Data and Image Models

Maneesh Agrawala

CS 448B: Visualization
Fall 2021

1

The big picture

task
questions, goals, assumptions

data
abstract type
nominal, ordinal, etc.

domain
metadata
semantics
conceptual model
conventions

processing algorithms

mapping
visual encoding

image
graphical marks
visual channel

15

Topics

Properties of data
Properties of the image
Mapping data to images

16

Data

17

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 Models/Types

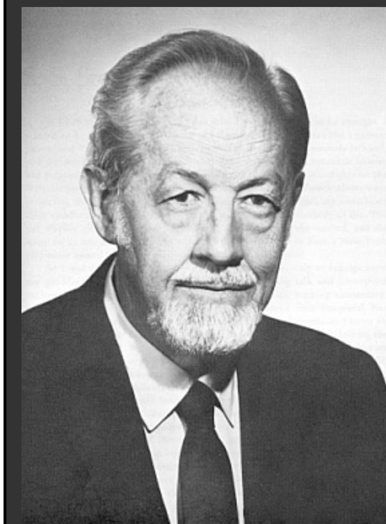
- **1D (sets and sequences)**
- **Temporal**
- **2D (maps)**
- **3D (shapes)**
- **nD (relational)**
- **Trees (hierarchies)**
- **Networks (graphs)**

Are there others?

The eyes have it: A task by data type taxonomy for information visualization [Schneiderman 96]

19

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, …
Counts and amounts
Like a geometric vector, origin is meaningful
Operations: =, ≠, <, >, -, ÷

21

From data model to N,O,Q

Data model
- 32.5, 54.0, -17.3, …
- Floating point numbers

Conceptual model
- Temperature (°C)

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

22

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.

23

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, …

25

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

27

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

Data Tables and Transformations

32

Relational data model

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

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

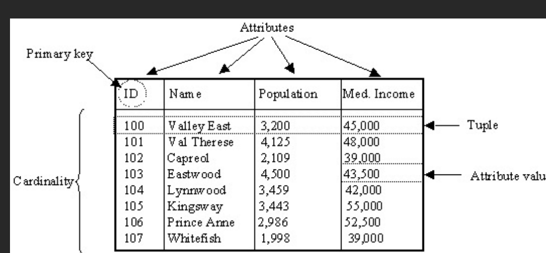
- Each record is a fixed-length tuple

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

33

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

34

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

35

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

36

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

37

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	957.10
MSFT	74.26

38

Roll-Up and Drill-Down

Want to examine population by year and age?
Roll-up the data (i.e. aggregate) along marst.

SELECT $\underbrace{\text{year, age,}}_{\text{Dimensions}}$ $\underbrace{\text{sum(people)}}_{\text{Measure}}$
FROM census
GROUP BY $\underbrace{\text{year, age}}_{\text{Dimensions}}$

