人人网
renren.com

基于Hadoop的 SNS统计平台和聚类推荐

白伯纯 张叶银

renren.com

人人网

- 2.2亿用户
- 平均190好友
- 月40亿照片访问

- 一成付费用户
- 五成用户每天使用
- 八成有真实的资料

人人网 renren.com

机遇

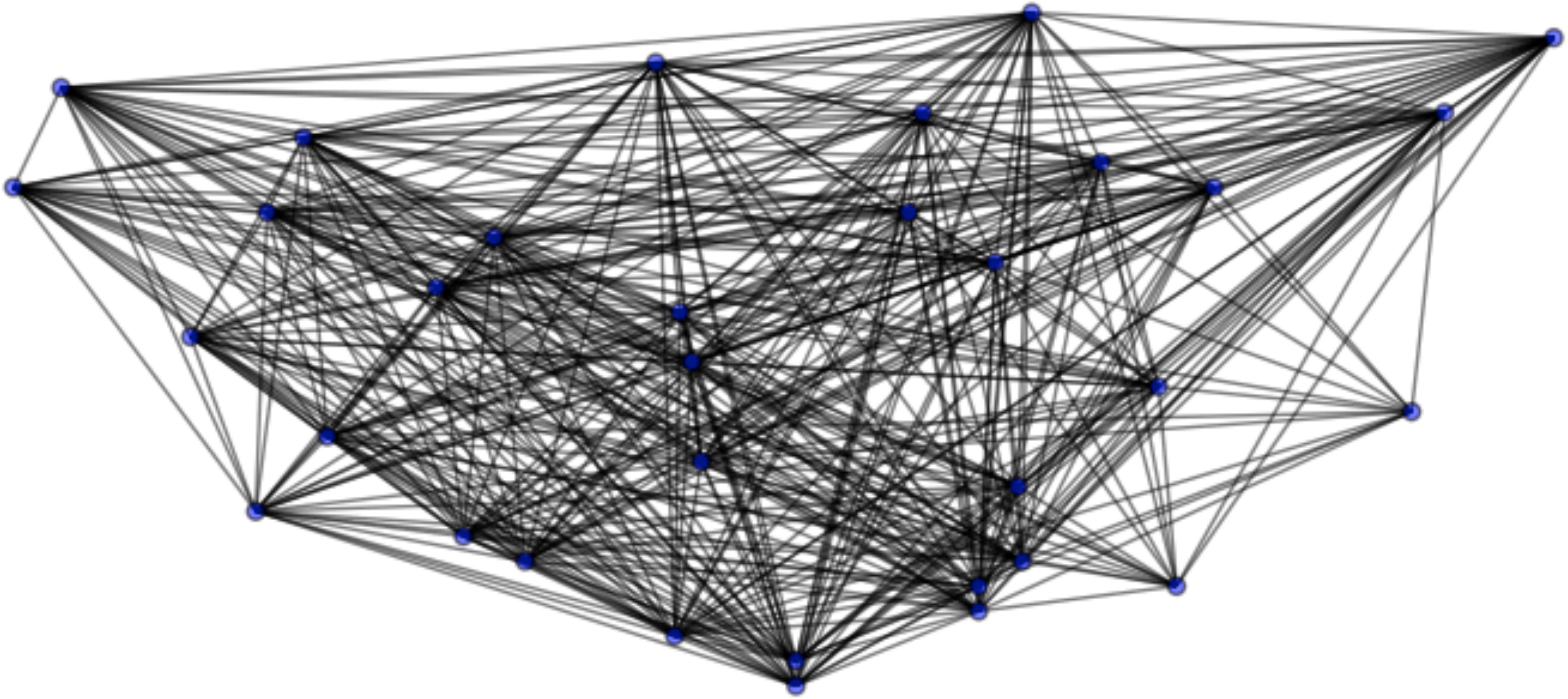
- 唯一标识

机遇

- 唯一标识

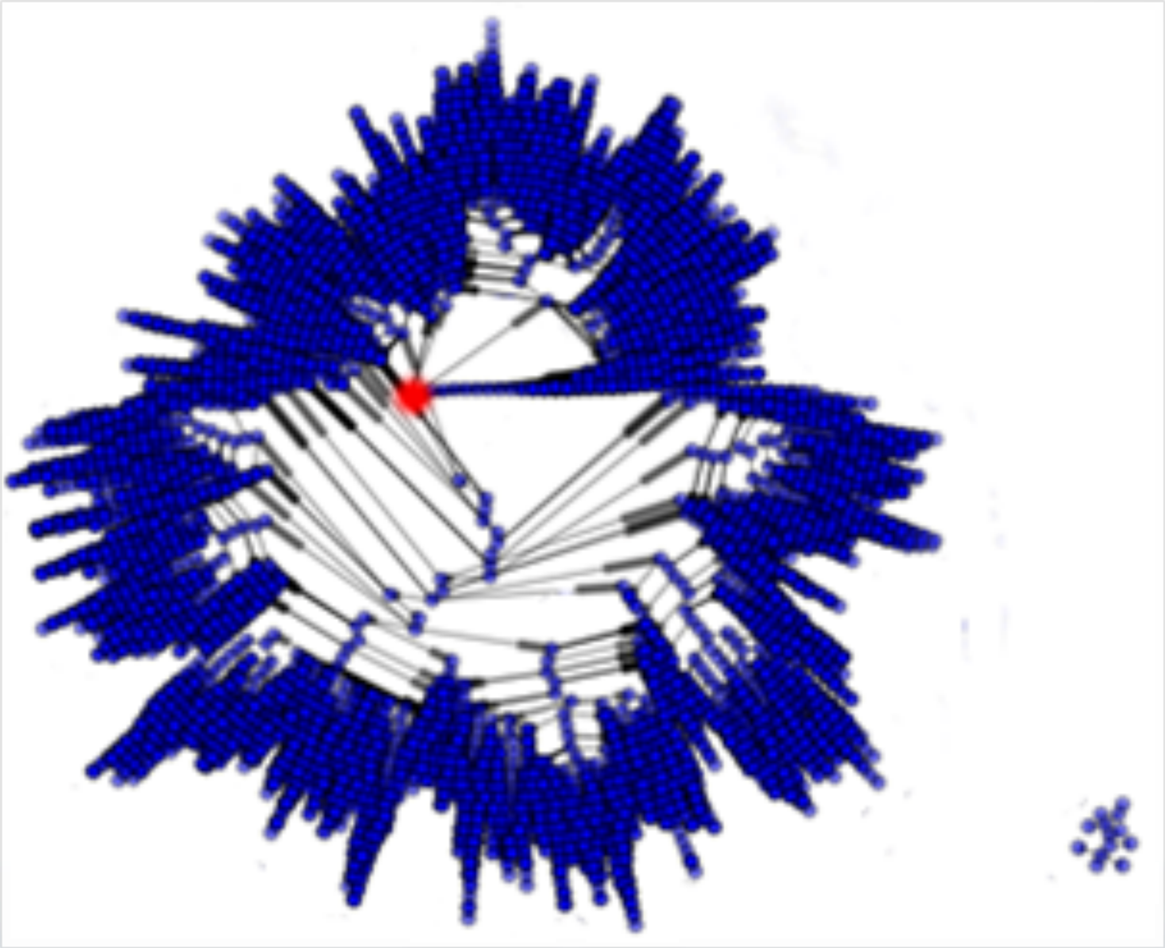
机遇

- 结构化数据

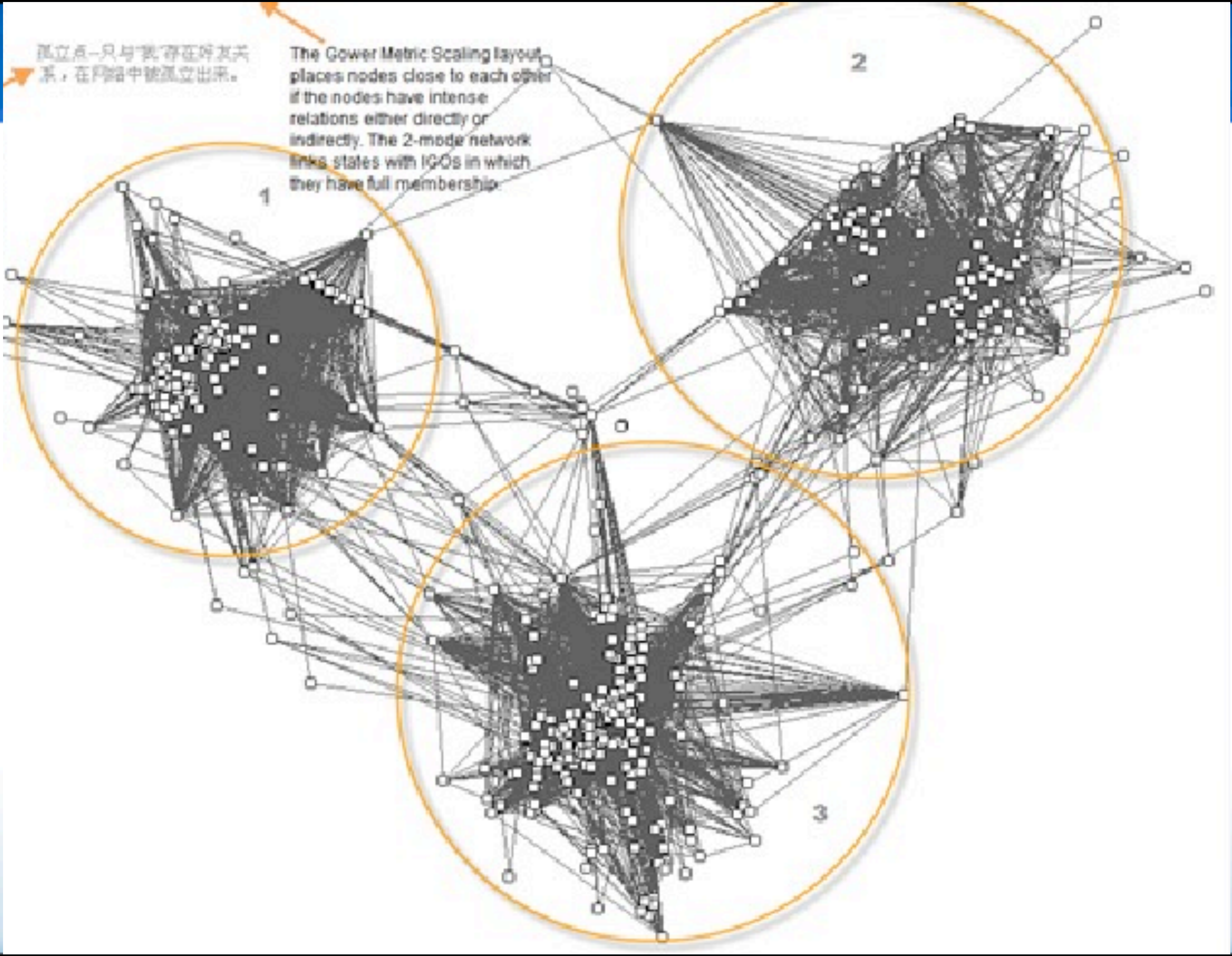


挑战

- 高复杂度计算



Saturday, December 3, 11



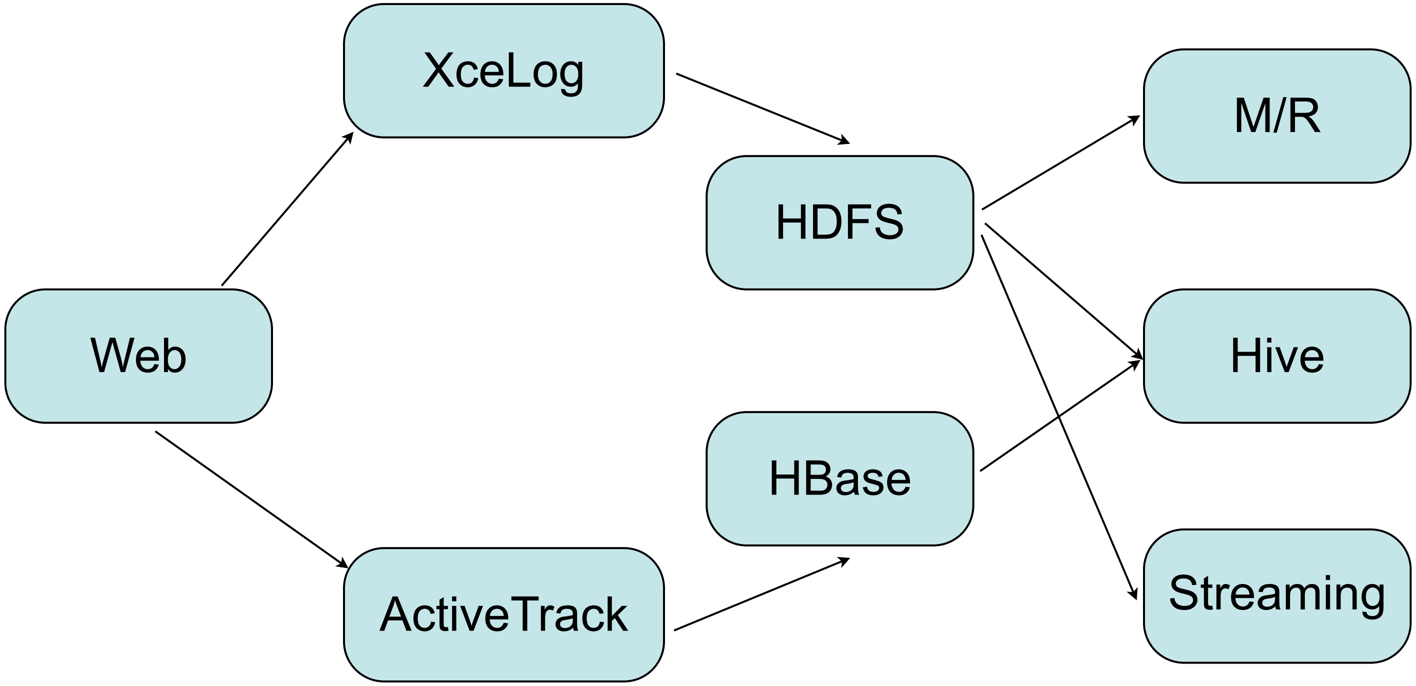
孤立点–只与"我"存在朋友关系，在网络中被孤立出来。

The Gower Metric Scaling layout places nodes close to each other if the nodes have intense relations either directly or indirectly. The 2-mode network links states with IGOs in which they have full membership.

