

Data and Image Models

DSC 106: Data Visualization

Sam Lau

UC San Diego

Announcements

Lab 1 due **today!**

Project 1 checkpoint due Tuesday, 4/7.

Final Project Showcase scheduled for Thurs Apr 11, 11:30am-2:30pm in the HDSI MPR.

FAQs on course logistics:

1. Can I use ChatGPT / CoPilot? Yes!
2. I already have a personal website, can I reuse it for Lab 1 (and beyond)? Yes, just make sure everything we ask for in the Lab video is there.

Announcements

Sam traveling Weeks 2-3, no lecture. See Piazza post for details on lecture videos and participation.

If baby arrives early, course will run as planned with staff steering the ship.

A note about using ChatGPT (for writing prose)

We need to be having high quality conversations about AI: what it can and can't do, its many risks and pitfalls and how to integrate it into society in the most beneficial ways possible.

Strawman: "Don't call it AI! It's not actually intelligent—it's just spicy autocomplete."

Which one was generated by ChatGPT?

As artificial intelligence continues to transform our world, it is crucial to approach its development and deployment with caution, recognizing the potential for unintended harms alongside its vast promise.

From biased decisions to privacy risks, AI misuse demands proactive oversight and thoughtful regulation.

A note about using ChatGPT (for writing prose)

We need to be having high quality conversations about AI: what it can and can't do, its many risks and pitfalls and how to integrate it into society in the most beneficial ways.

As artificial intelligence continues to transform our world, it is crucial to approach its development and deployment with caution, recognizing the

If it's super obvious your writeup is AI generated without any editing, you will lose 50% of the writeup score.

If it looks obviously AI-generated to the staff, it will also look AI-generated to professional data scientists.

Which one was generated by ChatGPT?

proactive oversight and thoughtful regulation.

Name that chart!

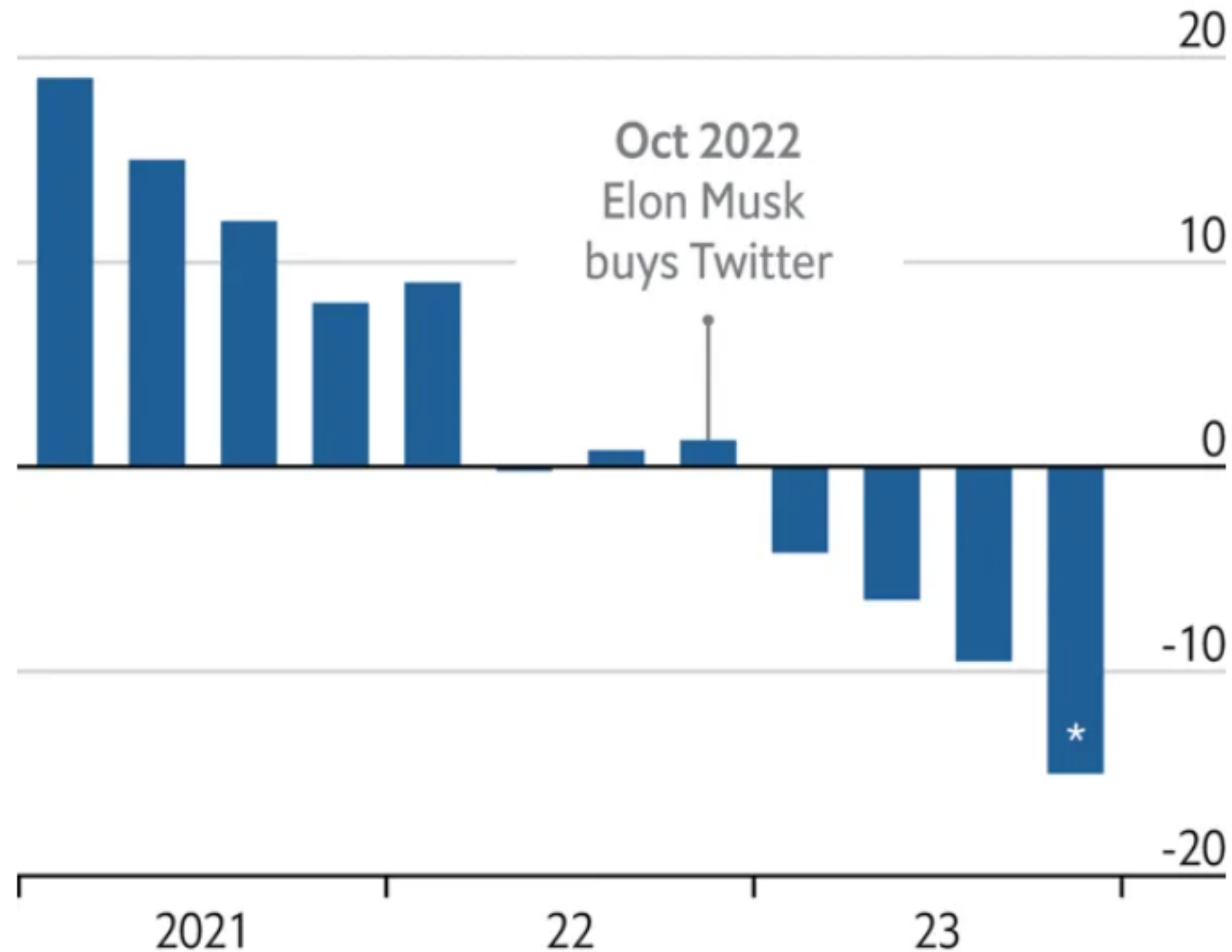
Percent of working-age people who said they had “serious difficulty” with ...



