DataVis et InfoVis

Jean-Daniel Fekete (+ Petra Isenberg)
INRIA
Jean-Daniel.Fekete@inria.fr

Dataviz, InfoGraphics, InfoVis ? [Wikipedia]

- Dataviz: Une **représentation graphique de données statistiques** ou **visualisation de données statistiques** est un résumé visuel des données chiffrées
- InfoGraphic: L'infographie de presse désigne le domaine professionnel ayant pour objet les graphes destinés à mettre en image des informations généralement statistiques au moyens de diagrammes.
- InfoVis: La **Visualisation d'Information** est un domaine informatique pluri-disciplinaire dont l'objet d'étude est la représentation visuelle de données, principalement abstraites, sur une Interface graphique [interactive]

Why

INFORMATION VISUALIZATION

It is estimated that 800 exabyte (**800x 10^{19}**)
of digital information will be generated this year

4

[source: The Diverse and Exploding Digital Universe, IDC, 2008]
[credit: Did You Know; Fisch, McLeod, Brenman]

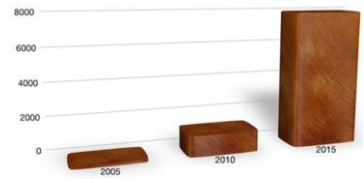
The Big Data Revolution

Sensors and loggers generate more and more data

- Pollution, computer logs, temperature, photos, videos, etc.

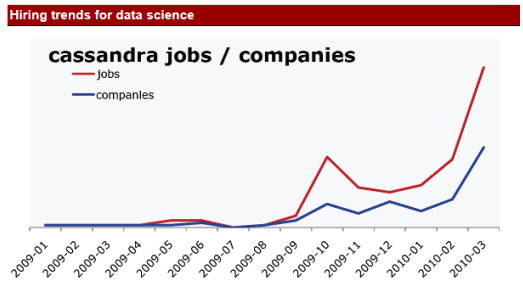
The Digital Universe Explodes:

- 2007: 281 Exabytes (281 billions of Gigabytes)
- 2010: Zetabytes barrier passed
- 2011: 1.8 Zetabytes
- 2015: 7,910 Zetabytes



http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm

6



It's not easy to get a handle on jobs in data science. However, data from O'Reilly Research shows a steady year-over-year increase in Hadoop and Cassandra job listings, which are good proxies for the "data science" market as a whole. This graph shows the increase in Cassandra jobs, and the companies listing Cassandra positions, over time.

"The ability to take data -- to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it - that's going to be a hugely important skill in the next decades."

Hal Varian, chief economist at Google

Question

how can we effectively access data?

- understand its structure?
- make comparisons?
- make decisions?
- gain new knowledge?
- convince others?
- ...

8

Many possible ways to address…

Example

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Raw Data from Anscombe's Quartet

[Source: Anscombe's quartet, Wikipedia]

Statistical Analysis

For all four columns, the statistics are identical

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

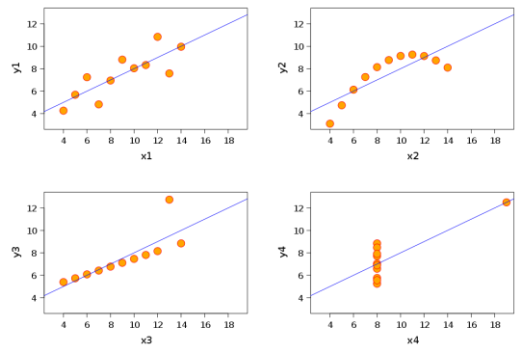
Mean of x	9.0
Variance of x	11.0
Mean of y	7.5
Variance of y	4.12
Correlation between x and y	0.816
Linear regression line	$y = 3 + 0.5x$

[Source: Anscombe's quartet, Wikipedia]

Visual Representation of the Data

Visual representation reveals a different story

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



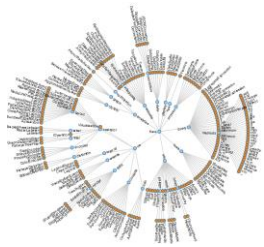
12

[Source: Anscombe's quartet, Wikipedia]

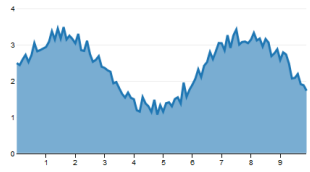
Why visual data representations?

- Vision is our most dominant sense
- We are very good at recognizing visual patterns
- We need to see and understand in order to explain, reason, and make decisions

common examples:



graphs / hierarchies



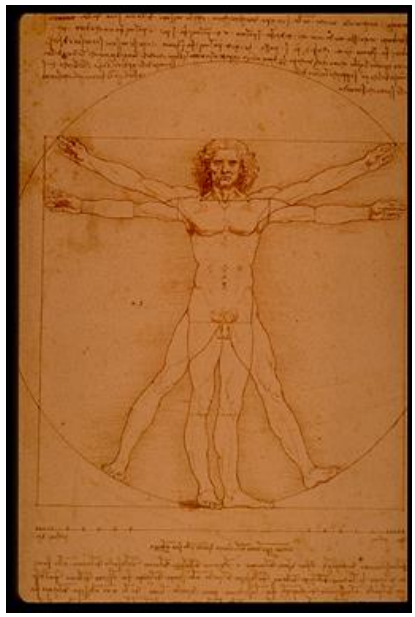
charts

maps

all examples from: http://vis.stanford.edu/protovis/

Other benefits of visualization

- expand human working memory
  - offload cognitive resources to the visual system,
- reduce search
  - by representing a large amount of data in a small space,
- enhance the recognition of patterns
  - by making them visually explicit
- aid monitoring of a large number of potential events
- provides a manipulable medium & allows exploration of a space of parameter values.



L'occhio,
che si dice finestra dell'anima,
è la principale via donde il comune
senso può piú copiosamente e
magnificamente considerare
le infinite opere di natura.

Leonardo da Vinci
(1452 - 1519)

The eye…
the window of the soul,
is the principal means
by which the central sense
can most completely and
abundantly appreciate
the infinite works of nature.

百聞不如一見

"One hundred rumors are not comparable to one look."

