Statistical Text Analysis for Social Science

Brendan O'Connor
Machine Learning Department
Carnegie Mellon University

Thesis defense presentation, Aug. 19, 2014

SOCIAL SCIENCE

Computational Social Science

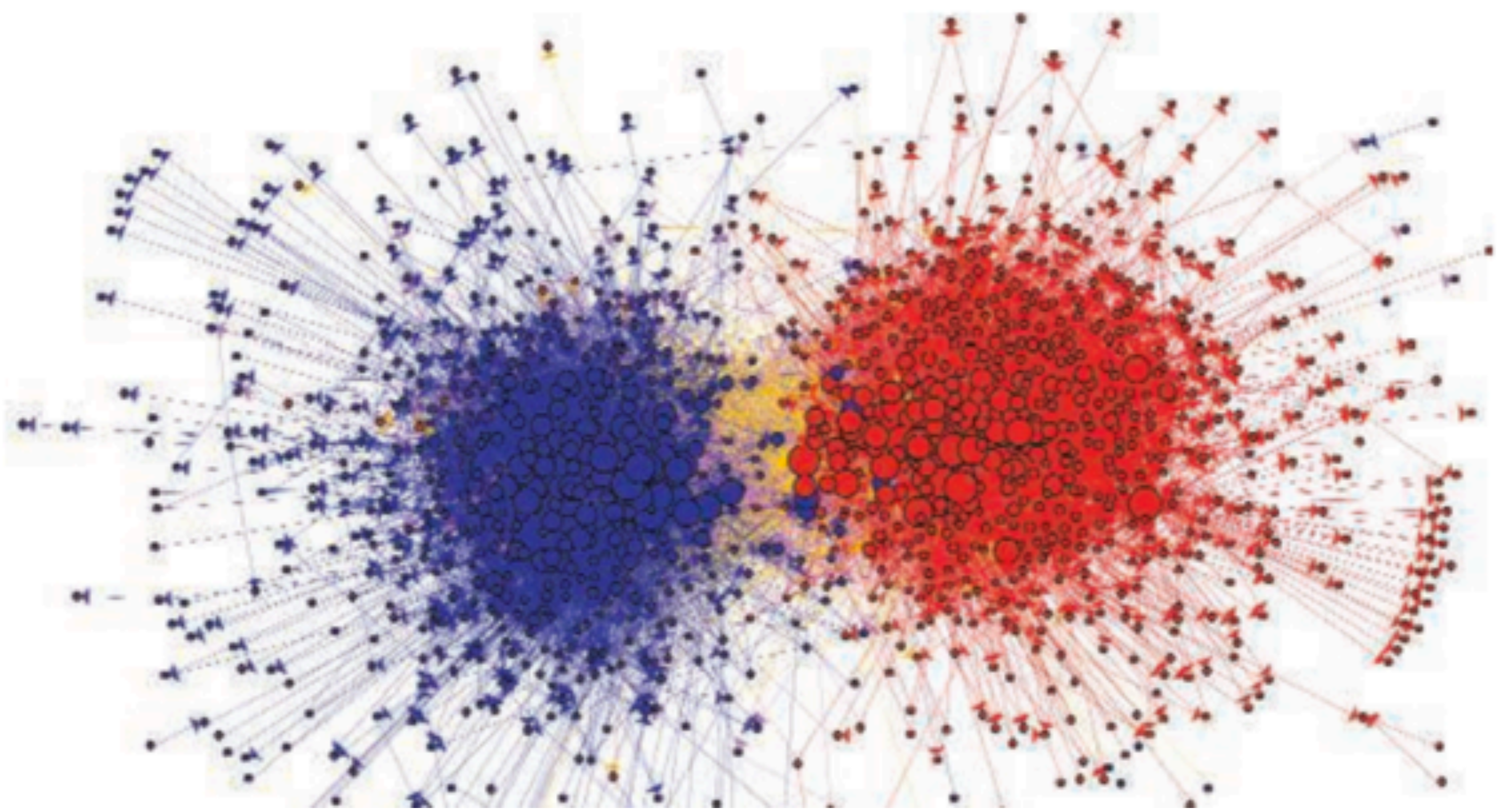
A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5] Devon Brewer,[6] Nicholas Christakis,[1] Noshir Contractor,[7] James Fowler,[8] Myron Gutmann,[3] Tony Jebara,[9] Gary King,[1] Michael Macy,[10] Deb Roy,[2] Marshall Van Alstyne[2,11]

We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven "computational social science" has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational social science is occurring—in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge.

What value might a computational social science—based in an open academic environment—offer society, by enhancing understanding of individuals and collectives? What are the




facebook
December 2010

Computational Social Science

Official social data

Data collection

100 BCE

Data analysis

1829

Data computation

1890

Semi-structured social data

Digitized behavior
Billions of users,
messages/day

Digitized news
Thousands of articles/day

HATHI TRUST Digital Library
Google books

Digitized archives
Millions of books/century

1900

2000

Text as "Data"?

Details Agreed on Nuclear Deal With Iran, Set to Start Jan. 20
PARIS — Iran and six world powers have agreed on how to put in place an accord that would temporarily freeze much of Iran's nuclear program, American and Iranian officials said on Sunday. That accord would go into effect on Jan. 20. International negotiators worked out an agreement in November to constrain much of Iran's program for six months so that diplomats would have time to pursue a more comprehensive follow-up accord. But before the temporary agreement could take effect, negotiators had to work out the technical procedures for carrying it out and resolve some of its ambiguities in concert with the International Atomic Energy Agency.
Antigovernment Protesters Try to Shut Down Bangkok
BANGKOK — Antigovernment protesters seeking to block next month's elections in Thailand took over major roads in Bangkok on Sunday as they began their campaign to shut down the city. In this vast metropolis of well over 10 million people, the protesters were unlikely to paralyze all movement and commerce. But they vowed that by Monday morning they would close busy intersections, make major government offices inaccessible and besiege the homes of top officials in the administration of Prime Minister Yingluck Shinawatra, whose party is most likely to win the general elections that are scheduled for Feb. 2. "We have to shut down Bangkok," said Ratchanee Saengarun, a protester who stood in the middle of an intersection in the city. "This is our last resort." By late Sunday, protesters had blocked several roads using double-decker buses and sandbags, and had diverted traffic.

Text as "Data"?

46 183 3388 43 135 2727 35258 149 14001 69 24 225
37 57124 7 9641 176 252 15 2086 183 3388 218 14001 161 10830 97 2128 33
5268 1459 28 5 449 14210 6966 43 45564 360 9641 3 363 3734 3388 39465
5268 33 1459 165 570 90 3388 24 7097 261 11 48 611 2128 197 10830 42
14001 2 449 14210 16347 398 5338 176 442 499 5268 5 1459 2086 480
14001 26 12709 1251 23 1 27181 2248 338 30775 28 197 739 248 38678 11
1139 14001 257 611 30775 37 24 5338 20 3837 611 9641 17 1073 14210
2341 2 10830 3 2727 30775 261 1 85 88741
17877 10 70 14001 11 438 2
2 65417 59555 10 87 14001 40 427 43199 31 10830 3 152 560 367 7 10830 2
3388 19 2857 1639 129 1159 73 14001 11 438 30775 47956 10830 1529 15
75989 14210 260 560 327 2692 51472 30775 10 1177 23 14001 90351 717 30
9641 24040 2248 1639 9 5268 2811 135 39 1639 1459 199 20 13554 406 367
552 51 1 9641 35951 30775 37 14210 121 363 10830 30775 165 14210 57 59
90525 87723 108 78 4750 597 179 14001 60 30775 257 31 5268 2563 68
5338 14 15012 2679 2086 14001 11 438 14456 3734 16286 44733 12709 1
1031 14 10830 30775 25 14210 2128 49392 10830 30775 20260 738 4750
250 797 32407 2811 195 90338 10 1139 4 244 7 111 3 7 9641 75964 9641
1139 5 95973

TextGenerator(SocialAttributes) → *Text*

Data generation process

Society *(SocialAttributes)*

Writing *(TextGenerator)*

Text Data *(Text)*

Language generation as social process
P(TextGen | Text, SocAttr)

Language for social measurement
P(SocAttr | Text, TextGen)

TextGenerator(SocialAttributes) → *Text*

Society *(SocialAttributes)*

Writing *(TextGenerator)*

Text Data *(Text)*

Data generation process

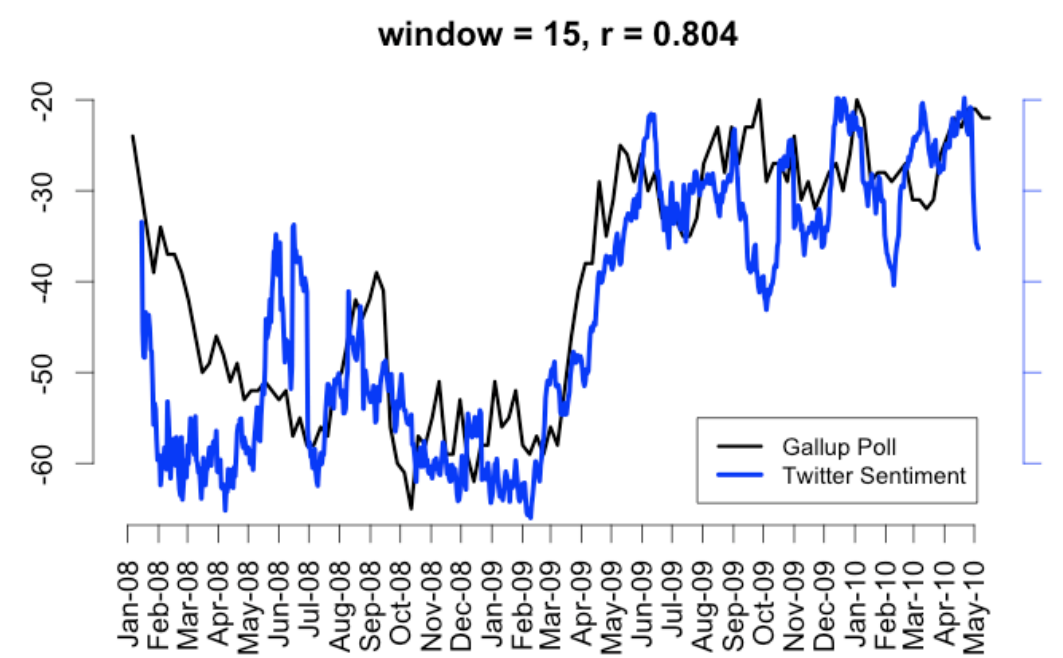
Public opinion

Social media usage

Model assumptions

Language for social measurement

P(SocAttr | Text, TextGen)



[O'Connor et al., 2010]

TextGenerator(SocialAttributes) → *Text*

Data generation process

Society *(SocialAttributes)*

Writing *(TextGenerator)*

Text Data *(Text)*

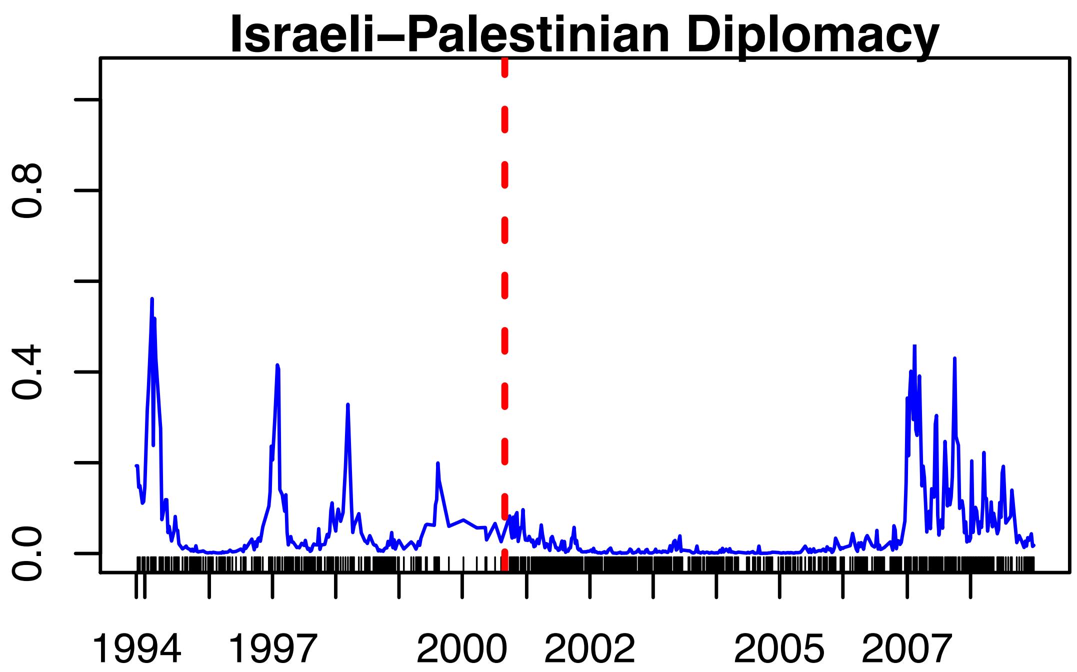
Real-world political events

News media process

Model assumptions

Language for social measurement

P(SocAttr | Text, TextGen)



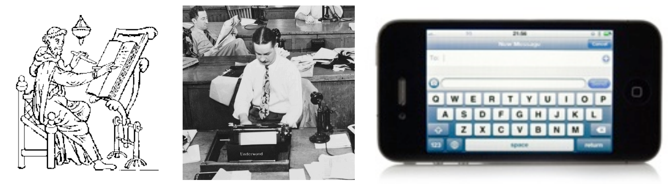
[O'Connor, Stewart, Smith 2013]

TextGenerator(SocialAttributes) → *Text*

Data generation process

Society *(SocialAttributes)*

Writing *(TextGenerator)*



Text Data *(Text)*

HATHI TRUST Digital Library

Real-world political events

News media process

Language generation as social process

P(TextGen | Text, SocAttr)

