

Introduction to Data Visualization

Adopted from Slides for CSE 512 – Data Visualization,
University of Washington, by Jeffrey Heer

Data & Image Models

The Big Picture

task

questions, goals
assumptions

data

physical data type
conceptual data type

domain

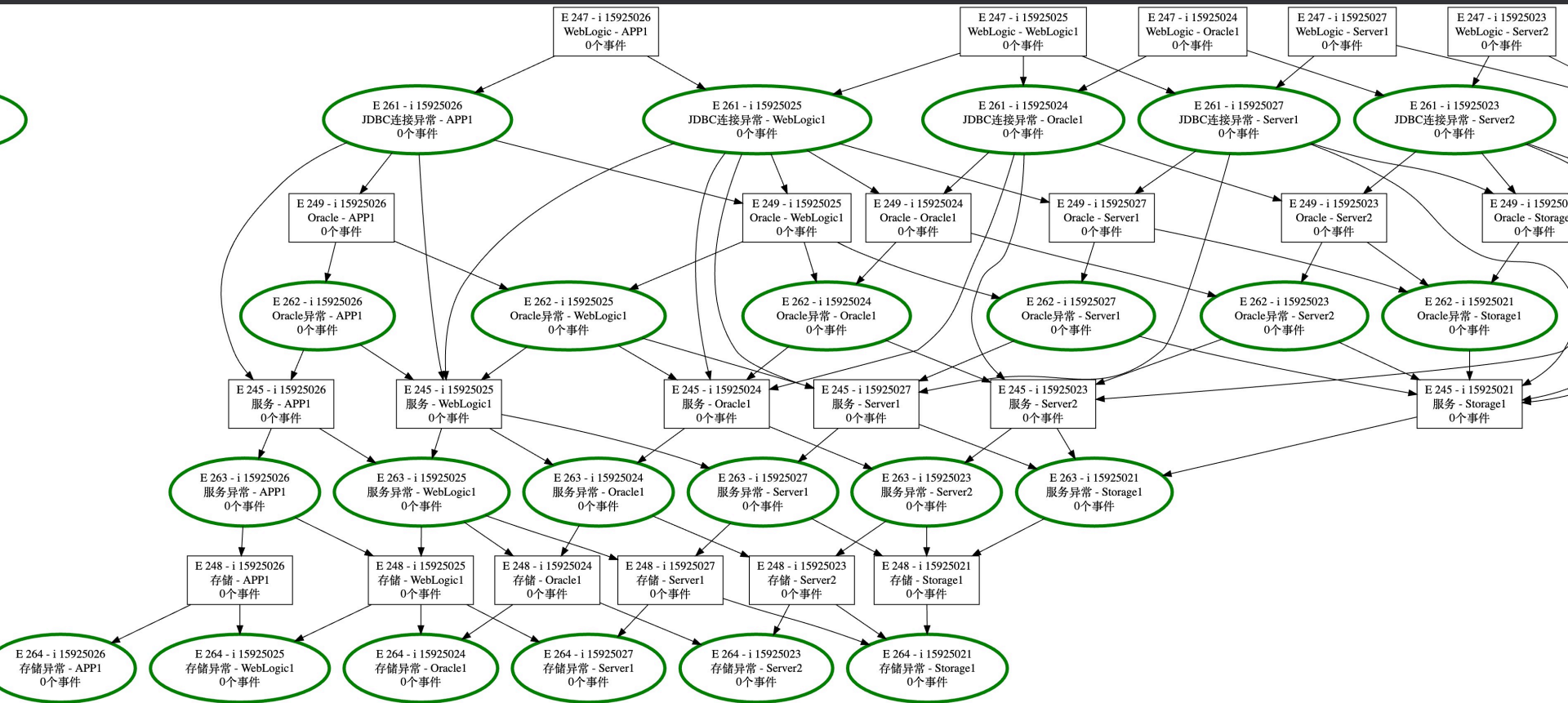
metadata
semantics
conventions

processing
algorithms

mapping
visual encoding

image

visual channel
graphical marks



Topics

Properties of Data

Properties of Images

Mapping Data to Images

Data

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 (Celsius, Fahrenheit)

3D vector of floats vs. spatial location

Taxonomy of Data Types (?)

1D (sets and sequences)

Temporal

2D (maps)

3D (shapes)

nD (relational)

Trees (hierarchies)

Networks (graphs)

Are there others?

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

Nominal, Ordinal & Quantitative

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Fruits: apples, oranges, ...

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Fruits: apples, oranges, ...

O - Ordered

- Quality of meat: Grade A, AA, AAA

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Fruits: apples, oranges, ...

O - Ordered

- Quality of meat: Grade A, AA, AAA

Q - Interval (location of zero arbitrary)

- Dates: Jan, 19, 2006; Location: (LAT 33.98, LONG -118.45)

-

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Fruits: apples, oranges, ...

O - Ordered

- Quality of meat: Grade A, AA, AAA

Q - Interval (location of zero arbitrary)

- Dates: Jan, 19, 2006; Location: (LAT 33.98, LONG -118.45)
-

Q - Ratio (zero fixed)

- Physical measurement: Length, Mass, Temp, ...
- Counts and amounts

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Operations: =, ≠

O - Ordered

- Operations: =, ≠, <, >

Q - Interval (location of zero arbitrary)

- Operations: =, ≠, <, >, -
- Can measure distances or spans

Q - Ratio (zero fixed)

- Operations: =, ≠, <, >, -, %
- Can measure ratios or proportions

From Data Model to N, O, Q

Data Model

32.5, 54.0, -17.3, ...

Floating point numbers

Conceptual Model

Temperature (°C)

Data Type

Burned vs. Not-Burned (N)

Hot, Warm, Cold (O)

Temperature Value (Q)

Dimensions & Measures

Dimensions (~ independent variables)

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...

Example: U.S. Census Data

Example: U.S. Census Data

People Count: # of people in group

Year: 1850 - 2000 (every decade)

Age: 0 - 90+

Sex: Male, Female

Marital Status: Single, Married, Divorced, ...

Example: U.S. Census

People Count

Year

Age

Sex

Marital Status

2,348 data points

	A	B	C	D	E
1	year	age	marst	sex	people
2	1850	0	0	1	1483789
3	1850	0	0	2	1450376
4	1850	5	0	1	1411067
5	1850	5	0	2	1359668
6	1850	10	0	1	1260099
7	1850	10	0	2	1216114
8	1850	15	0	1	1077133
9	1850	15	0	2	1110619
10	1850	20	0	1	1017281
11	1850	20	0	2	1003841
12	1850	25	0	1	862547
13	1850	25	0	2	799482
14	1850	30	0	1	730638
15	1850	30	0	2	639636
16	1850	35	0	1	588487
17	1850	35	0	2	505012
18	1850	40	0	1	475911
19	1850	40	0	2	428185
20	1850	45	0	1	384211
21	1850	45	0	2	341254
22	1850	50	0	1	321343
23	1850	50	0	2	286580
24	1850	55	0	1	194080
25	1850	55	0	2	187208
26	1850	60	0	1	174976
27	1850	60	0	2	162236
28	1850	65	0	1	106827
29	1850	65	0	2	105534
30	1850	70	0	1	73677
31	1850	70	0	2	71762
32	1850	75	0	1	40834
33	1850	75	0	2	40229
34	1850	80	0	1	23449
35	1850	80	0	2	22949
36	1850	85	0	1	8186
37	1850	85	0	2	10511
38	1850	90	0	1	5259
39	1850	90	0	2	6569
40	1860	0	0	1	2120846
41	1860	0	0	2	2092162

Census: N, O, Q?

People Count

Q-Ratio

Year

Q-Interval (O)

Age

Q-Ratio (O)

Sex

N

Marital Status

N

Census: Dimension or Measure?

People Count

Measure

Year

Dimension

Age

Depends!

Sex

Dimension

Marital Status

Dimension

Data 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]

Data Transformations (sql)

Projection (`select`) - selects columns

Selection (`where`) - filters rows

Sorting (`order by`)


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

Combine relations (`union, join, ...`)

Roll-Up and Drill-Down

Want to examine marital status in each decade?

Roll-up the data along the desired dimensions



The diagram consists of two horizontal curly braces. The left brace is labeled 'Dimensions' and spans the words 'year, marst,' in the SQL query below. The right brace is labeled 'Measure' and spans the words 'sum(people)' in the same query.

