Text analytics
and
new visualization techniques

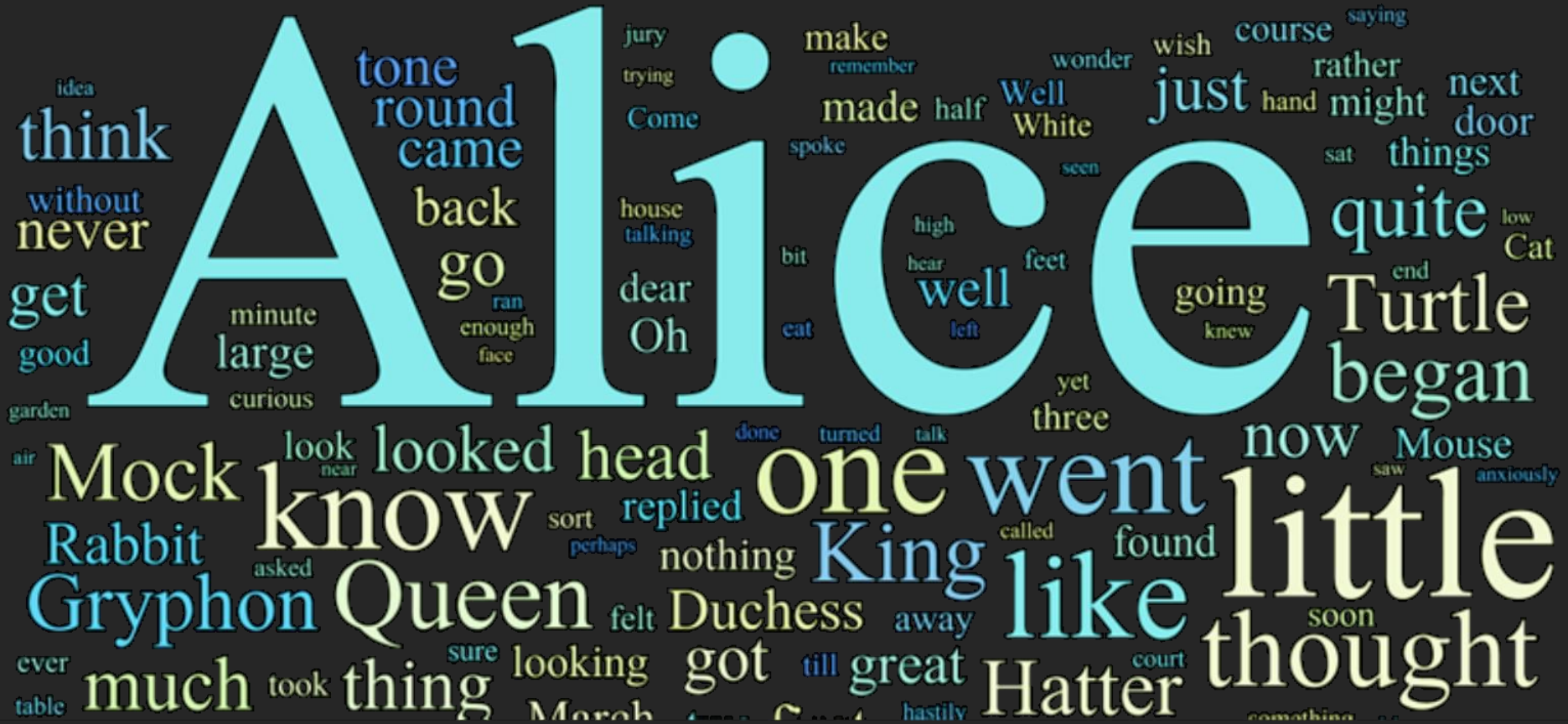
Richard Brath
rbrath@unchartedsoftware.com
brathr@lsbu.ac.uk
@rkbrath
richardbrath.wordpress.com

Scott Langevin
slangevin@unchartedsoftware.com
blog.unchartedsoftware.com
@slangevi

© 2017 Uncharted Software Inc.

uncharted

Most Common Text Visualization



Tag cloud via wordle.com. Note color is random, location is random and size is ambiguous – does Alice occur 10x or 100x more than Queen?

© 2017 Uncharted Software Inc.

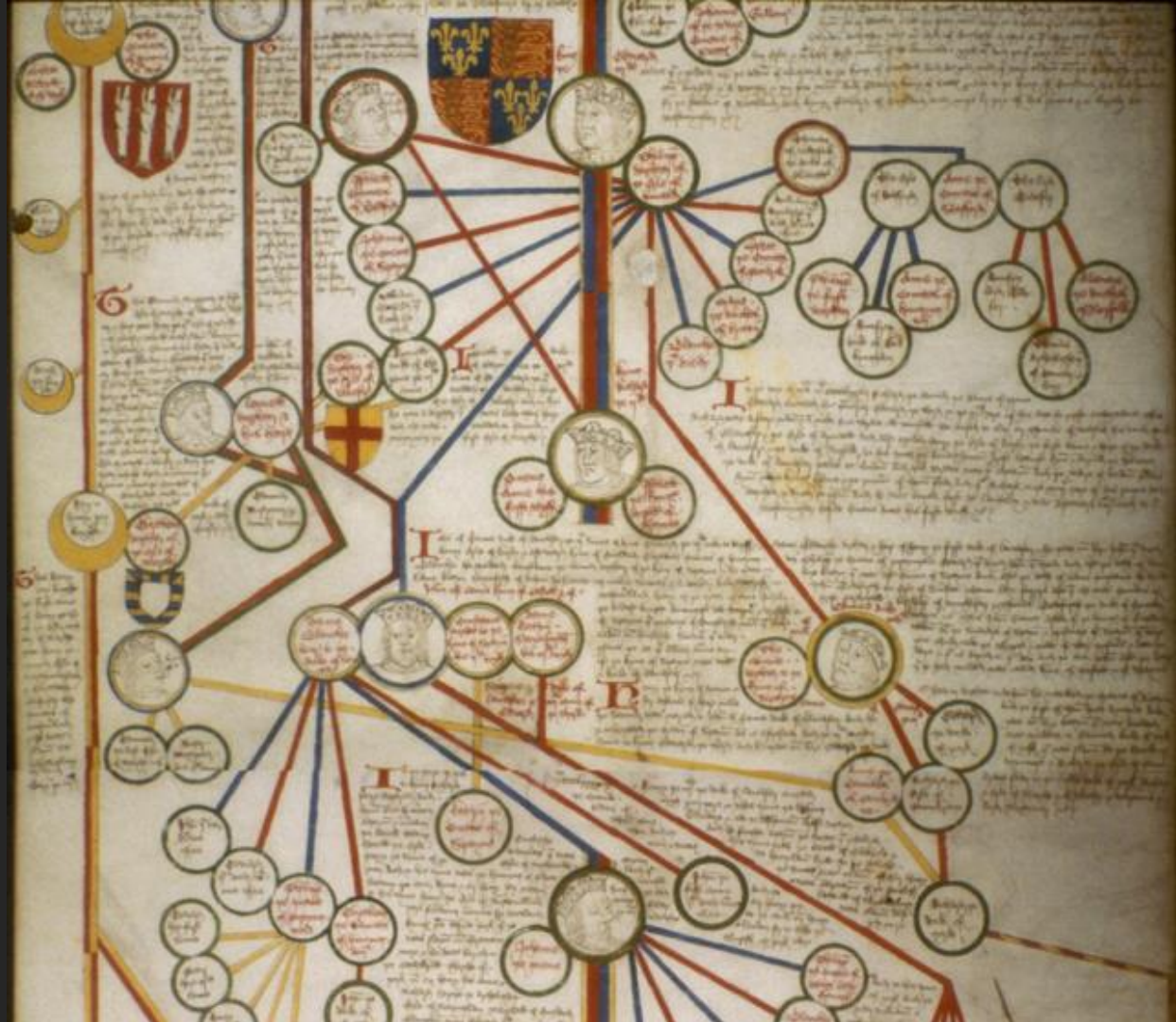
uncharted

1. THE 500 YEAR CONSPIRACY

against text visualization

© 2017 Uncharted Software Inc.

Medieval text and visualization



Genealogical tree from late 1400's. Note graph nodes use of image (people, shield) or text, where text may be black, red or start with a red initial. The nodes can vary in size, color, or shape (circle, crescent, shield). Textual commentary is intertwined throughout.

via Bodleian Library. University of Oxford., Genealogy of the kings of England to Richard III. Chronicle of the Percy family to 1485. MS. Bodl. Rolls 5. view 35.

© 2017 Uncharted Software Inc.

1439
Johannes Gutenberg changes everything

1573: Image separates from text.
from William Bullein's A Dialogue… Against the Fever Pestilence. Author photo from Bodleian exhibition "Shakespeare's Dead".

© 2017 Uncharted Software Inc.

1751
Enlightenment complete separation

Diderot's Encyclopedia has great illustrations of various occupations – all neatly labeled, but the viewer has to cross-reference the text to understand.

The *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers, par une Société de Gens de lettres* was published under the direction of Diderot and d'Alembert, with 17 volumes of text and 11 volumes of plates between 1751 and 1772. Containing 74,000 articles written by more than 130 contributors, the *Encyclopédie* was a massive reference work for the arts and sciences, as well as a *machine de guerre* which served to propagate the ideas of the French Enlightenment. The impact of the *Encyclopédie* was enormous. Through its attempt to classify learning and to open all domains of human activity to its readers, the *Encyclopédie* gave expression to many of the most important intellectual and social developments of its time.

Encyclopédie, ou dictionnaire raisonné des sciences, des arts et des métiers, etc., eds. Denis Diderot and Jean le Rond d'Alembert. University of Chicago: ARTFL Encyclopédie Project (Spring 2016 Edition), Robert Morrissey and Glenn Roe (eds), http://encyclopedie.uchicago.edu/.

© 2017 Uncharted Software Inc.

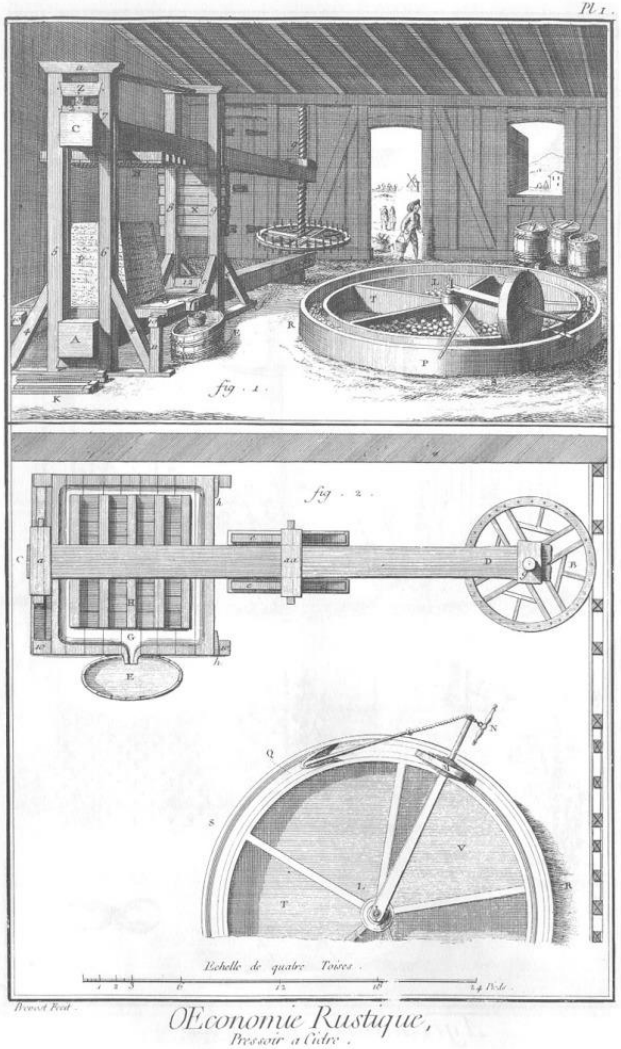


PLANCHE III.

Fig. 1. Plan & profil de l'un des coffres du preſſoir. P P, chantier. *r r*, brebis. *y*, doſſier. *q*, le mulet. *y, y, y, y, y*, pieces de maie. Z, coins. D, mouton. E E, coins ou pouſſeculs. *u u*, écrou. C D, vis. A B, grande roue.

2. Coupe ſuivant la longueur d'un des coffres du preſſoir. L L, faux chantier. 13, 13, jumelles. *t s*, chaines. *y*, doſſier. *r r*, brebis. q, le mulet. 1, 2, 3, 4, 5, 6, 7, 8, 9, pieces de maie. Z, coins. D, mouton. E E, coins ou pouſſeculs. *p, p, p*, appuis du doſſier. 10, mouleau. G G, planches à couteaux. *x, x, x, x*, chevrons. *u u*, écrou. C D, vis. A B, grande roue. M, bouquets ou piés-deſtaux de pierre. F, le marc.

PRESSOIR A CIDRE.

PLANCHE Ire.

Fig. 1. Une perſpective & plan du preſſoir. La vignette repréſente l'endroit où le preſſoir & la pile ſont établis. A B, la brebis. C D, le mouton. 5, 6, 7, 8, 9, les jumelles. 4, 4, *e, e*, contrevens. Y, Z, 2, 12, entretoiſes. *a, b*, chapeau. K X, les clés. *g*, la vis. E, le barlong. F, marc empilé ſur la maie ou l'emoy. G, la maie ou l'emoy. 10, 10, pieces qui ſupportent les pieces de maie. 11, pieces qui ſoûtiennent les couches. H, le bec. R S Q, auge circulaire de la pile. Q, le rabot. T L V, caſes ou ſéparations à différentes ſortes de pommes. M, la meule. L N, axe de la meule. N, palonnier. V P, conducteur du cheval ou guide.

2. Plan du preſſoir & de la pile. Les mêmes lettres déſignent les mêmes parties.

PLANCHE II.

Fig. 3. Profil & détail du preſſoir à cidre. Élévation géométrale du preſſoir vû de face. Les mêmes lettres déſignent auſſi les mêmes parties.

4. Élévation des jumelles qui embraſſent le gros bout du mouton & de la brebis.

5. Élévation des deux jumelles qui ſont placées vers le milieu du mouton & de la brebis, & qui ſervent à relever le mouton.

6. Partie inférieure de la vis qui entre dans la brebis.

7. Plan & profil d'une des clés.

INDIGOTERIE ET MANIOC.

Le haut de la Planche ou la vignette repréſente la vûe d'une indigoterie. A, reſervoir d'eau claire. B, la trempoire. C, la batterie. D, le repoſoir qu'on nomme auſſi diabloin. E E, robinets d'où la teinture d'une cuve paſſe dans la cuve qui eſt au-deſſous. E F, trous que l'on débouche ſucceſſivement, pour vuider l'eau claire de la batterie, lorſque la fécule bleue s'eſt précipitée au fond. G, indigot dont on a rempli des ſacs de toile en forme de chauſſes pour le faire égoutter. H, hangard ouvert & à claire voie ſous lequel on met l'indigot dans des caiſſons, pour achever de le faire ſécher à l'ombre. I, negre qui porte la plante dans la trempoire. K K, negres qui agitent continuellement la teinture de la batterie avec des ſeaux percés & attachés à de longues perches. L, plantes d'indigot. M, maiſon du maître de l'habitation. N, campagne ſemée d'indigot.

Fig. 1. o, o, caiſſons de bois élevés ſur des treteaux, ſervans à faire ſécher l'indigot à l'ombre ſous le hangard de la vignette.

2. P, couteau courbé en forme de ſerpette, pour couper l'indigot ſur pié.

3. Q, taſſe d'argent bien polie, ſervant à examiner la formation du grain dans la teinture de la batterie.

4. Preſſe à manioc. A, tronc d'arbre percé en-travers. B, branche fourchue diſpoſée en bras de levier & chargée de groſſes pierres. C, ſacs d'écorce d'arbre remplis de la rapure du manioc. D, bouts de planche ſervant à preſſer les ſacs également. E, couy ou coupe de calebaſſe recevant le ſuc du manioc dont on fait la mouchoche.

Fig. 5. Maniere d'exprimer le ſuc du manioc à la façon des Caraïbes. G, couleuvre ou eſpece de panier d'un tiſſu lâche & flexible, rempli de rapure de manioc. H, poids attaché au bas de la couleuvre qui la contraint de s'allonger en diminuant ſa groſſeur; ce qui ſuffit pour exprimer le ſuc de la rapure.

SUCRERIE ET AFFINAGE des Sucres.

PLANCHE Ire.

La vignette repréſente la vûe d'une habitation. 1, maiſon du maître & ſes dépendances. 2, 2, 2, partie des caſes à negres formant une ou pluſieurs rues, ſuivant le nombre & l'emplacement. 3, 3, 3, partie de ſavanne ou pâturage. 4, 4, liſiere ou forte haie qui ſépare la ſavanne des plantations de cannes. 5, 5, 5, partie de pieces plantées en cannes à ſucre à mi-côte & en plat-pays. 6, moulin à eau. 7, ſucrerie avec ſa cheminée, & ſon hangard pour les fourneaux. 8, gouttiere qui conduit l'eau du canal ſur la roue du moulin. 9, décharge de l'eau du moulin. 10, une des caſes à bagaſſe ou cannes écraſées. 11, purgerie ou grand magaſin ſervant à mettre les ſucres quand ils ſont en forme, pour les purger de leur ſyrop ſuperflu & les terrer. 12, étuve pour faire ſécher les pains de ſucre. 13, hauteurs entre leſquels ſont les plantations de manioc, les bananiers & l'habitation à vivre. 14, morne: c'eſt ainſi qu'on nomme aux îles Antilles les montagnes qui paroiſſent détachées des autres.

Fig. 1. Coupe verticale d'une étuve à mettre ſécher les pains de ſucre terrés. A, comble de l'étuve. B, murs de l'étuve. C, porte. D, coffre de fer ſervant de fourneau. E, bouches du foyer & du cendrier. F, rayons ou tablettes en grillage, ſur leſquelles on range les pains de ſucre. G, plancher couvert de cinq à ſix pouces de maçonnerie. H, trape que l'on ouvre pour laiſſer aller l'humidité qui s'éleve des pains de ſucre, & qui s'échappe au-dehors par les conduits i, i, pratiqués ſous le larmier.

2. K, canne à ſucre. L, feuille dentelée ſur les bords. M, fleche ou fleur de la canne portant la graine. N, partie inférieure de la canne avec ſa racine.

3. O, ſerpe pour ſarcler & couper les cannes.

4. P, houe à fouiller la terre.

5. Q, pelle de fer pour le même uſage & ramaſſer le ſucre pilé dans le canot.

6. R, pince de fer ſervant de levier.

7. S, canot avec ſes pilons, pour mêler le ſucre en poudre, & le fouler dans les futailles.

PLANCHE II.

Deux moulins, dont un à eau.

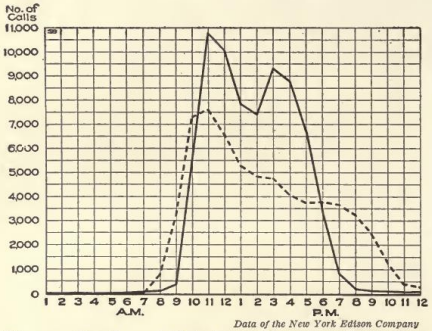
Fig. 1. Moulin mû par des animaux. A A, chaſſis de charpente très-ſolide. B B, table du moulin, communément faite d'un ſeul bloc creuſé & revêtu de plomb. C, C, C, trois roles couverts chacun d'un tambour ou cylindre de métal, & traverſés d'un axe de fer coulé, dont l'extrémité inférieure eſt garnie d'un pivot portant ſur une crapaudine. D, D, D, D, ouvertures faites à la table pour pouvoir changer & réparer les pivots & les crapaudines. E, E, entailles aux deux ouvertures des côtés, ſervant à chaſſer des coins de bois, pour ſerrer & rapprocher les tambours. F, F, autres ouvertures ſur les moiſes, avec des coins pour ſerrer les pivots ſupérieurs. G G, hériſſons dont les roles ſont couronnés, & qui engrenant les uns dans les autres, font tourner les tambours en ſens contraires. H, axe ou arbre prolongé du principal role. I, demoiſelle,

1920 Same for charts

1919 book explaining charts. Labels and notes are pushed to the periphery of the chart. (Willard C. Brinton, Graphic Methods for Presenting Facts, 1919)

© 2017 Uncharted Software Inc.

more than would be required for the tabulated figures, and if there is any serious limitation on space, Fig. 97 could be much reduced in size without detracting from its clearness.



