Classification
&
Clustering

IS4200/CS6200
Information Retrieval

Spam

Spam

Spam

```
To:  ...
From:  ...
Subject: non profit debt
X-Spam-Checked: This message probably not SPAM
X-Spam-Score: 3.853, Required: 5
X-Spam-Level: *** (3.853)
X-Spam-Tests: BAYES_50,DATE_IN_FUTURE_06_12,URIBL_BLACK
X-Spam-Report-rig: ---- Start SpamAssassin (v2.6xx-cscf) results
        2.0 URIBL_BLACK          Contains an URL listed in the URIBL blacklist
                                 [URIs: bad-debtyh.net.cn]
        1.9 DATE_IN_FUTURE_06_12   Date: is 6 to 12 hours after Received: date
        0.0 BAYES_50             BODY: Bayesian spam probability is 40 to 60%
                                 [score: 0.4857]

Say good bye to debt
Acceptable Unsecured Debt includes All Major Credit Cards, No-collateral
Bank Loans, Personal Loans,
Medical Bills etc.
http://www.bad-debtyh.net.cn
```

Spam

Website:

BETTING NFL FOOTBALL PRO FOOTBALL SPORTSBOOKS NFL FOOTBALL LINE ONLINE NFL SPORTSBOOKS NFL

Players Super Book

When It Comes To Secure NFL Betting And Finding The Best Football Lines Players Super Book Is The Best Option! Sign Up And Ask For 30 % In Bonuses.

MVP Sportsbook

Football Betting Has Never been so easy and secure! MVP Sportsbook has all the NFL odds you are looking for. Sign Up Now and ask for up to

30 % in Cash bonuses.

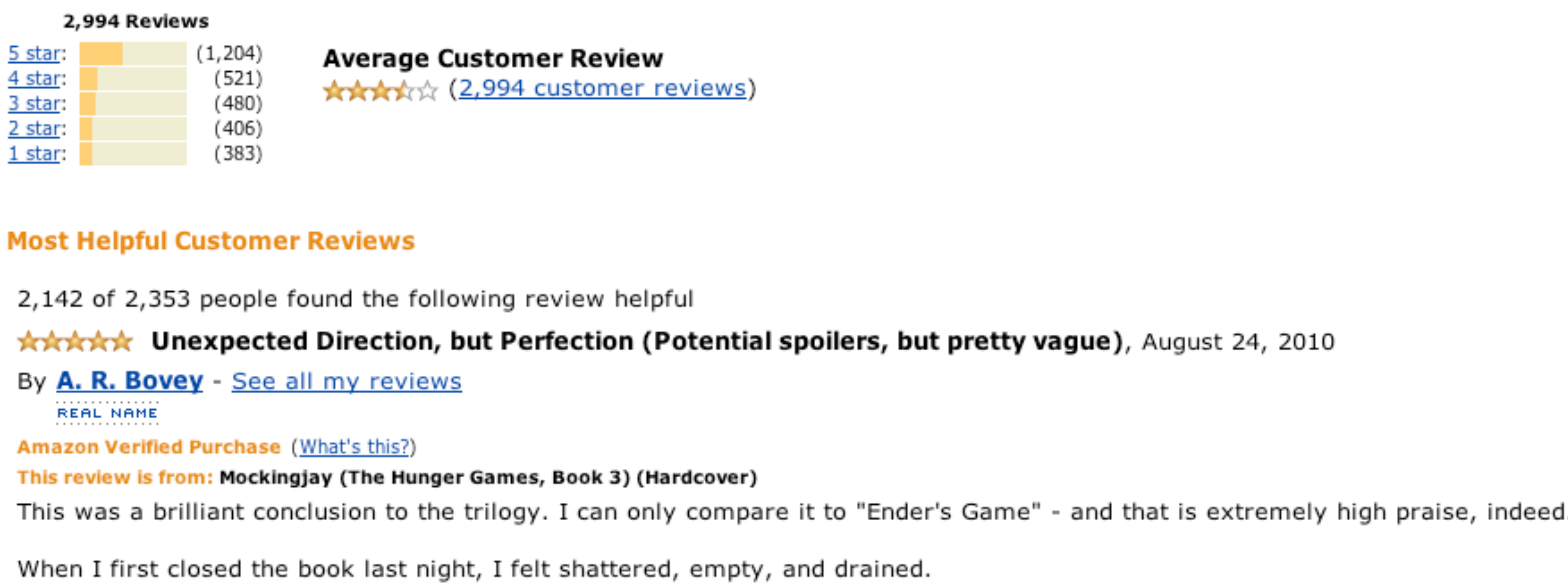
Term spam:

pro football sportsbooks nfl football line online nfl sportsbooks nfl football gambling odds online pro nfl betting pro nfl gambling online nfl football spreads offshore football gambling online nfl gamblibg spreads online football gambling line online nfl betting nfl sportsbook online online nfl betting spreads betting nfl football online online football wagering online gambling online gambling football online nfl football betting odds offshore football sportsbook online nfl football gambling ...

Link spam:

MVP Sportsbook Football Gambling Beverly Hills Football Sportsbook
Players SB Football Wagering Popular Poker Football Odds
Virtual Bookmaker Football Lines V Wager Football Spreads
Bogarts Casino Football Point Spreads Gecko Casino Online Football Betting
Jackpot Hour Online Football Gambling MVP Casino Online Football Wagering
Toucan Casino NFL Betting Popular Poker NFL Gambling
All Tracks NFL Wagering Bet Jockey NFL Odds
Live Horse Betting NFL Lines MVP Racebook NFL Point Spreads
Popular Poker NFL Spreads Bogarts Poker NFL Sportsbook ...

Sentiment



2,994 Reviews

5 star:		(1,204)
4 star:		(521)
3 star:		(480)
2 star:		(406)
1 star:		(383)

Average Customer Review

(2,994 customer reviews)

Most Helpful Customer Reviews

2,142 of 2,353 people found the following review helpful

Unexpected Direction, but Perfection (Potential spoilers, but pretty vague), August 24, 2010

By **A. R. Bovey** - See all my reviews

REAL NAME

Amazon Verified Purchase (What's this?)

This review is from: Mockingjay (The Hunger Games, Book 3) (Hardcover)

This was a brilliant conclusion to the trilogy. I can only compare it to "Ender's Game" - and that is extremely high praise, indeed.

When I first closed the book last night, I felt shattered, empty, and drained.

Sentiment

2,994 Reviews

5 star:		(1,204)
4 star:		(521)
3 star:		(480)
2 star:		(406)
1 star:		(383)

Average Customer Review
(2,994 customer reviews)

Most Helpful Customer Reviews

2,142 of 2,353 people found the following review helpful

Unexpected Direction, but Perfection (Potential spoilers, but pretty vague), August 24, 2010

By **A. R. Bovey** - See all my reviews

REAL NAME

Amazon Verified Purchase (What's this?)

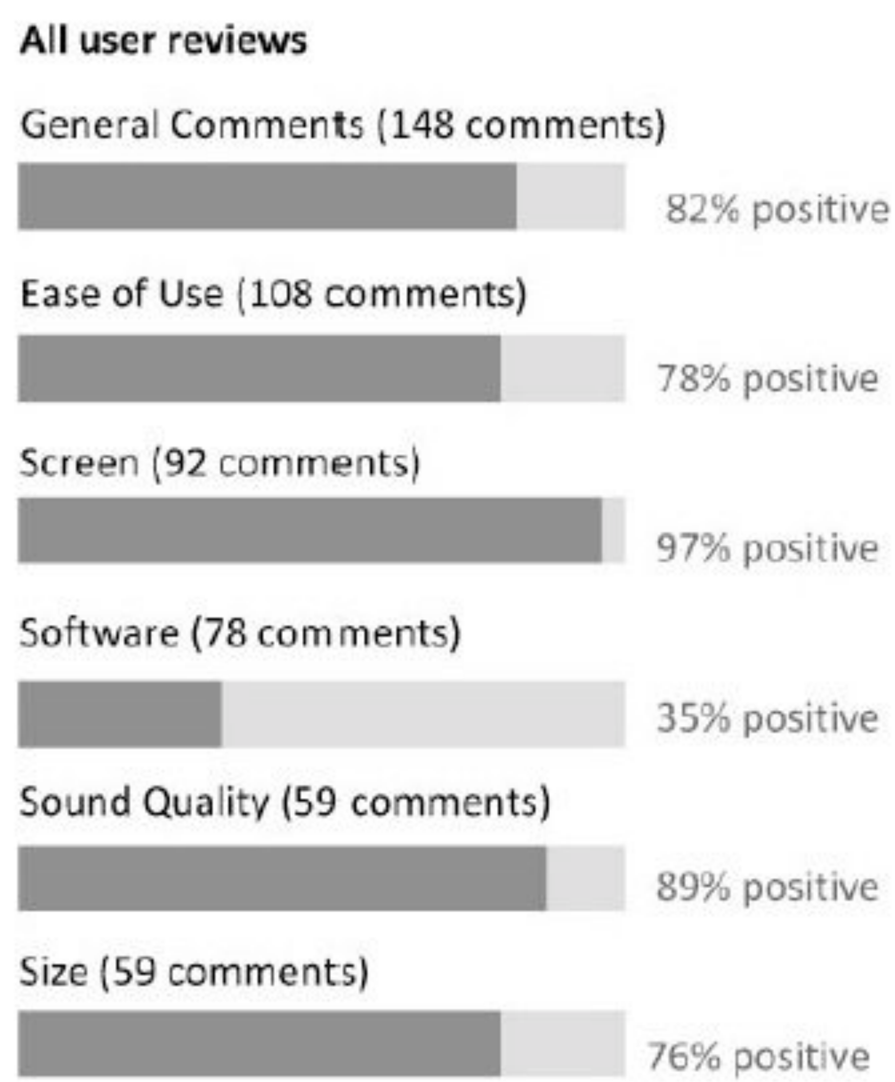
This review is from: **Mockingjay (The Hunger Games, Book 3) (Hardcover)**

This was a brilliant conclusion to the trilogy. I can only compare it to "Ender's Game" - and that is extremely high praise, indeed.

When I first closed the book last night, I felt shattered, empty, and drained.

Maybe not so good if found in a camera review

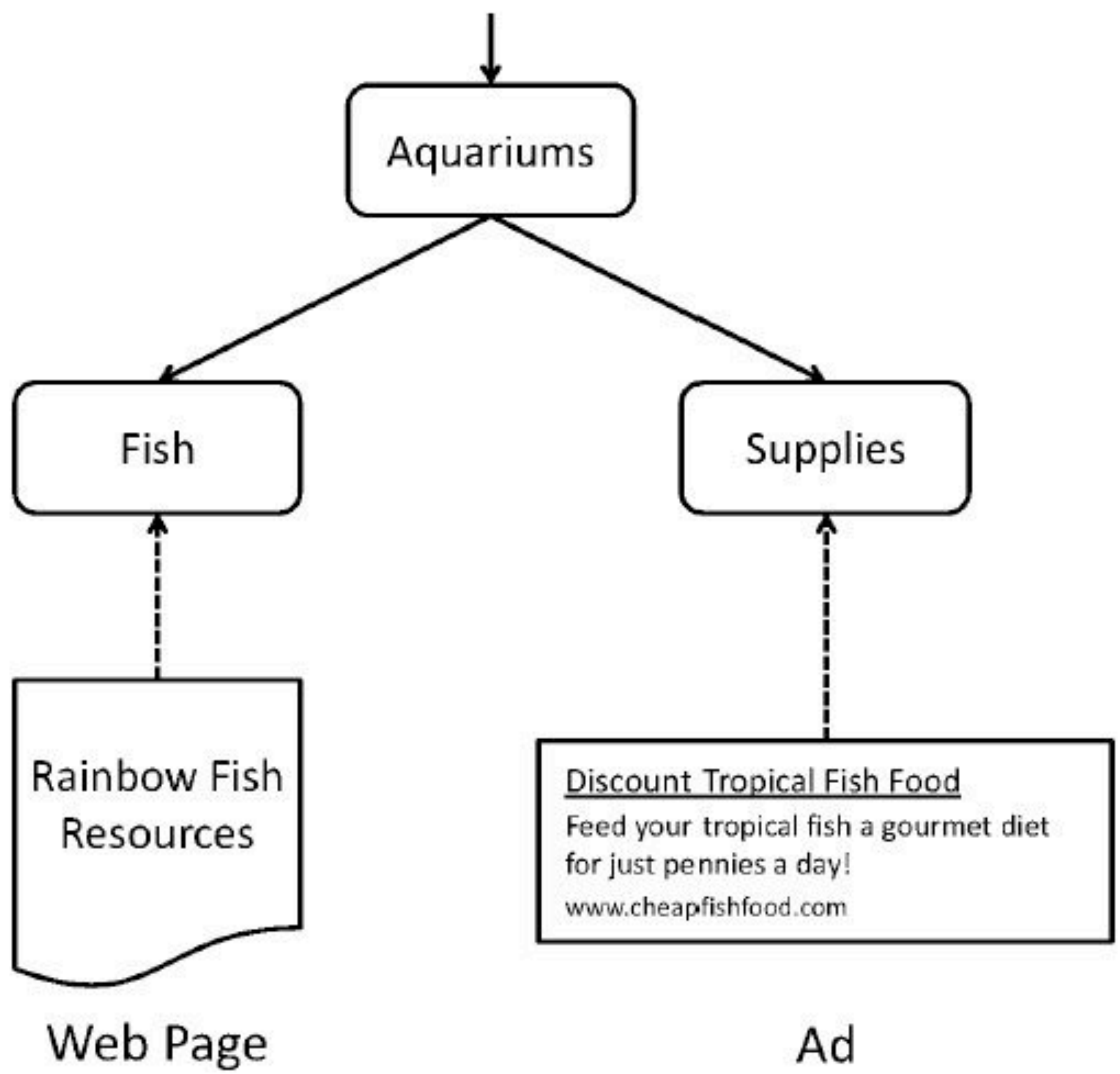
Sentiment


All user reviews
General Comments (148 comments)
82% positive
Ease of Use (108 comments)
78% positive
Screen (92 comments)
97% positive
Software (78 comments)
35% positive
Sound Quality (59 comments)
89% positive
Size (59 comments)
76% positive

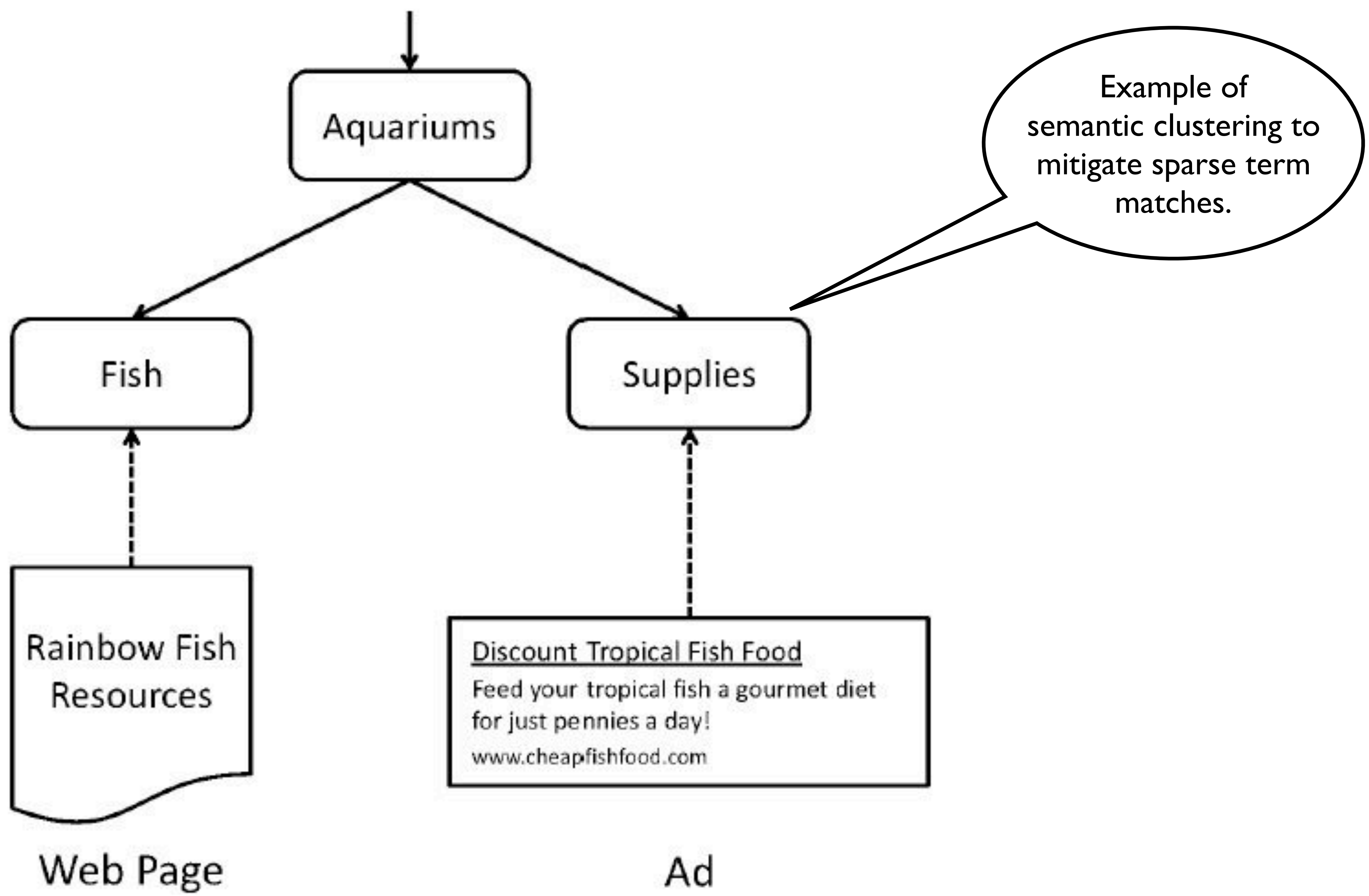
Advertising

- Search engines sell customer clicks from
  - Sponsored search
  - Content match
- Just retrieve ads topically like other docs?
  - Ads are very short and targeted
- Build specialized classifiers

