# Clustering in Practice

Micha Elsner

April 4, 2013

# Unsupervised learning

Several reasons to do unsupervised learning:

- Data exploration: what is the structure of a large dataset?
- Downstream application: learn features to use for another task
- Classification but classes keep changing
- Cognitive models of learning

# A biased survey of the literature

Models using unsupervised learning:

- Some classic “clustering”
- Some using more structured models
- Not all using EM to learn
  - We’ll focusing on applications, not learning
  - In several cases, the advanced technique being used is “hierarchical Bayesian non-parametrics”
  - I know a reasonable amount about this, but it’s beyond a single lecture
- Organized by motivation

# Exploratory analyses

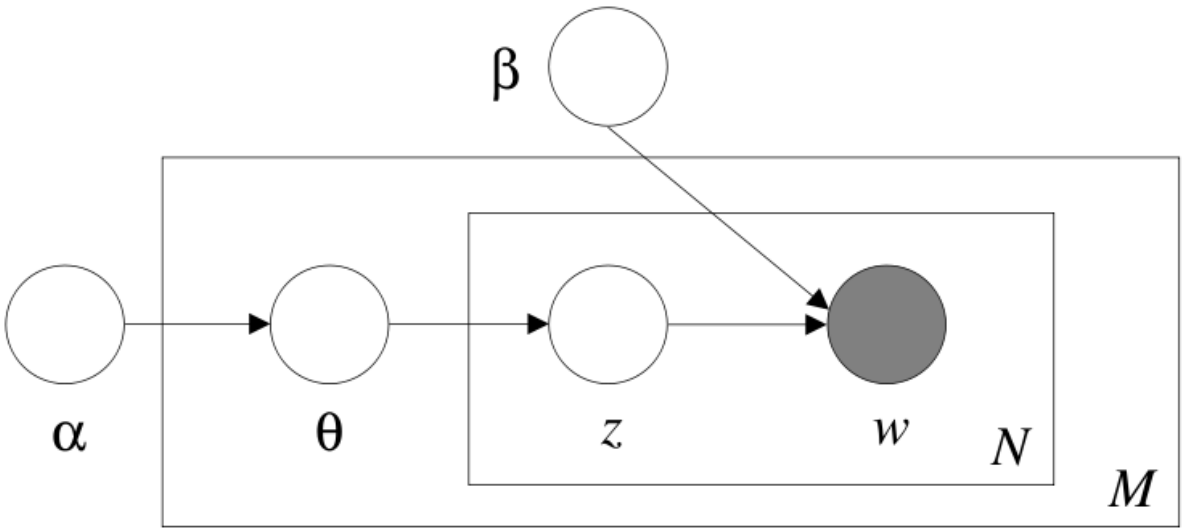
You have a lot of data... and you want to know what's going on in it (in some generic way)

- Can talk about *what instances are similar*
- Or *what features co-occur*

# Topic models

Latent Dirichlet allocation: Blei, Ng, Jordan

Wildly popular: Google Scholar lists 5692 citations




- Corpus of $M$ documents each with words $w^j_{1:N}$
- Want to cluster words based on document in which they occur
  - Word clusters are “topics” of related words
- Each word has cluster membership $z$
- Document-specific prior over $z$ denoted $\theta$
- Global prior over $\theta$ denoted $\alpha$

# Example output

(Original LDA paper)

| “arts” | “budget” | “children” | “education” |
|---|---|---|---|
| new | million | children | school |
| film | tax | women | students |
| show | program | people | schools |
| music | budget | child | education |
| movie | billion | years | teachers |

# Topic models for digital humanities

Robert Nelson “Mining the Dispatch”

`http://dsl.richmond.edu/dispatch/pages/intro`

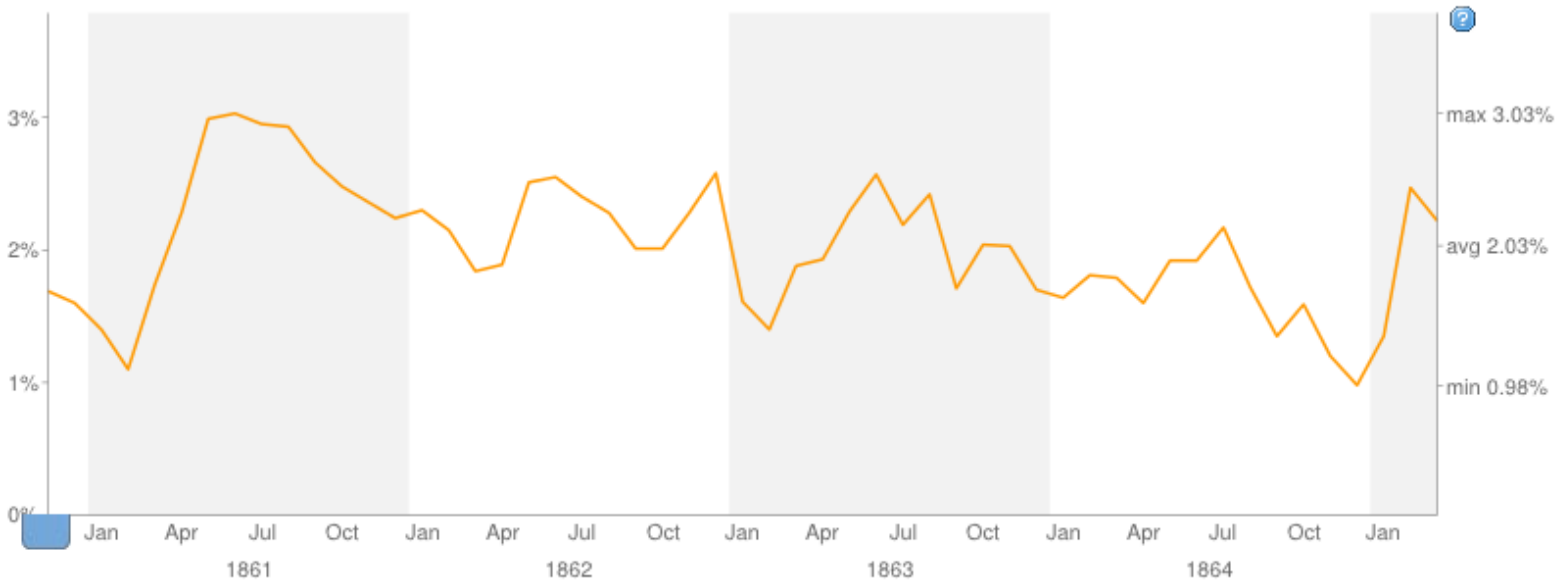
- Extended analysis of pre-Civil War Richmond Dispatch

## “Anti-Northern diatribes”

WAR PEOPLE SOUTH MEN SOUTHERN MAN NORTH WORLD YANKEE NORTHERN LINCOLN YANKEES COUNTRY BLOOD NATION HANDS MAKE TRUE HUMAN HISTORY POWER ENEMIES RACE HUMANITY

The brutal tyranny exercised by the Lincolnites in Maryland, Missouri, and wherever they have been able to obtain a foothold; the atrocities equalled only by those of the Sepays which their soldiery have committed, and the unheard of deed of despotism and darkness with which they threaten the . .

