Visualización de Información IIC2026

Visualización de Texto II

Denis Parra
(algunas slides de Lucas Valenzuela)

En esta clase, ejemplos y casos de estudio de visualización de texto, documents, corpora (document corpus)

Word Clouds

summaries
keyword
set
visual
data
design
present
frequency
aimed
document
peoples
performance
given
collection
work
word
clouds
paper
generate
sets
people
derive
use
often
intuition
values
user
call
exploring
lists
four
information
value
quantitative
used
extracting
words
studies
associated

Ejemplos

- http://www.science.smith.edu/dftwiki/index.php/Visualizations:_Lexical/Text
- Wordle.net

Ejemplos

- http://www.science.smith.edu/dftwiki/index.php/Visualizations:_Lexical/Text

Category: Lexical/Text
Author/Source: Money.cnn.com
Implementation: NA
Date: Feb. 2011

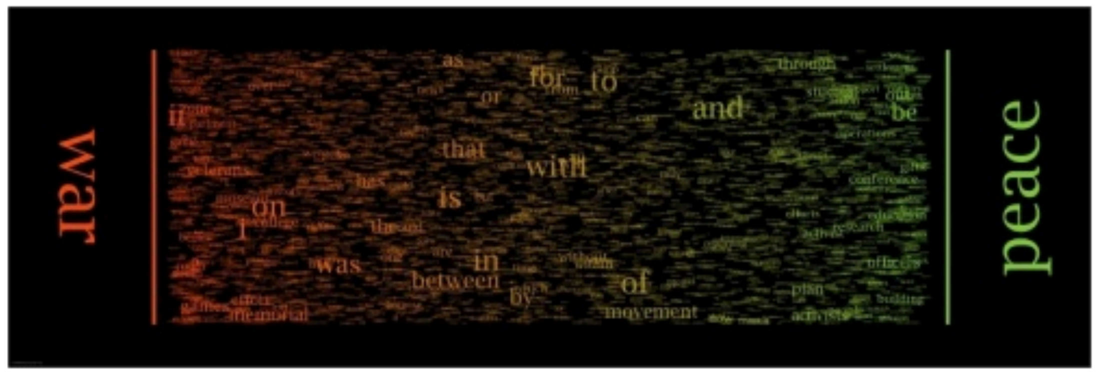
The best companies to work at, in the States. The left diagram shows the companies, the bubble size indicating the measure of goodness: large = good. The chart on the right indicates the most frequent words used when describing these companies. When one clicks on a word, sample phrases containing the word fly by...

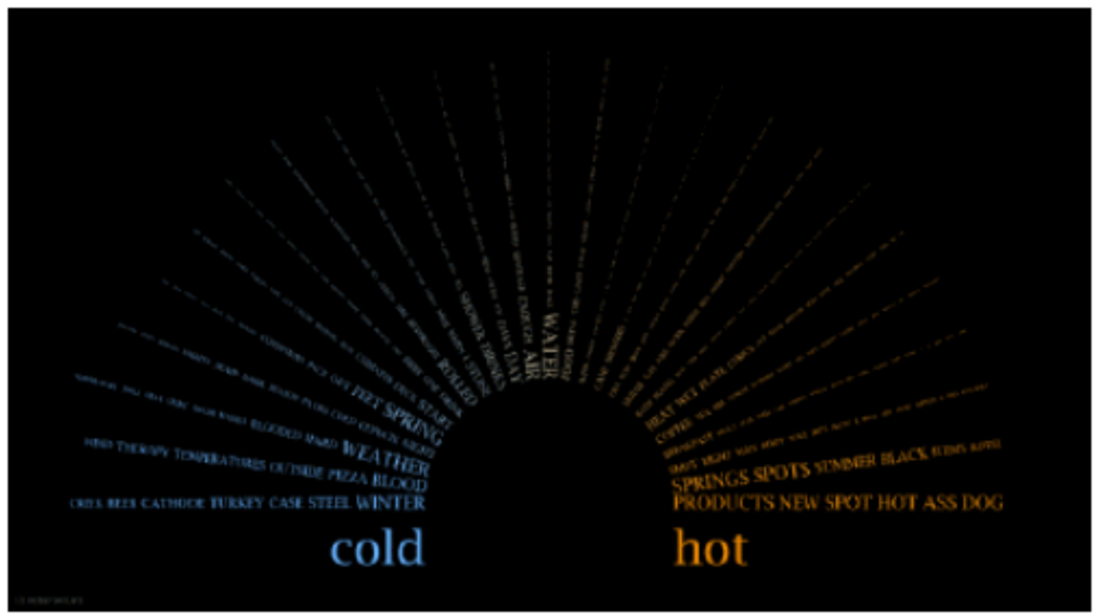
Ejemplos

- http://www.science.smith.edu/dftwiki/index.php/Visualizations:_Lexical/Text

Category: Lexical/Text
Author/Source: chrisharrison.net
Implementation: NA
Date: 2010

From Chris Harrison's page: *Using Google's enormous bigram dataset, I produced a series of visualizations that explore word associations. Each visualization pits two primary terms against each other. Then, the use frequency of words that follow these two terms are analyzed. For example, "war memorial" occurs 531,205 times, while "peace memorial" occurs only 25,699. A position for each word is generated by looking at the ratio of the two frequencies. If they are equal, the word is placed in the middle of the scale. However, if there is a imbalance in the uses, the word is drawn towards the more frequently related term. This process is repeated for thousands of other word combinations, creating a spectrum of word associations. Font size is based on a inverse power function (uniquely set for each visualization, so you can't compare across pieces). Vertical positioning is random.*





Felix, C., Franconeri, S., & Bertini, E. (2017).
Taking Word Clouds Apart: An Empirical Investigation of the Design Space for Keyword Summaries.
IEEE Transactions on Visualization and Computer Graphics.

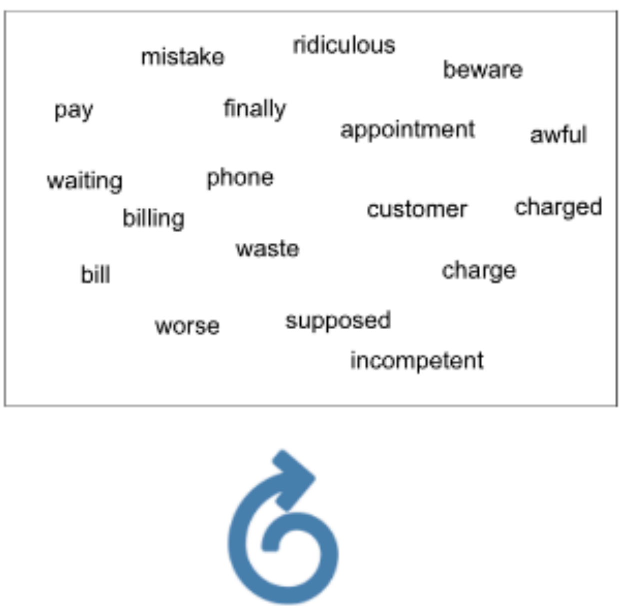
Espacio de Diseño

- Layout: Estrategia usada para posicionar las palabras

- Codificación visual: Canal visual usada para codificar magnitude

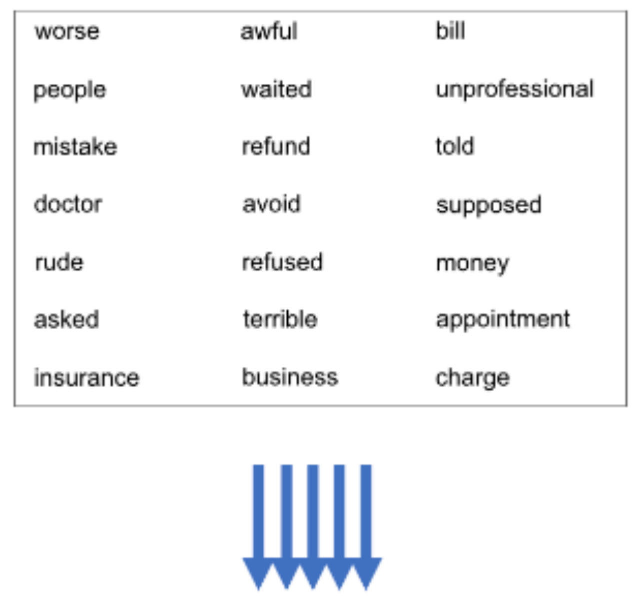
Layouts

- Spatial: Palabras ubicadas sin alineación particular
- Row: Palabras alineadas respecto a múltiples filas
- Column: Palabras alineadas respecto a múltiples columnas



