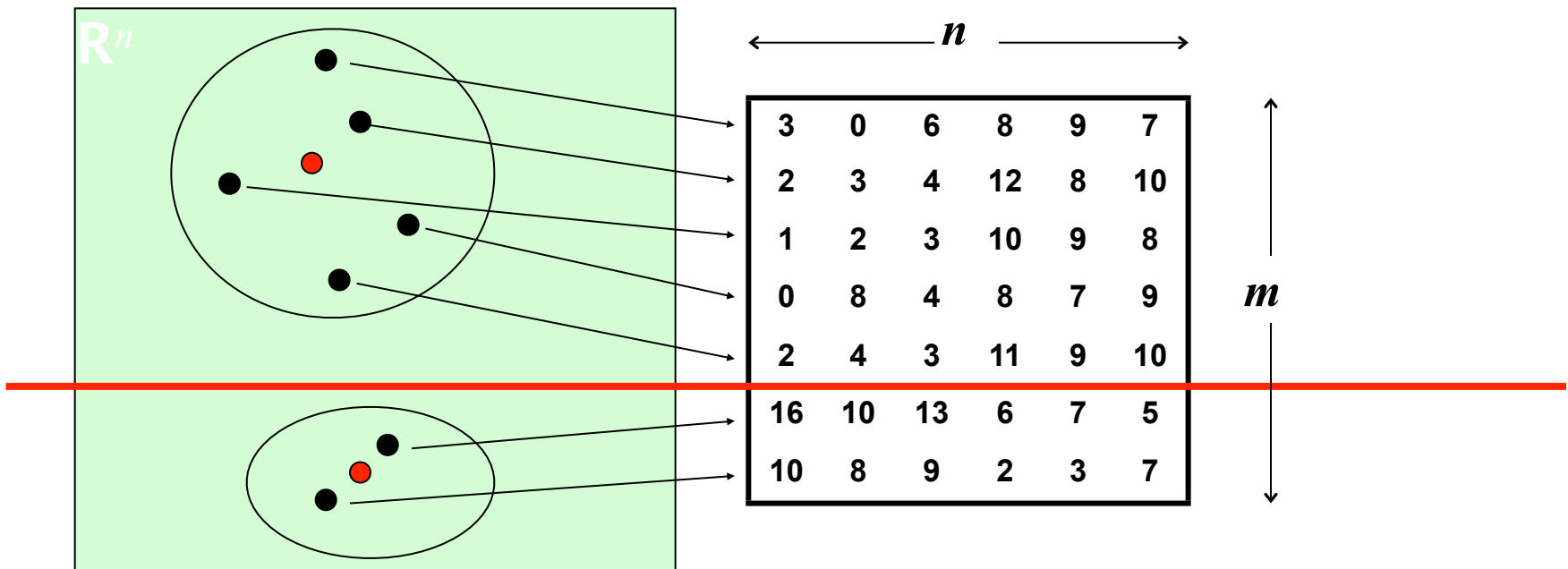


Co-clustering or Bi-clustering

- 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



Co-Clustering

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

$\ell = 2$

$k = 2$

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

A

Co-cluster

Motivation: Sponsored Search

The screenshot shows a Yahoo! search results page for the query 'car insurance'. The search bar at the top contains 'car insurance' and the Yahoo! logo is in the top right. Below the search bar, there are navigation links for 'Web', 'Images', 'Video', 'Local', and 'Shopping'. The search results are displayed in a list format. A red box highlights a section of sponsored results on the right side of the page, which includes:

- 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.

