Scalable Training of Mixture Models via Coresets

Dan Feldman
MIT

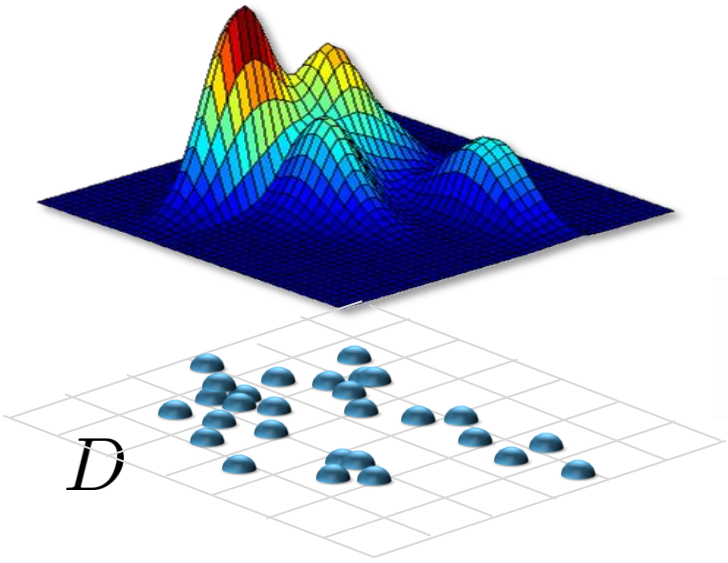
Matthew Faulkner
Caltech

Andreas Krause
ETH

Fitting Mixtures to Massive Data

EM, generally expensive

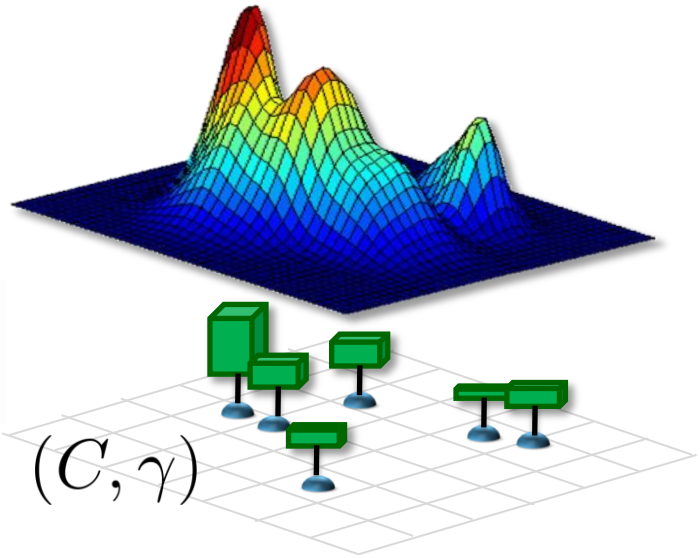
$$\theta^* = \arg\min_\theta \mathcal{L}(\theta|D)$$
$$\mathcal{L}(\theta|D) = -\sum_{i \in D} \ln P(x_j|\theta)$$



Small Sample

Weighted EM, fast!

$$\hat{\theta} = \arg\min_\theta \mathcal{L}(\theta|C)$$
$$\mathcal{L}(\theta|C) = -\sum_{j \in C} \gamma_j \ln P(x_j|\theta)$$



How to choose C to bound $|\mathcal{L}(\theta|D) - \mathcal{L}(\theta|C)|$?

Coresets for Mixture Models

A weighted set (C, γ) is an ϵ-**coreset** for $\text{cost}(\theta|D)$ if

$$(1-\epsilon)\text{cost}(\theta|D) \leq \text{cost}(\theta|C) \leq (1+\epsilon)\text{cost}(\theta|D)$$

k-medians: $\text{cost}(\theta|D) = \sum_{i=1}^{k} \sum_{x \in D_i} \text{dist}(x, \text{center}_i)$

k-means: $\text{cost}(\theta|D) = \sum_{i=1}^{k} \sum_{x \in D_i} \text{dist}(x, \text{center}_i)^2$

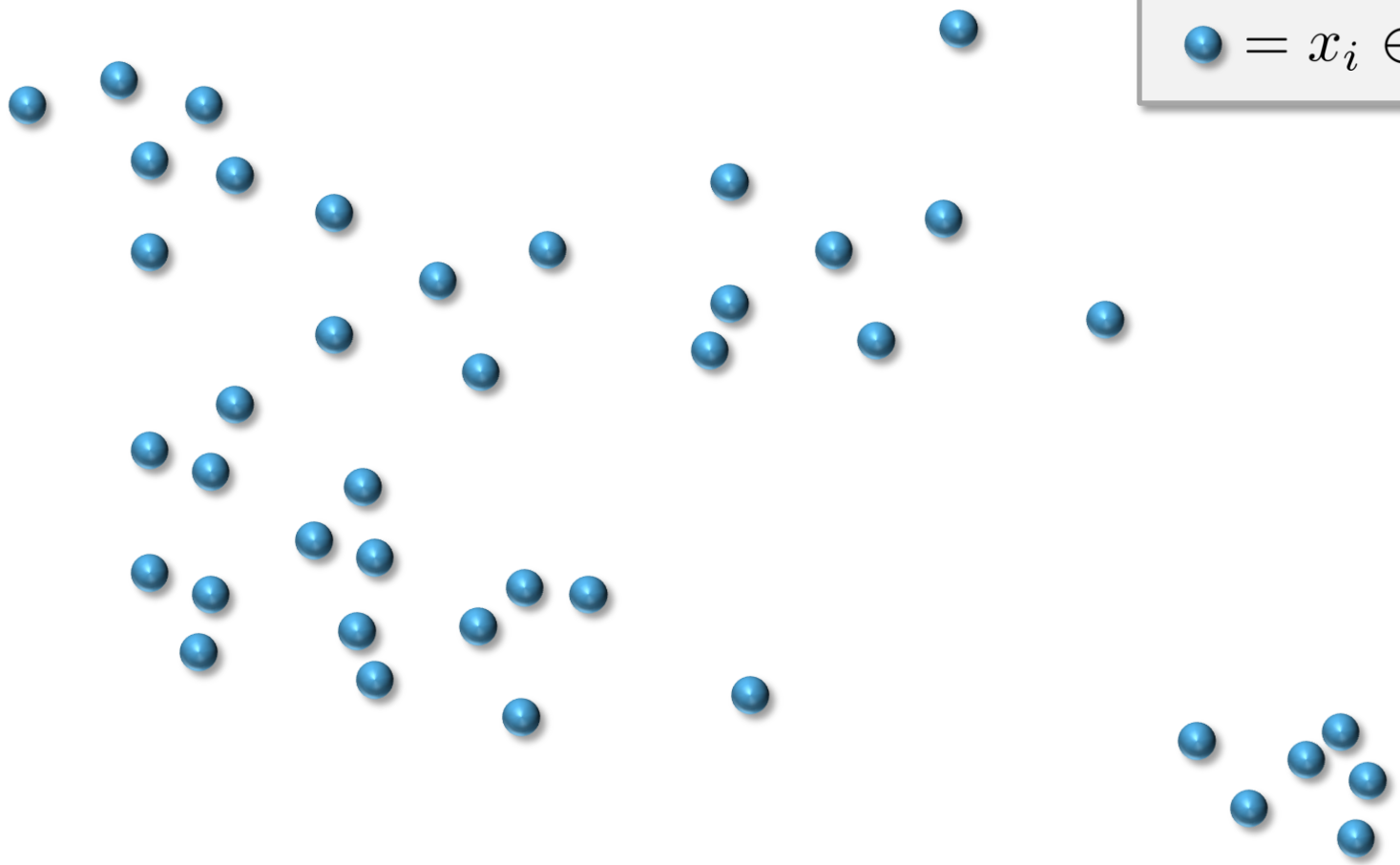
GMM: $\text{cost}(\theta|D) = \mathcal{L}(\theta|D) = -\sum_{i \in D} \ln P(x_j|\theta)$?

(C, γ) is a $(\mathbf{k}, \epsilon)$-**coreset** for GMMs if with prob. $> 1-\delta$

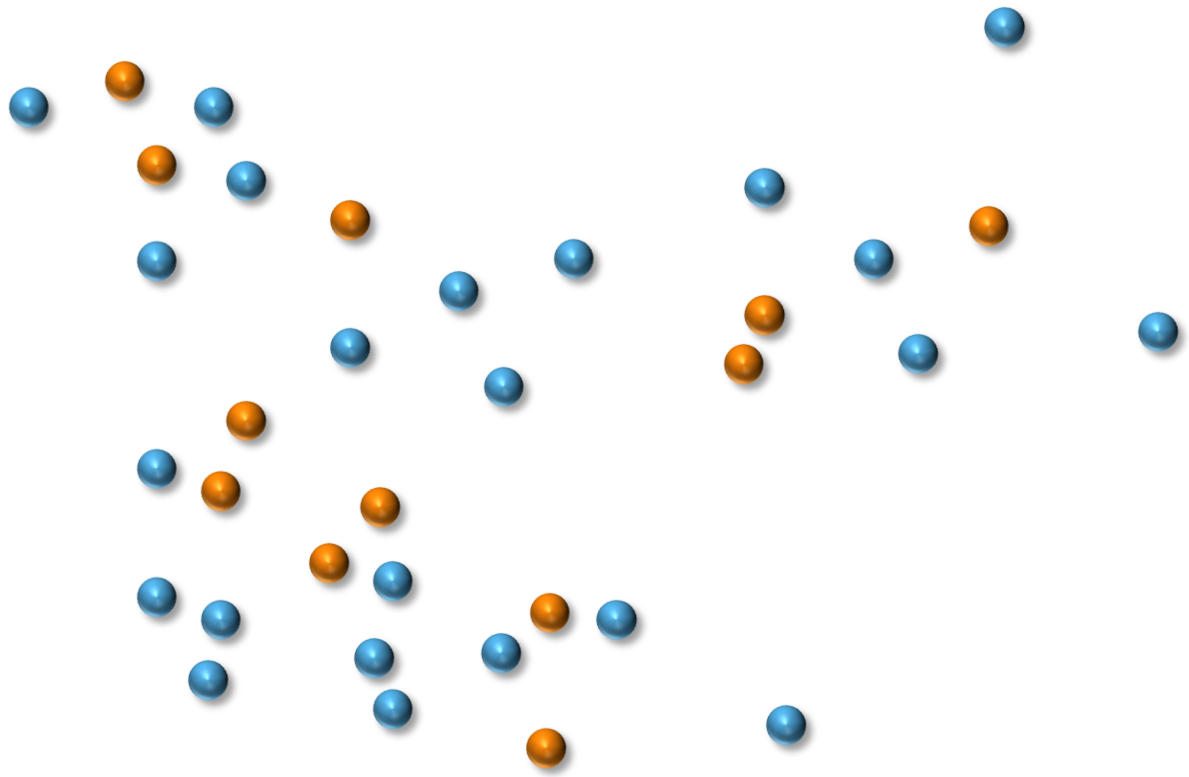
$$(1-\epsilon)\mathcal{L}(\theta|D) \leq \mathcal{L}(\theta|C) \leq (1+\epsilon)\mathcal{L}(\theta|D) \ *$$