Spatial

Row

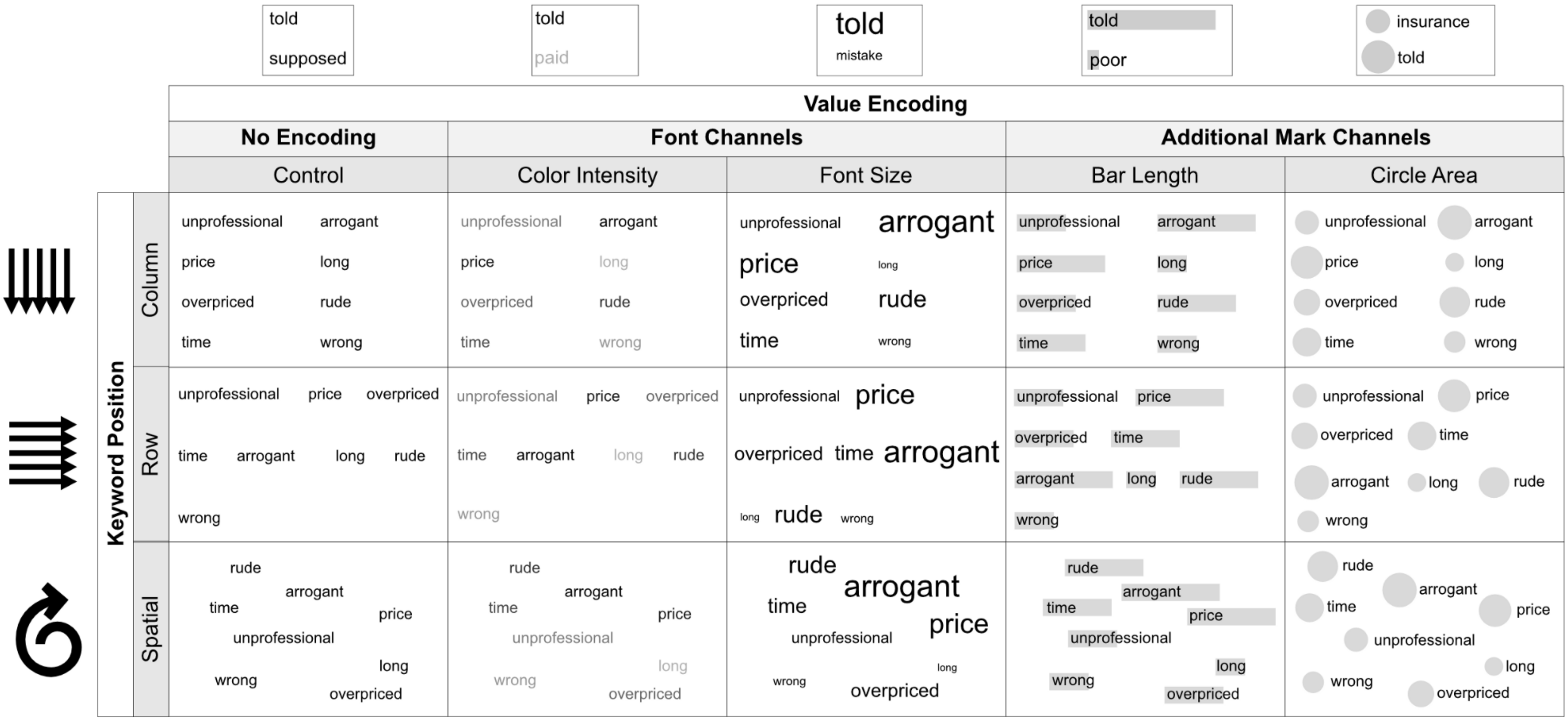


Column

Estrategias de codificación

- **Tamaño de Fuente**
- **Intensidad de Color**
- **Largo de Barra**
- **Tamaño de Círculo**
- **No Encoding**

Espacio de Diseño Combinado



¿Cuál Funciona mejor?

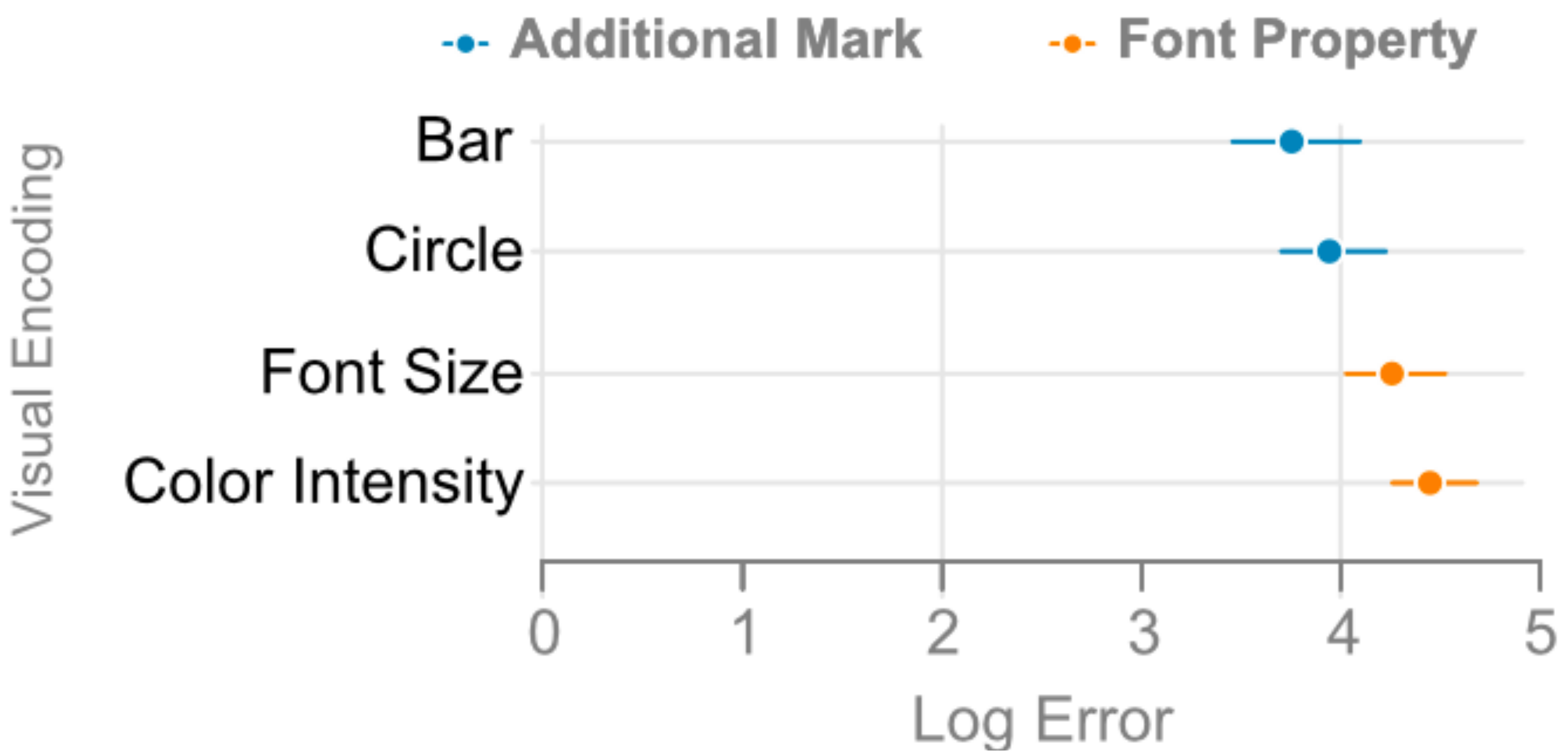
¿mejor para qué?

Tareas

- **Comparar Valores:** Comparar magnitudes de dos palabras seleccionadas
- **Buscar Palabras:** Buscar por una palabra específica en el cloud
- **Identificar Tópicos:** Elegir a qué tópico corresponde una palabra (de una lista)
- **Construir Tópicos:** Describir qué tópicos hay incluidos en una word cloud.

Resultados

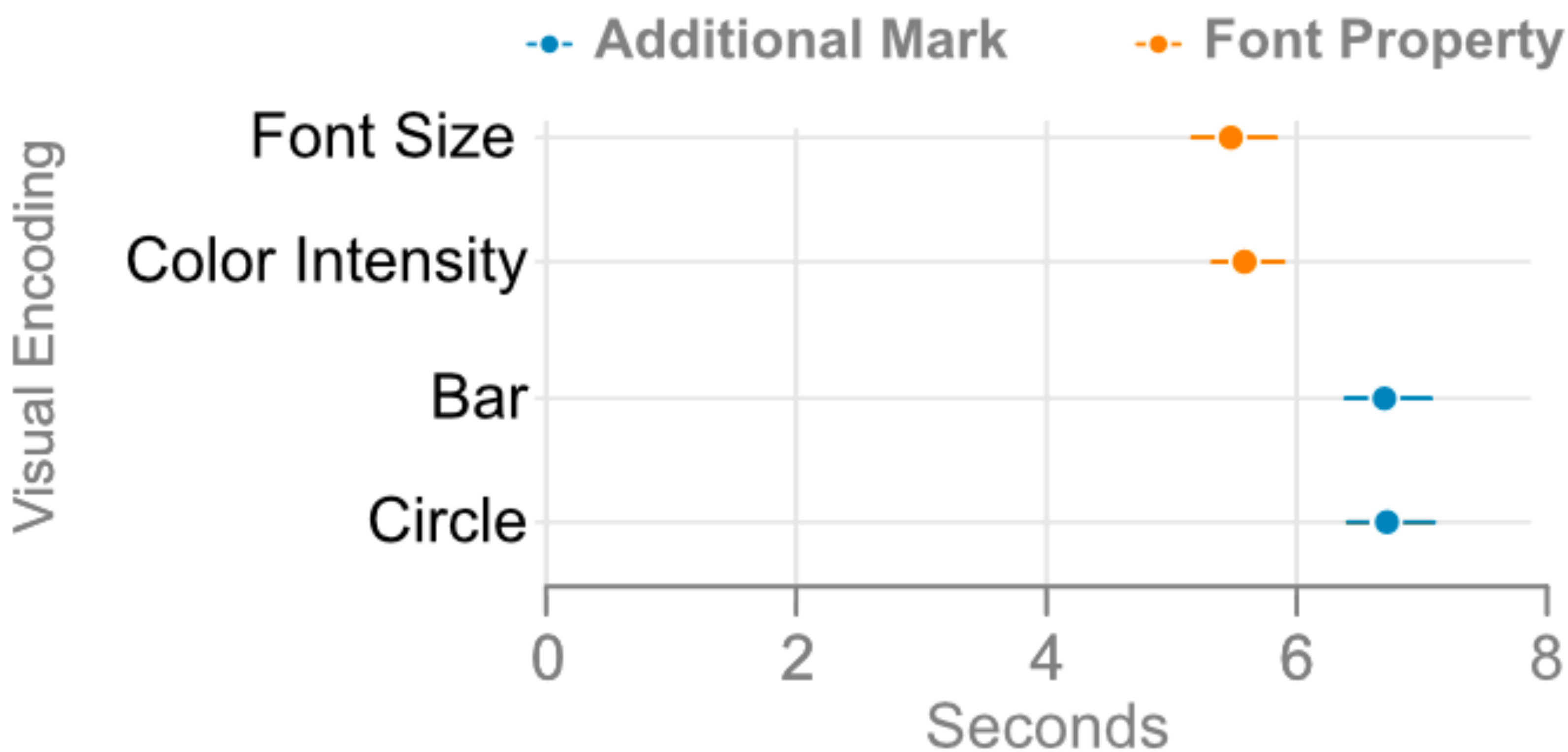
● **Barras alineadas son mejores para leer valores con precisión** (al comparar palabras)



