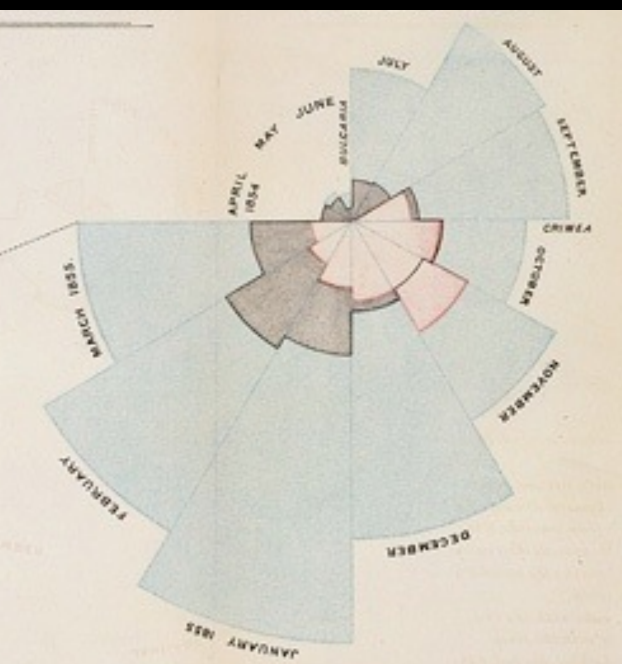
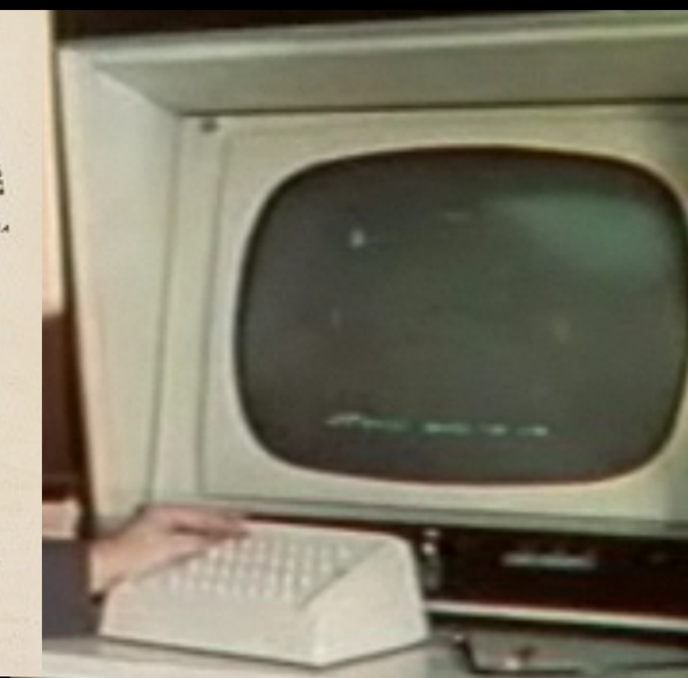
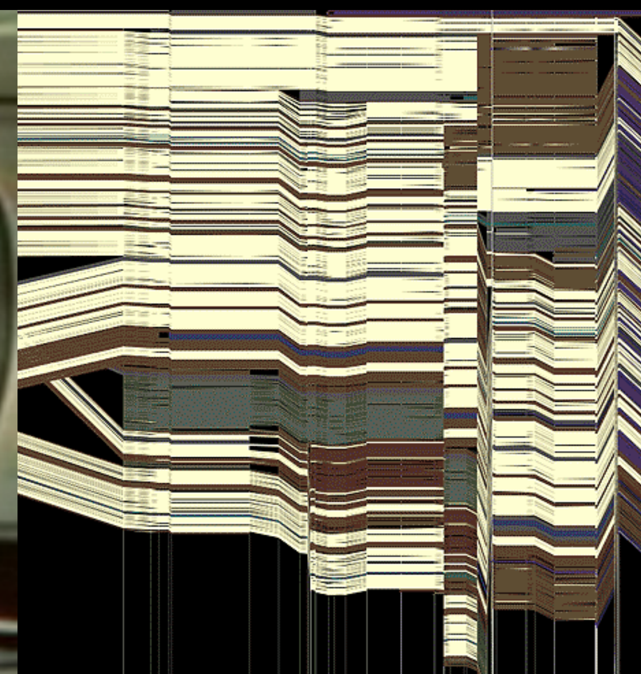
CSE 512 - Data Visualization

Data and Image Models







Jeffrey Heer University of Washington

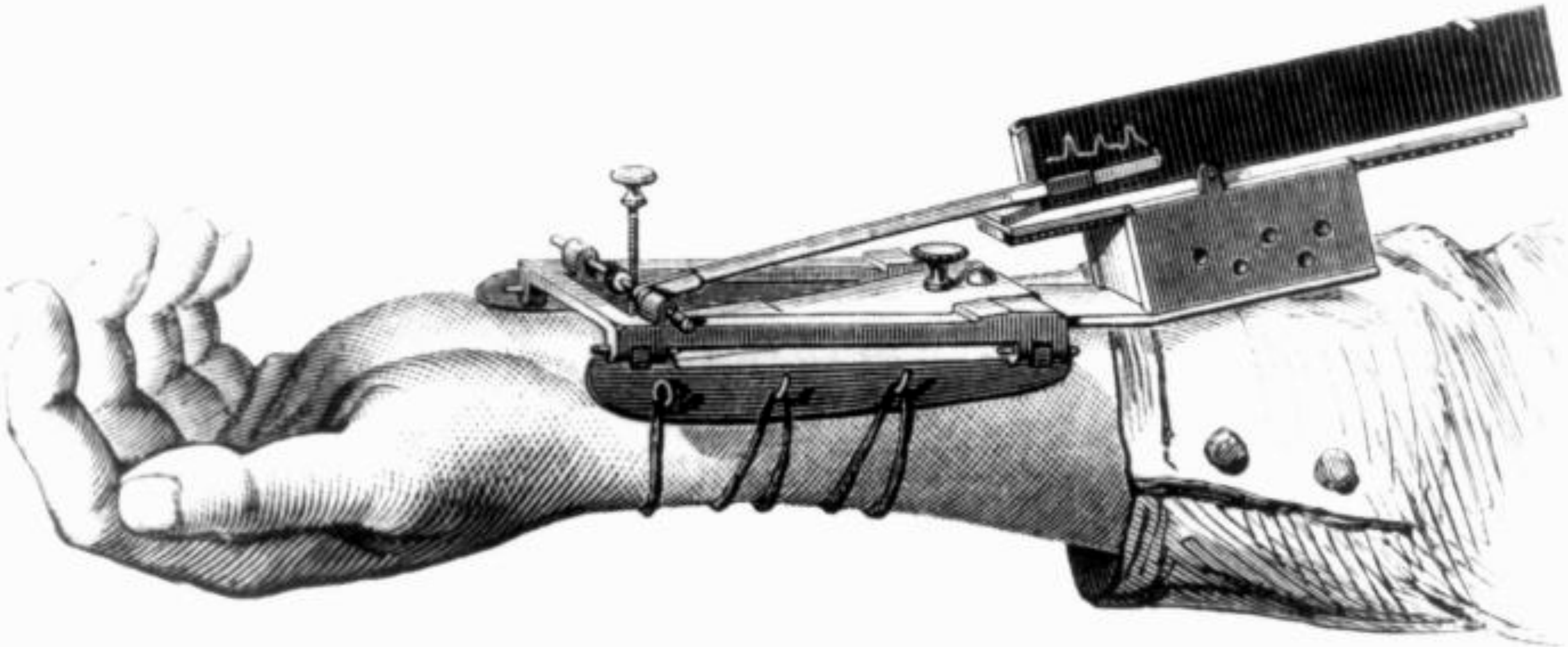
Last Time:
Value of Visualization

The Value of Visualization

Record information
- Blueprints, photographs, seismographs, …

Analyze data to support reasoning
- Develop and assess hypotheses
- Find patterns / Discover errors in data
- Expand memory

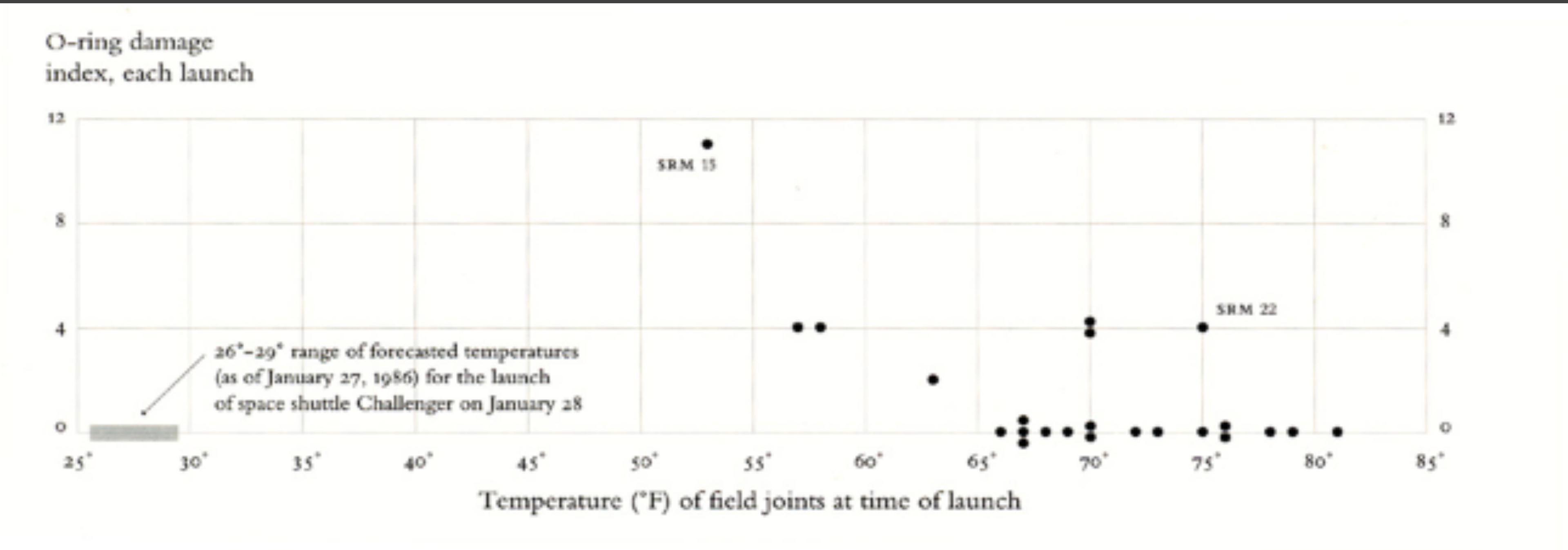
Communicate information to others
- Share and persuade
- Collaborate and revise



