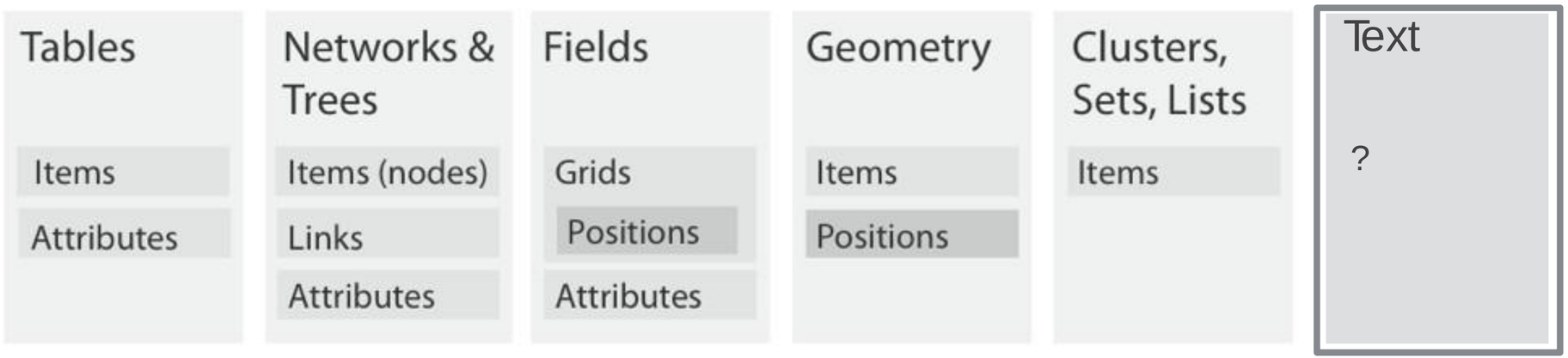
Paul Rosen
paul.rosen@utah.edu
@paulrosenphd
https://cspaul.com

Visualization for Data Science
DS-4630 / CS-5630 / CS-6630

VISUALIZING TEXT

what does it mean to be an “item”?



text data type

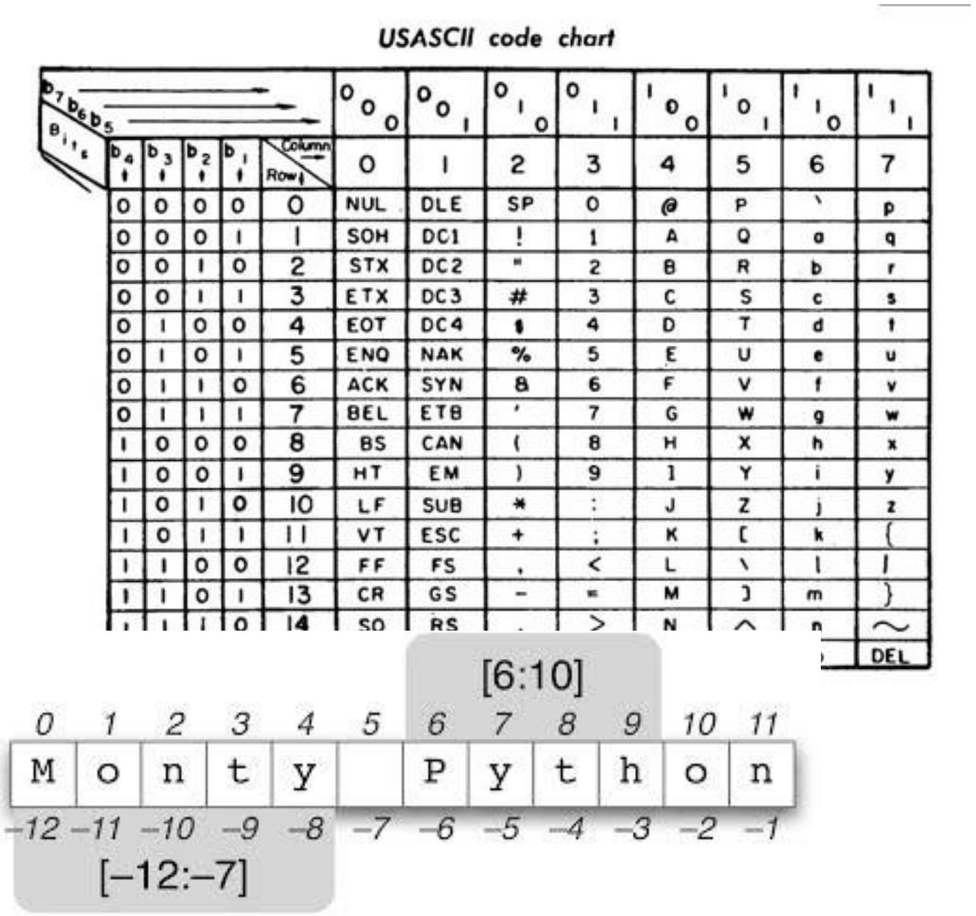
- no numbers (implicitly)
- characters (ASCII)
- strings

USASCII code chart

Bits $b_7 b_6 b_5$ →	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
b_4 ↓ b_3 ↓ b_2 ↓ b_1 ↓	Column 0	1	2	3	4	5	6	7
0 0 0 0	Row 0	NUL	DLE	SP	0	@	P	`	p
0 0 0 1	1	SOH	DC1	!	1	A	Q	a	q
0 0 1 0	2	STX	DC2	"	2	B	R	b	r
0 0 1 1	3	ETX	DC3	#	3	C	S	c	s
0 1 0 0	4	EOT	DC4	$	4	D	T	d	t
0 1 0 1	5	ENQ	NAK	%	5	E	U	e	u
0 1 1 0	6	ACK	SYN	&	6	F	V	f	v
0 1 1 1	7	BEL	ETB	'	7	G	W	g	w
1 0 0 0	8	BS	CAN	(	8	H	X	h	x
1 0 0 1	9	HT	EM	)	9	I	Y	i	y
1 0 1 0	10	LF	SUB	*	:	J	Z	j	z
1 0 1 1	11	VT	ESC	+	;	K	[	k	{
1 1 0 0	12	FF	FS	,	<	L	\	l	|
1 1 0 1	13	CR	GS	-	=	M	]	m	}
1 1 1 0	14	SO	RS	.	>	N	^	n	~
1 1 1 1	15	SI	US	/	?	O	_	o	DEL

