Mid-Term

- Remote exams:
  - Should already be arranged via **IIT-Online** or other IIT dept.
  - Email questions to Charles (Chuck) Scott at IIT-Online
- Covers readings and lecture slides (with focus on material in that appears in both)
- Format:
  - Closed Book – No printouts or cheat sheet, no calculator (not needed)
  - Multiple Choice, True/False, or similar
  - "Bubble" answer sheet (Fill in circles)
  - **Bring pencils (with erasers)!**

Mid-Term

How can I prepare this weekend?

- **Buy pencils (with erasers)!**
- Catch up on readings (see syllabus on Blackboard)
- Revisit slides – can you explain concepts to a friend?
- Formulas:
  - What is being defined or modeled?
  - How and where would it be applied?
  - For measures: What does a high and low value represent? What is a valid range (and why)?

Unsupervised Learning

CS-585

Natural Language Processing

Sonjia Waxmonsky

(with slides from William W. Cohen and Chris Manning)

REVISITING UNSUPERVISED METHODS

Latent semantic analysis

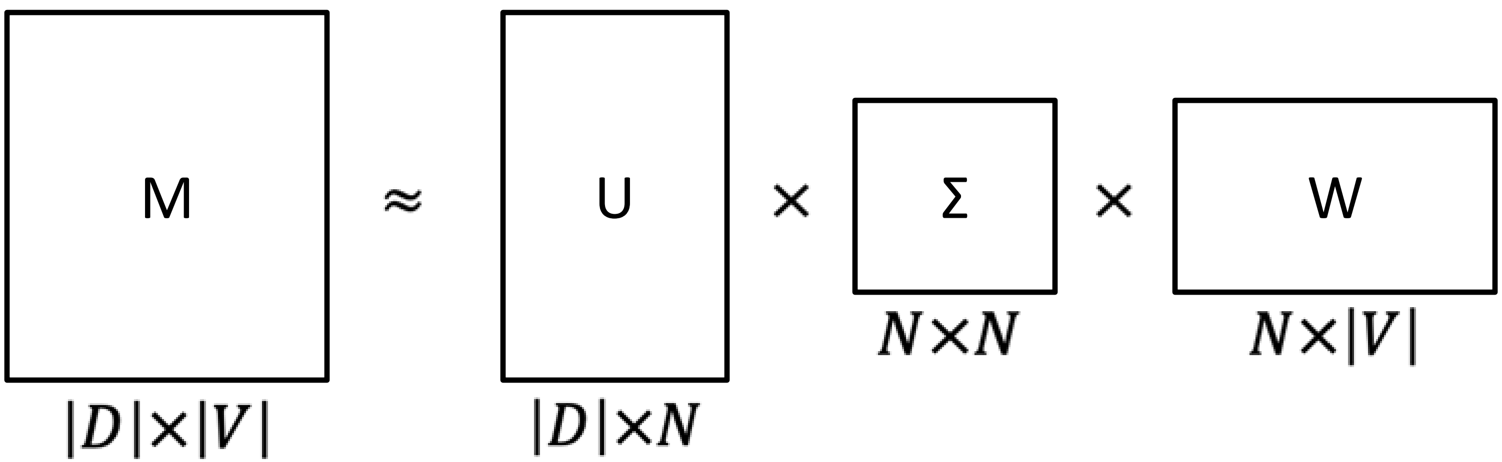
From Session 6

- The idea is to "compress" the representation of a word, using only $M \ll |D|$ dimensions for each vector
  - Compress for more efficient representation (smaller memory footprint)
  - Compress for *generalization*: retain only most important information, and allow distinctions between similar words to be obscured
- How to do this automatically?

Singular value decomposition

From Session 6

For a term-by-document matrix M, based on documents D and vocabulary V, we approximate



