

# VISUAL ANALYSIS & MINING OF SOCIAL MEDIA – PART 1

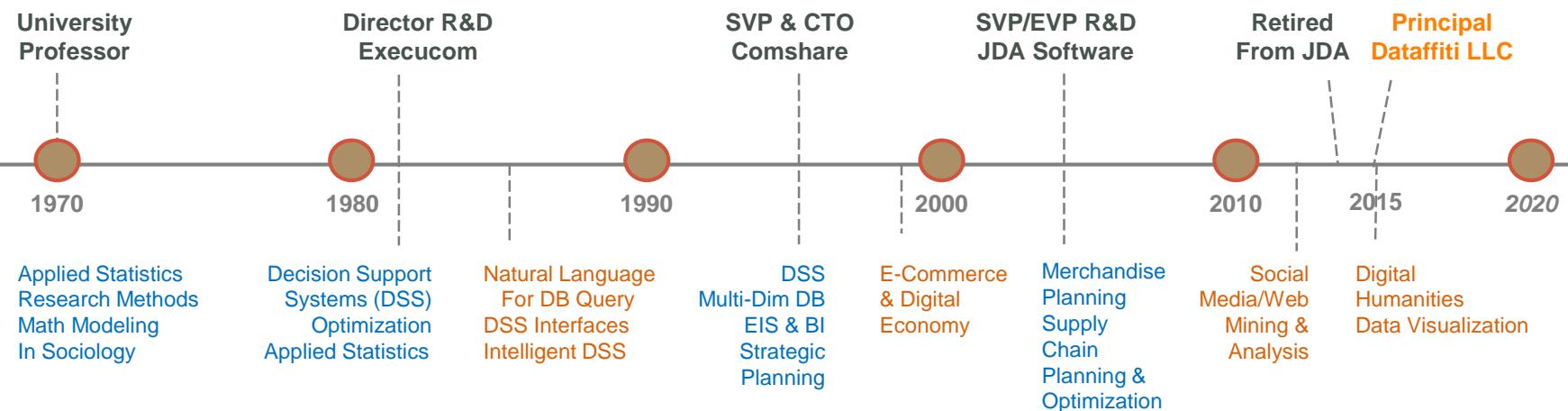
---

Dave King  
HICSS 48  
2015

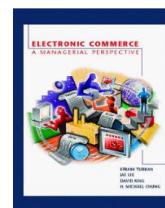
# Agenda

- Definitions
- Example 1: Quantification of Self
- Visual Framework(s)
- Example 2: Text analysis of Rap lyrics
- Example 3: Social Network Analysis

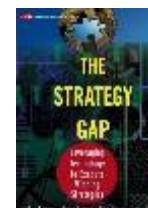
# Biography



- Formal Positions
- Work Focus
- Research Focus
- Books



1<sup>st</sup> Edition '99



'08



2nd Edition '11



3rd Edition '14



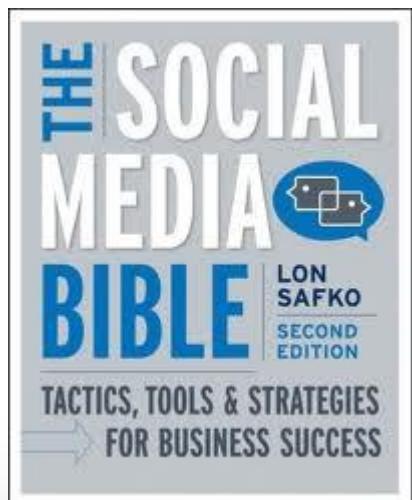
8<sup>th</sup> Edition '15

# DEFINITIONS

---

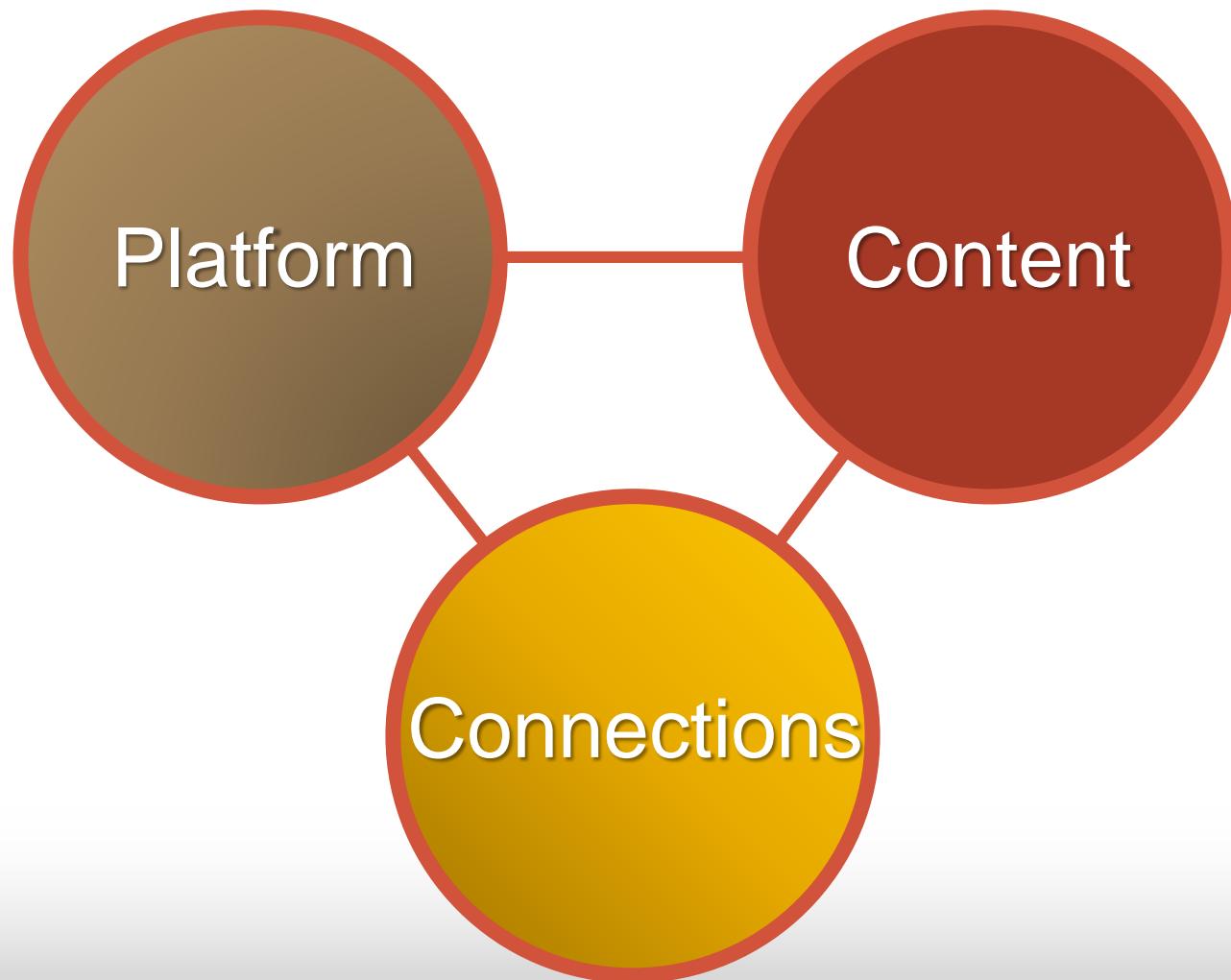
Social Media, Data Mining & Visualization

# Defined



is the media we use to  
be social. That's it.

# Media elements



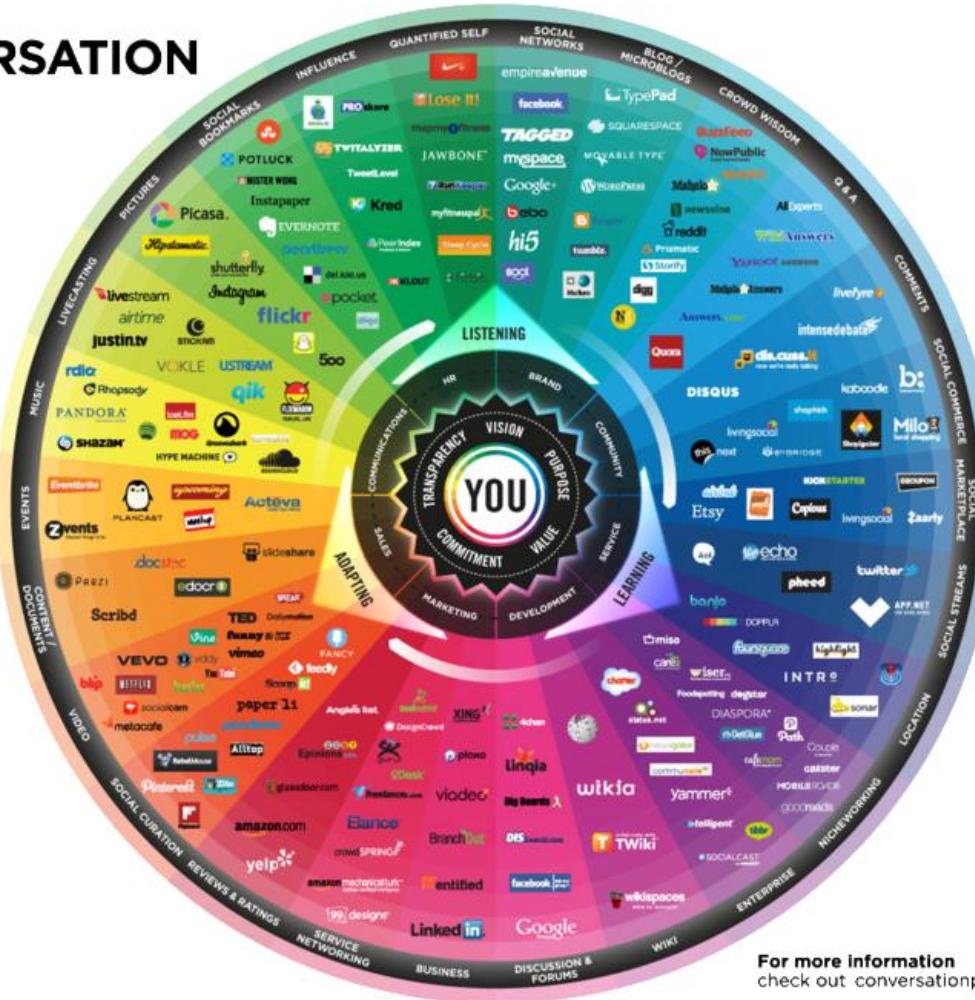
# Social media defined

- A group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, which allows the creation and exchange of user-generated content (Kaplan & Haenlein, 2008)
- Media for social interaction, using highly accessible and scalable publishing techniques (Morgan et al., 2012)
- The means of interaction among people in which information is created, shared, and exchanged, most often in virtual communities and digital networks (Paolo Ciuccarelli et al., 2014)

# Social media landscape (2013)

# THE CONVERSATION PRISM

Brought to you by  
Brian Solis & JESS3



# Mining and analyzing social media: *Key elements*

## Content



## Connections

# Social media content:

*Text, tags, profiles, location, images, video & sound*



# Social media connections:

## *Roles and relationships*



# Social media mining defined



The process of representing, analyzing, and extracting (actionable) patterns from social media content, relations, and interactions.

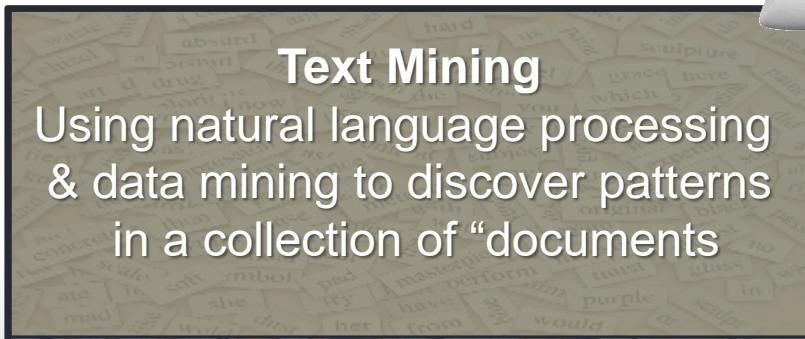
# Mining and analyzing: *Social media content*



**Data Mining**  
Discovering meaningful patterns from large data sets using pattern recognition technologies



**Web Mining**  
Data Mining focused on the analysis of Web usage, structure & content



**Text Mining**  
Using natural language processing & data mining to discover patterns in a collection of “documents”



**Image/Video Mining**  
Using image/video processing & analysis to uncover patterns in collections of images/videos

# Some data mining models...



- Summarization
- Regression
- Anomaly Detection
- Association Rule Learning
- Clustering
- Classification

# Data mining models assume the data is...



- Structured
- Transformed
- Well-formed

# Data mining example: *Affinity analysis*

Frequently Bought Together



Price for all three: \$63.45

This item: Visual Insights: A Practical Guide to Making Sense of Data by Katy Börner Paperback \$19.77

The Book of Trees: Visualizing Branches of Knowledge by Manuel Lima Hardcover \$18.84

Visual Complexity: Mapping Patterns of Information by Manuel Lima Paperback \$24.84

Show availability and shipping details

Customers Who Bought This Item Also Bought

Page 1 of 25



Visual Complexity: Mapping Patterns of...  
› Manuel Lima  
★★★★★ 18  
Paperback  
\$24.84 

The Book of Trees: Visualizing Branches...  
› Manuel Lima  
★★★★★ 14  
Hardcover  
\$18.84 

The Visual Organization: Data Visualization,...  
› Phil Simon  
★★★★★ 37  
Hardcover  
\$32.89 

Design for Information: An Introduction to the...  
› Isabel Meirelles  
★★★★★ 6  
Paperback  
\$25.17 

Transaction  
Records



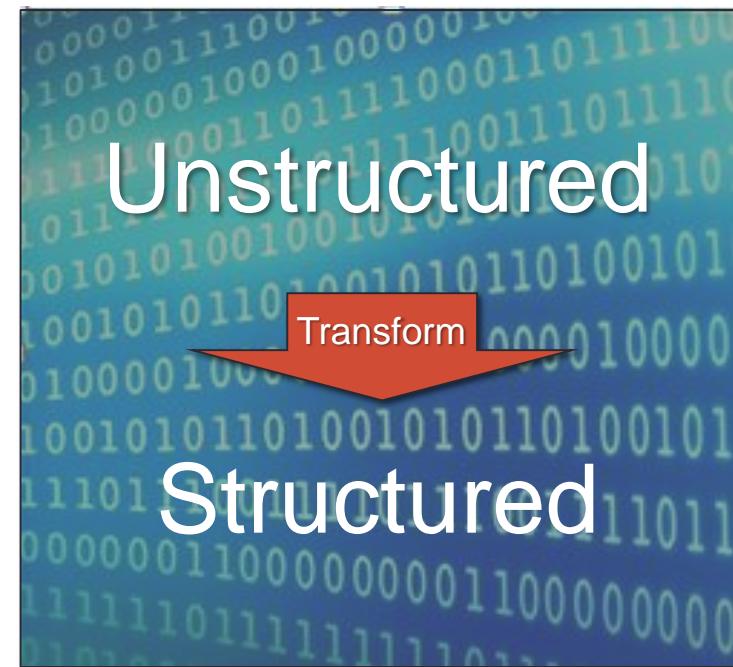
Association  
Rules

# Social media content is usually...



- Structured
- Transformed
- Well-formed

# Mining and analyzing Content requires...



# Mining social media content

## *Simple example – political blog text*

Holla-tricking-JayZ, the Obamacare Web site makeover ROCKS. Holla-tricking-JayZ, the long-awaited makeover of the Obamacare Web site ROCKS.

Three clicks and anyone can browse health care plans & in detail, and with prices. No more having to create an account simply in order to browse plans. No more having to have Homeland Security give you a probable cause before approving you to simply browse plans.

The swamp of the Affordable Care Act (ACA) federal exchange site is mostly over when they needed to do it's even easier than the DC exchange (and the DC one is great).

We've written before about the problems we (Becca and I both) personally had navigating the federal exchange. While the DC exchange worked quite well if you simply had to create a user name and password before you could browse plans, the Federal site was a nightmare. You had to go through a long process to create an application, simply to browse plans. Then you had to wait several days for the application-to-browse to be approved by angry little gnomes locked in the basement of HHS.

The new site? Click 3 buttons. That's it. Let me walk you through how easy the new Obamacare Web site actually is. The federal health care exchange home page is now short and simple, 3 options, making it very clear what you can do with the site. For our purposes, I chose option 1: I can see plans before I apply.

The Colorado #obamacare exchange is a disaster.

I'm sorry, but there's no other way to describe these enrollment numbers. Which are, by the way, from a state exchange – meaning that the problems of healthcare.gov should theoretically be irrelevant to the conversation:

Enrollment have grown to 15,074, marketplace officials announced Monday. Marketplace officials set a goal of 136,000 people covered on exchanged-based plans by the end of 2014, but so far the exchange has failed to reach even worst-case enrollment projections.

The Democratic-controlled Coloradan state government is blaming this mess on people supposedly not having to replace their insurance now and bad press about the national exchange. Which is easier than blaming the Democrats now running the state exchange, particularly since the head of the exchange wanted a merit raise for all her work on it. It's certainly easier than simply admitting that hey, maybe there wasn't all that much of a burning need for Obamacare in the first place.

But surely that's crazy talk.

Yes, I am sorry. Bad numbers = people not being insured next year. What I am not responsible for any of this.

Data Transformation

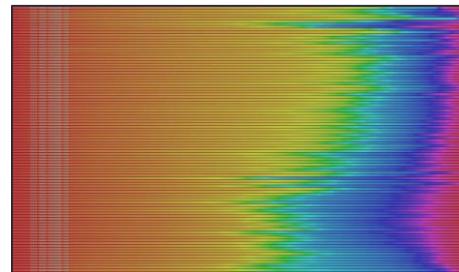
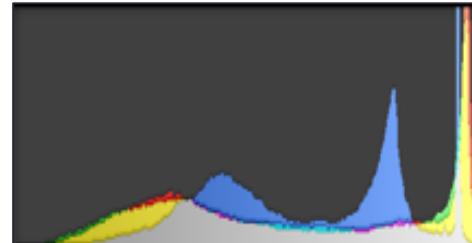
Index	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Column 11	Column 12	Column 13	Column 14	Column 15	Column 16	Column 17	Column 18	Column 19	Column 20	Column 21	Column 22	Column 23	Column 24	Column 25	Column 26	Column 27	Column 28	Column 29	Column 30	Column 31	Column 32	Column 33	Column 34	Column 35	Column 36	Column 37	Column 38	Column 39	Column 40	Column 41	Column 42	Column 43	Column 44	Column 45	Column 46	Column 47	Column 48	Column 49	Column 50	Column 51	Column 52	Column 53	Column 54	Column 55	Column 56	Column 57	Column 58	Column 59	Column 60	Column 61	Column 62	Column 63	Column 64	Column 65	Column 66	Column 67	Column 68	Column 69	Column 70	Column 71	Column 72	Column 73	Column 74	Column 75	Column 76	Column 77	Column 78	Column 79	Column 80	Column 81	Column 82	Column 83	Column 84	Column 85	Column 86	Column 87	Column 88	Column 89	Column 90	Column 91	Column 92	Column 93	Column 94	Column 95	Column 96	Column 97	Column 98	Column 99	Column 100
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88</												

# Mining social media content

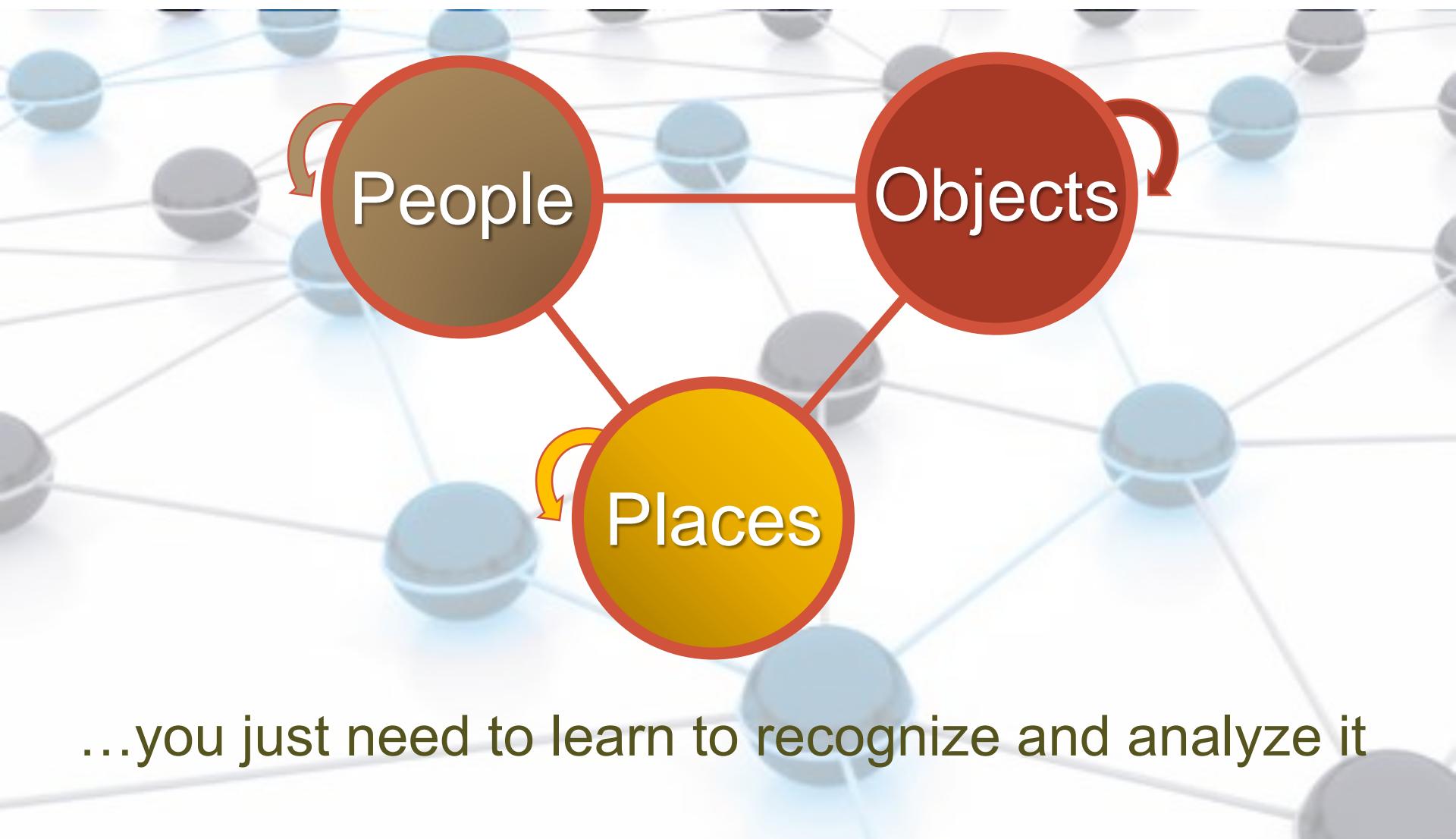
## *Simple example – graffiti collection*



Data Transformation



# Mining and analyzing Connections

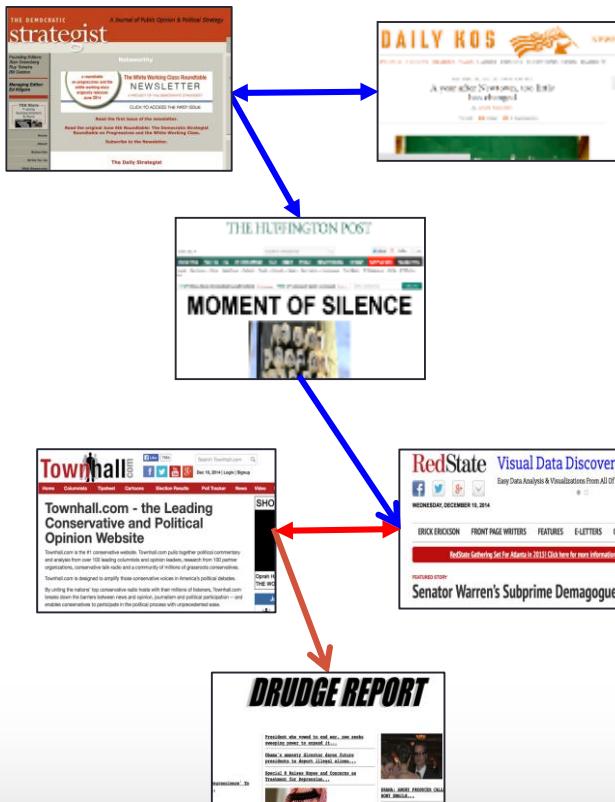


...you just need to learn to recognize and analyze it

# Mining social media connections

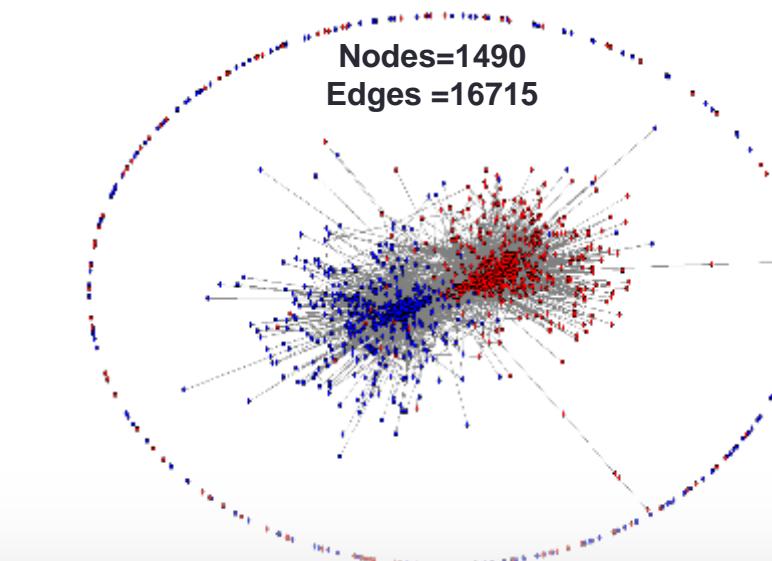
## Simple example – political blogrolls test

Individual  
Political Blogrolls



Data Transformation

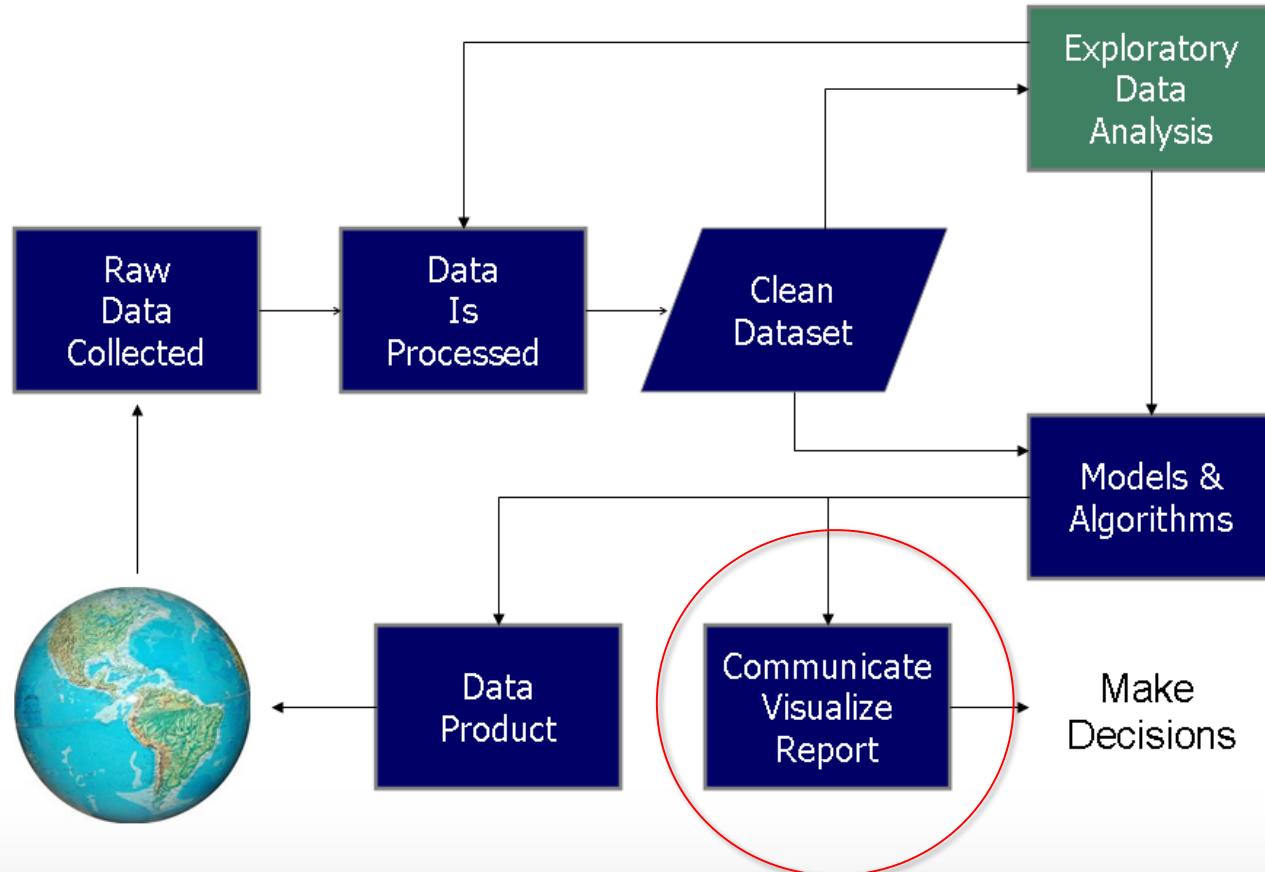
	Blog1	Blog2	Blog3	Blog4	...	BlogN
Blog1	-	1	0	0	...	0
Blog2	1	-	0	1	...	0
Blog3	0	0	-	1	...	1
Blog4	0	1	1	-	...	1
...	...	...	...	...	...	...
BlogN	0	0	1	1	...	-



Who links to whom?

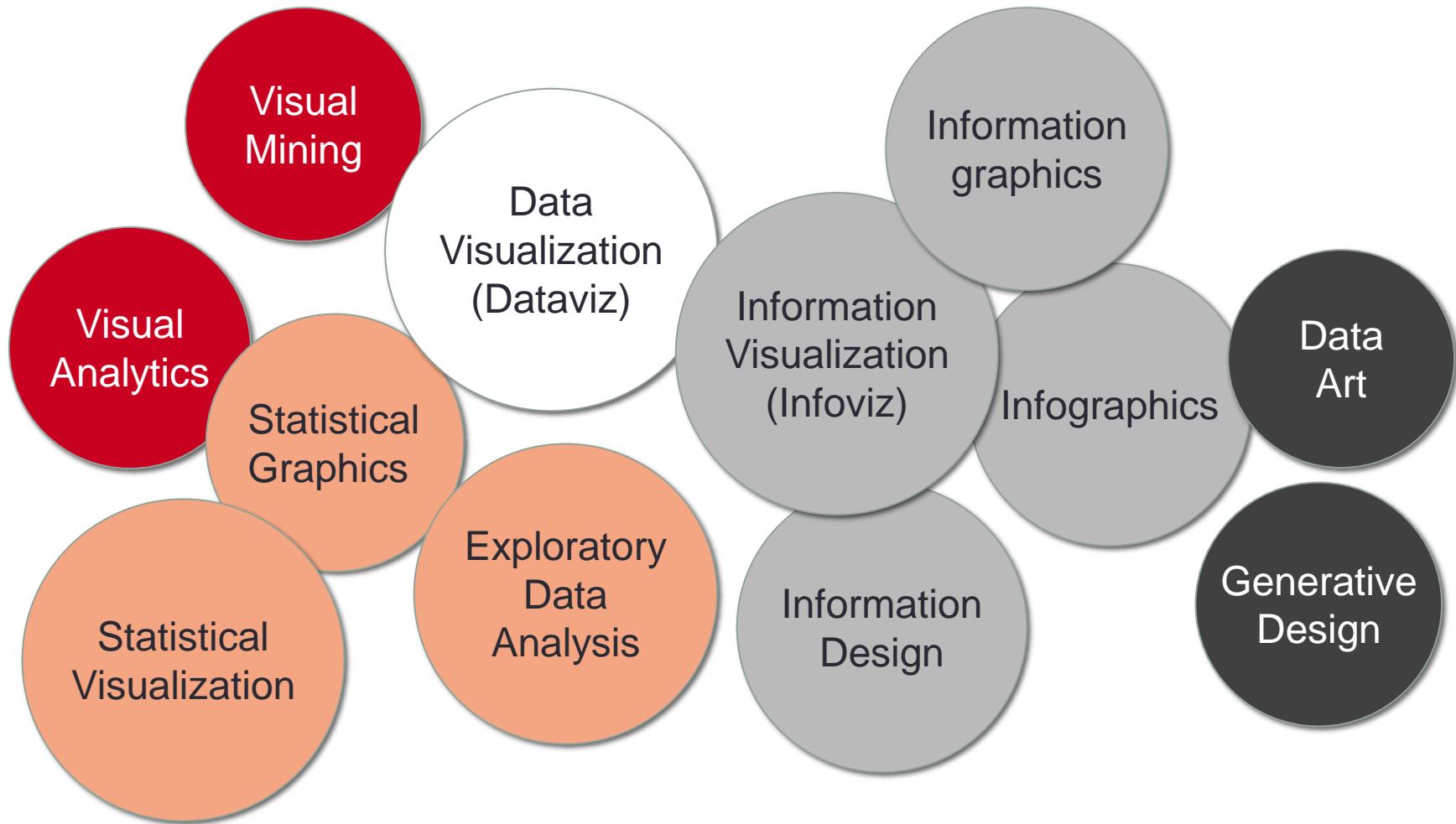
# Data Science Process

## *Mining and analysis workflow*



# Visualization by any other name...

*Multiple terms with multiple definitions*

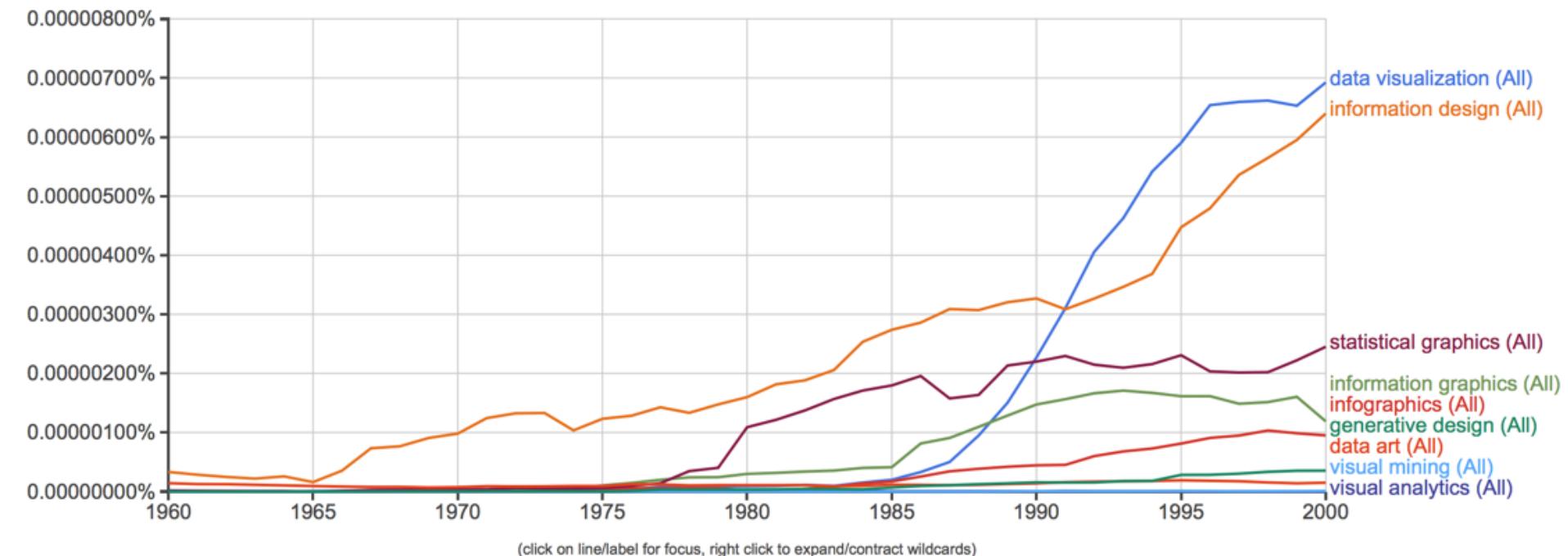


# Interest in visualization et al.

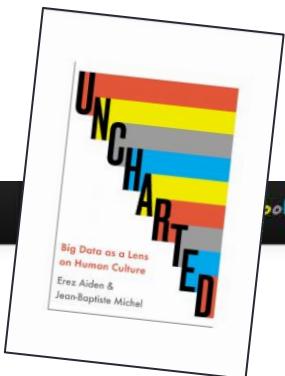
Google books Ngram Viewer

Graph these comma-separated phrases:   case-insensitive

between  and  from the corpus  with smoothing of



# Interest in visualization et al.



Google Ngrams + XKCD

About

Source Code

Culturomics

Plot Google Ngrams in XKCD Style

Graph these comma-separated phrases:

data visualization,infographics,information graphics,information design,visual analytics,statistical graphics,visual mining,data art,generative c

case-insensitive

between 1960

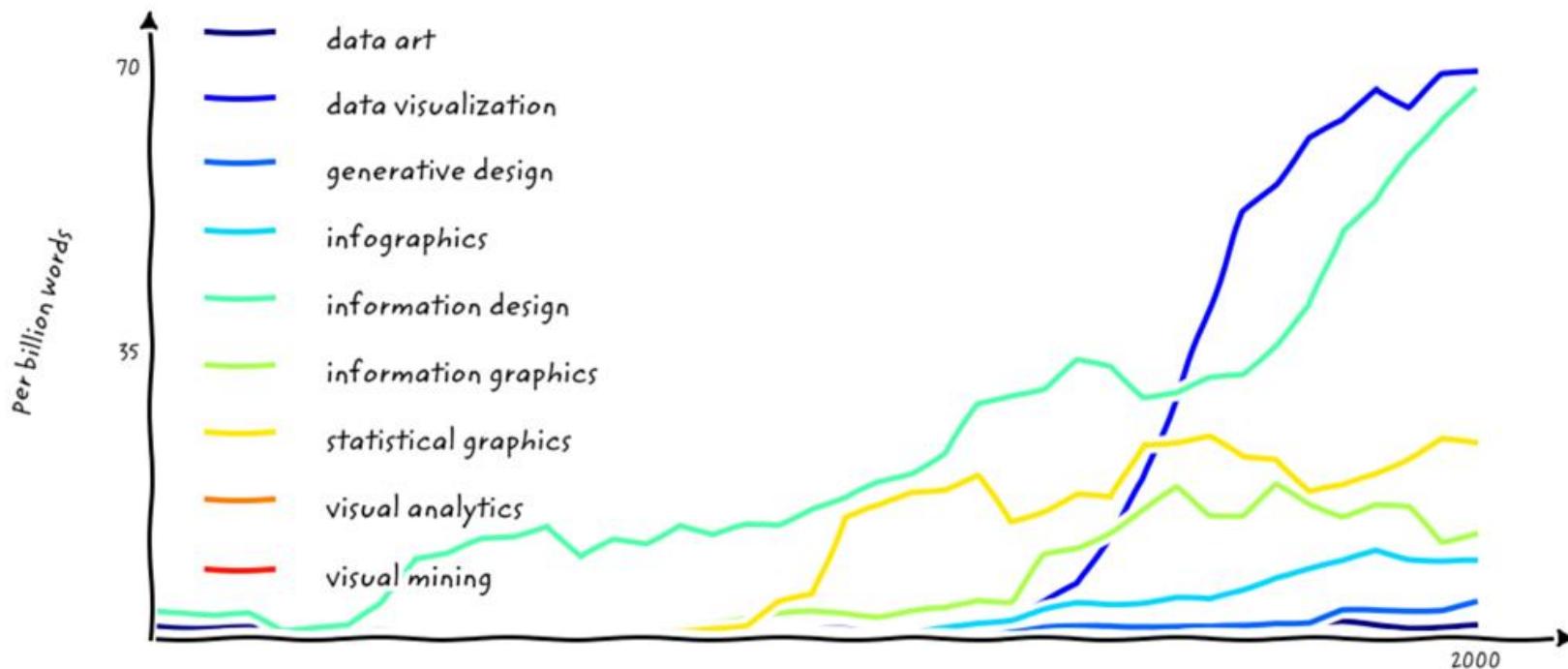
and 2000

from the corpus

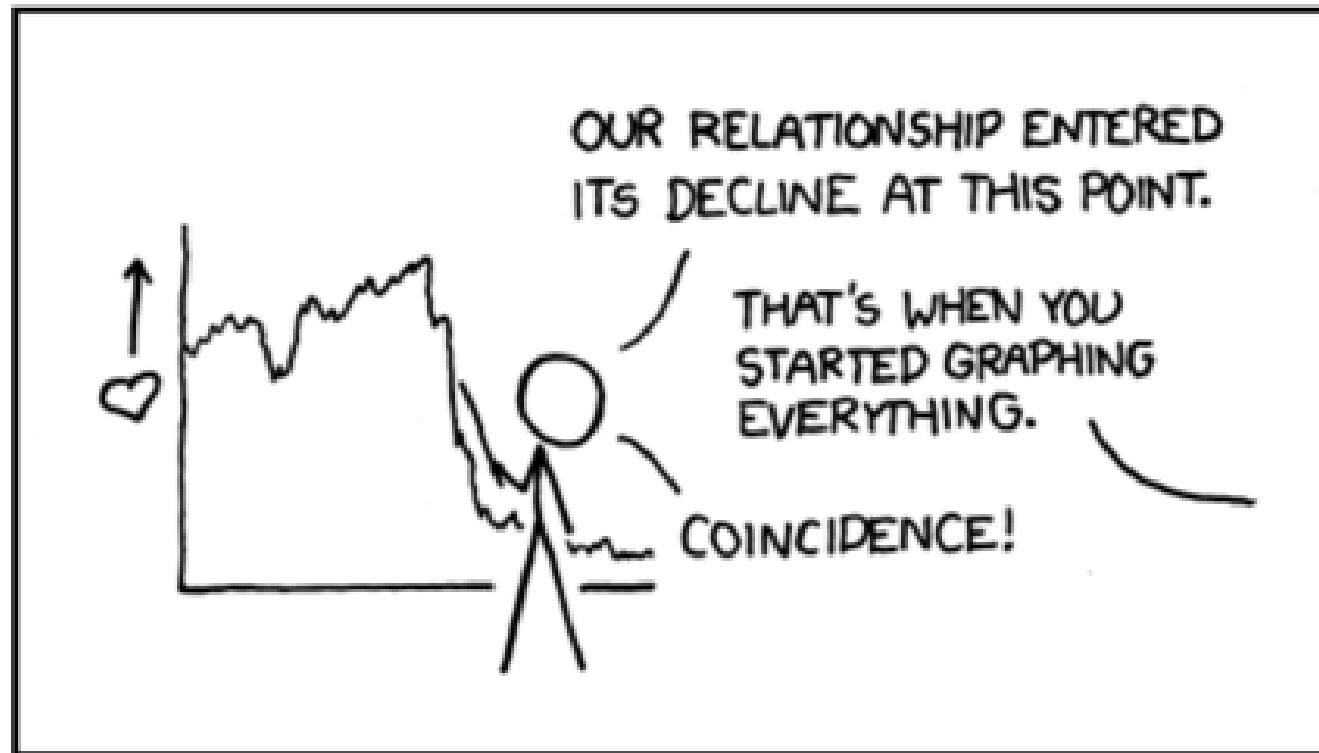
English

With smoothing of 2

Search lots of books



# Interest in Data Visualization and Infographics



Randall

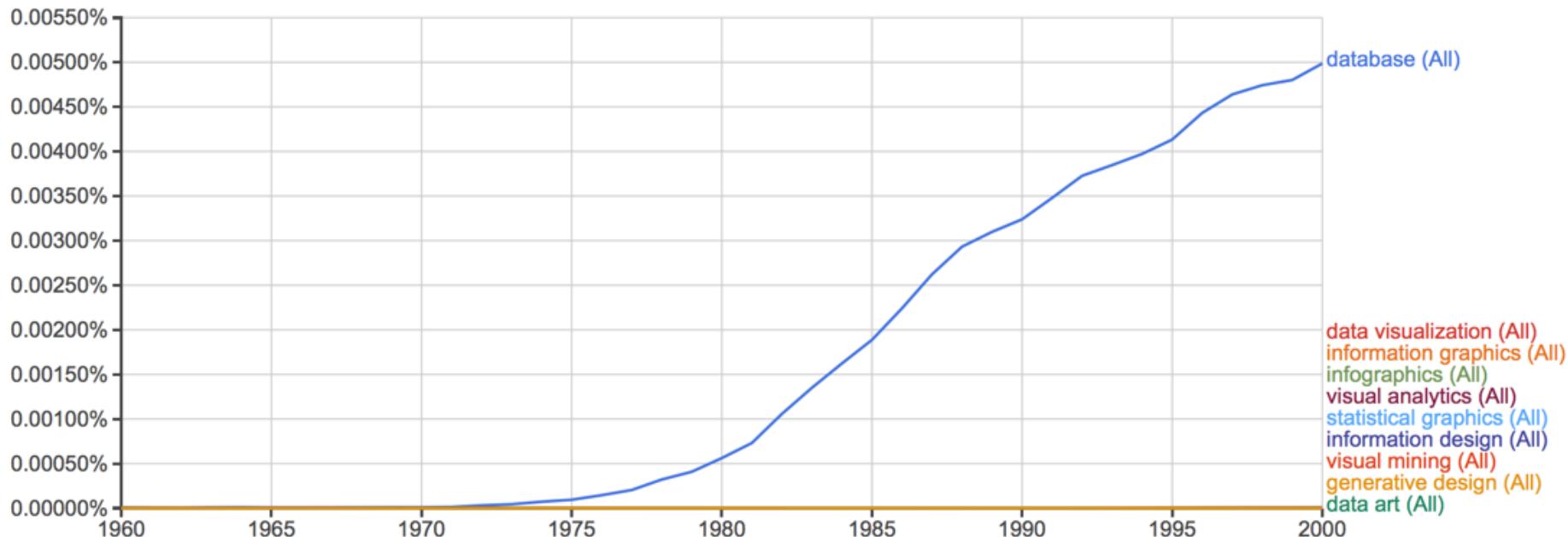
<http://xkcd.com/523/>

# Interest in visualization et al. *a bit of perspective*

## Google books Ngram Viewer

Graph these comma-separated phrases: database,data visualization,infographics,information graphics,inform:  case-insensitive

between  and  from the corpus  with smoothing of  .



# Interest in visualization

## *Google Trends*

Topics

[Subscribe](#)



data visu...

Search term

informati...

Search term

statistica...

Search term

informati...

