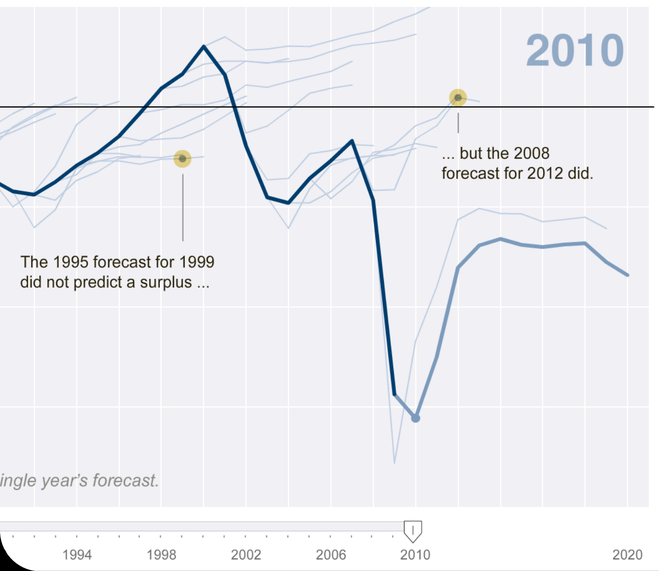
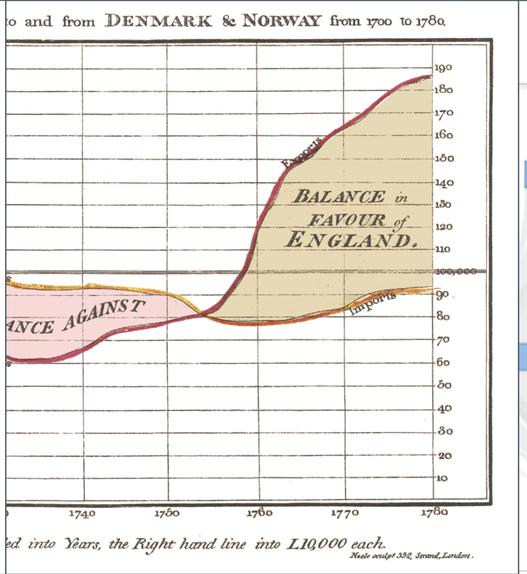
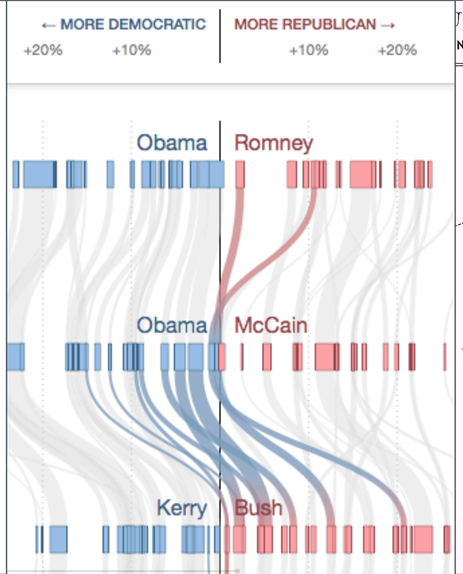
6.894: Interactive Data Visualization

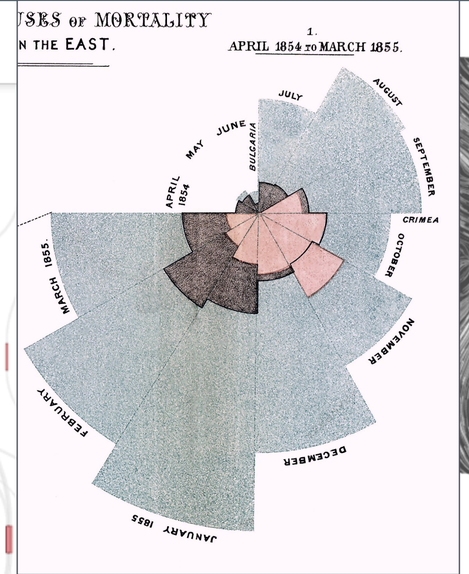
Data & Image Models

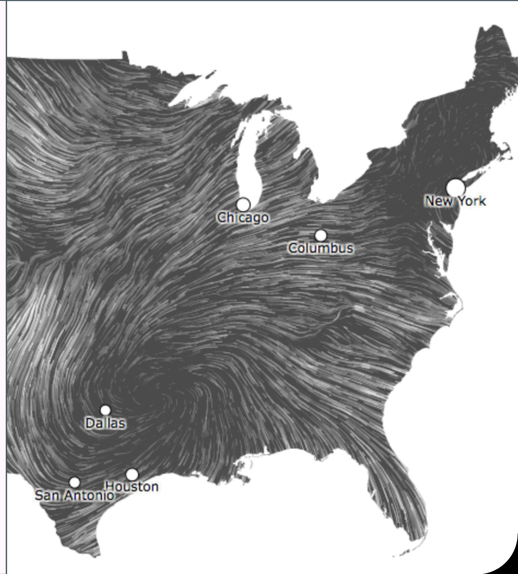
Arvind Satyanarayan











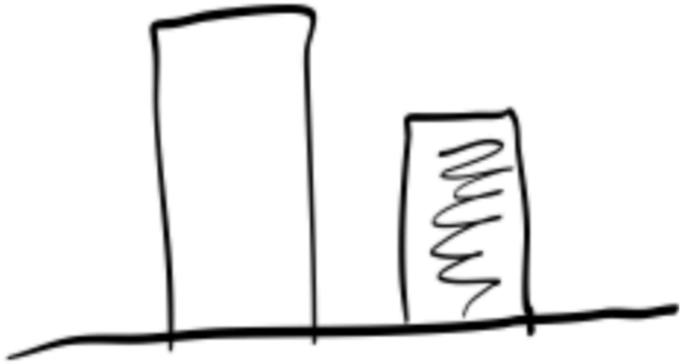
Activity!

In **1 minute**, sketch as many visualizations as possible of these two numbers:

75 37

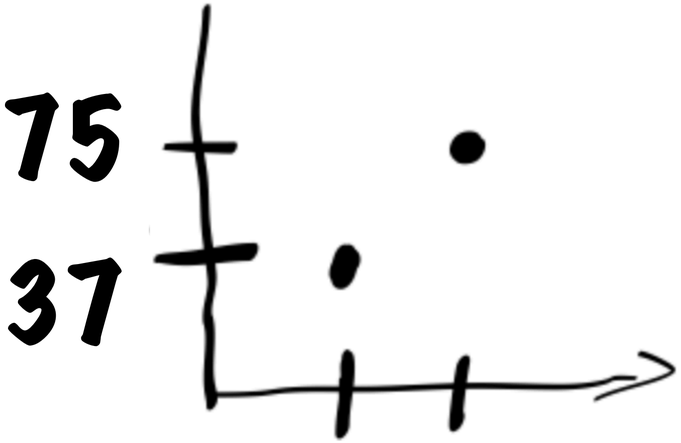
Most Likely Results

Pie Charts



Bar Charts

Scatterplot



75

37

Arabic Numbers

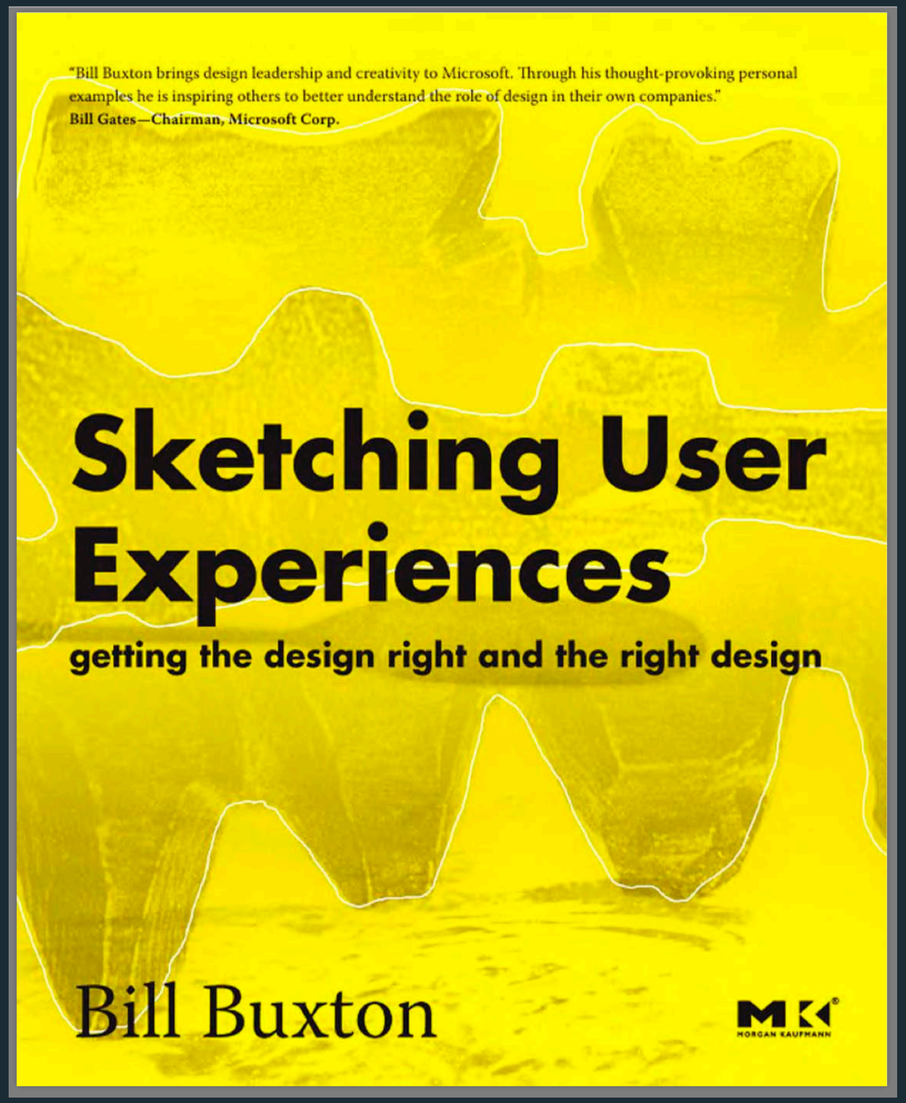
Design Fixation

"A blind adherence to a set of ideas or concepts limiting the output of conceptual design" [Jansson & Smith 1991]

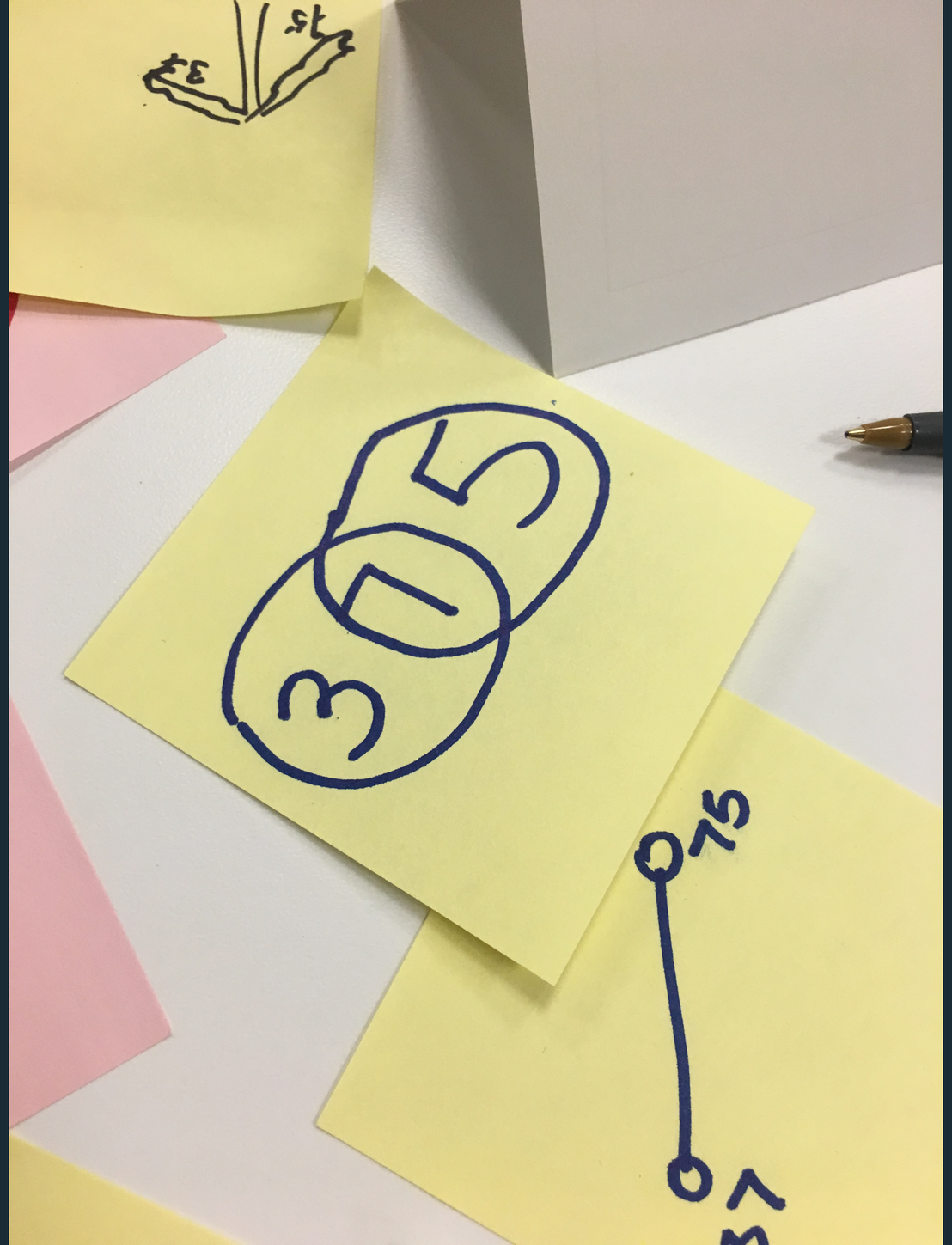
To overcome fixation:

sketch: quick, inexpensive, disposable ways of generating, evaluating, and sharing ideas [Buxton 2007]

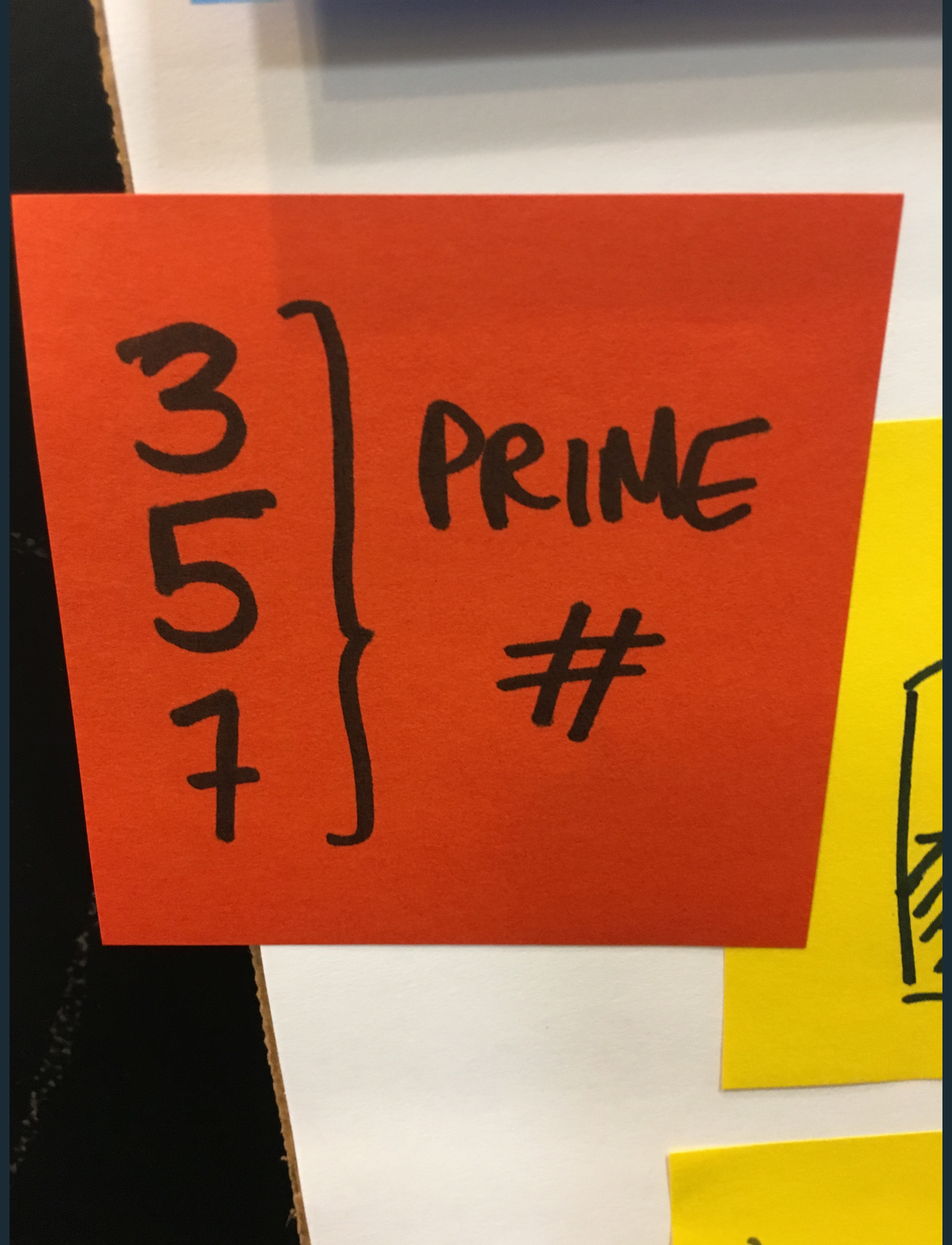
consult examples: early and repeated exposure to examples improves creativity [Kulkarni 2012]



