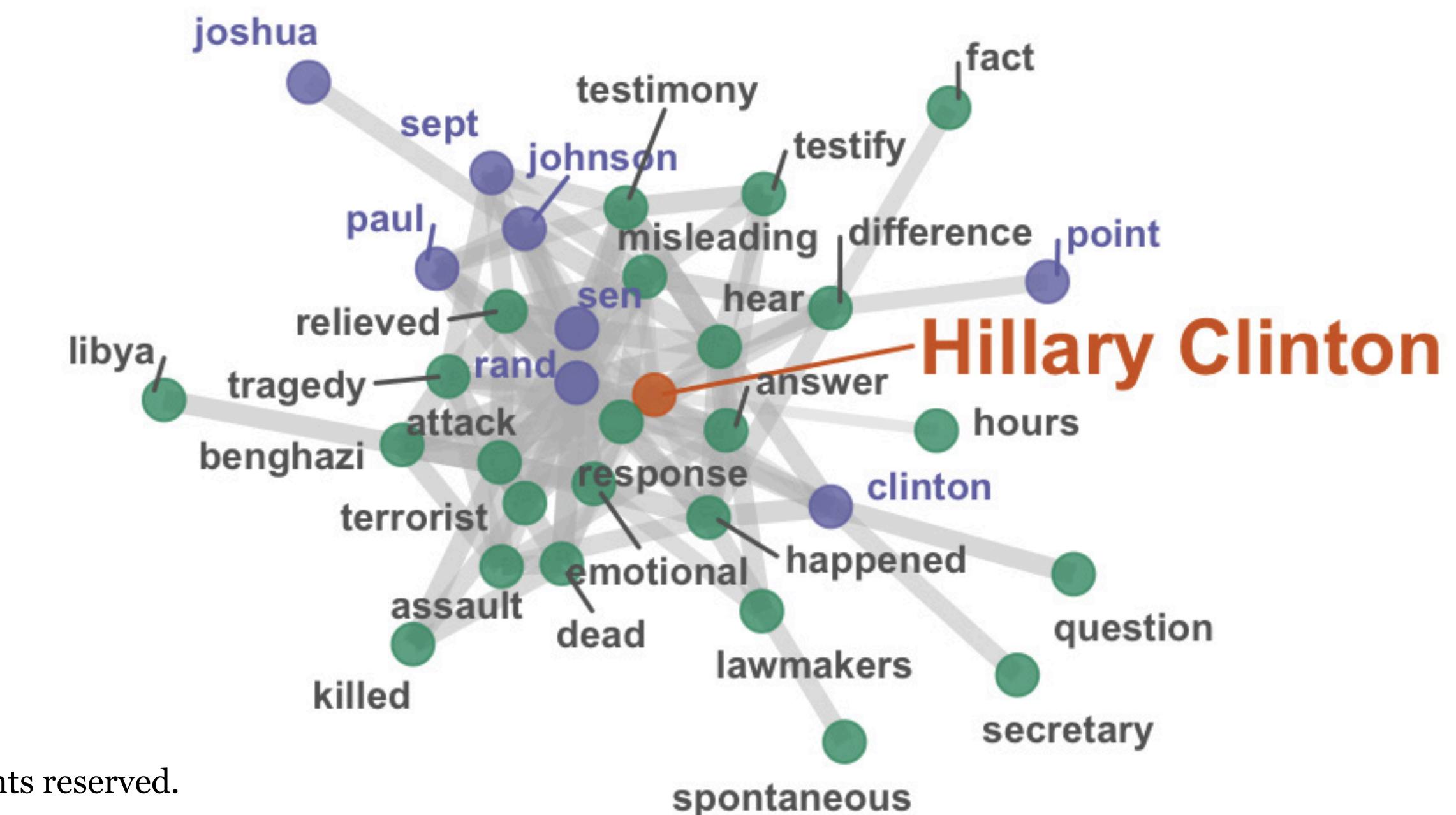


Mike Safar
KMS Product Strategies
August 21, 2018

Improved Insights and Visualizations for Unstructured Data Mining

Creative Data Mining of Hillary Clinton's State Department Emails





Sophisticated data analytics and visualization are key features of mainstream applications

AI and unsupervised machine-learning are being democratized

However...

*The reality for many is **usability and insight** aren't fully delivering on the potential*

Low user traction...

50% *of users indicate they are using
analytics technologies
less than a couple of times per year*

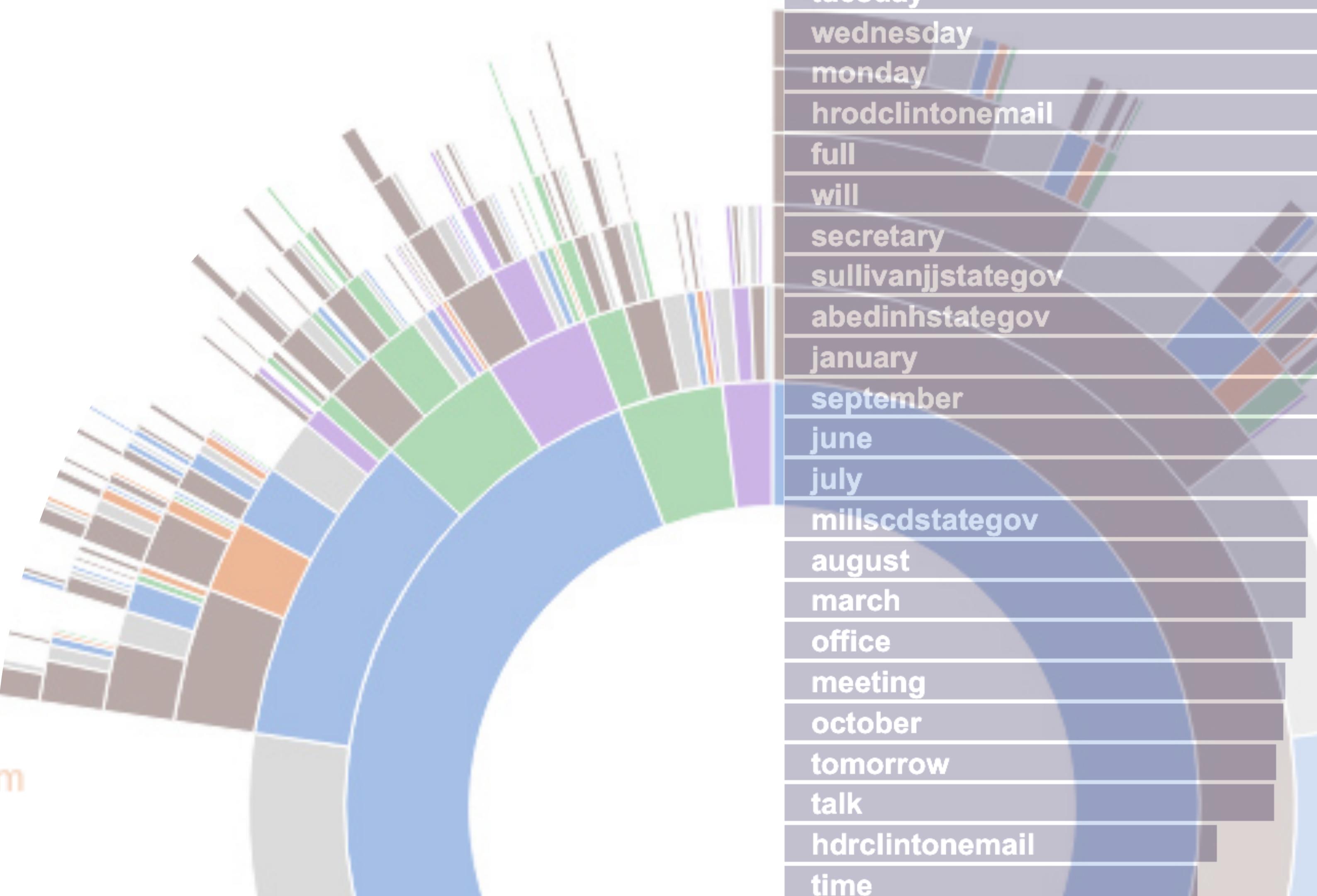
10% *are using social media analytics*

Why?