Search term

infograph

Search term

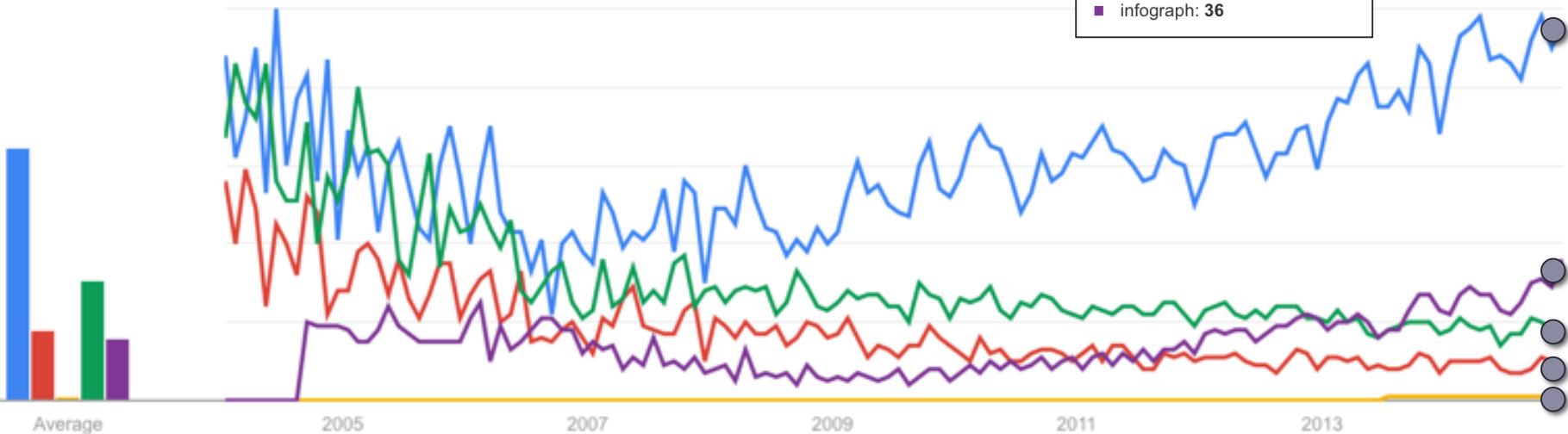
Interest over time



December 2014 (partial data)

- data visualization: 96
- information visualization: 7
- statistical visualization: 0
- information graphics: 15
- infograph: 36

Forecast



# Take one example

*Data Visualization - same term, multiple definitions*

- General term that describes any effort to help people understand the significance of data by placing it in a visual context.
- Overarching word for both the visual representation of data and the study of the presentation of data in a visual way.
- Pictorial representation of data that may take the form of an animation, a cloud, a map, a chart, or a simple picture.
- Study of the visual representation of data, meaning "information that has been abstracted in some schematic form, including attributes or variables for the units of information".
- Data visualization and information visualization refer to the use of computer-supported, interactive, visual representations of abstract data to amplify cognition

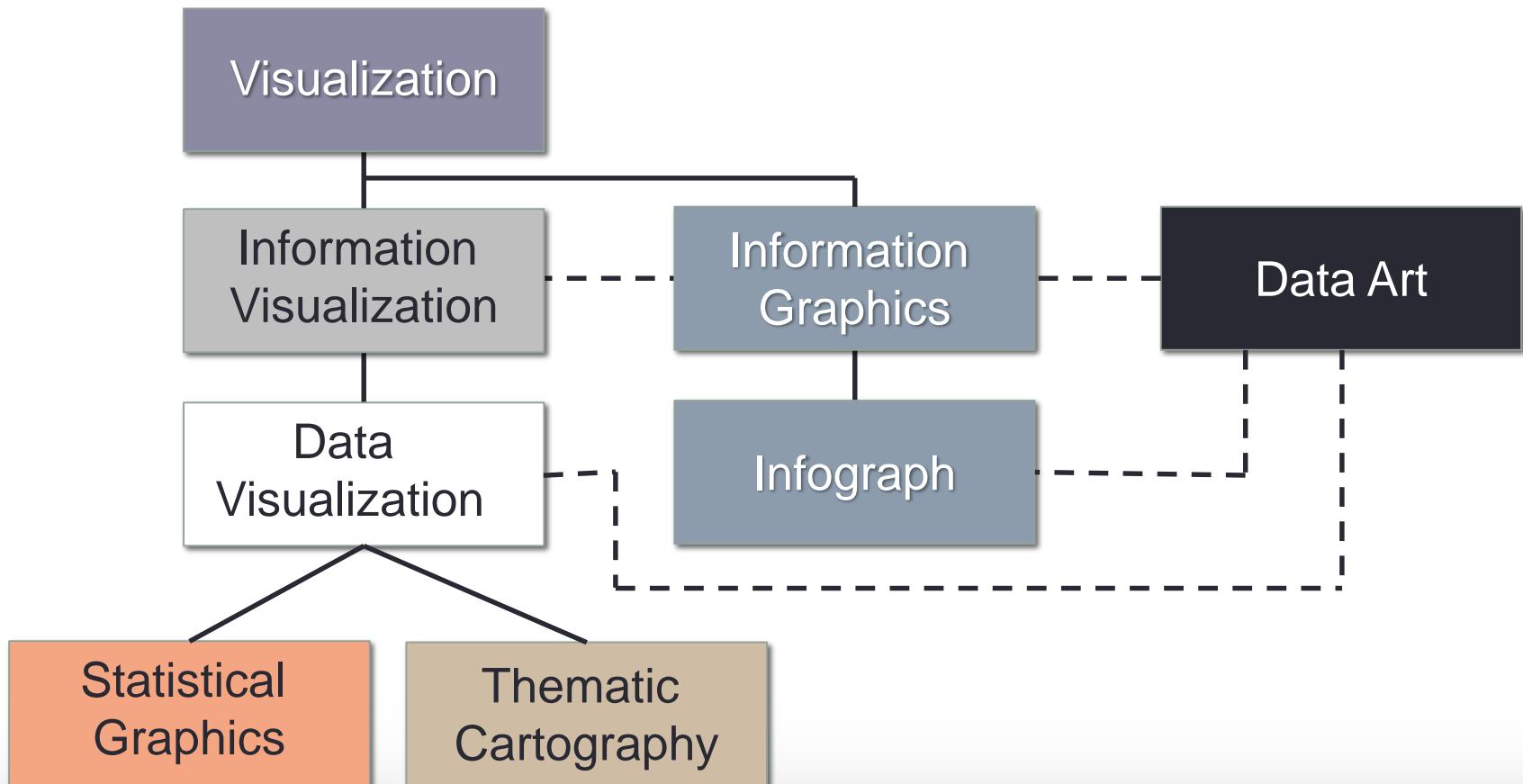
# Take one example

*Data Visualization - same term, multiple definitions*

- General term that describes any effort to help people understand the significance of **data** by placing it in a **visual** context.
- Overarching word for both the **visual representation** of **data** and the study of the presentation of data in a visual way.
- Pictorial **representation** of **data** that may take the form of an animation, a cloud, a map, a chart, or a simple picture.
- Study of the **visual representation** of **data**, meaning "information that has been abstracted in some schematic form, including attributes or variables for the units of information".
- Data visualization and information visualization refer to the use of computer-supported, interactive, **visual representations** of abstract **data** to amplify cognition

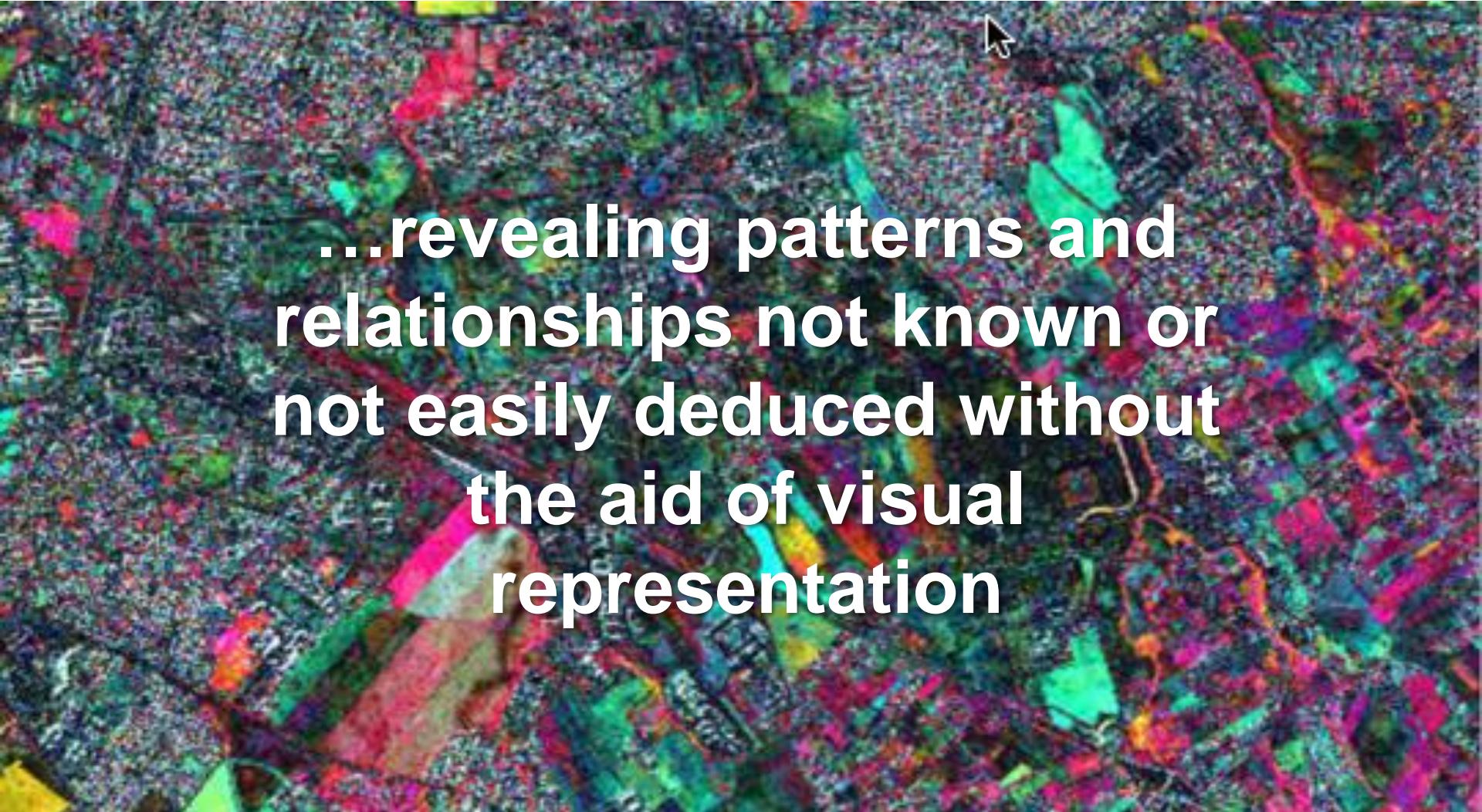
# Interrelated Terms

*One viewpoint*



# Purpose of Visualization

*“...Insight, not art”*

A vibrant, abstract visualization of a complex dataset, likely a 3D rendering of a brain or a network. The image is filled with a dense, granular texture of numerous small, multi-colored dots (red, green, blue, yellow) of varying sizes. Overlaid on this are several larger, more distinct shapes and patterns: a prominent red and yellow diagonal band, a green and yellow curved shape, and a blue and white curved shape. A small, dark cursor arrow is visible in the upper right corner, pointing towards the center of the visualization.

**...revealing patterns and  
relationships not known or  
not easily deduced without  
the aid of visual  
representation**

# Benefits of Data Visualization

## *Some claims*

- *Increases aware of essential facts*
- *Quickly see regularities*
- *Simpler to develop a deeper understanding of data*
- *Easier to convey information*
- *Plus it's less boring*

A <object1> is worth <#><object2>

A Drawing = 10 Pages in a Book

A Look = 1000 Words

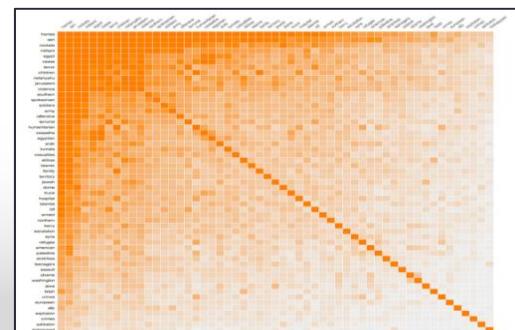
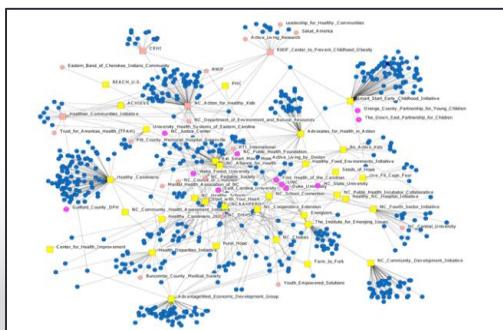
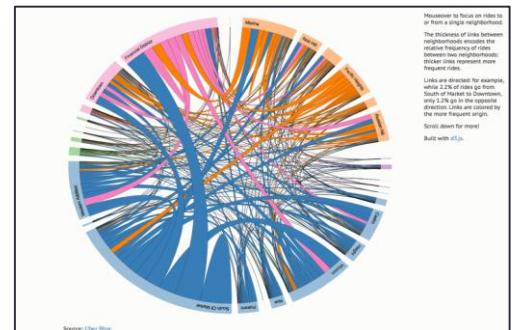
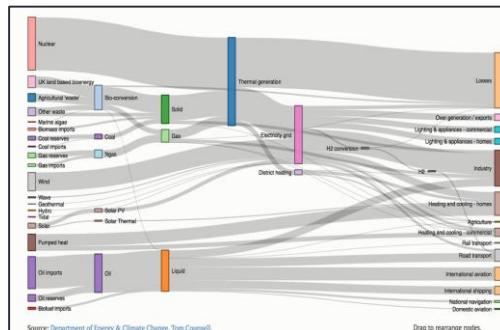
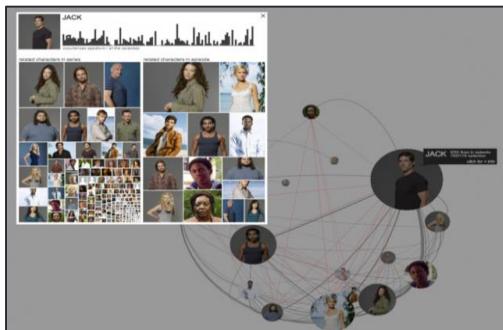
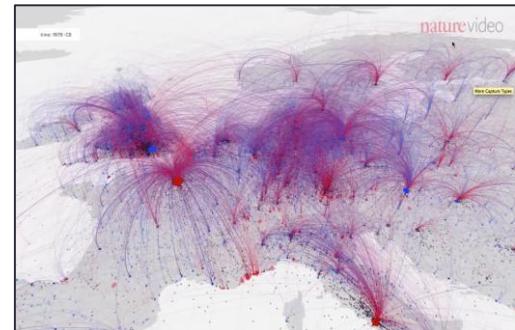
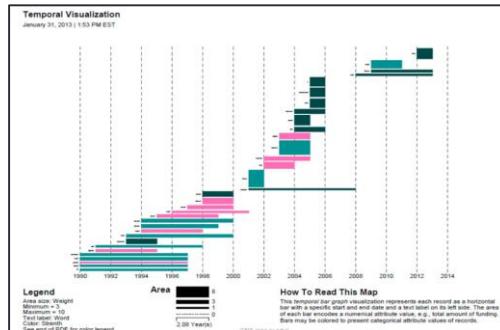
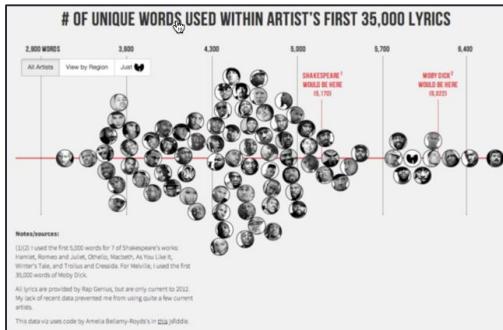
A Look = 100 Personal Letters/Ads

A Picture = 10,000 Words

A Picture = 84.1 Words

A picture tells a story just as well, or better, as a large amount of descriptive text.

# Is a visualization worth a 1000, 10000, 100000, ... of words, links, numbers,...?



# Some simple questions

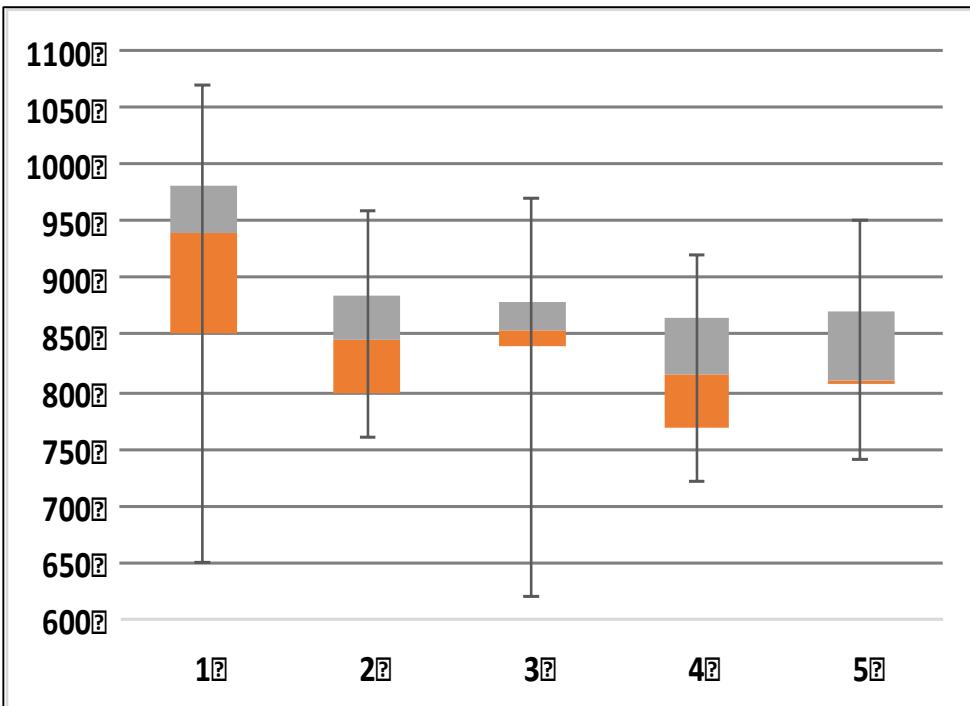
Michelson-Morley Data

Run 1	Run 2	Run 3	Run 4	Run 5
850	960	880	890	890
740	940	880	810	840
900	960	880	810	780
1070	940	860	820	810
930	880	720	800	760
850	800	720	770	810
950	850	620	760	790
980	880	860	740	810
980	900	970	750	820
880	840	950	760	850
1000	830	880	910	870
980	790	910	920	870
930	810	850	890	810
650	880	870	860	740
760	880	840	880	810
810	830	840	720	940
1000	800	850	840	950
1000	790	840	850	800
960	760	840	850	810
960	800	840	780	870

## What is the:

- Median value of Run 1?
- Minimum of Run 3? Max of Run 5?
- Which Run has the widest range?
- Which Run has the smallest interquartile range?
- ...

# Some simple questions



## What is the:

- Median value of Run 1?
- Minimum of Run 3? Max of Run 5?
- Which Run has the widest range?
- Which Run has the smallest interquartile range?
- ...

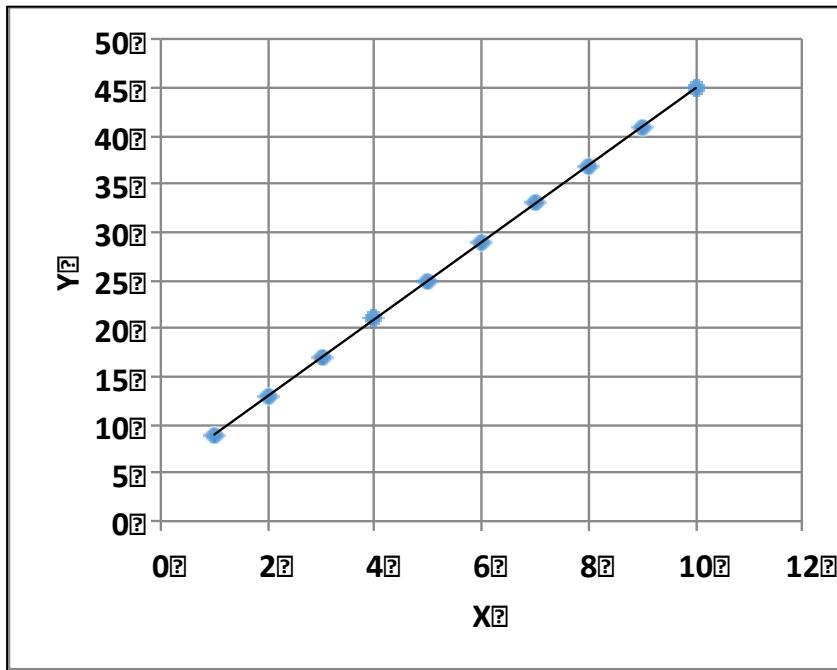
# Some additional simple questions

X	Y
1	9
10	45
2	13
9	41
3	17
8	37
4	21
7	33
5	25
6	29

**What is the  
relationship between  
X&Y?**

- Upward linear
- Downward linear
- Curvilinear
- Basically random

# Some additional simple questions



What is the  
relationship between  
X&Y?

- Upward linear
- Downward linear
- Curvilinear
- Basically random

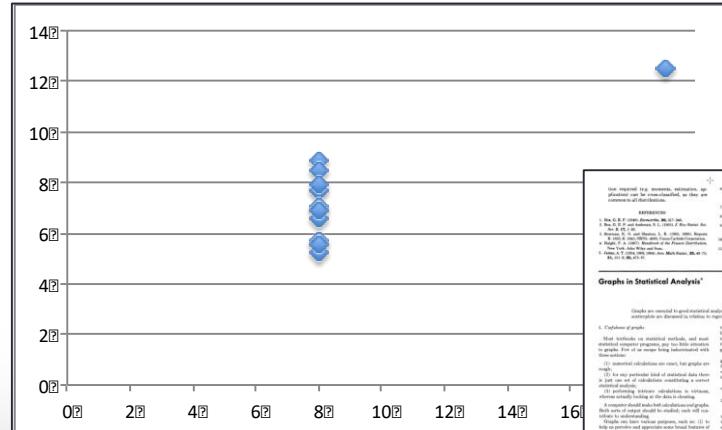
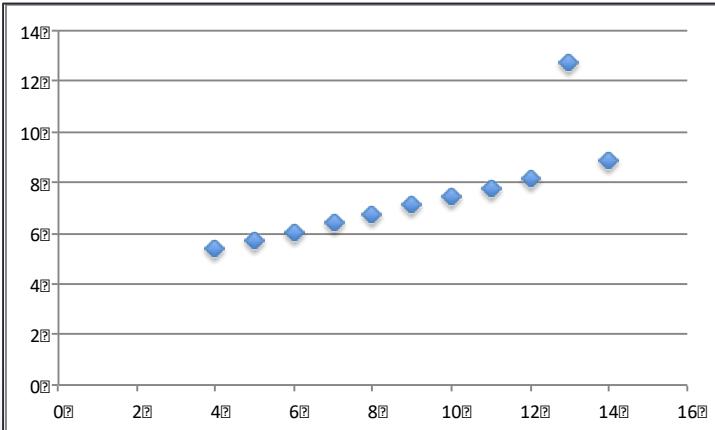
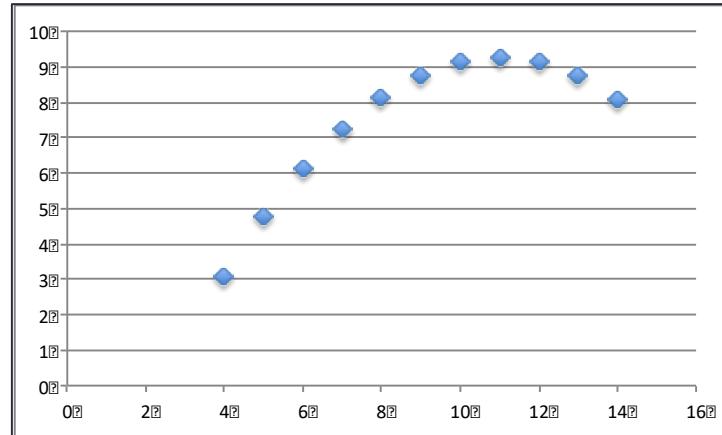
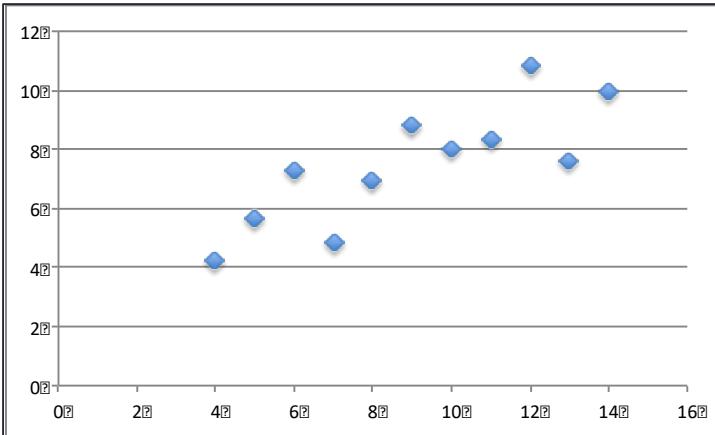
# Power of visualization

*Are you familiar with this data?*

	DataSet 1		DataSet 2		DataSet 3		DataSet 4	
	X1	Y1	X2	Y2	X3	Y3	X4	X4
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Std.Dev.	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
Correlation	0.82		0.82		0.82		0.82	

# The answer

## *Anscombe's Quartet (1973)*



2

4

## A Picture is Worth a Thousand Numbers

[SHARE](#) +

Posted on July 10, 2014 by CFOC Admin



By Douglas A. Glenn, Deputy Chief Financial Officer, Department of the Interior

There is no doubt the Federal financial industry has made great gains in financial reporting since the passage of the CFO Act and Government Management Reform Act. Material weaknesses are less than half of what they were in the mid-1990s and all but one CFO Act Agency has achieved an audit opinion. I am extremely proud of our industry's accomplishments. What makes me even more proud is the unsung value our industry provides the American people. The cumulative, government-wide, cost of the Federal Financial Community totaled billions of dollars including our IT and audit friends - critical partners in our industry's success. If you compare this cost to the total budget we execute, we cost less than 1/3 of 1% of the total. I am confident that our internal controls, reconciliations, audit support, and other efforts have prevented Billions of dollars in fraud, waste, and abuse. Just imagine the fraud, waste, and abuse that would occur if we weren't reconciling, reviewing, comparing, approving, and reporting Federal financial activity. We offer a significant return on investment to the American taxpayers and we take significant pride in that accomplishment.

Reporting is a key part of our community and with the recent passage of the DATA Act, there have been many discussions about what our industry reports as key financial data and an equally important consideration on how we report the information. Typically we use spreadsheets and slide decks to share information but as fascinating as spreadsheets and slide decks are, they are cumbersome at best. At the Department of the Interior (DOI), we are pushing the boundaries of communicating quantitative information and are fans of the concept that "A picture is worth a thousand numbers". For example, with one picture we can show where our 2013 expenditures went, who they went to, and what they were for (see <http://www.doi.gov/pfm/afr/2013/maps/index.cfm>). It is a "heat map" of the United States where each State is shaded based on the DOI funding disbursed to that State. If you run your cursor over a State, you can see the total expenditures to that State, the top object classes related to those expenditures, or who were the top recipients (i.e., vendors) of our disbursements.



### News/Events

- » Blog
- » Events
- » News

### Archives

- » November 2014
- » October 2014
- » August 2014
- » July 2014
- » June 2014
- » May 2014
- » April 2014
- » January 2014
- » December 2013

## A Picture is Worth a Thousand Numbers

SHARE +

Posted on July 10, 2014 by CFOC Admin



By David A. Glavin, CFOC Admin - Department of the Interior

There  
Manager  
audit  
the A  
friend  
am c  
Just I

To illustrate the value of pictures versus spreadsheets, I handed out the 60 row spreadsheet that supports the picture in the heat map noted above and asked a room of 25 accountants: "Who were the top 10 states receiving DOI disbursements?" They dutifully scanned their spreadsheets and after 106 seconds, the first brave hand was raised and the accountant correctly answered the question. Next a previously selected volunteer was pulled (back) into the room, shown the heat map, and asked the same question. The one volunteer (versus 25 other accountants with spreadsheets) reached the same conclusion, to the same question, in 1/3 of the time. Such is the power of pictures versus spreadsheets.

activity. We offer a significant return on investment to the American taxpayers and we take significant pride in that accomplishment.

Reporting is a key part of our community and with the recent passage of the DATA Act, there have been many discussions about what our industry reports as key financial data and an equally important consideration on how we report the information. Typically we use spreadsheets and slide decks to share information but as fascinating as spreadsheets and slide decks are, they are cumbersome at best. At the Department of the Interior (DOI), we are pushing the boundaries of communicating quantitative information and are fans of the concept that "A picture is worth a thousand numbers". For example, with one picture we can show where our 2013 expenditures went, who they went to, and what they were for (see <http://www.doi.gov/pfm/afr/2013/maps/index.cfm>). It is a "heat map" of the United States where each State is shaded based on the DOI funding disbursed to that State. If you run your cursor over a State, you can see the total expenditures to that State, the top object classes related to those expenditures, or who were the top recipients (i.e., vendors) of our disbursements.



### News/Events

- » Blog
- » Events
- » News

- » April 2014
- » January 2014
- » December 2013

# DOI Spreadsheet

BOC-Data-Tables.xls

Search in Sheet

Home Layout Tables Charts SmartArt Formulas Data Review

C53 fx 9271

All Of BOC

State/Territory	Top Budget Object Classes									
	1		2		3		4		5	
Description	Amount	Description	Amount	Description	Amount	Description	Amount	Description	Amount	More Captur
Alabama	Grants	37,892	DOI Salaries and Related Expenses	16,170	Contracts - Professional Services	4,221	Oil, Gas and Mineral Revenues	2,349	Cooperative Agreements	2,010
Alaska	DOI Salaries and Related Expenses	244,081	Contracts - Indian Self-Determination Services	88,295	Grants	56,565	Cooperative Agreements	36,789	PILT - Other Revenue Sharing	26,717
American Samoa	Grants	9,038	Space Rental Payments To Others	587	Freight - Other	24	Other	15	Grants to Insular Areas	11
Arizona	DOI Salaries and Related Expenses	343,350	Determination Services	209,920	Indian Tribal Government Grant	142,152	PILT - Other Revenue Sharing	32,575	Grants	27,859
Arkansas	DOI Salaries and Related Expenses	29,917	Grants	22,135	PILT - Other Revenue Sharing	6,508	Contracts - Airplanes & Helicopters	3,858	Capitalized - Land Acquisition	3,019
California	DOI Salaries and Related Expenses	526,373	Contracts - Professional Services	126,015	Oil, Gas and Mineral Revenues	93,817	Grants	84,120	Contracts - Training Services	53,677
Colorado	DOI Salaries and Related Expenses	596,784	Oil, Gas and Mineral Revenues	129,661	Contracts - Professional Services	78,310	Reimbursable Agreements - Internal	46,794	Other	42,971
Connecticut	Grants	20,897	DOI Salaries and Related Expenses	5,871	Determination Services	2,685	Cooperative Agreements	1,472	Other	1,101
Delaware	Grants	7,661	DOI Salaries and Related Expenses	4,052	Cooperative Agreements	664	Private Sector - R & D	512	Contracts - Professional Services	283
District of Columbia	Reimbursable Agreements - Internal	297,934	Agency	221,692	Medical and Health Care Services	197,408	Contracts - Professional Services	164,958	Other	162,091
Florida	DOI Salaries and Related Expenses	123,379	Grants	32,836	Other	13,013	Contracts - Indian Self-Determination Services	11,023	Cooperative Agreements	10,921
Georgia	DOI Salaries and Related Expenses	96,837	Grants	29,055	Contracts - Airplanes & Helicopters	12,912	Contracts - Professional Services	6,858	Other	5,744
Guam	Grants to Insular Areas	84,744	Grants	26,848	Capitalized - Buildings - Constructed	120	Cooperative Agreements	84	Operations Maintenance & Repairs - Other Structures	64
Hawaii	DOI Salaries and Related Expenses	\$ 55,277	Grants	\$ 26,911	Cooperative Agreements	\$ 5,644	Agency	\$ 3,090	Contracts - Professional Services	\$ 1,315
Idaho	DOI Salaries and Related Expenses	157,131	PILT - Other Revenue Sharing	26,622	Contracts - Airplanes & Helicopters	23,486	Contracts - Indian Self-Determination Services	18,348	Grants	17,421
Illinois	Grants	36,215	DOI Salaries and Related Expenses	21,859	Cooperative Agreements	10,743	Capitalized - Heavy Machinery	6,666	Private Sector - R & D	4,529
Indiana	Grants	46,766	DOI Salaries and Related Expenses	22,814	Space Rental Payments To Others	3,393	Contracts - Professional Services	2,694	Cooperative Agreements	2,392
Iowa	Grants	18,918	DOI Salaries and Related Expenses	14,673	Indian Tribal Government Grant	5,548	Cooperative Agreements	2,838	Non-Capitalized - Furniture & Fixtures Non-Contri	989
Kansas	DOI Salaries and Related Expenses	34,311	Grants	23,833	Contracts - Professional Services	9,041	Contracts - Indian Self-Determination Services	3,798	Space Rental Payments To Others	2,794

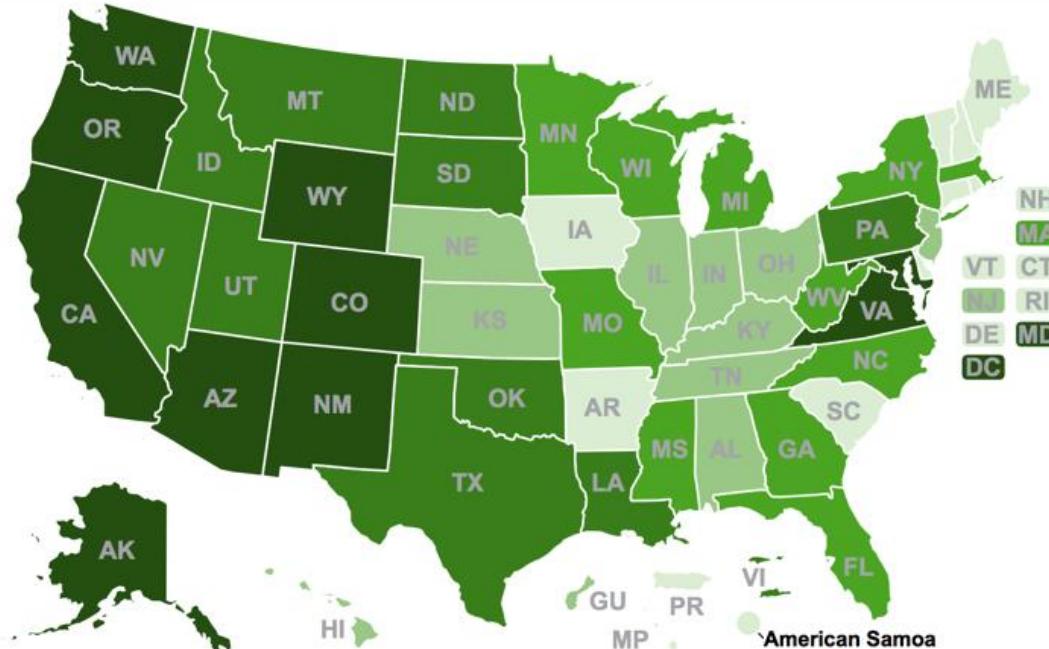
BOC Data Table 1 BOC Data Table 2 BOC Descriptions +

Normal View Ready Sum= 9,271

## FY 2013 Department of the Interior Data Maps



### Select Map Data:

[Expenditures by Budget Object Class](#)[Expenditures by Recipient](#)[Oil and Gas Petroleum Royalties](#)

### Domestic Expenditures by State and Budget Object Class (in thousands)

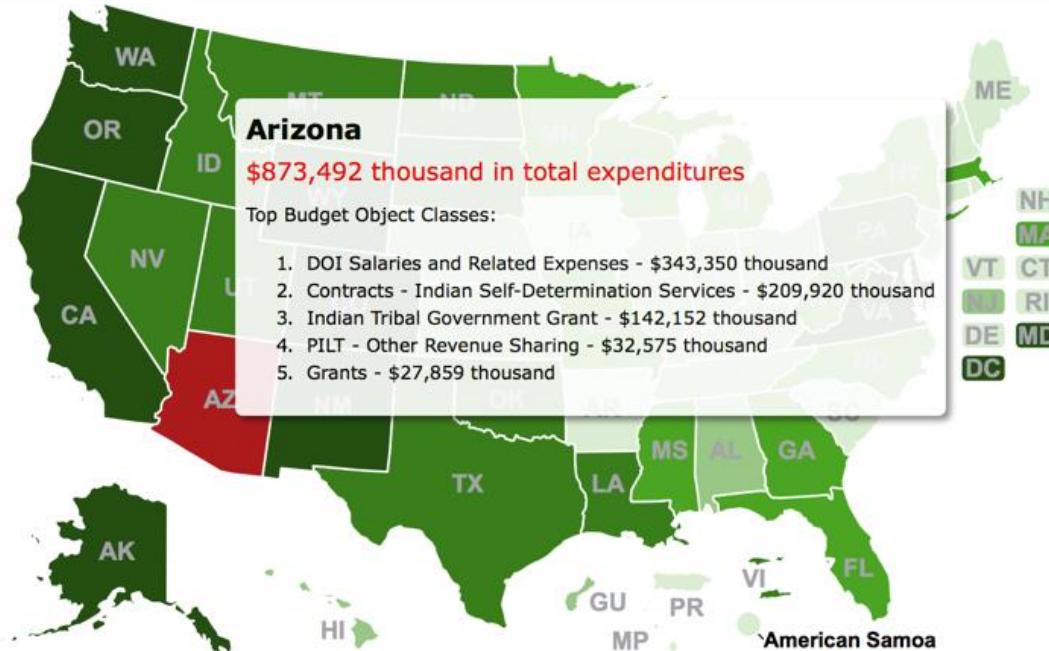
\$0 -- \$72,000	\$72,001 -- \$119,000	\$119,001 -- \$257,000	\$257,001 -- \$450,000	\$450,001 -- \$1,819,000
-----------------	-----------------------	------------------------	------------------------	--------------------------



## FY 2013 Department of the Interior Data Maps



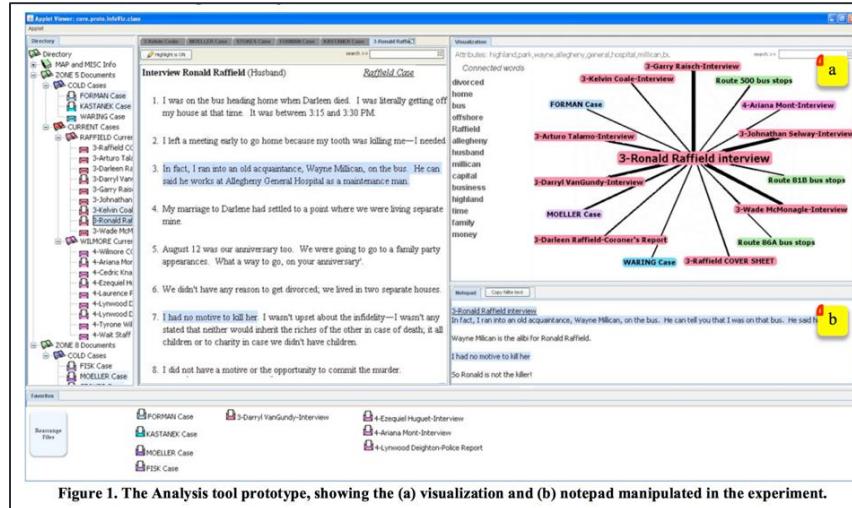
## Select Map Data:

[Expenditures by Budget Object Class](#) [Expenditures by Recipient](#) [Oil and Gas Petroleum Royalties](#)**Domestic Expenditures by State and Budget Object Class (in thousands)**

\$0 --- \$72,000	\$72,001 --- \$119,000	\$119,001 --- \$257,000	\$257,001 --- \$450,000	\$450,001 --- \$1,819,000
------------------	------------------------	-------------------------	-------------------------	---------------------------



# A visualization is worth a 100 words, but a few numbers are worth 1 visualization?



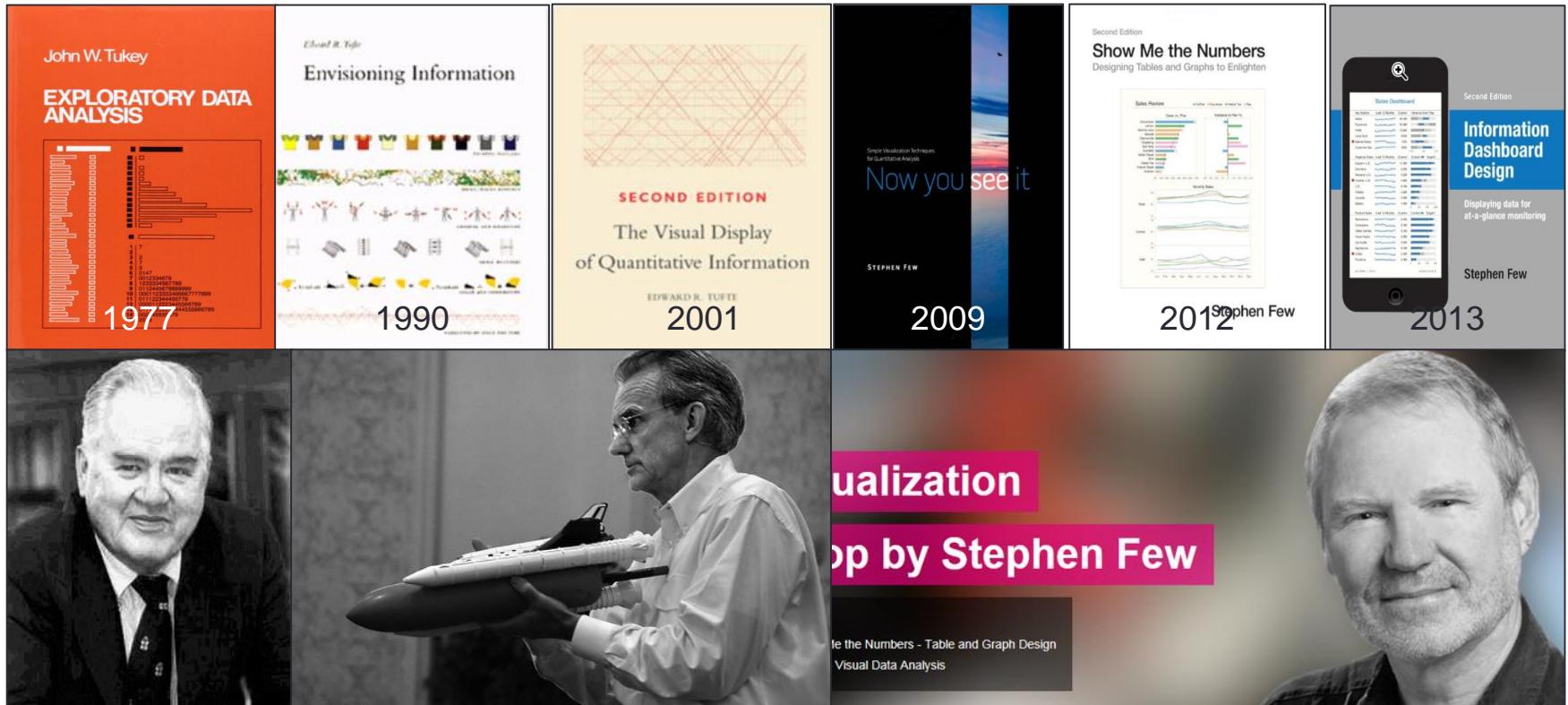
Feature	Serial Killer Identification (%)	Clue Recall (0 to 9)	Relationship Detection (0 to 1)	Time Spent on Features (%)	Perceived Feature Usefulness (0 to 5)
Vis. Only	90 (SD=10)	4.5 (SD=0.34)	0.85 (SD=0.07)	12.61 (SD=1.97)	3.66 (SD=0.44)
Notepad Only	30 (SD=15)	1.4 (SD=0.47)	0.35 (SD=0.10)	8.64 (SD=2.82)	3.90 (SD=0.37)
Visualization and Notepad	70 (SD=15)	3.9 (SD=0.73)	0.60 (SD=0.10)	V:16.11 (SD=5.98) N:14.61 (SD=6.01)	V:4.30 (SD=0.21) N:3.70 (SD=0.42)
Control	50 (SD=16)	2.0 (SD=0.63)	0.35 (SD=0.10)	N/A	N/A

**Table 1 Descriptive Summary (Means and Std. Deviations) across the four conditions for the five measures**

# A fundamental clash Data Art vs. Data Visualization

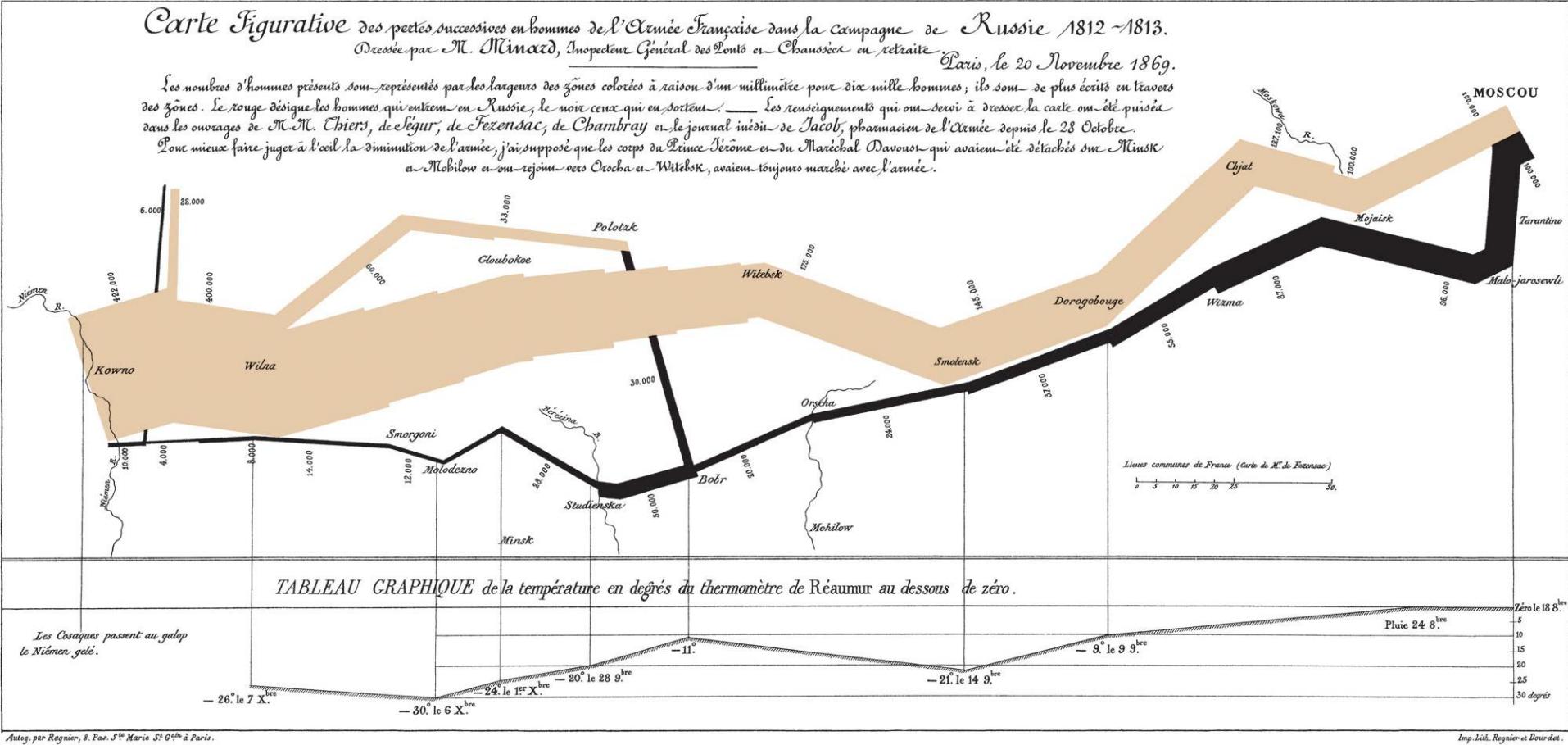
- There has always been a fundamental clash in information graphics and visualization between those who favor a rational, scientific approach to the profession, emphasizing functionality, and those who consider themselves “artists,” placing emphasis on emotion and esthetics (Albert Cairo)
- Two distinct approaches to presenting data graphically exist today—data visualization and data art—and rarely do the twain meet. They differ in purpose and in design. When we fail to distinguish them from one another, we not only create confusion, but do great harm as well (Stephen Few)
  -

# On one side

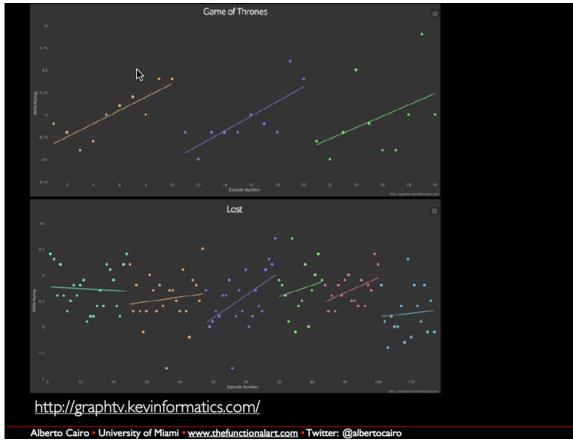


# The perfect information graphic

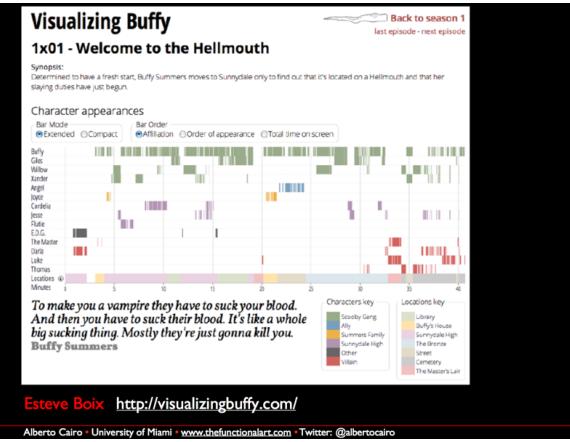
## The designer and how many variables?



