

*machine learning*

---

# Introduction to Clustering

Lecture VII

פיתוח:  
ד"ר יהונתן שלר  
משה פרידמן



---

# נושאי השיעור

---

- למידה לא מונחית
- מבוא ל-clustering
- אלגוריתם K-means
- Dimensionality Reduction – בהמשך ...



# למידה מונחית מול למידה לא מונחית

❖ למידה מונחית – לאלגוריתם יש מטרה ברורה: לחזות פלט רצוי, בהנתן קלט מסוים. בשלב האימון נתונים דגימות של זוגות  $\{(X^{(i)}, y^{(i)})\}$  ועל פיהם נבנה מודל החיזוי

❖ למידה לא מונחית – מטרת האלגוריתם ברורה פחות (אין פידבק ברור האם הפלט הנוצר הינו נכון). בשלב האימון נתונים דגימות של  $\{x^{(i)}\}$  (האם ללא ה  $y$  שלהם)



---

# סוגי בעיות בלמידה לא מונחית

---

**Clustering (אישכול):** נייצג כל דוגמה על ידי "אב-טיפוס" (prototype), למשל k-means, GMM ואחרים.

**Dimensionality reduction (הורדת המימדיות):** נייצג כל דוגמה על ידי מספר קטן יותר של מאפיינים. למשל Principal Components Analysis, Factor Analysis ואחרים.

**Density estimation (הערכת צפיפות):** נעריך את ההתפלגות מעל מרחב ה-data



---

# אישכול "Clustering"

---

❖ Cluster Analysis היא הפעולה של חלוקת קבוצה לתתי קבוצות ("אשכולות"/Clusters) כך ש:

❖ אובייקטים באותו אשכול "דומים" זה לזה

❖ אובייקטים באשכולות שונים, אינם "דומים" זה לזה.

## Unsupervised❖



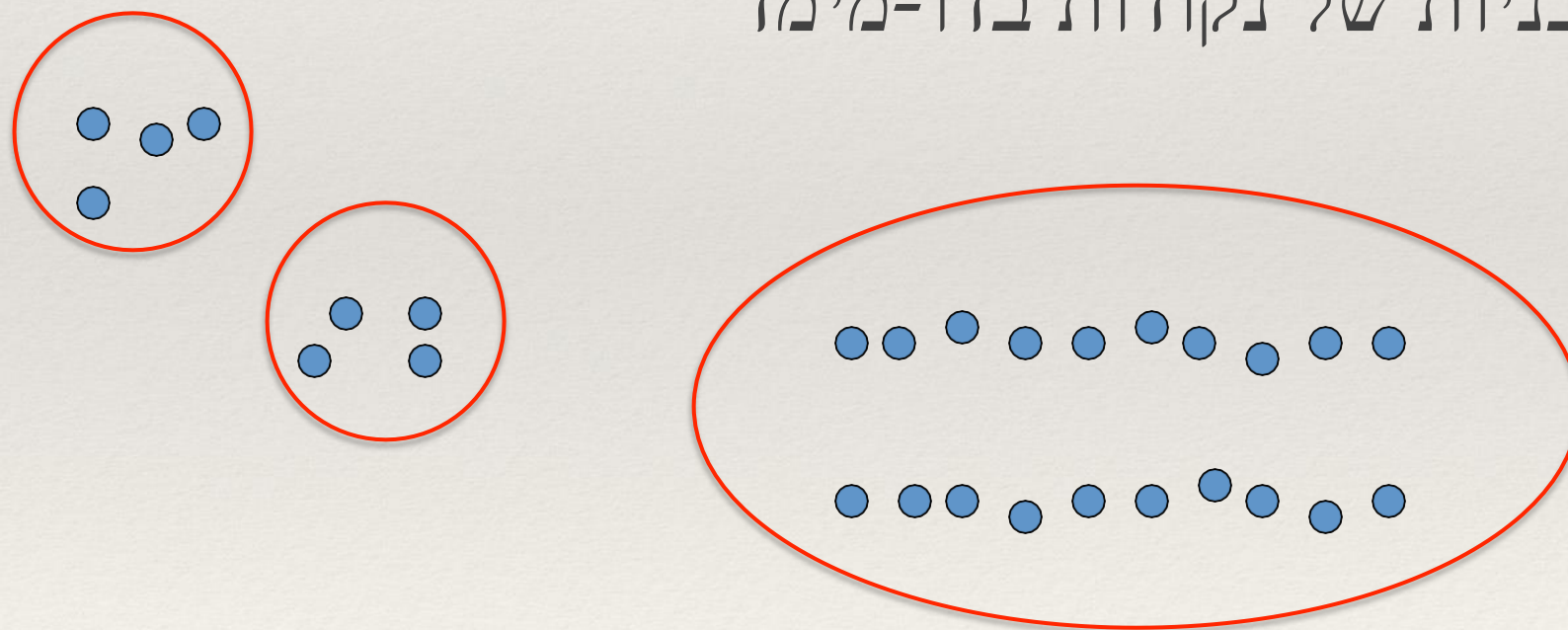
# Clustering – שאלה 1: איזו חלוקה מהווה חלוקה "נכונה"?

