INFO 290T

Human-Centered Data Management

A Brief Data-Centric Visualization Primer

Slide Credits to Jeff Heer & Arvind Satyanarayanan & Tamara Munzner

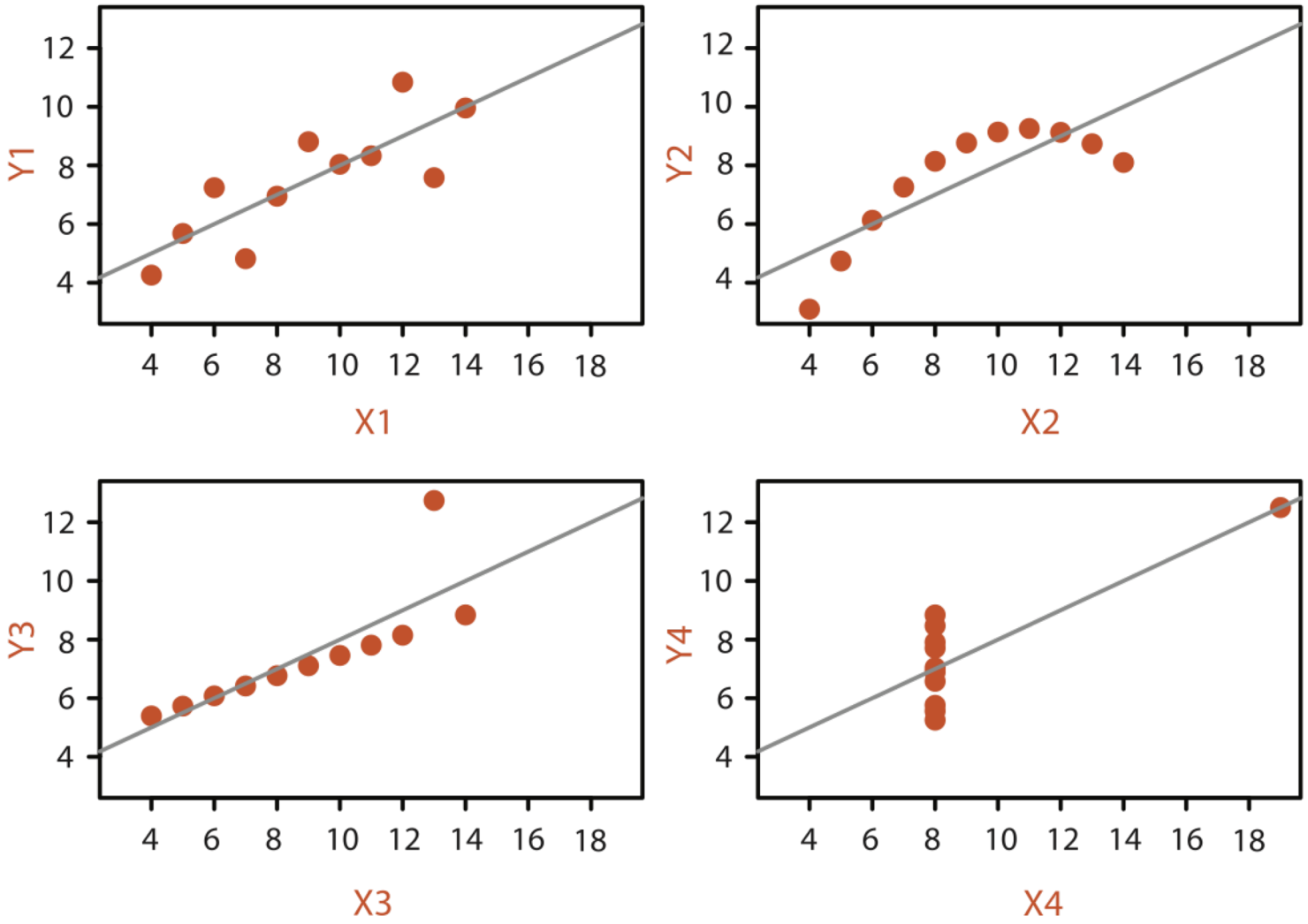
Resources:
Tamara Munzner's Book: Visualization Analysis and Design

Why Visualizations?

- Analyze (aka *exploration*)
  - Discover trends
    - Stock price is going up/down
  - Develop & check hypotheses
    - House prices are down due to the downturn
  - Detect errors
    - Null values in a column

- Share, record, communicate & collaborate (aka *explanation*)

Why Not Statistics?



Anscombe's Quartet

Identical statistics	
x mean	9
x variance	10
y mean	7.5
y variance	3.75
x/y correlation	0.816