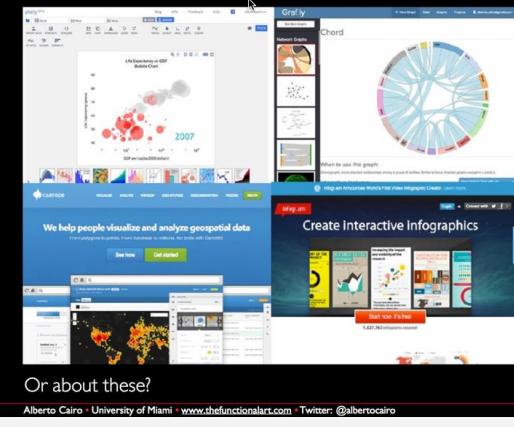
# The imperfect visualizations



Alberto Cairo • University of Miami • [www.thefunctionalart.com](http://www.thefunctionalart.com) • Twitter: @albertocairo



Alberto Cairo • University of Miami • [www.thefunctionalart.com](http://www.thefunctionalart.com) • Twitter: @albertocairo



Alberto Cairo • University of Miami • [www.thefunctionalart.com](http://www.thefunctionalart.com) • Twitter: @albertocairo

do projects like this represent the future of visualization?

front the point of view of tools, sure, but...

Alberto Cairo • University of Miami • [www.thefunctionalart.com](http://www.thefunctionalart.com) • Twitter: @albertocairo

But I think that the future of visualization doesn't depend on the tools we use, but on solid principles

Instead of thinking what the future of visualization will be, I perhaps should talk about what I would like the future of news visualization to be

Alberto Cairo • University of Miami • [www.thefunctionalart.com](http://www.thefunctionalart.com) • Twitter: @albertocairo

Aren't we talking too much about these?

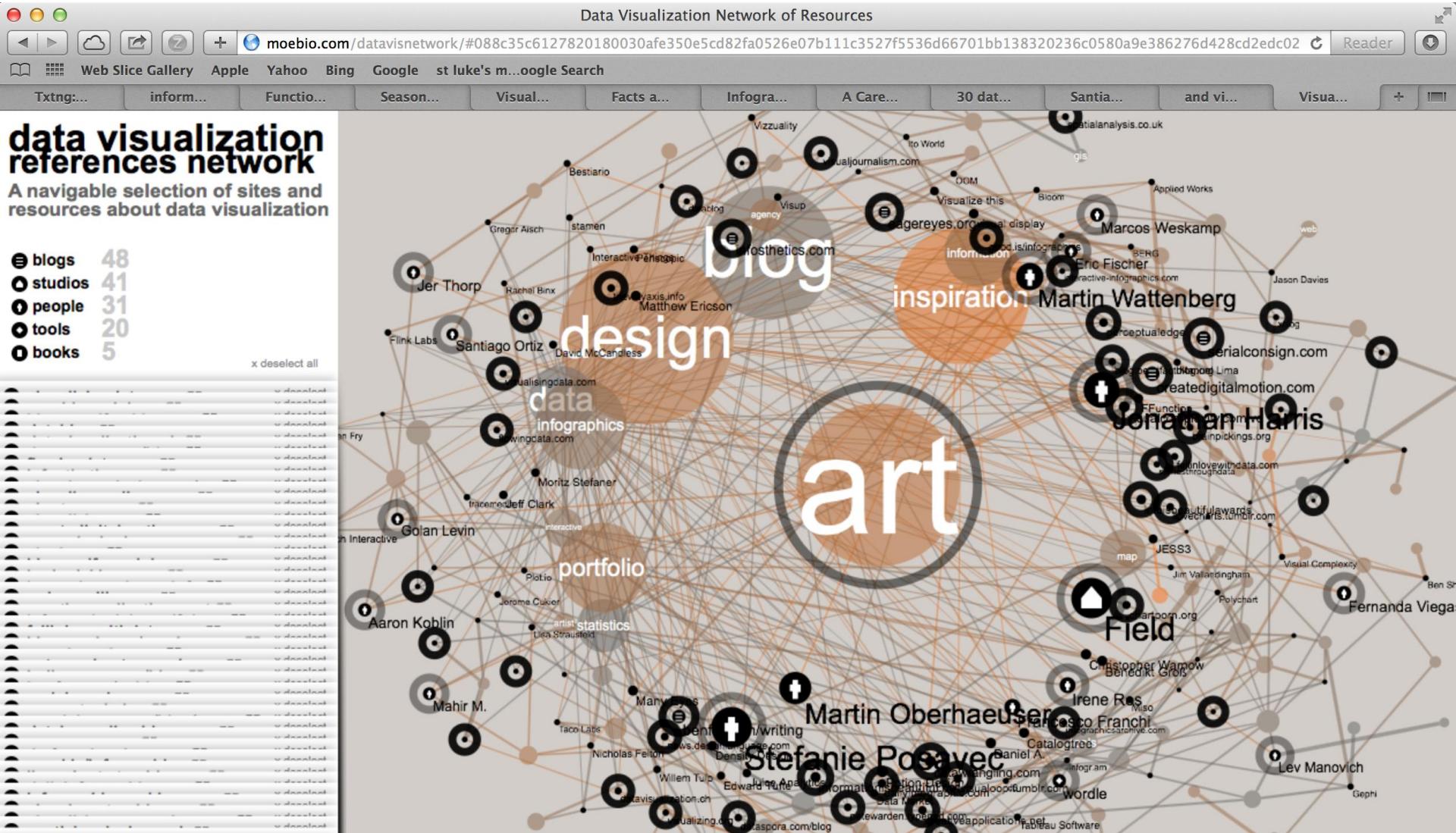
Alberto Cairo • University of Miami • [www.thefunctionalart.com](http://www.thefunctionalart.com) • Twitter: @albertocairo

# What's the harm?

- It suggests that data cannot be visualized without training in the graphic arts. As such, it works against the democratization of data
- It features ineffective practices as exemplars of data visualization
- It keeps the practice of data visualization spinning its wheels, never able to progress beyond the mistakes of the past.

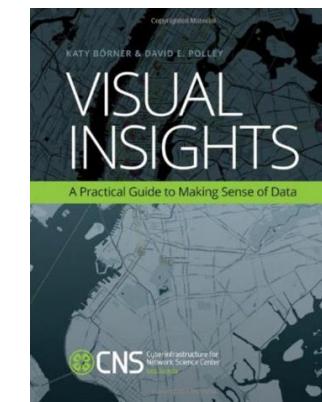
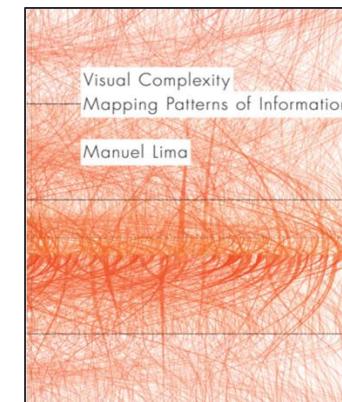
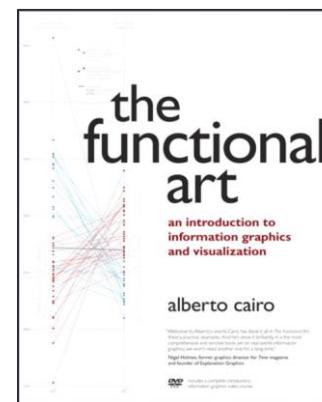
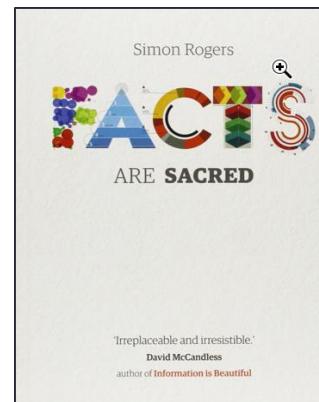
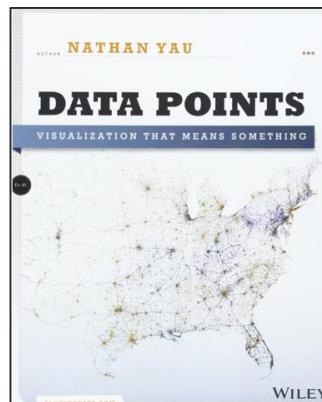
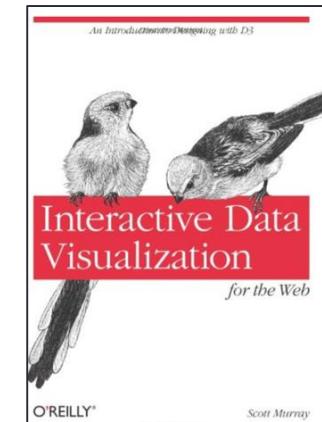
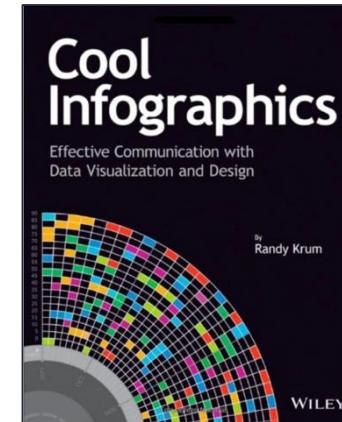
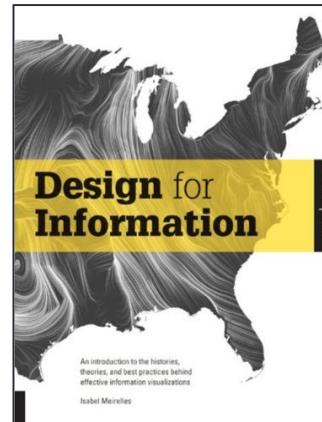
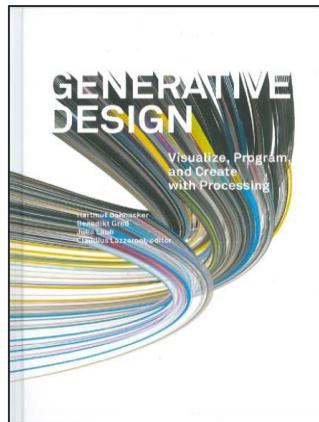
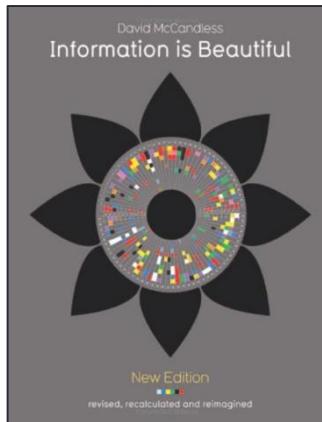
# On the other side

## *Proponents of Data Art*



# On the other side

## *Proponents of (some variation of) Data Art*



# Artifying Data

*It's more than just "Chartjunk"*

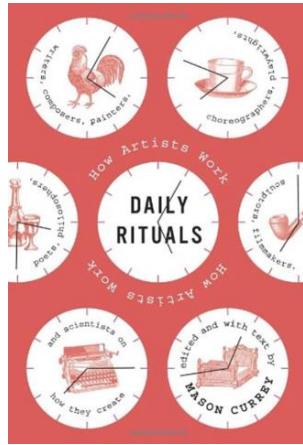
- More than adornment (or ornamentation)
- Applying artistic techniques to design choices
  - Strengthen by learning basic art elements and design principles
- Expand and enrich the experience
- Open new ways of seeing the world around us
- Encourage others and share ideas with fresh new vision

# SIMPLE EXAMPLE: QUANTIFICATION OF SELF

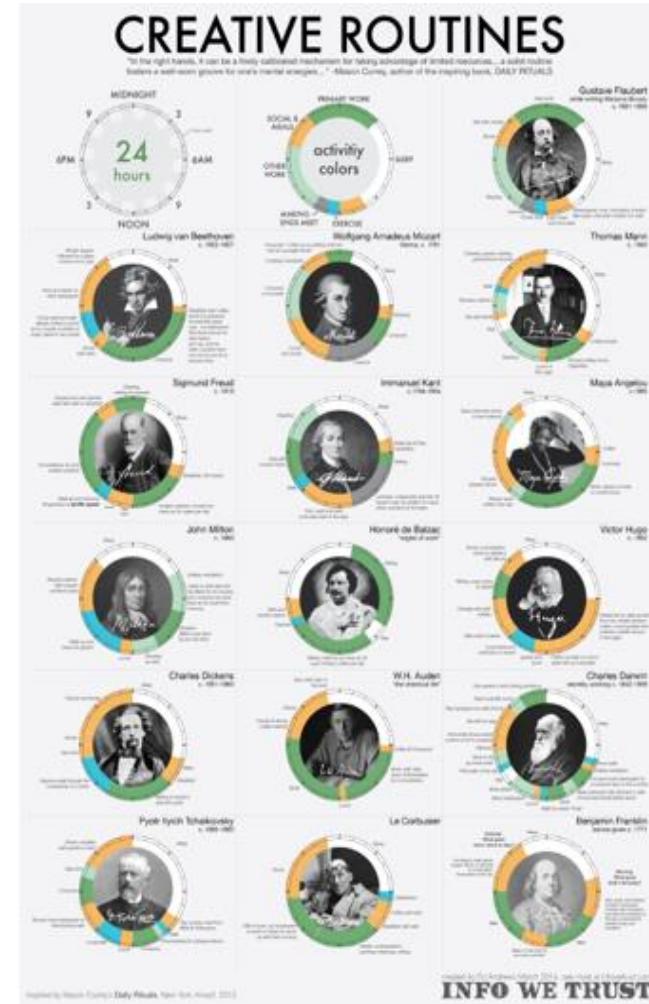
---

# Daily rituals of how (161) artists work

## Mason Currey



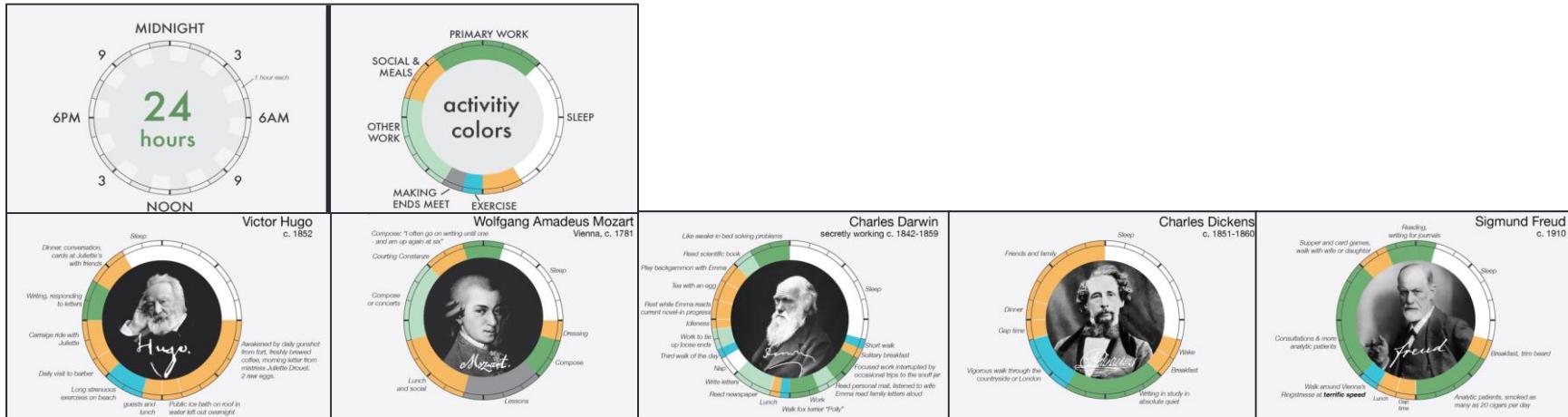
It's about the circumstances of creative activity, not the product; it deals with manufacturing rather than meaning.



INFO WE TRUST

# Daily rituals of how (161) artists work

## *Some simple questions*



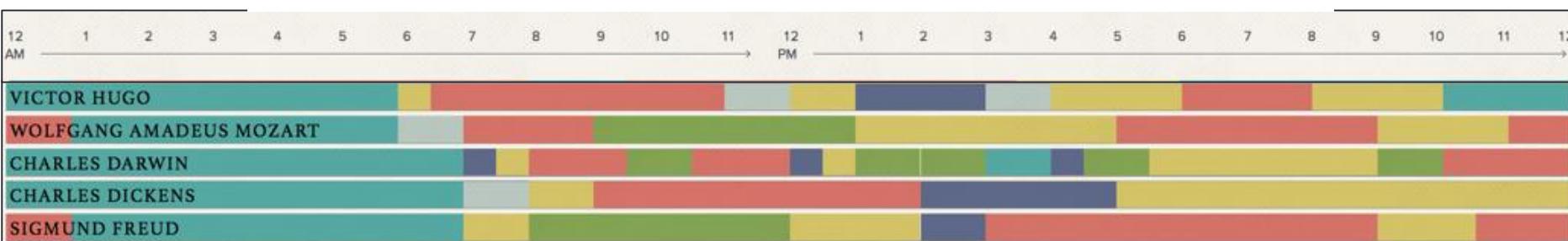
SLEEP  
OTHER

CREATIVE WORK

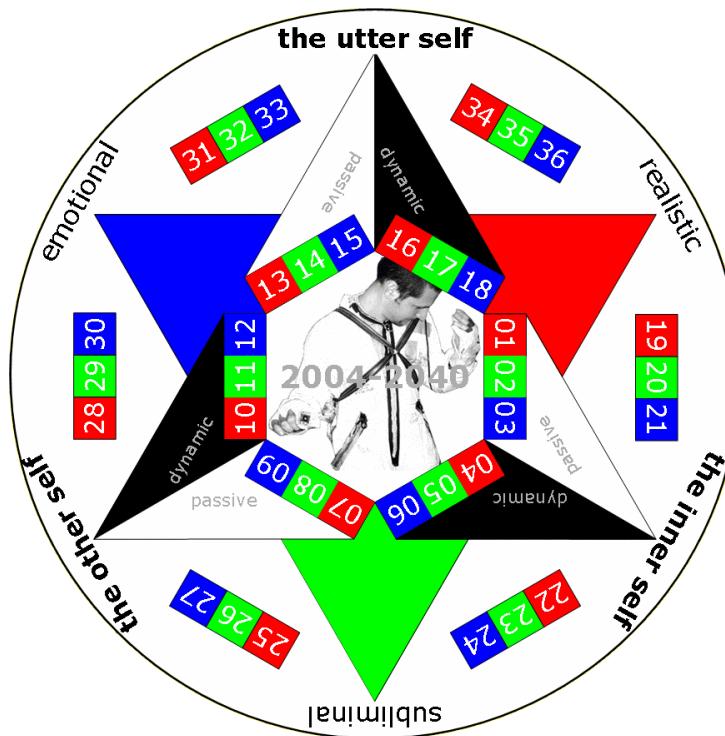
DAY JOB/ADMIN

FOOD/LEISURE

EXERCISE



# You've heard of "My Left Foot" How about Albert Frigo's Right Hand?



*It has been more than 10 years since I have started that project, to be precise today the 11th of May 2014, is my 3.882nd day I have been photographing every object my right hand uses.*

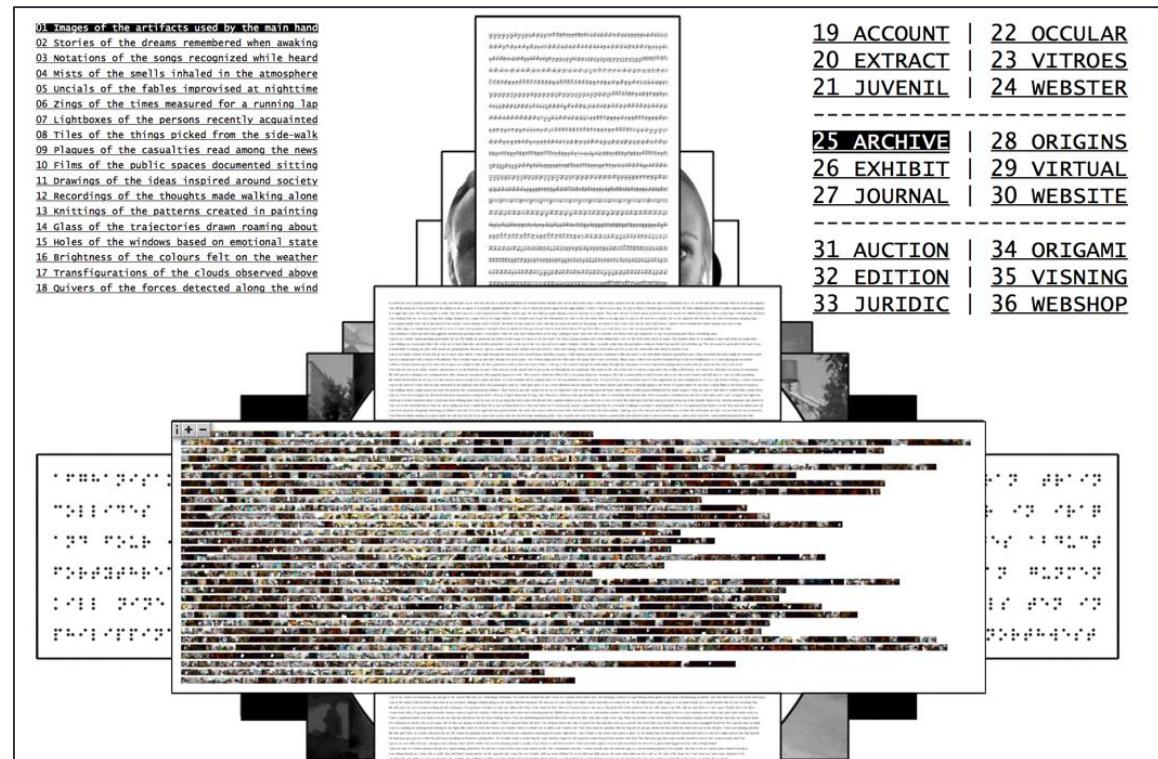
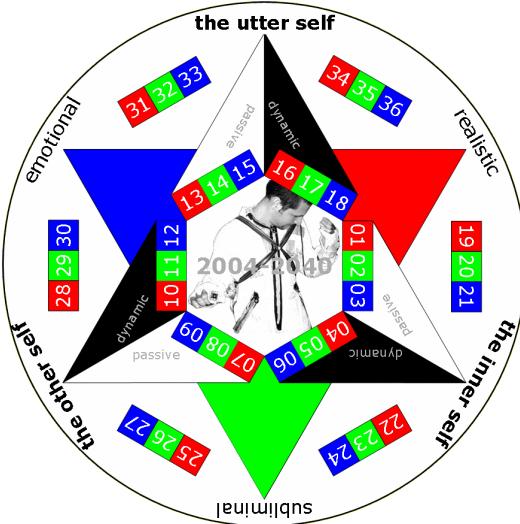
# Alfredo Frigo

*Pause in the action*



*If he had been left handed, he'd be taking a picture right now.*

# Visualizing the Right Hand



# Larger Movements



- Health, Wellness & Fitness
- Personal and Professional Productivity
- Education and Learning
- Gamification
- Child Rearing (Quantified Babies)
- ...

# Lifelogging Visualizations

## *Mapping movements from 2008-2012*



[aaronparecki.com](http://aaronparecki.com)

Aaron Parecki

[Articles](#)

[Pages](#)

[Notes](#)

[Bookmarks](#)

[Replies](#)

[Presentations](#)

[Metrics](#)

[Contact](#)

Aaron Parecki is CTO of [Esri's Portland R&D Center](#), and the co-founder of [IndieWebCamp](#). He is known for having tracked his location at 5 second intervals since 2008, and for co-founding Geoloqi, a location-based software company acquired by Esri in 2012. His work has been featured in [Wired](#), [Fast Company](#) and more. He was featured in [Inc.](#) Magazine's 30 Under 30 with Geoloqi co-founder Amber Case.

### Things I track consistently

- **Location: GPS** (since 2008) iPhone
- **Location: Checkins** (since 2009) Foursquare
- **Sleep** (since November 2011) Jawbone UP
- **Weight** (since October 2011) Withings Scale
- **Steps / Activity** (since November 2011) Jawbone

@aaronpk

# Lifelogging Visualization

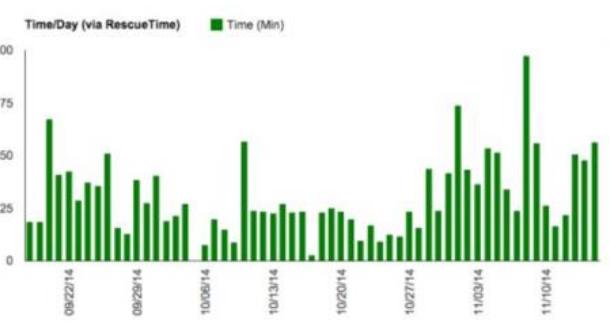
*Mapping movements from 2008-2012*

2.5 million GPS points over 3 1/2 years between 2008 and 2012, about one point every 2 to 6 mins

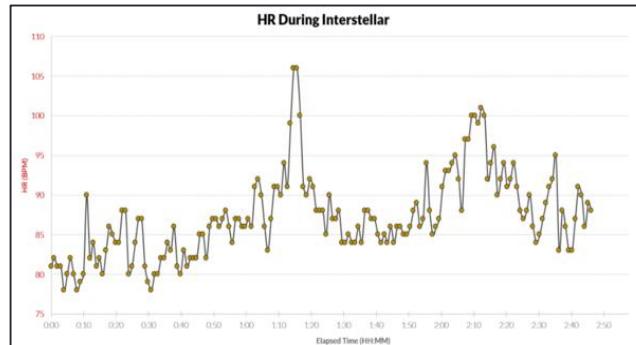


# Lifelogging Visualizations *BAPPS*

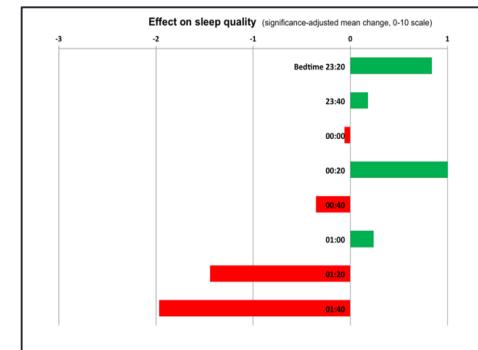
Mins Writing - Daily



Heart Rate - Mins



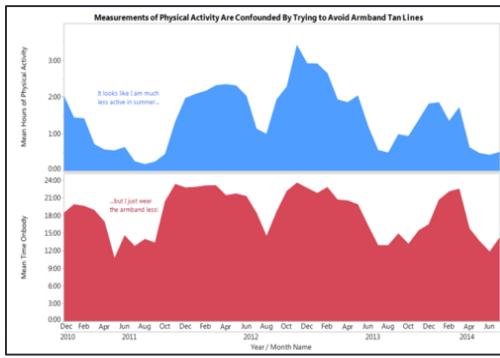
Sleep Quality - Hourly



Weight 1988 - 2014



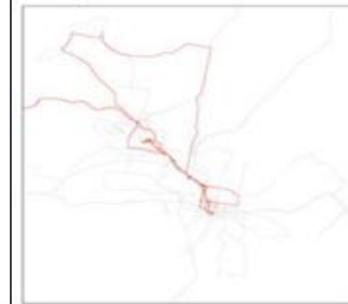
Measurements of Physical Activity Are Confounded By Trying to Avoid Armband Tan Lines



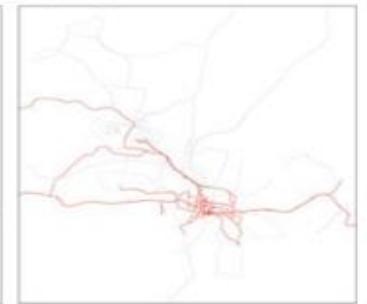
Mins Writing - Daily

Avg HRs Phy Act - Mnth

Dating

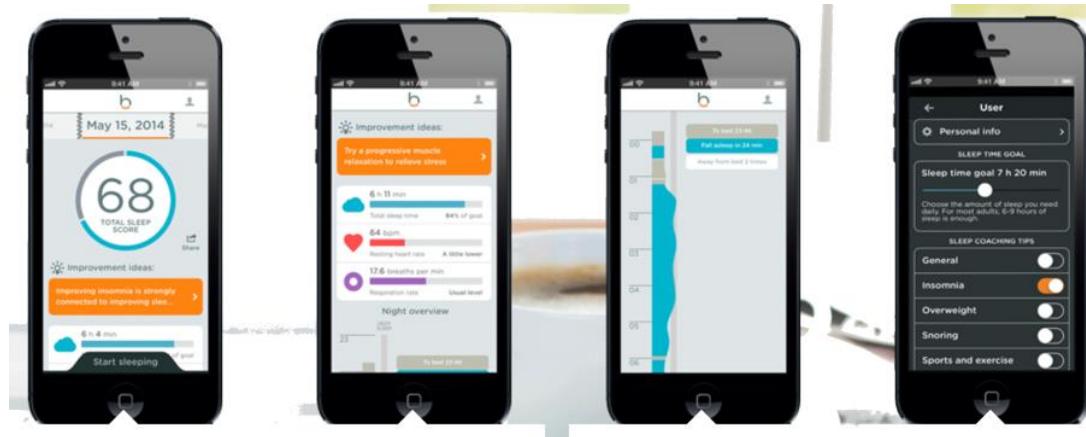


Exploring

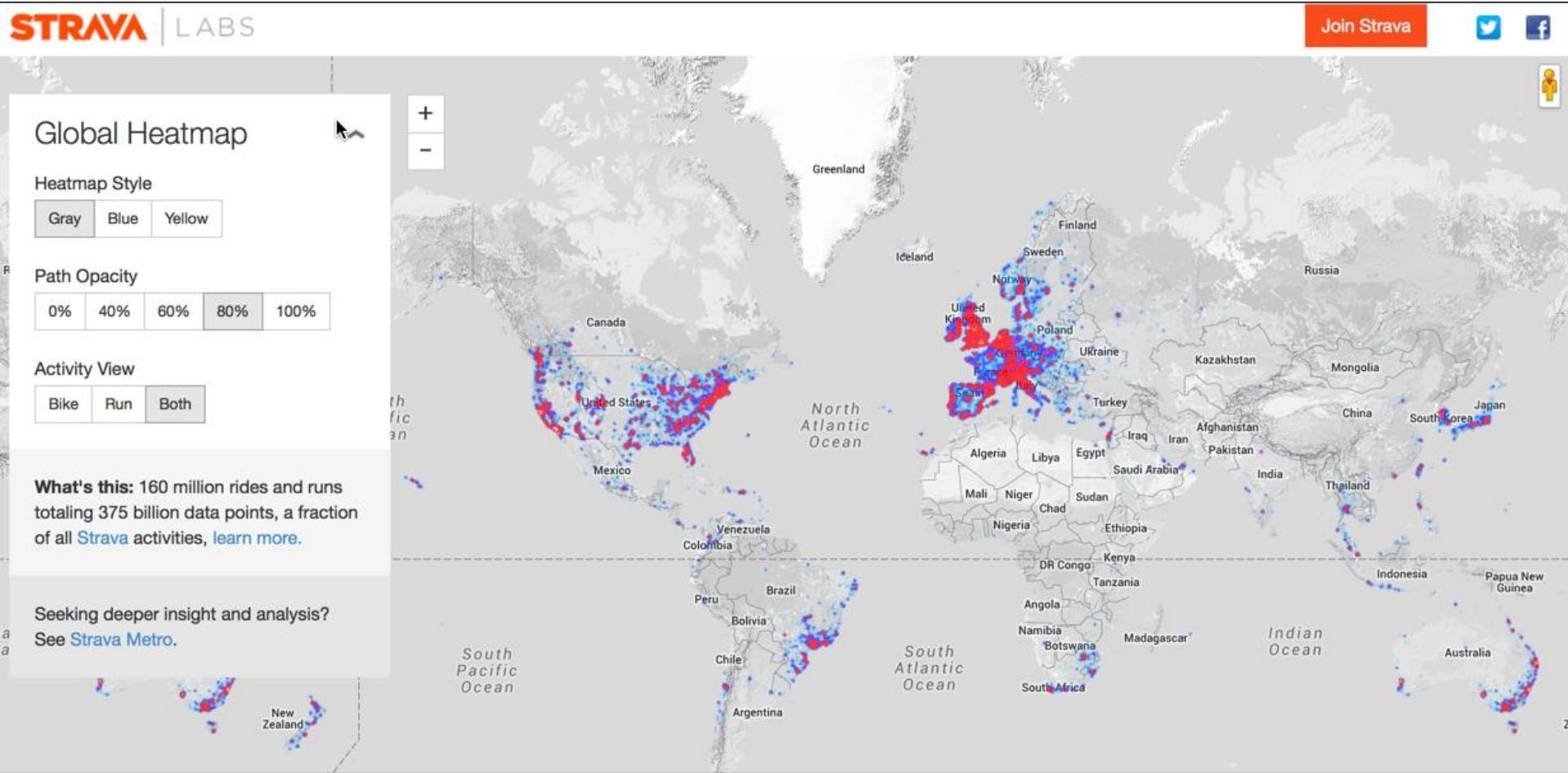


GPS Traces for 200 Days

# Lifelogging Visualizations APPSA



# A view of 160 million rides and runs



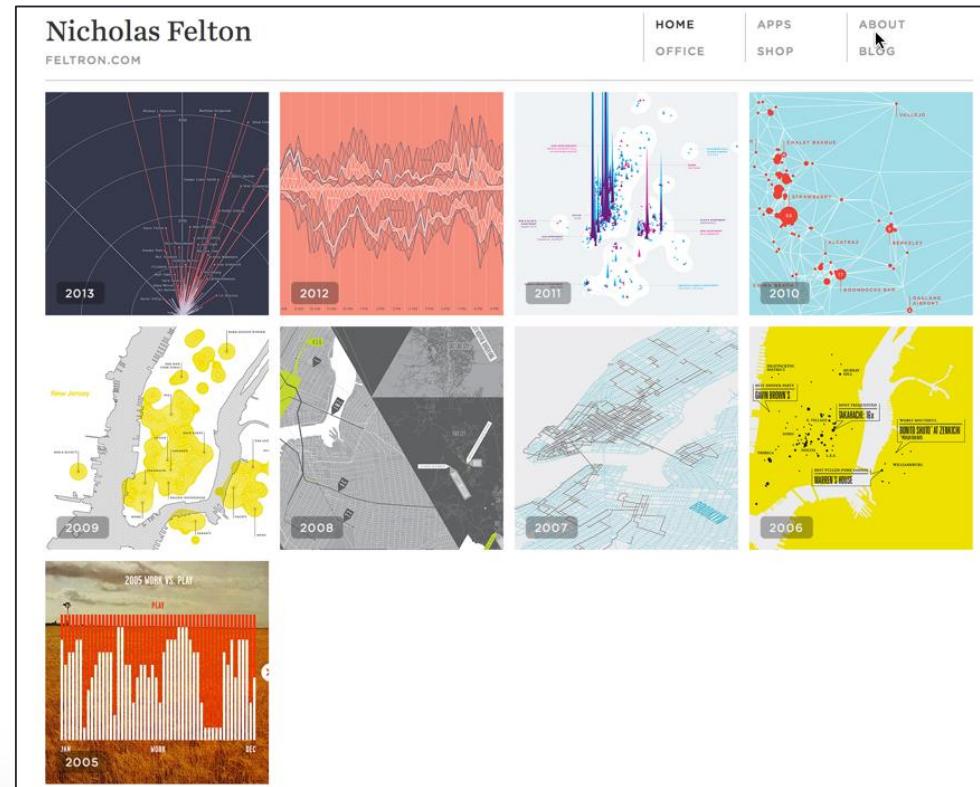
# Lifelogging Visualization

## *Infographing a year of activities*



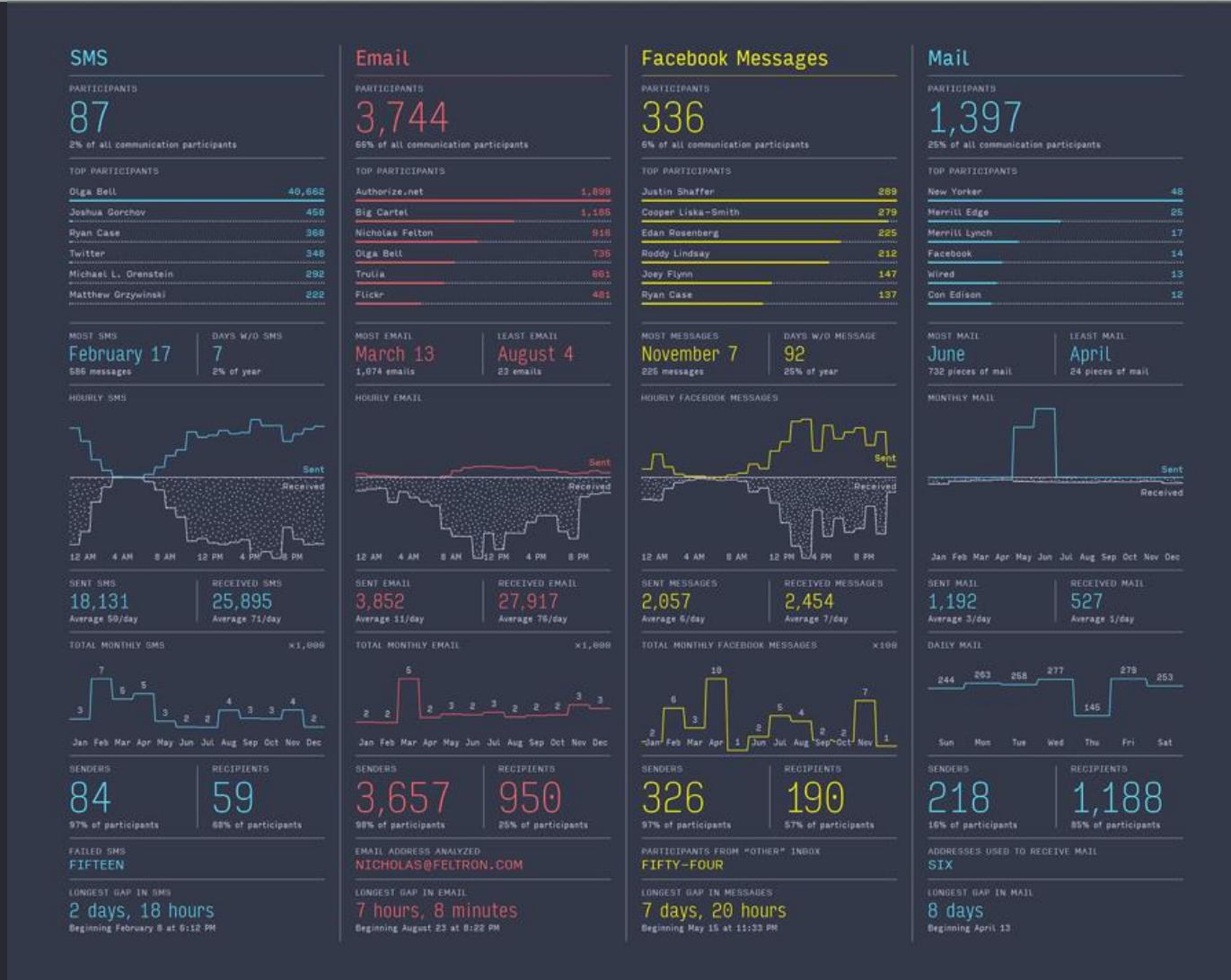
### Biography

Nicholas Felton spends much of his time thinking about data, charts and our daily routines. He is the author of many [Personal Annual Reports](#) that weave numerous measurements into a tapestry of graphs, maps and statistics reflecting the year's activities. He was one of the lead designers of Facebook's timeline and the co-founder of [Daytum.com](#). His most recent product is [Reporter](#), an iPhone app designed to record and visualize subtle aspects of our lives. His work is a part of the permanent collection at MoMA. He has also been profiled by the Wall Street Journal, Wired and Good Magazine and recognized as one of the 50 most influential designers in America by Fast Company.



