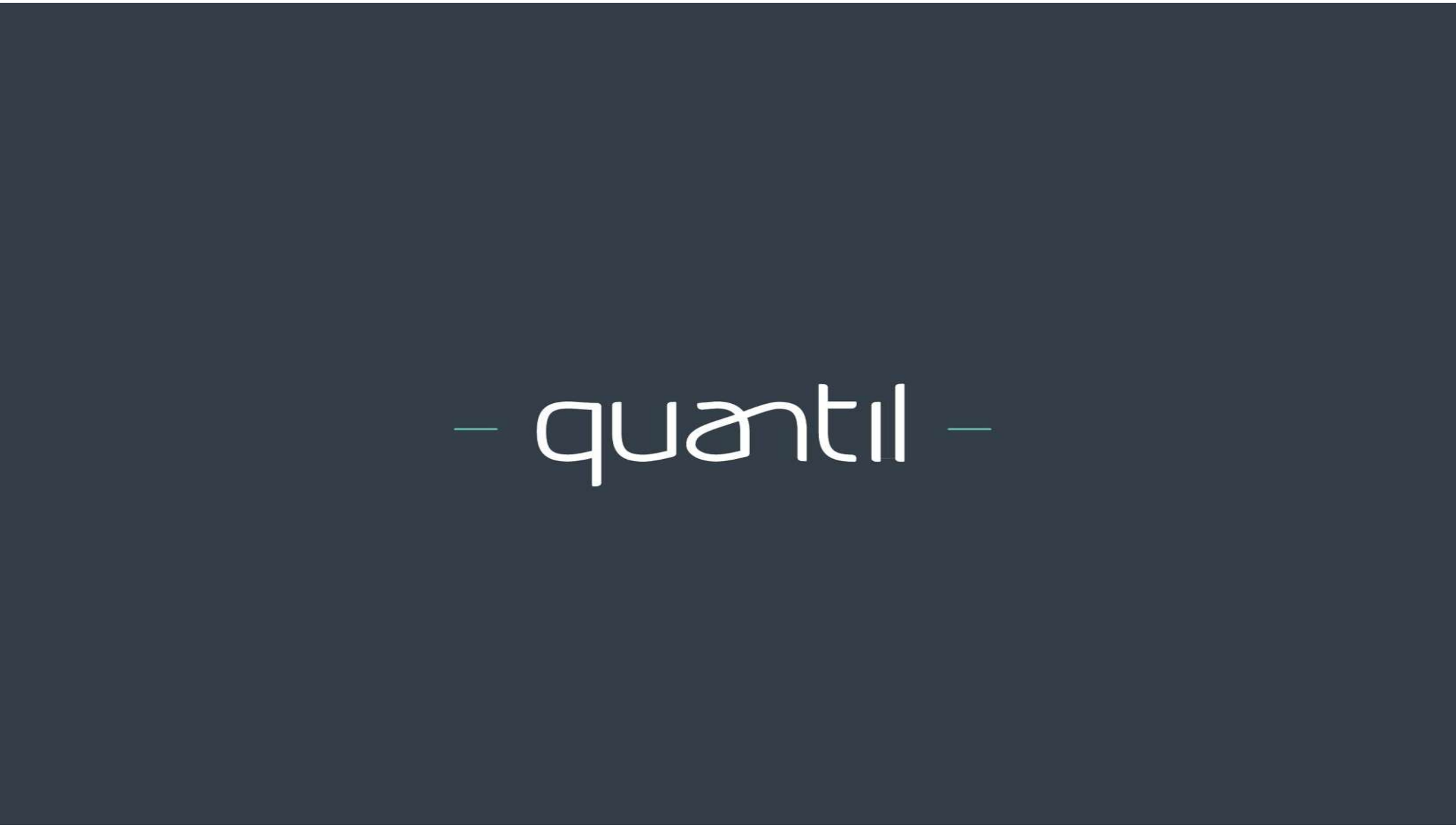
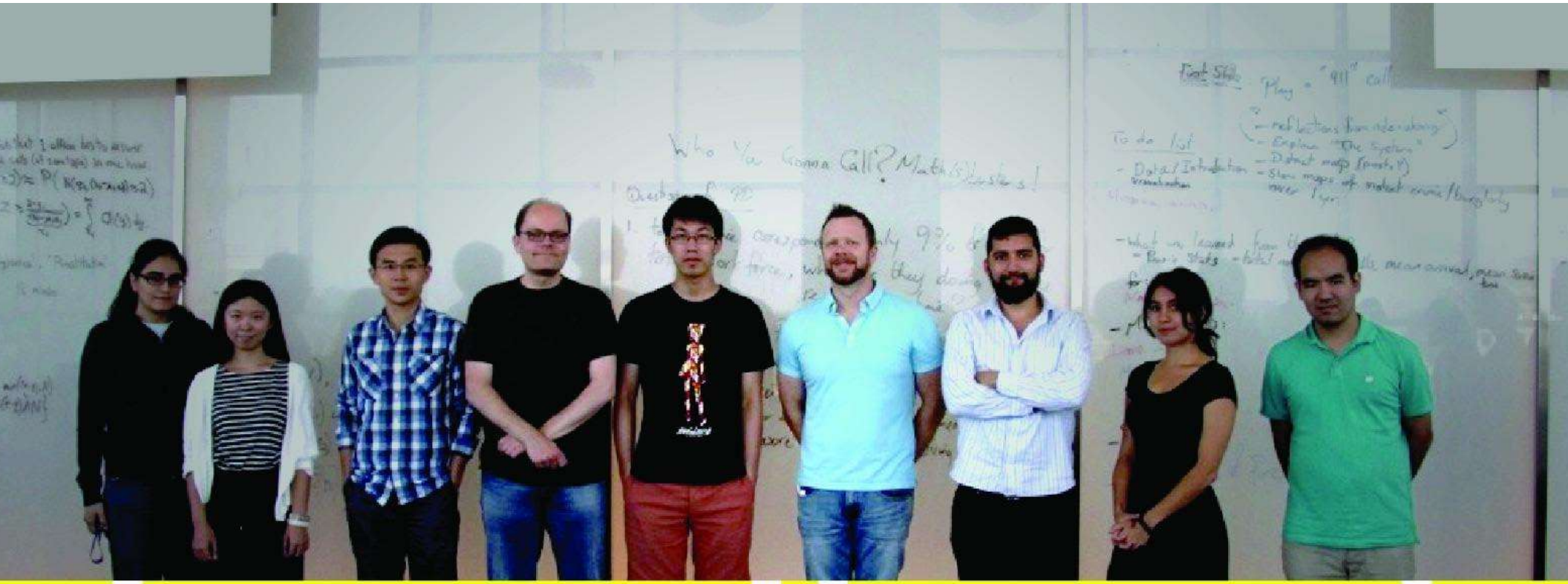

quantil

Guided NMF for Topic Detection

And other stuff with embeddings



AMERICA'S NEXT topic model

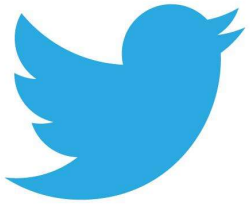
Introducción

¿Cuál es el sentido?

- Estructura a datos no estructurados.
- Análisis no supervisado.
- Explotar coocurrencia y contexto.
- Explotar propiedades de espacios vectoriales

Diferentes tipos de textos

- Reportes o textos jurídicos
  - Mucha longitud
  - Buena ortografía
  - Palabras repetidas y sofisticadas
- Artículos o blogs
  - Información sucinta
  - Longitud media
  - Lenguaje común y algo de repetición.
- Microblogs y SN status
  - No hay tópicos muy complejos
  - Corto.
  - No hay repetición pero si problemas ortográficos.