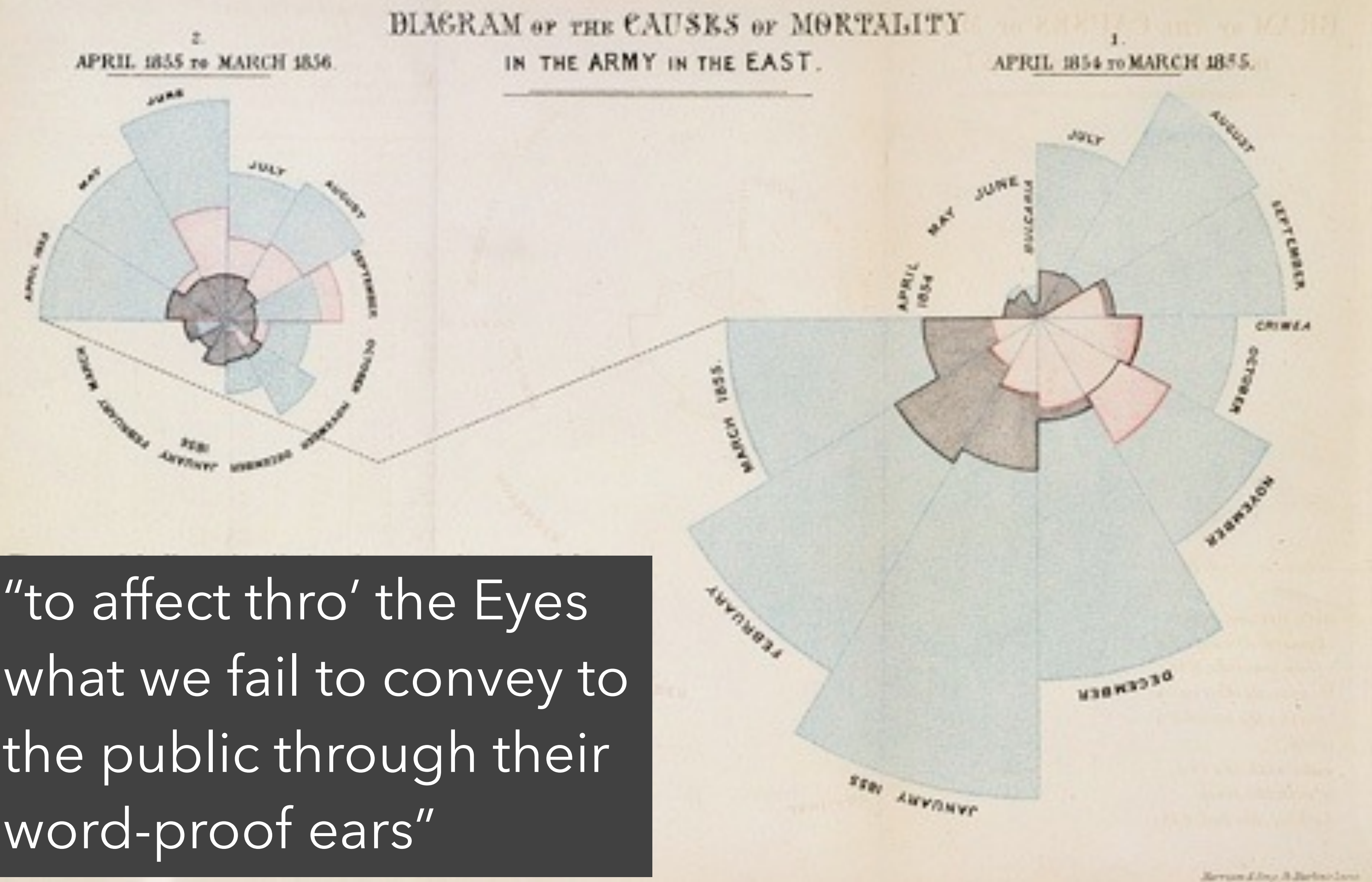
1.
Marey's sphygmograph in use, 1860. *La méthode graphique dans les sciences expérimentales et principalement en physiologie et en médecine.*

E.J. Marey's sphygmograph [from Braun 83]

Make a Decision: Challenger



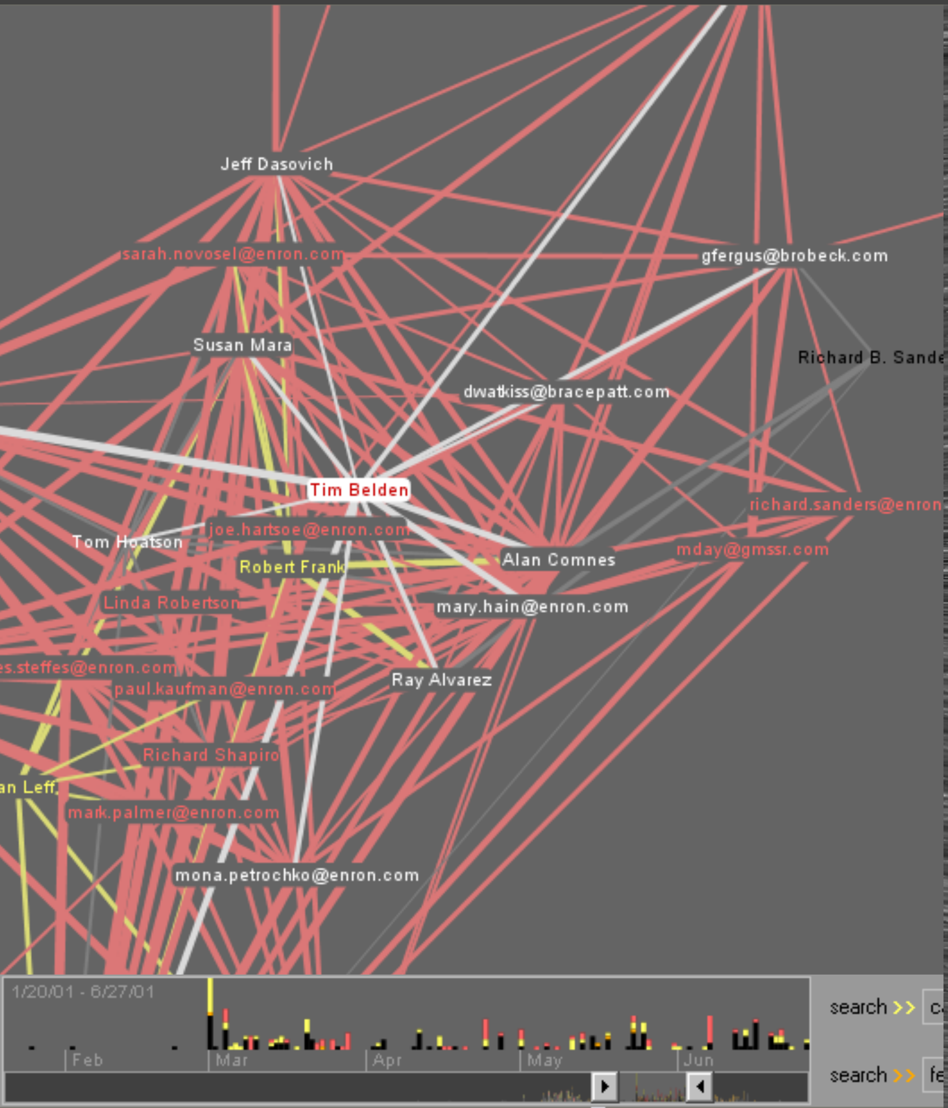
Visualizations drawn by Tufte show how low temperatures damage O-rings [Tufte 97]

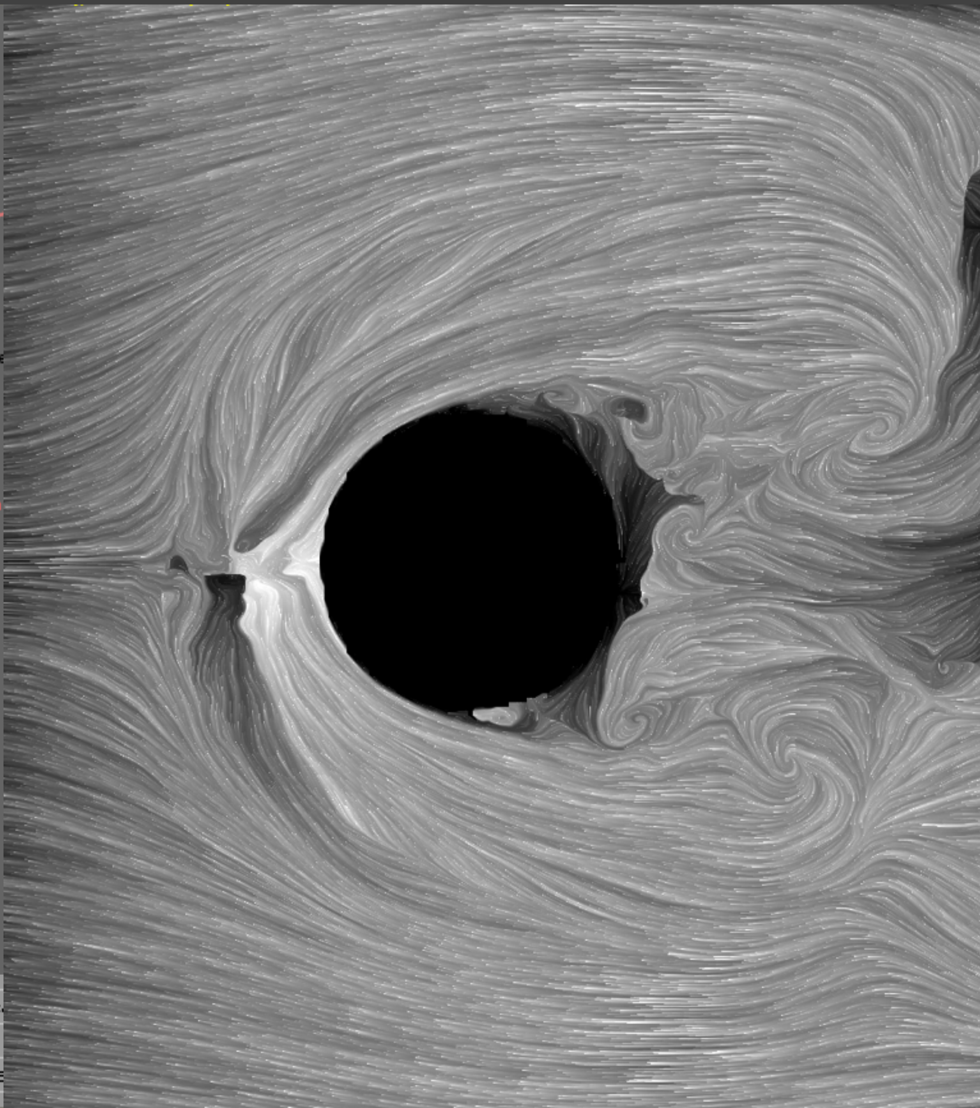


"to affect thro' the Eyes what we fail to convey to the public through their word-proof ears"

1856 "Coxcomb" of Crimean War Deaths, Florence Nightingale

InfoVis vs. SciVis?





Informative vs. Aesthetic?

Data & Image Models

The Big Picture

task
questions, goals
assumptions

data
physical data type
conceptual data type

domain
metadata
semantics
conventions

processing
algorithms

mapping
visual encoding

image
visual channel
graphical marks

Topics

Properties of Data

Properties of Images

Mapping Data to Images

Data

Data Models / Conceptual Models

Data models are low-level descriptions
Math: sets with operations on them
Example: integers with + and x operators

Conceptual models are mental constructions
Include semantics and support reasoning

Examples (data vs. conceptual)
1D floats vs. temperatures
3D vector of floats vs. spatial location

Taxonomy of Data Types (?)

