Data and Image Models

Maneesh Agrawala

CS 294-10: Visualization
Fall 2013

Last Time: The Purpose of Visualization

Three functions of visualizations

Record information

- Photographs, blueprints, …

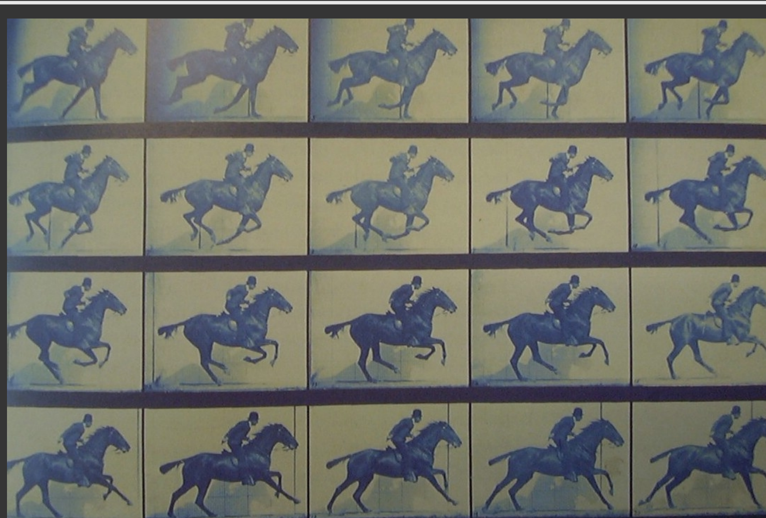
Support reasoning about information (analyze)

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

Convey information to others (present)

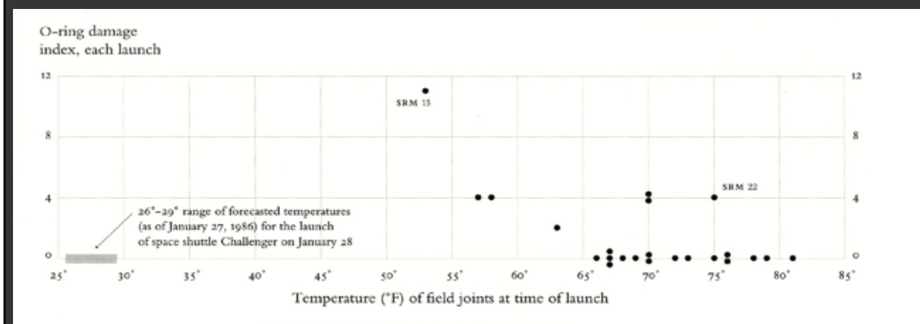
- Share and persuade
- Collaborate and revise
- Emphasize important aspects of data

Record information



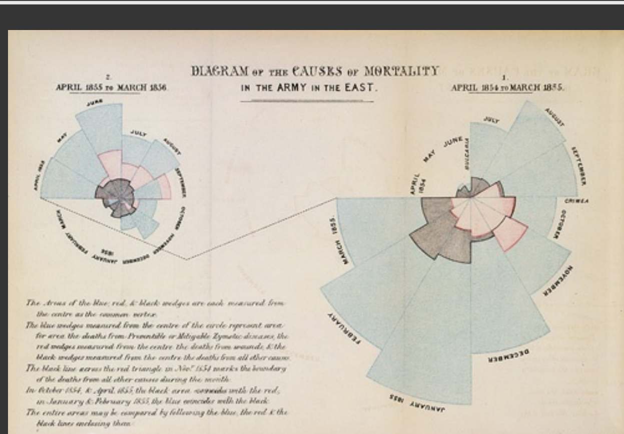
Gallop, Bay Horse “Daisy” [Muybridge 1884-86]

Analysis: Challenger



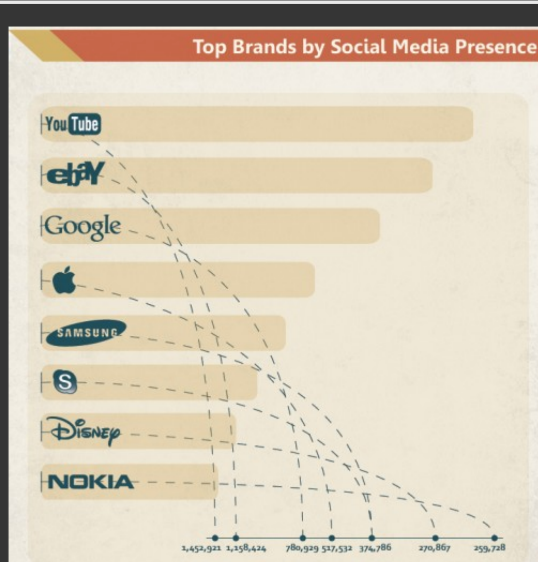
Visualizations drawn by Tufte show how low temperatures damage O-rings [Tufte 97]

Communicate: War Deaths



Crimean War Deaths [Nightingale 1858]