An Old Chinese Inscription

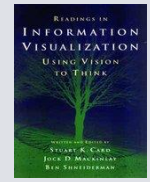
Via Brinton, Graphic Presentation, 1939

Information visualization

- Create visual representation
- Concentrates on abstract data
- Includes interaction

Official Definition:

The use of computer-supported, interactive, visual representations of abstract data to amplify cognition.
[Card et al., 1999]



Functions of Visualizations

- Recording information
  - Tables, blueprints, satellite images
- Processing information
  - needs feedback and interaction
- Presenting information
  - share, collaborate, revise
  - for oneself, for one's peers and to teach
- Seeing the unseen

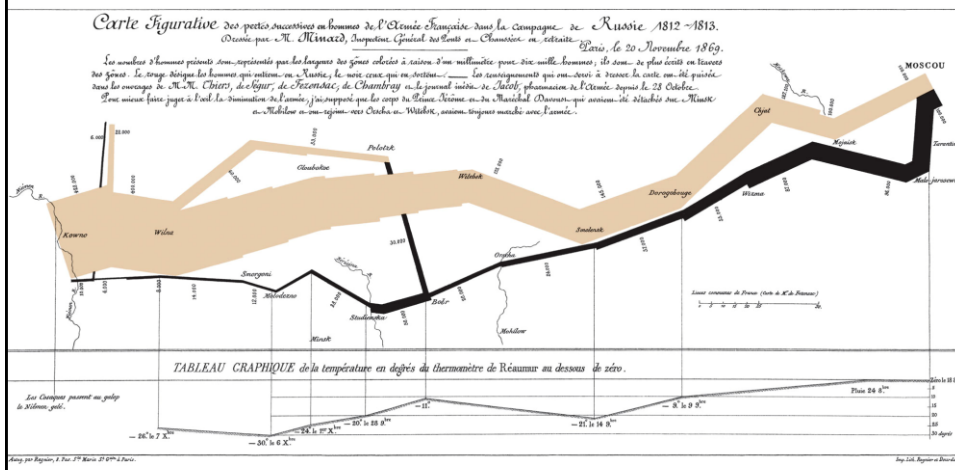
Visualization of abstract data has been practiced for hundreds of years…

HISTORICAL EXAMPLES

Napoleon's March on Moscow Charles Minard, 1869

Named the best statistical graphic ever drawn (by Edward Tufte)

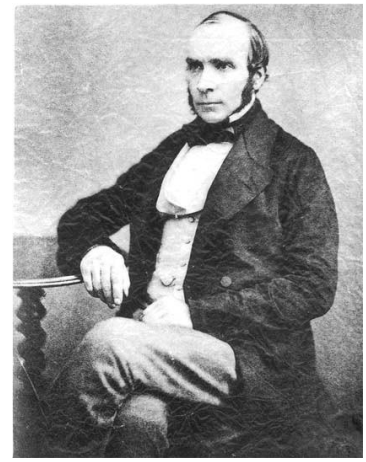
- Includes: spatial layout linked with stats on: army size, temperature, time
- Tells a story in one overview



More info: The Visual Display of Quantitative Information (Tufte)

The Broadway Street Pump

- In 1854 cholera broke out in London
  - 127 people near Broad Street died within 3 days
  - 616 people died within 30 days
- "Miasma in the atmosphere"
- Dr. John Snow was the first to link contaminated water to the outbreak of cholera
- How did he do it?
  - he talked to local residents
  - identified a water pump as a likely source
  - used maps to illustrate his theory
  - convinced authorities to disable the pump



More info here: http://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

22

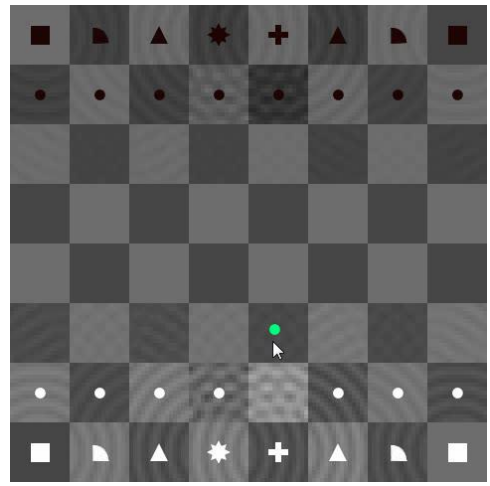
John Snow, 1854

… AND VERY RECENTLY

TrashTrack

Winner of the NSF International Science & Engineering Visualization Challenge!
http://senseable.mit.edu/trashtrack/

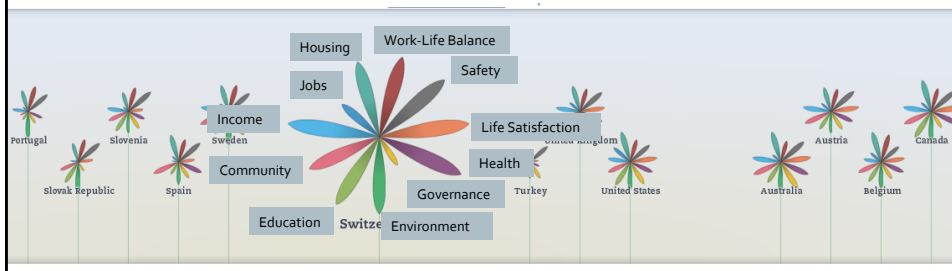
Artificial Intelligence



http://www.turbulence.org/spotlight/thinking/chess.html

Open Data

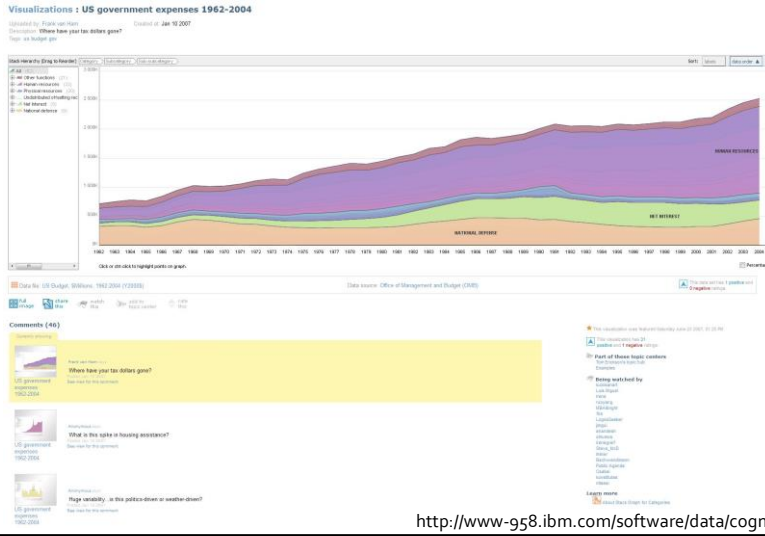
- Movement making government data freely available
- Encourage participation by everyone



OECD Better Life Index: http://www.oecdbetterlifeindex.org/

Many Eyes

- Upload data, create visualizations, discuss
- Distributed asynchronous collaboration

