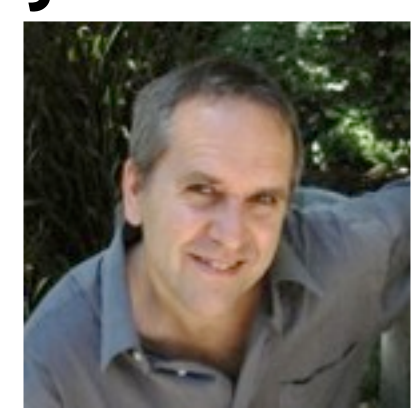
MAD-Bayes:
MAP-based Asymptotic Derivations from Bayes

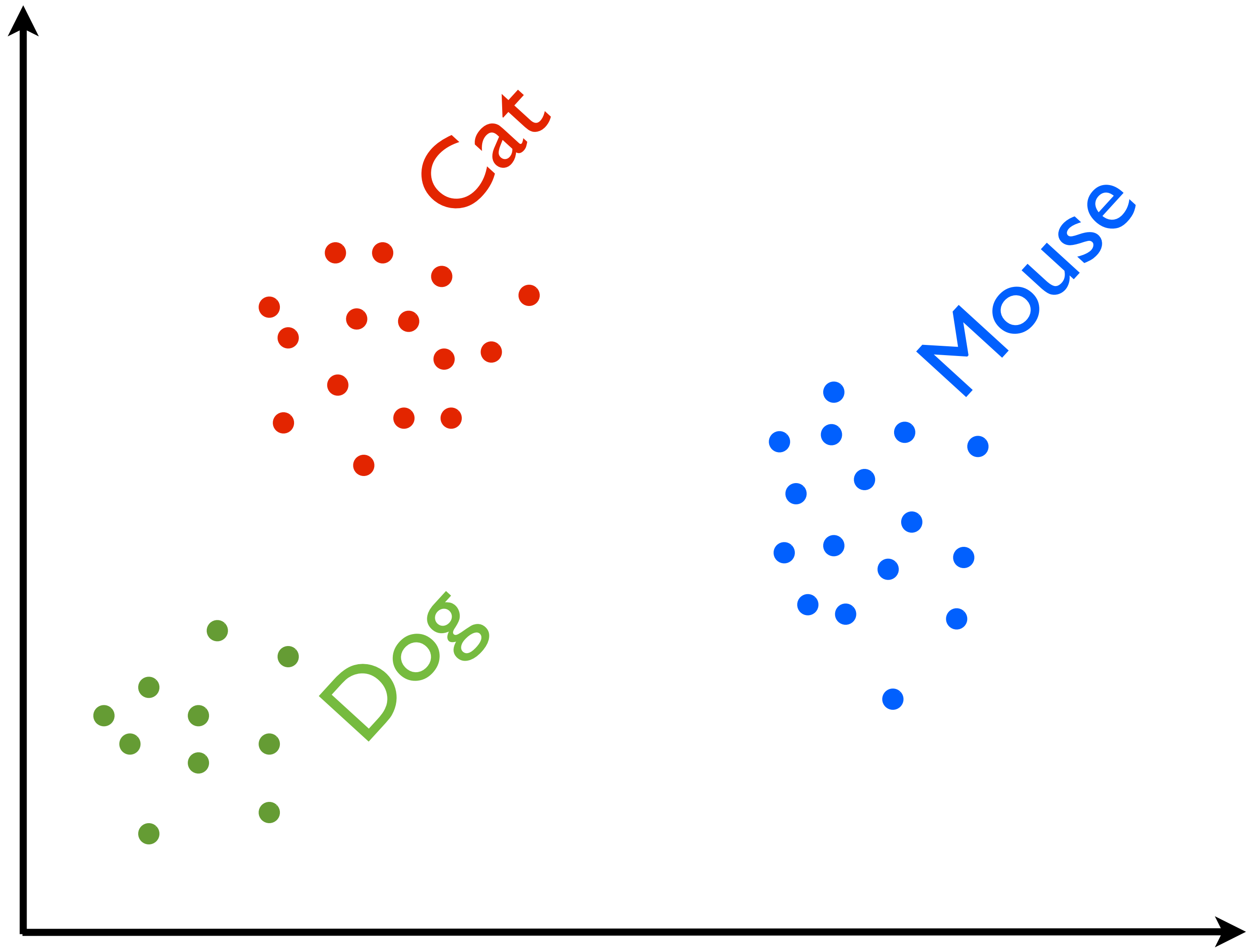
Tamara Broderick

Brian Kulis

Michael I. Jordan

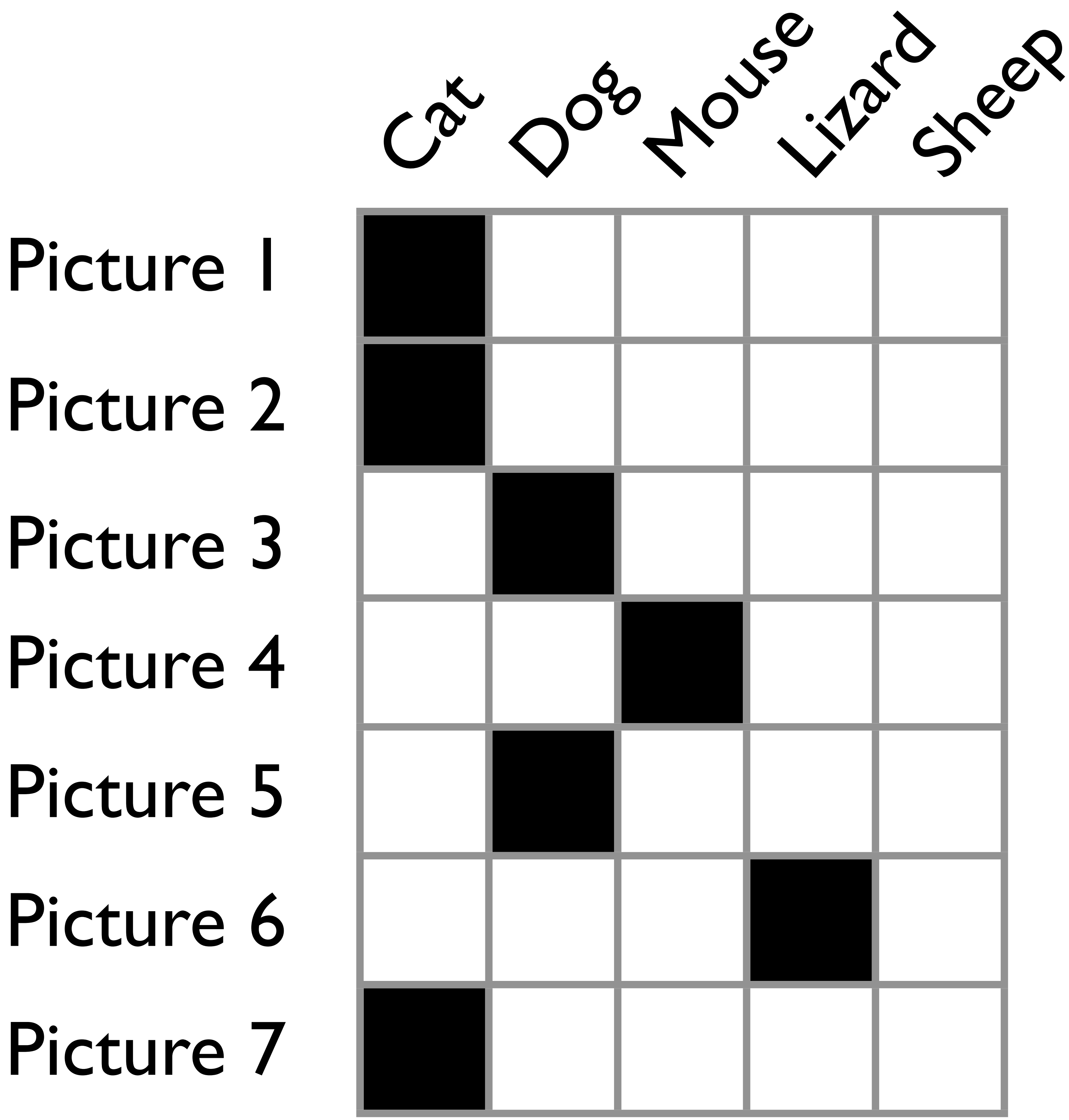


Clusters

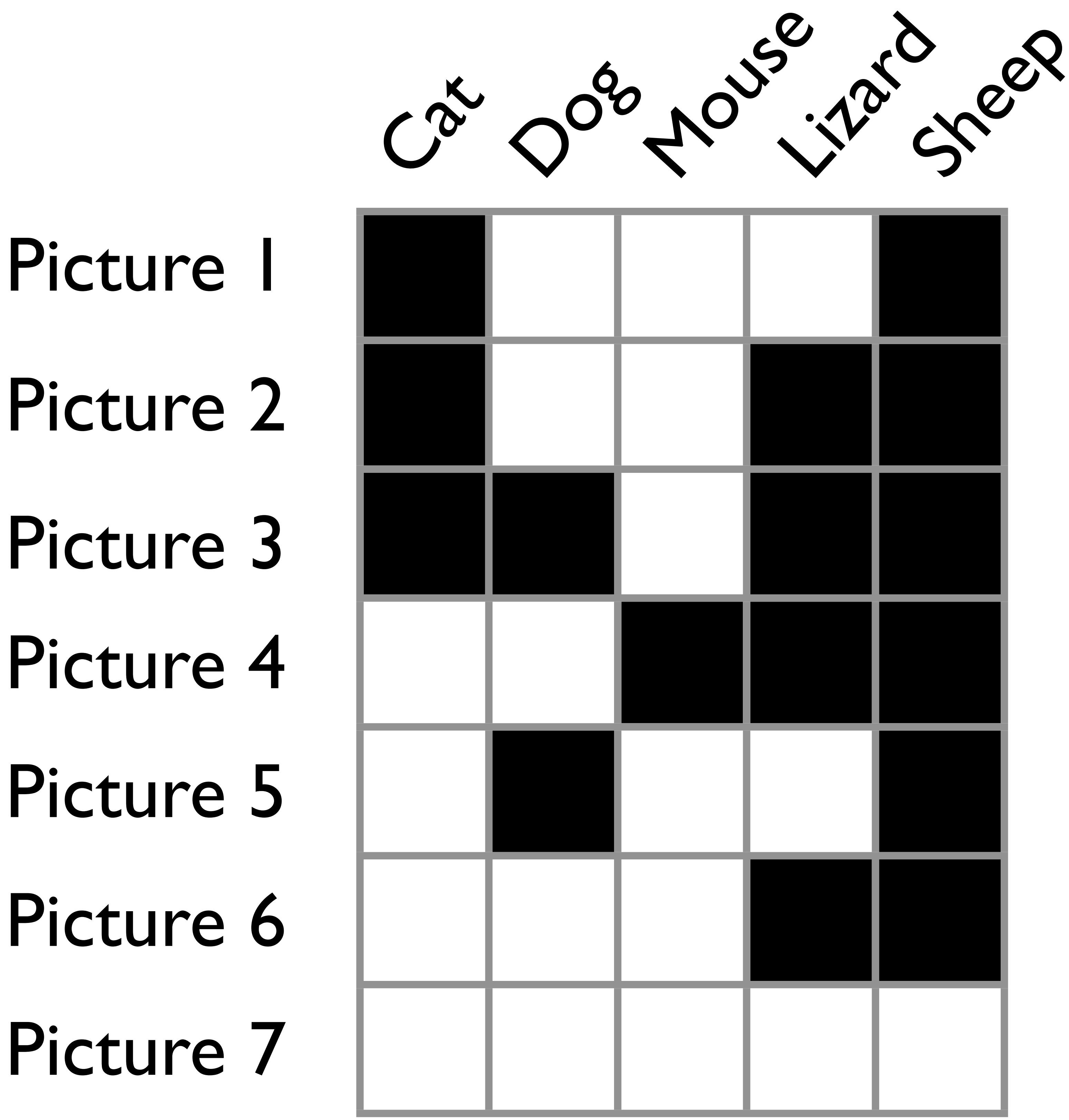


“clusters”