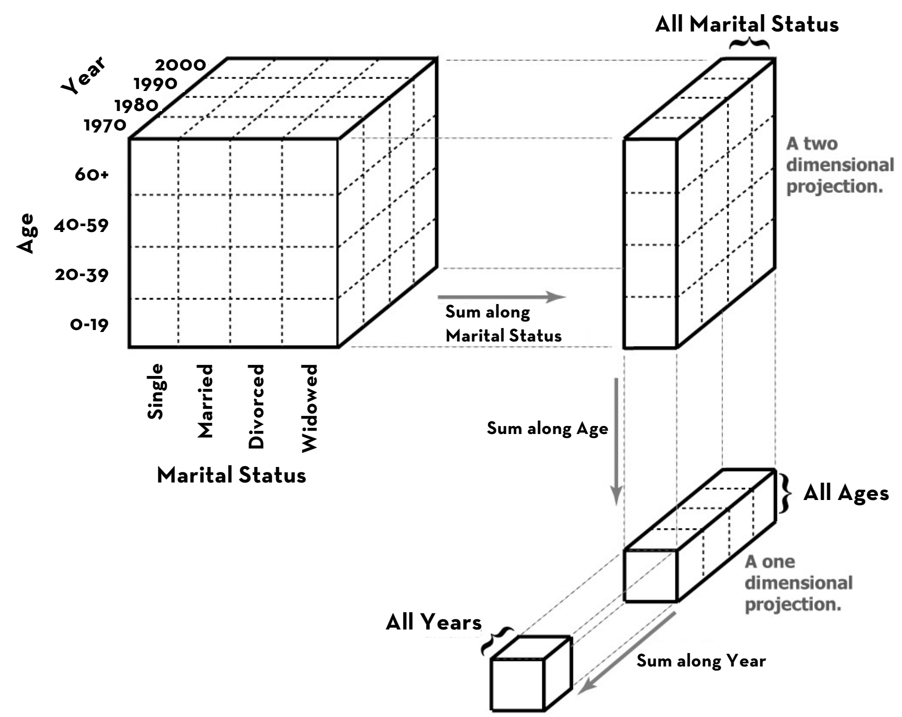
42

Roll-Up and Drill-Down

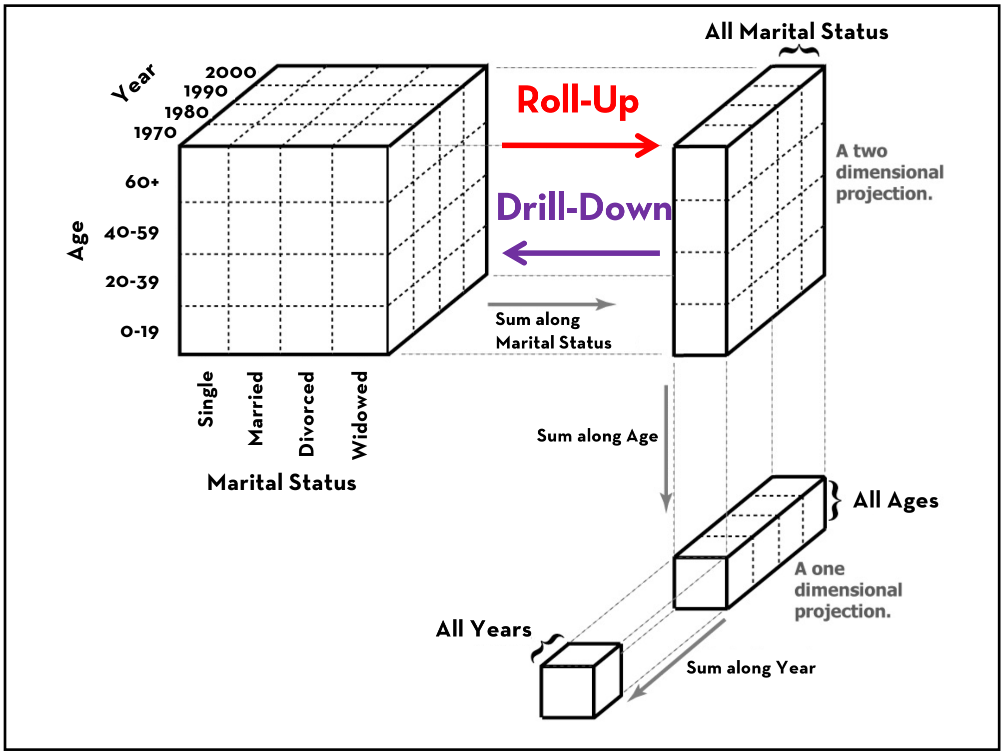
Want to breakdown by marital status?
Drill-down into additional dimensions

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

43



44



45

Original

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
. . .				

Pivoted or Cross-Tabulation

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? Why?

46

Tidy Data [Wickham 2014]

How do rows, columns, and tables match up with observations, variables, and types? In "tidy" data:

1. Each variable forms a column
2. Each observation forms a row
3. Each type of observational unit forms a table

Advantage: Flexible starting point for analysis, transformation, and visualization. Our pivoted table variant was not "tidy"!

