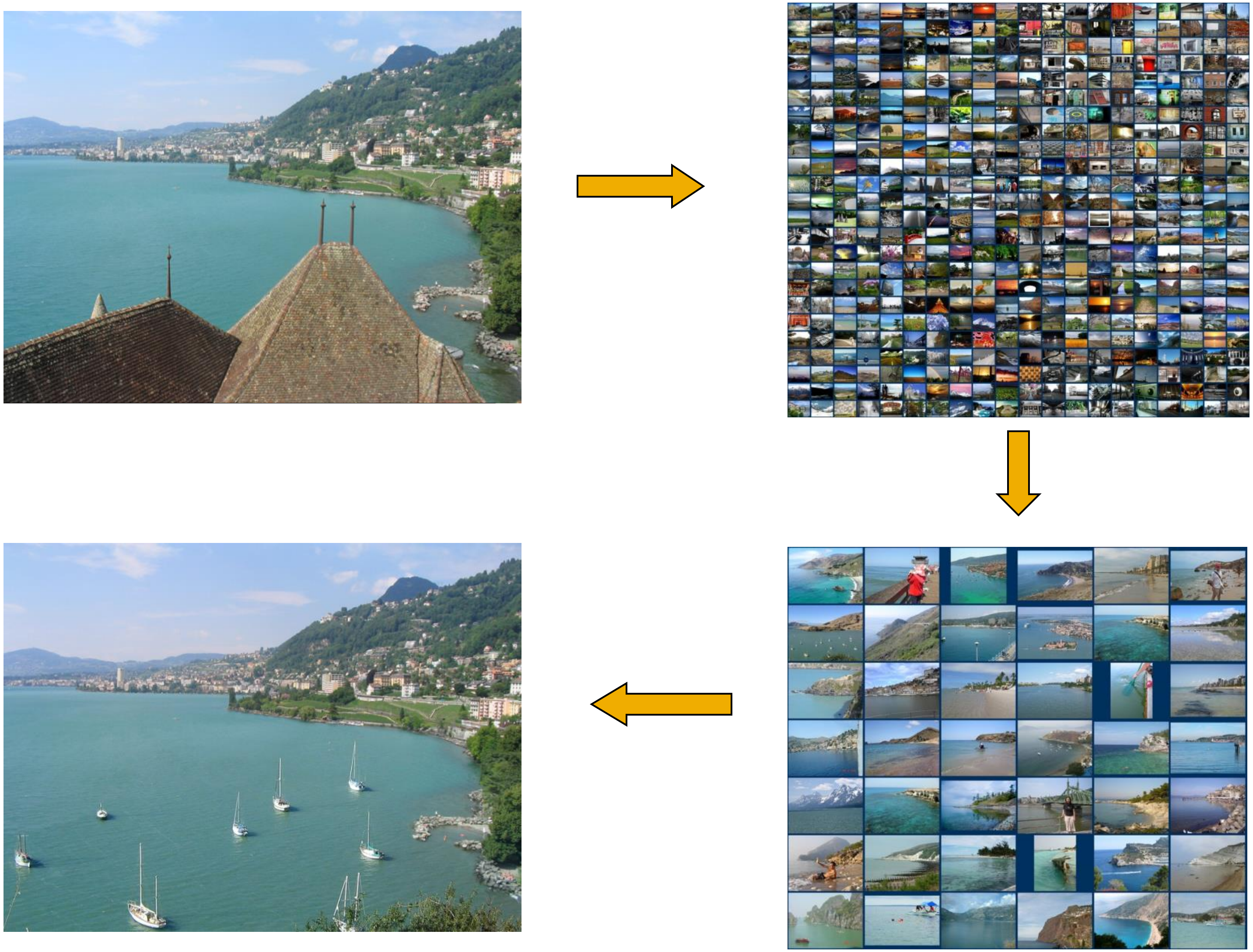
#16 - Finding Similar Items (1)

