαθμό σε δωμάτια ξενοδοχείων, πλοία αλλά και για την ασφάλεια αεροδρομίων.

- Αξιοποιούν την επεξεργασία φυσικής γλώσσας (NLP) και είναι ιδιαίτερα σημαντικοί μιας και επιτρέπουν στους επισκέπτες σε κάποιο ξενοδοχείο για παράδειγμα, να κάνουν ερωτήσεις ή να υποβάλλουν αιτήματα και να λαμβάνουν απάντηση άμεσα και όλο το εικοσιτετράωρο.

- Ένα τέτοιο παράδειγμα αποτελεί το PolyAI Voice Assistant.

ΤΕΧΝΟΛΟΓΙΑ #ChatGPT #Τεχνητή νοημοσύνη #ταξιδιωτικοί πράκτορες #Τουρισμός #Τεχνολογία

ChatGPT και τεχνητή νοημοσύνη αλλάζουν ήδη τον τρόπο με τον οποίο κάνουμε διακοπές

Expedia, Kayak και Trip.com αναπτύσσουν plug-ins ως «εικονικούς βοηθούς» - Τα αεροδρόμια χρησιμοποιούν την τεχνητή νοημοσύνη για να βελτιώνουν τις υπηρεσίες τους

Σπύρος Αλεξίου

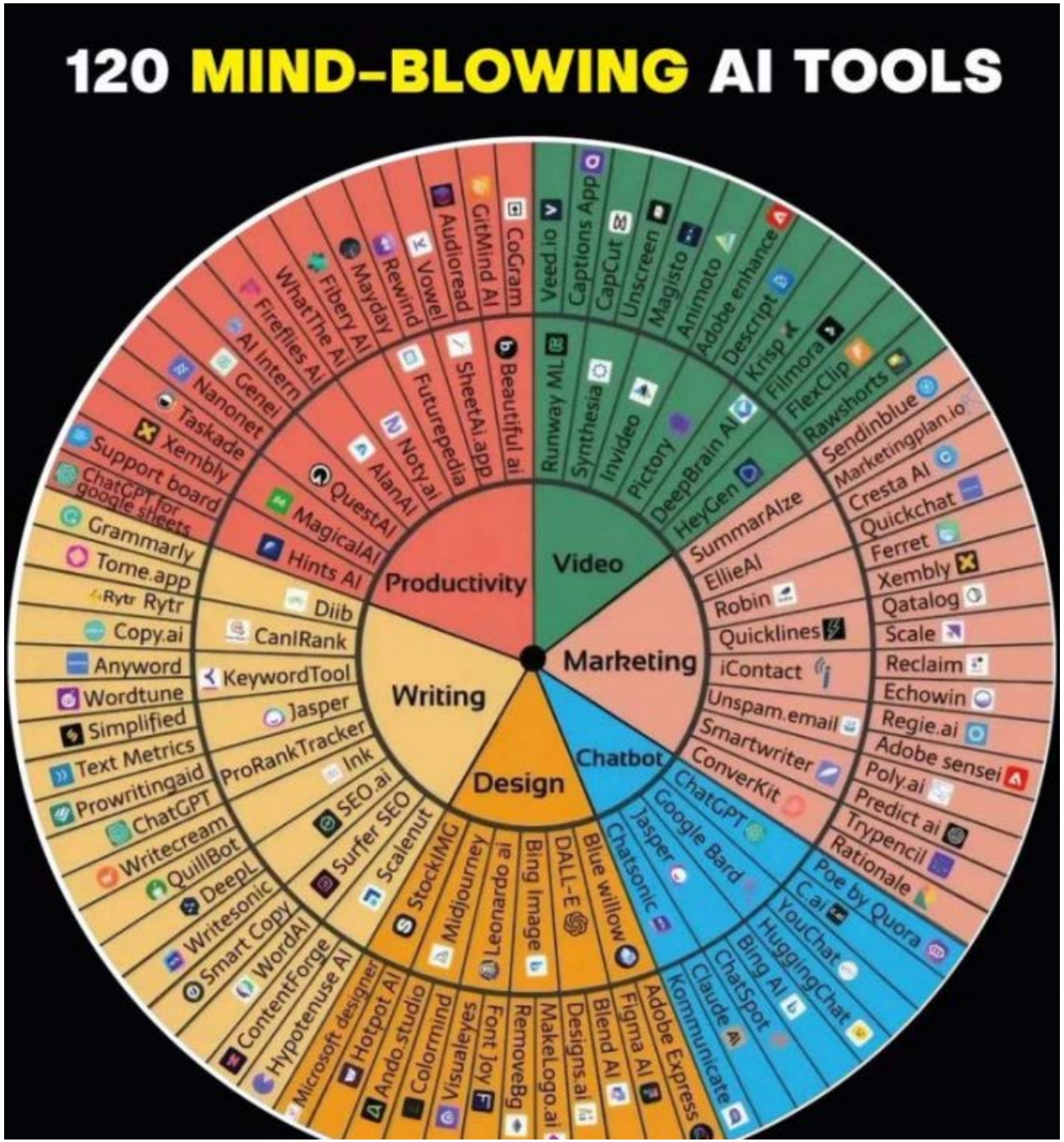
21.05.2023, 06:40

EXAMPLES OF AI IN TRAVEL

Τουρισμός: Τι φοβούνται οι πελάτες σχετικά με την τεχνητή νοημοσύνη στα ξενοδοχεία

Τουρισμός: Τι φοβούνται οι πελάτες σχετικά με την τεχνητή νοημοσύνη στα ξενοδοχεία

- Απώλειας Ανθρώπινης Επαφής
- Παραβίασης της Ιδιωτικότητας
- Τεχνικών Αποτυχιών
- Ανεπιθύμητης Παρακολούθησης
- Αναποτελεσματικότητας ή Μη Αξιοπιστίας


120 MIND-BLOWING AI TOOLS
Productivity
Video
Marketing
Writing
Design
Chatbot
CoGram
GitMind AI
Audioread
Vowel
Rewind
Mayday
Fibery AI
WhatThe AI
Fireflies AI
AI Intern
Genei
Nanonet
Taskade
Xembly
Support board
ChatGPT for google sheets
Grammarly
Tome.app
Rytr
Copy.ai
Anyword
Wordtune
Simplified
Text Metrics
Prowritingaid
ChatGPT
Writecream
QuillBot
DeepL
Writesonic
Smart Copy
WordAI
ContentForge
Hypotenuse AI
Beautiful ai
SheetAI.app
Futurepedia
Noty.ai
AlanAI
QuestAI
MagicalAI
Hints AI
Diib
CanIRank
KeywordTool
Jasper
ProRankTracker
Ink
SEO.ai
Surfer SEO
Scalenut
StockIMG
Midjourney
Leonardo ai
Bing Image
DALL-E
Blue willow
Microsoft designer
Hotpot AI
Ando.studio
Colormind
Visualeyes
Font Joy
RemoveBg
MakeLogo.ai
Designs.ai
Figma AI
Blend AI
Adobe Express
Veed.io
Captions App
CapCut
Unscreen
Magisto
Animoto
Adobe enhance
Descript
Krisp
Filmora
FlexClip
Rawshorts
Runway ML
Synthesia
Invideo
Pictory
DeepBrain AI
HeyGen
SummarAIze
EllieAI
Robin
Quicklines
iContact
Unspam.email
Smartwriter
ConverKit
Sendinblue
Marketingplan.io
Cresta AI
Quickchat
Ferret
Xembly
Qatalog
Scale
Reclaim
Echowin
Regie.ai
Adobe sensei
Poly.ai
Predict ai
Trypencil
Rationale
ChatGPT
Google Bard
Jasper
Chatsonic
Poe by Quora
C.ai
YouChat
HuggingChat
Bing AI
ChatSpot
Claude AI
Kommunicate


CHAT GPT
AI

Κίνδυνοι – Ηθικά Διλήμματα

Αντικατάσταση Βασικών Αναπτυξιακών Δεξιοτήτων από ΤΝ στον Άνθρωπο:

Έλλειψη κριτικής σκέψης

- Ανεπαρκής διαπροσωπικές δεξιότητες
- Μειωμένη δημιουργικότητα
- Εξάρτηση από τεχνολογία
- Ανικανότητα αυτο-ρύθμισης και αυτο-ελέγχου
- Χαμηλή αυτο-αποτίμηση και αυτο-αποδοχή

ΣΩΣΤΗ ΧΡΗΣΗ ΤΗΣ ΤΝ

Κίνδυνοι – Ηθικά Διλήμματα

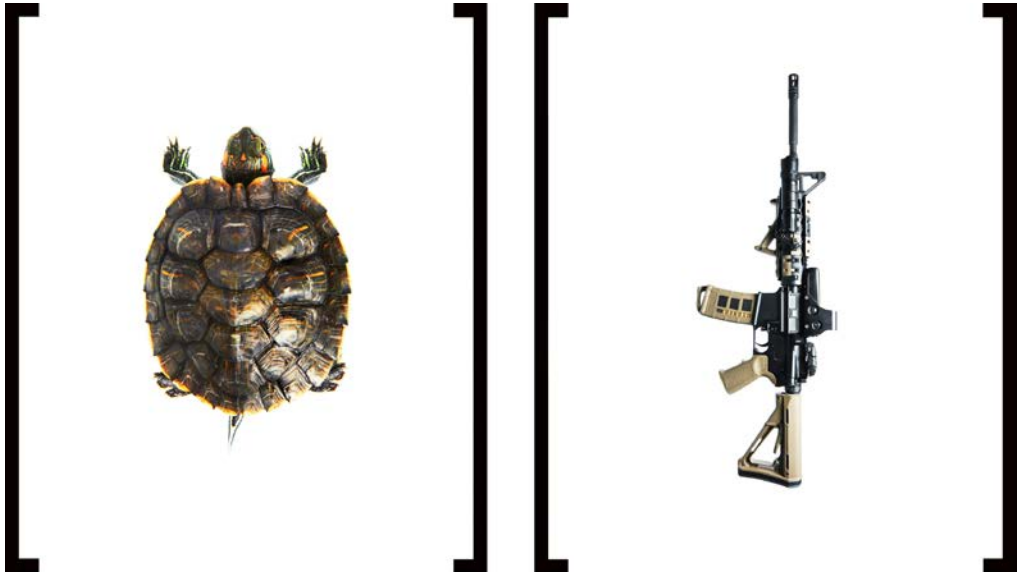
Chihuahua or muffin

Goldendoodle or Fried Chicken

Κίνδυνοι – Ηθικά Διλήμματα

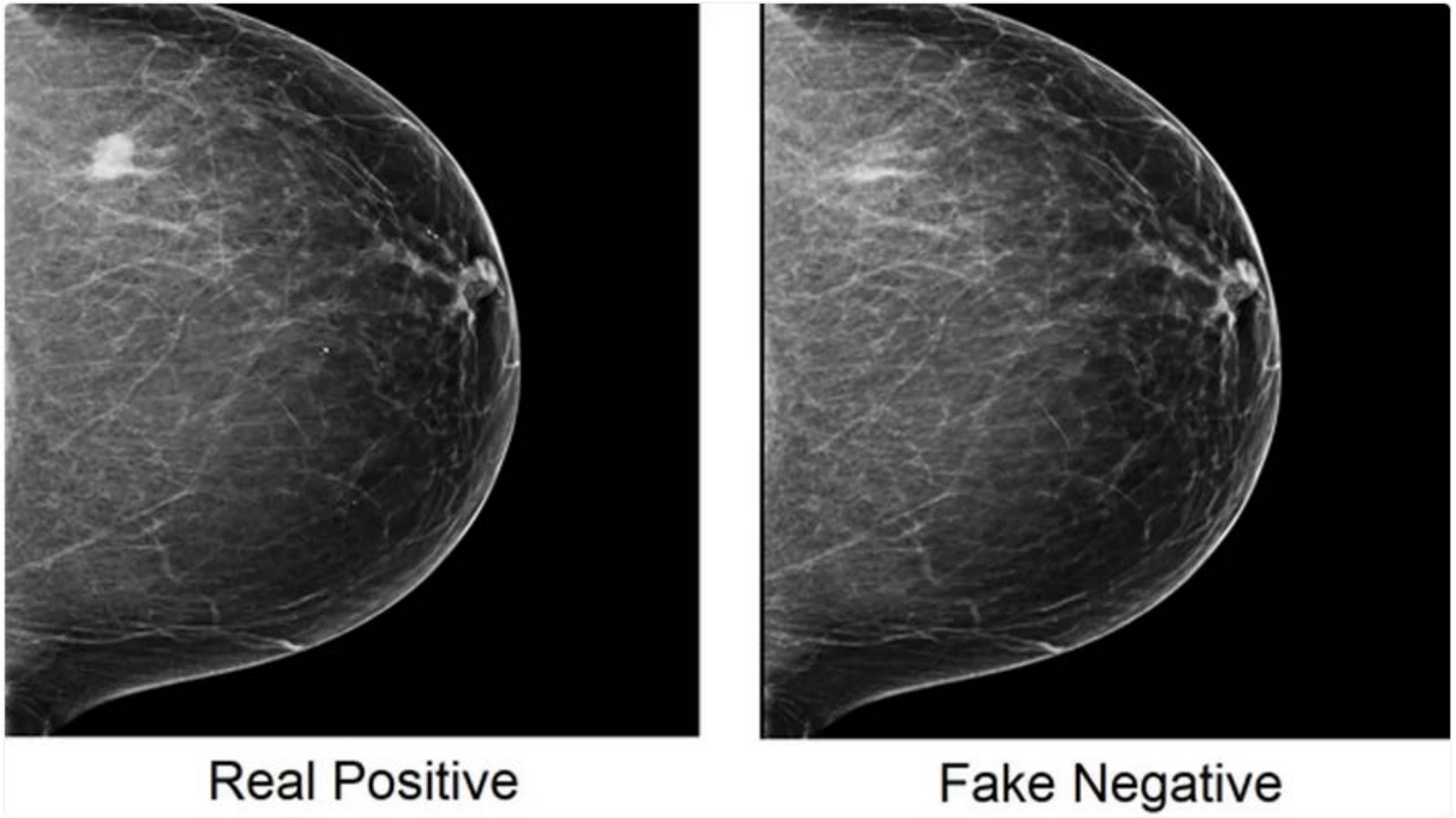
Google's AI thinks this turtle looks like a gun, which is a problem

New research shows how machine vision systems of all kinds can be tricked into misidentifying 3D objects



Cancer-Spotting AI Is Vulnerable To Cyberattacks

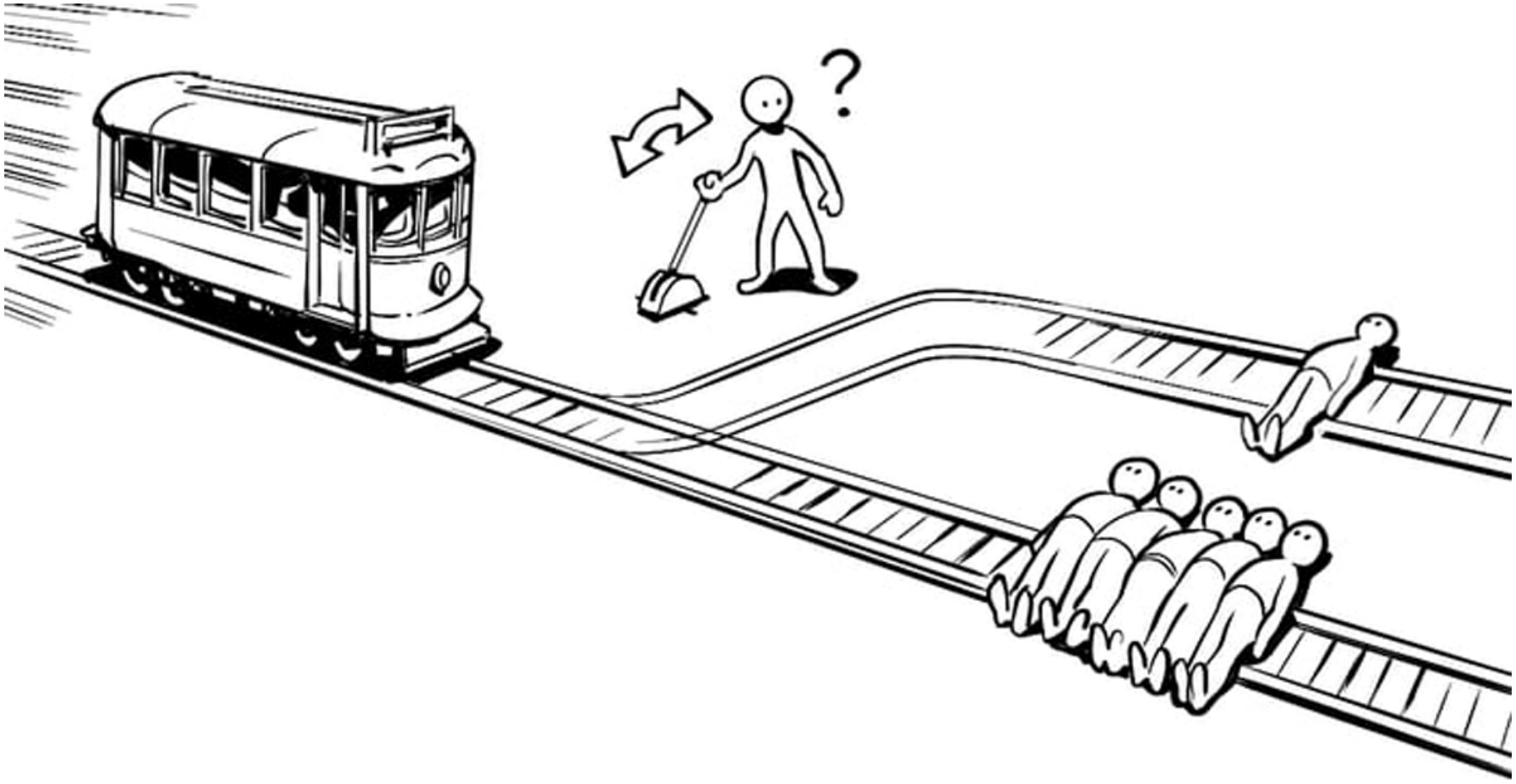
News Published: December 15, 2021 | Original story from University of Pittsburgh



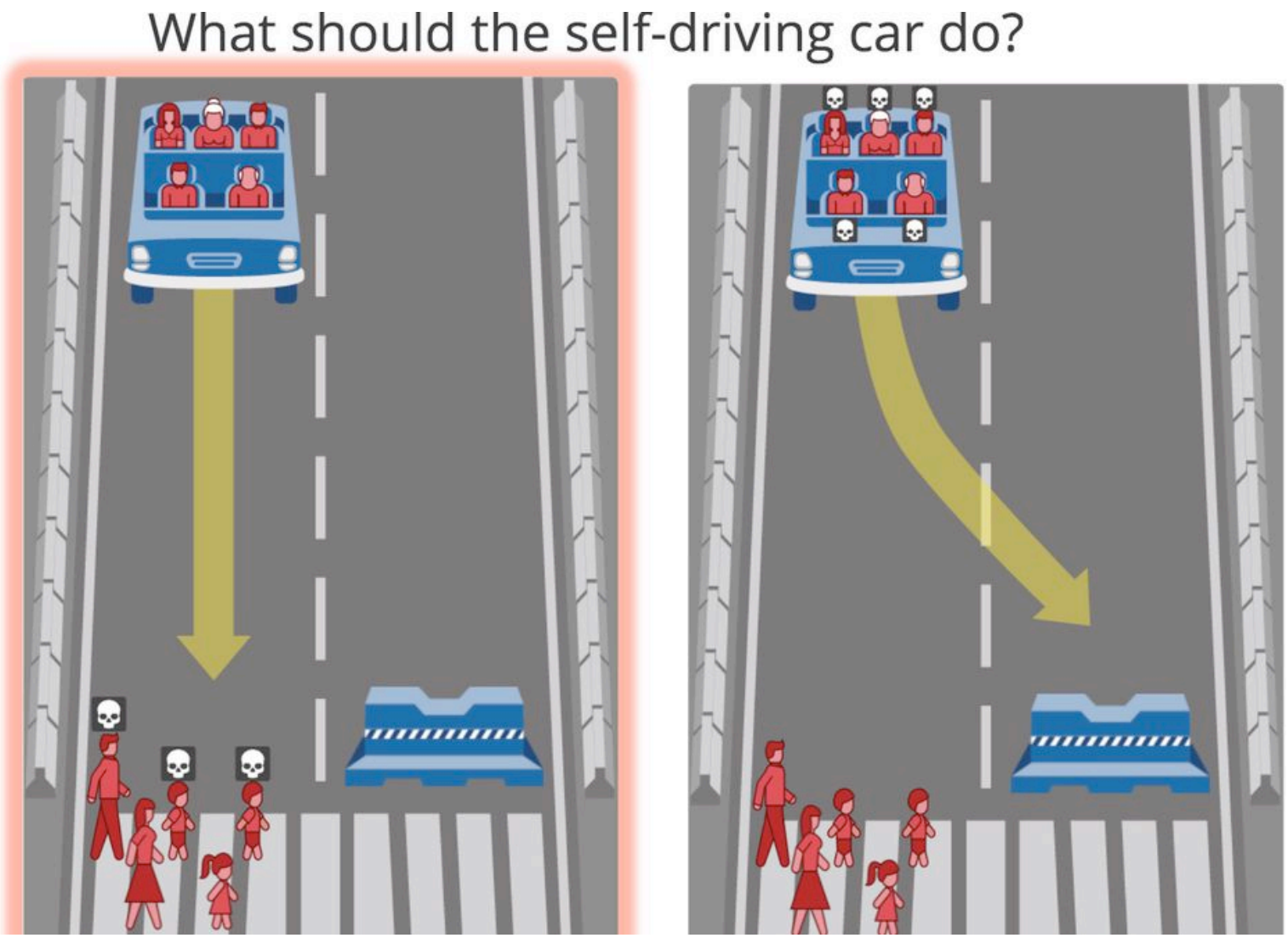
Mammogram images showing a real cancer-positive (left) case, with cancerous tissue indicated by white spot. A 'generative adversarial network' program removed cancerous regions from the cancer-positive image, creating a fake negative image (right). Credit: Q. Zhou et al., Nat. Commun. 2021