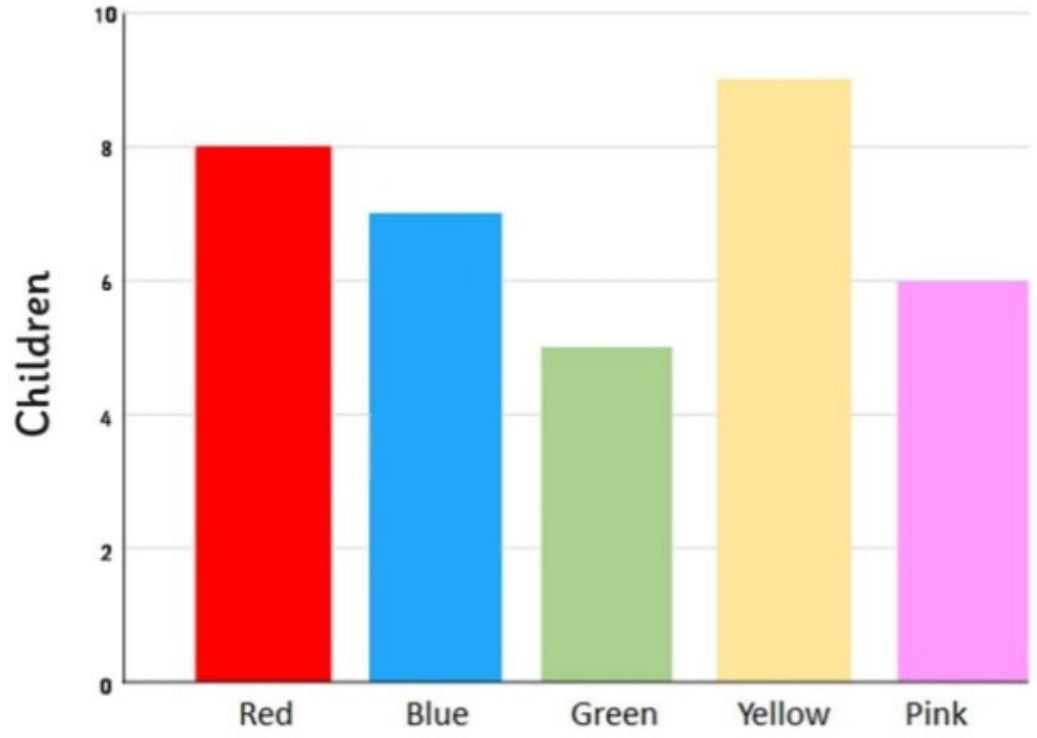
Anscombe, 1973

Visualizations ←→ SQL Queries

Most visualizations are group-by queries

```
SELECT AGG(M), D
FROM R
WHERE …
GROUP BY D
```

```
SELECT COUNT(*), Color
FROM R
GROUP BY Color
```



Types of Data: The Data Processing Viewpoint

Dimensions

- Independent variables
- Usually discrete, e.g., categories, dates, bins
- Can include numeric data, but usually doesn't make sense to aggregate
- Usually the GROUP BY columns in a SQL query

Measures

- Dependent variables (a function of one or more dimension vars)
- Usually continuous – can be analyzed and aggregated
- These are aggregated columns in a GROUP BY query

Dimensions/Measures?

US Census Data

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

	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

Dimensions/Measures?

US Census Data

- People Count: Measure
- Year: Dimension
- Age: Dimension (could vary in general!)
- Marital Status: Dimension
- Sex: Dimension

	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

Types of Data: The Visualization Viewpoint

- Nominal
  - =, ≠
- Ordinal
  - =, ≠, <, >
- Quantitative Interval
  - =, ≠, <, >, –
  - Arbitrary zero
- Quantitative Ratio
  - =, ≠, <, >, –, %
  - Physical quantities

Airlines, Genre

Film ratings, Batteries

Year, Location

Sales, Profit, Temperature

Types of Data: The Visualization Viewpoint

- Nominal
  - =, ≠
- Ordinal
  - =, ≠, <, >
- Quantitative Interval
  - =, ≠, <, >, –
  - Arbitrary zero
- Quantitative Ratio
  - =, ≠, <, >, –, %
  - Physical quantities

Hot, cold

Temperature

Good, OK, Bad

Grade

Score

N/O/QI/QR?

Order ID

Order Date

Order Priority

Product Container

Product Base Margin

Ship Date

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

N/O/QI/QR?

Order ID: N / O

Order Date: QI

Order Priority: O

Product Container: O

Product Base Margin: QR

Ship Date: QI

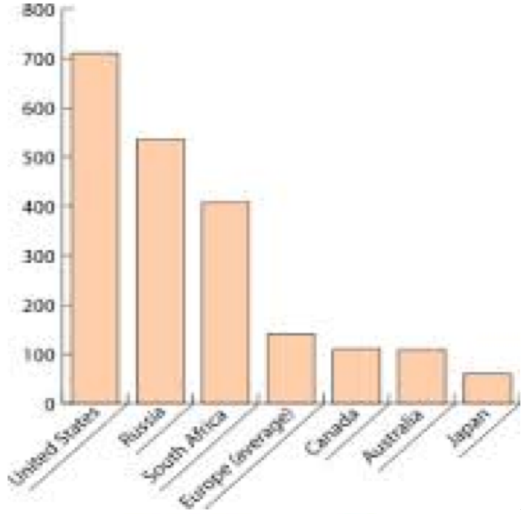
A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

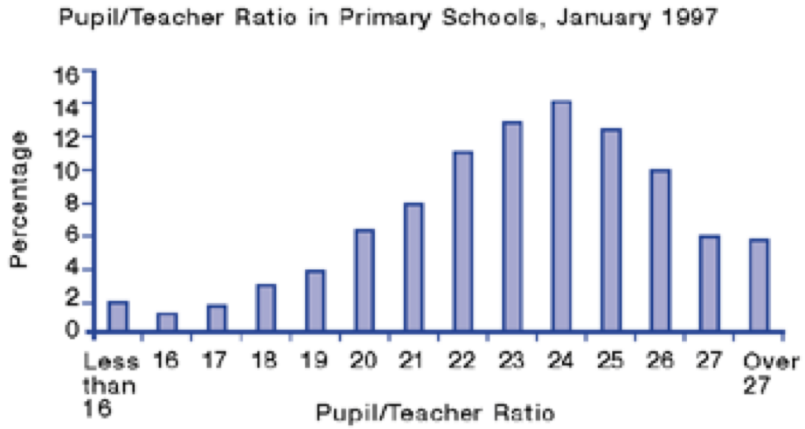
A Very Quick Primer on Visualization Types