47

Common Data Formats

CSV: Comma-Separated Values

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

48

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},
  ...
]
```

49

Announcements

Class participation requirements

- Complete readings and notebooks before class
- In-class discussion
- Post at least 1 discussion substantive comment/question per week.
- Due by 7am the following Monday
- 1 pass for the quarter

Class home page
https://magrawala.github.io/cs448b-fa21/

54

Reading/Notebook/Lecture Responses

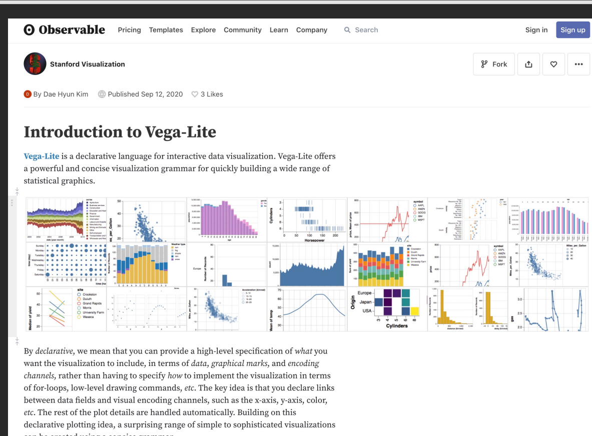
Good responses typically exhibit one or more

- 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

55

Observable Notebooks – Vega-Lite



Vega-Lite is a *declarative* API for producing visualizations
Make sure to go through a do exercises (fork the notebook)

Monday 9/27 lecture will assume you've done 1st three notebooks

56

Office Hours

Maneesh: 2-3pm Wed, Coupa Café Y2E2 and Canvas/Zoom
Dae Hyun: 10-11am Thu, CEMEX Aud and Canvas/Zoom
Shana Hadi: 7-8:00pm Sun, via Canvas/Zoom

Happy to schedule other OH by appointment
Outside of OH use Slack to connect with us

https://canvas.stanford.edu/courses/144332/external_tools/11232

57

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 Undergraduate Majors

The Stanford Daily publishes a variety of datasets through the Stanford Open Data Portal. They have published a data table containing information about the number of Stanford students majoring in 70 different subject areas from 2011-2019. We have filtered and wrangled this data to the top 10 majors over the time period to produce a dataset with the following variables:

Number of records:
Variable Names:
- **Year:** Academic year between 2011-2012 and 2018-2019.
- **Subject:** Subject areas in which students majored.
- **Number of Students:** Number of students majoring in the area.

The extracted dataset is available in csv format: StanfordTopTenMajors2010s.csv

Due by 7am on Mon Sep 27

59

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)

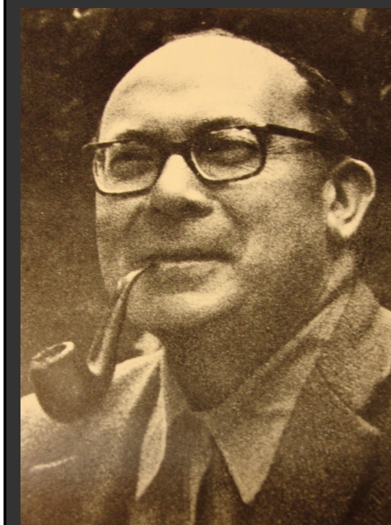
Due by 7am on Mon Sep 27

60

Image

62

Marks and Visual Variables

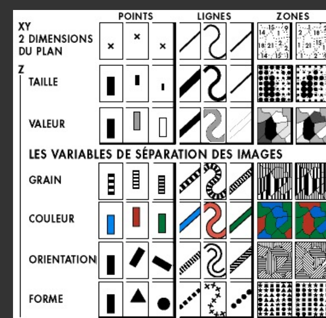


Semiology of Graphics
J. Bertin, 1967

Marks: geometric primitives

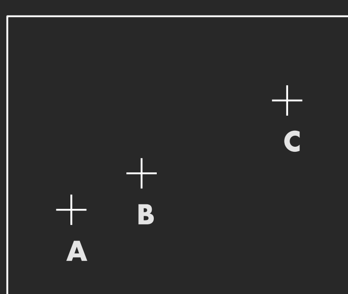
Visual Variables: control mark appearance

Position (2x)
Size
Value
Texture
Color
Orientation
Shape



63

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

66

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

69

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		

N Nominal
O Ordered
Q Quantitative

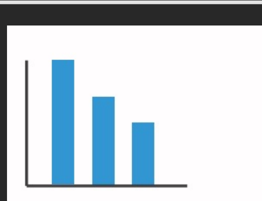
Note: Q < O < N

70

Visual Encoding

71

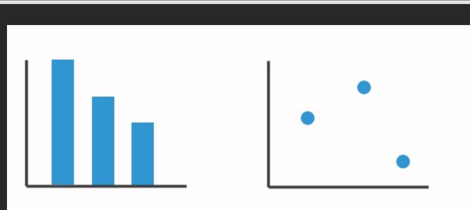
Encodings: Map Data to Mark Attr.



mark: rect
data → size (height)

72

Encodings: Map Data to Mark Attr.

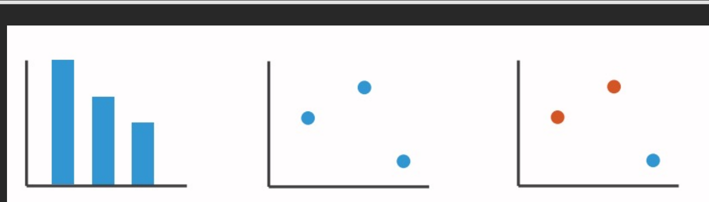


mark: rect
data → size (height)

mark: point
data$_1$ → x-pos
data$_2$ → y-pos

73

Encodings: Map Data to Mark Attr.



mark: rect
data → size (height)

mark: point
data$_1$ → x-pos
data$_2$ → y-pos

