生命科学中的
随机动力学模型

姚　远

2011.2.22

I. Course Website:

I. http://www.math.pku.edu.cn/teachers/yaoy/Spring2011/

II. Group Email:

I. yuanypku@googlegroups.com

II. My own email: yuany@math.pku.edu.cn

III. No final exam, yes final projects

I. Choose the topic interested, then work on it.

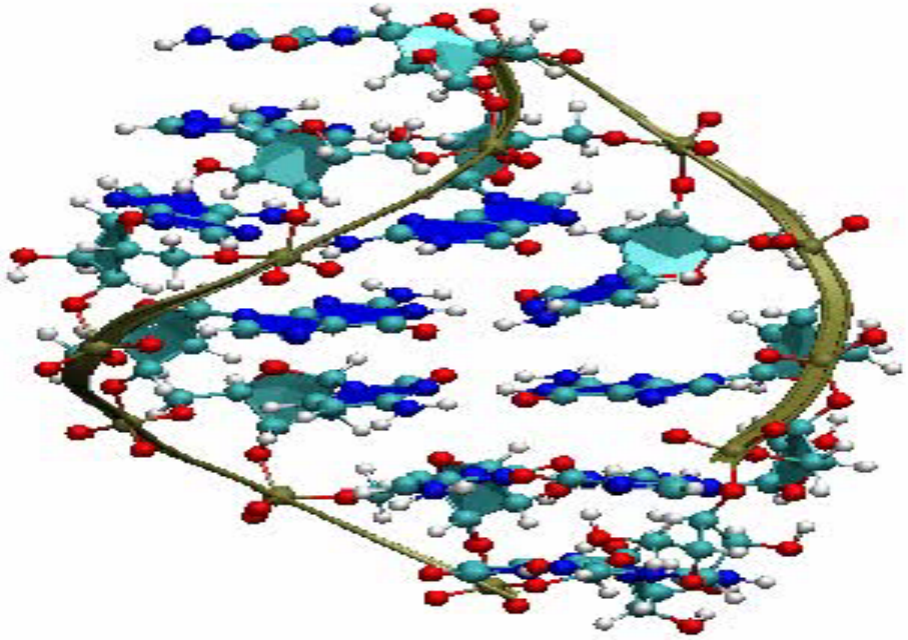
摘　　要

I. 高维动态数据分析
   I. Biomolecular folding
   II. Video (image sequence) analysis
   III. Dynamic (biological/social) networks

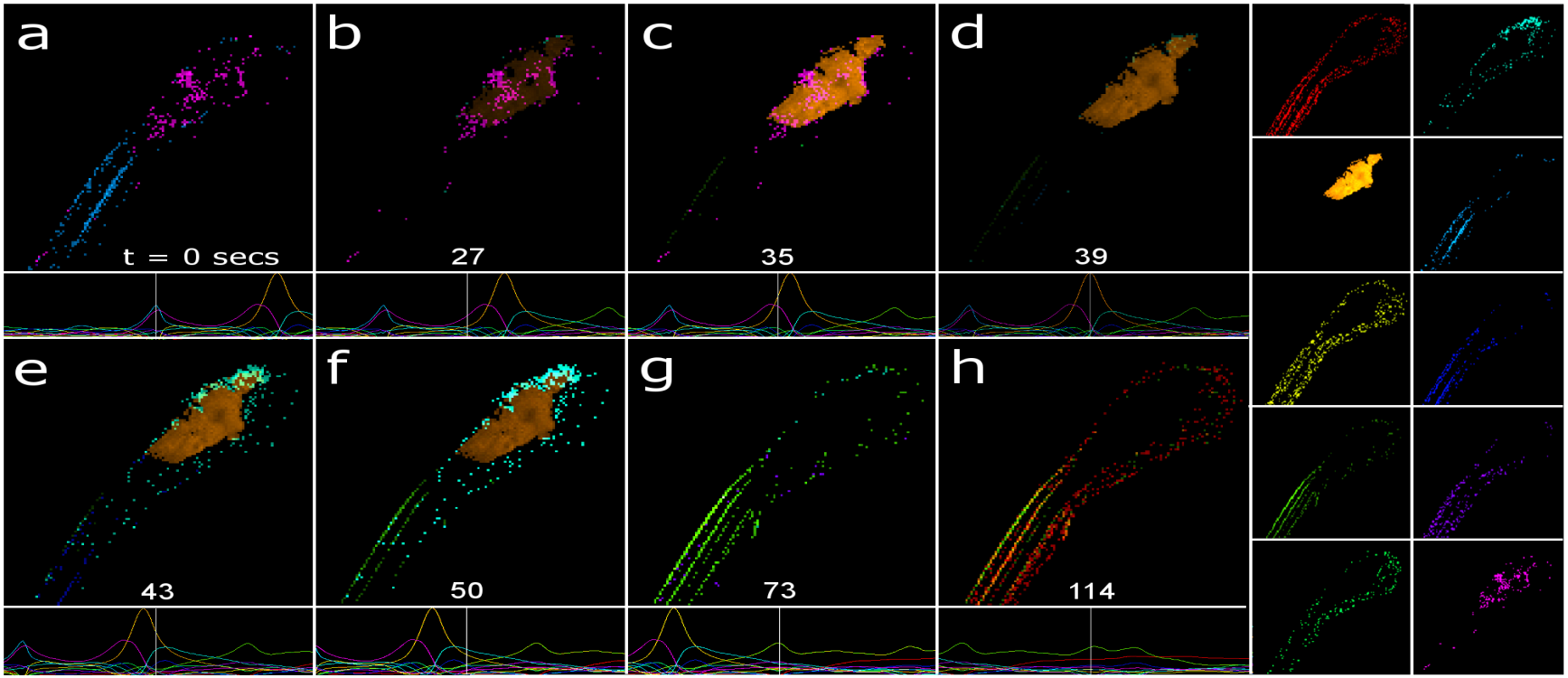
II. 随机动力学模型:
   I. Markov models
   II. Nonlinear diffusive models, etc.

III. 一个基本问题:
   I. How to reconstruct models given data generated (approximately) from such models?

Biomolecular Folding



Neuron Signaling in Zebrafish

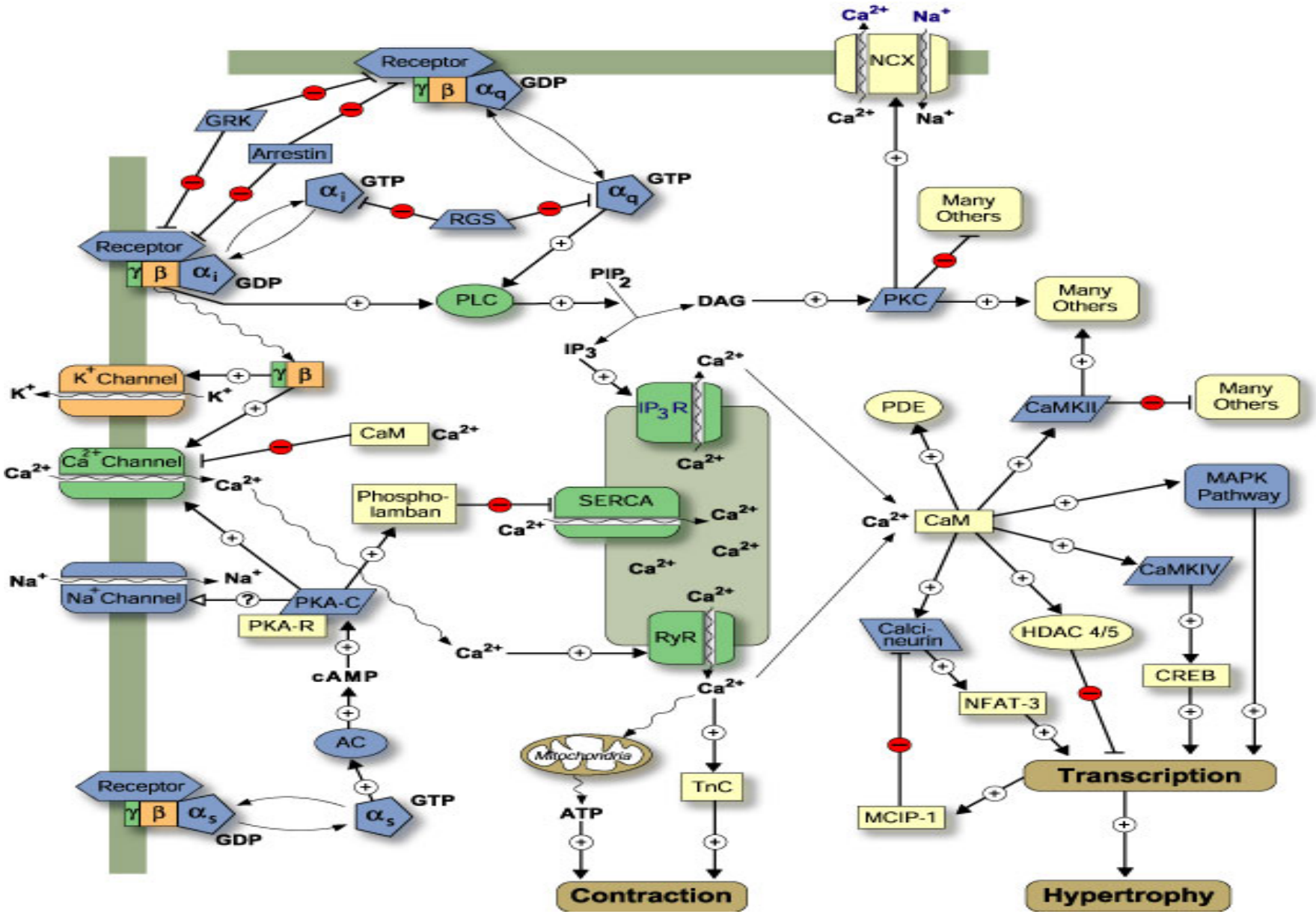


Courtesy of Le-Tian Tao

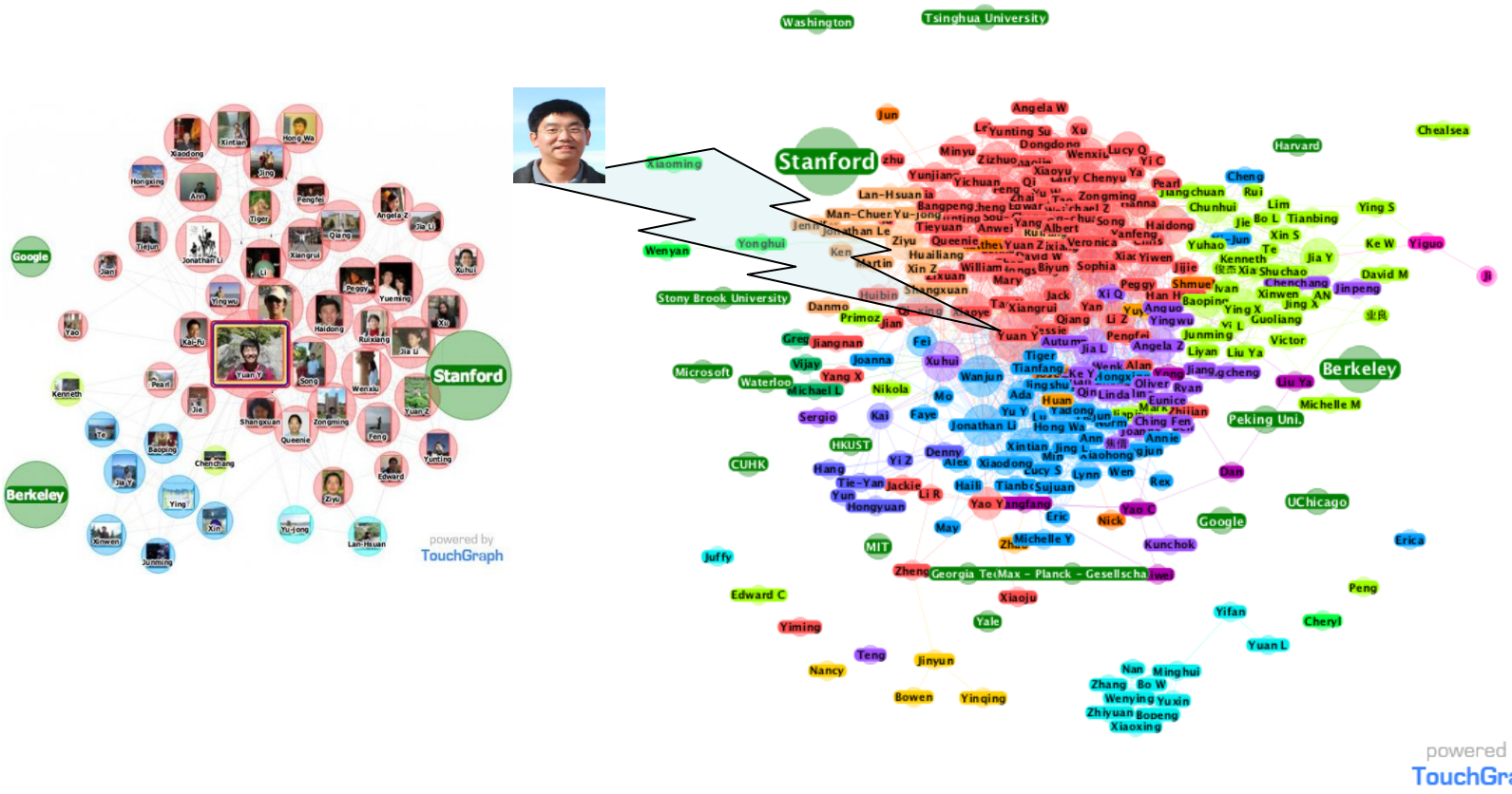
Surveillance Video

Courtesy of Yi Ma

Networks are Dynamic!



