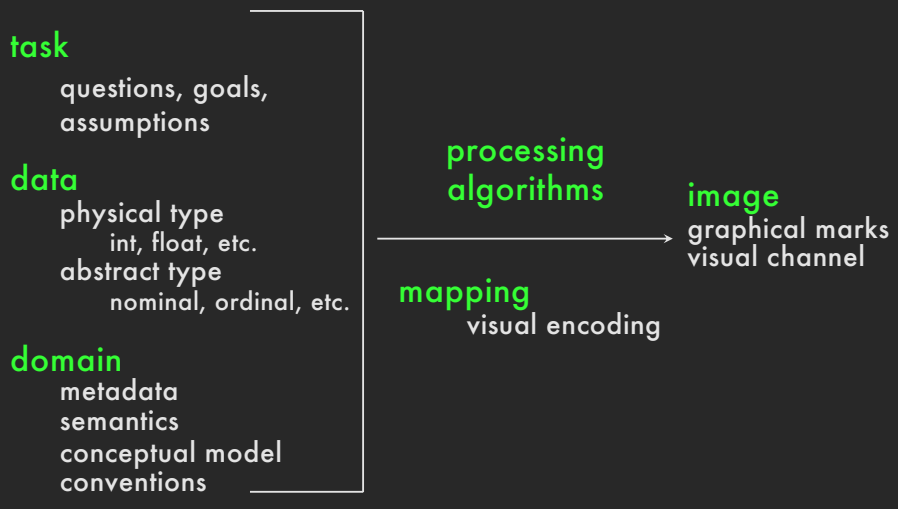
Data and Image Models

Maneesh Agrawala

CS 448B: Visualization
Winter 2020

The big picture



Topics

Properties of data
Properties of the image
Mapping data to images

Data

Data models vs. Conceptual models

Data models are formal descriptions

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

Conceptual models are mental constructions

- Include semantics and support reasoning

Examples (data vs. conceptual)

- 1D floats vs. temperature
- 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 [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, 2016; 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 (°C)

Data type

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

Dimensions and measures

Dimensions: (~ independent variables)

Often discrete variables describing data (N, O)

Categories, dates, binned values

Measures: (~ dependent variables)

Data values that can be aggregated (Q)

Numbers to be analyzed

Aggregate as sum, count, average, std. deviation

Distinction is not strict. The same variable may be treated either way depending on the task.

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, …

Census: N, O, Q?

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

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

Census: N, O, Q?

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

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

Relational data model

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

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

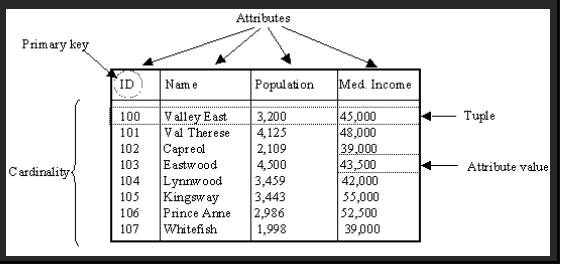
Each record is a fixed-length tuple

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 1970] / 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 rows
- Aggregation (GROUP BY, SUM, MIN, ...)
  partition rows into groups and summarize
- Combination (JOIN, UNION, ...)
  integrate data from multiple tables

Relational algebra [Codd 1970] / 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 1970] / 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 1970] / 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 1970] / SQL