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]

Nominal, Ordinal & Quantitative

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Fruits: apples, oranges, …

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Fruits: apples, oranges, …

O - Ordered

- Quality of meat: Grade A, AA, AAA

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

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

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

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

Q - Ratio (zero fixed)

- Physical measurement: Length, Mass, Temp, …
- Counts and amounts

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- 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

From Data Model to N, O, Q

Data Model

32.5, 54.0, -17.3, …

Floating point numbers

Conceptual Model

Temperature (°C)

Data Type

Burned vs. Not-Burned (N)

Hot, Warm, Cold (O)

Temperature Value (Q)

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].

Microsoft Excel - fischer.iris.2.colored.xls

File Edit View Insert Format Tools Data Window Help

H270

	A	B	C	D	E	F	G
1	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

fischer.iris

Ready

N

O

Q

Dimensions & Measures

Dimensions (~ independent variables)
Discrete variables describing data (N, O)
Categories, dates, binned quantities

Measures (~ dependent variables)
Data values that can be aggregated (Q)
Numbers to be analyzed
Aggregate as sum, count, avg, std. dev…

Example: U.S. Census Data

Example: U.S. Census Data

People Count: # 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 Count

Year

Age

Sex

Marital Status

2,348 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	N
Marital Status	N

Census: Dimension or Measure?

People Count	Measure
Year	Dimension
Age	Depends!
Sex	Dimension
Marital Status	Dimension

Data Transformation

Relational Data Model

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

Each **row** (*tuple*) represents a record
- Each record is a fixed-length tuple

Each **column** (*attribute*) represents a variable
- Each attribute has a *name* and a *data type*

A table's **schema** is the set of names and types

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

Relational Algebra [Codd '70]

Data Transformations (`sql`)

Projection (`select`) - selects columns

Selection (`where`) - filters rows

Sorting (`order by`)

Aggregation (`group by`, `sum`, `min`, `max`, …)

Combine relations (`union`, `join`, …)

Roll-Up and Drill-Down

Want to examine marital status in each decade?

Roll-up the data along the desired dimensions

Dimensions **Measure**

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

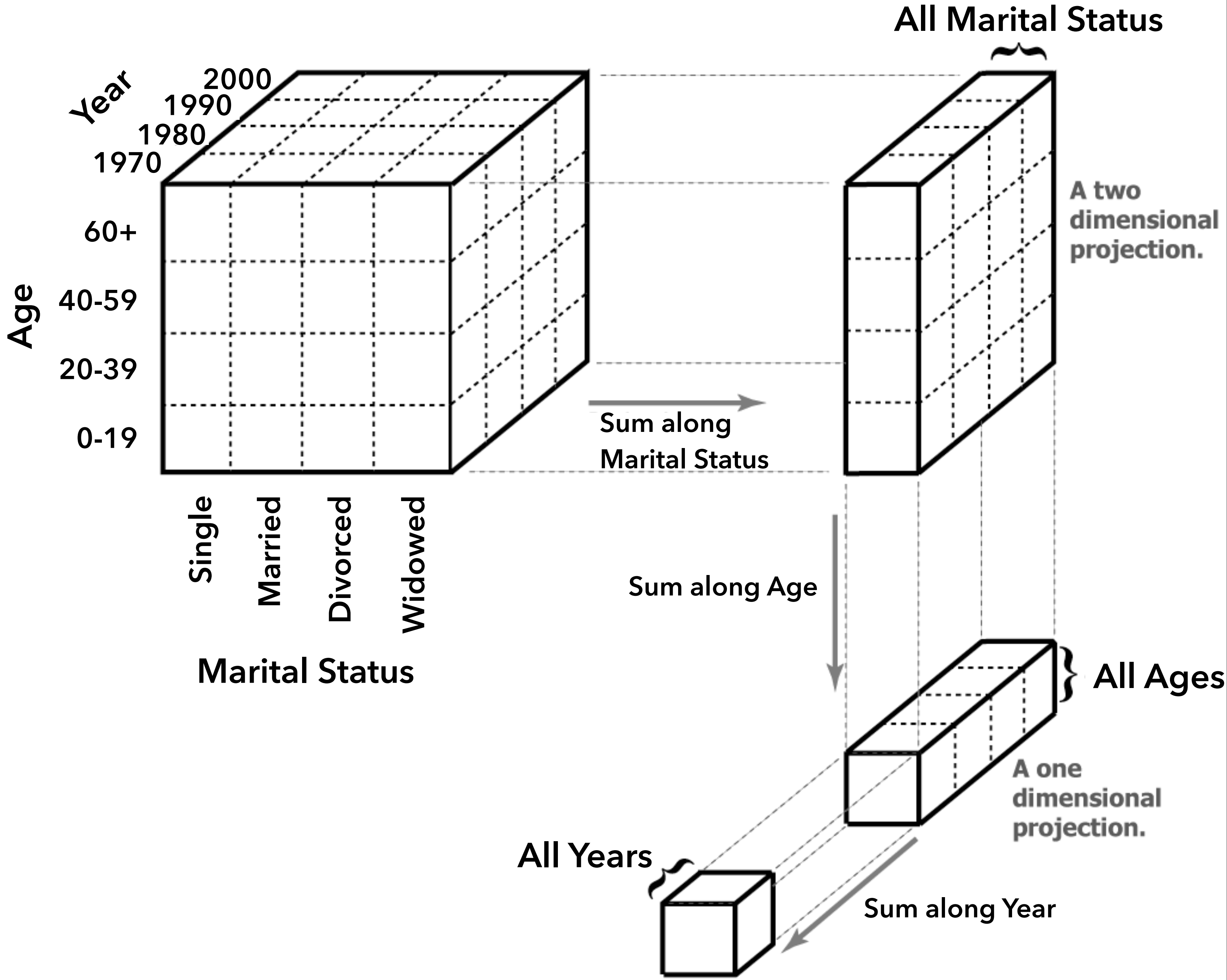
Dimensions

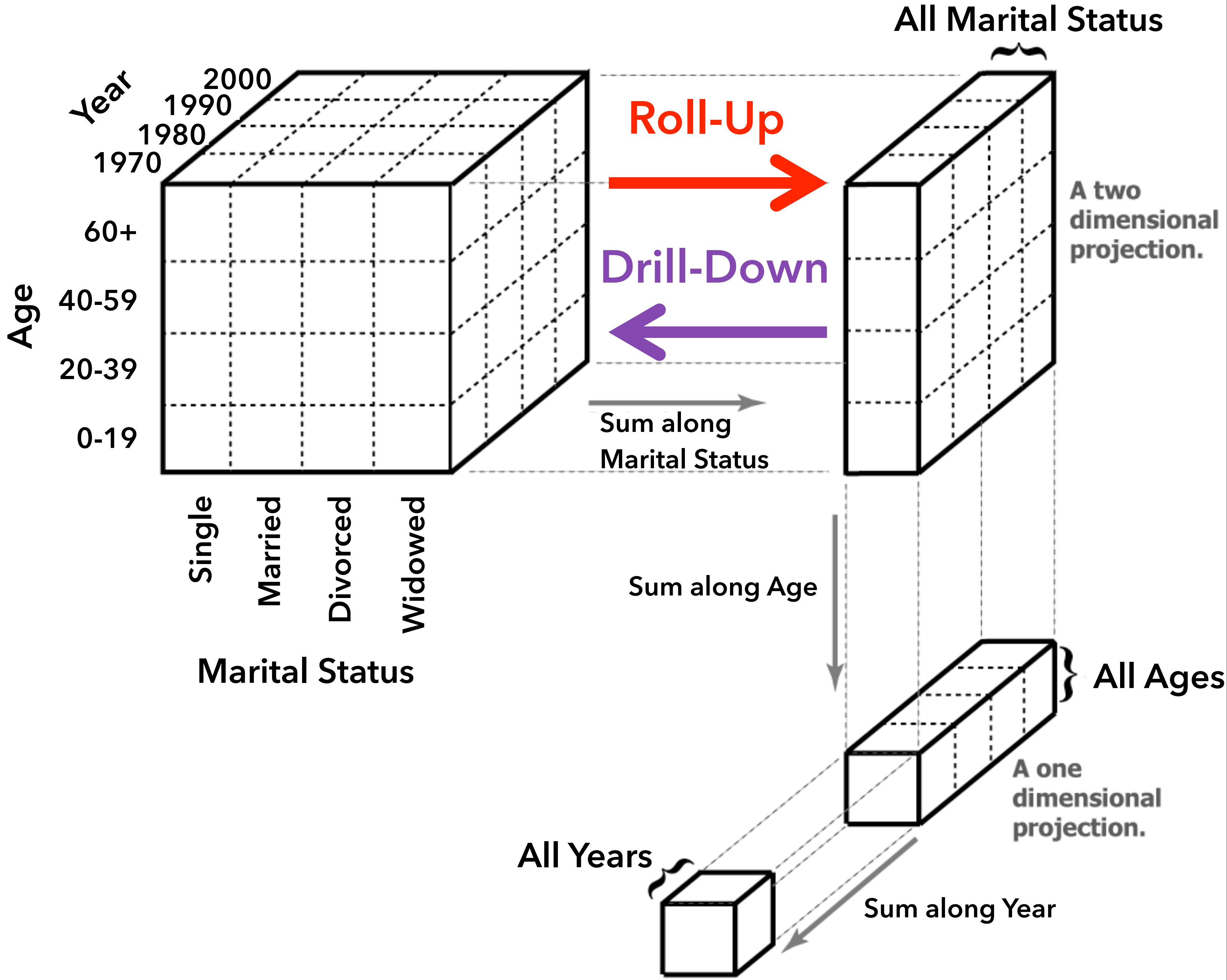
Roll-Up and Drill-Down

Need more detailed information?

Drill-down into additional dimensions

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


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


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

YEAR	AGE	MARST	SEX	PEOPLE
1850	0	0	1	1,483,789
1850	5	0	1	1,411,067
1860	0	0	1	2,120,846
1860	5	0	1	1,804,467
. . .				

AGE	MARST	SEX	1850	1860	. . .
0	0	1	1,483,789	2,120,846	. . .
5	0	1	1,411,067	1,804,467	. . .
. . .					

Which format might we prefer?

Administrivia

A1: Visualization Design

Design a static visualization for a data set.

After the World War II, antibiotics were considered “wonder drugs.” To learn which drug is most effective for which bacterial infection, performance of the three most popular antibiotics were gathered.

You must choose the message you want to convey. What task do you want to support? What insight do you want to communicate?

A1: Visualization Design

Design a **static visualization** for the data set.

You are free to **use any tools** (inc. pen & paper).

Deliverables (upload via Canvas; see A1 page)

Image of your visualization (PNG or JPG format)

Short description + design rationale (≤ 4 paragraphs)

Due by **5:00 pm, Monday April 6**.

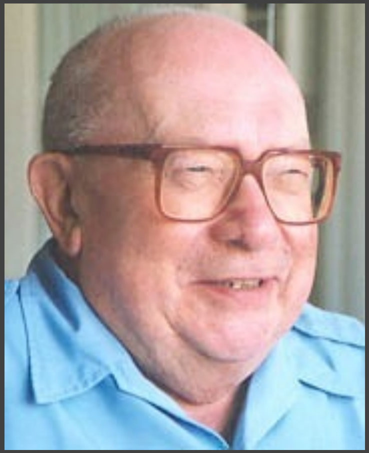
Next Tuesday: Design Exercise

We will **review A1 submissions**
So be sure to turn yours in on time!