Descripción de los datos en ejercicio ICERM

- 2.2 Millones de tweets **geolocalizados** de Baltimore.
- Durante todo el 2015, año de los disturbios de Baltimore.
  - `#FreddieGray #BlackLivesMatter`
  - Una empresa vendía tweets geolocalizados históricos **baratos**.
- 1.1 Millones de términos diferentes sin preprocesamiento.
- Coordenadas Lat/Lon
- ID usuario
- Fecha

Tweetxamples

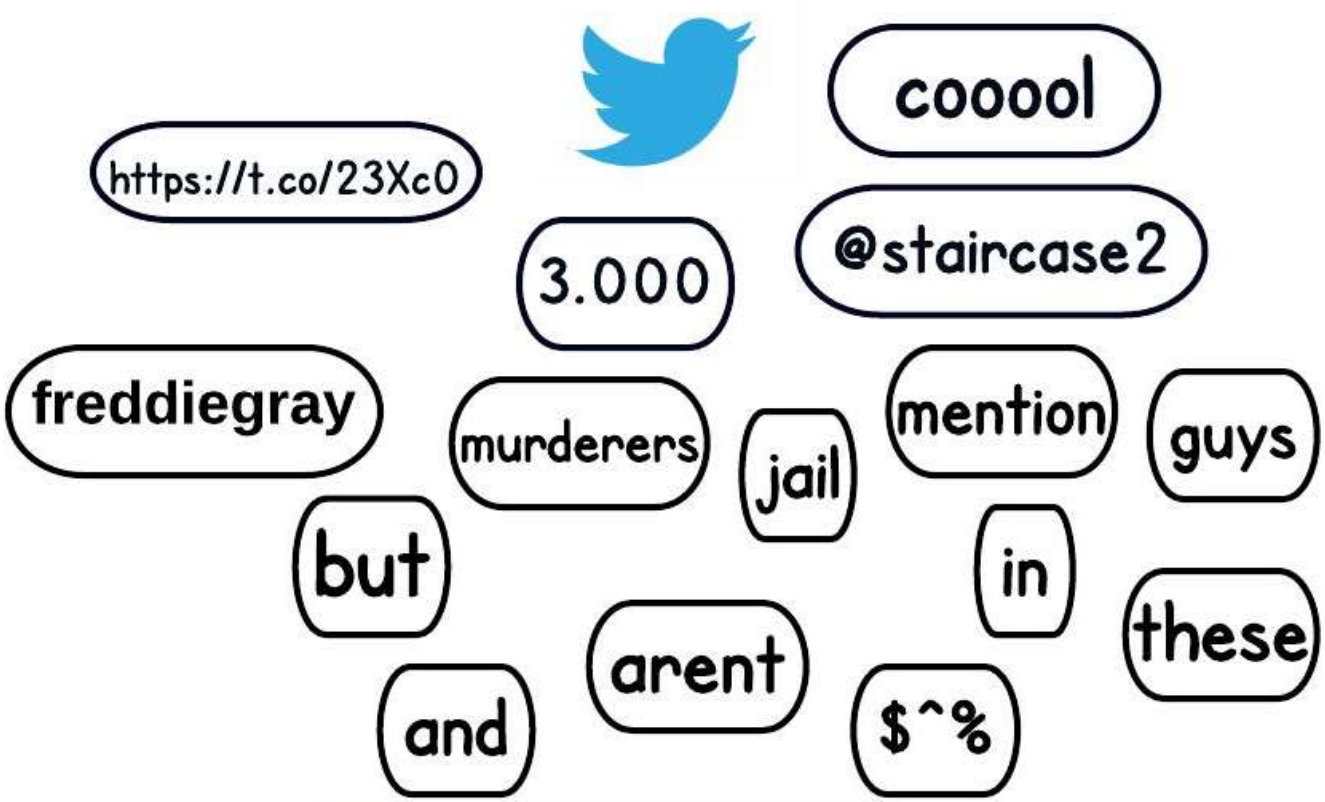
Where words go to die

A drinking straw only has one hole

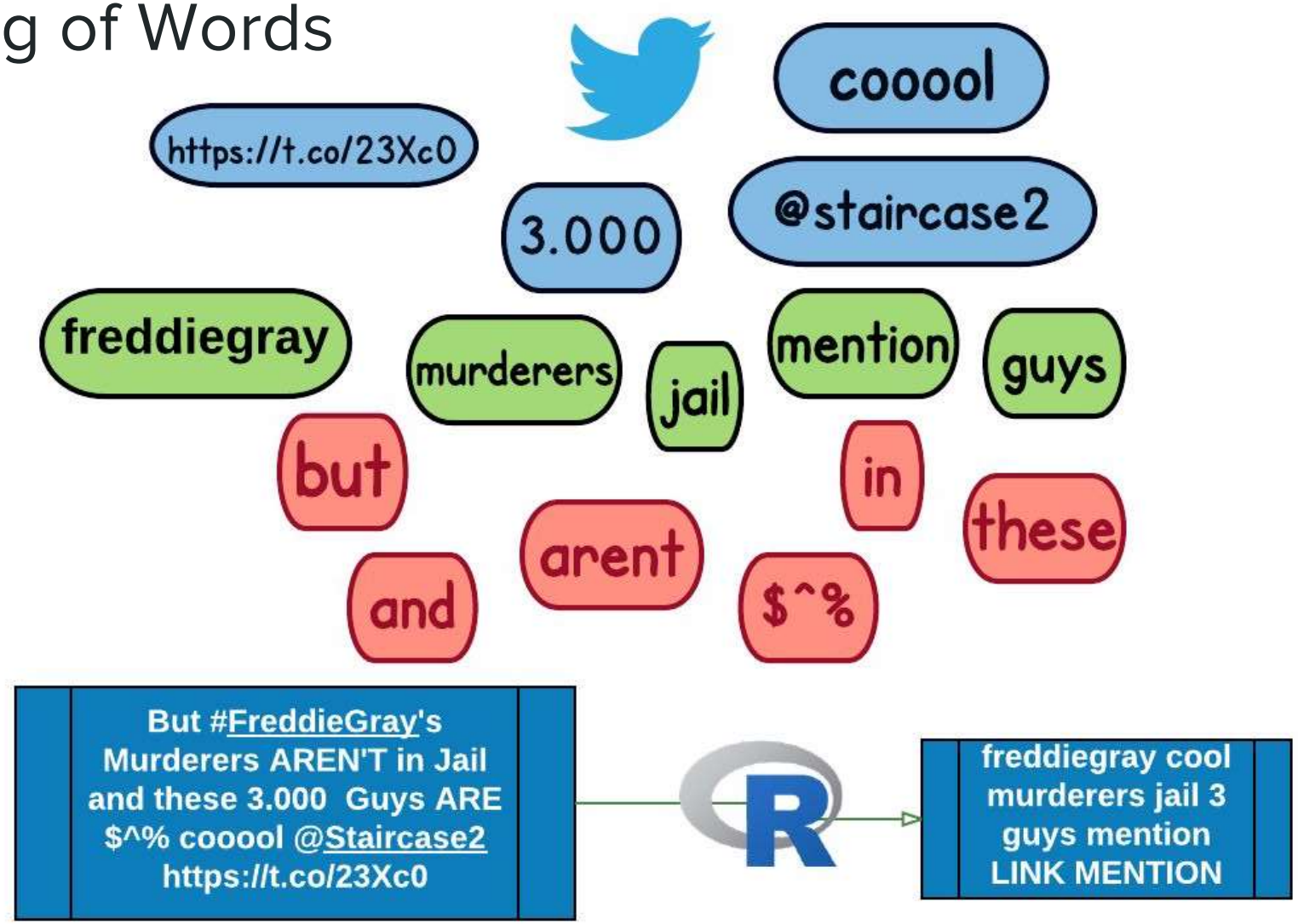
Where do I get the roast of beiber online. I'm drunk. Tryna party

â€œ@_xkayykayyx_: I'm rey start using @yeahmarco_ picture http://t.co/rsdYgyDAljâ€□????

they better not try and ban this video like the banned "man down" @rihanna â€” or we gonna start a **phuckin** RIOT!!


cooool
https://t.co/23Xc0
@staircase2
3.000
freddiegray
murderers
jail
mention
guys
but
in
these
arent
and
$^%
But #FreddieGray's
Murderers AREN'T in Jail
and these 3.000 Guys ARE
$^% cooool @Staircase2
https://t.co/23Xc0

Bag of Words


cooool
https://t.co/23Xc0
@staircase2
3.000
freddiegray
murderers
jail
mention
guys
but
in
these
arent
and
$^%
But #FreddieGray's Murderers AREN'T in Jail and these 3.000 Guys ARE $^% cooool @Staircase2 https://t.co/23Xc0
R
freddiegray cool murderers jail 3 guys mention LINK MENTION

Qué tanto se puede embeber el lenguaje en R^n ?

- Palabras, Tweets y Tópicos.
- Clusters, Estructura y Ángulos.
- Unidades intercambiables:
  - Problemas análogos.
  - Funciones multipropósito.