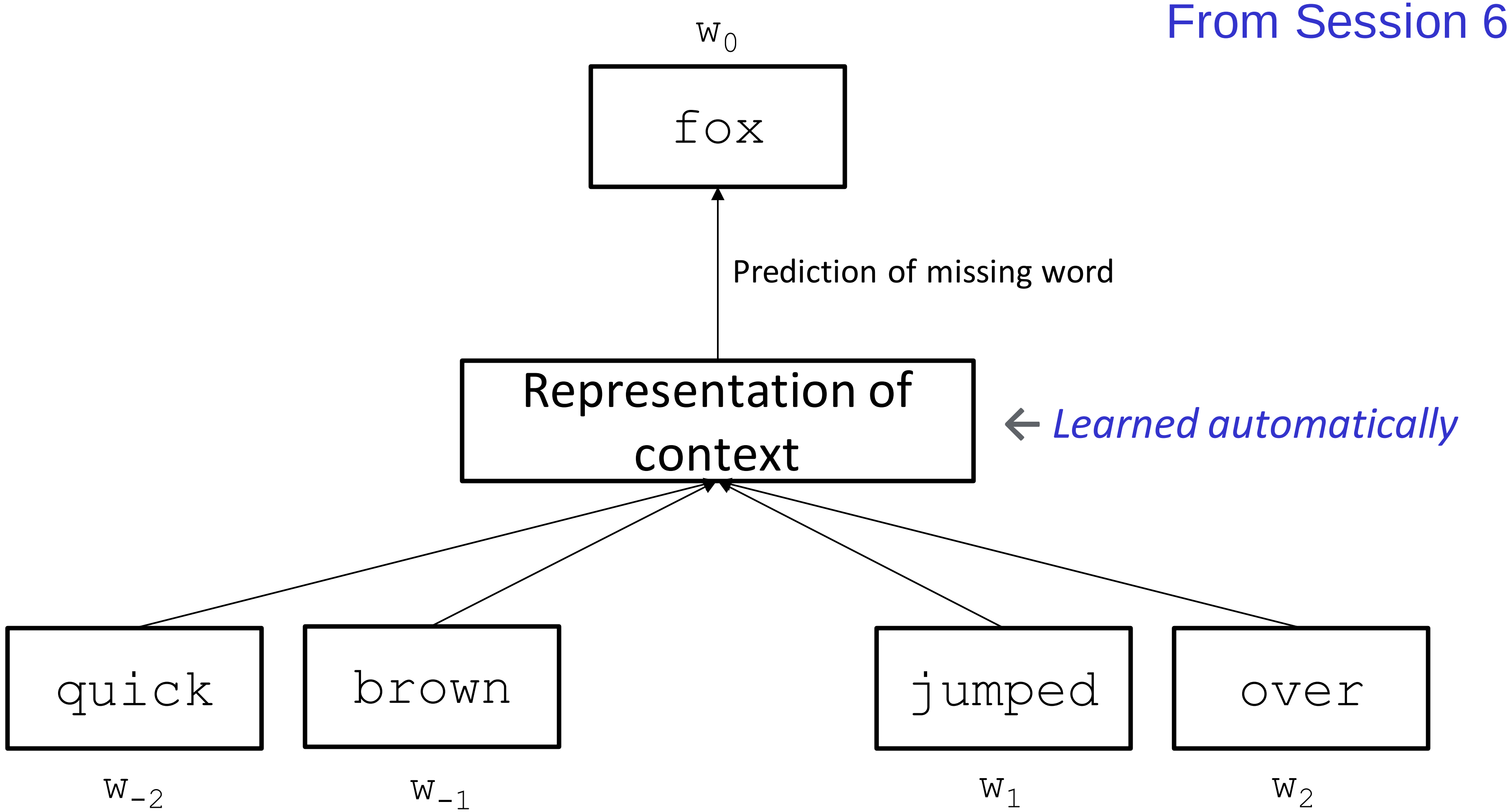
U is an orthogonal matrix with one row per document

W is an orthogonal matrix with one column per word

Σ is a diagonal matrix of "singular values"