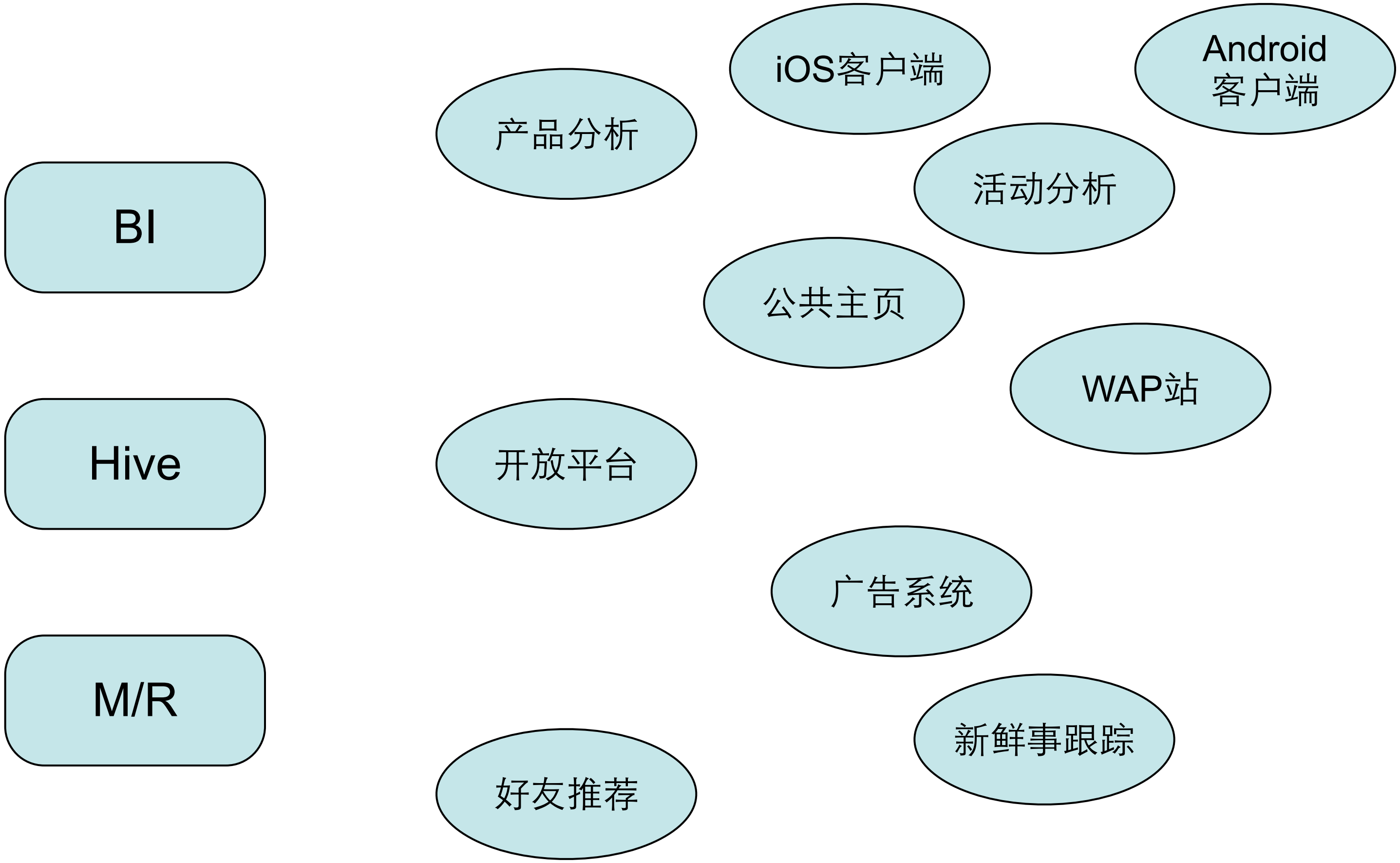
部署

- 200台
  - Hadoop 0.21.0
  - 4k+任务/天
  - 700TB Used/1.2PB Total
  - Hive/HBase/Streaming
- 30台
  - Hadoop 0.20.3
  - HBase only

结构



体系结构



BI

Hive

M/R

产品分析

iOS客户端

Android 客户端

活动分析

公共主页

WAP站

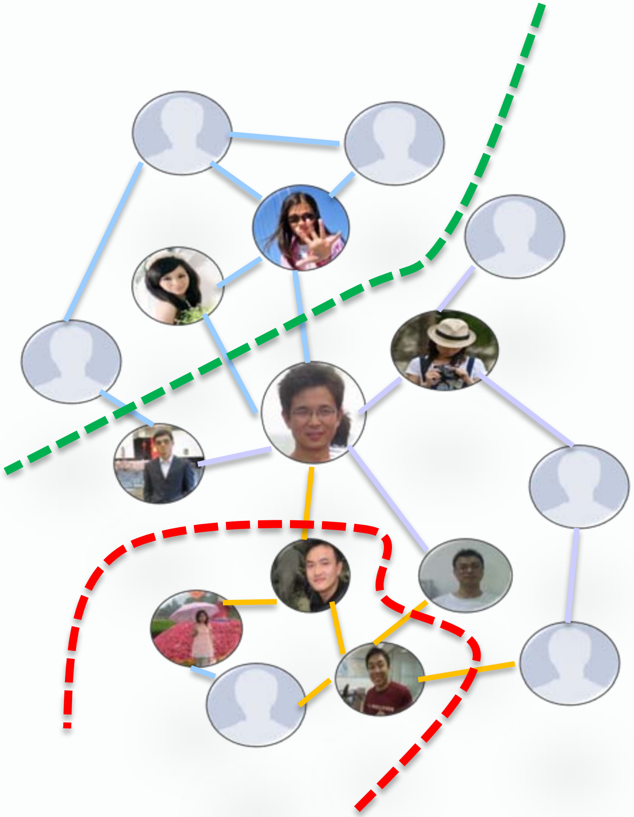
开放平台

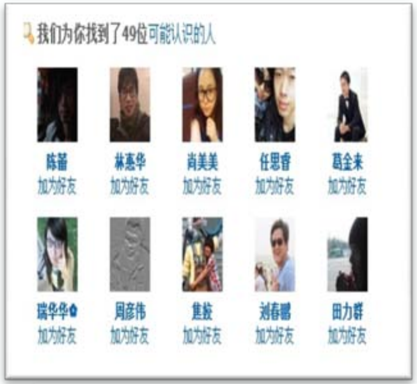
广告系统

新鲜事跟踪

好友推荐

Social computing at Renren





我们为你找到了49位可能认识的人

陈蕾 加为好友	林惠华 加为好友	肖美美 加为好友	任思香 加为好友	葛全来 加为好友
瑞华华 加为好友	周彦伟 加为好友	焦航 加为好友	刘春鹏 加为好友	田力群 加为好友

热门视频

拍客]结账偶遇开
享:66456 分享

当B-BOX牛人遇上诈骗集团打来的电话
分享:57929 分享

《你猜你猜你猜猜猜》夜店女王素颜太惊人 …
分享:68530 分享

《索命DV》延边大学探险社团鬼楼集体失踪死…
分享:28231 分享

热门相册

查看更多

热门日志

查看更多

【某A的折纸课堂】如何折出世界上最强的三大纸飞机之二：复仇者

自:汗尕州 Victo

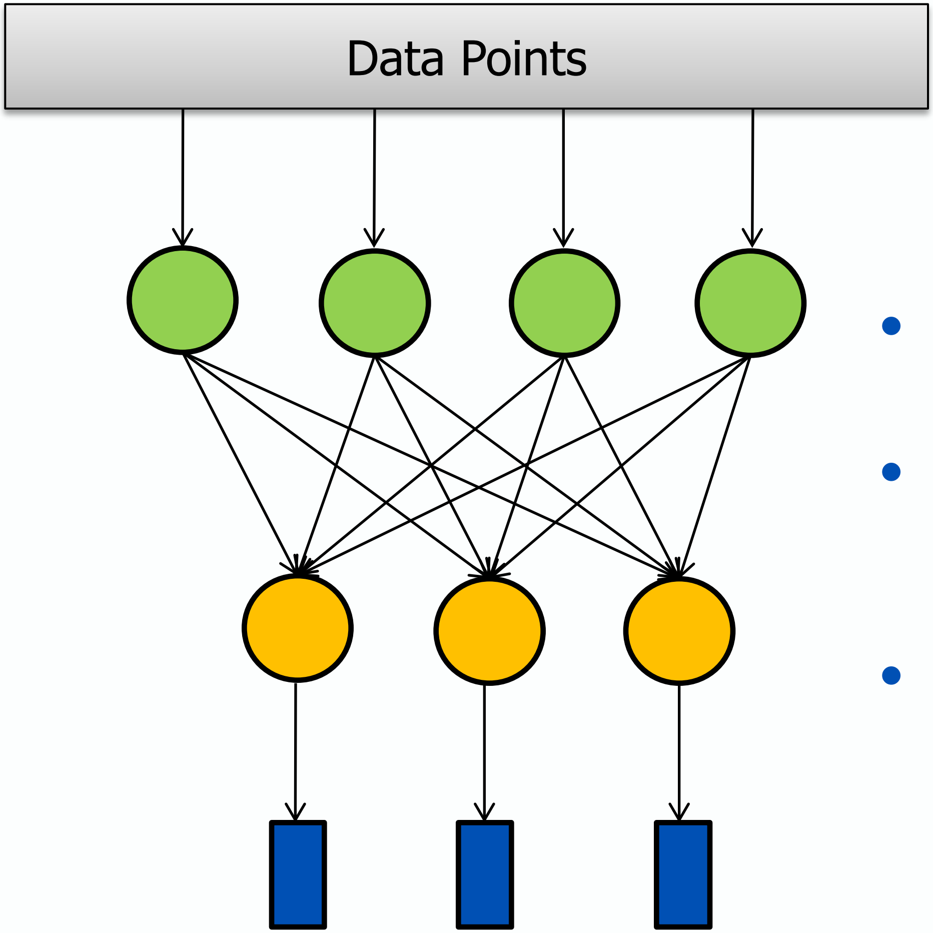
Data driven applications

Distribute

- TB of daily log data to be analyzed
- Millions of blogs, videos to be recommended
- Hundreds of millions of friends to be recommended

The most Computational-Intensive applications
with highly structured big data

MapReduce



Adjacent list data structure

- A sparse representation facilitate to pass graph structures from iteration to the next iteration

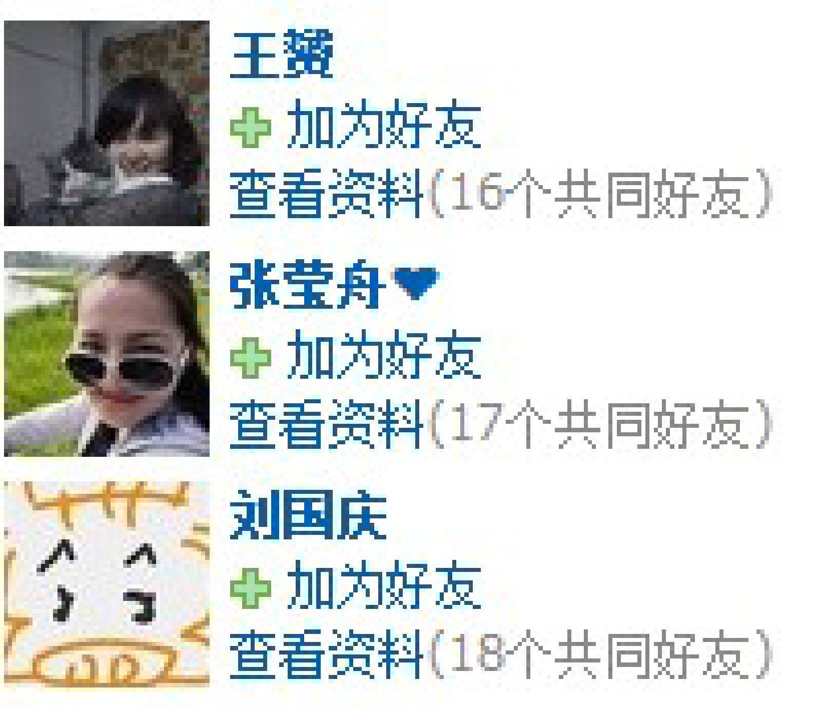
- Parallel Breadth-first search
  - To find shortest path in the graph
- Google page rank
  - Impact index passing through graph links
- Distributed k-means clustering
  - Clustering large data into pre-defined number of groups