for any k-mixture GMM with parameters θ

Naïve Uniform Sampling

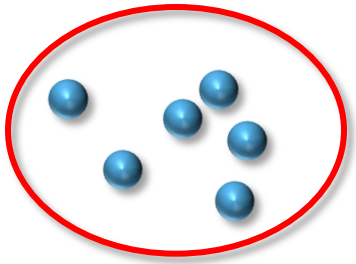


Naïve Uniform Sampling



$= x_i \in \mathbb{R}^d$

Small cluster is missed

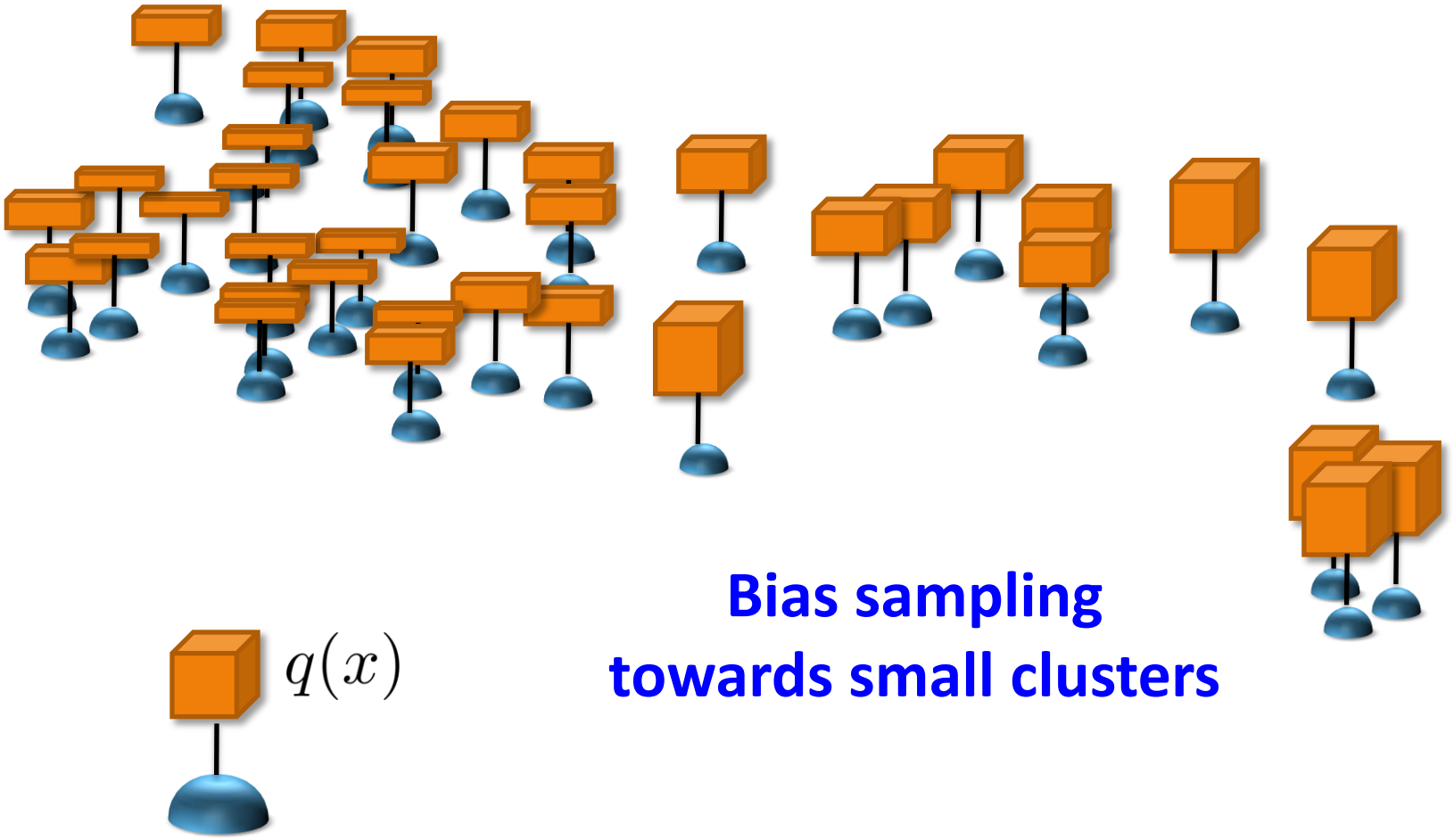


Sample a set *U* of *m* points uniformly

$$\mathcal{L}(\theta|U) = -\sum_{j \in U} \frac{n}{m} \ln P(x_j|\theta)$$

← High variance

Sampling Distribution

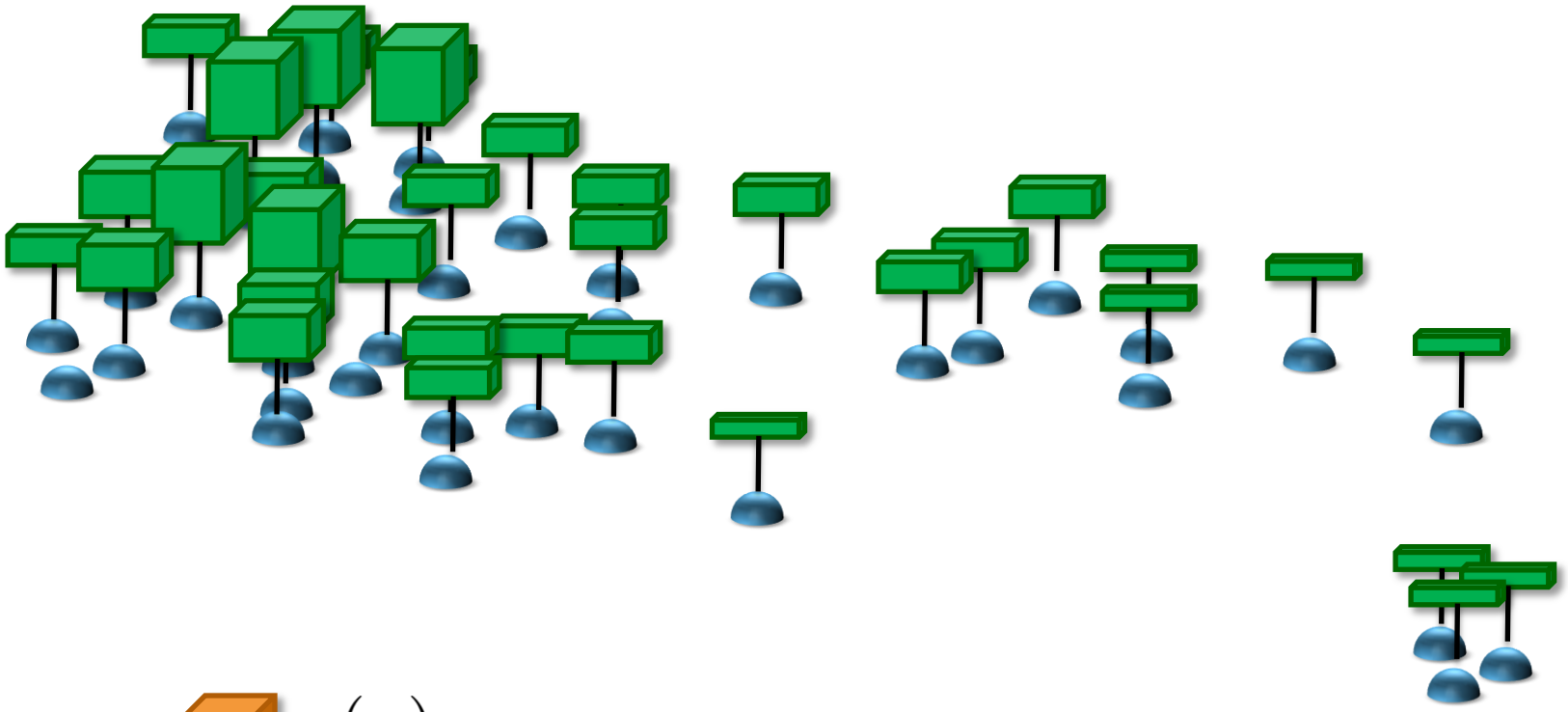


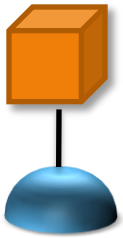
$q(x)$

Bias sampling towards small clusters

Sampling distribution

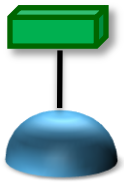
Importance Weights





$q(x)$

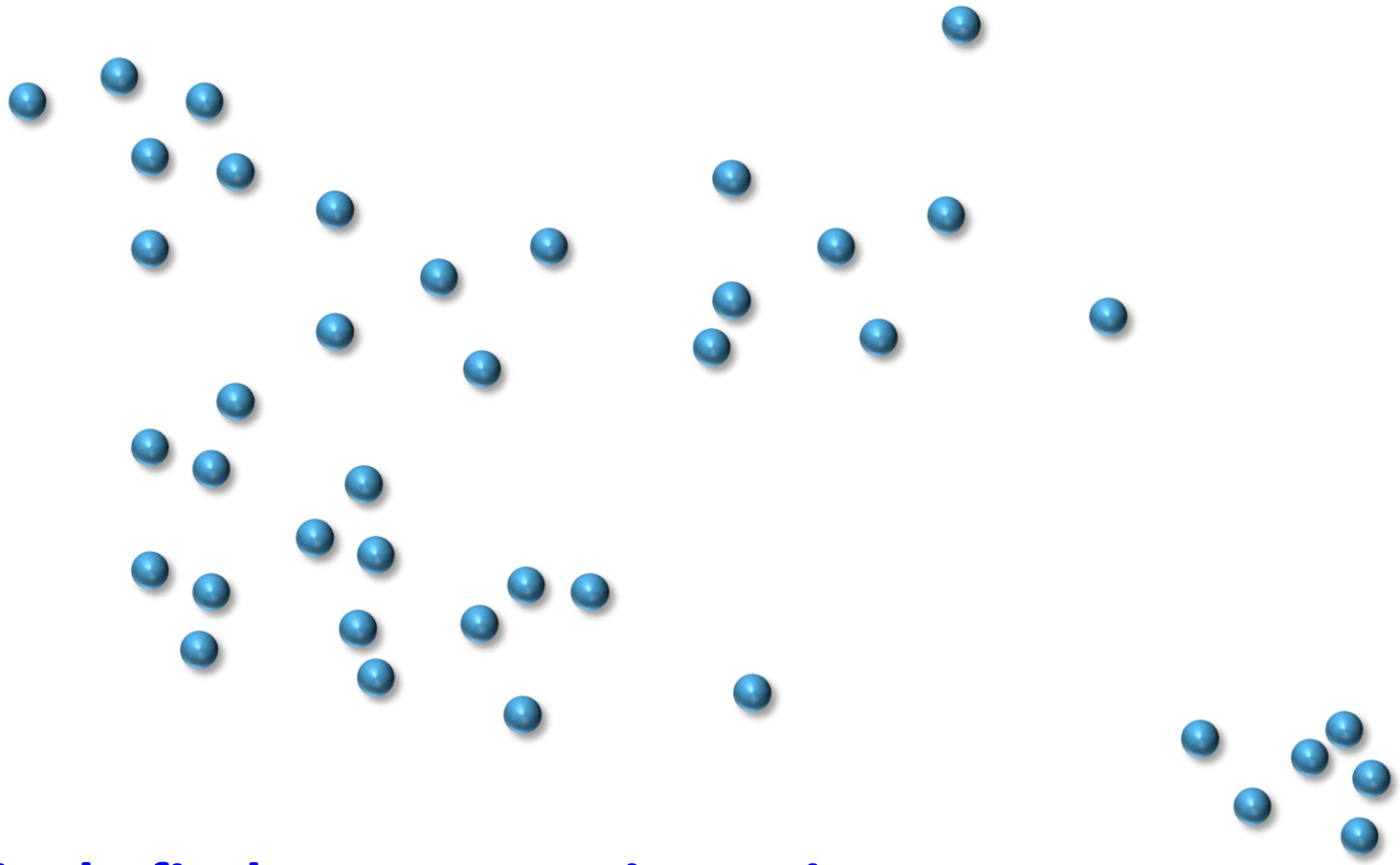
Sampling distribution



$\gamma(x)$

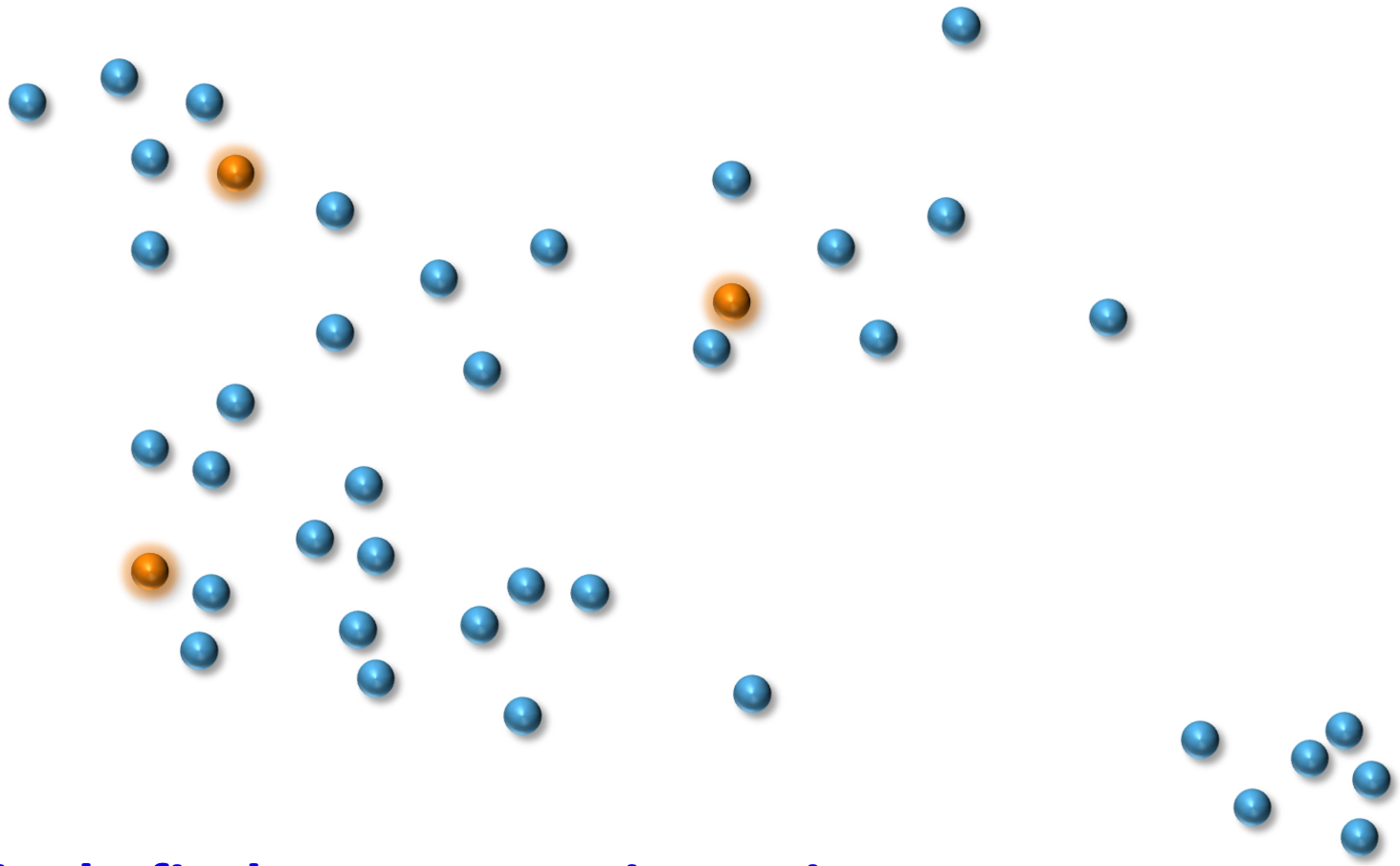
Weights $\gamma \propto \frac{1}{q}$

Creating a Sampling Distribution



