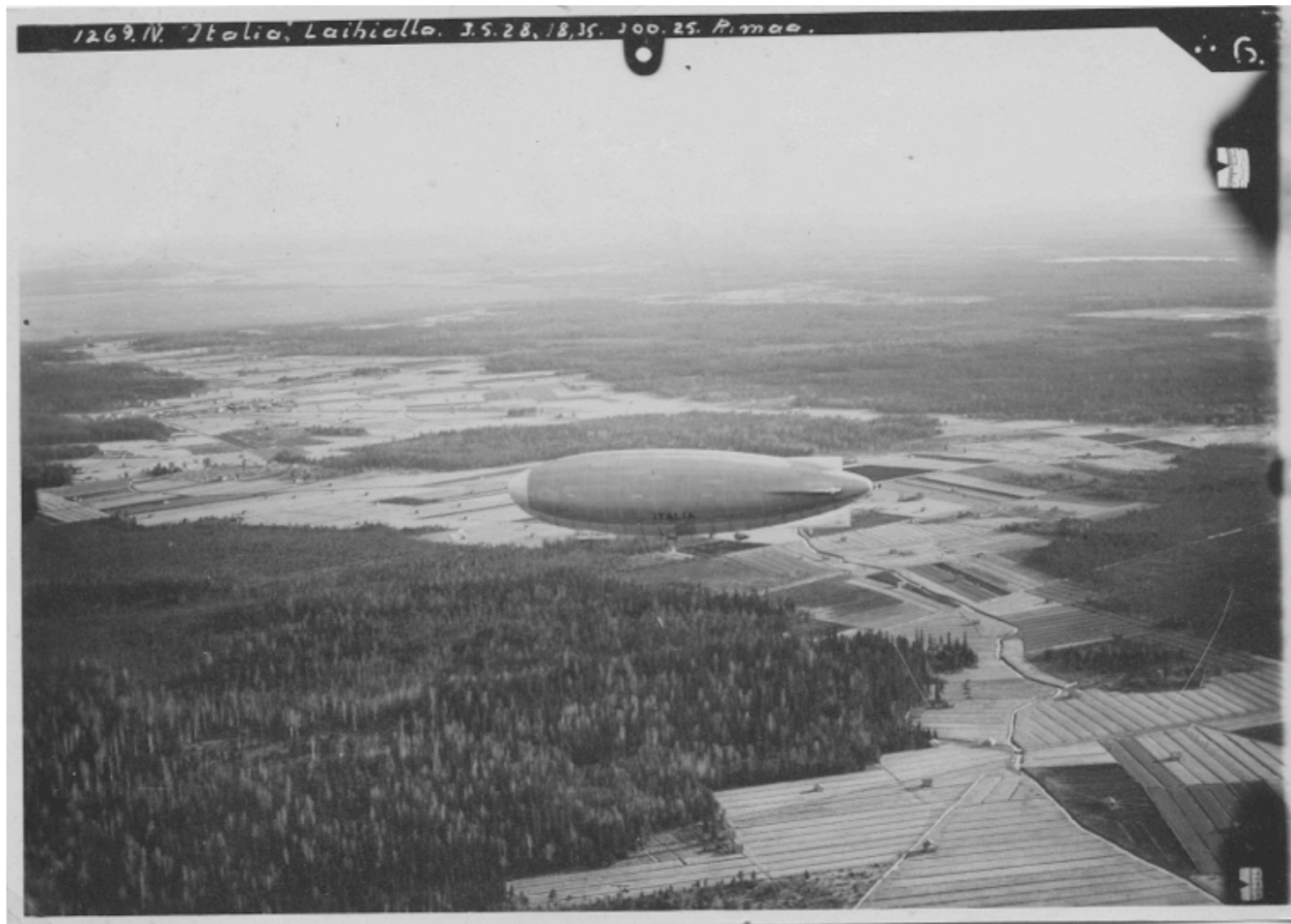


useful large-scale sentiment analysis

jussi karlgren, december 2013





gavagai



gavagai

recent startup - spinoff from sics



gavagai

recent startup - spinoff from sics

about 10 employees

Gavagai





gavagai

recent startup - spinoff from sics

about 10 employees




extracts actionable intelligence from very large text streams

Gavagai



we like the idea of huge human-generated data streams

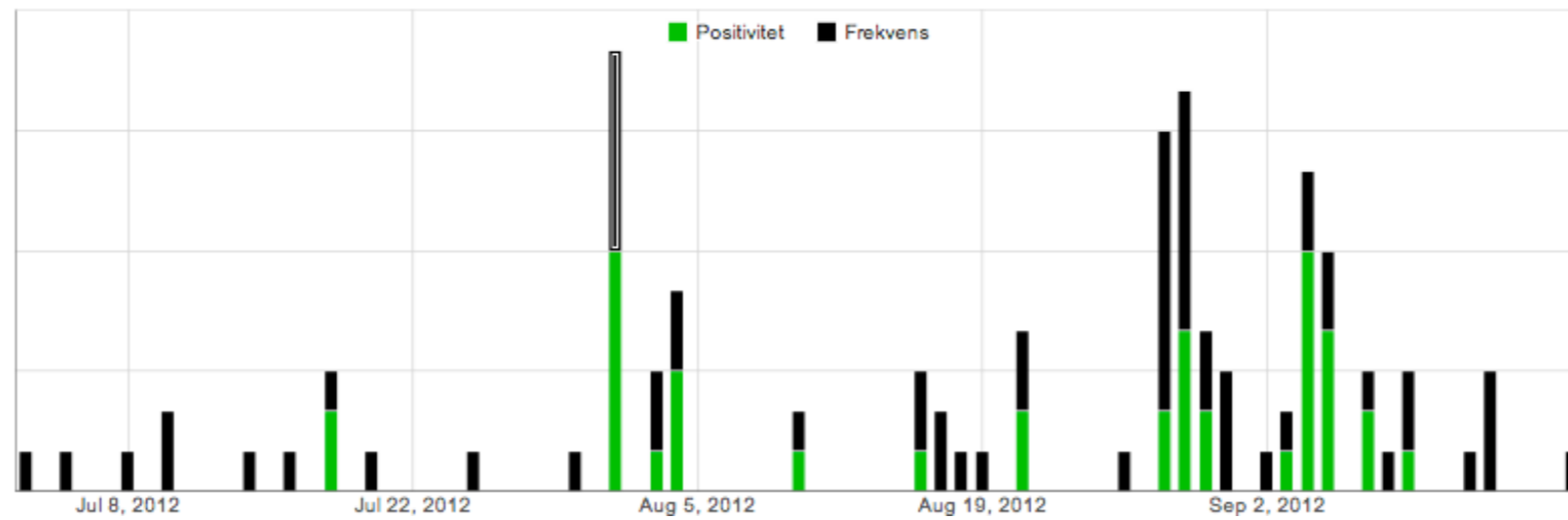
market watch: what is going on?

<i>My Premium Brand</i>		<i>Customer satisfaction +32%</i>
<i>My Other Brand</i>		<i>Quality perception -72%</i>
<i>Another Brand</i>		<i>Irritation +7%</i>

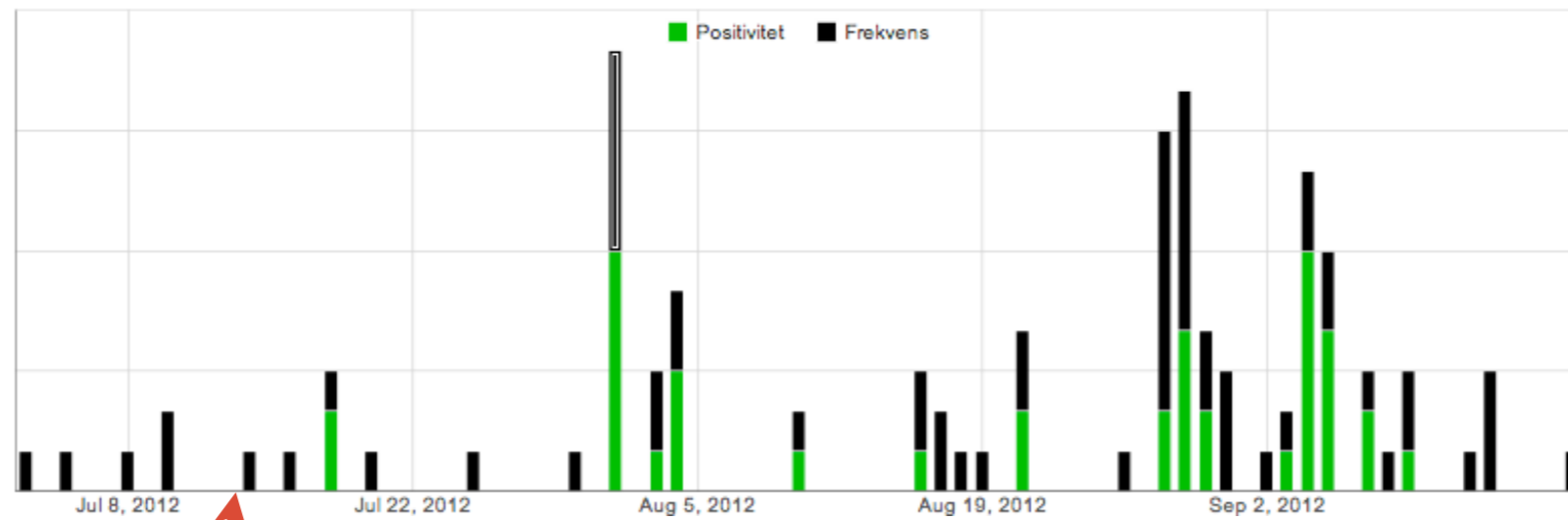
market watch: what are we associated with?



evaluating marketing

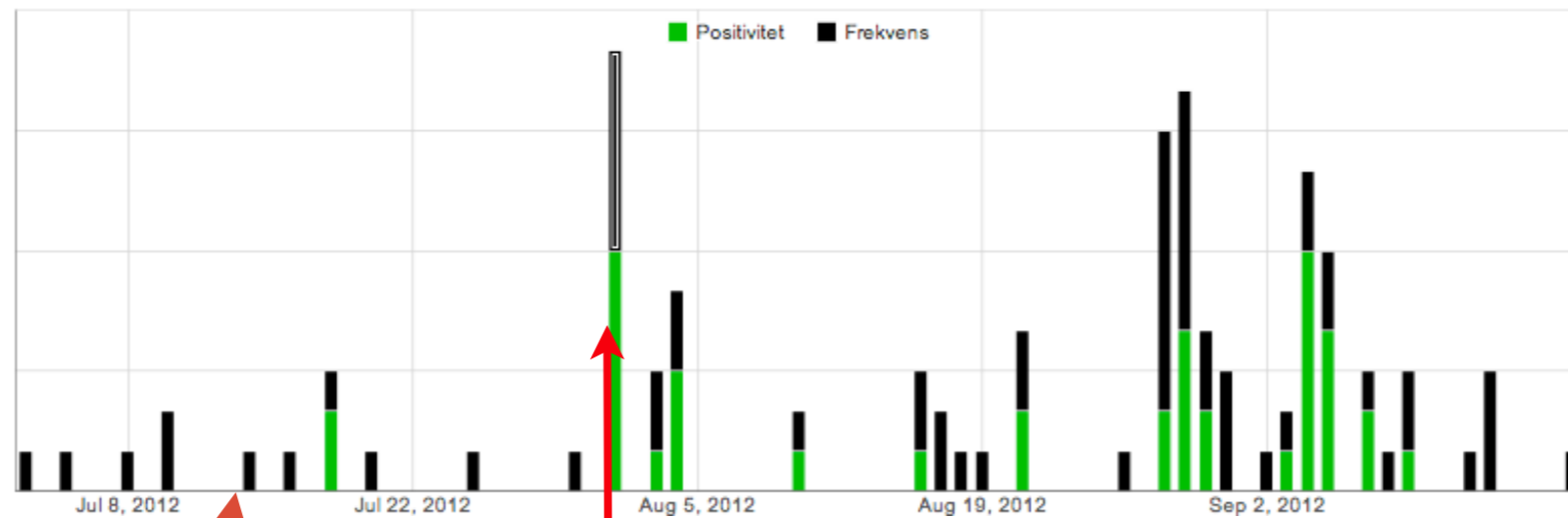


evaluating marketing



traditional marketing campaign

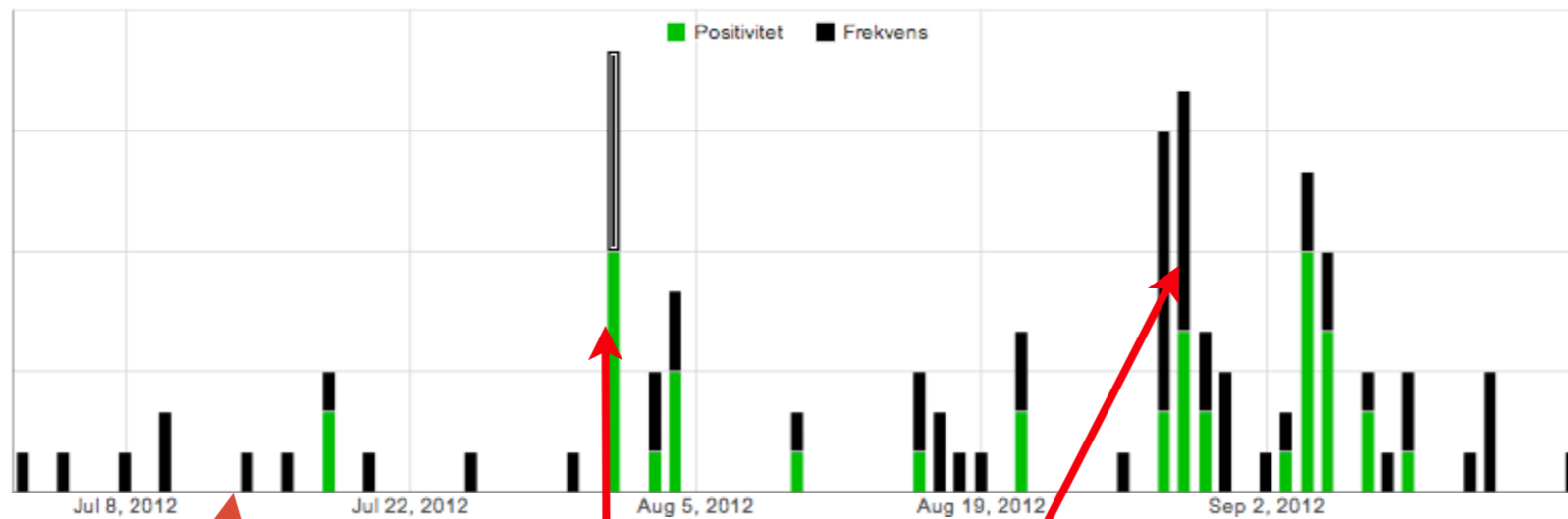
evaluating marketing



traditional marketing campaign

giveaways to cosmetics subscribers

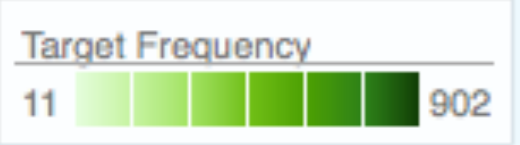
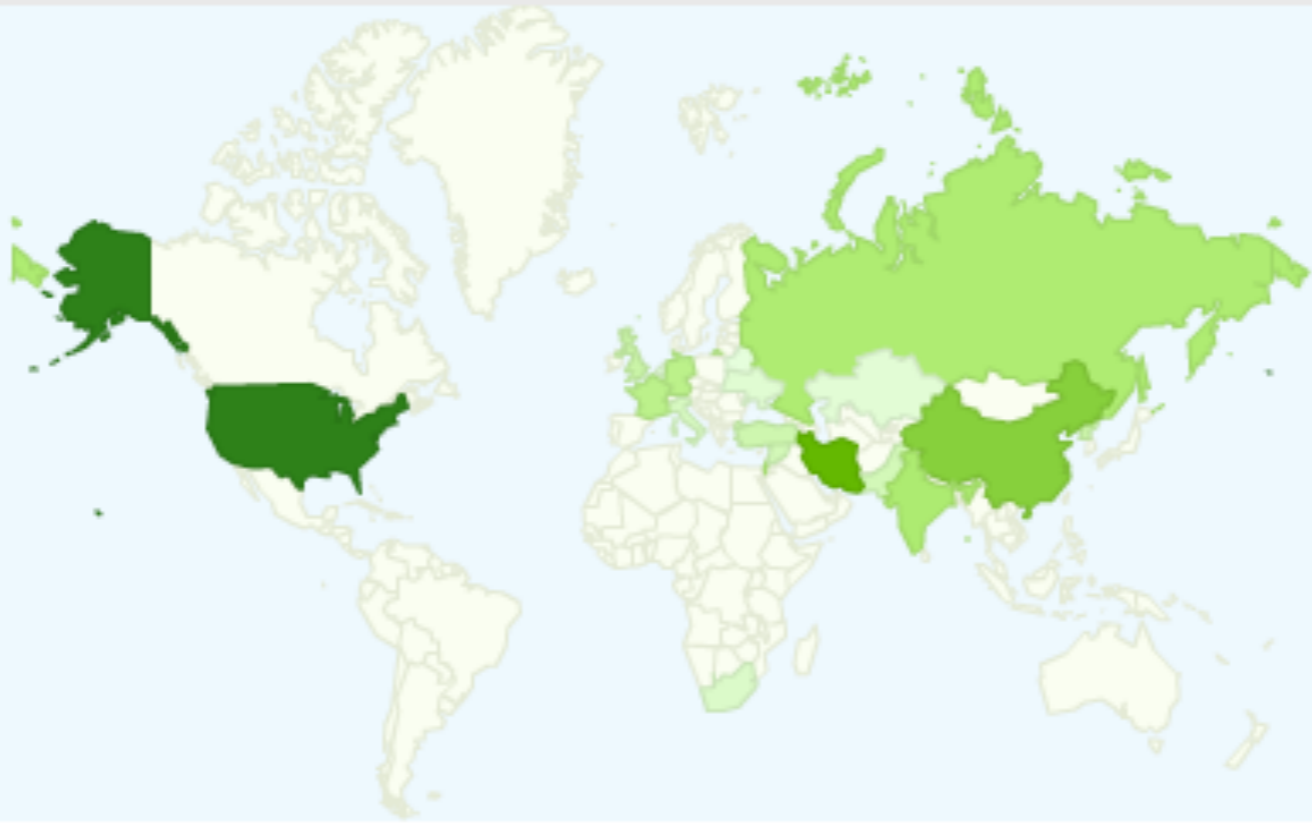
evaluating marketing

