SI 671/721:
Introduction to Data Mining (II)

Lecture 2

Fall 2021

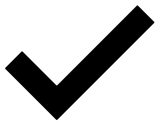
Instructor: Prof. Paramveer Dhillon

dhillonp@umich.edu

University of Michigan

UMSI

Four-Dimensions of Data Mining

- Data to be mined 
- Knowledge to be discovered
- Techniques utilized
- Applications adopted

Four-Dimensions of Data Mining

- Data to be mined
- Knowledge to be discovered (also known as “data mining functionalities”)
- Techniques utilized
- Applications adopted

Data Mining Functionalities

- Online Analytical Processing (OLAP)
- Association
- Classification
- Prediction
- Clustering
- Ranking
- Outlier/Anomaly detection

You will learn about each of these data mining functionalities, including their definitions, techniques to implement, and basic applications.

1. OLAP: Online Analytical Processing

- **OLAP** is a software for fast answering of multidimensional analytical queries from databases and data warehouses.

- Performs Analytical processing instead of transaction processing.

Transactional vs Analytical Processing

Typical transactional database

T_id	Datetime	Product	Country	Quantity
1001	09/01/11, 01:00:23	TV	USA	1
1002	09/01/11, 02:15:16	TV	Canada	2
1003	09/02/11, 15:10:41	PC	USA	1
…	…	…	…	…
15691	03/03/12, 23:59:07	VCR	Mexico	1
…	…	…	…	…

Transactional vs Analytical Processing

- Easy to query instances of transactions
- "How many TVs were sold in total?"

T_id	Datetime	Product	Country	Quantity
1001	09/01/11, 01:00:23	TV	USA	1
1002	09/01/11, 02:15:16	TV	Canada	2
1003	09/02/11, 15:10:41	PC	USA	1
…	…	…	…	…
15691	03/03/12, 23:59:07	VCR	Mexico	1
…	…	…	…	…

Transactional vs Analytical Processing

What about analytical queries that support business decisions?

T_id	Datetime	Product	Country	Quantity
1001	09/01/11, 01:00:23	TV	USA	1
1002	09/01/11, 02:15:16	TV	Canada	2
1003	09/02/11, 15:10:41	PC	USA	1
…	…	…	…	…
15691	03/03/12, 23:59:07	VCR	Mexico	1
…	…	…	…	…

Query: how many TVs were sold in the second quarter of 2011 in North America?

Transactional vs Analytical Processing

Transactional databases are not suitable for answering analytical queries.

T_id	Datetime	Product	Country	Quantity
1001	09/01/11, 01:00:23	TV	USA	1
1002	09/01/11, 02:15:16	TV	Canada	2
1003	09/02/11, 15:10:41	PC	USA	1
…	…	…	…	…
15691	03/03/12, 23:59:07	VCR	Mexico	1
…	…	…	…	…

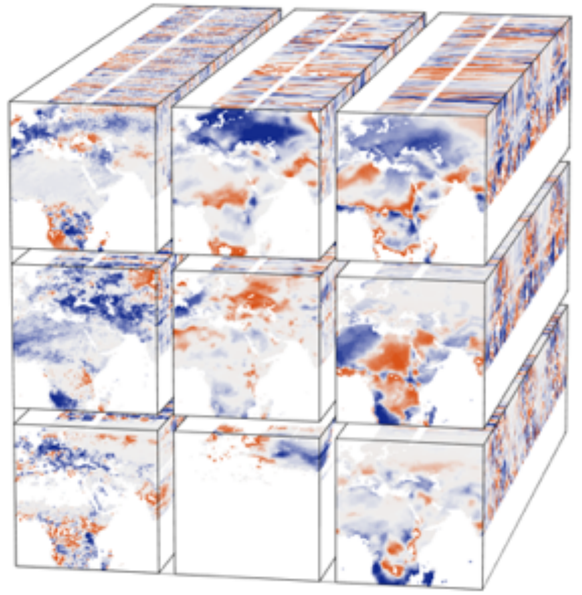
Query: how many TVs were sold in the second quarter of 2011 in North America?

OLAP: Online Analytical Processing

OLAP Techniques

- Data Warehouse: Infrastructure that integrates multiple types of transactional data and processes data for analysis.

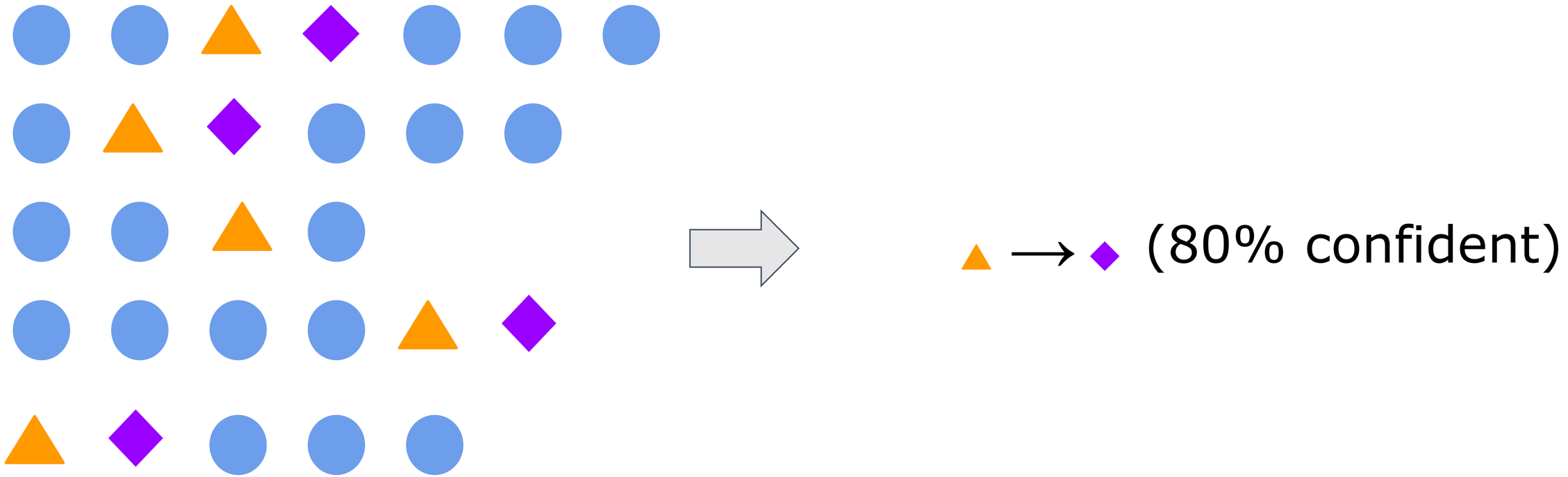
- Data Cubing: Pre-aggregate transactional data for use in future queries; improves efficiencies.



http://earthsystemdatacube.net/

2. Association

Finding inherent regularities and relations in big data



2. Association

Techniques: Co-occurrence analysis, correlation analysis, and causal inference.

Applications: Basket data analysis, catalog design, advertisement, and search engine log analysis.

2. Association

Applications: Basket data analysis, catalog design, advertisement, and search engine log analysis.

2. Association

Applications: Basket data analysis, catalog design, advertisement, and search engine log analysis.

"What are the best combination of keywords that are associated with my search engine ad?"

3. Classification

Classifier:

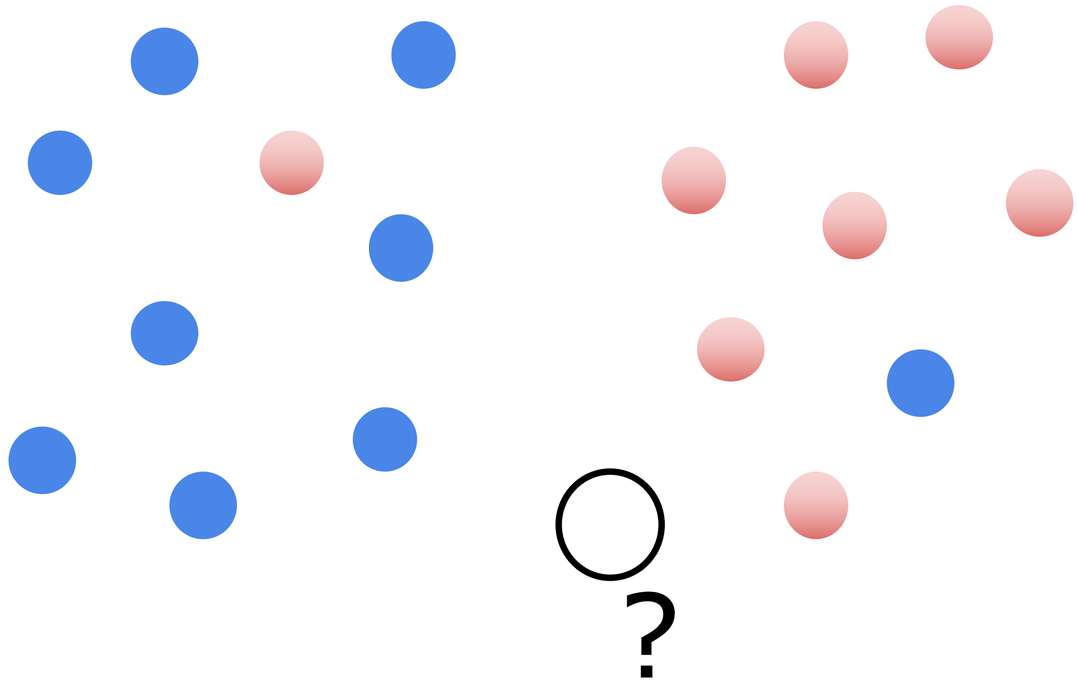
attributes (X) → categorical class label (Y)

The attributes of email spam are (X)	→	Is (Y) email spam?
The attributes of a PhD candidate are (X)	→	Should student (Y) be admitted to the PhD program?

