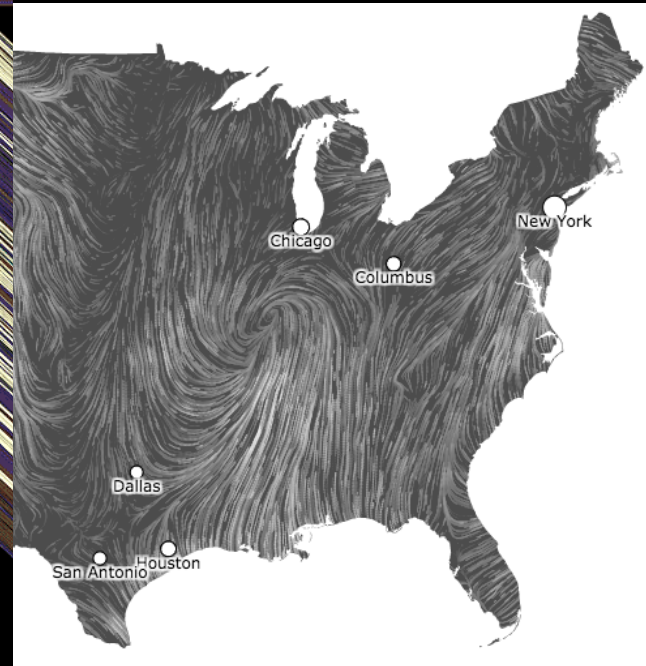
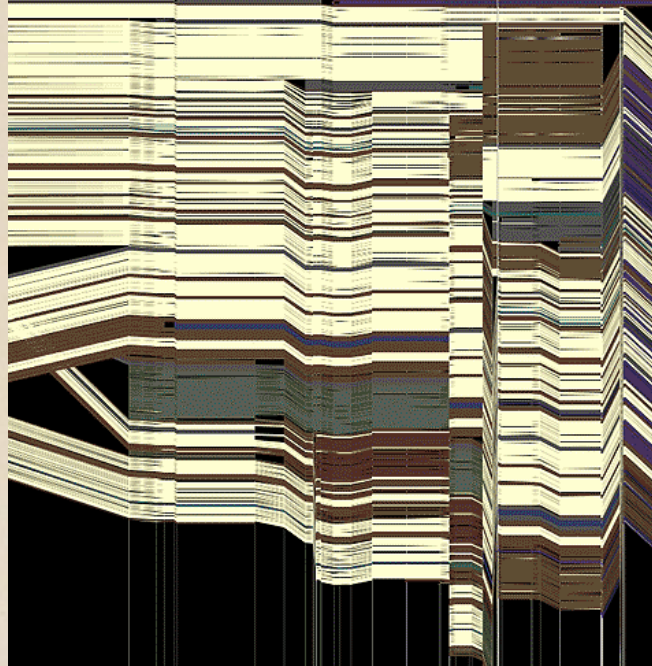
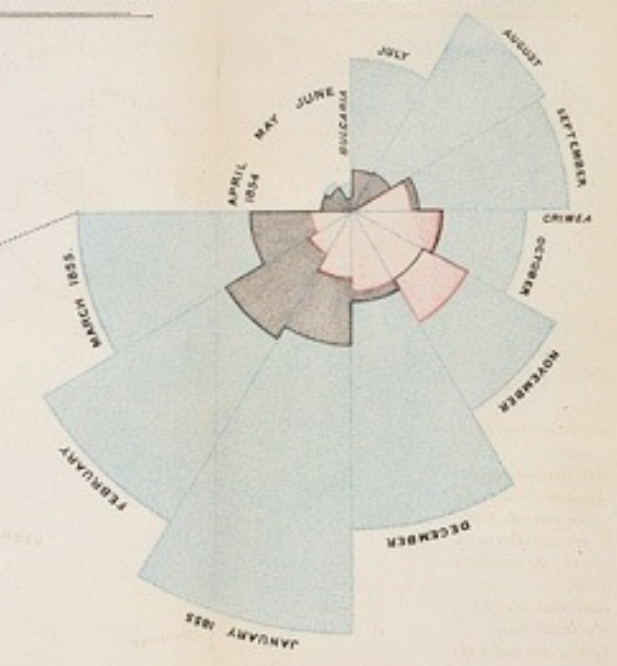


CSE 442 - Data Visualization

# Data and Image Models



Matthew Conlen University of Washington

# 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 Models

# Data Models / Conceptual Models

**Data models** are formal 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)

# Dimensions & Measures

**Dimensions** (~ independent variables)

Often 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...

Not a strict distinction. The same variable may be treated either way depending on the task.

# 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 Tables & Transformations

# Relational Data Model

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

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

Each record is a fixed-length tuple

Each **column** (or *field*) represents a variable

Each field 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] / SQL

Operations on Data Tables: table(s) in, table out

# Relational Algebra [Codd '70] / SQL

**Operations on Data Tables: table(s) in, table out**

Projection (`select`): select a set of columns

Selection (`where`): filter rows

Sorting (`order by`): order records

Aggregation (`group by, sum, min, max, ...`):

partition rows into groups + summarize

Combination (`join, union, ...`):

integrate data from multiple tables

# Relational Algebra [Codd '70] / SQL

**Projection** (select): select a set of columns  
select day, stock

day	stock	price
10/3	AMZN	957.10
10/3	MSFT	74.26
10/4	AMZN	965.45
10/4	MSFT	74.69



day	stock
10/3	AMZN
10/3	MSFT
10/4	AMZN
10/4	MSFT



# Relational Algebra [Codd '70] / SQL

Sorting (order by): order records

```
select * order by stock
```

day	stock	price
10/3	AMZN	957.10
10/3	MSFT	74.26
10/4	AMZN	965.45
10/4	MSFT	74.69



day	stock	price
10/3	AMZN	957.10
10/4	AMZN	965.45
10/3	MSFT	74.26
10/4	MSFT	74.69

# Relational Algebra [Codd '70] / SQL

**Selection (where):** filter rows

```
select * where price > 100
```

day	stock	price
10/3	AMZN	957.10
10/3	MSFT	74.26
10/4	AMZN	965.45
10/4	MSFT	74.69



day	stock	price
10/3	AMZN	957.10
10/4	AMZN	965.45

# Relational Algebra [Codd '70] / SQL

Aggregation (group by, sum, min, max, ...):

```
select stock, min(price) group by stock
```

day	stock	price
10/3	AMZN	957.10
10/3	MSFT	74.26
10/4	AMZN	965.45
10/4	MSFT	74.69



stock	min(price)
AMZN	957.10
MSFT	74.26

# Relational Algebra [Codd '70] / SQL

Join (join) multiple tables together

day	stock	price
10/3	AMZN	957.10
10/3	MSFT	74.26
10/4	AMZN	965.45
10/4	MSFT	74.69



day	stock	price	min
10/3	AMZN	957.10	957.10
10/3	MSFT	74.26	74.26
10/4	AMZN	965.45	957.10
10/4	MSFT	74.69	74.26

stock	min
AMZN	957.10
MSFT	74.26

```
select t.day, t.stock, t.price, a.min
from table as t, aggregate as a
where t.stock = a.stock
```

# Roll-Up and Drill-Down

Want to examine population by year and age?

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

Dimensions                      Measure

SELECT year, age, sum(people)

FROM census

GROUP BY year, age

Dimensions

The diagram consists of two horizontal curly braces. The first brace is positioned above the words 'Dimensions' and 'Measure' and spans the width of the 'SELECT' clause, grouping 'year, age, sum(people)'. The second brace is positioned below the words 'Dimensions' and 'Measure' and spans the width of the 'GROUP BY' clause, grouping 'year, age'.