We will then have a **redesign exercise**
*Please bring **paper, pens, etc for sketching***

Prof. Heer will be out attending **OpenVisConf**

Visual Language is a Sign System



Jacques Bertin

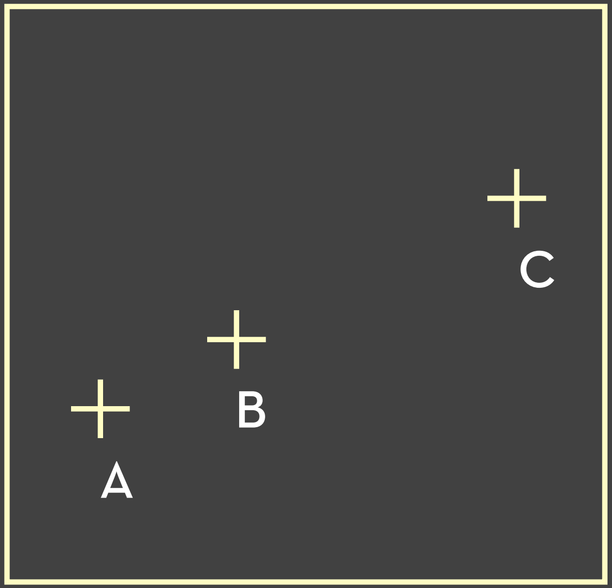
Images perceived as a set of signs

Sender encodes information in signs

Receiver decodes information from signs

Sémiologie Graphique, 1967

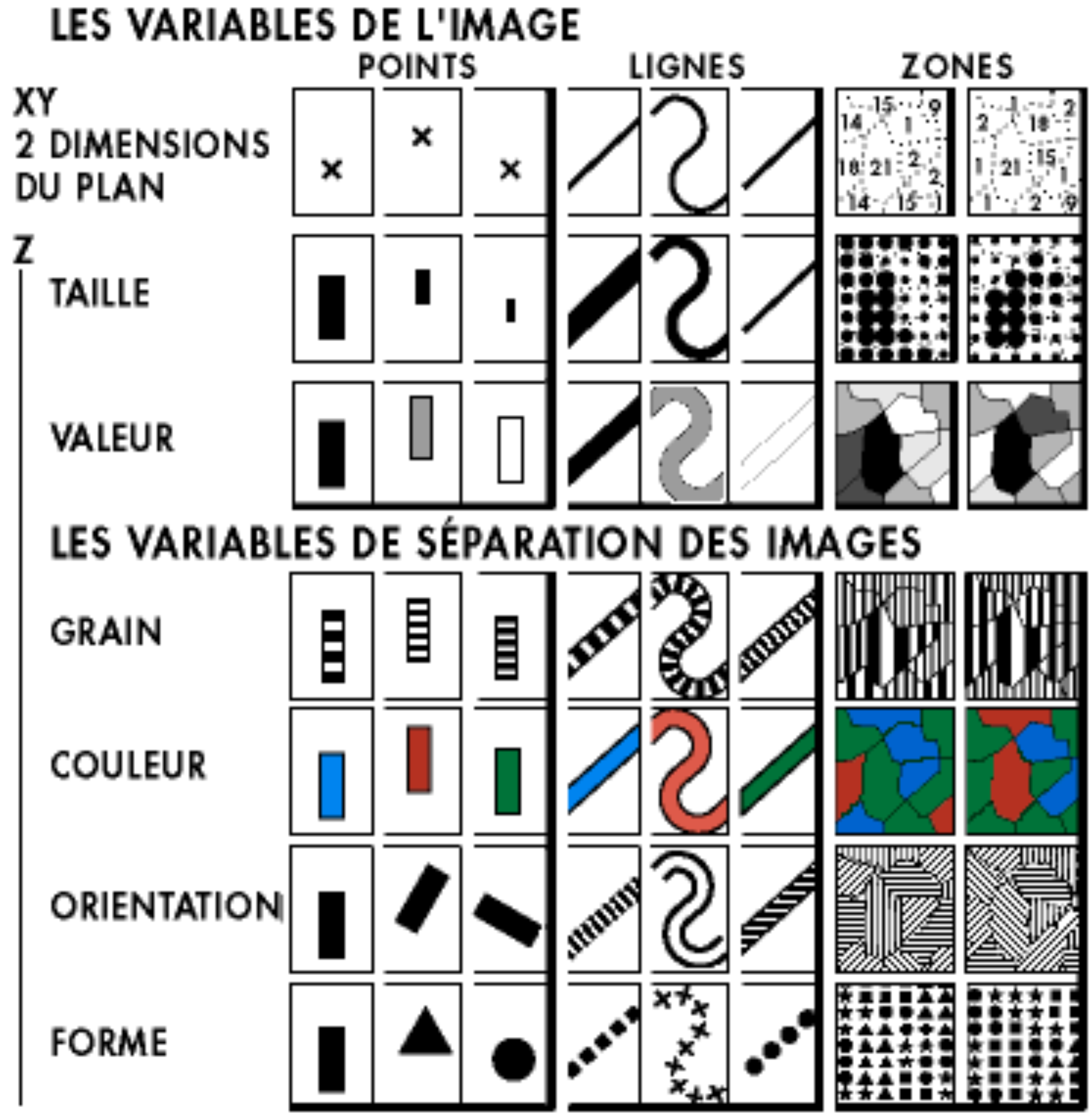
Bertin's Semiology of Graphics



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

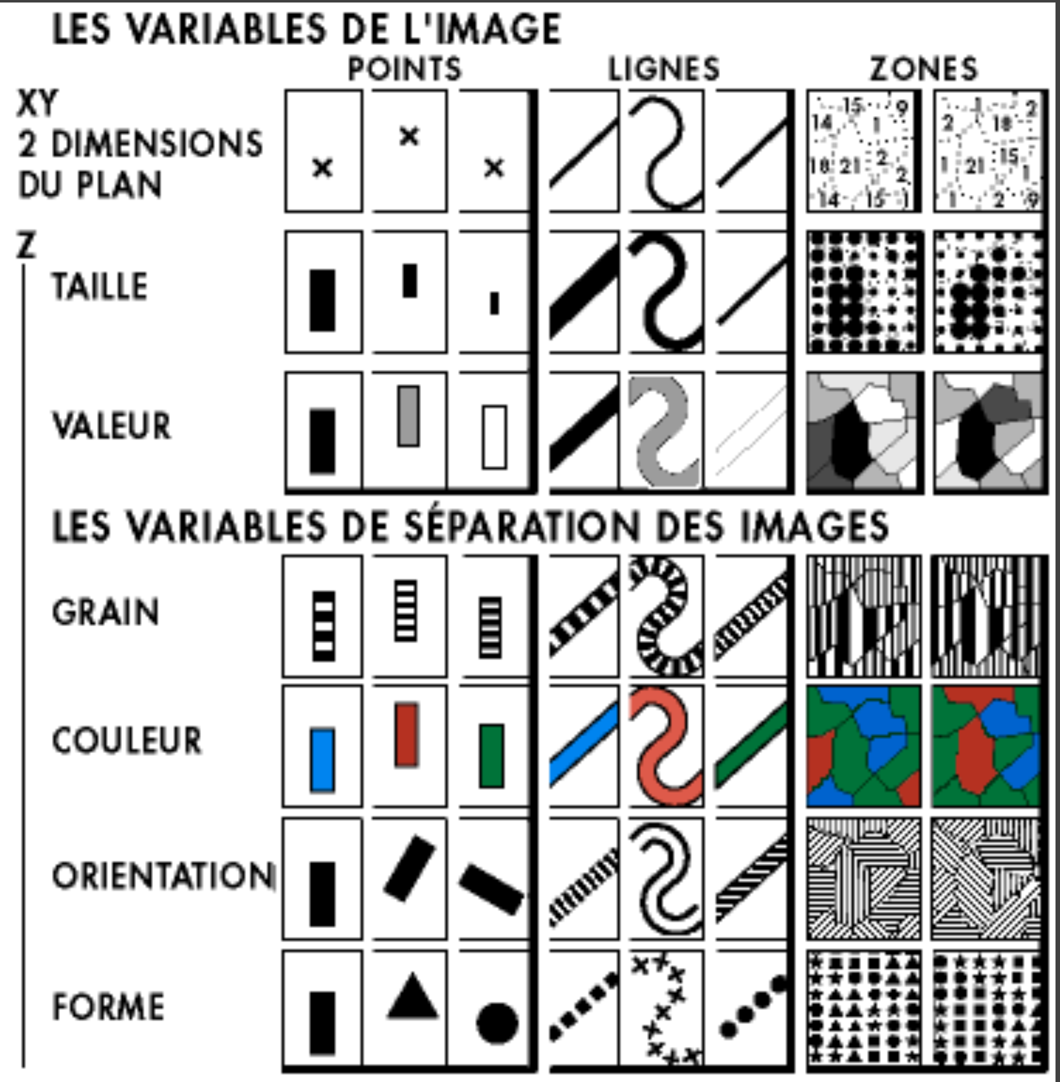
∴ Encode quantitative variables

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


LES VARIABLES DE L'IMAGE
POINTS
LIGNES
ZONES
XY
2 DIMENSIONS
DU PLAN
Z
TAILLE
VALEUR
LES VARIABLES DE SÉPARATION DES IMAGES
GRAIN
COULEUR
ORIENTATION
FORME