TextGenerator(SocialAttributes) → *Text*

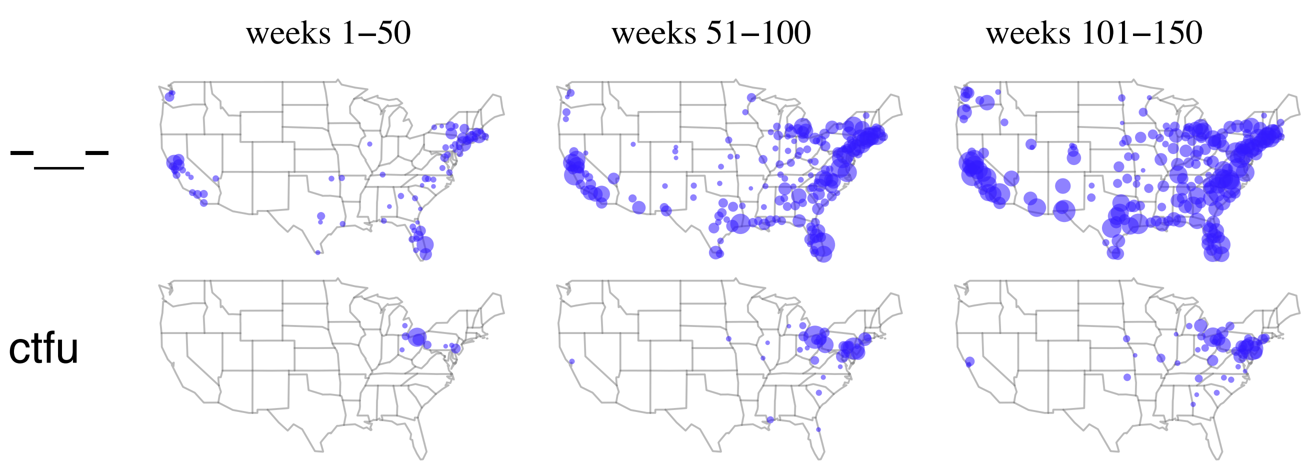
Society *(SocialAttributes)* | Writing *(TextGenerator)* | Text Data *(Text)*

Data generation process

→

→

Geography of authors → Social media usage



Language generation as social process

P(TextGen | Text, SocAttr)

[Eisenstein et al. 2010, O'Connor et al. 2010, Eisenstein et al. 2012]

TextGenerator(SocialAttributes) $\rightarrow$ *Text*

Data generation process

Society *(SocialAttributes)*

→

Writing *(TextGenerator)*

→

Text Data *(Text)*

HATHI TRUST Digital Library

- Social media and polls
- Geography and language
- Social determinants of lexical diffusion
- Events in international relations
- Text exploration on document covariates

TextGenerator(SocialAttributes) → Text

Society *(SocialAttributes)*

Data generation process

Writing *(TextGenerator)*

Text Data *(Text)*

HATHI TRUST Digital Library

Descriptive analysis:
Statistical associations

- Text exploration on document covariates

[O'Connor, 2014]

MiTextExplorer:

a Mutual Information

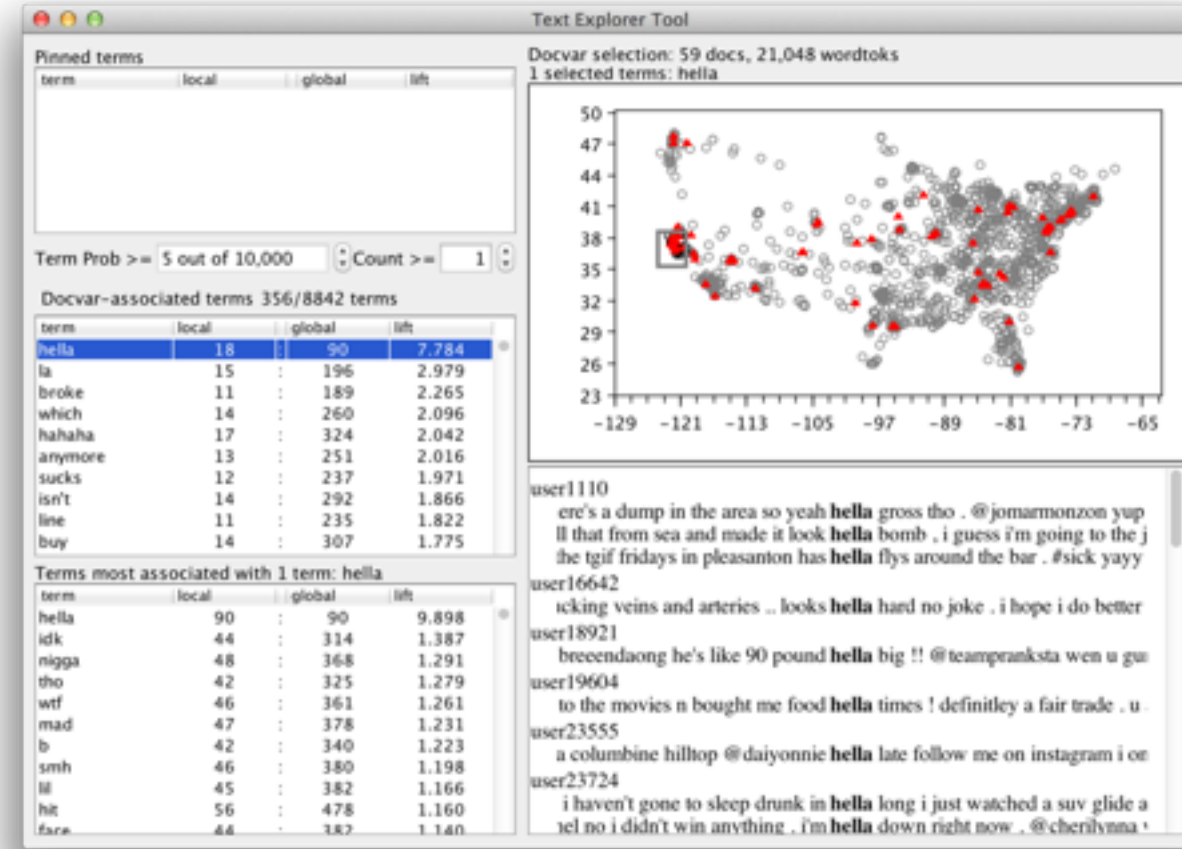
Text Explorer

using

Linked Brushing

with

Document Covariates



http://brenocon.com/mte

[O'Connor, 2014]

How are *X* and *Y* related? (Anscombe 1973)

x	y
10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68

x	y
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.10
6	6.13
4	3.10
12	9.13
7	7.26
5	4.74

x	y
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

x	y
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.25
19	12.50
8	5.56
8	7.91
8	6.89

How are X and Y related? (Anscombe 1973)

x	y
10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68

$r = 0.82$

x	y
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.10
6	6.13
4	3.10
12	9.13
7	7.26
5	4.74

$r = 0.82$

x	y
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

$r = 0.82$

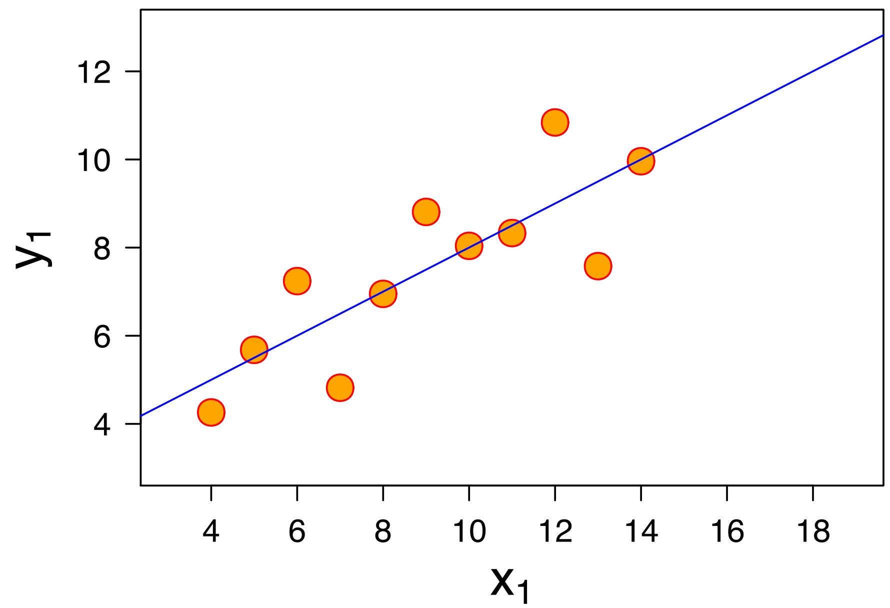
x	y
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.25
19	12.50
8	5.56
8	7.91
8	6.89

$r = 0.82$

How are X and Y related? (Anscombe 1973)

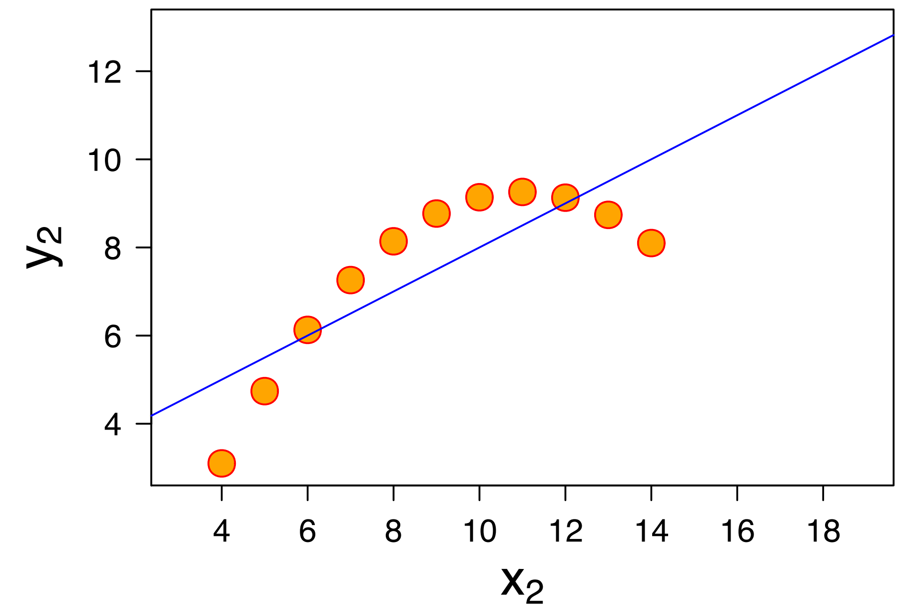
x	y
10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68

$r = 0.82$



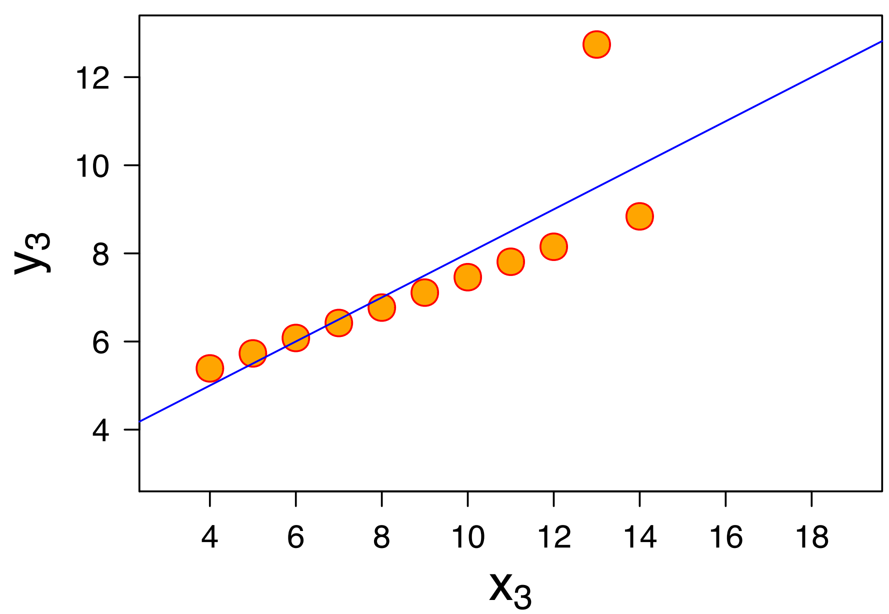
x	y
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.10
6	6.13
4	3.10
12	9.13
7	7.26
5	4.74

$r = 0.82$



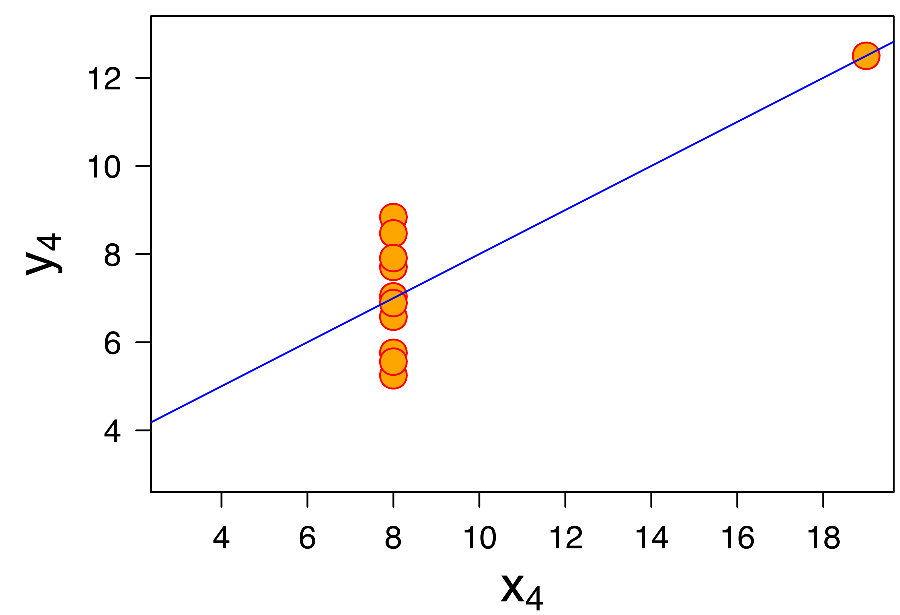
x	y
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

$r = 0.82$



x	y
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.25
19	12.50
8	5.56
8	7.91
8	6.89

$r = 0.82$



How are X and Y related? (Anscombe 1973)