# Over time




- Model doesn't have inherent temporal structure
- Can still detect some temporal patterns
- More sophisticated topic models can explicitly model time in various ways

# Unsupervised modeling of Twitter conversations

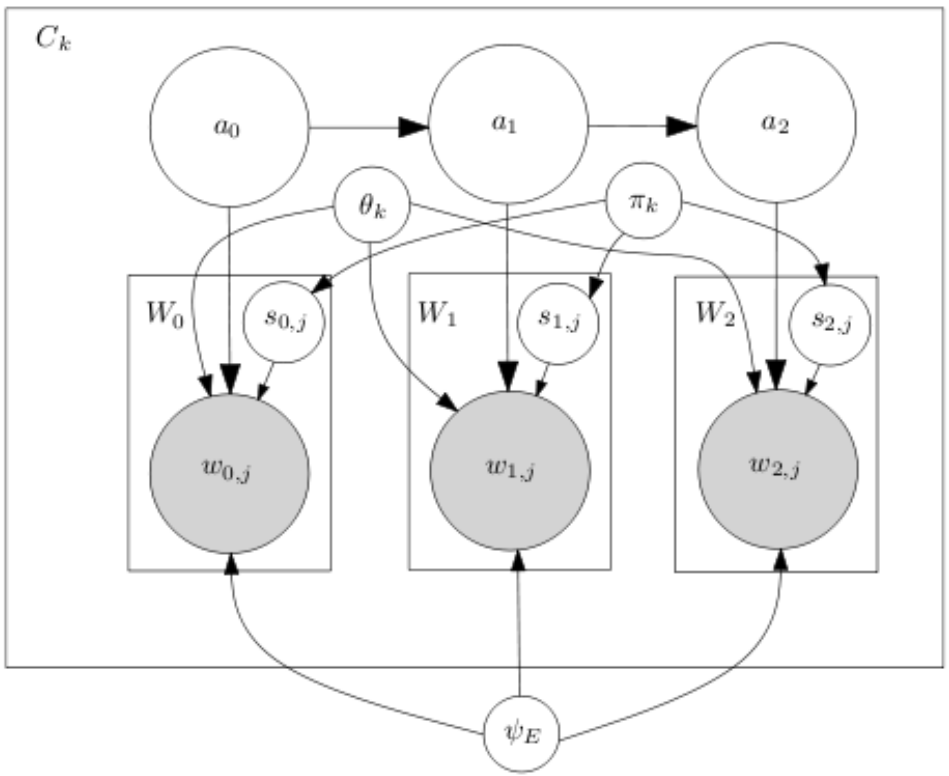
Ritter, Cherry and Dolan 2010

We've discussed dialogue acts briefly:

- Things like "statement", "question", "reply"
- Used in tutorial systems, generic-you classifier...
- Usually supervised

Are dialogues on Twitter similar to other dialogues?

Clustering tweets with complicated structured mixture model:



# Model

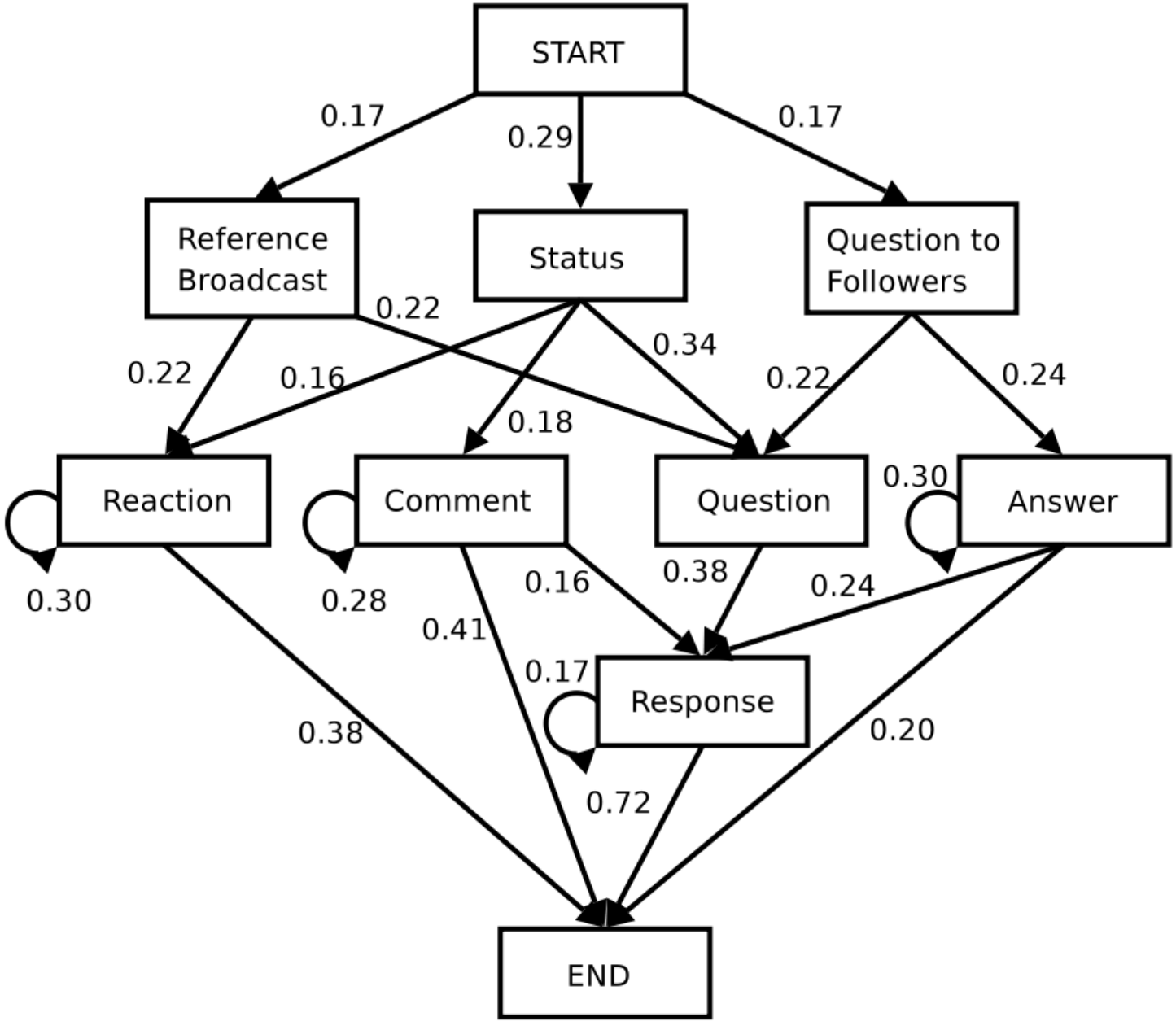
"Each conversation *C* is a sequence of acts *a* and each act produces a post, represented by a bag of words show [as *W*]... Starting with a random assignment of acts, we train our conversation model using EM, with forward-backward providing act distributions during the E-step"

- But this model could discover overall topics
  - Talking about food vs sports
- Instead of acts (questions and answers)

So add more structure:
"Each word ... is generated from one of three **sources**: the current post's dialogue act; the conversation topic; general English"

- New latent variable for each word indicates source

# Dialogue structure



# Downstream application

There's a task you really care about

- Information extraction, coreference, dialogue structure

You wish you had a

- tagger, parser, dictionary, etc

Build one using unsupervised learning!