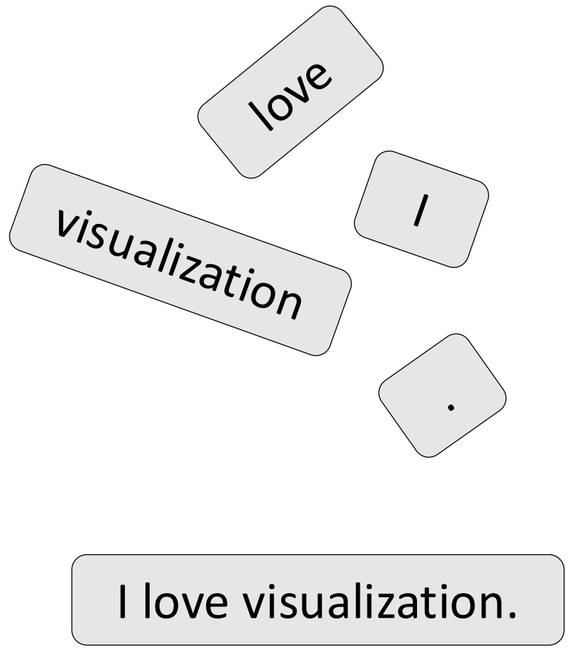
text data type

- no numbers (implicitly)
- characters (ASCII)
- strings



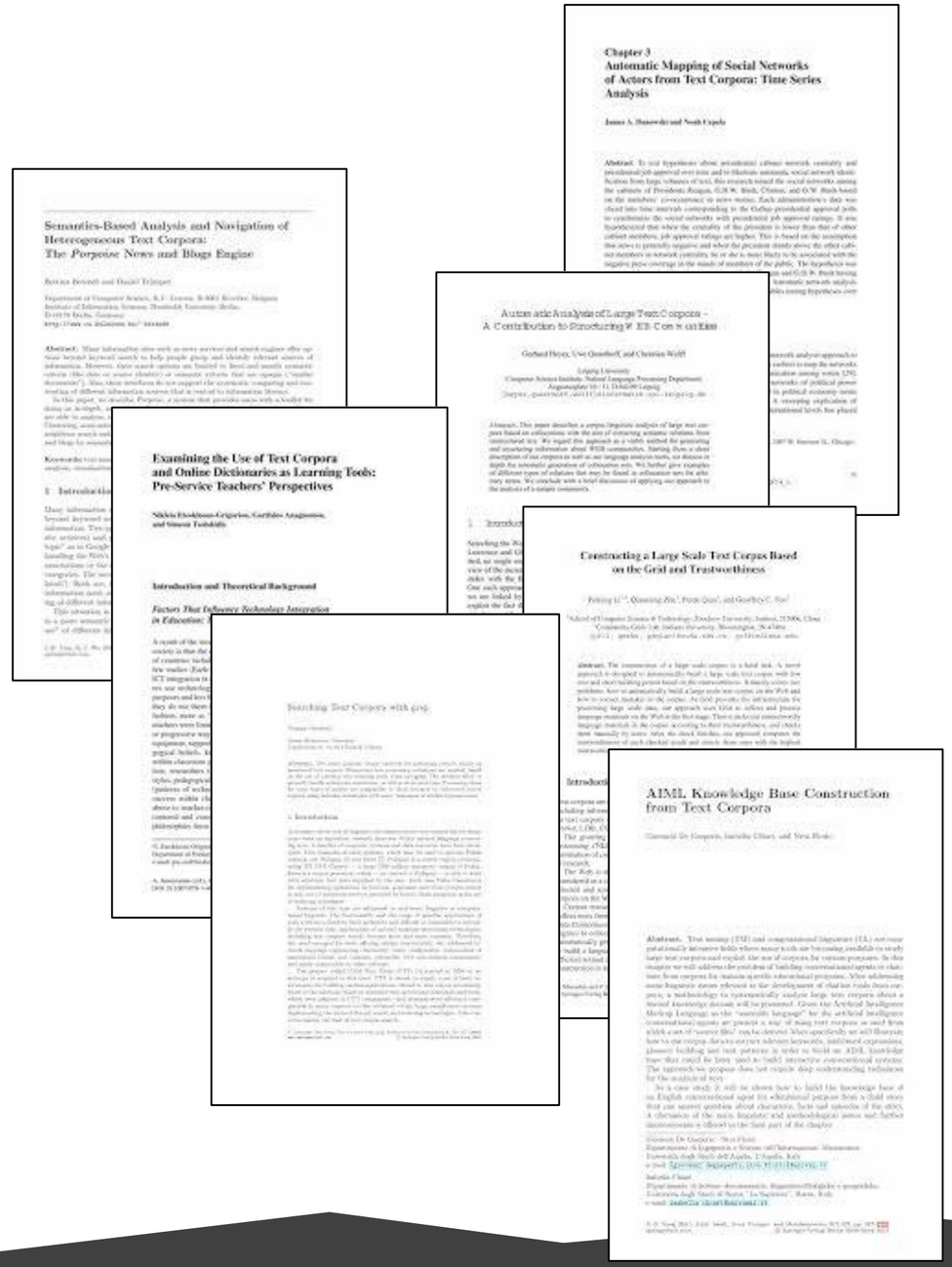
text data semantics

- words
- lines
- sentences
- paragraphs
- chapters

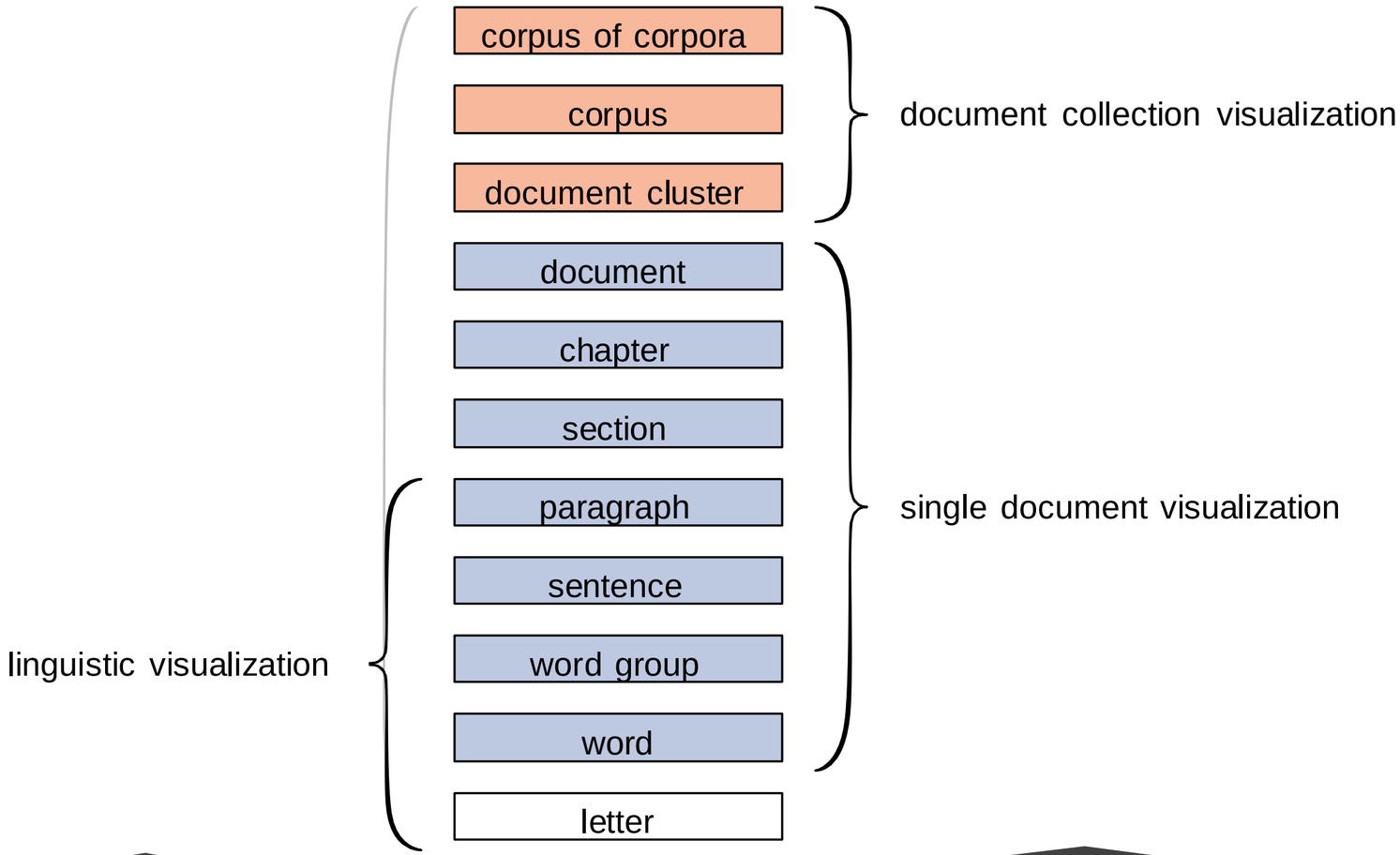


text data semantics

- documents: books, papers, webpages, e-mails, twitter post
- corpus: collection of documents



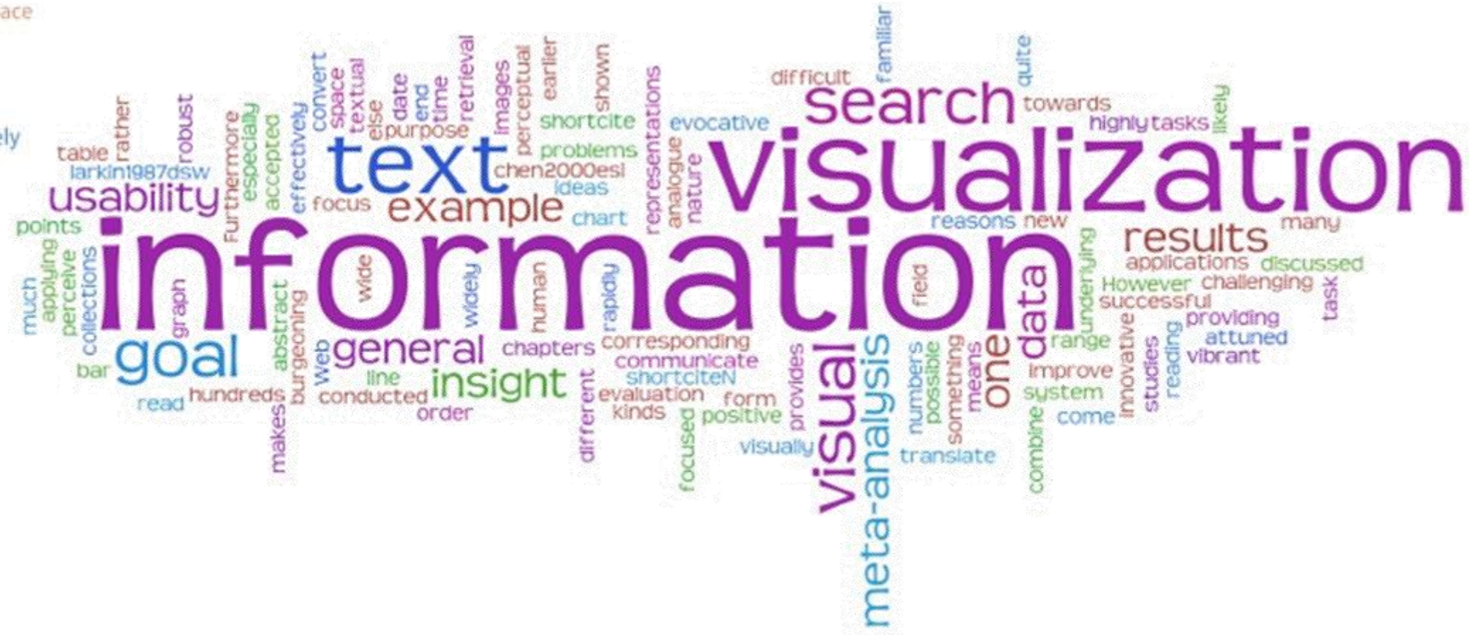
Text Units Hierarchy



single document

Tag Cloud / Word cloud

abstract accepted analogue applications applying attuned bar burgeoning challenging chapters chart collections combine communicate conducted convert data date difficult discussed earlier effectively end evaluation evocative familiar field