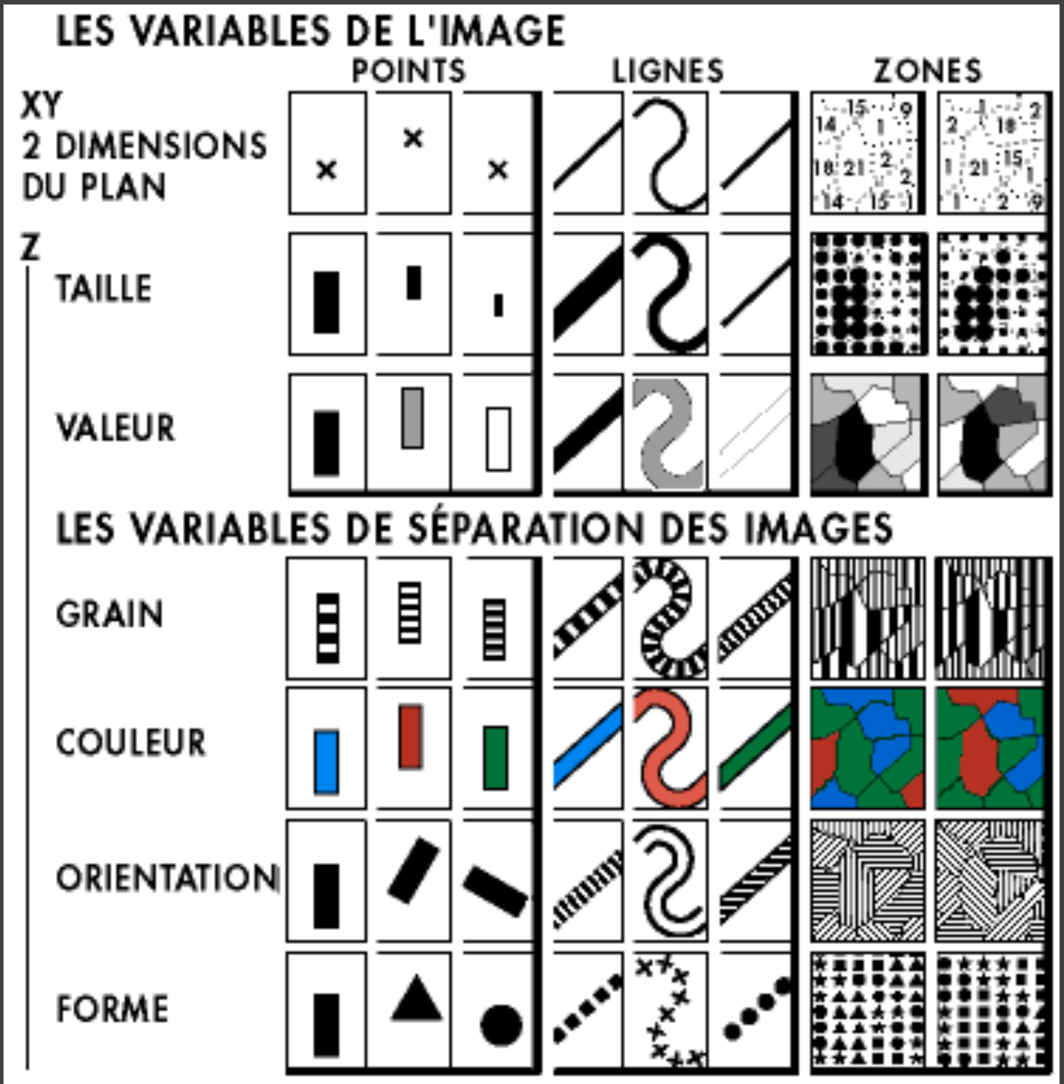
Visual Encoding Variables

Position (x 2)
Size
Value
Texture
Color
Orientation
Shape



Visual Encoding Variables

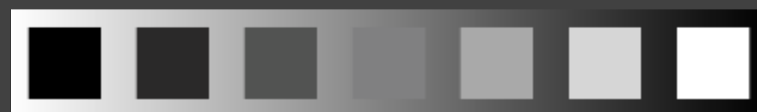
Position
Length
Area
Volume
Value
Texture
Color
Orientation
Shape
Transparency
Blur / Focus …



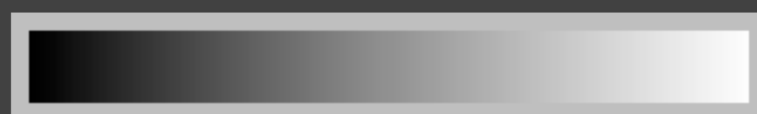
Information in Hue and Value

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

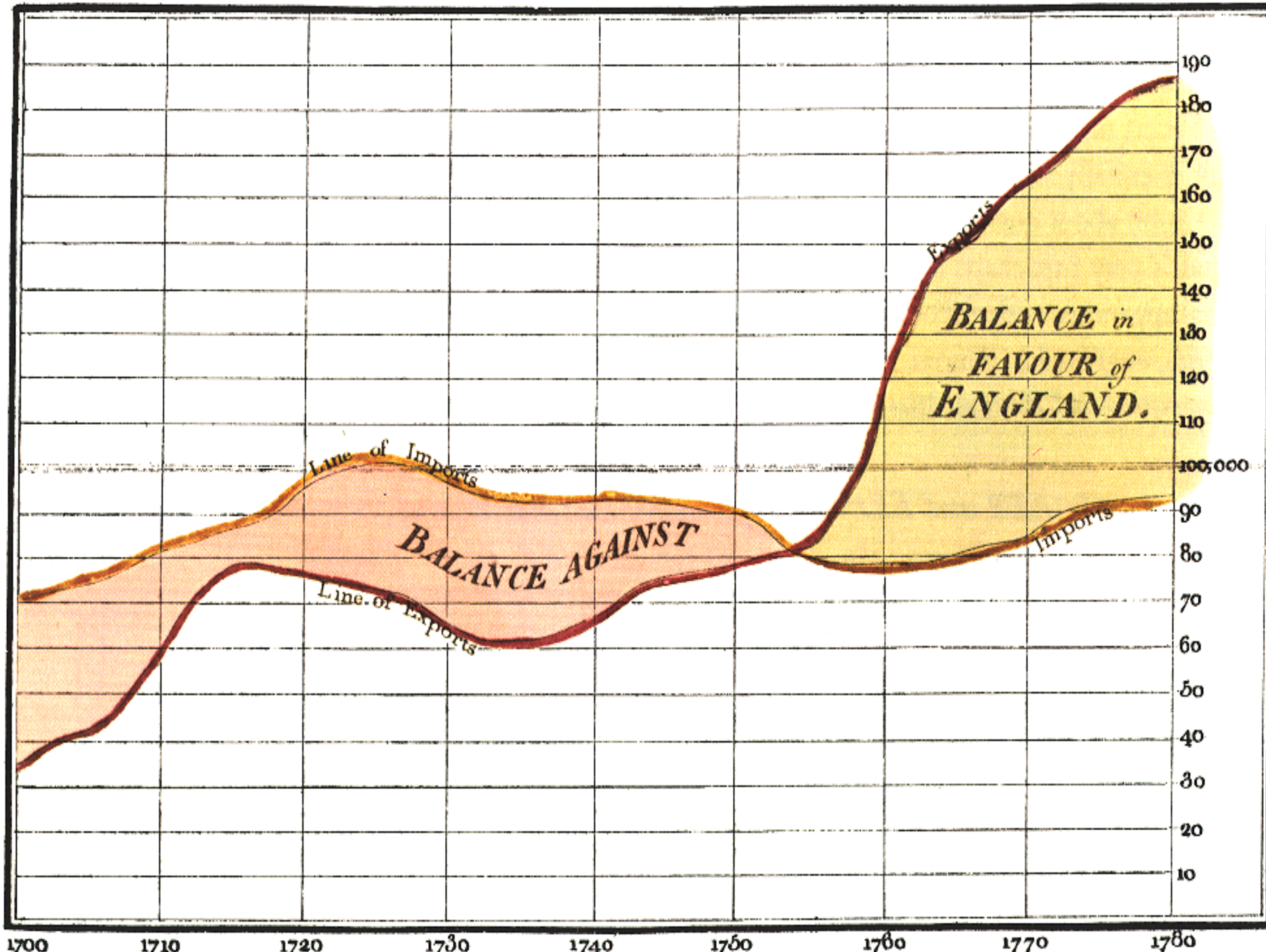
Ordinal

Quantitative

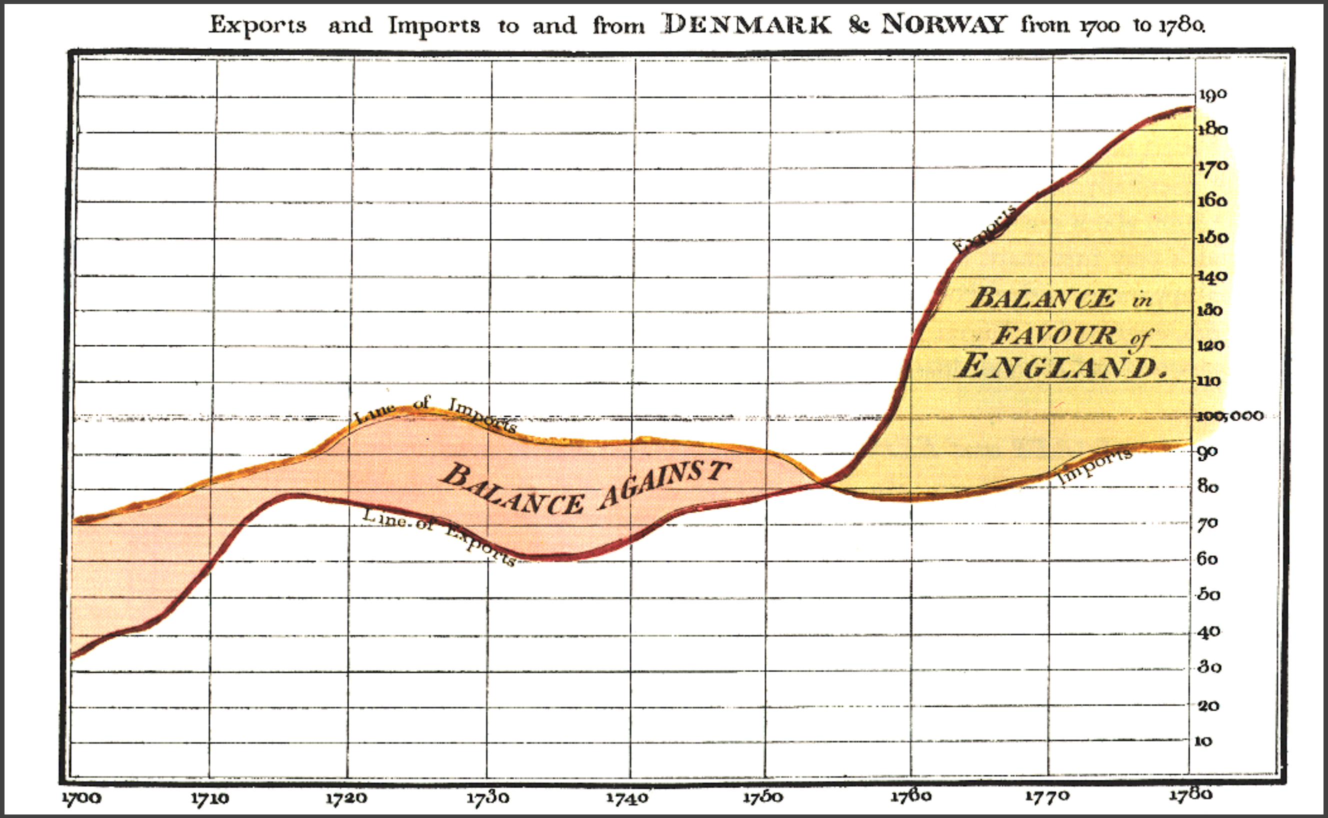
Note: **Q** $\subset$ **O** $\subset$ **N**