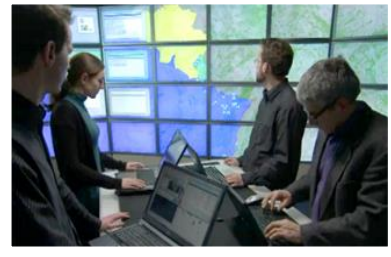


http://www-958.ibm.com/software/data/cognos/manyeyes/

Specific Visualization Environments

Molecular visualisation in the Reality Cube
University of Groningen, NL

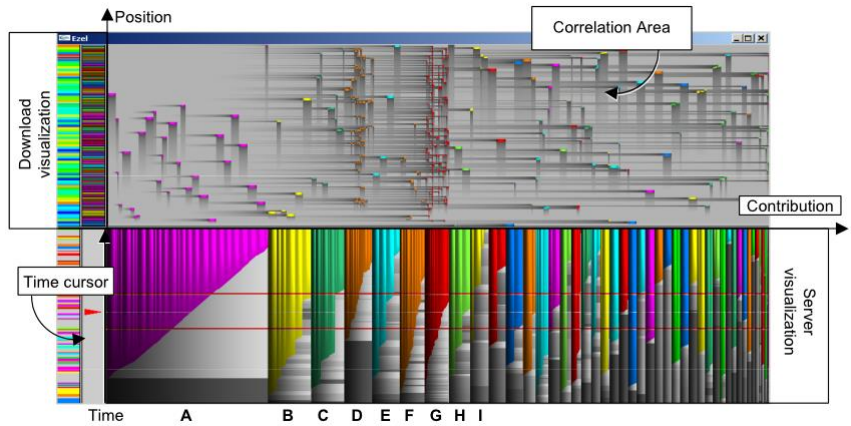
Tabletops for Visualization
University of Calgary



WILD Wall, INRIA

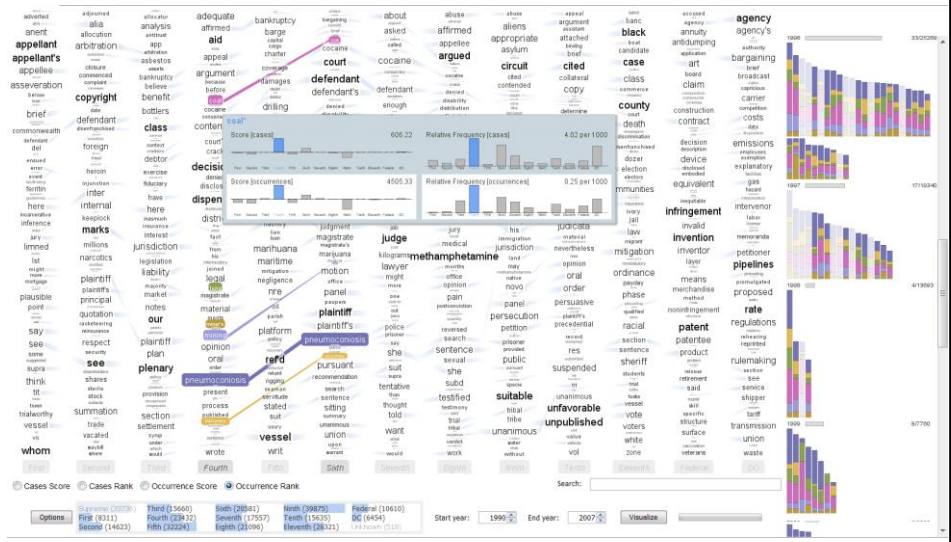
Software Visualization

EZEL: a Visual Tool for Performance Assessment of Peer-to-Peer File-Sharing Networks (Voinea et al., InfoVis, 2004)



Text Visualization

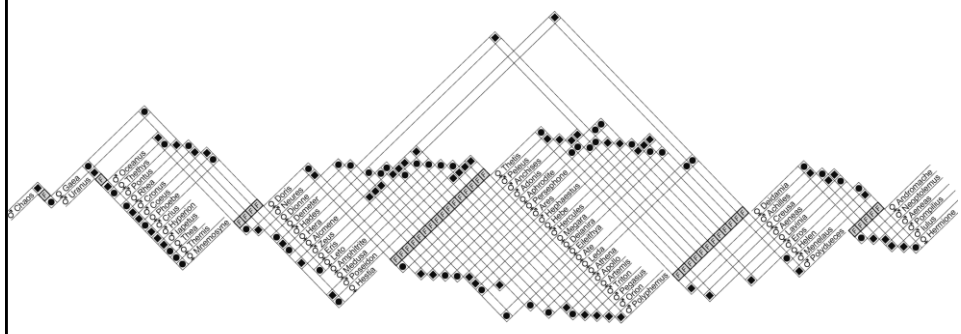
Parallel Tag Clouds to Explore Faceted Text Corpora (Collins et al., VAST 2009)



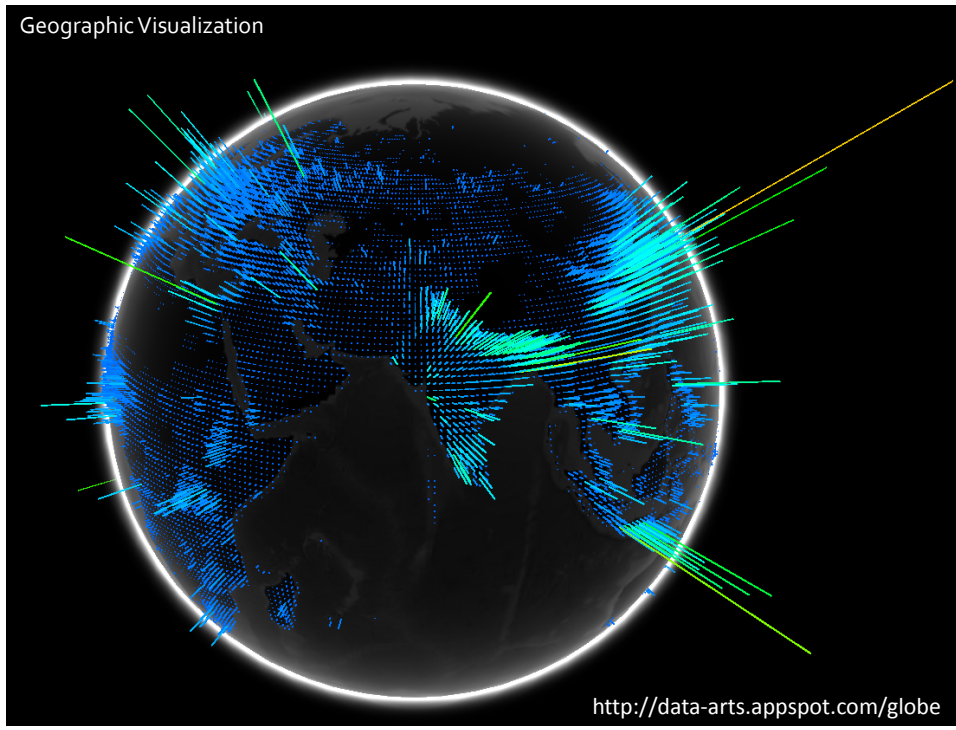
Graphs

http://www.facebook.com/note.php?note_id=469716398919
Visualizing Friendships by Paul Butler on Tuesday, December 14, 2010