## אפשרות א'

❖ רעיון בסיסי: לקבץ יחד דוגמאות "דומות"

❖ למשל רוצים לקבץ ביחד לקוחות "דומים" לקבוצות (למשל ע"מ שנוכל למכור ולתמוך בהם באופן דומה).

❖ דוגמא: תבניות של נקודות בדו-מימד





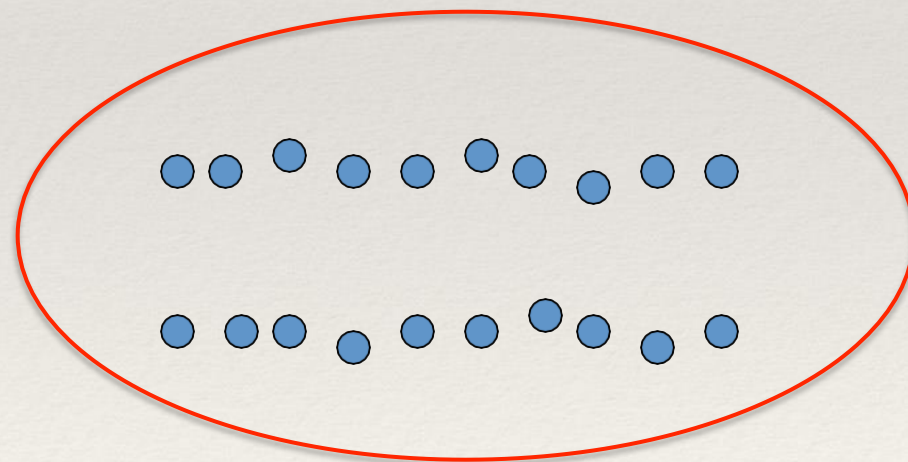
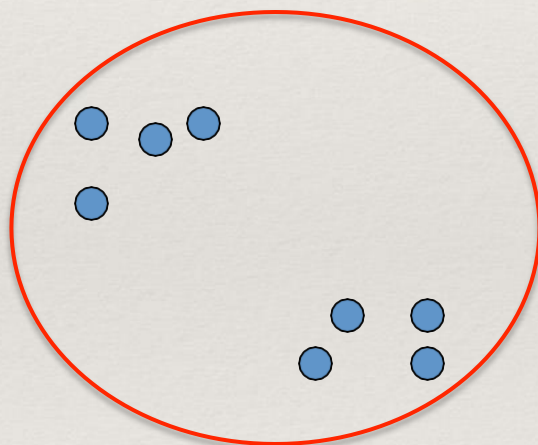
# Clustering – שאלה 1: איזו חלוקה מהווה חלוקה "נכונה"?

