Data and Image Models

Maneesh Agrawala

CS 448B: Visualization
Fall 2017

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

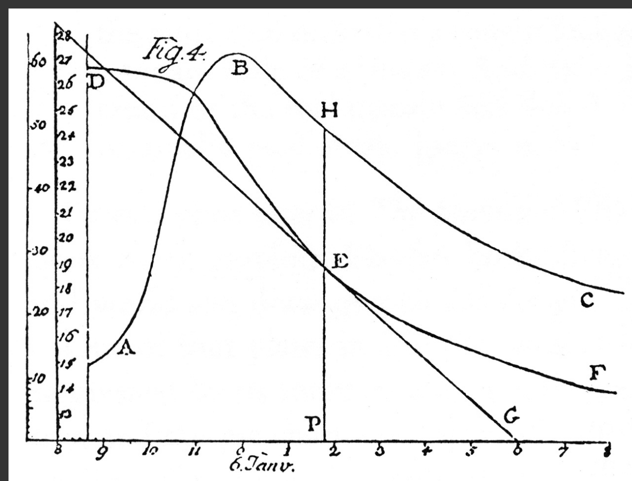
Convey information to others (present)

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

Record information

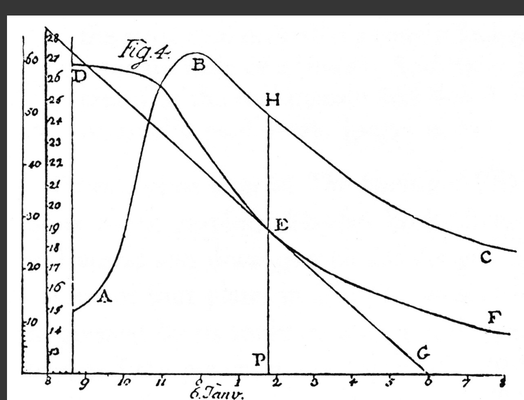
Gallop, Bay Horse “Daisy” [Muybridge 1884-86]

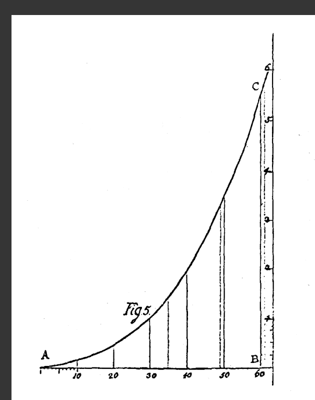
Graphical calculation: Evaporation



Johannes Lambert used graphs to study the rate of water evaporation as function of temperature [from Tufte 83]

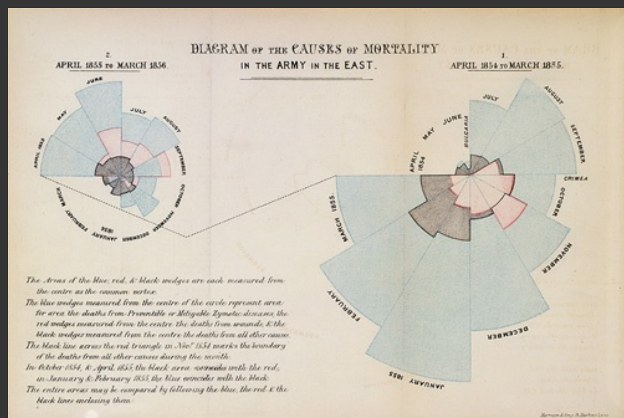
Graphical calculation: Evaporation





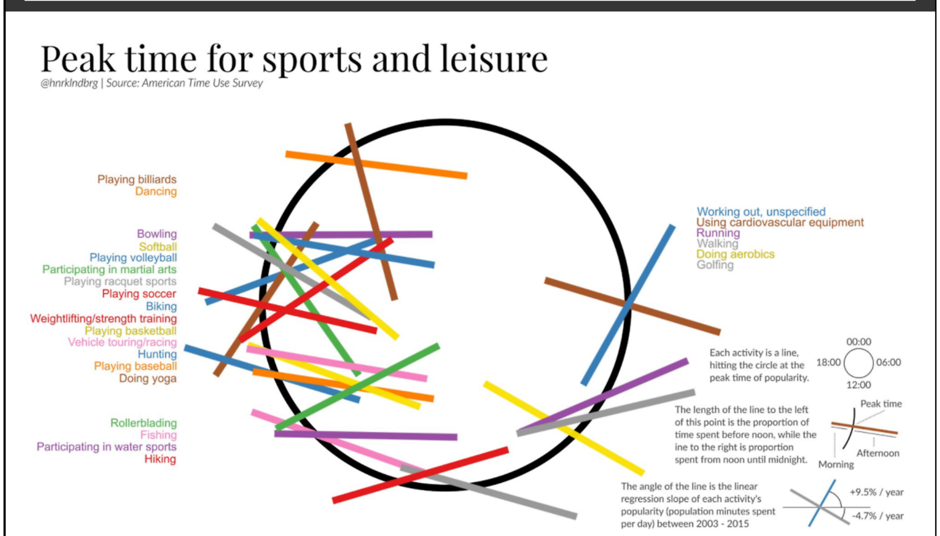
Johannes Lambert used graphs to study the rate of water evaporation as function of temperature [from Tufte 83]

