Introduction to Information Visualization

James EAGAN

TELECOM ParisTech

Slides adapted from John Stasko (Georgia Tech), Petra Isenberg & Jean-Daniel Fekete (INRIA), Chris North (Virginia Tech), Tamara Munzner (UBC)

Who am I?

James EAGAN

MAÎTRE DE CONFÉRENCES EN INTERACTION HOMME-MACHINE

Associate Prof. at Télécom ParisTech
Adjunct Researcher at LTCI

Ph.D. 2008 — Georgia Tech

B.A. 2000 — Lawrence University

james.eagan@telecom-paristech.fr

Research

Human-Computer Interaction

Information Visualization

Multi-surface Interaction

Exercise

- House directions

Data Exploration

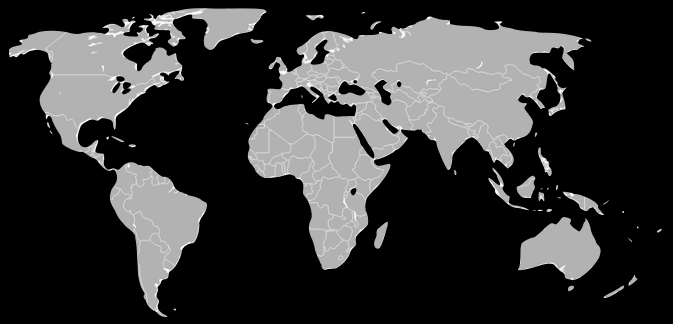
- Society is more complex
  - There is simply more “stuff”
- Computers, internet, and web give people access to an incredible amount of data
  - news, sports, financial, sales, demographics, etc.
  - pollution, computer logs, weather, photos, videos, etc.

How much data?

- Between 1 and 2 exabytes of unique info produced per year
  - 1000000000000000000 (10^{18}) bytes
  - 250 meg for every man, woman and child
  - Printed documents only .003% of total

Peter Lyman and Hal Varian, 2000
Cal-Berkeley, Info Mgmt & Systems
`www.sims.berkeley.edu/how-much-info`

2008

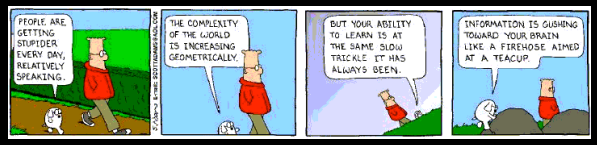


800 exabytes per year

[The Diverse and Exploding Digital Universe, IDC, 2008]

Data Overload

- How can we make use of the data?
  - How do we make sense of the data?
  - How do we harness this data in decision-making processes?
  - How do we avoid being overwhelmed?



Question

- How can we effectively access data?
  - understand its structure?
  - make comparisons?
  - make decisions?
  - gain new knowledge?
  - *convince others?*
  - ....

The need is there

"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's going to be a hugely important skill in the next decades."

— Hal Varian, chief economist, Google

[The McKinsey Quarterly, January 2009]

The Problem

Web,
Books,
Papers,
Game scores,
Scientific data,
Biotech,
Shopping
People
Stock/finance
News

Vision: 100 MB/s
Ears: <100 b/s
Telepathy
Haptic/tactile
Smell
Taste

[Courtesy of Chris North, Virginia Tech]

Human Vision

- Highest bandwidth sense
- Fast, parallel
- Pattern recognition
- Pre-attentive
- Extends memory and cognitive capacity

(Multiplication test)

- People think visually

Impressive. Lets use it!

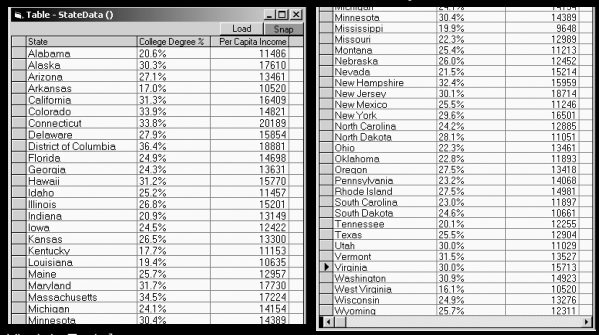
[Courtesy of Chris North, Virginia Tech]

The Challenge

- Transform the data into information (understanding, insight) thus making it useful to people

Example

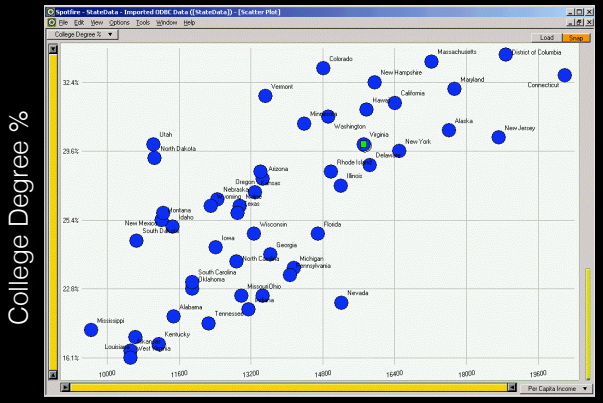
Questions:
Which state has the highest income?
Is there a relationship between income and education?
Are there any outliers?



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	28.8%	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%	10620
Wisconsin	24.9%	13276
Wyoming	25.7%	12311

[Courtesy of Chris North, Virginia Tech]

Visualize the Data



[Courtesy of Chris North, Virginia Tech]

Even Tougher?

- What if you could only see one state's data at a time? (e.g. U.S. Census Bureau's website)
- What if I read the data to you?

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

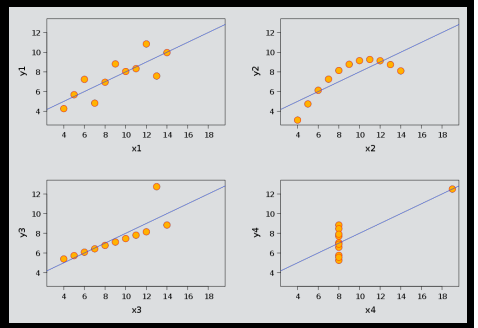
Anscombe's Quartet

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	$y = 3 + 0.5x$

Anscombe's Quartet

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



Illustrates Our Approach

- Provide tools that present data in a way to help people understand and gain insight from it
- Clichés
  - "Seeing is believing"
  - "A picture is worth a thousand words"

Visualization

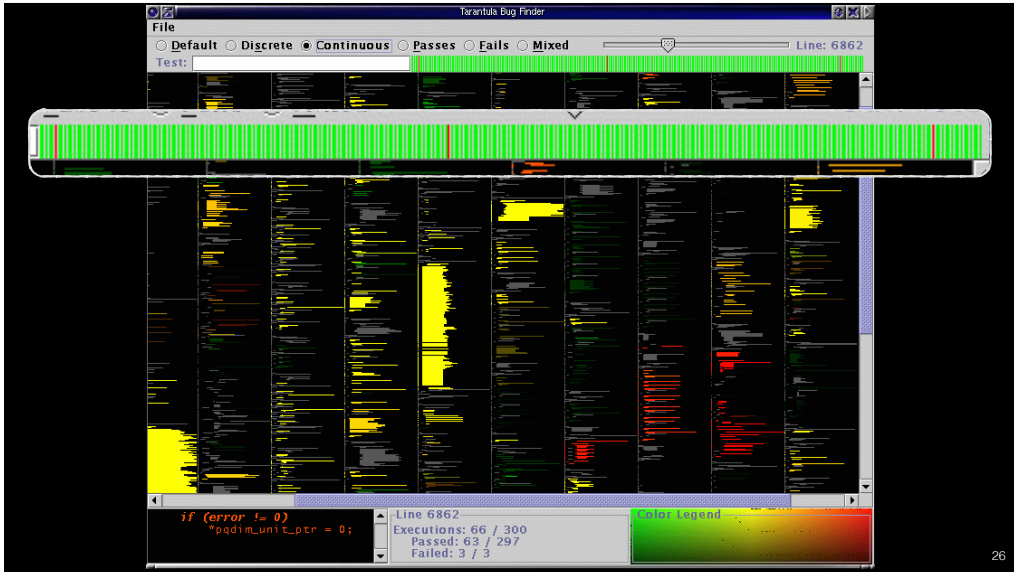
- Often thought of as process of making a graphic or an image
- Really is a cognitive process
  - Form a mental image of something
  - Internalize an understanding
- "The purpose of visualization is insight, not pictures"
- Insight: discovery, decision making, explanation

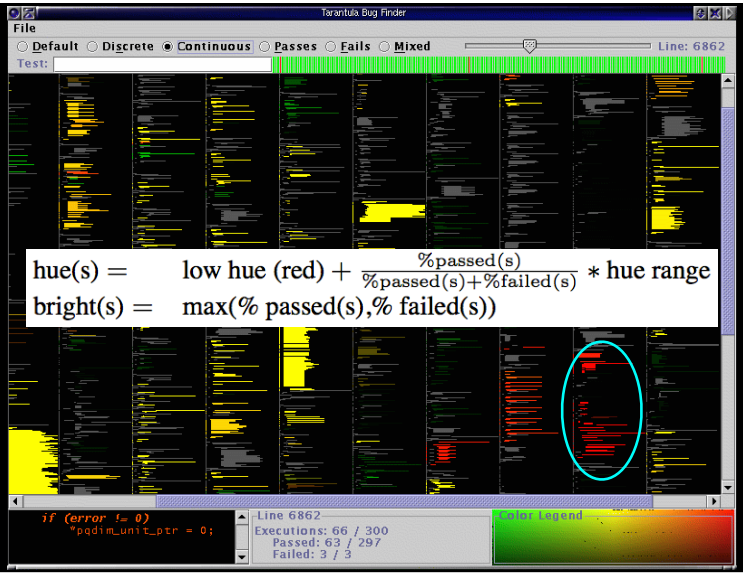
Main Idea

- Visuals help us think
  - Provide a frame of reference, a temporary storage area
- External cognition
  - Role of external world in thinking and reason
- Examples?