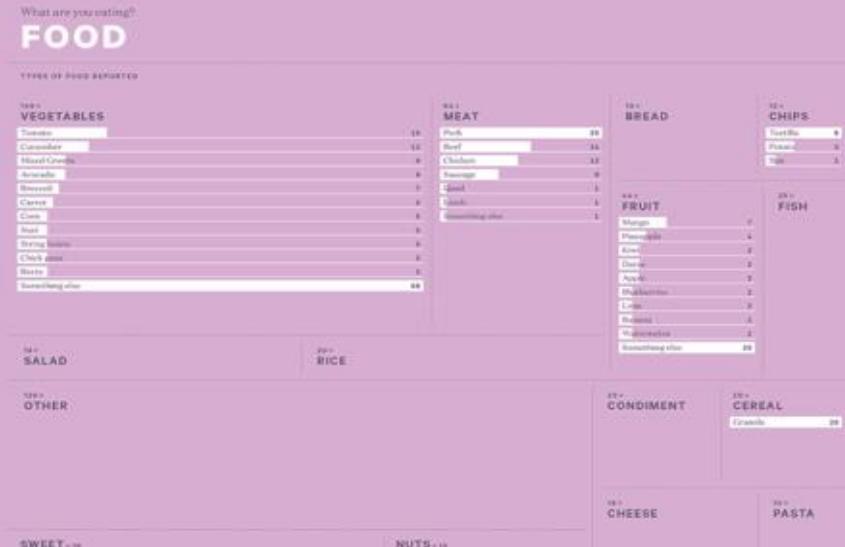
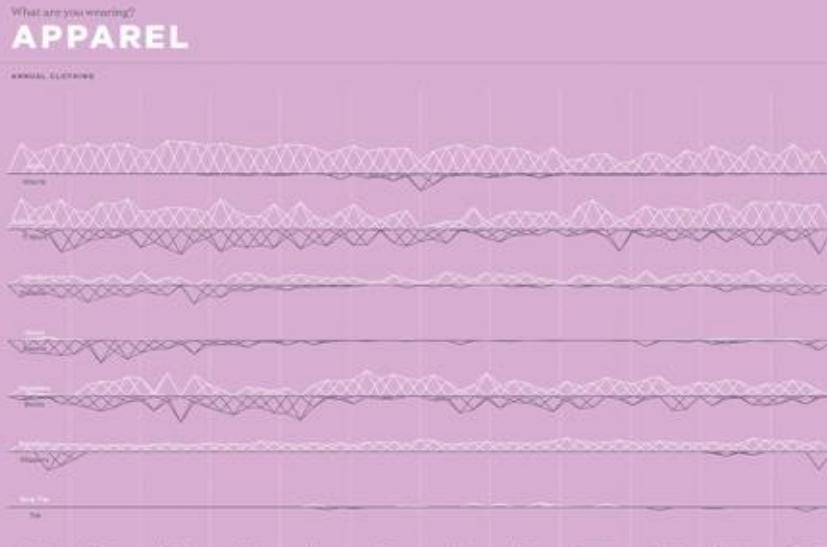
# Lifelogging Visualization

## Felton's 2013 Annual Report



# Lifelogging Visualization

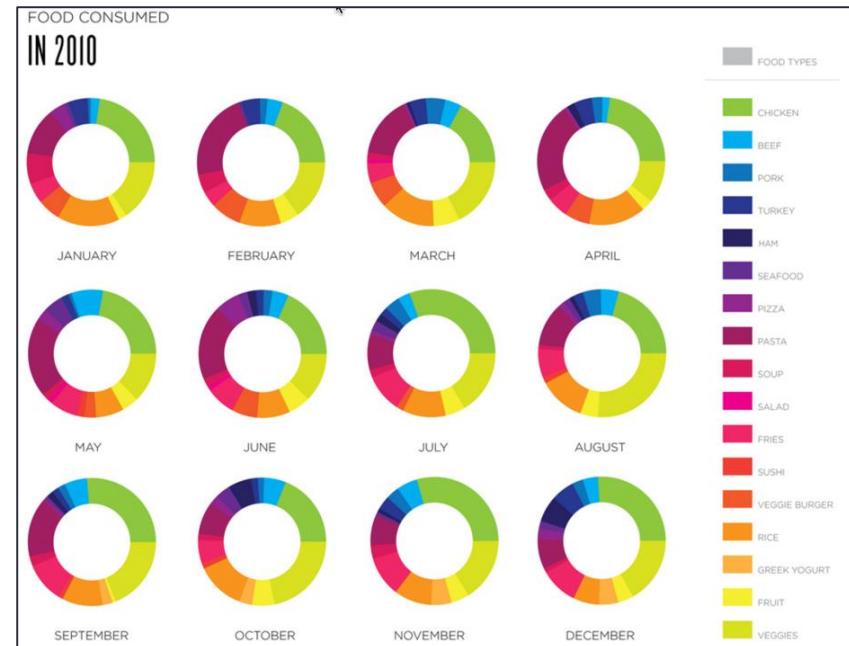
## *Felton's 2013 Annual Report*



# Lauren Manning

## *Food consumed in 2010*

Date	Meal	Food Type
1-Jan	Breakfast	Yogurt
1-Jan	Breakfast	Banana
1-Jan	Lunch	Salad
1-Jan	Lunch	Chicken
1-Jan	Dinner	Shrimp
1-Jan	Dinner	Pasta
1-Jan	Dinner	Zucchini
...		
31-Dec	Breakfast	Yogurt
31-Dec	Breakfast	Strawberries
31-Dec	Lunch	Turkey Sandwich
31-Dec	Lunch	Tomato Soup
31-Dec	Dinner	Miso Soup
31-Dec	Dinner	Mixed Salad
31-Dec	Dinner	Sashimi



# Lauren Manning

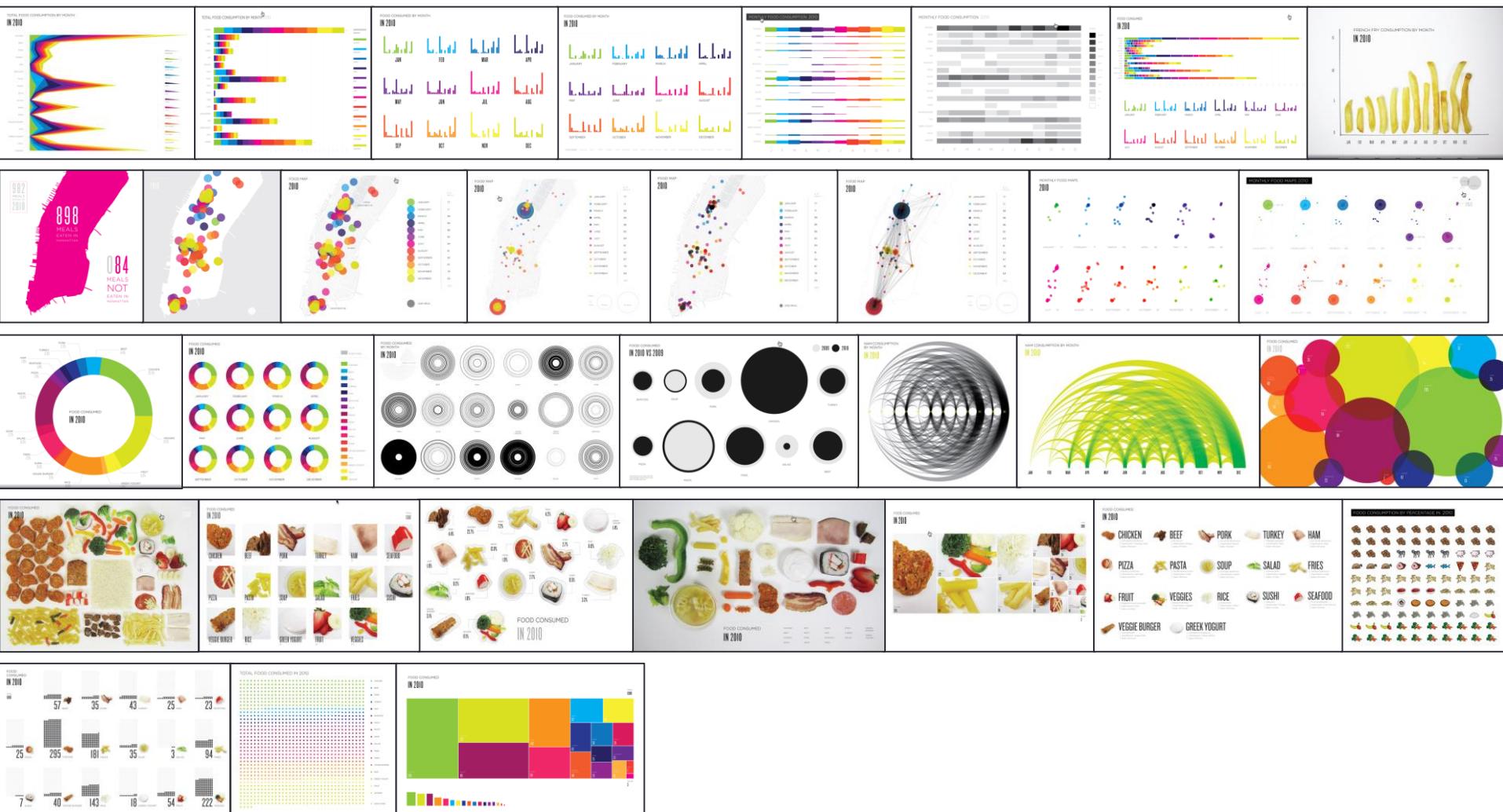
## *40 ways to love your liver (or chicken, or ham...)*



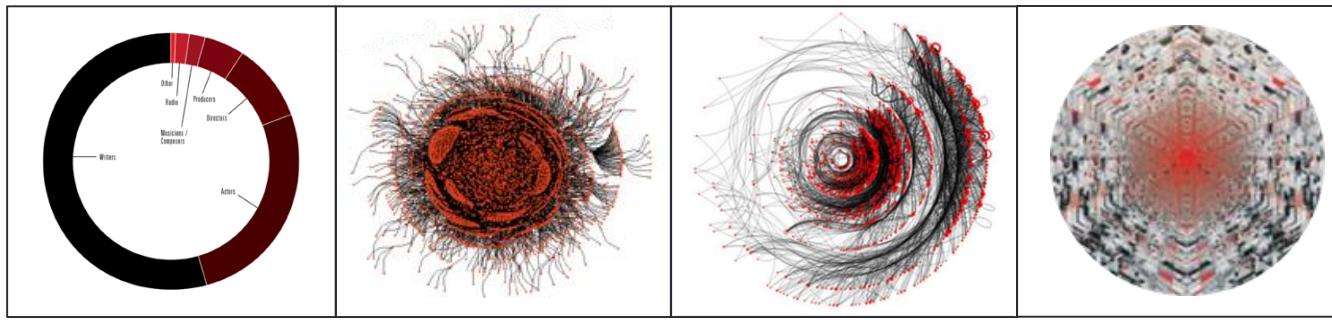
“It’s like comparing apples to oranges.” ... For the designer, it’s easy to find good visualizations and bad ones, but how to apply the successful elements of particular designs to one’s own data set starts to get a little more complicated

# Lauren Manning

## *Visualizing Yearly Food Consumption*



# Data Visualization



Statistical  
Graphics

Data  
Visualization

Data  
Art

Geometric  
Patterns

*“You can think of visualization as a continuous spectrum that stretches from statistical graphics to data art”*

*Yau, Data Points (2013)*

# VISUALIZATION FRAMEWORK

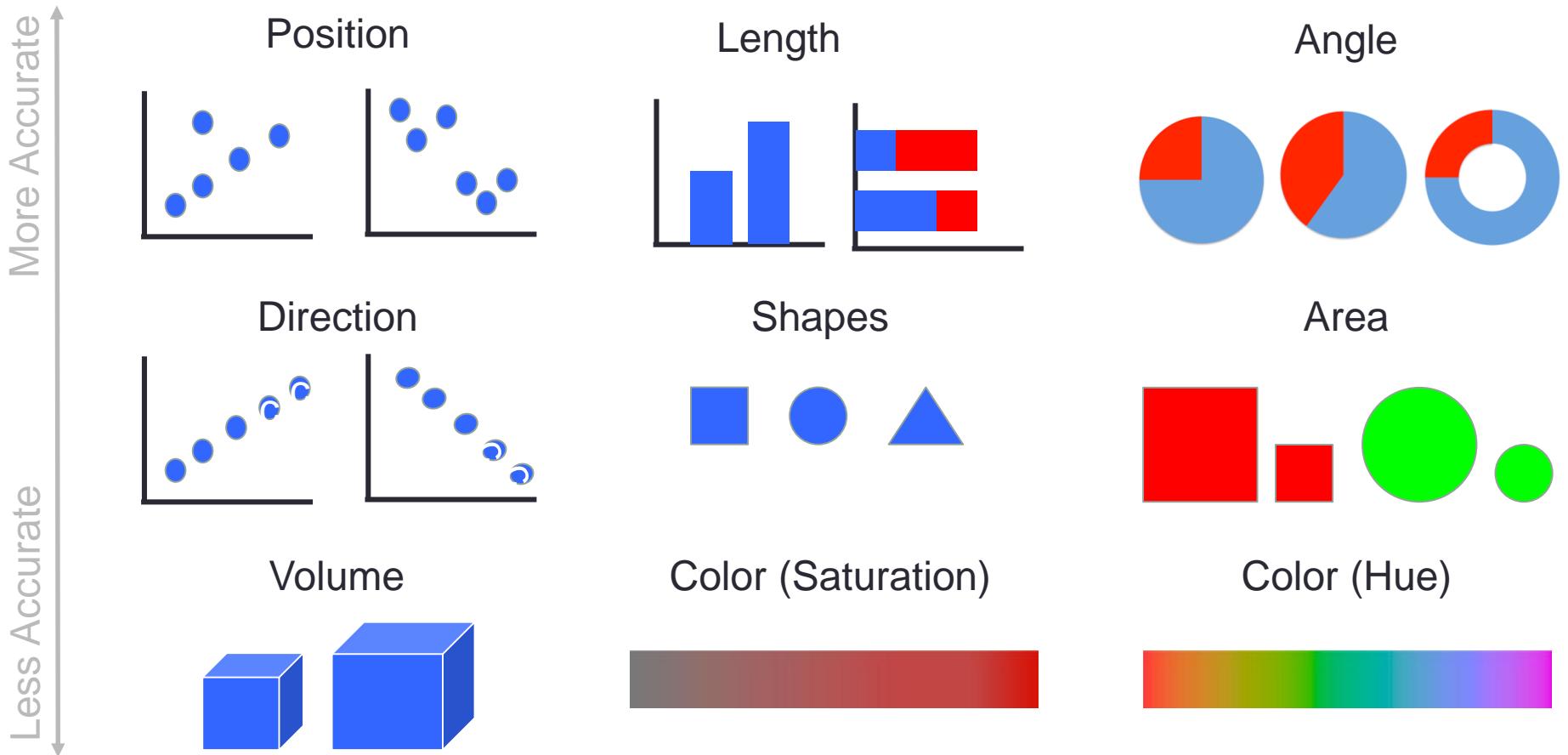
---

# Components of a data visualization

<h2>Working parts</h2> <p>Several pieces work together to make a graph. Sometimes these are explicitly shown in the visualization and other times they form a visual in the background. They all depend on the data.</p>	<p>Title of this Graph A description of the data or something worth highlighting to set the stage.</p> <table border="1"><thead><tr><th>Month</th><th>Value</th></tr></thead><tbody><tr><td>Jan. 2012</td><td>100 units</td></tr><tr><td>Feb. 2012</td><td>45 units</td></tr><tr><td>Mar. 2012</td><td>20 units</td></tr><tr><td>Apr. 2012</td><td>10 units</td></tr><tr><td>May 2012</td><td>5 units</td></tr><tr><td>June 2012</td><td>2 units</td></tr><tr><td>July 2012</td><td>1 unit</td></tr></tbody></table> <p>Source: Somewhere reputable</p>	Month	Value	Jan. 2012	100 units	Feb. 2012	45 units	Mar. 2012	20 units	Apr. 2012	10 units	May 2012	5 units	June 2012	2 units	July 2012	1 unit	<h2>Visual Cues</h2> <p>Visualization involves encoding data with shapes, colors, and sizes. Which cues you choose depends on your data and your goals.</p>	<p>Title of this Graph A description of the data or something worth highlighting to set the stage.</p> <table border="1"><thead><tr><th>Category</th><th>Proportion</th></tr></thead><tbody><tr><td>Light Blue</td><td>~45%</td></tr><tr><td>Orange</td><td>~25%</td></tr><tr><td>Green</td><td>~15%</td></tr><tr><td>Red</td><td>~15%</td></tr></tbody></table>	Category	Proportion	Light Blue	~45%	Orange	~25%	Green	~15%	Red	~15%	<h2>Coordinate System</h2> <p>You map data differently with a scatterplot than you do with a pie chart. It's x- and y-coordinates in one and angles with the other; it's cartesian versus polar.</p>																																						
Month	Value																																																																			
Jan. 2012	100 units																																																																			
Feb. 2012	45 units																																																																			
Mar. 2012	20 units																																																																			
Apr. 2012	10 units																																																																			
May 2012	5 units																																																																			
June 2012	2 units																																																																			
July 2012	1 unit																																																																			
Category	Proportion																																																																			
Light Blue	~45%																																																																			
Orange	~25%																																																																			
Green	~15%																																																																			
Red	~15%																																																																			
<p><b>Title of this Graph</b> A description of the data or something worth highlighting to set the stage.</p> <table border="1"><thead><tr><th>Month</th><th>Value</th></tr></thead><tbody><tr><td>Jan. 2012</td><td>100 units</td></tr><tr><td>Feb. 2012</td><td>45 units</td></tr><tr><td>Mar. 2012</td><td>20 units</td></tr><tr><td>Apr. 2012</td><td>10 units</td></tr><tr><td>May 2012</td><td>5 units</td></tr><tr><td>June 2012</td><td>2 units</td></tr><tr><td>July 2012</td><td>1 unit</td></tr></tbody></table> <p>Source: Somewhere reputable</p>	Month	Value	Jan. 2012	100 units	Feb. 2012	45 units	Mar. 2012	20 units	Apr. 2012	10 units	May 2012	5 units	June 2012	2 units	July 2012	1 unit	<p><b>Title of this Graph</b> A description of the data or something worth highlighting to set the stage.</p> <table border="1"><thead><tr><th>Month</th><th>Value</th></tr></thead><tbody><tr><td>Jan. 2012</td><td>100 units</td></tr><tr><td>Feb. 2012</td><td>45 units</td></tr><tr><td>Mar. 2012</td><td>20 units</td></tr><tr><td>Apr. 2012</td><td>10 units</td></tr><tr><td>May 2012</td><td>5 units</td></tr><tr><td>June 2012</td><td>2 units</td></tr><tr><td>July 2012</td><td>1 unit</td></tr></tbody></table>	Month	Value	Jan. 2012	100 units	Feb. 2012	45 units	Mar. 2012	20 units	Apr. 2012	10 units	May 2012	5 units	June 2012	2 units	July 2012	1 unit	<h2>Scale</h2> <p>Increments that make sense can increase readability, as well as shift focus.</p> <table border="1"><thead><tr><th>Month</th><th>Value</th></tr></thead><tbody><tr><td>Jan. 2012</td><td>100 units</td></tr><tr><td>Feb. 2012</td><td>45 units</td></tr><tr><td>Mar. 2012</td><td>20 units</td></tr><tr><td>Apr. 2012</td><td>10 units</td></tr><tr><td>May 2012</td><td>5 units</td></tr><tr><td>June 2012</td><td>2 units</td></tr><tr><td>July 2012</td><td>1 unit</td></tr></tbody></table>	Month	Value	Jan. 2012	100 units	Feb. 2012	45 units	Mar. 2012	20 units	Apr. 2012	10 units	May 2012	5 units	June 2012	2 units	July 2012	1 unit	<p><b>Title of this Graph</b> A description of the data or something worth highlighting to set the stage.</p> <table border="1"><thead><tr><th>Month</th><th>Value</th></tr></thead><tbody><tr><td>Jan. 2012</td><td>100 units</td></tr><tr><td>Feb. 2012</td><td>45 units</td></tr><tr><td>Mar. 2012</td><td>20 units</td></tr><tr><td>Apr. 2012</td><td>10 units</td></tr><tr><td>May 2012</td><td>5 units</td></tr><tr><td>June 2012</td><td>2 units</td></tr><tr><td>July 2012</td><td>1 unit</td></tr></tbody></table>	Month	Value	Jan. 2012	100 units	Feb. 2012	45 units	Mar. 2012	20 units	Apr. 2012	10 units	May 2012	5 units	June 2012	2 units	July 2012	1 unit	<h2>Context</h2> <p>If your audience is unfamiliar with the data, it's your job to clarify what values represent and explain how people should read your visualization.</p>
Month	Value																																																																			
Jan. 2012	100 units																																																																			
Feb. 2012	45 units																																																																			
Mar. 2012	20 units																																																																			
Apr. 2012	10 units																																																																			
May 2012	5 units																																																																			
June 2012	2 units																																																																			
July 2012	1 unit																																																																			
Month	Value																																																																			
Jan. 2012	100 units																																																																			
Feb. 2012	45 units																																																																			
Mar. 2012	20 units																																																																			
Apr. 2012	10 units																																																																			
May 2012	5 units																																																																			
June 2012	2 units																																																																			
July 2012	1 unit																																																																			
Month	Value																																																																			
Jan. 2012	100 units																																																																			
Feb. 2012	45 units																																																																			
Mar. 2012	20 units																																																																			
Apr. 2012	10 units																																																																			
May 2012	5 units																																																																			
June 2012	2 units																																																																			
July 2012	1 unit																																																																			
Month	Value																																																																			
Jan. 2012	100 units																																																																			
Feb. 2012	45 units																																																																			
Mar. 2012	20 units																																																																			
Apr. 2012	10 units																																																																			
May 2012	5 units																																																																			
June 2012	2 units																																																																			
July 2012	1 unit																																																																			

*Each visualization, regardless of where it is on the spectrum, is built on data and these five components.*

# Visual cues



# How many \_'s are there?

7	9	2	5	8	5	0	4	1	5	9	0	1	0	0
2	2	3	6	1	0	6	0	6	7	0	7	6	9	5
9	4	8	4	6	1	4	2	9	2	8	6	4	6	8
0	9	1	3	1	1	7	6	7	5	2	5	9	4	1

6's

7	9	2	5	8	5	0	4	1	5	9	0	1	0	0
2	2	3	<b>6</b>	1	0	<b>6</b>	0	<b>6</b>	7	0	7	<b>6</b>	9	5
9	4	8	4	<b>6</b>	1	4	2	9	2	8	<b>6</b>	4	<b>6</b>	8
0	9	1	3	1	1	7	<b>6</b>	7	5	2	5	9	4	1

2's

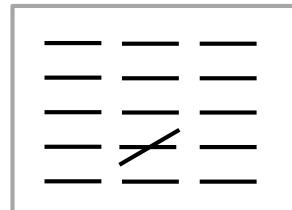
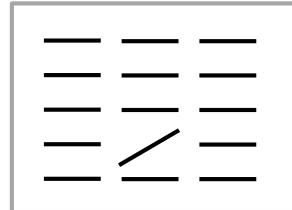
7	9	<b>2</b>	5	8	5	0	4	1	5	9	0	1	0	0
<b>2</b>	<b>2</b>	3	6	1	0	6	0	6	7	0	7	6	9	5
9	4	8	4	6	1	4	<b>2</b>	9	<b>2</b>	8	6	4	6	8
0	9	1	3	1	1	7	6	7	5	<b>2</b>	5	9	4	1

1's

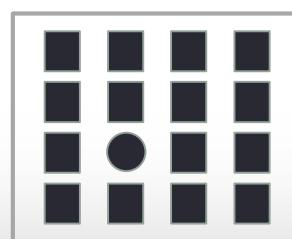
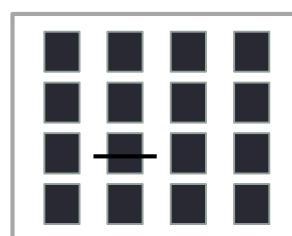
7	9	2	5	8	5	0	4	<b>1</b>	5	9	0	<b>1</b>	0	0
2	2	3	6	<b>1</b>	0	6	0	6	7	0	7	6	9	5
9	4	8	4	6	<b>1</b>	4	2	9	2	8	6	4	6	8
0	9	1	3	1	1	7	6	7	5	2	5	9	4	<b>1</b>

# Detection and Recognition

Line Orientation

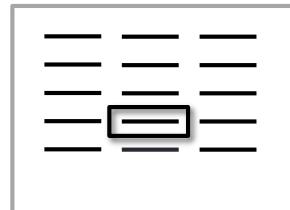
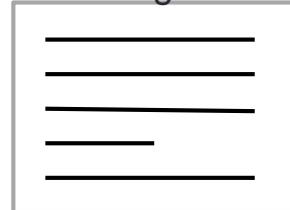


Added Markers

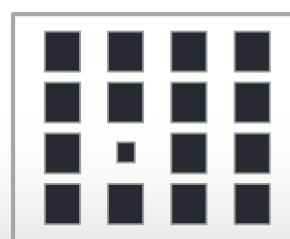
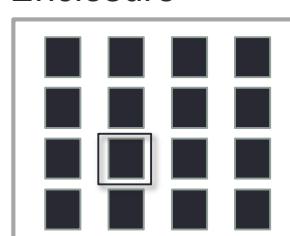


Shape

Line Length

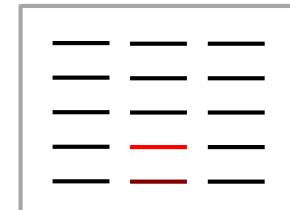
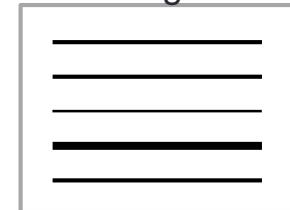


Enclosure

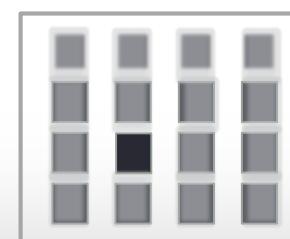
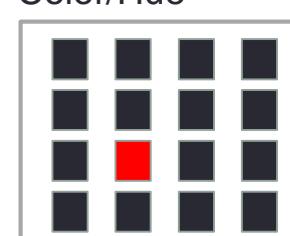


Size

Line Weight

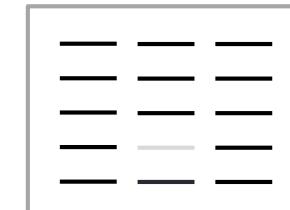
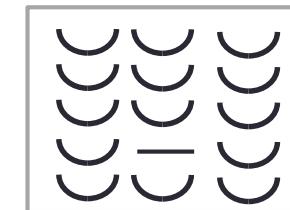


Color/Hue

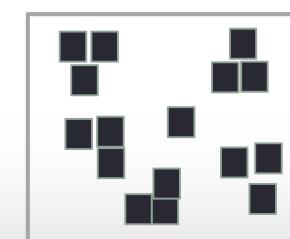
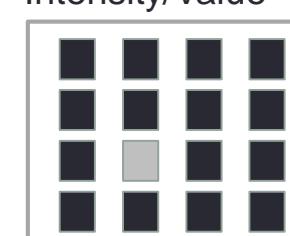


Sharpness

Curvature



Intensity/Value

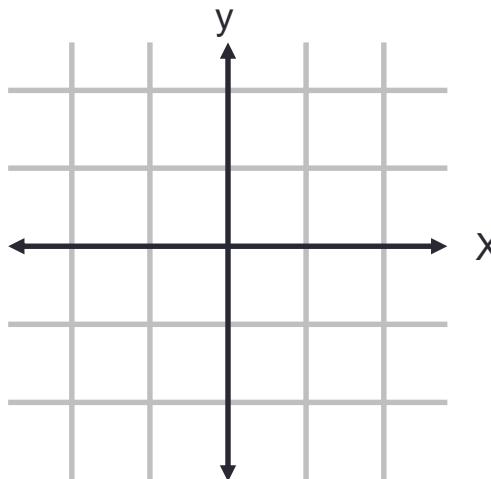


Numerosity

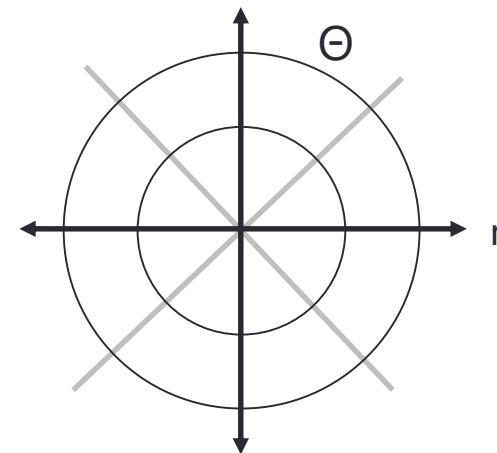
# Coordinate Systems

*Most common*

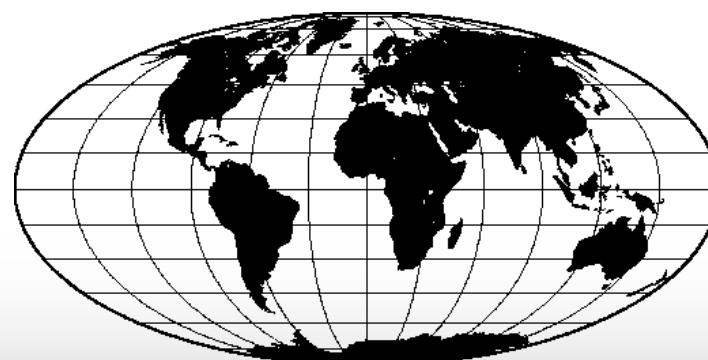
Cartesian



Polar



Geographic  
Projection



# Scale

*Most common*

Linear (even space)



Logarithmic (% change)



Categorical (discrete)



Ordinal (ranked categories)



Percent (of whole)

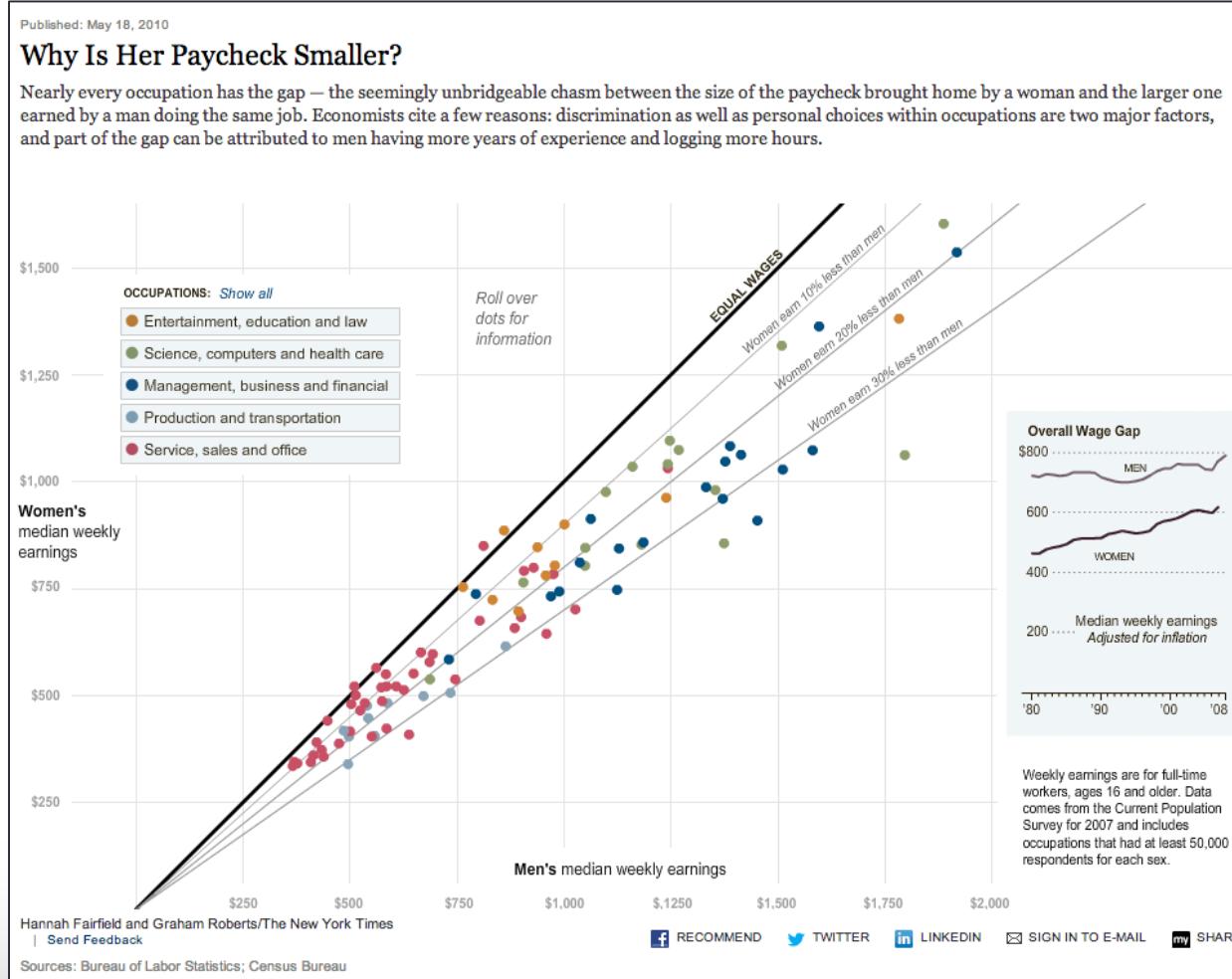


Time (hours, days, mnths, years ...)



# The role of context

*1 word is worth ...*



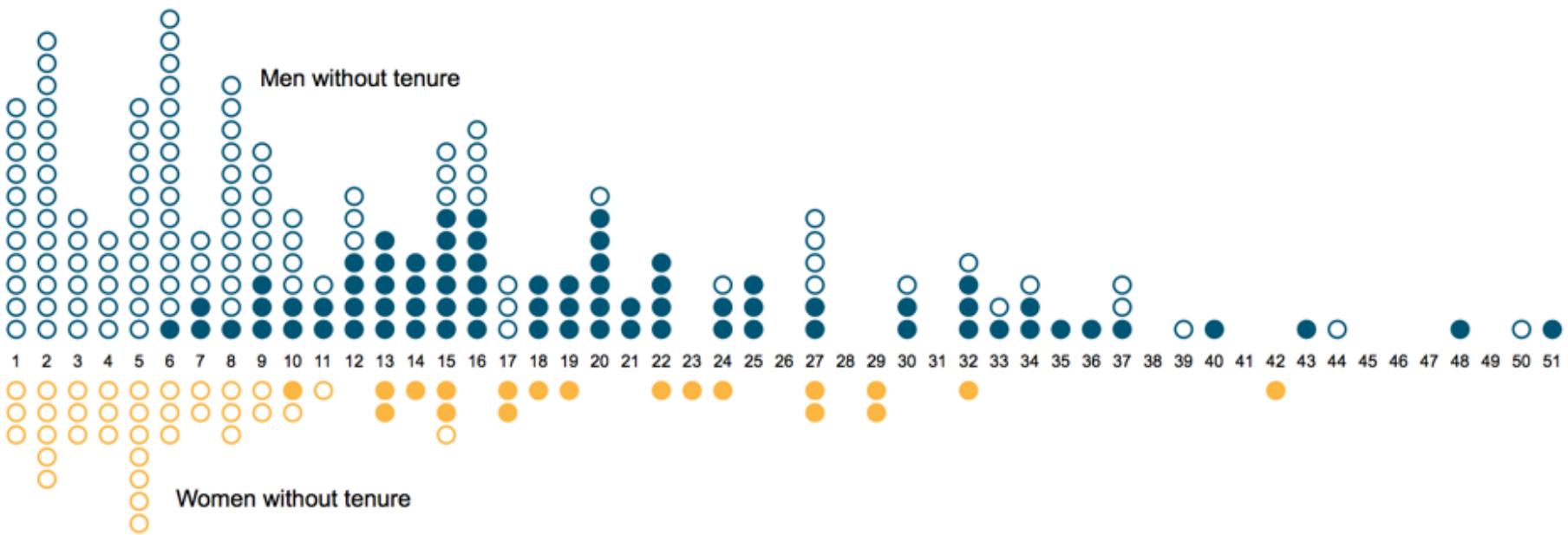
# Context

*1 word is worth ...*

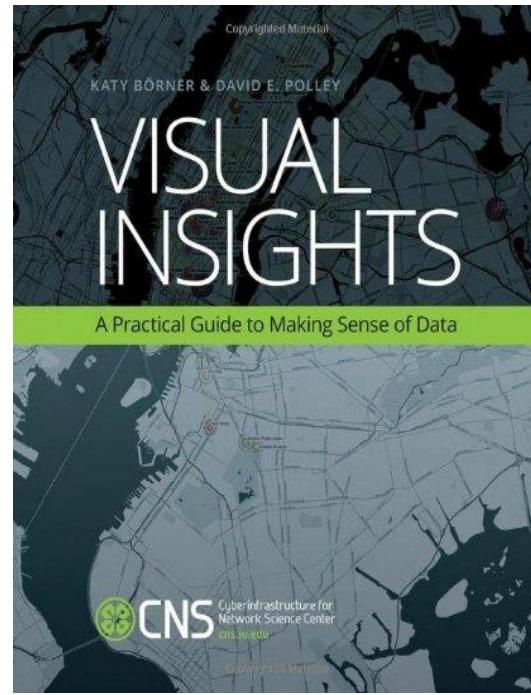
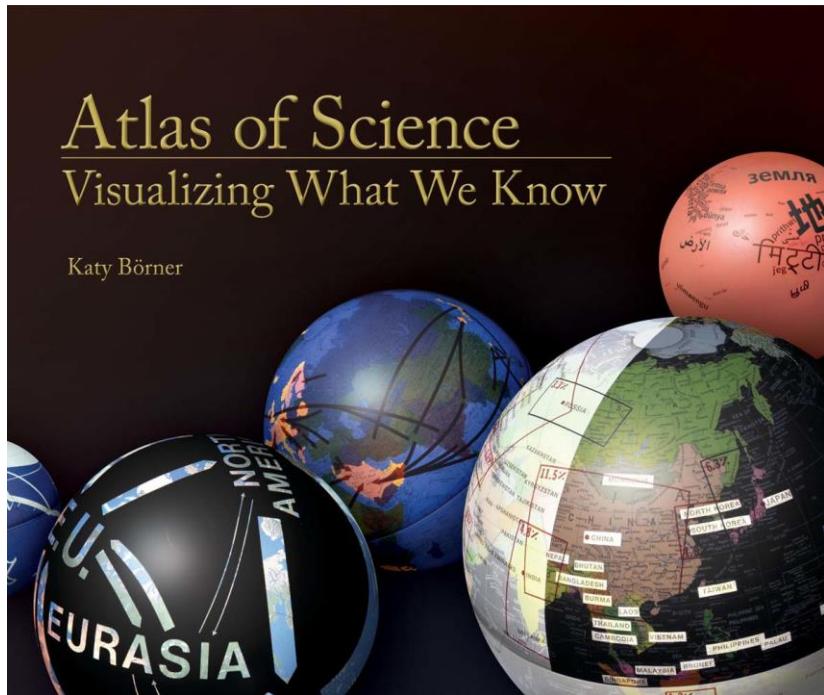
## The Tenure Pipeline at Harvard Business School

1 2 3 4 5 NEXT >

But the pipeline for women is small. Fewer than a third of untenured faculty members are women, making it unlikely that the gender imbalance in tenured faculty will shift in the near future.

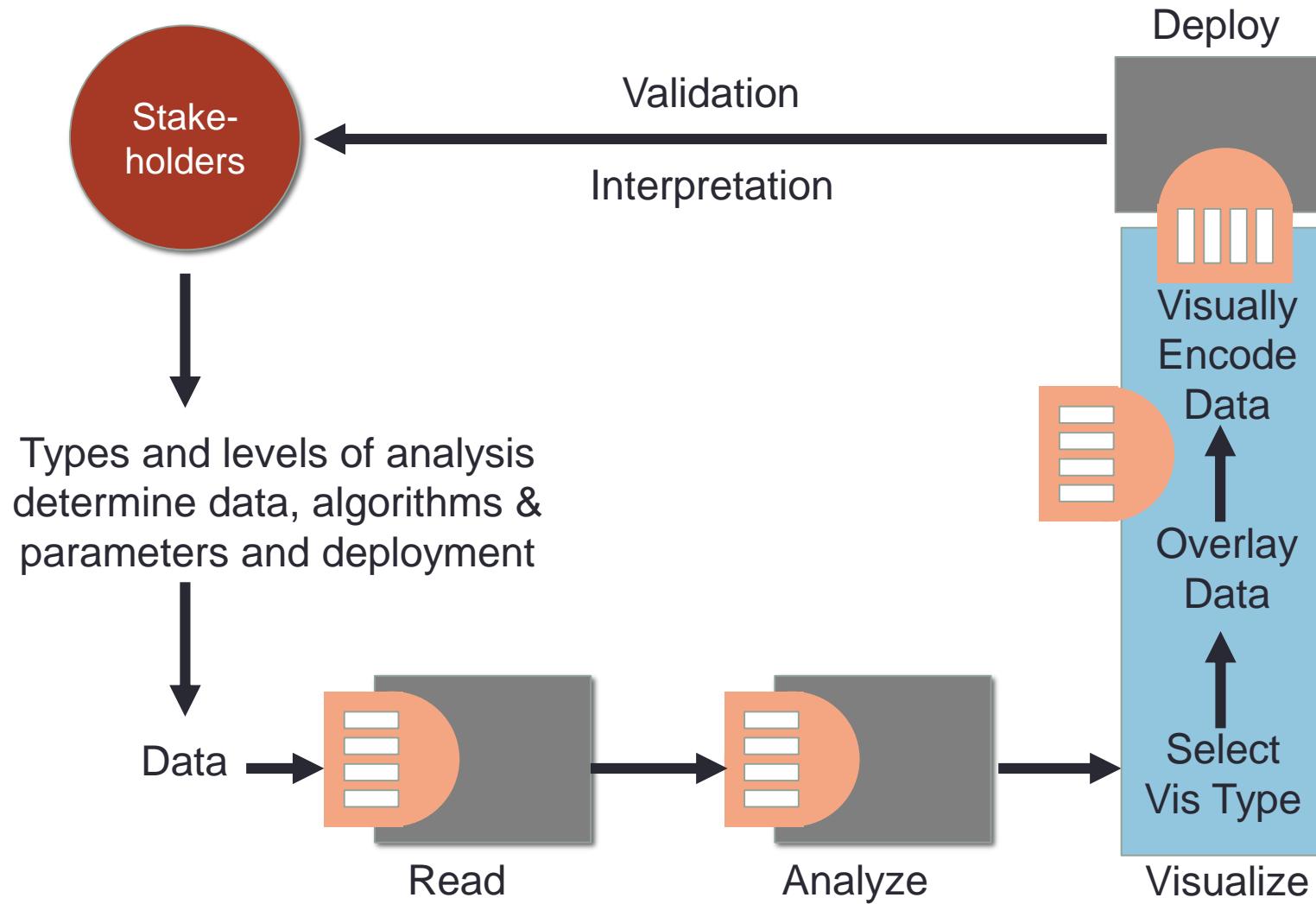


# Visualization Framework & Workflow



[cns.iu.edu/home.html](http://cns.iu.edu/home.html)  
[ella.slis.indiana.edu/~katy/](http://ella.slis.indiana.edu/~katy/)

# Visualization Framework & Workflow



# Analysis: Type vs. Level

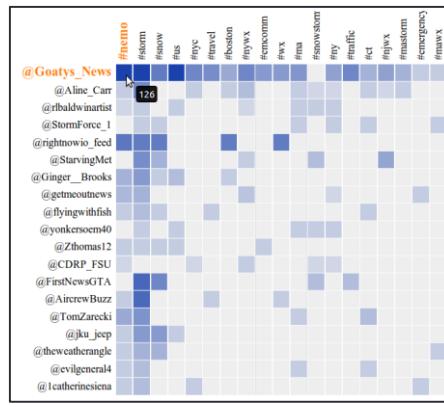
Types of Analysis	Levels of Analysis		
	Micro/Individual 1-100 records	Meso/Local 101-10,000 records	Macro/Global 10,000+ records
Statistical Analysis/Profiling			
Temporal Analysis (When)			
Geospatial Analysis (Where)			
Topical Analysis (What)			
Network Analysis (with Whom)			

# Visualization Types

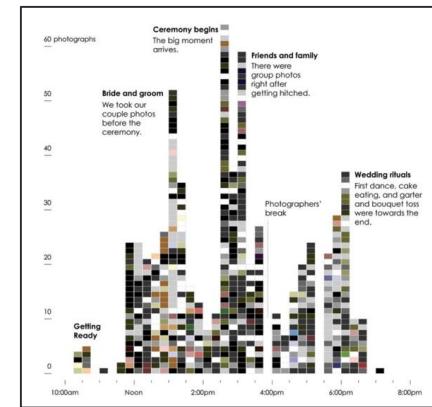
## Charts



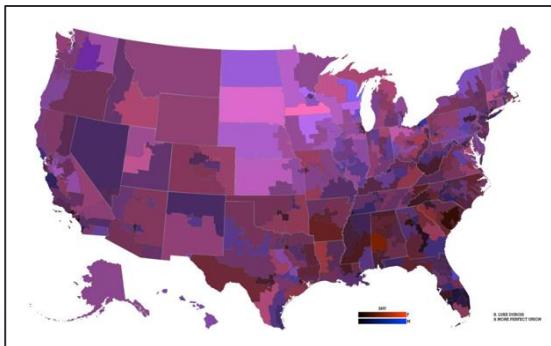
## Tables



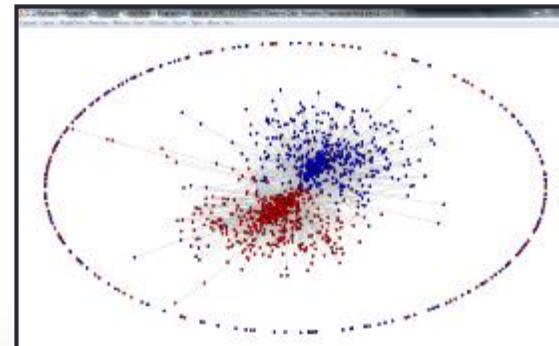
# Graphs



## Geospatial Maps



# Network Graphs

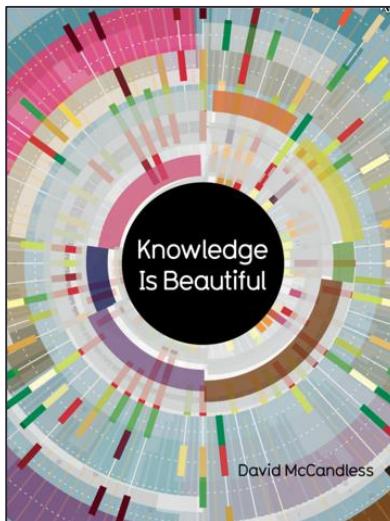
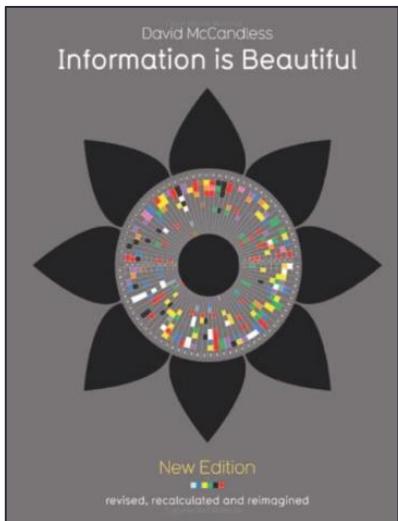


# EXAMPLE: RAP LYRICS

---

# Example

## *Analyzing the lyrics of Rap/Hip Hop songs*



A screenshot of the Informationisbeautiful.net website. The header features the site's name and a tagline 'ideas, issues, knowledge, data – visualized!'. It includes social media links (Facebook, Twitter, LinkedIn) and navigation links for Home, About, Blog, Our Data, Events, Contact, Books, Jobs, and Store. The main content area has a dark background with white text. It starts with a 'Hello' section featuring a small portrait of David McCandless. Below it is a 'data journalist and information designer' section, followed by a 'Our mission' section. On the right side, there are two sidebar boxes: one for 'Our Beautiful Books' (with a link to 'NEW UPDATED 2013 EDITIONS') and another for 'PRINTS, POSTERS AND PDFs' (with a small thumbnail image).

David McCandless – [Informationisbeautiful.net](http://Informationisbeautiful.net)

# Example

## Analyzing the lyrics of Rap/Hip Hop songs

KANTAR  
Information is Beautiful  
Awards 2014

Home About News Awards Challenges Showcase Sponsor

**Global alcohol consumption**  
Here in the UK it's Friday afternoon. And for lots of us, that means it's nearly "beer o'clock". Fed by data from the World Health Organisation, a new online application by Zenoid shows how much and what types of alcohol are consumed by the world's different countries on a weekly basis. Check it out here. Cheers!

