Data and Image Models

DSC 106: Data Visualization

Jared Wilber

UC San Diego

Announcements

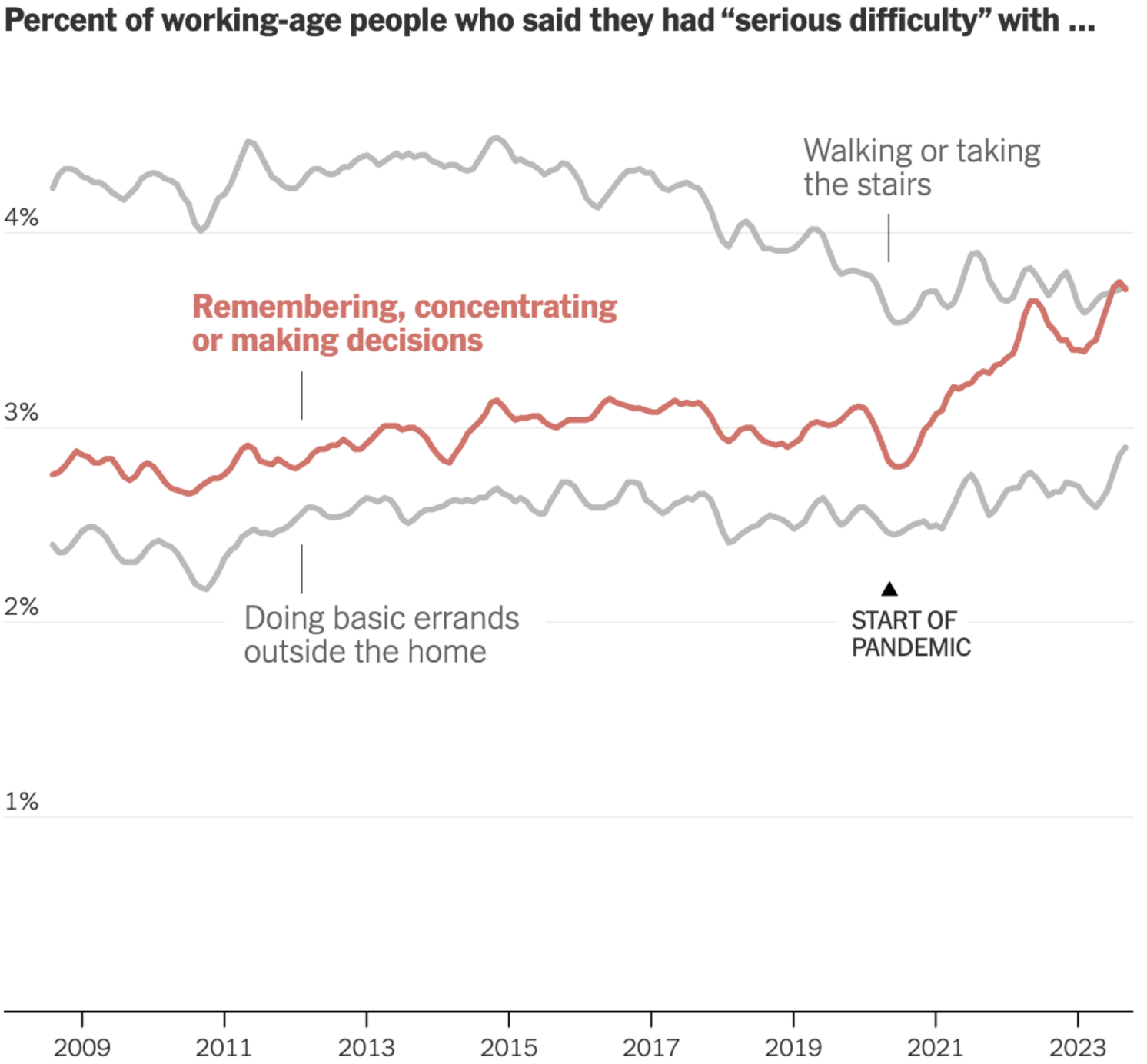
Lab 1 and Welcome Survey due **tomorrow**!

Project 1 due next week Friday, 4/12.

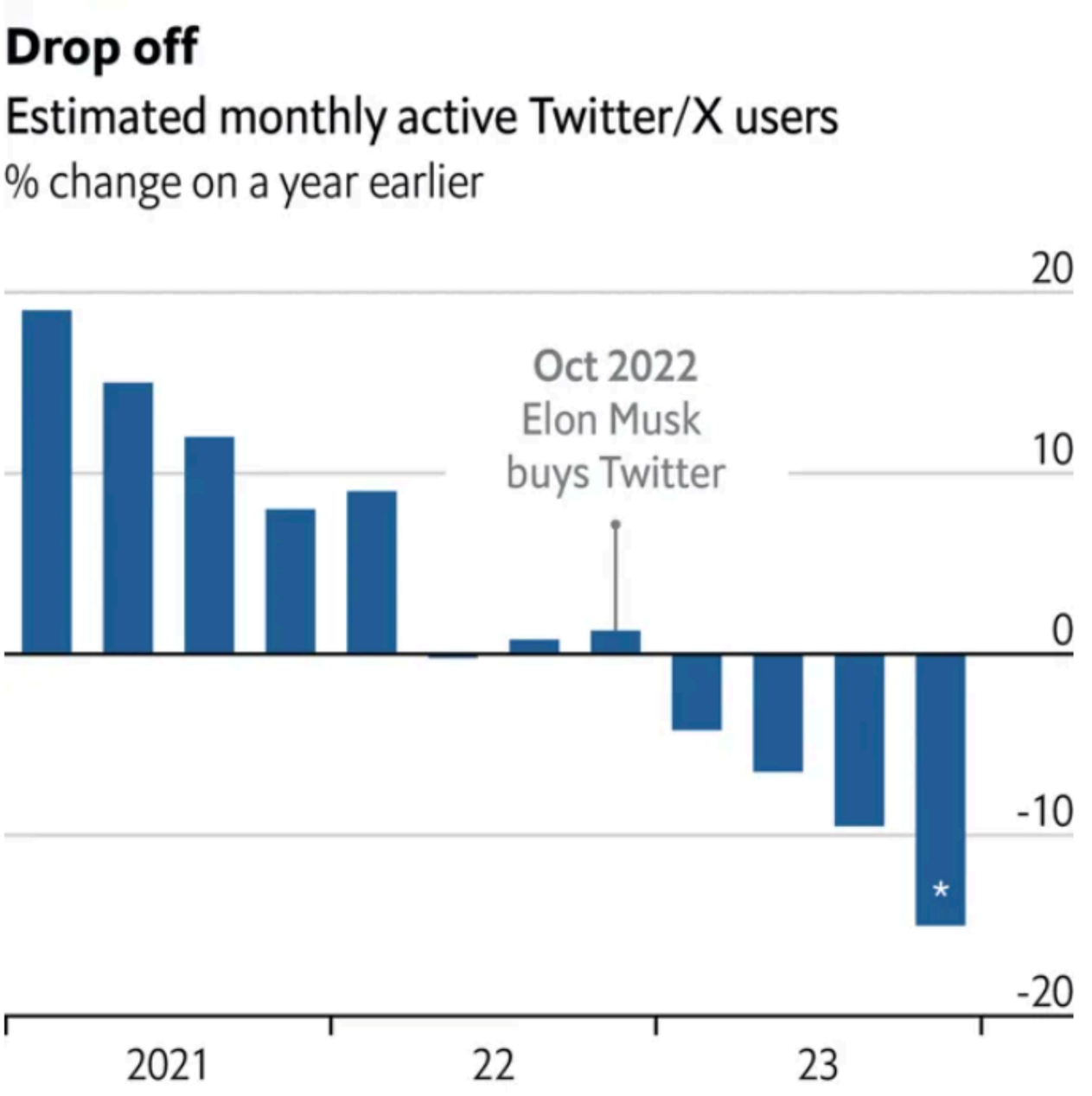
FAQs on course logistics:

1. Are lectures podcasted? Not yet, but ***once I get access to UCSD systems***
2. Participation? Attendance not mandatory
3. Can I use ChatGPT / CoPilot? Yes, but use with caution!
4. OH: TBD

Name that chart!



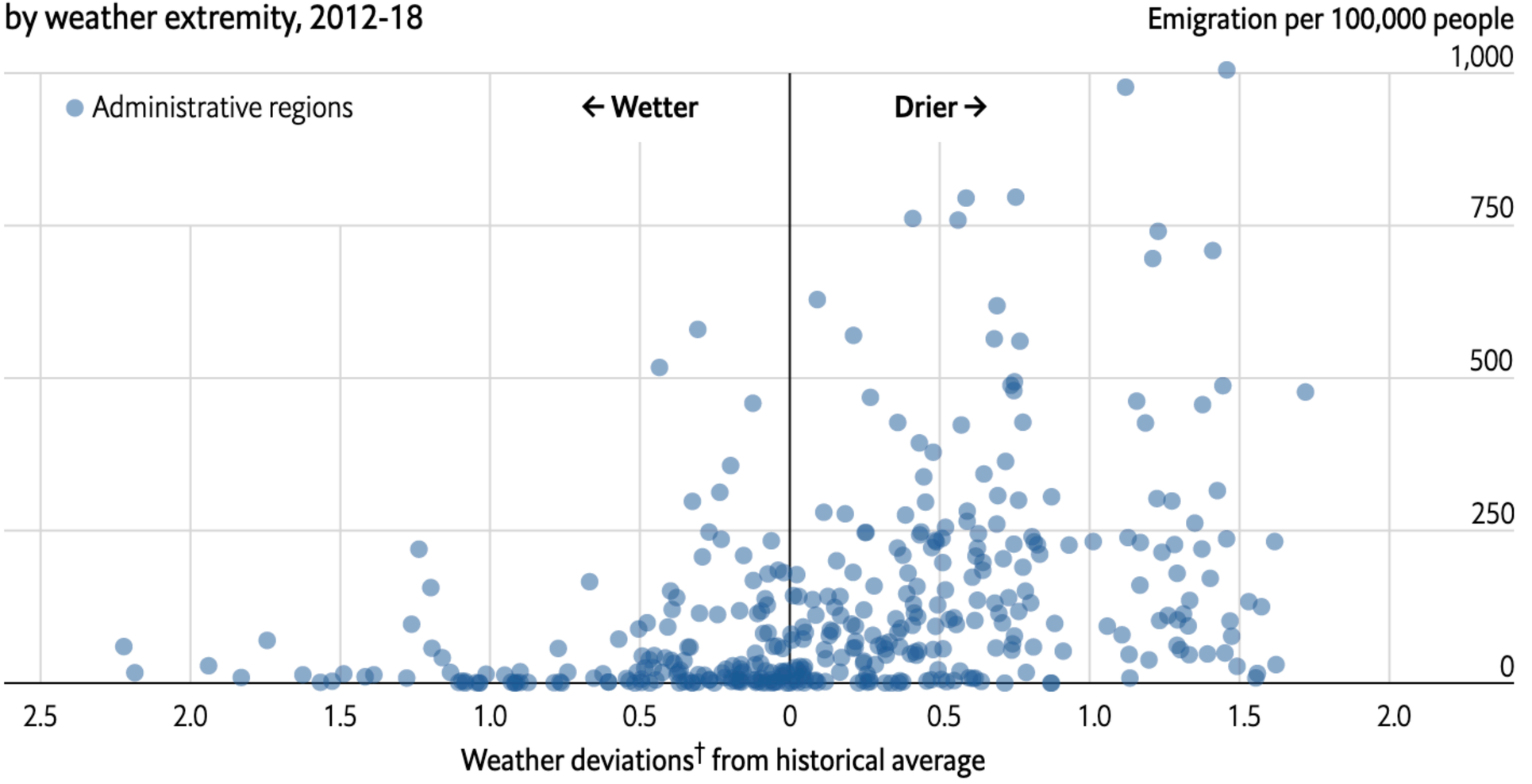
https://www.nytimes.com/2024/01/10/upshot/covid-pandemic-wave.html



https://www.economist.com/graphic-detail/2023/12/20/has-twitter-now-x-become-more-right-wing

Spotting a trend

Emigration from the Northern Triangle* to United States, by weather extremity, 2012-18

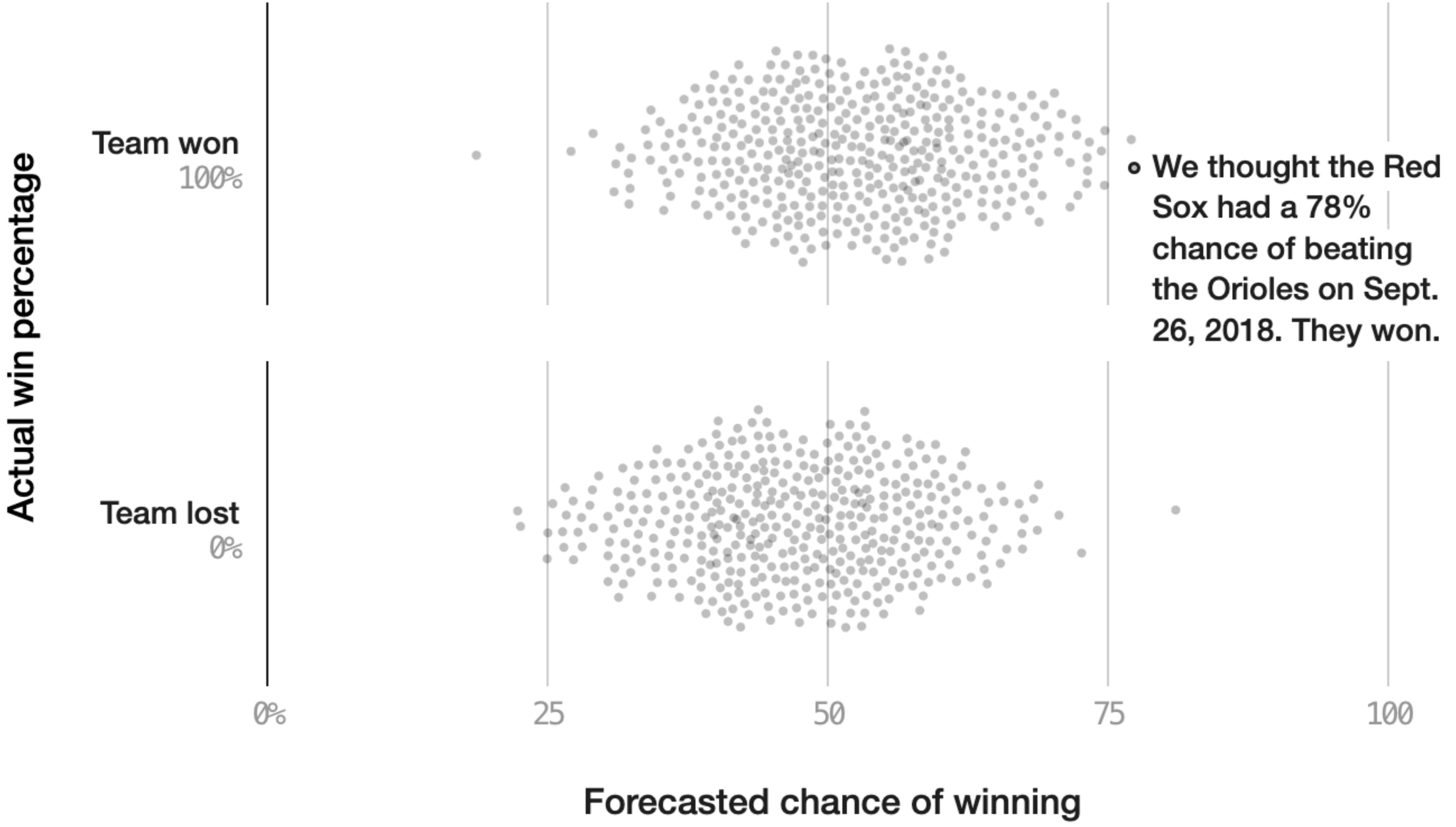


*El Salvador, Guatemala and Honduras †Using the Standardised Precipitation-Evapotranspiration Index three-month average