# Roll-Up and Drill-Down

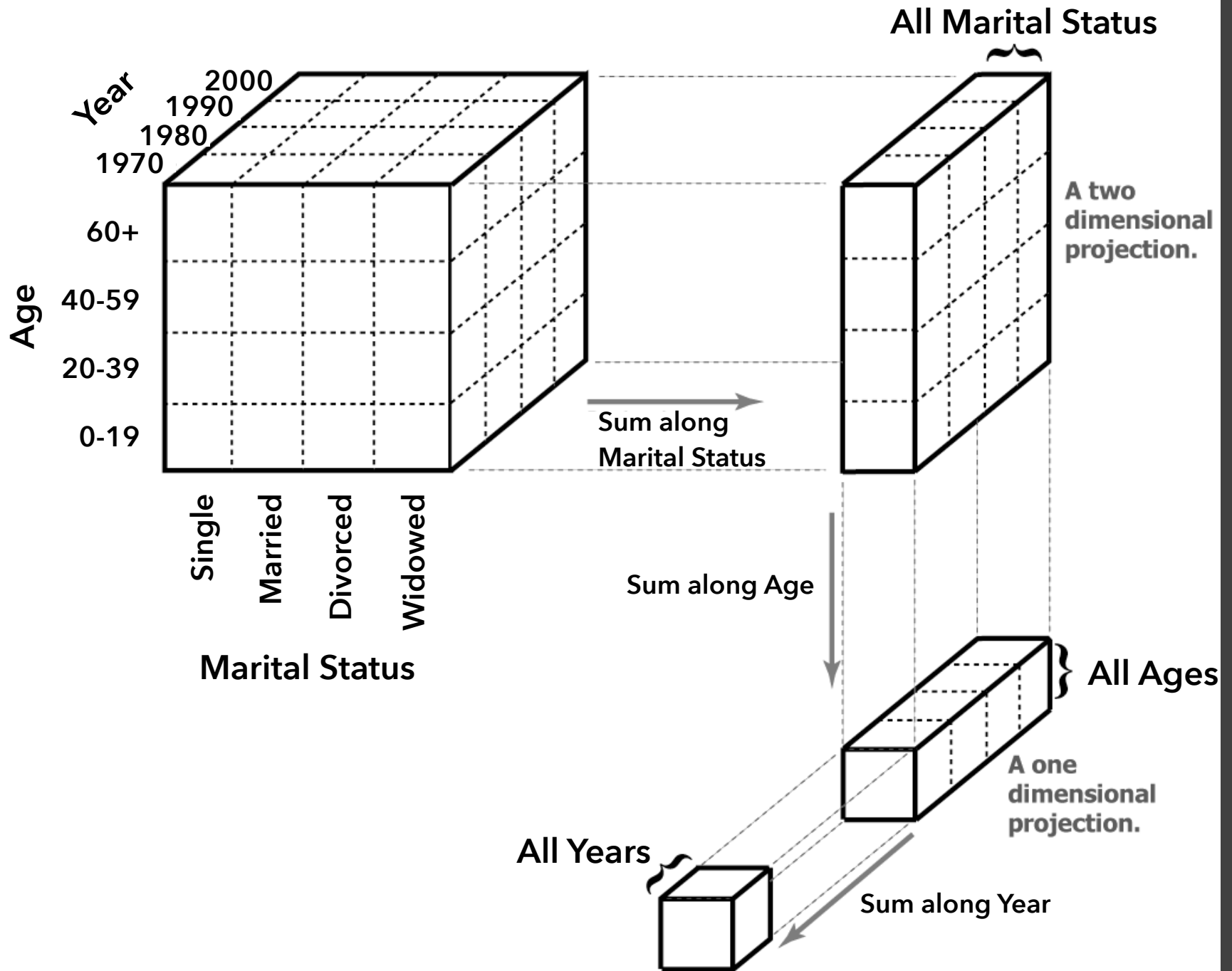
Want to see the breakdown by marital status?

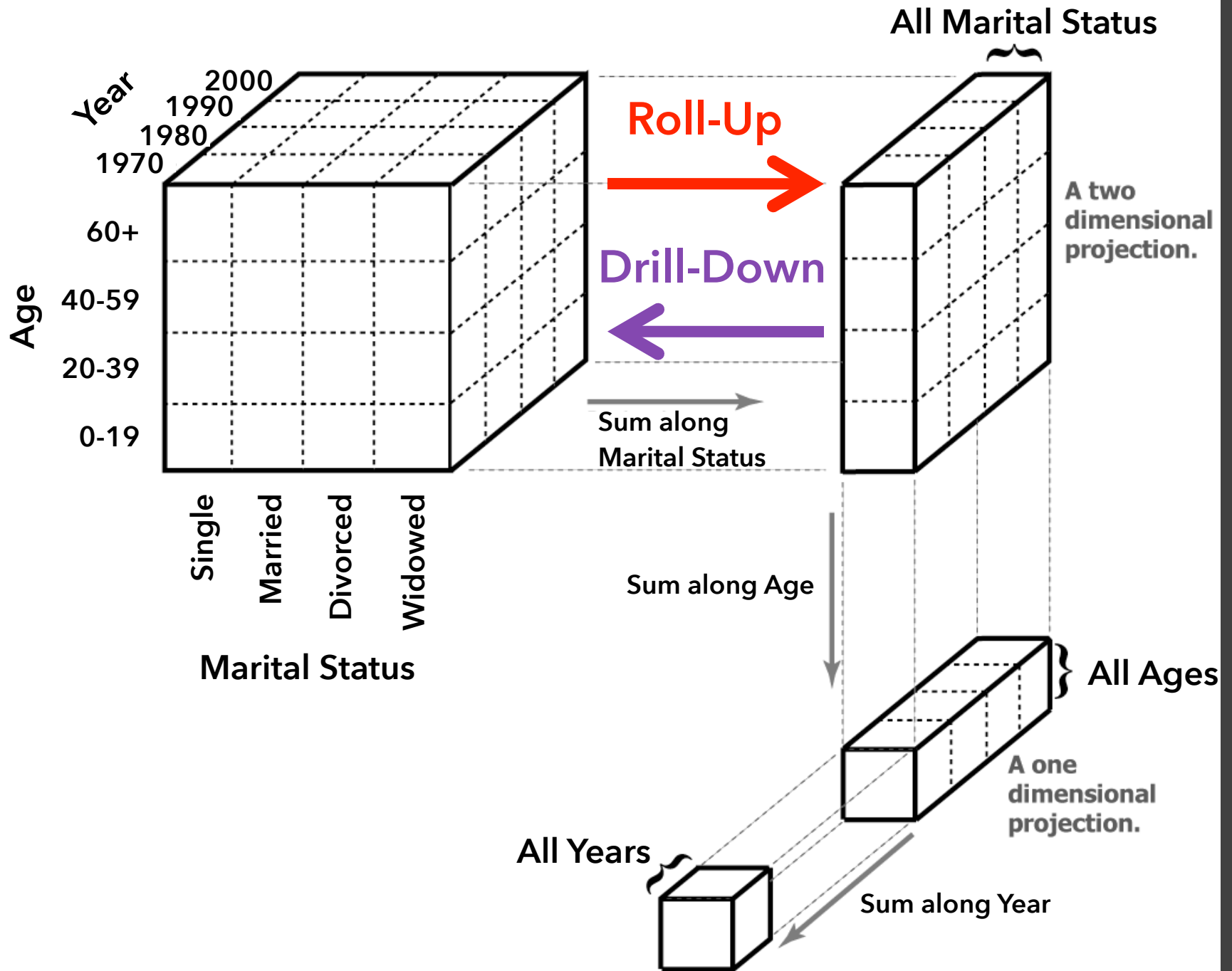
**Drill-down** into additional dimensions

```
SELECT year, age, marst, sum(people)
```

```
FROM census
```

```
GROUP BY year, age, marst
```







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?

# Common Data Formats

## CSV: Comma-Separated Values (d3.csv)

```
year,age,marst,sex,people
```

```
1850,0,0,1,1483789
```

```
1850,5,0,1,1411067
```

```
...
```

# Common Data Formats

## CSV: Comma-Separated Values (d3.csv)

```
year,age,marst,sex,people  
1850,0,0,1,1483789  
1850,5,0,1,1411067  
...
```

## JSON: JavaScript Object Notation (d3.json)

```
[  
  {"year":1850,"age":0,"marst":0,"sex":1,"people":1483789},  
  {"year":1850,"age":5,"marst":0,"sex":1,"people":1411067},  
  ...  
]
```

# Administrivia

# Administrative

Office hours now available on course website

Mon 2:30, Tues 1:30, Wed 10:30, Fri 1:00

Course staff email list ([cse442@cs](mailto:cse442@cs))

Canvas discussion and A1 submission is live

Lecture recordings starting next week

# A1: Visualization Design

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

Every 10 years, the census documents the demographic make-up of the U.S., influencing congressional districting and social services. This dataset contains a summary of census data for two years a century apart: 1900 and 2000.

You must choose the message you want to convey. What question(s) do you want to answer? What insight do you want to communicate?