Family Trees



http://www.aviz.fr/geneaquilts/



Weather

http://weatherspark.com/

Data Dashboards

http://globalspirometry.com

Resources for more examples

- Visualization conferences
- Blogs
  - http://infosthetics.com/
  - http://fellinlovewithdata.com/
  - http://eagereyes.org/
  - http://flowingdata.com/
  - http://www.informationisbeautiful.net/
- Books
  - Textbooks
    - Readings in Information Visualization: Using Vision to Think (a bit old now but good intro)
    - Information Visualization (Robert Spence – a light intro, I recommend as a start)
    - Information Visualization Perception for Design (Colin Ware, focused on perception and cognition)
    - Interactive Data Visualization: Foundations, Techniques, and Applications (Ward et al. – most recent)
  - Examples
    - Beautiful Data (McCandless)
    - Now You See it (Few)
    - Tufte Books: Visual Display of Quantitative Information (and others)
    - … (many more, ask me for details)

It is difficult to create

CREATE VISUALIZATIONS

GOOD

What is a representation?

- A representation is
  - a formal system or mapping by which the information can be specified (D. Marr)
  - a sign system in that it stands for something other than its self.
- for example: the number thirty-four

34	100010	XXXIV
decimal	binary	roman

Presentation

- different representations reveal different aspects of the information

  decimal: counting & information about powers of 10,
  binary: counting & information about powers of 2,
  roman: impress your friends (outperformed by positional system)

- presentation

  how the representation is placed or organized on the screen

  34, **34**, 34

Principles of Graphical Excellence

- Well-designed presentation of interesting data – a matter of *substance*, *statistics*, *design*
- Complex ideas communicated with clarity, precision, efficiency
- Gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space
- Involves almost always multiple variables
- Tell the truth about the data

The Visual Display of Quantitative Information, Tufte

41

Or a bit more simply…

- Solving a problem simply means representing it so as to make the solution transparent … (Simon, 1981)
- Good representations:
  - allow people to find relevant information
    - information may be present but hard to find
  - allow people to compute desired conclusions
    - computations may be difficult or "for free" depending on representations

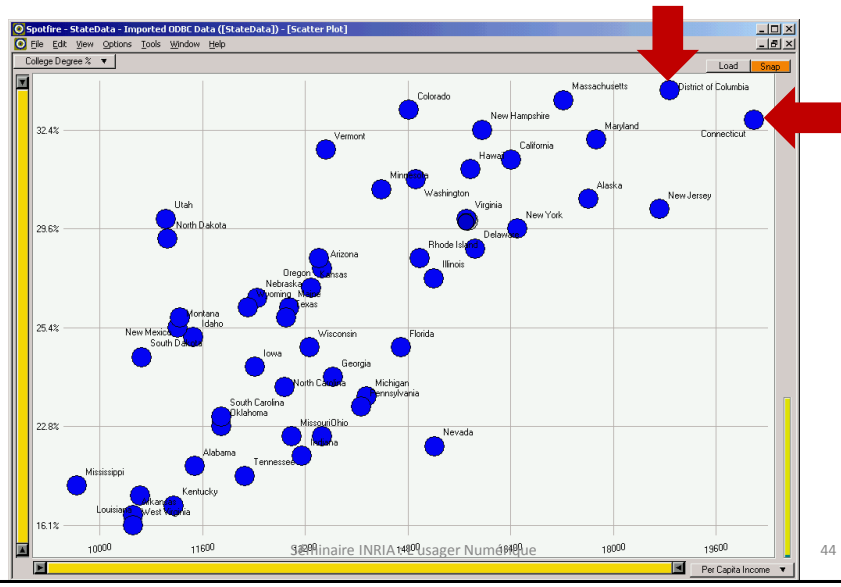
Good representation?

Table - StateData ()

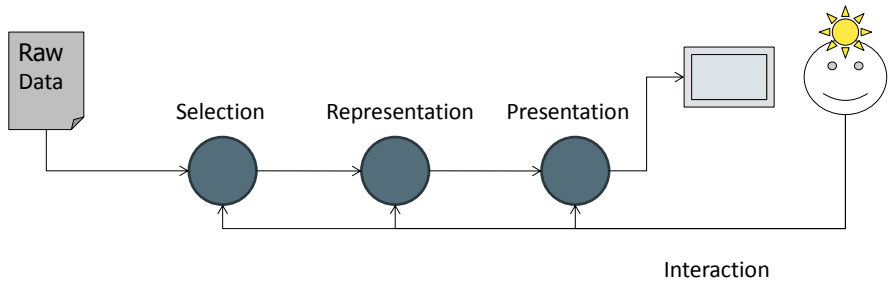
Load | Snap

State	College Degree %	Per Capita Income
Alabama	20.6%	11486
Alaska	30.3%	17610
Arizona	27.1%	13461
Arkansas	17.0%	10520
California	31.3%	16409
Colorado	33.9%	14821
Connecticut	33.8%	20189
Delaware	27.9%	15854
District of Columbia	36.4%	18881
Florida	24.9%	14698
Georgia	24.3%	13631
Hawaii	31.2%	15770
Idaho	25.2%	11457
Illinois	26.8%	15201
Indiana	20.9%	13149
Iowa	24.5%	12422
Kansas	26.5%	13300
Kentucky	17.7%	11153
Louisiana	19.4%	10635
Maine	25.7%	12957
Maryland	31.7%	17730
Massachusetts	34.5%	17224
Michigan	24.1%	14154
Minnesota	30.4%	14389
Mississippi	19.9%	9648
Missouri	22.3%	12989
Montana	25.4%	11213
Nebraska	26.0%	12452
Nevada	21.5%	15214
New Hampshire	32.4%	15959
New Jersey	30.1%	18714
New Mexico	25.5%	11246
New York	29.6%	16501
North Carolina	24.2%	12885
North Dakota	28.1%	11051
Ohio	22.3%	13461
Oklahoma	22.8%	11893
Oregon	27.5%	13418
Pennsylvania	23.2%	14068
Rhode Island	27.5%	14981
South Carolina	23.0%	11897
South Dakota	24.6%	10661
Tennessee	20.1%	12255
Texas	25.5%	12904
Utah	30.0%	11029
Vermont	31.5%	13527
Virginia	30.0%	15713
Washington	30.9%	14923
West Virginia	16.1%	10520
Wisconsin	24.9%	13276
Wyoming	25.7%	12311

Good representation!



