Data and Image Models

Maneesh Agrawala

CS 448B: Visualization
Fall 2020

1

Last Time: The Purpose of Visualization

2

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
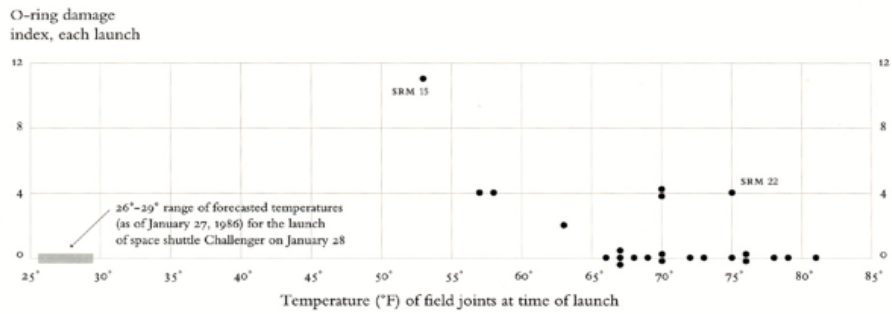- Collaborate and revise
- Emphasize important aspects of data

3

Record information

Gallop, Bay Horse "Daisy" [Muybridge 1884-86]

4

Analysis: Challenger

O-ring damage
index, each launch

Temperature (°F) of field joints at time of launch

26°–29° range of forecasted temperatures (as of January 27, 1986) for the launch of space shuttle Challenger on January 28

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

5

Communicate: Most Powerful Brain

BRAIN MASS grams

BODY MASS kilograms
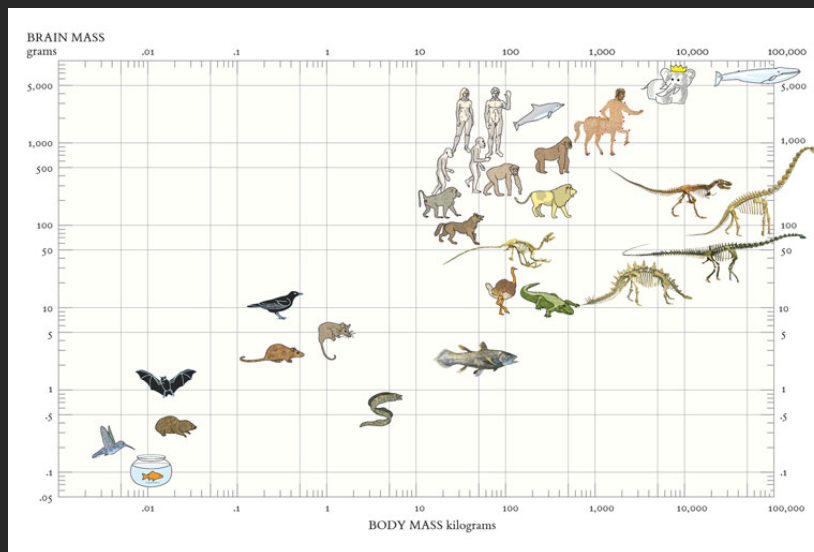
Beautiful Evidence [Tufte]

8

3

Data and Image Models

The big picture

task
  questions, goals,
  assumptions

data
  physical type
    int, float, etc.
  abstract type
    nominal, ordinal, etc.

domain
  metadata
  semantics
  conceptual model
  conventions

processing
algorithms

mapping
  visual encoding

image
graphical marks
visual channel

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

18

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]

19

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, …

20

Nominal, ordinal and quantitative

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: =, ≠, <, >, -, ÷

On the theory of scales of measurements
S. S. Stevens, 1946

21

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: Dim. or Meas.?