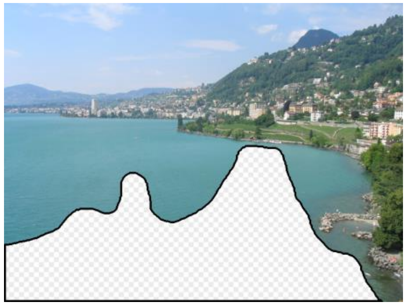
Some motivations…

[Hays and Efros, SIGGRAPH 2007]

Scene Completion Problem

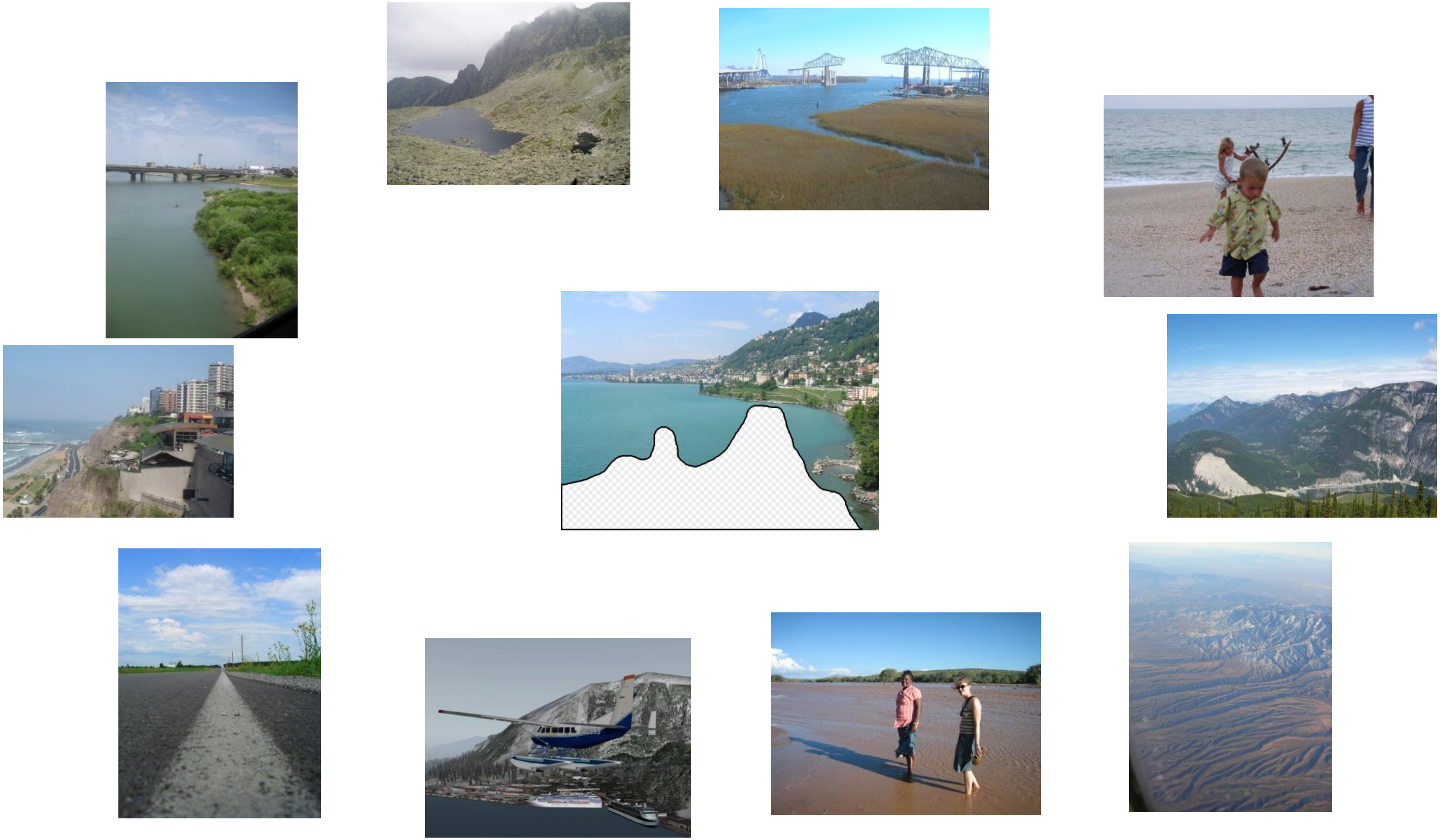


Scene Completion Problem



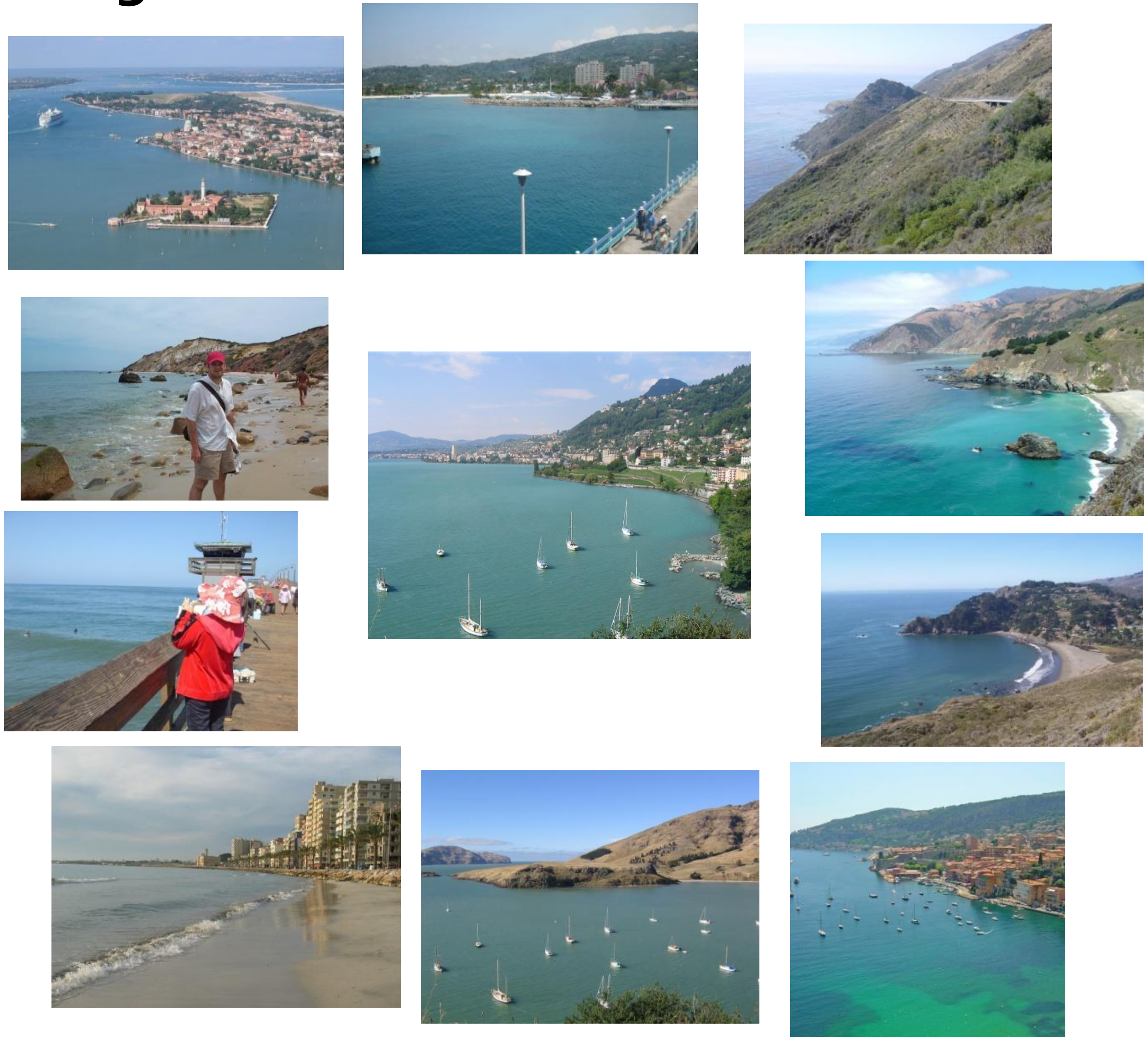
Scene Completion Problem

10 nearest neighbors from a collection of 20,000 images



Scene Completion Problem

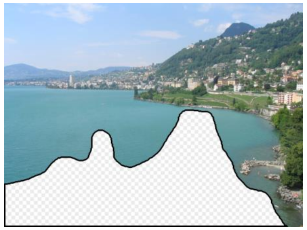
10 nearest neighbors from a collection of 2 million images