## אפשרות ב'

❖ רעיון בסיסי: לקבץ יחד דוגמאות "דומות"

❖ למשל רוצים לקבץ ביחד לקוחות "דומים" לקבוצות (למשל ע"מ שנוכל למכור ולתמוך בהם באופן דומה).

❖ דוגמא: תבניות של נקודות בדו-מימד





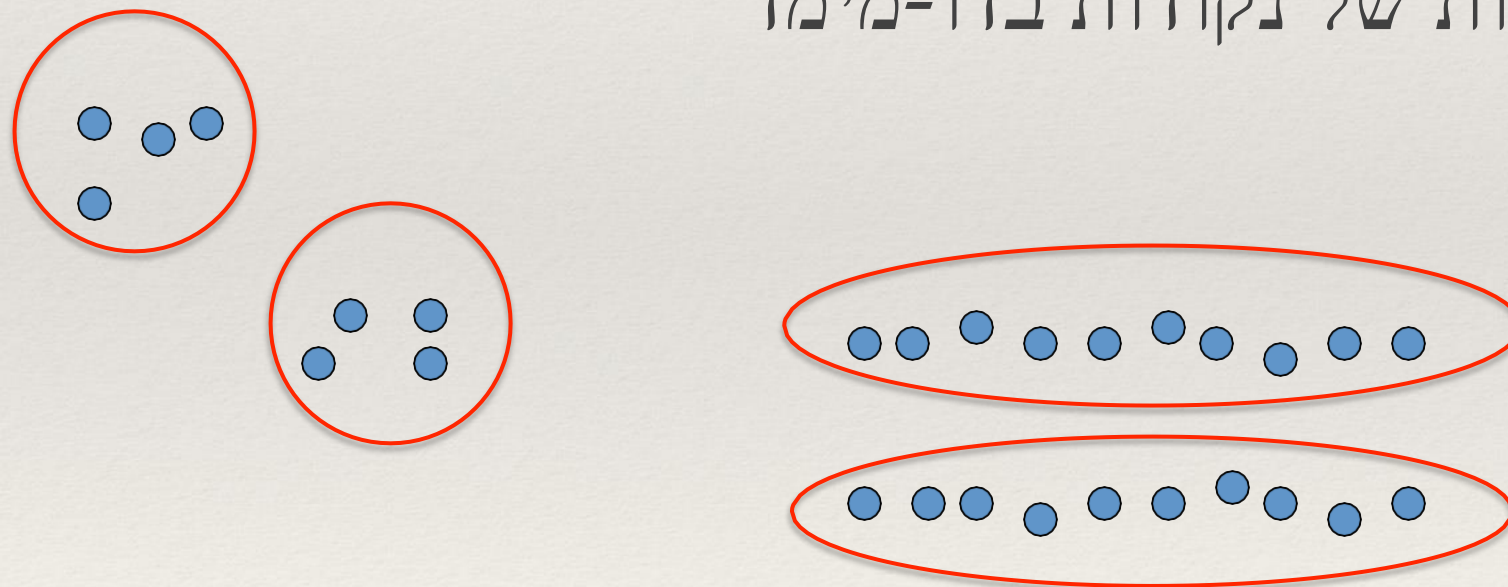
# Clustering – שאלה 1: איזו חלוקה מהווה חלוקה "נכונה"?

## אפשרות ג'

❖ רעיון בסיסי: לקבץ יחד דוגמאות "דומות"

❖ למשל רוצים לקבץ ביחד לקוחות "דומים" לקבוצות (למשל ע"מ שנוכל למכור ולתמוך בהם באופן דומה).

❖ דוגמא: תבניות של נקודות בדו-מימד





# Clustering - שאלה 2: כיצד נמדוד דמיון?



Similarity is hard to define, but...

*"We know it when we see it"*

Credit:  
Eamonn  
Keogh



# Clustering - שאלה 2: כיצד נמדוד דמיון?

❖ רעיון בסיסי: לקבץ יחד דוגמאות דומות

❖ דוגמאות:

❖ תבניות של נקודות בדו-מימד

❖ דוגמה נוספת – קיבוץ לקוחות דומים.

