Lecture outline

- Clustering aggregation
  - Reference: A. Gionis, H. Mannila, P. Tsaparas: Clustering aggregation, ICDE 2004

- Co-clustering (or bi-clustering)
- References:
  - A. Anagnostopoulos, A. Dasgupta and R. Kumar: Approximation Algorithms for co-clustering, PODS 2008.
  - K. Puolamaki. S. Hanhijarvi and G. Garriga: An approximation ratio for biclustering, Information Processing Letters 2008.

Clustering aggregation

- Many different clusterings for the same dataset!
  - Different objective functions
  - Different algorithms
  - Different number of clusters
- Which clustering is the best?
  - Aggregation: we do not need to decide, but rather find a reconciliation between different outputs

The clustering-aggregation problem

- Input
  - **n** objects $\mathbf{X = \{x_1,x_2,\ldots,x_n\}}$
  - **m** clusterings of the objects $\mathbf{C_1,\ldots,C_m}$
    - partition: a collection of disjoint sets that cover **X**
- Output
  - a **single partition C**, that is as close as possible to all input partitions
- How do we measure ***closeness of clusterings***?
  - disagreement distance

Disagreement distance

- For object **x** and clustering **C, C(x)** is the index of set in the partition that contains **x**

- For two partitions **C** and **P**, and objects **x,y** in **X** define

$$I_{C,P}(x,\ y) = \begin{cases} 1 & \text{if } C(x) = C(y) \text{ and } P(x) \neq P(y) \\ & \text{OR} \\ & \text{if } C(x) \neq C(y) \text{ AND } P(x) = P(y) \\ 0 & \text{otherwise} \end{cases}$$

U	***C***	***P***
x_1	**1**	**1**
x_2	**1**	**2**
x_3	**2**	**1**
x_4	**3**	**3**
x_5	**3**	**4**

- if $I_{P,Q}(x,y) = 1$ we say that **x,y** create a disagreement between partitions **P** and **Q**

- $$D(P,\ Q) = \sum_{(x,\ y)} I_{P,Q}(x,\ y)$$

Metric property for disagreement distance

- For clustering **C**: **$D(C,C) = 0$**
- **$D(C,C')\geq 0$** for every pair of clusterings **C, C'**
- **$D(C,C') = D(C',C)$**
- Triangle inequality?
- It is sufficient to show that for each pair of points **x,y ϵX**: **$I_{x,y}(C_1,C_3)\leq I_{x,y}(C_1,C_2) + I_{x,y}(C_2,C_3)$**
- **$I_{x,y}$** takes values 0/1; triangle inequality can only be violated when
  - **$I_{x,y}(C_1,C_3)=1$** and **$I_{x,y}(C_1,C_2) = 0$** and **$I_{x,y}(C_2,C_3)=0$**
  - Is this possible?

Clustering aggregation

- Given partitions $C_1,\ldots,C_m$ find C such that

$$D(C) = \sum_{i=1}^{m} D(C, C_i)$$

the aggregation cost

is minimized

U	C_1	C_2	C_3	C
x_1	1	1	1	1
x_2	1	2	2	2
x_3	2	1	1	1
x_4	2	2	2	2
x_5	3	3	3	3
x_6	3	4	3	3

Why clustering aggregation?

- Clustering categorical data

U	*City*	*Profession*	*Nationality*
x_1	New York	Doctor	U.S.
x_2	New York	Teacher	Canada
x_3	Boston	Doctor	U.S.
x_4	Boston	Teacher	Canada
x_5	Los Angeles	Lawer	Mexican
x_6	Los Angeles	Actor	Mexican

- The two problems are equivalent

Why clustering aggregation?

- Identify the correct number of clusters
  - the optimization function does not require an explicit number of clusters

- Detect outliers
  - outliers are defined as points for which there is no consensus

Why clustering aggregation?

- Improve the robustness of clustering algorithms
  - different algorithms have different weaknesses.
  - combining them can produce a better result.

Why clustering aggregation?

- Privacy preserving clustering
  - different companies have data for the same users. They can compute an aggregate clustering without sharing the actual data.

Complexity of Clustering Aggregation

- The clustering aggregation problem is NP-hard
  - the median partition problem [Barthelemy and LeClerc 1995].

- Look for heuristics and approximate solutions.

A simple 2-approximation algorithm

- The disagreement distance **D(C,P)** is a metric

- The algorithm **BEST**: Select among the input clusterings the clustering **C^*** that minimizes **$D(C^*)$.**
  - a 2-approximate solution. Why?