Iteratively find representative points

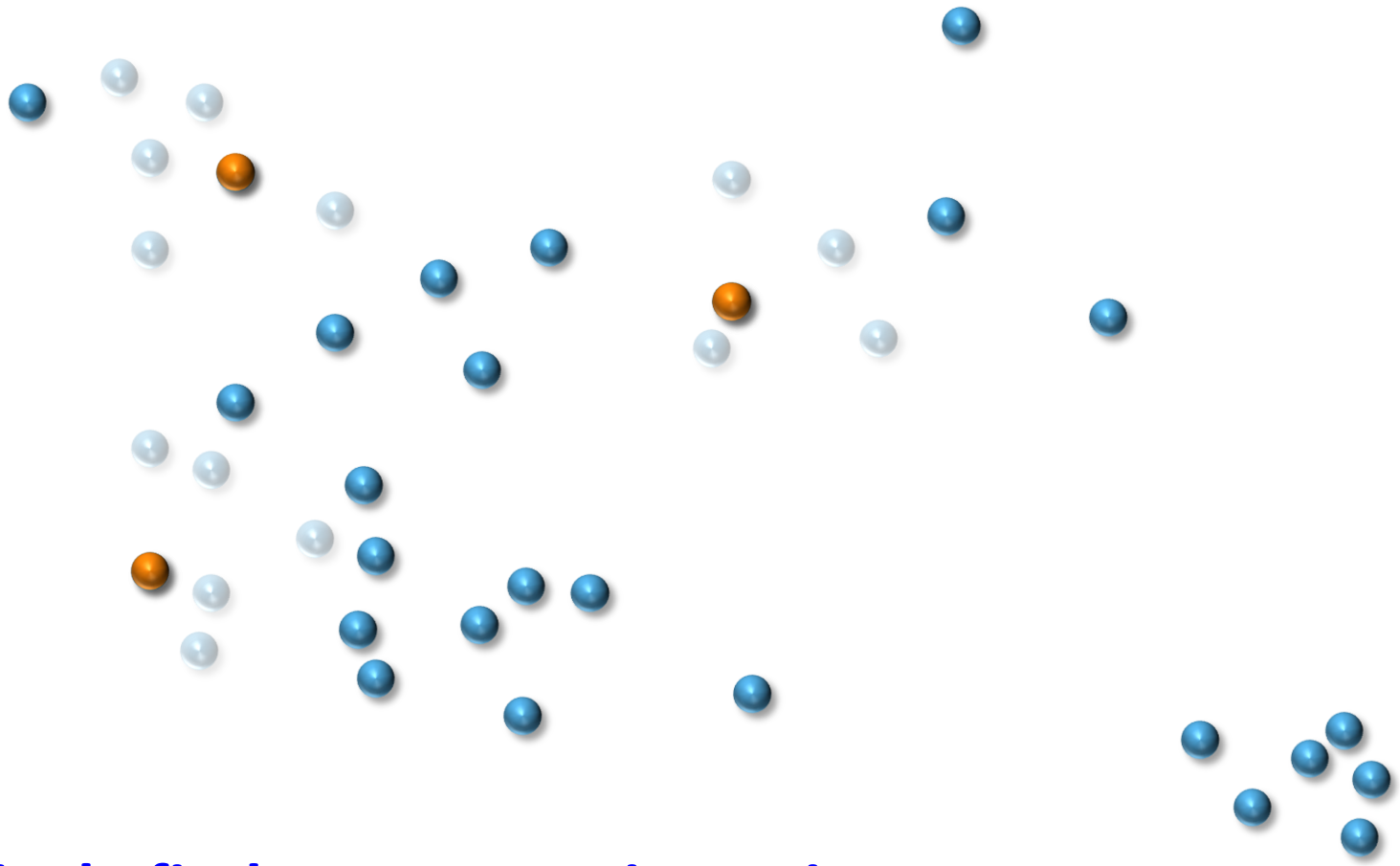
Creating a Sampling Distribution



Iteratively find representative points

- Sample a small set uniformly at random

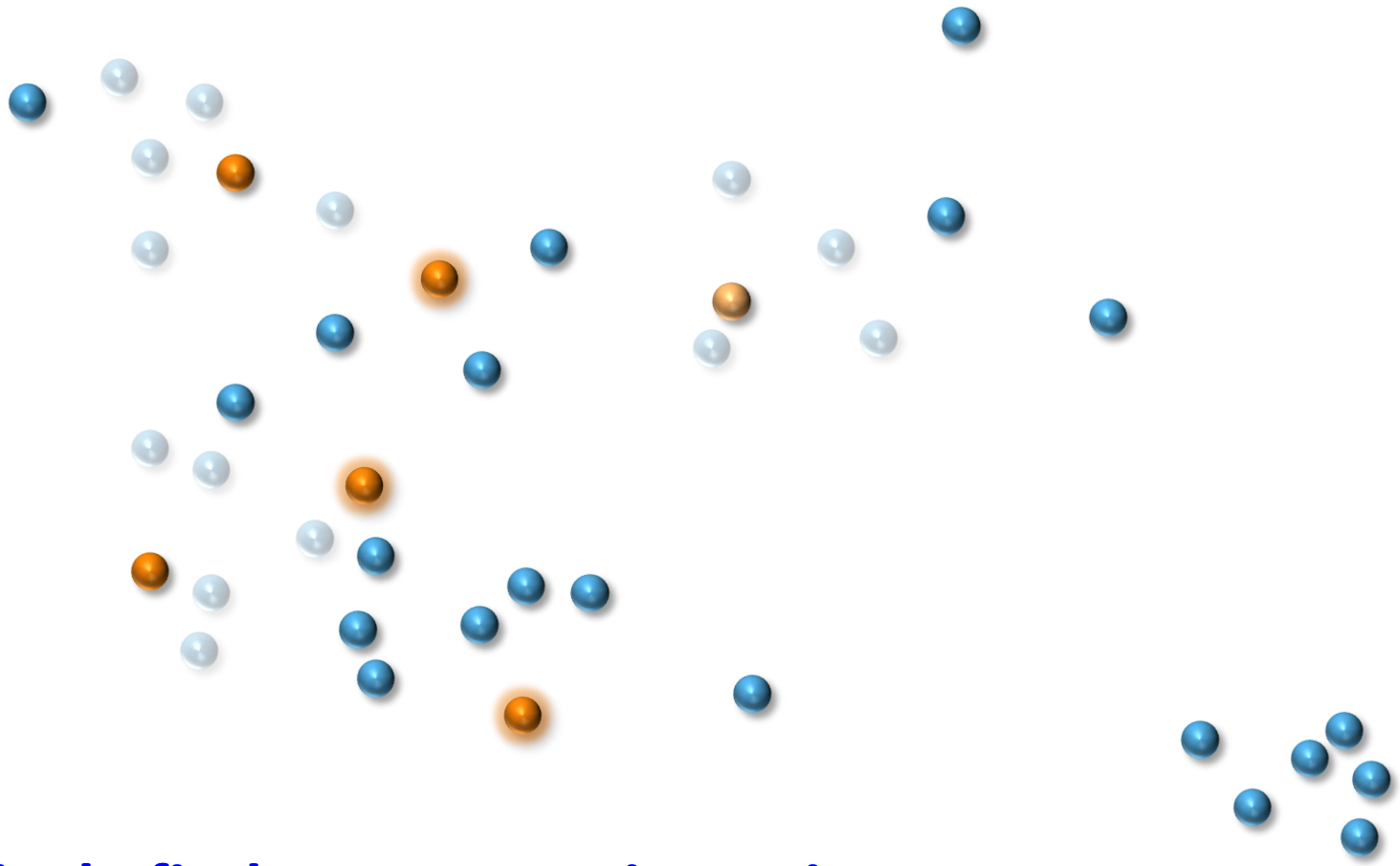
Creating a Sampling Distribution



Iteratively find representative points

- Sample a small set uniformly at random
- Remove half the blue points nearest the samples

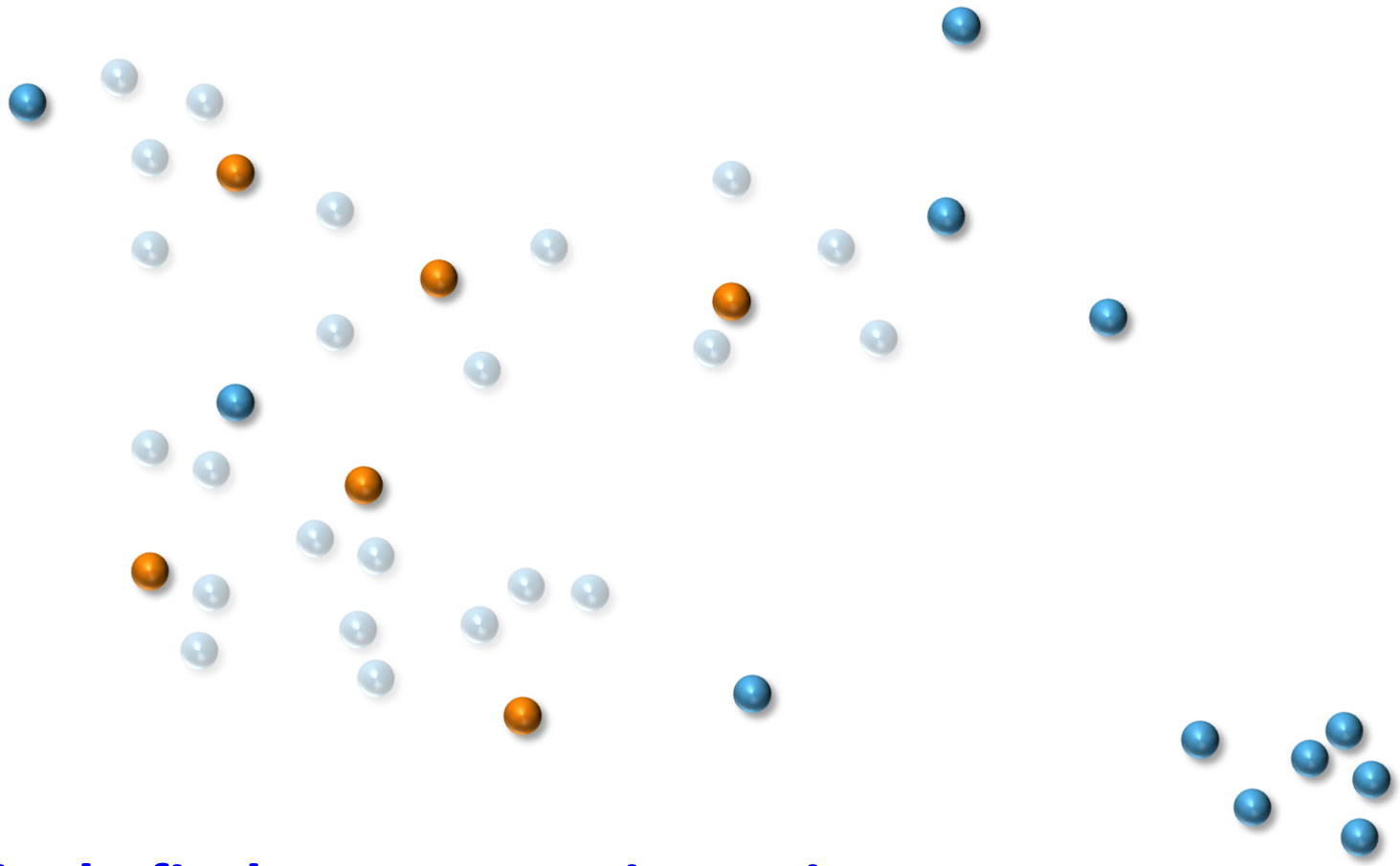
Creating a Sampling Distribution



Iteratively find representative points

- Sample a small set uniformly at random
- Remove half the blue points nearest the samples

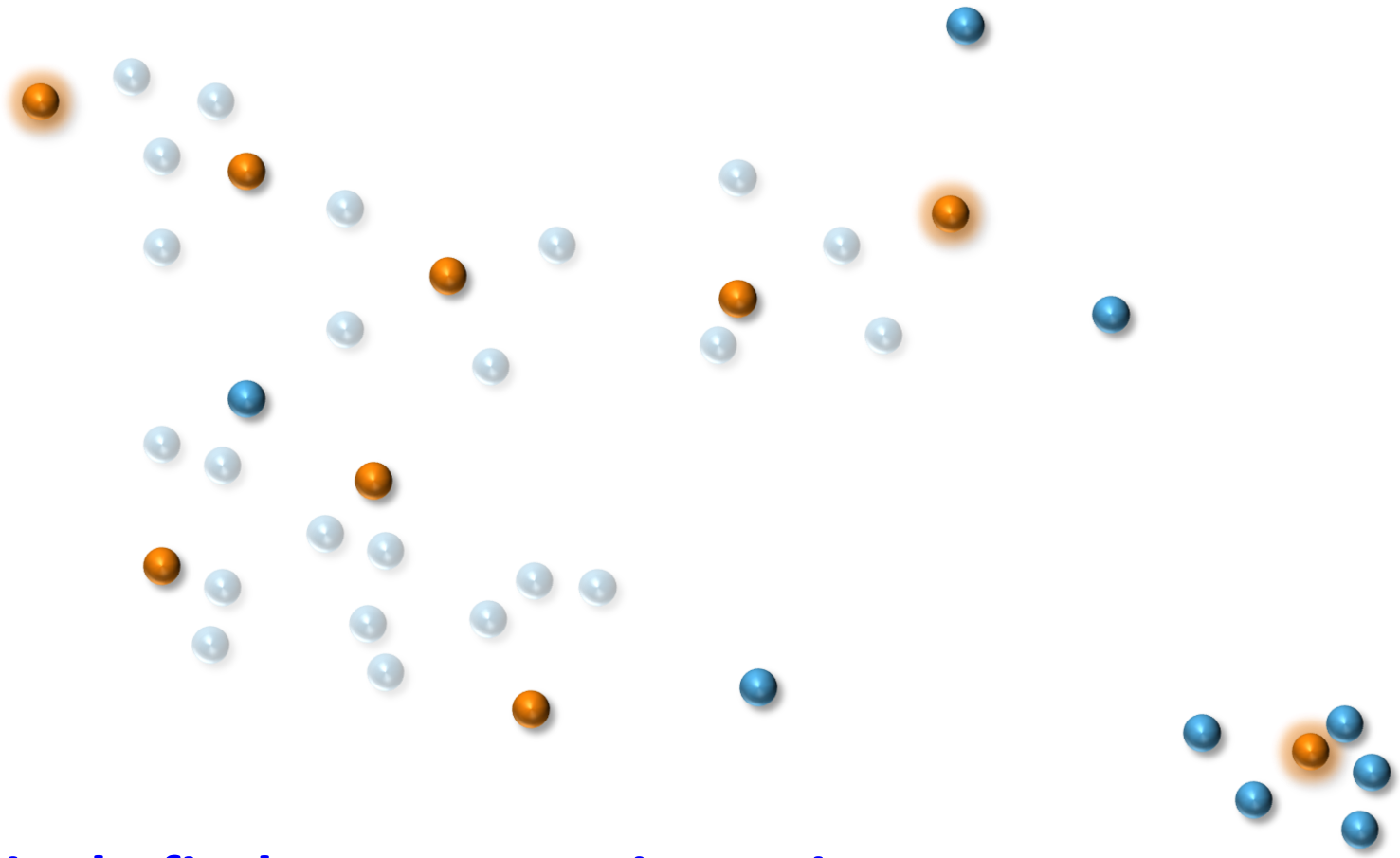
Creating a Sampling Distribution



Iteratively find representative points

- Sample a small set uniformly at random
- Remove half the blue points nearest the samples

