Data and Image Models

Maneesh Agrawala

CS 448B: Visualization
Spring 2016

Last Time: The Purpose of Visualization

Three functions of visualizations

Record information

- Photographs, blueprints, …

Support reasoning about information (analyze)

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

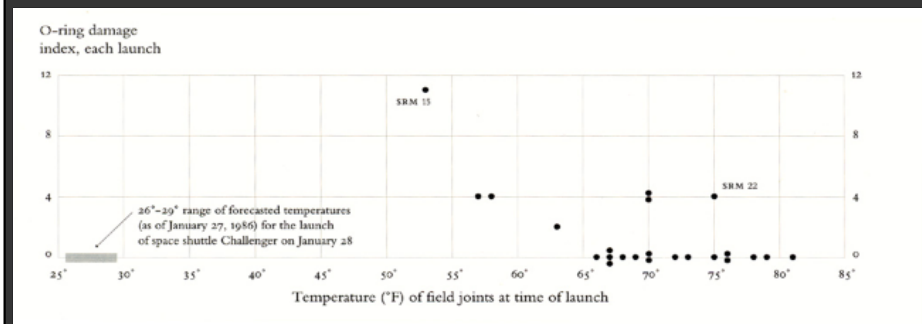
Convey information to others (present)

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

Record information

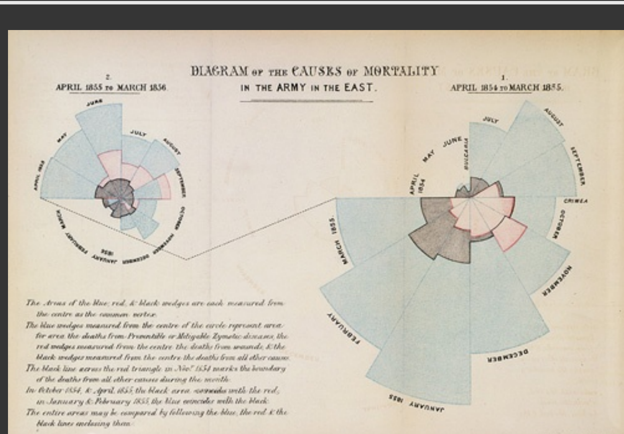
Gallop, Bay Horse “Daisy” [Muybridge 1884-86]

Analysis: Challenger



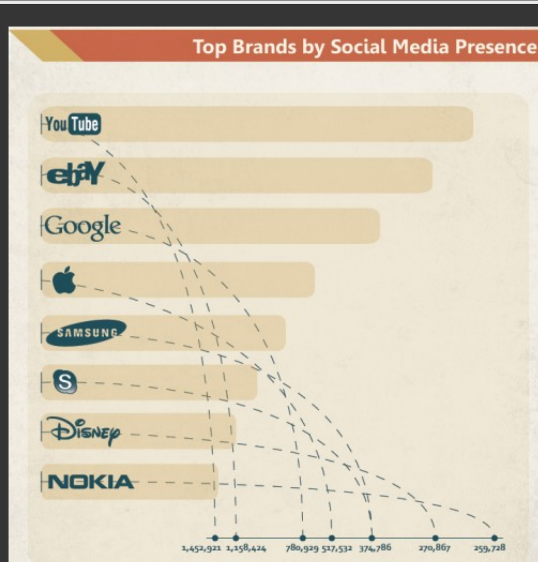
Visualizations drawn by Tufte show how low temperatures damage O-rings [Tufte 97]

Communicate: War Deaths



Crimean War Deaths [Nightingale 1858]

Confuse: Top Brands



from wtfviz.net

Announcements

Class participation requirements

- Complete readings before class
- In-class discussion
- Post at least 1 discussion substantive comment/question by 1:30pm on day of lecture

Sitting in on lectures is ok but we cannot grade your work if you choose to do the assignments