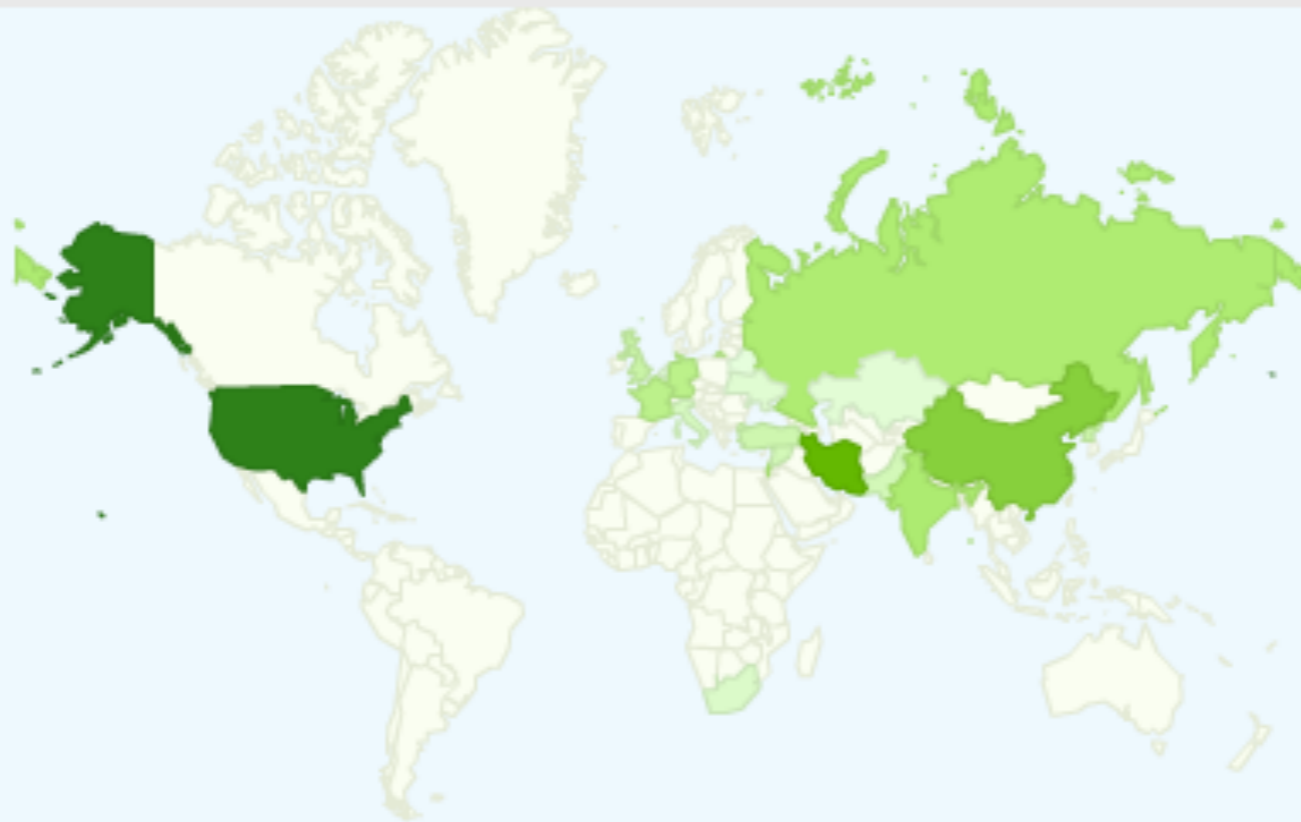


traditional marketing campaign

giveaways to cosmetics subscribers

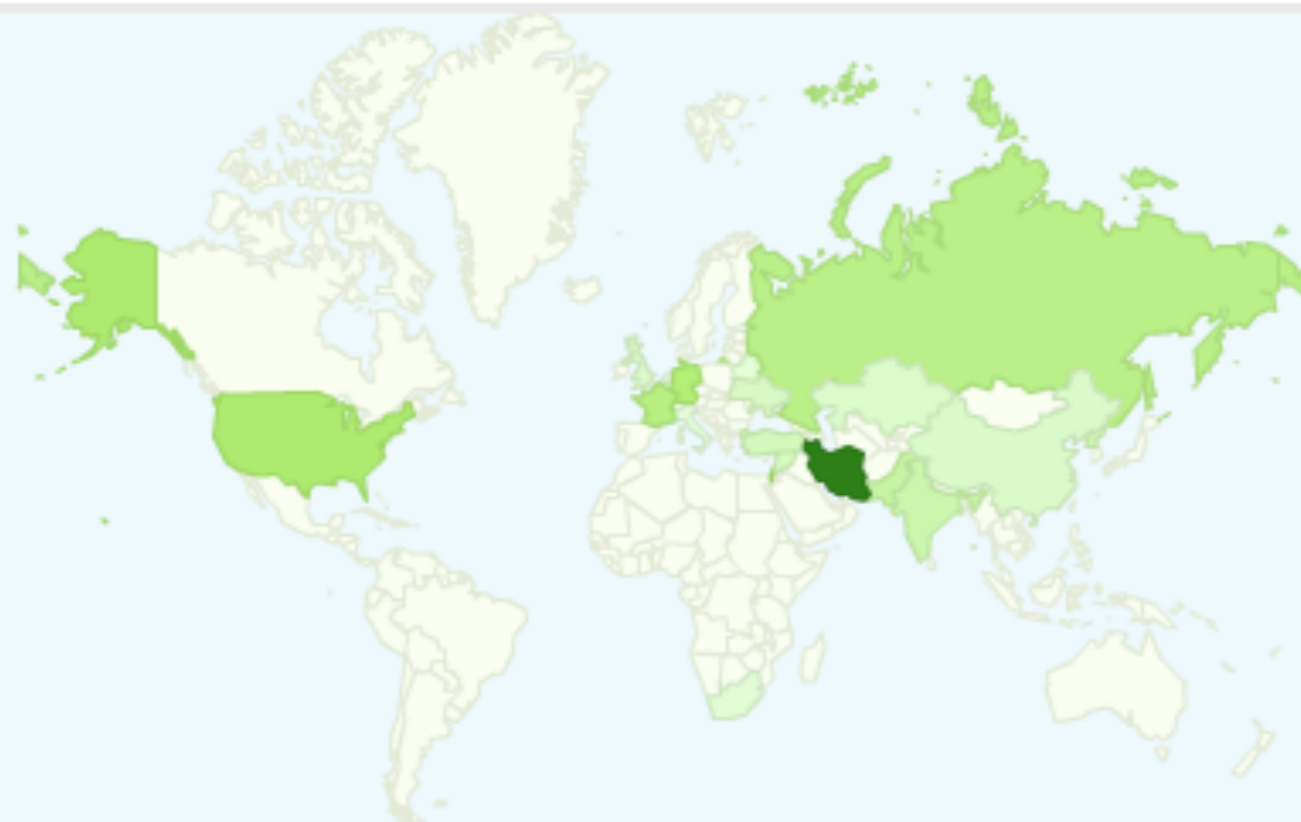
giveaways to bloggers





Target Frequency

11 902

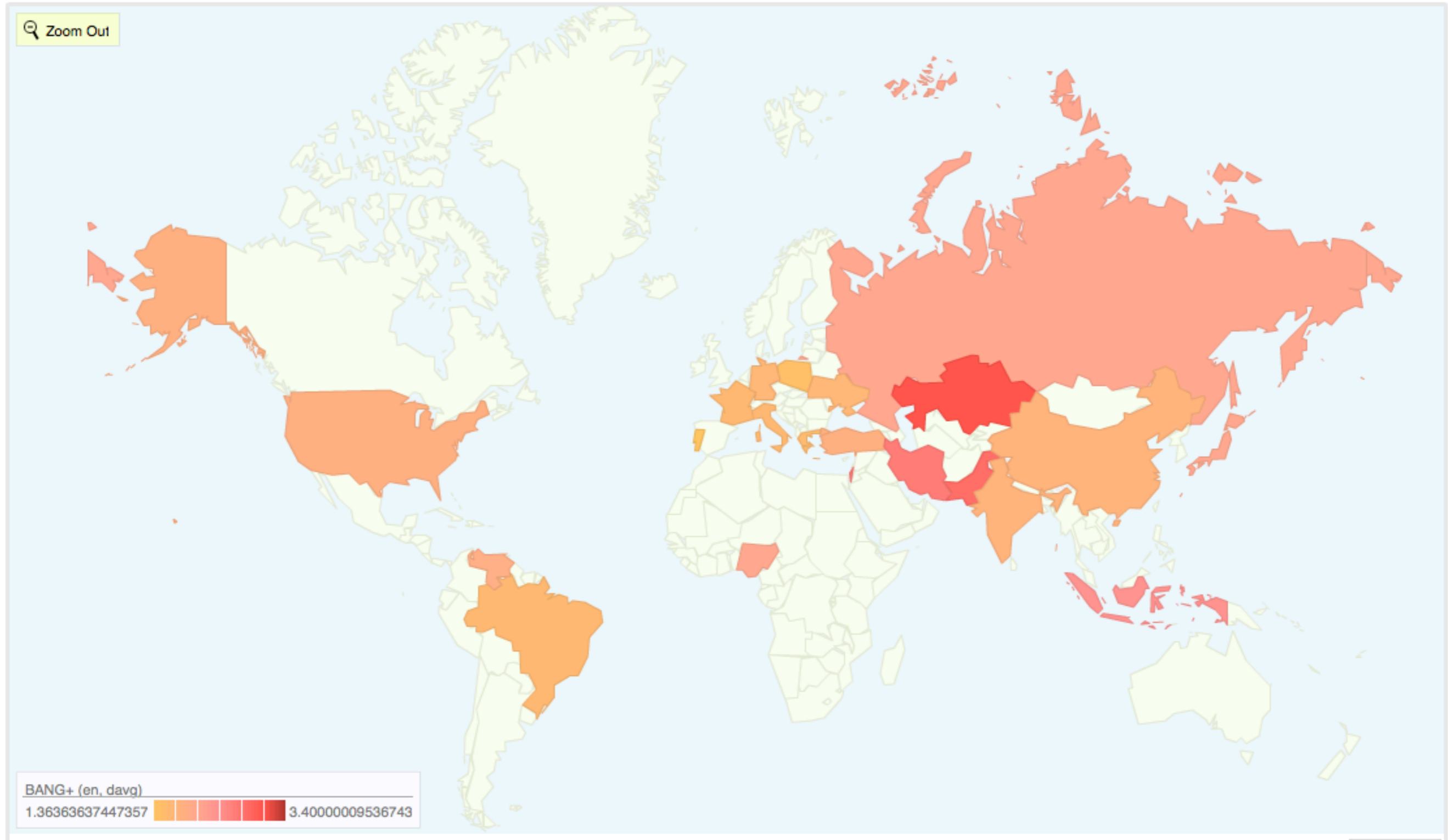


Maximal Violence

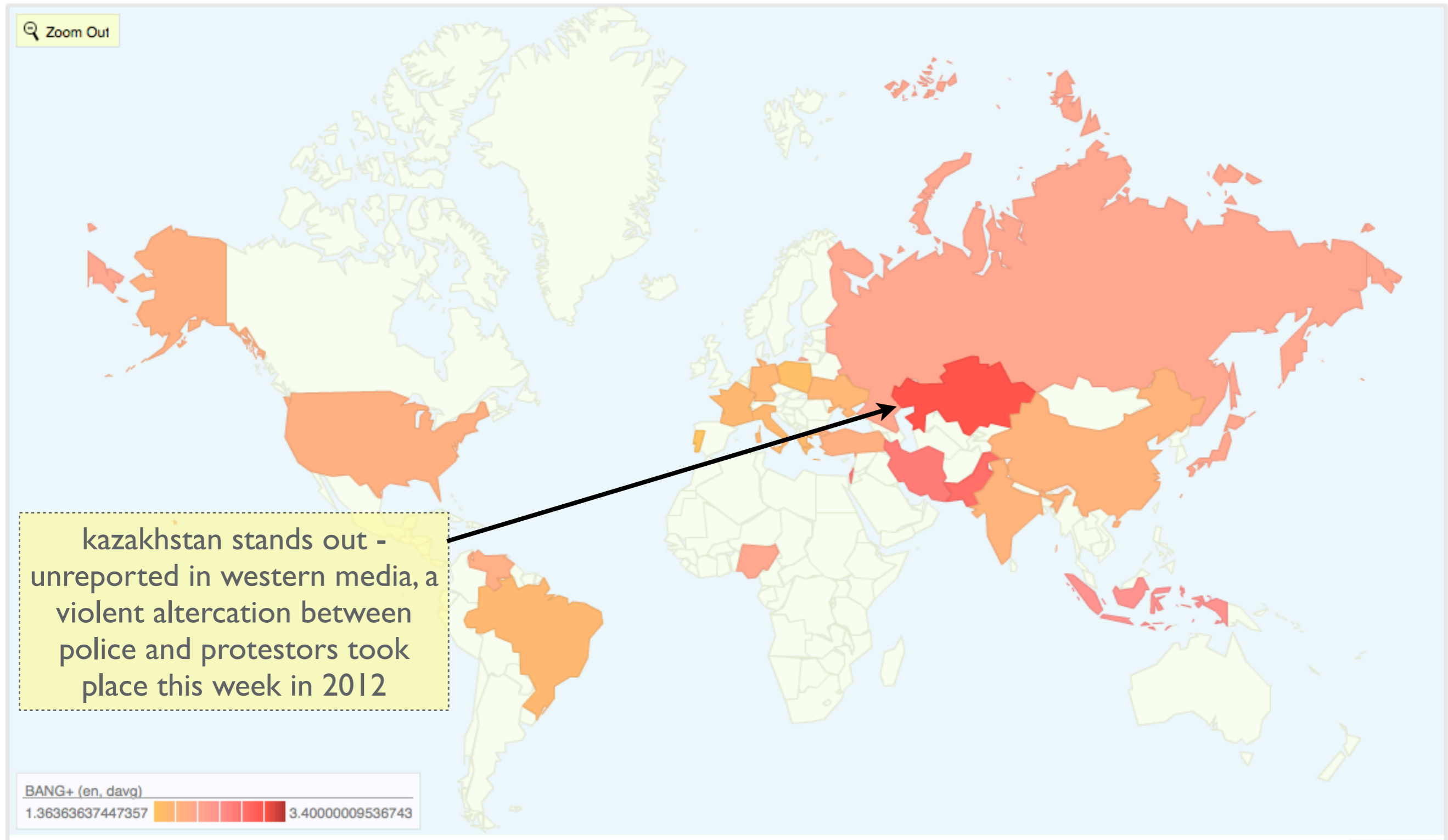
2 176



tracking violence in the world



tracking violence in the world



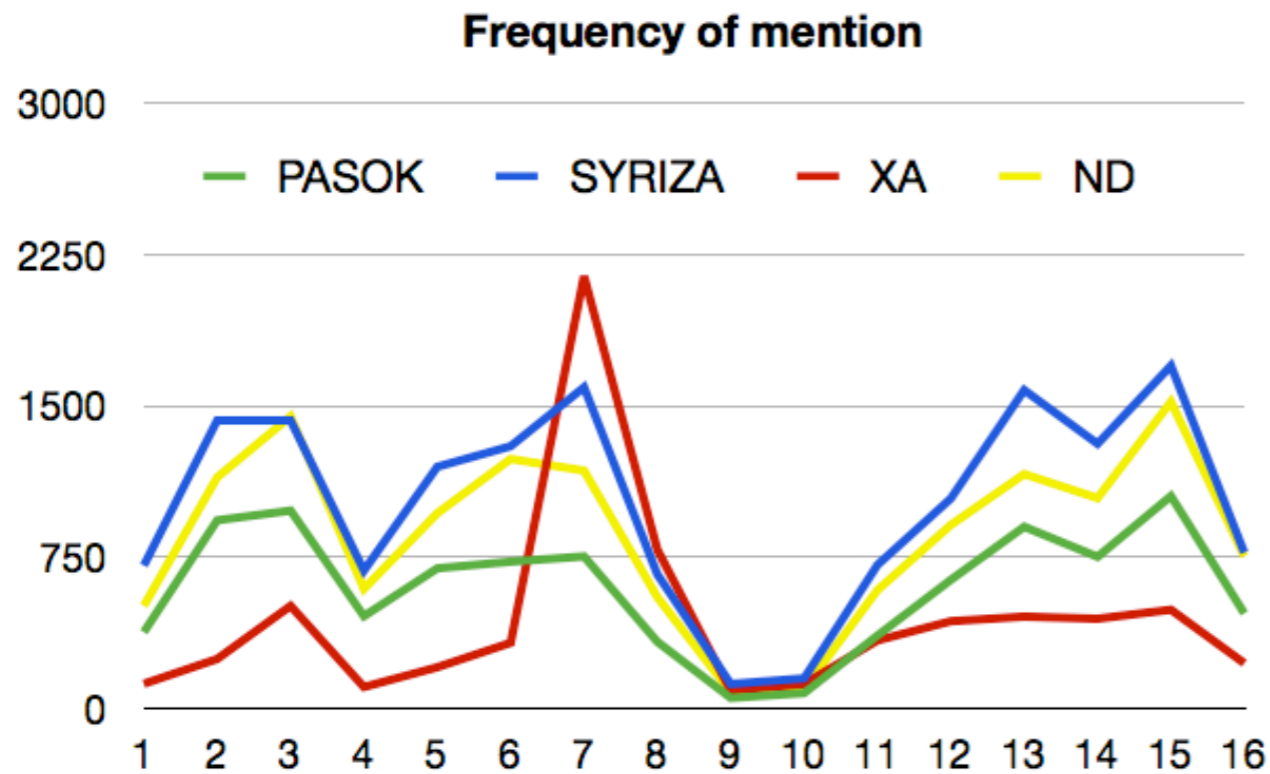
forecasting is not only counting mentions!

the day before the Greek election in June 2012



forecasting is not only counting mentions!

the day before the Greek election in June 2012