Case I: Friend recommendation by agglomerative hierarchical clustering

- Primary problem of friend recommendation
  - User familiarity
    - Common friends
    - User profile
    - User access
    - User interest

People you may know

- Friends’ friends



王赟
✚ 加为好友
查看资料(16个共同好友)

张莹舟
✚ 加为好友
查看资料(17个共同好友)

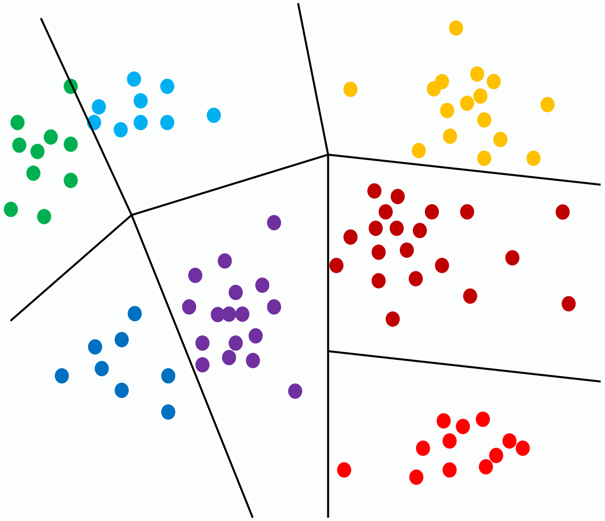
刘国庆
✚ 加为好友
查看资料(18个共同好友)

$$similarity(user1, user2) = |friendset1 \cap friendset2|$$

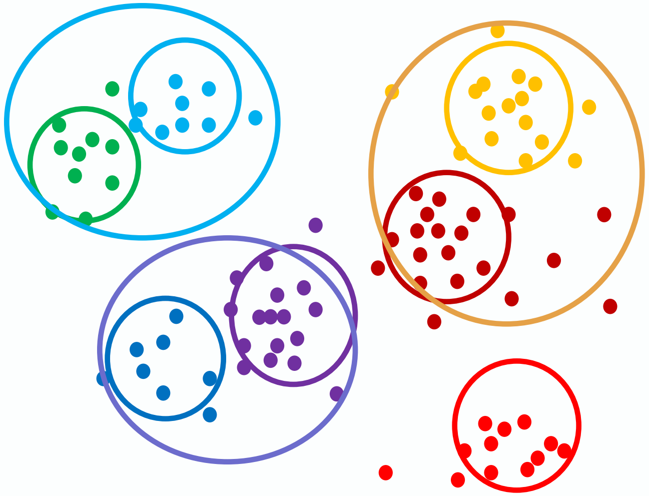
Hierarchy

- Clustering to find communities in social network
  - All in one community share some properties.
  - These overlapping communities reveal some social relationship of different levels.
  - They help to building new friendships in the social network.