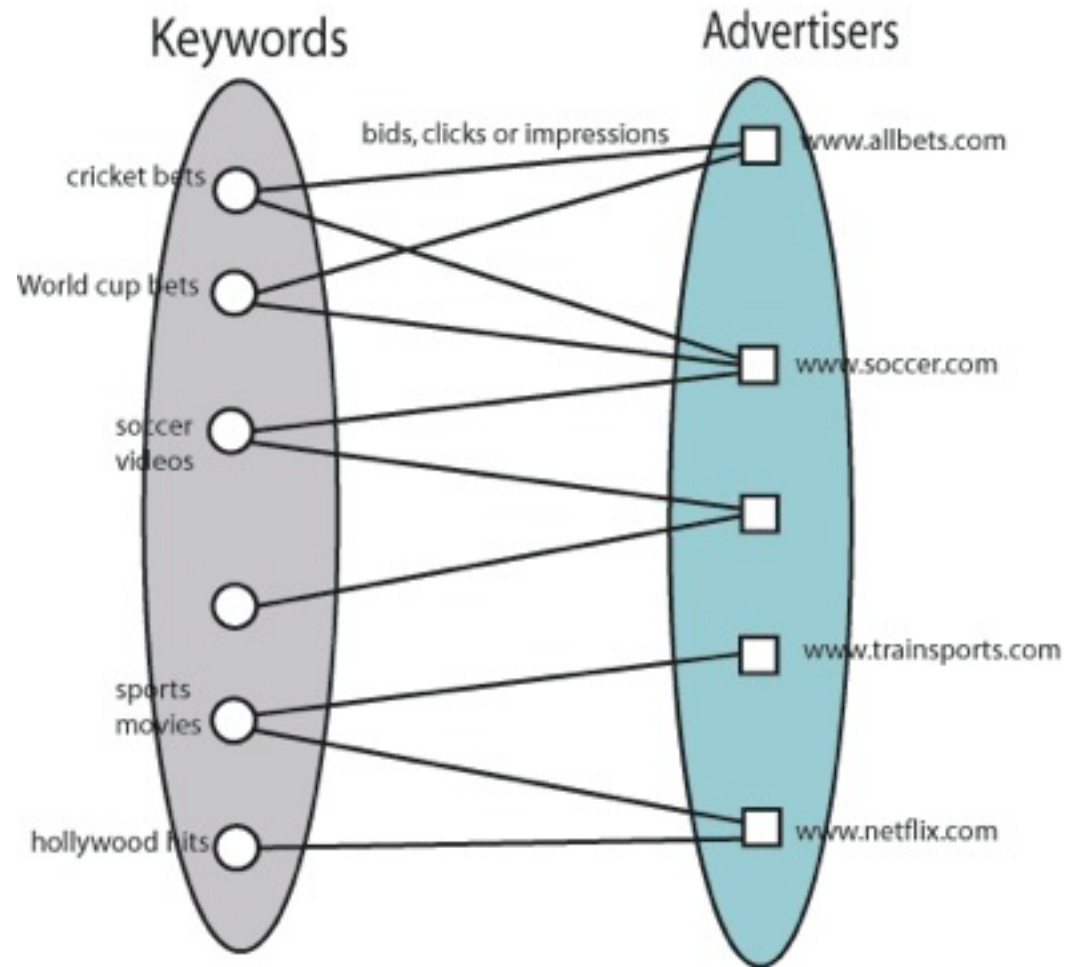
Red arrows point from the word 'Ads' to the highlighted sponsored results section.

