CDS 6324 DATA VISUALIZATION

LECTURE 2: DATA & IMAGE MODELS

“Think of yourself as a craftsperson”

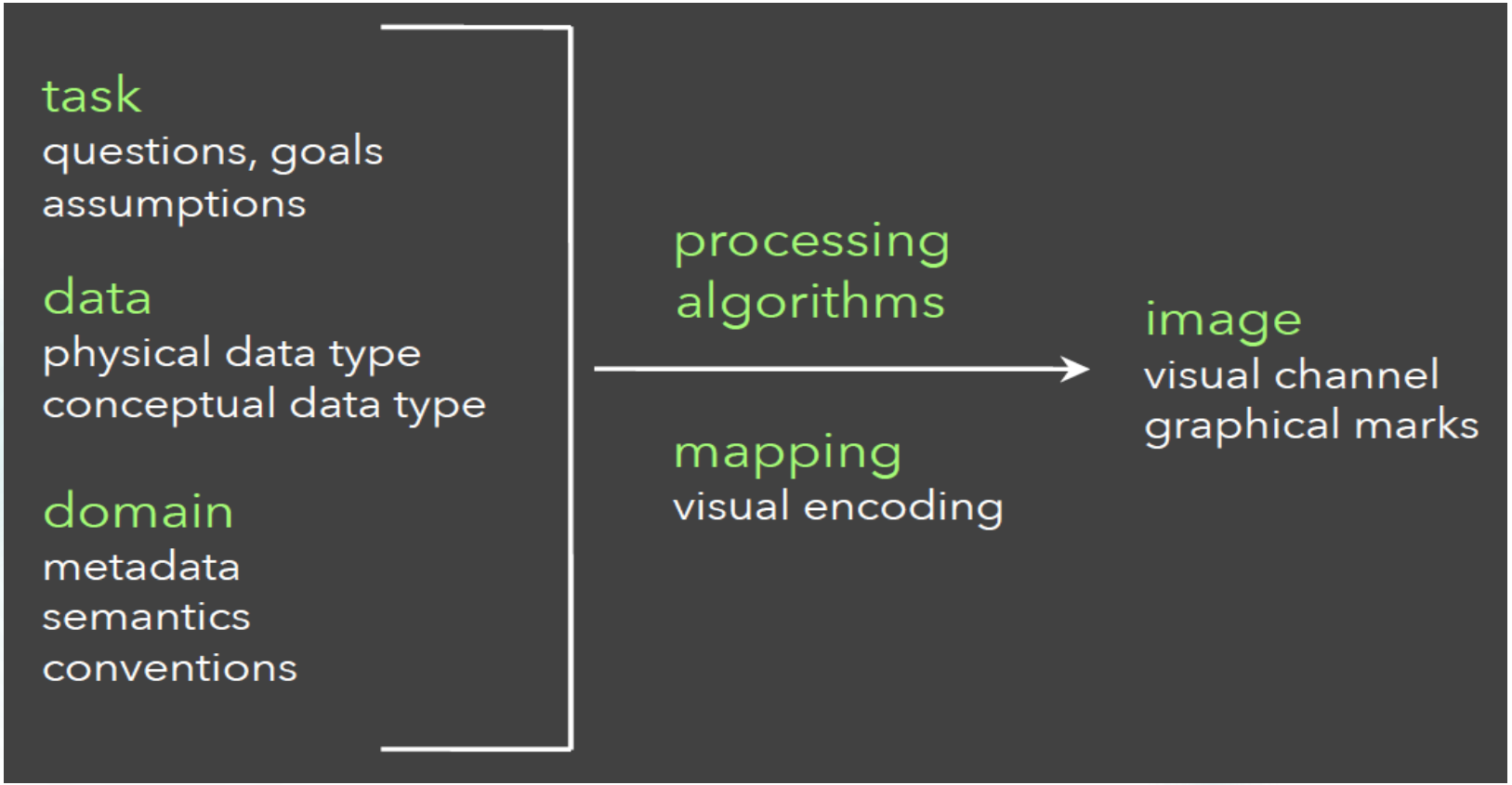
--Stephen Few

Data and Image Models

Topics

- Properties of Data
- Properties of Images
- Mapping Data to Images

The Big Picture



Data

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)

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

Qualitative vs. Quantitative Data

Categorical / Qualitative

- Describes qualities or characteristics that cannot be measured numerically
- Can be observed but not measured.
- Examples:
  - Color, gender, race, hair color, country, taste, smell

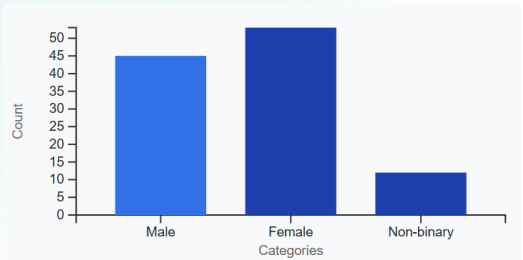
Quantitative

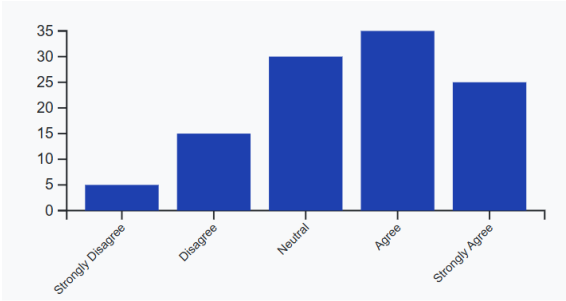
- Can be counted or measured using numbers.
- Represents quantities, amounts, or ranges.
- Examples:
  - Height, weight, age, temperature, scores, counts, prices

Level of Measurement : Qualitative

- **N** – Nominal
  - Labels or categories without no order or rank
  - Ex: Gender: male, female, …
  - Operations: mode, freq count, %
  - Common visualization: bar chart, pie chart, treemap

- **O** – Ordinal
  - Categories with clear, meaningful order
  - Ex: Medal: Gold, Silver, Bronze
  - Operations: mode, median, freq count, %, rank ordering
  - Common visualization: bar chart, stacked chart, dot plots

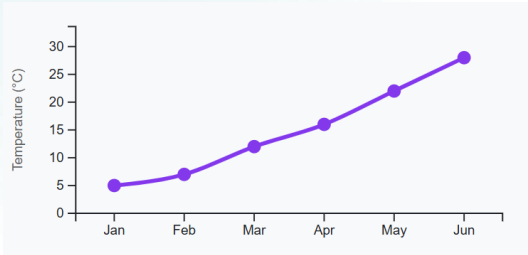


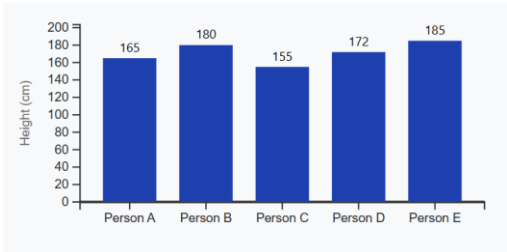


Level of Measurement : Quantitative

- Q – Interval
  - Has order and equal distance between values, but zero point is arbitrary
  - Ex: Temperature
  - Operations: Mean, median, mode, std dev, addition, subtraction
  - Can measure distances or spans
  - Common visualization: line chart, histogram, area chart

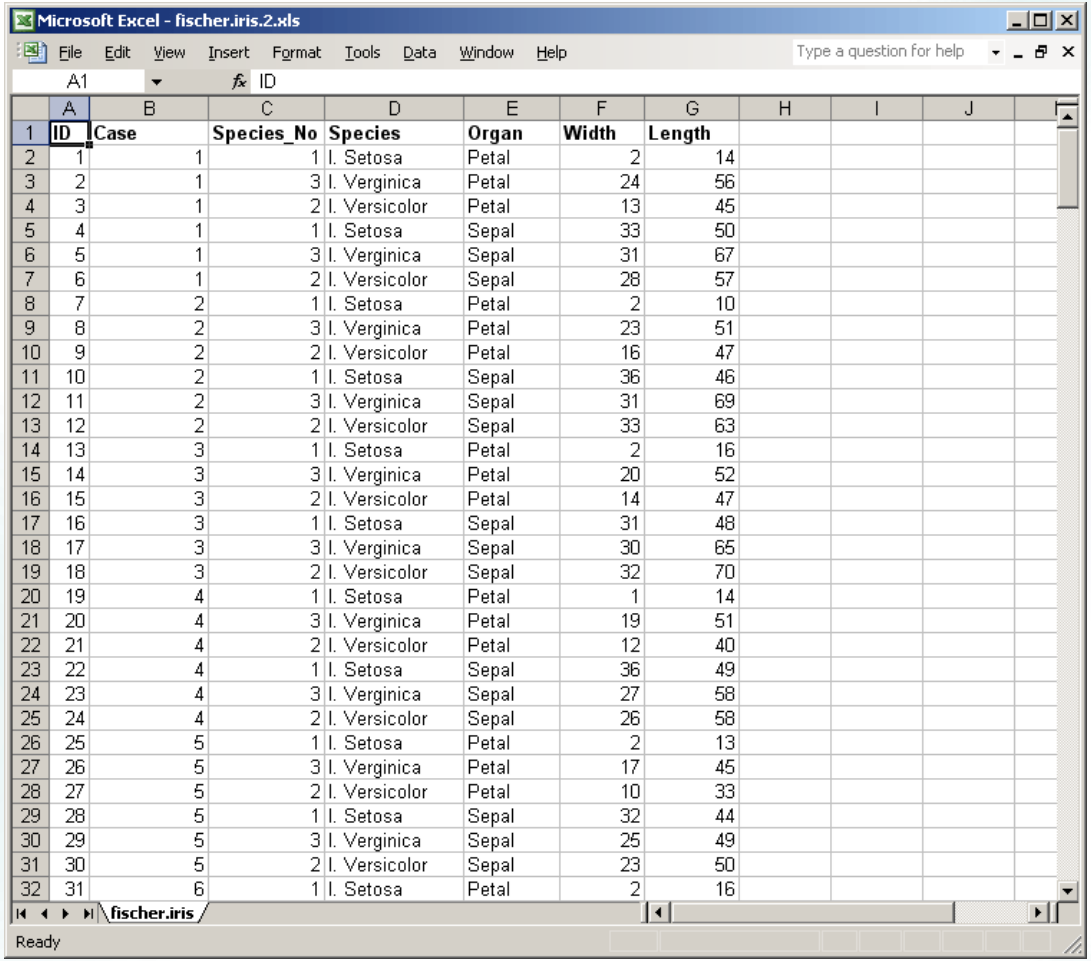
- Q – Ratio
  - Has order, equal distance and true zero point
  - Ex: count of items
  - Operations: All mathematical operations
  - Can measure ratios or proportions
  - Common visualization: bar chart, line chart, scatter plots





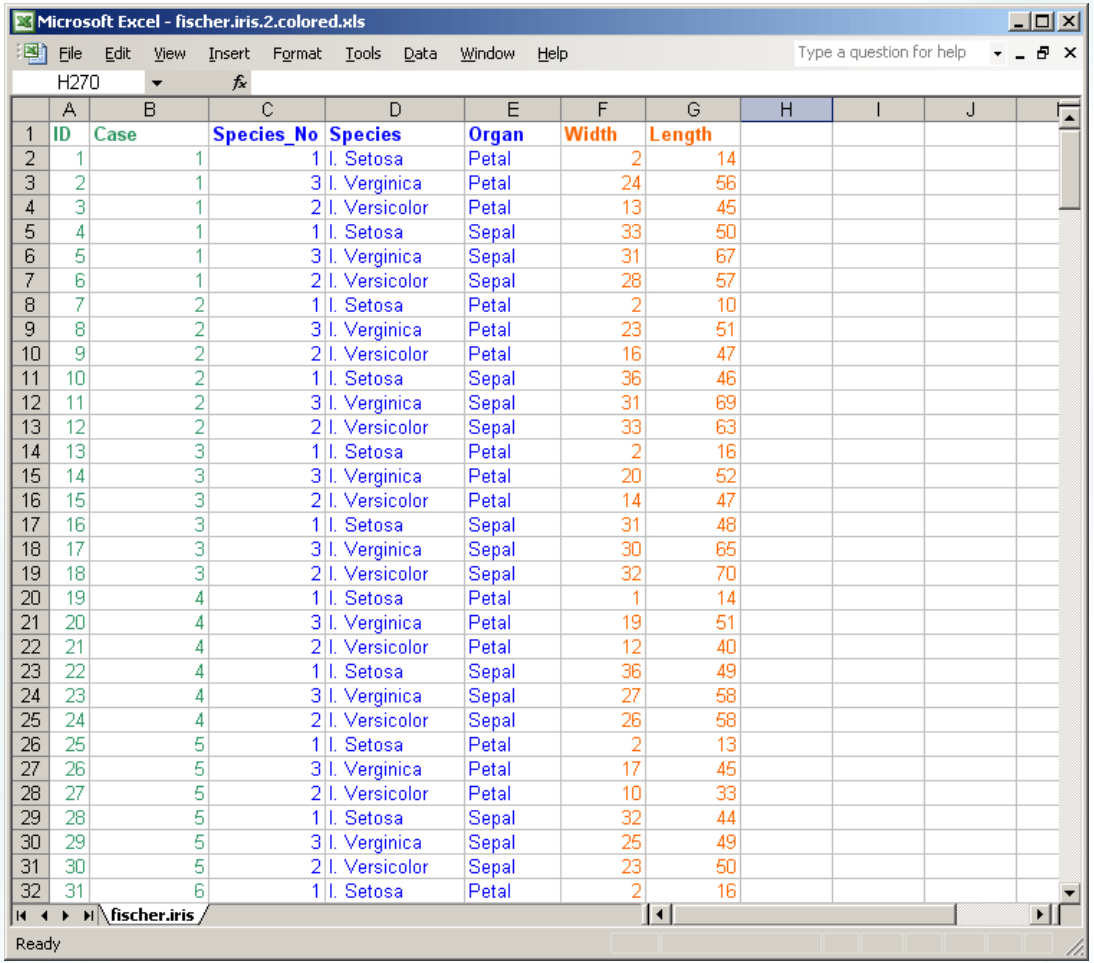
From Data Model to N, O, Q

- Data Model
  - 32.5, 54.0, -17.3, …
  - Floating point numbers
- Conceptual Model
  - Temperature (°C)
- Data Type
  - Burned vs. Not-Burned (**N**)
  - Hot, Warm, Cold (**O**)
  - Temperature Value (**Q**)