Advertising



Advertising



Person Classification

Joseph Dwyer and David Smith headshots for Scientific American Inbox x

Chin, Ann achin@sciam.com via cs.umass.edu 3:48 PM (3 minutes ago)

to jdwyer, dasmith

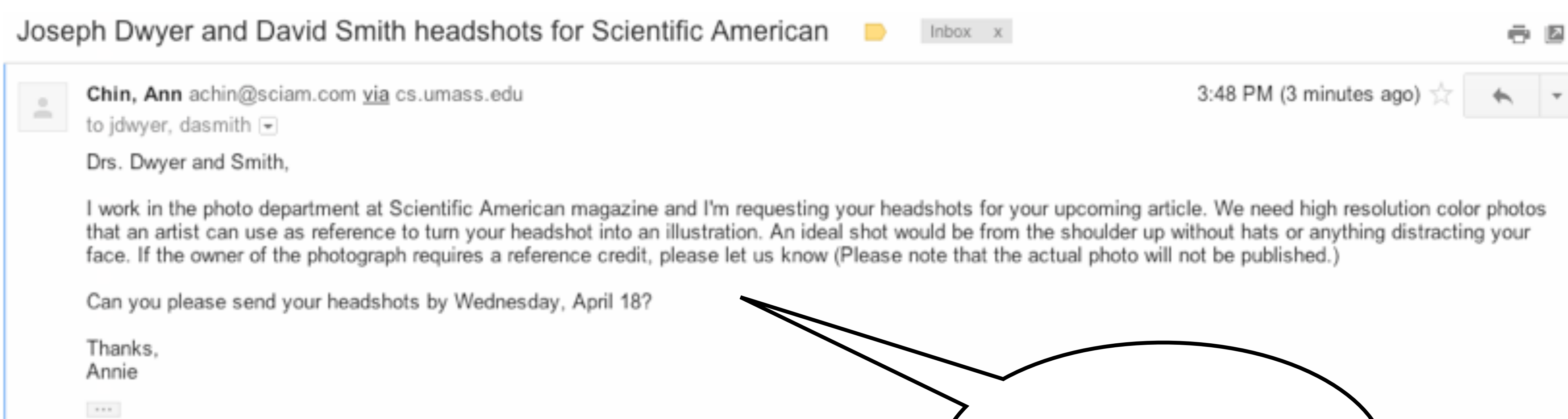
Drs. Dwyer and Smith,

I work in the photo department at Scientific American magazine and I'm requesting your headshots for your upcoming article. We need high resolution color photos that an artist can use as reference to turn your headshot into an illustration. An ideal shot would be from the shoulder up without hats or anything distracting your face. If the owner of the photograph requires a reference credit, please let us know (Please note that the actual photo will not be published.)

Can you please send your headshots by Wednesday, April 18?

Thanks,
Annie

Person Classification



Joseph Dwyer and David Smith headshots for Scientific American   Inbox x

Chin, Ann achin@sciam.com via cs.umass.edu   3:48 PM (3 minutes ago)

to jdwyer, dasmith

Drs. Dwyer and Smith,

I work in the photo department at Scientific American magazine and I'm requesting your headshots for your upcoming article. We need high resolution color photos that an artist can use as reference to turn your headshot into an illustration. An ideal shot would be from the shoulder up without hats or anything distracting your face. If the owner of the photograph requires a reference credit, please let us know (Please note that the actual photo will not be published.)

Can you please send your headshots by Wednesday, April 18?

Thanks,
Annie

I don't have a *Scientific American* article coming out.

Classification

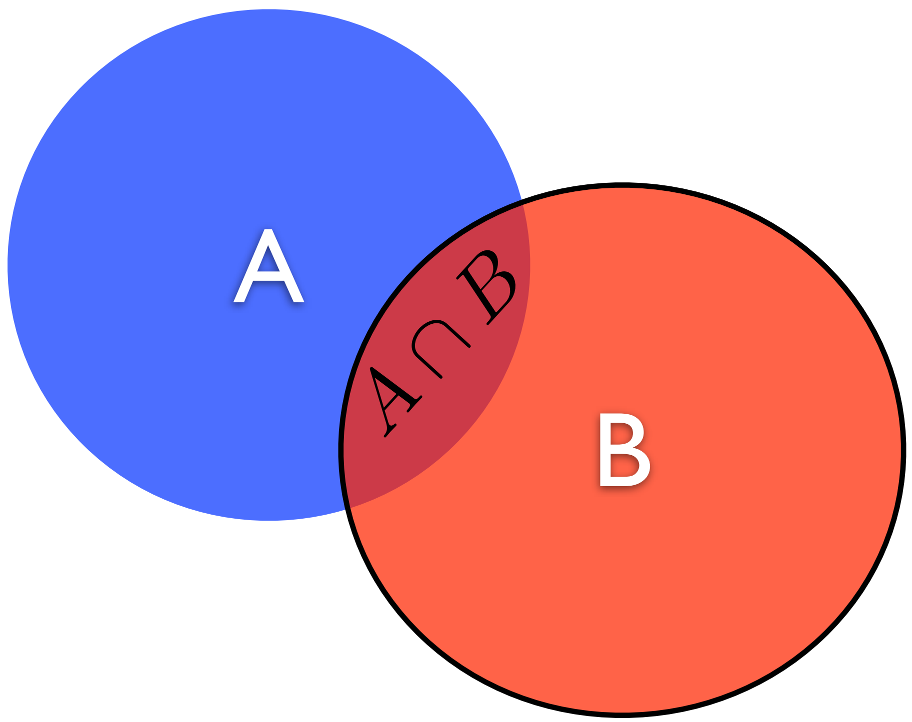
- Mapping from inputs to a finite output space
  - Contrast: *regression* and *ranking*
- Usually evaluated by *accuracy*
- Evaluated precision and recall if classes are very asymmetric in numbers or costliness (e.g., spam)
- Example: Naive Bayes
  - Simple, effective, similar to BM25
- Lots more: see book for SVM, nearest-neighbor

Axioms of Probability

- Define event space $\bigcup_i \mathcal{F}_i = \Omega$
- Probability function, s.t. $P : \mathcal{F} \rightarrow [0, 1]$
  - Disjoint events sum $A \cap B = \emptyset \Leftrightarrow P(A \cup B) = P(A) + P(B)$
  - All events sum to one $P(\Omega) = 1$
- Show that: $P(\bar{A}) = 1 - P(A)$

Conditional Probability

$$P(A \mid B) = \frac{P(A,B)}{P(B)}$$



$$P(A,B) = P(B)P(A \mid B) = P(A)P(B \mid A)$$

$$\begin{aligned} P(A_1, A_2, \ldots, A_n) \quad &= \quad P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1, A_2) \\ &\qquad \cdots P(A_n \mid A_1, \ldots, A_{n-1}) \end{aligned}$$

Chain rule

Independence

$$
\begin{aligned}
P(A, B) &= P(A)P(B) \\
&\Leftrightarrow \\
P(A \mid B) = P(A) \quad &\wedge \quad P(B \mid A) = P(B)
\end{aligned}
$$

In coding terms, knowing *B* doesn't help in decoding *A*, and vice versa.

Movie Reviews

Movie Reviews

there ' s some movies i enjoy even though i know i probably shouldn ' t and have a difficult time trying to explain why i did . " lucky numbers " is a perfect example of this because it ' s such a blatant rip - off of " fargo " and every movie based on an elmore leonard novel and yet it somehow still works for me . i know i ' m in the minority here but let me explain . the film takes place in harrisburg , pa in 1988 during an unseasonably warm winter . ...

Movie Reviews

☺

there ' s some movies i enjoy even though i know i probably shouldn ' t and have a difficult time trying to explain why i did . " lucky numbers " is a perfect example of this because it ' s such a blatant rip - off of " fargo " and every movie based on an elmore leonard novel and yet it somehow still works for me . i know i ' m in the minority here but let me explain . the film takes place in harrisburg , pa in 1988 during an unseasonably warm winter . ...

Movie Reviews

☺

there ' s some movies i enjoy even though i know i probably shouldn ' t and have a difficult time trying to explain why i did . " lucky numbers " is a perfect example of this because it ' s such a blatant rip - off of " fargo " and every movie based on an elmore leonard novel and yet it somehow still works for me . i know i ' m in the minority here but let me explain . the film takes place in harrisburg , pa in 1988 during an unseasonably warm winter . ...

seen at : amc old pasadena 8 , pasadena , ca (in sdds) paul verhoeven ' s last movie , showgirls , had a bad script , bad acting , and a " plot " (i use the word in its loosest possible sense) that served only to allow lots of sex and nudity . it stank . starship troopers has a bad script , bad acting , and a " plot " that serves only to allow lots of violence and gore . it stinks . nobody will watch this movie for the plot , ...

Movie Reviews

☺

there ' s some movies i enjoy even though i know i probably shouldn ' t and have a difficult time trying to explain why i did . " lucky numbers " is a perfect example of this because it ' s such a blatant rip - off of " fargo " and every movie based on an elmore leonard novel and yet it somehow still works for me . i know i ' m in the minority here but let me explain . the film takes place in harrisburg , pa in 1988 during an unseasonably warm winter . ...

☹

seen at : amc old pasadena 8 , pasadena , ca (in sdds) paul verhoeven ' s last movie , showgirls , had a bad script , bad acting , and a " plot " (i use the word in its loosest possible sense) that served only to allow lots of sex and nudity . it stank . starship troopers has a bad script , bad acting , and a " plot " that serves only to allow lots of violence and gore . it stinks . nobody will watch this movie for the plot , ...

Movie Reviews

☺

there ' s some movies i enjoy even though i know i probably shouldn ' t and have a difficult time trying to explain why i did . " lucky numbers " is a perfect example of this because it ' s such a blatant rip - off of " fargo " and every movie based on an elmore leonard novel and yet it somehow still works for me . i know i ' m in the minority here but let me explain . the film takes place in harrisburg , pa in 1988 during an unseasonably warm winter . ...

☹

seen at : amc old pasadena 8 , pasadena , ca (in sdds) paul verhoeven ' s last movie , showgirls , had a bad script , bad acting , and a " plot " (i use the word in its loosest possible sense) that served only to allow lots of sex and nudity . it stank . starship troopers has a bad script , bad acting , and a " plot " that serves only to allow lots of violence and gore . it stinks . nobody will watch this movie for the plot , ...

the rich legacy of cinema has left us with certain indelible images . the tinkling christmas tree bell in " it ' s a wonderful life . " bogie ' s speech at the airport in " casablanca . " little elliott ' s flying bicycle , silhouetted by the moon in " e . t . " and now , " starship troopers " director paul verhoeven adds one more image that will live in our memories forever : doogie houser doing a vulcan mind meld with a giant slug . " starship troopers , " loosely based on

Movie Reviews

☺ there ' s some movies i enjoy even though i know i probably shouldn ' t and have a difficult time trying to explain why i did . " lucky numbers " is a perfect example of this because it ' s such a blatant rip - off of " fargo " and every movie based on an elmore leonard novel and yet it somehow still works for me . i know i ' m in the minority here but let me explain . the film takes place in harrisburg , pa in 1988 during an unseasonably warm winter . ...

☹ seen at : amc old pasadena 8 , pasadena , ca (in sdds) paul verhoeven ' s last movie , showgirls , had a bad script , bad acting , and a " plot " (i use the word in its loosest possible sense) that served only to allow lots of sex and nudity . it stank . starship troopers has a bad script , bad acting , and a " plot " that serves only to allow lots of violence and gore . it stinks . nobody will watch this movie for the plot , ...

☹ the rich legacy of cinema has left us with certain indelible images . the tinkling christmas tree bell in " it ' s a wonderful life . " bogie ' s speech at the airport in " casablanca . " little elliott ' s flying bicycle , silhouetted by the moon in " e . t . " and now , " starship troopers " director paul verhoeven adds one more image that will live in our memories forever : doogie houser doing a vulcan mind meld with a giant slug . " starship troopers , " loosely based on

Setting up a Classifier

Setting up a Classifier

- What we want:

  $p(☺ \mid w_1, w_2, ..., w_n) > p(☹ \mid w_1, w_2, ..., w_n)$?

Setting up a Classifier

- What we want:

  $p(☺ \mid w_1, w_2, ..., w_n) > p(☹ \mid w_1, w_2, ..., w_n)$?

- What we know how to build:

Setting up a Classifier

- What we want:

  $p(☺ \mid w_1, w_2, \ldots, w_n) > p(☹ \mid w_1, w_2, \ldots, w_n)$?

- What we know how to build:
  - A language model for each class

Setting up a Classifier

- What we want:

  $p(☺ \mid w_1, w_2, ..., w_n) > p(☹ \mid w_1, w_2, ..., w_n)$?

- What we know how to build:
  - A language model for each class
    - $p(w_1, w_2, ..., w_n \mid ☺)$

Setting up a Classifier

- What we want:

  p(☺ | $w_1, w_2, \ldots, w_n$) > p(☹ | $w_1, w_2, \ldots, w_n$) ?

- What we know how to build:
  - A language model for each class
    - p($w_1, w_2, \ldots, w_n$ | ☺)
    - p($w_1, w_2, \ldots, w_n$ | ☹)

Bayes' Theorem

By the definition of conditional probability:

$$P(A, B) = P(B)P(A \mid B) = P(A)P(B \mid A)$$

we can show:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Seemingly trivial result from 1763;
interesting consequences...

