Social meanings from social media

Jacob Eisenstein

Georgia Institute of Technology

July 21, 2012

Taming the **SOCIAL** web

- Language is the dominant modality for web content.
- Syntax (parsing) provides the scaffolding that enables the construction of propositional meaning from words.
- But social dynamics play an important, underappreciated role:
  - Sociolinguistic **variation** affects the expression of propositional content.
  - People use language to create **social meaning**, thus situating themselves in the social world.

Sociolinguistic variation

- **Variation** describes systematic differences in linguistic expression, e.g.
  - regional dialects
  - socioeconomic class differences
  - genre and register

Sociolinguistic variation

- **Variation** describes systematic differences in linguistic expression, e.g.
  - regional dialects
  - socioeconomic class differences
  - genre and register
- For example, from Twitter:
  - ayee i'm telling uu @LilTwist and i are getting married lol
  - Omqq =0 I Love uu Leel Wayne
  - Happy 21st birthday @NAME ! Love uu and miss you, sad I can't be there!

Sociolinguistic variation

- **Variation** describes systematic differences in linguistic expression, e.g.
  - regional dialects
  - socioeconomic class differences
  - genre and register
- For example, from Twitter:
  - ayee i'm telling uu @LilTwist and i are getting married lol
  - Omqq =0 I Love uu Leel Wayne
  - Happy 21st birthday @NAME ! Love uu and miss you, sad I can't be there!
- There's propositional content here!
  But variation makes it difficult to get.

Sociolinguistic variation

- **Variation** describes systematic differences in linguistic expression, e.g.
  - regional dialects
  - socioeconomic class differences
  - genre and register
- For example, from Twitter:
  - ayee i'm telling **uu** @LilTwist and i are getting married lol
  - Omqq =0 I Love **uu** Leel Wayne
  - Happy 21st birthday @NAME ! Love **uu** and miss you, sad I can't be there!
- There's propositional content here!
  But variation makes it difficult to get.
- uu is neither shorter nor easier than u.
  This extra effort suggests that uu carries social meaning.

Here's looking at uu

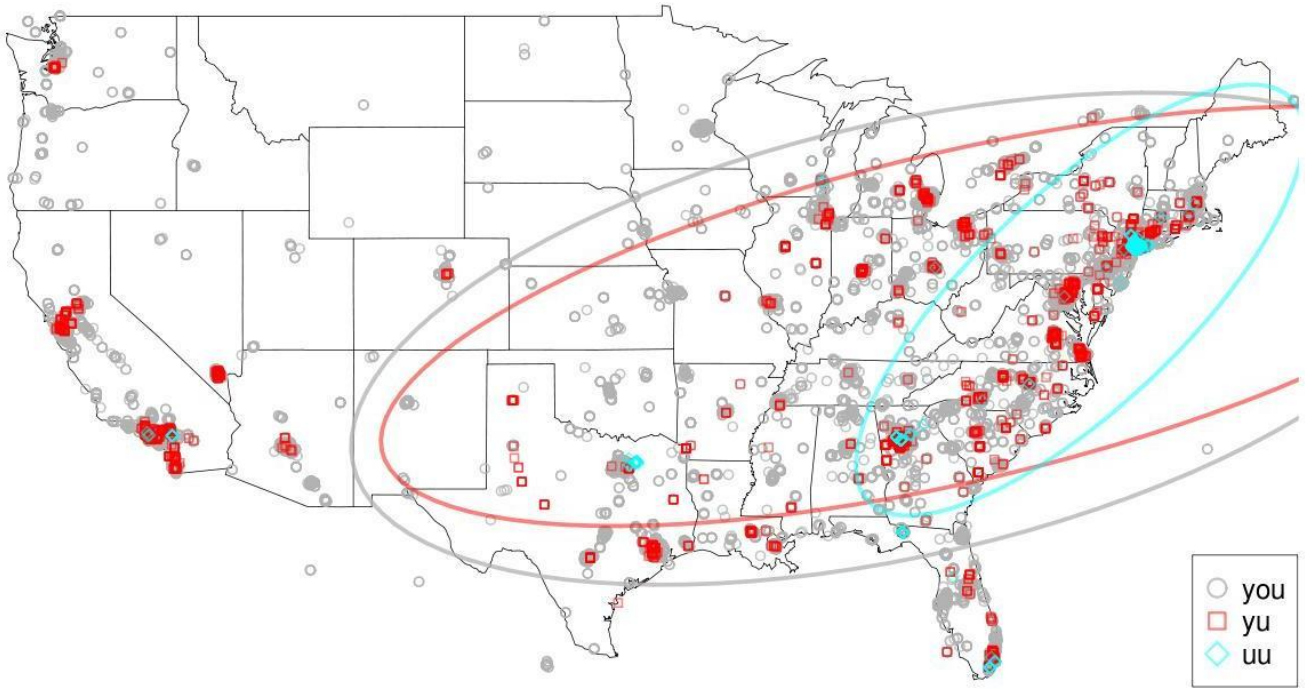
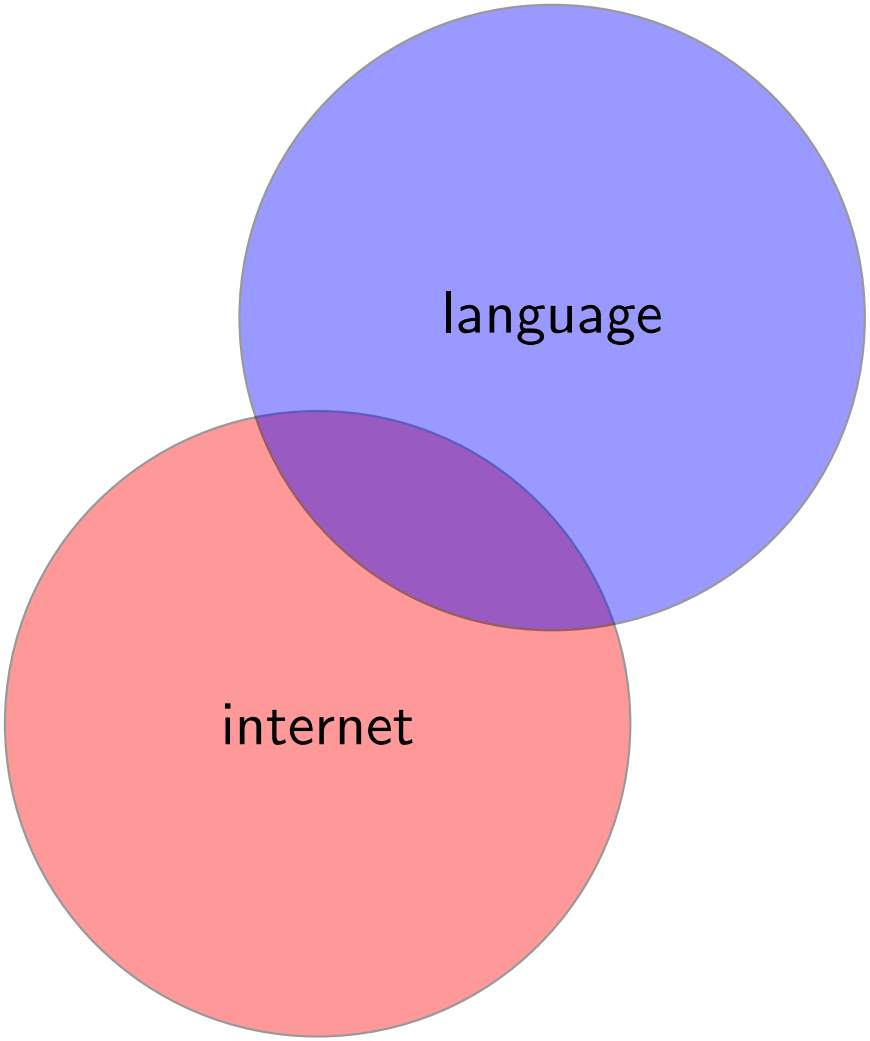


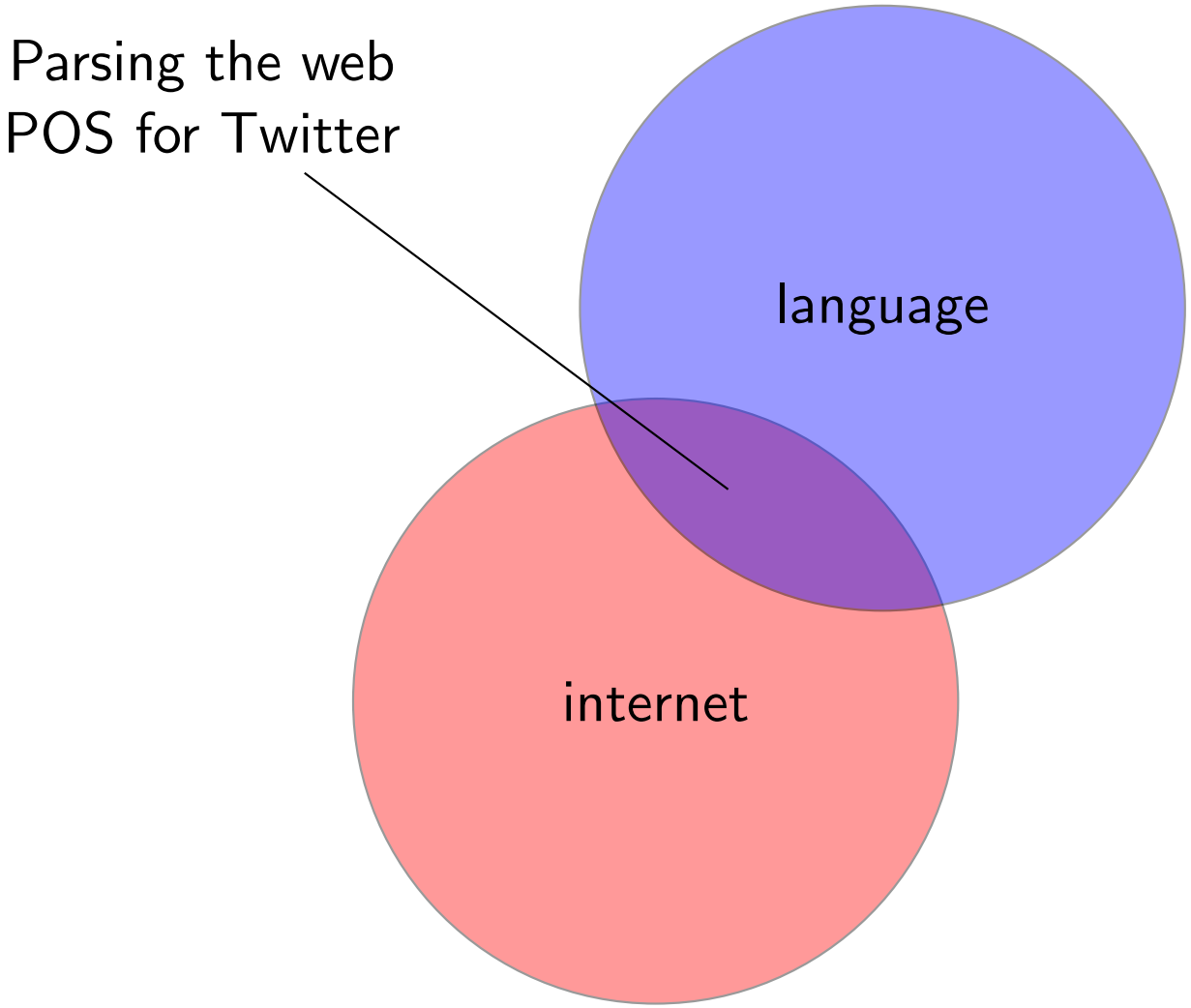
Figure: You and variants in March 2010

- Why uu instead of u?
- Why New York?
- Is it everyone in New York?

Taming the social web

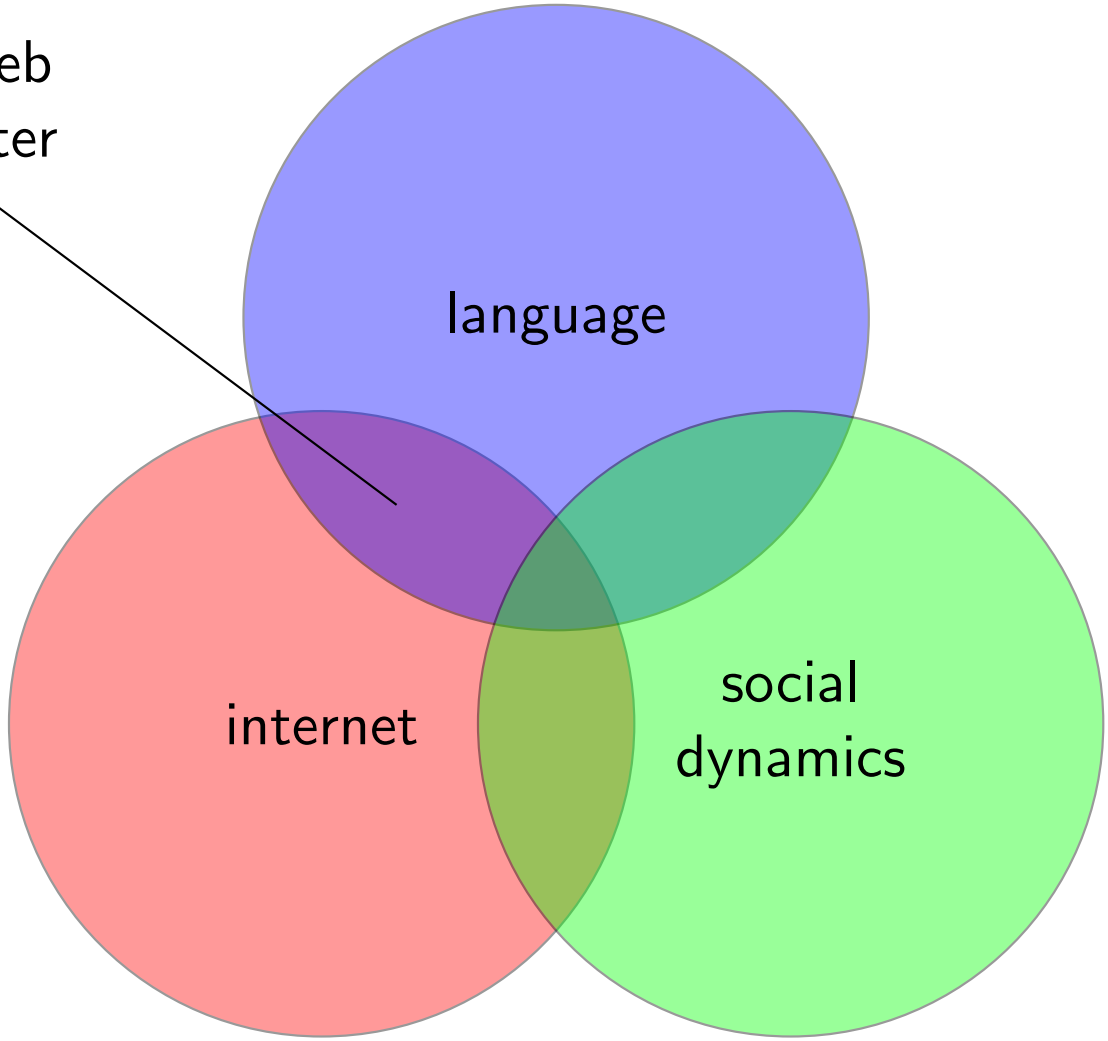


Taming the social web

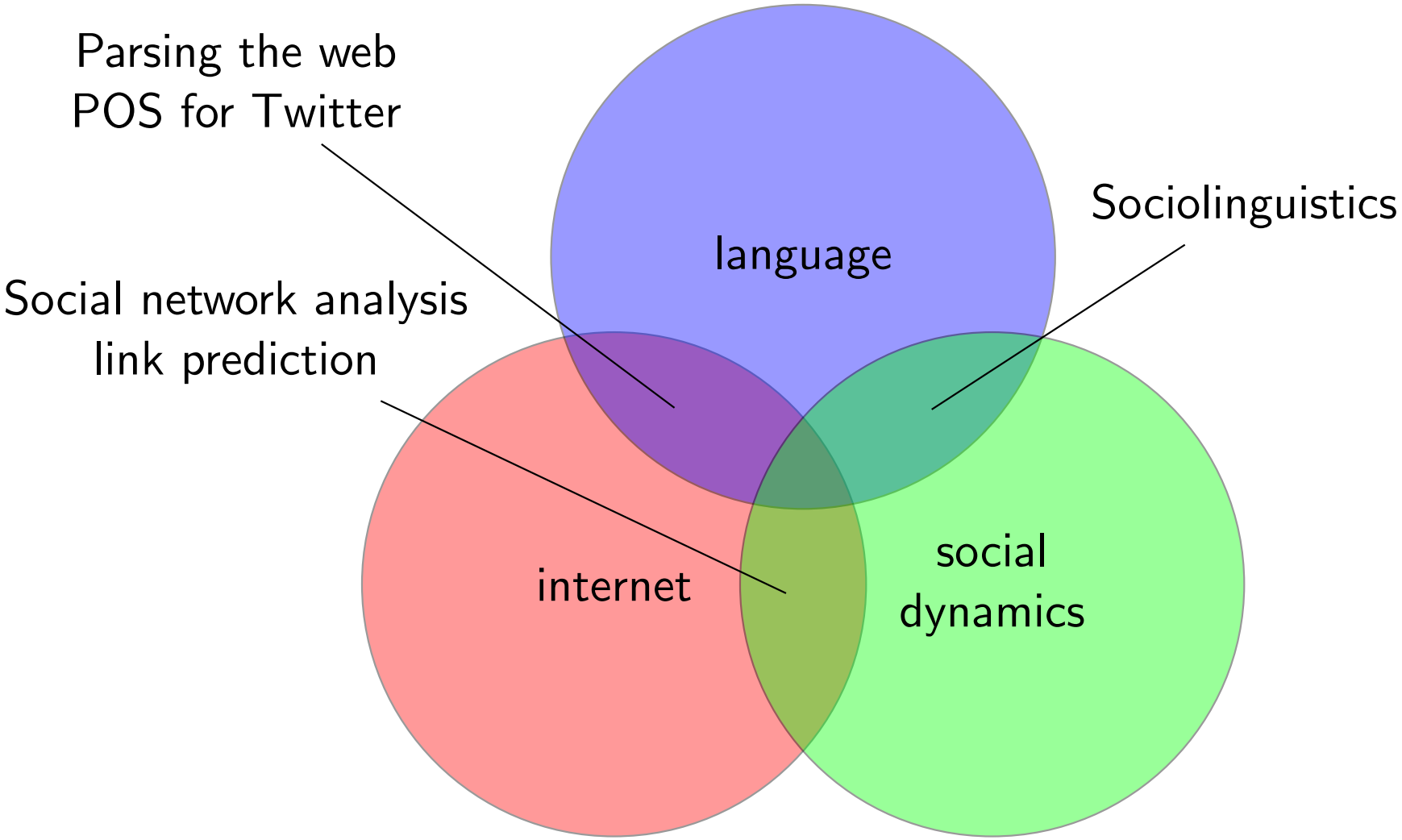


Taming the social web

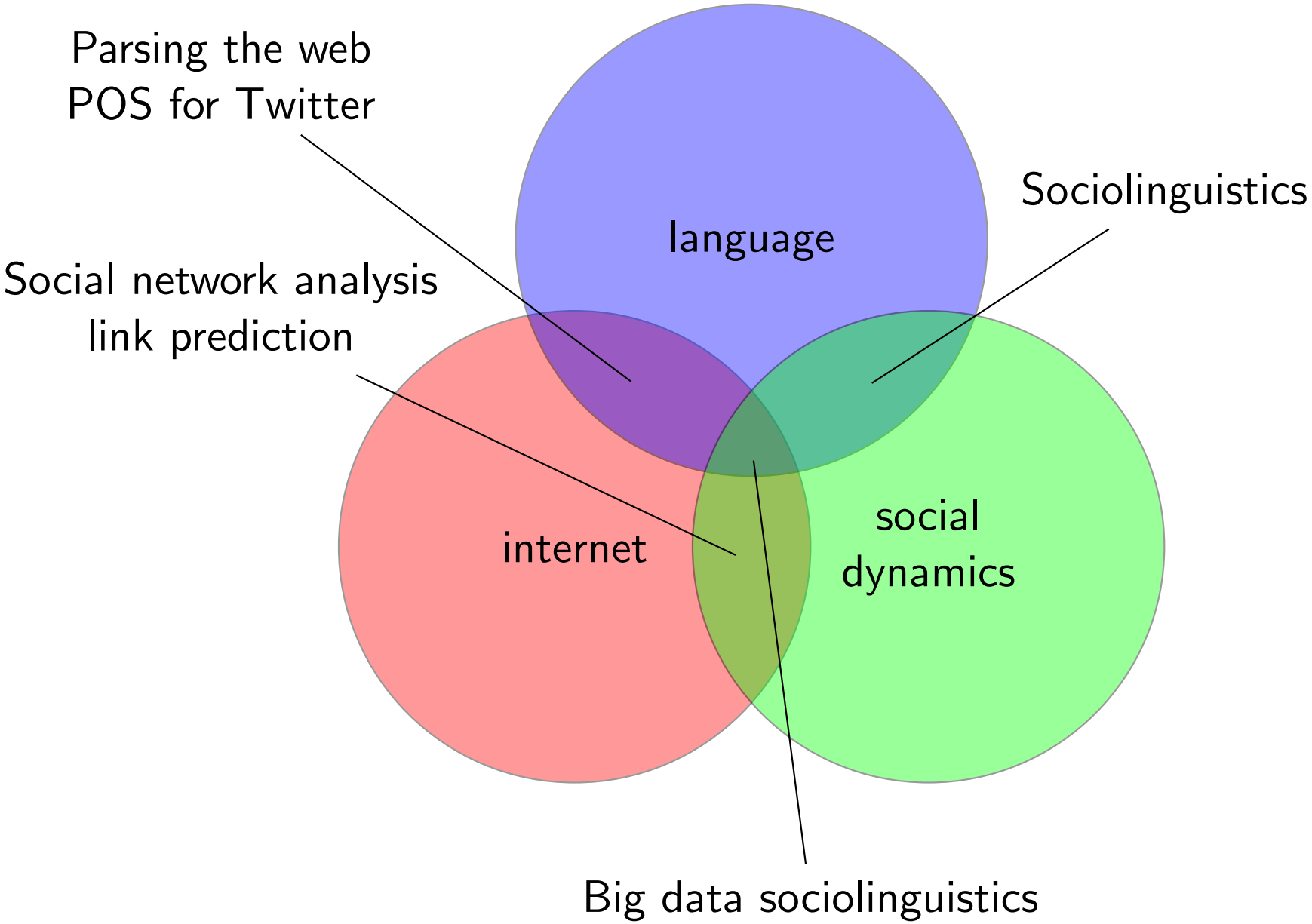
Parsing the web
POS for Twitter



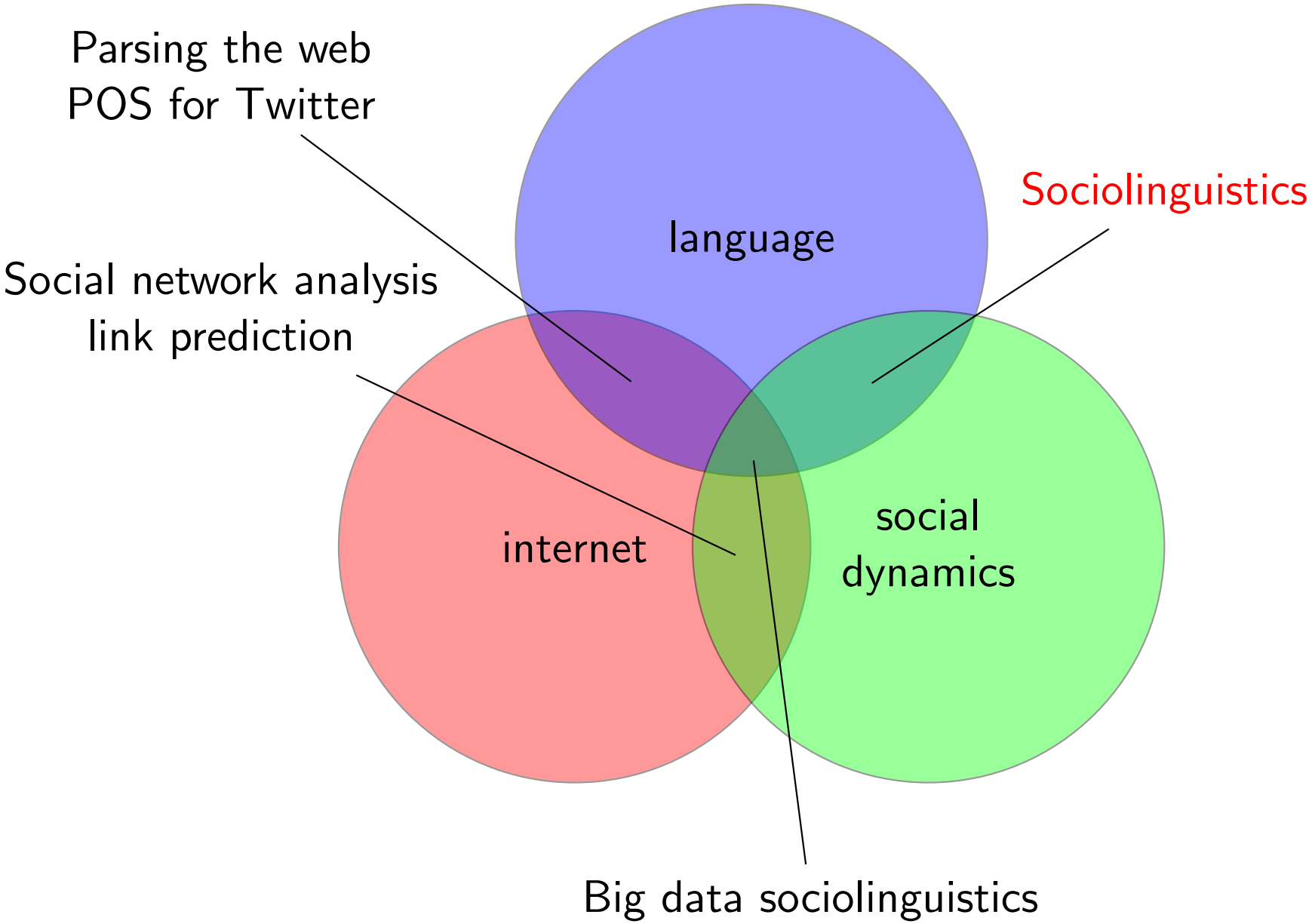
Taming the social web



Taming the social web



Taming the social web



Why sociolinguistics?