# Parsing

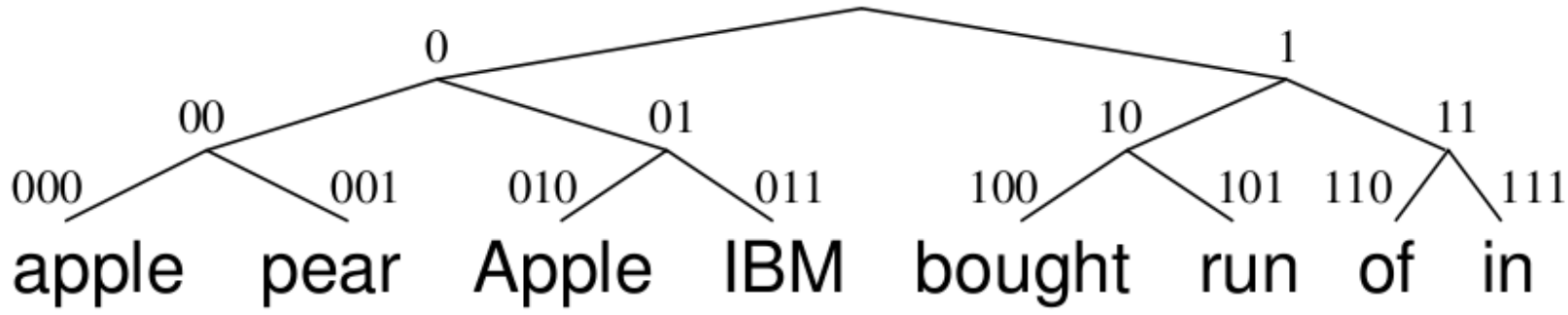
Koo, Carreras and Collins “Simple semi-supervised dependency parsing” 2008

- What we really care about: dependency parsing
- What we wish we had: a dictionary/thesaurus
  - So we could deal with unknown words— “cassowary” is like “chicken”
- What we’ll do:
  - Use a simple clustering algorithm to build word clusters
  - Use cluster of each word as feature in our dependency parsing model

# Clustering words

(Brown et al 1992)

- Not EM-based
- Build bigram LM over the corpus
- For every two words— if merged together, how much is LM likelihood reduced?
  - Merge the pair which reduces LL least
- Extract flat clustering from tree



# Dump this into a dependency parser

Previous arc prediction accuracy: 90.32
With clusters: 91.24
Various other results on Czech, different base parsers, etc...

# Where are you from?

A Latent Variable Model for Geographic Lexical Variation: Eisenstein, O'Connor, Smith and Xing 2009

- Watch someone on Twitter
- What we care about: where they come from
  - Application? Maybe we want drones to blow up their house?
- What we wish we had: a slang atlas
- Also exploratory: "what do people from California sound like?"

# The model

Training data: geotagged tweets
A cross between topic models and mixture of Gaussians:

- As in mixture of Gaussians, probability of being part of cluster $C$ depends on distance from your address $d$ to center $C_i$
- But then!
- The words in your tweet depend on word clusters $z$ (eg "Californian sports")
- The prior over $z$ depends on your regional cluster $C_i$
- It's actually a little more complicated than this
  - Clusters of clusters: "Californian sports" is part of "sports"
- But never mind

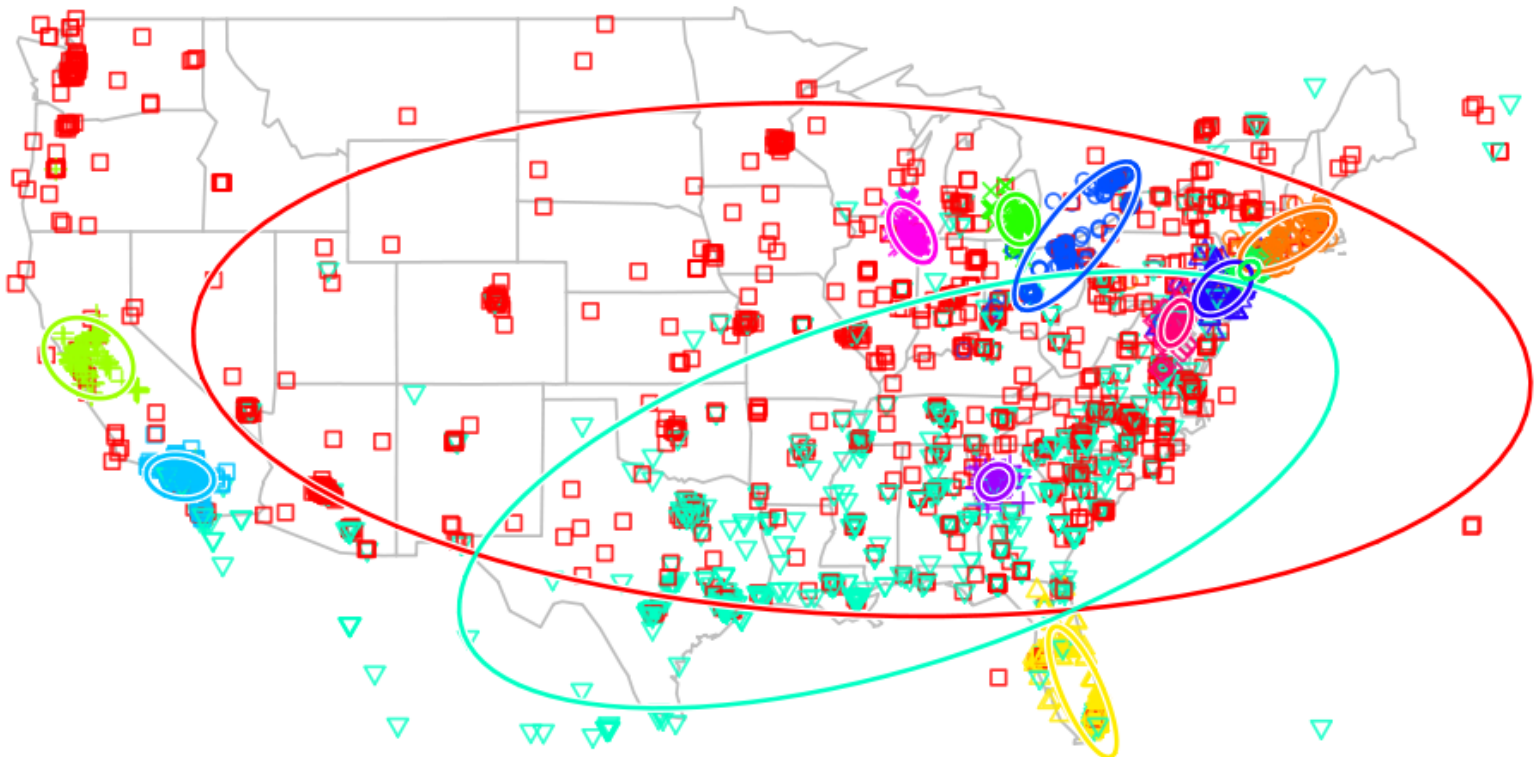
# Results

- Can we bomb your house?
- Median dist of 494 km between predicted and true location
- Center of the Twitter-verse is dist of 1018 km from median tweeter

What do people sound like?

| Main topic | "basketball" | "chatter" |
|---|---|---|
| Boston | Celtics, Boston, Charlotte | exam, suttin, sipping |
| No.Cal. | Thunder, Kings, Giants | hella, flirt, iono, Oakland |
| NYC | Nets, Knicks | wassup, nm |
| Lake Erie | Cavs, Cleveland, Ohio, Bucks, Columbus | foul, Wiz, salty, excuses, lames, officer, lastnight |

# Twitterverse dialectology



# Classification, but classes change

There is a ground truth labeling you want to recover

- You'd use classification
- But the set of classes for each instance is different

We discussed coreference as an application

# Disentangling chat

Elsner and Charniak (ACL 2011, 2010, 2008), Elsner and Schudy (2009)
(Sorry... but the slides are really easy to get!)