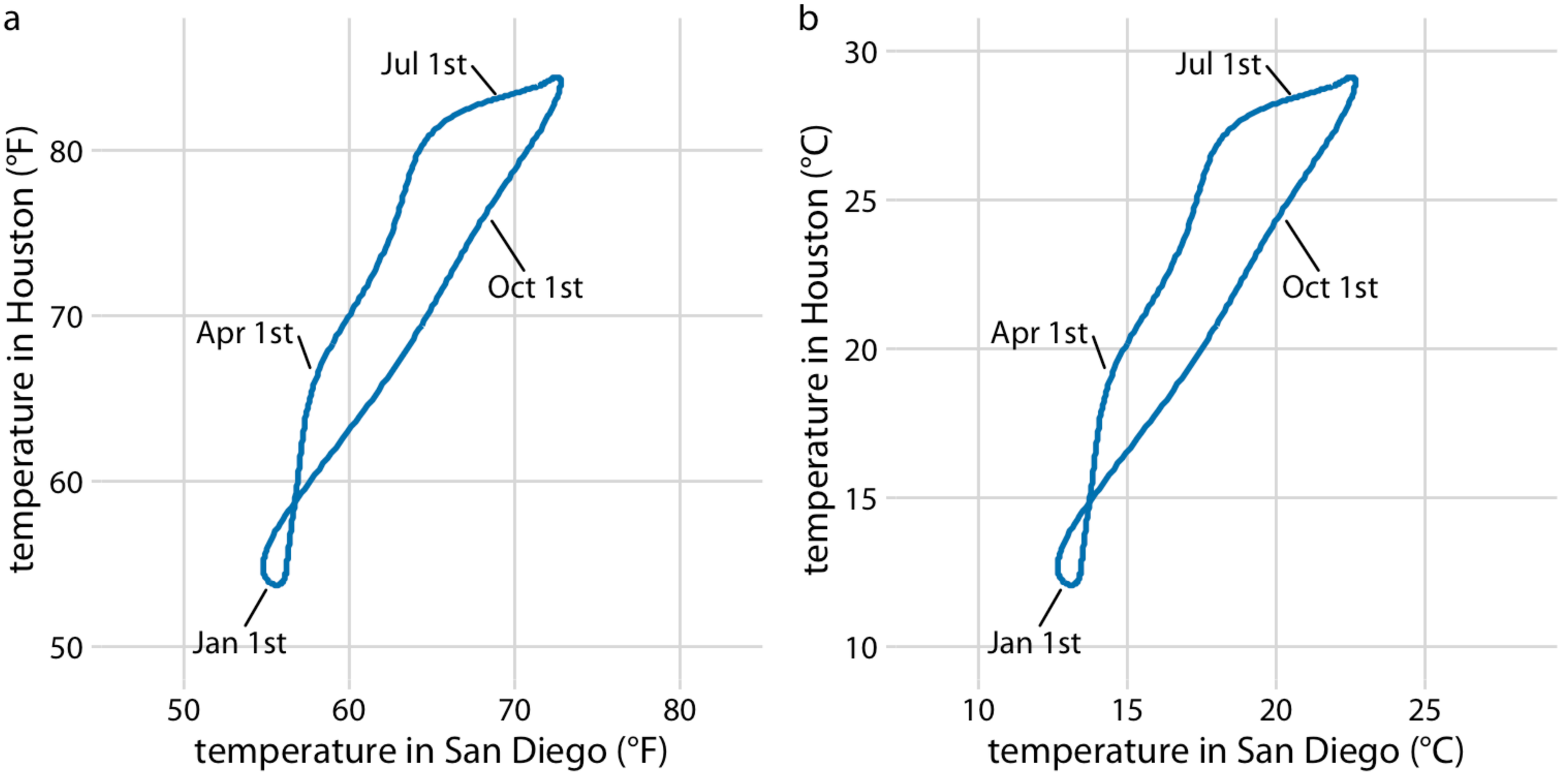
Source: "Dry growing seasons predicted Central American migration to the US from 2012 to 2018", by A. Linke et al., 2023

https://www.economist.com/graphic-detail/2023/11/17/why-central-americans-migrate-to-the-united-states-when-they-do

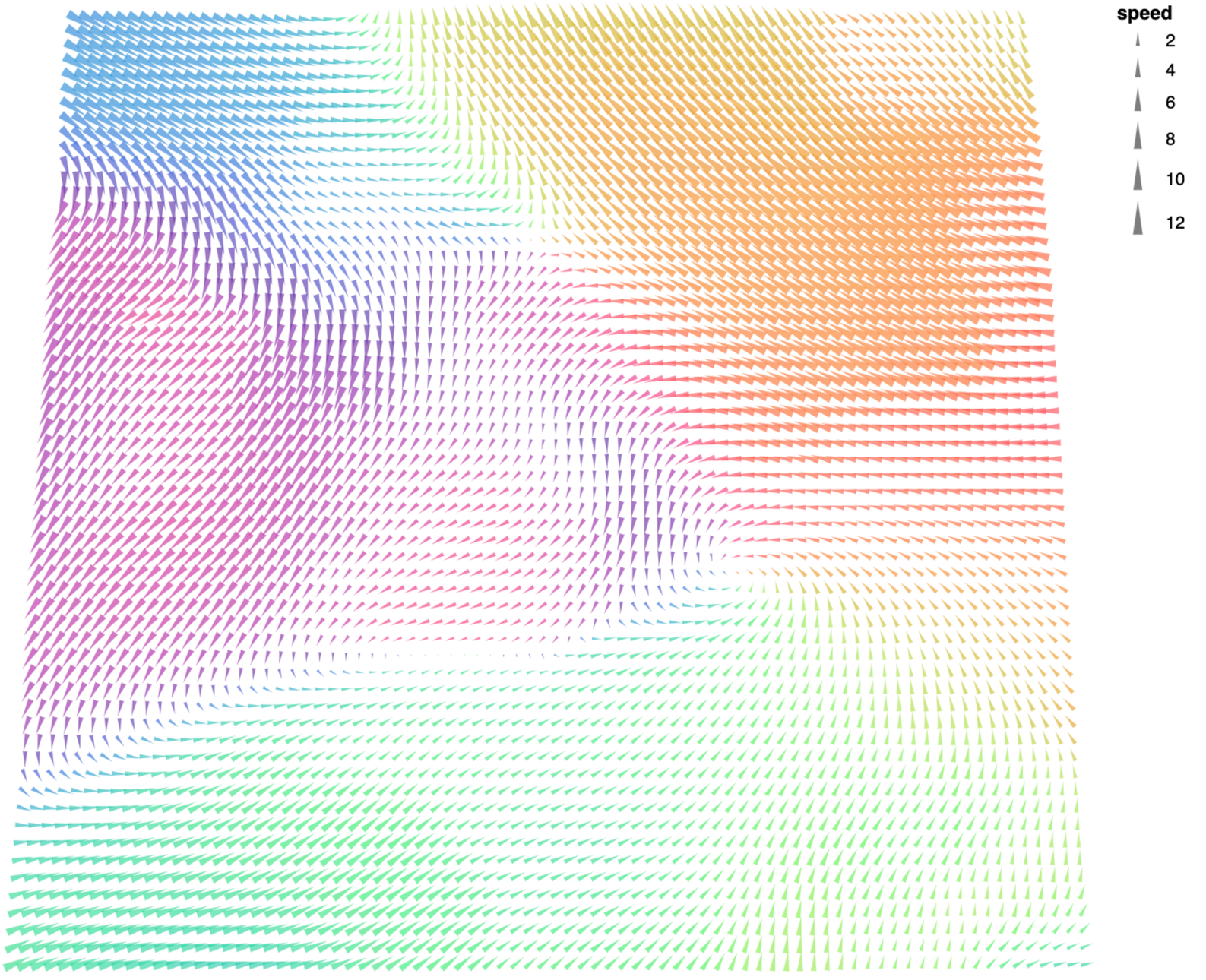
https://www.nature.com/articles/s41598-023-43668-9



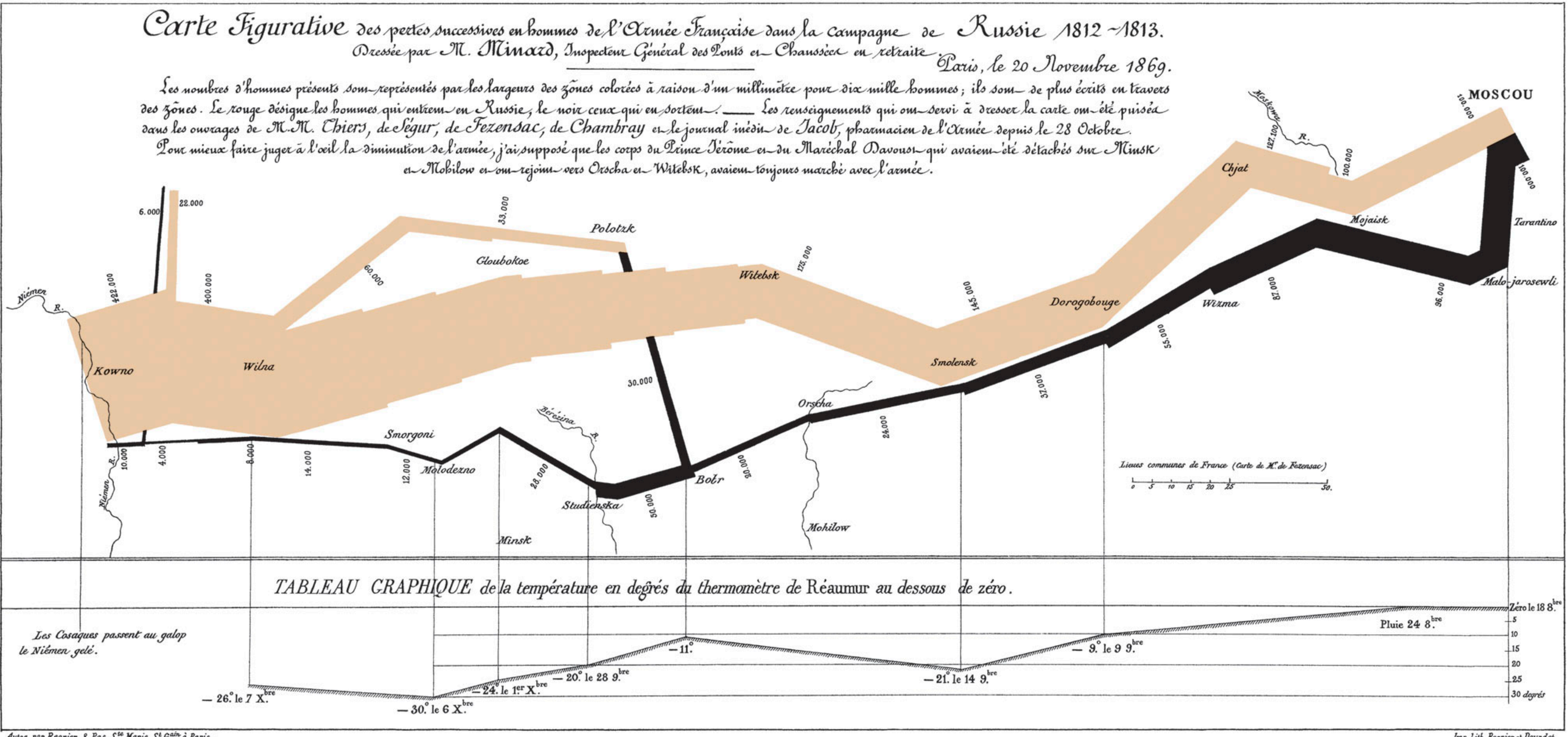
https://projects.fivethirtyeight.com/checking-our-work/



https://clauswilke.com/dataviz/coordinate-systems-axes.html#fig:temperature-normals-Houston-San-Diego



https://vega.github.io/vega-lite/examples/point_angle_windvector.html



[Charles Minard, 1869]