Deconstructions

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



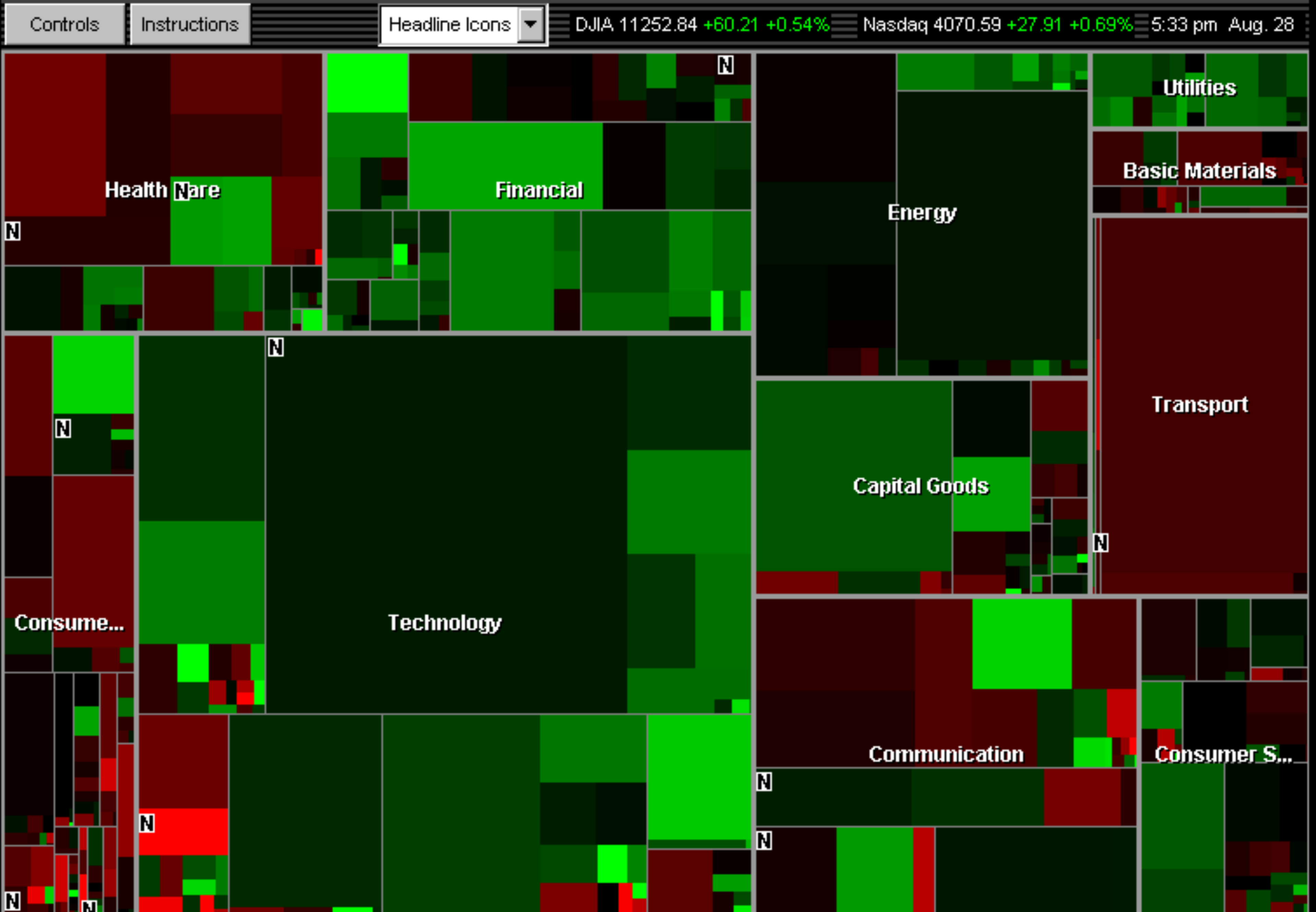
William Playfair, 1786



X-axis: year (Q)

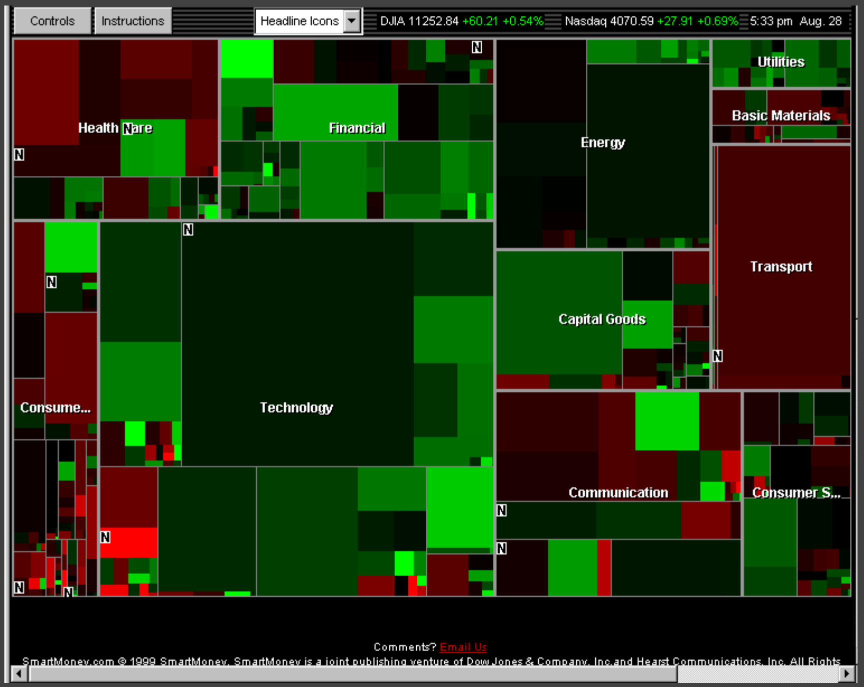
Y-axis: currency (Q)

Color: imports/exports (N, O)



http://www.smartmoney.com/marketmap/

Wattenberg's Map of the Market



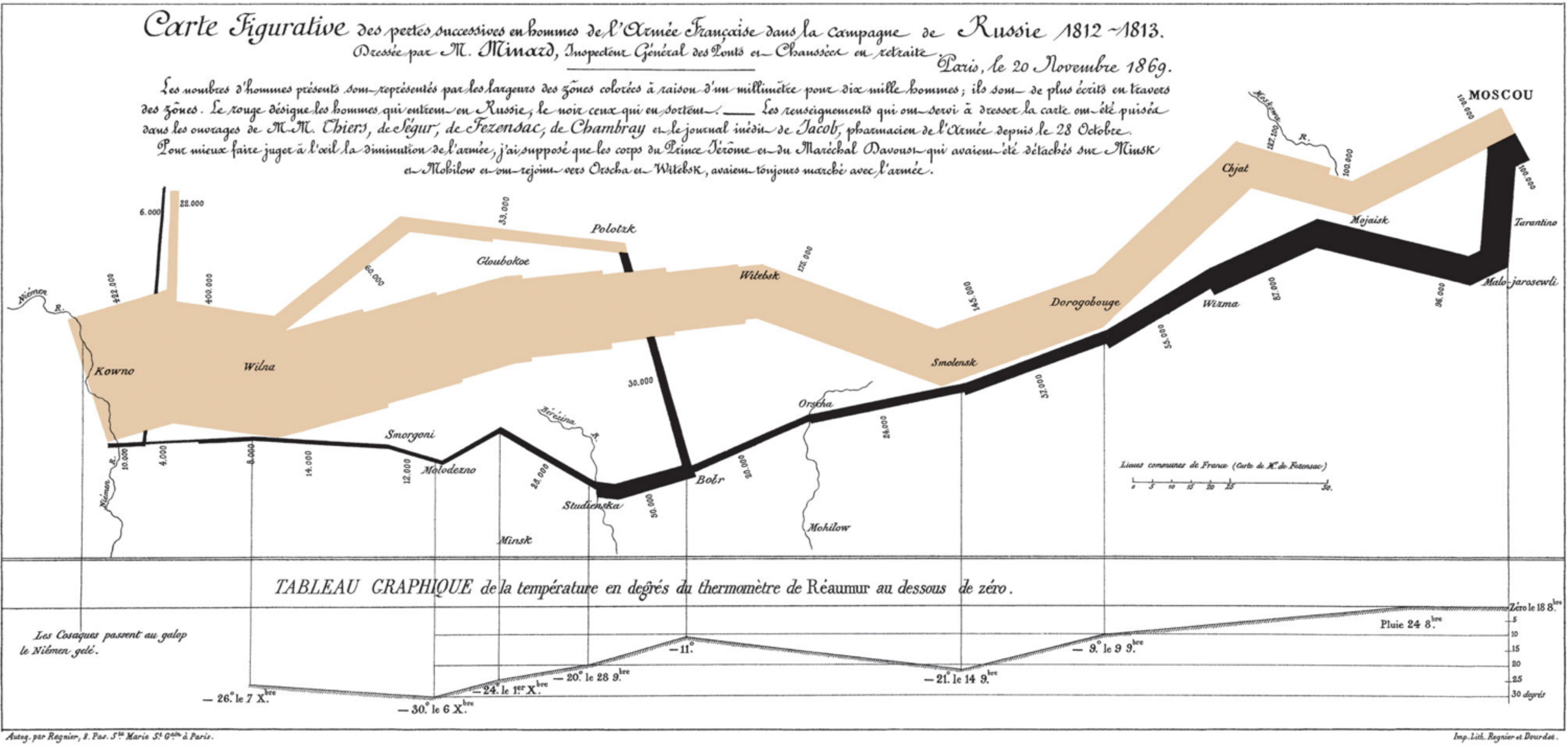
Rectangle Area: market cap (Q)

Rectangle Position: market sector (N), market cap (Q)

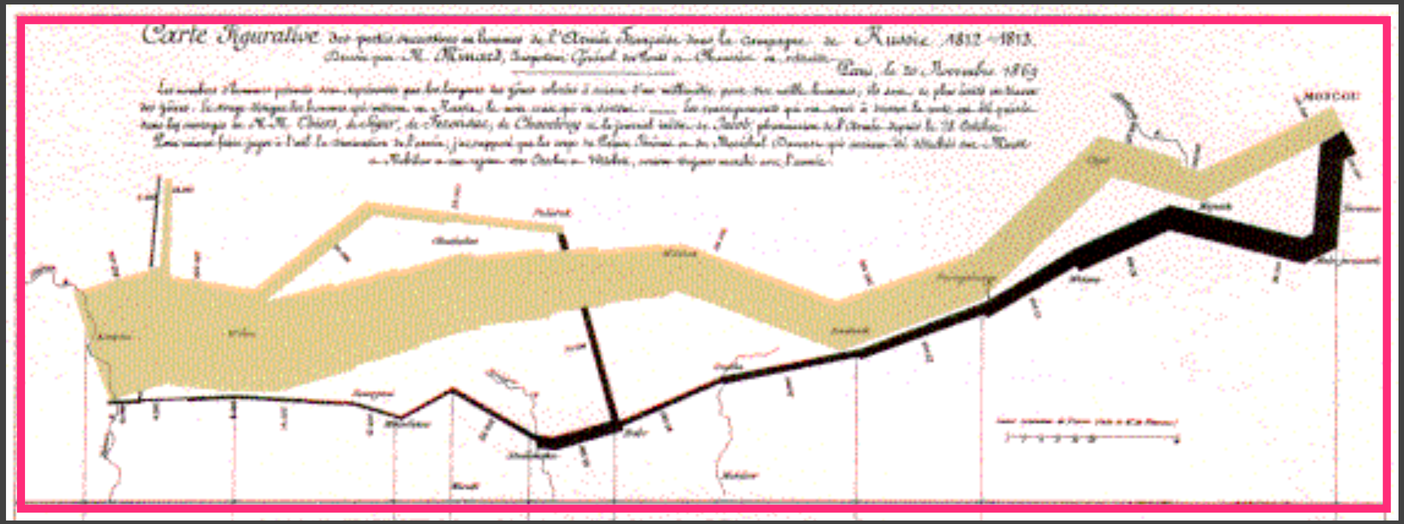
Color Hue: loss vs. gain (N, O)

Color Value: magnitude of loss or gain (Q)