Aggregation (GROUP BY, SUM, MIN, ...)

```
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	965.45
MSFT	74.26

Relational algebra [Codd 1970] / SQL

Combination (JOIN, UNION, ...)

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

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

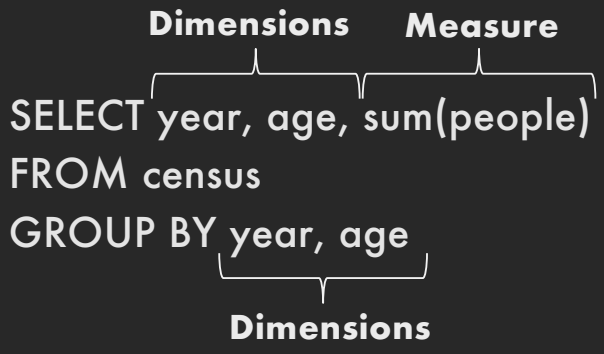
stock	min
AMZN	965.45
MSFT	74.26

→

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

Roll-Up and Drill-Down

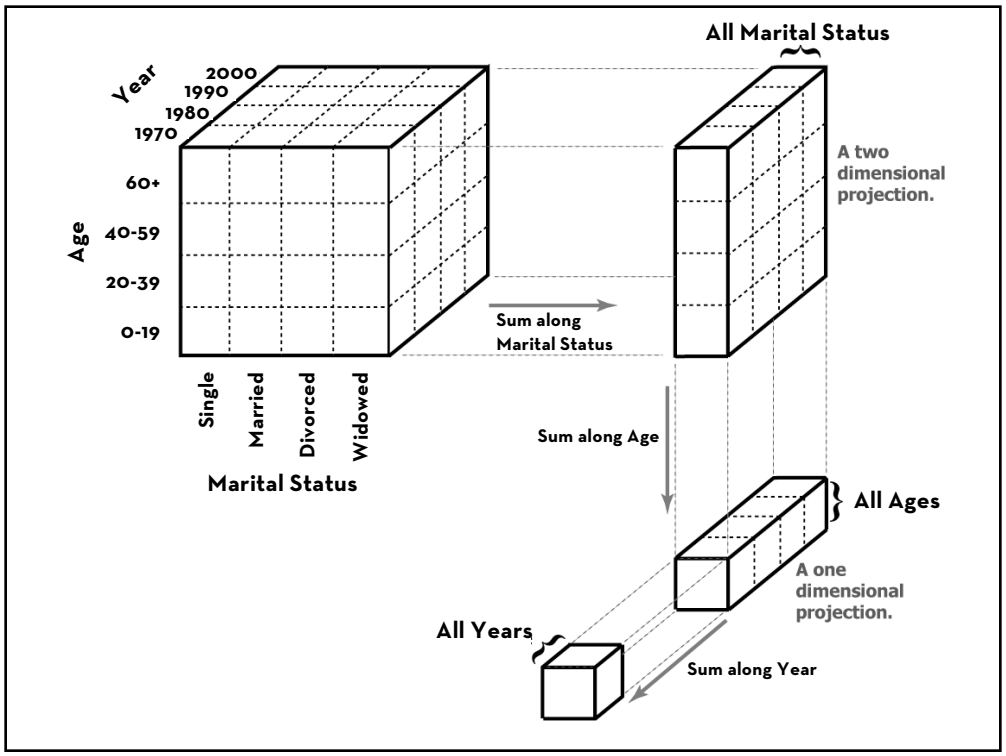
Want to examine population by year and age?
Roll-up the data along the desired dimensions

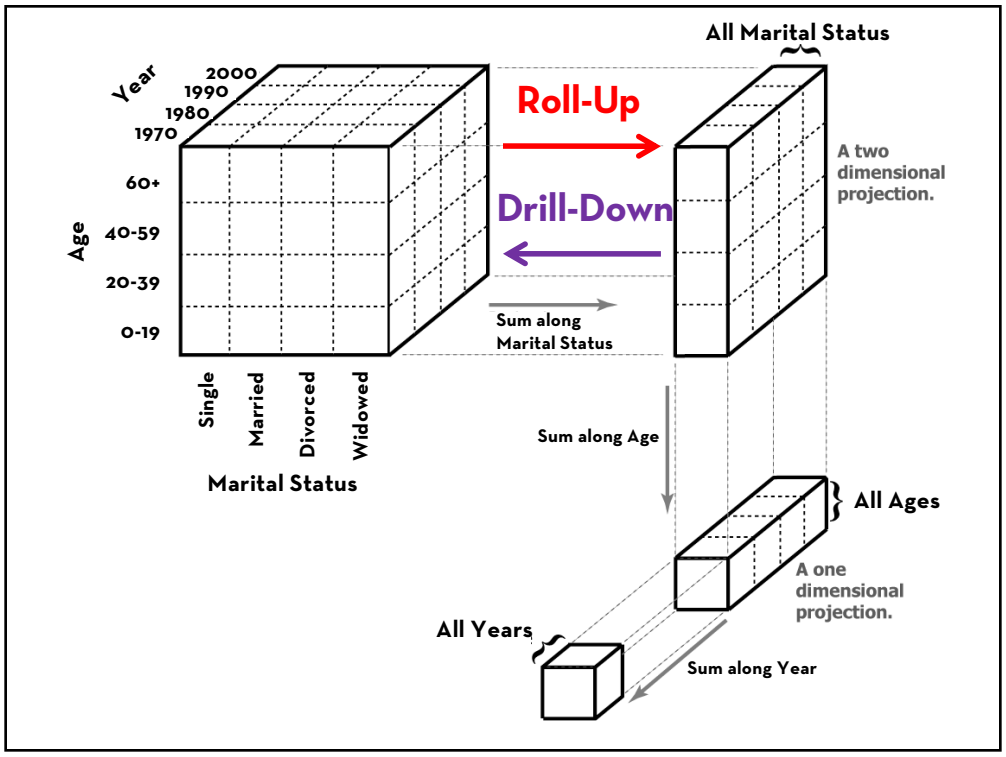


Roll-Up and Drill-Down

Want to breakdown by marital status?
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
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

Common Data Formats

CSV: Comma-Separated Values

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

Common Data Formats

CSV: Comma-Separated Values

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

JSON: JavaScript Object Notation

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

Announcements

Class participation requirements

- Complete readings and notebooks before class
- In-class discussion