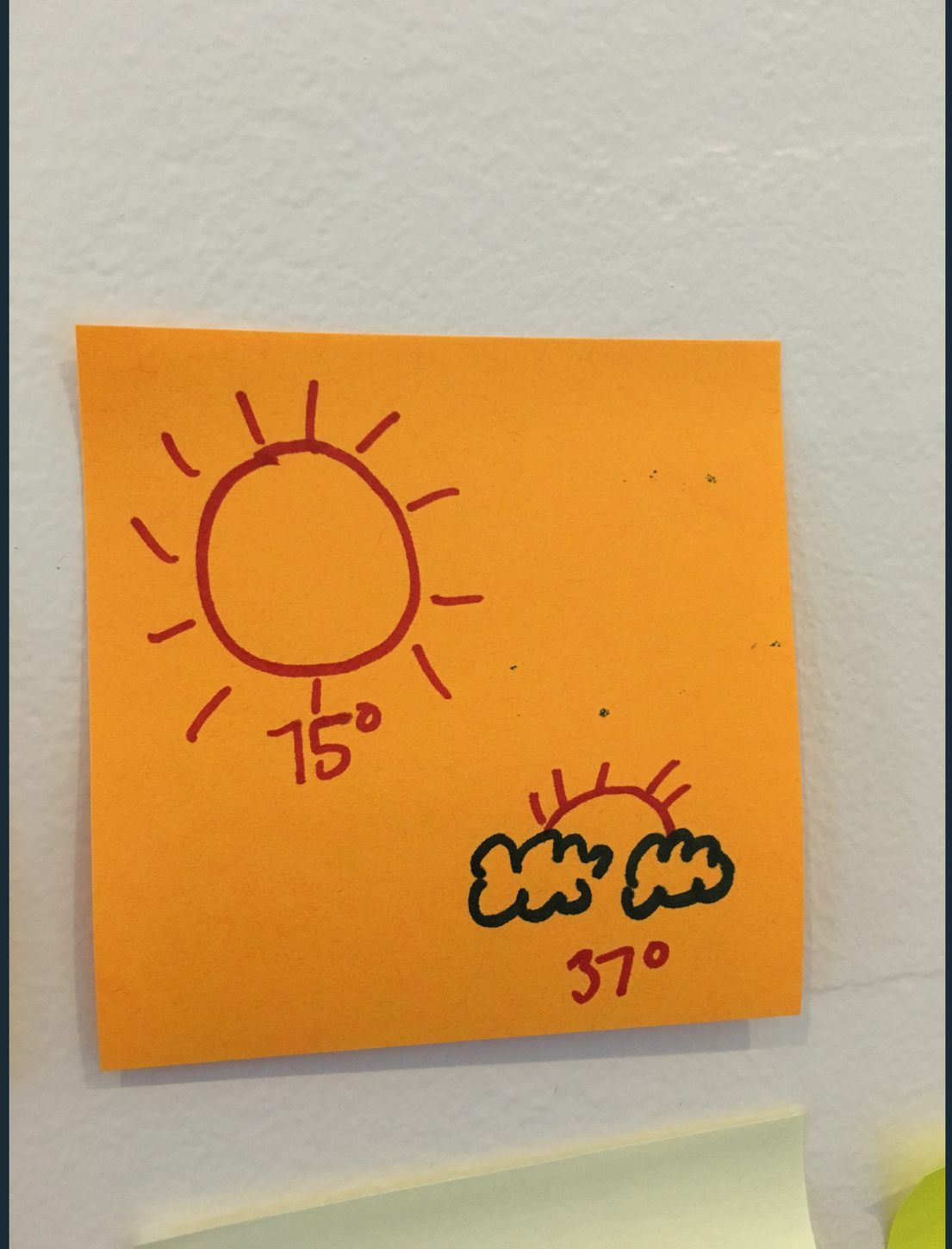
Examples from Jon Schwabish.



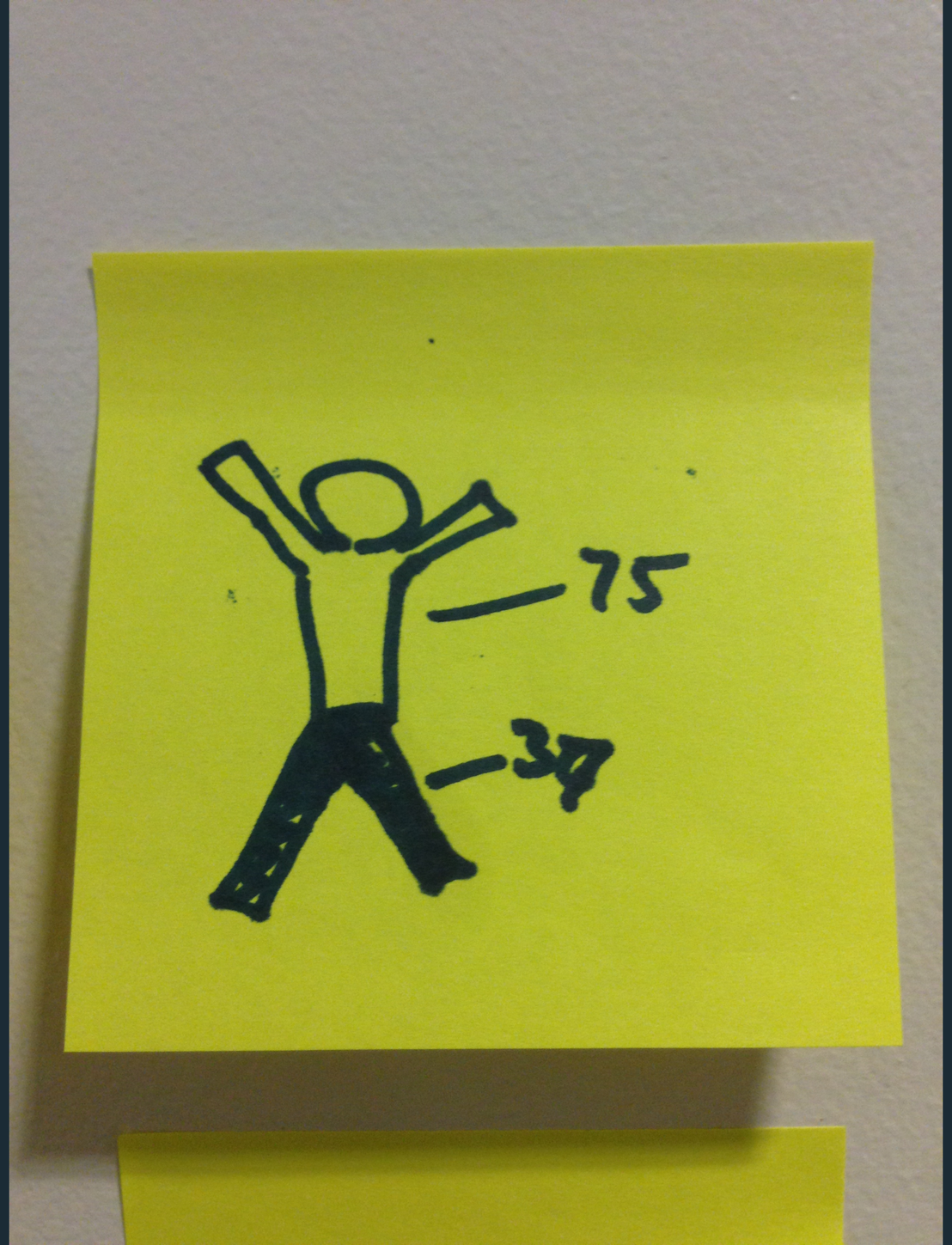
Examples from Jon Schwabish.



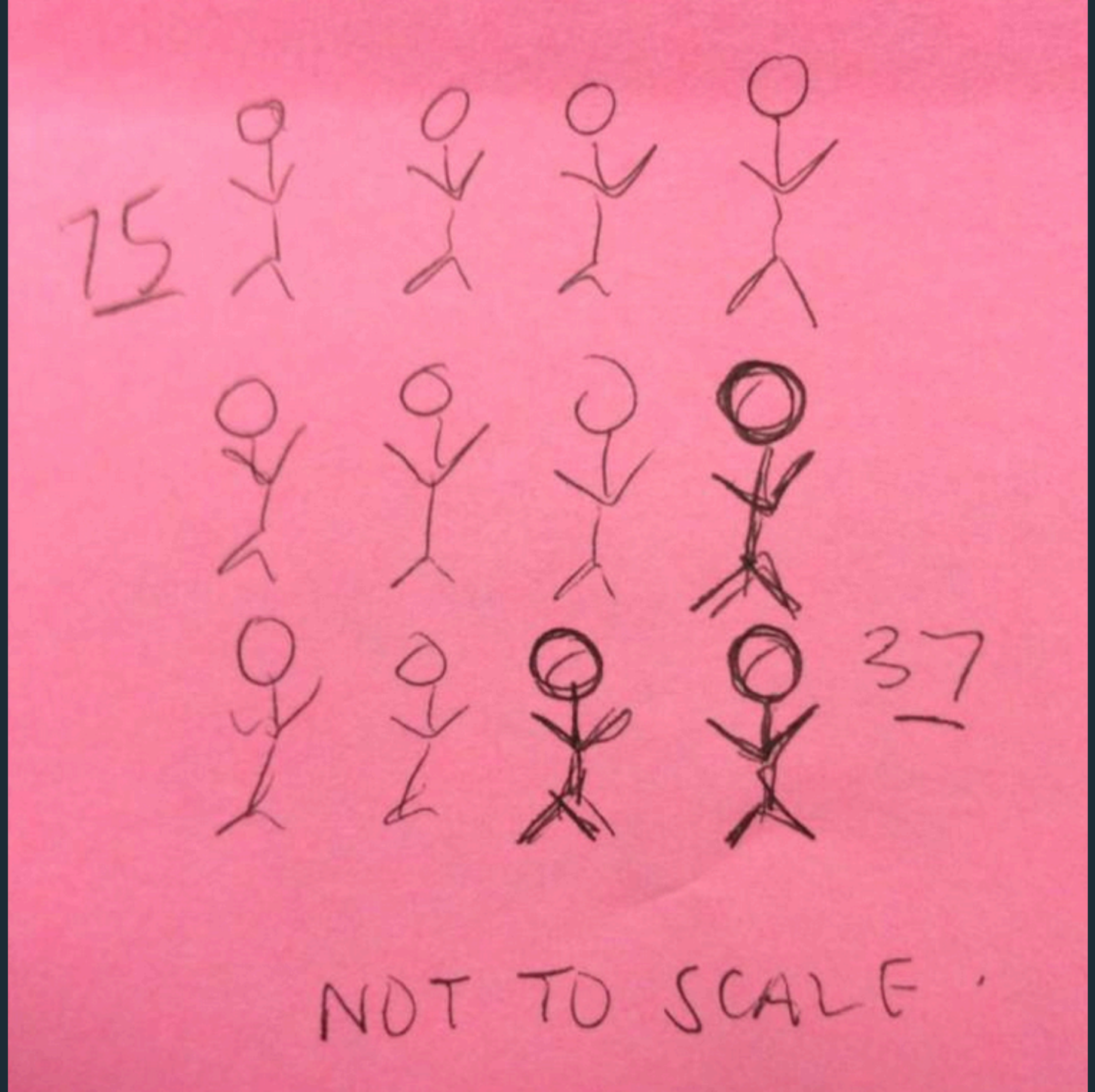
Examples from Jon Schwabish.



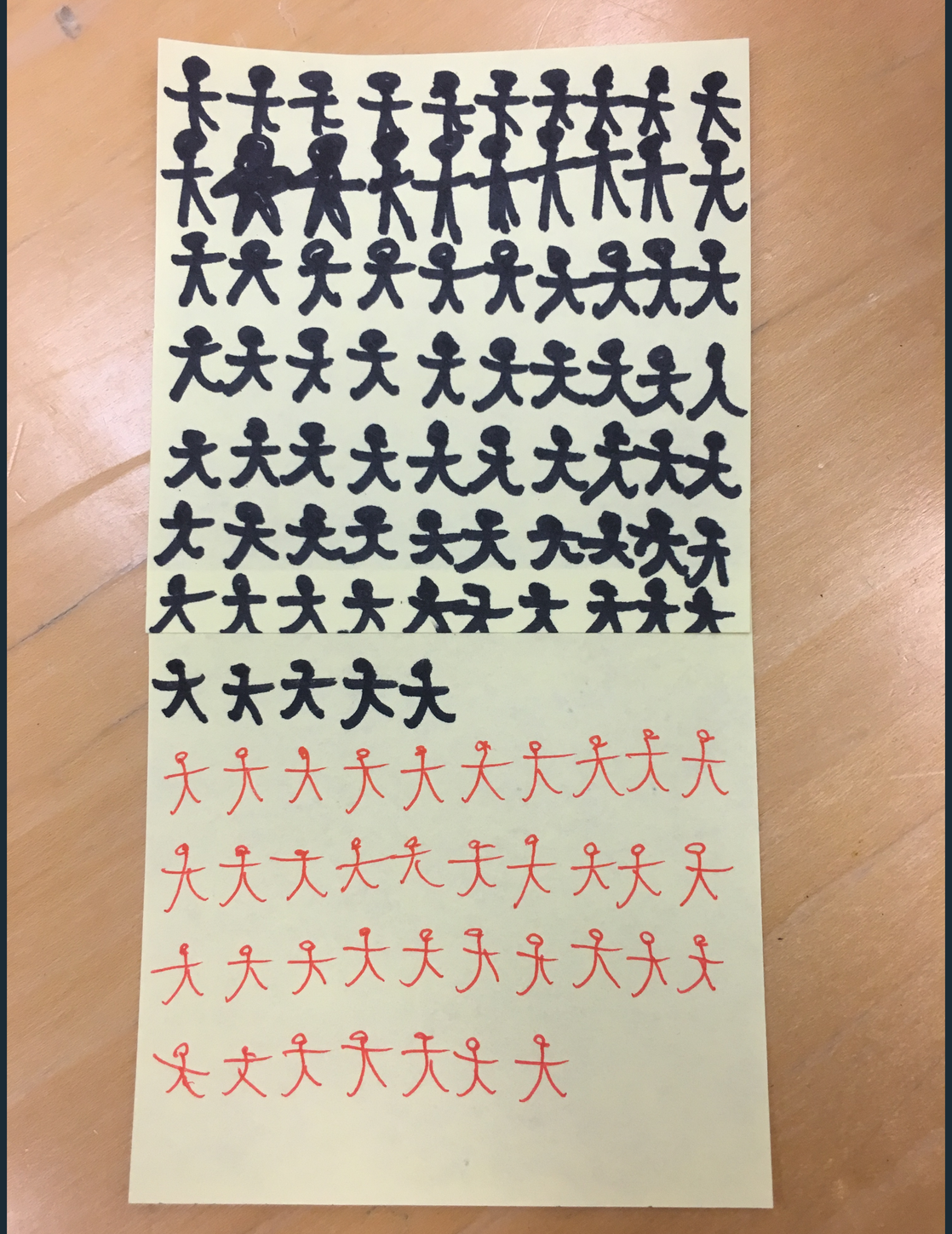
Examples from Jon Schwabish.