Those who do not understand sociolinguistics are condemned to say dumb things about language.

- **"40% of Twitter is meaningless babble."**
  – *a social media consulting firm*

Why sociolinguistics?

Those who do not understand sociolinguistics are condemned to say dumb things about language.

- **"40% of Twitter is meaningless babble."**
  – *a social media consulting firm*
- **Twitter is full of "rampant illiteracy."**
  – *the author of a guide to proper English*

Why sociolinguistics?

Those who do not understand sociolinguistics are condemned to say dumb things about language.

- **"40% of Twitter is meaningless babble."**
  – *a social media consulting firm*
- **Twitter is full of "rampant illiteracy."**
  – *the author of a guide to proper English*
- **Social media language is "noisy," "ungrammatical," "unstructured," or "urban."**
  – *papers in top NLP and machine learning venues!*

Why sociolinguistics?

Those who do not understand sociolinguistics are condemned to say dumb things about language.

- **"40% of Twitter is meaningless babble."**
  – *a social media consulting firm*
- **Twitter is full of "rampant illiteracy."**
  – *the author of a guide to proper English*
- **Social media language is "noisy," "ungrammatical," "unstructured," or "urban."**
  – *papers in top NLP and machine learning venues!*

Why sociolinguistics?

Those who do not understand sociolinguistics are condemned to say dumb things about language.

- **"40% of Twitter is meaningless babble."**
  – *a social media consulting firm*
- **Twitter is full of "rampant illiteracy."**
  – *the author of a guide to proper English*
- **Social media language is "noisy," "ungrammatical," "unstructured," or "urban."**
  – *papers in top NLP and machine learning venues!*

ChuckGrassley @ChuckGrassley 13 Feb

I now h v an iphone

Retweeted by Jacob Eisenstein

Collapse Reply Retweeted Favorite

How sociolinguists see language variation

Sociolinguists (and linguists generally) believe:

- Language standards are real. Obeying them can be important.

How sociolinguists see language variation

Sociolinguists (and linguists generally) believe:

- Language standards are real. Obeying them can be important.
- But non-standard language is not illiteracy.

How sociolinguists see language variation

Sociolinguists (and linguists generally) believe:

- Language standards are real. Obeying them can be important.
- But non-standard language is not illiteracy.
- Linguistic variation is not random noise; it carries social meaning.
- Our job is to understand the ways people use language, not to defend the linguistic status quo.

The department store study

"The Social Stratification of /r/ in New York City" [Lab66]

- Where can I return this tie?
- The fourth floor.
- Sorry, where was that?
- The FOURTH FLOOR.

The department store study

"The Social Stratification of /r/ in New York City" [Lab66]

- Where can I return this tie?
- The fourth floor.
- Sorry, where was that?
- The FOURTH FLOOR.

Findings of this study:

- Use of /r/ correlates with higher prices (Sak's > Macy's > Klein's)
- Use of /r/ increases with "attention to speech," especially in middle-class settings.

Variables, speakers, and context

- What did Labov have to know to do this research?
  - **language**: /r/-dropping as a linguistic *variable*
  - **speaker**: socioeconomic status as modulating regional dialect
  - **context**: workplace customer-service dialogue

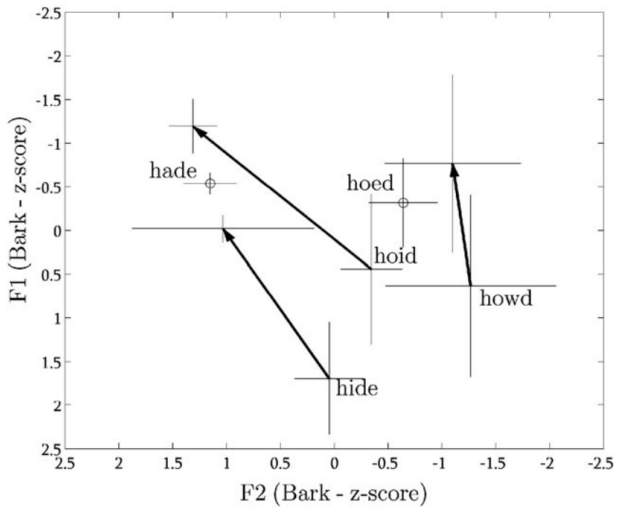
Variables, speakers, and context

- What did Labov have to know to do this research?
  - **language**: /r/-dropping as a linguistic *variable*
  - **speaker**: socioeconomic status as modulating regional dialect
  - **context**: workplace customer-service dialogue
- These 3 elements define the playing field of sociolinguistics.
- Labov's work was seminal to *variationist* sociolinguistics, an approach which embodies a specific set of methological and theoretical commitments with respect to this playing field.

The methodology of "first-wave" variationism [Eck12]

language: focused on phonological variables

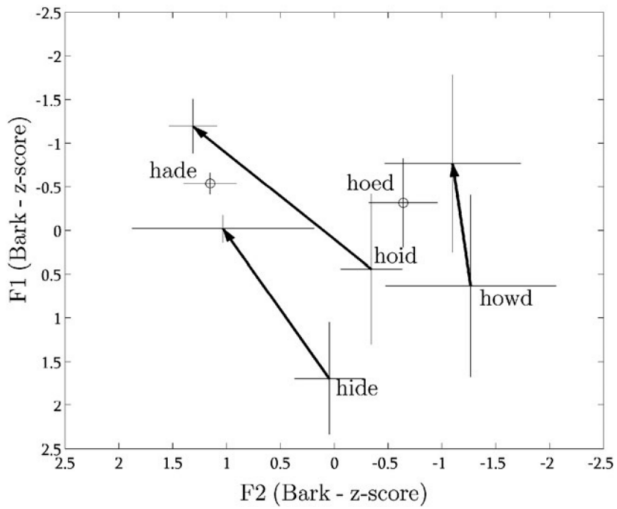
- can often be measured objectively
- can be obtained from little text and with minimal prompting (unlike lexical variables).



The methodology of "first-wave" variationism [Eck12]

language: focused on phonological variables

- can often be measured objectively
- can be obtained from little text and with minimal prompting (unlike lexical variables).



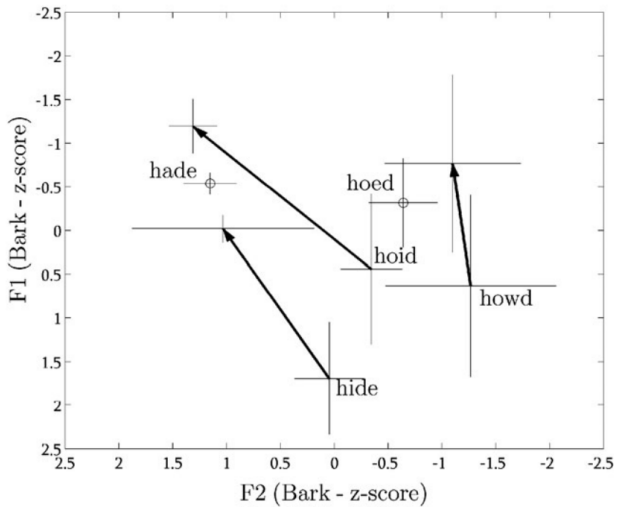
speaker: emphasis on broad categories, like class and gender

- originally focused on urban working class
- data obtained through telephone surveys and interviews

The methodology of "first-wave" variationism [Eck12]

language: focused on phonological variables

- can often be measured objectively
- can be obtained from little text and with minimal prompting (unlike lexical variables).



speaker: emphasis on broad categories, like class and gender

- originally focused on urban working class
- data obtained through telephone surveys and interviews

context: distinction between *vernacular* and *prestige* forms

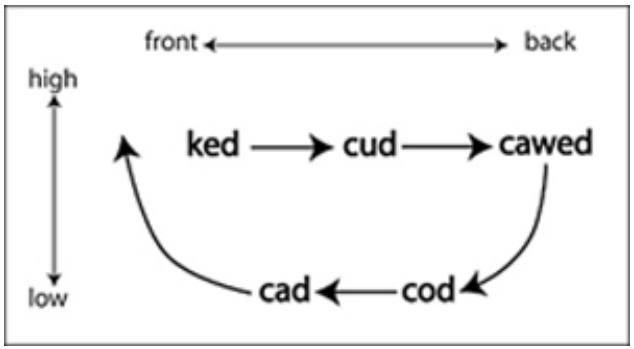
- The vernacular is inhibited by attention to speech.
- Careful interview methodology to elicit "truly" vernacular speech.

Insights from variationist sociolinguistics

- Language variation can be studied mathematically.
  (Logistic regression in the 1970s!)
- Cross-speaker variation is not purely geographical,
  but depends on multiple interacting social factors.
- Within-speaker variation is systematic, not random.
- Most importantly: **a coherent theory of language change**.

Insights from variationist sociolinguistics

- Language variation can be studied mathematically.
  (Logistic regression in the 1970s!)
- Cross-speaker variation is not purely geographical,
  but depends on multiple interacting social factors.
- Within-speaker variation is systematic, not random.
- Most importantly: **a coherent theory of language change**.



- Phonology is an integrated system.
- Change is triggered by social and historical factors.
- These factors can be understood by identifying the leaders of language change.

The Philadelphia study

- In the 1970s, a large-scale study of language change in Philadelphia [Lab01] asked the question: **who is changing language**?

The Philadelphia study

- In the 1970s, a large-scale study of language change in Philadelphia [Lab01] asked the question: **who is changing language**?
- A detective story with lots of likely culprits:
  - **Change from above**: Only the upper class can change language, because they hold the political, economic, and cultural power.

The Philadelphia study

- In the 1970s, a large-scale study of language change in Philadelphia [Lab01] asked the question: **who is changing language**?
- A detective story with lots of likely culprits:
  - **Change from above**: Only the upper class can change language, because they hold the political, economic, and cultural power.
  - **Change from below**: The lower class changes language because they are disadvantaged by the existing social order and therefore reject it.

The Philadelphia study

- In the 1970s, a large-scale study of language change in Philadelphia [Lab01] asked the question: **who is changing language**?
- A detective story with lots of likely culprits:
  - **Change from above**: Only the upper class can change language, because they hold the political, economic, and cultural power.
  - **Change from below**: The lower class changes language because they are disadvantaged by the existing social order and therefore reject it.
  - **Change from the bourgeois**: The upper middle class because they are the most economically dynamic, have most diverse social connections.

The Philadelphia study

- In the 1970s, a large-scale study of language change in Philadelphia [Lab01] asked the question: **who is changing language**?
- A detective story with lots of likely culprits:
  - **Change from above**: Only the upper class can change language, because they hold the political, economic, and cultural power.
  - **Change from below**: The lower class changes language because they are disadvantaged by the existing social order and therefore reject it.
  - **Change from the bourgeois**: The upper middle class because they are the most economically dynamic, have most diverse social connections.
  - As in any good detective story, none of the usual suspects are guilty.

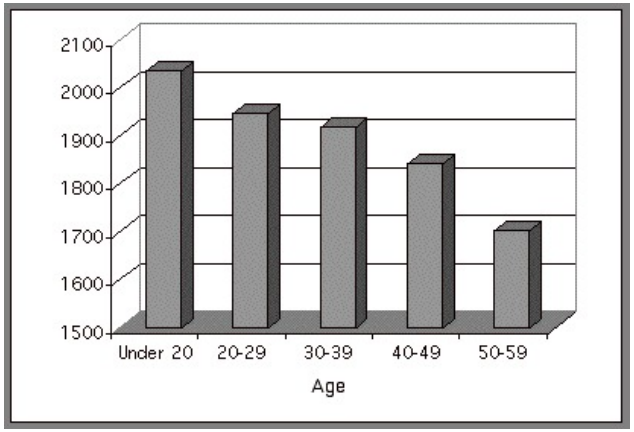
The Philadelphia study: methodology

How can we identify changes in progress without waiting for years?

The Philadelphia study: methodology

How can we identify changes in progress without waiting for years?

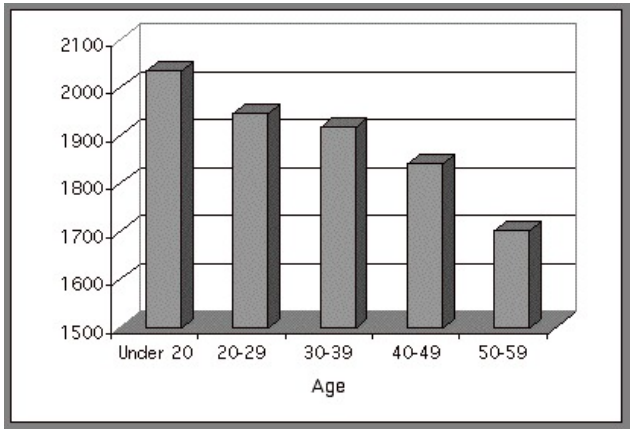
- **Apparent time**: differential language patterns by speaker age.



The Philadelphia study: methodology

How can we identify changes in progress without waiting for years?

- **Apparent time**: differential language patterns by speaker age.

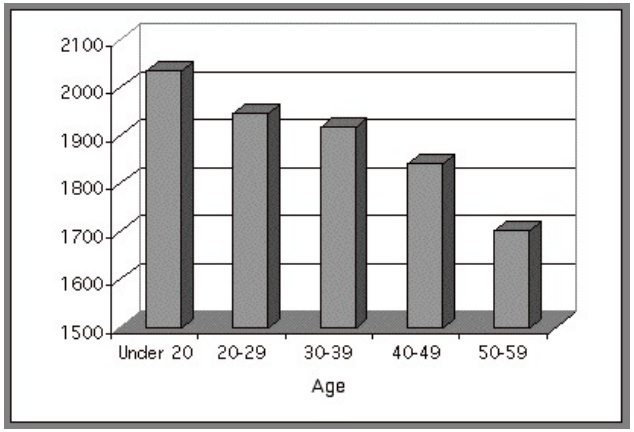


How to find the leaders of language change?

The Philadelphia study: methodology

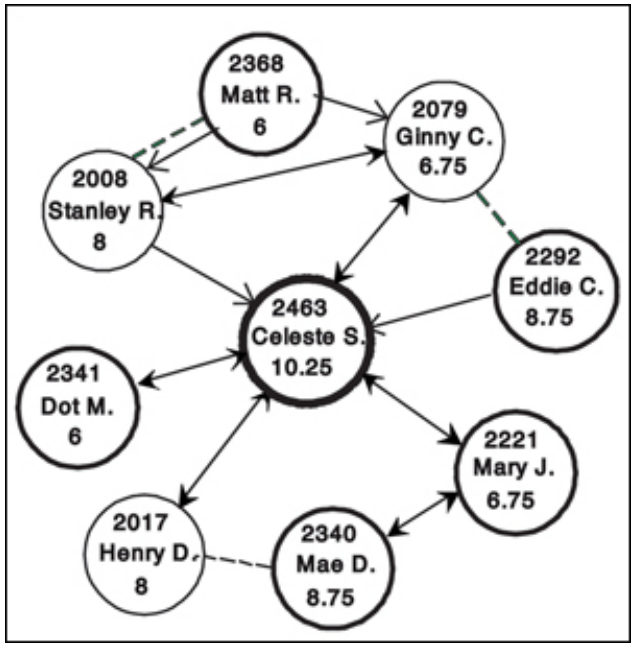
How can we identify changes in progress without waiting for years?

- **Apparent time**: differential language patterns by speaker age.



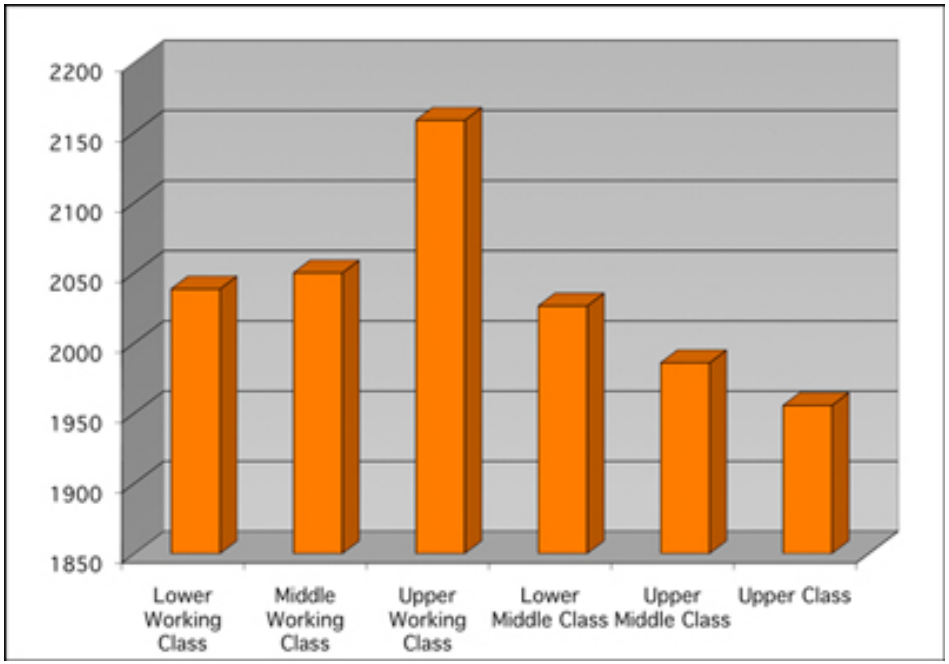
How to find the leaders of language change?

- Multiple local communities were studied over several months.
- The communities were differentiated by class and ethnicity, to measure the broad characteristics of language change.
- In-depth study within each community reveals the personal characteristics of change leaders, and their position in their local social network.



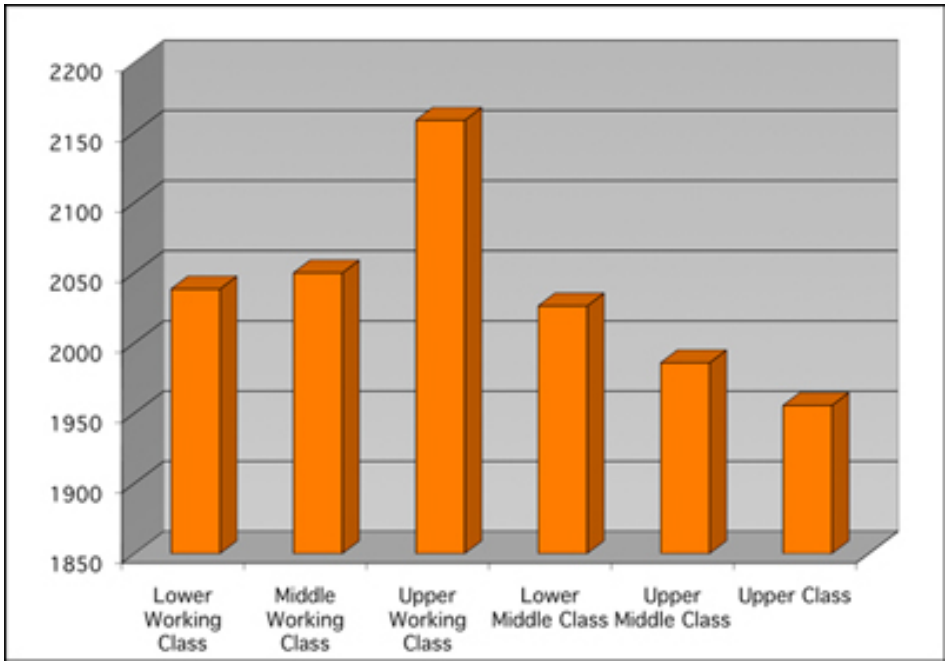
The Philadelphia study: results

Change is triggered by the **upper working class**.



The Philadelphia study: results

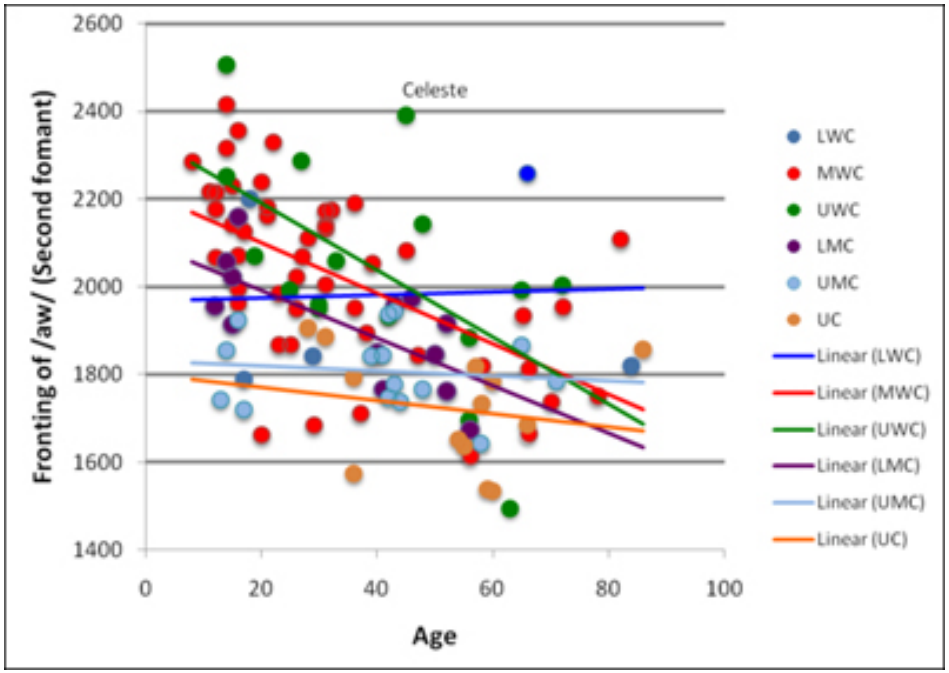
Change is triggered by the **upper working class**.