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

28

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 attribute 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

32

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

33

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

34

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

35

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

36

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

37

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

→

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

stock	min
AMZN	965.45
MSFT	74.26

38

13

Roll-Up and Drill-Down

Want to examine population by year and age?
Roll-up the data along the desired dimension(s)

Dimensions **Measure**

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

Dimensions

41

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

42

14
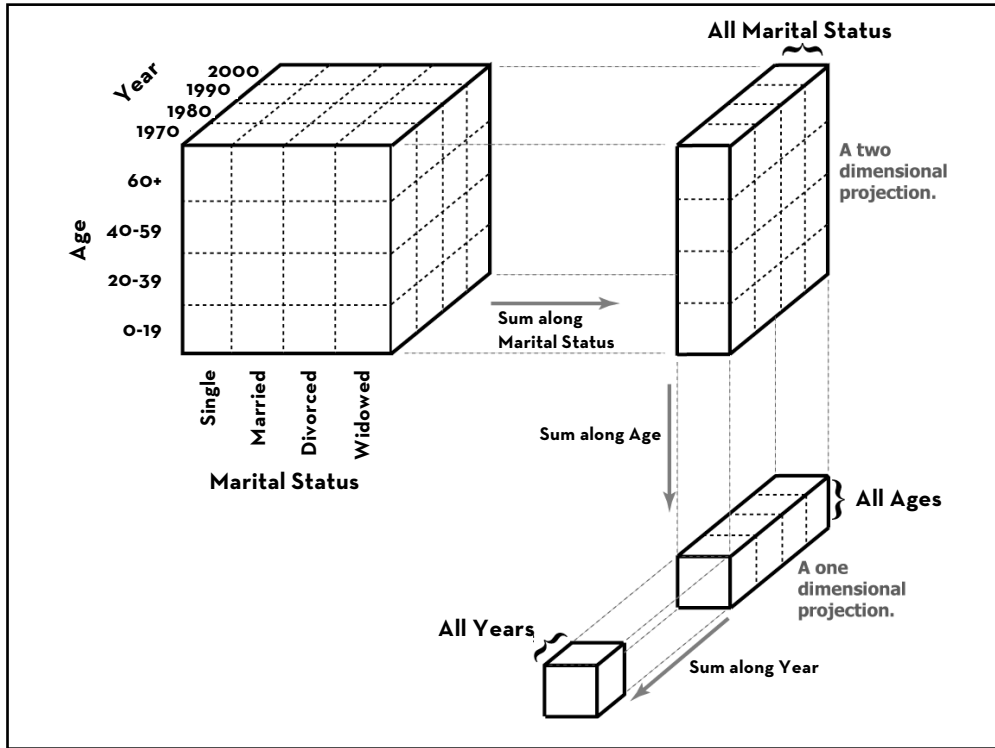
43
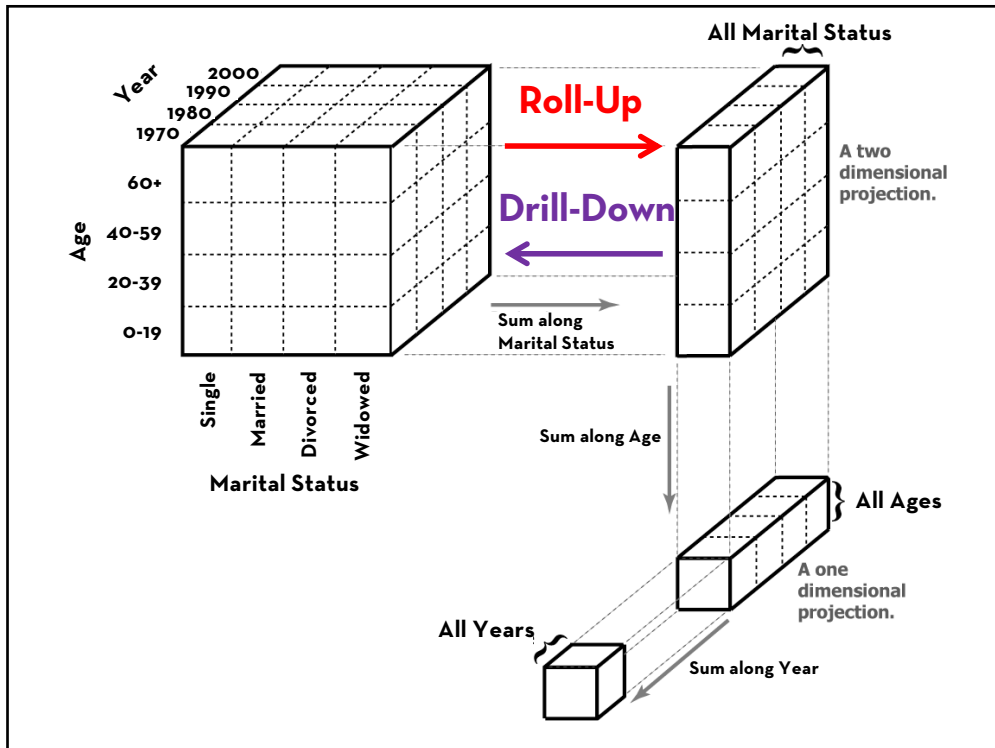
44