Networks are Dynamic!



My facebook TouchGraph from the year of 2009 to 2011

The following tools may be relavant...

- 降维和粗粒化方法，拓扑和几何方法
  - 非线性降维 nonlinear dimensionality reduction
  - 压缩感知 compressed sensing
  - 低秩矩阵分解 low-rank matrix factorization
  - 在线学习 online learning of dictionary
  - …
- 粗粒化动态系统近似
  - Markov models
  - …

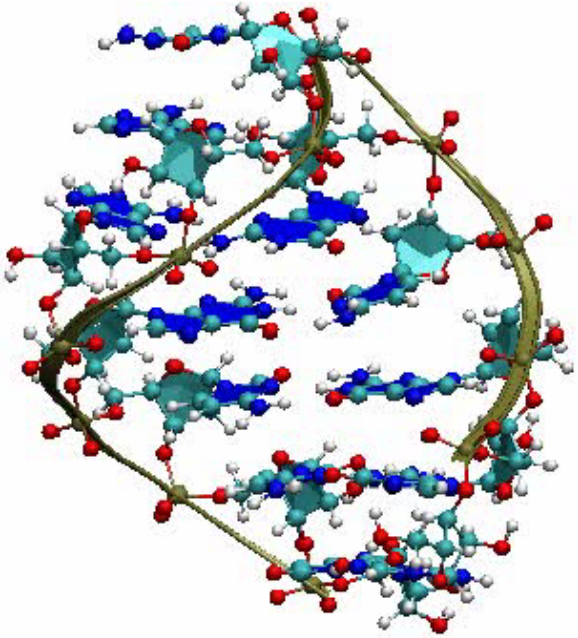
生物分子动力学

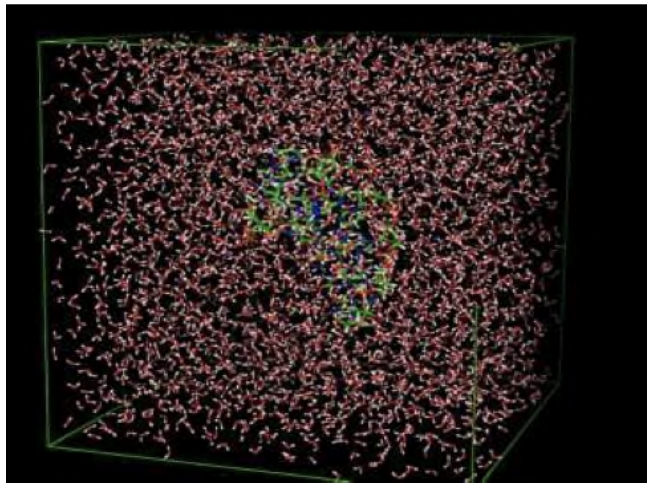
I. 分子动力学仿真数据特点

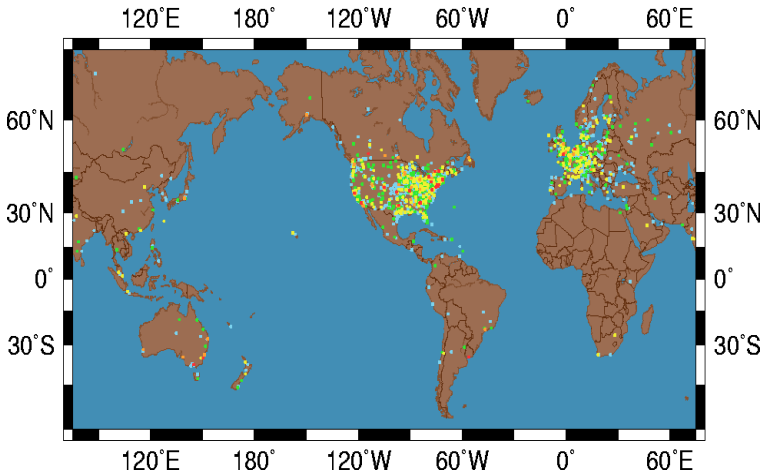
II. Clustering Analysis

III. Bayesian Inference of Markov Models

生物分子动力系统仿真



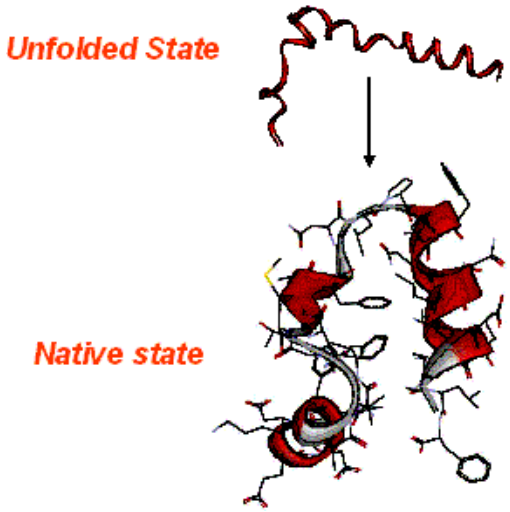




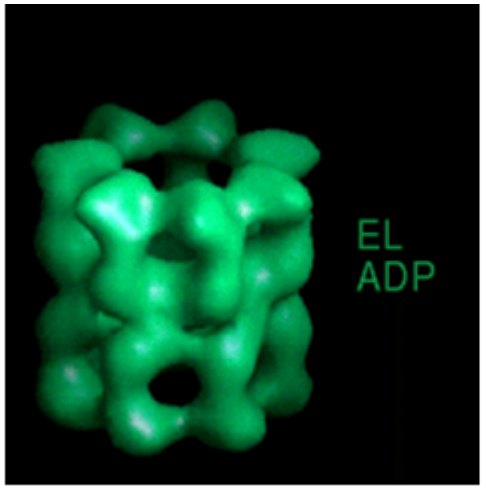
Folding@home distributed computing

Shirts and Pande, ***Science***, 290, 1903 , 2000

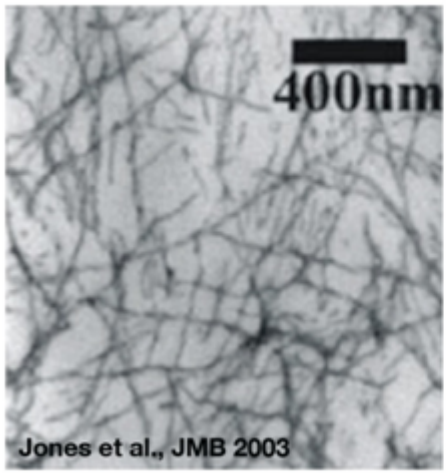
生物问题: Conformational Changes



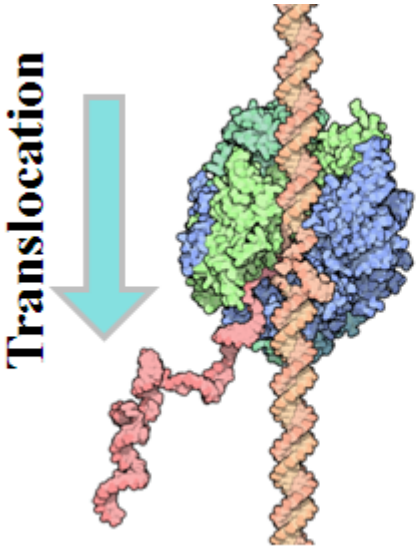
Protein Folding



Protein folding in Chaperone



Protein misfolding and aggregation

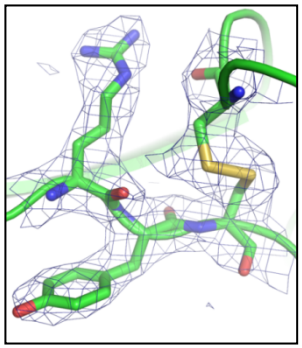


RNA Polymerase Translocation

Illustrations by David Goodsell

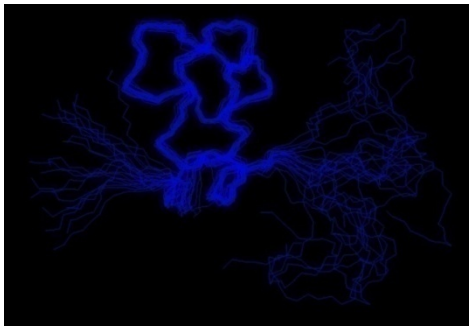
Understanding Conformational Changes at Atomic Resolution is Difficult Experimentally

Computer simulations may complement experiments!



X-ray

structures are static snapshots

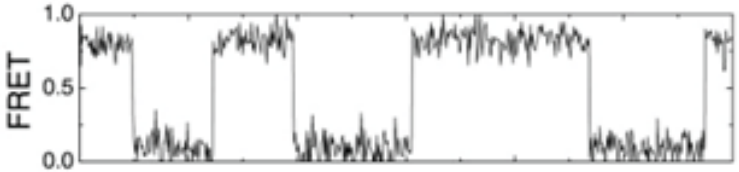


NMR

can provide dynamics, but difficult for large systems

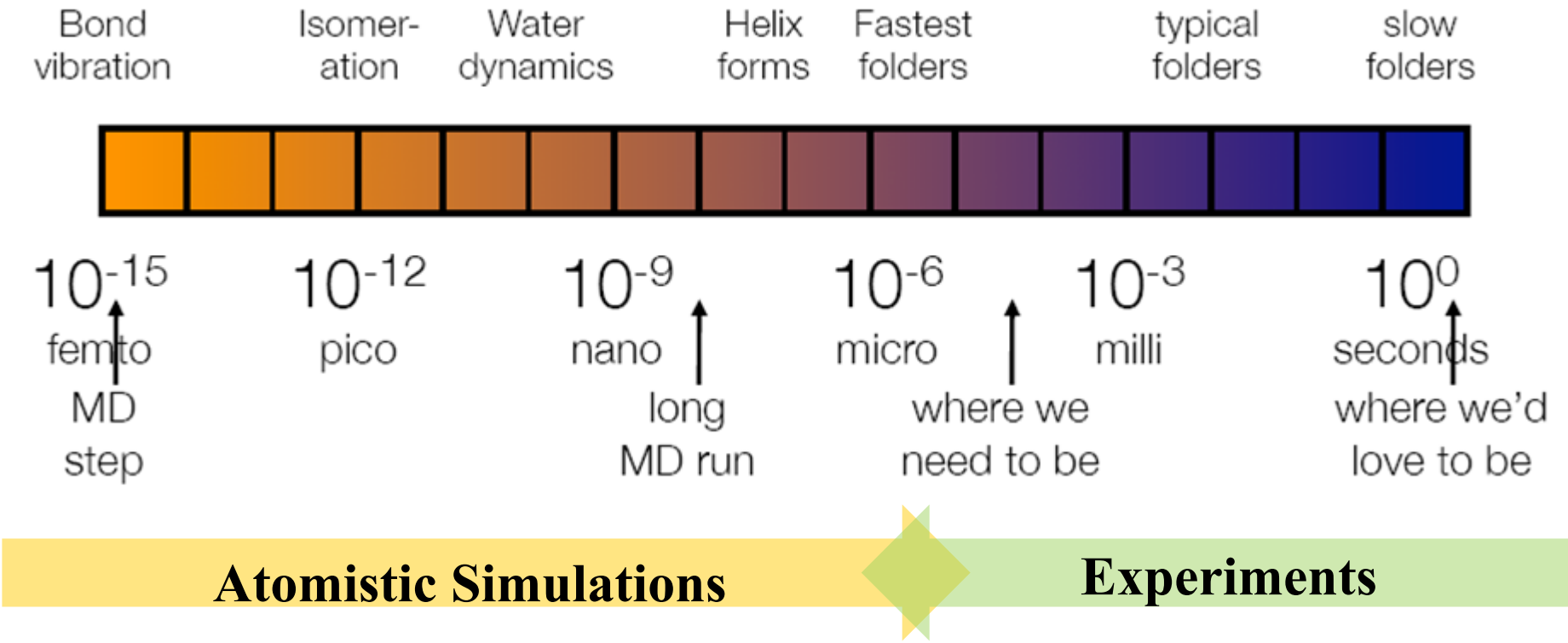
Single Molecule FRET

Provide information of an order parameter



Zhuang, Science 296:1473, 2002.