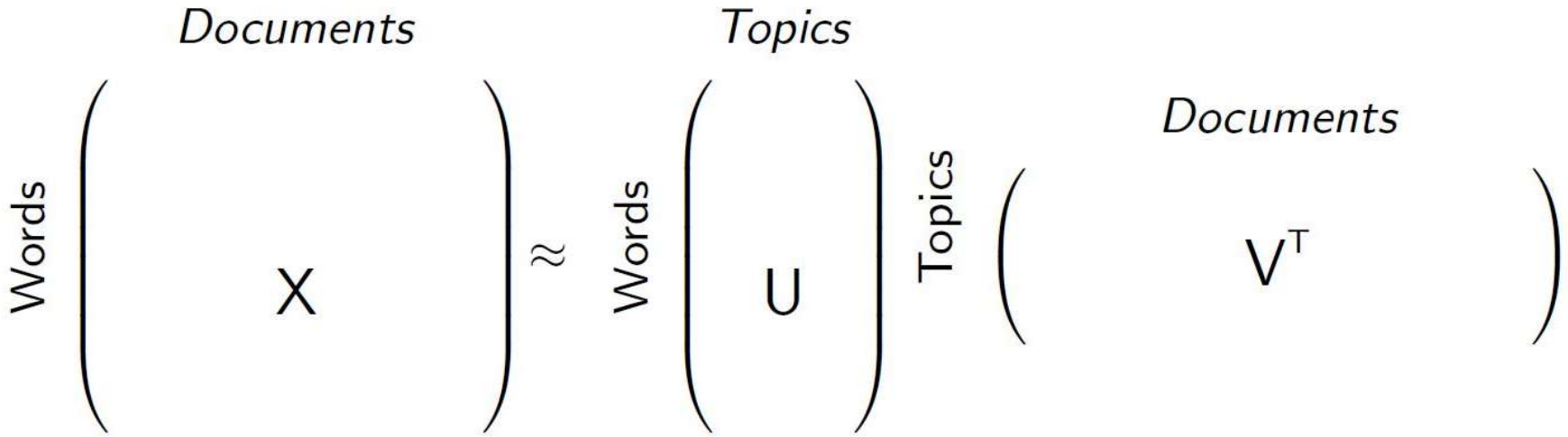
Teoría

Topic Model

Non-Negative Matrix Factorization

$$\min_{U,V} \|X - UV^T\|$$

where $U = [U]_+, V = [V]_+$.



Un poco de Álgebra: La mejor restricción

$$\begin{aligned}\|A\| &= \sup\{\|Ax\| : x \in K^n \text{ with } \|x\| = 1\} \\ &= \sup\left\{\frac{\|Ax\|}{\|x\|} : x \in K^n \text{ with } x \neq 0\right\}.\end{aligned}$$

VARIAS NORMAS DE MATRICES

$$\lim_{p\to 0} \|A\|_{p,1} = \sum_{j=1}^{n}\left(\sum_{i=1}^{m}|a_{ij}|^p\right)^{1/p}$$

Norma “Cero”: Ultracóncava

$$\|A\|_1 = \max_{1\le j\le n}\sum_{i=1}^{m}|a_{ij}|,$$

Norma 1: Eficiente

$$\|A\|_2 \le \left(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2\right)^{1/2} = \|A\|_F,$$

Norma Frobenius: De vector

Buscando matrices dispersas y soluciones rápidas

$$\text{Minimize } \frac{1}{2}\left(\|X - UV\|_F^2 + \alpha \|U\|_F^2 + \beta \left(\sum_{i=1}^{n} ||V(:,i)||_1^2\right)\right)$$

- La norma de Frobenius se puede minimizar fácilmente.
- La norma 1 es la mejor norma que produce optimización convexa para lograr Sparse Matrices. (La norma 0 es *intractable*)
- Queremos factorizar X.
- Que U no se haga muy grande (soluciones muy **extremas**)
- Y queremos que V sea dispersa → los tweets tienen **pocos tópicos**.

Precauciones con NMF

Y en general con la optimización convexa

- R es muy malo para optimizar.
- Cuidado: Mínimos locales.
- Crece exponencialmente con el número de tópicos.
- No es trivialmente paralelizable.
- Cómo se calibran los parámetros? Cómo se sabe que son buenos tópicos?

Input y Output de NMF Topic Modelling

- Una matriz de términos y documentos.
- Debe venir de un Corpus procesado.
- Usualmente se recibe una matriz dispersa (Simple Triplet Matrix) junto a un diccionario.

- El resultado son un par de matrices que minimizan la expresión.
- La matriz de tópicos por tweets es un *Embedding* de tweets.
- La matriz de términos por tópicos es una distribución de probabilidad.
- La matriz de términos por tópicos es un *Embedding* de palabras.

Other embeddings...

GloVe

- Trata de factorizar la matriz de coocurrencias.
- Co-ocurrencia se define en cierta ventana + *decay*

$$J = \sum_{i=1}^{V} \sum_{j=1}^{V} f(X_{ij})(w_i^T w_j + b_i + b_j - \log X_{ij})^2$$

S.A.

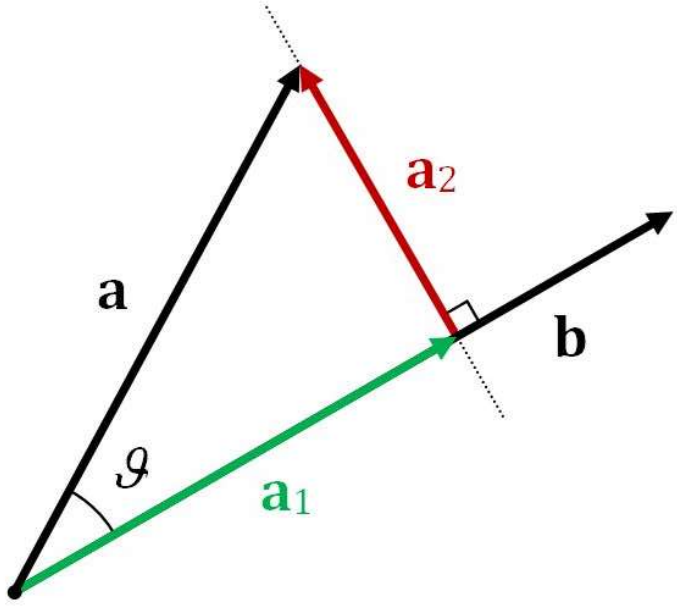
$$w_i^T w_j + b_i + b_j = log(X_{ij})$$

LDA

- Modelo generativo de tópicos.
- Expresa cada documento como una distribución de tópicos.
- Cada tópico es una distribución de palabras.
- Estima las distribuciones óptimas usando máxima verosimilitud.

GloVe vs NMF

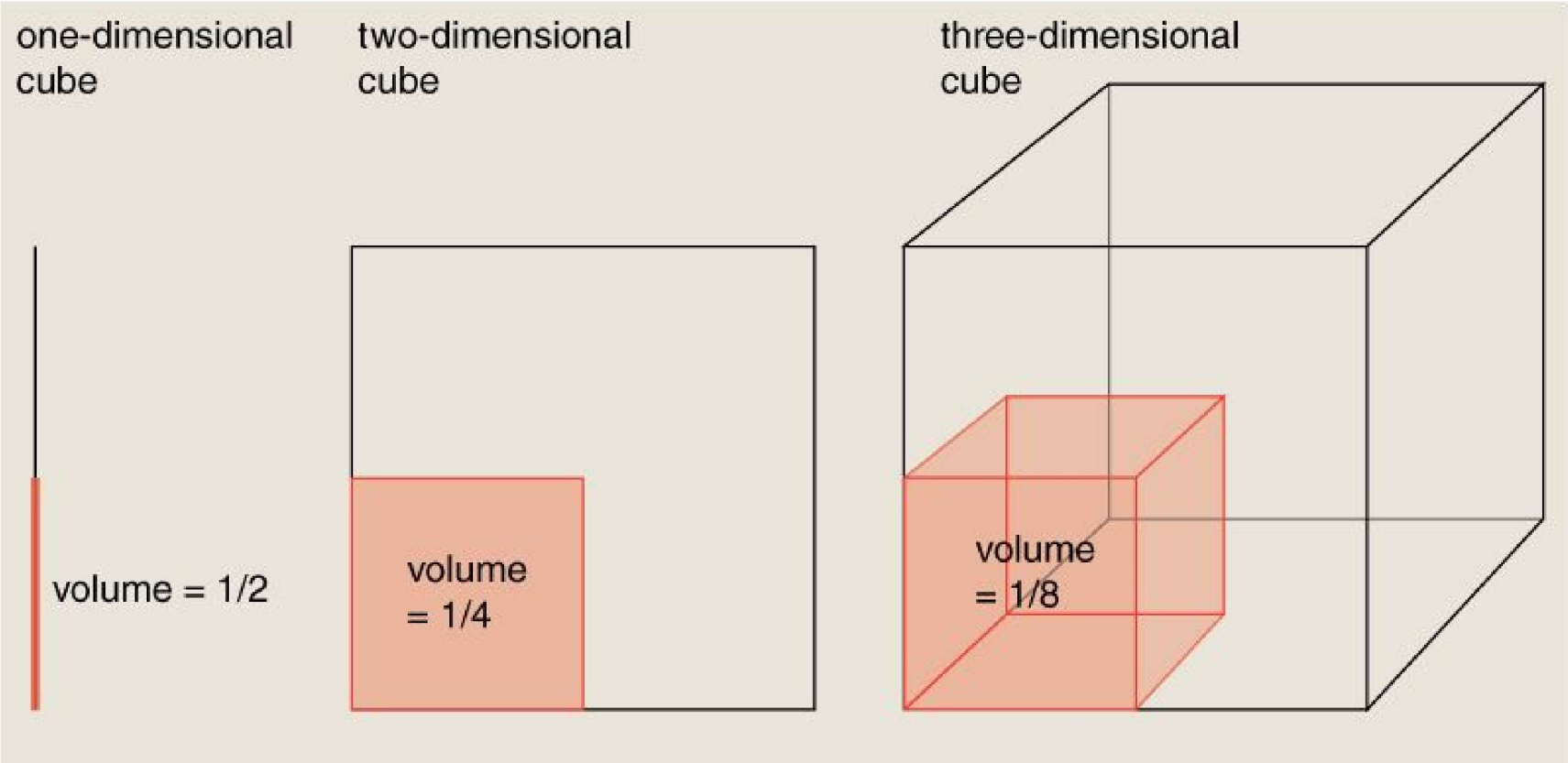
- Ambos tienen optimización convexa.
- Ambos son Word Embeddings.
- Ambos pueden usarse para representar tópicos.
- Ambos usan matrices iniciales aleatorias.
- NMF puede garantizar *sparsity*
- Con NMF se puede conocer de los tópicos.
- GloVe hace analogías.
- GloVe le da buen uso a los StopWords.
- GloVe tiene una forma estandarizada de evaluar su calidad.