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

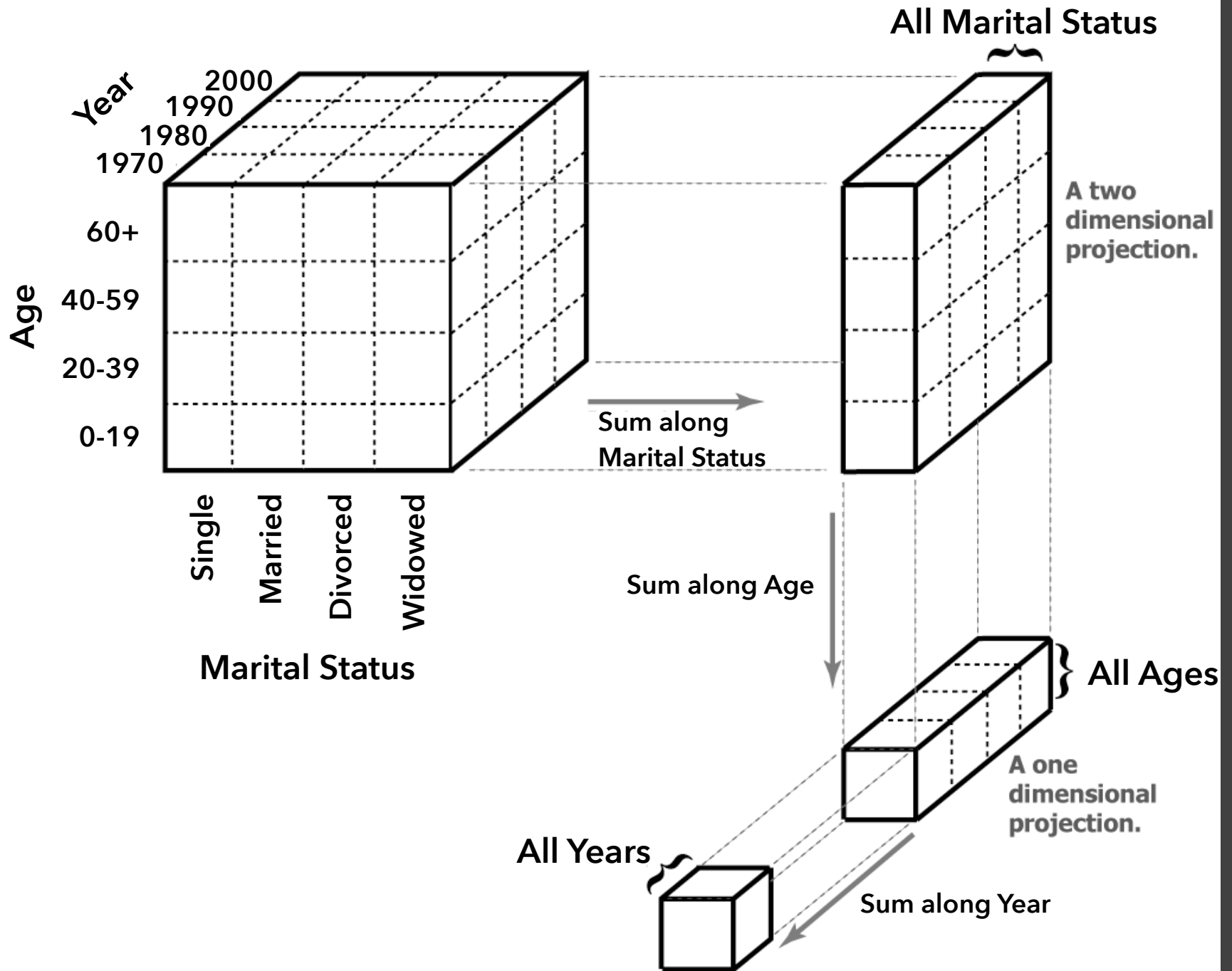
Dimensions

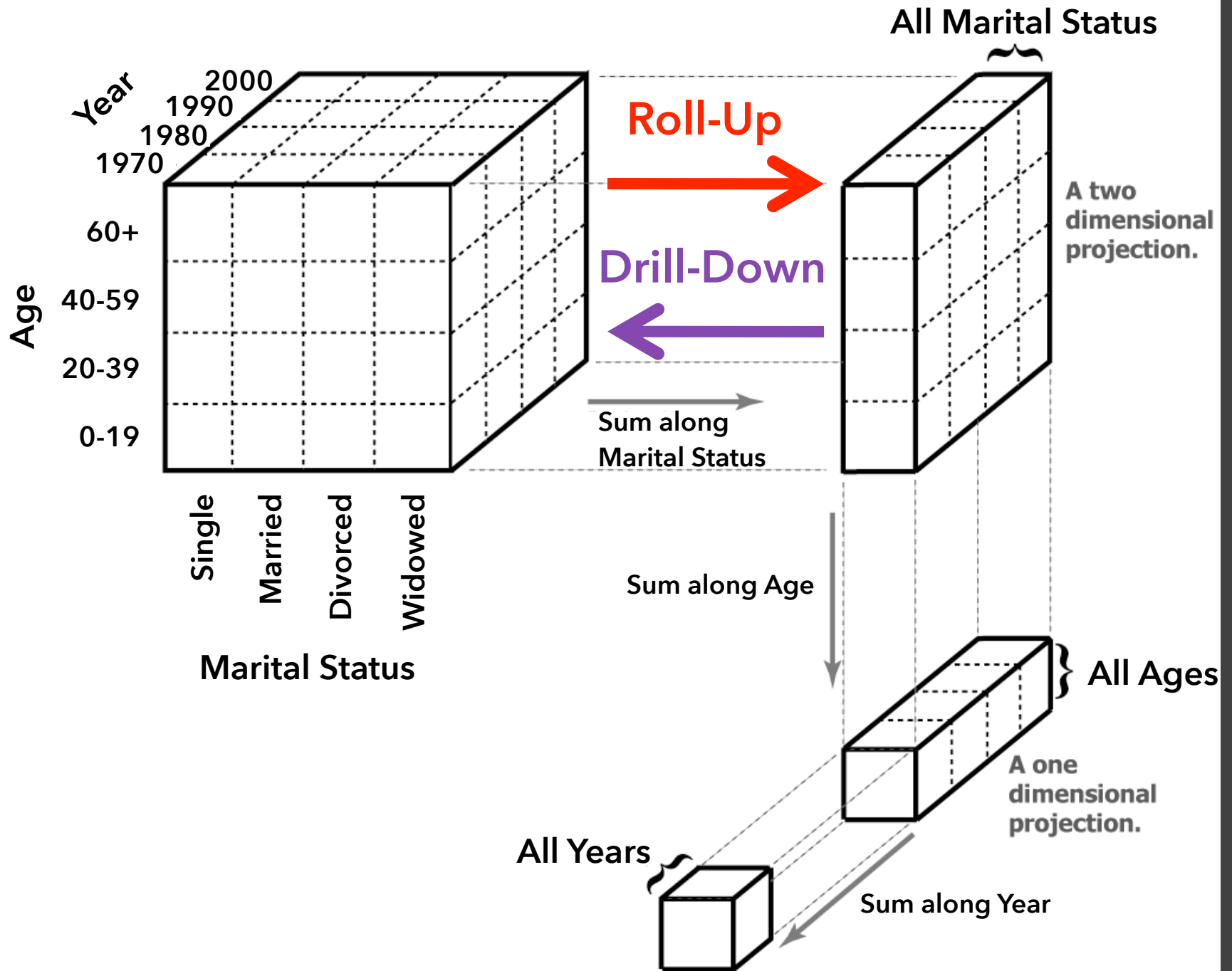
Roll-Up and Drill-Down

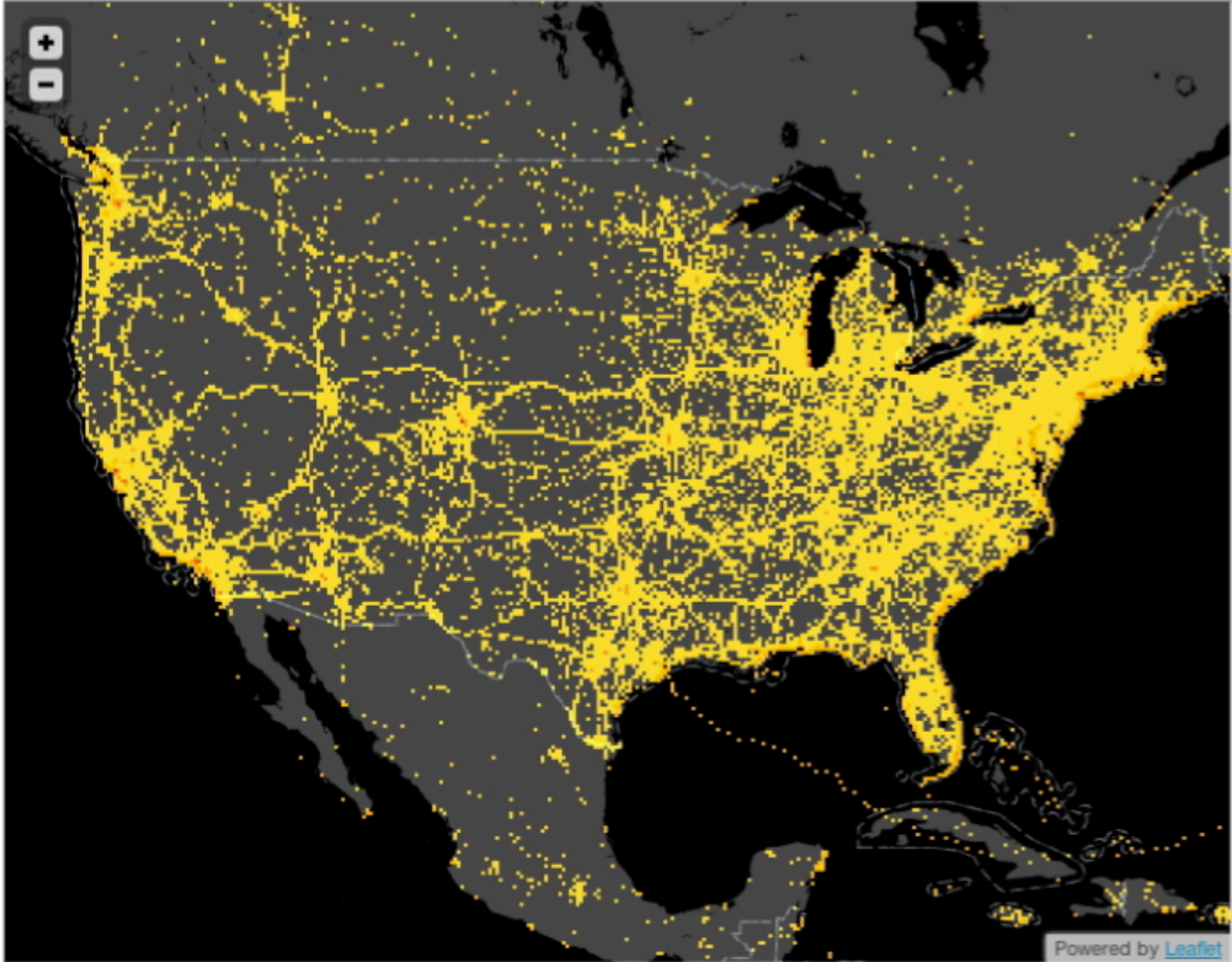
Need more detailed information?