Clustering: unsupervised learning

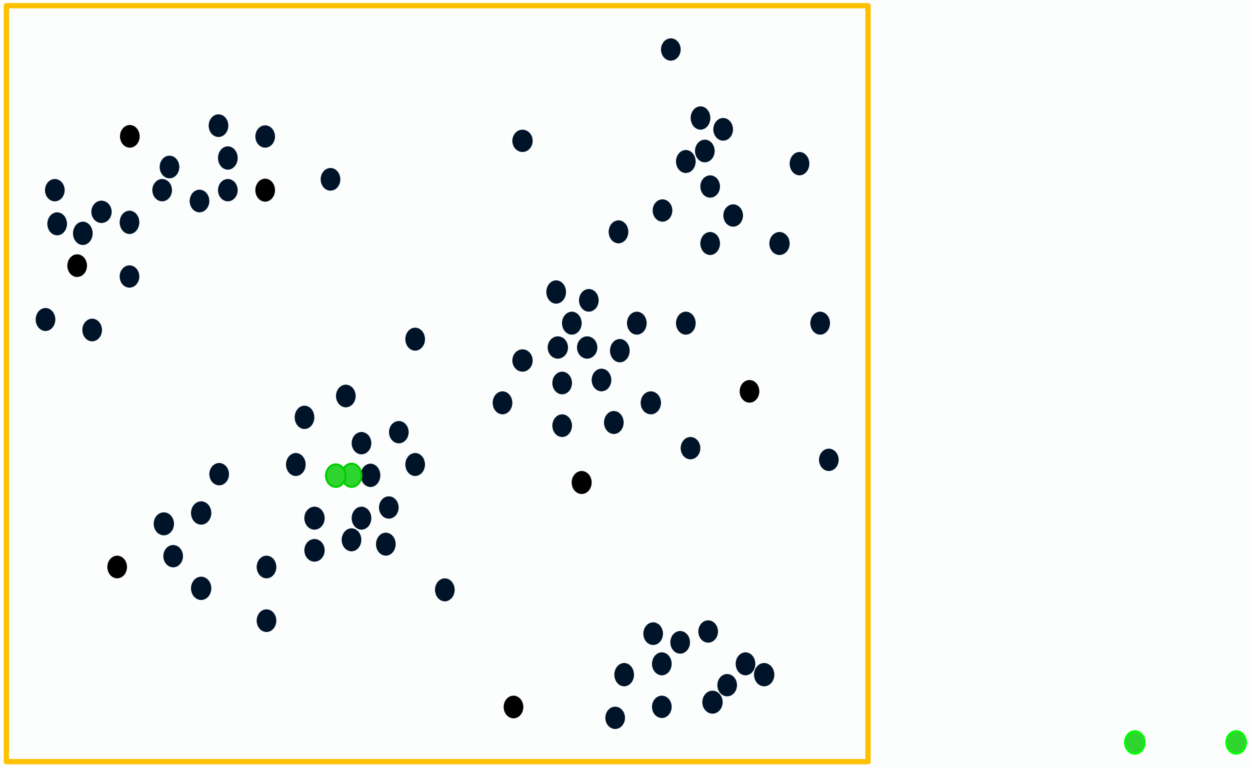


flat clustering

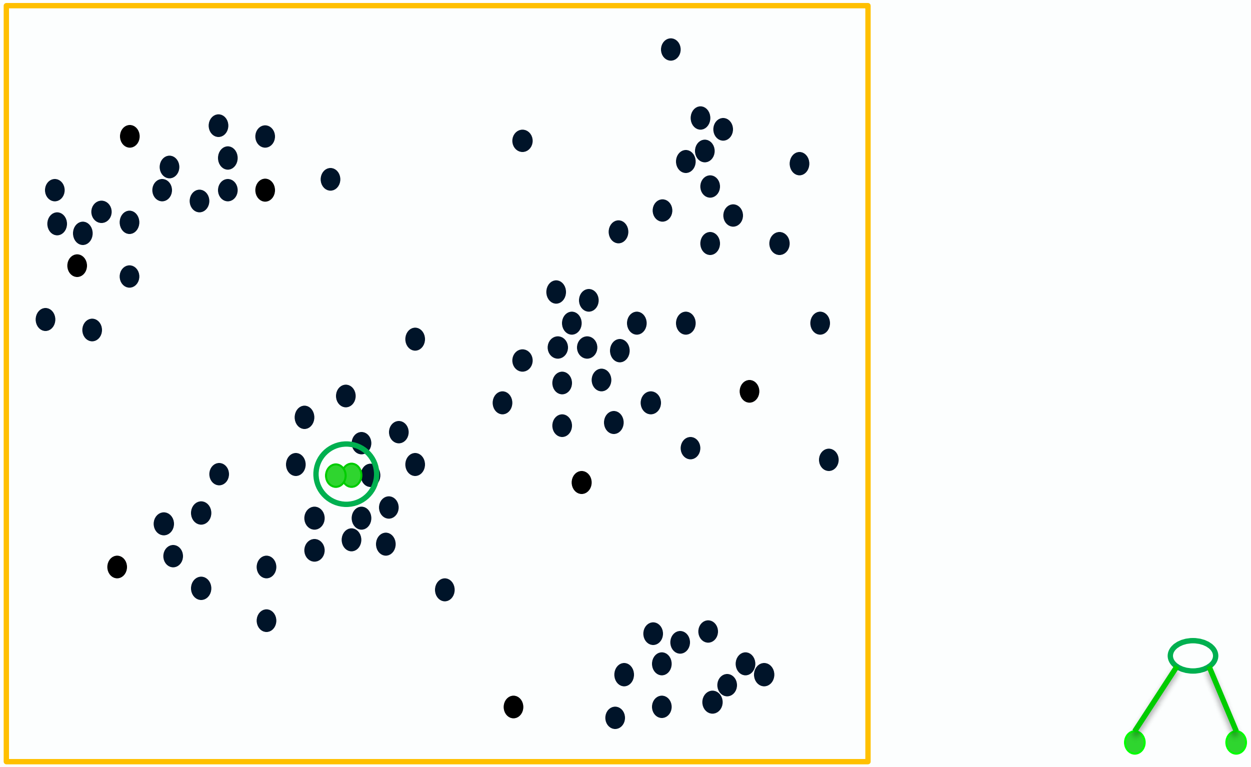


hierarchical clustering

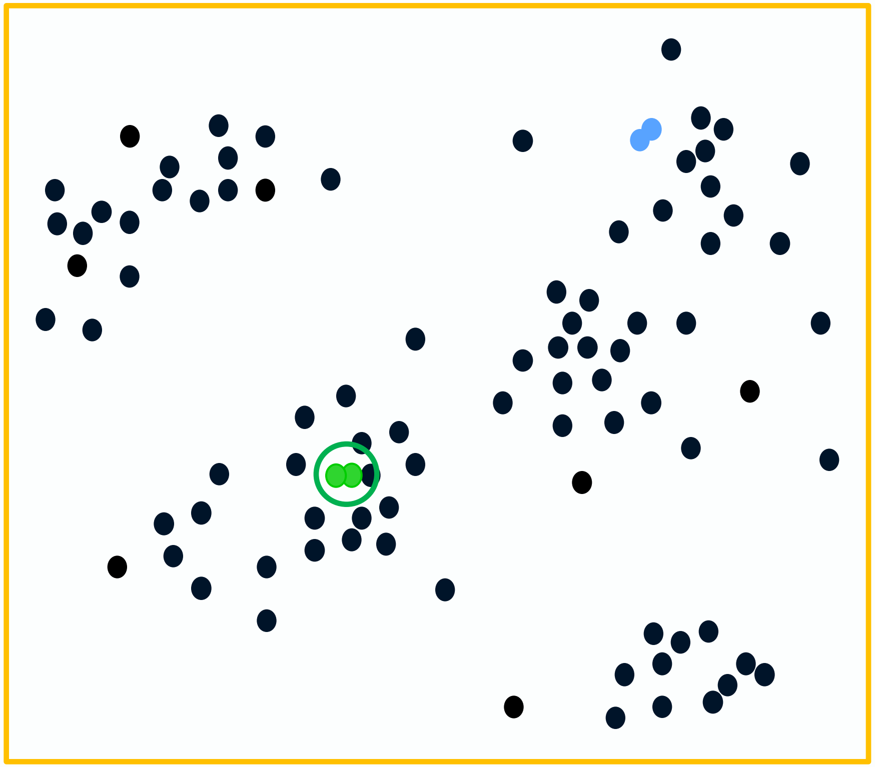
Hierarchical clustering

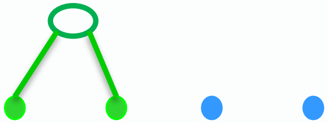


Hierarchical clustering

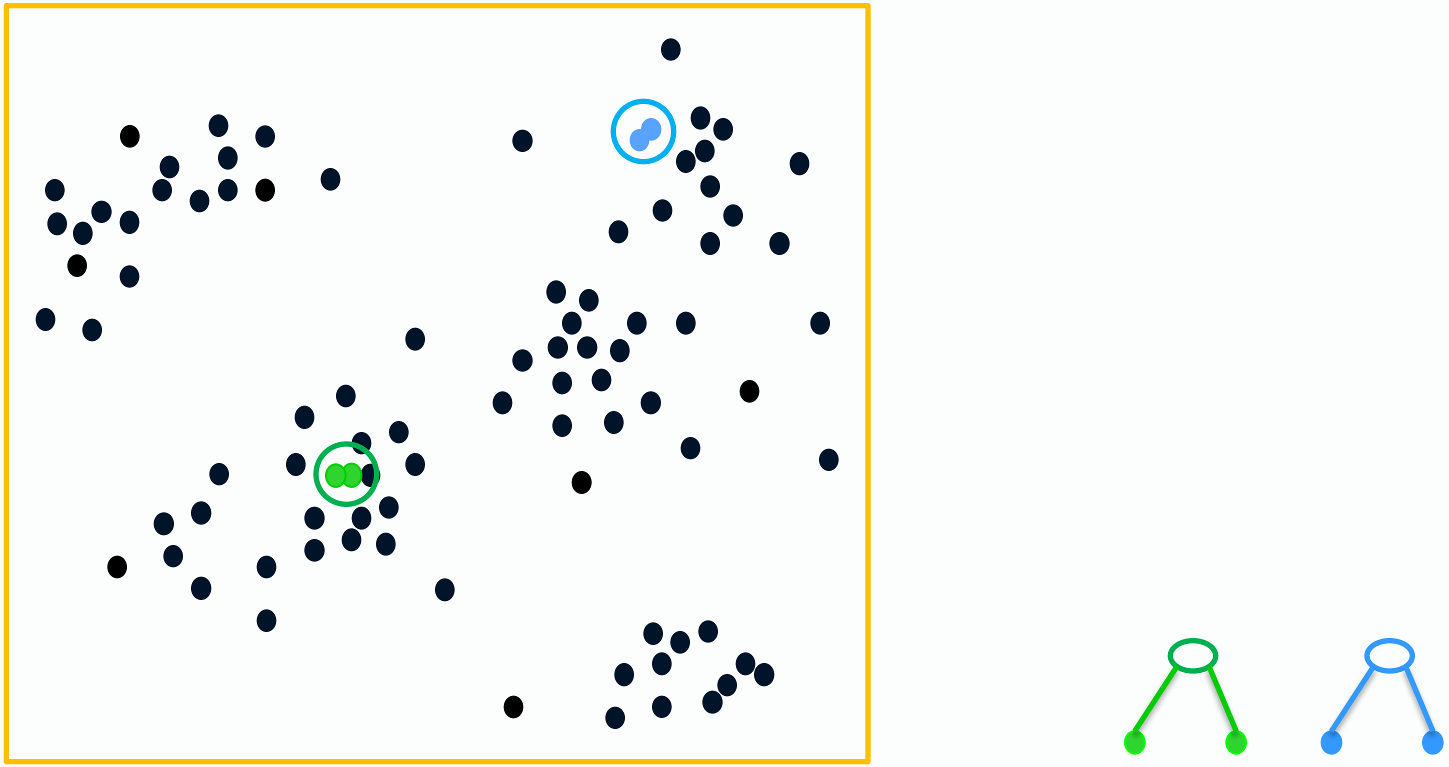


Hierarchical clustering

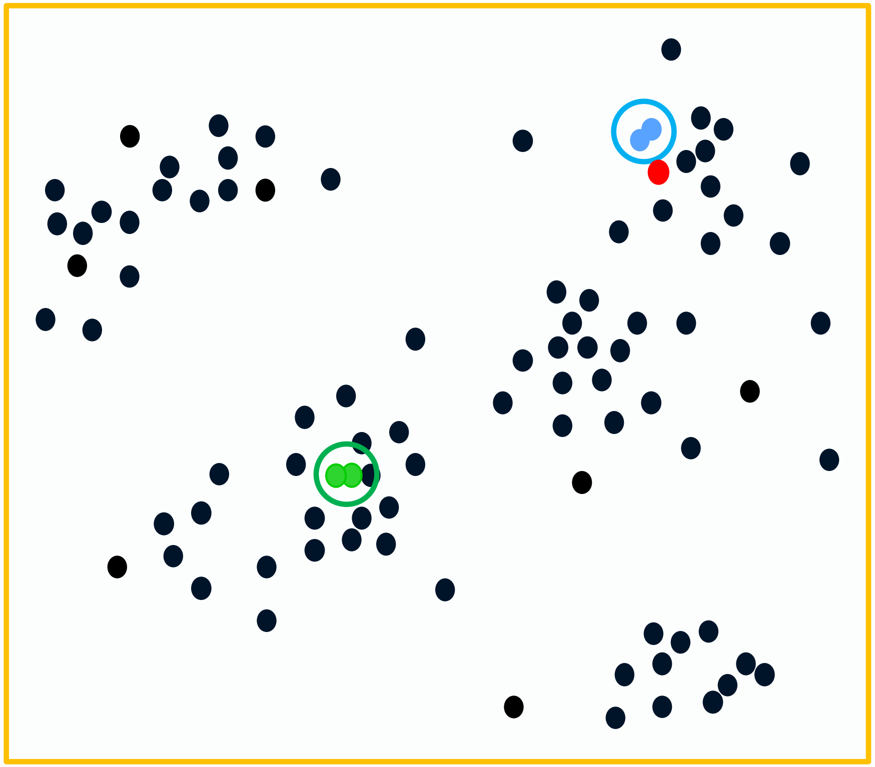


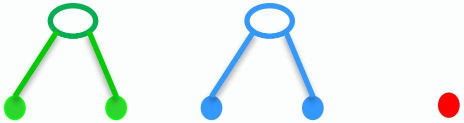


Hierarchical clustering

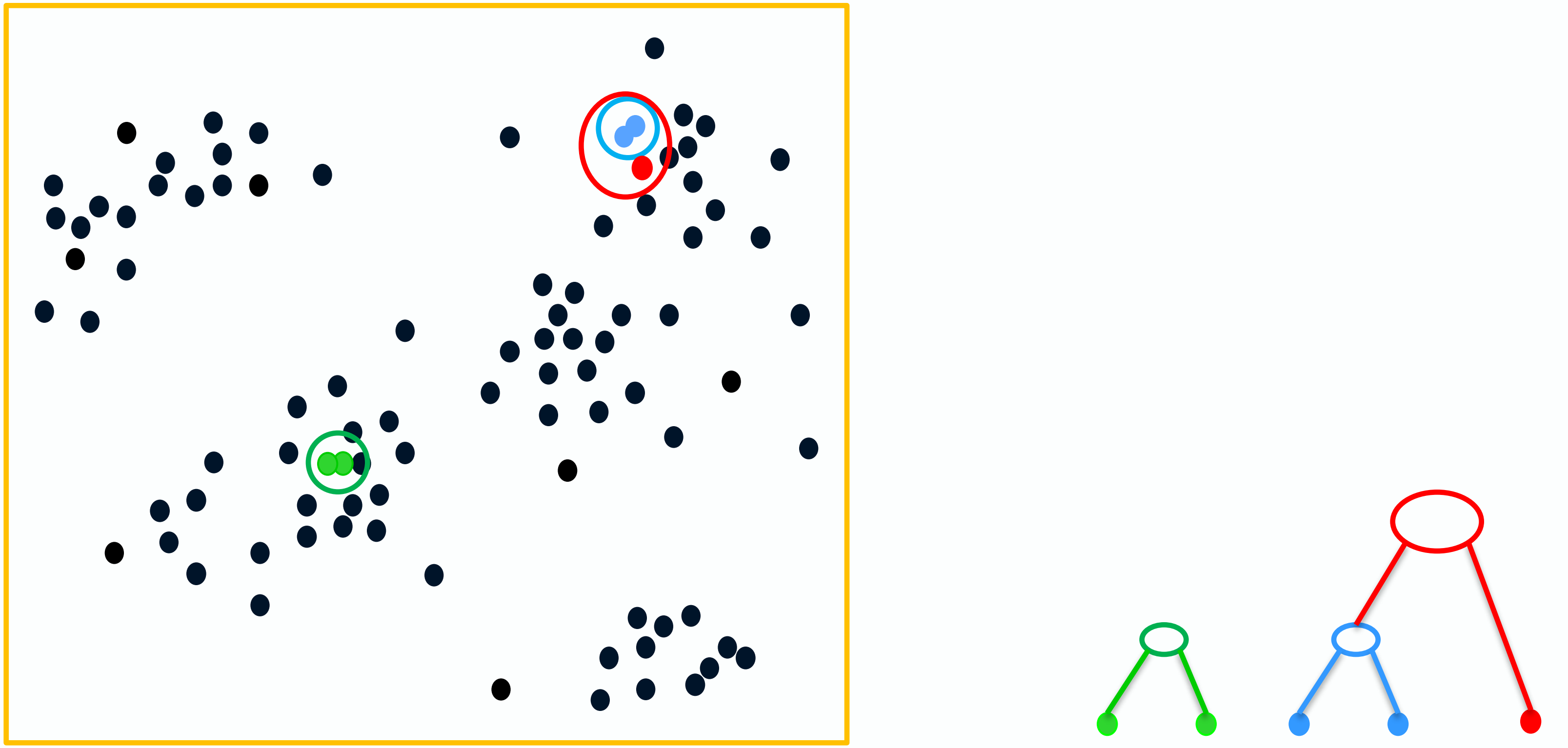


Hierarchical clustering

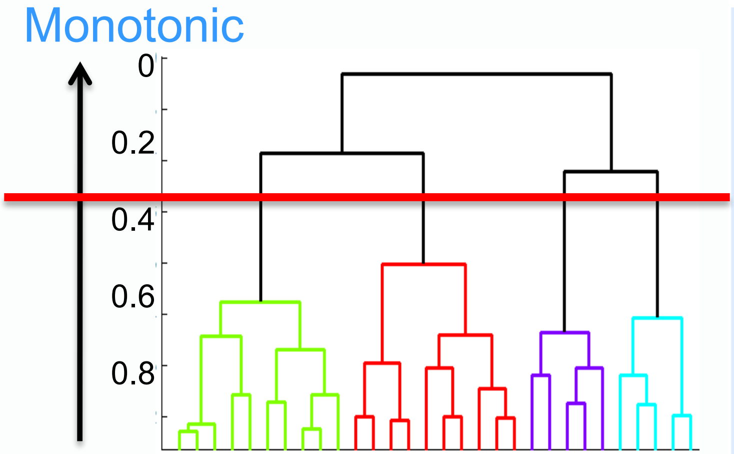




Hierarchical clustering



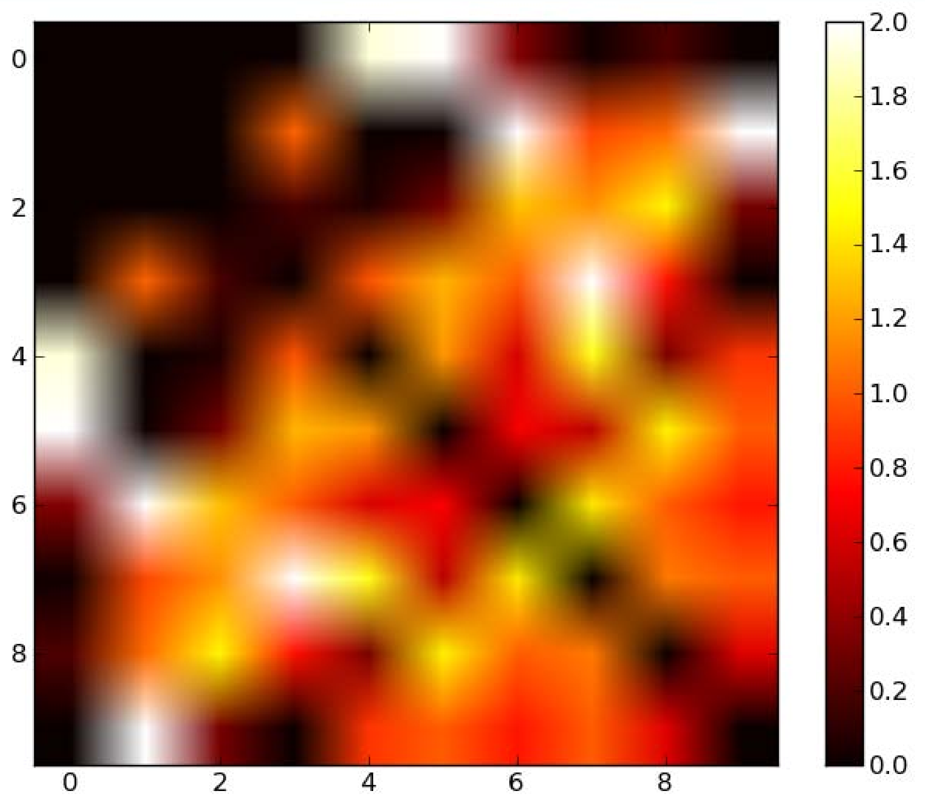
Hierarchical agglomerative clustering

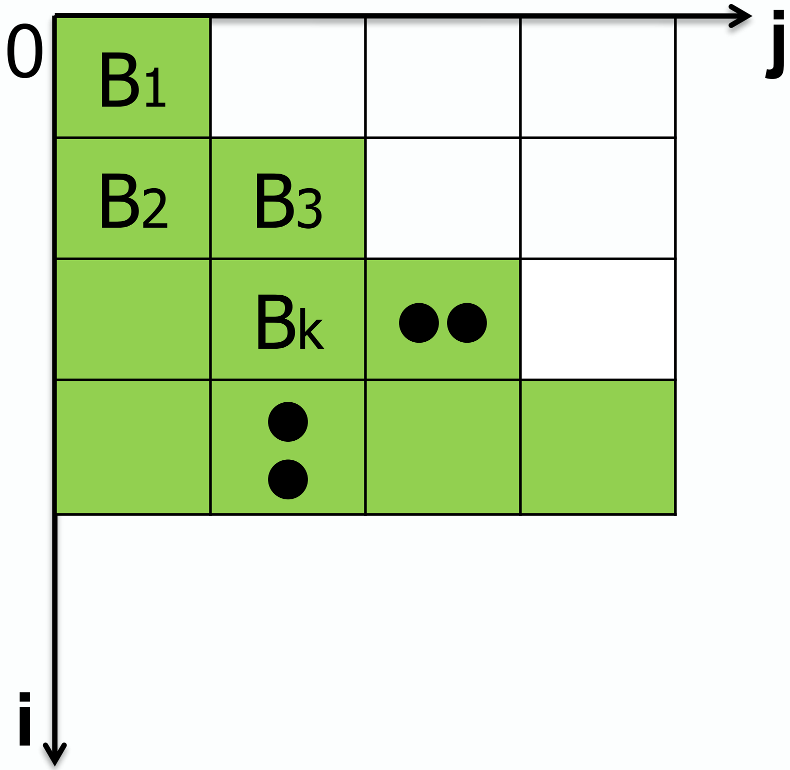


Method: **Merge the nearest clusters until a single cluster is left**

```
Procedure HAC (N points, stop criterion)
{
(1) Initialize n points as n cluster centers;
(2) Iterate over centers until stop criterion is satisfied:
        a. Compute pair-wise similarity between any two centers $sim(c_i, c_j)$
        b. Find the nearest pair of centers
        c. Merge the two centers
            $< i, j > \leftarrow \arg\max_{i,j} sim(c_i, c_j)$
(3) Output the hierarchical clusters.
}
```

