

CS448B :: 23 Sep 2009

# Data and Image Models



Jeffrey Heer Stanford University

## Last Time: Value of Visualization

### Three functions of visualizations

**Record:** store information

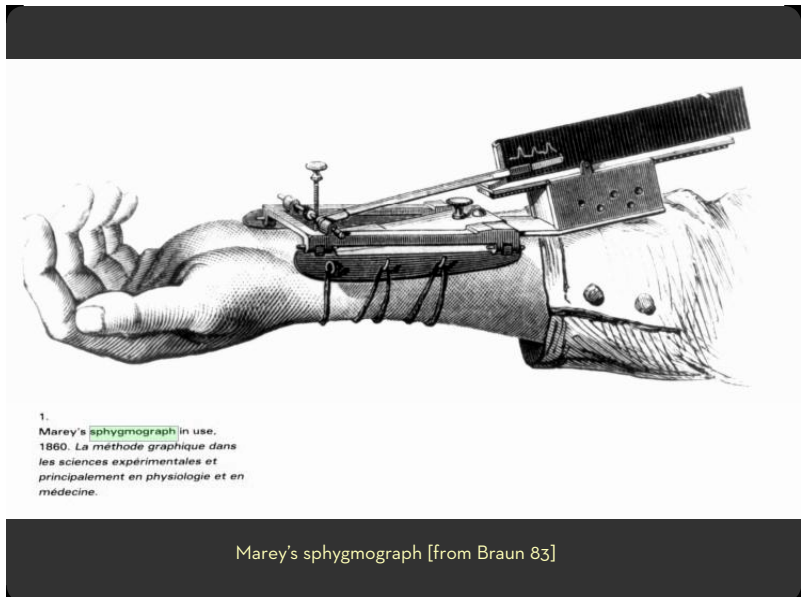
- Photographs, blueprints, ...

**Analyze:** support reasoning about information

- Process and calculate
- Reason about data
- Feedback and interaction

**Communicate:** convey information to others

- Share and persuade
- Collaborate and revise
- Emphasize important aspects of data

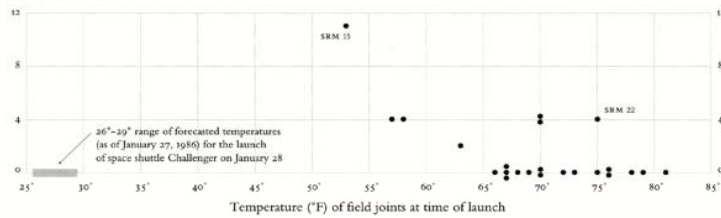


1.  
Marey's sphygmograph in use.  
1860. *La methode graphique dans  
les sciences experimentales et  
principalement en physiologie et en  
medecine.*

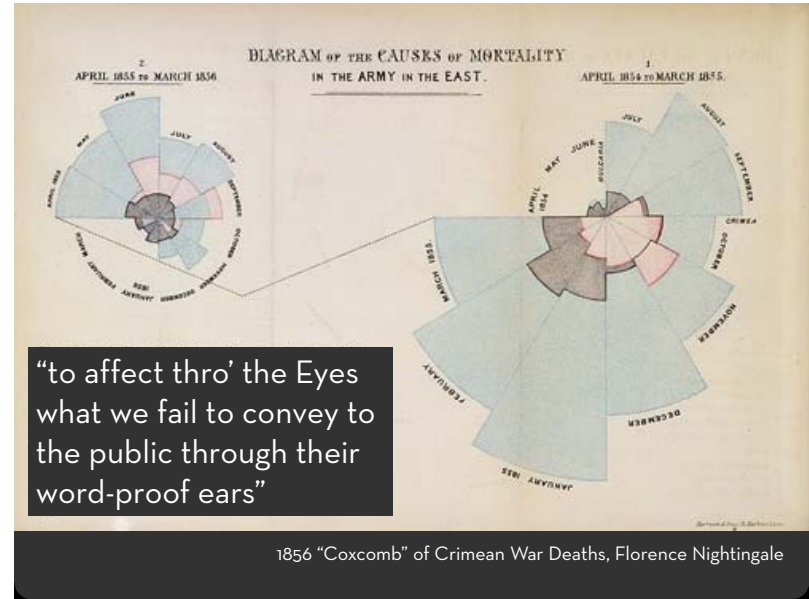
Marey's sphygmograph [from Braun 83]