Ethics Challenges in AI

The Trolley Dilemma



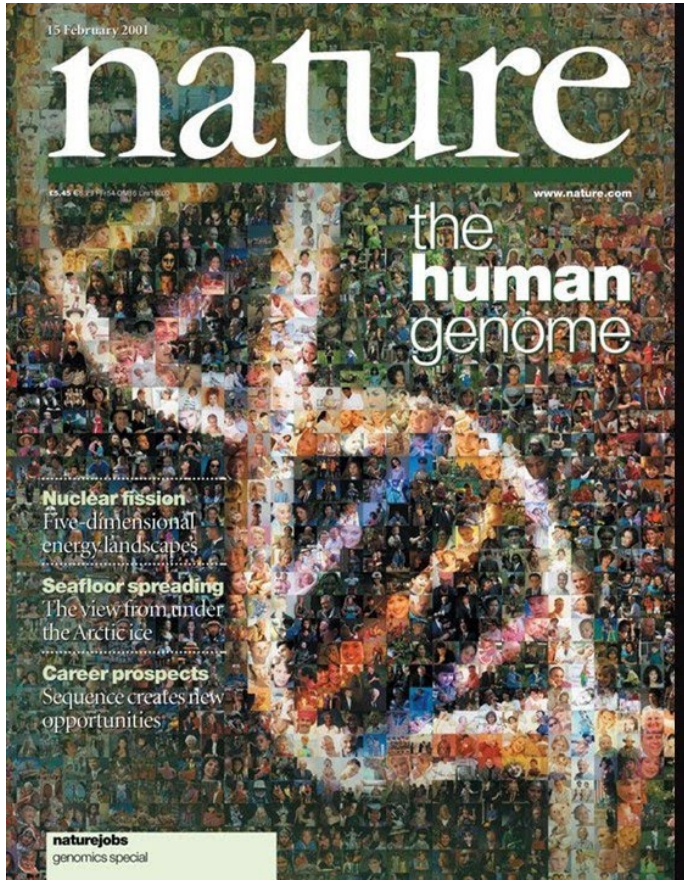
Κίνδυνοι – Ηθικά Διλήμματα

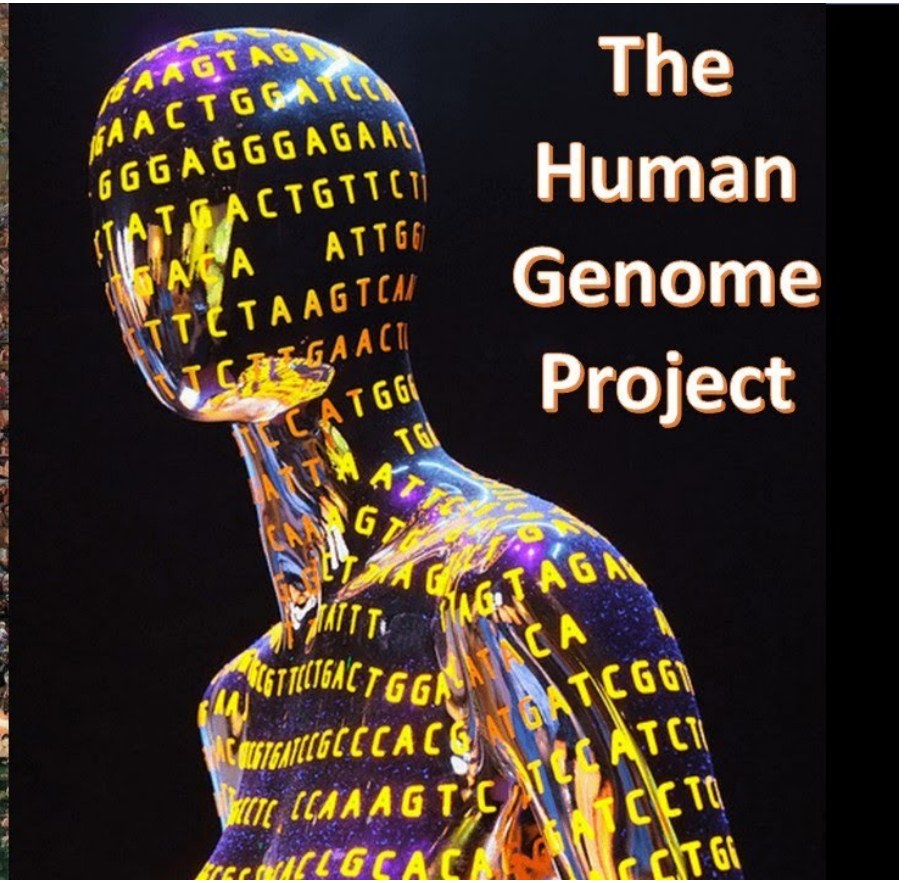


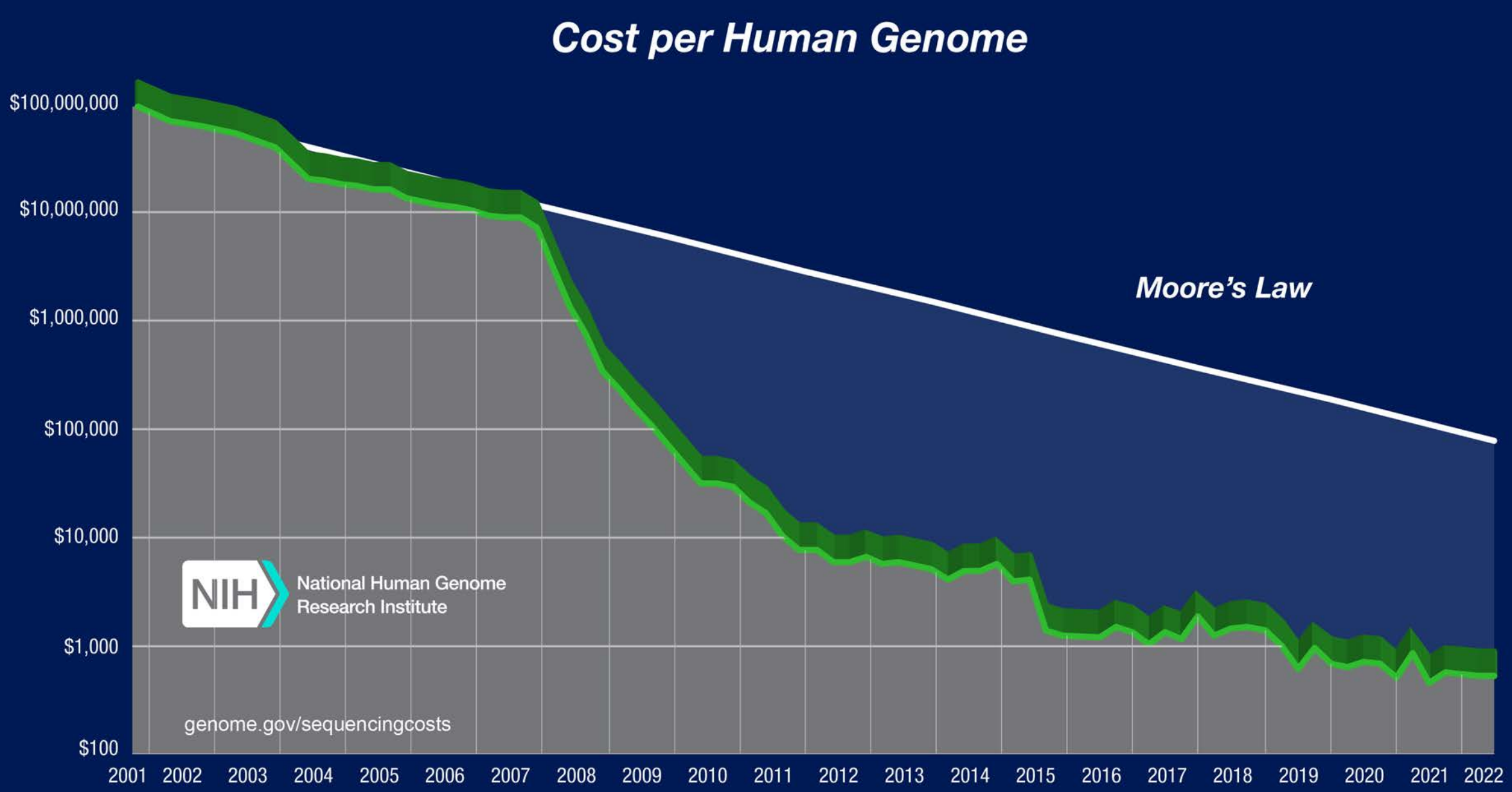
Ethics of Using Smart City/Home AI

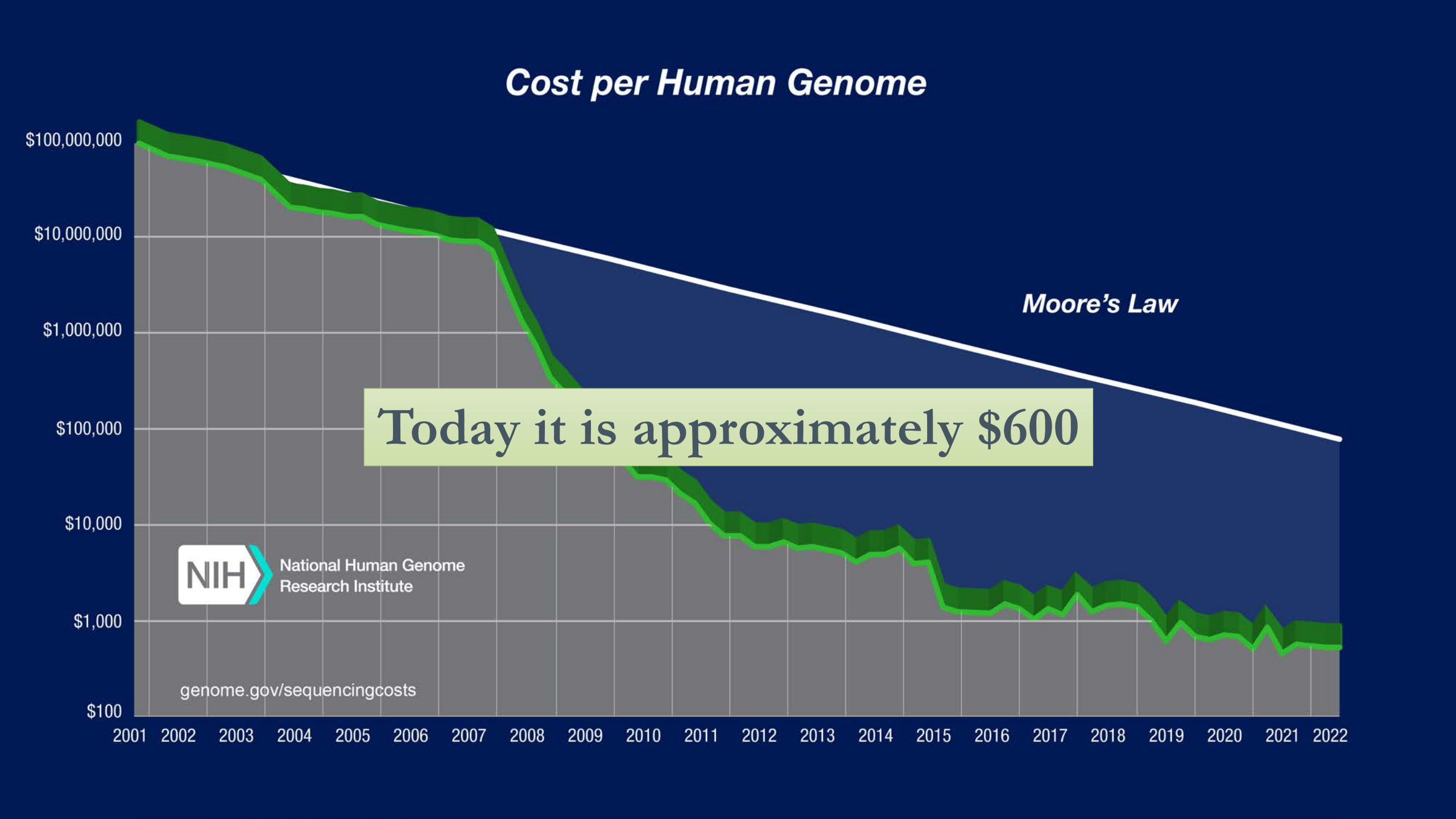
Human Genome Project (April 2003)

Μία διεθνής ερευνητική προσπάθεια χαρτογράφησης κάθε ανθρώπινου γονιδίου και αλληλουχίας των 3,1 δισεκατομμυρίων βάσεων που συνιστούν το ανθρώπινο DNA

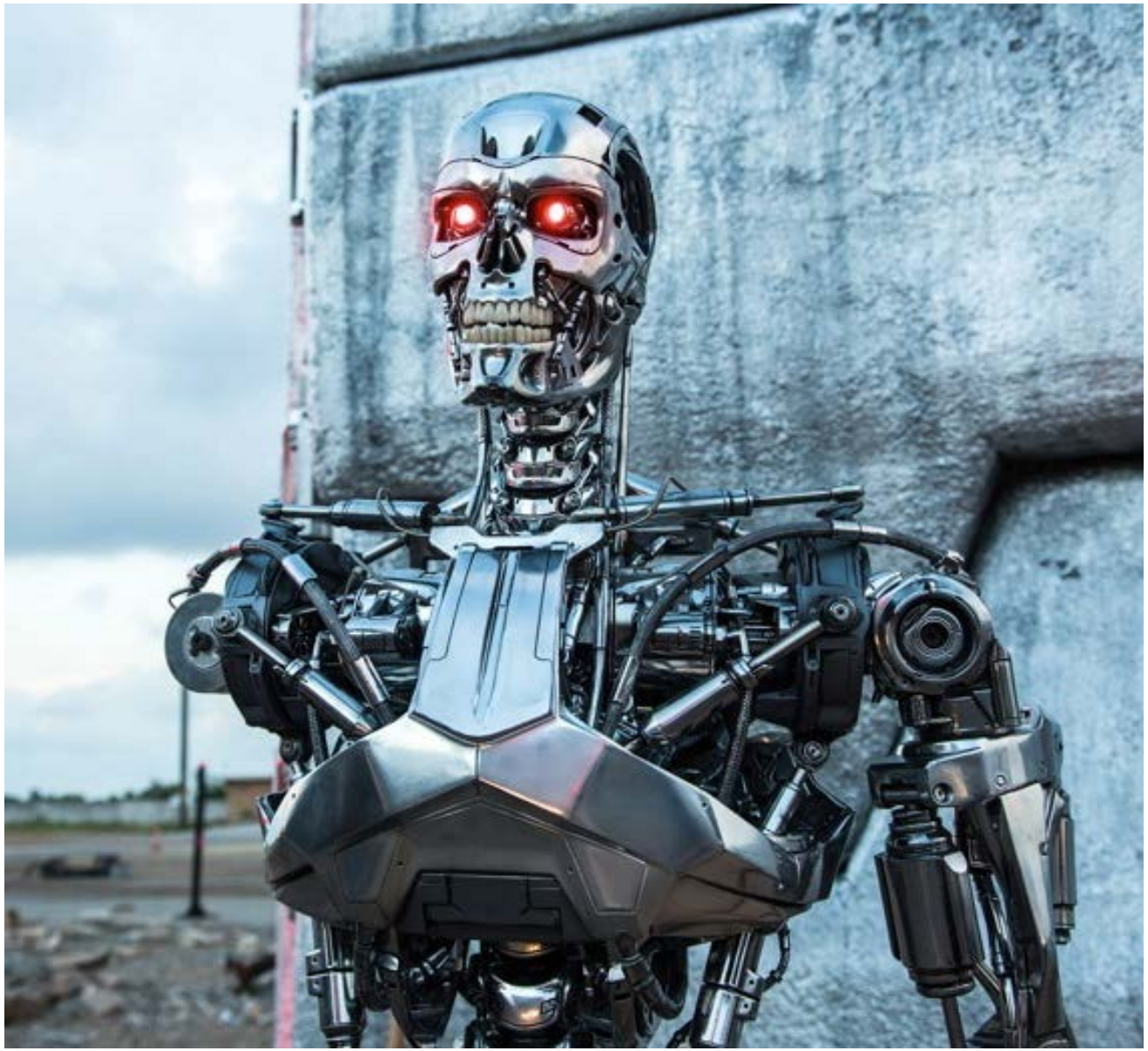


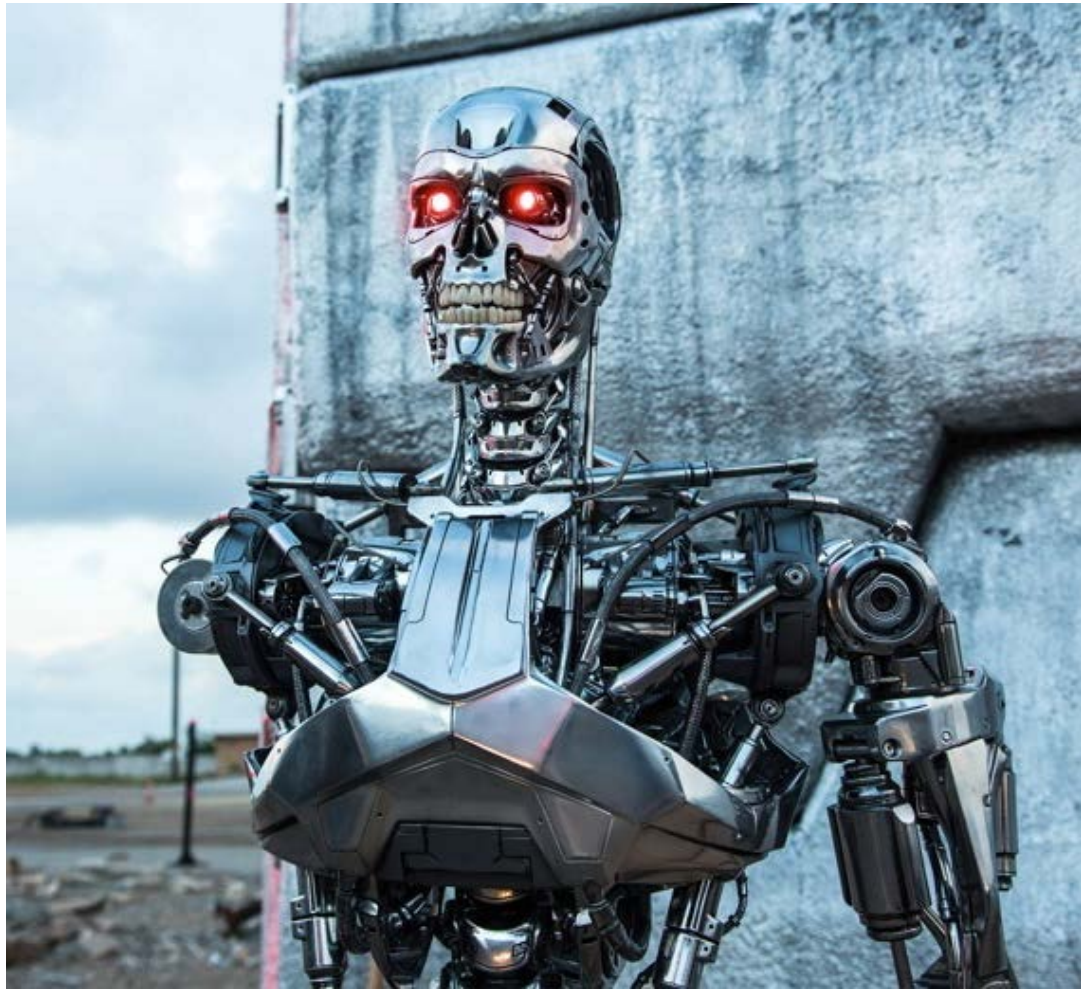



Cost per Human Genome
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
$1,000
$100
Moore's Law
NIH
National Human Genome
Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022

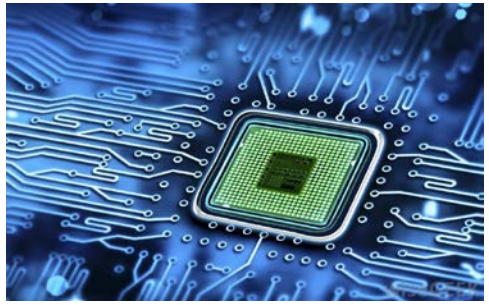

Cost per Human Genome
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
$1,000
$100
Moore's Law
Today it is approximately $600
NIH
National Human Genome
Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022

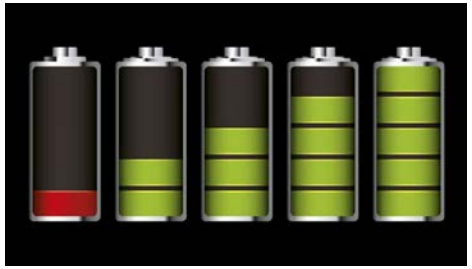
ΘΑ ΜΑΣ ΚΑΤΑΚΤΗΣΕΙ ΤΟ ΑΙ ???