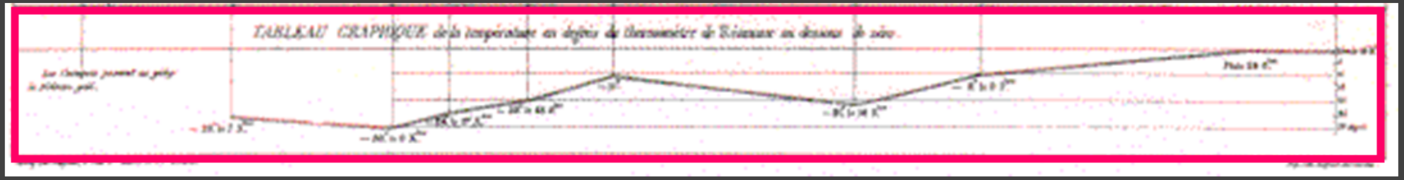
Minard 1869: Napoleon's March



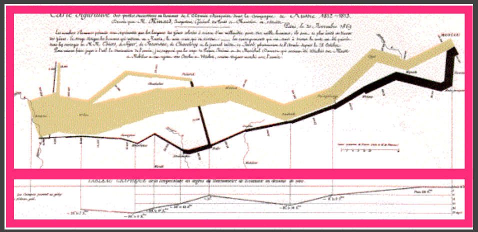
Single-Axis Composition



+



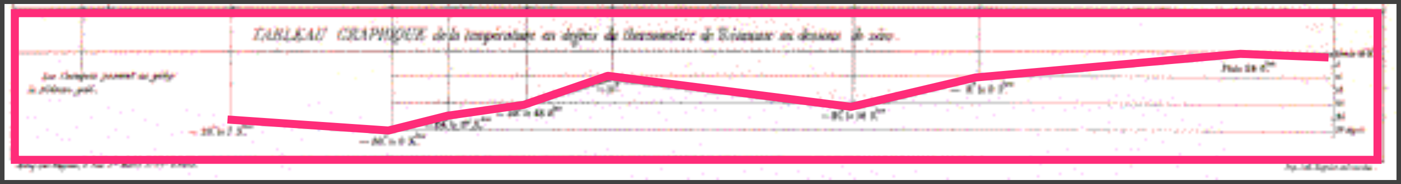
=



Mark Composition

Y-axis: temperature (Q)

\+ X-axis: longitude (Q) / time (O)

= 

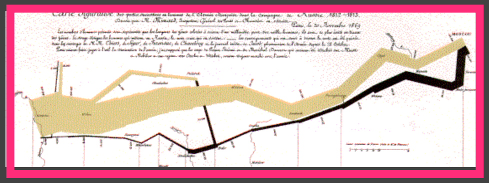
Temp over space/time (Q x Q)

Mark Composition

Y-axis: longitude (Q)

\+ X-axis: latitude (Q)

\+ Width: army size (Q)

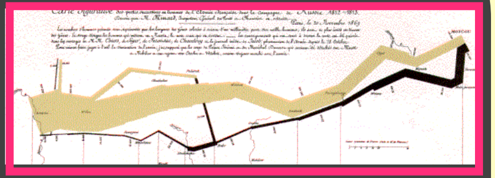
= 

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

longitude (Q)

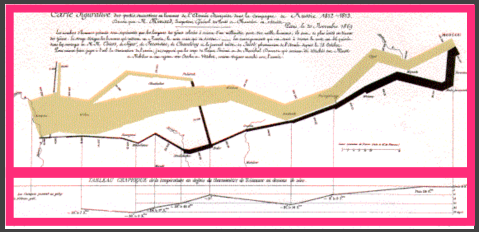
latitude (Q)

army size (Q)

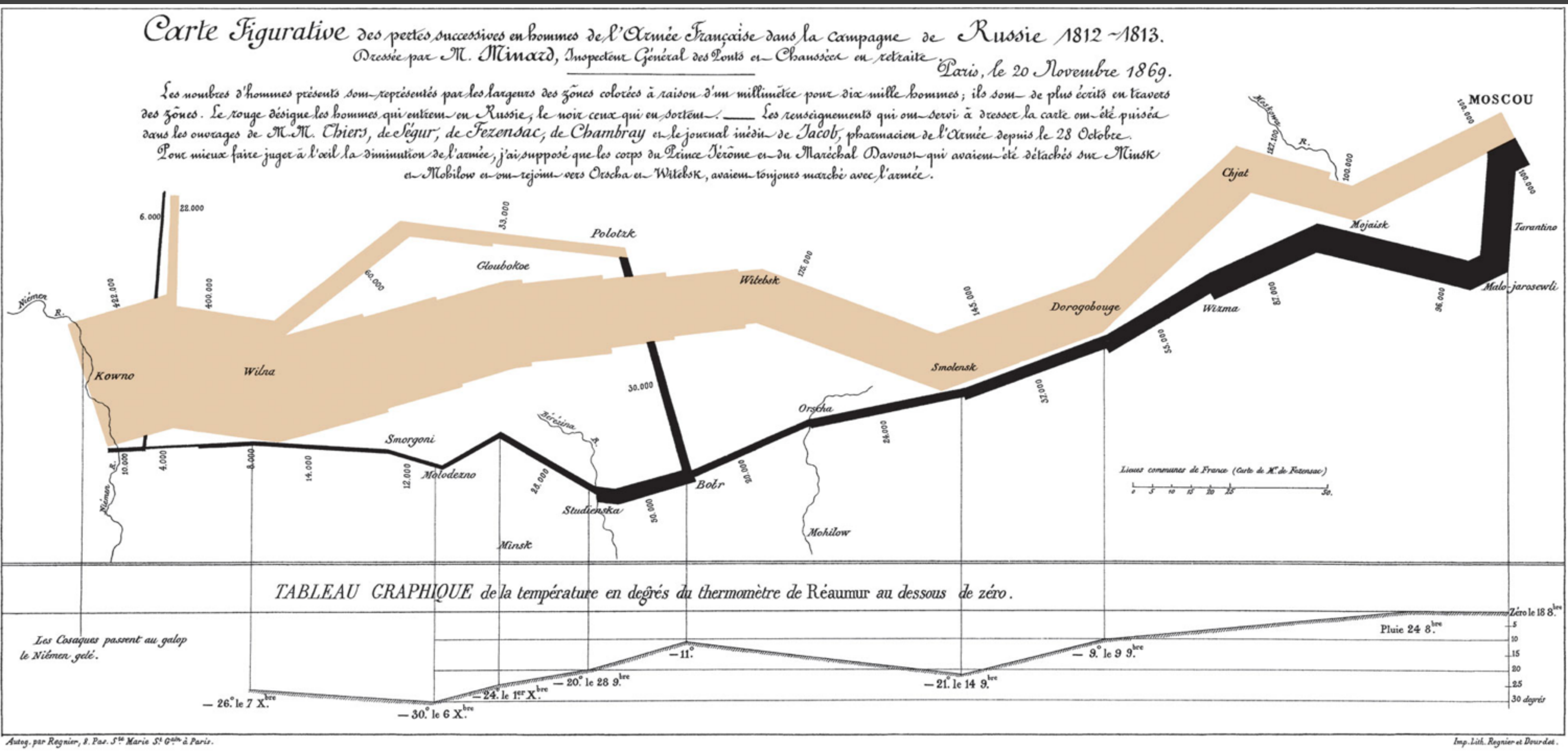


temperature (Q)

latitude (Q) / time (O)



Minard 1869: Napoleon’s March



Depicts at least 5 quantitative variables. Any others?

Formalizing Design

Choosing Visual Encodings

Assume *k* visual encodings and *n* data attributes. We would like to pick the "best" encoding among a combinatorial set of possibilities of size $(n+1)^k$

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 86]

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.

Design Criteria [Mackinlay 86]

Expressiveness

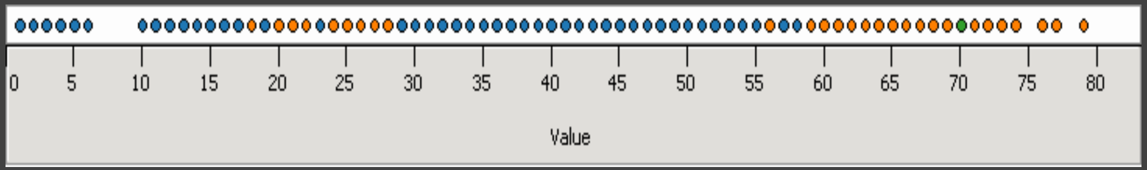
A set of facts is *expressible* in a visual language if the sentences (i.e. the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

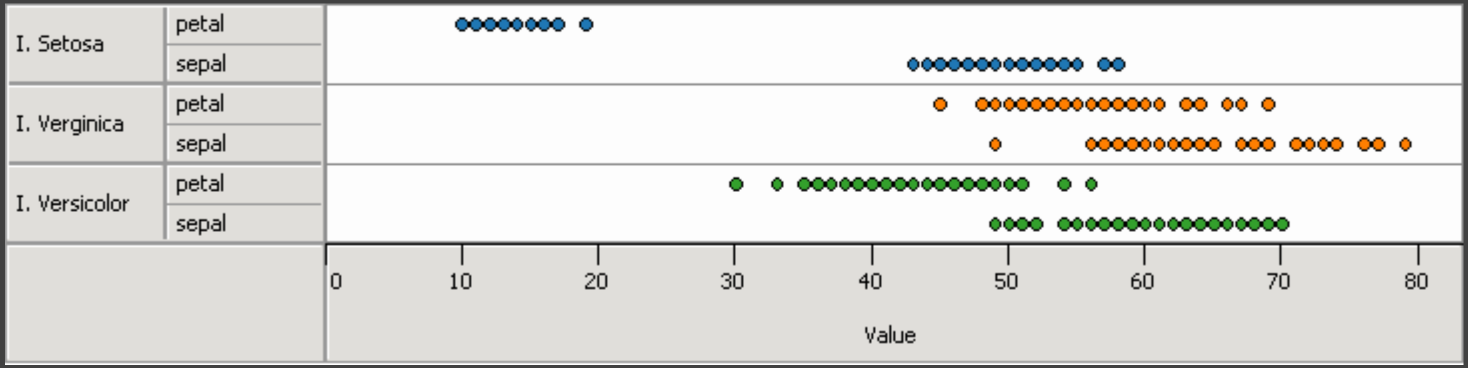
Effectiveness

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.

Can not express the facts

A multivariate relation may be *inexpressive* in a single horizontal dot plot because multiple records are mapped to the same position.