Insights Obscured in Noise

- *Low-value data*
 - *Low-dimensionality*
 - *Lack of context*

request
message
philippe thomas
dos best cheryl bill
pis jan oschedule mon tom
letter tomorrow today office
jake thu lauren jiloty
speech email ionavalmoro
discuss state will sunfull work
sat update humaabedin
p fyi cherylmills
update t: clinton
sent house
sure william
tue david
nt house
nt

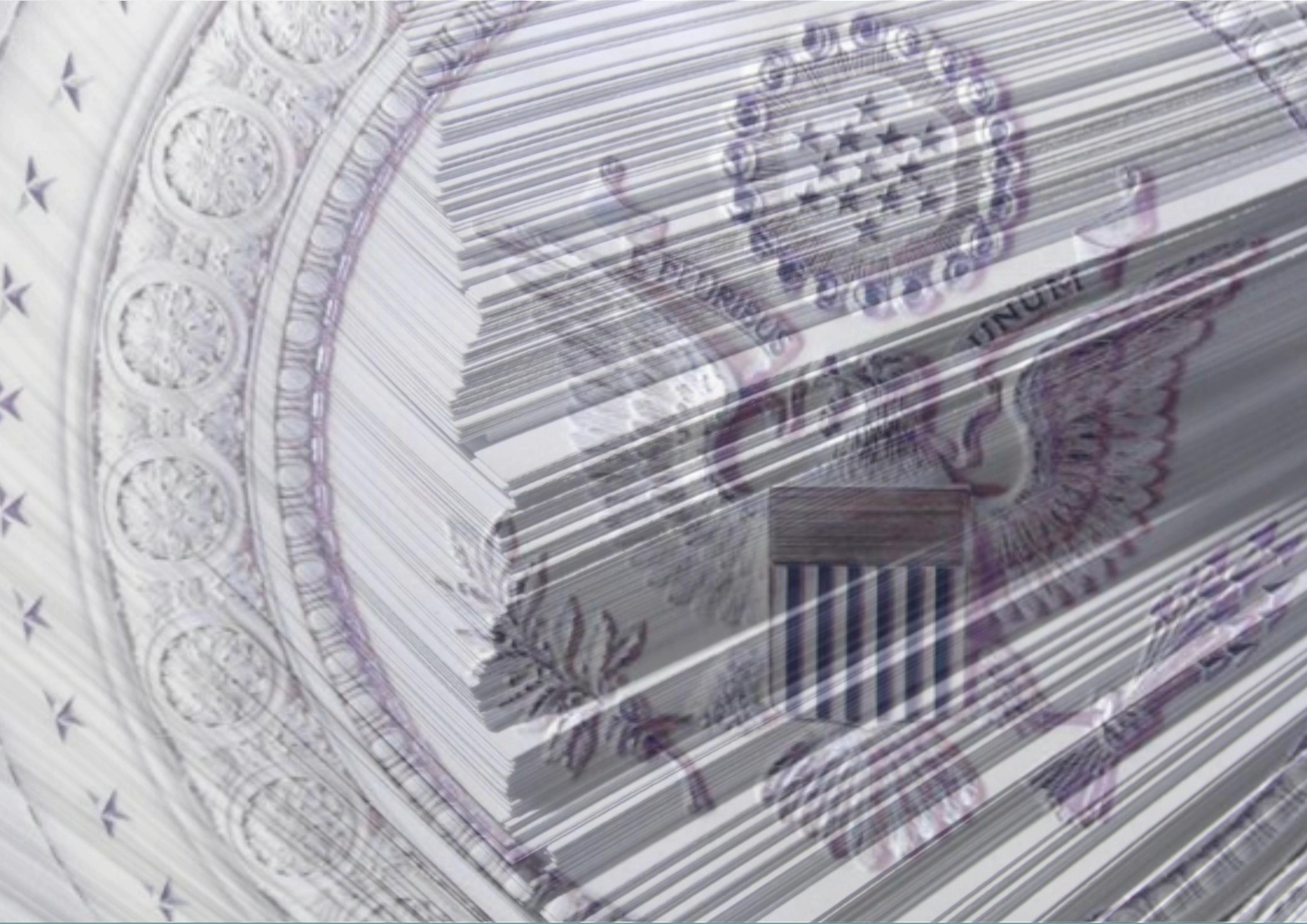


The Challenge

- Improve usability of standard text-mining visualizations
- Derive summary-level insights quickly
- Useful even on demanding datasets
- Repeatable methodology

Subject:
Hillary Clinton's
State Department Email Release



A large stack of US one-hundred-dollar bills is shown from a top-down perspective, slightly angled. The bills are tightly packed, creating a textured, layered effect. The 'In U.S. Mint' and 'U.S. Note' text is visible on the bills.

The Dataset

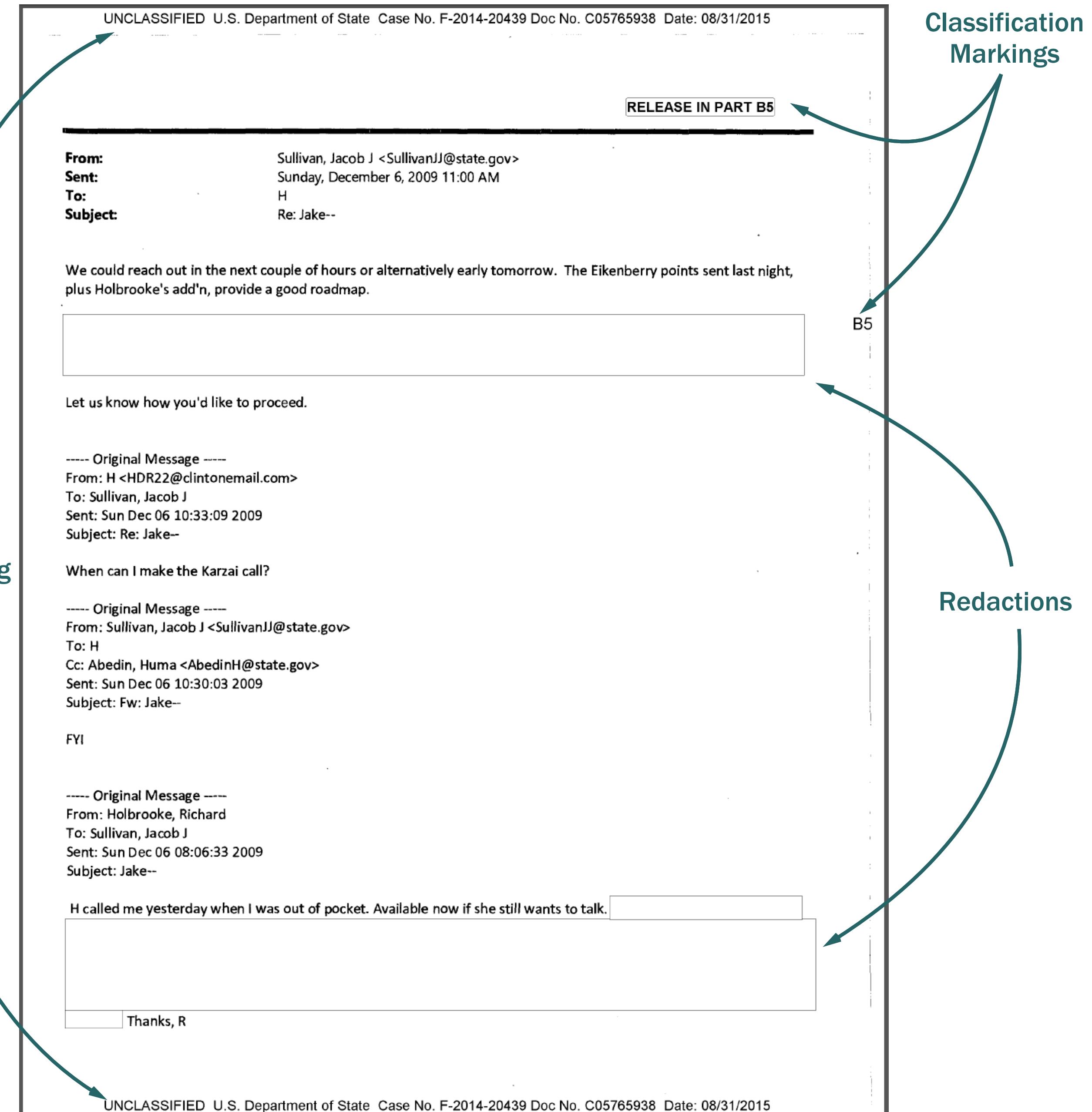
- Sec. Hillary Clinton's private email server (@clintonemail.com)
- 27,159 emails & docs
- 2009-2012
- No metadata
- Redacted and marked