Data of the New York Edison Company

Fig. 97. Number of Telephone Messages Each Hour for One Day in a Business District in New York City as Compared with a Residence District

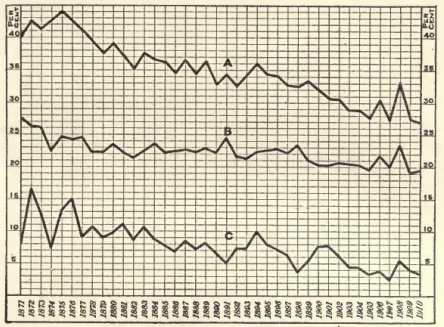
Solid line, the "Broad" exchange—a typical business exchange
Dotted line, the "Riverside" exchange—a typical residence exchange
Note the great number of business calls after mail has been opened in the morning and after return from lunch. In the residence district there is much morning shopping by telephone

Health-department reports are not usually of interest to the layman. Yet health-department reports, well presented, may be of as much popular interest as a well-written magazine article. Fig. 98 is taken from a commendable report by the health department of the city of Boston. In the report itself, different colors of ink were used for the various curves, with the effect of emphasizing the contrast.

Though the colored inks assisted in catching the eye of the reader, the chart with curves designated by letters is usually sufficient for all practical purposes whenever the curves do not cross over each other in such manner as to be confusing. As mentioned elsewhere, a certain slope of a curve plotted on rectangular co-ordinate paper does not in itself indicate a greater or less amount of increase or decrease than holds true for some other curve having a different slope. The slope of a curve plotted on paper with ordinary co-ordinate ruling depends largely on whether the data of the curve are in large figures, so as to bring the curve near the top of the chart, or in small figures, bringing the curve near the bottom of the chart. In Fig. 98 the fact that curve A slopes more than curve C is due to the fact that curve A is placed higher on the vertical scale of the chart than curve C. A little study will show that the reduction in mortality portrayed by curve A is much less on a percentage basis than that depicted by curve C, yet curve A has the steeper slope. The slope of these two

curves can best be compared if a pencil line is drawn in such manner that the peaks above the pencil line are approximately equal to the valleys below the pencil line for each curve.

Fig. 99 contains some interesting information. Though the chart proves fairly well the close dependence of the price of cast-iron pipe upon the price of the pig iron from which it is made, the chart is nevertheless misleading in that the first glance would indicate a much greater fluctuation in the price of pig iron than actually occurred.

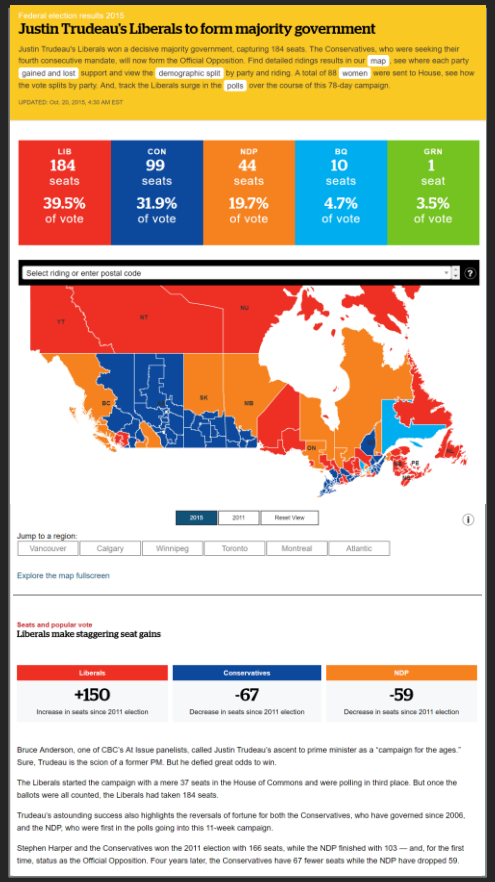


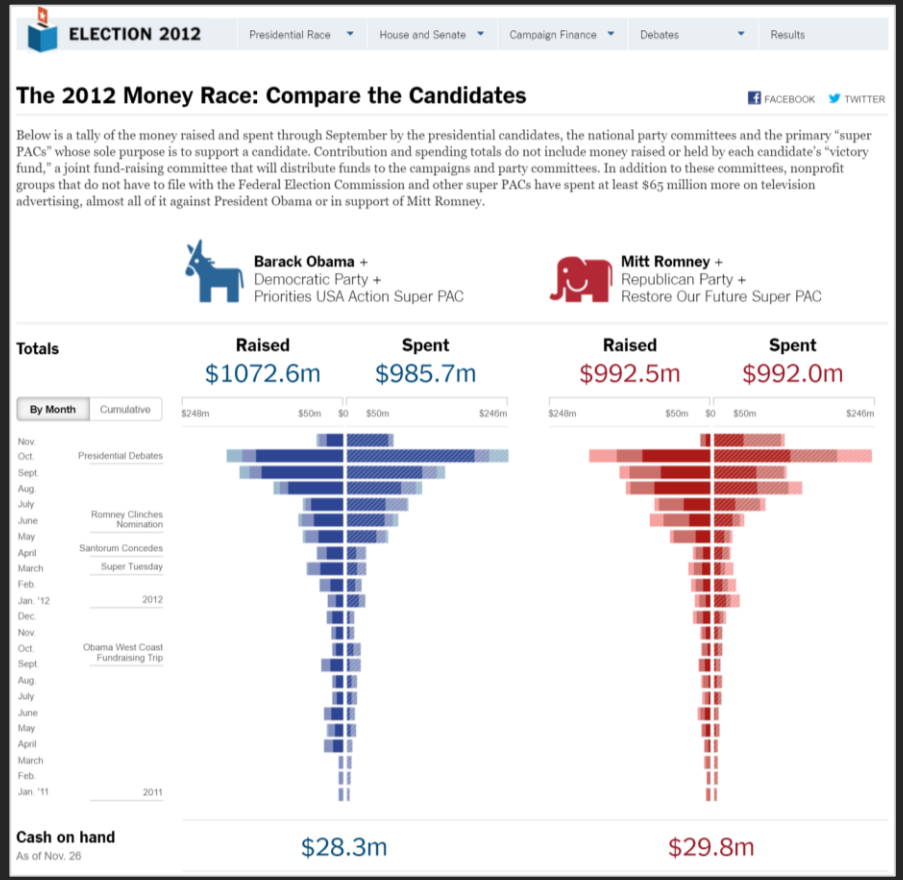
From the 1910 Annual Report of the Health Department, City of Boston

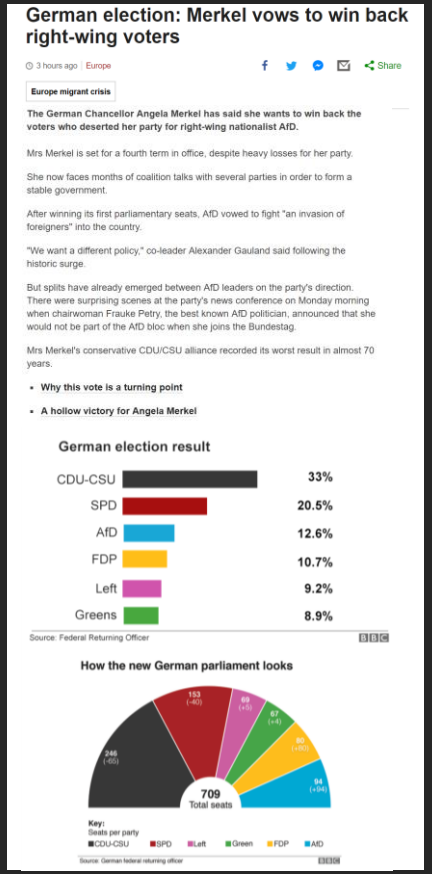
Fig. 98. Deaths in Boston of Children under Five Years of Age, under One Year, and from Five of the Principal Infectious Diseases, Expressed as a Percentage of the Total Mortality

Curve A. Deaths of children under five years of age as a percentage of the total mortality
Curve B. Deaths of children under one year as a percentage of the total mortality
Curve C. Deaths from Diphtheria, Scarlet Fever, Measles, Typhoid Fever and Smallpox as a percentage of the total mortality
Note that Curve A shows a much steeper slope than Curve C, yet Curve C drops in 1910 to less than half the figure for 1871. Curves plotted by rectangular co-ordinates should not be compared by the slope of the different curve lines

The reader is apt to overlook the fact that the vertical scale of the chart does not extend below $11 per ton. He is quite likely to think that the price of pig iron had all the rapid fluctuations which would be indicated by the changing vertical distances between the pig-iron curve and the bottom line of the chart itself. The amount of fluctuation would look much less if the chart extended to the zero line of the vertical scale.







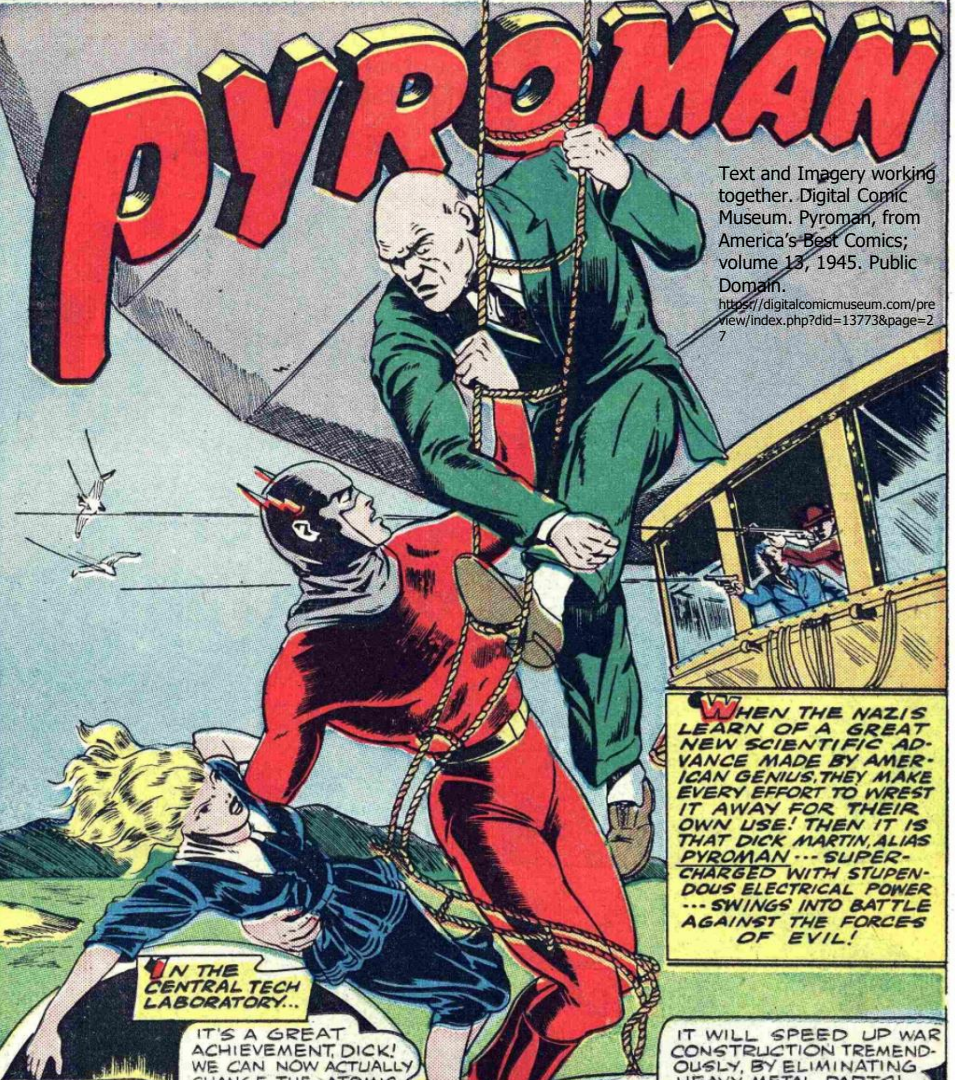
Modern visualizations in news. Note how there is a lot of text – but moved out and round the charts.

2015. --, Canada Votes: Justin Trudeau's Liberals to form majority government. CBC.ca. Oct. 15, 2015. Accessed 2017/09/25

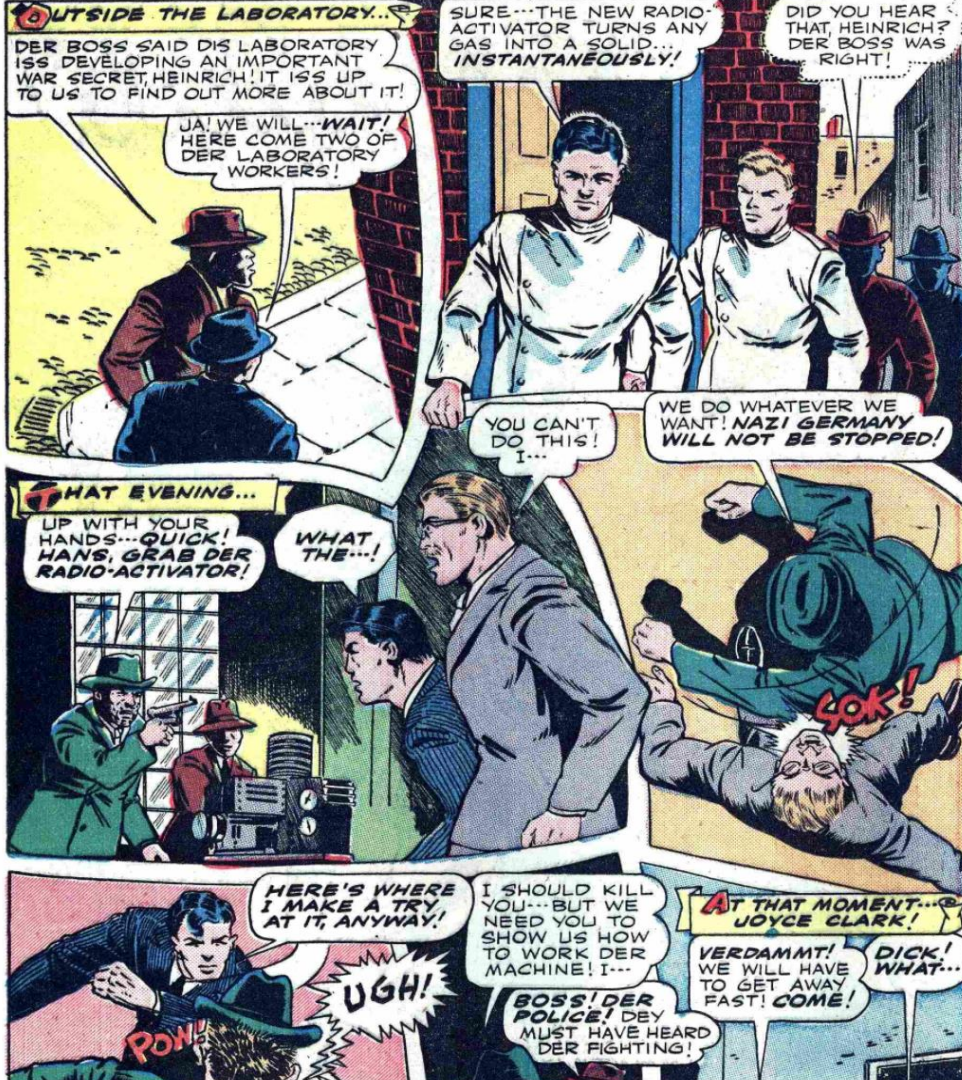
2012. Jeremy Ashkenas, Matthew Ericson, Alicia Parlapiano and Derek Willis. **The 2012 Money Race: Compare the Candidates**. The New York Times. Accessed 2017/07/24

2017. --, German Election: Merkel vows to win back right-wing voters. BBC.com. Sep. 25, 2017. Accessed 2017/09/25.

© 2017 Uncharted Software Inc.
uncharted



Text and Imagery working together. Digital Comic Museum. Pyroman, from America's Best Comics; volume 13, 1945. Public Domain. https://digitalcomicmuseum.com/preview/index.php?did=13773&page=27



2. Bringing **TEXT** back into **VISUALIZATION**

© 2017 Uncharted Software Inc.

Plain Dots

- Simple scatterplot
- Macro pattern e.g. crescent shape
- Also color and size set to population

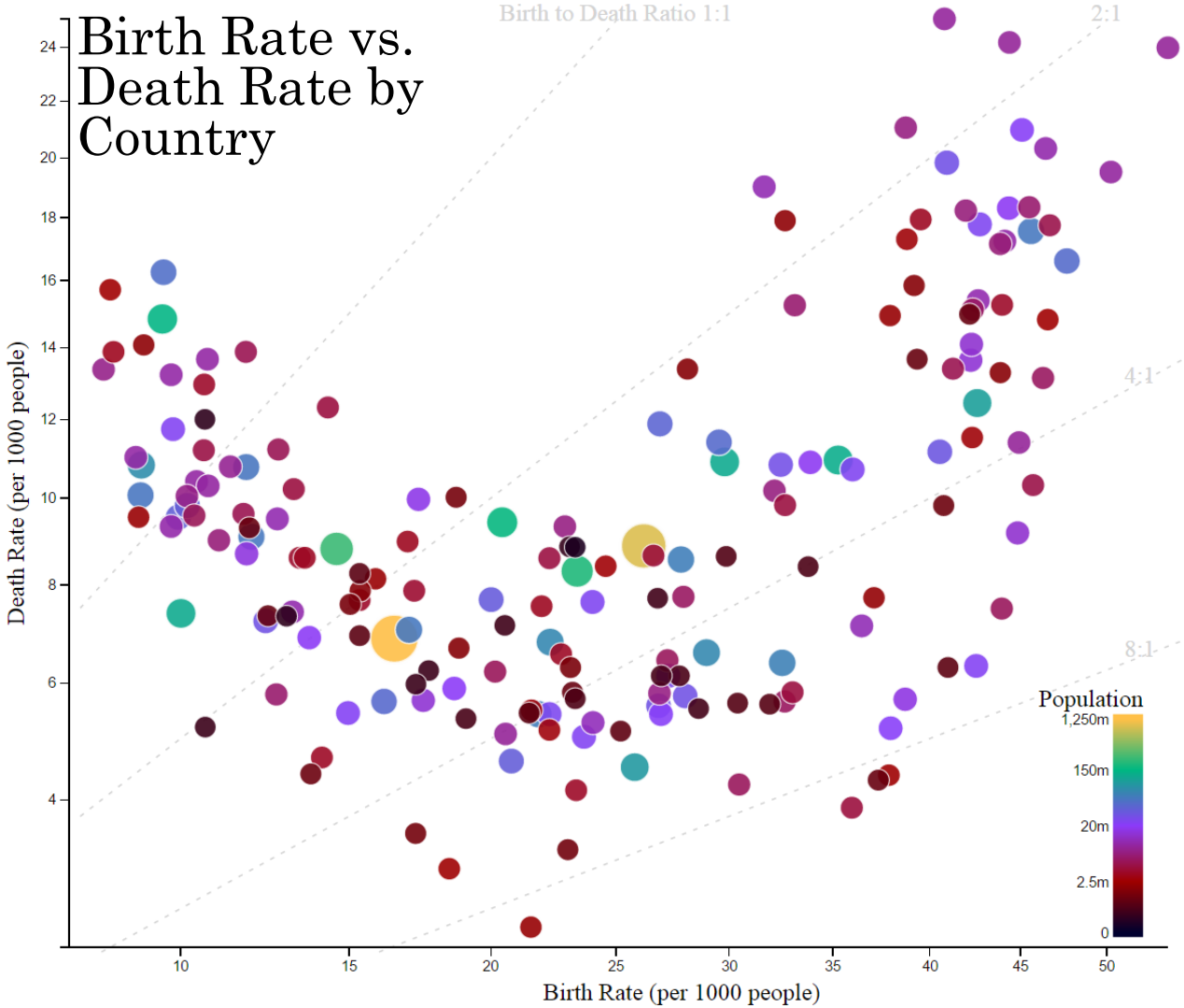


Image: Birth rate vs. Death rate per country, with size and color indicating population; and diagonal lines indicating birth to death ratio. Note visible crescent shape. Image by author.

© 2017 Uncharted Software Inc.