focus focused form general goal graph highly human hundreds ideas images improve information innovative insight kinds line makes means meta-analysis nature new numbers order ost perceive perceptual points positive problems providing purpose range rapidly read reading reasons representations results retrieval robust search shortciten{chen2000esi} shortcite{larkin1987dsw} shown space studies successful system table task tasks text textual time translate underlying usability vibrant visual visualization visually web wide widely



Word Tree

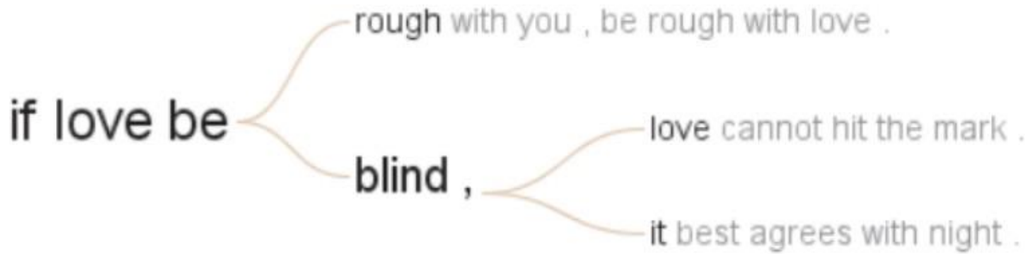
Text

if love be rough with you , be rough with love .

if love be blind , love cannot hit the mark .

if love be blind , it best agrees with night .