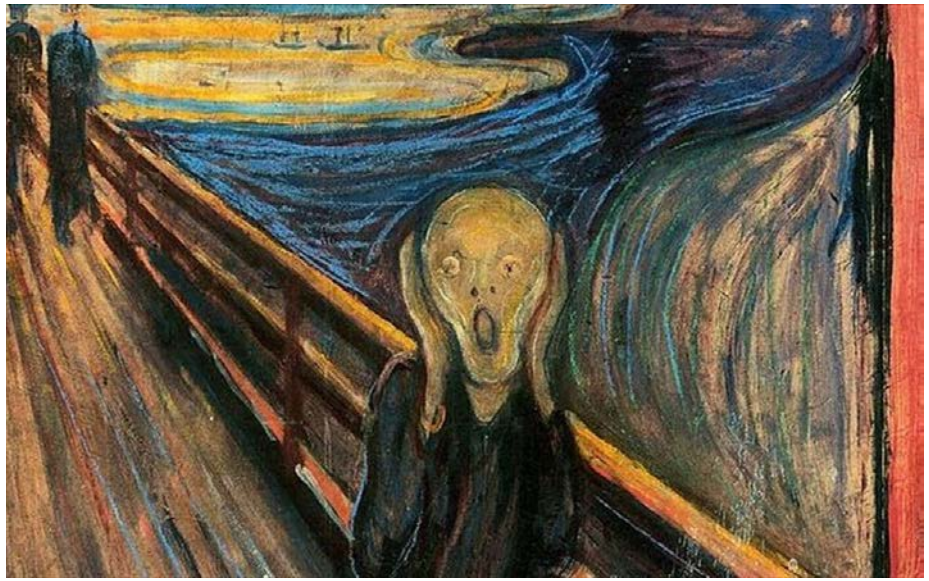


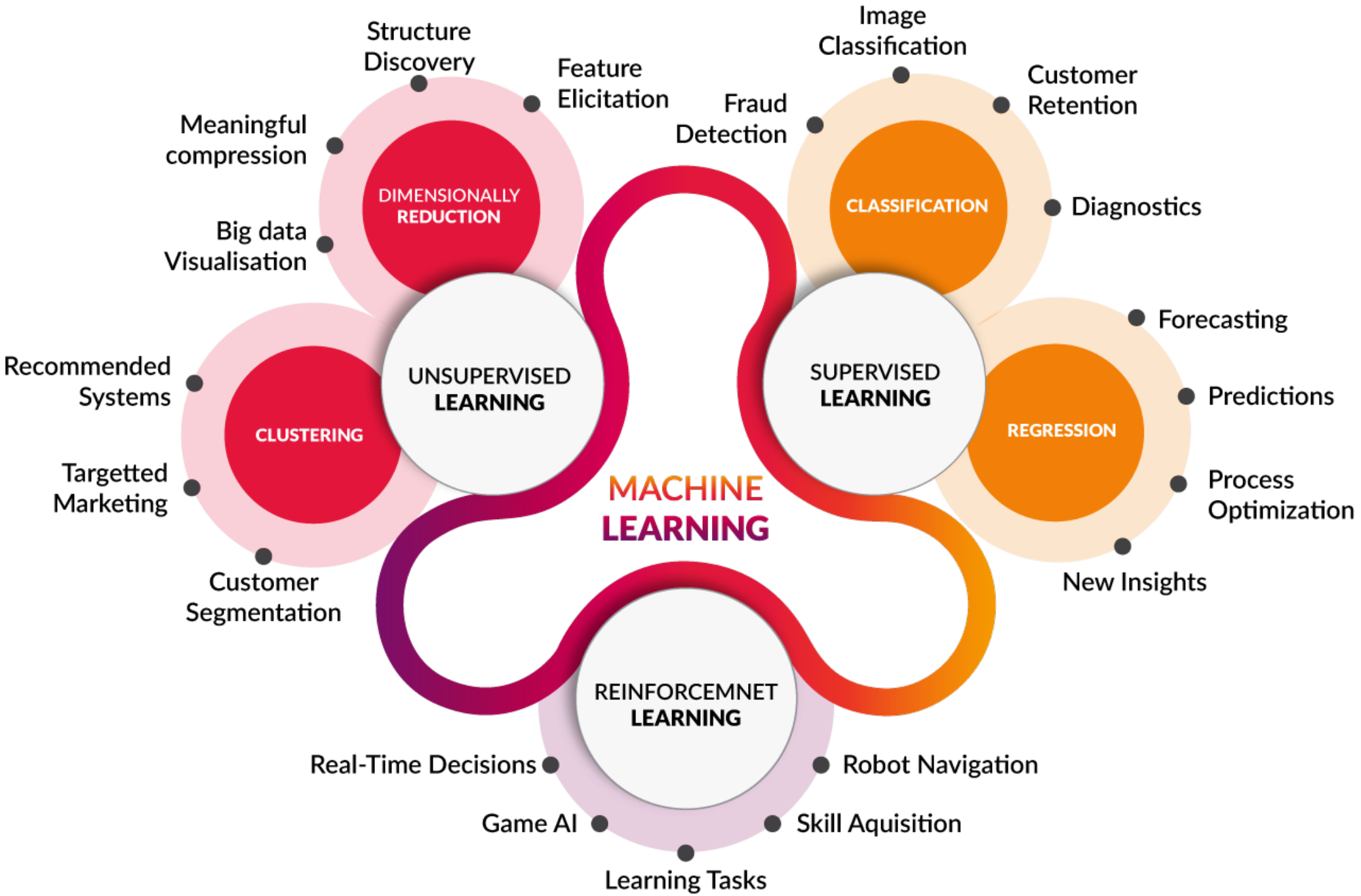



120mm





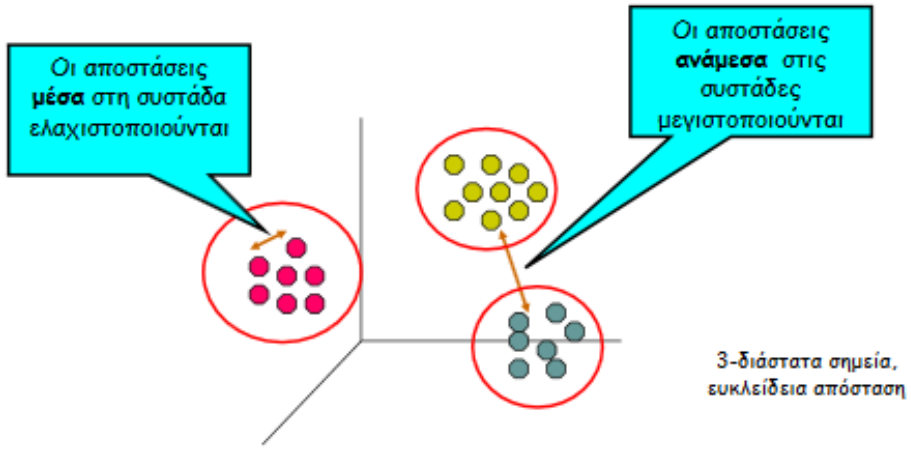



MACHINE LEARNING
UNSUPERVISED LEARNING
SUPERVISED LEARNING
REINFORCEMNET LEARNING
DIMENSIONALLY REDUCTION
CLUSTERING
CLASSIFICATION
REGRESSION
Structure Discovery
Feature Elicitation
Meaningful compression
Big data Visualisation
Recommended Systems
Targetted Marketing
Customer Segmentation
Image Classification
Customer Retention
Fraud Detection
Diagnostics
Forecasting
Predictions
Process Optimization
New Insights
Real-Time Decisions
Robot Navigation
Game AI
Skill Aquisition
Learning Tasks

Συσταδοποίηση - Clustering

- Είναι η διαδικασία της κατηγοριοποίησης των δεδομένων σε σύνολα ομοειδών αντικειμένων καλούμενα ομάδες (clusters)

- Στόχος
  - Να παράγει ένα σύνολο από ομάδες με υψηλή εντός των ομάδων ομοιότητα (intra-cluster similarity), ενώ παράλληλα να διατηρείται χαμηλή η ομοιότητα μεταξύ των διαφόρων ομάδων (inter-cluster similarity)



- Εφαρμογές
  - Ευρύ φάσμα εφαρμογών, από τις κοινωνικές επιστήμες, την οικονομία, την αναγνώριση προτύπων έως την βιοπληροφορική, την αστροφυσική και σεισμολογία

Ομαδοποίηση - Clustering

- Well Separated
  - μία συστάδα είναι το σύνολο των αντικειμένων όπου κάθε αντικείμενο είναι πιο κοντά σε κάθε άλλο αντικείμενο της συστάδας, από ότι σε κάποιο άλλο αντικείμενο.

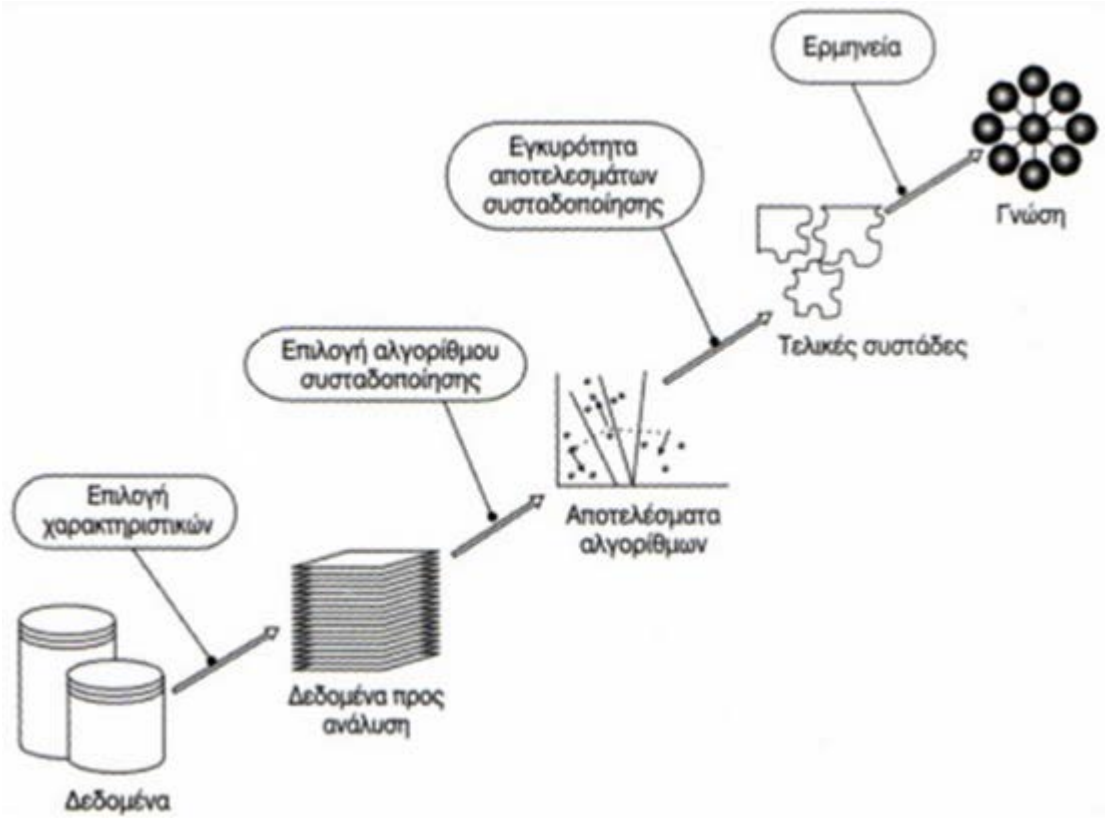
- Prototype Based
  - μία συστάδα είναι τα αντικείμενα που είναι πιο κοντά σε ένα πρωτότυπο (prototype) από ότι κάποιο άλλο αντικείμενο. Συνήθως σαν πρωτότυπο επιλέγεται το μέσο των σημείων μίας συστάδας.

- Graph Based
  - μία συνεκτική συνιστώσα ή μία κλίκα του γραφήματος.

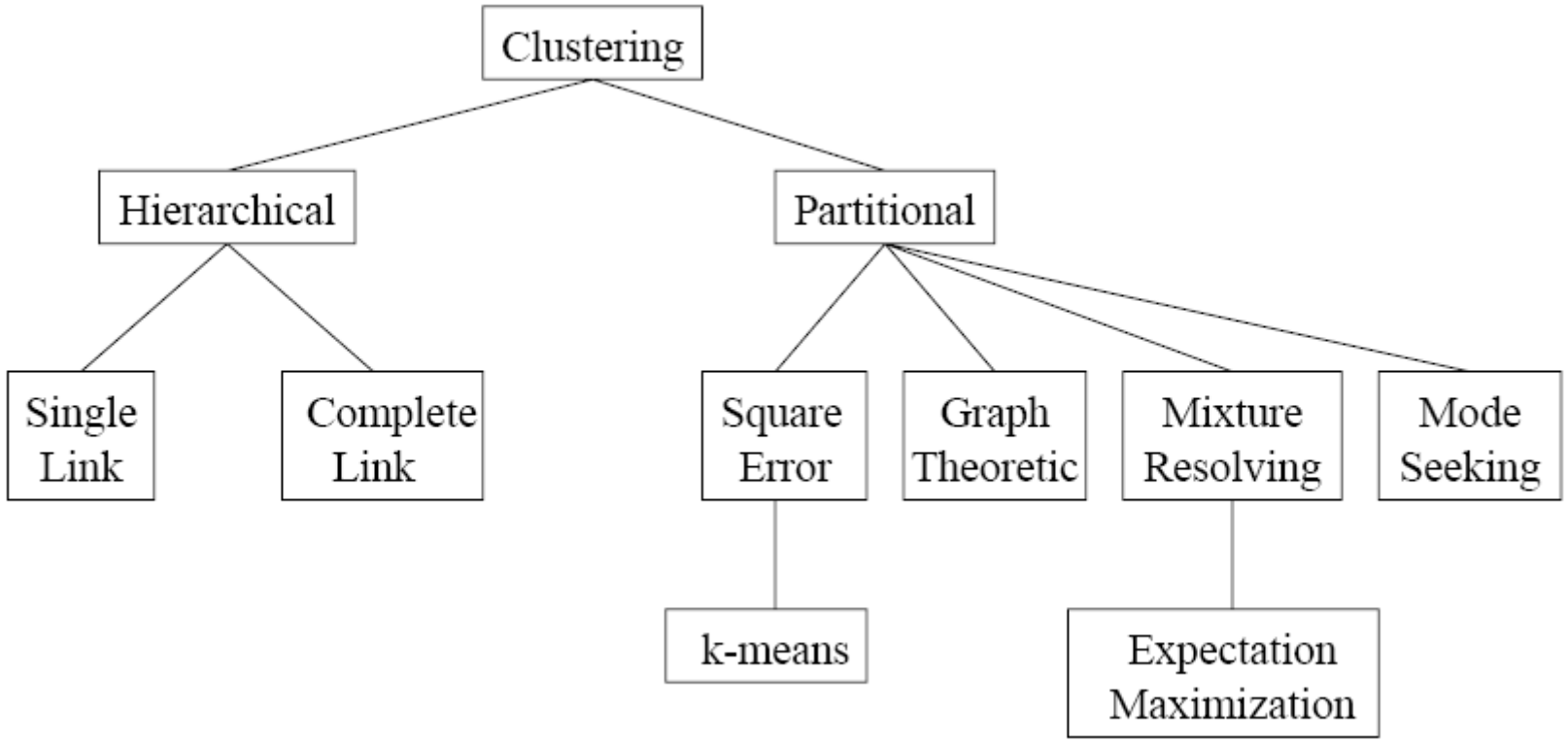
- Density Based
  - μία πυκνή περιοχή αντικειμένων που περιβάλλεται από μία αραιή

- Shared Property (conceptual clusters)
  - σύνολο αντικειμένων που μοιράζονται μία ιδιότητα – έχει εφαρμογή κυρίως σε κατηγορικά αντικείμενα

Βήματα Διαδικασίας Συσταδοποίησης



Κατηγοριοποίηση των Αλγορίθμων Ομαδοποίησης



Είδη Ομαδοποίησης

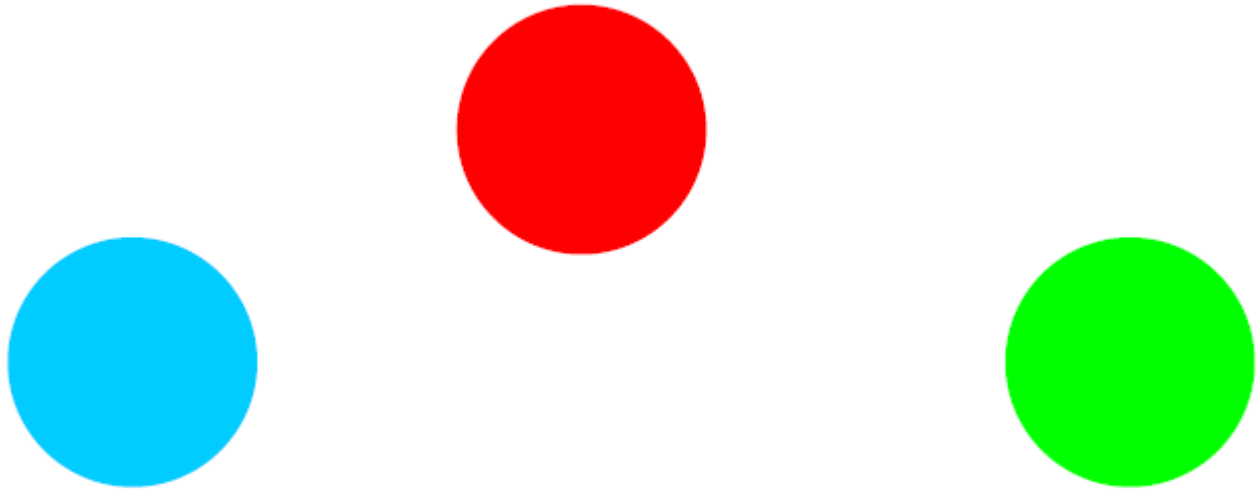
- Βασική διάκριση ανάμεσα στο ιεραρχικό (hierarchical) και διαχωριστικό (partitional) σύνολο από ομάδες

- Διαχωριστική Συσταδοποίηση (Partitional Clustering)
  - Ένας διαμερισμός των αντικειμένων σε μη επικαλυπτόμενα -non-overlapping - υποσύνολα (συστάδες) τέτοιος ώστε κάθε αντικείμενο ανήκει σε ακριβώς ένα υποσύνολο

- Ιεραρχική Συσταδοποίηση (Hierarchical clustering)
  - Ένα σύνολο από εμφωλευμένες (nested) ομάδες Επιτρέπουμε σε μια συστάδα να έχει υπο-συστάδες οργανωμένες σε ένα ιεραρχικό δέντρο

Τύποι συστάδων: Καλώς Διαχωρισμένες Συστάδες

Μια συστάδα είναι ένα σύνολο από σημεία τέτοια ώστε κάθε σημείο μιας συστάδας είναι κοντινότερο σε (ή πιο όμοιο με) όλα τα άλλα σημεία της συστάδας από ότι σε οποιοδήποτε άλλο σημείο που δεν ανήκει στη συστάδα.



3 καλώς-διαχωρισμένες συστάδες

Συχνά υπάρχει η έννοια του κατωφλιού (threshold)

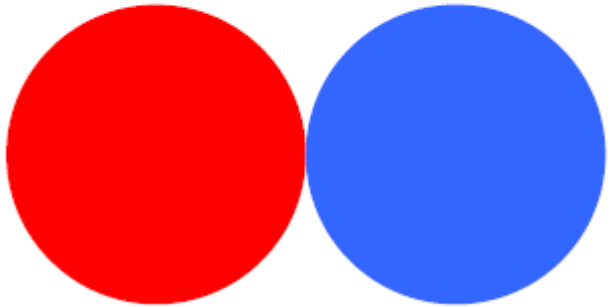
Όχι απαραίτητα κυκλικοί (οποιοδήποτε σχήμα)

Τύποι συστάδων: Συστάδες βασισμένες σε κέντρο ή πρότυπο

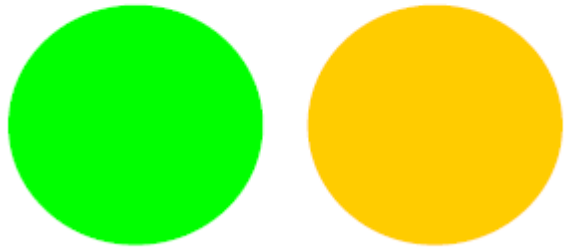
Μια συστάδα είναι ένα σύνολο από αντικείμενα τέτοιο ώστε ένα αντικείμενο στην συστάδα είναι κοντινότερο σε (ή πιο όμοιο με) το «κέντρο» ή πρότυπο της συστάδας από ότι από το κέντρο οποιασδήποτε άλλης συστάδας.

Το κέντρο της ομάδας είναι συχνά

- centroid, ο μέσος όρος των σημείων της συστάδας, ή
- a medoid, το πιο «αντιπροσωπευτικό» σημείο της συστάδας (πχ όταν κατηγορικά γνωρίσματα)



4 συστάδες βασισμένες σε κέντρο



Τείνουν στο να είναι κυκλικοί

Τύποι συστάδων: Συνεχής Συστάδες

Συνεχής Συστάδες (Contiguous Cluster) (Κοντινότερος γείτονα ή μεταβατικά) – Βάσει γειτνίασης

Μια συστάδα είναι ένα σύνολο σημείων τέτοιο ώστε κάθε σημείο είναι πιο κοντά σε ένα ή περισσότερα σημεία της συστάδας από ό,τι σε οποιοδήποτε άλλο σημείο εκτός συστάδας

Συχνά σε περιπτώσεις συστάδων με μη κανονικό σχήμα ή με αλληλοπλεκόμενα σχήματα – ή όταν έχουμε γραφήματα και θέλουμε να βρούμε συνεκτικά υπογραφήματα

Πρόβλημα με θόρυβο

8 συνεχείς συστάδες

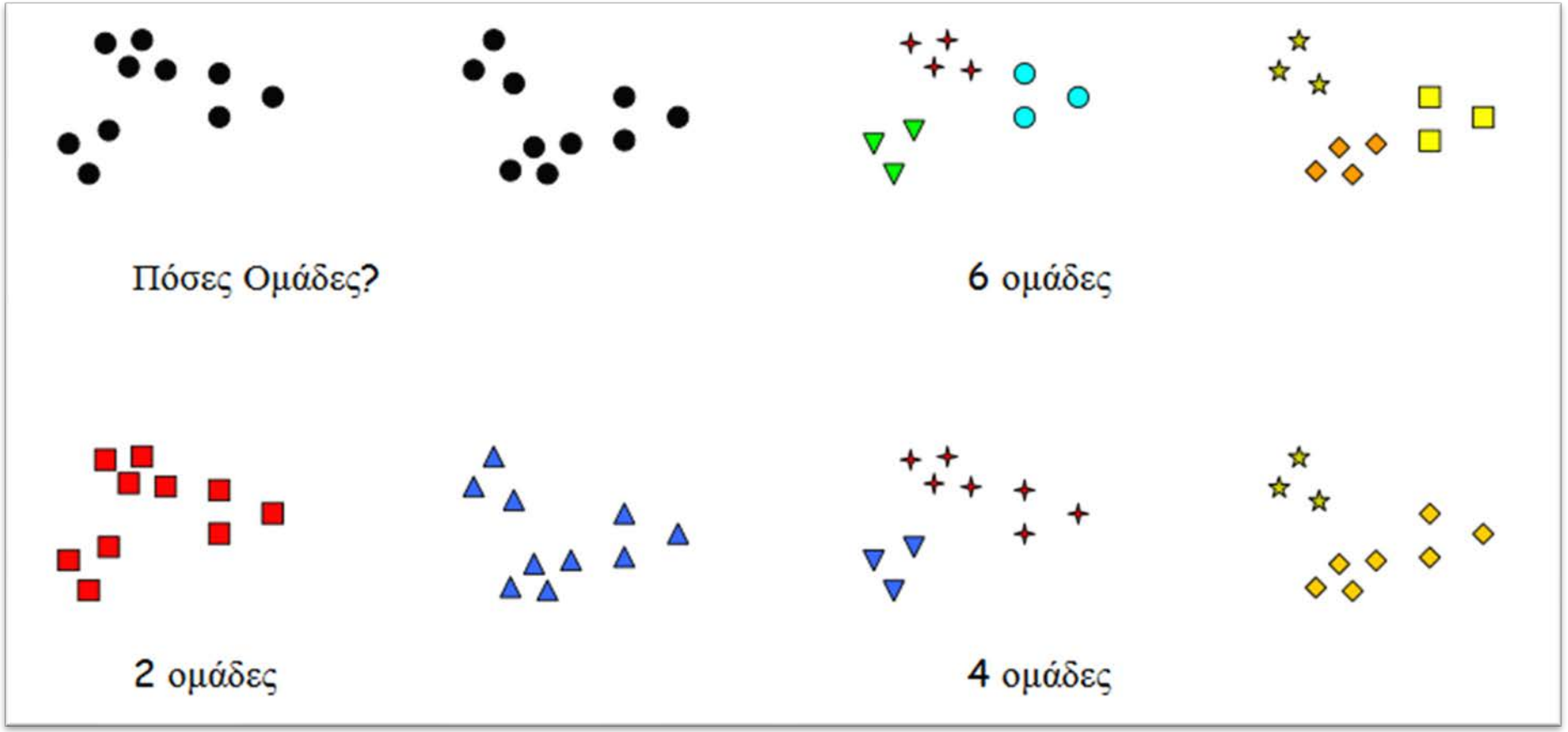
Τύποι συστάδων: Συστάδες βασισμένες στην πυκνότητα