Class wiki
http://web.stanford.edu/class/cs448b/

Assignment 1: Visualization Design

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	name	age	gender	raceethnicity	month	day	year	streetaddres	city	state	latitude	longitude	state_fp	county_fp	tr
2	A'donte Was	16	Male	Black	February	23	2015	Clearview Ln	Millbrook	AL	32.529577	-86.362829	1	51	
3	Aaron Rutlec	27	Male	White	April	2	2015	300 block Iris	Pineville	LA	31.3217392	-92.43486	22	79	
4	Aaron Siler	26	Male	White	March	14	2015	22nd Ave an	Kenosha	WI	42.5835597	-87.83571	55	59	
5	Aaron Valdez	25	Male	Hispanic/Lati	March	11	2015	3000 Seminc	South Gate	CA	33.9392976	-118.21946	6	37	
6	Adam Jovicic	29	Male	White	March	19	2015	364 Hiwood	Munroe Falls	OH	41.1485748	-81.429878	39	153	
7	Adam Reinha	29	Male	White	March	7	2015	18th St and F	Phoenix	AZ	33.4693799	-112.04332	4	13	
8	Adrian Hernz	22	Male	Hispanic/Lati	March	27	2015	4000 Union	Bakersfield	CA	35.3956975	-119.00274	6	29	
9	Adrian Solis	35	Male	Hispanic/Lati	March	26	2015	1500 Bayview	Wilmington	CA	33.7930495	-118.27093	6	37	
10	Alan Alverso	44	Male	White	January	28	2015	Pickett Runn	Sunset	TX	30.6653042	-96.401482	48	41	
11	Alan James	31	Male	White	February	7	2015	200 Abbie St	Wyoming	MI	42.8932381	-85.660584	26	81	
12	Albert Hanso	76	Male	White	April	26	2015	7th Ave and	Hanford	CA	36.2109603	-119.58288	6	31	
13	Alec Ouzoun	40	Male	White	May	12	2015	28 Paseo Vie	Rancho Santa	CA	33.6533852	-117.61337	6	59	
14	Alejandro Sa	Unknown	Male	Hispanic/Lati	February	20	2015	1200 E Airte	Houston	TX	29.9832049	-95.403857	48	201	
15	Alexander Lo	31	Male	White	February	25	2015	25th St and F	Terre Haute	IN	39.4629302	-87.37886	18	167	
16	Alexander M	23	Male	White	April	6	2015	5700 block A	Indianapolis	IN	39.7669106	-86.149963	18	97	
17	Alexander Ri	39	Male	Hispanic/Lati	May	30	2015	1128 Murfre	Nashville	TN	36.1259117	-86.709015	47	37	
18	Alexia Christi	25	Female	Black	April	30	2015	141 Pryor St	Atlanta	GA	33.7512627	-84.393028	13	121	
19	Alfredo Rials	54	Male	Hispanic/Lati	May	19	2015	4219 2nd Ro	Arlington	VA	38.8731527	-77.10501	51	13	
20	Alice Brown	24	Female	White	March	17	2015	Van Ness Av	San Francisco	CA	37.7894309	-122.4221	6	75	
21	Alvin Haynes	57	Male	Black	January	26	2015	1 Moreland	San Francisco	CA	37.6279793	-122.45393	6	81	

Police Killings 2015

Due by 11:59pm on Sun Apr 3

Submissions via Wiki: Look on Piazza for instructions for logging into wiki

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: low level descriptions of the data

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

Conceptual models: 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 [Schneiderman 96]

Types of variables

Physical types

- Characterized by storage format
- Characterized by machine operations

Example:

bool, short, int32, float, double, string, …

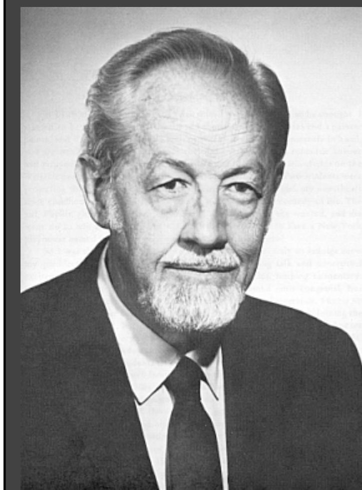
Abstract types

- 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



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

N - Nominal (labels)

Fruits: Apples, oranges, …

Operations: =, ≠

O - Ordered

Quality of meat: Grade A, AA, AAA

Operations: =, ≠, <, >, ≤, ≥

Q - Interval (location of zero arbitrary)

Dates: Jan, 19, 2006; Loc.: (LAT 33.98, LON -118.45)

Like a geometric point. Cannot compare directly

Only differences (i.e. intervals) may be compared

Operations: =, ≠, <, >, ≤, ≥, -

Q - Ratio (location of zero fixed)