Drop off

Estimated monthly active Twitter/X users

% change on a year earlier

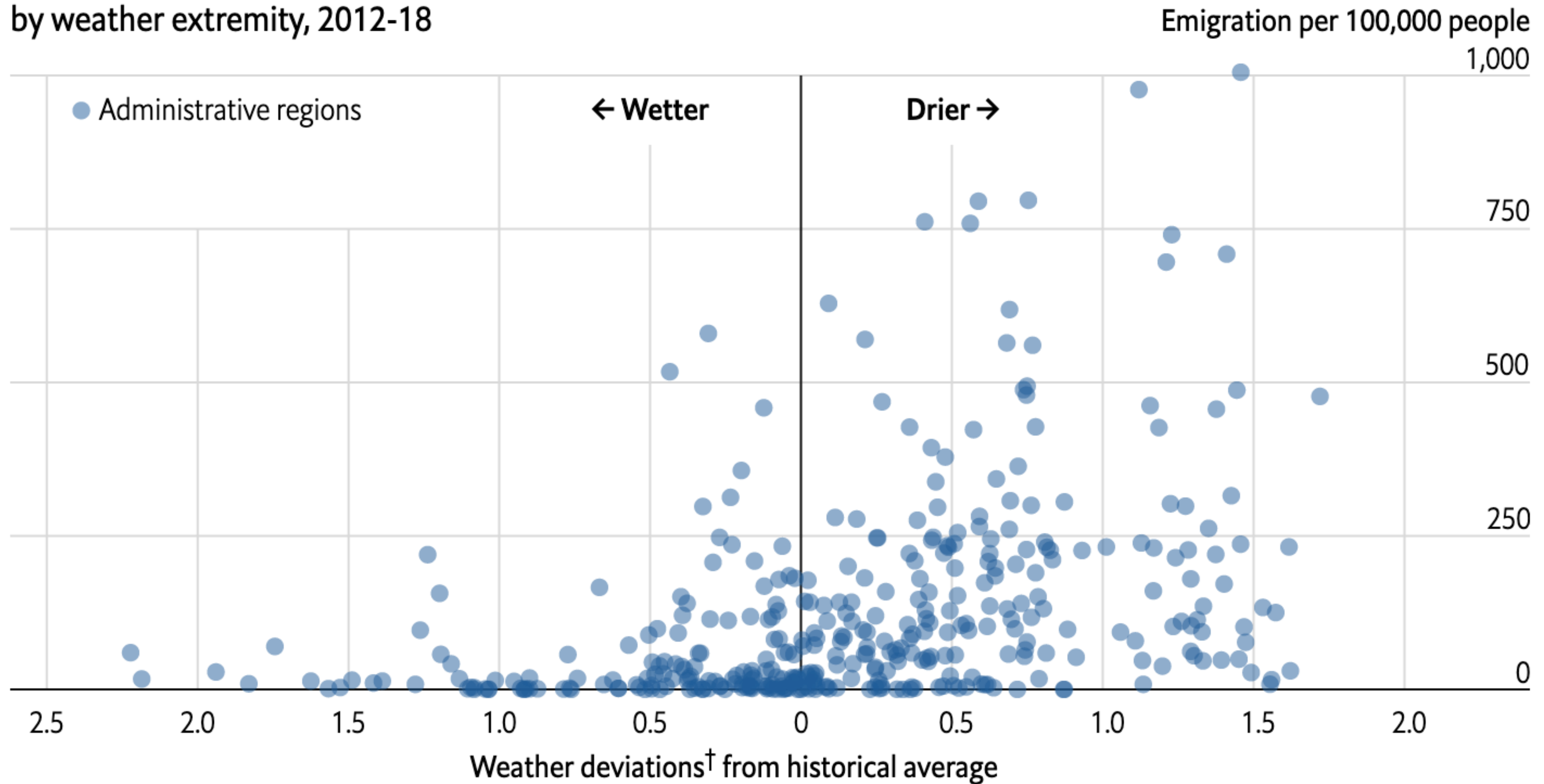


*To December 5th

Source: Sensor Tower