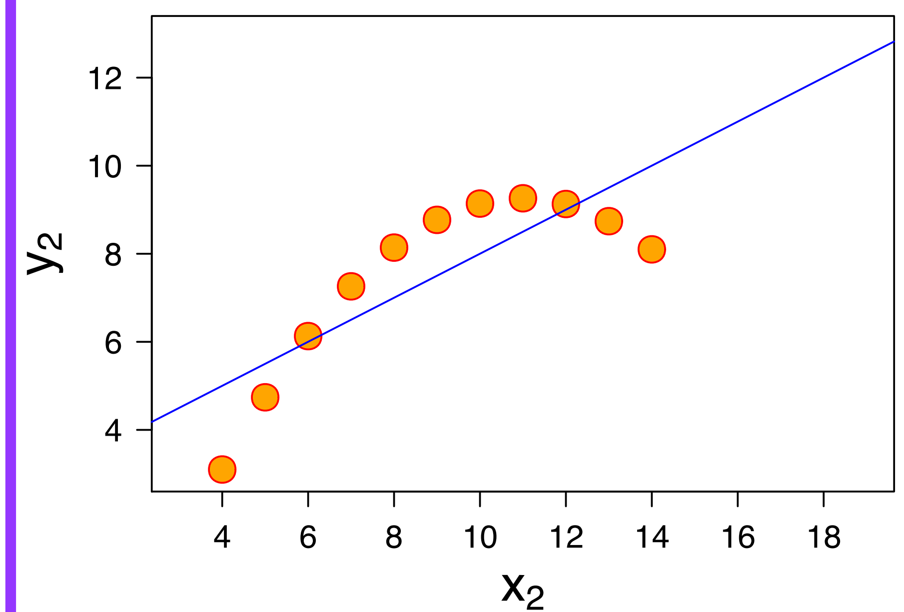
Pearson correlation

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

assumes $(x, y) \sim N(\mu, \Sigma)$

x	y
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.10
6	6.13
4	3.10
12	9.13
7	7.26
5	4.74

$r = 0.82$



Scatterplot:

x = horizontal position

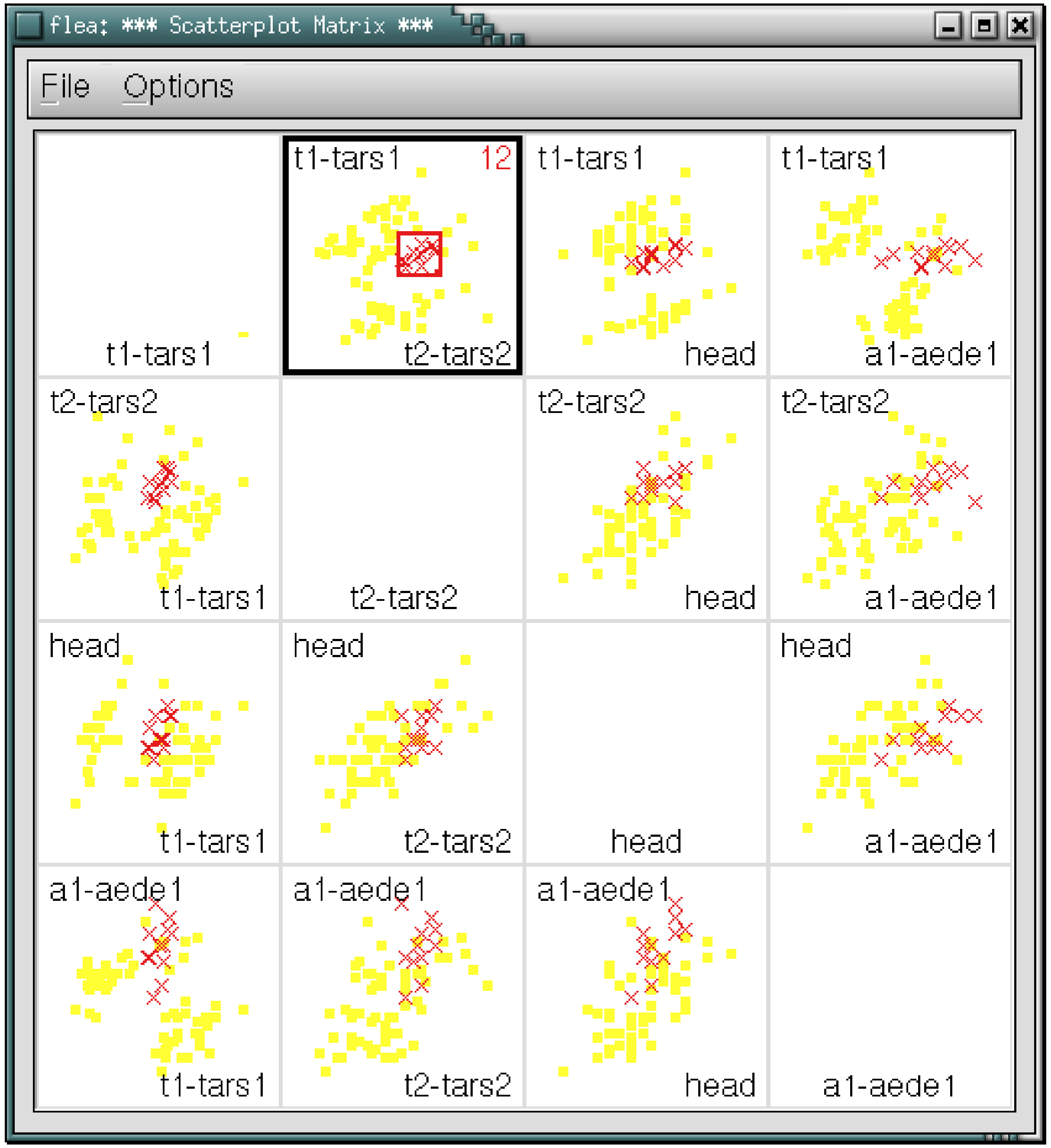
y = vertical position

Simple

Non-parametric(?)

Is there an analogue to the scatterplot, when text is a variable?

Linking and brushing



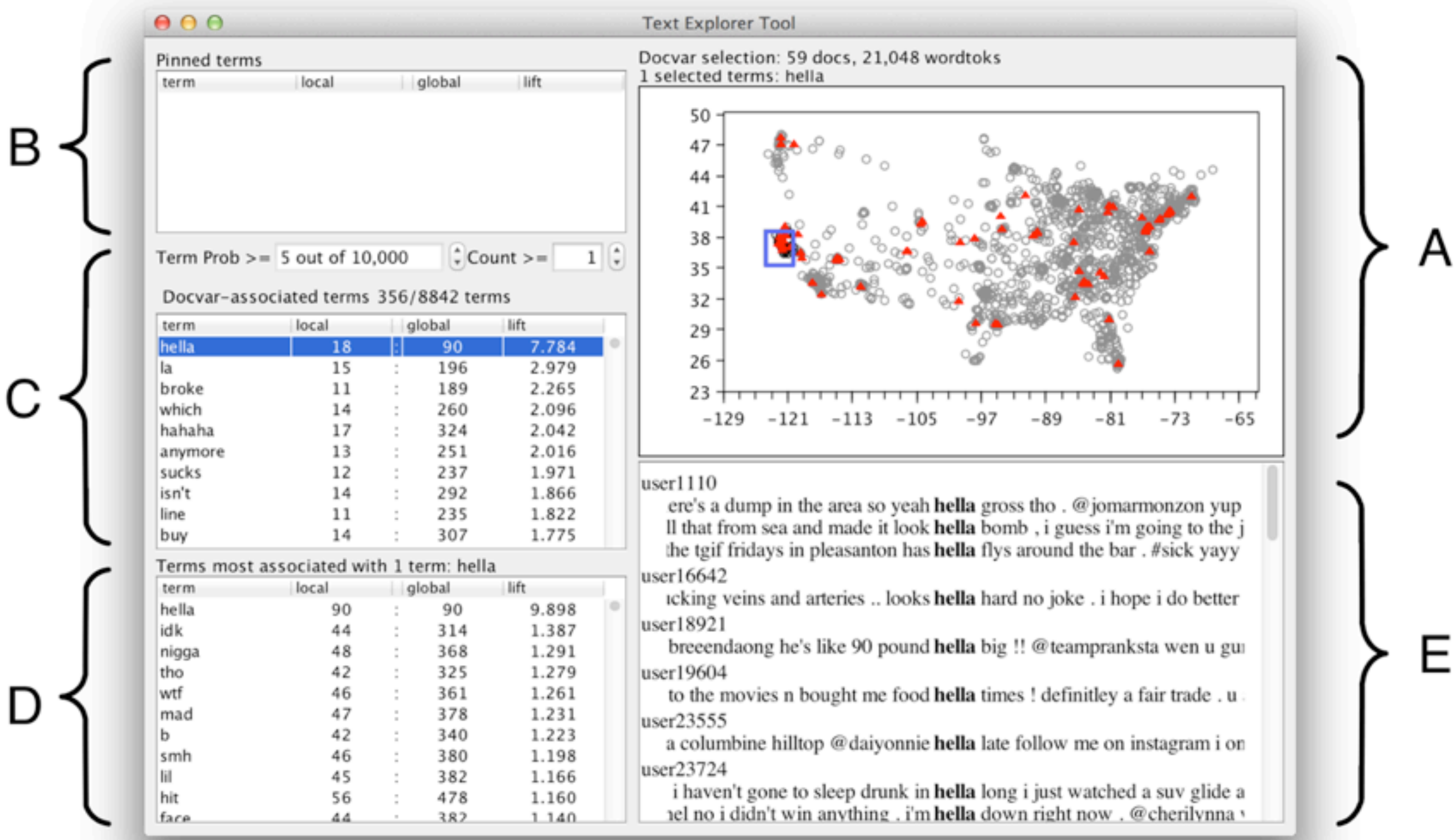
GGobi software
(Cook and Swayne 2007, Buja et. al 1996, etc.)

Is there an analogue to linking/brushing, when text is a variable?

Text and document covariates

- ***X***: Text
  - Discrete, high-dimensional (e.g. bag of words)
- ***Y***: Document covariates (metadata)
  - Time, author attributes, social context, geography, community membership...
  - Discrete or continuous
  - Lower dimensional
- Goal is <u>exploratory data analysis</u>: first-cut insight into *relationship(X,Y)*
- Requirement: speed for interactivity

Demo

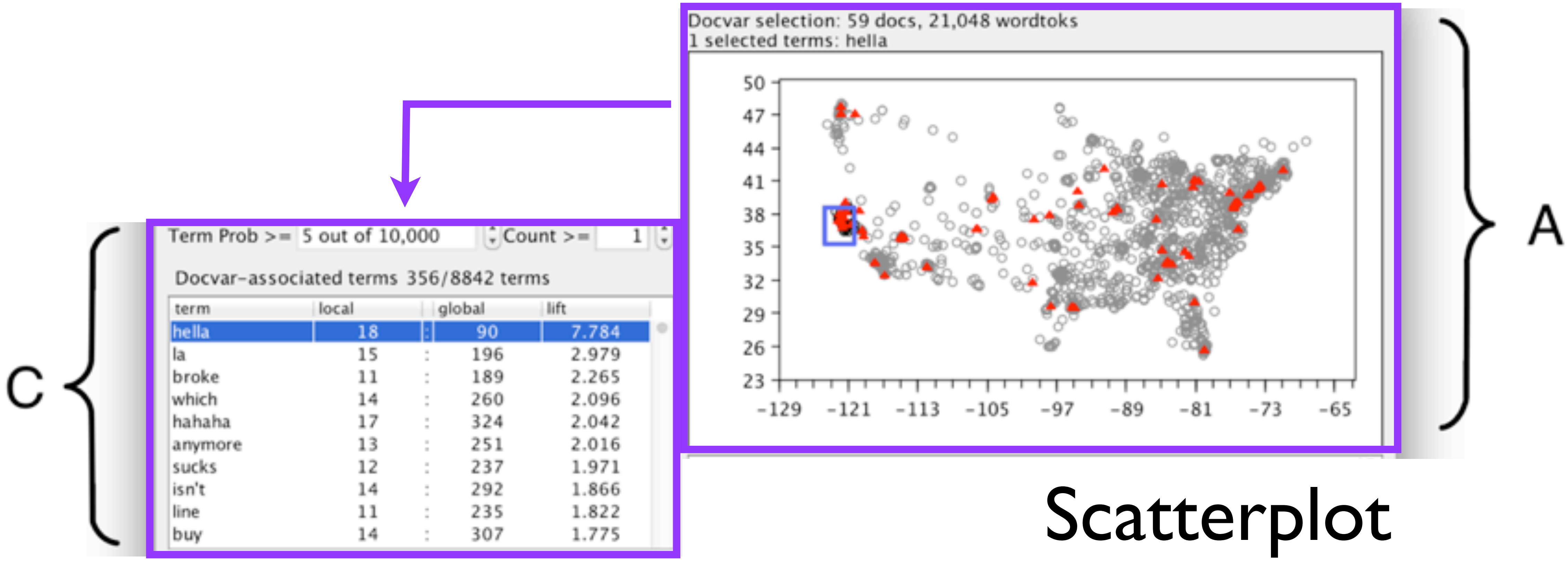


Linked views of the data

(A) Covariate display
(C) Covariate-word associations
(E) Keyword-in-context text display

[A] → *[C]*: words related to covariate query *Q*

Q selection: "brushing"