Physical measurement: Length, Mass, Temp, …

Counts and amounts

Like a geometric vector, origin is meaningful

Operations: =, ≠, <, >, ≤, ≥, -, ÷

From data model to N,O,Q data type

Data model

- 32.5, 54.0, -17.3, …
- floats

Conceptual model

- Temperature

Data type

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

Iris Setosa

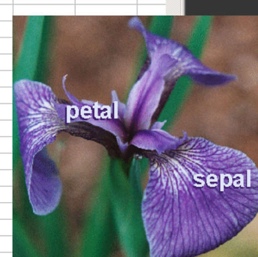
Iris Versicolor

Iris Virginica

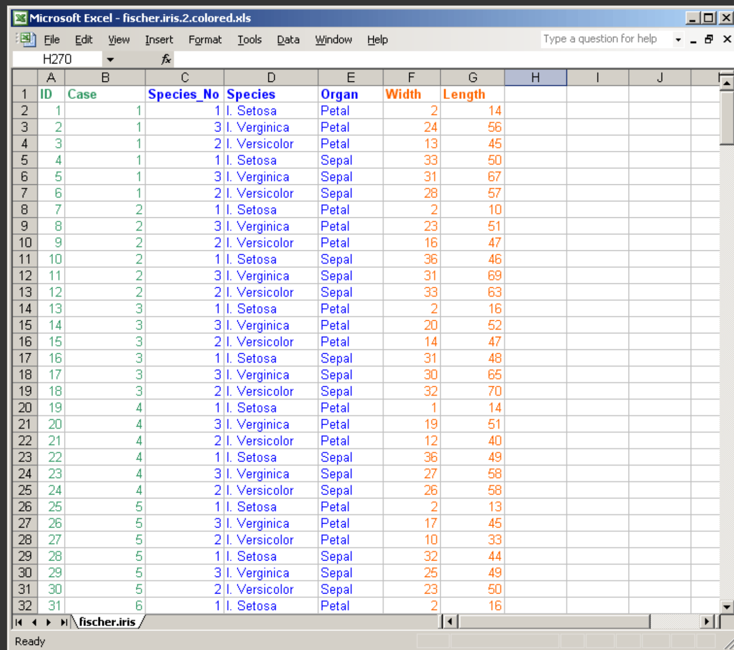
Microsoft Excel - fischer.iris.2.xls

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

fischer.iris



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



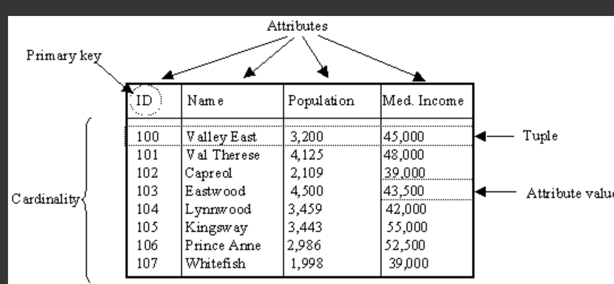
ID	Case	Species_No	Species	Organ	Width	Length
1	1	1	I. Setosa	Petal	2	14
2	1	3	I. Verginica	Petal	24	56
3	1	2	I. Versicolor	Petal	13	45
4	1	1	I. Setosa	Sepal	33	50
5	1	3	I. Verginica	Sepal	31	67
6	1	2	I. Versicolor	Sepal	28	57
7	2	1	I. Setosa	Petal	2	10
8	2	3	I. Verginica	Petal	23	51
9	2	2	I. Versicolor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

Q
O
N

Relational data model

Represent data as a **table** (*relation*)
Each **row** (*tuple*) represents a single record
  Each record is a fixed-length tuple
Each **column** (*attribute*) represents a single *variable*
  Each attribute has a *name* and a *data type*
A table's **schema** is the set of names and data types

A **database** is a collection of tables (relations)



ID	Name	Population	Med. Income
100	Valley East	3,200	45,000
101	Val Therese	4,125	48,000
102	Capreol	2,109	39,000
103	Eastwood	4,500	43,500
104	Lynnwood	3,459	42,000
105	Kingsway	3,443	55,000
106	Prince Anne	2,986	52,500
107	Whitefish	1,998	39,000

Relational algebra [Codd 1970]

Data transformations (SQL)

- Selection (WHERE) – restrict values
- Projection (SELECT) – choose subset of attributes
- Sorting (ORDER BY)
- Aggregation (GROUP BY, SUM, MIN, …)
- Set operations (UNION, …)
- Combine (INNER JOIN, OUTER 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

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

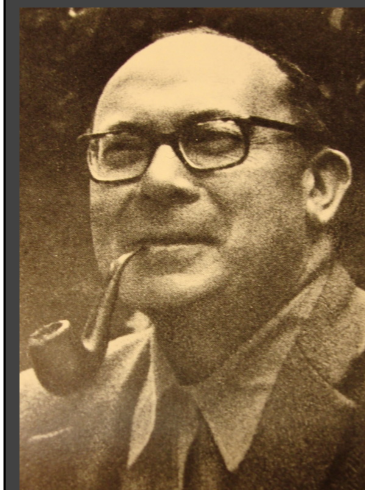
Dimensions and measures

Independent vs. dependent variables

- Example: y = f(x,a)
- Dimensions: Domain(x) × Domain(a)
- Measures: Range(y)