# A1: Visualization Design

Pick a **guiding question**, use it to title your vis.  
Design a **static visualization** for that question.  
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 ( $\leq 4$  paragraphs)

Due by **11:59 pm, Jan 14.**

# Next Thursday: 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!*



# Online Discussion

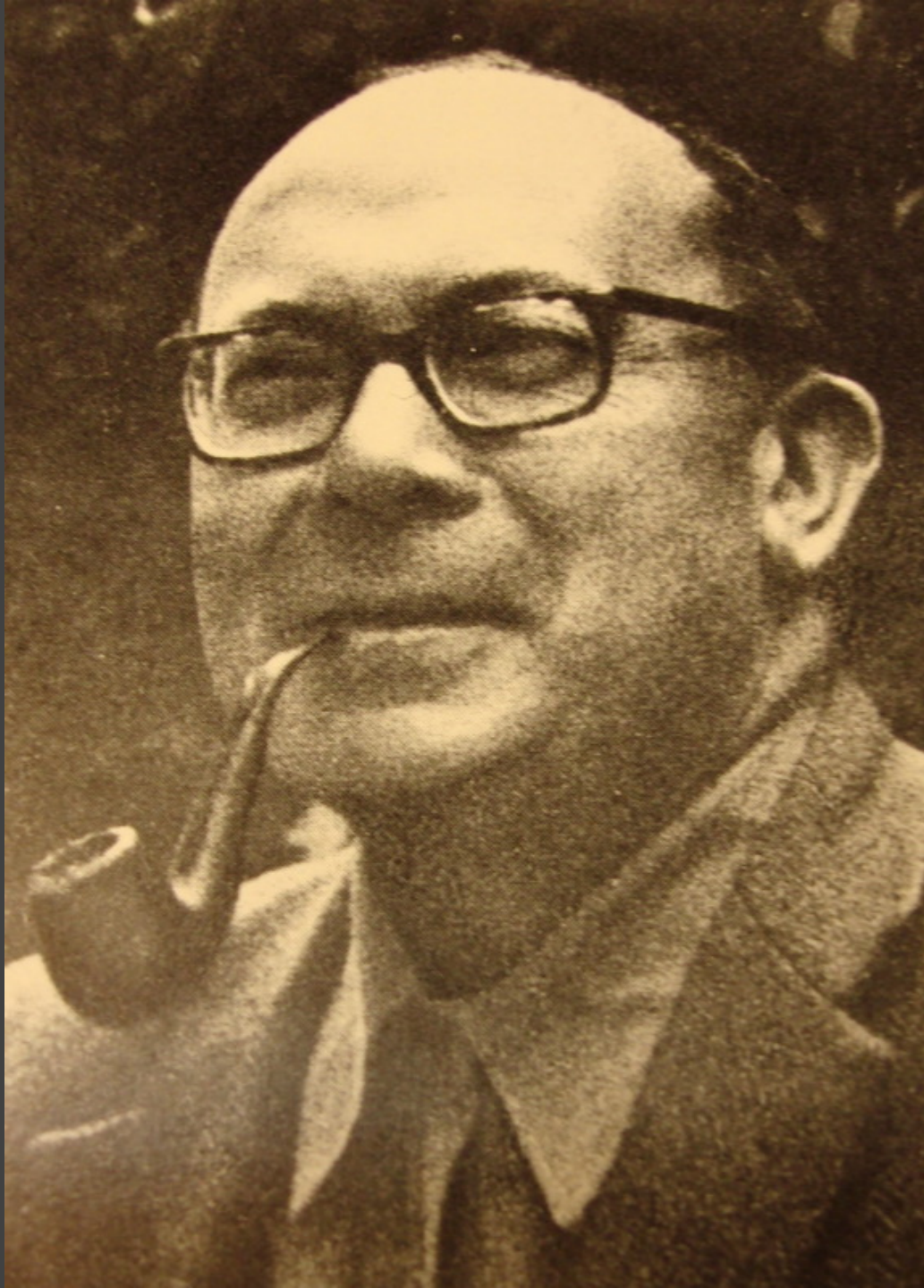
Post discussion comments on class Canvas forum.

One comment per week (ending week 8).

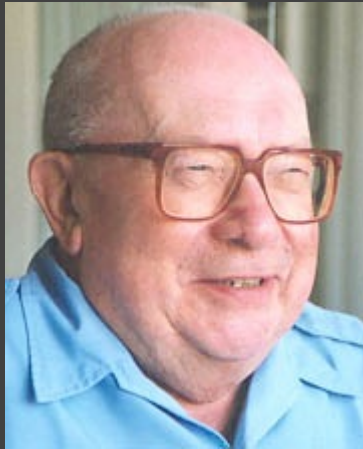
Comments must be posted by Friday 11:59pm.

You have 1 "pass" for the quarter.

# Image Models



# Visual Language is a Sign System



Jacques Bertin

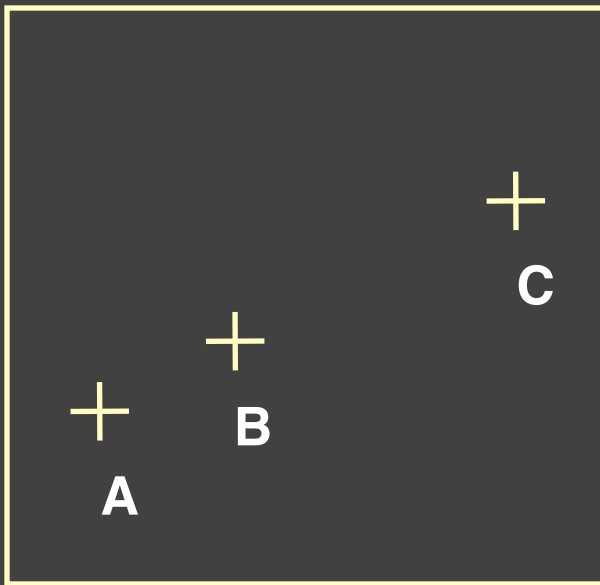
Images perceived as a set of signs

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 proportional 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

