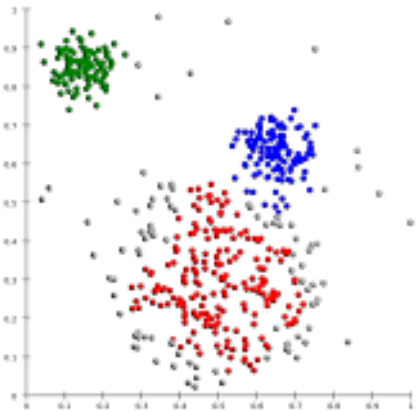
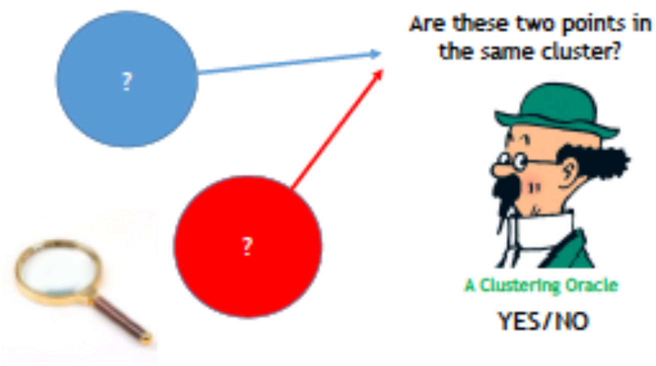
Interactive Clustering

Barna Saha

Clustering



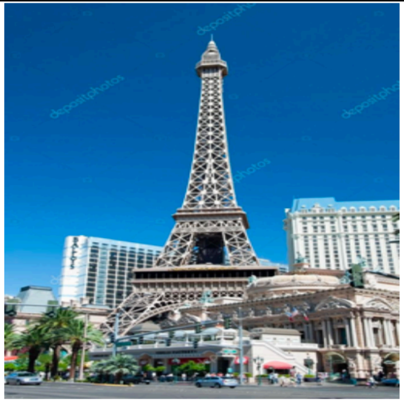


- n points k clusters
- Clustering Oracle Knows the ground truth clustering or can solve the clustering under some optimization rule
- A Query to a Clustering Oracle: Are points A and B in the same cluster?

Query Complexity?

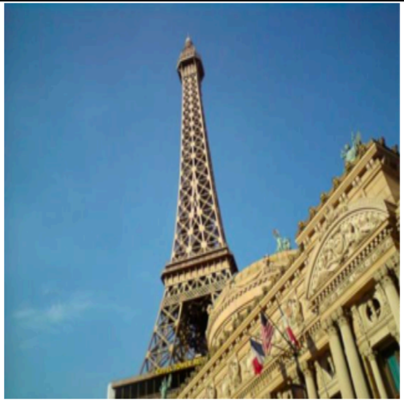
Learning over Noisy Data

Learn a classifier or find clusters over noisy/uncertain data



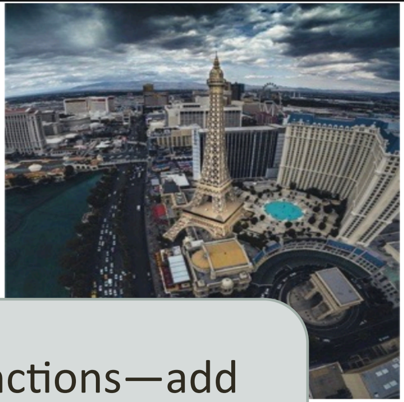
Type-1

Type-1



Type-1

Type-1



Noise comes from using similarity functions—add an edge between two images if they represent the same monument—clusters could be erroneous

Learning over Noisy Data

Learn a classifier or find clusters over noisy/uncertain data

Noise comes from inherent data errors/missing attributes—clustering collaboration network obtained from DBLP could be erroneous.