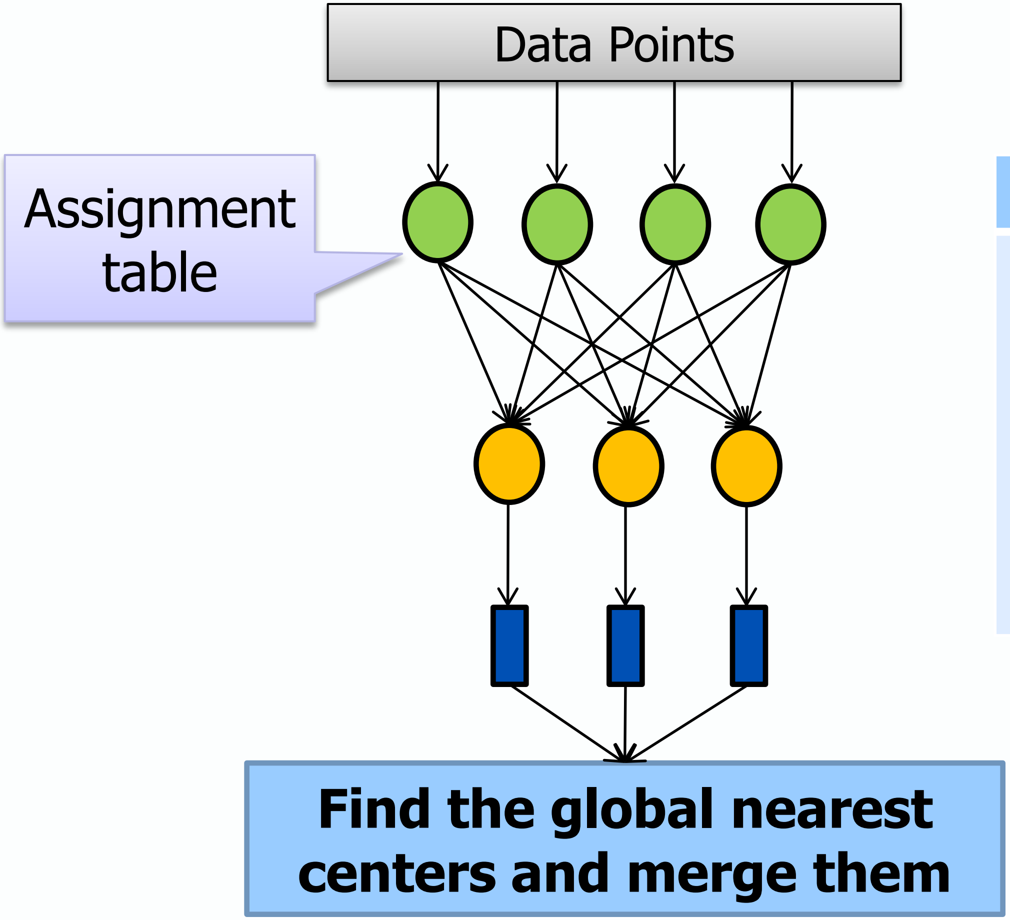
Pair-wise distances





Symmetric similarity matrix

Iterative map/reduce



Iterative map/reduce

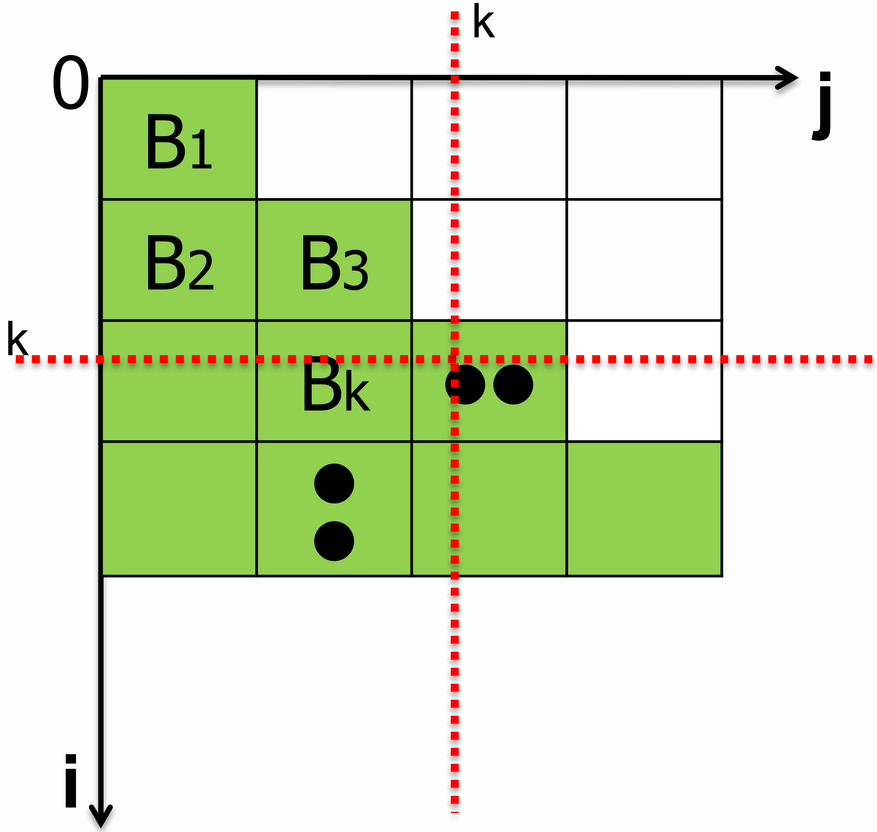
Mapper:

assign Block IDs to each data point.

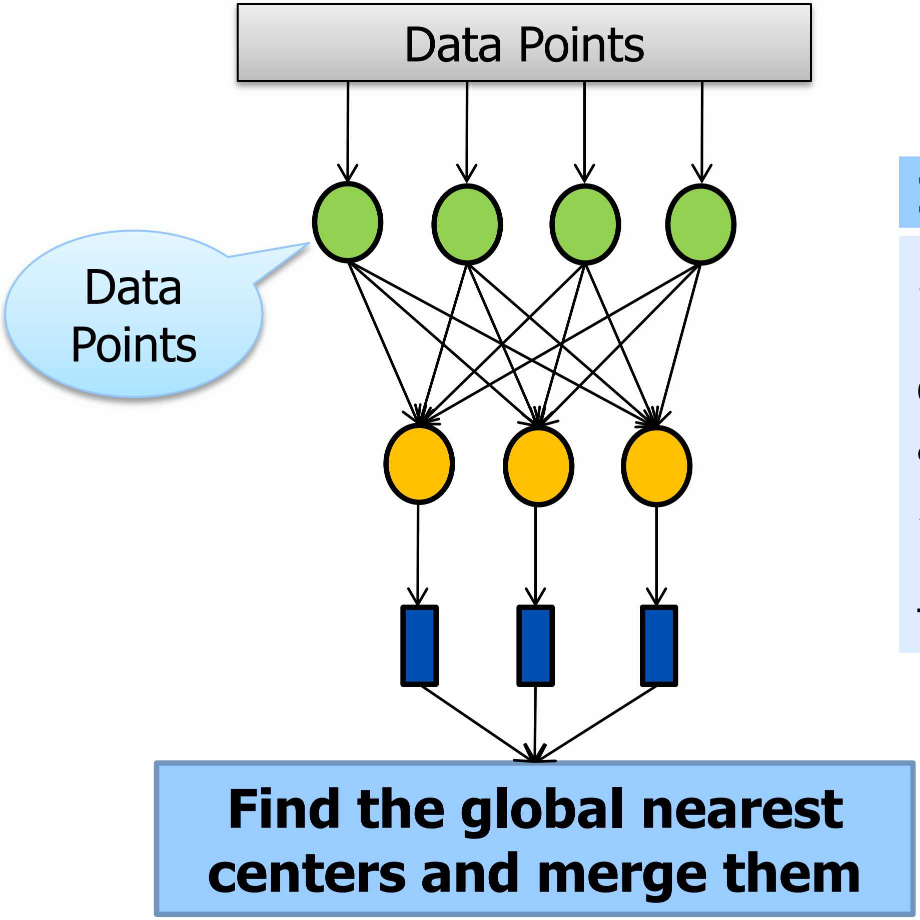
Reducer:

each reducer is responsible for find the local nearest pair of centers.

Assignment table



Iterative map/reduce


Data Points

Data Points

Find the global nearest centers and merge them

Iterative map/reduce

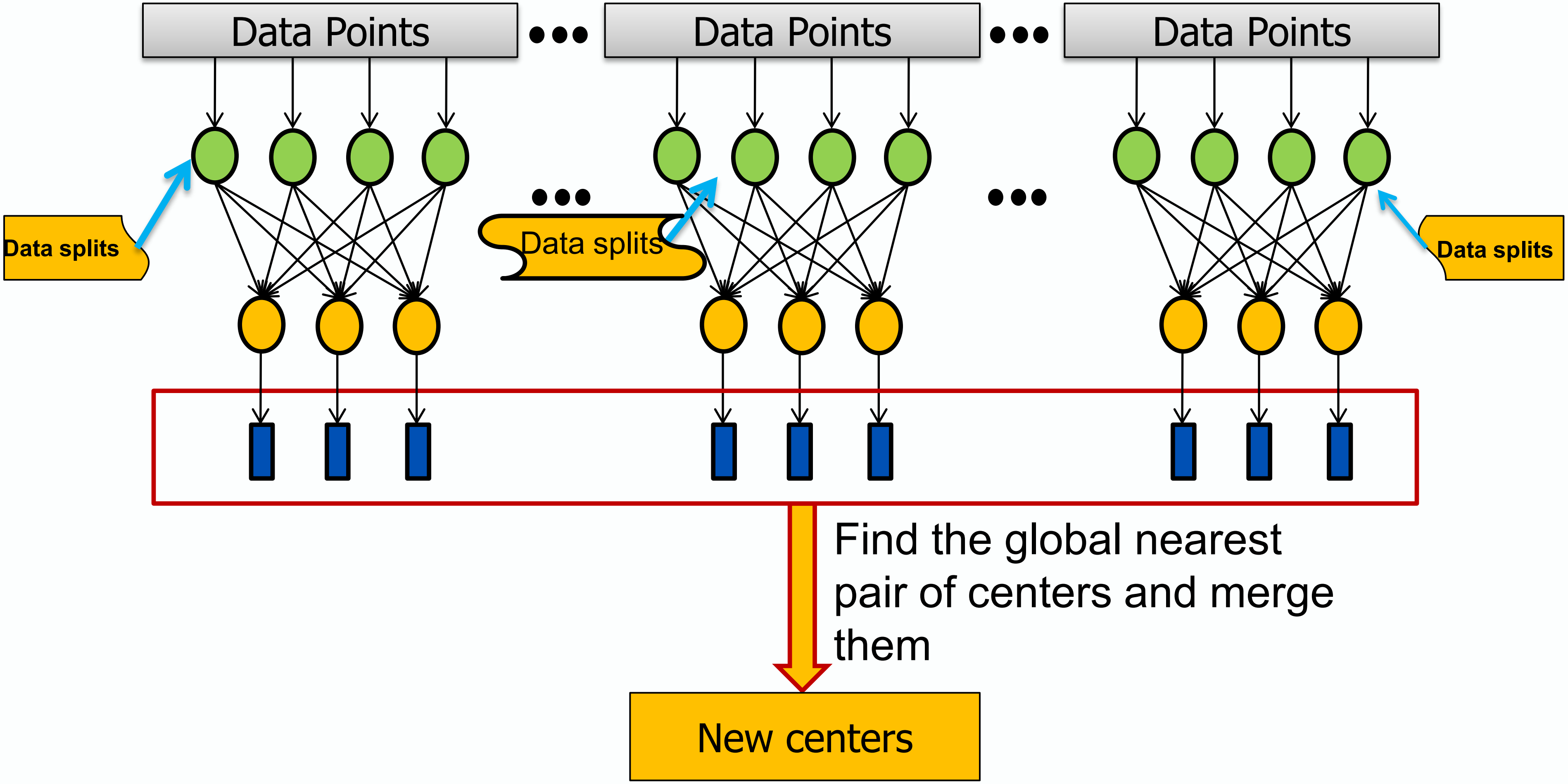
Mapper:

preload centers in memory and compute distances from each center to all the centers coming into mappers.