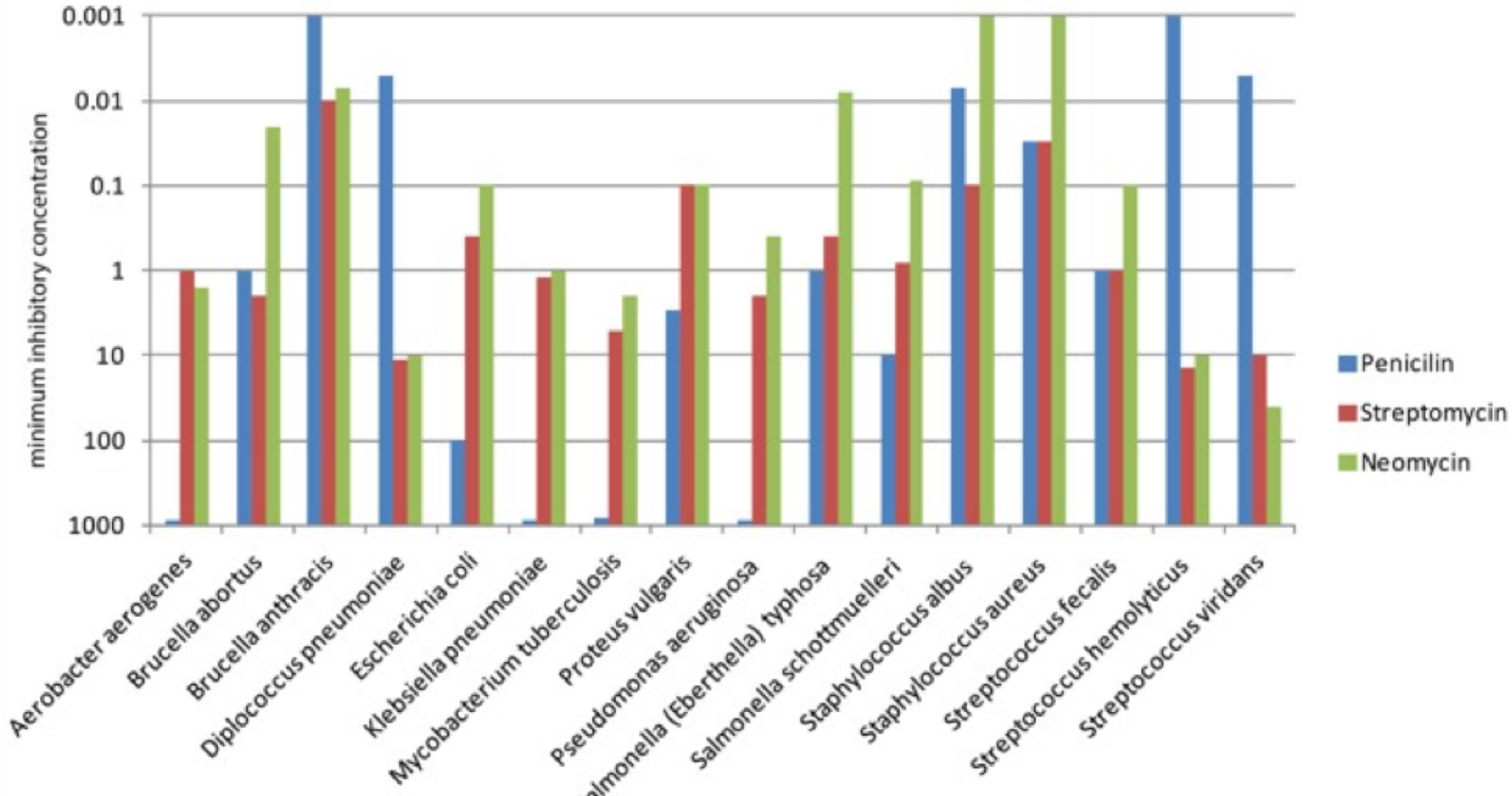
The most basic visualization is a table!