ID	Case	Species_No	Species	Organ	Width	Length
1	1	1	I. Setosa	Petal	2	14
2	1	3	I. Verginica	Petal	24	56
3	1	2	I. Versicolor	Petal	13	45
4	1	1	I. Setosa	Sepal	33	50
5	1	3	I. Verginica	Sepal	31	67
6	1	2	I. Versicolor	Sepal	28	57
7	2	1	I. Setosa	Petal	2	10
8	2	3	I. Verginica	Petal	23	51
9	2	2	I. Versicolor	Petal	16	47
10	2	1	I. Setosa	Sepal	36	46
11	2	3	I. Verginica	Sepal	31	69
12	2	2	I. Versicolor	Sepal	33	63
13	3	1	I. Setosa	Petal	2	16
14	3	3	I. Verginica	Petal	20	52
15	3	2	I. Versicolor	Petal	14	47
16	3	1	I. Setosa	Sepal	31	48
17	3	3	I. Verginica	Sepal	30	65
18	3	2	I. Versicolor	Sepal	32	70
19	4	1	I. Setosa	Petal	1	14
20	4	3	I. Verginica	Petal	19	51
21	4	2	I. Versicolor	Petal	12	40
22	4	1	I. Setosa	Sepal	36	49
23	4	3	I. Verginica	Sepal	27	58
24	4	2	I. Versicolor	Sepal	26	58
25	5	1	I. Setosa	Petal	2	13
26	5	3	I. Verginica	Petal	17	45
27	5	2	I. Versicolor	Petal	10	33
28	5	1	I. Setosa	Sepal	32	44
29	5	3	I. Verginica	Sepal	25	49
30	5	2	I. Versicolor	Sepal	23	50
31	6	1	I. Setosa	Petal	2	16

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



ID	Case	Species_No	Species	Organ	Width	Length
1	1	1	I. Setosa	Petal	2	14
2	1	3	I. Verginica	Petal	24	56
3	1	2	I. Versicolor	Petal	13	45
4	1	1	I. Setosa	Sepal	33	50
5	1	3	I. Verginica	Sepal	31	67
6	1	2	I. Versicolor	Sepal	28	57
7	2	1	I. Setosa	Petal	2	10
8	2	3	I. Verginica	Petal	23	51
9	2	2	I. Versicolor	Petal	16	47
10	2	1	I. Setosa	Sepal	36	46
11	2	3	I. Verginica	Sepal	31	69
12	2	2	I. Versicolor	Sepal	33	63
13	3	1	I. Setosa	Petal	2	16
14	3	3	I. Verginica	Petal	20	52
15	3	2	I. Versicolor	Petal	14	47
16	3	1	I. Setosa	Sepal	31	48
17	3	3	I. Verginica	Sepal	30	65
18	3	2	I. Versicolor	Sepal	32	70
19	4	1	I. Setosa	Petal	1	14
20	4	3	I. Verginica	Petal	19	51
21	4	2	I. Versicolor	Petal	12	40
22	4	1	I. Setosa	Sepal	36	49
23	4	3	I. Verginica	Sepal	27	58
24	4	2	I. Versicolor	Sepal	26	58
25	5	1	I. Setosa	Petal	2	13
26	5	3	I. Verginica	Petal	17	45
27	5	2	I. Versicolor	Petal	10	33
28	5	1	I. Setosa	Sepal	32	44
29	5	3	I. Verginica	Sepal	25	49
30	5	2	I. Versicolor	Sepal	23	50
31	6	1	I. Setosa	Petal	2	16

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

Dimensions & Measures

- Dimensions (~ independent variables)
  - Discrete variables describing data (**N**, **O**)
  - Categories, dates, binned quantities
- Measures (~ dependent variables)
  - Data values that can be aggregated (**Q**)
  - Numbers to be analyzed
  - *Aggregate as sum, count, avg, std. dev…*

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, …

U.S. Census Data

- People Count
- Year
- Age
- Sex
- Marital Status

	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 Data Demo: https://www.census.gov/dataviz/

Census: N, O, Q?

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

Census: Dimension or Measure?

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

Data Transformation

Relational Data Model

- Represent data as a table *(relation)*
- Each **row (*tuple*)** *represents a record*
  - Each record is a fixed-length tuple
- Each **column (*attribute*)** *represents a variable*
  - Each attribute has a *name and a data type*
- A **table's schema** is the set of names and types
- A **database** is a collection of tables (relations)

Relational Algebra [Codd '70] / SQL

- **Operations on Data Tables**: table(s) in, table out
  - Projection (`select`) - selects columns
  - Selection (`where`) - filters rows
  - Sorting (`order by`)
  - Aggregation (`group by, sum, min, max, ...`)
    - partition rows into groups + summarize
  - Combine relations (`union, join,` ...)
    - integrate data from multiple tables

Relational Algebra [Codd '70] / SQL

- Project (`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

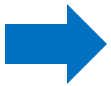
day	stock
10/3	AMZN
10/3	MSFT
10/4	AMZN
10/4	MSFT

Relational Algebra [Codd '70] / SQL

- Filter (where): remove unwanted 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

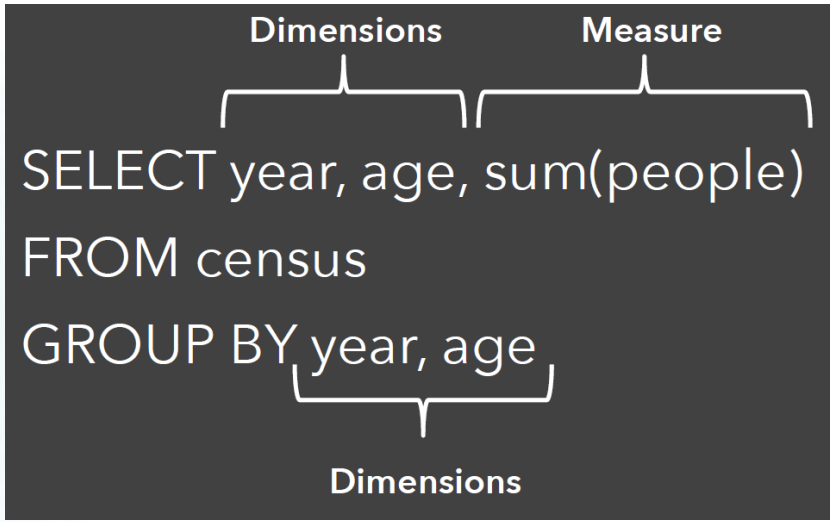
- Aggregate (group by, sum, min, max, …):
  select stock, min(price) group by stock

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

Roll-Up and Drill-Down

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

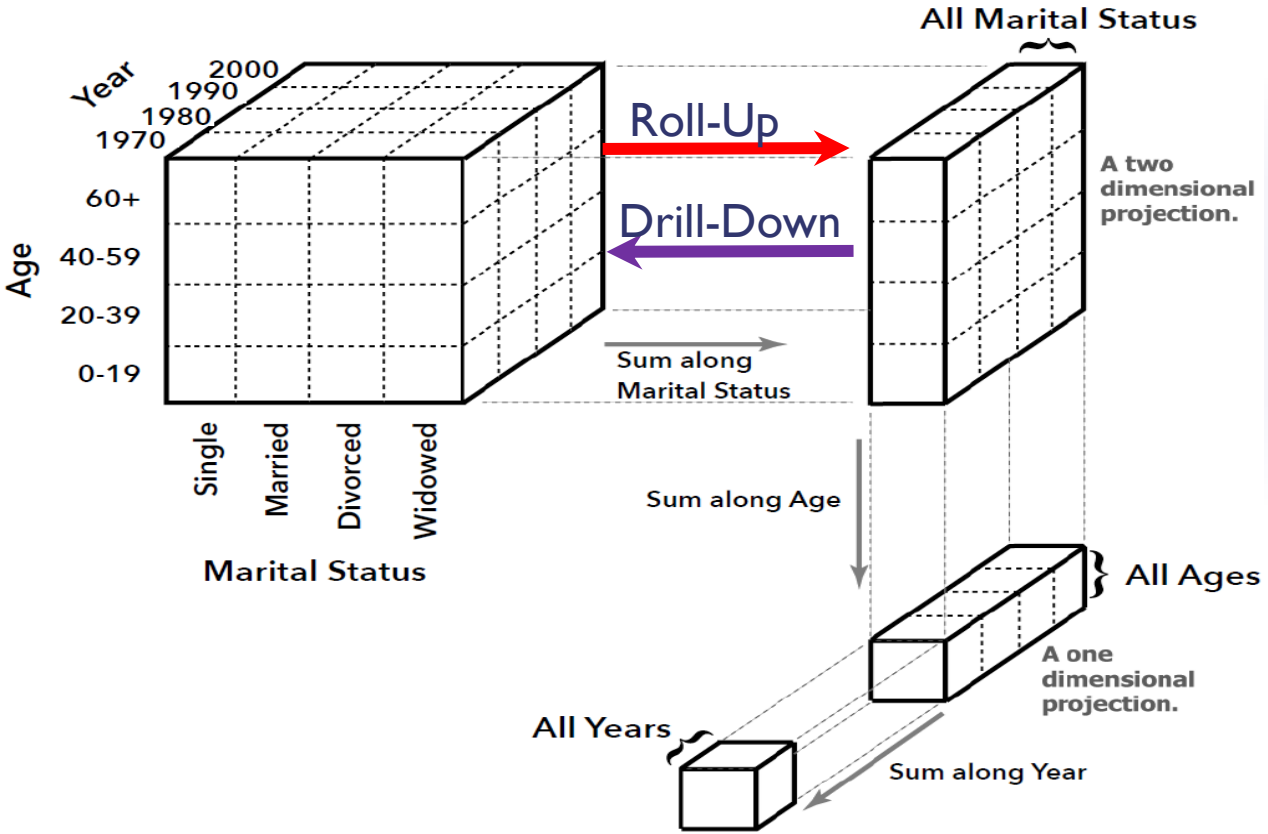


Roll-Up and Drill-Down

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

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

Roll-Up and Drill-Down



Roll-Up and Drill-Down

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?

Tidy Data [Wickham 2014]

- How do rows, columns, and tables match up with observations, variables, and types?
- In **"tidy" data**:
  - Each variable forms a **column**.
  - Each observation forms a **row**.
  - Each type of observational unit forms a **table**.
- The advantage is that this provides a flexible starting point for analysis, transformation, and visualization

Image

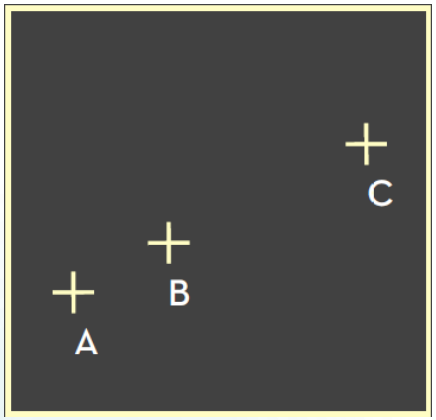
Visual Language is a Sign System

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



- A, B, C are distinguishable
- B is between A and C.
- BC is twice as long as AB.