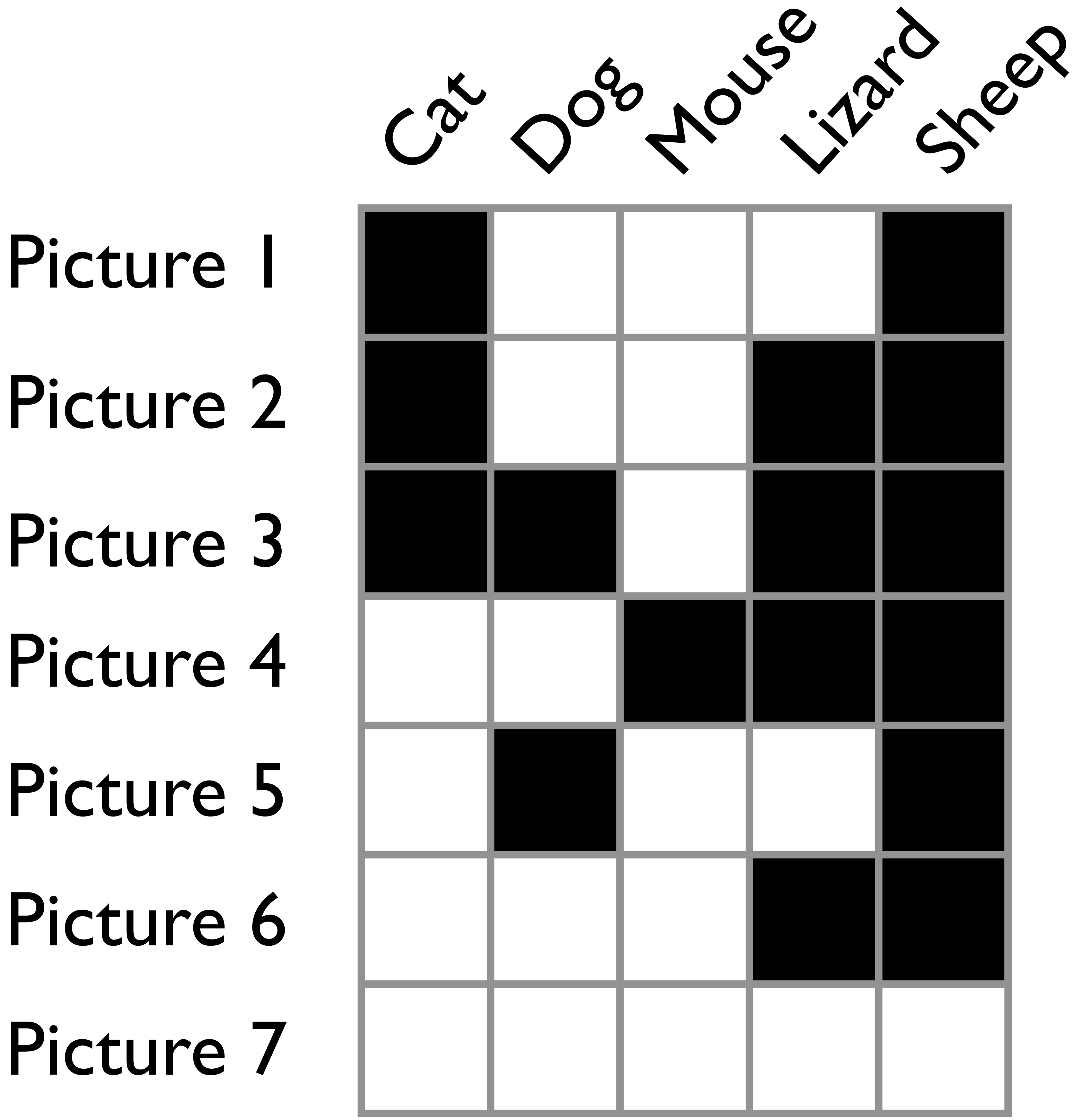
Clusters


Cat
Dog
Mouse
Lizard
Sheep
Picture 1
Picture 2
Picture 3
Picture 4
Picture 5
Picture 6
Picture 7

Features


Cat
Dog
Mouse
Lizard
Sheep
Picture 1
Picture 2
Picture 3
Picture 4
Picture 5
Picture 6
Picture 7

Features



Many other possible latent structures in data

How do we learn latent structure?

How do we learn latent structure?

K-means

How do we learn latent structure?

K-means

- Fast

How do we learn latent structure?

K-means

- Fast
- Can parallelize

How do we learn latent structure?

K-means

- Fast
- Can parallelize
- Straightforward

How do we learn latent structure?

K-means

- Fast
- Can parallelize
- Straightforward
- Only works for K clusters

How do we learn latent structure?

K-means

- Fast
- Can parallelize
- Straightforward
- Only works for K clusters

Nonparametric Bayes

How do we learn latent structure?

K-means

- Fast
- Can parallelize
- Straightforward
- Only works for K clusters

Nonparametric Bayes

- Modular (general latent structure)

How do we learn latent structure?

K-means

- Fast
- Can parallelize
- Straightforward
- Only works for K clusters

Nonparametric Bayes

- Modular (general latent structure)
- Flexible (K can grow as data grows)

How do we learn latent structure?

K-means

- Fast
- Can parallelize
- Straightforward
- Only works for K clusters

Nonparametric Bayes

- Modular (general latent structure)
- Flexible (K can grow as data grows)
- Coherent treatment of uncertainty

How do we learn latent structure?

K-means

- Fast
- Can parallelize
- Straightforward
- Only works for K clusters

Nonparametric Bayes

- Modular (general latent structure)
- Flexible (K can grow as data grows)
- Coherent treatment of uncertainty

But...

- E.g., Silicon Valley: can have petabytes of data
- Practitioners turn to what runs

MAD-Bayes Perspectives

- Bayesian nonparametrics assists the optimization-based inference community

MAD-Bayes Perspectives

- Bayesian nonparametrics assists the optimization-based inference community
  - New, modular, flexible, nonparametric objectives & regularizers

MAD-Bayes Perspectives

- Bayesian nonparametrics assists the optimization-based inference community
  - New, modular, flexible, nonparametric objectives & regularizers
  - Alternative perspective: fast initialization

MAD-Bayes Perspectives

- Bayesian nonparametrics assists the optimization-based inference community
  - New, modular, flexible, nonparametric objectives & regularizers
  - Alternative perspective: fast initialization

Inspiration

- Consider a finite Gaussian mixture model

MAD-Bayes Perspectives

- Bayesian nonparametrics assists the optimization-based inference community
  - New, modular, flexible, nonparametric objectives & regularizers
  - Alternative perspective: fast initialization

Inspiration

- Consider a finite Gaussian mixture model
- The steps of the EM algorithm limit to the steps of the K-means algorithm as the Gaussian variance is taken to 0

MAD-Bayes

The MAD-Bayes idea

- Start with nonparametric Bayes model
- Take a similar limit to get a K-means-like objective

MAD-Bayes

The MAD-Bayes idea

- Start with nonparametric Bayes model
- Take a similar limit to get a K-means-like objective

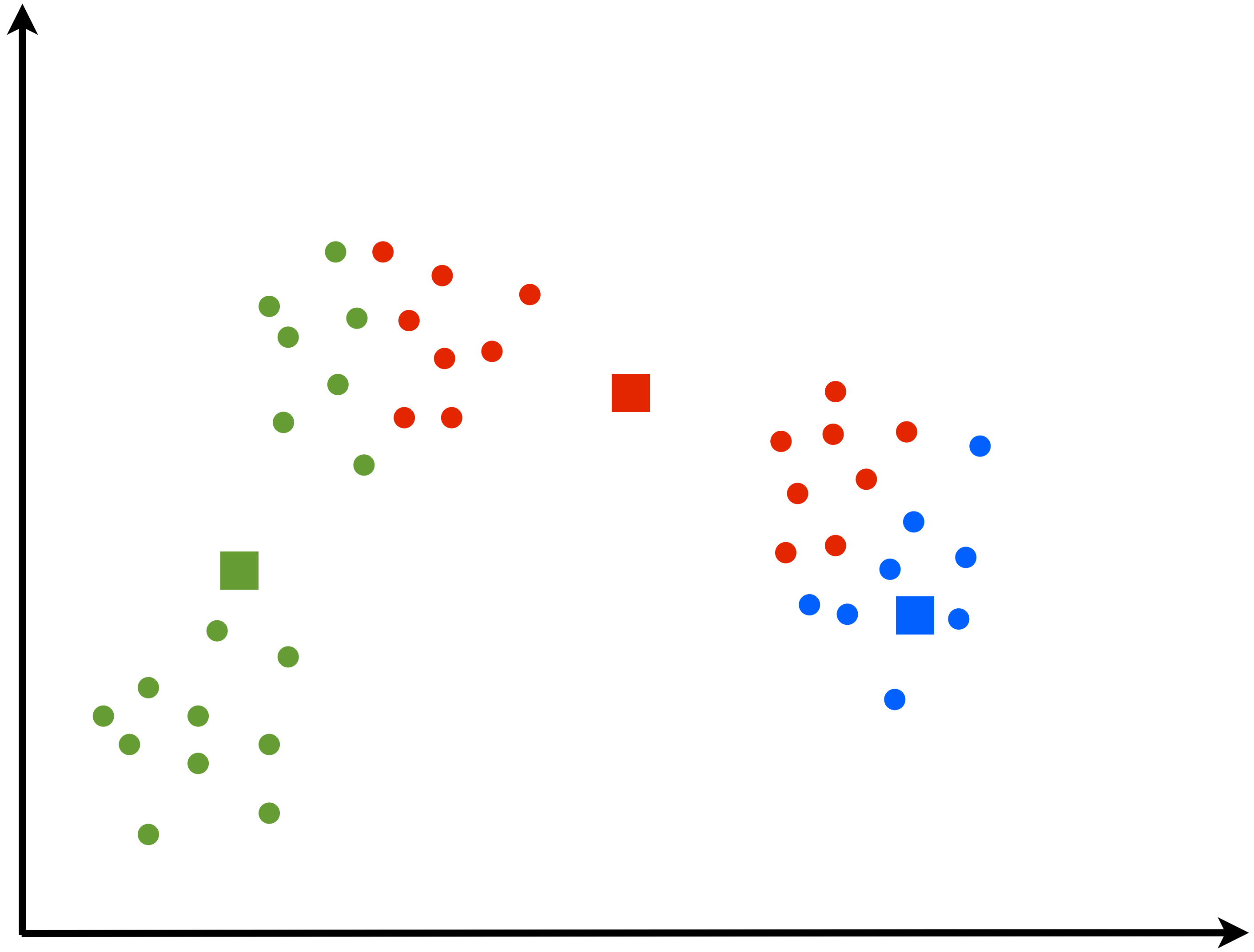
MAD-Bayes

The MAD-Bayes idea

- Start with nonparametric Bayes model
- Take a similar limit to get a K-means-like objective