# Make a decision: Challenger

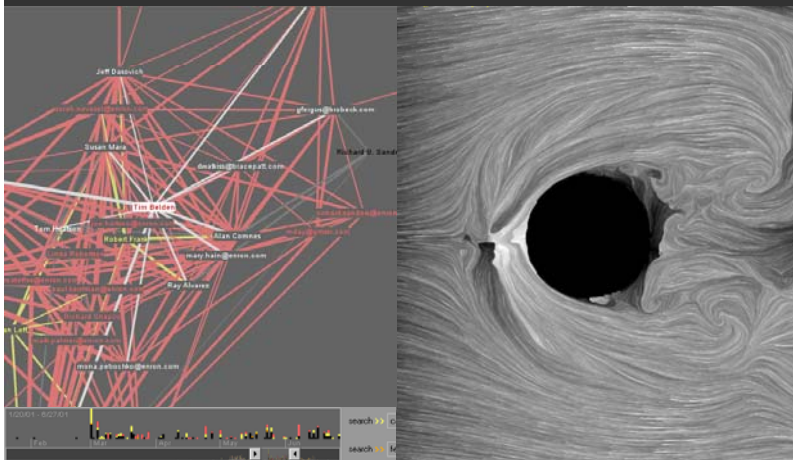
O-ring damage index, each launch



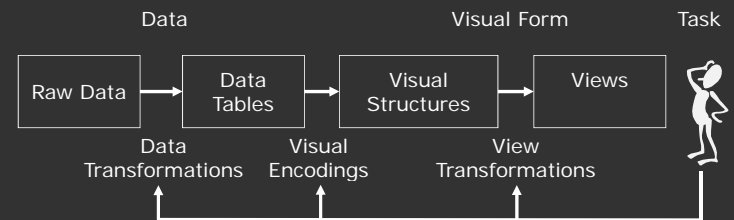
Visualizations drawn by Tufte show how low temperatures damage O-rings [Tufte 97]



# Info-Vis vs. Sci-Vis?



# Visualization Reference Model



## Data and Image Models

## The Big Picture

task

data

physical type  
int, float, etc.  
abstract type  
nominal, ordinal, etc.

domain

metadata  
semantics  
conceptual model

processing  
algorithms

mapping  
visual encoding  
visual metaphor

image

visual channel  
retinal variables

## Topics

Properties of data or information

Properties of the image

Mapping data to images

## Data

## Data models vs. Conceptual models

Data models are low level descriptions of the data

- Math: Sets with operations on them
- Example: integers with + and × operators

Conceptual models are mental constructions

- Include semantics and support reasoning

Examples (data vs. conceptual)

- (1D floats) vs. Temperature
- (3D vector of floats) vs. Space

## Taxonomy (?)

- 1D (sets and sequences)
- Temporal
- 2D (maps)
- 3D (shapes)
- nD (relational)
- Trees (hierarchies)
- Networks (graphs)
- Are there others?

The eyes have it: A task by data type taxonomy for information visualization [Shneiderman 96]

## Types of variables

Physical types

- Characterized by storage format
  - Characterized by machine operations
- Example: bool, short, int32, float, double, string, ...

Abstract types

- Provide descriptions of the data
  - May be characterized by methods/attributes
  - May be organized into a hierarchy
- Example: plants, animals, metazoans, ...

## Nominal, Ordinal and Quantitative

N - Nominal (labels)

- Fruits: Apples, oranges, ...

O - Ordered

- Quality of meat: Grade A, AA, AAA

Q - Interval (Location of zero arbitrary)

- Dates: Jan, 19, 2006; Location: (LAT 33.98, LONG -118.45)
- Like a geometric point. Cannot compare directly
- Only differences (i.e. intervals) may be compared

Q - Ratio (zero fixed)

- Physical measurement: Length, Mass, Temp, ...
- Counts and amounts
- Like a geometric vector, origin is meaningful

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

# Nominal, Ordinal and Quantitative

N - Nominal (labels)

- Operations: =, ≠

O - Ordered

- Operations: =, ≠, <, >

Q - Interval (Location of zero arbitrary)

- Operations: =, ≠, <, >, -
- Can measure distances or spans

Q - Ratio (zero fixed)

- Operations: =, ≠, <, >, -, ÷
- Can measure ratios or proportions

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

# From data model to N,O,Q data type

Data model

- 32.5, 54.0, -17.3, ...
- floats

Conceptual model

- Temperature (°C)

Data type

- Burned vs. Not burned (N)
- Hot, warm, cold (O)
- Continuous range of values (Q)

ID	Case	Species_No	Species	Organ	Width	Length
1	1	1	I. Setosa	Petal	2	14
2	2	1	3 I. Versicolour	Petal	24	56
3	3	1	2 I. Versicolour	Petal	13	45
4	4	1	1 I. Setosa	Sepal	33	50
5	5	1	3 I. Versicolour	Sepal	31	67
6	6	1	2 I. Versicolour	Sepal	20	57
7	7	2	1 I. Setosa	Petal	2	10
8	8	2	3 I. Versicolour	Petal	23	51
9	9	2	2 I. Versicolour	Petal	16	47
10	10	2	1 I. Setosa	Sepal	36	46
11	11	2	3 I. Versicolour	Sepal	31	69
12	12	2	2 I. Versicolour	Sepal	33	63
13	13	3	1 I. Setosa	Petal	2	16
14	14	3	3 I. Versicolour	Petal	20	52
15	15	3	2 I. Versicolour	Petal	14	47
16	16	3	1 I. Setosa	Sepal	31	48
17	17	3	3 I. Versicolour	Sepal	30	65
18	18	3	2 I. Versicolour	Sepal	32	70
19	19	4	1 I. Setosa	Petal	1	14
20	20	4	3 I. Versicolour	Petal	19	51
21	21	4	2 I. Versicolour	Petal	12	40
22	22	4	1 I. Setosa	Sepal	36	49
23	23	4	3 I. Versicolour	Sepal	27	58
24	24	4	2 I. Versicolour	Sepal	26	50
25	25	5	1 I. Setosa	Petal	2	13
26	26	5	3 I. Versicolour	Petal	17	45
27	27	5	2 I. Versicolour	Petal	10	33
28	28	5	1 I. Setosa	Sepal	32	44
29	29	5	3 I. Versicolour	Sepal	25	49
30	30	5	2 I. Versicolour	Sepal	23	50
31	31	6	1 I. Setosa	Petal	2	16
32	32	6	3 I. Versicolour	Petal	23	59
33	33	6	2 I. Versicolour	Petal	16	46
34	34	6	1 I. Setosa	Sepal	30	44
35	35	6	3 I. Versicolour	Sepal	26	49
36	36	6	2 I. Versicolour	Sepal	23	50
37	37	6	1 I. Setosa	Petal	2	16