A 3-approximation algorithm

- The **BALLS** algorithm:
  - Select a point **x** and look at the set of points **B** within distance ½ of **x**
  - If the average distance of **x** to **B** is less than ¼ then create the cluster **BU{p}**
  - Otherwise, create a singleton cluster **{p}**
  - Repeat until all points are exhausted

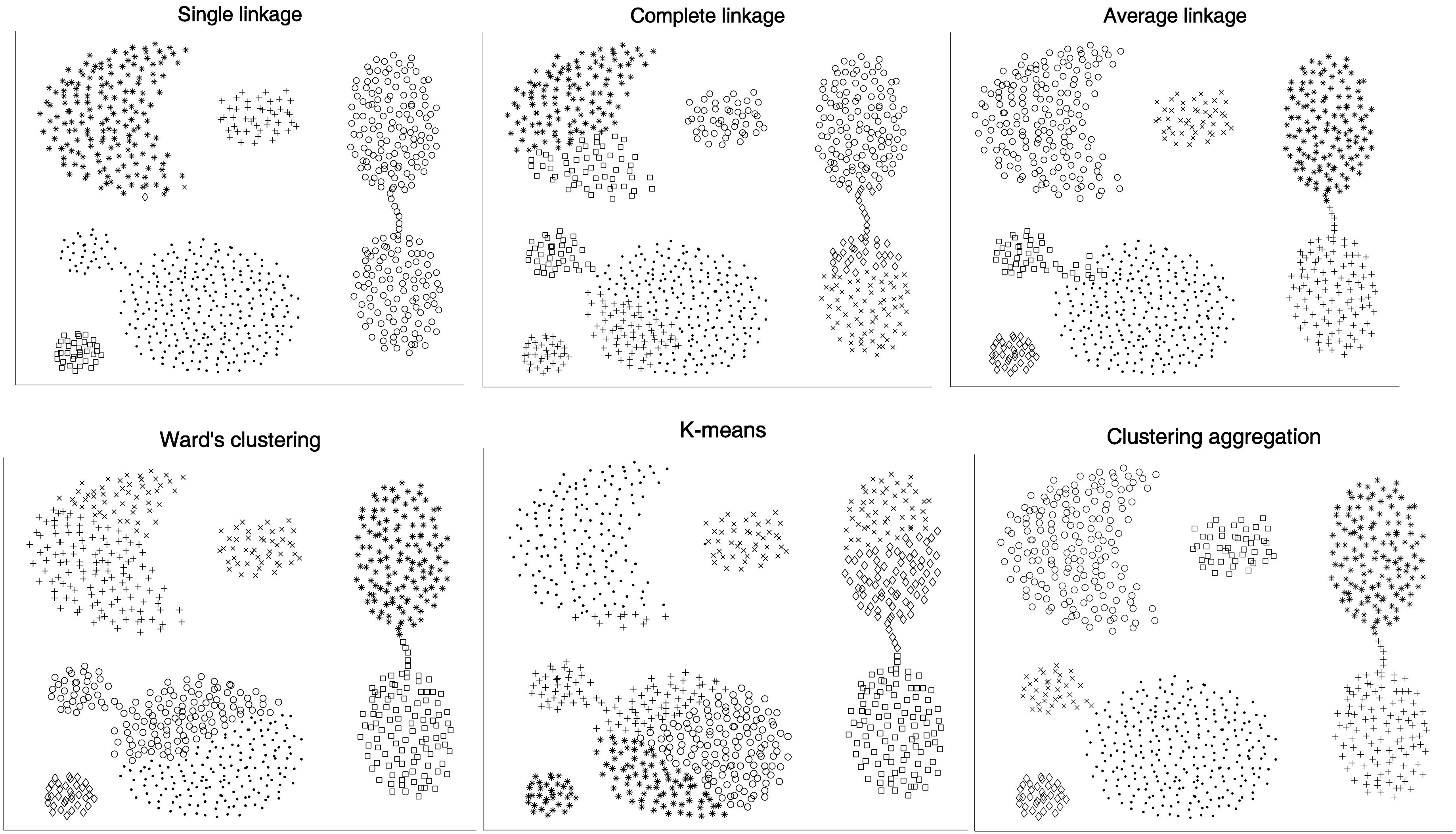
- Theorem: The **BALLS** algorithm has worst-case approximation factor **3**

Other algorithms

- **AGGLO**:
  - Start with all points in singleton clusters
  - Merge the two clusters with the smallest average inter-cluster edge weight
  - Repeat until the average weight is more than ½

- **LOCAL**:
  - Start with a random partition of the points
  - Remove a point from a cluster and try to merge it to another cluster, or create a singleton to improve the cost of aggregation.
  - Repeat until no further improvements are possible