K-means

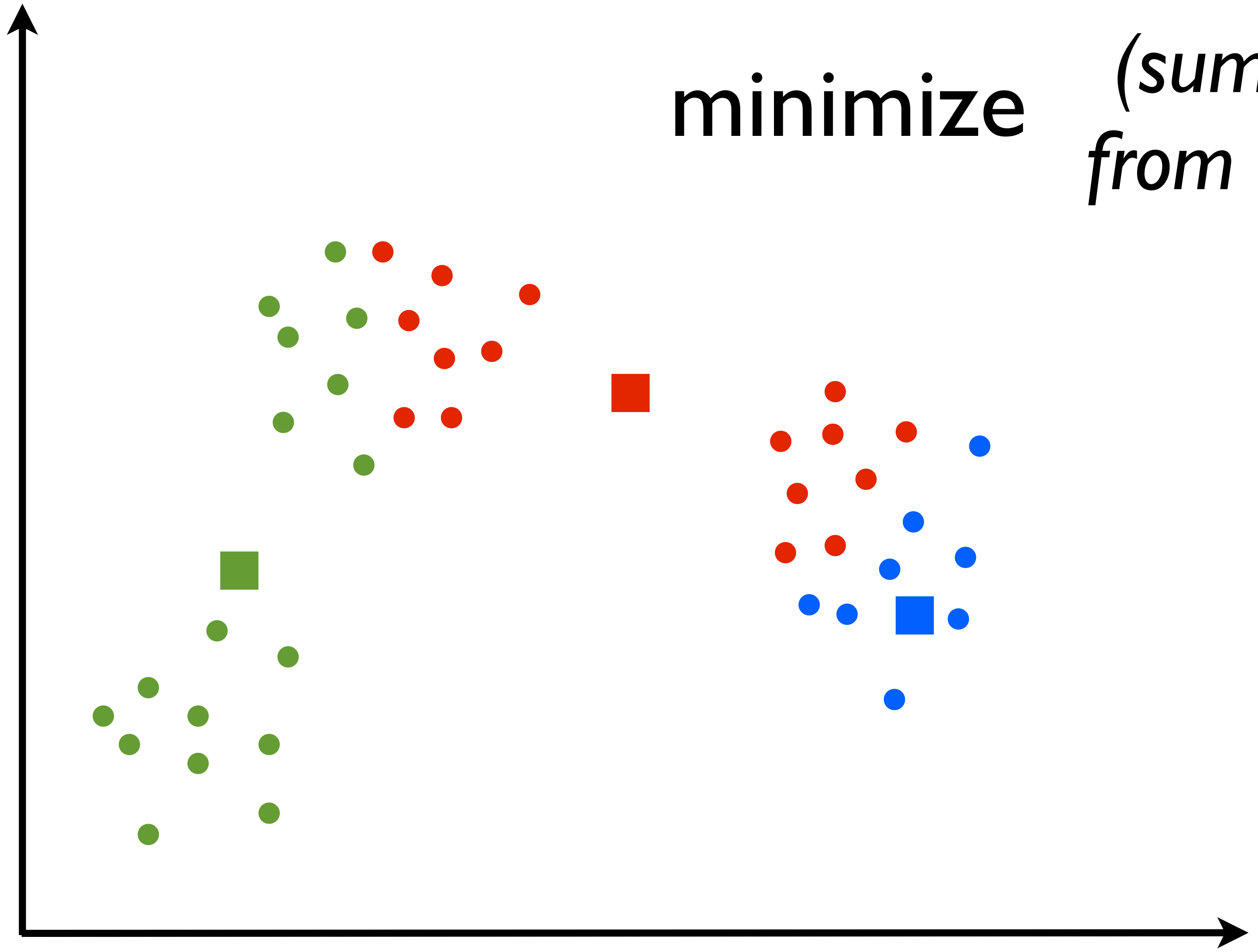
K-means clustering problem



K-means

K-means clustering problem

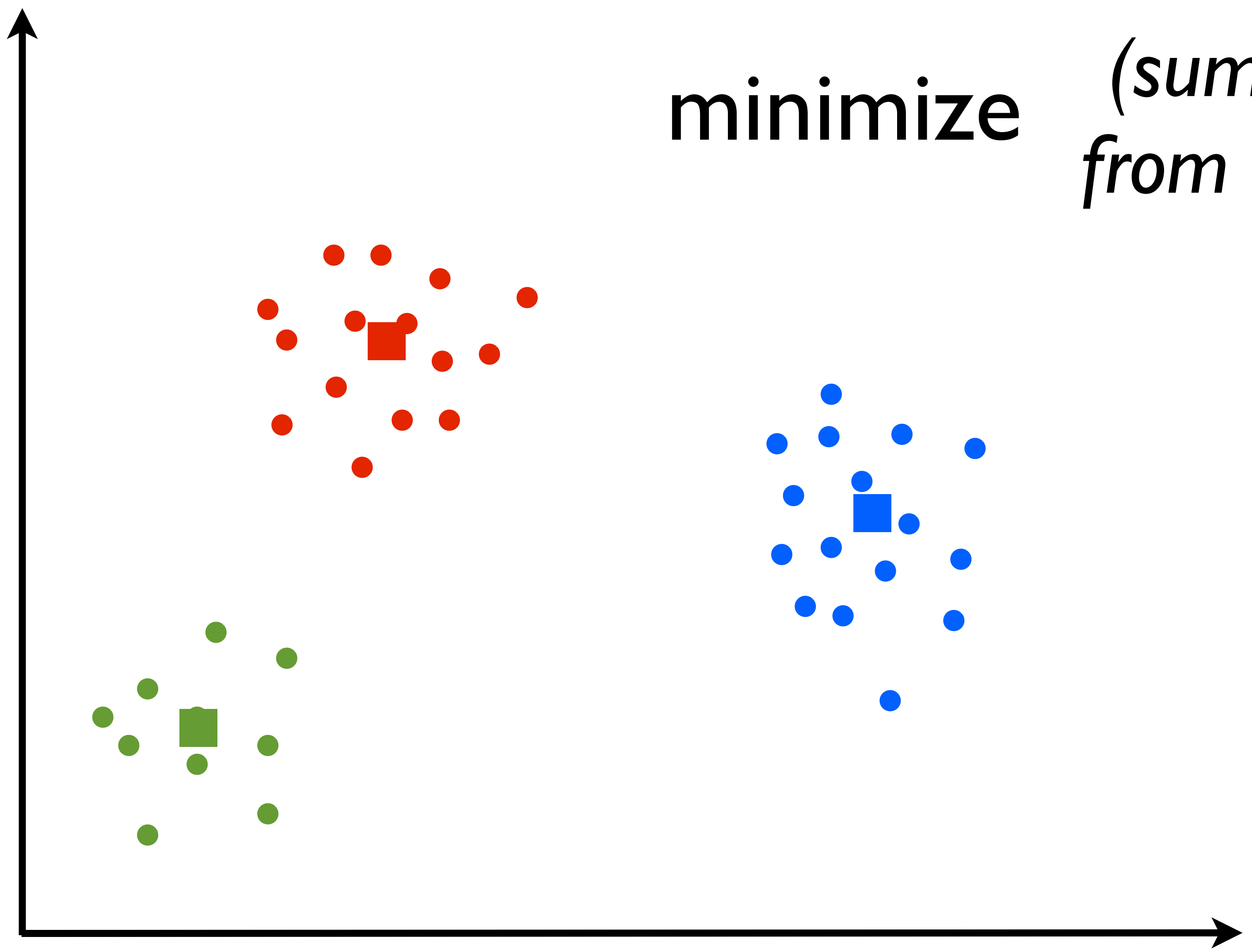
minimize *(sum of square distances from data points to cluster centers)*



K-means

K-means clustering problem

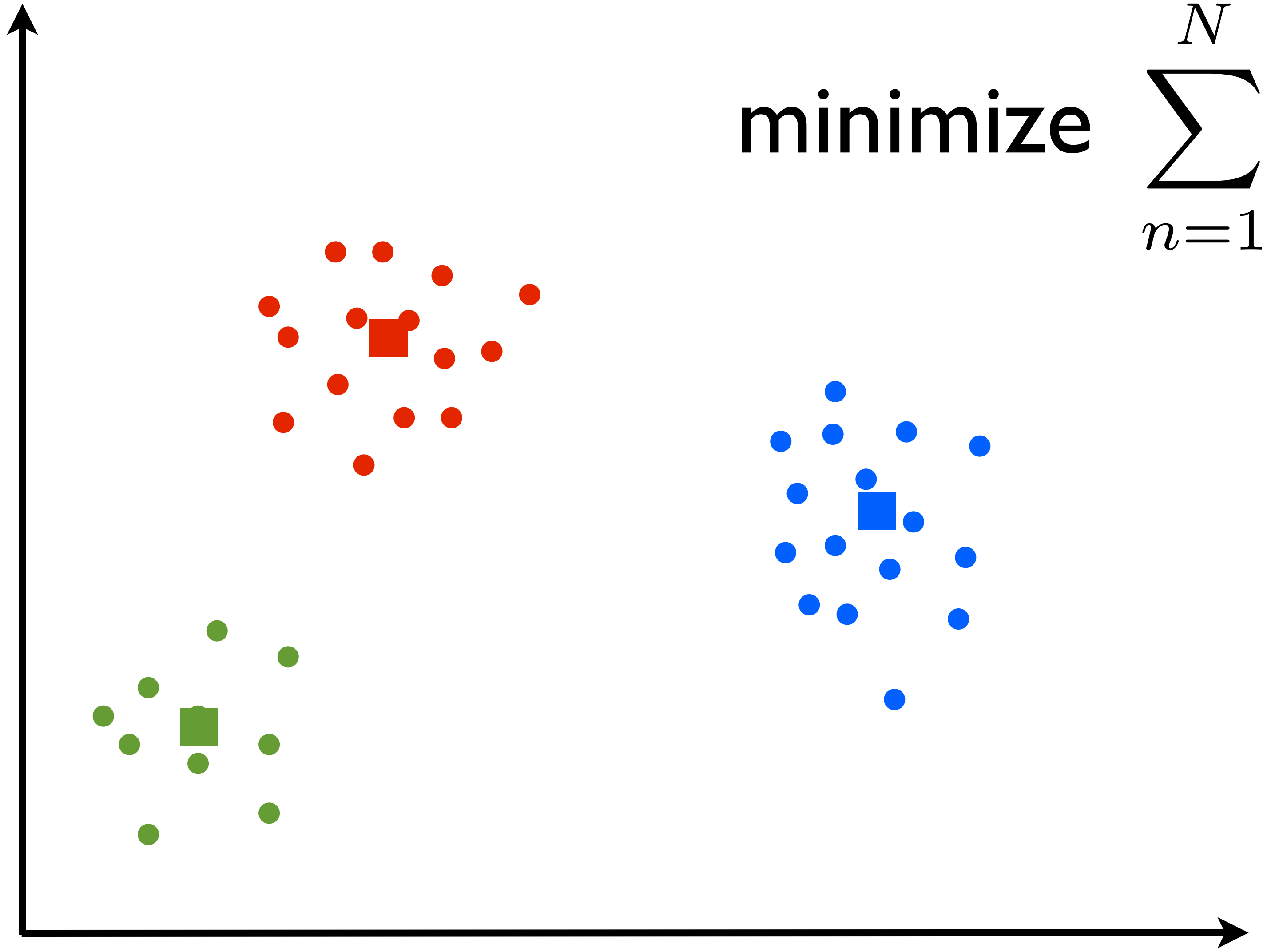
minimize *(sum of square distances from data points to cluster centers)*