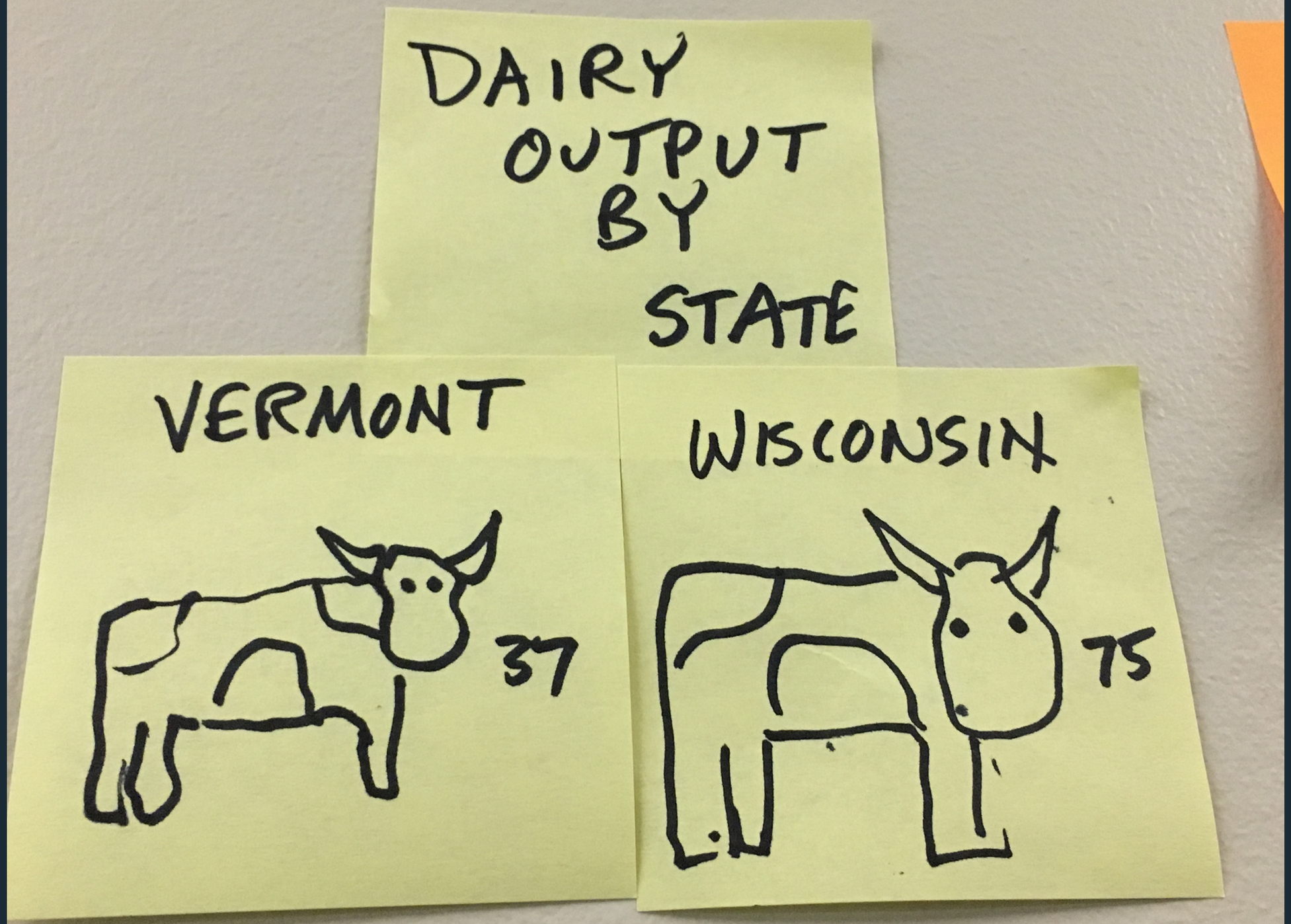
Examples from Jon Schwabish.



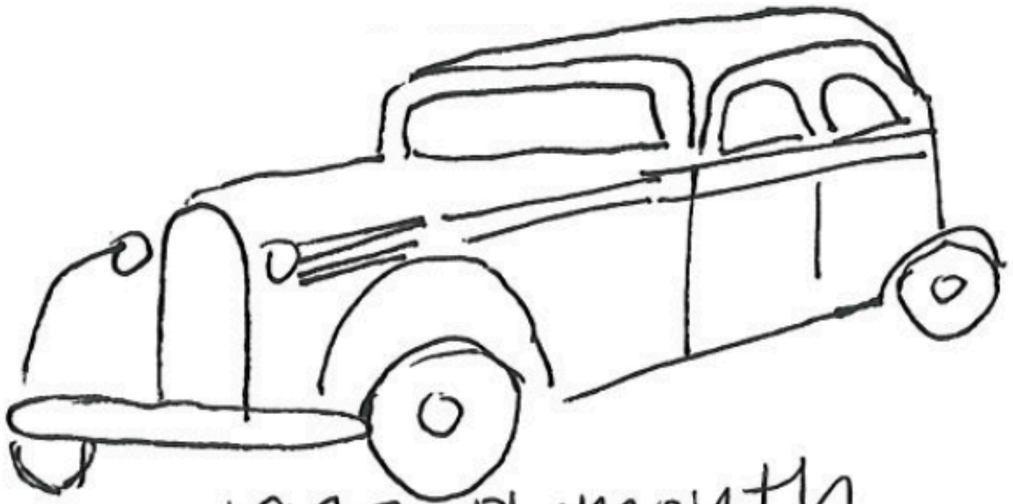
Examples from Jon Schwabish.



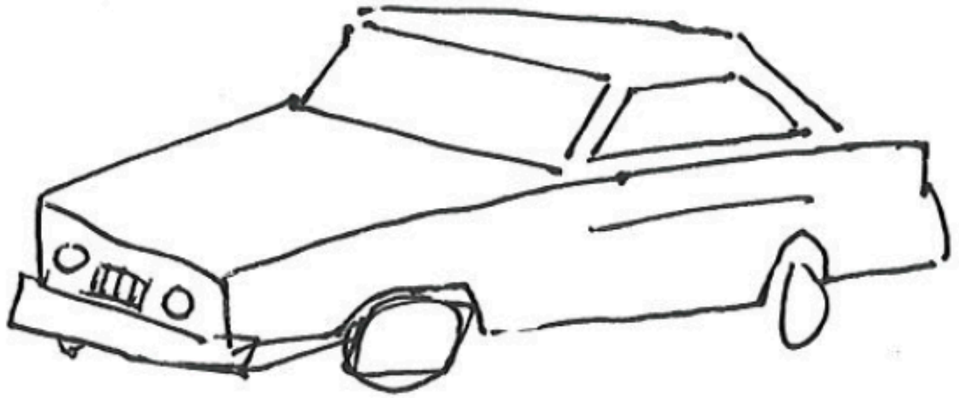
Examples from Jon Schwabish.



Examples from Jon Schwabish.



1937 Plymouth



1975 Plymouth

Design Fixation

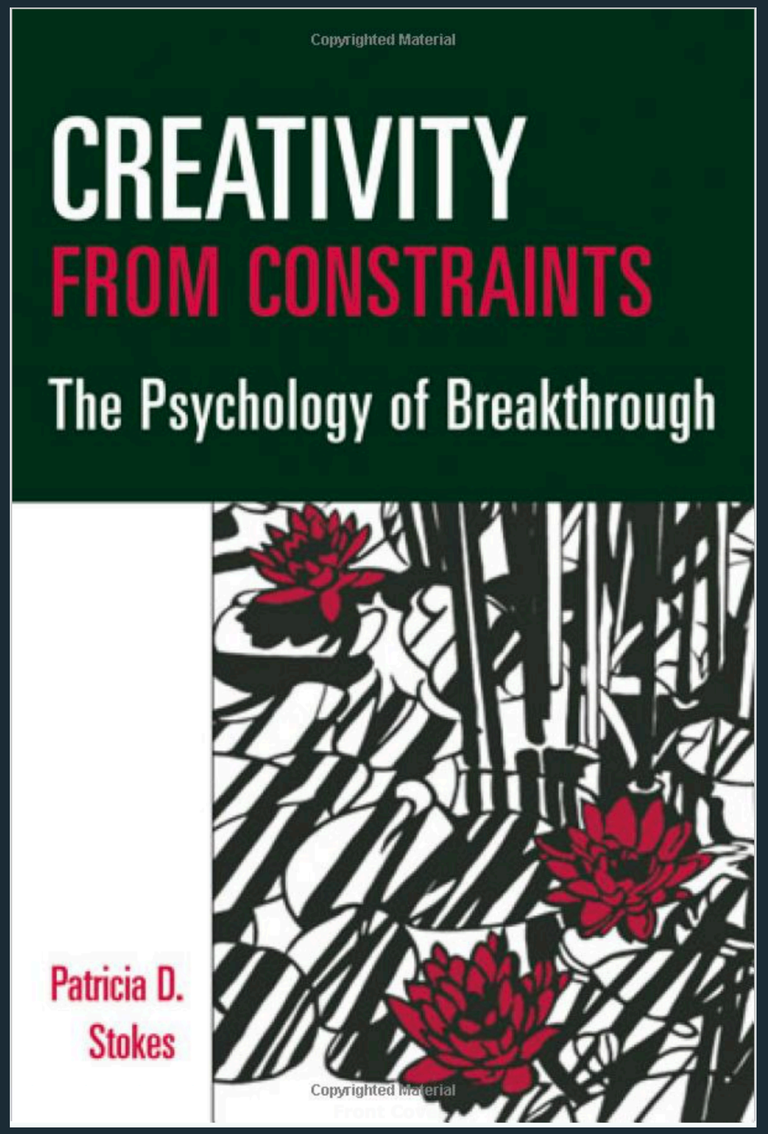
"A blind adherence to a set of ideas or concepts limiting the output of conceptual design" [Jansson & Smith 1991]

To overcome fixation:

sketch: quick, inexpensive, disposable ways of generating, evaluating, and sharing ideas [Buxton 2007]

consult examples: early and repeated exposure to examples improves creativity [Kulkarni 2012]

introduce a constraint: impose new structures to the problem to spur creativity [Stokes 2006]



Activity!

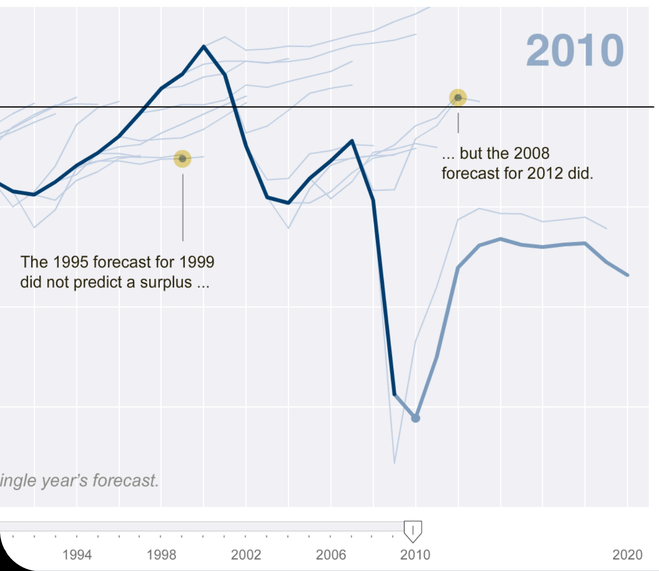
In **5 minutes**, sketch as many **new visualizations** as possible that are different from your previous ideas. If you're stuck, introduce a constraint -- e.g., one line, only black/white, only round objects, etc.

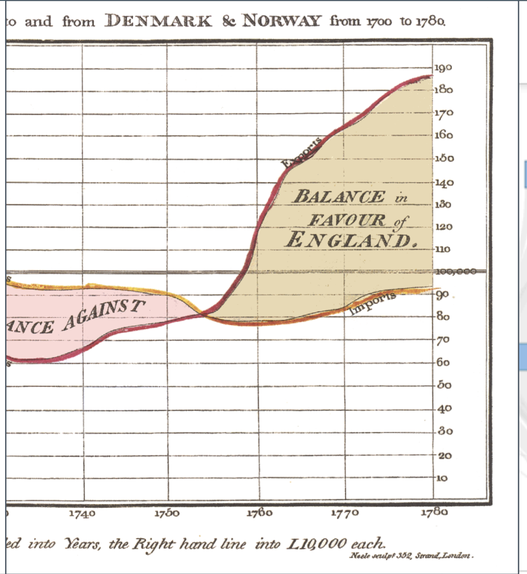
75 37

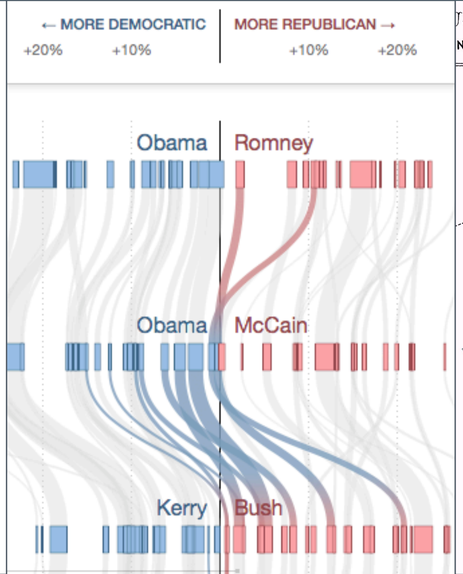
6.894: Interactive Data Visualization

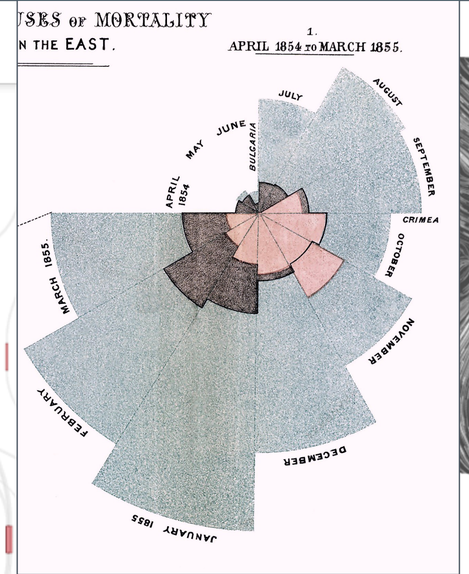
Data & Image Models

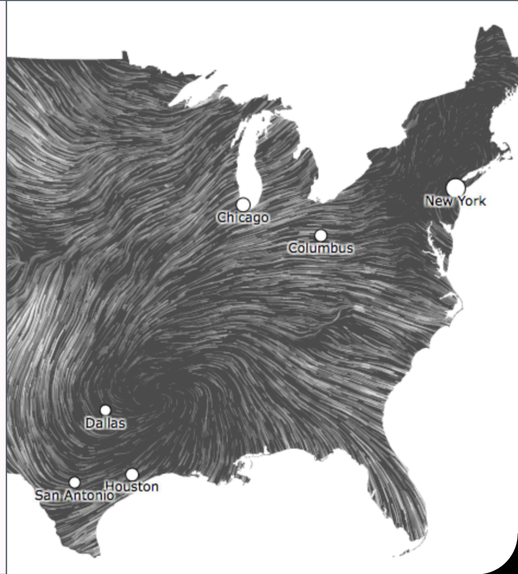
Arvind Satyanarayan











Data Visualization

Mapping or Visual Encoding

Data

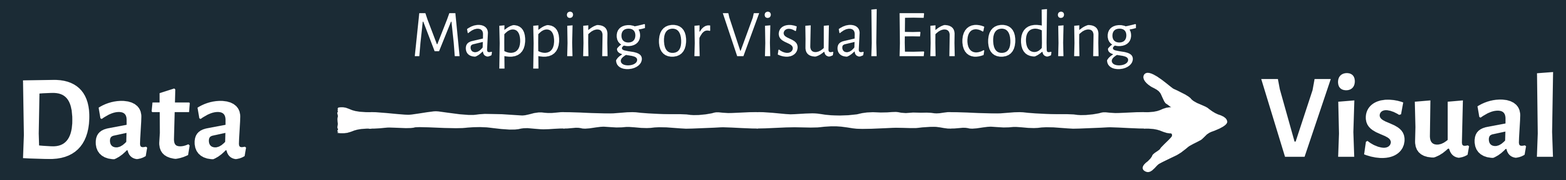
Physical Data Types
`int, float, string`

Conceptual Data Types
`temperature, location`

Visual

Visual **Channels**
`x, y, color, opacity`

Graphical **Marks**
`rect, line, point, area`



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.