# A crowded chat room

- Shared communications channel
- Utterances appear one by one

**Chanel:** How do I limit my internet connection speed?
**Felicia:** Use the keyword “throttling” in google.
**Chanel:** Felicia, google solved my problem.
**Gale:** You guys have never worked in a factory, have you?
**Gale:** There’s some real unethical stuff that goes on.
**Arlie:** Of course, that’s how they make money.
**Chanel:** You deserve a trophy!
**Gale:** People lose limbs, or get killed.
**Felicia:** Excellent!

# Participants know the structure is this:

**Chanel:** How do I limit my internet connection speed?
**Felicia:** Use the keyword “throttling” in google.
**Chanel:** Felicia, google solved my problem.
**Gale:** You guys have never worked in a factory...
**Gale:** There’s some real unethical stuff that goes on.
**Arlie:** Of course, that’s how they make money.
**Chanel:** You deserve a trophy!
**Gale:** People lose limbs, or get killed.
**Felicia:** Excellent!

# Not this:

**Chanel:** How do I limit my internet connection speed?
**Felicia:** Use the keyword “throttling” in google.
**Chanel:** Felicia, google solved my problem.
**Gale:** You guys have never worked in a factory...
**Gale:** There’s some real unethical stuff that goes on.
**Arlie:** Of course, that’s how they make money.
**Chanel:** You deserve a trophy!
**Gale:** People lose limbs, or get killed.
**Felicia:** Excellent!

# Especially not this!

**Chanel:** How do I limit my internet connection speed?

**Felicia:** Use the keyword “throttling” in google.

**Chanel:** Felicia, google solved my problem.

**Gale:** You guys have never worked in a factory...

**Gale:** There’s some real unethical stuff that goes on.

**Arlie:** Of course, that’s how they make money.

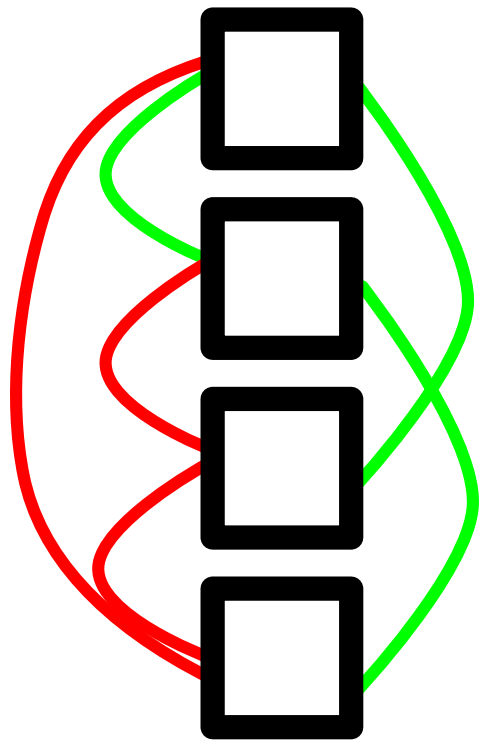
**Chanel:** You deserve a trophy!

**Gale:** People lose limbs, or get killed.

**Felicia:** Excellent!

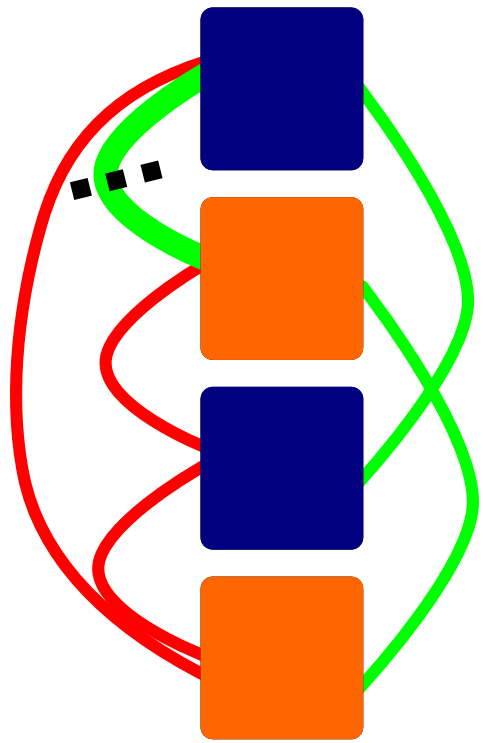
# Correlation clustering framework

- Classify each utterance pair “same thread” or “different”
- Partition to keep “same” utterances together and split “different” ones apart



# Correlation clustering framework

- Classify each utterance pair “same thread” or “different”
- Partition to keep “same” utterances together and split “different” ones apart



# Classifying pairs

Pair of utterances: same conversation or different?

## Chat-specific features (F 66%)

- Same speaker
- Time between utterances
- Speaker's name mentioned:

  **Sara:** Can you extract files from a patch?

  **Carly:** Sara, no

# Classifying pairs

Pair of utterances: same conversation or different?

## Dialogue features (F 58%)

- Question/answer
  - "...?" "Yes, ..."
- Greetings, goodbyes, thanks

## Word overlap (F 56%)

- Repeated terms
  - "release the lions!" "...opens the lions' gate"

# Classifying pairs

Pair of utterances: same conversation or different?

## Dialogue features (F 58%)

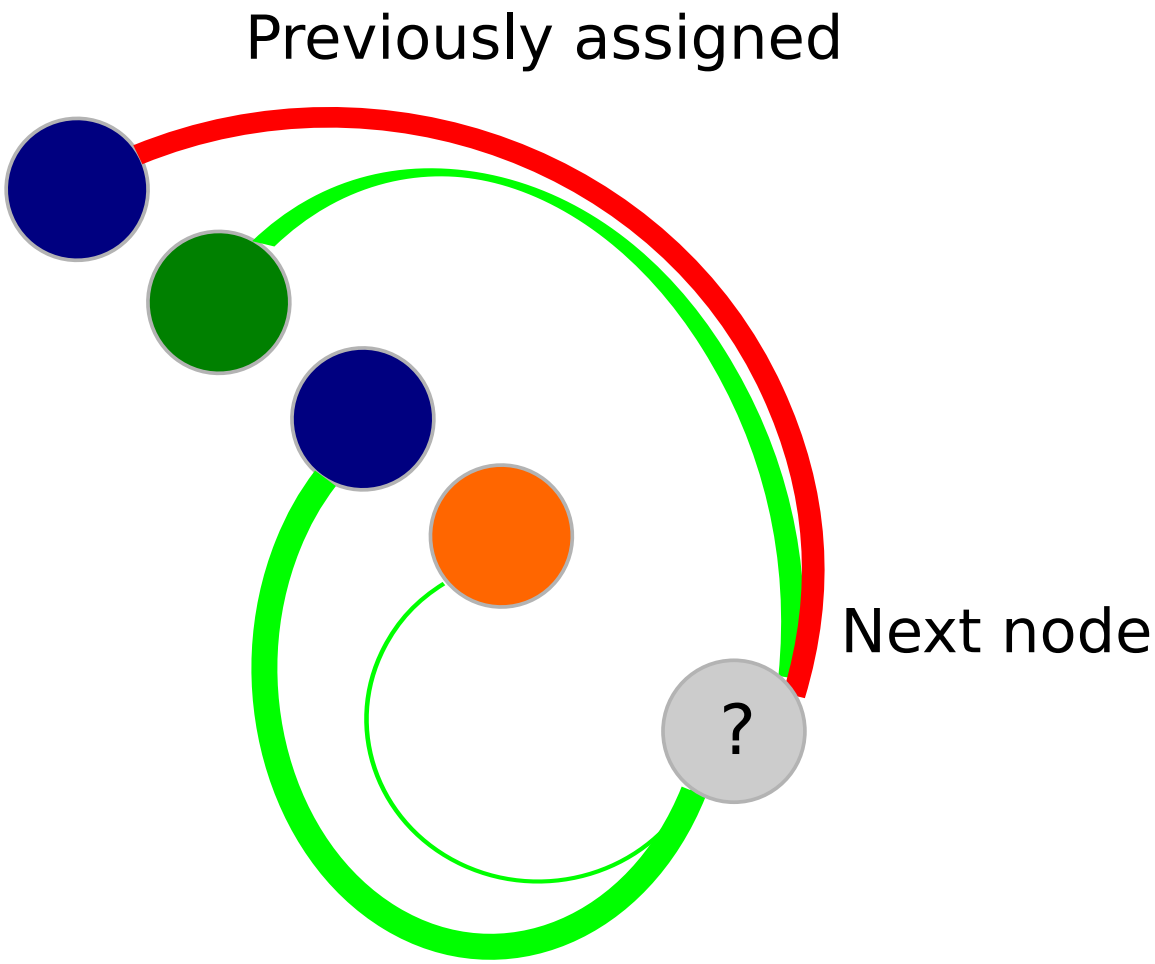
- Question/answer
  - "...?" "Yes, ..."
- Greetings, goodbyes, thanks