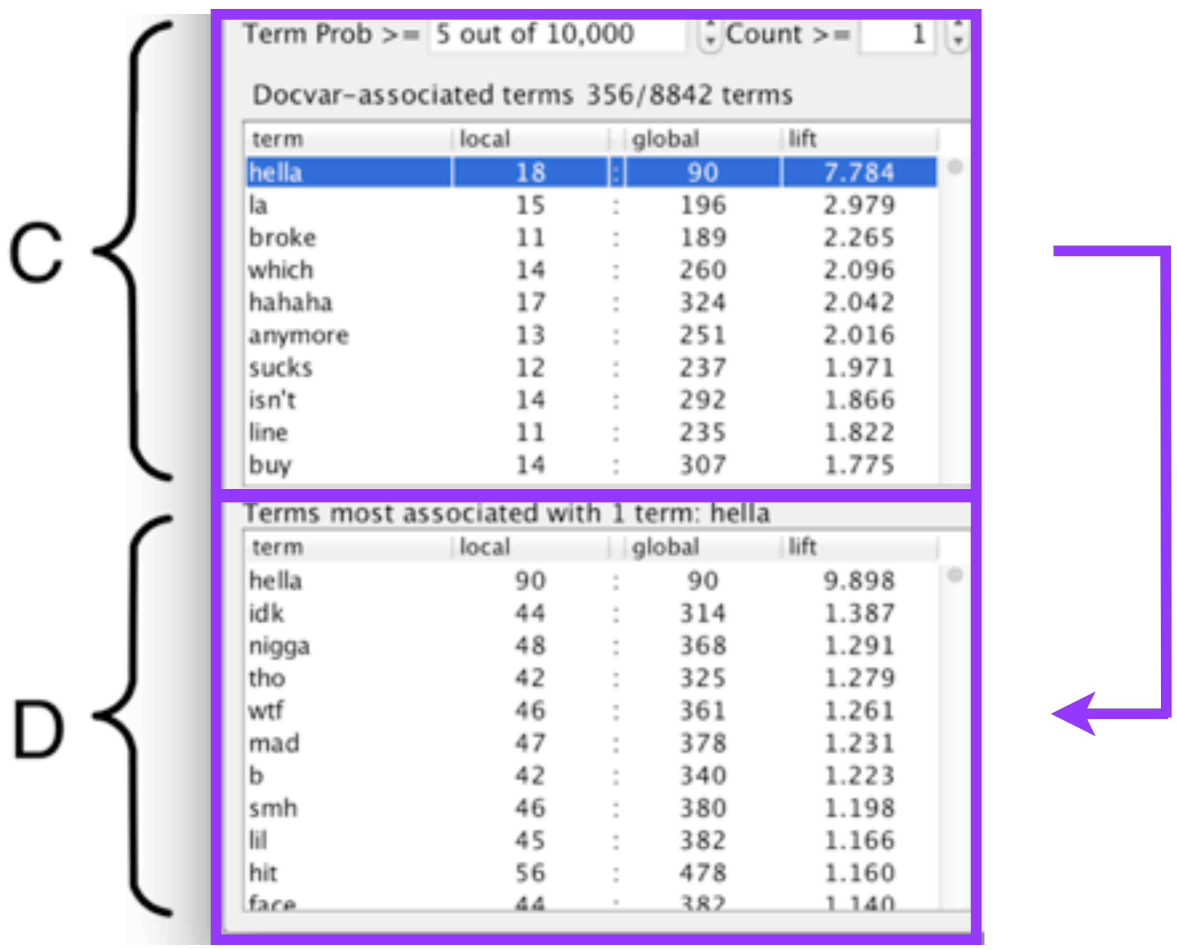
Scatterplot

Ranked list

$$\mathrm{rank}_w \frac{p(w|Q)}{p(w)} \quad \text{where} \quad \begin{aligned} p(w|Q) &\geq \mathrm{TermProbThresh} \\ \mathrm{count}_Q(w) &\geq \mathrm{TermCountThresh} \end{aligned}$$

(Exponentiated) Pointwise Mutual Information (a.k.a. *lift*)

[C] → *[D]*: word-word associations



Term Prob >= 5 out of 10,000 Count >= 1

Docvar-associated terms 356/8842 terms

term	local	global	lift
hella	18	90	7.784
la	15	196	2.979
broke	11	189	2.265
which	14	260	2.096
hahaha	17	324	2.042
anymore	13	251	2.016
sucks	12	237	1.971
isn't	14	292	1.866
line	11	235	1.822
buy	14	307	1.775

Terms most associated with 1 term: hella

term	local	global	lift
hella	90	90	9.898
idk	44	314	1.387
nigga	48	368	1.291
tho	42	325	1.279
wtf	46	361	1.261
mad	47	378	1.231
b	42	340	1.223
smh	46	380	1.198
lil	45	382	1.166
hit	56	478	1.160
face	44	382	1.140

$$\text{rank}_v \frac{p(v|w \in \text{doc})}{p(v)}$$

(Exponentiated) Pointwise Mutual Information (a.k.a. *lift*)

Term Prob >= 5 out of 10,000 Cou

Docvar selection: 39 docs, 34,002 wordtoks

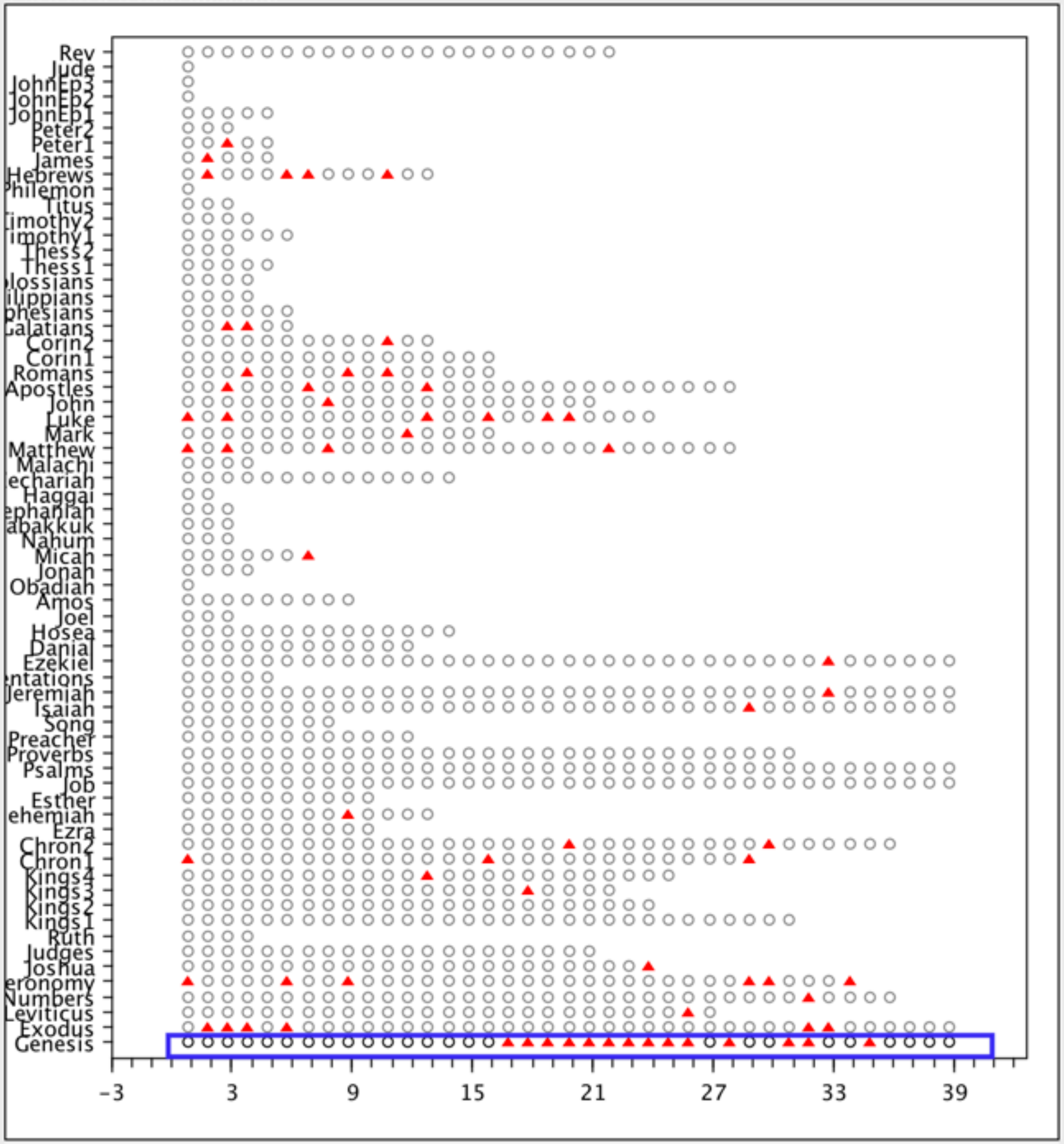
1 selected terms: abraham

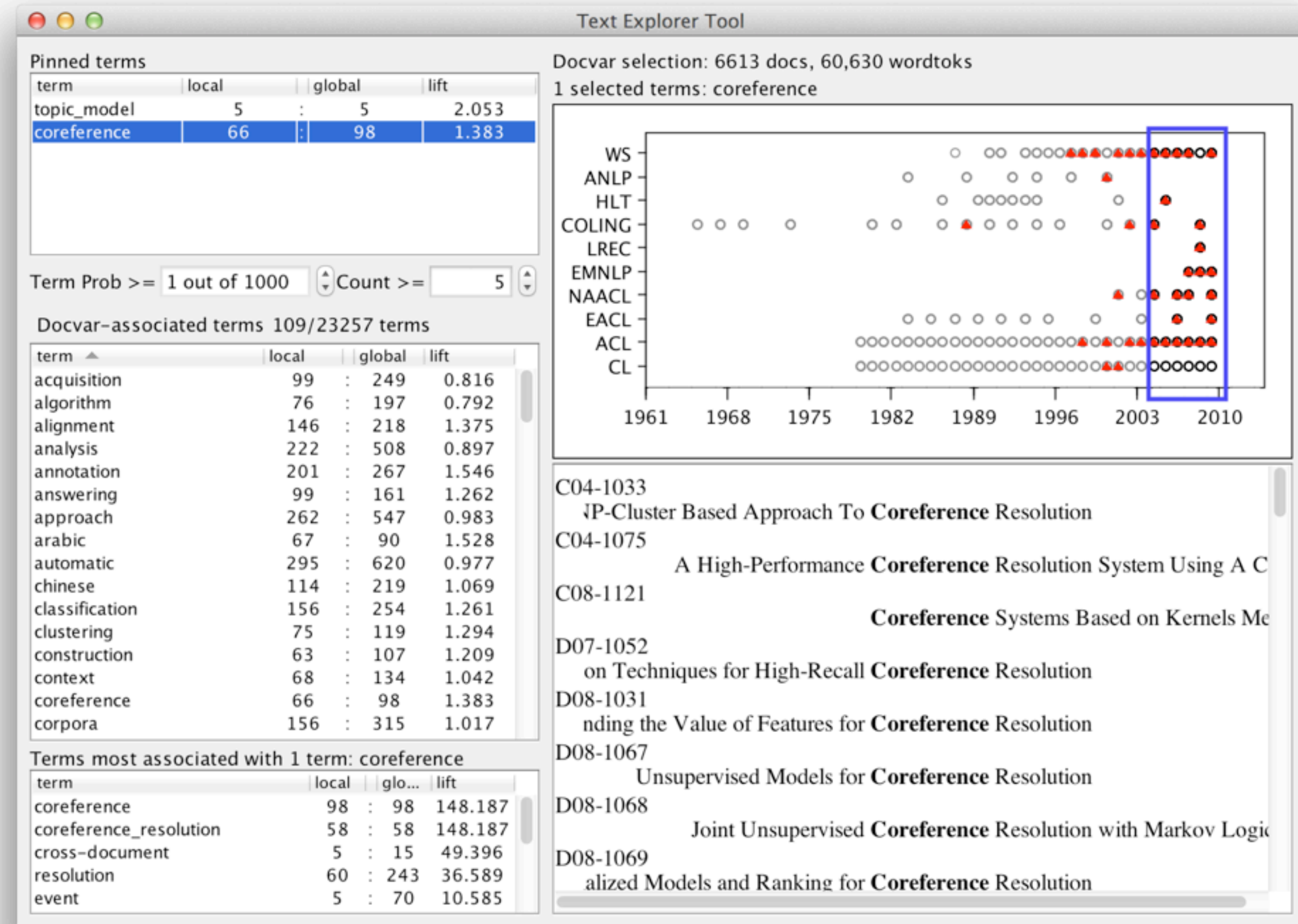
Docvar-associated terms

term	local	gl...	lift
rebekah	29	: 29	24.228
laban	52	: 53	23.771
leah	30	: 31	23.446
abram	59	: 61	23.433
rachel	40	: 43	22.537
sarah	36	: 39	22.364
noah	41	: 51	19.477
esau	76	: 98	18.789
duke	32	: 43	18.030
lived	36	: 54	16.152
isaac	75	: 126	14.421
adam	18	: 31	14.068
abraham	128	: 237	13.085
jacob	143	: 278	12.462
camels	23	: 46	12.114
conceived	22	: 45	11.845
sodom	21	: 46	11.060
canaan	25	: 67	9.040
abimelech	24	: 66	8.810
cattle	43	: 135	7.717
begat	67	: 225	7.214
bare	50	: 172	7.043
lot	30	: 106	6.857
flocks	18	: 67	6.509
tent	22	: 82	6.500
joseph	31	: 117	6.419
daughters	58	: 232	6.057
wife	99	: 398	6.026
bless	18	: 84	5.192
seed	50	: 239	5.069