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

74

Encodings: Map Data to Mark Attr.

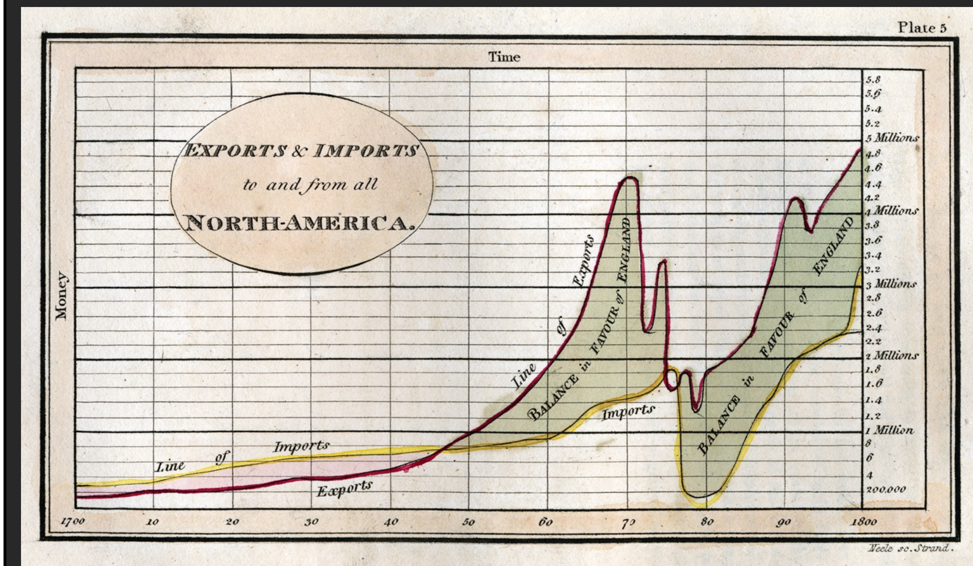
mark: rect	mark: point	mark: point	mark: point
data → size (height)	$data_1$ → x-pos	$data_1$ → x-pos	$data_1$ → x-pos
	$data_2$ → y-pos	$data_2$ → y-pos	$data_2$ → y-pos
		$data_3$ → color	$data_3$ → color
			$data_4$ → size

75

Deconstructions

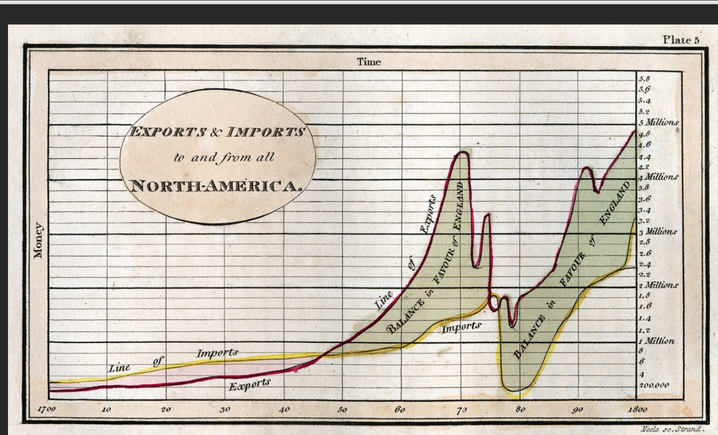
84

Playfair 1786/1801



88

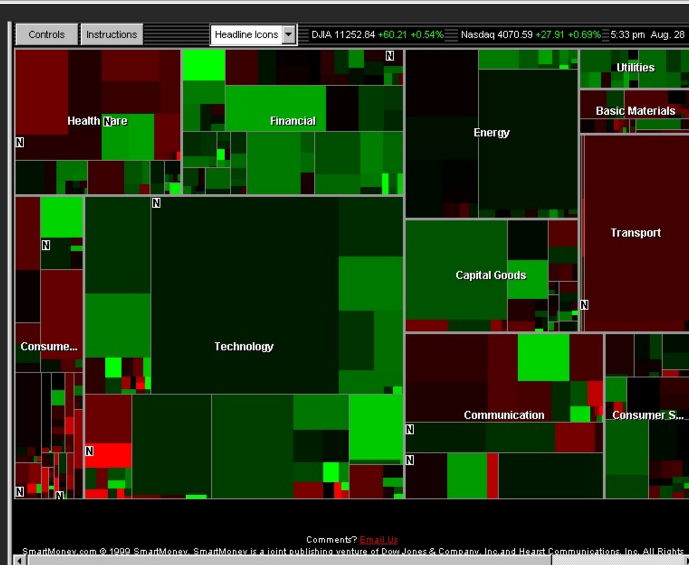
Playfair 1786/1801



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

89

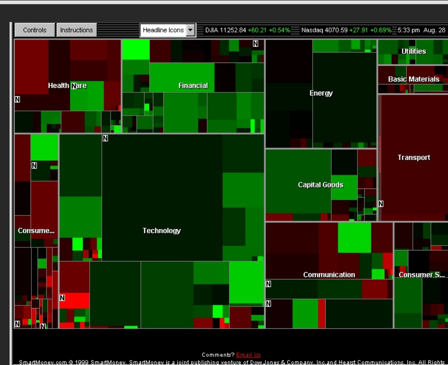
Map of the Market [Wattenberg 1998]



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

90

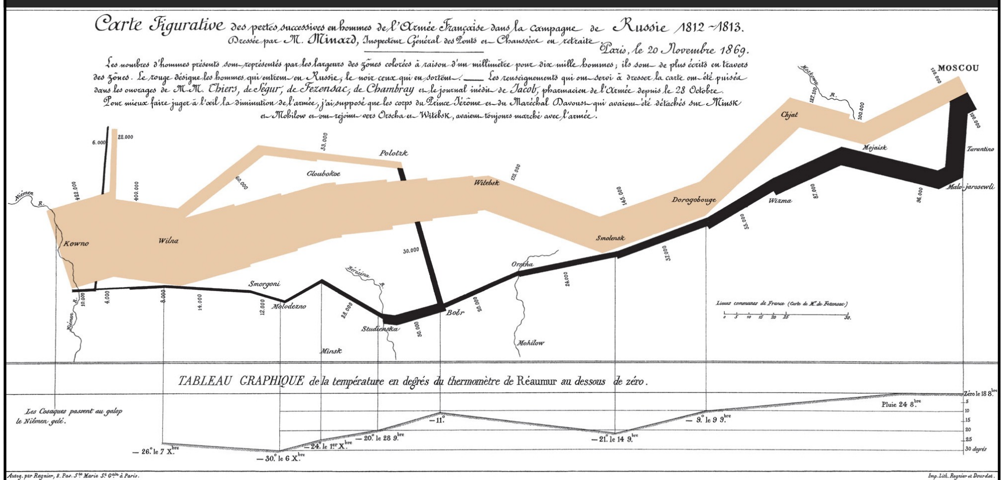
Map of the Market [Wattenberg 1998]



- rectangle size: 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)

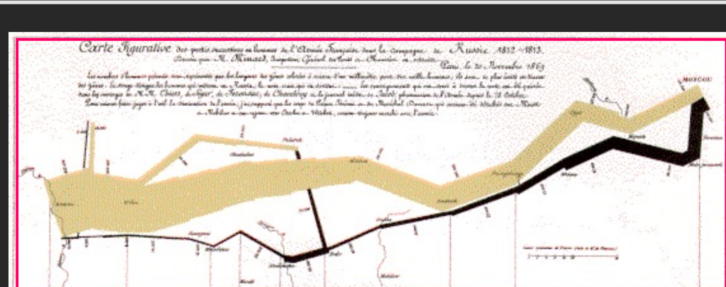
91

Minard 1869: Napoleon's march

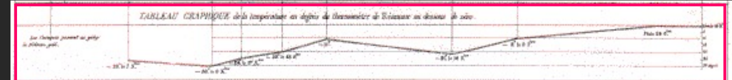


92

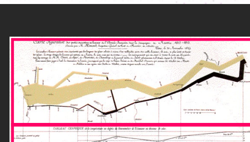
Single axis composition



+



=



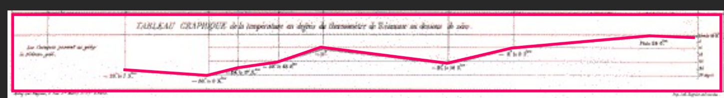
[based on slide from Mackinlay]

93

Mark composition

temperature → y-position (Q, linear)

\+ longitude → x-position (Q, linear)

= 

temp over longitude (Q x Q)

[based on slide from Mackinlay]

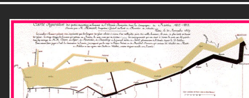
94

Mark composition

latitude → y-position (Q)

\+ longitude → x-position (Q)

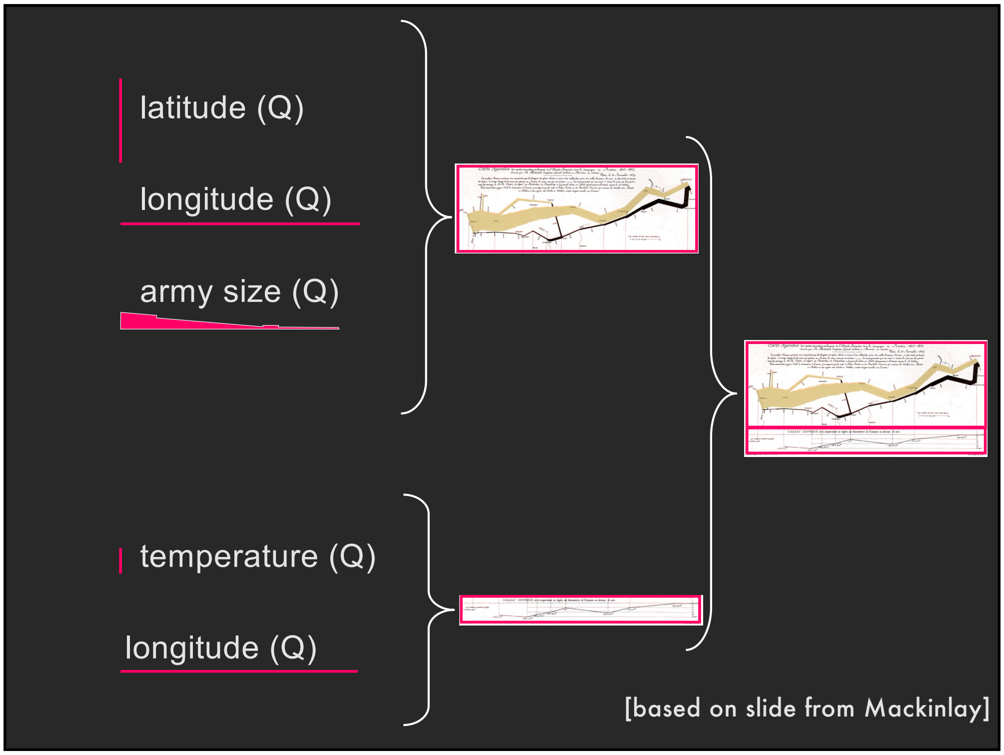
\+ army size → width (Q)

= 

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

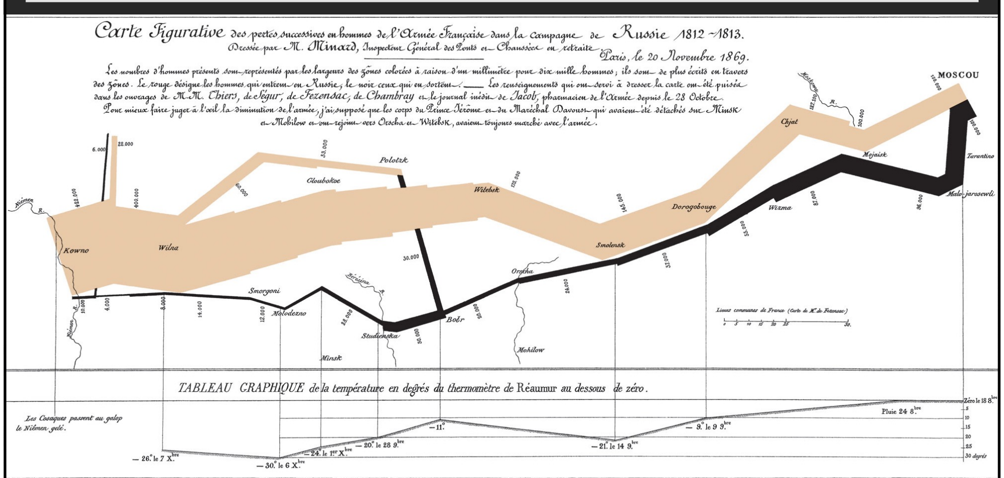