[Mackinlay 1986]

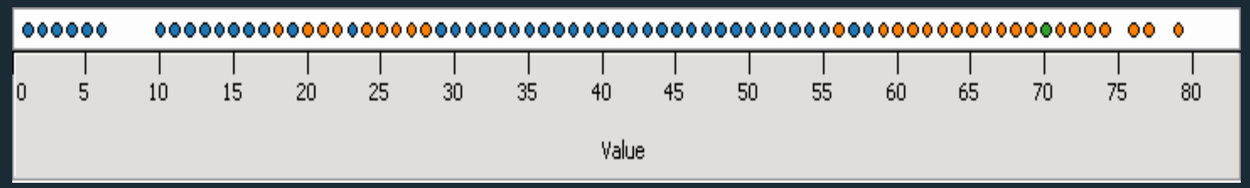
Mapping or Visual Encoding

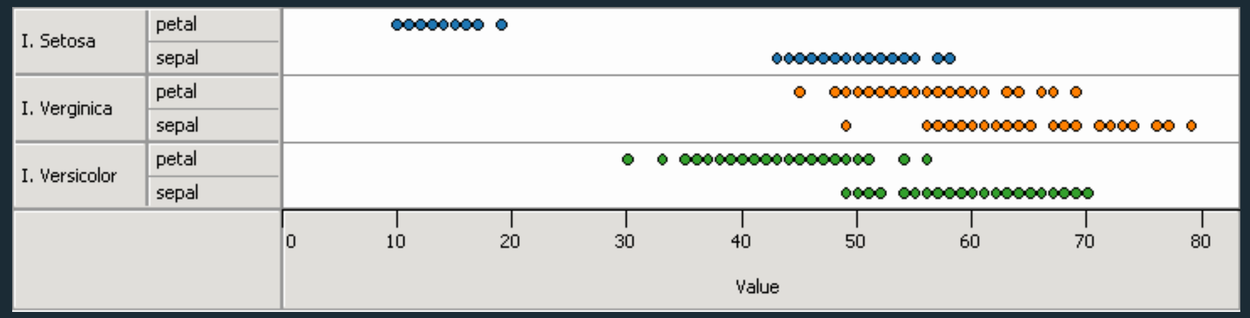
Data → **Visual**

Expressiveness

Cannot express the facts

A multivariate dataset may be *inexpressive* in a single horizontal dot plot because multiple records are mapped to the same position.





Mapping or Visual Encoding

Data → **Visual**

Expressiveness

Express facts not in the data

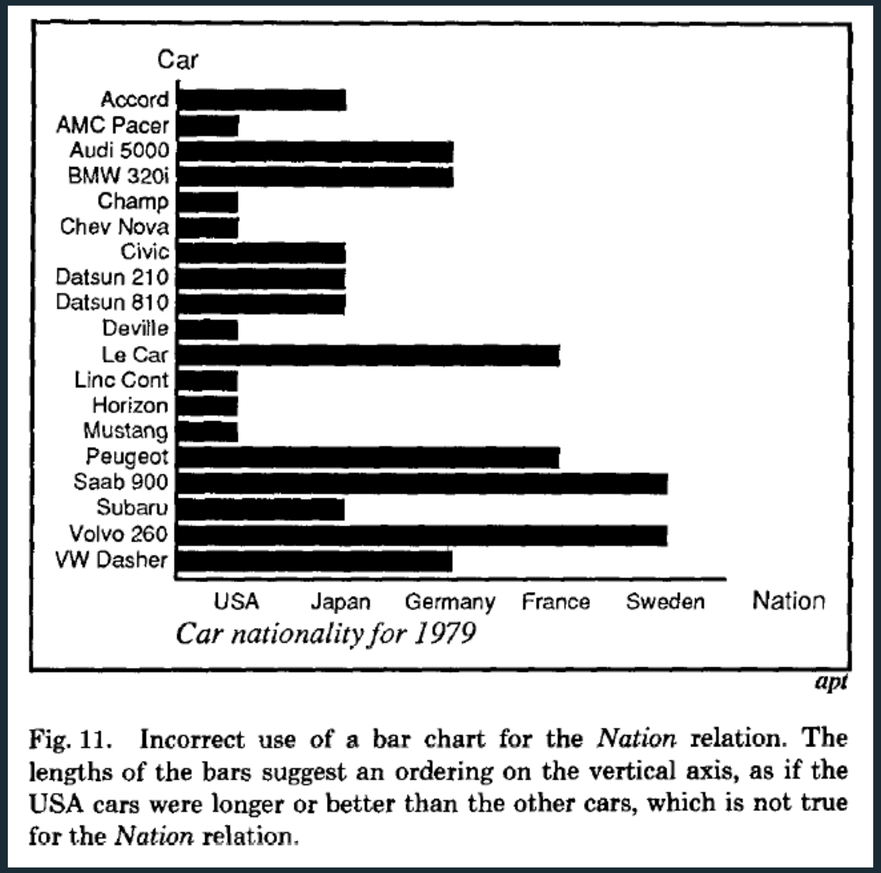
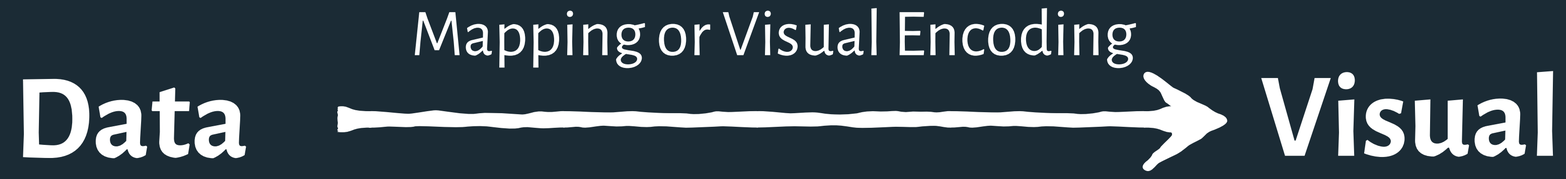


Fig. 11. Incorrect use of a bar chart for the *Nation* relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the *Nation* relation.



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*.

Data models give us a way of talking about this.

[Mackinlay 1986]

Data Models

Conceptual Models vs Data Models

By "default", data is described in terms of a specific *domain*.

- E.g., The average amount of *rain* or *snow* in different *towns, cities, countries.*
- E.g., *friends, followers, connections* depending on the social network (or *citations* in academia!).

To effectively map data to visuals, we need a level of *abstraction*.

- Data abstraction allows us to consistently encode the same "types" of data, even if different domains use different terminology to describe it.

Dataset Types

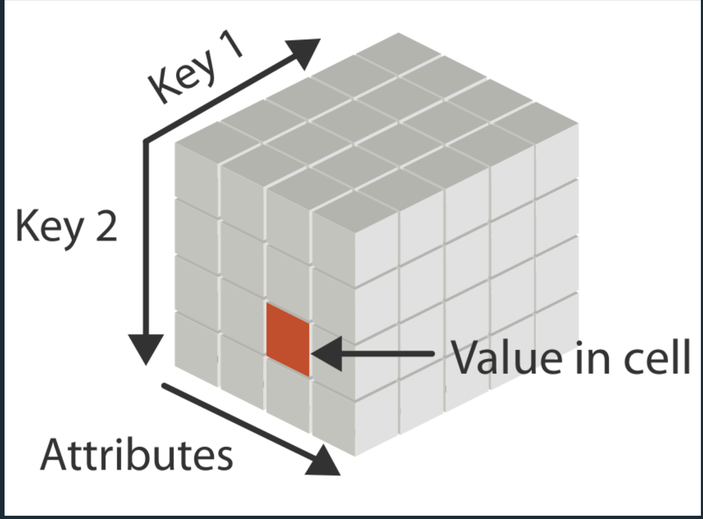
1. Tabular

rows/records/items

columns/attributes/variables

	A	B	C	D	E	F
1	**EmployerName**	**Address**	**DiffMeanHourlyPercent**	**DiffMeanBonusPercent**	**MaleBonusPercent**	**FemaleBonusPercent**
2	1ST CHOICE STAFF RECRUITMENT LIMITED	8, St. Loyes Street, Bedford, MK40 1EP	-4.5	206.9	2	1
3	23.5 DEGREES LIMITED	Charles Watts Way, Hedge End, Southampton,	10	79	4	3
4	A. & B. GLASS COMPANY LIMITED	Chilton Industrial Estate, Sudbury, Suffolk,	15	85	61	32
5	ABACUS HOTELS LIMITED	20 Station Street, Swaffham, Norfolk,	37.8	-6.6	19.2	16.2
6	Abbeyfield Wales Society	24 Gold Tops, Newport, NP20 4PG	21.9	0	0	0
7	ABERDEEN JOURNALS LIMITED	Mastrick, Aberdeen, United Kingdom,	15.7	44.7	17.1	39.7
8	ACCESSIBLE TRANSPORT GROUP CONTRACT SERVICES LIMITED	Birmingham, West Midlands, United Kingdom,		0	0	0
9	ACEGOLD LIMITED	Norcliffe House, Station Road, Wilmslow, SK9 1BU	-5.1	0	0	0
10	Acorns Children's Hospice Trust	Wythall, Birmingham, United Kingdom,	11.2	0	0	0
11	AD Astra Academy Trust	Davison Drive, Hartlepool, Cleveland,	9.5	0	0	0
12	ADAPT BUSINESS SERVICES LIMITED	Drive, Gorseinon, Swansea, SA4 4QN	3.3	0	0	0
13	ADARE INTERNATIONAL LIMITED	Two Colton Square, Leicester, England,	18.8	71.3	11.6	10.5

cell containing value



Tamara Munzner, *Visualization Analysis and Design* (2014).

https://gender-pay-gap.service.gov.uk

Dataset Types

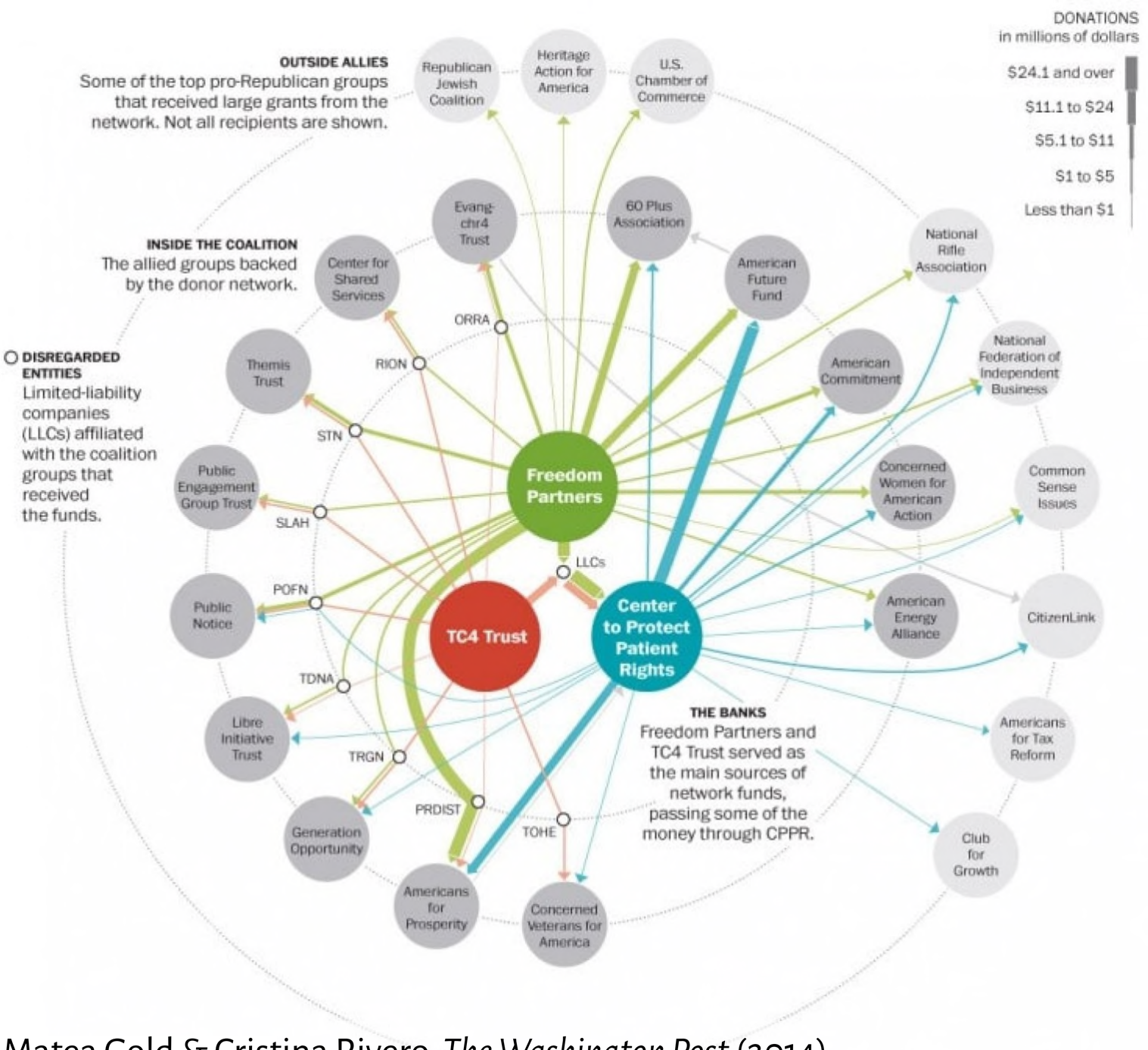
1. **Tabular**

   A collection of records with named attributes.

2. **Networks**

   Nodes and links can also have attributes (e.g., size of nodes, thickness/directionality of links).

   Trees are special networks where each node has only one parent.



Matea Gold & Cristina Rivero, *The Washington Post* (2014).