Sepal and petal lengths and widths for three species of iris [Fisher 1936].

ID	Case	Species_No	Species	Organ	Width	Length
1	1	1	I. Setosa	Petal	2	14
2	2	1	3 I. Versicolour	Petal	24	56
3	3	1	2 I. Versicolour	Petal	13	45
4	4	1	1 I. Setosa	Sepal	33	50
5	5	1	3 I. Versicolour	Sepal	31	67
6	6	1	2 I. Versicolour	Sepal	20	57
7	7	2	1 I. Setosa	Petal	2	10
8	8	2	3 I. Versicolour	Petal	23	51
9	9	2	2 I. Versicolour	Petal	16	47
10	10	2	1 I. Setosa	Sepal	36	46
11	11	2	3 I. Versicolour	Sepal	31	69
12	12	2	2 I. Versicolour	Sepal	33	63
13	13	3	1 I. Setosa	Petal	2	16
14	14	3	3 I. Versicolour	Petal	20	52
15	15	3	2 I. Versicolour	Petal	14	47
16	16	3	1 I. Setosa	Sepal	31	48
17	17	3	3 I. Versicolour	Sepal	30	65
18	18	3	2 I. Versicolour	Sepal	32	70
19	19	4	1 I. Setosa	Petal	1	14
20	20	4	3 I. Versicolour	Petal	19	51
21	21	4	2 I. Versicolour	Petal	12	40
22	22	4	1 I. Setosa	Sepal	36	49
23	23	4	3 I. Versicolour	Sepal	27	58
24	24	4	2 I. Versicolour	Sepal	26	50
25	25	5	1 I. Setosa	Petal	2	13
26	26	5	3 I. Versicolour	Petal	17	45
27	27	5	2 I. Versicolour	Petal	10	33
28	28	5	1 I. Setosa	Sepal	32	44
29	29	5	3 I. Versicolour	Sepal	25	49
30	30	5	2 I. Versicolour	Sepal	23	50
31	31	6	1 I. Setosa	Petal	2	16
32	32	6	3 I. Versicolour	Petal	23	59
33	33	6	2 I. Versicolour	Petal	16	46
34	34	6	1 I. Setosa	Sepal	30	44
35	35	6	3 I. Versicolour	Sepal	26	49
36	36	6	2 I. Versicolour	Sepal	23	50
37	37	6	1 I. Setosa	Petal	2	16

Legend: N (Nominal), O (Ordinal), Q (Quantitative)

## Relational data model

- Records are fixed-length tuples
- Each column (attribute) of tuple has a domain (type)
- Relation is schema and a table of tuples
- Database is a collection of relations

## Relational Algebra [Codd]

- Data transformations (SQL)
- Selection (SELECT)
- Projection (WHERE)
- Sorting (ORDER BY)
- Aggregation (GROUP BY, SUM, MIN, ...)
- Set operations (UNION, ...)
- Join (INNER JOIN)

## Statistical data model

- Variables or measurements
- Categories or factors or dimensions
- Observations or cases

## Statistical data model

- Variables or measurements
- Categories or factors or dimensions
- Observations or cases

Month	Control	Placebo	300 mg	450 mg
March	165	163	166	168
April	162	159	161	163
May	164	158	161	153
June	162	161	158	160
July	166	158	160	148
August	163	158	157	150

Blood Pressure Study (4 treatments, 6 months)

## Dimensions and Measures

Independent vs. dependent variables

- Example:  $y = f(x, a)$
- Dimensions:  $\text{Domain}(x) \times \text{Domain}(a)$
- Measures:  $\text{Range}(y)$

## Dimensions and Measures

**Dimensions:** Discrete variables describing data  
Dates, categories of values (independent vars)

**Measures:** Data values that can be aggregated  
Numbers to be analyzed (dependent vars)  
Aggregate as sum, count, average, std. deviation