Bigger Picture

- NMF extrae el **ADN** de textos cortos.
  - Distribución de palabras.
  - Permite inferir estructura entre textos y palabras.
- GloVe extrae el **ADN** de las palabras.
  - Distribución sobre factores ocultos.
  - Permite inferir estructura entre palabras y factores ocultos.
- A qué más se le puede extraer el **ADN**? Cuál es la conexión?

Maldición de dimensionalidad



- Nuestro espacio tiene dimensiones gigantes. Debemos usar **ángulos y no distancias**.

Algunas funcionalidades

- Rankear tweets de acuerdo a tópico.
- Rankear tópicos de acuerdo a palabras.
- Rankear tópicos de acuerdo a tweets.
- Hallar el promedio ponderado de algunas palabras.
- Hallar el promedio ponderado de algunos tópicos.
- **Generar un tópico según la distancia coseno a un punto en el hiperespacio.**

Guiar la creación de un tópico

- **Ex Post:**
  - Se usa el word embedding implícito.
  - Los tópicos son basados en distancia coseno a cierto punto.
  - Al promediar muchas palabras de cierto tópico, se **cancela** el ruido y queda la **esencia** del tópico.
  - No factoriza el corpus, pero si respeta las palabras que uno le dé.
- **Ex Ante:**
  - Se inicializa la matriz U de NMF con sesgo en ciertos tópicos hacia ciertas palabras.
  - Los tópicos podrían de todas formas **apartarse** de la sugerencia.
  - Factoriza el corpus, entonces la sugerencia sólo va a permanecer si es **un tópico real** en el texto.

Queda claro que se pueden diseñar tópicos con **Glove...**

También se pueden hacer word embeddings con **LDA**...

Buenas preguntas

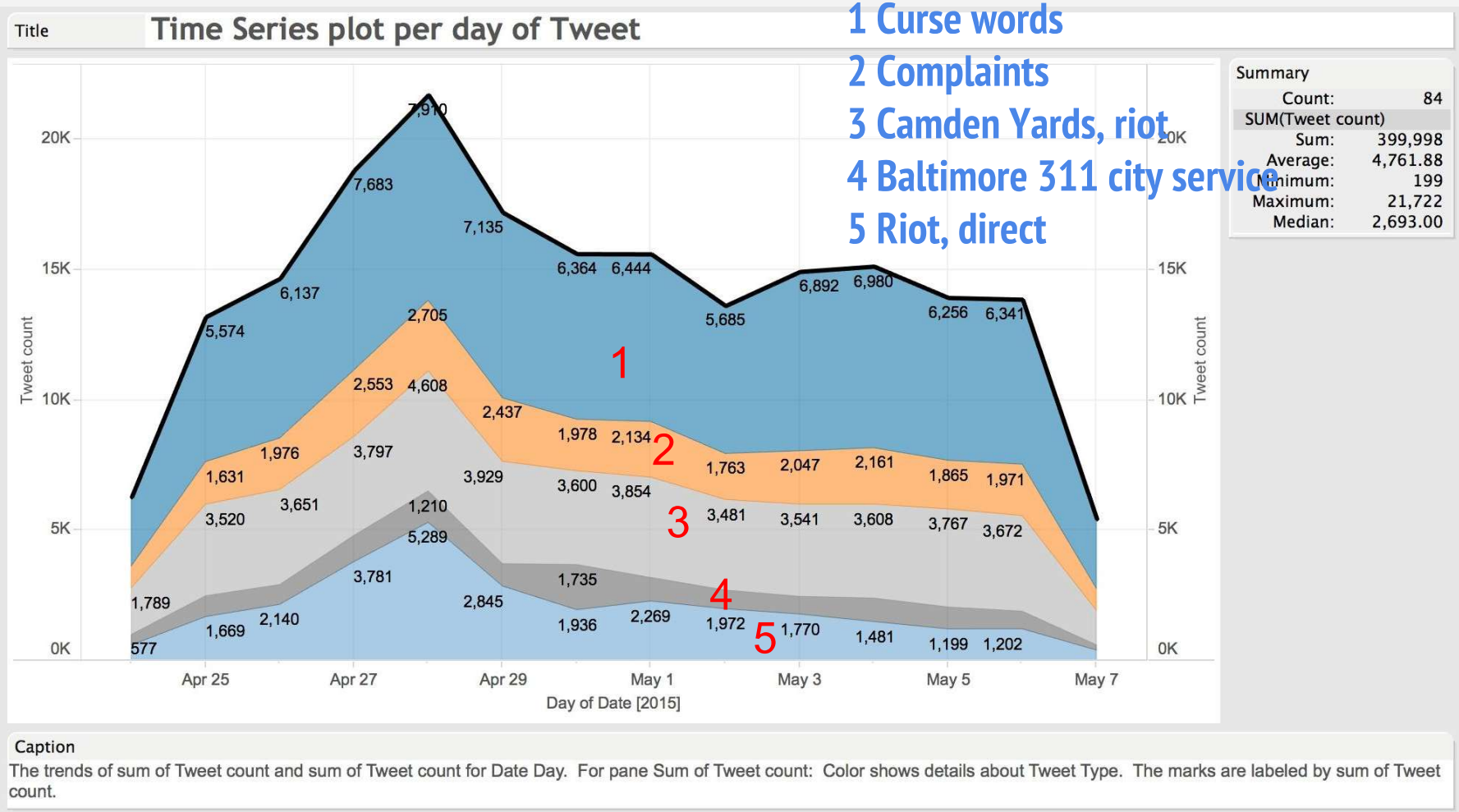
Y futura investigación en NLP teórico

- Se puede forzar a que LDA sea Sparse?
- Se pueden rankear las palabras de acuerdo a las dimensiones de GloVe?
- Se puede combinar más GloVe con NMF?
- Se puede extrapolar la idea de factorizar a Párrafos? Música? Codificación automática?
- Para hacer tópicos cuál de todas las formas da algo “mejor”.