Communicate: War Deaths



Crimean War Deaths [Nightingale 1858]

Confuse



from wtfviz.net

Announcements

Class participation requirements

- Complete readings before class
- In-class discussion
- Post at least 1 discussion substantive comment/question by noon the day after lecture

Office Hours on website

Class wiki
https://magrawala.github.io/cs448b-fa17

Assignment 1: Visualization Design

Barley Yield Data

In 1931 and 1932 Minnesota collected data on the yield in bushels per acre of 10 varieties of barley grown in 1/40 acre plots at University Farm, St. Paul, and at the five branch experiment stations located at Waseca, Morris, Crookston, Grand Rapids, and Duluth (all in Minnesota). The varieties were grown in three randomized blocks at each of the six stations during 1931 and 1932, different land being used each year of the test.

Number of records: 120
Variable Names:
Site: Crookston, Duluth, Grand Rapids, Morris, University Farm, Waseca
Variety: Glabron, Manchuria, No 457, No 462, No 475, Peatland, Svansota, Trebi, Velvet, Wisc. No 38
Yield: bushels/acre
Year: 1931, 1932

We've cleaned up this dataset and posted in csv format: barley2.csv

Barley Yields

Due by noon on Mon Oct 2

Submissions of PDF via Canvas, bring printout to class

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, …

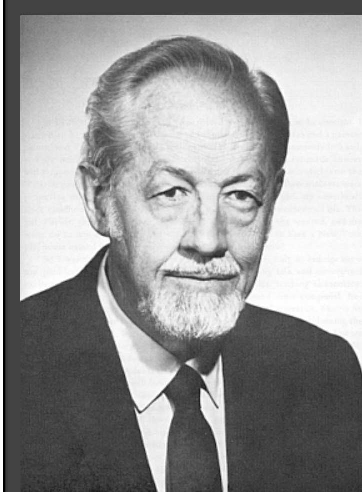
Abstract types

- 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

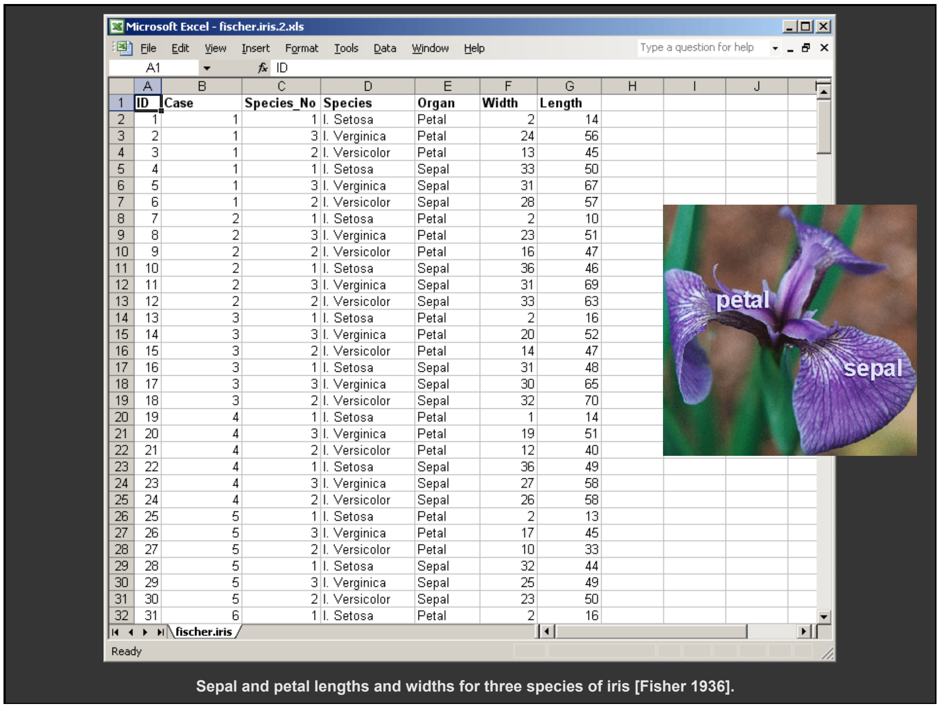
Data type

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

Iris Setosa

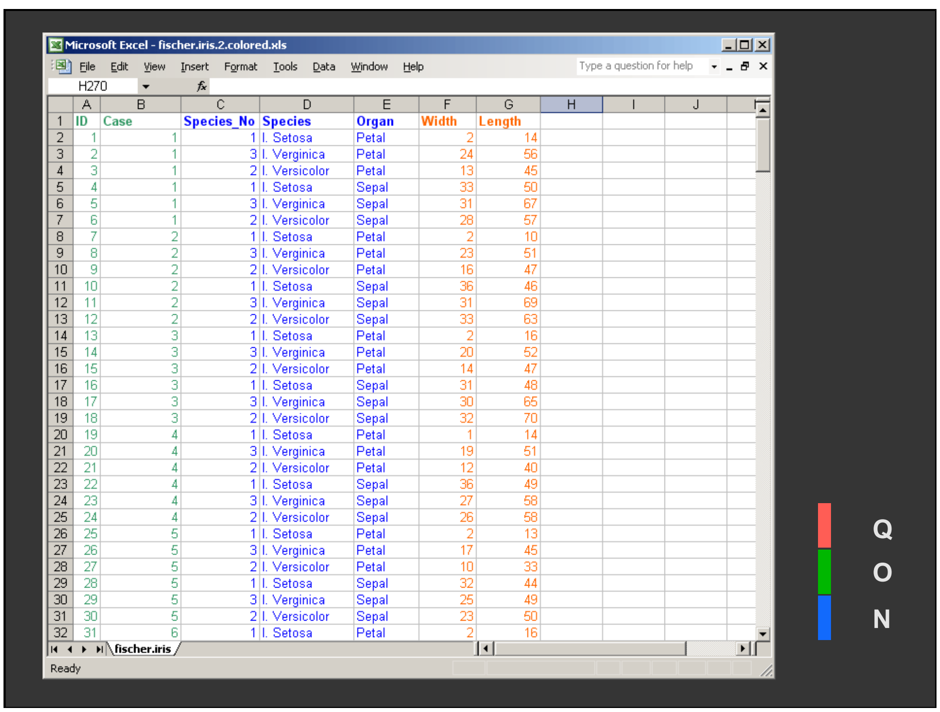
Iris Versicolor

Iris Virginica



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

Relational data model

Represent data as a **table** (*relation*)