How do we arrive at a visualization?



The Visualization Pipeline

From [Spence, 2000]

64

Perception Préattentive

•Qu'est que c'est ?

•Mise en évidence par Anne Treisman (1985), psychologue de la perception

•propriétés visuelles détectées très rapidement par le système visuel (200 – 250 msec)

•exemples: teinte, orientation de ligne, longueur, épaisseur, taille, courbure, cardinalité, clignotement, direction de mouvement … etc.

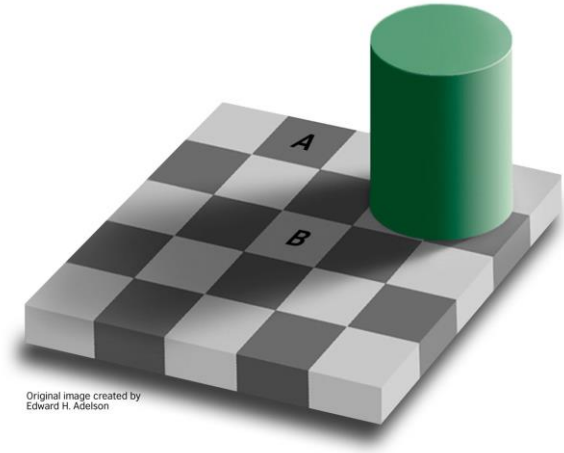
65

Propriétés de la vision

- Sens ayant la plus grande bande passante
- Rapide, reconnaissance de formes
- Préattentif (dans certaines limites)
- Etend les capacités cognitives et mémorielles
- On pense visuellement

- Beaucoup d'avantages, mais quelques inconvénients

Les cases A et B ont des couleurs différentes ?



67

Qu'est-ce qui change entre ces deux images ?

- Qu'est-ce qui change entre ces deux images ?

68

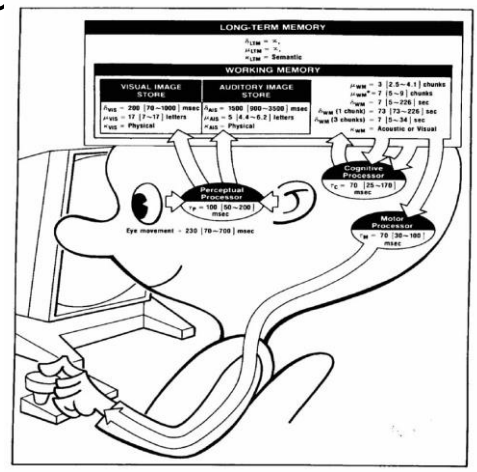
Montrer ce qui change

- Essayez encore

Et la visualisation dans tout ça ?

- On perçoit rapidement certaines caractéristiques graphiques
  - Donc il faut les utiliser pour coder des données
- On doit interagir pour explorer les données
  - Donc il faut offrir des méthodes d'interaction
- Lorsqu'on va vite, on utilise notre mémoire à court terme
  - 7 items +/- 2
  - Il faut utiliser cette mémoire avec parcimonie

Optimiser pour le processeur humain
afin qu'il comprenne les données



Data

- Data is the foundation of any visualization
- The visualization designer needs to understand
  - the data properties
  - know what meta-data is available
  - know what people want from the data

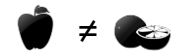
Nominal, Ordinal and Quantitative

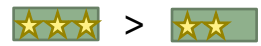
- Nominal (labels)
  - Fruits: apples, oranges
- Ordered
  - Quality of meat: grade A, AA, AAA
  - Can be counted and ordered, but not measured
- Quantitative: Interval
  - no clear zero (or arbitrary)
  - e.g. dates, longitude, latitude
  - usually compare differences (intervals)
- Quantitative: Ratio
  - meaningful origin (zero)
  - physical measurements (temperature, mass, length)
  - counts and amounts

S.S. Stevens, On the theory of scales of measurements, 1946

Nominal, Ordinal and Quantitative

- Nominal (labels)
  - Operations: =, ≠
- Ordered
  - Operations: =, ≠, <, >
- Quantitative: Interval
  - Operations: =, ≠, <, >, -, +
  - Can measure distances or spans
- Quantitative: Ratio
  - Operationrs: =, ≠, <, >, - , +, ×, ÷
  - Can measure ratios or proportions





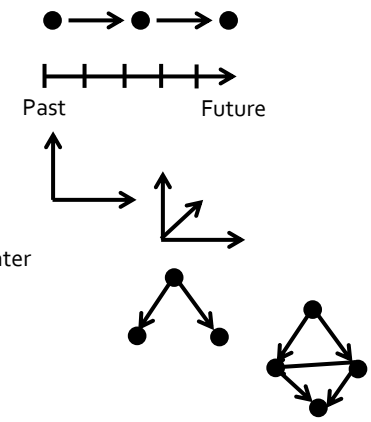
[1989 – 1999] + [2002 – 2012]

10kg / 5kg

S.S. Stevens, On the theory of scales of measurements, 1946

Data-Type Taxonomy

- 1D (linear)
- Temporal
- 2D (maps)
- 3D
- nD (relational) vis examples later
- Trees (hierarchies)
- Networks (graphs)



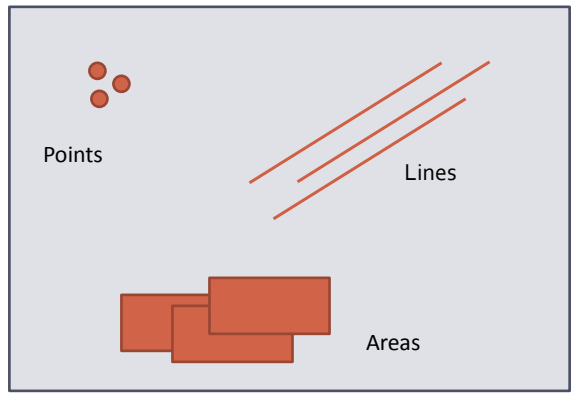
Shneiderman: The Eyes Have It

Why is this important?

- Nominal, ordinal, and quantitative data are best expressed in different ways visually
- Data types often have inherent tasks
  - temporal data (comparison of events)
  - trees (understand parent-child relationships)
  - …
- But:
  - any data type (1D, 2D,…) can be expressed in a multitude of ways!