Information Visualization

- What is “visualization”?
  - The use of computer-supported, interactive visual representations of data to amplify cognition.
  - From [Card, Mackinlay, Shneiderman ’98]





Three Subfields

- Scientific Visualization
- Information Visualization
- Visual Analytics

Scientific Visualization

- Primarily relates to and represents something physical or geometric
- Examples
  - Air flow over a wing
  - Stresses on a girder
  - Weather over Pennsylvania

Information Visualization

- Components:
  - Taking items without a direct physical correspondence and mapping them to a 2-D or 3-D physical space.
  - Giving information a visual representation that is useful for analysis and decision-making

Two Key Attributes

- Scale
  - Challenge often arises when data sets become very large
- **Interactivity**
  - Want to show multiple different perspectives on the data

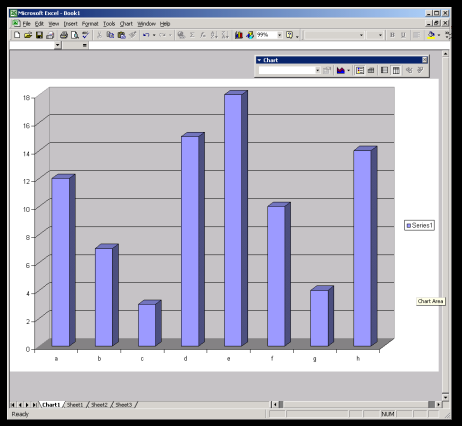
Domains for InfoVis

- Text
- Statistics
- Financial/business data
- Internet information
- Software
- ...

Examples

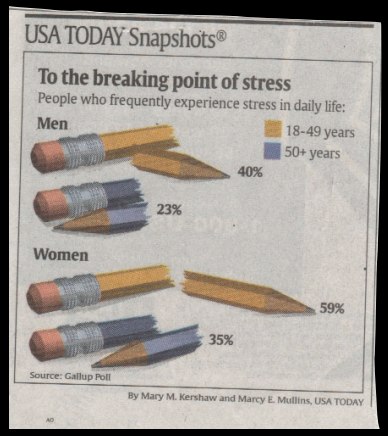
- Images
- Are these static pictures information visualizations?

Excel



USA Today Graphics

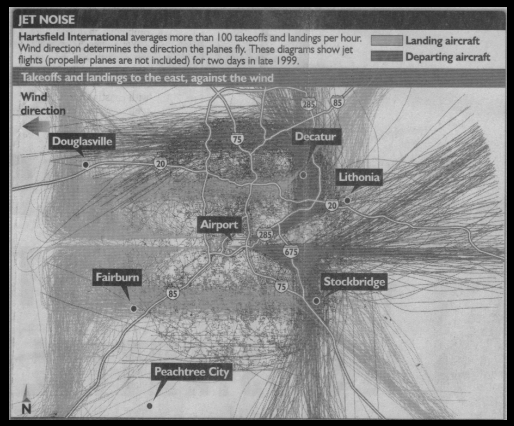
Or worse yet…



Bullet Train Schedule

Atlanta Flight Traffic

Atlanta Journal
April 30, 2000



Country Music

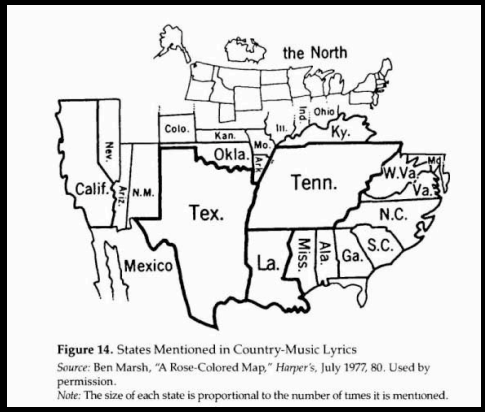
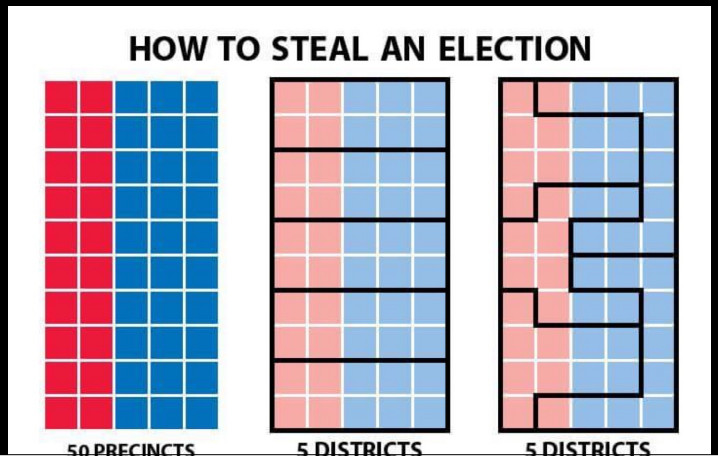
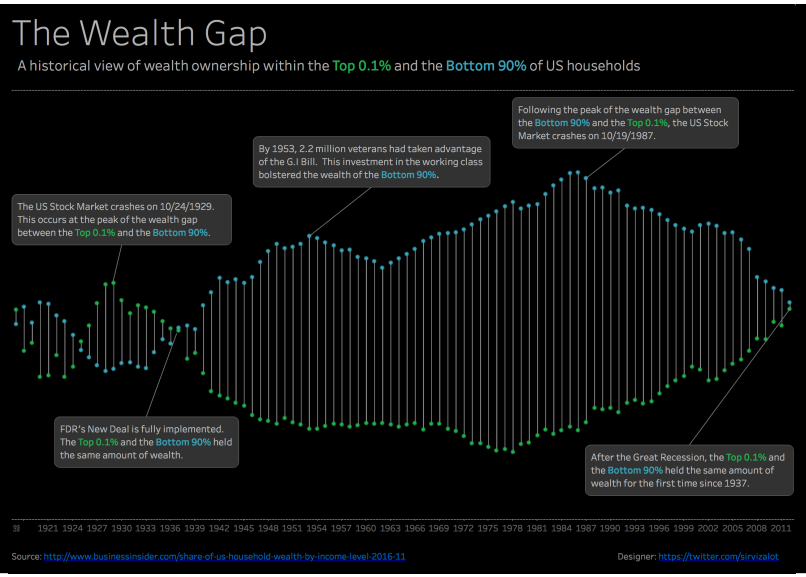
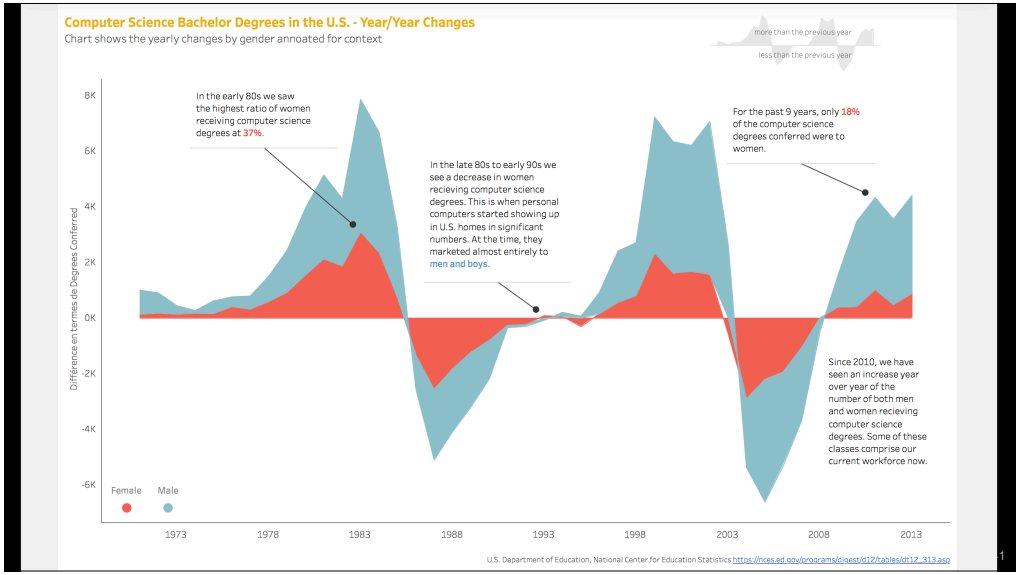


Figure 14. States Mentioned in Country-Music Lyrics
Source: Ben Marsh, "A Rose-Colored Map," *Harper's*, July 1977, 80. Used by permission.
Note: The size of each state is proportional to the number of times it is mentioned.



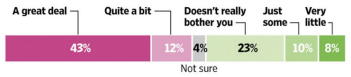




Critique

1. Original WSJ graphic

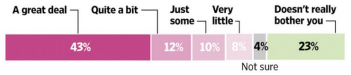
Q: How much does it bother you that hackers working in connection with a foreign government were involved in trying to influence our elections?



Looks like purple means "bothered" and green means "not bothered." The graphic conveys near balance between bothered and not, as well as a consistent left/right scale.

But why is "just some" darker than "doesn't really bother you"? In fact, why is it even green at all, and so far to the right?

2. Redesign: consistent colors and order



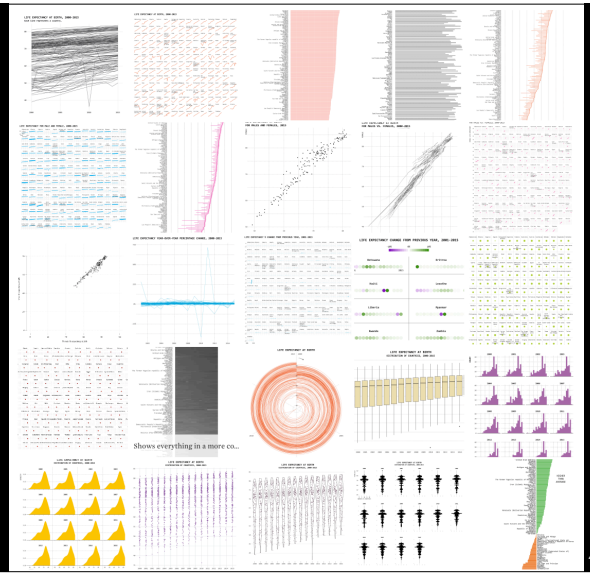
A redesigned graphic makes it easier to see how many people are at least a little bothered.

There's still room for debate: What's the best color for "very little"? The right darkness for "quite a bit"? Were these even useful categories to ask about?

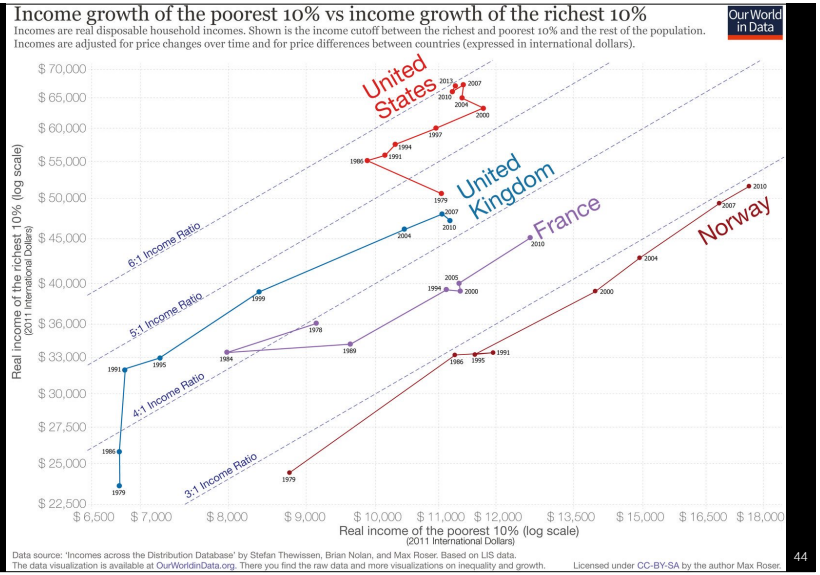
But design really matters here.