Each **row** (*tuple*) represents a single record

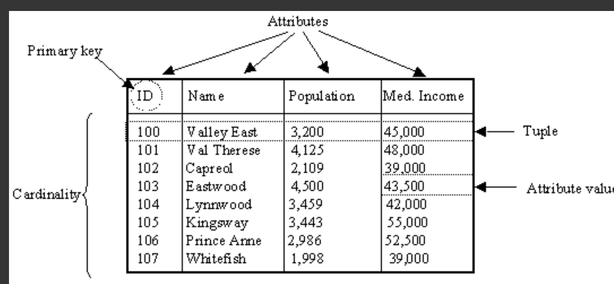
  Each record is a fixed-length tuple

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

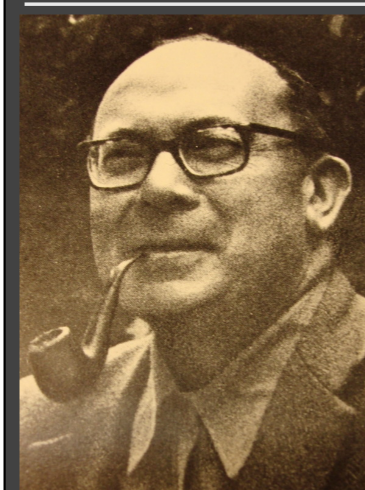
Dimensions and measures

Independent vs. dependent variables

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

Image

Marks and Visual Variables



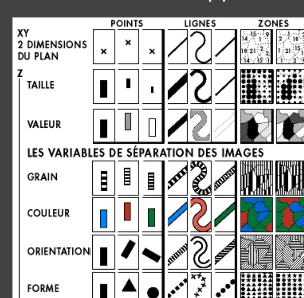