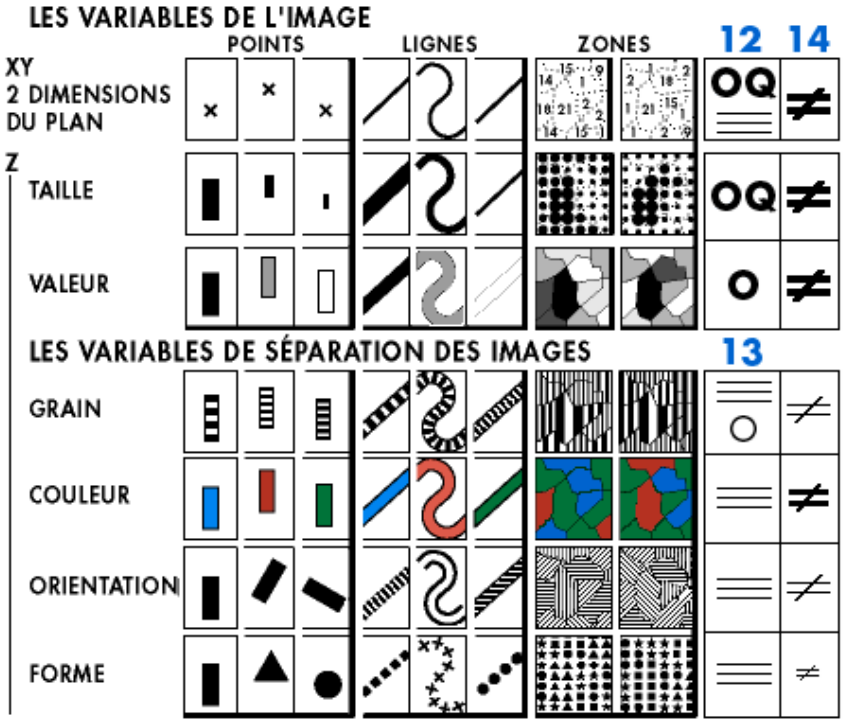
∴ Encode quantitative variables

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

Visual Encoding Variables

Position (x 2)
Size
Value

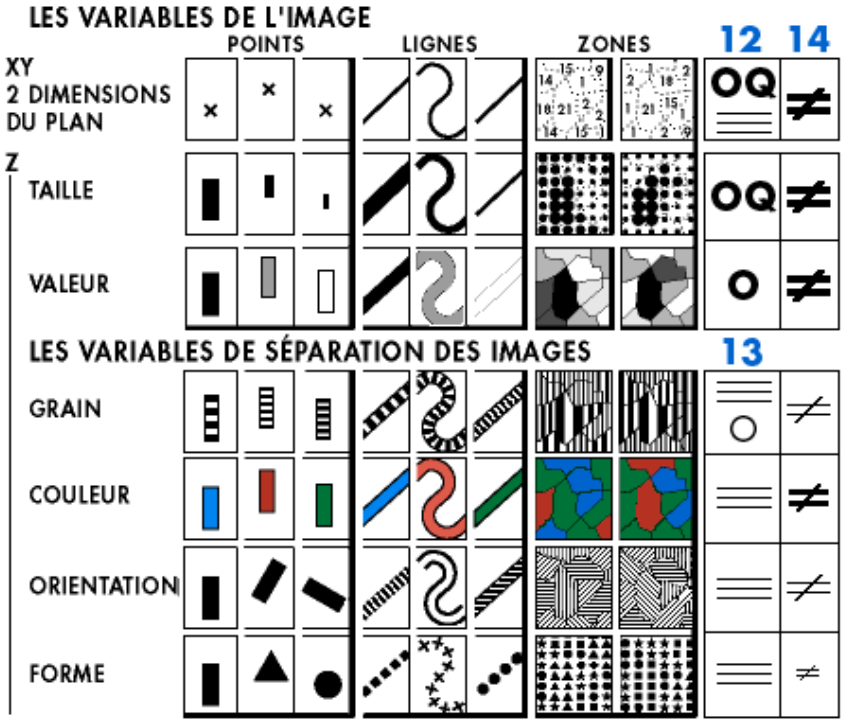
Texture
Color
Orientation
Shape



[Bertin, Graphics and Graphic Information Processing 1981]

Visual Encoding Variables

Position (x 2)
Length
Area
Volume
Size
Value
Texture
Color
Orientation
Shape
Transparency
Blur / Focus …



[Bertin, Graphics and Graphic Information Processing 1981]

Information in Hue and Value

Value is perceived as ordered

∴ Encode ordinal variables (O)

∴ Encode continuous variables (Q) [not as well]

Hue (Color) is normally perceived as unordered

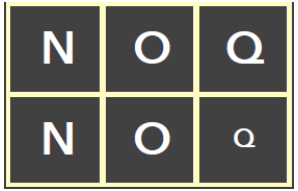
∴ Encode nominal variables (N) using color

Bertin's "Levels of Organization"

Position

Size

Value

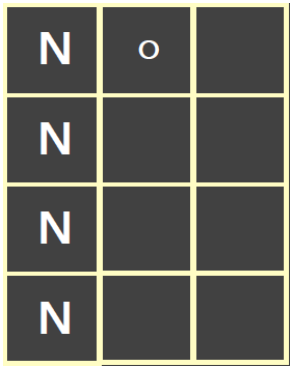


Texture

Color

Orientation

Shape



Nominal

Ordinal

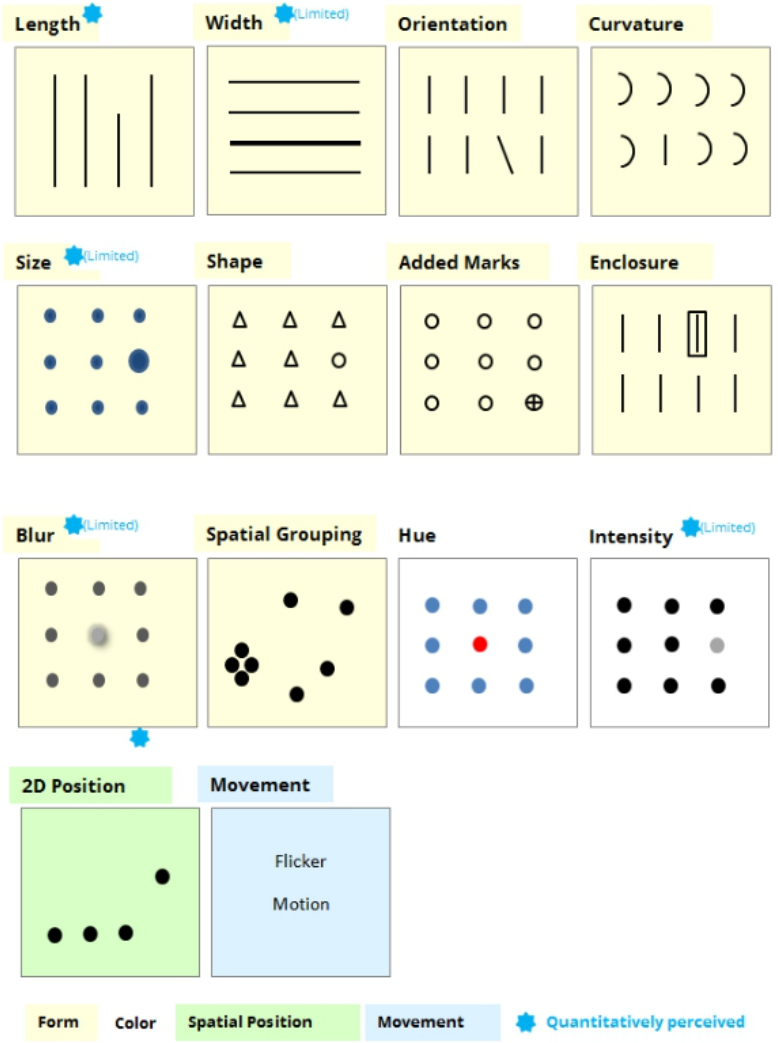
Quantitative

Note: **Q** ⊂ **O** ⊂ **N**

Sample Encodings

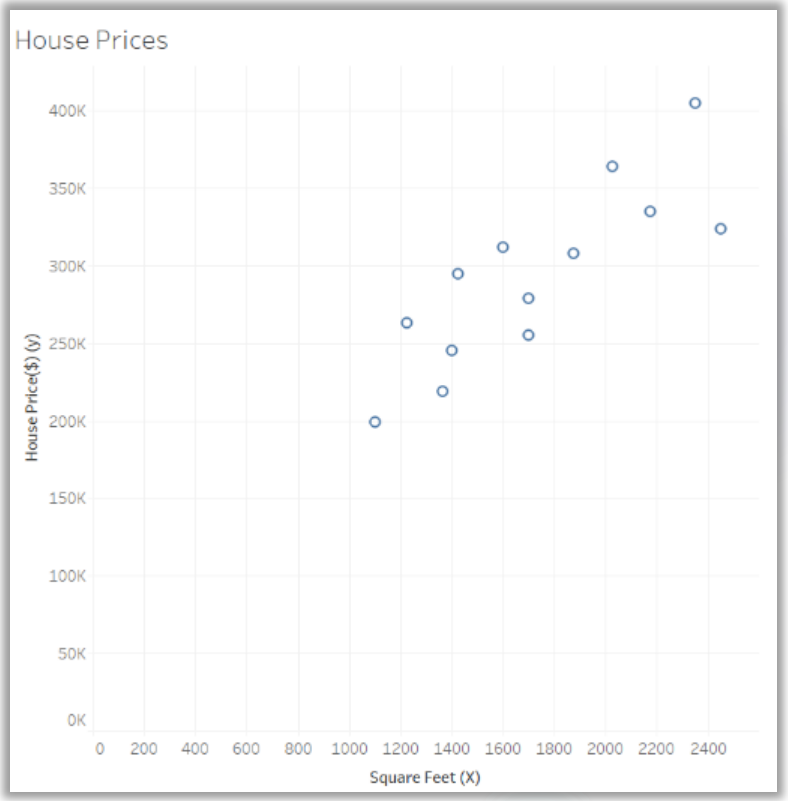
Preattentive

attributes determine what information catches our attention. This is important in visualization because it enables us to direct our viewer's attention towards the most important information in our visual.



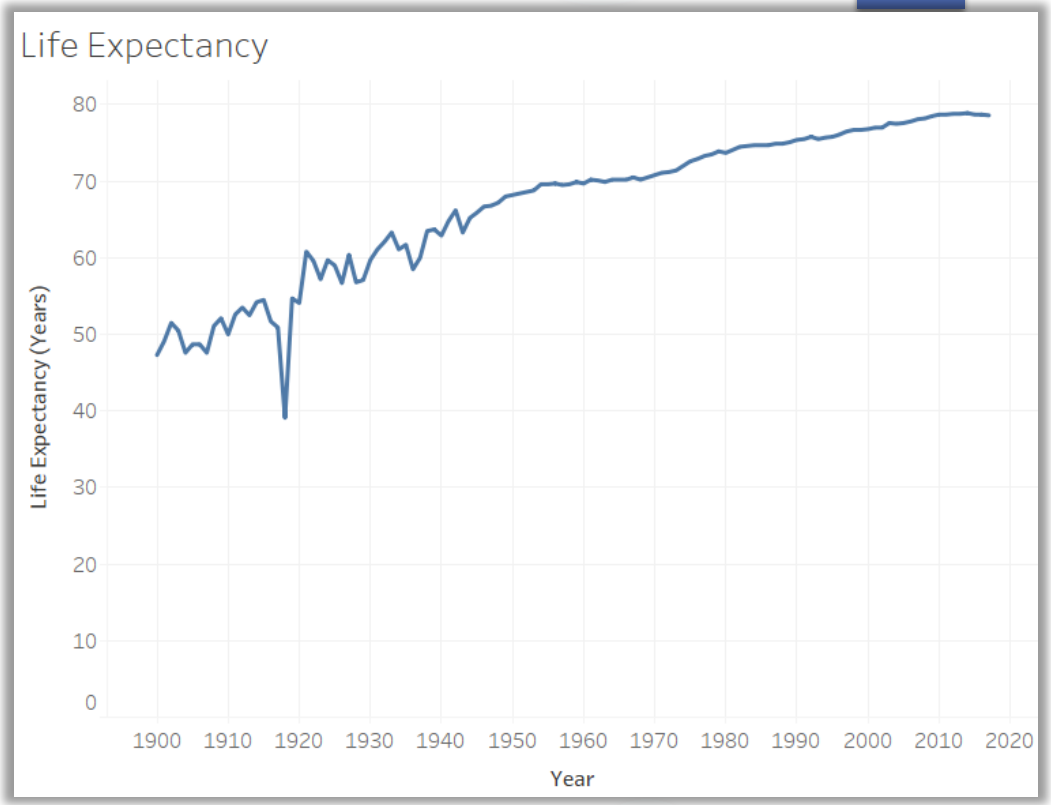
Marks to Encode Quantitative Values

- **Points**
- Lines
- Bars
- Boxes
- Shapes with 2-D areas
- Shapes with color intensity