The amount of error we found in different designs, when participants are asked to compare magnitudes associated to two words. Bar and circle score better than font size and color intensity.

Resultados

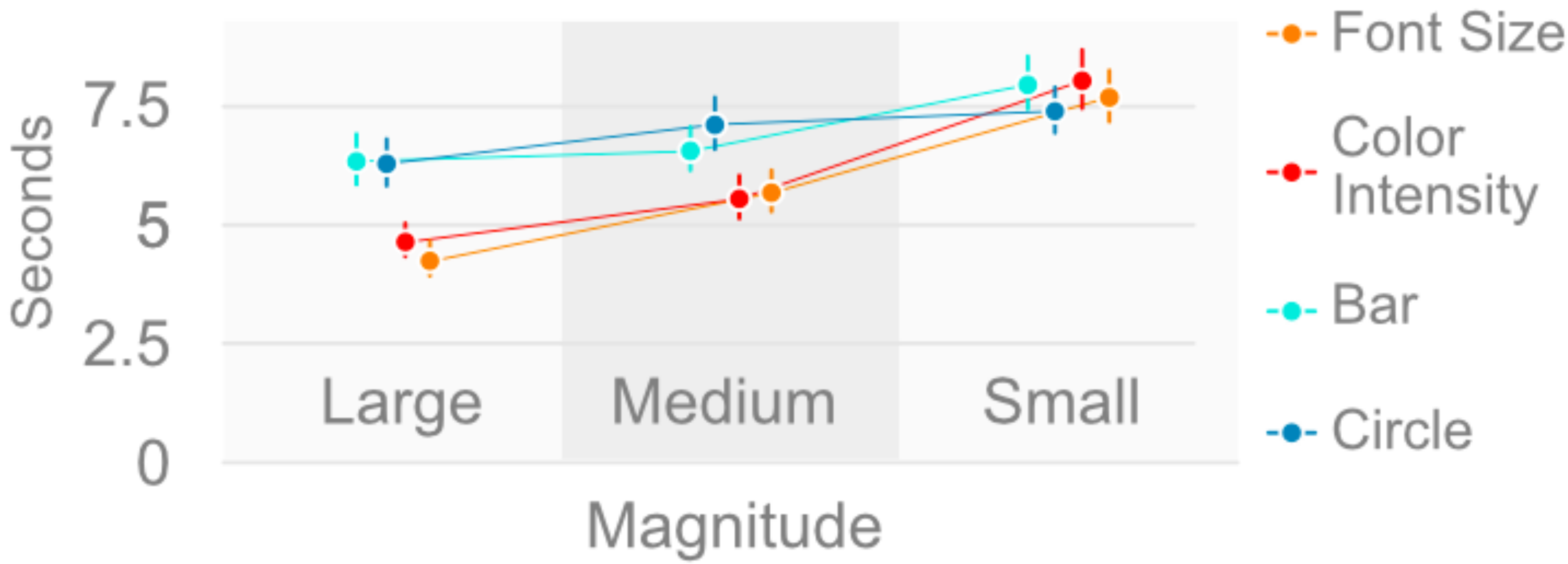
- **Tamaño de la fuente e intensidad de color funcionan mejor para búsqueda de palabras**



The amount of time it took participants to find specific words with different designs. Font size and color intensity score better than bar and circle.

Resultados

● Ventaja de la “propiedades de fuente” se desvanecen cuando la pabra objetivo e pequeña

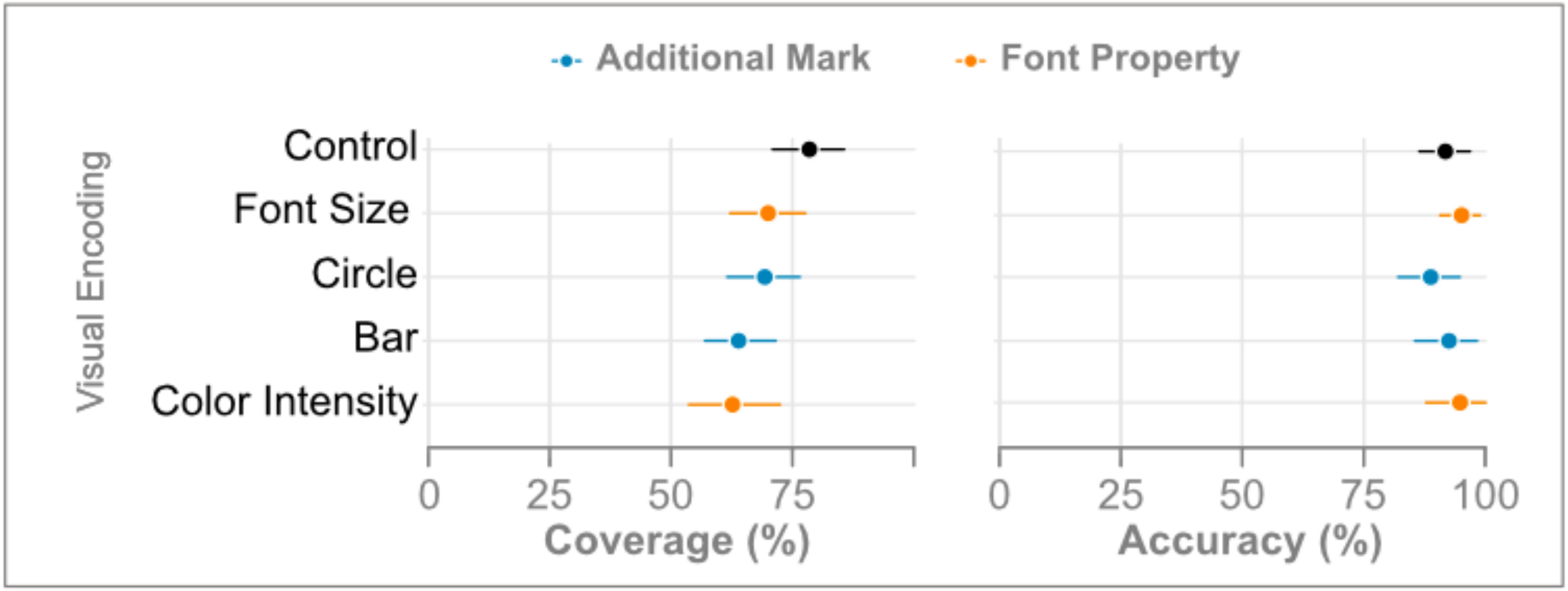


Time it takes to find a word when associated to different magnitudes. The advantage of font properties vanished when the target gets smaller.

Resultados

● En Tareas más complejas, los resultados no se ven igualmente reflejados (tópicos)

●Listas simples de palabras pueden funcionar bastante bien



Coverage and accuracy of topics detected when asked to identify topics in the word cloud. Performance is similar across all conditions. Simple lists (control) work pretty well.