WordTree

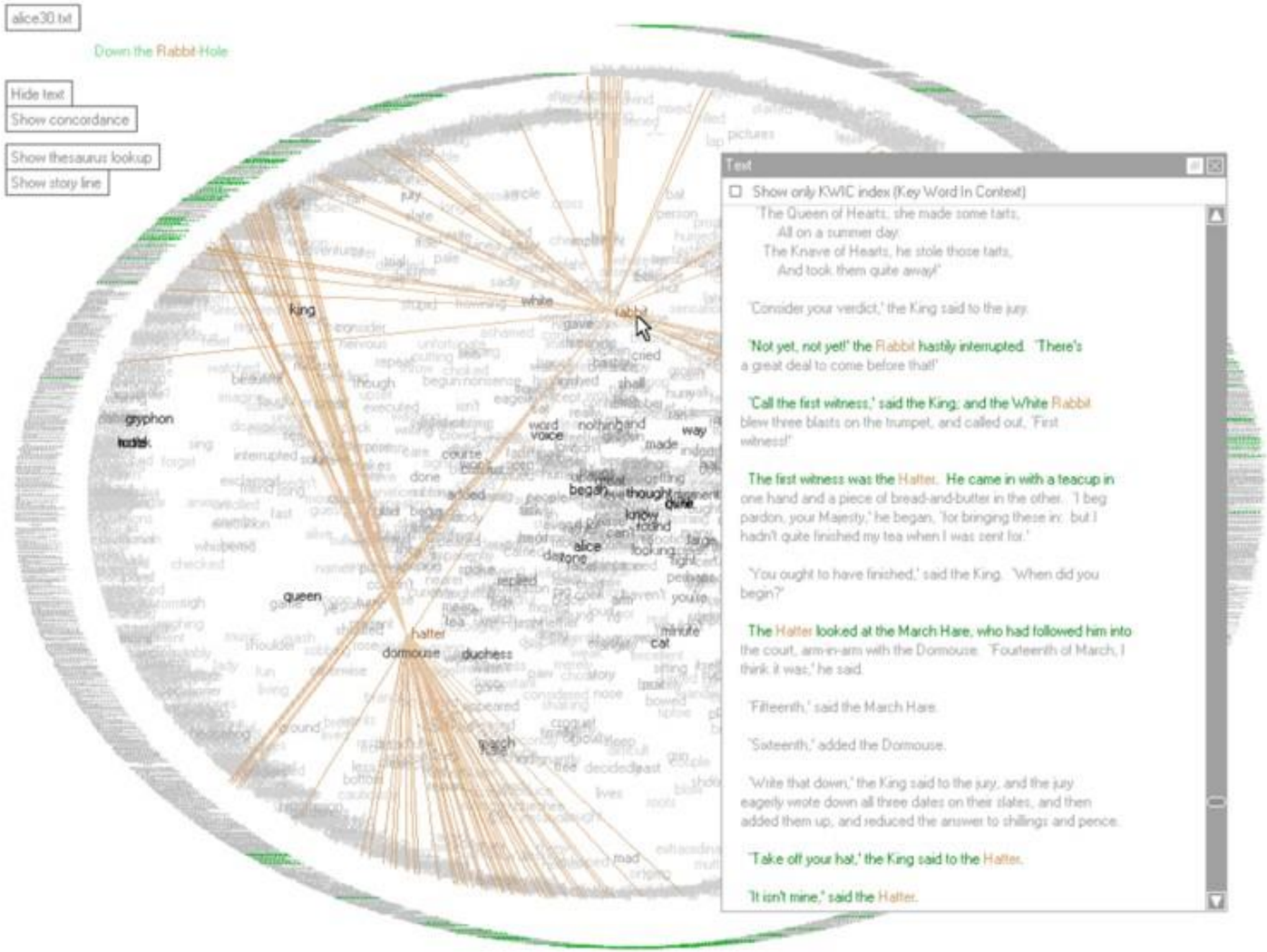


[Wattenberg 2008]

Visualizations : definitions of visualization word tree

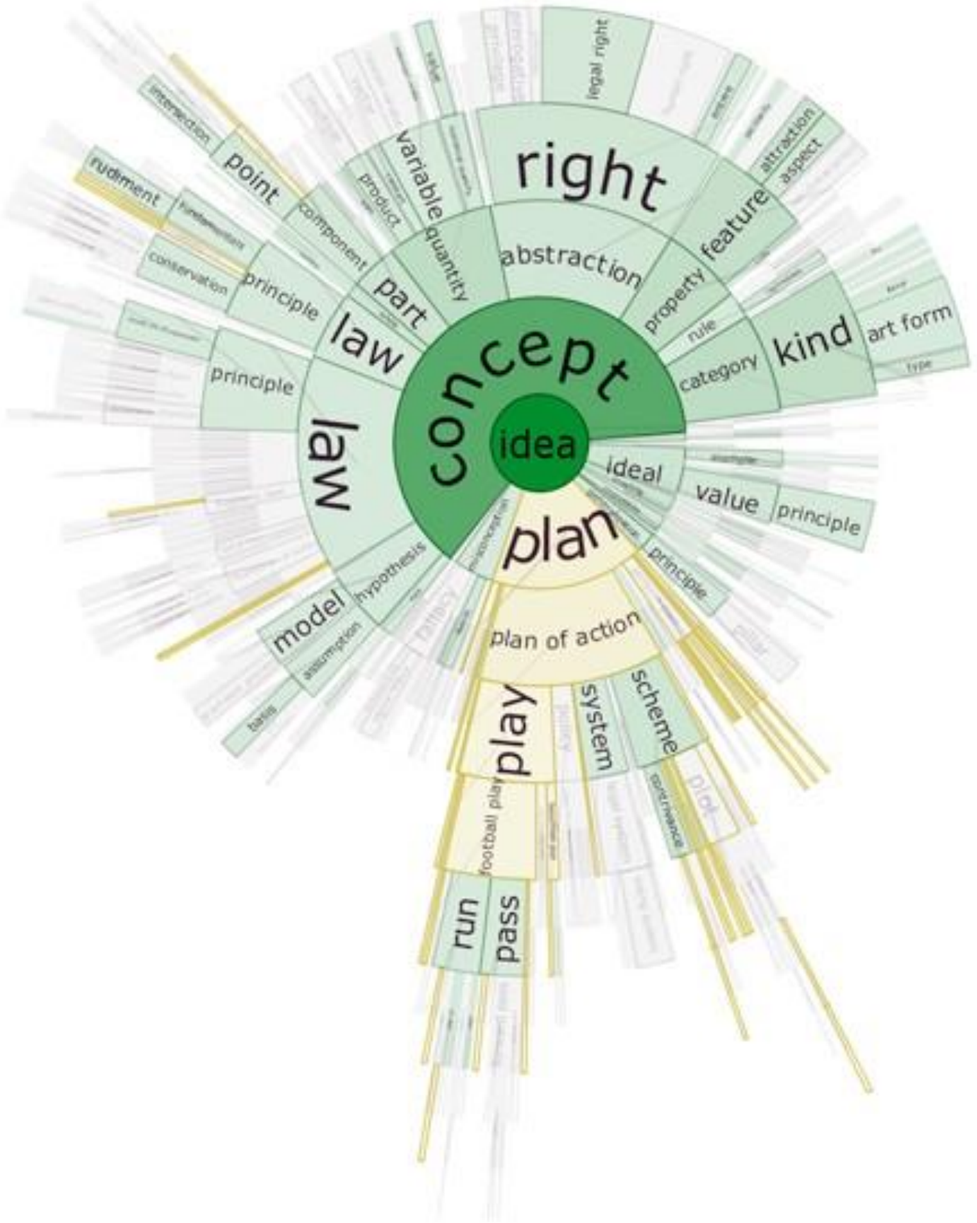
Uploaded by: mhalle
Created at: Wednesday May 21 2008, 11:37 PM
Tags: text