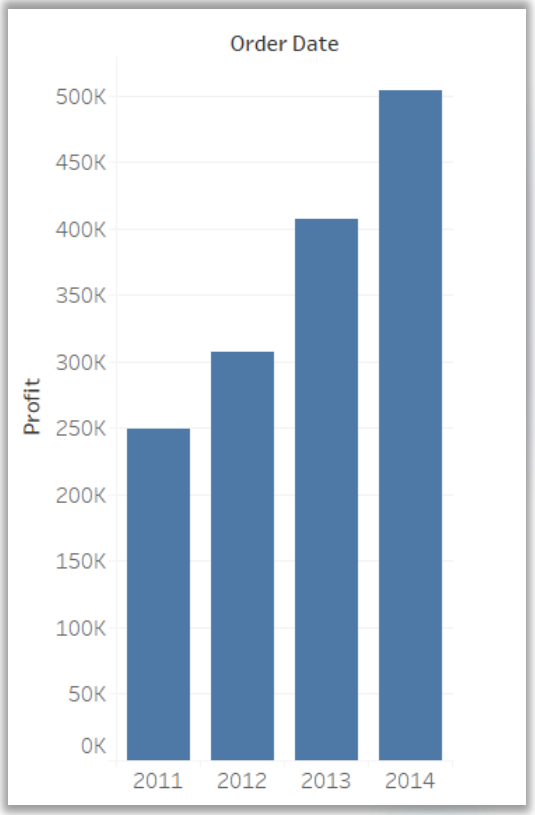
Marks to Encode Quantitative Values

- Points
- **Lines**
- Bars
- Boxes
- Shapes with 2-D areas
- Shapes with color intensity



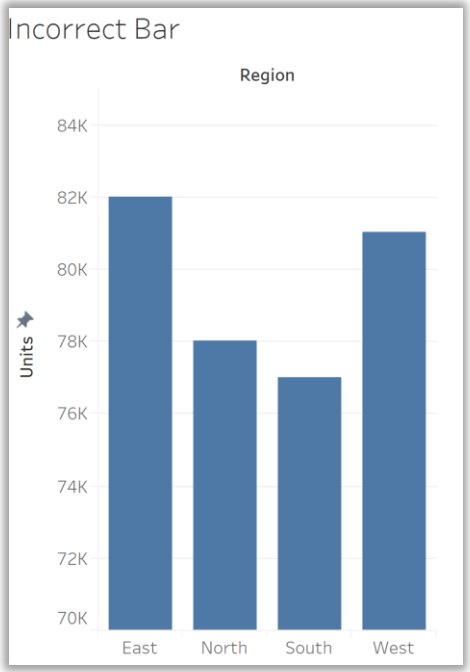
Marks to Encode Quantitative Values

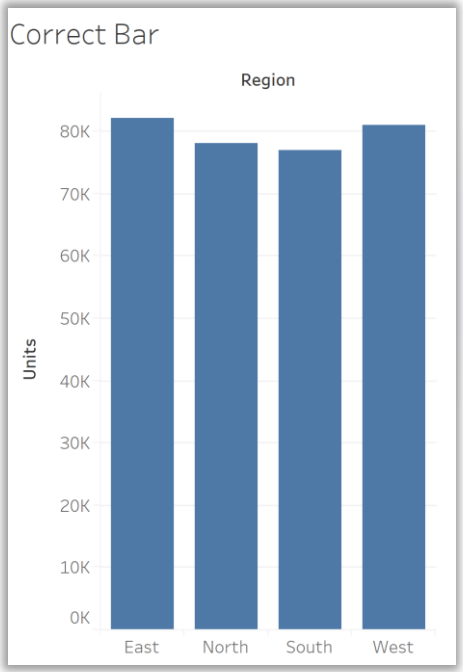
- Points
- Lines
- **Bars**
- Boxes
- Shapes with 2-D areas
- Shapes with color intensity



Marks to Encode Quantitative Values

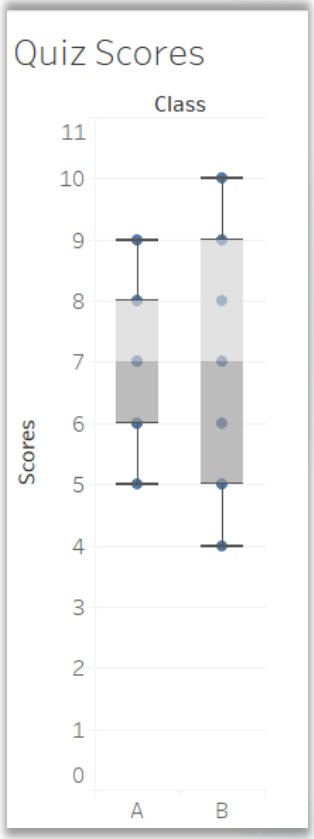
- Points
- Lines
- **Bars**
- Boxes
- Shapes with 2-D areas
- Shapes with color intensity





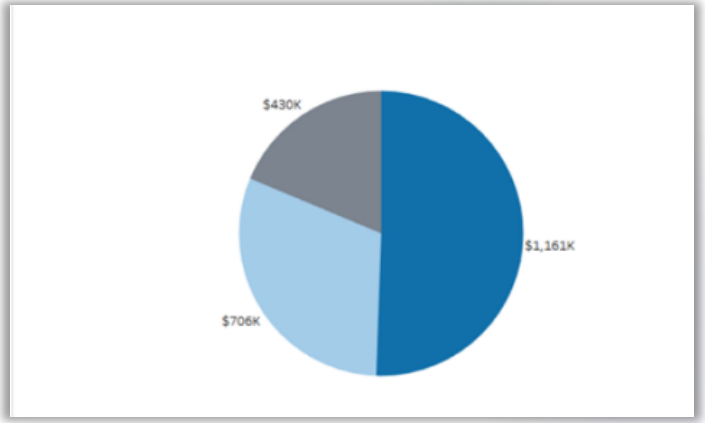
Marks to Encode Quantitative Values

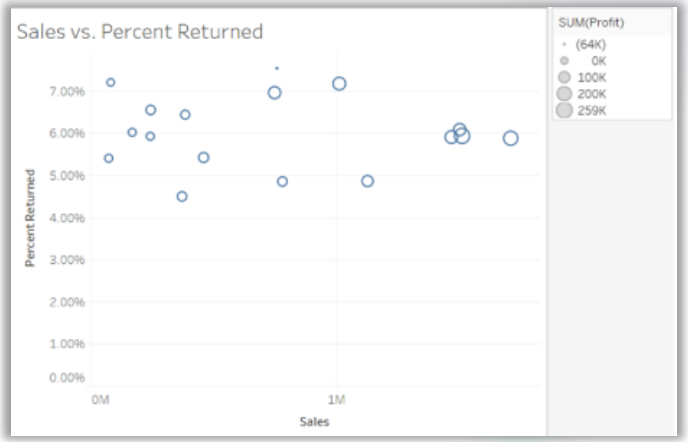
- Points
- Lines
- Bars
- **Boxes**
- Shapes with 2-D areas
- Shapes with color intensity



Marks to Encode Quantitative Values

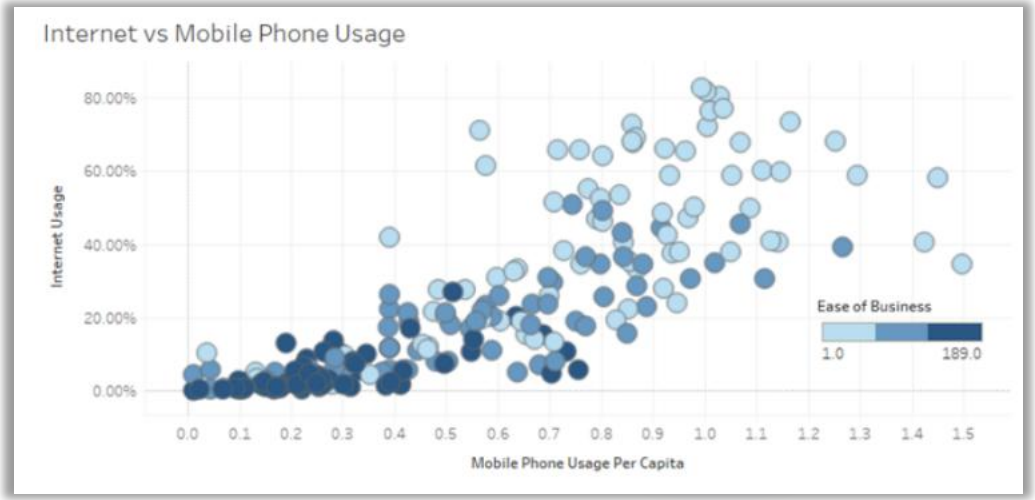
- Points
- Lines
- Bars
- Boxes
- **Shapes with 2-D areas** (use to add third or fourth set of values)
- Shapes with color intensity





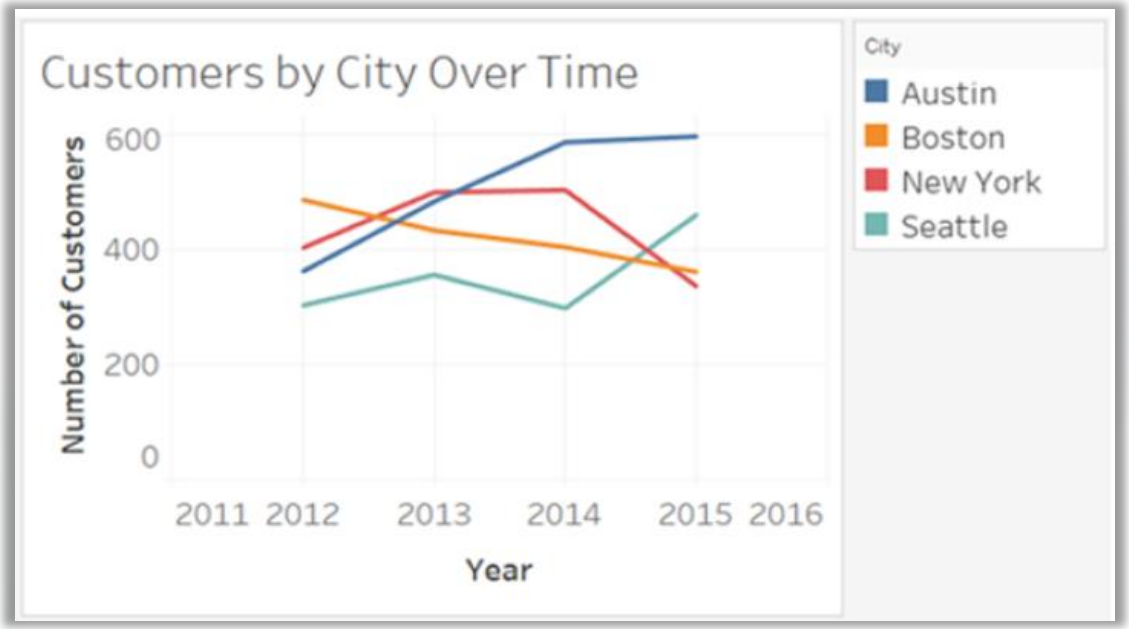
Marks to Encode Quantitative Values

- Points
- Lines
- Bars
- Boxes
- Shapes with 2-D areas
- **Shapes with color intensity**
  (use to add third or fourth set of values)