Recomendaciones Finales

- **Experimentar con diseños diferentes.**

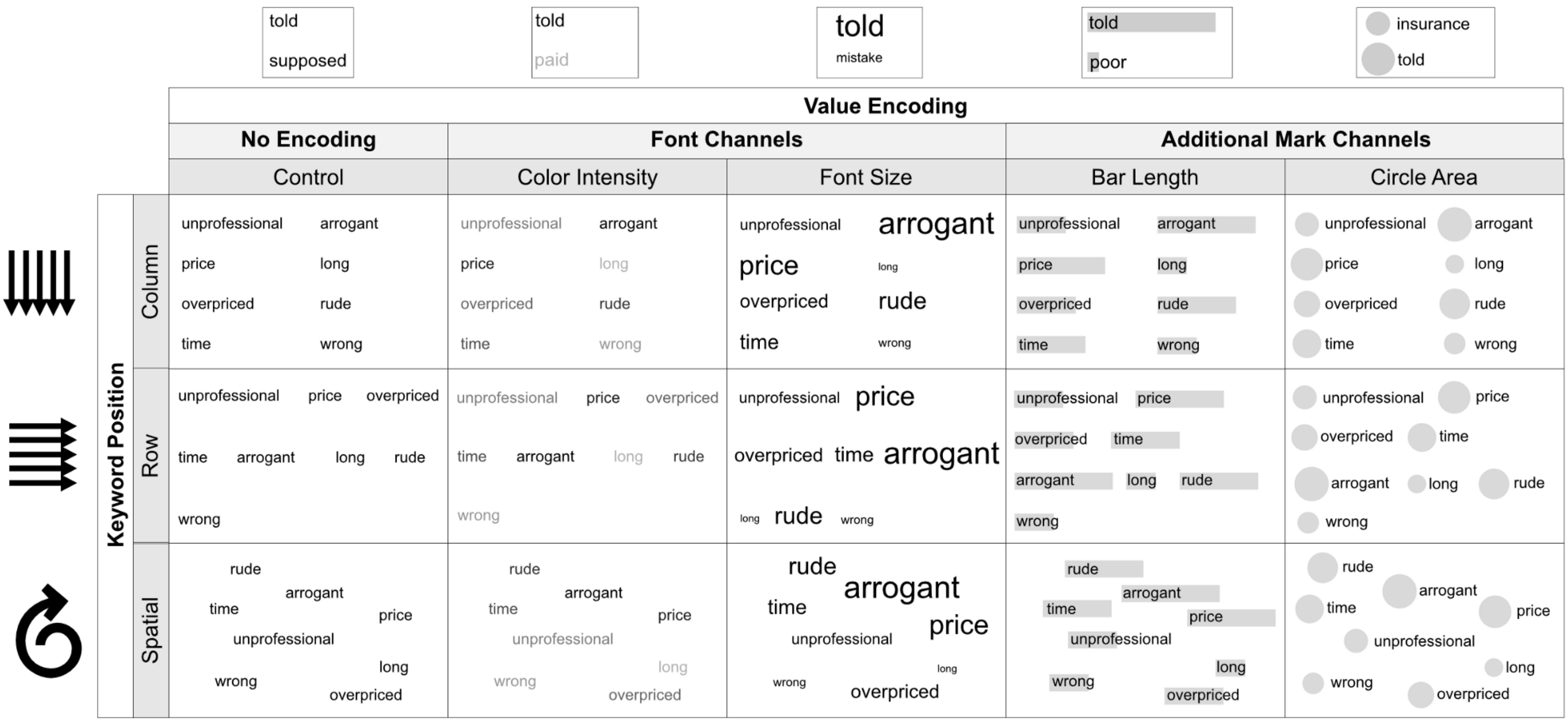
- **Give column bars a chance.** Agregar barras a las palabras, alinear en columnas y permitir al usuarios cierta interacción.

- **Probar otras disposiciones espaciales.** Manteniendo tamaño de fuente e intesidad de color parece funcionar bien para algunas tareas, pero se puede experimentar con otros *layout*.

- **Listas simples de palabras funcionan bien.** No se ven muy fashion, pero especialmente en tareas como encontrar tópicos o temas, parecen funcionar bien.

- **Pensar en ordenar.** En una aplicación interactiva, puede ser buena idea agregar funcionalidad para ordernar alfabéticamente o por tamaño. Esto puede ayudar a responder diferentes preguntas en un ambiente de análisis de datos.

Espacio de Diseño Combinado



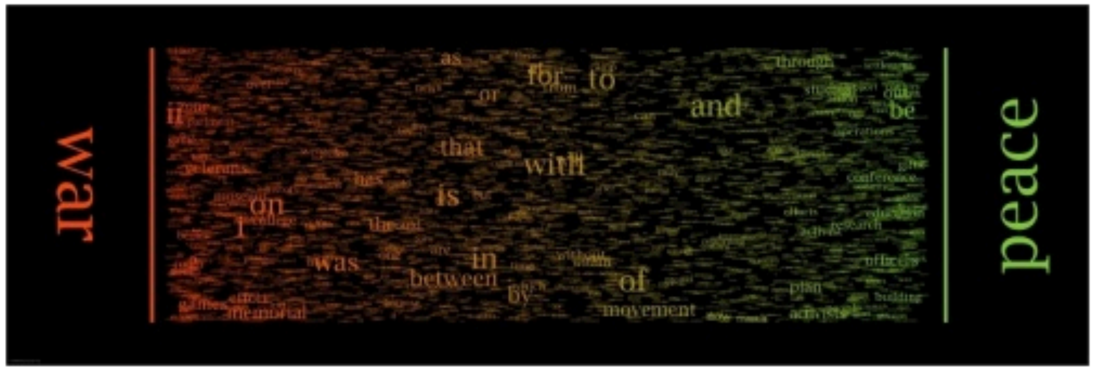
¿Cómo medir **asociación**
entre palabras y conceptos ?
(para usarlos en visualización)

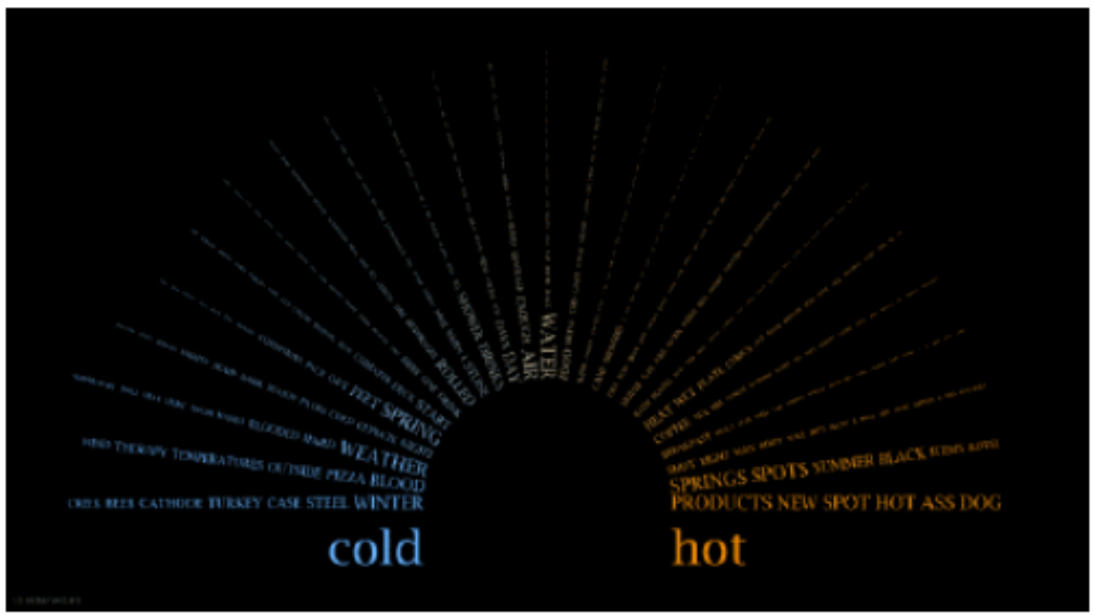
Ejemplos

- http://www.science.smith.edu/dftwiki/index.php/Visualizations:_Lexical/Text

Category: Lexical/Text
Author/Source: chrisharrison.net
Implementation: NA
Date: 2010

From Chris Harrison's page: *Using Google's enormous bigram dataset, I produced a series of visualizations that explore word associations. Each visualization pits two primary terms against each other. Then, the use frequency of words that follow these two terms are analyzed. For example, "war memorial" occurs 531,205 times, while "peace memorial" occurs only 25,699. A position for each word is generated by looking at the ratio of the two frequencies. If they are equal, the word is placed in the middle of the scale. However, if there is a imbalance in the uses, the word is drawn towards the more frequently related term. This process is repeated for thousands of other word combinations, creating a spectrum of word associations. Font size is based on a inverse power function (uniquely set for each visualization, so you can't compare across pieces). Vertical positioning is random.*





Ejemplos

- Wagner, C., Graells-Garrido, E., Garcia, D., & Menczer, F. (2016). Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Science*, *5*(1), 5.

Figure 4 Topical bias. Word clouds for biographies of women (top) and men (bottom), with birth date before 1900 (left) and since 1900 (right). Spaces in bi-grams are replaced with an underscore. Font size is proportional to PMI with each gender. Colors depict the four categories: *gender* in orange, *family* in green, *relationship* in violet, and *other* in blue. Beside professional and topical areas, words in the gender, relationship, and family categories are more dominant in articles about women born before 1900. Gender-specific differences are much less pronounced in articles about people born since 1900.