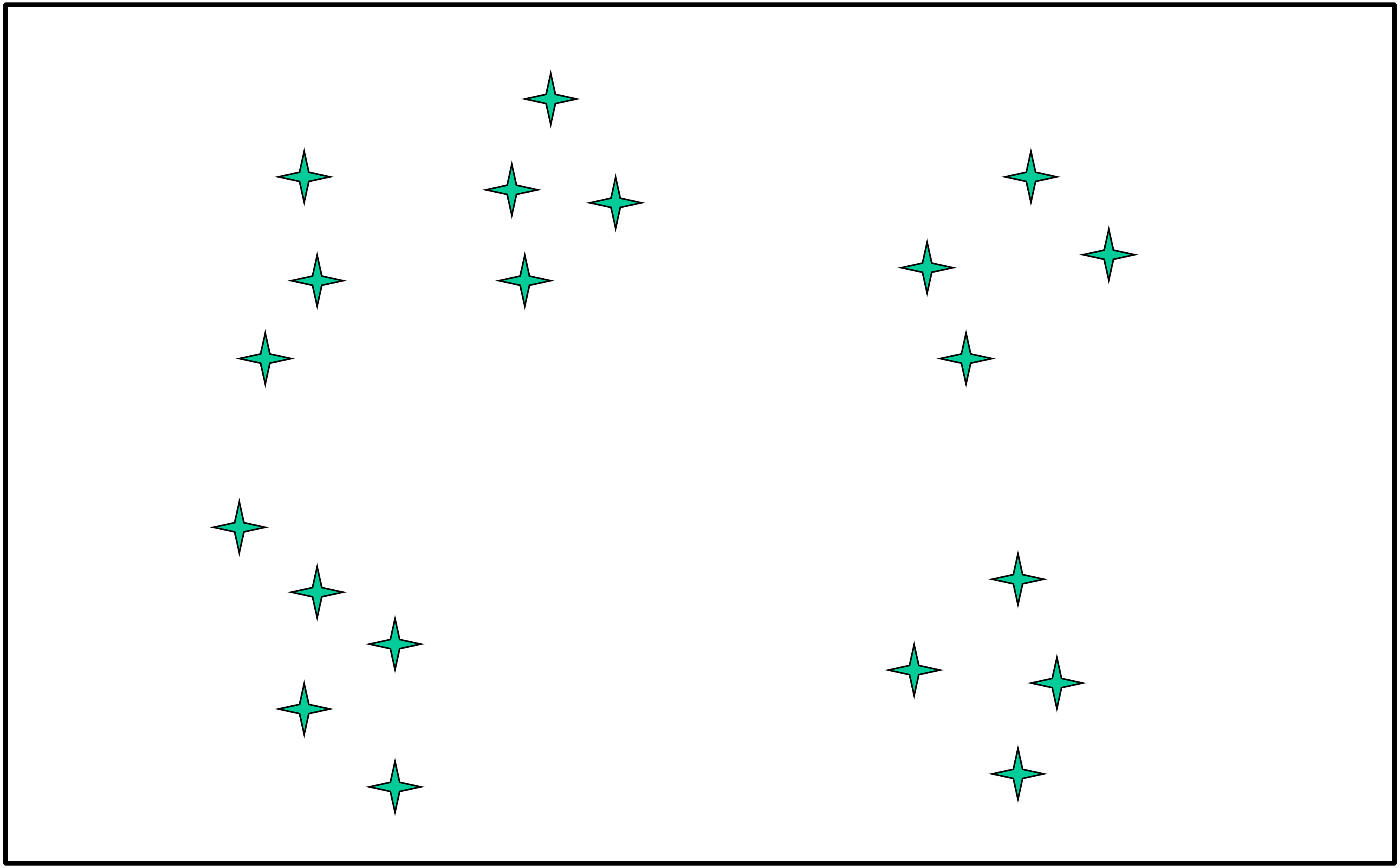
word2vec
Continuous Bag of Words (CBOW)