## Word overlap (F 56%)

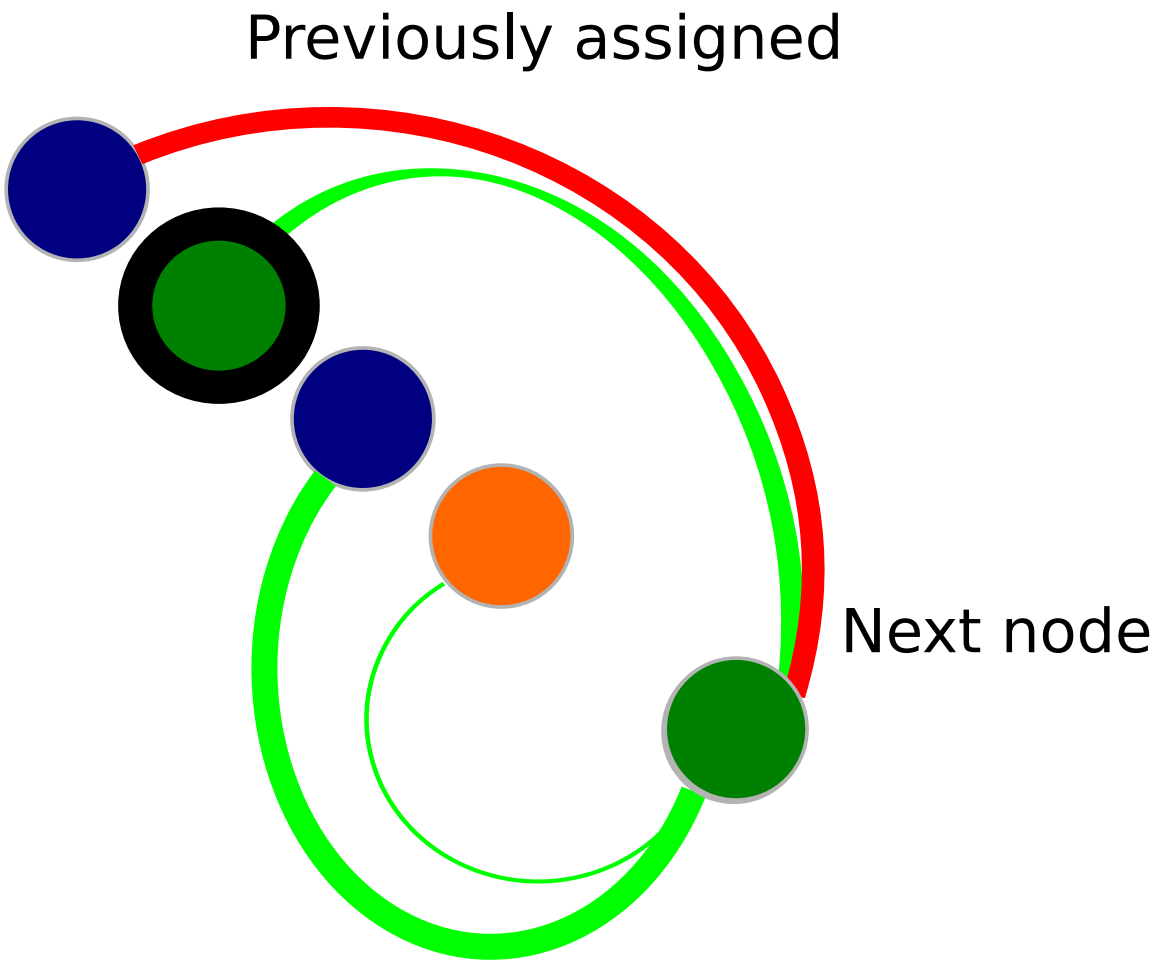
- Repeated terms
  - "release the lions!" "...opens the lions' gate"