Clustering Robustness



Single linkage

Complete linkage

Average linkage

Ward's clustering

K-means

Clustering aggregation

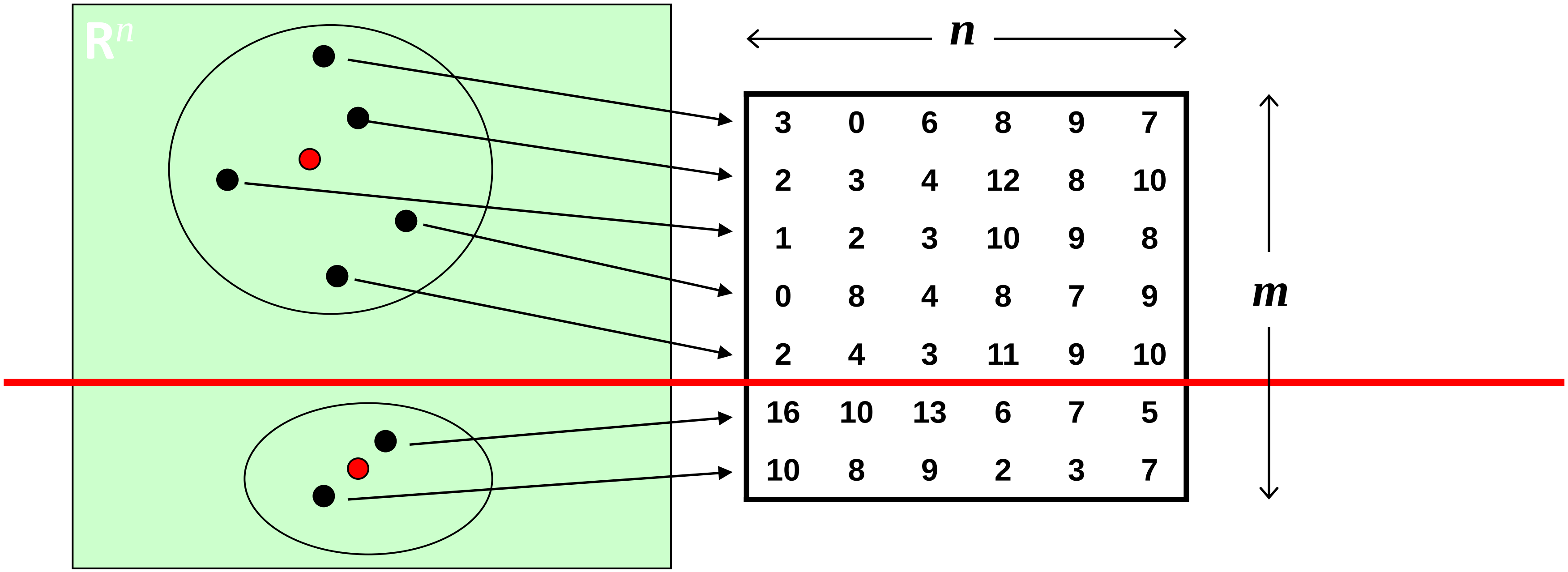
Lecture outline

- Clustering aggregation
  - Reference: A. Gionis, H. Mannila, P. Tsaparas: Clustering aggregation, ICDE 2004

- Co-clustering (or bi-clustering)
- References:
  - A. Anagnostopoulos, A. Dasgupta and R. Kumar: Approximation Algorithms for co-clustering, PODS 2008.
  - K. Puolamaki. S. Hanhijarvi and G. Garriga: An approximation ratio for biclustering, Information Processing Letters 2008.

Clustering

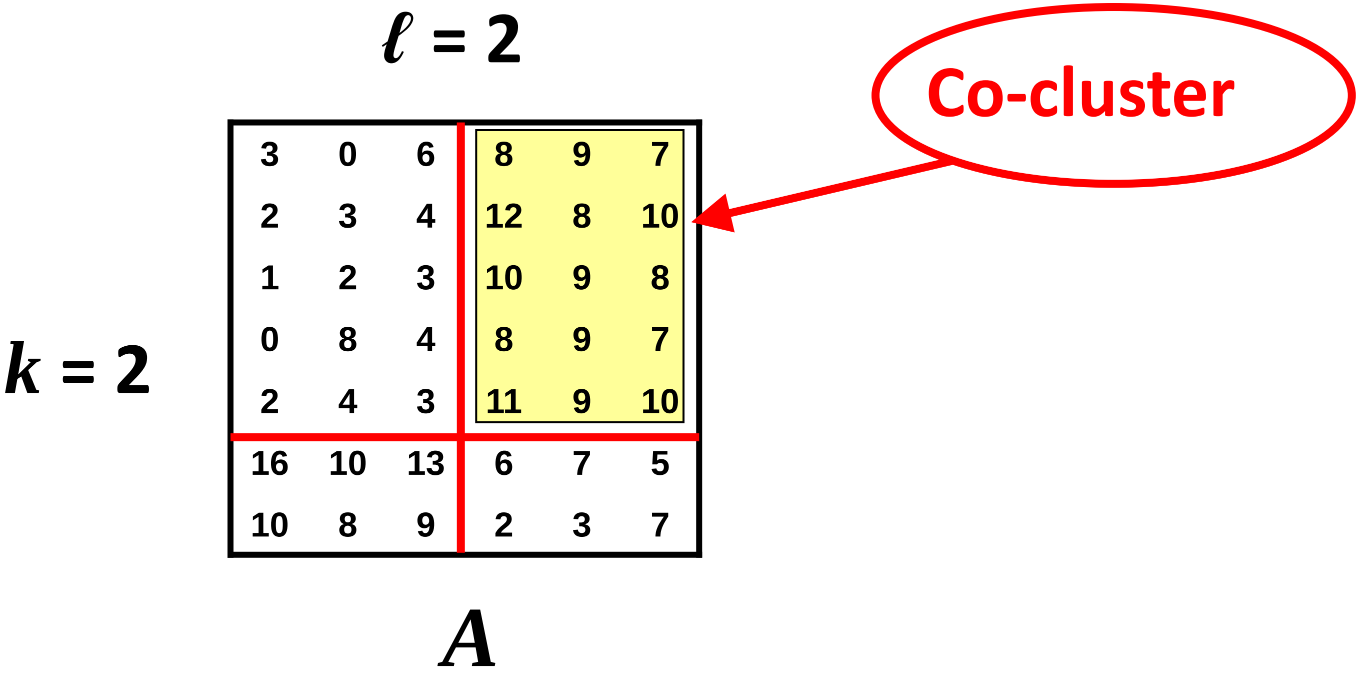
- m points in $\mathbf{R}^n$
- Group them to k clusters
- Represent them by a matrix $A \in \mathbf{R}^{m \times n}$
  - – A point corresponds to a row of A
- **Cluster:** Partition the rows to k