Confuse: Top Brands



from wtfviz.net

Announcements

Auditors, *please* enroll in the class (1 unit, P/NP)

- Requirements: Come to class and participate (online as well)
- Requirements: Assignment 1

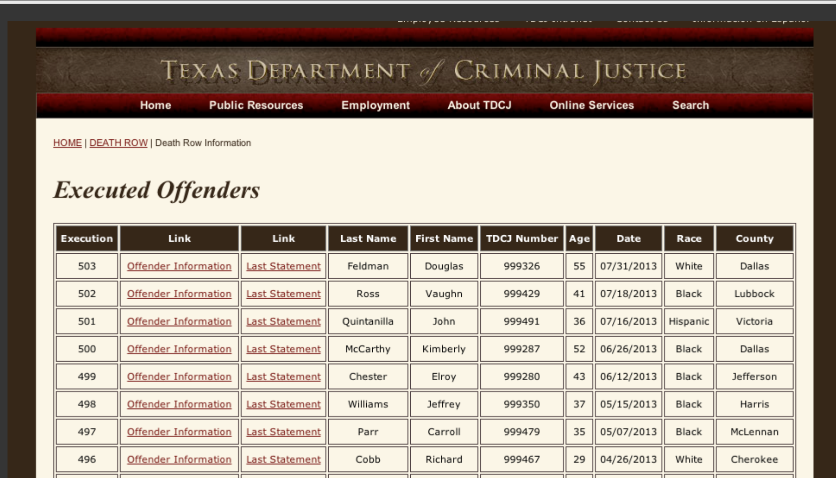
Class participation requirements

- Complete readings before class
- In-class discussion
- Post at least 1 discussion substantive comment/question by 3pm on day of lecture

All, add yourself to participants page on the wiki

Class wiki
http://vis.berkeley.edu/courses/cs294-10-fa13/wiki/

Assignment 1: Visualization Design



TEXAS DEPARTMENT of CRIMINAL JUSTICE

Home | Public Resources | Employment | About TDCJ | Online Services | Search

HOME | DEATH ROW | Death Row Information

Executed Offenders

Execution	Link	Link	Last Name	First Name	TDCJ Number	Age	Date	Race	County
503	Offender Information	Last Statement	Feldman	Douglas	999326	55	07/31/2013	White	Dallas
502	Offender Information	Last Statement	Ross	Vaughn	999429	41	07/18/2013	Black	Lubbock
501	Offender Information	Last Statement	Quintanilla	John	999491	36	07/16/2013	Hispanic	Victoria
500	Offender Information	Last Statement	McCarthy	Kimberly	999287	52	06/26/2013	Black	Dallas
499	Offender Information	Last Statement	Chester	Elroy	999280	43	06/12/2013	Black	Jefferson
498	Offender Information	Last Statement	Williams	Jeffrey	999350	37	05/15/2013	Black	Harris
497	Offender Information	Last Statement	Parr	Carroll	999479	35	05/07/2013	Black	McLennan
496	Offender Information	Last Statement	Cobb	Richard	999467	29	04/26/2013	White	Cherokee