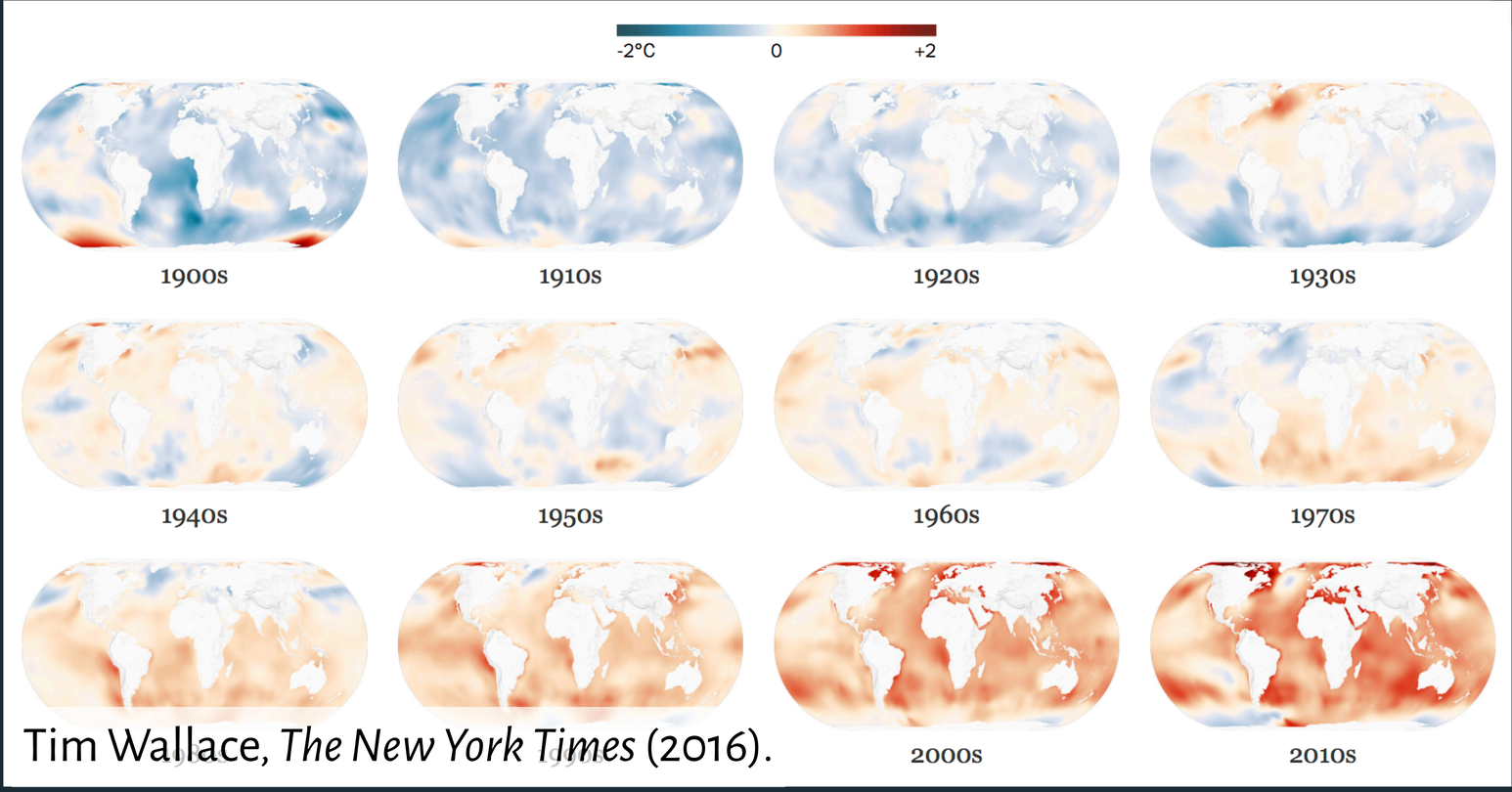
Dataset Types

1. **Tabular**
   A collection of records with named attributes.

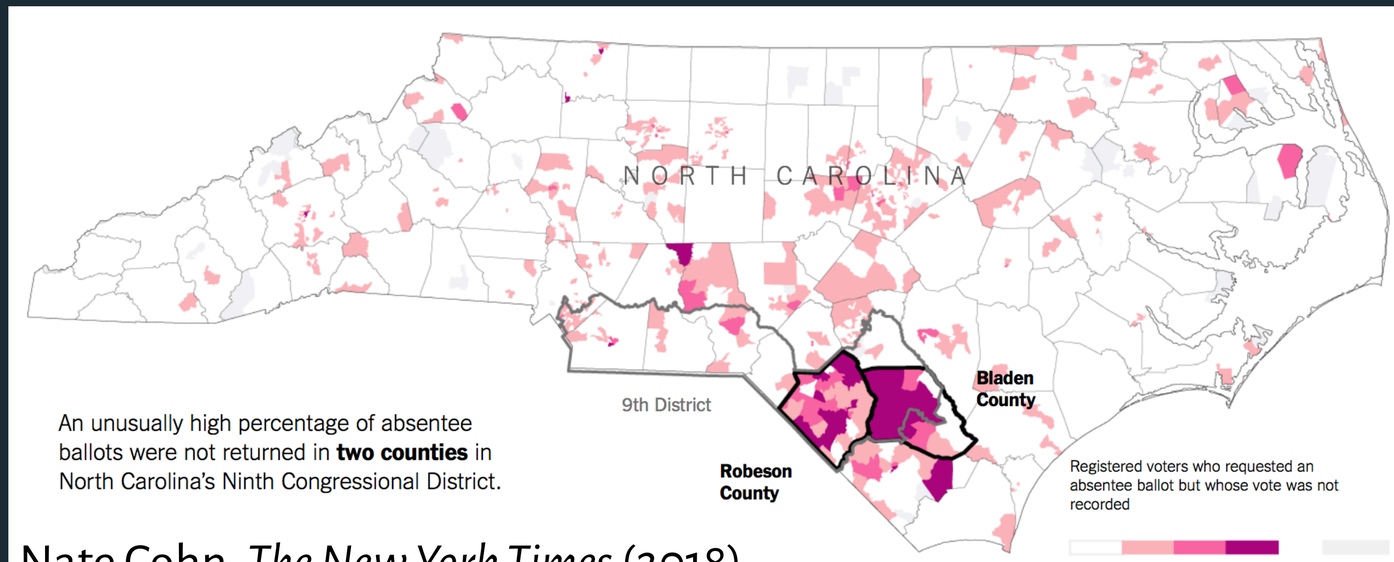
2. **Networks**
   Nodes and links can also have attributes (e.g., size of nodes, thickness/directionality of links).

   Trees are special networks where each node has only one parent.

3. **Spatial**
   Continuous "fields" vs discrete "positions"



Tim Wallace, *The New York Times* (2016).



Nate Cohn, *The New York Times* (2018).

Attribute Types

Dimensions

~ Independent variables.

Ways of describing the data, often discrete.

E.g., categories, dates, binned quantities.

Can include numerical data, but doesn't make sense to aggregate.

Measures

~ Dependent variables (i.e., their value is a function of one or more dimensions).

Numerical data that can be analyzed and aggregated.

Aggregations including sum, count, avg, std. dev, etc.

Tableau - World Indicators [Read-Only] - Tableau license expires in 14 days

Fit Width | Show Me

Data | Analytics

World Indicators

Dimensions

- Above Threshold?
- Birth Rate Bin
- Country
- Ease of Business (clusters)
- GDP per Capita (bin)
- Region

Measures

- Business
  - Business Tax Rate
  - Days to Start Business
  - Ease of Business
  - Hours to do Tax
  - Lending Interest
- Development
  - CO2 Emissions
  - Energy Usage
  - GDP
  - Internet Usage
  - Mobile Phone Usage
  - Tourism Inbound
  - Tourism Outbound
- Health
  - Health Exp % GDP
  - Health Exp/Capita
  - Infant Mortality Rate
  - Life Expectancy
  - Life Expectancy Female
  - Life Expectancy Male
- Population
  - Birth Rate

Parameters

- Health Exp Threshold

Pages

Filters
- Measure Names
- Country

Marks
Square
Color | Size | Label | Detail | Tooltip
- Measure Values
- AVG(Birth Rate)
- AVG(Infant Mortal..
- AVG(Health Exp ..
- AVG(Life Expecta..
- Region
- Measure Values

Measure Values
- AVG(Birth Rate)
- AVG(Infant Mortality Rat..
- AVG(Health Exp % GDP)
- AVG(Life Expectancy)

Columns: Measure Names

Rows: Country

Health Indicators

	Birth Rate	Infant Mortality	Health Exp % GDP	Life Expectancy
Ethiopia	3.8%	0.07	4.4%	58
Malawi	4.2%	0.07	7.4%	50
Eritrea	3.9%	0.05	3.3%	59
Congo, Dem. Rep.	4.6%	0.10	5.8%	48
Niger	5.1%	0.08	7.2%	55
Madagascar	3.7%	0.05	4.7%	62
Mozambique	4.3%	0.09	6.1%	48
Rwanda	3.8%	0.07	8.0%	56
Uganda	4.6%	0.07	8.4%	54
Afghanistan	4.3%	0.08	9.2%	58
Sierra Leone	4.1%	0.13	15.6%	42
Central African Republic	3.7%	0.11	4.1%	46
Guinea	4.0%	0.08	5.8%	53
Tanzania	4.1%	0.06	5.1%	55
Nepal	2.7%	0.05	5.7%	65
Togo	3.8%	0.07	6.5%	55
Guinea-Bissau	4.0%	0.09	5.9%	53
Burkina Faso	4.4%	0.08	6.5%	53
Tajikistan	3.1%	0.05	5.1%	66
Bangladesh	2.3%	0.05	3.3%	68
Mali	4.8%	0.10	6.5%	52
Gambia, The	4.4%	0.06	4.4%	57
Haiti	2.8%	0.07	5.9%	60
Zimbabwe	3.2%	0.06		48
Benin	4.0%	0.07	4.5%	57
Cambodia	2.6%	0.05	6.1%	68
Kenya	3.8%	0.06	4.5%	56
Chad	4.9%	0.10	4.8%	48
Timor-Leste	3.8%	0.06	6.9%	64
Kyrgyz Republic	2.3%	0.03	6.0%	69
Comoros	3.8%	0.07	4.0%	59

Birth Rate
0.8% 5.1%

Infant Mortality Rate
0.00 0.13

Health Exp % GDP
2.2% 18.0%

Life Expectancy
42 82

Data Source | Population | Health Indicators | Care Spend | Technology | Economy | Tourism | Business | Global Indicators

788 marks | 197 rows by 4 columns | SUM of Measure Values: 13,448.53

Attribute Types

Nominal
=, ≠

Labels or categories.
E.g., Fruits: apples, bananas, cantaloupes, ...

Ordinal
=, ≠, <, >

Ordered.
E.g., Quality of meat: Grade A, AA, AAA

Quantitative (Interval)
=, ≠, <, >, −

Interval (zero can be arbitrarily located).
E.g., Dates: Jan 19, 2018; Location: (Lat 42.36, -71.09)
Only differences can be calculated (e.g., distances or spans).

Quantitative (Ratio)
=, ≠, <, >, −, %

Ratio (fixed zero).
E.g., Physical measurement: length, mass, temperature
Counts and amounts. Can measure ratios or proportions.

Data Models

Physical Model

`32.5, 54.0, –17.3, ...`

Floating point numbers

Conceptual Model

Temperature (°C)

Attribute Type

Burned vs. Not-Burned (N)

Hot, Warm, Cold (O)

Temperature Value (Q)

Activity: U.S. Census

What are the types of these attributes (N/O/Q and dimension/measure)? Think (~1 min), Pair (~2 mins), Share.

People Count: # of people in group

Year: 1850 – 2000 (every decade)

Age: 0 – 90+

Sex: Male, Female

Marital Status: Single, Married, Divorced, …

	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

Activity: U.S. Census

What are the types of these attributes (N/O/Q and dimension/measure)?
Think (~1 min), Pair (~2 mins), Share.

People Count: Q-Ratio

Year: Q-Interval or Ordinal

Age: Q-Ratio or Ordinal

Sex: Nominal

Marital Status: Nominal

	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

Activity: U.S. Census

What are the types of these attributes (N/O/Q and dimension/measure)? Think (~1 min), Pair (~2 mins), Share.

People Count: Measure

Year: Dimension

Age: Depends!

Sex: Dimension

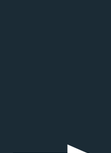
Marital Status: Dimension

	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

Data Transformation

Sorting

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

Filtering (`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

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

day	stock	price
10/3	AMZN	957.10
10/3	MSFT	74.26
10/4	AMZN	965.45
10/4	MSFT	74.69

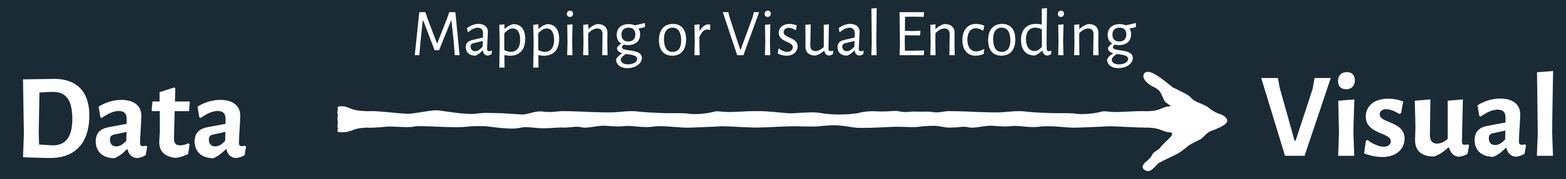
stock	min(price)
AMZN	965.45
MSFT	74.26

Join

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
AMZN	965.45
MSFT	74.26

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



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*.

Data models give us a way of talking about this.

[Mackinlay 1986]

Mapping or Visual Encoding

Data → **Visual**

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*.

Data models give us a way of talking about this.

[Mackinlay 1986]

Effectiveness

A visualization is more *effective* than another if the information it conveys is more readily perceived than the information in the other visualization

Mapping or Visual Encoding

Data → **Visual**

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*.

Data models give us a way of talking about this.

[Mackinlay 1986]

Effectiveness

A visualization is more *effective* than another if the information it conveys *is more readily perceived* than the information in the other visualization

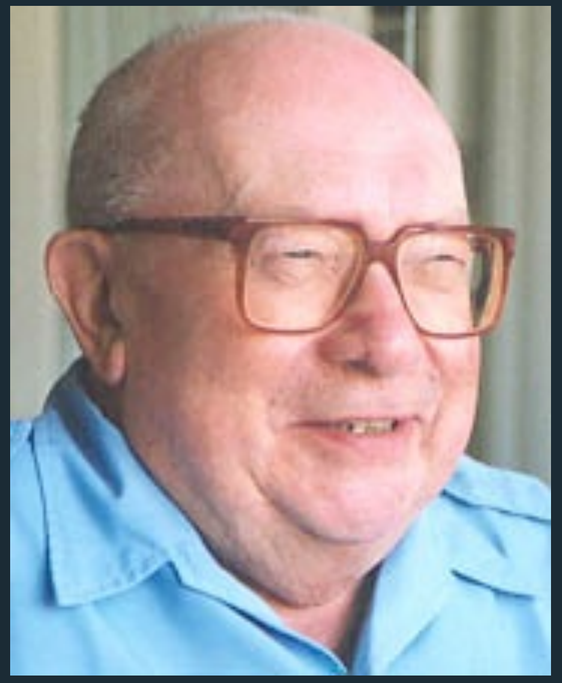
Image models give us a way of talking about this.

Image Models

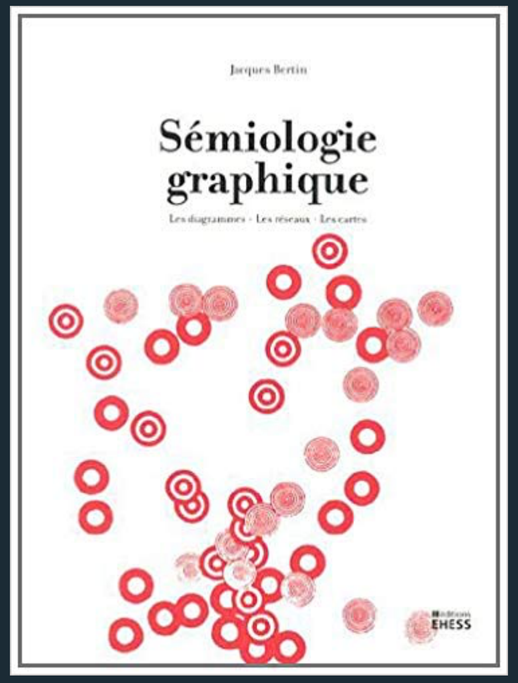
The Semiology of Graphics (1967)

*Study of **signs** and how cultures use them.*

Anything that stands for something other than itself.



Jacques Bertin (1918 – 2010)
French cartographer

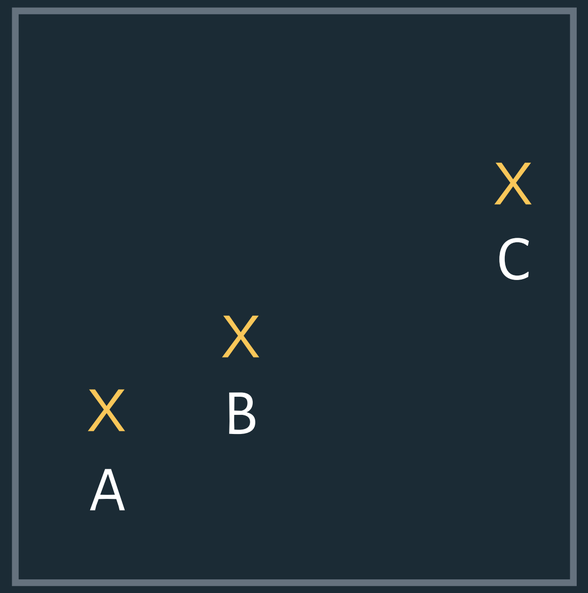


Images are perceived as a set of signs.