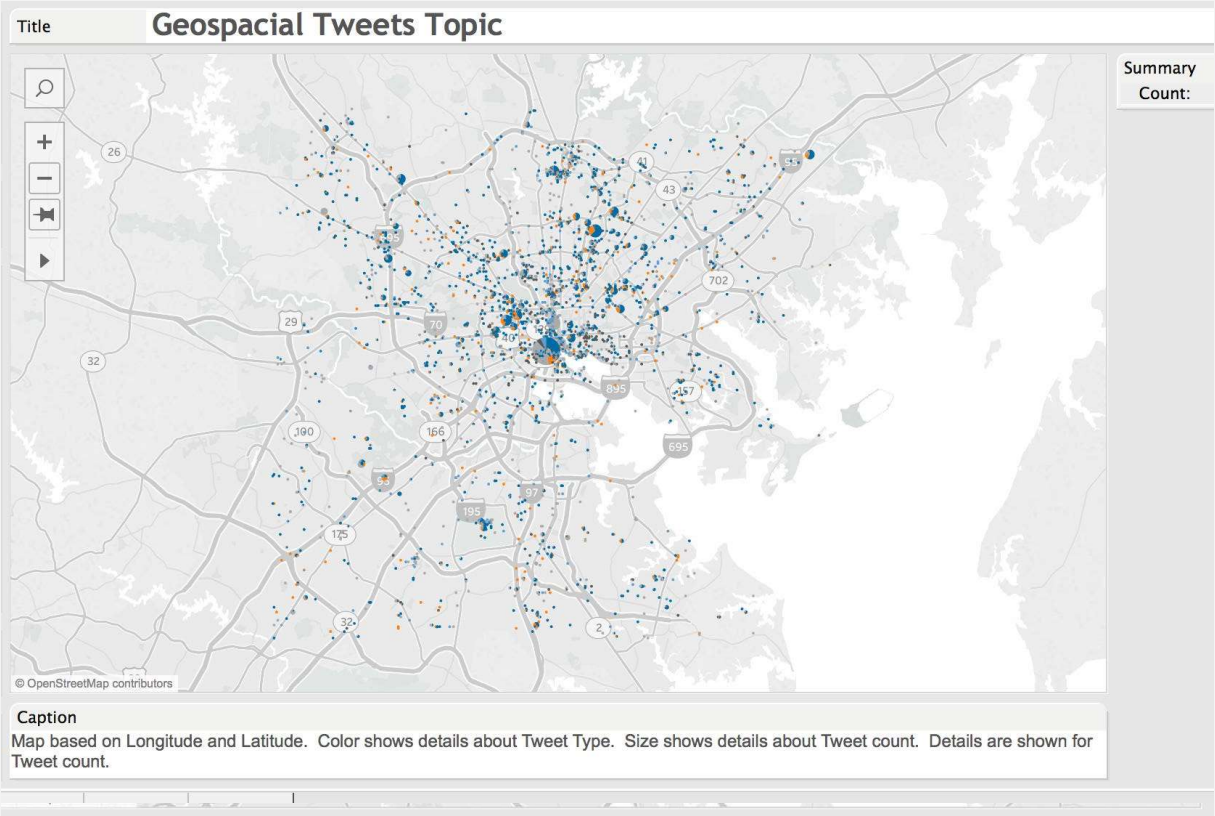
Topics based on tweets Apr 24 noon --May 7 11

1: 'like' 'feel' 'look' 'looks' 'bitch' 'act' 'hate' 'yall' 'nigga' 'fuck' 'ass' 'time' 'said' 'looking' 'girl' 'make' 'damn' 'bitches' 'acting' 'sounds' 'man' 'did' 'tf' 'aint' 'yo' 'niggas' 'girls' 'feeling' 'ppl' 'youre'
2: 'no' 'aint' 'parking' 'police' 'fuck' 'reason' 'bitch' 'complaint' 'violation' 'way' 'matter' 'justice' 'oh' 'peace' 'theres' 'hell' 'closed' 'nigga' 'say' 'idea' 'time' 'said' 'cause' 'problem' 'school' 'curfew' 'se' 'fqbegin' 'longer' 'make'
3: 'picturelink' 'md' 'park' 'nyc' 'new' 'drinking' 'orioles' 'maryland' 'freddiegray' 'baltimoreuprising' 'photo' 'post' 'camden' 'yards' 'night' 'hhtfitness' 'check' 'oriole' 'towson' 'daystodigifest' 'best' 'personaltrainer' 'sold' 'video' 'look' 'fqbegin' 'happy' 'posted' 'great' 'game'
4: 'request' 'closed' 'st' 'street' 'opened' 'alley' 'dirty' 'ave' 'trash' 'iphone' 'high' 'grass' 'weeds' 'n' 'fqbegin' 's' 'android' 'parking' 'complaint' 'close' 'light' 'w' 'debris' 'removed' 'cleaned' 'picturelink' 'graffiti' 'water' 'removal' 'rd'
5: 'baltimore' 'city' 'md' 'police' 'freddiegray' 'hall' 'maryland' 'curfew' 'trending' 'baltimoreriots' 'bwi' 'baltimoreuprising' 'android' 'washington' 'airport' 'international' 'marshall' 'thurgood' 'live' 'riots' 'mayor' 'protest' 'north' 'west' 'national' 'guard' 'harbor' 'peace' 'freddie' 'today'