Drill-down into additional dimensions

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



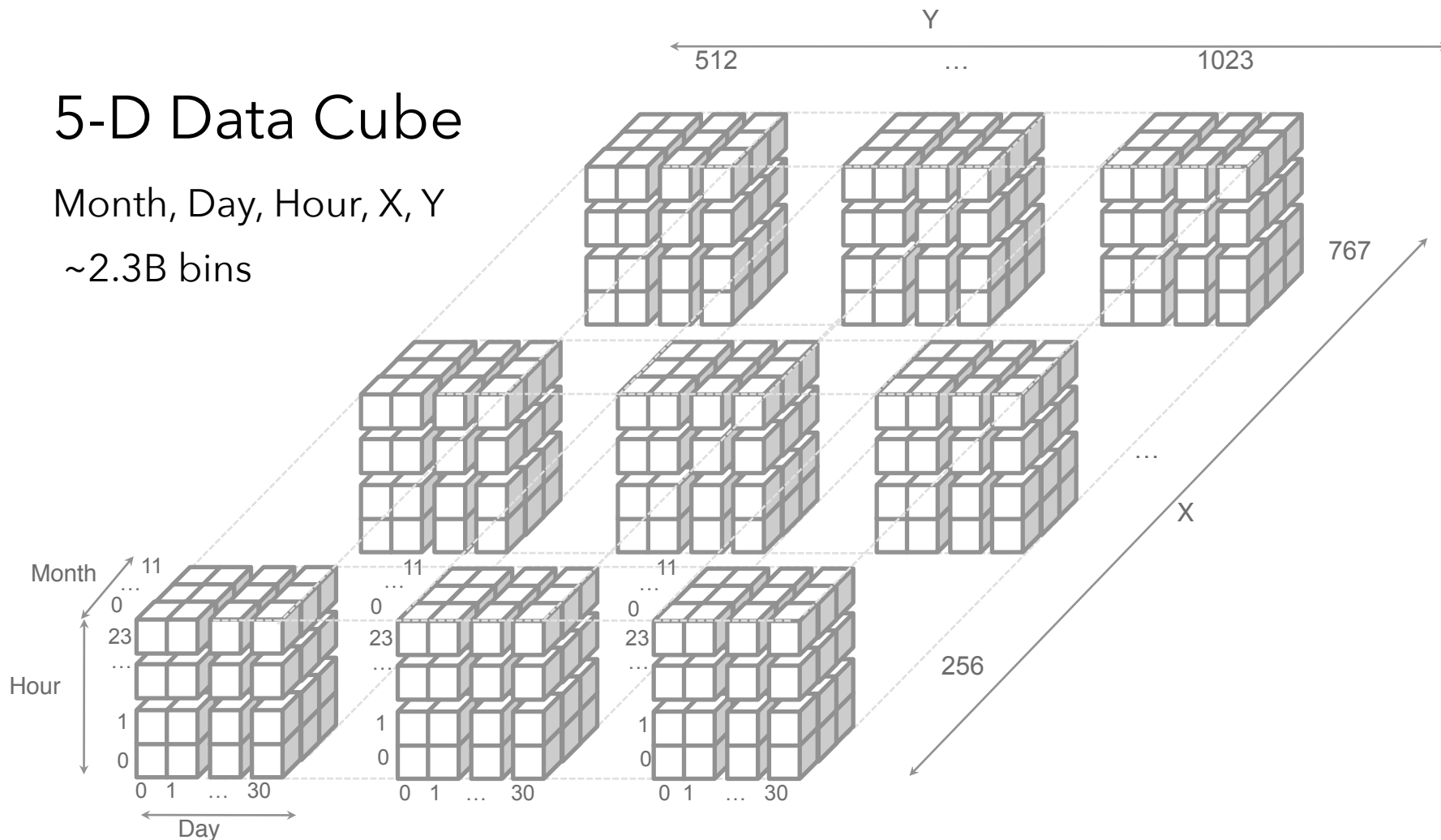




5-D Data Cube

Month, Day, Hour, X, Y

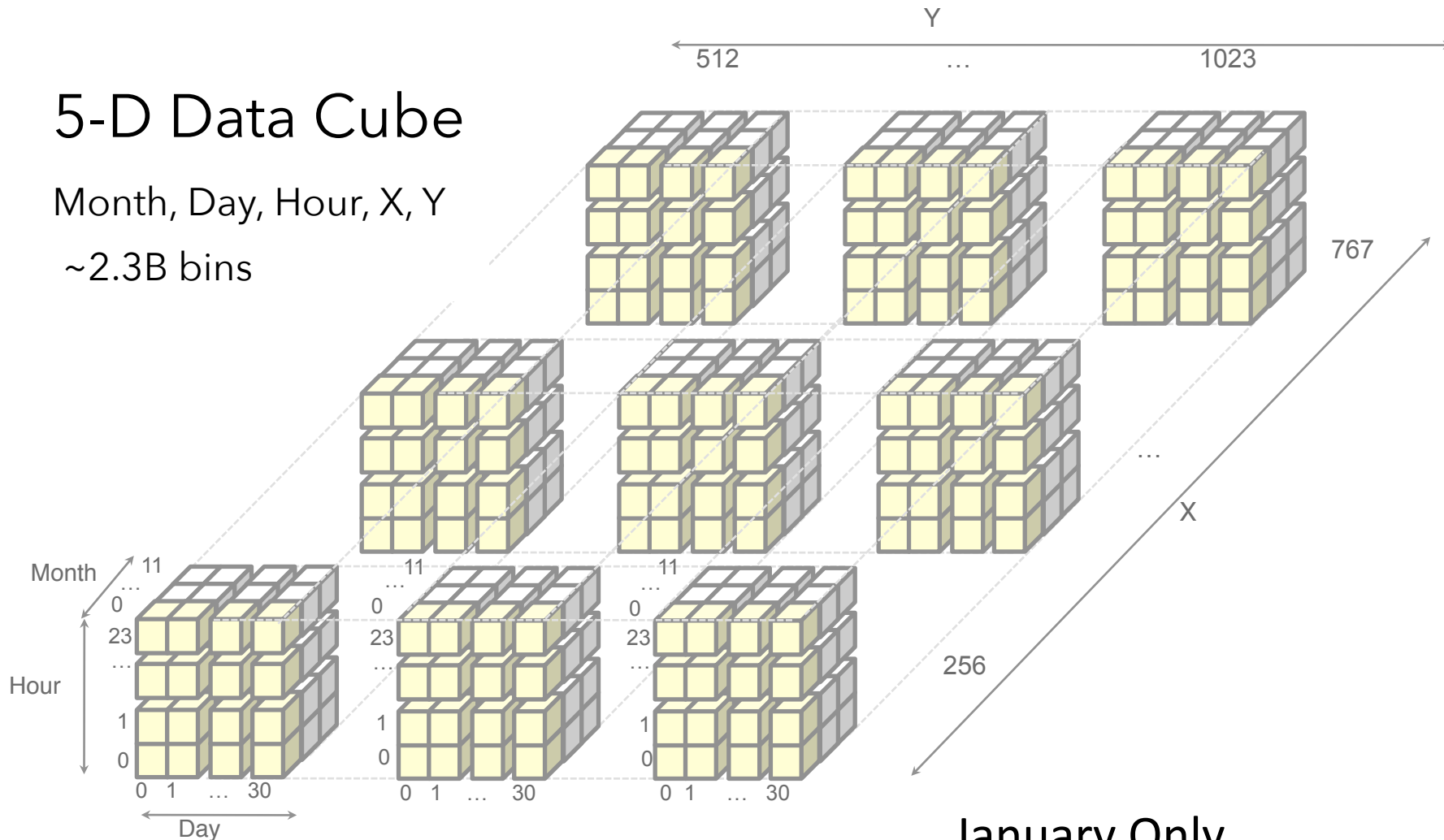
~2.3B bins



5-D Data Cube

Month, Day, Hour, X, Y

~2.3B bins



Visual Encoding Variables

Position (x 2)

Size

Value

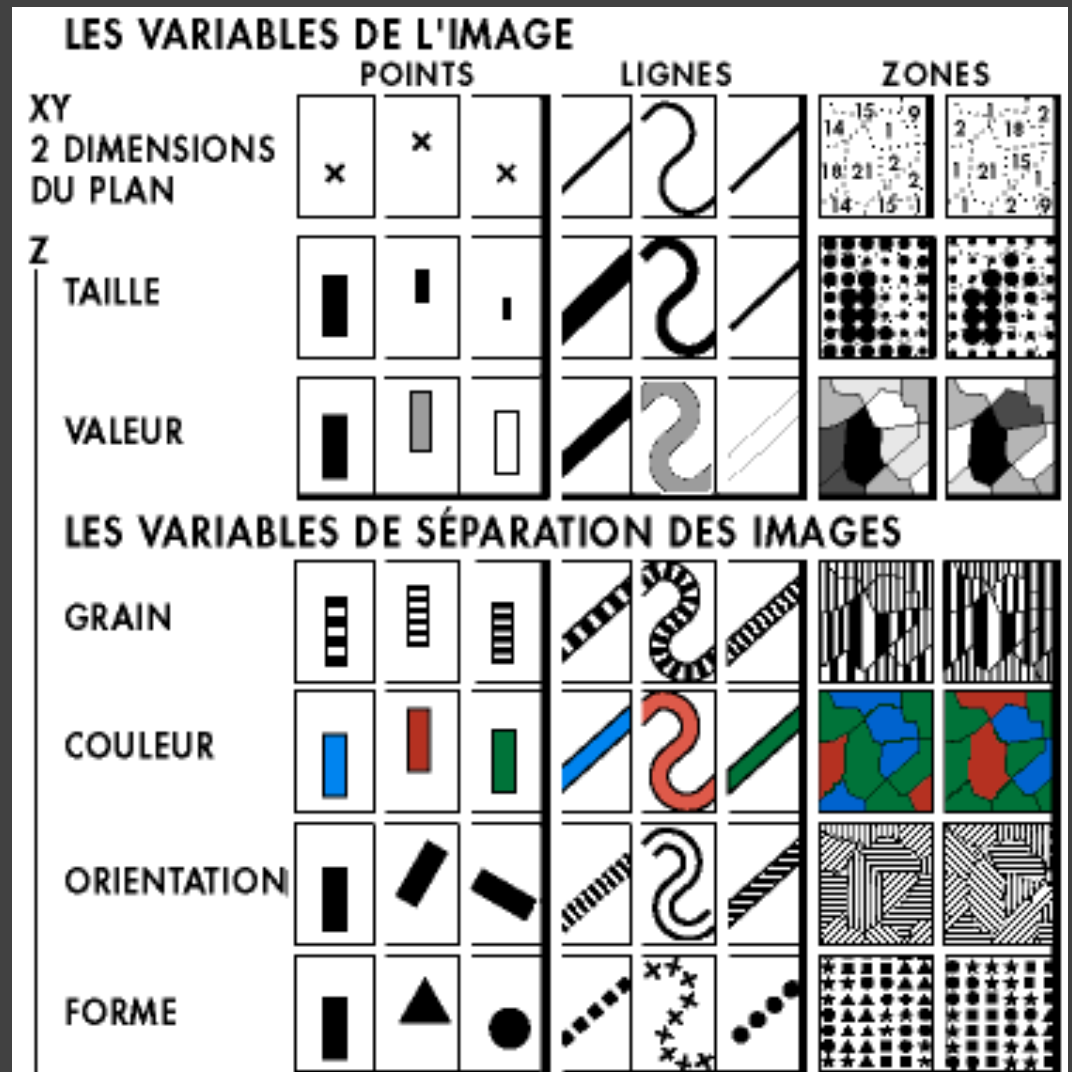
Texture

Color

Orientation

Shape

Others?



Bertin's "Levels of Organization"

Position

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

Nominal

Size

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

Ordinal

Value

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

Quantitative

Note: $Q \subset O \subset N$

Texture

N	o	
---	---	--

Color

N		
---	--	--

Orientation

N		
---	--	--

Shape

N		
---	--	--

Choosing Visual Encodings

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

Principle of Consistency

The properties of the image (visual variables) should match the properties of the data.

Principle of Importance Ordering

Encode the most important information in the most effective way.

Design Criteria [Mackinlay 86]

Expressiveness

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

Effectiveness

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

Design Criteria *Translated*

Tell the truth and nothing but the truth
(don't lie, and don't lie by omission)

Use encodings that people decode better
(where better = faster and/or more accurate)

Effectiveness Rankings [Mackinlay 86]

QUANTITATIVE

Position
Length
Angle
Slope