3. Classification

Classifier:

attributes (X) → categorical class label (Y)



3. Classification

Techniques: supervised machine learning

- Classifier is trained using labeled training data to identify unlabeled data.

Applications: loan approval, spam detection, clinical diagnosis, political leaning, Web page categorization.

4. Prediction

Regression:

attributes (X) → numerical outcome (Y)

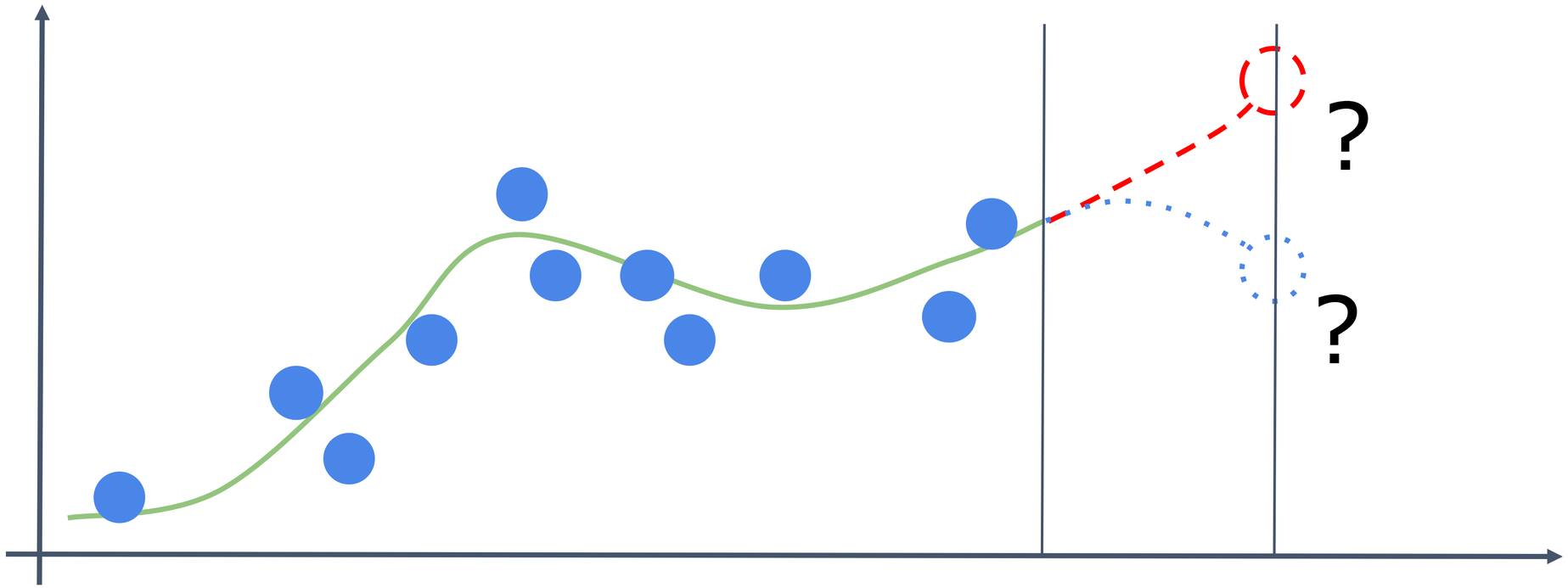
When (X) market conditions occur → Stock will increase in price by (Y) dollars

When player (X) average is above 1.5 goals/game → Team will score at least (Y) goals in their next game

4. Prediction

Regression:

attributes (X) → numerical outcome (Y)



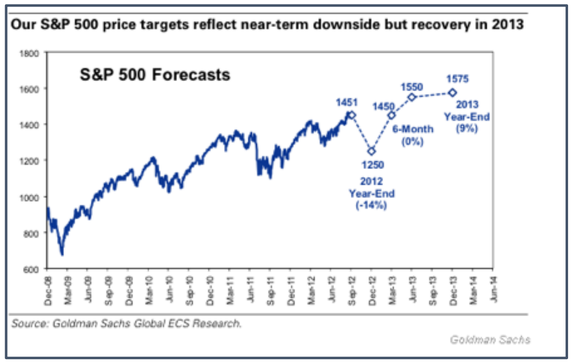
4. Prediction

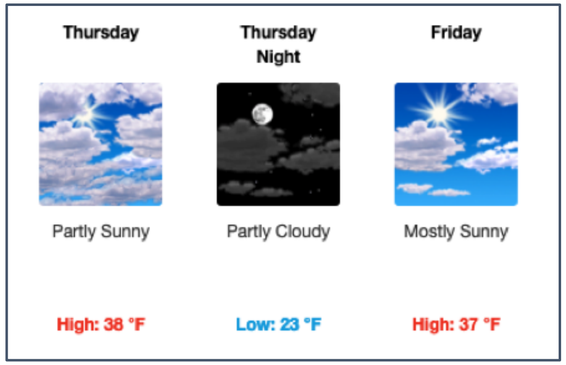
Techniques: supervised machine learning, regression analysis, time series analysis.

Applications: recommender systems, stock market prediction, weather forecasting and election prediction.

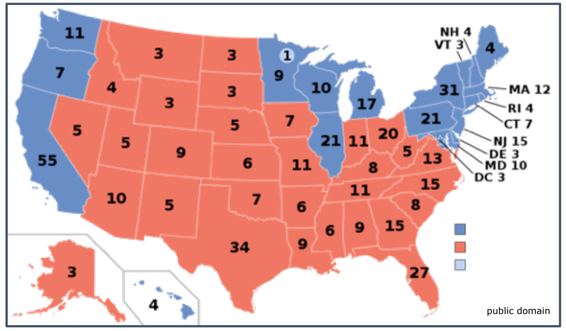
"How would a particular person rate this movie?"

★ ★ ★ ★ ★



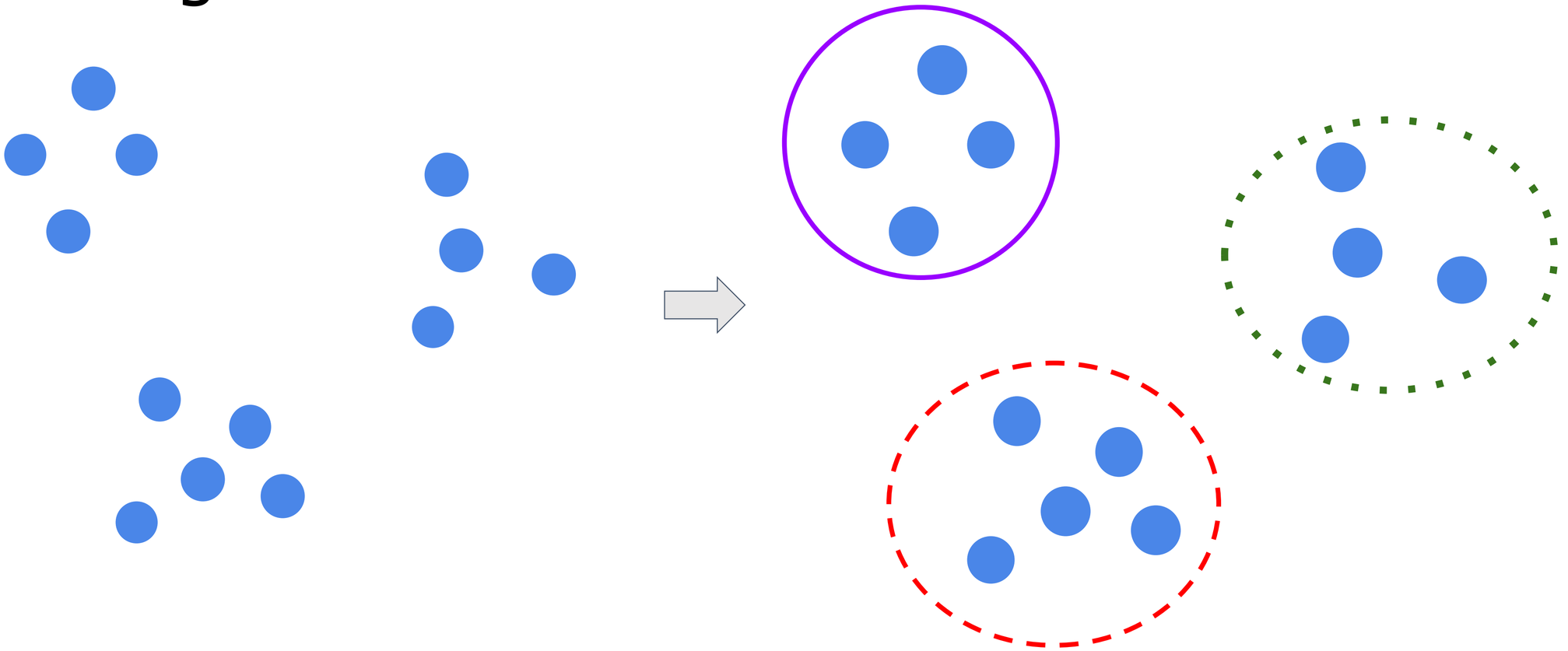


www.weather.gov



5. Clustering

Similar to classification, but no predefined classes or training data available.



