Lexical Semantics

Introduction to Natural Language Processing
Computer Science 585—Fall 2009
University of Massachusetts Amherst

David Smith
with slides from Jason Eisner

Overview

- Semantics so far: compositional semantics
  - How to put together propositions from atomic meanings (lexicon)?
- Now: lexical semantics
  - What are those atomic meanings?
- Today: clustering words with similar senses
- Next time: sense disambiguation, functional clustering
- Manning & Schütze ch. 5 & 7

A Concordance for “party”

- thing. She was talking at a party thrown at Daphne's restaurant in
- have turned it into the hot dinner-party topic. The comedy is the
- selection for the World Cup party, which will be announced on May 1
- in the 1983 general election for a party which, when it could not bear to
- to attack the Scottish National Party, who look set to seize Perth and
- that had been passed to a second party who made a financial decision
- the by-pass there will be a street party. "Then," he says, "we are going
- number-crunchers within the Labour party, there now seems little doubt
- political tradition and the same party. They are both relatively Anglophilic
- he told Tony Blair's modernised party they must not retreat into "warm
- "Oh no, I'm just here for the party," they said. "I think it's terrible
- A future obliges each party to the contract to fulfil it by
- be signed by or on behalf of each party to the contract." Mr David N

What Good are Word Senses?

- thing. She was talking at a party thrown at Daphne's restaurant in
- have turned it into the hot dinner-party topic. The comedy is the
- selection for the World Cup party, which will be announced on May 1
- in the 1983 general election for a party which, when it could not bear to
- to attack the Scottish National Party, who look set to seize Perth and
- that had been passed to a second party who made a financial decision
- the by-pass there will be a street party. "Then," he says, "we are going
- number-crunchers within the Labour party, there now seems little doubt
- political tradition and the same party. They are both relatively Anglophilic
- he told Tony Blair's modernised party they must not retreat into "warm
- "Oh no, I'm just here for the party," they said. "I think it's terrible
- A future obliges each party to the contract to fulfil it by
- be signed by or on behalf of each party to the contract." Mr David N

What Good are Word Senses?

- thing. She was talking at a party thrown at Daphne's restaurant in
- have turned it into the hot dinner-party topic. The comedy is the
- selection for the World Cup party, which will be announced on May 1
- the by-pass there will be a street party. "Then," he says, "we are going
- "Oh no, I'm just here for the party," they said. "I think it's terrible

- in the 1983 general election for a party which, when it could not bear to
- to attack the Scottish National Party, who look set to seize Perth and
- number-crunchers within the Labour party, there now seems little doubt
- political tradition and the same party. They are both relatively Anglophilic
- he told Tony Blair's modernised party they must not retreat into "warm

- that had been passed to a second party who made a financial decision
- A future obliges each party to the contract to fulfil it by
- be signed by or on behalf of each party to the contract." Mr David N

What Good are Word Senses?

- John threw a “rain forest” party last December. His living room was full of plants and his box was playing Brazilian music …

What Good are Word Senses?

- Replace word w with sense s
  - Splits w into senses: distinguishes this token of w from tokens with sense t
  - Groups w with other words: groups this token of w with tokens of x that also have sense s

What Good are Word Senses?

- number-crunchers within the Labour party, there now seems little doubt
- political tradition and the same party. They are both relatively Anglophilic
- he told Tony Blair's modernised party they must not retreat into "warm
- thing. She was talking at a party thrown at Daphne's restaurant in
- have turned it into the hot dinner-party topic. The comedy is the
- selection for the World Cup party, which will be announced on May 1
- the by-pass there will be a street party. "Then," he says, "we are going
- "Oh no, I'm just here for the party," they said. "I think it's terrible

- an appearance at the annual awards bash , but feels in no fit state to
- -known families at a fundraising bash on Thursday night for Learning
- Who was paying for the bash? The only clue was the name Asprey,
- Mail, always hosted the annual bash for the Scottish Labour front-
- popular. Their method is to bash sense into criminals with a short,
- just cut off people's heads and bash their brains out over the floor,

What Good are Word Senses?

- number-crunchers within the Labour party, there now seems little doubt
- political tradition and the same party. They are both relatively Anglophilic
- he told Tony Blair's modernised party they must not retreat into "warm

- thing. She was talking at a party thrown at Daphne's restaurant in
- have turned it into the hot dinner-party topic. The comedy is the
- selection for the World Cup party, which will be announced on May 1
- the by-pass there will be a street party. "Then," he says, "we are going
- "Oh no, I'm just here for the party," they said. "I think it's terrible
- an appearance at the annual awards bash, but feels in no fit state to
- -known families at a fundraising bash on Thursday night for Learning
- Who was paying for the bash? The only clue was the name Asprey,
- Mail, always hosted the annual bash for the Scottish Labour front-

- popular. Their method is to bash sense into criminals with a short,
- just cut off people's heads and bash their brains out over the floor,

What Good are Word Senses?

What Good are Word Senses?

- Semantics / Text understanding
  - Axioms about TRANSFER apply to (some tokens of) `throw`
  - Axioms about BUILDING apply to (some tokens of) `bank`

What Good are Word Senses?

- Semantics / Text understanding
  - Axioms about TRANSFER apply to (some tokens of) `throw`
  - Axioms about BUILDING apply to (some tokens of) `bank`
- Machine translation

What Good are Word Senses?

- Semantics / Text understanding
  - Axioms about TRANSFER apply to (some tokens of) `throw`
  - Axioms about BUILDING apply to (some tokens of) `bank`
- Machine translation
- Info retrieval / Question answering / Text categ.
  - Query or pattern might not match document exactly

What Good are Word Senses?

- Semantics / Text understanding
  - Axioms about TRANSFER apply to (some tokens of) `throw`
  - Axioms about BUILDING apply to (some tokens of) `bank`
- Machine translation
- Info retrieval / Question answering / Text categ.
  - Query or pattern might not match document exactly
- Backoff for just about anything
  - what word comes next? (speech recognition, language ID, …)
    - trigrams are sparse but tri-meanings might not be
  - bilexical PCFGs: p(S[`devour`] → NP[`lion`] VP[`devour`] | S[`devour`])
    - approximate by p(S[EAT] → NP[`lion`] VP[EAT] | S[EAT])

What Good are Word Senses?

- Semantics / Text understanding
  - Axioms about TRANSFER apply to (some tokens of) `throw`
  - Axioms about BUILDING apply to (some tokens of) `bank`
- Machine translation
- Info retrieval / Question answering / Text categ.
  - Query or pattern might not match document exactly
- Backoff for just about anything
  - what word comes next? (speech recognition, language ID, …)
    - trigrams are sparse but tri-meanings might not be
  - bilexical PCFGs: p(S[`devour`] → NP[`lion`] VP[`devour`] | S[`devour`])
    - approximate by p(S[EAT] → NP[`lion`] VP[EAT] | S[EAT])
- Speaker's real intention is senses; words are a noisy channel

Cues to Word Sense

Cues to Word Sense

- Adjacent words (or their senses)

Cues to Word Sense

- Adjacent words (or their senses)
- Grammatically related words (subject, object, …)

Cues to Word Sense

- Adjacent words (or their senses)
- Grammatically related words (subject, object, …)
- Other nearby words

Cues to Word Sense

- Adjacent words (or their senses)
- Grammatically related words (subject, object, …)
- Other nearby words
- Topic of document

Cues to Word Sense

- Adjacent words (or their senses)
- Grammatically related words (subject, object, …)
- Other nearby words
- Topic of document
- Sense of other tokens of the word in the same document

Word Classes by Tagging

- Every tag is a kind of class
- Tagger assigns a class to each word token

Word Classes by Tagging

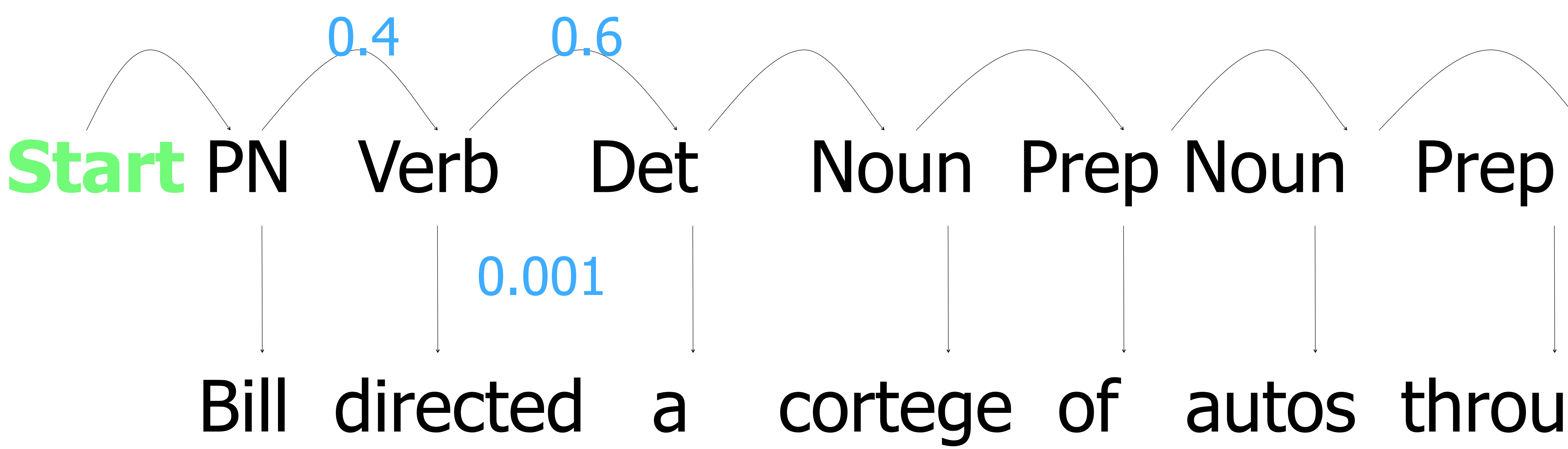
- Every tag is a kind of class
- Tagger assigns a class to each word token

Start PN Verb Det Noun Prep Noun Prep

Bill directed a cortege of autos throu

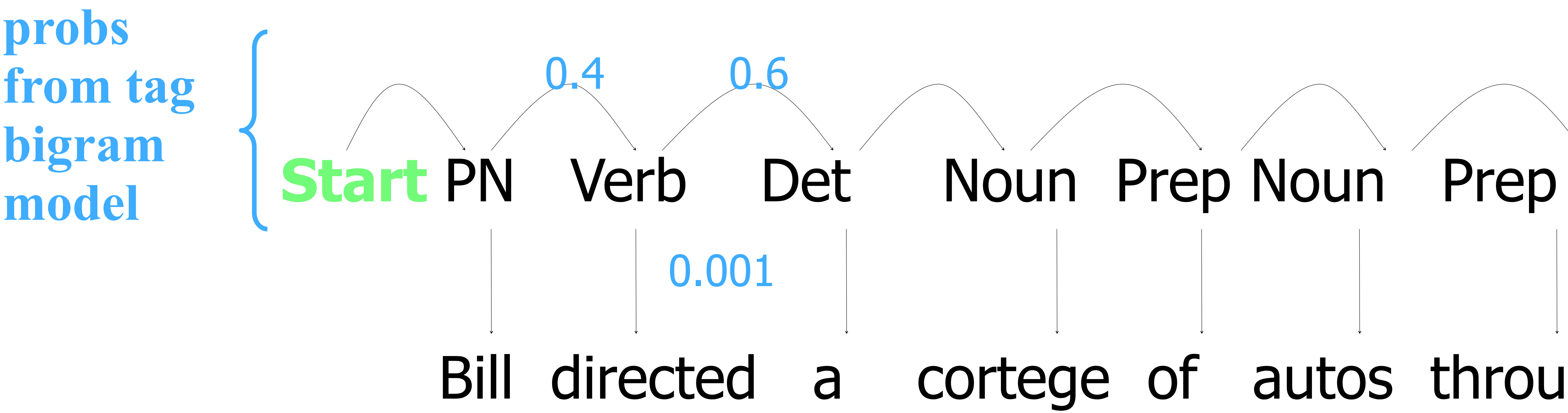
Word Classes by Tagging

- Every tag is a kind of class
- Tagger assigns a class to each word token



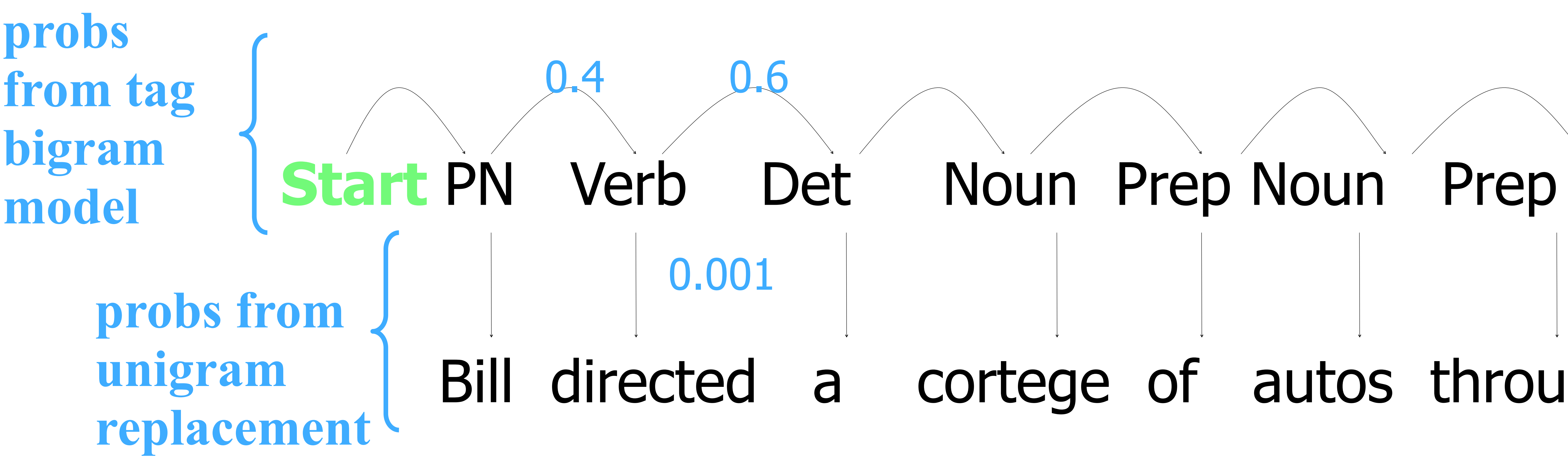
Word Classes by Tagging

- Every tag is a kind of class
- Tagger assigns a class to each word token



Word Classes by Tagging

- Every tag is a kind of class
- Tagger assigns a class to each word token



Word Classes by Tagging

- Every tag is a kind of class
- Tagger assigns a class to each word token
  - Simultaneously groups and splits words
  - "party" gets split into N and V senses
  - "bash" gets split into N and V senses
  - {party/N, bash/N} vs. {party/V, bash/V}
  - What good are these groupings?