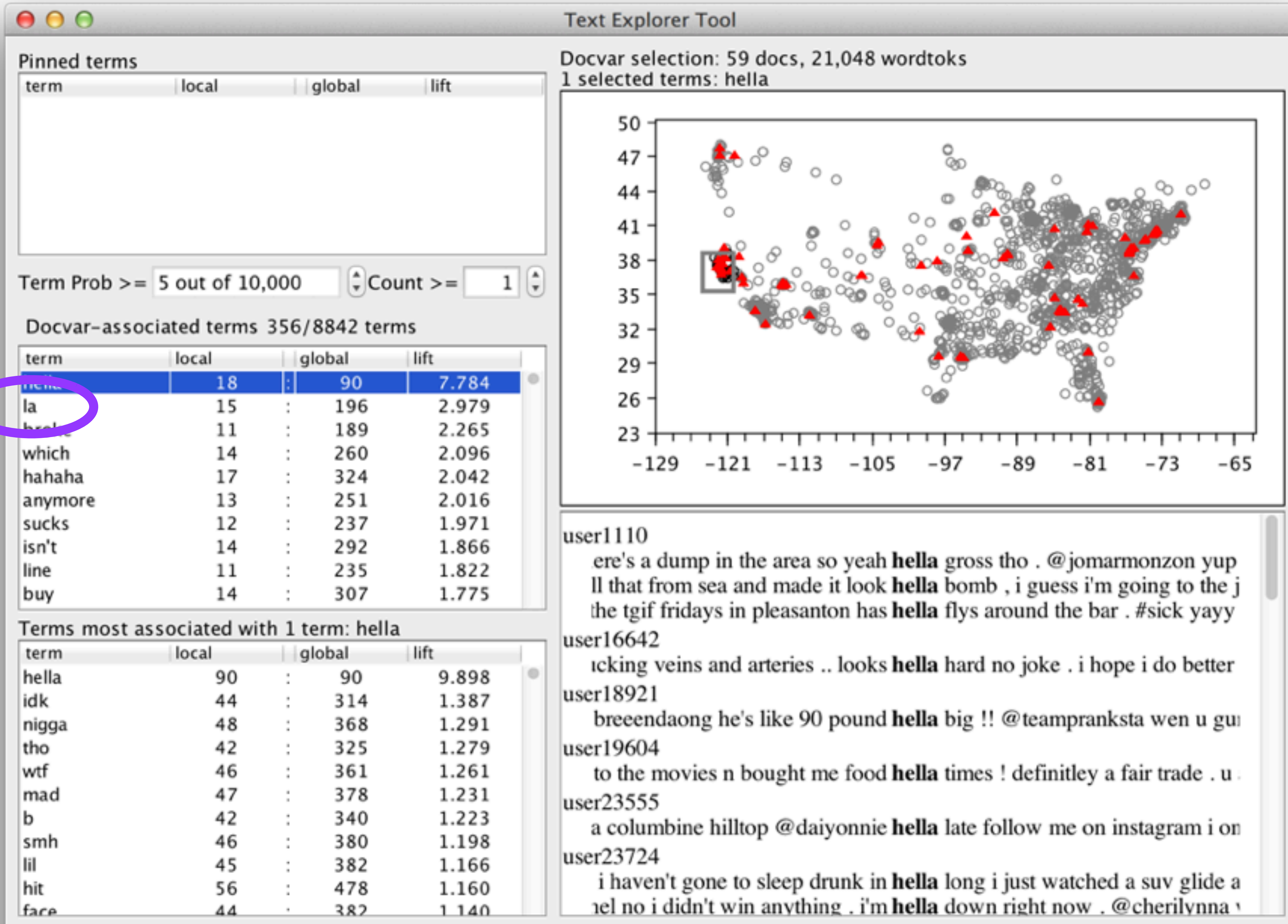
Terms most associated with 1 term:...

term	lo...	gl...	lift
abraham	237	: 237	12.148
sarah	38	: 39	11.836
isaac	111	: 126	10.702
jacob	130	: 278	5.681
seed	75	: 239	3.812
faith	71	: 247	3.492
begat	53	: 225	2.862
covenant	58	: 261	2.700




Text Explorer Tool
Pinned terms
term local global lift
topic_model 5 : 5 2.053
coreference 66 : 98 1.383
Term Prob >= 1 out of 1000
Count >= 5
Docvar-associated terms 109/23257 terms
term local global lift
acquisition 99 : 249 0.816
algorithm 76 : 197 0.792
alignment 146 : 218 1.375
analysis 222 : 508 0.897
annotation 201 : 267 1.546
answering 99 : 161 1.262
approach 262 : 547 0.983
arabic 67 : 90 1.528
automatic 295 : 620 0.977
chinese 114 : 219 1.069
classification 156 : 254 1.261
clustering 75 : 119 1.294
construction 63 : 107 1.209
context 68 : 134 1.042
coreference 66 : 98 1.383
corpora 156 : 315 1.017
Terms most associated with 1 term: coreference
term local glo... lift
coreference 98 : 98 148.187
coreference_resolution 58 : 58 148.187
cross-document 5 : 15 49.396
resolution 60 : 243 36.589
event 5 : 70 10.585
Docvar selection: 6613 docs, 60,630 wordtoks
1 selected terms: coreference
WS
ANLP
HLT
COLING
LREC
EMNLP
NAACL
EACL
ACL
CL
1961
1968
1975
1982
1989
1996
2003
2010
C04-1033
IP-Cluster Based Approach To Coreference Resolution
C04-1075
A High-Performance Coreference Resolution System Using A C
C08-1121
Coreference Systems Based on Kernels Me
D07-1052
on Techniques for High-Recall Coreference Resolution
D08-1031
nding the Value of Features for Coreference Resolution
D08-1067
Unsupervised Models for Coreference Resolution
D08-1068
Joint Unsupervised Coreference Resolution with Markov Logic
D08-1069
alized Models and Ranking for Coreference Resolution

KWIC (keyword-in-context)



KWIC reveals word senses

user1110

guess i'm going to the jungle (**la**) @killa_kimbo its totally tru

ch " (@seanygrey i will be in **la** by morning :) that's a fuckin

user29006

per . @gastelo12 did u bust ? **la** ! la la laa laa la la la laa . goc

. @gastelo12 did u bust ? la ! **la** la laa laa la la la laa . goodm

@gastelo12 did u bust ? la ! la **la** laa laa la la la laa . goodmorr

2 did u bust ? la ! la la laa laa **la** la la laa . goodmorning my li

did u bust ? la ! la la laa laa la **la** la laa . goodmorning my littl

d u bust ? la ! la la laa laa la la **la** laa . goodmorning my little r

user31473

me @cherylsatjipto ;) balik dr **la** kpn ? bb is a distraction , it k

user34771

y twiin sister is going to be in **la** for my bros middleschool gr

user47627

san fracisco is way better that **la** trust me . :) @teammahone y

user5149

king you a nuisance . i'll be in **la** this weekend hobnobbing w

yself right now . just drove to **la** from sf and back alone for th

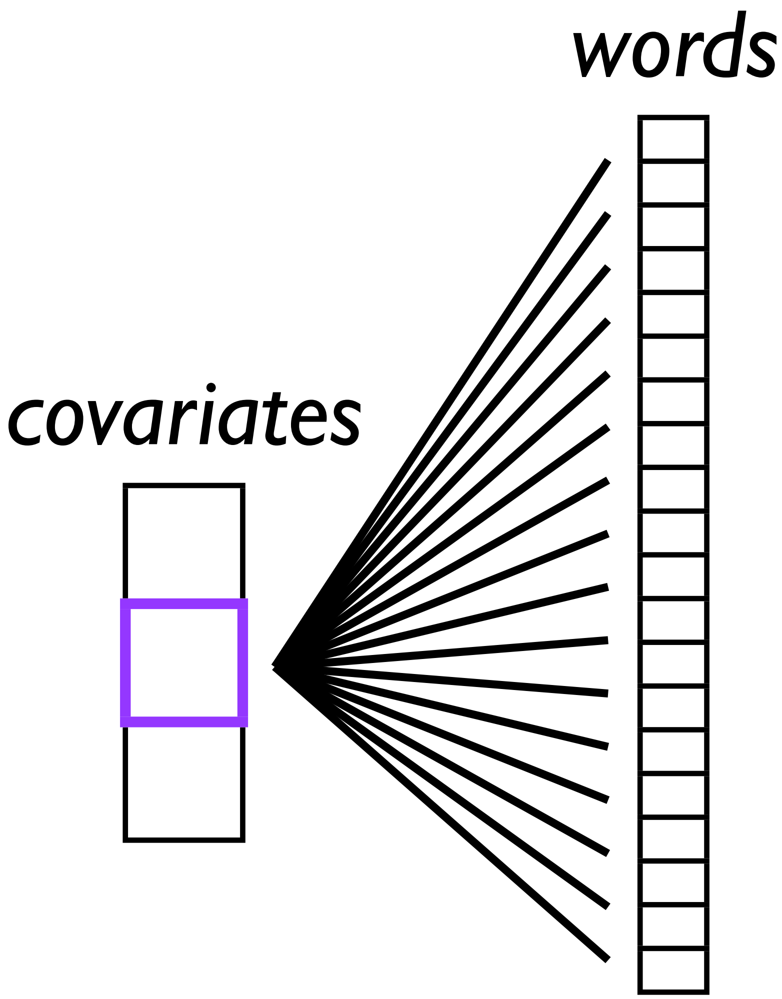
user5239

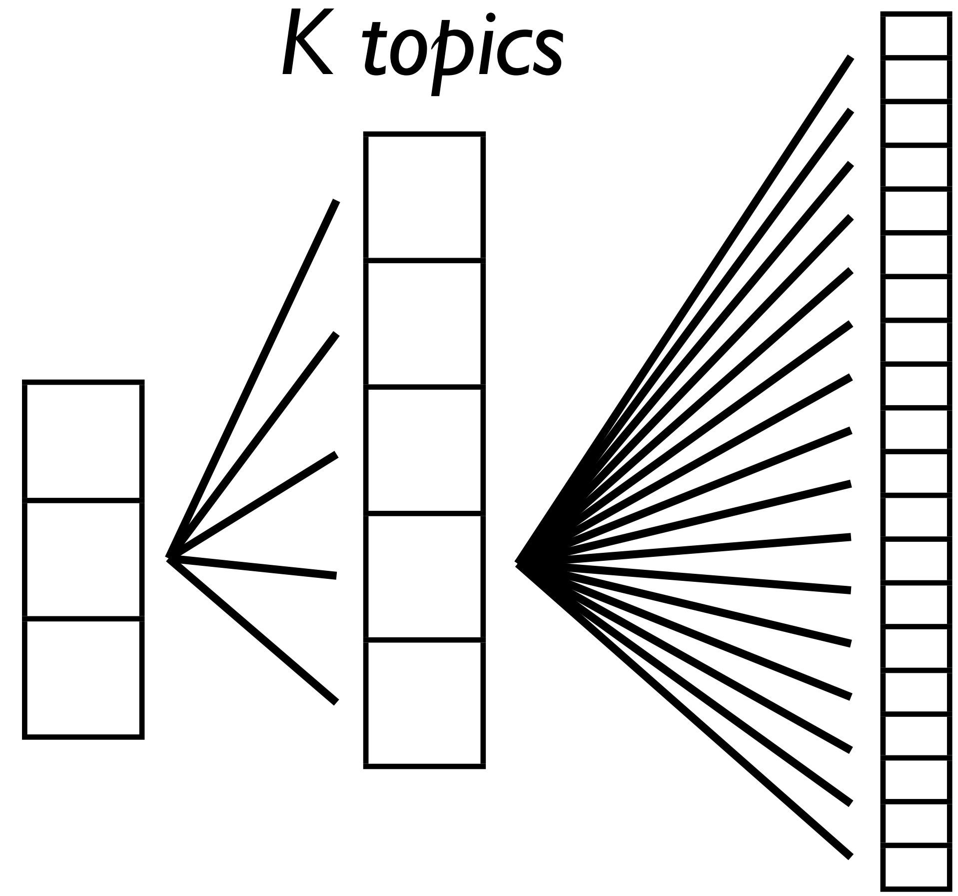
co @jorge_cortesc en pipolos **la** comida esta super grasosa #t

@s me voy a dormir aca ya es **la** 1am supongo q alla las 3am

Covariate -- word analysis

direct PMI *-vs-* topic model bottleneck





- Feature selection
- Monroe et al. (2008)

- p(text | covariates): Dirichlet-Multinomial Regression, Author-Topic Model, Labeled LDA, Structural Topic Model ...
- p(text, covariates): Supervised LDA, MedLDA, GeoTM ...

Related work: Text Exploration

- Voyant/Voyeur (Rockwell et al. 2010)
- WordSeer (Shrikumar 2013)
- Jigsaw (Görg et al. 2013)
- Topical Guide (Gardner et al. 2010)
- etc...

- Other uses
  - Figure out NLP models and parameters (what should be a stopword?)
  - Select documents to read in an intelligent way (by covariates)
  - What variables to use in a model?
  - Identify coding errors in the data
- Extensions
  - Structure from NLP tools
  - Interactive labeling and keyword query building *[King et al 2014]*

Prototype available: **http://brenocon.com/mte**

Text as "data"?

Details Agreed on Nuclear Deal With Iran, Set to Start Jan. 20
PARIS — Iran and six world powers have agreed on how to put in place an accord that would temporarily freeze much of Iran's nuclear program, American and Iranian officials said on Sunday. That accord would go into effect on Jan. 20. International negotiators worked out an agreement in November to constrain much of Iran's program for six months so that diplomats would have time to pursue a more comprehensive follow-up accord. But before the temporary agreement could take effect, negotiators had to work out the technical procedures for carrying it out and resolve some of its ambiguities in concert with the International Atomic Energy Agency.
Antigovernment Protesters Try to Shut Down Bangkok
BANGKOK — Antigovernment protesters seeking to block next month's elections in Thailand took over major roads in Bangkok on Sunday as they began their campaign to shut down the city. In this vast metropolis of well over 10 million people, the protesters were unlikely to paralyze all movement and commerce. But they vowed that by Monday morning they would close busy intersections, make major government offices inaccessible and besiege the homes of top officials in the administration of Prime Minister Yingluck Shinawatra, whose party is most likely to win the general elections that are scheduled for Feb. 2. "We have to shut down Bangkok," said Ratchanee Saengarun, a protester who stood in the middle of an intersection in the city. "This is our last resort." By late Sunday, protesters had blocked several roads using double-decker buses and sandbags, and had diverted traffic.

Text as "data"?

Details Agreed on Nuclear Deal With Iran, Set to Start Jan. 20

PARIS — Iran and six world powers have **agreed** on how to **put** in place an accord that would temporarily **freeze** much of Iran's nuclear program, American and Iranian officials said on Sunday. That accord would **go** into effect on Jan. 20. International negotiators worked out an agreement in November to **constrain** much of Iran's program for six months so that diplomats would **have** time to pursue a more comprehensive follow-up accord. But before the temporary agreement could **take effect**, negotiators had to **work out** the technical procedures for carrying it out and **resolve** some of its ambiguities in concert with the International Atomic Energy Agency.