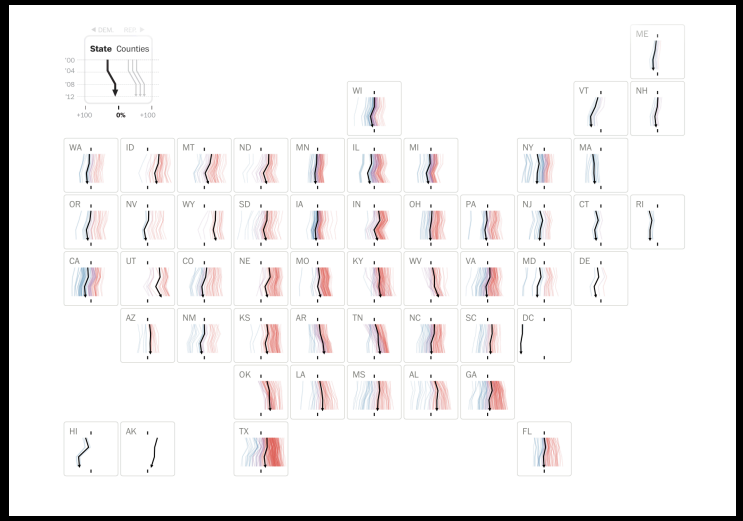
Life expectancy 2000–2015

flowingdata.com/2017/01/24/one-dataset-visualized-25-ways/



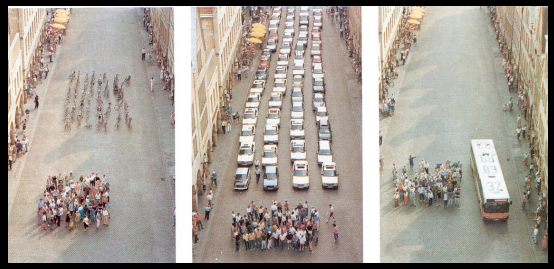
[Max Roser]





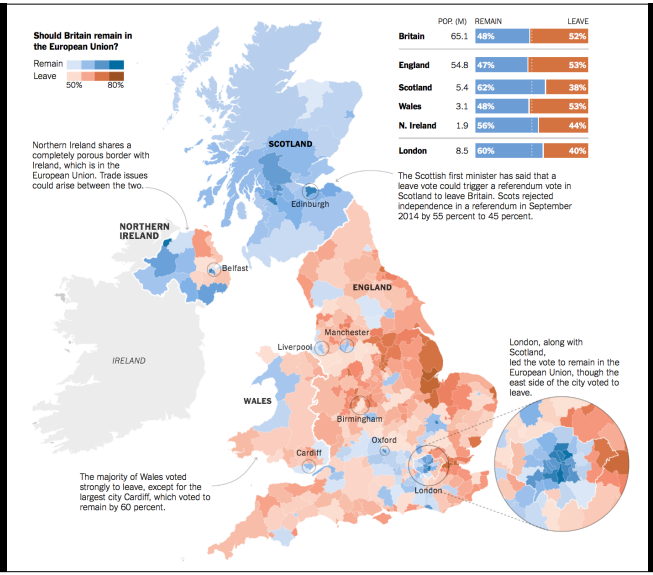
[Washington Post]

72 people

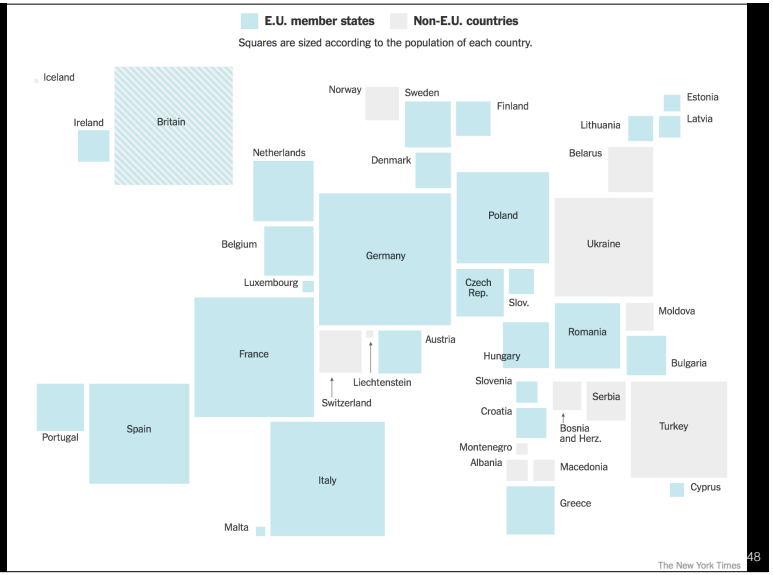


[www.visualnews.com]

Brexit

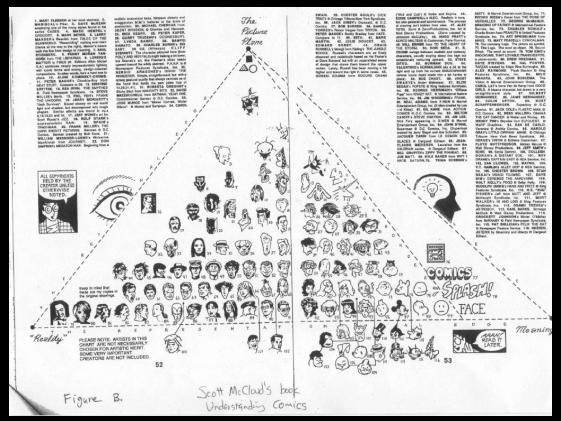


E.U.

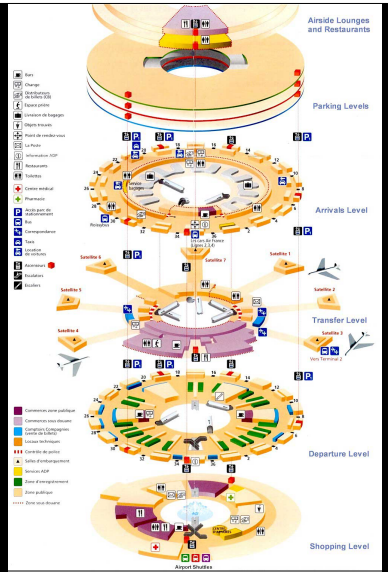


The Comics

Understanding Comics by Scott McCloud

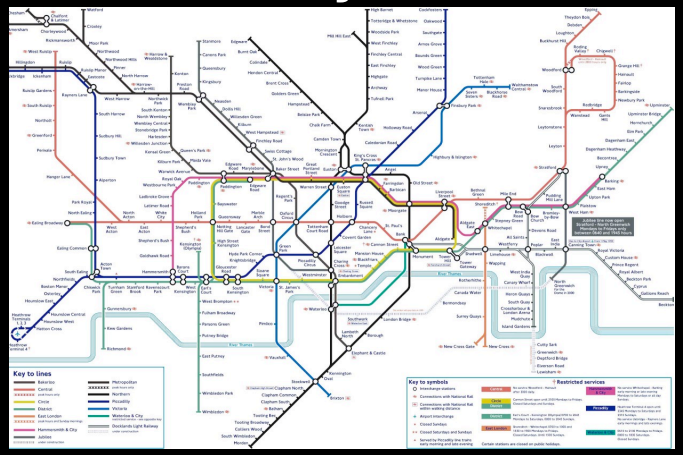


Charles de Gaulle (Paris)