5. Clustering

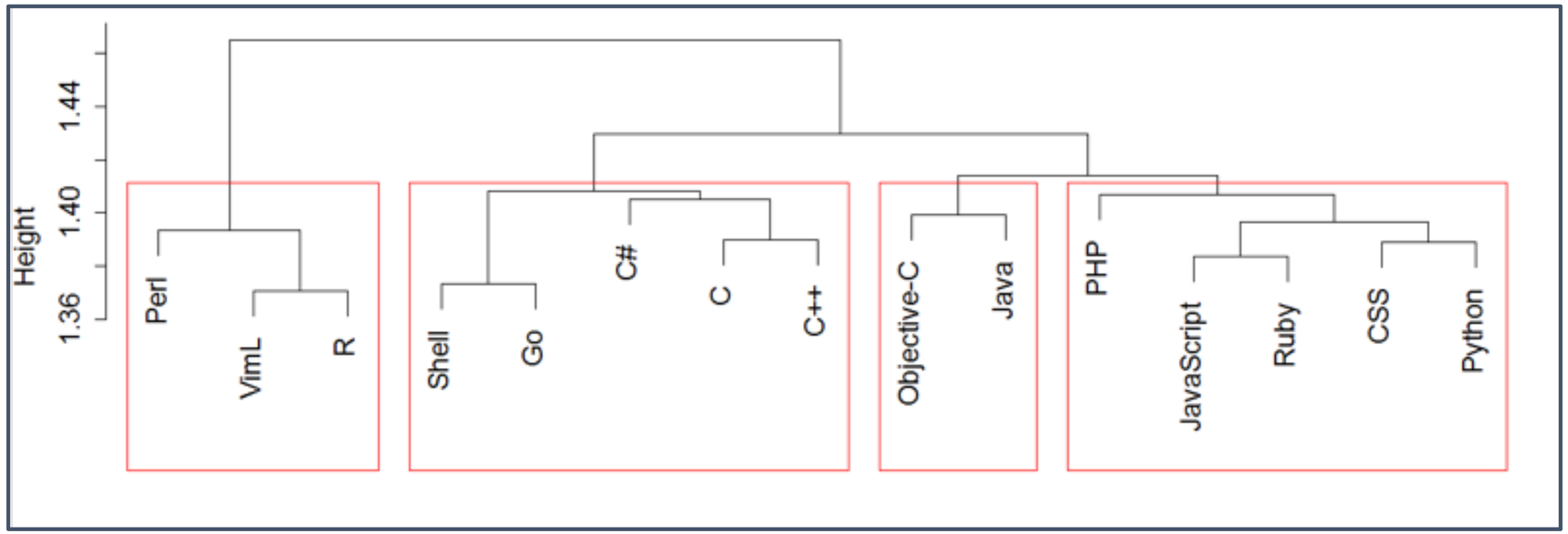
Techniques:

Unsupervised machine learning: Model uses unlabeled data examples for training.

Network analysis: A cluster of nodes is known as a “community.”

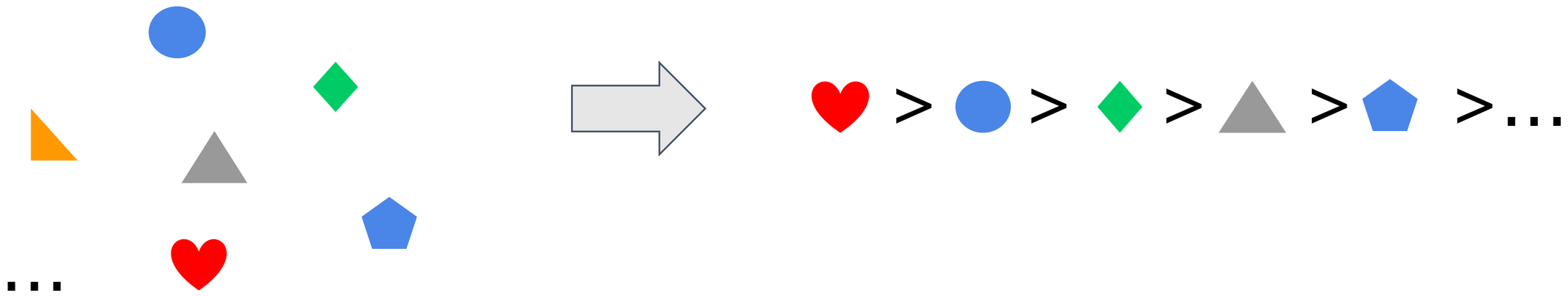
5. Clustering

Applications: community extraction (social networks), topic extraction (text), market segmentation, and taxonomies.



6. Ranking

Find a ranked order of data objects

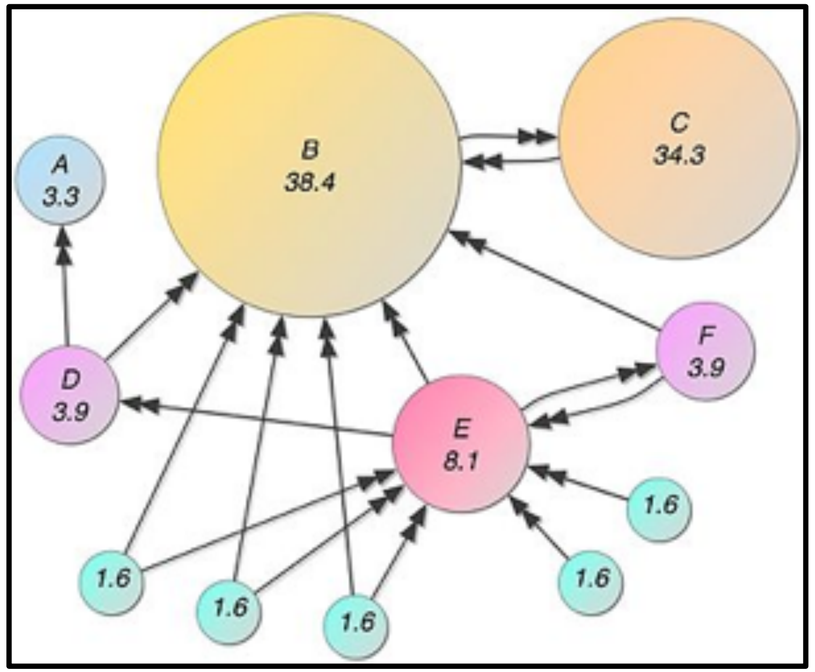


6. Ranking

Techniques:

Machine learning: Called “learning to rank” when applied to ranking.

Network analysis: Google’s “PageRank” algorithm is based on network analysis.



6. Ranking

Applications: Web search, recommender systems, bibliometrics, social networks (e.g., opinion leaders), and the NBA draft.

Instagram

Rank	Name	Followers
1	**Huda Kattan hudabeauty** $ 33,000/post	26,000,000
2	**Eleonora Pons** lelepons $ 32,500/post	25,600,000
3	**Zach King** zachking $ 30,000/post	21,500,000
4	**Sommer Ray** sommerray $ 29,000/post	18,700,000

6 more rows

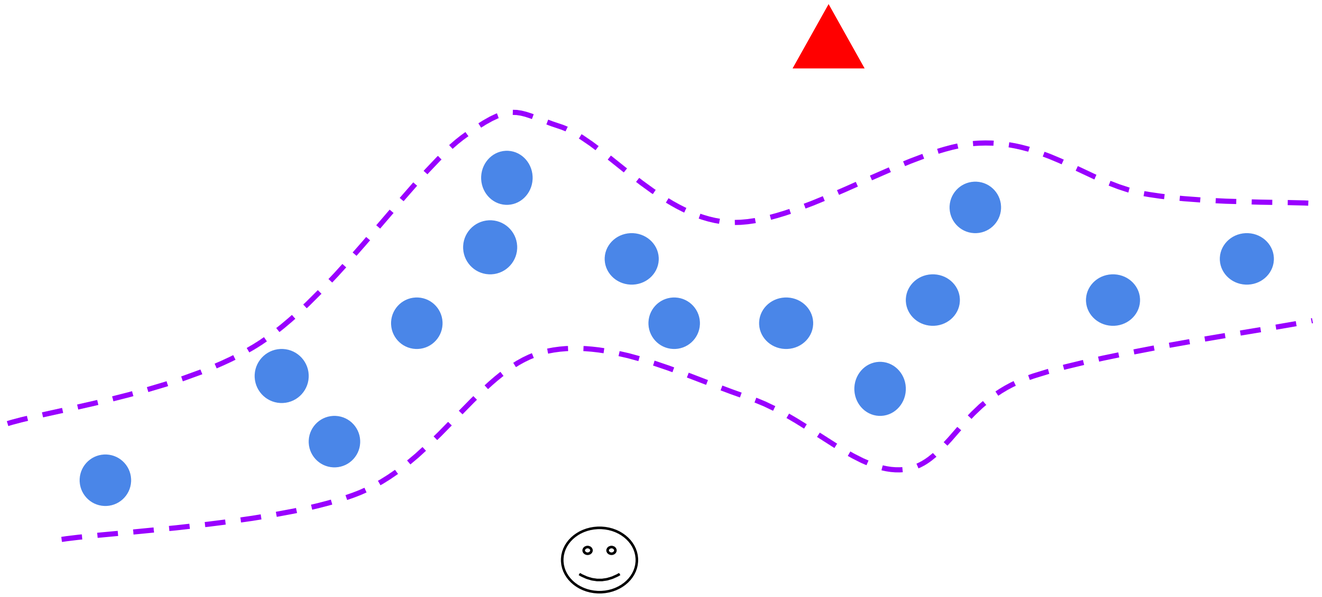
ource: Google search "social media influencers 2018"