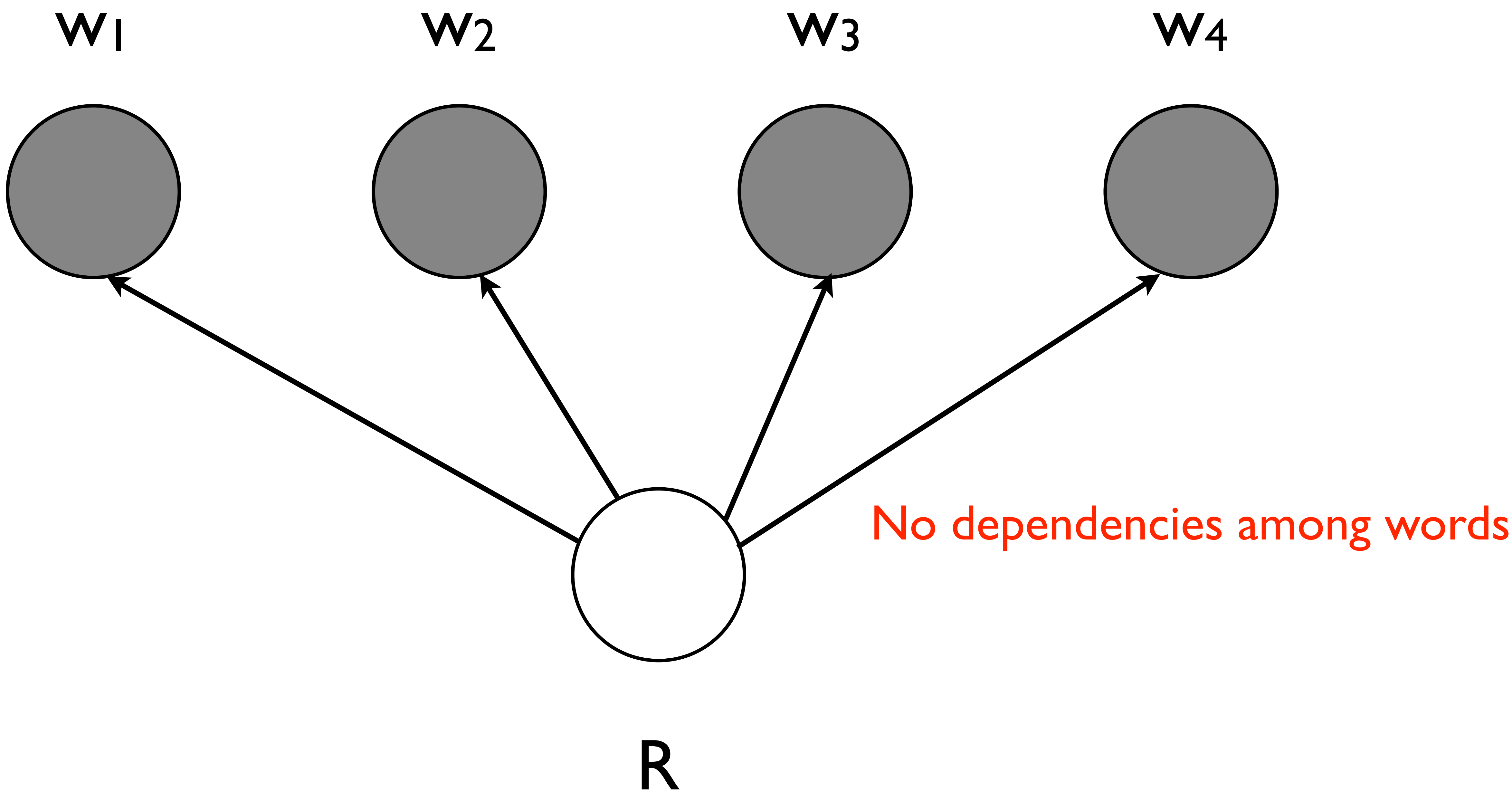
REV. T. BAYES

A "Bayesian" Classifier

$$p(R \mid w_1, w_2, \ldots, w_n) = \frac{p(R)p(w_1, w_2, \ldots, w_n \mid R)}{p(w_1, w_2, \ldots, w_n)}$$

$$\max_{R \in \{\ddot\smile, \ddot\frown\}} \underbrace{p(R \mid w_1, w_2, \ldots, w_n)}_{\text{Posterior}} = \max_{R \in \{\ddot\smile, \ddot\frown\}} \underbrace{p(R)}_{\text{Prior}} \underbrace{p(w_1, w_2, \ldots, w_n \mid R)}_{\text{Likelihood}}$$

Naive Bayes Classifier



NB on Movie Reviews

- Train models for positive, negative
- For each review, find higher posterior
- Which word probability ratios are highest?

```
>>> classifier.show_most_informative_features(5)

classifier.show_most_informative_features(5)
Most Informative Features
   contains(outstanding) = True              pos : neg    =     14.1 : 1.0
         contains(mulan) = True              pos : neg    =      8.3 : 1.0
        contains(seagal) = True              neg : pos    =      7.8 : 1.0
   contains(wonderfully) = True              pos : neg    =      6.6 : 1.0
         contains(damon) = True              pos : neg    =      6.1 : 1.0
```

What's Wrong With NB?

- What happens for word dependencies are strong?
- What happens when some words occur only once?
- What happens when the classifier sees a new word?

ML for Naive Bayes

- Recall: p(+ | Damon movie)

  = p(Damon | +) p(movie | +) p(+)

- If corpus of positive reviews has 1000 words, and "Damon" occurs 50 times,

  p_{ML}(Damon | +) = ?

- If pos. corpus has "Affleck" 0 times,

  p(+ | Affleck Damon movie) = ?

Will the Sun Rise Tomorrow?

Will the Sun Rise Tomorrow?

Laplace's Rule of Succession:
On day n+1, we've observed that the sun has risen s times before.

$$p_{Lap}(S_{n+1} = 1 \mid S_1 + \cdots + S_n = s) = \frac{s+1}{n+2}$$

What's the probability on day 0?
On day 1?
On day 10^6?
Start with prior assumption of equal rise/not-rise probabilities; *update* after every observation.

Clustering

Clustering

- Unsupervised structure discovery
- Exploratory data analysis
- Clustering for word senses
- Clustering for retrieval effectiveness
  - Some have also proposed clustering for efficiency

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

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 17^{th} 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
  = 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)

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)

count too low

From corpus:

Arlen Specter **abandoned** the Republican **party**.
There were lots of **abbot**s and nuns dancing at that **party**.
The **party** **above** the art gallery was, **above** all, a laboratory for synthesizing **zygote**s 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
  - 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?

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

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

how might you measure this?

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

- 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 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
- 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 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

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

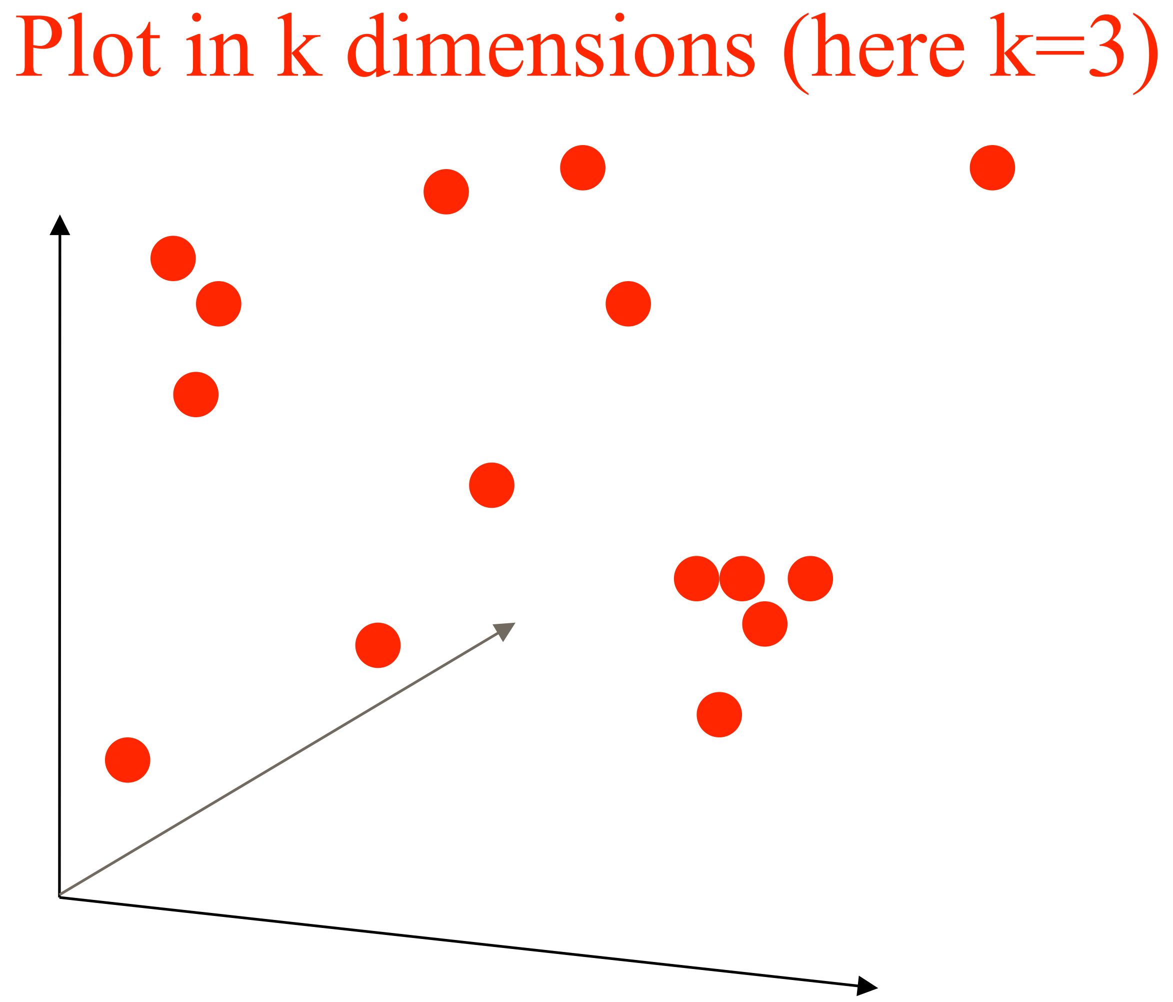
aardvark abacus abandoned abbot abduct above zygote zymurgy