Due by 9am on Sep 11

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

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

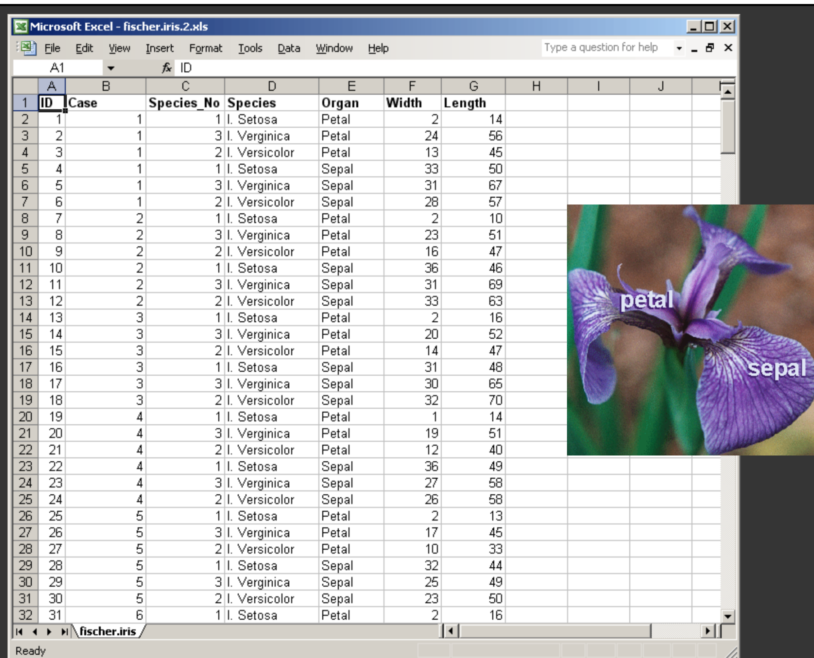
Data type

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

Iris Setosa

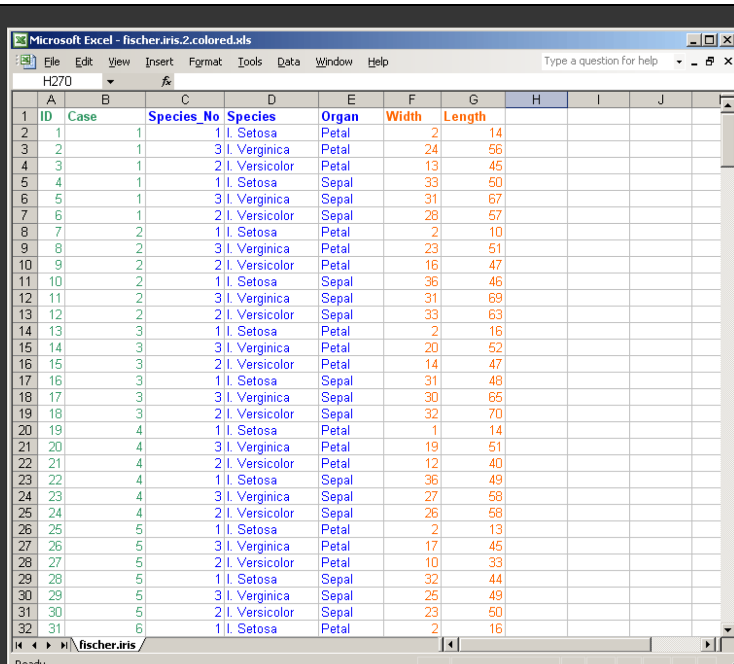
Iris Versicolor

Iris Virginica

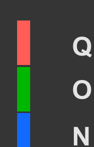


ID	Case	Species_No	Species	Organ	Width	Length
1	1	1	I. Setosa	Petal	2	14
2	1	3	I. Verginica	Petal	24	56
3	1	2	I. Versicolor	Petal	13	45
4	1	1	I. Setosa	Sepal	33	50
5	1	3	I. Verginica	Sepal	31	67
6	1	2	I. Versicolor	Sepal	28	57
7	2	1	I. Setosa	Petal	2	10
8	2	3	I. Verginica	Petal	23	51
9	2	2	I. Versicolor	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

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



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)

Relational algebra [Codd]

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 (INNNER 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)**

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	1450376
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	1077133
9	1850	15	0	2	1110619
10	1850	20	0	1	1017281
11	1850	20	0	2	1003841
12	1850	25	0	1	862547
13	1850	25	0	2	799482
14	1850	30	0	1	730638
15	1850	30	0	2	639636
16	1850	35	0	1	588487
17	1850	35	0	2	505012
18	1850	40	0	1	475911
19	1850	40	0	2	428185
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	162236
28	1850	65	0	1	106827
29	1850	65	0	2	105534
30	1850	70	0	1	73677
31	1850	70	0	2	71762
32	1850	75	0	1	40834
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	5259
39	1850	90	0	2	6569
40	1860	0	0	1	2120846
41	1860	0	0	2	2092162

Census: N, O, Q?

People Count	Q-Ratio
Year	Q-Interval *(O)*
Age	Q-Ratio *(O)*
Sex (M/F)	N
Marital Status	N

Census: Dimension or Measure?

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

Census: N, O, Q?

People	Q-Ratio
Year	Q-Interval (O)
Age	Q-Ratio (O)
Sex	N
Marital Status	N

Roll-Up and Drill-Down

Want to examine marital status in each decade?

Roll-up the data along the desired dimensions

Dimensions (year, marst) — Measure (sum(people))

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

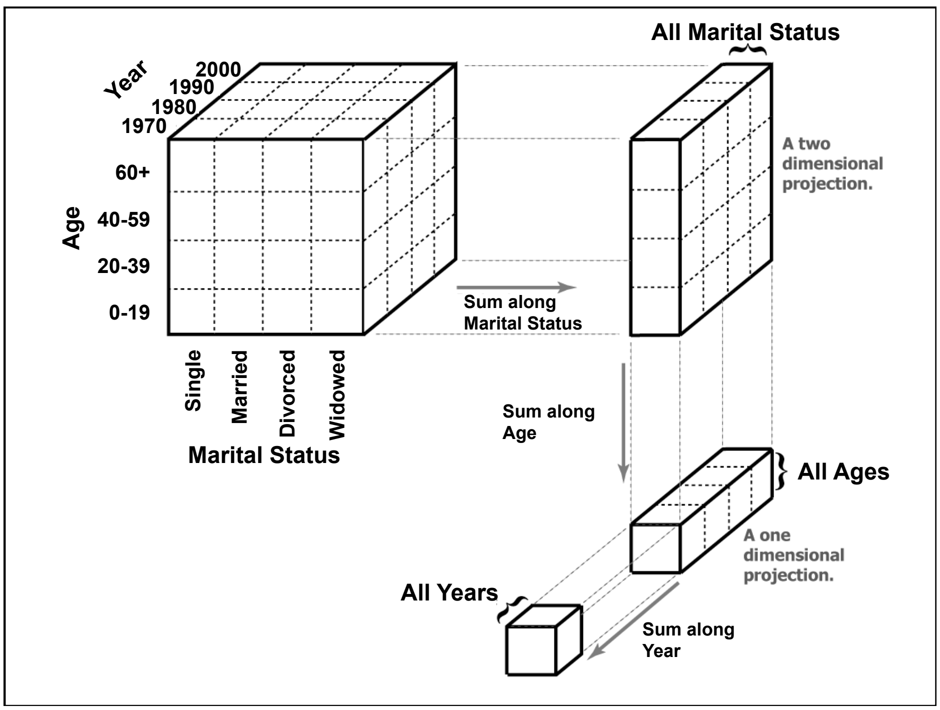
Dimensions (year, marst)