3	0	6	8	9	7
2	3	4	12	8	10
1	2	3	10	9	8
0	8	4	8	7	9
2	4	3	11	9	10
16	10	13	6	7	5
10	8	9	2	3	7

Co-Clustering

- **Co-Clustering: Cluster rows and columns of A simultaneously:**



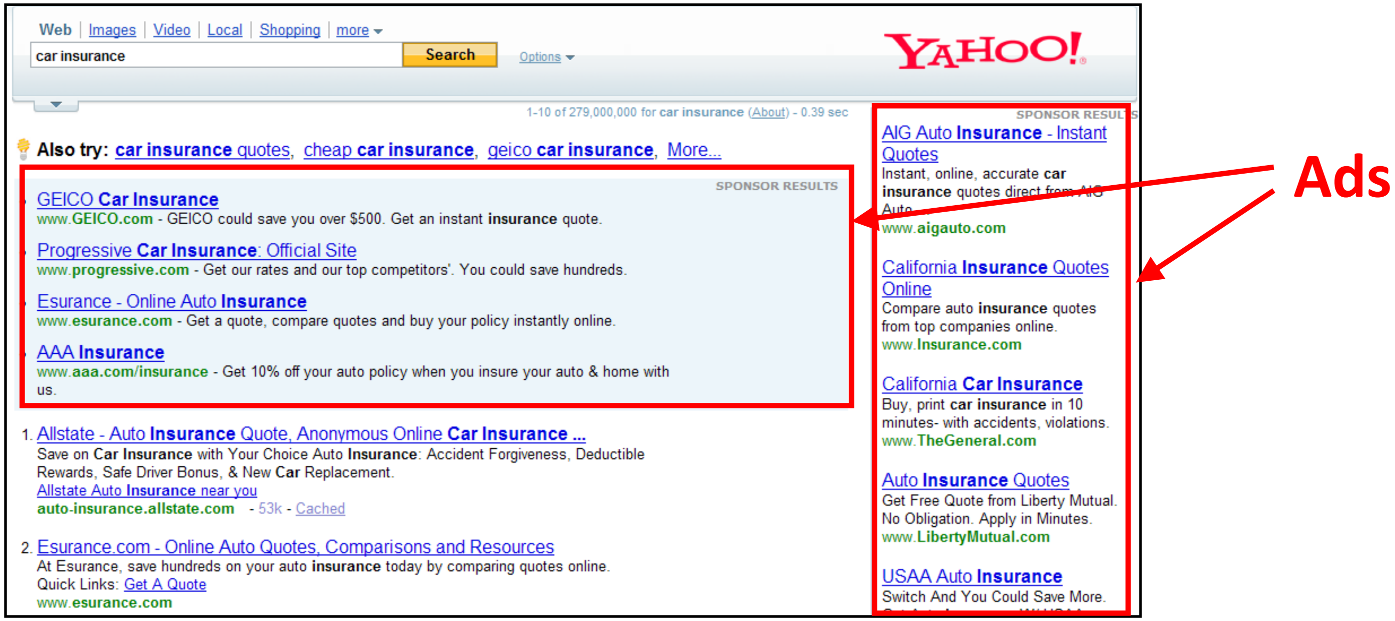
$\ell = 2$

3	0	6	8	9	7
2	3	4	12	8	10
1	2	3	10	9	8
0	8	4	8	9	7
2	4	3	11	9	10
16	10	13	6	7	5
10	8	9	2	3	7

$k = 2$

A

Motivation: Sponsored Search



1-10 of 279,000,000 for **car insurance** (About) - 0.39 sec

Also try: **car insurance** quotes, cheap **car insurance**, geico **car insurance**, More...

SPONSOR RESULTS

GEICO **Car Insurance**
www.GEICO.com - GEICO could save you over $500. Get an instant **insurance** quote.

Progressive **Car Insurance**: Official Site
www.progressive.com - Get our rates and our top competitors'. You could save hundreds.

Esurance - Online Auto **Insurance**
www.esurance.com - Get a quote, compare quotes and buy your policy instantly online.

AAA **Insurance**
www.aaa.com/insurance - Get 10% off your auto policy when you insure your auto & home with us.