Visualization's Main Building Blocks

Marks which represent:



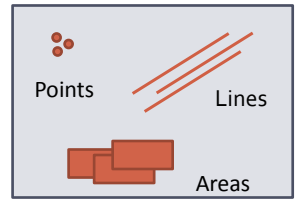
From Semiology of Graphics (Bertin)

77

The following slides on the topic adapted from Sheelagh Carpendale

Points

- "A point represents a location on the plane that has **no theoretical length or area**. This signification is independent of the size and character of the mark which renders it visible."
- a location
- marks that indicate points can vary in all visual variables

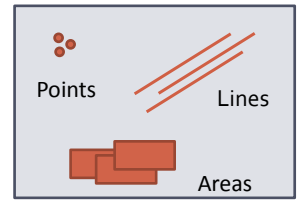


From Semiology of Graphics (Bertin)

78

Lines

- "A line signifies a phenomenon on the plane which has **measurable length but no area**. This signification is independent of the width and characteristics of the mark which renders it visible."
- a boundary, a route, a connection

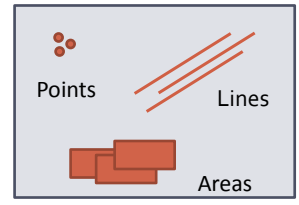


From Semiology of Graphics (Bertin)

79

Areas

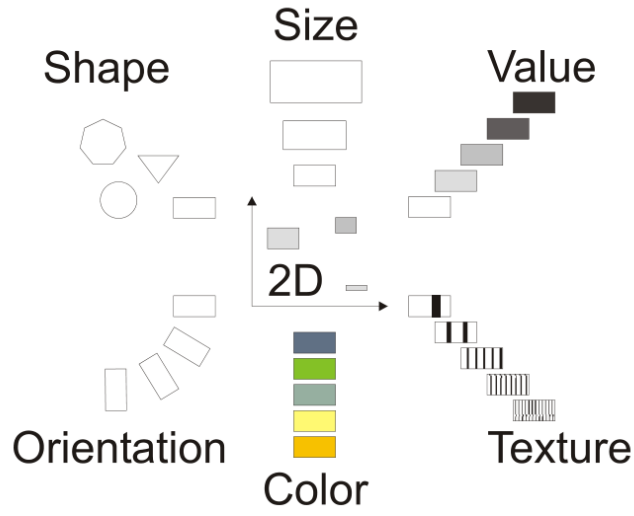
- "An area signifies something on the plane that **has measurable size**.
  This signification applies to the entire area covered by the visible mark."
- an area can change in position but not in size, shape or orientation without making the area itself have a different meaning



From Semiology of Graphics (Bertin)

80

Visual Variables Applicable to Marks

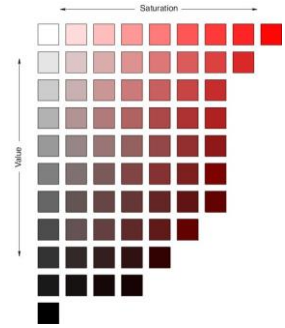


From Semiology of Graphics (Bertin)

Additional Variables for Computers

- **motion**
  - direction, acceleration, speed, frequency, onset, 'personality'

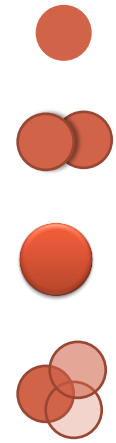
- **saturation**
  - colour as Bertin uses largely refers to hue, saturation != value



Extending those from Semiology of Graphics (Bertin)

Additional Variables for Computers

- **flicker**
  - frequency, rhythm, appearance
- **depth?** 'quasi' 3D
  - depth, occlusion, aerial perspective, binocular disparity
- **Illumination**
- **transparency**



From Semiology of Graphics (Bertin)

Characteristics of Visual Variables

- Selective:
  Can this variable allow us to spontaneously differentiate/isolate items from groups?
- Associative:
  Can this variable allow us to spontaneously group items in a group?
- Ordered:
  Can this variable allow us to spontaneously perceive an order?
- Quantitative:
  Is there a numerical reading obtainable from changes in this variable?
- Length (resolution):
  Across how many changes in this variable are distinctions possible?

From Semiology of Graphics (Bertin)

84

Motion

85

Visual Variables

Visual Variable	Selective	Associative	Quantitative	Order	Length
Position	Yes	Yes	Yes	Yes	Dependant on resolution
Size	Yes	Yes	Approximate	Yes	Association: 5; Distinction: 20
Shape	With Effort	With Effort	No	No	Infinite
Value	Yes	Yes	No	Yes	Association: 7; Distinction: 10
Hue	Yes	Yes	No	No	Association: 7; Distinction: 10
Orientation	Yes	Yes	No	No	4
Grain	Yes	Yes	No	No	5
Texture	Yes	Yes	No	No	Infinite
Motion	Yes	Yes	No	Yes	Unknown

Carpendale, 2003

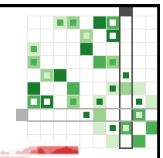
86

Summary

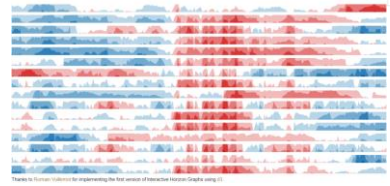
	Quantitative	Ordinal	Nominal
More Accurate	Position	Position	Position
↑	Length	Density	Hue
	Angle	Saturation	Density
	Slope	Hue	Saturation
	Area	Length	Shape
	Density	Angle	Length
	Saturation	Slope	Angle
↓	Hue	Area	Slope
Less Accurate	Shape	Shape	Area

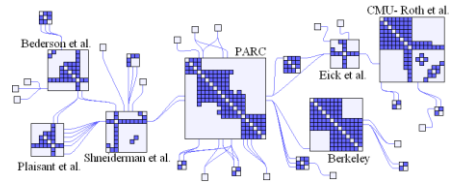
Jacques Bertin refined by Cleveland&McGill then by Card&Mackinlay

Visualisation avancée

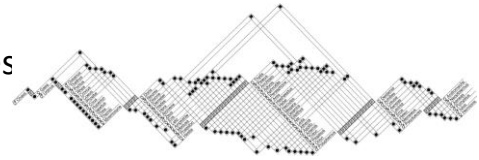


- AVIZ (et d'autres) conçoivent des visualisations très efficaces
- Comment les pousser vers le grand public / citoyen ?





- Attendre 15 ans que les techniques soient adoptées ?



Promouvoir une prise de décision éclairée par les citoyens

- La visualisation est efficace
- Donc, tous les citoyens veulent l'utiliser ?