PMI

- Frecuencia de las palabras no es una buena medida de asociación entre palabras (sesgada) – “el” y “de” pueden ser frecuentes, pero poco asociadas
- Medida de asociación entre dos eventos
- Dos eventos co-ocurren más que si fueran independientes?

$$\mathrm{PMI}(X,Y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

Vector Semantics

Positive Pointwise Mutual Information (PPMI)

Pointwise Mutual Information

Pointwise mutual information:

Do events x and y co-occur more than if they were independent?

$$\text{PMI}(X,Y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

PMI between two words: (Church & Hanks 1989)

Do words x and y co-occur more than if they were independent?

$$\text{PMI}(word_1, word_2) = \log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}$$

Positive Pointwise Mutual Information

- PMI ranges from $-\infty$ to $+\infty$
- But the negative values are problematic
  - Things are co-occurring **less than** we expect by chance
  - Unreliable without enormous corpora
    - Imagine w1 and w2 whose probability is each 10^{-6}
    - Hard to be sure p(w1,w2) is significantly different than 10^{-12}
  - Plus it’s not clear people are good at “unrelatedness”

Positive Pointwise Mutual Information

- So we just replace negative PMI values by 0
- Positive PMI (PPMI) between word1 and word2:

$$\mathrm{PPMI}(word_1, word_2) = \max\left(\log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}, 0\right)$$

Computing PPMI on a term-context matrix

- Matrix F with W rows (words) and C columns (contexts)
- f_{ij} is # of times w_i occurs in context c_j

	aardvark	computer	data	pinch	result	sugar
apricot	0	0	0	1	0	1
pineapple	0	0	0	1	0	1
digital	0	2	1	0	1	0
information	0	1	6	0	4	0

$$p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}} \qquad p_{i*} = \frac{\sum_{j=1}^{C} f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}} \qquad p_{*j} = \frac{\sum_{i=1}^{W} f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}}$$

$$pmi_{ij} = \log_2 \frac{p_{ij}}{p_{i*}p_{*j}} \qquad ppmi_{ij} = \begin{cases} pmi_{ij} & \text{if } pmi_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}}$$

Count(w,context)

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

p(w=information,c=data) = 6/19 = .32

p(w=information) = 11/19 = .58

p(c=data) = 7/19 = .37

$$p(w_i) = \frac{\sum_{j=1}^{C} f_{ij}}{N} \qquad p(c_j) = \frac{\sum_{i=1}^{W} f_{ij}}{N}$$

p(w,context) / **p(w)**

	computer	data	pinch	result	sugar	p(w)
apricot	0.00	0.00	0.05	0.00	0.05	0.11
pineapple	0.00	0.00	0.05	0.00	0.05	0.11
digital	0.11	0.05	0.00	0.05	0.00	0.21
information	0.05	0.32	0.00	0.21	0.00	0.58
p(context)	0.16	0.37	0.11	0.26	0.11	

$$pmi_{ij} = \log_2 \frac{p_{ij}}{p_{i*} p_{*j}}$$

	p(w,context)					p(w)
	computer	data	pinch	result	sugar	
apricot	0.00	0.00	0.05	0.00	0.05	0.11
pineapple	0.00	0.00	0.05	0.00	0.05	0.11
digital	0.11	0.05	0.00	0.05	0.00	0.21
information	0.05	0.32	0.00	0.21	0.00	0.58
p(context)	0.16	0.37	0.11	0.26	0.11	

- pmi(information,data) = $\log_2$ (.32 / (.37*.58)) = .58 *(.57 using full precision)*

PPMI(w,context)

	computer	data	pinch	result	sugar
apricot	-	-	2.25	-	2.25
pineapple	-	-	2.25	-	2.25
digital	1.66	0.00	-	0.00	-
information	0.00	0.57	-	0.47	-

Weighting PMI

- PMI is biased toward infrequent events
  - Very rare words have very high PMI values
- Two solutions:
  - Give rare words slightly higher probabilities
  - Use add-one smoothing (which has a similar effect)

Weighting PMI: Giving rare context words slightly higher probability

- Raise the context probabilities to $\alpha = 0.75$:

$$\mathrm{PPMI}_{\alpha}(w,c) = \max(\log_2 \frac{P(w,c)}{P(w)P_{\alpha}(c)},0)$$

$$P_{\alpha}(c) = \frac{count(c)^{\alpha}}{\sum_c count(c)^{\alpha}}$$

- This helps because $P_{\alpha}(c) > P(c)$ for rare *c*

- Consider two events, P(a) = .99 and P(b)=.01

- $P_{\alpha}(a) = \frac{.99^{.75}}{.99^{.75}+.01^{.75}} = .97 \quad P_{\alpha}(b) = \frac{.01^{.75}}{.01^{.75}+.01^{.75}} = .03$

Use Laplace (add-1) smoothing

Add-2 Smoothed Count(w,context)

	computer	data	pinch	result	sugar
apricot	2	2	3	2	3
pineapple	2	2	3	2	3
digital	4	3	2	3	2
information	3	8	2	6	2

p(w,context) [add-2]

	computer	data	pinch	result	sugar	**p(w)**
apricot	0.03	0.03	0.05	0.03	0.05	0.20
pineapple	0.03	0.03	0.05	0.03	0.05	0.20
digital	0.07	0.05	0.03	0.05	0.03	0.24
information	0.05	0.14	0.03	0.10	0.03	0.36
p(context)	0.19	0.25	0.17	0.22	0.17	

PPMI versus add-2 smoothed PPMI

PPMI(w,context)

	computer	data	pinch	result	sugar
apricot	-	-	2.25	-	2.25
pineapple	-	-	2.25	-	2.25
digital	1.66	0.00	-	0.00	-
information	0.00	0.57	-	0.47	-

PPMI(w,context) [add-2]

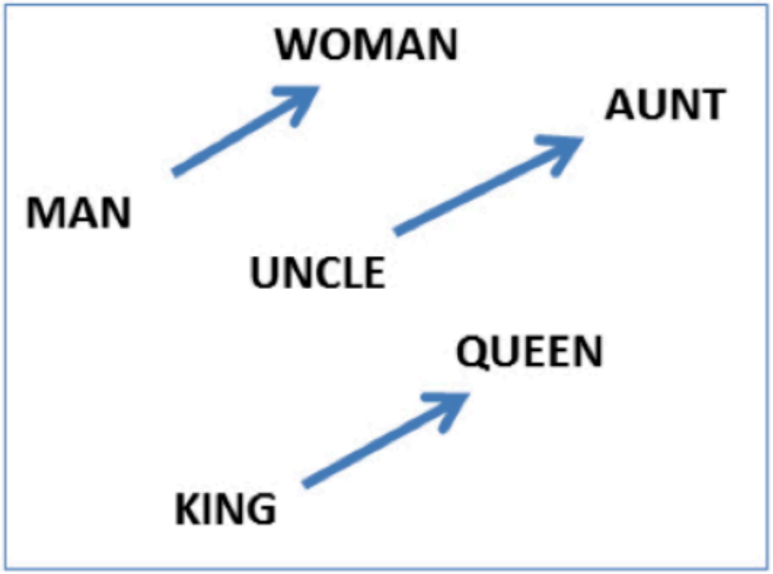
	computer	data	pinch	result	sugar
apricot	0.00	0.00	0.56	0.00	0.56
pineapple	0.00	0.00	0.56	0.00	0.56
digital	0.62	0.00	0.00	0.00	0.00
information	0.00	0.58	0.00	0.37	0.00

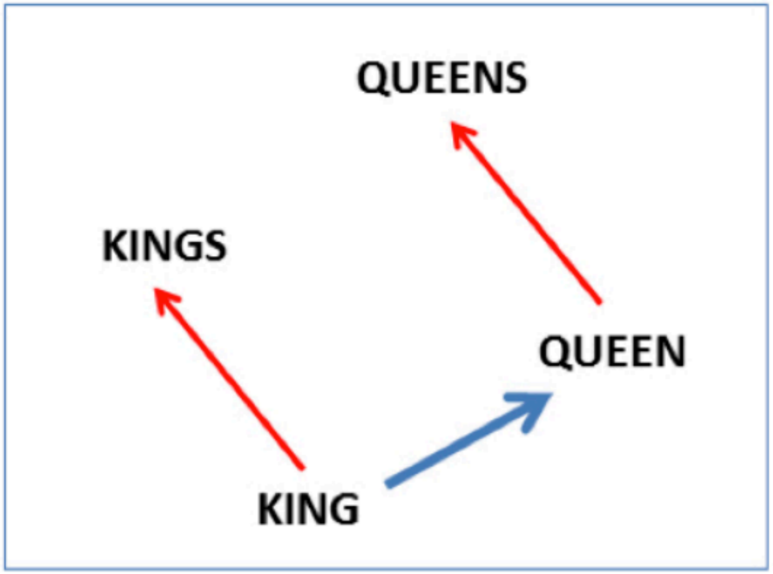
¿Puede obtenerse y visualizarse semática?