Image

Marks and Visual Variables

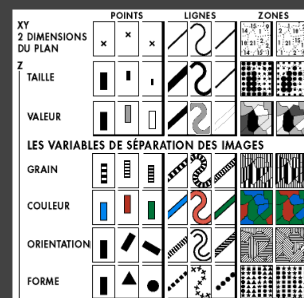


Semiology of Graphics
J. Bertin, 1967

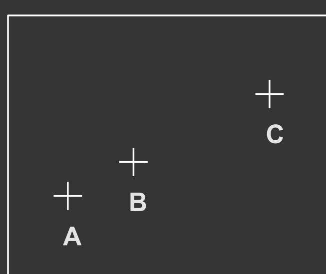
Marks: geometric primitives

Visual Variables: control mark appearance

- Position (2x)
- Size
- Value
- Texture
- Color
- Orientation
- Shape



Coding information in position



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

∴ Encode quantitative variables

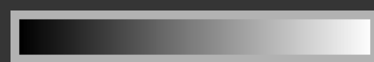
"Resemblance, order and proportional are the three signfields in graphics." - Bertin

Coding info 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

Bertins' "Levels of Organization"

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

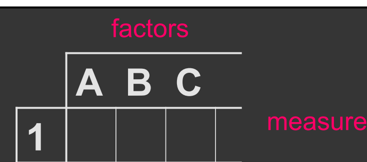
N Nominal
O Ordered
Q Quantitative

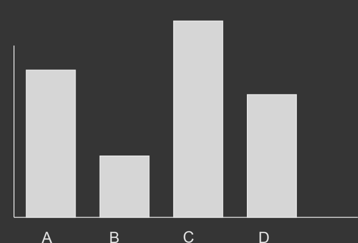
Note: Q < O < N

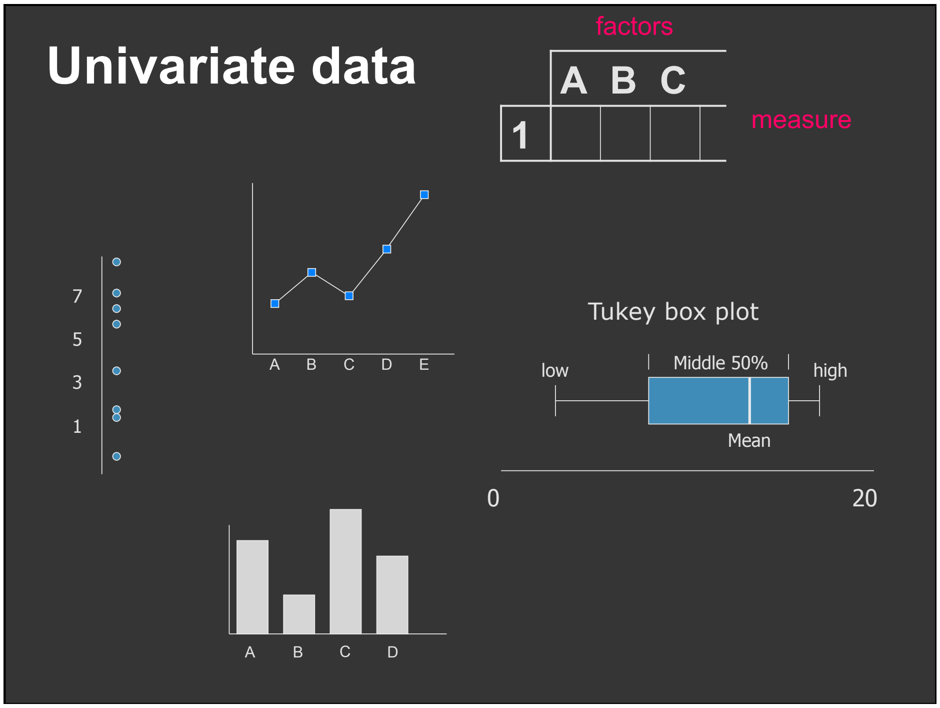
Note: Bertin actually breaks visual variables down into differentiating (≠) and associating (≡)