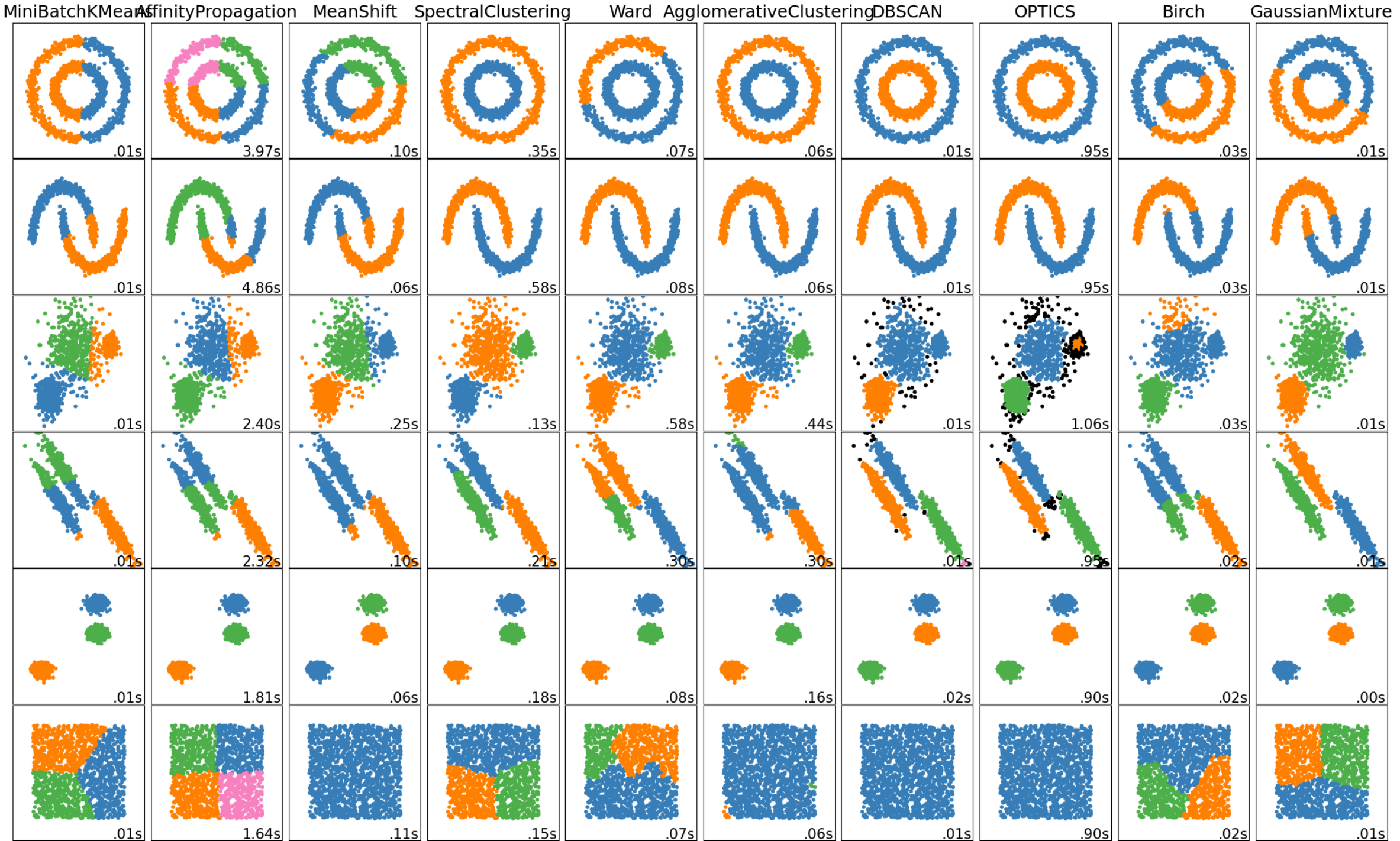
– Μια συστάδα είναι μια πυκνή περιοχή από σημεία την οποία χωρίζουν από άλλες περιοχές μεγάλης πυκνότητας περιοχές χαμηλής πυκνότητας

– Συχνά σε περιπτώσεις συστάδων με μη κανονικό σχήμα ή με αλληλοπλεκόμενα σχήματα ή όταν θόρυβος ή outliers

Πόσες Ομάδες Βλέπετε ?

Ασαφεια




MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
OPTICS
Birch
GaussianMixture
.01s
3.97s
.10s
.35s
.07s
.06s
.01s
.95s
.03s
.01s
.01s
4.86s
.06s
.58s
.08s
.06s
.01s
.95s
.03s
.01s
.01s
2.40s
.25s
.13s
.58s
.44s
.01s
1.06s
.03s
.01s
.01s
2.32s
.10s
.21s
.30s
.30s
.01s
.95s
.02s
.01s
.01s
1.81s
.06s
.18s
.08s
.16s
.02s
.90s
.02s
.00s
.01s
1.64s
.11s
.15s
.07s
.06s
.01s
.90s
.02s
.01s

Συσταδοποίηση βασισμένη σε αντιπροσώπους

- Αν δίνεται ένα σύνολο δεδομένων με *n* σημεία σε έναν *d*-διάστατο χώρο,

$$\mathbf{D} = \{\mathbf{x}_i\}_{i=1}^{n}$$

- καθώς και το πλήθος των επιθυμητών συστάδων *k*, ο στόχος της βασισμένης σε αντιπροσώπους συσταδοποίησης είναι ο διαμερισμός του συνόλου δεδομένων σε *k* ομάδες ή συστάδες, η οποία ονομάζεται *συσταδοποίηση* και συμβολίζεται με

  - $C = \{C_1, C_2, \ldots, C_k\}$.

- Για κάθε συστάδα C_i υπάρχει ένα αντιπροσωπευτικό σημείο που τη συνοψίζει· μια δημοφιλής επιλογή γι' αυτό το σημείο είναι ο μέσος $\boldsymbol{\mu}_i$ όλων των σημείων της συστάδας, που ονομάζεται επίσης *κέντρο βάρους*:

$$\mu_i = \frac{1}{n_i} \sum_{x_j \in C_i} \mathbf{x}_j$$

  - όπου $n_i = |C_i|$ είναι το πλήθος των σημείων που ανήκουν στη συστάδα C_i.

Συσταδοποίηση βασισμένη σε αντιπροσώπους

- Υπάρχει ένας απλοϊκός αλγόριθμος εξαντλητικής αναζήτησης για την εύρεση μιας καλής συσταδοποίησης:
  - Παράγουμε όλους τους πιθανούς διαμερισμούς των n σημείων σε k συστάδες,
  - Αξιολογούμε κάποια μετρική βελτιστοποίησης ώστε να προκύψει μια «βαθμολογία» για καθεμία από τις συστάδες, και
  - Επιλέγουμε τη συσταδοποίηση με την καλύτερη βαθμολογία.
- Όμως, αυτό είναι πρακτικά ανέφικτο επειδή υπάρχουν $O(k^n/k!)$ συσταδοποιήσεις των n σημείων σε k ομάδες.

Ο αλγόριθμος K-μέσων (k-means)

- Θέλουμε να βρούμε εκείνο το σύνολο k σημείων στον d-διάστατο χώρο, το οποίο ελαχιστοποιεί την μέση απόσταση ελαχίστων τετραγώνων κάθε σημείου από το κοντινότερό του κέντρο

- Η συνάρτηση βαθμολόγησης που βασίζεται στο άθροισμα των τετραγώνων των σφαλμάτων (SSE) ορίζεται ως:

$$SSE(\mathsf{C}) = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} \| \mathbf{x}_j - \boldsymbol{\mu}_i \|^2$$

- Ο στόχος μας είναι να βρούμε εκείνη τη συσταδοποίηση που ελαχιστοποιεί τη βαθμολογία SSE:

$$\mathsf{C}^* = \arg\min_{\mathsf{C}} \{SSE(\mathsf{C})\}$$

- Ο αλγόριθμος K μέσων χρησιμοποιεί μια άπληστη επαναληπτική τεχνική για να βρει μια συσταδοποίηση που ελαχιστοποιεί την αντικειμενική συνάρτηση SSE.

- Κατά συνέπεια, μπορεί να συγκλίνει σε τοπικά βέλτιστα και όχι σε μια καθολικά βέλτιστη συσταδοποίηση.

Ο αλγόριθμος K-μέσων (k-means)

- Ο αλγόριθμος K μέσων καθορίζει τις αρχικές τιμές των μέσων για τις συστάδες παράγοντας με τυχαίο τρόπο k σημεία στον χώρο δεδομένων. Κάθε επανάληψη του αλγορίθμου K μέσων αποτελείται από δύο βήματα: (1) την αντιστοίχιση σε συστάδες και (2) την ενημέρωση των κέντρων βάρους.

- Με την προϋπόθεση ότι δίνονται οι μέσοι των k συστάδων, κάθε σημείο $x_j \in D$ αντιστοιχίζεται στον πλησιέστερο μέσο κατά τη διάρκεια του πρώτου βήματος του αλγορίθμου· αυτό προκαλεί μια συσταδοποίηση, με κάθε συστάδα C_i να περιλαμβάνει σημεία που βρίσκονται πιο κοντά στον μέσο μ_i σε σύγκριση με τον μέσο οποιασδήποτε άλλης συστάδας. Δηλαδή, κάθε σημείο x_j αντιστοιχίζεται στη συστάδα C_{j*}, όπου

$$j^* = \arg\min_{k}^{i=1} \left\{ \| \mathbf{x}_j - \boldsymbol{\mu}_i \|^2 \right\}$$

- Για ένα καθορισμένο σύνολο συστάδων C_i, $i = 1,\ldots, k$, στο δεύτερο βήμα του αλγορίθμου (ενημέρωση των κέντρων βάρους) υπολογίζονται νέες μέσες τιμές για κάθε συστάδα από τα σημεία του συνόλου C_i.

- Τα βήματα της αντιστοίχισης σε συστάδες και της ενημέρωσης των κέντρων βάρους εκτελούνται επαναληπτικά μέχρι να καταλήξουμε σε ένα σταθερό σημείο ή σε τοπικά ελάχιστα.

Ο αλγόριθμος K-μέσων (k-means)

- ΣΚΟΠΟΣ : Εύρεση των κέντρων των ομάδων

- ΜΕΘΟΔΟΣ : Ελαχιστοποίηση του σφάλματος, J

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} (||x_i^{(j)} - c_j||)^2$$

- ΒΗΜΑΤΑ
  I. Ορισμός K κέντρων συστάδων με τυχαίο τρόπο
  II. Εισαγωγή αντικειμένου στη συστάδα με το πιο κοντινό κέντρο
  III. Ανανέωση του κέντρου της συστάδας
  IV. Επανάληψη των βημάτων 2,3 μέχρι τη σύγκλιση (αλλαγή στις συστάδες μικρότερη από ένα κατώφλι)

- Ουσιαστικά, ο αλγόριθμος προσπαθεί επαναληπτικά να «μειώσει» την απόσταση όλων των σημείων από ένα σημείο της συστάδας

Ο αλγόριθμος K-μέσων (k-means)

K-MEANS (D, k, ϵ **)**:

1 $t \leftarrow 0$

2 Καθορισμός αρχικής τιμής για k κέντρα βάρους με τυχαίο τρόπο: $\boldsymbol{\mu}_1^t, \boldsymbol{\mu}_2^t, \ldots, \boldsymbol{\mu}_k^t \in \mathbb{R}^d$

3 **repeat**

4 $\quad t \leftarrow t + 1$

5 $\quad C_j \leftarrow \emptyset$ για όλα τα $j = 1, \ldots, k$

$\quad$ // Βήμα αντιστοίχισης σε συστάδες

6 $\quad$ **foreach** $\mathbf{x}_j \in \mathbf{D}$ **do**

7 $\quad\quad j^* \leftarrow \arg\min_i \left\{ \left\| \mathbf{x}_j - \boldsymbol{\mu}_i^{t-1} \right\|^2 \right\}$ // Αντιστοίχιση του $\mathbf{x}_j$ στο πλησιέστερο κέντρο βάρους

8 $\quad\quad C_{j^*} \leftarrow C_{j^*} \cup \{\mathbf{x}_j\}$

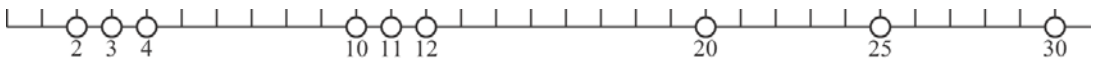
$\quad$ // Βήμα ενημέρωσης των κέντρων βάρους

9 $\quad$ **foreach** $i = 1$ *to* k **do**

10 $\quad\quad \boldsymbol{\mu}_i^t \leftarrow \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$

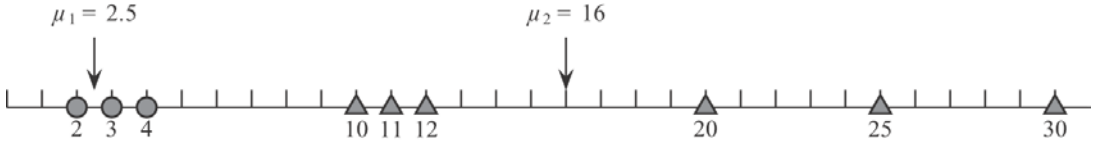
11 **until** $\sum_{i=1}^{k} \left\| \boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^{t-1} \right\|^2 \leq \epsilon$

Ο αλγόριθμος Κ μέσων στη μία διάσταση

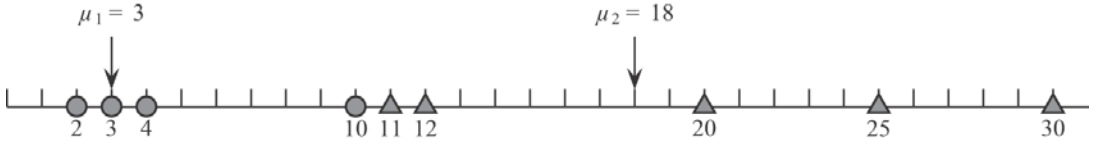


(α) Αρχικό σύνολο δεδομένων

(β) Επανάληψη: $t = 1$



(γ) Επανάληψη: $t = 2$



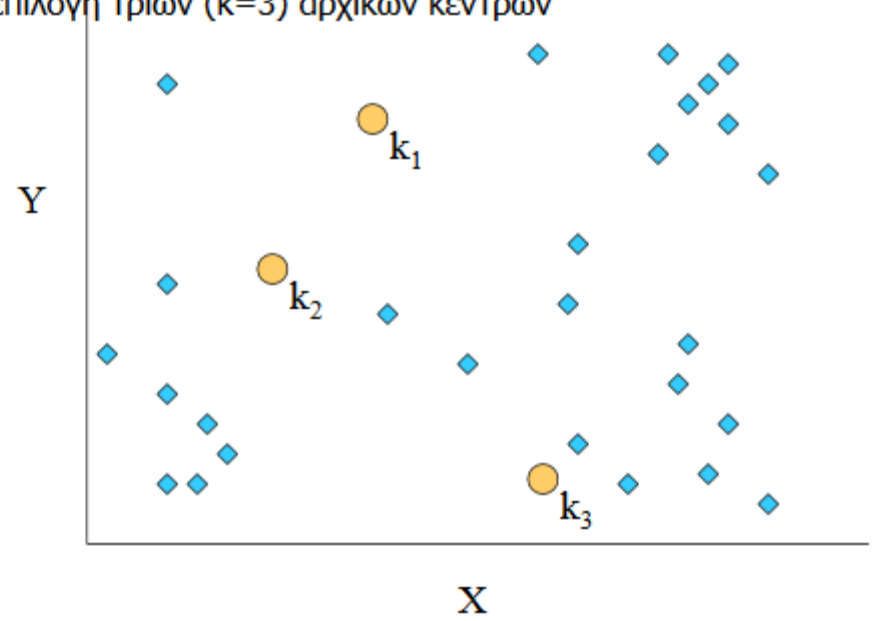
(δ) Επανάληψη: $t = 3$

(ε) Επανάληψη: $t = 4$

(στ) Επανάληψη: $t = 5$ (σύγκλιση)

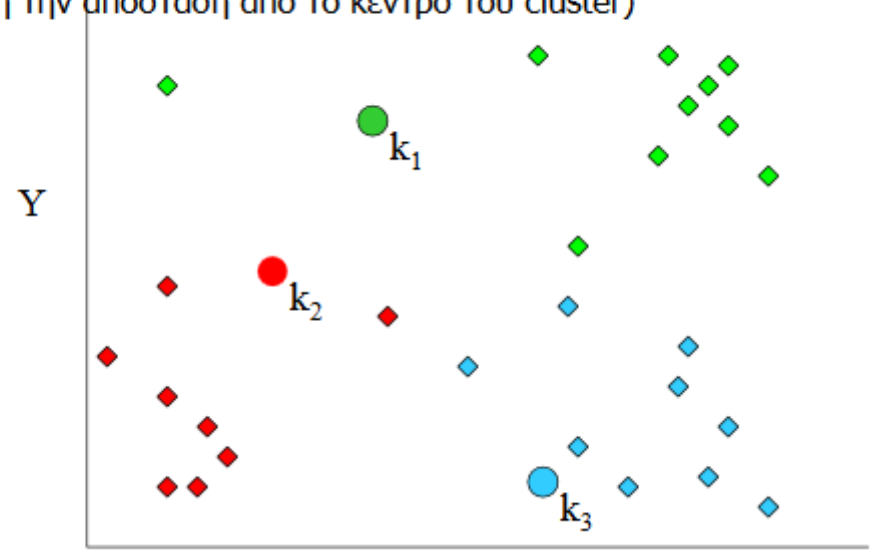
K-means σε 2 διαστάσεις

- Τυχαία επιλογή τριών (k=3) αρχικών κέντρων



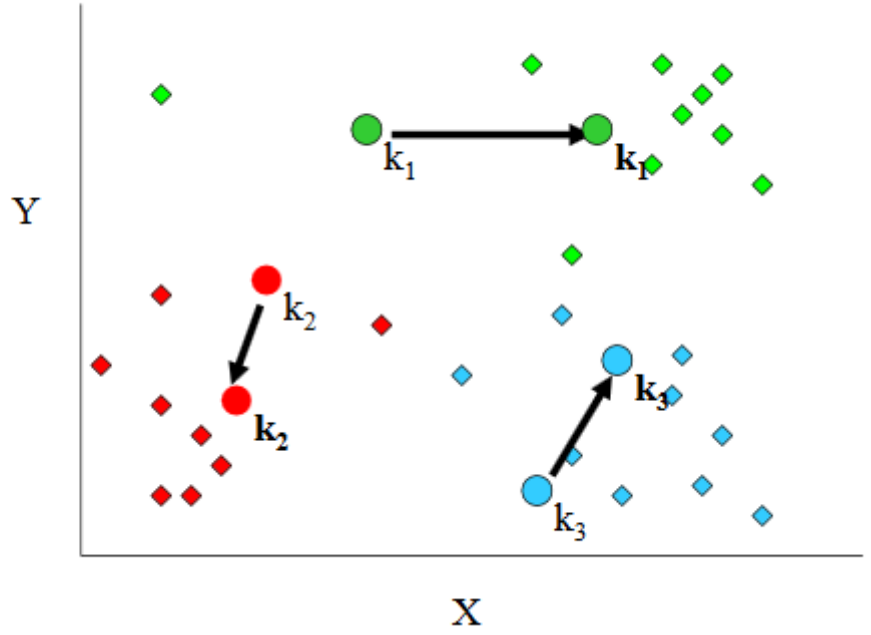
K-means σε 2 διαστάσεις

- Εκχώρηση κάθε στοιχείου στο πλησιέστερό του cluster (με βάση την απόσταση από το κέντρο του cluster)



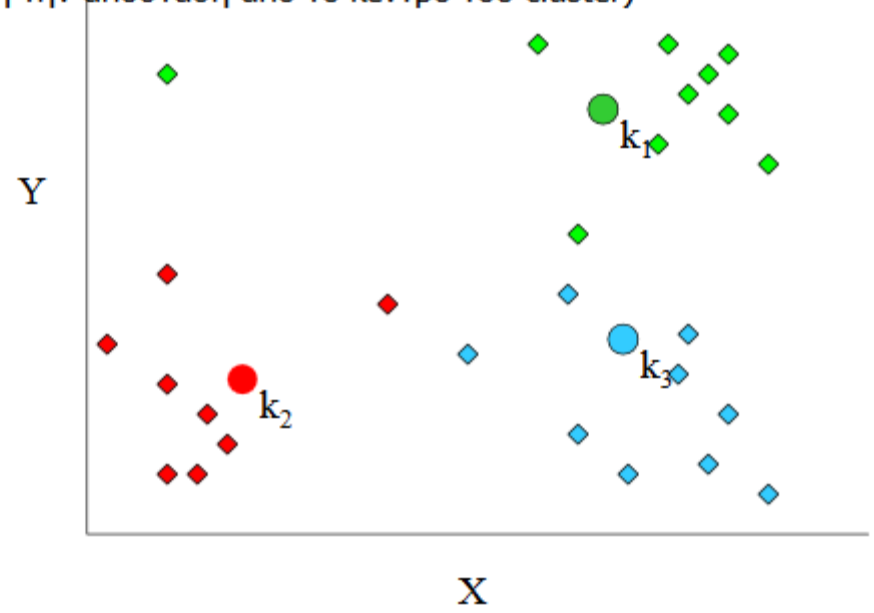
K-means σε 2 διαστάσεις

- Επανυπολογισμός του νέου κέντρου βάρους του κάθε cluster

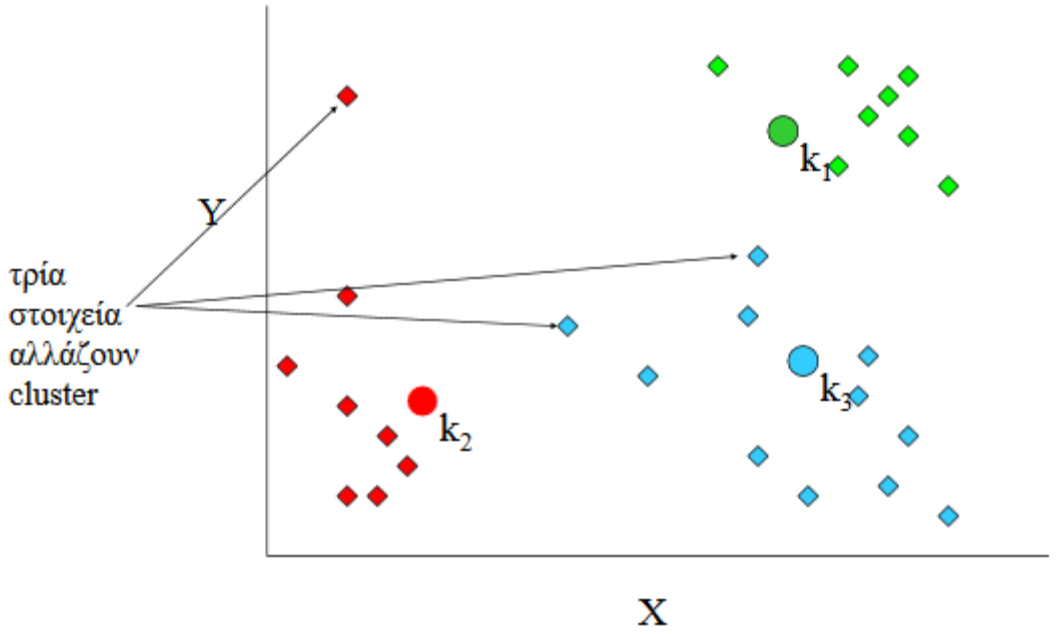


K-means σε 2 διαστάσεις

- Εκχώρηση κάθε στοιχείου στο πλησιέστερό του cluster
  (με βάση την απόσταση από το κέντρο του cluster)