Alphanumeric Points

- Same scatterplot
- 2-letter ISO codes added
- More information, more insight

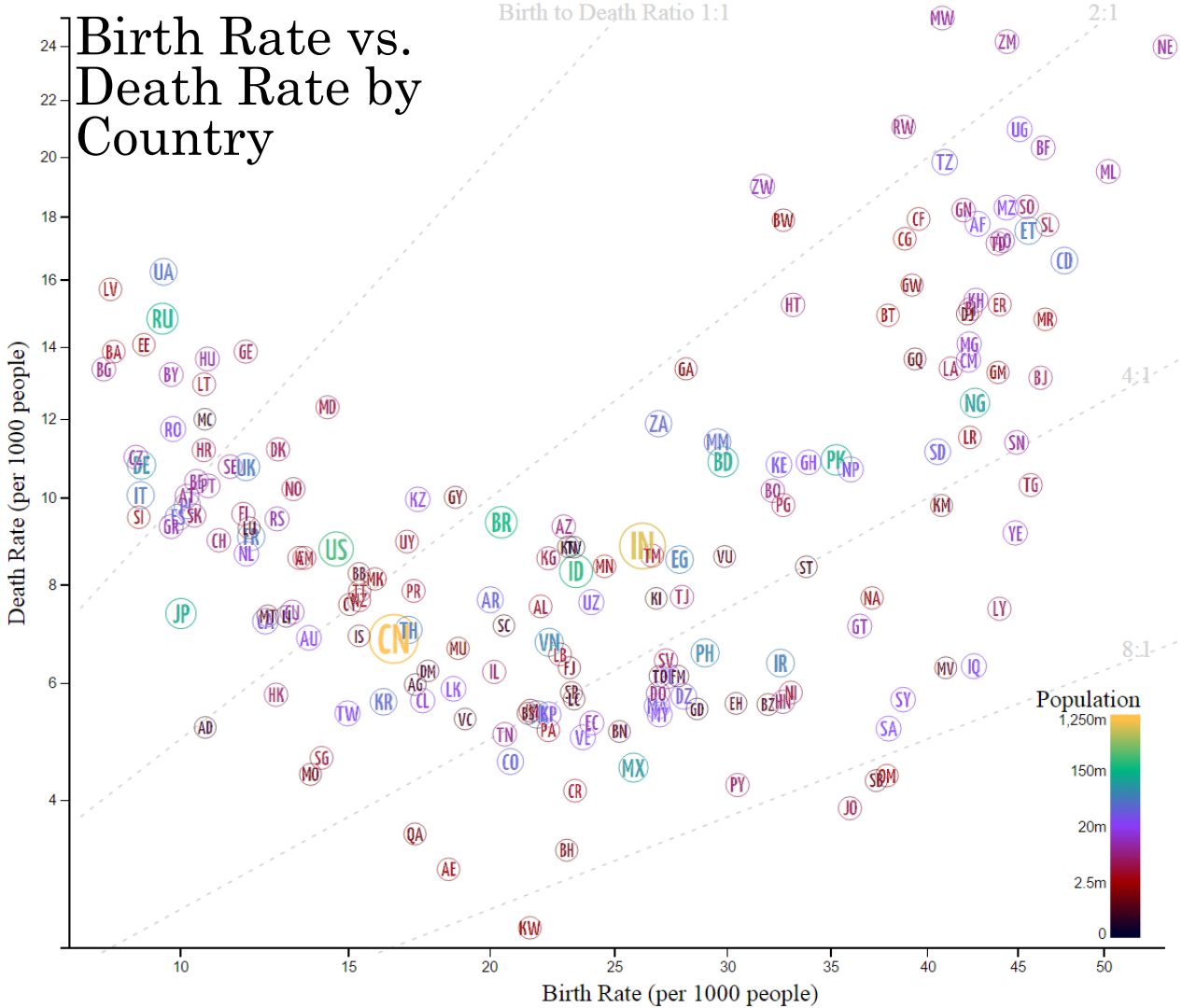


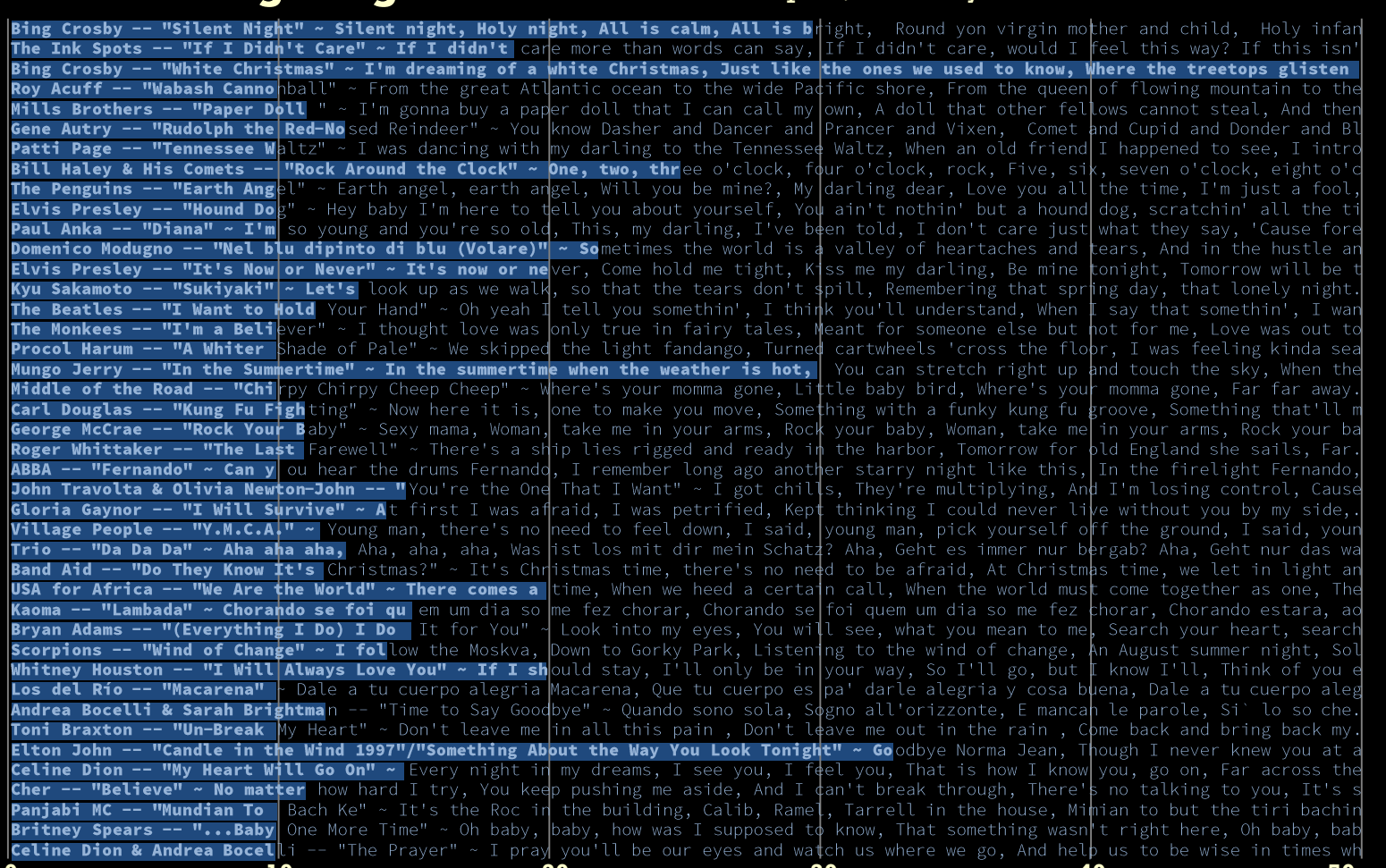
Image: 2 letter country ISO codes act as mnemonics. Note codes upper left: LV, UA, RU, EE – Latvia, Ukraine, Russia, Estonia, etc., all below population replacement rate. Image by author.

© 2017 Uncharted Software Inc.

Bar Chart of Text

Better than label and bar: See artist, title and first line.

Simple label has problems: Artist as label doesn't show the song. Song as label is ambiguous (e.g. who's version of *Silent Night*?). Label using artist + song is very long and leaves little space for bar (e.g. see entry for Elton John). First line provides rich context: maybe you don't know the *Wabash Cannonball*, but with the first line you know what it's about. Images by author.



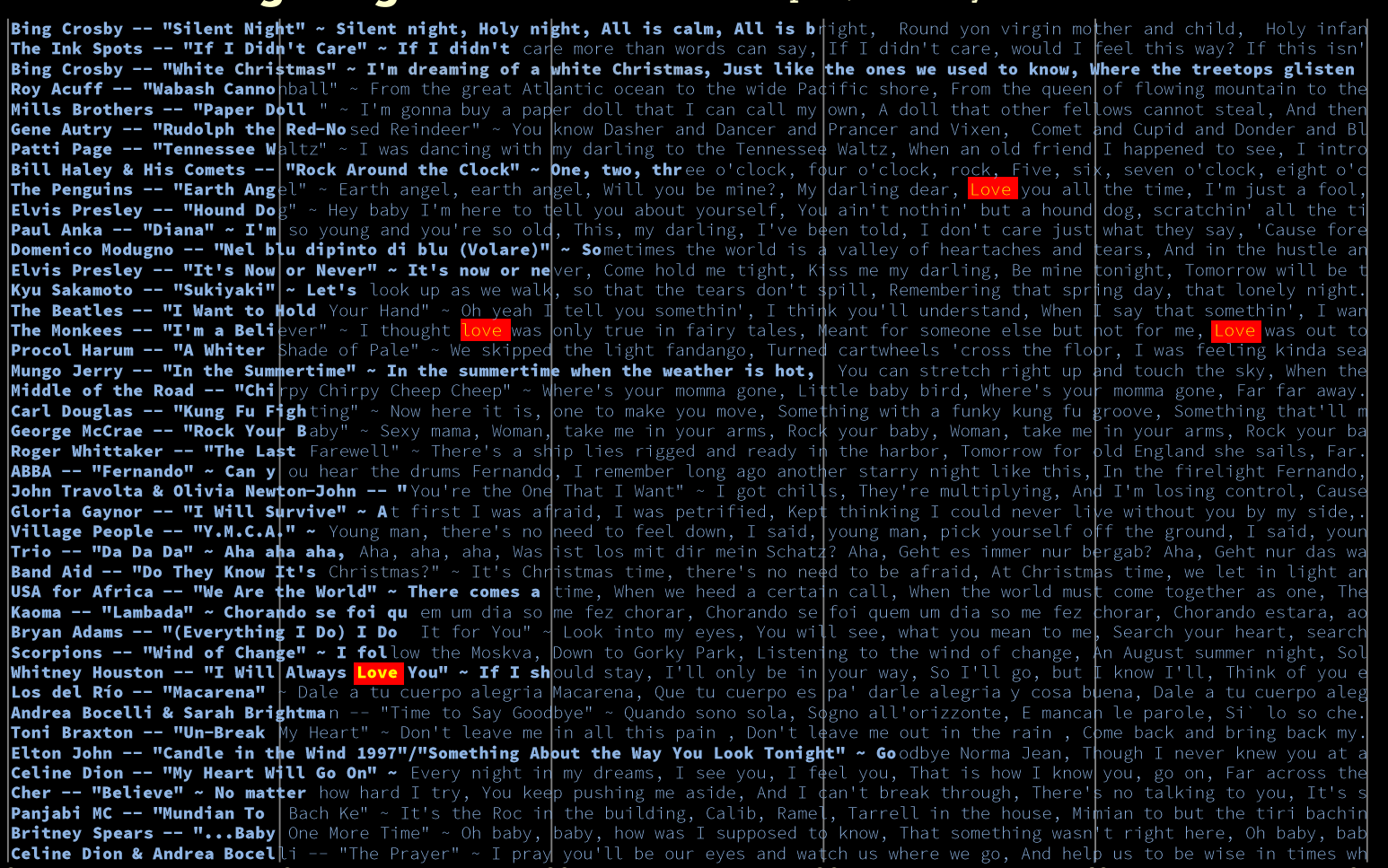
© 2017 Uncharted Software Inc

Bar Chart of Text

Plus content analysis:

Is "*love*" the top word in the top 40?

Best-Selling Singles more than 10 million copies, sorted by release date



Data source: https://en.wikipedia.org/wiki/List_of_best-selling_singles

© 2017 Uncharted Software Inc

Bar Chart of Text

No: "*Christmas*" beats love (but occurs in only 2 songs)

Best-Selling Singles more than 10 million copies, sorted by release date

Data source: https://en.wikipedia.org/wiki/List_of_best-selling_singles

© 2017 Uncharted Software Inc

Bar Chart of Text

And "*baby*" beats both in counts and songs

Best-Selling Singles more than 10 million copies, sorted by release date

Data source: https://en.wikipedia.org/wiki/List_of_best-selling_singles

© 2017 Uncharted Software Inc

3. But what about

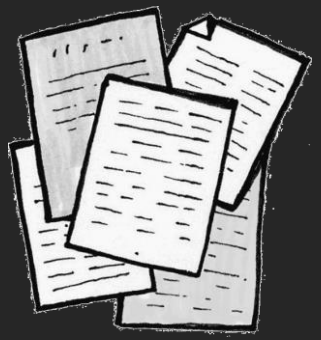
BIG DATA TEXT ANALYTICS?

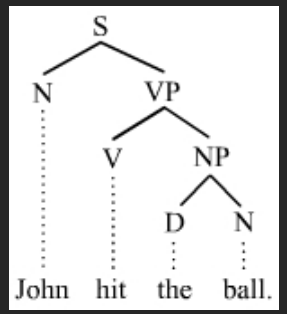
© 2017 Uncharted Software Inc.

Text Analytics 101

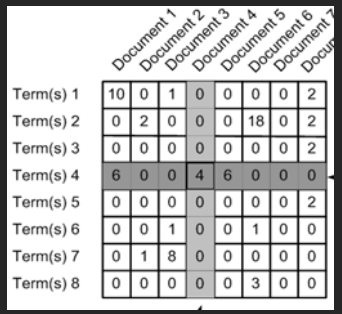
Typical "Natural Language Processing" Pipeline

Raw Text → Preprocessing → Numerical Representation → Analysis

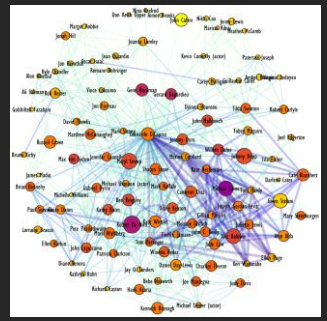




https://en.wikipedia.org/wiki/Parse_tree



http://brandonrose.org/clustering

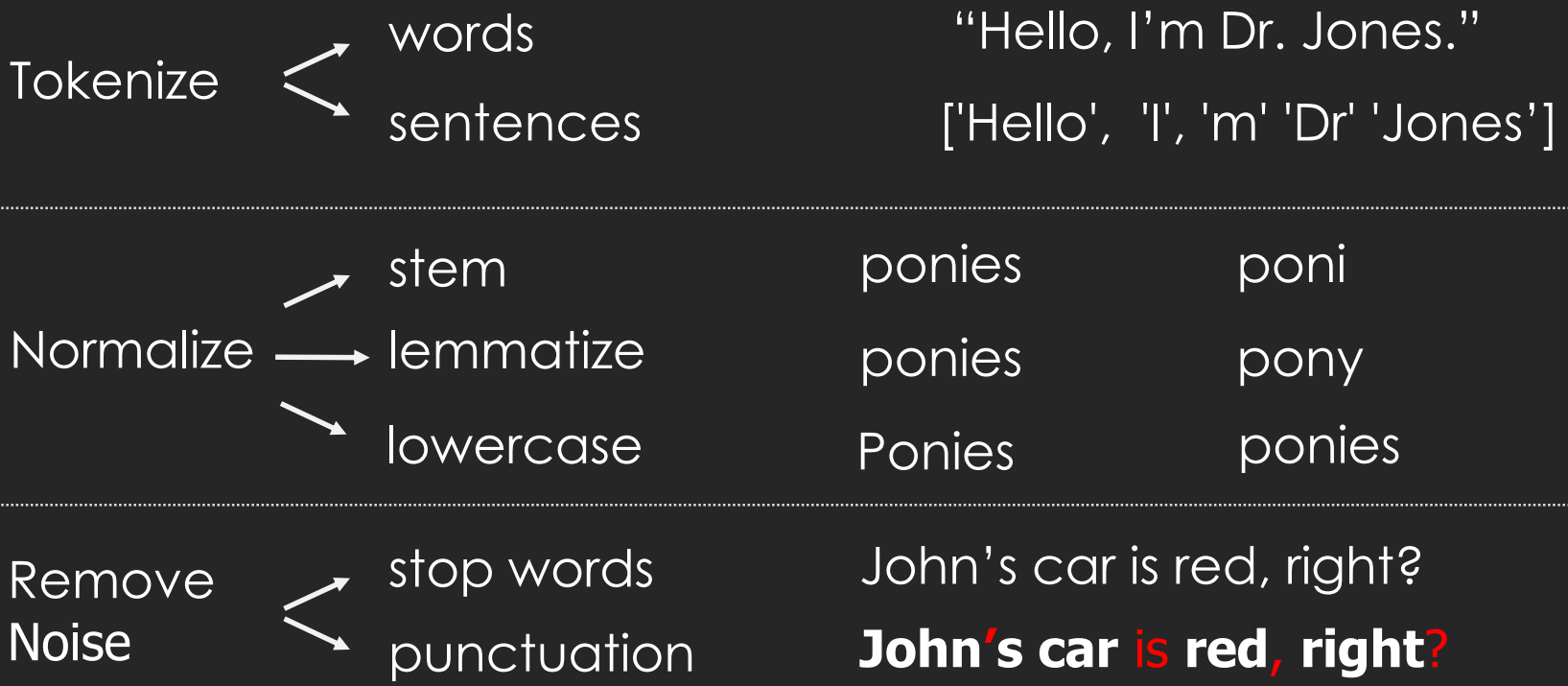


Brath and Jonker. Graph Analysis and Visualization. 2015. Wiley.
http://ca.wiley.com/WileyCDA/WileyTitle/productCd-1118845846.html

© 2017 Uncharted Software Inc.

uncharted

Preprocesing Raw Text



© 2017 Uncharted Software Inc.

Document Representation

Counting

word count
sentence count
syllable count

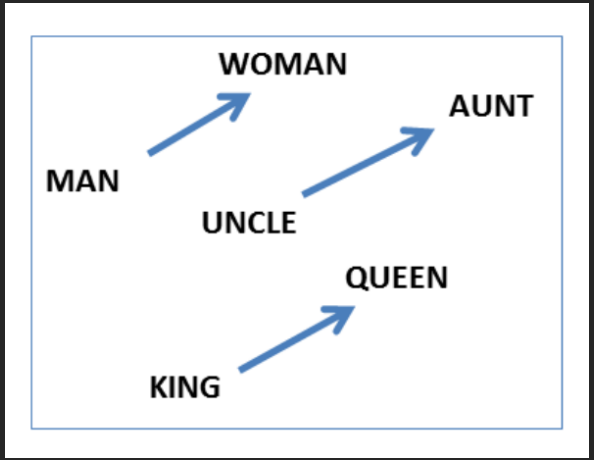
TFIDF

$$\text{\# times term t is in doc d} \times \frac{1}{\text{\# docs term t appears in}}$$

Word Embedding

Text Vectors

	king	queen	man	woman
count	32	147	14	83
TFIDF	0.001	0.097	0.051	0.147



Mikolov et al., NAACL HLT, 2013. https://www.aclweb.org/anthology/N13-1090

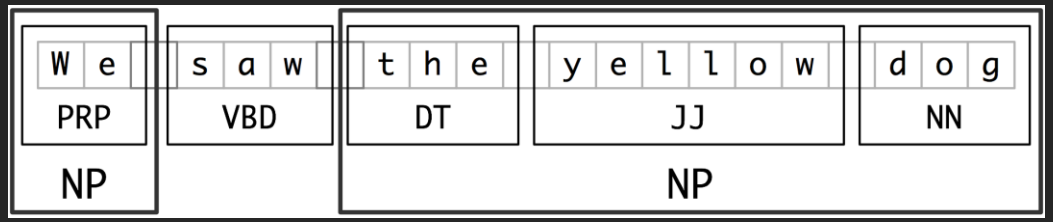
© 2017 Uncharted Software Inc.