[based on slide from Mackinlay]

95



96

Minard 1869: Napoleon's march

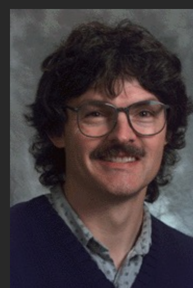


Depicts at least 5 quantitative variables
Any others?

97

Automated design

Jock Mackinlay's APT 86



98

Combinatorics of encodings

Challenge:

Assume 8 visual encodings and n data fields

Pick the best encoding from the exponential number of possibilities $(n+1)^8$

99

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

100

Mackinlay's expressiveness criteria

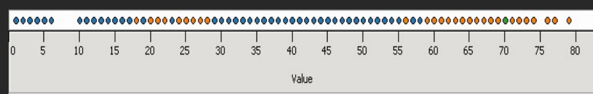
Expressiveness

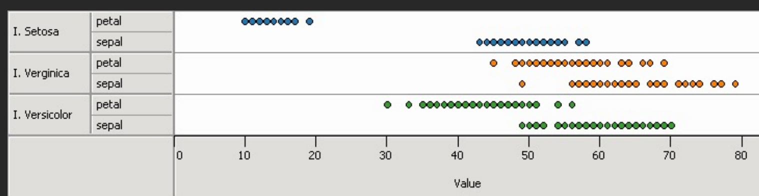
A set of facts is expressible in a visual language if the sentences (i.e. the visualizations) in the language express *all* the facts in the set of data, and *only* the facts in the data.

101

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





102

Expresses facts not in the data

A length is interpreted as a quantitative value;
∴ Length of bar says something untrue about N data

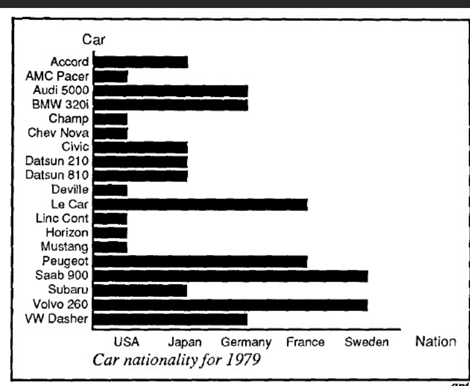


Fig. 11. Incorrect use of a bar chart for the *Nation* relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the *Nation* relation.

[Mackinlay, APT, 1986]

103

Mackinlay's effectiveness criteria

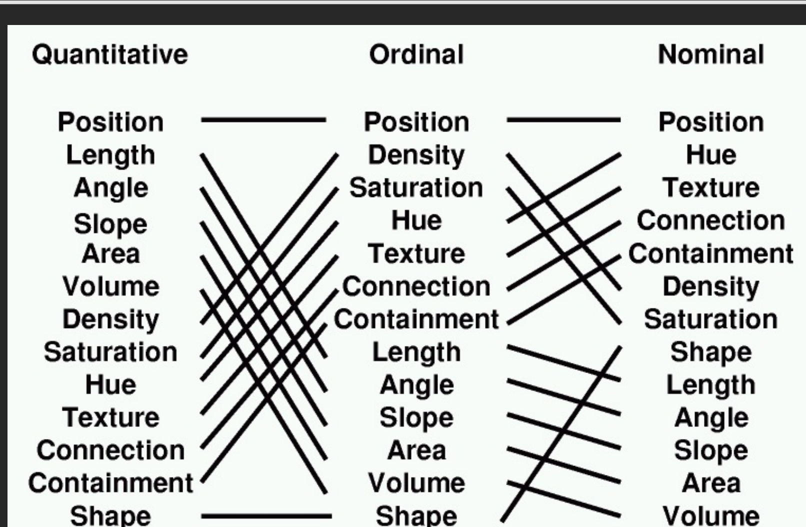
Effectiveness

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

104

Mackinlay's ranking



Conjectured *effectiveness* of the encoding

105

Mackinlay's Design Algorithm

User formally specifies data model and type

Input: list of data variables ordered by importance

APT searches over design space