❖ כיצד נמדוד "דמיון"?

❖ אפשרות אחת: 2 נקודות (דוגמאות) יחשבו "דומות", אם יהיה ביניהן מרחק קטן.

❖ למשל מרחק אוקלידי קטן:  $\text{dist}(\vec{x}_1, \vec{x}_2) = \|\vec{x}_1 - \vec{x}_2\|_2$

❖ מסקנה 1: כמו שכבר מבינים: תוצאות האישכול תלויות במידה רבה בפונקציות המרחק אותן נבחר..



# Clustering - נניח שנמדוד דמיון על ידי מרחק (קטן)

## שאלה 3: בין מי למי מודדים מרחק?

מוטיבציה: רוצים לחלק את הלקוחות לקבוצות.

❖ נוכל להחליט על "דמיון" בין הלקוחות, על ידי מציאת לקוחות עם מרחק (קטן ביניהם), אך בין מי למי מודדים את המרחק?

❖ חלק מהאלגוריתמים דורשים **מרחק בין נקודה  $x_i$**  (דוגמה  $x_i$ , או לקוח מסויים, כמו במקרה שלנו) **לבין קבוצת נקודות  $A$**  (קבוצת דוגמאות  $A$ , או קבוצת לקוחות, כמו במקרה שלנו).

❖ במקרה זה נמדוד את המרחק  $d(x, A)$

❖ אלגוריתמים אחרים דורשים **מרחק בין קבוצת נקודות  $A$**  (קבוצת דוגמאות  $A$ , או קבוצת לקוחות, במקרה שלנו) **לבין קבוצת נקודות  $B$**  (קבוצת דוגמאות  $B$ , או קבוצת לקוחות אחרת, במקרה שלנו).

❖ במקרה זה נמדוד את המרחק  $d(A, B)$



---

# Clustering - מוטיבציה אפליקטיבית

---

ניקה צעד אחד אחורה ...

מוטיבציה אפליקטיביות:

- עבור איזה סוגי מידע נרצה לבצע clustering?
- דוגמאות עבור אפליקציות ל-clustering