Antigove

BANGKOK … s in

Thailand … ign to

shut dow … otesters

were unli … Monday

morning … accessible

and besie … gluck

Shinawat … duled for

Feb. 2. " … r who

stood in … e Sunday,

protester … nd had

diverted

Structure from text:

Semantic parsing

Information extraction

[e.g. MUC-3: *Lehnert, Williams, Cardie, Riloff, Fisher 1991*]

Event data through knowledge engineering

[Schrodt 1994, Leetaru and Schrodt 2013]

Event classes
(~200)

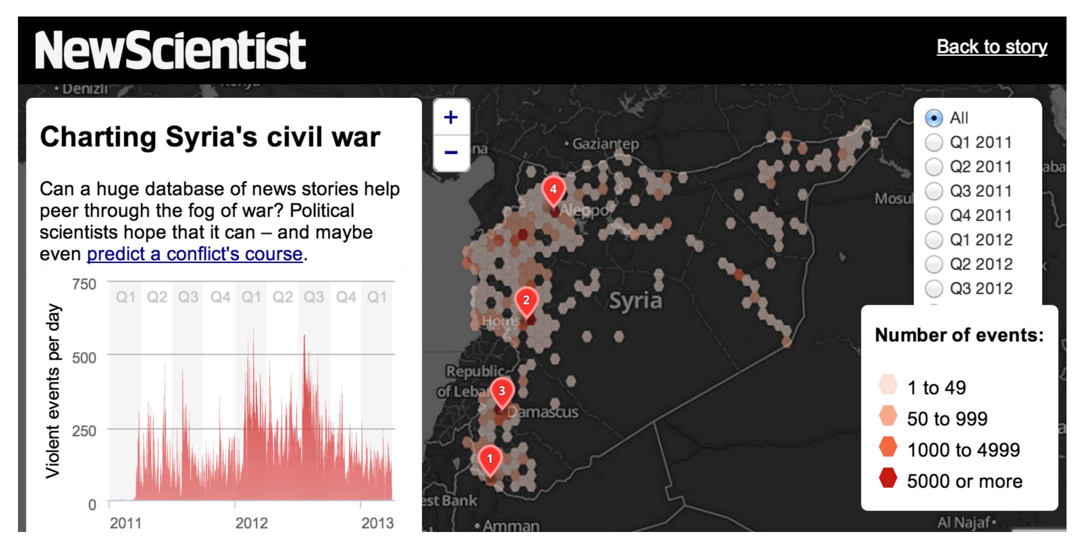
03 - EXPRESS INTENT TO COOPERATE
07 - PROVIDE AID
15 - EXHIBIT MILITARY POSTURE

Dictionary:
Verb patterns per event class
(~15000)

191 - Impose blockade, restrict movement

not_ allow to_ enter ;mj 02 aug 2006
barred travel
block traffic from ;ab 17 nov 2005
block road ;hux 1/7/98

Extract events from news text



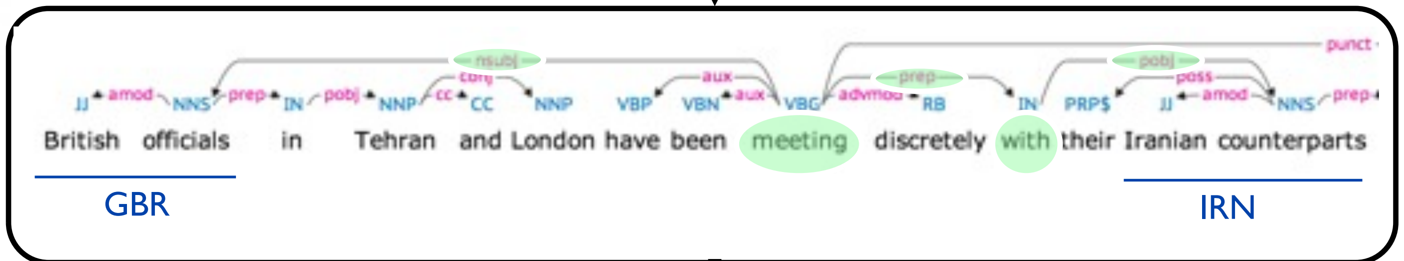
Issue: Hard to maintain and adapt to new domains

Our inference process

[O'Connor, Stewart, and Smith, 2013]

Data: twenty years of news articles

Natural Language Processing



Event phrases of actor interactions

Probabilistic Graphical Model

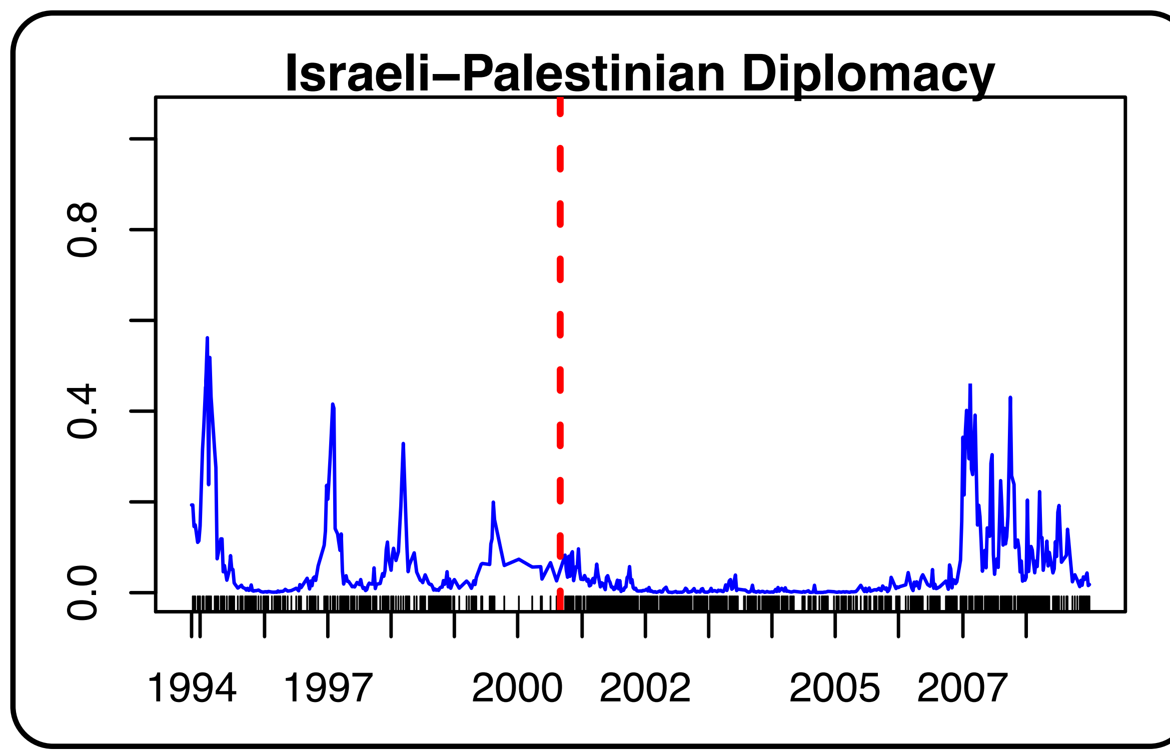
Purely from textual data, jointly learns both

(1) **Event class dictionaries**

(2) **Political dynamics**

"diplomacy"	arrive in, visit, meet with, travel to, leave, hold with, meet, meet in, fly to, be in, arrive for talk with, say in, arrive with, head to, hold in, due in, leave for, make to, arrive to,
"verbal conflict"	accuse, blame, say, break with, sever with, blame on, warn, call, attack, rule with, charge, say←ccomp come from, say ←ccomp, suspect, slam, accuse government ←poss,
"material conflict"	kill in, have troops in, die in, be in, wound in, have soldier in, hold in, kill in attack in, remain in, detain in, have in, capture in, stay in, about ←pobj troops in, kill, have troops

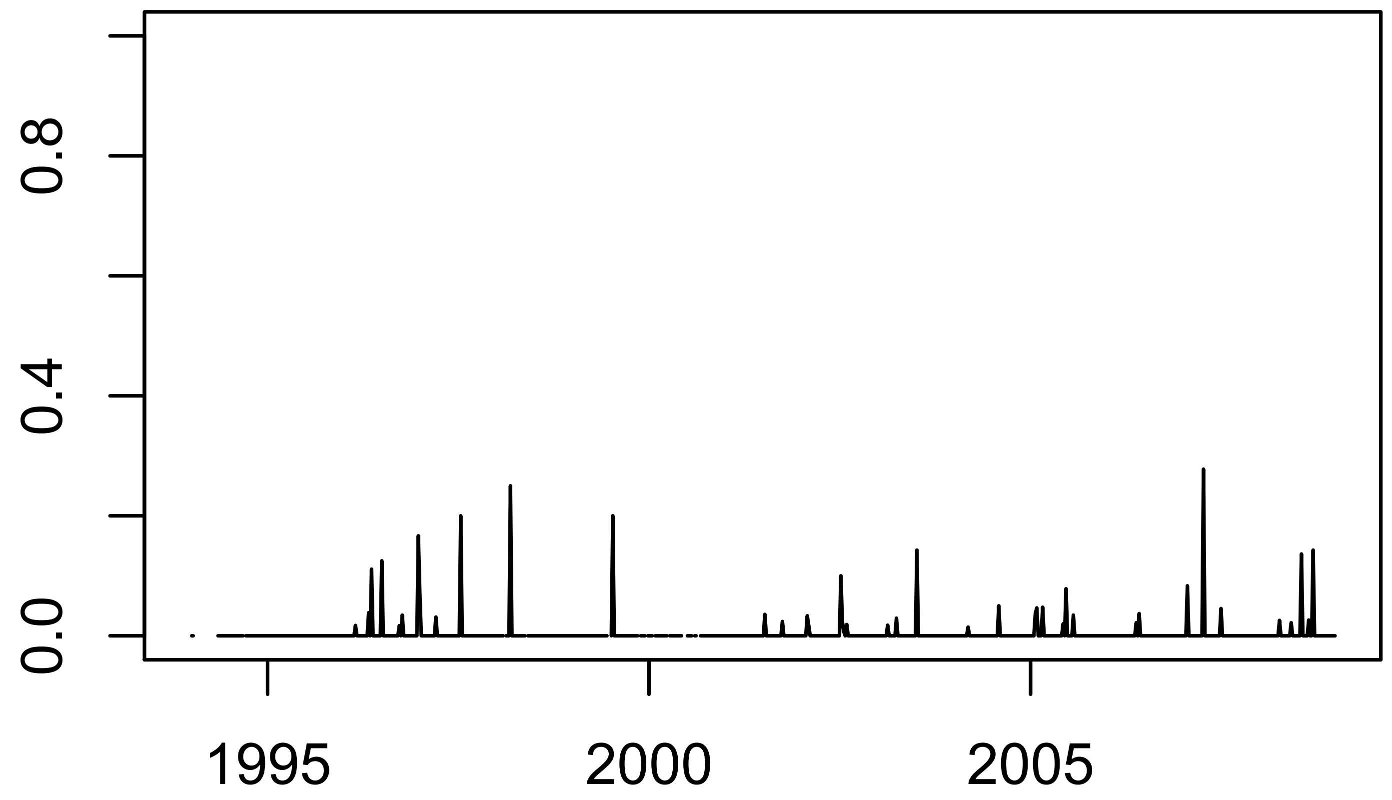
↔



Event phrases

"ISR meet with *PSE"*

P(w = "meet with" | t, s=*ISR*, r=*PSE*)



Too sparse for human interpretability

Do word semantics cluster on social context?

s=ISR, *r*=PSE

t= Jul 15-21, 2002

say <-ccomp be to
release to
take control of
occupy
wound in
scuffle with
be <-xcomp meet
meet with
meet with
arrest

t= Jul 3-9, 2006

commit to
strike
carry in
continue in
reject
fire at target in
start around
ratchet pressure on
shell
hit

s=USA, *r*=FRA

t= Feb 2-8, 1998

travel <-xcomp meet with
consider
meet with
meet with
meet with

t= Dec 22-28, 2003

release with
welcome
welcome by
win
agree with
indict
win from
concern over
win
indict

Do word semantics cluster on social context?

s=ISR, *r*=PSE

<u>*t*= Jul 15-21, 2002</u>

say <-ccomp be to
release to
take control of
occupy
wound in
scuffle with
be <-xcomp meet
meet with
meet with
arrest

<u>*t*= Jul 3-9, 2006</u>

commit to
strike
carry in
continue in
reject
fire at target in
start around
ratchet pressure on
shell
hit

s=USA, *r*=FRA

<u>*t*= Feb 2-8, 1998</u>

travel <-xcomp meet with
consider
meet with
meet with
meet with

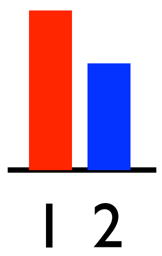
<u>*t*= Dec 22-28, 2003</u>

release with
welcome
welcome by
win
agree with
indict
win from
concern over
win
indict

Clustering approach: Mixed-membership models
("topic models," "admixtures")

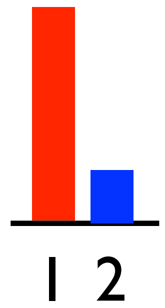
Contextual event class probabilities

s=ISR, *r*=PSE

$\theta_{s,r,t} =$ 

_t_= Jul 15-21, 2002

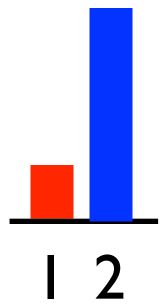
say <-ccomp be to
release to
take control of
occupy
wound in
scuffle with
be <-xcomp meet
meet with
meet with
arrest