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

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

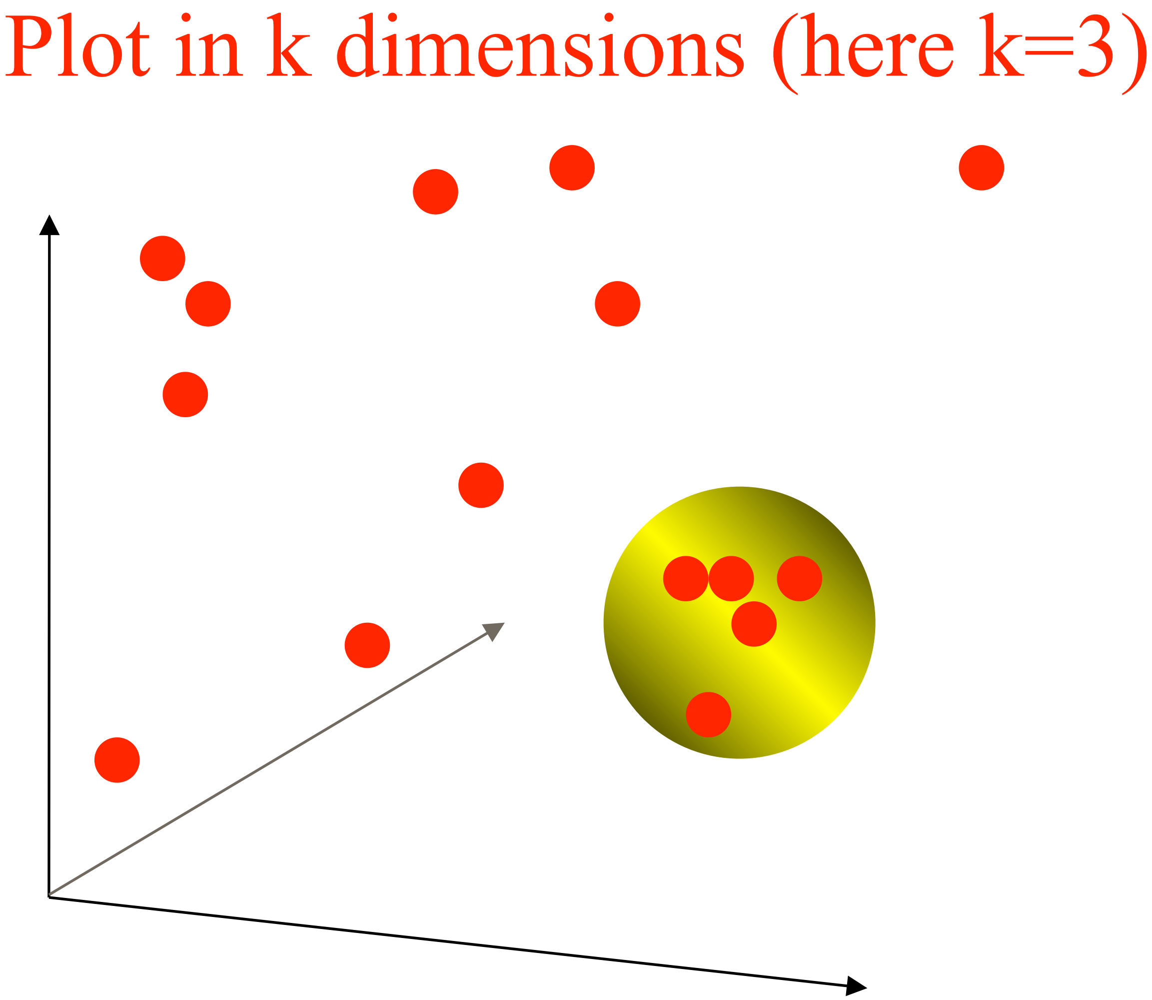
Learning Classes by Clustering

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



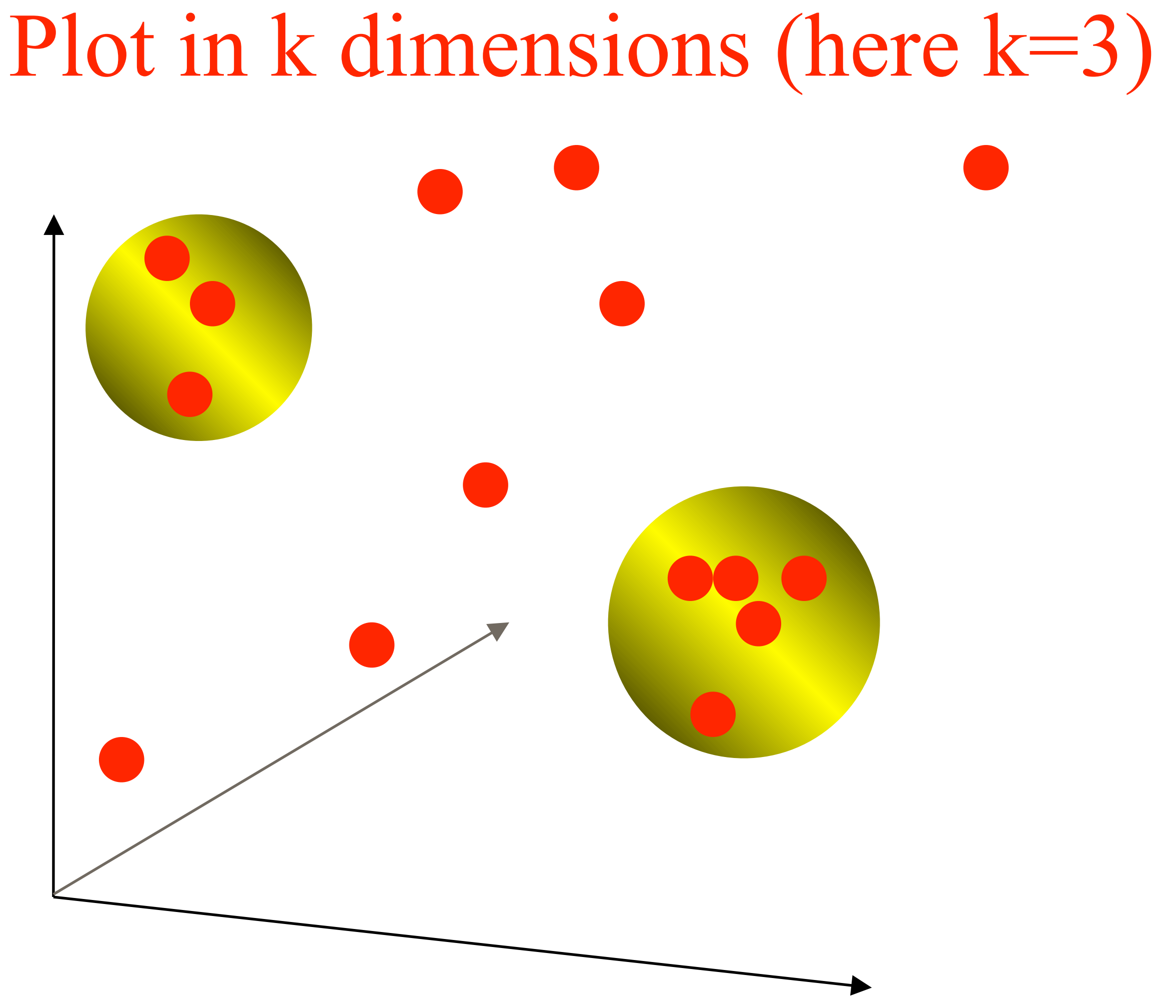
Learning Classes by Clustering

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



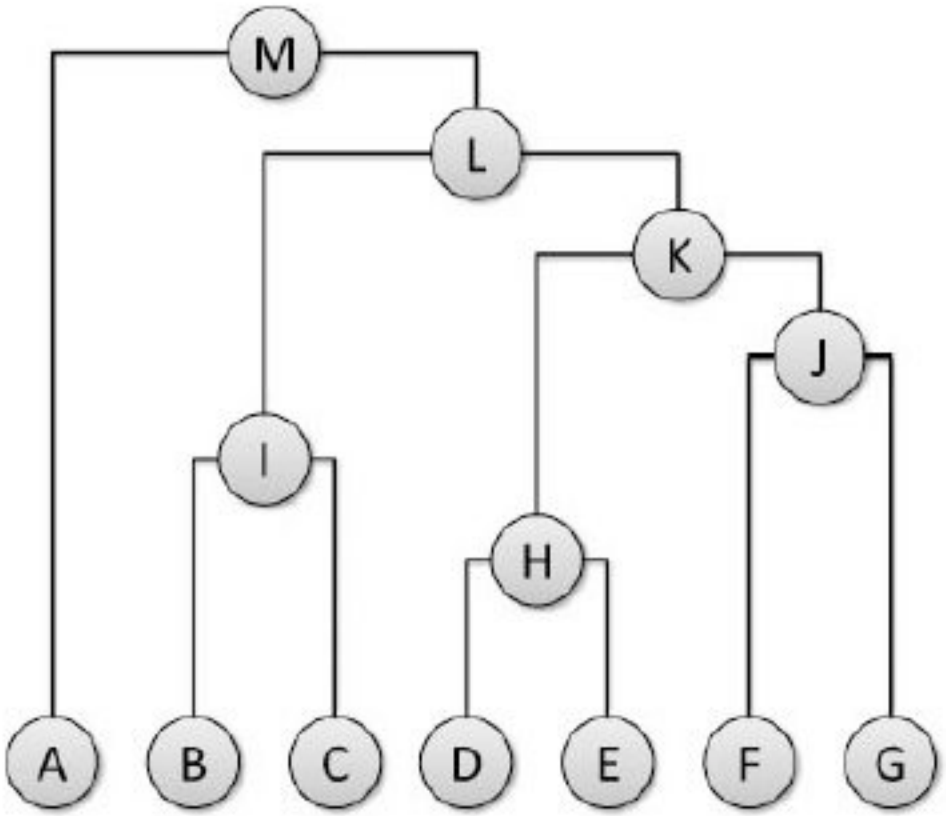
Learning Classes by Clustering

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



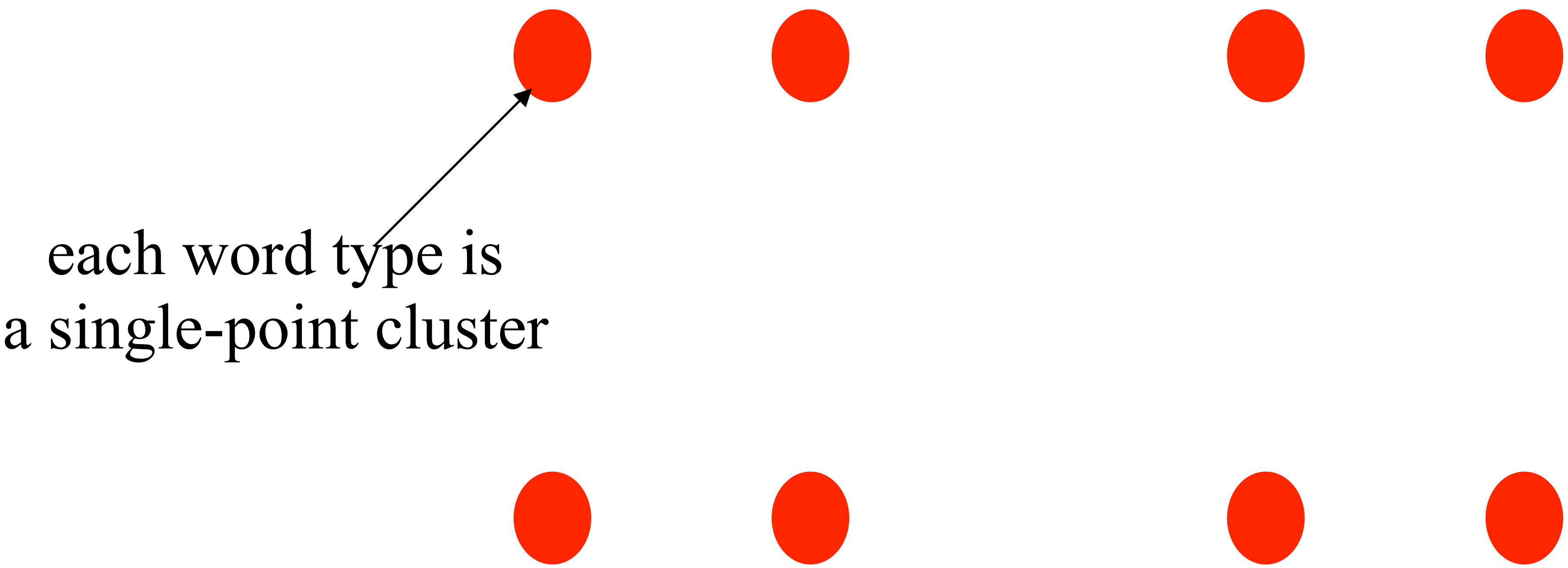
Bottom-Up Clustering

- 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 $a \in A$, $b \in B$
- Produces a dendrogram