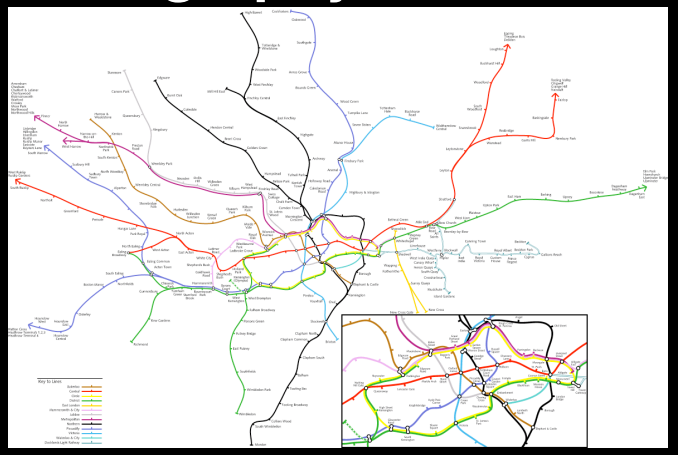


London Subway

www.thetube.com



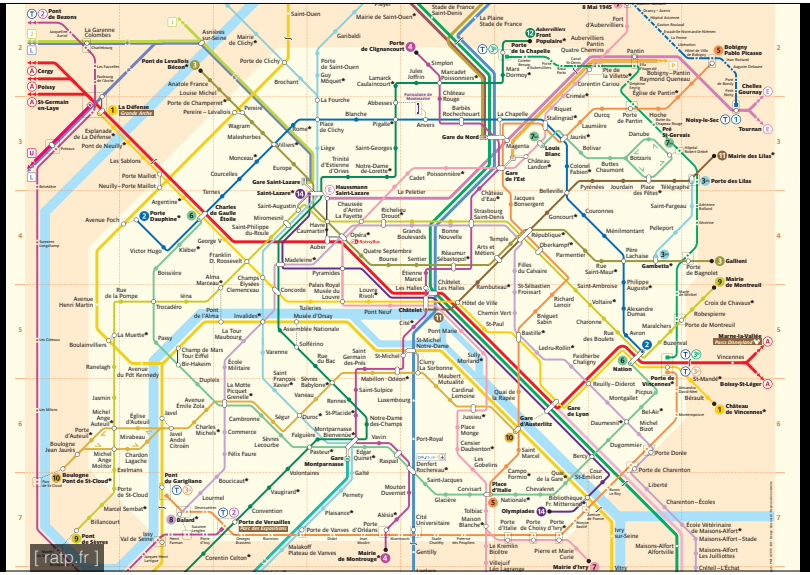
True Geography

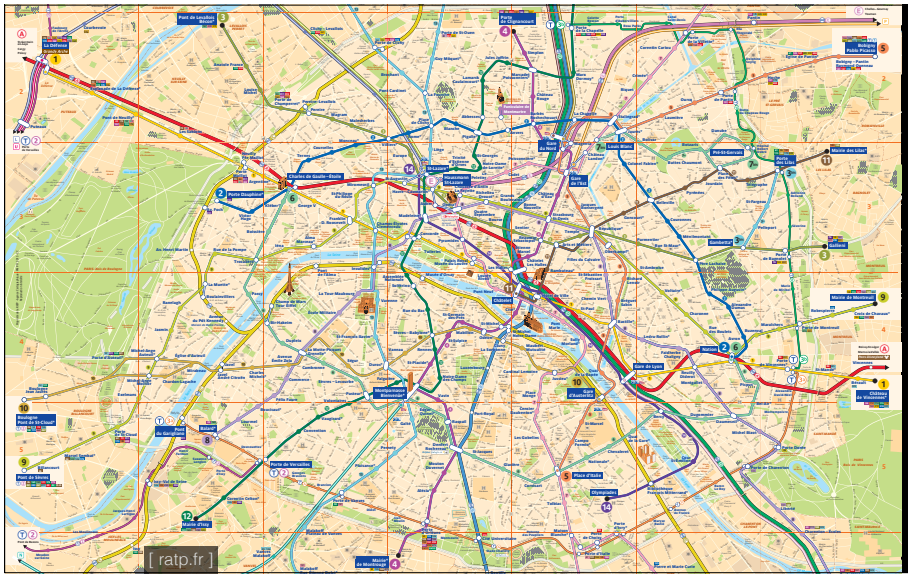


www.kottke.org/plus/misc/images/tubegeo.gif

Easy Walking Lines Added

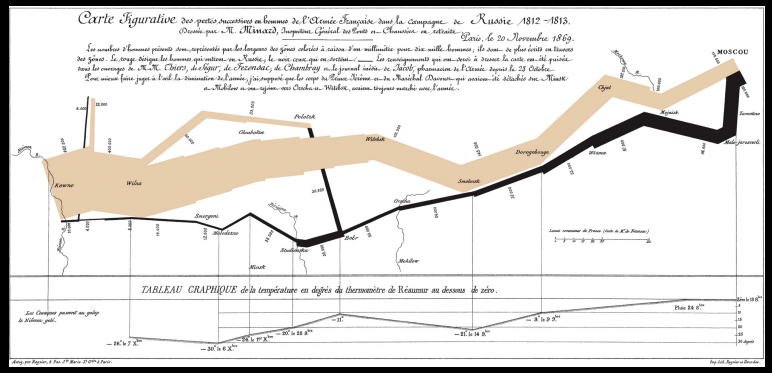
rodcorp.typepad.com/photos/art_2003/tube_walklines_final_lmfaint.html