Encoding rules

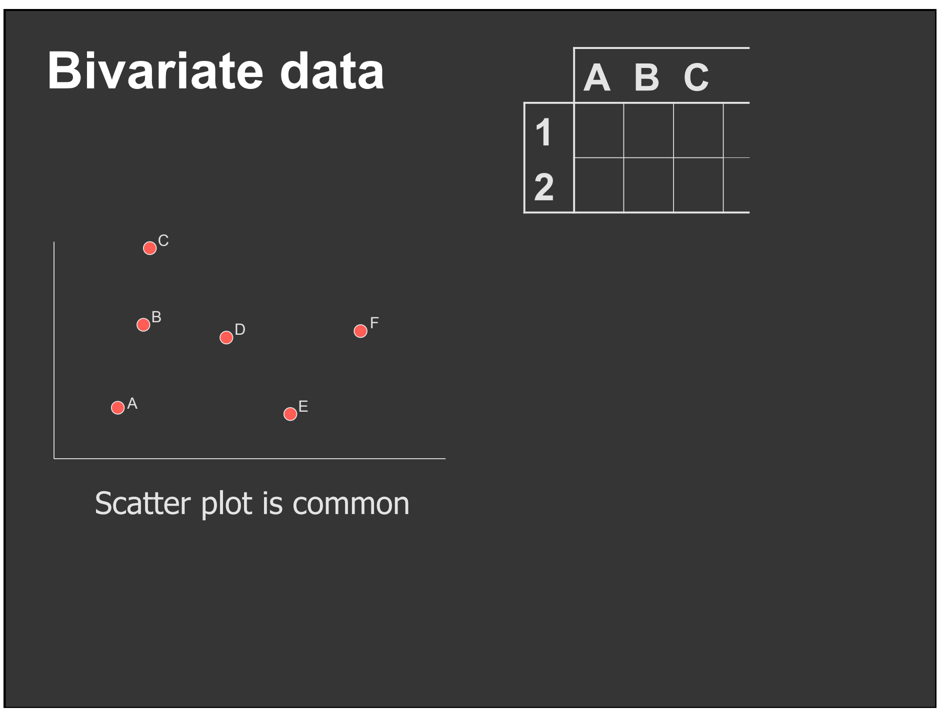
Univariate data







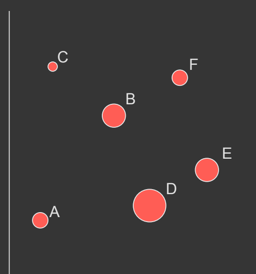
Univariate data
factors
A B C
1
measure
7
5
3
1
A B C D E
Tukey box plot
low
Middle 50%
high
Mean
0
20
A B C D



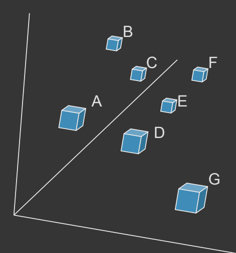
Bivariate data
A B C
1
2
C
B
D
F
A
E
Scatter plot is common

Trivariate data

	A	B	C
1			
2			
3			



3D scatter plot is possible



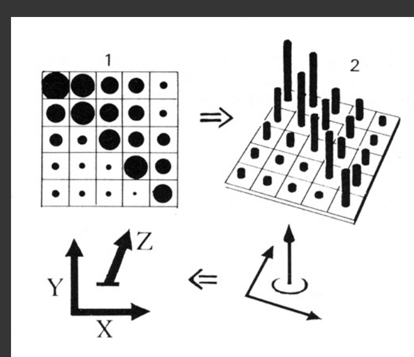
Three variables

Two variables [x,y] can map to points

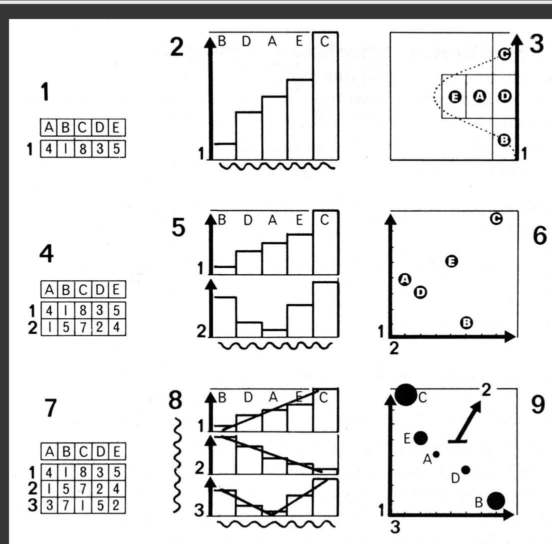
- Scatterplots, maps, …

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

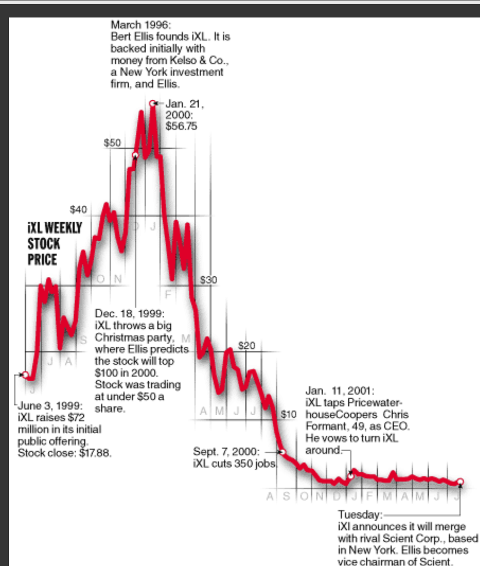
How many variables can be depicted in an image?