- Bar Charts
- Line Charts
- Scatter Plot
- Choropleth

Bar Charts





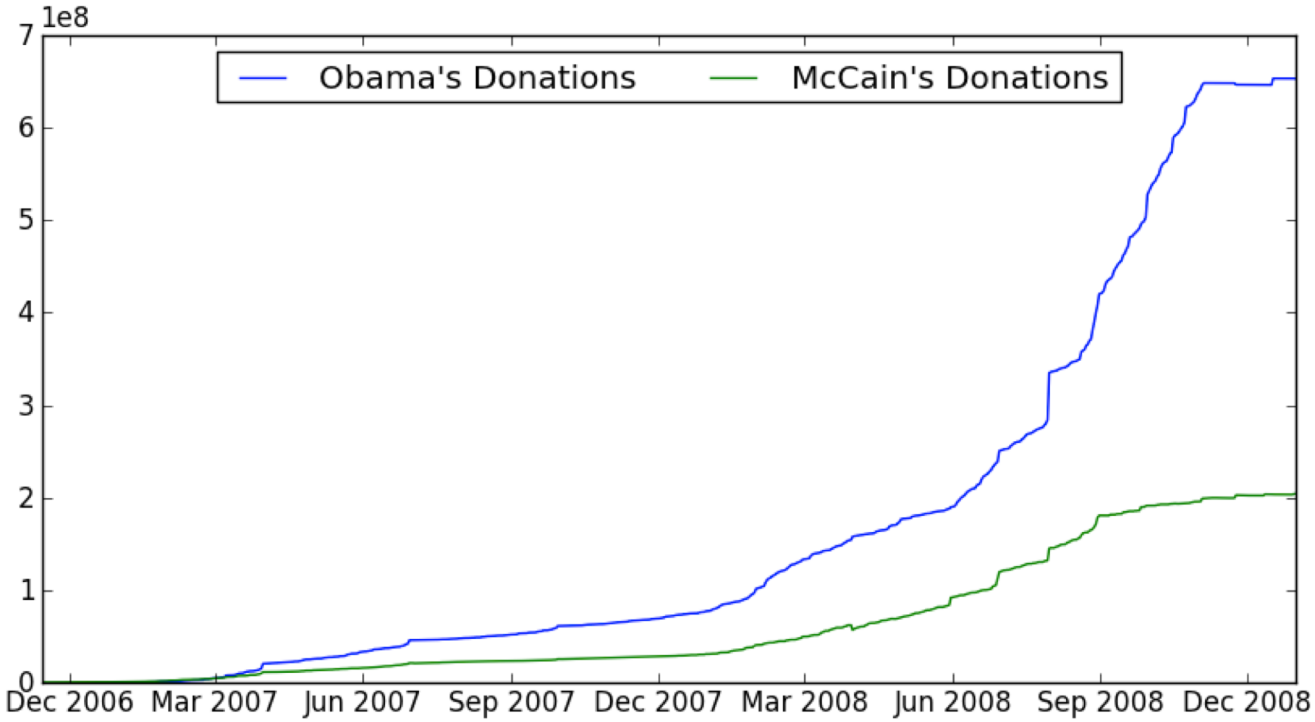


Q: What SQL query generates a multiple bar chart?

When are bar charts appropriate?

- When plotting a Q-R vs. either an N, O, Q-I, or Q-R
- Emphasizes the differences in height than differences in X axis
- Most fundamental chart
- From a SQL standpoint, simple aggregation of some Y axis measure, grouped by one or more dimensions
  - can generate results in the appropriate order in the X axis by doing an ORDER BY following the GROUP BY