Napolean's March

From E. Tufte
The Visual Display of
Quantitative
Information



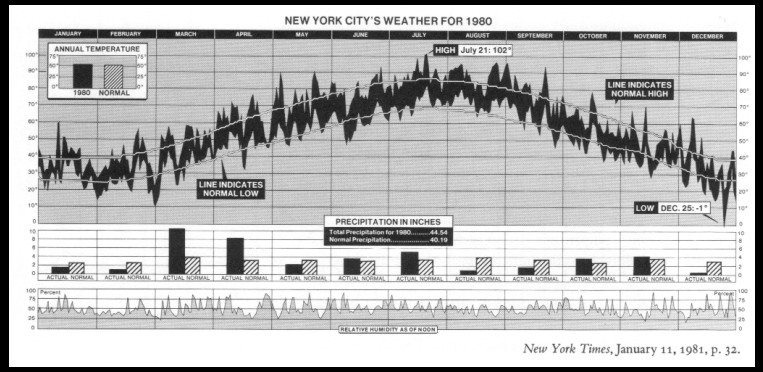
Graphic by Minard

size of army
direction

latitude
longitude

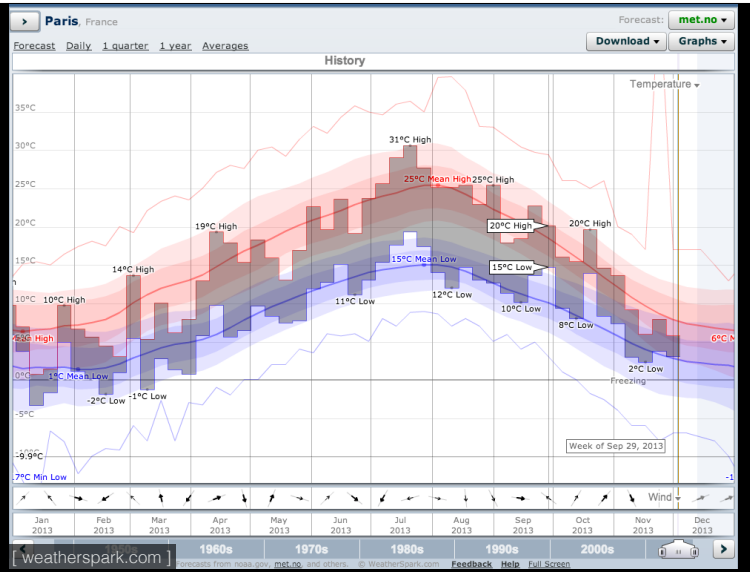
temperature
date

NYC Weather



Tufte, Vol. 1

2220 numbers

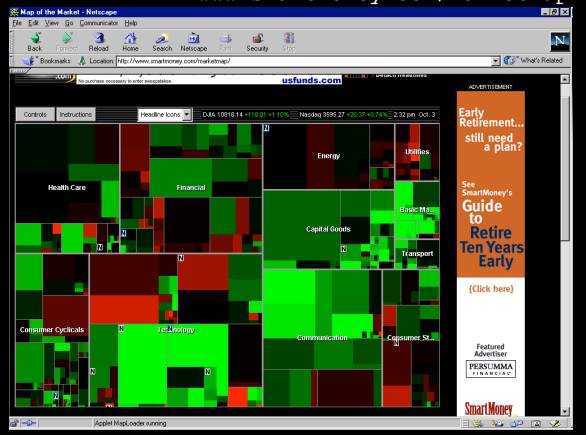


Examples

- Software

Map of the Market

www.smartmoney.com/marketmap

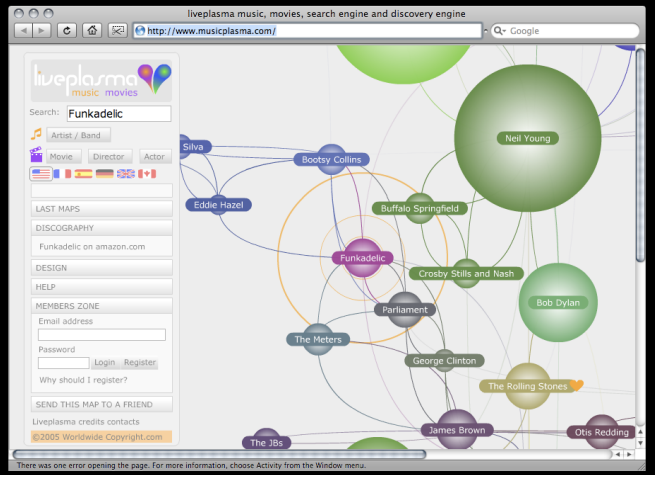


New York Timeformations

http://www.skyscraper.org/timeformations/transparent.html

Music Plasma

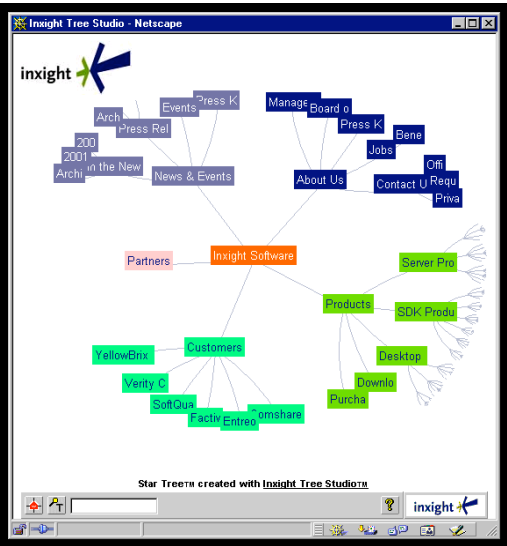
http://www.musicplasma.com/



StarTree

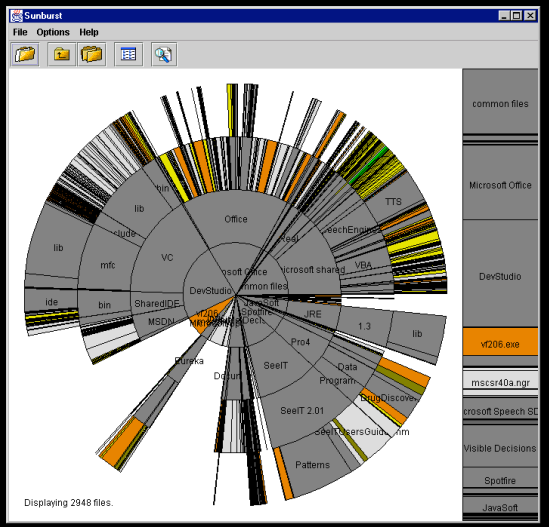
Hyperbolic Tree

www.inxight.com



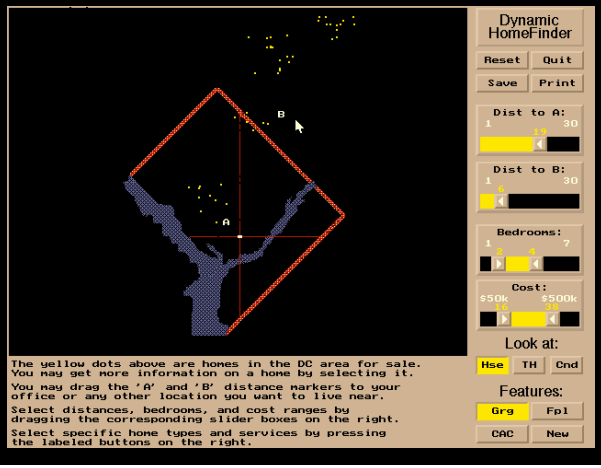
SunBurst

www.cc.gatech.edu/gvu/ii/sunburst

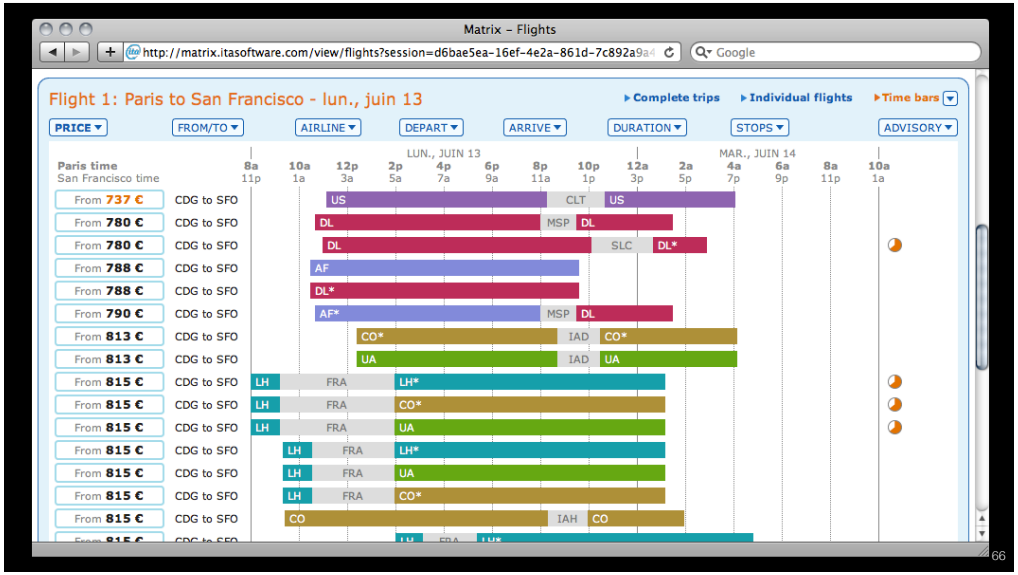


HomeFinder

HCIL
U of
Maryland

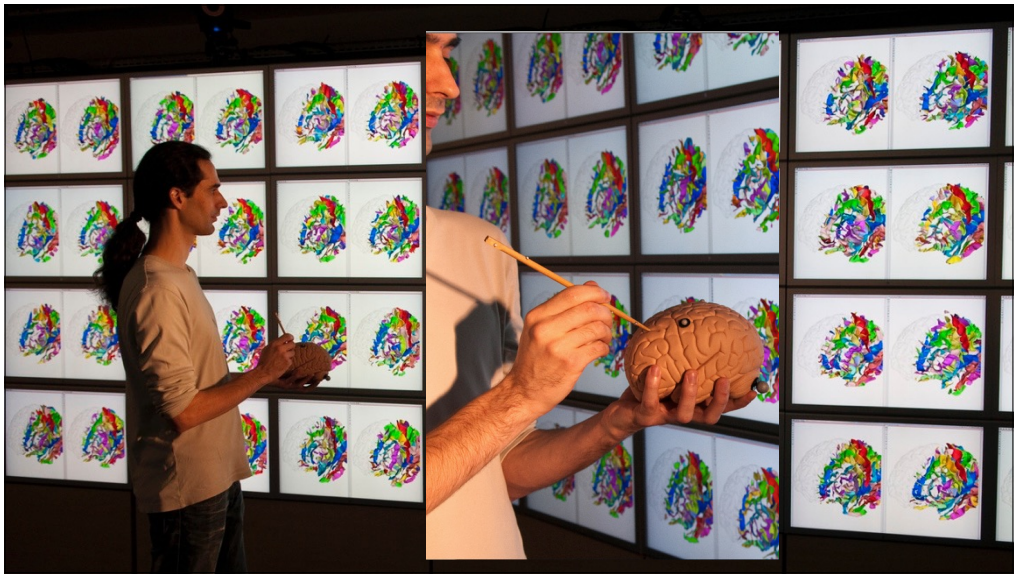


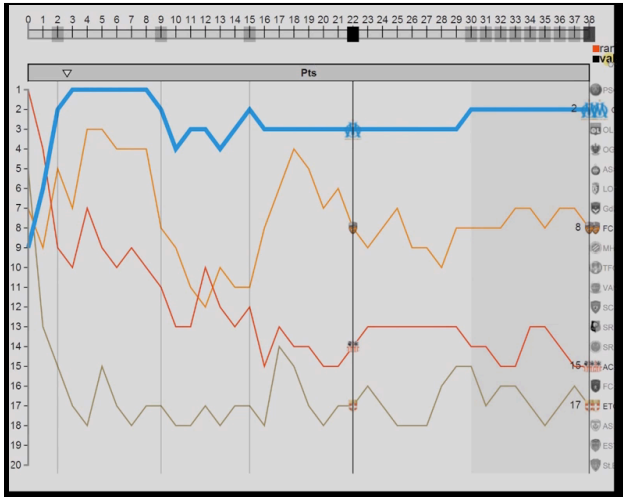
65



66

67





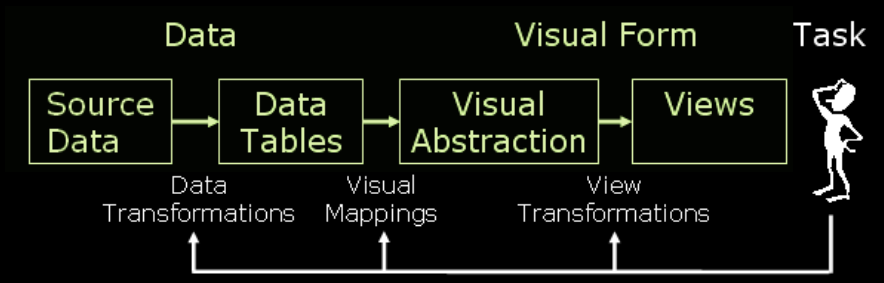
VIZ-RANK:

temporal exploration using transient line charts of team ranks

1- START: mouse click
2- columns widening
3- line-chart format
4- inspector
5- absolute and relative scales
6- END: mouse click

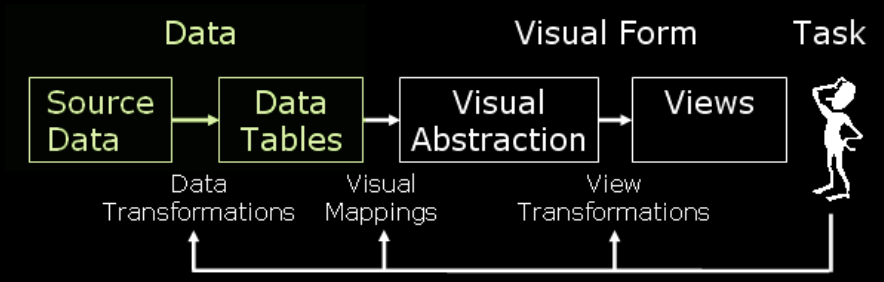
5-switching between Value mode (absolute) and Rank mode (Relative) and selecting teams

InfoVis Pipeline



Data, Marks & Spaces

InfoVis Pipeline



Data Sets

- Data comes in many different forms
- Typically, not in the way you want it
- How is stored (in the raw)?

Example

- Cars
  - make
  - model
  - year
  - miles per gallon
  - cost
  - number of cylinders
  - weights
  - ...

Data Tables

- Often, we take raw data and transform it into a form that is more workable
- Main idea:
  - Individual items are called cases
  - Cases have variables (attributes)

Data Table Format

Dimensions

	$Variable_1$	$Variable_2$	$Variable_3$	$Variable_4$	...
$Case_1$	$Value_{1,1}$	$Value_{1,2}$	$Value_{1,3}$	$Value_{1,4}$	
$Case_2$	$Value_{2,1}$	$Value_{2,2}$	$Value_{2,3}$	$Value_{2,4}$	
$Case_3$	$Value_{3,1}$	$Value_{3,2}$	$Value_{3,3}$	$Value_{3,4}$	
$Case_4$	$Value_{4,1}$	$Value_{4,2}$	$Value_{4,3}$	$Value_{4,4}$	
⋮					

Think of as a function:

$f(case_i) = \langle value_{i,1}, value_{i,2}, \ldots, value_{i,n} \rangle$

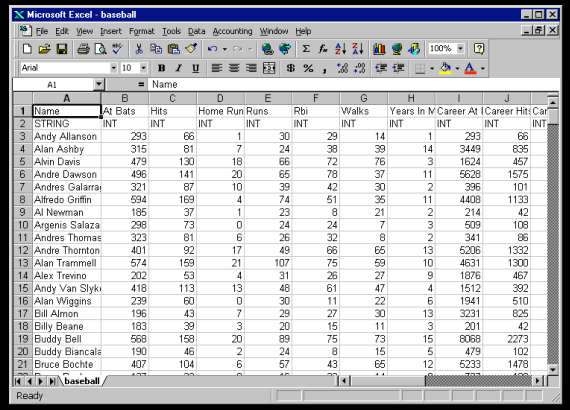
Data Table Example

People in Class

	Hair	Age	GPA	ID
Marie	brown	23	12,3	901-12-3456
Jean	black	17	14,6	901-12-4567
Henri	blond	47	10,2	901-12-5678
Bob	red	29	11,8	901-12-6789
⋮				

Example

Baseball Statistics



Variable Types

- Three main types of variables
  - N-Nominal (equal or not equal to other values)
    - Example: gender
  - O-Ordinal (obeys < relation, ordered set)
    - Example: fr,so,jr,sr
  - Q-Quantitative (can do math on them)
    - Can be *absolute* or *relative*
    - Example: age, temperature

Metadata

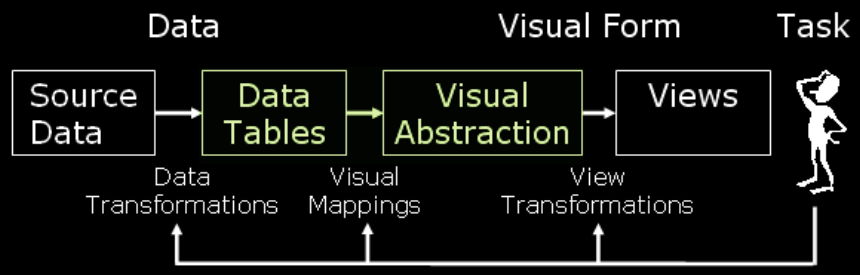
- Descriptive information about the data
  - Might be something as simple as the type of a variable, or could be more complex
  - For times when the table itself just isn't enough
  - Example:
    - if car motor is electric, then L/100km is meaningless.
    - number of home runs ≤ number of at-bats

How do we show the data?