Sender encodes information in signs.

Through visual perception, the receiver decodes the signs for information:

1. What are the elements in question?
2. What are the relationships between them?



What do these signs signify?

1. A, B, C are distinguishable.
2. B is between A and C.
3. BC is twice as long as AB.

Sender encodes information in signs.

Through visual perception, the receiver decodes the signs for information:

1. What are the elements in question?
2. What are the relationships between them?

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

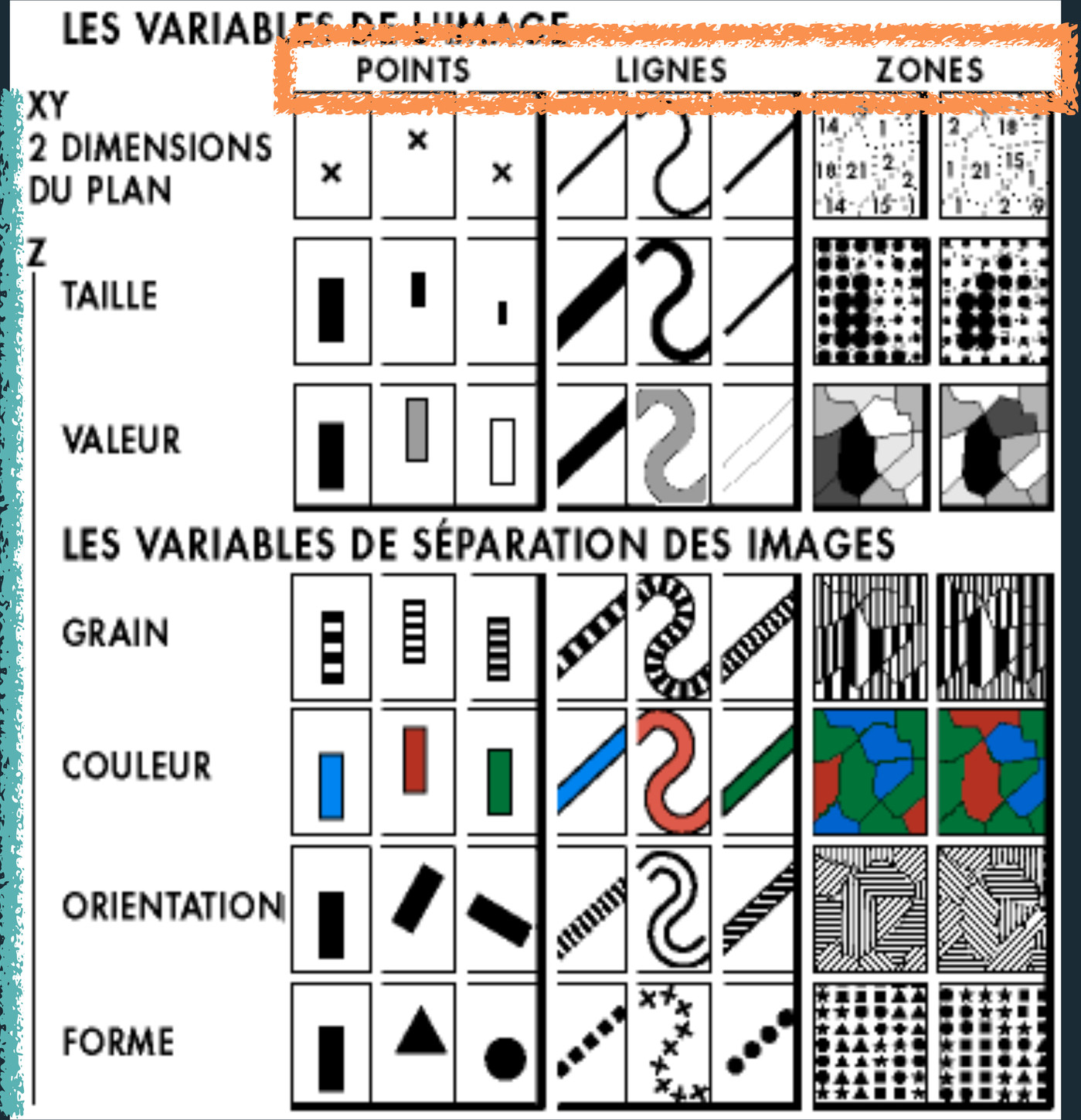
—Bertin

Visual Variables

Also called visual *channels*.

Used to encode data values as characteristics of marks.

** From 1967, so Bertin only accounted for visualizations that were printable, white paper.*



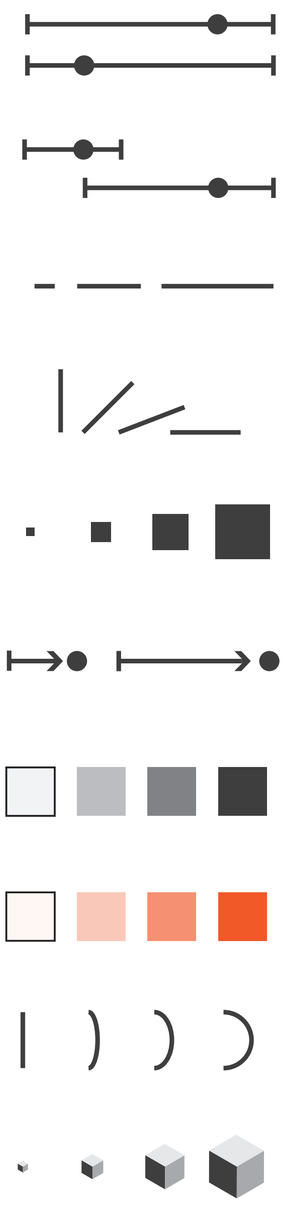
Marks

Basic graphical elements that represent data items.

Channels: Expressiveness Types and Effectiveness Ranks

Magnitude Channels: O or Q attributes

- Position on common scale
- Position on unaligned scale
- Length (1D size)
- Tilt/angle
- Area (2D size)
- Depth (3D position)
- Color luminance
- Color saturation
- Curvature
- Volume (3D size)



Most

Effectiveness

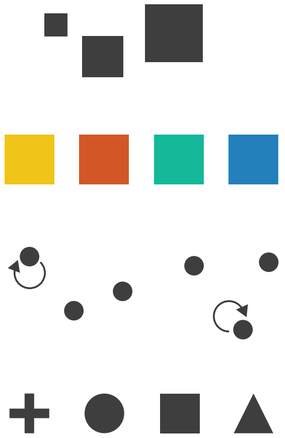
Least

Same

Same

Identity Channels: N attributes

- Spatial region
- Color hue
- Motion
- Shape



Tamara Munzner, *Visualization Analysis and Design* (2014).

Mapping or Visual Encoding

Data → **Visual**

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*.

Data models give us a way of talking about this.

[Mackinlay 1986]

Effectiveness

A visualization is more *effective* than another if the information it conveys *is more readily perceived* than the information in the other visualization

Image models give us a way of talking about this.

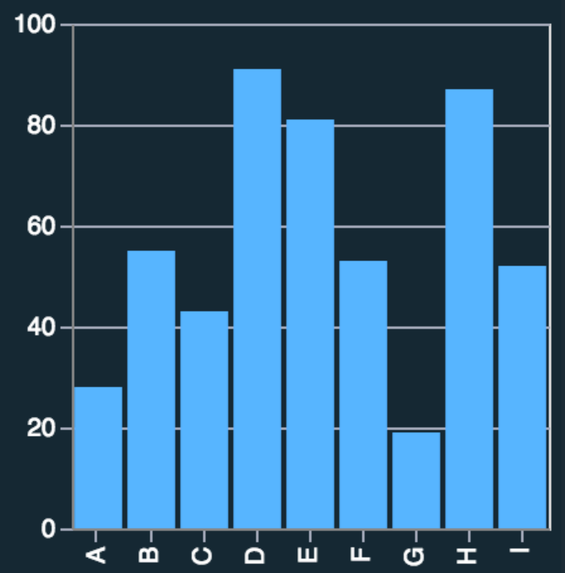
Activity!

Analyze which channels and marks you've used for the sketches so far. Create at least **3 new sketches** using different channels/marks. **3 minutes.**

75 37

Visual Encoding

Visual Encoding: 1 Nominal, 1 Quantitative

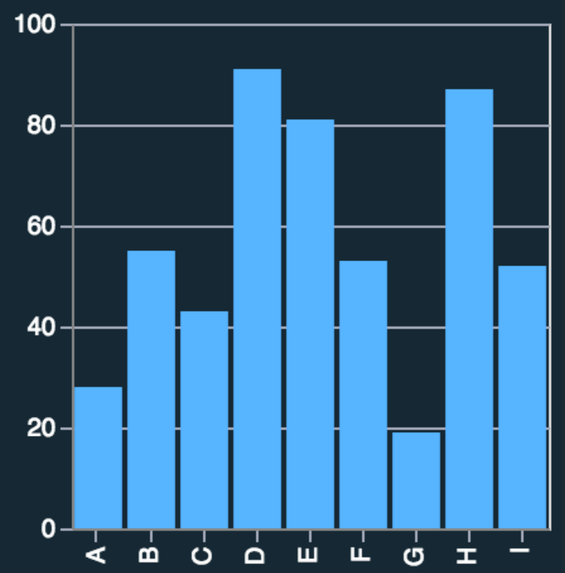


Mark: Bar

$d_{nominal} \rightarrow x$

$d_{quantitative} \rightarrow y$

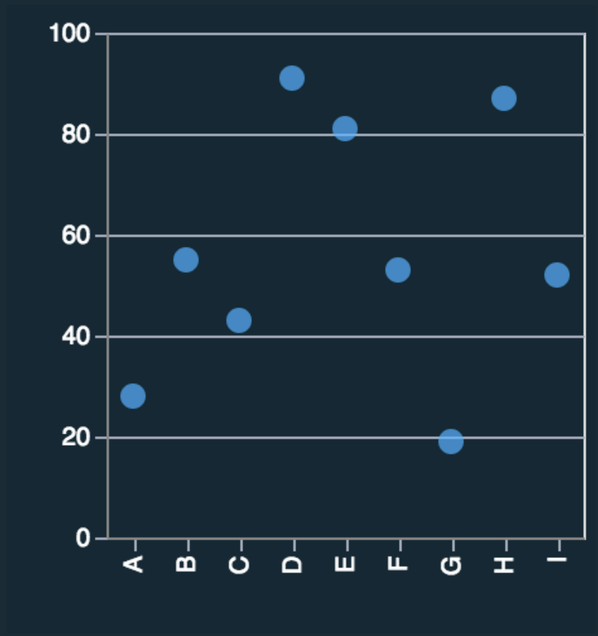
Visual Encoding: 1 Nominal, 1 Quantitative



Mark: Bar

$d_{nominal} \rightarrow x$

$d_{quantitative} \rightarrow y$

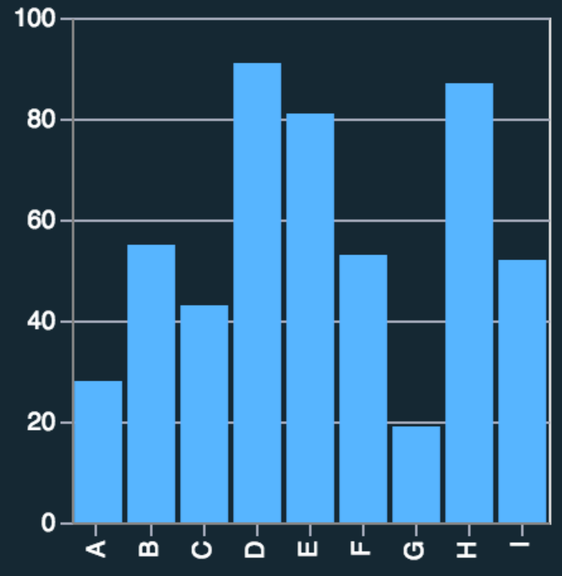


Mark: Point

$d_{nominal} \rightarrow x$

$d_{quantitative} \rightarrow y$

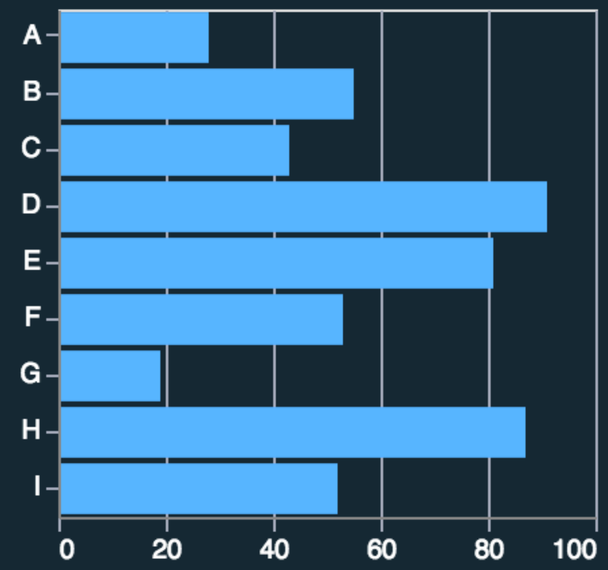
Visual Encoding: 1 Nominal, 1 Quantitative



Mark: Bar

$d_{nominal} \rightarrow x$

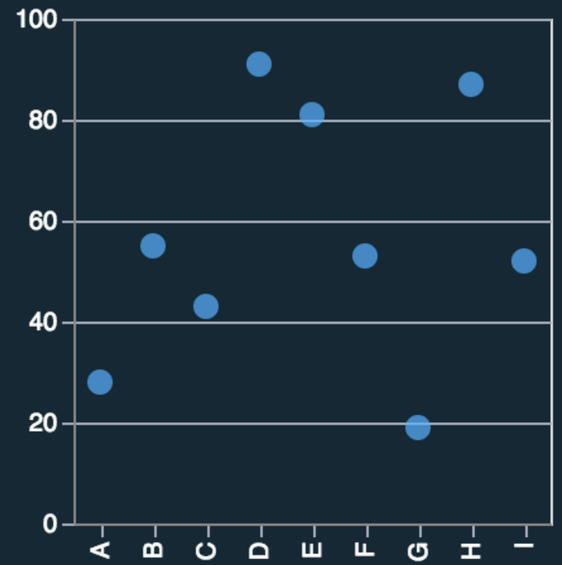
$d_{quantitative} \rightarrow y$



Mark: Bar

$d_{nominal} \rightarrow y$

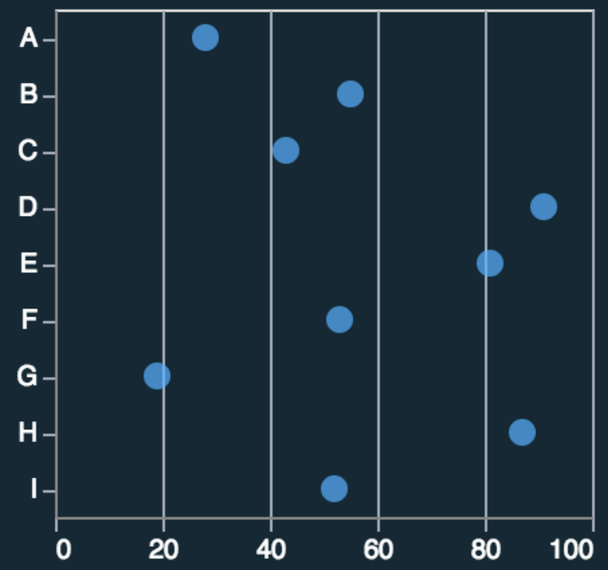
$d_{quantitative} \rightarrow x$



Mark: Point

$d_{nominal} \rightarrow x$

$d_{quantitative} \rightarrow y$

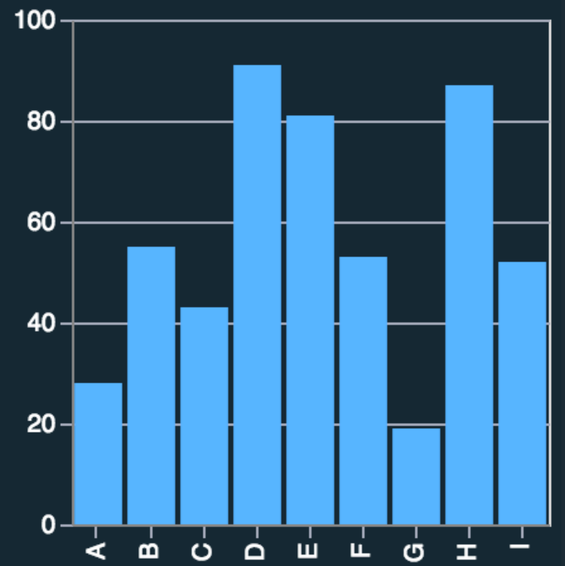


Mark: Point

$d_{nominal} \rightarrow y$

$d_{quantitative} \rightarrow x$

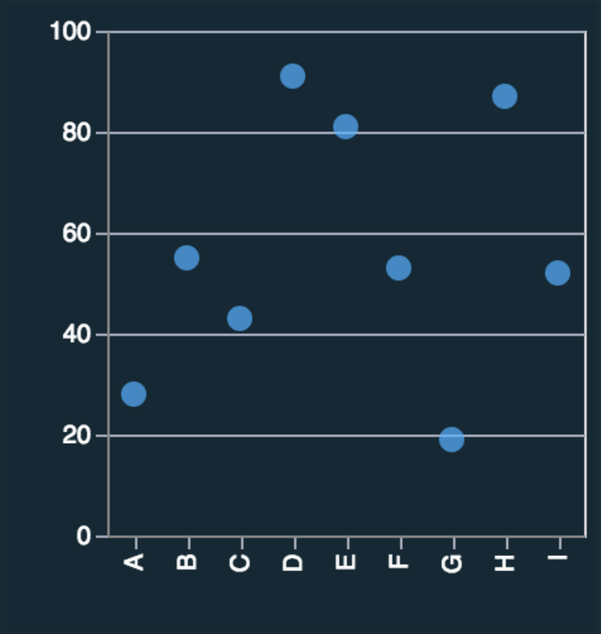
Visual Encoding: 1 Nominal, 1 Quantitative