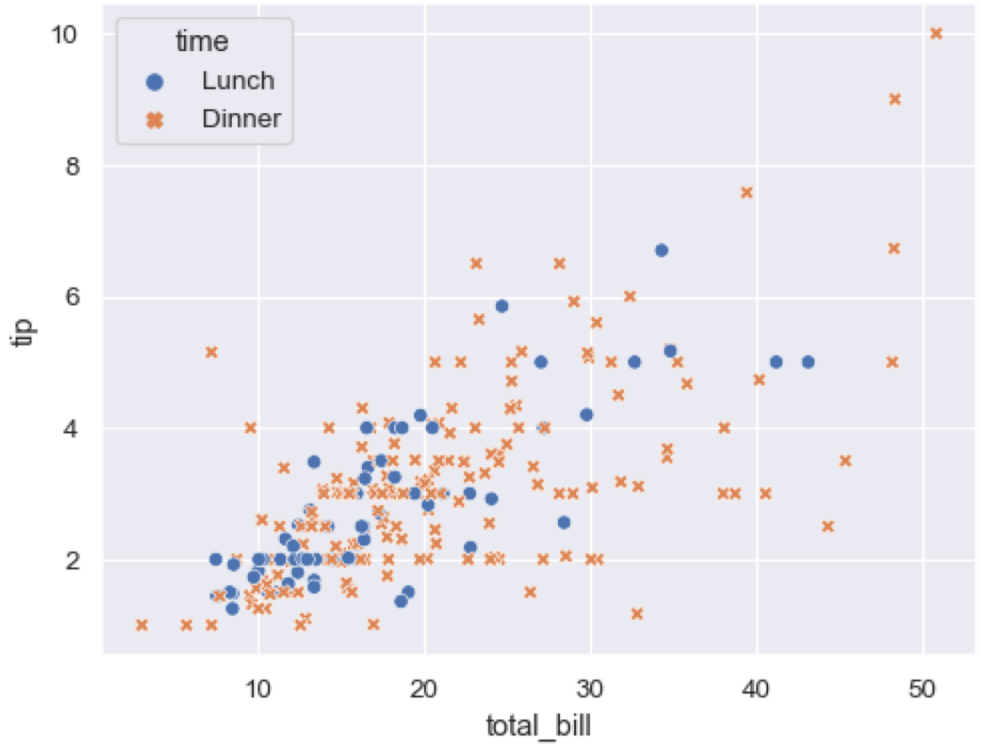
Line Charts

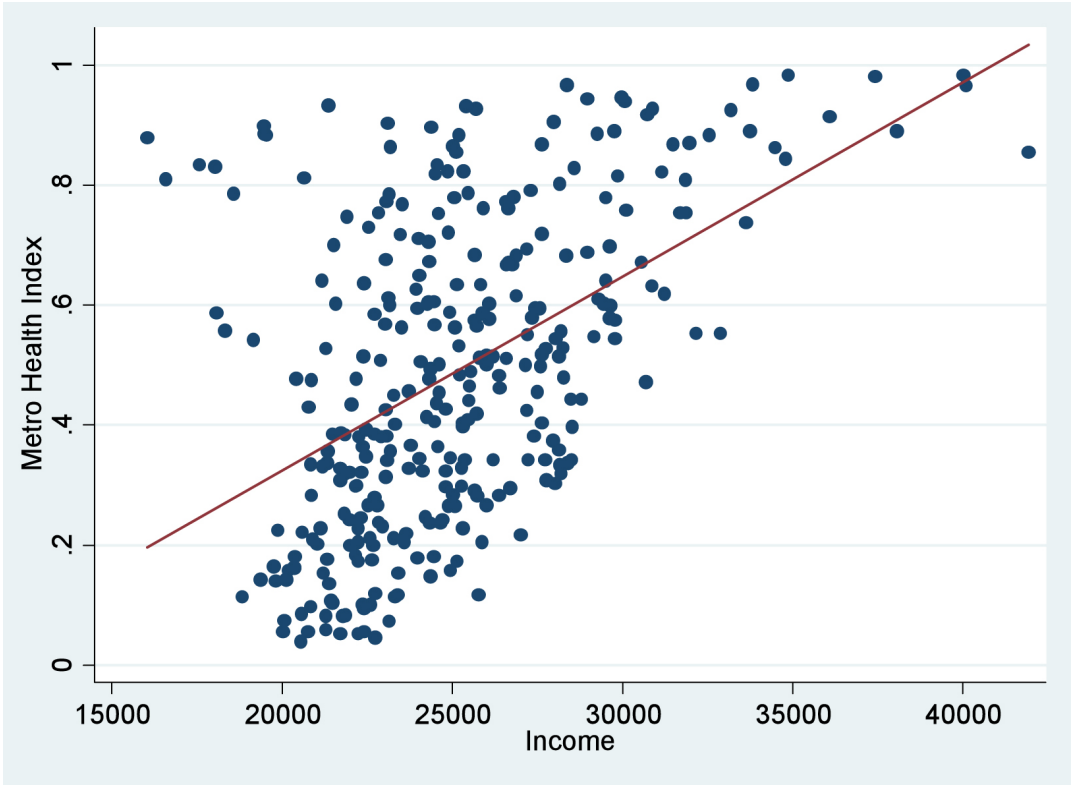

1e8
7
6
5
4
3
2
1
0
Obama's Donations
McCain's Donations
Dec 2006
Mar 2007
Jun 2007
Sep 2007
Dec 2007
Mar 2008
Jun 2008
Sep 2008
Dec 2008

When are line charts appropriate?

- When plotting a Q-R vs. a Q-I or a Q-R
- Mainly makes sense when the X axis is ordered in some way and distances between X axis points matter
  - e.g., is the rate of change in this interval the same as the other interval
- Want to be able to see "trends"
  - There is an assumption of interpolation between points and dependence of the Y-axis on the X-axis
- From a SQL standpoint, the query for generation is similar to bar charts, grouping by the X-axis