How Many Variables?

- Data sets of dimensions 1, 2, 3 are common
- Number of variables per class
  - 1 - Univariate data
  - 2 - Bivariate data
  - 3 - Trivariate data
  - >3 - Hypervariate data

InfoVis Pipeline



Map *data* to a *representation*

- Data is abstract
- Representation is more conceptual
- Use a space
- Create an *implantation* of *data* into the space

Visual Structures

- Composed of
  - Spatial substrate
  - Marks
  - Graphical properties of marks

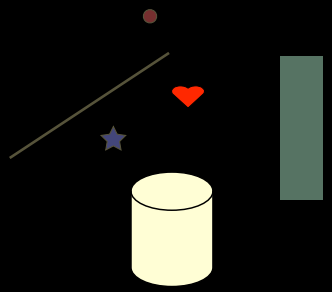
[Bertin, Sémiologie Graphique 1967]

Space

- Visually dominant
- Often put axes on space to assist
- Use techniques of composition, alignment, folding, recursion, overloading to
  1. increase use of space
  2. do data encodings

Marks

- Things that occur in space
  - Points
  - Lines
  - Areas
  - Volumes

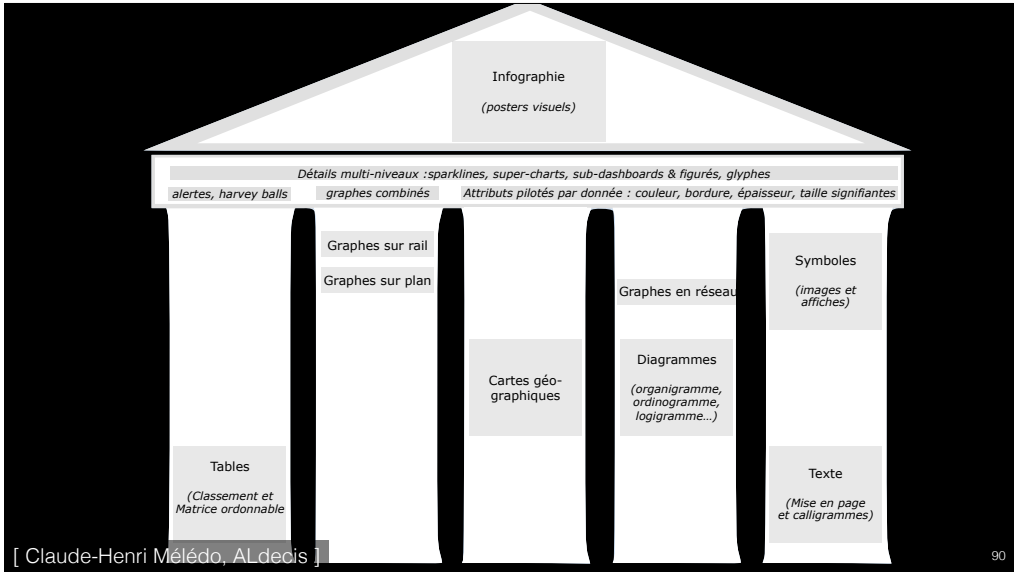


Graphical Properties

	Spatial properties	Object properties
Expressing extent	position size	greyscale
Differentiating marks	orientation	color shape texture

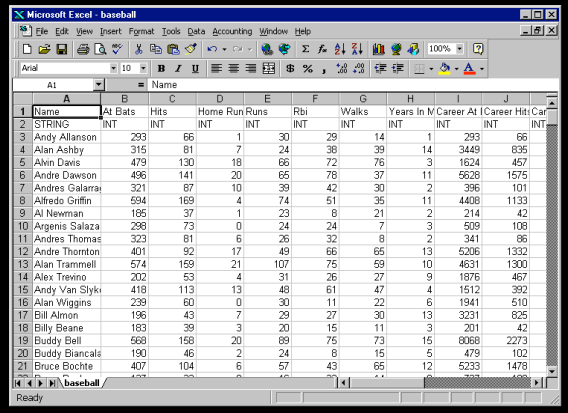
There are only 5 graphics

(sort of…)

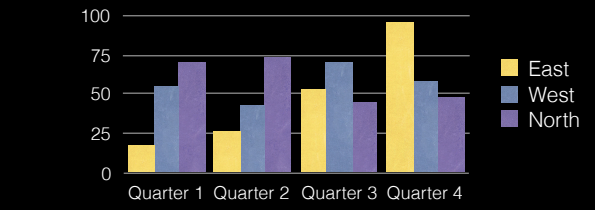


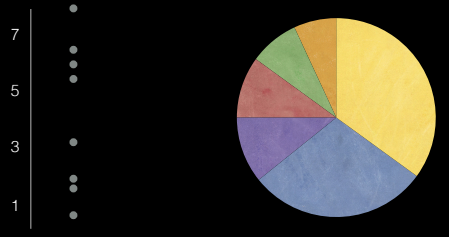
[Claude-Henri Mélédo, ALdecis]

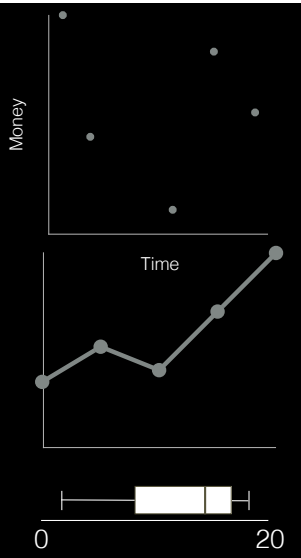
Tables



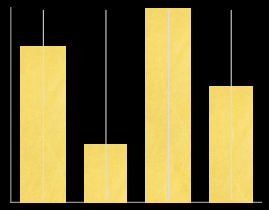
Graphs on rails



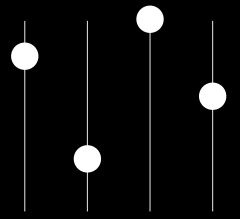




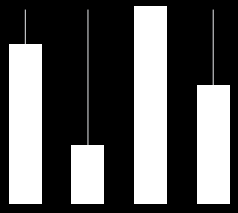
Graphs on rails



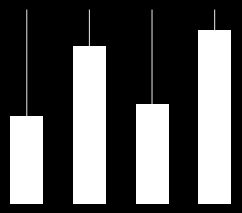
Graphs on rails



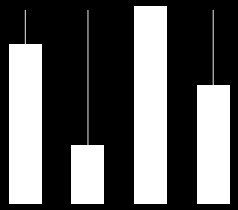
Graphs on rails



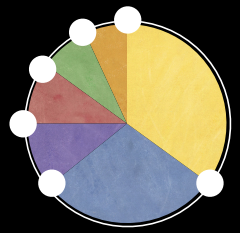
Graphs on rails



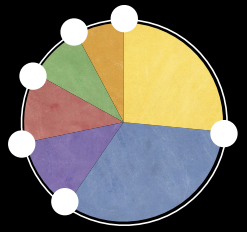
Graphs on rails



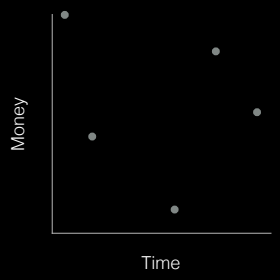
Graphs on rails

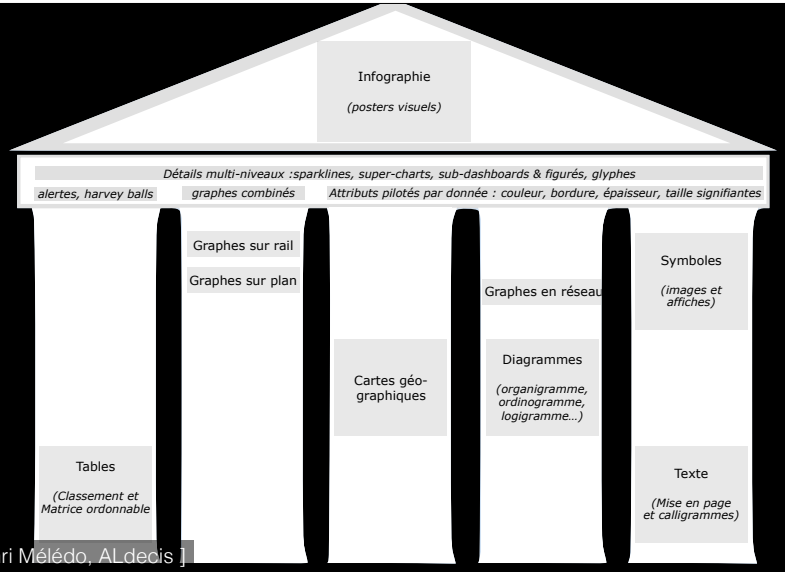


Graphs on rails



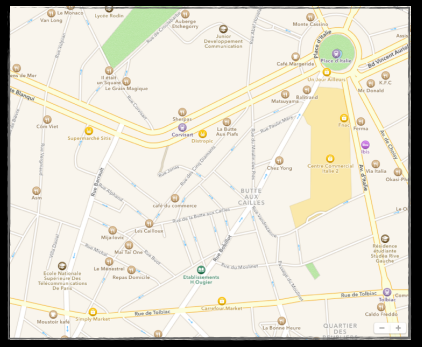
Graphs on the plane

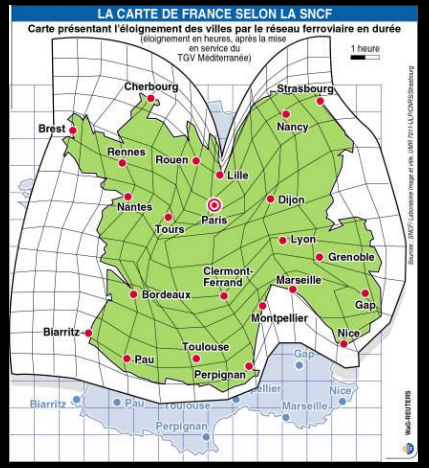




[Claude-Henri Mélédo, ALdecis]