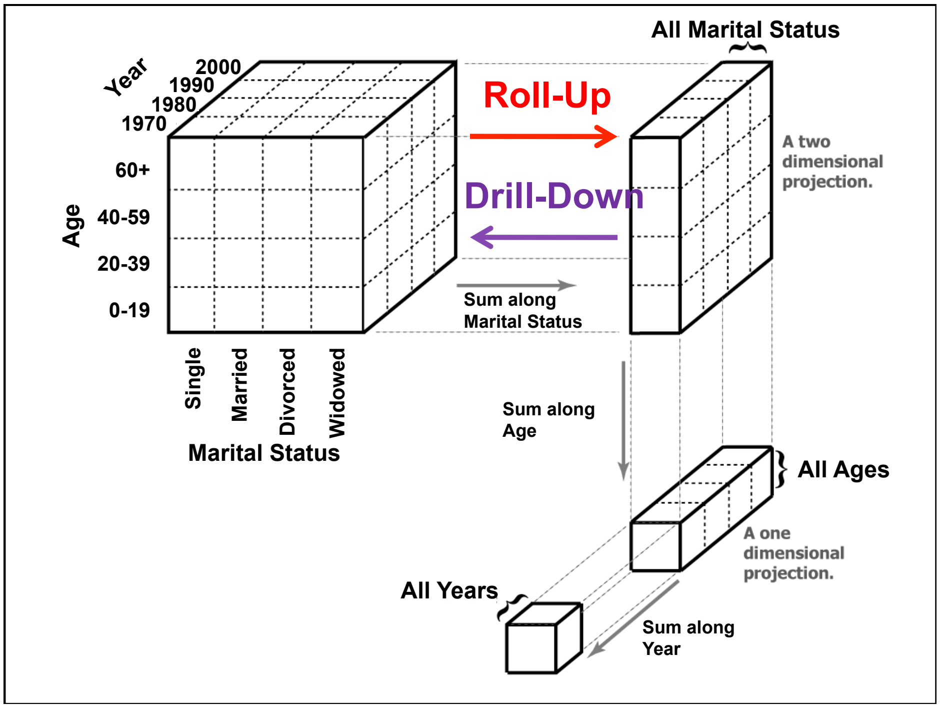
Roll-Up and Drill-Down

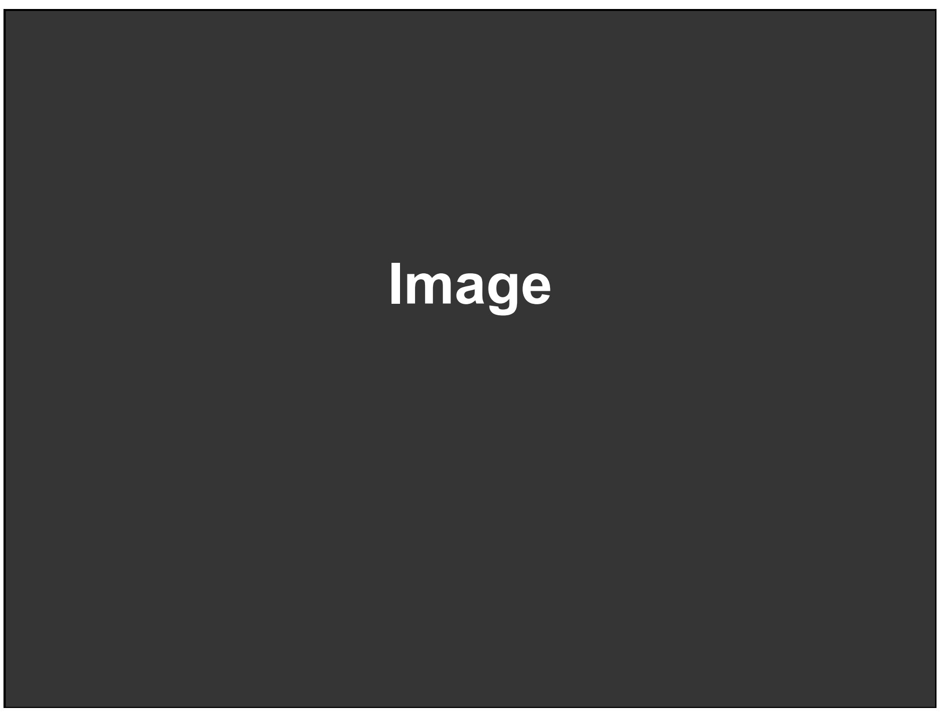
Need more detailed information?