Word2vec Embeddings (Mikolov, 2010)

vector(*'king'*) - vector(*'man'*) + vector(*'woman'*) ≈ vector('queen')

vector(*'Paris'*) - vector(*'France'*) + vector(*'Italy'*) ≈ vector('Rome')

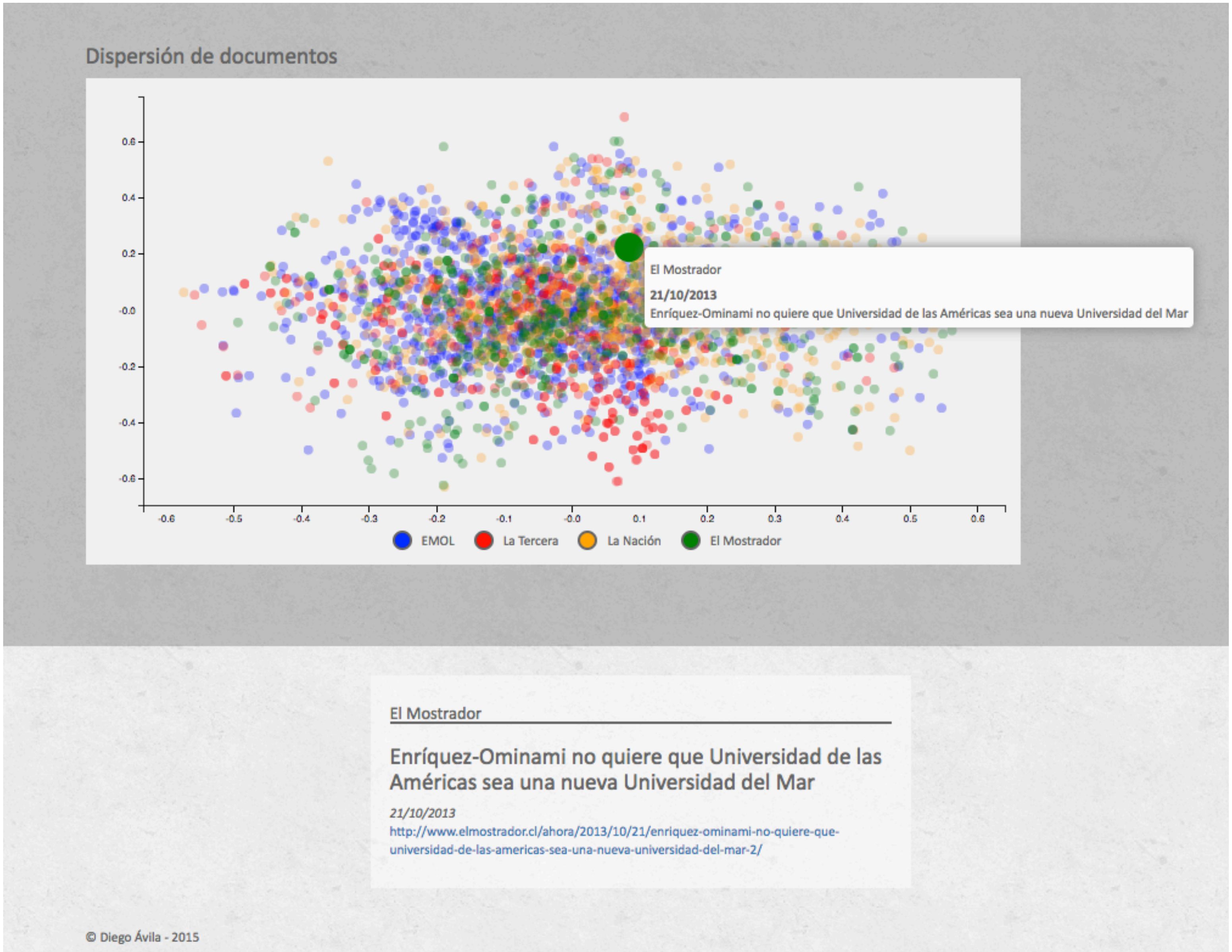




¿Cómo visualizar un corpus?

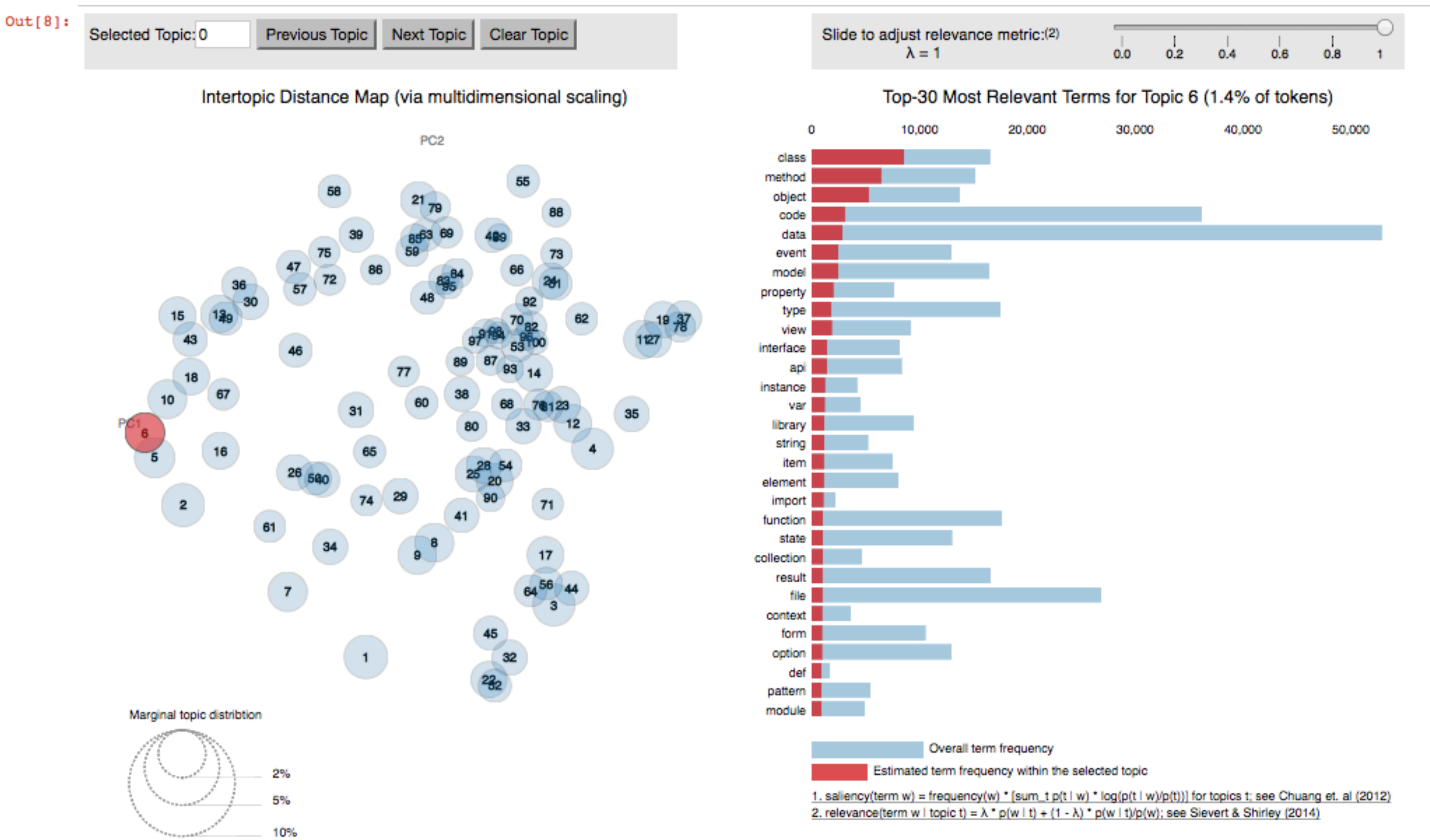
LSA (SVD)

http://dfao-uc.github.io/#articulo



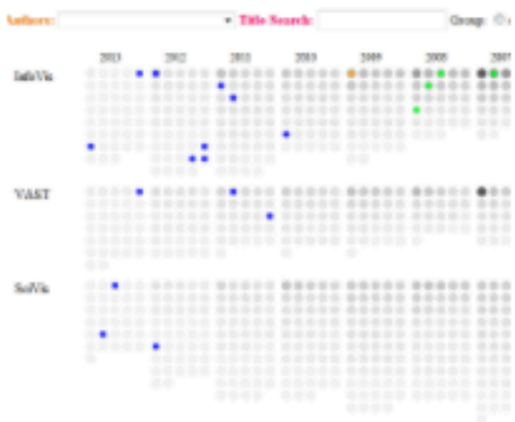
Topic Modeling

http://nbviewer.ipython.org/github/bmabey/hacker_news_topic_modelling/blob/master/HN%20Topic%20Model%20Talk.ipynb

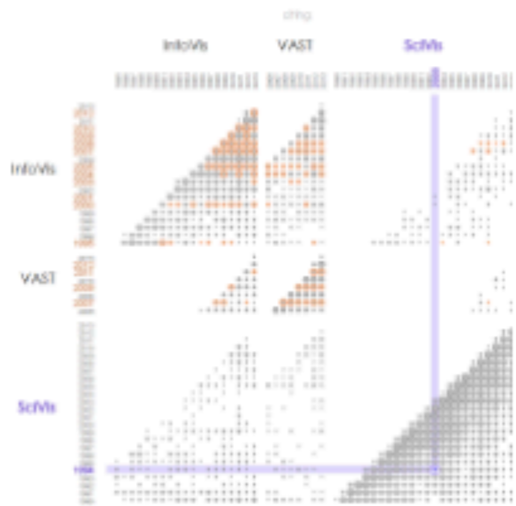


Interacción

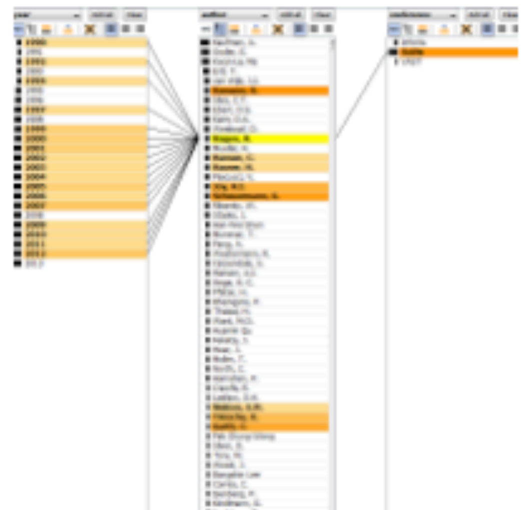
https://www.cc.gatech.edu/gvu/ii/citevis/VIS25/