$\theta_{s,r,t} =$ 

_t_= Jul 3-9, 2006

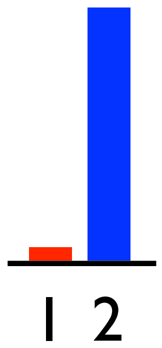
commit to
strike
carry in
continue in
reject
fire at target in
start around
ratchet pressure on
shell
hit

s=USA, *r*=FRA

$\theta_{s,r,t} =$ 

_t_= Feb 2-8, 1998

travel <-xcomp meet with
consider
meet with
meet with
meet with

$\theta_{s,r,t} =$ 

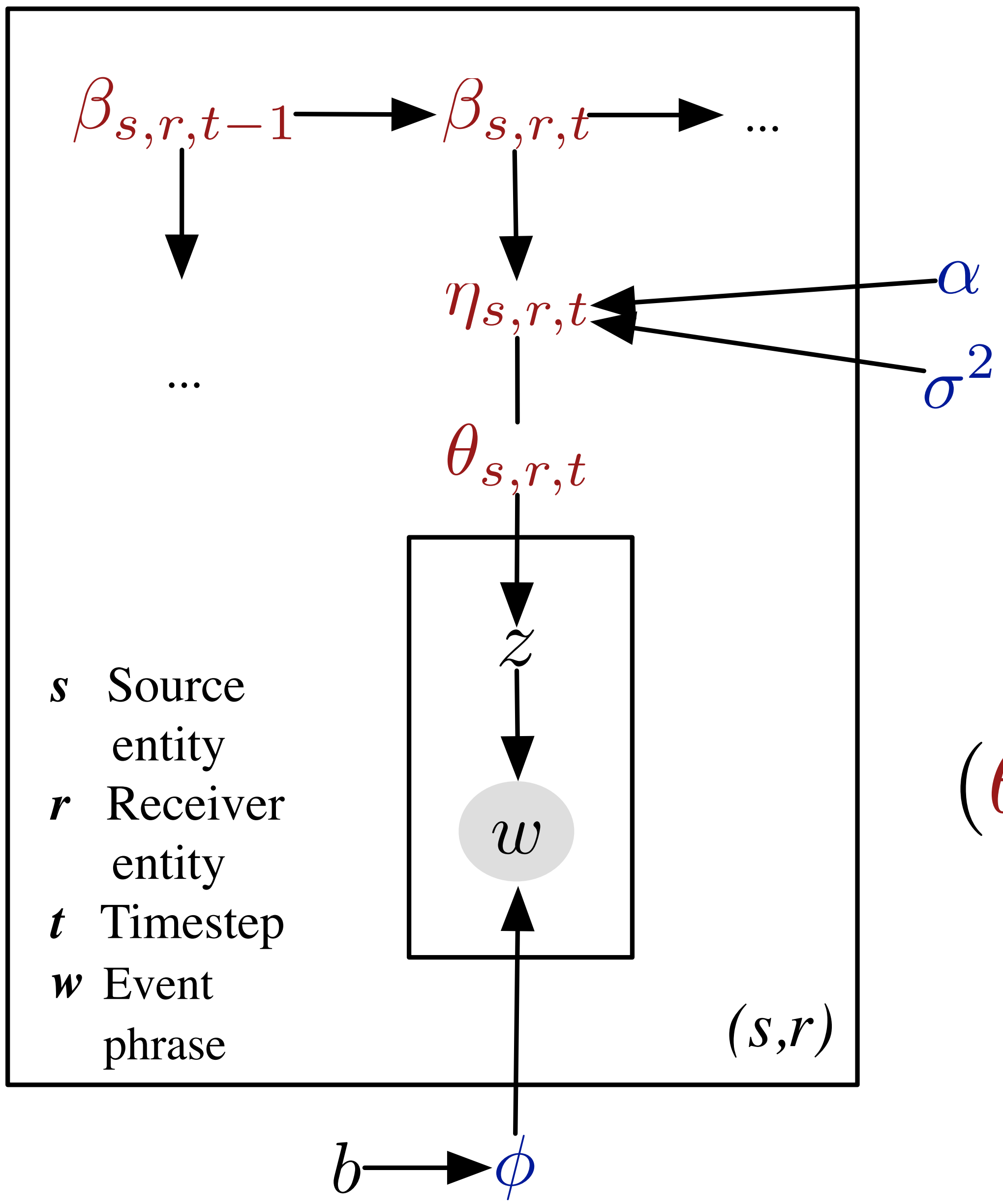
_t_= Dec 22-28, 2003

release with
welcome
welcome by
win
agree with
indict
win from
concern over
win
indict

Event class dictionaries ϕ_1 ϕ_2

agree with, arrest, be <-xcomp meet, carry in, commit to, concern over, consider, continue in, fire at target in, hit, indict, meet with, occupy, ratchet pressure on, reject, release to, release with, say <-ccomp be to, scuffle with, shell, start around, strike, take control of, travel <-xcomp meet with, welcome, welcome by, win, win from, wound in

Model



Event prior models

M1: independent contexts

M2: temporal smoothing

[Blei and Lafferty 2006, Quinn and Martin 2002]

Adjacent timestep similarity

$$\beta_{s,r,t} \sim N(\beta_{s,r,t-1}, \mathbb{I}\tau^2)$$

$$\eta_{s,r,t} \sim N(\alpha + \beta_{s,r,t}, \mathrm{Diag}[\sigma_1^2..\sigma_K^2])$$

$$(\theta_{s,r,t})_k \propto \exp(\eta_{s,r,t,k})$$

$$z \sim \mathrm{Mult}(\theta_{s,r,t})$$

$$w \sim \mathrm{Mult}(\phi_z)$$

$$w \sim \mathrm{Mult}(\Phi\theta_{s,r,t})$$

$$\phi_k \sim \mathrm{Dir}(b)$$

K=100 $\longrightarrow$ 80 million parameters

Learning: blocked Gibbs sampling

$$p(\beta, (\eta, \theta), \sigma_1^2..\sigma_K^2, z, \phi, b \mid w)$$

$$\begin{aligned}
\beta_{s,r,t} &\sim N(\beta_{s,r,t-1}, \mathbb{I}\tau^2) \\
\eta_{s,r,t} &\sim N(\alpha + \beta_{s,r,t}, \mathrm{Diag}[\sigma_1^2..\sigma_K^2]) \\
(\theta_{s,r,t})_k &\propto \exp(\eta_{s,r,t,k}) \\
z &\sim \mathrm{Mult}(\theta_{s,r,t}) \\
w &\sim \mathrm{Mult}(\phi_z) \\
\phi_k &\sim \mathrm{Dir}(b)
\end{aligned}$$

Learning: blocked Gibbs sampling

$$p(\beta, (\eta, \theta), \sigma_1^2 .. \sigma_K^2, z, \phi, b \mid w)$$

Conjugate normal

Linear dynamical system
Forward filter backward sampler (FFBS)
[Carter and Kohn 1994, West and Harrison 1997]

$$\beta_{s,r,t} \sim N(\beta_{s,r,t-1}, \mathbb{I}\tau^2)$$

Logistic normal
Metropolis-within-Gibbs,
Laplace approximation proposal
[Hoff 2003]

$$\eta_{s,r,t} \sim N(\alpha + \beta_{s,r,t}, \mathrm{Diag}[\sigma_1^2 .. \sigma_K^2])$$

$$(\theta_{s,r,t})_k \propto \exp(\eta_{s,r,t,k})$$

$$z \sim \mathrm{Mult}(\theta_{s,r,t})$$

$$w \sim \mathrm{Mult}(\phi_z)$$

Dirichlet-multinomial
Collapsed sampling
[Griffiths and Steyvers 2005]

$$\phi_k \sim \mathrm{Dir}(b)$$

Slice sampling
[Neal 2003]

Event classes: word posteriors

Most probable phrases in ϕ_k

arrive in, visit, meet with, travel to, leave, hold with, meet, meet in, fly to, be in, arrive for talk with, say in, arrive with, head to, hold in, due in, leave for, make to, arrive to, praise

accuse, blame, say, break with, sever with, blame on, warn, call, attack, rule with, charge, say←ccomp come from, say ←ccomp, suspect, slam, accuse government ←poss, accuse agency ←poss, criticize, identify

kill in, have troops in, die in, be in, wound in, have soldier in, hold in, kill in attack in, remain in, detain in, have in, capture in, stay in, about ←pobj troops in, kill, have troops ←partmod station in, station in, injure in, invade, shoot in

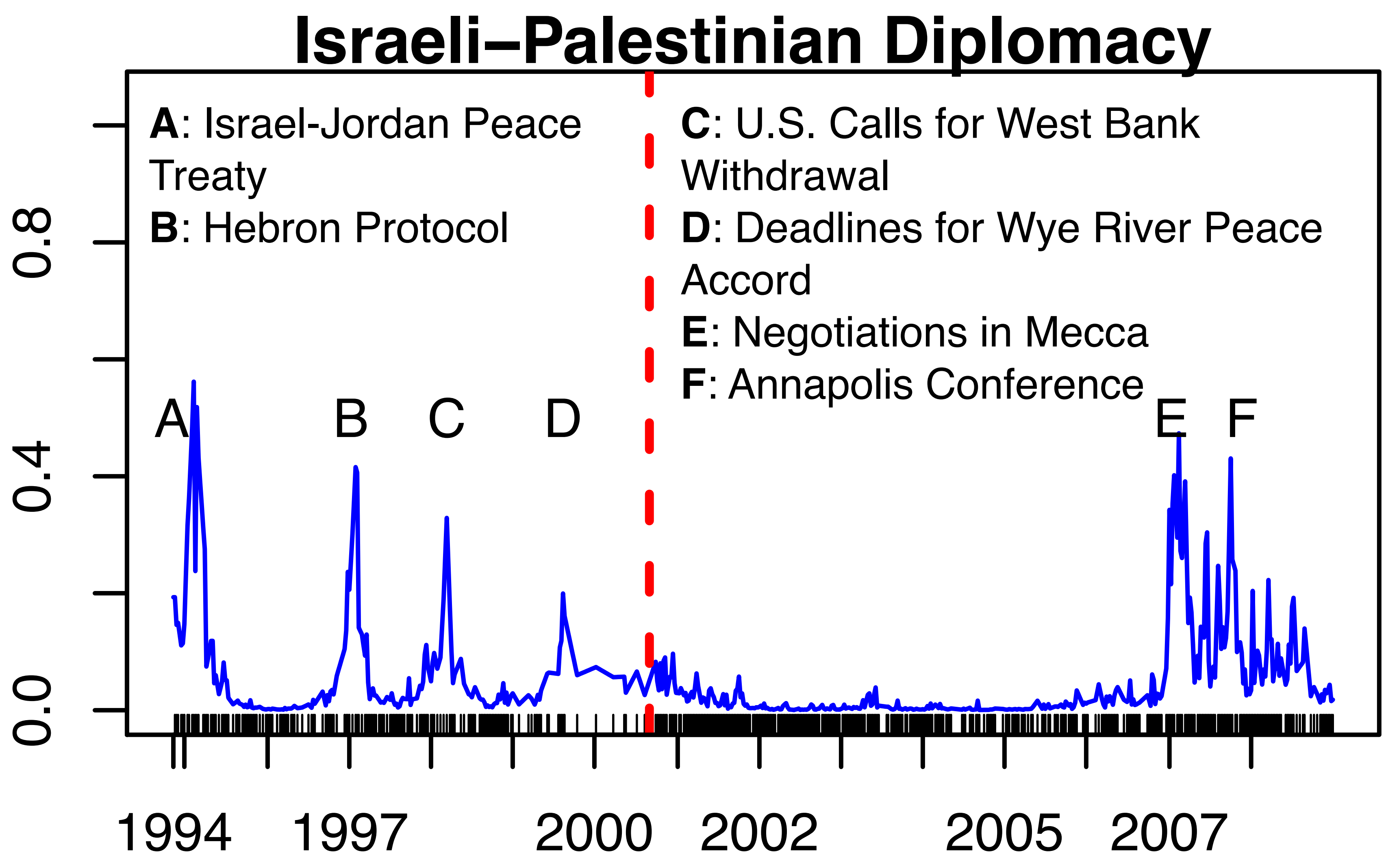
Event classes: word posteriors

Most probable phrases in ϕ_k

"diplomacy"	arrive in, visit, meet with, travel to, leave, hold with, meet, meet in, fly to, be in, arrive for talk with, say in, arrive with, head to, hold in, due in, leave for, make to, arrive to, praise
"verbal conflict"	accuse, blame, say, break with, sever with, blame on, warn, call, attack, rule with, charge, say←ccomp come from, say ←ccomp, suspect, slam, accuse government ←poss, accuse agency ←poss, criticize, identify
"material conflict"	kill in, have troops in, die in, be in, wound in, have soldier in, hold in, kill in attack in, remain in, detain in, have in, capture in, stay in, about ←pobj troops in, kill, have troops ←partmod station in, station in, injure in, invade, shoot in

Case study

meet with, sign with, praise, say with, arrive in, host, tell, welcome, join, thank, meet, travel to, criticize, leave, take to, begin to, begin with, summon, reach with, hold with



Evaluations

Society *(SocialAttributes)*

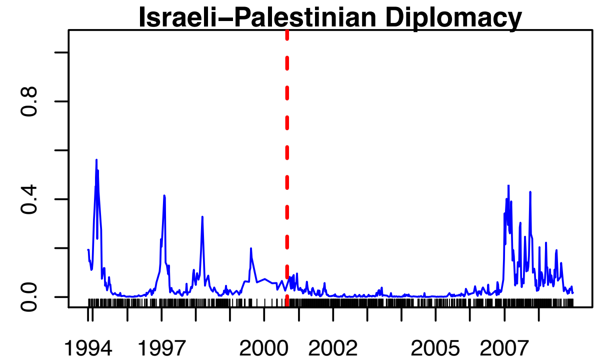
Data generation process

Writing *(TextGenerator)*

Text Data *(Text)*

Inference & collection

Inference

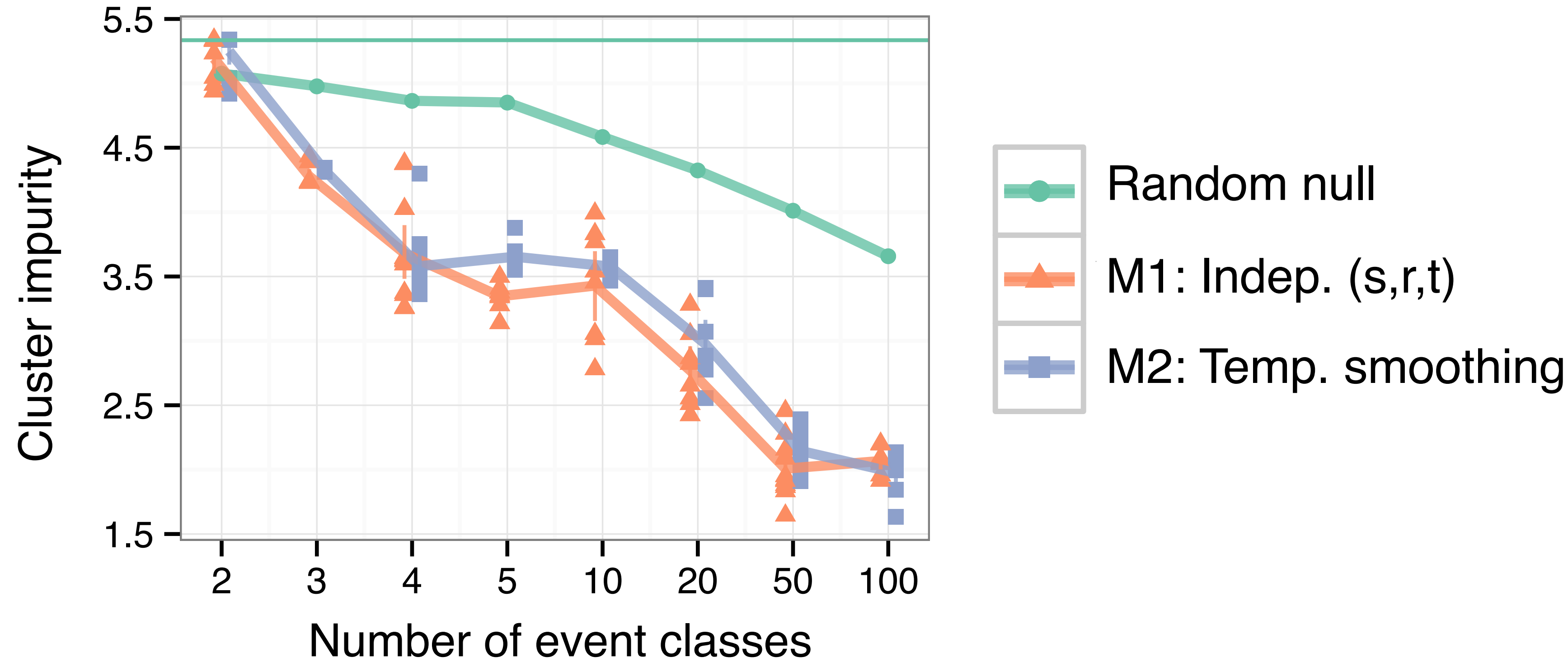


Event datasets from political science (Previous work)

Evaluation: do these correlate?