Semiology of Graphics
J. Bertin, 1967

Marks: geometric primitives

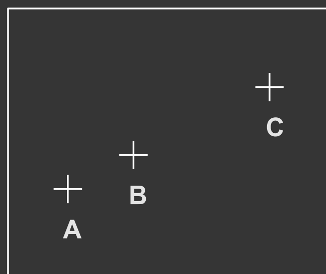
points lines areas

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		

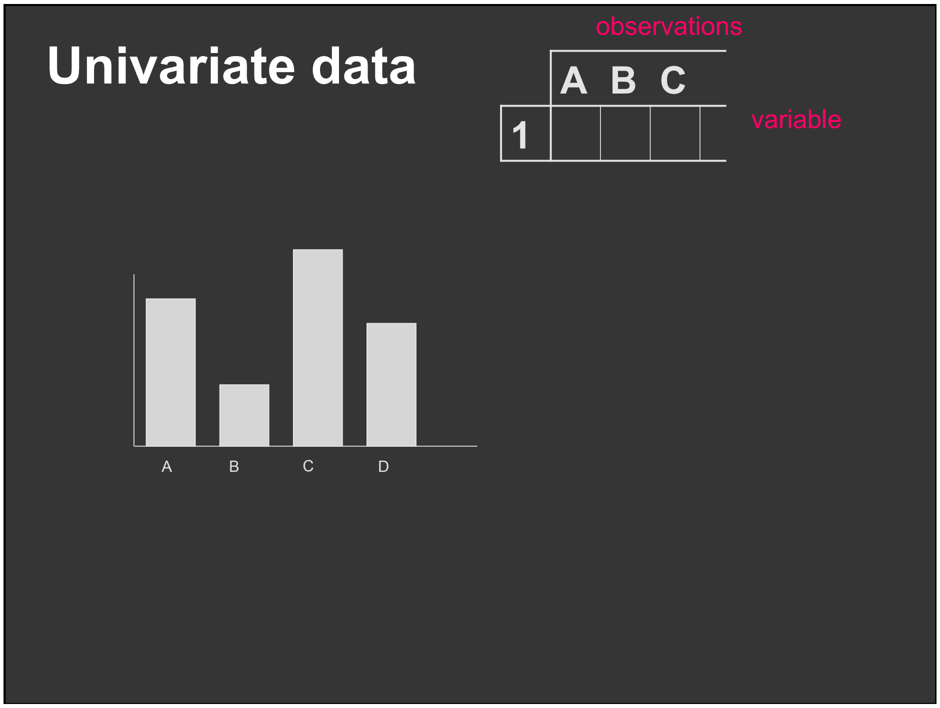
N Nominal
O Ordered
Q Quantitative

Note: Q < O < N

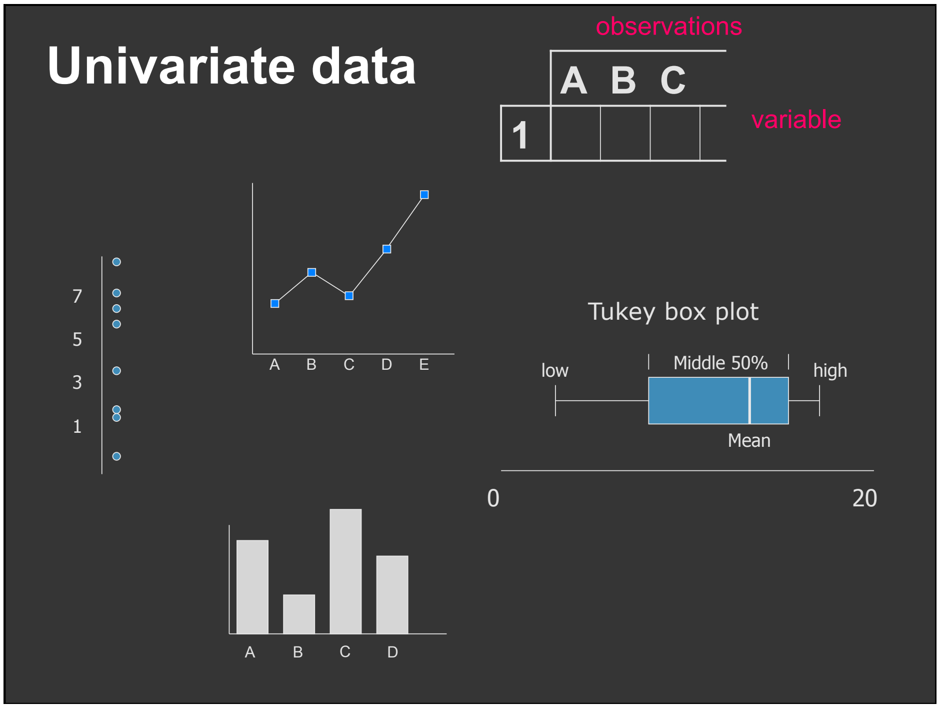
Note: Bertin actually breaks visual variables down into differentiating (≠) and associating (≡)