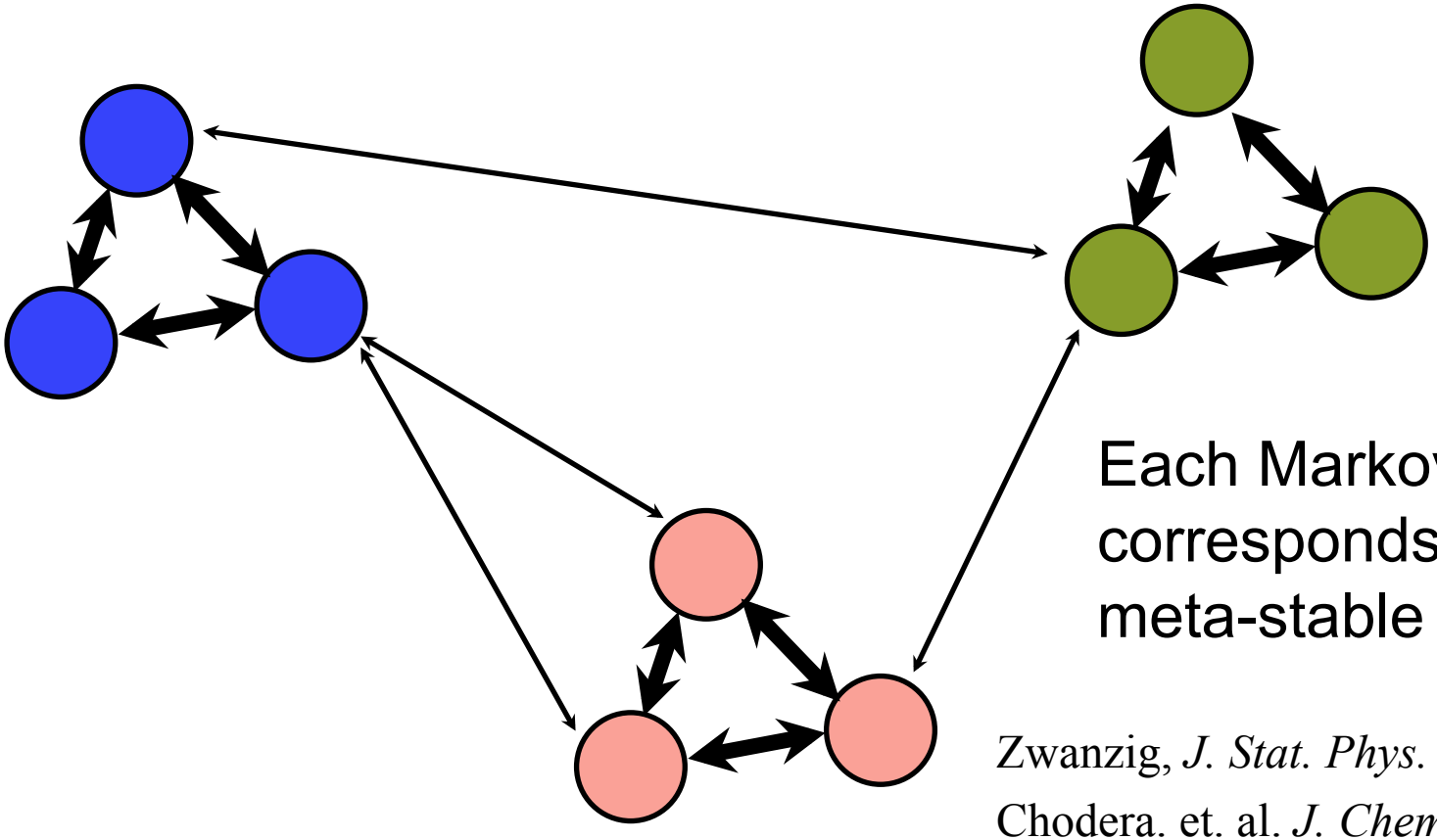
Key Challenge: Timescale Gap



Solution:

Use short simulations to predict long timescale dynamics

Conformational Dynamics: Nearly Uncoupled Markov Chains



Each Markov chain corresponds to a meta-stable state

Zwanzig, *J. Stat. Phys.* 1983

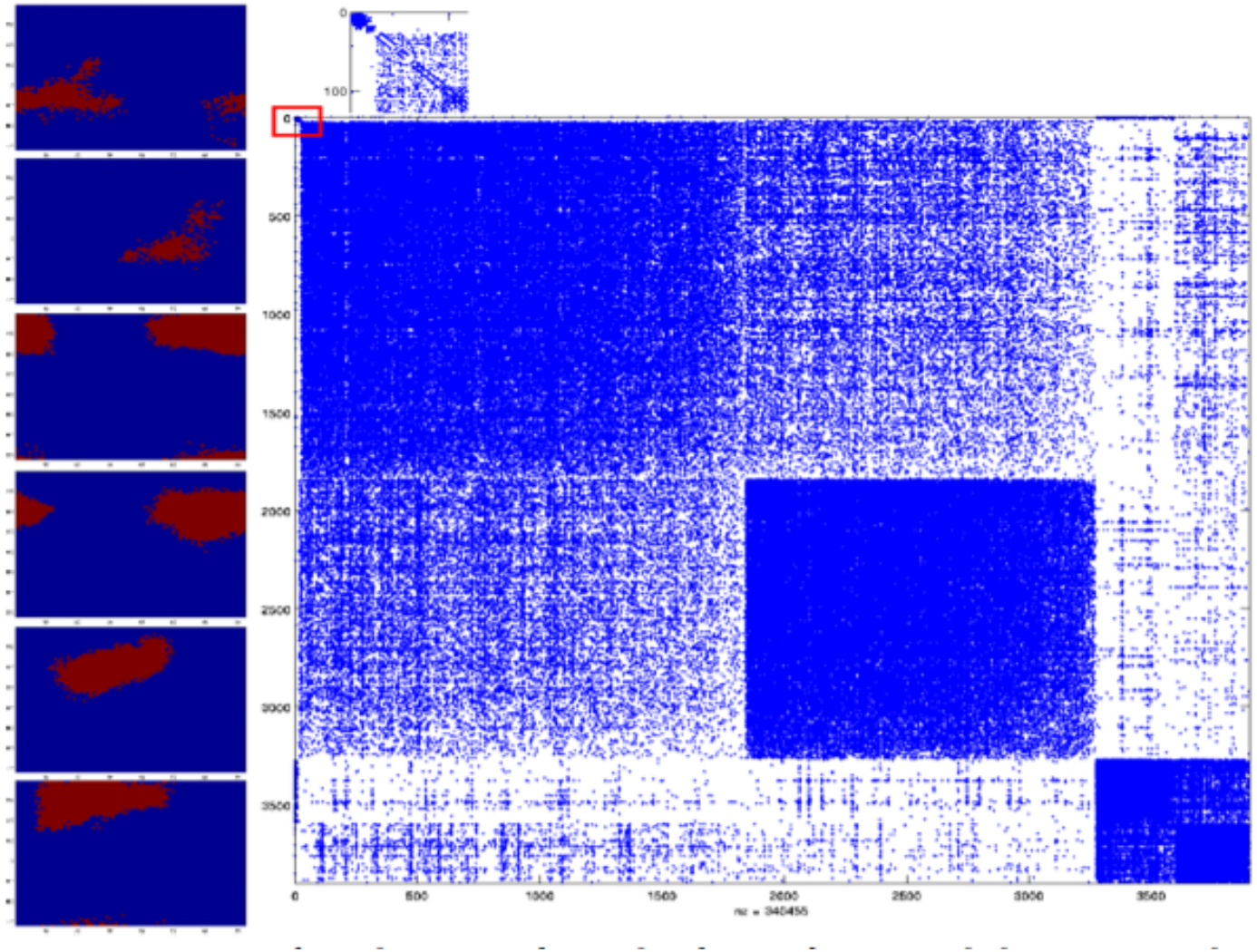
Chodera. et. al. *J. Chem. Phys.* 2007

Noé. et.al. *J. Chem. Phys.* 2007

Huang et.al. 2009, Hummer, Shuttle....

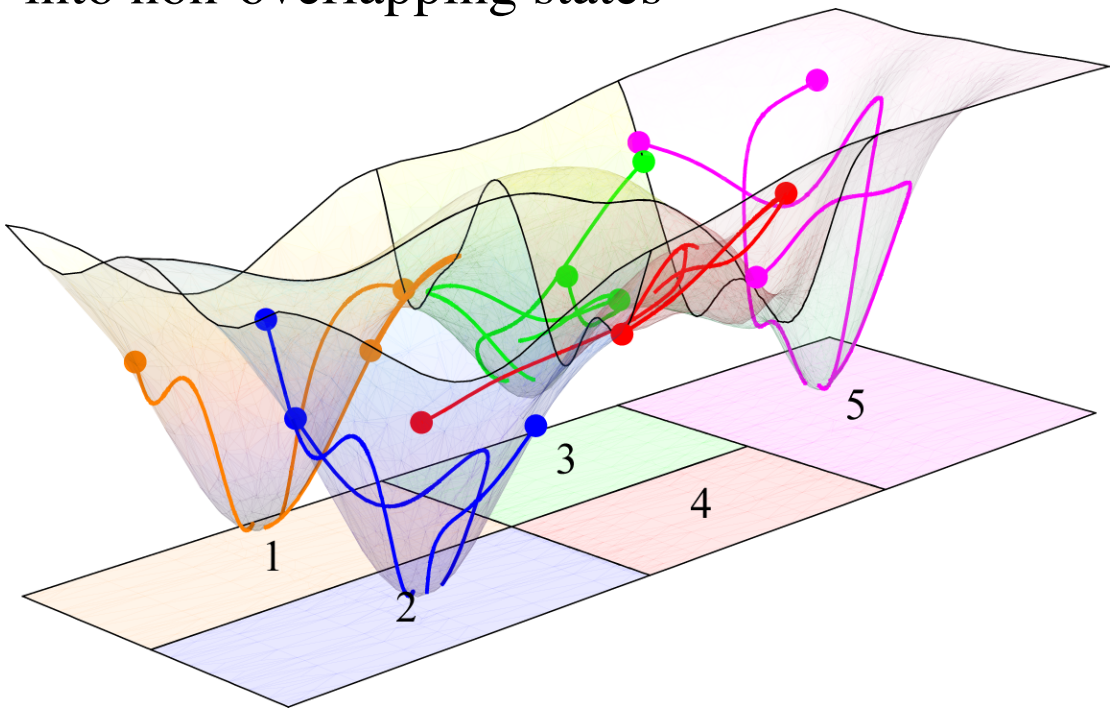
Figure Courtesy John Chodera

Block Structure of Transition Matrix



Markov State Models (MSMs)