Area (Size)
Volume
Density (Value)
Color Sat
Color Hue
Texture
Connection
Containment
Shape

ORDINAL

Position
Density (Value)
Color Sat
Color Hue
Texture
Connection
Containment
Length
Angle
Slope
Area (Size)
Volume
Shape

NOMINAL

Position
Color Hue
Texture
Connection
Containment
Density (Value)
Color Sat
Shape
Length
Angle
Slope
Area
Volume

Effectiveness Rankings [Mackinlay 86]

QUANTITATIVE

Position

Length
Angle
Slope
Area (Size)
Volume
Density (Value)
Color Sat
Color Hue
Texture
Connection
Containment
Shape

ORDINAL

Position

Density (Value)
Color Sat
Color Hue
Texture
Connection
Containment
Length
Angle
Slope
Area (Size)
Volume
Shape

NOMINAL

Position

Color Hue
Texture
Connection
Containment
Density (Value)
Color Sat
Shape
Length
Angle
Slope
Area
Volume

Design Considerations

Title, labels, legend, captions, source!

Expressiveness and Effectiveness

Avoid unexpressive marks (lines? gradients?)

Use perceptually effective encodings

Don't distract: faint gridlines, pastel highlights/fills

The "elimination diet" approach - start minimal

Support comparison and pattern perception

Between elements, to a reference line, or to totals

Design Considerations

Transform data (e.g., invert, log, normalize)

Are model choices (regression lines) appropriate?

Group / sort data by meaningful dimensions

Reduce cognitive overhead

Minimize visual search, minimize ambiguity

Avoid legend lookups if direct labeling works

Avoid color mappings with indiscernible colors

Be consistent! Visual inferences should consistently support data inferences.