→

4 MONTHS AGO

**Travel through topographical time**  
A nifty web app by The Swiss Federal Office of Topography (swisstopo) shows Swiss topographical changes over the years from 1844 up ... →

4 MONTHS AGO

**The biggest vocabularies in Hip Hop**  
Shakespeare utilised no fewer than 28,829 different words across his entire works. The brainy git. Intrigued, digital strategist Matt Daniels used token analysis to determine ... →

4 MONTHS AGO

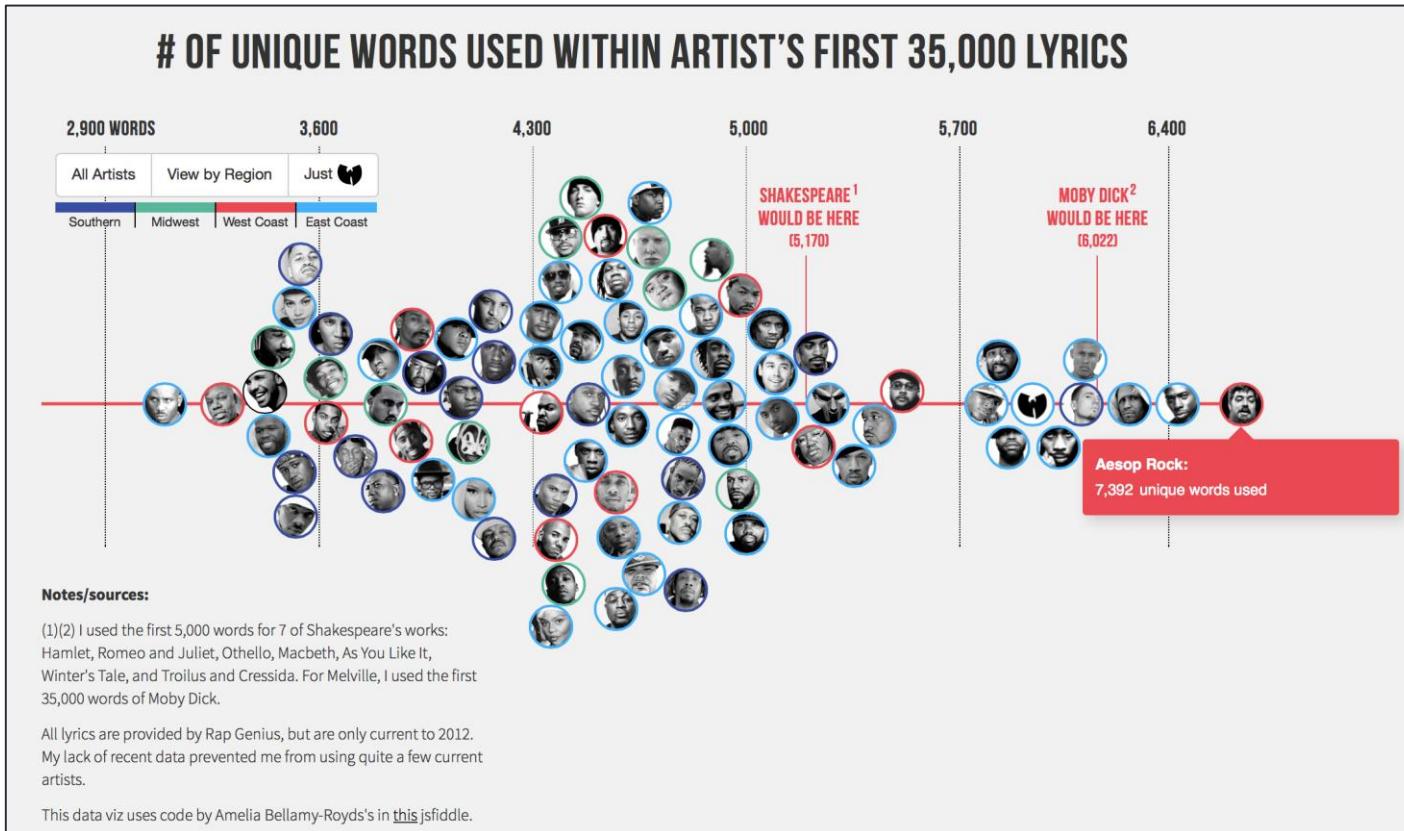
**Nate Silver's search for America's Best Burrito**  
As well as being the editor of data-driven journalism website, FiveThirtyEight, Nate Silver is a burrito fan. Big time. Over several years he's not only been eating and enthus... →

Open 24 hrs.

Informationisbeautifulawards.net

# Example: Lyrics of Rap Songs

## *Biggest Vocabularies and Unique Words*

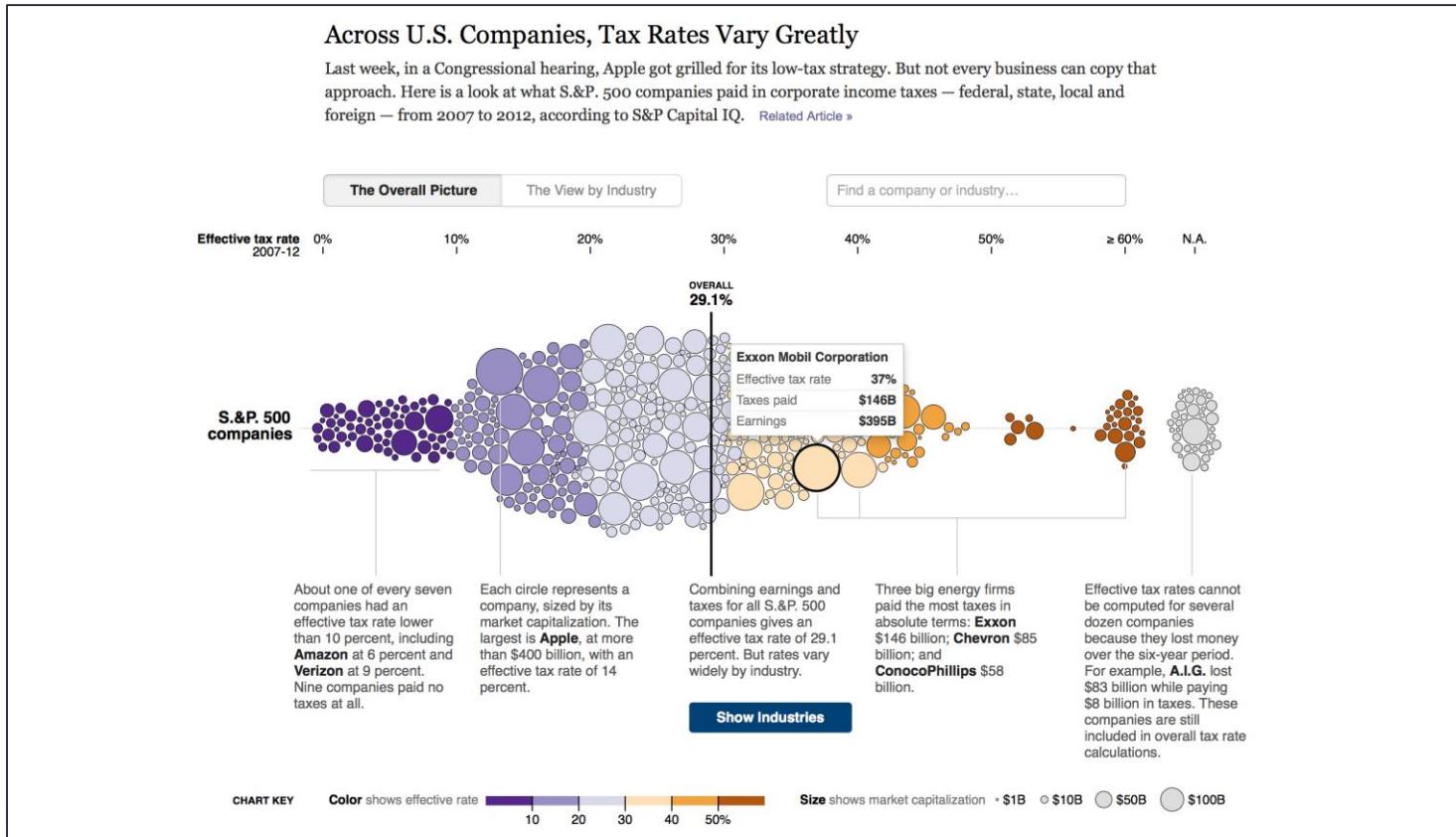


Research by: Matt Daniels, [mdaniels.com](http://mdaniels.com)

Visualization by: Amelia Bellamy-Royds, [fiddle.jshell.net/6cW9u/8/](http://fiddle.jshell.net/6cW9u/8/)

# Example: Lyrics of Rap Songs

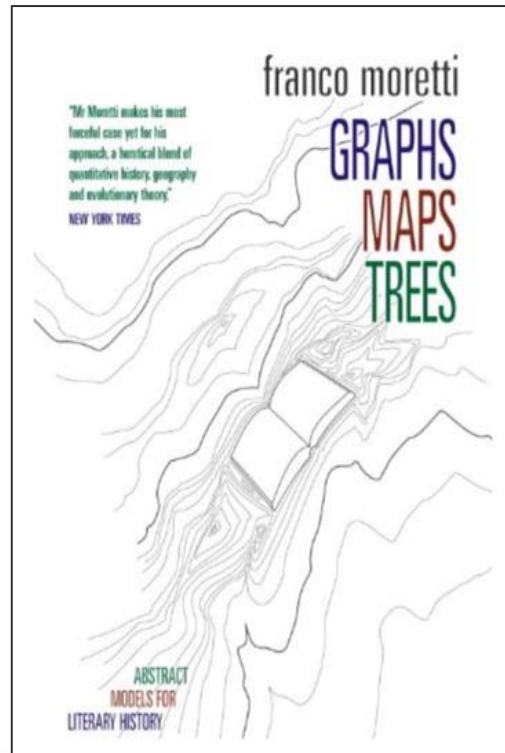
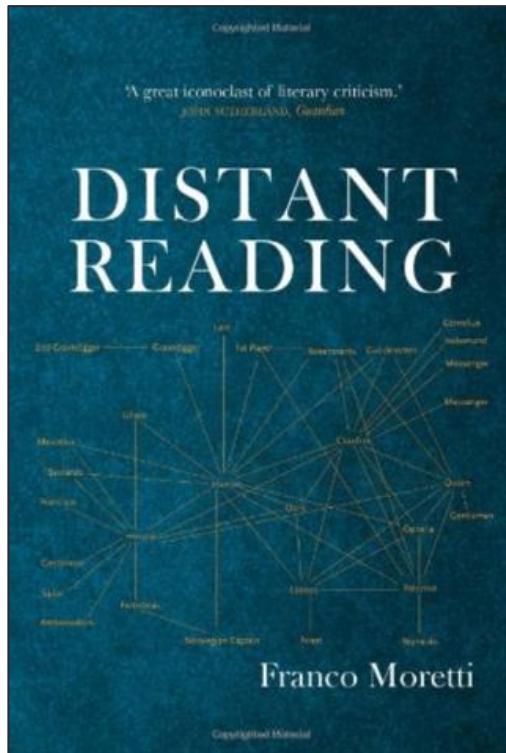
## Visualization inspired by...



Inspired by: [nytimes.com/interactive/2013/05/25/sunday-review/corporate-taxes.html?\\_r=2&](http://nytimes.com/interactive/2013/05/25/sunday-review/corporate-taxes.html?_r=2&)

# Example: Lyrics of Rap Songs

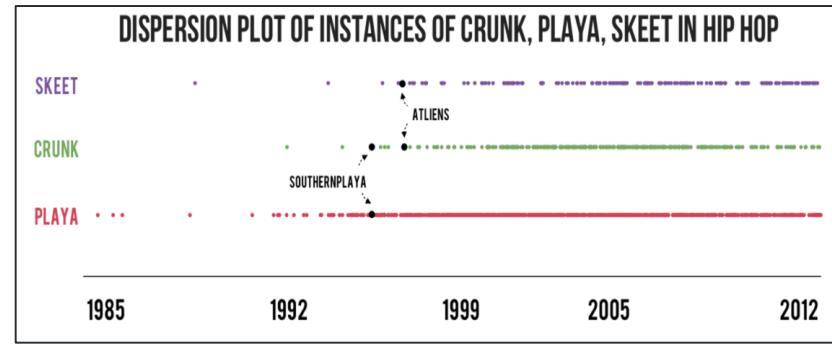
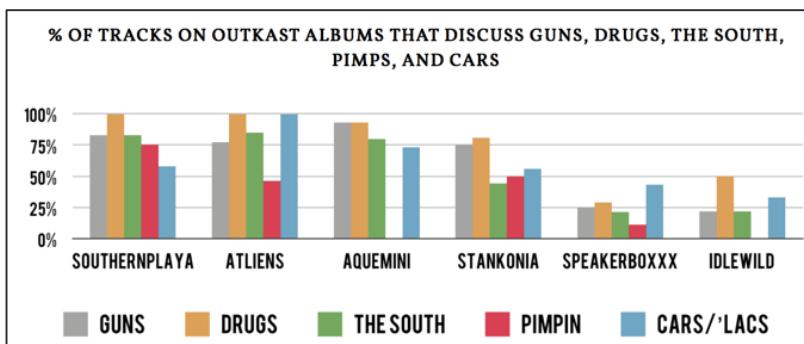
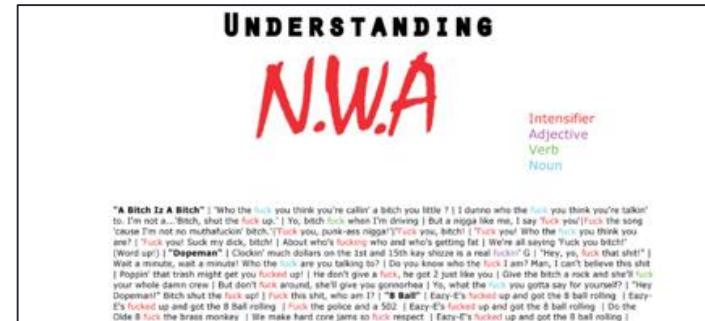
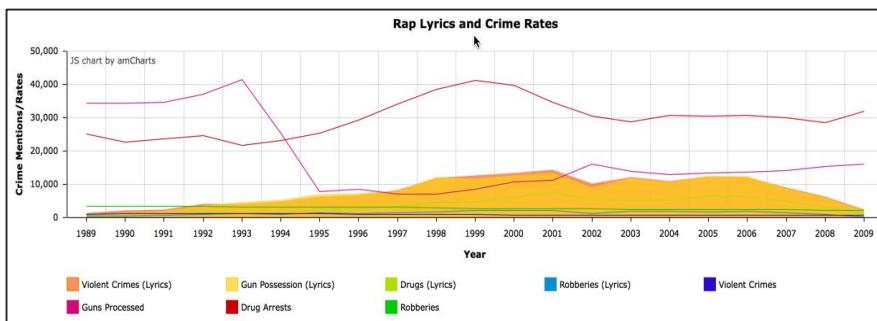
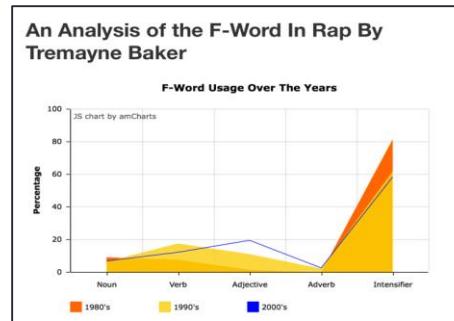
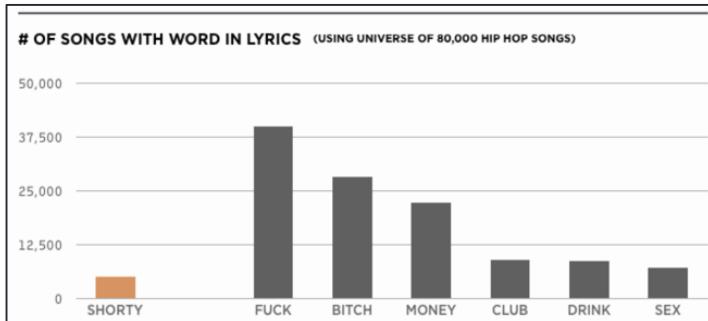
*In the same vein as ...*



Literature scholars should stop reading books and start counting, graphing, and mapping them instead.

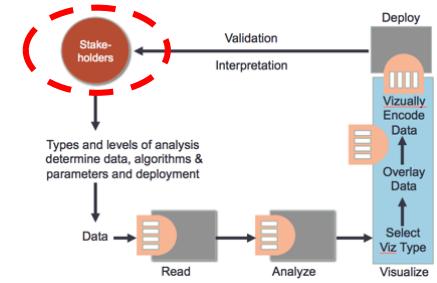
# Example: Lyrics of Rap Songs

*Types of analyses – word or topic frequency*



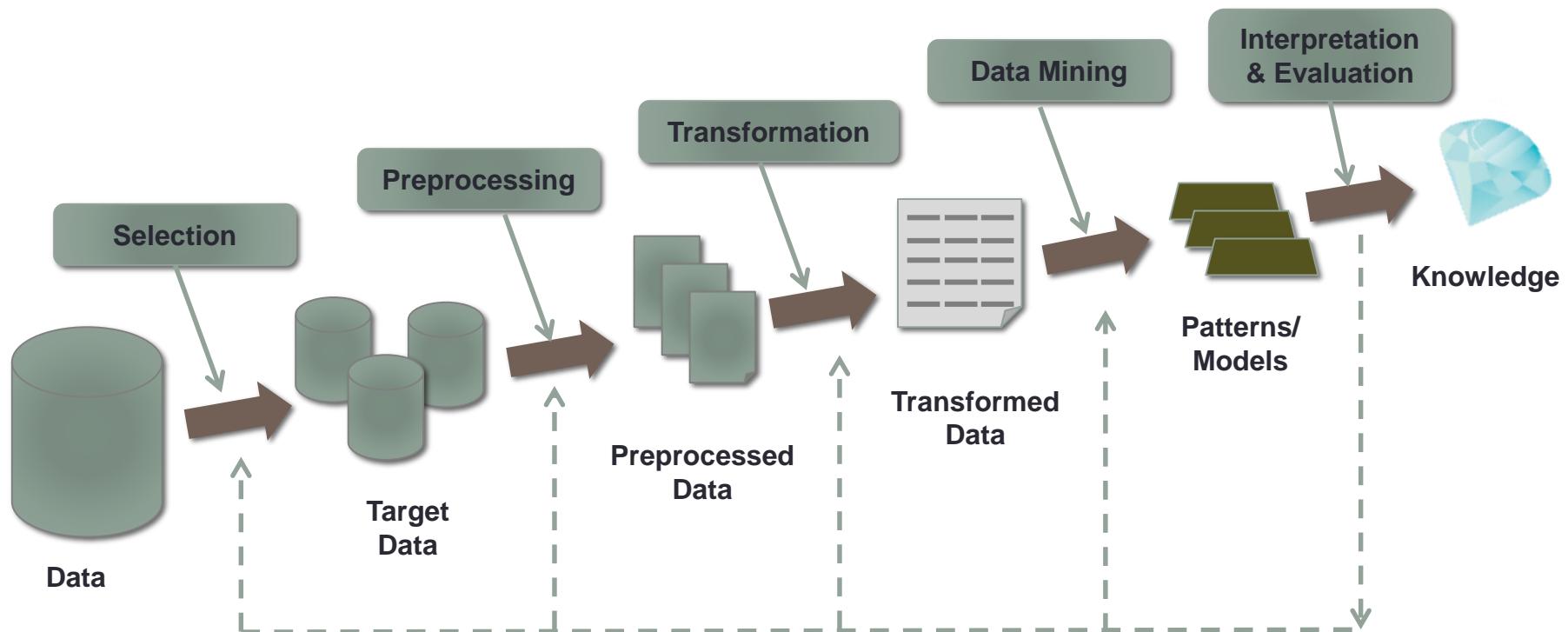
# Example: Lyrics of Rap Songs

*Textual analysis – the interest*



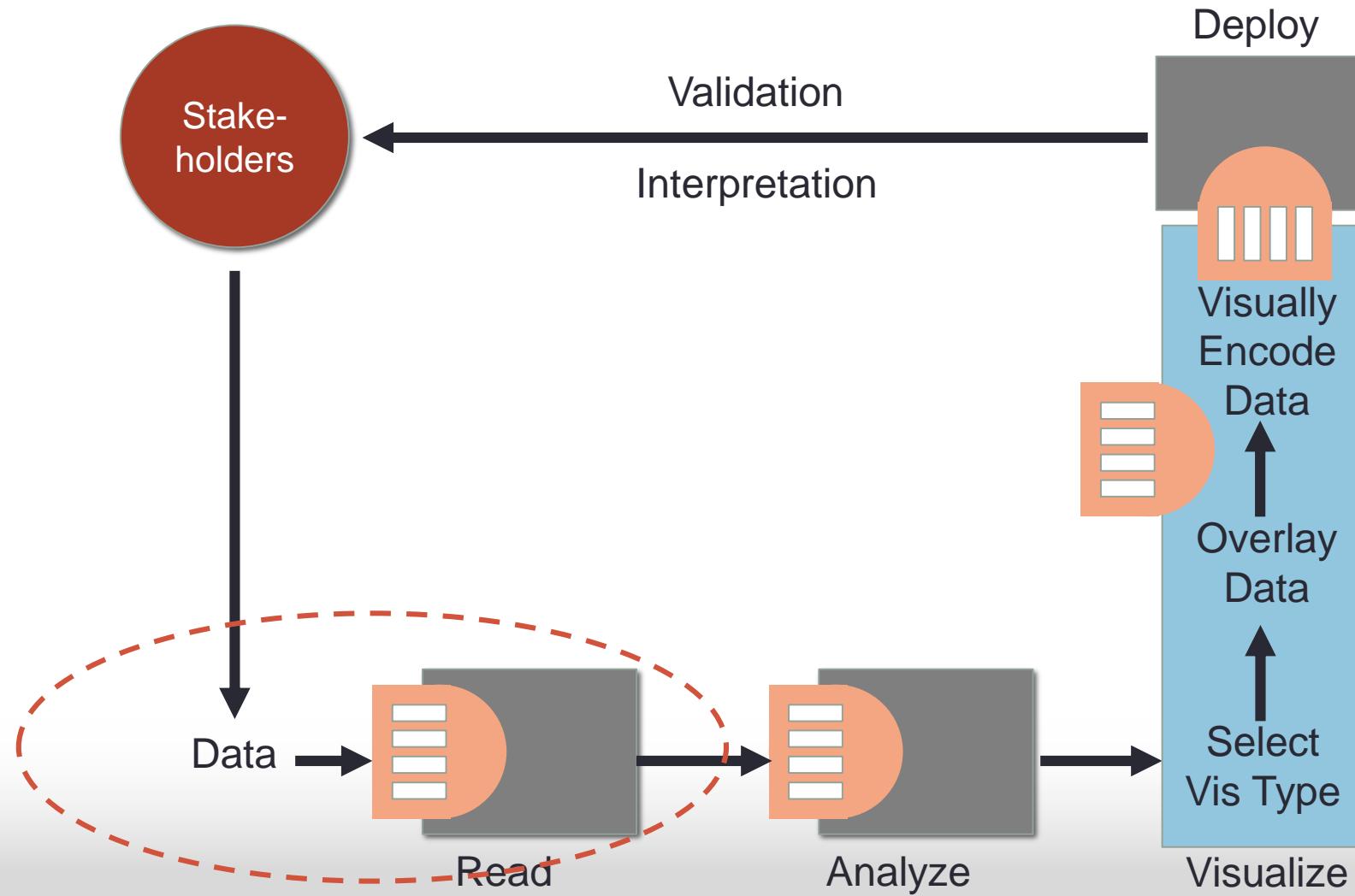
# Example: Lyrics of Rap Songs

## CRISP Data Mining Process



# Example: Lyrics of Rap Songs

*Difficult part is turning text into analyzable data*



# Example: Lyrics of Rap Songs

## *Moving from lyrics to...*

Everybody just rock it, don't stop it  
You gotta rock it, don't stop  
Keep tickin' and tockin', work it all around the clock  
Everybody keep rockin' and clockin' and shockin' and rockin', go house

Da leaders, lookin straight charming in our Giorgio Armani's  
You wanna harm me and Nas you gots ta conquer through a whole army  
The cee-lo rollers, money folders, sippin' Bolla, holdin mad payola  
Slangin that Coke without the Cola  
Me and black don't fake jacks but we might sling one  
It ain't no shame in our game we do our thing son

[Chorus – Eminem (repeat 2x):]  
'Cause I'm Slim Shady, yes I'm the real Shady  
All you other Slim Shadys are just imitating  
So won't the real Slim Shady please stand up,  
Please stand up, please stand up?

They were jammin off a record that said it best:  
"Now what you hear is not a test!"

Aowowowowowowowowowowo!  
A-hunga-hunga-hunga-hunga  
Aowowowowowowowowowowo!  
Aowowowowowowowowowowo!

Form(at)  
amenable  
to  
mathematical/stati  
stical analysis and  
visual display

# Example: Lyrics of Rap Songs

## *Issues with rap/hip hop lyrics*



- No specified format
- Variable length
- Variable spelling
- Punctuation and non-alphanumeric characters
- No predefined content or predefined set of values
- Slang, made up and misspelled words.
- Repetitive content(e.g. choruses)
- Some sources embed annotation (who is singing, how often to repeat a phrase, ...)

# Example: Lyrics of Rap Songs

## *General approach*



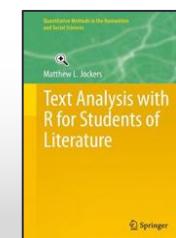
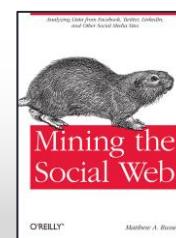
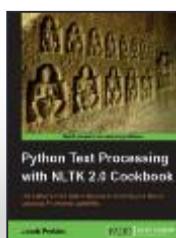
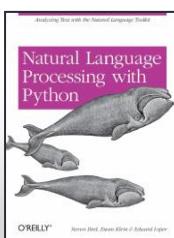
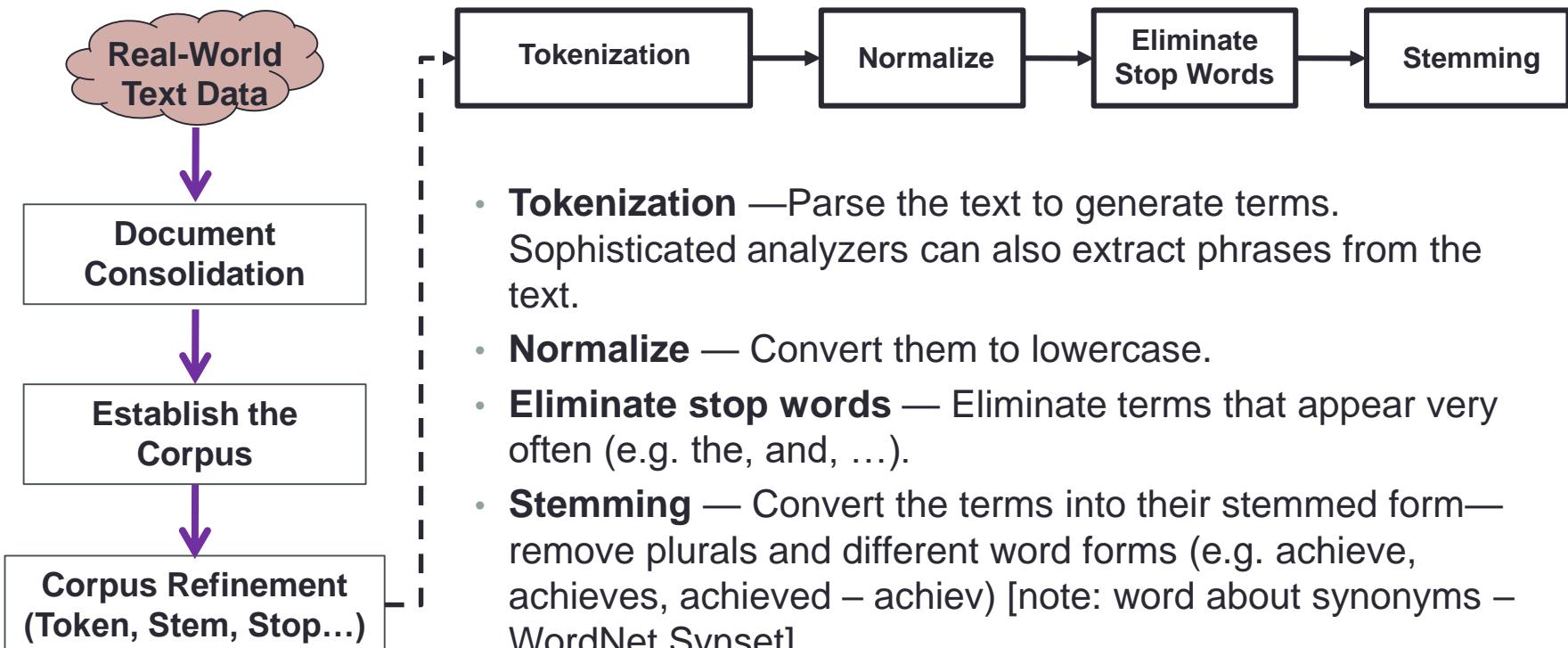
The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods.



It is often assumed that documents are a *bag of words*, where order does not inform our analyses... If this assumption is unpalatable, we can retain some word order by including bigrams (word pairs) or trigrams.

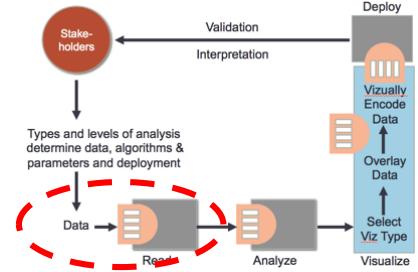
# Example: Lyrics of Rap Songs

## *Text conversion process*



# Example: Lyrics of Rap Songs

## *Text conversion process - Corpus*



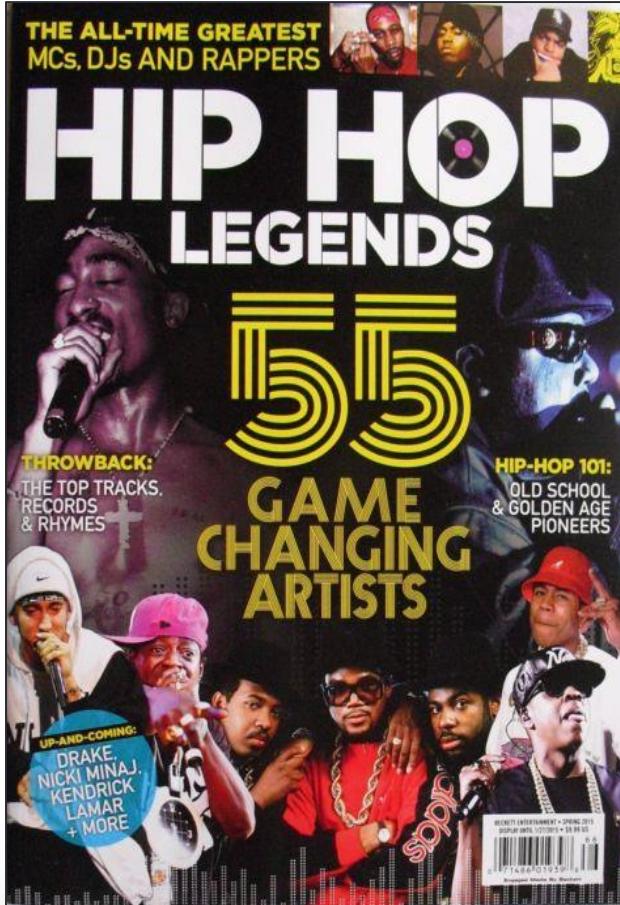
A screenshot of the OHHLA.com website, titled "THE ORIGINAL HIP-HOP LYRICS ARCHIVE". The left sidebar contains links such as Add Lyrics, All Artists, Compilations, Corrections, FAQ, Featured Artists, Links, New Lyrics, Press, RapReviews, RapGenerator, Soundtracks, Store, Support, Top 30 Songs, and Updates. The main content area shows a search bar with "# A B C D E F J K-O P-T U-Z" and a list of categories including 10 X.A.N.'s, 10ision, 11/11, 11/5, 1200 Techniques, 12 O'Clock, 1982, 1.4.0. Productions, 1st Infantry, 213, and 2 Chainz.

A screenshot of the Hip Hop Word Count website. The title "Hip Hop Word Count" is displayed prominently. Below it, a text block states: "A database of the lyrics to more than 50,000 rap songs dating back to 1979, "Hip Hop Word Count" is a tool that's catalyzing beautiful data visualizations and insights into the changing meanings of one of the most ubiquitous art forms of American culture."

A screenshot of the Genius website. The logo features a diamond icon followed by the word "GENIUS". Below the logo, the tagline "Annotate the world." is visible. There are two buttons: "Sign Up" and "Learn More ». At the bottom, a navigation bar includes links for All, Rap, Rock, Lit, Pop, Country, R&B, History, Sports, Law, Tech, and X.

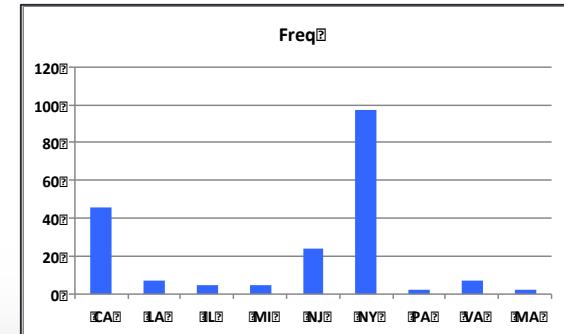
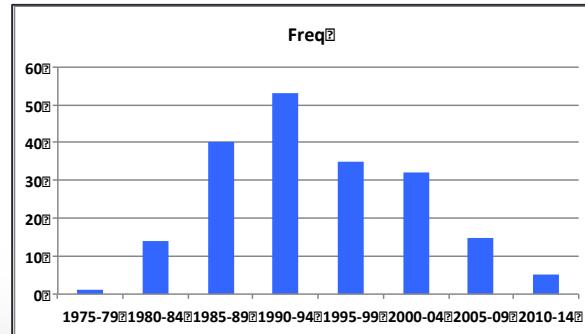
# Example: Lyrics of Rap Songs

*Text conversion process – Hip Hop corpus*



195 Hip Hop Artists (Individuals & Groups)					
Old School	15	Golden Age	86	Gangsta Rap	62
Afrika Bambaataa	3	2 Live Crew	6	2PAC	5 Eminem
Grandmaster Flash & The Furious Five	1	ATCQ	4	Dr. Dre	6 Jay-Z
Kool Moe Dee & Treacherous Three	2	Beastie Boys	6	Ice Cube	8 Kanye West
Kurtis Blow	4	Big Daddy Kane	2	Ice-T	6 Lil Wayne
Sugar Hill Gang	3	Biz Markie	2	Lil Kim	5 Missy Elliott
		Busta Rhymes	7	N.W.A	6
		De La Soul	4	NAS	5
		DJ Jazzy & The Fresh Prince	2	Puff Daddy, P. Diddy	6
		Eric B. & Rakim	7	Snoop Dogg	5
		Gang Starr	2	The Notorious Big	5
		KRS-One	4	Wu-Tang Clan	5
		LL Cool J	9		
		M.C. Hammer	4		
		Public Enemy	7		
		Queen Latifah	6		
		Run-D.M.C.	7		
		Salt-N-Pepa	5		
		Slick Rick	2		

195 Songs  
By  
40 Artists



# Example: Lyrics of Rap Songs

## Text conversion process – Hip Hop corpus

```
Something's Got to Give  
wish for peace between the races  
Someday we shall all be one  
Why fight yourself?  
This one's called 'Rectify'  
There's something coming to the su  
There's fire all around but this i  
I've seen better days than this on  
I've seen better nights than this  
Tension is rebuilding  
Something's got to give  
Something's got to give  
Someday we shall all be one  
Jesus Christ, we're nice
```

### Words - 69/45

[*wish*', *for*', *peace*', *between*', *the*', *races*', *Someday*', *we*', *shall*', *all*', *be*', *one*', *Why*', *fight*', *yourself*', *This*', *ones*', *called*', *Rectify*', *Theres*', *something*', *coming*', *to*', *the*', *surface*', *Theres*', *fire*', *all*', *around*', *but*', *this*', *is*', *all*', *illusion*', *Ive*', *seen*', *better*', *days*', *than*', *this*', *one*', *Ive*', *seen*', *better*', *nights*', *than*', *this*', *one*', *Tension*', *is*', *rebuilding*', *Somethings*', *got*', *to*', *give*', *Somethings*', *got*', *to*', *give*', *Someday*', *we*', *shall*', *all*', *be*', *one*', *Jesus*', *Christ*', *were*', *nice*']

### Lowers - 69/44

[*wish*', *for*', *peace*', *between*', *the*', *races*', *someday*', *we*', *shall*', *all*', *be*', *one*', *why*', *fight*', *yourself*', *this*', *ones*', *called*', *rectify*', *theres*', *something*', *coming*', *to*', *the*', *surface*', *theres*', *fire*', *all*', *around*', *but*', *this*', *is*', *all*', *illusion*', *Ive*', *seen*', *better*', *days*', *than*', *this*', *one*', *Ive*', *seen*', *better*', *nights*', *than*', *this*', *one*', *tension*', *is*', *rebuilding*', *somethings*', *got*', *to*', *give*', *somethings*', *got*', *to*', *give*', *someday*', *we*', *shall*', *all*', *be*', *one*', *jesus*', *christ*', *were*', *nice*']

### Alphas - 69/44

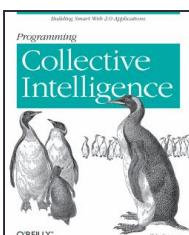
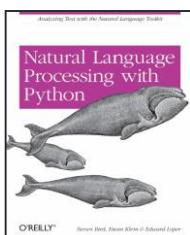
[*wish*', *for*', *peace*', *between*', *the*', *races*', *someday*', *we*', *shall*', *all*', *be*', *one*', *why*', *fight*', *yourself*', *this*', *ones*', *called*', *rectify*', *theres*', *something*', *coming*', *to*', *the*', *surface*', *theres*', *fire*', *all*', *around*', *but*', *this*', *is*', *all*', *illusion*', *Ive*', *seen*', *better*', *days*', *than*', *this*', *one*', *Ive*', *seen*', *better*', *nights*', *than*', *this*', *one*', *tension*', *is*', *rebuilding*', *somethings*', *got*', *to*', *give*', *somethings*', *got*', *to*', *give*', *someday*', *we*', *shall*', *all*', *be*', *one*', *jesus*', *christ*', *were*', *nice*']

### Nonstops - 42/30

[*wish*', *peace*', *races*', *someday*', *shall*', *one*', *fight*', *ones*', *called*', *rectify*', *theres*', *something*', *coming*', *surface*', *theres*', *fire*', *around*', *illusion*', *Ive*', *seen*', *better*', *days*', *one*', *Ive*', *seen*', *better*', *nights*', *one*', *tension*', *rebuilding*', *somethings*', *got*', *give*', *somethings*', *got*', *give*', *someday*', *shall*', *one*', *jesus*', *christ*', *nice*']

### Stems - 42/28

[*wish*', *peac*', *race*', *someday*', *shall*', *one*', *fight*', *one*', *call*', *rectifi*', *there*', *someth*', *come*', *surfac*', *there*', *fire*', *around*', *illus*', *Ive*', *seen*', *better*', *day*', *one*', *Ive*', *seen*', *better*', *night*', *one*', *tension*', *rebuild*', *someth*', *got*', *give*', *someth*', *got*', *give*', *someday*', *shall*', *one*', *jesu*', *christ*', *nice*']



Analysis from Python  
NLTK Program

# Example: Lyrics of Rap Songs

## *Text conversion process – Hip Hop corpus*

```
Something's Got to Give  
wish for peace between the races  
Someday we shall all be one  
Why fight yourself?  
This one's called 'Rectify'  
There's something coming to the  
There's fire all around but this  
I've seen better days than this  
I've seen better nights than thi  
Tension is rebuilding  
Something's got to give  
Something's got to give  
Someday we shall all be one  
Jesus Christ, we're nice
```

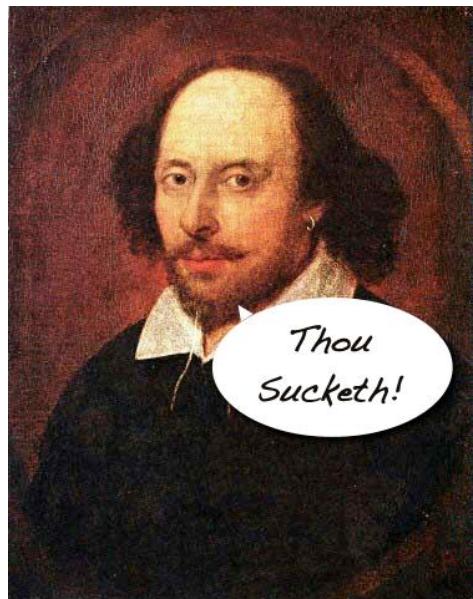
“Bag of Words,  
Terms, Tokens,  
Stems, ...”



```
wordsCnts  
{'all': 4, 'Christ': 1, 'rebuilding': 1, 'is': 2, 'Somethings': 2, 'surface': 1, 'yourself': 1, 'something': 1, 'seen': 2, 'Jesus': 1, 'fire': 1, 'Tension': 1, 'nights': 1, 'for': 1, 'peace': 1, 'fight': 1, 'better': 2, 'to': 3, 'I've': 2, 'between': 1, 'got': 2, 'Why': 1, 'nice': 1, 'be': 2, 'we': 2, 'around': 1, 'shall': 2, 'This': 1, 'but': 1, 'ones': 1, 'coming': 1, 'one': 4, 'than': 2, 'Theres': 2, 'Someday': 2, 'give': 2, 'Rectify': 1, 'this': 3, 'wish': 1, 'days': 1, 'illusion': 1, 'races': 1, 'were': 1, 'the': 2, 'called': 1}  
lowerCnts  
{'someday': 2, 'all': 4, 'rectify': 1, 'is': 2, 'surface': 1, 'yourself': 1, 'I've': 2, 'something': 1, 'seen': 2, 'jesus': 1, 'nights': 1, 'christ': 1, 'for': 1, 'peace': 1, 'somethings': 2, 'fight': 1, 'better': 2, 'to': 3, 'between': 1, 'got': 2, 'illusion': 1, 'nice': 1, 'be': 2, 'we': 2, 'around': 1, 'than': 2, 'shall': 2, 'fire': 1, 'rebuilding': 1, 'but': 1, 'ones': 1, 'coming': 1, 'one': 4, 'why': 1, 'give': 2, 'this': 4, 'wish': 1, 'days': 1, 'tension': 1, 'races': 1, 'were': 1, 'the': 2, 'called': 1, 'theres': 2}  
alphasCnts  
{'someday': 2, 'all': 4, 'rectify': 1, 'is': 2, 'surface': 1, 'yourself': 1, 'I've': 2, 'something': 1, 'seen': 2, 'jesus': 1, 'nights': 1, 'christ': 1, 'for': 1, 'peace': 1, 'somethings': 2, 'fight': 1, 'better': 2, 'to': 3, 'between': 1, 'got': 2, 'illusion': 1, 'nice': 1, 'be': 2, 'we': 2, 'around': 1, 'than': 2, 'shall': 2, 'fire': 1, 'rebuilding': 1, 'but': 1, 'ones': 1, 'coming': 1, 'one': 4, 'why': 1, 'give': 2, 'this': 4, 'wish': 1, 'days': 1, 'tension': 1, 'races': 1, 'were': 1, 'the': 2, 'called': 1, 'theres': 2}  
nonstopsCnts  
{'someday': 2, 'rectify': 1, 'surface': 1, 'one': 4, 'I've': 2, 'something': 1, 'seen': 2, 'nights': 1, 'christ': 1, 'peace': 1, 'somethings': 2, 'fight': 1, 'better': 2, 'got': 2, 'illusion': 1, 'nice': 1, 'tension': 1, 'around': 1, 'shall': 2, 'fire': 1, 'ones': 1, 'coming': 1, 'jesus': 1, 'give': 2, 'wish': 1, 'days': 1, 'races': 1, 'rebuilding': 1, 'called': 1, 'theres': 2}  
stemsCnts  
{'someday': 2, 'I've': 2, 'jesu': 1, 'one': 5, 'rectifi': 1, 'seen': 2, 'surfac': 1, 'christ': 1, 'come': 1, 'someth': 3, 'there': 2, 'fight': 1, 'better': 2, 'call': 1, 'got': 2, 'nice': 1, 'tension': 1, 'around': 1, 'shall': 2, 'fire': 1, 'illus': 1, 'day': 1, 'give': 2, 'wish': 1, 'peac': 1, 'race': 1, 'night': 1, 'rebuild': 1}
```

# Example: Lyrics of Rap Songs

*Text conversion process – structural features*



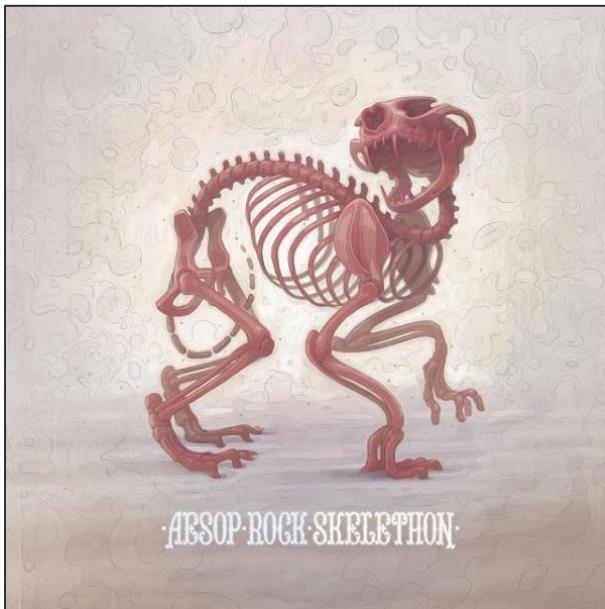
	Words	Alphas	Stems
Total	127643	120263	52094
TotUniq	12521	10182	7675
NumLines	16412	16412	16412
Chars	495063	487265	247340
AvgWrd/Ln	7.8	7.3	3.2
AvgCh/Word	3.9	4.1	4.7
LexDiv	10.2	11.8	6.8

Author	WordLen	AvgWords/Sent	Words/Voca
austen-emma.txt	4	21	26
austen-persuasion.txt	4	23	16
austen-sense.txt	4	24	22
bible-kjv.txt	4	33	79
blake-poems.txt	4	18	5
bryant-stories.txt	4	17	14
burgess-busterbrown.txt	4	14	12
carroll-alice.txt	4	16	12
chesterton-ball.txt	4	17	11
chesterton-brown.txt	4	19	11
chesterton-thursday.txt	4	16	10
edgeworth-parents.txt	4	18	24
melville-moby_dick.txt	4	24	15
milton-paradise.txt	4	52	10
shakespeare-caesar.txt	4	12	8
shakespeare-hamlet.txt	4	13	7
shakespeare-macbeth.txt	4	13	6
whitman-leaves.txt	4	35	12



# Example: Lyrics of Rap Songs

## *Text conversion process – structural features*



...

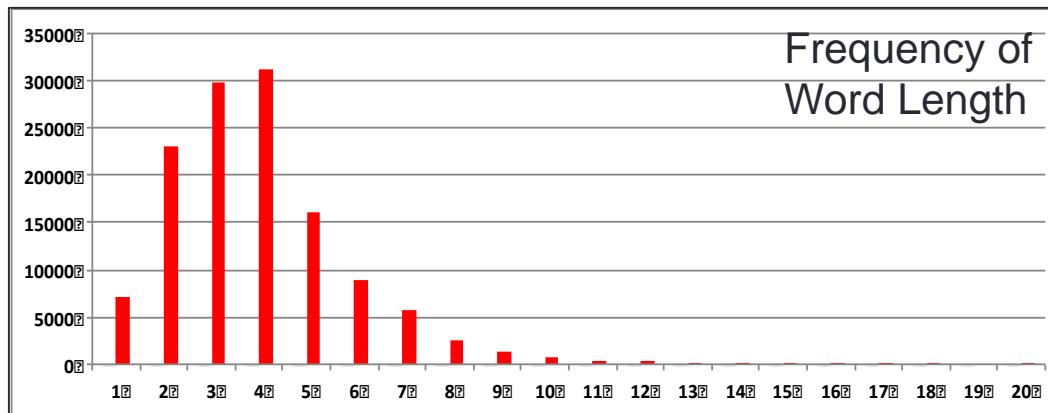
Final answer "not to be", "not to be" is right!  
Next question - to build winged shoes or autophagy  
Silk screen band tees, take apart a vcr, ringer off, canned peas  
Cabin fever mi amor  
Patiently adhering to the chandelier ta key-in-door  
To usher in the understated anarchy of leisureforce  
Led a purple tongue and ratty caballeros  
Up over the black rainbow into the house of mirrors  
To become a thousand zeroes  
Echoing a twisted alchemy, freak flags, fluttering to circadian free jazz  
Sleep apnea scratching "bring that beat back"  
I doze off, clothes on, noise in the feedbag  
Shhh.. om nom nom, blinds drawn  
Compost thrown to the spine pile, bygones, mangy  
Intimately spaced pylons on a plot of inhospitable terrain  
Hi mom!