- Strong local ties, but heterogeneous work connections
- Leaders are also usually: female, socially central, with a non-conformist personality (anecdotally)

The Philadelphia study: results

Change is triggered by the **upper working class**.



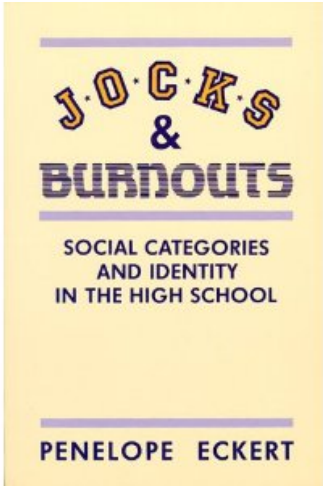
- Strong local ties, but heterogeneous work connections
- Leaders are also usually: female, socially central, with a non-conformist personality (anecdotally)

Jocks and burnouts

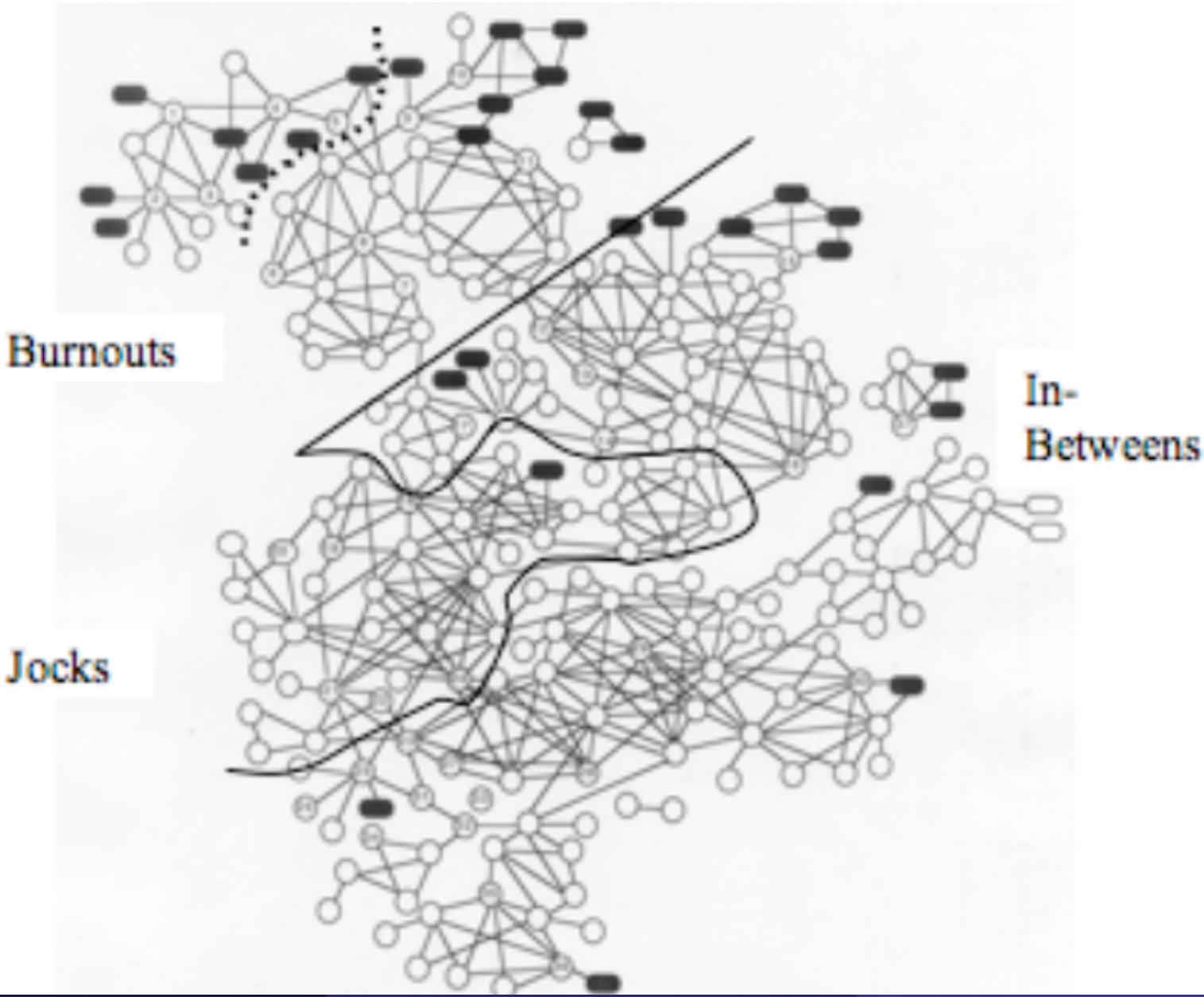
- First-wave variationist sociolinguistics focused on large-scale factors: class, race, and gender.
- But variation occurs even within homogeneous local communities.

Jocks and burnouts

- First-wave variationist sociolinguistics focused on large-scale factors: class, race, and gender.
- But variation occurs even within homogeneous local communities.
- In the 1980s, Eckert studied a high school in suburban Detroit. The school was sharply divided into two groups [Eck89].
  - **Jocks** participate in school-sponsored activities, plan to go to college, and cultivate good relations with authority figures.
  - **Burnouts** reject the school, plan to work in Detroit, and cultivate friendships in the larger metropolitan area.

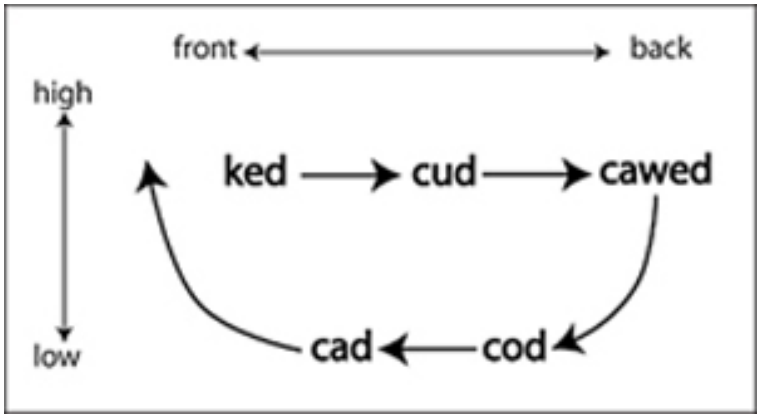


Jocks and burnouts



Jocks and burnouts

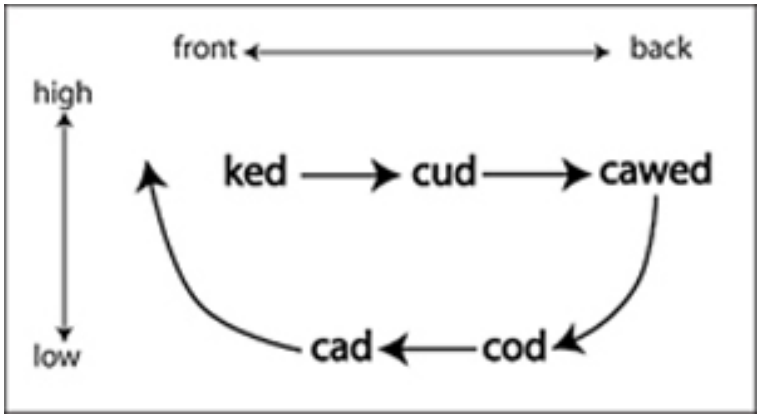
- In the 1980s, Detroit was undergoing the Northern Cities Vowel Shift, a change in the position of several vowels.



- Eckert asked whether these groups are responding to this shift differently.

Jocks and burnouts

- In the 1980s, Detroit was undergoing the Northern Cities Vowel Shift, a change in the position of several vowels.

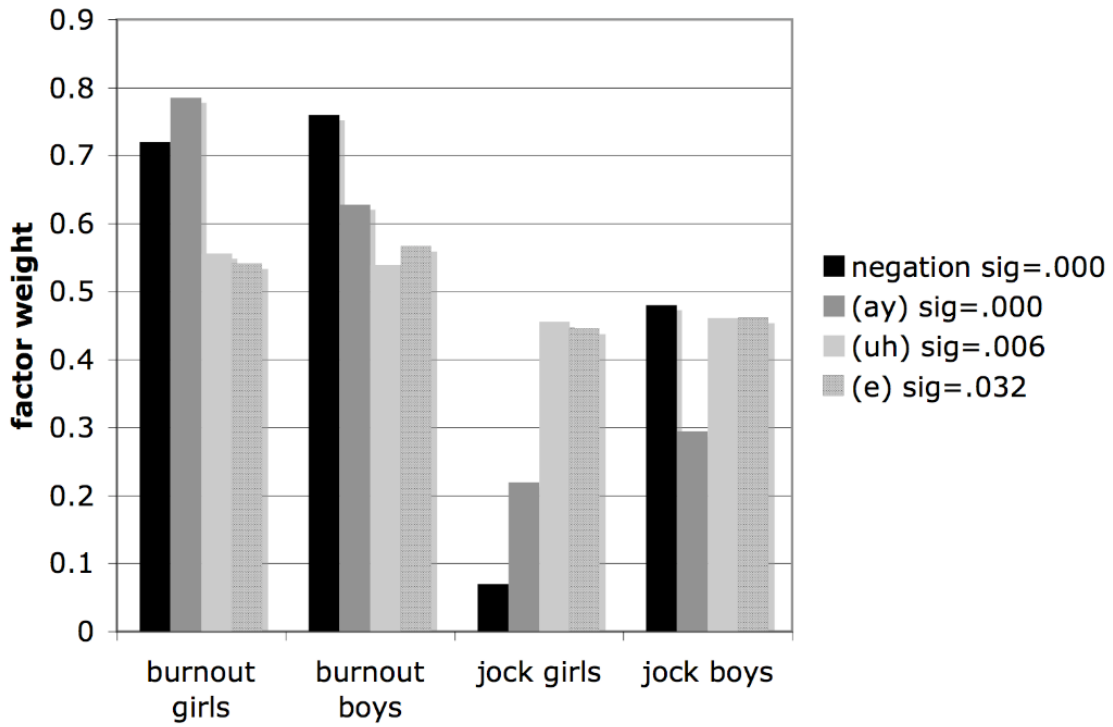


- Eckert asked whether these groups are responding to this shift differently.
  - **language**: Northern Cities shifted vowels, multiple negation
  - **speaker**: locally-defined jock and burnout groups
  - **context**: ethnographic interviews

Jocks and burnouts: results

Jocks and burnouts differ markedly on the northern cities shift:

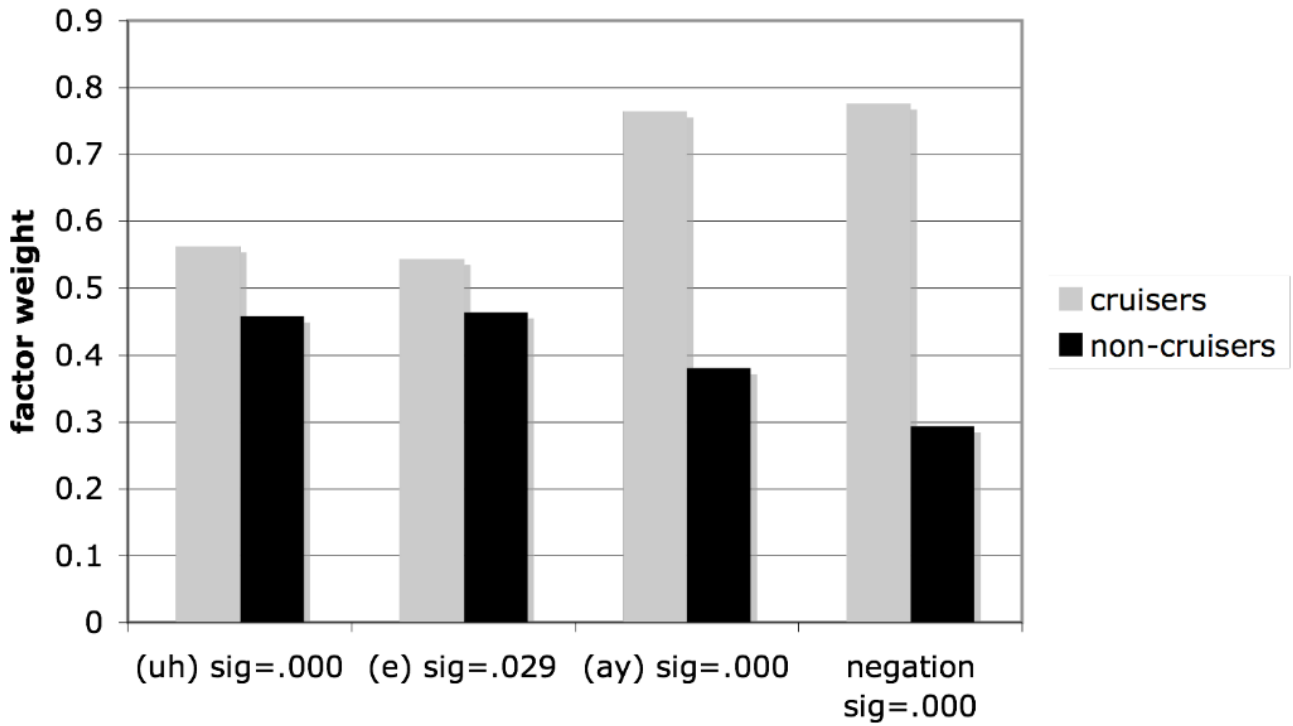
- Burnouts adopt a more Detroit-oriented style.
- Jocks adopt a style that is closer to the national standard.



Jocks and burnouts: results

Jocks and burnouts differ markedly on the northern cities shift:

- Burnouts adopt a more Detroit-oriented style.
- Jocks adopt a style that is closer to the national standard.



Variation in local communities

A related example from Ucieda, in the Spanish Pyrenees [Hol85]

- Castilian: El trabajo del campo no lo saben
- Ucieda: El trabaj**u** del camp**u** no lo saben

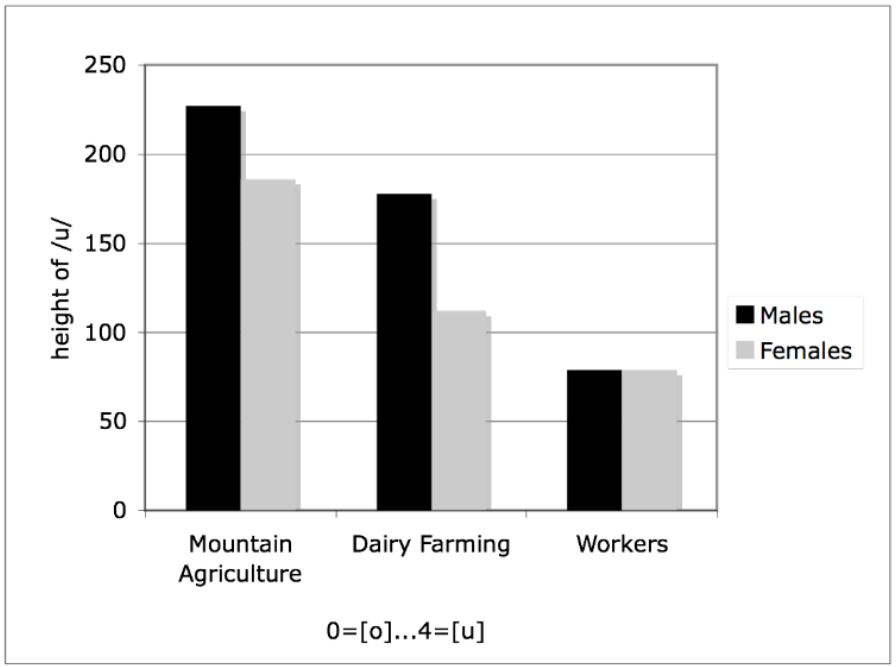
This is a change in progress, towards the Castilian standard.

Variation in local communities

A related example from Ucieda, in the Spanish Pyrenees [Hol85]

- Castilian: El trabajo del campo no lo saben
- Ucieda: El trabaj**u** del camp**u** no lo saben

This is a change in progress, towards the Castilian standard.



- Uciedans doing less traditional work are changing most quickly.
- The change is led by women. Eckert argues that this reflects the particular undesirability of mountain agriculture for women [Eck12].

Variationist sociolinguistics today

Recent variationist sociolinguistics focuses on **local meanings** [Buc11].

- How do people use language to define a social identity?
- How does language reflect and perpetuate social structures?

Variationist sociolinguistics today

Recent variationist sociolinguistics focuses on **local meanings** [Buc11].

- How do people use language to define a social identity?
- How does language reflect and perpetuate social structures?

Variationist sociolinguistics has become **less** quantitative over time.

- Data: from survey to structured interview to ethnography
- Analysis: from quantitative to discourse analysis

Variationist sociolinguistics today

Recent variationist sociolinguistics focuses on **local meanings** [Buc11].

- How do people use language to define a social identity?
- How does language reflect and perpetuate social structures?

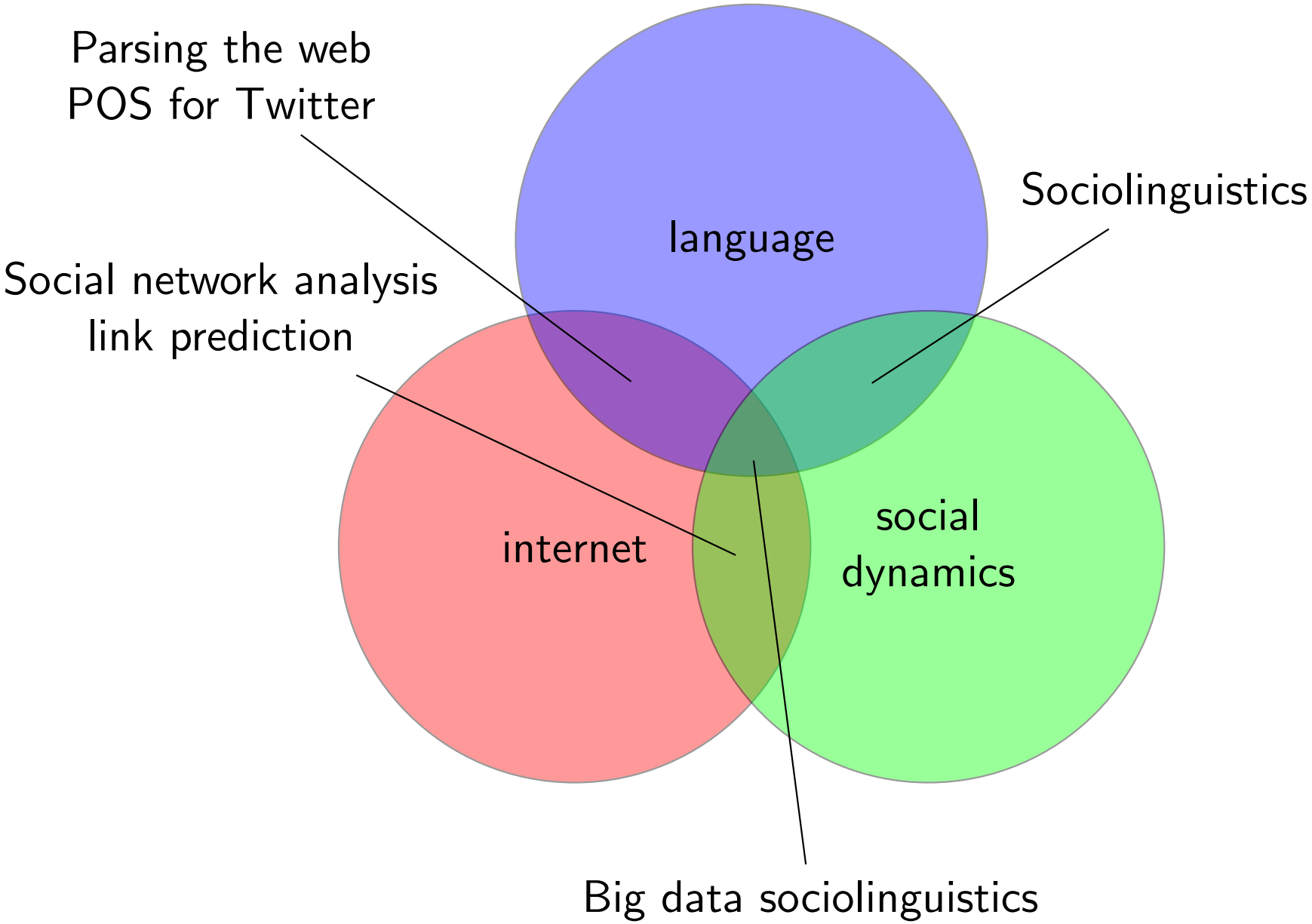
Variationist sociolinguistics has become **less** quantitative over time.

- Data: from survey to structured interview to ethnography
- Analysis: from quantitative to discourse analysis

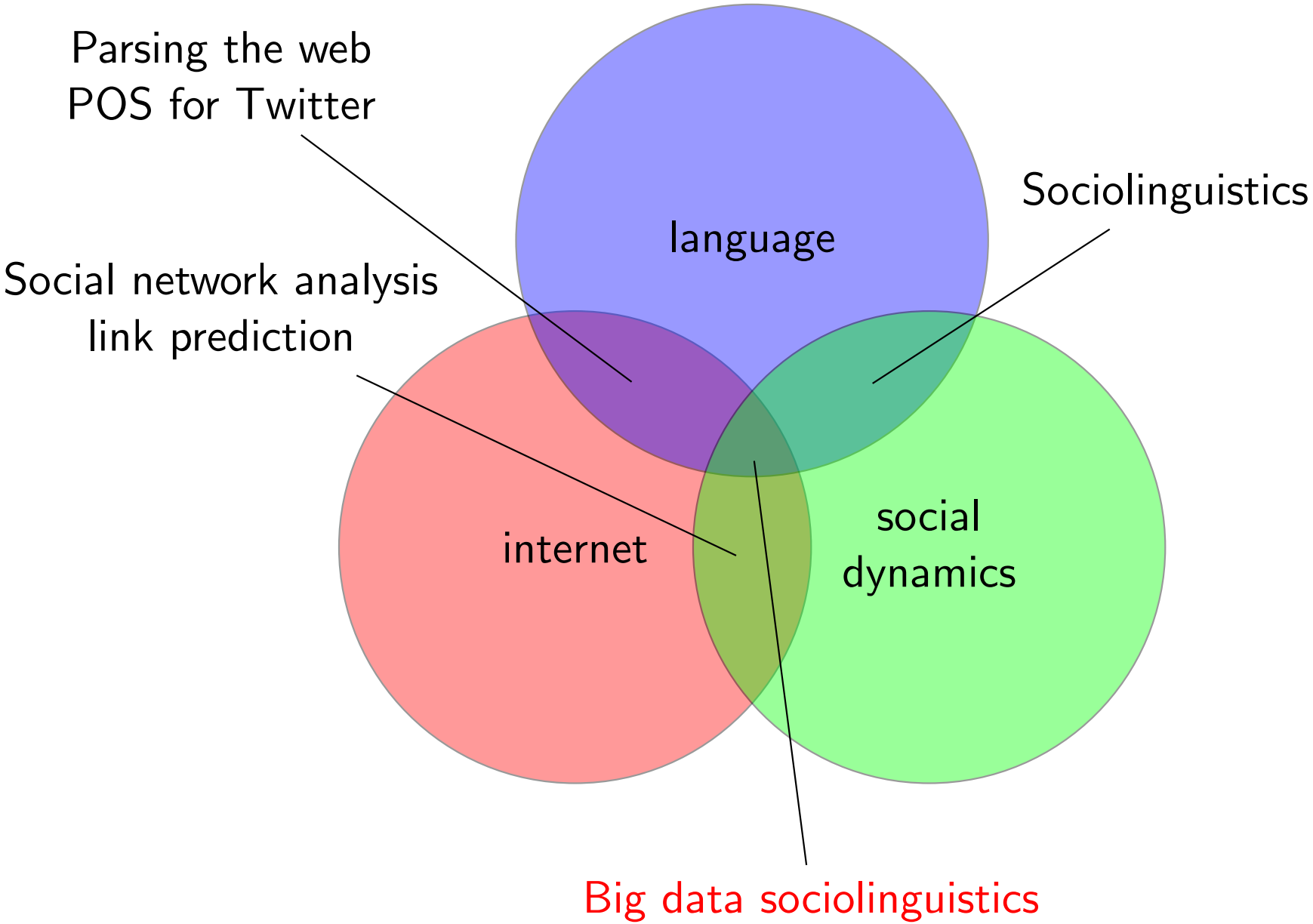
This is not a disavowal of quantitative "1st wave" sociolinguistics.