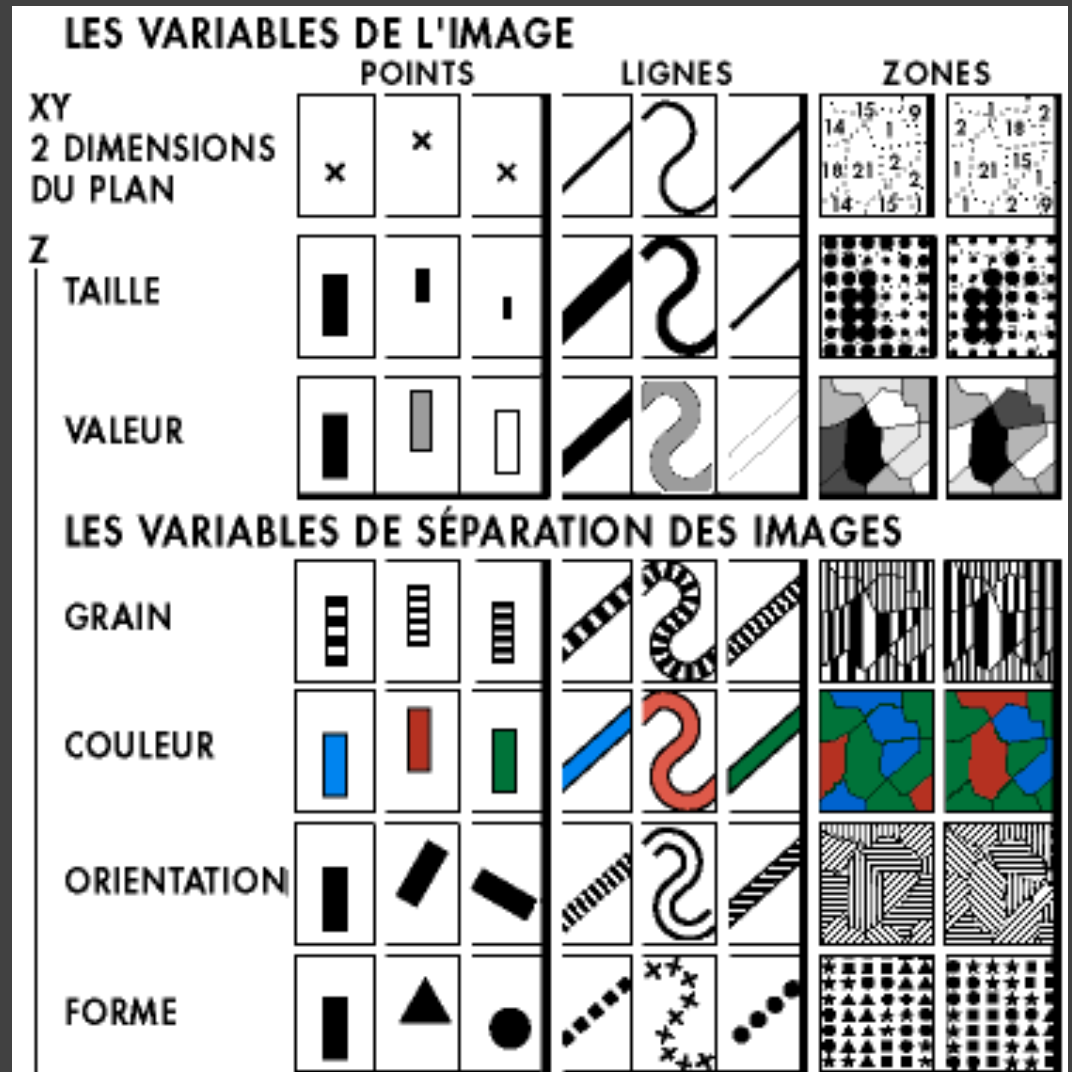
Value

Texture

Color

Orientation

Shape



# Visual Encoding Variables

Position

**Length**

**Area**

**Volume**

Value

Texture

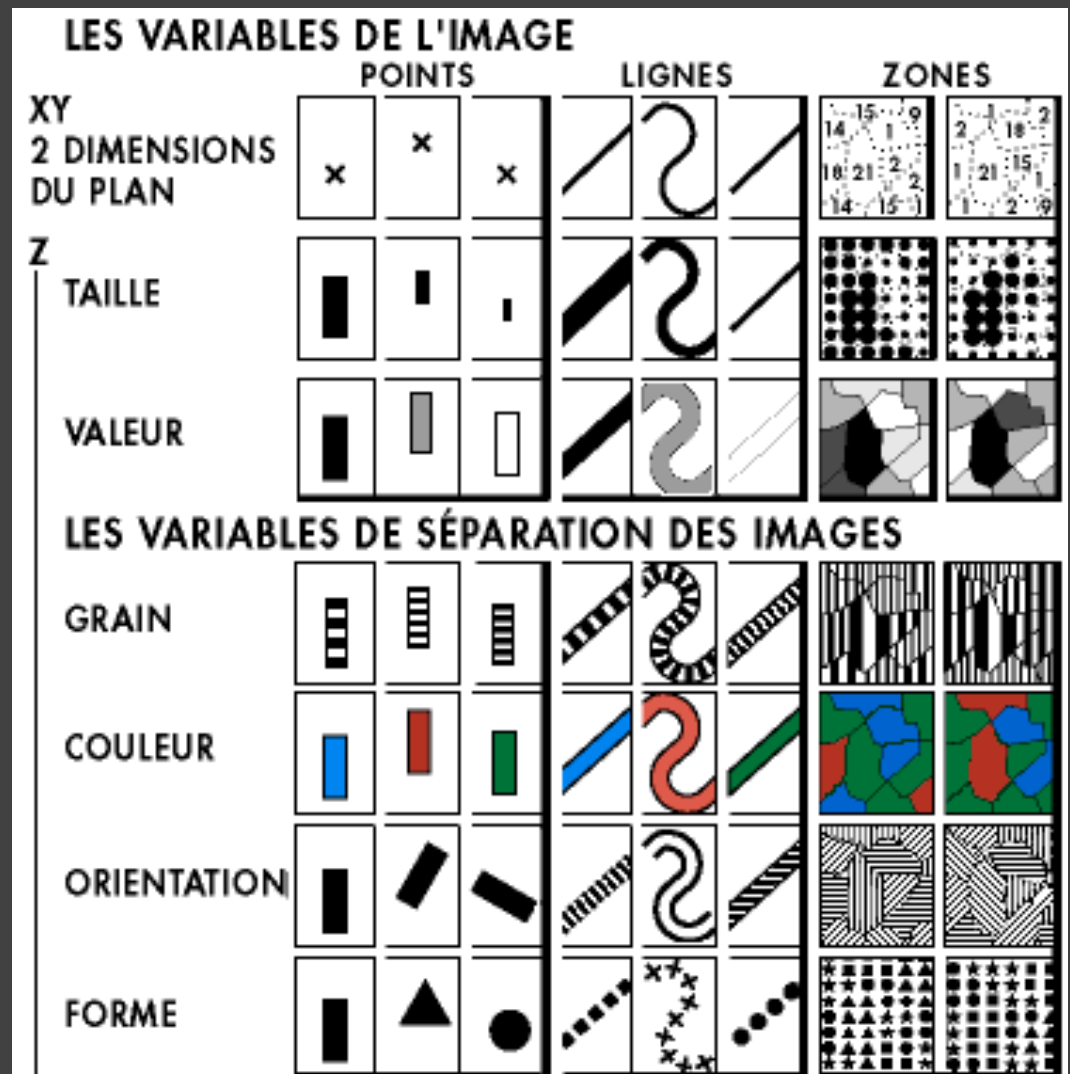
Color

Orientation

Shape

**Transparency**

**Blur / Focus ...**





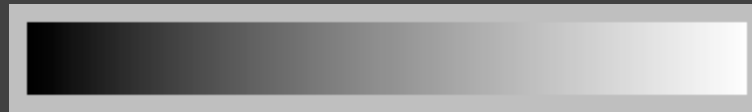
# Information in Hue and Value

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
---	---	---

Nominal

Size

N	O	Q
---	---	---

Ordinal

Value

N	O	Q
---	---	---