## All combined (F 71%)

# Greedy algorithm:

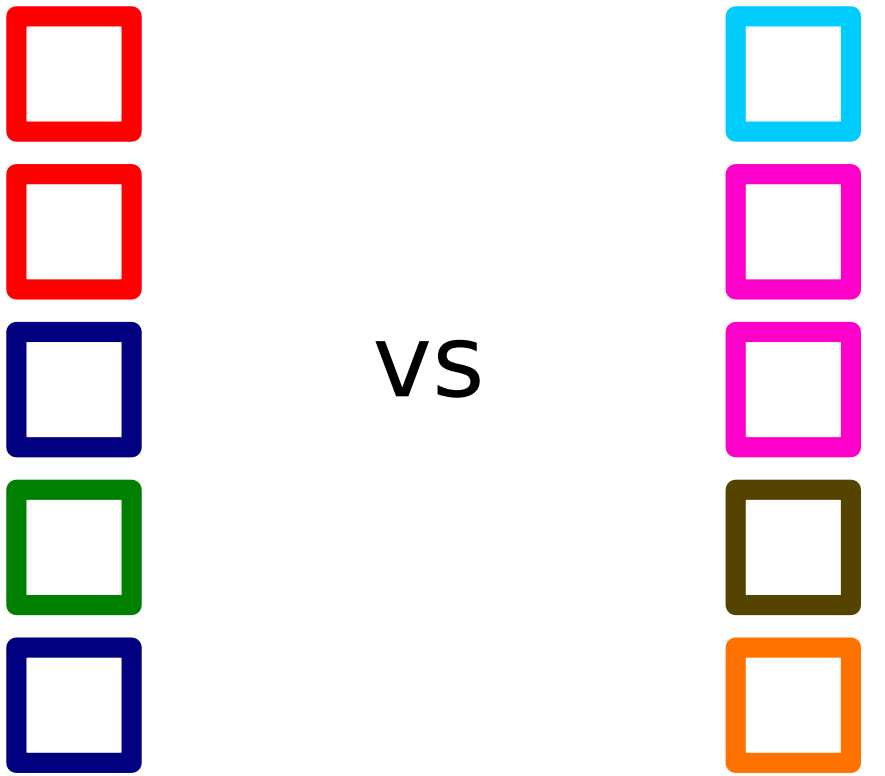


# Greedy algorithm:



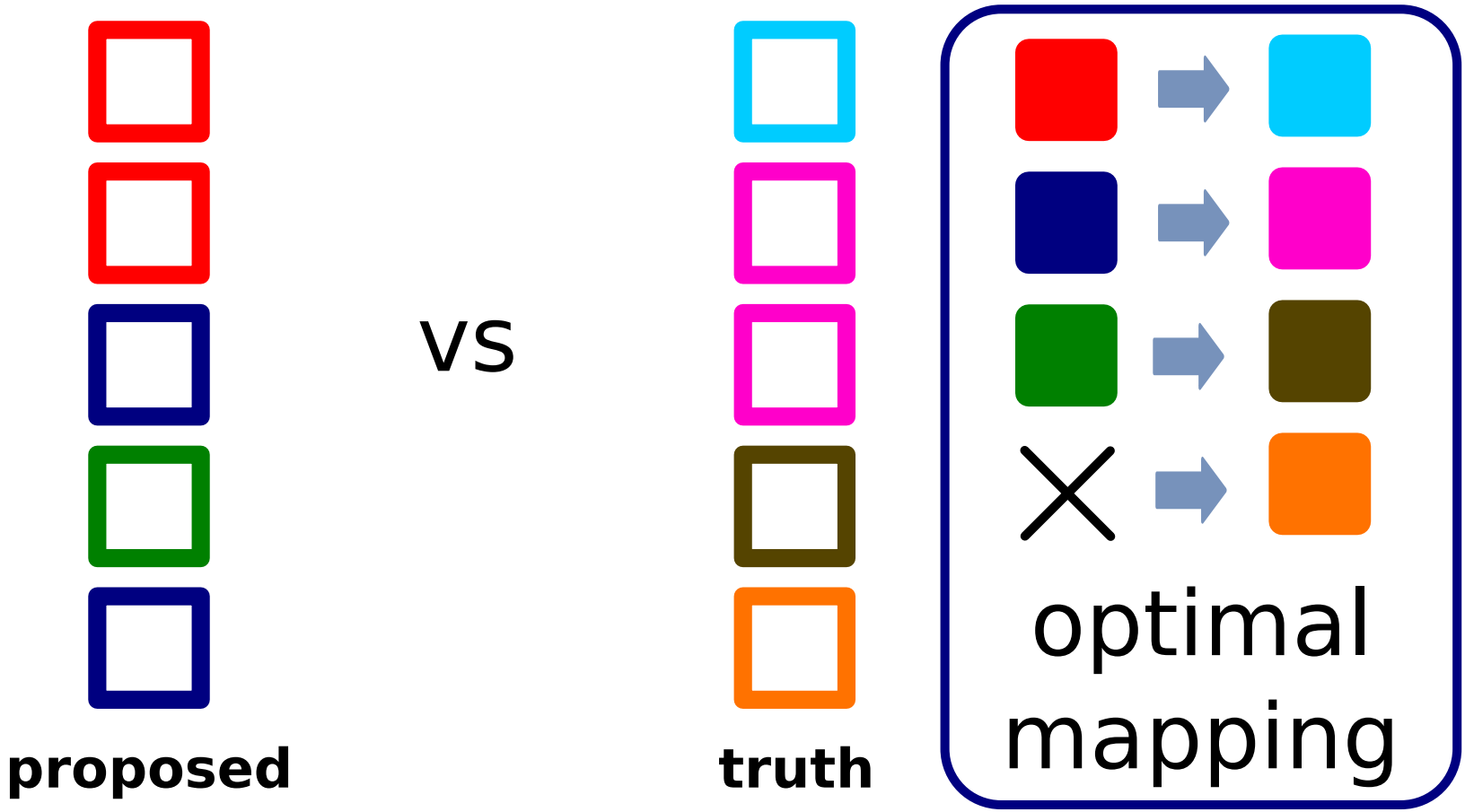

the cluster with **highest arc sum**

# One-to-one overlap

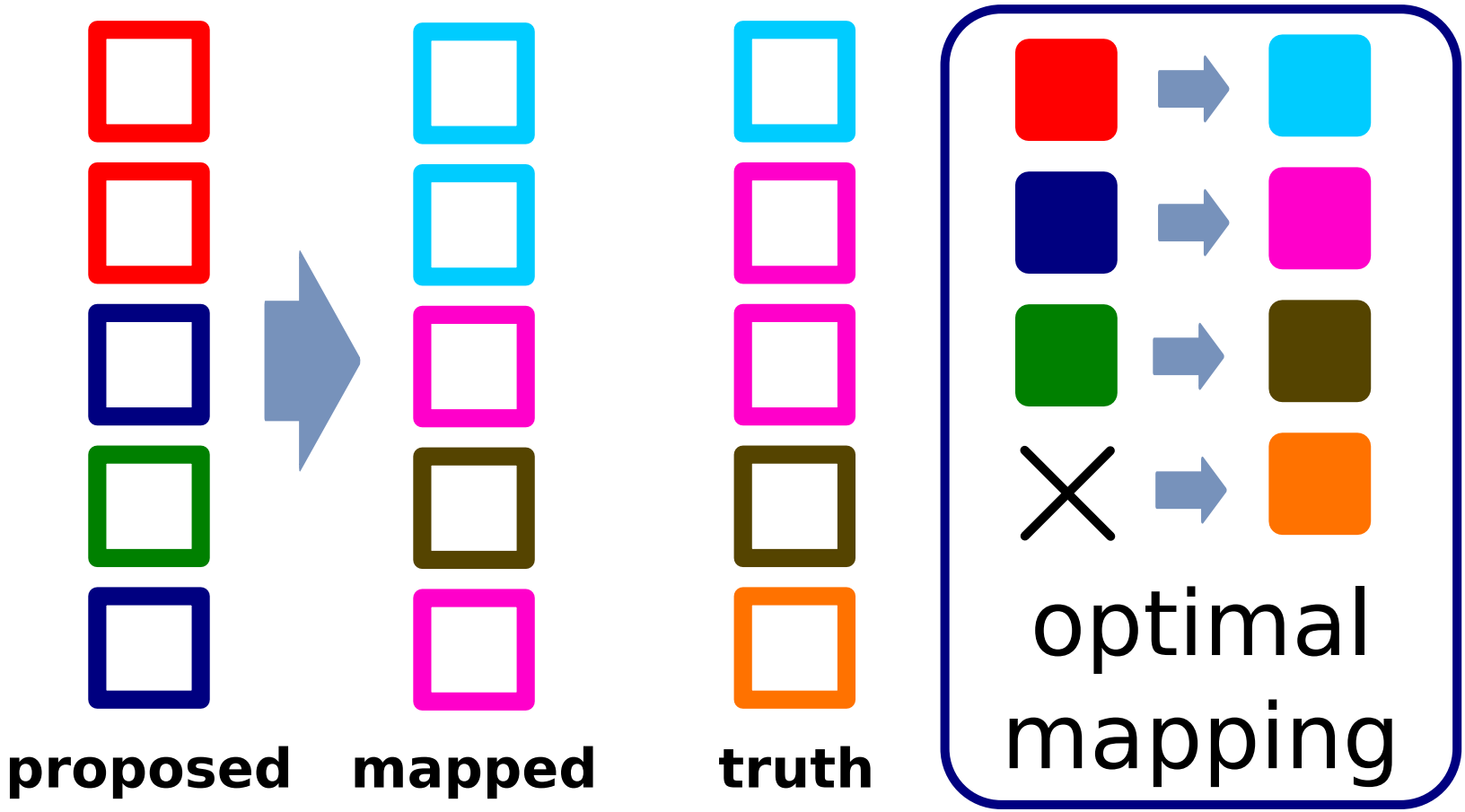



**proposed** **truth**

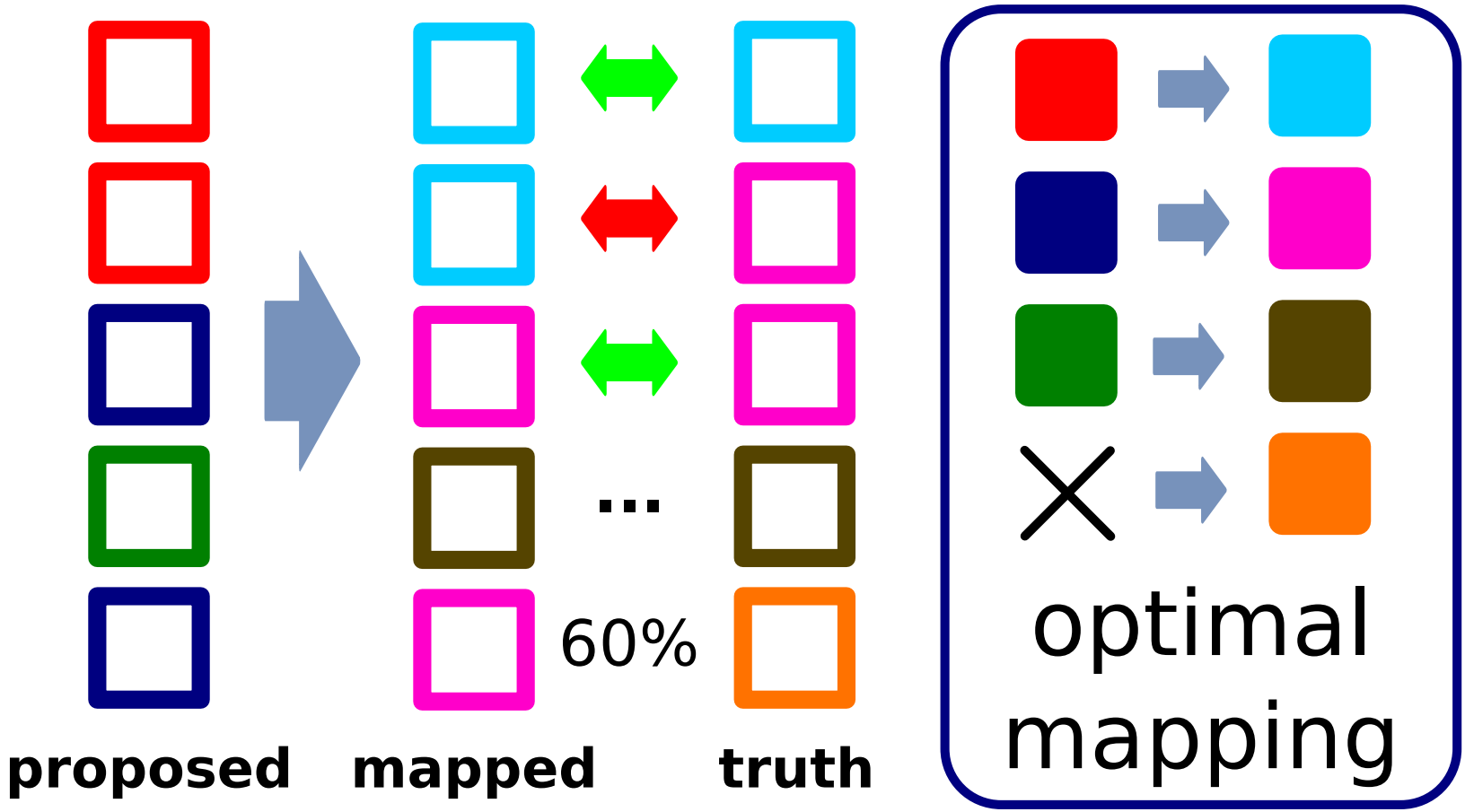
# One-to-one overlap



# One-to-one overlap



# One-to-one overlap



# Results

| | One-to-one |
|---|---|
| Annotators | 53 |

- Annotators don't always agree...
  - Some make finer distinctions than others

# Results

| | One-to-one |
|---|---|
| Annotators | 53 |
| Best Baseline | 35 (pause 35s) |

- Annotators don't always agree...
  - Some make finer distinctions than others
- Tested a variety of simple baselines

# Results

| | One-to-one |
|---|---|
| Annotators | 53 |
| Best Baseline | 35 (pause 35s) |
| Corr. Clustering | **46** |

- Annotators don't always agree...
  - Some make finer distinctions than others
- Tested a variety of simple baselines
- Model outperforms all baselines

# Word sense induction

"Discovering word senses from text" Pantel and Lin suit:

- Cluster 34: blouse, slack, legging, sweater
- Cluster 137: lawsuit, allegation, case, charge

plant:

- Cluster 215: plant, factory, facility, refinery
- Cluster 235: shrub, ground cover, perennial, bulb

# Features

" 'sip _' is a verb-object context. If the word *wine* occurred in this context, the context is a feature of *wine*."
Features weighted by mutual information (does "sip" strongly predict "wine" and vv?)
Use *cosine* distance between feature vectors:

$$|w_i, w_j|_{cos} = \frac{f(w_i) \cdot f(w_j)}{\sqrt{f(w_i)^2 \times f(w_j)^2}}$$

Not quite k-means:

- First find a set of "committees" (very tight clusters)
- For other words, find nearest committee
- Assign word a sense for that committee...
- Remove features explained by that sense

# Evaluation

Gold standard is WordNet manual dictionary

- To detect multiple senses with k-means...
- Check if a point is near two cluster centers in final clustering
- ...Detect two senses

| System | Prec | Rec | F-score |
|---|---|---|---|
| CBC (Pantel and Lin) | 60 | 50 | 55 |
| K-means | 48 | 44 | 46 |
| Average-link | 50 | 41 | 45 |