Document Representation

Part-of-Speech tagging

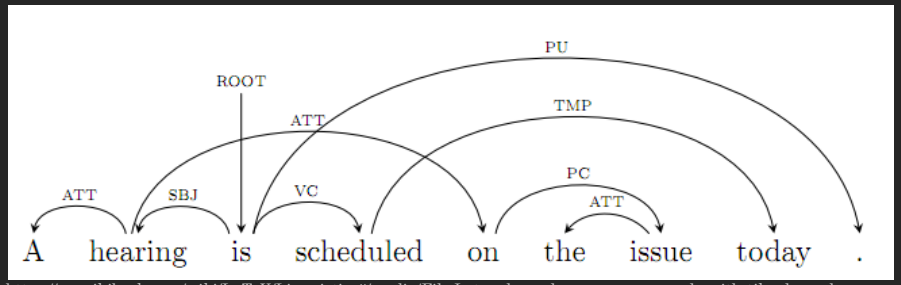
['And, 'now', 'for', 'something', 'completely', 'different']
'CC' 'RB' 'IN' 'NN' 'RB' 'JJ'

Chunking / Phrase extraction



Bird, Steven, Edward Loper and Ewan Klein (2009). Natural Language Processing with Python. O'Reilly http://www.nltk.org/book/ch07.html

Dependency parse tree



https://en.wikibooks.org/wiki/LaTeX/Linguistics#/media/File:Latex-dependency-parse-example-with-tikz-dependency.png

© 2017 Uncharted Software Inc.

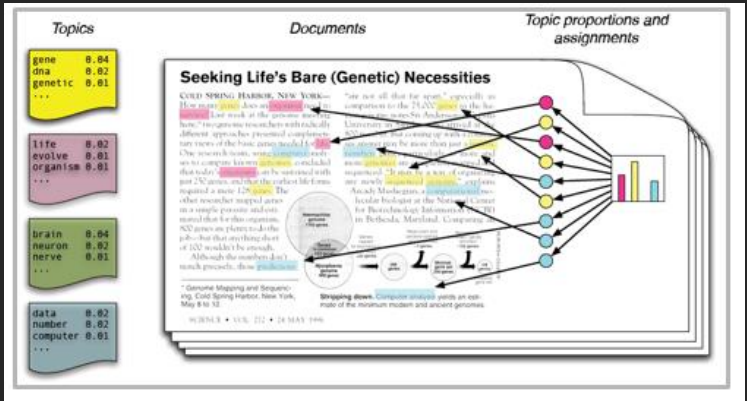
Text Analysis

Topic Modeling

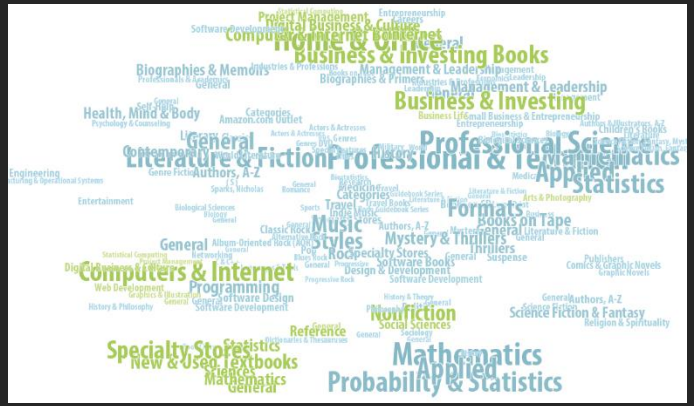
Using statistical features to discover abstract 'topics' that occur in a collection of documents

'topic' is a probability distribution of words

Topic 1: algorithm, computer, data, ...
Topic 2: gene, data, bio, ...
Topic 3: math, equation, matrix, ...
Topic 4: chemical, carbon, water, ...



Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.
http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf or
https://lists.cs.princeton.edu/pipermail/topic-models/2011-August/001535.html



Topic graph, showing topic nodes and edge connections removed.
Brath and Jonker. Graph Analysis and Visualization. 2015. Wiley.
http://ca.wiley.com/WileyCDA/WileyTitle/productCd-1118845846.html

© 2017 Uncharted Software Inc.
uncharted

Other NLP Tasks

Coreference resolution	"Scott took a flight. He said it was very crowded"
Text Classification	Spam / Not Spam
Text Similarity	(definitely, definately) = 1
Sentiment Analysis	Strata was **great** this year! :) = POSITIVE
Entity extraction	New York is a metropolis ➔ ('New York', 'city')
Text Summarization	Mostly harmless
Machine Translation	Mostly harmless -> ほとんど無害
Natural Language Generation	"Sounds good, talk to you then"

© 2017 Uncharted Software Inc.

4. TEXT ANALYTICS & VIZ

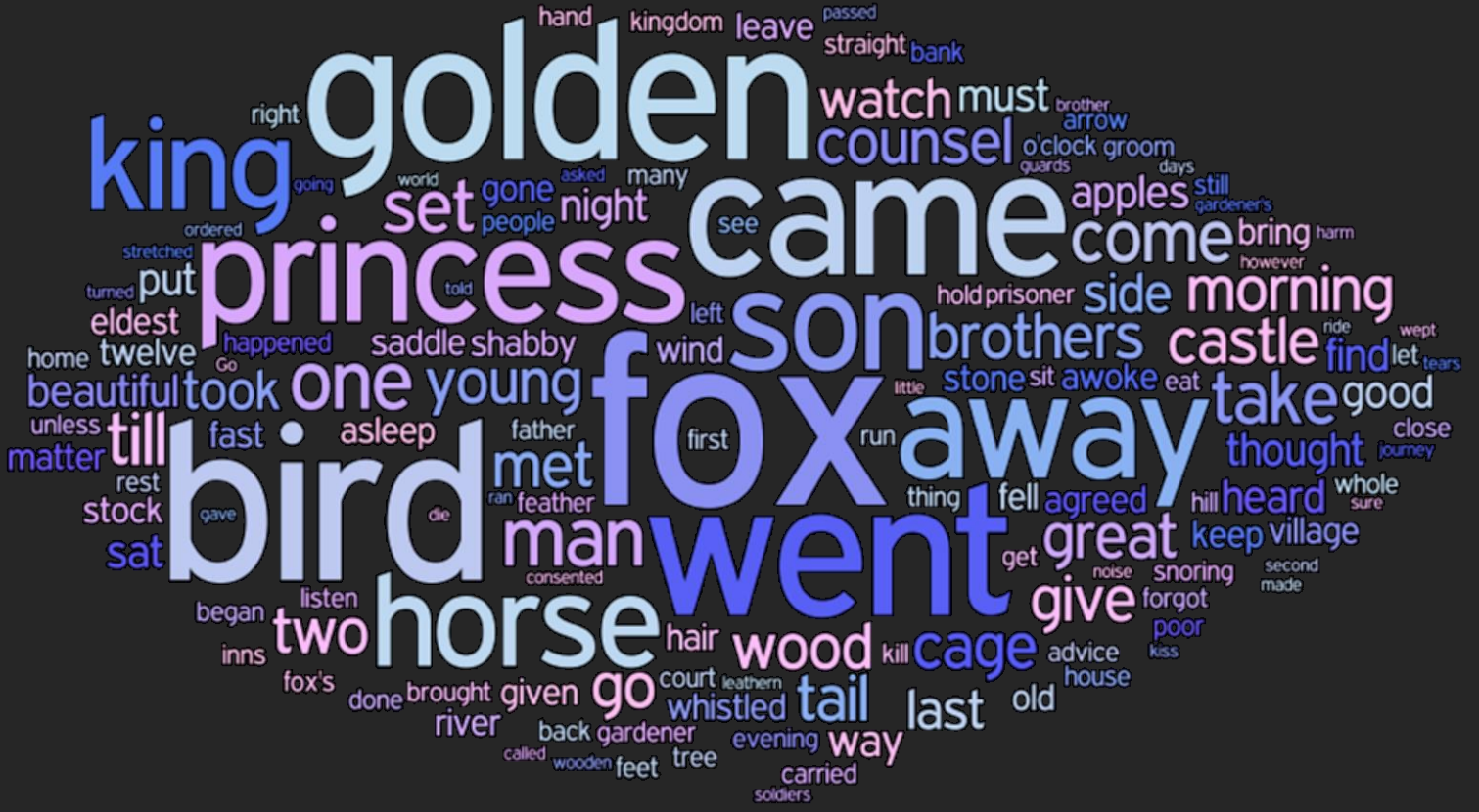
together!

© 2017 Uncharted Software Inc.

Character Traits

Who are the characters?

What are their traits?



Tag cloud showing frequency of words in *Grimms' Fairy Tales,* a collection of 50+ short stories. Tag clouds reveals some characters, e.g. king, princess, fox; but doesn't provide any insight into those characters. *(via wordle.net)*

© 2017 Uncharted Software Inc.

uncharted

NLP to find Traits

1. Extract the characters
2. Extract the adjectives near the characters
3. Visualize

e.g. This example, one sentence from *Grimms' Fairy Tales*.

When you come to the old king, and he asks for the beautiful princess, you must say, "Here she is!"

When you come to the old **king**, and he asks for the beautiful **princess**, you must say, "Here she is!"

When you come to the *old* **king**, and he asks for the *beautiful* **princess**, you must say, "Here she is!"

© 2017 Uncharted Software Inc.

Traits

Apply to whole text

Unexpected character traits?

Are witches really wicked, or maybe just old?

© 2017 Uncharted Software In

Characters from Grimms' Fairy Tales with Associated Adjectives Weighted by Frequency

Character	List of adjectives, weighted by frequency: 2 3 4-5 **6-9** **10+**
bird	**beautiful** splendid open wooden like hanging
cat	little one long
fox	**old** dead young first fast
gretel	little poor good
hansel	**little** like fat
bride	false true first right real
king	**old** **great** **one** **young** **three** angry beautiful married like third
princess	**beautiful** young **last** dear enchanted strange true third free
girl	**little** poor lazy pretty young dead beautiful silly
queen	beautiful late little far
wife	standing married next poor two beautiful new one dear true
witch	**old** wicked

0 5 10

Approximate number of unique adjectives

CEO's Use of Emotion Words During Earnings Calls, weighted by frequency

CEO	List of adjectives, weighted by frequency: 2 3-4 5-7 **8-11** **12+**
Carl Bass	pleased good
Carl Trowell	**demand** **lower** good debt improve cash older grow tender retirement late termination
Carlos A. Rodriguez	grow confidence pleased success excited proud progress
Carol M. Meyrowitz	pleased confident share margin grow powerful organization excellent shopping terrific passionate gai
Charles Alutto	**good** **kind** **waste** **cash** excited lines share safe **hazardous** **versus** **fee** pleased resources tough respe
Charles B. Stanley	**completion** improvement lower gross progress reject rejection
Charles Bunch	**improvement** **lower** **share** cash versus margin **unfavorable** **gross** good aggressive tax income kind i
Charlie Scharf	**kind** **good** **share** **change** **lower** **grow** pay pleased cross case versus **helpful** terrific progress excited
Chris Connor	**share** **kind** **margin** **good** **pretty** **demand** **gross** **john** **confidence** versus income cash bad interior re
Chris Sinclair	**good** **progress** pretty hot content grow share change organization **improvement** **gross** **finally** kind e
Christopher Crane	good create late hearing grow successful pleased share advocacy regulatory clean majority kind debt
Chuck Bunch	**good** **improvement** **share** **unfavorable** versus cash lower excellent **demand** **respect** difficult debt inc
Chuck Jones	**case** **cash** **good** **kind** improvement progress finally cutting inflation court improve share risk electric
Clay Williams	**lower** **demand** **pretty** **kind** good excluding **improve** **completion** **versus** share grow deflation tough di
Clifton Pemble	margin lower income gross finally excited share pretty
Craig A. Menear	share pleased

0 5 10

Approximate number of unique adjectives

Most CEO's tend to use the same words – not much insight.

© 2017 Uncharted Software Inc.

uncharted

CEO's Use of Emotion Words During Earnings Calls, weighted by frequency

CEO	List of adjectives, weighted by frequency: 2 3-4 5-7 **8-11 12+**
Andrew P. Wilson	**battlefield star** console entertainment dragon fun ultimate
Bruce Broussard	star tax enforce good progress successful
David Hallal	**disease complement devastating** progress **therapeutic** success majority hospital
Fabrizio Freda	beauty exciting **prestige** successful
Gracia Martore	**love** content **respect** terrific apologize lagging
J.P. Bilbrey	**chocolate shanghai lower** unfavorable force
Jim Tisch	**diamond cash** income special pay doubt lose
John D. Idol	**luxury** demand shopping watch bottom
John Ferriola	**pig** improvement **cash lower** challenge
John Lechleiter	**cancer** disease **deal** regulatory tax therapeutic
Kelly King	**mortgage risk** income **tough** money fee
Larry D. Young	**punch** case music **football love passion**
Les Moonves	**content deal** bang **success** terrific confident scorpion late
Mark Parker	**basketball** demand excited elite offense discipline
Mark R. Hunter	**beer** tax lower unfavorable loss organization banquet blue mad
Omid Kordestani	shopping **holiday progress pay music** watch friction
Richard J. Harshman	demand fell defect excellent labor wages aggressive falling
Steven Temares	**sue shopping baby** improve excited exciting share **success** pleased fulfillment
Thomas Kennedy	**missile defense demand government** award bottom threat

Electronic Arts

Estée Lauder

Alexion Pharmaceuticals

Loews

0 5 10

Approximate number of unique adjectives

Focusing on unique terms (TFIDF) characterizes the companies.

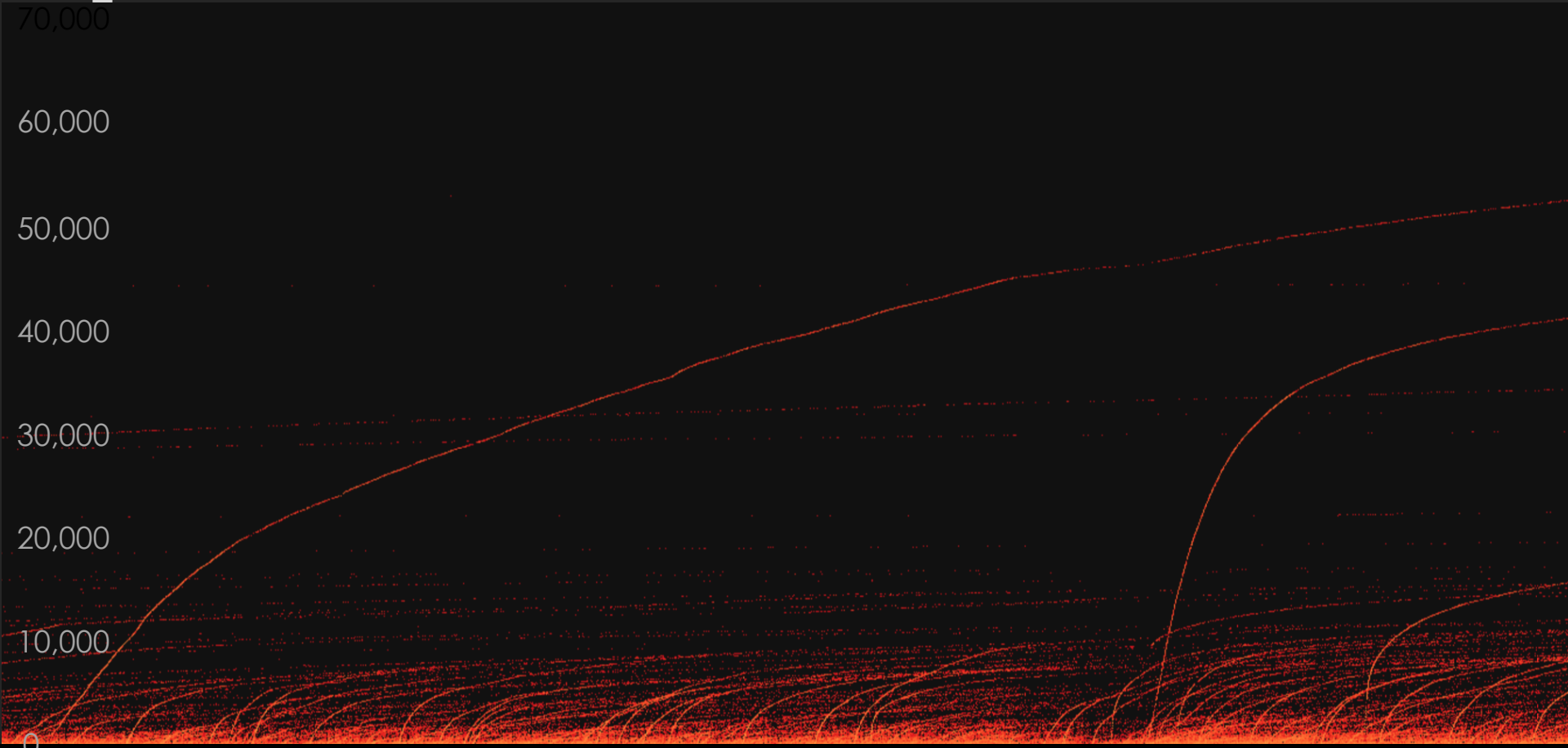
uncharted

5. STICKY social media

© 2017 Uncharted Software Inc.

Uncharted Salt visualization of tweets. E.g. https://unchartedsoftware.github.io/salt-core/

Superbowl 50 Retweets



Uncharted Flare Chart. Vertical axis is number of retweets. Lines are individual tweets. Some tweets flare up quickly then flatten, some go up, stutter, then go up more, etc..

Superbowl 50 Retweets

uncharted

Note: some retweet values are offscreen

Betty White
@BettyMWhite

Follow

I taught @CameronNewton everything he knows. #SuperBowl

7:44 PM - Feb 7, 2016

919 75,417 87,997

7:00 7:30 8:00 8:30 9:00

Uncharted Flare Chart. Betty White goes up flattens a bit, then up more, repeats. Live version: http://unchartedsoftware.github.io/salt-core/demos/superbowl-retweets/

Superbowl 50 Retweets

uncharted

Coldplay
@coldplay

Follow

#SB50 #PepsiHalftime

9:42 PM - Feb 7, 2016

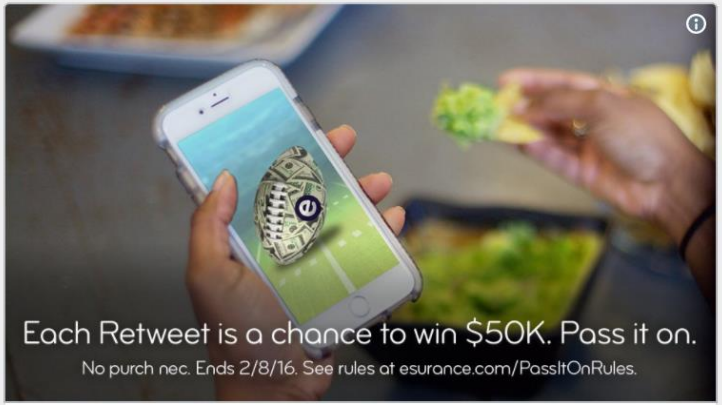
751 49,474 58,172

0 7:00 7:30 8:00 8:30 9:00

Uncharted Flare Chart. Coldplay has a popular tweet at 8:45 pm which then flattens out.

Superbowl 50 Retweets

uncharted



Esurance ✓
@esurance

Follow

Before you reach for another chip, RT for your chance to win. #EsuranceSweepstakes #SB50

9:12 PM - Feb 7, 2016

476 12,500 1,924

0 7:00 7:30 8:00 8:30 9:00

Uncharted Flare Chart. Esurance has a retweet campaign which pops up quickly but starts to flatten after 20 minutes.

Superbowl 50 Retweets

uncharted

Scott Kelly @StationCDRKelly

Follow

Got to see the #SuperBowl in person after all! But at 17,500MPH, it didn't last long. #YearInSpace

8:58 PM - Feb 7, 2016

445 11,831 22,973

0 7:00 7:30 8:00 8:30 9:00

Uncharted Flare Chart. This tweet from the International Space Station doesn't flare up, but keeps getting more and more retweets. Note how only one tweet can be examined at a time.