- But interest has shifted to more deeply contextual settings, and quantitative methods seem unable to contribute.
- (but is this true?)

Taming the social web



Taming the social web



Towards big data sociolinguistics

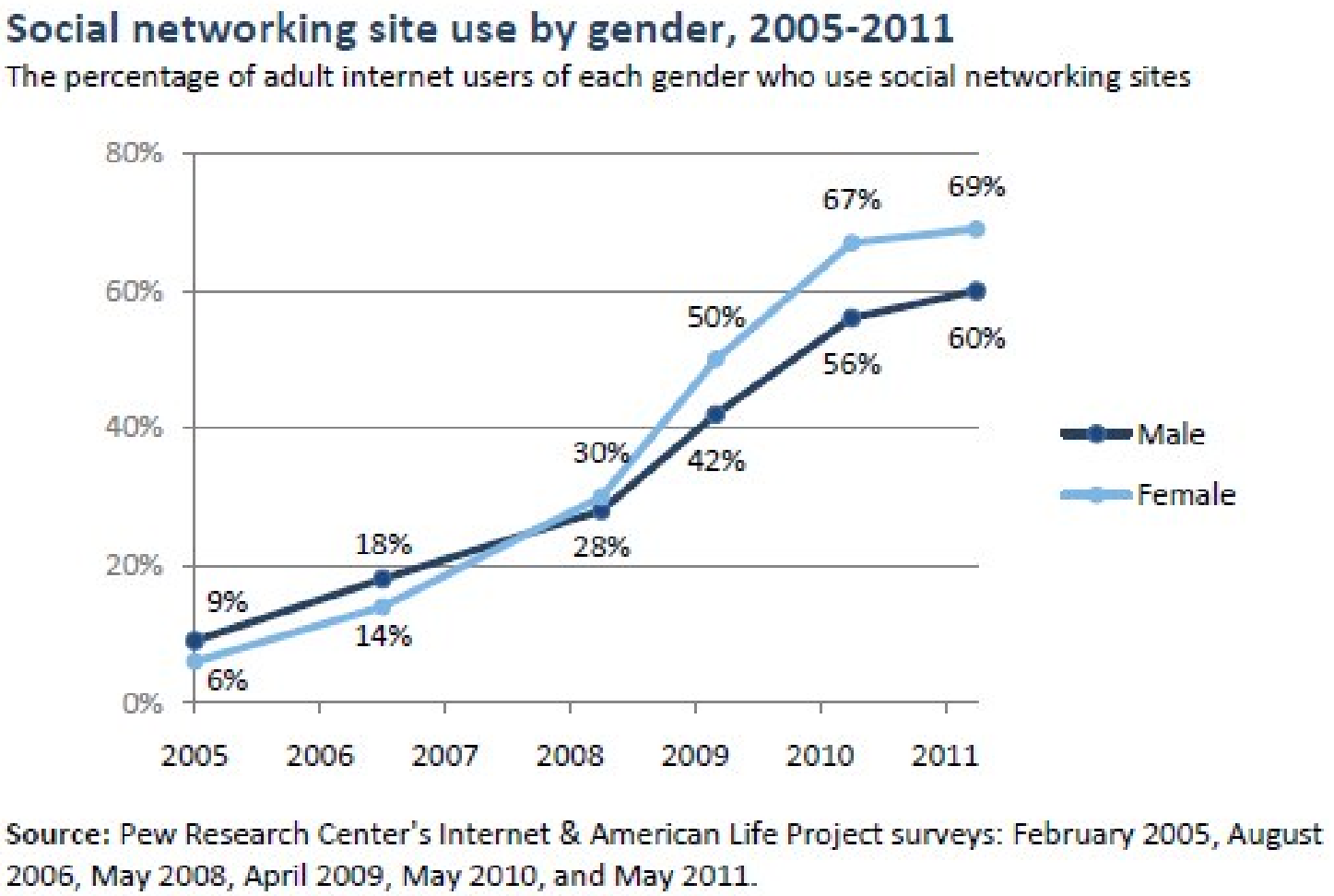
Social media corpora open the possibility of a new, "big data" methodology for sociolinguistics.

Towards big data sociolinguistics

Social media corpora open the possibility of a new, "big data" methodology for sociolinguistics.

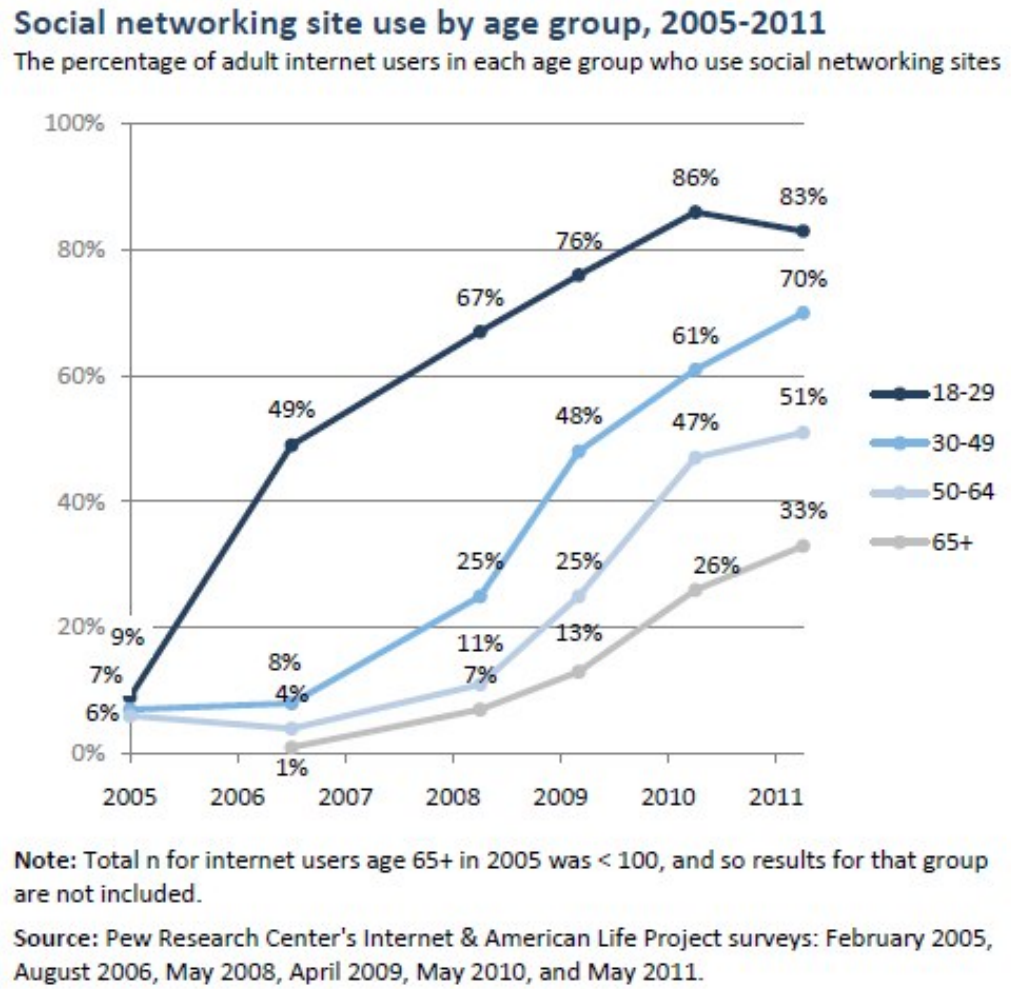
- **Exploratory analysis**
  find linguistic variables in the data, rather than relying on experimenter's intuition.
- **Limited observer bias**
  language from real (public) social interactions, outside a lab.
- **Law of large numbers**
  a big pool of participants means less sensitivity to outliers.

Who uses social media?



(Pew Research Center, Aug 2011)

Who uses social media?

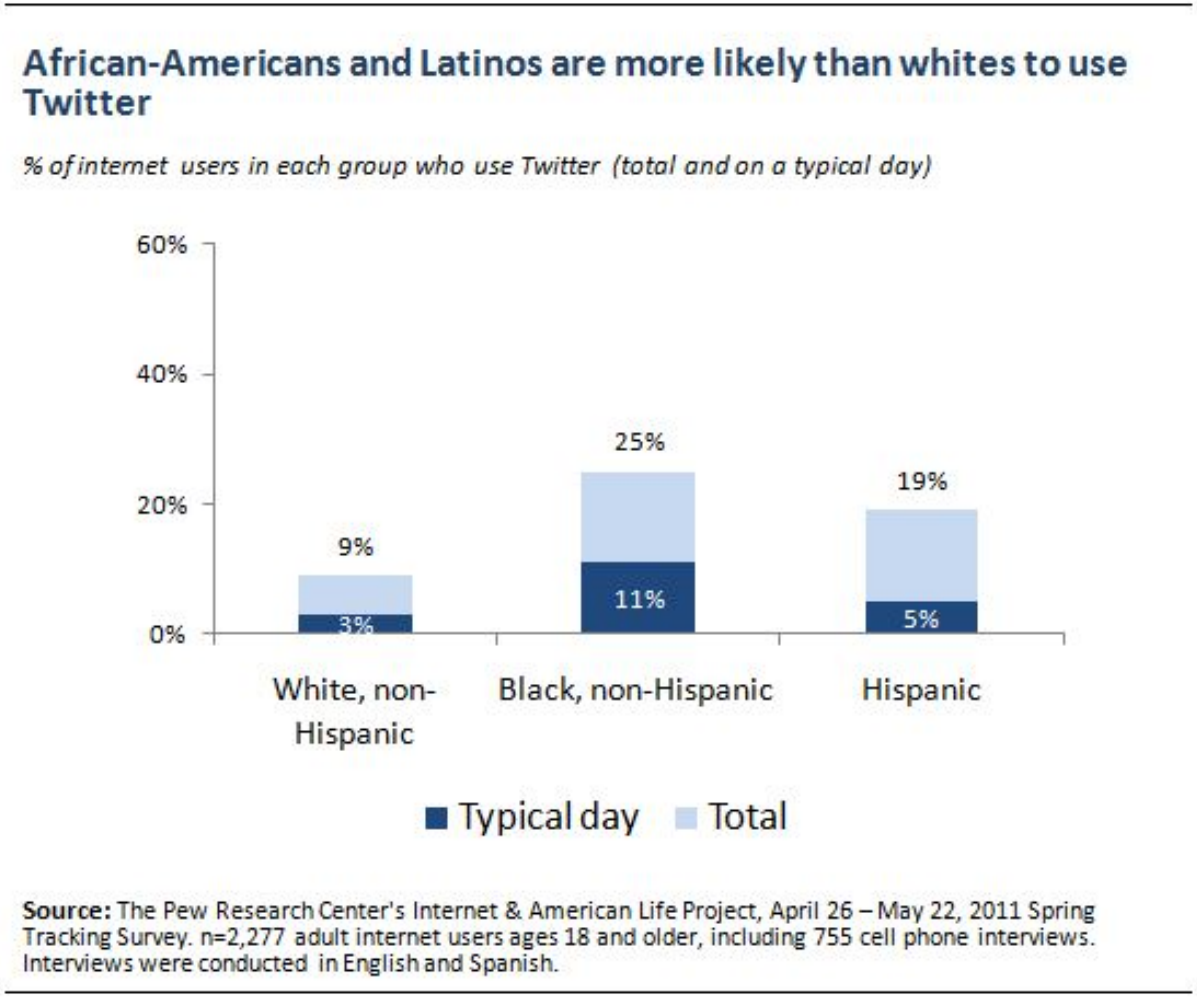


(Pew Research Center, Aug 2011)

Twitter

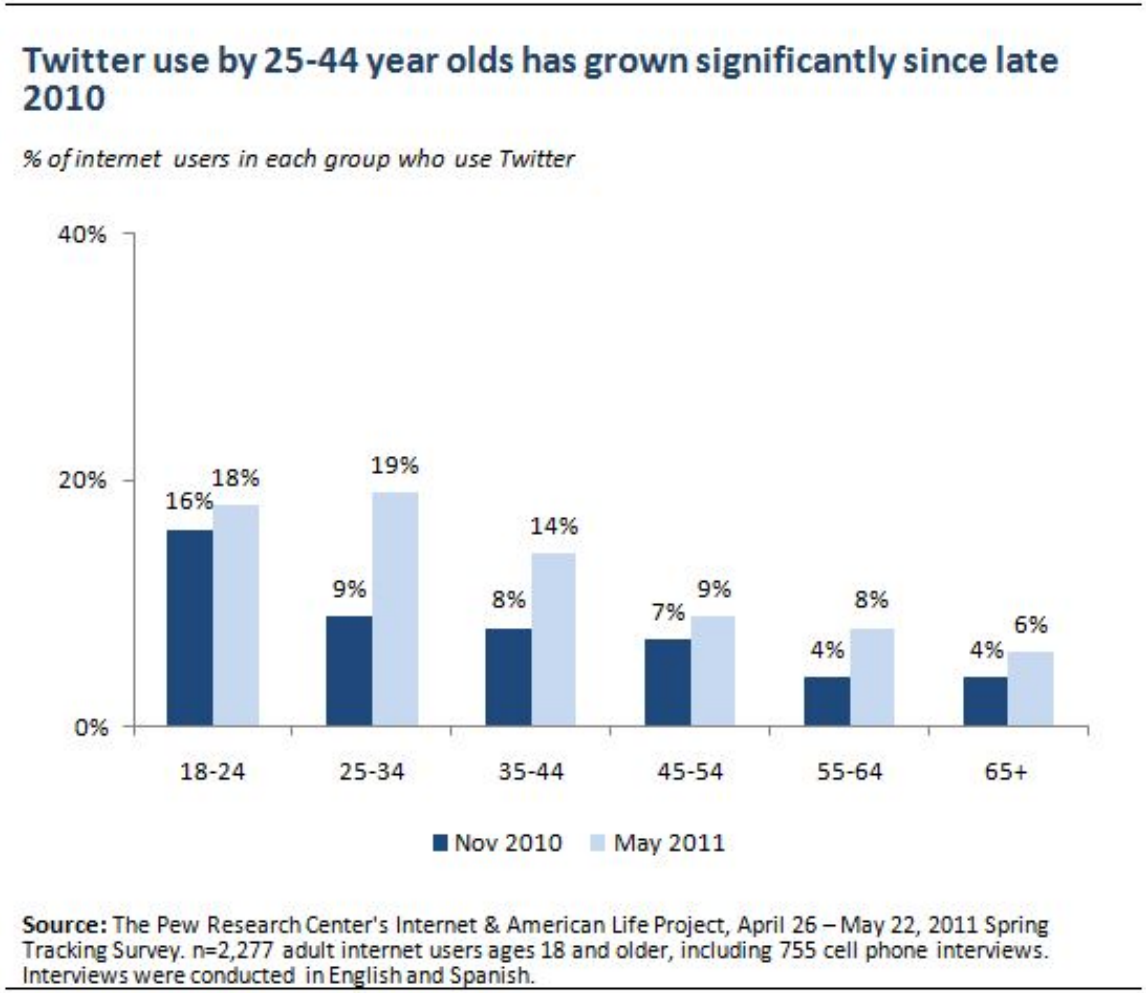
- Weak tie to real-life identity
- Short, unstructured content (140 characters)
- Unidirectional social network connections
- Public but not redistributable

Who uses Twitter?



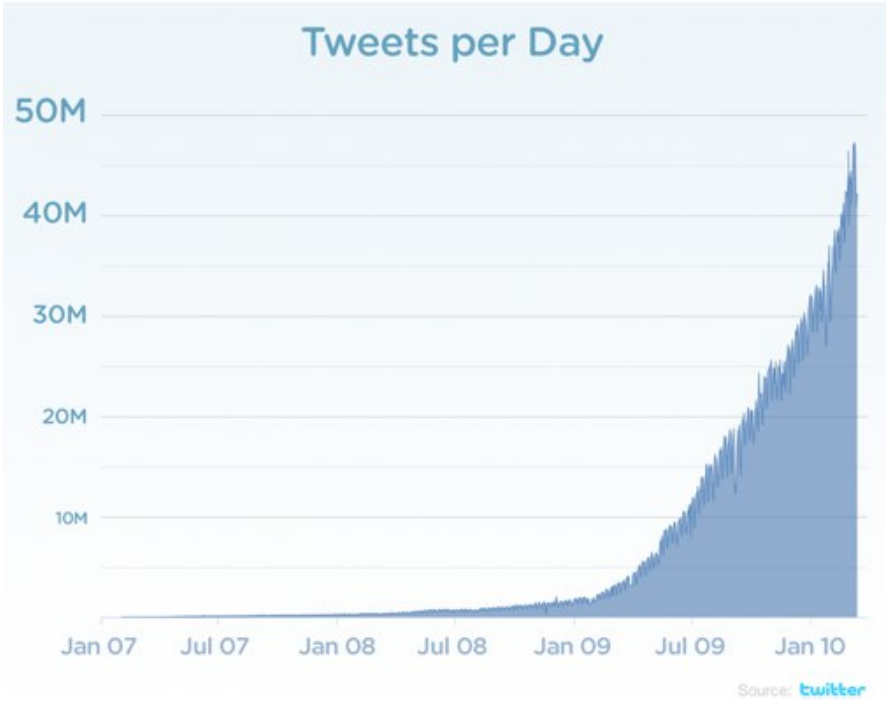
(Pew Research Center, June 2011)

Who uses Twitter?



(Pew Research Center, June 2011)

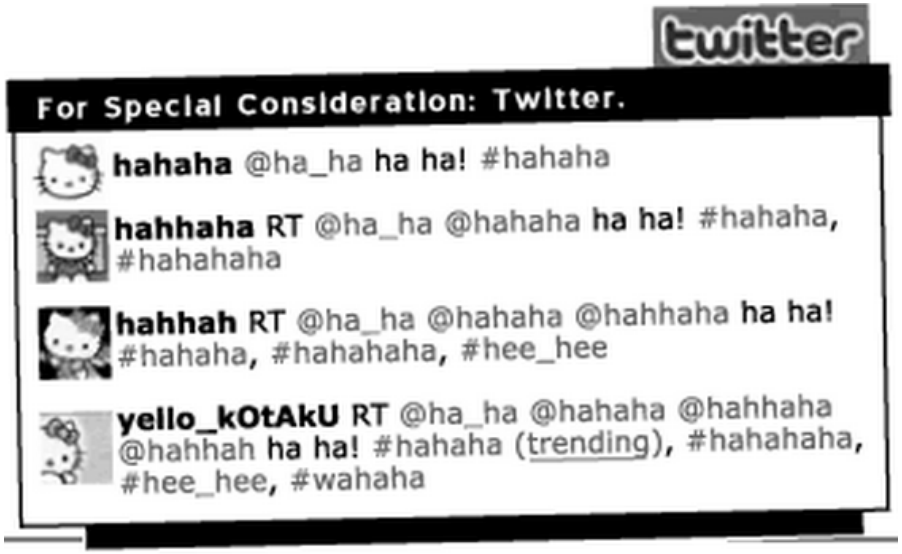
How much?



- Twitter claims:
  100 million active users, 177 million tweets per day in March 2011

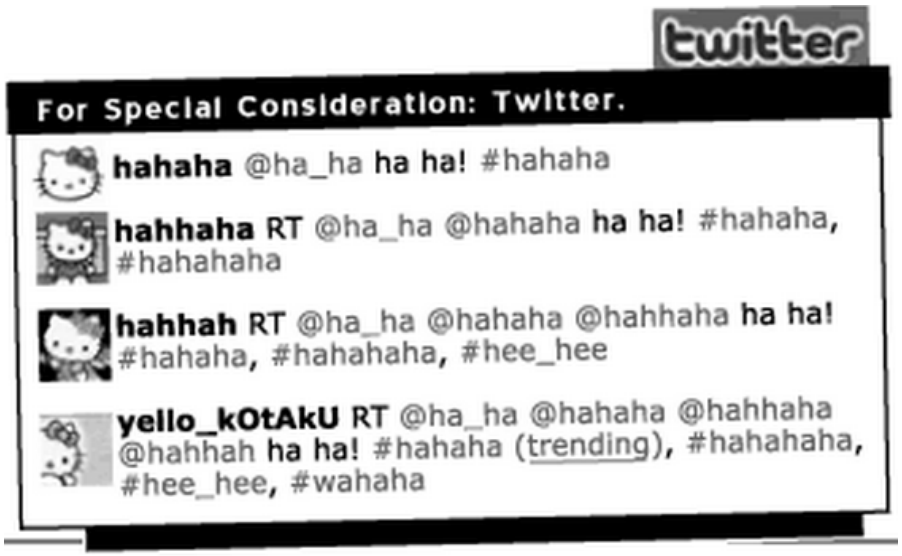
What does social media language look like?

What does social media language look like?



(**Achewood**, January 17, 2010)

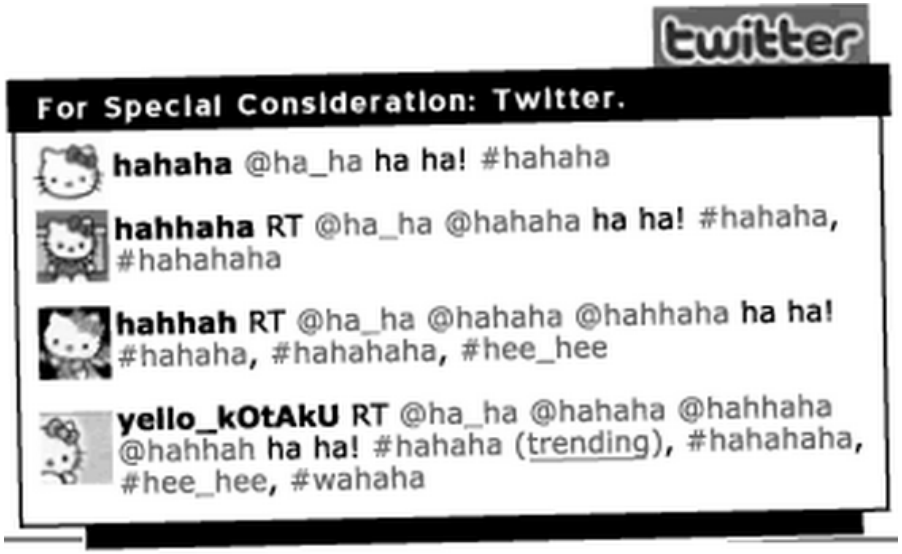
What does social media language look like?