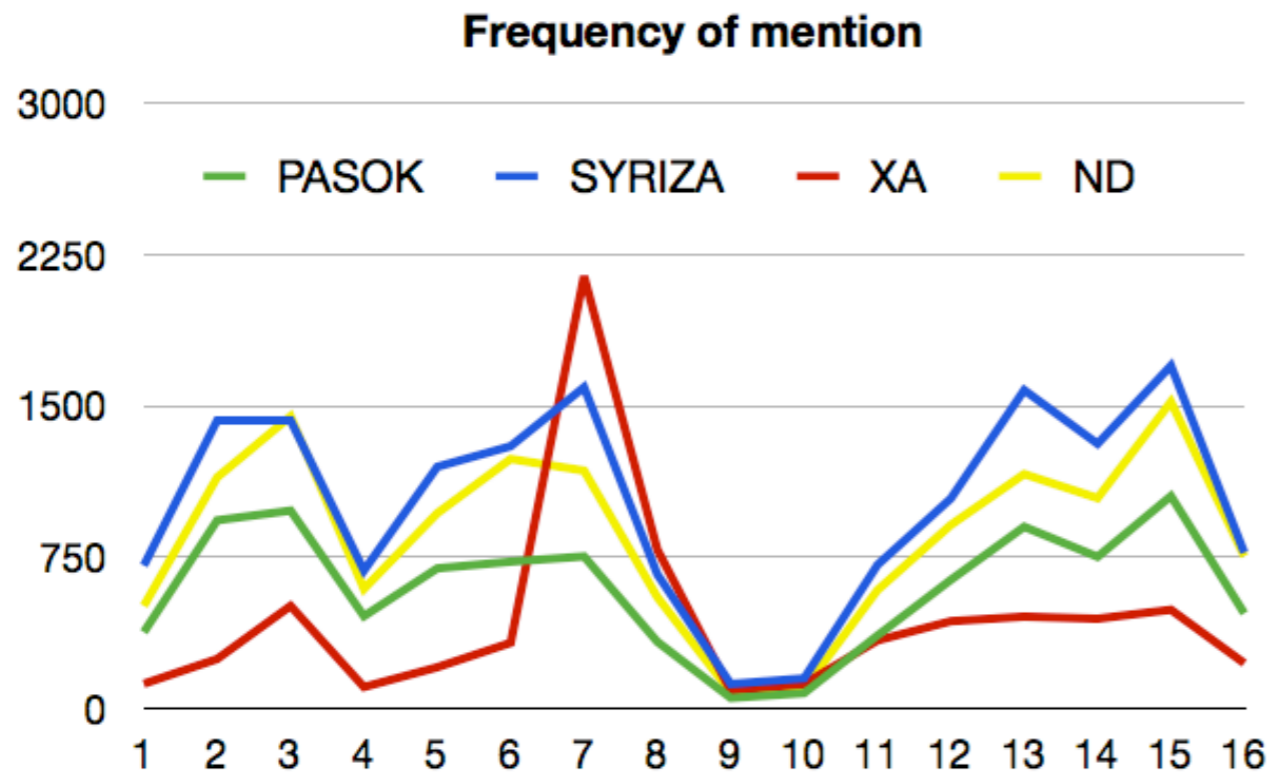


Syriza gets the most attention

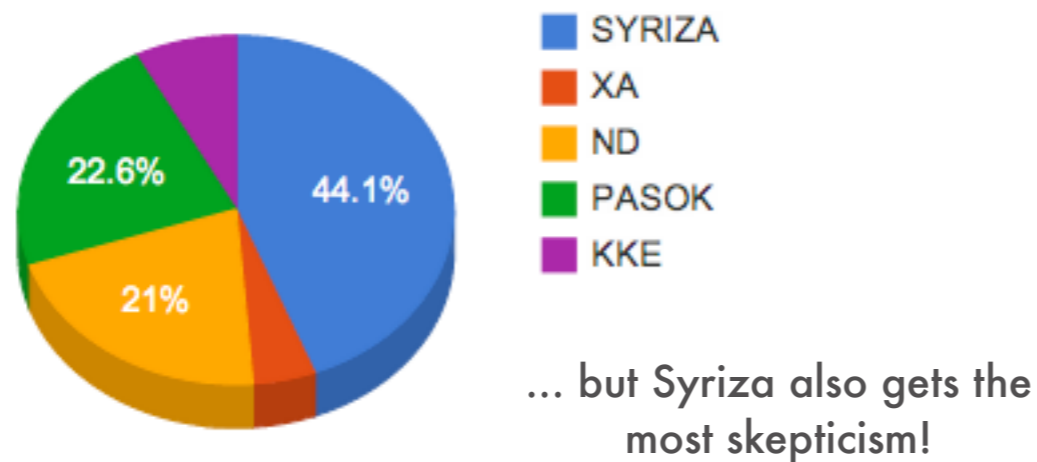


forecasting is not only counting mentions!

the day before the Greek election in June 2012



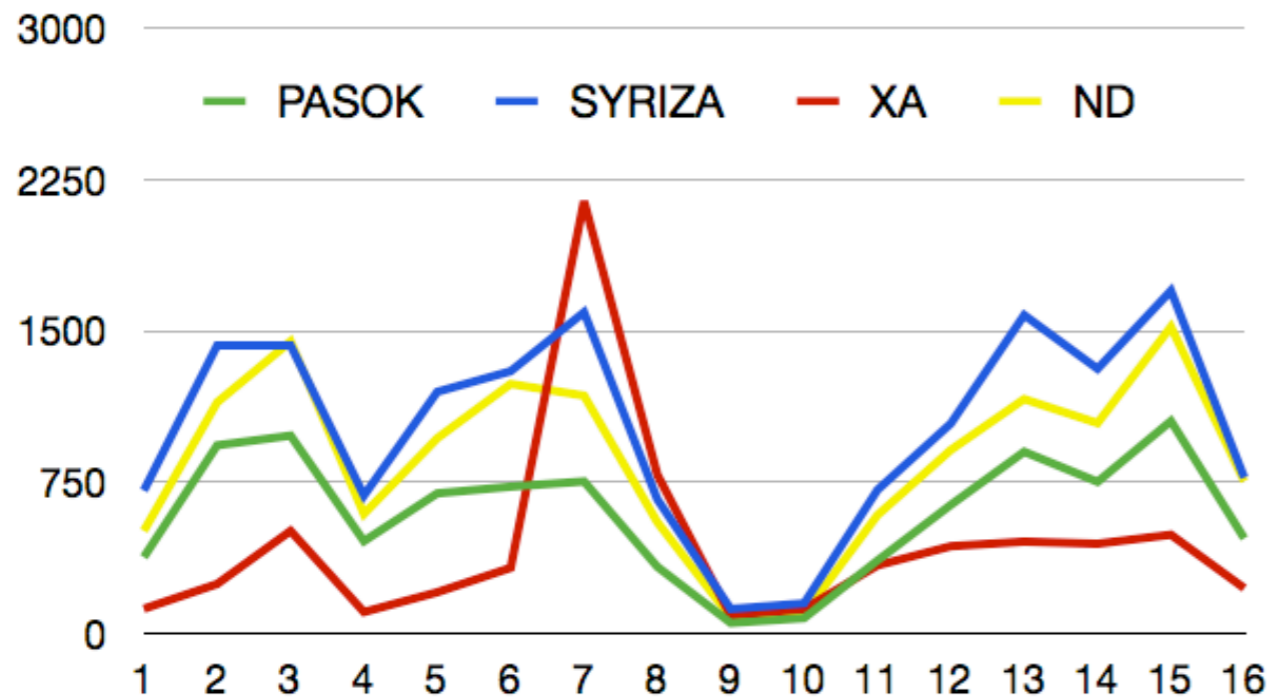
Syriza gets the most attention



forecasting is not only counting mentions!

the day before the Greek election in June 2012

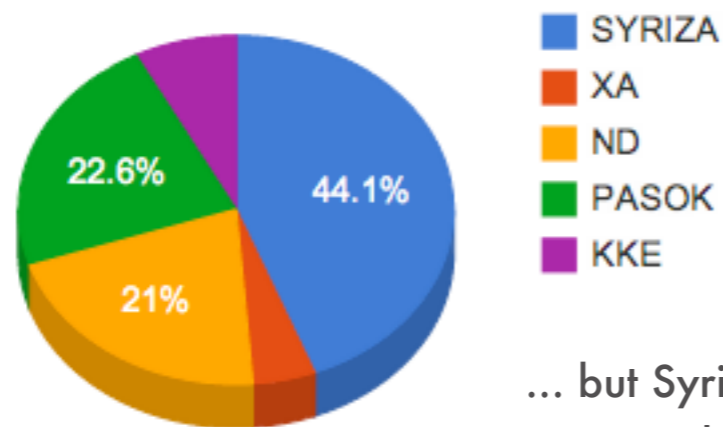
Frequency of mention



Syriza gets the most attention

Results

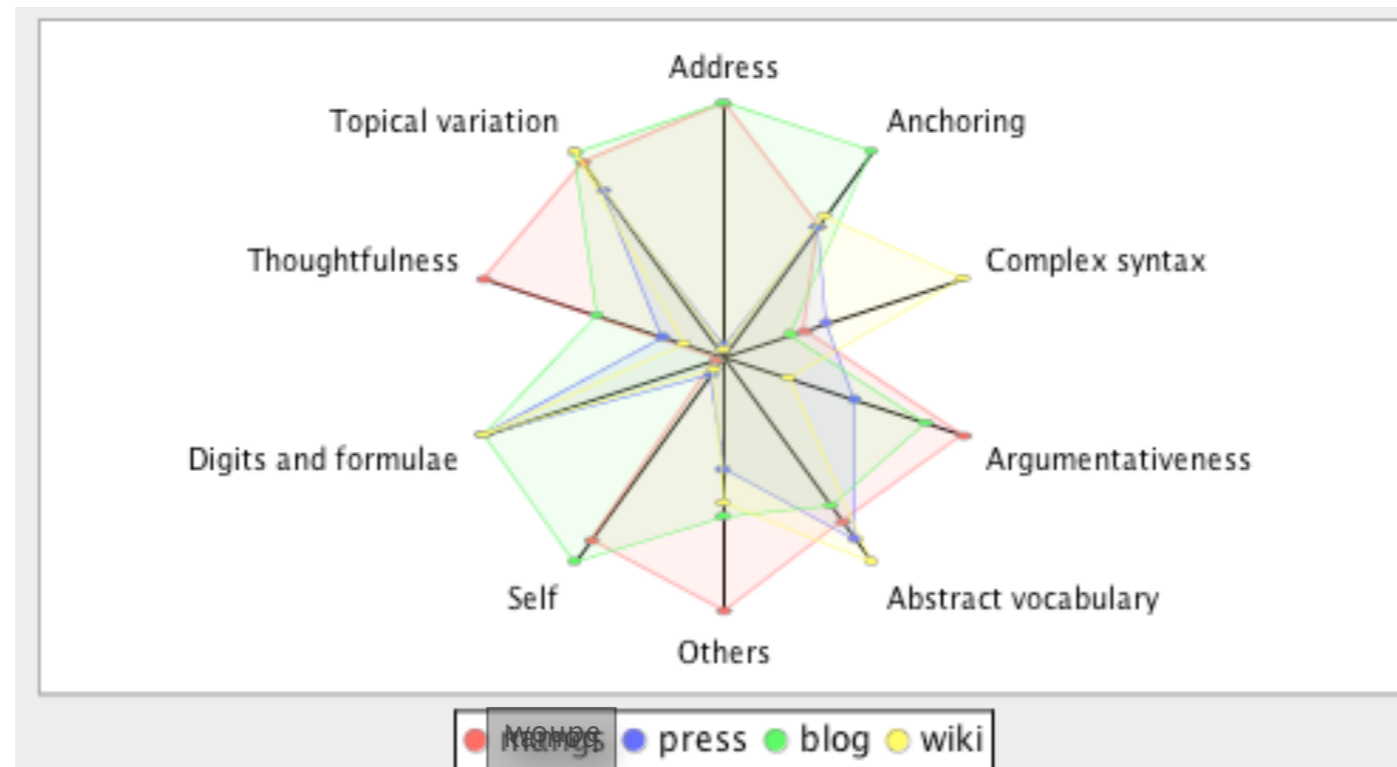
ND: 30%
Syriza: 27%
Pasok: 12%
XA: 6.9%
KKE: 4.9%



... but Syriza also gets the most skepticism!