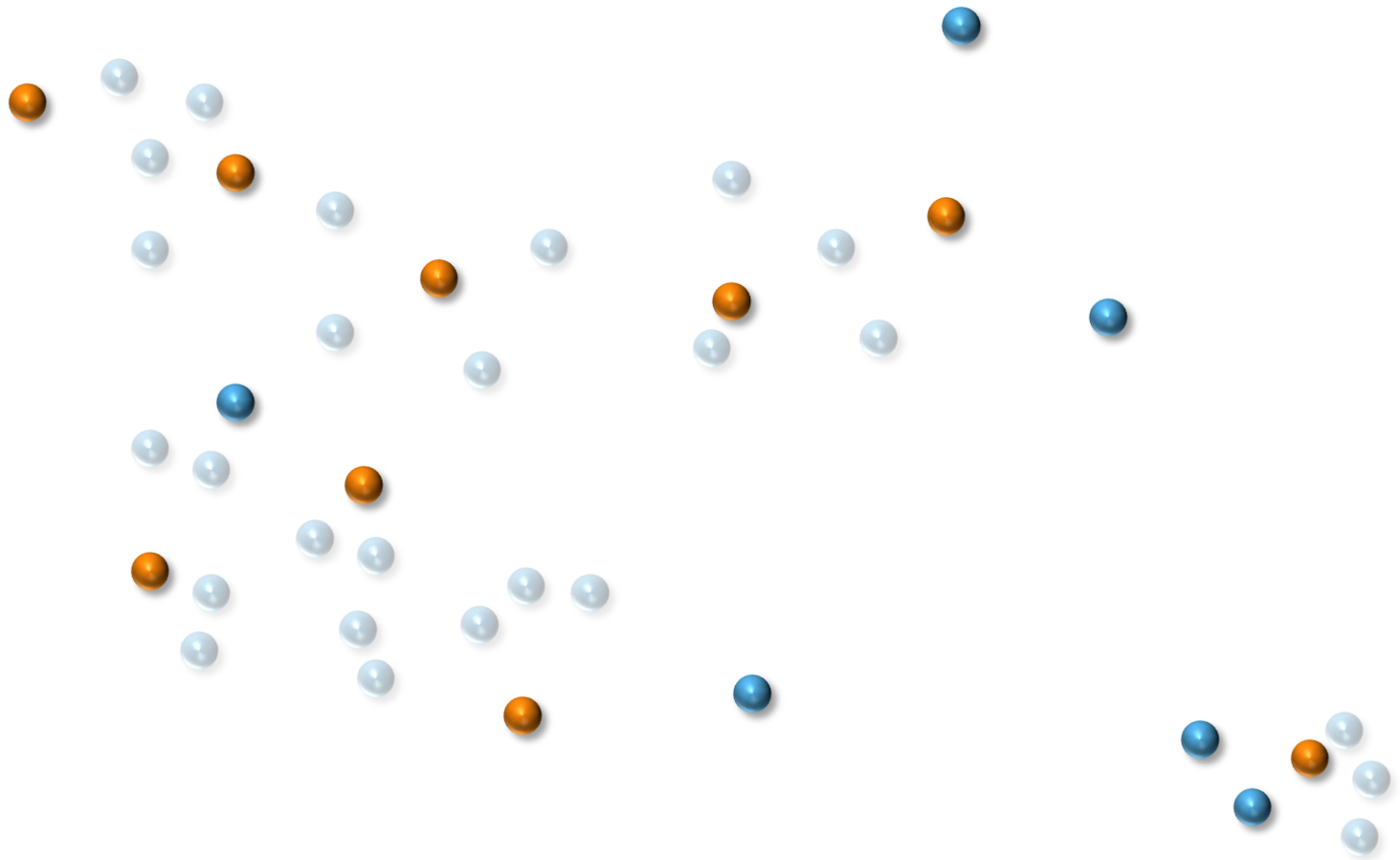
Creating a Sampling Distribution



Iteratively find representative points

- Sample a small set uniformly at random
- Remove half the blue points nearest the samples

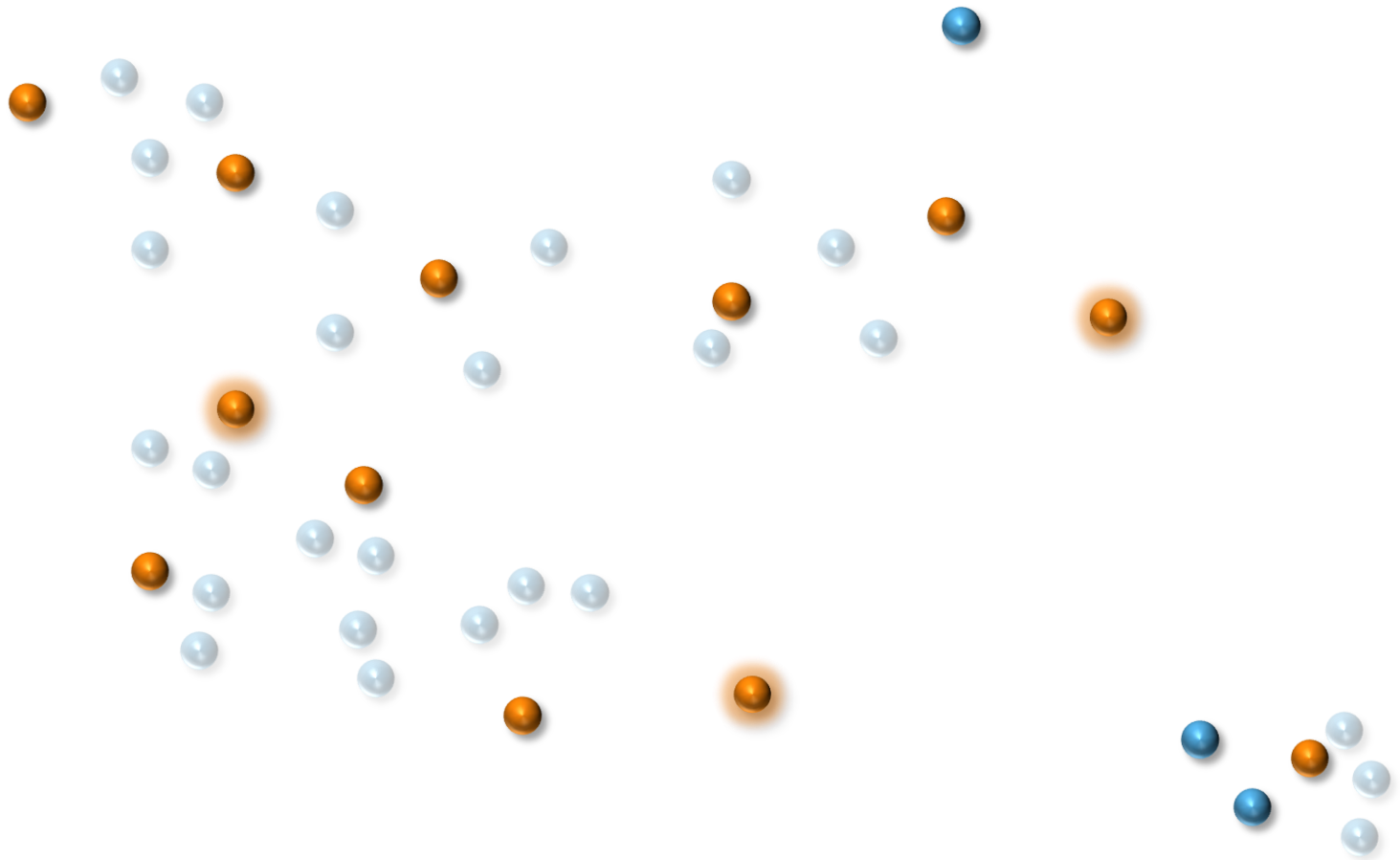
Creating a Sampling Distribution



Iteratively find representative points

- Sample a small set uniformly at random
- Remove half the blue points nearest the samples

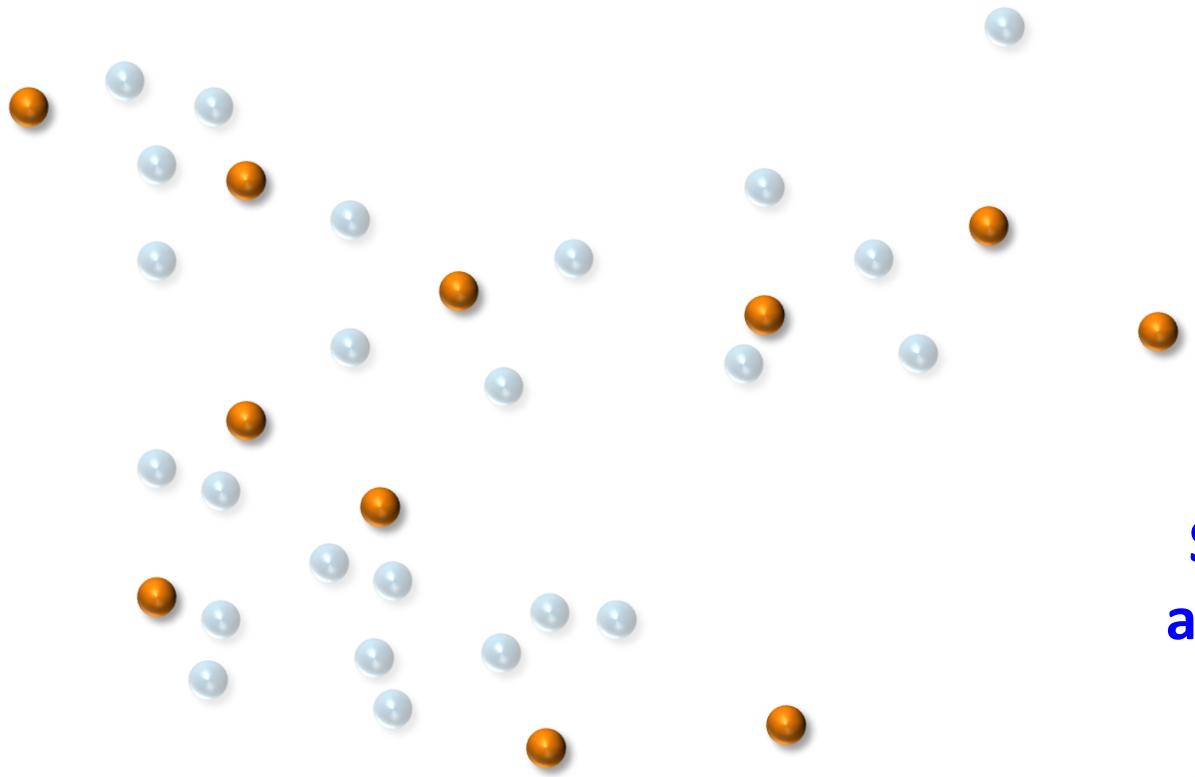
Creating a Sampling Distribution



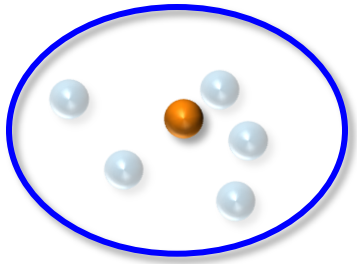
Iteratively find representative points

- Sample a small set uniformly at random
- Remove half the blue points nearest the samples