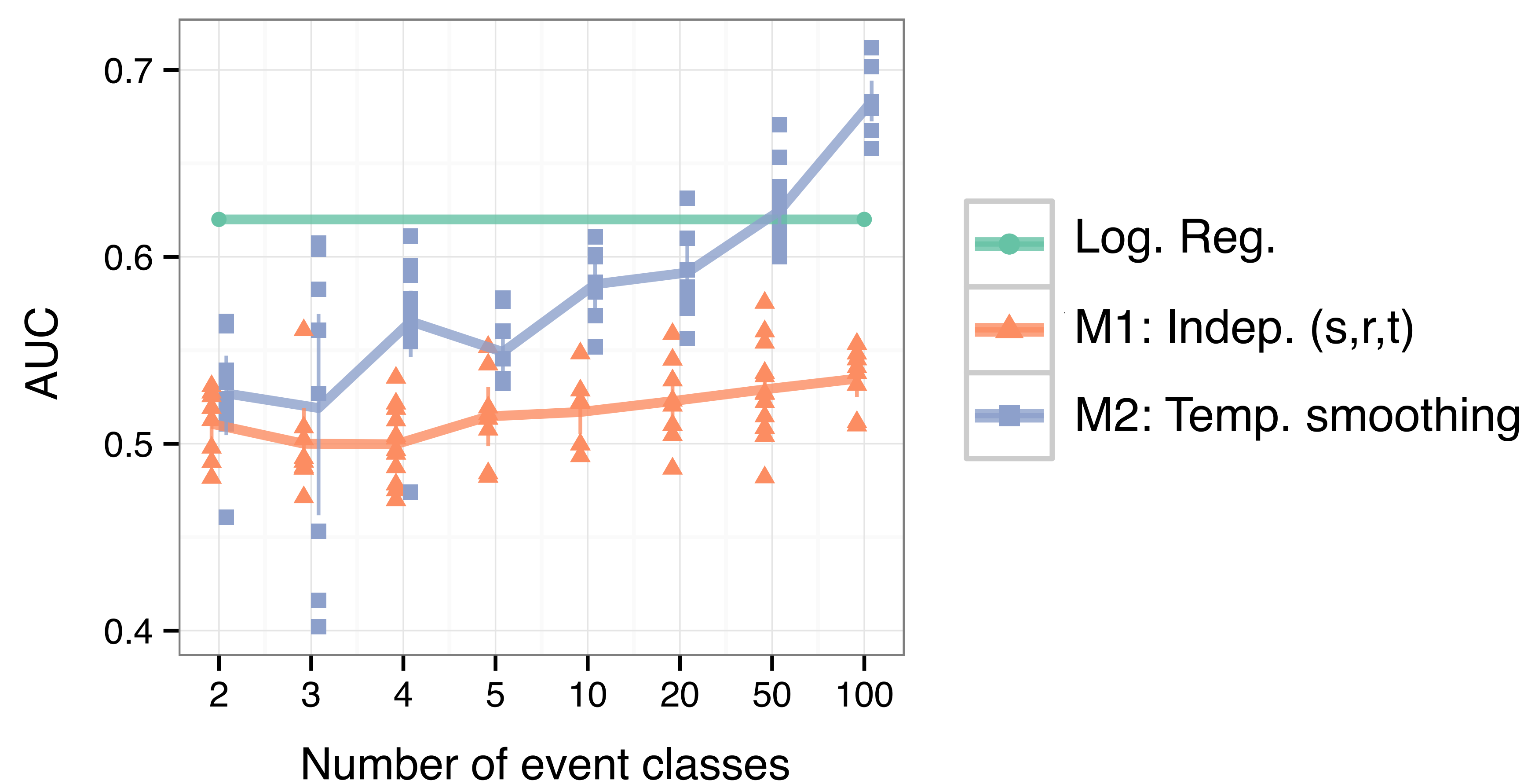
Evaluations

Better



Lexicon / Ontology reconstruction

Better



Real-world conflict reconstruction

Geographic lexical variation in Twitter

[Eisenstein, O'Connor, Smith, Xing 2010]

Geographic topic model

$$r \sim \vec{\pi}$$

$$(lat,\ lon) \sim N(\vec{\mu}_r, \mathbf{\Sigma}_r)$$

User's locations from DPMM Gaussian mixture

$$\theta \sim Dir(\vec{\alpha})$$

$$z \sim \vec{\theta}$$

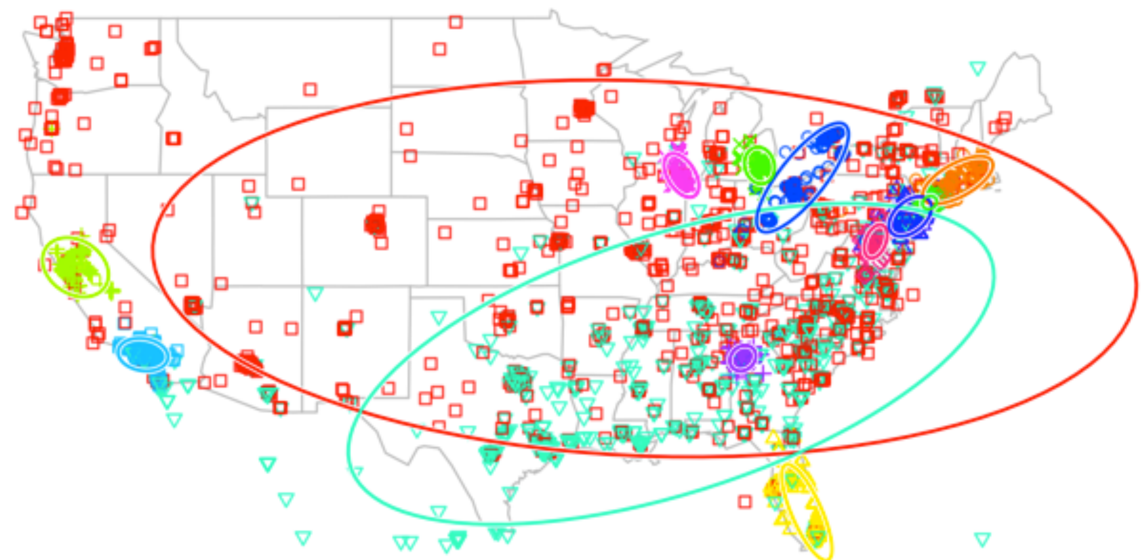
User's topics

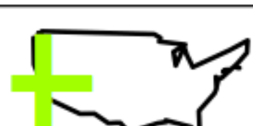
$$w \sim \exp(\vec{\eta}_{zr})$$

$$\vec{\phi}_k \sim N(\vec{a}, b^2\mathbf{I})$$

$$\vec{\eta}_{kj} \sim N(\vec{\phi}_k, s_k^2\mathbf{I})$$

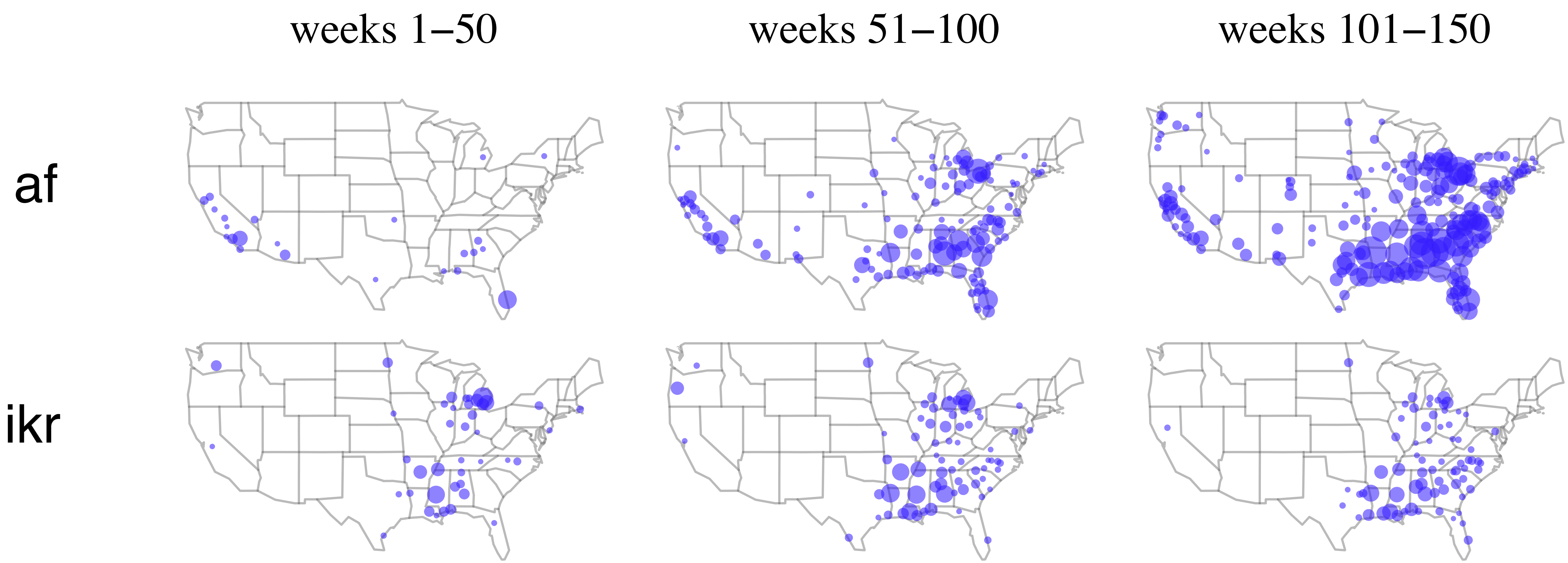
have regional variants



	"basketball"	"popular music"	"daily life"	"emoticons"	"chit chat"
	PISTONS KOBE LAKERS game DUKE NBA CAVS STUCKEY JETS KNICKS	album music beats artist video #LAKERS ITUNES tour produced vol	tonight shop weekend getting going chilling ready discount waiting iam	:) haha :d :(;) :p xd :/ hahaha hahah	lol smh jk yea wyd coo ima wassup somethin jp
Boston 	CELTICS victory BOSTON CHARLOTTE	playing daughter PEARL alive war comp	BOSTON	;p gna loveee	*ese* exam suttin sippin
N. California 	THUNDER KINGS GIANTS pimp trees clap	SIMON dl mountain seee	6am OAKLAND	*pues* hella koo SAN fckn	hella flirt hut iono OAKLAND

Social determinants of language change

[*Eisenstein, O'Connor, Smith, Xing 2012 and in review*]



Test sociolinguistic theories of how linguistic innovations diffuse

U.S. Census data

200 regions, 2600 words, 165 timesteps = 85M parameters

$$n_{w,r,t} \sim \mathrm{Binom}(N_{r,t},\ \sigma(\nu_w + \tau_{r,t} + \eta_{w,*,t} + \eta_{w,r,t})$$

$$\boldsymbol{\eta}_{w,t} \sim \mathrm{Normal}(\mathbf{A}\boldsymbol{\eta}_{w,t-1},\ \boldsymbol{\Gamma})$$

$\mathbf{A}$ autoregressive coefficients (size $R \times R$)

Social Media NLP
Part-of-speech tagger for Twitter

Example

ikr	smh	he	asked	fir	yo	last	name
!	G	O	V	P	D	A	N

HMM word cluster (features for CRF tagger)

yeah yea nah naw yeahh nooo yeh noo noooo yeaa **ikr** nvm yeahhh nahh nooooo yh yeaaa yeaah yupp naa yeahhhh yeaaahiknow werd noes nahhh naww yeaaaa shucks yeaaaah yeahhhhh naaa naah nawl nawww yehh ino yeaaaaa yeeah yeeeah wordd yeaahh nahhhh naaah yeahhhhhh yeaaaaah naaaa yeeeeah nall yeaaaaaa

http://www.ark.cs.cmu.edu/TweetNLP/

[*Gimpel, Schneider, O'Connor, Das, Mills, Eisenstein, Heilman, Yogatama, Smith, 2011*]
[*Owoputi, O'Connor, Dyer, Gimpel, Schneider, Smith, 2013*]

Text Analysis for Social Science

- Tools for discovery and measurement
  - Social, spatial, temporal context
  - Probabilistic models
  - A little bit of NLP can go a long way
- Future work
  - Text visualization / exploration tools
  - Semantics: belief structures from text
  - Incorporate a-priori knowledge
  - Causal inference