Expresses facts not in the data

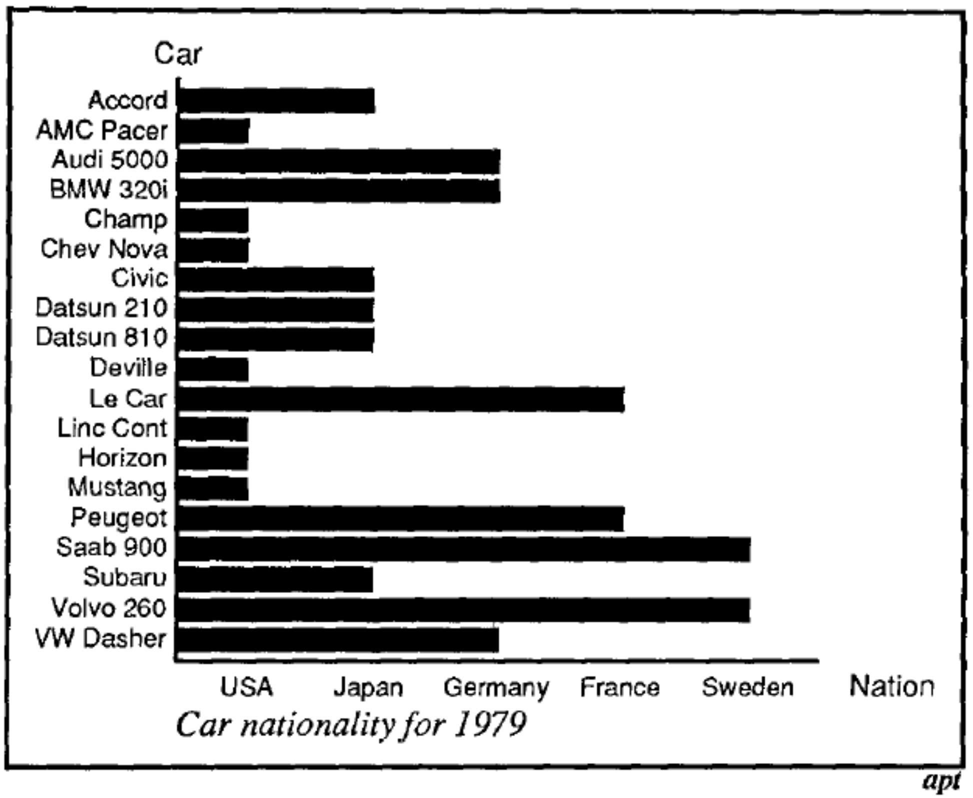


Fig. 11. Incorrect use of a bar chart for the *Nation* relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the *Nation* relation.

A length is interpreted as a quantitative value.

Design Criteria [Mackinlay 86]

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.

Design Criteria [Mackinlay 86]

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.

Design Criteria [Tversky 02]

Congruence
The structure and content of the external representation should correspond to the desired structure and content of the internal representation.

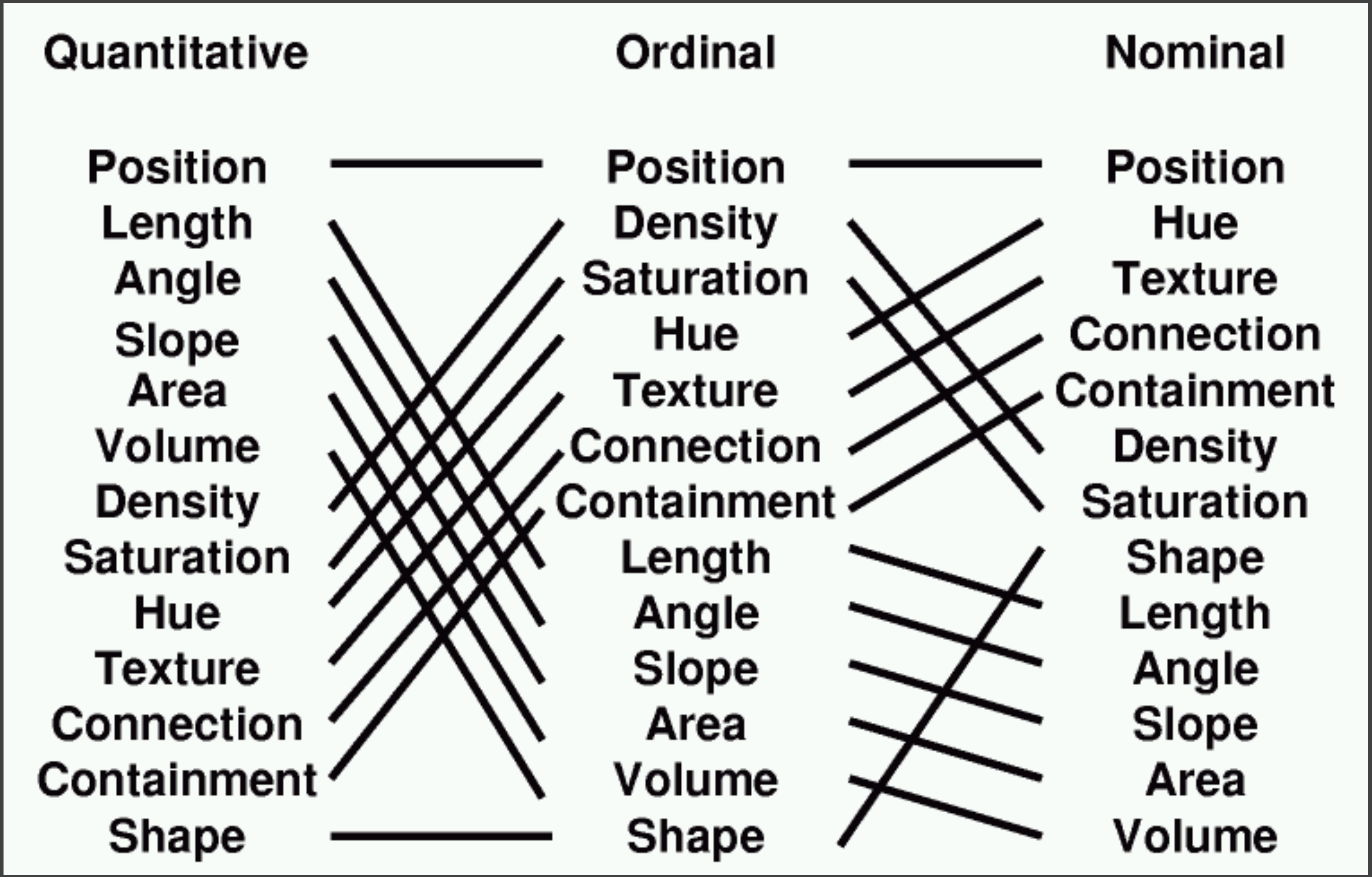
Apprehension
The structure and content of the external representation should be readily and accurately perceived and comprehended.

Design Criteria *Translated*

Tell the truth and nothing but the truth
(don’t lie, and don’t lie by omission)

Use encodings that people decode better
(where better = faster and/or more accurate)

Mackinlay's Ranking



Conjectured *effectiveness* of encodings by data type

Mackinlay’s Design Algorithm

APT - “A Presentation Tool”, 1986

User formally specifies data model and type
Input: ordered list of data variables to show

APT searches over design space
Test expressiveness of each visual encoding
Generate encodings that pass test
Rank by perceptual effectiveness criteria

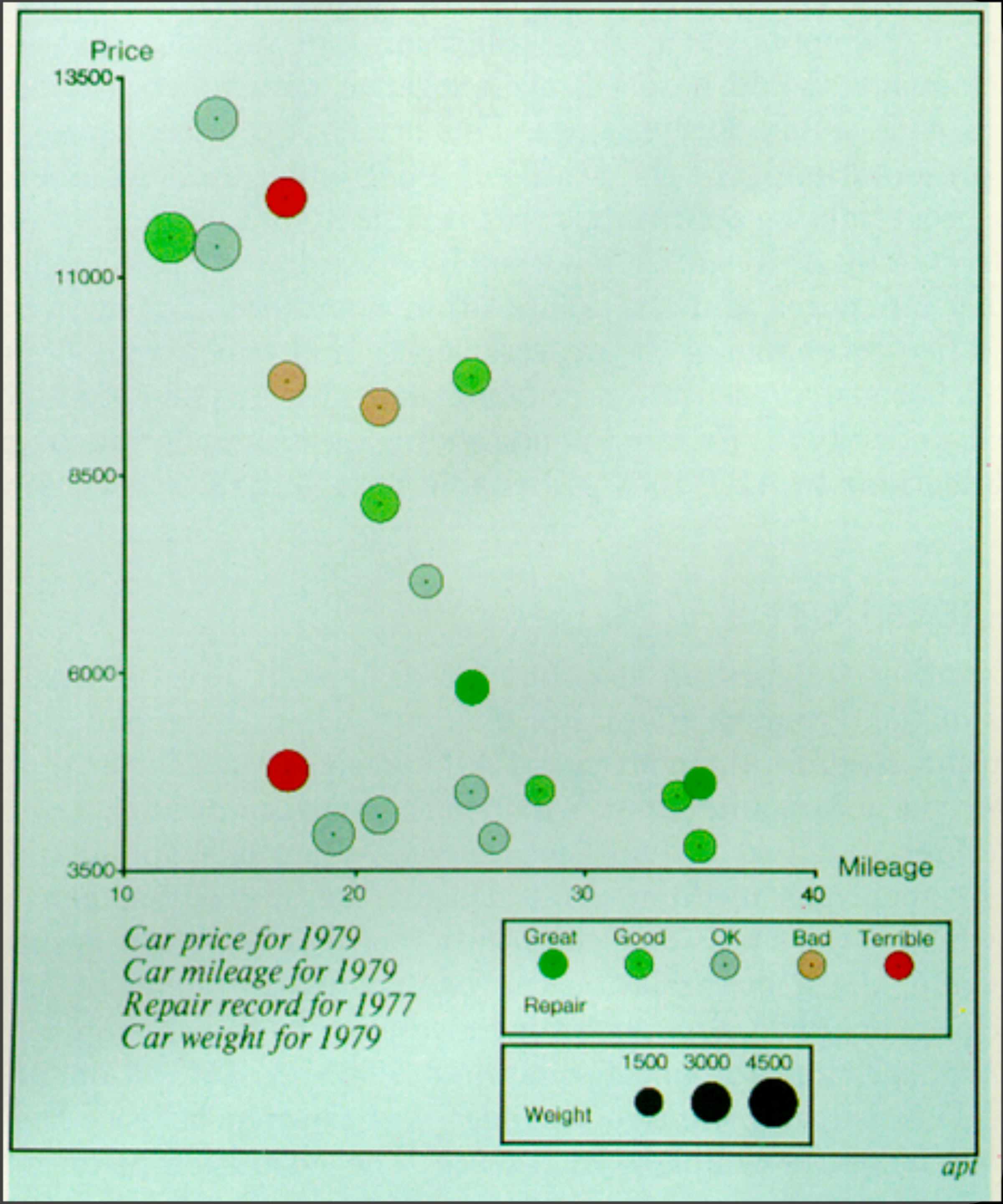
Output the “most effective” visualization

APT

Automatically generate chart for car data

Input variables:

1. Price
2. Mileage
3. Repair
4. Weight



Limitations of APT?

Limitations of APT

Does not cover many visualization techniques
Networks, hierarchies, maps, diagrams
Also: 3D structure, animation, illustration, …

Does not consider interaction

Does not consider semantics / conventions

Assumes single visualization as output

Summary: Data & Image Models

Formal specification
Data model: relational data; N,O,Q types
Image model: visual encoding channels
Encodings map data to visual variables

Choose expressive and effective encodings
Rule-based tests of expressiveness
Perceptual effectiveness rankings

Question: how do we establish effectiveness criteria? *Subject of perception lectures…*

A1: Visualization Design

Design a **static visualization** for the data set.

You are free to **use any tools** (inc. pen & paper).

Deliverables (upload via Canvas; see A1 page)

Image of your visualization (PNG or JPG format)

Short description + design rationale (≤ 4 paragraphs)

Due by **5:00 pm, Monday April 6**.