Text Arc



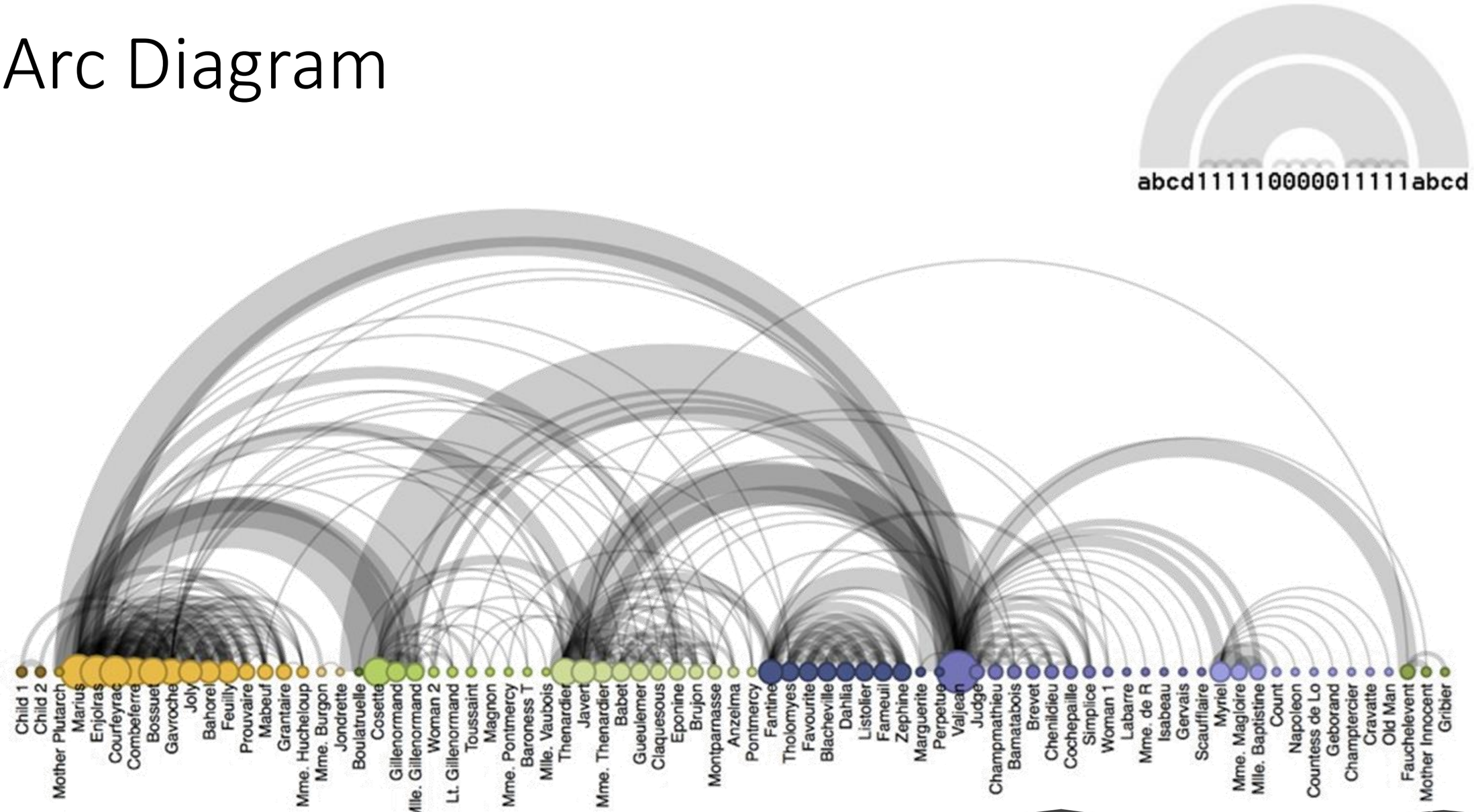
- Wattenberg, Viegas 2008

DocuBurst

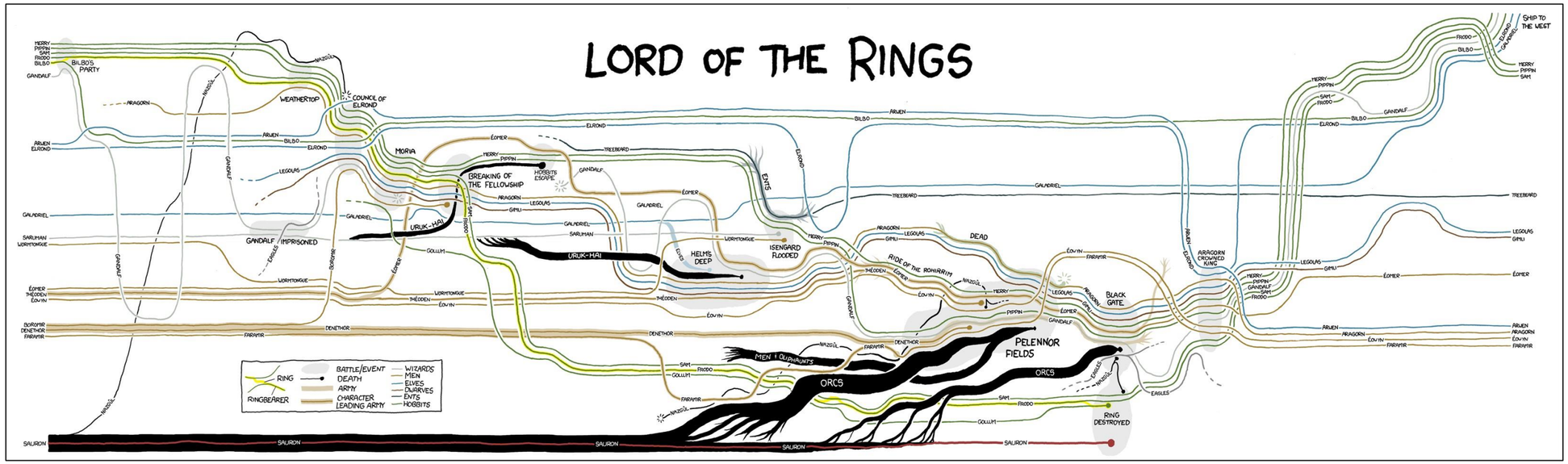


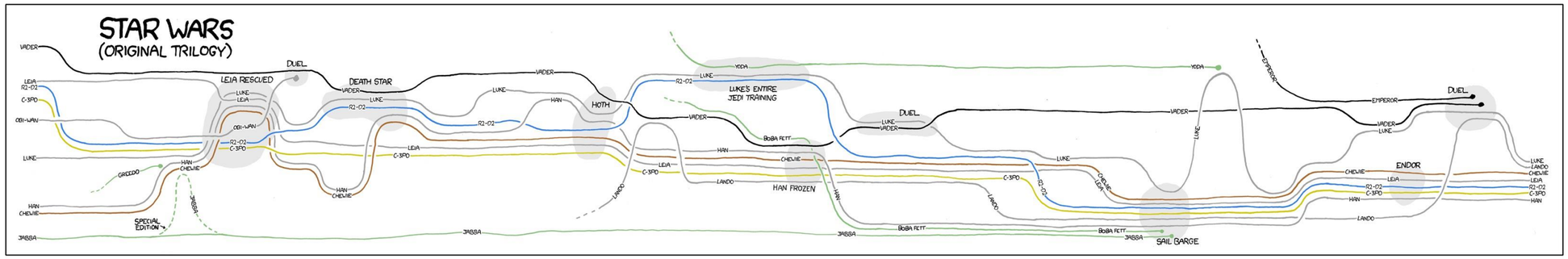
- Collins, Carpendale, Penn 2008

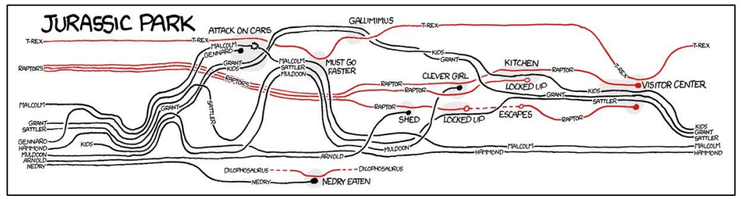
Arc Diagram

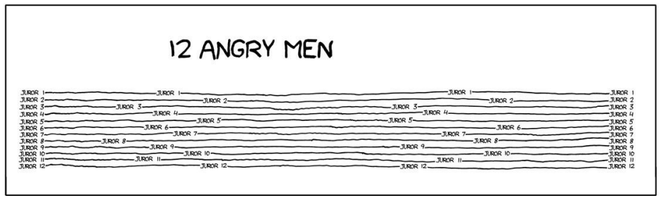

abcd1111100000111111abcd
Child 1
Child 2
Mother Plutarch
Marius
Enjolras
Courfeyrac
Combeferre
Bossuet
Gavroche
Joly
Bahorel
Feuilly
Prouvaire
Mabeuf
Grantaire
Mme. Hucheloup
Mme. Burgon
Jondrette
Boulatruelle
Cosette
Gillenormand
Mlle. Gillenormand
Woman 2
Lt. Gillenormand
Toussaint
Magnon
Mme. Pontmercy
Baroness T
Mlle. Vaubois
Thenardier
Javert
Mme. Thenardier