Learning Word Classes

- Every tag is a kind of class
- Tagger assigns a class to each word token
  - {party/N, bash/N} vs. {party/V, bash/V}
  - What good are these groupings?
  - Good for predicting next word or its class!
- Role of forward-backward algorithm?
  - It adjusts classes etc. in order to predict sequence of words better (with lower perplexity)

Words as Vectors

- Represent each word **type** w by a point in k-dimensional space
  - e.g., k is size of vocabulary
  - the 17^{th} coordinate of w represents **strength** of w's association with vocabulary word 17

Words as Vectors

- Represent each word **type** w by a point in k-dimensional space
  = party
  - e.g., k is size of vocabulary
  - the 17^{th} coordinate of w represents **strength** of w's association with vocabulary word 17

Words as Vectors

- Represent each word **type** w by a point in k-dimensional space
  - e.g., k is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

= party

(0, 0, 3, 1, 0, 7, . . . 1, 0)

Words as Vectors

- Represent each word **type** w by a point in k-dimensional space
  = party
  - e.g., k is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

aardvark abacus abandoned abbot abduct above zygote zymurgy

(0, 0, 3, 1, 0, 7, . . . 1, 0)

Words as Vectors

- Represent each word **type** w by a point in k-dimensional space
  - e.g., k is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

= party

aardvark abacus abandoned abbot abduct above ... zygote zymurgy

(0, 0, 3, 1, 0, 7, . . . 1, 0)

From corpus:

Arlen Specter **abandoned** the Republican **party**.
There were lots of **abbots** and nuns dancing at that **party**.
The **party** **above** the art gallery was, **above** all, a laboratory for synthesizing **zygotes** and beer.

Words as Vectors

- Represent each word **type** w by a point in k-dimensional space
  - e.g., k is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

= party

aardvark abacus abandoned abbot abduct above . . . zygote zymurgy

(0, 0, 3, 1, 0, 7, . . . 1, 0)

count too high
(too influential)

From corpus:

Arlen Specter **abandoned** the Republican **party**.
There were lots of **abbots** and nuns dancing at that **party**.
The **party** **above** the art gallery was, **above** all, a laboratory for synthesizing **zygotes** and beer.

Words as Vectors

- Represent each word **type** w by a point in k-dimensional space
  = **party**
  - e.g., k is size of vocabulary
  - the 17^{th} coordinate of w represents **strength** of w's association with vocabulary word 17

aardvark abacus abandoned abbot abduct above ... zygote zymurgy

(0, 0, 3, 1, 0, 7, . . . 1, 0)

count too high (too influential)

count too low

From corpus:

Arlen Specter **abandoned** the Republican **party**.
There were lots of **abbots** and nuns dancing at that **party**.
The **party** **above** the art gallery was, **above** all, a laboratory for synthesizing **zygotes** and beer.

Words as Vectors

- Represent each word **type** w by a point in k-dimensional space
  = party
  - e.g., k is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

aardvark	abacus	abandoned	abbot	abduct	above		zygote	zymurgy
(0,	0,	3,	1,	0,	7,	. . .	1,	0)

Words as Vectors

- Represent each word **type** w by a point in k-dimensional space
  = party
  - e.g., k is size of vocabulary
  - the 17^{th} coordinate of w represents **strength** of w's association with vocabulary word 17

how might you measure this?

aardvark abacus abandoned abbot abduct above zygote zymurgy

(0, 0, 3, 1, 0, 7, . . . 1, 0)

Words as Vectors

- Represent each word **type** w by a point in k-dimensional space
  - e.g., k is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

= party

aardvark abacus abandoned abbot abduct above ... zygote zymurgy

(0, 0, 3, 1, 0, 7, . . . 1, 0)

how might you measure this?

- how often words appear next to each other

Words as Vectors

- Represent each word **type** w by a point in k-dimensional space
  = party
  - e.g., k is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

aardvark abacus abandoned abbot abduct above ... zygote zymurgy

(0, 0, 3, 1, 0, 7, . . . 1, 0)

how might you measure this?

- how often words appear next to each other
- how often words appear near each other

Words as Vectors

- Represent each word **type** w by a point in k-dimensional space
  = party
  - e.g., k is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

aardvark	abacus	abandoned	abbot	abduct	above		zygote	zymurgy
(0,	0,	3,	1,	0,	7,	. . .	1,	0)

how might you measure this?

- how often words appear next to each other
- how often words appear near each other
- how often words are syntactically linked

Words as Vectors

- Represent each word **type** w by = party a point in k-dimensional space
  - e.g., k is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

aardvark	abacus	abandoned	abbot	abduct	above		zygote	zymurgy
(0,	0,	3,	1,	0,	7,	. . .	1,	0)

how might you measure this?

- how often words appear next to each other
- how often words appear near each other
- how often words are syntactically linked
- should correct for commonness of word (e.g., "above")

Words as Vectors

- Represent each word **type** w by a point in k-dimensional space
  - e.g., k is size of vocabulary
  - the 17^{th} coordinate of w represents **strength** of w's association with vocabulary word 17

aardvark	abacus	abandoned	abbot	abduct	above		zygote	zymurgy
(0,	0,	3,	1,	0,	7,	. . .	1,	0)

Words as Vectors

- Represent each word **type** w by a point in k-dimensional space
  - e.g., k is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

aardvark	abacus	abandoned	abbot	abduct	above		zygote	zymurgy
(0,	0,	3,	1,	0,	7,	. . .	1,	0)

- Plot all word types in k-dimensional space

Words as Vectors

- Represent each word **type** w by a point in k-dimensional space
  - e.g., k is size of vocabulary
  - the 17th coordinate of w represents **strength** of w's association with vocabulary word 17

aardvark abacus abandoned abbot abduct above . . . zygote zymurgy

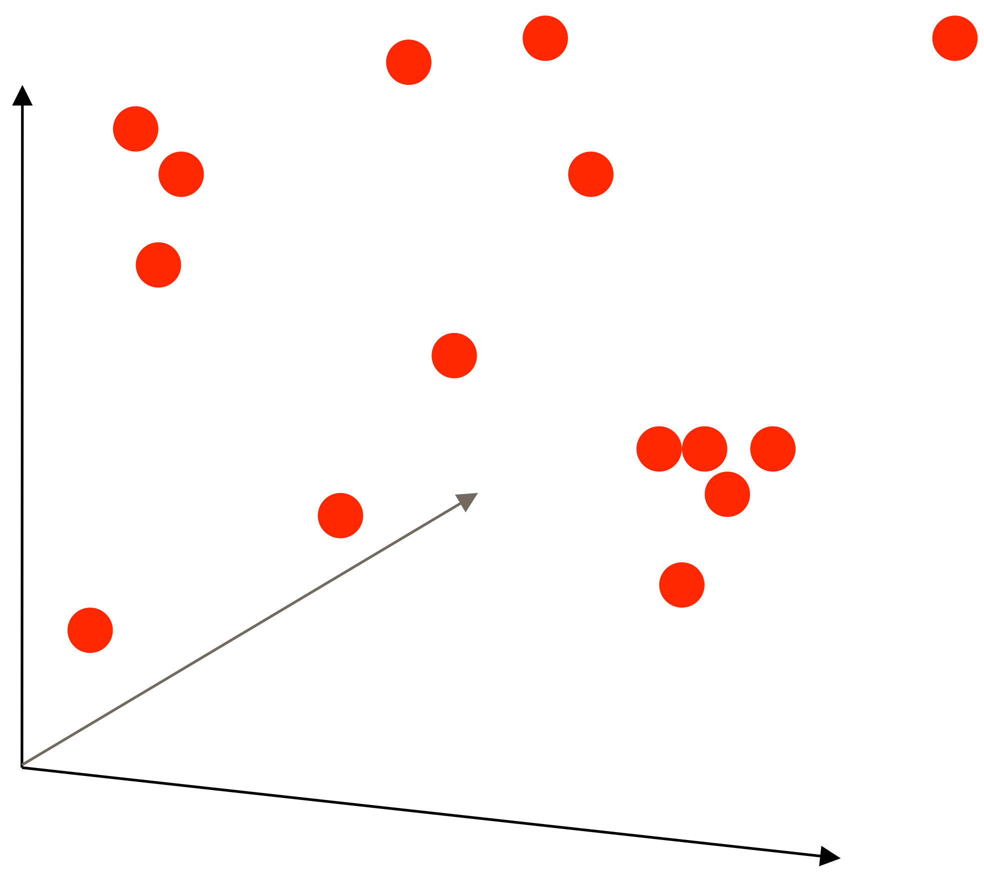
(0, 0, 3, 1, 0, 7, . . . 1, 0)

- Plot all word types in k-dimensional space
- Look for **clusters** of close-together types

Learning Classes by Clustering

- Plot all word types in k-dimensional space
- Look for **clusters** of close-together types

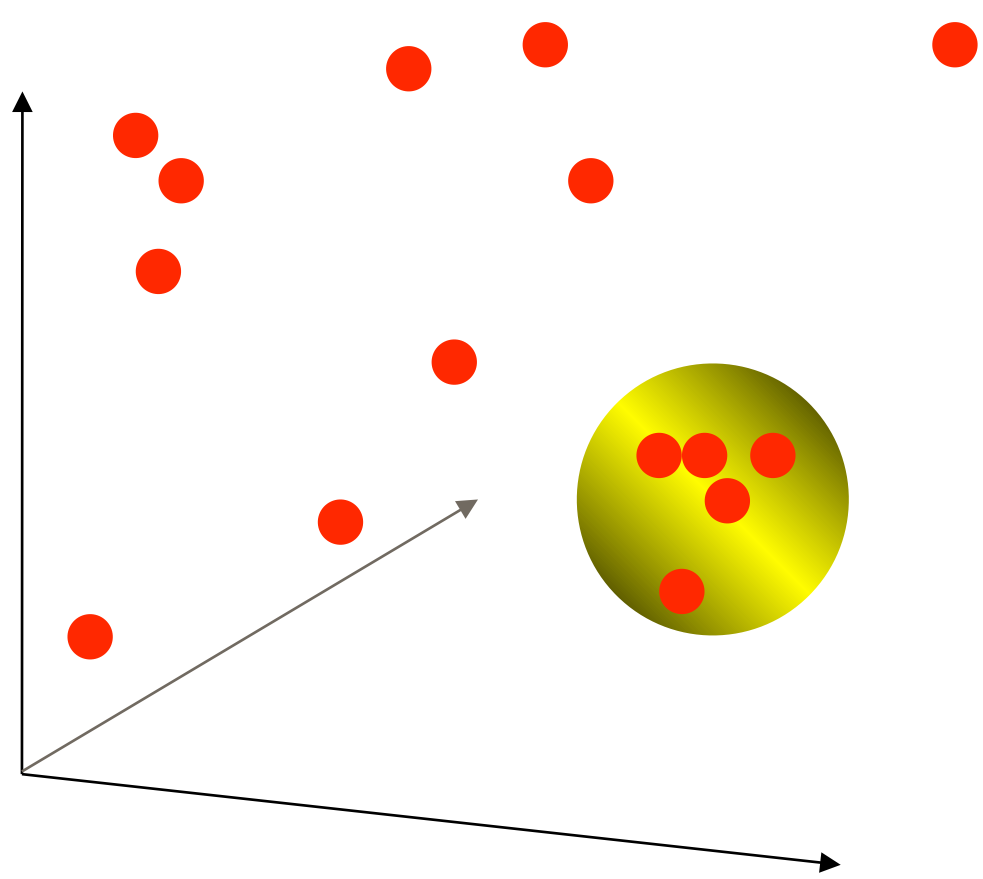
Plot in k dimensions (here k=3)



Learning Classes by Clustering

- Plot all word types in k-dimensional space
- Look for **clusters** of close-together types

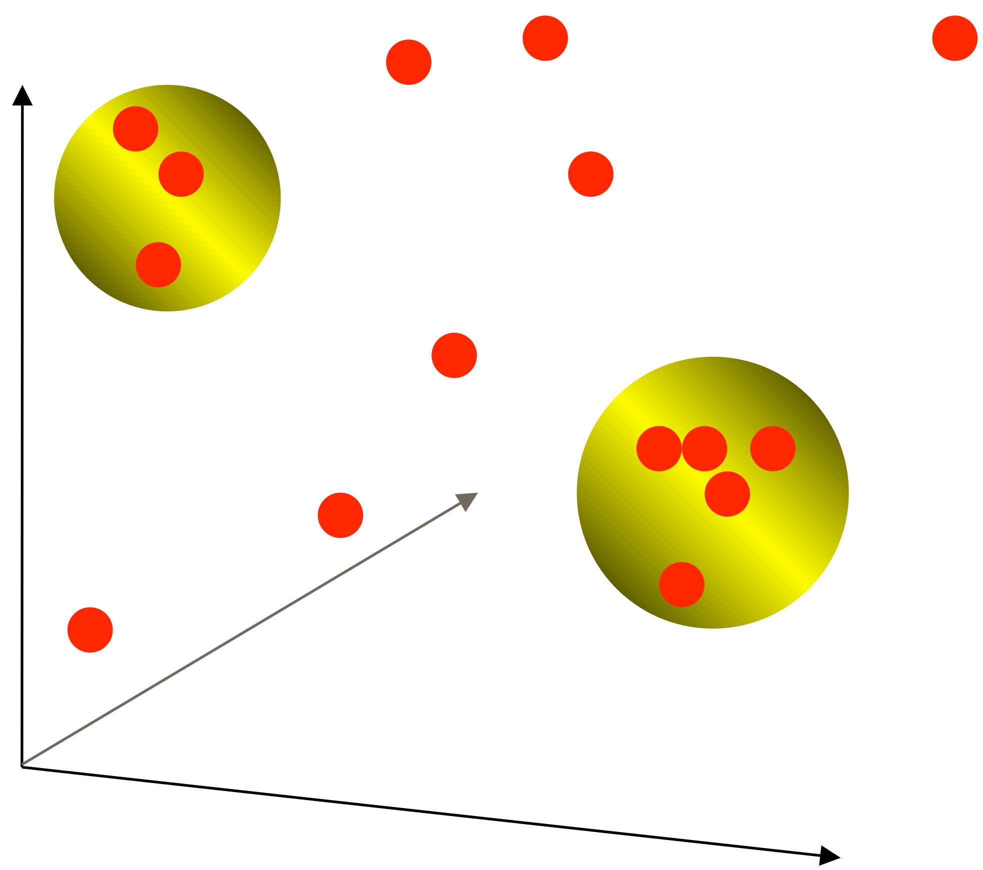
Plot in k dimensions (here k=3)