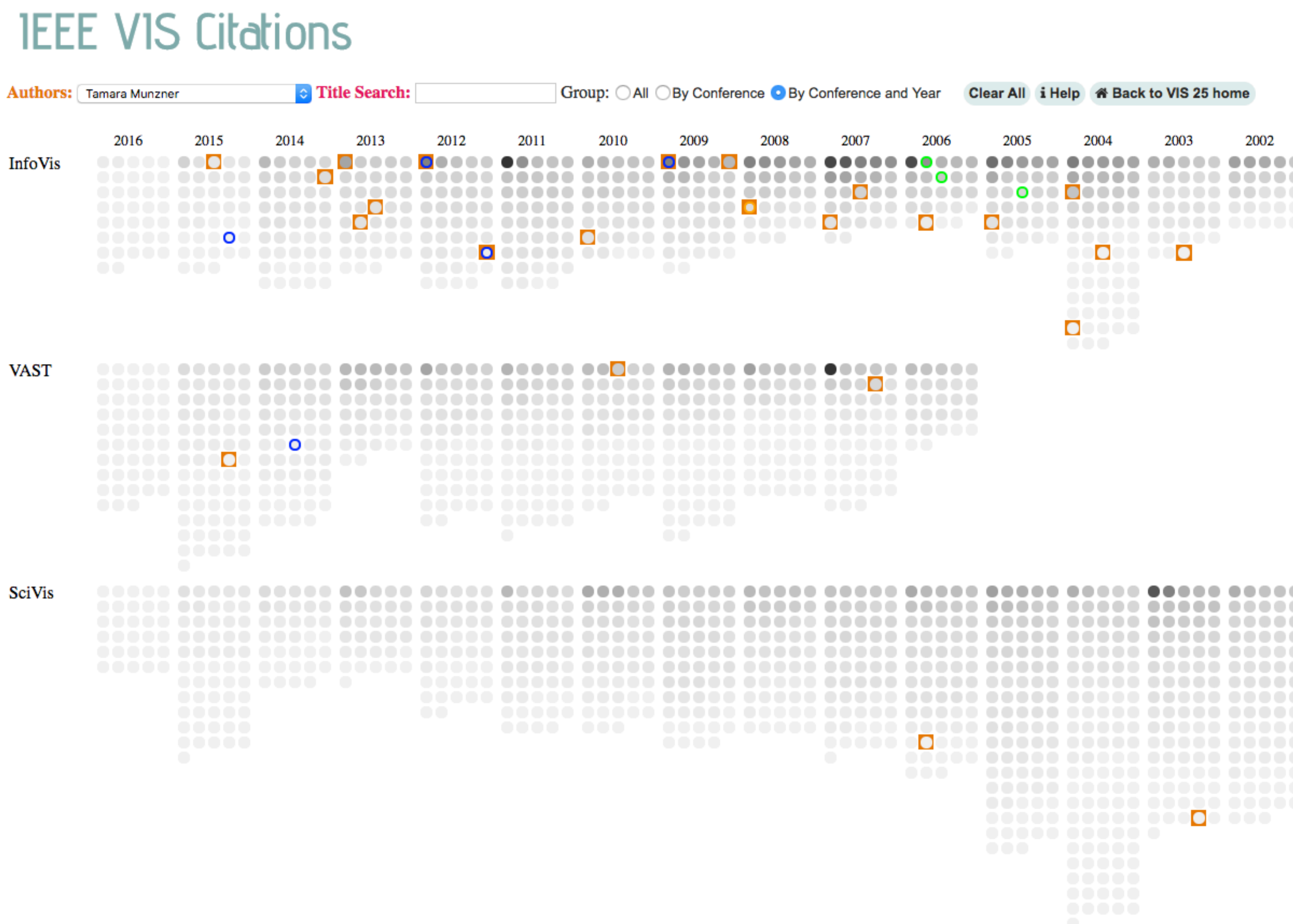
CiteVis2

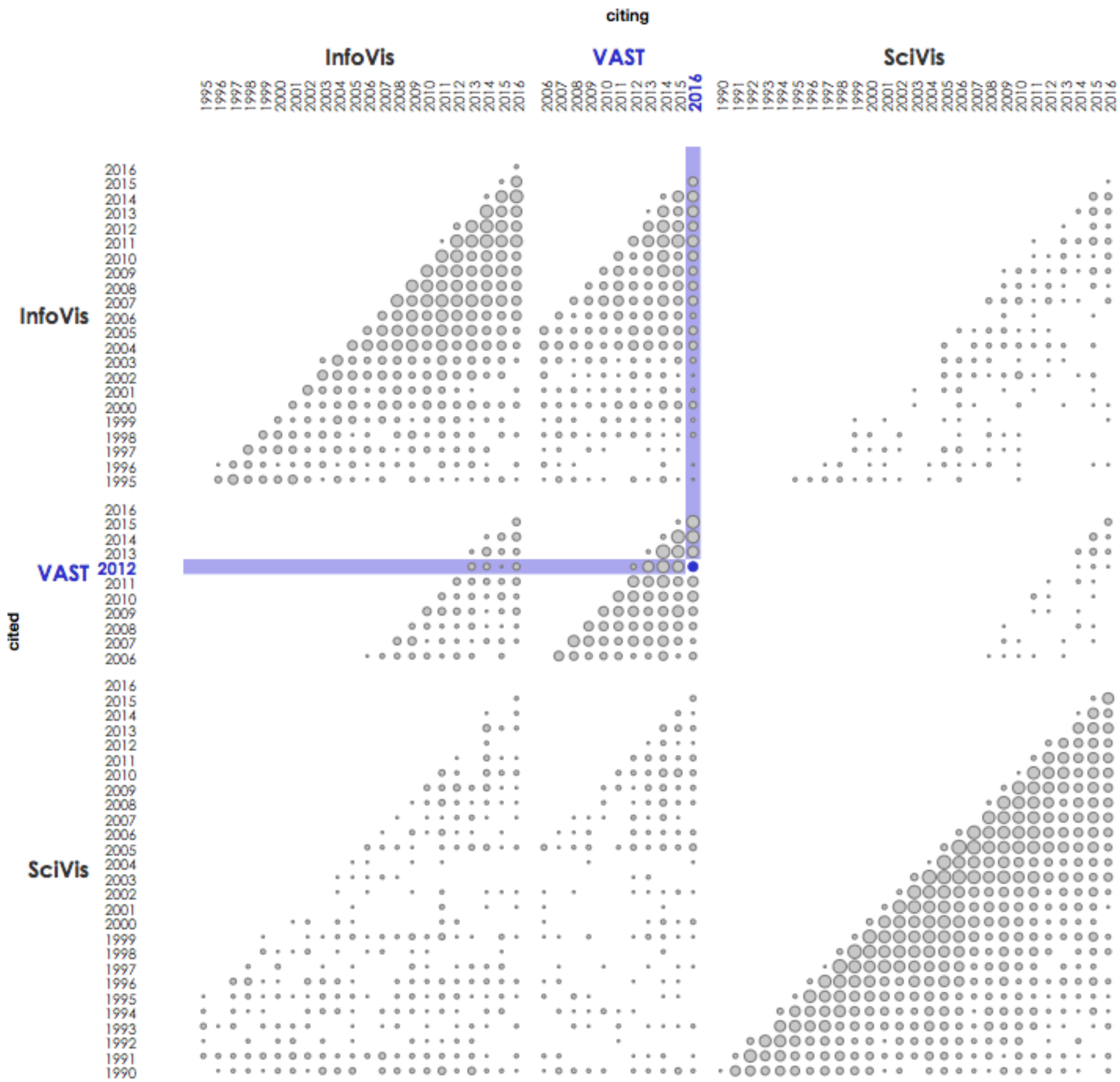


CiteMatrix



VISLists


IEEE VIS Citations
Authors: Tamara Munzner
Title Search:
Group: All By Conference By Conference and Year
Clear All
i Help
Back to VIS 25 home
2016
2015
2014
2013
2012
2011
2010
2009
2008
2007
2006
2005
2004
2003
2002
InfoVis
VAST
SciVis