Drill-down into additional dimensions

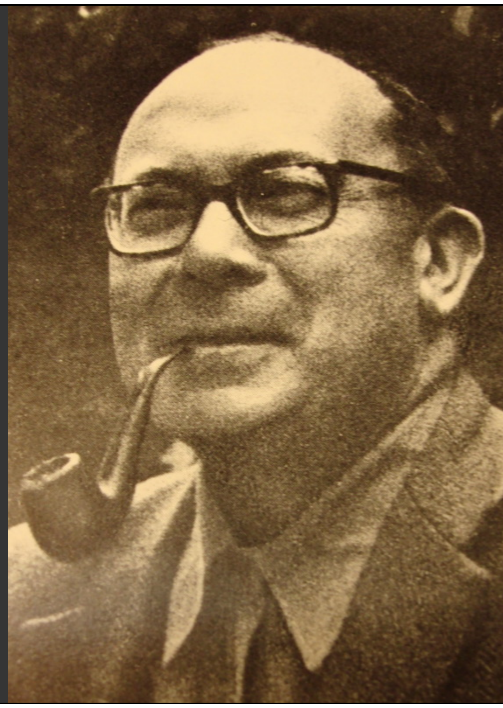
```
SELECT year, age, marst, sum(people)
FROM census
GROUP BY year, age, marst;
```




All Marital Status
Year
2000
1990
1980
1970
Roll-Up
60+
40-59
20-39
0-19
Age
Drill-Down
A two dimensional projection.
Sum along Marital Status
Single
Married
Divorced
Widowed
Marital Status
Sum along Age
All Ages
A one dimensional projection.
All Years
Sum along Year



Image



Visual language is a sign system

Jacques Bertin

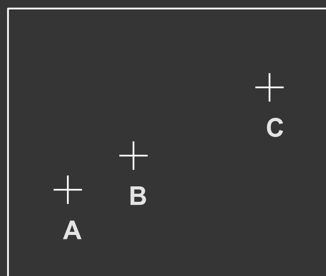
Images perceived as a set of signs

Sender encodes information in signs

Receiver decodes information from signs

Semiology of Graphics, 1967

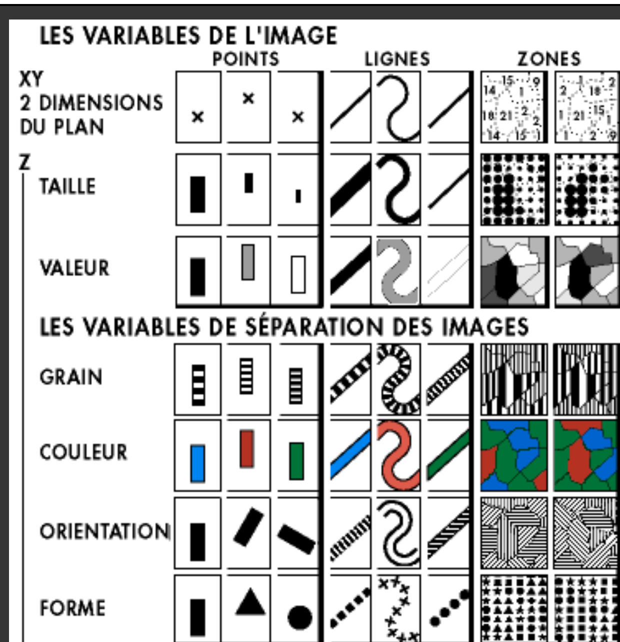
Information in position



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

∴ Encode quantitative variables

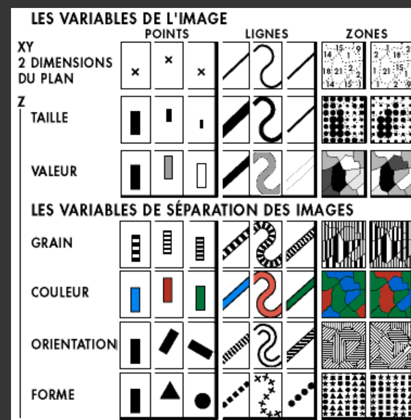
"Resemblance, order and proportional are the three signfields in graphics." - Bertin



[Bertin, Semiology of Graphics, 1983]

Visual variables

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



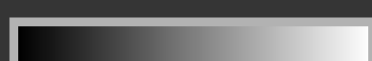
Note: Bertin does not consider 3D or time
Note: Card and Mackinlay extend the number of vars.