## Example: U.S. Census Data

**People:** # of people in group

**Year:** 1850 – 2000 (every decade)

**Age:** 0 – 90+

**Sex:** Male, Female

**Marital Status:** Single, Married, Divorced, ...

## Example: U.S. Census

**People**

**Year**

**Age**

**Sex**

**Marital Status**

2348 data points

	A	B	C	D	E
1	year	age	marst	sex	people
2	1850	0	0	1	1483789
3	1850	0	0	2	1450374
4	1850	5	0	1	1411067
5	1850	5	0	2	1359668
6	1850	10	0	1	1260099
7	1850	10	0	2	1216114
8	1850	15	0	1	1077193
9	1850	15	0	2	1110619
10	1850	20	0	1	1017281
11	1850	20	0	2	1005841
12	1850	25	0	1	862547
13	1850	25	0	2	799482
14	1850	30	0	1	730038
15	1850	30	0	2	639036
16	1850	35	0	1	588487
17	1850	35	0	2	505012
18	1850	40	0	1	475913
19	1850	40	0	2	428180
20	1850	45	0	1	384211
21	1850	45	0	2	341254
22	1850	50	0	1	321343
23	1850	50	0	2	286580
24	1850	55	0	1	194080
25	1850	55	0	2	187208
26	1850	60	0	1	174976
27	1850	60	0	2	162126
28	1850	65	0	1	106827
29	1850	65	0	2	105554
30	1850	70	0	1	78677
31	1850	70	0	2	71763
32	1850	75	0	1	40884
33	1850	75	0	2	40229
34	1850	80	0	1	23449
35	1850	80	0	2	22949
36	1850	85	0	1	8186
37	1850	85	0	2	10511
38	1850	90	0	1	5159
39	1850	90	0	2	6389
40	1860	0	0	1	2120846
41	1860	0	0	2	2092162

## Census: Dimension or Measure?

<b>People Count</b>	Measure
<b>Year</b>	Dimension
<b>Age</b>	Depends!
<b>Sex (M/F)</b>	Dimension
<b>Marital Status</b>	Dimension

## Roll-Up and Drill-Down

Want to examine marital status in each decade?

**Roll-up** the data along the desired dimensions

```
SELECT  year, marst,  sum(people)
FROM census
GROUP BY  year, marst;
```

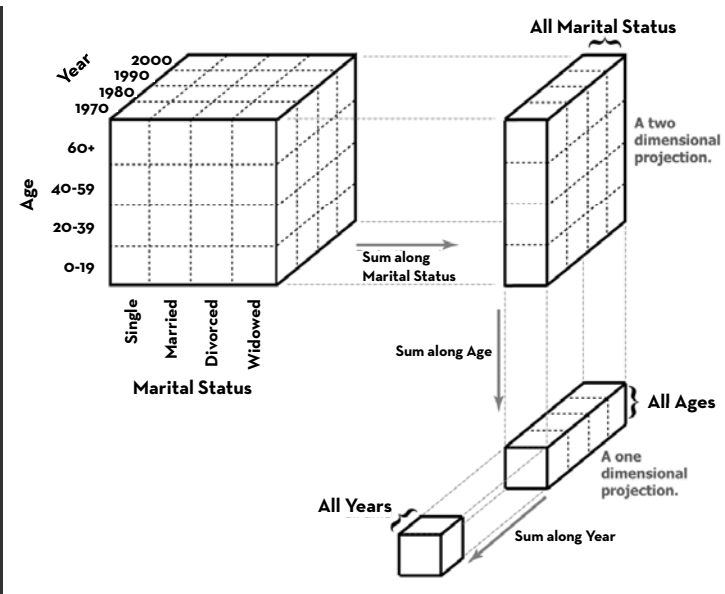
Dimensions

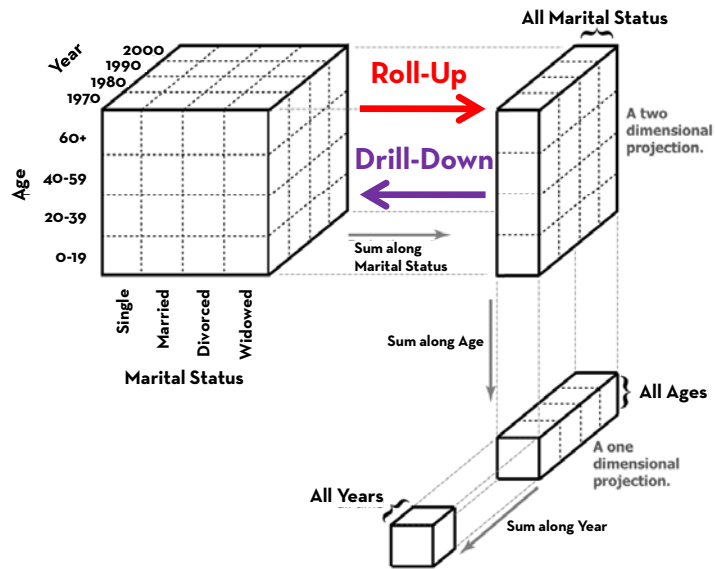
## Roll-Up and Drill-Down

Need more detailed information?

**Drill-down** into additional dimensions

```
SELECT year, age, marst, sum(people)
FROM census
GROUP BY year, age, marst;
```





## Row vs. Column-Oriented Databases

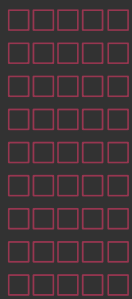
## Relational Data Organizations

Transactions

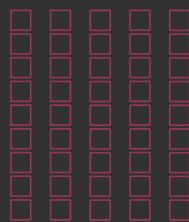
vs.