Learning Classes by Clustering

- Plot all word types in k-dimensional space
- Look for **clusters** of close-together types

Plot in k dimensions (here k=3)

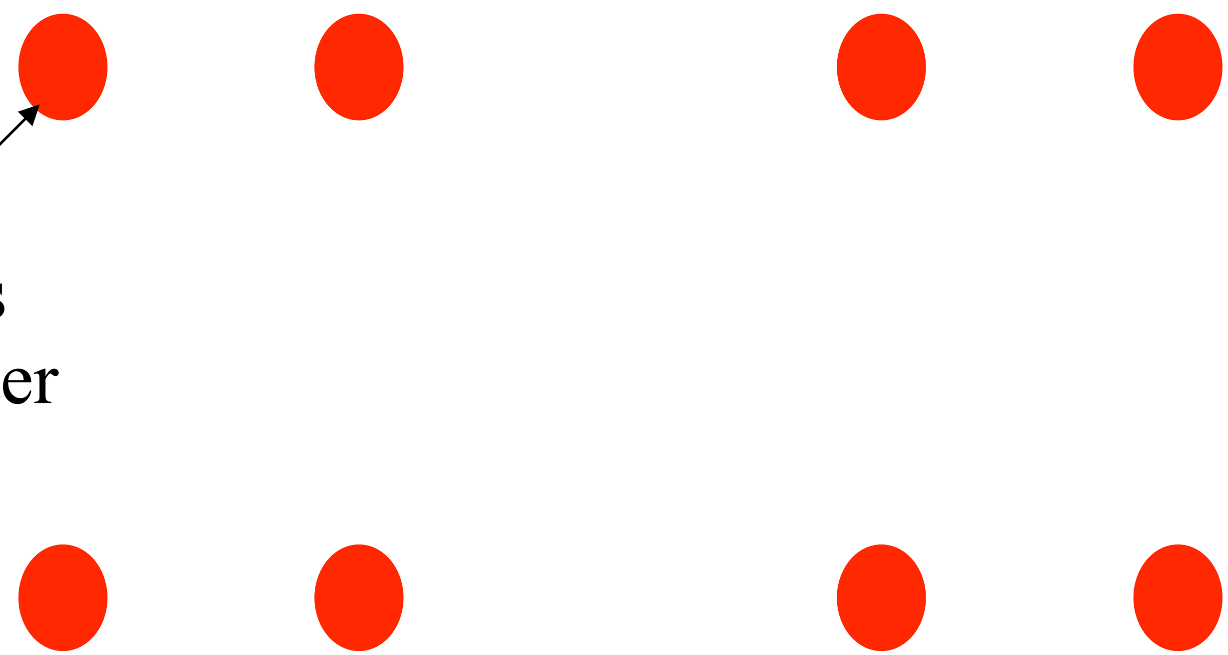


Bottom-Up Clustering

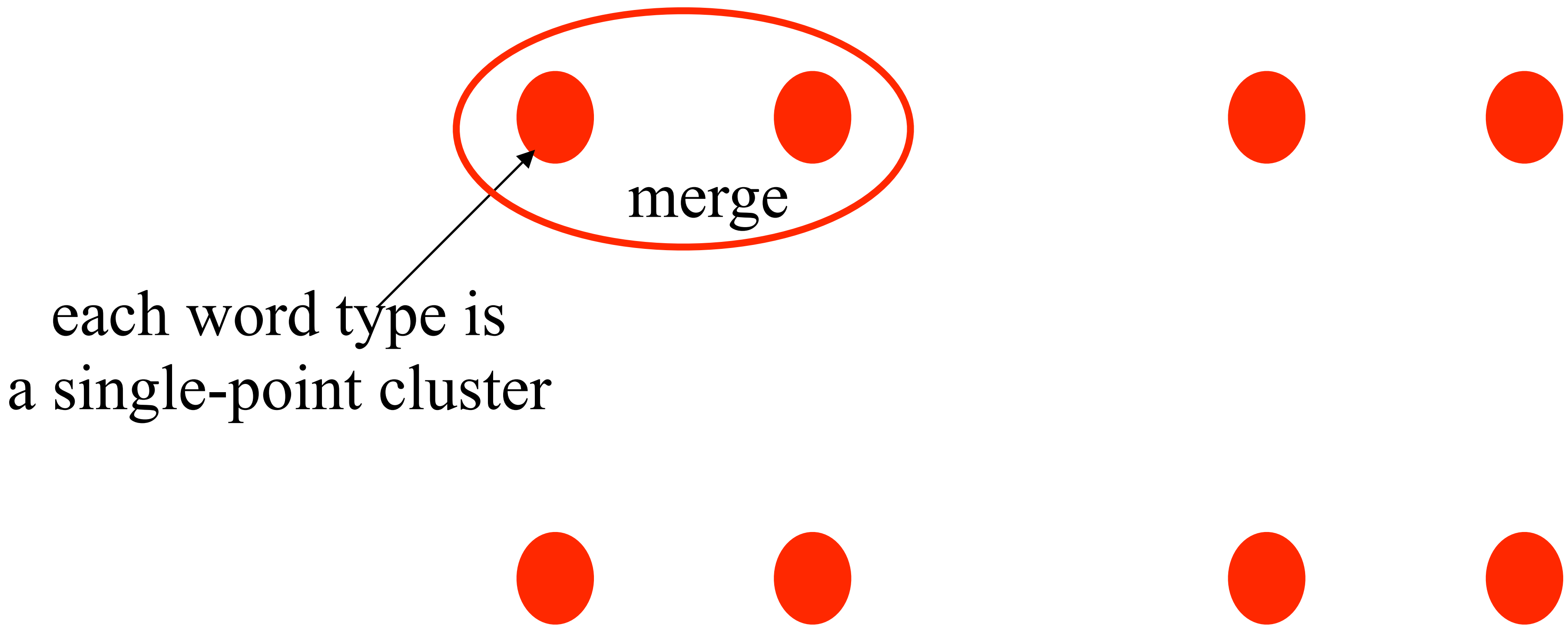
- Start with one cluster per point
- Repeatedly merge 2 closest clusters
  - Single-link: dist(A,B) = min dist(a,b) for $a \in A$, $b \in B$
  - Complete-link: dist(A,B) = max dist(a,b) for $a \in A$, $b \in B$