The configuration space is decomposed into non-overlapping states



Define transition probabilities between states

$$T(\tau) = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{15} \\ p_{21} & p_{22} & & \\ \vdots & & \ddots & \\ p_{51} & & & p_{55} \end{bmatrix}$$

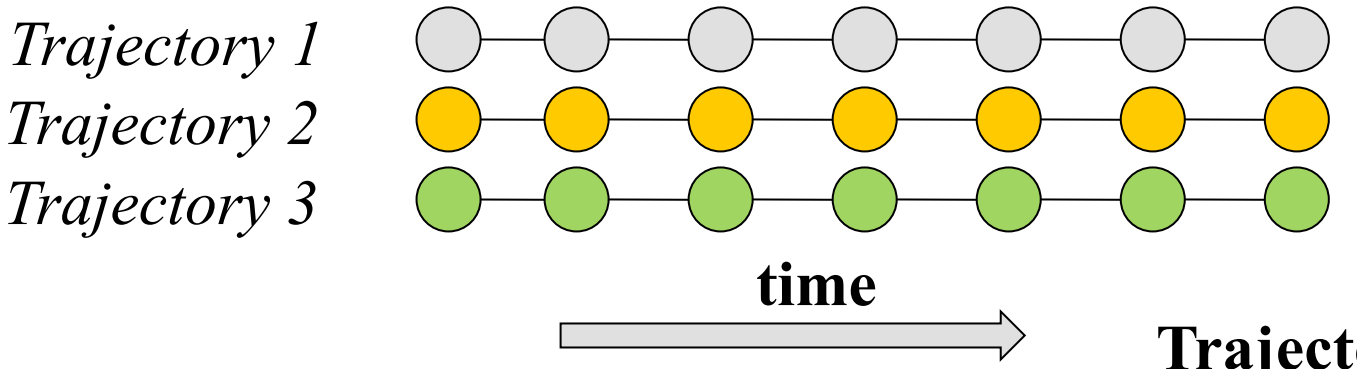
We can extract long time dynamics from MSMs built from short simulations

$$P(n\tau) = [T(\tau)]^n P(0)$$

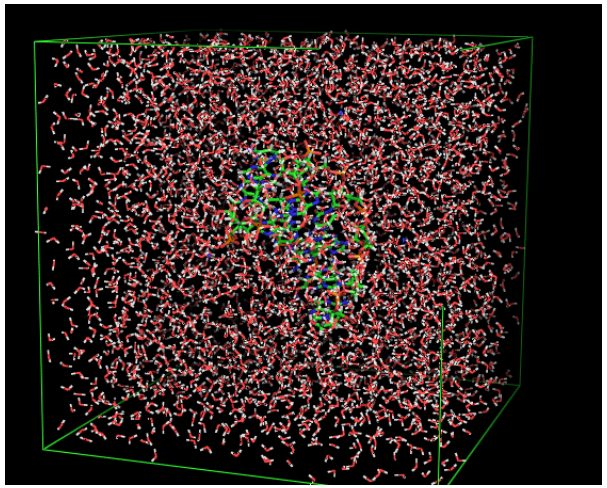
The time is coarse-grained in τ

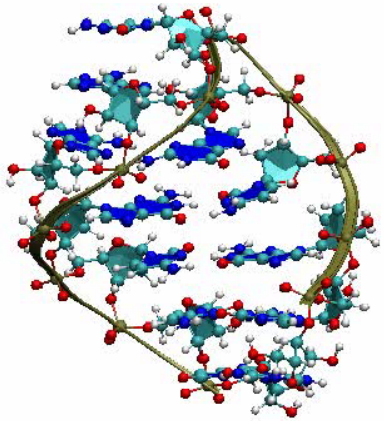
How to construct MSMs?

Dataset: Multiple trajectories with a lot of conformations.



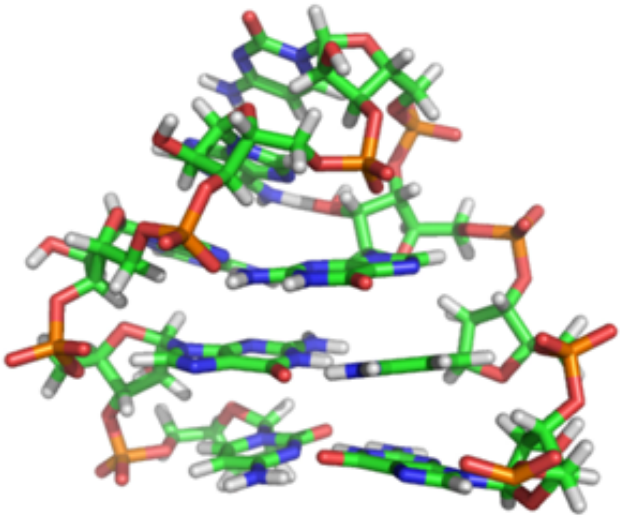
Conformation





Example: 8-RNA hairpin

An eight nucleotide RNA GCAA hairpin



- 2,543 TIP3P waters and 7 Na^+ ions
- 9963 45ns simulations

> 2.3 million conformations in total

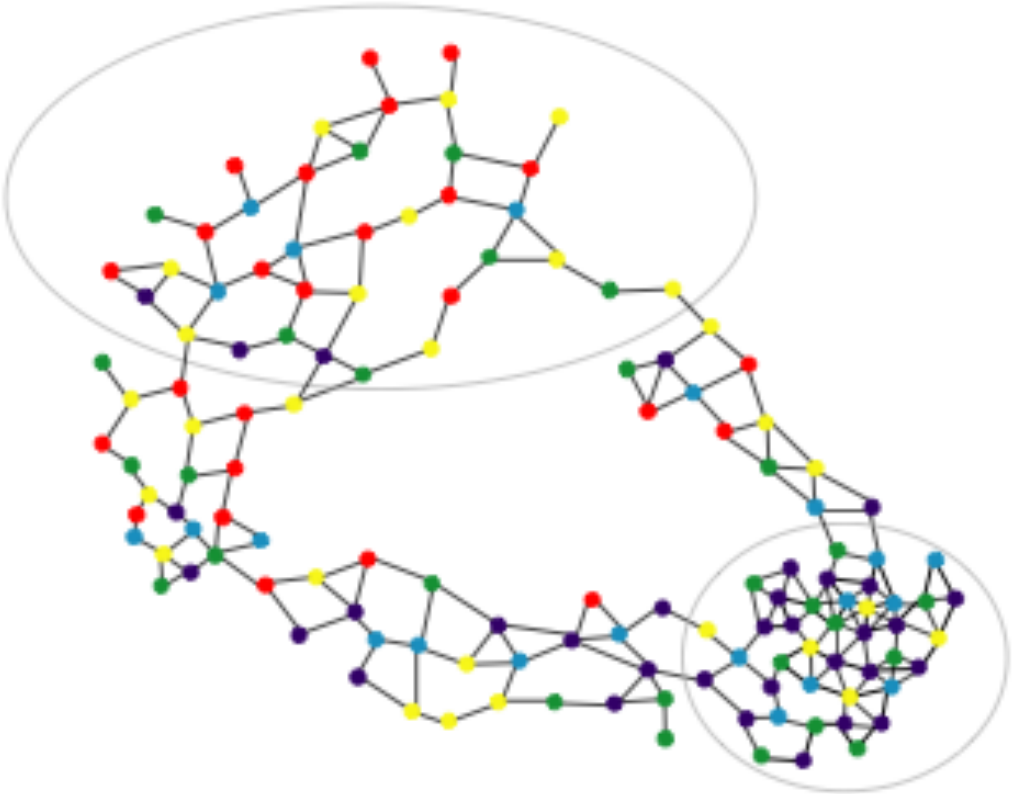
How to construct MSMs?

Data: A large amount of conformations

Directly work on conformations

Network nodes are snapshots from multiple simulations.

800,000 nodes, 7.4 billion edges



Very Expensive!

Andrec, Felts, Gallicchio & Levy (2005) PNAS, 102, 6801

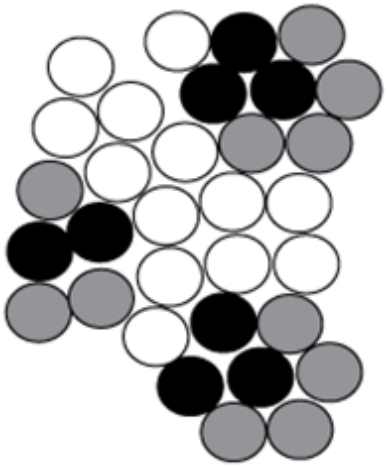
How to construct MSMs?

Conformations

Geometric Clustering
(K-means/K-center)

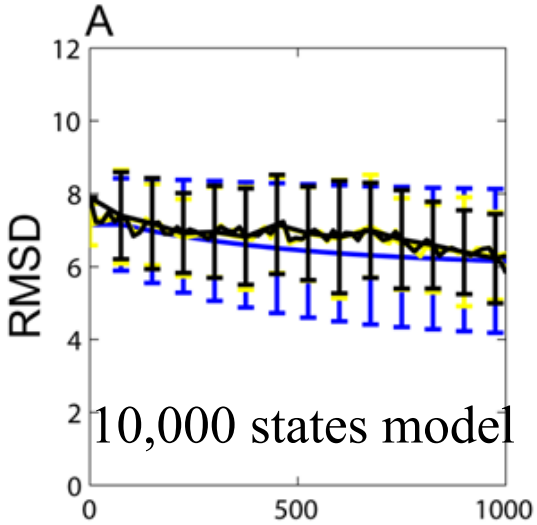
with distance metric
(RMSD …)

Microstates



Conformations within the same microstate interconvert quickly

Construct MSMs on microstates:

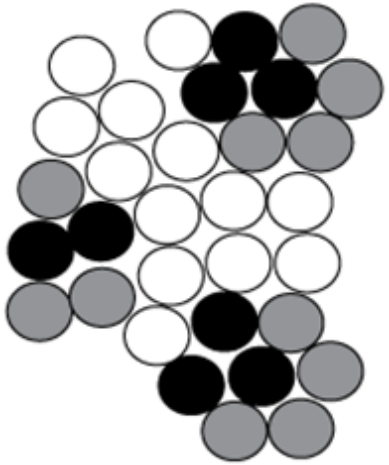


# of microstates is huge. Difficult for studying folding mechanisms!

Bowman, Beauchamp, Boxer, and Pande. Methods 2009.

How to construct MSMs?

Geometric Clustering (Splitting)



Spectral Clustering (Lumping)

Conformations

Microstates

Macrostates

Bayesian Inference of MSM

$$T(\tau) = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{15} \\ p_{21} & p_{22} & & \\ \vdots & & \ddots & \\ p_{51} & & & p_{55} \end{bmatrix}$$

Chodera. et. al. *J. Chem. Phys.* 2007

Noé. et.al. *J. Chem. Phys.* 2007

Deuflhard and Weber, *ZIB-report,* 2003

Weber, *ZIB-report,* 2004

Bowman, Huang, and Pande. *Methods* 2009.

Barcalado, et al. *J. Chem. Phys.* 2009

A Theory of Lumpability

- Lumpability
  - (Kemeny-Snell 1976) A finite Markov chain T is lumpable w.r.t. partition $S=(S_1,\ldots,S_n)$ iff its induced dynamics on S is Markovian
  - (Meila-Shi 2001) T is lumpable w.r.t. S iff T has n piece-wise constant right eigenvectors, T_{ij} is the transition probability from i to j.
  - If T is block diagonal, i.e. uncoupled Markov chain, then T is lumpable with piece-wise constant right eigenvectors associated with multiple eigenvalue 1.
  - (E-Li-Vanden-Eijnden 2007) For reversible chains, optimal approximation of lumpable Markov chains in Hilbert-Schmidt norms
  - An spectral algorithm to find lumpable states in nearly uncoupled systems:
    - find top n piece-wise constant eigenvectors as embedding coordinates
    - Use k-means to find n clusters
    - Other variants with spectral bipartition also works (PCCA)

But there are issues when using with k-center here!