Quantitative

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

Texture

N	o	
---	---	--

Color

N		
---	--	--

Orientation

N		
---	--	--

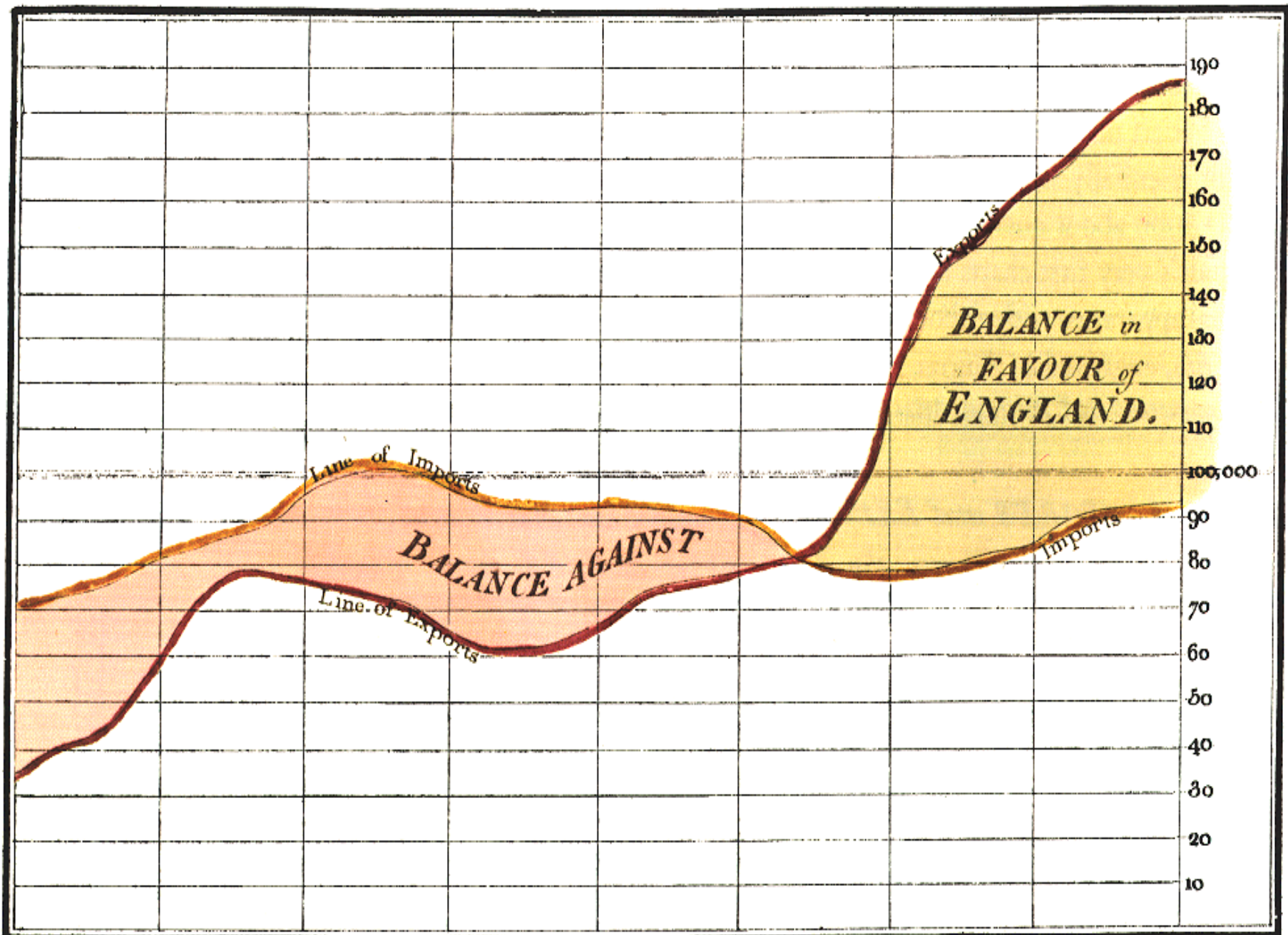
Shape

N		
---	--	--

# Altair Demo

# Deconstructions

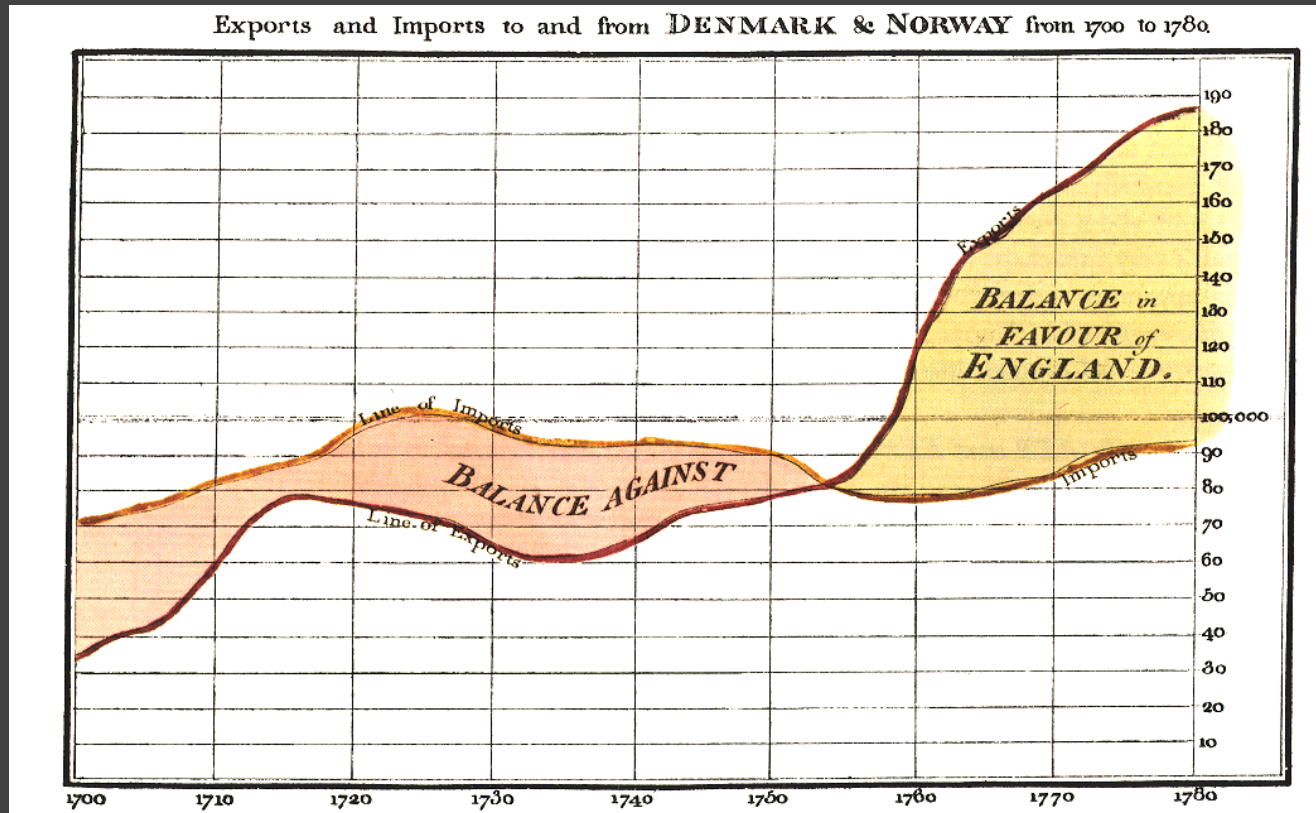
Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.



1700 1710 1720 1730 1740 1750 1760 1770 1780

190  
180  
170  
160  
150  
140  
130  
120  
110  
100,000  
90  
80  
70  
60  
50  
40  
30  
20  
10