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

Bertins' "Levels of Organization"

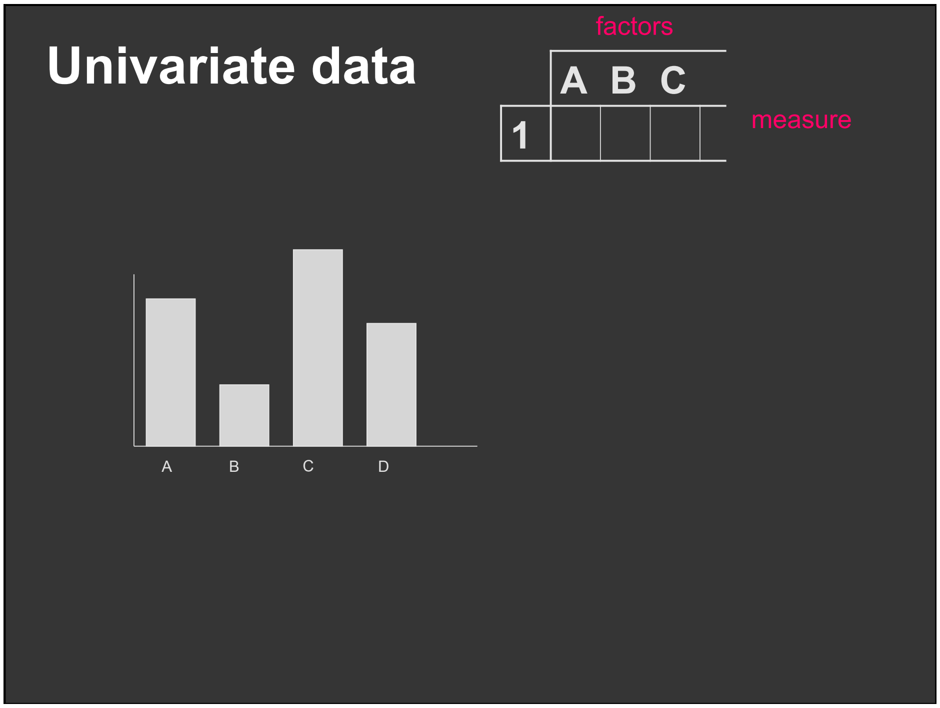
Position	N	O	Q
Size	N	O	Q
Value	N	O	Q
Texture	N	O	
Color	N		
Orientation	N		
Shape	N		