example from Manning & Schütze

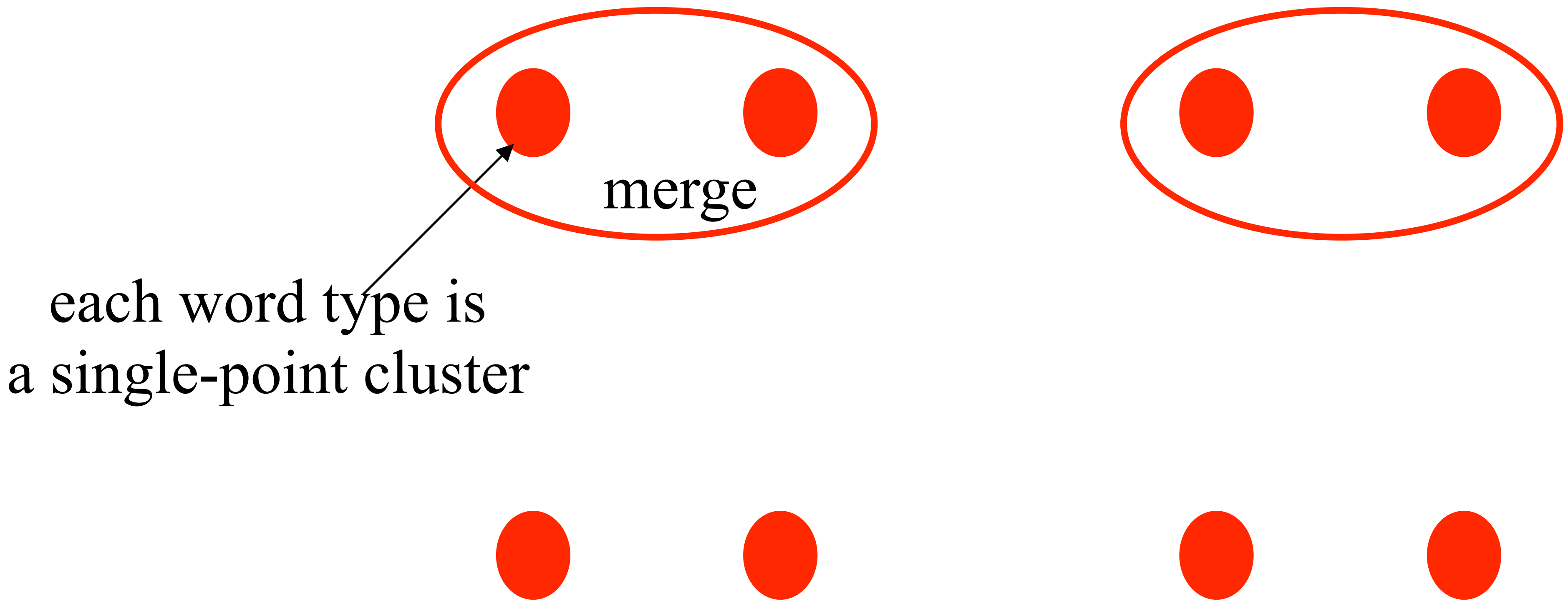
Bottom-Up Clustering – Single-Link



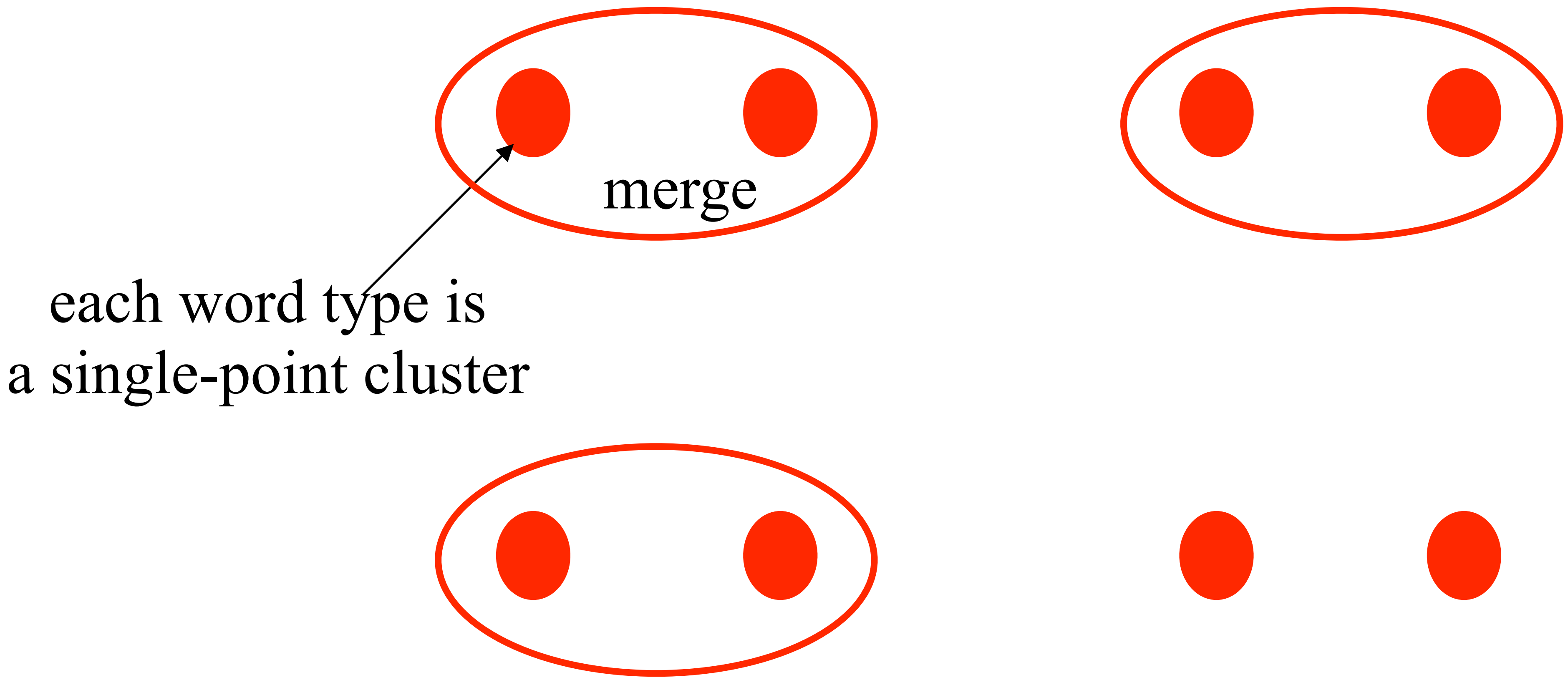
Bottom-Up Clustering – Single-Link



Bottom-Up Clustering – Single-Link

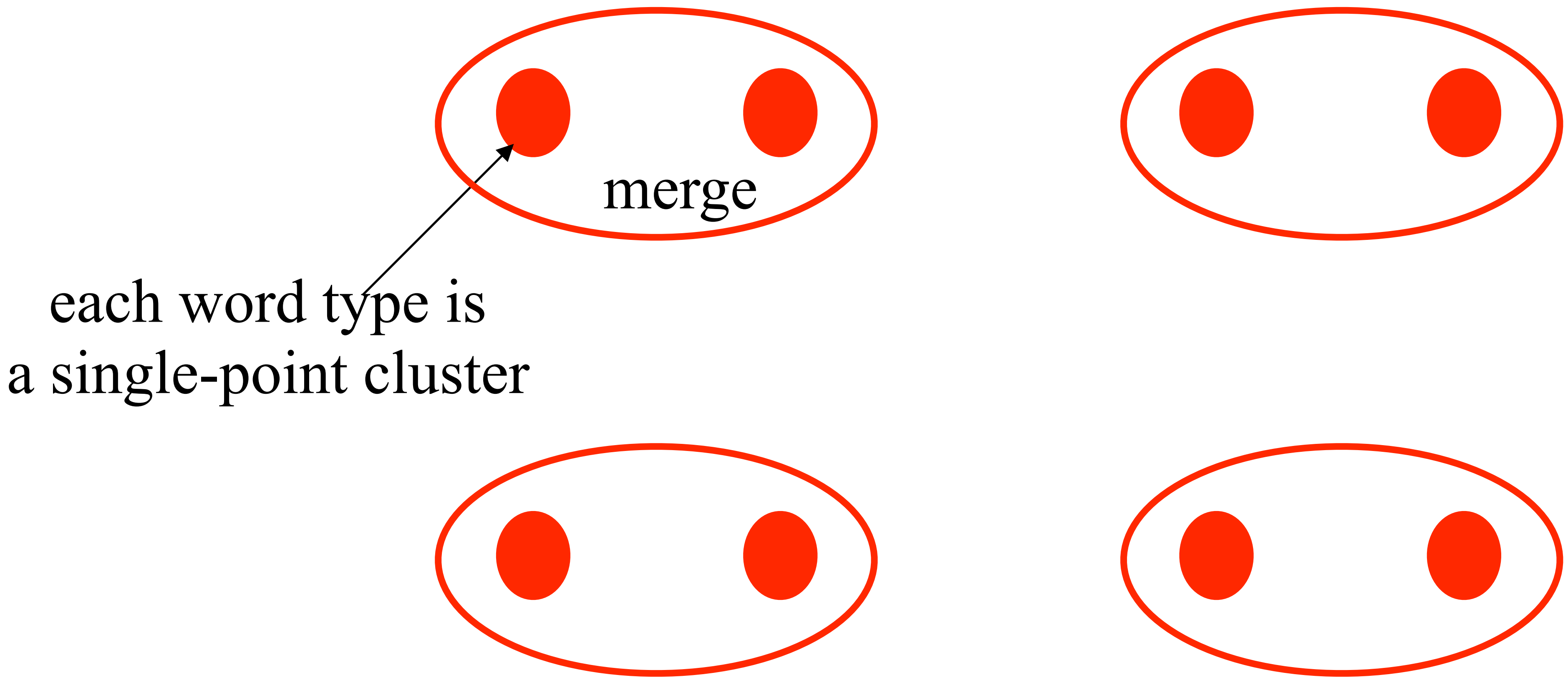


Bottom-Up Clustering – Single-Link

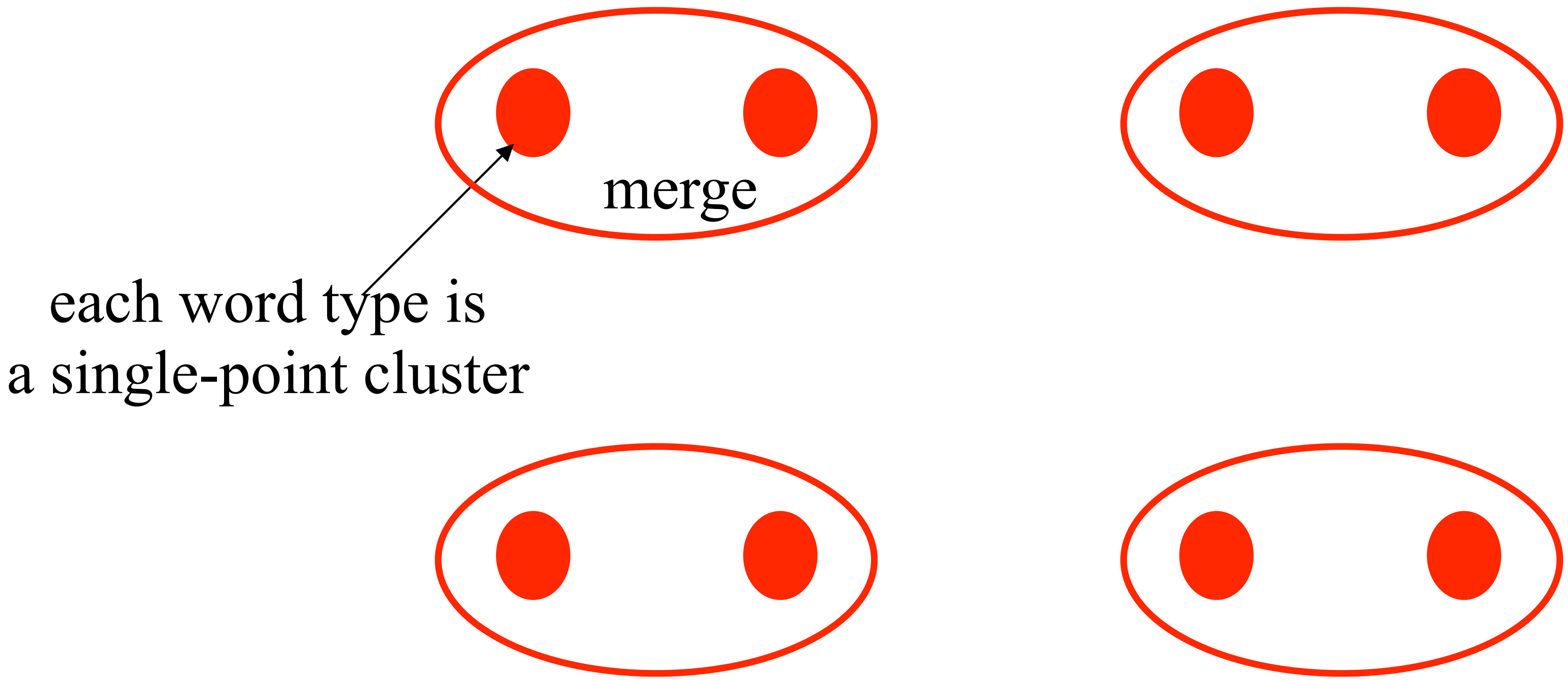


example from Manning & Schütze

Bottom-Up Clustering – Single-Link



Bottom-Up Clustering – Single-Link

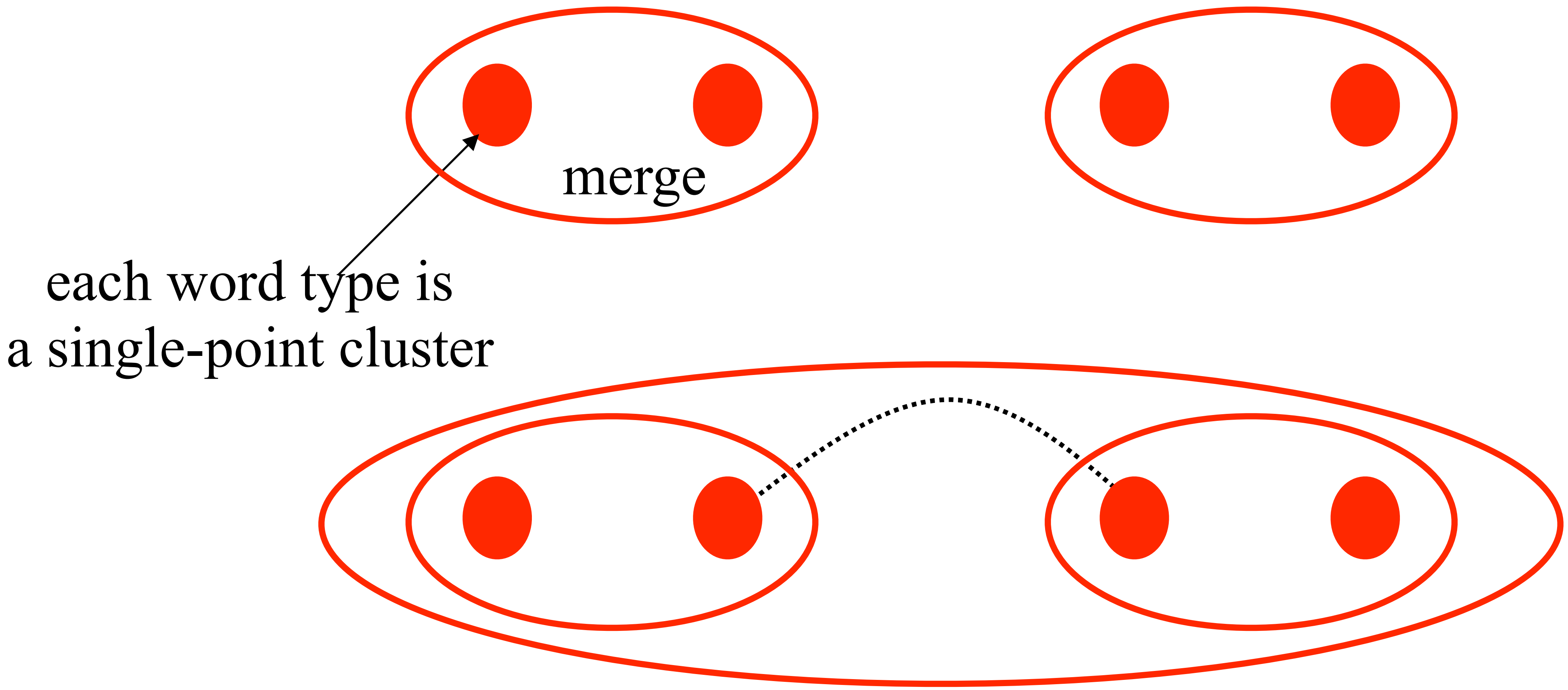


Again, merge closest pair of clusters:

Single-link: clusters are close if any of their points are

dist(A,B) = min dist(a,b) for $a \in A$, $b \in B$

Bottom-Up Clustering – Single-Link

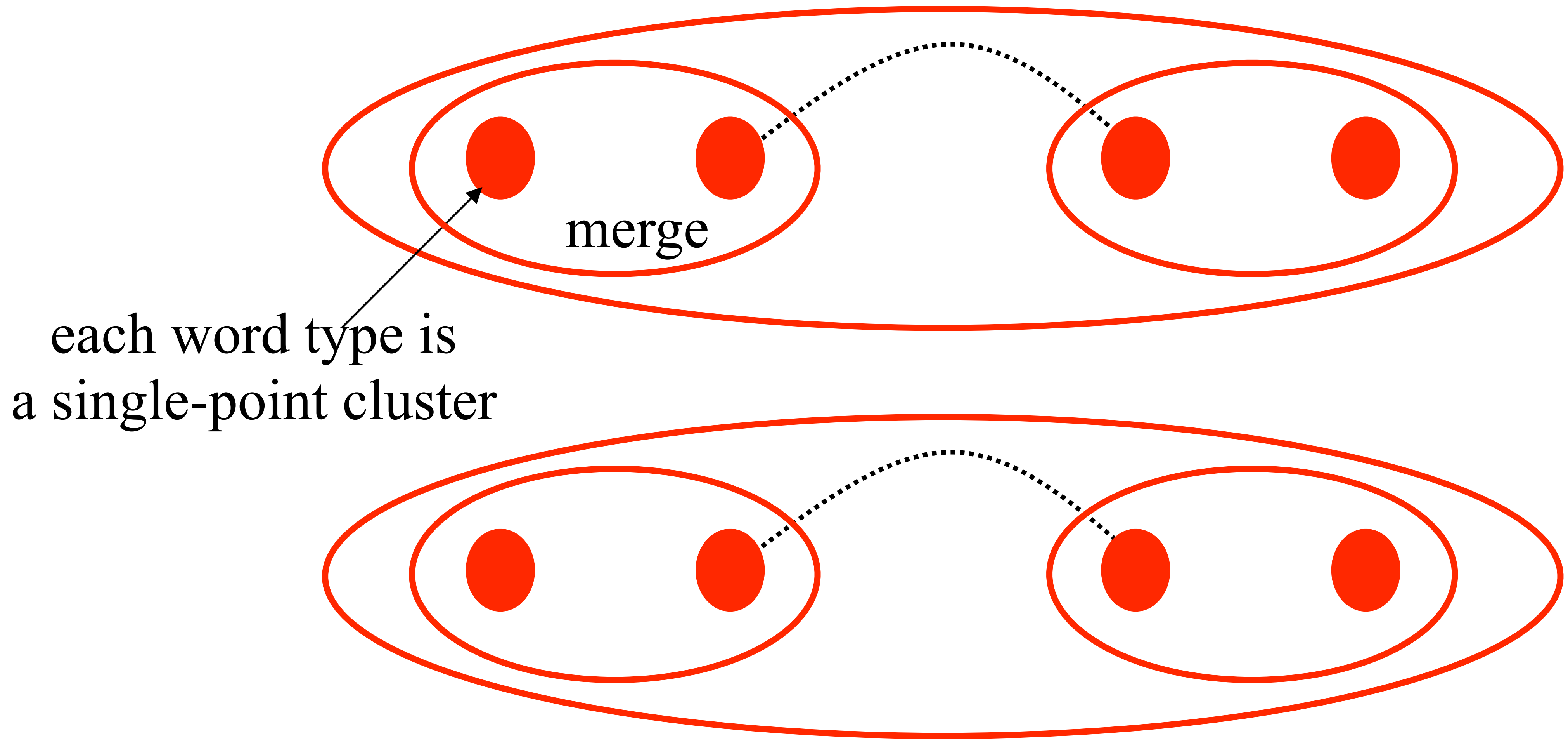


Again, merge closest pair of clusters:

Single-link: clusters are close if any of their points are

dist(A,B) = min dist(a,b) for $a \in A$, $b \in B$

Bottom-Up Clustering – Single-Link



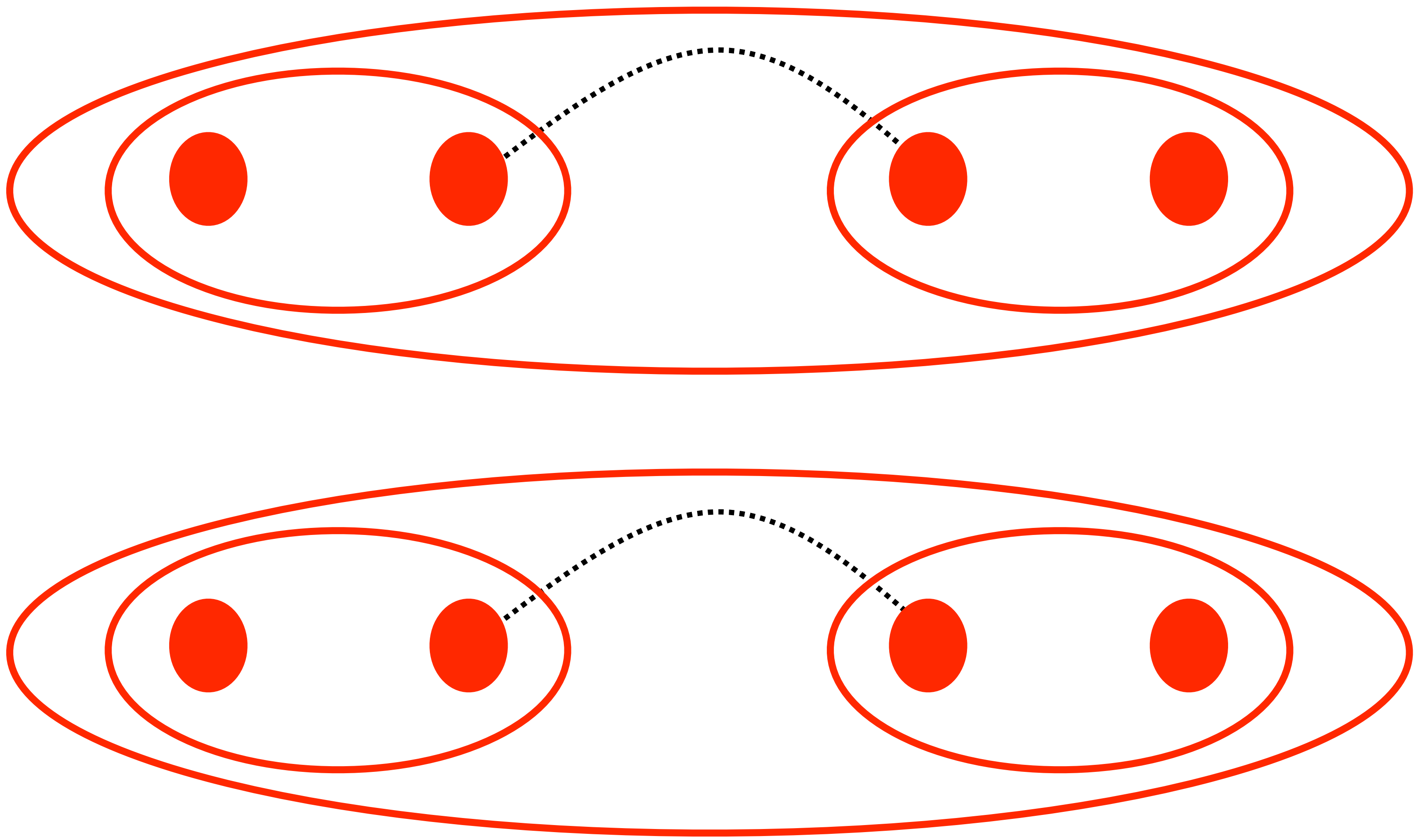
Again, merge closest pair of clusters:

Single-link: clusters are close if any of their points are

dist(A,B) = min dist(a,b) for a∈A, b∈B

example from Manning & Schütze

Bottom-Up Clustering – Single-Link



Again, merge closest pair of clusters:

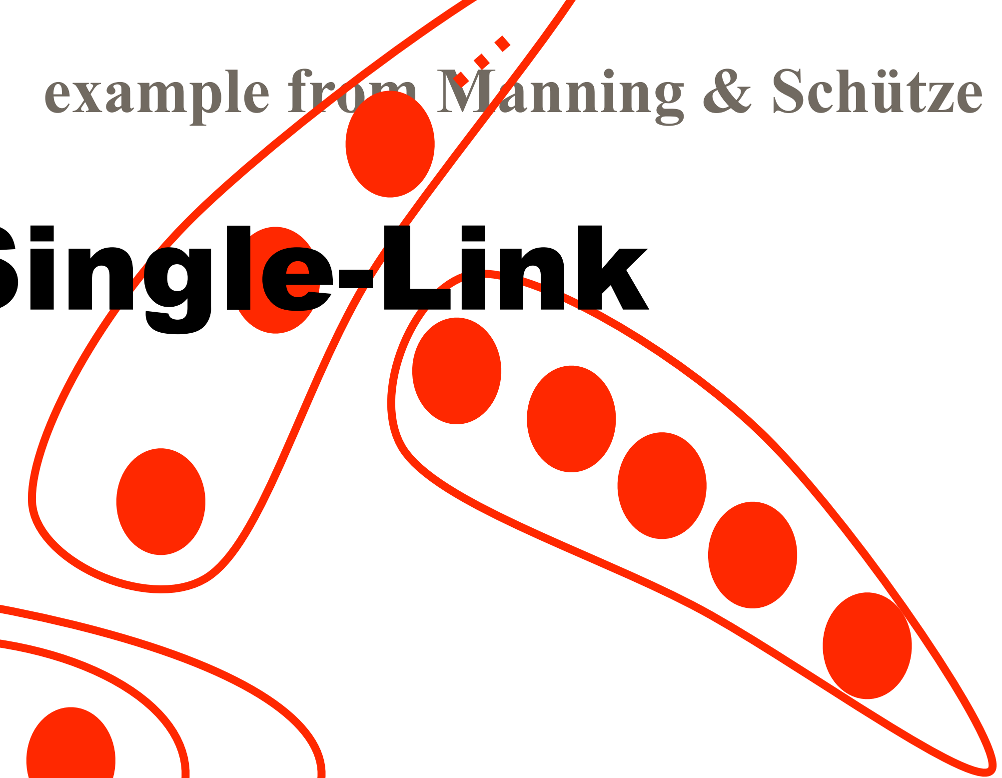
Single-link: clusters are close if any of their points are

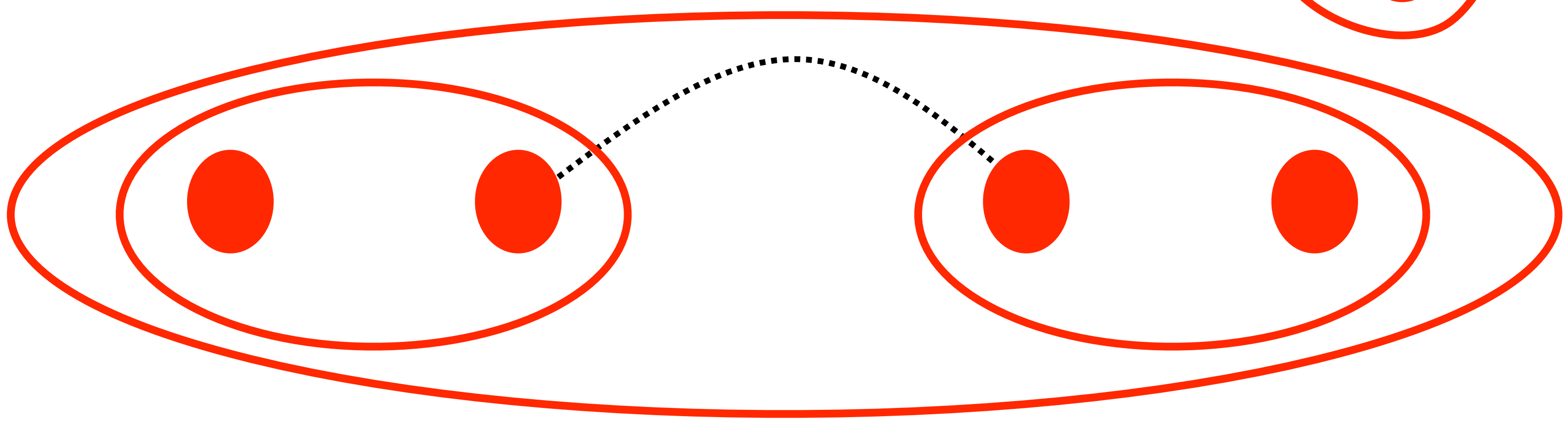
dist(A,B) = min dist(a,b) for $a \in A$, $b \in B$

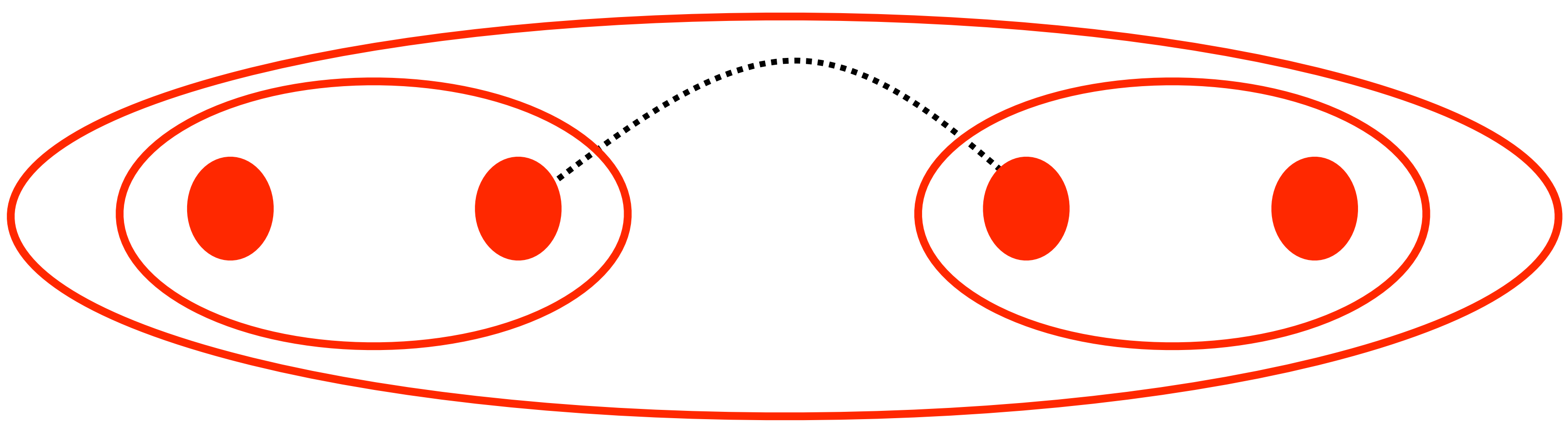
Fast, but tend to get long, stringy, meandering clusters

example from Manning & Schütze

Bottom-Up Clustering – Single-Link







Again, merge closest pair of clusters:

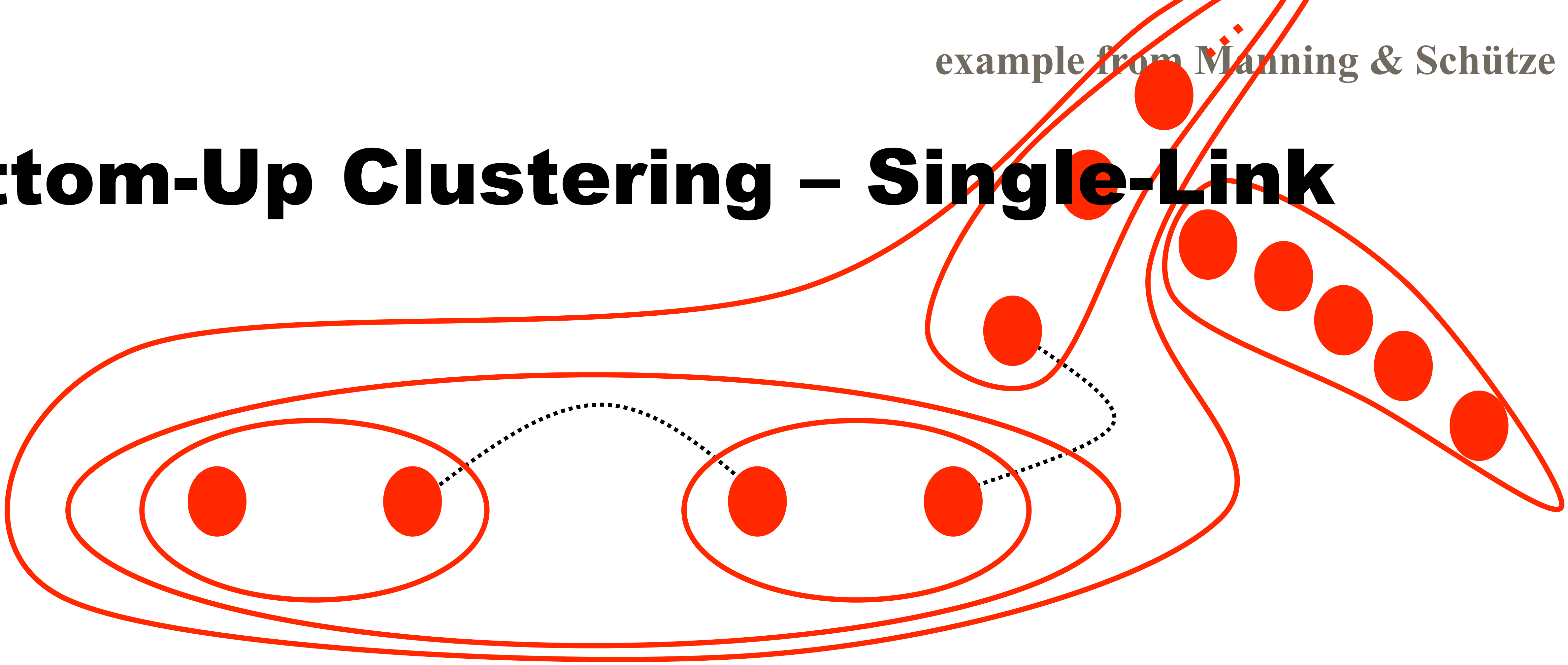
Single-link: clusters are close if any of their points are