K-means

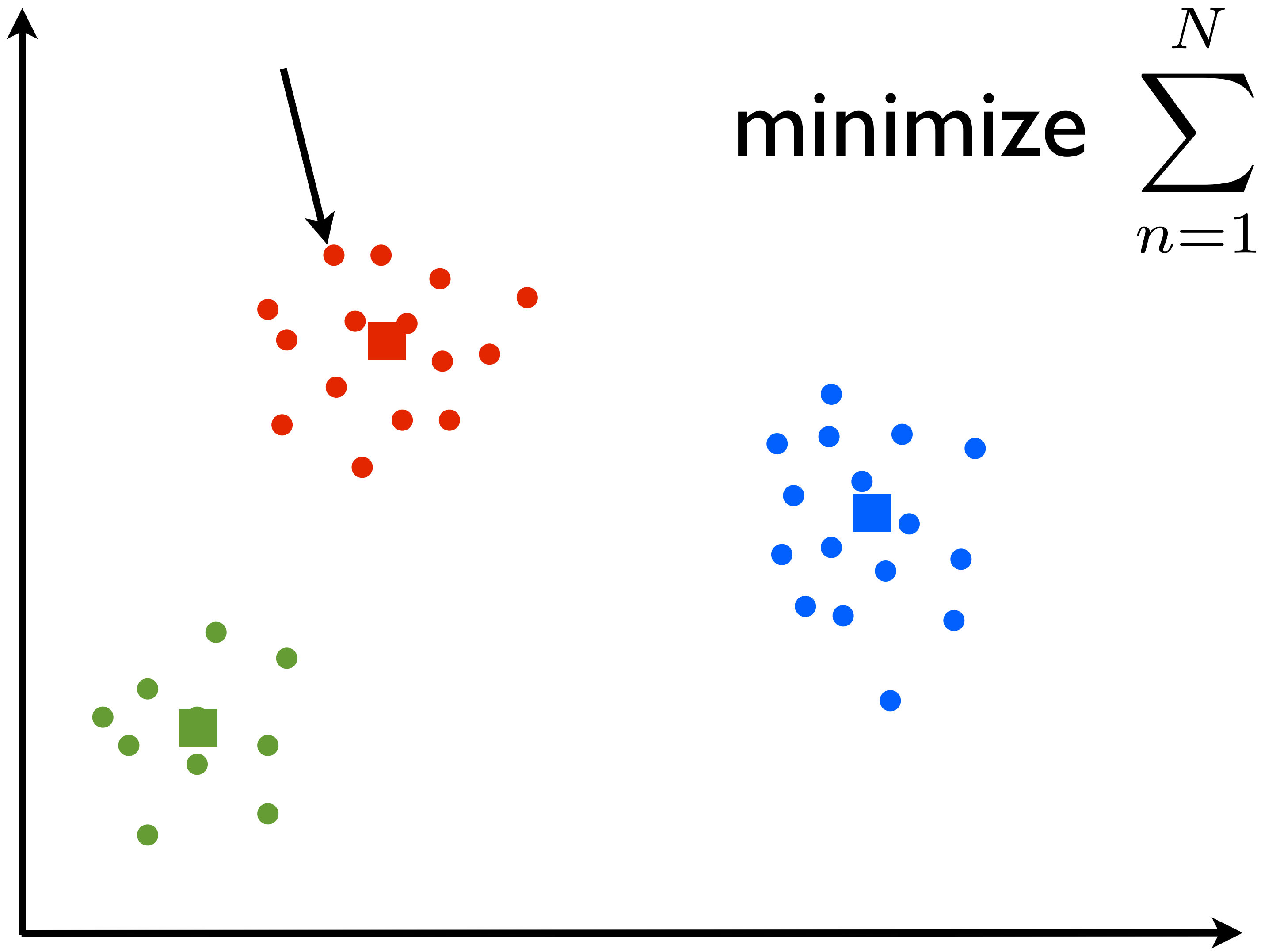
K-means clustering problem

minimize $\sum_{n=1}^{N} \|x_n - \text{center}_n\|^2$



K-means

K-means clustering problem

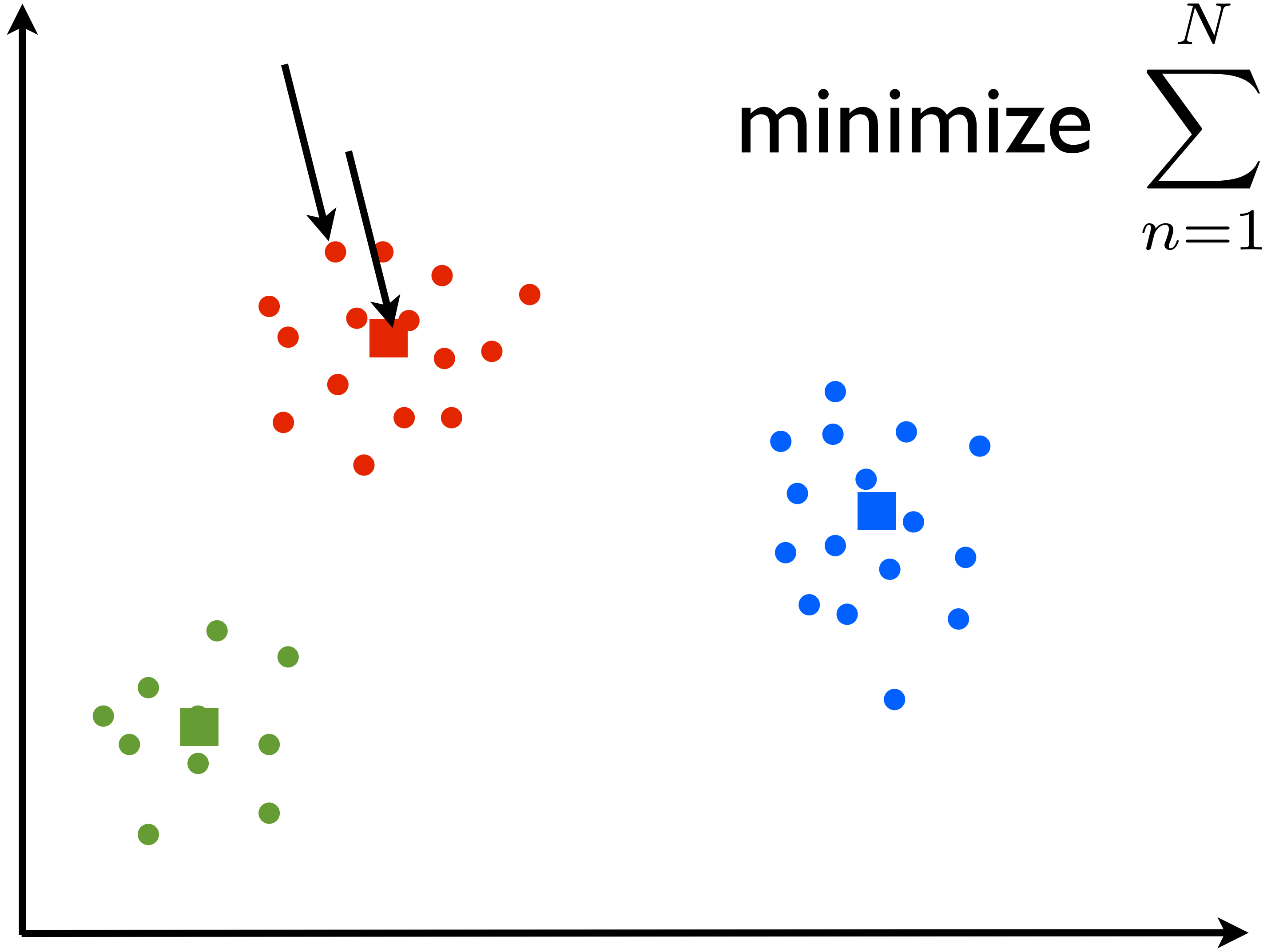


minimize $\sum_{n=1}^{N} \|x_n - \mathrm{center}_n\|^2$

K-means

K-means clustering problem

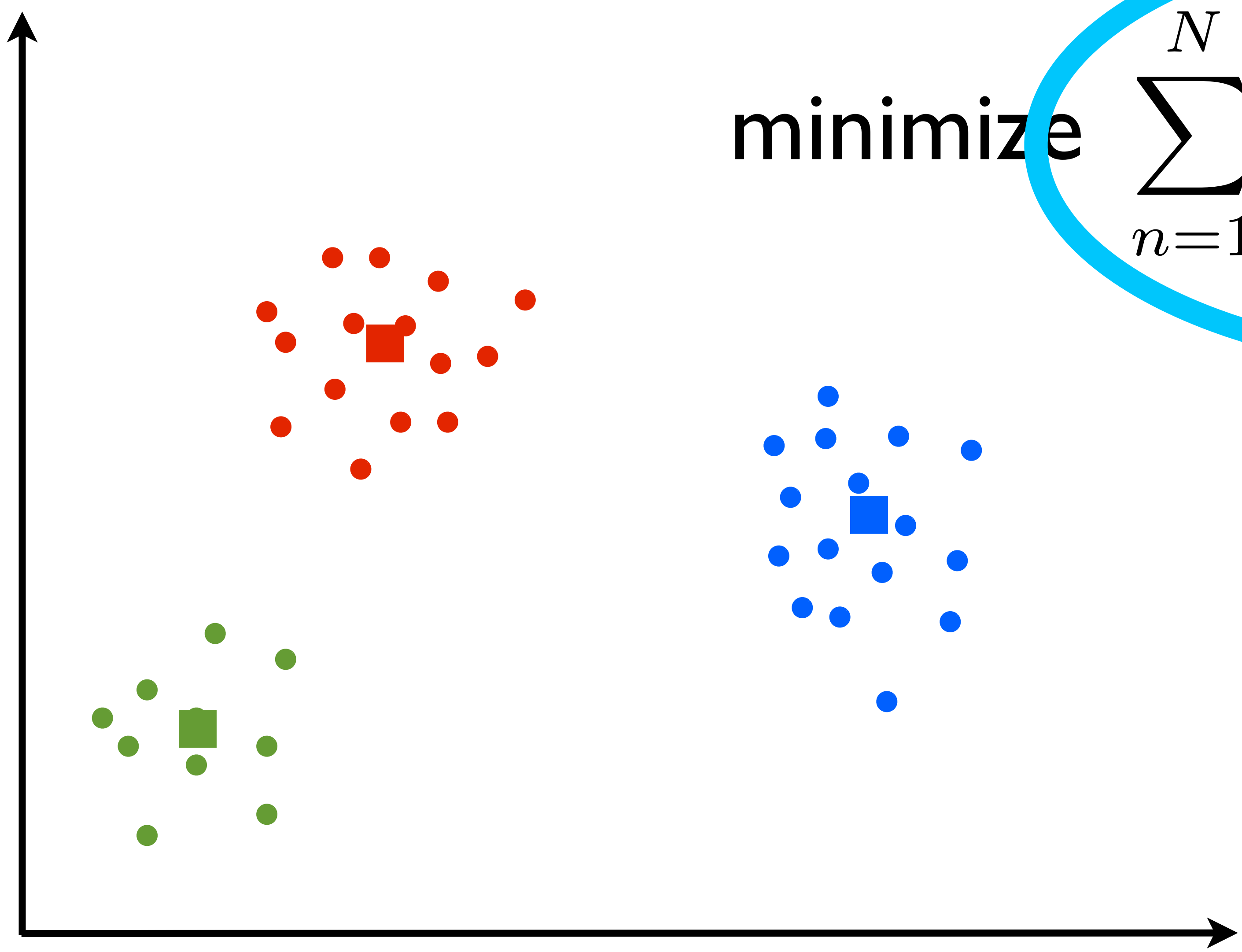
$$\text{minimize} \sum_{n=1}^{N} \|x_n - \text{center}_n\|^2$$



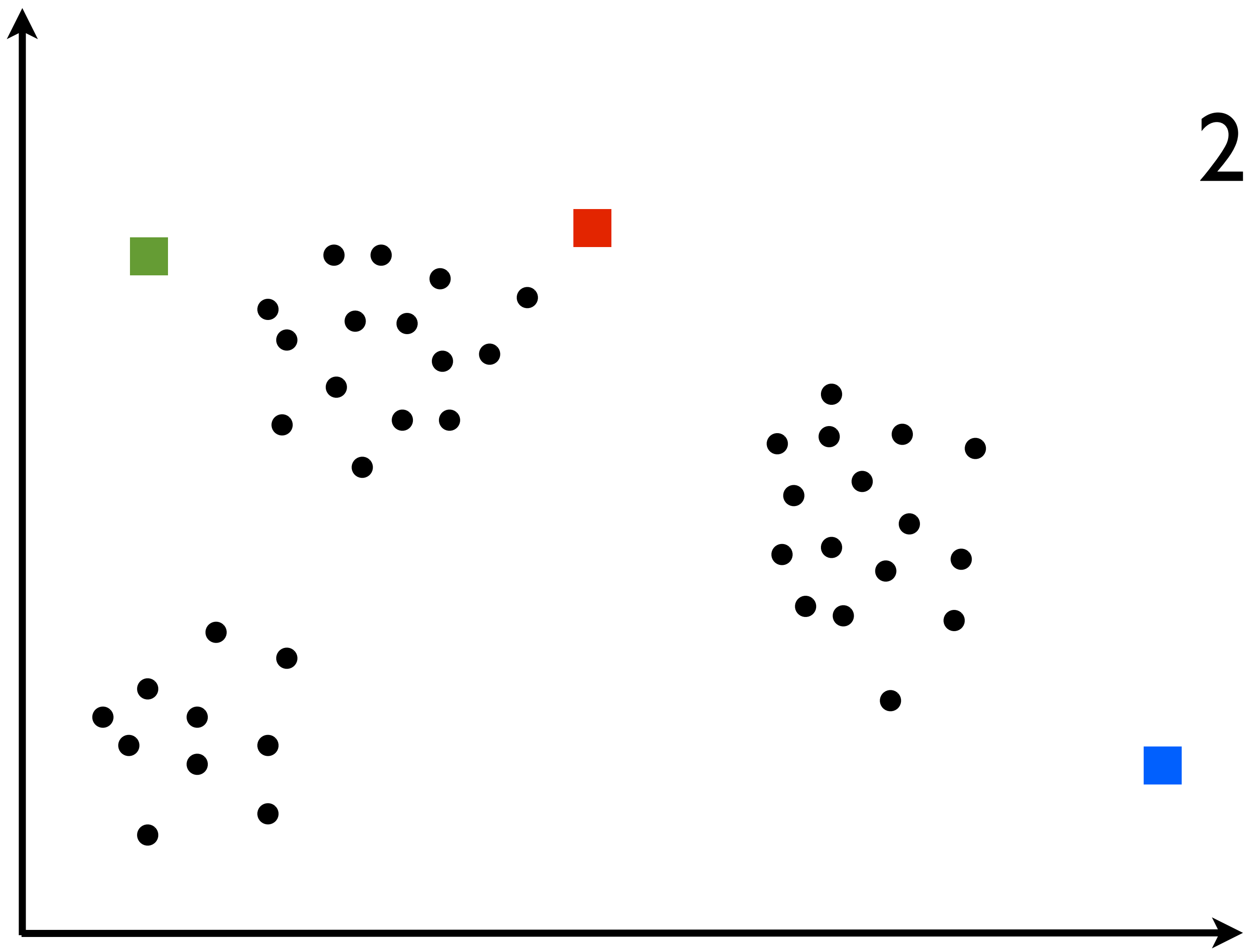
K-means

K-means objective

minimize $$\sum_{n=1}^{N} \|x_n - \text{center}_n\|^2$$



Lloyd's algorithm



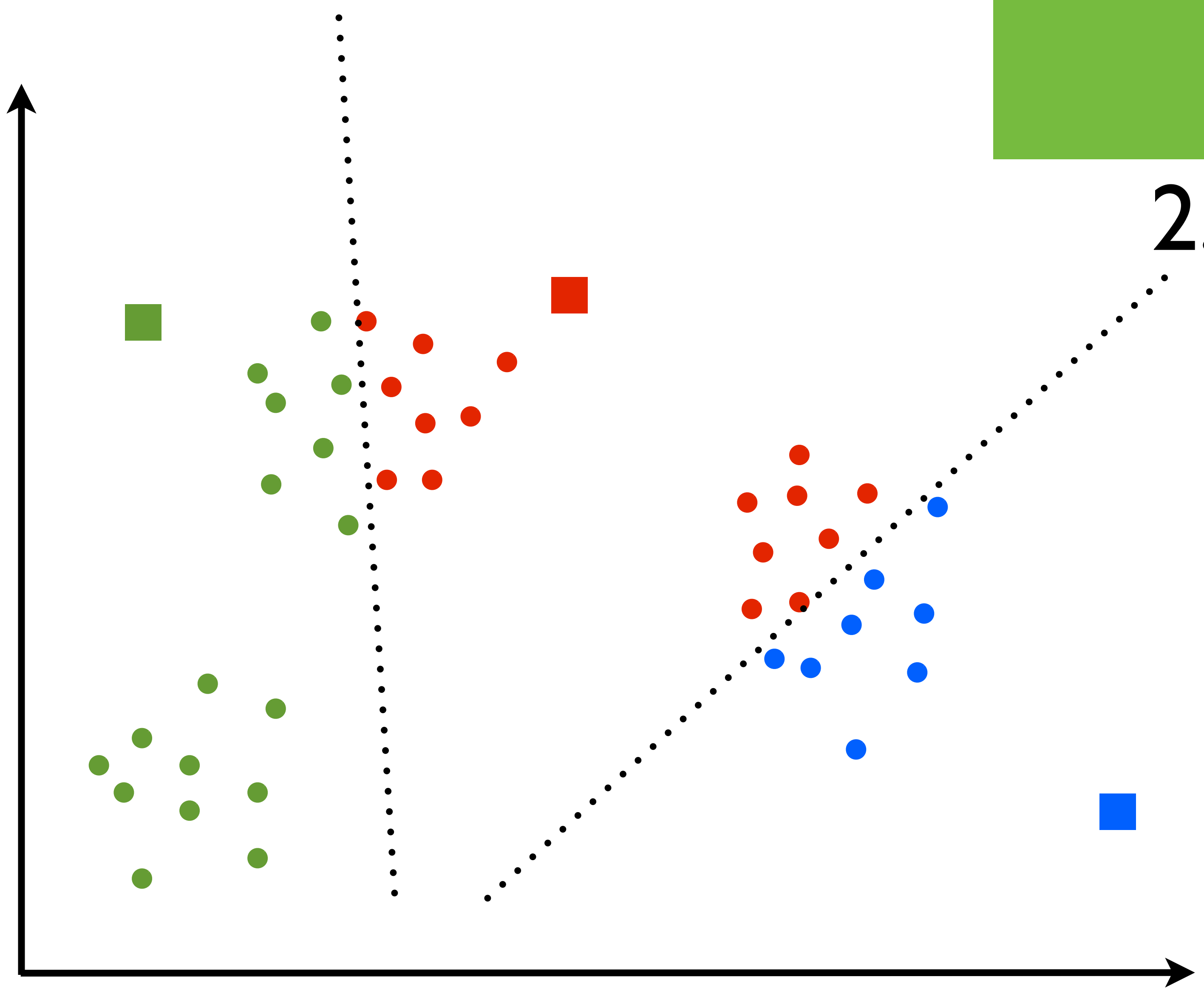
Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to a cluster
2. Update cluster means

Lloyd’s algorithm

Iterate until no changes:

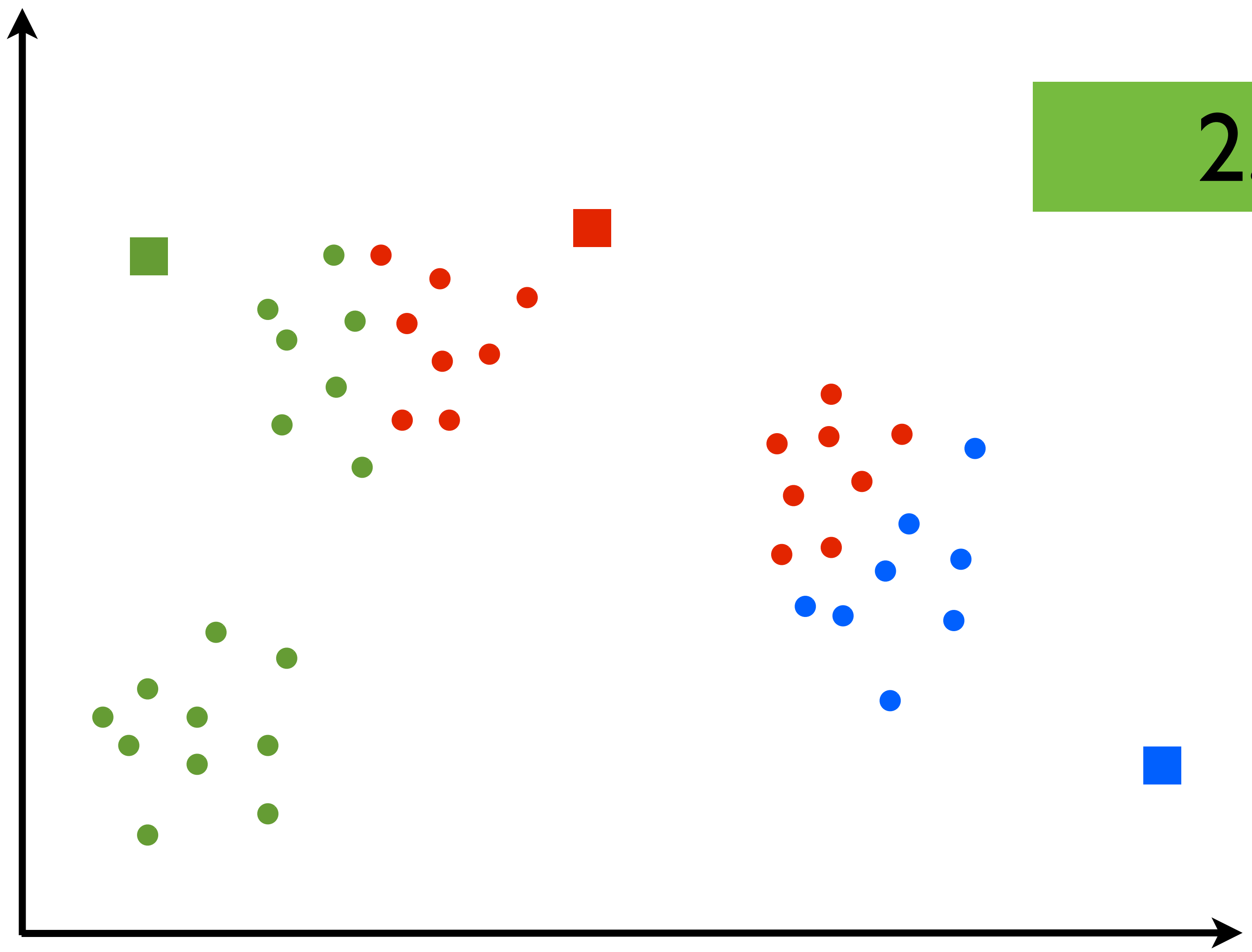
1. For n = 1, ..., N
   - Assign point n to a cluster
2. Update cluster means