A Common Metaphor

- **Many problems can be expressed as finding “similar” sets:**
  - **Find near-neighbors in high-dimensional space**

- **Examples:**
  - **Pages with similar words**
    - For duplicate detection, classification by topic
  - **Customers who purchased similar products**
    - Products with similar customer sets
  - **Images with similar features**
    - Users who visited similar websites
  - Clients that bought similar books
  - People that scored similar restaurants, movies… etc



Problem for Today's Lecture

- **Given: High dimensional data points $x_1, x_2, \ldots$**
  - **For example:** Image is a long vector of pixel colors

$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 2 & 1 \\ 0 & 1 & 0 \end{bmatrix} \rightarrow [1\ 2\ 1\ 0\ 2\ 1\ 0\ 1\ 0]$$

- **And some distance function $d(x_1, x_2)$**
  - Which quantifies the "distance" between x_1 and x_2
- **Goal:** Find **all pairs of data points** (x_i, x_j) that are within some distance threshold $d(x_i, x_j) \leq s$
- **Note:** Naïve solution would take $O(N^2)$ ☹ where N is the number of data points
- **MAGIC: This can be done in $O(N)$!! How?**

Distance Measures

- **Goal: Find near-neighbors in high-dim. space**
  - We formally define "near neighbors" as points that are a "small distance" apart

- For each application, we first need to define what "**distance**" means

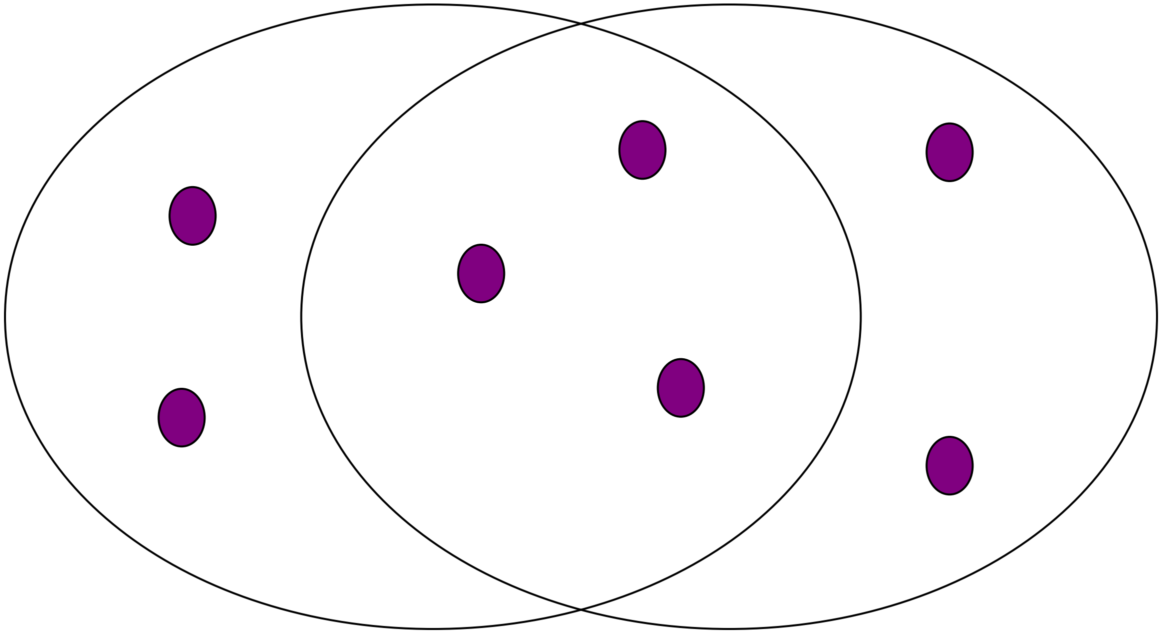
Jaccard distance/similarity

- The **Jaccard similarity** of two **sets** is the size of their intersection divided by the size of their union:

  $\boldsymbol{sim(C_1, C_2) = |C_1 \cap C_2| / |C_1 \cup C_2|}$

- **Jaccard distance:**
  - $\boldsymbol{d(C_1, C_2) = 1 - |C_1 \cap C_2| / |C_1 \cup C_2|}$

Example



3 in intersection
7 in union

Jaccard similarity= 3/7
Jaccard distance = 4/7

Application example #1 (Matlab)

- Detect similar texts

- Toy example

- Sets are the (unique) words found in the documents
  - No post-processing

- Direct application of Jaccard distance

Main tasks

1. Create Sets of Words for all documents

    ```
    Sets{1}=getSetOfWordsFromFile('texto1.txt')
    Sets{2}=getSetOfWordsFromFile('texto2.txt')
    …
    ```

2. Calculate Jaccard distance for each pair of documents

    ```
    distJ=calcDistancesJ(Sets);
    ```

3. Determine pairs that have distances below a predifined threshold

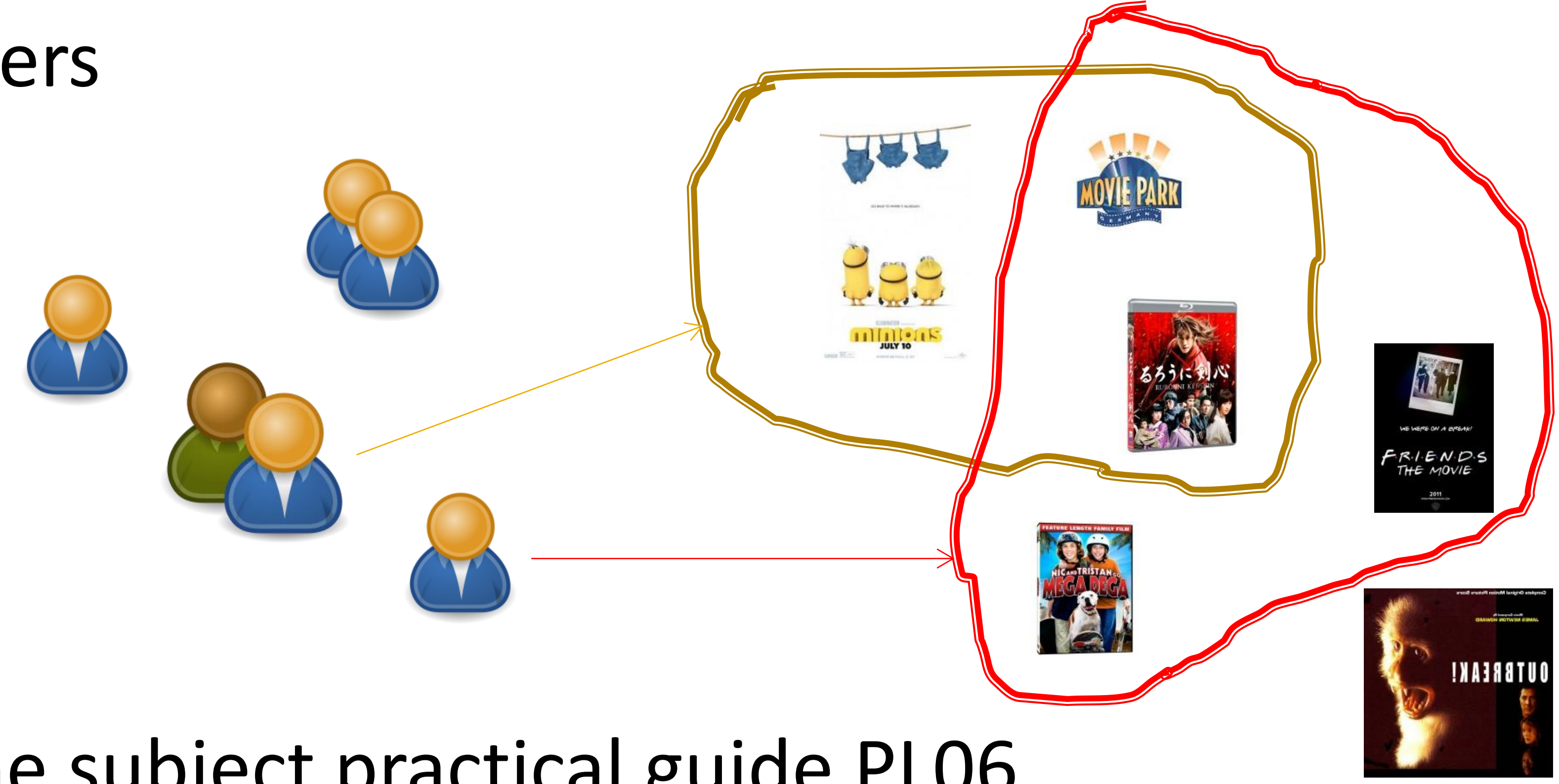
    ```
    Similar=findSimilar(distJ,threshold,ids);
    ```