To: H <hrod17@clintonemail.com>

Typical Image

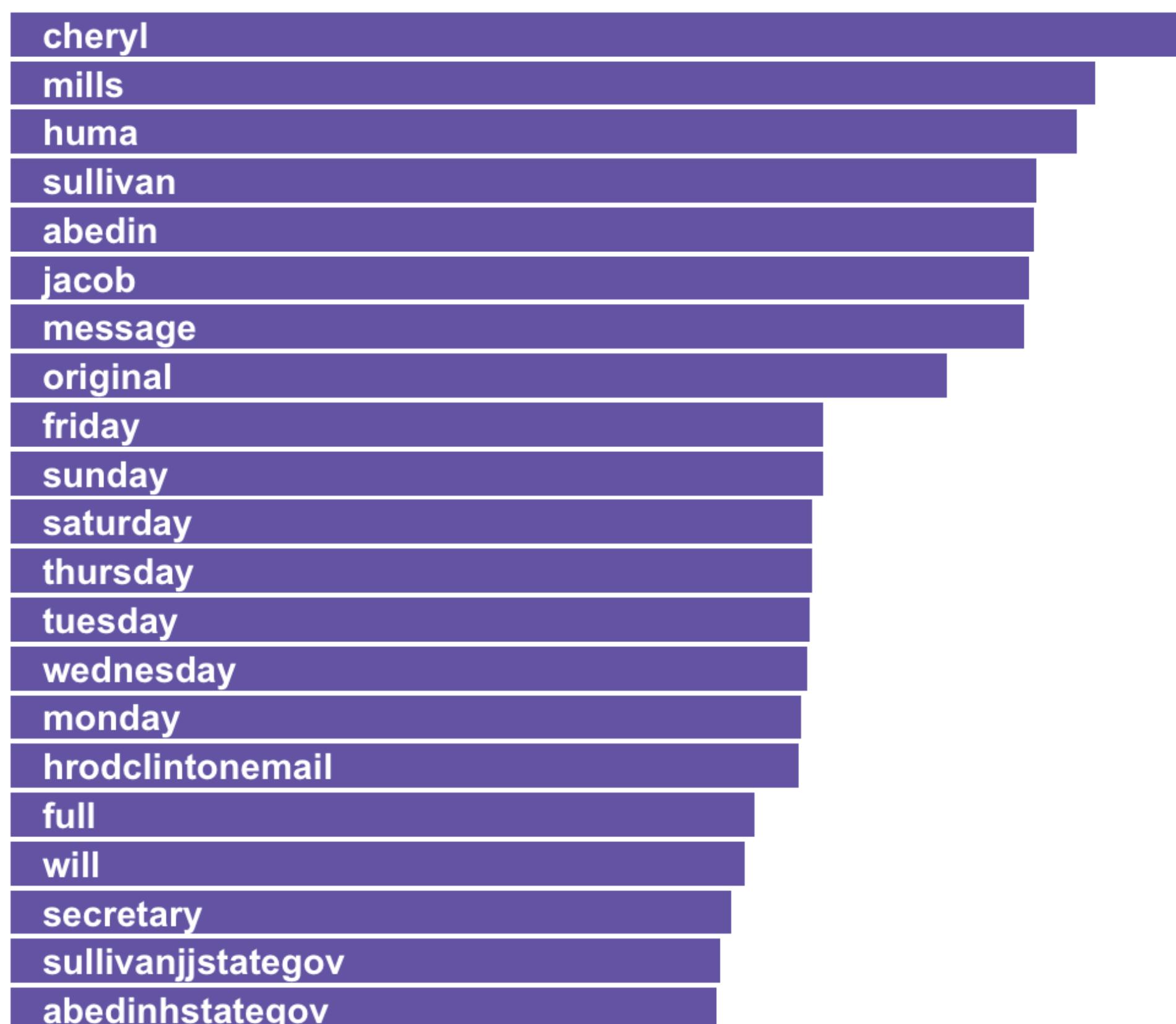
Unlike original email, these images come with no metadata, or even machine-readable text; OCR is required

Document Stamping and Numbering

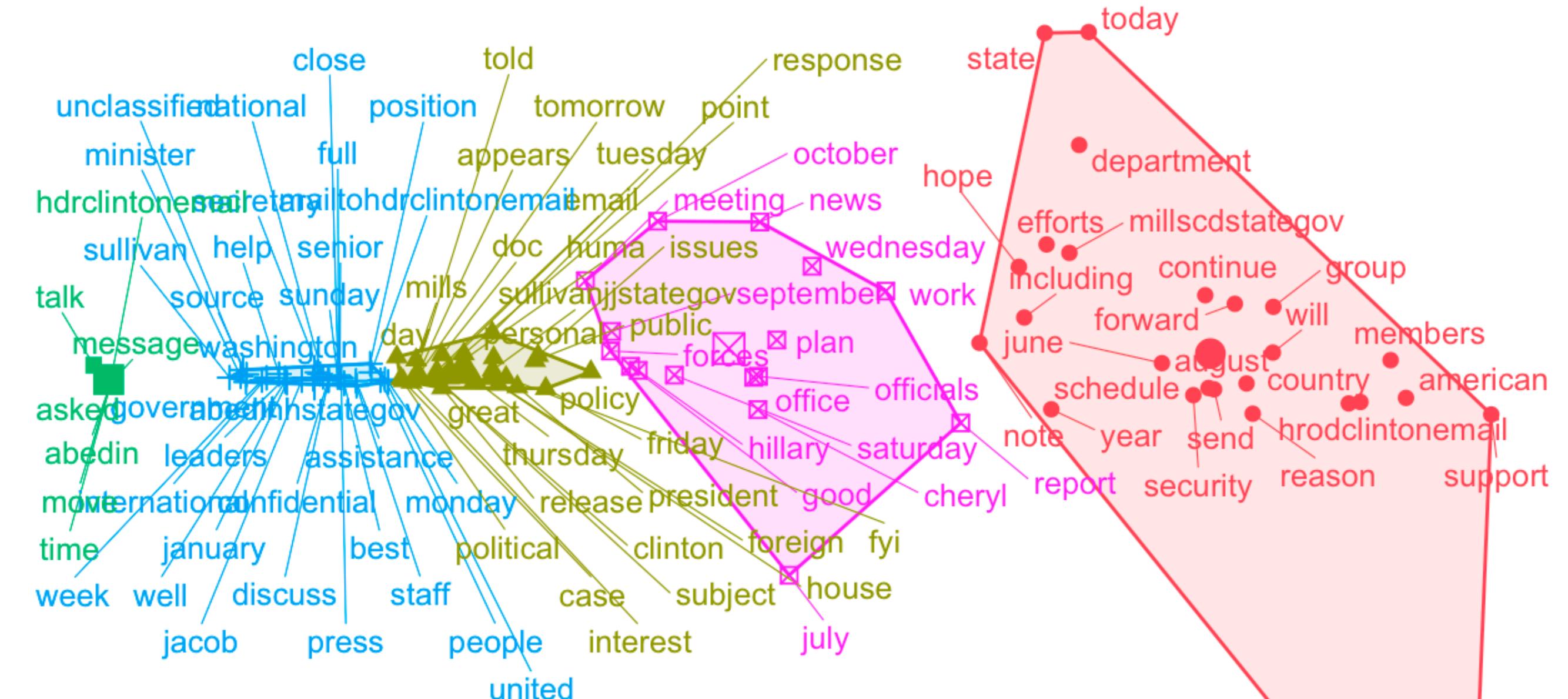


The data is searchable, but as far as visual analytics go, it's a bust...

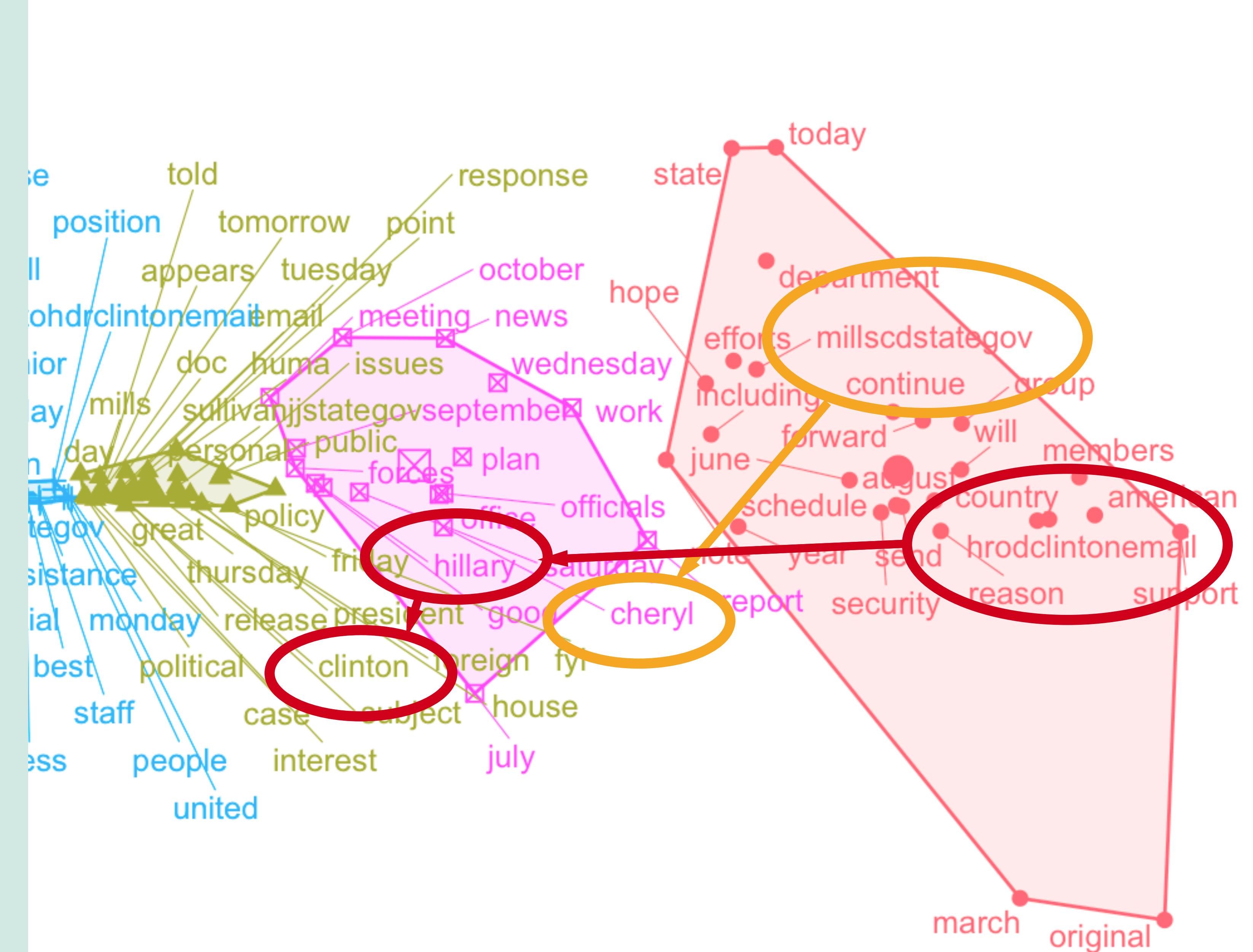
Relative Word Frequencies for HRC Dataset