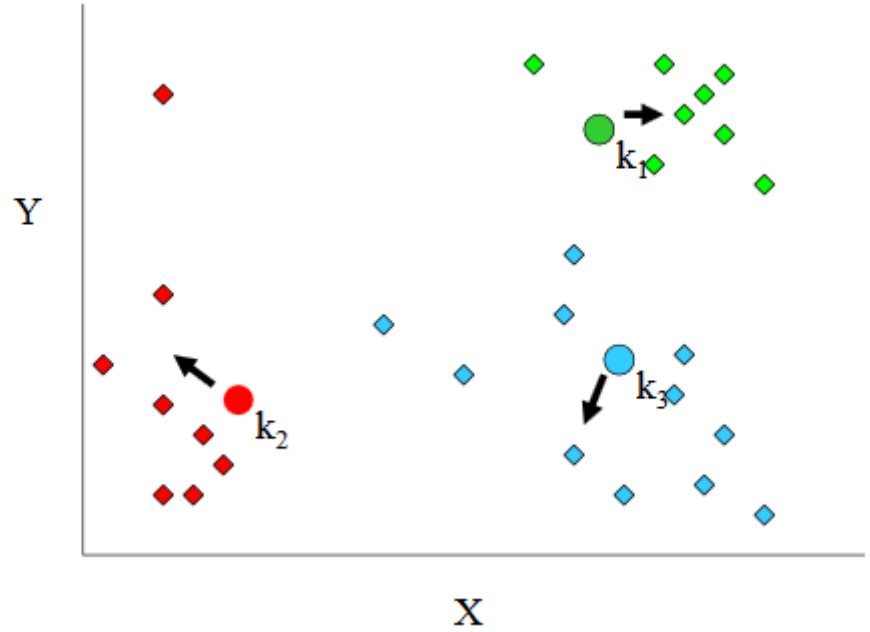


K-means σε 2 διαστάσεις

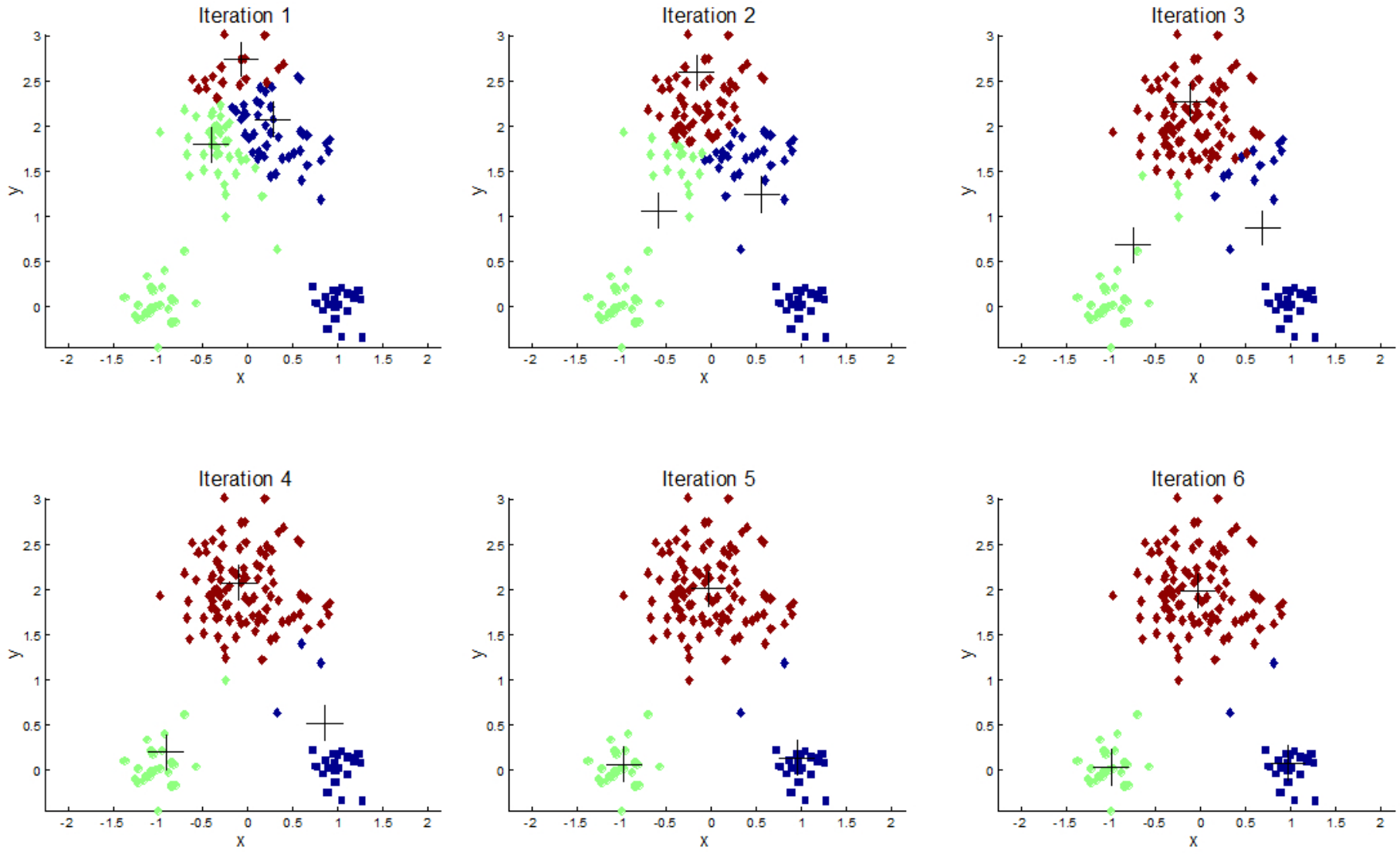


K-means σε 2 διαστάσεις

- Επανυπολογισμός του νέου κέντρου βάρους του κάθε cluster



Αλγόριθμος k-means - ΒΗΜΑΤΑ



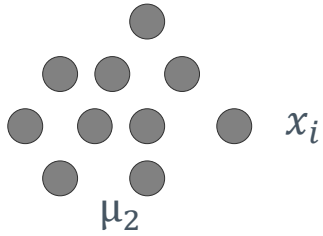
K-means - Συμπέρασμα

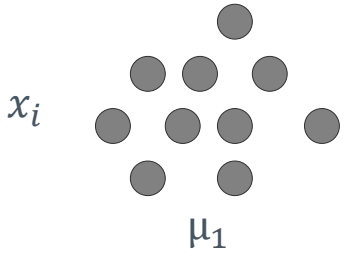
- Πλεονεκτήματα
  - Απλός, κατανοητός
  - Τα αντικείμενα ανατίθενται αυτόματα σε κάποιο cluster
  - Ταχύτητα σύγκλισης

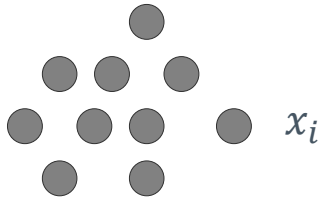
- Μειονεκτήματα
  - Πρέπει να οριστεί ο αριθμός των clusters
  - Όλα τα αντικείμενα πρέπει υποχρεωτικά να ανήκουν σε κάποιο cluster
  - Δε δουλεύει για μη αριθμητικά δεδομένα
  - Μη-ντετερμινιστικός

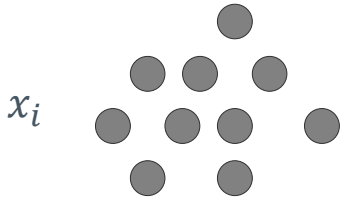
Ερωτηση

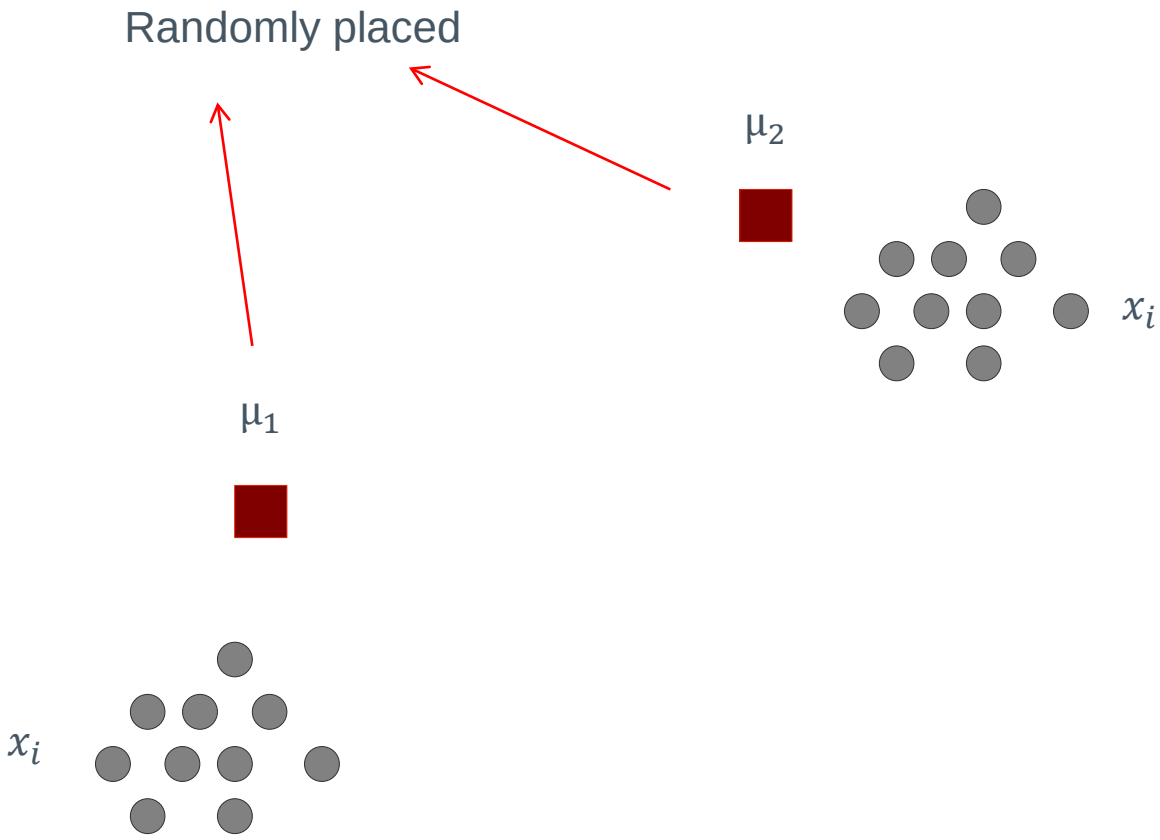
– Ποια αρχικοποίηση θα μπέρδευε τον K-means ? (εστω k = 2)



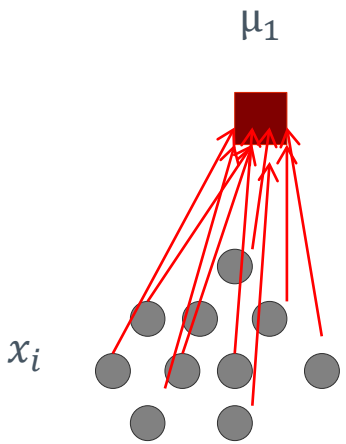


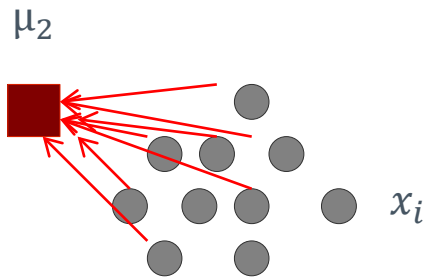

x_i


x_i

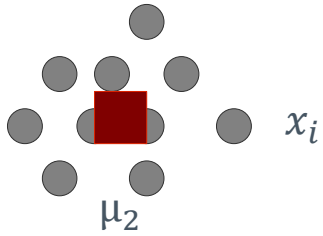

Randomly placed
μ_2
x_i
μ_1
x_i

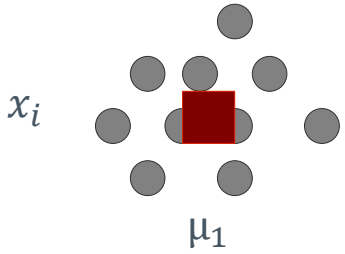
Find optimal allocation of points





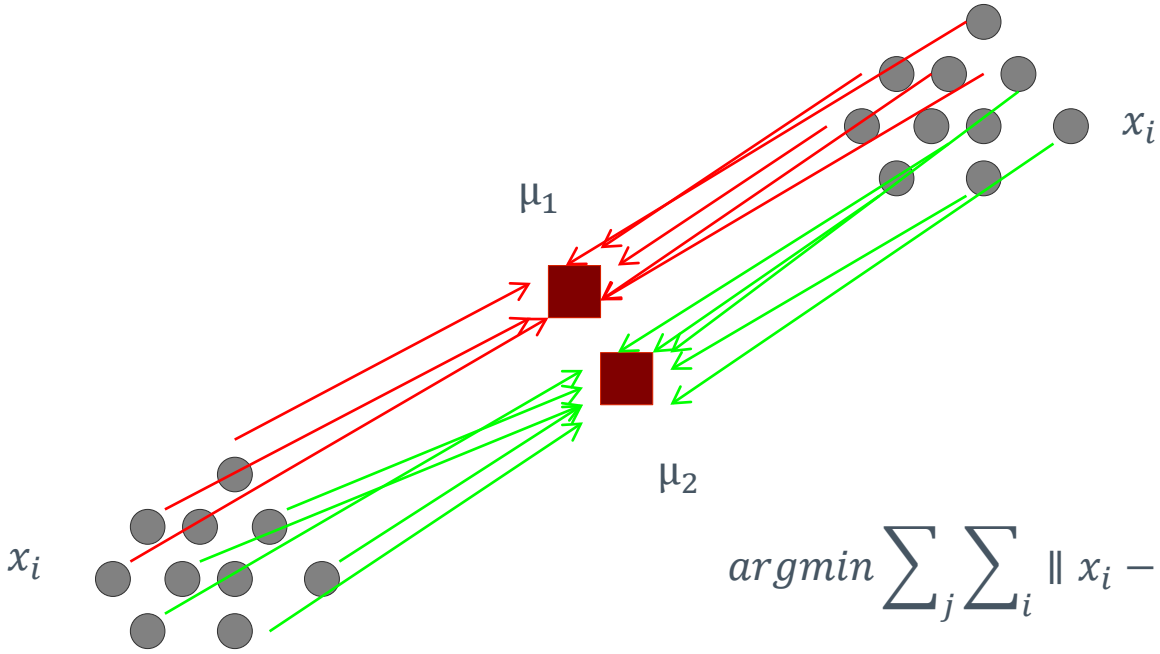
Reassign and repeat





$$argmin \sum_j \sum_i \| x_i - \mu_j \|$$

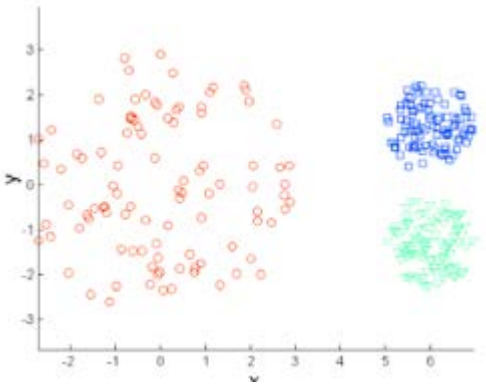
Problems with local optimum of the optimization function



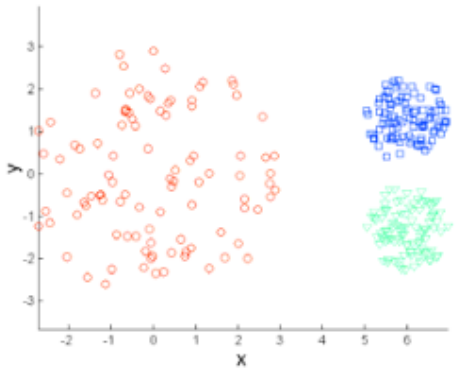
$$argmin \sum_j \sum_i \| x_i - \mu_j \|$$

Ερωτηση

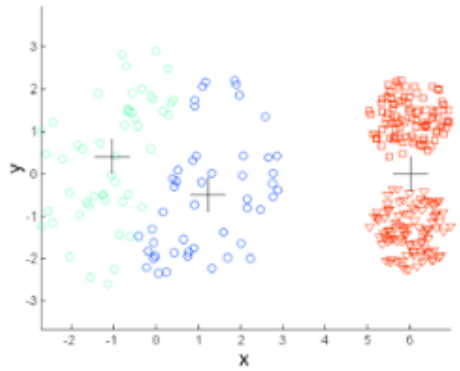
– Θα λειτουργούσε καλά ο k-means (εστω ότι δίναμε σωστό K)



K-means: Περιορισμοί – Διαφορετικές Πυκνότητες



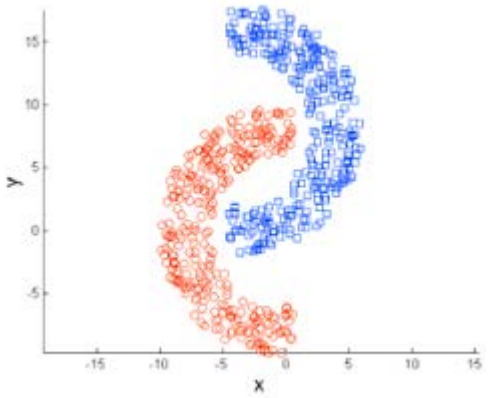
Αρχικά σημεία

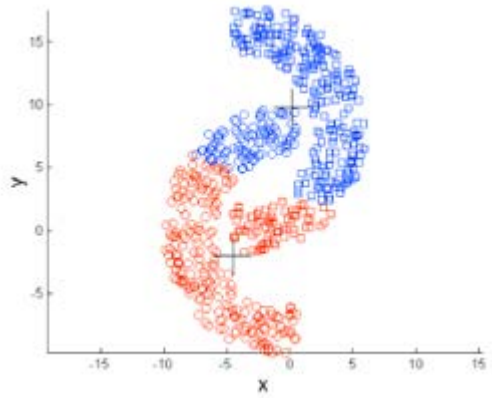


K-means (3 συστάδες)

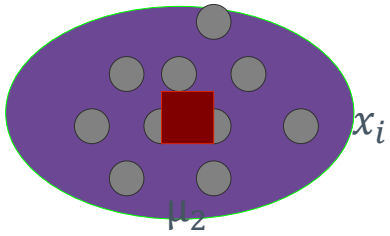
Δεν μπορεί να διαχωρίσει τους δυο μικρούς γιατί είναι πολύ πυκνοί σε σχέση με τον ένα μεγάλο

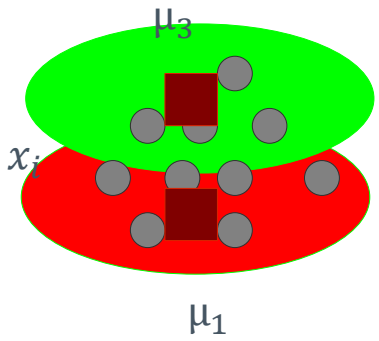
Δεν μπορεί να βρει τις δύο συστάδες γιατί έχουν μη κυκλικά σχήματα





Problems with wrong number of clusters

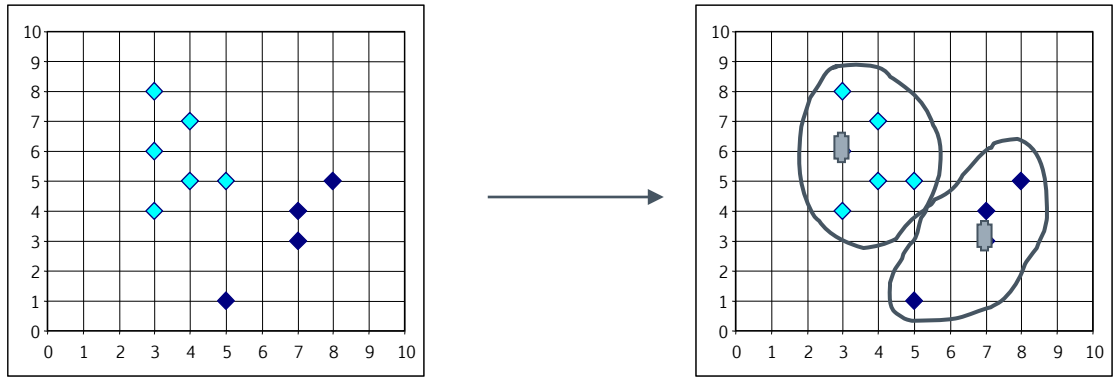




$$argmin \sum_j \sum_i \| x_i - \mu_j \|$$

K-medoid

- Συνήθως συνεχή d-διάστατο χώρο
- Διαλέγει ένα αντιπροσωπευτικό σημείο από τα δεδομένα και ελαχιστοποιεί την απόσταση από αυτό – Medoid: το πιο κεντρικό σημείο της συστάδας (αντί να χρησιμοποιεί το mean)
- Μειώνει την ευαισθησία σε outliers
- Μπορεί να εφαρμοστεί σε δεδομένα οποιουδήποτε τύπου (πχ και για κατηγορικά δεδομένα)



Ιεραρχική συσταδοποίηση: Ένθετες διαμερίσεις

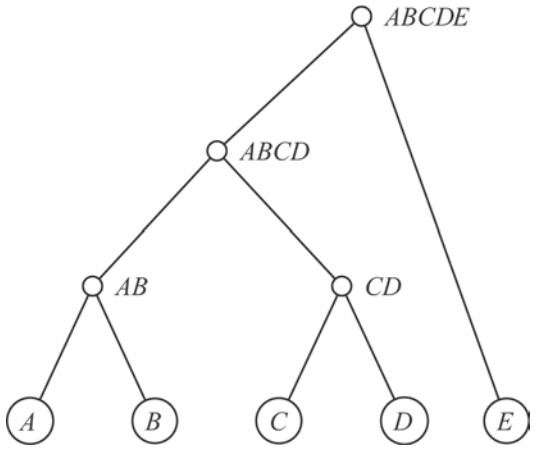
— Αν δίνεται ένα σύνολο δεδομένων $\mathbf{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, όπου $\mathbf{x}_i \in \mathbb{R}^d$, μια συσταδοποίηση $C = \{C_1, \ldots, C_k\}$ αποτελεί διαμέριση του $\mathbf{D}$.

— Λέμε ότι μια συσταδοποίηση $A = \{A_1, \ldots, A_r\}$ είναι ένθετη σε μια άλλη συσταδοποίηση $B = \{B_1, \ldots, B_s\}$ αν και μόνο αν ισχύει $r > s$, και για κάθε συστάδα $A_i \in A$ υπάρχει μια συστάδα $B_j \in B$ τέτοια ώστε να ισχύει $A_i \subseteq B_j$.

— Η ιεραρχική συσταδοποίηση παράγει μια ακολουθία n ένθετων διαμερίσεων $C_1, \ldots, C_N$. Η συσταδοποίηση είναι ένθετη στη συσταδοποίηση C_t.