- Post at least 1 discussion substantive comment/question per week. 1 pass for the quarter

Class website
https://magrawala.github.io/cs448b-wi20

Lecture/Reading Responses

Good responses typically exhibit one or more

- **Critiques of arguments made in the papers/lectures**
- **Analysis of implications or future directions for ideas in readings/lectures**
- **Insightful questions about the readings/lectures**

Responses should not be summaries

Discussion

Discussion is essential for effective design, evaluation and critique of visualizations

- Attendance for non-SCPD students is mandatory (you have 2 passes before it will affect your grade)
- Laptops not allowed (unless we specifically ask for them)

Assignment 1: Visualization Design

Design a static visualization for a data set.

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

Data: Coterminal Computer Science Master's Degrees at Stanford

Stanford Institutional Research and Decision Support collects a variety of data about the educational programs at Stanford. We have extracted and prepared a small data set about the the number of students who have pursued coterminal Master's degrees in Computer Science between 2014 and 2019. Our data set contains the following information:

Number of records: 112
Variable Names:
Completion Year: Year in which the coterminal degrees were completed.
Coterm Master's Plan: Name of Master's degree student completed.
Coterm Undergraduate Plan: Name of Undergraduate degree student completed.
Number of Completions: Num. of students that completed the corresponding (Master's, Undergraduate) coterminal degree plan.

The extracted dataset is available in csv format: StanfordCSCotermPlans_2014-1029.csv

Due by noon on Mon Jan 13

Assignment 1: Visualization Design

Pick a guiding question, use it to title your visualization

Design a static visualization for that question

You are free to use any tools (including pen & paper)

Deliverables (upload via Canvas; see A1 page)

PDF of your visualization with a short description including design rationale (≤ 4 paragraphs)

Due by noon on Mon Jan 13

Next Monday: Design Exercise

Will review A1submissions

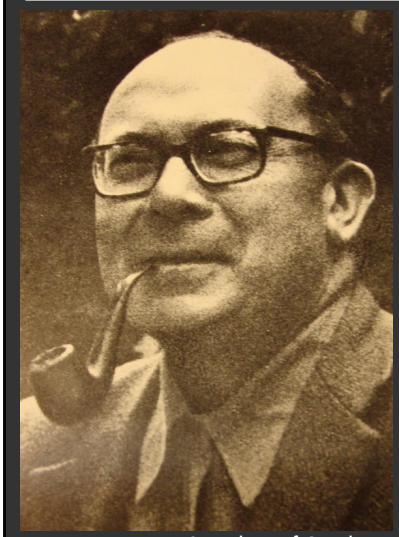
So make sure you get them in on time! (noon Mon)

Will then do a redesign exercise

Make sure to bring paper, pens, etc. for sketching!

Image

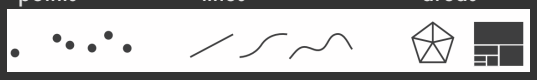
Marks and Visual Variables



Semiology of Graphics
J. Bertin, 1967

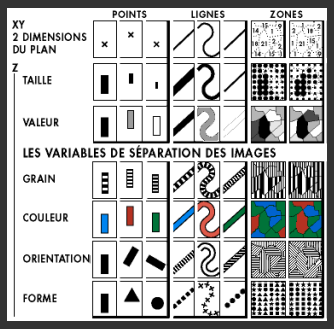
Marks: geometric primitives