Clustering: Single-word Co-occurrence by Document

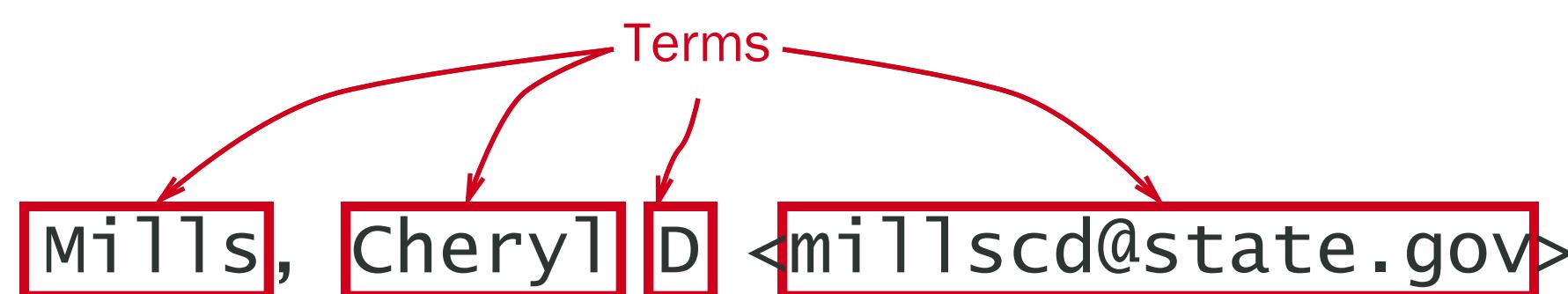


K-means clustering and related graph are saturated and confusing



Specific Challenges with Clinton Dataset

- Unfielded email addressing is split into multiple terms
- Document stamping and marking skews the data



- OCR errors in the data mask email addresses

hrod17@clintonemail.com

UNCLASSIFIED U.S. Department of State Case No. F-2014-20439

- Redaction removes some identifying info

----- Original Message -----
From: Holbrooke, Richard
To: Sullivan, Jacob J
Sent: Sun Dec 06 08:06:33 2009
Subject: Jake--

H called me yesterday when I was out of pocket. Available now if she still wants to talk.

Thanks, R

Email as a source poses additional challenges:

- Emoticons, abbreviations, spelling errors
- Repeated content in reply chains
- Boilerplates and signatures skew data
- Shifting topics and language over time



Summarization and visualization are hampered under these conditions!

Enhanced Data Prep

▶ Scrub

- Normalize
- Extract

■ Type

Tune Cleaning Regimen to Corpus

1. Remove common OCR errors in the data:

hrod17@clintonemail.com*i*

hrod17@clintonemail.com

2. Remove unnecessary doc-stamping and markings.

3. Discard irrelevant, but non-typical stop words

## [1]	afternoon	am	anytime	apr	arrive
## [6]	asap	attached	aug	availability	bcc
## [11]	btw	calls	cc	check	checking
## [16]	confirmed	connect	copies	copy	copying
## [21]	cscC	day	dec	depart	departing
## [26]	departs	drive	email	emailed	fax
## [31]	feb	finish	flight	forward	fri
## [36]	full	fw	fyi	going	good
## [41]	jan	jul	jun	list	mailto
## [46]	mar	meet	meeting	meetings	messages
## [51]	minutes	mon	month	morning	mtg
## [56]	noon	sat	offered	plane	plane

Enhanced Data Prep

- Scrub
- Normalize
- Extract
- Type

Normalize Around Known Entities

Collapse critical term sequences into unique identifiers:

[1] " UNCLASSIFIED U.S. Department of State Case No. F-2014-20439 Doc No. C057
65938 Date: 08/31/2015
4 TERMS RELEASE IN PART B5 From:
Sullivan, Jacob J <sullivanjj@state.gov> Sent: Sunday,
December 6, 2009 11:00 AM To: H Subject:
Re: Jake- We could reach out in the next couple of hours or alternatively early tomorrow. The Eikenberry points sent last night, plus Holbrooke's add'n, provide a good roadmap. Let us know how you'd like to proceed. Original Message From: H <HDR22@clintonemail.com> To: Sullivan, Jacob i Sent: Sun Dec 06 10:33:09 2009 Subject: Re: Jake-- When can I make the Karzai call? Original Message From: Sullivan, Jacob J <sullivanjj@state.gov> To: H Cc: Abedin, Huma <AbedinH@state.gov> Sent: Sun Dec 06 10:30:03 2009 Subject: Fw: Jake-- FYI Original Message From: Holbrooke, Richard To: Sullivan, Jacob i Sent: Sun Dec 06 08:06:33 2009 Subject: Jake-- H called me yesterday when I was out of pocket. Available now if she still wants to talk.
Thanks, R UNCLASSIFIED U.S. Department of State Case No. F-2014-20439 Doc No. C05765938 Date: 08/31/2015 \f"

Matched with Directory or other ID database

Moniker	First	Last	Email1
CHERYLMILLS	Cheryl	Mills	millscd@state.gov
HILLARYCLINTON	Hillary	Clinton	hrod17@clintonemail.com
JAKESULLIVAN	Jacob	Sullivan	sullivanjj@state.gov
HUMAABEDIN	Huma	Abedin	abedinh@state.gov
LAURENJILOTY	Lauren	Jiloty	JilotyLC@state.gov
LONAVALMORO	Lona	Valmoro	ValmoroLJ@state.gov
MONICAHANLEY	Monica	Hanley	HanleyMR@state.gov

. C05765938 Date: 08/31/2015
1 TERM RELEASE IN PART B5 From:
DAY, MONTH 6,2009 1
Re: Jake- We could reach out in the next couple of hours or alternatively early tomorrow. The Eikenberry points sent last night, plus Holbrooke's add'n, provide a good roadmap. Let us know how you'd like to proceed. From: H HILLARYCLINTON To: JAKESULLIVAN i Sent: Sun Dec 06 10:33:09 2009 Subject: Re: Jake-- When can I make the Karzai call? From: JAKESULLIVAN HILLARYCLINTON Cc: HUMAABEDIN Sent: Sun Dec 06 10:30:03 2009 Subject: Fw: Jake-- FYI From: Holbrooke, Richard To: JAKESULLIVAN i Sent: Sun Dec 06 08:06:33 2009 Subject: Jake-- H called me DAY when I was out of pocket. Available now if she still wants to talk. Thanks, R . C05765938 Date:

Enhanced Data Prep

- Scrub
- Normalize
- Extract
- Type

Extract Actionable Metadata

Identify metadata and extract...