— Το δενδρόγραμμα συστάδων είναι ένα δυαδικό δένδρο με ρίζα που αποτυπώνει την ένθετη δομή, και στο οποίο υπάρχουν ακμές μεταξύ της συστάδας $C_i \in C_{t-1}$ και της συστάδας $C_j \in C_t$ αν η C_i είναι ένθετη στη C_j, δηλαδή αν ισχύει $C_i \subset C_j$.

Ιεραρχική συσταδοποίηση: Ένθετες διαμερίσεις



Συσταδοποίηση	Συστάδες
	{A}, {B}, {C}, {D}, {E}
	{AB}, {C}, {D}, {E}
	{AB}, {CD}, {E}
	{ABCD}, {E}
	{ABCDE}

Το δενδρόγραμμα αναπαριστά την παρακάτω ακολουθία ένθετων διαμερίσεων

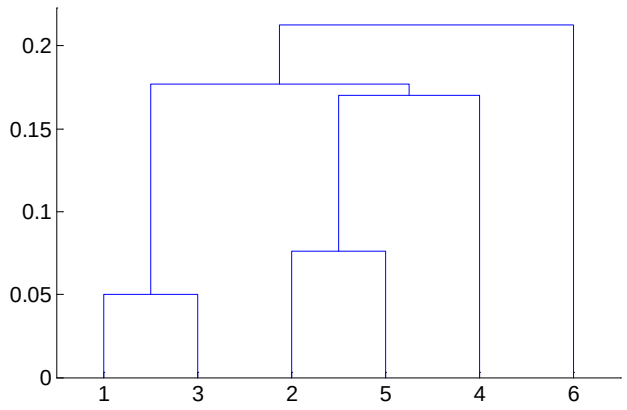
— με $C_{t-1} \in C_t$ για $t = 2,\ldots, 5$.

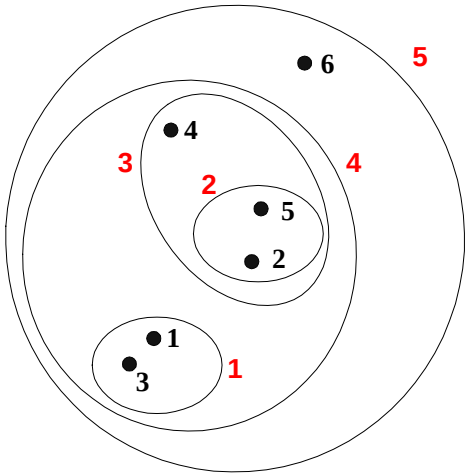
— Υποθέτουμε ότι οι συστάδες A και B συγχωνεύονται πριν από τις συστάδες C και D.

Ιεραρχική Συσταδοποίηση: Βασικά

- Παράγει ένα σύνολο από εμφωλευμένες συστάδες οργανωμένες σε ένα ιεραρχικό δέντρο

- Μπορεί να παρασταθεί με ένα δένδρο-γραμμα
  - Ένα διάγραμμα που μοιάζει με δένδρο και καταγράφει τις ακολουθίες από συγχωνεύσεις (merges) και διαχωρισμούς (splits)

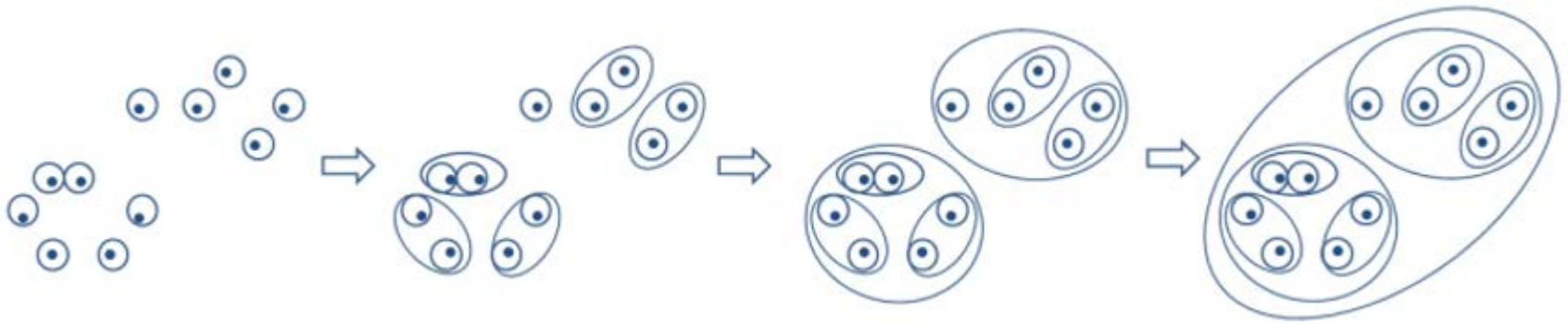




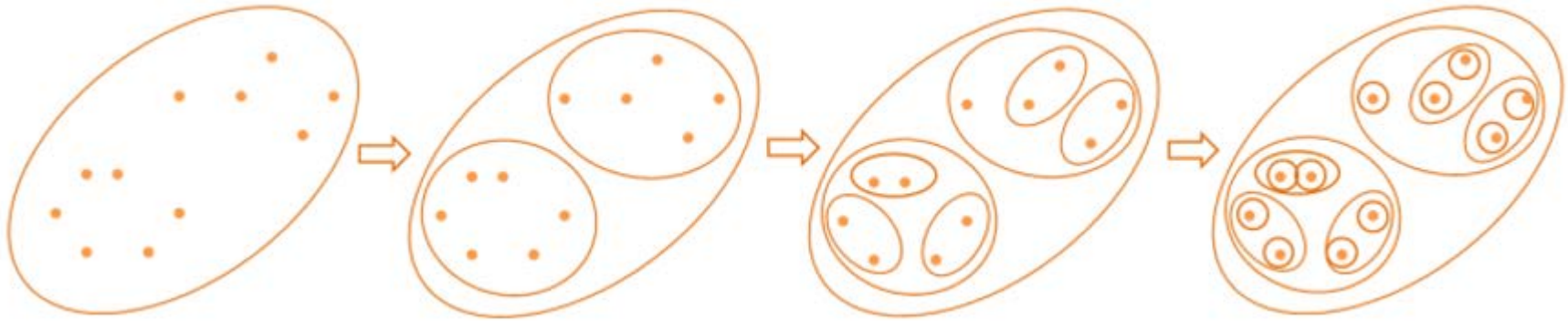
Ιεραρχική συσταδοποίηση

Δυο βασικοί τύποι ιεραρχικής συσταδοποίησης

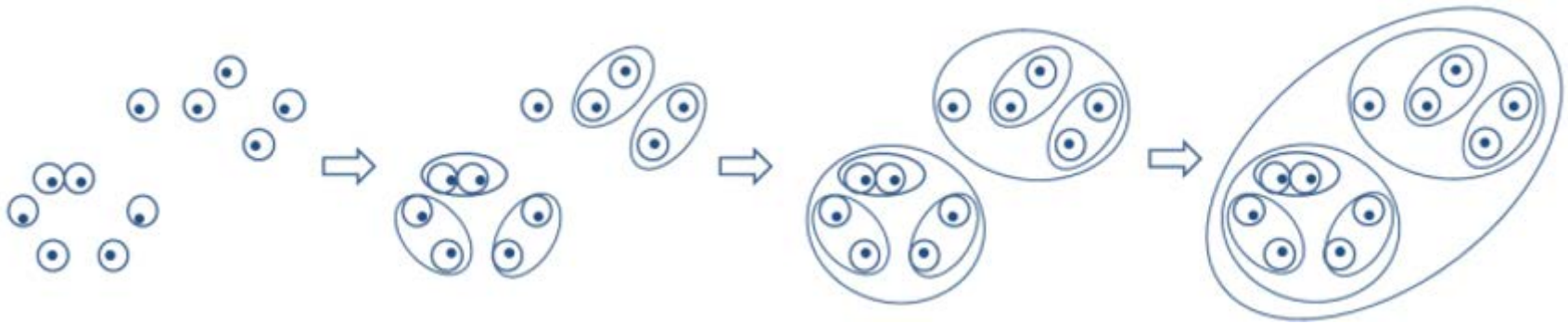
- Συσσωρευτικός (Agglomerative):
  - Αρχίζει με τα σημεία ως ξεχωριστές συστάδες
  - Σε κάθε βήμα, συγχωνεύει το πιο κοντινό ζευγάρι συστάδων μέχρι να μείνει μόνο μία (ή k) συστάδες



- Διαιρετικός (Divisive):
  - Αρχίζει με μία συστάδα που περιέχει όλα τα σημεία
  - Σε κάθε βήμα, διαχωρίζει μία συστάδα, έως κάθε συστάδα να περιέχει μόνο ένα σημείο (ή να δημιουργηθούν k συστάδες)



Συσσωρευτικός (Agglomerative)



Βασικός Αλγόριθμος

1: Υπολογισμός του Πίνακα Γειτνίασης
2: Έστω κάθε σημείο αποτελεί και μια συστάδα
3: **Repeat**
4: Συγχώνευση των δύο κοντινότερων συστάδων
5: Ενημέρωση του Πίνακα Γειτνίασης
6: **Until** να μείνει μία μόνο συστάδα

- Βασική λειτουργία είναι ο υπολογισμός της γειτνίασης δυο συστάδων
- **Διαφορετικοί αλγόριθμοι με βάση το πως ορίζεται η απόσταση ανάμεσα σε δύο συστάδες**

Κριτήριο Απόστασης

$$\delta(\mathbf{x},\mathbf{y}) = ||\,\mathbf{x}-\mathbf{y}||_2 = \left(\sum_{i=1}^{d} |x_i - y_i|^p\right)^{1/p}$$

Lp-Norm

$$\delta(\mathbf{x},\mathbf{y}) = ||\,\mathbf{x}-\mathbf{y}||_2 = \sum_{i=1}^{d} |x_i - y_i|$$

L1-Norm Manhattan

$$\delta(\mathbf{x},\mathbf{y}) = ||\,\mathbf{x}-\mathbf{y}||_2 = \sqrt{\sum_{i=1}^{d} |x_i - y_i|^2}$$

L2-Norm Euclidean

Απόσταση μεταξύ συστάδων: Μοναδικός σύνδεσμος, πλήρης σύνδεσμος και μέσος όρος ομάδας

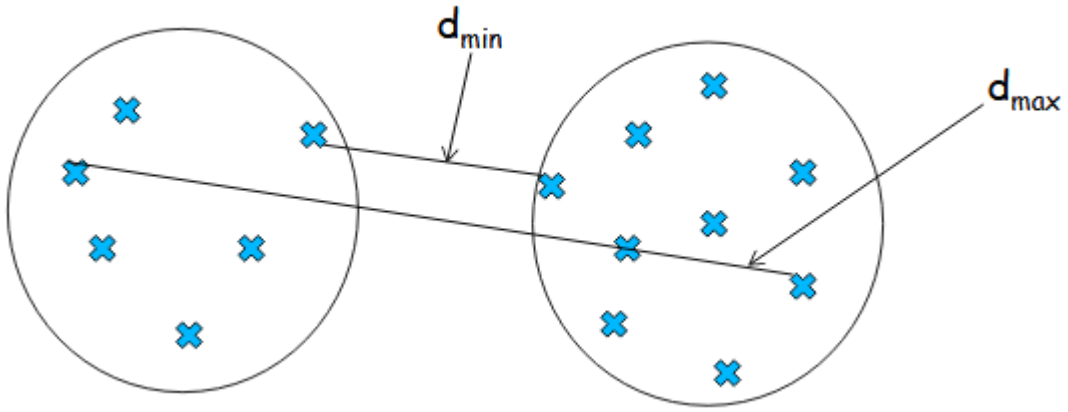
— Οι «διασυσταδικές» αποστάσεις υπολογίζονται ως εξής.

— **Μοναδικός σύνδεσμος:** Η ελάχιστη απόσταση ενός σημείου της συστάδας C_i από ένα σημείο της συστάδας C_j

$$\delta(C_i, C_j) = \min\{\delta(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$

— **Πλήρης σύνδεσμος:** Η μέγιστη απόσταση μεταξύ ενός σημείου της συστάδας C_i και ενός σημείου της συστάδας C_j

$$\delta(C_i, C_j) = \max\{\delta(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in C_i, \mathbf{y} \in C_j\}$$



Απόσταση μεταξύ συστάδων: Μοναδικός σύνδεσμος, πλήρης σύνδεσμος και μέσος όρος ομάδας

— **Μέσος όρος ομάδας :** Ο μέσος όρος της απόστασης ανά ζεύγη μεταξύ σημείων της συστάδας C_i και της συστάδας C_j

$$\delta(C_i, C_j) = \delta(\mu_i, \mu_j)$$

— Ελάχιστη διακύμανση ή μέθοδος του Ward: Η απόσταση δύο συστάδων ορίζεται ως η αύξηση που παρουσιάζει το άθροισμα των τετραγώνων των σφαλμάτων (SSE) όταν οι δύο συστάδες συγχωνευθούν. Το άθροισμα SSE για μια δεδομένη συστάδα C_i δίνεται από τη σχέση

$$\delta(C_i, C_j) = \Delta SSE_{ij} = SSE_{ij} - SSE_i - SSE_j$$

όπου $SSE_i = \sum_{\mathbf{x} \in C_i} ||\mathbf{x} - \mu_i||^2$

— Μετά από απλοποίηση, παίρνουμε

$$\delta(C_i, C_j) = \left(\frac{n_i n_j}{n_i + n_j}\right) ||\mu_i - \mu_j||^2$$

– Συνεπώς, το μέτρο του Ward αποτελεί μια σταθμισμένη εκδοχή του μέτρου της μέσης απόστασης.

Ιεραρχική Ομαδοποίηση

Στον πίνακα δίνονται οι τιμές από δύο μετρήσεις (X_1 και X_2).

1. Να ομαδοποιήσετε τα δεδομένα με τον αλγόριθμο «Συσσωρευτικής Ιεραρχικής Ομαδοποίησης» (Agglomerative Hierarchical Clustering), χρησιμοποιώντας ως κριτήριο της απόστασης μεταξύ ομάδων, την περίπτωση του απλού συνδέσμου (MIN ή single link) και πλήρους συνδέσμου (MAX ή complete linkage).
2. Να κατασκευαστεί το δενδρόγραμμα της κάθε ομαδοποίησης και να τα συγκρίνετε.

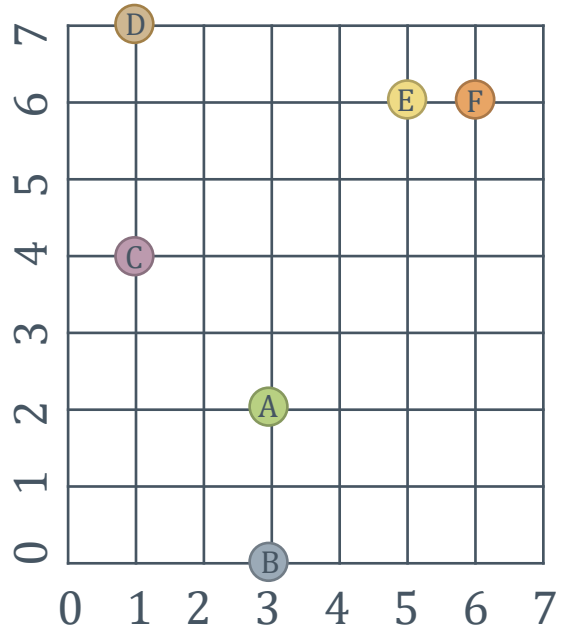
	X_1	X_2
D	1	7
B	3	0
A	3	2
E	5	6
F	6	6
C	1	4

Ιεραρχική Ομαδοποίηση

Στον πίνακα δίνονται οι τιμές από δύο μετρήσεις (X_1 και X_2).

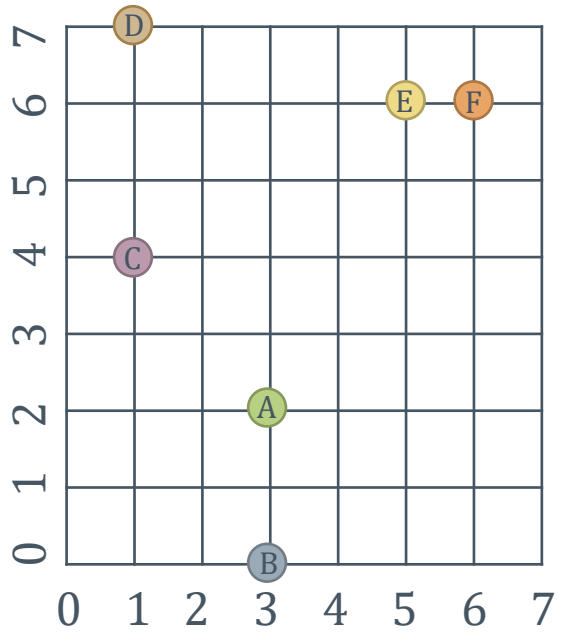
1. Να ομαδοποιήσετε τα δεδομένα με τον αλγόριθμο «Συσσωρευτικής Ιεραρχικής Ομαδοποίησης» (Agglomerative Hierarchical Clustering), χρησιμοποιώντας ως κριτήριο της απόστασης μεταξύ ομάδων, την περίπτωση του απλού συνδέσμου (MIN ή single link) και πλήρους συνδέσμου (MAX ή complete linkage).
2. Να κατασκευαστεί το δενδρόγραμμα της κάθε ομαδοποίησης και να τα συγκρίνετε.

	X_1	X_2
D	1	7
B	3	0
A	3	2
E	5	6
F	6	6
C	1	4



Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)

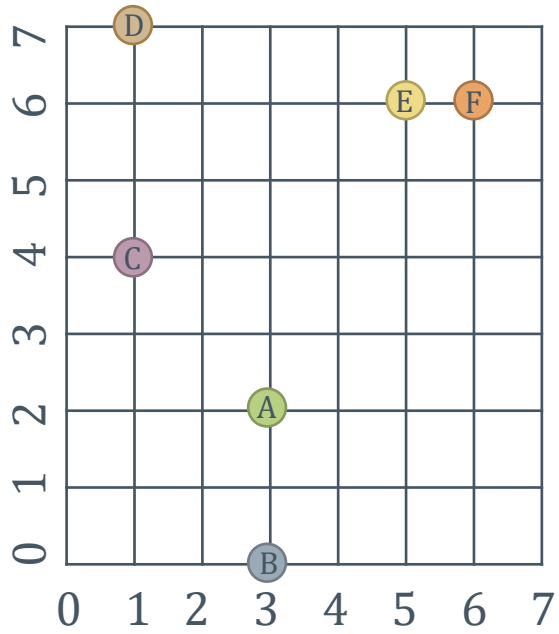


Πίνακας Αποστάσεων

	A	B	C	D	E	F
A	0	2	4	7	6	7
B	2	0	6	9	8	9
C	4	6	0	3	6	7
D	7	9	3	0	5	6
E	6	8	6	5	0	1
F	7	9	7	6	1	0

Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)

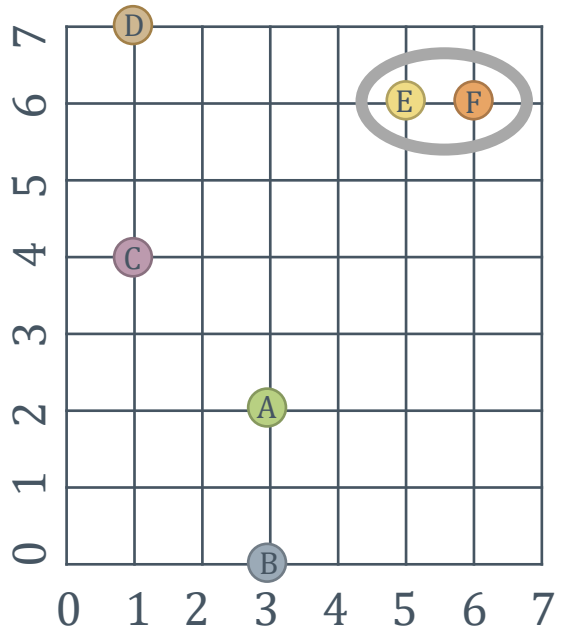


Πίνακας Αποστάσεων

	A	B	C	D	E	F
A	0	2	4	7	6	7
B	2	0	6	9	8	9
C	4	6	0	3	6	7
D	7	9	3	0	5	6
E	6	8	6	5	0	1
F	7	9	7	6	1	0

Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)

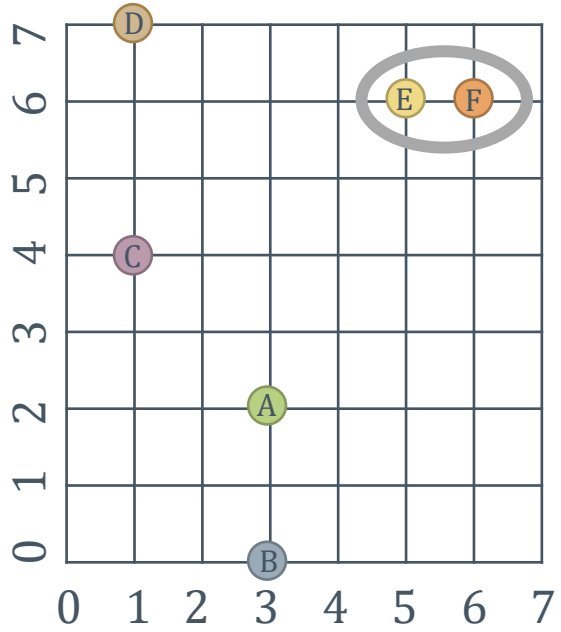


Πίνακας Αποστάσεων

	A	B	C	D	E	F
A	0	2	4	7	6	7
B	2	0	6	9	8	9
C	4	6	0	3	6	7
D	7	9	3	0	5	6
E	6	8	6	5	0	1
F	7	9	7	6	1	0

Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)

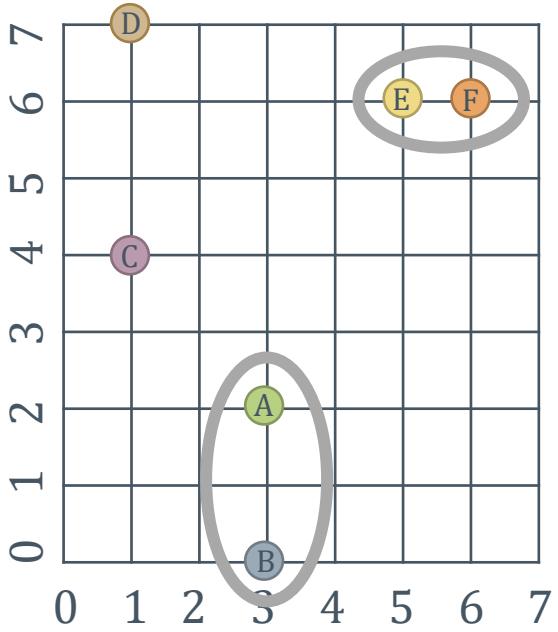


Πίνακας Αποστάσεων

	A	B	C	D	E/F
A	0	2	4	7	6
B	2	0	6	9	8
C	4	6	0	3	6
D	7	9	3	0	5
E/F	6	8	6	5	0

Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)