Babet
Gueulemer
Claquesous
Eponine
Brujon
Montparnasse
Anzelma
Pontmercy
Fantine
Tholomyes
Favourite
Blacheville
Dahlia
Listolier
Fameuil
Zephine
Marguerite
Perpetue
Valjean
Judge
Champmathieu
Bamatabois
Brevet
Chenildieu
Cochepaille
Simplice
Woman 1
Labarre
Mme. de R
Isabeau
Gervais
Scaufflaire
Myriel
Mme. Magloire
Mlle. Baptistine
Count
Napoleon
Countess de Lo
Geborand
Champtercier
Cravatte
Old Man
Fauchelevent
Mother Innocent
Gribier

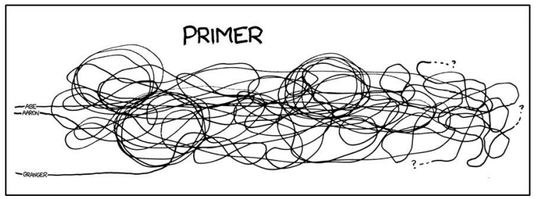
THESE CHARTS SHOW MOVIE CHARACTER INTERACTIONS.
THE HORIZONTAL AXIS IS TIME. THE VERTICAL GROUPING OF THE
LINES INDICATES WHICH CHARACTERS ARE TOGETHER AT A GIVEN TIME.