Reducer:

each reducer is responsible for find the local nearest pair of centers.

Block map/reduce



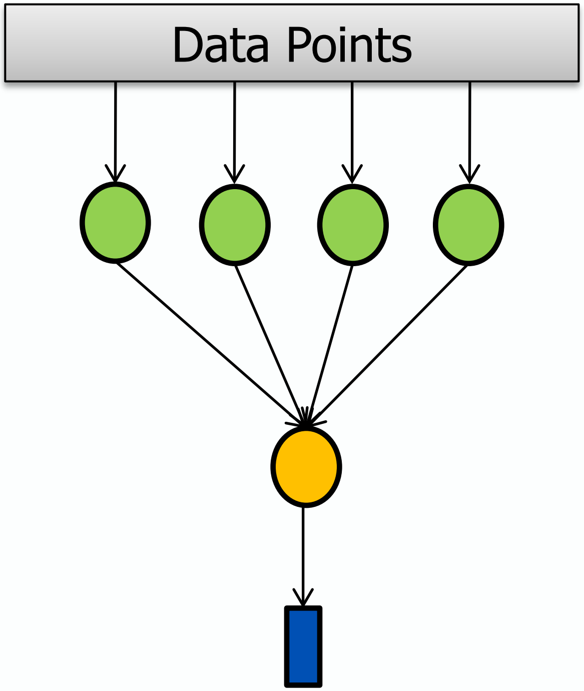
Partition

- 100,000,000 users have to be partitioned to blocks before clustering.
- User profile helps to partition users into overlapping blocks.
- There are millions of blocks and each block contains several thousands of users on average.

Speed-up

- For small blocks, iterative map/reduce is not efficient for the overhead of start and end of a job.
- Only few of elements of the similarity matrix need to be updated.

One-off map/reduce



One-off map/reduce

Mapper:

Passing friend list to reducers.

Reducer:

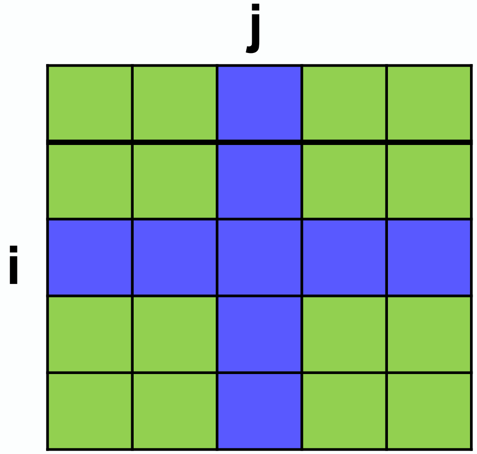
Agglomeration until clustering stops and output clustering results.

Scalability

- Only suitable for m*n matrix where m,n < 10^4

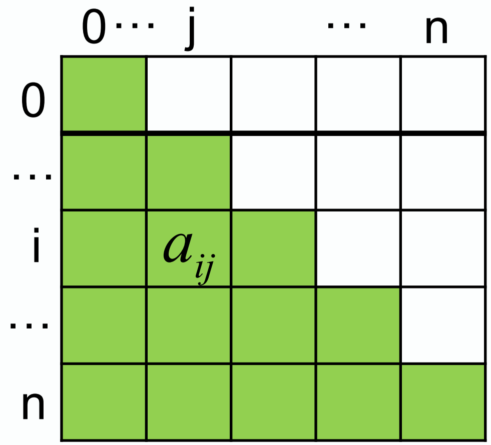
Distance caching

- Avoid re-calculating pair-wise distance between centers not for agglomeration from iteration to the next.

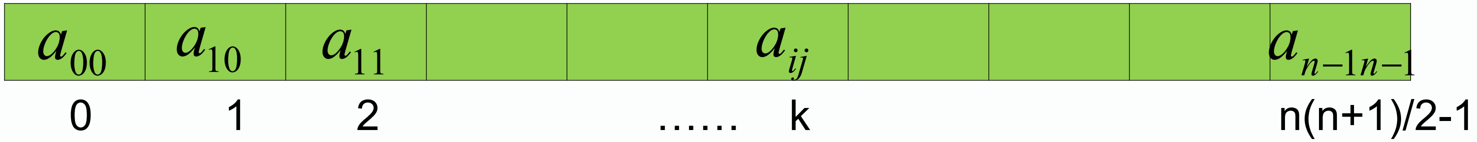


Compressed storage for lower triangle

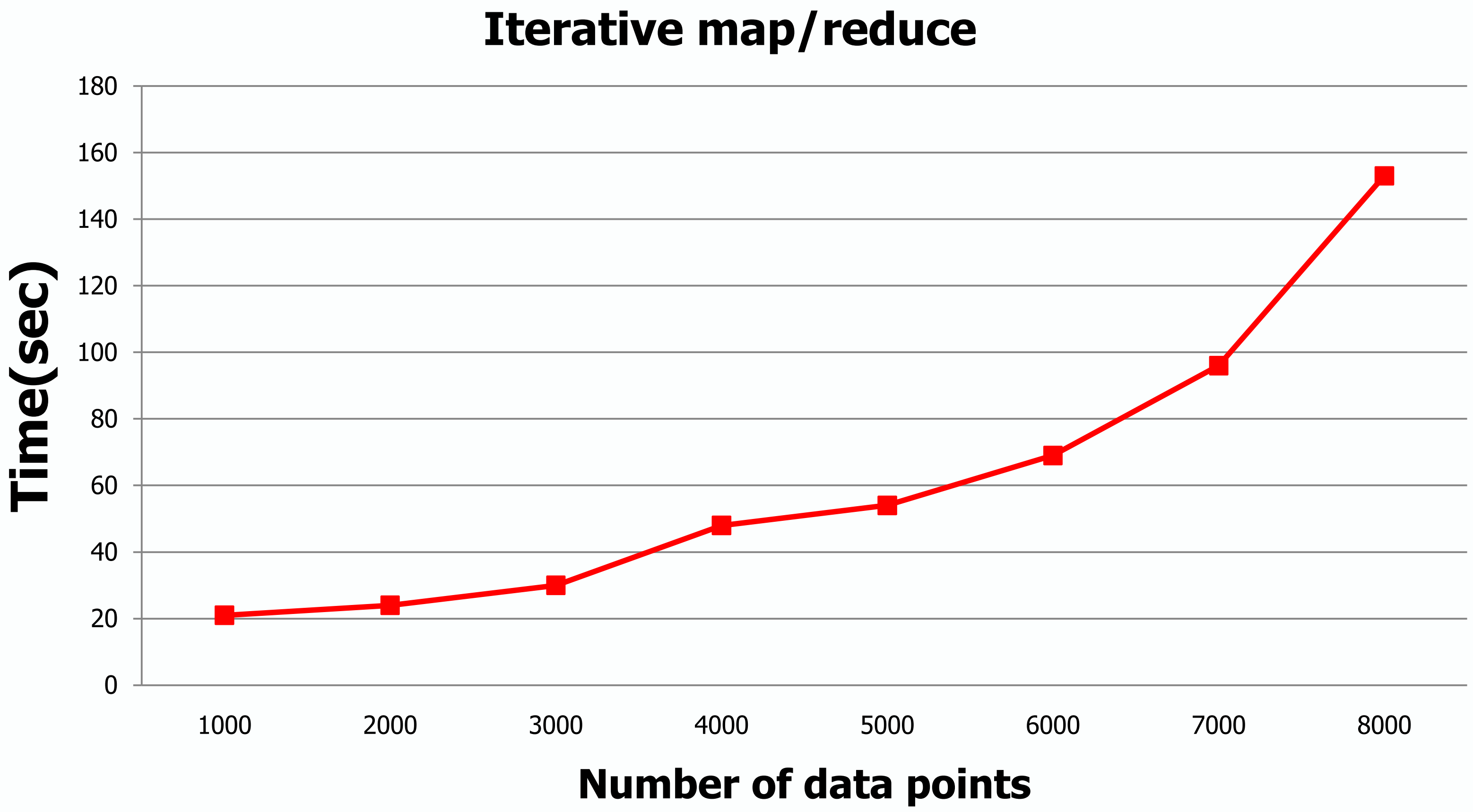
- Choose a row compressed storage mode to keep pair-wise distance between centers in memory.