Geospatial maps

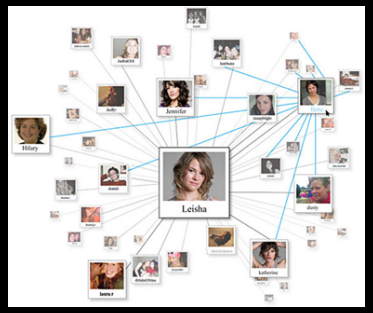


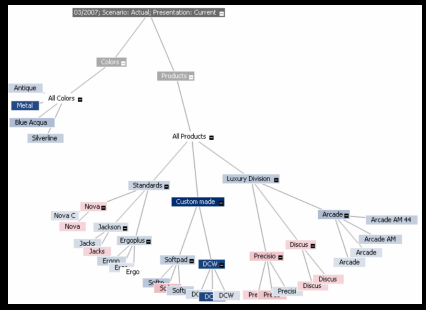


Well-studied area

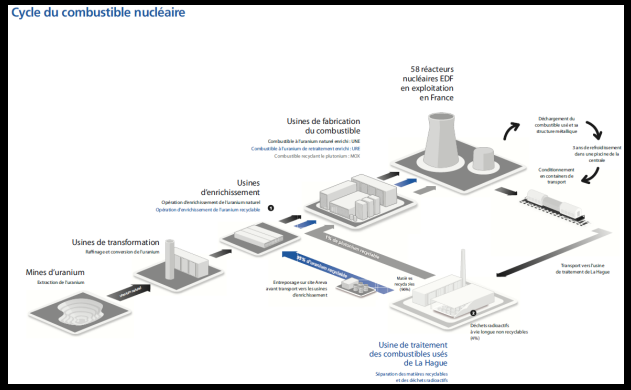
- Cartographers and map-makers have a wealth of knowledge about the design and creation of visual information artifacts
  - Labeling, color, layout, …
- Information visualization researchers should learn from this older, existing area

Networks & Diagrams

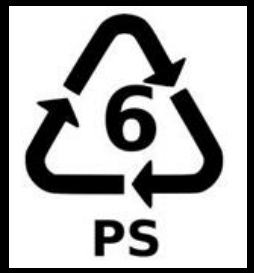


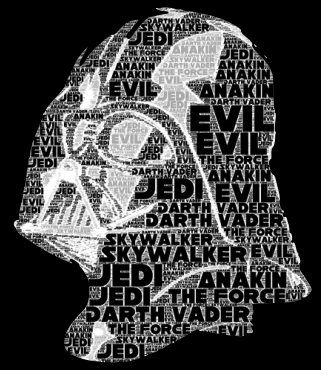


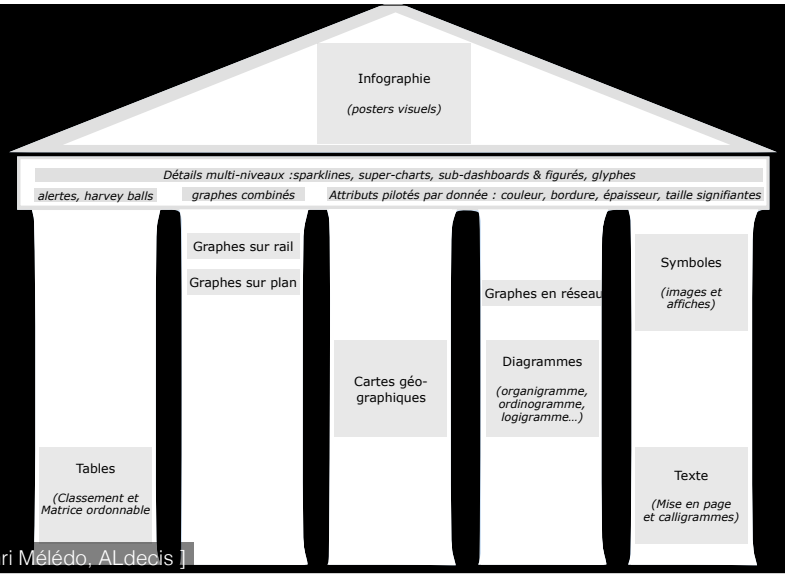
Networks & Diagrams



Conceptual Images







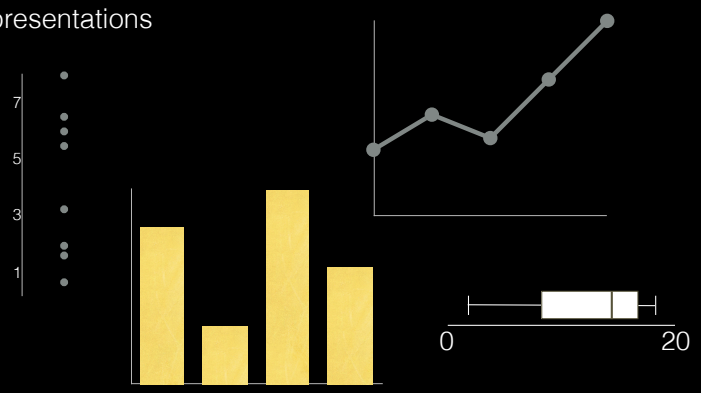
[Claude-Henri Mélédo, ALdecis]

Back to Data

- What were the different types of data sets?
- Number of variables per class
  - 1 – Univariate data
  - 2 – Bivariate data
  - 3 – Trivariate data
  - >3 – Hypervariate data

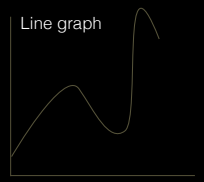
Univariate Data

- Representations

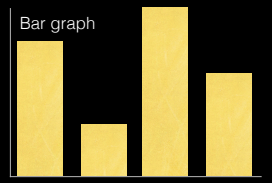


What Goes Where?

- In univariate representations, we often think of the data case as being shown along one dimension, and the value in another



- Y-axis is quantitative variable
- See changes over consecutive values



- Y-axis is quantitative variable
- Compare relative point values

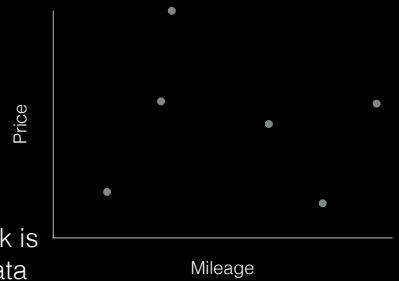
Alternative View

- We may think of graph as representing independent (data case) and dependent (value) variables
- Guideline:
  - Independent vs. dependent variables
    - Put independent on x-axis
    - See resultant dependent variables along y-axis

Bivariate Data

- Representations

Scatterplot is common



Each mark is now a data case

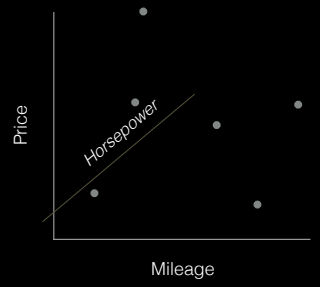
Two variables, want to see a relationship

Is there a linear, curved, or random relationship?

Trivariate Data

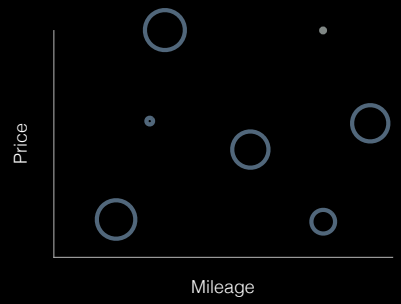
- Representations

3D scatterplot is possible



Alternative Representation

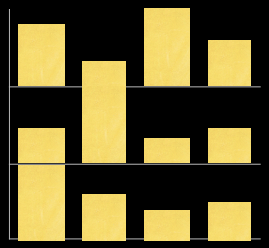
- Representations



Still use 2D but have mark property represent third variable

Alternative Representation

- Representations



Represent each variable in its own explicit way

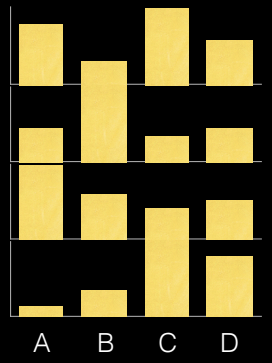
Hypervariate Data

- Ahhh, the tough one
- Number of well-known visualization techniques exist for data sets of 1-3 dimensions
  - line graphs, bar graphs, scatter plots OK
  - We see a 3-D world (4-D with time)
- What about data sets with more than 3 variables?
  - Often the interesting, challenging ones

Multiple Views

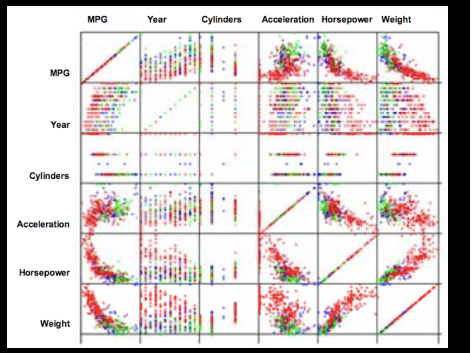
- Give each variable its own display

	A	B	C	D
1	8	3	10	6
2	4	9	3	4
3	10	6	4	5
4	1	3	9	7



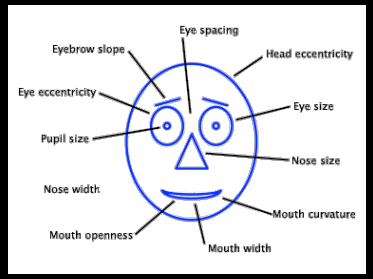
Scatterplot Matrix

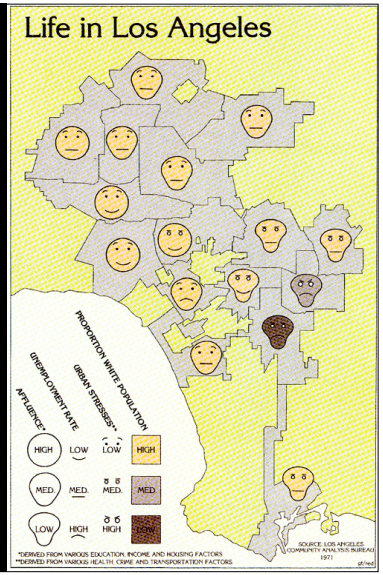
- Represent each possible pair of variables in their own 2-D scatterplot
- Useful for what?
- Misses what?