From Session 6

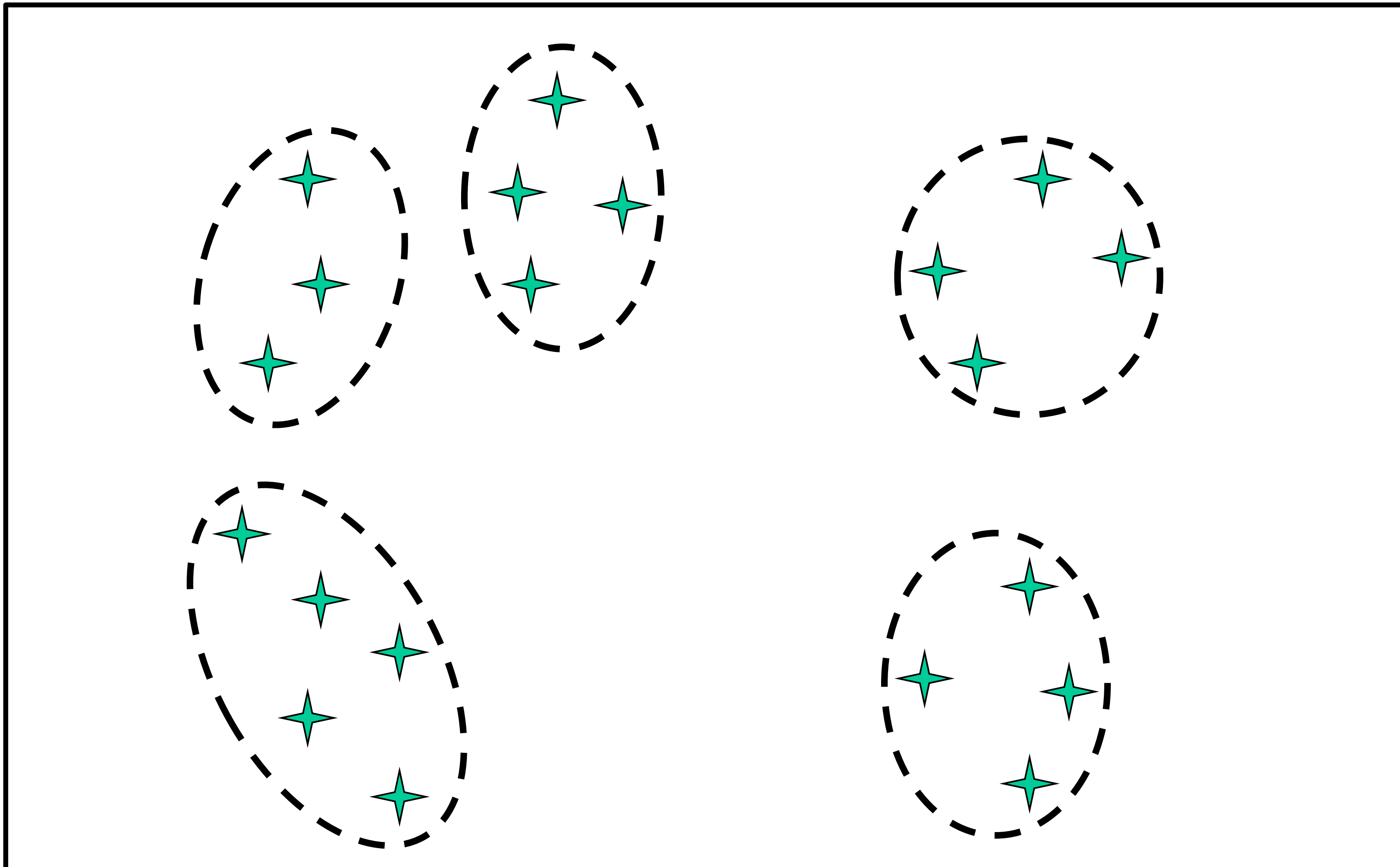


UNSUPERVISED CLASSIFICATION

Clustering



Clustering



Clustering on Text

FOR SALE: CHEAP LOGIC BOARDS!!! (update)	3
Ampex 456 2" Recording Tape For Sale	3
The Bob Dylan Baseball Abstract	1
Patient-Physician Diplomacy	2
Defensive Averages 1988-1992, Third Base	1
Dana-Faber Cancer Institute	3
Ryan rumor...	1
MS-Windows graphics viewer?	3
Jack Morris	1
Candida Albicans: what is it?	3

← 2 or 3?

Clustering

- Unsupervised: Input is features, or vector representation of data, rather than (feature, label) pair

- *Identify the* ***underlying structure of the observed data,*** *such that there are a few clusters of points, each of which is internally coherent* [Eisenstein-NLP]

- Hierarchical Clustering – nested clusters
- Non-hierarchical – clusters do not overlap

Non-Hierarchical Clustering

- **Iterative clustering:**
  - Start with initial (random) set of clusters
  - Assign each object to a cluster (or clusters)
  - Recompute cluster parameters
  - Stop when clustering is “good”

- Q: How many clusters?

  A: Who knows??

K-means Algorithm

Input:

- Set $X = \{x_1, \ldots, x_n\}$ of objects
- Distance measure d: $X \times X \rightarrow \mathbb{R}$
- Mean function μ

Select k initial cluster centers $f_1, \ldots, f_k$

while not finished do:

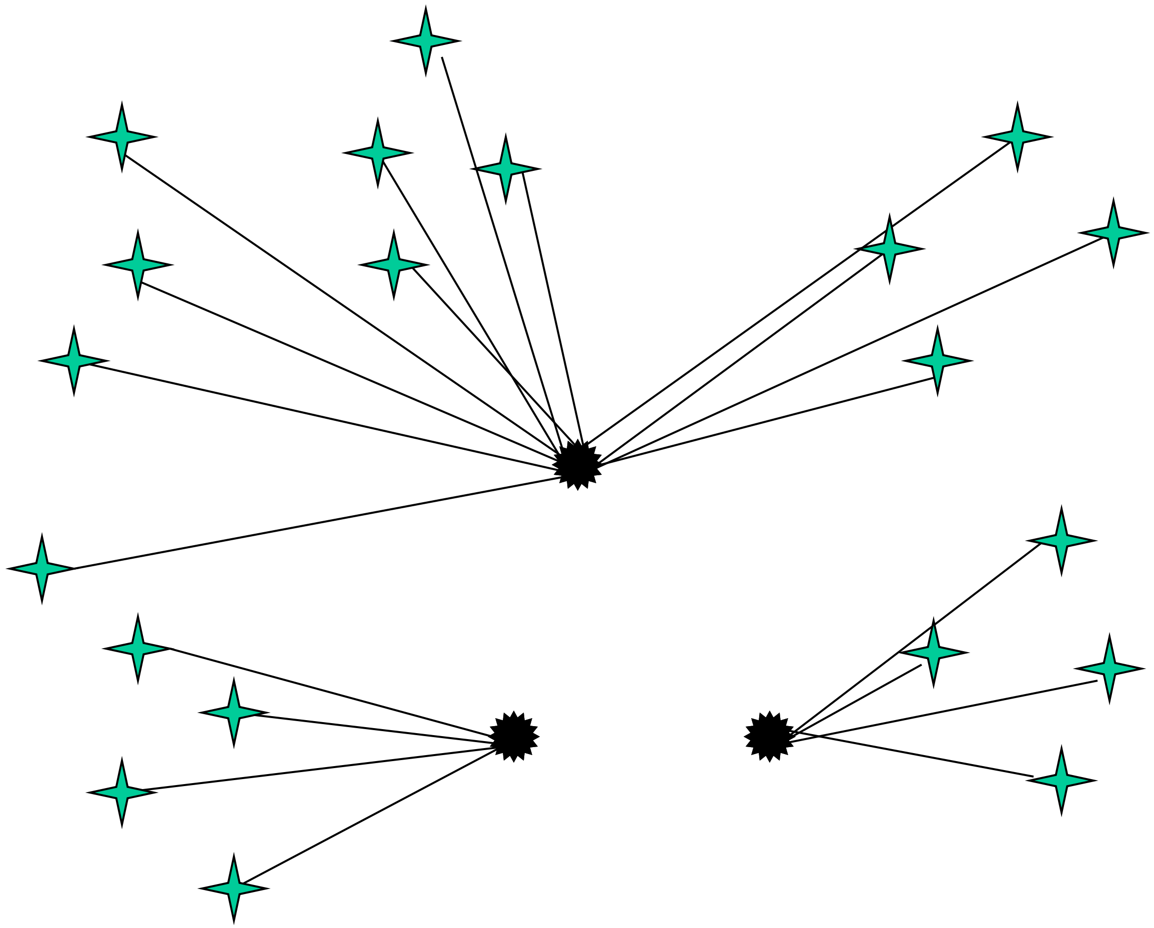
  for all clusters c_j do:

    $c_j \leftarrow \{ x_i \mid f_j = \mathbf{argmin}_f \, d(x_i, f) \}$

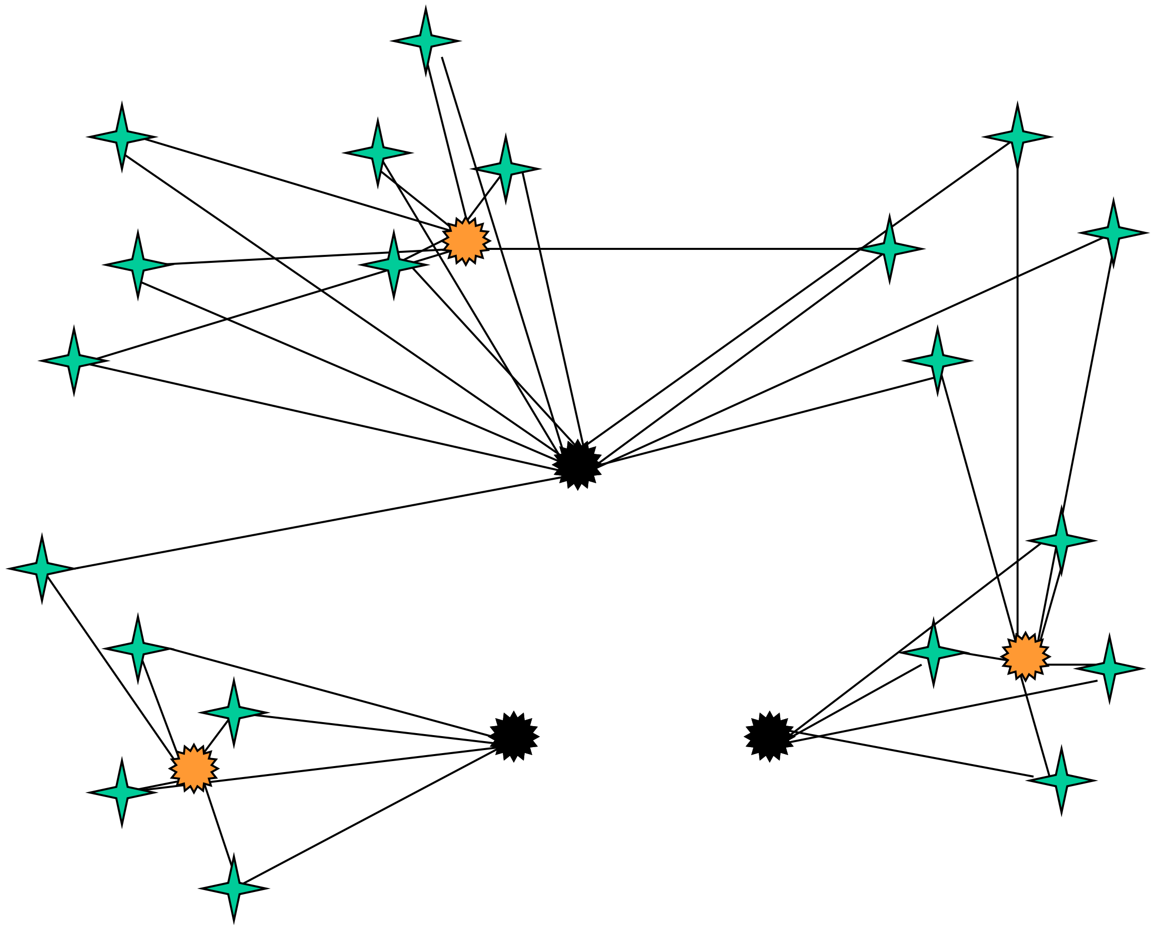
  for all means f_j do:

    $f_j \leftarrow \mu(c_j)$

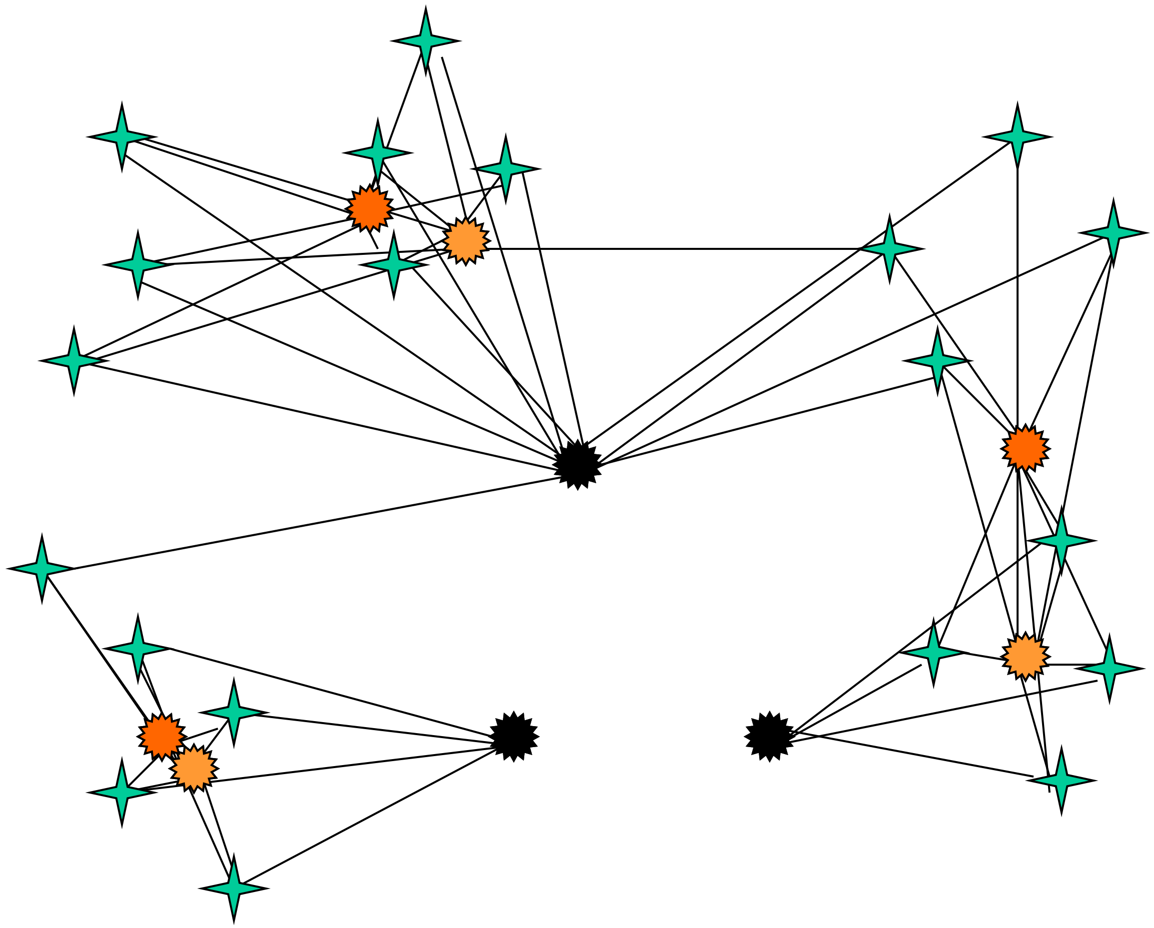
K-means Clustering