	rank	yr	age	rsci	apex	q25	q50	q75	q90	apex
Wendell Carter Jr./2018/duke	1	FR	19.1	7	9.8					
Jaren Jackson Jr./2018/michigan	2	FR	18.6	9	9.4					
Trae Young/2018/oklahoma	3	FR	19.6	21	9.2					
Luka Doncic/2018/real-mad	4		19.2		8.8					
DeAndre Ayton/2018/arizona	5	FR	19.8	3	8.6					
Shai Gilgeous-Alexander/2018/kentucky	6	FR	19.8	30	8.2					
Troy Brown/2018/oregon	7	FR	18.8	12	7.8					
Dzanan Musa/2018/cedevita	8		19.0		7.7					
Marvin Bagley III/2018/duke	9	FR	19.1	1	7.5					
Miles Bridges/2018/michigan	10	SO	20.1	10	7.5					
Robert Williams/2018/texas-a&	11	SO	20.5	51	7.4					
De'Anthony Melton/2017/usc	12	FR	18.9		7.3					
Mohamed Bamba/2018/texas	13	FR	20.0	4	7.2					
Rodions Kurucs/2017/fc-barce	14		19.2		7.1					

source: http://www.tothemean.com/2018/06/10/nba-draft-2018.html

7. Outlier Detection

- Find a set of objects that are considerably dissimilar from the remainder of the data.
- Usually coupled with trend analysis



7. Outlier Detection

Techniques: signal processing, time-series analysis, and also clustering.

Applications: fraud detection, network security, mining software bugs, event detection, and clinical diagnosis.

Open Questions in Data Mining

- Are there “basic” functionalities of data mining that most/all data mining tasks care about?

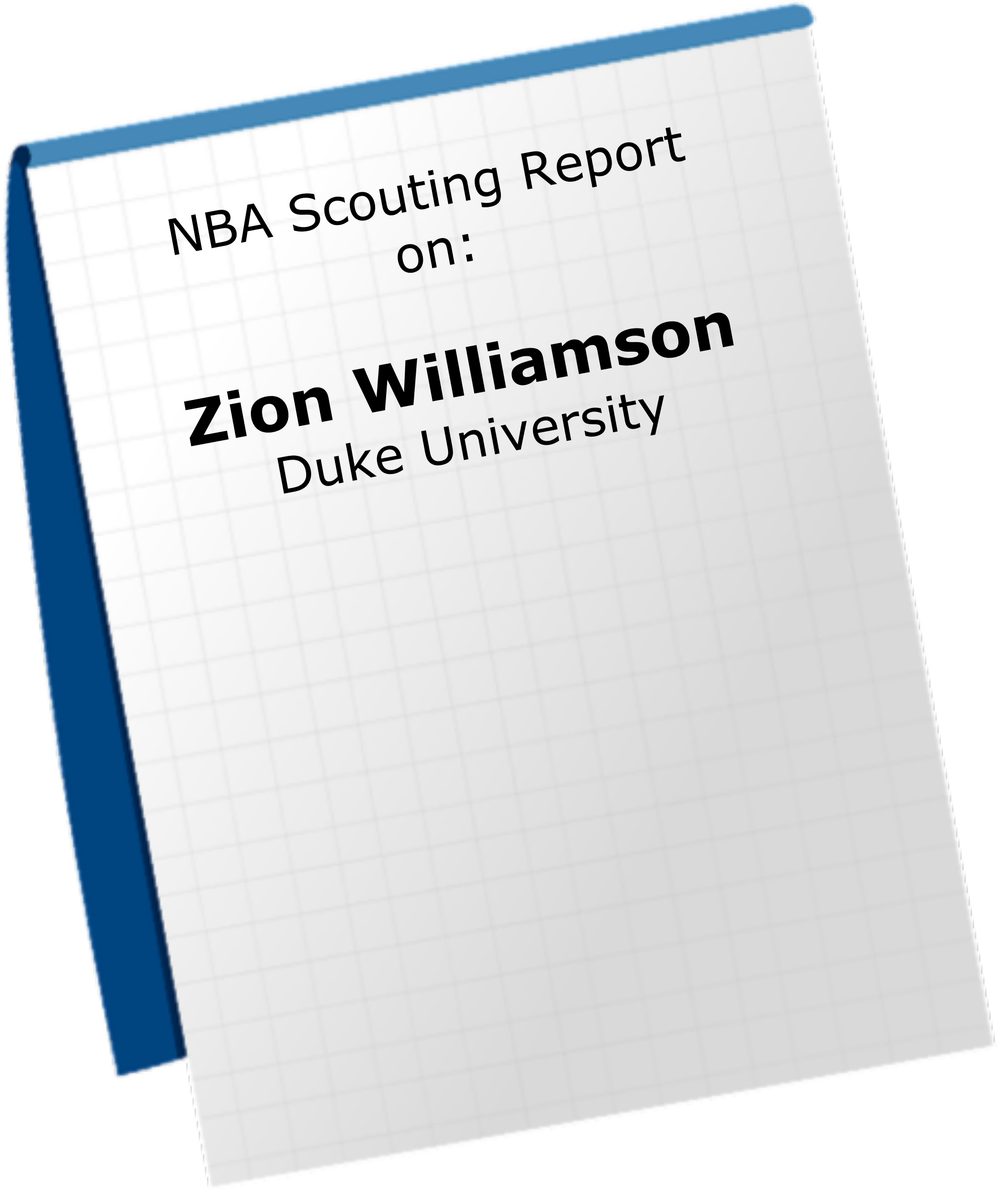
- Are there general ways to implement complex functionalities of data mining through basic ones?

Two ways of decision-making

- Make decisions (classification, prediction, clustering, ranking, ...) about an object
- Because it has a particular **pattern**
- Because it is **similar** to some objects

Example - NBA

Image by mohamed Hassan from Pixabay



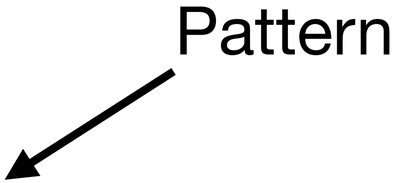
Example - NBA

"Zion is all about dominance, power and athleticism . . .

Williamson often gets compared to LeBron James, but he's more like Blake Griffin."

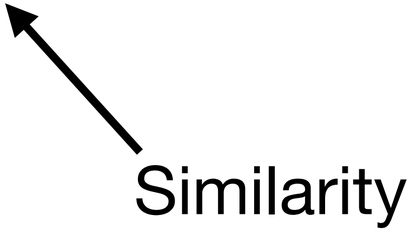
http://theleadsportsmedia.com/2019-nba-draft-top-5-prospect-pro-comparisons/

Example - NBA



"Zion is all about
dominance, power and athleticism . . .

Williamson often gets compared to
***LeBron James**, but he's more like*
***Blake Griffin**."*



Example - Movie Reviews

"I like the movie because it stars John Cusack (one of my favorite actors) and because it is set in Chicago (where I was born and raised) . . ." Pattern

"I like the movie because it shows Cowboy culture as it really is: a bunch of losers that use contrived violence with horses and cattle to make a big show off . . ." Pattern

"I like the movie because it reminds me of the movie 'The Players Club' . . . " Similarity

Pattern and Similarity

Pattern and similarity are two basic outputs of data mining that:

- apply to almost all data representations;
- can be used to build almost all other functionalities (even though it may not be the most optimal).

What is a “pattern”?

- A structure of attributes that represents the intrinsic and important properties of data objects.

- Particular formulation depends on data representation.