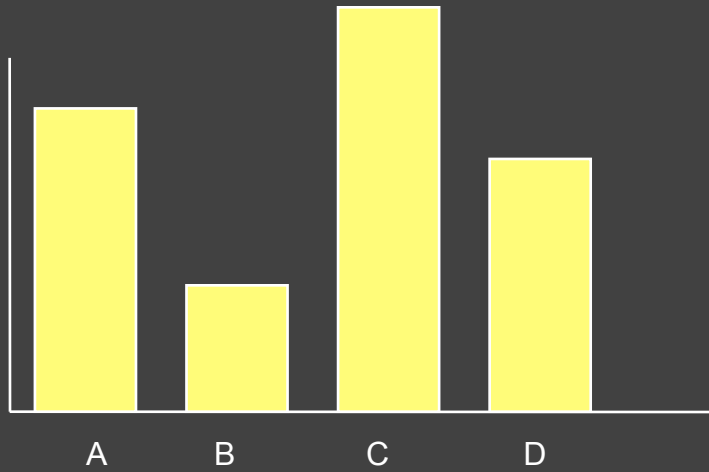
The Design Space of Visual Encodings

Univariate Data

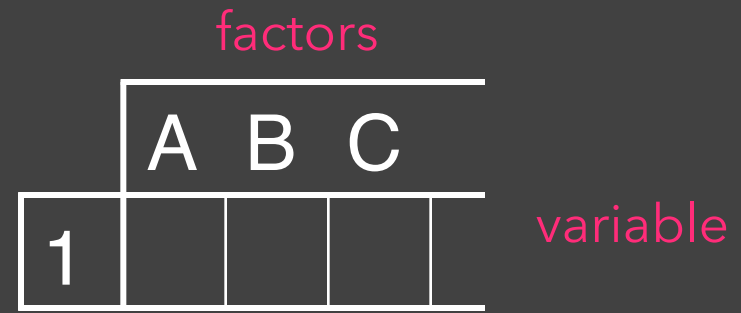
factors

	A	B	C	
1				

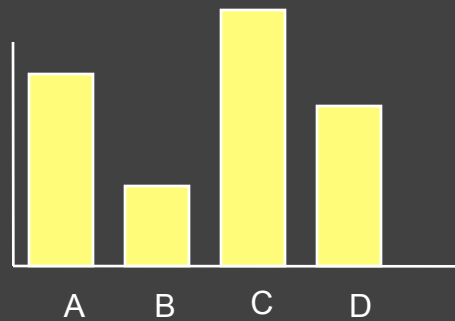
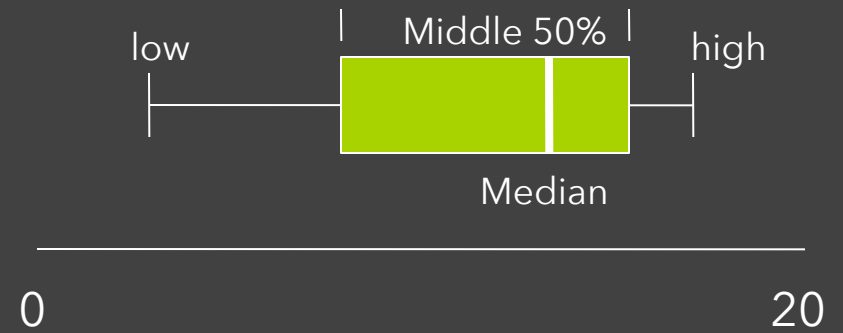
variable



Univariate Data

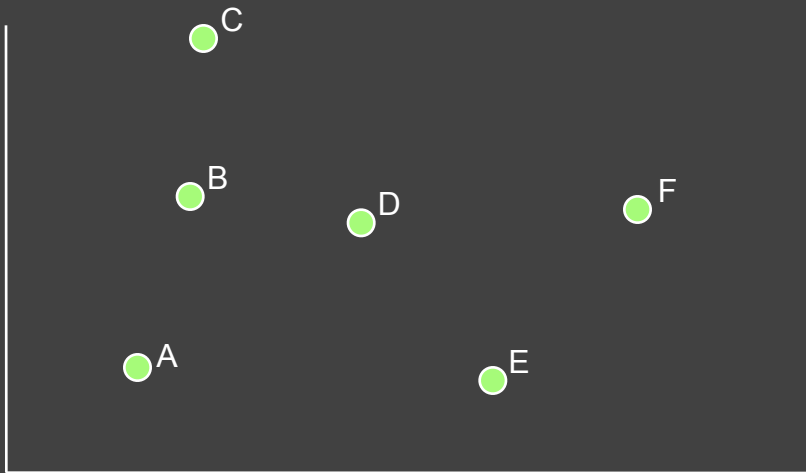


Tukey box plot



Bivariate Data

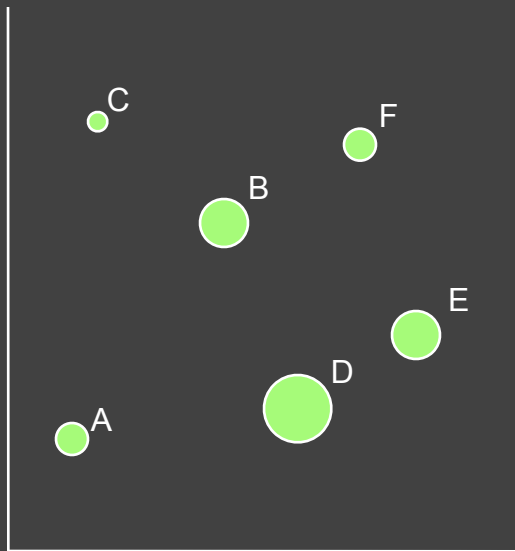
	A	B	C
1			
2			



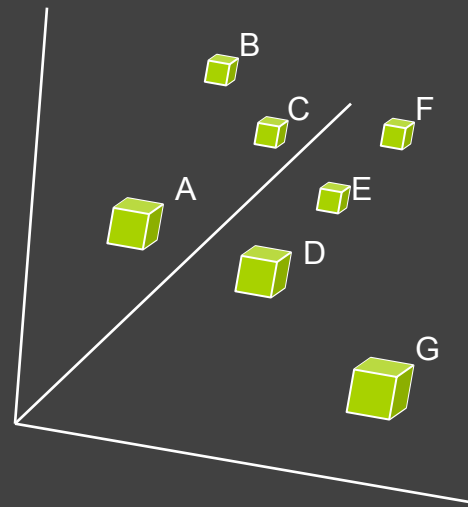
Scatter plot is common

Trivariate Data

	A	B	C	
1				
2				
3				



3D scatter plot is possible



Multidimensional Data

Visual Encoding Variables

Position (X)

Position (Y)

Size

Value

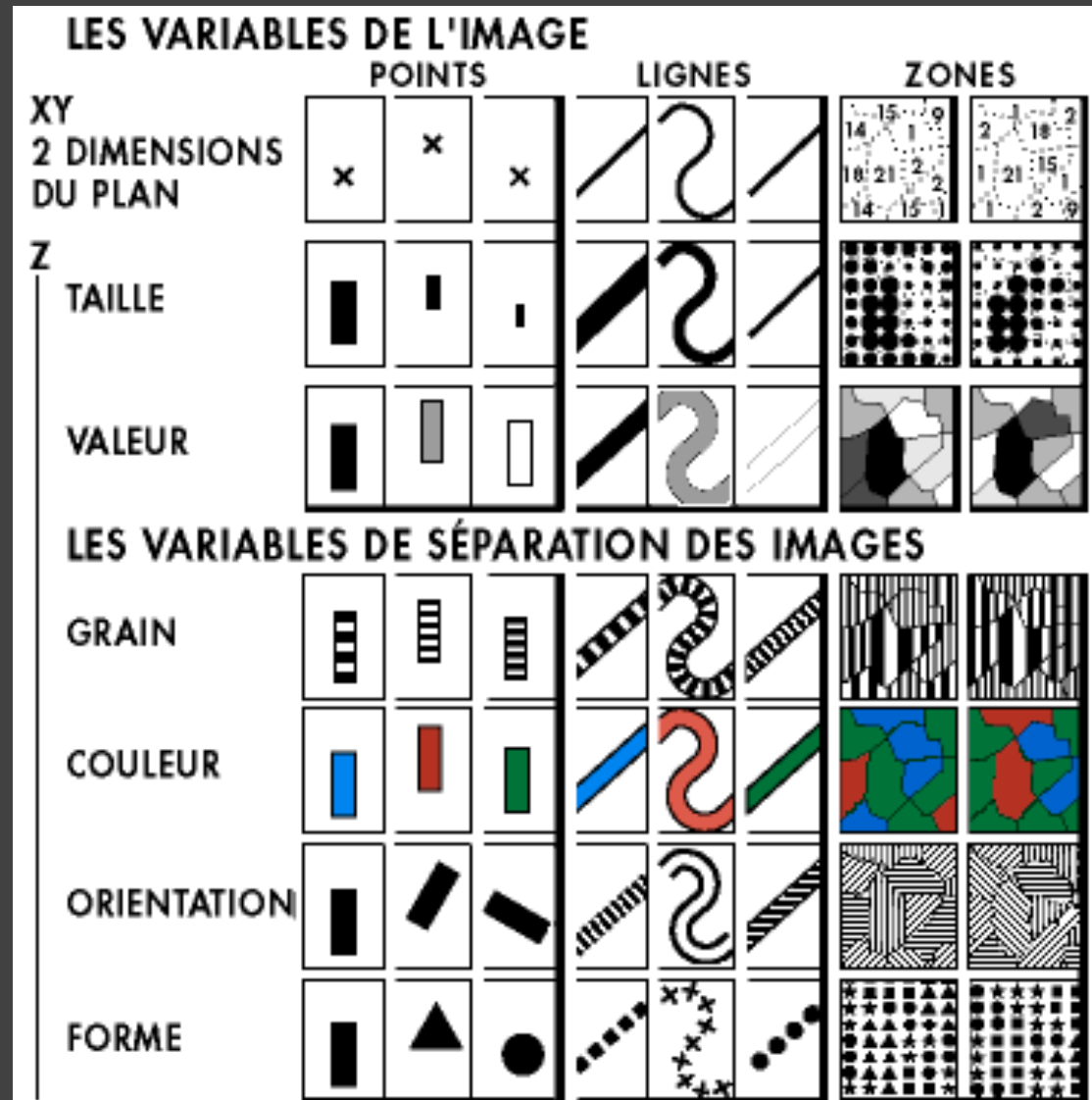
Texture

Color

Orientation

Shape

~8 dimensions?