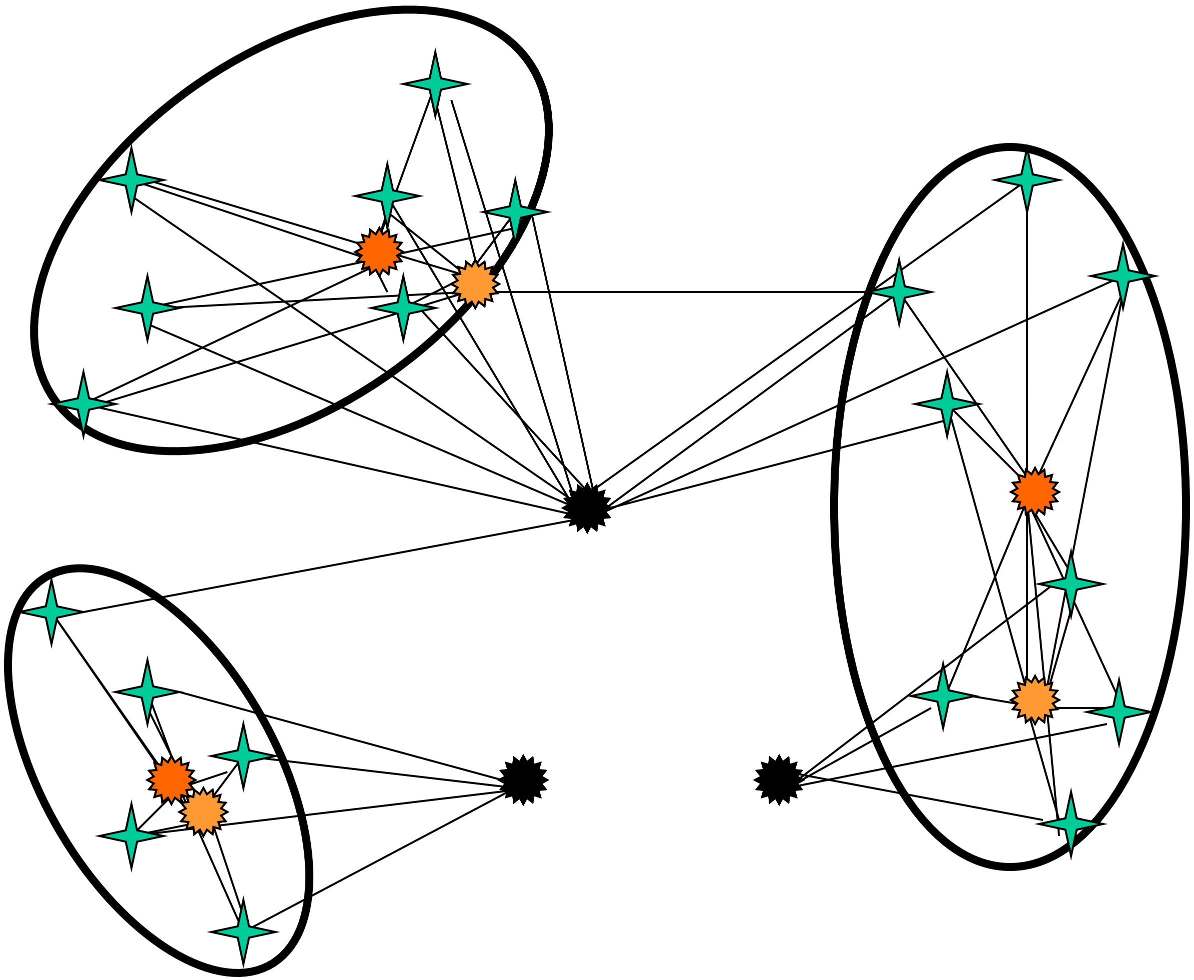
K-means Clustering



K-means Clustering



K-means Clustering



K-means as EM (ish)