---

# Clustering - מוטיבציה אפליקטיבית

## סוגי data שונים עבור clustering

---

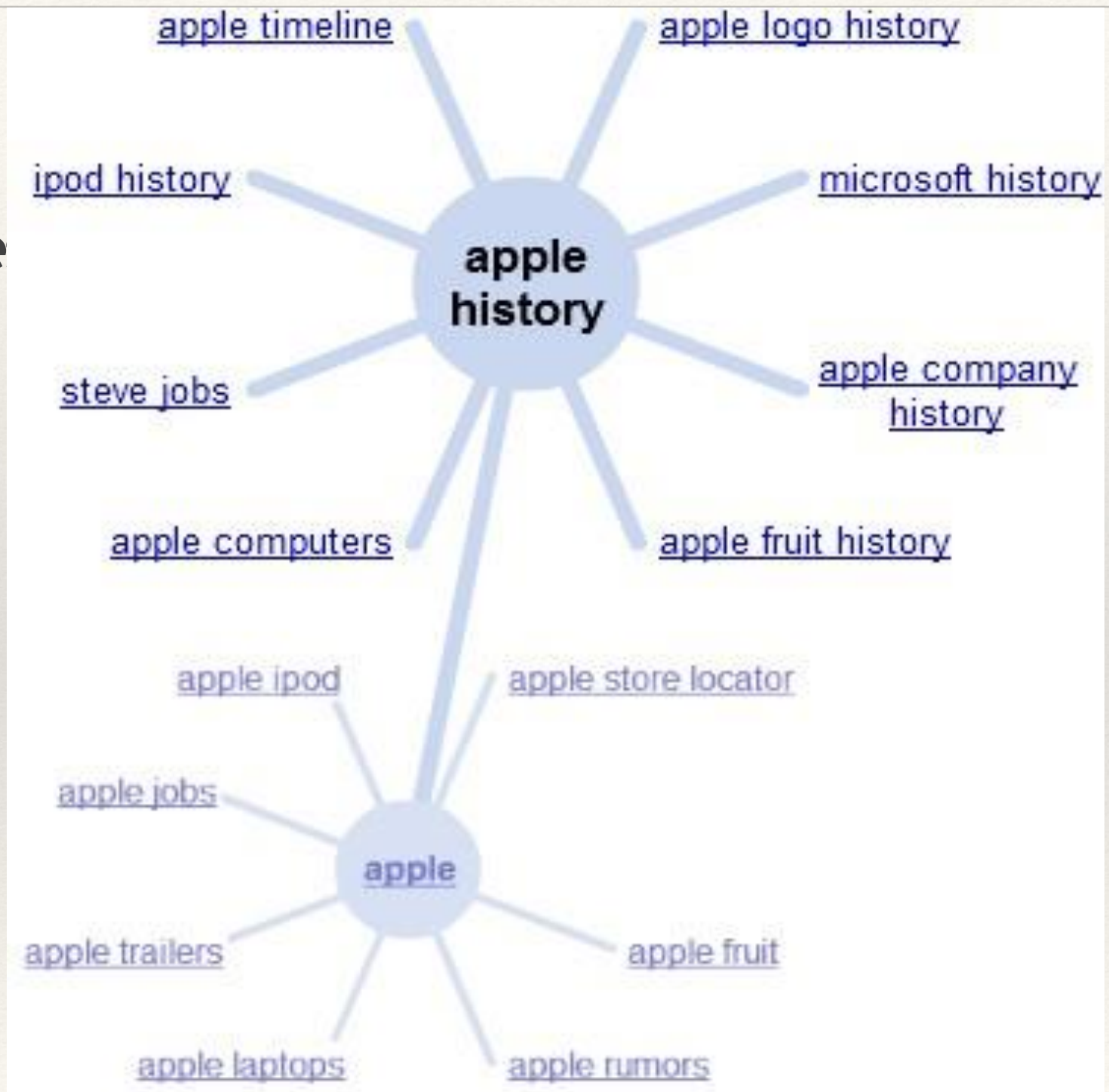
- ❖ Numerical data
- ❖ Categorical data: e.g. demographic, many times binary (has some category or not)
- ❖ Text data (popular in social media, web, social nets):
  - ❖ Features: high dimensional, sparse, values corresponding to word frequencies
  - ❖ Methods: combinations of: k-means, agglomerative (hierarchical); topic modeling; co-clustering



# Clustering - מוטיבציה אפליקטיבית

## דוגמאות אפליקטיביות – אשכול חיפושים לצורך שיפור החיפוש

Clustering search queries





---

# Clustering - מוטיבציה אפליקטיבית

## סוגי data שונים עבור clustering (2)

---

- ❖ Multimedia data [image, audio, video] (e.g., flicker, YouTube):
  - ❖ Multi-model (often combine with text data)
  - ❖ Contextual: containing both behavioral and contextual attributes
  - ❖ Images: position of pixel represents its context, value represents its behavior
  - ❖ Video & music: temporal ordering of records represent its meaning



---

# Clustering - מוטיבציה אפליקטיבית

## דוגמאות אפליקטיביות – אשכול מקטעים בתמונה

---

### Image segmentation

Goal: Break up the image into meaningful or perceptually similar regions





# Clustering - מוטיבציה אפליקטיבית

## דוגמאות אפליקטיביות – אשכול מאמרי עיתונות

### Cluster news articles

Google

News

U.S. edition

Classic

Top Stories

Boston Red Sox

Apple Inc.

Angela Merkel

Nokia Lumia

Bashar al-Assad

Republican Party

Facebook

Pets

Katy Perry

Bushfires in Australia

New York, New York

Recommended

U.S.