Common Data Formats

CSV: Comma-Separated Values

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

46

Common Data Formats

JSON: JavaScript Object Notation

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

47

16

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-fa20

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 is very highly recommended
- Video – please leave on if you are comfortable doing so

54

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: Stanford Olympic Medals

The Stanford Daily publishes a variety of datasets through the Stanford Open Data Portal. We have extracted and wrangled a small data table containing information about the Stanford students that have won medals at the summer Olympics since 1912. Our data contains the following information:

Number of records: 270
Variable Names:
   Athlete Name: Name of athlete.
   Host City: Name of city that hosted the Olympics.
   Host Country: Name of country that hosted the Olympics.
   Year: Year the summer Olympics took place.
   Athlete Team Country: Country whose team the athlete represented.
   Sport: Name of sport athlete competed in.
   Event: Name of event athlete competed in ("–" indicates event name is same as name of sport).
   Medal: Type of medal won (ties are indicated in parentheses).

The extracted dataset is available in csv format: olympic_athletes-wrangled.csv

Due by 7am on Tue Sep 22

55

18

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 7am on Tue Sep 22

56

Next Tuesday: Design Exercise

Will review A1 submissions
So make sure you get them in on time! (7am Tue)

Will then do a redesign exercise
Make sure you have paper, pens, etc. for sketching!

57

Image

58

Marks and Visual Variables

Marks: geometric primitives

points   lines   areas
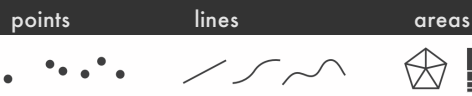
Visual Variables: control mark appearance

Position (2x)

Size

Value

Texture

Color

Orientation

Shape

Semiology of Graphics
J. Bertin, 1967

59

20

Coding information in position
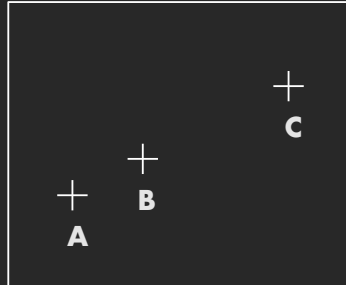
1. A, B, C are distinguishable
2. Three pts colinear: B between A and C
3. BC is twice as long as AB

∴ Encode quantitative variables

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

62

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

65

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

66

Visual Encoding

67

Encodings: Map Data to Mark Attr.
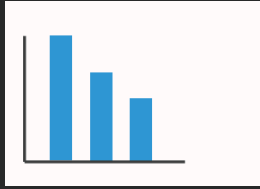
mark: lines
data → size (length)

68

Encodings: Map Data to Mark Attr.
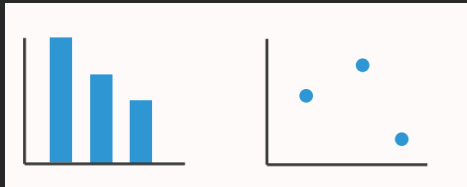
mark: lines
data → size (length)

mark: points
data₁ → x-pos
data₂ → y-pos

69

Encodings: Map Data to Mark Attr.
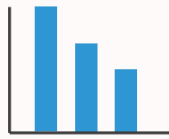
mark: lines
data → size (length)

mark: points
data$_1$ → x-pos
data$_2$ → y-pos
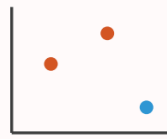
mark: points
data$_1$ → x-pos
data$_2$ → y-pos
data$_3$ → color