- Tests expressiveness of each visual encoding (rule-based)
- Generates encodings that pass test
- Rank by perceptual effectiveness criteria

Outputs *most effective* visualization

107

Automatic chart construction

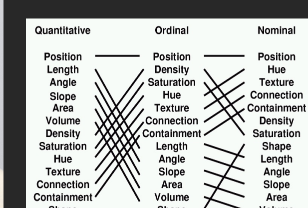
Automating the design of graphical presentation of relational information
J. Mackinlay, 1986

Encode most important data using highest ranking visual variable for the data type

Price	Mileage	Weight	Repair
13,500	22	3000	great
7,200	31	1500	ok
11,300	12	4200	terrible
...	...	...	...

→

1. Price (Q)
2. Mileage (Q)
3. Weight (Q)
4. Repair (N)



→

mark: lines

Price → y-pos (Q)
Mileage → x-pos (Q)
Weight → size (Q)
Repair → color (N)

109

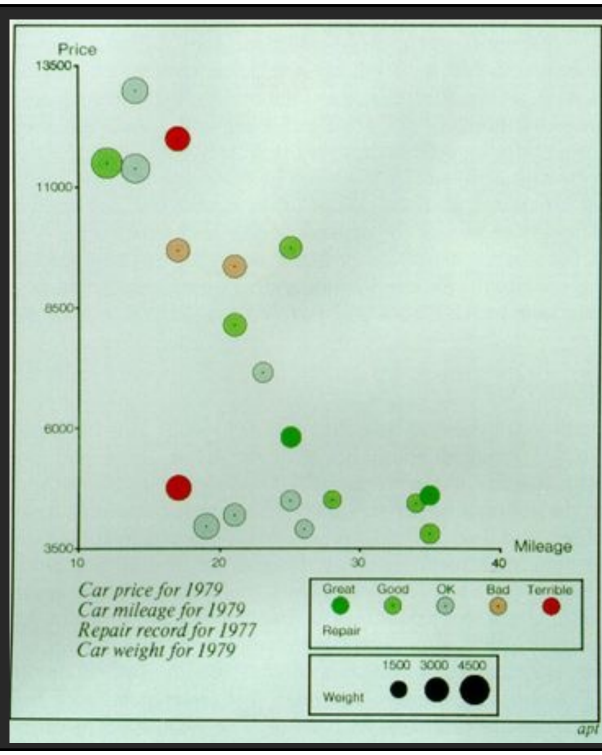
Cars Data

1. Price (Q)
2. Mileage (Q)
3. Weight (Q)
4. Repair (Q)

Quantitative

Position
Length
Angle
Slope
Area
Volume
Density
Saturation
Hue
Texture
Connection
Containment
Shape

[Mackinlay, APT, 1986]



110

Limitations

Does not cover many visualization techniques

- Networks, maps, diagrams
- Also, 3D, animation, illustration, ...

Does not consider interaction
Does not consider semantics or conventions
Assumes single visualization as output

111

Summary

Formal specification

- **Data model:** relational data, N,O,Q types
- **Image model:** marks, attributes, encodings
- **Encodings mapping data to image**

Choose expressive and effective encodings

- **Rule-based test of expressiveness**
- **Perceptual effectiveness rankings**

112

113

Microsoft Excel - fischer.iris.2.xls