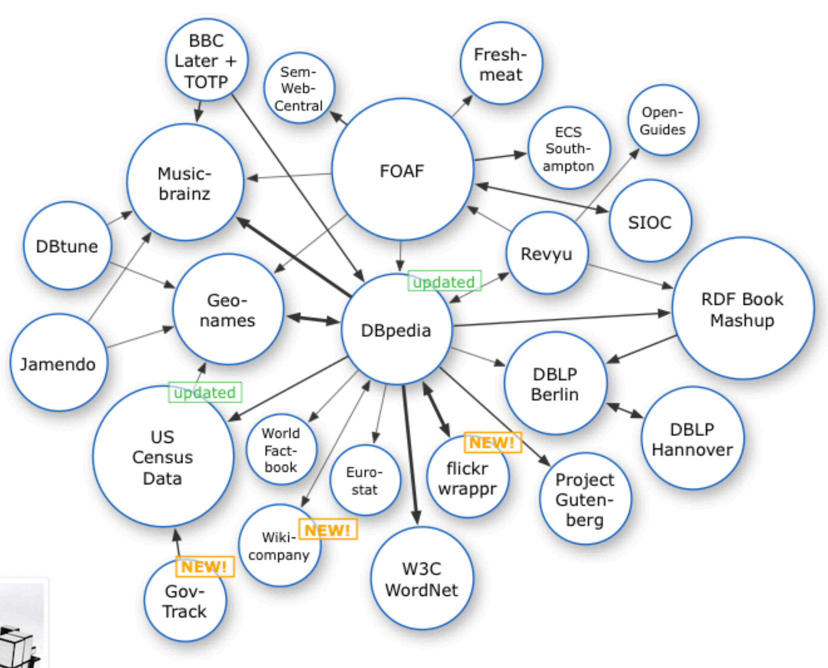
[–] **2010 – today**

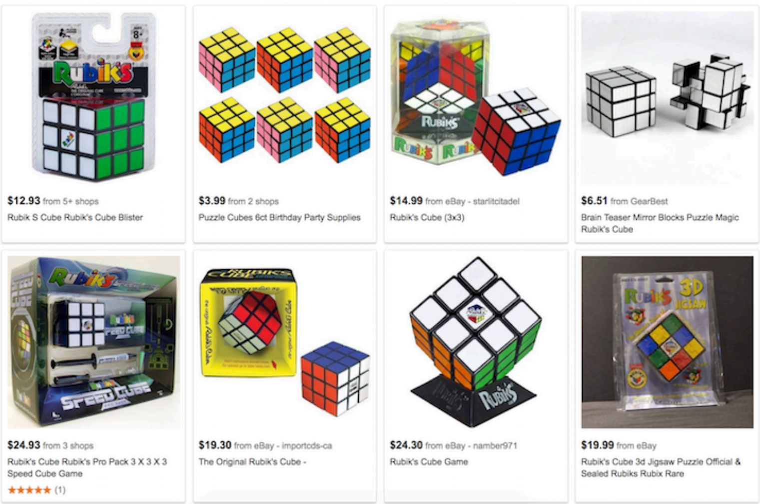
2014

- [j221] Xiao Xia, Xiaodong Wang, Jian Li, Xingming Zhou: **Multi-objective mobile app recommendation: A system-level collaboration approach.** Computers & Electrical Engineering 40(1): 203-215 (2014)
- [j220] Qingjia Huang, Kai Shuang, Peng Xu, Jian Li, Xu Liu, Sen Su: **Prediction-based Dynamic Resource Scheduling for Virtualized Cloud Systems.** JNW 9(2): 375-383 (2014)
- [j219] Jingdong Wang, Naiyan Wang, You Jia, Jian Li, Gang Zeng, Hongbin Zha, Xian-Sheng Hua: **Trinary-Projection Trees for Approximate Nearest Neighbor Search.** IEEE Trans. Pattern Anal. Mach. Intell. 36(2): 388-403 (2014)
- [j218] George-Othon Glentis, Kexin Zhao, Andreas Jakobsson, Habti Abeida, Jian Li: **SAR imaging via efficient implementations of sparse ML approaches.** Signal Processing 95: 15-26 (2014)
- [c258] MohammadTaghi Hajiaghayi, Wei Hu, Jian Li, Shi Li, Barna Saha: **A Constant Factor Approximation Algorithm for Fault-Tolerant *k*-Median.** SODA 2014: 1-12

Further Applications

- Linking Census Records
- Public Health
- Web search
- Comparison shopping
- Spam Detection
- Machine Reading
- IP Aliasing
- ....





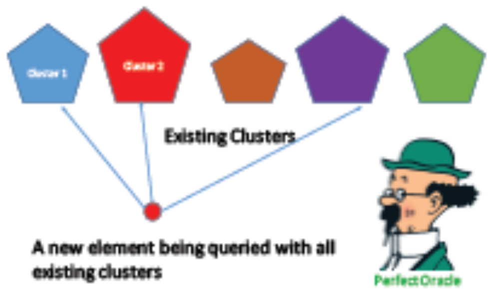
Query complexity of optimal strategy?