points lines areas

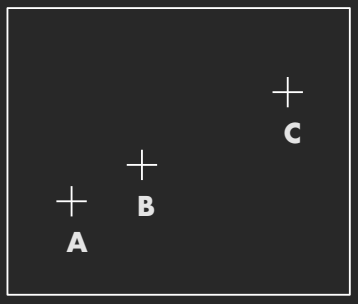


Visual Variables: control mark appearance

- 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

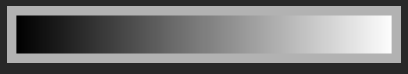
"Resemblance, order and proportional are the three signfields in graphics." - Bertin

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"

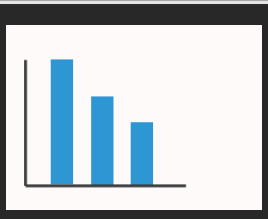
Position	N	O	Q
Size	N	O	Q
Value	N	O	Q
Texture	N	O	
Color	N		
Orientation	N		
Shape	N		

N Nominal
O Ordered
Q Quantitative

Note: Q < O < N

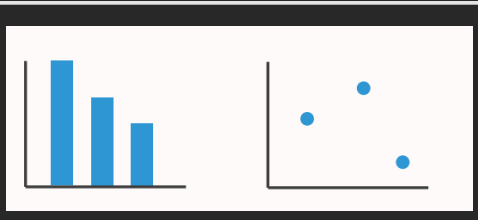
Visual Encoding

Encodings: Map Data to Mark Attr.



mark: lines
data → size (length)

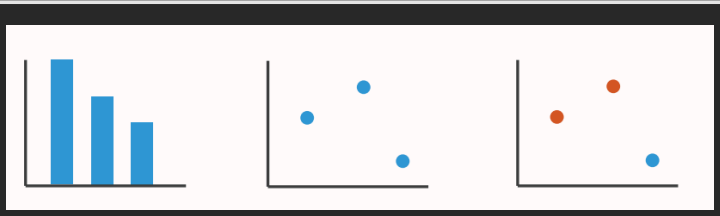
Encodings: Map Data to Mark Attr.



mark: lines
data → size (length)

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

Encodings: Map Data to Mark Attr.

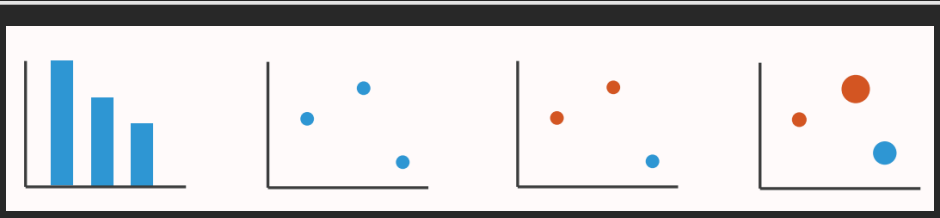


mark: lines
data → size (length)

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

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

Encodings: Map Data to Mark Attr.



mark: lines
data → size (length)

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

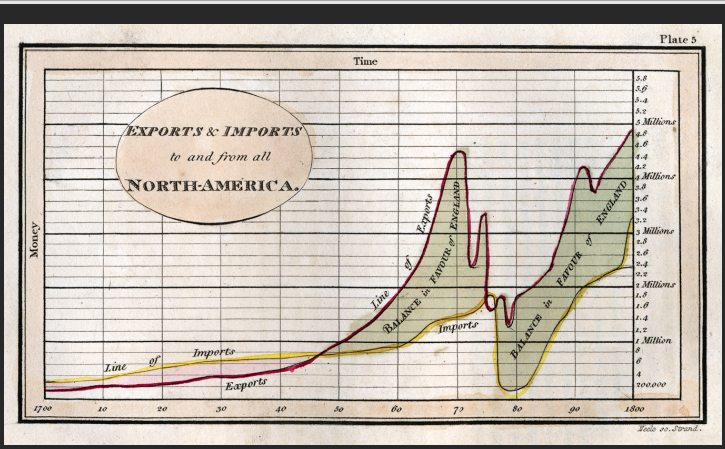
mark: points
$data_1$ → x-pos