Example: Coffee Sales

Sales figures for a fictional coffee chain

Sales	Q-Ratio
Profit	Q-Ratio
Marketing	Q-Ratio
Product Type	N {Coffee, Espresso, Herbal Tea, Tea}
Market	N {Central, East, South, West}

Filters

YEAR(Date): 2010

Marks

x+ Automatic

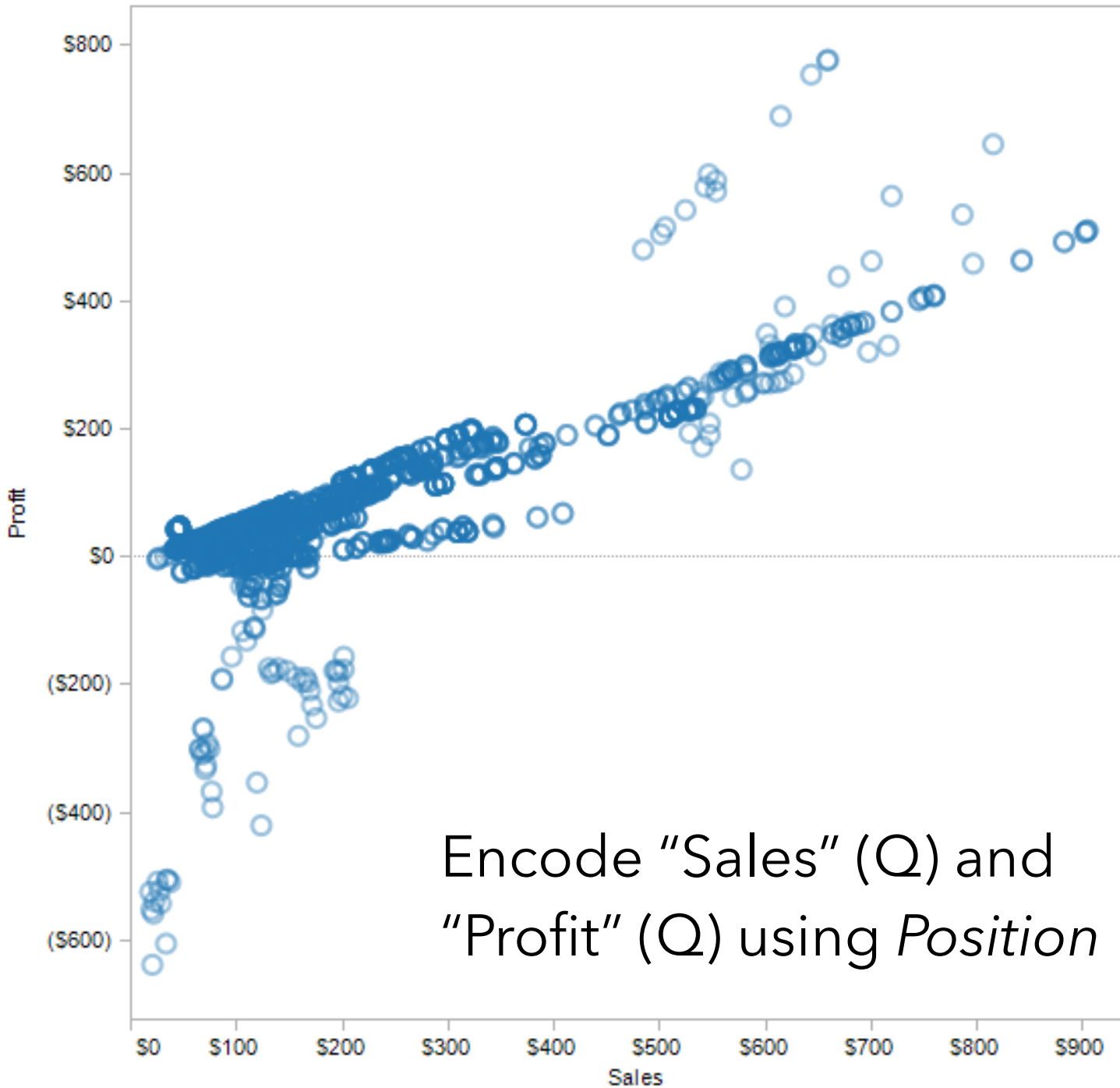
Shape Circle

Label

Color

Size

Level of Detail



Filters

YEAR(Date): 2010

Marks

x+ Automatic

Shape

Label

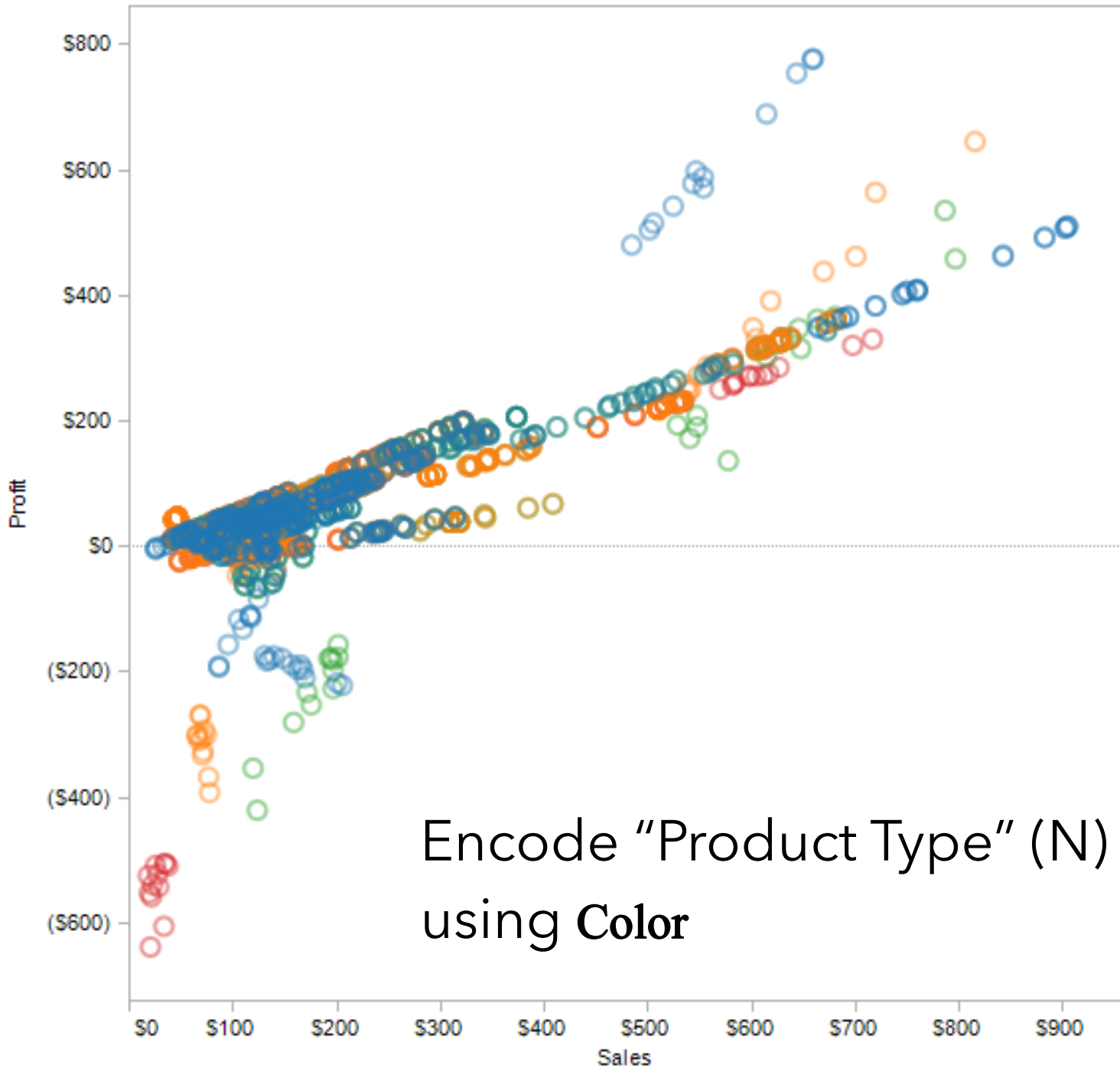
Color

Size

Level of Detail

Product Type

- Coffee
- Espresso
- Herbal Tea
- Tea



Filters

YEAR(Date): 2010

Marks

Automatic

Shape Market

Label Market

Color Product Type

Size

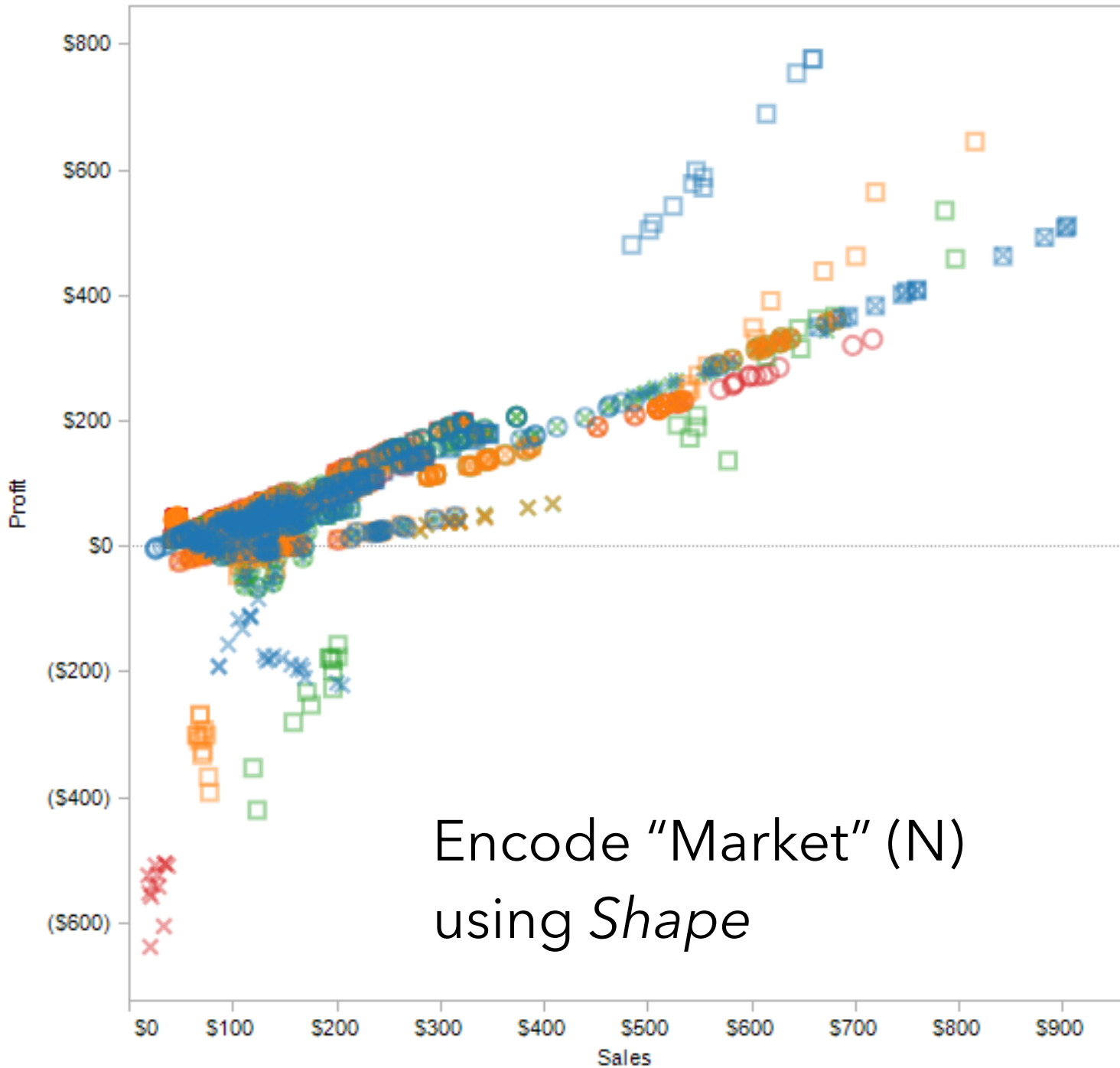