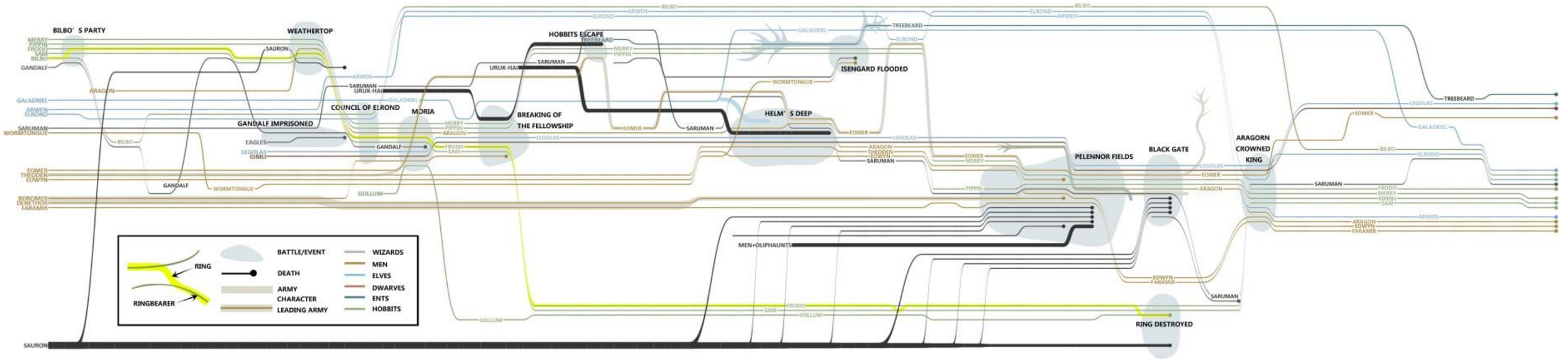






https://xkcd.com/657/

StoryFlow: Tracking the Evolution of Stories



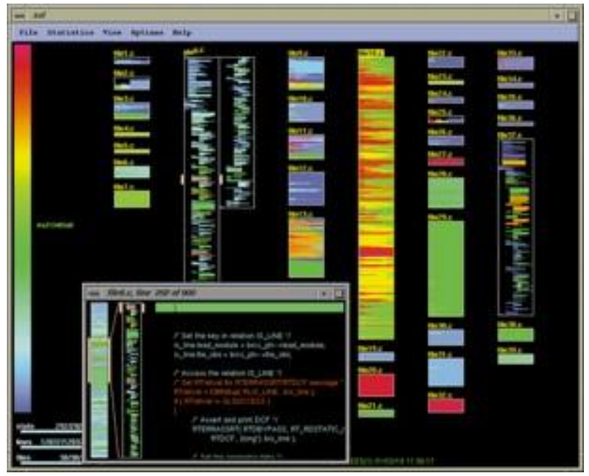
[Liu 2013]

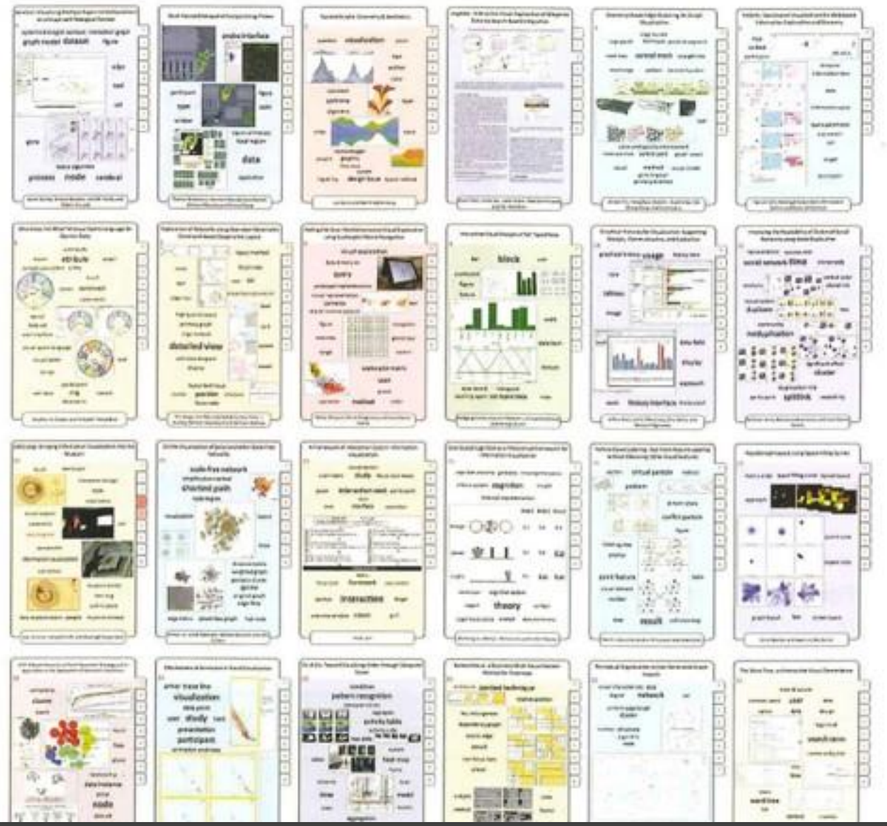
collection of documents

Text Corpora

- Varied Goals:
  - Discover interesting documents Summarize Documents Classify Documents
  - Extract Facts (Intelligence Analysis)
- Rich Information:
  - Document Metadata Authors, date, type,
  - Paragraphs, figures…
  - Revisions, annotations, comments,

Document Cards (small multiples)





Strobelt et al. 2009

MDS Approaches

- use bag-of-word to project documents w.r.t. text similarity into a landscape
- (only) one example