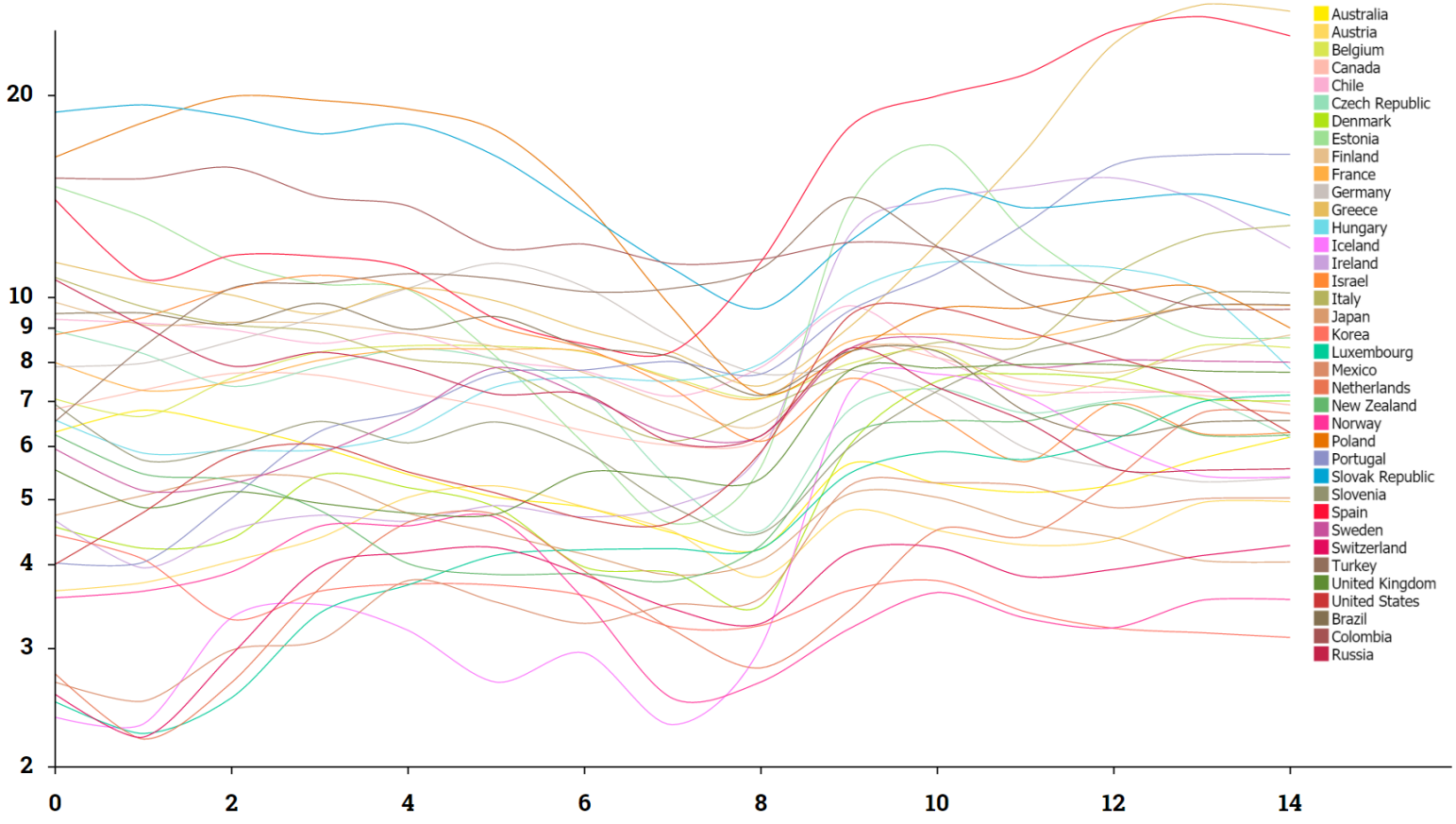
Line Chart

- Which line is which?

Similar problem: many lines on a line chart are difficult to disambiguate. Only way to disambiguate is to interact with each line individually. Slow!

Unemployment Rate as Percent Civilian Labour Force

2000-2014 by Country.



Source: OECD (stats.oecd.org, retrieved Oct 24, 2015). Author: Richard Brath/ London South Bank University 2015 (richardbrath.wordpress.com)

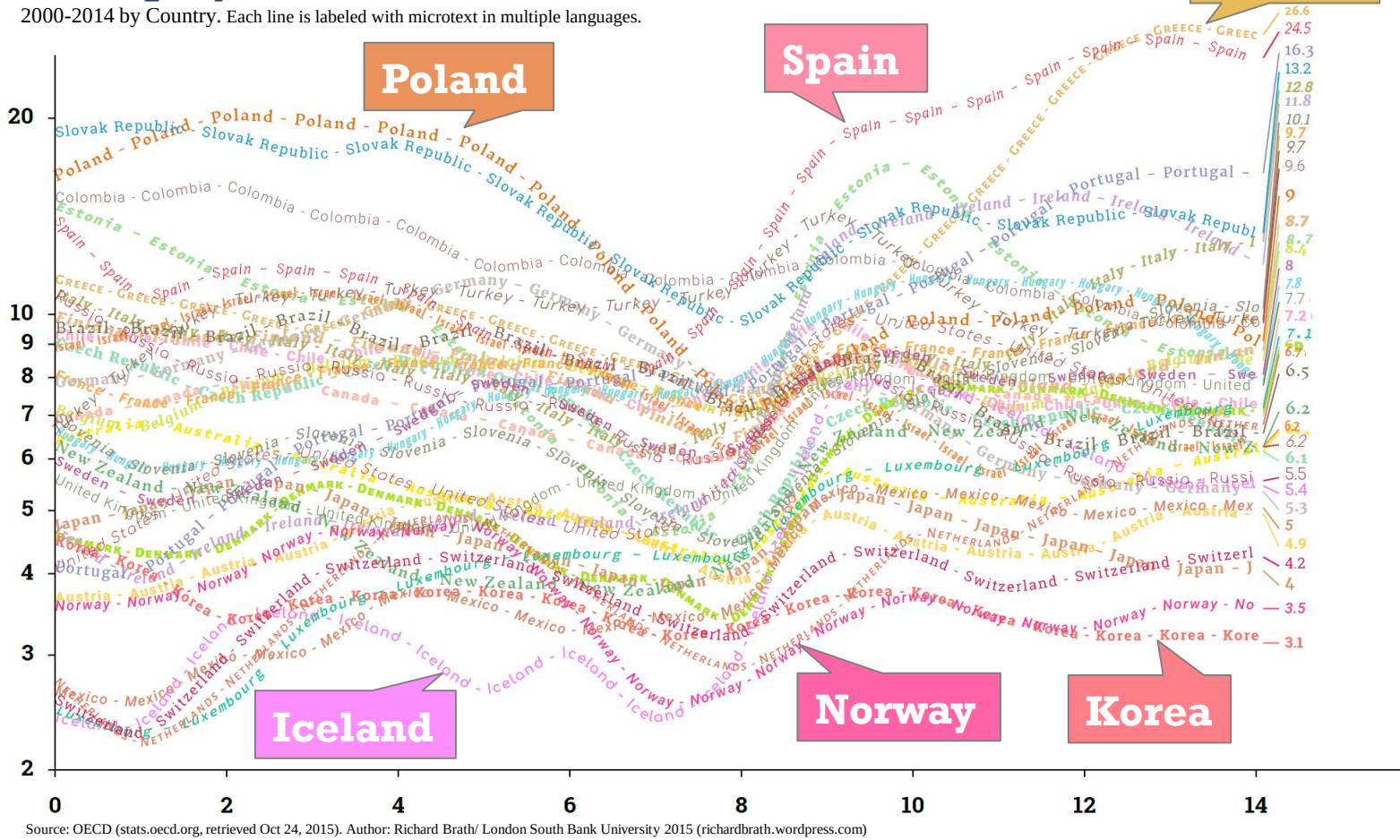
© 2017 Uncharted Software Inc.

Line Chart

Turn the line into text

Unemployment Rate as Percent Civilian Labour Force

2000-2014 by Country. Each line is labeled with microtext in multiple languages.



Source: OECD (stats.oecd.org, retrieved Oct 24, 2015). Author: Richard Brath/ London South Bank University 2015 (richardbrath.wordpress.com)

When lines are text, identity can be directly read. Callouts for audience at back of room.

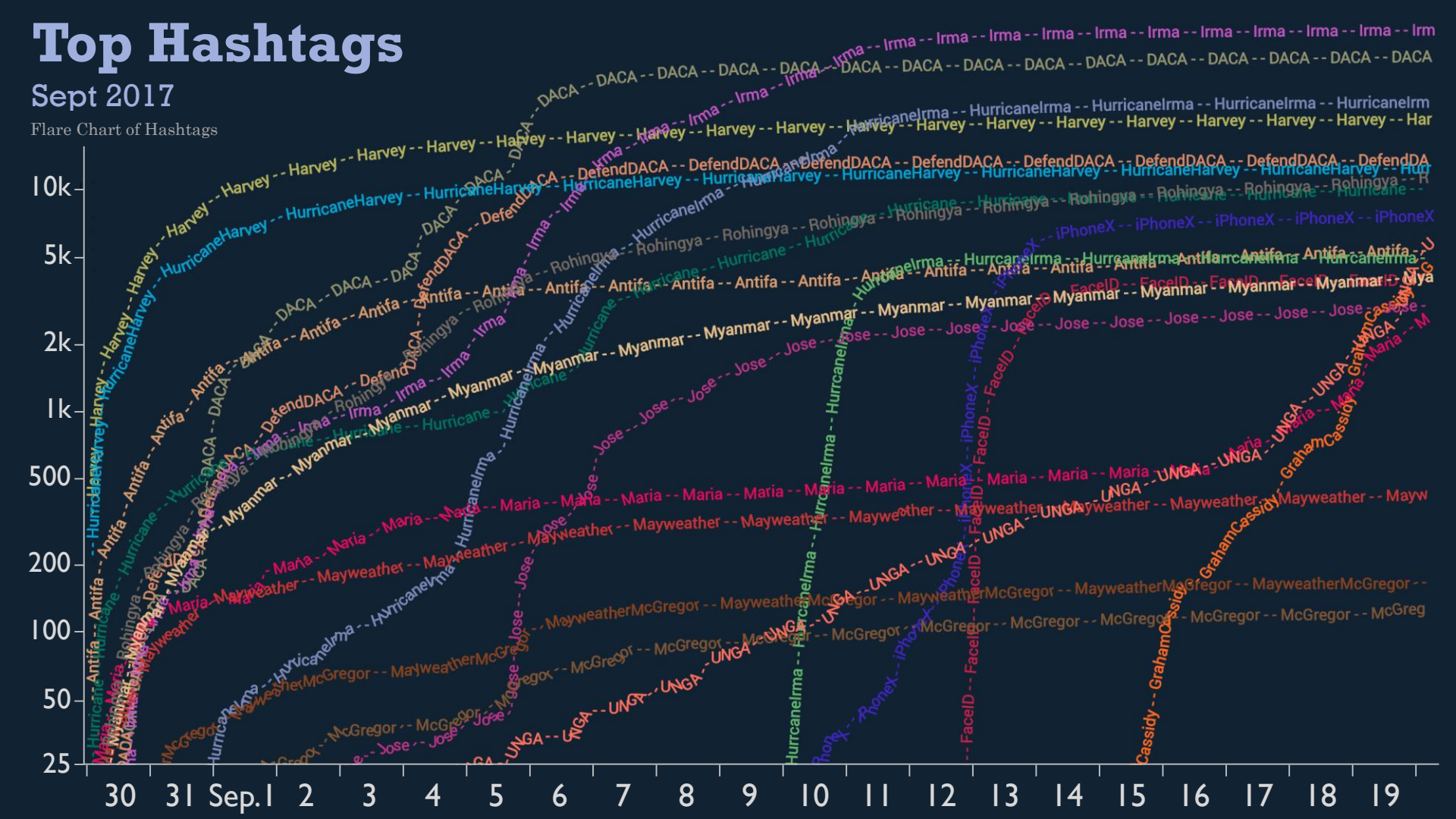
© 2017 Uncharted Software Inc.

uncharted

Top Hashtags

Sept 2017

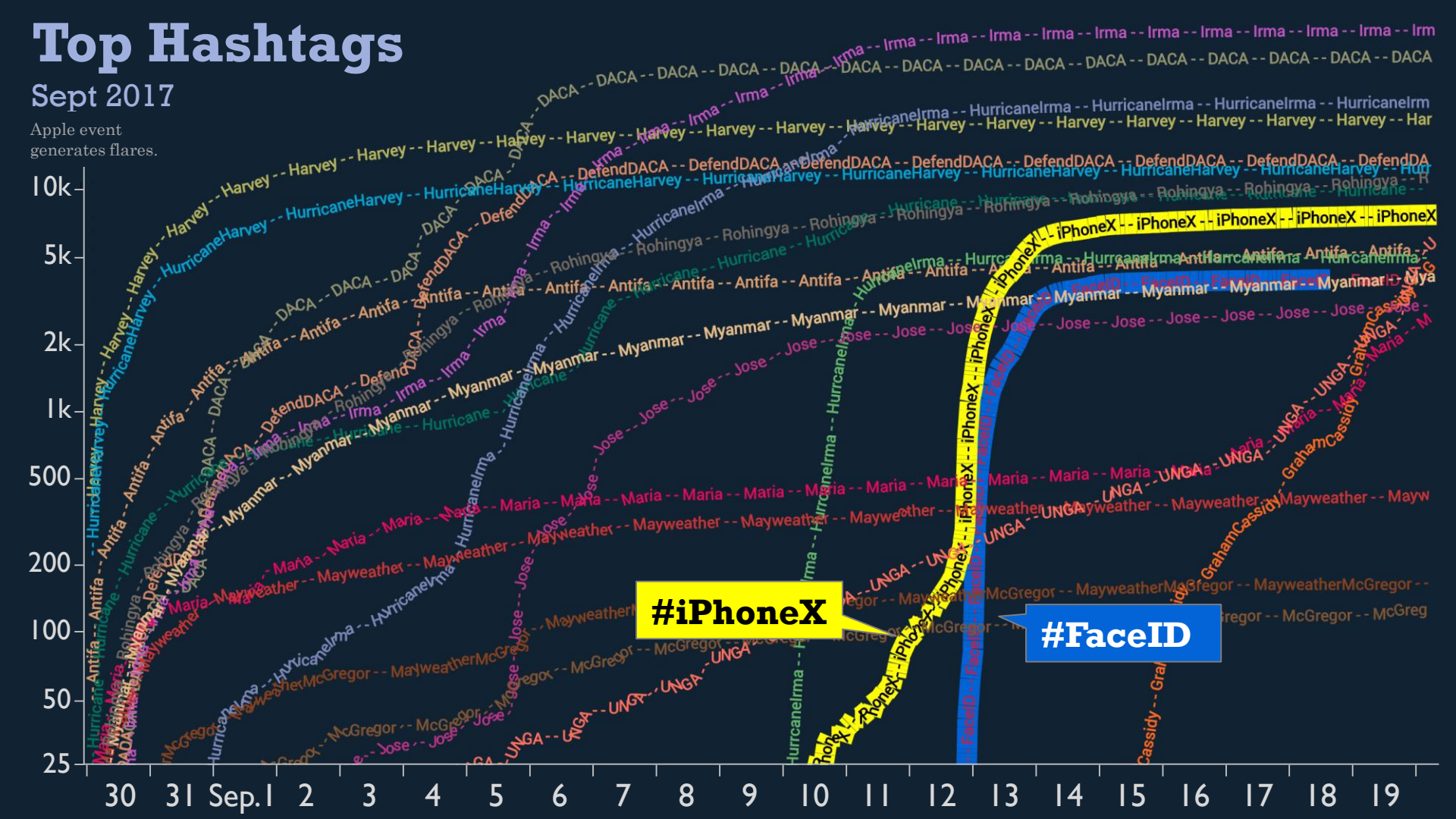
Flare Chart of Hashtags



Top Hashtags

Sept 2017

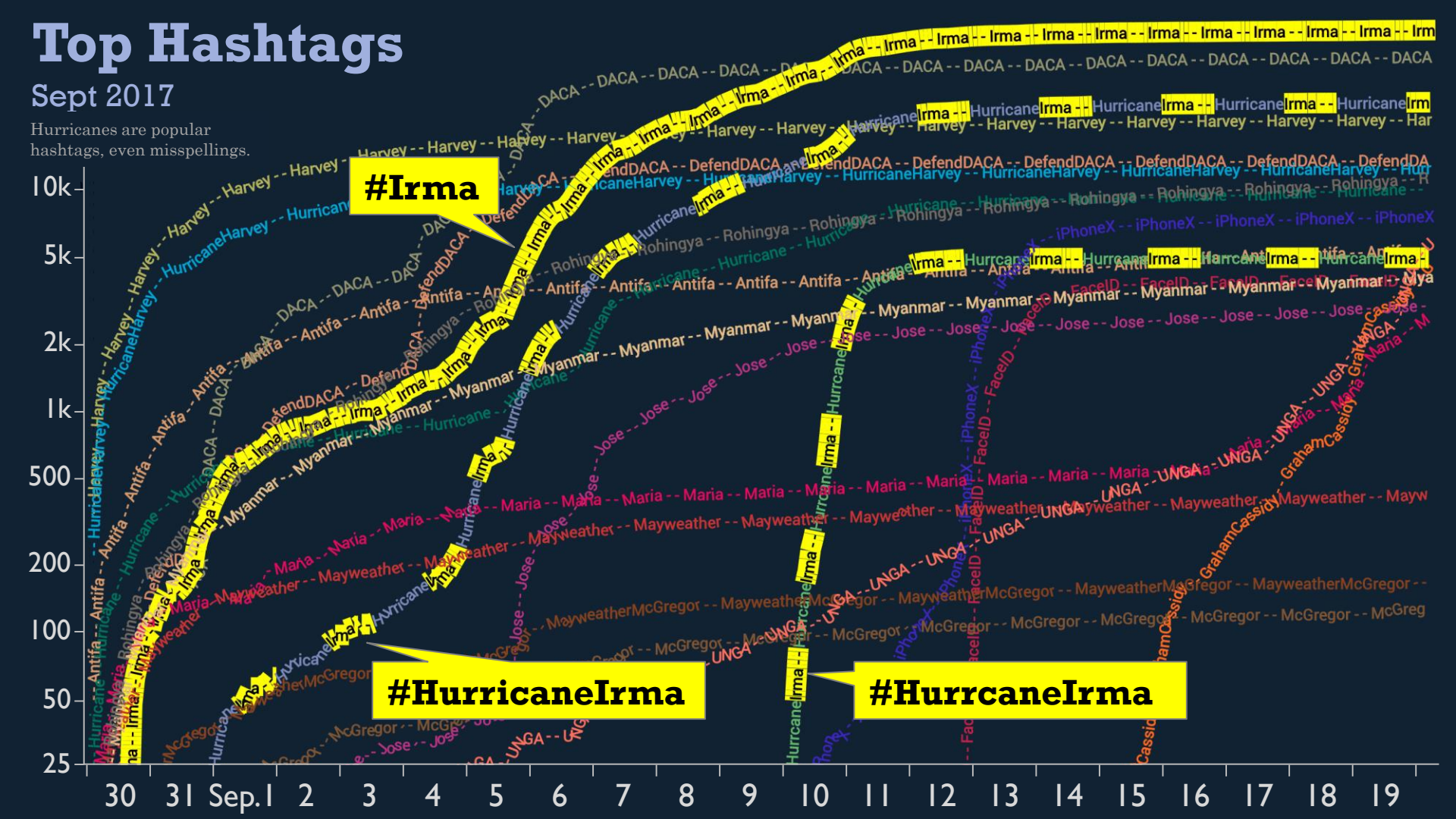
Apple event generates flares.



Top Hashtags

Sept 2017

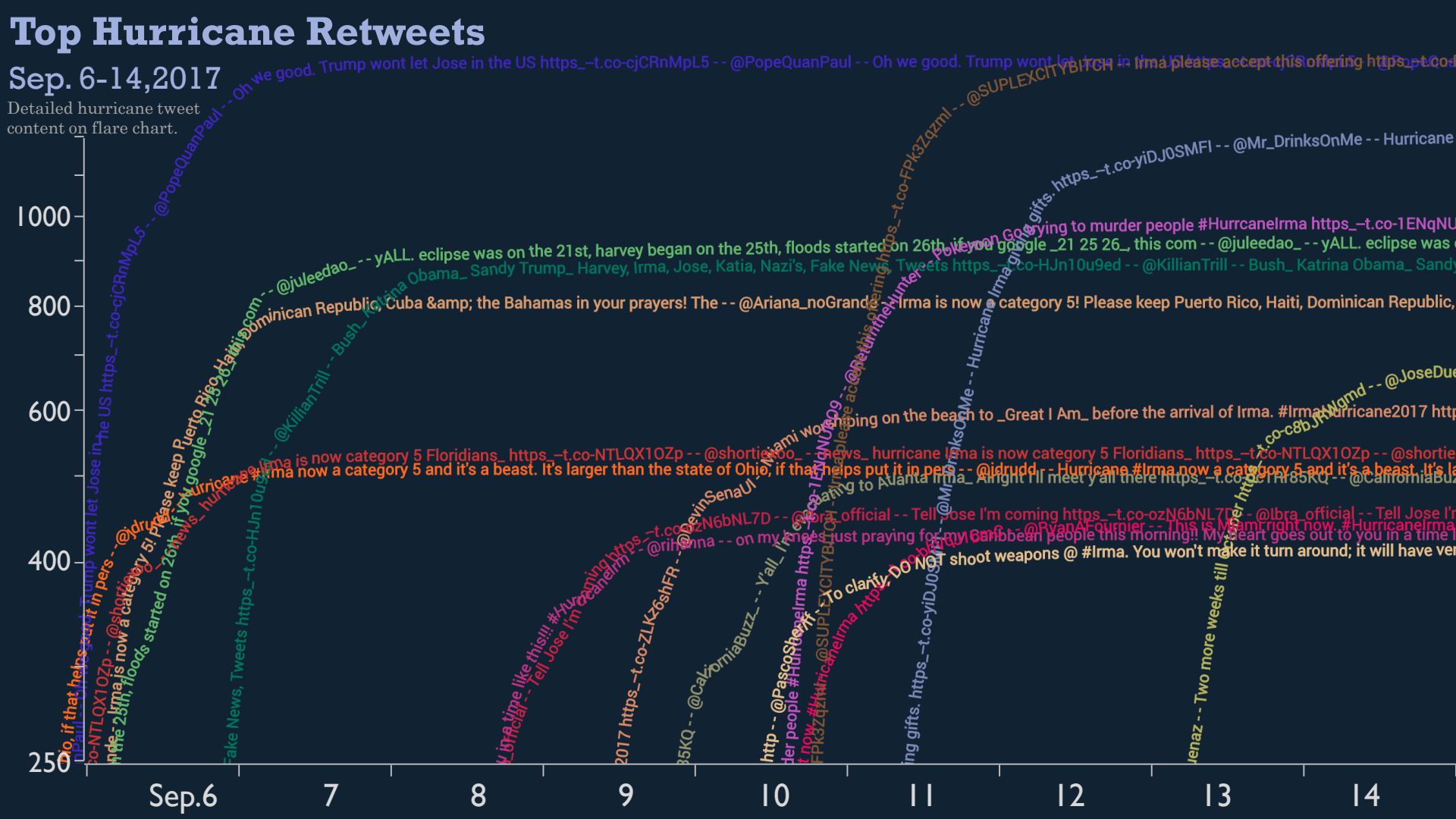
Hurricanes are popular hashtags, even misspellings.