Level of Detail

Product Type

- Coffee
- Espresso
- Herbal Tea
- Tea

Market

- Central
- East
- South
- West



Encode "Market" (N)
using *Shape*

Filters

YEAR(Date): 2010

Marks

Automatic

Shape Market

Label

Color Product Type

Size Marketing

Marketing

Level of Detail

Product Type

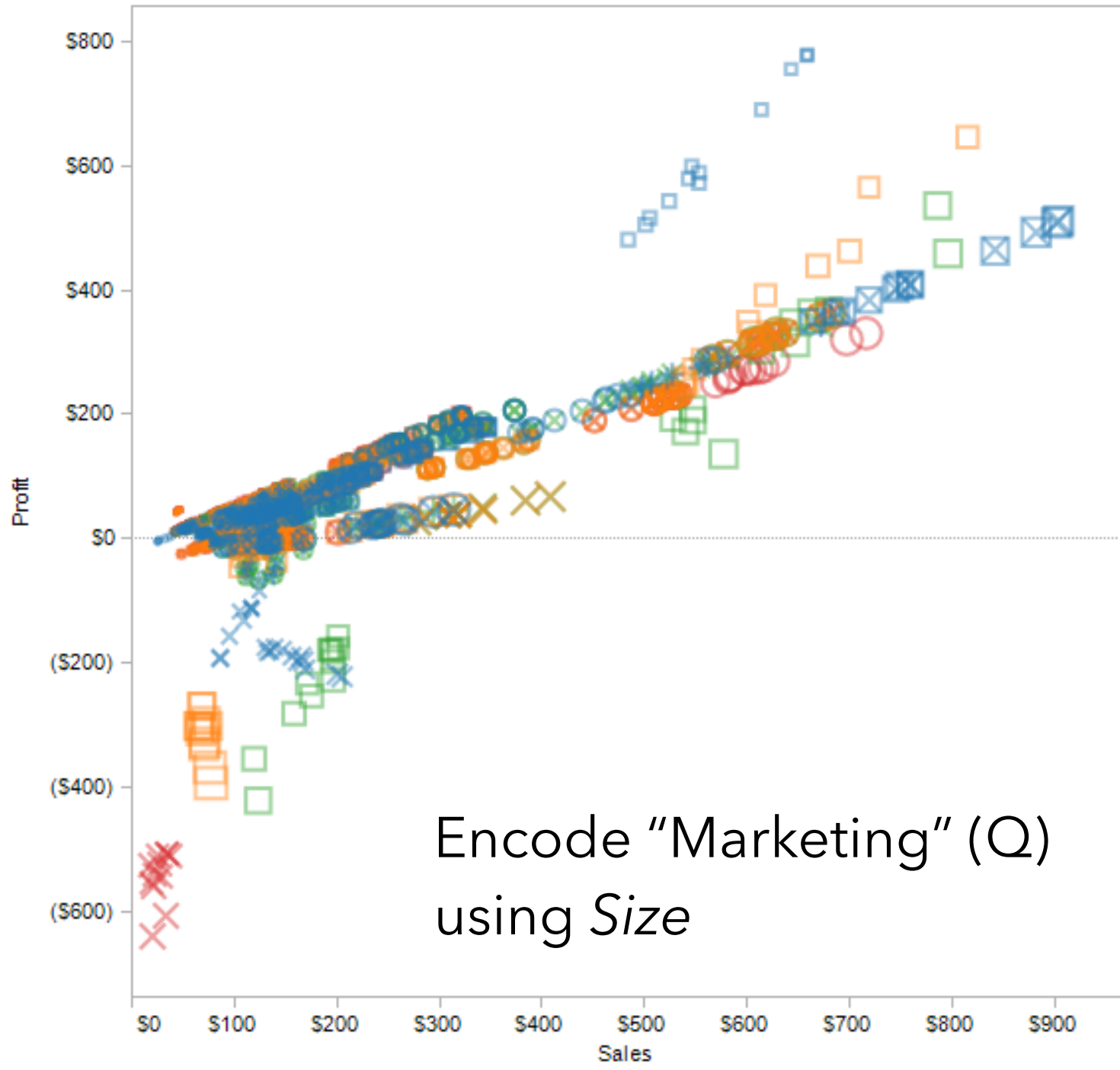
- Coffee
- Espresso
- Herbal Tea

Market

- Central
- East
- South

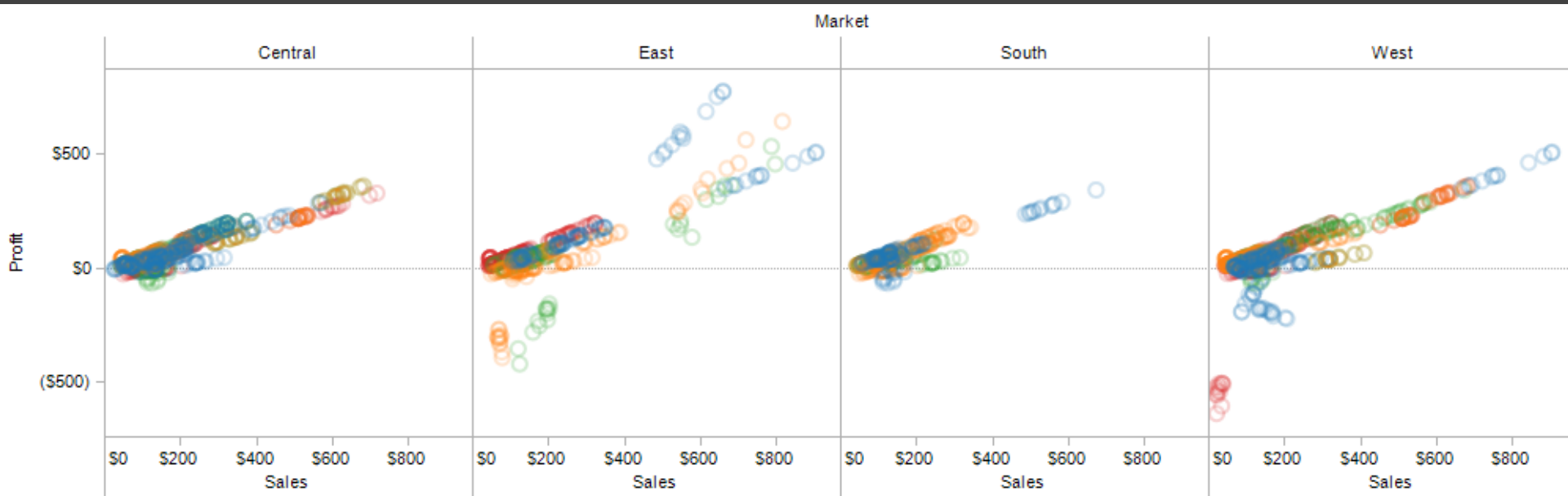
Marketing

- \$0
- \$50
- \$100



Encode "Marketing" (Q) using *Size*

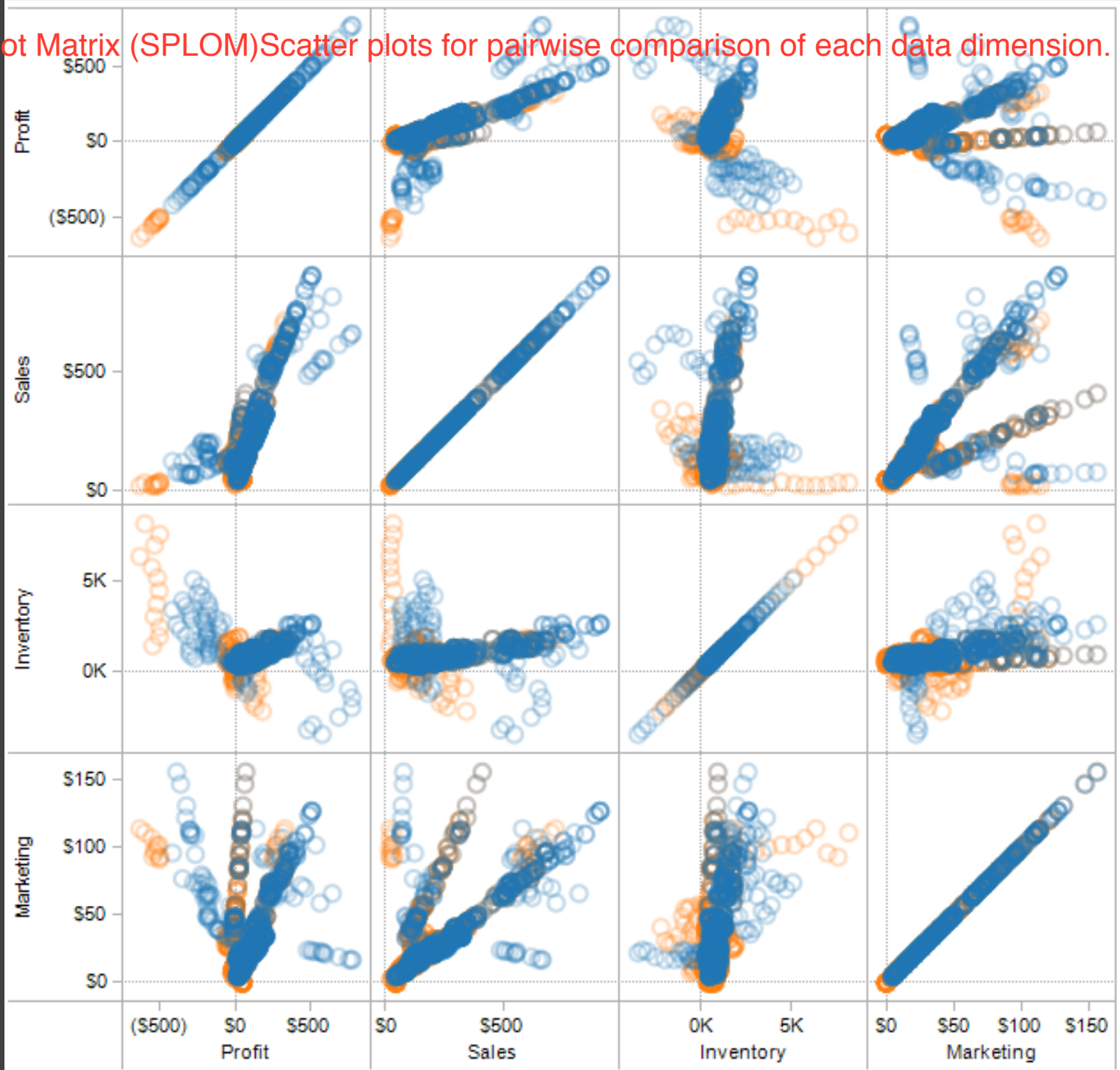
Trellis Plots



A *trellis plot* subdivides space to enable comparison across multiple plots.