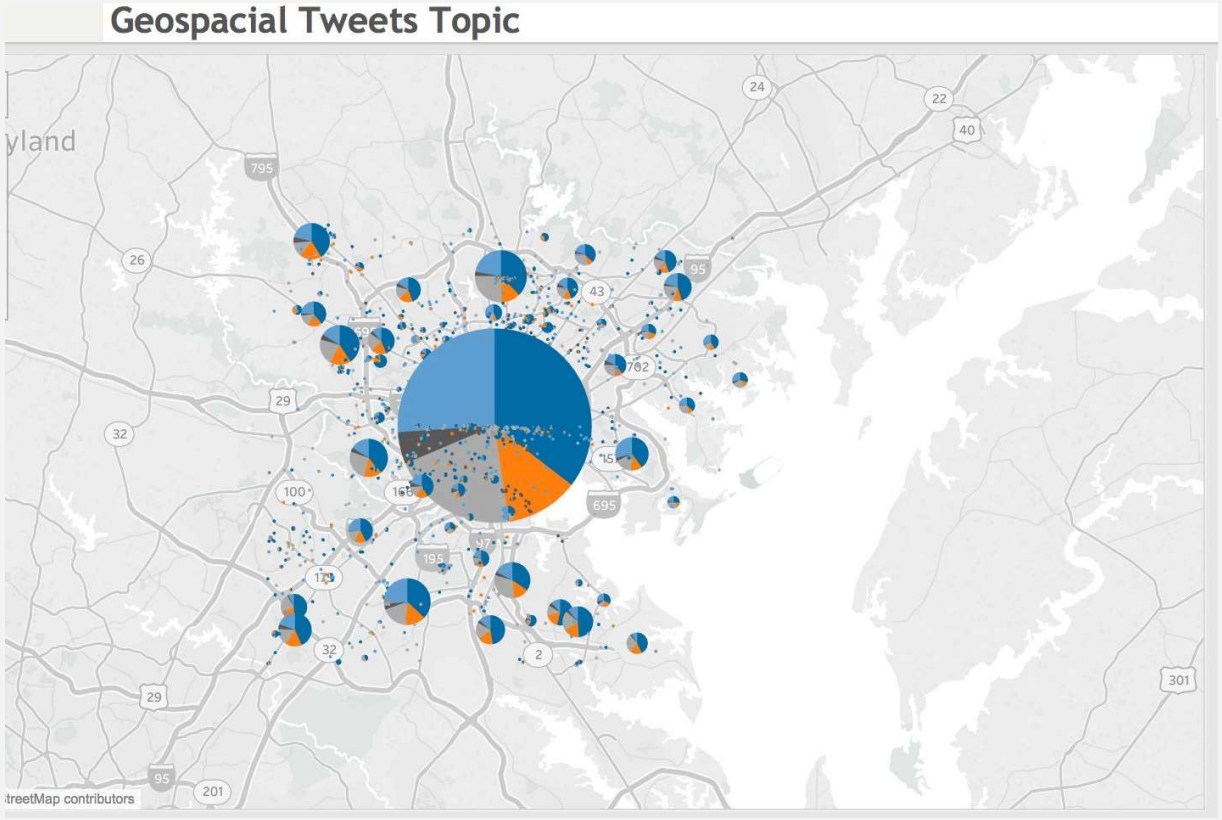
Topic Analyzing-Time Series



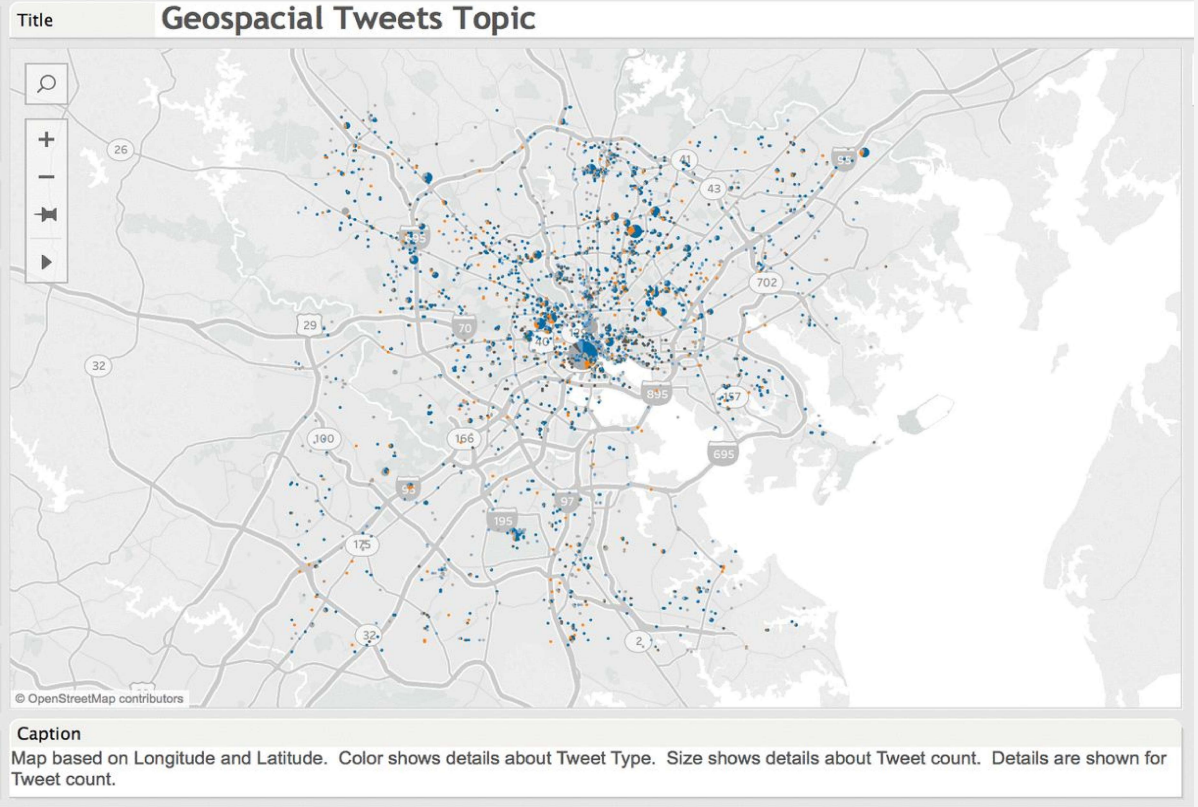
Topic Analysing-Geospatial



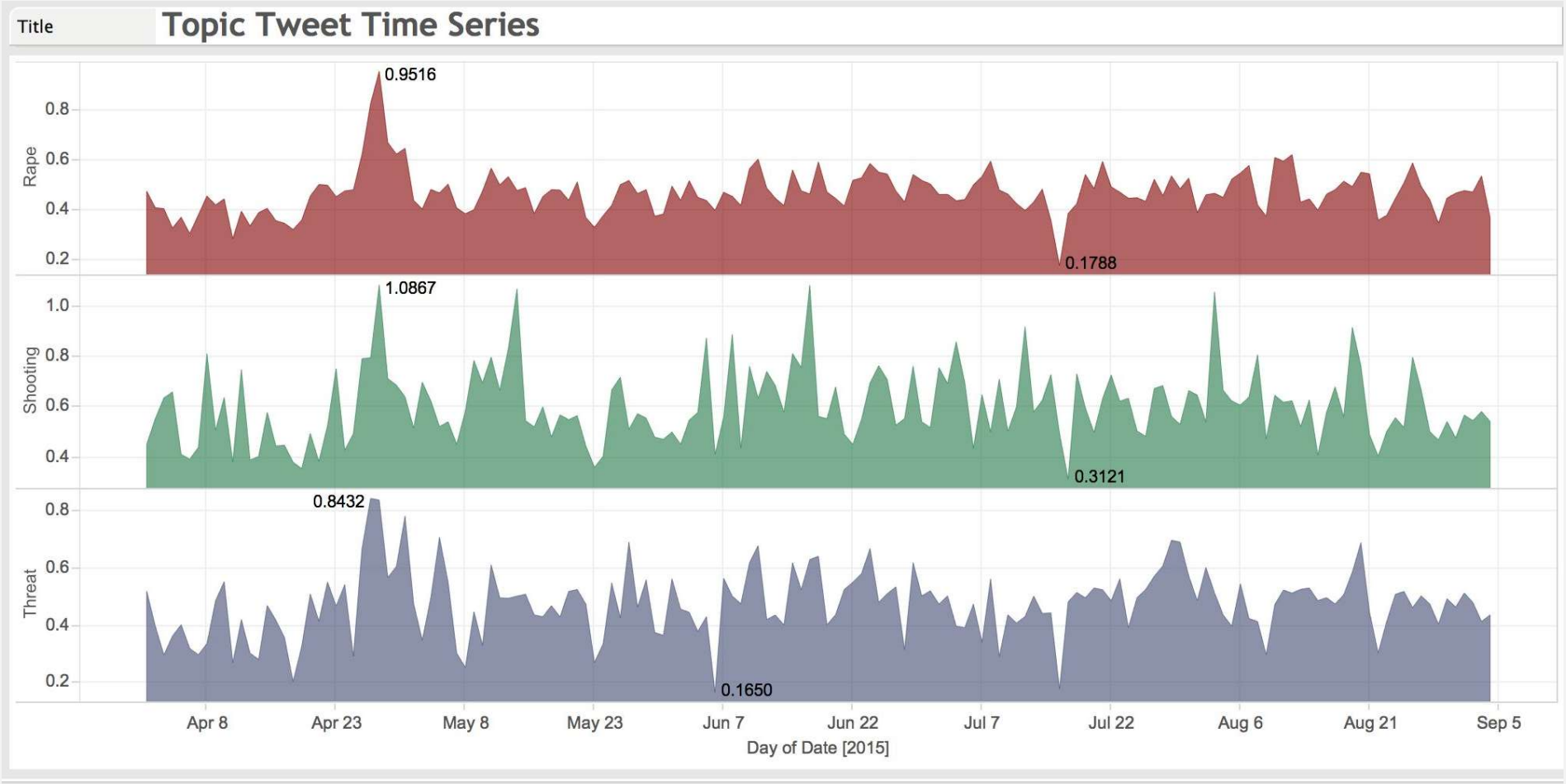
Topic Analysing-Geospatial



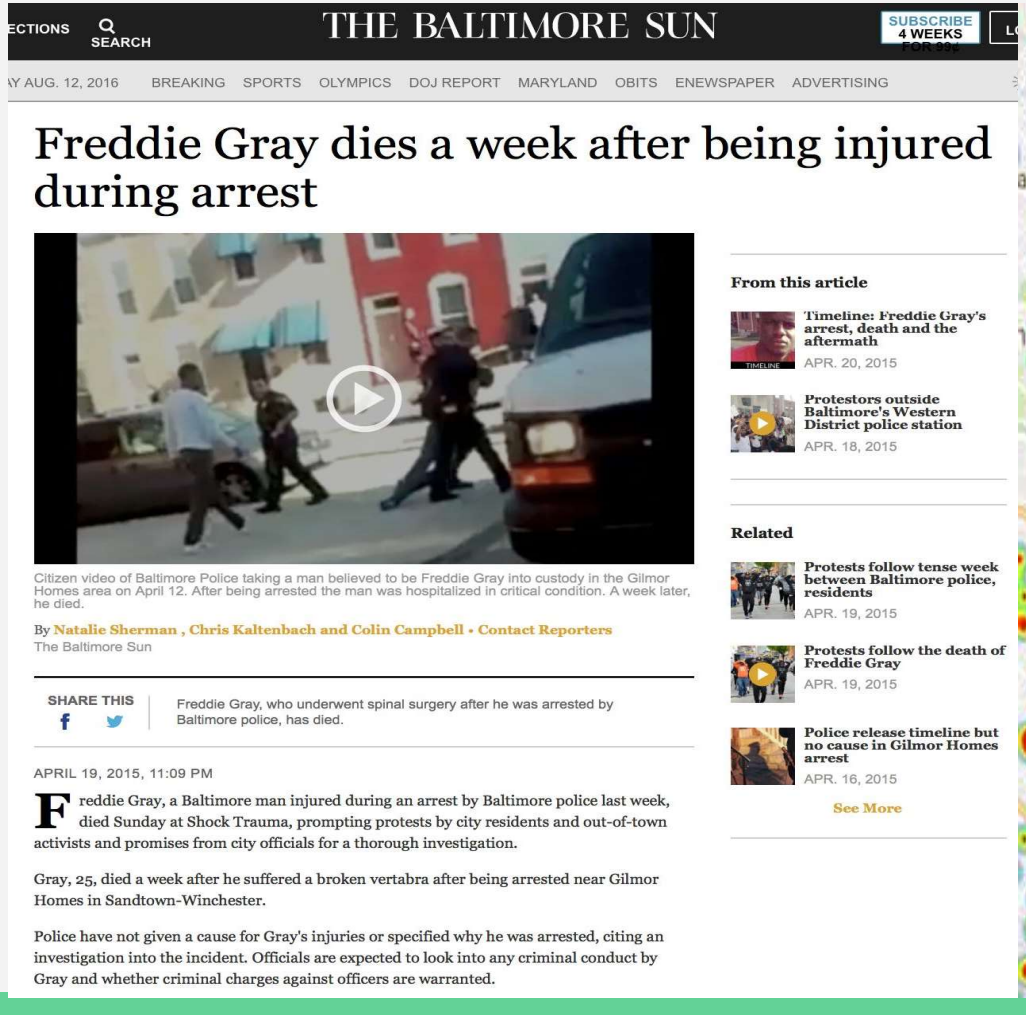
Topic Analysing-Geospatial



Topic Time Series Analysis



News Articles to Tweets



ECTIONS SEARCH THE BALTIMORE SUN SUBSCRIBE 4 WEEKS FOR 99¢

AY AUG. 12, 2016 BREAKING SPORTS OLYMPICS DOJ REPORT MARYLAND OBITS ENEWSPAPER ADVERTISING

Freddie Gray dies a week after being injured during arrest

Citizen video of Baltimore Police taking a man believed to be Freddie Gray into custody in the Gilmor Homes area on April 12. After being arrested the man was hospitalized in critical condition. A week later, he died.