Help

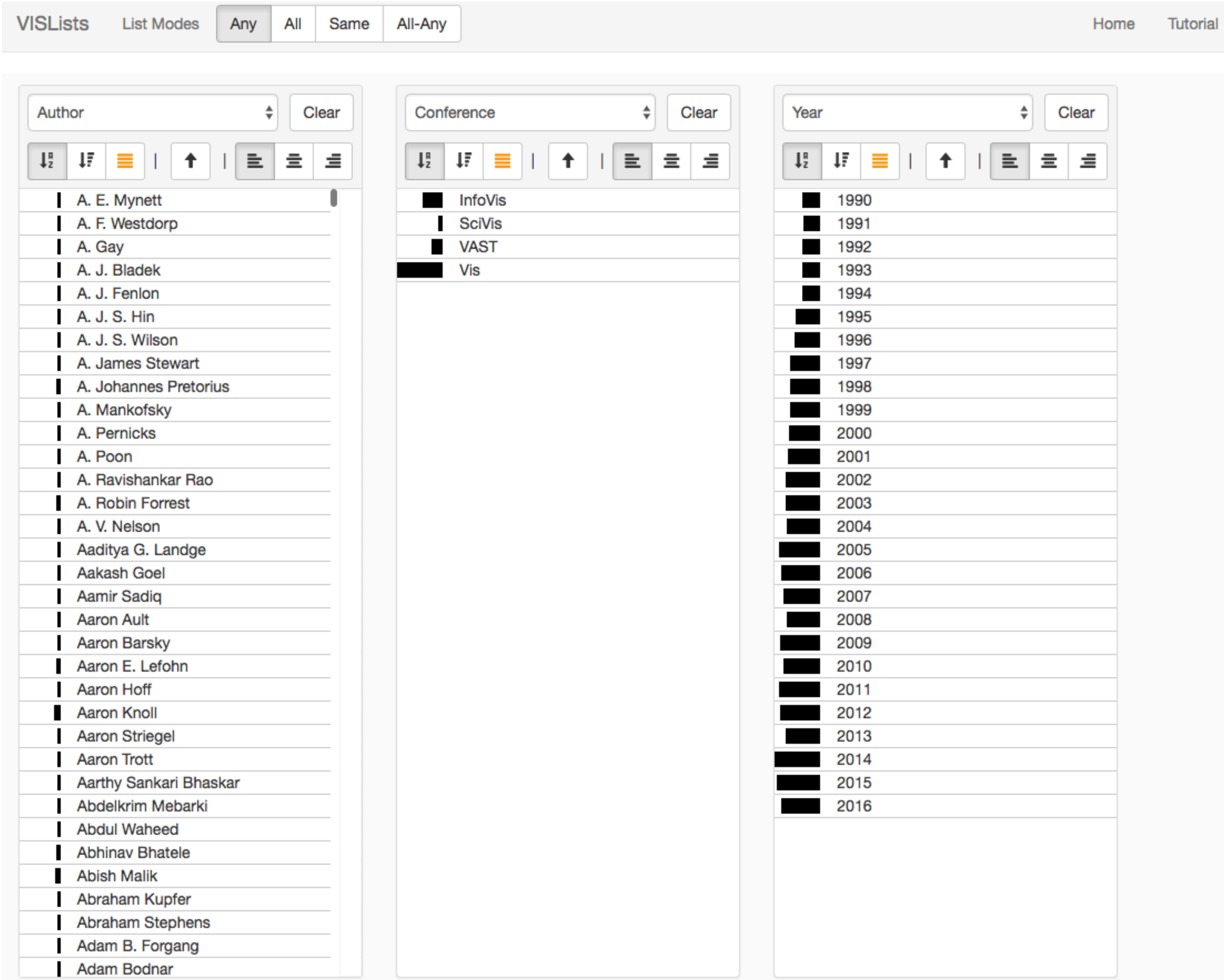
Back to VIS 25 home

Author:

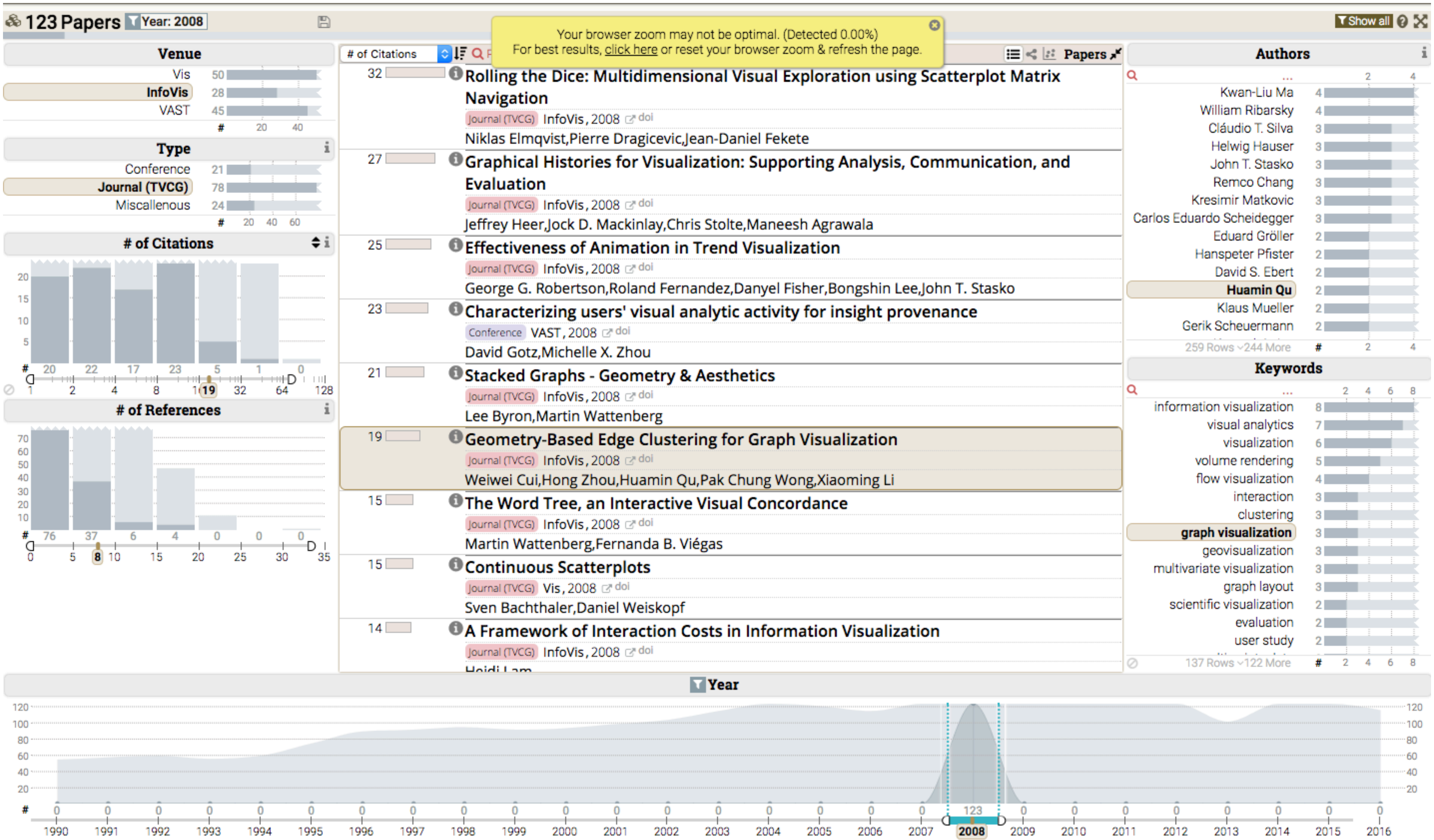
Clear All

Cited Papers:

3--Dis-function: Learning distance functions interactively
2--An Affordance-Based Framework for Human Computation and Human-Computer Collaboration
2--Visual Classifier Training for Text Document Retrieval
2--LeadLine: Interactive visual analysis of text data through event identification and exploration
1--Enterprise Data Analysis and Visualization: An Interview Study
1--Examining the Use of a Visual Analytics System for Sensemaking Tasks: Case Studies with Domain Experts
1--Just-in-time annotation of clusters, outliers, and trends in point-based data visualizations
1--Subspace search and visualization to make sense of alternative clusterings in high-dimensional data
1--Inter-active learning of ad-hoc classifiers for video visual analytics
1--Visual cluster exploration of web clickstream data
1--Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition


VISLists
List Modes
Any
All
Same
All-Any
Home
Tutorial
Author
Clear
A. E. Mynett
A. F. Westdorp
A. Gay
A. J. Bladek
A. J. Fenlon
A. J. S. Hin
A. J. S. Wilson
A. James Stewart
A. Johannes Pretorius
A. Mankofsky
A. Pernicks
A. Poon
A. Ravishankar Rao
A. Robin Forrest
A. V. Nelson
Aaditya G. Landge
Aakash Goel
Aamir Sadiq
Aaron Ault
Aaron Barsky
Aaron E. Lefohn
Aaron Hoff
Aaron Knoll
Aaron Striegel
Aaron Trott
Aarthy Sankari Bhaskar
Abdelkrim Mebarki
Abdul Waheed
Abhinav Bhatele
Abish Malik
Abraham Kupfer
Abraham Stephens
Adam B. Forgang
Adam Bodnar
Conference
Clear
InfoVis
SciVis
VAST
Vis
Year
Clear
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016

https://keshif.me/demo/vispubdata



http://charles.perin.free.fr/misc/vis/dragVisPub/dragtable.html#data=VAST

Rank		Author	Papers	**Citations▽**	Cited	Co-Authors(sum)	Co-Authors(uniq)
1	■	Ribarsky, W.	5	23	3	25	15
2	□	Jing Yang	5	23	3	25	15
3	□	Stasko, J.	2	18	0	5	5
4	■	Elmqvist, N.	1	15	0	2	2
5	□	Ward, M.O.	3	14	2	11	10
6	□	Keim, D.A.	4	13	0	10	8
7	□	Hangzai Luo	3	13	3	14	7
8	□	Jianping Fan	3	13	3	14	7
9	□	Schneidewind, J.	3	12	0	9	7
10	□	Rundensteiner, E.A.	2	9	1	5	4
11	■	Ebert, D.S.	3	7	1	19	17
12	□	Yun Jang	3	7	1	19	17
13	□	Mueller, K.	2	6	2	9	9
14	□	Heer, J.	2	6	0	5	5
15	□	Sips, M.	1	5	0	2	2
16	□	Gotz, D.	1	4	2	2	2
17	□	Zhou, M.X.	1	4	2	2	2
18	□	Weaver, C.	2	4	1	8	8
19	□	Chang, R.	1	4	0	8	8
20	□	Munzner, T.	1	4	0	3	3

http://paperquest.github.io/