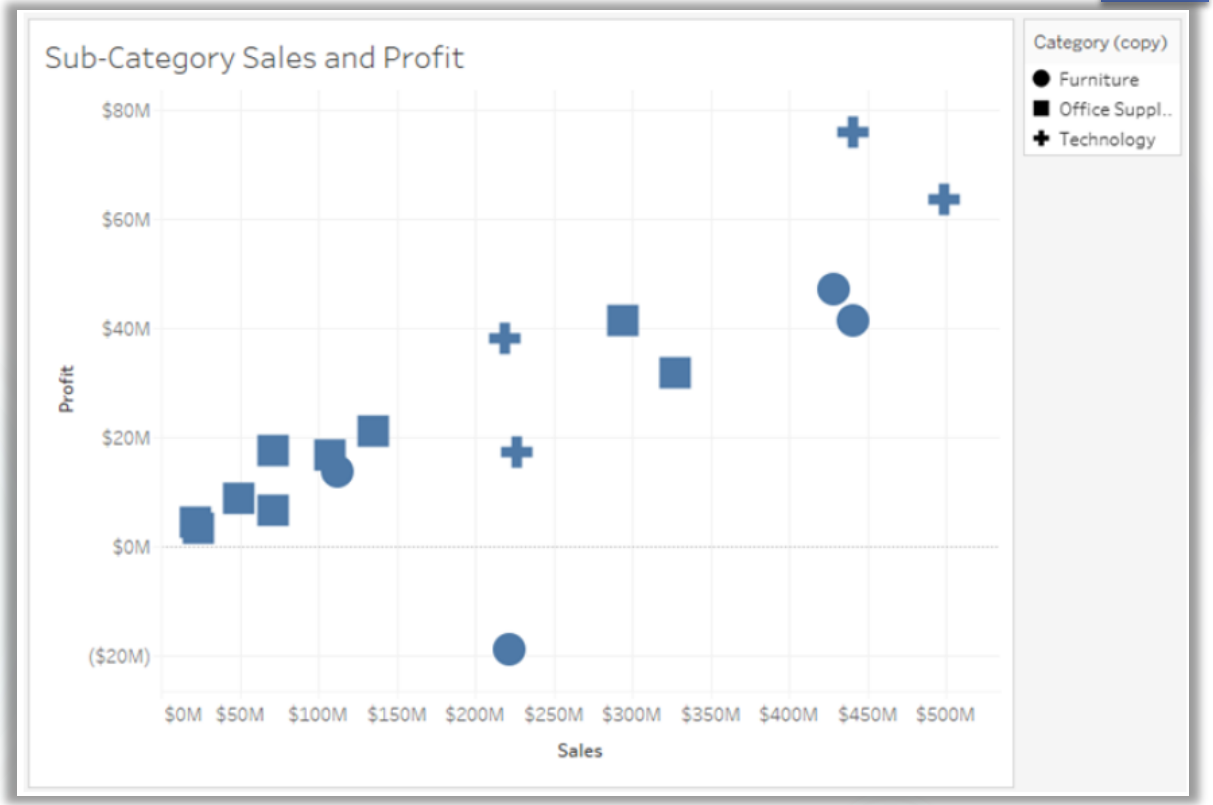
Encoding Nominal Items

- **Hue**
- Point shape
- 2-D Position



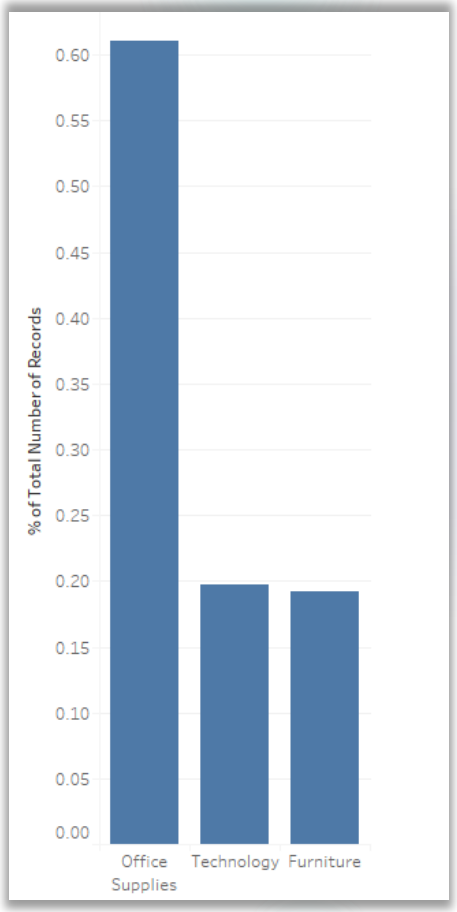
Encoding Nominal Items

- Hue
- **Point shape**
- 2-D Position



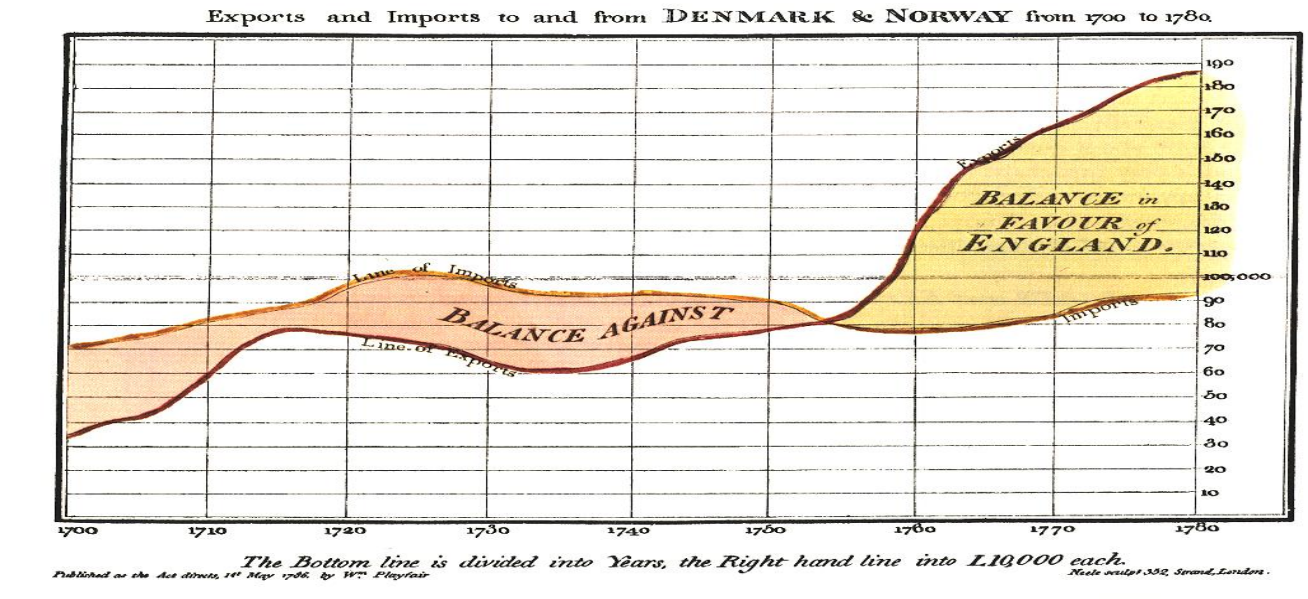
Encoding Nominal Items

- Hue
- Point shape
- **2-D Position**

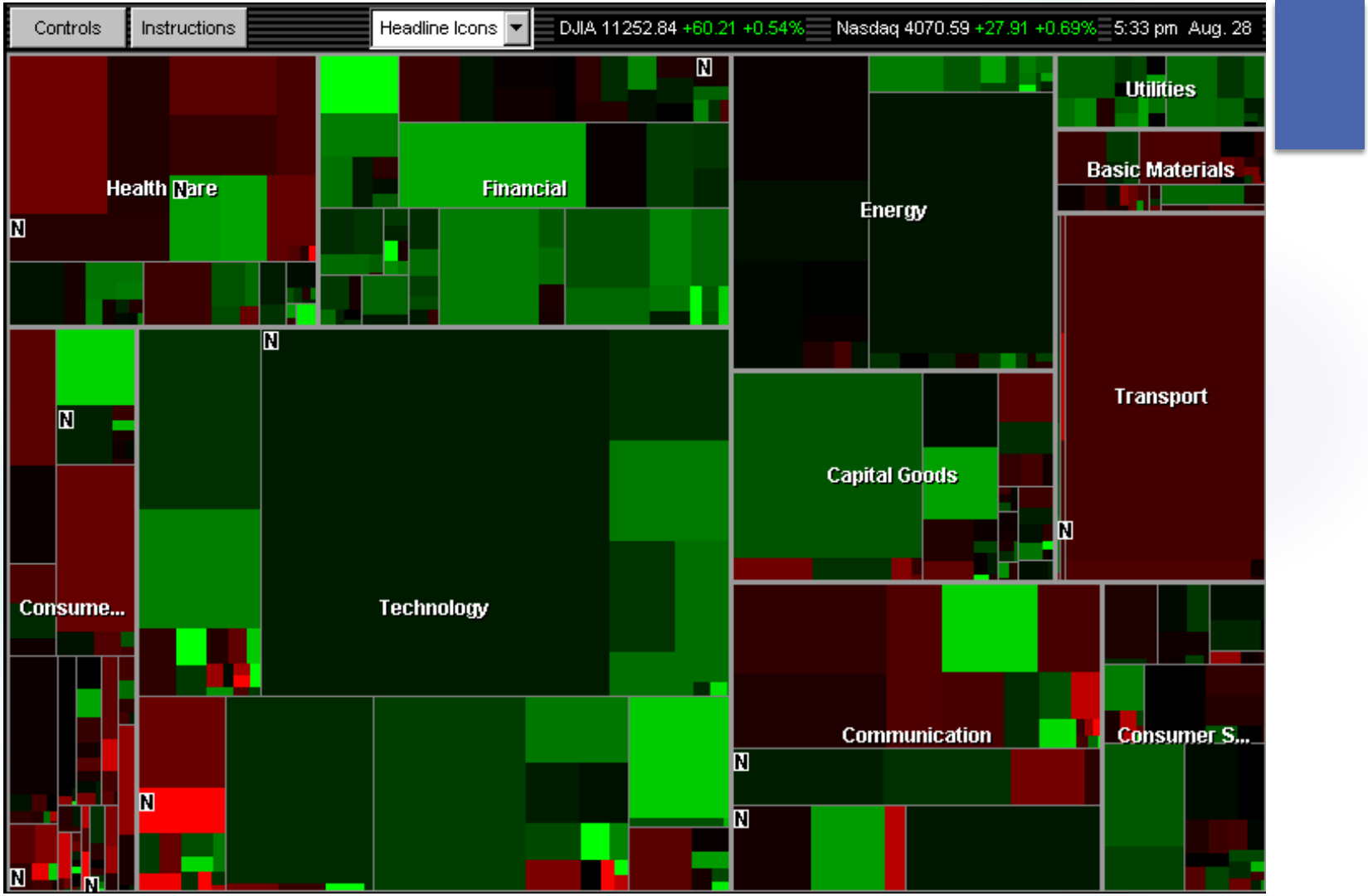


Deconstructions

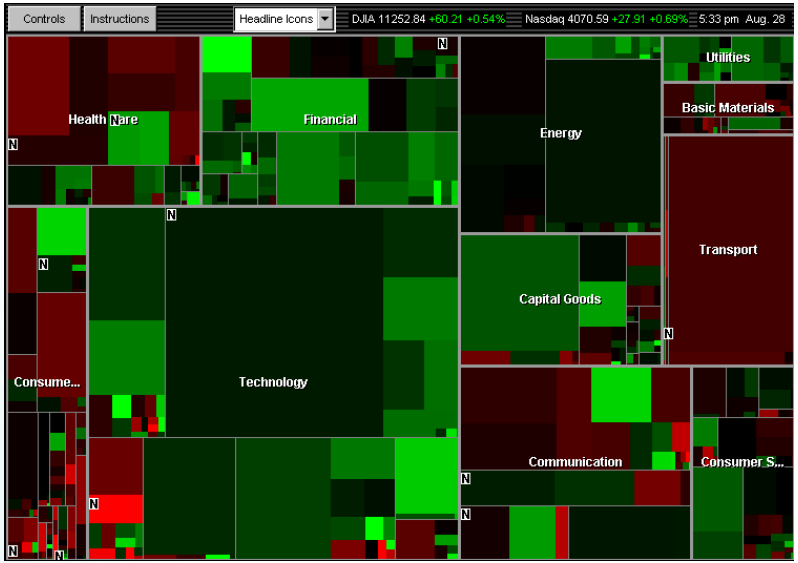
William Playfair, 1786



X-axis: year (Q)
Y-axis: currency (Q)
Color: imports/exports (N, O)

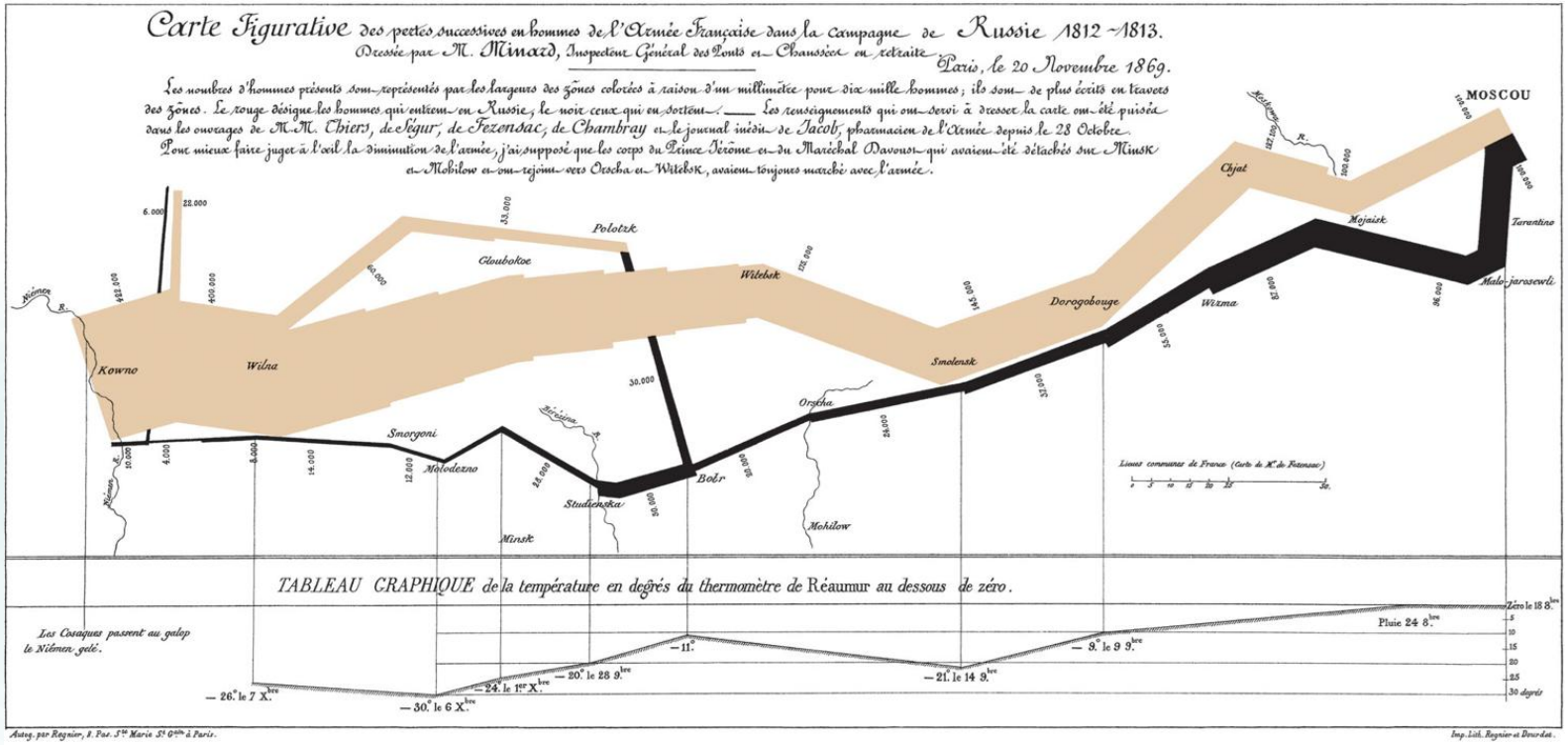

Controls
Instructions
Headline Icons
DJIA 11252.84 +60.21 +0.54%
Nasdaq 4070.59 +27.91 +0.69%
5:33 pm Aug. 28
Health Care
Financial
Energy
Utilities
Basic Materials
Transport
Capital Goods
Consume...
Technology
Communication
Consumer S...