Mark: Bar

$d_{nominal} \rightarrow x$

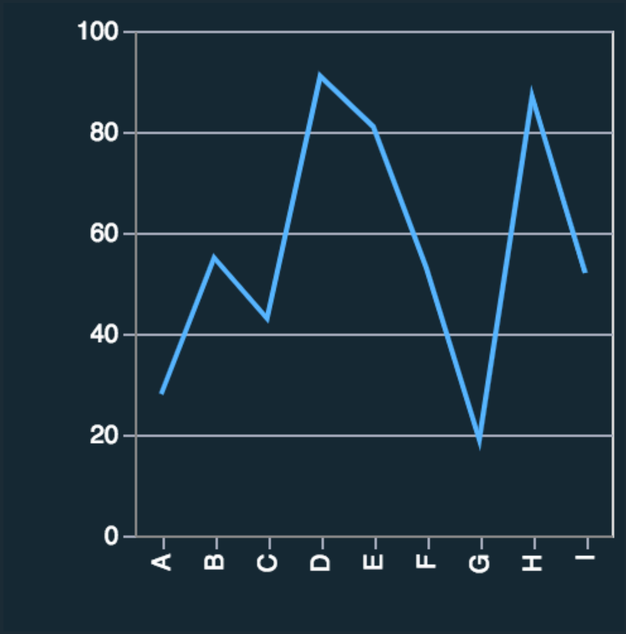
$d_{quantitative} \rightarrow y$



Mark: Point

$d_{nominal} \rightarrow x$

$d_{quantitative} \rightarrow y$

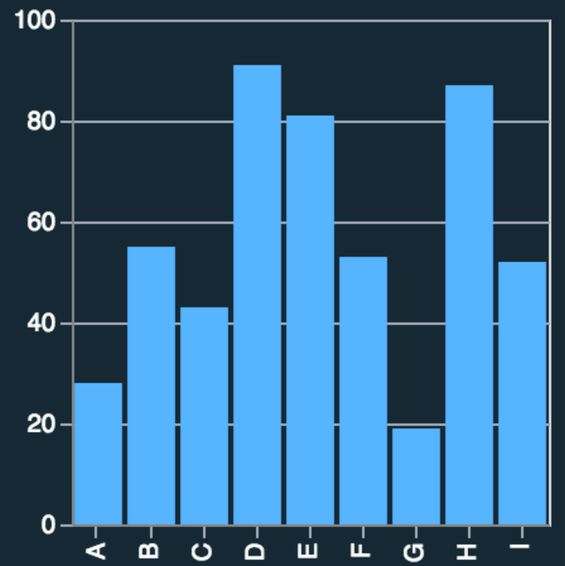


Mark: Line

$d_{nominal} \rightarrow x$

$d_{quantitative} \rightarrow y$

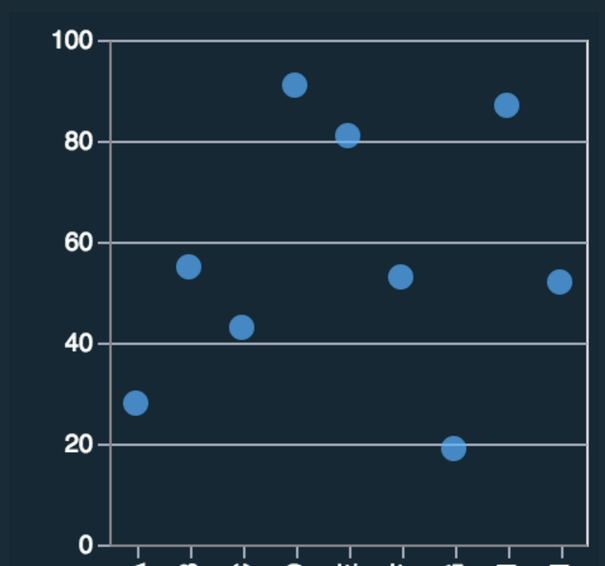
Visual Encoding: 1 Nominal, 1 Quantitative



Mark: Bar

$d_{nominal} \rightarrow x$

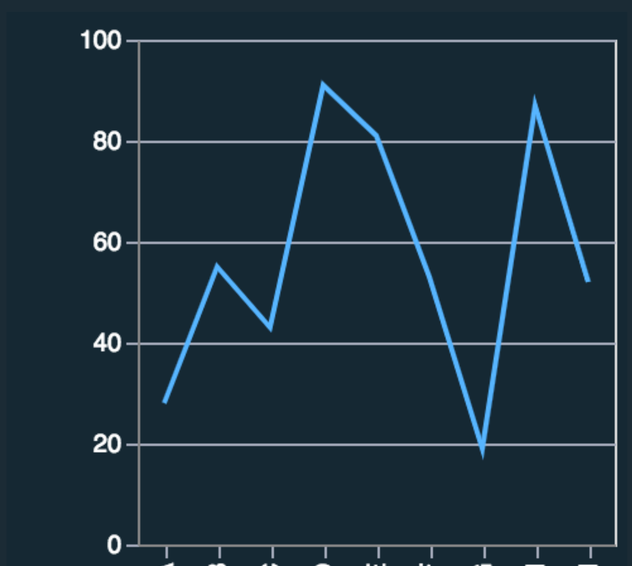
$d_{quantitative} \rightarrow y$



Mark: Point

$d_{nominal} \rightarrow x$

$d_{quantitative} \rightarrow y$



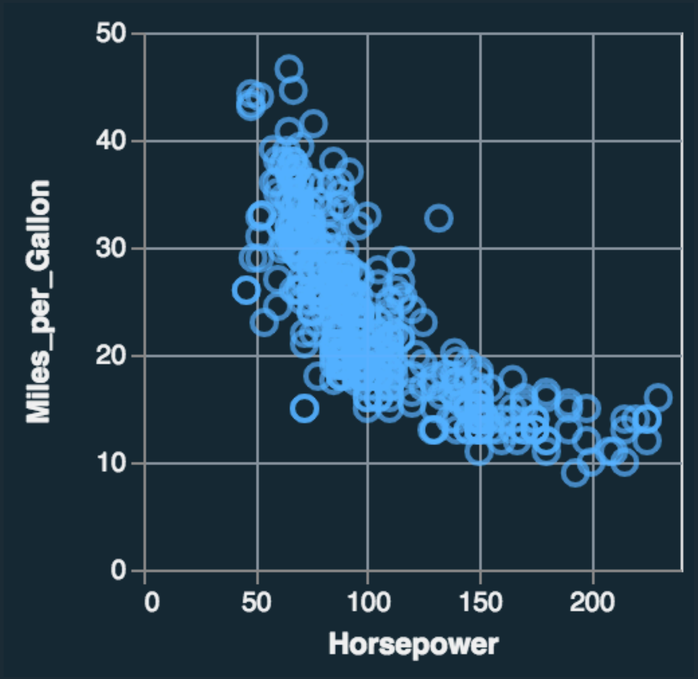
Mark: Line

$d_{nominal} \rightarrow x$

$d_{quantitative} \rightarrow y$

Violates expressiveness: the line mark implies a trend across the various categories.