Scatterplots

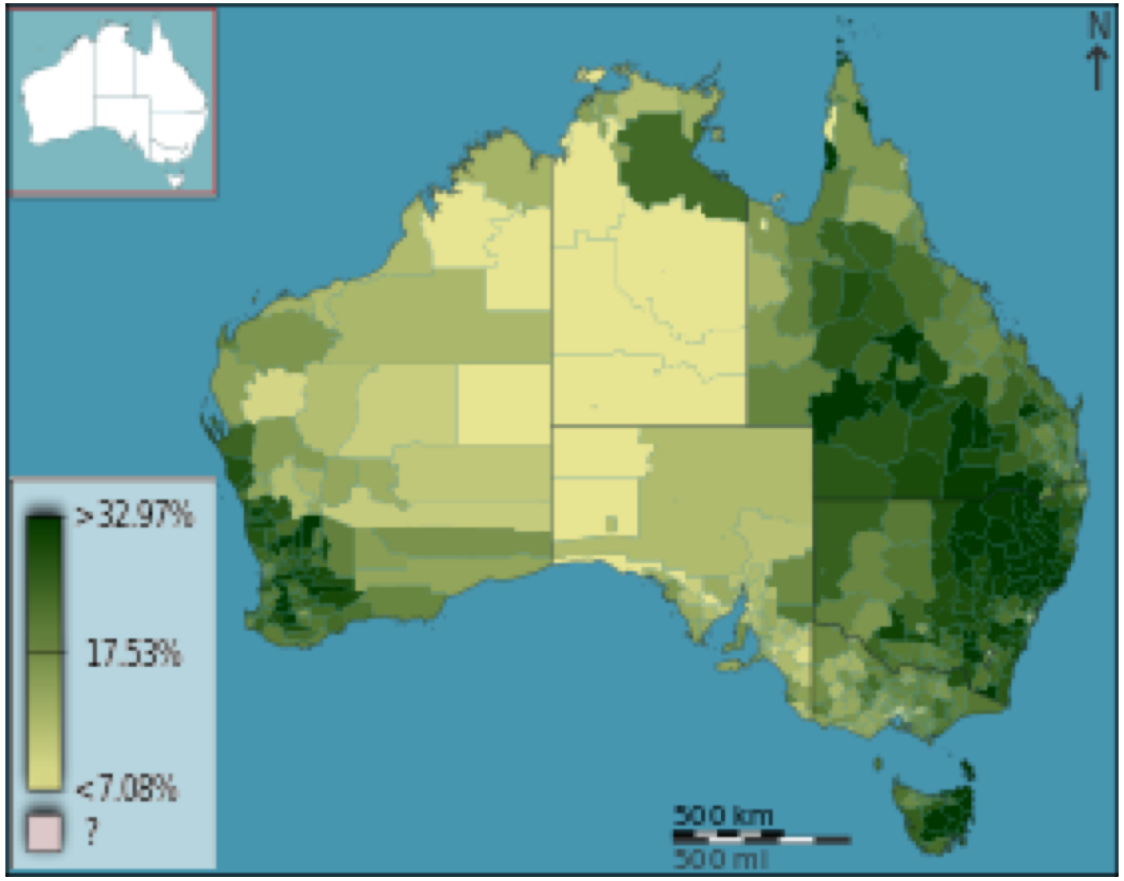




When are scatterplots appropriate?

- When plotting a Q-R vs. a Q-R
- No assumption of interpolation unlike line charts
- Care more about "density", understanding of "correlation"
- From a SQL query standpoint, one way to plot a scatterplot is to simply perform a SELECT X, Y FROM R with no grouping.
  - Additional aspects (e.g., color) can also be selected if needed
- However, there is a danger of too many rows being returned.
  - Imagine a relation of size 1B: 1B pairs returned
  - A safer option in that case is to bin the scatterplot into grid cells
  - Q: How would we do this in SQL?
  - A: CTE to add new "binned" columns corresponding to a CASE statement, followed by a GROUP BY on the new columns

Choropleths



When are choropleths appropriate?

- Choropleths map a Q-R vs. a two-dimensional Q-I variable

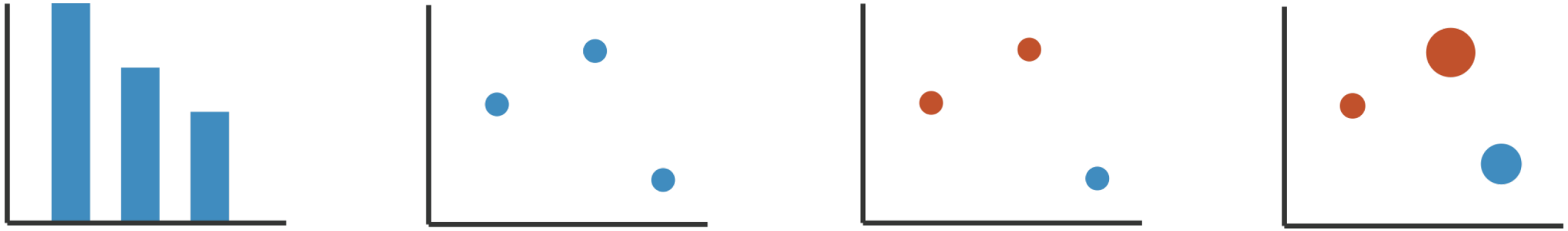
- From a SQL query standpoint, grouping can be done on a per-geographical region basis followed by overlaying on a map.

What type of visualization would you use?

- A plot of rainfall by location on a map
- A plot of average age by department
- A plot of total sales by year
- A plot of rainfall by temperature

We just …

- Saw a bunch of primitive visualization types & relationships to SQL
- But there are lots more variants!



- We need a way to think about visualization types more formally
- And compare between them
- Enter visual encodings!

From Data to Visual Encodings

- Given a dataset, we apply a mapping or visual encoding to transform it into a visualization
- As part of this visual encoding, we select:
  - *Marks*: basic items / geometric primitives
  - *Channels:* visual aspects that change appearance of marks based on values
- This visual encoding process allows us to reason about a variety of visualization types, and compose them "from the bottom up"