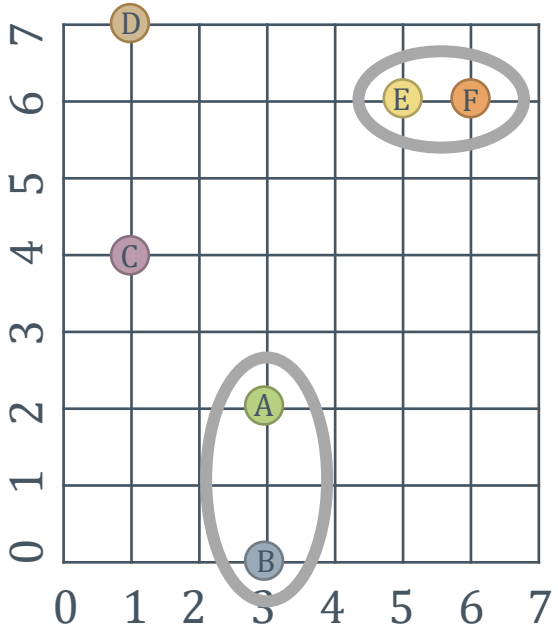


Πίνακας Αποστάσεων

	A	B	C	D	E/F
A	0	2	4	7	6
B	2	0	6	9	8
C	4	6	0	3	6
D	7	9	3	0	5
E/F	6	8	6	5	0

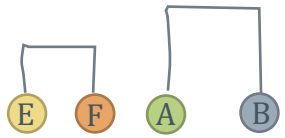
Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)



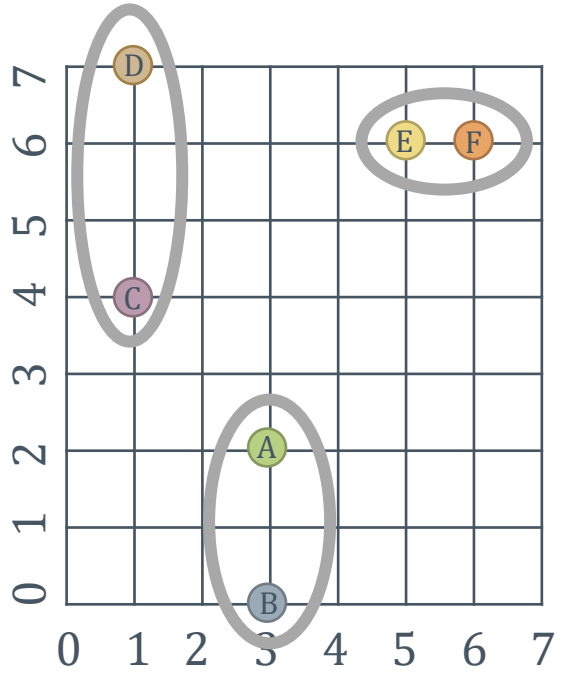
Πίνακας Αποστάσεων

	A/B	C	D	E/F
A/B	0	4	7	6
C	4	0	3	6
D	7	3	0	5
E/F	6	6	5	0



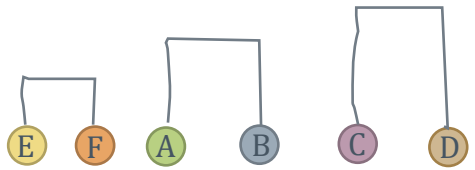
Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)



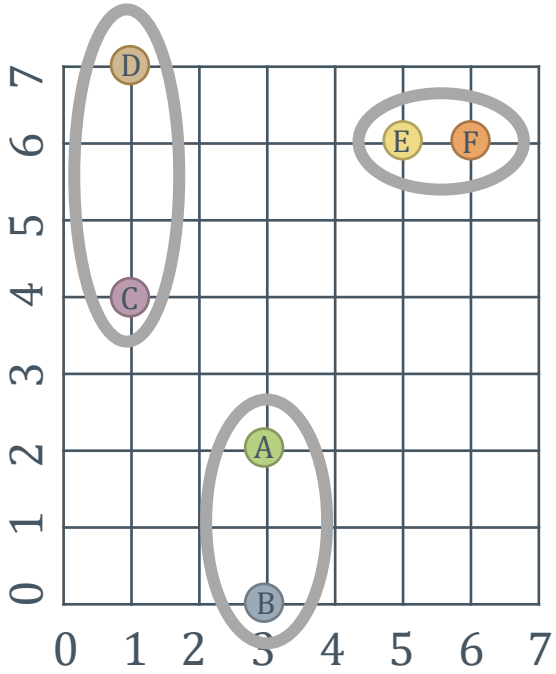
Πίνακας Αποστάσεων

	A/B	C	D	E/F
A/B	0	4	7	6
C	4	0	3	6
D	7	3	0	5
E/F	6	6	5	0



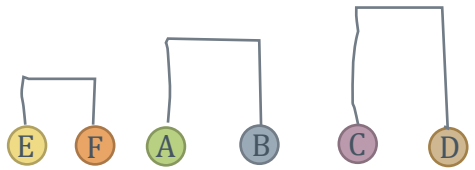
Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)



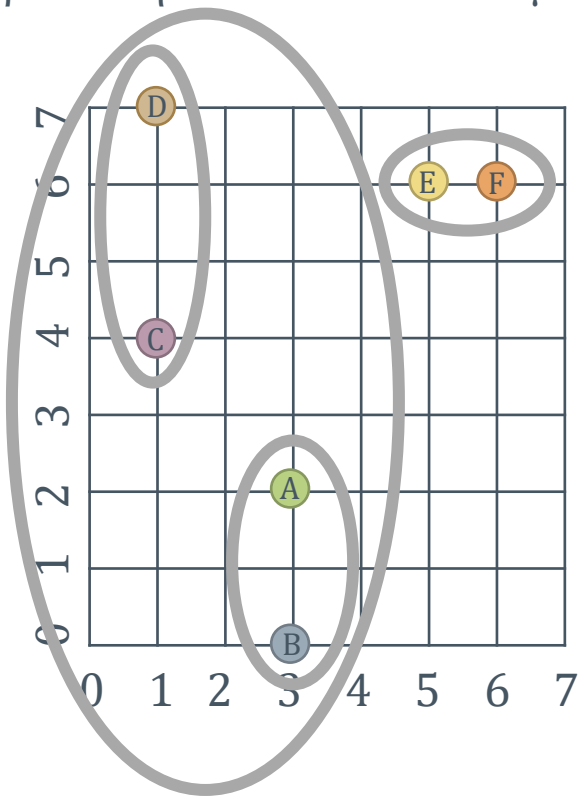
Πίνακας Αποστάσεων

	A/B	C/D	E/F
A/B	0	4	6
C/D	4	0	5
E/F	6	5	0



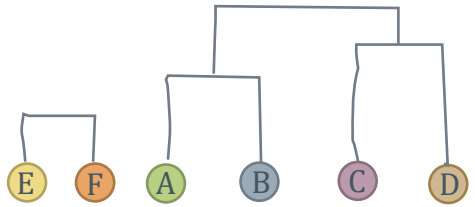
Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)



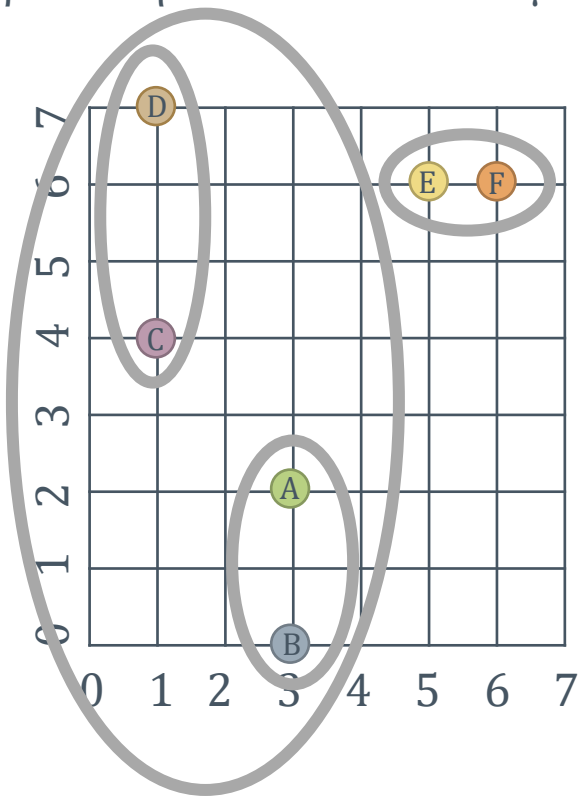
Πίνακας Αποστάσεων

	A/B	C/D	E/F
A/B	0	4	6
C/D	4	0	5
E/F	6	5	0



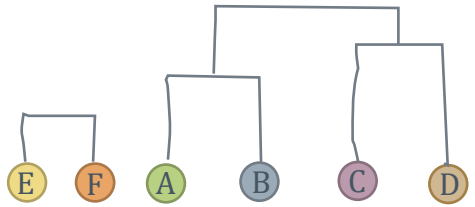
Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)



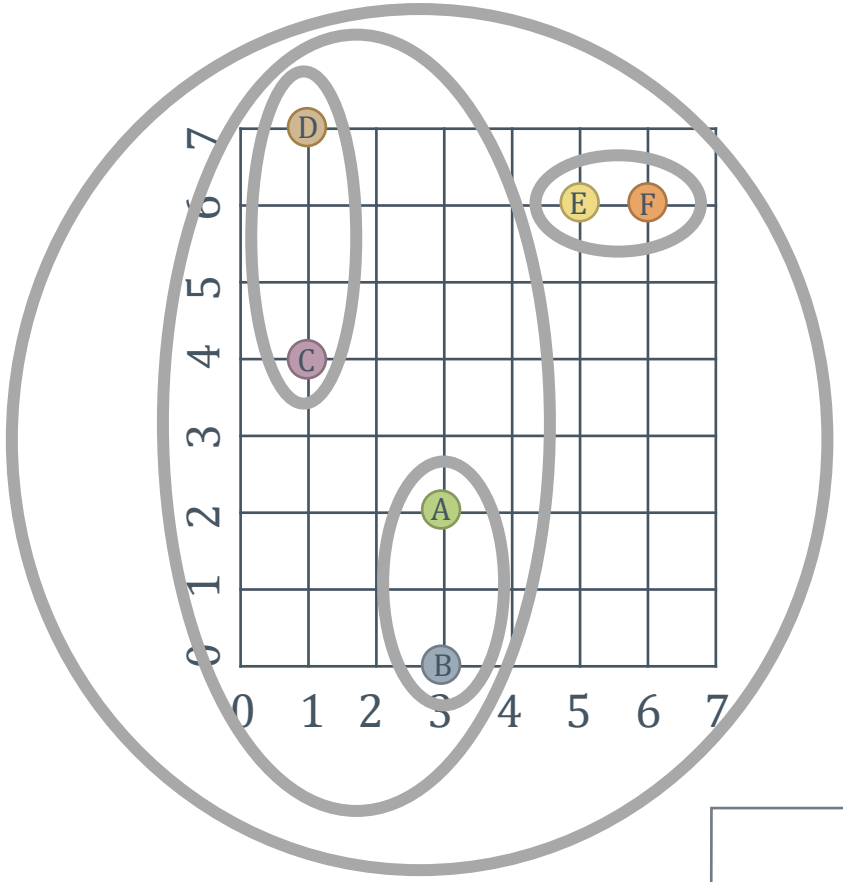
Πίνακας Αποστάσεων

	A/B/C/D	E/F
A/B/C/D	0	5
E/F	5	0



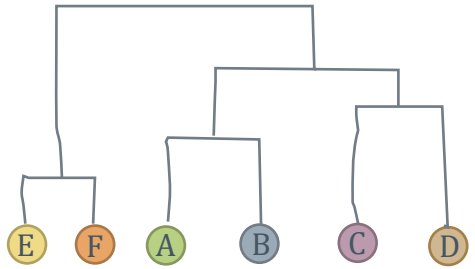
Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)



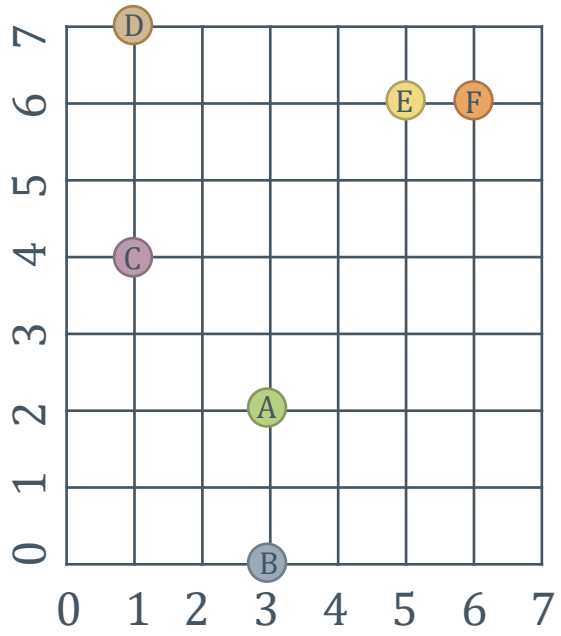
Πίνακας Αποστάσεων

	A/B/C/DE/F
A/B/C/D/E/F	0



Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)

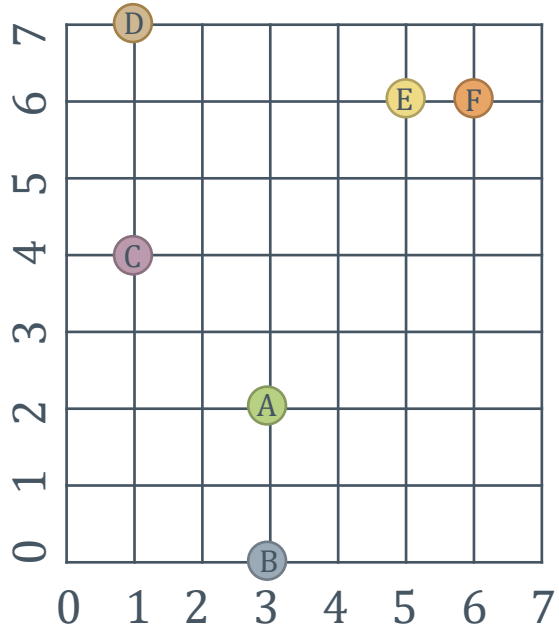


Πίνακας Αποστάσεων

	A	B	C	D	E	F
A	0	2	4	7	6	7
B	2	0	6	9	8	9
C	4	6	0	3	6	7
D	7	9	3	0	5	6
E	6	8	6	5	0	1
F	7	9	7	6	1	0

Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)

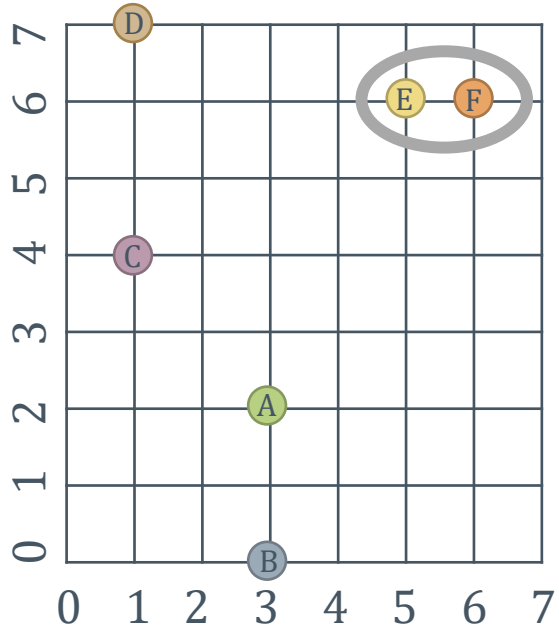


Πίνακας Αποστάσεων

	A	B	C	D	E	F
A	0	2	4	7	6	7
B	2	0	6	9	8	9
C	4	6	0	3	6	7
D	7	9	3	0	5	6
E	6	8	6	5	0	1
F	7	9	7	6	1	0

Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)

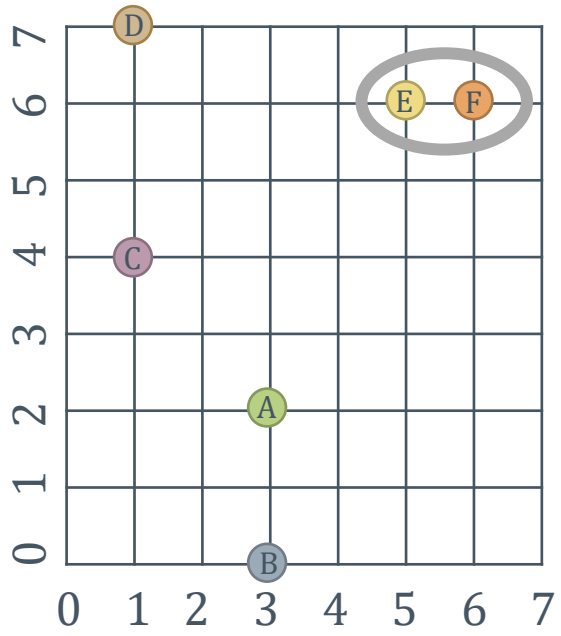


Πίνακας Αποστάσεων

	A	B	C	D	E	F
A	0	2	4	7	6	7
B	2	0	6	9	8	9
C	4	6	0	3	6	7
D	7	9	3	0	5	6
E	6	8	6	5	0	1
F	7	9	7	6	1	0

Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)

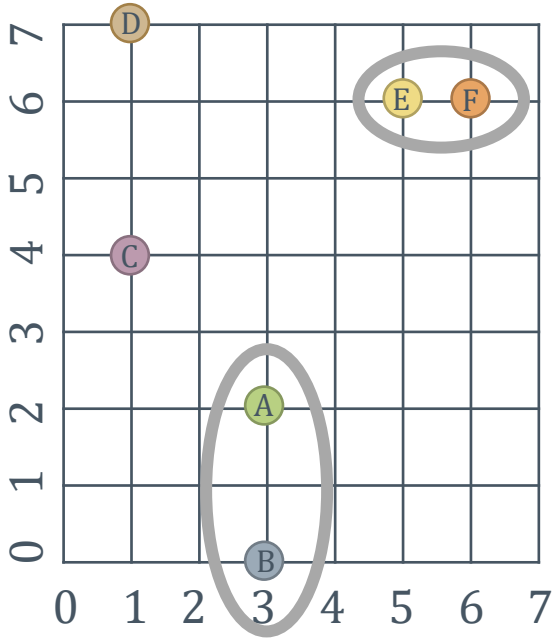


Πίνακας Αποστάσεων

	A	B	C	D	E/F
A	0	2	4	7	7
B	2	0	6	9	9
C	4	6	0	3	7
D	7	9	3	0	6
E/F	7	9	7	6	0

Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)

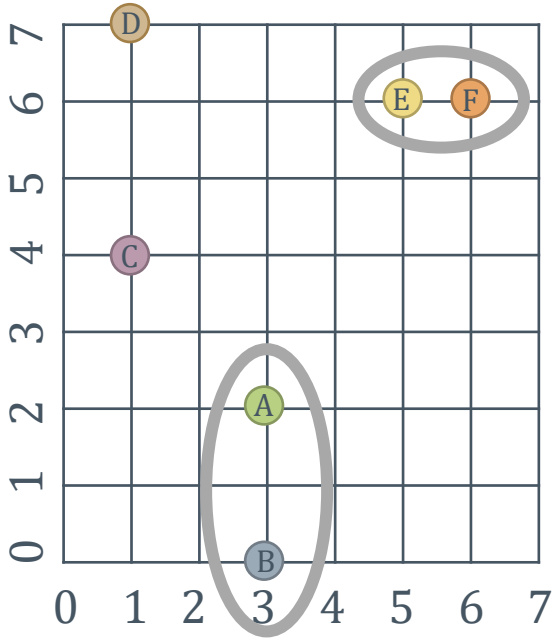


Πίνακας Αποστάσεων

	A	B	C	D	E/F
A	0	2	4	7	7
B	2	0	6	9	9
C	4	6	0	3	7
D	7	9	3	0	6
E/F	7	9	7	6	0

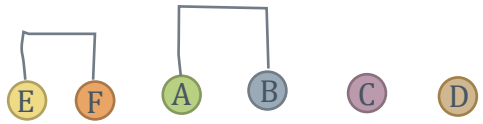
Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)



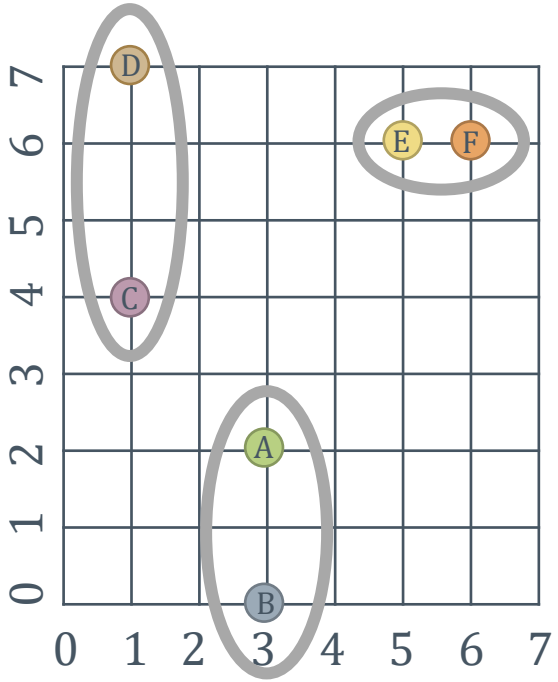
Πίνακας Αποστάσεων

	A/B	C	D	E/F
A/B	0	6	9	9
C	6	0	3	7
D	9	3	0	6
E/F	9	7	6	0



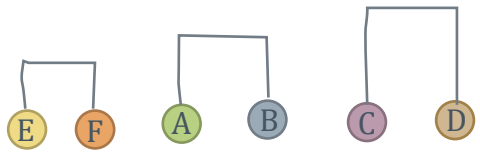
Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)



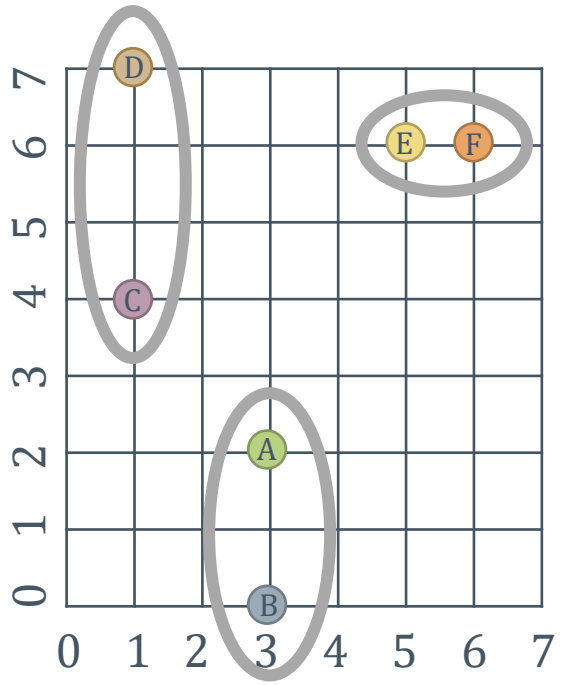
Πίνακας Αποστάσεων

	A/B	C	D	E/F
A/B	0	6	9	9
C	6	0	3	7
D	9	3	0	6
E/F	9	7	6	0



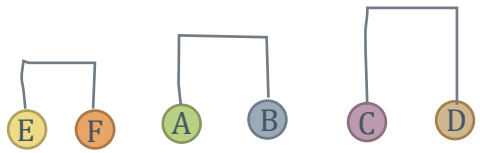
Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)