Visual Encoding

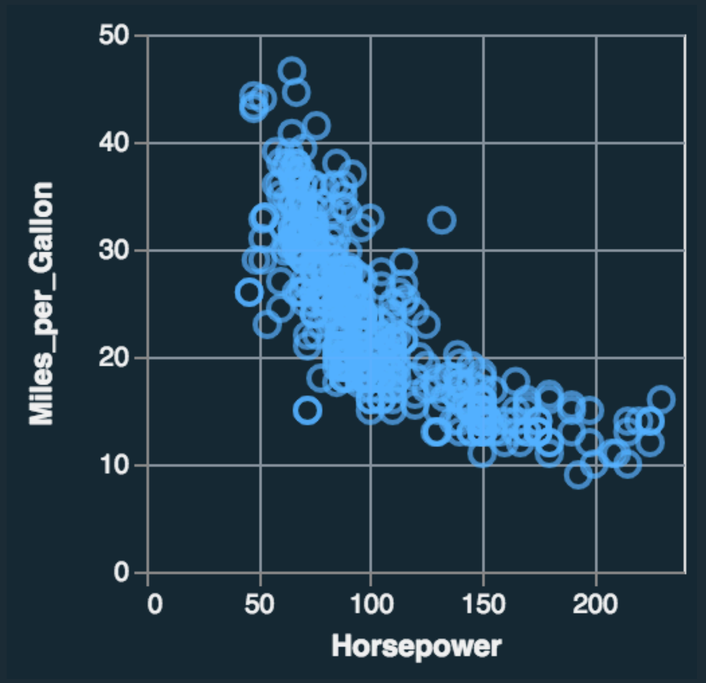


Mark: Point

$d_{quantitative} \rightarrow x$

$d_{quantitative} \rightarrow y$

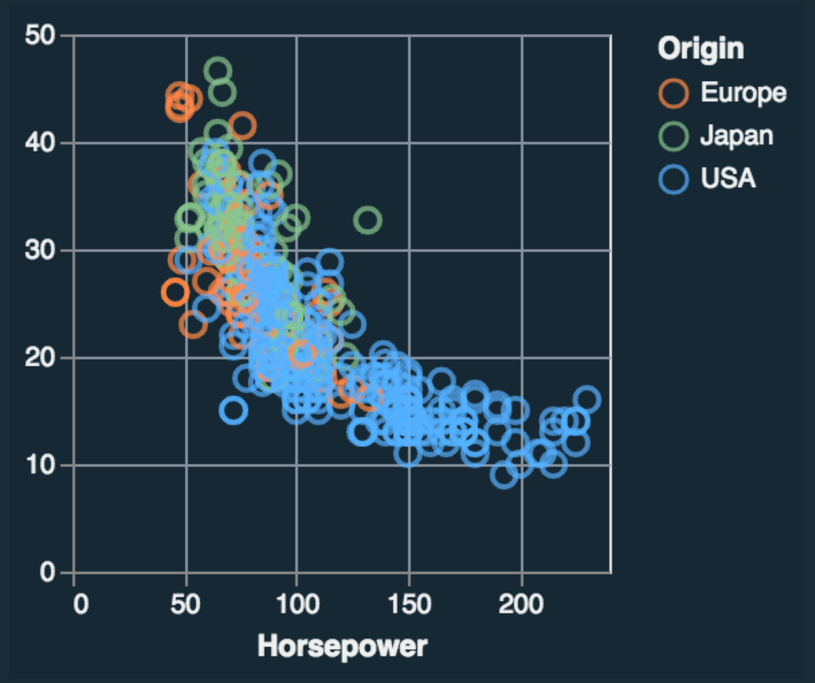
Visual Encoding



Mark: Point

$d_{quantitative} \rightarrow x$

$d_{quantitative} \rightarrow y$



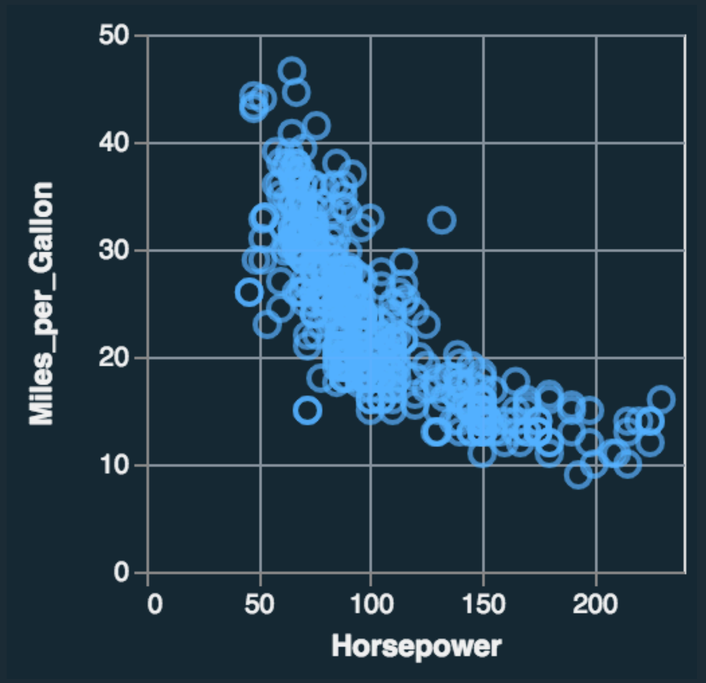
Mark: Point

$d_{quantitative} \rightarrow x$

$d_{quantitative} \rightarrow y$

$d_{nominal} \rightarrow color$

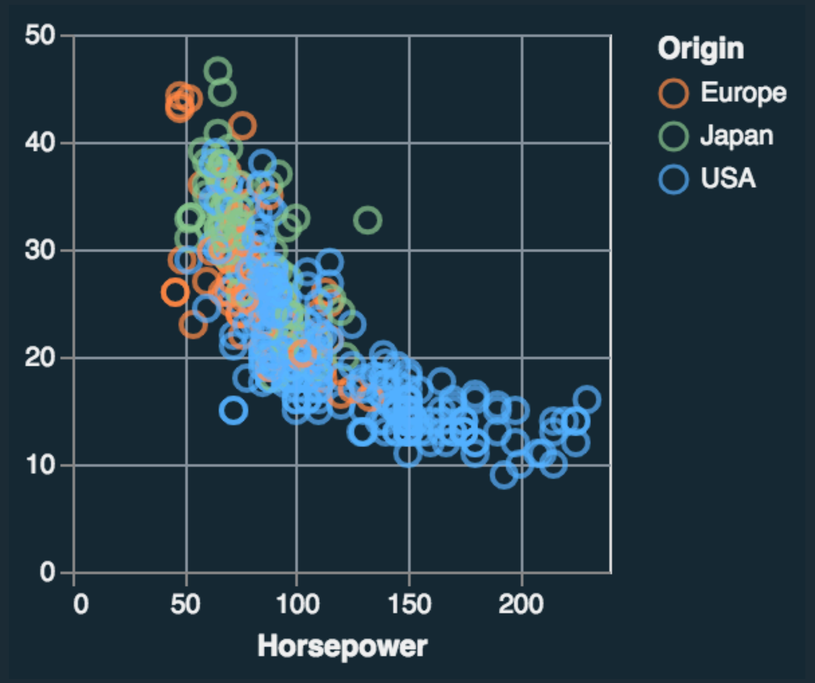
Visual Encoding



Mark: Point

$d_{quantitative} \rightarrow x$

$d_{quantitative} \rightarrow y$

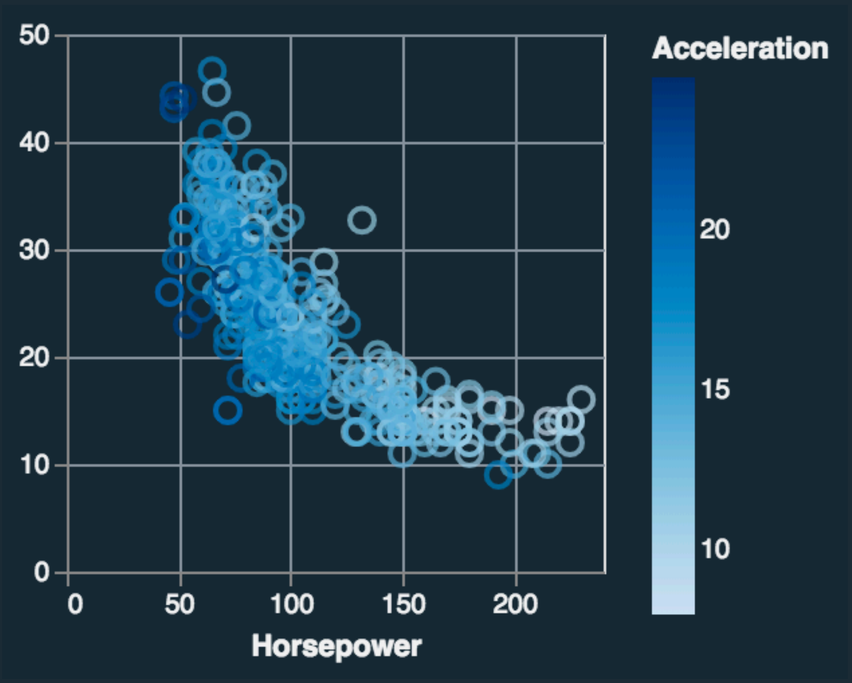


Mark: Point

$d_{quantitative} \rightarrow x$

$d_{quantitative} \rightarrow y$

$d_{nominal} \rightarrow color$



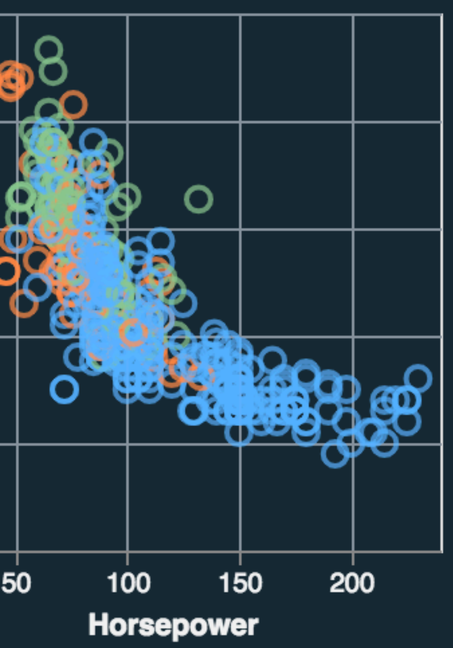
Mark: Point

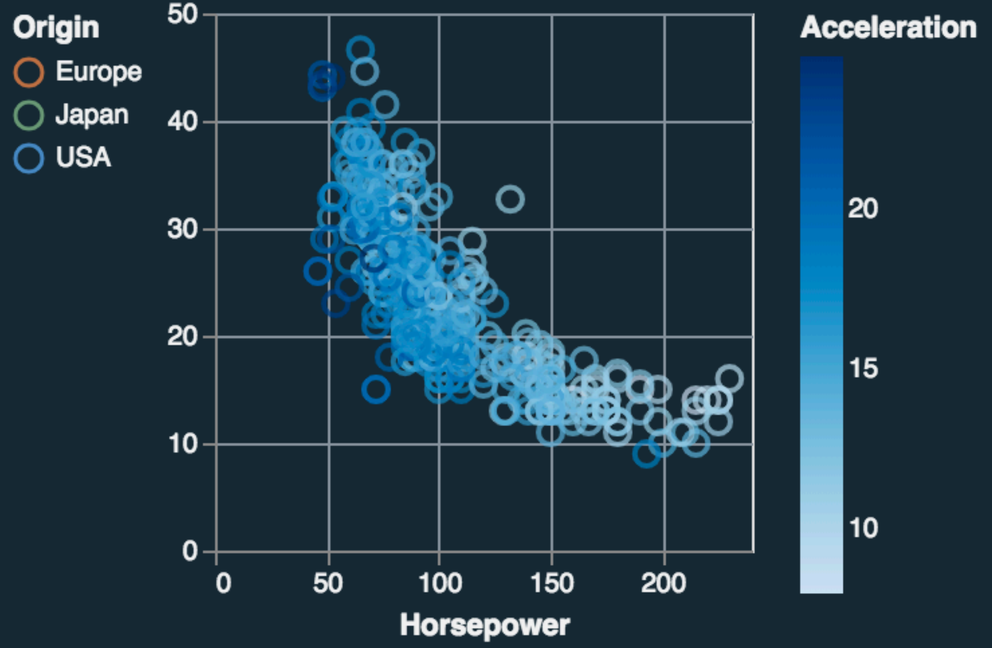
$d_{quantitative} \rightarrow x$

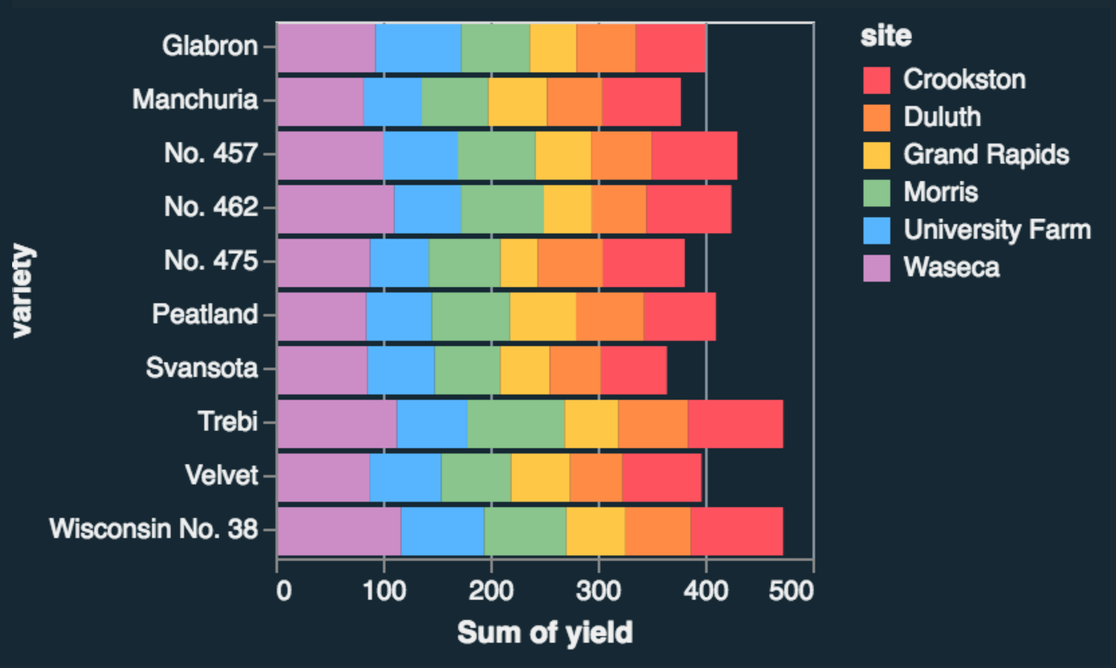
$d_{quantitative} \rightarrow y$

$d_{quantitative} \rightarrow color$

Visual Encoding







Mark: Point

$d_{quantitative} \rightarrow x$

$d_{quantitative} \rightarrow y$

$d_{nominal} \rightarrow color$

Mark: Point

$d_{quantitative} \rightarrow x$

$d_{quantitative} \rightarrow y$

$d_{quantitative} \rightarrow color$

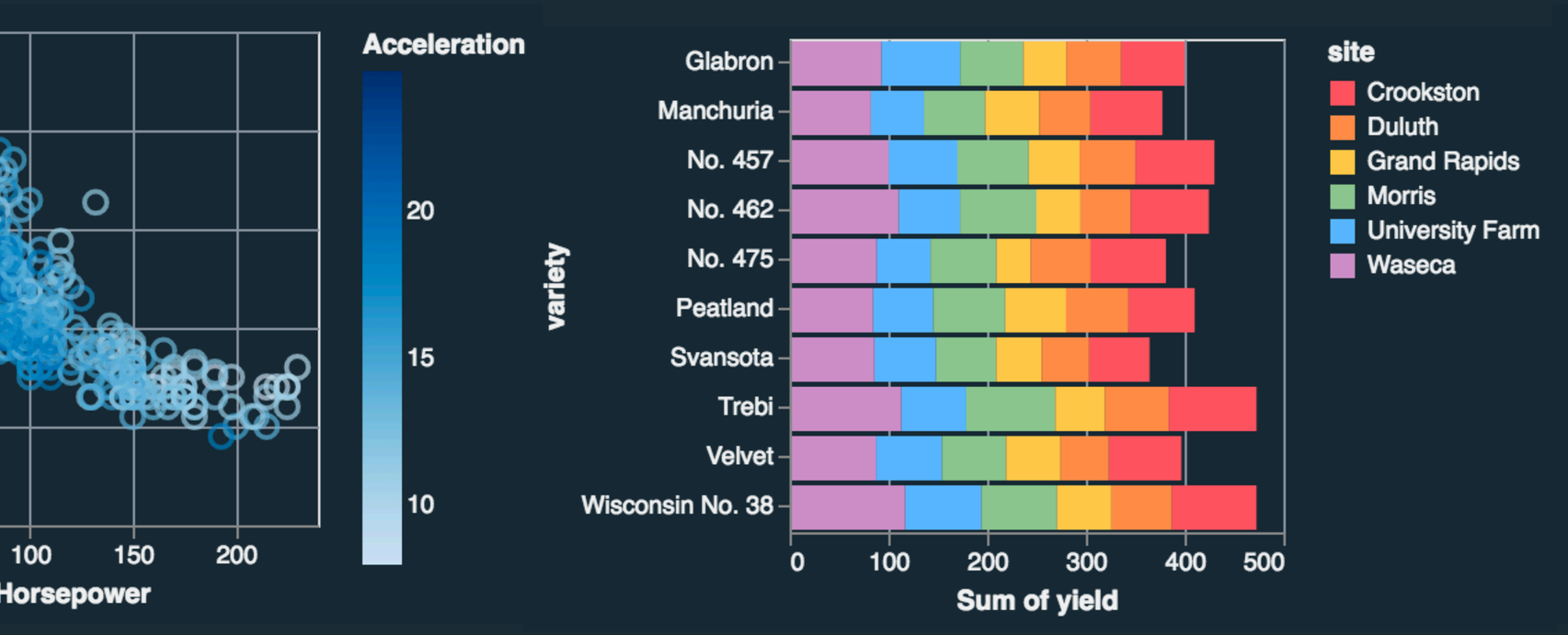
Mark: Bar

$d_{quantitative} \rightarrow x$

$d_{nominal} \rightarrow y$

$d_{nominal} \rightarrow color$

Visual Encoding



Mark: Point

$d_{quantitative} \rightarrow x$

$d_{quantitative} \rightarrow y$

$d_{quantitative} \rightarrow color$

Mark: Bar

$d_{quantitative} \rightarrow x$

$d_{nominal} \rightarrow y$

$d_{nominal} \rightarrow color$

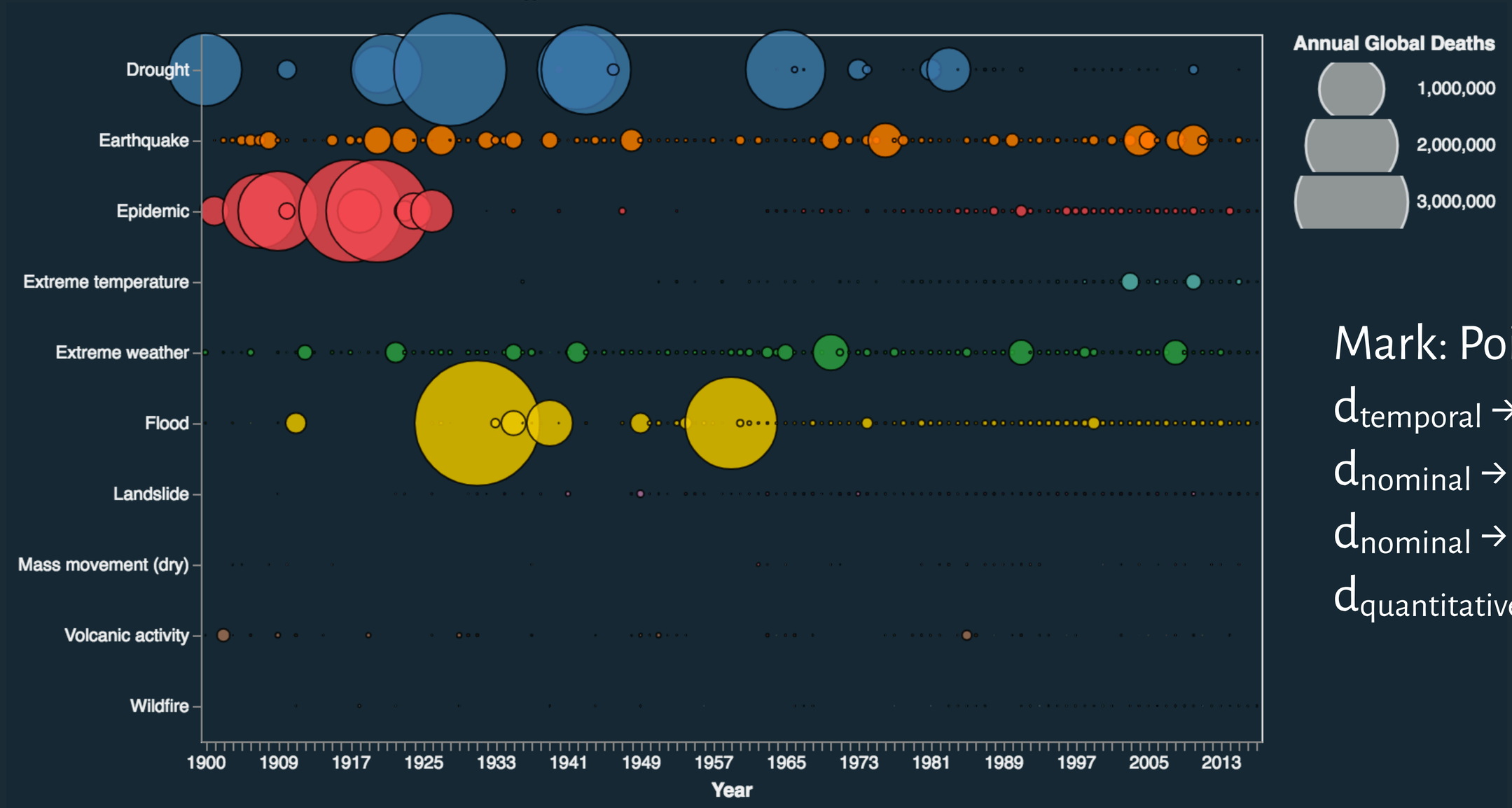
Mark: Line

$d_{temporal} \rightarrow x$

$d_{quantitative} \rightarrow y$

$d_{nominal} \rightarrow color$

Visual Encoding



Mark: Point

$d_{temporal} \rightarrow x$

$d_{nominal} \rightarrow y$

$d_{nominal} \rightarrow color$

$d_{quantitative} \rightarrow size$