Analysis

Row-oriented

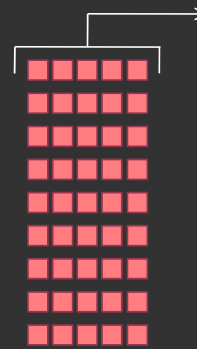


Column-oriented

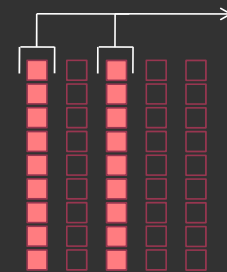


## Relational Data Organizations

Row-oriented



Column-oriented

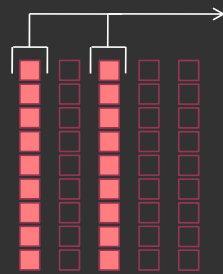


## Relational Data Organizations

### Speed-up Analysis

- Reduce data transfer
- Improved locality
- Better data compression

Column-oriented



## Administrivia

## Announcements

### Auditors

- Requirements: Come to class and participate (online as well)

### Class participation requirements

- Complete readings before class
- In-class discussion
- Post at least 1 discussion substantive comment/question on wiki within a week of each lecture

**Class wiki:** <http://cs448b.stanford.edu>

## Assignment 1: Visualization Design

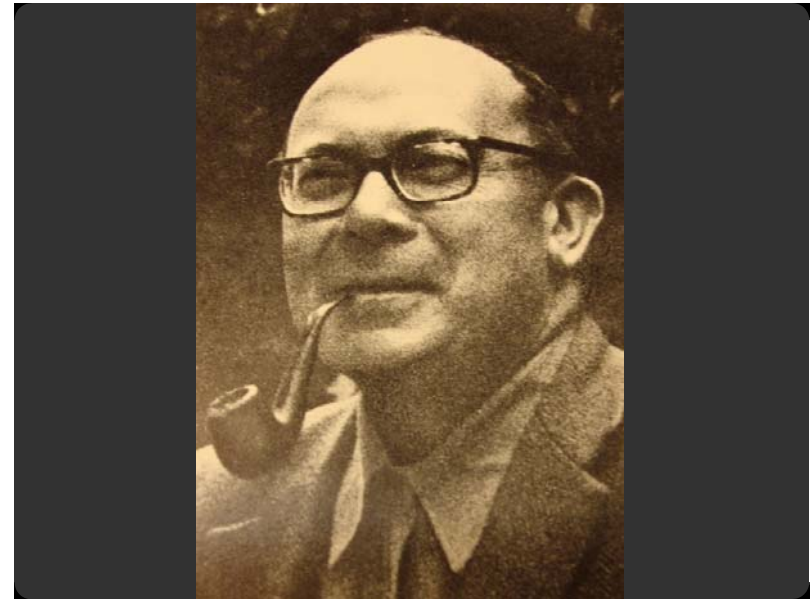
Design a static visualization for a given data set.

Deliverables (post to the course wiki)

- Image of your visualization
- Short description and design rationale ( $\leq 4$  para.)

Due by **7:00am** on **Monday 9/28**.

# Image



## Visual language is a sign system

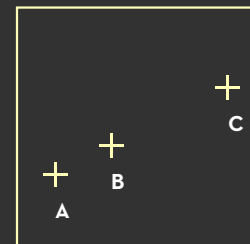


Jacques Bertin

Images perceived as a set of signs  
Sender encodes information in signs  
Receiver decodes information from signs

Sémiologie Graphique, 1967

## Bertin's Semiology of Graphics



1. A, B, C are distinguishable
2. B is between A and C.
3. BC is twice as long as AB.

∴ Encode quantitative variables

*"Resemblance, order and proportion are the three signifieds in graphics."* - Bertin

### LES VARIABLES DE L'IMAGE

	POINTS			LIGNES		ZONES	
XY 2 DIMENSIONS DU PLAN	x	x	x	/	?	/	?
Z TAILLE	█	█	█	█	█	█	█
VALEUR	█	█	█	█	█	█	█
LES VARIABLES DE SÉPARATION DES IMAGES							
GRAIN	█	█	█	█	█	█	█
COULEUR	█	█	█	█	█	█	█
ORIENTATION	█	█	█	█	█	█	█
FORME	█	█	█	█	█	█	█

## Visual encoding variables

Position (x 2)  
Size  
Value  
Texture  
Color  
Orientation  
Shape

### LES VARIABLES DE L'IMAGE

	POINTS			LIGNES		ZONES	
XY 2 DIMENSIONS DU PLAN	x	x	x	/	?	/	?
Z TAILLE	█	█	█	█	█	█	█
VALEUR	█	█	█	█	█	█	█
LES VARIABLES DE SÉPARATION DES IMAGES							
GRAIN	█	█	█	█	█	█	█
COULEUR	█	█	█	█	█	█	█
ORIENTATION	█	█	█	█	█	█	█
FORME	█	█	█	█	█	█	█

## Visual encoding variables

Position  
Length  
Area  
Volume  
Value  
Texture  
Color  
Orientation  
Shape  
Transparency  
Blur / Focus ...