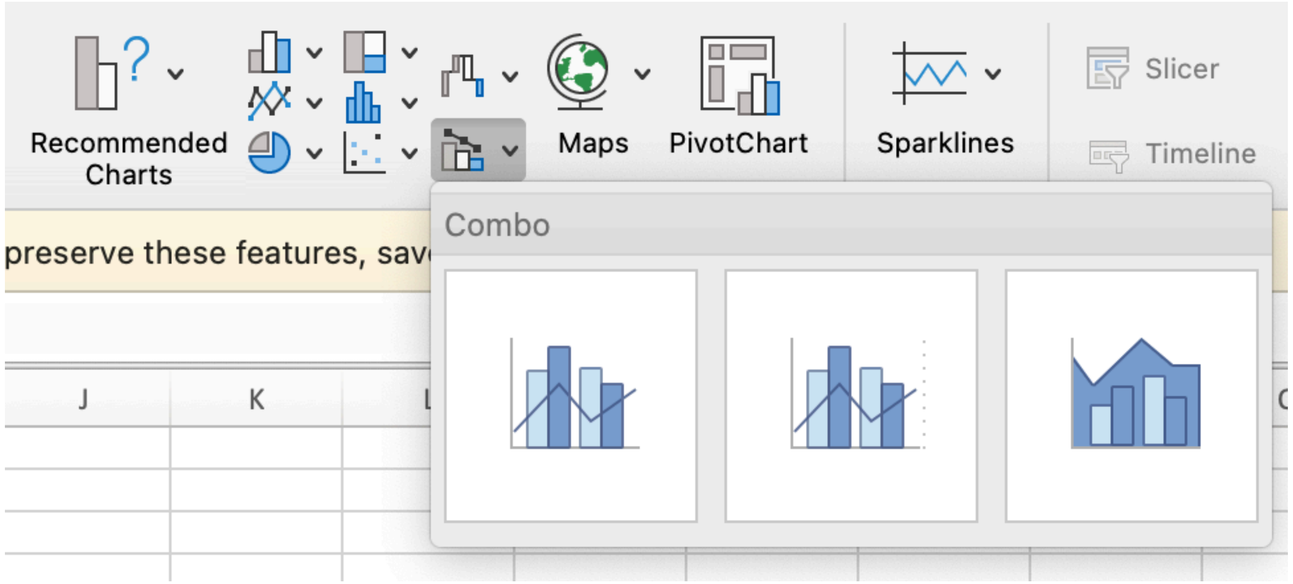

Recommended Charts
Maps
PivotChart
Sparklines
Slicer
Timeline
Combo
preserve these features, sav
J
K

Chart editor
Setup
Customize
Chart type
Pie chart
Line
Area
Column
Bar
Pie
Scatter
Map
Other

Visualizing Data

Physical Data Types

`int, float, string`

Graphical **Marks**

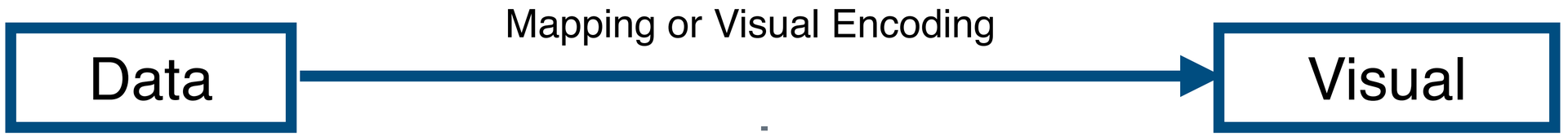
`rect, line, point, area`

Conceptual Data Types

`temperature, location`

Visual **Channels**

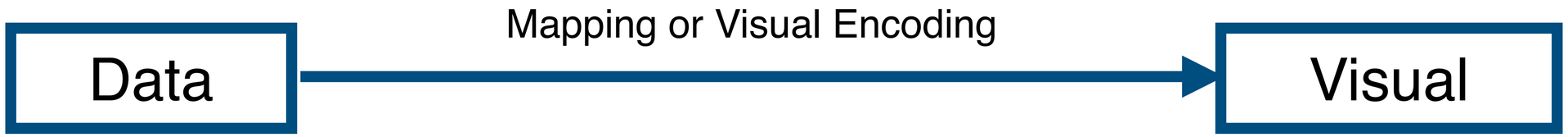
`x, y, color, opacity`



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, Jock. "Automating the design of graphical presentations of relational information." *Acm Transactions On Graphics (Tog)* 5.2 (1986): 110-141.

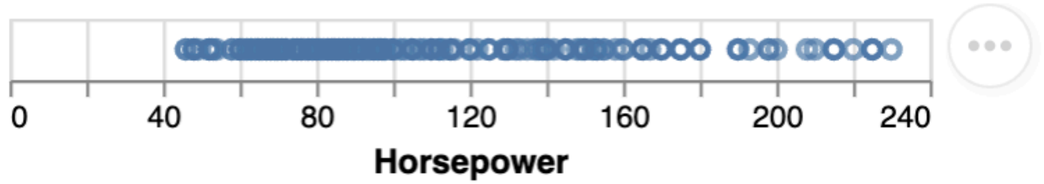


Expressiveness

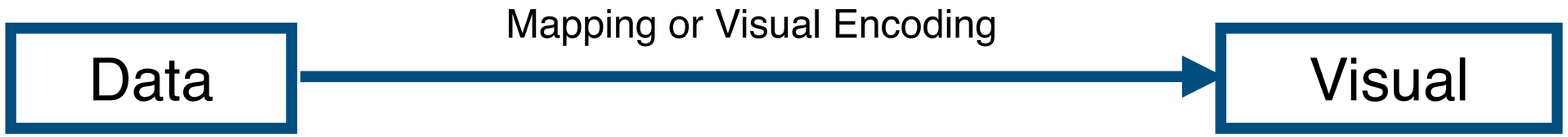
Can't express the facts

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

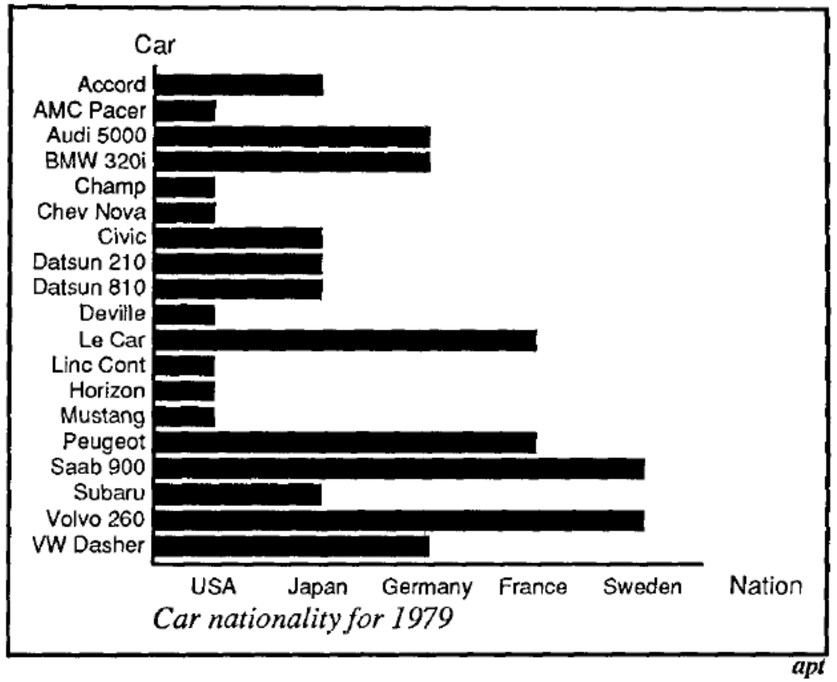
```
alt.Chart(source).mark_point().encode(
    x='Horsepower'
)
```

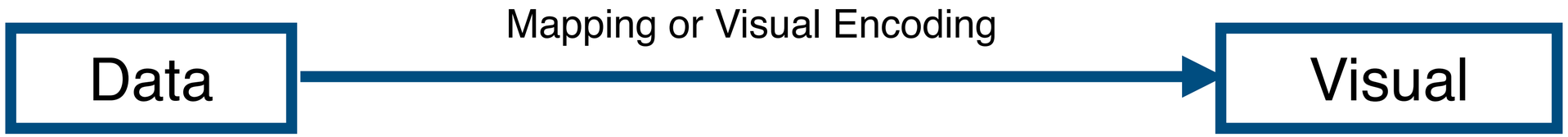


Mackinlay, Jock. "Automating the design of graphical presentations of relational information." *Acm Transactions On Graphics (Tog)* 5.2 (1986): 110-141.



Expressiveness





Expressiveness

Expresses facts not in the data

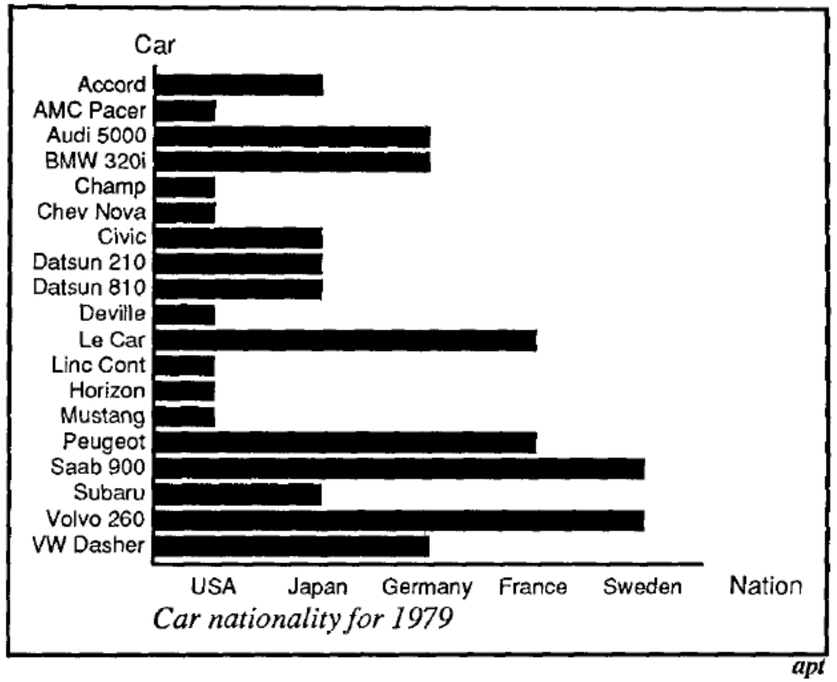
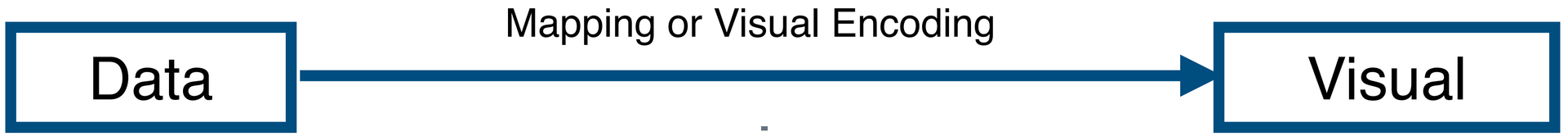


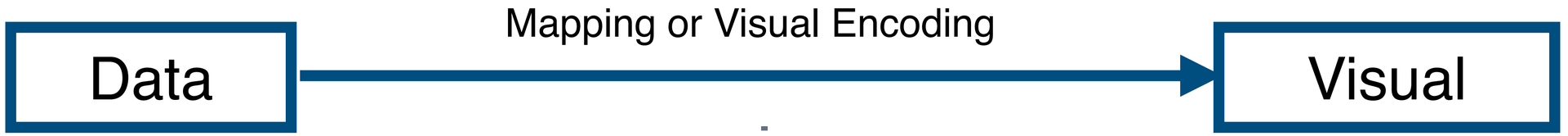
Fig. 11. Incorrect use of a bar chart for the *Nation* relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the *Nation* relation.



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.

Mackinlay, Jock. "Automating the design of graphical presentations of relational information." *Acm Transactions On Graphics (Tog)* 5.2 (1986): 110-141.



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 what the facts are.

Mackinlay, Jock. "Automating the design of graphical presentations of relational information." *Acm Transactions On Graphics (Tog)* 5.2 (1986): 110-141.

Data Models

Conceptual Models vs. Data Models

```
df = pd.read_csv('projects/proj01/weather.csv')
df
```

	city					sunshine	rain
0	San Diego				1	217	1.53
1	San Diego				2	255	0.15
2	San Diego	32.715736	-117.161087	Mar	3	234	0.57
3	San Diego	32.715736	-117.161087	Apr	4	236	1.01
4	San Diego	32.715736	-117.161087	May	5	277	0.02
...	...	...	...	...	...	...	...
67	Miami	25.761681	-80.191788	Aug	8	263	8.88
68	Miami	25.761681	-80.191788	Sep	9	216	9.86
69	Miami	25.761681	-80.191788	Oct	10	215	6.33
70	Miami	25.761681	-80.191788	Nov	11	212	3.27
71	Miami	25.761681	-80.191788	Dec	12	209	2.04

Conceptual Model: column represents hours of sunshine

Conceptual Models vs. Data Models

```
df = pd.read_csv('projects/proj01/weather.csv')
df
```

	city	lat	lon	month	monthnum	sunshine	rain
0	S				1	217	1.53
1	S				2	255	0.15
2	S				3	234	0.57
3	San Diego	32.715736	-117.161087	Apr	4	236	1.01
5					5	277	0.02
...	...	...	...	...	...	...	...
67	Miami	25.761681	-80.191788	Aug	8	263	8.88
68	Miami	25.761681	-80.191788	Sep	9	216	9.86
69	Miami	25.761681	-80.191788	Oct	10	215	6.33
70	Miami	25.761681	-80.191788	Nov	11	212	3.27
71	Miami	25.761681	-80.191788	Dec	12	209	2.04

Data Model:
column contains numbers

Higher level of abstraction!