Similar concepts to “Pattern”

- Property
- Characteristic
- Regularity
- Feature

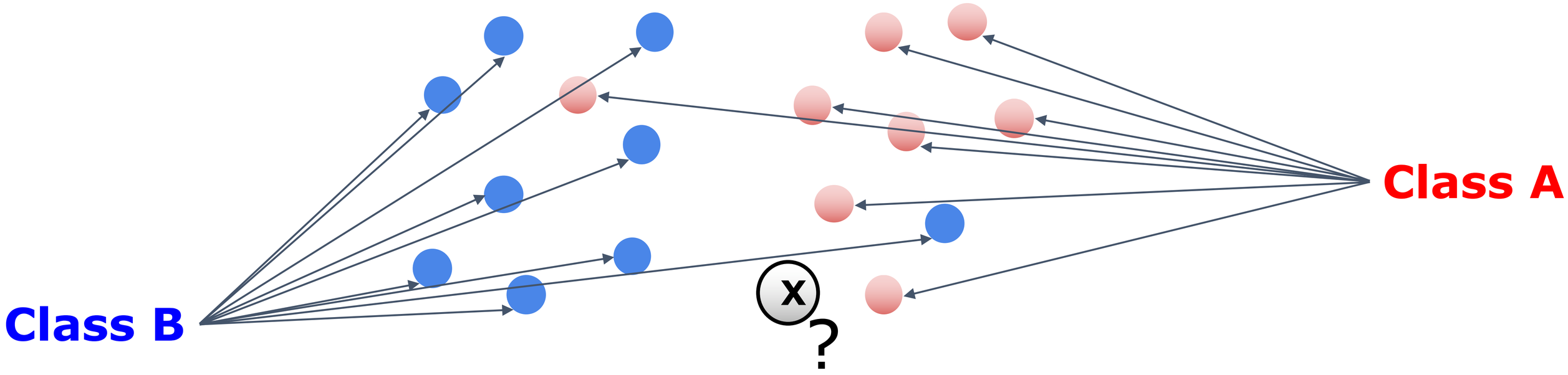
Use Patterns for Prediction

*"The **Cup and Handle** is a bullish continuation pattern . . . Once the handle is complete, the stock may breakout to new highs and resume its trend higher."*

- Investopedia (https://www.investopedia.com/university/technical/techanalysis8.asp)

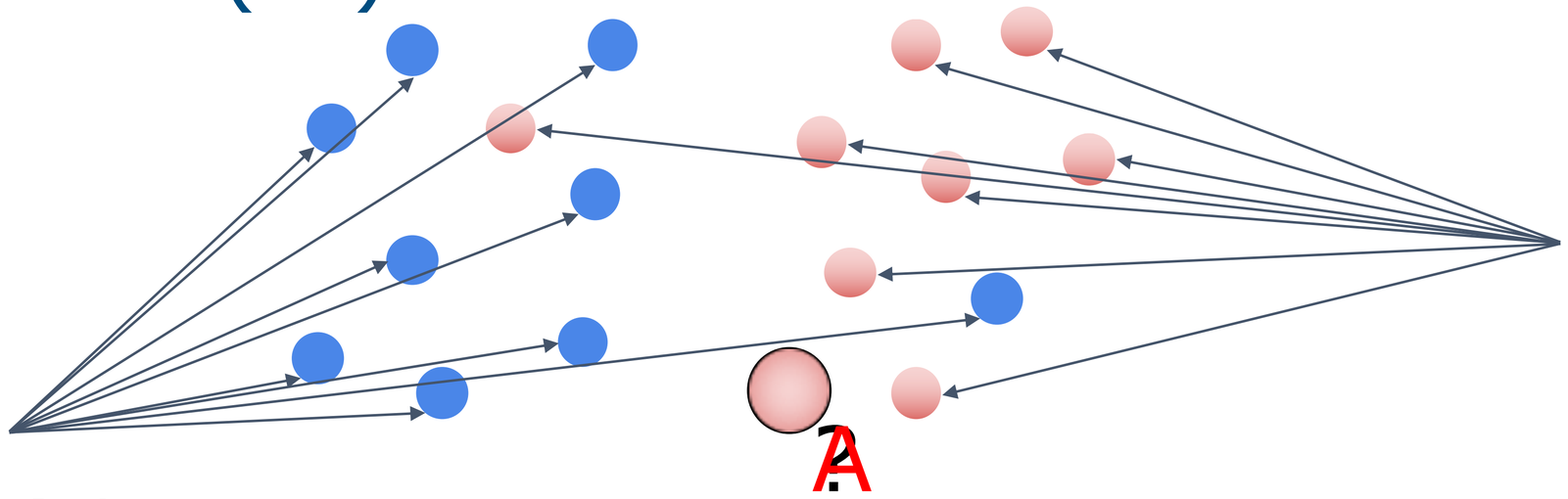
Use Patterns for Classification

Given labeled training examples, assign a new object to a class(es)



Use Patterns for Classification

Given labeled training examples, assign a new object to a class(es)



Class B

- 10% of class B have **pattern *P***

Class A

- 90% of class A have **pattern *P***

- X has **pattern *P***
- **X → A**

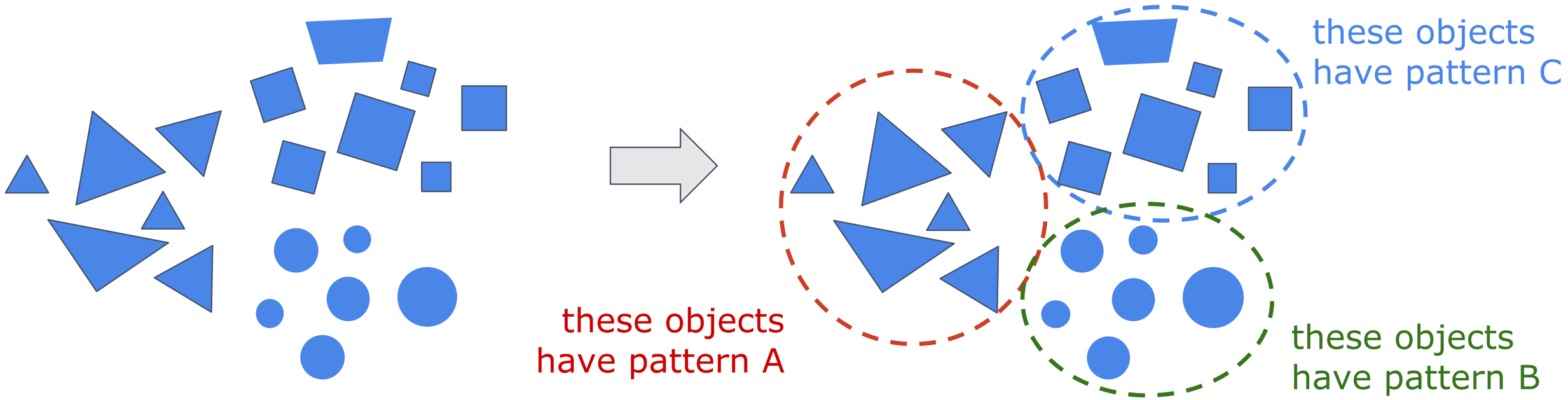
Use Patterns for Classification

- Facial features as a pattern.
- Multiple patterns to classify emotions.
- Features can be combined by a machine learning algorithm.

http://www.thefiscaltimes.com/2015/09/03/How-Tell-If-Your-Candidate-Lying

Use Patterns for Clustering

Group data objects into classes with no predefined classes or training examples

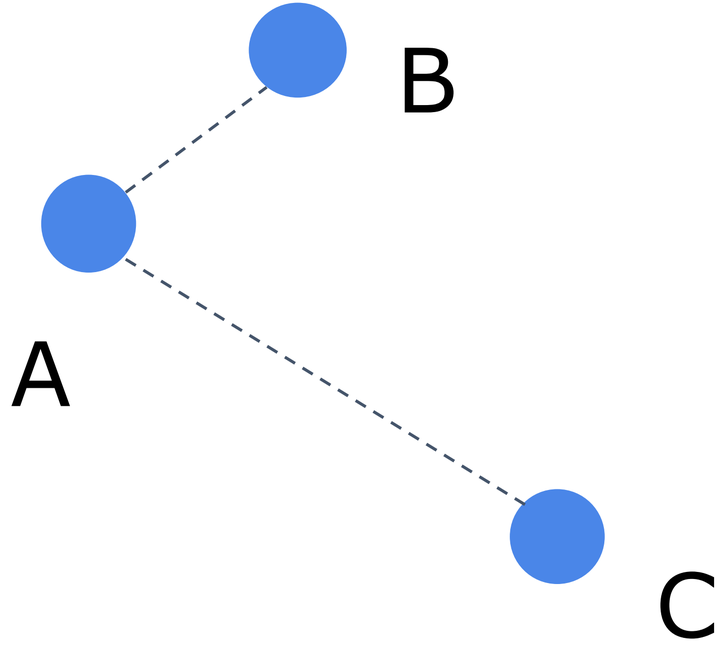


What is “similarity”?

- **Similarity** is a measure of how much two data objects are alike.

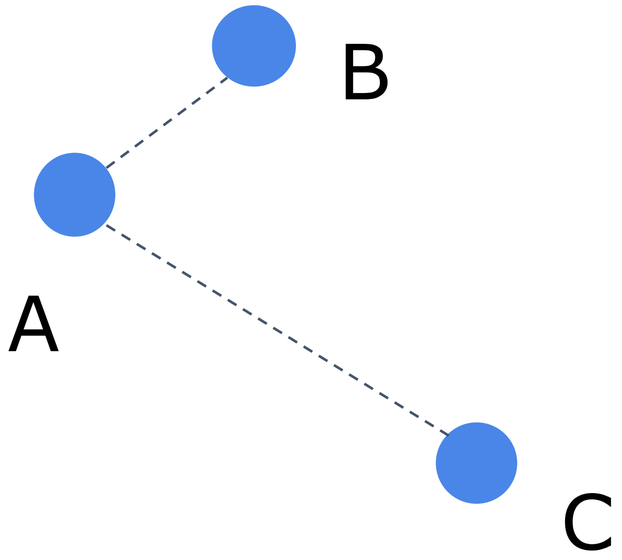
- **Distance** measures the opposite: how much two objects are dissimilar.