Visual Encoding

Univariate data

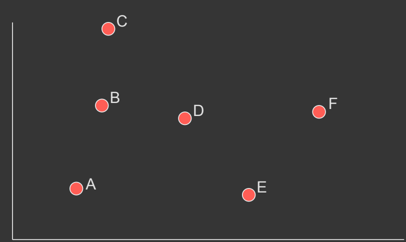


Univariate data



Bivariate data

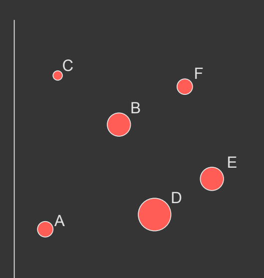
	A	B	C
1			
2			



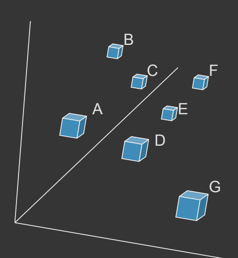
Scatter plot is common

Trivariate data

	A	B	C
1			
2			
3			



3D scatter plot is possible



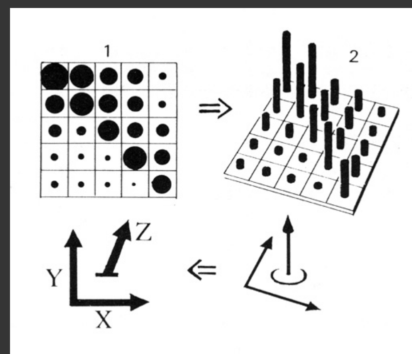
Three variables

Two variables [x,y] can map to points

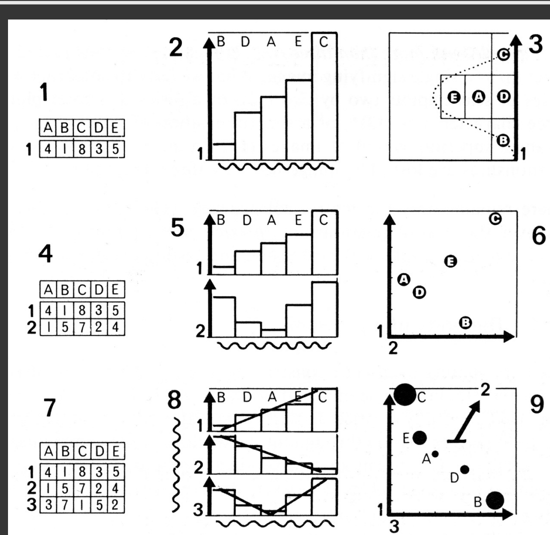
- Scatterplots, maps, …

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				

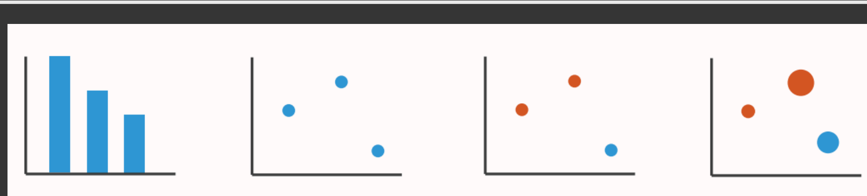
Multidimensional data

How many variables can be depicted in an image?

"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

	A	B	C	
1				
2				
3				
4				
5				
6				
7				
8				

Encodings Map Data to Mark Attr.



mark: lines
data → size
(length)

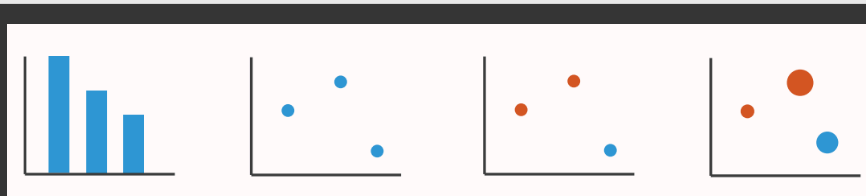
mark: points
$data_1$ → x-pos
$data_2$ → y-pos