Ads

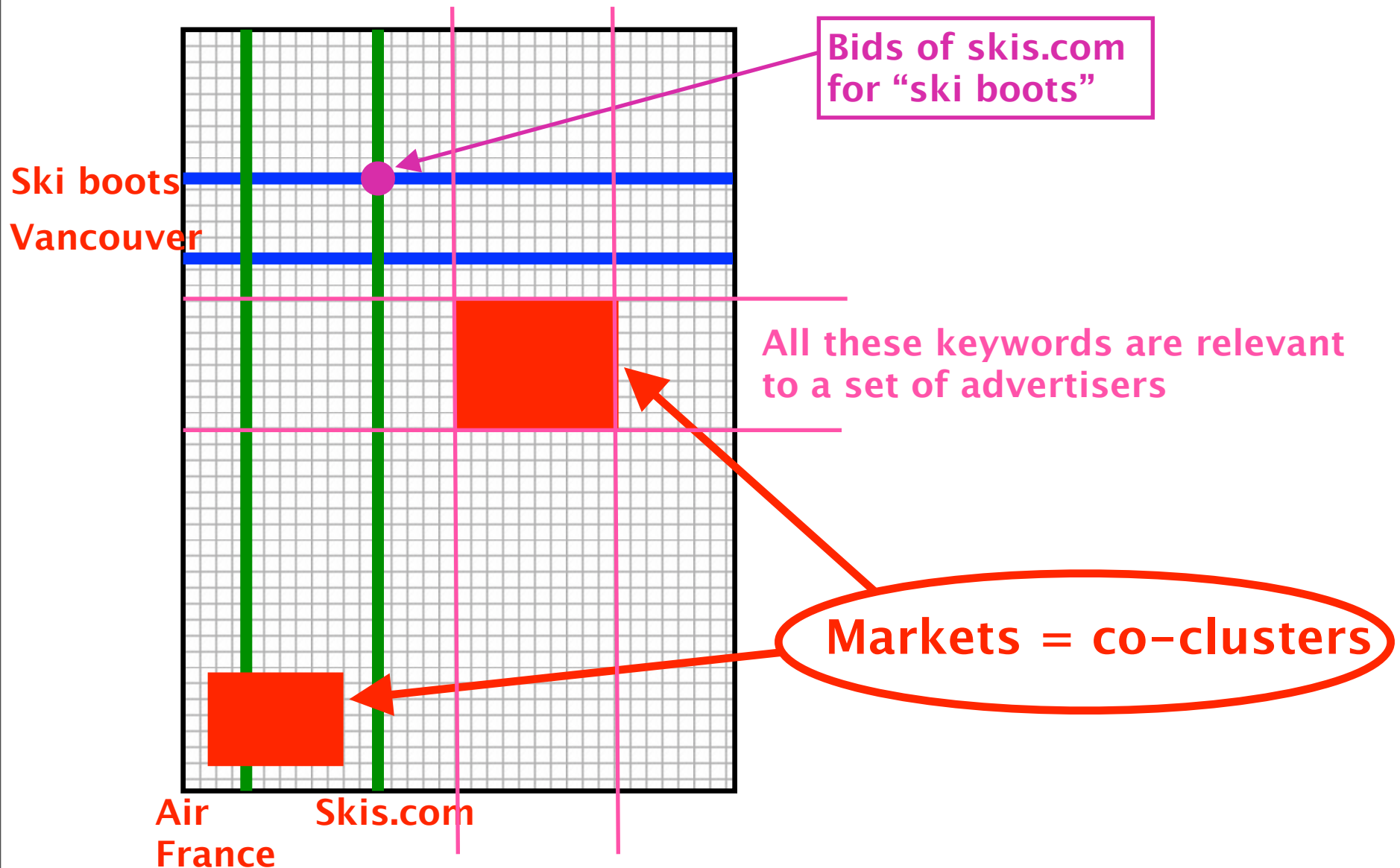
- Advertisers bid on keywords
- A user makes a query
- Show ads of advertisers that are relevant and have high bids
- User clicks or not an ad