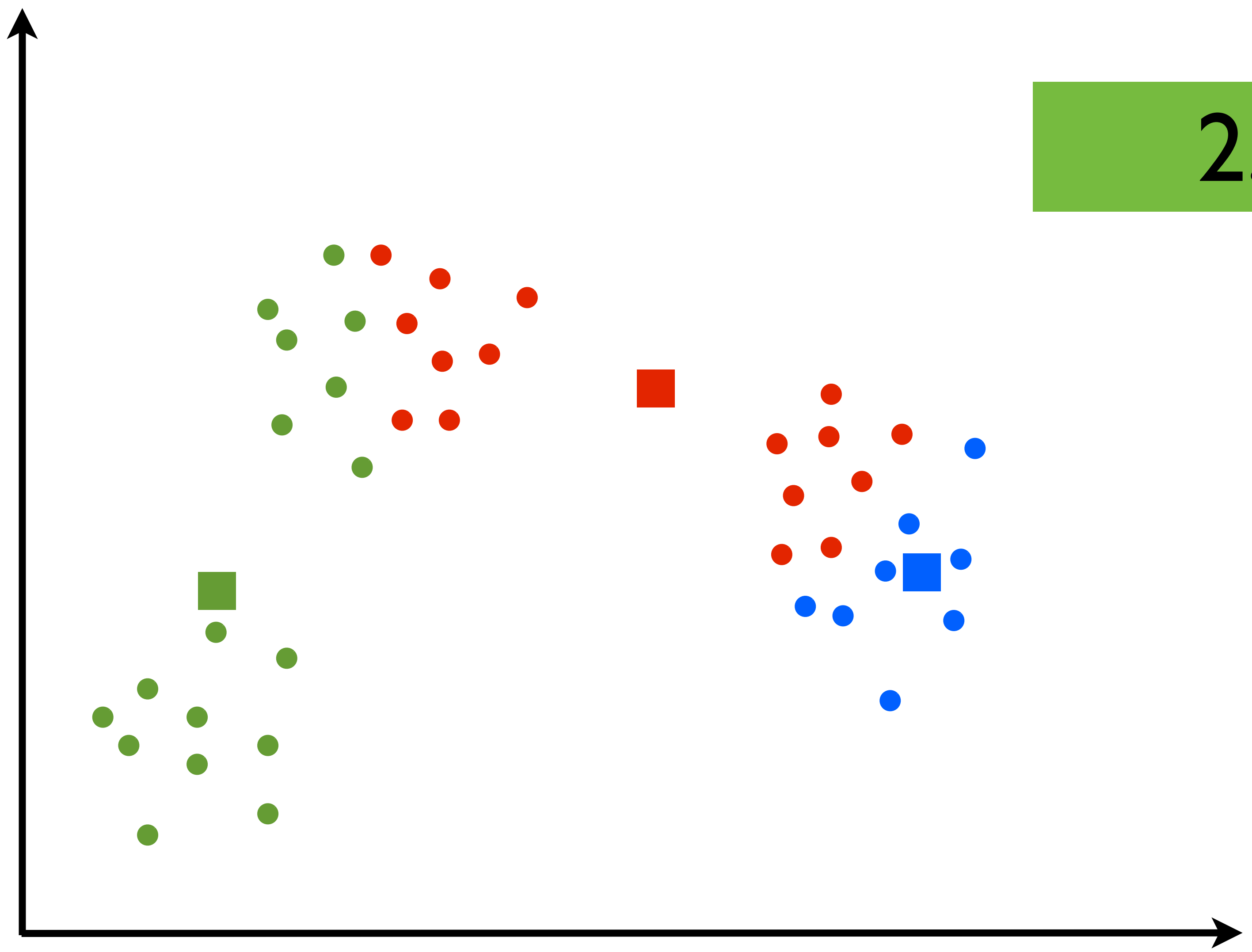
Lloyd's algorithm

Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to a cluster
2. Update cluster means



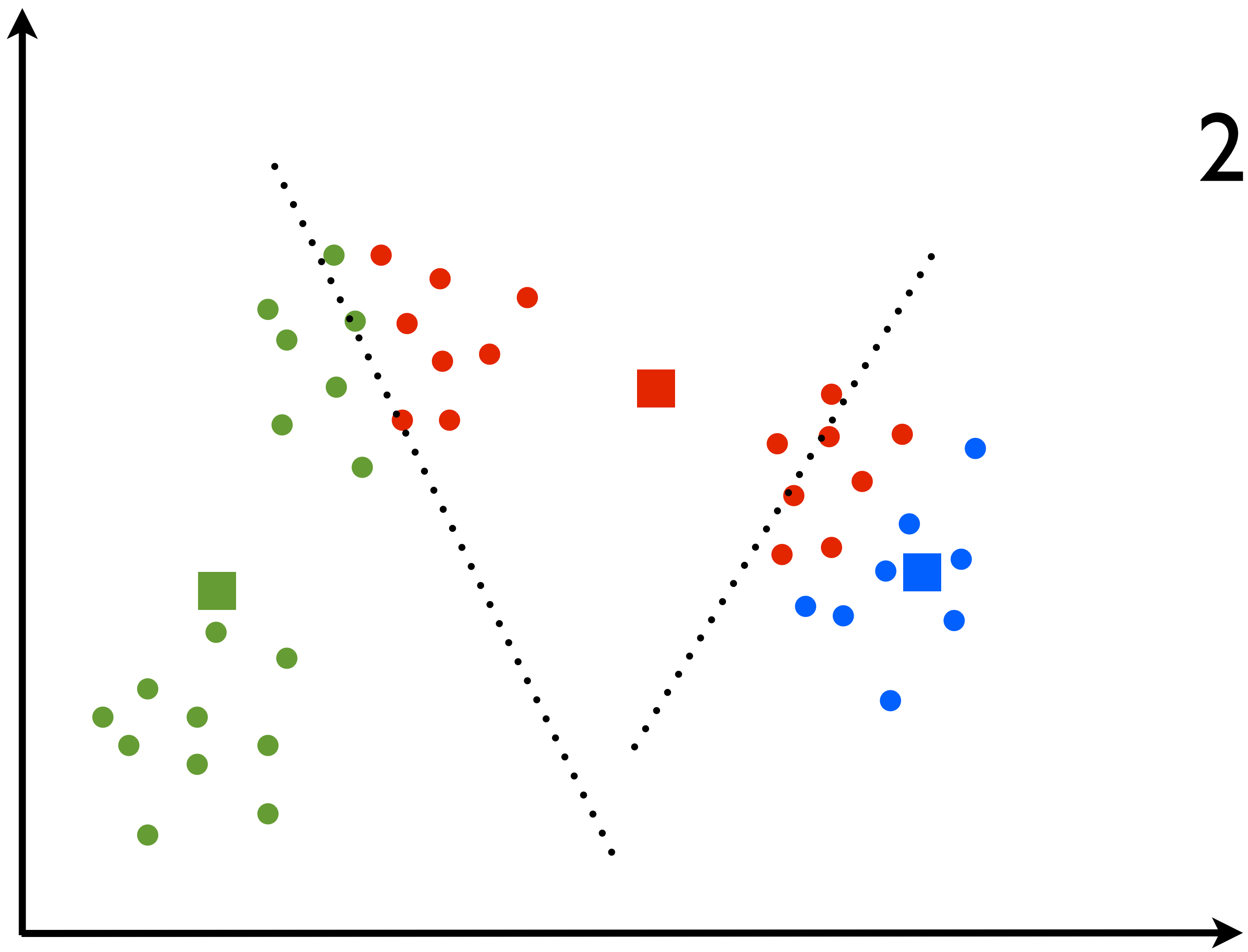
Lloyd's algorithm



Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to a cluster
2. Update cluster means

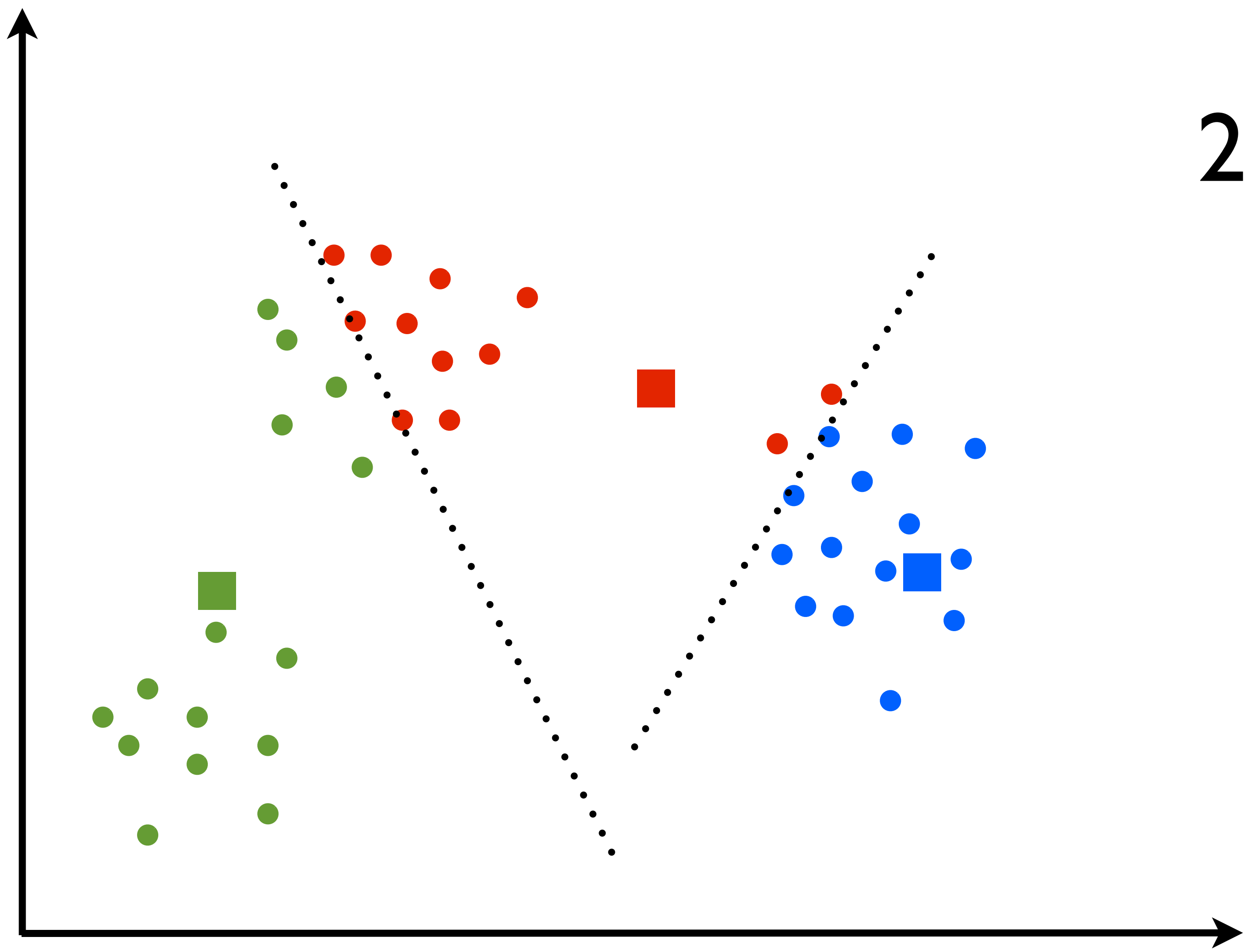
Lloyd's algorithm



Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to a cluster
2. Update cluster means