[1] " UNCLASSIFIED U.S. Department of State Case No. F-2014-20439 Doc No. C057
65938 Date: 08/31/2015
RELEASE IN PART B5 From:
Sullivan, Jacob J <SullivanJJ@state.gov> Sent: Sunday,
December 6, 2009 11:00 AM To: H Subject:
Re: Jake- We could reach out in the next couple of hours or alternatively early tomorrow. The Eikenberry points sent last night, plus Holbrooke's add'n, provide a good roadmap. Let us know how you'd like to proceed. Original Message From: H <HDR22@clintonemail.com> To: Sullivan, Jacob i Sent: Sun Dec 06 10:33:09 2009 Subject: Re: Jake-- When can I make the Karzai call? Original Message From: Sullivan, Jacob J <Sullivanii@state.gov> To: H Cc: Abedin, Huma <AbedinH@state.gov> Sent: Sun Dec 06 10:30:03 2009 Subject: Fw: Jake-- FYI Original Message From: Holbrooke, Richard To: Sullivan, Jacobi Sent: Sun Dec 06 08:06:33 2009 Subject: Jake-- H called me yesterday when I was out of pocket. Available now if she still wants to talk.
Thanks, R UNCLASSIFIED U.S. Department of State Case No. F-2014-20439 Doc No. C05765938 Date: 08/31/2015 \f"

Sometimes, metadata such as sent date can be effectively extracted

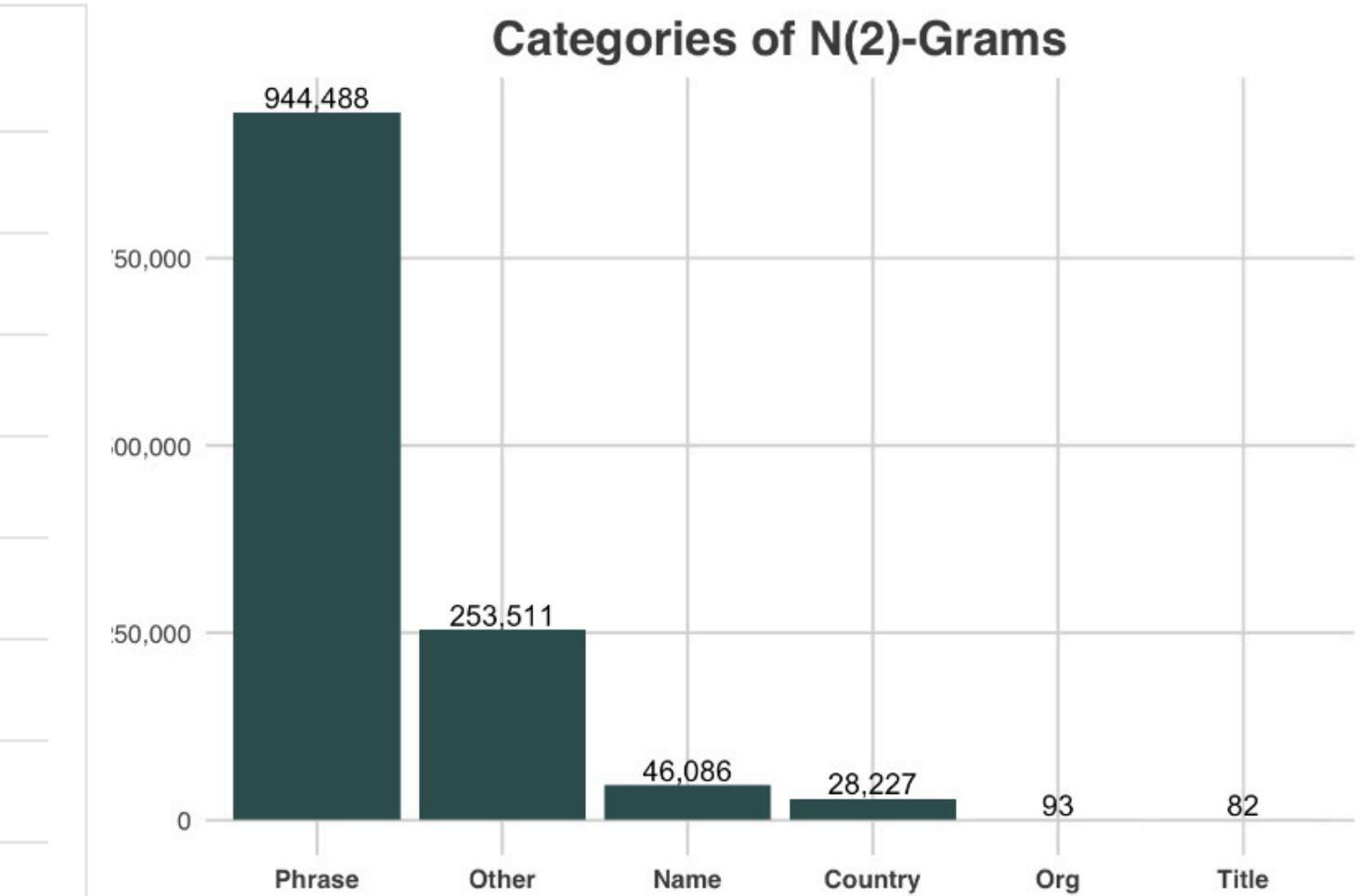
Enhanced Data Prep

- Scrub
- Normalize
- Extract
- ▶ Type

Add Dimension Through Typing

Apply context to content by matching N-grams with dictionaries and other algorithms to achieve entity extraction:

word1	word2	n	type
united	states	7,151	Country
state	department	6,233	Org
white	house	4,699	Org
department	state	3,812	Org
prime	minister	3,392	Title
secretary	state	3,187	Title
secretary	office	3,014	Org
human	rights	2,544	Phrase
secretary	clinton	2,448	Phrase
foreign	policy	2,029	Phrase
middle	east	1,988	Phrase

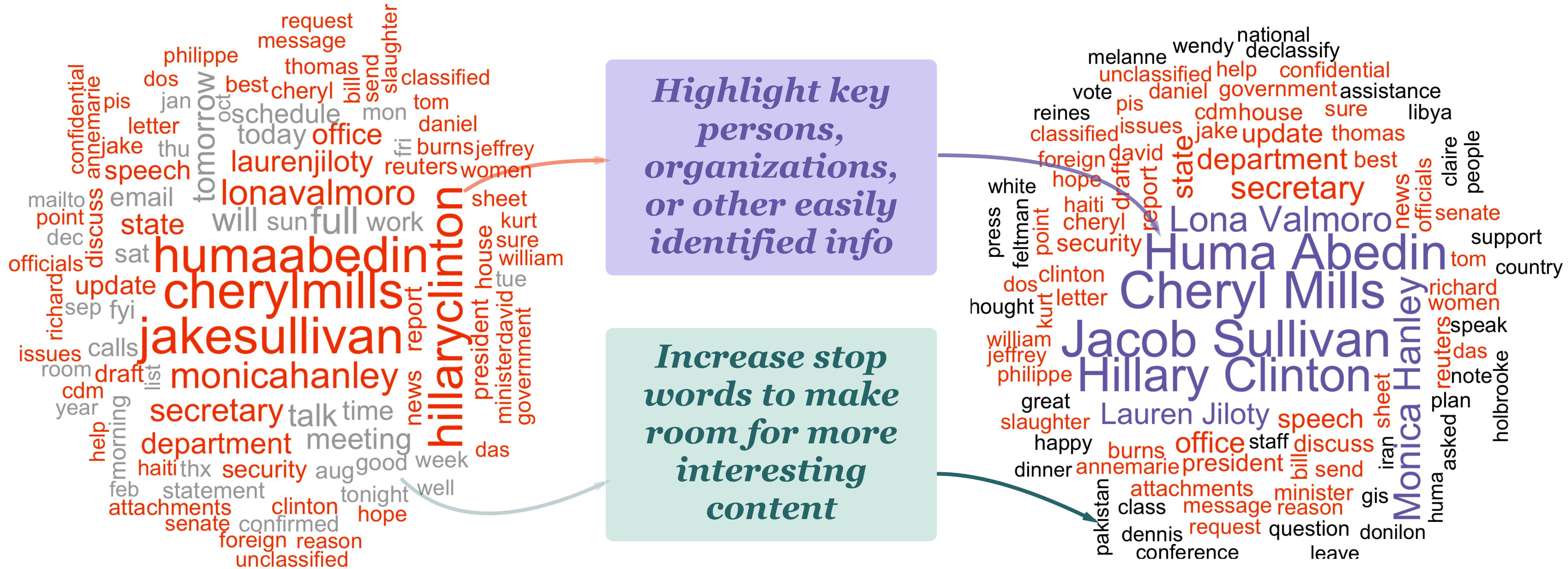


Building Blocks

Steps towards better understanding...