Top Hurricane Retweets

Sep. 6-14,2017

Detailed hurricane tweet content on flare chart.



Top Hurricane Retweets

Sep. 6-14,2017

Comedic flare…

@PopeQuanPaul – Oh we good. Trump wont let Jose in the US.

1000

800

600

400

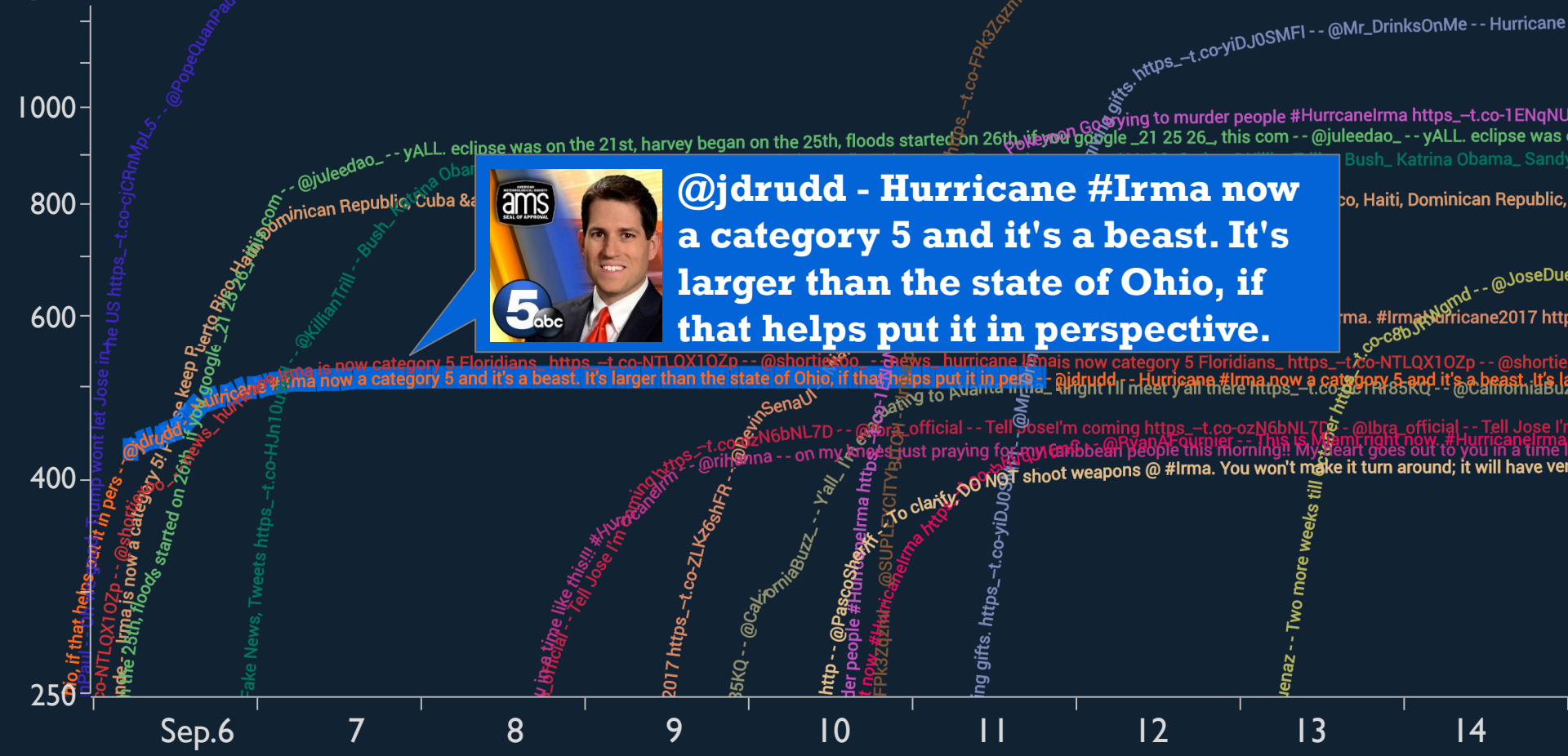
250

Sep.6 7 8 9 10 11 12 13 14

Top Hurricane Retweets

Sep. 6-14,2017

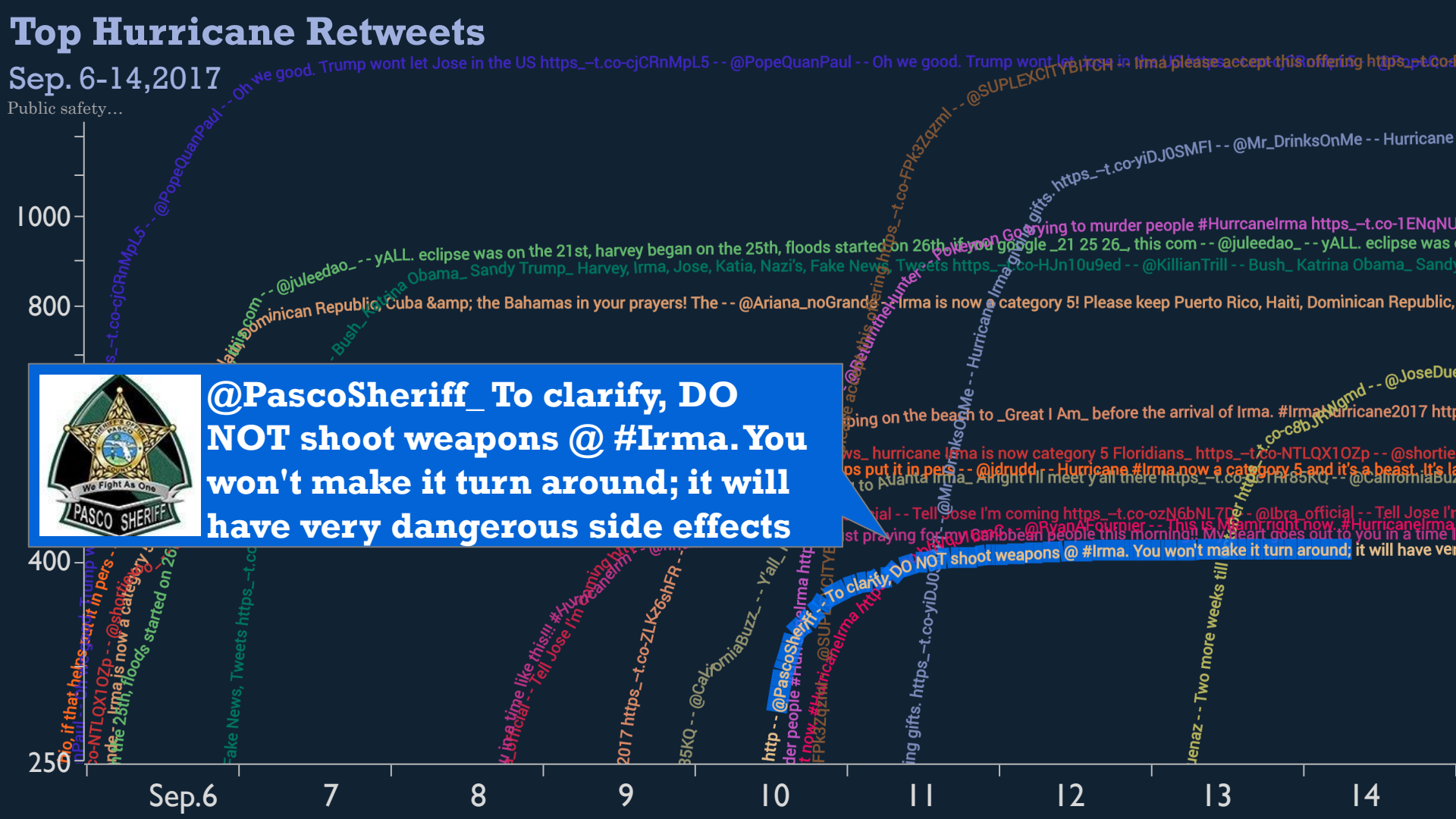
Important news...



Top Hurricane Retweets

Sep. 6-14,2017

Public safety...



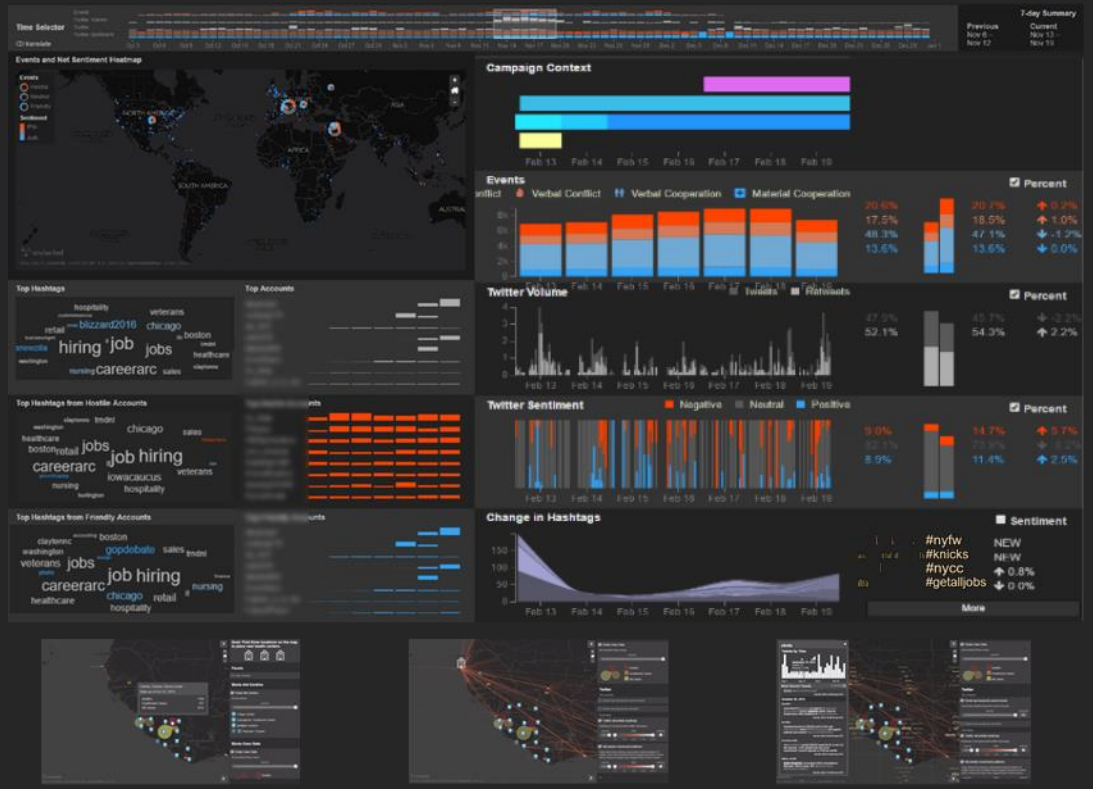
So What

Sticky is about a lot of things

- number retweets
- longevity
- reach across communities
- who's saying it
- what they're saying

And…

- geographic reach
- topics
- demographic segmentation
- message impact
- drift over time



© 2017 Uncharted Software Inc.

uncharted

Salt

TILE-BASED VISUAL ANALYTICS

- Hierarchical data tiling using cluster computing.
- Interactive on-demand image tile generation.
- Layers of raw data and derivative analytics.

OPEN SOURCE

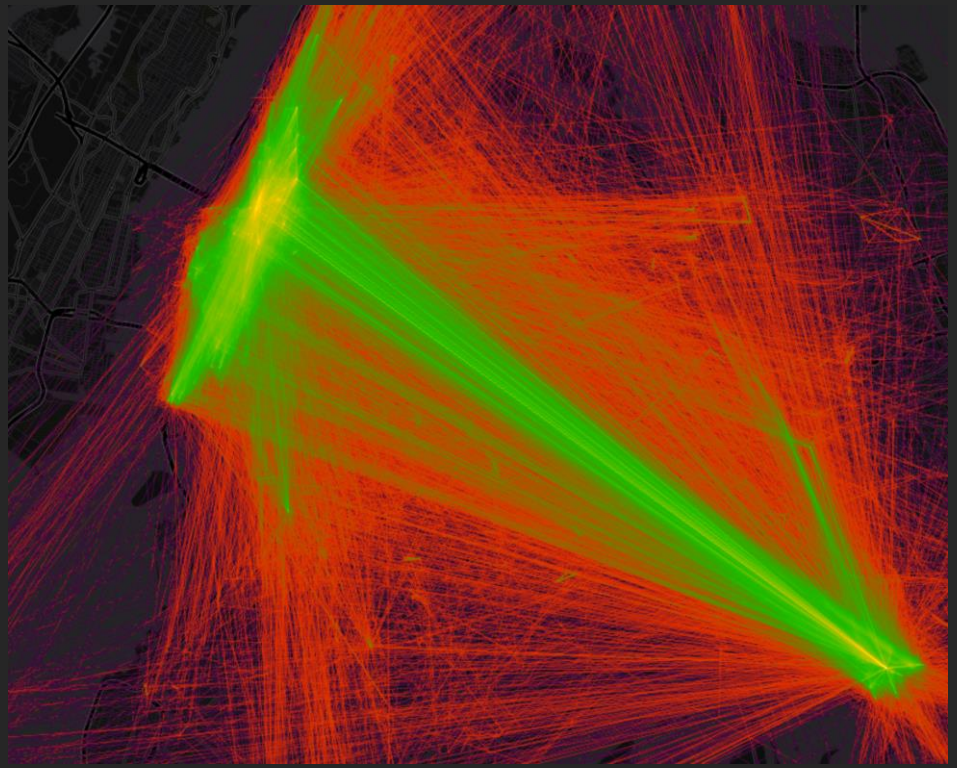
- Saltlib.com
- Built on Apache Spark, Hadoop and ElasticSearch.
  - Cross Plots, Geospatial, Time-series, Graph analysis

© 2017 Uncharted Software Inc.

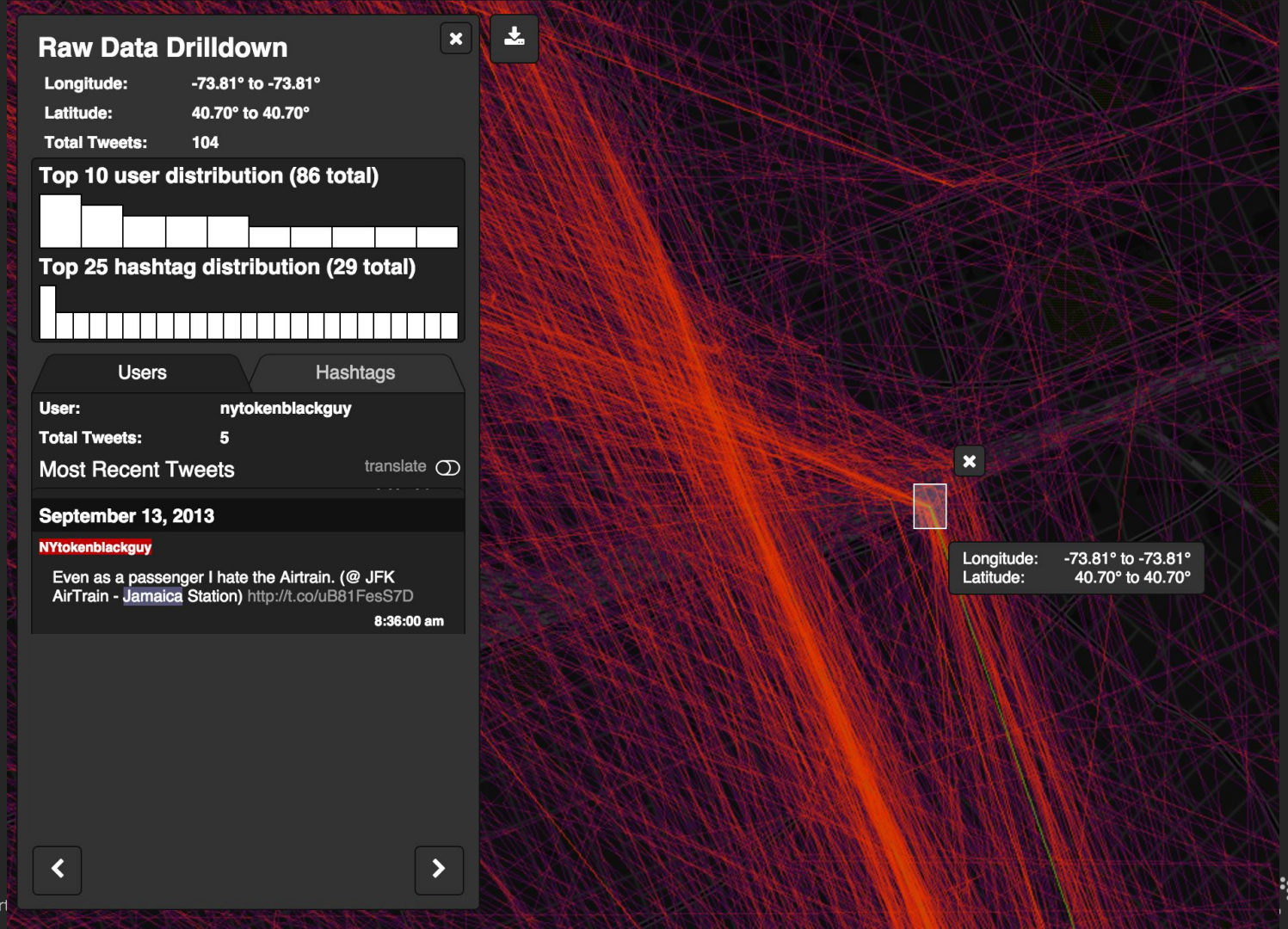
Geo-temporal Movement and Events View

People Trails

- By connecting users geo-coded social media posts ordered temporally aggregate "people trails" can be extracted and aggregated to see movement patterns
- User provides selection criteria to specify population segment to generate trails
  - Based upon topic or keywords, time and location or start with seed social media accounts and generate trails for community
  - E.g. generate trails for all people going through Kennedy Airport June 1 – June 20



© 2017 Uncharted Software Inc.