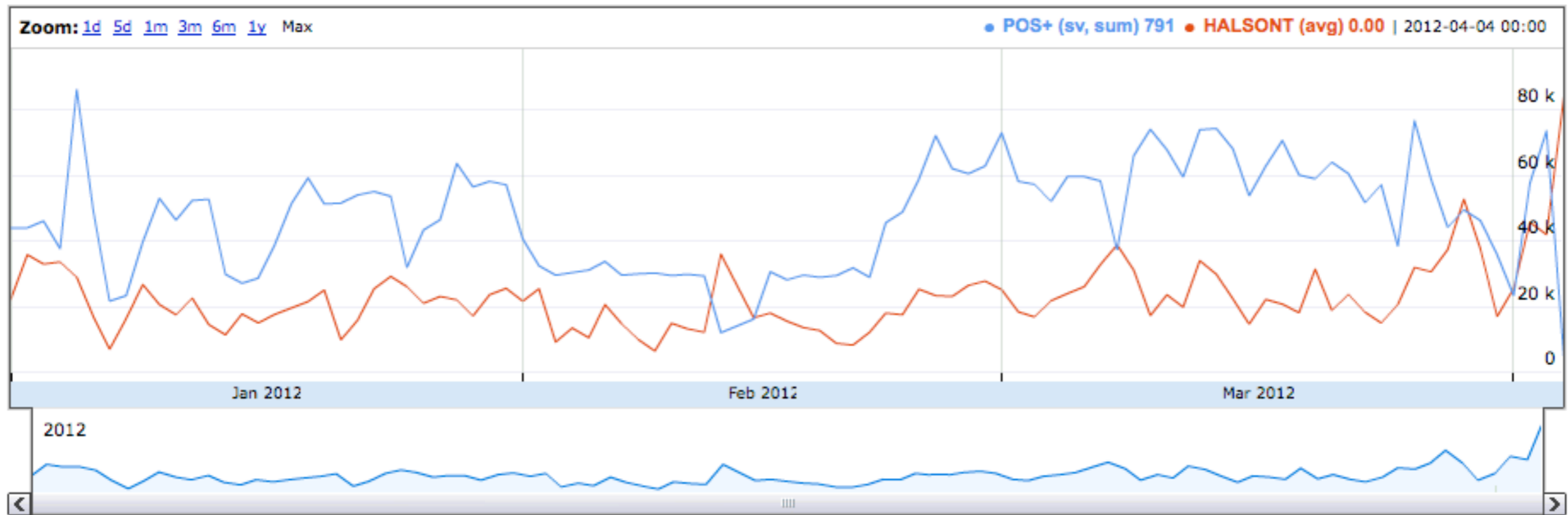


authorship profiling

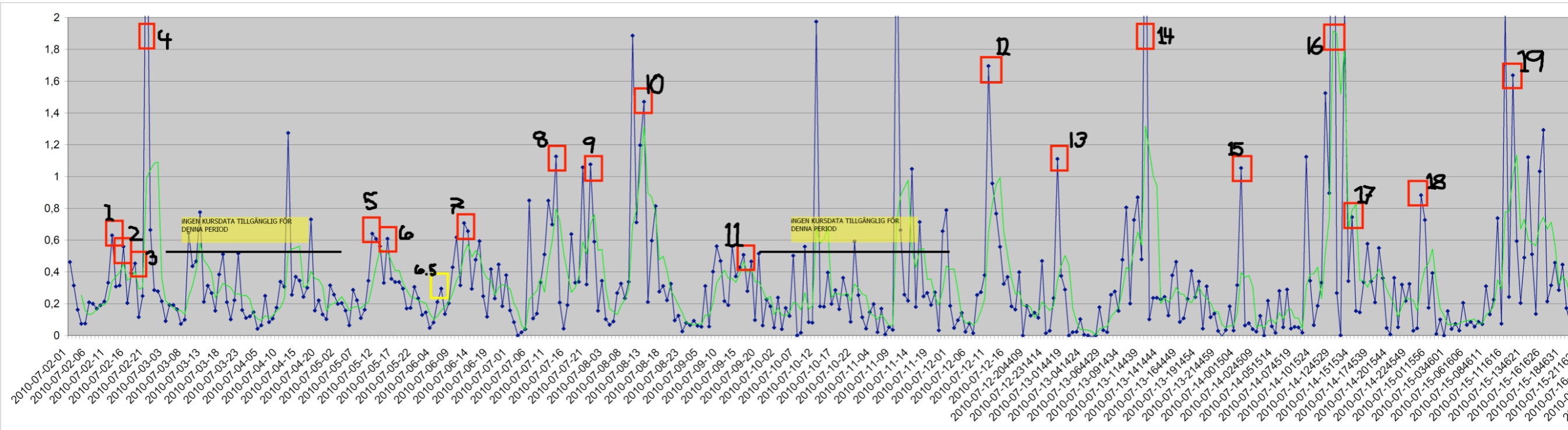


how does the writing of some individual differ from standard text?

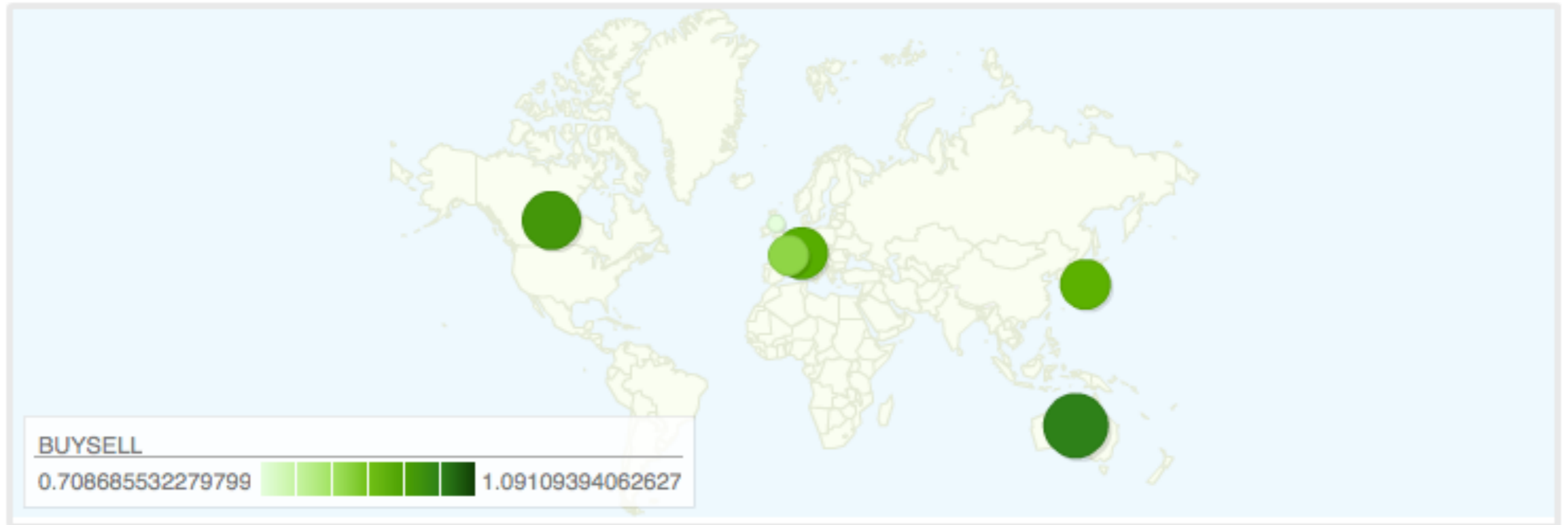
what's the mood out there?



trading on information

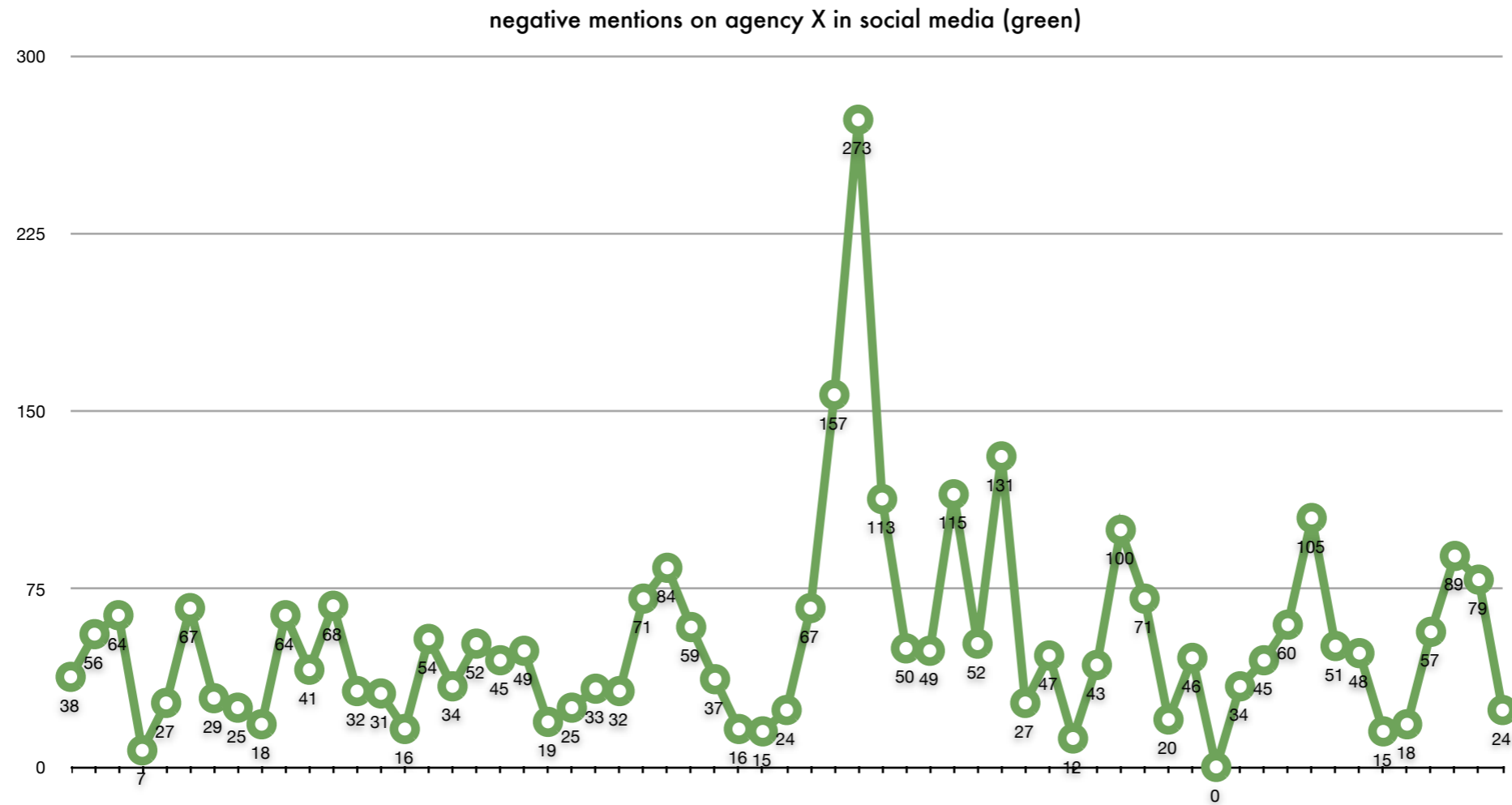


financial analysis: buy or sell?