1. Allstate - Auto **Insurance** Quote, Anonymous Online **Car Insurance** ...
Save on **Car Insurance** with Your Choice Auto **Insurance**: Accident Forgiveness, Deductible Rewards, Safe Driver Bonus, & New **Car** Replacement.
Allstate Auto Insurance near you
auto-insurance.allstate.com - 53k - Cached

2. Esurance.com - Online Auto Quotes, Comparisons and Resources
At Esurance, save hundreds on your auto **insurance** today by comparing quotes online.
Quick Links: Get A Quote
www.esurance.com

SPONSOR RESULTS

AIG Auto **Insurance** - Instant Quotes
Instant, online, accurate **car insurance** quotes direct from AIG Auto.
www.aigauto.com

California **Insurance** Quotes Online
Compare auto **insurance** quotes from top companies online.
www.Insurance.com

California **Car Insurance**
Buy, print **car insurance** in 10 minutes- with accidents, violations.
www.TheGeneral.com

Auto **Insurance** Quotes
Get Free Quote from Liberty Mutual. No Obligation. Apply in Minutes.
www.LibertyMutual.com

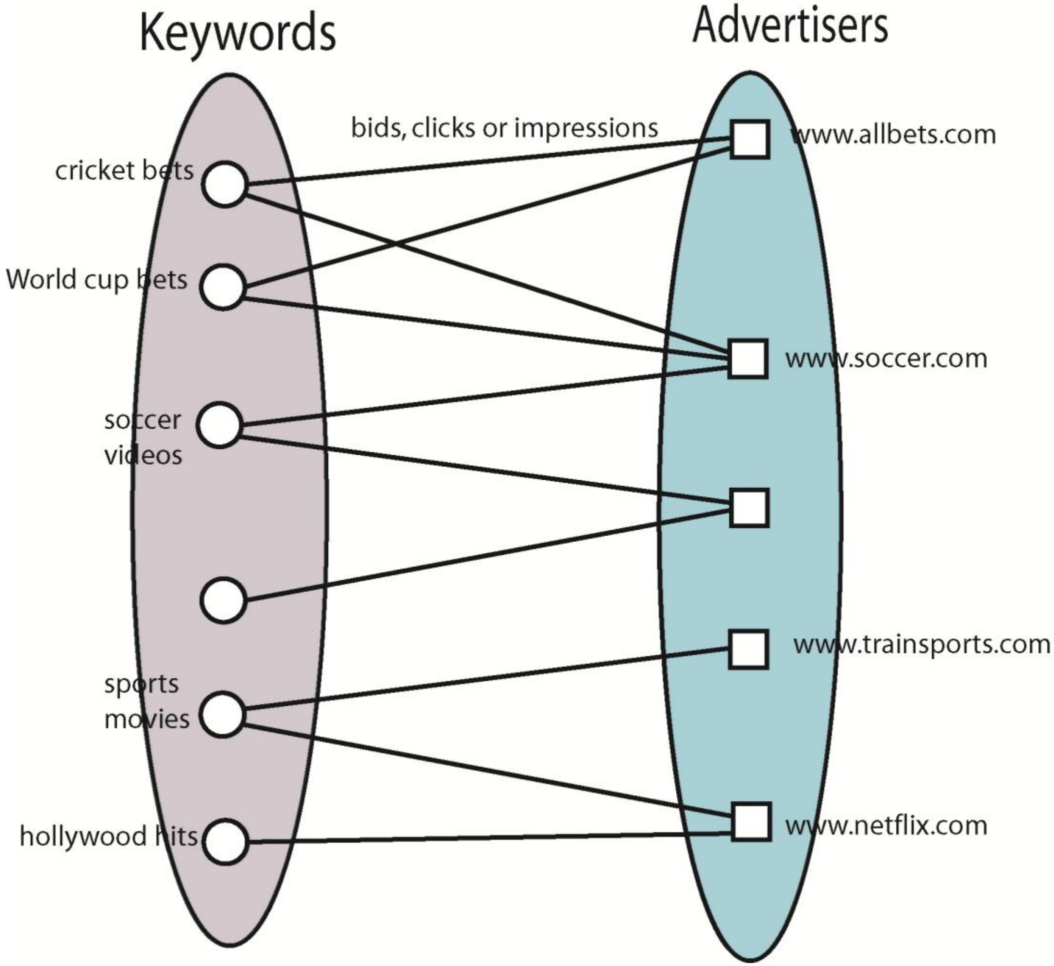
USAA Auto **Insurance**
Switch And You Could Save More.