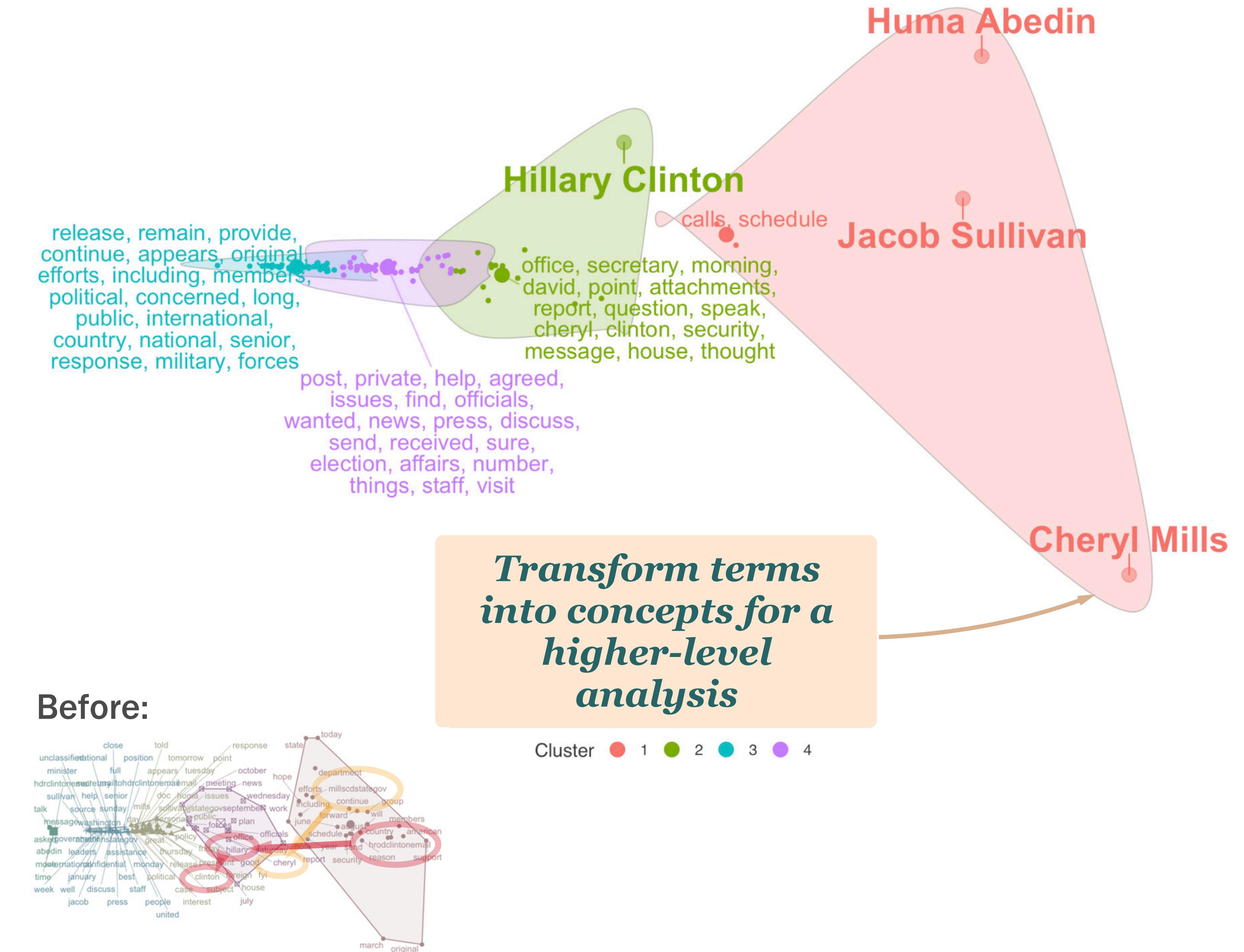
Filter Noisy Content

Fixed stop word lists are insufficient; tune stop words to the content and use case



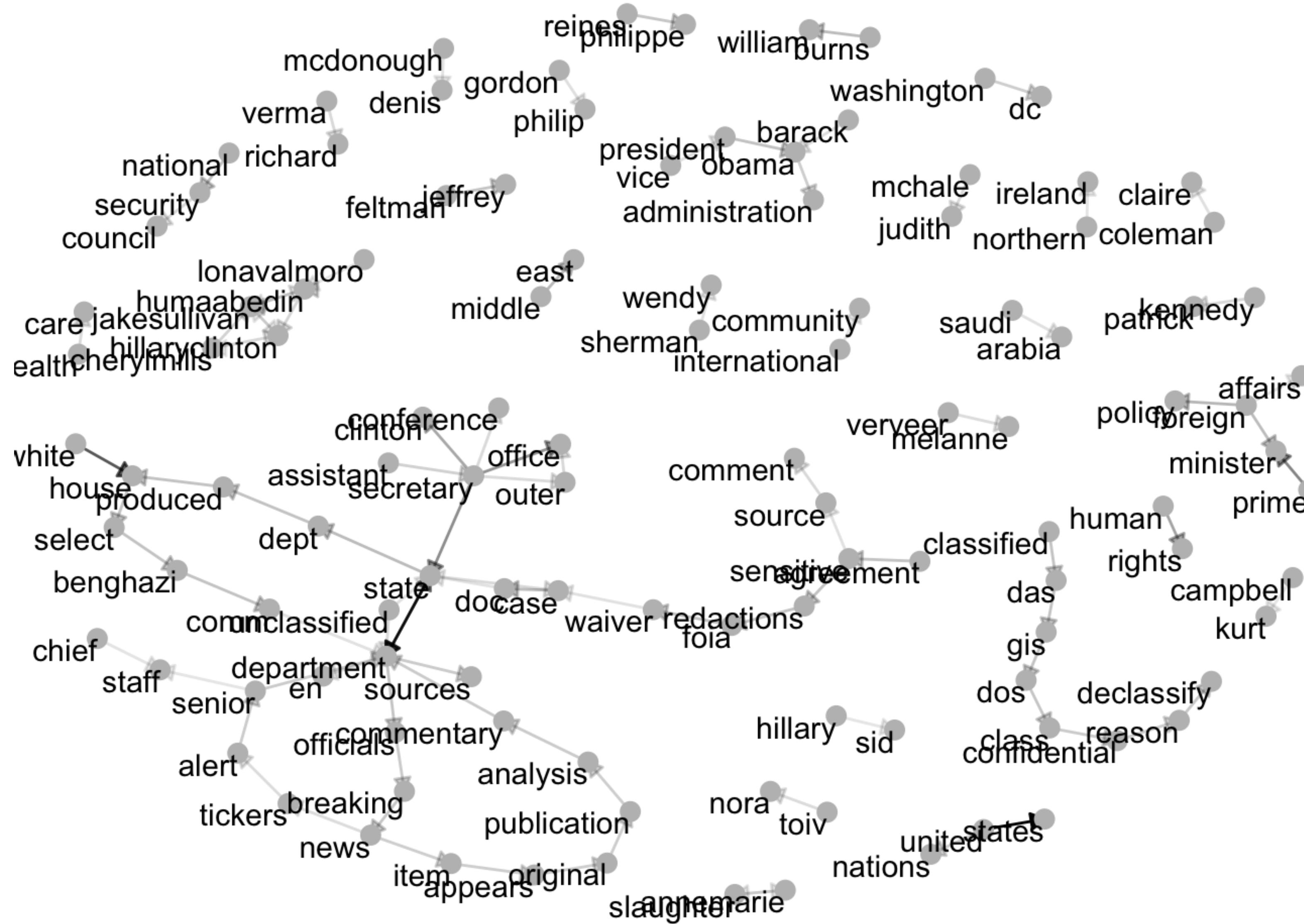
Focus on Logical Relationships

*Normalization
gets you closer to
the concepts*



Before

1 and 2 dimensional graphs have limits on the volume of summary information the viewer can consume



Country	Name	Org	Other	Phrase	Title
● saudi-arabia	● barack-obama ● mcdonough-denis ● kennedy-patrick ● campbell-kurt ● gordon-philip	● obama-administration ● united-nations ● security-council	● sensitive-redactions ● redactions-foia ● news-tickers ● tickers-alert ● verveer-melanne	● vice-president ● president-obama ● foreign-policy ● northern-ireland ● foreign-affairs ● agreement-sensitive ● sensitive-source ● source-comment ● health-care ● national-security ● reason-declassify ● middle-east ● news-item ● item-appears ● confidential-reason ● class-confidential ● appears-original ● foia-waiver ● breaking-news ● dos-class ● original-publication ● senior-staff ● officials-breaking ● alert-senior ● gis-dos ● waiver-case ● publication-analysis ● das-gis ● classified-das ● analysis-commentary ● senior-department ● department-officials ● commentary-department ● case-doc	● prime-minister ● foreign-minister ● chief-staff
● united-states	● sherman-wendy ● burns-william ● mchale-judith ● toiv-nora				