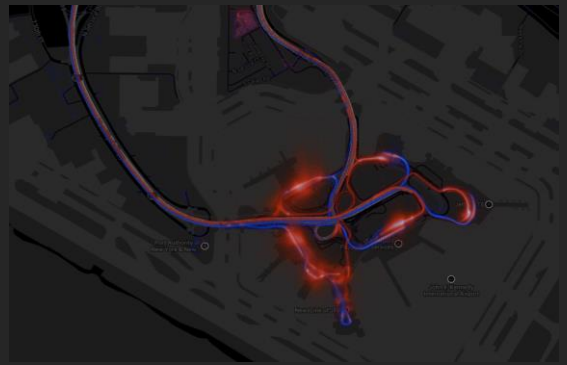


© 2017 Uncharted

Geo-temporal Movement and Events View

Delta-plots and heatmaps

- Use heatmaps of activity in NYC taxi or social media to summarize overall activity and hot spots
- Delta plots show differences from average behavior or previous time periods
  - See increases or decreases in typical behavior for selected time period
- Supports identifying abnormal patterns and spatial extents
- Filtering by frequency allows focusing in on areas of largest activity or change, e.g. traffic problems during construction at the airport or change in usage during a street festival

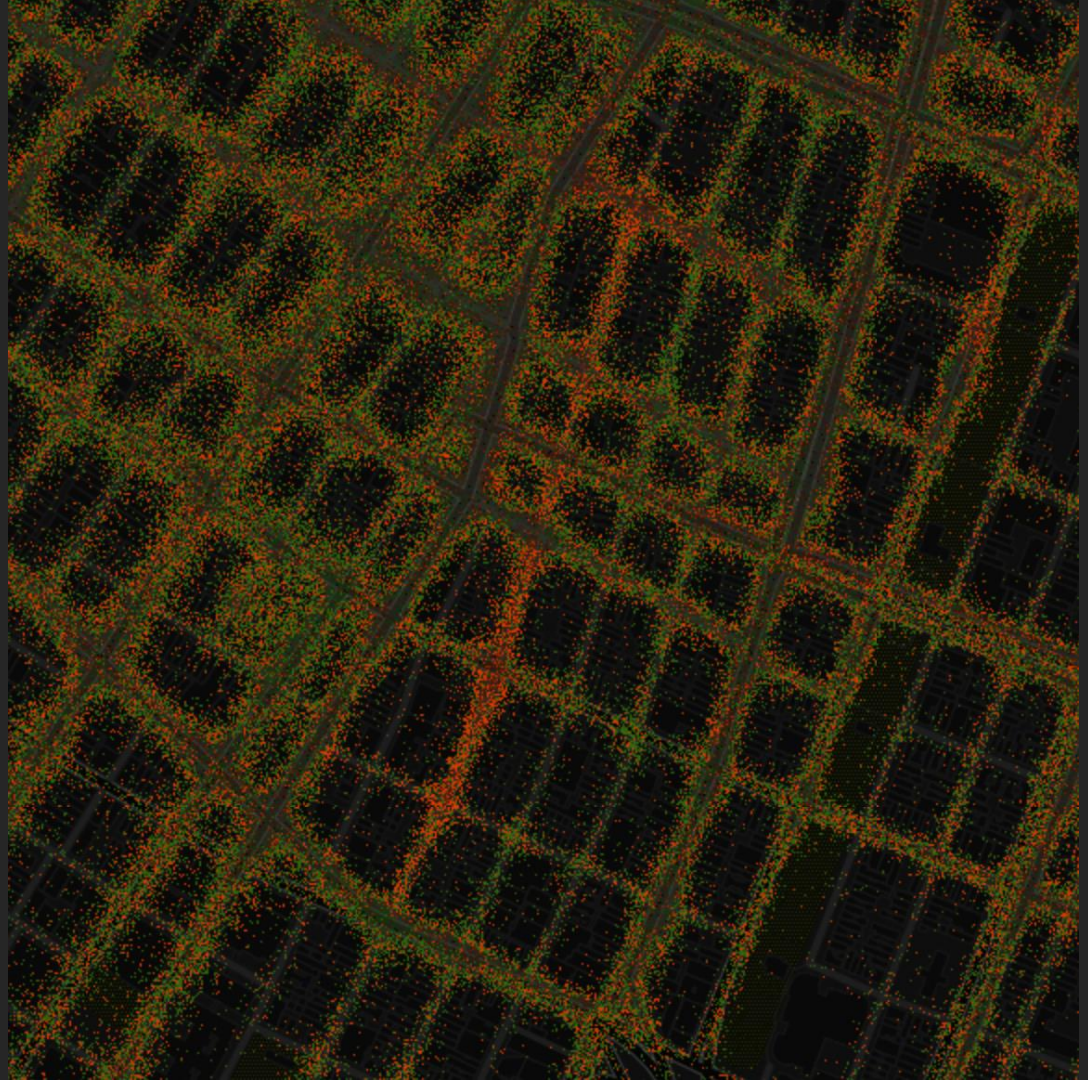


Polar color ramps show positive / negative difference from "normal" or pervious time periods

Heatmaps with a perceptually correct color ramp summarize overall activity

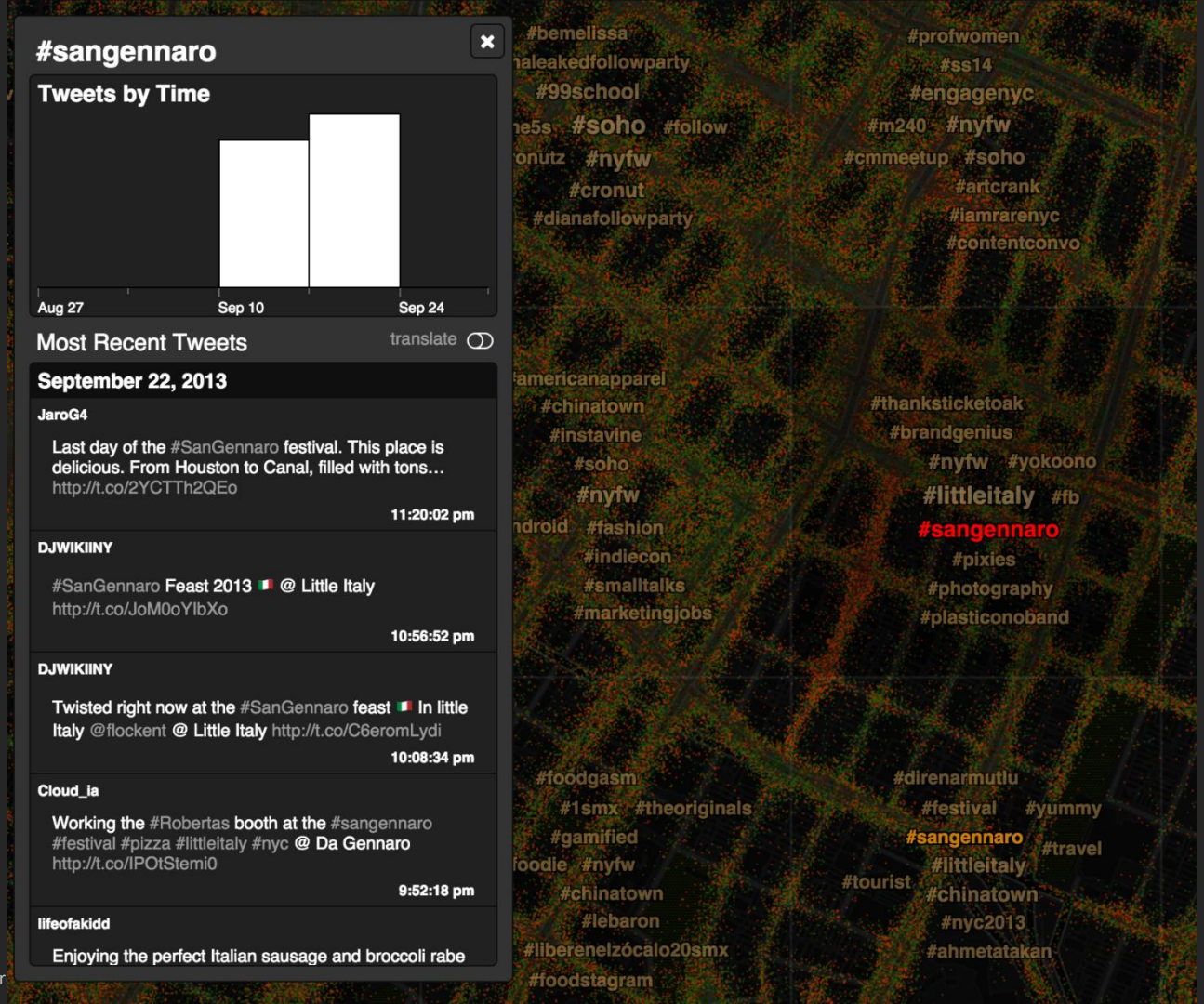
© 2017 Uncharted Software Inc.

The delta plot data is noisy, but still a red stripe is clearly visible in this plot indicating a big change such as an event occurring.



© 2017 Uncharted Software Inc.

The most frequent hashtags (TFIDF) can be used to quickly characterize the event – in this case a street festival.



© 2017 Uncharted Softwar

uncharted

6. LITIGIOUS TOPICS

© 2017 Uncharted Software Inc.

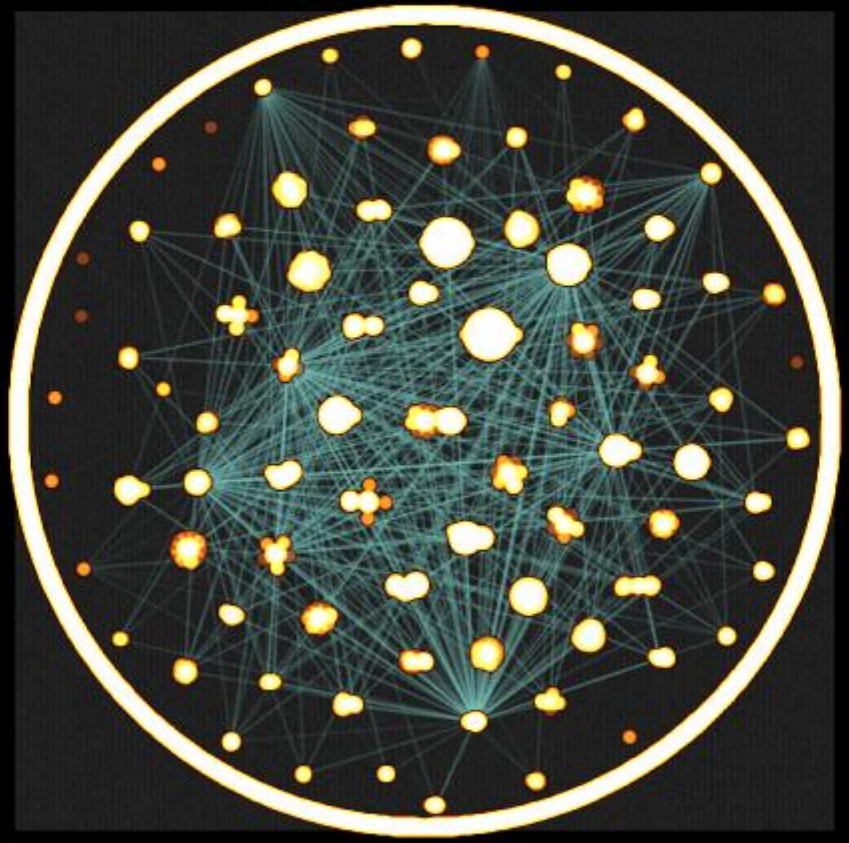
Intellectual Property

- Where are the hot spots?
  → What skills and expertise do we need?

- Where are the problem spots?
  → More diligent in crafting patent
  (more time, more resources)

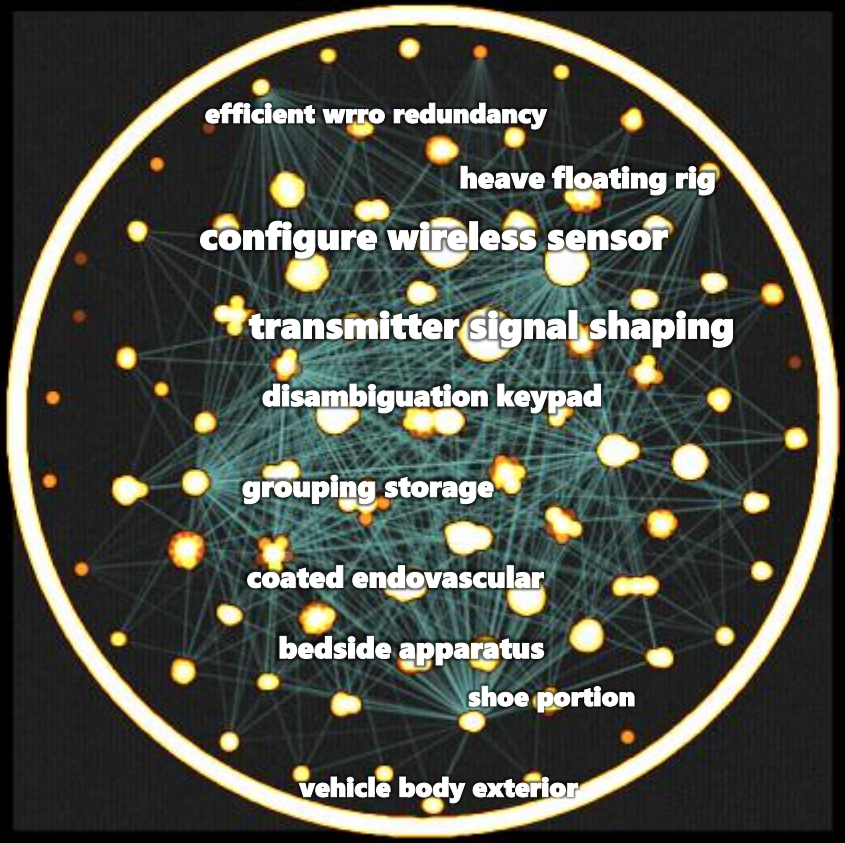
© 2017 Uncharted Software Inc.

uncharted



Plot showing every patent since 1982, grouped into communities based on each patent's citation of earlier relevant patents.

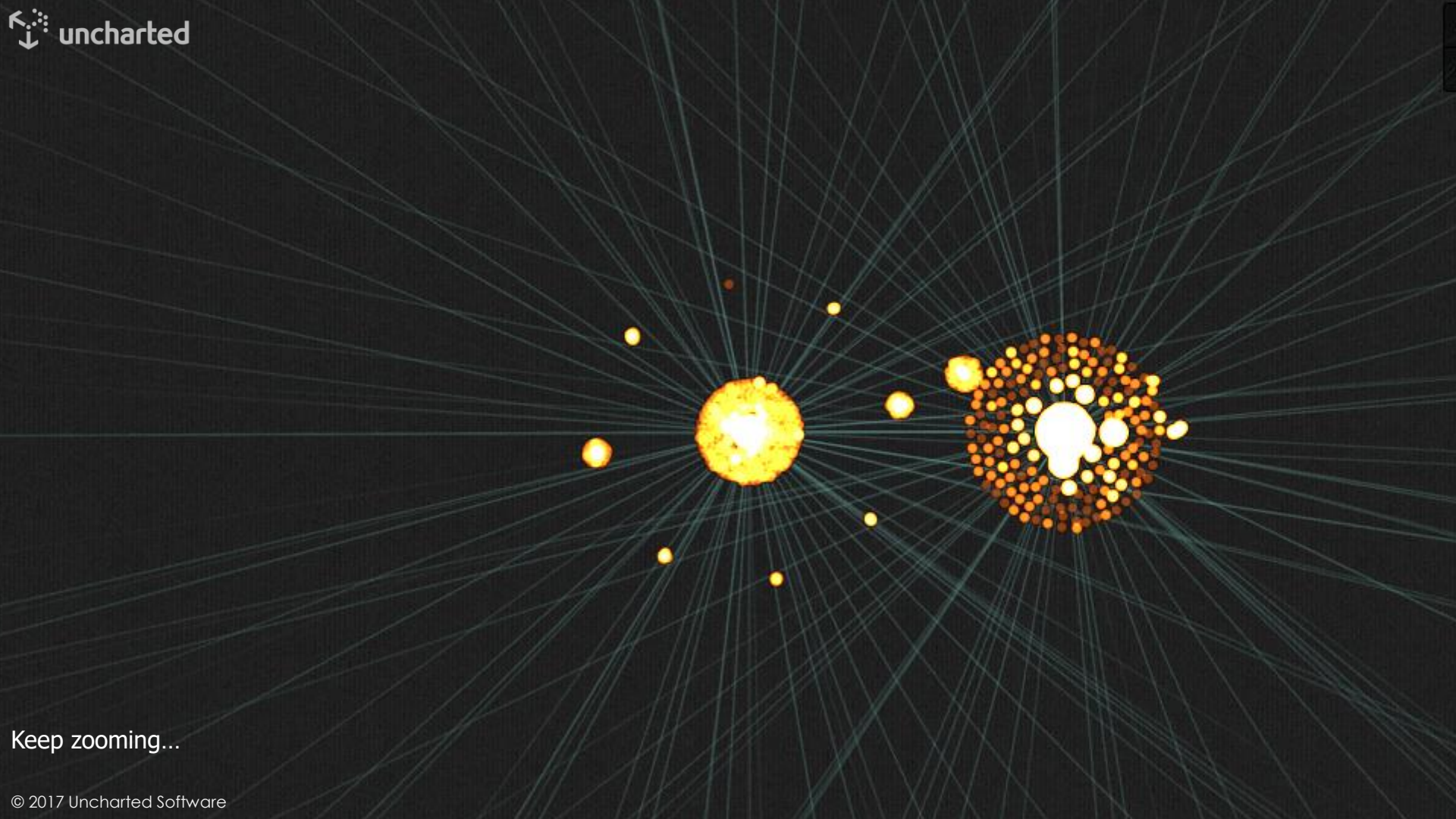
© 2017 Uncharted Software



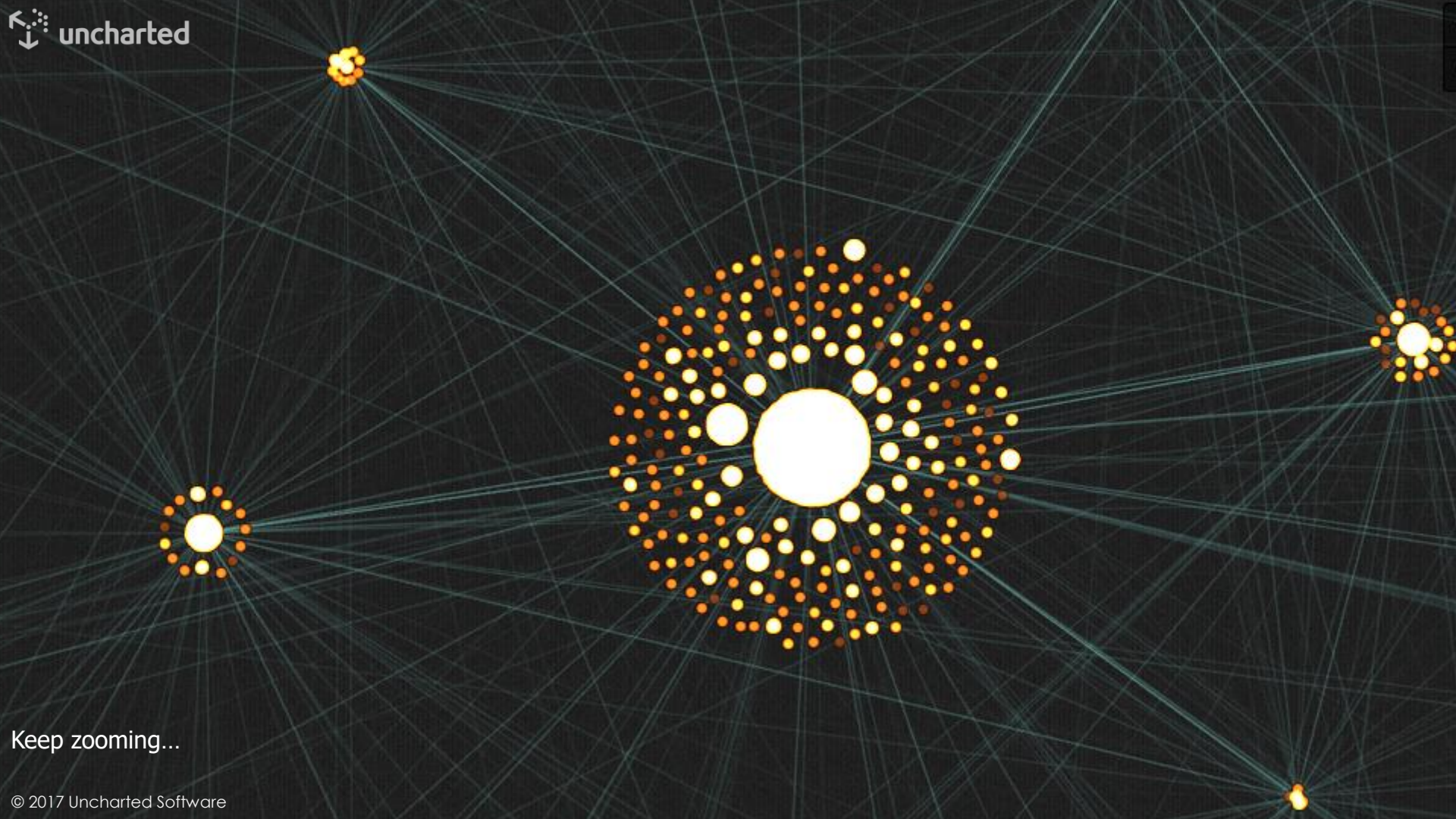
Labels, extracted using TFIDF, show the most unique terms on the most heavily cited patent in each community. There terms give an indication of the topics in each community.