Dataset Types

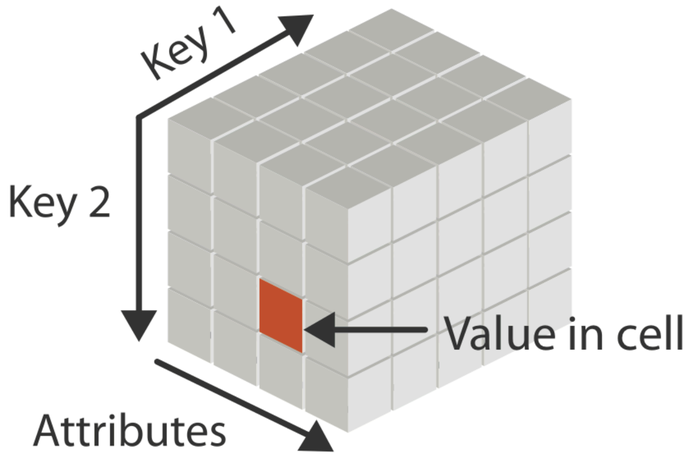
1. Tabular

columns/attributes/ variables

rows/records/items

	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



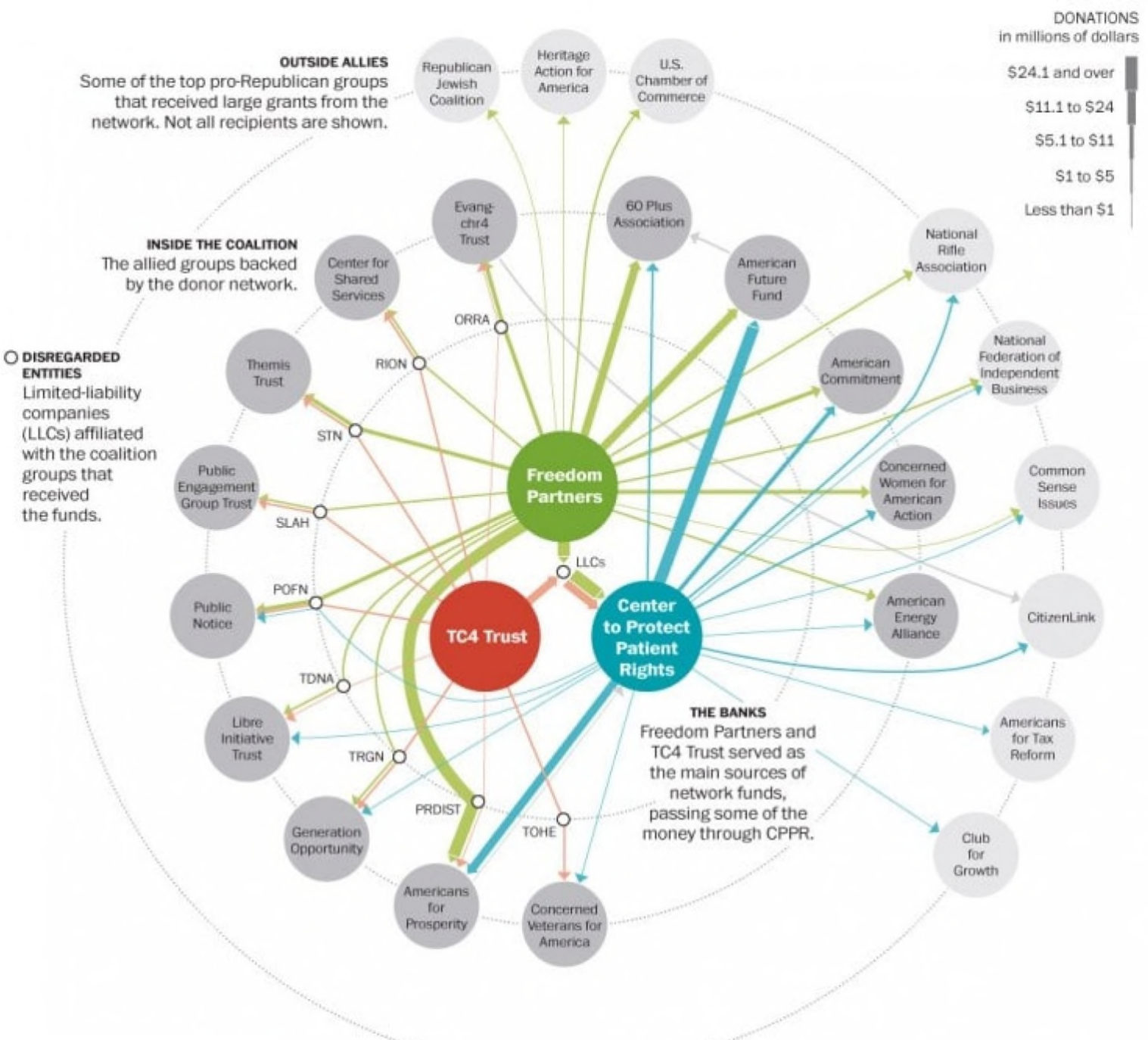
Tamara Munzner, *Visualization Analysis and Design* (2014).

Dataset Types

1. Tabular:
   collection of records with named attributes

2. Network:
   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.



https://www.washingtonpost.com/politics/inside-the-koch-backed-political-donor-network/2014/01/05/94719296-7661-11e3-b1c5-739e63e9c9a7_graphic.html

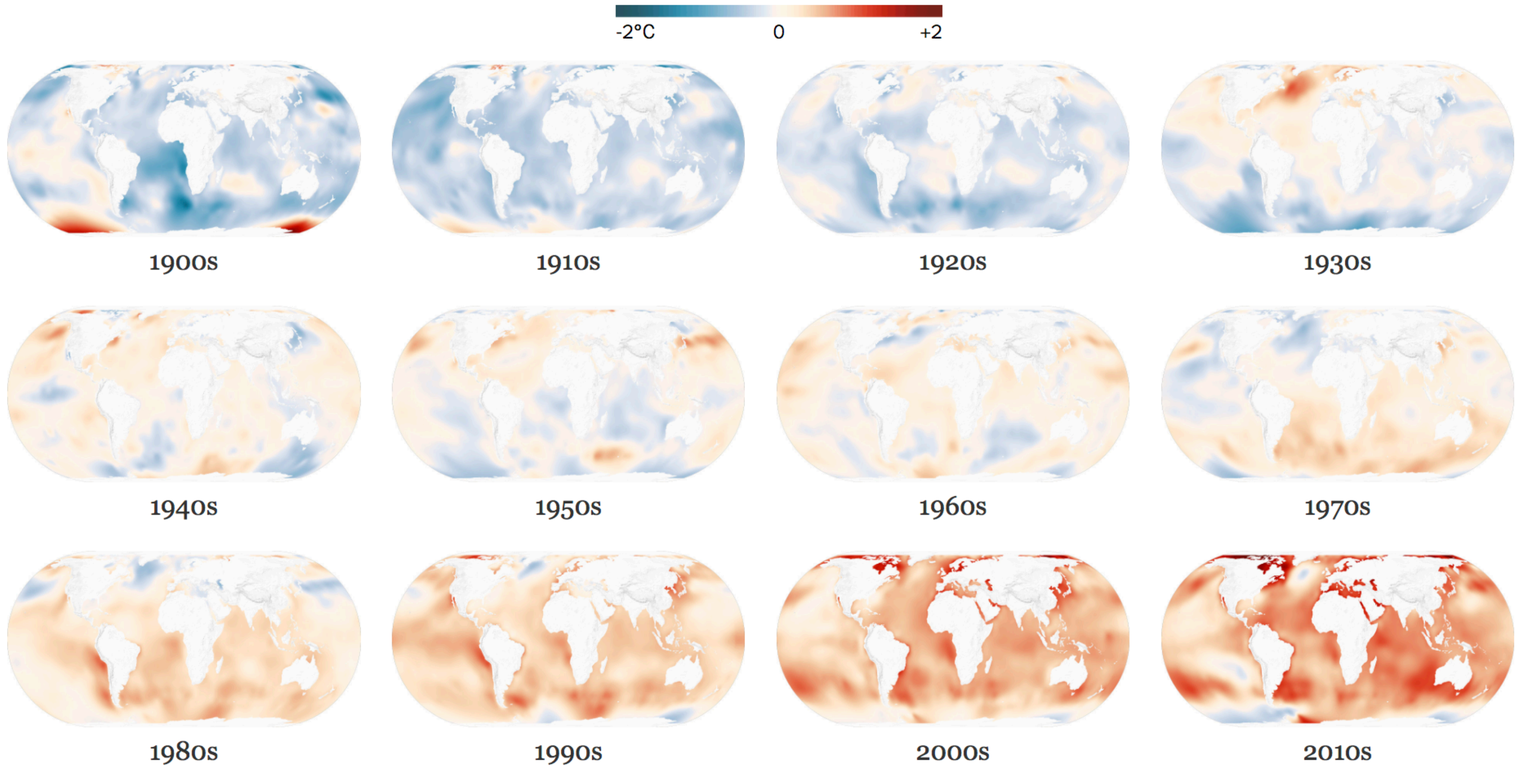
Dataset Types

1. Tabular:
   collection of records with named attributes

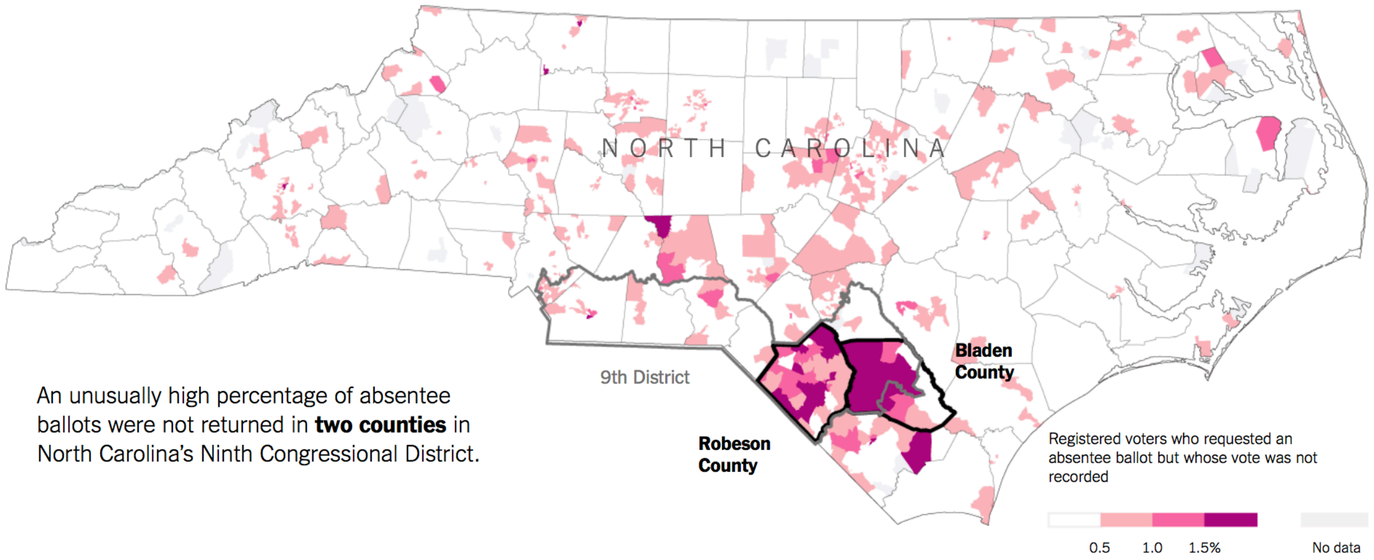
2. Network:
   Nodes and links can also have attributes (e.g., size of nodes, thickness/directionality of links).

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

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



https://www.nytimes.com/interactive/2016/09/12/science/earth/ocean-warming-climate-change.html



https://www.nytimes.com/2018/12/07/upshot/mapped-why-voting-anomalies-are-impossible-to-ignore-in-north-carolina.html

Attribute / Data Types (remember DSC 80?)

Nominal **=, ≠**	Labels or categories. E.g., Fruits: apples, bananas, cantaloupes, ...
Ordinal =, ≠, **<, >**	Ordered. E.g., Quality of eggs: Grade AA, A, B
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 / meaningful baseline). 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

Attribute Type
Burned vs. Not-Burned (N)
Hot, Warm, Cold (O)
Temperature Value (Q)

Conceptual Model
Temperature (°C)

Activity: U.S. Census

What are the types of these attributes (N/O/Q)?

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					34211
21	1850	45	0	2	341254
22	1850	50	0	1	321348

Think on your own for 1 minute

Activity: U.S. Census

What are the types of these attributes (N/O/Q)?

People Count: # of people in group

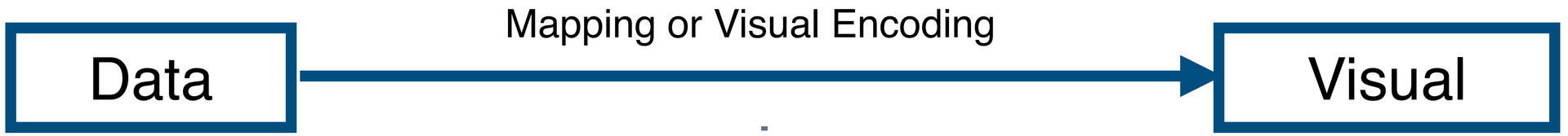
Year: 1850 – 2000 (every decade)

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

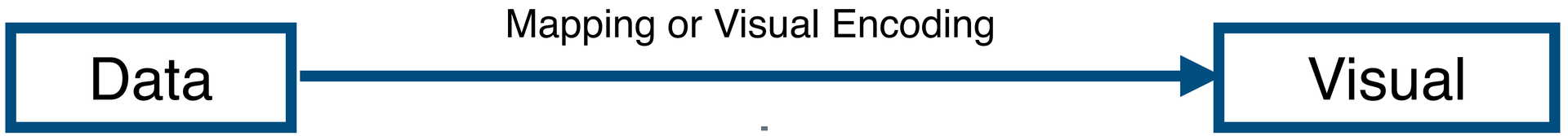


Expressiveness

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 what the facts are.

Mackinlay, Jock. "Automating the design of graphical presentations of relational information." *Acm Transactions On Graphics (Tog)* 5.2 (1986): 110-141.



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

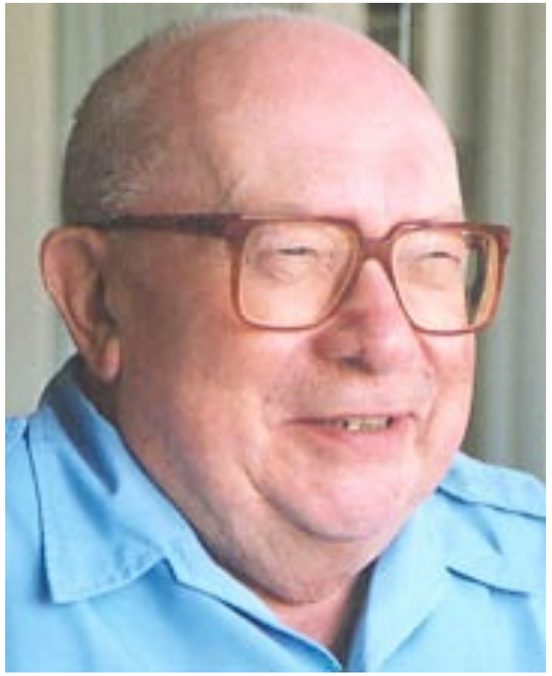
A visualization is more *effective* than another if the information it conveys *is more readily perceived* than the information in the other visualization

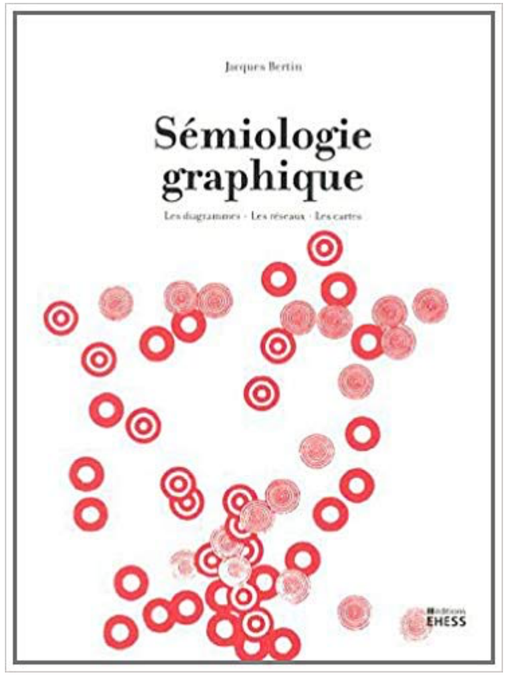
Image models give us a way of talking about what is more readily perceived.