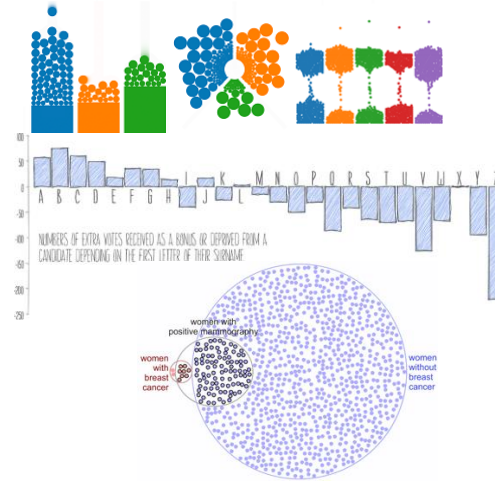
- Non hélas :
  - Pas beaucoup de sources de visualisation pour le grand public
  - Beaucoup de mauvaises représentations visuelles (pas préattentive donc pas mieux que tu texte)
  - Problème d'« **illettrisme visuel** »

Illettrisme visuel

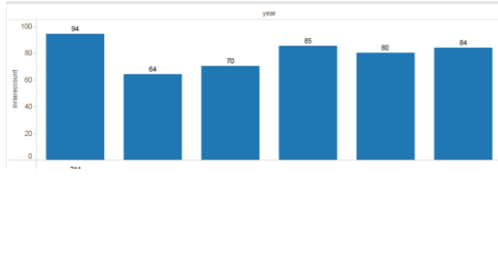
- Dès le CP, nous apprenons à lire, à écrire et à compter
  - Représentations symboliques des concepts
- Nous n'apprenons pas à lire les graphiques
  - Sauf les cartes
- Les jeunes apprennent les graphiques avec les jeux vidéo
- Les moins jeunes n'ont presque jamais appris.

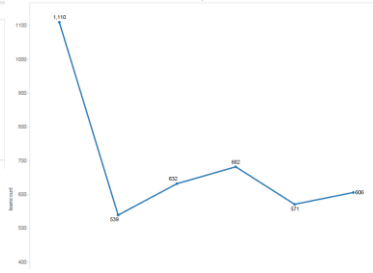
Recherches sur les visualisations engageantes à AVI7

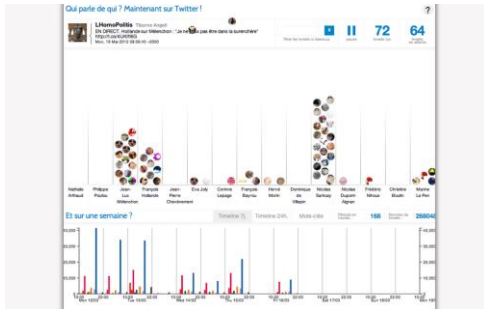
- Nouvelles métaphores pour visualiser les données temporelles
  - –http://www.visualsedimentation.org/
- Utilisation du style crayonné
  - –http://www.aviz.fr/Research/SketchyRendering
- Raisonnement Bayesien compréhensible
  - –http://www.aviz.fr/bayes
- Visualisation pour le peuple
  - –http://peopleviz.gforge.inria.fr/trunk/