$data_2$ → y-pos
$data_3$ → color

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

Deconstructions

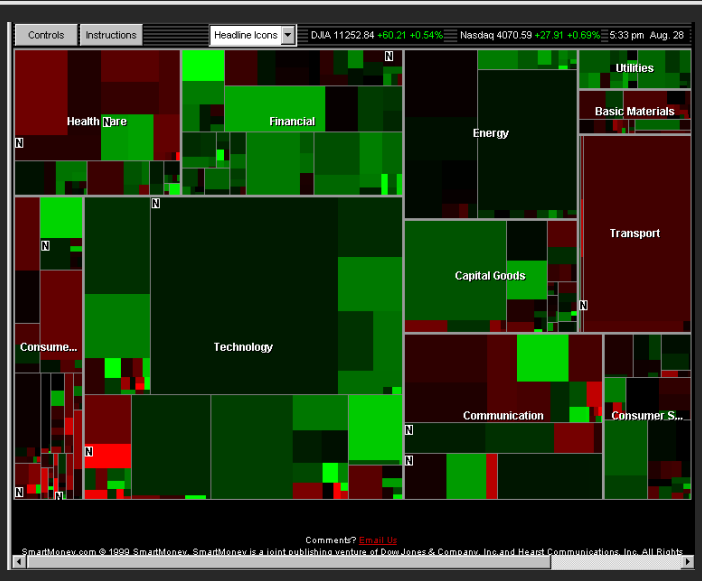
Playfair 1786/1801

Playfair 1786/1801



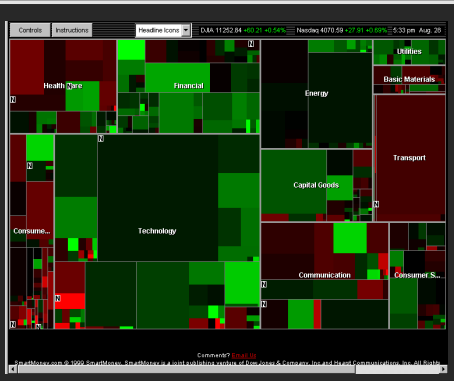
- Time → x-position (Q, linear)
- Exports/Imports Values → y-position (Q, linear)
- Exports/Imports → color (N, O)
- Balance for/against → area (maybe length??) (Q, linear)
- Balance for/against → color (N, O)

Map of the Market [Wattenberg 1998]



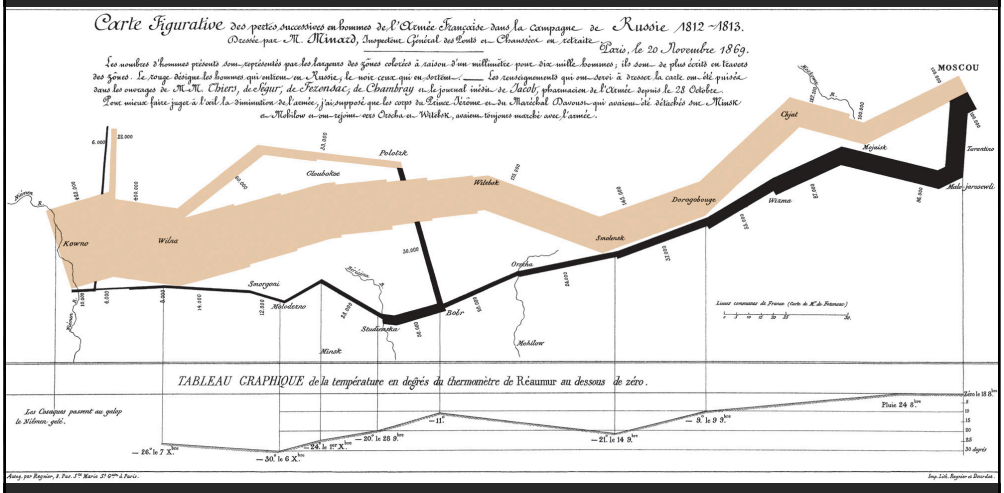
http://www.smartmoney.com/marketmap/

Map of the Market [Wattenberg 1998]

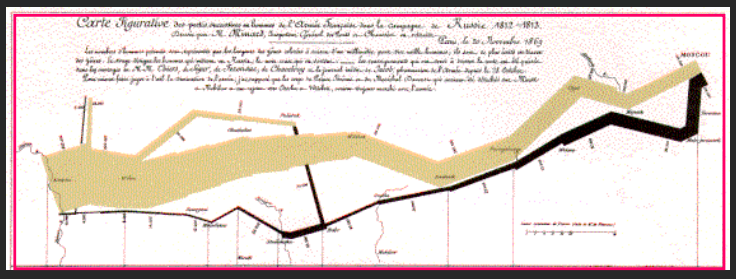


- rectangle size: market cap (Q, linear)
- rectangle position: market sector (N), market cap (Q, linear)
- color hue: loss vs. gain (N, O)
- color value: magnitude of loss or gain (Q, linear)

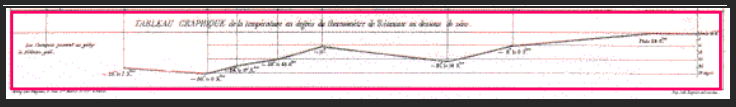
Minard 1869: Napoleon's march



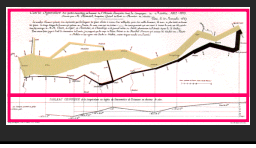
Single axis composition



+



=



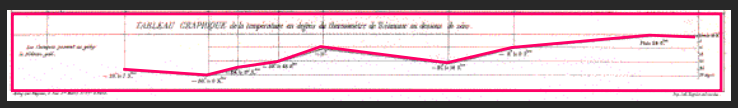
[based on slide from Mackinlay]

Mark composition

temperature → y-position (Q, linear)

+ longitude → x-position (Q, linear)

=



temp over longitude (Q x Q)

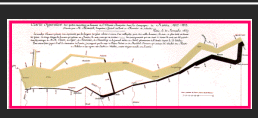
[based on slide from Mackinlay]