Marks

Basic Geometric Elements

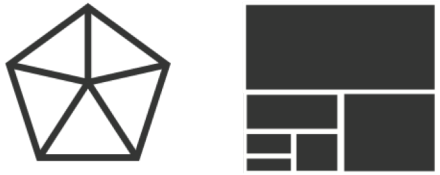
- Points

0D

- Lines

1D

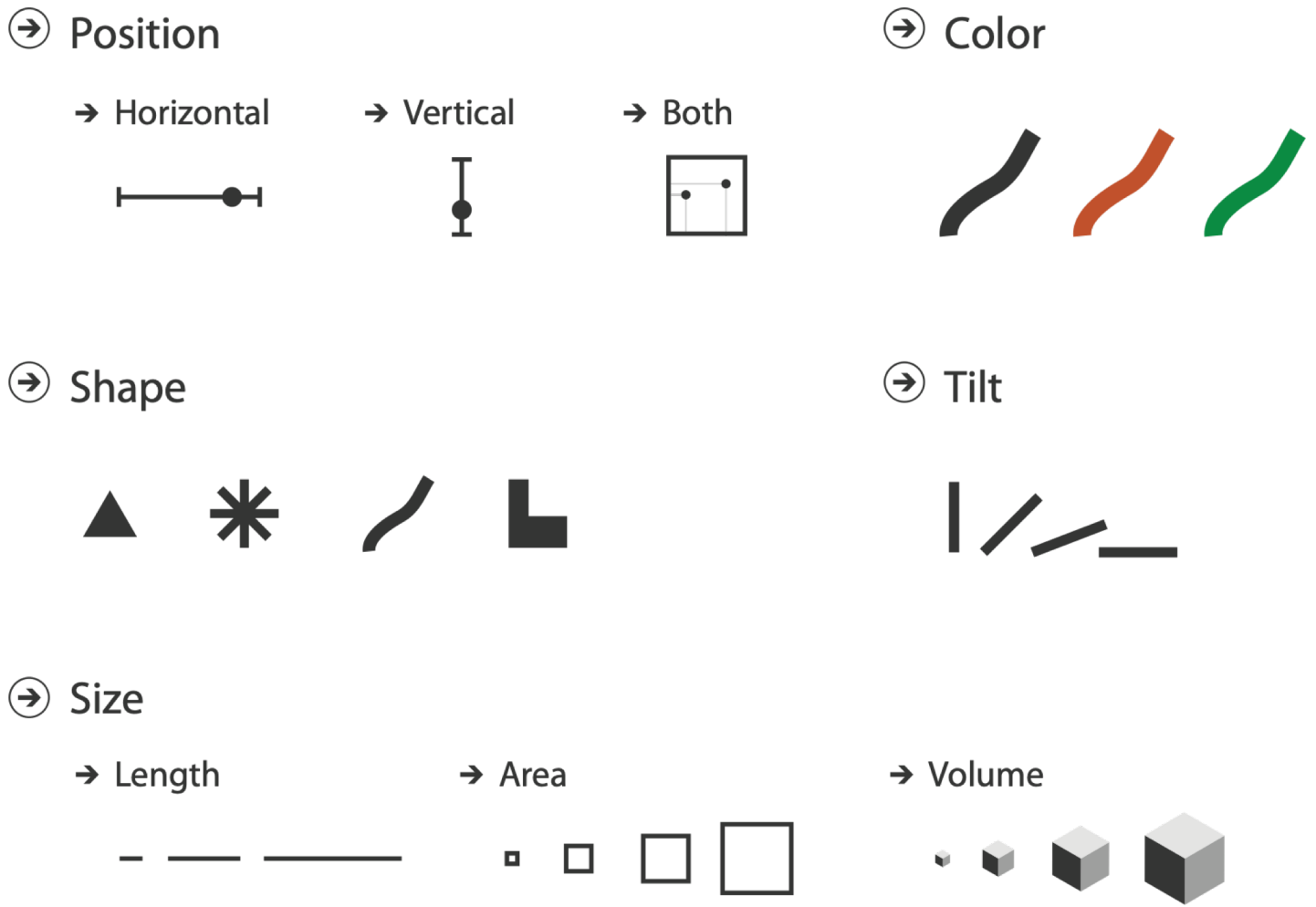
- Interlocking Areas



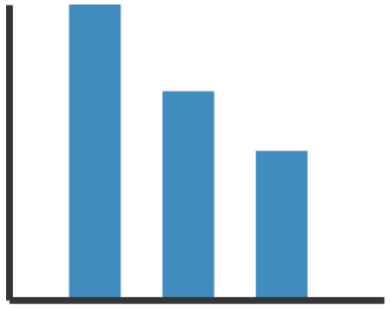
2D

Channels

- control appearance of marks
  - proportional to or based on attributes
- many names
  - **visual channels**
  - visual variables
  - retinal channels
  - visual dimensions
  - ...

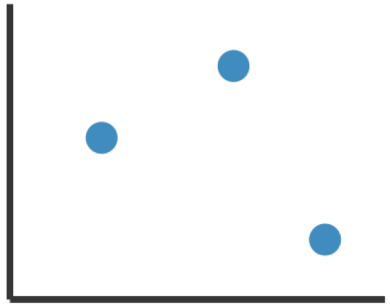


Visual Encodings Example



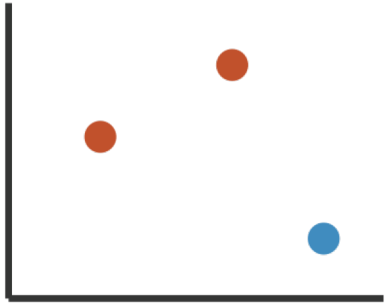
1:
QR: vertical position
N: horizontal position

mark: line



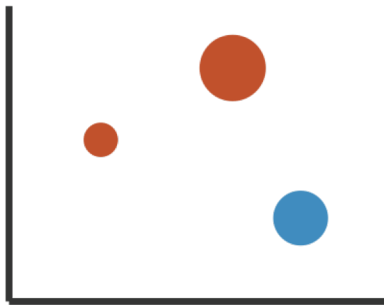
2:
QR: vertical position
QR: horizontal position

mark: point



3:
QR: vertical position
QR: horizontal position
N: color hue

mark: point



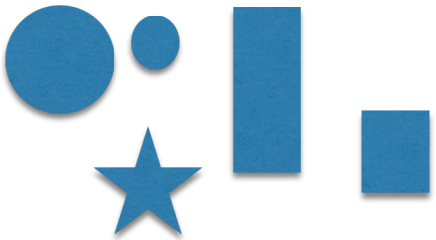
4:
QR: vertical position
QR: horizontal position
N: color hue
QR: size (area)