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



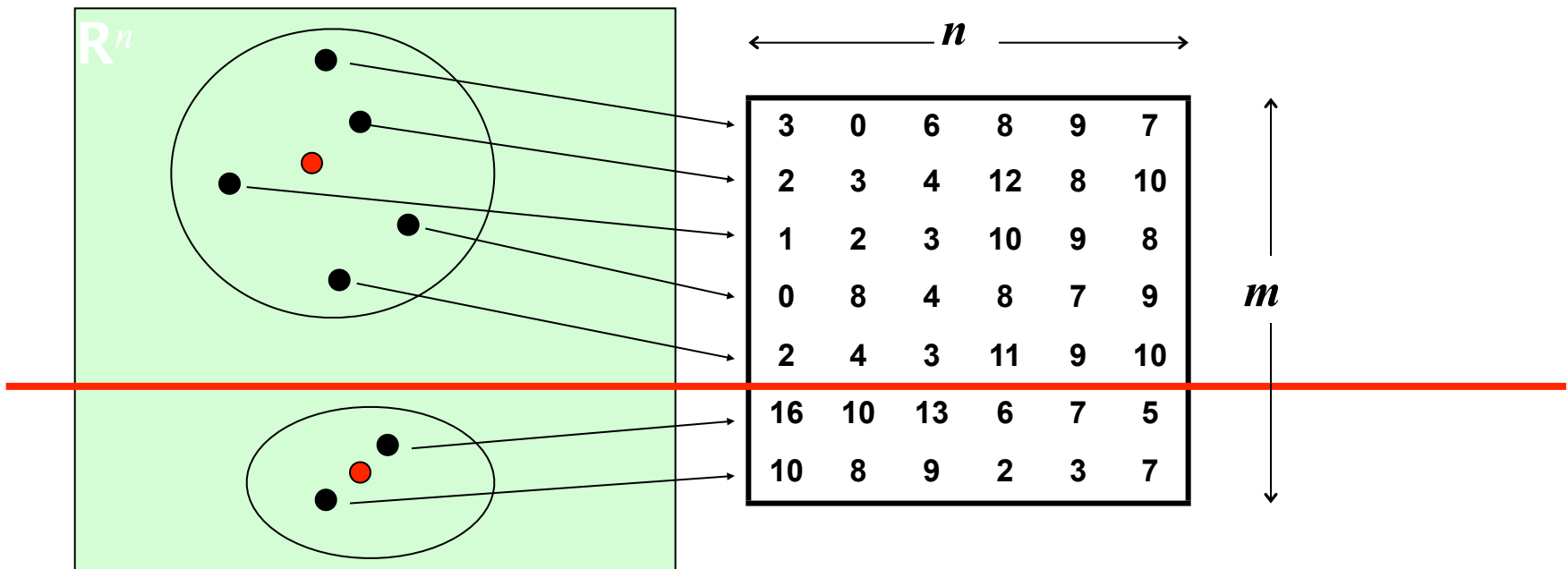
Co-Clustering in Sponsored Search

Applications:

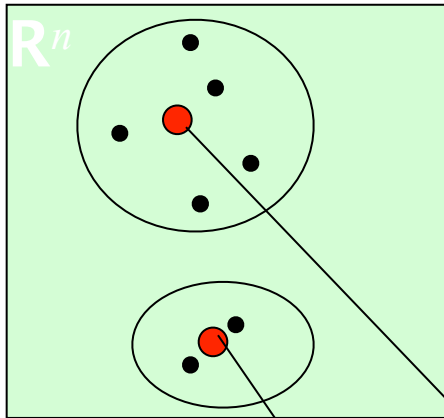
- Keyword suggestion
 - Recommend to advertisers other relevant keywords
- Broad matching / market expansion
 - Include more advertisers to a query
- Isolate submarkets
 - Important for economists
 - Apply different advertising approaches
- Build taxonomies of advertisers / keywords