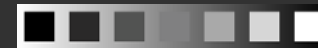
### LES VARIABLES DE L'IMAGE

	POINTS			LIGNES		ZONES	
XY 2 DIMENSIONS DU PLAN	x	x	x	/	?	/	?
Z TAILLE	█	█	█	█	█	█	█
VALEUR	█	█	█	█	█	█	█
LES VARIABLES DE SÉPARATION DES IMAGES							
GRAIN	█	█	█	█	█	█	█
COULEUR	█	█	█	█	█	█	█
ORIENTATION	█	█	█	█	█	█	█
FORME	█	█	█	█	█	█	█

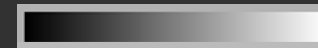
## Information in color and value

Value is perceived as ordered

∴ Encode ordinal variables (O)



∴ Encode continuous variables (Q) [not as well]



Hue is normally perceived as unordered

∴ Encode nominal variables (N) using color



## Bertin's "Levels of Organization"

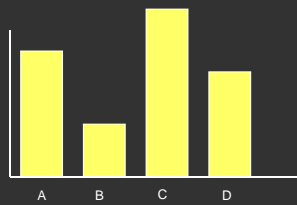
Position	N	O	Q
Size	N	O	Q
Value	N	O	q
Texture	N	o	
Color	N		
Orientation	N		
Shape	N		

Nominal  
 Ordered  
 Quantitative  
 Note: Q < O < N

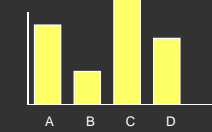
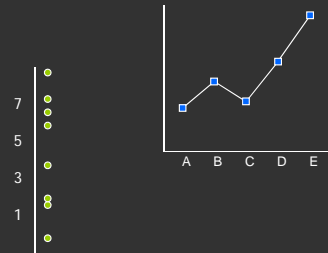
Note: Bertin actually breaks visual variables down into differentiating ( $\neq$ ) and associating ( $\equiv$ )

## Design Space of Visual Encodings

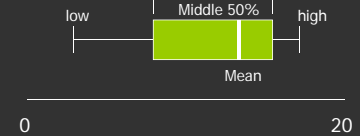
## Univariate data



## Univariate data



Tukey box plot



## Bivariate data

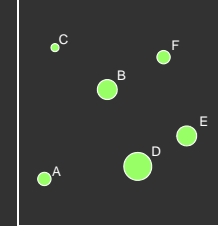
	A	B	C
1			
2			



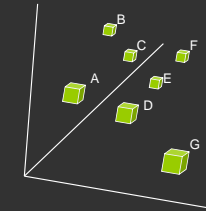
Scatter plot is common

## Trivariate data

	A	B	C
1			
2			
3			



3D scatter plot is possible



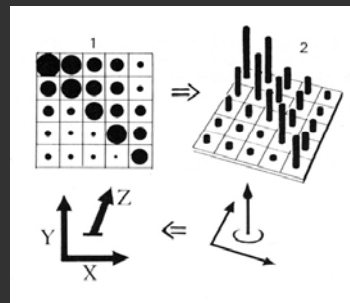
## Three variables

Two variables [x,y] can map to points

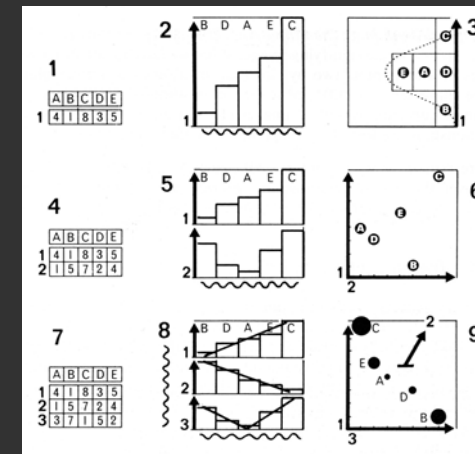
- Scatterplots, maps, ...

Third variable [z] must use

- Color, size, shape, ...



## Large design space (visual metaphors)



[Bertin, Graphics and Graphic Info. Processing, 1981]

## Multidimensional data

How many variables can be depicted in an image?

	A	B	C
1			
2			
3			
4			
5			
6			
7			
8			

## Multidimensional data

How many variables can be depicted in an image?

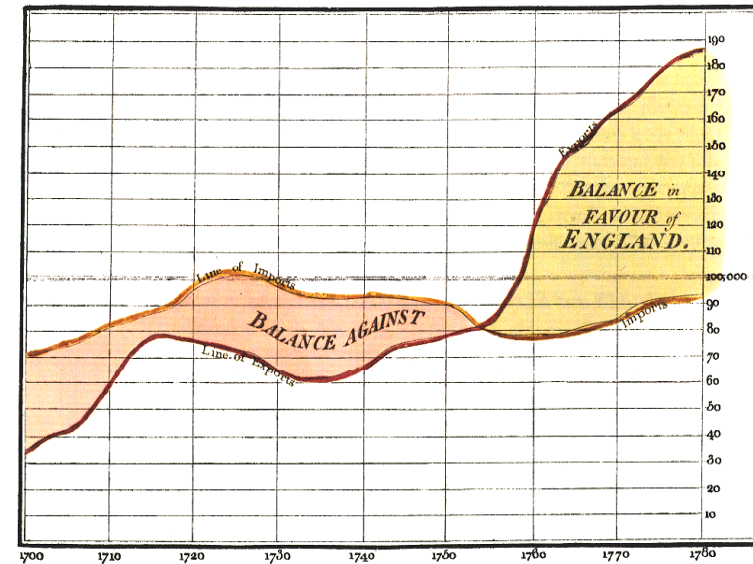
	A	B	C
1			
2			
3			
4			
5			
6			
7			
8			