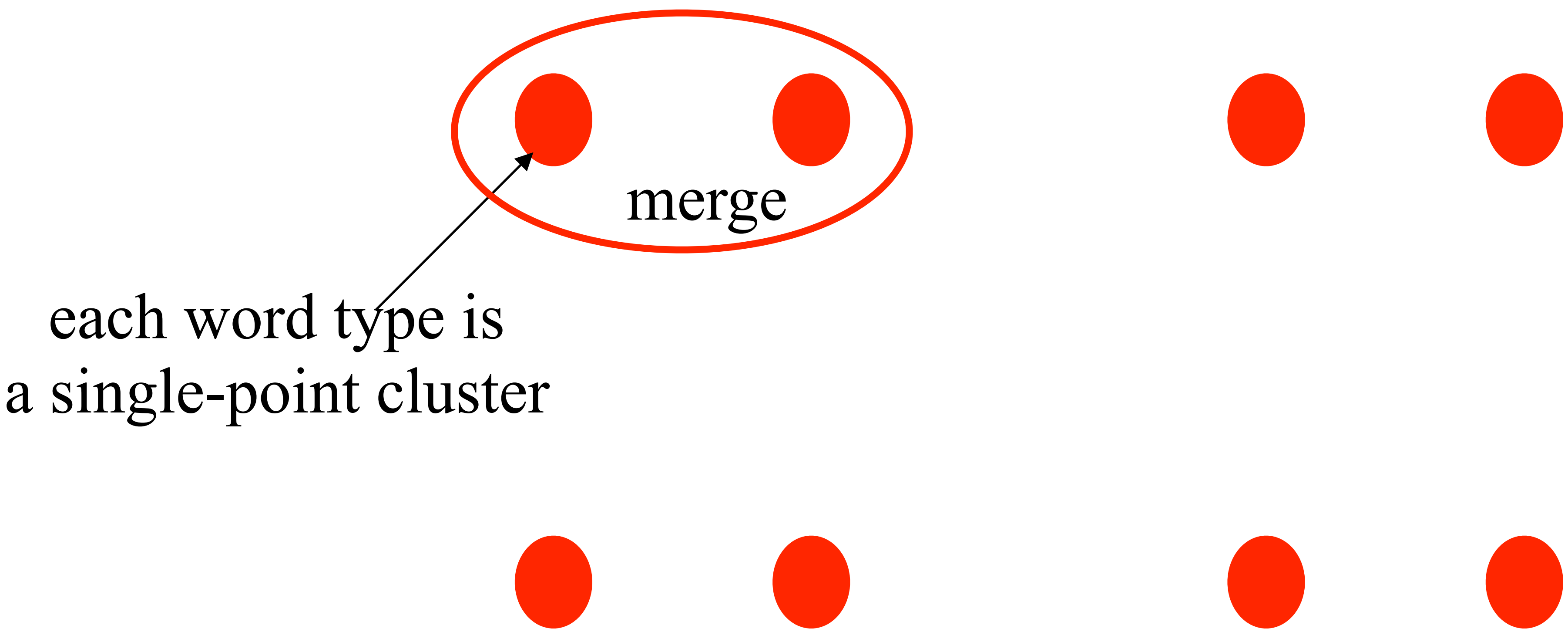
example from Manning & Schütze

Bottom-Up Clustering – Single-Link



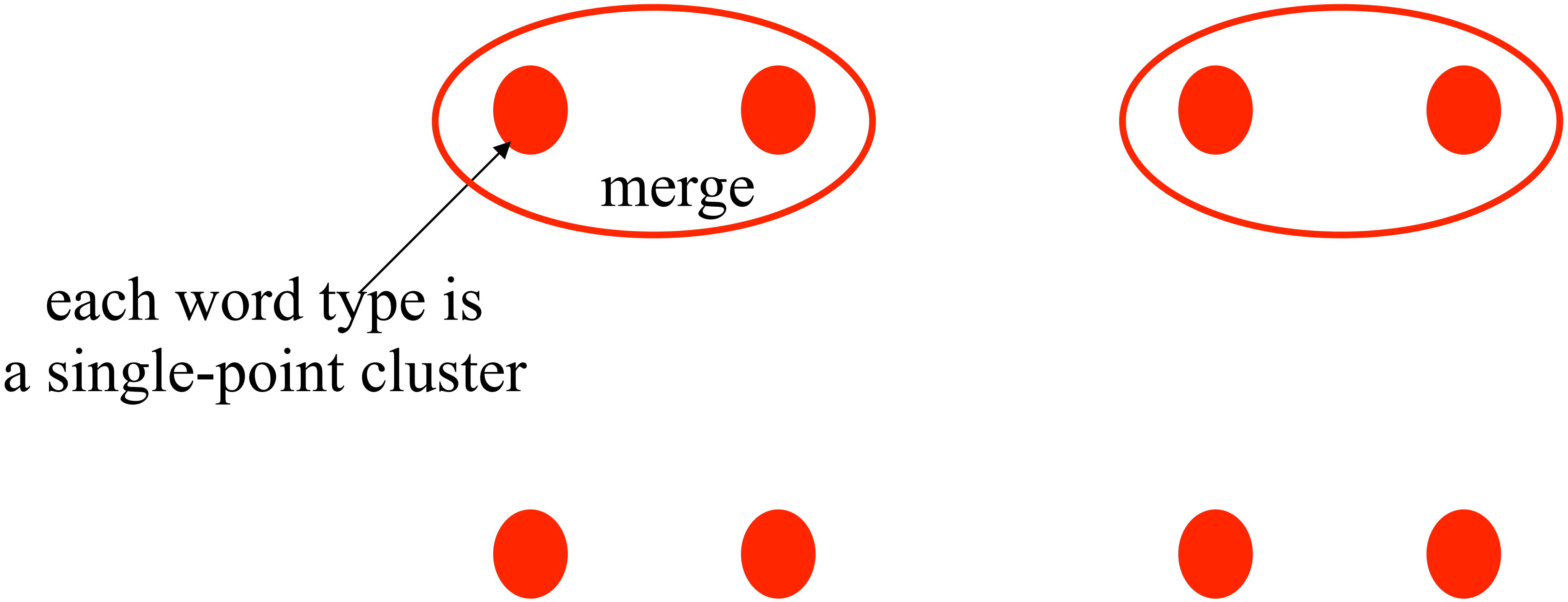
example from Manning & Schütze

Bottom-Up Clustering – Single-Link



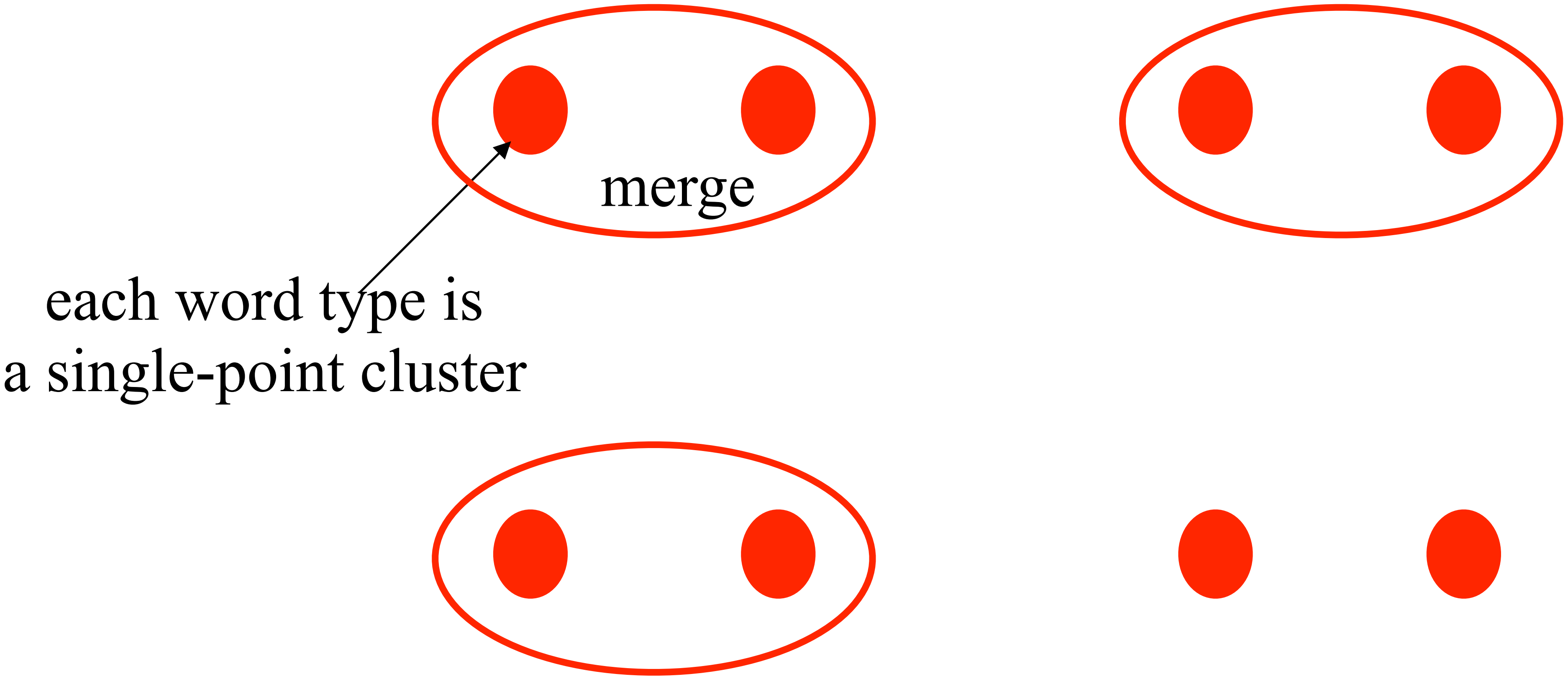
example from Manning & Schütze

Bottom-Up Clustering – Single-Link



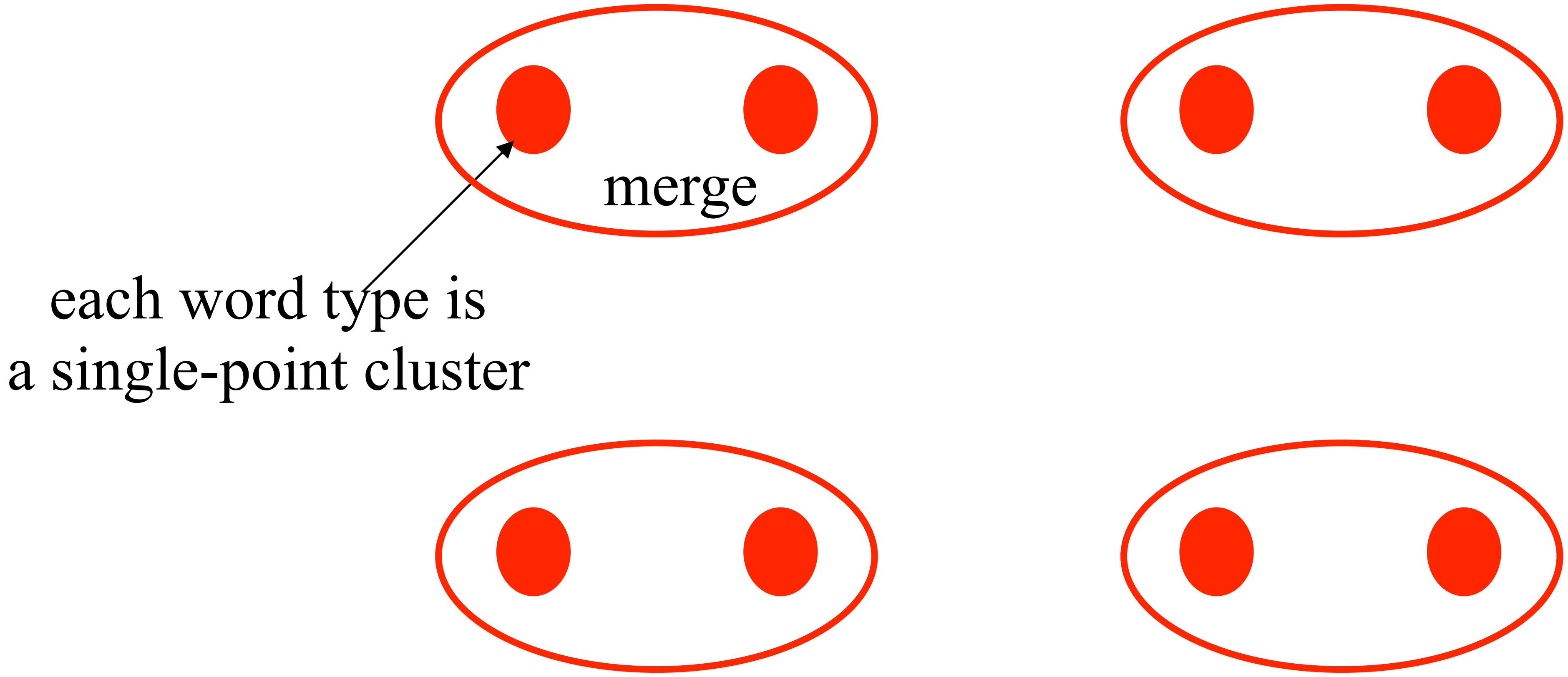
example from Manning & Schütze

Bottom-Up Clustering – Single-Link



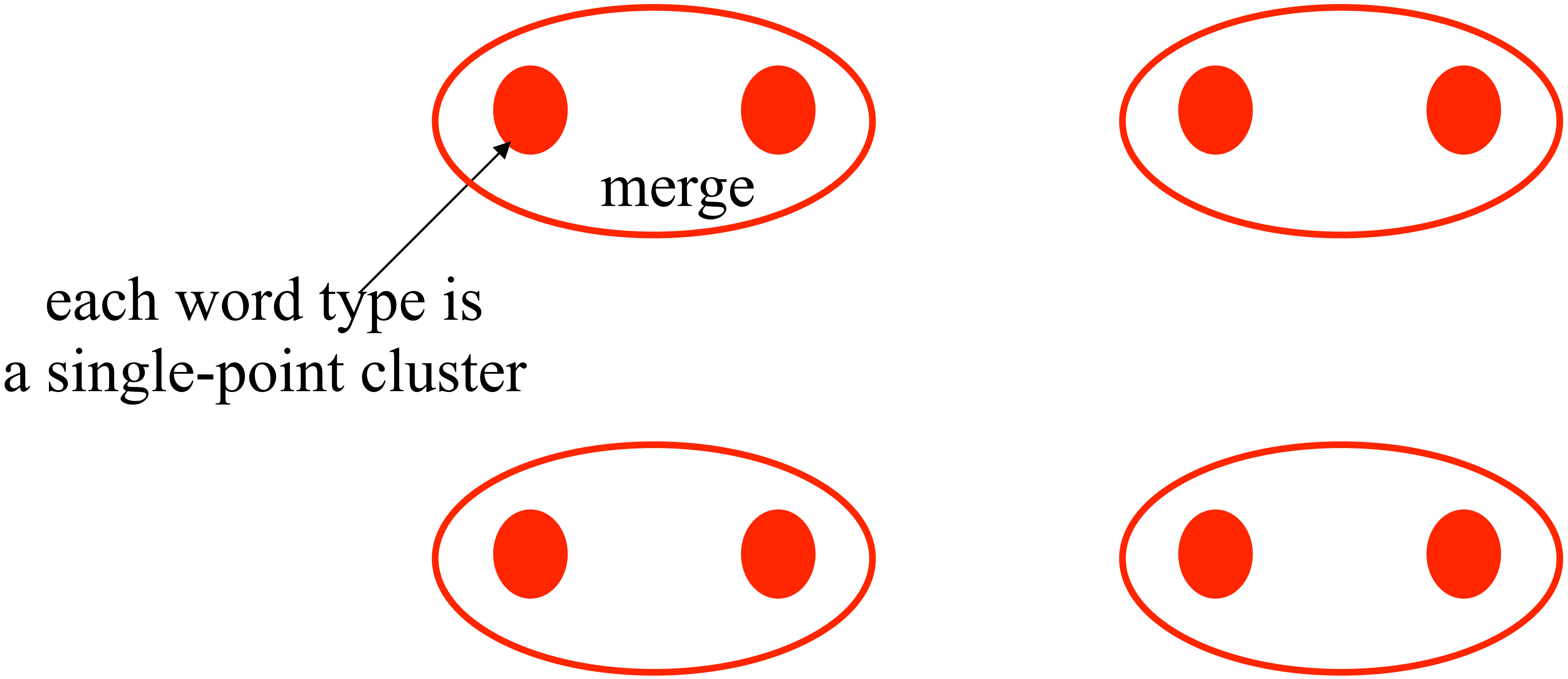
example from Manning & Schütze

Bottom-Up Clustering – Single-Link



example from Manning & Schütze

Bottom-Up Clustering – Single-Link

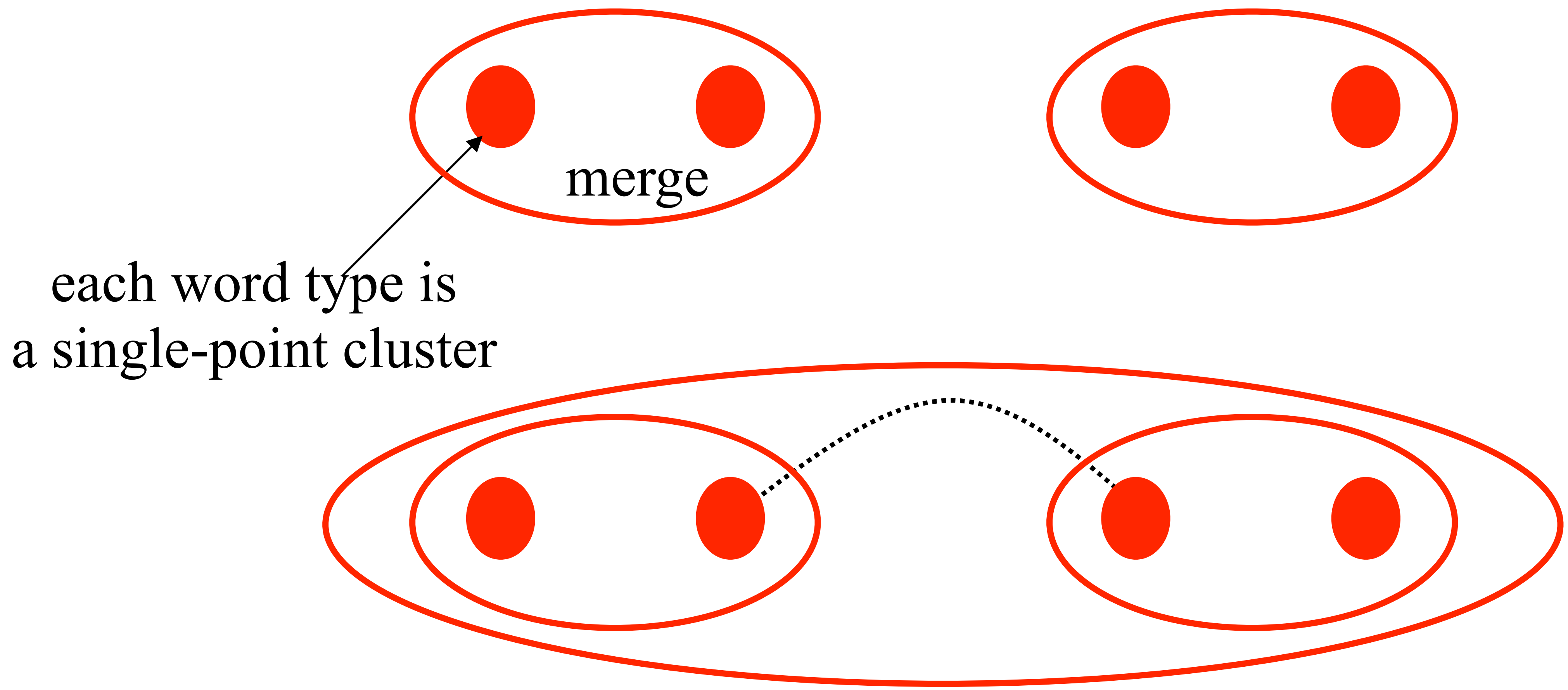


Again, merge closest pair of clusters:

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

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

Bottom-Up Clustering – Single-Link



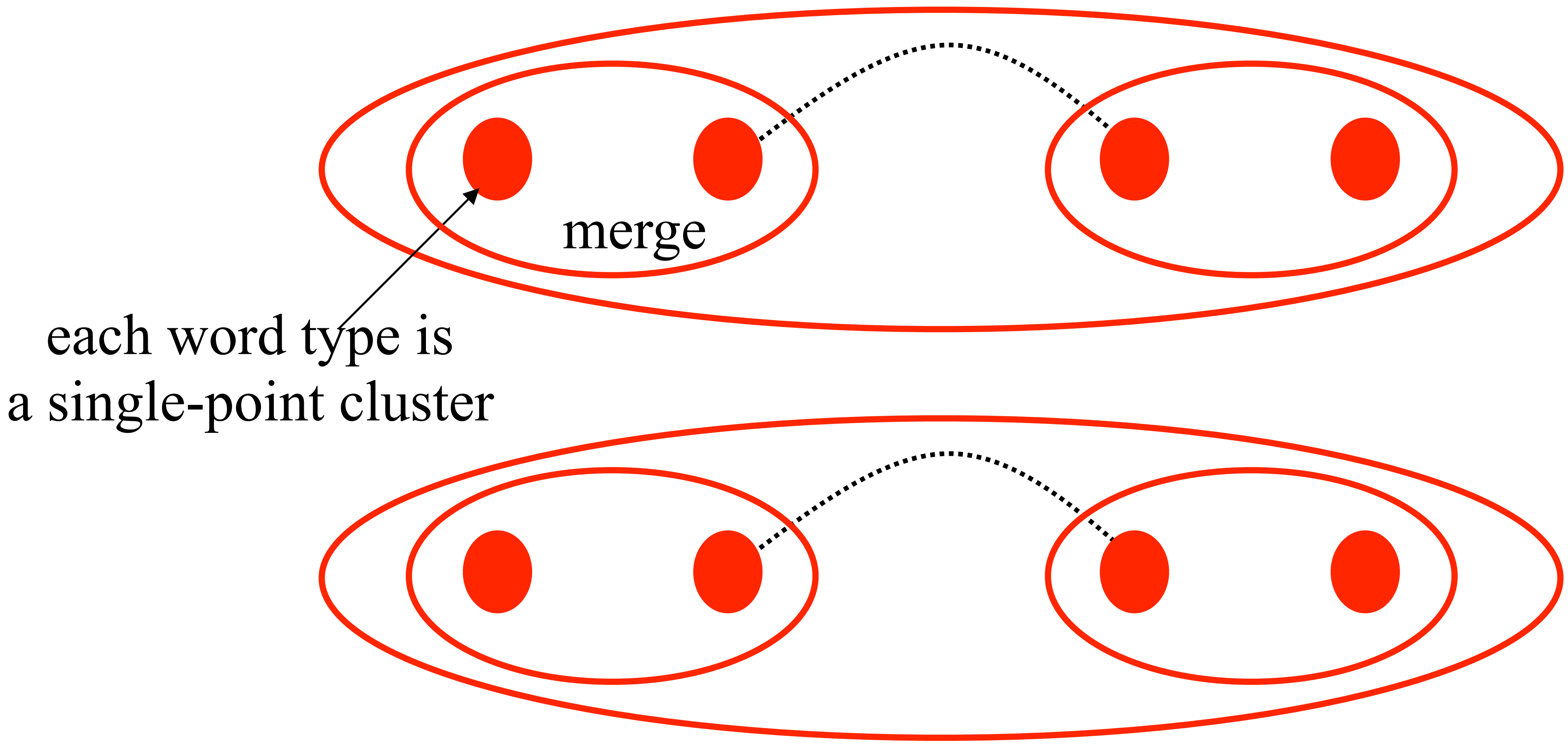
Again, merge closest pair of clusters:

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

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

example from Manning & Schütze

Bottom-Up Clustering – Single-Link

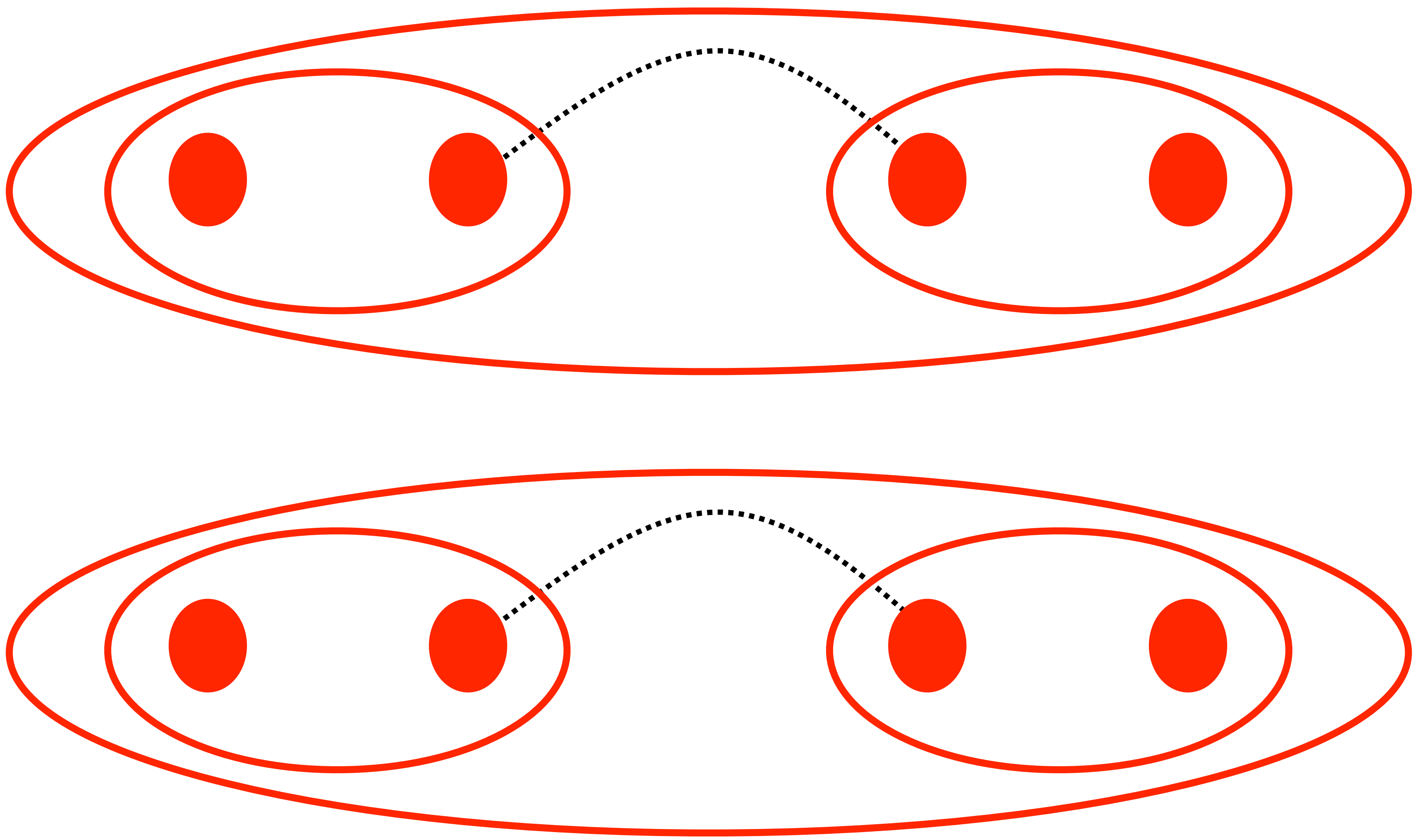


Again, merge closest pair of clusters:

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

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

Bottom-Up Clustering – Single-Link



Again, merge closest pair of clusters:

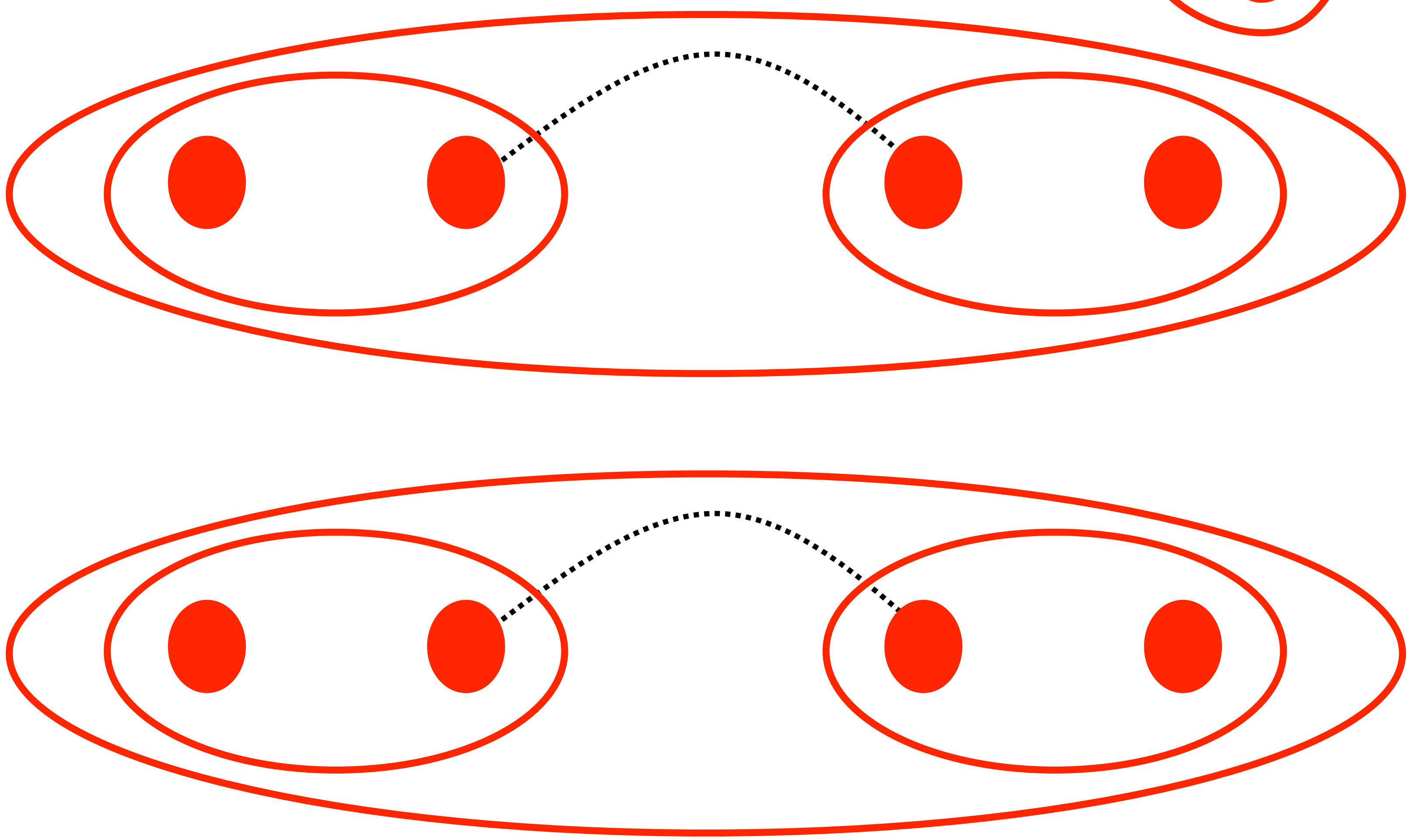
Single-link: clusters are close if any of their points are

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

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

example from Manning & Schütze

Bottom-Up Clustering – Single-Link



Again, merge closest pair of clusters:

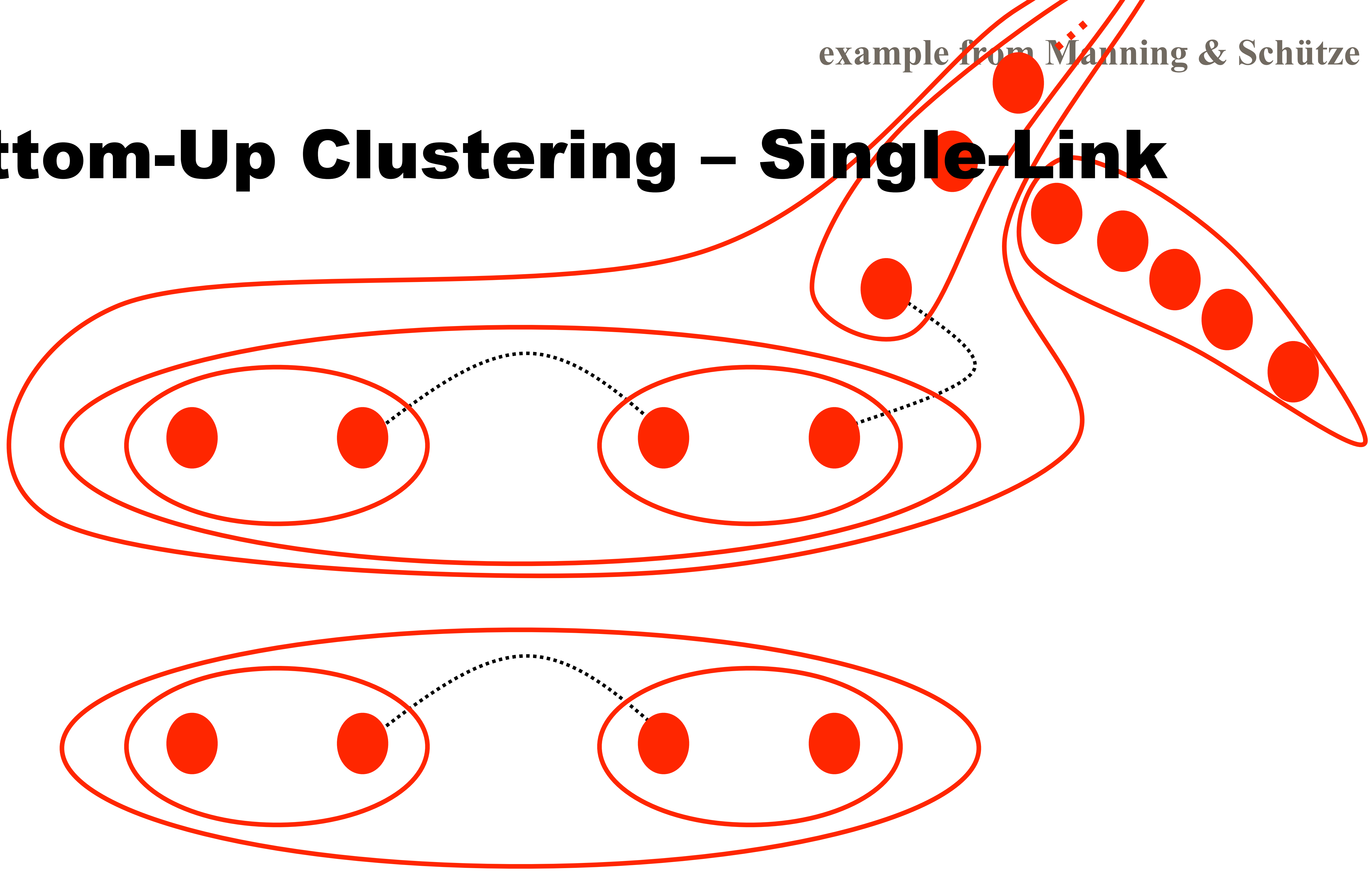
Single-link: clusters are close if any of their points are

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:

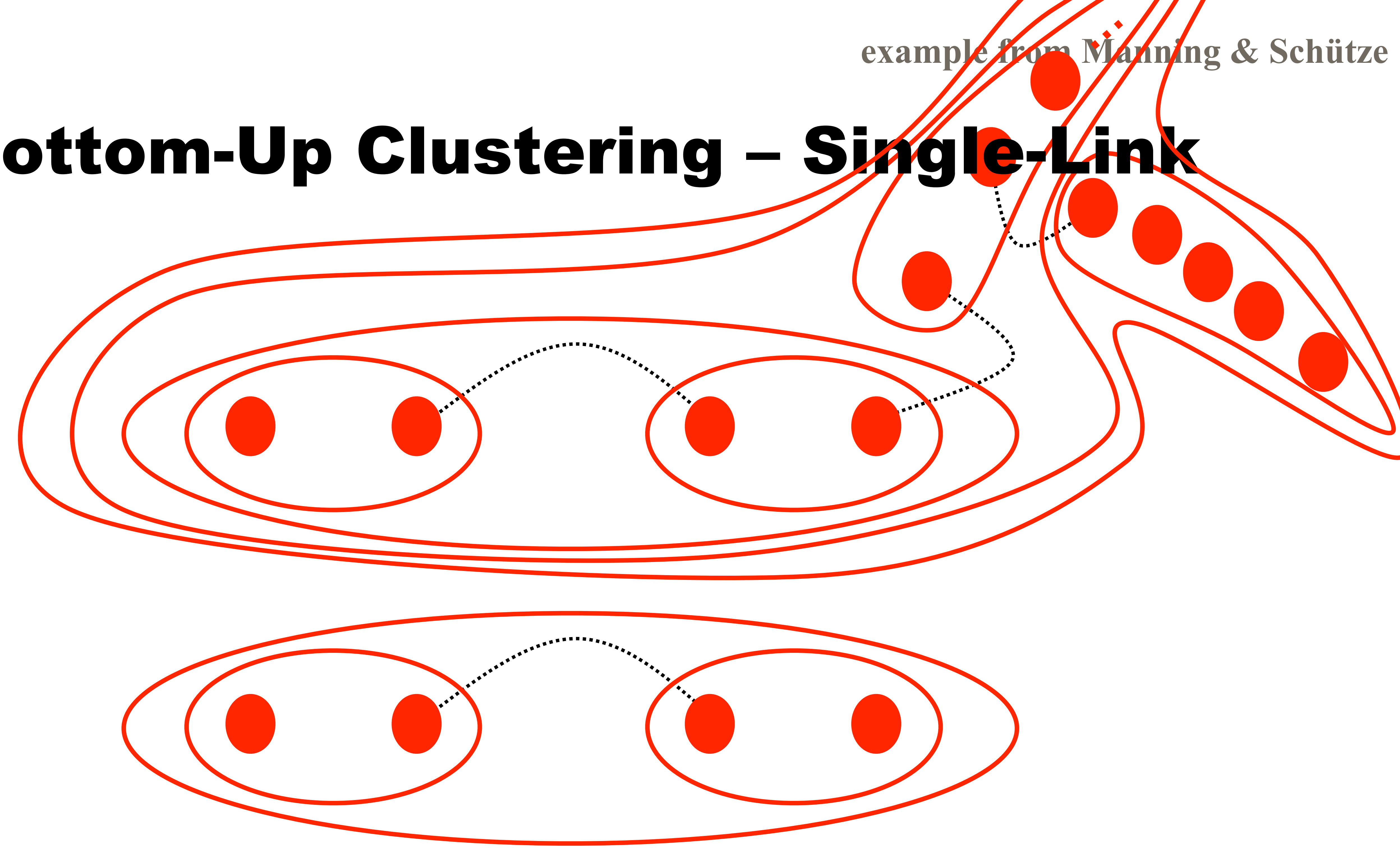
Single-link: clusters are close if any of their points are

$\text{dist}(A,B) = \min \text{dist}(a,b)$ 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:

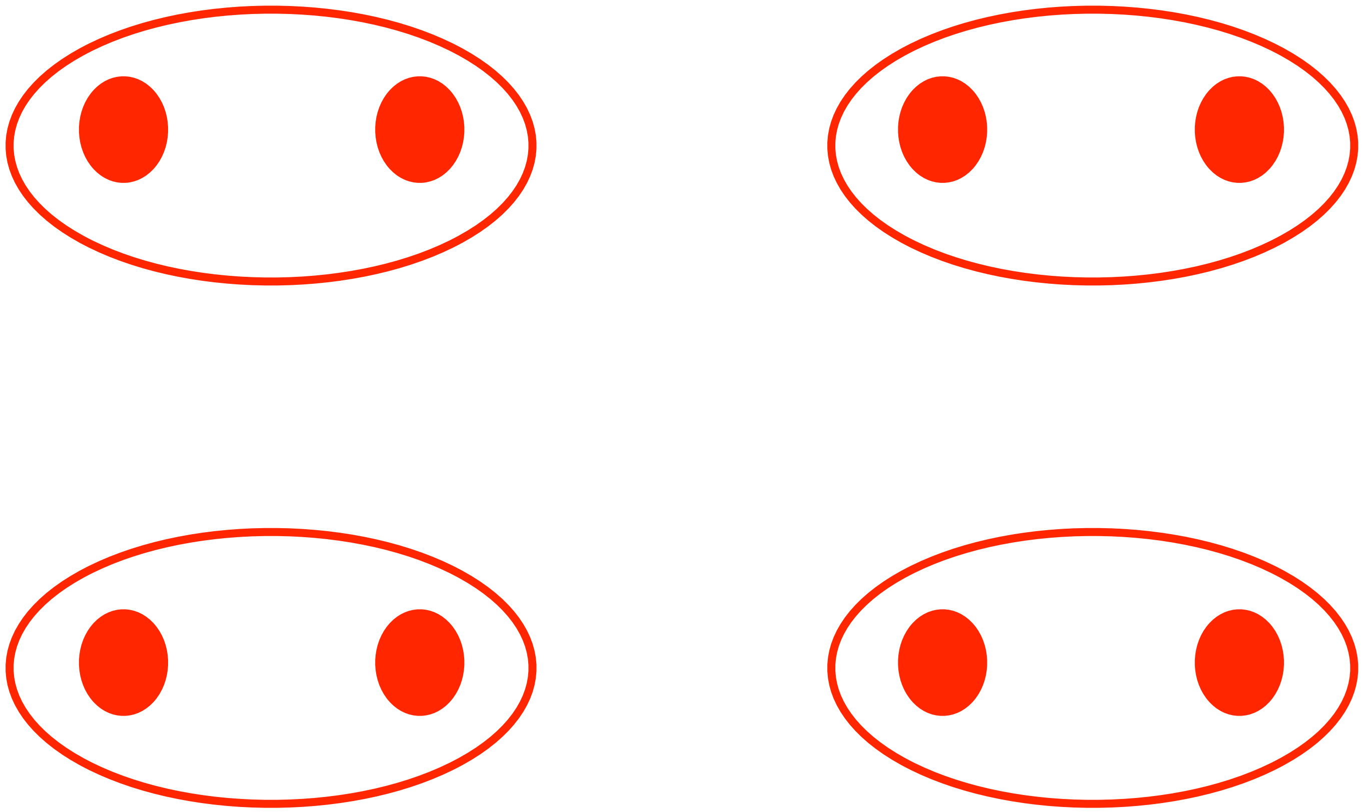
Single-link: clusters are close if any of their points are