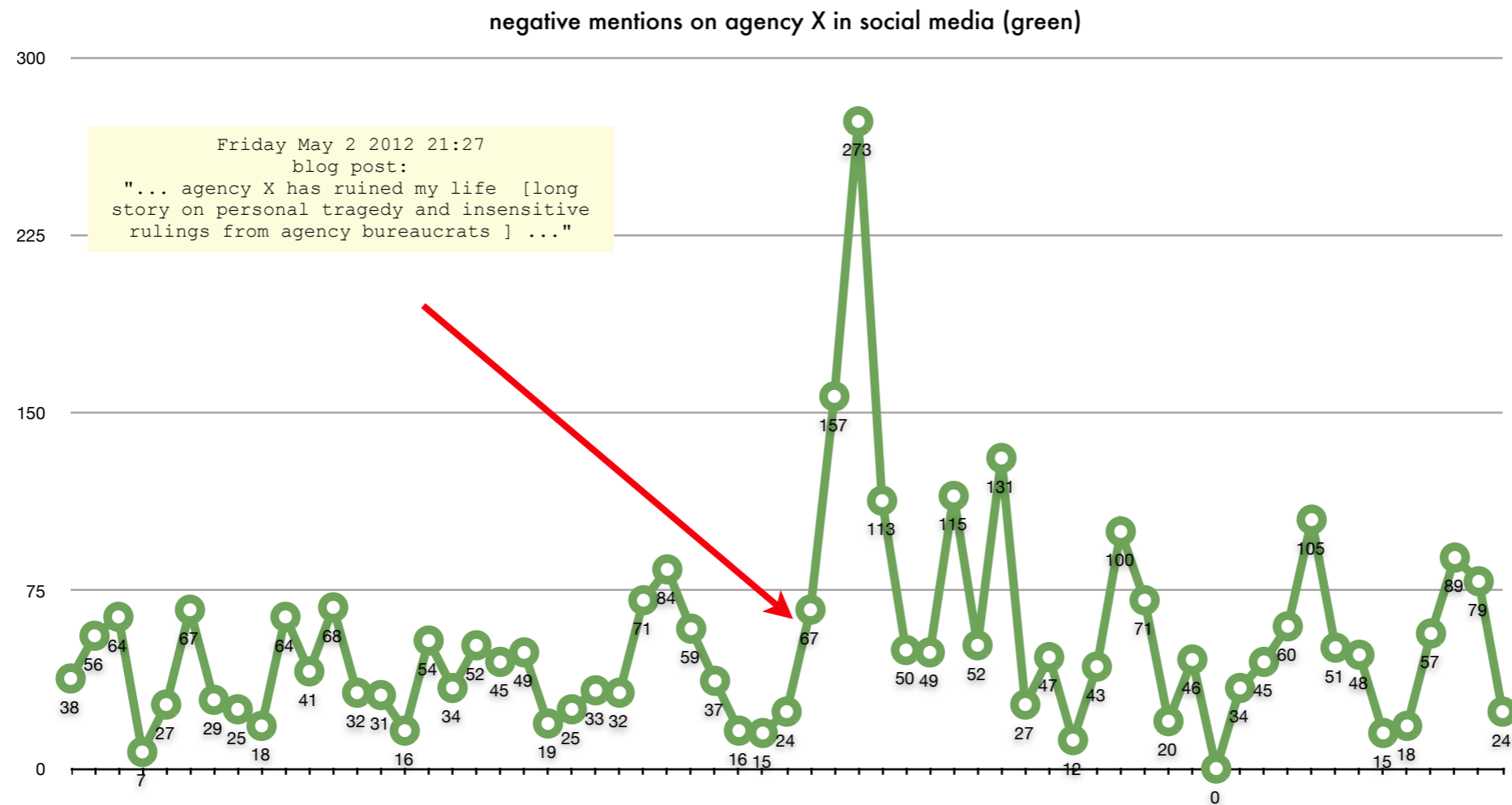


early warning in social media

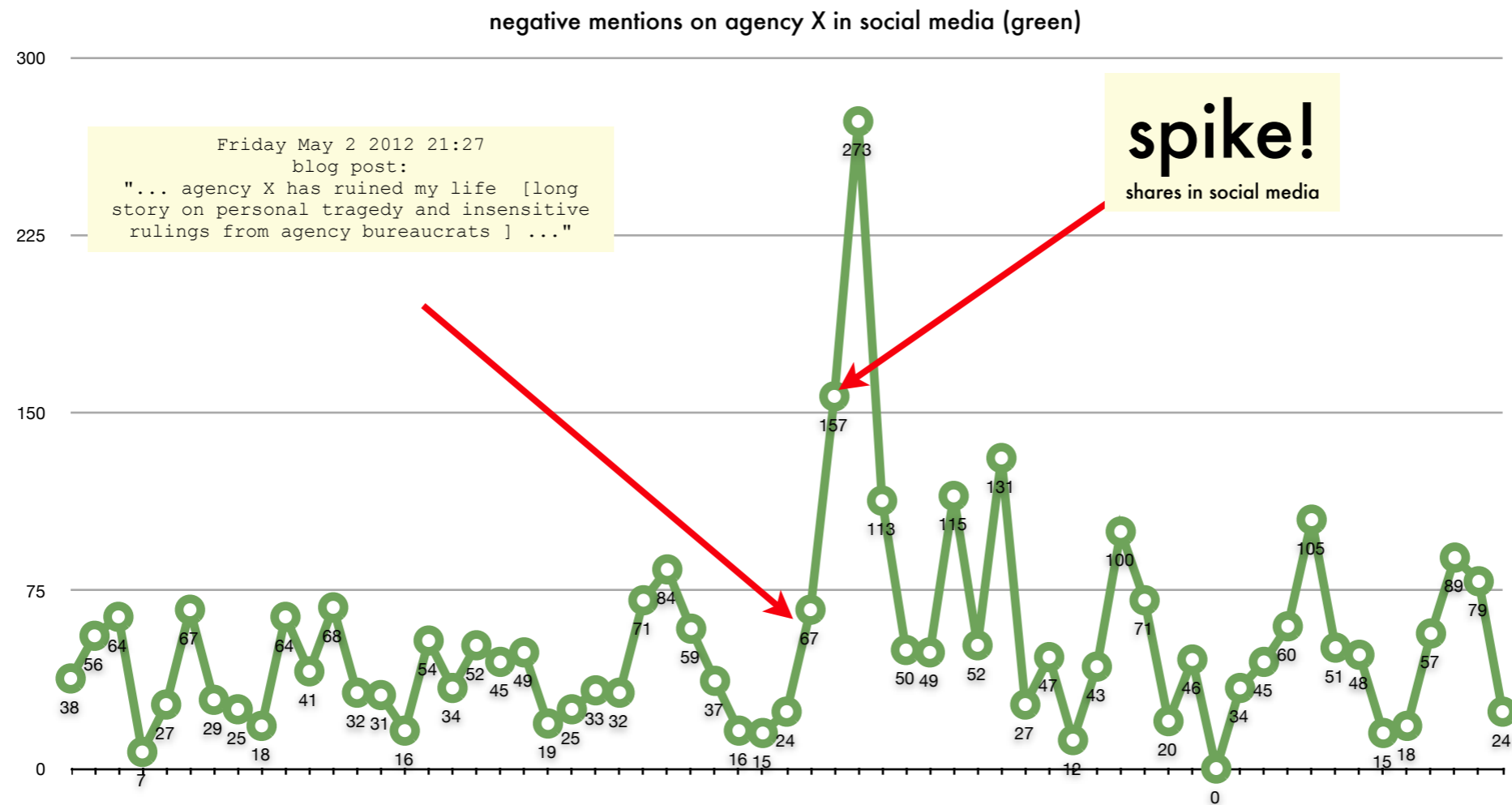
early warning in social media



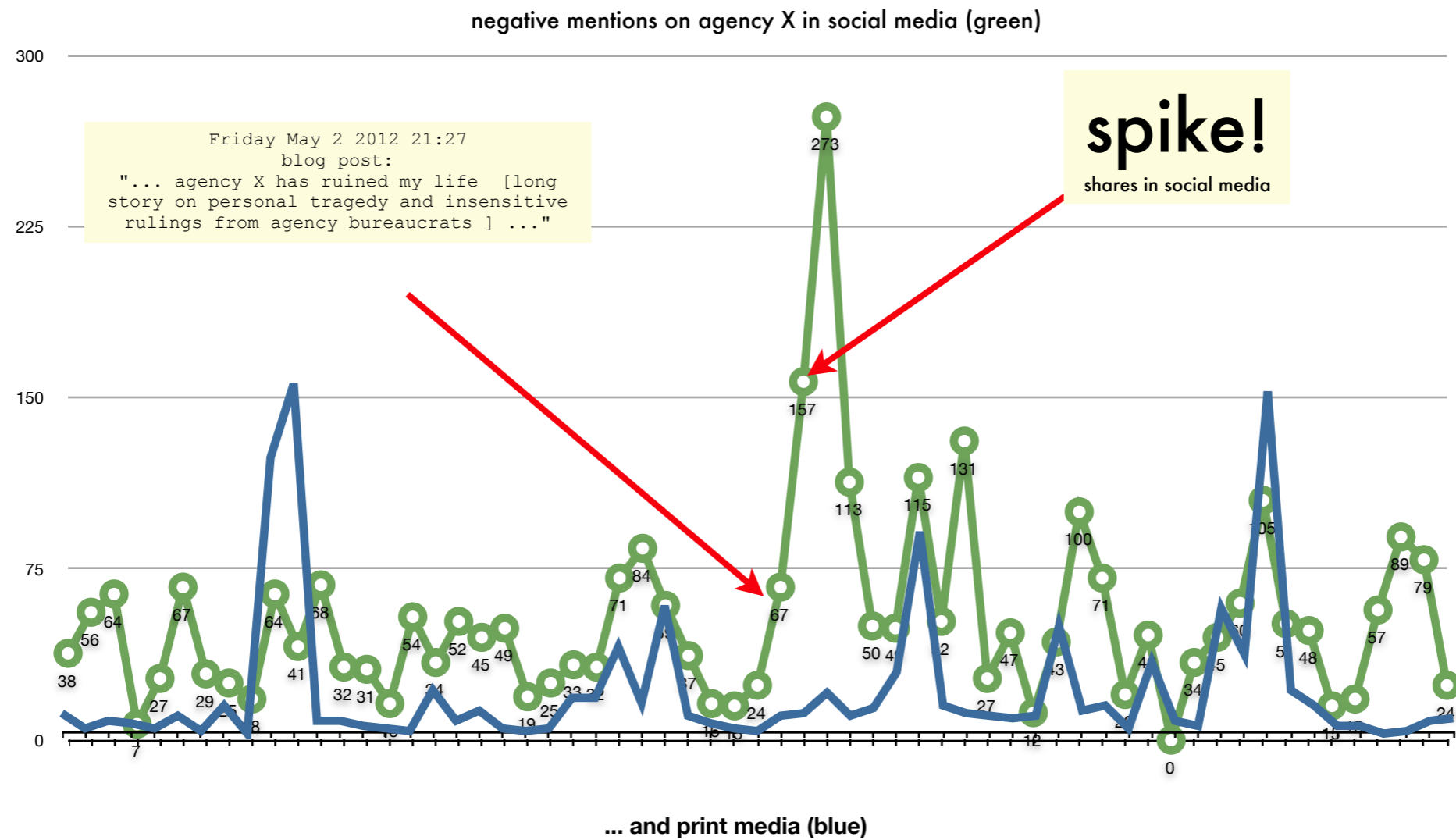
early warning in social media



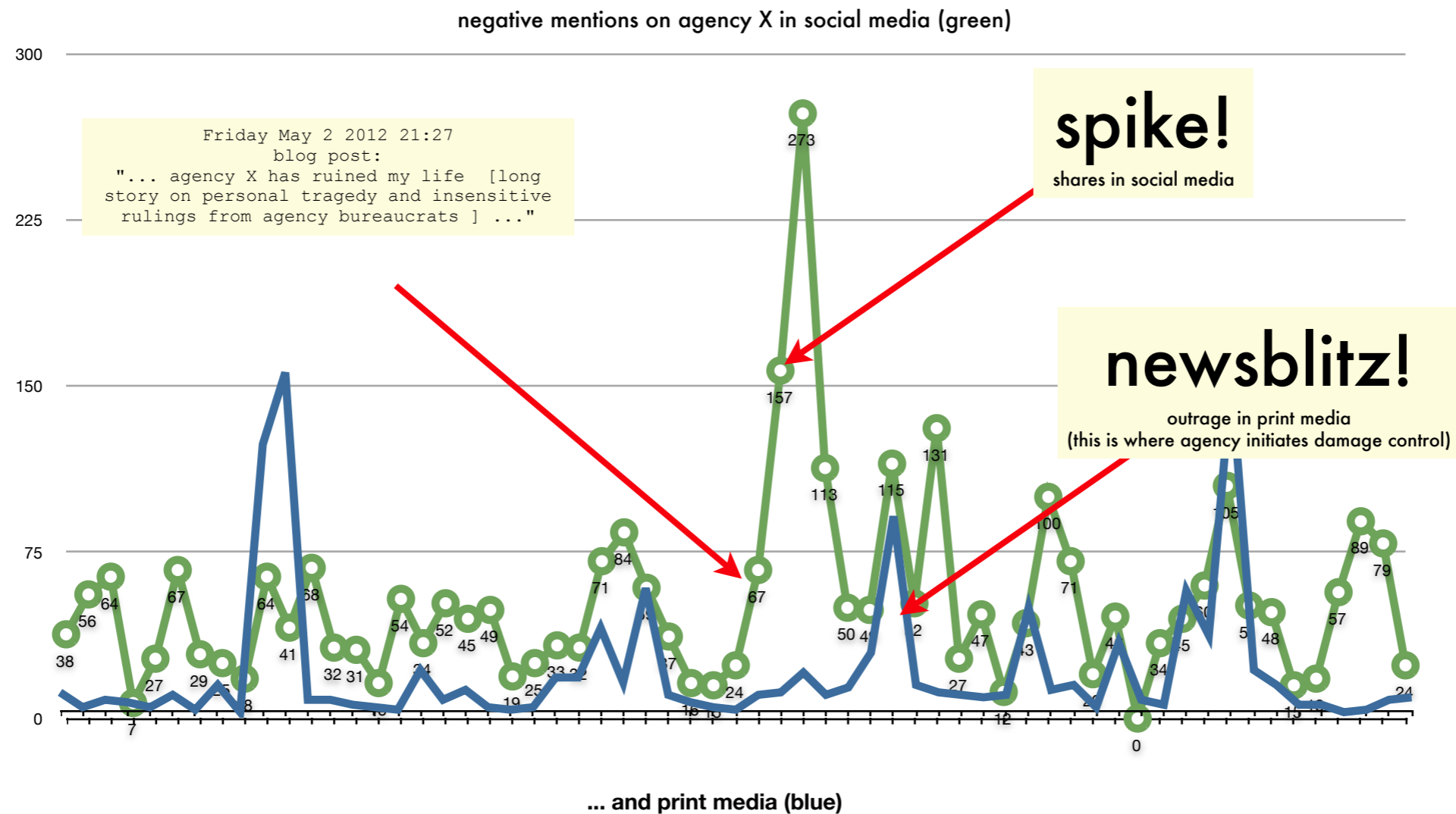
early warning in social media



early warning in social media



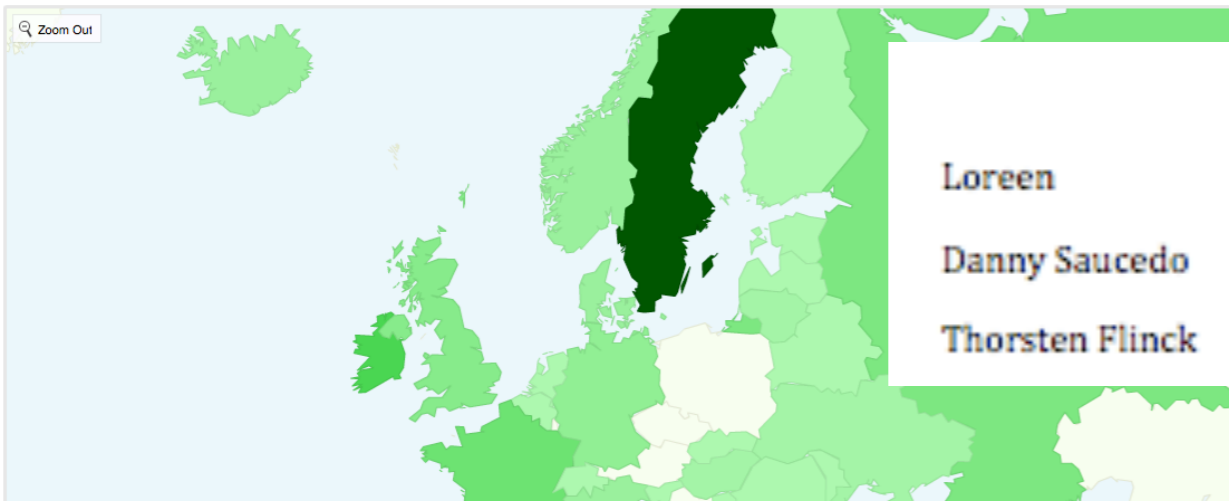
early warning in social media



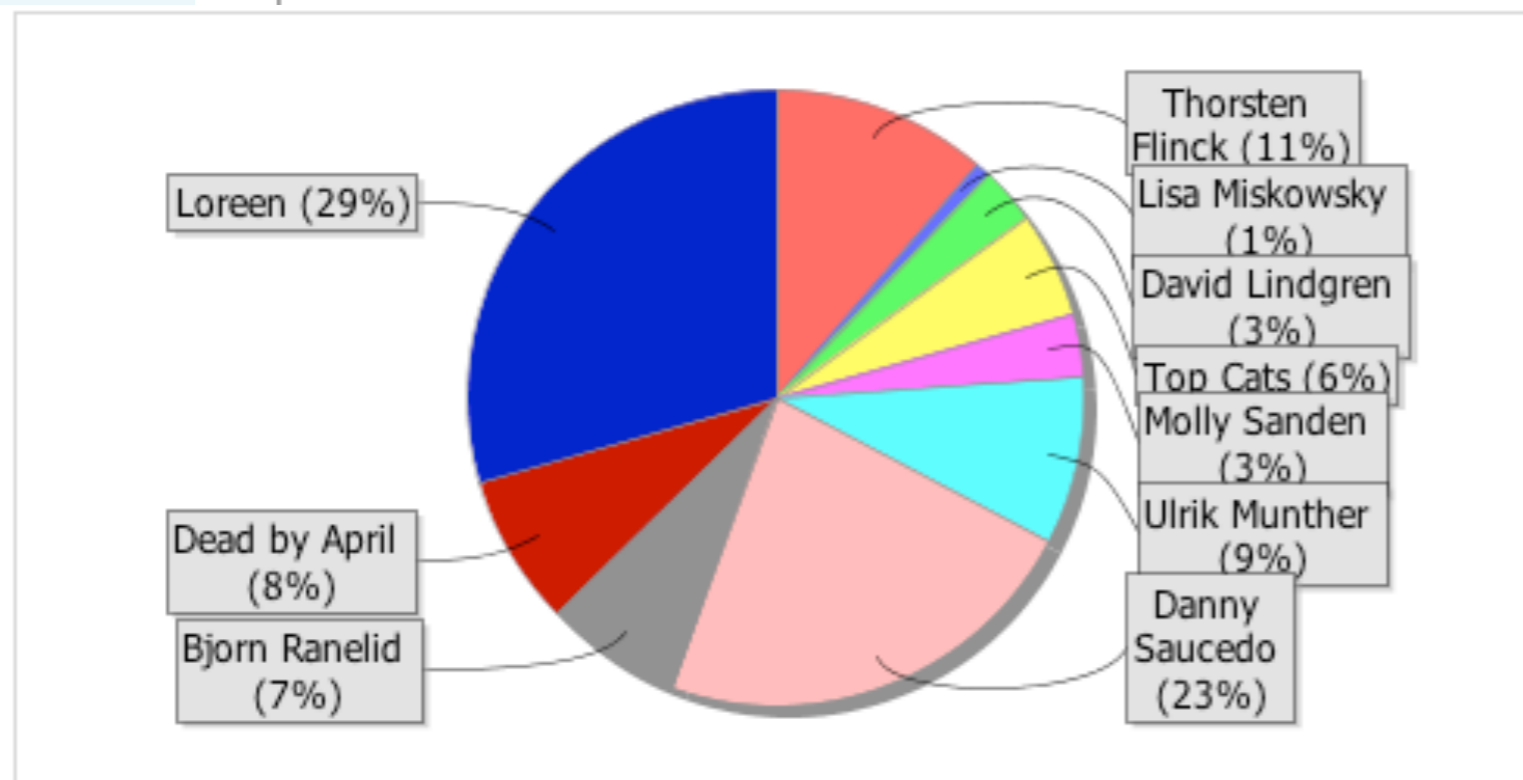
political analysis: unrest or distrust?



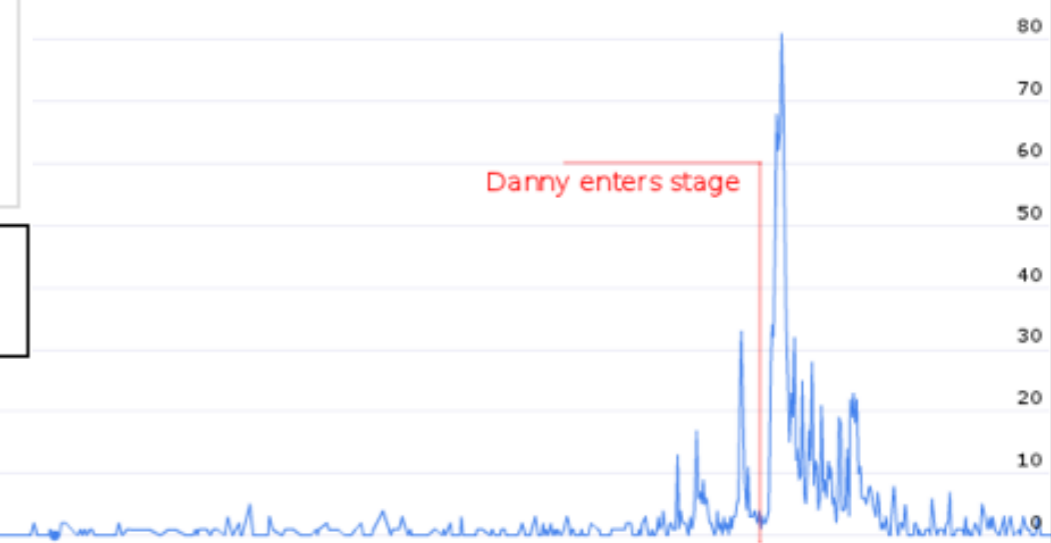
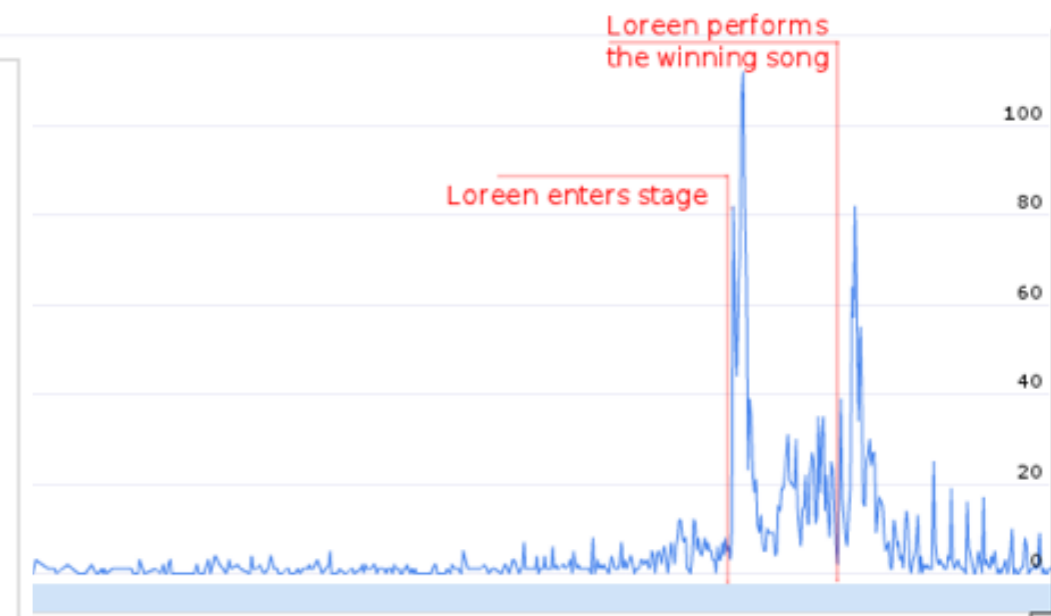
european song contest 2012