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)

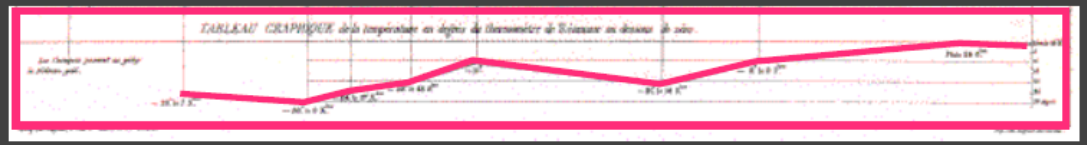
Minard 1869: Napoleon's March



Mark Composition

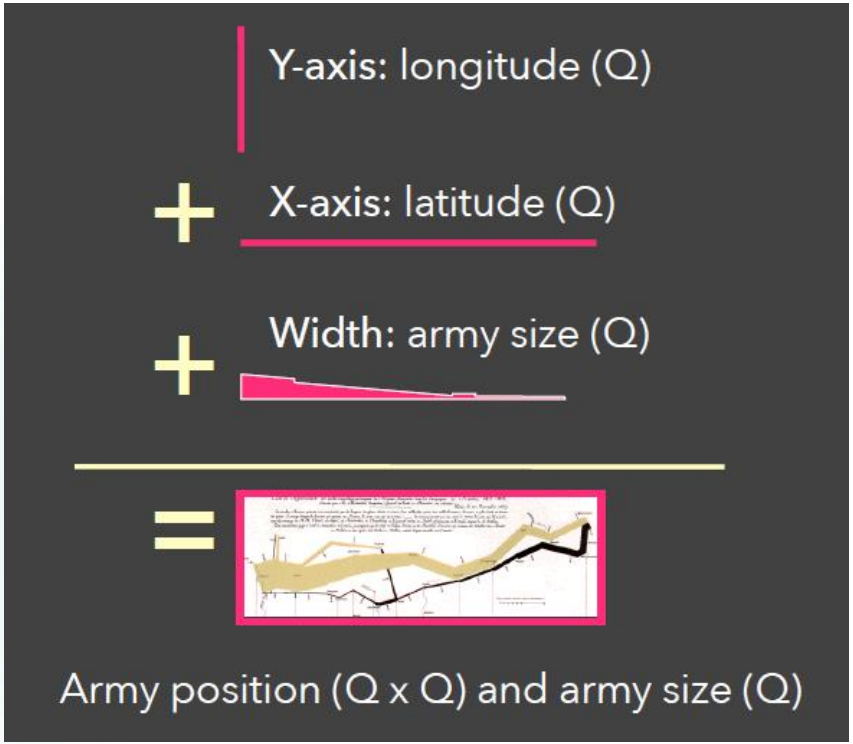
Y-axis: temperature (Q)

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

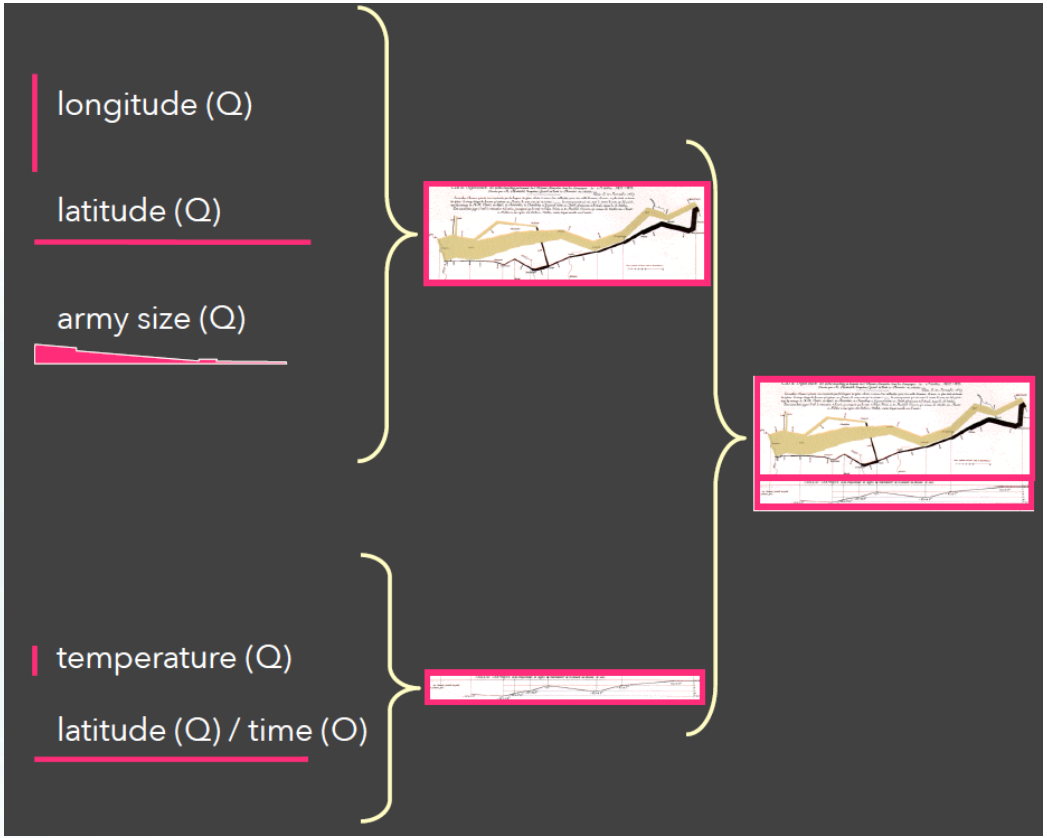
\= 

Temp over space/time (Q x Q)

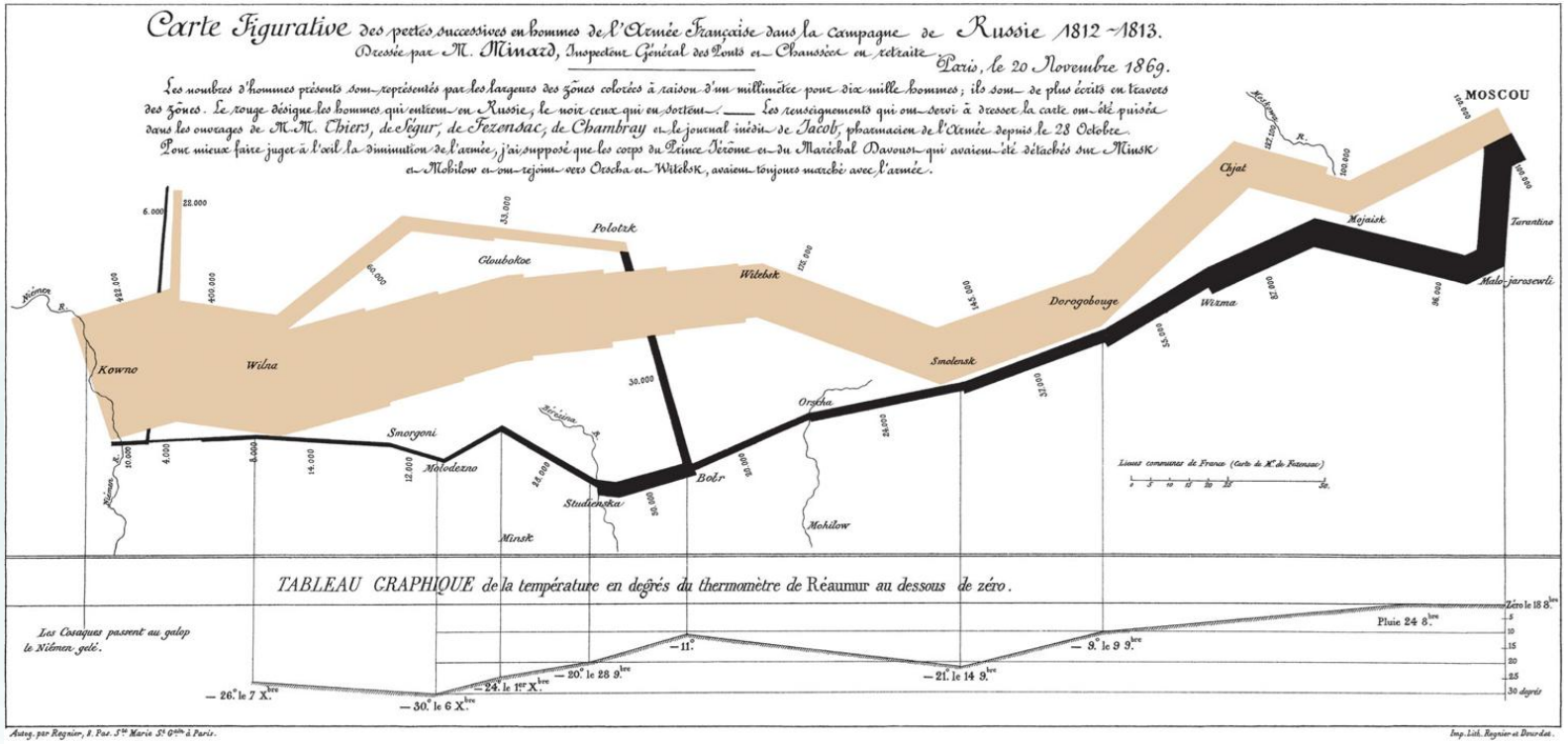
Mark Composition



Mark Composition



Minard 1869: Napoleon's March



Depicts at least 5 quantitative variables. Any others?

Formalizing Design

Choosing Visual Encodings

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

- **Principle of Consistency**
  - The properties of the image (visual variables) should match the properties of the data.
- **Principle of Importance Ordering**
  - Encode the most important information in the most effective way.

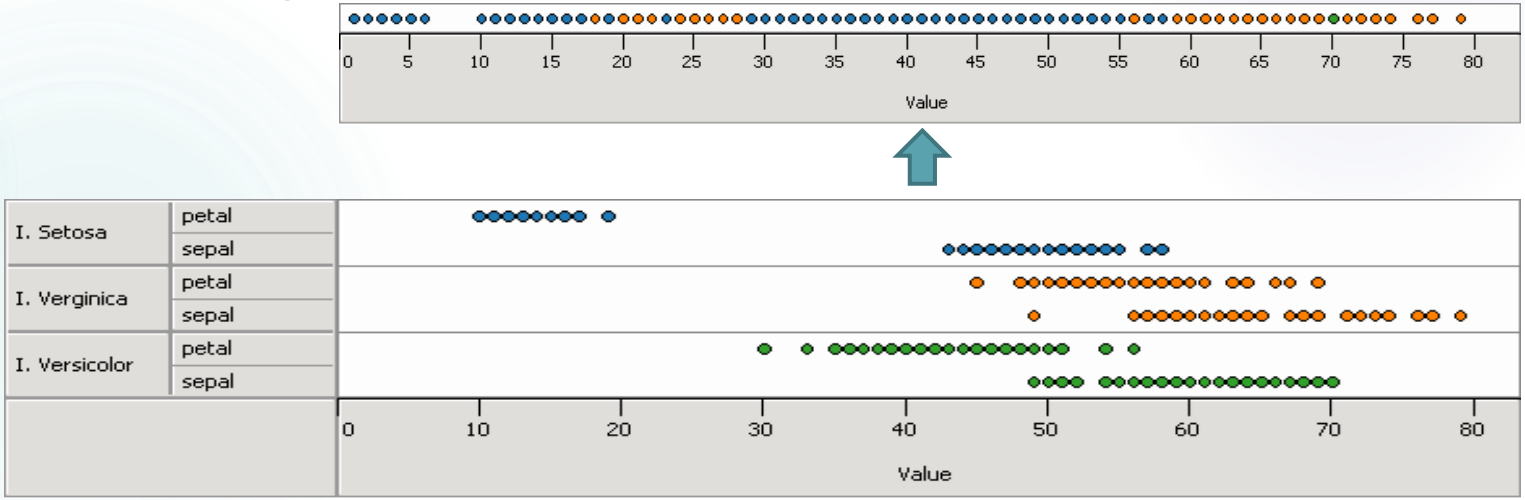
Design Criteria [Mackinlay 86]

- **Expressiveness**

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

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.