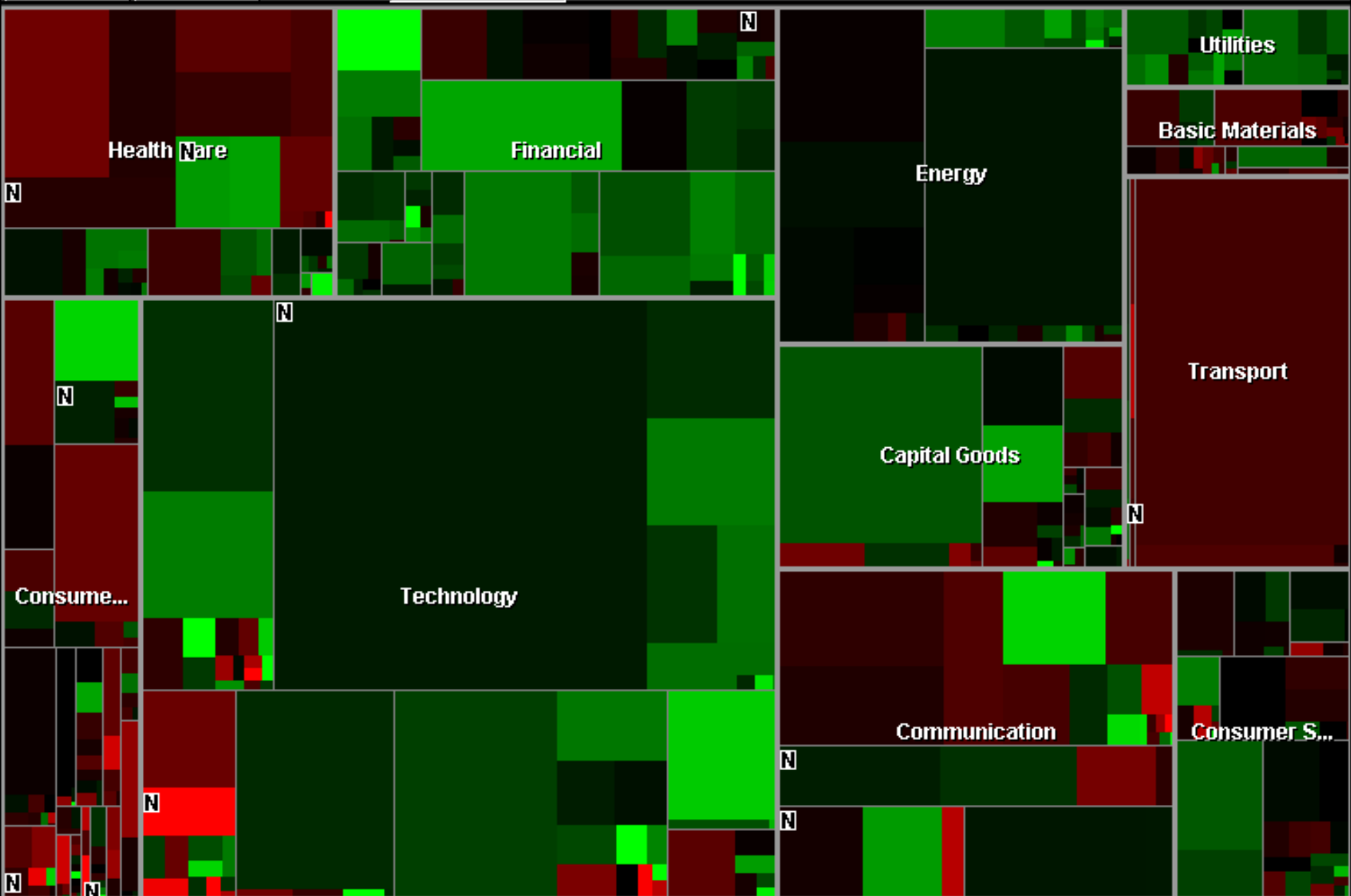
# William Playfair, 1786



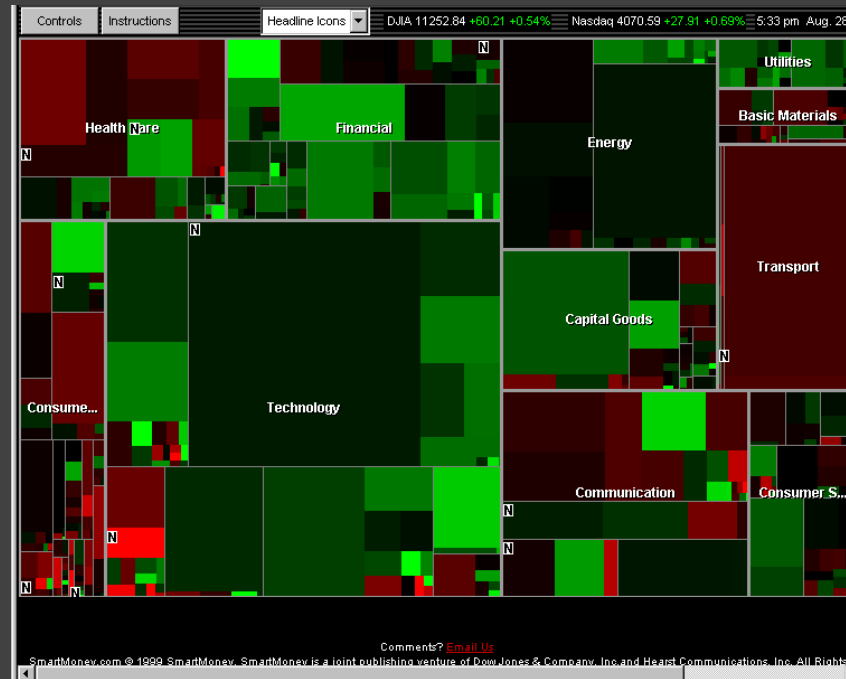
X-axis: year (Q)

Y-axis: currency (Q)

Color: imports/exports (N, O)



# 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

## Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M. M. Chiers, de Légar, de Fezensac, de Chambrey et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davout qui avaient été détachés sur Minsk et Mohilow et qui rejoignent vers Orscha et Witebsk, avaient toujours marché avec l'armée.

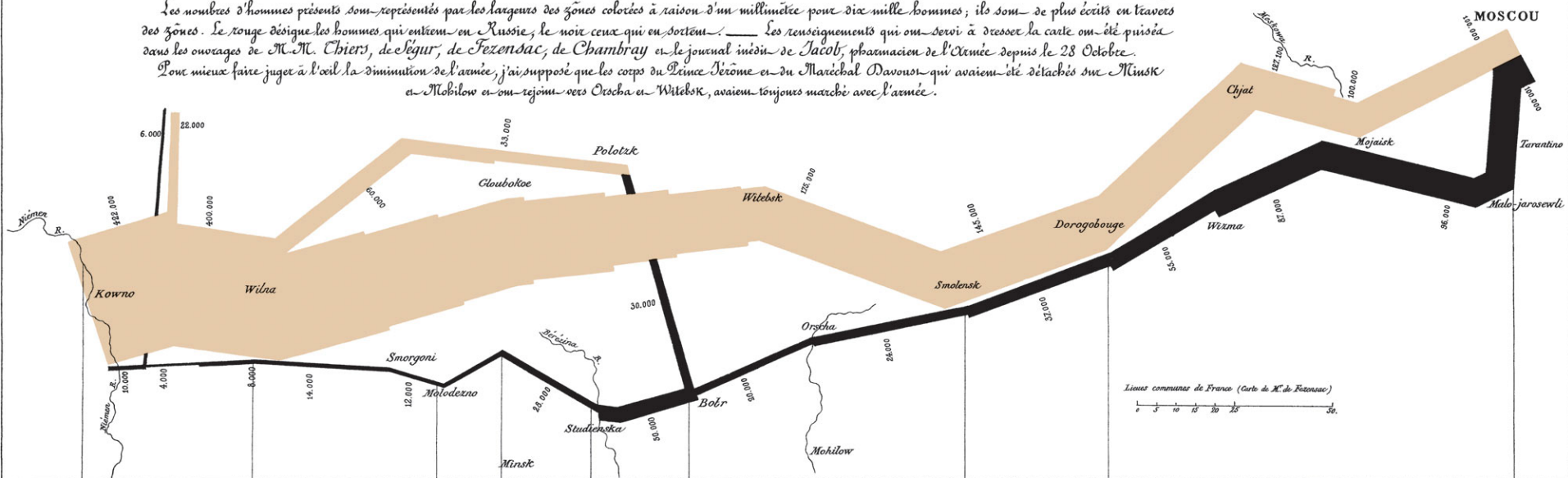
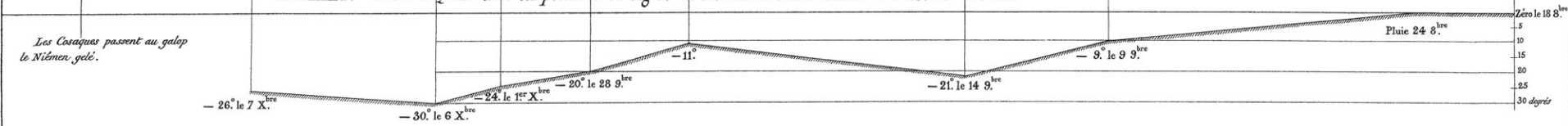


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.



Les Cosaques passent au galop le Niémen, gelé.



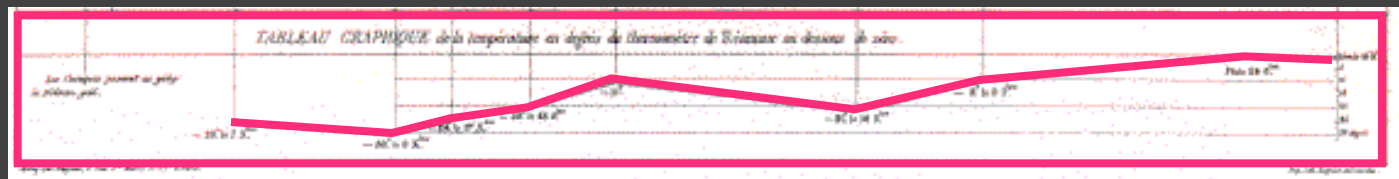
# Mark Composition

Y-axis: temperature (Q)

+

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

=



Temp over space/time (Q x Q)

# Mark Composition

Y-axis: longitude (Q)

+

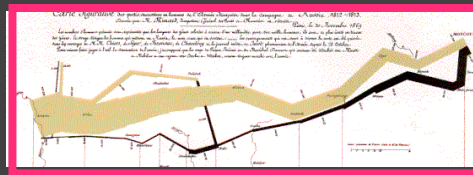
X-axis: latitude (Q)

+

Width: army size (Q)



=

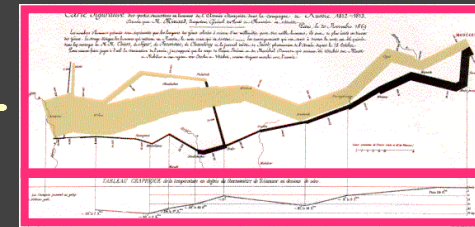
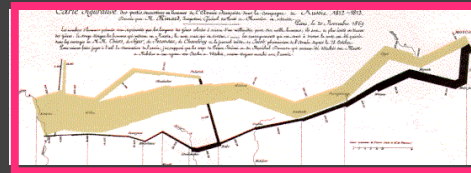