Gavagai Forecast		Viewer votes	
Loreen	30% (1 st)	33% (1 st)	
Danny Saucedo	22% (2 nd)	22% (2 nd)	
Thorsten Flinck	12% (3 rd)	8% (3 rd)	



- Thorsten Flinck ● Lisa Miskowsky ● David Lindgren ● Top Cats ● Molly Sanden
- Ulrik Munther ● Danny Saucedo ● Bjorn Ranelid ● Dead by April ● Loreen



so what are some of the challenges?

big data

@netnod @lynnstamour @klindqvist To some degree more important what other not named people have done, Matti Rendahl for example. #ind I I

grins evil
sup
lol well.....dunno lol
wat
i wud tap that
u lik?
rly?
thx stranger

iverapport: Bruce Schneier först ut på Internetdagarna. <http://t.co/0Q7yXIJ> #ind I I

ruce Schneier on powers that want to change Internet infrastructure to facilitate surveillance and censorship #ind I I #telecom

new text needs robust analysis

ynn St. Amour is speaking at #Internet Days in Stockholm this morning <http://t.co/74AHJMfZ> #ind I I

the yr 5's enit i wos callin ya but u were 2 busy eatin all da cakes lol.

xamples of #cyberwar, we love to use the word war when there's no war in sight. Bruce Schneier <http://xwalck.se/j/pic/PB219252.JPG> #ind I I

we r the sissters adn we wil cill u al bcuz u dun belay our story!!!

ruce Schneier at #ind I I, we like to use "war" in peace time, like cyber war. We don't use it for actual war. No good def of cyber war. #ind I I (@ Stockholm Waterfront Congress Centre w/ 15 others)

is bcuz she is dead she tried 2 come bak then and sopedly got out of 2 sexy 4 u? and all the g's and a's are similar to each other ...

http://t.co/hgG7BEwJ

heeh
had a kewl day morph?

ot feeling well - so no #ind I I today. Meet @martinaalm from @tripbirds at instead. Ping @jonasl @jocke

=)
Greeting
i came home 2 am .and went up at 9
sum bastard woke me up @ 2pm

Internetdagarna: #ind I I (@ Stockholm Waterfront Congress Centre w/ 7 others) <http://t.co/lf3eLdOT>

=P
heh naah.
Irritation i h8 dat guy
lol
phoner on the sms?
:)

bjr, Adele!! koi29?
lut, ca va?

Gavagai 

whutta dowk

scale and crowd is essential to meaning



Gavagai



noise is not something you wash off



Gavagai



change is normal

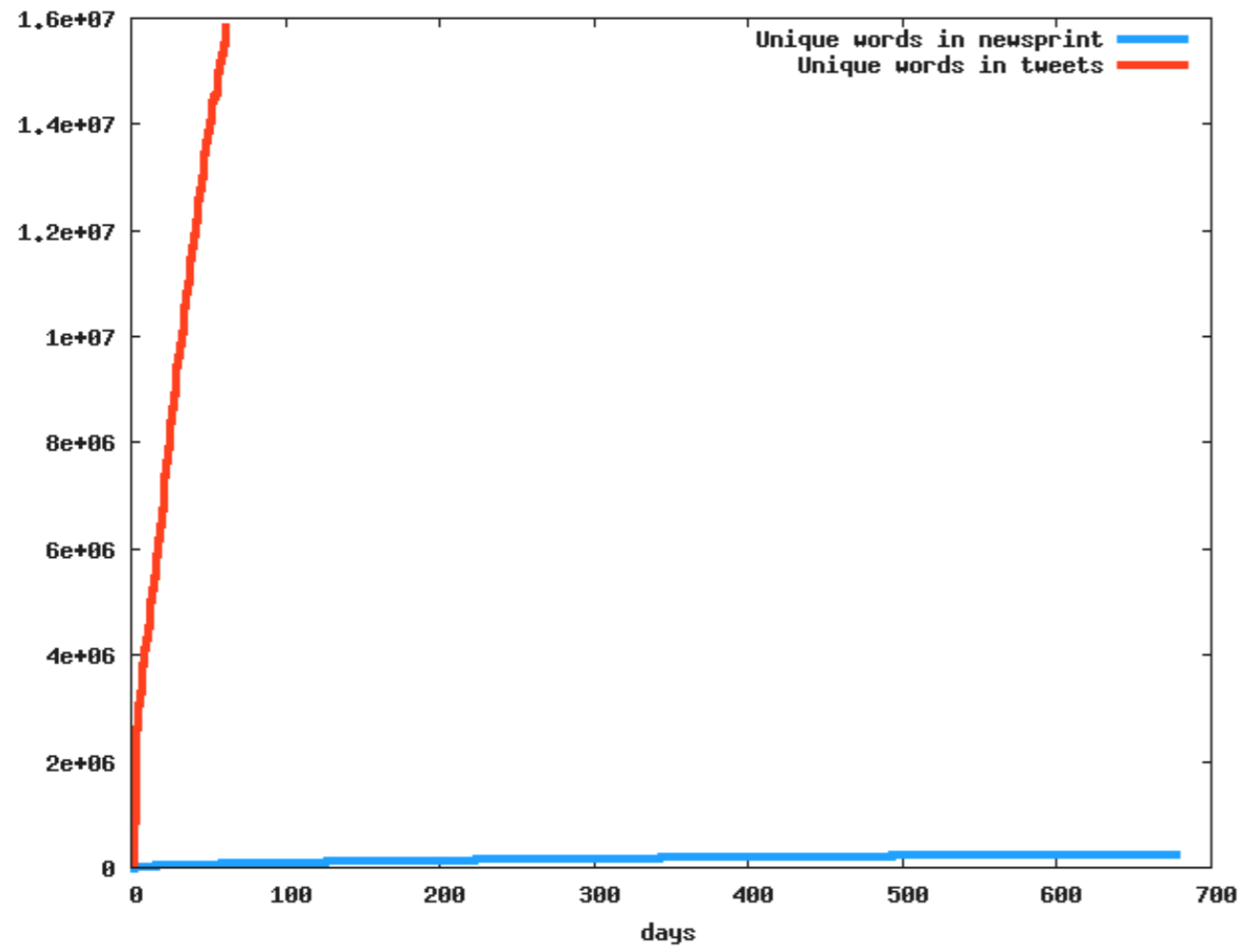


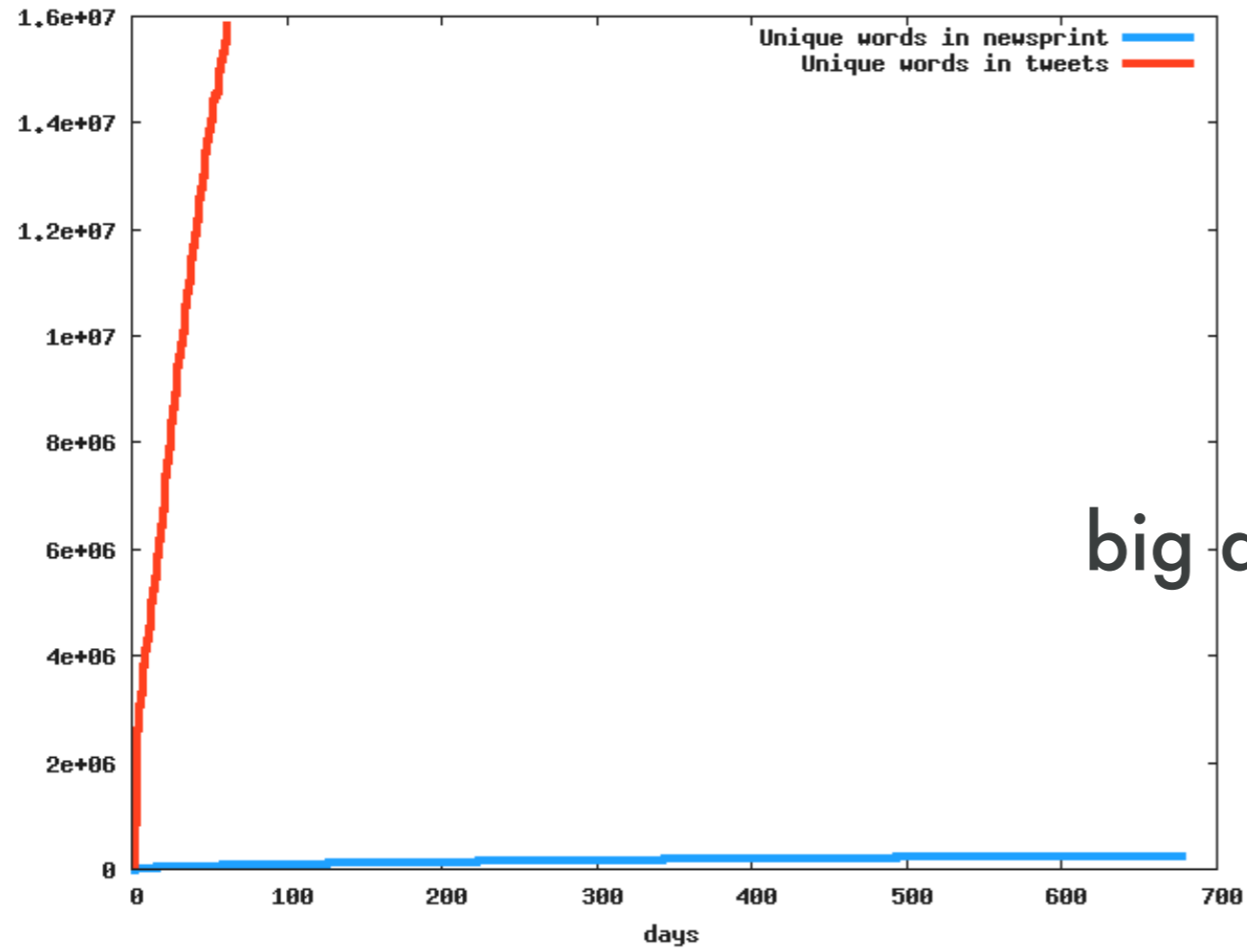
Gavagai



learning, not training

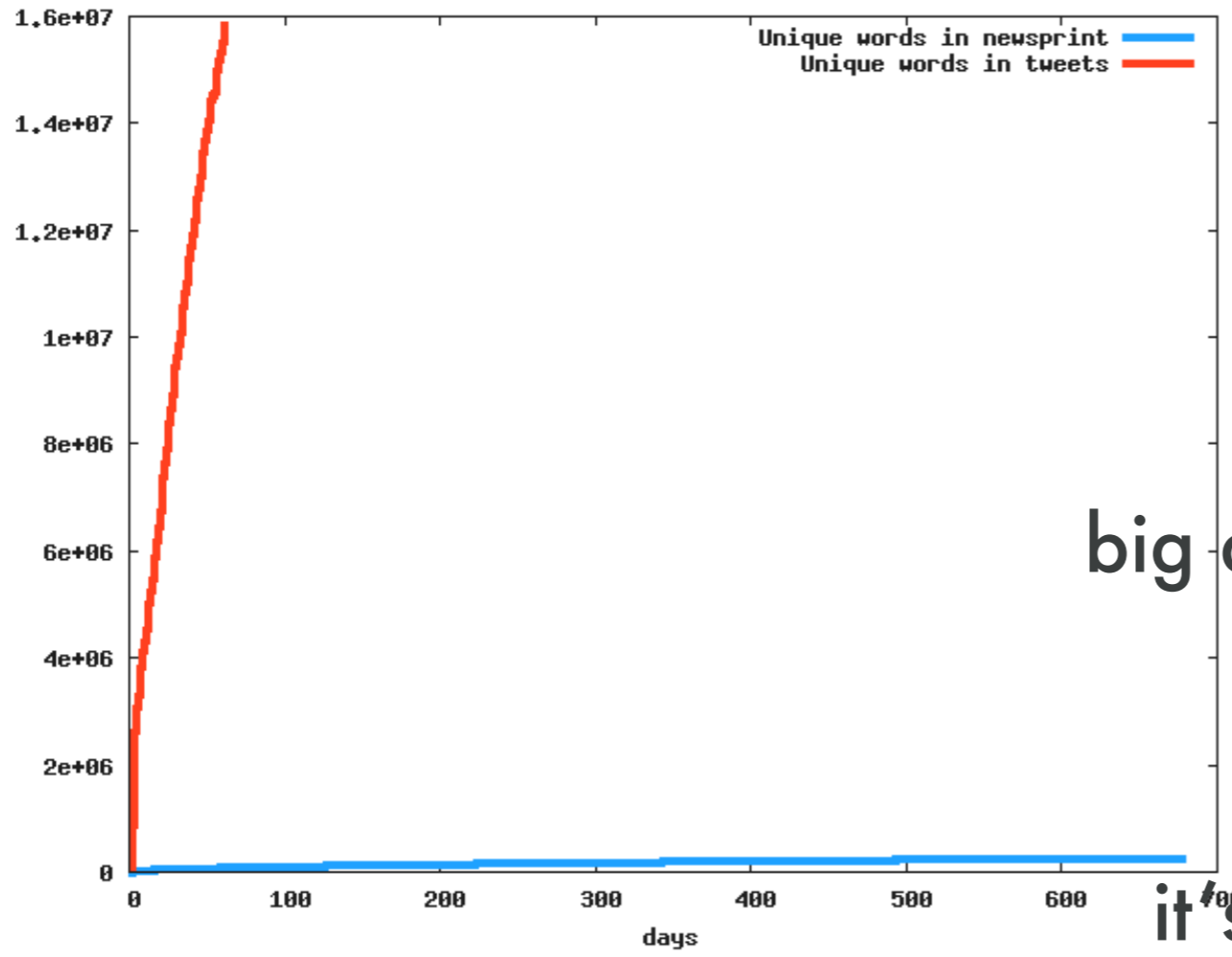






big data changes everything

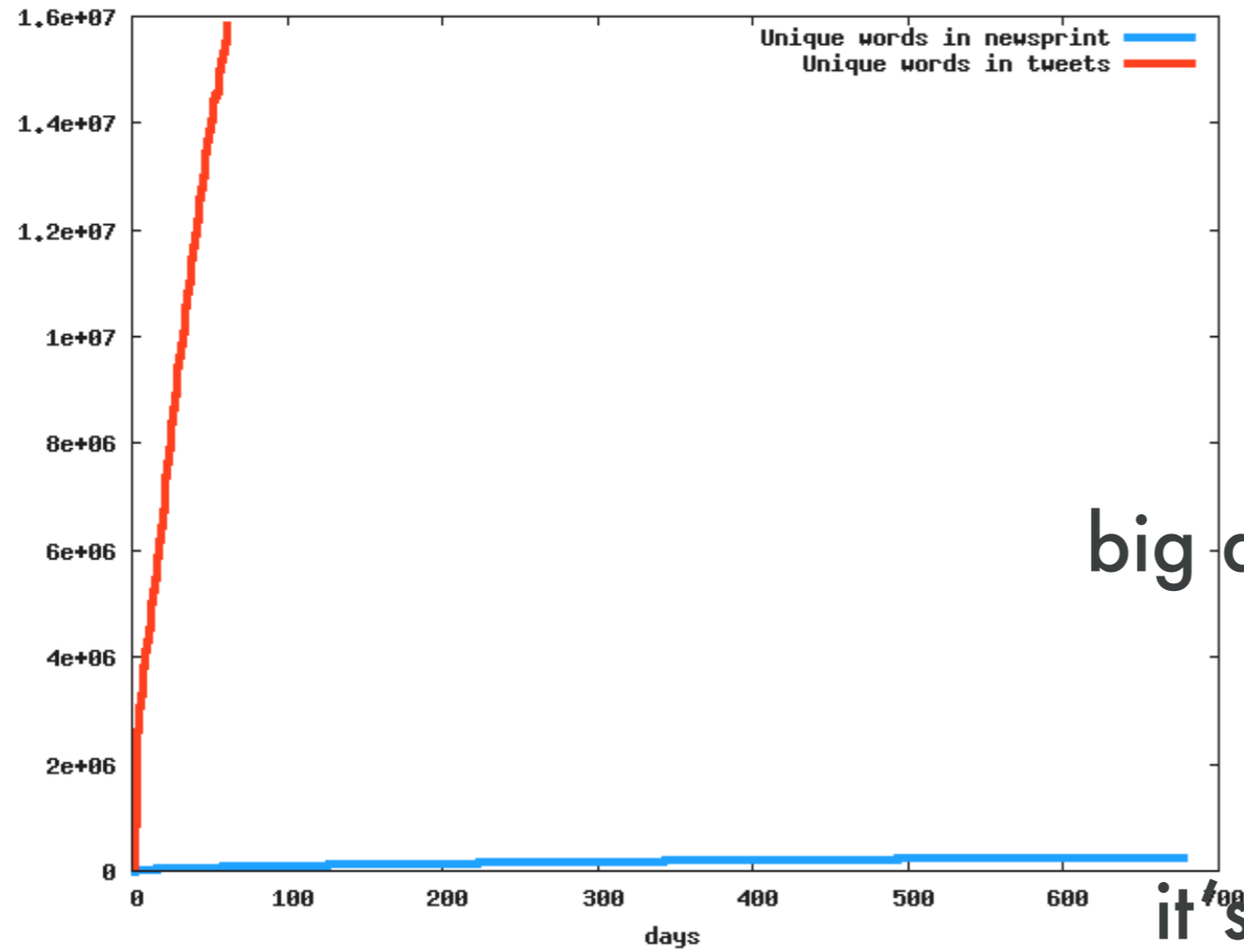




big data changes everything

it's not a scalability issue





big data changes everything

it's not a scalability issue

the stream and the scale *is* the data



humans excel at this sort of task

human information processing handles near misses and variation

human information processing handles near misses and variation

humans not work by definitions but by analogy to the typical



human information processing handles near misses and variation

humans not work by definitions but by analogy to the typical

humans classify previously unseen items after few observations



human information processing handles near misses and variation

humans not work by definitions but by analogy to the typical

humans classify previously unseen items after few observations

human information processing does not rely on explicitly stored
and recallable experience



how often have you learnt a term by its definition?