Clustering: n points k clusters

- Sufficient: nk
  - Compare any element with all the previously formed clusters
  - Any item needs to be queried at most k times before it is assigned to a cluster

Query complexity of optimal strategy?

Clustering: n points k clusters



- Sufficient: nk
  - Compare any element with all the previously formed clusters
  - Any item needs to be queried at most k times before it is assigned to a cluster
- Necessary: $\Omega(nk)$ Davidson, Khanna, Milo, Roy, 2014
  - Deterministic Algorithms: Needs to query $\Theta(n)$ points at least $k-1$ times
  - Randomized Algorithms (find the clustering exactly whp): Same lower bound applies

Faulty Oracle

- Each query answer can be independently wrong with probability p (when the points are in same cluster) or $1 - q$ (when in different clusters)

Faulty Oracle

- Each query answer can be independently wrong with probability p (when the points are in same cluster) or $1 - q$ (when in different clusters)

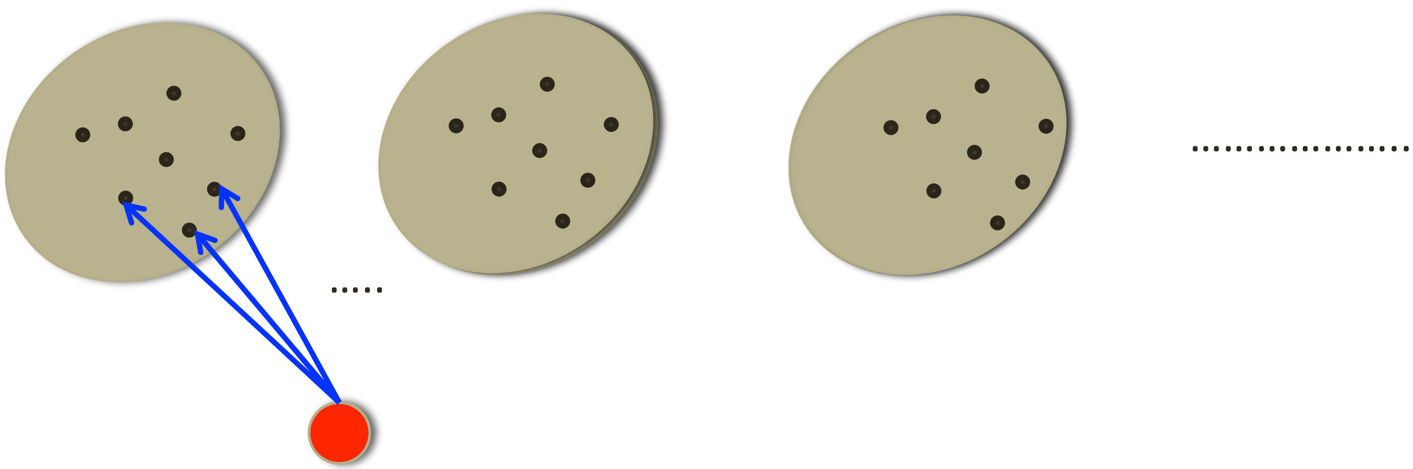
Repeat the same question. Assuming p=q, repeat each question (say) 24log n/$(1-2p)^2$ times

Faulty Oracle

- Each query answer can be independently wrong with probability p (when the points are in same cluster) or $1 - q$ (when in different clusters)
- Resampling is not allowed
  1) It is not theoretically interesting 2) Also not practical (only 20% reduction via resampling, Gruenheid et al. 2015, error increases upon aggregation Prelec et al. (Nature 2017))

Faulty Oracle: No Resampling

- Find seed nodes for each cluster
- If we can find $24 \log n/(1\text{-}2p)^2$ seed nodes from each cluster then we are done! [Why?]



Faulty Oracle: How to find seed nodes?

- Let $N=O(k^2 \log n/(1-2p)^4)$
- Select N nodes and ask all possible pairwise queries among these nodes.
- Run correlation clustering algorithm in this small set of nodes
- Each cluster returned by the correlation clustering that has size at least $24 \log n/(1-2p)^2$ act as a seed

Faulty Oracle: How to find seed nodes?

- Let $N=O(k^2 \log n/(1-2p)^4)$
- Select N nodes and ask all possible pairwise queries among these nodes.
- Run correlation clustering algorithm in this small set of nodes
- Each cluster returned by the correlation clustering that has size at least $24 \log n/(1-2p)^2$ act as a seed

Some intuition on the analysis: If we know all the query results, correlation clustering gives the maximum likelihood estimator.
Moreover, it is an instance of correlation clustering where errors are random—we know how to solve it!