Army position (Q x Q) and army size (Q)

longitude (Q)

latitude (Q)

army size (Q)



temperature (Q)

latitude (Q) / time (O)



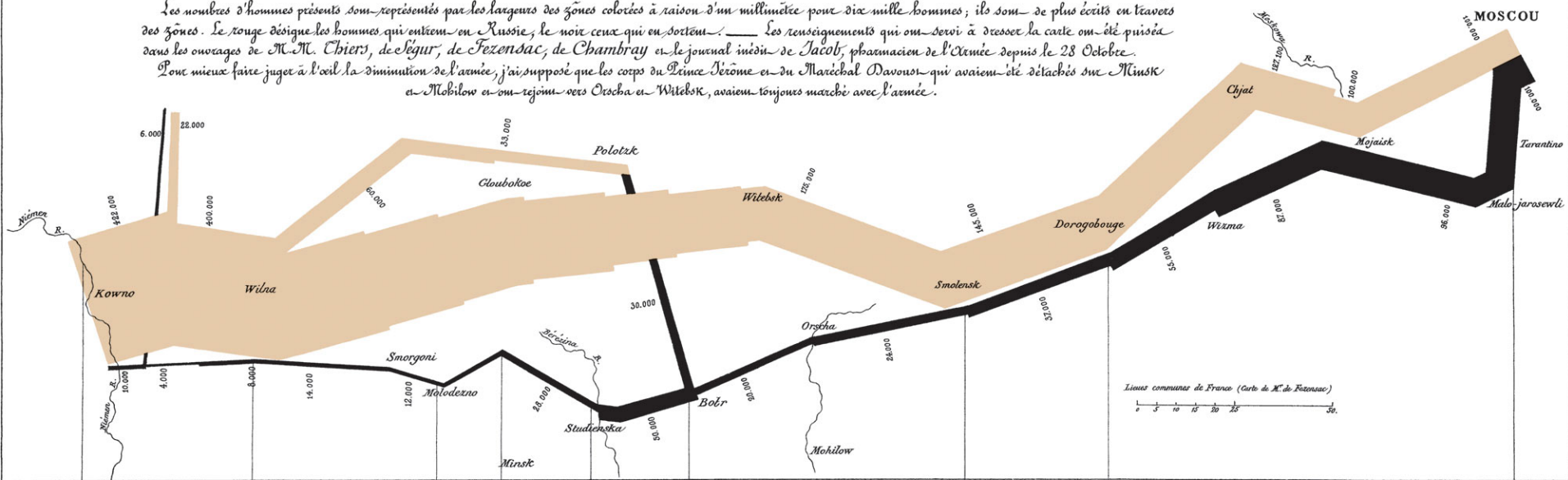
# Minard 1869: Napoleon's March

## Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

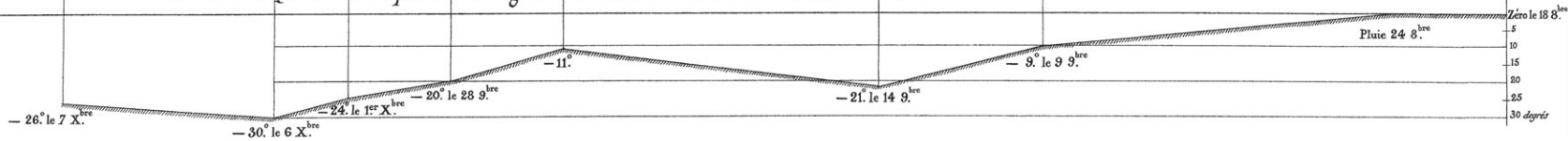
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M. M. Chiers, de Légar, de Fezensac, de Chambrey et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davout qui avaient été détachés sur Minsk et Mohilow et qui rejoignent vers Orscha et Witebsk, avaient toujours marché avec l'armée.



## TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.



Les Cosaques passent au galop le Niémen gelé.

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

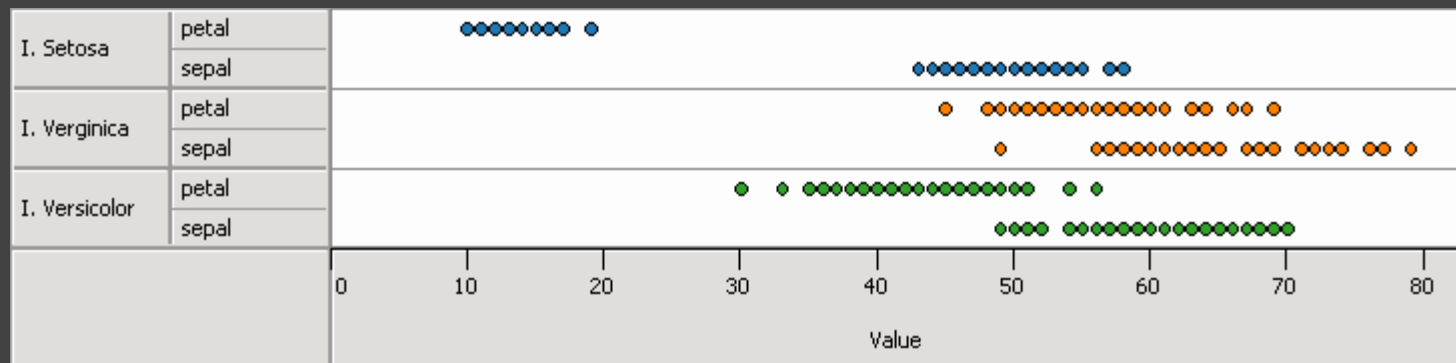
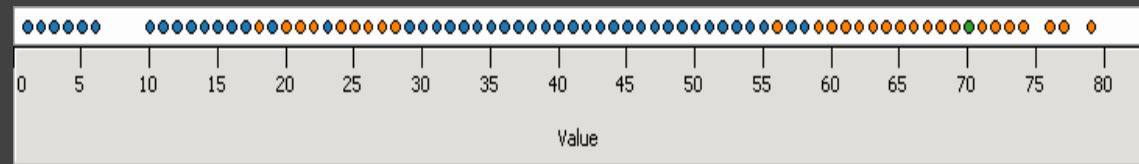
A set of facts is *expressible* in a visual language if the sentences (i.e. the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

## Effectiveness

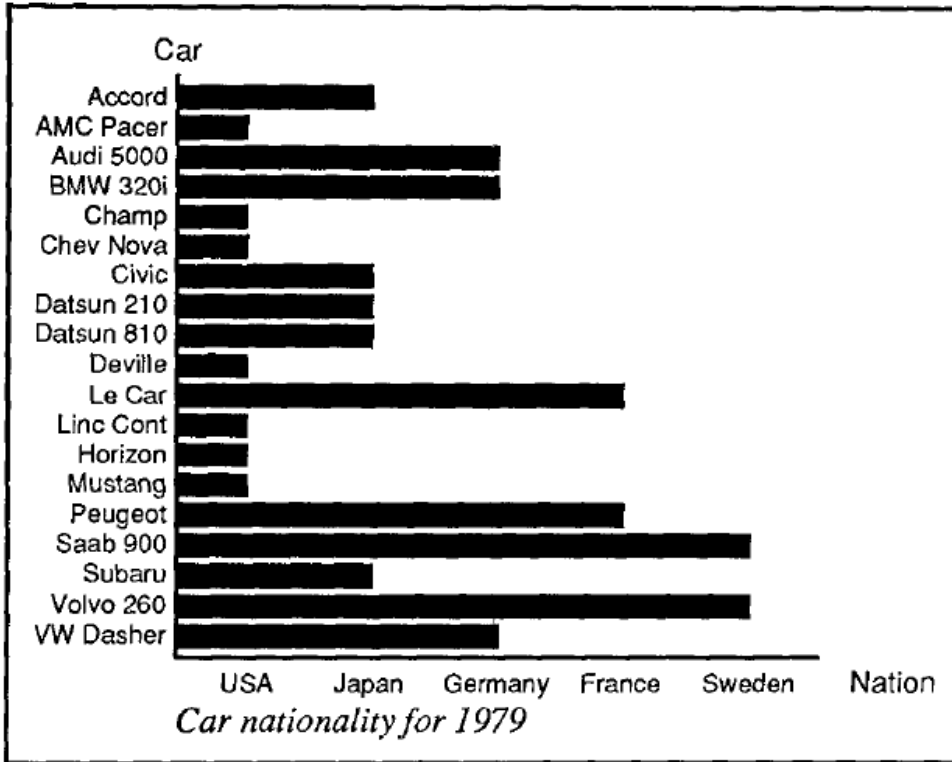
Visualization A is more effective than another visualization if the information conveyed by one is perceived more effectively than the other in the other visualization.