how often have you learnt a term by its definition?

how often do you need to ask what someone means?

how often have you learnt a term by its definition?

how often do you need to ask what someone means?

can you tell us about all the times you have encountered some
certain term in use?

emotion is a complex notion

models of emotion and affect



models of emotion and affect

{POS,NEG,NEU}



models of emotion and affect

{POS,NEG,NEU}

*{anger, fear, sadness, enjoyment, disgust,
surprise, contempt}*



models of emotion and affect

{POS,NEG,NEU}

*{anger, fear, sadness, enjoyment, disgust,
surprise, contempt}*

[arousal x valence x strength]



And the sound quality - my God!

And the sound quality - my God!

Raymond left no room for error on his recordings and it shows.



And the sound quality - my God!

Raymond left no room for error on his recordings and it shows.

Definitely one of the better tracks on the album.



And the sound quality - my God!

Raymond left no room for error on his recordings and it shows.

Definitely one of the better tracks on the album.

Wow, could have been a expansion pack.



And the sound quality - my God!

Raymond left no room for error on his recordings and it shows.

Definitely one of the better tracks on the album.

Wow, could have been a expansion pack.

I loved The Spy Who Came In From The Cold but the movie is a bit dated in a way the book never will be.



And the sound quality - my God!

Raymond left no room for error on his recordings and it shows.

Definitely one of the better tracks on the album.

Wow, could have been a expansion pack.

I loved The Spy Who Came In From The Cold but the movie is a bit dated in a way the book never will be.

Meat is more environmentally friendly than seafood.



And the sound quality - my God!

Raymond left no room for error on his recordings and it shows.

Definitely one of the better tracks on the album.

Wow, could have been a expansion pack.

I loved The Spy Who Came In From The Cold but the movie is a bit dated in a way the book never will be.

Meat is more environmentally friendly than seafood.

I am unsure about the feasibility of this knitting pattern.



And the sound quality - my God!

Raymond left no room for error on his recordings and it shows.

Definitely one of the better tracks on the album.

Wow, could have been an expansion pack.

I loved *The Spy Who Came In From The Cold* but the movie is a bit dated in a way the book never will be.

Meat is more environmentally friendly than seafood.

I am unsure about the feasibility of this knitting pattern.

I love the Samsung B2710 but I would not recommend it to my colleagues.



And the sound quality - my God!

Raymond left no room for error on his recordings and it shows.

Definitely one of the better tracks on the album.

Wow, could have been a expansion pack.

I loved The Spy Who Came In From The Cold but the movie is a bit dated in a way the book never will be.

Meat is more environmentally friendly than seafood.

I am unsure about the feasibility of this knitting pattern.

I love the Samsung B2710 but I would not recommend it to my colleagues.

I don't know if I should call her up - I liked her when I met her last weekend.



And the sound quality - my God!

Raymond left no room for error on his recordings and it shows.

Definitely one of the better tracks on the album.

Wow, could have been a expansion pack.

I loved The Spy Who Came In From The Cold but the movie is a bit dated in a way the book never will be.

Meat is more environmentally friendly than seafood.

I am unsure about the feasibility of this knitting pattern.

I love the Samsung B2710 but I would not recommend it to my colleagues.

I don't know if I should call her up - I liked her when I met her last weekend.

This is true.



And the sound quality - my God!

Raymond left no room for error on his recordings and it shows.

Definitely one of the better tracks on the album.

Wow, could have been a expansion pack.

I loved The Spy Who Came In From The Cold but the movie is a bit dated in a way the book never will be.

Meat is more environmentally friendly than seafood.

I am unsure about the feasibility of this knitting pattern.

I love the Samsung B2710 but I would not recommend it to my colleagues.

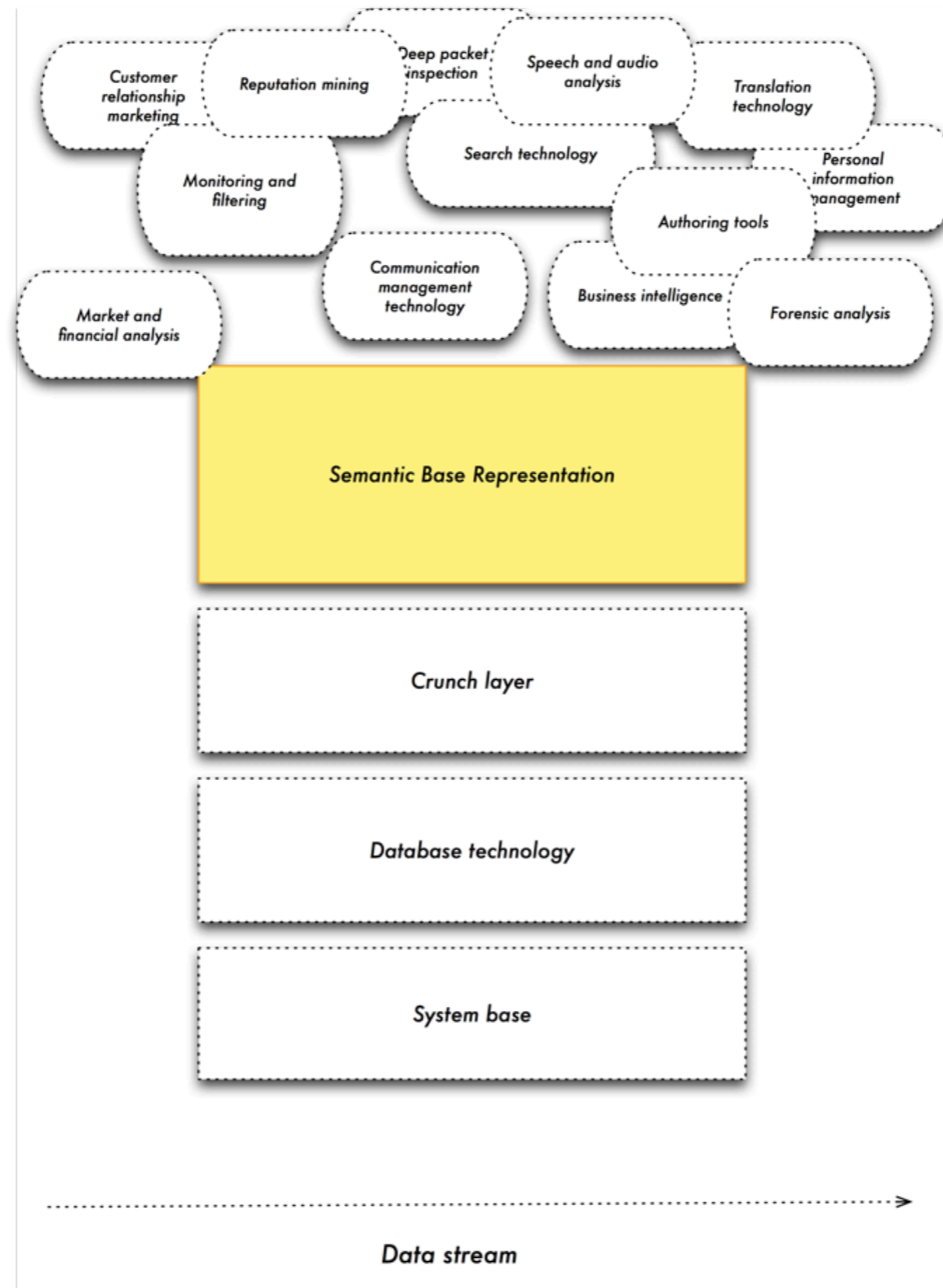
I don't know if I should call her up - I liked her when I met her last weekend.

This is true.



we believe in building a base technology

technology enabler: big data stack



a semantic base technology?

is this an example of that?

are these two the same?

has this changed? how?

what is the relation of this and that?

is this a new way of saying that?

are these or those more like this?

is this typical or strange?

can we trust this?

does the author believe this to be true?



so what do we have to do to meet the
challenges?

so what do we have to do to meet the
challenges?

so what do we have to do to meet the
challenges?

- challenge: robustness

so what do we have to do to meet the
challenges?

- challenge: robustness
- challenge: scale

distributional semantics



distributional semantics

observations with similar distributions have similar meanings



distributional semantics

observations with similar distributions have similar meanings

or



distributional semantics

observations with similar distributions have similar meanings

or

observations that occur in the same contexts have similar meaning



distributional semantics

observations with similar distributions have similar meanings

or

observations that occur in the same contexts have similar meaning

and



distributional semantics

observations with similar distributions have similar meanings

or

observations that occur in the same contexts have similar meaning

and

observations that co-occur near each other build meaning together



distributional semantics

observations with similar distributions have similar meanings

or

observations that occur in the same contexts have similar meaning

and

observations that co-occur near each other build meaning together

and



distributional semantics

observations with similar distributions have similar meanings

or

observations that occur in the same contexts have similar meaning

and

observations that co-occur near each other build meaning together

and

observations that co-occur over some distance have topical relations



distributional models as vector spaces

the weather is great in barcelona

the weather is chilly in bergen

the weather is gray in stockholm

the weather is nippy in moscow

the weather is nice in hong kong

the climate is passable in nice

the weather in barcelona is balmy

...

the climate is chilly at the office

...

	winter texts	water texts	night life texts
STHLM	+	+	-
BCN	-	+	+
MOS	+	-	-
HKK	-	+	-

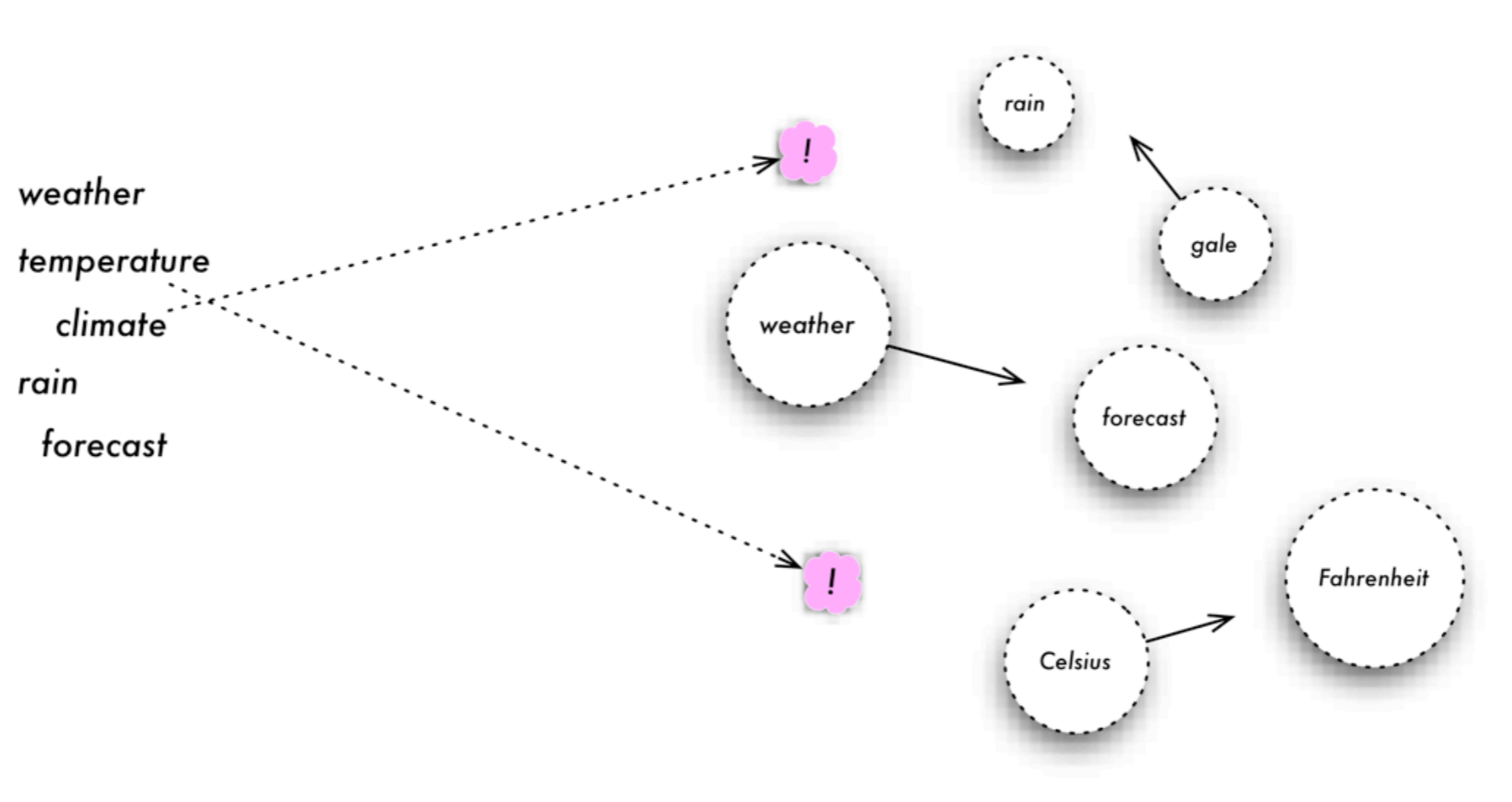
	STHLM	BCN	MOS	HKK	cold	warm	hot	harbour
STHLM	0	+	+	+	++	+	-	+
BCN	+	0	+	+	-	++	+	++
MOS	+	+	0	+	+++	-	+	-
HKK	+	+	+	0	-	++	++	+++
weather	+	+	+	+	+++	+	+++	+
climate	+	+	+	+	+++	+	+++	-

parameters of variation

- observations: words? multi-word-terms? constructions?
- contexts: clauses, paragraphs, N neighbouring words, documents?
- frequency weighting: idf? entropy? information measures? odds?
- dimension reduction: random projections? principal components?
- similarity measure?