(**Achewood**, January 17, 2010)

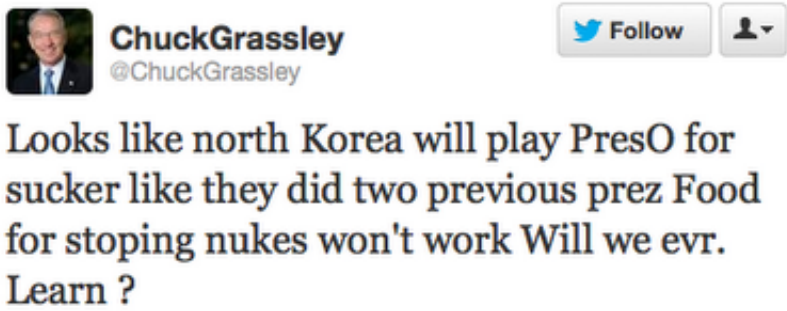
- Traditionally, non-standard language has been strongly inhibited in writing.
- In social media, that is no longer true.

What does social media language look like?



(**Achewood**, January 17, 2010)

- Traditionally, non-standard language has been strongly inhibited in writing.
- In social media, that is no longer true.



Three studies in computational sociolinguistics

The remainder of this talk describes three large-scale studies of the language variation.

- **Language**: we use machine learning to discover linguistic variables, rather than defining the them a priori. We focus on **lexical** variables.

Three studies in computational sociolinguistics

The remainder of this talk describes three large-scale studies of the language variation.

- **Language**: we use machine learning to discover linguistic variables, rather than defining the them a priori. We focus on **lexical** variables.
- **Speaker**: various sources of metadata define speaker characteristics. We strive for a flexible, multifaceted treatment of speaker identity.

Three studies in computational sociolinguistics

The remainder of this talk describes three large-scale studies of the language variation.

- **Language**: we use machine learning to discover linguistic variables, rather than defining the them a priori. We focus on **lexical** variables.
- **Speaker**: various sources of metadata define speaker characteristics. We strive for a flexible, multifaceted treatment of speaker identity.
- **Context**: public Twitter messages.

Geographical Language Variation

A Latent Variable Model for Geographical Lexical Variation
Eisenstein, O'Connor, Smith, and Xing. EMNLP 2010.

Geographical Language Variation

A Latent Variable Model for Geographical Lexical Variation
Eisenstein, O'Connor, Smith, and Xing. EMNLP 2010.

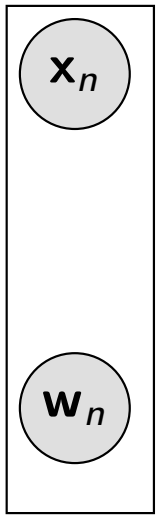
- Does language display geographical variation in social media?
- If so, does it match spoken language variation?
- What are the main linguistic divisions of the United States?
- Can we predict where people are from based on only their text?

Dataset

- 9250 authors with GPS locations
- 380K messages from one week in March 2010
- 4.9M tokens
- Vocabulary limited to 5000 words (expanded later)
- Filters
  - At least 20 messages (in Gardenhose)
  - Must include GPS within a USA zipcode
  - No more than 1000 followers, followees

A mixture model for dialect

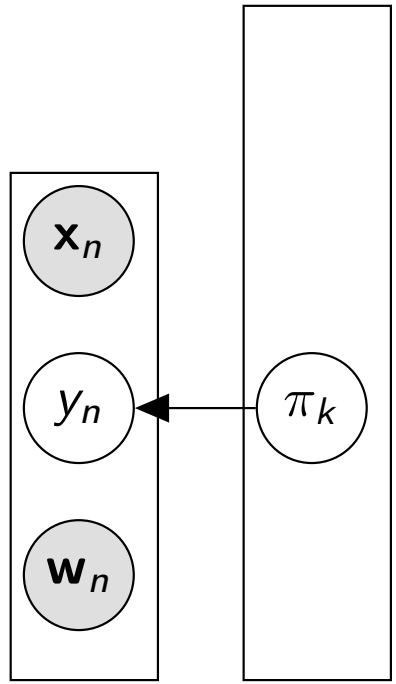
- For each author, we observe text $\mathbf{w}_n$ and GPS $\mathbf{x}_n$. Imagine these observations are generated from a random process:



A mixture model for dialect

- For each author, we observe text $\mathbf{w}_n$ and GPS $\mathbf{x}_n$. Imagine these observations are generated from a random process:
  - For each author n, draw latent region $y_n \sim \pi$,

    where π_k is the prior likelihood of region k.



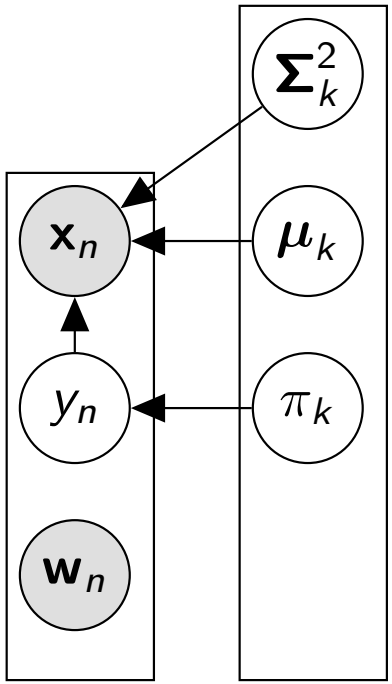
A mixture model for dialect

- For each author, we observe text $\mathbf{w}_n$ and GPS $\mathbf{x}_n$. Imagine these observations are generated from a random process:
  - For each author n, draw latent region $y_n \sim \pi$,

    where π_k is the prior likelihood of region k.

  - Draw the location $\mathbf{x}_n \sim \mathcal{N}(\mu_{y_n}, \Sigma^2_{y_n})$,

    where μ_{y_n} and $\Sigma^2_{y,n}$ are the spatial mean and covariance for region y_n.



A mixture model for dialect

- For each author, we observe text $\mathbf{w}_n$ and GPS $\mathbf{x}_n$. Imagine these observations are generated from a random process:
  - For each author n, draw latent region $y_n \sim \pi$,

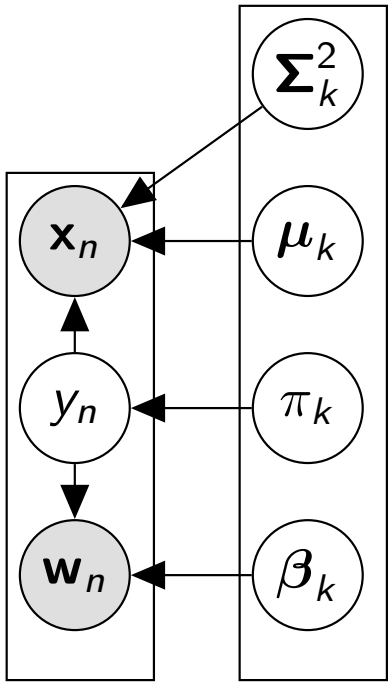
    where π_k is the prior likelihood of region k.

  - Draw the location $\mathbf{x}_n \sim \mathcal{N}(\mu_{y_n}, \Sigma^2_{y_n})$,

    where μ_{y_n} and $\Sigma^2_{y,n}$ are the spatial mean and covariance for region y_n.

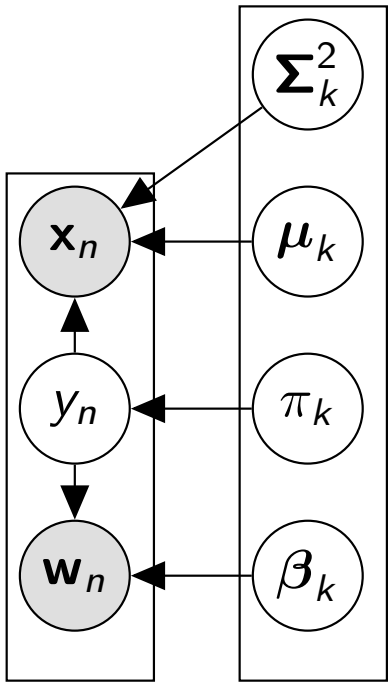
  - Draw the text $\mathbf{w}_n \sim \beta_{y_n}$,

    where β_{y_n} is a multinomial word distribution for y_n.



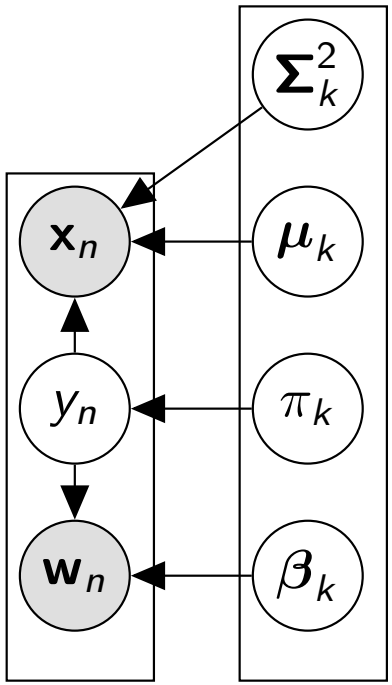
Inference

- We have defined a **generative process** that explains our observed data.
- We now want to recover distributions over the hidden variables: $\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^2$.



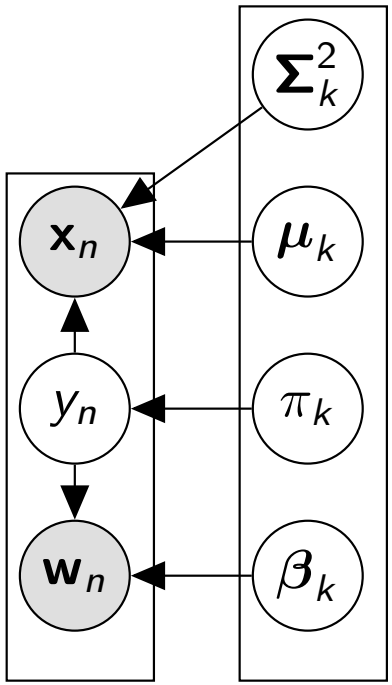
Inference

- We have defined a **generative process** that explains our observed data.
- We now want to recover distributions over the hidden variables: $\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^2$.
- This is accomplished through Bayesian inference over the joint distribution $P(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^2 | \mathbf{w}, \mathbf{x})$.



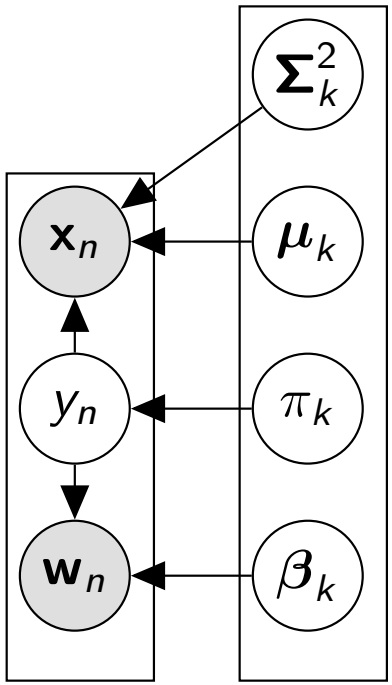
Inference

- We have defined a **generative process** that explains our observed data.
- We now want to recover distributions over the hidden variables: $\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^2$.
- This is accomplished through Bayesian inference over the joint distribution $P(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^2 | \mathbf{w}, \mathbf{x})$.
  - **Gibbs sampling**: randomly sample from the posterior for each variable. Iterate until convergence.



Inference

- We have defined a **generative process** that explains our observed data.
- We now want to recover distributions over the hidden variables: $\mathbf{y}, \beta, \mu, \mathbf{\Sigma}^2$.
- This is accomplished through Bayesian inference over the joint distribution $P(\mathbf{y}, \beta, \mu, \mathbf{\Sigma}^2 | \mathbf{w}, \mathbf{x})$.
  - **Gibbs sampling**: randomly sample from the posterior for each variable. Iterate until convergence.
  - **Variational inference**: design a tractable approximation, and optimize its similarity to the true distribution.



Adding topics

- The mixture model assumes all lexical differences are either geographical, or IID noise.
  - This makes the predictions less accurate, and confuses feature analysis.
  - For example, LA has more Spanish speakers than NYC, so Spanish words become associated with LA.

Adding topics

- The mixture model assumes all lexical differences are either geographical, or IID noise.
  - This makes the predictions less accurate, and confuses feature analysis.
  - For example, LA has more Spanish speakers than NYC, so Spanish words become associated with LA.
- To account for non-geographical variation, we add **latent topics**: groups of words which are used by the same authors.

Adding topics

- The mixture model assumes all lexical differences are either geographical, or IID noise.
  - This makes the predictions less accurate, and confuses feature analysis.
  - For example, LA has more Spanish speakers than NYC, so Spanish words become associated with LA.
- To account for non-geographical variation, we add **latent topics**: groups of words which are used by the same authors.
  - album, music, beats, artist, video, #lakers, itunes, tour
  - bieber, justin, gaga, jonas, pants, beiber, ring, annoying
  - da, dat, dis, wat, dats, dey, gud, watz, wats

Adding topics

The observations are still the text $\mathbf{w}_n$ and GPS $\mathbf{x}_n$.
Again, imagine these observations are generated from a random process.

Adding topics

The observations are still the text $\mathbf{w}_n$ and GPS $\mathbf{x}_n$.
Again, imagine these observations are generated from a random process.

- For each author n
  - draw latent region $y_n \sim \pi_k$
  - draw latent topic vector $\boldsymbol{\theta}_n \sim \text{Dirichlet}(\alpha)$
  - draw location $\mathbf{x}_n \sim \mathcal{N}(\mu_{y_n}, \Sigma^2_{y_n})$.
  - For each token t,
    - draw topic $z_t^{(n)} \sim \boldsymbol{\theta}_n$
    - draw text $w_t^{(n)} \sim P(w_n | \beta^{(T)}_{z_t}, \beta^{(R)}_{y_t})$.

Adding topics

The observations are still the text $\mathbf{w}_n$ and GPS $\mathbf{x}_n$.
Again, imagine these observations are generated from a random process.

- For each author n
  - draw latent region $y_n \sim \pi_k$
  - draw latent topic vector $\boldsymbol{\theta}_n \sim \text{Dirichlet}(\alpha)$
  - draw location $\mathbf{x}_n \sim \mathcal{N}(\mu_{y_n}, \Sigma^2_{y_n})$.
  - For each token t,
    - draw topic $z_t^{(n)} \sim \boldsymbol{\theta}_n$
    - draw text $w_t^{(n)} \sim P(w_n | \beta^{(T)}_{z_t}, \beta^{(R)}_{y_t})$.
- We must now maximize an approximation to $P(y, \beta, \boldsymbol{\mu}, \boldsymbol{\Sigma}^2, \mathbf{z}, \boldsymbol{\theta} | \mathbf{w}, \mathbf{x})$.
- This is more work, but it's worth it.

Predictive accuracy

We can predict the location of unlabeled authors by summing over regions:

$$\hat{x}_n = \arg\max_x \sum_k P(x|\mu_k, \Sigma_k^2) Pr(y_n = k|\mathbf{w}_n, \boldsymbol{\beta}_\mathbf{k})$$

error in kilometers $\rightarrow$	**mean**	**median**
population center	1148	1018

Predictive accuracy

We can predict the location of unlabeled authors by summing over regions:

$$\hat{x}_n = \arg\max_x \sum_k P(x|\mu_k, \Sigma_k^2) Pr(y_n = k|\mathbf{w}_n, \boldsymbol{\beta}_\mathbf{k})$$

error in kilometers $\rightarrow$	**mean**	**median**
population center	1148	1018
text regression	948	712
supervised LDA	1055	728

Predictive accuracy

We can predict the location of unlabeled authors by summing over regions:

$$\hat{x}_n = \arg\max_x \sum_k P(x|\mu_k, \Sigma_k^2) Pr(y_n = k|\mathbf{w}_n, \boldsymbol{\beta}_\mathbf{k})$$

error in kilometers $\rightarrow$	**mean**	**median**
population center	1148	1018
text regression	948	712
supervised LDA	1055	728
mixture model	947	644
+topics	900	494

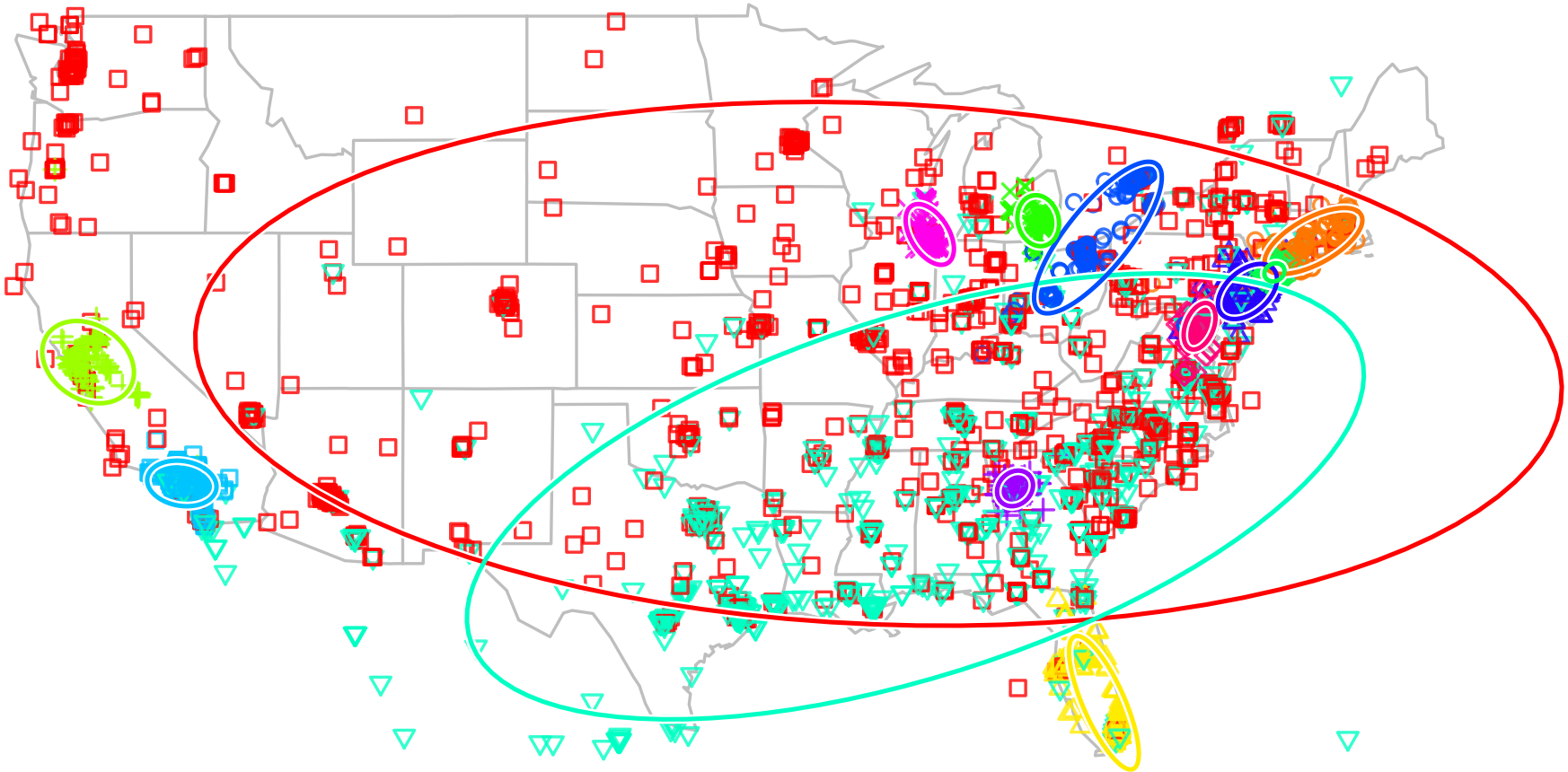
Predictive accuracy

We can predict the location of unlabeled authors by summing over regions:

$$\hat{x}_n = \arg\max_x \sum_k P(x|\mu_k, \Sigma_k^2) Pr(y_n = k|\mathbf{w}_n, \boldsymbol{\beta}_\mathbf{k})$$

error in kilometers →	**mean**	**median**
population center	1148	1018
text regression	948	712
supervised LDA	1055	728
mixture model	947	644
+topics	900	494
+sparsity [EAX11]	845	501
+larger vocab	791	461

Text+geography model output



Text+geography model output

For each cluster,[1] we rank words by log-odds: $\log \beta_i - \log \frac{1}{K} \sum_j \beta_j$:

- **New York**: brib, lml, wassupp, uu, werd, deadass, flatbush, odee, dha
- **So. Cal**: disneyland, cuh, fucken, af, fasho, faded, wyd, freeway, bomb
- **No. Cal**: sac, oakland, sf, hella, warriors, pleasure, bay, koo
- **Atlanta**: atlanta, atl, georgia, ga, $1, waffle, af, nun, shawty
- **Cleveland/Detroit**: ctfu, detroit, foolin, .!!, cleveland, geeked, salty, ikr
- **Northwest**: seattle, portland, oregon, olympic, heh, canada, stoked

[1] nb: clusters do not match previous slide.

Text+geography model output

For each cluster,[1] we rank words by log-odds: $\log \beta_i - \log \frac{1}{K} \sum_j \beta_j$:

- **New York**: brib, lml, wassupp, uu, werd, deadass, flatbush, odee, dha
- **So. Cal**: disneyland, cuh, fucken, af, fasho, faded, wyd, freeway, bomb
- **No. Cal**: sac, oakland, sf, hella, warriors, pleasure, bay, koo
- **Atlanta**: atlanta, atl, georgia, ga, $1, waffle, af, nun, shawty
- **Cleveland/Detroit**: ctfu, detroit, foolin, .!!, cleveland, geeked, salty, ikr
- **Northwest**: seattle, portland, oregon, olympic, heh, canada, stoked

[1] nb: clusters do not match previous slide.