*"With up to three rows, a data table can be constructed directly as a single image ... However, an image has only three dimensions. And this barrier is impassible."*

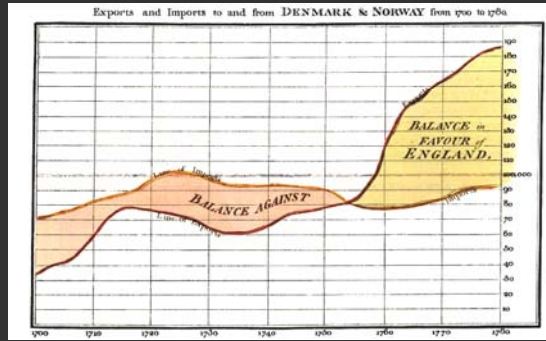
Bertin

## Deconstructions

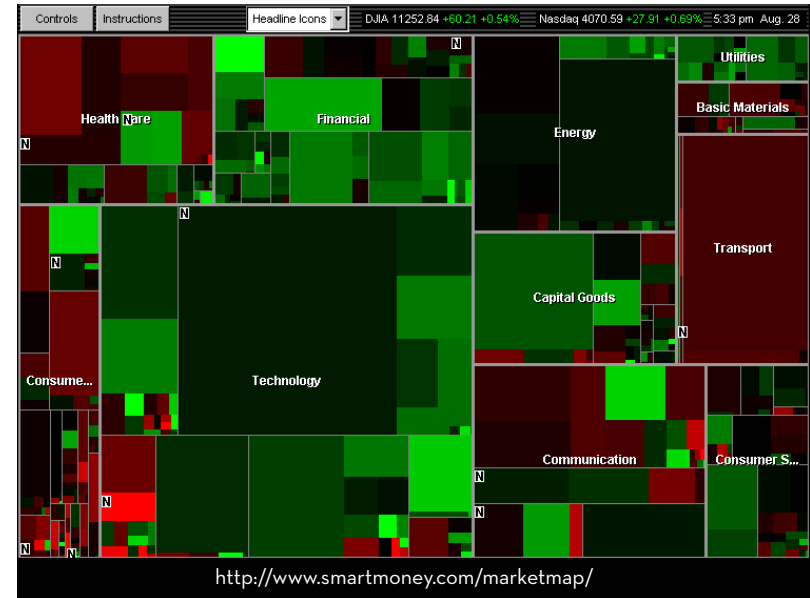
Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780



# Playfair 1786



x-axis: year (Q)  
 y-axis: currency (Q)  
 color: imports/exports (N, O)

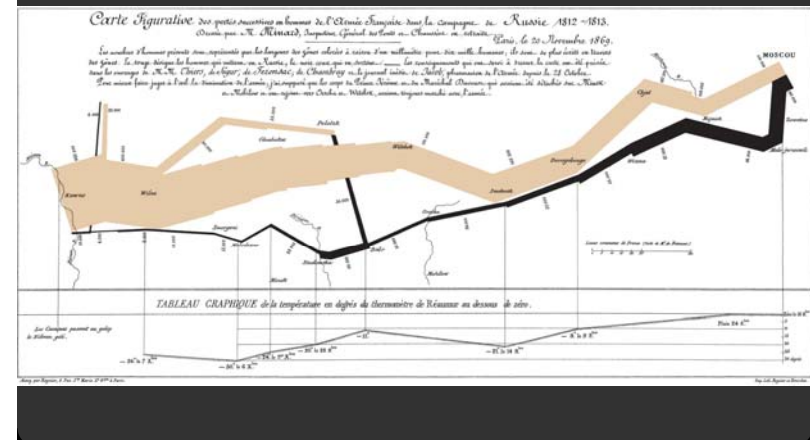


# Wattenberg 1998

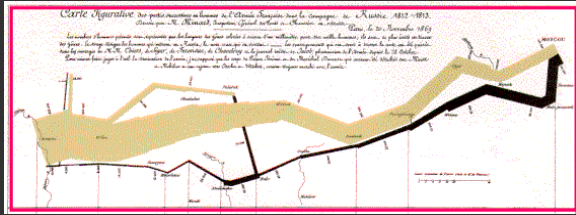


rectangle size: market cap (Q)  
 rectangle position: market sector (N), market cap (Q)  
 color hue: loss vs. gain (N, O)  
 color value: magnitude of loss or gain (Q)

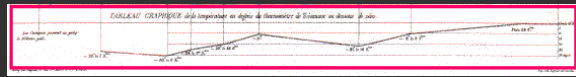
# Minard 1869: Napoleon's march



# Single axis composition



+



=



[based on slide from Mackinlay]