Creating a Sampling Distribution



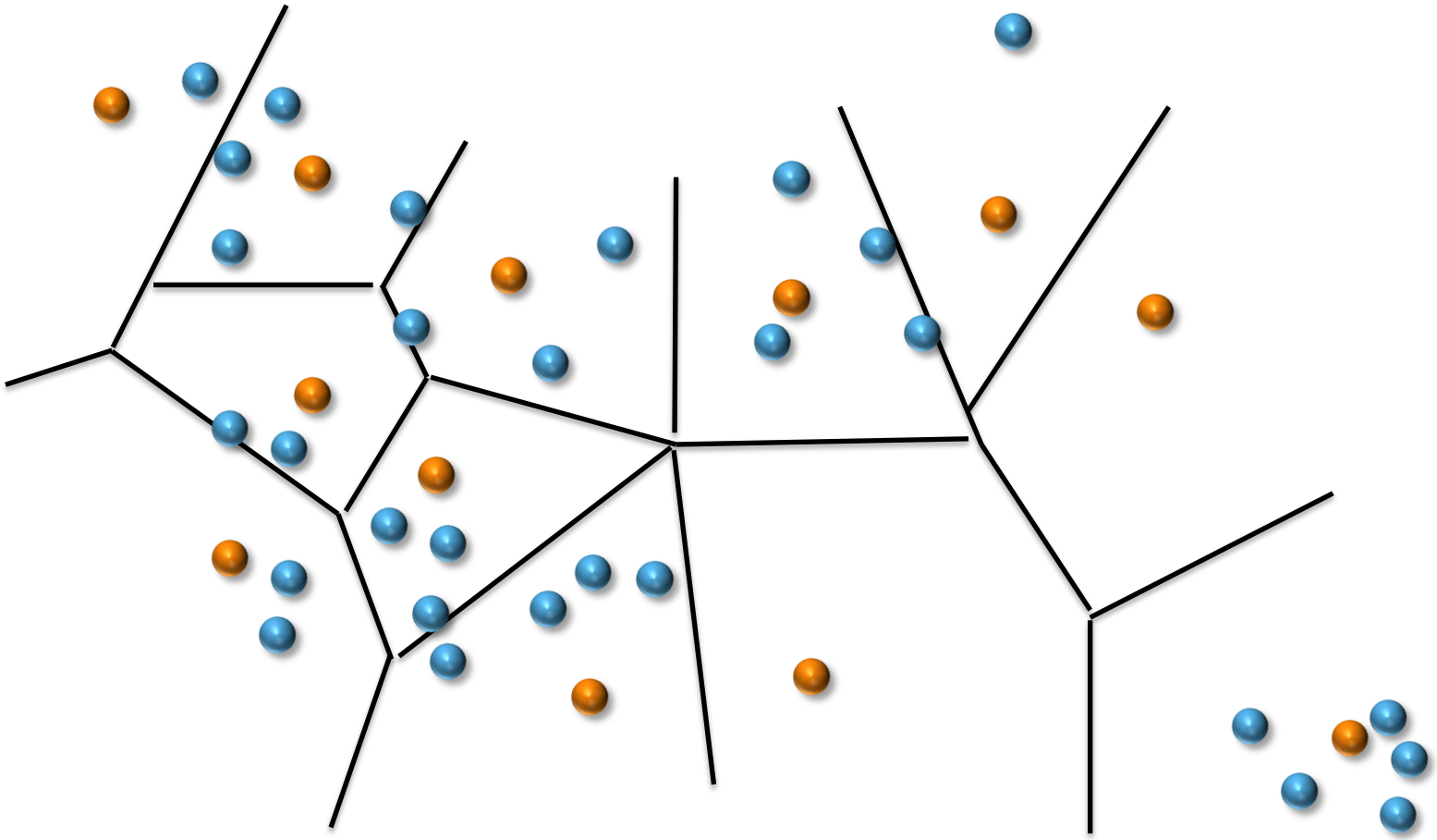
Small clusters are represented



Iteratively find representative points

- Sample a small set uniformly at random
- Remove half the blue points nearest the samples

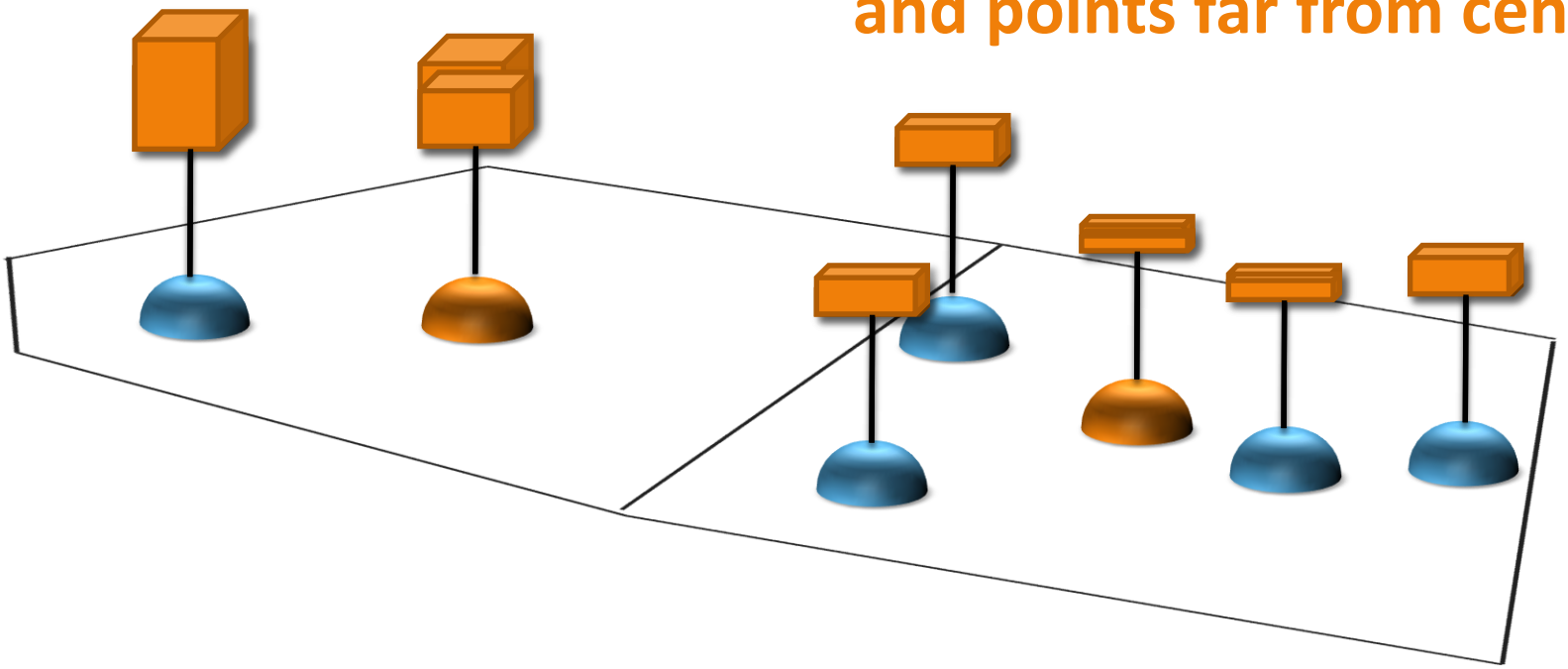
Creating a Sampling Distribution

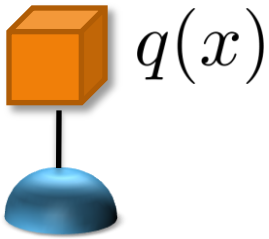


Partition data via a Voronoi diagram centered at points

Creating a Sampling Distribution

Points in sparse cells get more mass and points far from centers

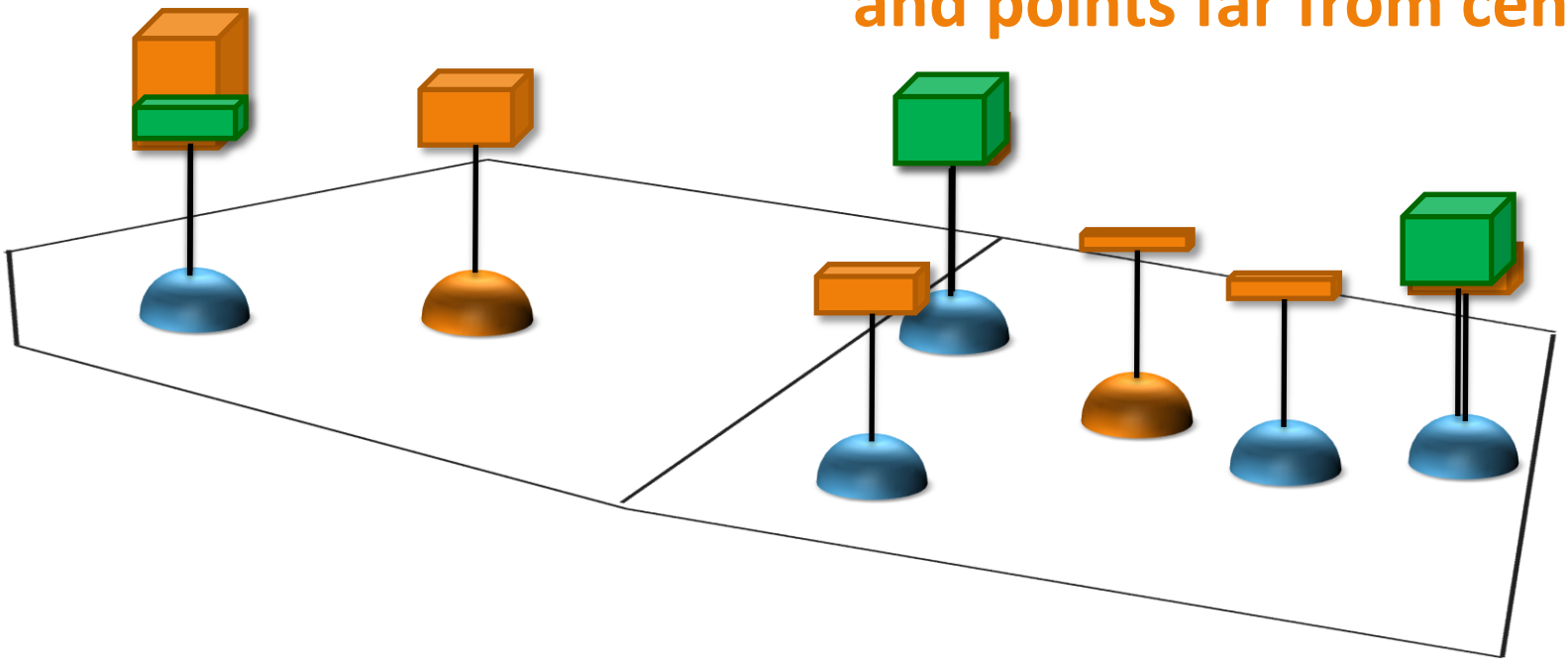


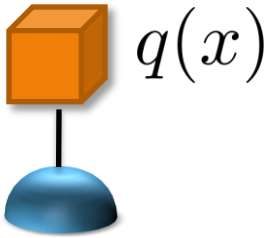


Sampling distribution q

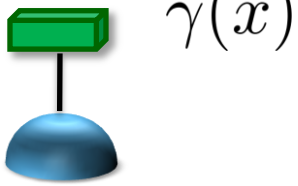
Importance Weights

Points in sparse cells get more mass and points far from centers



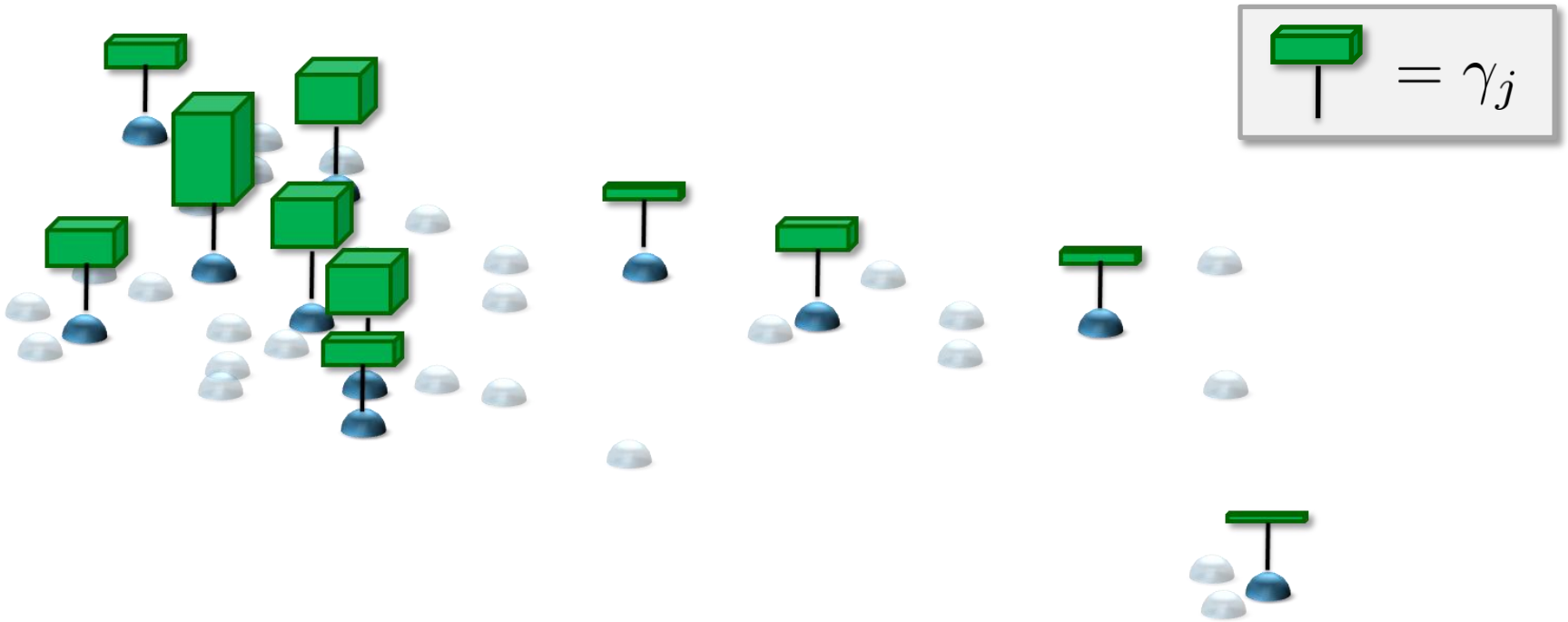


Sampling distribution q



Weights $\gamma \propto \frac{1}{q}$

Importance Sample



$$\mathcal{L}(\theta|C) = -\sum_{j \in U} \gamma_j \ln P(x_j|\theta)$$

Unbiased, but low variance?