分子动力系统中的聚类分析

I. Geometric Clustering (距离度量)

- K-means/K-medoids vs. K-center, etc.

II. Kinetic Clustering

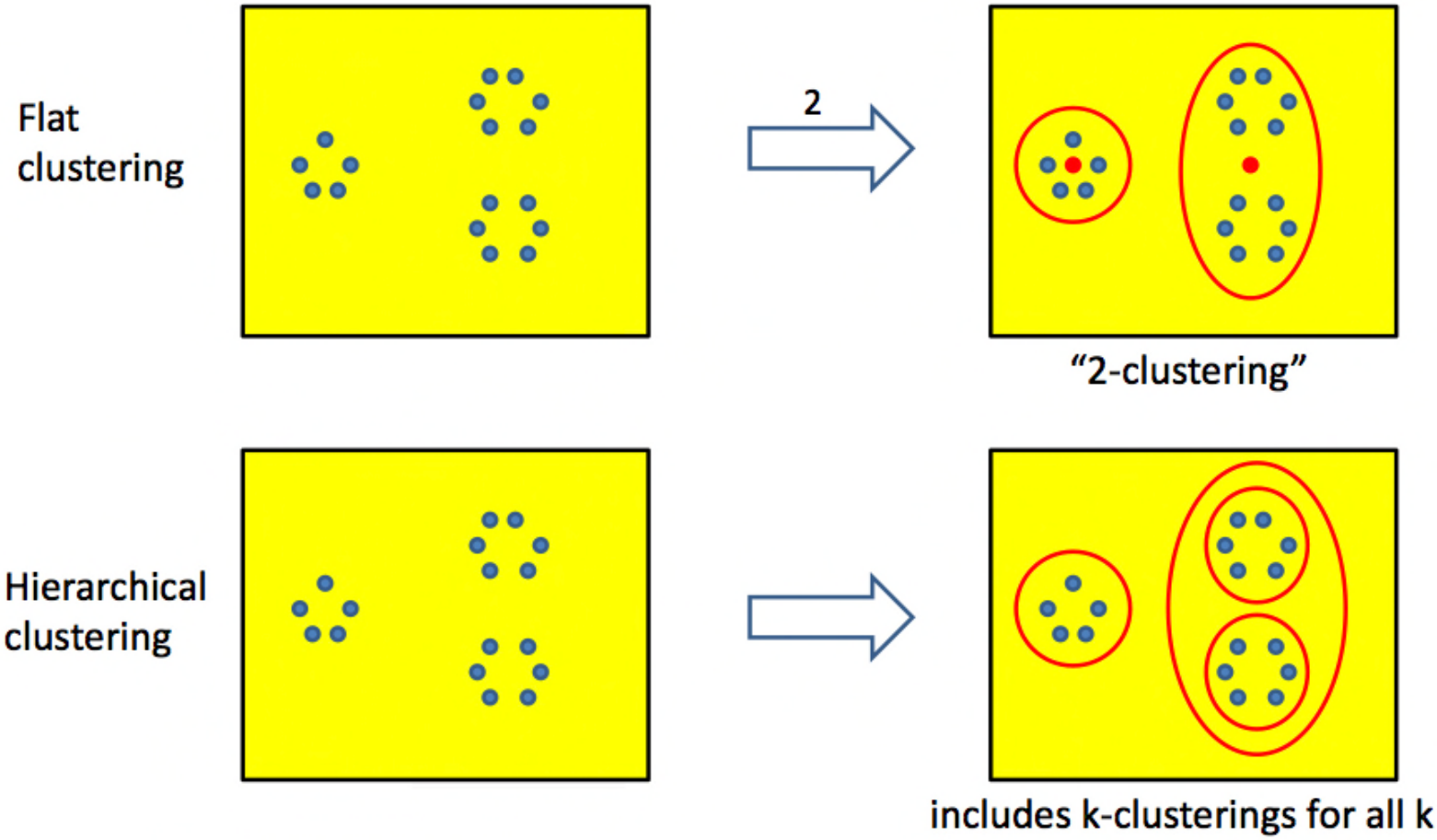
- Spectral clustering, etc.

III. 聚类分析的性质

1) Flat clustering vs. Hierarchical clustering
2) Batch vs.Streaming (online) data
3) 近似算法和计算复杂性
4) 统计性质

几何聚类分析
GEOMETRIC CLUSTERING

Two Types of Clustering



How is the data presented

- Batch

  n data point, all at once

  (can store all of them in memory)

- Online/streaming

  n or endless data point, one at once

  (o(1) or o(n) memory, can NOT store all of them)

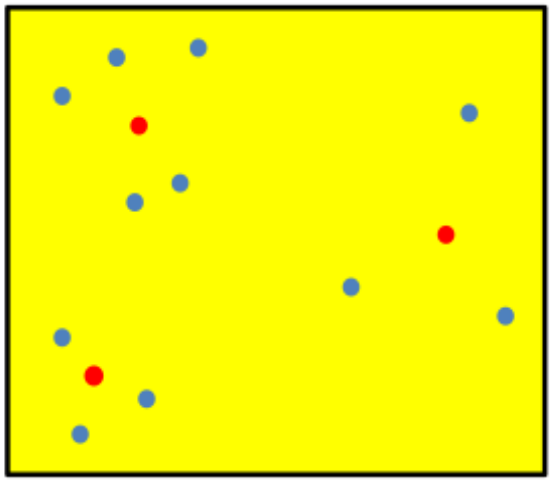
Molecular dynamics data is online/streaming in nature!

K-center vs. K-means

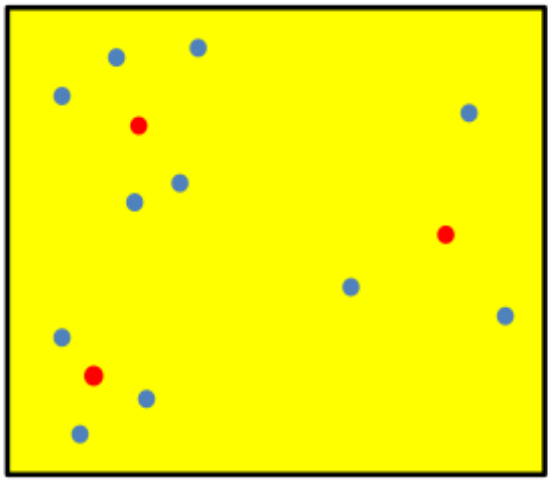
Input: Data set $X \subset R^D$, desired # of clusters k
Goal: Summarize data using a few representatives $C = \{c_1, c_2, \ldots, c_k\} \subset R^D$, to minimize overall distortion.

The *distortion* on a particular x is $d(x,C) = \min\{\|x - c\| : c \text{ in } C\}$

Max distortion (k-center)
$\max \{d(x,C) : x \text{ in } X\}$



Average distortion (k-means)
$\text{sum } \{d(x,C)^2 : x \text{ in } X\}$



A Greedy Algorithm for K-center

Farthest-first traversal

[Gonzalez, 1985]

Input: data set X, integer k

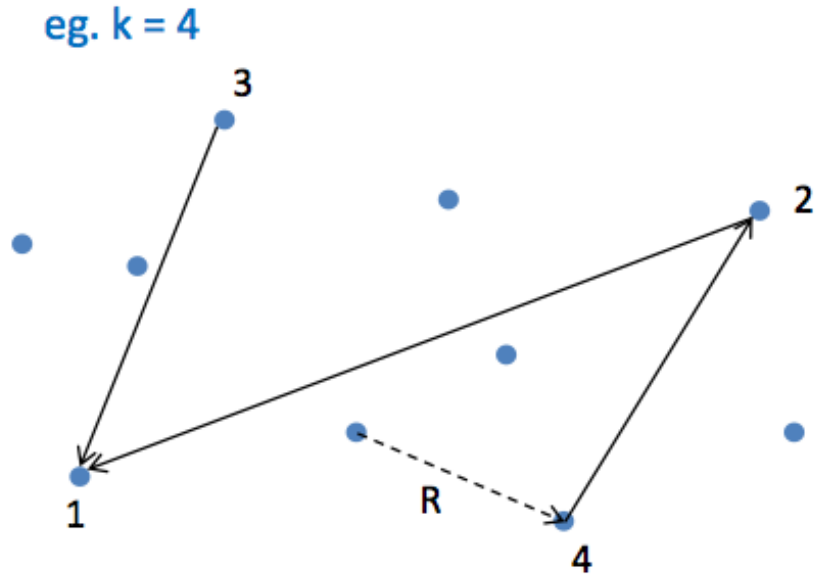
Pick any x in X and set C = {x}

for i = 2 to k:

    find x in X with largest d(x,C)

    add x to C

return centers C



Claim: cost(C) $\leq$ 2 OPT

Proof:

(i) Let x be the point in X that is farthest from C; and let R = d(x,C). Thus cost(C) = R.

(ii) The k+1 points C $\cup$ {x} are all at distance $\geq$ R from each other.

(iii) Any k-clustering must put two of these points in the same cluster; and this cluster must therefore have radius $\geq$ R/2. Therefore OPT $\geq$ R/2.

K-center 几何性质

- K-center形成了样本空间的一个epsilon-net
  - Any two points in C are R-distance away
  - Points in C form a R-cover of sample space
- 只依赖于度量结构
- K-center is NP-hard, but greedy algorithm is O(kn)
- K-center在ISOMAP(TdL'2000, Science)中被采用，称为Landmark技术
- Molecular dynamics application [Sun, Y, Huang, et al. JPC, 09]
- 缺点：
  - 对样本空间边缘的outlier和noise比较敏感

Approximability of K-center

Upper bounds [Gonzalez, 1985]

Farthest-first traversal achieves factor 2 approximation for data in any metric space.

Lower bounds [Feder and Greene, 1988]

Unless P = NP, no polynomial time algorithm achieves a factor:

better than 2 in a metric space

better than 1.82 in Euclidean space

Open problems:

1. Close the gap in the Euclidean case.
2. Other algorithms that are better in practice than farthest-first traversal?

A Greedy Algorithm for K-means

A stochastic farthest-first traversal

kmeans++

[Arthur and Vassilvitskii, 2006]

Input: data set X, integer k

Pick x in X at random, set C = {x}

for i = 2 to k:

    pick x in X at random, with probability $\propto d(x,C)^2$

    add x to C

return centers C

Claim: E[cost(C)] $\leq$ O(log k) · OPT

A Constant-factor Approximation

local search

[Kanungo et al, 2003]

Input: data set X, integer k

Pick initial centers C arbitrarily from X

while $\exists$ c in C, x in X with

    cost(C – {c} + {x}) < cost(C):

    C = C – {c} + {x}

return C

<u>Claim</u>: cost(C) $\leq$ 50 · OPT

Complexity of K-means

Upper bounds [Inaba et al, 1989]

Can solve optimally in time $O(n^{kd})$, where

n = number of points

d = dimension

Lower bounds [D. et al, 2009; Mahajan et al, 2009]

NP-hard in the following cases:

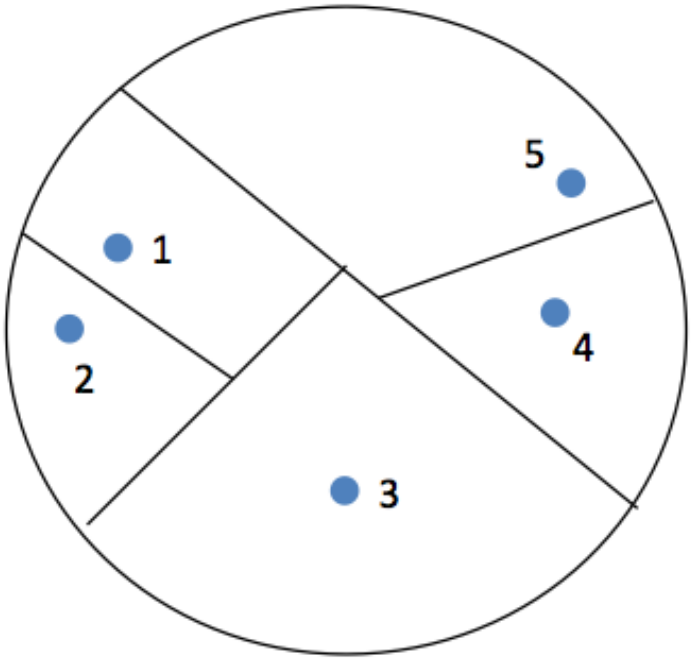
k = 2, arbitrary d

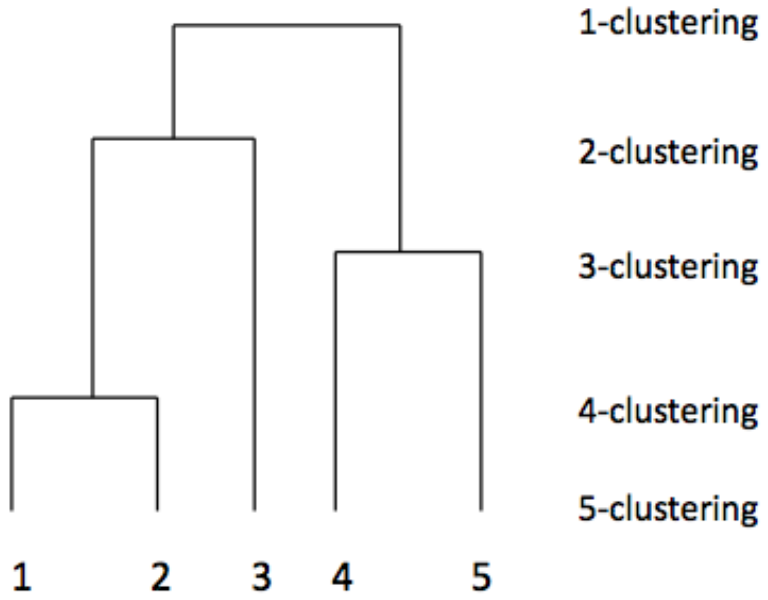
d = 2, arbitrary k

Open problems:

1. Better approximation algorithms?
2. Hardness of approximation results?

Hierarchical Clustering



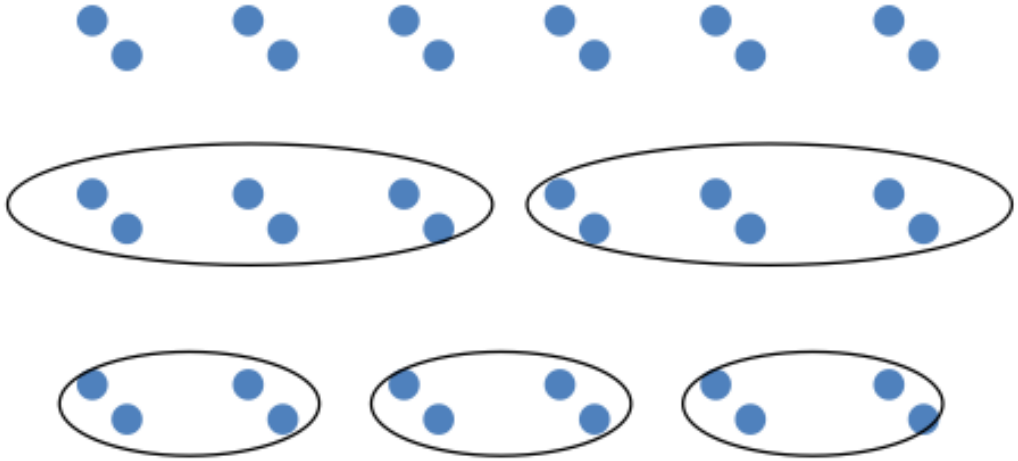


Popular form of data analysis:

- No need to specify number of clusters
- Can view data at many levels of granularity, all at the same time
- Simple greedy agglomerative heuristics for constructing these clusterings

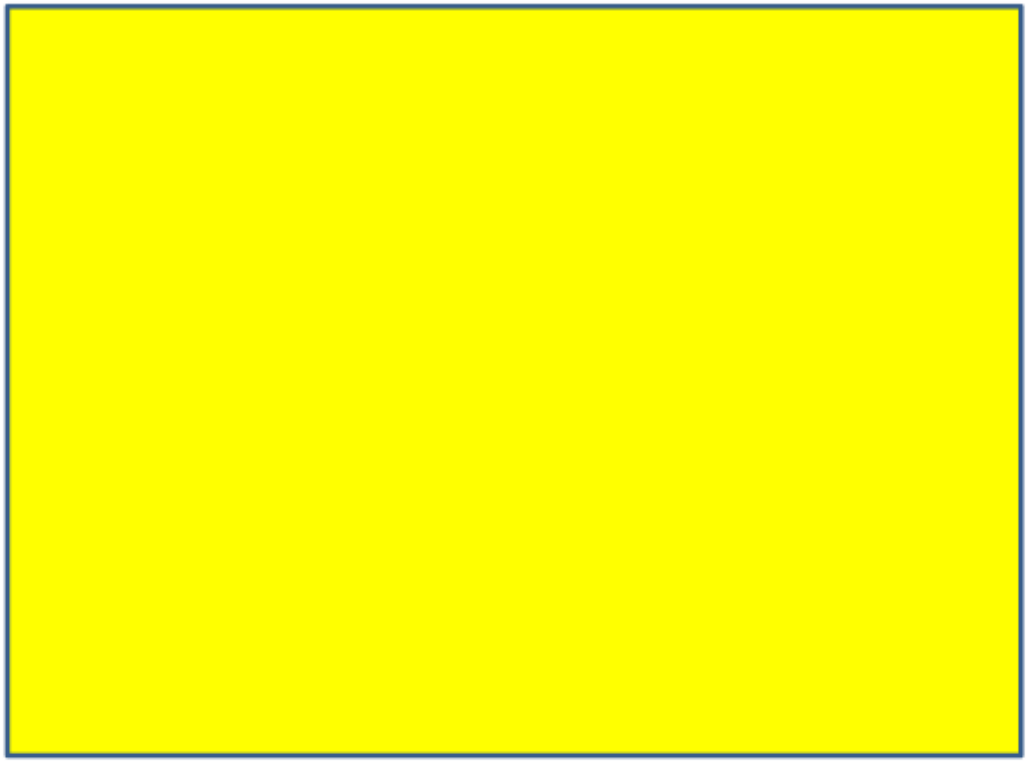
A Basic Existence Problem

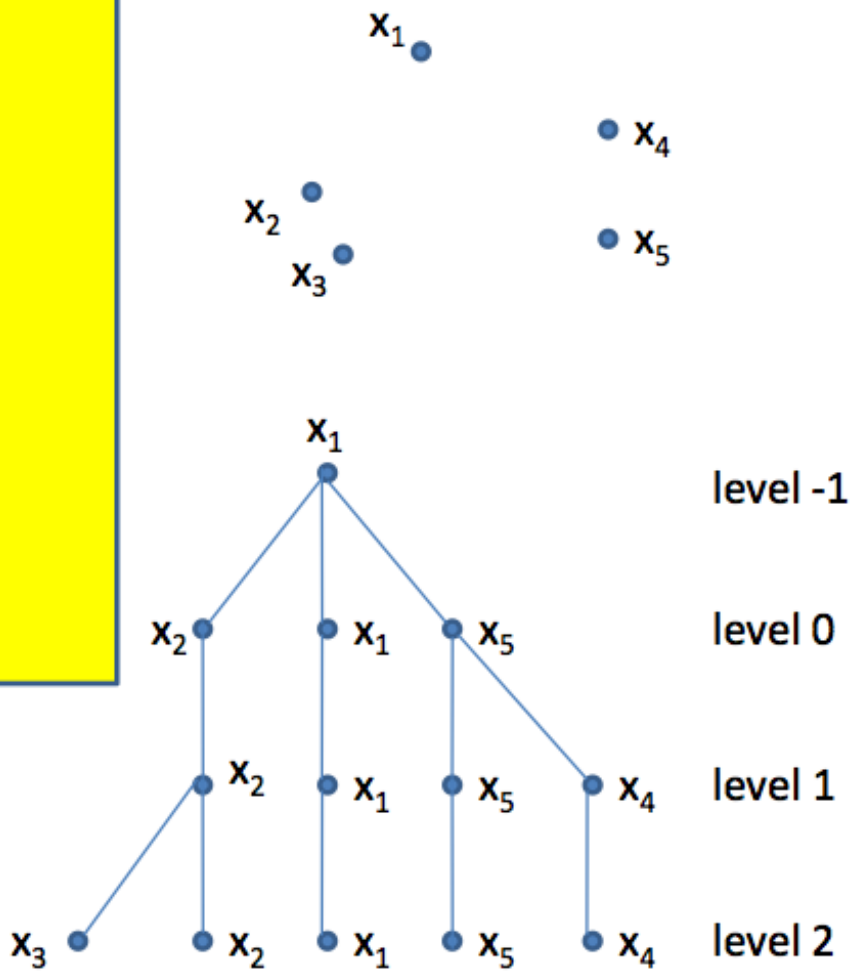
The whole enterprise of hierarchical clustering could use some more justification.



Must there always exist a hierarchical clustering which is close to optimal at *every* level of granularity, simultaneously? [such that for *all k*, the induced *k*-clustering is close to the best *k*-clustering?]

Hierarchical K-center

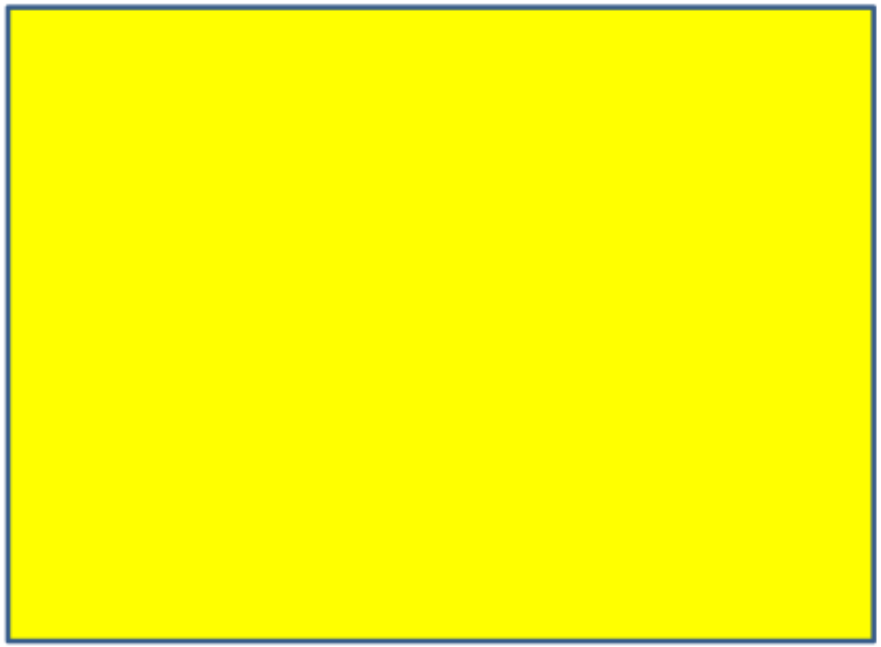


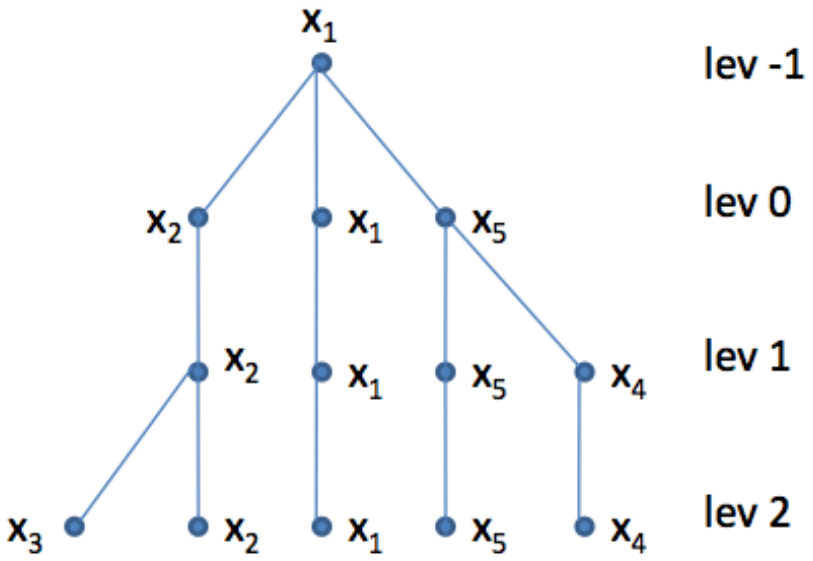


Build online! When new point x arrives:

1. Find largest j such that x is within dist $1/2^j$ of some node p at level j
2. Make x a child of p

Hierarchical K-center: Complexity





Claim: For any k, consider the lowest level with $\leq$ k nodes, and let C_k be those nodes. Then cost(C_k) $\leq$ 8 OPT_k.