E: Calculate cluster assignments given current centroid locations

Data point	Location	Closest cluster centroid
1	(-1,1)	2
2	(-1,-1)	3
3	(1,2)	1
4	(2,2)	1
5	(-2,1)	2
6	(-2,-2)	3
7	(-3,-1)	3
8	(4,2)	1
9	(-1,0)	2

Can this be a “soft” assignment?
(Soft K-means)

K-means as EM (ish)

M: Move the cluster centroids to the center of their associated data points (making the data more “likely”)

Data point	Location	Closest cluster centroid
1	(-1,1)	2
2	(-1,-1)	3
3	(1,2)	1
4	(2,2)	1
5	(-2,1)	2
6	(-2,-2)	3
7	(-3,-1)	3
8	(4,2)	1
9	(-1,0)	2

Cluster	New centroid
1	(2.33,2)
2	(-1.33,0.67)
3	(-2,-1.33)

$$\mathrm{mean}([1,2],[2,2],[4,2])$$

$$\mathrm{mean}([-1,1],[-2,1],[-1,0])$$

$$\mathrm{mean}([-1,-1],[-2,-2],[-3,-1])$$

EXPECTATION MAXIMIZATION

Expectation-Maximization

- Combines **generative** modeling of Naïve Bayes with **latent variable** modeling of soft K-means
- Review:
  - Generative model – Models joint probability of class labels and observations
  - Discriminative model – Models conditional probability of labels given observations
- E-M framework (not a single algorithm) is **iterative** framework to find **maximum likelihood estimate** given a set of latent variables

The EM Algorithm

Soft clustering method to solve

$$\theta^* = \arg\max_\theta P_{model}(X \mid \theta)$$

Note: Any occurrence of the data consists of:

- **Observable variables:** The objects we see
  - *Bags of words*
  - *Word sequences in tagging tasks*
- **Hidden variables:** Which cluster generated which object
  - *Document categories*
  - *Underlying tag sequences*

Two Principles

Expectation: If we knew θ we could compute the expected values of the hidden variables (e.g, probability of x belonging to some cluster)

Maximization: If we knew the hidden structure, we could compute the maximum likelihood value of θ

Iterative Solution

Initialize: Choose an initial θ_0

Then iterate until convergence:

– E-step: Compute $(X, Z_i) = \text{Exp}[X, Z \mid \theta_i]$

– M-step: Choose θ_{i+1} to maximize $P(X, Z_i, \theta_{i+1})$

M-step sometimes cannot be computed, but moving along its gradient also works

EM for Naive Bayes Text Classification

E-step: Compute $P(c_k \mid d_i)$ for each document d_i and category c_k given current model

M-step: Re-estimate the model parameters $P(w_j \mid c_k)$ and $P(c_k)$

Continue as long as log-likelihood of corpus increases:

$$\log \prod_i \sum_k P(d_i \mid c_k) P(c_k) = \sum_i \log \sum_k P(d_i \mid c_k) P(c_k)$$

E-Step

- For each document d_i and each category c_k, estimate the posterior probability $h_{ik} = P(c_k \mid d_i)$:

$$h_{ik} = \frac{P(d_i \mid c_k)P(c_k)}{\sum_{k'} P(d_i \mid c_{k'})P(c_{k'})}$$

- To compute $P(d_i \mid c_k)$, use naive Bayes:

$$P(d_i \mid c_k) = \prod_{w_j \in d_k} P(w_j \mid c_k)$$

M-Step

Re-estimate parameters using maximum-likelihood estimation:

$$P(w_j|c_k) = \frac{\sum_{d_i:w_j \in d_i} h_{ik}}{\sum_{d_i,\forall w_{j'} \in d_i} h_{ik}}$$

$$P(c_k) = \frac{\sum_i h_{ik}}{\sum_k \sum_i h_{ik}}$$

Decision Procedure

Assign categories by:

$$cat(d_i) = \arg\max_{c_k} \left[\log P(c_k) + \sum_{w_j \in d_i} \log P(w_j \mid c_k) \right]$$

- Can adjust number of categories *k* to get finer or coarser distinctions
- If adding more categories doesn't increase log-likelihood of data much, then stop

Applications of E-M

Clustering based on local context applicable to many NLP tasks:

- Word-sense induction – Overcome limitations of hand-annotation
- Part-of-speech tagging – For low-resource languages (limited annotated data)
- … and other NLP tasks