World

Sci/Tech

Business

More Top Stories

Health

Spotlight

Elections

Entertainment

Sports

Technology

Science

Top Stories

Teen suspect saw movie moments after allegedly killing beloved Massachusetts ...

Fox News - 8 minutes ago

The 14-year-old student who authorities say murdered a beloved math teacher at a Massachusetts high school admitted to police that he slashed her throat with a box cutter, a source told MyFoxBoston.

Colleen Ritzer, slain Danvers High School teacher, remembered as passionate ... CBS News

14-Year-Old Charged in Brutal Murder of Massachusetts Teacher New York Magazine

Highly Cited: 14-year-old student held without bail in slaying of Danvers High teacher Boston.com

Opinion: Heslam: Heartbroken friends say Colleen was born to teach Boston Herald

In Depth: Student, 14, arraigned in murder of Mass. teacher USA TODAY

Wikipedia: Danvers, Massachusetts

See realtime coverage »

Obamacare contractors tell their stories at congressional hearing

CNN - 40 minutes ago

Washington (CNN) -- [Breaking news update at 10:09 a.m.]. [URGENT - Congress-Obamacare-Testing]. (CNN) -- A contractor on the problem-plagued government website for President Barack Obama's signature health care reforms said Thursday his ...

Hearing on health care website today to focus on blame WXIA-TV

Contractors Point Fingers Over Health-Law Website AllThingsD

See realtime coverage »

EU leaders meet amid concern about US spying claims

CNN - 1 hour ago

(CNN) -- European Union leaders are meeting Thursday in Brussels for a summit that may be overshadowed by anger about allegations that the United States has been spying on its European allies.

Germany summons US ambassador over spying claims USA TODAY

Germany Summons US Envoy Over Alleged NSA Spying ABC News

Highly Cited: Readout of the President's Phone Call with Chancellor Merkel of Germany Whitehouse.gov (press release)

From Germany: Press Review: Outrage over NSA eavesdropping Deutsche Welle

Opinion: The Handyüberwachung Disaster New York Times

In Depth: US ambassador to Germany summoned in Merkel mobile row BBC News

See realtime coverage »

US jobless claims miss forecasts, trade deficit widens slightly

Reuters - 59 minutes ago

WASHINGTON | Thu Oct 24, 2013 9:19am EDT. WASHINGTON (Reuters) - The number of Americans filing new claims for unemployment benefits fell less than expected last week, but a lingering backlog of applications in California makes it difficult to get a ...


Weekly Jobless Claims Fall to 350000 Fox Business

How States Fared on Unemployment Benefit Claims ABC News


In Depth: More Americans Than Forecast Filed Jobless Claims Businessweek

See realtime coverage »


Kennedy cousin gets new trial in 1975 killing of neighbor; victim's mother ...




ABC News



Wall Street Journal



National Post



The Olympian

17



---

# Clustering - מוטיבציה אפליקטיבית

## סוגי data שונים עבור clustering (3)

---

- ❖ Time-series data: sensor data, stock market, temporal tracking, forecasting and so on data is temporal dependent
- ❖ time: context, data: behavioral
- ❖ correlation based online analysis (e.g., online clustering of stocks to find stock trickers)
- ❖ shape-based offline analysis (e.g., cluster ECG based on overall shapes)