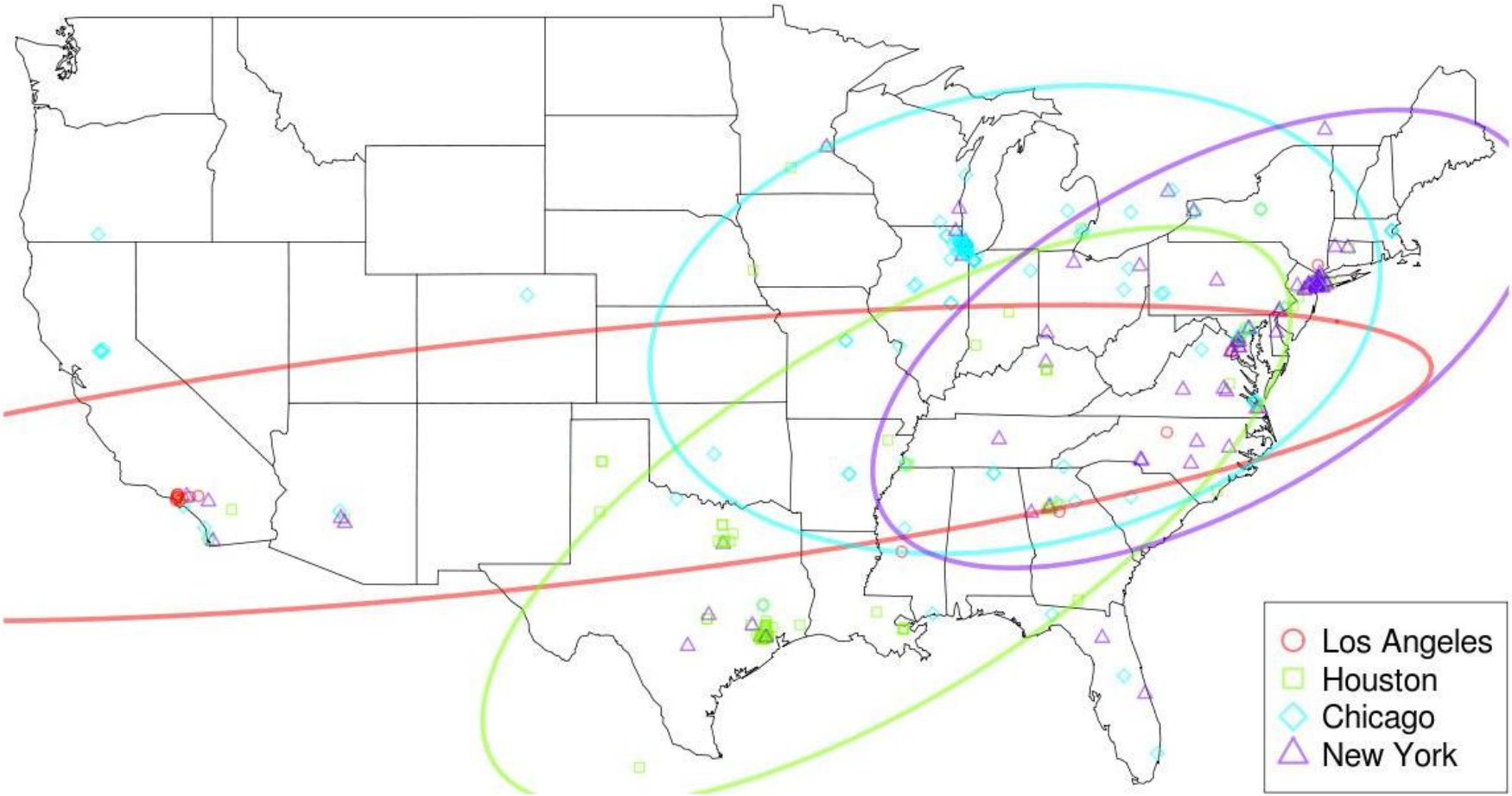
Text+geography model output

For each cluster,[1] we rank words by log-odds: $\log \beta_i - \log \frac{1}{K} \sum_j \beta_j$:

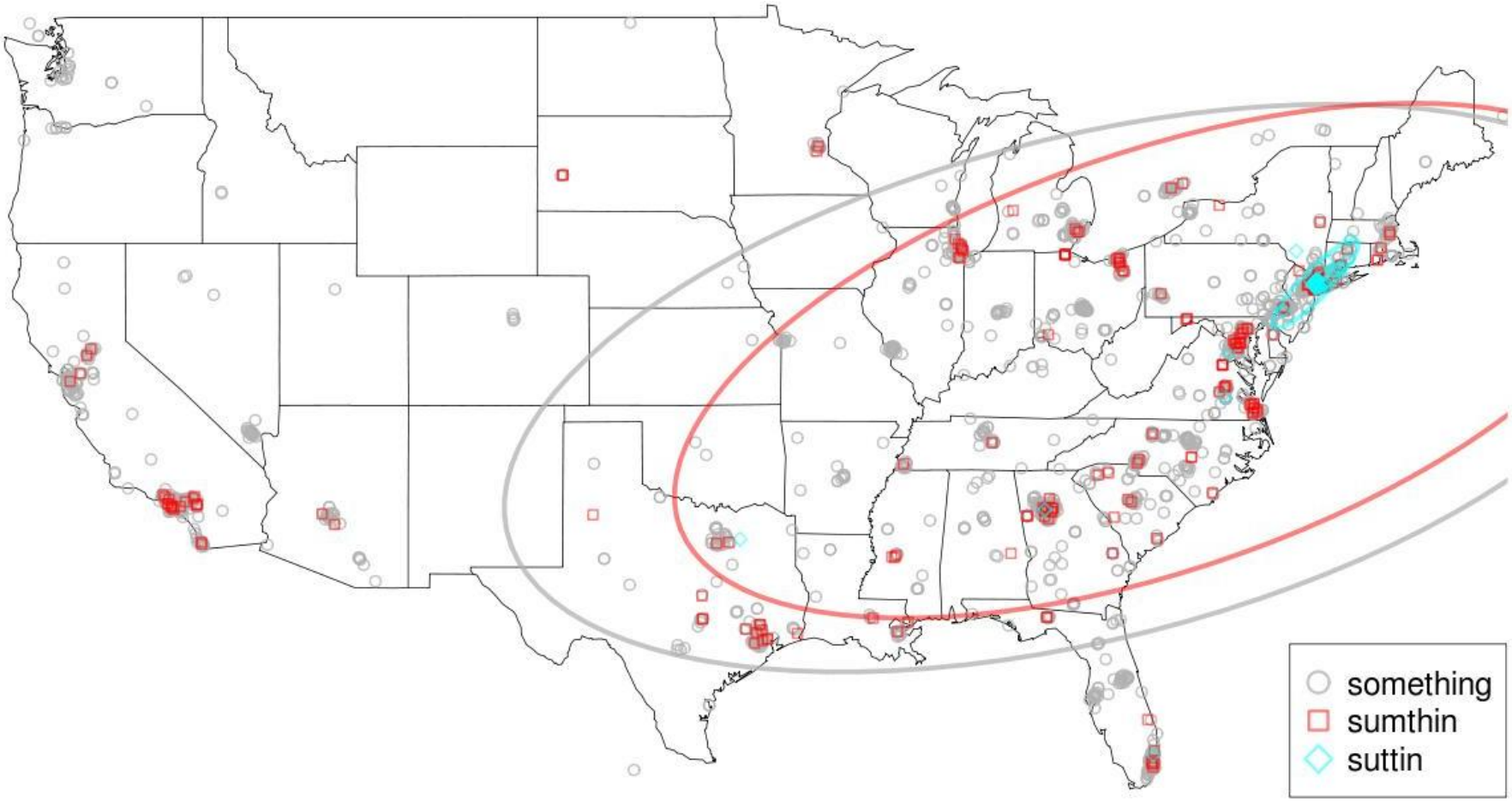
- **New York**: brib, lml, wassupp, uu, werd, deadass, flatbush, odee, dha
- **So. Cal**: disneyland, cuh, fucken, af, fasho, faded, wyd, freeway, bomb
- **No. Cal**: sac, oakland, sf, hella, warriors, pleasure, bay, koo
- **Atlanta**: atlanta, atl, georgia, ga, $1, waffle, af, nun, shawty
- **Cleveland/Detroit**: ctfu, detroit, foolin, .!!, cleveland, geeked, salty, ikr
- **Northwest**: seattle, portland, oregon, olympic, heh, canada, stoked

[1] nb: clusters do not match previous slide.

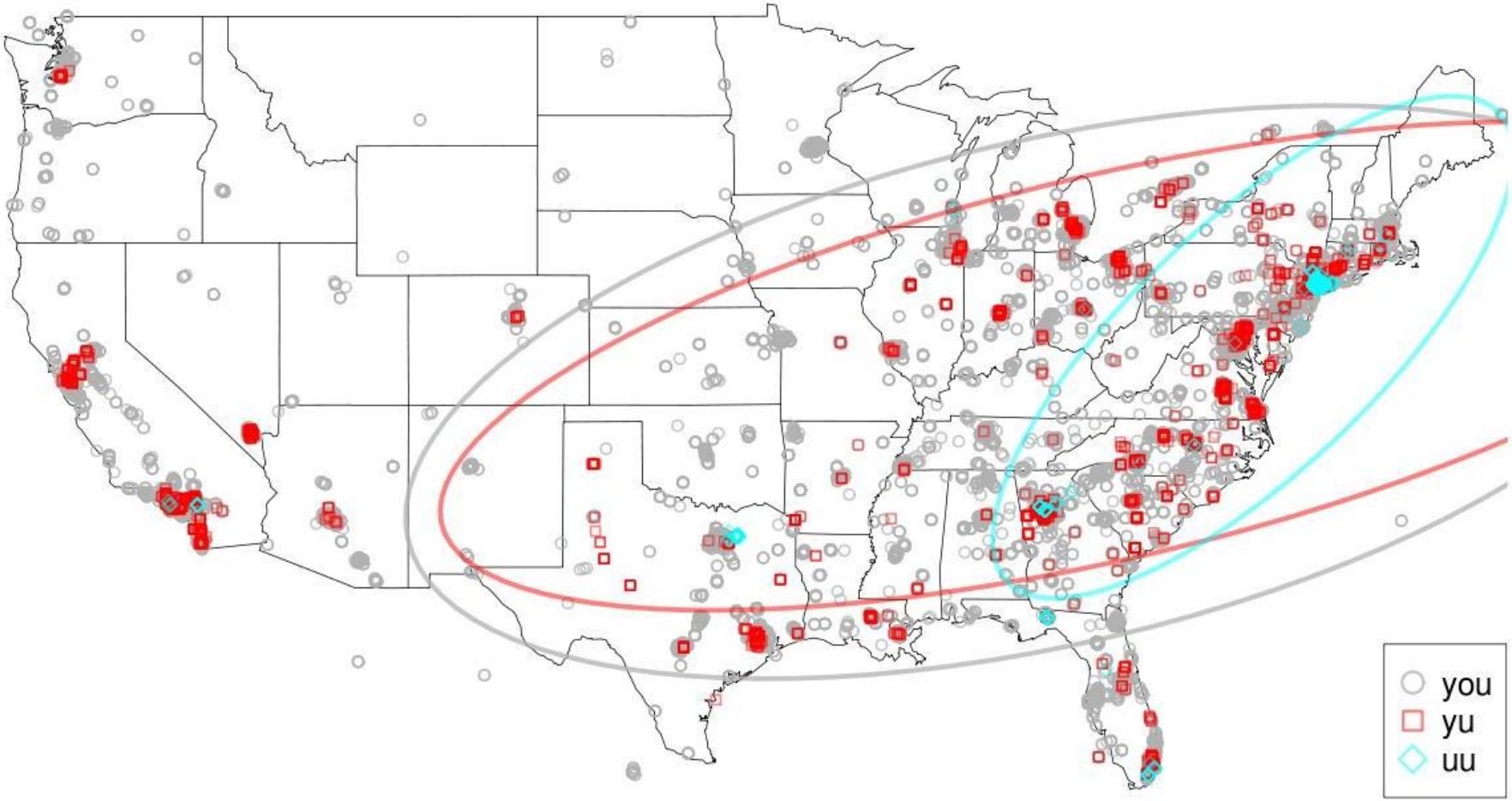
Mentions of city names



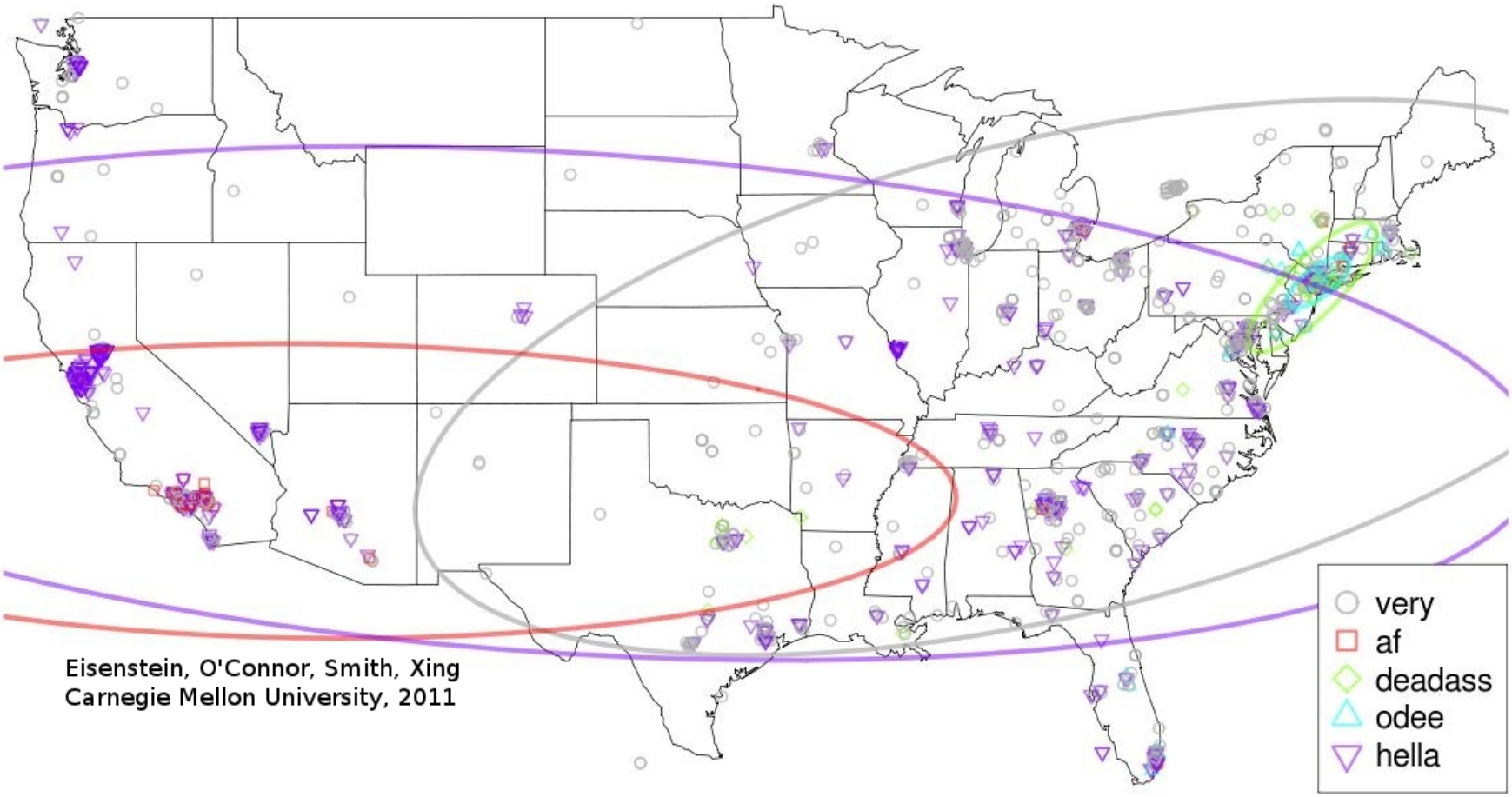
Something & variants



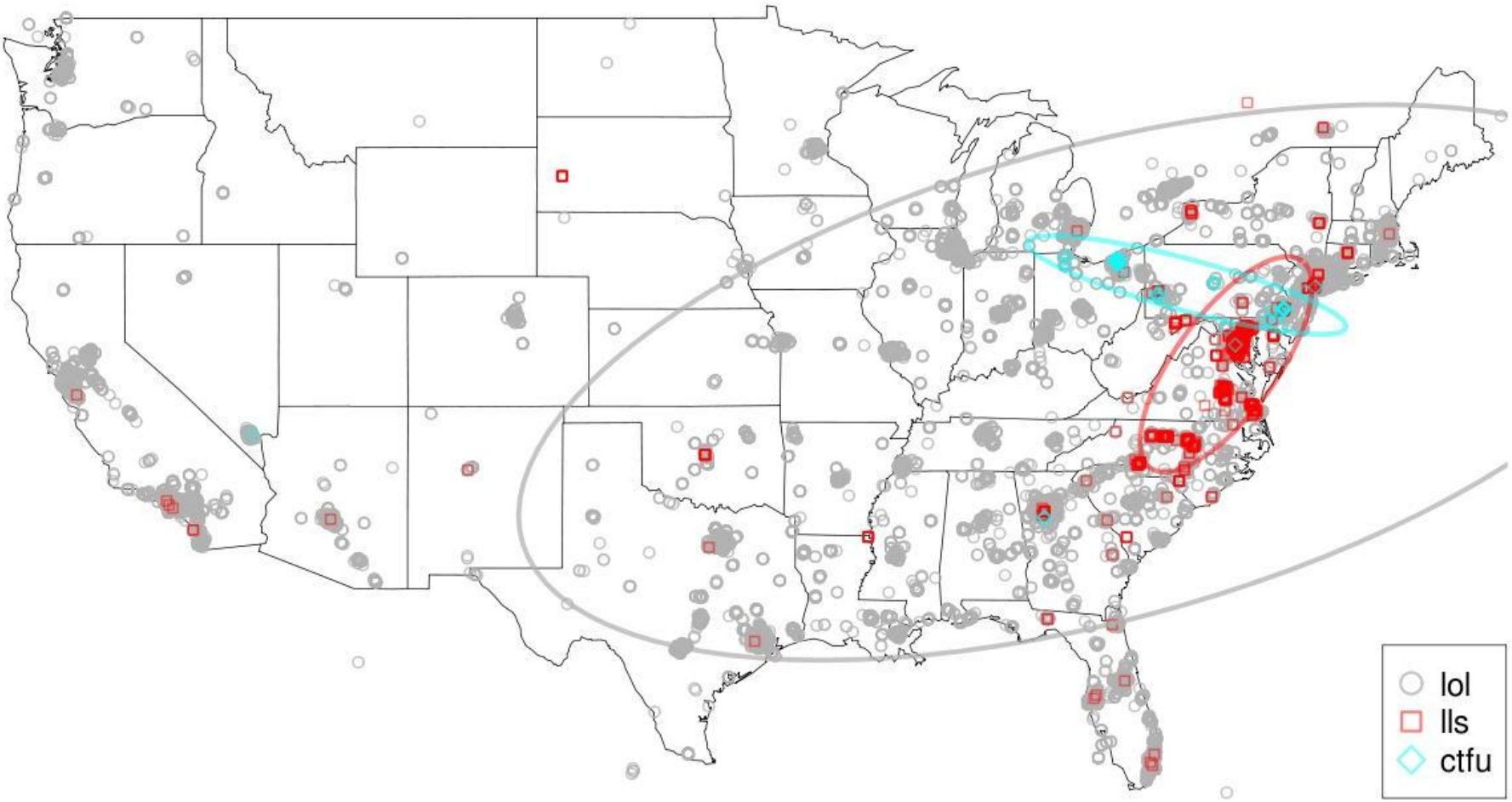
You & variants



Intensifiers



LOL & variants



Overview of geographical variation

- Social media introduces new lexical variables which are not possible in speech, e.g. ctfu, koo, uu
- The geographical distribution of these variables may not follow speech dialect regions.
- A more flexible models of author identity is both more plausible and more accurate.

Demographic language variation

Discovering Sociolinguistic Associations with Structured Sparsity, Eisenstein, Smith, and Xing. ACL 2011.

Demographic language variation

Discovering Sociolinguistic Associations with Structured Sparsity, Eisenstein, Smith, and Xing. ACL 2011.

- Los Angeles dialect: coo, af, wyd, fasho, bomb
  My LA in-laws don't use any of these words, ever.

Demographic language variation

Discovering Sociolinguistic Associations with Structured Sparsity, Eisenstein, Smith, and Xing. ACL 2011.

- Los Angeles dialect: coo, af, wyd, fasho, bomb
  My LA in-laws don't use any of these words, ever.
- Geographical variation is modulated by other factors:
  - Social class in New York [Lab66]
  - High school cliques in Detroit [Eck89]
  - Farming versus herding in Ucieda [Hol85]

Demographic language variation

Discovering Sociolinguistic Associations with Structured Sparsity, Eisenstein, Smith, and Xing. ACL 2011.

- Los Angeles dialect: coo, af, wyd, fasho, bomb
  My LA in-laws don't use any of these words, ever.
- Geographical variation is modulated by other factors:
  - Social class in New York [Lab66]
  - High school cliques in Detroit [Eck89]
  - Farming versus herding in Ucieda [Hol85]
- Such variation is governed by the speaker's sense of *identity*, but identity is a complex combination of demographic factors.

Demographic language variation

Discovering Sociolinguistic Associations with Structured Sparsity, Eisenstein, Smith, and Xing. ACL 2011.

- Los Angeles dialect: coo, af, wyd, fasho, bomb
  My LA in-laws don't use any of these words, ever.
- Geographical variation is modulated by other factors:
  - Social class in New York [Lab66]
  - High school cliques in Detroit [Eck89]
  - Farming versus herding in Ucieda [Hol85]
- Such variation is governed by the speaker's sense of *identity*, but identity is a complex combination of demographic factors.

Our goal: rather than directly model personal identity, can we induce a vocabulary of lexical variables **jointly** across many demographic attributes.

Data

We use the same Twitter data from the geography study, but augment it with demographic attributes.

Data

We use the same Twitter data from the geography study, but augment it with demographic attributes.

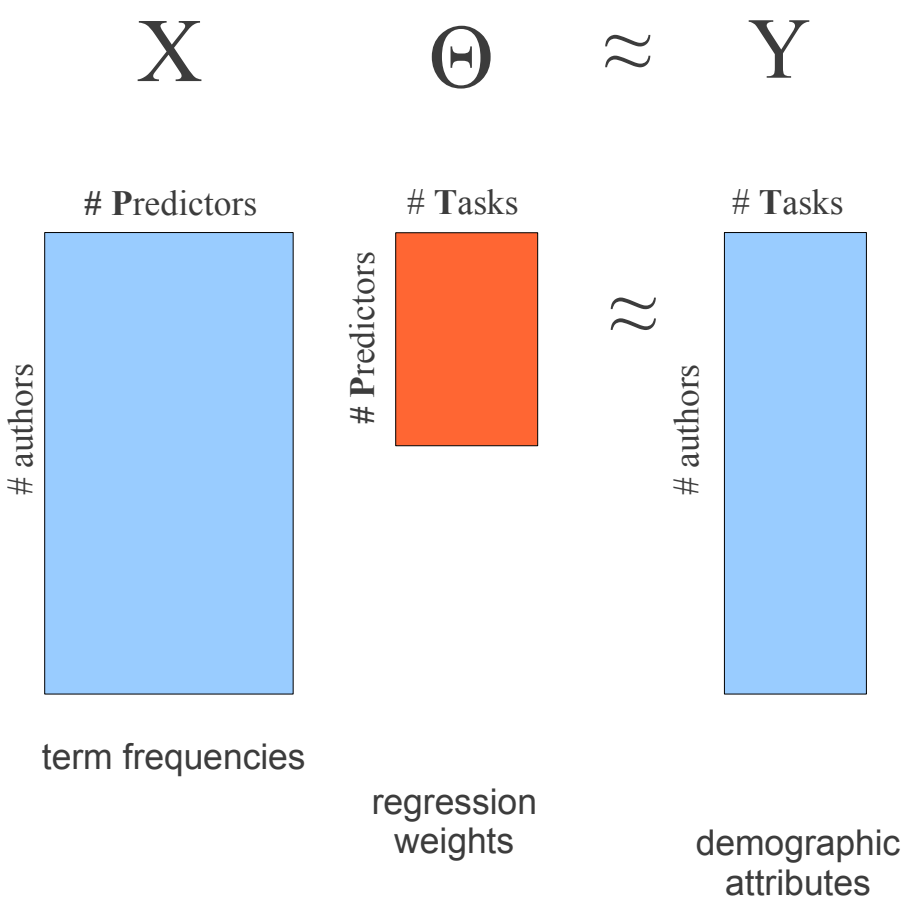
	mean	std. dev.
race & ethnicity		
% white	52.1	29.0
% African American	32.2	29.1
% Hispanic	15.7	18.3
language		
% English speakers	73.7	18.4
% Spanish speakers	14.6	15.6
% other language speakers	11.7	9.2
socioeconomic		
% urban	95.1	14.3
% with family	64.1	14.4
% renters	48.9	23.4
median income ($)	42,500	18,100

- We are using zipcode to proxy for demographics.
- This is standard in public health research, but...
- Careful! Twitter users are not an unbiased sample from a zipcode.