Typically nominal or ordinal variables are used as dimensions for subdivision.

Scatterplot Matrix (SPLOM) Scatter plots for pairwise comparison of each data dimension.



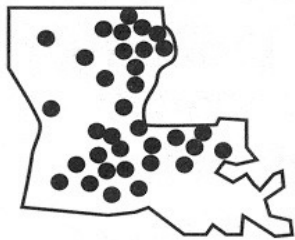
Small Multiples



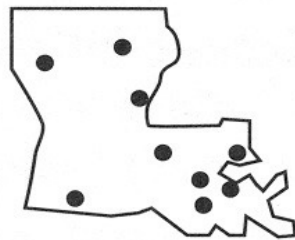
[MacEachren 95, Figure 2.11, p. 38]

Small Multiples

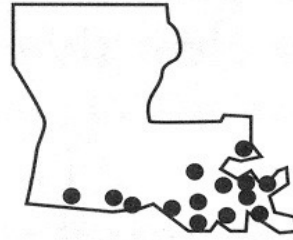
alfisol



entisol



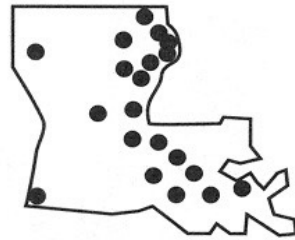
histosol



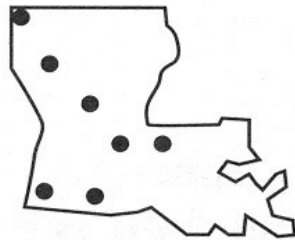
geographic distribution of different soil



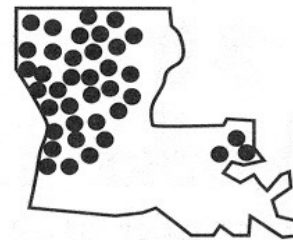
inceptisol



mollisol



ultisol



[MacEachren 95, Figure 2.11, p. 38]