Example of Similarity



- A has a higher similarity to B than to C
- A is closer to B than to C
- A has a lower distance to B than to C

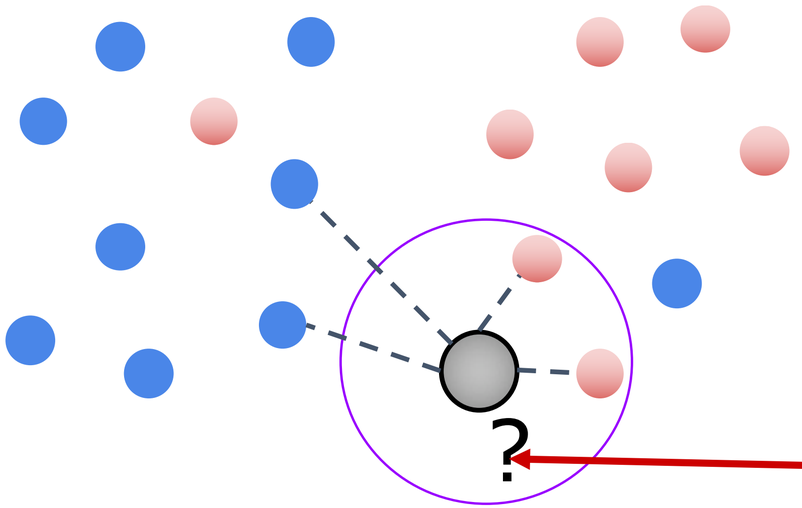
Example of Similarity



The measurement of similarity/distance depends on the data representation (itemsets or vectors).

Use Similarity for Classification

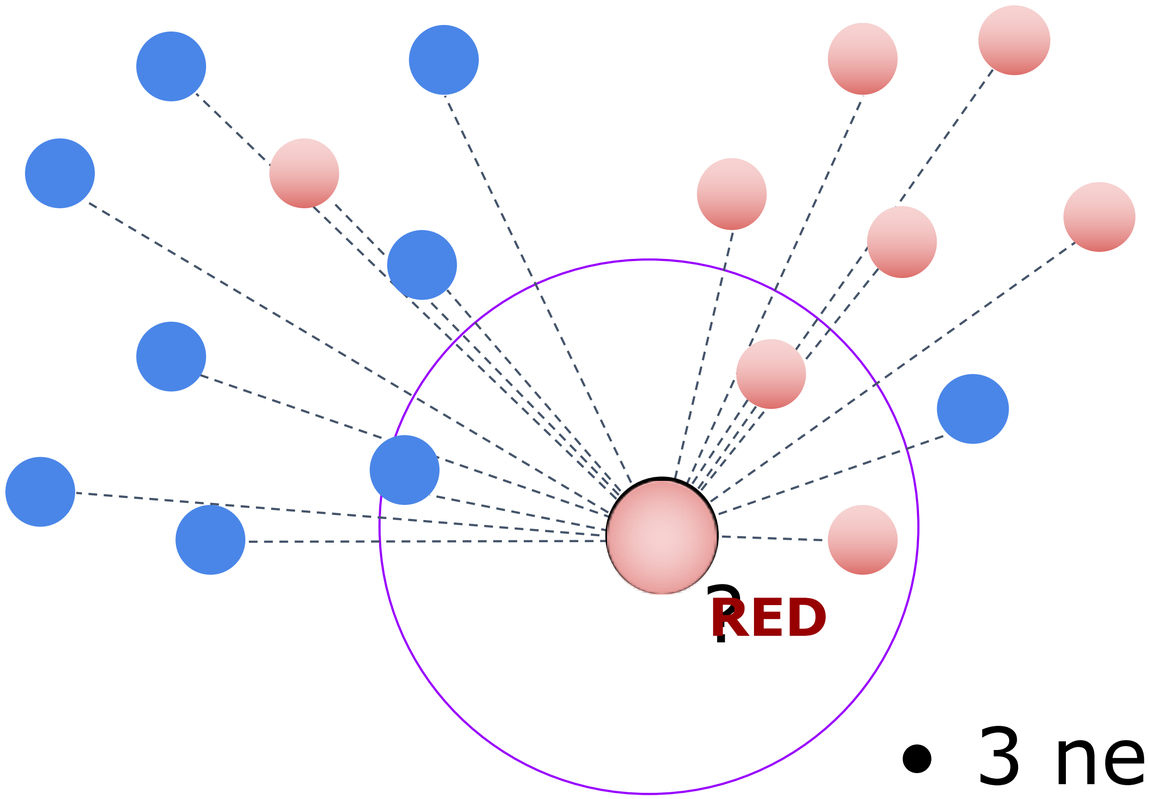
Given labeled training examples, assign a new object to a class(es).



Compute the similarity of new object to existing classified objects

New data object is closer to the red class; assign it to RED

K Nearest Neighbor Classifier



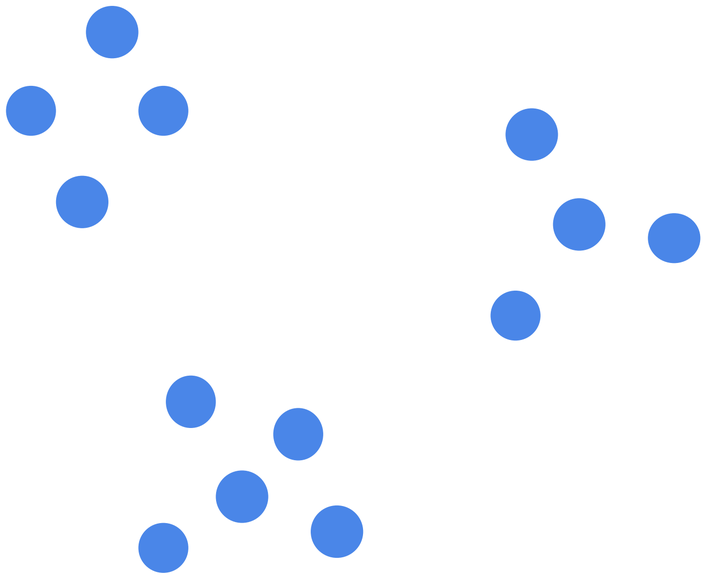
Starting with unlabeled object X:

- Calculate the similarities (or distances) between X to all existing objects
- Find K labeled objects that are nearest to X: KNN(X)
- Label(X) = majority of class labels in KNN(X)

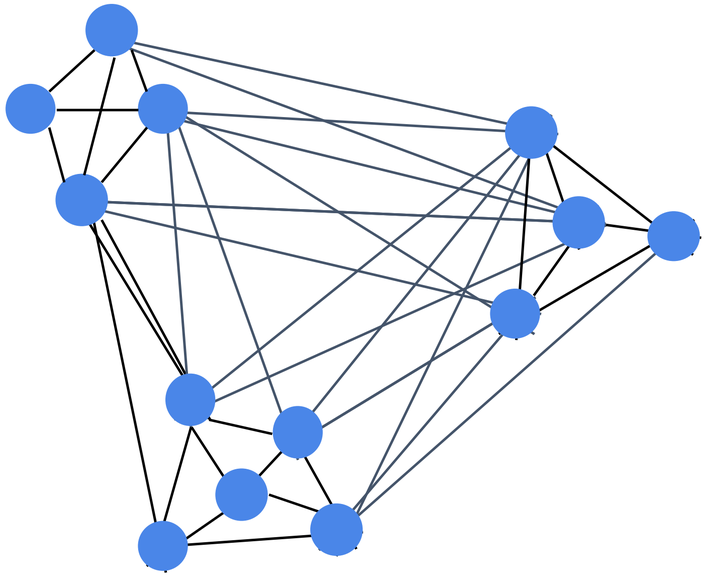
- 3 nearest neighbors: 2 RED, 1 BLUE
- Label of X = RED