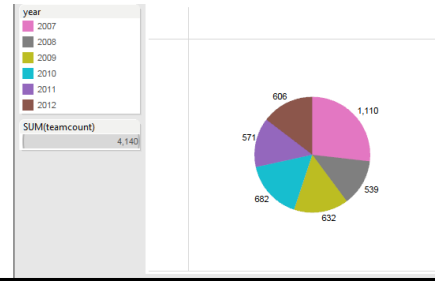


Montrer 1 dimension à la fois: ok

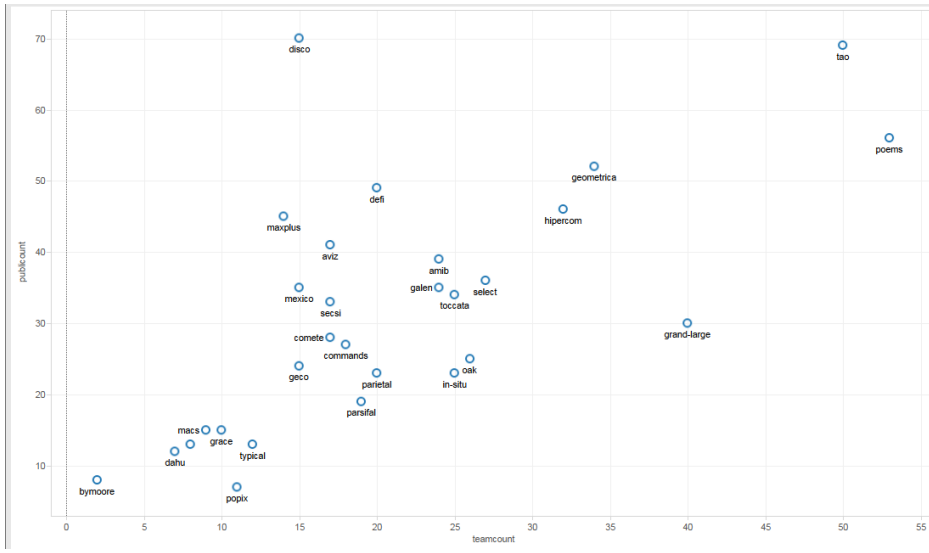




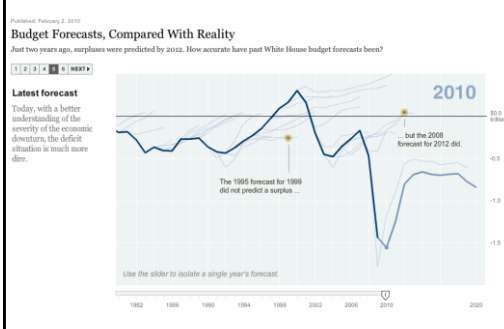


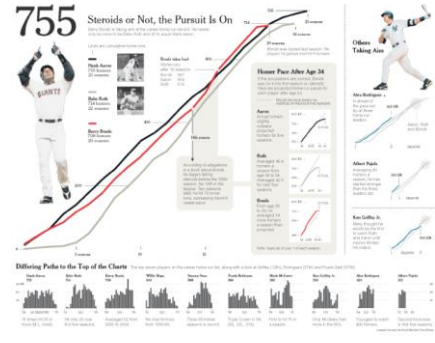


Montrer 2 dimensions à la fois…



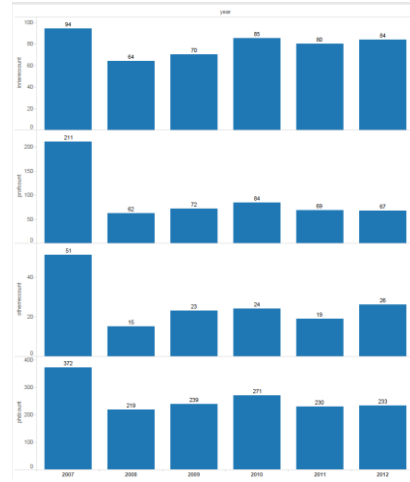
Le New York Times s'en mêle



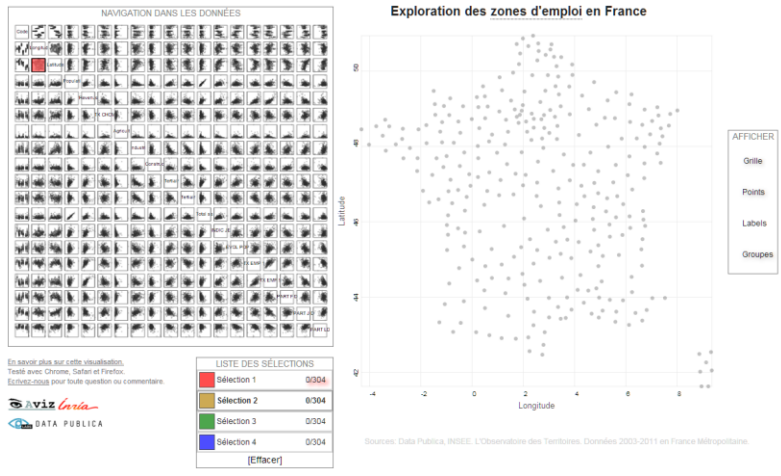


Montrer > 2 dimensions

- N x 1 dimension
  - Matrice de Bertin
- N-1 x 2 dimensions
  - Matrice de scatterplot
- Coordonnées parallèles
  - Beaucoup d'apprentissage



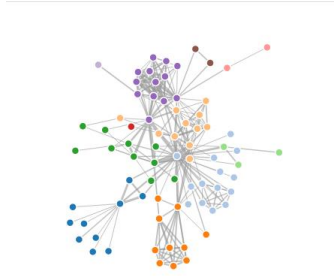
Matrices de Scatter Plots

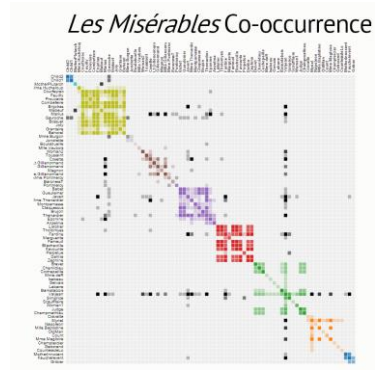


http://labs.data-publica.com/emploi/

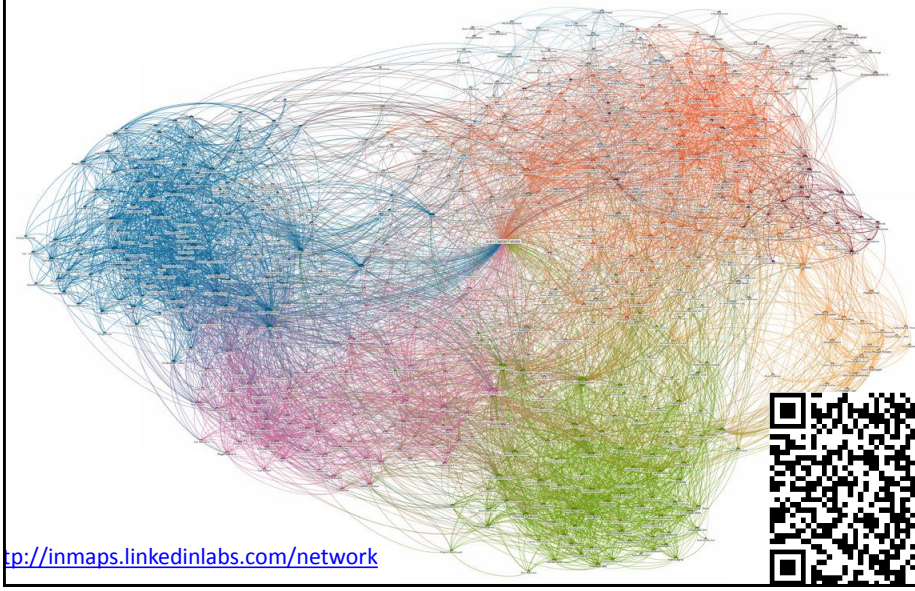
Graphes

- Diagrammes en nœuds et liens

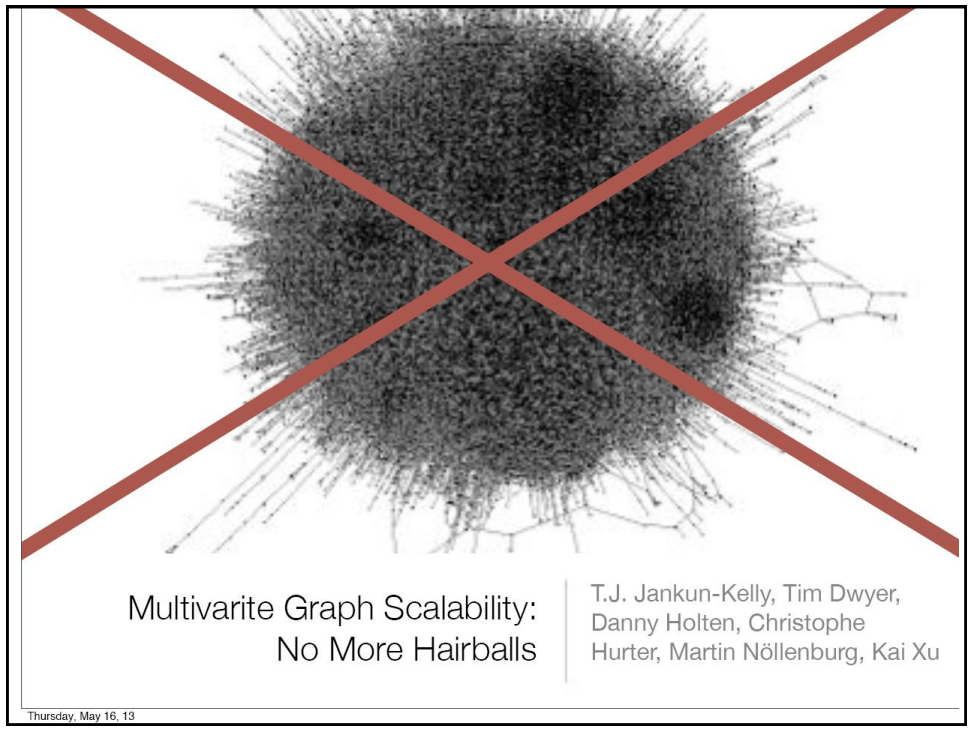




Réseau Professionnel LinkedIn



tp://inmaps.linkedinlabs.com/network



Multivarite Graph Scalability:
No More Hairballs

T.J. Jankun-Kelly, Tim Dwyer,
Danny Holten, Christophe
Hurter, Martin Nöllenburg, Kai Xu

Thursday, May 16, 13