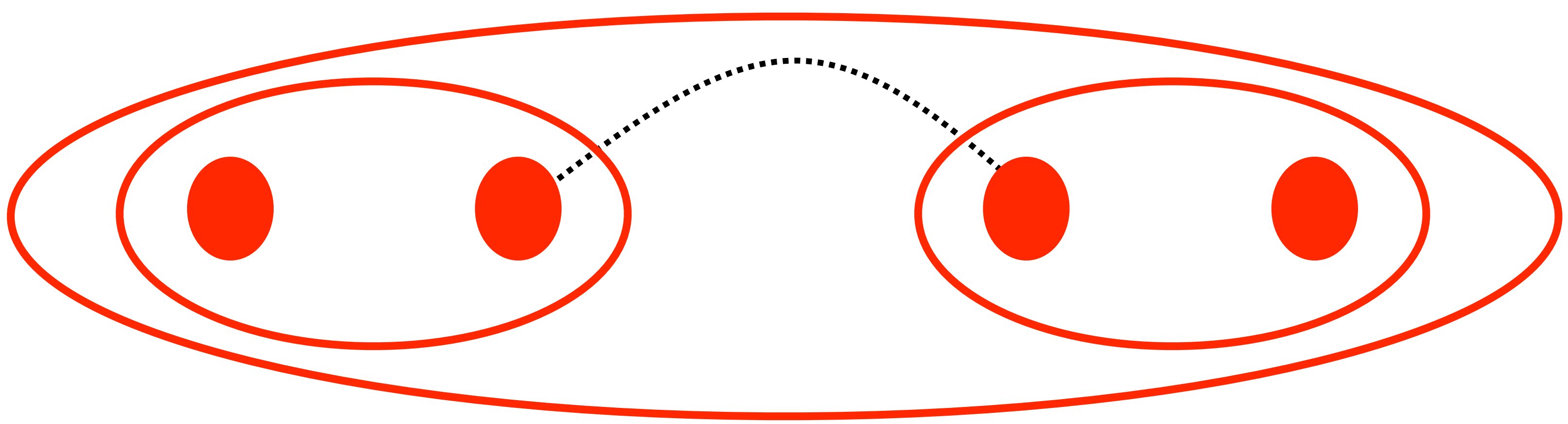
dist(A,B) = min dist(a,b) for $a \in A$, $b \in B$

Fast, but tend to get long, stringy, meandering clusters

example from Manning & Schütze

Bottom-Up Clustering – Single-Link





Again, merge closest pair of clusters:

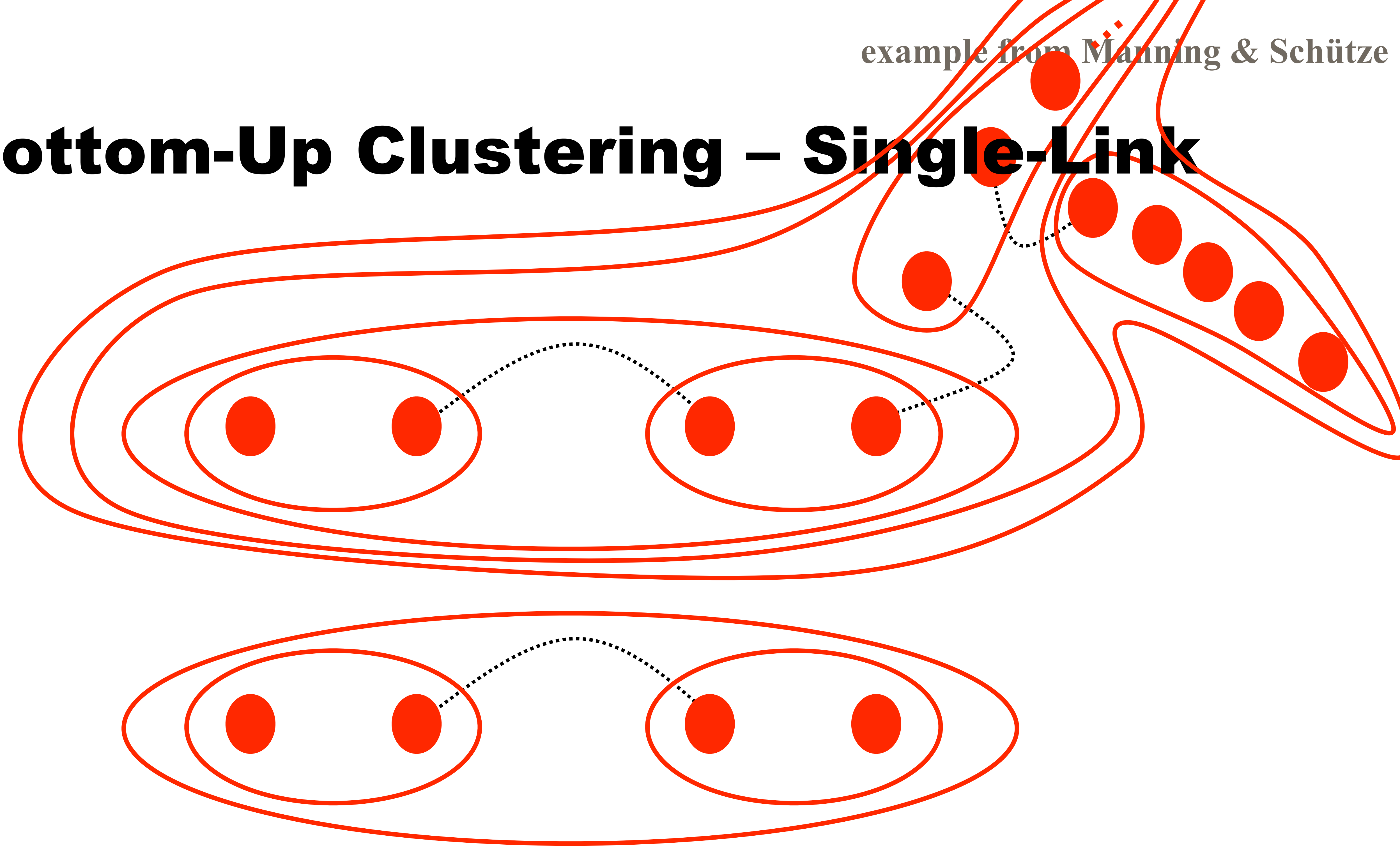
Single-link: clusters are close if any of their points are

dist(A,B) = min dist(a,b) for a∈A, b∈B

Fast, but tend to get long, stringy, meandering clusters

example from Manning & Schütze

Bottom-Up Clustering – Single-Link



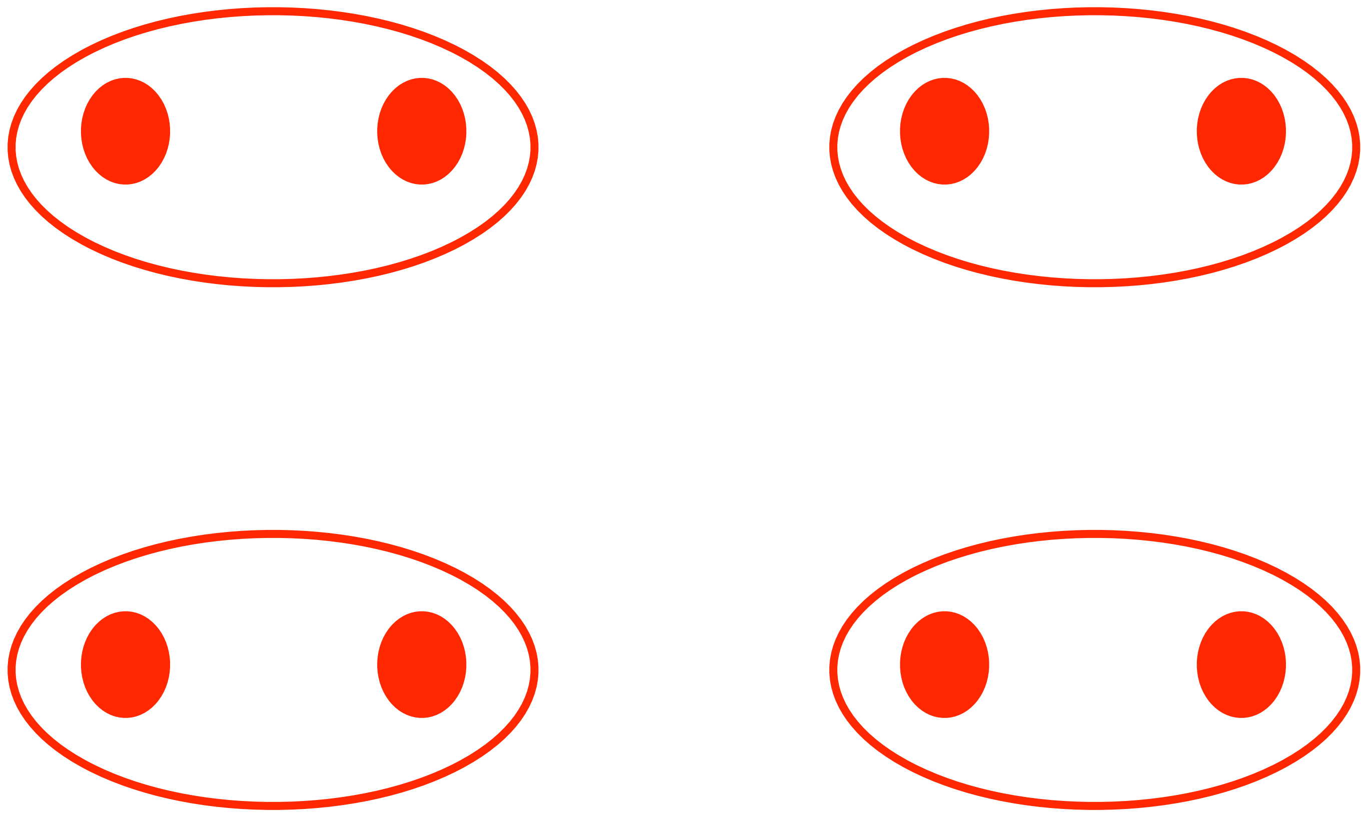
Again, merge closest pair of clusters:

Single-link: clusters are close if any of their points are

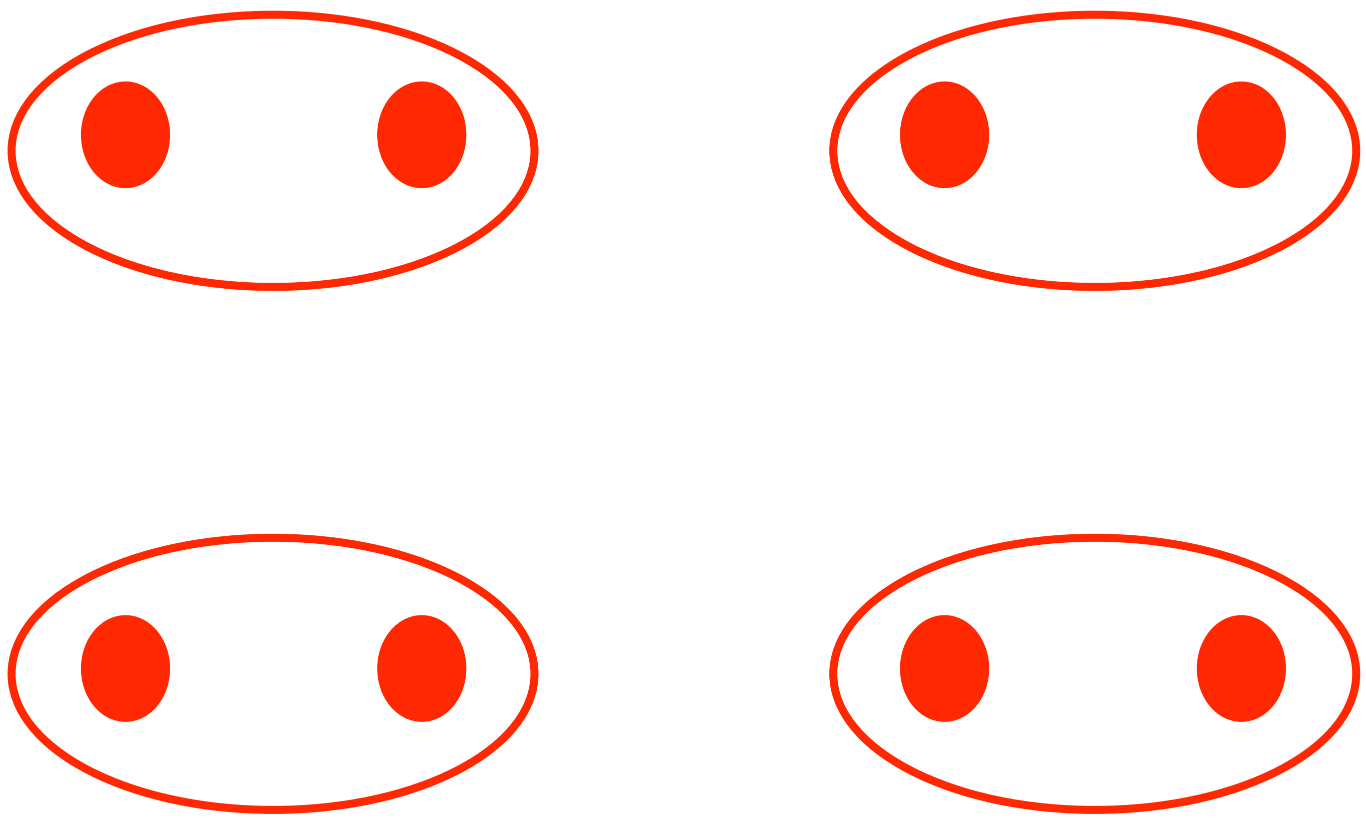
dist(A,B) = min dist(a,b) for a∈A, b∈B

Fast, but tend to get long, stringy, meandering clusters

Bottom-Up Clustering – Complete-Link



Bottom-Up Clustering – Complete-Link

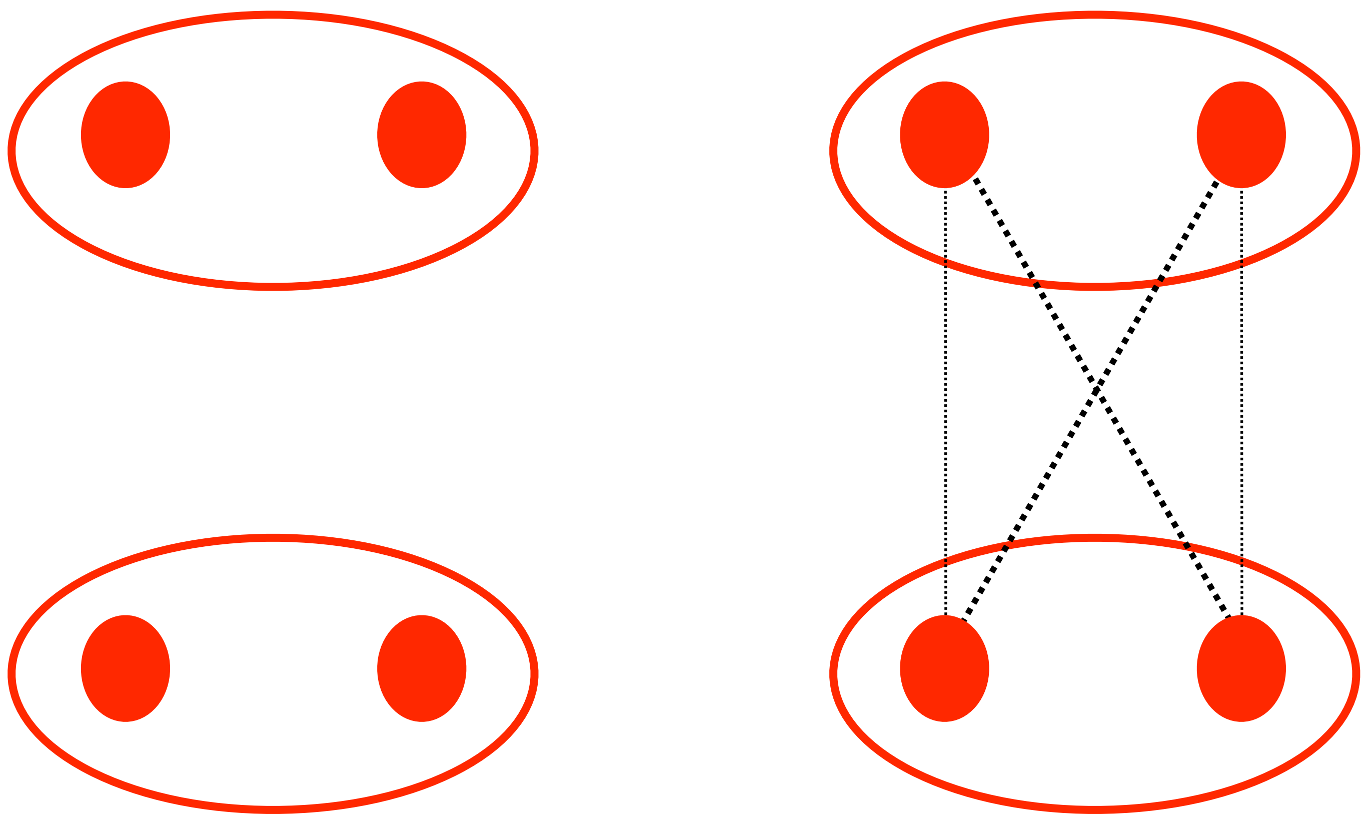


Again, merge closest pair of clusters:

Complete-link: clusters are close only if all of their points are

$$\text{dist}(A,B) = \max \text{dist}(a,b) \text{ for } a \in A,\ b \in B$$

Bottom-Up Clustering – Complete-Link

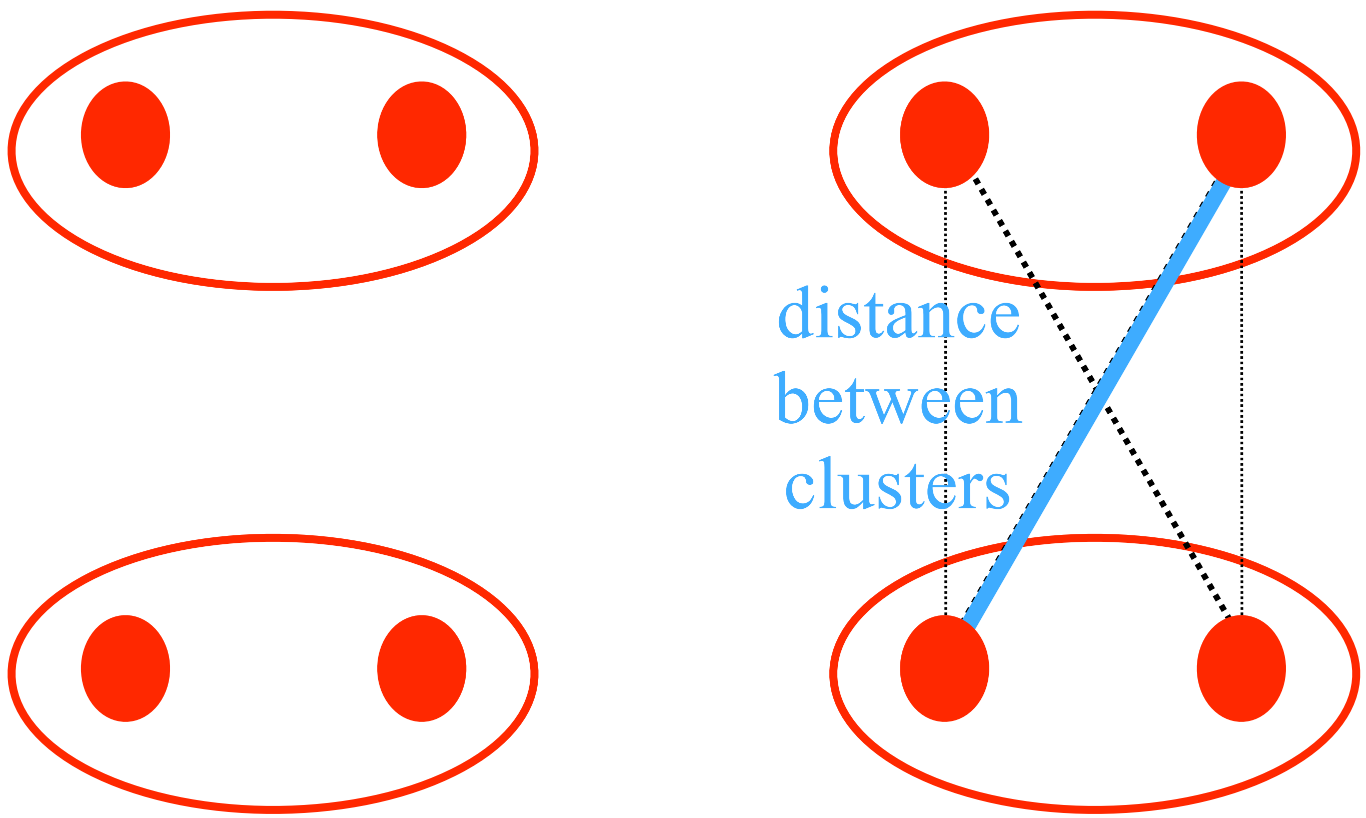


Again, merge closest pair of clusters:

Complete-link: clusters are close only if all of their points are

dist(A,B) = max dist(a,b) for $a \in A$, $b \in B$

Bottom-Up Clustering – Complete-Link



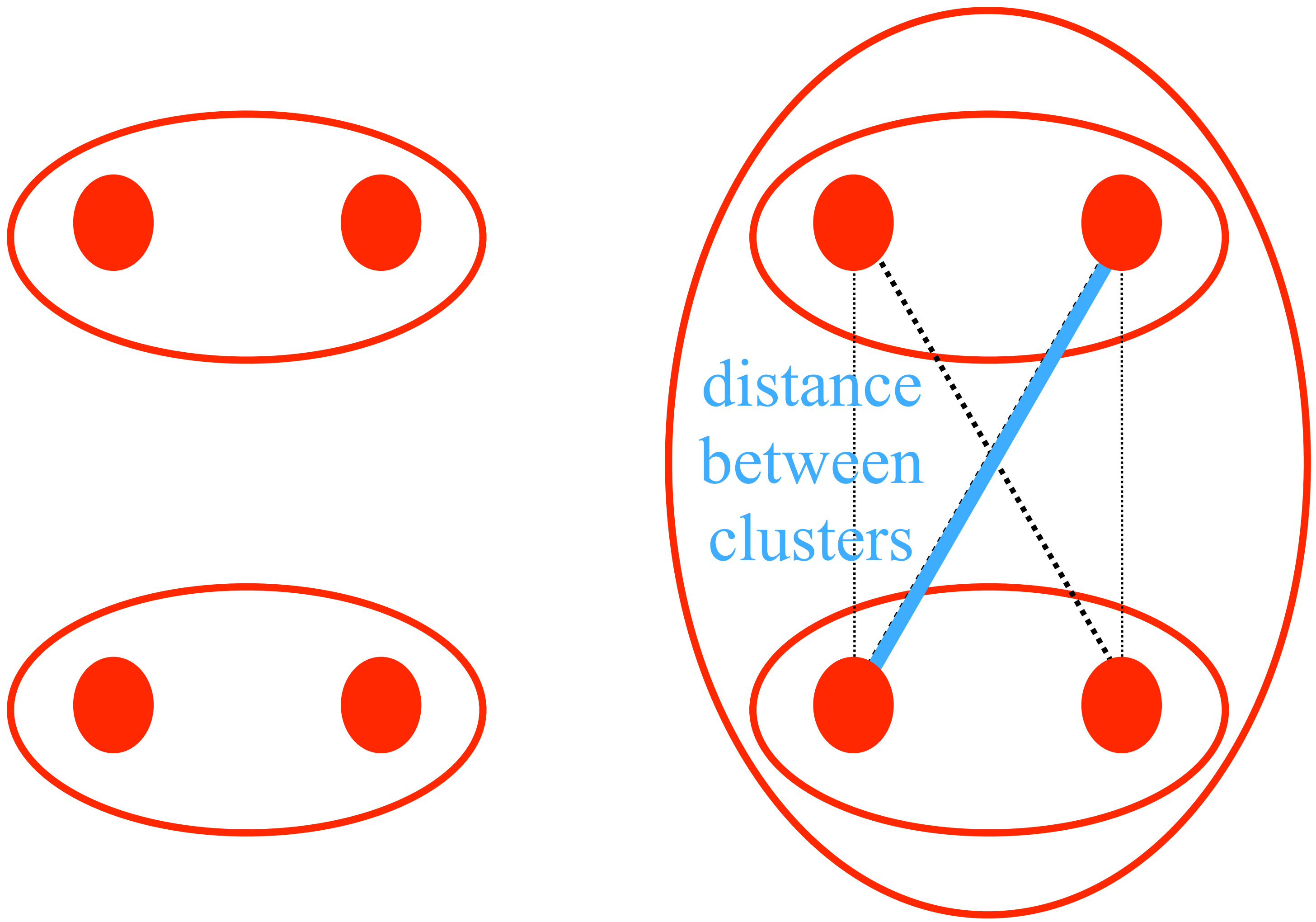
Again, merge closest pair of clusters:

Complete-link: clusters are close only if all of their points are

dist(A,B) = max dist(a,b) for a∈A, b∈B

example from Manning & Schütze

Bottom-Up Clustering – Complete-Link

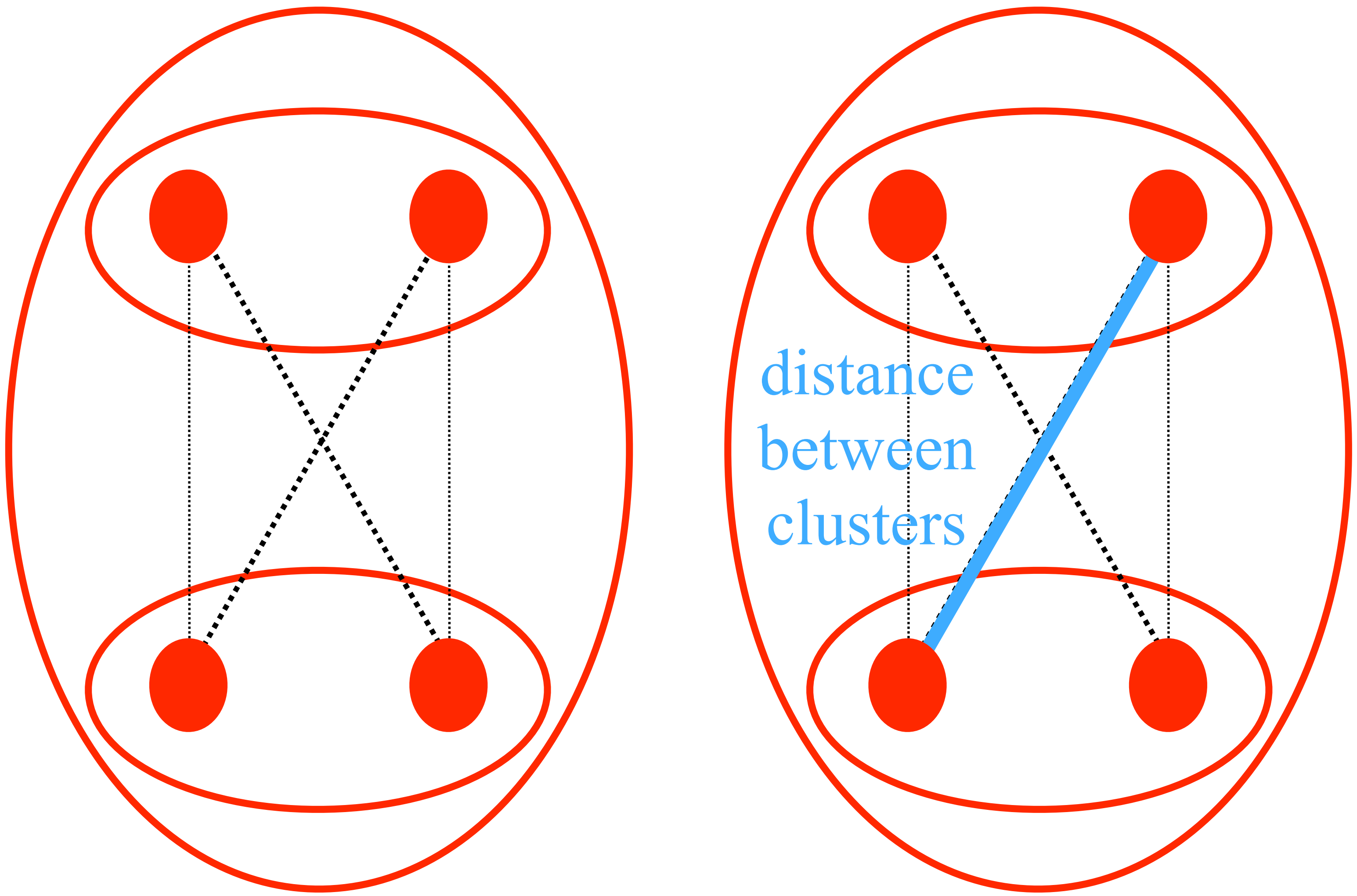


Again, merge closest pair of clusters:

Complete-link: clusters are close only if all of their points are

dist(A,B) = max dist(a,b) for $a \in A$, $b \in B$

Bottom-Up Clustering – Complete-Link

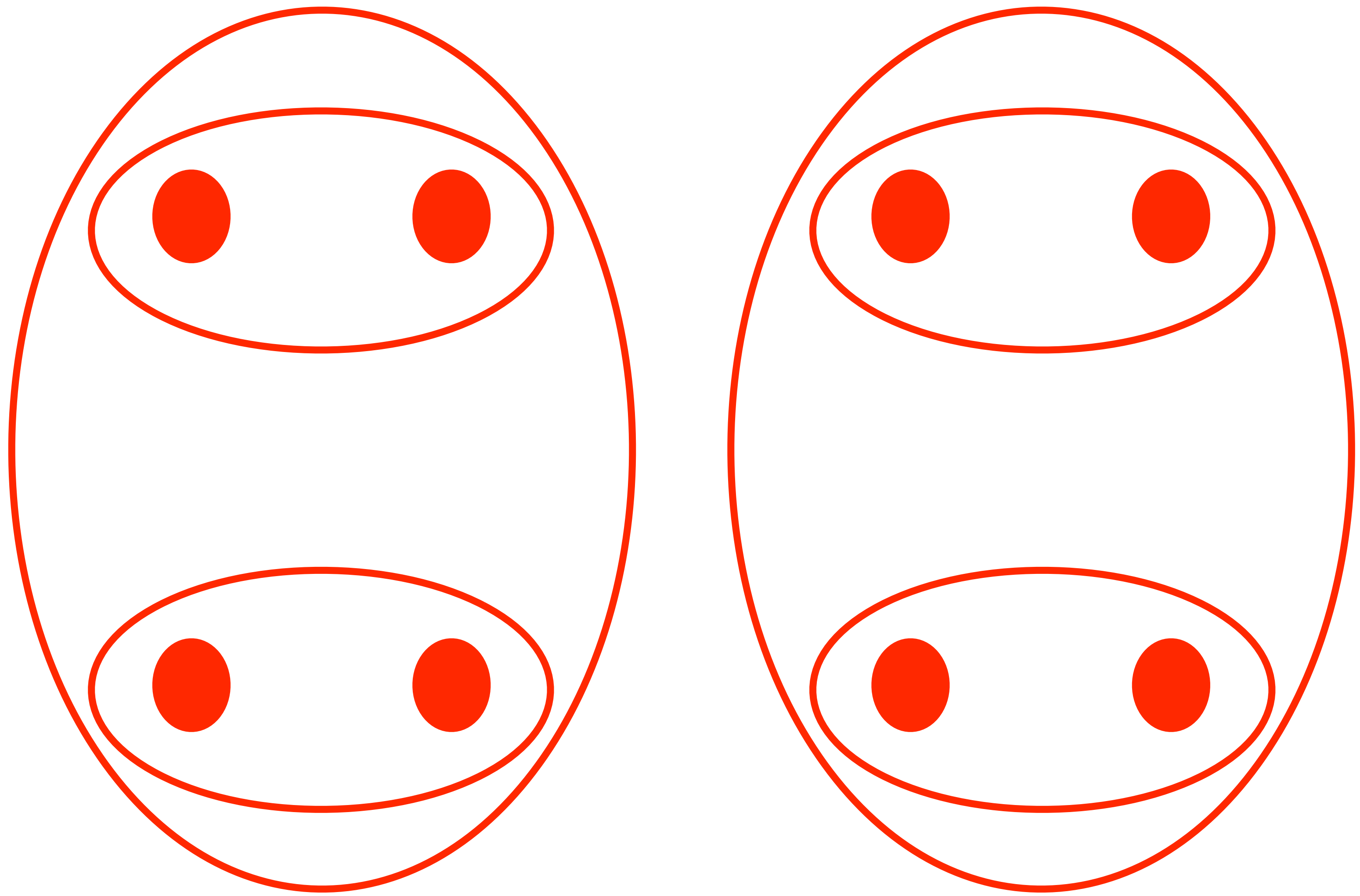


Again, merge closest pair of clusters:

Complete-link: clusters are close only if all of their points are

dist(A,B) = max dist(a,b) for $a \in A$, $b \in B$

Bottom-Up Clustering – Complete-Link



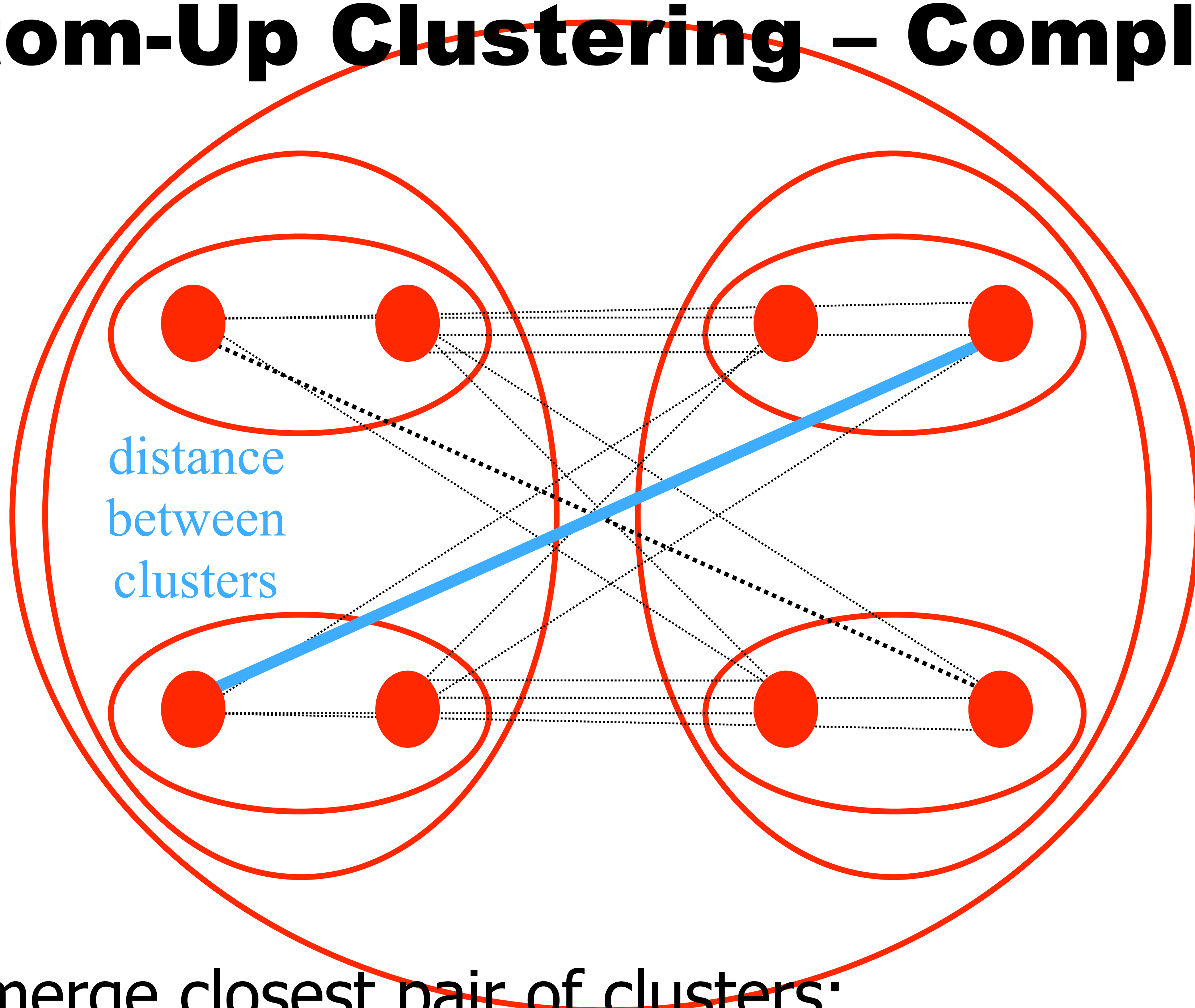
Again, merge closest pair of clusters:

Complete-link: clusters are close only if all of their points are

$\text{dist}(A,B) = \max \text{dist}(a,b)$ for $a \in A$, $b \in B$

Slow to find closest pair – need quadratically many distances

Bottom-Up Clustering – Complete-Link



Again, merge closest pair of clusters:

Complete-link: clusters are close only if all of their points are

dist(A,B) = max dist(a,b) for $a \in A$, $b \in B$

Slow to find closest pair – need quadratically many distances

Bottom-Up Clustering

- Start with one cluster per point
- Repeatedly merge 2 closest clusters
  - Single-link: dist(A,B) = min dist(a,b) for $a \in A$, $b \in B$
  - Complete-link: dist(A,B) = max dist(a,b) for $a \in A$, $b \in B$
    - too slow to update cluster distances after each merge; but $\exists$ alternatives!

Bottom-Up Clustering

- Start with one cluster per point
- Repeatedly merge 2 closest clusters
  - Single-link: dist(A,B) = min dist(a,b) for $a \in A$, $b \in B$
  - Complete-link: dist(A,B) = max dist(a,b) for $a \in A$, $b \in B$
    - too slow to update cluster distances after each merge; but $\exists$ alternatives!
  - Average-link: dist(A,B) = mean dist(a,b) for $a \in A$, $b \in B$
  - Centroid-link: dist(A,B) = dist(mean(A),mean(B))

Bottom-Up Clustering

- Start with one cluster per point
- Repeatedly merge 2 closest clusters
  - Single-link: dist(A,B) = min dist(a,b) for $a \in A$, $b \in B$
  - Complete-link: dist(A,B) = max dist(a,b) for $a \in A$, $b \in B$
    - too slow to update cluster distances after each merge; but $\exists$ alternatives!
  - Average-link: dist(A,B) = mean dist(a,b) for $a \in A$, $b \in B$
  - Centroid-link: dist(A,B) = dist(mean(A),mean(B))
- Stop when clusters are "big enough"
  - e.g., provide adequate support for backoff (on a development corpus)

Bottom-Up Clustering

- Start with one cluster per point
- Repeatedly merge 2 closest clusters
  - Single-link: dist(A,B) = min dist(a,b) for $a \in A$, $b \in B$
  - Complete-link: dist(A,B) = max dist(a,b) for $a \in A$, $b \in B$
    - too slow to update cluster distances after each merge; but $\exists$ alternatives!
  - Average-link: dist(A,B) = mean dist(a,b) for $a \in A$, $b \in B$
  - Centroid-link: dist(A,B) = dist(mean(A),mean(B))
- Stop when clusters are "big enough"
  - e.g., provide adequate support for backoff (on a development corpus)
- Some flexibility in defining dist(a,b)
  - Might not be Euclidean distance; e.g., use vector angle

EM Clustering (for k clusters)

EM Clustering (for k clusters)

- EM algorithm
  - Viterbi version – called "k-means clustering"
  - Full EM version – called "Gaussian mixtures"

EM Clustering (for k clusters)

- EM algorithm
  - Viterbi version – called “k-means clustering”
  - Full EM version – called “Gaussian mixtures”

EM Clustering (for k clusters)

- EM algorithm
  - Viterbi version – called “k-means clustering”
  - Full EM version – called “Gaussian mixtures”

- **Expectation step:** Use current parameters (and observations) to reconstruct hidden structure

EM Clustering (for k clusters)

- EM algorithm
  - Viterbi version – called "k-means clustering"
  - Full EM version – called "Gaussian mixtures"

- **Expectation step:** Use current parameters (and observations) to reconstruct hidden structure
- **Maximization step:** Use that hidden structure (and observations) to reestimate parameters

EM Clustering (for k clusters)

- EM algorithm
  - Viterbi version – called “k-means clustering”
  - Full EM version – called “Gaussian mixtures”

- **Expectation step:** Use current parameters (and observations) to reconstruct hidden structure
- **Maximization step:** Use that hidden structure (and observations) to reestimate parameters

EM Clustering (for k clusters)

- EM algorithm
  - Viterbi version – called "k-means clustering"
  - Full EM version – called "Gaussian mixtures"

- **Expectation step:** Use current parameters (and observations) to reconstruct hidden structure
- **Maximization step:** Use that hidden structure (and observations) to reestimate parameters

- Parameters: k points representing cluster centers

EM Clustering (for k clusters)

- EM algorithm
  - Viterbi version – called “k-means clustering”
  - Full EM version – called “Gaussian mixtures”

- **Expectation step:** Use current parameters (and observations) to reconstruct hidden structure
- **Maximization step:** Use that hidden structure (and observations) to reestimate parameters

- Parameters: k points representing cluster centers
- Hidden structure: for each data point (word type), which center generated it?