Mark composition

latitude → y-position (Q, linear)

\+ longitude → x-position (Q, linear)

\+ army size → width (Q, linear)

\= 

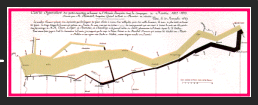
army position (Q x Q) and army size (Q)

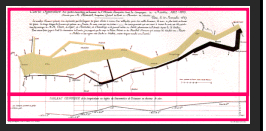
[based on slide from Mackinlay]

latitude (Q, lin)

longitude (Q, lin)

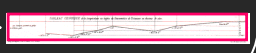
army size (Q, lin)





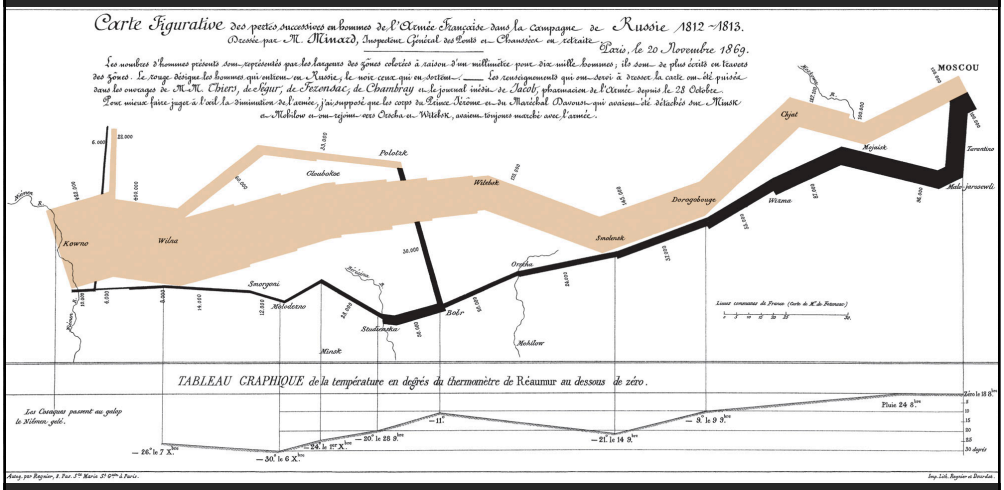
temperature (Q, lin)

longitude (Q, lin)



[based on slide from Mackinlay]

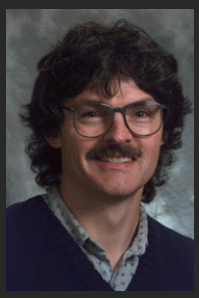
Minard 1869: Napoleon's march



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 $(n+1)^8$

Principles

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

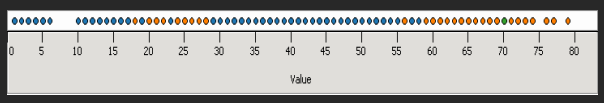
Mackinlay's expressiveness criteria

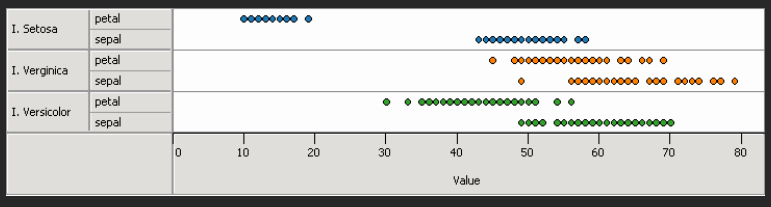
Expressiveness

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

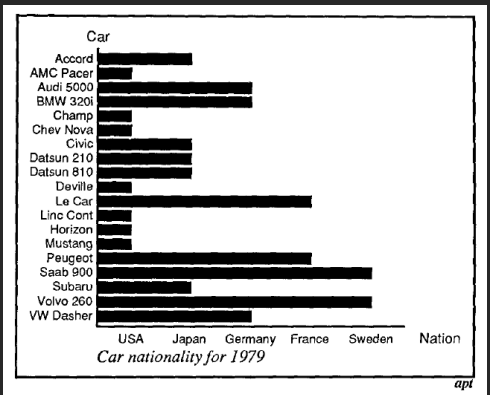


Fig. 11. Incorrect use of a bar chart for the *Nation* relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the *Nation* relation.

[Mackinlay, APT, 1986]

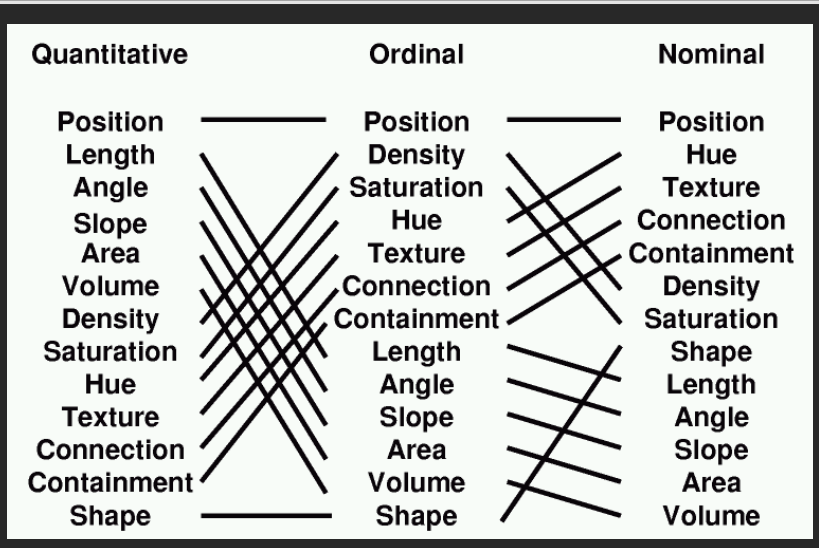
Mackinlay's effectiveness criteria

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