Ads

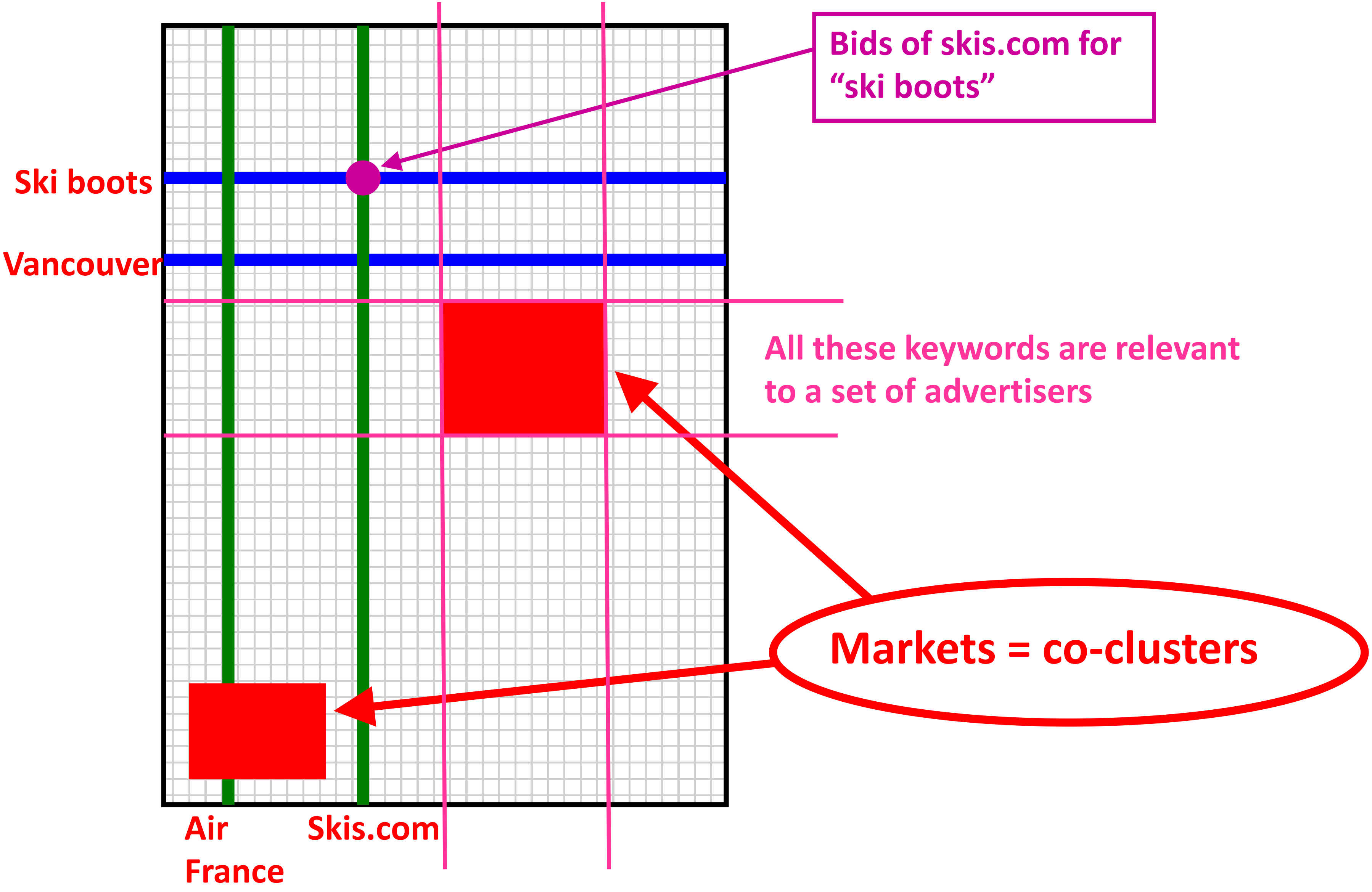
- Advertisers bid on keywords
- A user makes a query
- Show ads of advertisers that are relevant and have high bids
- User clicks or not an ad

Motivation: Sponsored Search

- For every (*advertiser, keyword*) pair we have:
    - Bid amount
    - Impressions
    - # clicks
- Mine information at query time
    - Maximize # clicks / revenue



Co-Clusters in Sponsored Search



Ski boots

Vancouver

Air France

Skis.com

Bids of skis.com for "ski boots"