# 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



apt

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

Visualization A is more effective than another visualization if the information conveyed by one is perceived more effectively than the other 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

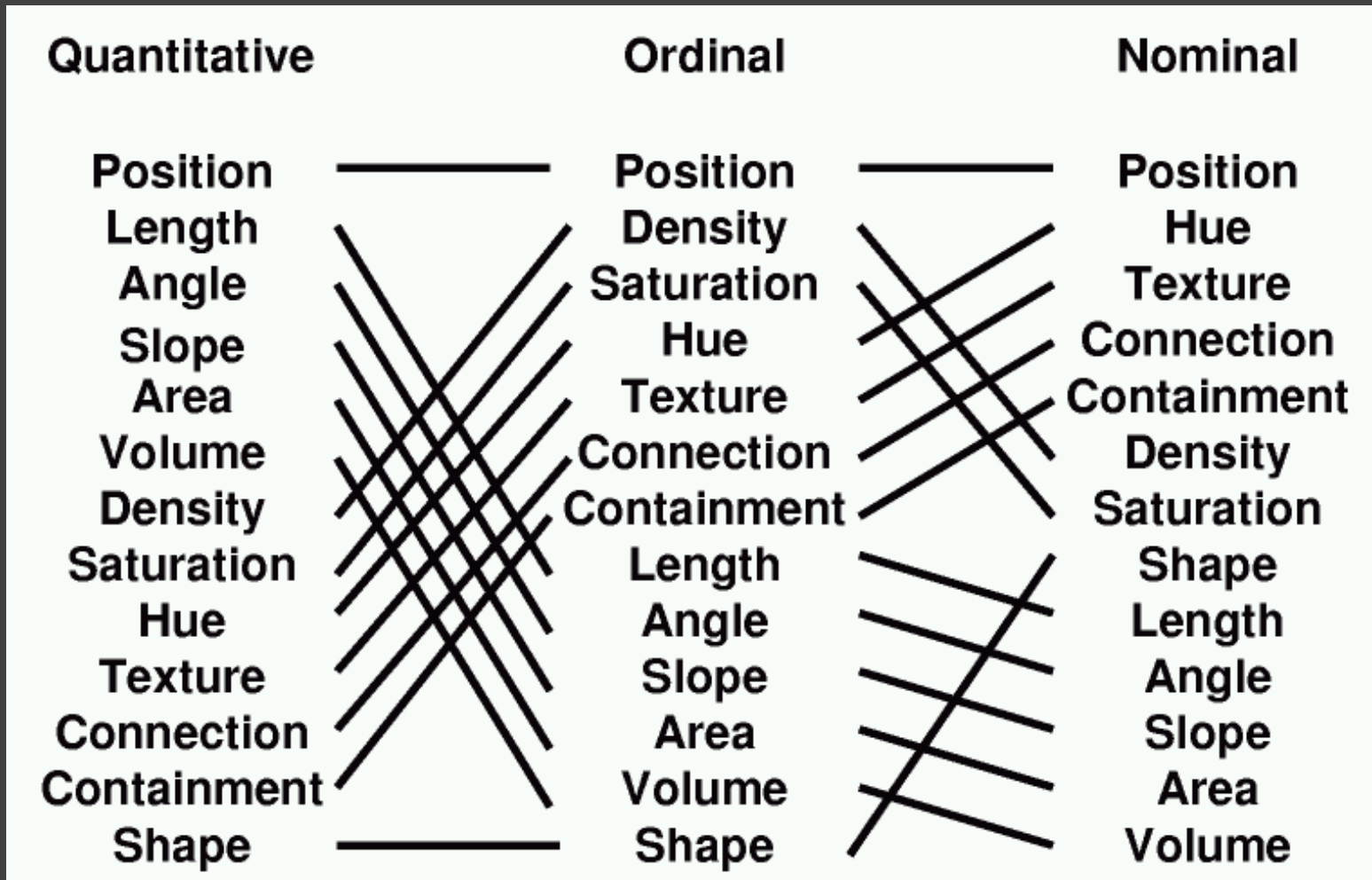
A visualization is more *effective* than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.

# Design Criteria *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

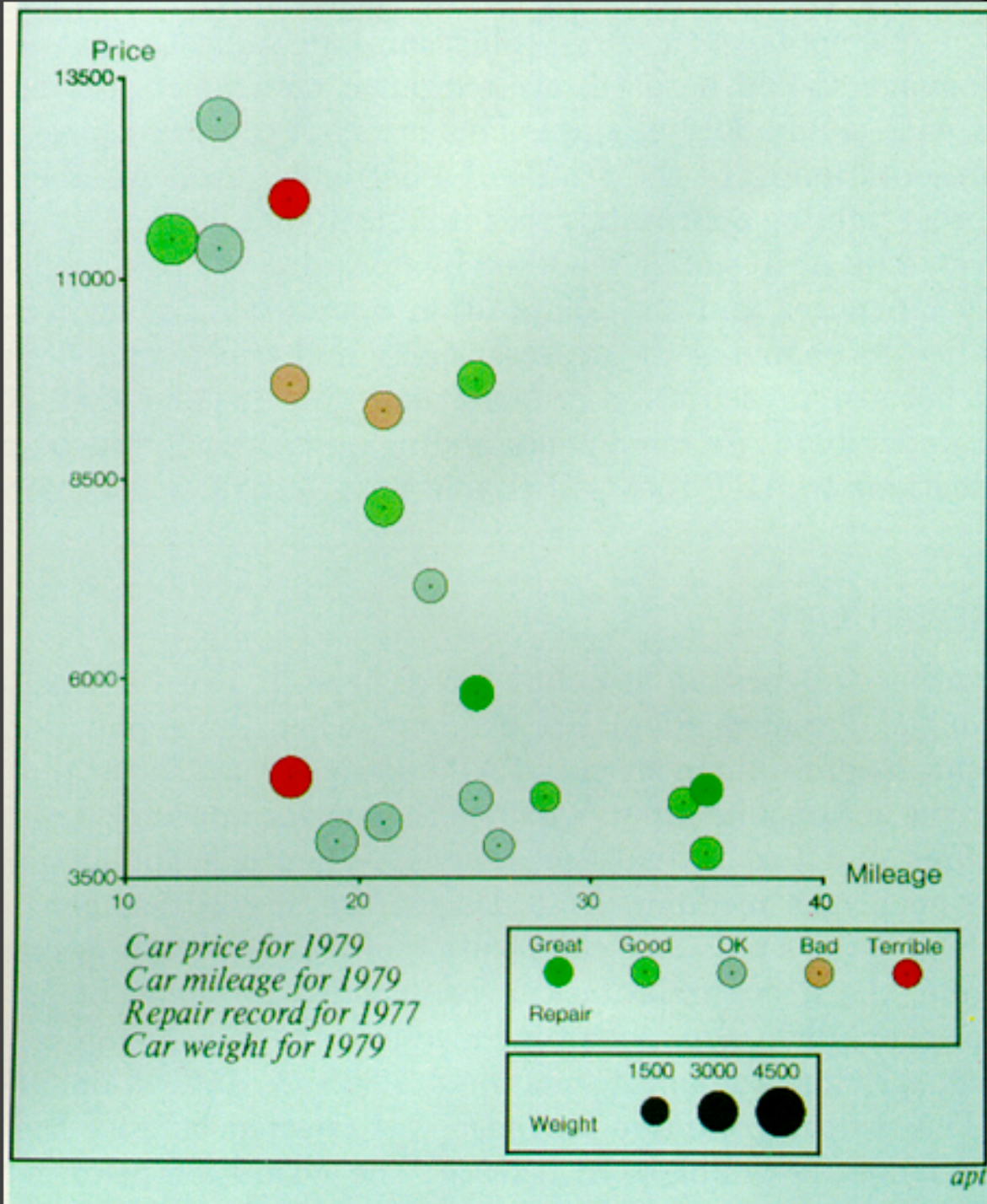
**Output the "most effective" visualization**

# 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...*