70

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

71

Deconstructions

80

Playfair 1786/1801
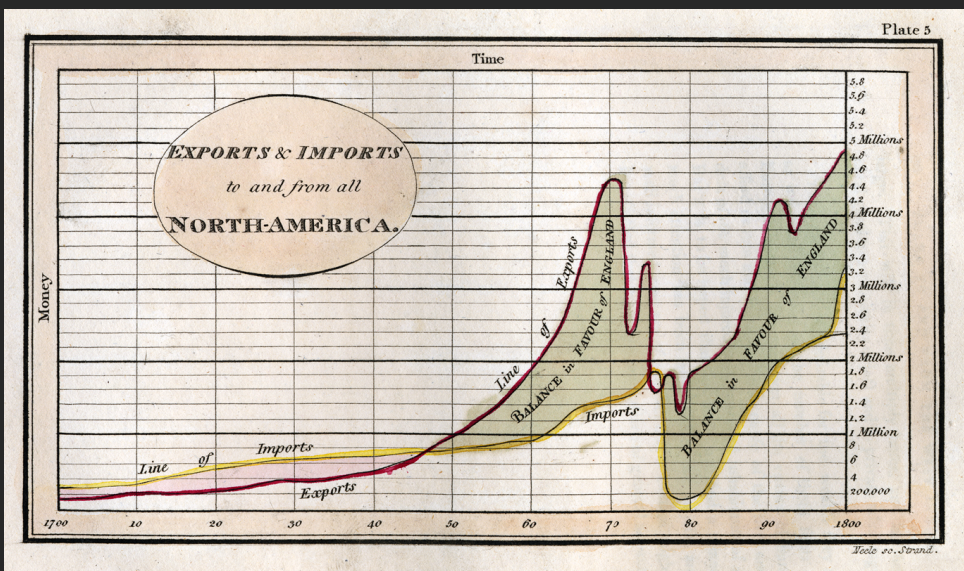
84

Playfair 1786/1801
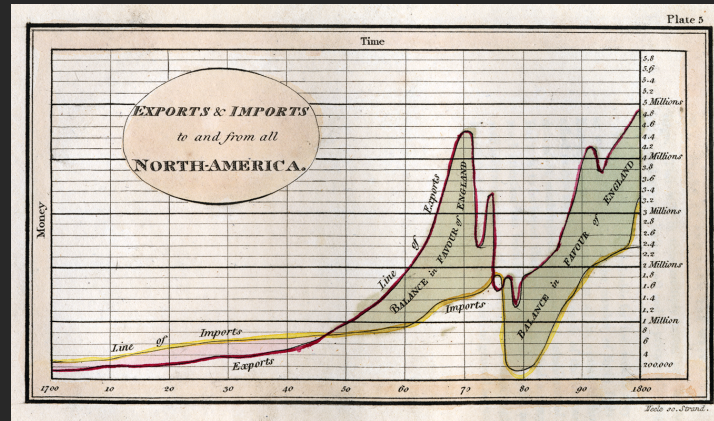
- Time → x-position (Q, linear)
- Exports/Imports Values → y-position (Q, linear)
- Exports/Imports → color (N, O)
- Balance for/against → area (maybe length??) (Q, linear)
- Balance for/against → color (N, O)
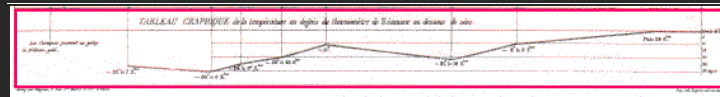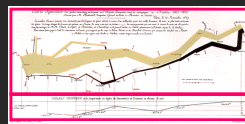
85

Map of the Market [Wattenberg 1998]

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

86

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)