Proof: (Suppose it is level j.) C_k's children are within $1/2^j$ of it, and its grandchildren are within $1/2^j + 1/2^{j+1}$ of it, and so on. Therefore:

$$\text{cost}(C_k) \leq 1/2^j + 1/2^{j+1} + 1/2^{j+2} + \ldots \leq 1/2^{j-1}$$

Meanwhile, level j+1 has $\geq$ k+1 nodes, at dist $\geq 1/2^{j+1}$ from each other. Any k-clustering puts two of these in the same cluster, and thus has radius $\geq 1/2^{j+2}$.

Hierarchical Clustering: Open Problems

1. Hierarchical k-center: closing the gap

   Upper bound: we have a factor 8 approximation. Can we do better?

   Two sources of lower bounds:
   - Hardness of approximation of k-center (factor of 2)
   - Hierarchical incompatibility of optimal k-clusterings (factor of 2?)

   Can these be combined to give a lower bound greater than 2?

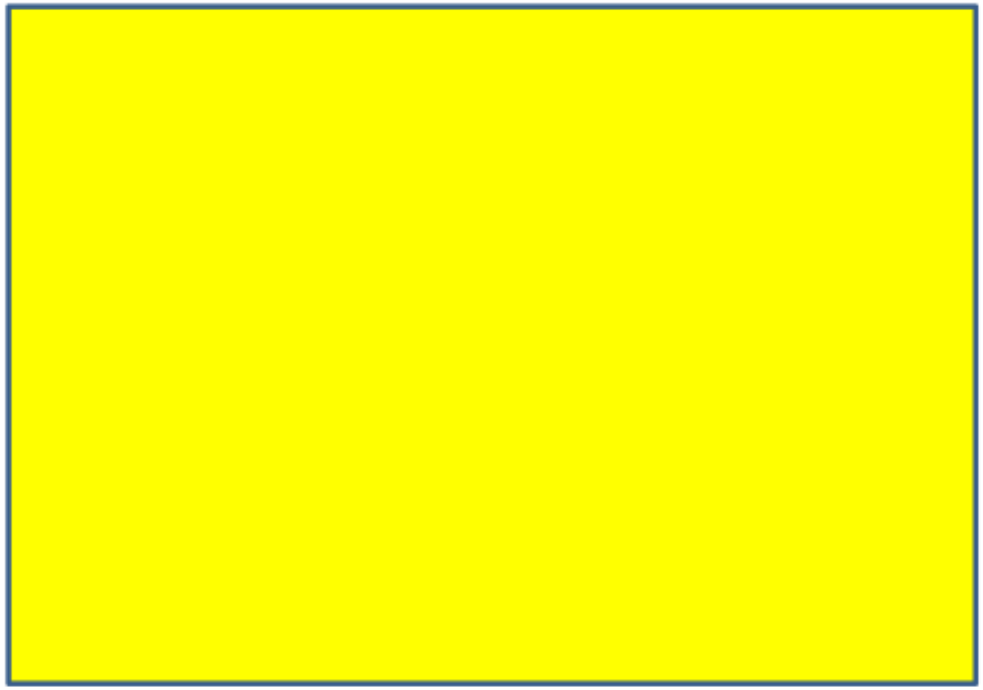
2. Hierarchical k-means

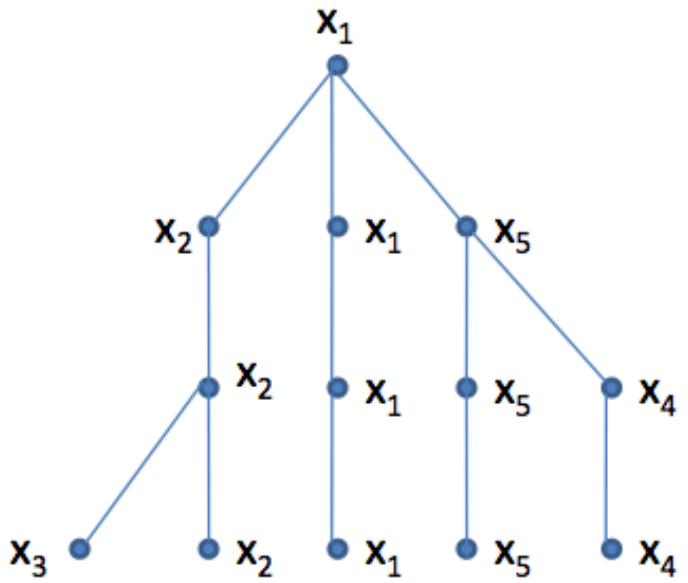
   Good algorithms for this?

Clustering online/streaming data

Online	Streaming
Endless stream of data	Stream of (known) length n
Fixed amount of memory	Memory is o(n), e.g. sqrt(n)
Tested at every time step	Tested only at the very end
Each point is only seen once	More than one pass may be possible

Online K-center





For each new point x that arrives:

Find largest j such that x is within dist $1/2^j$ of some node p at level j

Make x a child of p

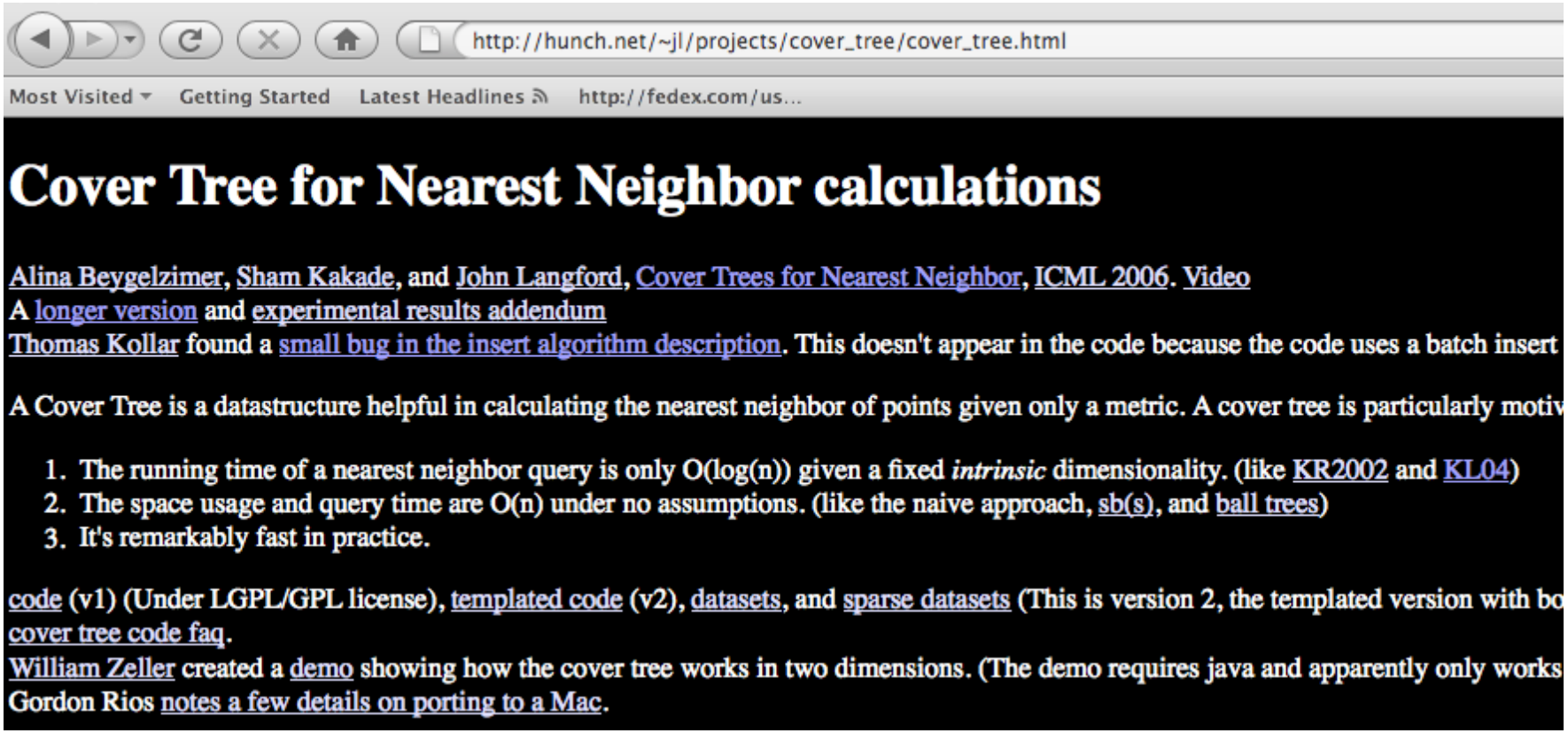
Problem: requires O(n) space – all points are stored

Solution: only maintain levels upto the first level j with $\geq$ k nodes

Open problem: online k-means.

Online K-center Implementation

- Cover Tree

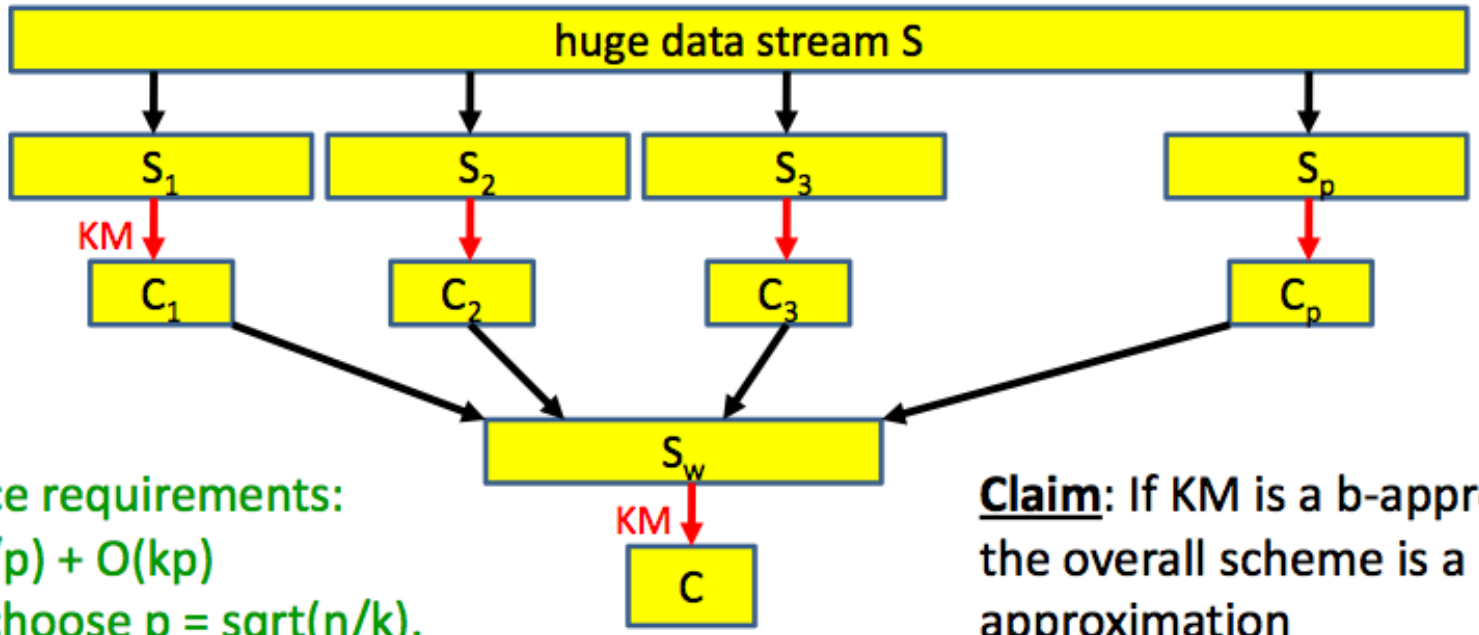


Streaming K-means: I

Strategy #1: *divide and conquer.*
[Guha et al 03]
Start with approx alg KM for *weighted* k-means: each point x has a weight w(x) and cost of k-clustering C is:

$$\text{cost}(C) = \text{sum}\{ w(x)\, d(x,C)^2 \}$$

Divide stream S into p groups $S_1, \ldots, S_p$
for each i = 1, 2, ..., p:
  KM(S_i) yields centers $C_i = \{c_{i1}, \ldots, c_{ik}\}$
  and clusters $S_{i1}, \ldots, S_{ik} \subseteq S_i$
S_w = {all c_{ij}}, with weights $w(c_{ij}) = |S_{ij}|$
return KM(S_w)