	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

Sepal and petal lengths and widths for three species of iris [Fisher 1936].

115

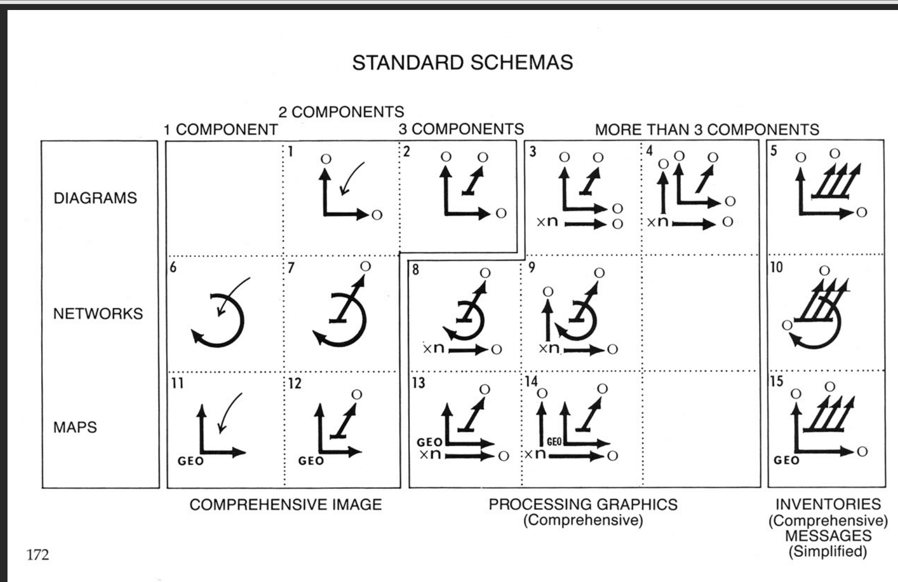
	I. Setosa				I. Verginica				I. Versicolor			
	petal		sepal		petal		sepal		petal		sepal	
	length	width	length	width	length	width	length	width	length	width	length	width
1	14	2	50	33	56	24	67	31	45	13	57	28
2	10	2	46	36	51	23	69	31	47	16	63	33
3	16	2	48	31	52	20	65	30	47	14	70	32
4	14	1	49	36	51	19	58	27	40	12	58	26
5	13	2	44	32	45	17	49	25	33	10	50	23
6	16	2	51	38	50	19	63	25	41	10	58	27
7	16	2	50	30	49	18	63	27	45	15	60	29
8	19	4	51	38	56	21	64	28	33	10	49	24
9	14	2	49	30	51	19	58	27	39	14	52	27
10	14	2	50	36	55	18	64	31	39	12	58	27
11	15	4	54	34	50	15	60	22	42	15	59	30
12	14	2	55	42	57	23	69	32	44	13	63	23
13	14	2	44	29	49	20	56	28	49	15	63	25
14	14	1	48	30	58	18	67	25	30	11	51	25
15	17	3	57	38	54	21	69	31	36	13	56	29
16	15	4	51	37	61	25	72	36	44	14	66	30
17	13	2	55	35	55	21	68	30	50	17	67	30
18	13	2	44	30	56	22	64	28	45	15	62	22
19	16	2	47	32	51	15	63	28	46	14	61	30
20	12	2	50	32	59	23	68	32	39	11	56	25

Format of the data in Appendix 14, pp. 365-366

Chambers, Cleveland, Kleiner, Tukey, *Graphical Methods for Data Ana*

116

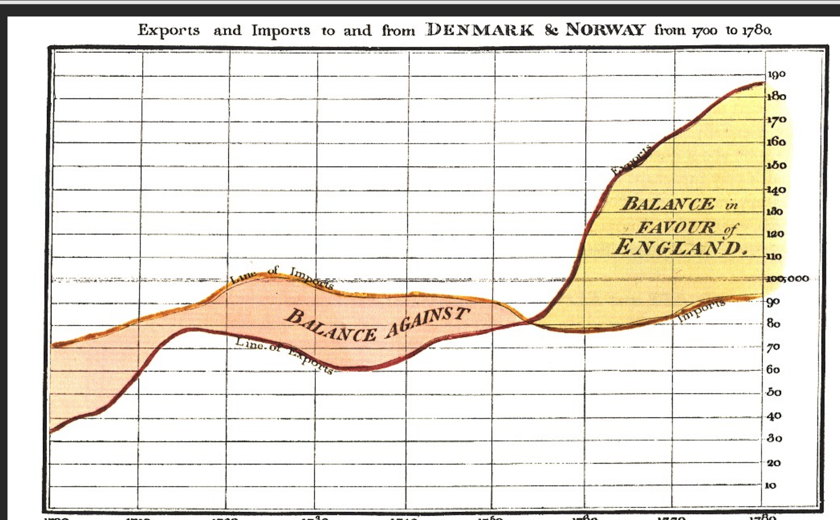
Bertin's specification



[Bertin, Semiology of Graphics, 1967]

117

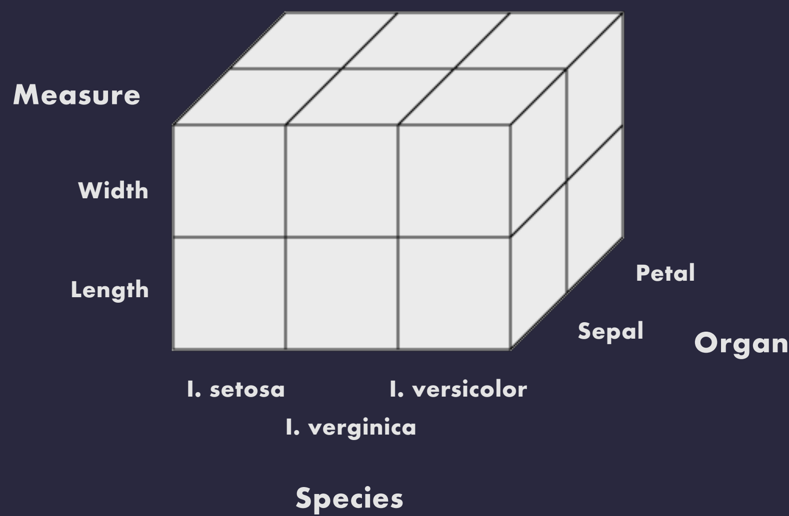
Communicate: Exports and Imports



[Playfair 1786]

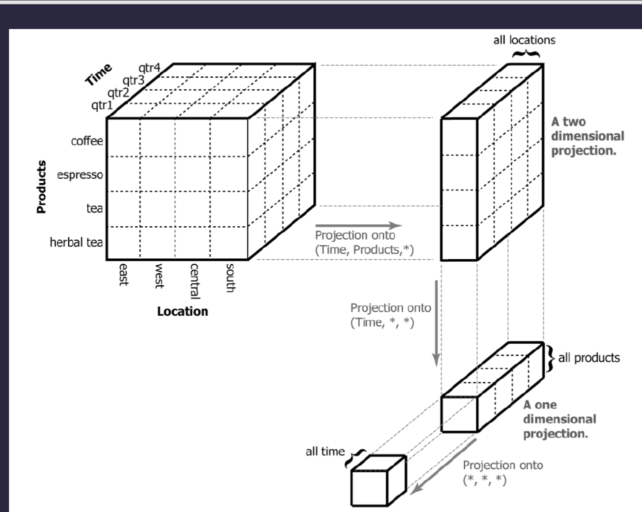
118

Data cube



119

Projections summarize data



Multiscale visualization using data cubes [Stolte et al. 02]

120