...

(from Leisureforce)

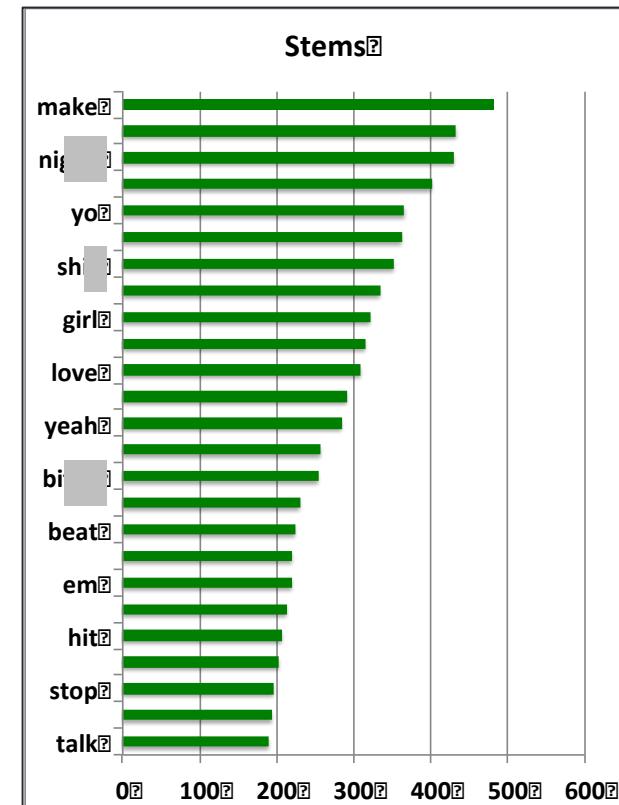
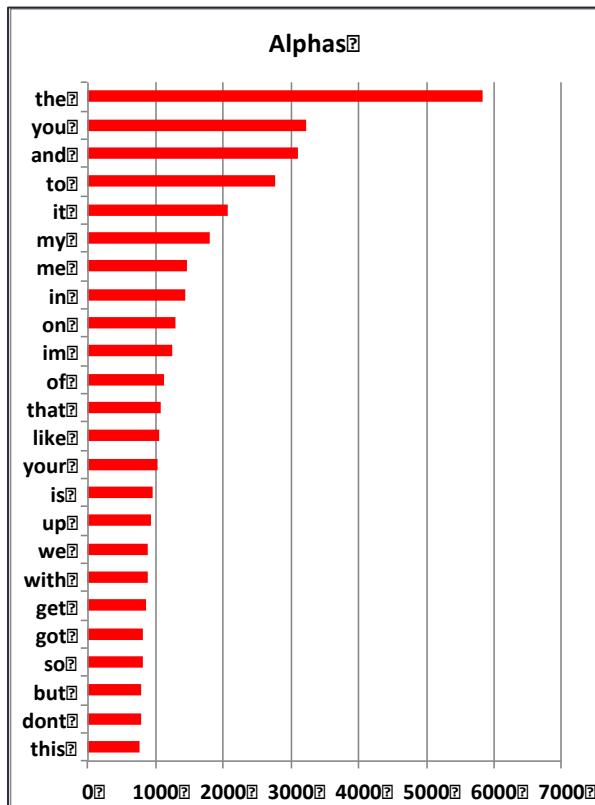
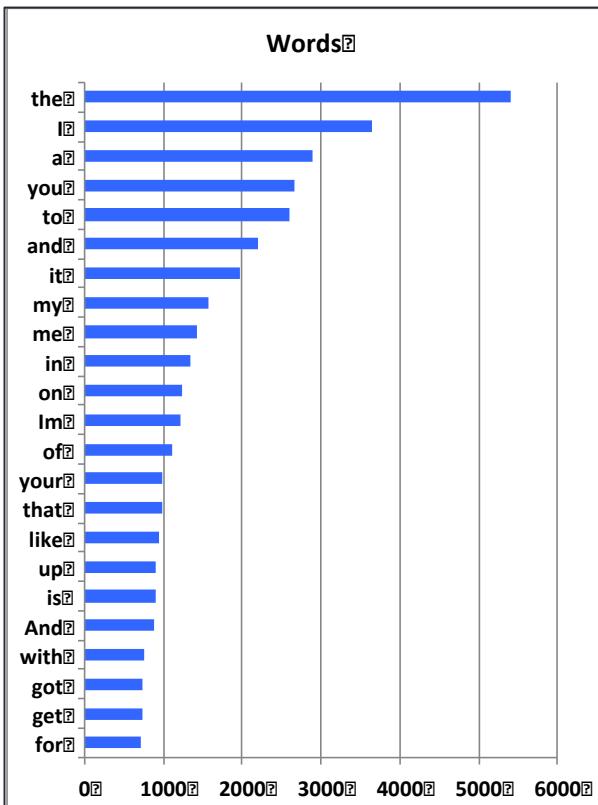
# Example: Lyrics of Rap Songs

## *Text conversion process – word length*



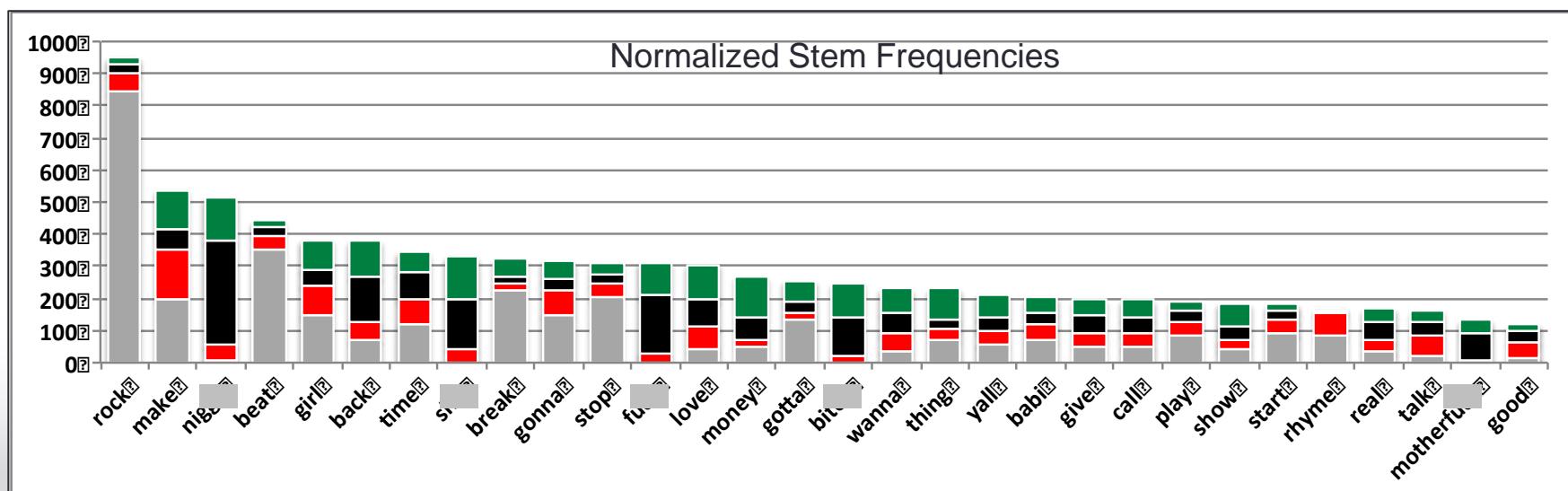
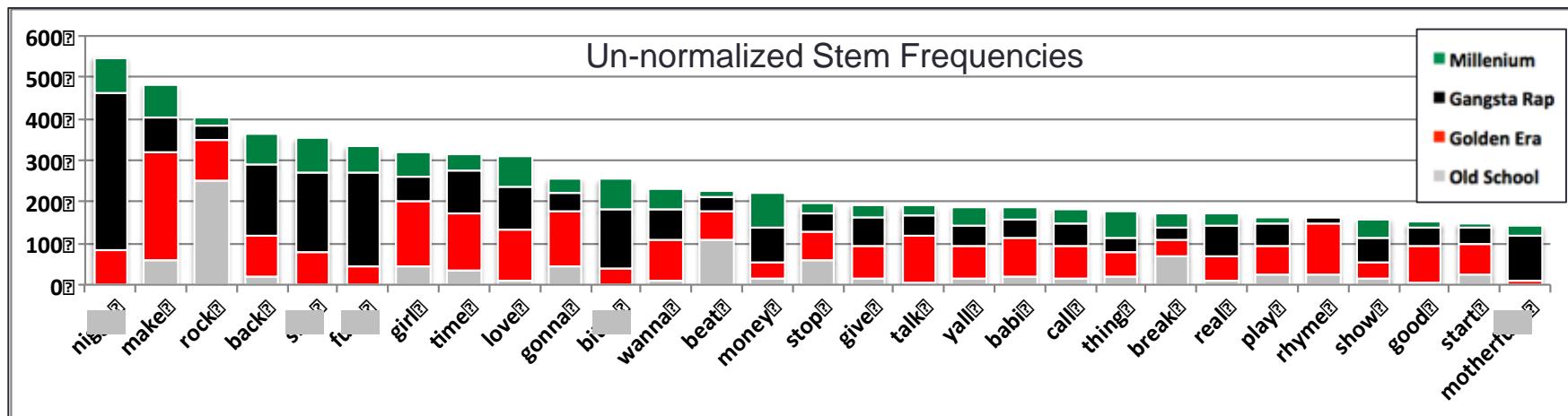
# Example: Lyrics of Rap Songs

## *Text conversion process – frequencies*



# Example: Lyrics of Rap Songs

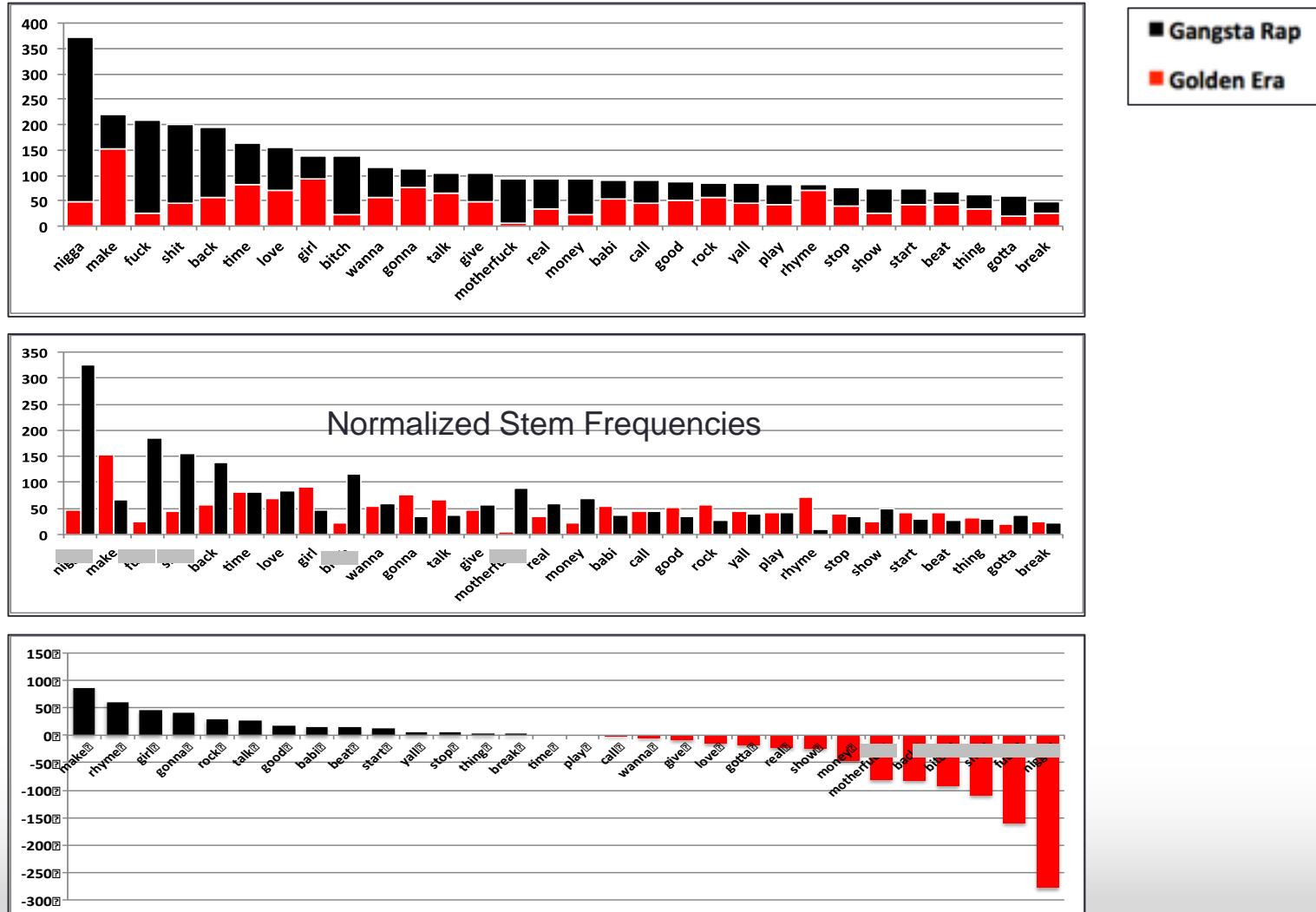
*Text analysis process – stem frequencies by era*



# Example: Lyrics of Rap Songs

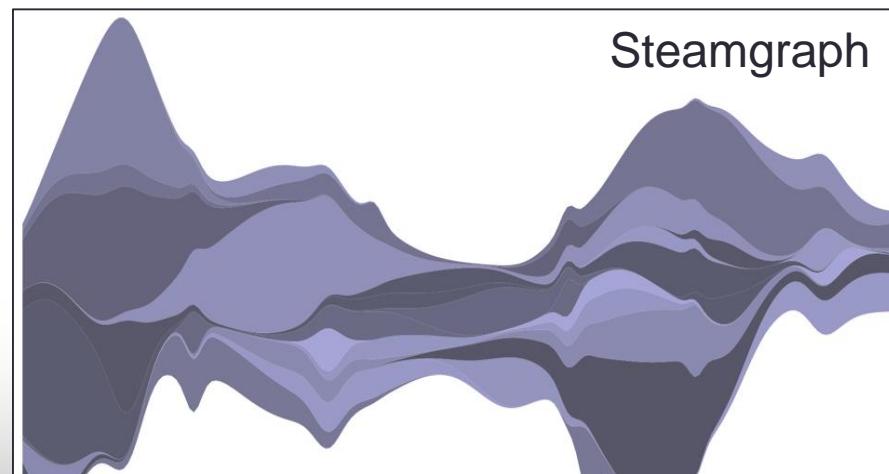
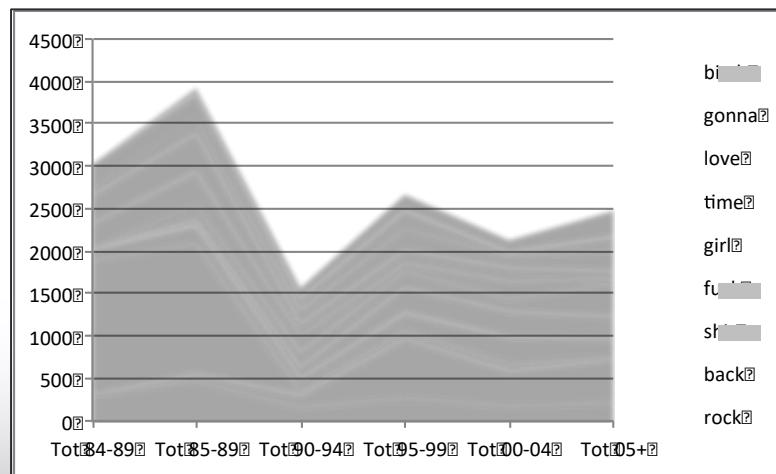
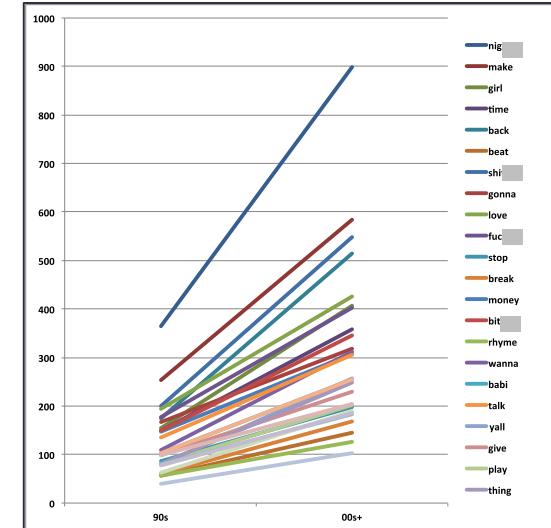
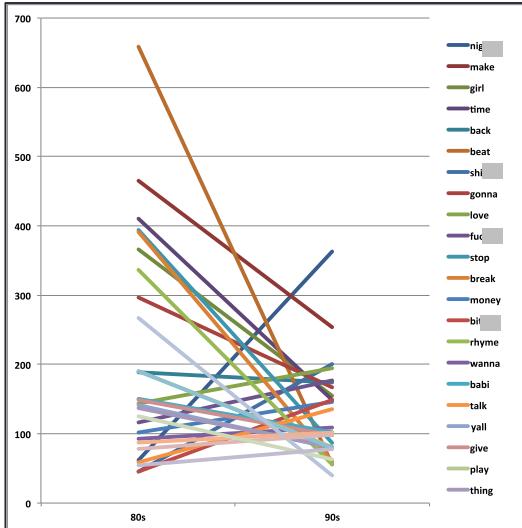
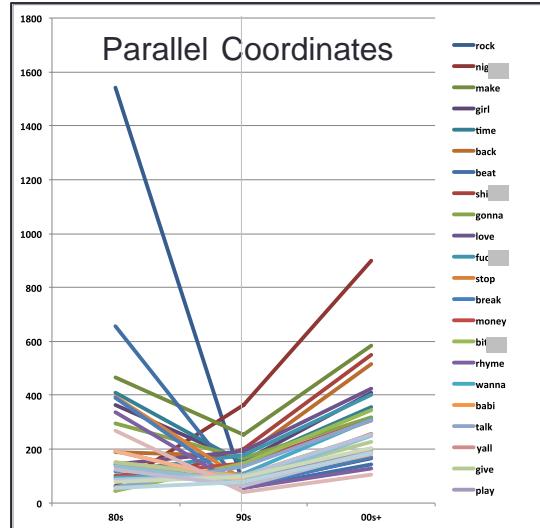
*Text analysis process – stem frequencies by era*

Normalized  
Stem  
Frequencies



# Example: Lyrics of Rap Songs

*Text analysis process – stem frequencies by time*



# Example: Lyrics of Rap Songs

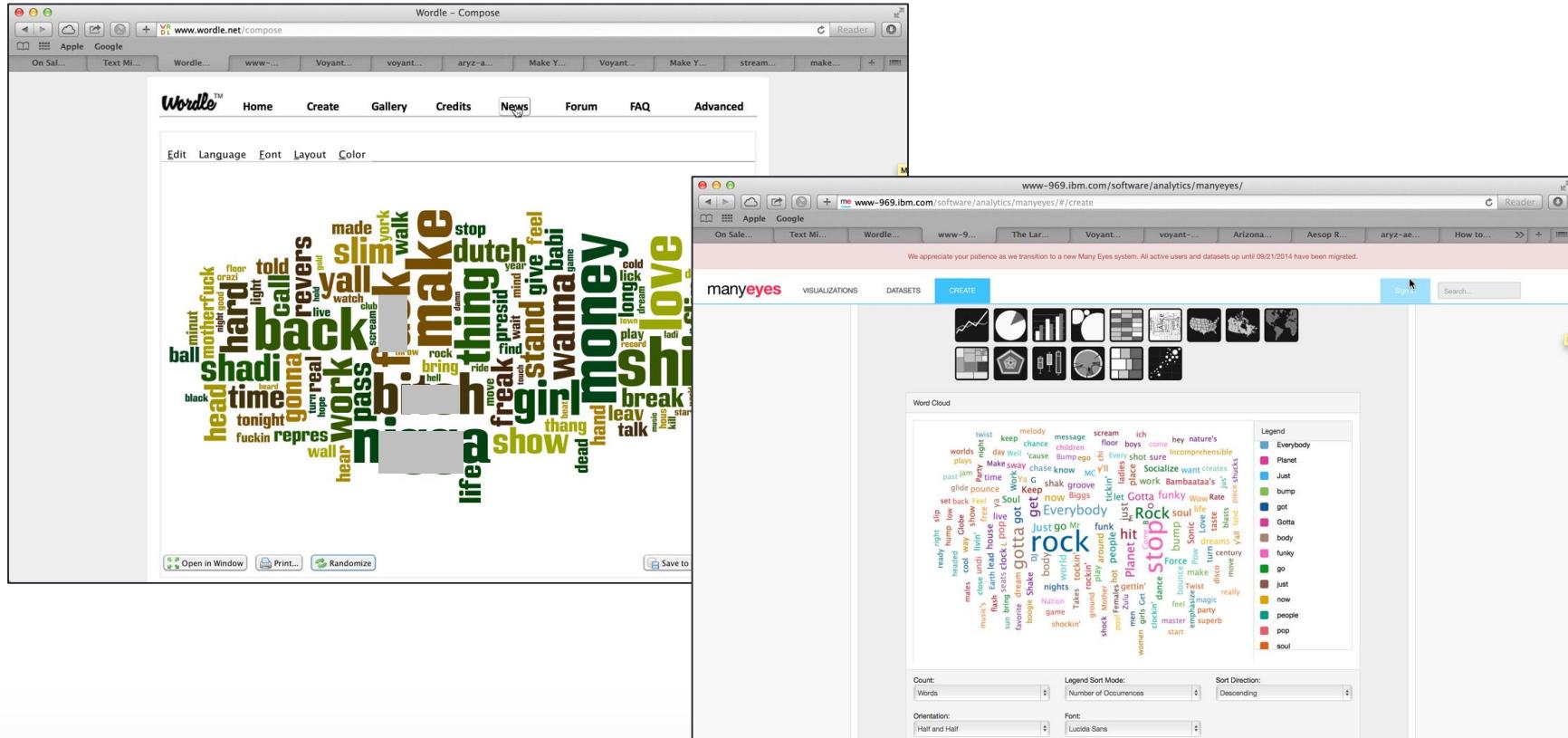
## *Text analysis process – frequencies with less precision*



Word Clouds – They're everywhere along with subway maps, periodic charts...

# Example: Lyrics of Rap Songs

*Text analysis process – frequencies with less precision*

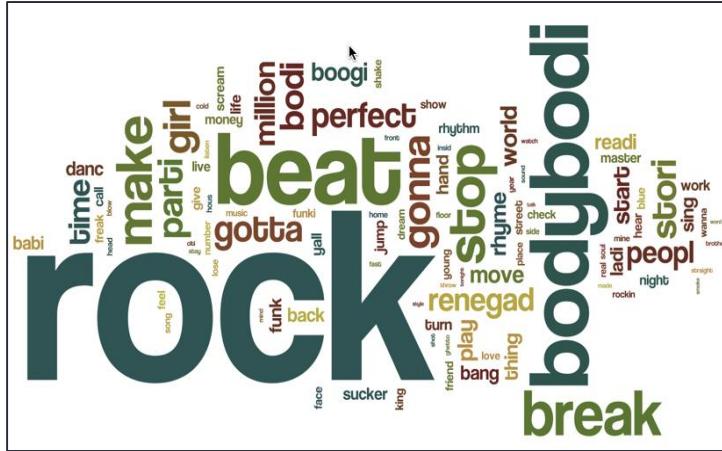


Word Clouds – They're everywhere along with subway maps, periodic charts...

# Example: Lyrics of Rap Songs

## *Text analysis process – frequencies with less precision*

# Old School



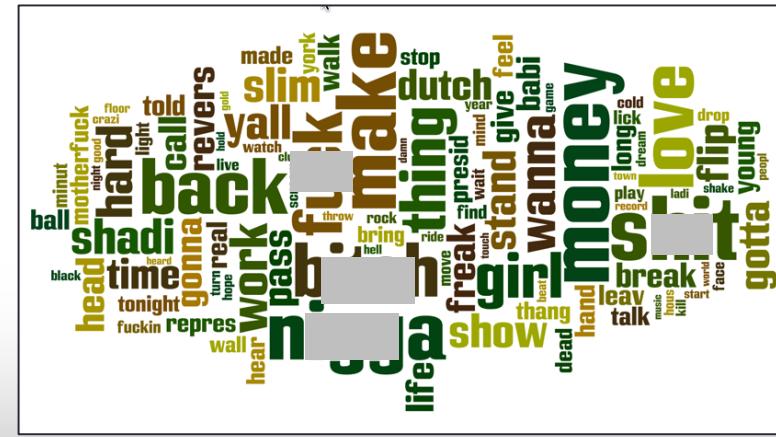
# Golden Era



# Gangsta

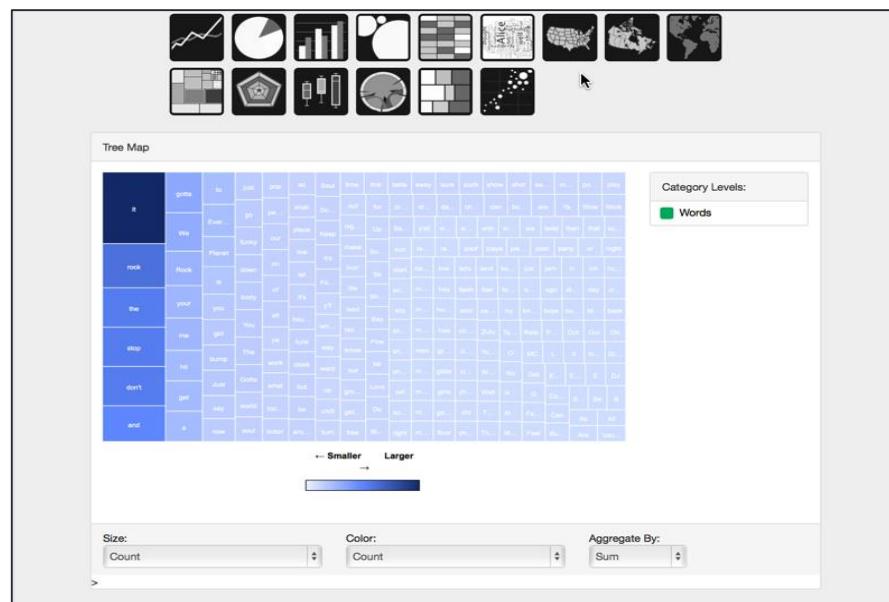
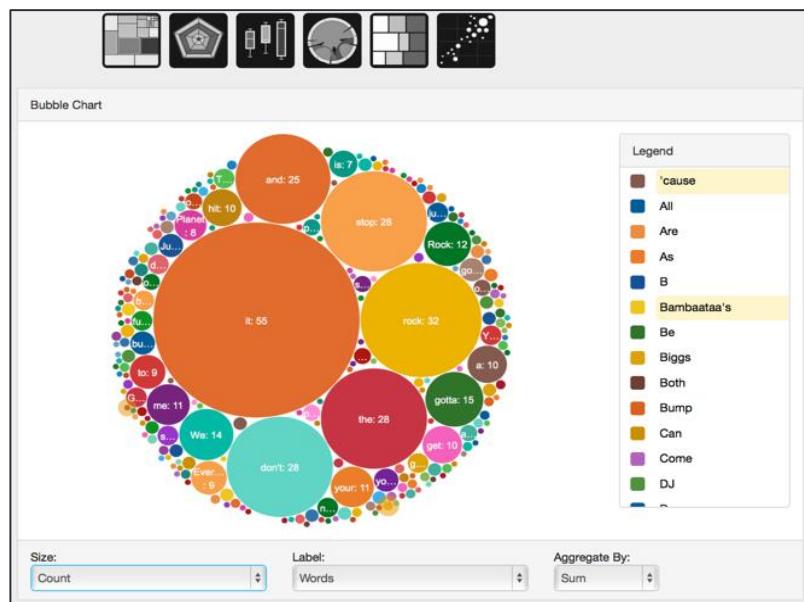


# Millenium



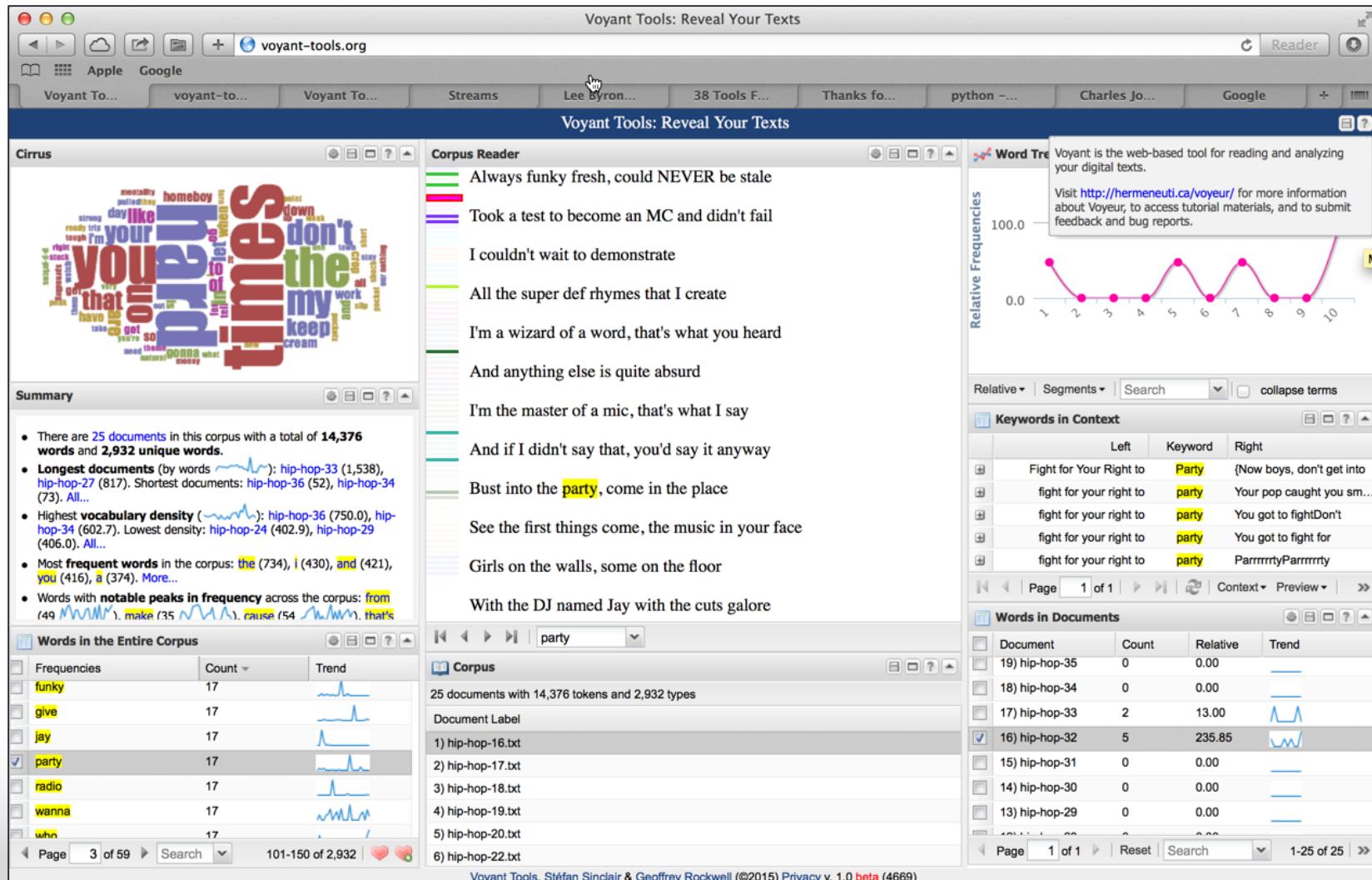
# Example: Lyrics of Rap Songs

## *Text analysis process – frequencies with less precision*



# Example: Lyrics of Rap Songs

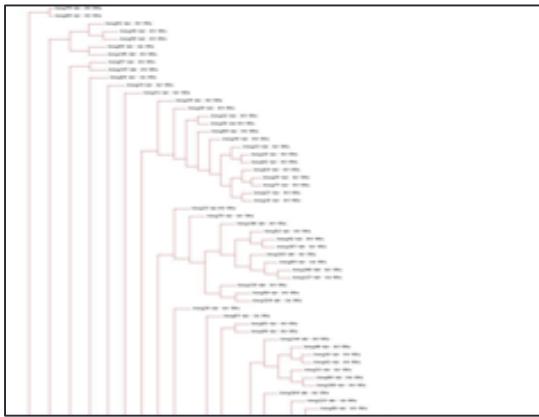
*Text analysis process – general interactive platform*



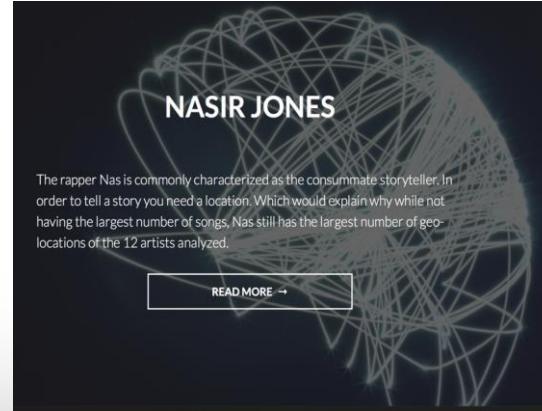
# Example: Lyrics of Rap Songs

## *Types of (visual) analysis*

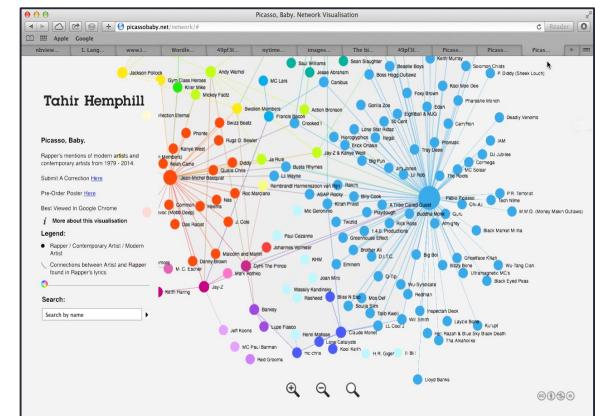
### Cluster/Classification



### Geospatial

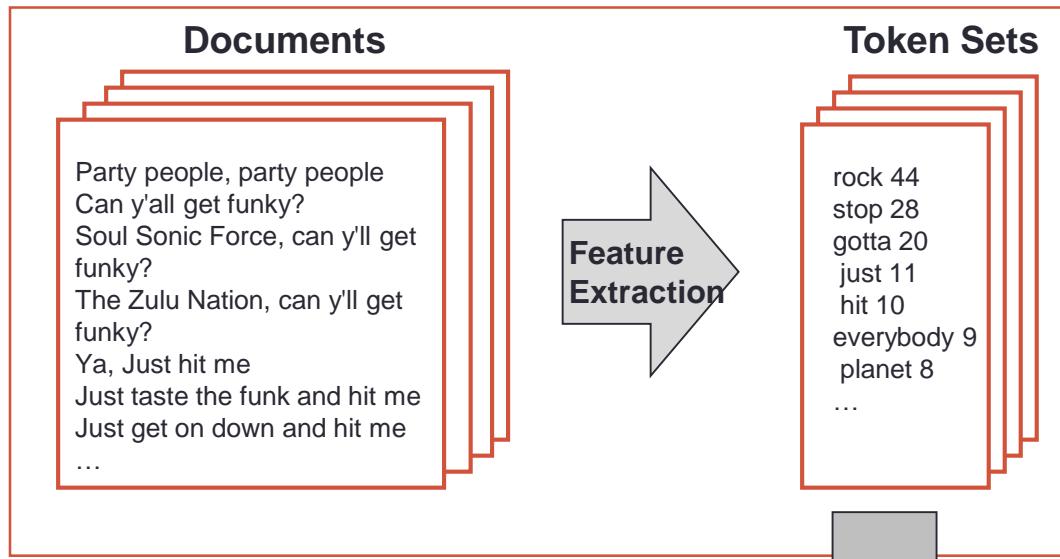


### Network



# Example: Lyrics of Rap Songs

*Text analysis process – Role of Document-Term Matrix*



*Feature Extraction & Weighting*

**“Bag of Words, Terms or Tokens”**

**Doc/Token Matrix:  
Vectors of Words, Terms or Tokens by Doc**

	Token1	Token2	Token3	Token4	...
Doc1	1	2	2	4	
Doc2	4	2	3	0	
Doc3	1	1	1	0	
Doc4	1	1	1	2	
...					

**“Bag of Words” (BOW) or  
Vector Space Model (VSM):  
Words or Tokens are  
attributes and documents  
are examples**

# Example: Lyrics of Rap Songs

*Text analysis process – Document-Term Matrix*

SongCode	make	nigga	rock	back	shit	fuck	...	girl	sell	suck	sweat	thug	uhuh	week	worri	Total Items	0	0	3	2
Song1-OS-INY-80s	3	0	43	1	0	0	1	2	2	0	0	0	0	0	0	193	1	0	1	6
Song2-OS-INY-80s	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0	116	2	1	3	1
Song3-OS-INY-80s	3	0	8	0	0	0	...	1	0	0	0	0	0	0	0	174	3	1	3	0
Song4-OS-INY-80s	3	0	1	0	0	0	...	2	0	0	0	0	0	0	1	82	1	0	0	1
Song5-OS-INY-80s	0	0	1	1	0	0	...	3	0	0	0	0	0	0	0	166	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Song16-GA-INY-80s	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	55	0	0	1	0
Song17-GA-INY-80s	2	0	2	3	0	0	...	3	0	0	0	0	0	0	0	93	0	0	0	1
Song18-GA-INY-80s	2	0	6	1	0	0	...	3	0	0	0	0	0	0	0	108	1	0	0	2
Song19-GA-INY-80s	1	0	10	1	0	0	...	0	1	0	0	0	0	0	1	98	0	0	0	0
Song20-GA-INY-80s	0	0	6	1	0	0	...	0	0	0	0	0	0	0	0	103	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Song102-GR-INJ-90s	0	0	0	0	1	32	...	0	0	0	0	0	0	0	0	106	0	0	3	4
Song103-GR-INJ-90s	0	3	0	2	1	1	...	1	0	0	1	0	0	0	0	98	0	0	2	0
Song104-GR-INJ-90s	1	0	3	5	5	0	...	2	0	0	0	0	0	0	0	190	1	0	13	17
Song105-GR-INJ-90s	1	0	1	0	0	0	...	0	0	0	0	0	0	0	0	56	0	0	2	1
Song106-GR-INJ-90s	1	0	0	1	0	0	...	1	1	0	1	0	0	0	0	101	8	3	13	11
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Song164-ML-ILA-00s	9	4	0	6	1	0	...	0	0	0	6	0	0	0	0	123	1	0	12	1
Song165-ML-ILA-00s	7	0	0	1	3	2	...	0	0	0	0	0	0	0	1	84	0	11	0	0
Song166-ML-ILA-00s	0	0	0	1	0	0	...	0	0	0	0	0	0	0	0	49	0	0	0	0
Song167-ML-ILA-00s	0	1	1	0	4	0	...	1	0	0	0	0	0	0	0	66	1	0	12	1
Song168-ML-ILA-00s	0	6	0	3	2	1	...	0	0	0	0	0	0	0	0	77	0	11	0	0
Total Items	482	430	401	362	352	334	...	321	31	31	31	31	31	31	31	12	10	5	3	2

Song1-OS-INJ'	3	0	43	1	0	0	1	2	2	0	0	0	0	0	28	0	0	1
Song2-OS-INJ'	1	0	0	0	0	0	0	6	0	0	0	2	0	2	0	1	0	0

Song13-OS-IL	5	0	112	2	0	0	0	3	0	11	0	0	7	1	1	0	0	0
Song14-OS-IL	8	0	39	2	0	0	12	5	1	10	0	1	4	1	2	5	2	2

Which pair of song vectors is more correlated? (.05 vs .49)

Which pair of word vectors is more correlated? (.7 vs. .6)

# Example: Lyrics of Rap Songs

*Text analysis process – Document-Term Matrix*

## Normalizing the data

- Binary Frequencies:  $tf = 1$  for  $tf > 0$ ; otherwise 0
- Term Frequencies:  $tf(i,j) / \text{Sum of } tf(i,j) \text{ in Doc K}$
- Log Frequencies:  $1 + \log(tf)$  for  $tf > 0$ ; otherwise 0
- Normalized Frequencies: Divide each frequency by  $\sqrt{\text{Sum of Squares of the frequencies within the vector (column)}}$
- Term Frequency–Inverse Document Frequency
  - TF: Freq of term for given doc/ sum of words for given doc
  - Inverse Document Frequency:  $\log(N/(1+D))$  where N is total number of docs and D is number with term
  - TF \* IDF

# Example: Lyrics of Rap Songs

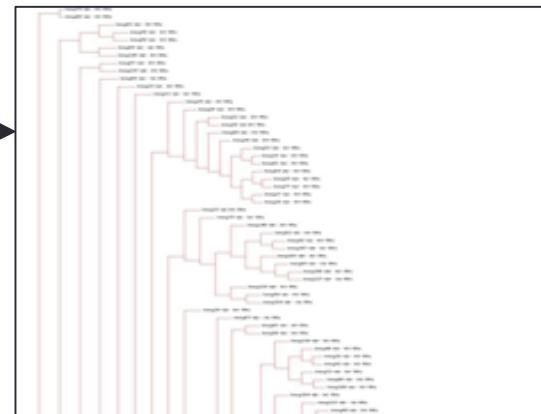
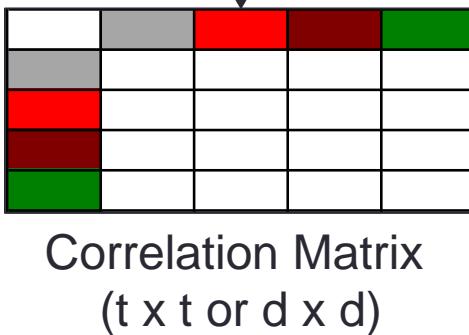
## *Text analysis process – Clustering*

SongCode	reku	nigga	rock	buck	shh	fuck	---	girl	self	rock	sweat	thug	uhhh	wow!	TotItems
Song1-GA-NY-90s	3	0	43	1	0	0	-	1	0	0	0	0	0	0	10
Song2-GA-NY-90s	1	0	0	0	0	0	-	0	0	0	0	0	0	0	5
Song3-GA-NY-90s	1	0	0	0	0	0	-	2	0	0	0	0	0	0	5
Song4-GA-NY-90s	3	0	1	0	0	0	-	2	0	0	0	0	1	0	8
Song5-GA-NY-90s	0	0	1	1	0	0	-	3	0	0	0	0	0	0	10
Song6-GA-NY-90s	0	0	0	0	0	0	-	0	0	0	0	0	0	0	5
Song7-GA-NY-90s	2	0	2	3	0	0	-	3	0	0	0	0	0	0	9
Song8-GA-NY-90s	1	0	0	1	0	0	-	2	0	0	0	0	0	0	6
Song9-GA-NY-90s	1	0	10	1	0	0	-	0	1	0	0	0	1	0	10
Song10-GA-NY-90s	0	0	6	1	0	0	-	0	0	0	0	0	0	0	7
Song11-GA-NY-90s	0	0	0	0	0	0	-	0	0	0	0	0	0	0	5
Song12-GA-NY-90s	0	0	0	0	1	32	-	0	0	0	0	0	0	0	106
Song13-GA-NY-90s	0	3	0	2	1	-	-	1	0	0	0	0	0	0	98
Song14-GA-NY-90s	1	0	0	0	0	0	-	2	0	0	0	0	0	0	100
Song15-GA-NY-90s	1	0	1	0	0	0	-	0	0	0	0	0	0	0	56
Song16-GA-NY-90s	1	0	0	1	0	0	-	0	0	0	0	0	0	0	101
Song17-GA-NY-90s	1	0	0	1	0	0	-	1	1	0	0	0	0	0	57
Song18-GA-NY-90s	0	0	0	0	0	0	-	0	0	0	0	0	0	0	57
Song19-GA-NY-90s	9	4	0	6	1	0	-	0	0	0	0	6	0	0	123
Song20-GA-NY-90s	7	0	0	1	3	2	-	0	0	0	0	0	1	0	84
Song21-GA-NY-90s	0	0	0	1	0	0	-	0	0	0	0	0	0	0	56
Song22-GA-NY-90s	0	1	1	0	4	0	-	1	0	0	0	0	0	0	66
Song23-GA-NY-90s	0	6	0	3	2	1	-	0	0	0	0	0	0	0	77
Total Items	482	430	401	362	352	344	-	221	31	31	31	31	31	31	1000

Doc-Term Matrix

SongCode	make	nigga	rock	buck	shh	---	girl	self	rock	sweat	thug	uhhh	wow!	TotItems	
Song1-GA-NY-90s	100	55	100	100	100	100	100	100	100	100	100	100	100	100	
Song2-GA-NY-90s	95	55	100	100	100	100	100	100	100	100	100	100	100	100	
Song3-GA-NY-90s	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
Song4-GA-NY-90s	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
Song5-GA-NY-90s	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
Song6-GA-NY-90s	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
Song7-GA-NY-90s	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
Song8-GA-NY-90s	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
Song9-GA-NY-90s	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
Song10-GA-NY-90s	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
Song11-GA-NY-90s	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
Song12-GA-NY-90s	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
Song13-GA-NY-90s	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
Song14-GA-NY-90s	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
Song15-GA-NY-90s	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
Song16-GA-NY-90s	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
Song17-GA-NY-90s	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
Song18-GA-NY-90s	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
Song19-GA-NY-90s	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
Song20-GA-NY-90s	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
Song21-GA-NY-90s	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
Song22-GA-NY-90s	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
Song23-GA-NY-90s	100	100	100	100	100	100	100	100	100	100	100	100	100	100	
Total Items	482	430	401	362	352	344	-	221	31	31	31	31	31	31	1000