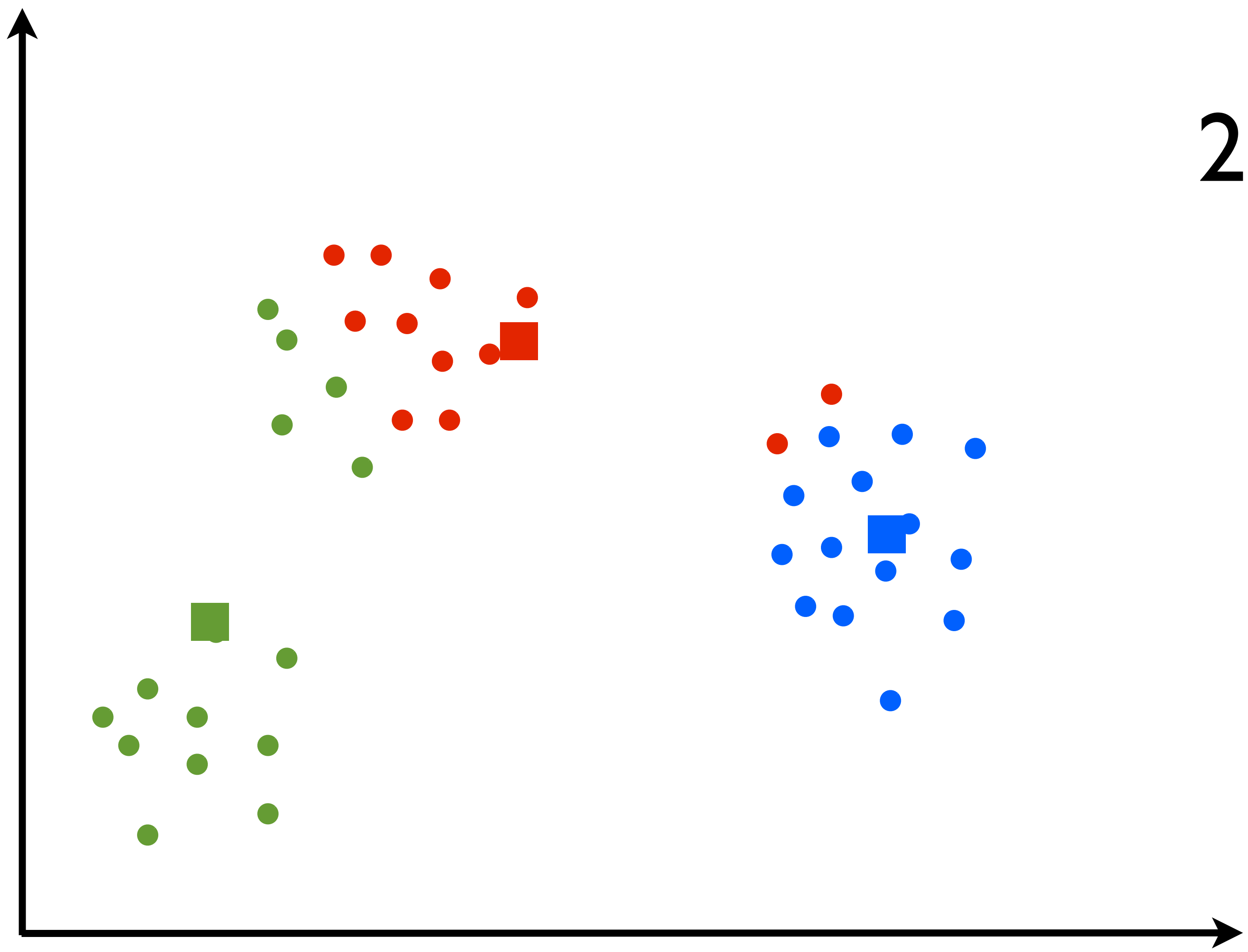
Lloyd’s algorithm



Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to a cluster
2. Update cluster means

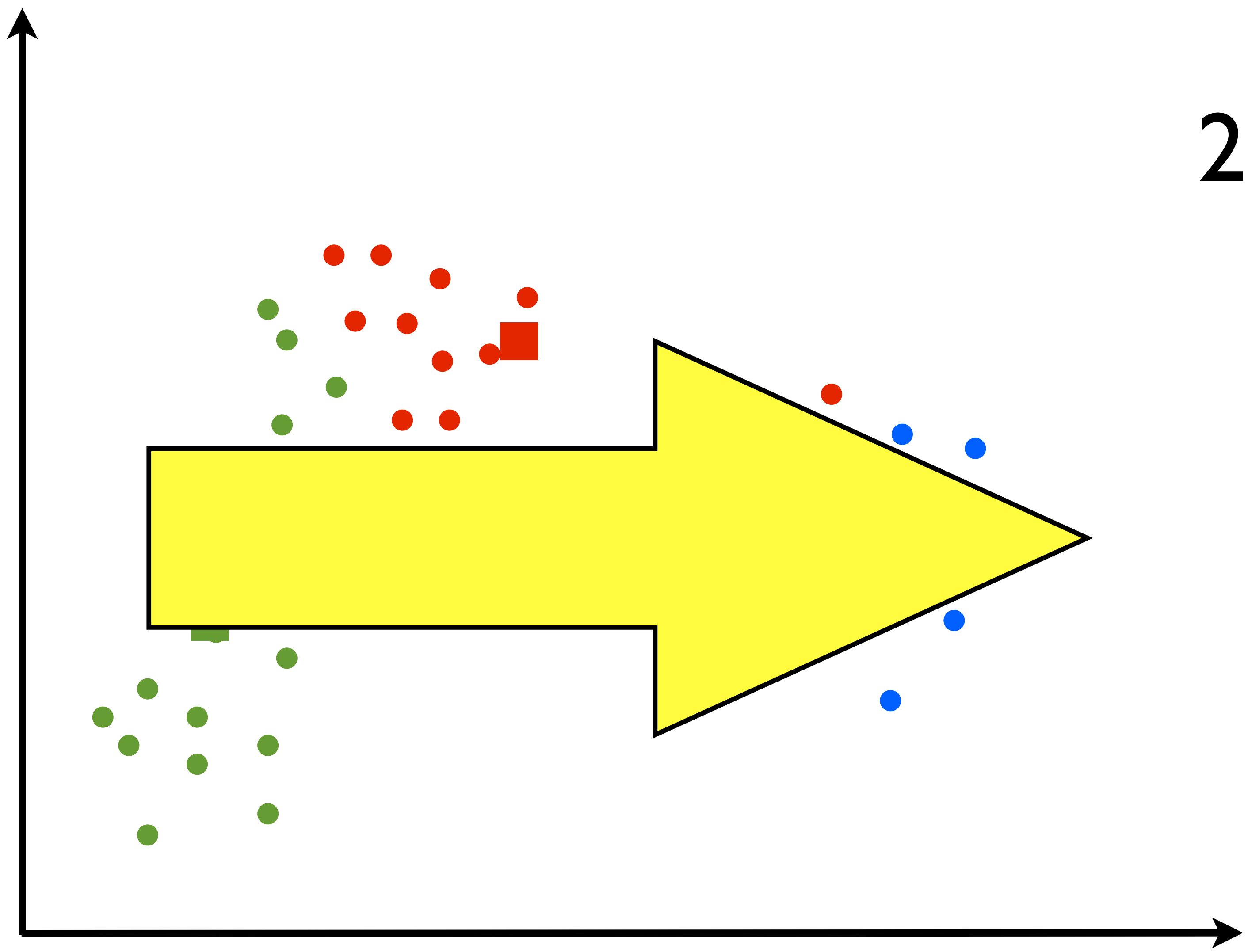
Lloyd's algorithm



Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to a cluster
2. Update cluster means

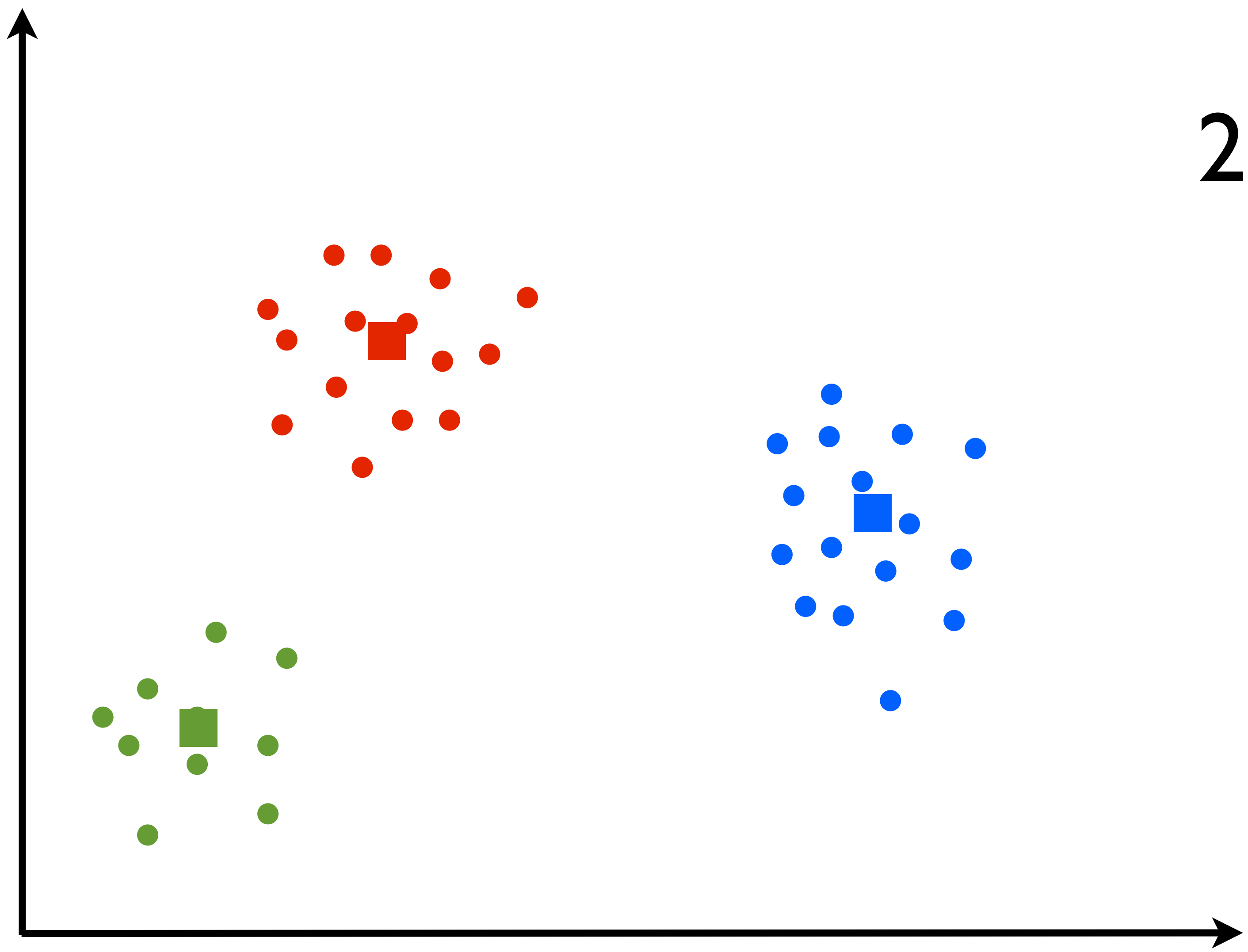
Lloyd's algorithm



Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to a cluster
2. Update cluster means

Lloyd's algorithm



Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to a cluster
2. Update cluster means

MAD-Bayes

The MAD-Bayes idea

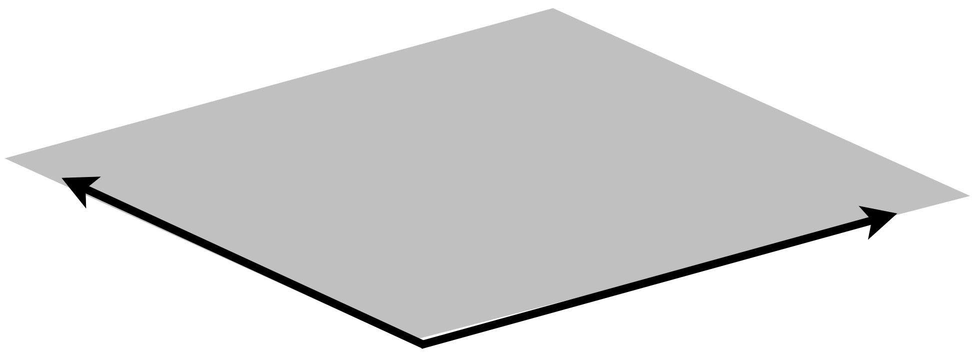
- Start with nonparametric Bayes model
- Take a similar limit to get a K-means-like objective

MAD-Bayes

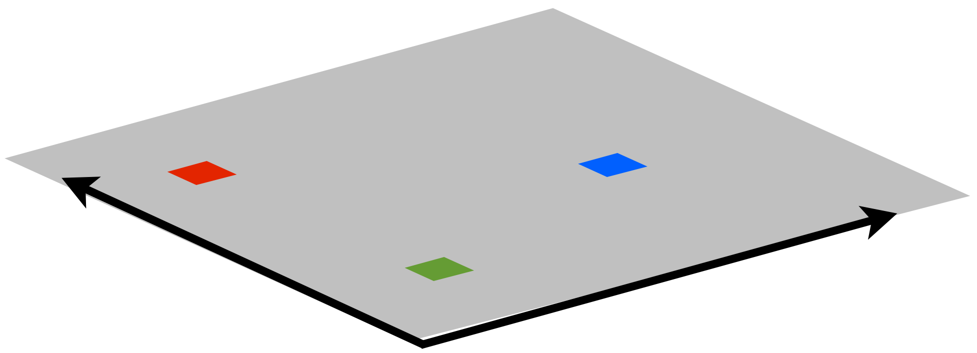
The MAD-Bayes idea

- Start with nonparametric Bayes model
- Take a similar limit to get a K-means-like objective