Design Criteria [Mackinlay 86]

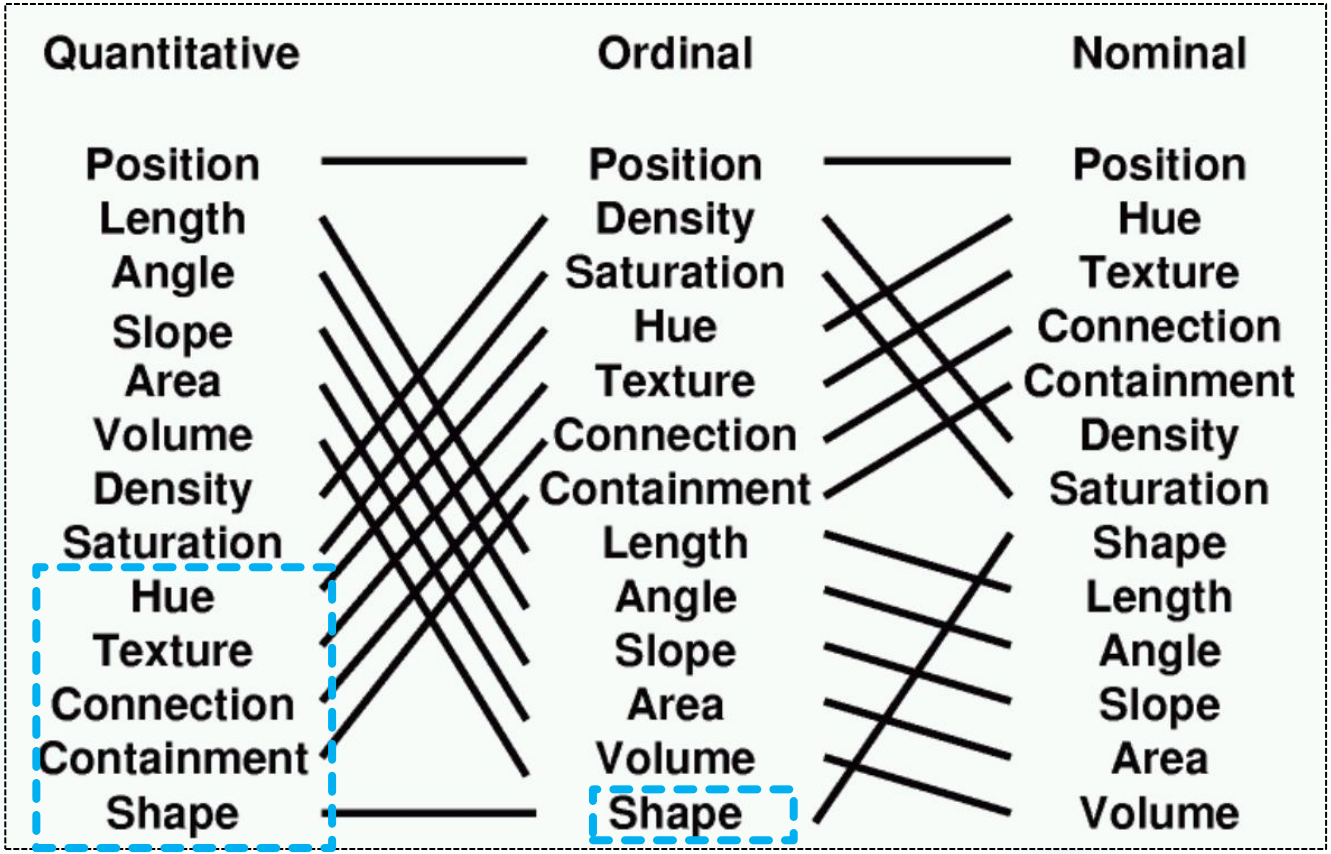
- **Expressiveness**

  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.

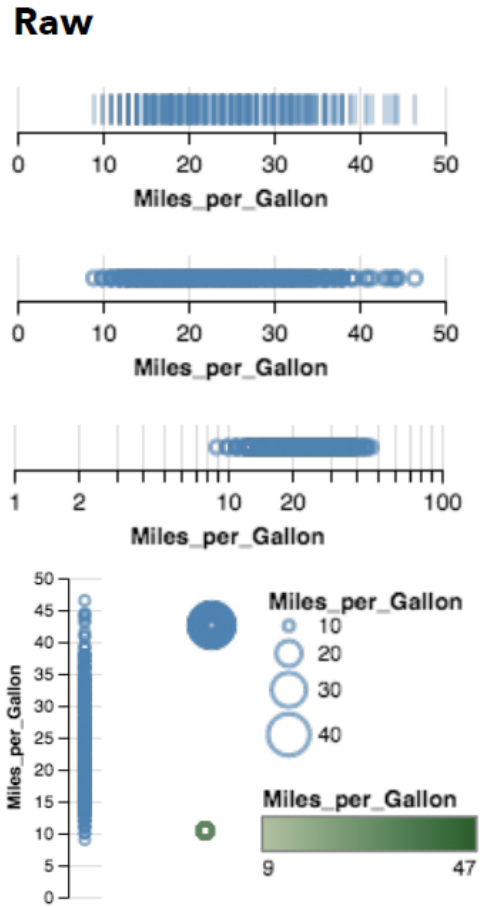
Mackinlay's Ranking

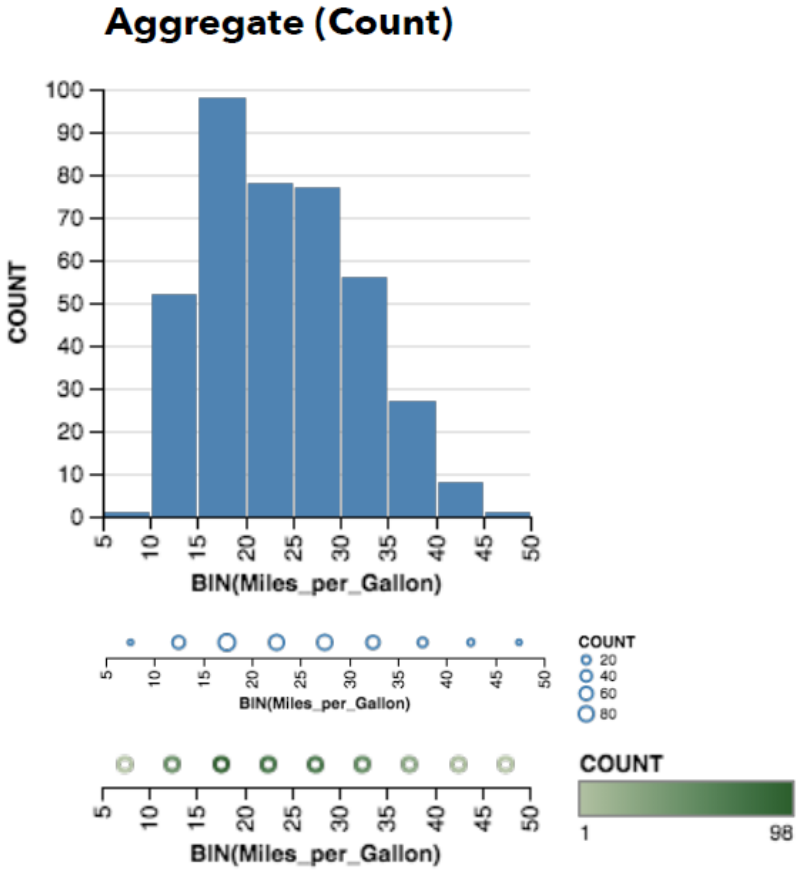


Mackinlay's Design Algorithm

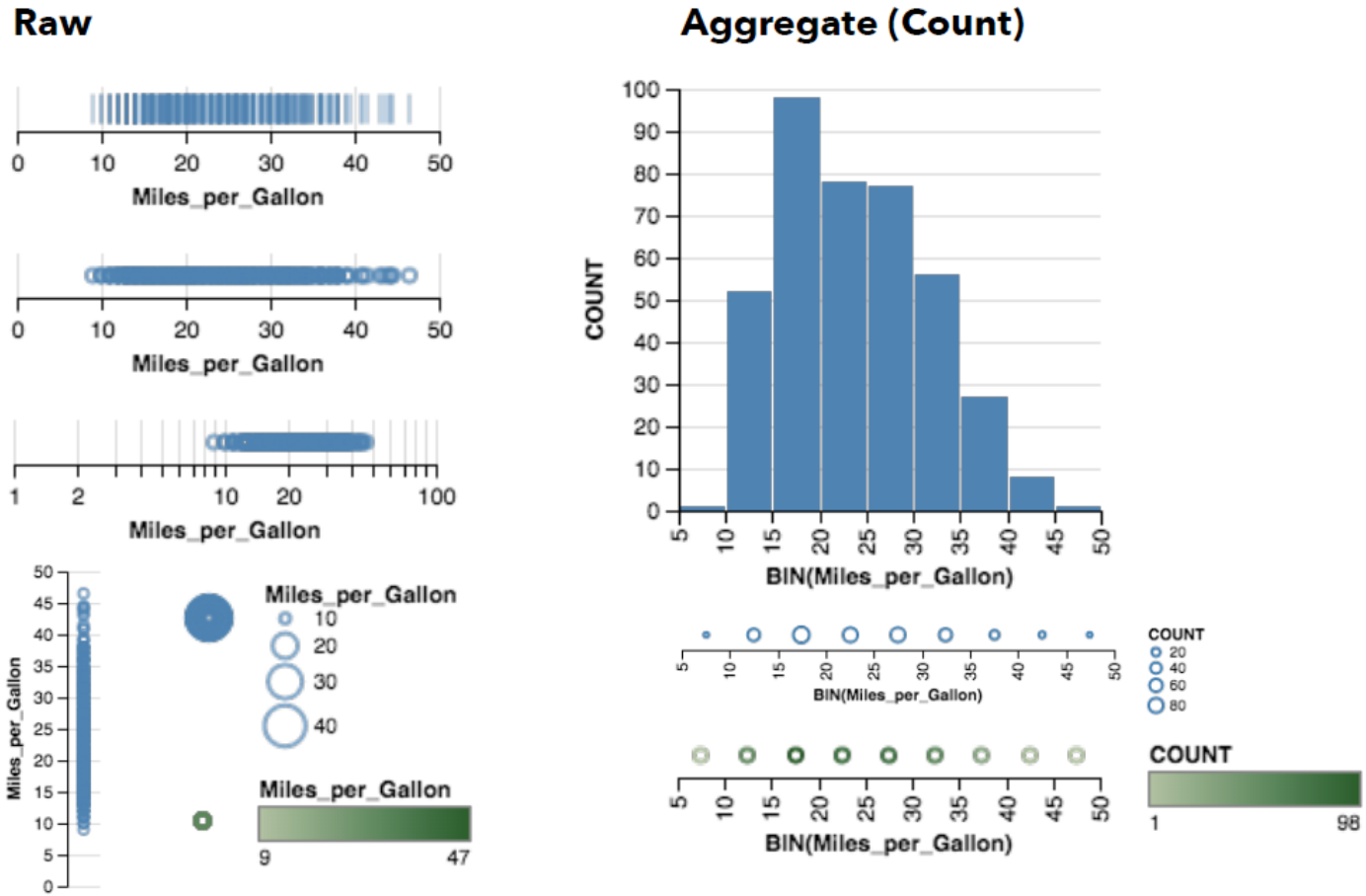
- **APT -** "**A P**resentation **T**ool", 1986
- User formally specifies data model and type
  - **Input**: ordered list of data variables to show
  - **Algorithm**:
    - APT **searches** over design space
    - **Test** expressiveness of each visual encoding
    - **Generate encodings** that pass test Rank by perceptual effectiveness criteria
  - **Output:** the "most effective" visualization