By Natalie Sherman , Chris Kaltenbach and Colin Campbell • Contact Reporters
The Baltimore Sun

SHARE THIS

Freddie Gray, who underwent spinal surgery after he was arrested by Baltimore police, has died.

APRIL 19, 2015, 11:09 PM

Freddie Gray, a Baltimore man injured during an arrest by Baltimore police last week, died Sunday at Shock Trauma, prompting protests by city residents and out-of-town activists and promises from city officials for a thorough investigation.

Gray, 25, died a week after he suffered a broken vertabra after being arrested near Gilmor Homes in Sandtown-Winchester.

Police have not given a cause for Gray's injuries or specified why he was arrested, citing an investigation into the incident. Officials are expected to look into any criminal conduct by Gray and whether criminal charges against officers are warranted.

From this article

Timeline: Freddie Gray's arrest, death and the aftermath
APR. 20, 2015

Protestors outside Baltimore's Western District police station
APR. 18, 2015

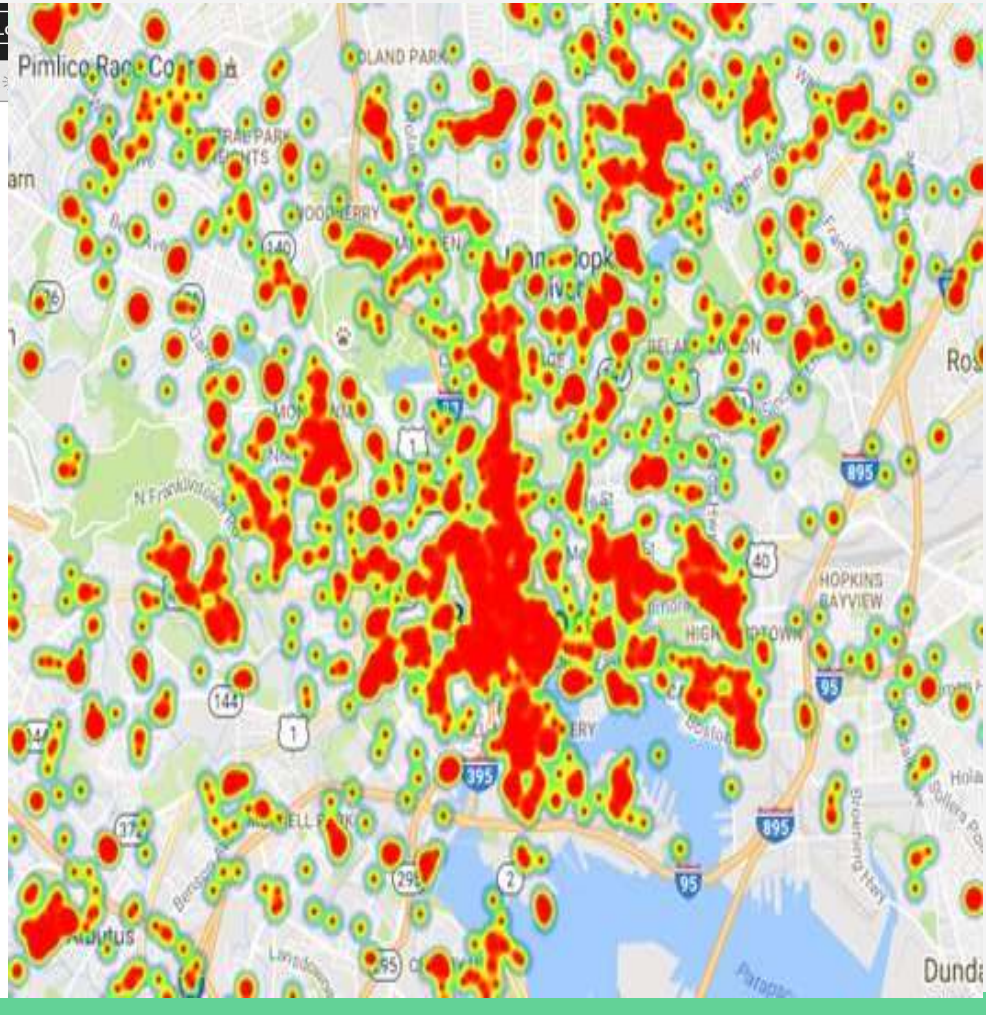
Related

Protests follow tense week between Baltimore police, residents
APR. 19, 2015

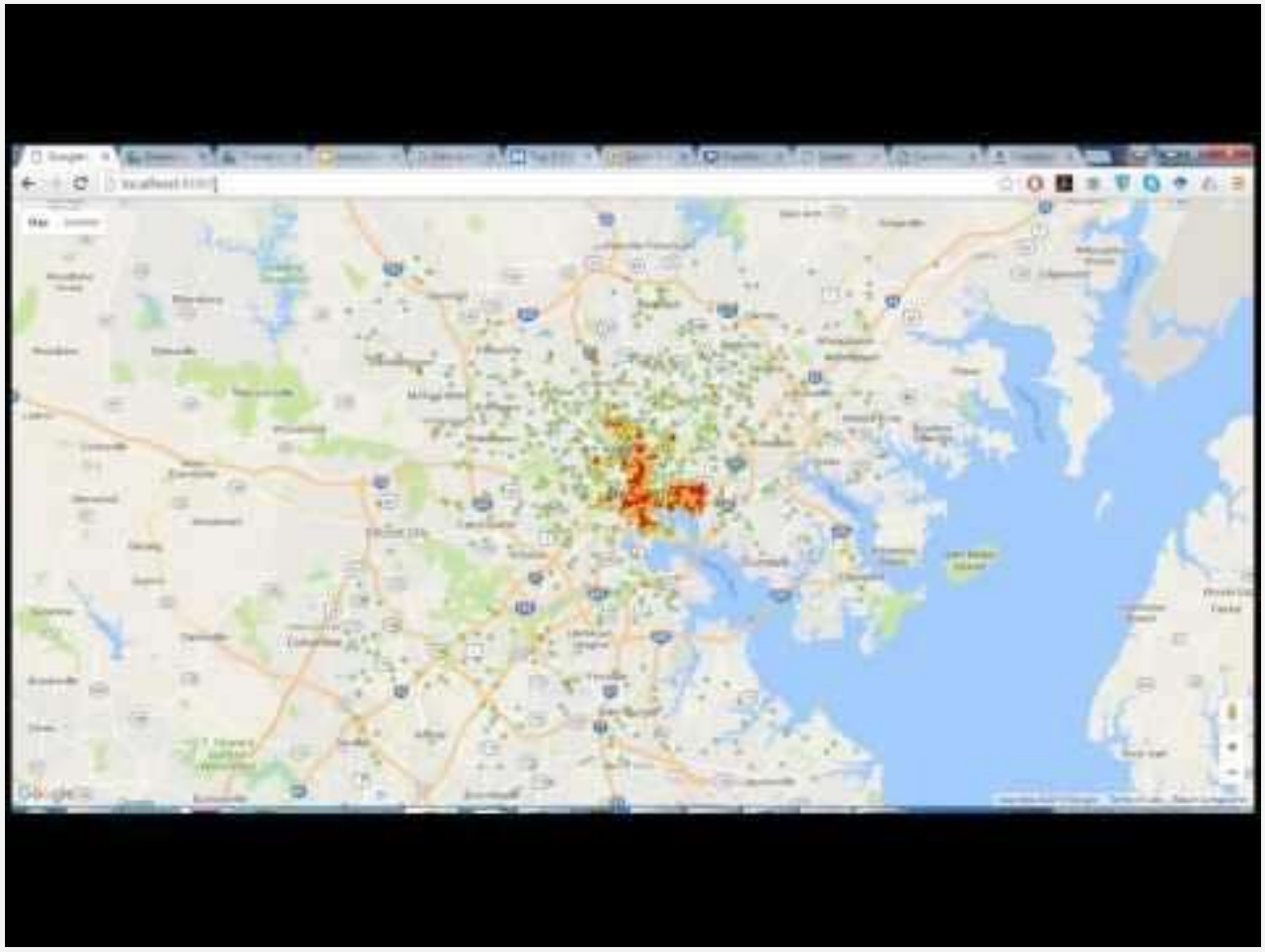
Protests follow the death of Freddie Gray
APR. 19, 2015