TD-IDF Scores



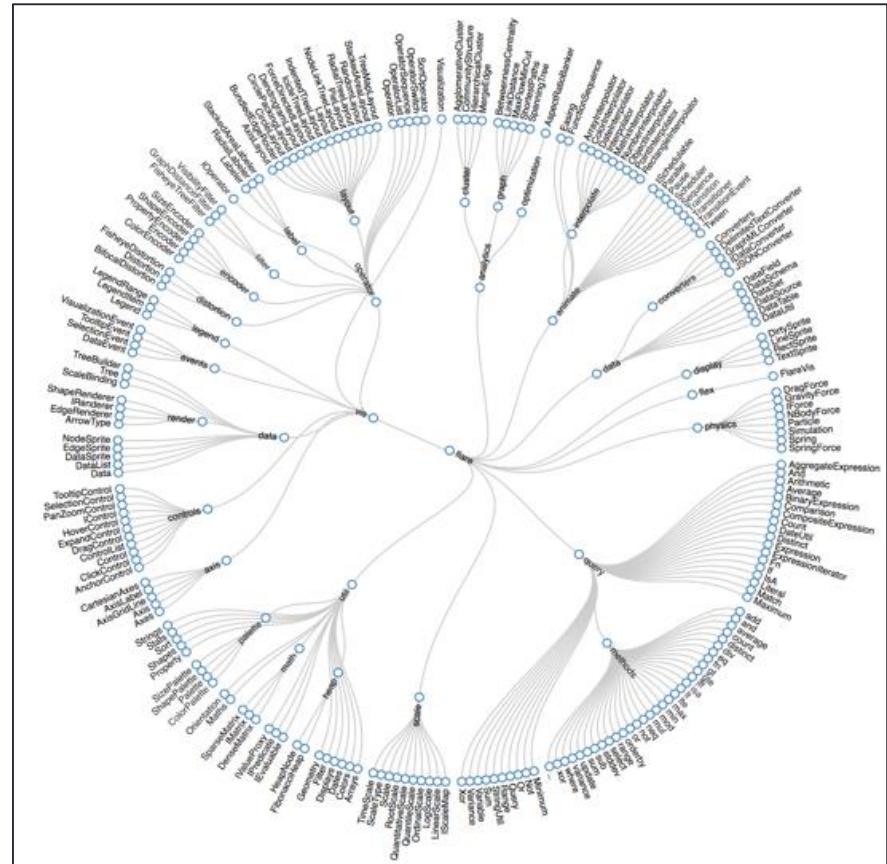
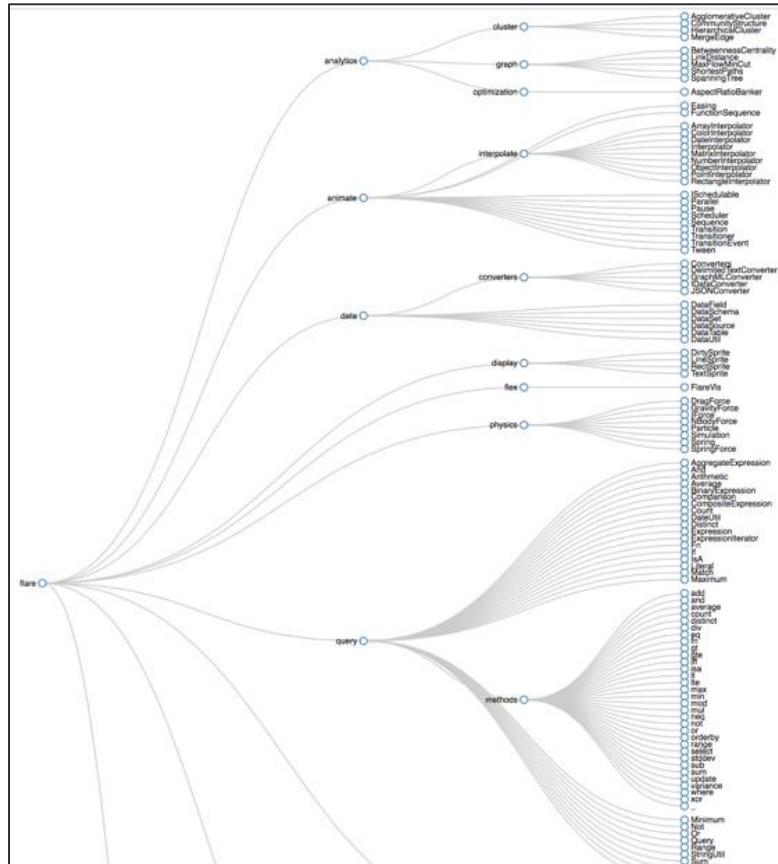
# Example: Lyrics of Rap Songs

## *Text analysis process – Clustering*



# Example: Lyrics of Rap Songs

## *Text analysis process – Clustering*



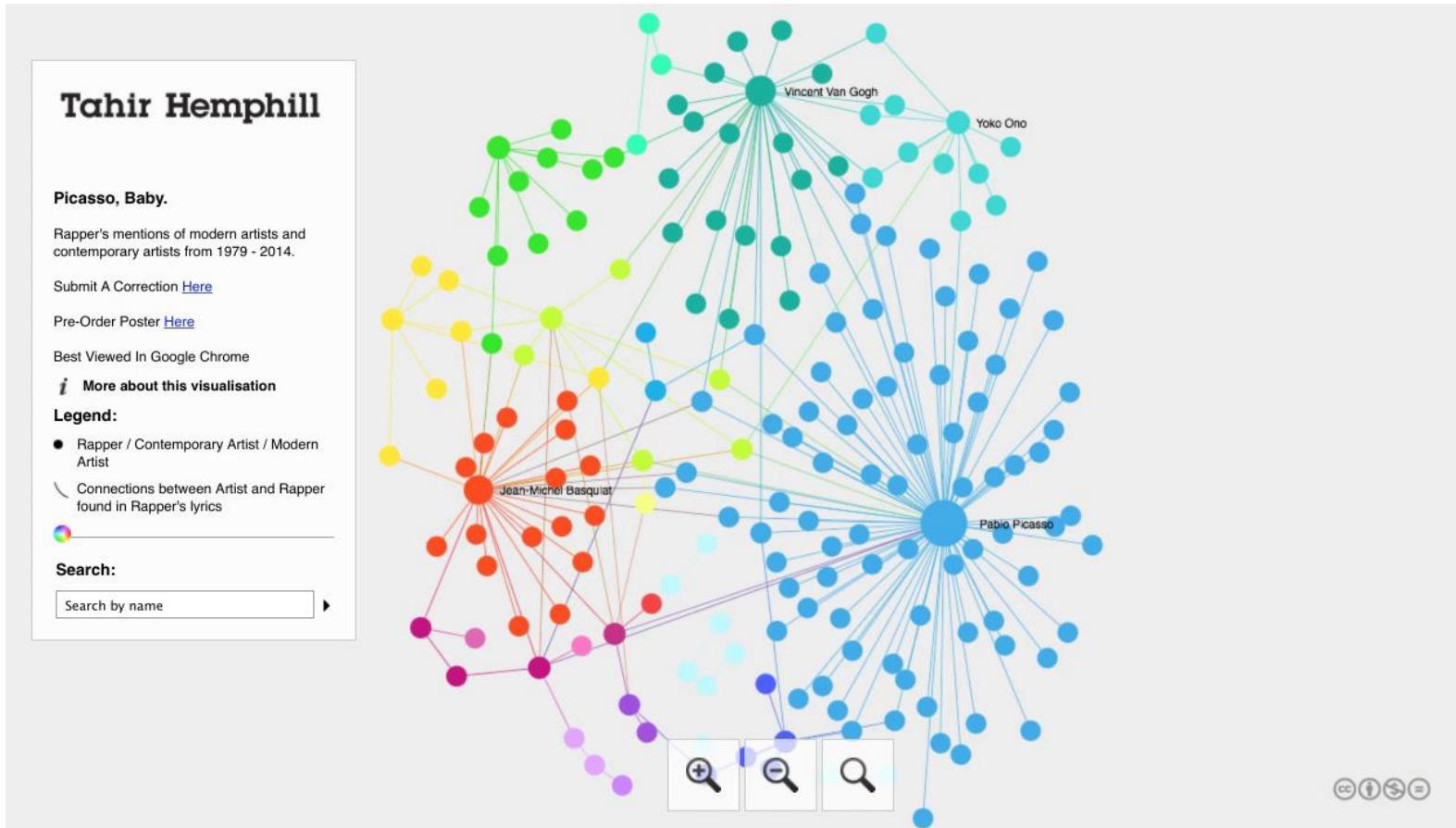
New Versions

# EXAMPLE: SOCIAL NETWORK ANALYSIS

---

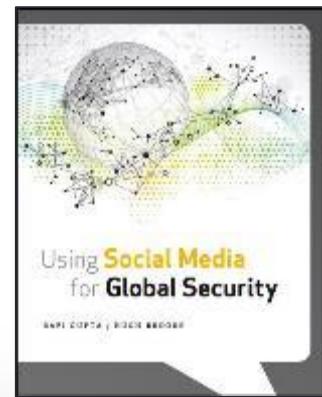
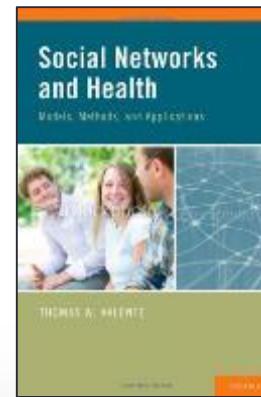
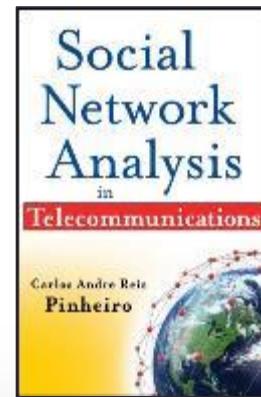
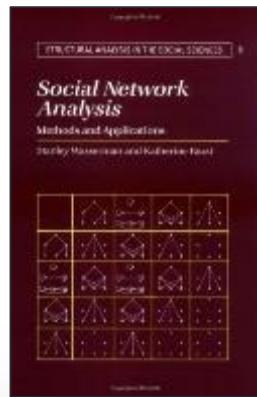
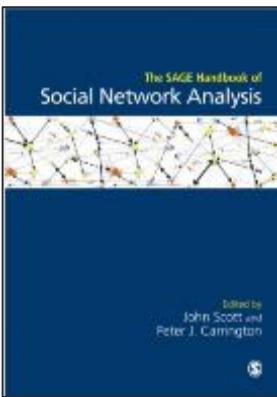
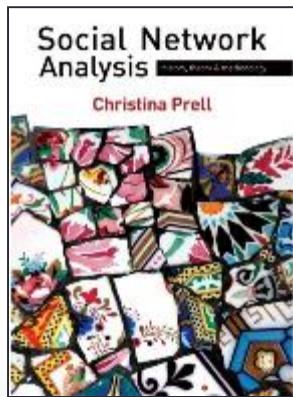
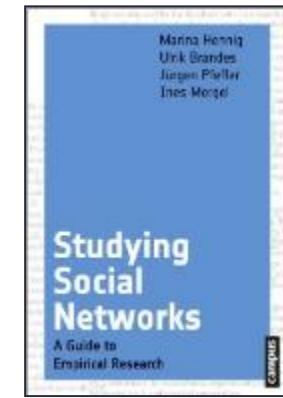
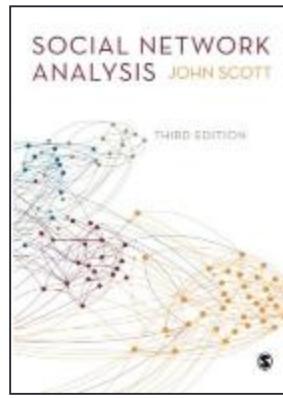
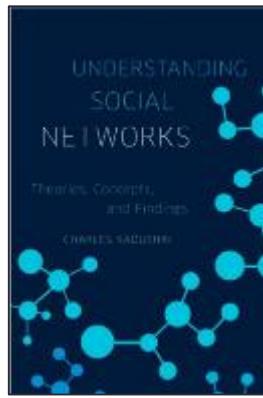
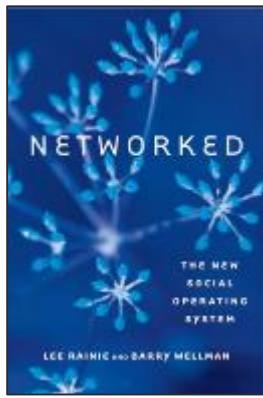
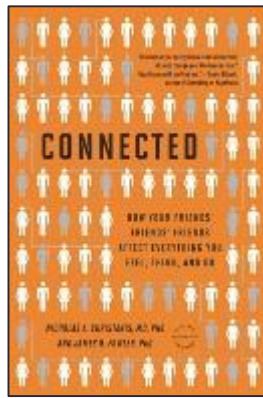
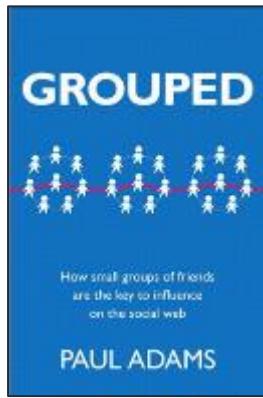
# Example: Lyrics of Rap Songs

## Network analysis



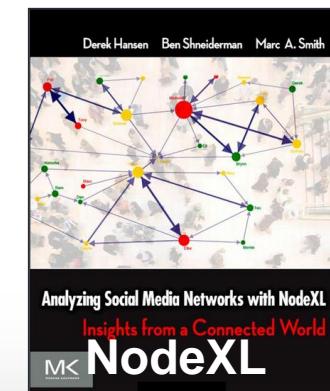
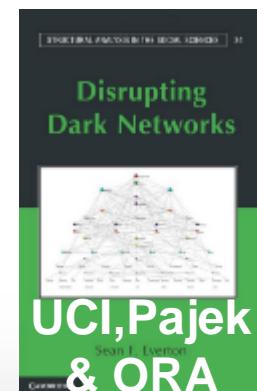
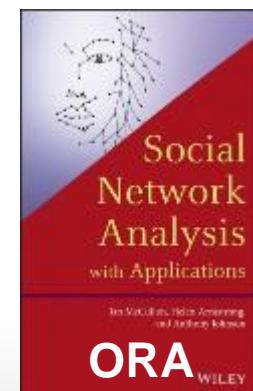
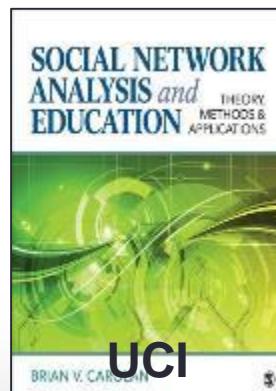
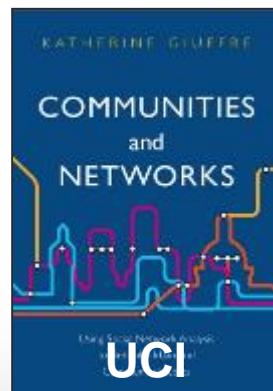
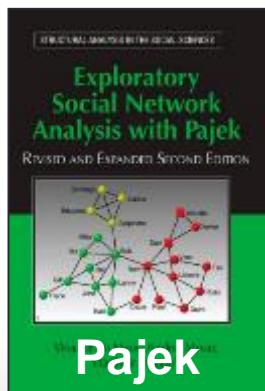
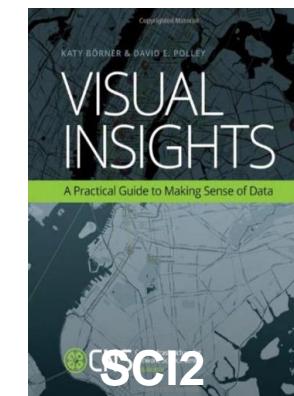
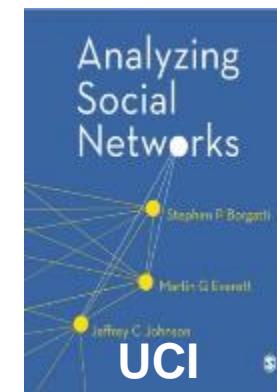
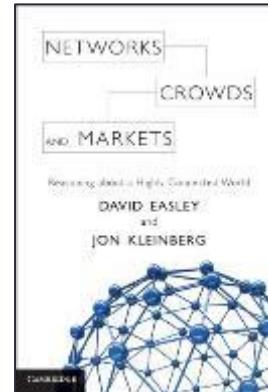
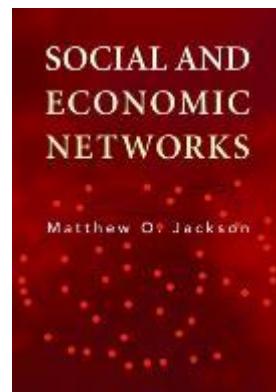
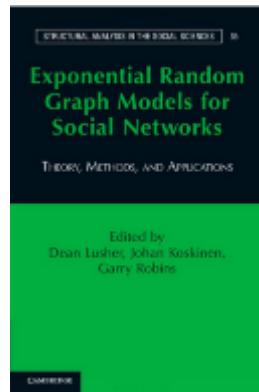
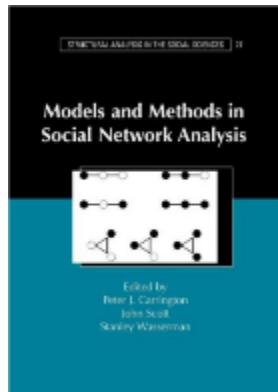
# Social Network Analysis

## General Resources



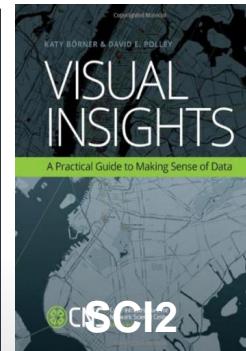
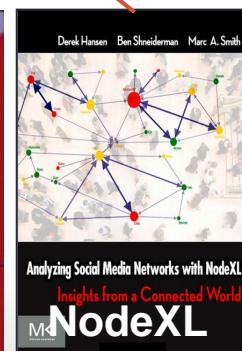
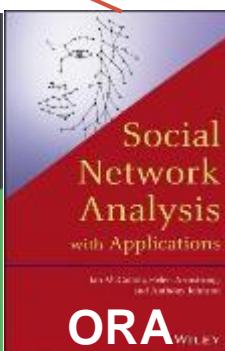
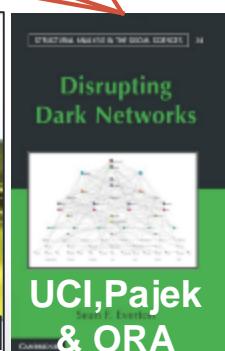
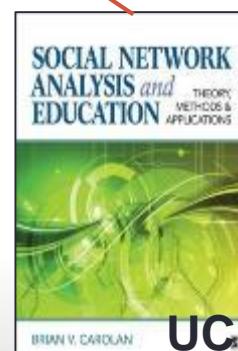
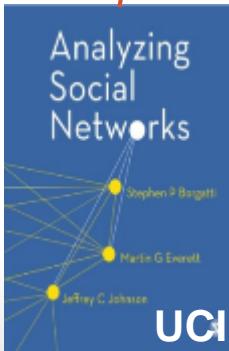
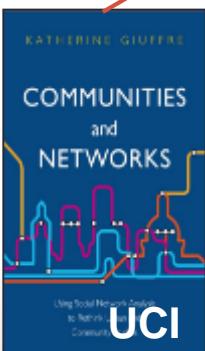
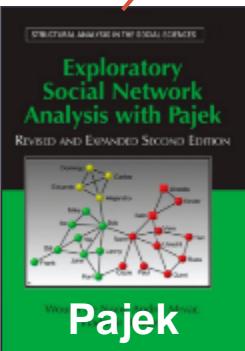
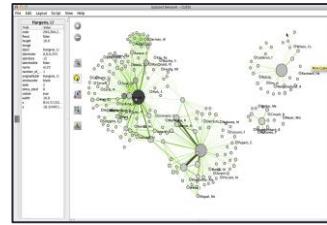
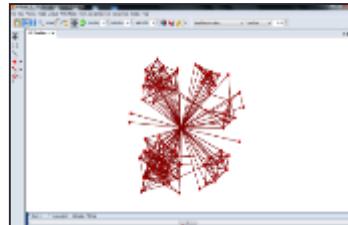
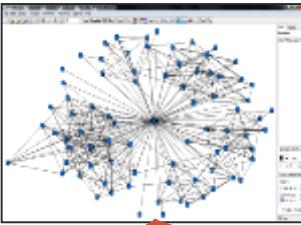
# Social Network Analysis

## General Resources with Packages



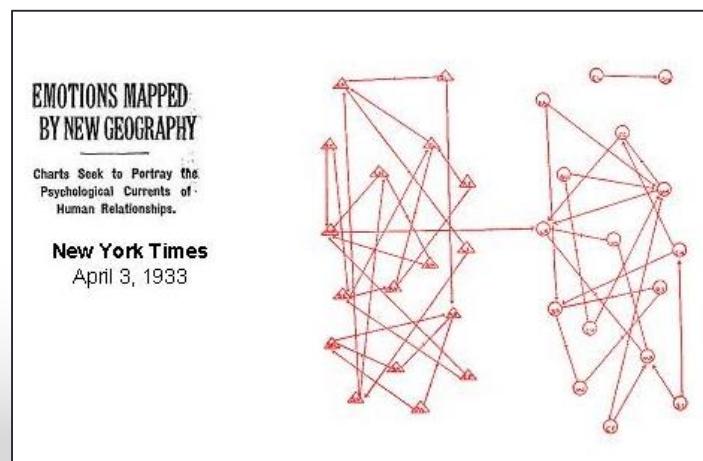
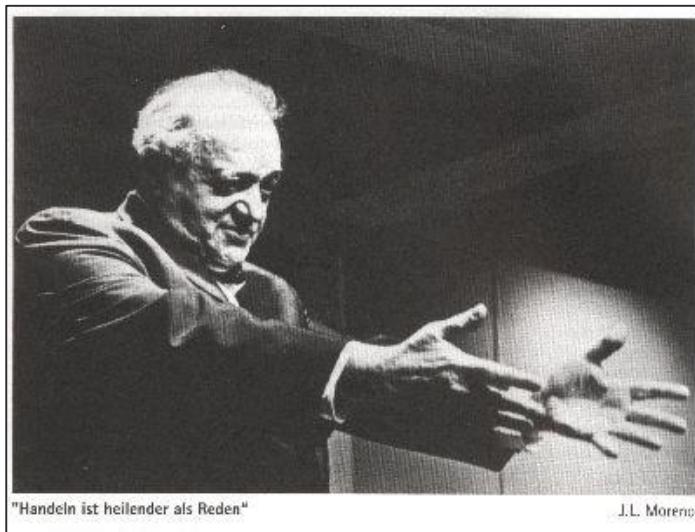
# Social Network Analysis

## *Social Network Packages*



# Social Network Analysis

*Long tie to visualization*



Decade	Scholar(s)	Innovations
1900	Simmel	Dyads and Triads
1930	Jacob Moreno	Sociometry, Sociograms
1930	Mayo & Warner	Hawthorne Study
1940	Forsyth & Katz	(Adjacency) Matrix
1940	Luce & Festinger	Matrix Algebra, n-cliques
1940	Bavelas	Centrality, Centralization
1950	Radcliff-Browne	Social Structure as a Network of Social Relations
1950	Harary & Norman	Graph Theory, Structural Balance
1950	Manchester School	Ego Networks
1950	Bott	Connectedness, Density
1950	Barnes	Social Network <sup>1</sup>
1950	Homans	Social Exchange
1960	James Davis	Clustering, Transitivity
1960	Coleman	Diffusion in Social Networks
1960	Milgram	Small world
1970	Blau	Homophily
1970	White	Block models, Vacancy Chains
1970	Granovetter	Weak ties
1980	Holland & Leinhardt	Exponential Random Graph Models
1980	Frank & Strauss	Markov dependency graphs
1990	Friedkin	Social Influence Network Theory
1990	Bonacich	Eigenvector centrality, Power centrality
1990	Putnam	Social capital
1990	Watts & Strogatz	Small world simulation
2000	Snijders & Huisman	Longitudinal network data

# Social Network Analysis

## *Key Elements*

**Vertices or Nodes**

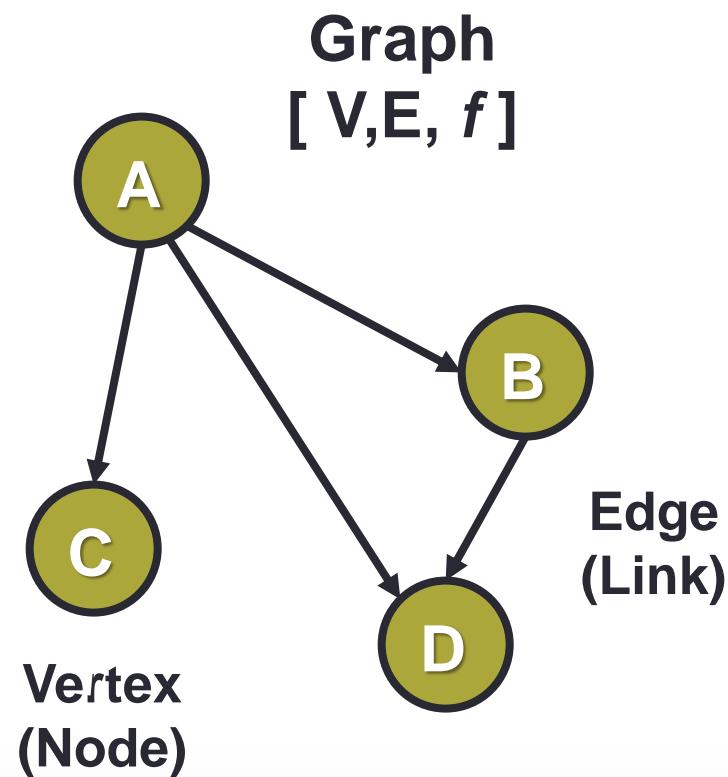
*The “things”*

**Edges or Links**

*The “relationships”*

**Graph or Network**

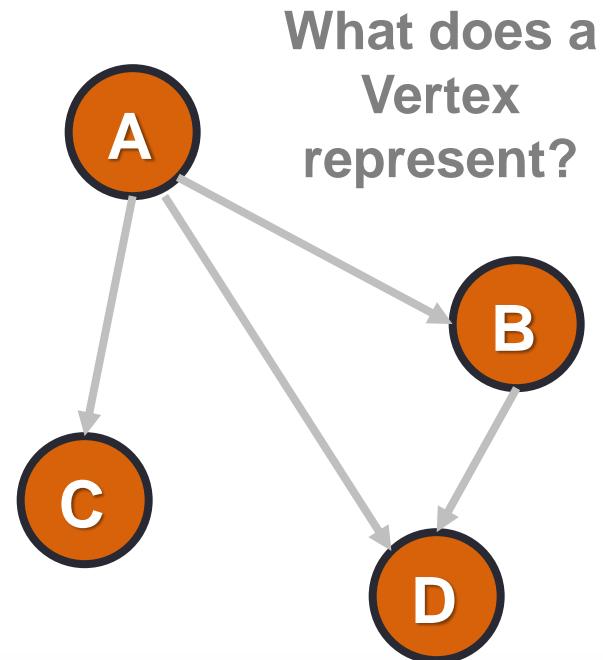
*The set of vertices/nodes, edges/links and the relationship/function connecting them.*



# Social Network Analysis

## *Types of Nodes or Vertices*

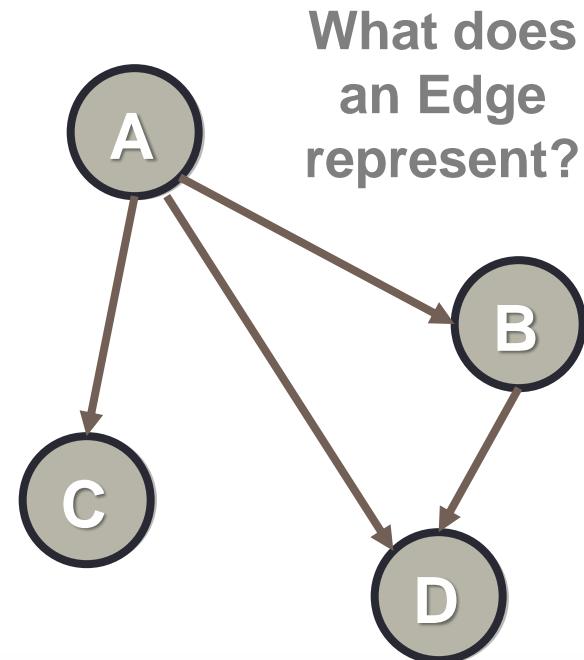
- Social Entities
  - People or social structures such as workgroups, teams, organizations, institutions, states, or even countries.
- Content
  - Web pages, keyword tags, or videos.
- Locations
  - Physical or virtual locations or events.
- Primary building blocks of social media
  - Friends in social networking sites, posts or authors in blogs, or pages in wikis.



# Social Network Analysis

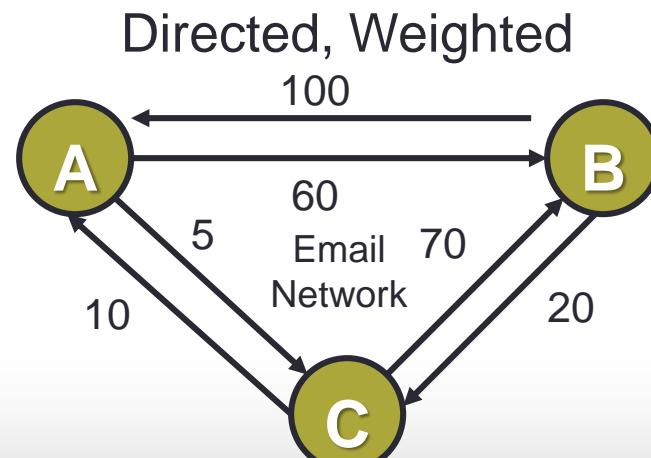
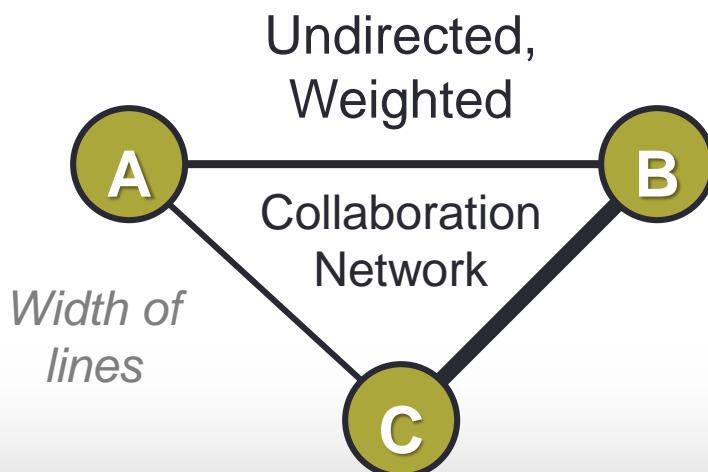
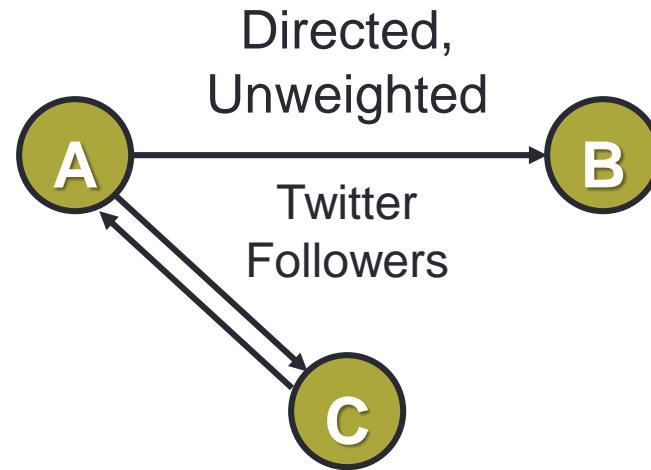
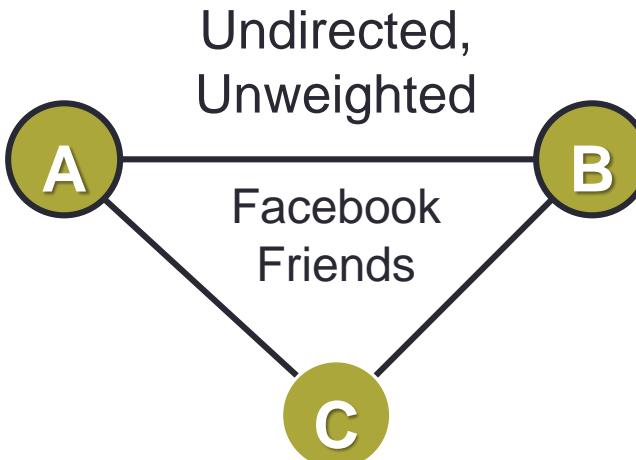
## *Types of Relations or Links*

- Similarities
  - Location (same spatial and temporal space)
  - Participation (same club, same event, ...)
  - Attributes (Age, gender, same attitudes, ...)
- Relational Roles
  - Kinship (mother of, sibling of, ...)
  - Other Roles (friend of, boss of, ...)
- Relational Cognition
  - Affective (Likes, Hates...)
  - Perceptual (Knows, Knows of, ...)
- Relational Events
  - Interactions (Sold to, talked to, helped, ...)
  - Flows (Information, beliefs, money, ...)



# Social Network Analysis

## *Types of Edges or Links*



# Social Network Analysis

## Bipartite or Bimodal Networks

Bipartite or Bimodal Network

Linking individuals to events  
(participation, membership, topic, tag,  
post, ...)

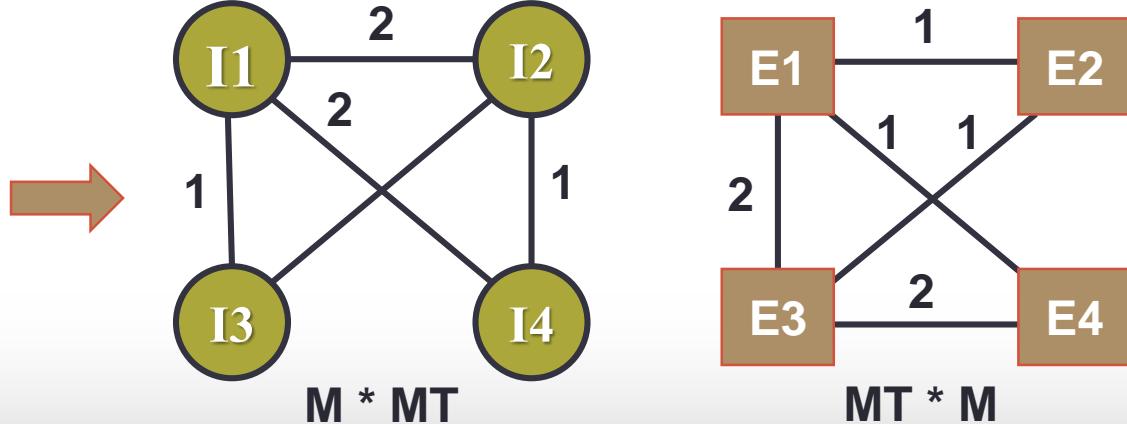
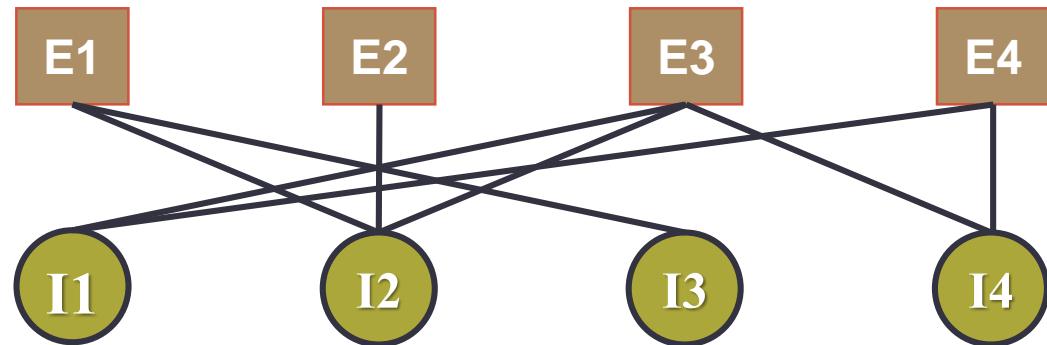
Examples:

Authors-to-papers (they authored)  
Actors-to-Movies (they appeared in)  
Users-to-Movies (they rated)

“Folded” networks:

Author collaboration networks  
Movie co-rating networks

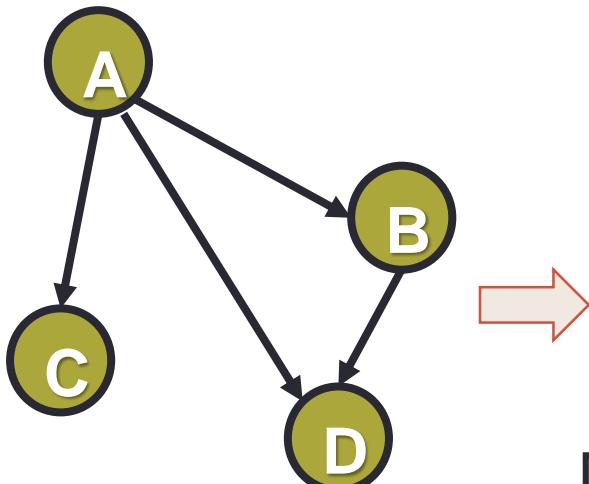
	E1	E2	E3	E4
I1	1	0	1	1
I2	1	1	1	0
I3	1	0	0	0
I4	0	0	1	1



# Social Network Analysis

## *Types of Alternative Representations*

Graph



Edge List

A	B
A	C
A	D
B	D

Adjacency Matrix

	A	B	C	D
A	-	1	1	1
B	0	-	0	1
C	0	0	-	0
D	0	0	0	-

Adjacency List

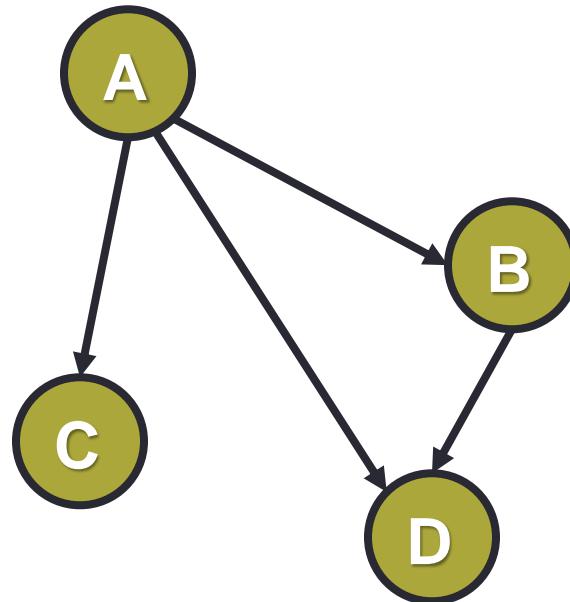
A	B, C, D
C	D

XML

<Node>	
<Label>	A </Label>
<Connection>	B </Connection>
<Connection>	C </Connection>
<Connection>	D </Connection>
</Node>	
<Node>	
<Label>	C </Label>
<Connection>	D </Connection>
</Node>	

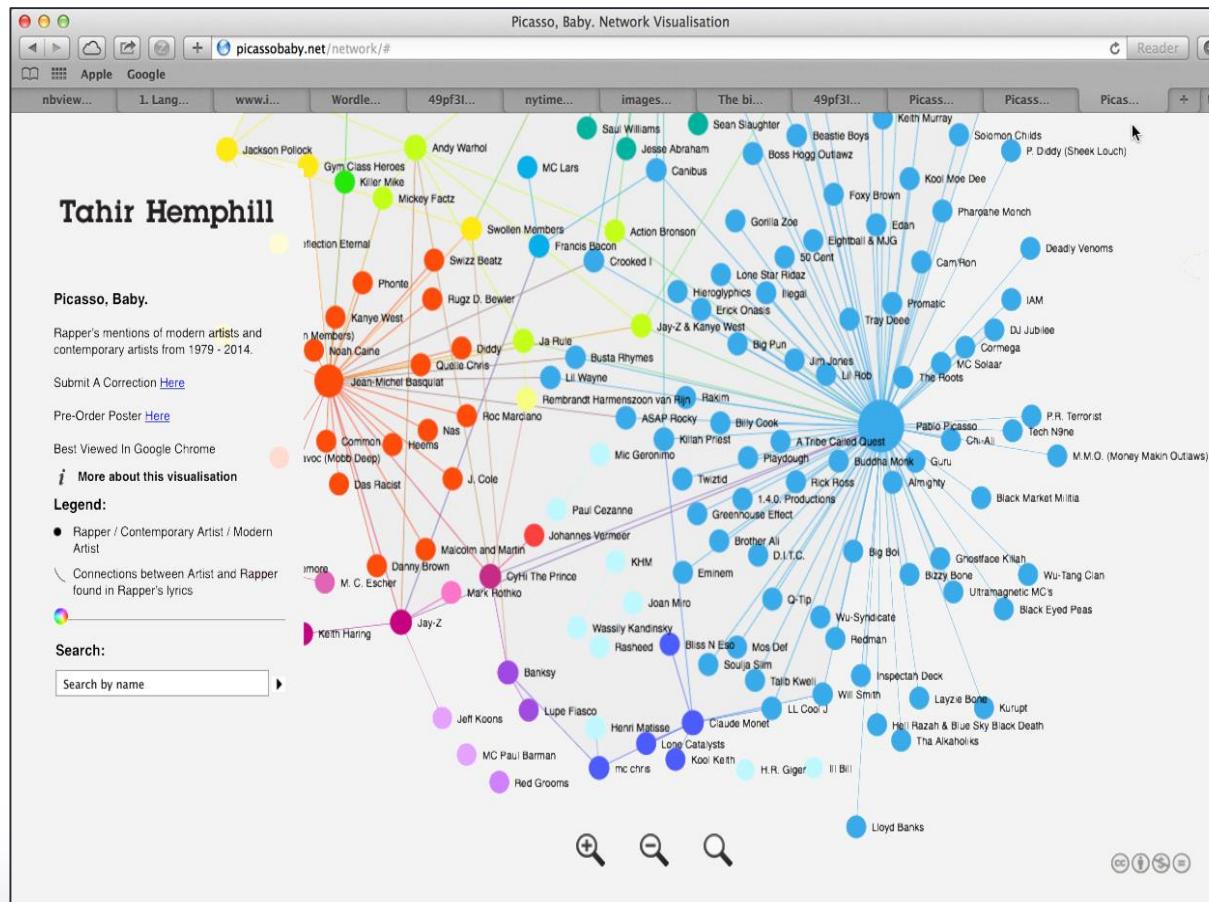
# Social Network Analysis

*How many variables can we represent?*



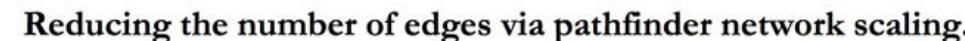
# Social Network Analysis

## *How many variables in this network?*



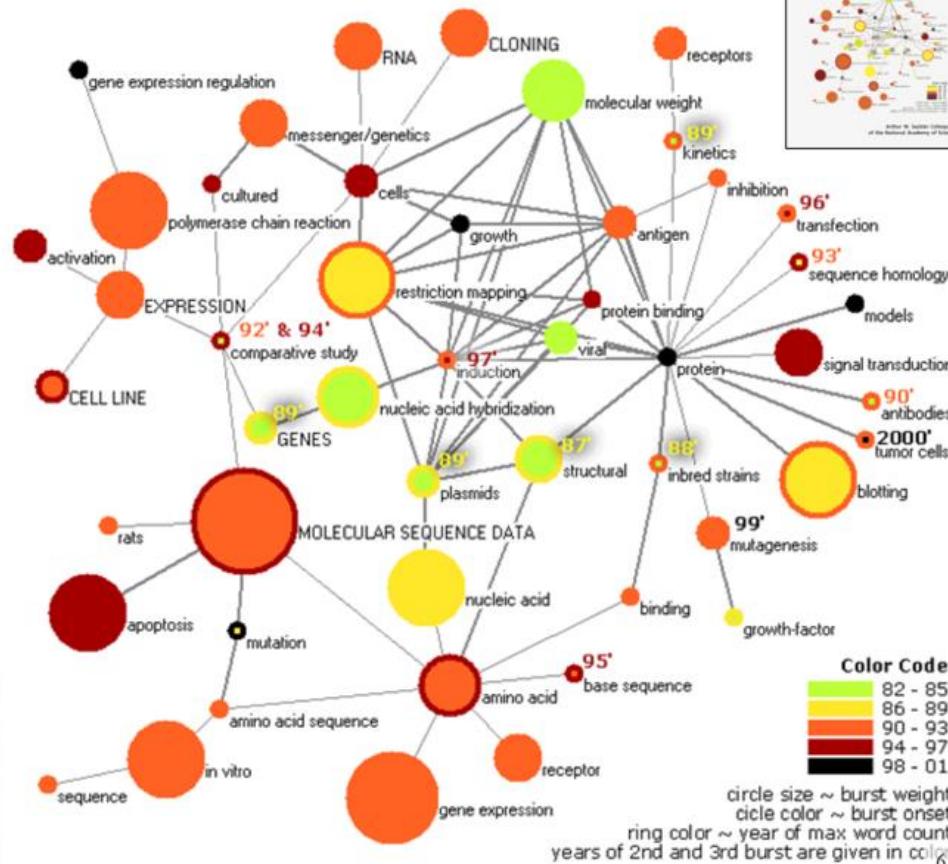
# Social Network Analysis

# *How many variables in this network?*



Co-word space of the top 50 highly frequent and bursty words used in the top 10% most highly cited PNAS publications in 1982-2001.

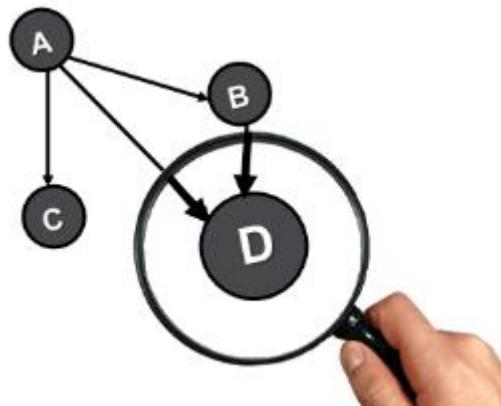
(Mane & Börner, 2004)



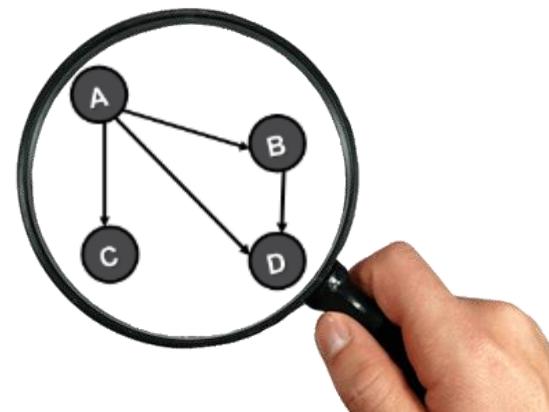
# Social Network Analysis

*Bifurcated measures*

Local  
Measures



Global  
Measures



# Social Network Analysis

## *Wide Variety of Metrics Available*



Measures Manager

Select Measures Set Measure Inputs Contains ▾

	Measure Title	Network Level	Node Level	Computation...	U...
<input type="checkbox"/>	Redundancy, Access	true	false	normal	fa
<input type="checkbox"/>	Actual Workload	false	true	normal	fz
<input type="checkbox"/>	Socio Economic Power, Agent	false	true	normal	fa
<input type="checkbox"/>	Redundancy, Assignment	true	false	normal	fa
<input type="checkbox"/>	Centrality, Authority	false	true	normal	tr
<input type="checkbox"/>	Characteristic Path Length	true	false	normal	tr
<input type="checkbox"/>	Speed, Average	true	false	normal	tr
<input checked="" type="checkbox"/>	Centrality, Betweenness	false	true	normal	tr
<input checked="" type="checkbox"/>	Network Centralization, Betweenness	true	false	normal	tr
<input type="checkbox"/>	Centrality, Bonacich Power	false	true	normal	tr
<input type="checkbox"/>	Capability	false	true	normal	tr
<input type="checkbox"/>	Clique Count	false	true	normal	fa
<input checked="" type="checkbox"/>	Centrality, Closeness	false	true	normal	tr
<input checked="" type="checkbox"/>	Network Centralization, Closeness	true	false	normal	tr
<input checked="" type="checkbox"/>	Density, Clustering Coefficient	true	true	normal	fa
<input type="checkbox"/>	Cognitive Demand	false	true	normal	tr
<input type="checkbox"/>	Cognitive Distinctiveness	false	true	normal	fa
<input type="checkbox"/>	Cognitive Expertise	false	true	normal	fa
<input type="checkbox"/>	Cognitive Resemblance	false	true	normal	fa
<input type="checkbox"/>	Cognitive Similarity	false	true	normal	fa
<input type="checkbox"/>	Breadth, Column	true	false	normal	fa
<input type="checkbox"/>	Count, Column	true	false	normal	fa
<input type="checkbox"/>	Centrality, Column Dearee	false	true	normal	tr

!!!