Use Distance for Clustering

Group data objects into classes with no predefined classes or training examples

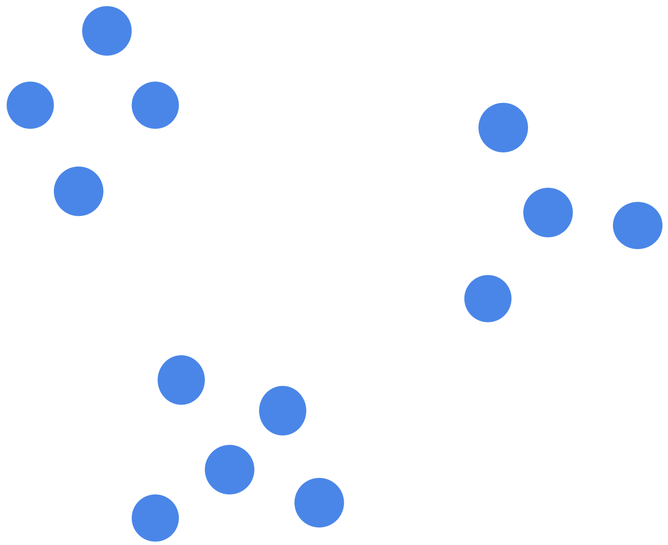


Compute distance between data objects



Use Distance (inverse of similarity) for Clustering

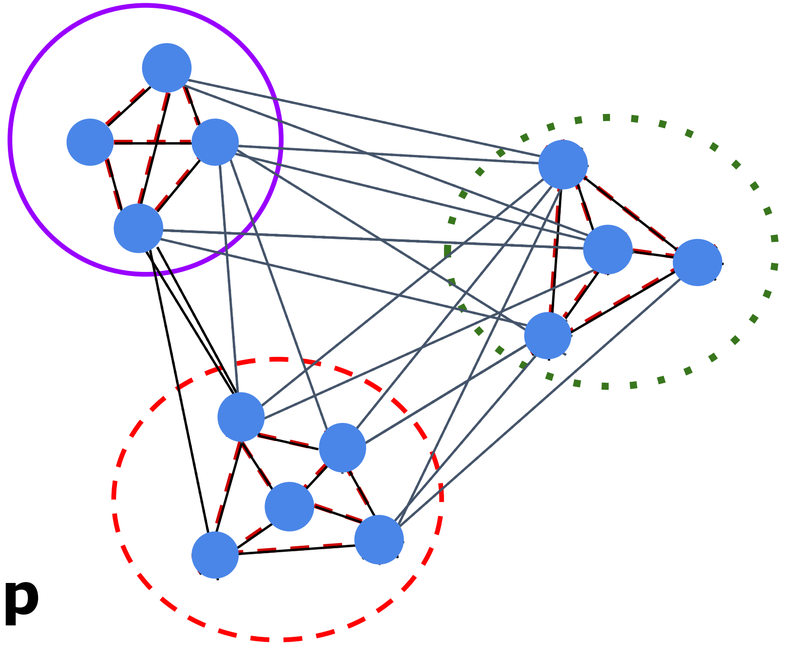
Group data objects into classes with no predefined classes or training examples.



Compute distance between data objects

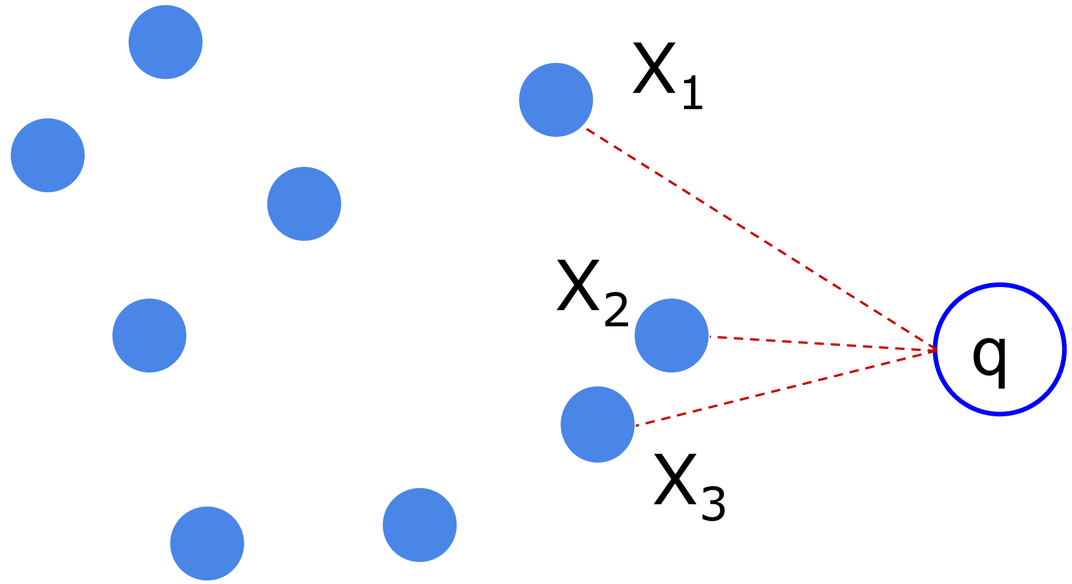
Minimize **in-group distances**;

Maximize **cross-group distances**



Use Similarity for Ranking

With query: objects closer (more similar) to the query should be ranked higher



Compute:

$sim(X_1, q)$;

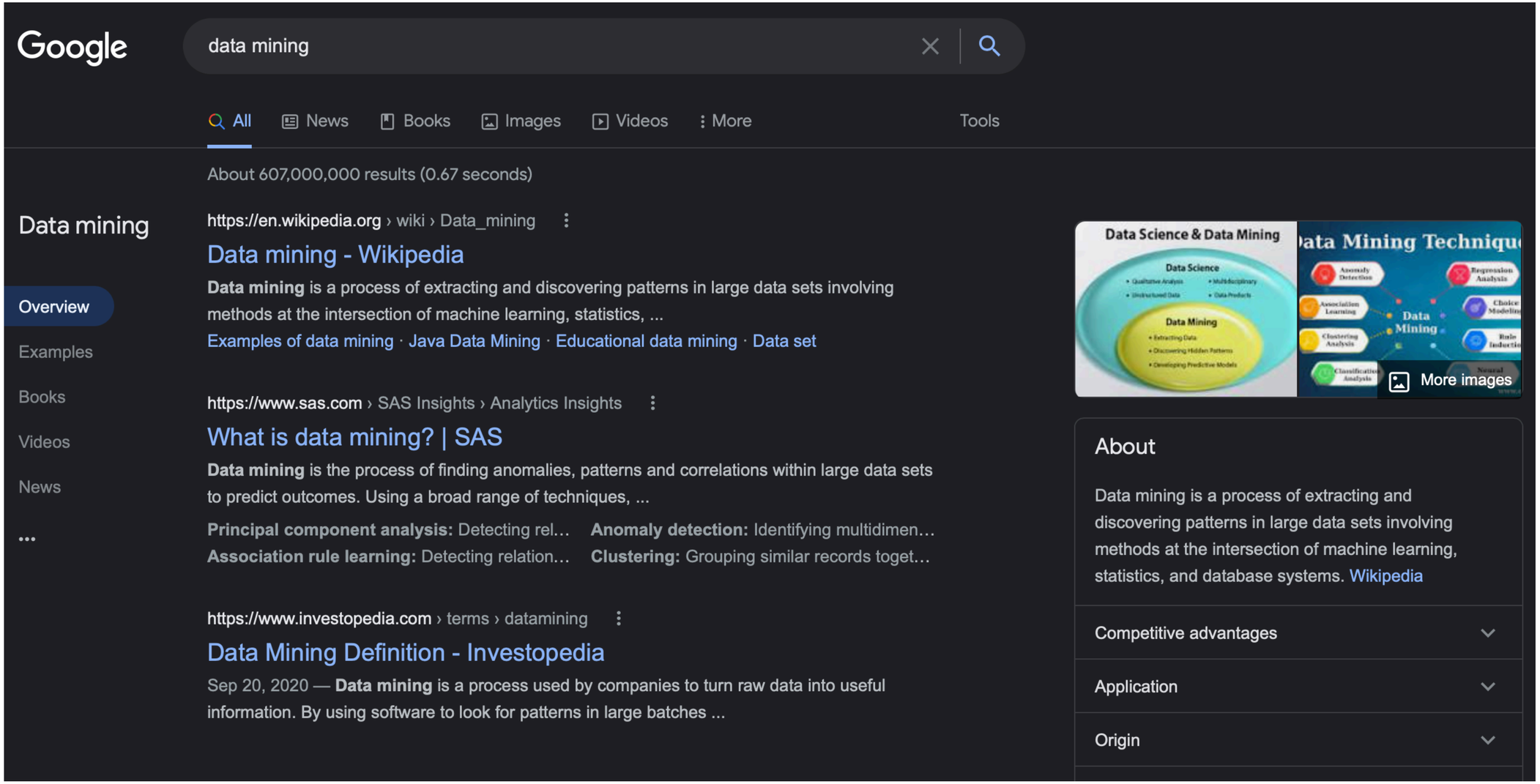
$sim(X_2, q)$;

$sim(X_3, q)$;

Then we rank:

$X_2 > X_3 > X_1$

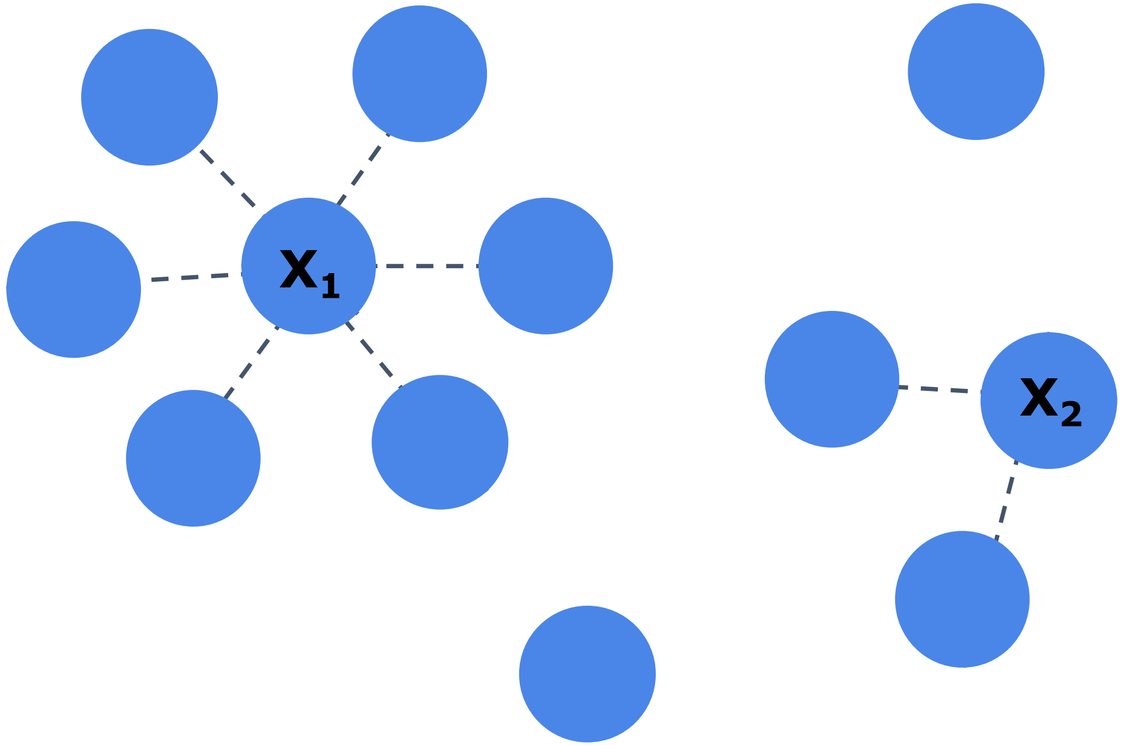
Application: Search Engine



Use Similarity for Ranking

Without query: an object X is important if it is close to many objects.

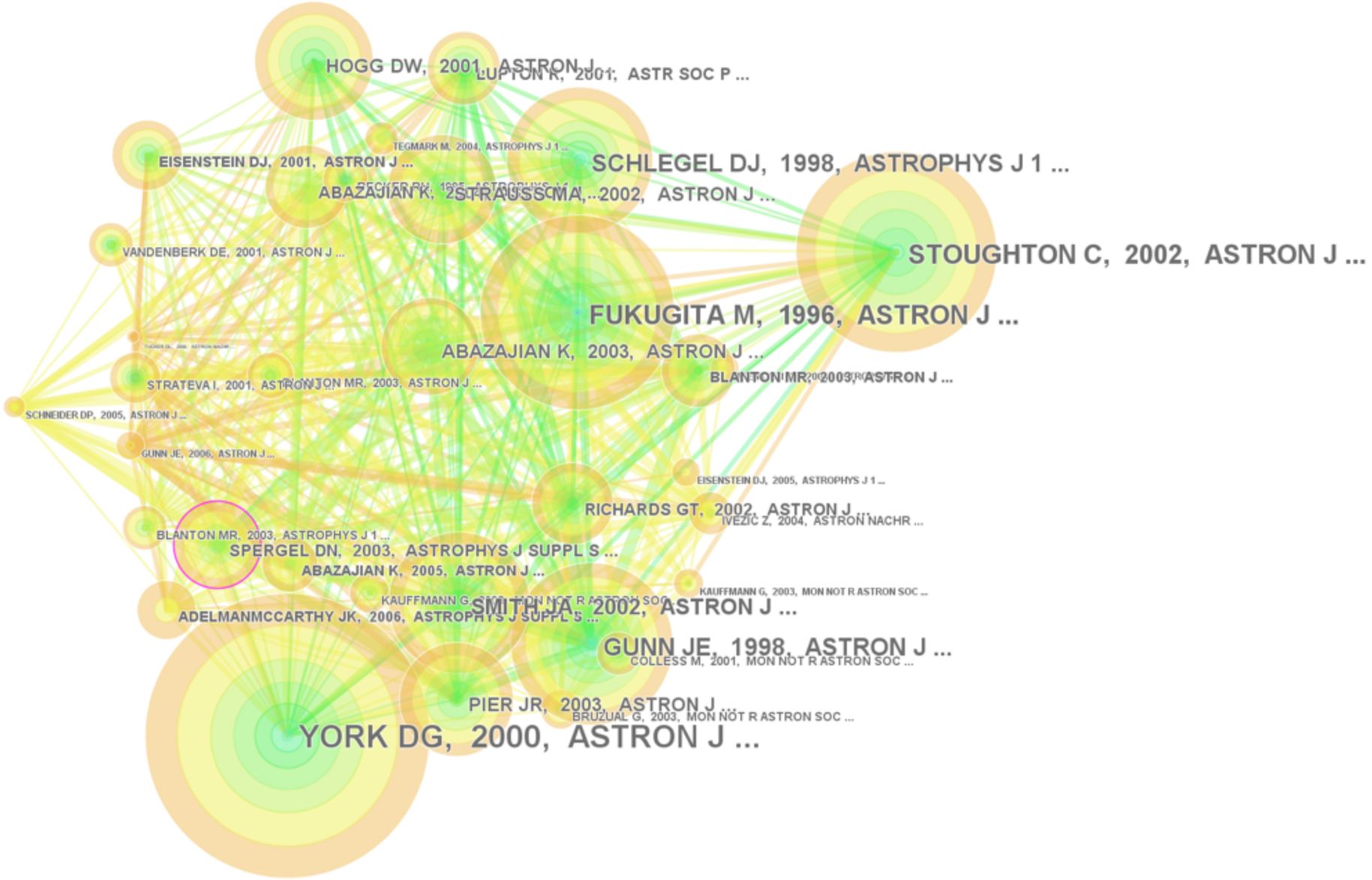
X_1 is close to 5 objects



X_1 is more important than X_2

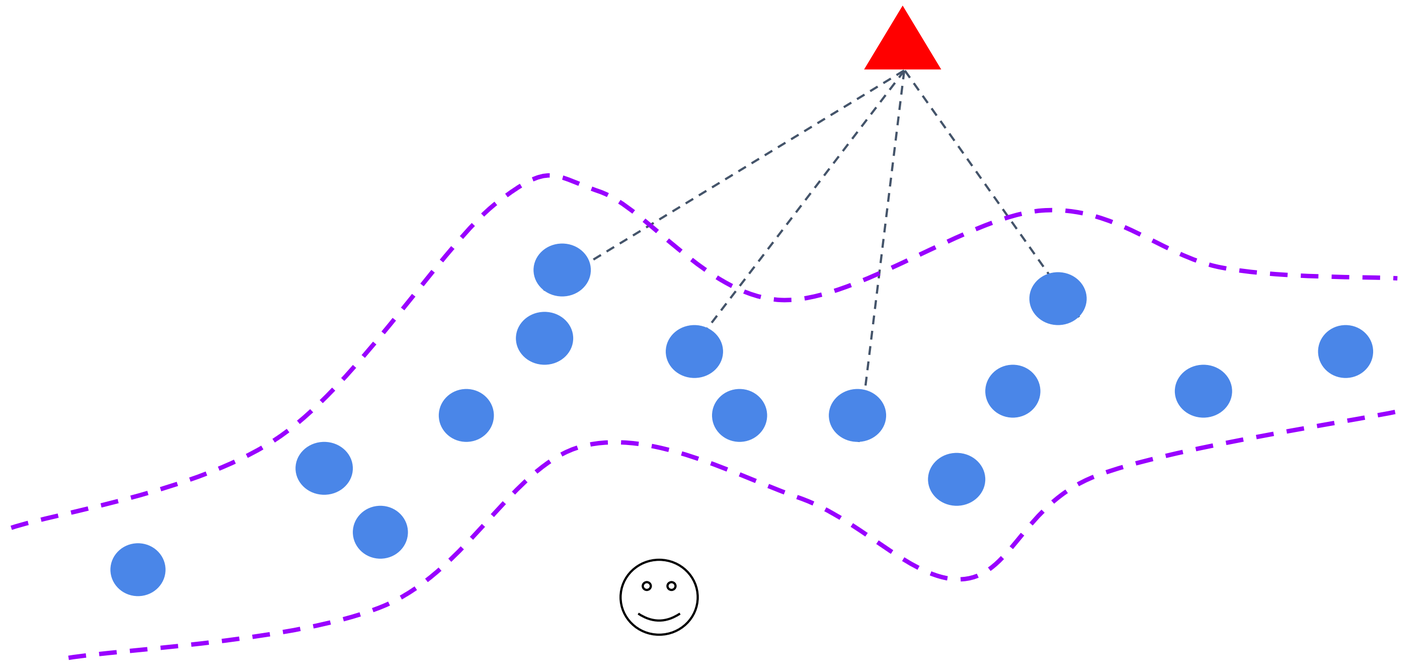
X_2 is close 2 objects

Application:
Co-citation Network of Sloan Digital Sky Survey



Use Distance for Outlier Detection

Find objects that are considerably dissimilar from the remainder of the data.



The red triangle is an outlier as its distance to all other data objects are above a threshold

What you should know

- Definition(s) of data mining.
- Relation to other concepts.
- Multiple views of data mining.
- Four dimensions of a data mining task.
- Basic functionalities of data mining.
- How functionalities differ from each other.

What you should know

- Data formulation is the first task of data mining.
- Different representations of data may be applied.
- How to represent data as item sets, matrix, time series, sequences, networks, and streams.
- Particular choice depends on the task and application.
- Patterns and similarity are two basic data mining outputs.
- Complex functionalities can be produced by patterns and similarities.

Thank You

Questions?