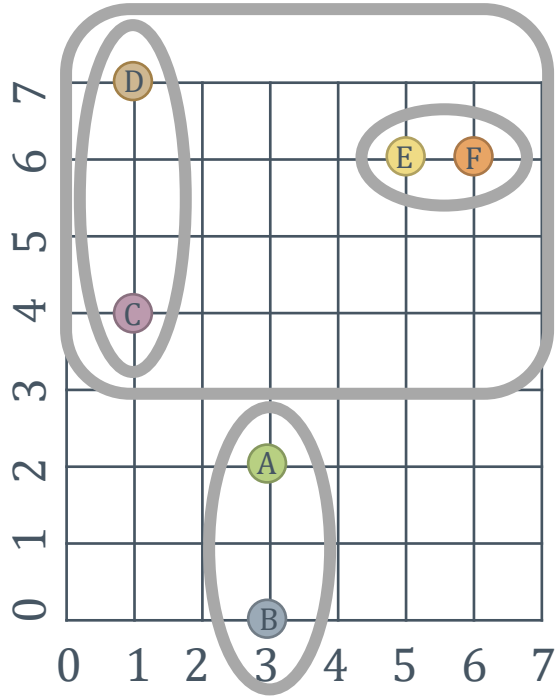
Πίνακας Αποστάσεων

	A/B	C/D	E/F
A/B	0	9	9
C/D	9	0	7
E/F	9	7	0



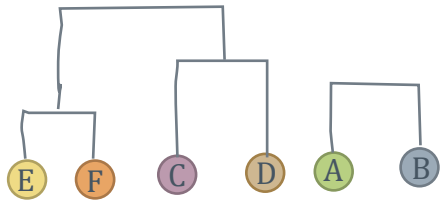
Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)



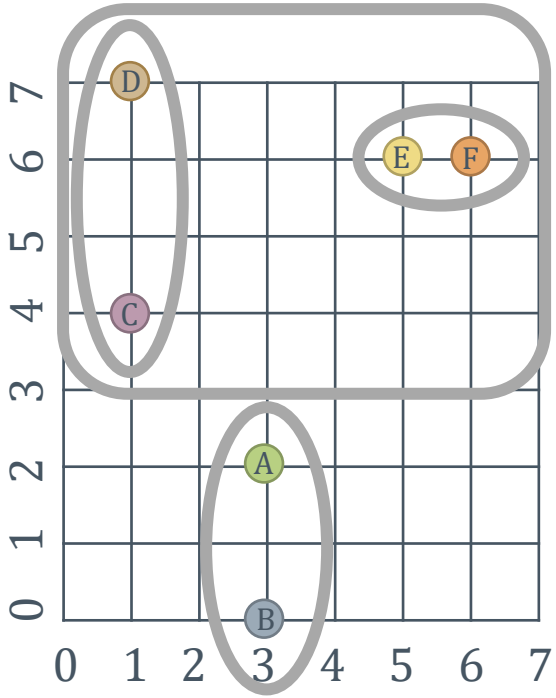
Πίνακας Αποστάσεων

	A/B	C/D	E/F
A/B	0	9	9
C/D	9	0	7
E/F	9	7	0



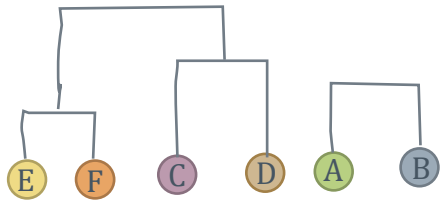
Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)



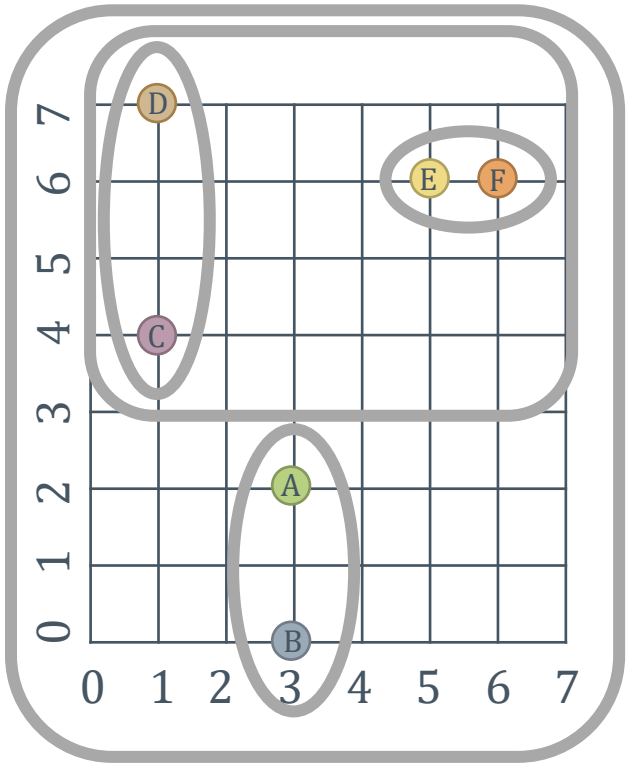
Πίνακας Αποστάσεων

	A/B	C/D/E/F
A/B	0	9
C/D/E/F	9	0



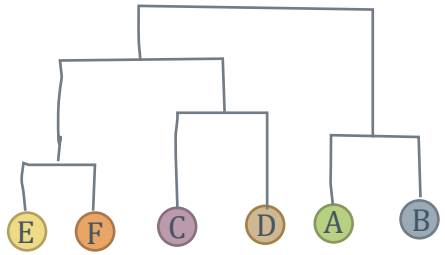
Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)

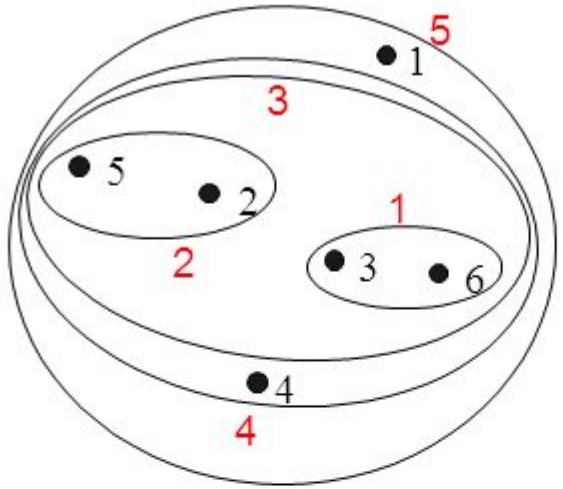


Πίνακας Αποστάσεων

	A/BC/D/E/F
A/BC/D/E/F	0

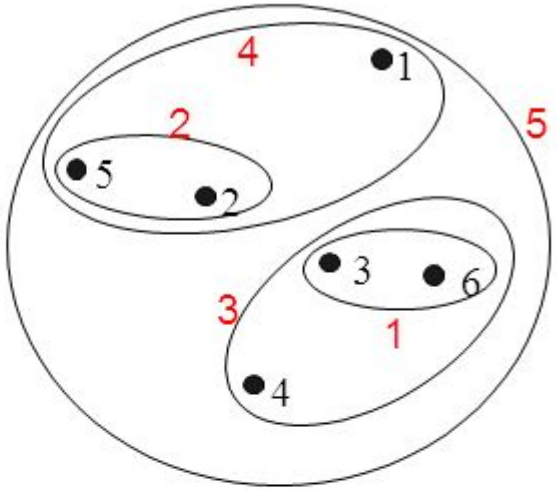


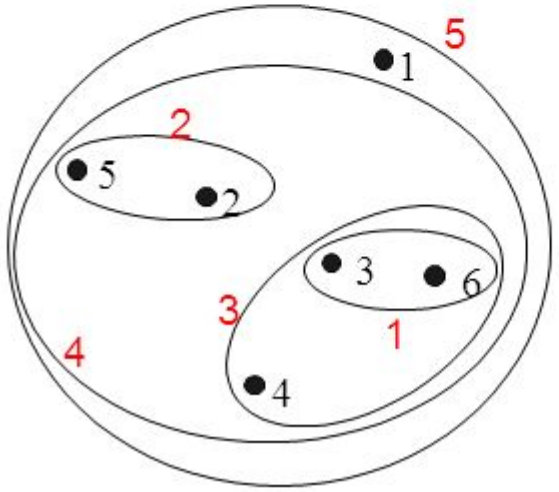
Ιεραρχική Ομαδοποίηση



MIN

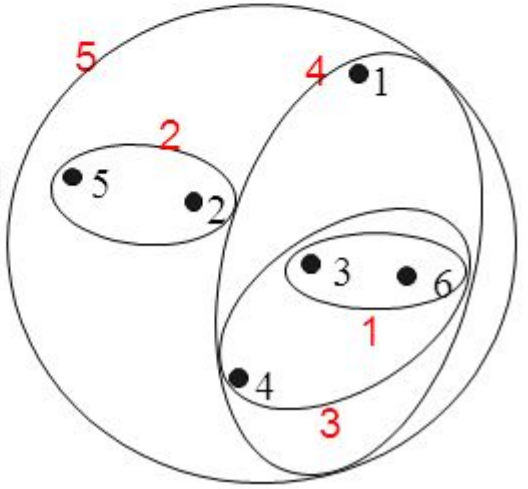
MAX





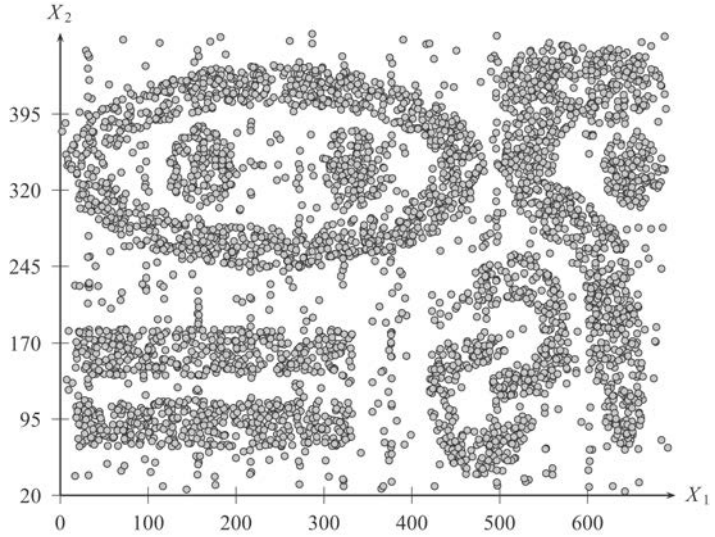
Group Average

Ward's Method



Συσταδοποίηση βασισμένη στην πυκνότητα

– Οι μέθοδοι που βασίζονται στην πυκνότητα είναι σε θέση να εξορύξουν κοίλες συστάδες, τις οποίες ενδέχεται να δυσκολευτούν να βρουν οι μέθοδοι που βασίζονται στην απόσταση.



Η μέθοδος DBSCAN: Γειτονιά και σημεία-πυρήνες

- Ορίζουμε μια μπάλα ακτίνας ϵ γύρω από ένα σημείο $\mathbf{x} \in \mathbb{R}^d$, η οποία ονομάζεται ϵ–γειτονιά του $\mathbf{x}$, ως εξής:

$$N_\epsilon(\mathbf{x}) = B_d(\mathbf{x}, \epsilon) = \{\mathbf{y} \mid \ \delta(\mathbf{x}, \mathbf{y}) \leq \epsilon\}$$

- Εδώ, το $\delta(\mathbf{x}, \mathbf{y})$ αναπαριστά την απόσταση των σημείων $\mathbf{x}$ και $\mathbf{y}$· συνήθως υποθέτουμε ότι πρόκειται για την Ευκλείδεια απόσταση.

- Λέμε ότι το $\mathbf{x}$ είναι ένα *σημείο-πυρήνας* αν υπάρχουν τουλάχιστον *minpts* σημεία στην ϵ–γειτονιά του σημείου, δηλαδή αν ισχύει $|N_\epsilon(\mathbf{x})| \geq minpts$.

- Τα *οριακά σημεία* δεν ικανοποιούν το κατώφλι *minpts*· για τα σημεία αυτά ισχύει μεν η ανισότητα $|N_\epsilon(\mathbf{x})| < minpts$, αλλά ανήκουν στην ϵ–γειτονιά κάποιου σημείου-πυρήνα $\mathbf{z}$, δηλαδή το $\mathbf{x} \in N_\epsilon(\mathbf{z})$.

- Αν ένα σημείο δεν είναι ούτε σημείο-πυρήνας ούτε οριακό σημείο, τότε ονομάζεται *σημείο θορύβου* ή έκτοπη παρατήρηση (outlier).

Η μέθοδος DBSCAN: Προσπελασιμότητα και συστάδα βασισμένη στην πυκνότητα

- Ένα σημείο **x** είναι *άμεσα προσπελάσιμο ως προς την πυκνότητα* από κάποιο άλλο σημείο **y** αν το $\mathbf{x} \in N_\epsilon(\mathbf{y})$ και το **y** είναι σημείο-πυρήνας.

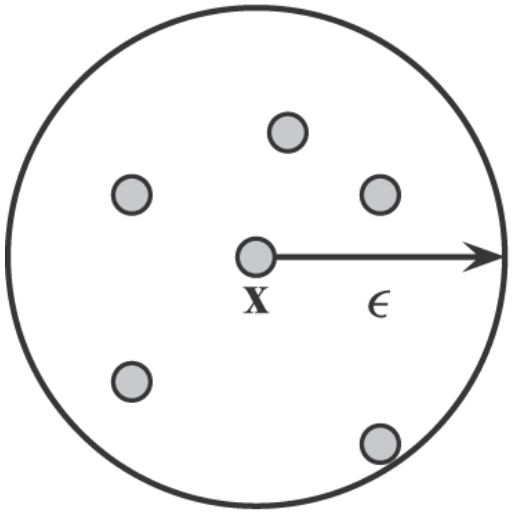
- Ένα σημείο **x** είναι *προσπελάσιμο ως προς την πυκνότητα* από το **y** αν υπάρχει μια αλυσίδα σημείων $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_l$ τέτοια ώστε να ισχύει $\mathbf{x} = \mathbf{x}_0$ και $\mathbf{y} = \mathbf{x}_l$, και το $\mathbf{x}_i$ να είναι άμεσα προσπελάσιμο ως προς την πυκνότητα από το $\mathbf{x}_{i-1}$ για όλα τα $i = 1, \ldots, l$.

- Με άλλα λόγια, υπάρχει ένα σύνολο σημείων-πυρήνων που οδηγεί από το **y** στο **x**.

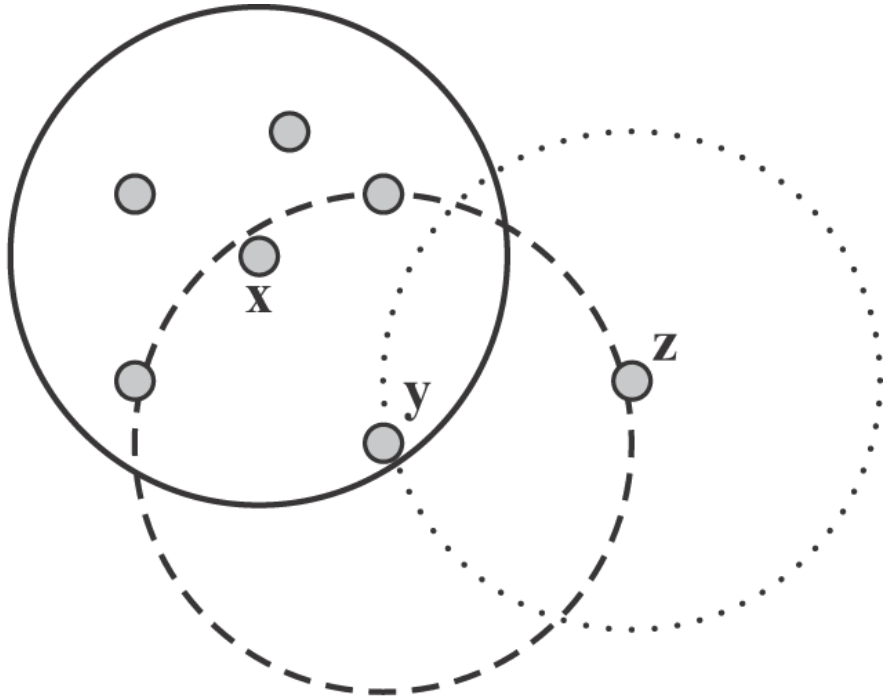
- Ορίζουμε ότι δύο οποιαδήποτε σημεία **x** και **y** είναι *συνδεδεμένα ως προς την πυκνότητα* αν υπάρχει ένα σημείο-πυρήνας **z** τέτοιο ώστε τόσο το **x** όσο και το **y** να είναι προσπελάσιμα ως προς την πυκνότητα από το **z**.

- Μια *συστάδα βασισμένη στην πυκνότητα* ορίζεται ως ένα μέγιστο σύνολο σημείων συνδεδεμένων ως προς την πυκνότητα.

Σημεία-πυρήνες, οριακά σημεία και σημεία θορύβου



(α)



(β)

Ο αλγόριθμος DBSCAN

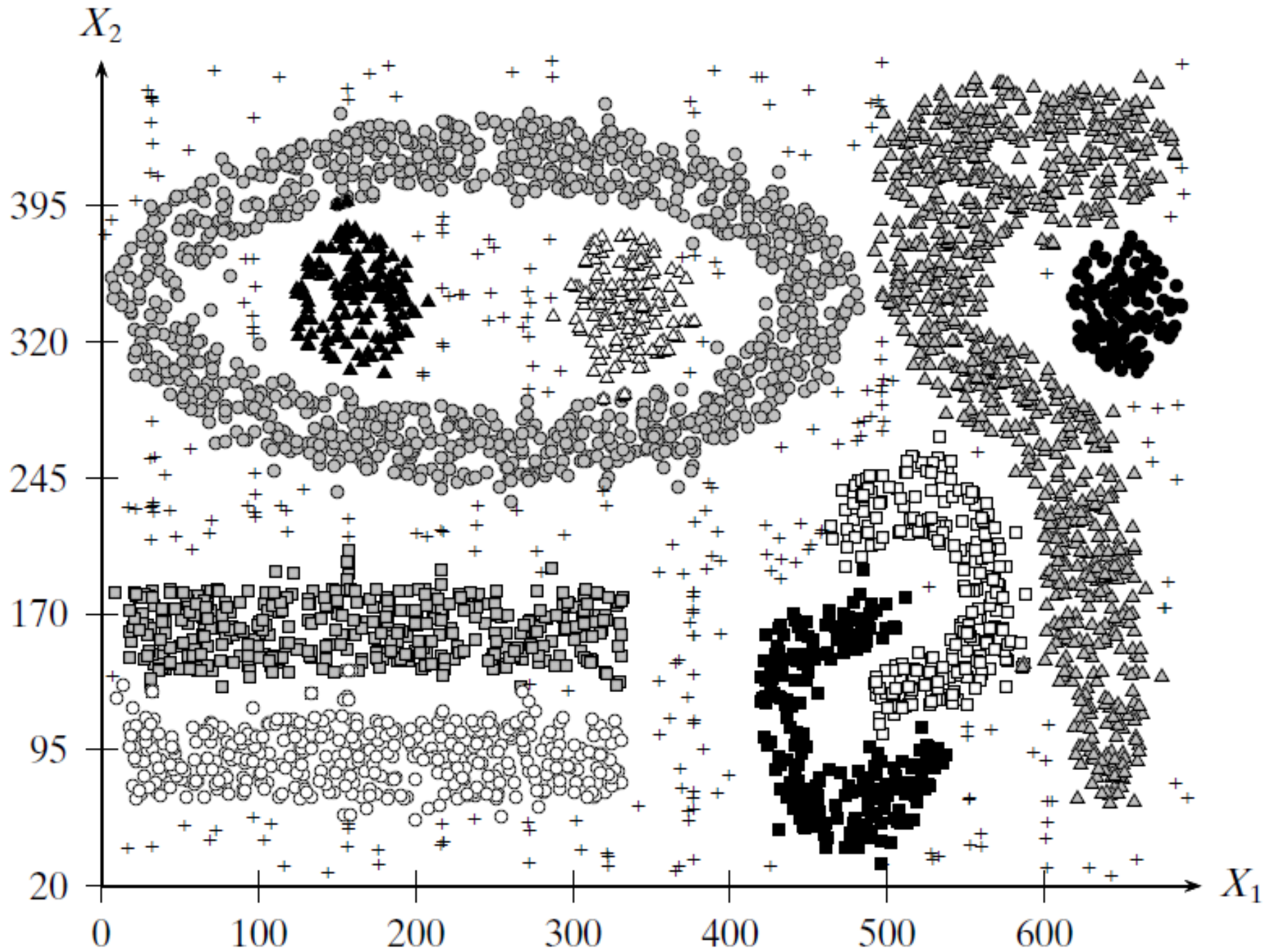
- Αρχικά, ο αλγόριθμος DBSCAN υπολογίζει την ϵ−γειτονιά για κάθε σημείο $\mathbf{x}_i$ του συνόλου δεδομένων **D**, και ελέγχει αν πρόκειται για σημείο-πυρήνα.
  - Επίσης, ορίζει ότι το αναγνωριστικό συστάδας είναι $id(\mathbf{x}_i) = Ø$ για όλα τα σημεία· έτσι υποδεικνύει ότι τα σημεία δεν έχουν αντιστοιχιστεί σε κάποια συστάδα.

- Ξεκινώντας από κάθε μη αντιστοιχισμένο σημείο-πυρήνα, η μέθοδος βρίσκει με αναδρομικό τρόπο όλα τα συνδεδεμένα σημεία ως προς την πυκνότητα για το κάθε σημείο, τα οποία αντιστοιχίζονται στην ίδια συστάδα.

- Κάποιο οριακό σημείο ενδέχεται να είναι προσπελάσιμο από σημεία-πυρήνες που ανήκουν σε περισσότερες από μία συστάδες, τα οποία μπορεί να έχουν αντιστοιχιστεί αυθαίρετα σε μία από τις συστάδες ή σε όλες τις συστάδες (αν επιτρέπονται οι επικαλυπτόμενες συστάδες).

- Τα συγκεκριμένα σημεία δεν ανήκουν σε οποιαδήποτε συστάδα και είτε θεωρούνται έκτοπες παρατηρήσεις είτε εκλαμβάνονται ως θόρυβος.

- Κάθε συστάδα του DBSCAN είναι μια μέγιστη συνεκτική συνιστώσα του γραφήματος των οριακών σημείων.
  - Ο αλγόριθμος εμφανίζει ευαισθησία στην επιλογή του ϵ, ειδικά αν οι συστάδες έχουν διαφορετικές πυκνότητες.


X_2
395
320
245
170
95
20
0
100
200
300
400
500
600
X_1

https://www.statskingdom.com/cluster-analysis.html

https://maayanlab.cloud/clustergrammer/

https://biit.cs.ut.ee/clustvis/