4. Show results

demoJaccard.m

Application example #2

- Determine similar sets of movies rated by users

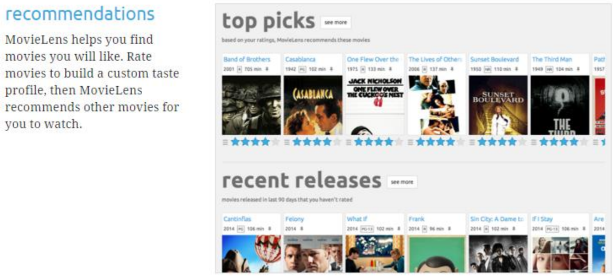


- The subject practical guide PL06

MovieLens

- MovieLens (http://movielens.org) is a web site that helps people find movies to watch.
  - It has hundreds of thousands of registered users.

Non-commercial, personalized movie recommendations.


recommendations

MovieLens helps you find movies you will like. Rate movies to build a custom taste profile, then MovieLens recommends other movies for you to watch.

top picks

based on your ratings, MovieLens recommends these movies

Band of Brothers 2001 705 min
Casablanca 1942 PG 102 min
One Flew Over the 1975 R 133 min
The Lives of Others 2006 R 137 min
Sunset Boulevard 1950 NR 110 min
The Third Man 1949 NR 104 min
Path 1957

recent releases

movies released in last 90 days that you haven't rated

Cantinflas 2014 PG 106 min
Felony 2014
What If 2014 PG-13 102 min
Frank 2014 R 96 min
Sin City: A Dame to 2014 R 102 min
If I Stay 2014 PG-13 106 min
Are 2014

MovieLens Datasets

- GroupLens Research has collected and made available rating data sets from the MovieLens web site (http://movielens.org).
- The data sets were collected over various periods of time, depending on the size of the set.
  - Available from: http://grouplens.org/datasets/movielens/

- There are several:
  - **MovieLens 100K Dataset**
  - MovieLens 1M Dataset
  - …

MovieLens 100K Dataset

- Stable benchmark dataset.
  - Released 4/1998.

- Aprox. 100 000 ratings
- from 943 users on 1682 movies.