Data summary

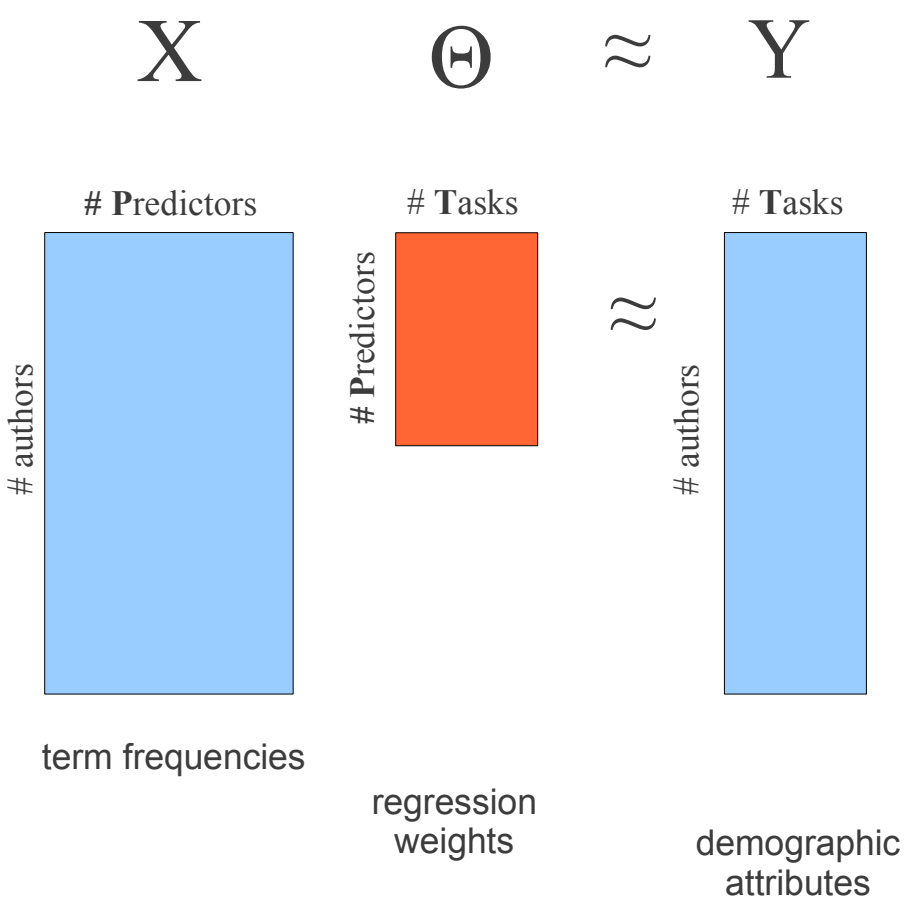
	mean	std. dev.
race & ethnicity		
% white	52.1	29.0
% African American	32.2	29.1
% Hispanic	15.7	18.3
language		
% English speakers	73.7	18.4
% Spanish speakers	14.6	15.6
% other language speakers	11.7	9.2
socioeconomic		
% urban	95.1	14.3
% with family	64.1	14.4
% renters	48.9	23.4
median income ($)	42,500	18,100

Model



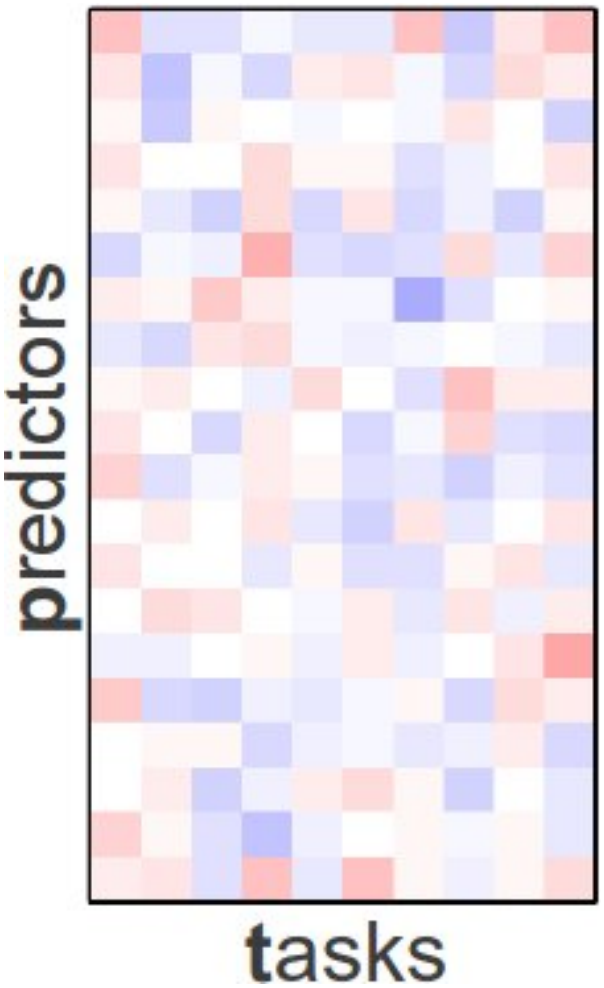
- Regress demographics against term frequencies
- $\min_{\mathbf{\Theta}} \ell(\mathbf{X\Theta} - \mathbf{Y}) + \Omega(\mathbf{\Theta})$
  - Loss: $\ell(\mathbf{X\Theta} - \mathbf{Y})$
  - Regularizer: $\Omega(\mathbf{\Theta})$

Model



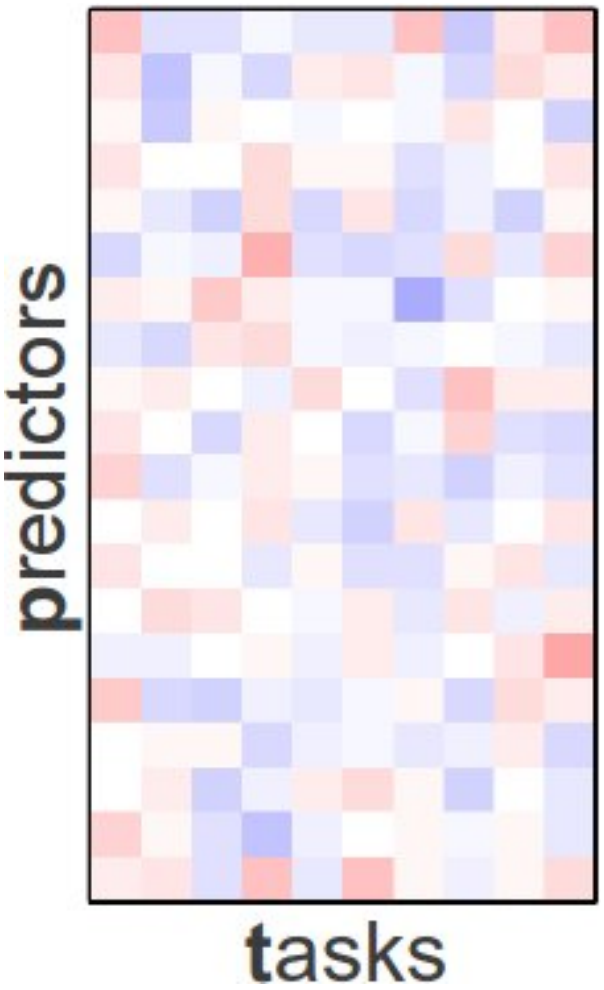
- Regress demographics against term frequencies
- $\min_{\mathbf{\Theta}} \ell(\mathbf{X\Theta} - \mathbf{Y}) + \Omega(\mathbf{\Theta})$
  - Loss: $\ell(\mathbf{X\Theta} - \mathbf{Y})$
  - Regularizer: $\Omega(\mathbf{\Theta})$
- We'll use the regularizer to make the $P \times T$ weights into an interpretable model.

Three regularizers



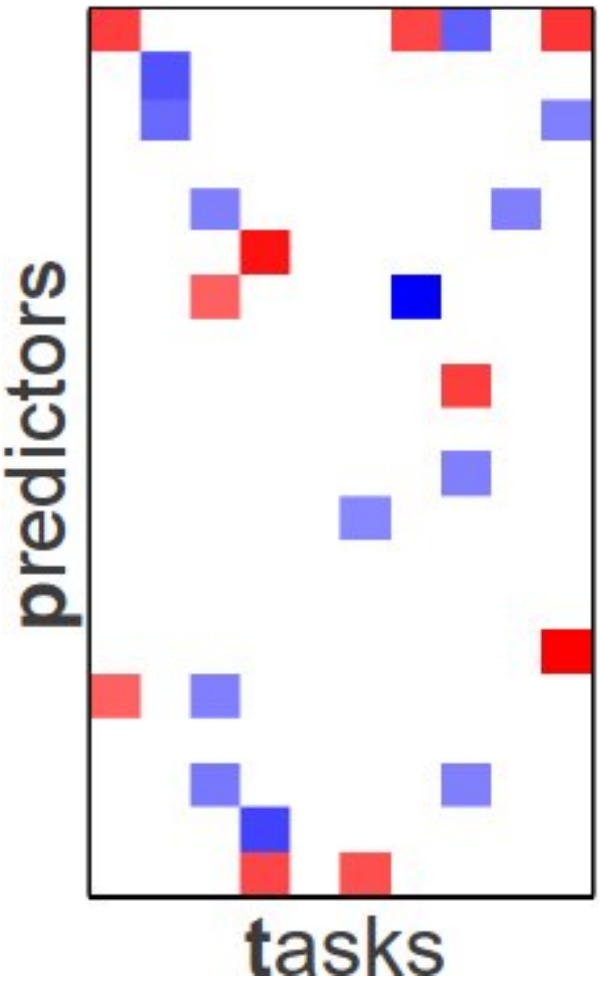
- L2 regularizer: $\Omega(\mathbf{\Theta}) = \sum_t \sum_p \theta_{pt}^2$
- encourages small regression coefficients
- equivalent to running T separate ridge regressions

Three regularizers



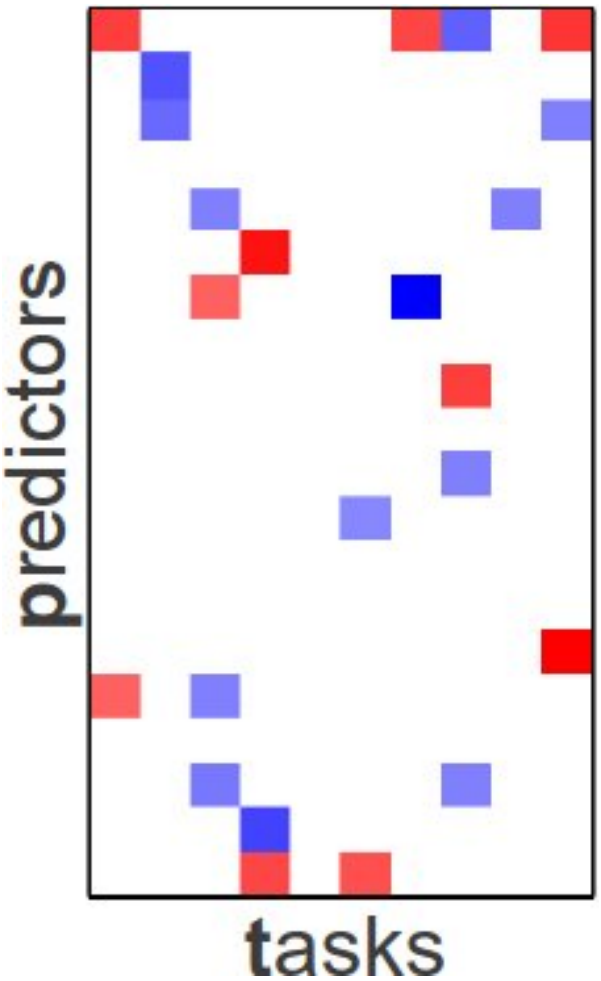
- L2 regularizer:
  $\Omega(\mathbf{\Theta}) = \sum_t \sum_p \theta_{pt}^2$
- encourages small regression coefficients
- equivalent to running T separate ridge regressions
- Interpretable? No.
  Every word-demographic association has a non-zero weight.

Three regularizers



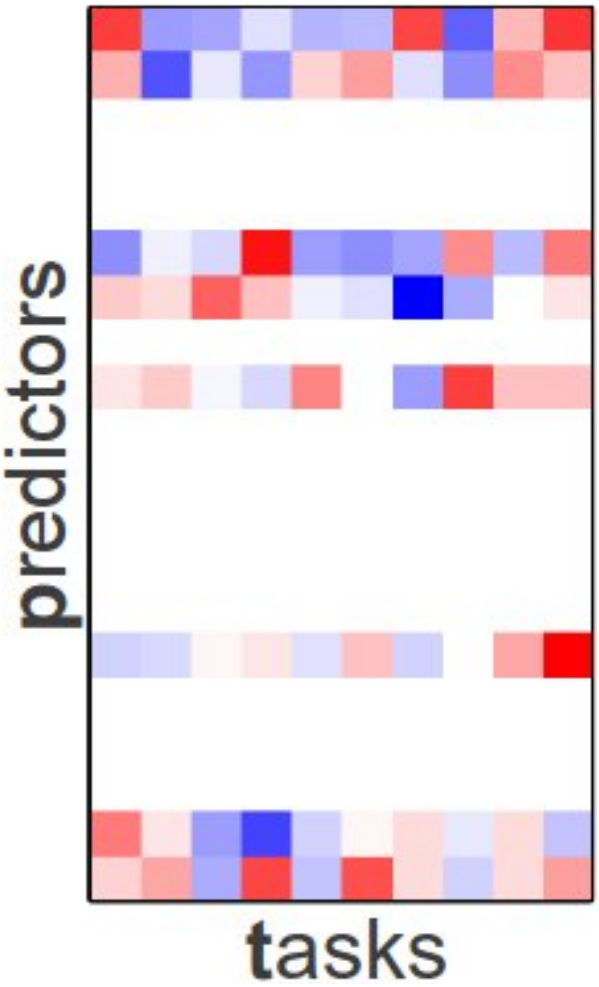
- $L1$ regularizer [FHT10]: $\Omega(\mathbf{\Theta}) = \sum_t \sum_p |\theta_{pt}|$
- encourages regression coefficients to be zero
- equivalent to running T separate lasso regressions

Three regularizers



- $L1$ regularizer [FHT10]: $\Omega(\mathbf{\Theta}) = \sum_t \sum_p |\theta_{pt}|$
- encourages regression coefficients to be zero
- equivalent to running T separate lasso regressions
- Interpretable? Sort of. L1 isolates a few robust associations, but it forces us to think about each demographic attribute separately.

Structured regularization

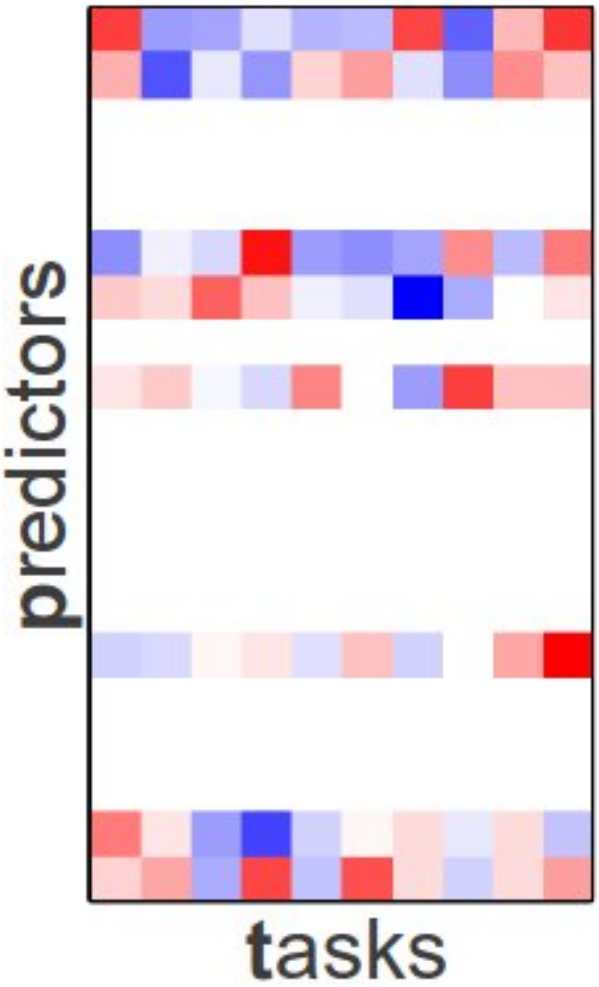


$L1/L\infty$ regularizer

$$\Omega(\mathbf{\Theta}) = \sum_t \max_p |\theta_{pt}|$$

aka **multi-output lasso** [TVW05]

Structured regularization



$L1/L\infty$ regularizer

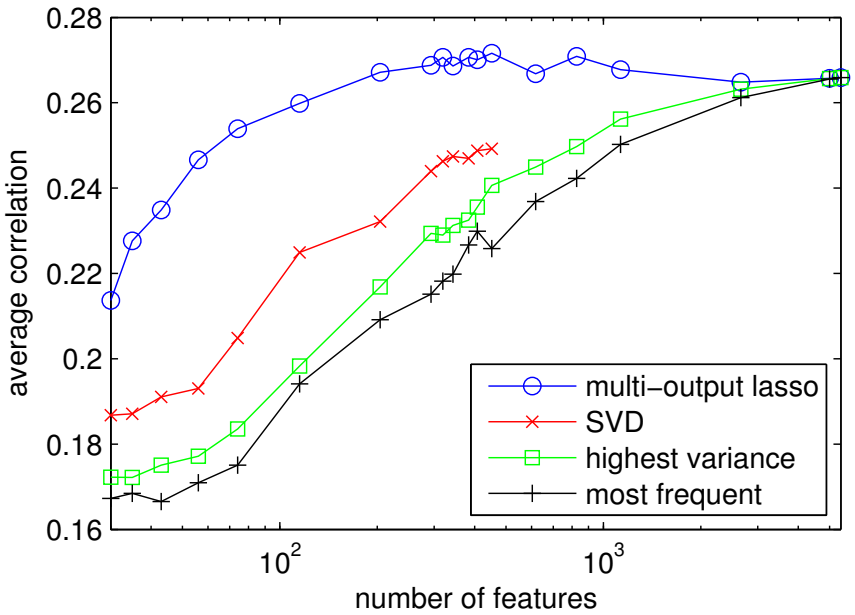
$$\Omega(\mathbf{\Theta}) = \sum_t \max_p |\theta_{pt}|$$

aka **multi-output lasso** [TVW05]

- pushes entire rows to zero
- not decomposable, must be solved jointly
- Interpretable? Yes: it identifies words with strong associations across a range of demographic attributes

Regularization path

By increasing the strength of the regularizer, we discard more terms from the vocabulary.



- Predictive accuracy remains high even when the vocabulary is reduced by more than an order of magnitude.
- Other dimensionality-reduction techniques suffer dramatically.

Predictive accuracy

vocabulary	**# features**	average	white	Afr. Am.	Hisp.
full	5418	0.260	0.337	0.318	0.296
multi-output lasso	394.6	0.260	0.326	0.308	0.304
SVD		0.237	0.321	0.299	0.269
highest variance		0.220	0.309	0.287	0.245
most frequent		0.204	0.294	0.264	0.222

- **metric**: Pearson correlation between predicted and true values of demographic attributes
- **baselines**:
  - SVD of author-term matrix
  - highest variance words
  - most frequent words

Predictive accuracy

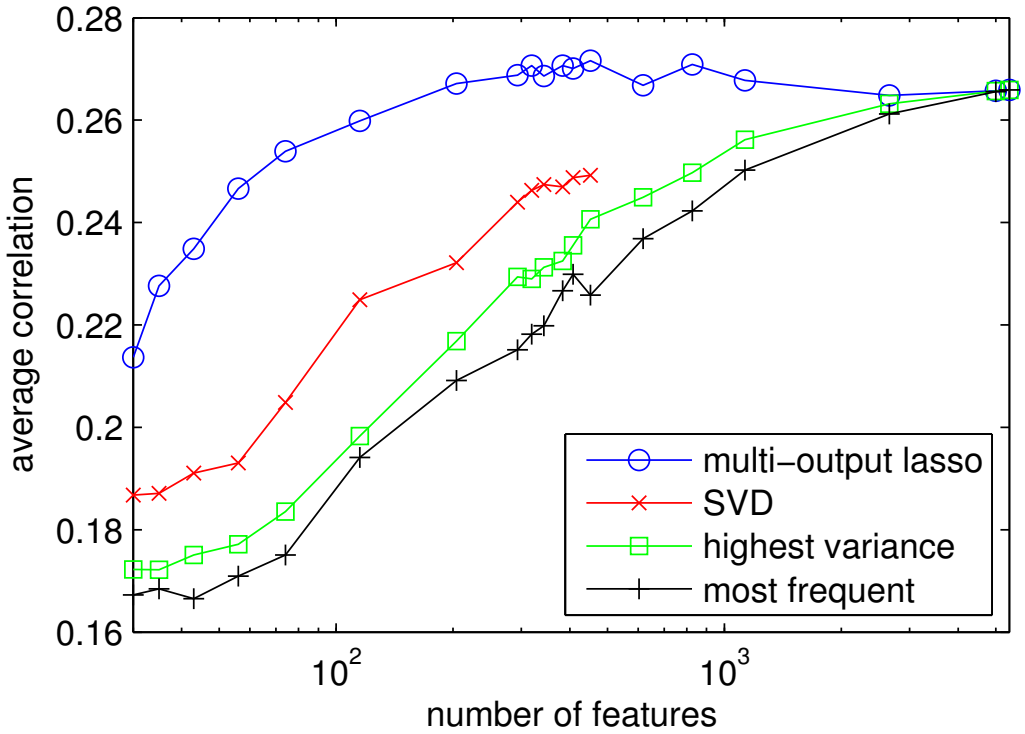
vocabulary	**# features**	average	Eng. lang.	Span. lang.	other lang.
full	5418	0.260	0.384	0.296	0.256
multi-output lasso	394.6	0.260	0.383	0.303	0.249
SVD		0.237	0.352	0.272	0.226
highest variance		0.220	0.315	0.248	0.199
most frequent		0.204	0.293	0.229	0.178

Predictive accuracy

vocabulary	**# features**	average	urban	family	renter	med. inc.
full	5418	0.260	0.155	0.113	0.295	0.152
multi-output lasso	394.6	0.260	0.153	0.113	0.302	0.156
SVD		0.237	0.138	0.081	0.278	0.136
highest variance		0.220	0.132	0.085	0.250	0.135
most frequent		0.204	0.129	0.073	0.228	0.126

- All confidence intervals are tighter than ± 0.02.
- 93% reduction in model size, no loss in performance.
- Socioeconomic attributes are more difficult to predict than race, ethnicity, and language.

Qualitative analysis



- In predictive experiments, regularization chosen on dev set
- Now we tune it to identify a more compact set of 69 terms

Place names and foreign language

	white	Afr. Am.	Hisp.	Eng. lang.	Span. lang.	other lang.	urban	family	renter	med. inc.
atlanta			-	+	-	-				
famu		+	-	+	-	-				-
harlem				-					+	
con			+	-	+				+	
la		-	+	-	+					
si		-	+	-	+					

The symbols + and - indicate significant positive or negative association, as measured by a Wald Test, with Bonferroni correction.