Coresets via Adaptive Sampling

$B \leftarrow \emptyset \quad D' \leftarrow D$

while $D' \neq \emptyset$

- $S \leftarrow$ uniformly sample $10dk \ln(\frac{1}{\epsilon})$ points from D'
  Remove $\frac{|D'|}{2}$ points nearest to S from D'
  $B \leftarrow B \cup S$

Partition D into Voronoi cells D_b centered at $b \in B$

$$q(x) \propto \left\lceil \frac{5}{|D_b|} + \frac{\mathrm{dist}(x,B)^2}{\sum_{x'} \mathrm{dist}(x',B)^2} \right\rceil, \quad \gamma(x) = \frac{1}{|C|q(x)}$$

$$C \leftarrow \text{sample } 10\lceil dk \log^2 n \log(\tfrac{1}{\delta})/\epsilon^2 \rceil \text{ from } D \text{ via } q$$

Thm. (C, γ) is a (k, ϵ)-coreset for semi-spherical GMMs whose covariance matrices have bounded eigenvalues $\lambda_{min} \leq \lambda_i \leq \lambda_{max}$

A General Coreset Framework

Use a general coreset framework for additive costs $\text{cost}(D|\theta) = \sum_{x \in D} \text{cost}(x|\theta)$ [Feldman & Langberg '11]

Contributions for Mixture Models:

Coreset size depends on *pseudo-dimensionality* of cost functions $F = \{\text{cost}(x|\cdot),\ x \in D\}$

New complexity result for pseudo-dimensionality of mixtures of k Gaussians $\in \mathbb{R}^d$

Efficient coreset constructions for specific cost functions

Efficient algorithm for GMM coresets via reduction to computational geometry

A Geometric Perspective

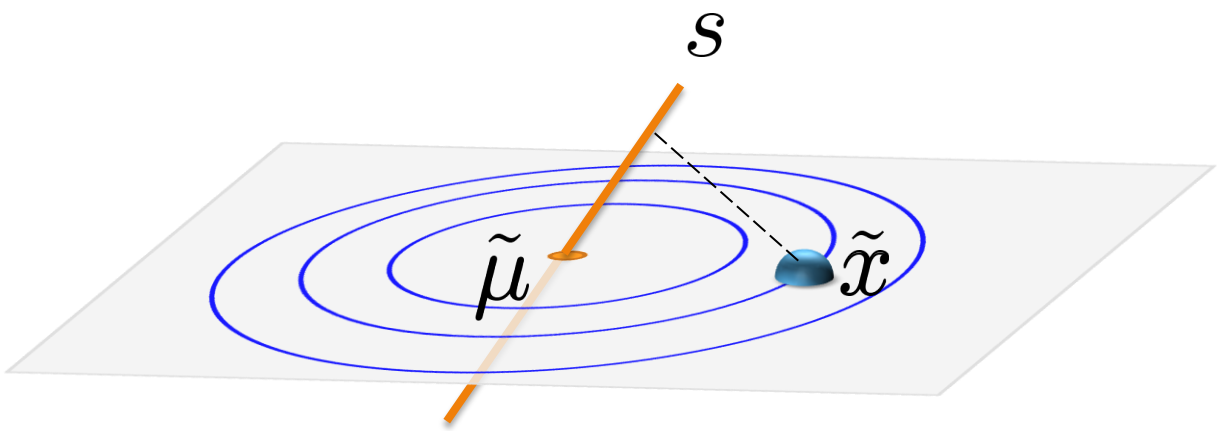
Gaussian level sets can be expressed purely geometrically:

$$\mathcal{N}(x;\mu,\Sigma) = \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

$$= \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-W\,\mathrm{dist}(\tilde{x},\mathbf{s})^2\right)$$

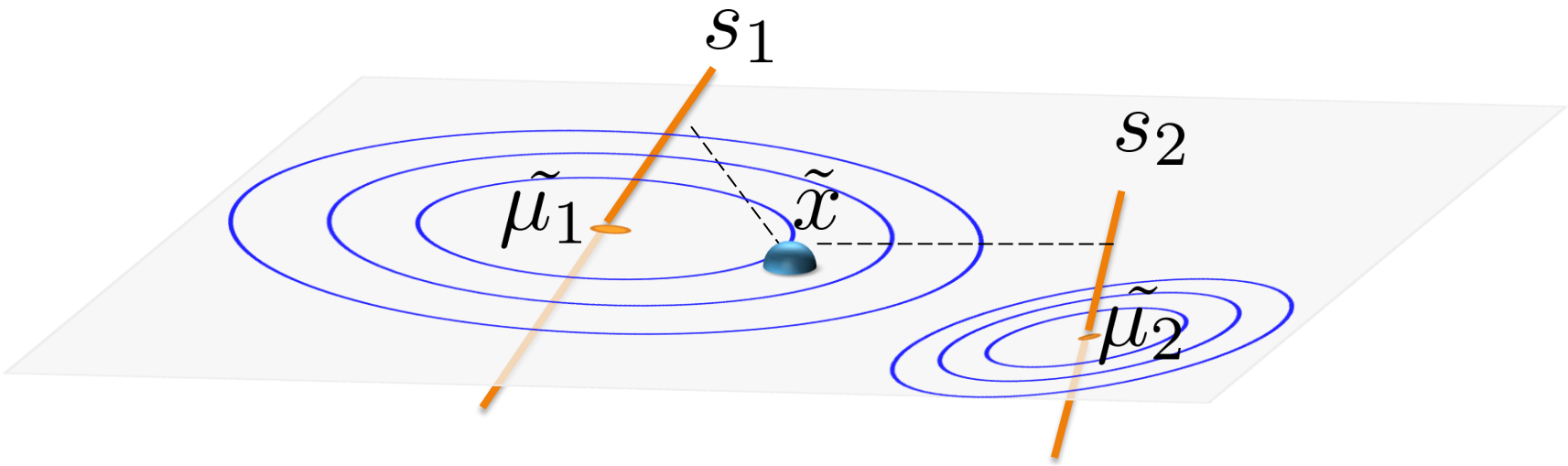
affine subspace

$\mathbf{s} = \mathbf{s}(\mu,\Sigma) \subset \mathbb{R}^{2d}$



$\tilde{x}, \tilde{\mu} \in \mathbb{R}^{2d}$

Geometric Reduction



$$\ln P(x|\theta) \propto -\underbrace{\ln \textstyle\sum_{i=1}^{k} w_i' \exp\left(-W_i \mathrm{dist}(\tilde{x}, \mathbf{s_i})^2\right)}_{\text{Soft-min}}$$

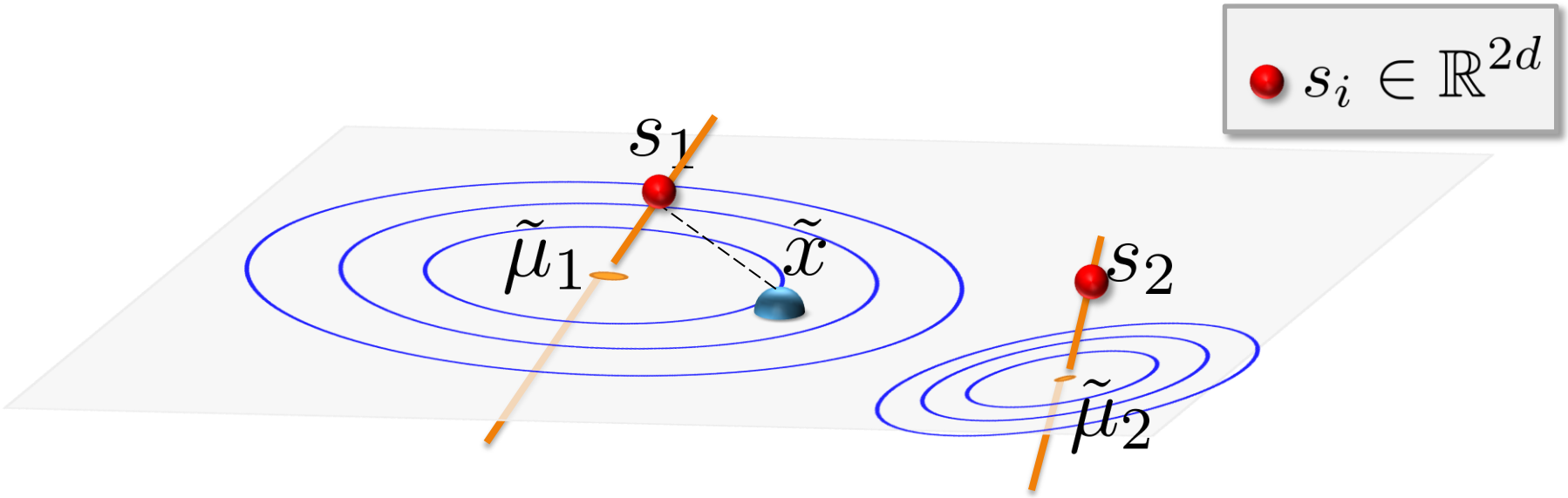
$\ln P(x|\theta) \geq \min_i W_i \mathrm{dist}(\tilde{x}, \mathbf{s_i})$ Projective Clustering!

Bound using generalized $\triangle$-inequality

Lifts geometric coreset tools to mixture models

Semi-Spherical Gaussian Mixtures

Subspaces $\mathbf{s_i}$ can be chosen as points for Semi-spherical GMMs (covariance eigenvalues $\lambda_{\text{min}} \leq \lambda_i \leq \lambda_{\text{max}}$)



Thm. An ϵ-coreset for k-means in the transformed space gives a $(k, \epsilon\lambda_{max}^2/\lambda_{min}^2)$-coreset for semi-spherical GMMs

$$(1 - \epsilon\frac{\lambda_{\text{max}}^2}{\lambda_{\text{min}}^2})\mathcal{L}(\theta|D) \leq \mathcal{L}(\theta|C) \leq (1 + \epsilon\frac{\lambda_{\text{max}}^2}{\lambda_{\text{min}}^2})\mathcal{L}(\theta|D) \text{ w.h.p}$$

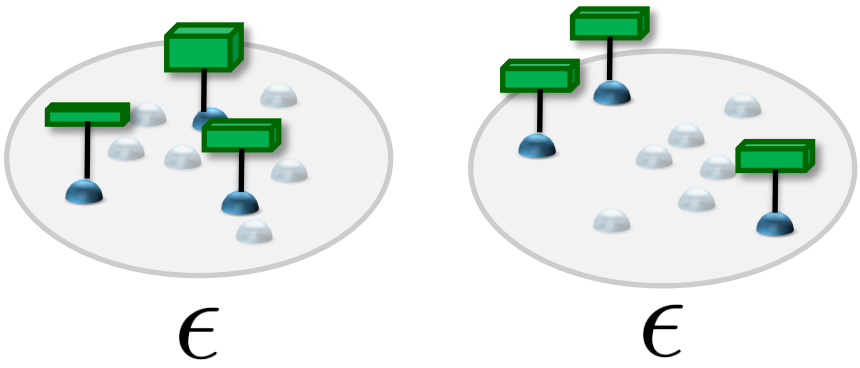
Extensions and Generalizations

- Coresets for **non-spherical GMMs** can be obtained via reduction to recent projective clustering coresets
- Other mixtures (e.g. Laplace) based on ℓ_q **distances** and **other norms** via generalized $\triangle$-inequality
- Coresets of size **independent of n**
- Efficient implementations in **Parallel (MapReduce)** and **Streaming** settings

Composition of Coresets

[c.f. Har-Peled, Mazumdar 04]

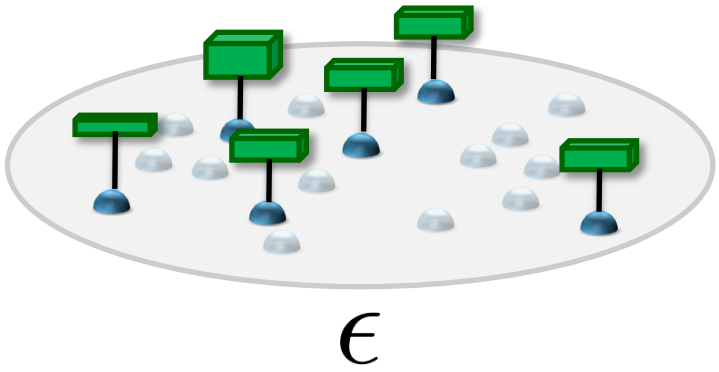
Merge The union of two (k, ϵ)-coresets is a (k, ϵ)-coreset.



Composition of Coresets

[c.f. Har-Peled, Mazumdar 04]

Merge The union of two (k, ϵ)-coresets is a (k, ϵ)-coreset.



ϵ

Composition of Coresets

[Har-Peled, Mazumdar 04]

Merge The union of two (k, ϵ)-coresets is a (k, ϵ)-coreset.

Compress A (k, δ)-coreset of a (k, ϵ)-coreset is a $(k, \epsilon + \delta + \epsilon\delta)$-coreset

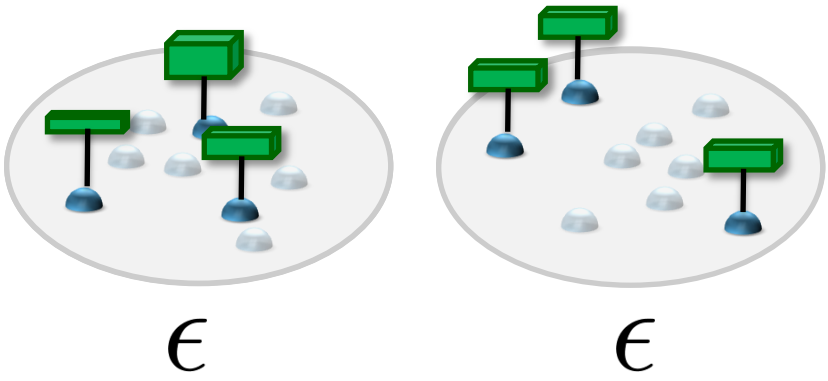
$$\epsilon + \delta + \epsilon\delta$$

Coresets on Streams

[Har-Peled, Mazumdar 04]

Merge The union of two (k, ϵ)-coresets is a (k, ϵ)-coreset.