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

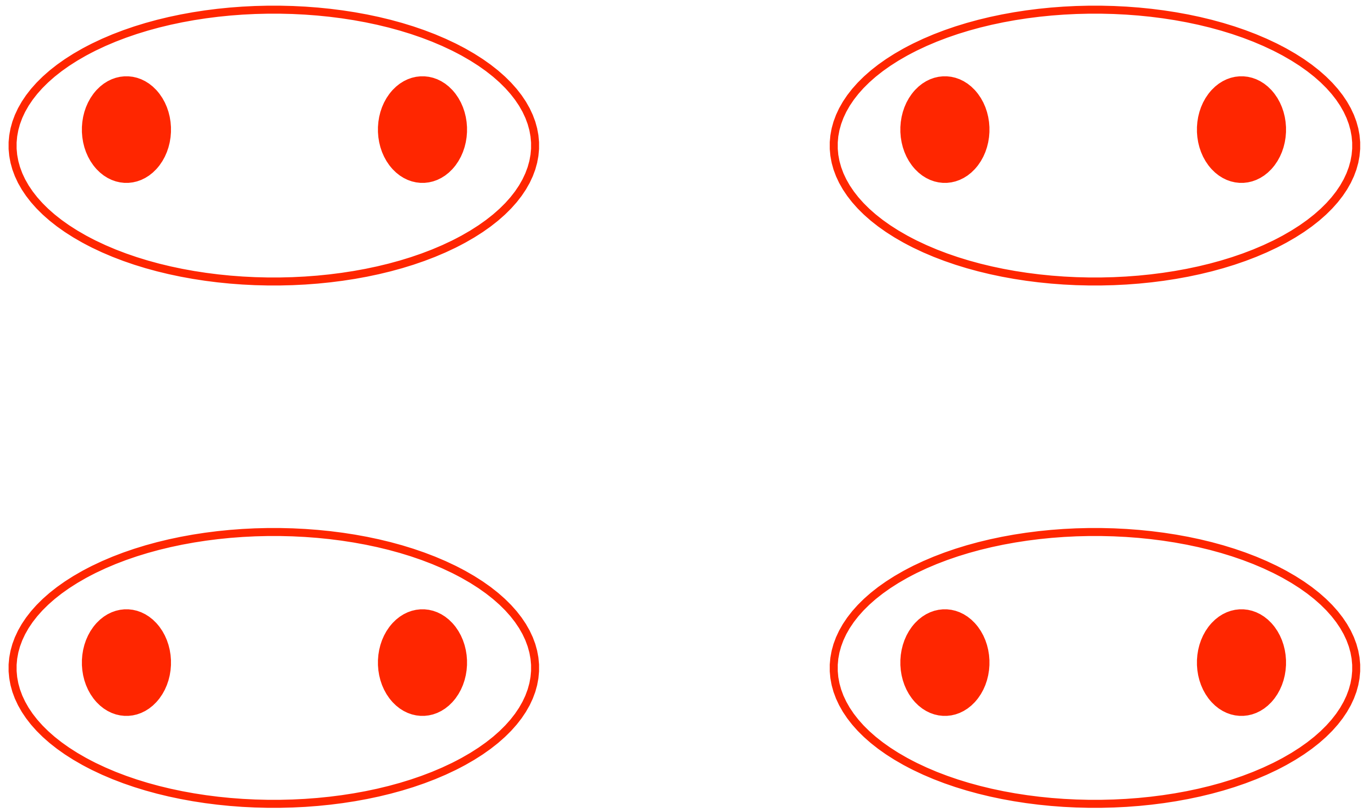
Fast, but tend to get long, stringy, meandering clusters

example from Manning & Schütze

Bottom-Up Clustering – Complete-Link



Bottom-Up Clustering – Complete-Link

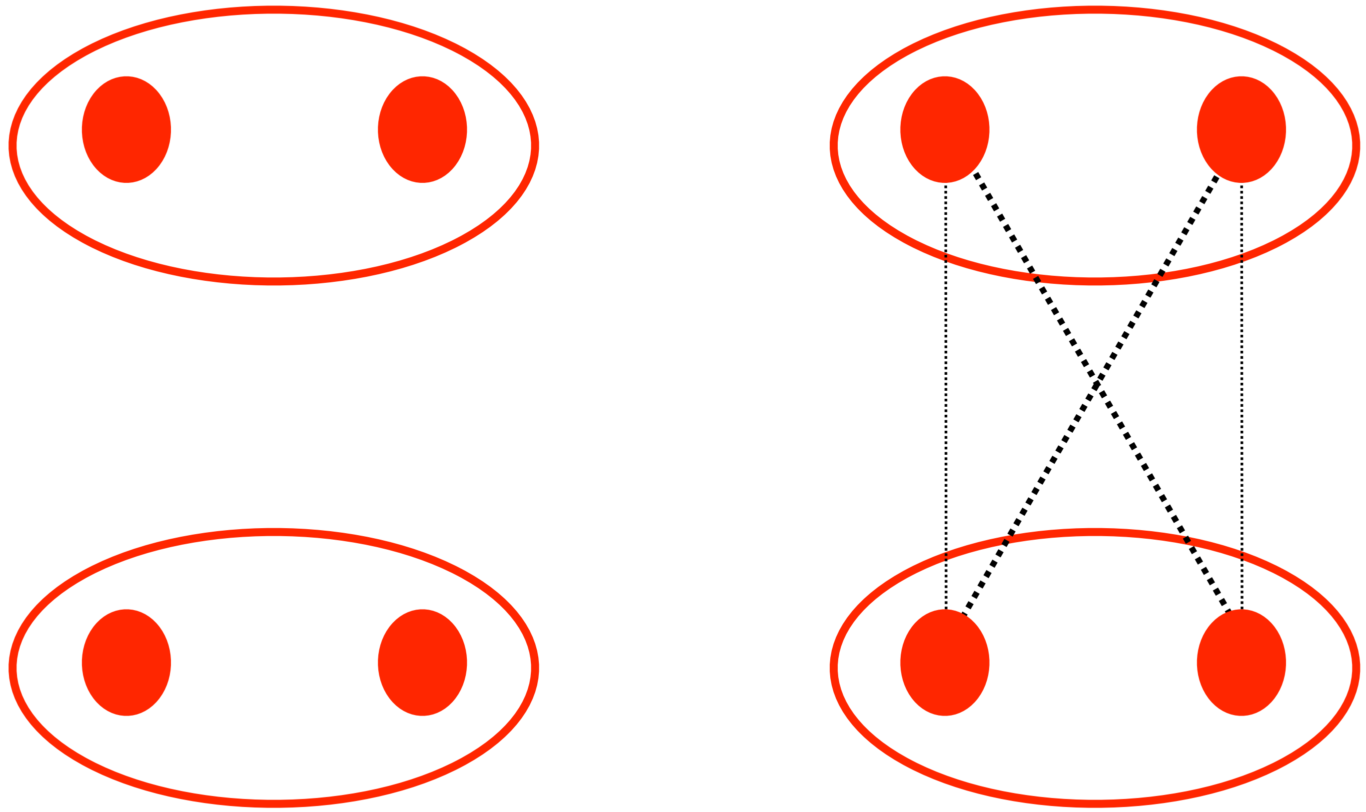


Again, merge closest pair of clusters:

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

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

Bottom-Up Clustering – Complete-Link



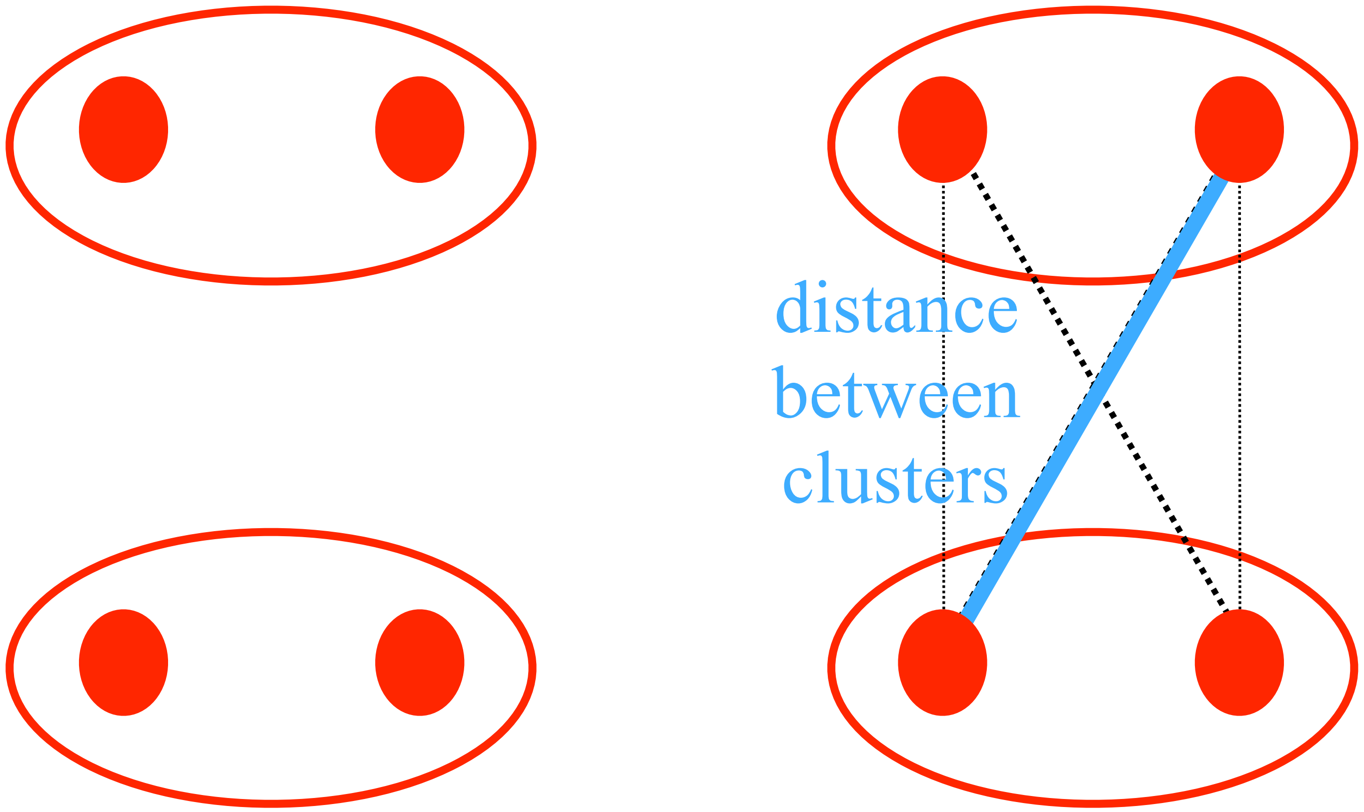
Again, merge closest pair of clusters:

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

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

example from Manning & Schütze

Bottom-Up Clustering – Complete-Link



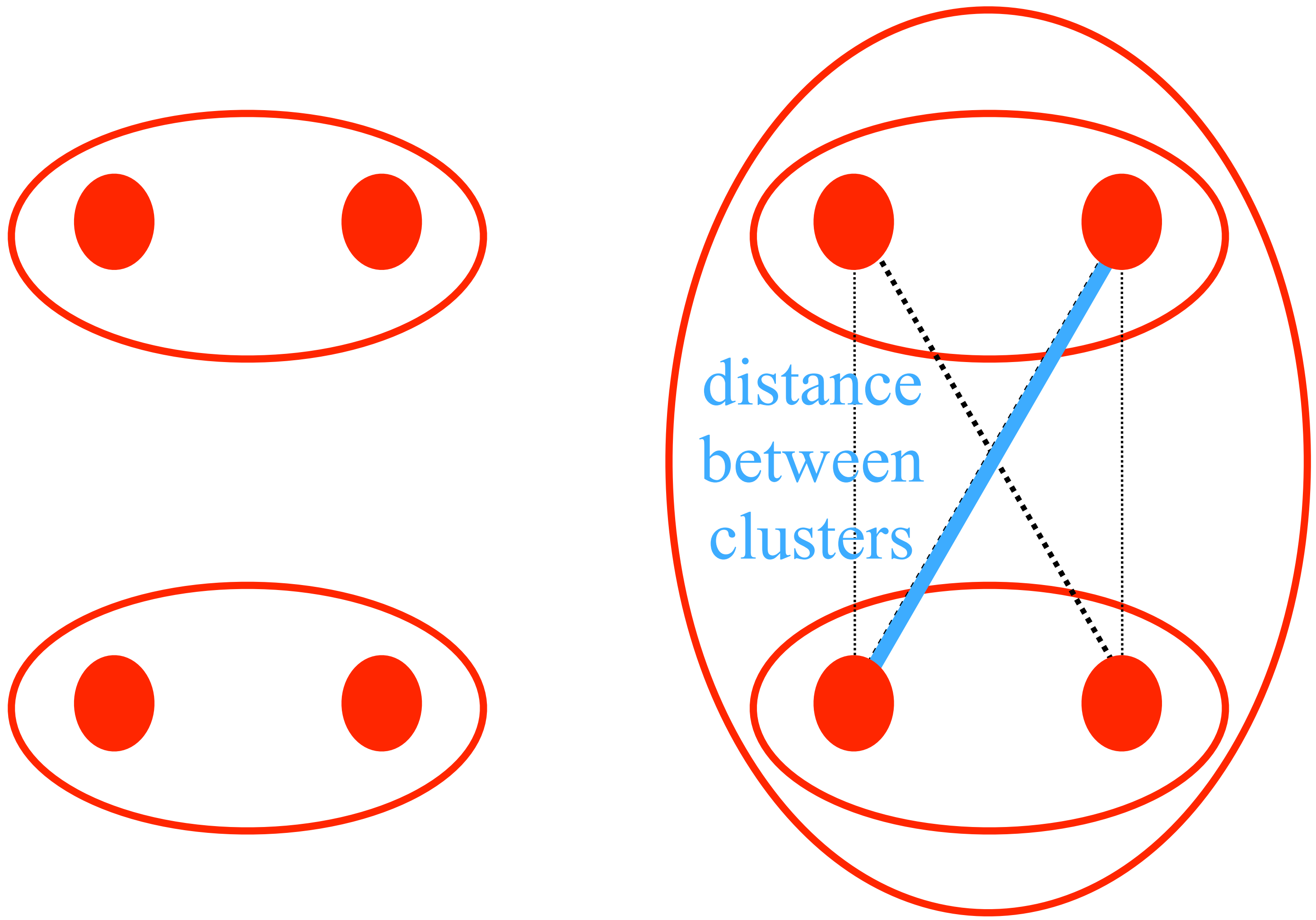
Again, merge closest pair of clusters:

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

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

example from Manning & Schütze

Bottom-Up Clustering – Complete-Link

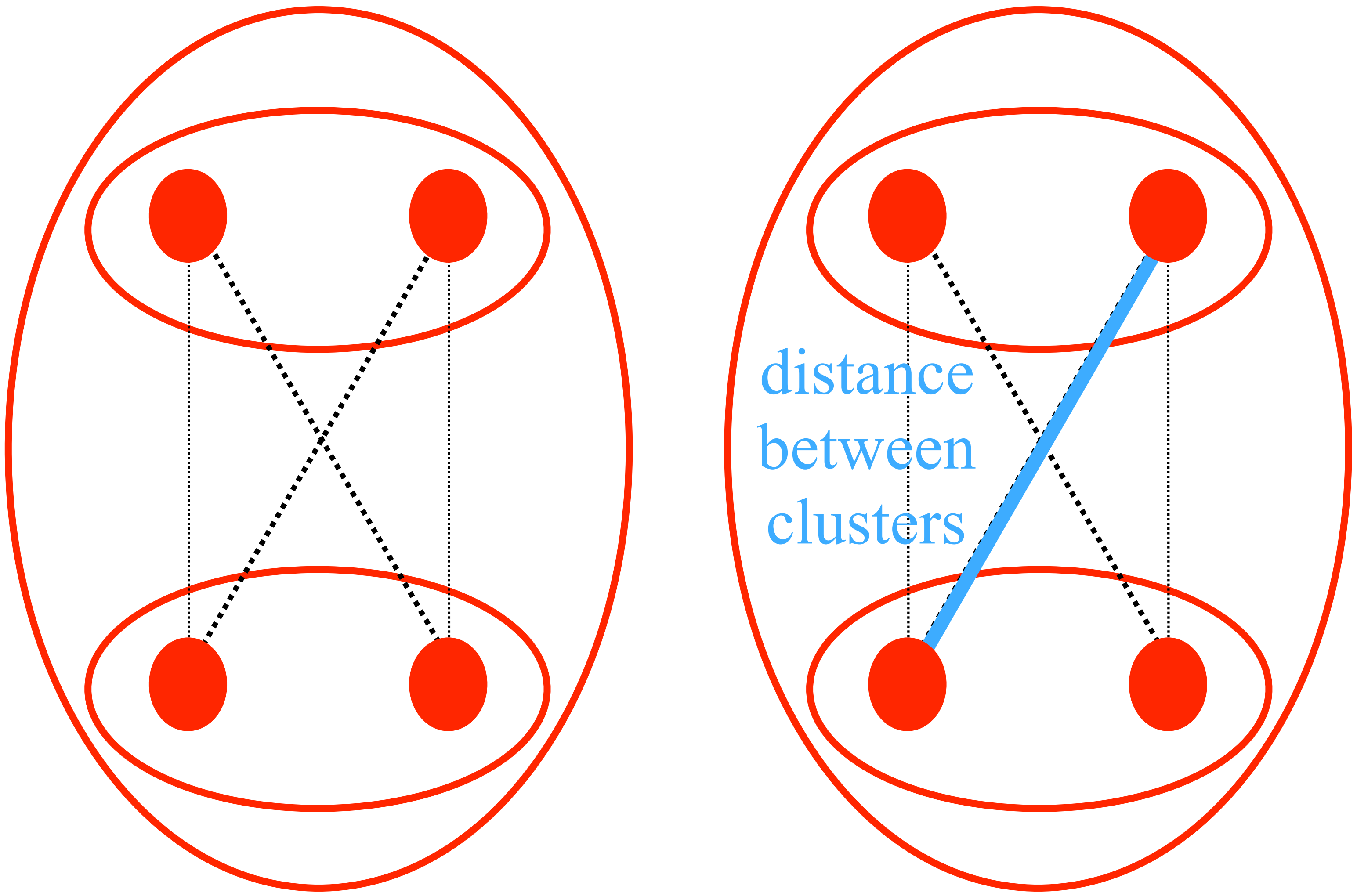


Again, merge closest pair of clusters:

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



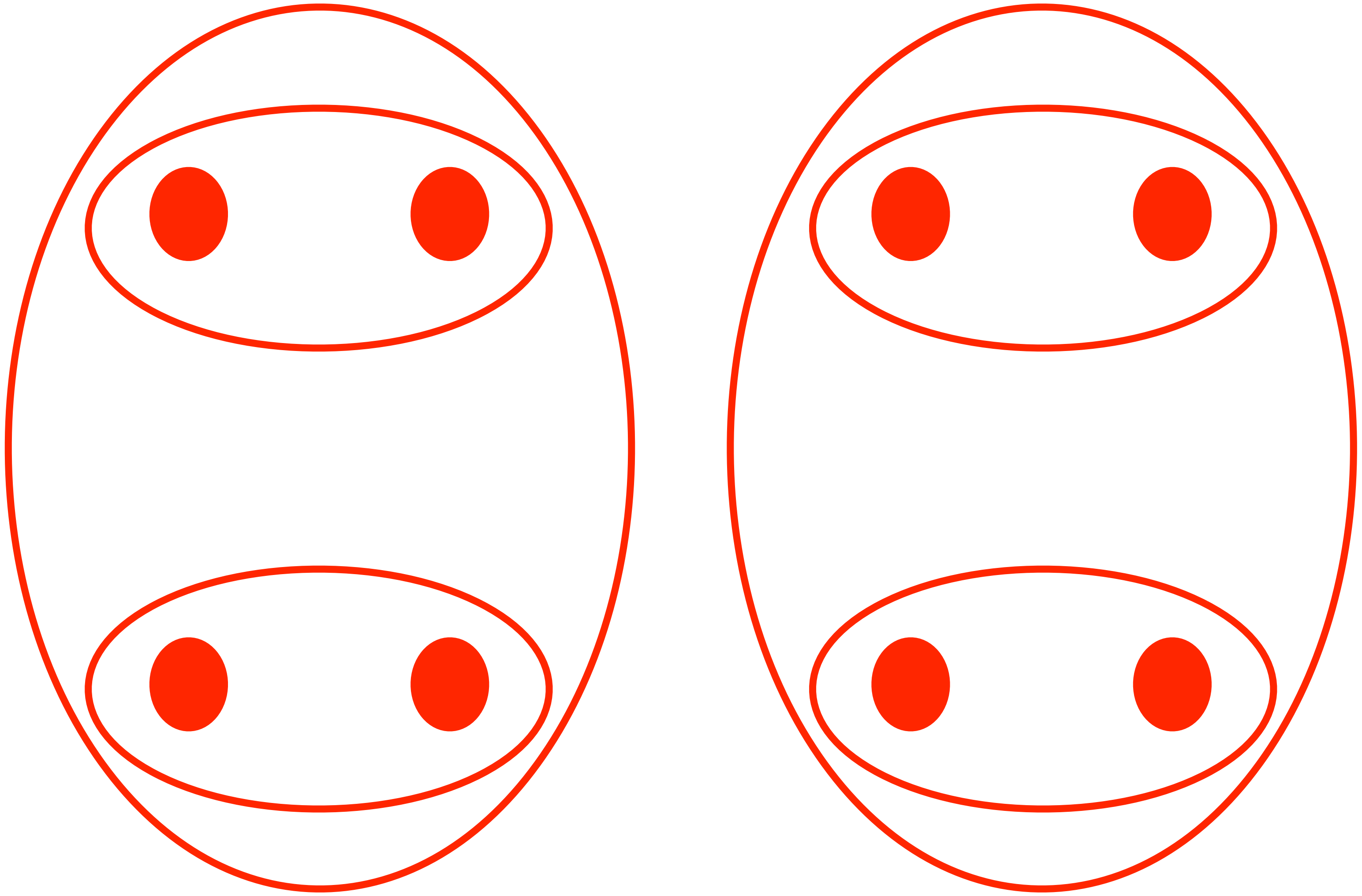
Again, merge closest pair of clusters:

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

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

example from Manning & Schütze

Bottom-Up Clustering – Complete-Link



Again, merge closest pair of clusters:

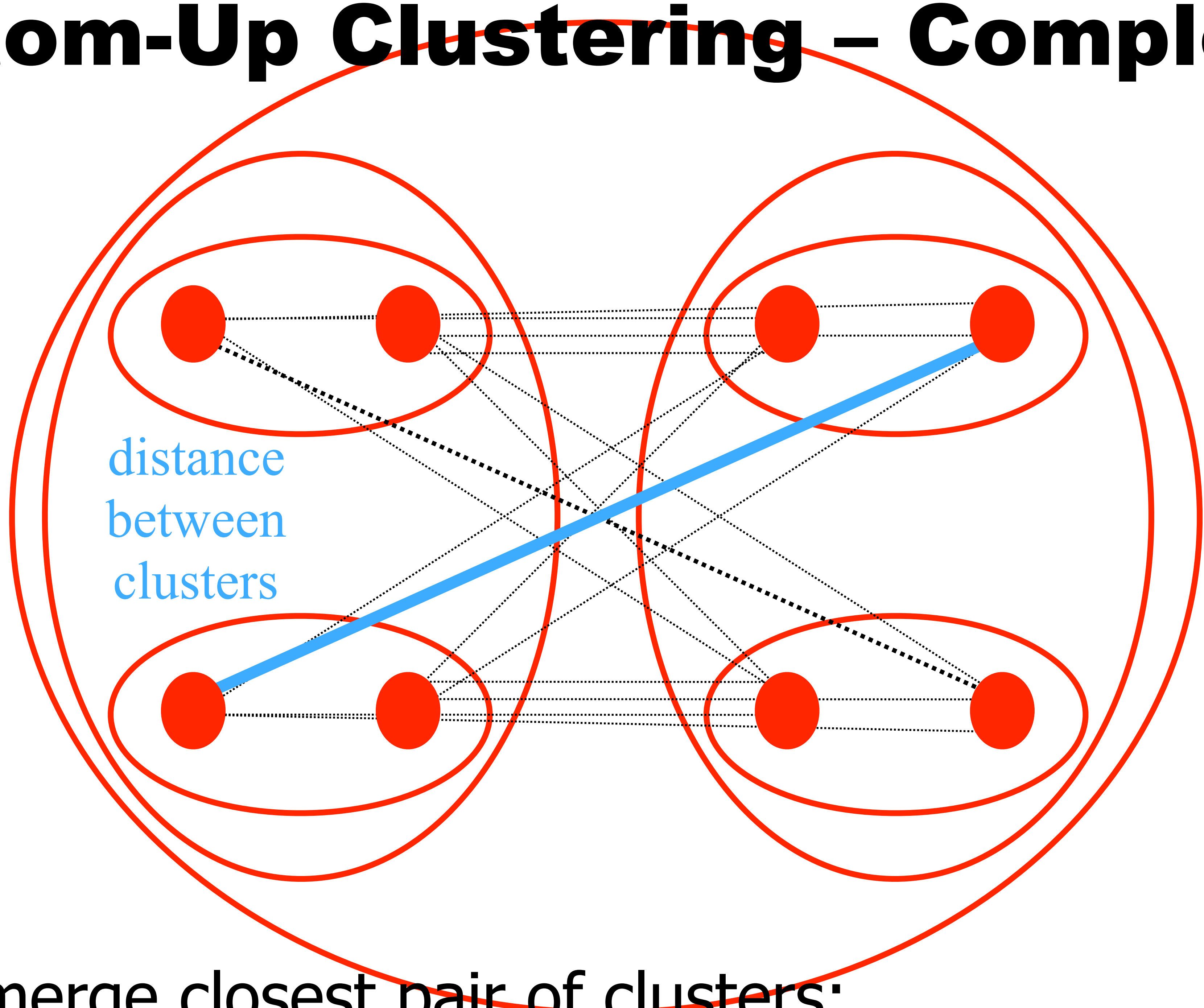
Complete-link: clusters are close only if all of their points are

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

Slow to find closest pair – need quadratically many distances

example from Manning & Schütze

Bottom-Up Clustering – Complete-Link



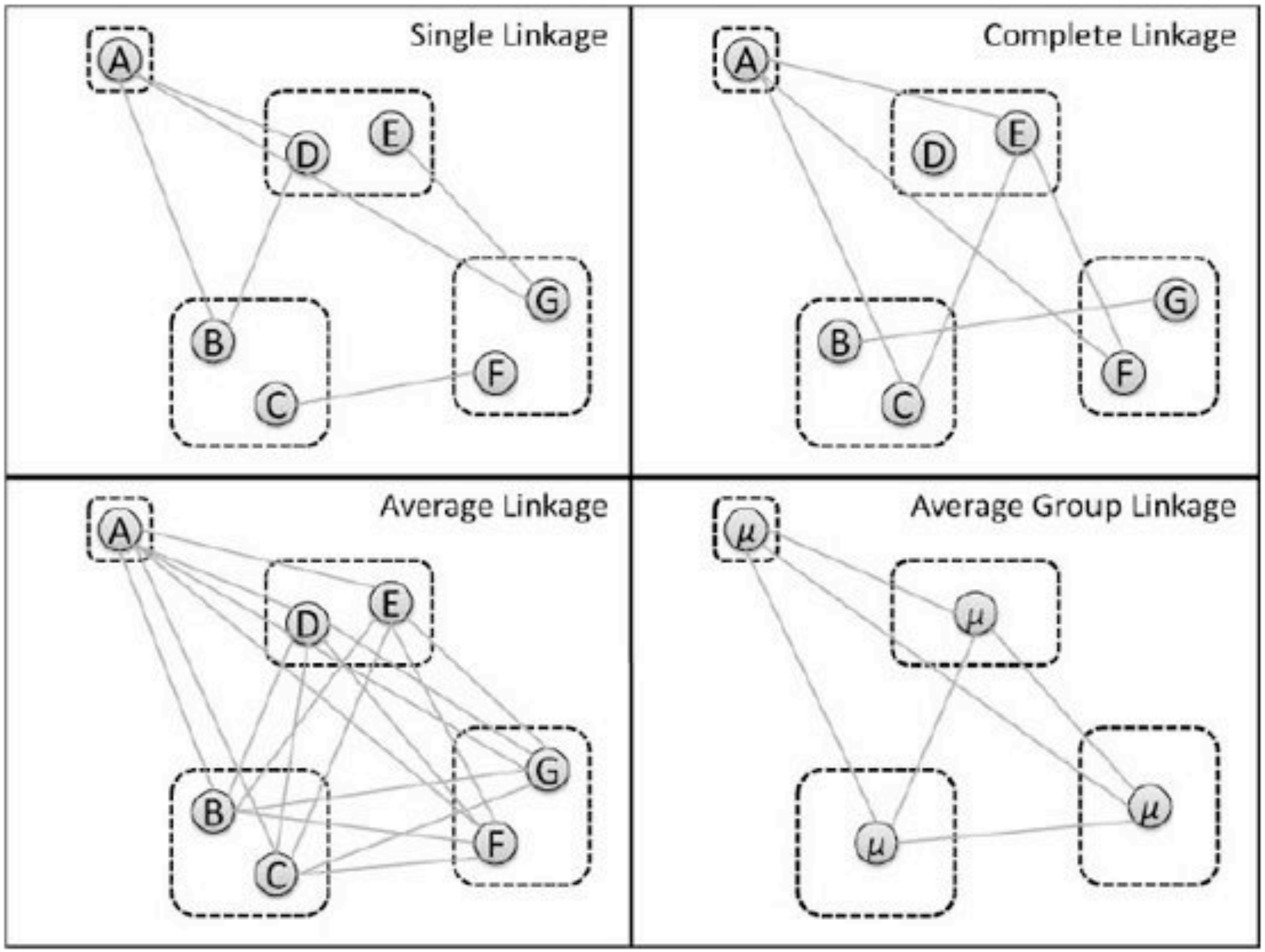
Again, merge closest pair of clusters:

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

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

Slow to find closest pair – need quadratically many distances

Bottom-Up Clustering Heuristics



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 ∃ 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 ∃ 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∈A, b∈B
  - Complete-link: dist(A,B) = max dist(a,b) for a∈A, b∈B
    - too slow to update cluster distances after each merge; but ∃ alternatives!
  - Average-link: dist(A,B) = mean dist(a,b) for a∈A, b∈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"

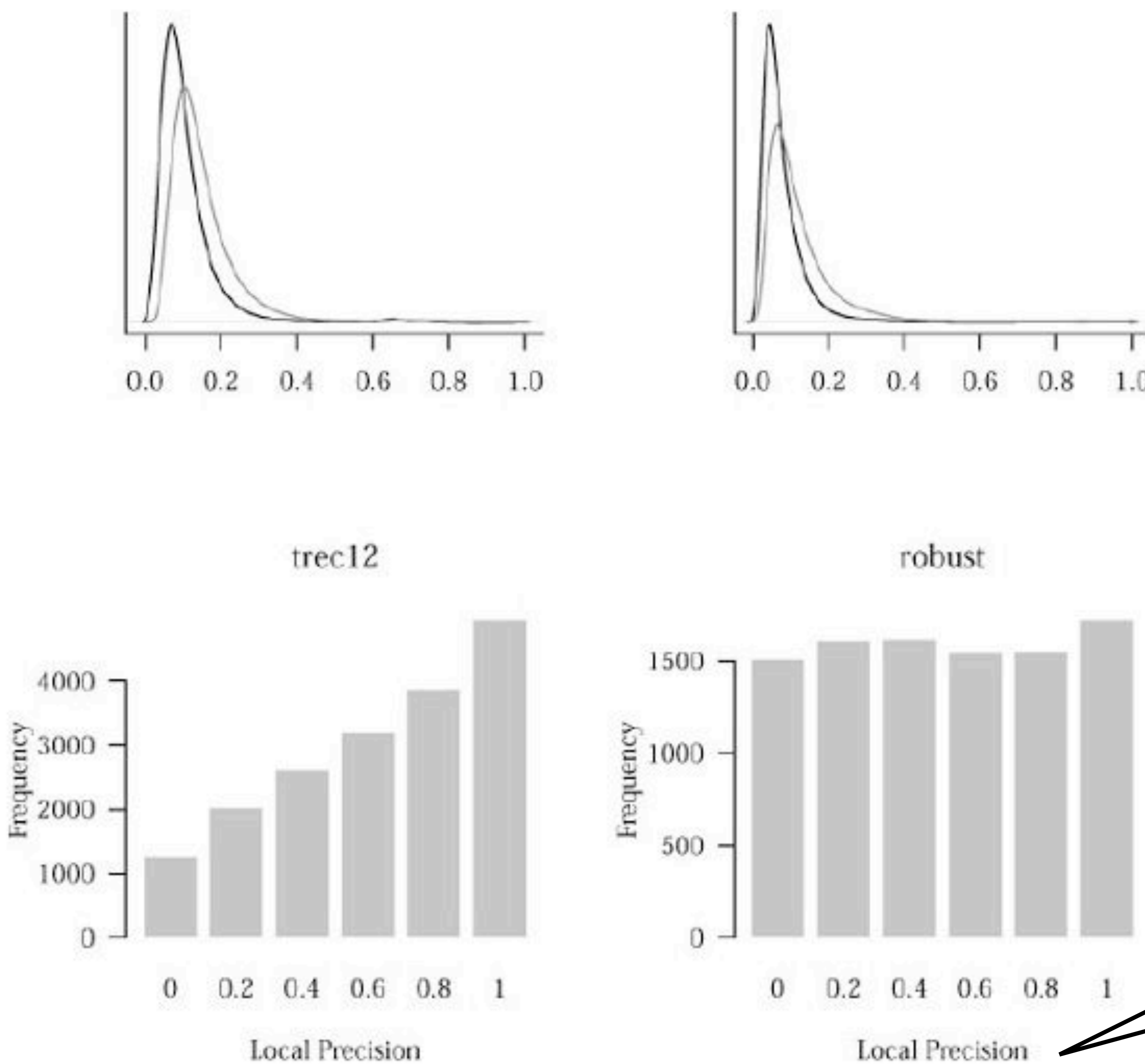
- **Expectation step:** Use current parameters (and observations) to reconstruct hidden structure
- **Maximization step:** Use that hidden structure (and observations) to reestimate parameters

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

Cluster Hypothesis

- Keith van Rijsbergen: "Closely associated documents tend to be relevant to the same requests."

Cluster Hypothesis



Precision in of the 5 nearest neighbors of relevant documents

But Does It Help Retrieval?

- Cluster retrieval

$$P(Q|C_j) = \prod_{i=1}^{n} P(q_i|C_j)$$

- Smoothing with hard clusters

$$P(w|D) = (1 - \lambda - \delta)\frac{f_{w,D}}{|D|} + \delta\frac{f_{w,C_j}}{|C_j|} + \lambda\frac{f_{w,Coll}}{|Coll|}$$

- Smoothing with soft clusters

$$P(w|D) = (1 - \lambda - \delta)\frac{f_{w,D}}{|D|} + \delta\sum_{C_j}\frac{f_{w,C_j}}{|C_j|}P(D|C_j) + \lambda\frac{f_{w,Coll}}{|Coll|}$$

- Last two more effective (cf. topic models)