Chernoff Faces

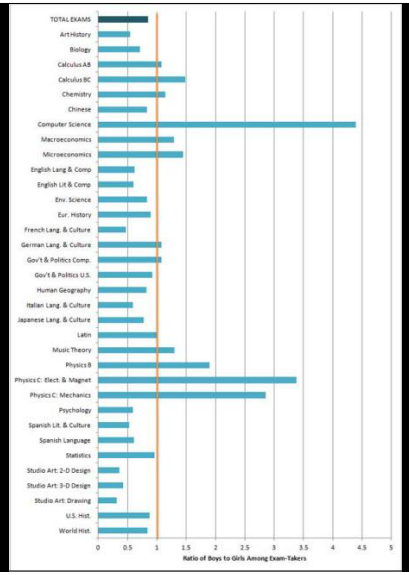
- Encode different variables' values in characteristics of human face

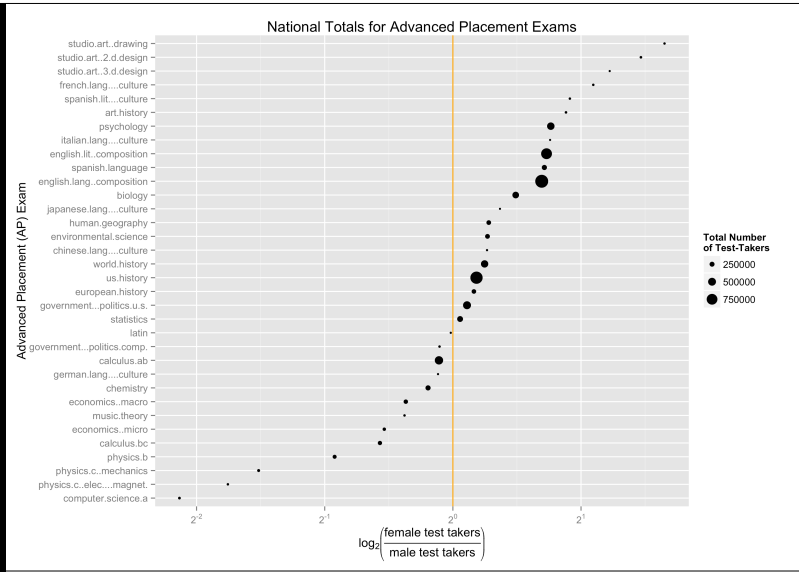




Critique du moment

Ratio of boys to girls among exam takers





So far…

- We examined a number of tried-and-true techniques/visualizations for presenting multivariate (typically ≤3) data sets
  - Hinted at how to go above 3 dimensions

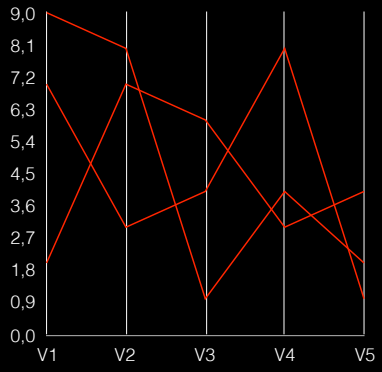
More Dimensions

- Fundamentally, we have 2 display dimensions
- For data sets with >2 variables, we must project data down to 2D
- Come up with visual mapping that locates each dimension into 2D plane
- Computer graphics 3D→2D projections

Wait a moment...

- A spreadsheet already does that
  - Each variable is positioned into a column
  - Data cases in rows
  - This is a projection (mapping)

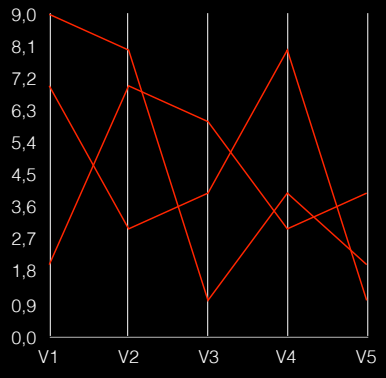
Parallel Coordinates



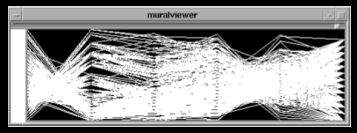
	V1	V2	V3	V4	V5
C1	7	3	4	8	1
C2	2	7	6	3	4
C3	9	8	1	4	2

Parallel Coordinates

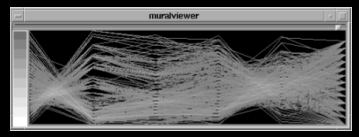
- Encode variables along a horizontal row
- Vertical line specifies different values that variable can take
- Data point represented as a polyline



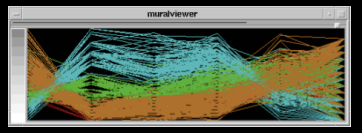
Parallel Coordinates Example



Basic



Greyscale



Color

Issue

- Different variables can have values taking on quite different ranges
- Must normalize all down (e.g., f(x): $\mathbb{N} \rightarrow [0, 1]$)

Example

- VLSI chip manufacture
- Want high quality chips (high speed) and a high yield batch (% of useful chips)
- Able to track defects
- Hypothesis: No defects gives desired chip types
- 473 batches of data

The Data

- 16 variables
- X1 — yield
- X2 — quality
- X3–X12 - # defects (inverted)
- X13–X16 - physical parameters

Parallel Coordinates

yield & quality ⟷ defects ⟷ parameters

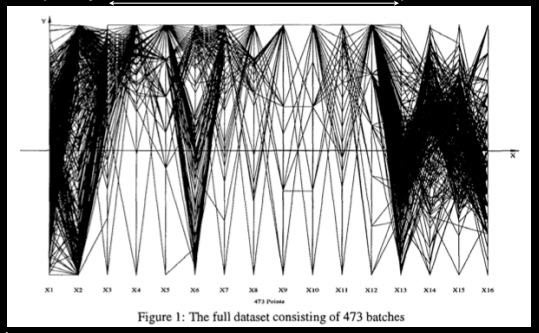


Figure 1: The full dataset consisting of 473 batches

Yikes!
But not
that bad

Distributions
x1 - normal
x2 - bipolar

Top Yield & Quality

defects

split

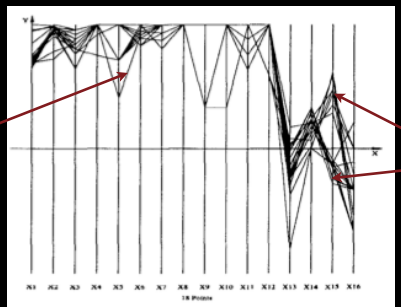


Figure 2: The batches high in Yield, $X1$, and Quality, $X2$.

Minimal Defects

Not the highest yields and quality

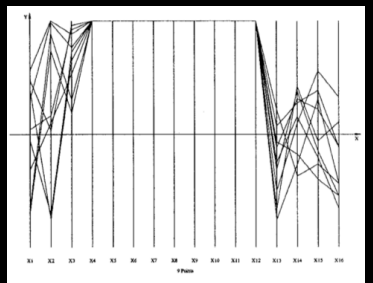


Figure 3: The batches with zero in 9 out of the ten defect types.

Best Yields

Appears that some defects are necessary to produce the best chips

Non-intuitive!

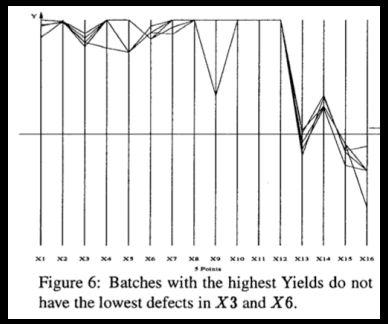
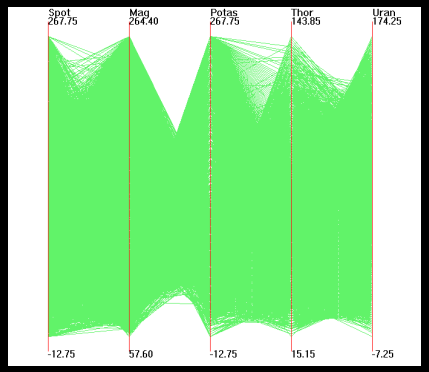


Figure 6: Batches with the highest Yields do not have the lowest defects in $X3$ and $X6$.

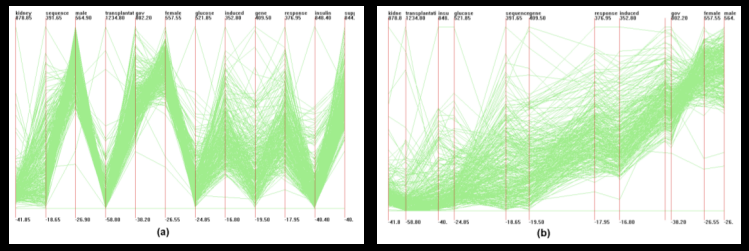
Challenges



Out5d dataset (5 dimensions, 16384 data items)

Dimensional Reordering

Which dimensions are most like each other?



Same dimensions ordered by similarity

Dimensional Reordering

Can you reduce clutter and highlight other interesting features in data?

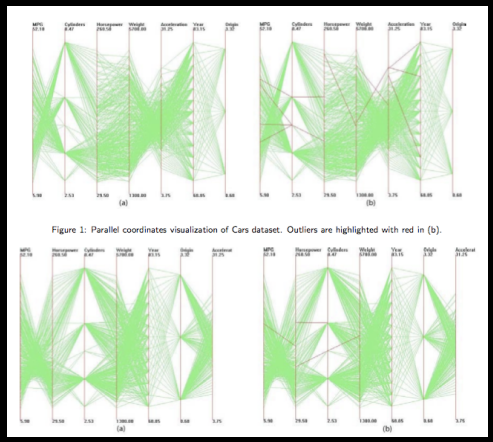
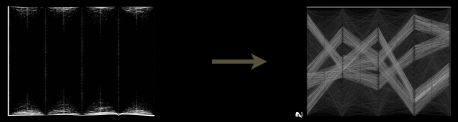
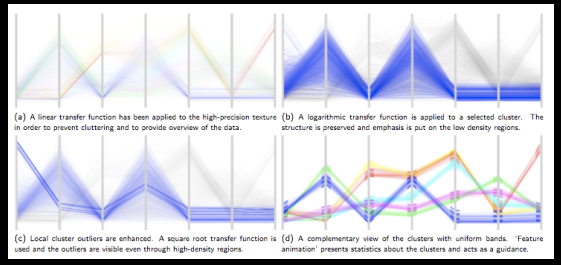


Figure 1: Parallel coordinates visualization of Cars dataset. Outliers are highlighted with red in (b).

Reducing Density



[Artero et al, InfoVis 2004]



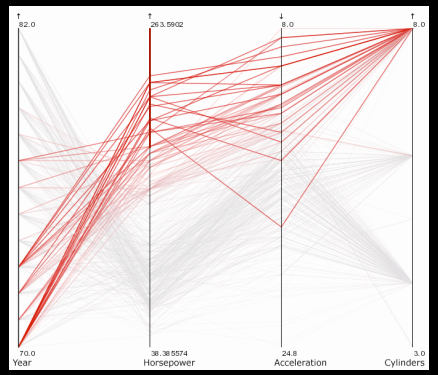
[Johansson et al, InfoVis 2005]

Querying the Display

- Provide a variety of techniques to pick out the “interesting” data points from the display

Smooth Brushing

- Specify a region of interest along one axis



Interaction