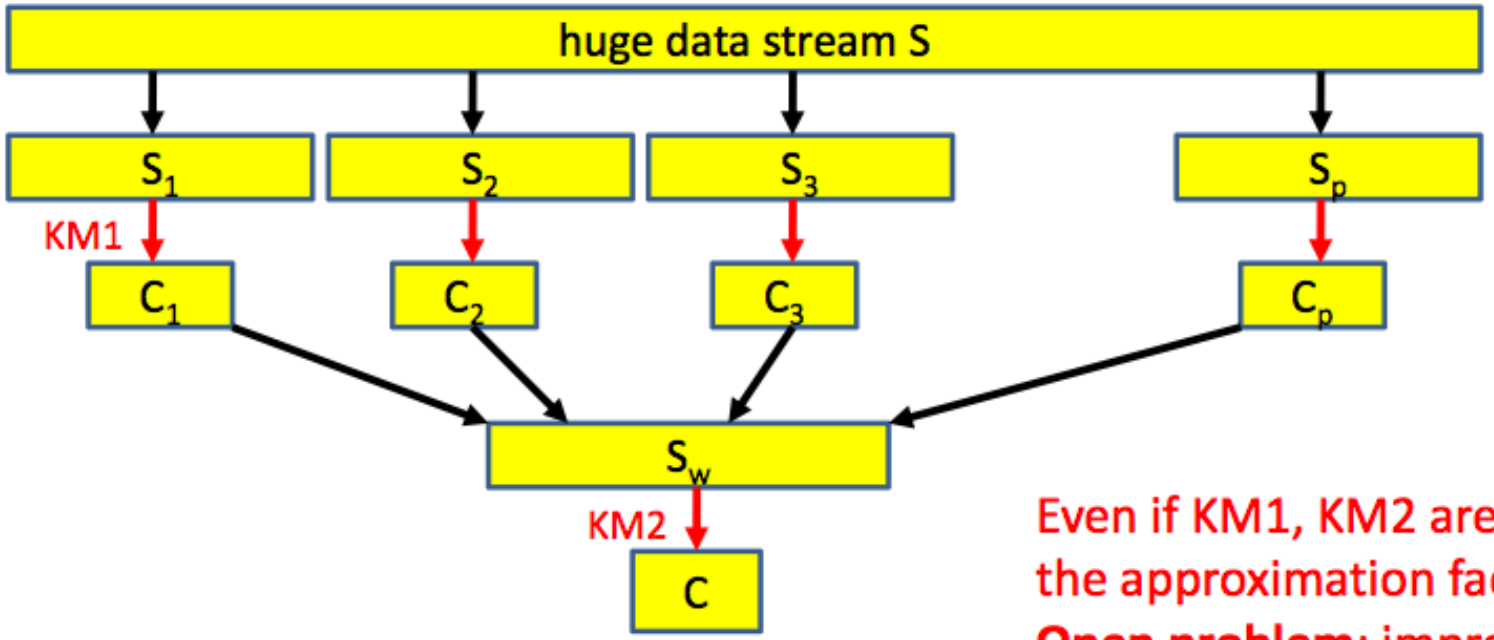
Space requirements:
O(n/p) + O(kp)
eg. choose p = sqrt(n/k).

Claim: If KM is a b-approximation, the overall scheme is a $(6b + 4b^2)$ approximation

Streaming K-means: I (bicriterion)

Bicriterion version:
An (a,b)-approximation for k-means yields ak centers with cost at most b times that of the best k-means solution.

Claim: If KM1 is an (a,b)-approximation and KM2 is an (a',b')-approximation, the overall scheme is an (a', 2b + 4b'(b+1)) approximation.



Even if KM1, KM2 are perfect, the approximation factor is 10.
Open problem: improve this!

Streaming K-means: II

Strategy #2: *random sampling*.
[Indyk 99]

Assume we have an (a,b)-approx alg KM.

Pass 1:
S' = random subset of S, of size s
C' = KM(S')

All large clusters are sufficiently sampled in S', and covered by C'

Pass 2:
S'' = (8kn/s) log (k/δ) points farthest from C'
C'' = KM(S'')

Now pick up *all* points not adequately covered by C'

return C' $\cup$ C''

Claim: With probability $\geq$ 1- δ, this is a (2a, 2(b + 1)(1+ 4/δ)) approximation.

Hierarchical Agglomerative Clustering

Building a hierarchical clustering:

1. Start with each data point in its own cluster.
2. Repeatedly merge two "closest" clusters.

Notion of distance between clusters:

Single linkage
closest pair of points
Complete linkage
furthest pair of points
Average linkage – several variants
(i) distance between centers
(i) average pairwise distance
(ii) *Ward's method*: increase in k-means cost due to merger

Guarantees for Agglomerative Clustering

Complete linkage has underlying k-center cost function.

Approximability characterization: for all k, the induced k-clustering is within factor $\alpha(k)$ of the optimal k-center solution… what is $\alpha(k)$?

Claim: [Dasgupta 09] k ≤ a(k) ≤ k^log3

[Recall: cover tree has a(k) = 8.]

Open Problem: Ward's method of average linkage has the underlying k-means cost function… what is its approximation ratio?

Statistical Theory for Clustering

I. Consistency of K-means

II. Density Cluster Tree and Consistency of Single-linkage

Consistency of K-means

Suppose data $D_n=\{X_i: I=1,\ldots,n\}$ is drawn iid from an underlying distribution P.

Let C_k be the optimal k-means centers with respect to P.

Let C_{nk} be the optimal k-means centers for D_n.

Claim: [Pollard 81] If C_j is unique for $1 \le j \le k$, then dist(Ck,Cnk)->0 a.s.

Here $\text{dist}(S,T) = \max_{s \in S} \min_{t \in T} \|s - t\|$

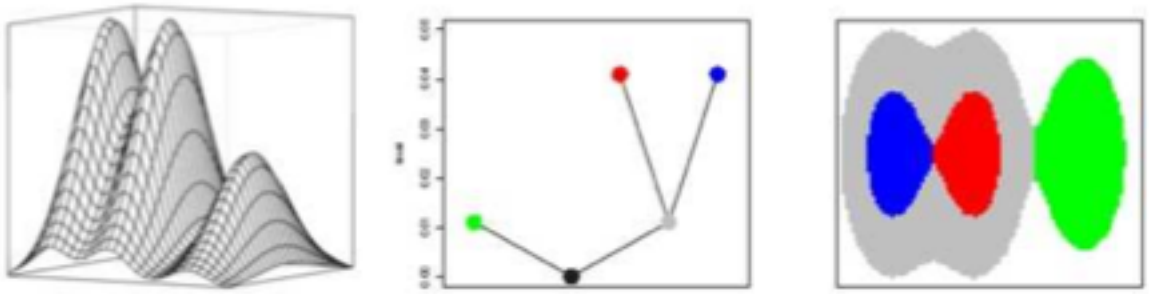
Issues:

1. C_{nk} is NP-hard to compute.
2. Is C_k something truly useful?

Density Cluster Tree

For any density p(x), consider the super-level set $\{x: p(x) \geq r\}$ and let C_r be the connected components of this super-level set.

Claim: [Hartigan 81] If $r \leq s$,then $C_s \subseteq C_r$, ie Hierarchical clustering with tree structure.



Which clustering converges to Cluster Tree

Robust Single Linkage: Build a neighborhood graph G_r,

nodes $\{X_i\}$,

edges $\{(i,j): \text{dist}(X_i, X_j) \leq r\}$,

discard nodes with degree $< c \log n$,

Let C_{nr} be the connected components of such a graph.

Claim: [Stuetzle 03, Zhou-Wong 08] C_{nr} converges to density cluster tree.

In fact: this is equivalent to the 1-skeleton Rips complex with persistent 0-homology, a special case in computational topology.

Other methods: Witness complex?

Bibliography I

D. Arthur, B. Manthey, and H. Roglin. k-Means has polynomial smoothed complexity. arXiv:0904.1113v1, 2009.

D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. ACM-SIAM Symposium on Discrete Algorithms, 2007.

A. Beygelzimer, S. Kakade, and J. Langford. Cover trees for nearest neighbor, 23rd International Conference on Machine Learning, pp. 97-104, 2006.

S. Dasgupta and Y. Freund. Random projection trees for vector quantization. IEEE Transactions on Information Theory, 55(7), 2009.

D. Arthur, B. Manthey, and H. Roglin. k-Means has polynomial smoothed complexity. arXiv:0904.1113v1, 2009.

T. Feder and D.H. Greene. Optimal algorithms for approximate clustering. ACM Symposium on Theory of Computing, pp 434-444, 1988.

T.F. Gonzalez. Clustering to minimize the maximum intercluster distance. Theoretical Computer Science, 38:293-306, 1985.

S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams: theory and practice. IEEE Transactions on Knowledge and Data Engineering, 15(3):515- 528, 2003.

J.A. Hartigan. Consistency of single linkage for high-density clusters. Journal of the American Statistical Association, 76:388-394, 1981.

M. Inaba, N. Katoh, H. Imai. Variance-based k-clustering algorithms by Voronoi diagrams and randomization. IEICE Transactions on Information and Systems, E83-D:1199-1206, 2000.

P. Indyk. Sublinear time algorithms for metric space problems. ACM Symposium on Theory of Computing, 1999.

T. Kanungo, D.M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A.Y. Wu. A local search approximation algorithm for k-means clustering. Computational Geometry: Theory and Applications, 28:89-112, 2004.

R. Krauthgamer and J.R. Lee. Navigating nets: simple algorithms for proximity search. ACM-SIAM Symposium on Discrete Algorithms, 2004.

Bibliography II

J. MacQueen. Some methods for classification and analysis of multivariate observations. Proceedings of the 5thBerkeley Symposium on Mathematics, Statistics, and Probability, pp 281-296, 1967.

M. Mahajan, P. Nimbhorkar, and K.R. Varadarajan. The planar k-means problem is NP-hard. Proceedings of 3rd Annual Workshop on Algorithms and Computation, pp.274--285, 2009.

R. Nugent and W. Stuetzle. A Generalized Single Linkage Method for Estimating the Cluster Cluster Tree of a Density. Journal of Classification, 20(5): 25-47, preprint.

D. Pollard. Strong consistency of k-means clustering. Annals of Statistics, 9(1):135-140, 1981.

W. Stuetzle. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. Journal of Classification, 20(5):25-47, 2003.

J. Sun, Y. Yao, X. Huang, V. Pande, G. Carlsson, and L. Guibas, A Fast Geometric Clustering Method on Conformation Space of Biomolecules, submitted to Journal of Physical Chemistry, 2009

A. Vattani. k-means requires exponentially many iterations even in the plane. ACM Symposium on Computational Geometry, 2009.

J.H. Ward. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58(201):236-244, 1963.

Q. Zhou and W.-H. Wong. Reconstructing the energy landscape of a distribution from Monte Carlo samples. Annals of Applied Statistics, 2(4): 1307-1331, 2008. doi: 10.1214/08-AOAS196.