1D Quantitative

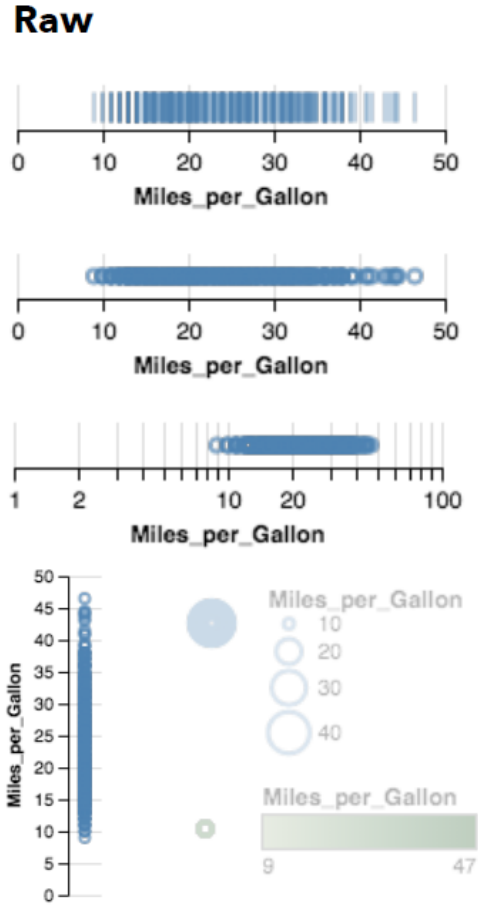


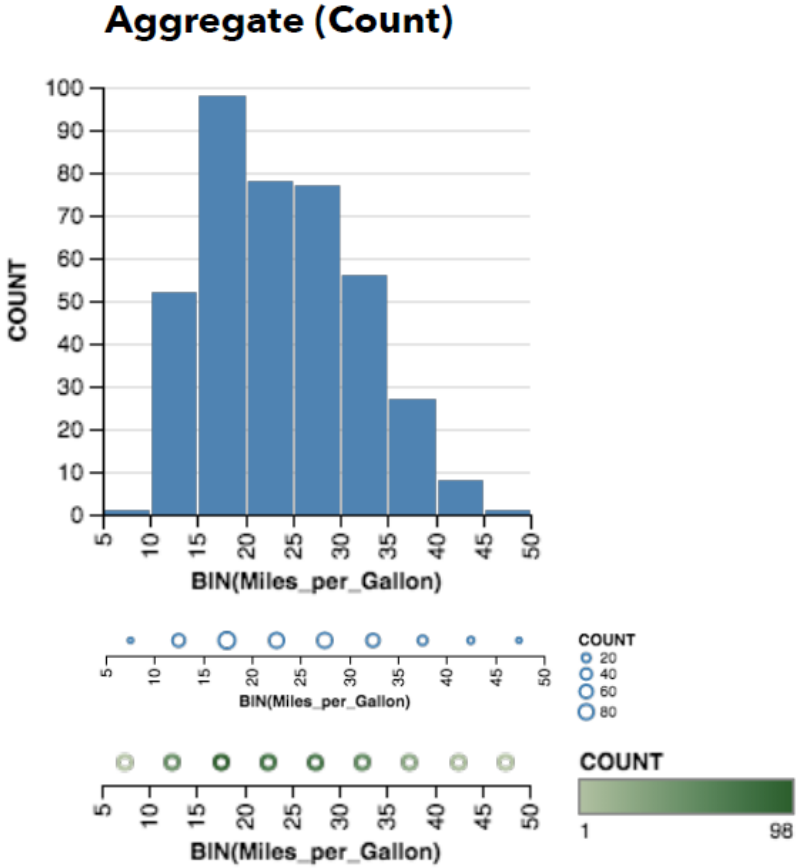


Expressive?

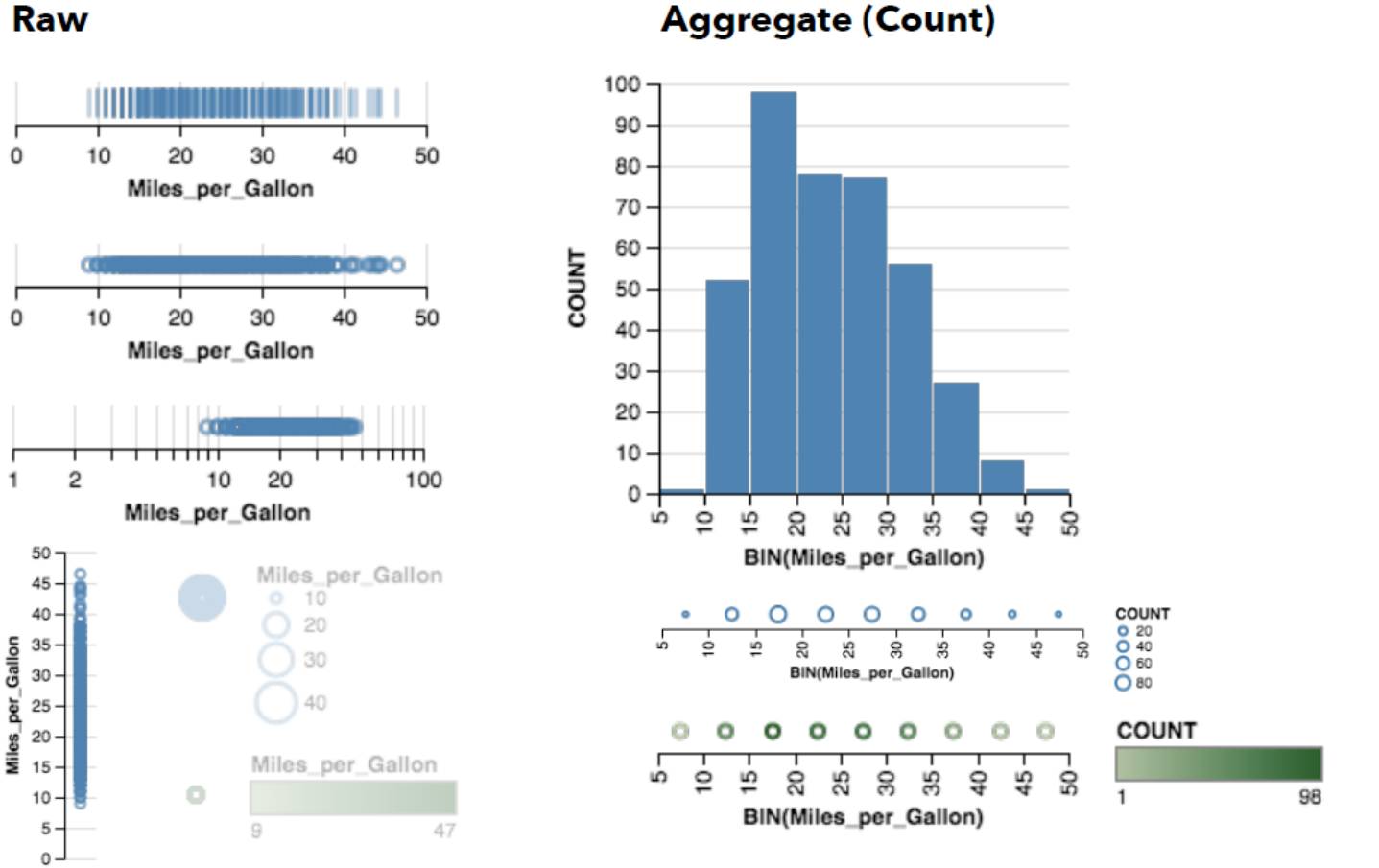

Raw
Aggregate (Count)
Miles_per_Gallon
BIN(Miles_per_Gallon)
COUNT
Miles_per_Gallon
10
20
30
40
Miles_per_Gallon
9
47
COUNT
20
40
60
80
COUNT
1
98

Expressive?

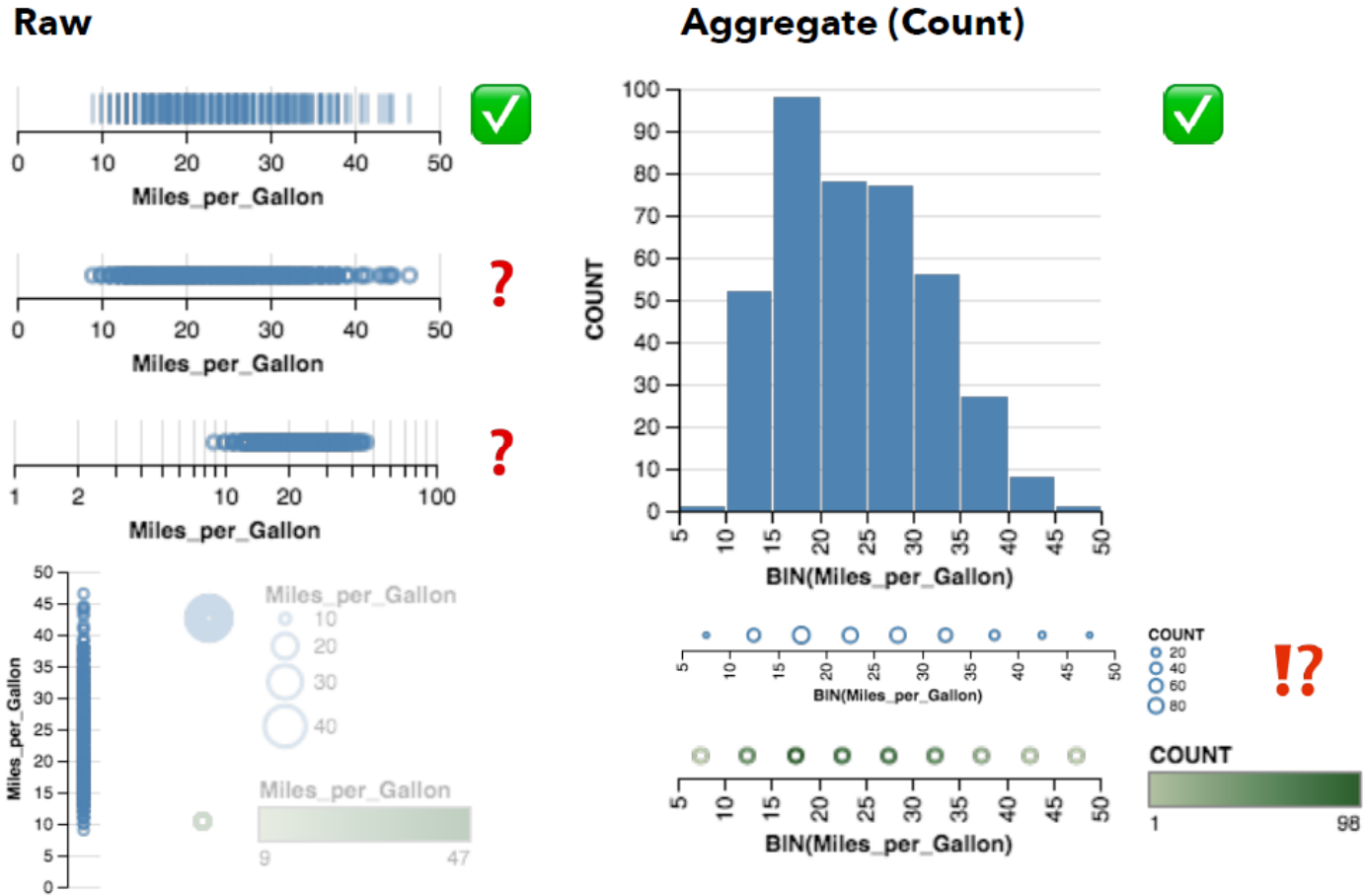

Raw
0
10
20
30
40
50
Miles_per_Gallon
0
10
20
30
40
50
Miles_per_Gallon
1
2
10
20
100
Miles_per_Gallon
Miles_per_Gallon
50
45
40
35
30
25
20
15
10
5
0
Miles_per_Gallon
10
20
30
40
Miles_per_Gallon
9
47


Aggregate (Count)
COUNT
100
90
80
70
60
50
40
30
20
10
0
5
10
15
20
25
30
35
40
45
50
BIN(Miles_per_Gallon)
BIN(Miles_per_Gallon)
COUNT
20
40
60
80
BIN(Miles_per_Gallon)
COUNT
1
98

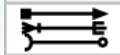
Effective?


Raw
Aggregate (Count)
Miles_per_Gallon
BIN(Miles_per_Gallon)
COUNT
Miles_per_Gallon
COUNT
1
98
9
47

Effective?


Raw
Aggregate (Count)
Miles_per_Gallon
COUNT
BIN(Miles_per_Gallon)
?
?
!?

Visual Encoding Variables

Example	Encoding	Ordered	Useful values	Quantitative	Ordinal	Categorical	Relational
	position, placement	yes	infinite	Good	Good	Good	Good
1, 2, 3; A, B, C	text labels	optional alpha or num	infinite	Good	Good	Good	Good
	length	yes	many	Good	Good		
	size, area	yes	many	Good	Good		
	angle	yes	medium	Good	Good		
	pattern density	yes	few	Good	Good		
	weight, boldness	yes	few		Good		
	saturation, brightness	yes	few		Good		
	color	no	few (<20)			Good	
	shape, icon	no	medium			Good	
	pattern texture	no	medium			Good	
	enclosure, connection	no	infinite			Good	Good
	line pattern	no	few				Good
	line endings	no	few				Good
	line weight	yes	few		Good		

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

Additional Reading

- Text Book: Fundamentals of Data Visualization
  - Chapter 2: Visualizing data: Mapping data onto aesthetics
    https://clauswilke.com/dataviz/aesthetic-mapping.html

References

- M Tufte, E. (2001). *The Visual Display of Quantitative Information (2nd Edition)*. Graphics Press.
  - https://www.edwardtufte.com/
- *Data Visualization Course*, University of Washington
  - https://courses.cs.washington.edu/courses/cse442
- *Visual variables*
  - http://www.infovis-wiki.net/index.php?title=Visual_Variables