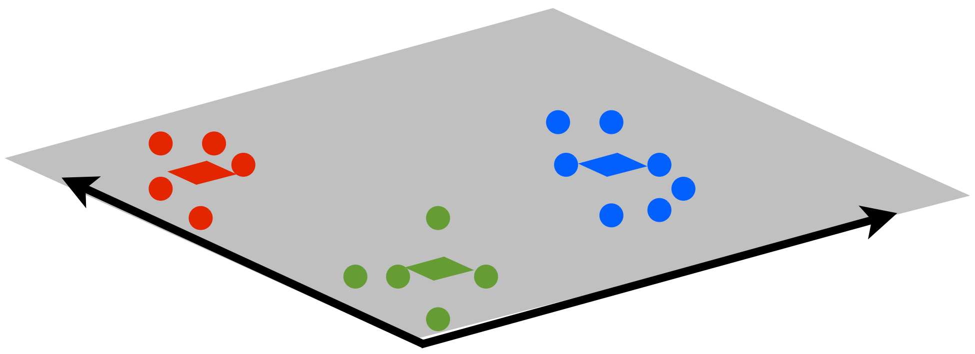
Bayesian model



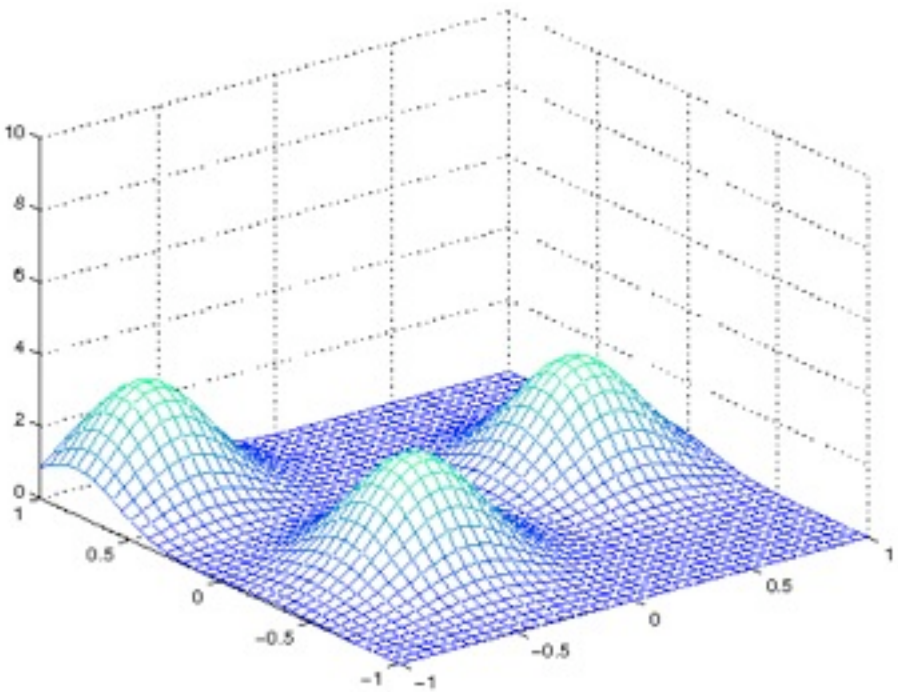
Bayesian model



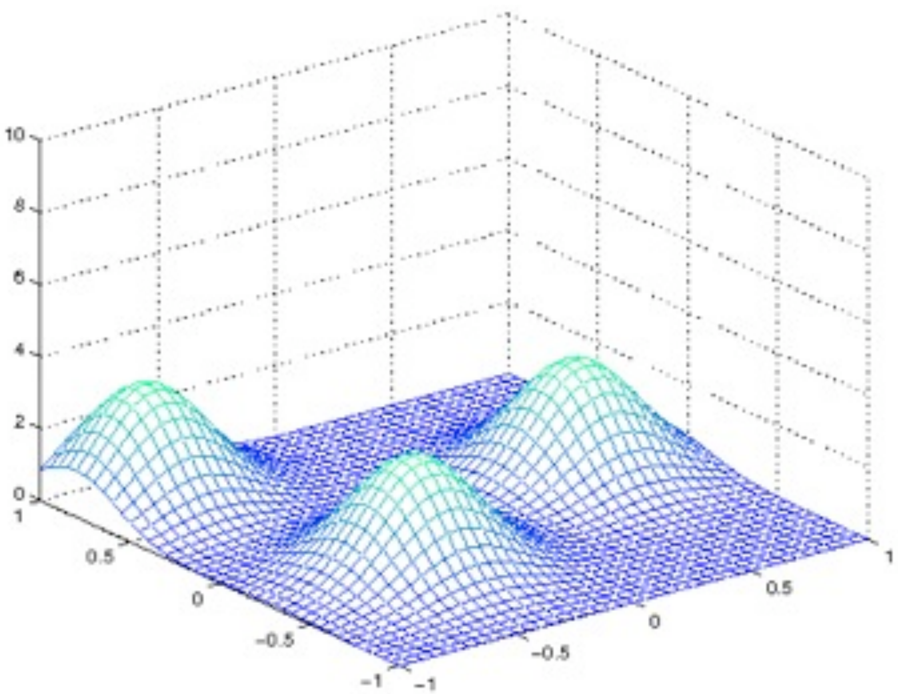
Bayesian model



Bayesian model



Bayesian model



Nonparametric

- number of parameters can grow with the number of data points

MAD-Bayes

The MAD-Bayes idea

- Start with nonparametric Bayes model
- Take a similar limit to get a K-means-like objective

MAD-Bayes

The MAD-Bayes idea

- Start with nonparametric Bayes model
- Take a similar limit to get a K-means-like objective

MAD-Bayes

MAD-Bayes

- *Maximum a Posteriori* (MAP) is an optimization problem

$$\mathrm{argmax}_{\mathrm{parameters}} \mathbb{P}(\mathrm{parameters} | \mathrm{data})$$

MAD-Bayes

- *Maximum a Posteriori* (MAP) is an optimization problem

$$\text{argmax}_{\text{parameters}} \mathbb{P}(\text{parameters}|\text{data})$$

- We take a limit of the objective (posterior) and get one like K-means

MAD-Bayes

- *Maximum a Posteriori* (MAP) is an optimization problem

$$\text{argmax}_{\text{parameters}} \mathbb{P}(\text{parameters}|\text{data})$$

- We take a limit of the objective (posterior) and get one like K-means
  - “Small-variance asymptotics”

MAD-Bayes

Bayesian posterior

K-means-like objectives

MAD-Bayes

<u>Bayesian posterior</u> <u>K-means-like objectives</u>

Mixture of K Gaussians → K-means

MAD-Bayes

Bayesian posterior | K-means-like objectives

Mixture of K Gaussians → K-means

Dirichlet process mixture → Unbounded number of clusters

MAD-Bayes

Bayesian posterior		K-means-like objectives
Mixture of K Gaussians	→	K-means
Dirichlet process mixture	→	Unbounded number of clusters
Hierarchical Dirichlet process	→	Multiple data sets share cluster centers

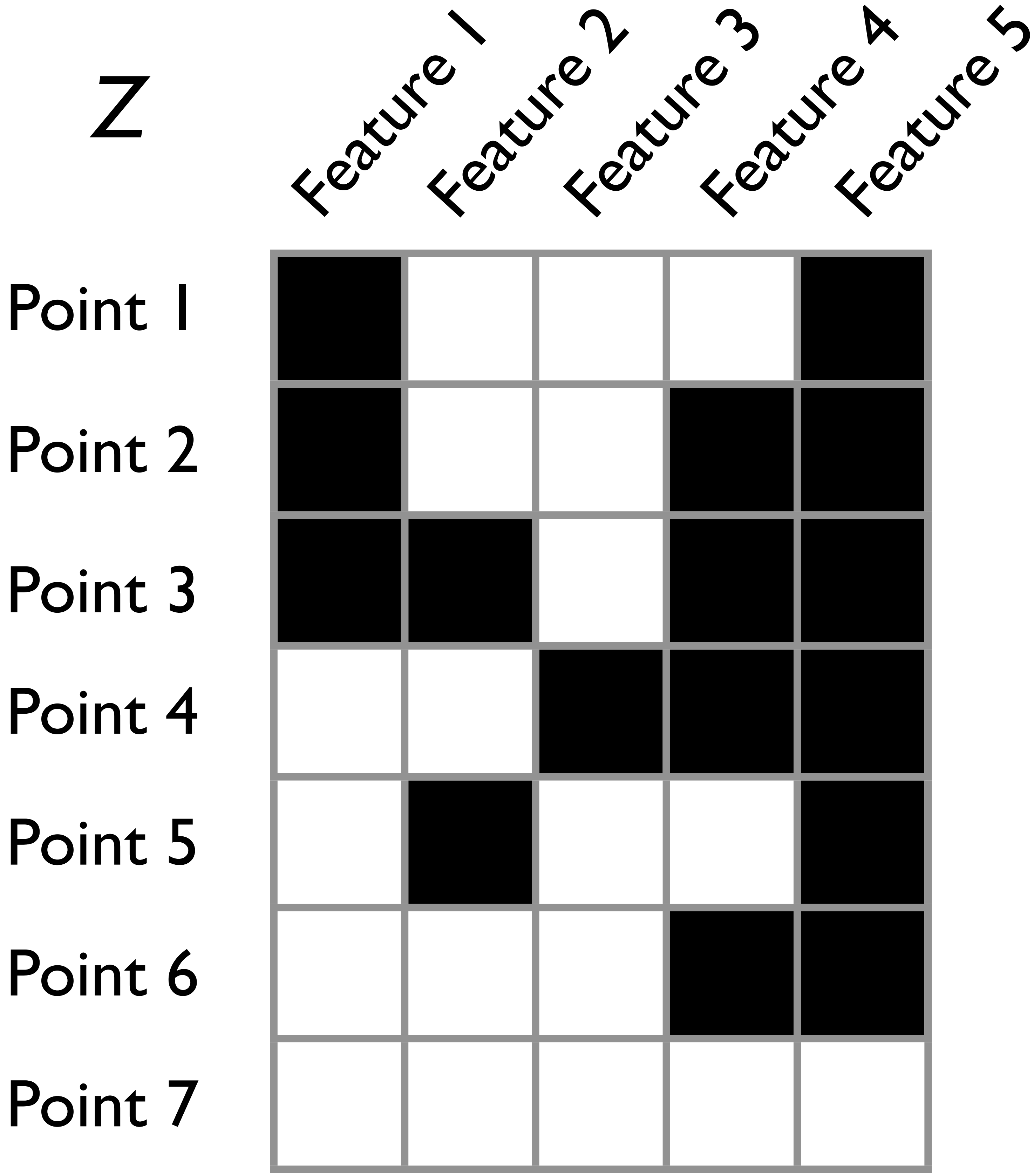
MAD-Bayes

<u>Bayesian posterior</u>		<u>K-means-like objectives</u>
Mixture of K Gaussians	→	K-means
Dirichlet process mixture	→	Unbounded number of clusters
Hierarchical Dirichlet process	→	Multiple data sets share cluster centers
⋮		⋮

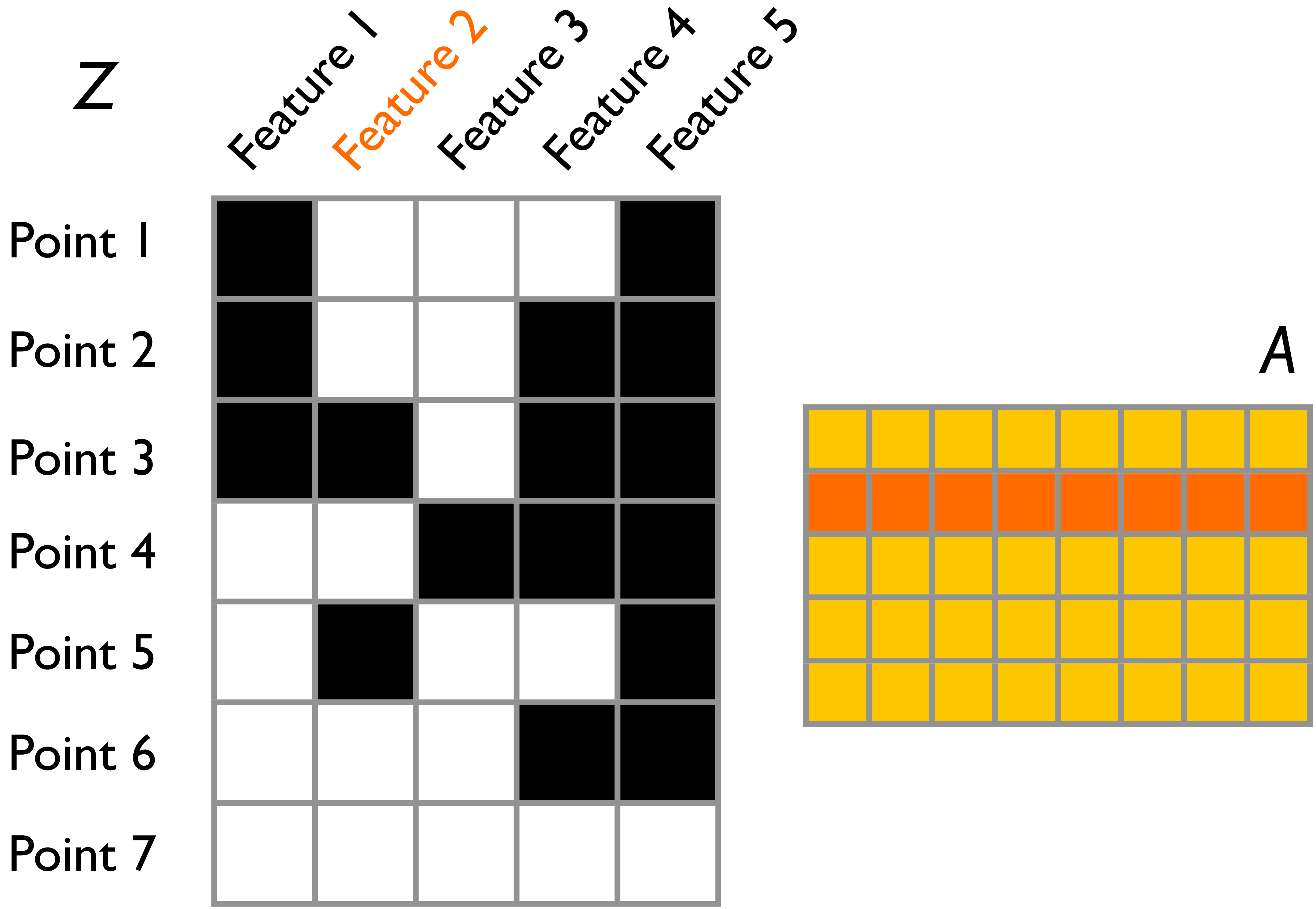
MAD-Bayes

Bayesian posterior		K-means-like objectives
Mixture of K Gaussians	→	K-means
Dirichlet process mixture	→	Unbounded number of clusters
Hierarchical Dirichlet process	→	Multiple data sets share cluster centers
⋮		⋮
Beta process	→	Features