# Mark composition

y-axis: temperature (Q)

+

x-axis: longitude (Q) / time (O)

=



temp over space/time (Q x Q)

[based on slide from Mackinlay]

# Mark composition

y-axis: longitude (Q)

+

x-axis: latitude (Q)

+

width: army size (Q)

=



army position (Q x Q) and army size (Q)

[based on slide from Mackinlay]

longitude (Q)

latitude (Q)

army size (Q)

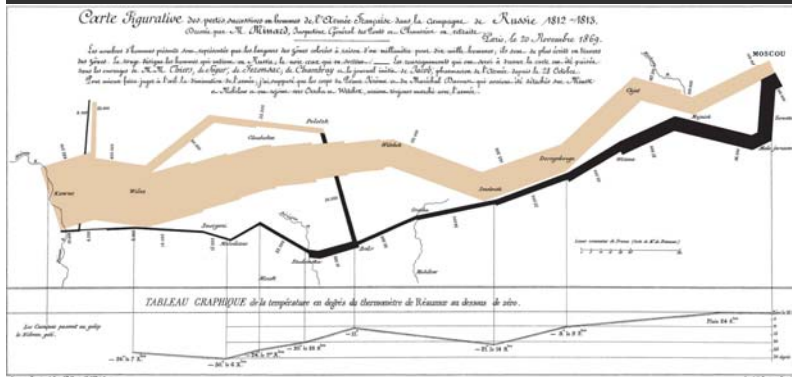
temperature (Q)

latitude (Q) / time (O)



[based on slide from Mackinlay]

## Minard 1869: Napoleon's march



Depicts at least 5 quantitative variables. Any others?

## Automated design

Jock Mackinlay's APT 86



## Combinatorics of Encodings

### Challenge:

Pick the best encoding from the exponential number of possibilities  $(n+1)^8$

### Principle of Consistency:

The properties of the image (visual variables) should match the properties of the data.

### Principle of Importance Ordering:

Encode the most important information in the most effective way.

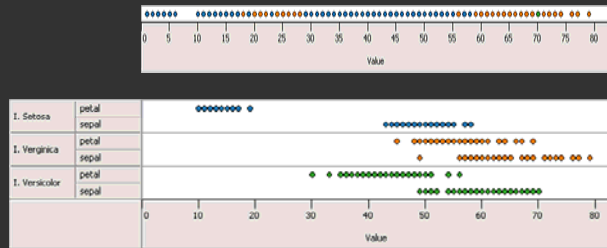
## Design Criteria (Mackinlay)

### Expressiveness

A set of facts is expressible in a visual language if the sentences (i.e. the visualizations) in the language express *all* the facts in the set of data, and *only* the facts in the data.

## Cannot express the facts

A one-to-many (1 → N) relation cannot be expressed in a single horizontal dot plot because multiple tuples are mapped to the same position



## Expresses facts not in the data

A length is interpreted as a quantitative value;  
∴ Length of bar says something untrue about N data

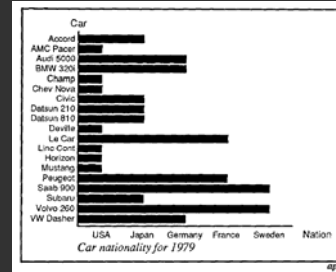


Fig. 11. Incorrect use of a bar chart for the Nation relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the Nation relation.

[Mackinlay, APT, 1986]

## Design Criteria (Mackinlay)

### Expressiveness

A set of facts is expressible in a visual language if the sentences (i.e. the visualizations) in the language express *all* the facts in the set of data, and *only* the facts in the data.

### Effectiveness

A visualization is more effective than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.

(Effectiveness subject of the *Graphical Perception* lecture)

## Mackinlay's Ranking

Quantitative	Ordinal	Nominal
Position	Position	Position
Length	Density	Hue
Angle	Saturation	Texture
Slope	Hue	Connection
Area	Texture	Containment
Volume	Connection	Density
Density	Containment	Saturation
Saturation	Length	Shape
Hue	Angle	Length
Texture	Slope	Angle
Connection	Area	Slope
Containment	Volume	Area
Shape	Shape	Volume

Conjectured effectiveness of the encoding

## Mackinlay's Design Algorithm

User formally specifies data model and type

- Additional input: ordered list of data variables to show

APT searches over design space

- Tests expressiveness of each visual encoding
- Generates image for encodings that pass test
- Tests perceptual effectiveness of resulting image

Outputs the “most effective” visualization

## Limitations

Does not cover many visualization techniques

- Bertin and others discuss networks, maps, diagrams
- Does not consider 3D, animation, illustration, photography, ...

Does not model interaction

## Summary

Formal specification

- Data model
- Image model
- Encodings mapping data to image

Choose expressive and effective encodings

- Formal test of expressiveness
- Experimental tests of perceptual effectiveness

## Assignment 1: Visualization Design

Design a static visualization for a given data set.

Deliverables (post to the course wiki)

- Image of your visualization
- Short description and design rationale ( $\leq 4$  para.)

Due by **7:00am** on **Monday 9/28**.