Mackinlay, Jock. "Automating the design of graphical presentations of relational information." *Acm Transactions On Graphics (Tog)* 5.2 (1986): 110-141.

Image Models

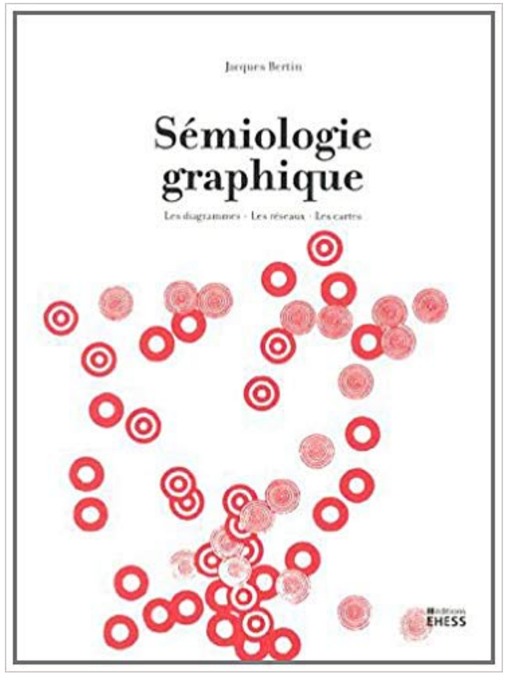
The Semiology of Graphics (1967)





Jacques Bertin (1918 – 2010)
French cartographer

The Semiology of Graphics (1967)

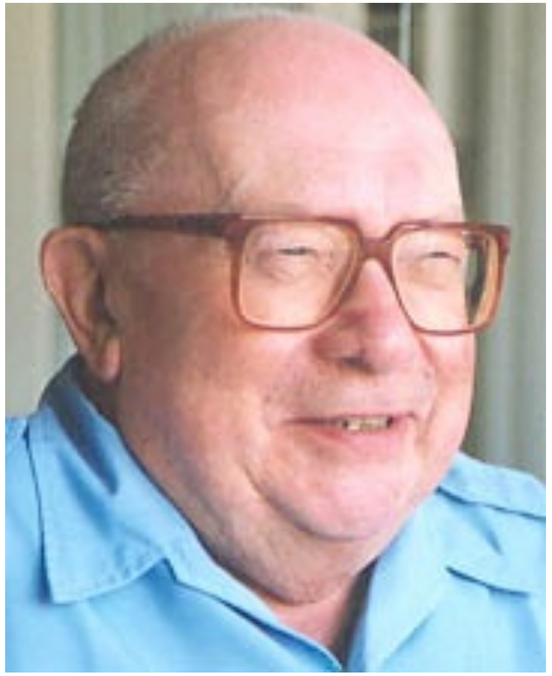


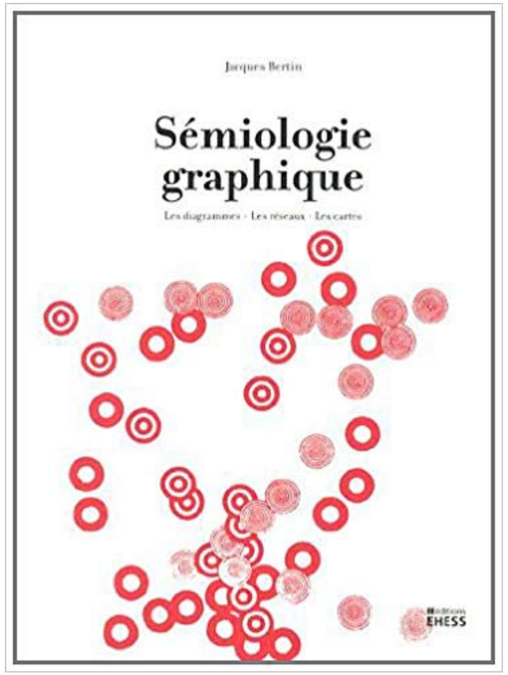
Jacques Bertin (1918 – 2010)
French cartographer

Study of visual signs and symbols used in graphical representations to convey information effectively.

The Semiology of Graphics (1967)

Study of signs and how cultures use them.



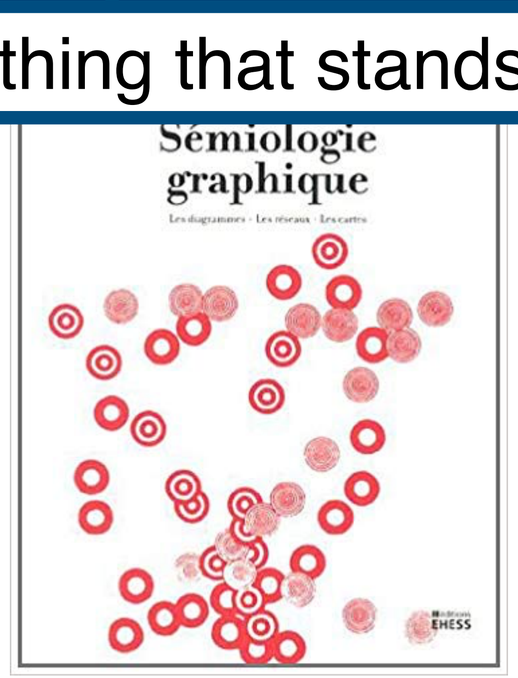


Jacques Bertin (1918 – 2010)
French cartographer

The Semiology of Graphics (1967)

Study of signs and how cultures use them.

Anything that stands for something other than itself.



"Metal painted red"?

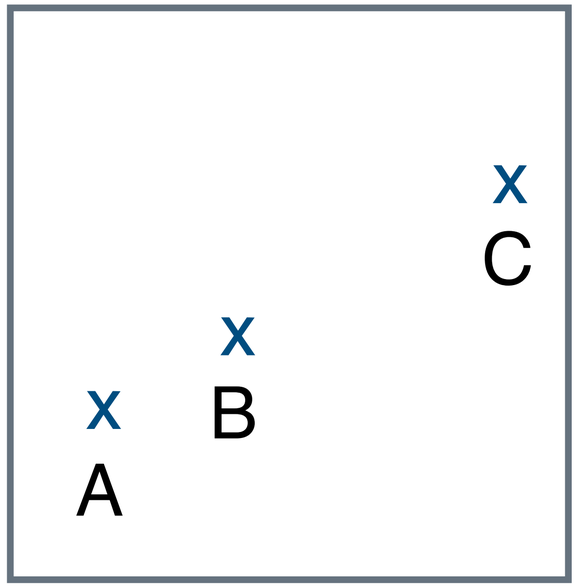
or

"Hit the brakes!"

Jacques Bertin (1918 – 2010)
French cartographer

Bertin proposed that graphical elements such as points, lines, shapes, colors, and textures could be systematically organized and manipulated to represent different attributes of data

This work laid the foundation for modern data visualization techniques.



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.

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

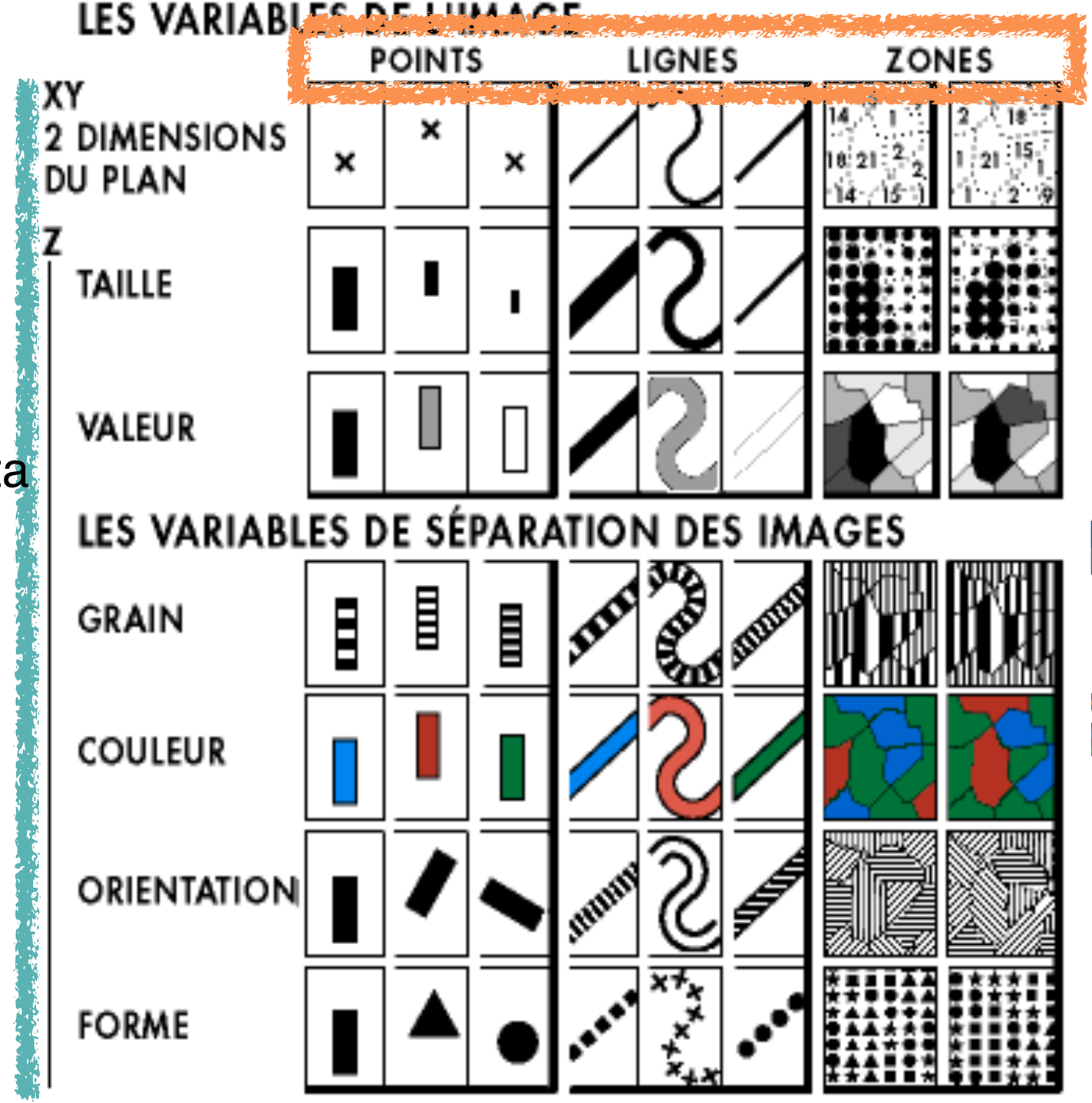
–Bertin

Visual Variables

Also called visual *channels*.

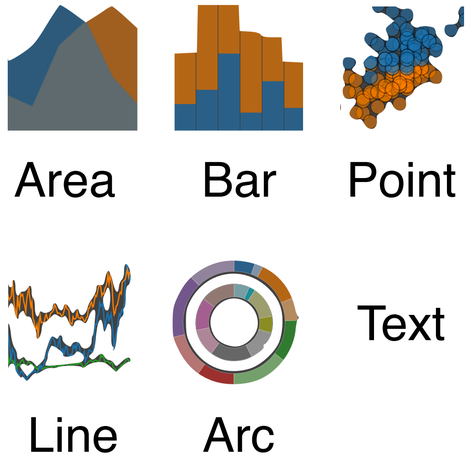
Used to encode data values as characteristics of marks.

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



Marks

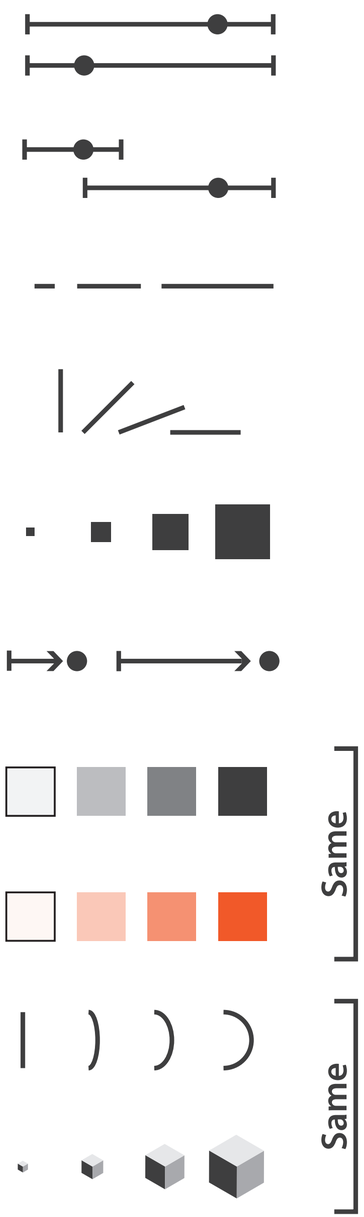
Basic graphical elements that represent data items.