All these keywords are relevant to a set of advertisers

Markets = co-clusters

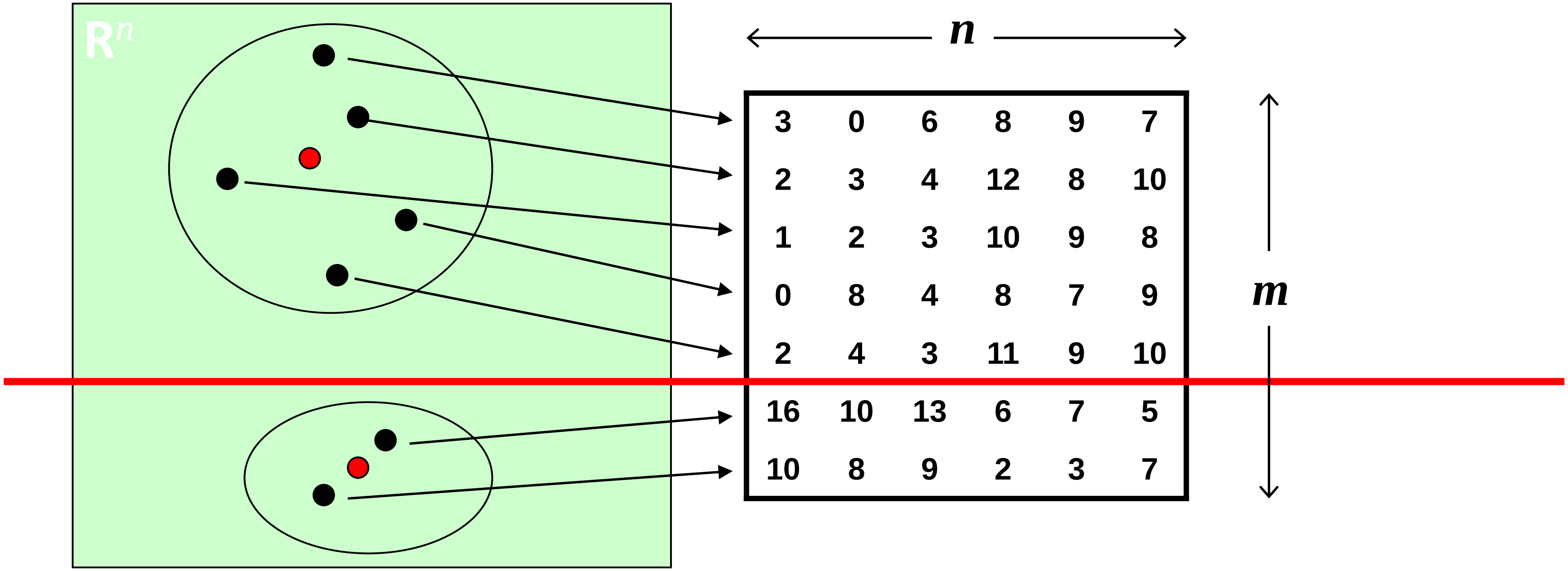
Co-Clustering in Sponsored Search

Applications:

- Keyword suggestion
  - Recommend to advertisers other relevant keywords
- Broad matching / market expansion
  - Include more advertisers to a query
- Isolate submarkets
  - Important for economists
  - Apply different advertising approaches
- Build taxonomies of advertisers / keywords

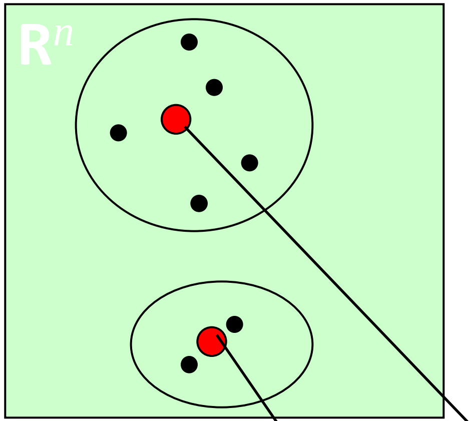
Clustering of the rows

- m points in $\mathbf{R}^n$
- Group them to k clusters
- Represent them by a matrix $A \in \mathbf{R}^{m \times n}$
  - A point corresponds to a row of A
- **Clustering:** Partitioning of the rows into k groups



3	0	6	8	9	7
2	3	4	12	8	10
1	2	3	10	9	8
0	8	4	8	7	9
2	4	3	11	9	10
16	10	13	6	7	5
10	8	9	2	3	7

Clustering of the columns



- n points in $\mathbf{R}^m$
- Group them to k clusters
- Represent them by a matrix $A \in \mathbf{R}^{m\times n}$
  - A point corresponds to a column of A
- **Clustering:** Partitioning of the columns into k groups

3	0	6	8	9	7
2	3	4	12	8	10
1	2	3	10	9	8
0	8	4	8	7	9
2	4	3	11	9	10
16	10	13	6	7	5
10	8	9	2	3	7

$$\begin{array}{c|cccccc} & \multicolumn{6}{c}{n} \\ \hline & 3 & 3 & 3 & 9 & 9 & 9 \\ & 3 & 3 & 3 & 9 & 9 & 9 \\ & 3 & 3 & 3 & 9 & 9 & 9 \\ m & 3 & 3 & 3 & 9 & 9 & 9 \\ & 3 & 3 & 3 & 9 & 9 & 9 \\ & 11 & 11 & 11 & 5 & 5 & 5 \\ & 11 & 11 & 11 & 5 & 5 & 5 \end{array}$$

Cost of clustering

3	0	6	8	9	7
2	3	4	12	8	10
1	2	3	10	9	8
0	8	4	8	7	9
2	4	3	11	9	10
16	10	13	6	7	5
10	8	9	2	3	7

Original data points **A**

1.6	3.4	4	9.8	8.4	8.8
1.6	3.4	4	9.8	8.4	8.8
1.6	3.4	4	9.8	8.4	8.8
1.6	3.4	4	9.8	8.4	8.8
1.6	3.4	4	9.8	8.4	8.8
13	9	11	4	5	6
13	9	11	4	5	6

Data representation **A’**

- In **A’** every point in **A** (row or column) is replaced by the corresponding representative (row or column)
- The quality of the clustering is measured by computing distances between the data in the cells of **A** and **A’**.

- **k-means clustering**: **cost =** $\sum_{i=1\ldots n} \sum_{j=1\ldots m} \left(A(i,j)-A'(i,j)\right)^2$
- **k-median clustering**: **cost =** $\sum_{i=1\ldots n} \sum_{j=1\ldots m} \left|A(i,j)-A'(i,j)\right|$

Co-Clustering

- **Co-Clustering:** Cluster rows and columns of $A \in \mathbf{R}^{m\times n}$ simultaneously
- ***k*** row clusters, $\boldsymbol{\ell}$ column clusters
- Every cell in **A** is represented by a cell in **A'**
- All cells in the same co-cluster are represented by the same value in the cells of **A'**