Select All  Select Visible

13 / 160 Selected, 160 / 160 Visible

OK Close

# Social Network Analysis

## *Who is most influential*

Measure	Definition	Interpretation	Reasoning
Degree	Number of edges or links. In degree- links in, Out-degree - links out	How connected is a node? How many people can this person reach directly?	Higher probability of receiving and transmitting information flows in the network. Nodes considered to have influence over larger number of nodes and are capable of communicating quickly with the nodes in their neighborhood.
Betweenness	Number of times node or vertex lies on shortest path between 2 nodes divided by number of all the shortest paths	How important is a node in terms of connecting other nodes? How likely is this person to be the most direct route between two people in the network?	Degree to which node controls flow of information in the network. Those with high betweenness function as brokers. Useful where a network is vulnerable.
Closeness	1 over the average distance between a node and every other node in the network	How easily can a node reach other nodes? How fast can this person reach everyone in the network?	Measure of reach. Importance based on how close a node is located with respect to every other node in the network. Nodes able to reach most or be reached by most all other nodes in the network through geodesic paths.
Eigenvector	Proportional to the sum of the eigenvector centralities of all the nodes directly connected to it.	How important, central, or influential are a node's neighbors? How well is this person connected to other well-connected people?	Evaluates a player's popularity. Identifies centers of large cliques. Node with more connections to higher scoring nodes is more important.

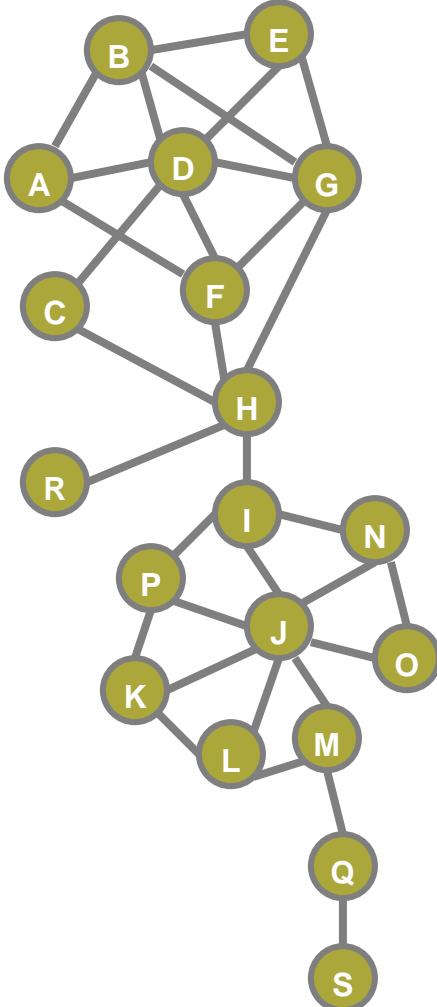
# Social Network Analysis

## *How well connected is the network?*

Cohesion	Definition	Interpretation	Reasoning
Density	Ratio of the number of edges in the network over the total number of possible edges between all pairs of nodes	How well connected is the overall network?	Perfectly connected network is called a "clique" and has a density of 1.
Clustering	A node's <i>clustering coefficient</i> is the density of its 1.5 degree egocentric network (ratio of connecting among ego's alters). For entire network it is the average of all the coefficients for the individual nodes.	What proportion of ego's alters are connected? More technically, how many nodes form triangular subgraphs with their adjacent nodes?	Measures certain aspects of "cliquishness." Proportion of your friends that are also friends with each other. Another way to measure is to determine (in a undirected) graph the ratio of the number of times that two links emanating from the same node are also linked.
Average Path Length (Distance)	Average number of edges or links between any two nodes (along the shortest path)	On average, how far apart are any two nodes?	This is synonymous with the "degrees of separation" in a network.
Diameter	Longest (shortest path) between any two nodes	At most, how long will it take to reach any node in the network? Sparse networks usually have greater diameters.	Measure of the reach of the network
Centralization	Normalize ratio of the sum of the variances of the centrality of each node from the most central node to the maximum sum possible	Indicates how unequal the distribution of centrality is in a network.	Measures how much variance there is in the distribution of centrality in a network. The measure applies to all forms of centrality.

# Social Network Analysis

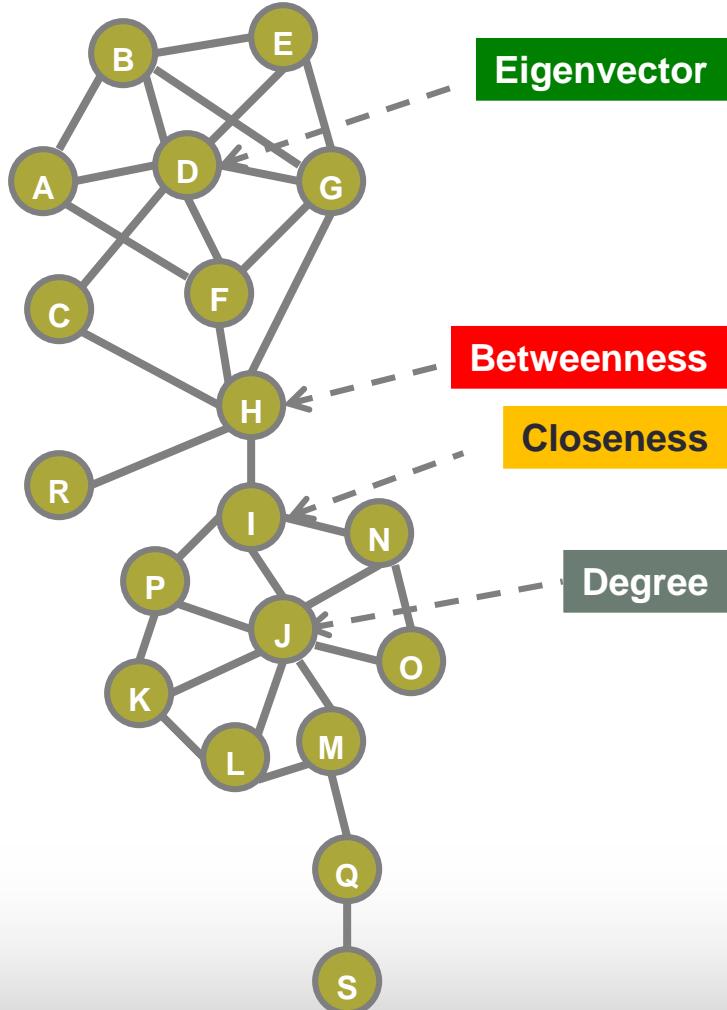
## *Generic Example*



- For each of the nodes what is it's
  - Degree Centrality
  - Betweenness Centrality
  - Closeness Centrality
  - Eigenvector Centrality
- For the entire network what is it's
  - Degree Centralization
  - Betweenness Centralization
  - Closeness Centralization
  - Eigenvector Centralization

# Social Network Analysis

*Who's most influential?*

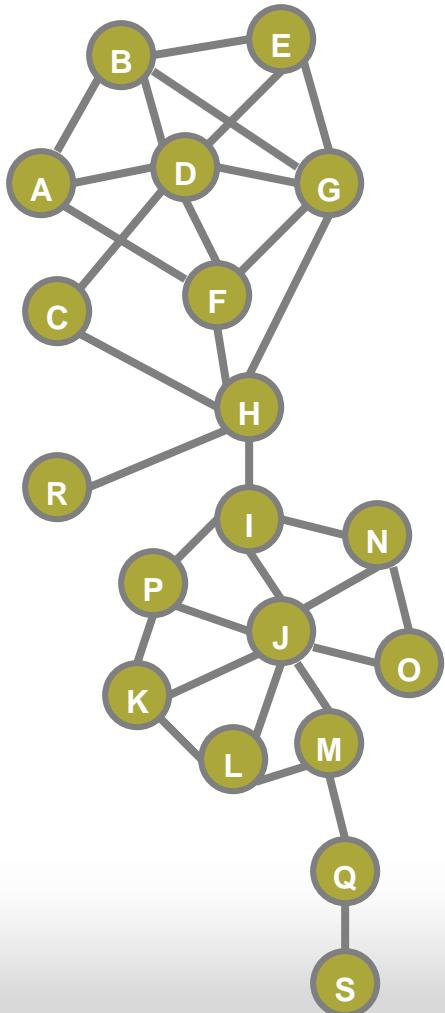


Node	Degree	Normed Degree	Betweenness	Closeness	Eigen Vector
A	3	0.17	0.00	0.29	0.29
B	4	0.22	0.01	0.30	0.36
C	2	0.11	0.03	0.35	0.18
D	6	0.33	0.04	0.31	0.46
E	3	0.17	0.00	0.29	0.30
F	4	0.22	0.11	0.36	0.35
G	5	0.28	0.19	0.37	0.43
H	5	0.28	0.58	0.45	0.28
I	4	0.22	0.53	0.46	0.13
J	7	0.39	0.43	0.43	0.12
K	3	0.17	0.00	0.32	0.06
L	3	0.17	0.01	0.33	0.05
M	3	0.17	0.21	0.33	0.04
N	3	0.17	0.03	0.38	0.07
O	2	0.11	0.00	0.31	0.05
P	3	0.17	0.03	0.38	0.08
Q	2	0.11	0.11	0.26	0.01
R	1	0.06	0.00	0.32	0.07
S	1	0.06	0.00	0.21	0.00

Correlations	Degree	Betweenness	Closeness	Eigenvector
Degree	-	0.57	0.59	0.59
Betweenness		-	0.79	0.07
Closeness			-	0.13
Eigenvector				-

# Social Network Analysis

## How cohesive is the network?



Measure	Value
Network Size	19
Average Degree	3.37
Degree Centralization	0.22
Betweenness Centralization	0.48
Closeness Centralization	0.27
Eigenvector Centralization	0.56
Clustering Coefficient	0.43
Density	0.19
Average Distance	3.06
Diameter	8
Number of Unreachable Nodes	0

Node	Clustering
A	0.67
B	0.67
C	0.00
D	0.40
E	1.00
F	0.50
G	0.50
H	0.10
I	0.33
J	0.29
K	0.67
L	0.67
M	0.33
N	0.67
O	1.00
P	0.67
Q	0.00
R	NA
S	NA

6

Frigyes Karinthy  
1929



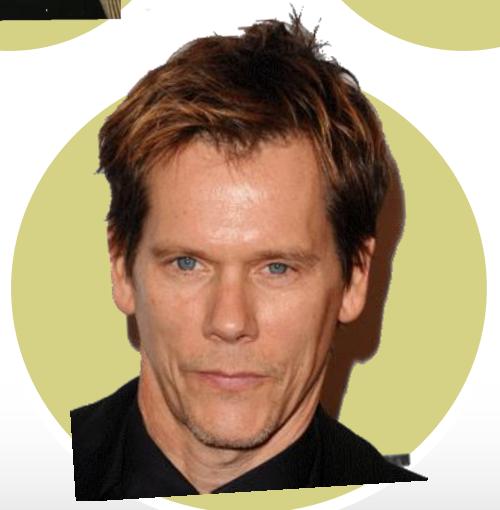
John Guare  
1990



Stanley Milgram  
1967



Duncan Watts  
1998



**Six Degrees of Kevin Bacon (1994)**

# Social Network Analysis

## *Who is the center of Hollywood?*

### The Center of the Hollywood Universe

Click on a name to see that person's table.

1. [Harvey Keitel](#) (2.848635)
2. [Dennis Hopper](#) (2.849329)
3. [Robert De Niro](#) (2.855810)
4. [David Carradine](#) (2.857729)
5. [Martin Sheen](#) (2.858291)
6. [Udo Kier](#) (2.859489)
7. [Michael Madsen](#) (I) (2.860010)
8. [Donald Sutherland](#) (I) (2.860447)
9. [Michael Caine](#) (I) (2.862189)
10. [Eric Roberts](#) (I) (2.867675)
11. [Seymour Cassel](#) (2.869415)
12. [Malcolm McDowell](#) (2.870208)
13. [Max von Sydow](#) (I) (2.872338)
14. [Willem Dafoe](#) (2.873805)
15. [Samuel L. Jackson](#) (2.873819)
16. [Danny Trejo](#) (2.876002)
17. [John Hurt](#) (2.878378)
18. [Christopher Lee](#) (I) (2.879217)
19. [Harry Dean Stanton](#) (2.880725)
20. [Bruce Willis](#) (2.886364)
21. [Christopher Plummer](#) (I) (2.886928)
22. [John Malkovich](#) (2.888575)
23. [Morgan Freeman](#) (I) (2.891003)
24. [Christopher Walken](#) (2.894212)
25. [John Savage](#) (I) (2.894873)

Kevin Bacon Number	# of People
0	1
1	2799
2	313045
3	1078865
4	276680
5	22296
6	2361
7	251
8	24

Total number of linkable actors: 1696322  
Weighted total of linkable actors: 5099799  
Average Kevin Bacon number: 3.006

Kyra Sedgwick Number	# of People
0	1
1	1353
2	229226
3	1083255
4	350117
5	29167
6	2845
7	331
8	27

Total number of linkable actors: 1696322  
Weighted total of linkable actors: 5275476  
Average Kyra Sedgwick number: 3.110

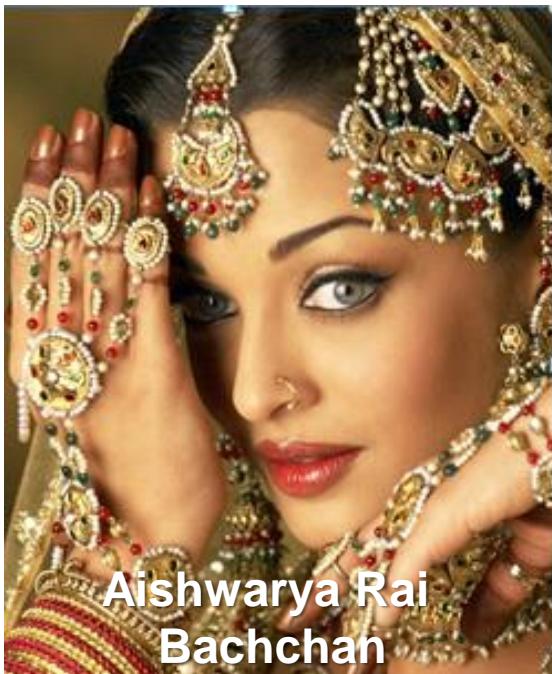
Harvey Keitel (I) Number	# of People
0	1
1	4128
2	454260
3	1051685
4	169704
5	14709
6	1679
7	141
8	15

Total number of linkable actors: 1696322  
Weighted total of linkable actors: 4831245  
Average Harvey Keitel (I) number: 2.848

John Savage (I) Number	# of People
0	1
1	3766
2	408573
3	1073294
4	192718
5	16000
6	1764
7	195
8	11

Total number of linkable actors: 1696322  
Weighted total of linkable actors: 4903703  
Average John Savage (I) number: 2.891

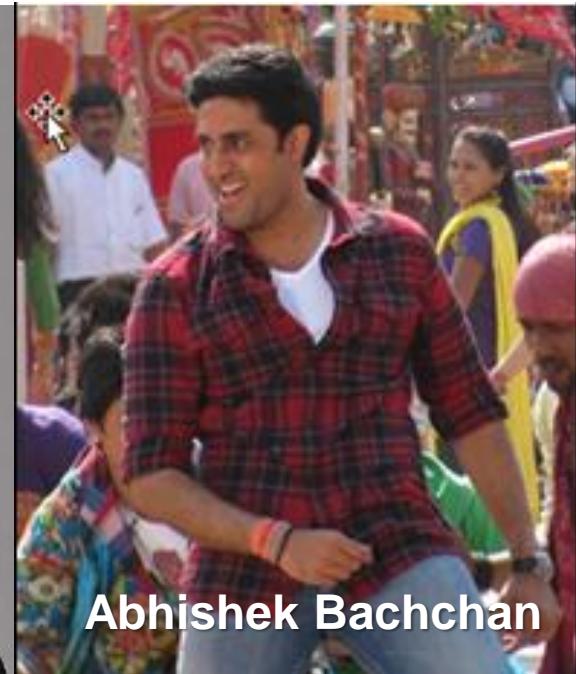
# Social Network Analysis Bollywood?



Aishwarya Rai  
Bachchan



Amitabh Bachchan



Abhishek Bachchan

जुदाई की छह डिग्री

judā'ī kī chaha digrī

# Social Network Analysis

## Source of Bollywood Data

Amitabh Bachchan - Google

List of Bollywood films - Wikipedia

en.wikipedia.org/wiki/List\_of\_Bollywood\_films#2010s

**List of Bollywood films**

From Wikipedia, the free encyclopedia

This article does not cite any references or sources. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. (September 2012)

This is a list of films produced by the Bollywood film industry of Mumbai ordered by year and decade of release and also contains the top ten or forty superhit films of respective years as the case may be. Although "Bollywood" films are generally listed under the Hindi language, most are in mixed Hindi, Urdu and Punjabi and occasionally other languages. There is a range of mixtures from mostly Urdu to mostly Hindi to mostly Punjabi. Speakers of Hindi, Urdu, and Punjabi understand the mixed language usage of Bollywood thus extending the viewership to people all over the Indian subcontinent (throughout India and its neighboring countries). Here are some examples - Partly Hindi: *Om Shanti Om*, *Dhoom 2*, *No Entry* and *Kabhi Alvida Naa Kehna*, Partly Urdu: *Jodhaa Akbar*, *Fanaa*, *Saawariya* and *Kurbaan*, Partly Punjabi: *Singh Is Kinng*, *Jab We Met*, *Patiala House* and *Rab Ne Bana Di Jodi*. The film *Veer Zaara* is an equal mix of Hindi, Punjabi and Urdu.

Contents [hide]

- 1 2010s
- 2 2000s
- 3 1990s
- 4 1980s
- 5 1970s
- 6 1960s
- 7 1950s
- 8 1940s
- 9 1930s

**2010s [edit]**

- List of Bollywood films of 2010
- List of Bollywood films of 2011
- List of Bollywood films of 2012
- List of Bollywood films of 2013
- List of Bollywood films of 2014

**2000s [edit]**

- List of Bollywood films of 2000

*Alam Ara* (1931), the first Indian sound film



WIKIPEDIA  
The Free Encyclopedia

Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia  
Wikimedia Shop

Interaction  
Help  
About Wikipedia  
Community portal  
Recent changes  
Contact page

Tools  
Print/export

Languages  
हिन्दी  
Português  
ଓଡ଼ିଆ

Edit links

# Social Network Analysis

## *Source of Bollywood Data*

# Wikipedia Web Page

```
HTML- Page Source
```

# HTML- Page Source

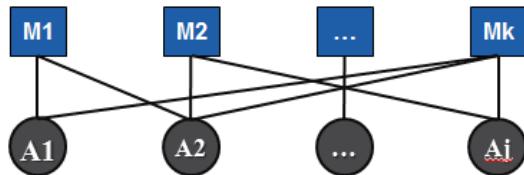
ID	Title	Subtitle	Year	Director	Cast	
1	Movies 1	SubMovie 1	2010	Director 1	Actor 1	
2	Movies 2	SubMovie 2	2011	Director 2	Actor 2	
3	Movies 3	SubMovie 3	2012	Director 3	Actor 3	
4	Movies 4	SubMovie 4	2013	Director 4	Actor 4	
5	Movies 5	SubMovie 5	2014	Director 5	Actor 5	
6	Movies 6	SubMovie 6	2015	Director 6	Actor 6	
7	Movies 7	SubMovie 7	2016	Director 7	Actor 7	
8	Movies 8	SubMovie 8	2017	Director 8	Actor 8	
9	Movies 9	SubMovie 9	2018	Director 9	Actor 9	
10	Movies 10	SubMovie 10	2019	Director 10	Actor 10	
11	Movies 11	SubMovie 11	2020	Director 11	Actor 11	
12	Movies 12	SubMovie 12	2021	Director 12	Actor 12	
13	Movies 13	SubMovie 13	2022	Director 13	Actor 13	
14	Movies 14	SubMovie 14	2023	Director 14	Actor 14	
15	Movies 15	SubMovie 15	2024	Director 15	Actor 15	
16	Movies 16	SubMovie 16	2025	Director 16	Actor 16	
17	Movies 17	SubMovie 17	2026	Director 17	Actor 17	
18	Movies 18	SubMovie 18	2027	Director 18	Actor 18	
19	Movies 19	SubMovie 19	2028	Director 19	Actor 19	
20	Movies 20	SubMovie 20	2029	Director 20	Actor 20	
21	Movies 21	SubMovie 21	2030	Director 21	Actor 21	
22	Movies 22	SubMovie 22	2031	Director 22	Actor 22	
23	Movies 23	SubMovie 23	2032	Director 23	Actor 23	
24	Movies 24	SubMovie 24	2033	Director 24	Actor 24	
25	Movies 25	SubMovie 25	2034	Director 25	Actor 25	
26	Movies 26	SubMovie 26	2035	Director 26	Actor 26	
27	Movies 27	SubMovie 27	2036	Director 27	Actor 27	
28	Movies 28	SubMovie 28	2037	Director 28	Actor 28	
29	Movies 29	SubMovie 29	2038	Director 29	Actor 29	
30	Movies 30	SubMovie 30	2039	Director 30	Actor 30	
31	Movies 31	SubMovie 31	2040	Director 31	Actor 31	
32	Movies 32	SubMovie 32	2041	Director 32	Actor 32	
33	Movies 33	SubMovie 33	2042	Director 33	Actor 33	
34	Movies 34	SubMovie 34	2043	Director 34	Actor 34	
35	Movies 35	SubMovie 35	2044	Director 35	Actor 35	
36	Movies 36	SubMovie 36	2045	Director 36	Actor 36	
37	Movies 37	SubMovie 37	2046	Director 37	Actor 37	
38	Movies 38	SubMovie 38	2047	Director 38	Actor 38	
39	Movies 39	SubMovie 39	2048	Director 39	Actor 39	
40	Movies 40	SubMovie 40	2049	Director 40	Actor 40	
41	Movies 41	SubMovie 41	2050	Director 41	Actor 41	
42	Movies 42	SubMovie 42	2051	Director 42	Actor 42	
43	Movies 43	SubMovie 43	2052	Director 43	Actor 43	
44	Movies 44	SubMovie 44	2053	Director 44	Actor 44	
45	Movies 45	SubMovie 45	2054	Director 45	Actor 45	
46	Movies 46	SubMovie 46	2055	Director 46	Actor 46	
47	Movies 47	SubMovie 47	2056	Director 47	Actor 47	
48	Movies 48	SubMovie 48	2057	Director 48	Actor 48	
49	Movies 49	SubMovie 49	2058	Director 49	Actor 49	
50	Movies 50	SubMovie 50	2059	Director 50	Actor 50	
51	Movies 51	SubMovie 51	2060	Director 51	Actor 51	
52	Movies 52	SubMovie 52	2061	Director 52	Actor 52	
53	Movies 53	SubMovie 53	2062	Director 53	Actor 53	
54	Movies 54	SubMovie 54	2063	Director 54	Actor 54	
55	Movies 55	SubMovie 55	2064	Director 55	Actor 55	
56	Movies 56	SubMovie 56	2065	Director 56	Actor 56	
57	Movies 57	SubMovie 57	2066	Director 57	Actor 57	
58	Movies 58	SubMovie 58	2067	Director 58	Actor 58	
59	Movies 59	SubMovie 59	2068	Director 59	Actor 59	
60	Movies 60	SubMovie 60	2069	Director 60	Actor 60	
61	Movies 61	SubMovie 61	2070	Director 61	Actor 61	
62	Movies 62	SubMovie 62	2071	Director 62	Actor 62	
63	Movies 63	SubMovie 63	2072	Director 63	Actor 63	
64	Movies 64	SubMovie 64	2073	Director 64	Actor 64	
65	Movies 65	SubMovie 65	2074	Director 65	Actor 65	
66	Movies 66	SubMovie 66	2075	Director 66	Actor 66	
67	Movies 67	SubMovie 67	2076	Director 67	Actor 67	
68	Movies 68	SubMovie 68	2077	Director 68	Actor 68	
69	Movies 69	SubMovie 69	2078	Director 69	Actor 69	
70	Movies 70	SubMovie 70	2079	Director 70	Actor 70	
71	Movies 71	SubMovie 71	2080	Director 71	Actor 71	
72	Movies 72	SubMovie 72	2081	Director 72	Actor 72	
73	Movies 73	SubMovie 73	2082	Director 73	Actor 73	
74	Movies 74	SubMovie 74	2083	Director 74	Actor 74	
75	Movies 75	SubMovie 75	2084	Director 75	Actor 75	
76	Movies 76	SubMovie 76	2085	Director 76	Actor 76	
77	Movies 77	SubMovie 77	2086	Director 77	Actor 77	
78	Movies 78	SubMovie 78	2087	Director 78	Actor 78	
79	Movies 79	SubMovie 79	2088	Director 79	Actor 79	
80	Movies 80	SubMovie 80	2089	Director 80	Actor 80	
81	Movies 81	SubMovie 81	2090	Director 81	Actor 81	
82	Movies 82	SubMovie 82	2091	Director 82	Actor 82	
83	Movies 83	SubMovie 83	2092	Director 83	Actor 83	
84	Movies 84	SubMovie 84	2093	Director 84	Actor 84	
85	Movies 85	SubMovie 85	2094	Director 85	Actor 85	
86	Movies 86	SubMovie 86	2095	Director 86	Actor 86	
87	Movies 87	SubMovie 87	2096	Director 87	Actor 87	
88	Movies 88	SubMovie 88	2097	Director 88	Actor 88	
89	Movies 89	SubMovie 89	2098	Director 89	Actor 89	
90	Movies 90	SubMovie 90	2099	Director 90	Actor 90	
91	Movies 91	SubMovie 91	2100	Director 91	Actor 91	
92	Movies 92	SubMovie 92	2101	Director 92	Actor 92	
93	Movies 93	SubMovie 93	2102	Director 93	Actor 93	
94	Movies 94	SubMovie 94	2103	Director 94	Actor 94	
95	Movies 95	SubMovie 95	2104	Director 95	Actor 95	
96	Movies 96	SubMovie 96	2105	Director 96	Actor 96	
97	Movies 97	SubMovie 97	2106	Director 97	Actor 97	
98	Movies 98	SubMovie 98	2107	Director 98	Actor 98	
99	Movies 99	SubMovie 99	2108	Director 99	Actor 99	
100	Movies 100	SubMovie 100	2109	Director 100	Actor 100	
101	Movies 101	SubMovie 101	2110	Director 101	Actor 101	
102	Movies 102	SubMovie 102	2111	Director 102	Actor 102	
103	Movies 103	SubMovie 103	2112	Director 103	Actor 103	
104	Movies 104	SubMovie 104	2113	Director 104	Actor 104	
105	Movies 105	SubMovie 105	2114	Director 105	Actor 105	
106	Movies 106	SubMovie 106	2115	Director 106	Actor 106	
107	Movies 107	SubMovie 107	2116	Director 107	Actor 107	
108	Movies 108	SubMovie 108	2117	Director 108	Actor 108	
109	Movies 109	SubMovie 109	2118	Director 109	Actor 109	
110	Movies 110	SubMovie 110	2119	Director 110	Actor 110	
111	Movies 111	SubMovie 111	2120	Director 111	Actor 111	
112	Movies 112	SubMovie 112	2121	Director 112	Actor 112	
113	Movies 113	SubMovie 113	2122	Director 113	Actor 113	
114	Movies 114	SubMovie 114	2123	Director 114	Actor 114	
115	Movies 115	SubMovie 115	2124	Director 115	Actor 115	
116	Movies 116	SubMovie 116	2125	Director 116	Actor 116	
117	Movies 117	SubMovie 117	2126	Director 117	Actor 117	
118	Movies 118	SubMovie 118	2127	Director 118	Actor 118	
119	Movies 119	SubMovie 119	2128	Director 119	Actor 119	
120	Movies 120	SubMovie 120	2129	Director 120	Actor 120	
121	Movies 121	SubMovie 121	2130	Director 121	Actor 121	
122	Movies 122	SubMovie 122	2131	Director 122	Actor 122	
123	Movies 123	SubMovie 123	2132	Director 123	Actor 123	
124	Movies 124	SubMovie 124	2133	Director 124	Actor 124	
125	Movies 125	SubMovie 125	2134	Director 125	Actor 125	
126	Movies 126	SubMovie 126	2135	Director 126	Actor 126	
127	Movies 127	SubMovie 127	2136	Director 127	Actor 127	
128	Movies 128	SubMovie 128	2137	Director 128	Actor 128	
129	Movies 129	SubMovie 129	2138	Director 129	Actor 129	
130	Movies 130	SubMovie 130	2139	Director 130	Actor 130	
131	Movies 131	SubMovie 131	2140	Director 131	Actor 131	
132	Movies 132	SubMovie 132	2141	Director 132	Actor 132	
133	Movies 133	SubMovie 133	2142	Director 133	Actor 133	
134	Movies 134	SubMovie 134	2143	Director 134	Actor 134	
135	Movies 135	SubMovie 135	2144	Director 135	Actor 135	
136	Movies 136	SubMovie 136	2145	Director 136	Actor 136	
137	Movies 137	SubMovie 137	2146	Director 137	Actor 137	
138	Movies 138	SubMovie 138	2147	Director 138	Actor 138	
139	Movies 139	SubMovie 139	2148	Director 139	Actor 139	
140	Movies 140	SubMovie 140	2149	Director 140	Actor 140	
141	Movies 141	SubMovie 141	2150	Director 141	Actor 141	
142	Movies 142	SubMovie 142	2151	Director 142	Actor 142	
143	Movies 143	SubMovie 143	2152	Director 143	Actor 143	
144	Movies 144	SubMovie 144	2153	Director 144	Actor 144	
145	Movies 145	SubMovie 145	2154	Director 145	Actor 145	
146	Movies 146	SubMovie 146	2155	Director 146	Actor 146	
147	Movies 147	SubMovie 147	2156	Director 147	Actor 147	
148	Movies 148	SubMovie 148	2157	Director 148	Actor 148	
149	Movies 149	SubMovie 149	2158	Director 149	Actor 149	
150	Movies 150	SubMovie 150	2159	Director 150	Actor 150	
151	Movies 151	SubMovie 151	2160	Director 151	Actor 151	
152	Movies 152	SubMovie 152	2161	Director 152	Actor 152	
153	Movies 153	SubMovie 153	2162	Director 153	Actor 153	
154	Movies 154	SubMovie 154	2163	Director 154	Actor 154	
155	Movies 155	SubMovie 155	2164	Director 155	Actor 155	
156	Movies 156	SubMovie 156	2165	Director 156	Actor 156	
157	Movies 157	SubMovie 157	2166	Director 157	Actor 157	
158	Movies 158	SubMovie 158	2167	Director 158	Actor 158	
159	Movies 159	SubMovie 159	2168	Director 159	Actor 159	
160	Movies 160	SubMovie 160	2169	Director 160	Actor 160	
161	Movies 161	SubMovie 161	2170	Director 161	Actor 161	
162	Movies 162	SubMovie 162	2171	Director 162	Actor 162	
163	Movies 163	SubMovie 163	2172	Director 163	Actor 163	
164	Movies 164	SubMovie 164	2173	Director 164	Actor 164	
165	Movies 165	SubMovie 165	2174	Director 165	Actor 165	
166	Movies 166	SubMovie 166	2175	Director 166	Actor 166	
167	Movies 167	SubMovie 167	2176	Director 167	Actor 167	
168	Movies 168	SubMovie 168	2177	Director 168	Actor 168	
169	Movies 169	SubMovie 169	2178	Director 169	Actor 169	
170	Movies 170	SubMovie 170	2179	Director 170	Actor 170	
171	Movies 171	SubMovie 171	2180	Director 171	Actor 171	
172	Movies 172	SubMovie 172	2181	Director 172	Actor 172	
173	Movies 173	SubMovie 173	2182	Director 173	Actor 173	
174	Movies 174	SubMovie 174	2183	Director 174	Actor 174	
175	Movies 175	SubMovie 175	2184	Director 175	Actor 175	
176	Movies 176	SubMovie 176	2185	Director 176	Actor 176	
177	Movies 177	SubMovie 177	2186	Director 177	Actor 177	
178	Movies 178	SubMovie 178	2187	Director 178	Actor 178	
179	Movies 179	SubMovie 179	2188	Director 179	Actor 179	
180	Movies 180	SubMovie 180	2189	Director 180	Actor 180	
181	Movies 181	SubMovie 181	2190	Director 181	Actor 181	
182	Movies 182	SubMovie 182	2191	Director 182	Actor 182	
183	Movies 183	SubMovie 183	2192	Director 183	Actor 183	
184	Movies 184	SubMovie 184	2193	Director 184	Actor 184	
185	Movies 185	SubMovie 185	2194	Director 185	Actor 185	
186	Movies 186	SubMovie 186	2195	Director 186	Actor 186	
187	Movies 187	SubMovie 187	2196	Director 187	Actor 187	
188	Movies 188	SubMovie 188	2197	Director 188	Actor 188	
189	Movies 189	SubMovie 189	2198	Director 189</		

Relational DB – 627 Movies – 1061 Actors (2010-2013)

```
*Network bollywood.net [2-Mode]
*Vertices 1643 627
1 "Mumbai Mirror" 0.0000 0.0000 0.5000
2 "Vishwaroop" 0.0000 0.0000 0.5000
...
628 "A. K. Hangal" 0.0000 0.0000 0.5000
629 "Aamir Ali" 0.0000 0.0000 0.5000
...
*Arcs
*Edges
1 896 1
1 1220 1
...
627 856 1
627 1053 1
```

# Social Network Data

## Bipartite Network – Movies and Actors



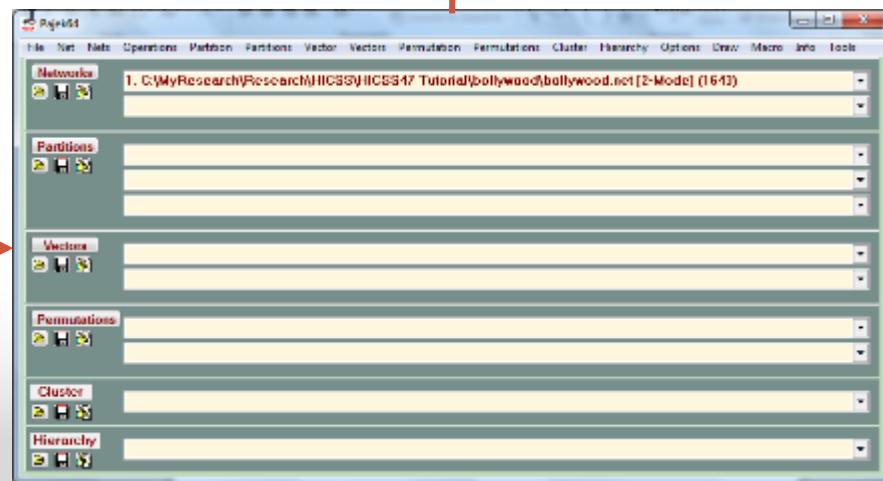
	M1	M2	...	Mk
A1	1	0		1
A2	1	1	...	1
...	...	...	...	...
Aj	0	1		1

	A1	A2	...	Aj
A1	-	2	...	1
A2	2	-	...	1
...	...	...	...	...
Aj	1	1	...	-

	M1	M2	...	Mk
M1	-	1	...	2
M2	1	-	...	2
...	...	...	...	...
Mk	2	2	...	-

```
*Network bollywood.net [2-Mode]
*Vertices 1643 627
1 "Mumbai Mirror" 0.0000 0.0000 0.5000
2 "Vishwaroop" 0.0000 0.0000 0.5000
...
628 "A. K. Hangal" 0.0000 0.0000 0.5000
629 "Aamir Ali" 0.0000 0.0000 0.5000
...
*Arcs
*Edges
1 896 1
1 1220 1
...
627 856 1
627 1053 1
```

Pajek .NET File

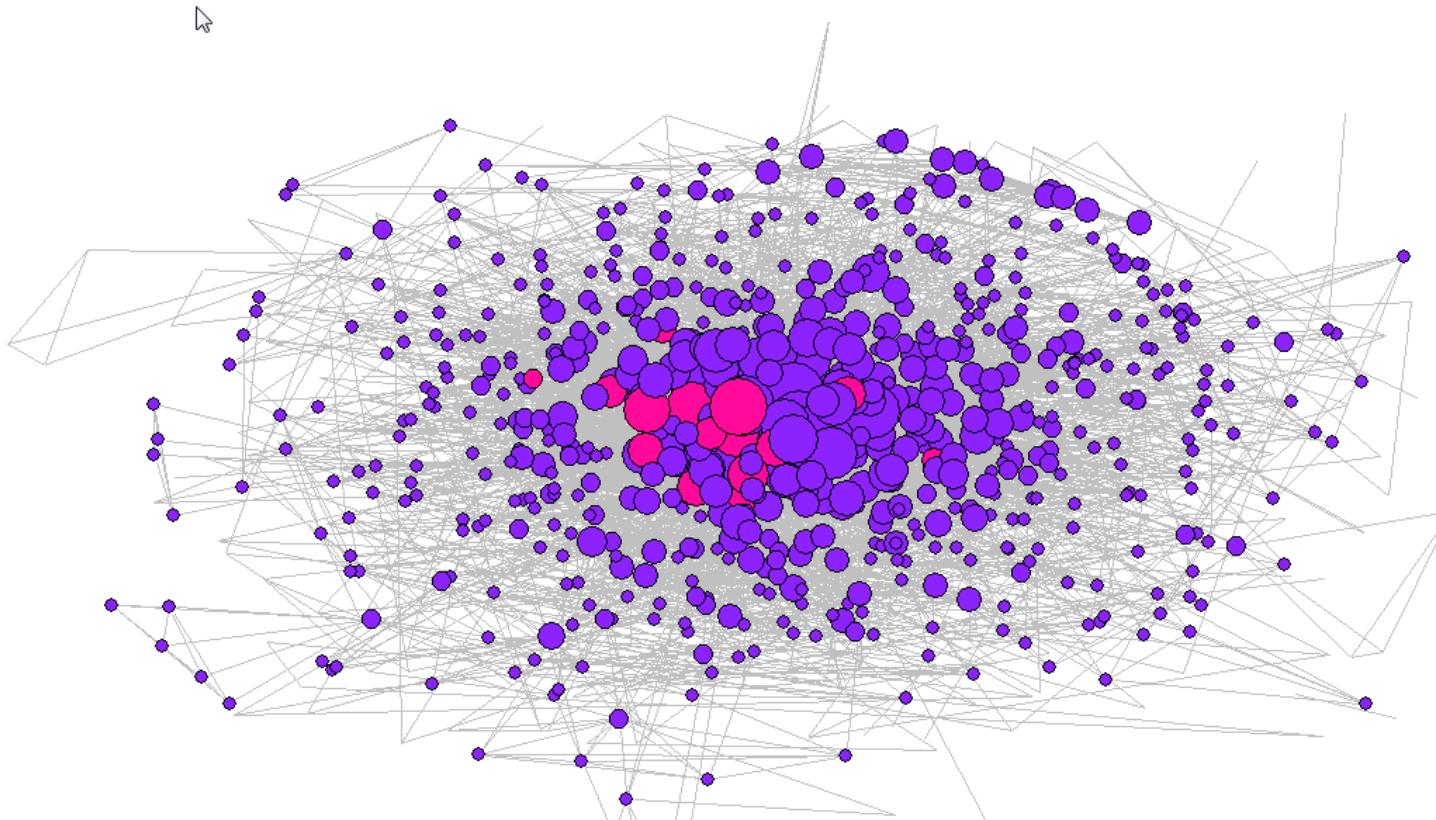


# Social Network Analysis

*And the answer is...*

3. Network from COLS in affiliation network N1 (1016) / C2. C:\MyResearch\Research\HICSS\HICSS47 Tutorial\bollywood\MostPopularActor.clu (1016) / V7. C:\MyResearch\Research\HICSS\HICSS47 Tutorial\bollywood\Vect-D-Reduc...

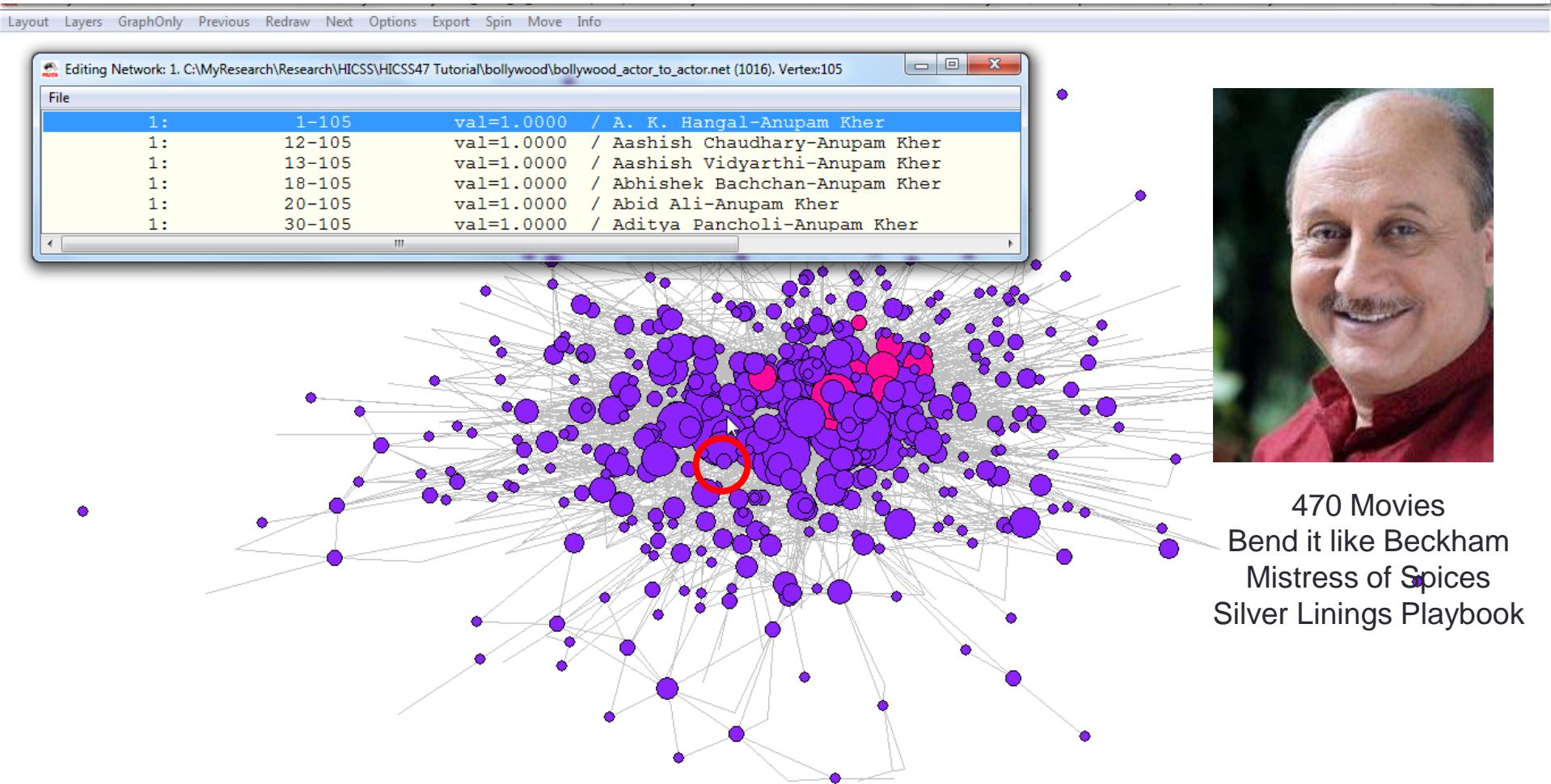
Layout Layers GraphOnly Previous Redraw Next Options Export Spin Move Info



Average Distance is 3.44

# Social Network Analysis

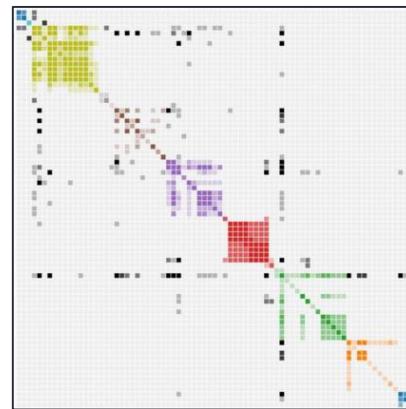
## *and the center of Bollywood is*



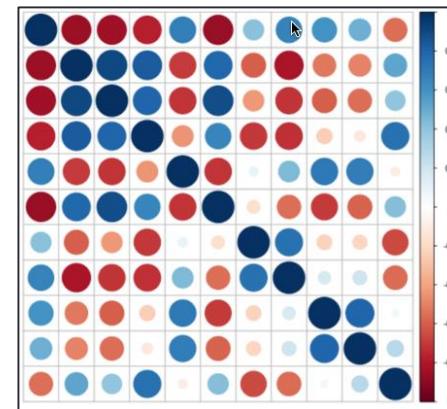
# Alternative Visualizations

*Once in matrix form ...*

Co-occurrence



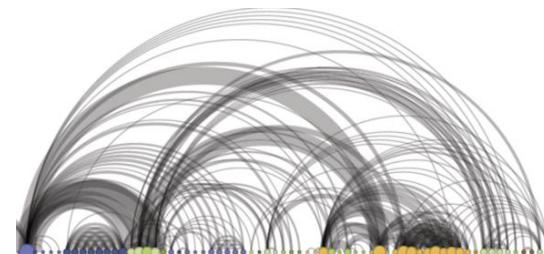
Corrplot



Chord



Arc



# Alternative Visualizations

## *Chord Diagram*

### Migration flows in the United States

This interactive graphic shows migration patterns among states in 2012.

Select a state by mousing over the light-colored edge of the circle. Then mouseover a link to see the number of people moving between your selected state and another state.

Thicker links mean more people moving. States are linked only if at least 10,000 people moved between them. If a state does not appear in the graphic, it is because it did not exchange at least 10,000 people with any other state in 2012.

Source: U.S. Census American Community Survey (ACS)

Notes: Migration figures are estimates based on the 2012 ACS. Estimates of migratory flows are subject to a margin of error that varies by state.