mark: points
$data_1$ → x-pos
$data_2$ → y-pos
$data_3$ → color

mark: points
$data_1$ → x-pos
$data_2$ → y-pos
$data_3$ → color
$data_4$ → size

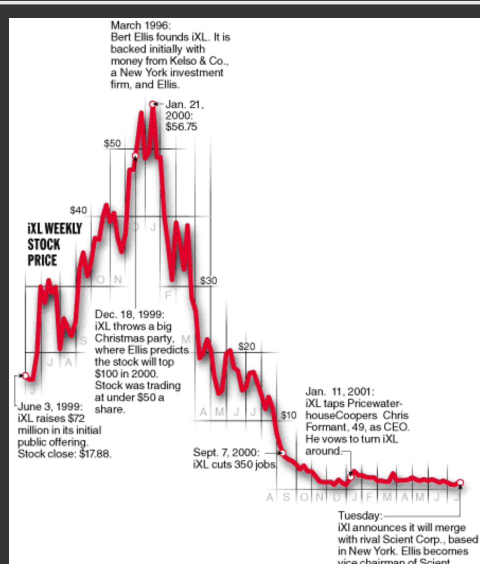
Deconstructions

Given Image Describe Encodings

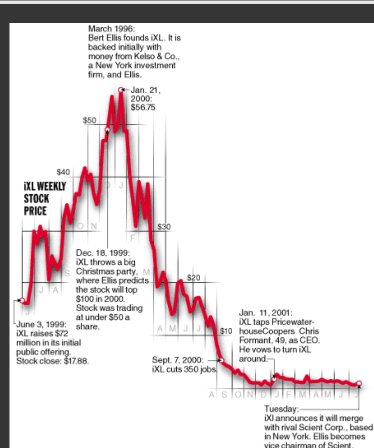


mark: lines	mark: points	mark: points	mark: points
data → size (length)	$data_1$ → x-pos	$data_1$ → x-pos	$data_1$ → x-pos
	$data_2$ → y-pos	$data_2$ → y-pos	$data_2$ → y-pos
		$data_3$ → color	$data_3$ → color
			$data_4$ → size

Stock chart from the late 90s

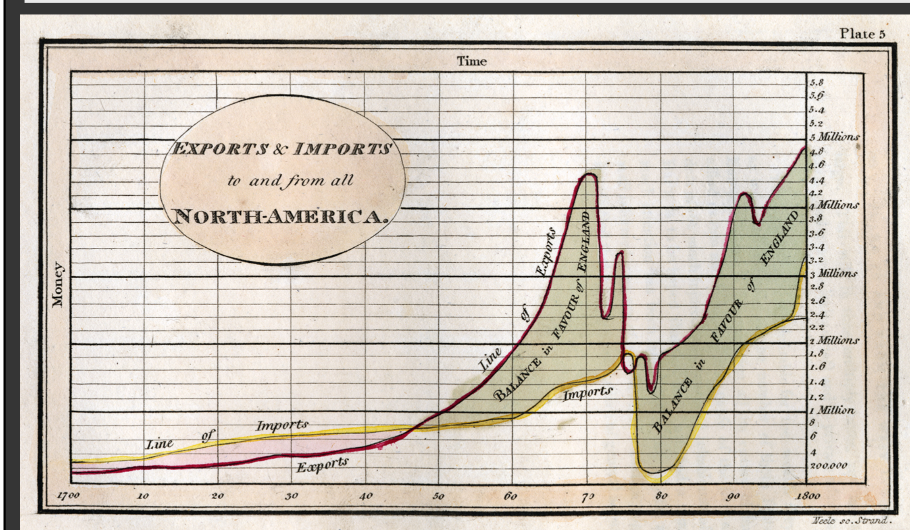


Stock chart from the late 90s

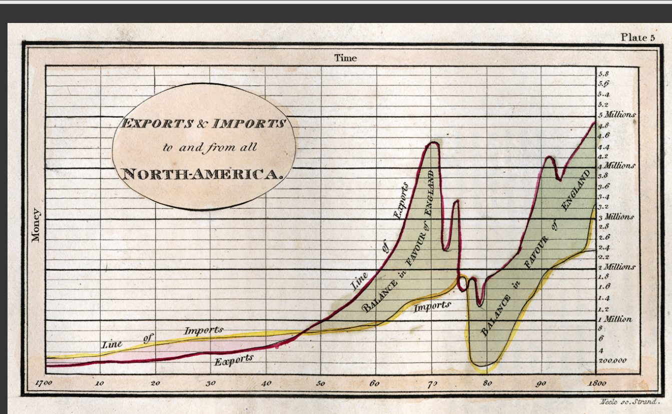


- Time → x-position (Q, linear)
- Price → y-position (Q, linear)