mark: point

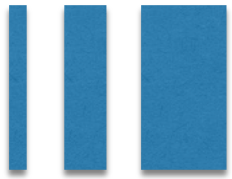
Constraints on Marks → Channels

- Marks have dimensions, so dimensions impose constraints

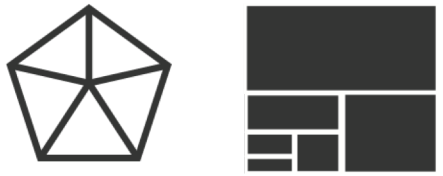
Points



Lines



Interlocking Areas



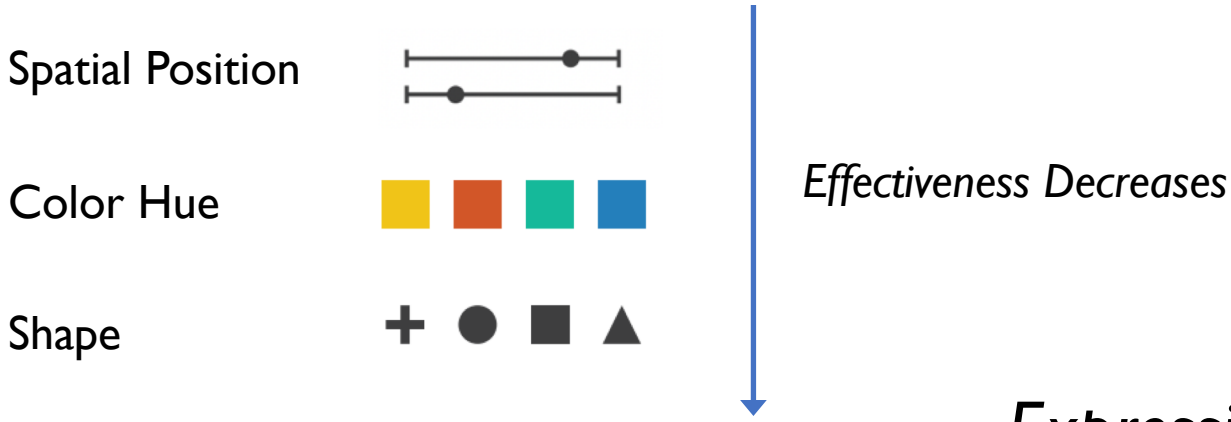
- constraint view: mark type constrains what else can be encoded
  - points: 0 constraints on size, can encode more attributes w/ size & shape
  - lines: 1 constraint on size (length), can still encode size other way (width)
  - interlocking areas: 2 constraints on size (length/width), cannot code for size or shape

When to use which channel

- Expressiveness
  - Match channel type to data type

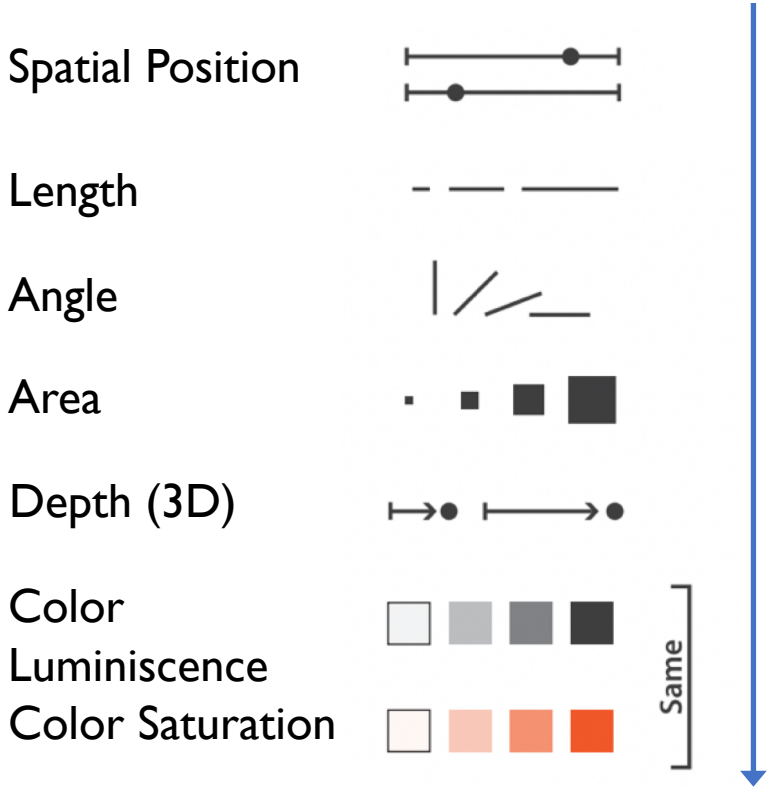
- Effectiveness
  - Some channels are better than others

Nominal Attributes



Expressiveness principle:
Don't use Shape to encode a Quantitative attribute!

Ordinal/Quantitative Attributes



Expressiveness principle:
Don't use Area to encode a Nominal attribute!
(imposing an order on something that isn't ordered)

Visualization Tools

- Many good visualization packages: these help you generate a visualization on your data from within a programming language
  - Matplotlib
  - Plotly
  - D3/Vega/Vega-lite
  - ggplot2
  - Gnuplot
- Usually, compose visualizations "bottom up", starting from the marks, assigning attributes to channels, etc.

- Plus visual analytics tools: these are tools that provide an interactive environment to explore your data visually without writing code
  - Looker
  - PowerBI
  - Spotfire
  - Tableau ← This is the focus of the paper you'll be reading!
- Here, the visual encoding is a bit more automatic, but with users able to override

Takeaways

- Visualization is an essential means for data exploration — hypothesis generation and confirmation, spotting of outliers and trends, among others.
- Data types dictate how the data should be visualized
- A lot can be accomplished with a small number of visualization types: often these suffice during data exploration
- Visual encodings provide a useful way to compose visualizations from the ground up
- Visual analytics tools provide interactive visualization capabilities via simple interactions