	A	B	C	
1				
2				
3				
4				
5				
6				
7				
8				

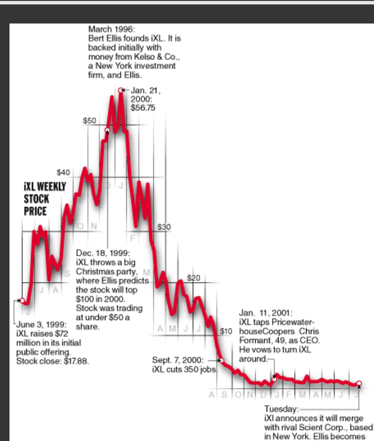
"With up to three rows, a data table can be constructed directly as a single image … However, an image has only three dimensions. And this barrier is impassible." **Bertin**

Deconstructions

Stock chart from the late 90s

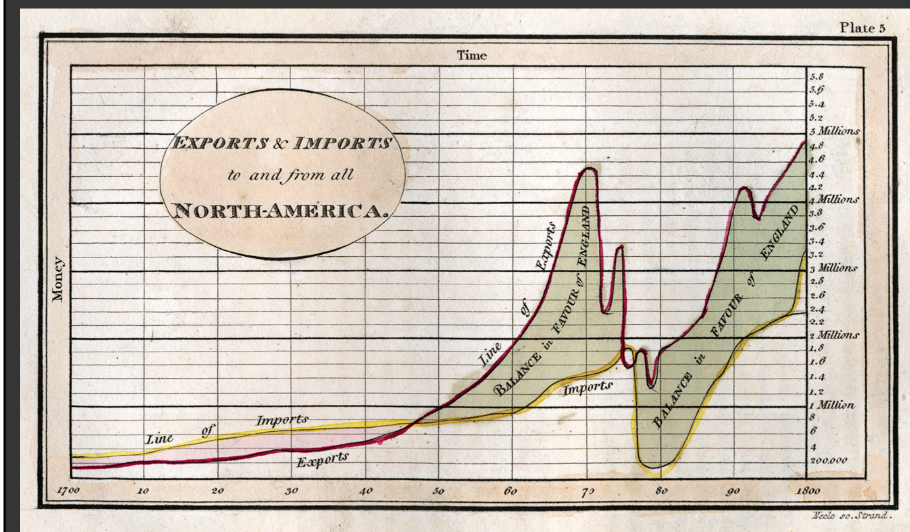


Stock chart from the late 90s

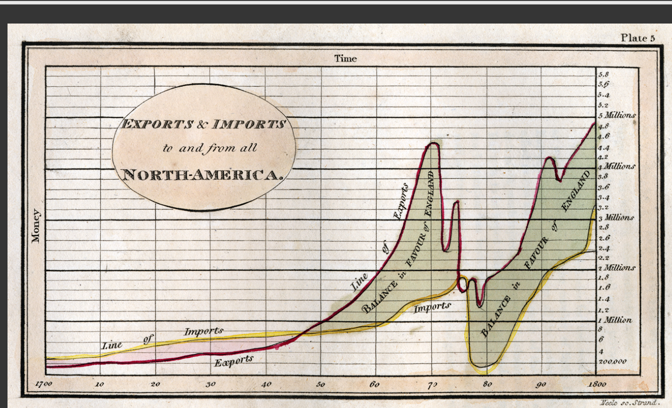


- Time → x-position (Q, linear)
- Price → y-position (Q, linear)

Playfair 1786/1801

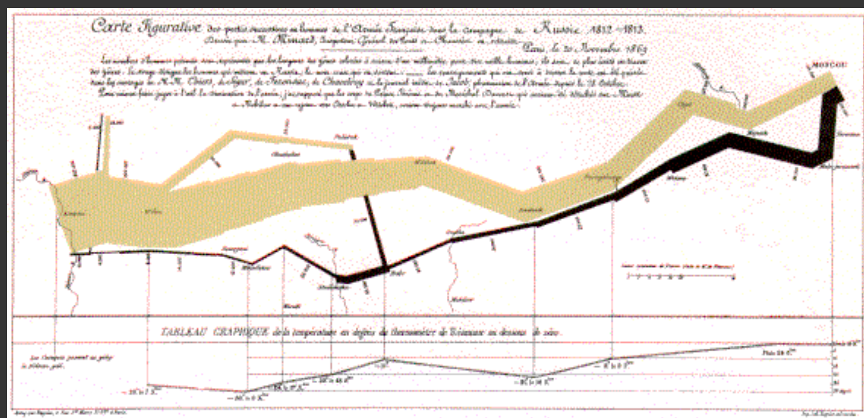


Playfair 1786/1801

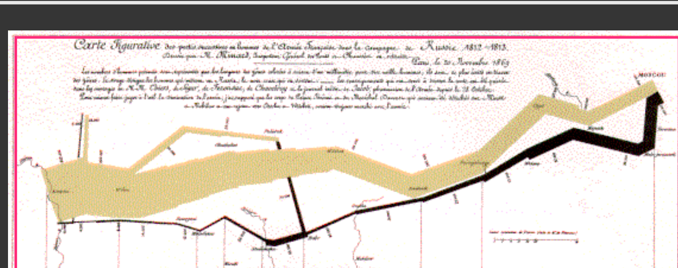


- Time → x-position (Q, linear)
- Exports/Imports Values → y-position (Q, linear)
- Exports/Imports → color (N, O, nominal)
- Balance for/against → area (maybe length??) (Q, linear)
- Balance for/against → color (N, O, nominal)