Features

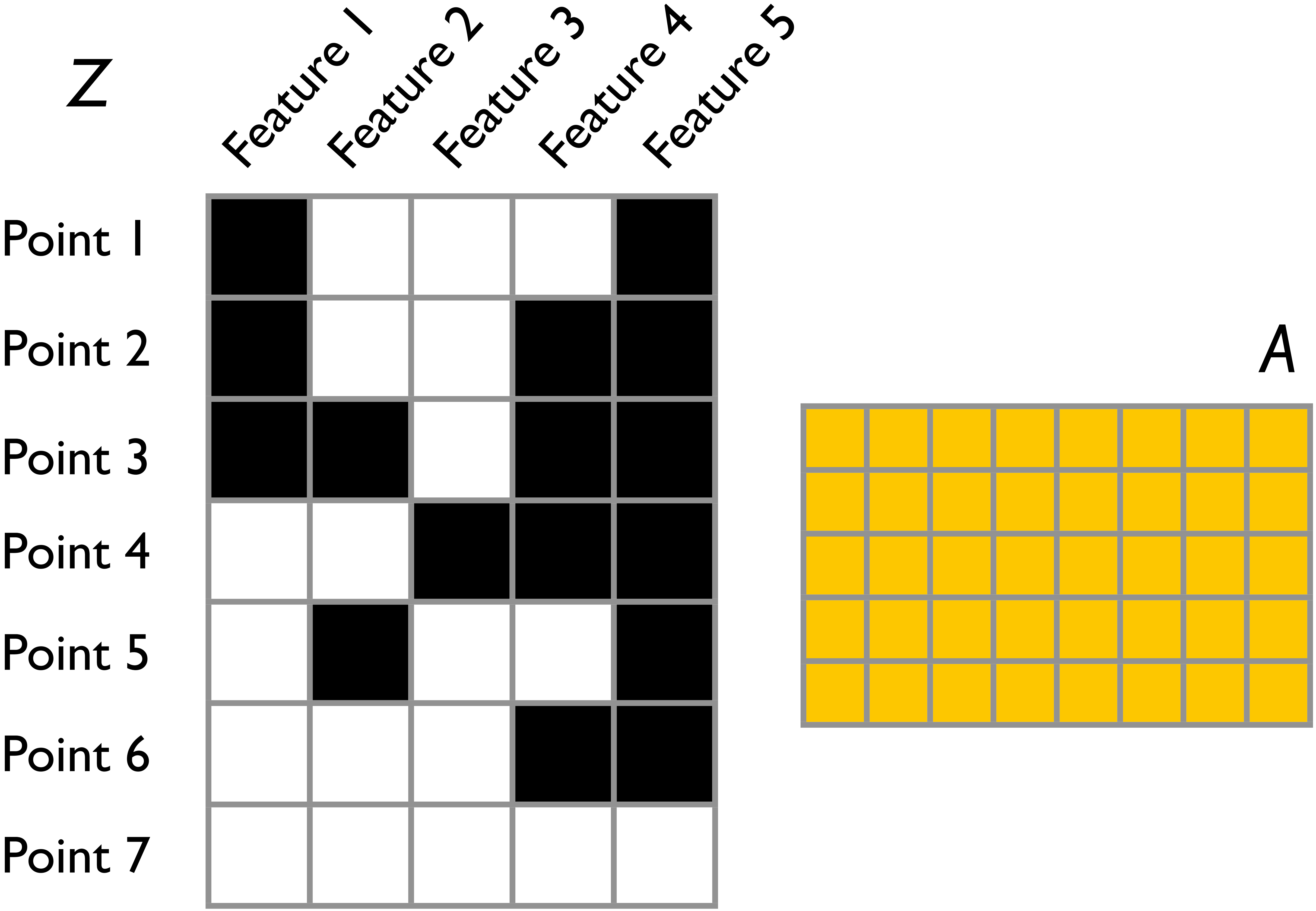

Z
Feature 1
Feature 2
Feature 3
Feature 4
Feature 5
Point 1
Point 2
Point 3
Point 4
Point 5
Point 6
Point 7

Features



Z
Feature 1
Feature 2
Feature 3
Feature 4
Feature 5
Point 1
Point 2
Point 3
Point 4
Point 5
Point 6
Point 7
A

Features


Z
Feature 1
Feature 2
Feature 3
Feature 4
Feature 5
Point 1
Point 2
Point 3
Point 4
Point 5
Point 6
Point 7
A

MAD-Bayes

Bayesian posterior

$$
\begin{aligned}
&\mathbb{P}(Z, A|X) \\
&\qquad \propto \frac{1}{(2\pi\sigma^2)^{ND/2}} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{tr}((X-ZA)'(X-ZA))\right\} \\
&\qquad \cdot \frac{\gamma^{K^+} \exp\left\{-\sum_{n=1}^N \frac{\gamma}{n}\right\}}{\prod_{h=1}^H \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k}-1)!(N-S_{N,k})!}{N!} \\
&\qquad \cdot \frac{1}{(2\pi\rho^2)^{K^+D/2}} \exp\left\{-\frac{1}{2\rho^2}\mathbf{tr}(A'A)\right\}.
\end{aligned}
$$

MAD-Bayes

Bayesian posterior

$$\mathbb{P}(Z, A|X)$$

$$\propto \frac{1}{(2\pi\sigma^2)^{ND/2}} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{tr}((X - ZA)'(X - ZA))\right\}$$

$$\cdot \frac{\gamma^{K^+} \exp\left\{-\sum_{n=1}^{N} \frac{\gamma}{n}\right\}}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k} - 1)!(N - S_{N,k})!}{N!}$$

$$\cdot \frac{1}{(2\pi\rho^2)^{K^+D/2}} \exp\left\{-\frac{1}{2\rho^2}\mathbf{tr}(A'A)\right\}.$$

MAD-Bayes

Bayesian posterior

$$
\begin{aligned}
\mathbb{P}(Z, A | X) \\
&\propto \frac{1}{(2\pi\sigma^2)^{ND/2}} \exp\left\{ -\frac{1}{2\sigma^2} \mathbf{tr}((X - ZA)'(X - ZA)) \right\} \\
&\cdot \frac{\gamma^{K^+} \exp\left\{ -\sum_{n=1}^{N} \frac{\gamma}{n} \right\}}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k} - 1)!(N - S_{N,k})!}{N!} \\
&\cdot \frac{1}{(2\pi\rho^2)^{K^+ D/2}} \exp\left\{ -\frac{1}{2\rho^2} \mathbf{tr}(A'A) \right\}.
\end{aligned}
$$

MAD-Bayes

Bayesian posterior

$$\mathbb{P}(Z, A | X)$$

$$\propto \frac{1}{(2\pi\sigma^2)^{ND/2}} \exp\left\{ -\frac{1}{2\sigma^2} \mathbf{tr}((X - ZA)'(X - ZA)) \right\}$$

$$\cdot \frac{\gamma^{K^+} \exp\left\{ -\sum_{n=1}^{N} \frac{\gamma}{n} \right\}}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k} - 1)!(N - S_{N,k})!}{N!}$$

$$\cdot \frac{1}{(2\pi\rho^2)^{K^+ D/2}} \exp\left\{ -\frac{1}{2\rho^2} \mathbf{tr}(A'A) \right\}.$$

MAD-Bayes

Bayesian posterior

$$
\begin{aligned}
&\mathbb{P}(Z, A|X) \\
&\qquad \propto \frac{1}{(2\pi\sigma^2)^{ND/2}} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{tr}((X - ZA)'(X - ZA))\right\} \\
&\qquad \cdot \frac{\gamma^{K^+} \exp\left\{-\sum_{n=1}^{N} \frac{\gamma}{n}\right\}}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k} - 1)!(N - S_{N,k})!}{N!} \\
&\qquad \cdot \frac{1}{(2\pi\rho^2)^{K^+ D/2}} \exp\left\{-\frac{1}{2\rho^2}\mathbf{tr}(A'A)\right\}.
\end{aligned}
$$

MAD-Bayes

Bayesian posterior

$$\mathbb{P}(Z, A|X)$$

$$\propto \frac{1}{(2\pi\sigma^2)^{ND/2}} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{tr}((X - ZA)'(X - ZA))\right\}$$

$$\cdot \frac{\gamma^{K^+} \exp\left\{-\sum_{n=1}^{N} \frac{\gamma}{n}\right\}}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k} - 1)!(N - S_{N,k})!}{N!}$$

$$\cdot \frac{1}{(2\pi\rho^2)^{K^+D/2}} \exp\left\{-\frac{1}{2\rho^2}\mathbf{tr}(A'A)\right\}.$$

MAD-Bayes

Bayesian posterior

$$
\begin{aligned}
&\mathbb{P}(Z, A \mid X) \\
&\qquad \propto \frac{1}{(2\pi\sigma^2)^{ND/2}} \exp\left\{ -\frac{1}{2\sigma^2} \mathbf{tr}((X - ZA)'(X - ZA)) \right\} \\
&\qquad \cdot \frac{\gamma^{K^+} \exp\left\{ -\sum_{n=1}^{N} \frac{\gamma}{n} \right\}}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k} - 1)!(N - S_{N,k})!}{N!} \\
&\qquad \cdot \frac{1}{(2\pi\rho^2)^{K^+ D/2}} \exp\left\{ -\frac{1}{2\rho^2} \mathbf{tr}(A'A) \right\}.
\end{aligned}
$$

MAD-Bayes

BP-means objective

$$\mathrm{argmin}_{K^+,Z,A}\mathbf{tr}[(X-ZA)'(X-ZA)]+K^+\lambda^2.$$

MAD-Bayes

BP-means objective

$$\mathrm{argmin}_{K^+,Z,A} \mathbf{tr}[(X - ZA)'(X - ZA)] + K^+\lambda^2.$$

MAD-Bayes

BP-means objective

$$\mathrm{argmin}_{K^+,Z,A}\mathbf{tr}[(X - ZA)'(X - ZA)] + K^+\lambda^2.$$

MAD-Bayes

BP-means objective

$$\text{argmin}_{K^+,Z,A}\mathbf{tr}[(X - ZA)'(X - ZA)] + K^+\lambda^2.$$

BP-means algorithm

Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to features
   - Create a new feature if it lowers the objective
2. Update feature means $A \leftarrow (Z'Z)^{-1}Z'X$

MAD-Bayes

BP-means objective

$$\text{argmin}_{K^+,Z,A}\mathbf{tr}[(X-ZA)'(X-ZA)]+K^+\lambda^2.$$

BP-means algorithm

Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to features
   - Create a new feature if it lowers the objective
2. Update feature means $A \leftarrow (Z'Z)^{-1}Z'X$

MAD-Bayes

BP-means objective

$$\mathrm{argmin}_{K^+, Z, A} \mathbf{tr}[(X - ZA)'(X - ZA)] + K^+ \lambda^2.$$

BP-means algorithm

Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to features
   - Create a new feature if it lowers the objective
2. Update feature means $A \leftarrow (Z'Z)^{-1} Z'X$

MAD-Bayes

BP-means objective

$$\text{argmin}_{K^+,Z,A}\mathbf{tr}[(X - ZA)'(X - ZA)] + K^+\lambda^2.$$

BP-means algorithm

Iterate until no changes:

1. For n = 1, ..., N
   - Assign point n to features
   - Create a new feature if it lowers the objective
2. Update feature means $A \leftarrow (Z'Z)^{-1}Z'X$

MAD-Bayes

Griffiths & Ghahramani (2006) computer vision problem "tabletop data"

Bayesian posterior
Gibbs sampler

BP-means algorithm

$8.5 * 10^3$ sec

0.36 sec

Still faster by order of magnitude
if restart 1000 times

MAD-Bayes

Parallelism and optimistic concurrency control

	DP-means alg.	BP-means alg.
# data points	**134M**	**8M**
time per iteration	**5.5 min**	**4.3 min**

MAD-Bayes

Bayesian posterior	K-means-like objectives
Mixture of K Gaussians	K-means
Dirichlet process mixture	Unbounded number of clusters
Hierarchical Dirichlet process	Multiple data sets share cluster centers
⋮	⋮
Beta process	Features

MAD-Bayes conclusions

MAD-Bayes conclusions

- We provide new optimization objectives and regularizers

MAD-Bayes conclusions

- We provide new optimization objectives and regularizers
  - ◇ In fact, general means of obtaining more

MAD-Bayes conclusions

- We provide new optimization objectives and regularizers
  - In fact, general means of obtaining more
  - Straightforward, fast algorithms

References

T. Broderick, B. Kulis, and M. I. Jordan. MAD-Bayes: MAP-based asymptotic derivations from Bayes. In *International Conference on Machine Learning*, 2013.

X. Pan, J. E. Gonzales, S. Jegelka, T. Broderick, and M. I. Jordan. Optimistic concurrency control for distributed unsupervised learning. In *Neural Information Processing Systems*, 2013.

T. Broderick, N. Boyd, A. Wibisono, A. C. Wilson, and M. I. Jordan. Streaming variational Bayes. In *Neural Information Processing Systems*, 2013.

Further References

T. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *Neural Information Processing Systems*, 2006.

N. L. Hjort. Nonparametric Bayes estimators based on beta processes in models for life history data. *Annals of Statistics*, 18(3):1259–1294, 1990.

J. F. C. Kingman. The representation of partition structures. *Journal of the London Mathematical Society*, 2(2):374, 1978.

B. Kulis and M. I. Jordan. Revisiting k-means: New algorithms via Bayesian nonparametrics. In *International Conference on Machine Learning*, 2012.

J. Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158, 1995.

R. Thibaux and M. I. Jordan. Hierarchical beta processes and the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, 2007.