---

# Clustering - מוטיבציה אפליקטיבית

## סוגי data שונים עבור clustering (4)

---

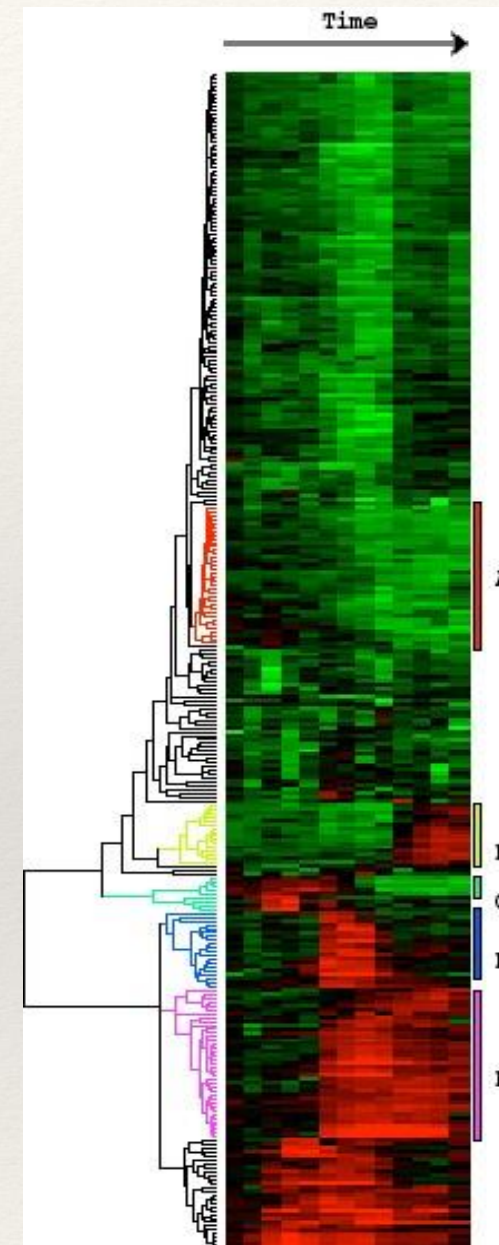
- ❖ sequence data: weblogs, biological sequences, system command sequences
  - ❖ contextual attributes: Placement (rather than time)
  - ❖ Similarity attributes : hamming distance, edit distance, longest common sequence
  - ❖ sequence clustering: suffix trees, generative model (e.g. HMM - hidden markov model)
- ❖ Stream data:
  - ❖ Real time, evolution and concept drift, single pass algorithm
  - ❖ Create efficient intermediate representation, e.g., micro-clustering



# Clustering - מוטיבציה אפליקטיבית

## דוגמאות אפליקטיביות – אשכול micro-arrays

### Clustering gene expression data



Eisen et al, PNAS 1998



# Clustering – סיכום ביניים

מה הבנו עד כה?

❖ כמה שאלות בסיסיות, כמו:

- איך ניצור את ה-clusters (לא ענינו על השאלה הזו)

- איך נמדוד דימיון

- בין מה למה נמדוד דימיון

❖ בנוסף, הבנו את המוטיבציה האפליקטיבית לשימוש ב-clustering

הנושאים (והשאלות) הבאים בהם נדון:

- התכונות הרצויות של אלגוריתם clustering

- הגישות המרכזיות לביצוע clustering



# Clustering - תכונות רצויות של אלגוריתם clustering

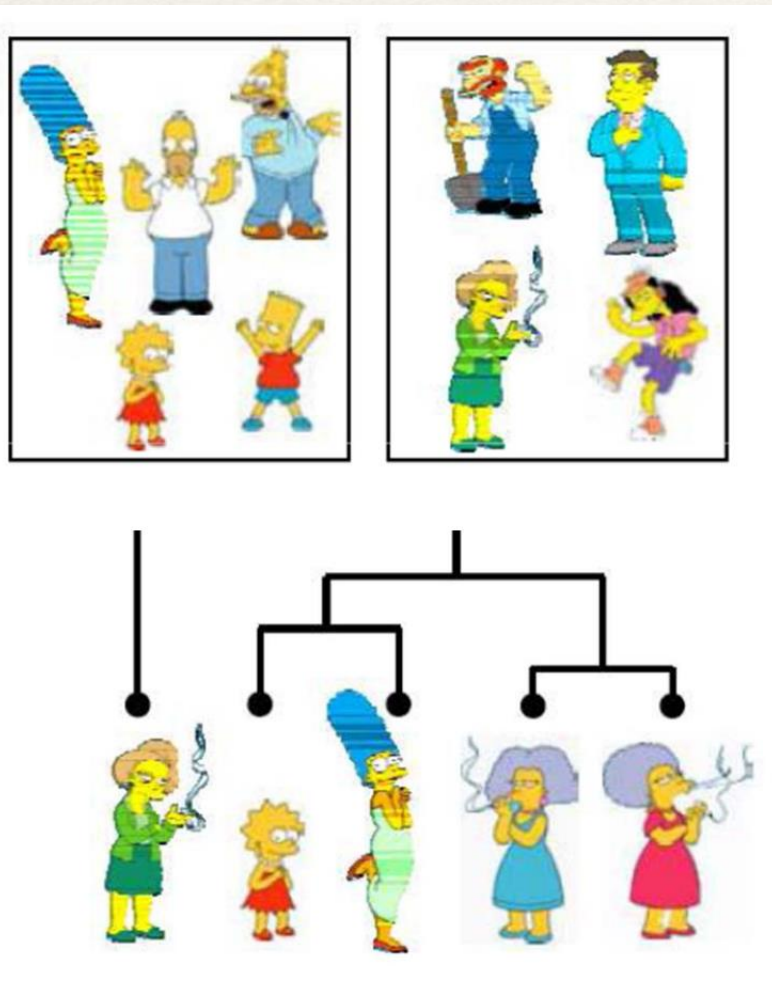
- ❖ Scalability – התמודדות גם עם זמן וגם עם מקום.
- ❖ היכולת להתמודד עם סוגי data שונים (כפי שראינו בדוגמאות הרבות לעיל).
- ❖ ידע מינימלי של התחום שאיתו מתעסקים (ה-domain, למשל, ניהול לקוחות), כדי לקבוע את הפרמטרים של הקלט.
- ❖ היכולת להתמודד עם רעש בקלט (noisy data)
- ❖ התוצאה ניתנת לפירוש (Interpretability)
- ❖ למשל מובן מה הקשר בין הלקוחות שבקבוצה
- ❖ התוצאה צריכה להיות ברת שימוש (usability)
- ❖ כתלות באפליקציה – למשל, האם הקשר בין הלקוחות עוזר לנו?



# גישות מרכזיות ב-Clustering

## 1. שיטות מבוססות חלוקה (Partitioning) –

- ❖ בהינתן קבוצה של  $n$  אובייקטים, חלק ל- $k$  תתי קבוצות ( $k \leq n$ ). כל תתי קבוצה צריכה להכיל אובייקט אחד לפחות וכל אובייקט משויך לקבוצה אחת בלבד.



## 2. שיטות היררכיות –

- ❖ בונים מבנה היררכי של תתי הקבוצות - גישות:

Agglomerative (bottom-up) ❖

Divisive(top-down) ❖



# גישות מרכזיות ב-Clustering

3. מבוססות צפיפות (Density Based) – לא נסתכל רק על המרחק בין הנקודות  
על גם האם יש "מסלול" ביניהן

4. הסתברותי וגנרטיבי:

- ❖ מניחים תצורה מסויימת של מודל גנרטיבי (mixture of Gaussian)
- ❖ שערך הפרמטרים בעזרת אלגוריתם expectation maximization (EM)  
ומשתמשים ב-dataset, כדי לשערך maximum likelihood
- ❖ שערך ההסתברות הגנרטיבית של נקודת נתונות.
- ❖ יש גמישות לכל נקודה להיות שייכת לכמה clusters



---

# שאלת סקר

---

1. איזו מהבעיות הבאות נרצה לפתור בעזרת clustering?

תשובות אפשרויות:

- א. בניית מודל שיחליט האם תמונה מסויימת היא של הולך רגל או לא
- ב. בניית מודל שימצא קבוצות חברים ברשת חברתית
- ג. בניית מודל שיחזה את תחזית מזג האוויר מחר

תשובה – ב.