© 2017 Uncharted Software

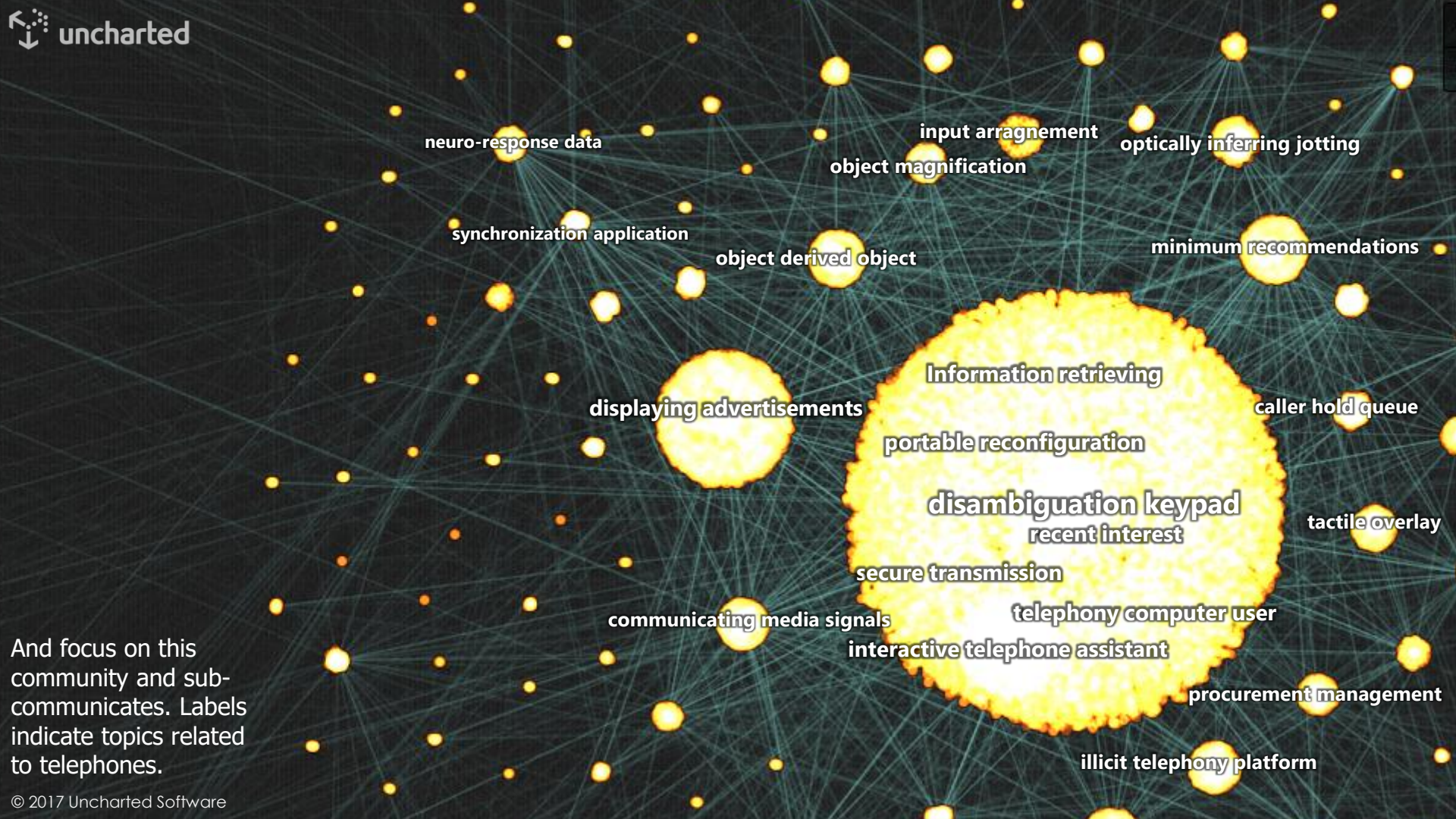
uncharted
Zoom in...
© 2017 Uncharted Software



© 2017 Uncharted Software

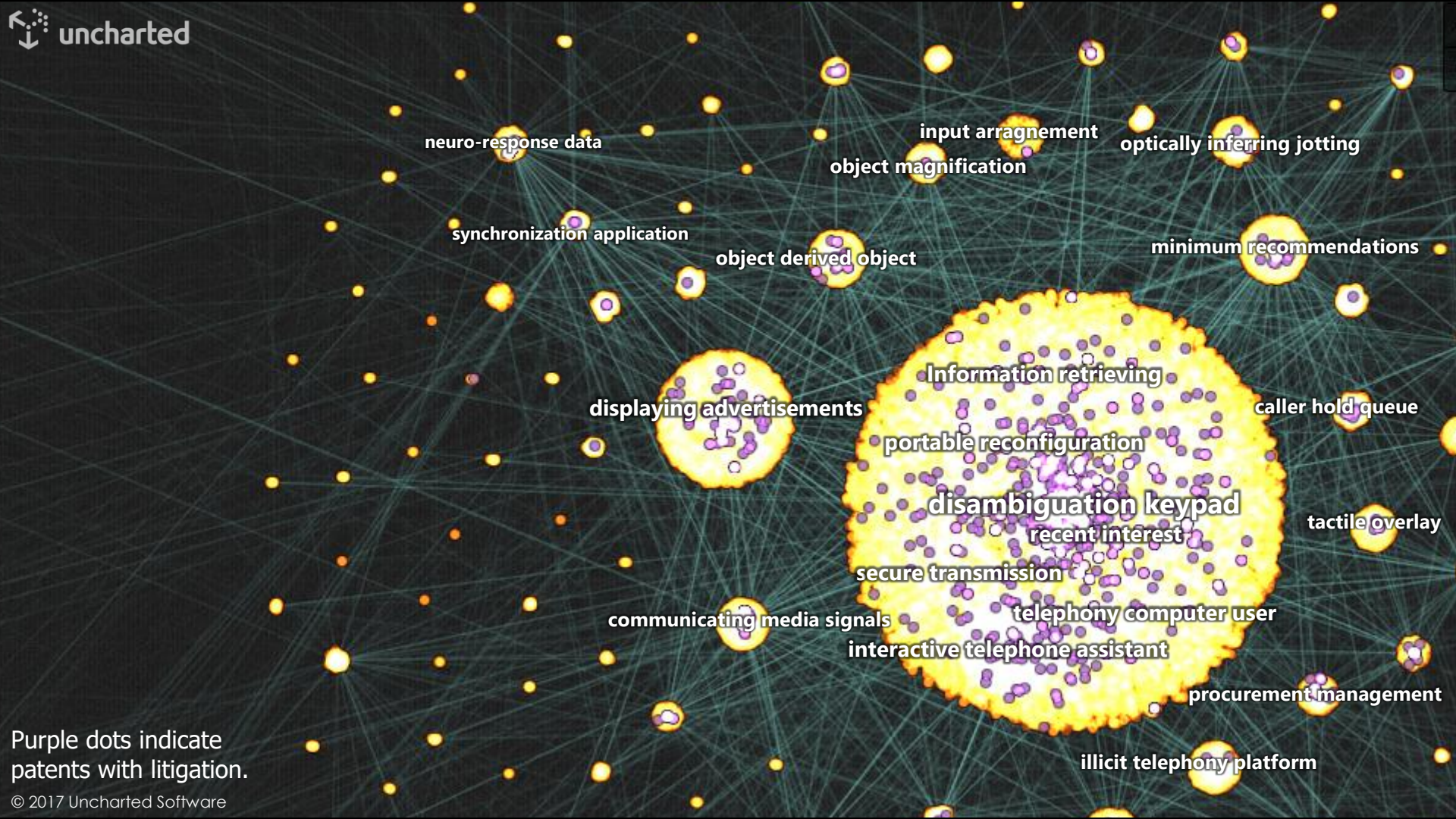


© 2017 Uncharted Software

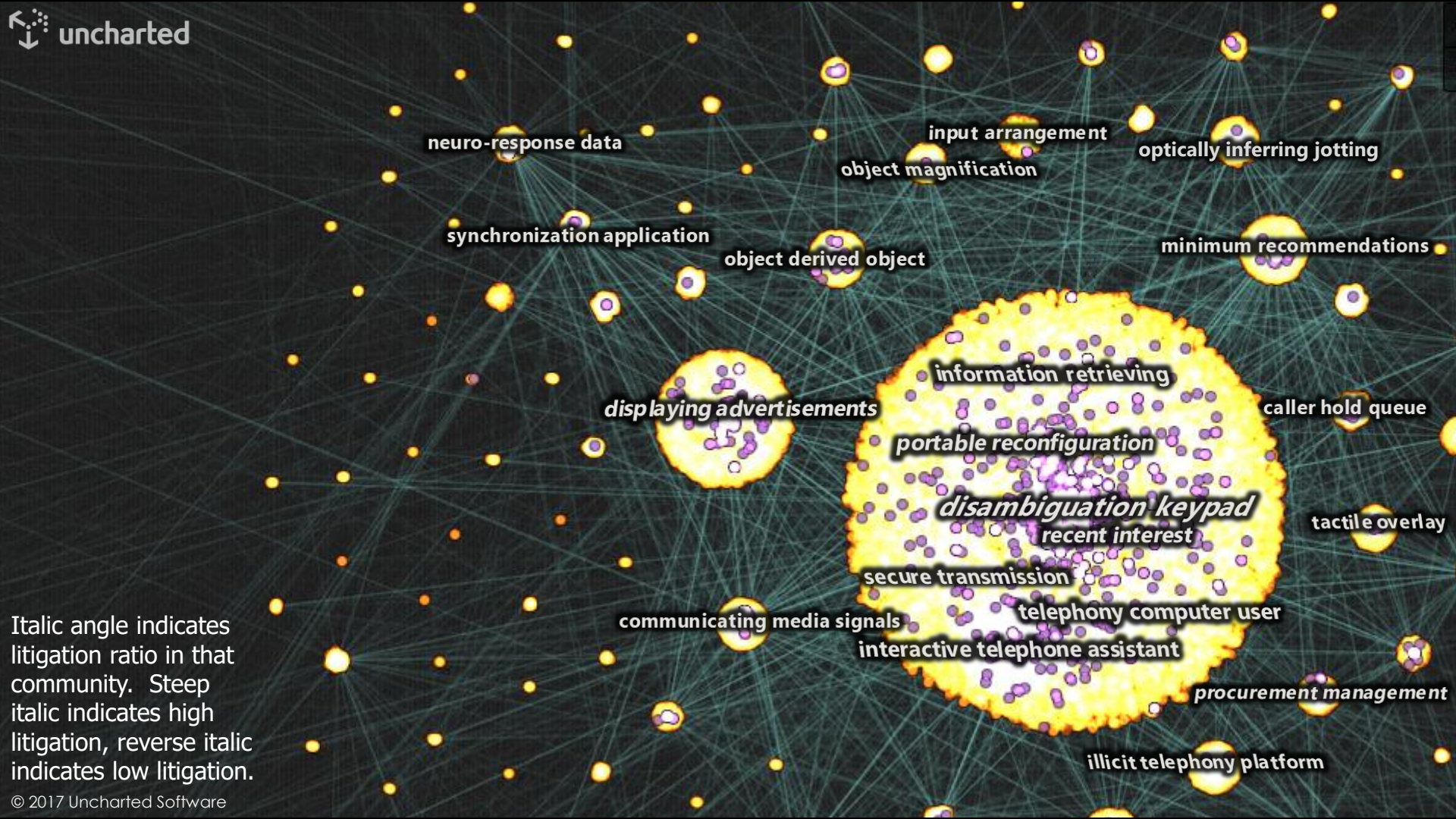


And focus on this community and sub-communicates. Labels indicate topics related to telephones.

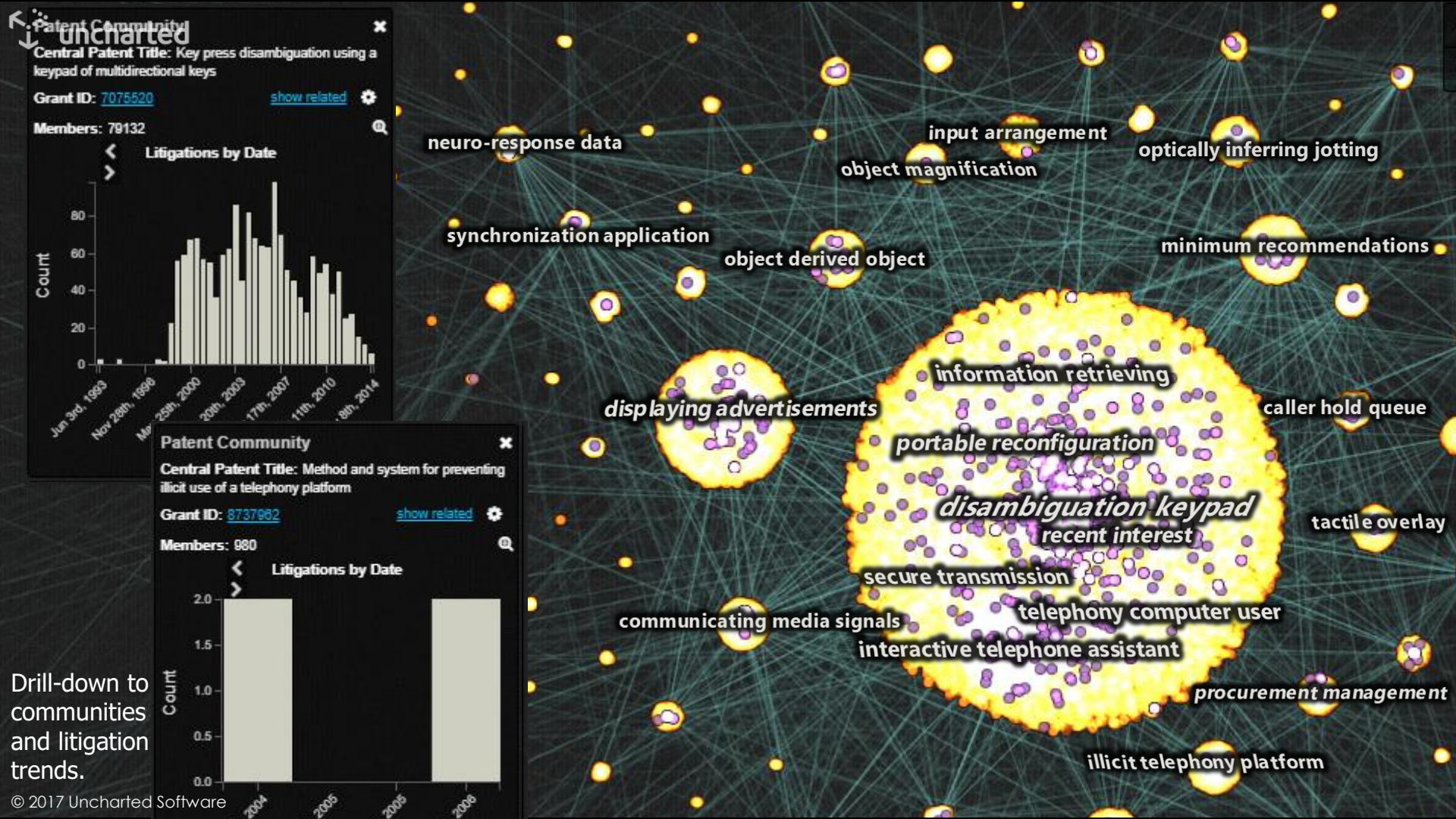
© 2017 Uncharted Software



© 2017 Uncharted Software



© 2017 Uncharted Software



Drill-down to communities and litigation trends.

© 2017 Uncharted Software

See citations to other communities.

© 2017 Uncharted Software

Google

Patents Application Grant Find prior art Discuss this patent View PDF Download PDF

Method and system for preventing illicit use of a telephony platform

US 8737962 B2

Publication number	US8737962 B2
Publication type	Grant
Application number	US 13/949,984
Publication date	27 May 2014
Filing date	24 Jul 2013
Priority date	24 Jul 2012
Also published as	US9270833, 5 More »
Inventors	Adam Ballai, Robert C. Hagemann, III, Daniel Zarick
Original Assignee	Twilio, Inc.
Export Citation	BiBTeX, EndNote, RefMan

Patent Citations (280), Non-Patent Citations (1), Referenced by (3), Classifications (11), Legal Events (1)

External Links: USPTO, USPTO Assignment, Espacenet

ABSTRACT

A system and method for preventing illicit use of a telephony platform that includes enrolling a plurality of accounts on a telecommunications platform, wherein an account includes account configuration; at a fraud detection system of the telecommunications platform, receiving account usage data, wherein the usage data includes at least communication configuration data and billing configuration data of account configuration and further includes communication history of the plurality of accounts; calculating fraud scores of a set of fraud rules from the usage data, wherein at least a sub-set of the fraud rules include conditions of usage data patterns between at least two accounts; detecting when the fraud scores of an account satisfy a fraud threshold; and initiating an action response when a fraud score satisfies the fraud threshold.

Drill to patent publication.

IMAGES (9)

6. SO WHAT?

© 2017 Uncharted Software Inc.

So What?

1. Visualization isn't just about numbers.
   Text repositories may be 4x or more > numeric data.

2. The next big leap in analytics and visualization will be much deeper integration of *content* throughout the full *analytics + visualization* analysis cycle.

3. Real-world applications
   - Tone analysis of your (external) communications
   - Ongoing stickiness of (marketing) messages
   - Local topics and event detection
   - Litigious communities

© 2017 Uncharted Software Inc.

Text Analytics and Visualization is Much Broader

Letters

Words

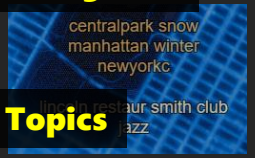
Phrases

Sentences

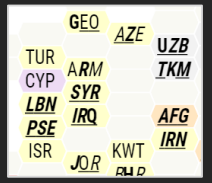
Paragraphs

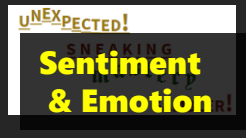
Documents

Categories

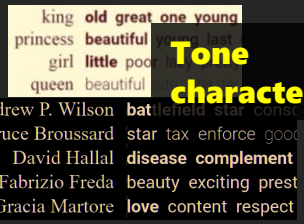


Topics





Sentiment & Emotion

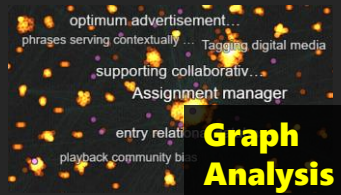


Tone characterization

Prosody Display

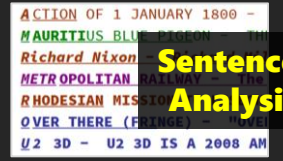
Social Stickiness

Classifiers & Sets



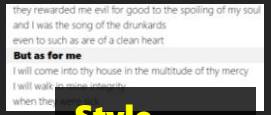
Graph Analysis

Proximity Analysis

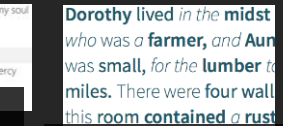


Sentence Analysis

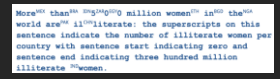
Entity Recognition & Tagging



Style Analysis



Semantics/ Summarization



Opinion Analysis

© 2017 Uncharted Software Inc.

uncharted

What can you do?

- Understand what text data you have
- Define what business insights you want to extract

- Use open source tools for analytics and visualization
  e.g. Saltlib, D3js, NLTK, Stanford Core NLP
- Incrementally experiment towards a solution
- Productionize the parts that are successful

© 2017 Uncharted Software Inc.

THANK YOU

Richard Brath

rbrath@unchartedsoftware.com
@rkbrath
richardbrath.wordpress.com

Scott Langevin

slangevin@unchartedsoftware.com
@slangevi
blog.uncharted.software

© 2017 Uncharted Software Inc.

uncharted

About Uncharted: Profoundly visual software for improved awareness, analysis and decision-making, used by hundreds of thousands in government, law enforcement, finance, and more.

© 2017 Uncharted Software Inc.