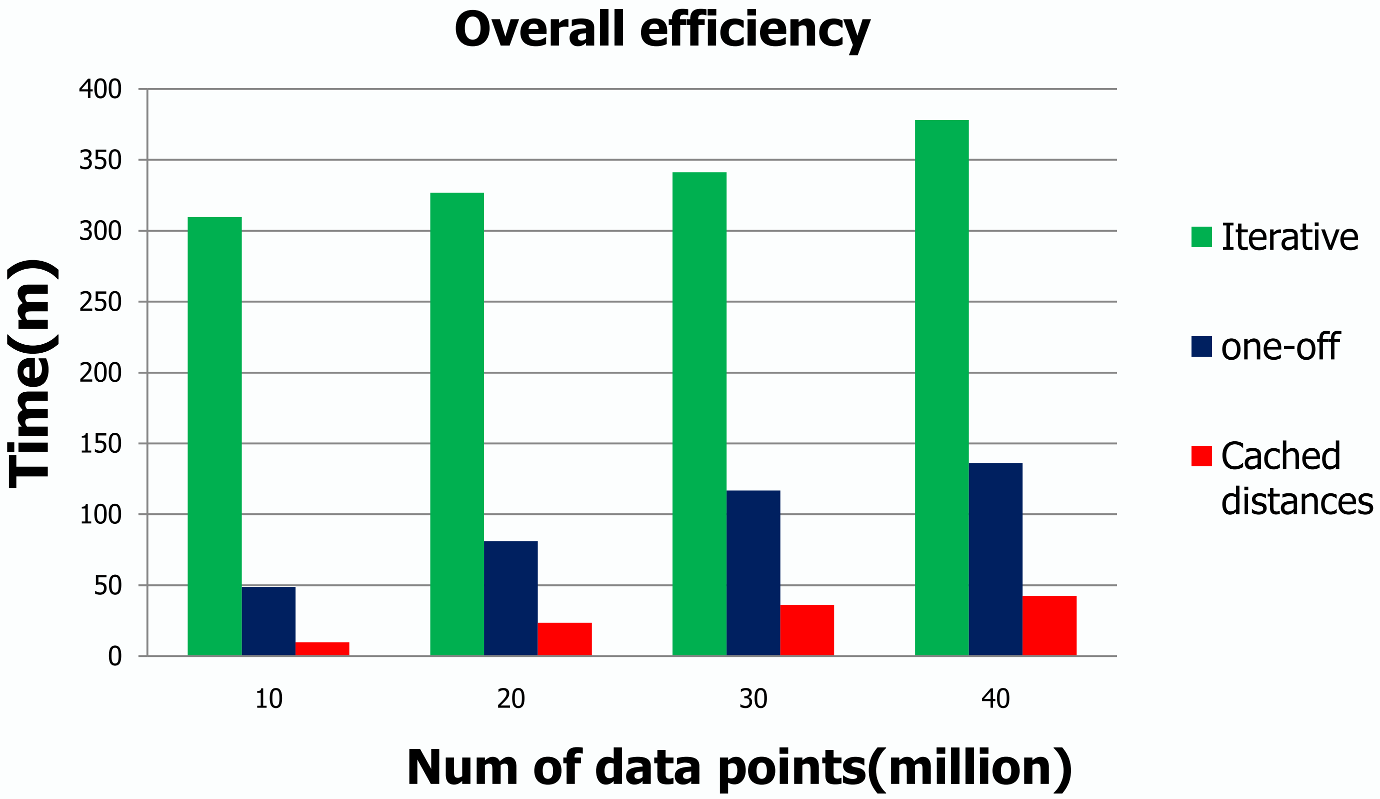
K=(i+1)*i/2+j



Performance



Performance

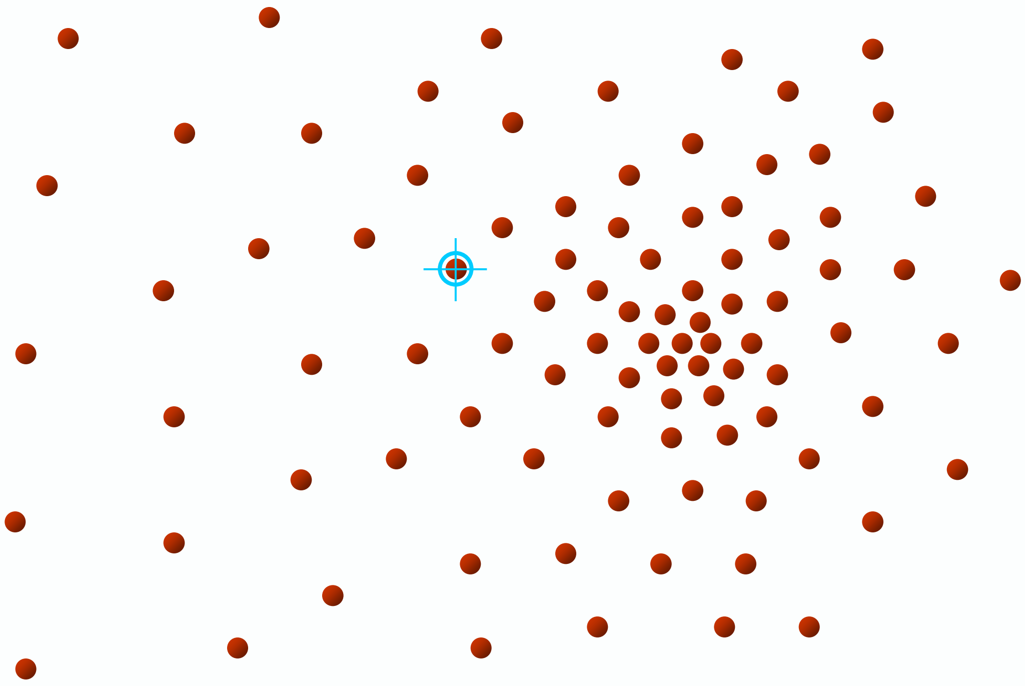


Case II: Topic detection by Mean-shift clustering

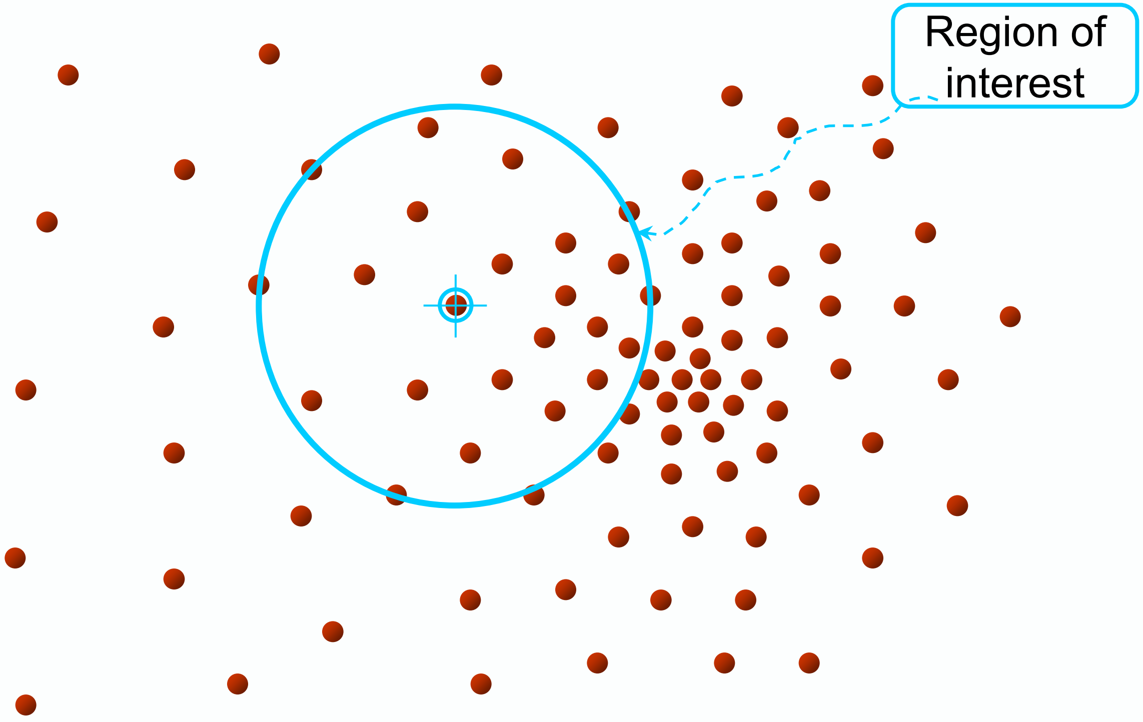
- Clustering to find topics in news feed
  - High density indicates hot news.
  - Without knowing the number of topics.

澳大利亚
清华
世界
技术
小王子
唐诗
汇编
生活
家族
颓废
网络
考试
微积分
英语
数学
编程
艺术
物理
忧伤
上海
java
思想
人文
诺贝尔奖
时间
少年
导演
穿越
单词
中国人