Minard 1869: Napoleon's march

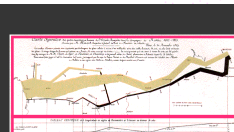


Single axis composition



+

=



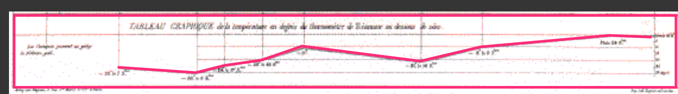
[based on slide from Mackinlay]

Mark composition

temperature → y-position (Q, linear)

\+ longitude → x-position (Q, linear)

=



temp over longitude (Q x Q)

[based on slide from Mackinlay]

Mark composition

latitude → y-position (Q, linear)

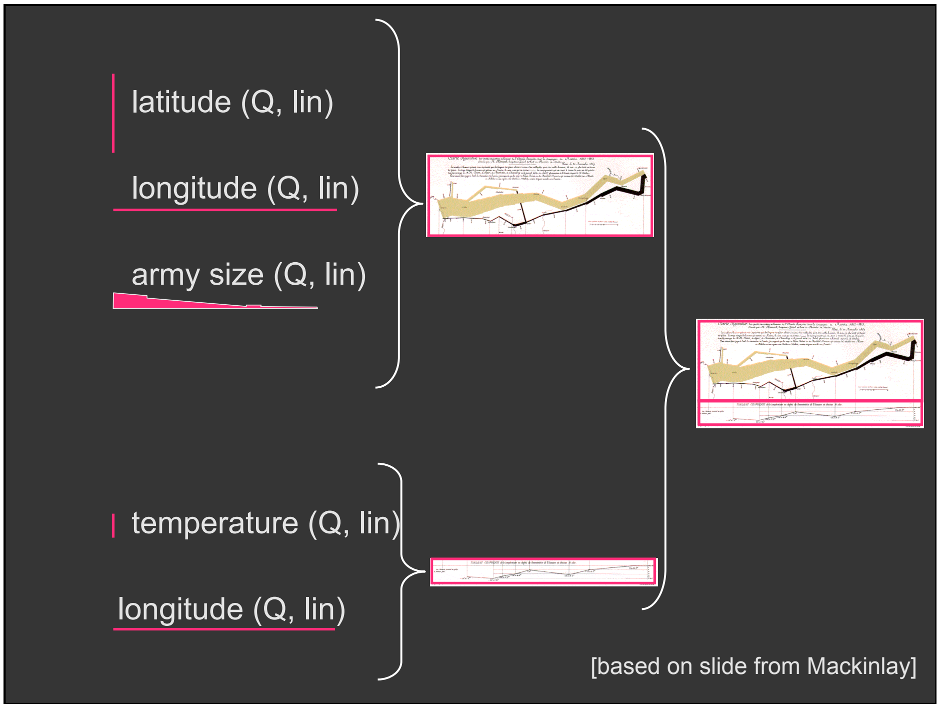
\+ longitude → x-position (Q, linear)

\+ army size → width (Q, linear)

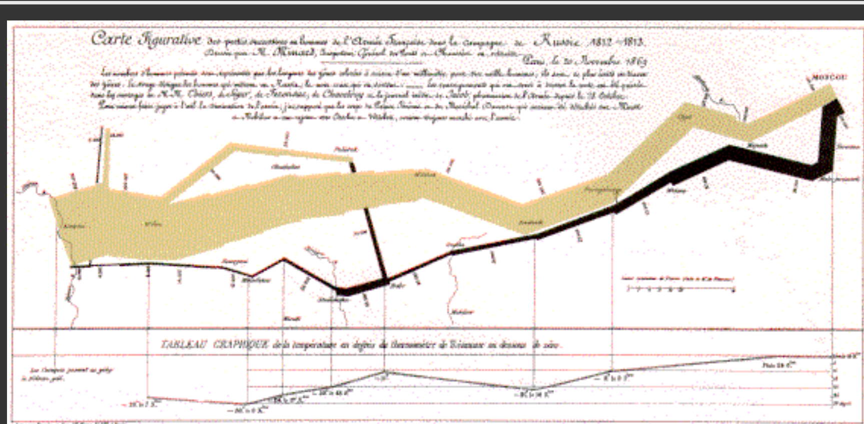
=

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

[based on slide from Mackinlay]



Minard 1869: Napoleon's march

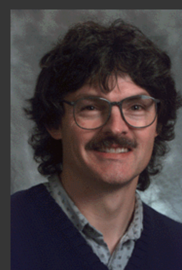


Depicts at least 4 quantitative variables

Any others?