Compress A (k, δ)-coreset of a (k, ϵ)-coreset is a $(k, \epsilon + \delta + \epsilon\delta)$-coreset

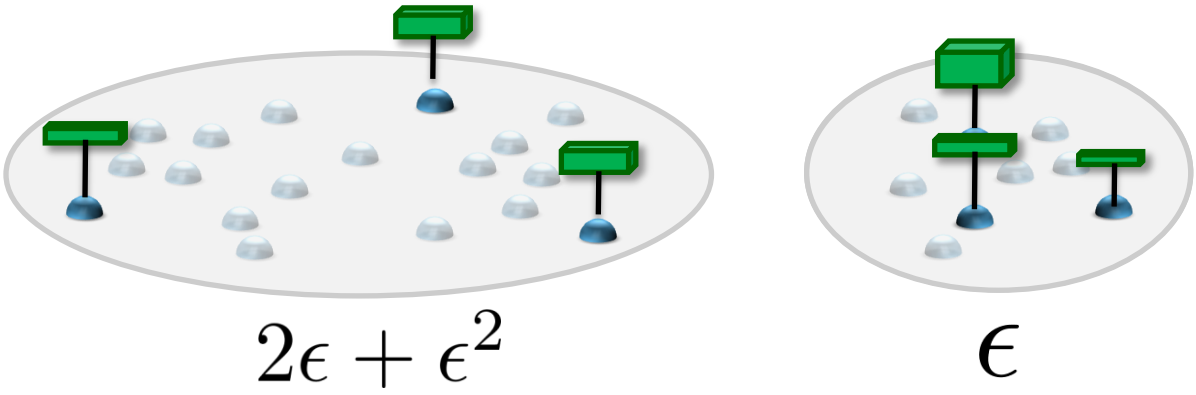


Coresets on Streams

[Har-Peled, Mazumdar 04]

Merge The union of two (k, ϵ)-coresets is a (k, ϵ)-coreset.

Compress A (k, δ)-coreset of a (k, ϵ)-coreset is a $(k, \epsilon + \delta + \epsilon\delta)$-coreset



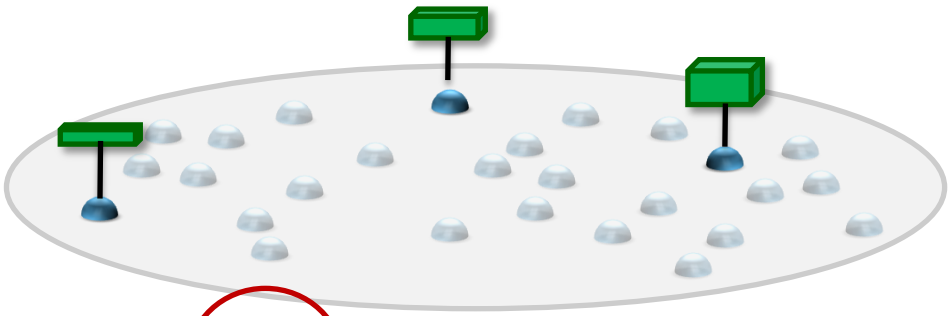
$2\epsilon + \epsilon^2$ ϵ

Coresets on Streams

[Har-Peled, Mazumdar 04]

Merge The union of two (k, ϵ)-coresets is a (k, ϵ)-coreset.

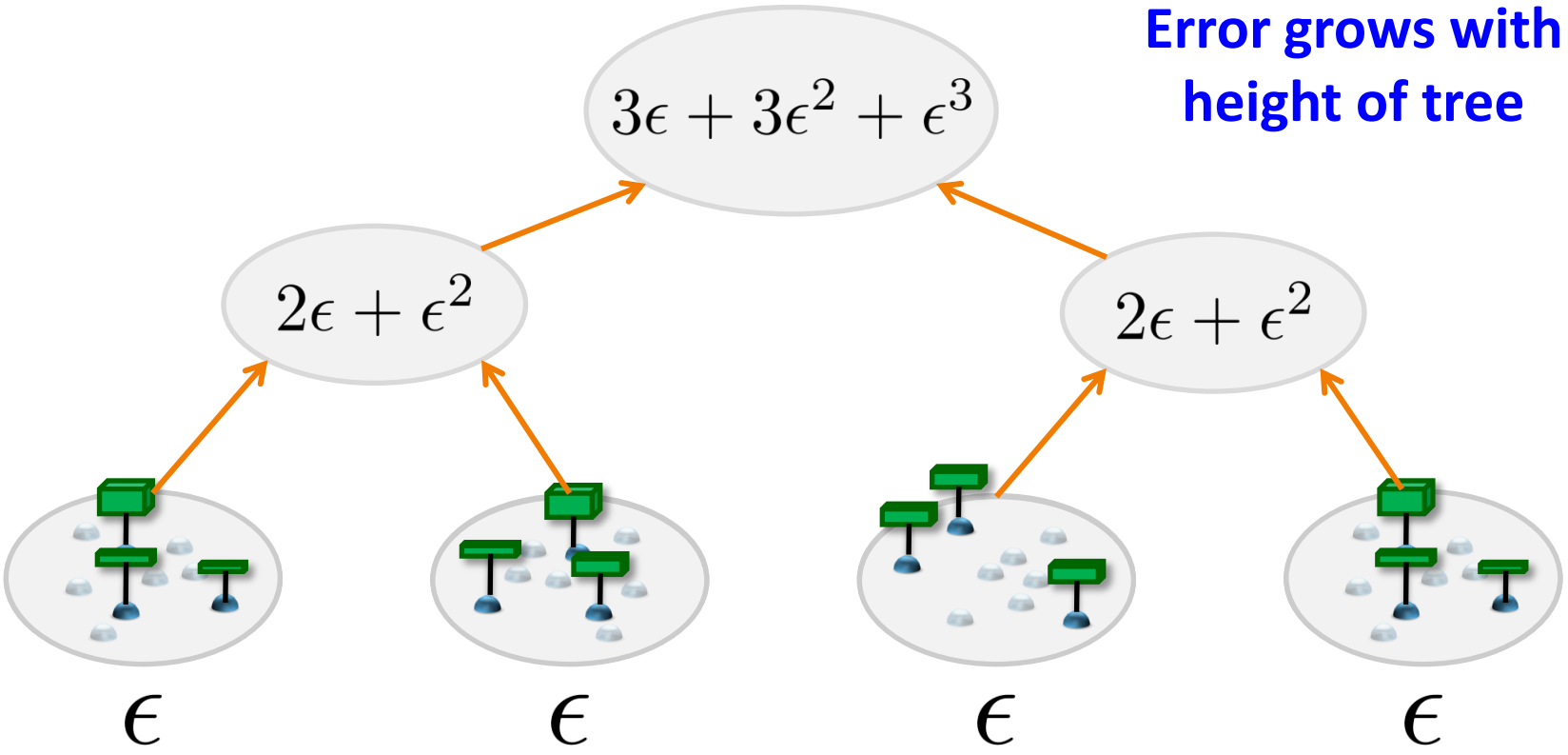
Compress A (k, δ)-coreset of a (k, ϵ)-coreset is a $(k, \epsilon + \delta + \epsilon\delta)$-coreset



$3\epsilon + 3\epsilon^2 + \ldots$

Error grows linearly with number of compressions

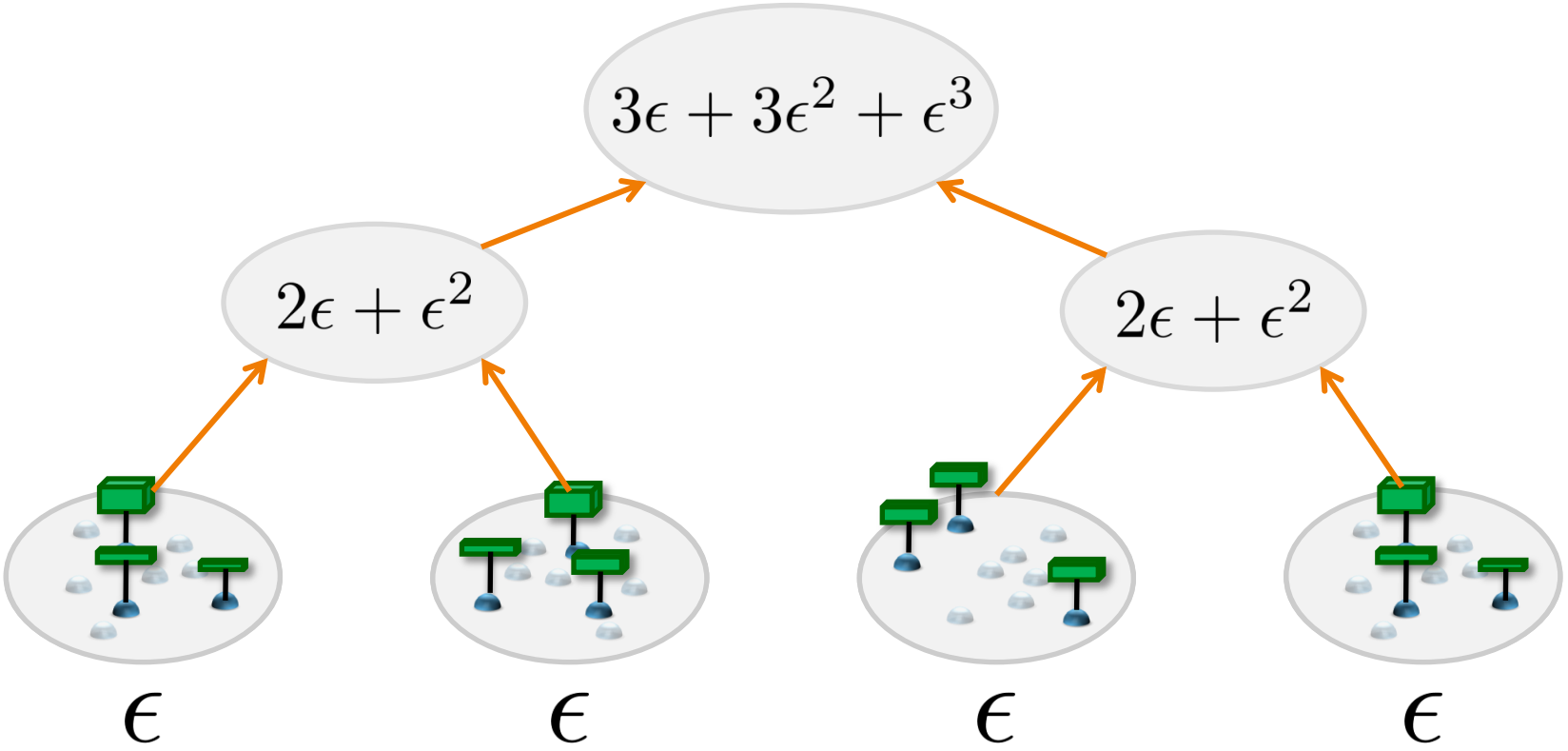
Coresets on Streams



Error grows with height of tree

THM: a (k, ϵ)-coreset for a stream of n points $\in \mathbb{R}^d$ can be computed for ϵ-semi-spherical GMM with prob. $\geq (1 - \delta)$ in space and update time $\text{poly}(dk\epsilon^{-1} \log(1/\delta) \log n)$

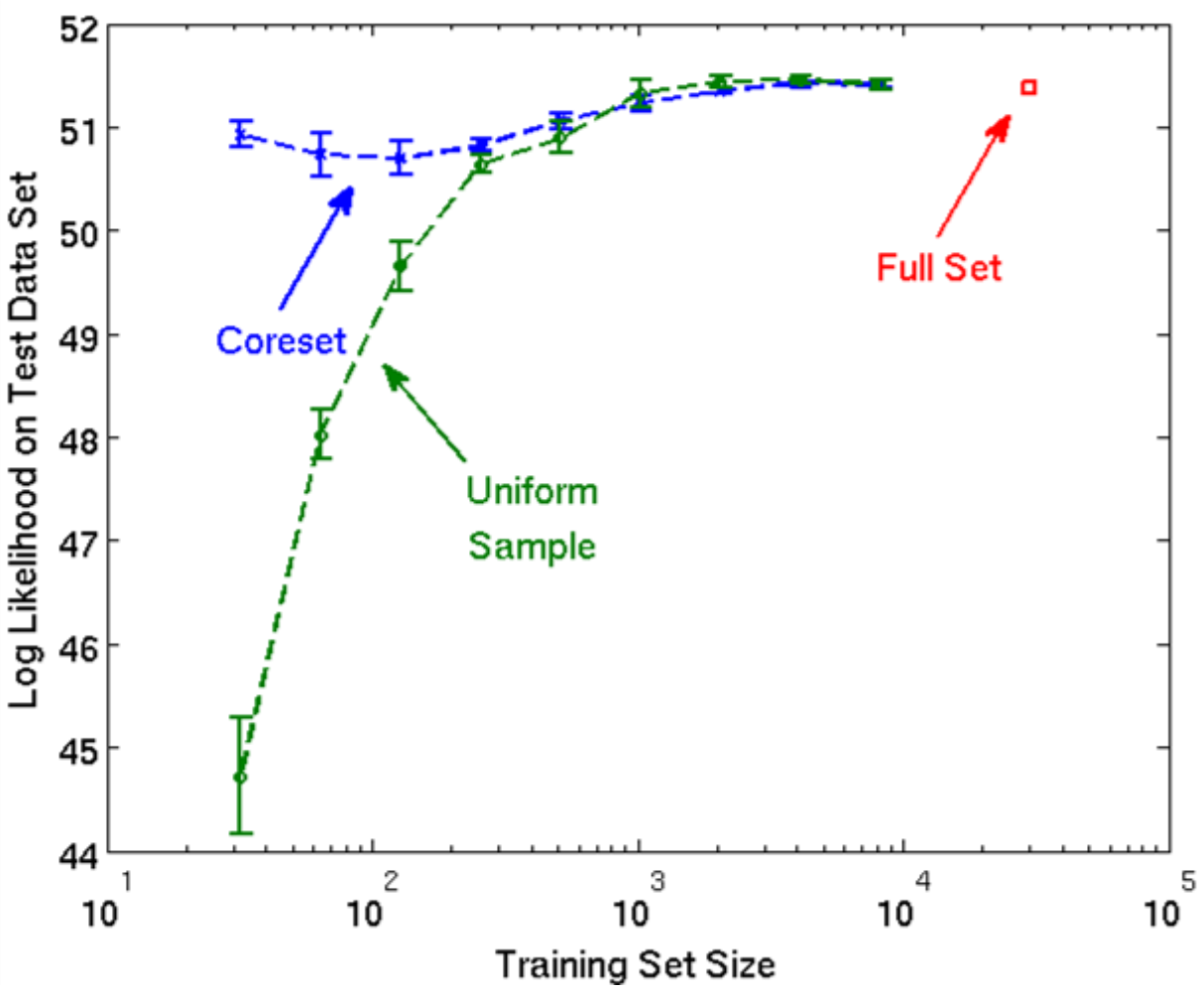
Coresets in Parallel

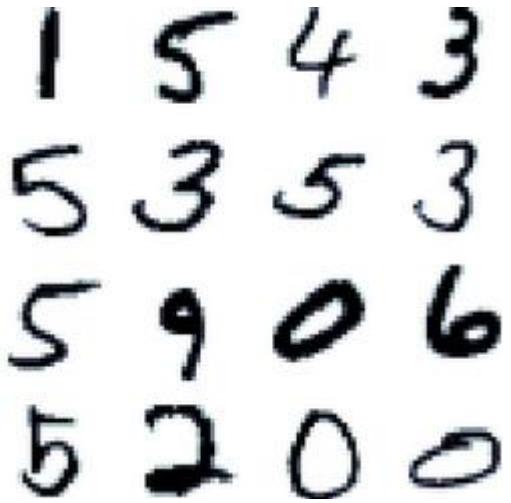


THM: a (k, ϵ)-coreset for n points $\in \mathbb{R}^d$ can be computed for ϵ-semi-spherical GMM with prob. $\geq 1 - \delta$ using **m machines** in time $(\mathbf{n/m})\text{poly}(dk\epsilon^{-1} \log(1/\delta) \log n)$

Handwritten Digits

Obtain 100-dimensional features from 28x28 pixel images via PCA. Fit GMM with k=10 components.