Dictionary Words

	white	Afr. Am.	Hisp.	Eng. lang.	Span. lang.	other lang.	urban	family	renter	med. inc.
as			-	+	-					
awesome	+	-					-		-	+
break			-	+	-	-				
campus			-	+	-	-				
dead	-	+		-	+		+		+	
hell			-	+	-	-				
shit	-								+	
train				-	+				+	
will			-	+	-					
would				+					-	

Abbreviations

	white	Afr. Am.	Hisp.	Eng. lang.	Span. lang.	other lang.	urban	family	renter	med. inc.
bbm	-	+		-		+	+		+	
lls		+	-	+	-	-				
lmaoo	-	+	+	-	+	+	+		+	
lmaooo	-	+	+	-	+	+	+		+	
lmaoooo	-	+	+	-	+	+			+	
lmfaoo	-		+	-	+	+			+	
lmfaooo	-		+	-	+	+			+	
lml	-	+	+	-	+	+	+		+	-
odee	-		+	-	+		+		+	
omw	-	+	+	-	+	+	+		+	
smfh	-	+	+	-	+	+	+		+	
smh	-	+					+		+	
w	-		+	-	+	+	+		+	

Emoticons

	white	Afr. Am.	Hisp.	Eng. lang.	Span. lang.	other lang.	urban	family	renter	med. inc.
-_-	-		+	-	+	+	+			
;)		-	+	-	+					
:(		-								
:)		-								
:d	+	-	+	-	+					

Other

	white	Afr. Am.	Hisp.	Eng. lang.	Span. lang.	other lang.	urban	family	renter	med. inc.
dats	-	+		-					+	-
deadass	-	+	+	-	+	+	+		+	
haha	+	-							-	
hahah	+	-								
hahaha	+	-							-	+
ima	-		+	-	+				+	
madd	-			-		+		+		
nah	-		+	-	+	+			+	
ova	-	+		-					+	
sis	-	+							+	
skool	-	+		-		+	+		+	-
wassup	-	+	+	-	+	+	+		+	-
wat	-	+	+	-	+	+	+		+	-
ya	-	+							+	
yall	-	+								
yep			-	+	-	-	-		-	
yoo	-	+	+	-	+	+	+		+	
yooo	-	+		-	+				+	

Overview of demographic variation

- Some observations
  - **Paralinguistic commentary** is an area of major differences between demographic groups.
    - **Emoticons** correlate with % Whites, English speakers, Hispanics
    - **Abbreviations** correlate % African Americans, Hispanics, renters
  - **Phoneticization** correlates with % African American
    - May reflect phonological features of AAE (“dats”, “wassup”, “ima”)

Overview of demographic variation

- Some observations
  - **Paralinguistic commentary** is an area of major differences between demographic groups.
    - **Emoticons** correlate with % Whites, English speakers, Hispanics
    - **Abbreviations** correlate % African Americans, Hispanics, renters
  - **Phoneticization** correlates with % African American
    - May reflect phonological features of AAE ("dats", "wassup", "ima")
- Jointly modeling many demographic attributes makes lexical analysis more robust, limiting multiple comparisons.

Gender, language and social networks

Gender in Twitter: Styles, stances, and social networks.
Bamman, Eisenstein, and Schnoebelen. In preparation.

Gender, language and social networks

Gender in Twitter: Styles, stances, and social networks.
Bamman, Eisenstein, and Schnoebelen. In preparation.

- Social networks are often homophilous with respect to gender.
- We started with a simple idea:
  use the social network to improve gender prediction.

Data

- 14,464 Twitter users (56% male)
  - geolocation in USA
  - must use 50 of 1000 most frequent words
  - no more than 1000 follow connections
- 9.2M tweets, from January to June 2011
- Author gender induced from first name and census records
- Social network induced from mutual @-mentions
  - Women have 58% female friends
  - Men have 67% male friends

Adding network features

Logistic regression, 10-fold cross-validation:

- Text alone: 89% accuracy

Adding network features

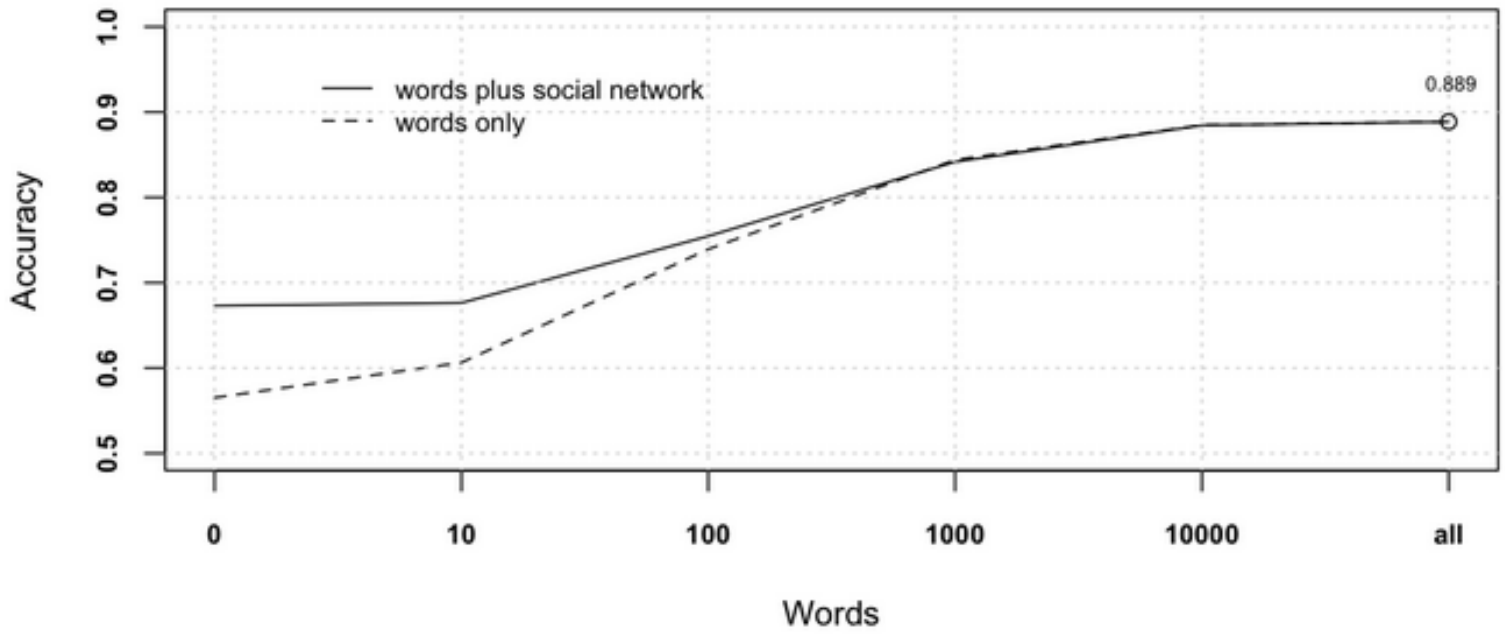
Logistic regression, 10-fold cross-validation:

- Text alone: 89% accuracy
- Text+network: 89% accurate

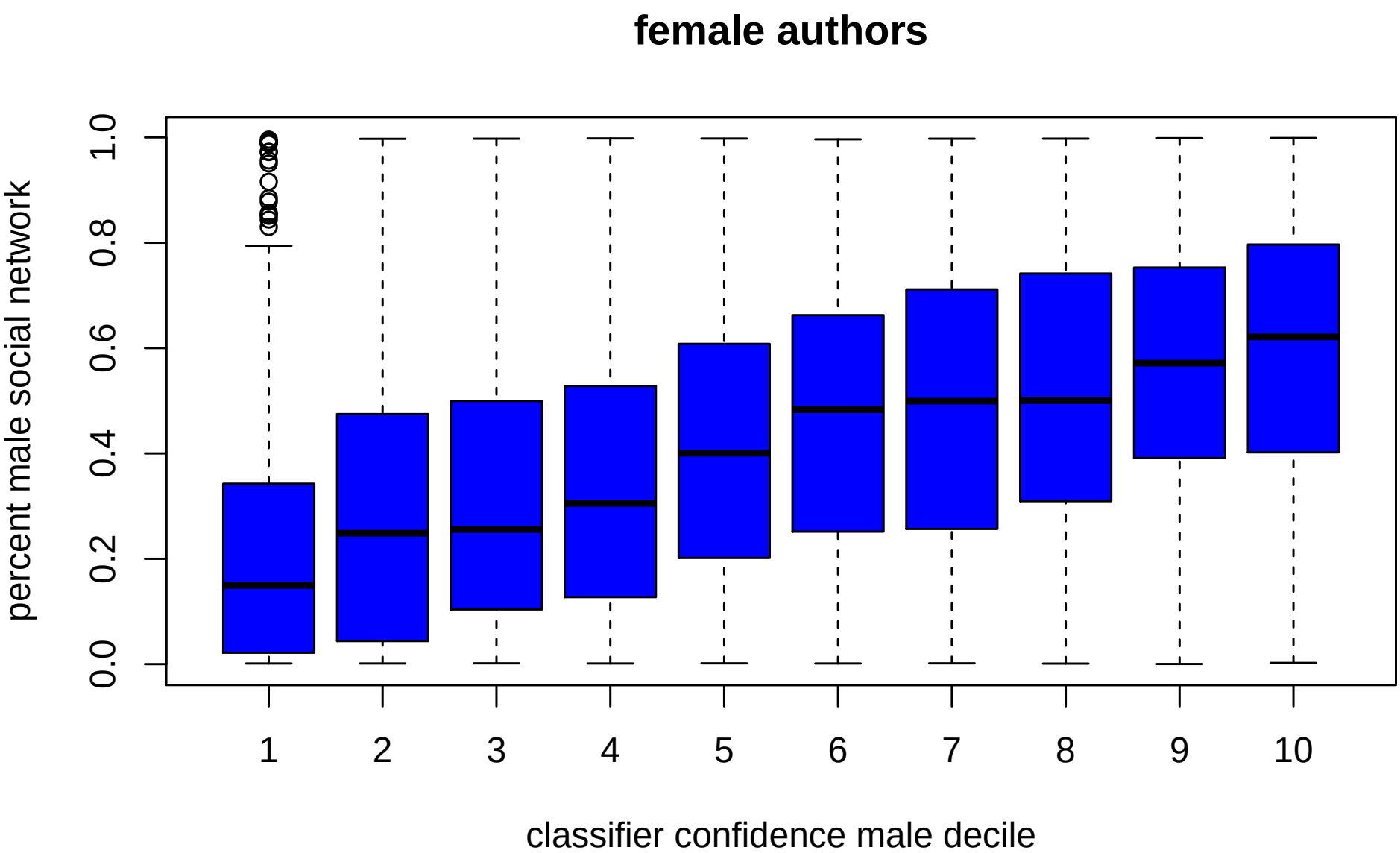
Adding network features

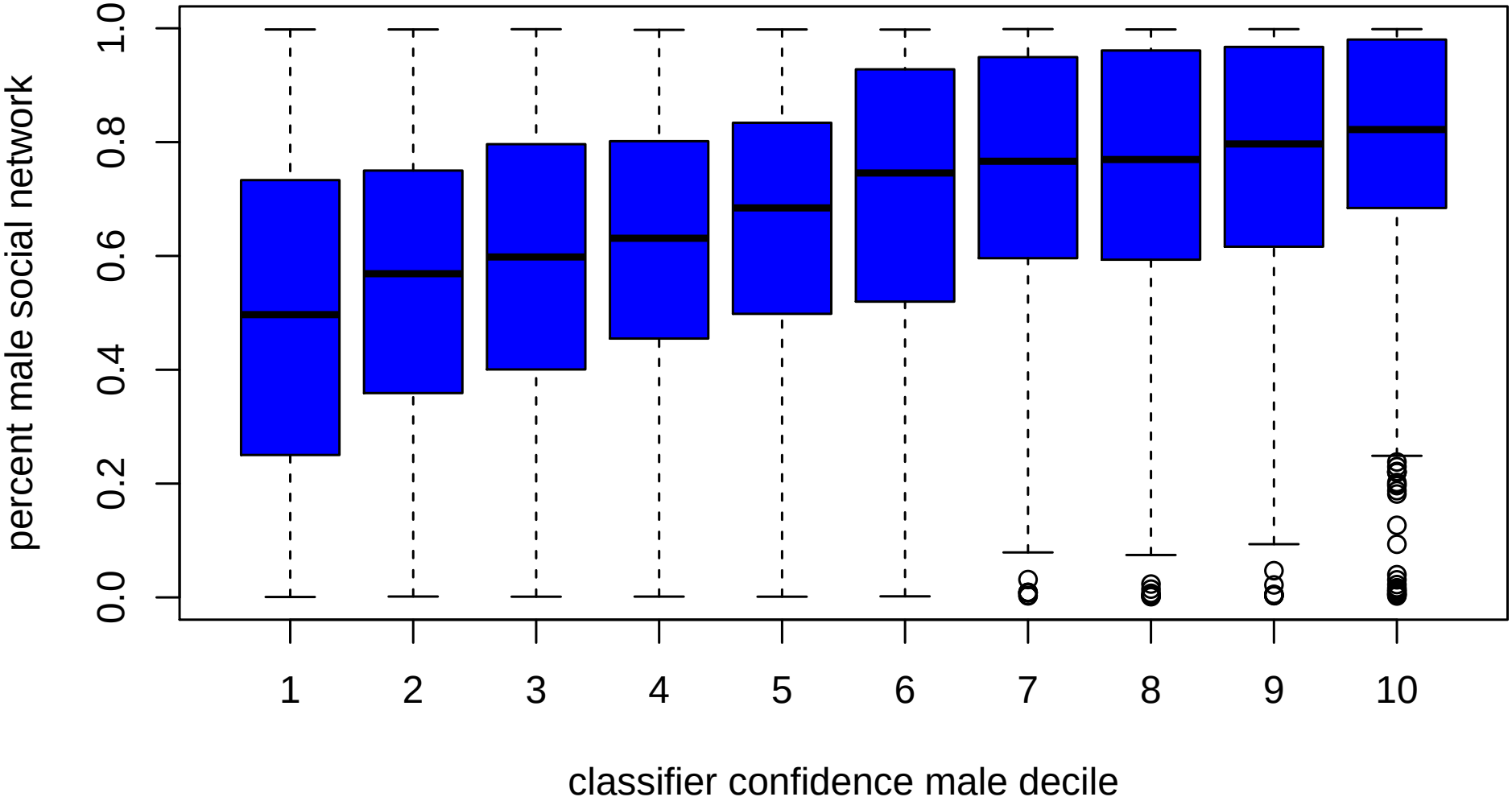
Logistic regression, 10-fold cross-validation:

- Text alone: 89% accuracy
- Text+network: 89% accurate



Once we have 1000 words per author, adding network information does not improve performance. **Why not?**


female authors
percent male social network
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
7
8
9
10
classifier confidence male decile


male authors
percent male social network
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
7
8
9
10
classifier confidence male decile

Why social network features don't help

correlation	female authors	male authors
classifier vs. network	0.38 ($.35 \leq r \leq .40$)	0.33 ($.30 \leq r \leq .36$)

Why social network features don't help

correlation	female authors	male authors
classifier vs. network	0.38 ($.35 \leq r \leq .40$)	0.33 ($.30 \leq r \leq .36$)

- Network features will improve gender classification only to the extent that they are adding new information.

Why social network features don't help

correlation	**female authors**	**male authors**
classifier vs. network	0.38 ($.35 \leq r \leq .40$)	0.33 ($.30 \leq r \leq .36$)

- Network features will improve gender classification only to the extent that they are adding new information.
- But language and social network are **not** independent views on gender as a discrete, binary variable [EMG03].

Why social network features don't help

correlation	female authors	male authors
classifier vs. network	0.38 ($.35 \leq r \leq .40$)	0.33 ($.30 \leq r \leq .36$)

- Network features will improve gender classification only to the extent that they are adding new information.
- But language and social network are **not** independent views on gender as a discrete, binary variable [EMG03].
- Language and the social network are correlated because they are noisy views on gender as a continuous space of possibilities.

Text-based gender prediction

- To determine the language features that drive the classifier's performance, we computed the 500 words that were the strongest predictor of each gender.
- We then manually divided these words into categories.

		female	male
Dictionary words		325	295
Shortenings	ballin	19	16
Named entities	ipad	13	134
Taboo/swear words	fuck	1	10
Other, pronounceable	awww	80	8
Other, unpronounceable	<3	46	23
(not categorized)		16	14

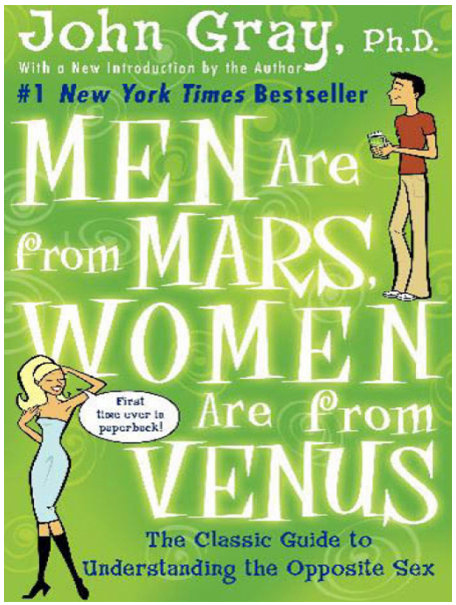
Text-based gender prediction

- To determine the language features that drive the classifier's performance, we computed the 500 words that were the strongest predictor of each gender.
- We then manually divided these words into categories.

		female	male
Dictionary words		325	295
Shortenings	ballin	19	16
Named entities	ipad	13	134
Taboo/swear words	fuck	1	10
Other, pronounceable	awww	80	8
Other, unpronounceable	<3	46	23
(not categorized)		16	14

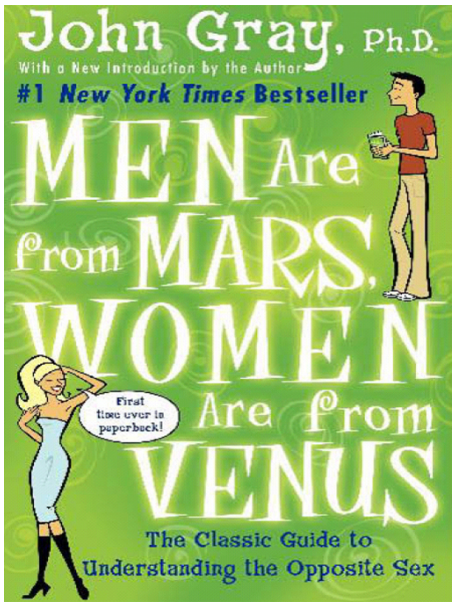
Clustering by content

- At the corpus level, women use more non-dictionary words and men mention more named entities.
- This fits snugly into an airport bookstore pop psycholinguistics of gender...



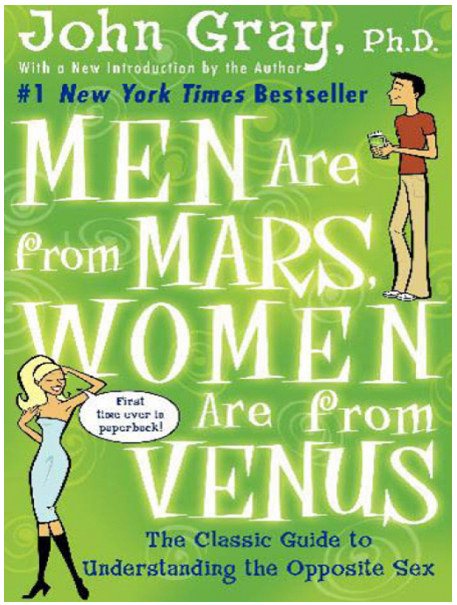
Clustering by content

- At the corpus level, women use more non-dictionary words and men mention more named entities.
- This fits snugly into an airport bookstore pop psycholinguistics of gender...
- But we have already seen that we should distrust the binary male/female opposition.



Clustering by content

- At the corpus level, women use more non-dictionary words and men mention more named entities.
- This fits snugly into an airport bookstore pop psycholinguistics of gender...
- But we have already seen that we should distrust the binary male/female opposition.
- For a closer look, we clustered all authors by text.
  - K-means ($K = 20$)
  - Clusters represent shared interests and/or styles.
  - Many clusters **happen to have** strong demographic orientations, including gender, race, and age.



Female clusters

% fem	words
0.84	fabric blogged hubs recipe recipes delish @starbucks almond howdy baking cocktails
0.79	;o xx hun xxx hump sweetie x xoxoxo cena becky
0.78	xo elizabeth gr8 -) ranked ty blessings thnx fr 2day
0.76	muah darren bo sry xoxoxo sux ,,, scotty lmbo hun
0.75	**clark pokemon ash arc #idol authors unicorns terrifying romance chapter**
0.75	:') (: <333 @justinbieber (; xxx <33333 </3 <33 ;d

Female clusters

% fem	words
0.84	fabric blogged hubs recipe recipes delish @starbucks almond howdy baking cocktails
0.79	;o xx hun xxx hump sweetie x xoxoxo cena becky
0.78	xo elizabeth gr8 -) ranked ty blessings thnx fr 2day
0.76	muah darren bo sry xoxoxo sux ,,, scotty lmbo hun
0.75	**clark pokemon ash arc #idol authors unicorns terrifying romance chapter**
0.75	:') (: <333 @justinbieber (; xxx <33333 </3 <33 ;d

- At the population level, women use few named entities and many non-dictionary words.
- But there are clusters of (mostly) women who do the opposite.

Male clusters

% fem	words
0.29	dems gop democrats unions conservative senate muslim israel liberal republicans
0.28	**niggaz shyt dats dey wats lmmfao lik dis neva lls**
0.19	e3 gears psn 360 kombat halo gaming portal console marvel
0.19	bama @darrenrovell @espn severe auburn ky #heat thunderstorm au #marchmadness
0.15	#nba mets #jets #mavs #knicks crawford @ochocinco pacers #lakers wright
0.14	api ui ios apple's developers developer dev hardware plugin interface
0.07	#nhl nhl prospect #bruins qb roster timeout 2-1 boozer 1-0

- At the population level, men use many named entities and few non-dictionary words.
- But there are clusters of men who do the opposite.

Overview of gender variation

- From a theoretical perspective, the sociolinguistics of gender has moved beyond the binary male/female opposition [EMG03]
- Our work provides new quantitative support for this theory.
  - The social network analysis supports the view of gender as a continuous – rather than binary – variable.
  - The clustering analysis supports the view that gender is a social construction, with linguistic properties that are contingent on the *local meaning* of broader gender categories.

Summary

We have examined sociolinguistics on three main axes

- **language** – what aspects of language are affected by social dynamics?
- **speakers** – how does personal identity relate to language variation?
- **context** – how does language carry social meaning in various social and linguistic contexts?

Summary

Computational social media analysis can provide a new perspective:

- **language** – automatically identifying new lexical variables through structured sparsity
- **speakers** – modeling speaker identity through latent variables (such as geographical regions, topics, and clusters).
- **context** – the next frontier?

Mary Bucholtz.
White kids: language, race, and styles of youth identity.
Cambridge University Press, 2011.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing.
Sparse additive generative models of text.
In Lise Getoor, Tobias Scheffer, Lise Getoor, and Tobias Scheffer, editors, *ICML*, pages 1041–1048. Omnipress, 2011.

Penelope Eckert.
Jocks and Burnouts: Social Categories and Identity in the High School.
Teachers College Press, 1989.

Penelope Eckert.
Three waves of variation study: The emergence of meaning in the study of variation.
working paper, 2012.

Penelope Eckert and Susan McConnell-Ginet.
Language and Gender.
Cambridge University Press, 2003.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani.
Regularization paths for generalized linear models via coordinate descent.
Journal of Statistical Software, 33(1):1–22, 2010.

J. Holmquist.
Social correlates of a linguistic variable: A study in a spanish village.
Language in Society, 14:191–203, 1985.

William Labov.
The Social Stratification of English in New York City.
Center for Applied Linguistics, 1966.

William Labov.
Principles of Linguistic Change, Volume 2: Social Factors.
Blackwell, 2001.


Berwin A. Turlach, William N. Venables, and Stephen J. Wright.
Simultaneous variable selection.
Technometrics, 47(3):349–363, 2005.