- README

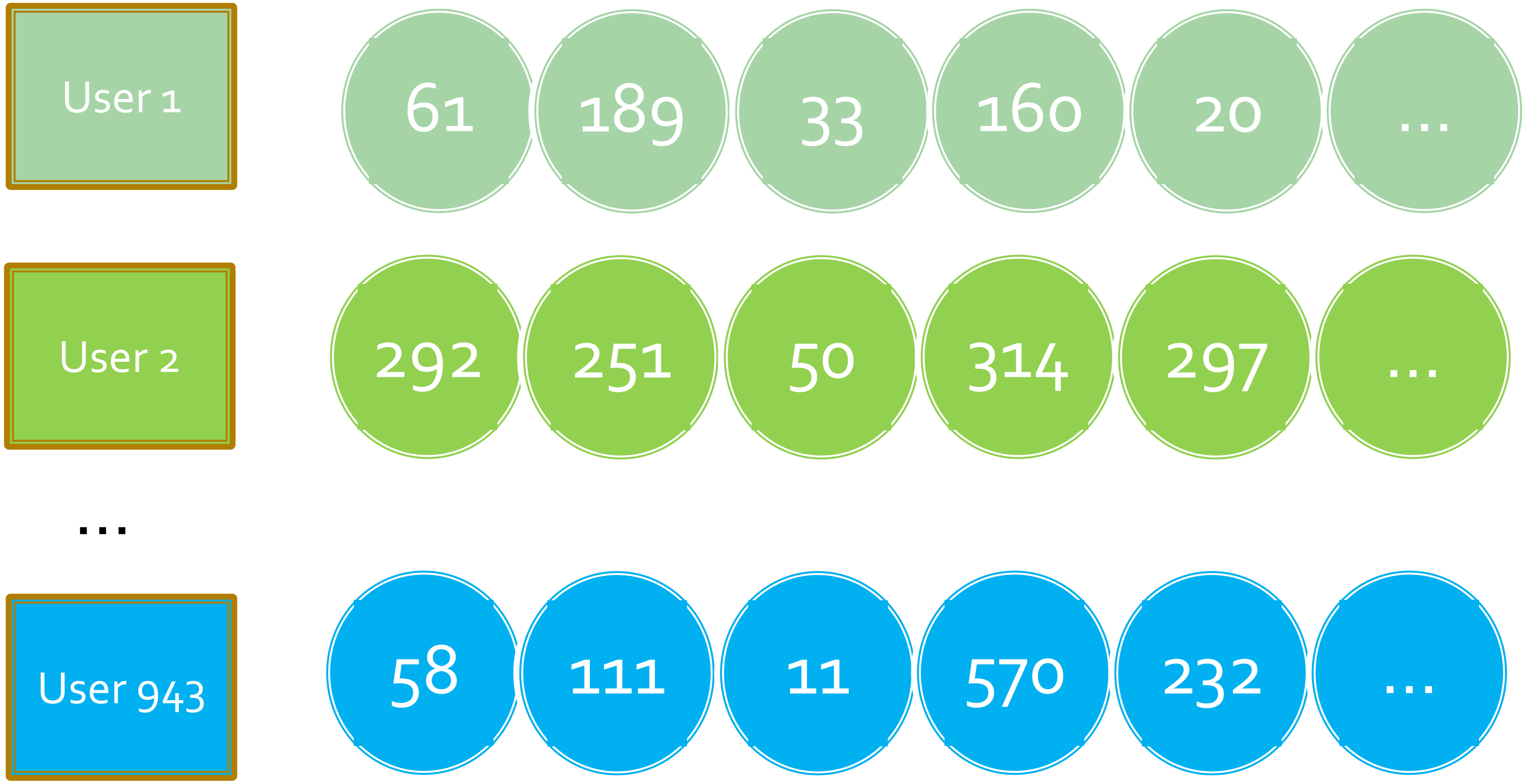
- Permalink: http://grouplens.org/datasets/movielens/100k/

File u.data

```
196         242         3           881250949
186         302         3           891717742
22          377         1           878887116
244         51          2           880606923
166         346         1           886397596
298         474         4           884182806
115         265         2           881171488
253         465         5           891628467
305         451         3           886324817
…
```

- first column contains the user ID
- second column the ID of a movie
  - rated by the user in the first column
- rating is in the third column
- fourth column is a timestamp.

Sets of movies (reviewed by each user)



User 1	61	189	33	160	20	…
User 2	292	251	50	314	297	…
…						
User 943	58	111	11	570	232	…

Demonstration

- First "solution"
  - Very slow
  - Direct use of Jaccard distance

demoJaccardMovies.m

Demonstration

- Results (similar sets):

```
328  788    distance = 0.327
408  898    distance = 0.161
489  587    distance = 0.370
```

Handling Large and Huge sets

→ Next Class