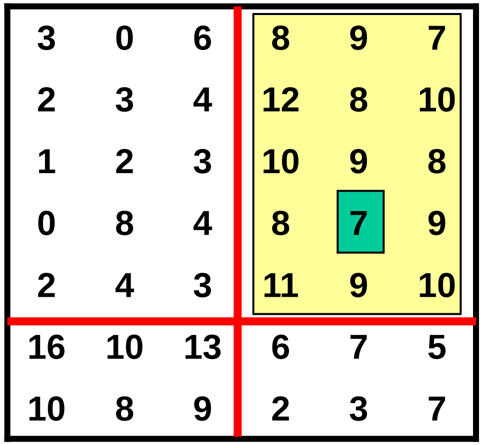
$$\left[\begin{array}{ccc|ccc}
3 & 0 & 6 & 8 & 9 & 7 \\
2 & 3 & 4 & 12 & 8 & 10 \\
1 & 2 & 3 & 10 & 9 & 8 \\
0 & 8 & 4 & 8 & 9 & 7 \\
2 & 4 & 3 & 11 & 9 & 10 \\
\hline
16 & 10 & 13 & 6 & 7 & 5 \\
10 & 8 & 9 & 2 & 3 & 7
\end{array}\right]$$

Original data **A**

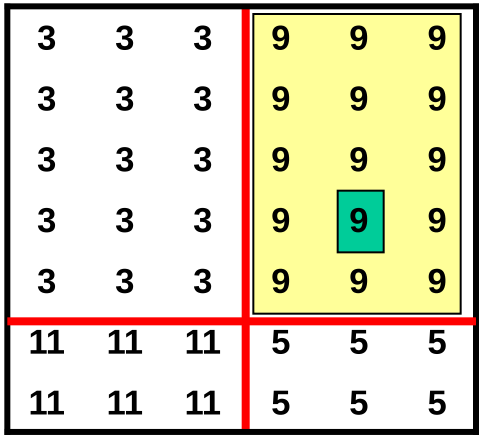
$$\left[\begin{array}{ccc|ccc}
3 & 3 & 3 & 9 & 9 & 9 \\
3 & 3 & 3 & 9 & 9 & 9 \\
3 & 3 & 3 & 9 & 9 & 9 \\
3 & 3 & 3 & 9 & 9 & 9 \\
3 & 3 & 3 & 9 & 9 & 9 \\
\hline
11 & 11 & 11 & 5 & 5 & 5 \\
11 & 11 & 11 & 5 & 5 & 5
\end{array}\right]$$

Co-cluster representation **A'**

Co-Clustering Objective Function



3	0	6	8	9	7
2	3	4	12	8	10
1	2	3	10	9	8
0	8	4	8	7	9
2	4	3	11	9	10
16	10	13	6	7	5
10	8	9	2	3	7



3	3	3	9	9	9
3	3	3	9	9	9
3	3	3	9	9	9
3	3	3	9	9	9
3	3	3	9	9	9
11	11	11	5	5	5
11	11	11	5	5	5

- In **A'** every point in **A** (row or column) is replaced by the corresponding representative (row or column)
- The quality of the clustering is measured by computing distances between the data in the cells of **A** and **A'**.

- **k-means Co-clustering**: **cost** $= \sum_{i=1\ldots n} \sum_{j=1\ldots m} (A(i,j)-A'(i,j))^2$
- **k-median Co-clustering**: **cost** $= \sum_{i=1\ldots n} \sum_{j=1\ldots m} |A(i,j)-A'(i,j)|$

Some Background

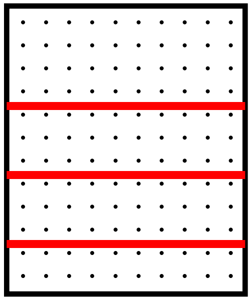
- A.k.a.: biclustering, block clustering, …

- Many objective functions in co-clustering
  - This is one of the easier
  - Others factor out row-column average (priors)
  - Others based on information theoretic ideas (e.g. KL divergence)

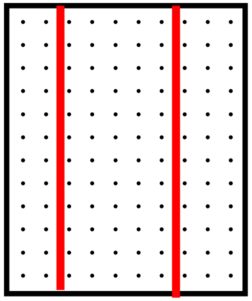
- A lot of existing work, but mostly heuristic
  - k-means style, alternate between rows/columns
  - Spectral techniques

Algorithm

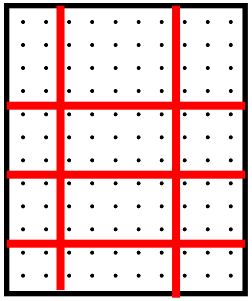
1. Cluster rows of $\boldsymbol{A}$



2. Cluster columns of $\boldsymbol{A}$



3. Combine



Properties of the algorithm

Theorem 1. Algorithm with optimal row/column clusterings is 3-approximation to co-clustering optimum.

Theorem 2. For L_2 distance function, the algorithm with optimal row/column clusterings is a 2-approximation.

Algorithm--details

- Clustering of the **n** rows of **A** assigns every row to a cluster with cluster name **{1,…,k}**
  - $R(i)= r_i$ with $1\leq r_i \leq k$
- Clustering of the **m** columns of **A** assigns every column to a cluster with cluster name **{1,…,ℓ}**
  - $C(j)=c_j$ with $1\leq c_j \leq \ell$
- $A'(i,j) = \{r_i,c_j\}$
- **(i,j)** is in the same co-cluster as **(i’,j’)** if $A'(i,j)=A'(i',j')$