Channels: Expressiveness Types and Effectiveness Ranks

Magnitude Channels: Ordered 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)



Identity Channels: Categorical Attributes

- Spatial region
- Color hue
- Motion
- Shape

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

Channels: Expressiveness Types and Effectiveness Ranks

Magnitude Channels: Ordered Attributes

Position on common scale

Position on unaligned scale

Color saturation

Curvature

Volume (3D size)

Most

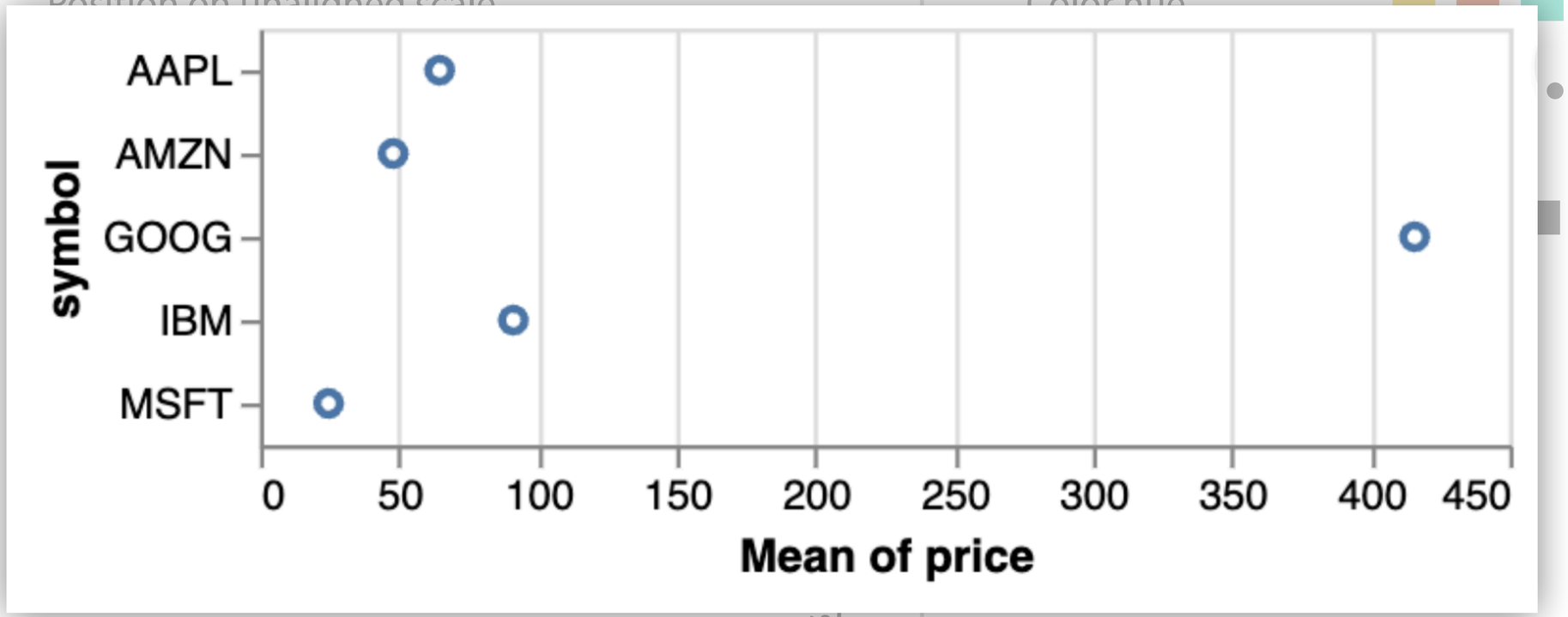
Same

Least

Identity Channels: Categorical Attributes

Spatial region

Color hue



Perceive dot positions on common x-axis scale

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

Channels: Expressiveness Types and Effectiveness Ranks

Magnitude Channels: Ordered Attributes

Position on common scale

Position on unaligned scale

Length (1D size)

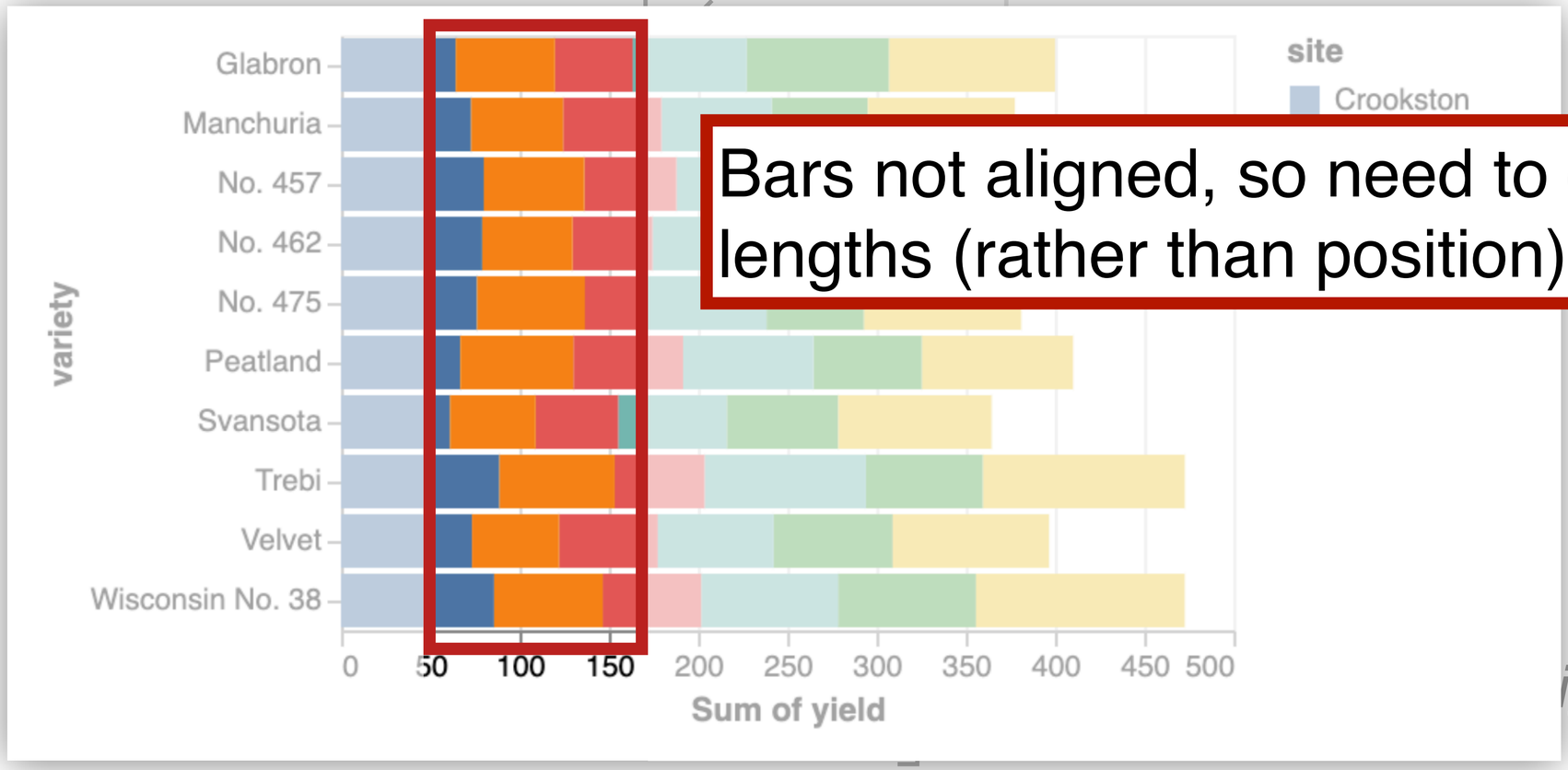
Most

Identity Channels: Categorical Attributes

Spatial region

Color hue

Motion

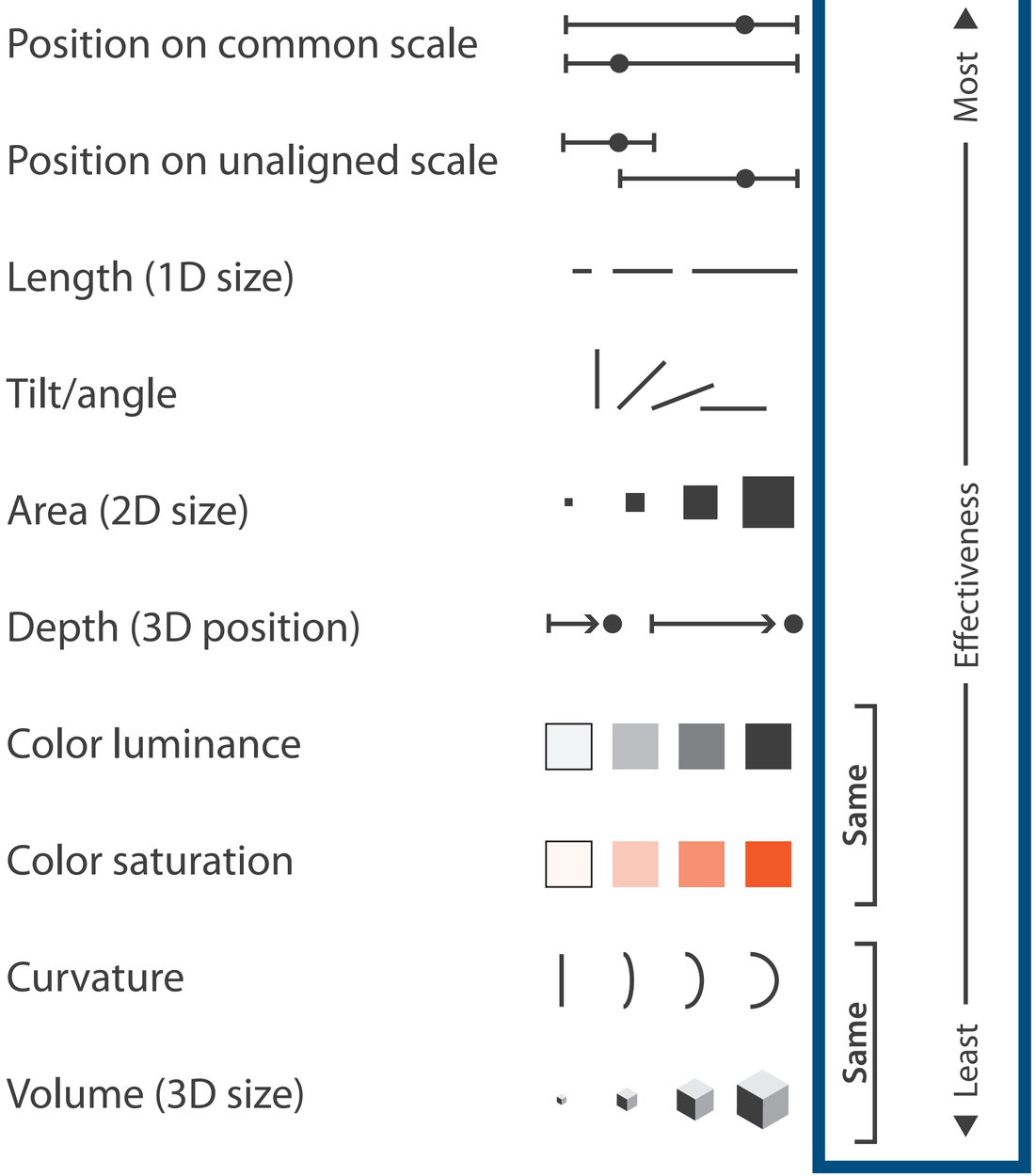


Bars not aligned, so need to compare lengths (rather than position)

Visualization
ign (2014).

Channels: Expressiveness Types and Effectiveness Ranks

Magnitude Channels: Ordered Attributes



Identity Channels: Categorical Attributes

Most effective to least effective

Color hue

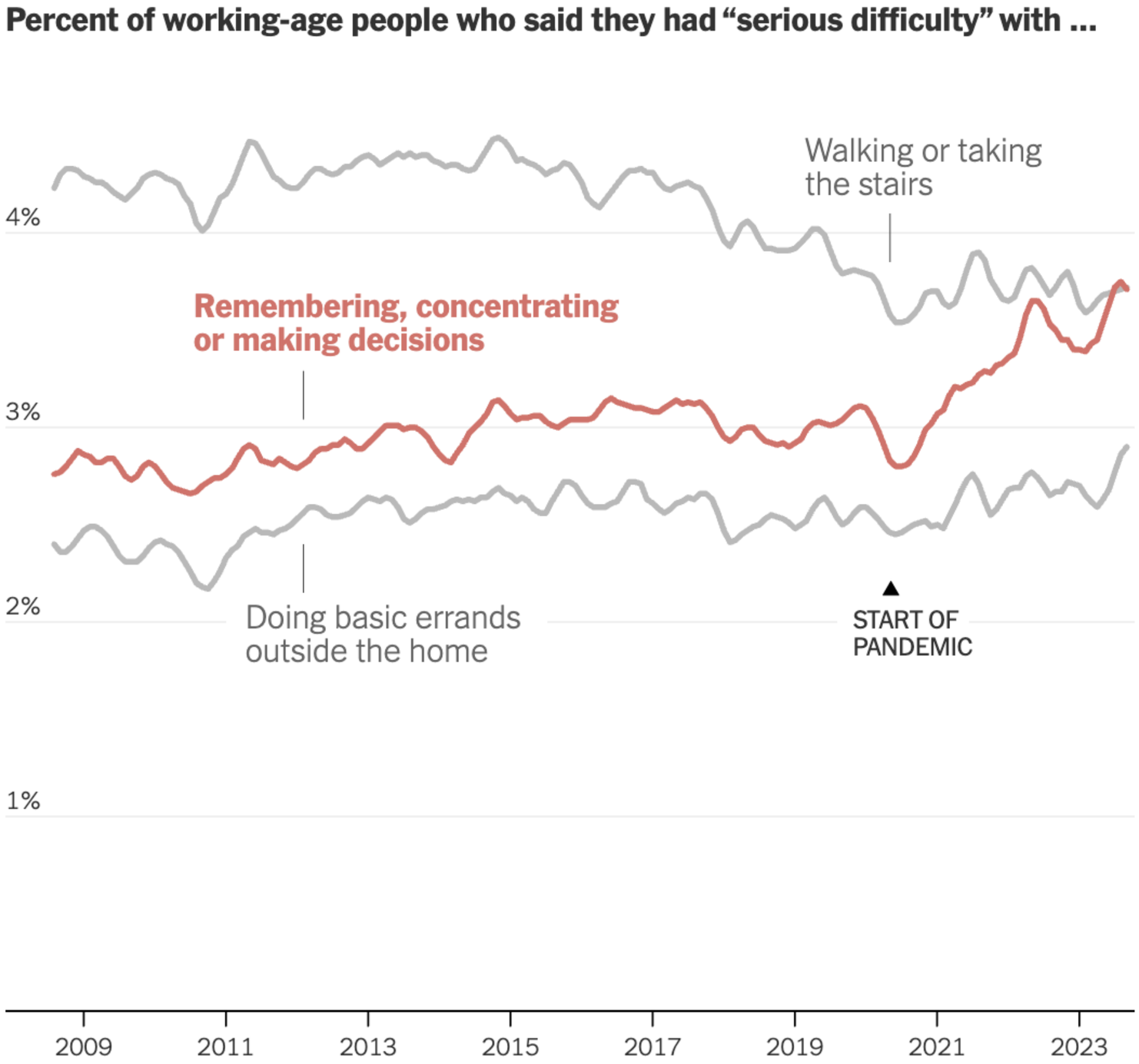
Top of scale = easiest for people to make accurate comparisons

Shape

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

Name that ~~chart!~~

Visual Encoding!