Playfair 1786/1801

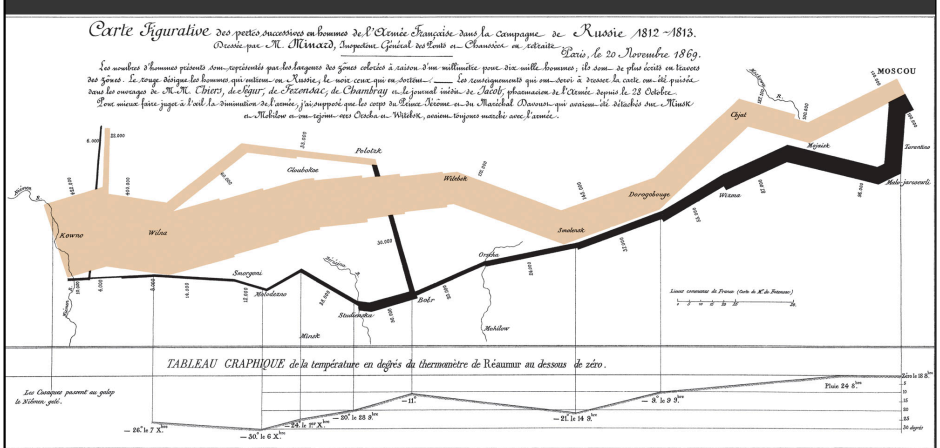


Playfair 1786/1801

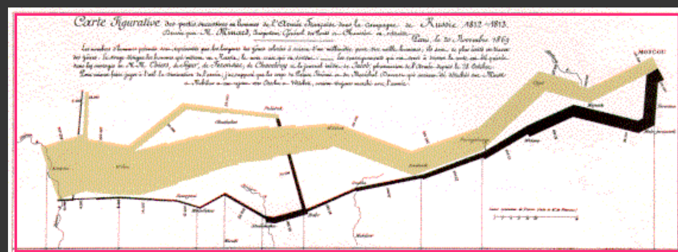


- Time → x-position (Q, linear)
- Exports/Imports Values → y-position (Q, linear)
- Exports/Imports → color (N, O, nominal)
- Balance for/against → area (maybe length??) (Q, linear)
- Balance for/against → color (N, O, nominal)

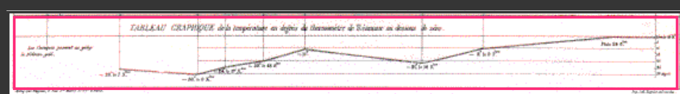
Minard 1869: Napoleon's march



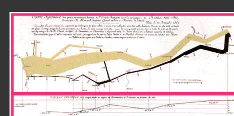
Single axis composition



+



=



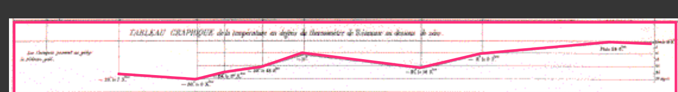
[based on slide from Mackinlay]

Mark composition

temperature → y-position (Q, linear)

\+ longitude → x-position (Q, linear)

=



temp over longitude (Q x Q)

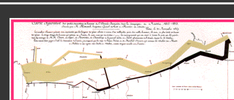
[based on slide from Mackinlay]

Mark composition

latitude → y-position (Q, linear)

\+ longitude → x-position (Q, linear)

\+ army size → width (Q, linear)

= 

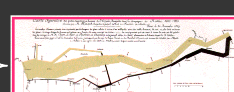
army position (Q x Q) and army size (Q)

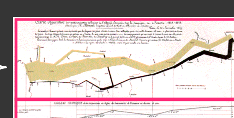
[based on slide from Mackinlay]

latitude (Q, lin)

longitude (Q, lin)

army size (Q, lin)



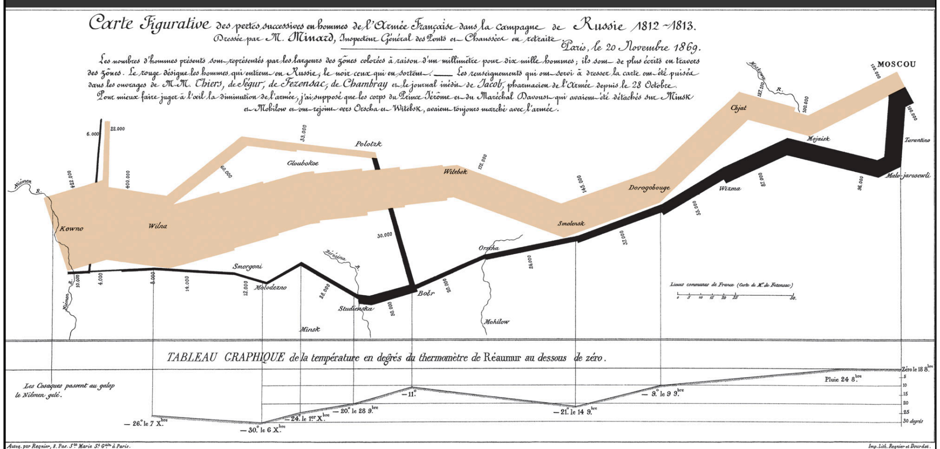


temperature (Q, lin)

longitude (Q, lin)

[based on slide from Mackinlay]

Minard 1869: Napoleon's march

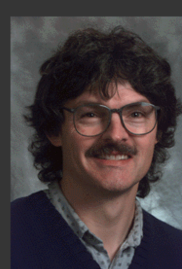


Depicts at least 4 quantitative variables

Any others?