Add Depth

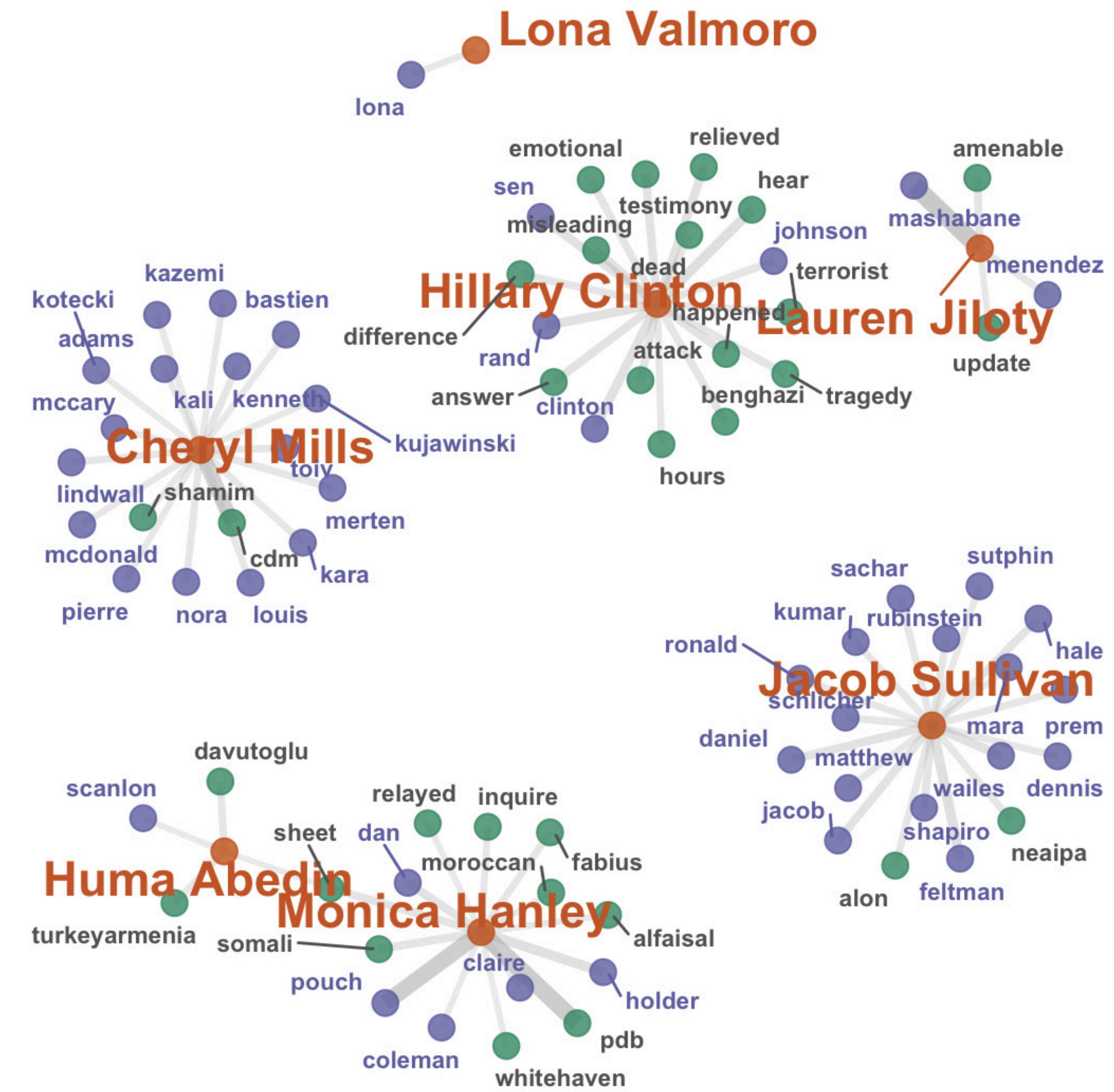
*Within limits,
multi-variate
graphs enhance
not only
contextual
understanding,
but basic
readability*

Boosting Signal in Unstructured Data Analytics

Surfacing buried insights...

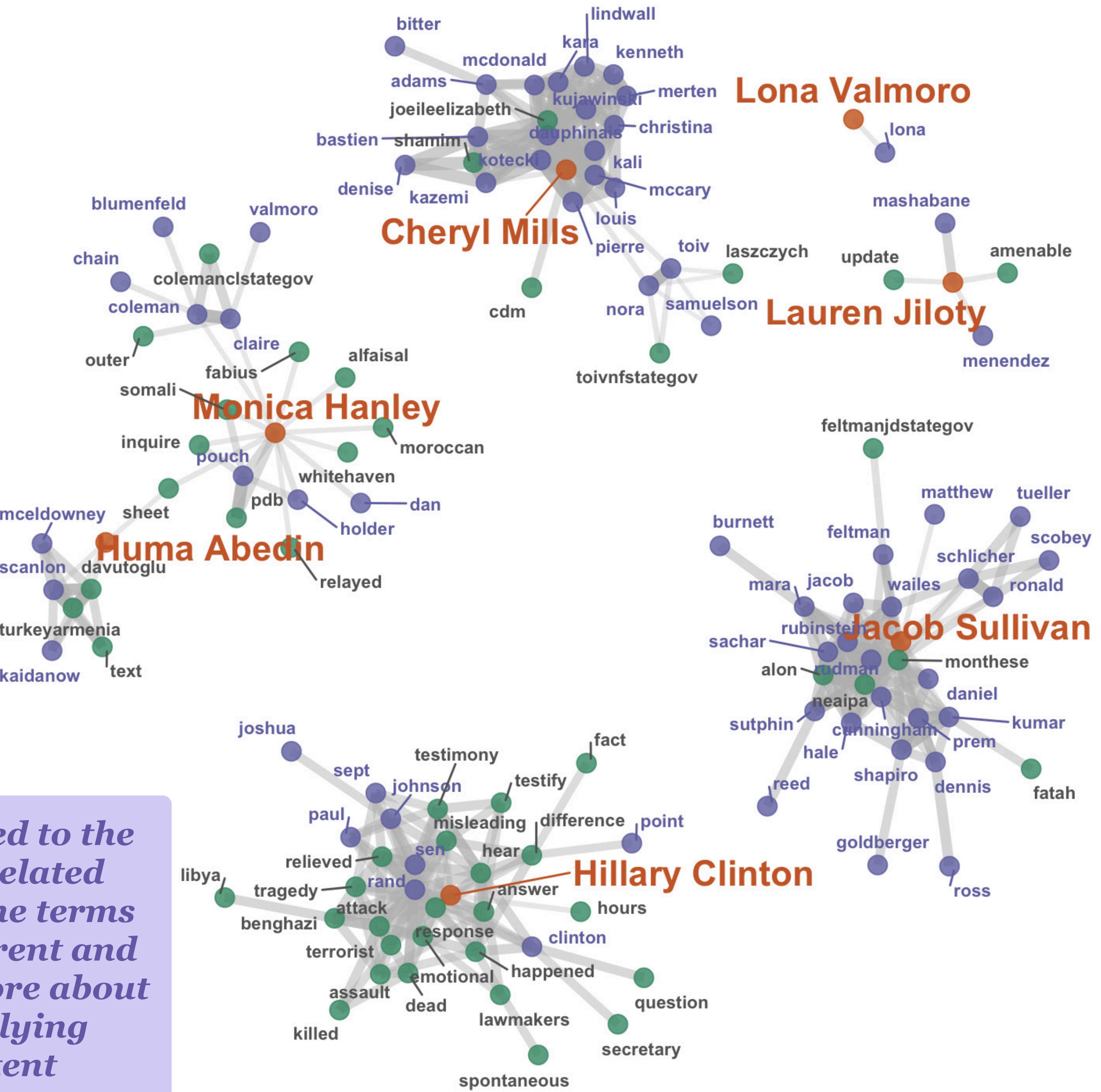
Term Clustering by Person

*Build topical
interest clouds by
correlating term
relationships to
normalized
persons*



N-Gram Clustering by Person

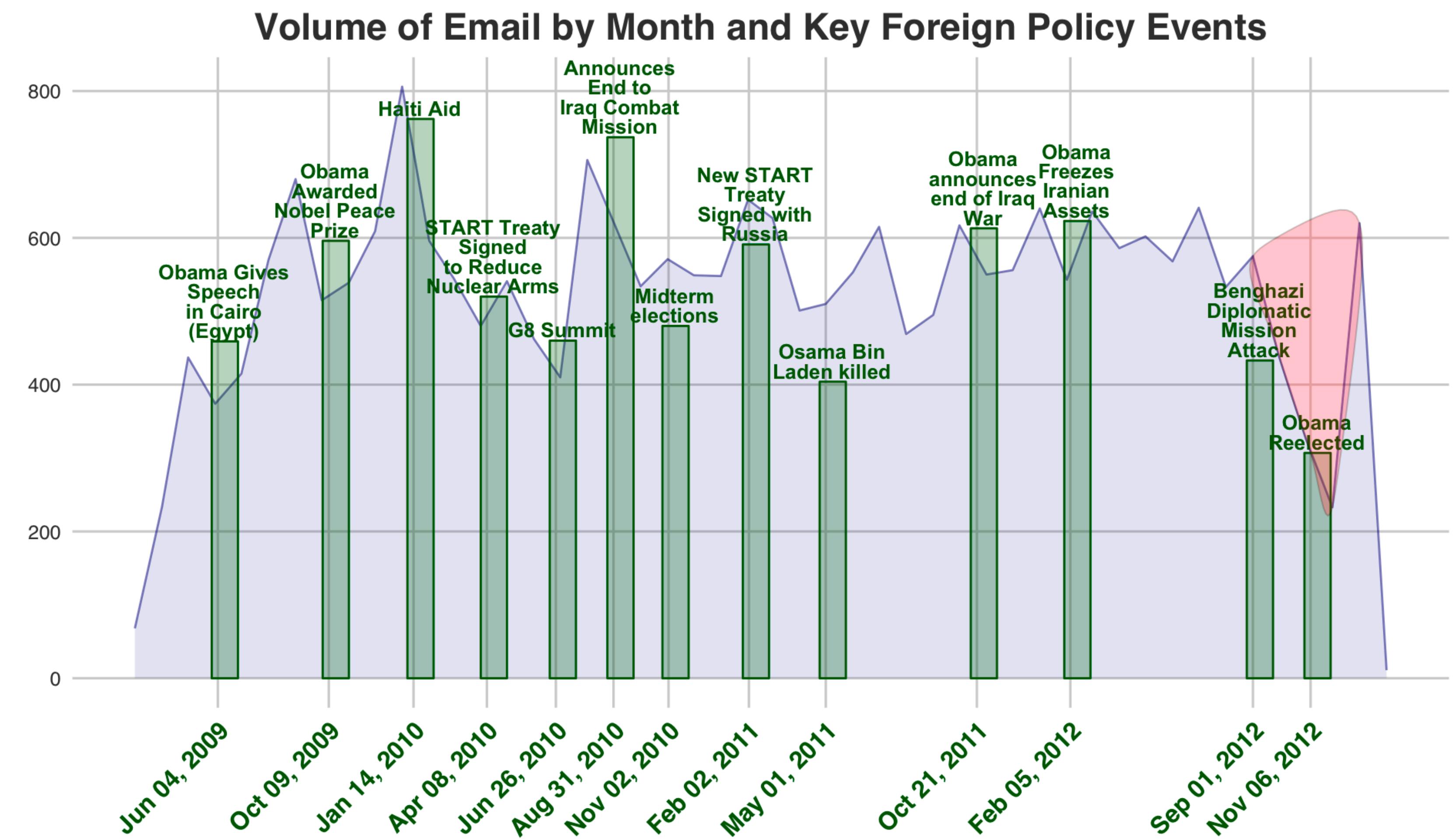
By correlating with relevant info, a completely different set of data emerges, as well as patterns