- CBC doesn't propose as many bogus senses
- And recovers more actual senses

# Cognitive models of learning

There is a ground truth labeling you want to recover

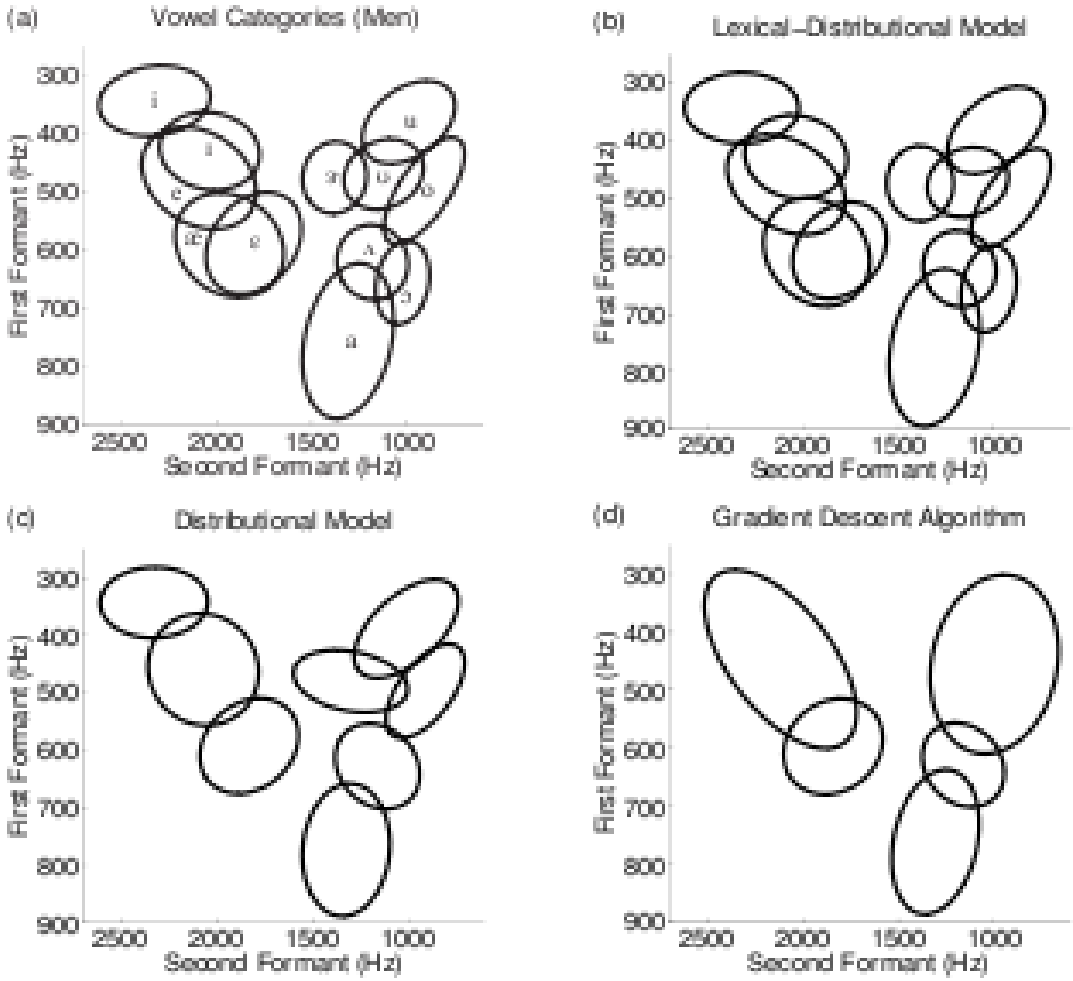
- But how do people learn it?
- They don't see labeled examples...
  - But they still get the right grammar

We discussed some models of grammar and POS tag learning

- Mirella talked about a model like this for categories

# Vowels

Vallabha, McClelland, Pons, Werker and Amano “Unsupervised learning of vowel categories from infant-directed speech”
Feldman, Griffiths and Morgan “Learning phonetic categories by learning a lexicon”



# Models

People often learn vowels with mixtures of Gaussians (as in Vallabha et al)

- Very much like our toy vowel models
- But the i/u problem is trivial
- The full problem is very hard
- Lots of category overlap, individual variability, etc
- Feldman: knowing the lexicon can help
  - "r[V]t" is a word: /V/ sounds like /a/ but not /e/: /a/ is a different vowel than /e/
- Feldman's model: mixture of Gaussians with word-specific prior over vowels

# Semantics

A Probabilistic Model of Syntactic and Semantic Acquisition from Child-Directed Utterances and their Meanings: Kwiatkowski, Goldwater, Zettlemoyer and Steedman
Problem setting:

## Data

Utterance: you have another cookie
Candidate meanings (in some kind of lambda-format):

- have(you, another(x, cookie(x)))
- eat(you, your(x, cake(x)))
- want(i, another(x, cookie(x)))

Learn the correct meaning for each sentence, semantics of each word, and syntactic parser

# Basic setup

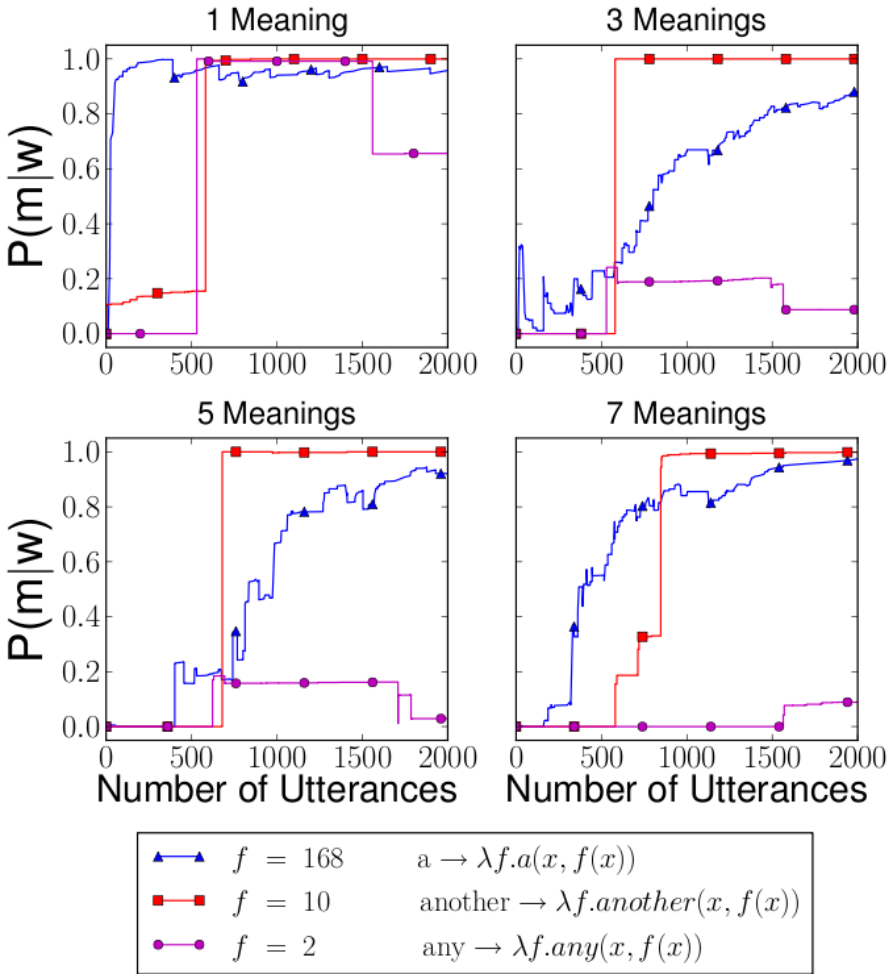
- CCG (combinatory categorial grammar) defines syntax
- CCG rules specify compositional semantics
- Main issue is lexical semantics (what does each word mean?)
  - And what is its logical type?

EM-like algorithm switches between:

- E-step: use current grammar to parse sentences
  - Involves computing an MT-like alignment between nodes in the parse tree and candidate meanings
  - Assign responsibilities for the meanings to parse tree nodes
  - ...add counts to tree/semantics events
- M-step: revise grammar probabilities
- Splitting step: propose new lexical entries based on current parses
  - If we think "the cat" means $\lambda x : \mathit{the}(x, \mathit{cat}(x))$, propose that "the" means $\lambda f : \mathit{the}(x, f(x))$

# Results

"Fast mapping" of rare words




60% of unseen sentences correct with full training set