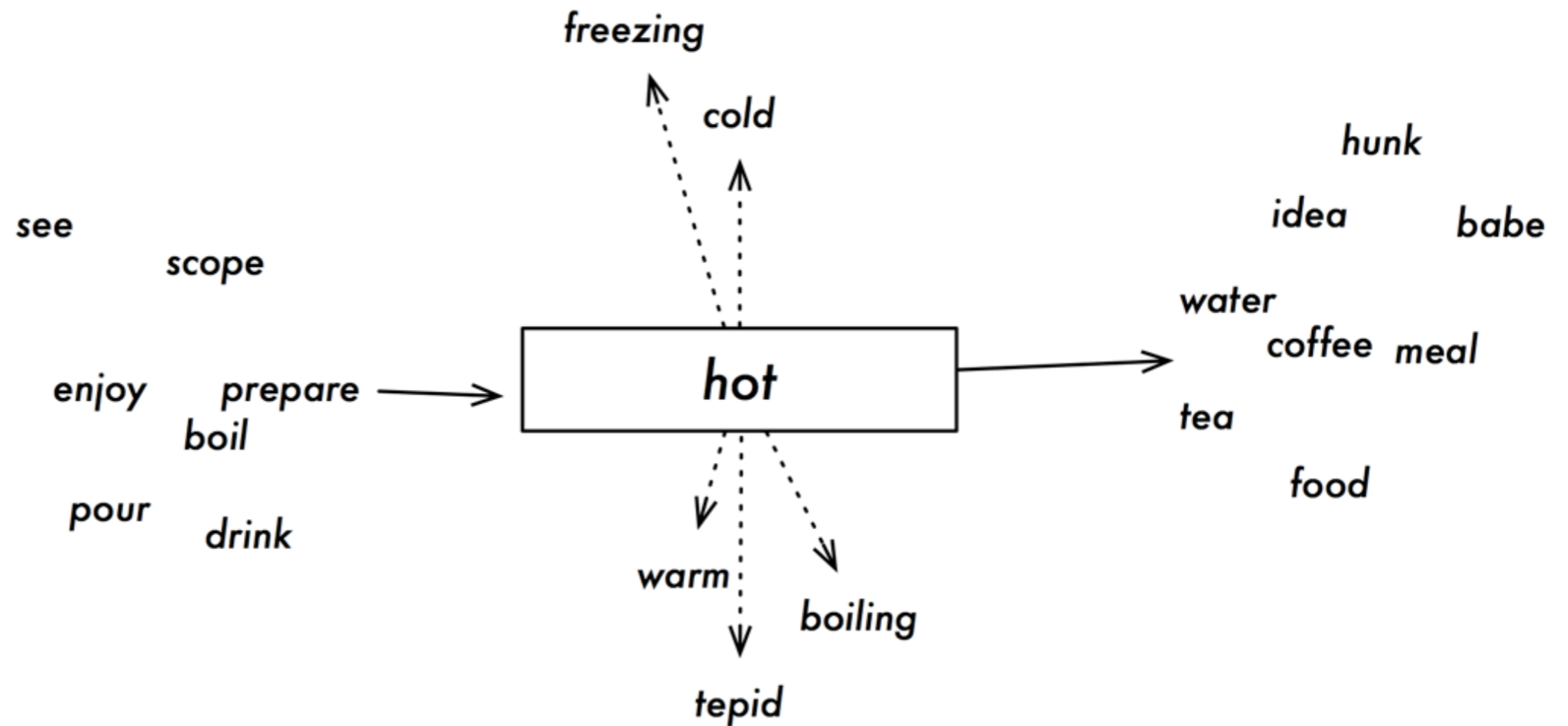
random indexing



randomness is the path of least assumption

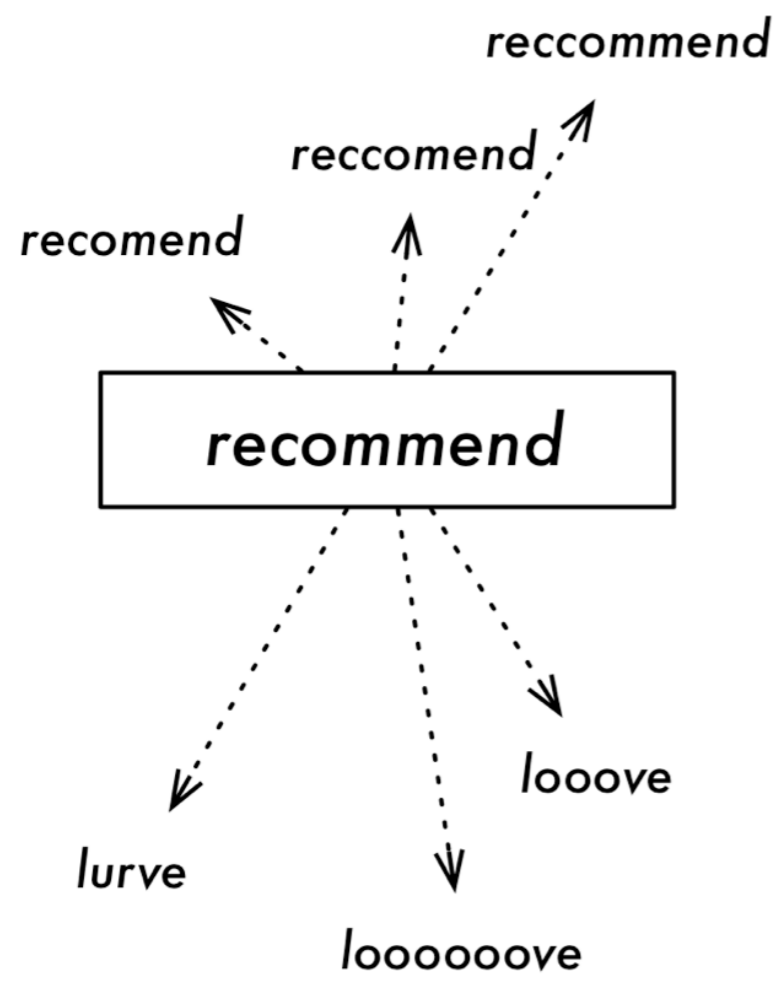
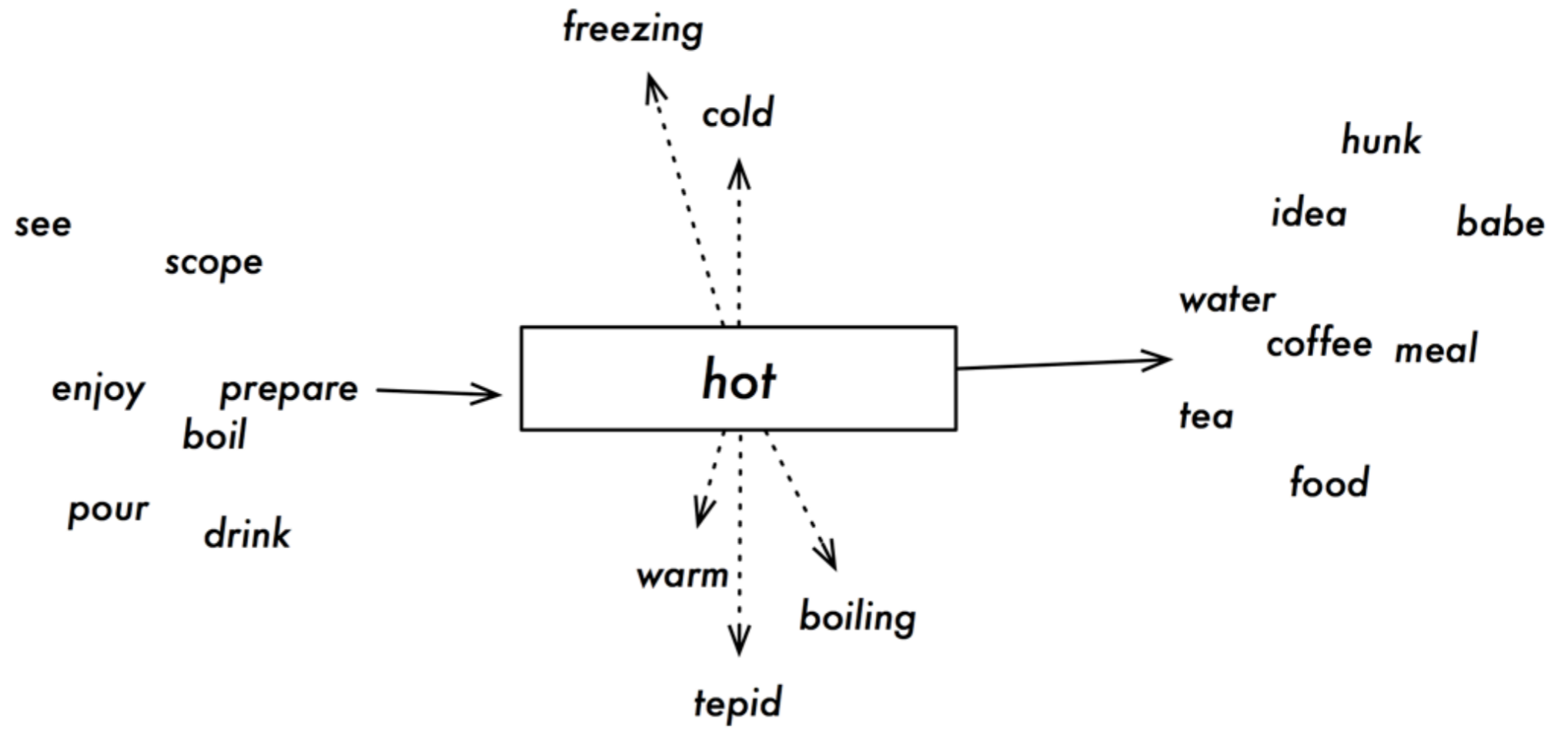
- online learning of term relations
- from distributional data
- complete – not based on sampling
- robust by virtue of processing model
- language-independent





- online learning of term relations
- from distributional data
- complete – not based on sampling
- robust by virtue of processing model
- language-independent

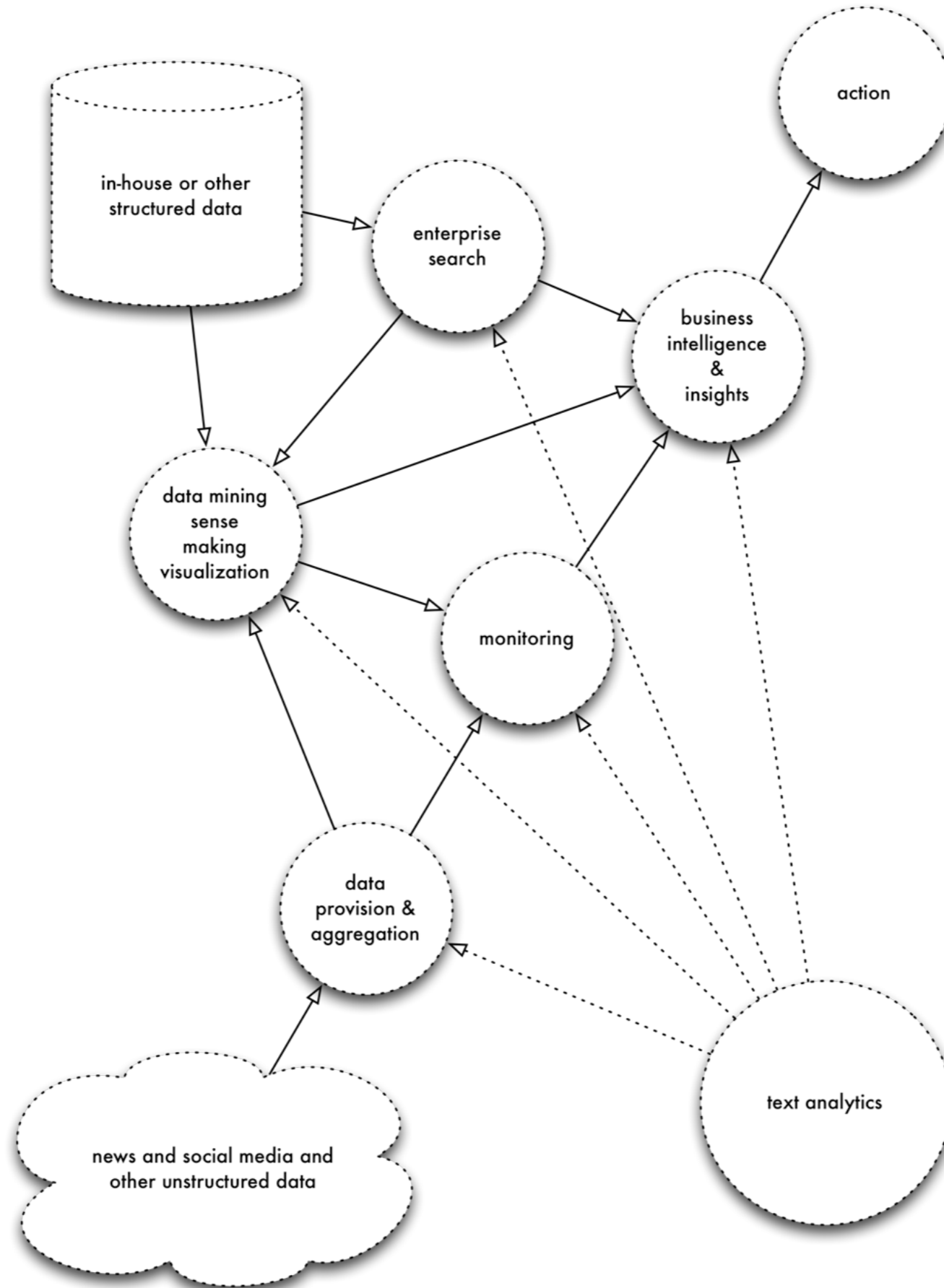




- online learning of term relations
- from distributional data
- complete – not based on sampling
- robust by virtue of processing model
- language-indepedent

remaining challenges

rapidly evolving business ecosystem



new task space

beyond search



beyond search



it is not about the needle in the hay

beyond search



it is not about the needle in the hay

it is not about search and retrieval, it is about
understanding



business processes not mature yet

domain issues

application: finance

application: finance

predict price movements from social media

application: finance

predict price movements from social media

- do traders tweet?



application: finance

predict price movements from social media

- do traders tweet?
- do people tweet about assets?



application: finance

predict price movements from social media

- do traders tweet?
- do people tweet about assets?
- does public mood predict trading price?



application: finance

predict price movements from social media

- do traders tweet?
- do people tweet about assets?
- does public mood predict trading price?

mood? opinion? purchase propensity?



application: consumer purchase decision

application: consumer purchase decision

find reviews or comments made about some
product or service by others

application: consumer purchase decision

find reviews or comments made about some
product or service by others

attitude and opinion!



application: consumer purchase decision

find reviews or comments made about some
product or service by others

attitude and opinion!

coverage, relevant target, diversity, source quality



application: security

application: security

- do bad guys tweet?

application: security

- do bad guys tweet?
- does expression of public mood predict unrest?

application: security

- do bad guys tweet?
- does expression of public mood predict unrest?
- is bluster different from bite?



application: security

- do bad guys tweet?
- does expression of public mood predict unrest?
- is bluster different from bite?



application: customer sentiment

application: customer sentiment

consumers comments wrt a product or service

application: customer sentiment

consumers comments wrt a product or service

attitude and opinion - beyond POS-NEG-NEU!



application: customer sentiment

consumers comments wrt a product or service

attitude and opinion - beyond POS-NEG-NEU!

target relevance; source clout; weak signal coverage;

actionability



back to usefulness of sentiment analysis

back to usefulness of sentiment analysis

big data overwhelms current approaches: not just more disk!



back to usefulness of sentiment analysis

big data overwhelms current approaches: not just more disk!

the stream is the signal: no sampling!



back to usefulness of sentiment analysis

big data overwhelms current approaches: not just more disk!

the stream is the signal: no sampling!

answer right now: no off-line training!



back to usefulness of sentiment analysis

big data overwhelms current approaches: not just more disk!

the stream is the signal: no sampling!

answer right now: no off-line training!

data is messy: no washing or static resources!



back to usefulness of sentiment analysis

big data overwhelms current approaches: not just more disk!

the stream is the signal: no sampling!

answer right now: no off-line training!

data is messy: no washing or static resources!

not needles in haystacks but the state of the world as it happens



back to usefulness of sentiment analysis

big data overwhelms current approaches: not just more disk!

the stream is the signal: no sampling!

answer right now: no off-line training!

data is messy: no washing or static resources!

not needles in haystacks but the state of the world as it happens

not positive and negative - way more complex!



back to usefulness of sentiment analysis

big data overwhelms current approaches: not just more disk!

the stream is the signal: no sampling!

answer right now: no off-line training!

data is messy: no washing or static resources!

not needles in haystacks but the state of the world as it happens

not positive and negative - way more complex!

can enable a new generation of applications



back to usefulness of sentiment analysis

big data overwhelms current approaches: not just more disk!

the stream is the signal: no sampling!

answer right now: no off-line training!

data is messy: no washing or static resources!

not needles in haystacks but the state of the world as it happens

not positive and negative - way more complex!

can enable a new generation of applications

but we do not yet know which or how, which means ...



back to usefulness of sentiment analysis

big data overwhelms current approaches: not just more disk!

the stream is the signal: no sampling!

answer right now: no off-line training!

data is messy: no washing or static resources!

not needles in haystacks but the state of the world as it happens

not positive and negative - way more complex!

can enable a new generation of applications

but we do not yet know which or how, which means ...

analyses needs to be built from principles, not from grep



back to usefulness of sentiment analysis

big data overwhelms current approaches: not just more disk!

the stream is the signal: no sampling!

answer right now: no off-line training!

data is messy: no washing or static resources!

not needles in haystacks but the state of the world as it happens

not positive and negative - way more complex!

can enable a new generation of applications

but we do not yet know which or how, which means ...

analyses needs to be built from principles, not from grep