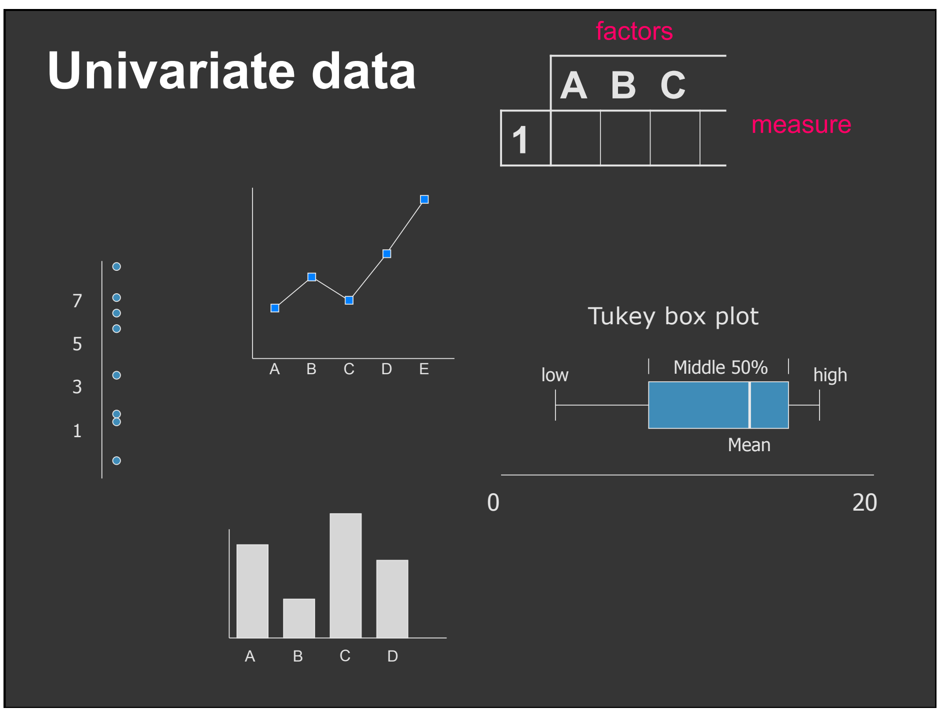
N Nominal
O Ordered
Q Quantitative

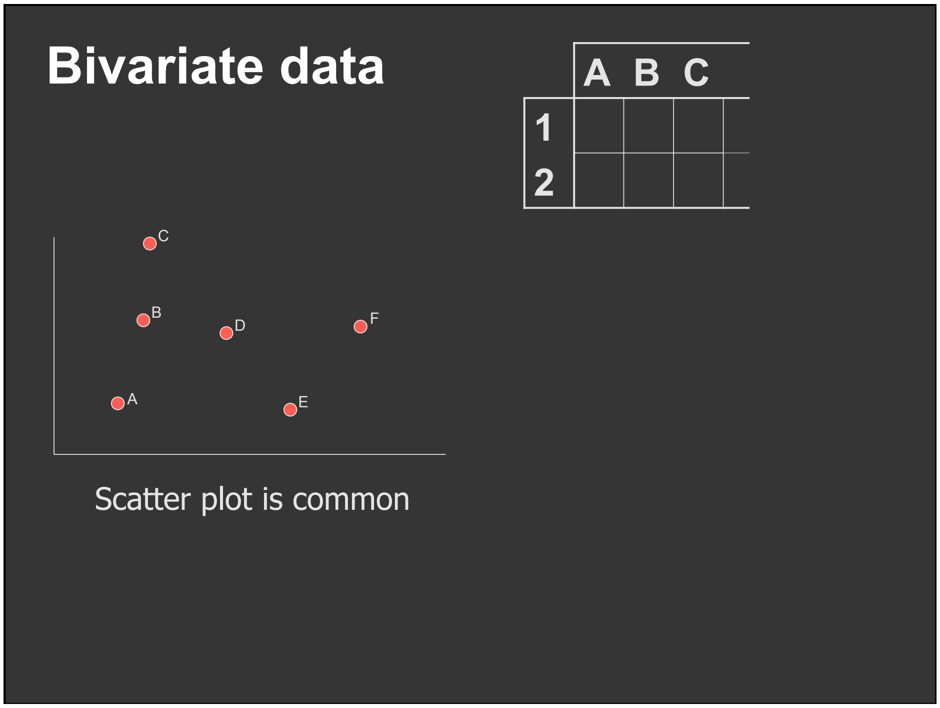
Note: Q < O < N

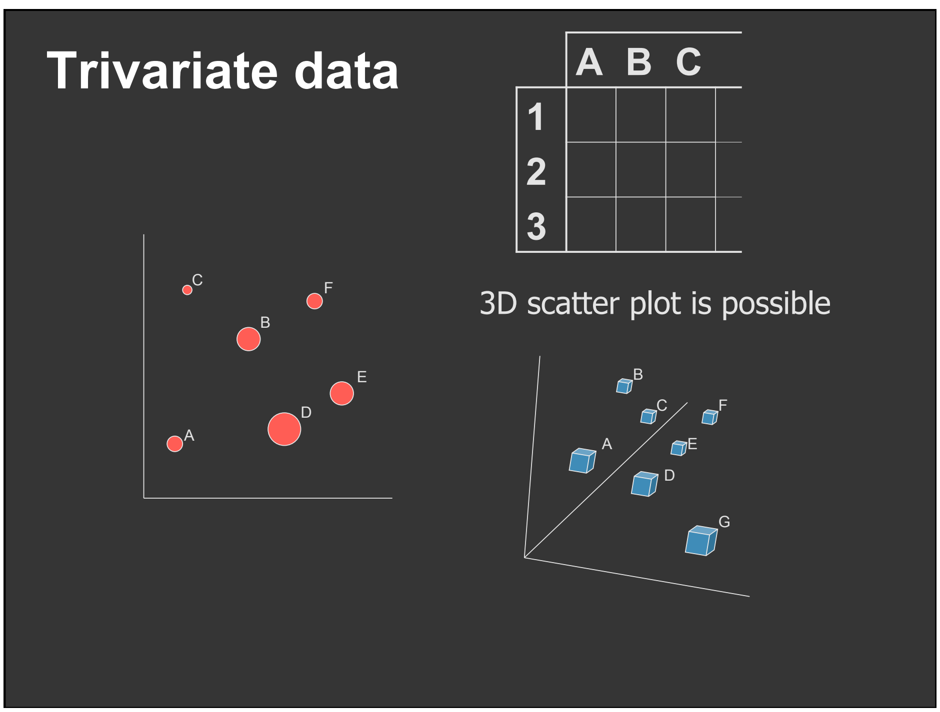
Note: Bertin actually breaks visual variables down into differentiating (≠) and associating (≡)

Encoding rules


Univariate data
factors
A B C
1
measure
A
B
C
D


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
C
F
B
E
D
A
3D scatter plot is possible
B
C
F
A
E
D
G

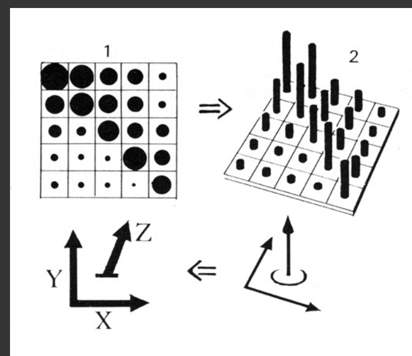
Three variables

Two variables [x,y] can map to points

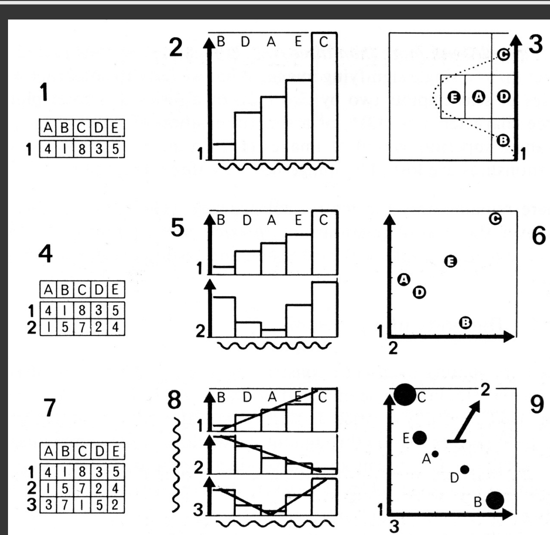
- Scatterplots, maps, …

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?