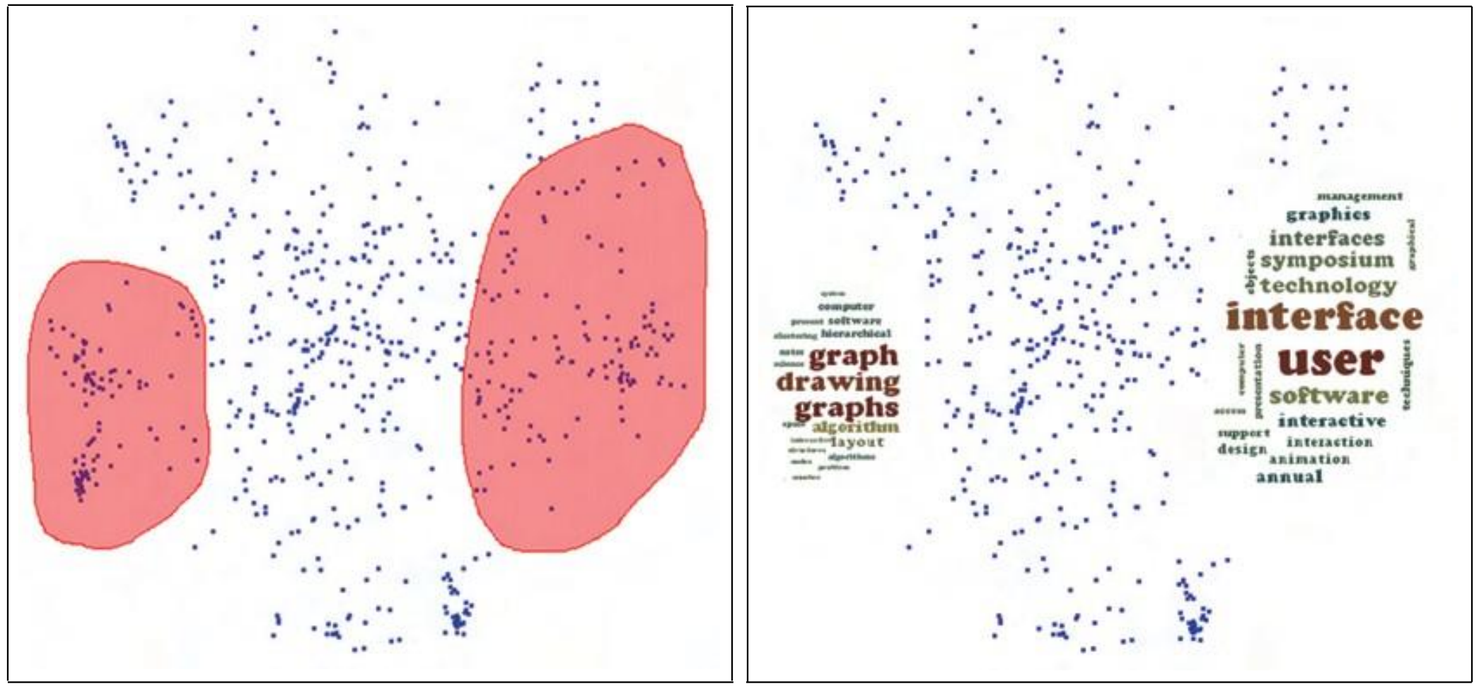
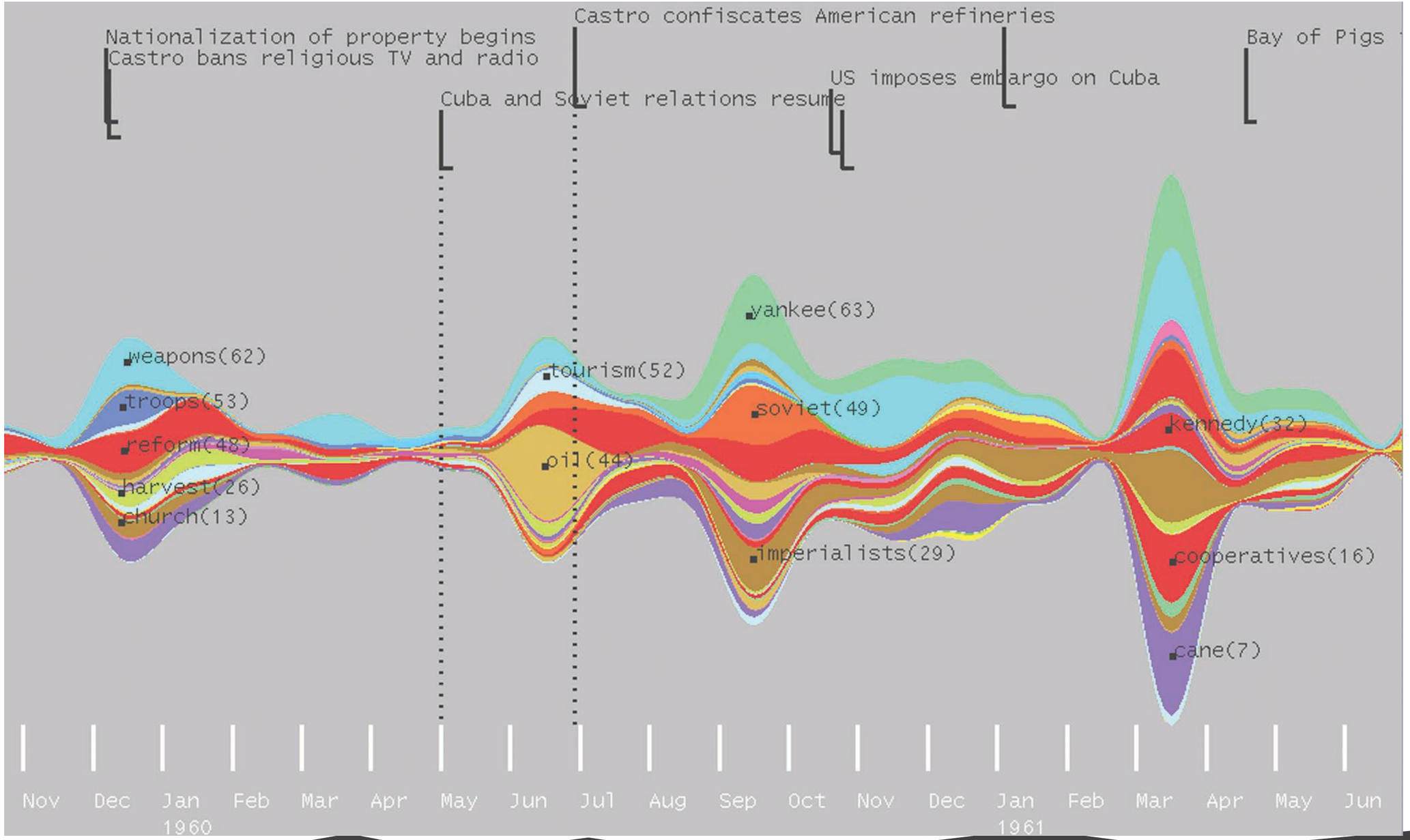


Figure 5: *A user can interactively draw a region (polygon) containing a subset of documents of interest (top figure). Keywords are extracted from the selected document and their corresponding word could is built inside the user-defined region (bottom figure).*



ThemeRiver – Havre, Hetzler, Whitney, Nowell 2002

Parallel Tag Clouds to Explore and Analyze Faceted Text Corpora

Christopher Collins

Subscribe 26

2,768

Add to Share More 12 0

Uploaded on Aug 19, 2009

This video accompanies a full publication which appeared in the Proceedings of the 2009 IEEE Symposium on Visual Analytics Science and

Jigsaw: Many Linked Views

Visual Analytics Support for Intelligence Analysis
Case Study: The 9/11 Report

Carsten Görg
Youn-ah Kang
Zhicheng Liu
John Stasko

Information Interfaces Group
Georgia Institute of Technology

http://textvis.lnu.se/