Clustering of the rows

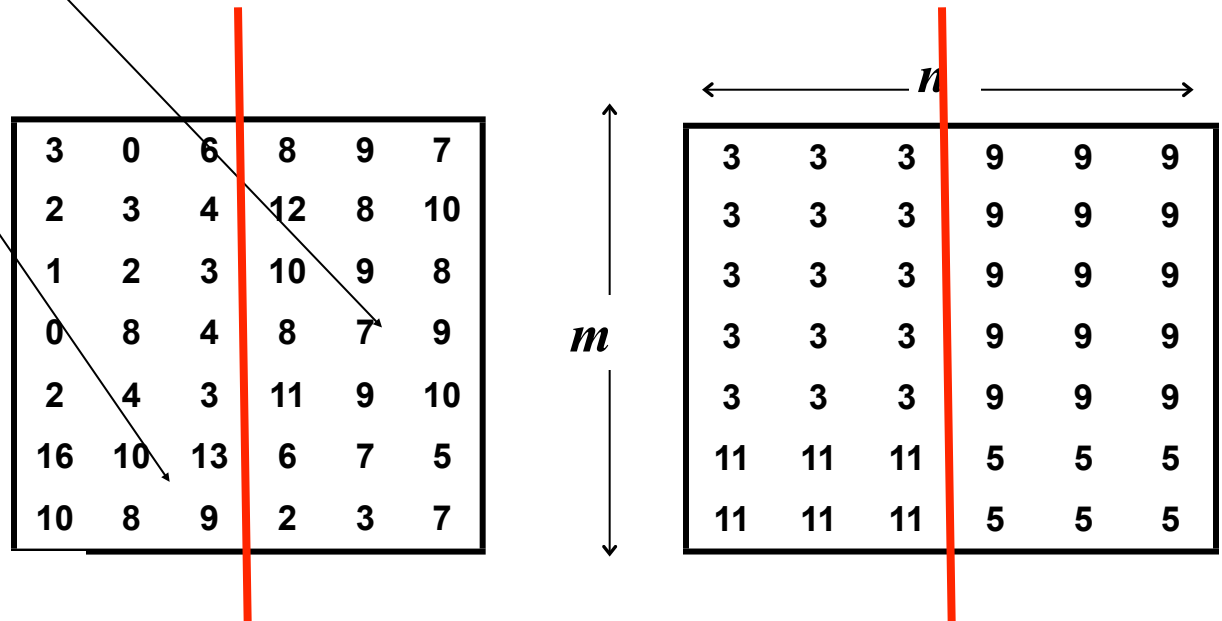
- m points in \mathbf{R}^n
- Group them to k clusters
- Represent them by a matrix $A \in \mathbf{R}^{m \times n}$
 - A point corresponds to a row of A
- **Clustering:** Partitioning of the rows into k groups



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



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:**

$$\text{cost} = \sum_{i=1 \dots n} \sum_{j=1 \dots m} (A(i,j) - A'(i,j))^2$$

- **k-median clustering:**

$$\text{cost} = \sum_{i=1 \dots n} \sum_{j=1 \dots m} |A(i,j) - A'(i,j)|$$

Co-Clustering

- **Co-Clustering:** Cluster rows and columns of $A \in \mathbb{R}^{m \times n}$ simultaneously
- k row clusters, ℓ 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'

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

Original data A

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

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:**

$$\text{cost} = \sum_{i=1 \dots n} \sum_{j=1 \dots m} (A(i,j) - A'(i,j))^2$$

- **k-median Co-clustering:**

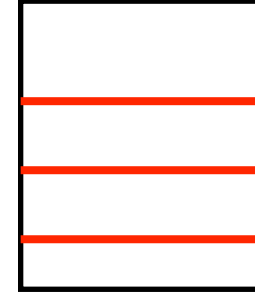
$$\text{cost} = \sum_{i=1 \dots n} \sum_{j=1 \dots m} |A(i,j) - A'(i,j)|$$

Some Background

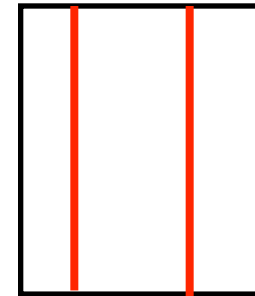
- A.k.a.: biclustering, block clustering, ...
- Many objective functions in co-clustering
 - This is one of the easier
 - Others factor out row-column average (priors)
 - Others based on information theoretic ideas (e.g. KL divergence)
- A lot of existing work, but mostly heuristic
 - k -means style, alternate between rows/columns
 - Spectral techniques

Algorithm

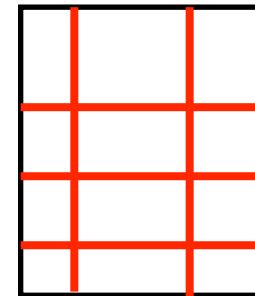
1. Cluster rows of A



2. Cluster columns of 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 $\{1, \dots, k\}$
 - $R(i) = r_i$ with $1 \leq r_i \leq k$
- Clustering of the m columns of A assigns every column to a cluster with cluster name $\{1, \dots, \ell\}$
 - $C(j) = c_j$ 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 $A'(i,j) = A'(i',j')$