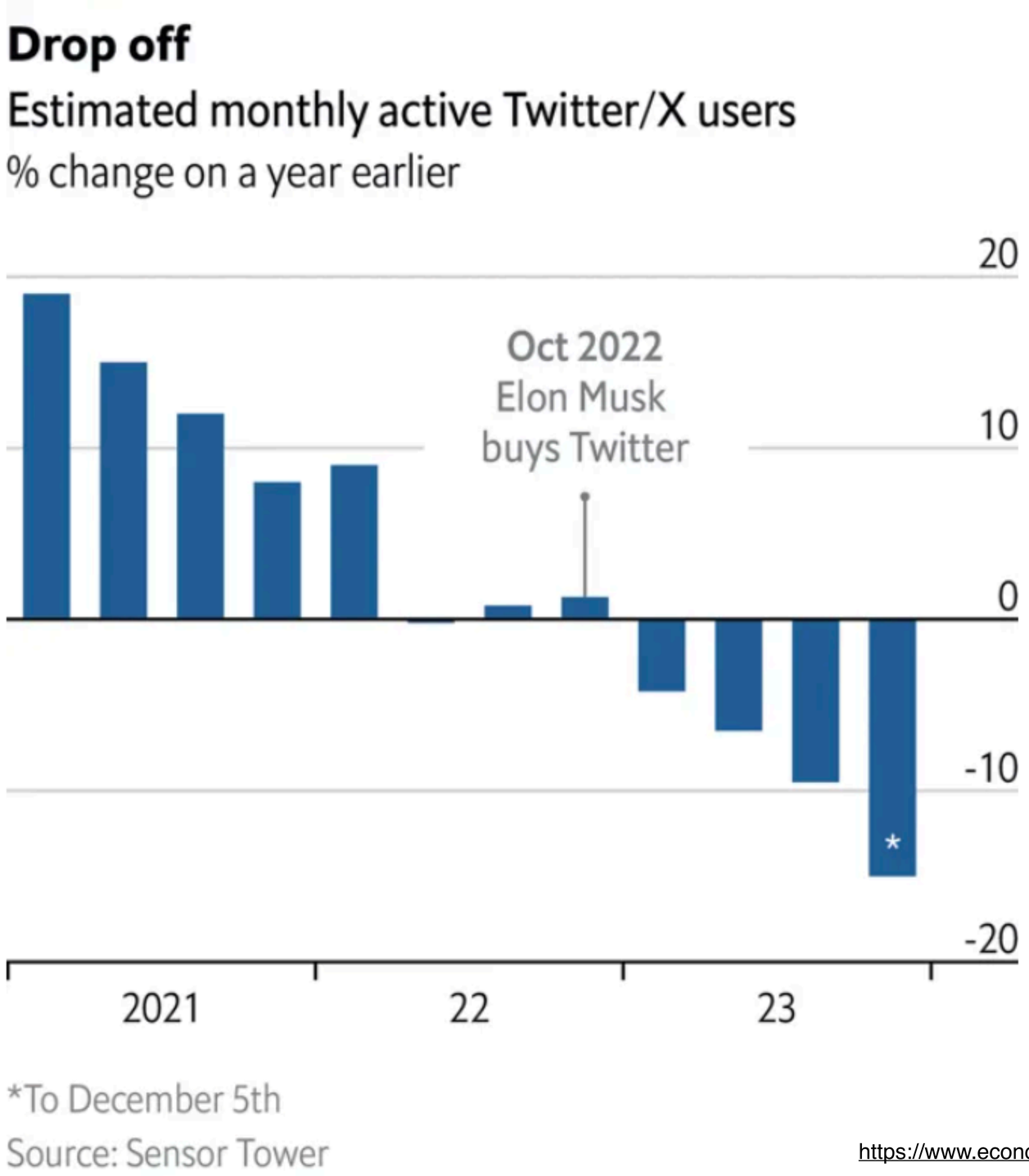


Mark: `line`

X-axis: `date` (Q-interval)
Y-axis: `percent` (Q-ratio)

What about color?

https://www.nytimes.com/2024/01/10/upshot/covid-pandemic-wave.html



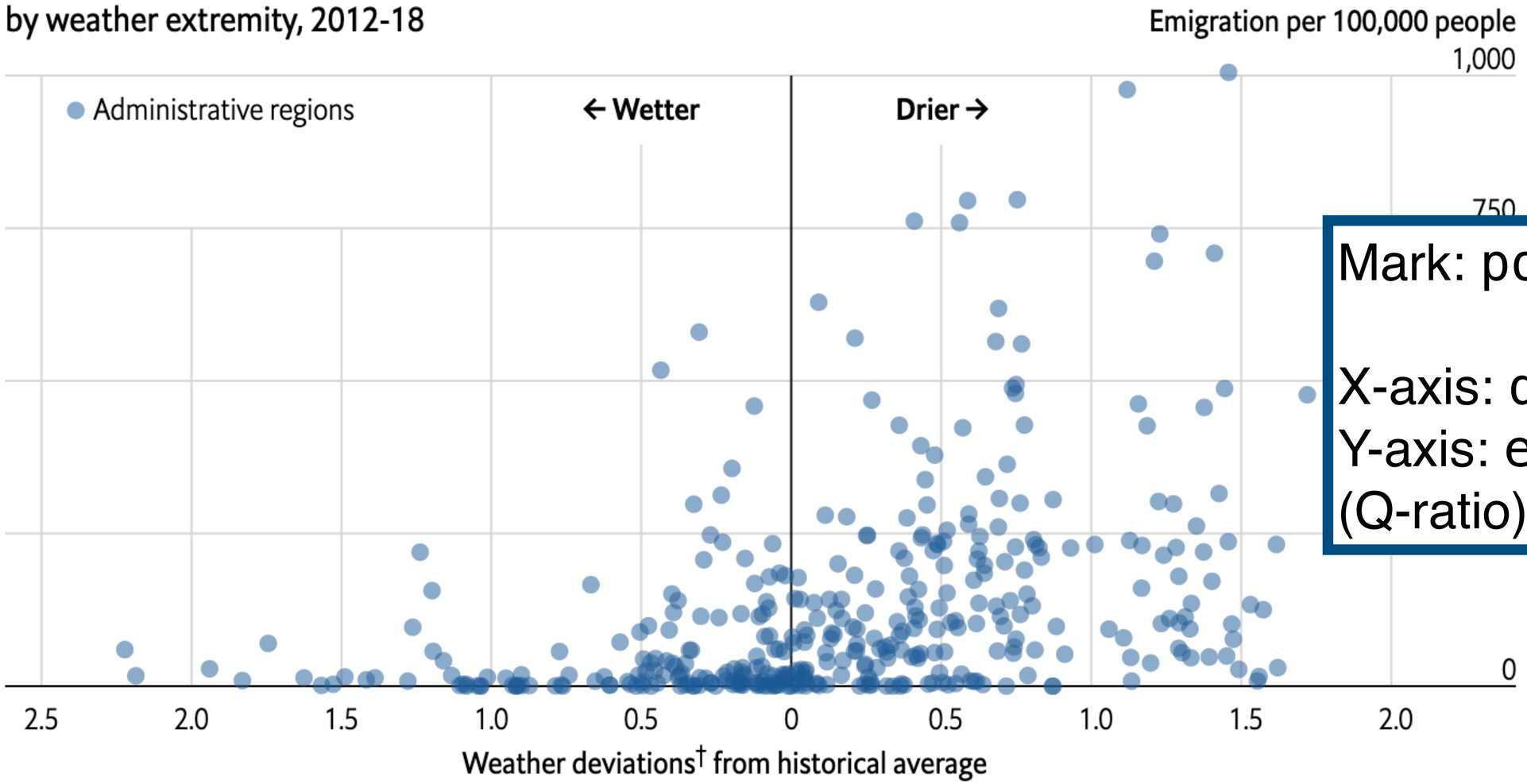
Mark: `bar`

X-axis: `date` (Q-interval)
Y-axis: `percent` (Q-ratio)

https://www.economist.com/graphic-detail/2023/12/20/has-twitter-now-x-become-more-right-wing

Spotting a trend

Emigration from the Northern Triangle* to United States, by weather extremity, 2012-18



Mark: `point`

X-axis: `dev` (Q-ratio)
Y-axis: `emigration` (Q-ratio)

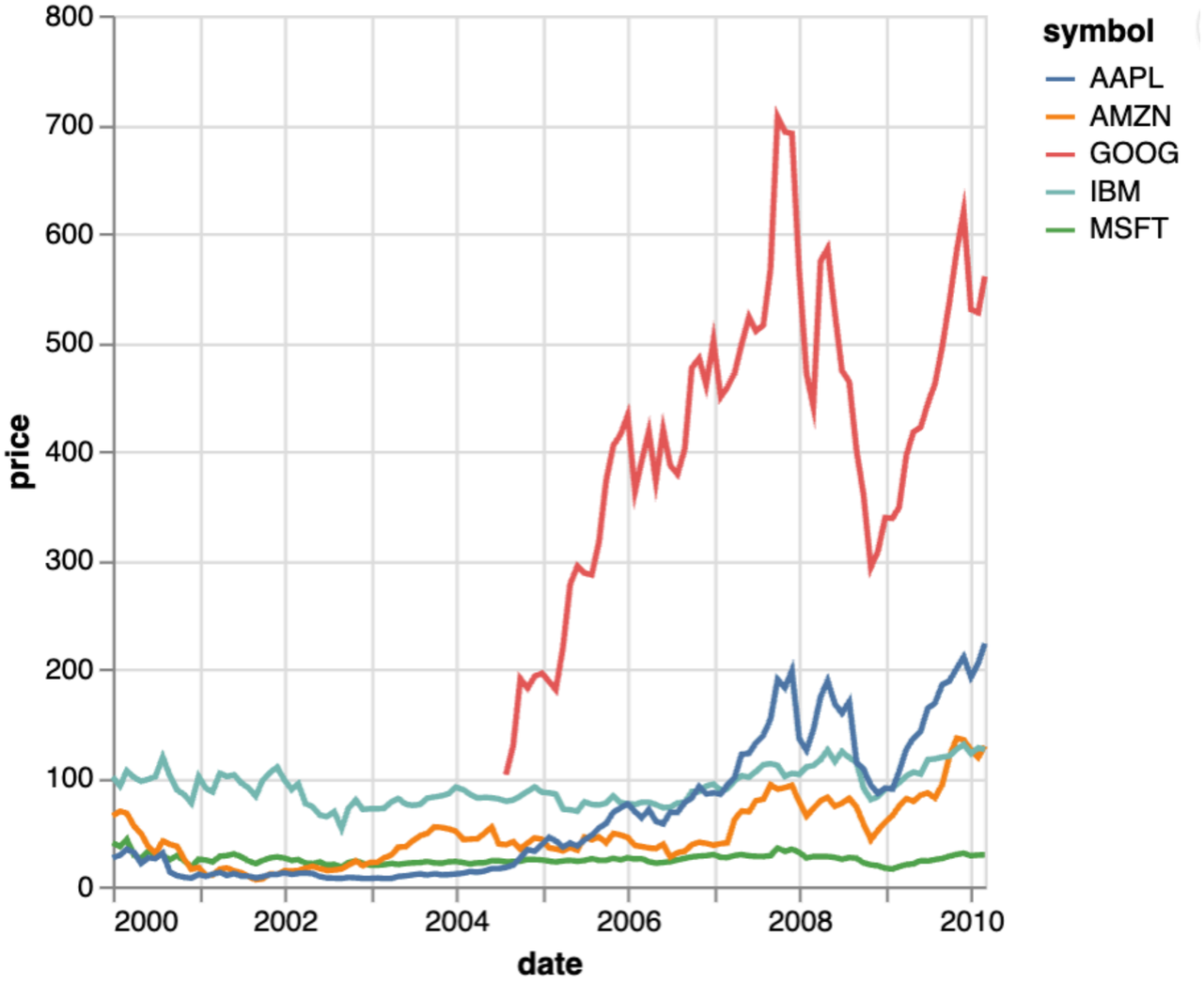
*El Salvador, Guatemala and Honduras †Using the Standardised Precipitation-Evapotranspiration Index three-month average