87

Minard 1869: Napoleon's march
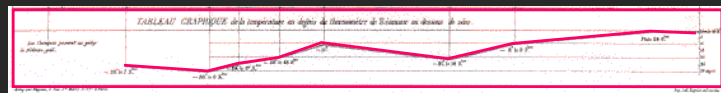

88
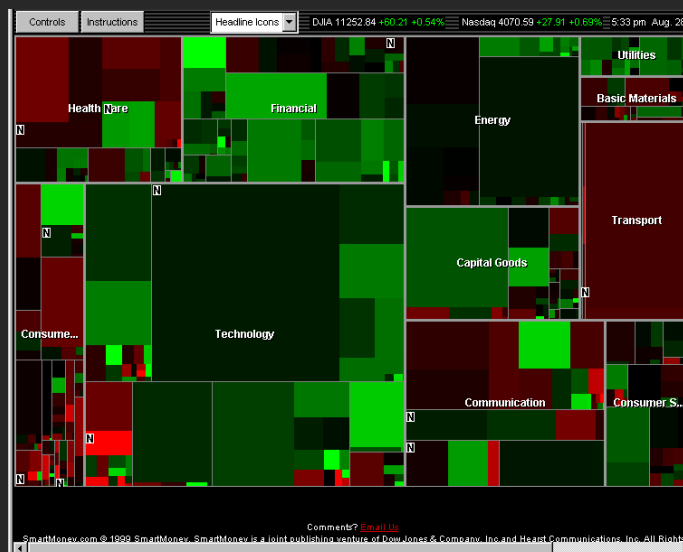
Single axis composition
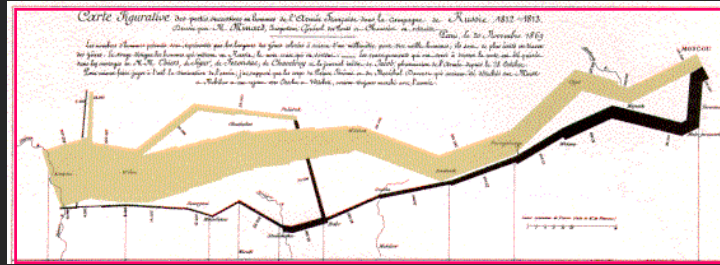
89

Mark composition

|temperature → y-position (Q, linear)

+   longitude → x-position (Q, linear)

=

temp over longitude (Q x Q)

[based on slide from Mackinlay]

90

Mark composition

latitude → y-position (Q, linear)

+ longitude → x-position (Q, linear)

+ army size → width (Q, linear)
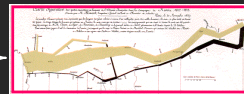

=

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

[based on slide from Mackinlay]

91

---

latitude (Q, lin)
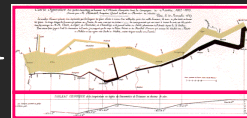
longitude (Q, lin)

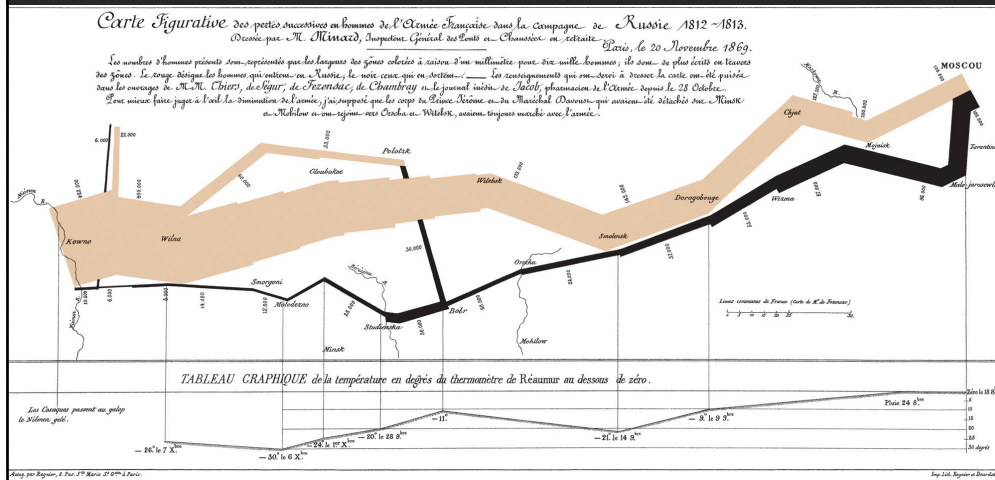army size (Q, lin)

temperature (Q, lin)

longitude (Q, lin)

[based on slide from Mackinlay]

92