Automated design

Jock Mackinlay's APT 86



Combinatorics of encodings

Challenge:

Assume 8 visual encodings and n data attributes

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

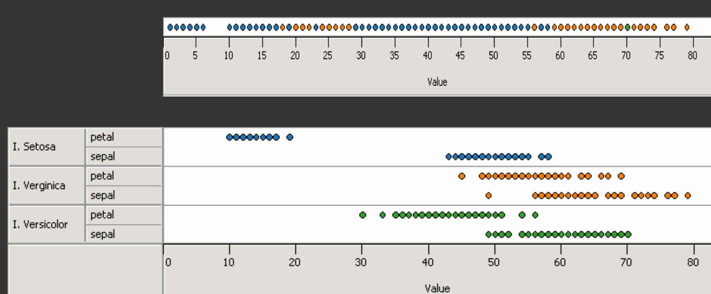
Mackinlay's expressiveness criteria

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

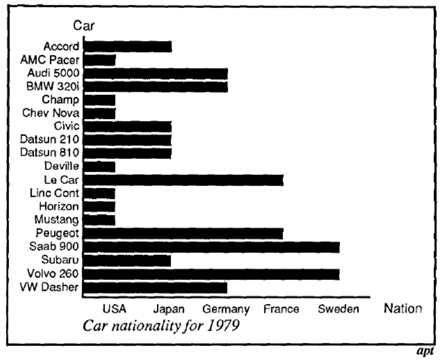


Fig. 11. Incorrect use of a bar chart for the *Nation* relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the *Nation* relation.

[Mackinlay, APT, 1986]

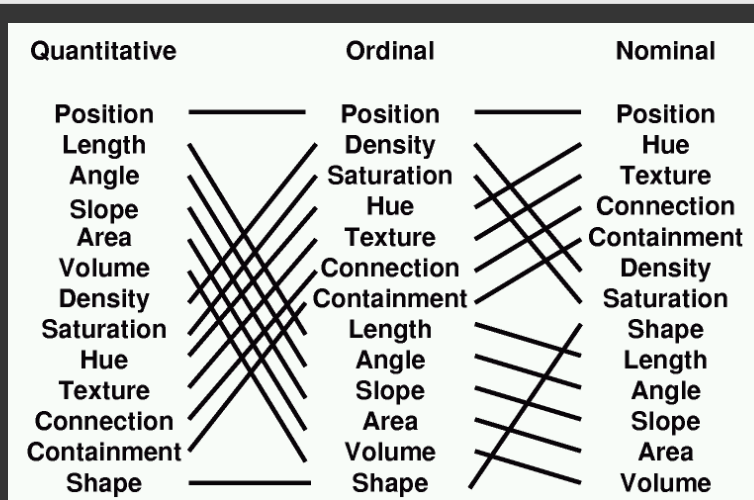
Mackinlay's effectiveness criteria

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.**

Subject of perception lecture

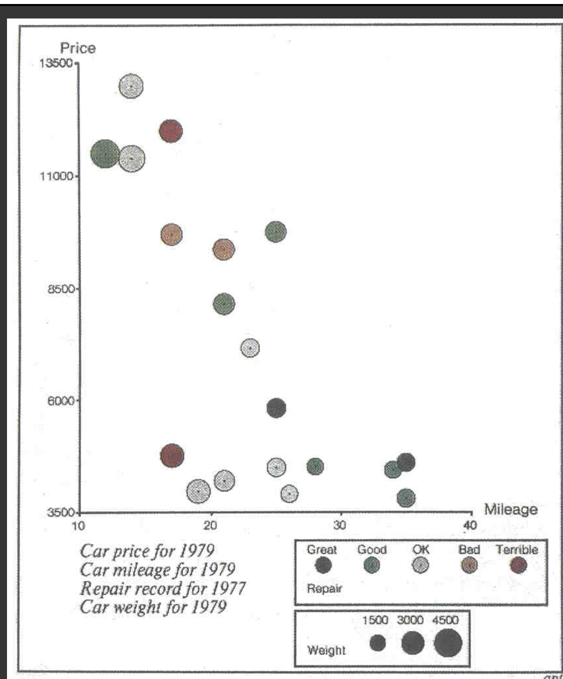
Mackinlay's ranking



Conjectured *effectiveness* of the encoding

Mackinlay's design algorithm

- User formally specifies data model and type
- 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 most effective visualization



[Mackinlay, APT, 1986]

Limitations

Does not cover many visualization techniques

- Bertin and others discuss networks, maps, diagrams
- They do 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