Automated design

Jock Mackinlay's APT 86



Combinatorics of encodings

Challenge:

- Assume 8 visual encodings and n data fields
- 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

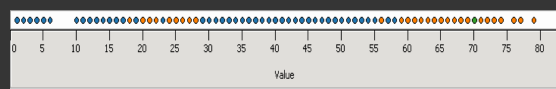
Mackinlay's expressiveness criteria

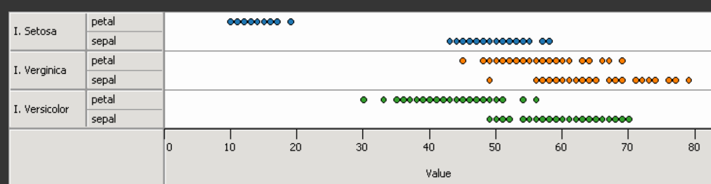
Expressiveness

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

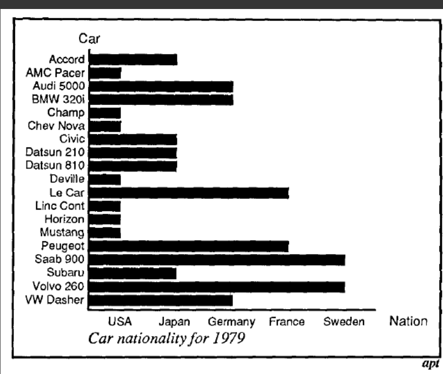


Fig. 11. Incorrect use of a bar chart for the *Nation* relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the *Nation* relation.

[Mackinlay, APT, 1986]

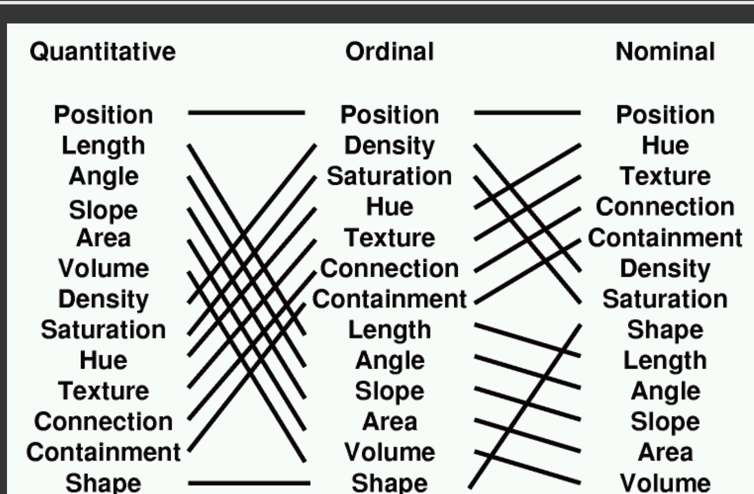
Mackinlay's effectiveness criteria

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

Graphical Perception

Most accurate ↑ Least accurate

- Position (common) scale
- Position (non-aligned) scale
- Length
- Slope
- Angle
- Area
- Volume
- Color hue-saturation-density

Automatic chart construction

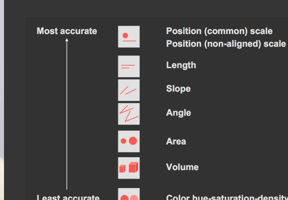
Automating the design of graphical presentation of relational information
J. Mackinlay, 1986

Encode most important data using highest ranking visual variable for the data type

Year	Exports	Imports
1700	170,000	300,000
1701	171,000	302,000
1702	176,000	303,000
…	…	…

→

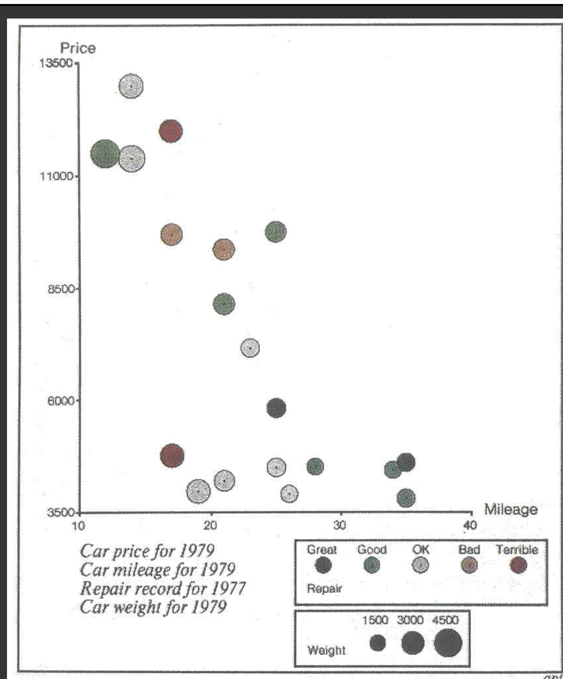
1. Year (Q)
2. Exports (Q)
3. Imports (Q)



→

mark: lines

Year → x-pos (Q)
Exports → y-pos (Q)
Imports → y-pos (Q)



[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