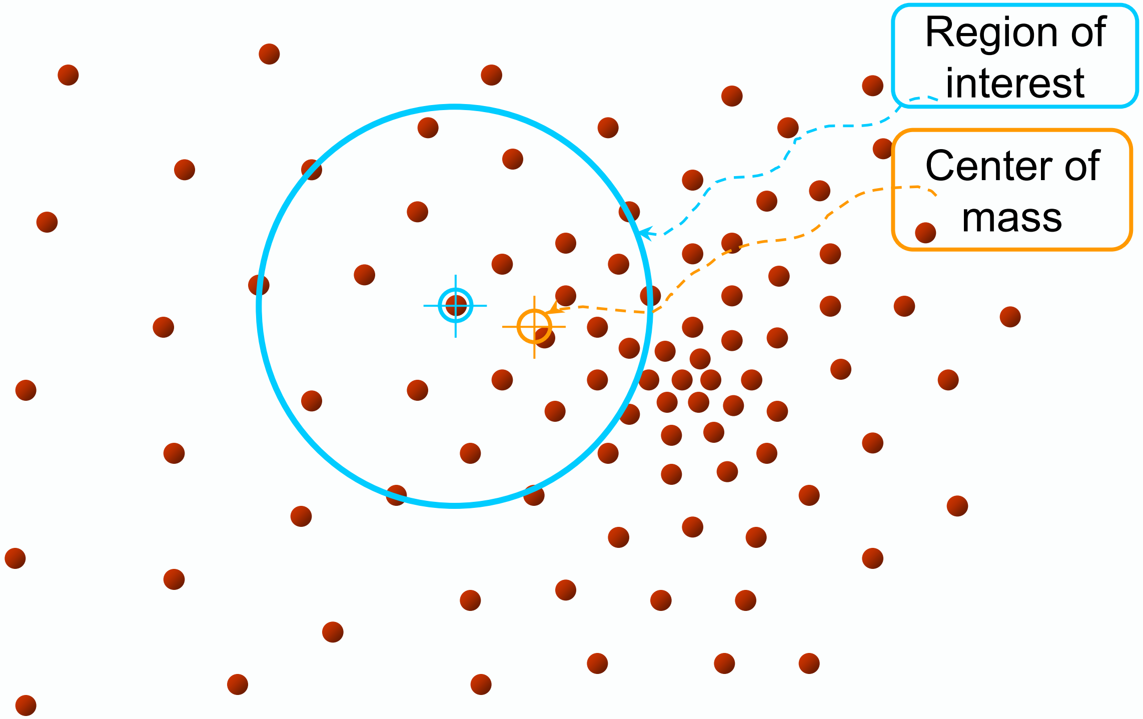
Mean shift



Mean shift



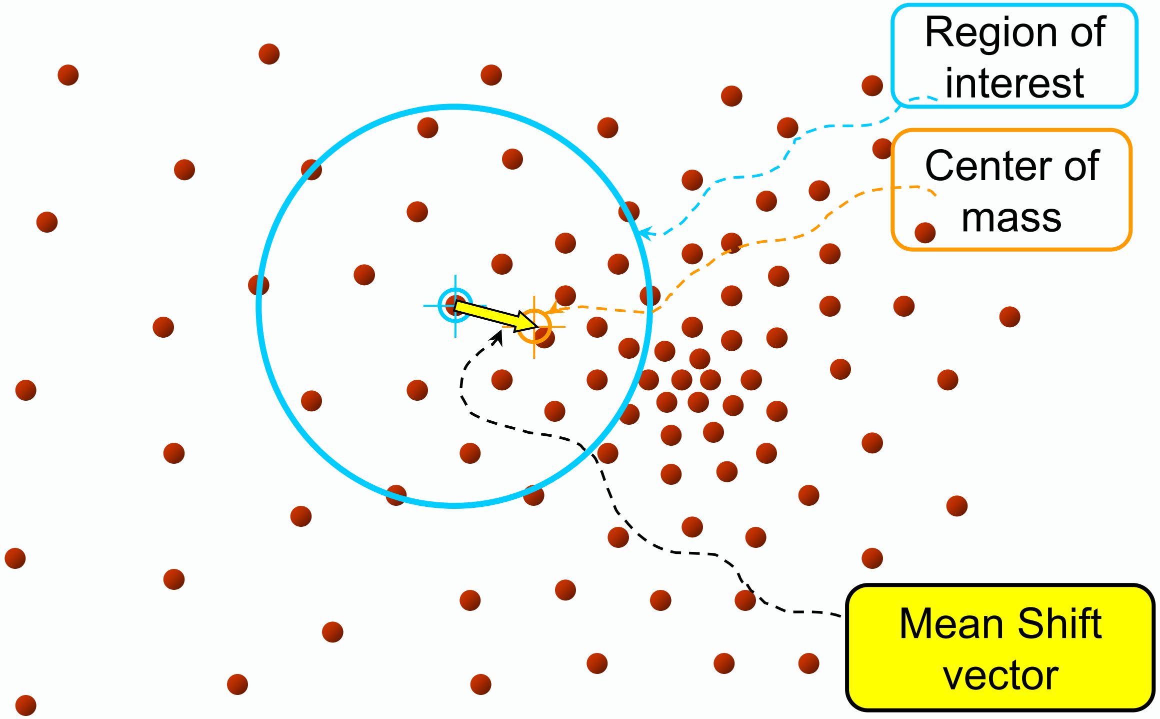
Mean shift



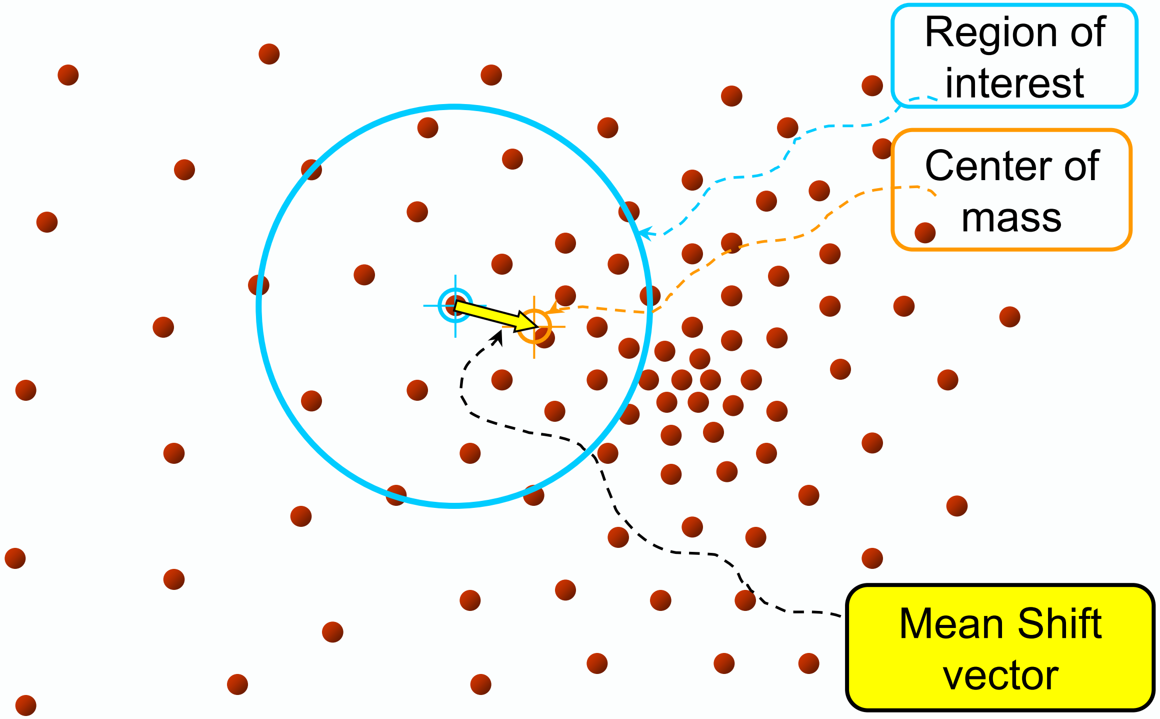
Region of interest

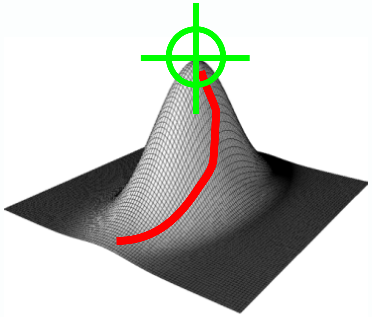
Center of mass

Mean shift

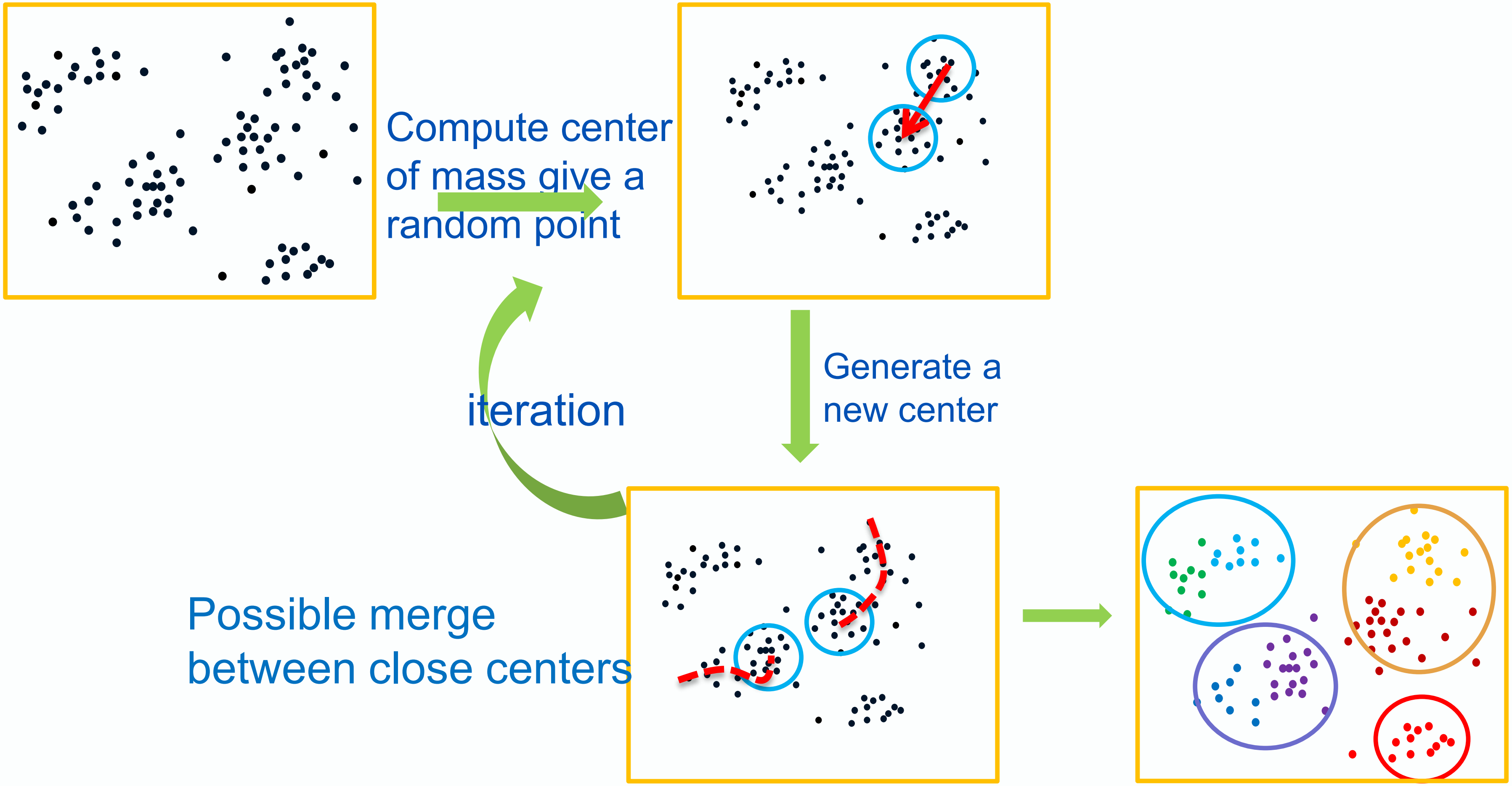


Mean shift





Mean shift clustering



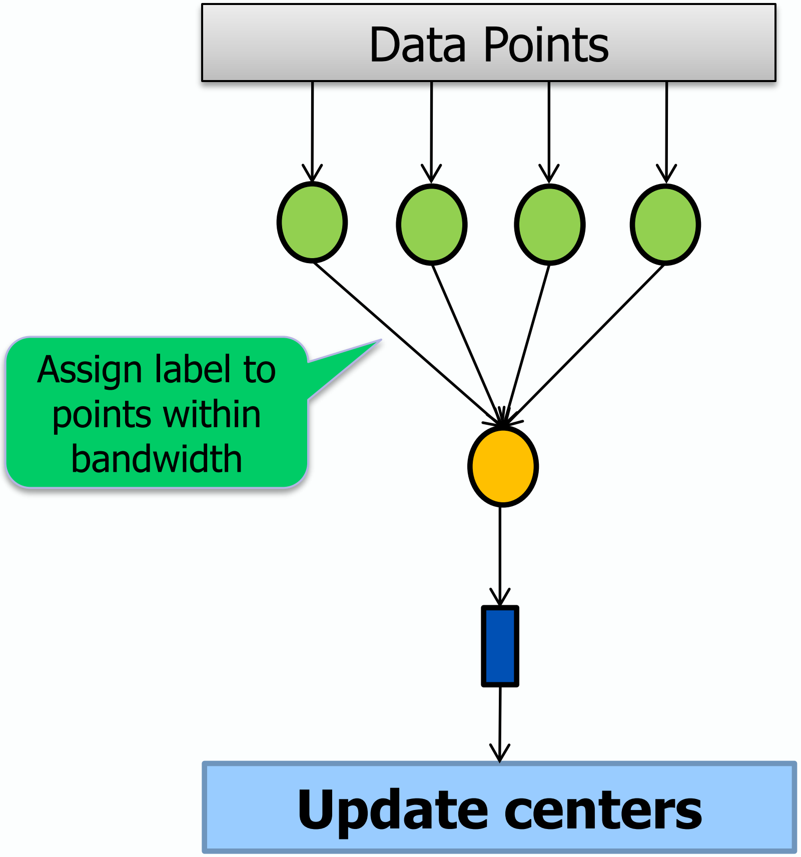
Compute center of mass give a random point

Generate a new center

iteration

Possible merge between close centers

Iterative map/reduce



Iterative map/reduce

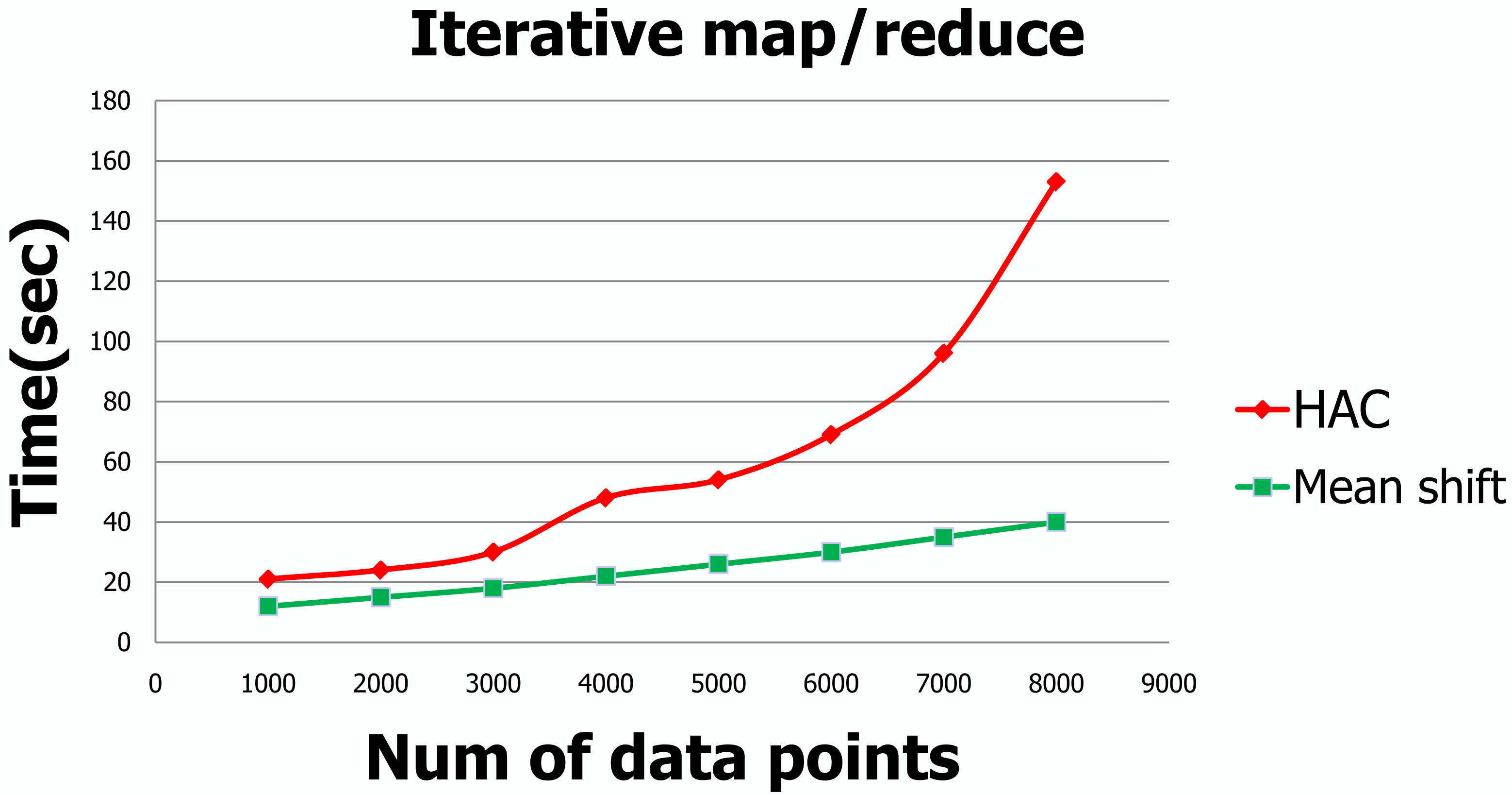
Mapper:

select a point as the center and compute distances from the center to all the data points, and assign a label to the points within the bandwidth.

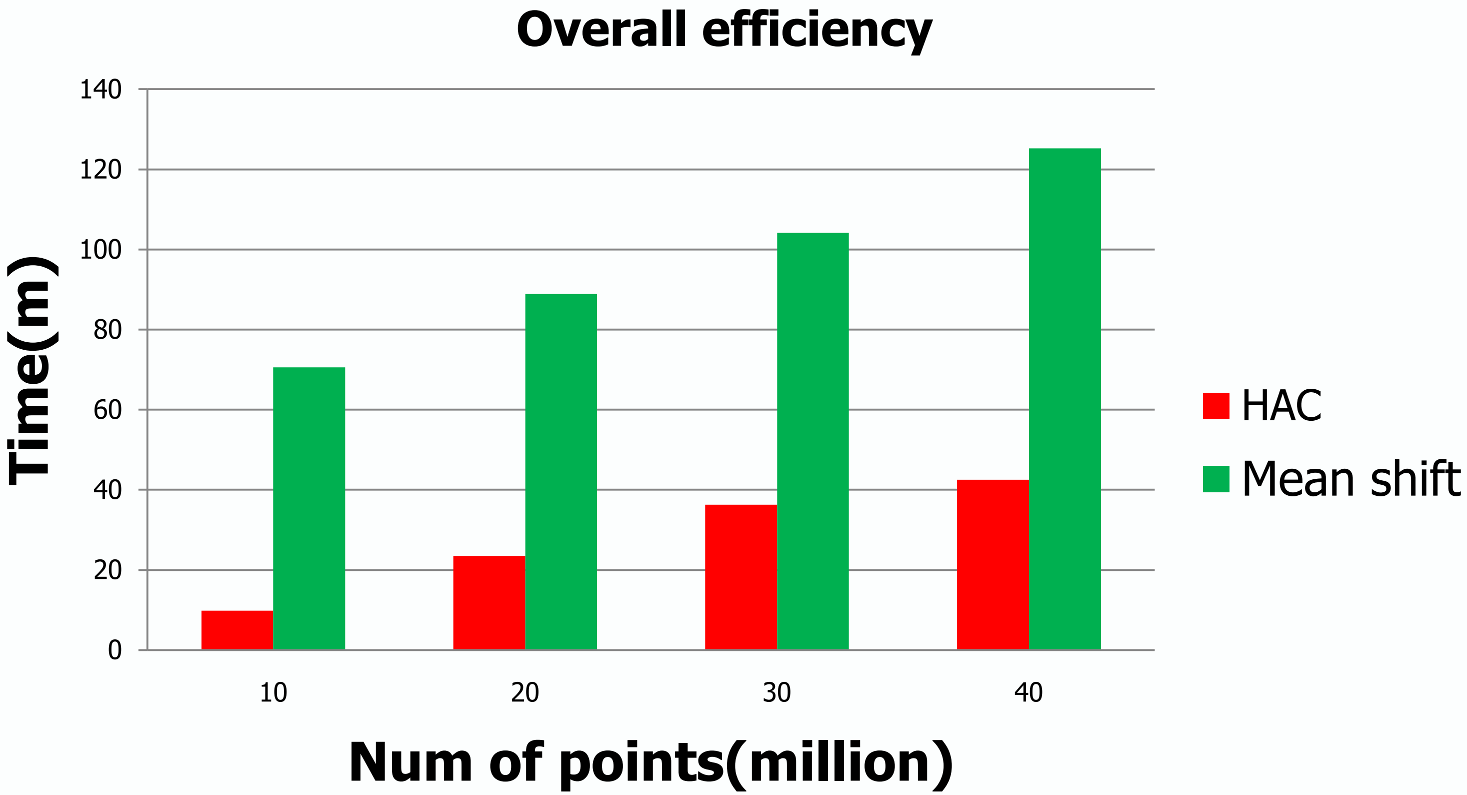
Reducer:

collect data points of the same label and compute the center of mass of them.

Performance



Performance



人人数据平台

白伯纯
bochun.bai@renren-inc.com
http://renren.com/bbc

张叶银
yeyin.zhang@renren-inc.com
http://renren.com/hartebeest