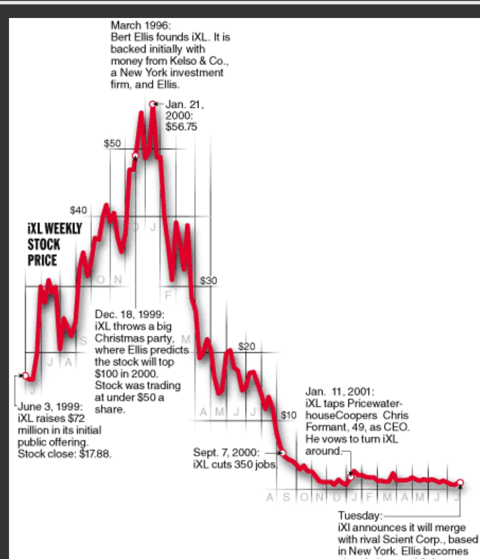
	A	B	C	
1				
2				
3				
4				
5				
6				
7				
8				

"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."

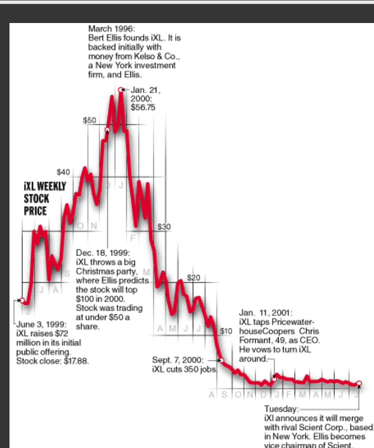
Bertin

Deconstructions

Stock chart from the late 90s

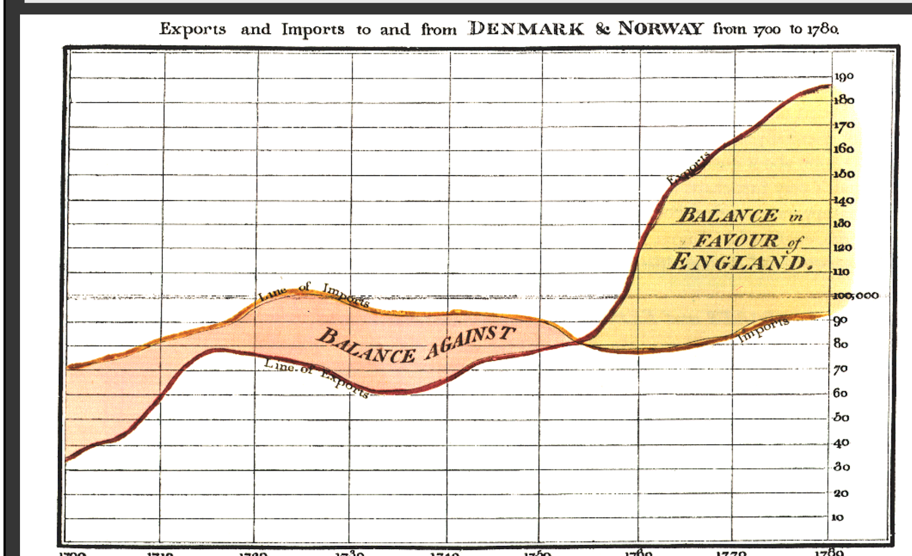


Stock chart from the late 90s

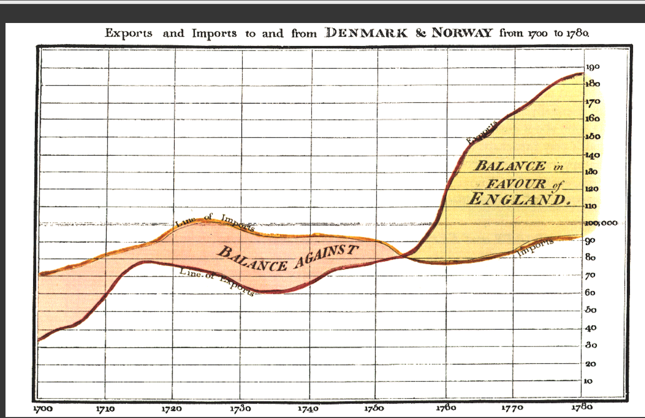


- x-axis: time (Q)
- y-axis: price (Q)

Playfair 1786



Playfair 1786



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