Police release timeline but no cause in Gilmor Homes arrest
APR. 16, 2015

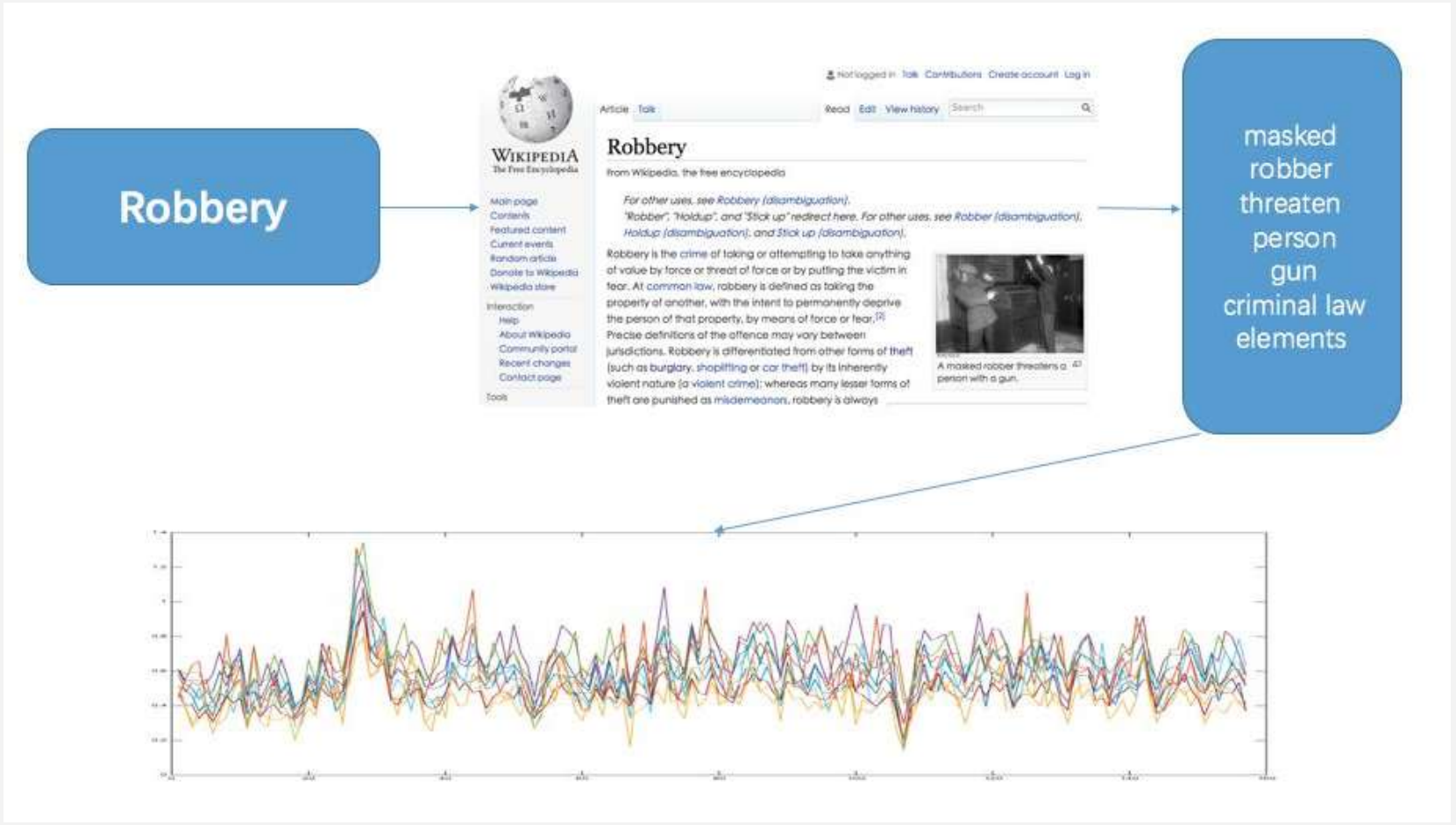
See More



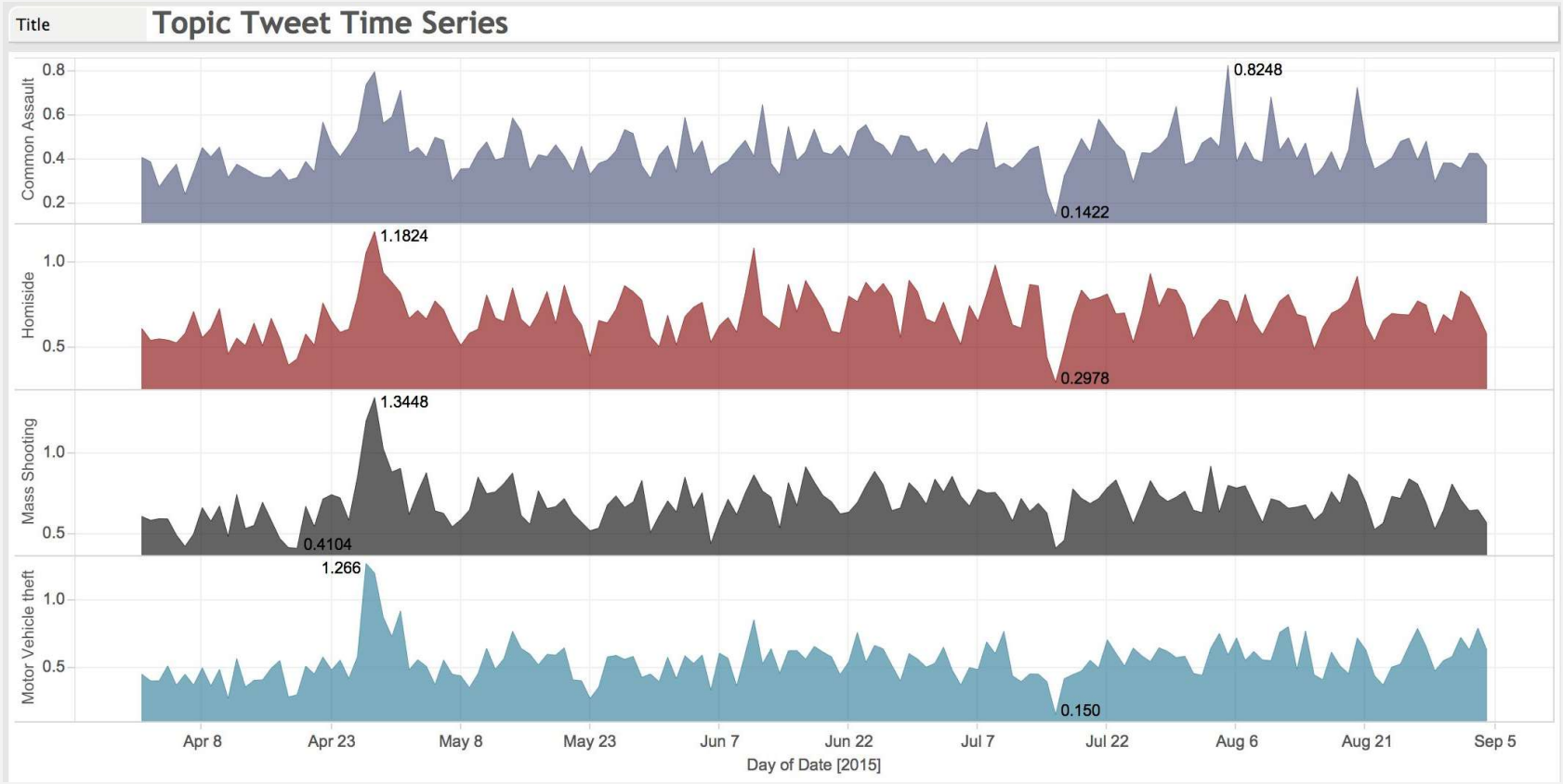
News Articles Topic based tweets heatmaps



Uncurated to curated



Learned Crime Topics



Spatio-temporal Agression Meter

- Crear tópicos generados a partir de diccionarios de emociones de WordNet.
- Crear “Polos” que sean dos tópicos (puntos en el espacio).
- Proyectar cada tweet a la linea que une esos dos tweets (O normalizar las distancias coseno).
- Graficar un heatmap que diga que tan Frío o Caliente está cada tweet.
- ¿UIAF?

What could we do in the future

Agression spatio-temporal meter.

Event detection

Emerging topic detection

Predict popular opinions and trends

Network analysis for users and topics