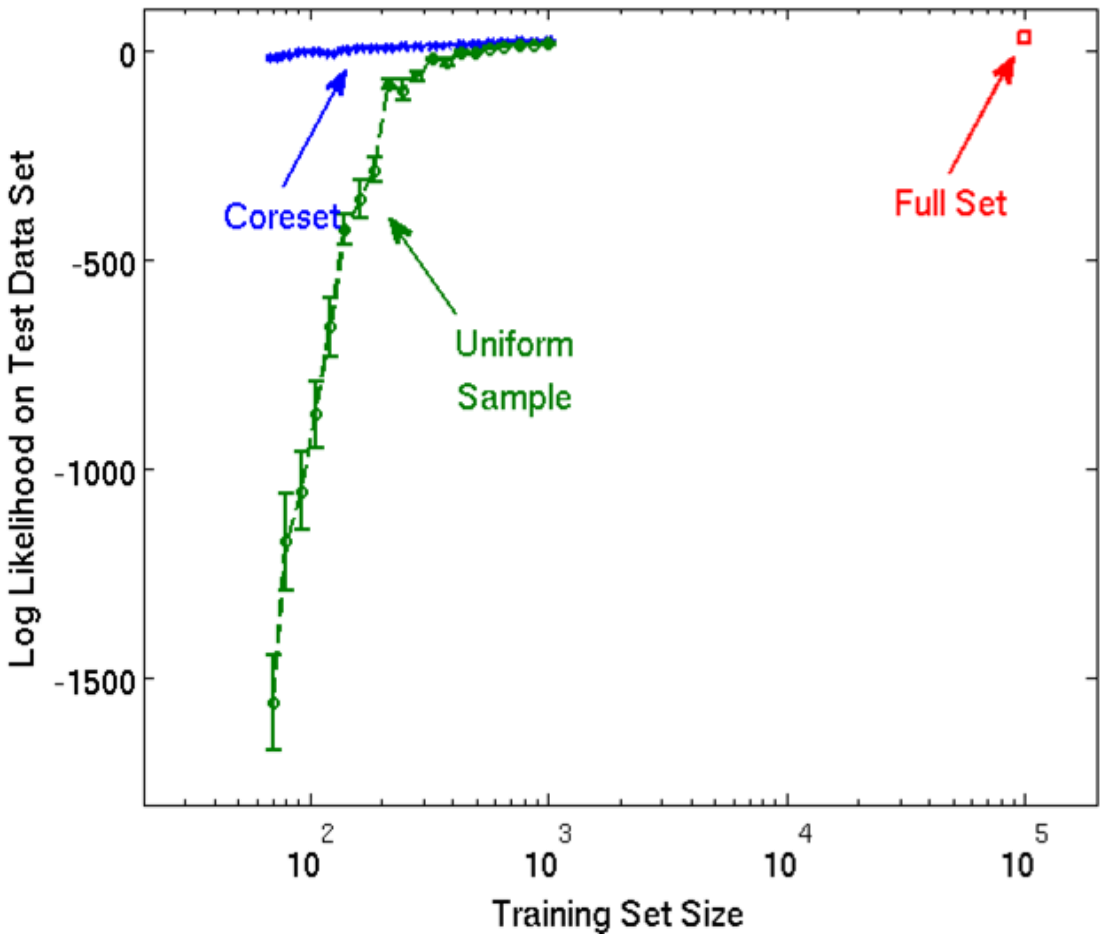


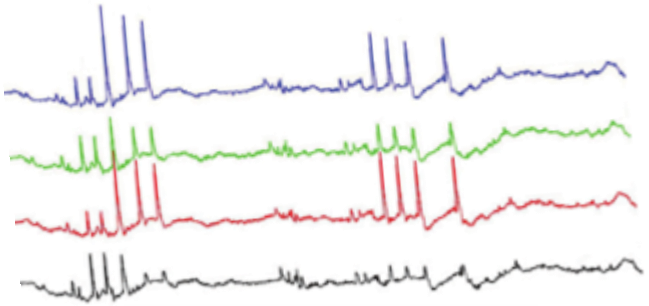


MNIST data:
60,000 training,
10,000 testing

Neural Tetrode Recordings

Waveforms of neural activity at four co-located electrodes in a live rat hippocampus. 4 x 38 samples = 152 dimensions.





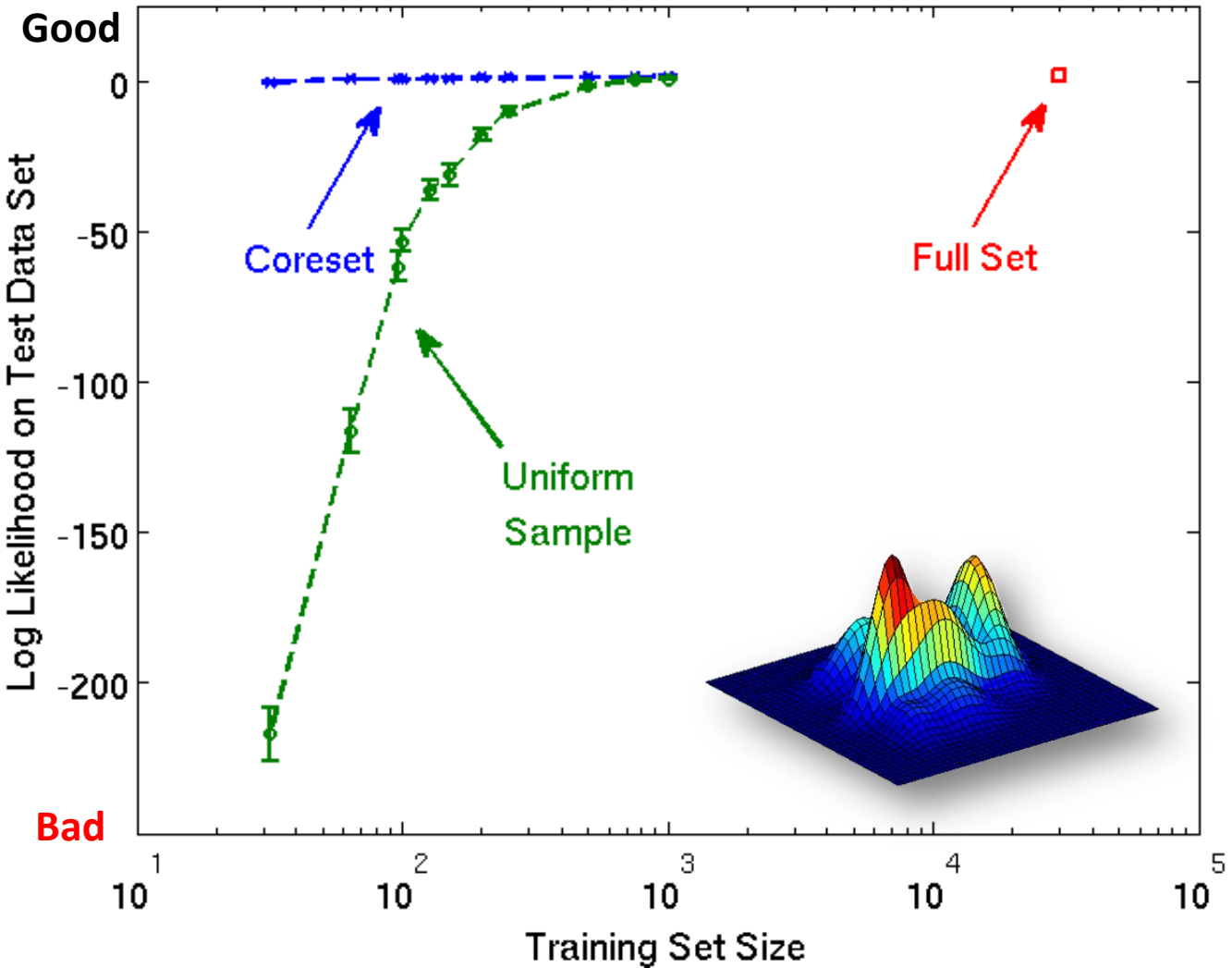
T. Siapas et al, Caltech

Community Seismic Network

Detect and monitor earthquakes using smart phones, USB sensors, and cloud computing.

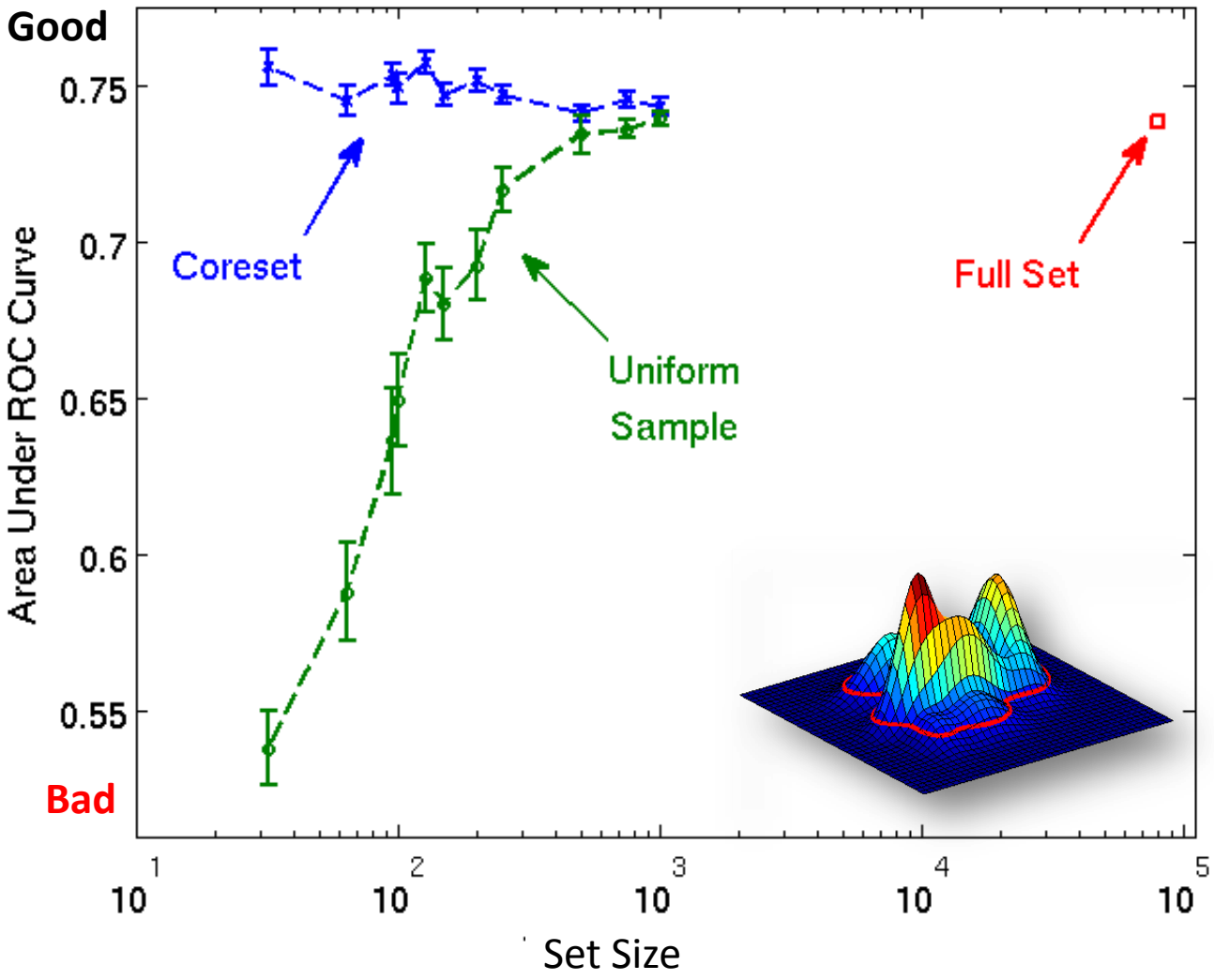
Learning User Acceleration

17-dimensional acceleration feature vectors



Seismic Anomaly Detection

GMM used for anomaly detection



Conclusions

- **GMMs admit coresets of size *independent of n***
  - Extensions for other mixture models
- **Lift geometric coreset tools to the statistical realm**
  - New complexity result for GMM level sets
- **Parallel (MapReduce) and Streaming implementations**
- **Strong empirical performance, enables learning on mobile devices**