Source: "Dry growing seasons predicted Central American migration to the US from 2012 to 2018", by A. Linke et al., 2023

https://www.economist.com/graphic-detail/2023/11/17/why-central-americans-migrate-to-the-united-states-when-they-do

https://www.nature.com/articles/s41598-023-43668-9

Example from Lab 1

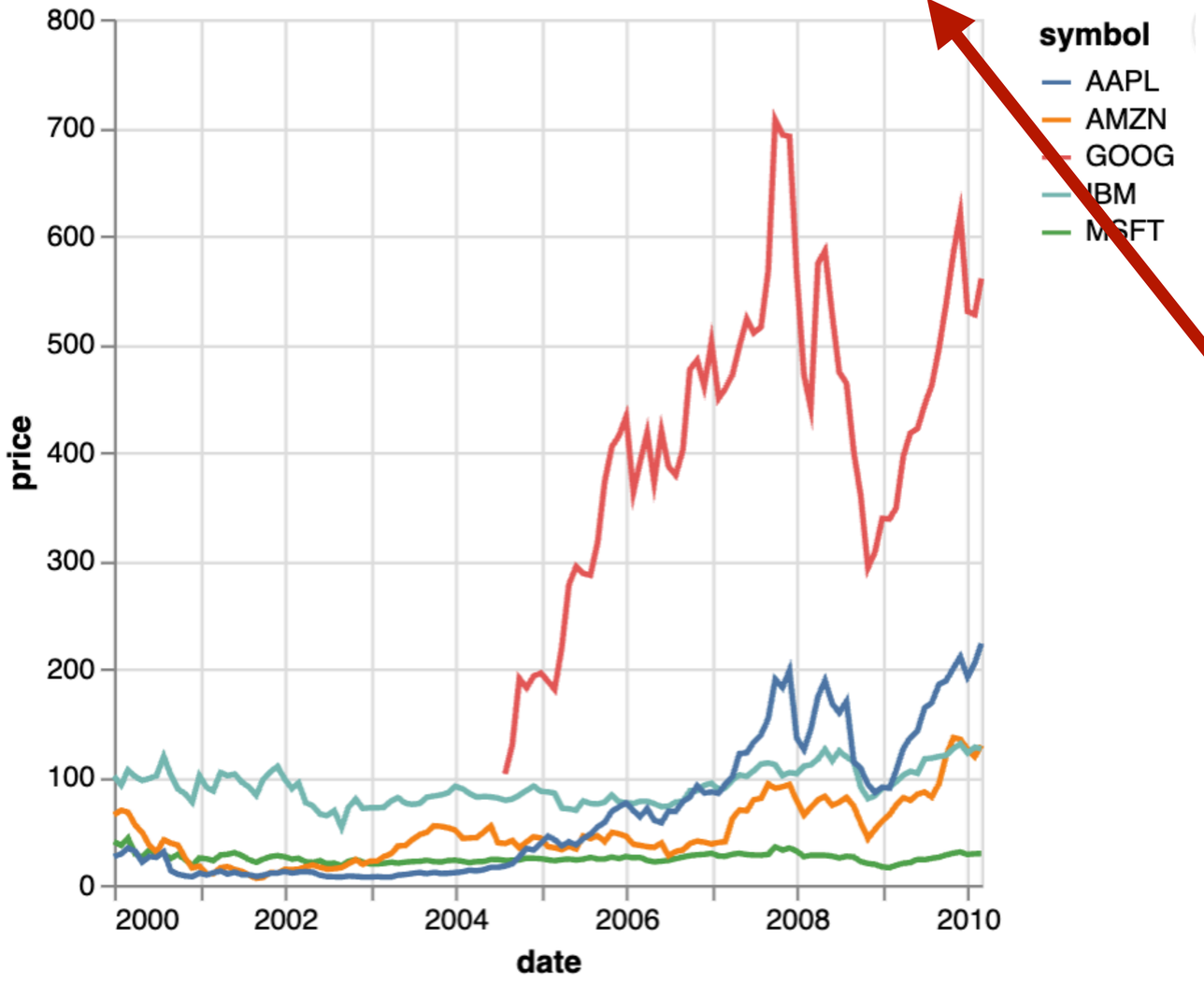


Mark: `line`

X-axis: `date` (Q-interval)
Y-axis: `price` (Q-ratio)
Color: `symbol` (N)

Example from Lab 1

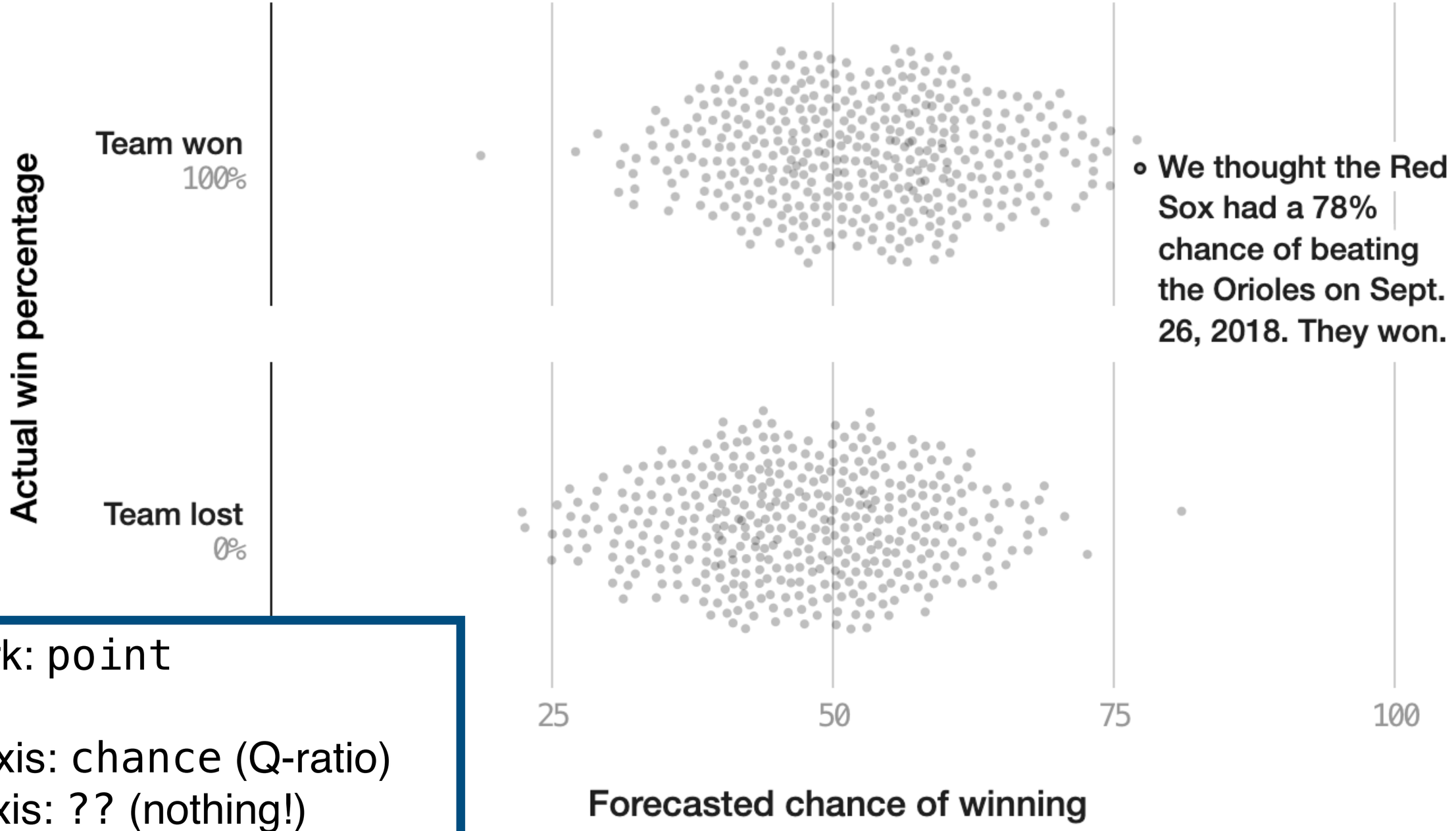
```
alt.Chart(stocks_df).mark_line().encode(
    x="date:T",
    y="price",
    color="symbol",
)
```



Mark: `line`

X-axis: `date` (Q-interval)
Y-axis: `price` (Q-ratio)
Color: `symbol` (N)

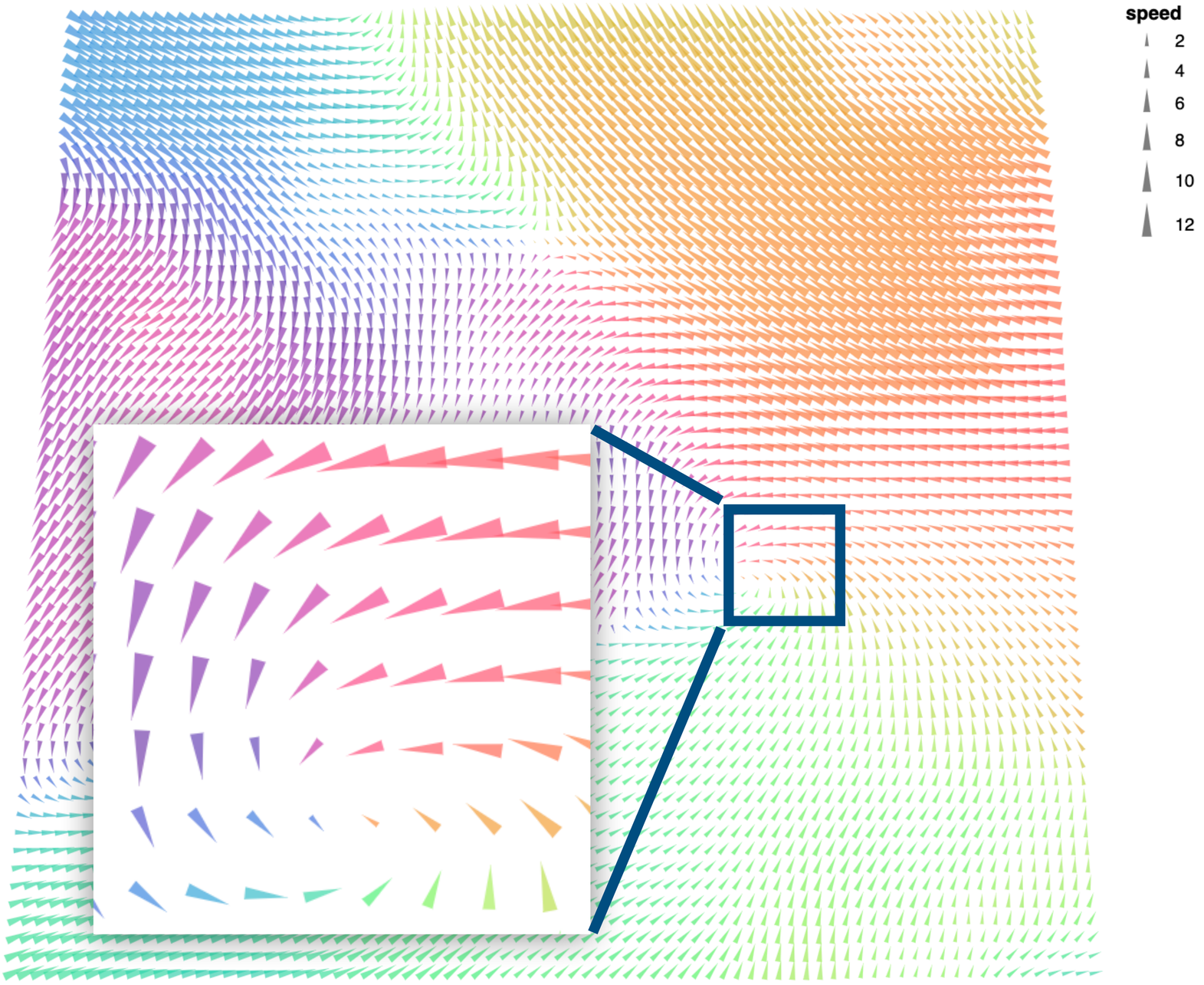
Notice how Altair lets us specify the mark, then the encodings!



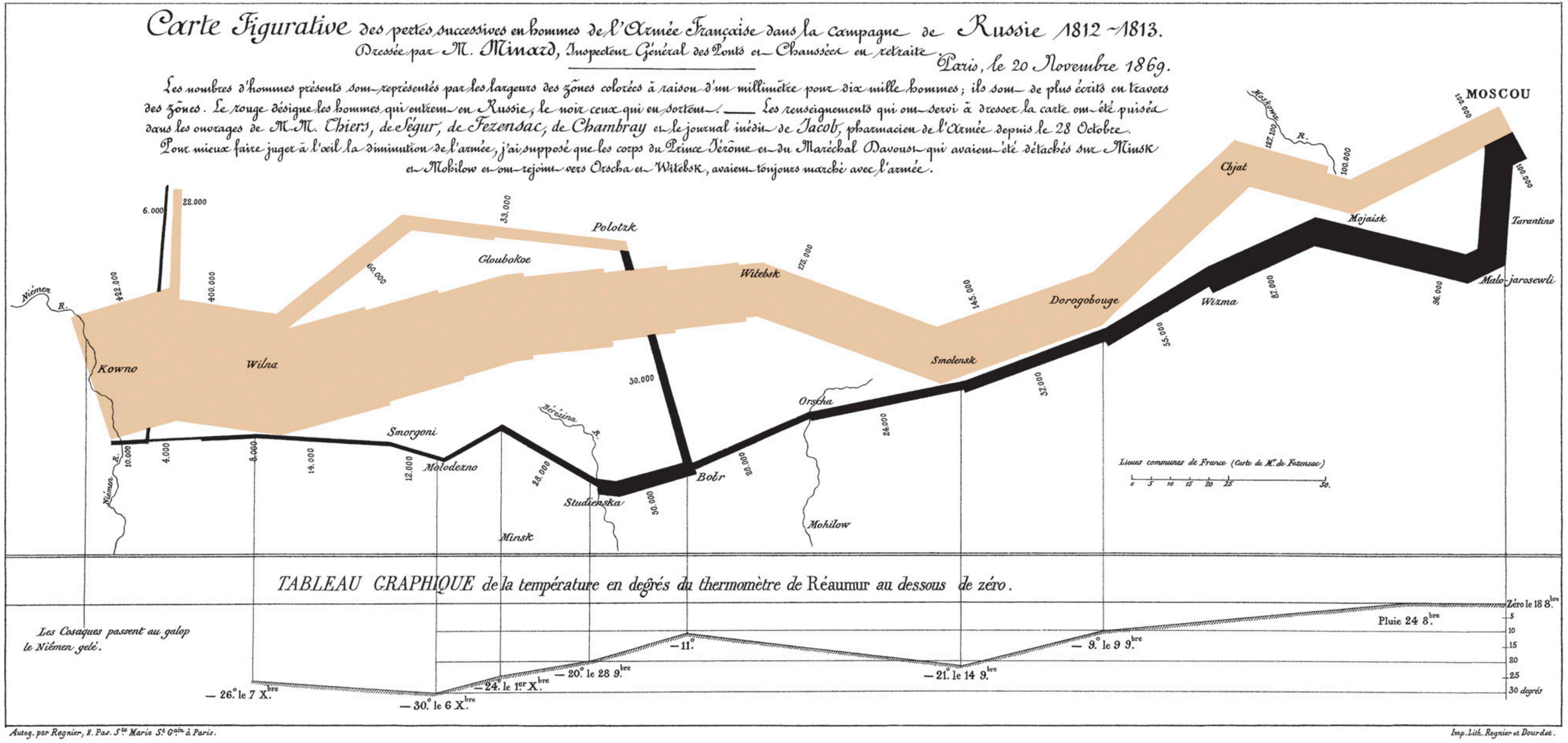
Mark: `point`

X-axis: `chance` (Q-ratio)
Y-axis: ?? (nothing!)

https://projects.fivethirtyeight.com/checking-our-work/



https://vega.github.io/vega-lite/examples/point_angle_windvector.html



[Charles Minard, 1869]

Next time: Visual Encoding & Design