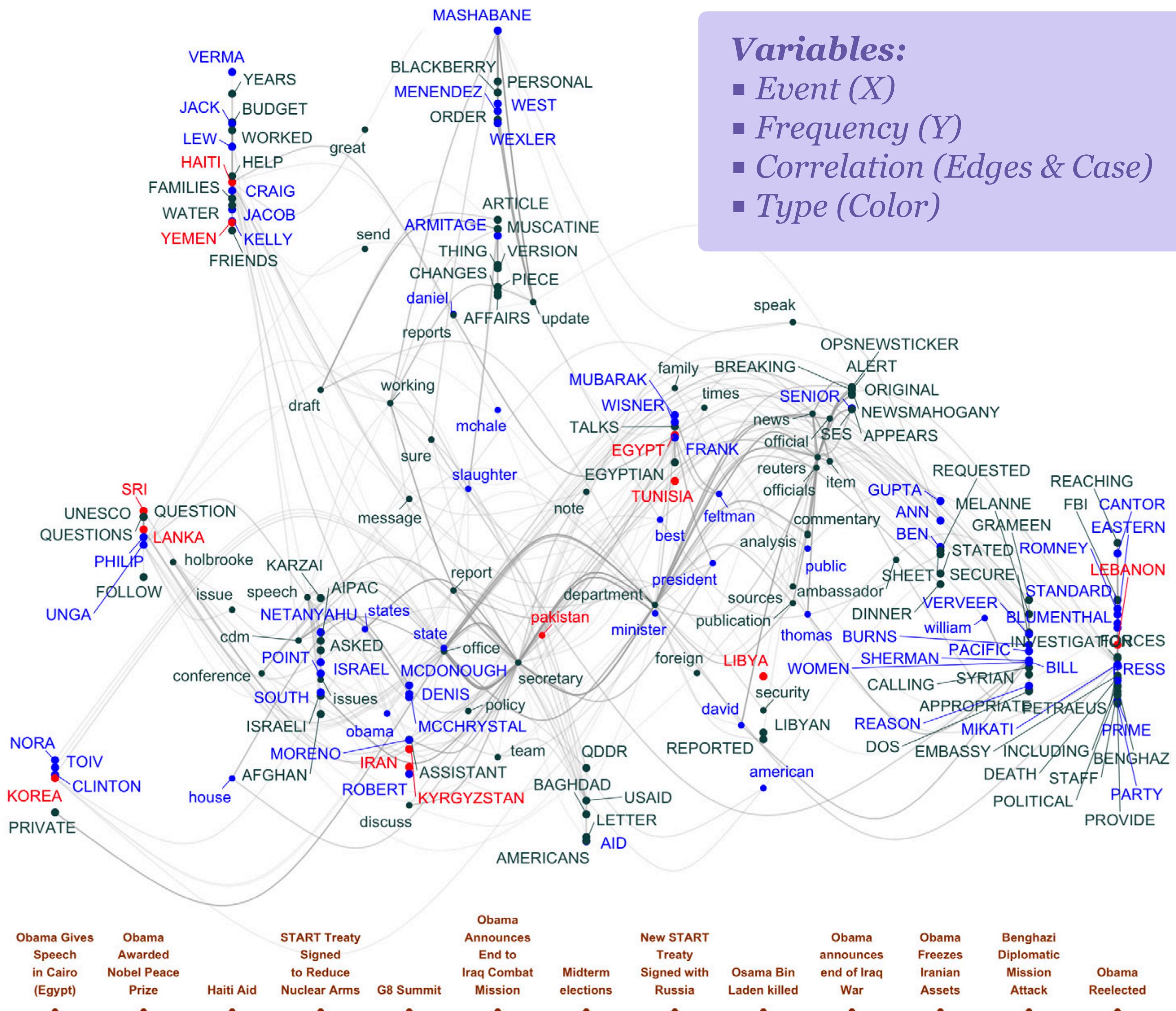
Compared to the uncorrelated graph, the terms are different and reveal more about underlying content

Analyzing by Time

Collect key event data and tag content as to its event relationship, or correlation in time



Relevant Keywords Per Major Foreign Policy Event



N-Grams by Key Event

*Static graphs
won't always
lend themselves
to high density;
interactivity
would greatly
enhance this
presentation*

mikati
reaching
cantor
blumenthal
times
best
fbi

romney

standard

eastern

ambassador

lebanon

minister

forces

investigation

president

security

foreign

family

embassy

petraeus

death

press

benghazi

including

political

staff

american

obama

prime

party

provide

•

11/06/2012

Obama

Reelected

•

11/06/2012

Obama

Reelected

•

melanne
verveer
grameen
state
best
sherman
secretary
pacific
secure
women
thomas
burns
calling
president
department
william
bill
syrian
reason
appropriate
dos
david
sure

office
secretary
ann
department
gupta
ben
update
minister
requested
states
state
times
sheet
working
stated
william
note
dinner

reuters
department
news
feltman
thomas
libya
official
officials
item
pakistani
secretary
sources
commentary
analysis
alert
breaking
states
original
newsmahogany
senior
appears
ses
public
publication

reuters
department
news
feltman
thomas
libya
official
officials
item
pakistani
secretary
sources
commentary
analysis
alert
breaking
states
original
newsmahogany
senior
appears
ses
public
publication

egypt
feltman
wisner
mubarak
talks
frank
secretary
department
reuters
news
president
message
office
egyptian
report
reports
thomas
tunisia
minister
conference
working
daniel
ambassador

mashabane
update
times
working
menendez
wexler
sure
speak
great
personal
blackberry
west
slaughter
muscatine
order
article
draft
armitage
thing
version
send
piece
changes
affairs

qddr
usaid
slaughter
team
states
secretary
report
discuss
office
baghdad
pakistan
president
times
americans
american
letter
aid
note
working

aipac
speech
report
netanyahu
karzai
working
point
israel
slaughter
states
state
president
asked
update
secretary
policy
south
issues
issue
team
israeli
obama
pakistan
best
security
holbrooke
iran
david
robert

lanka
sri
holbrooke
secretary
questions
unesco
unga
state
office
report
speech
issues
issue
philip
update
sure
message
minister
department
working
foreign
follow
best
states
friends
president
afghan

verma
haiti
worked
working
lew
budget
jack
years
best
help
daniel
speech
craig
reports
report
families
family
mchale
water
great
jacob
cdm
kelly
note
yemen
send
state
states
friends
president
afghan

office
secretary
department
state
korea
states
obama
working
draft
nora
toiv
cdm
clinton
discuss
sure
conference
house
private

•
06/04/2009
Obama Gives
Speech
in Cairo
(Egypt)
•
10/09/2009
Obama
Awarded
Nobel Peace
Prize
•
01/14/2010
Haiti Aid
•
04/08/2010
START Treaty
Signed
to Reduce
Nuclear Arms
•
06/26/2010
G8 Summit
•
08/31/2010
Obama
Announces
End to
Iraq Combat
Mission
•
11/02/2010
Midterm
elections
•
02/02/2011
New START
Treaty
Signed with
Russia
•
05/01/2011
Osama Bin
Laden killed
•
10/21/2011
Obama
announces
end of Iraq
War
•
02/05/2012
Obama
Freezes
Iranian
Assets
•
09/01/2012
Benghazi
Diplomatic
Mission
Attack
•
11/06/2012
Obama
Reelected
•

Type ● Country ● Name ● Other

Simple Version: High Relevancy Terms by Event

Mike Safar
KMS Product Strategies

www.linkedin.com/in/msafar

Learn More:

*For a full article, including
complete R-language source code
and additional examples, see:*

<http://www.corpus.live/signal-boost>