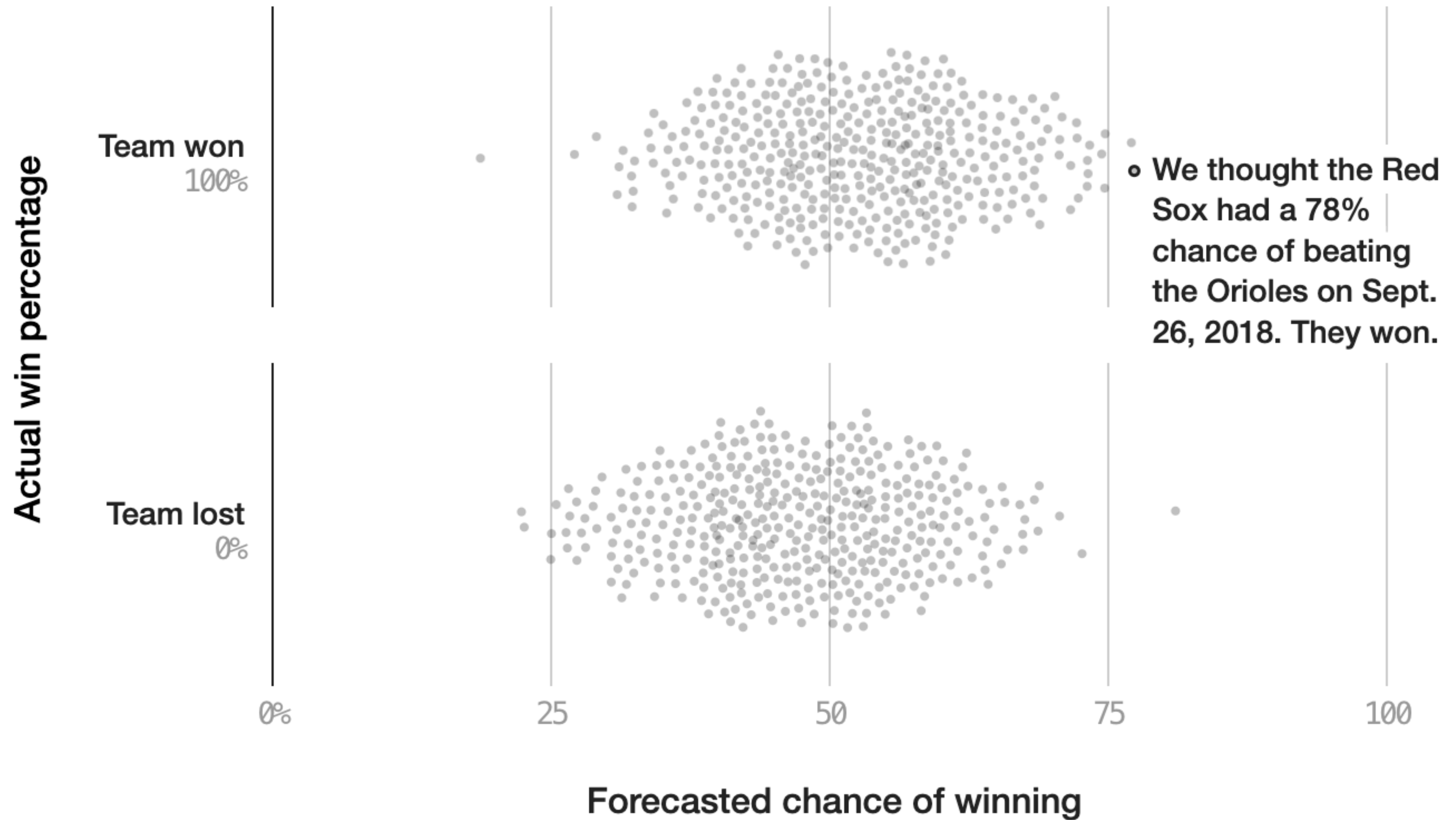
Spotting a trend

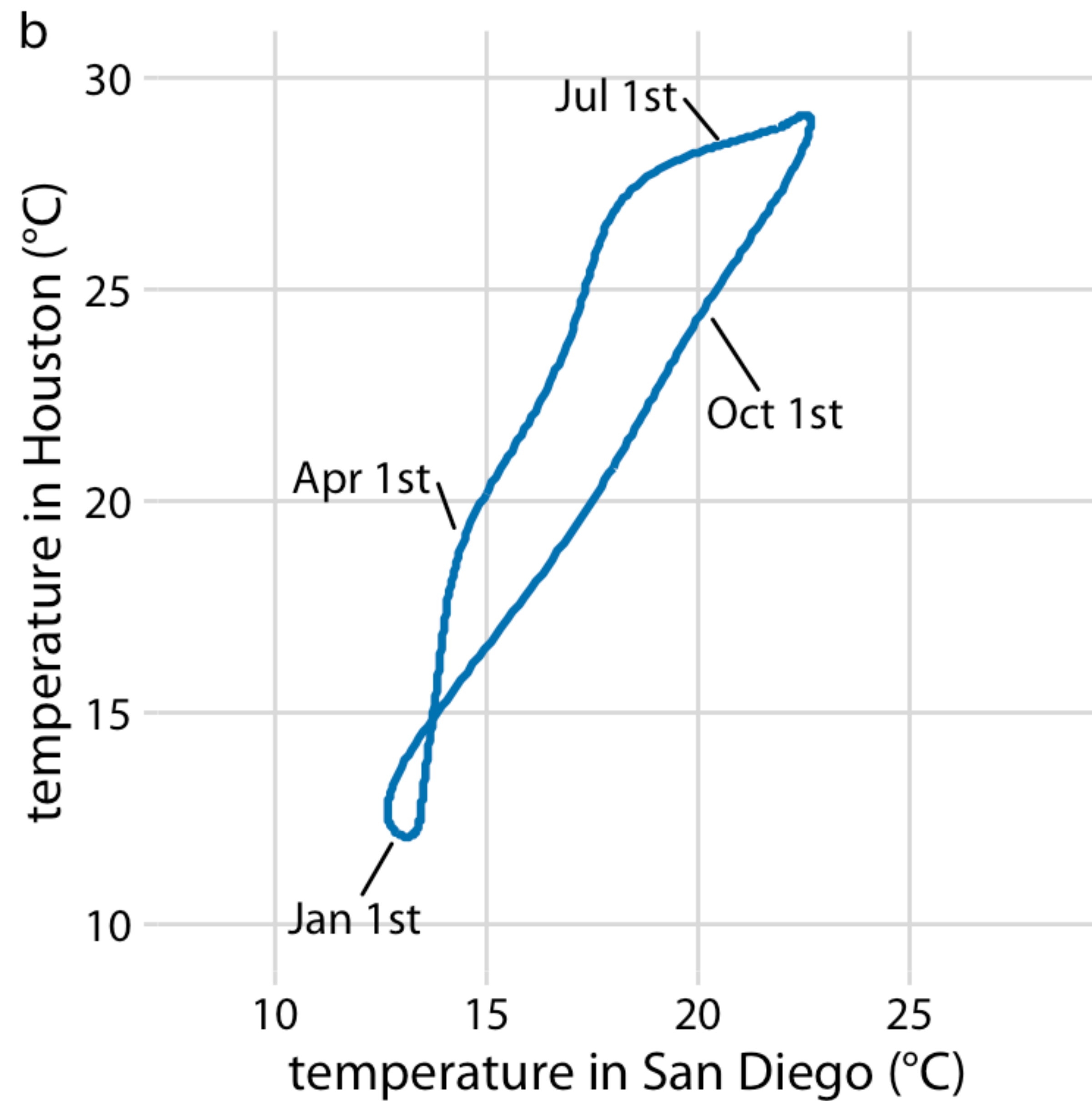
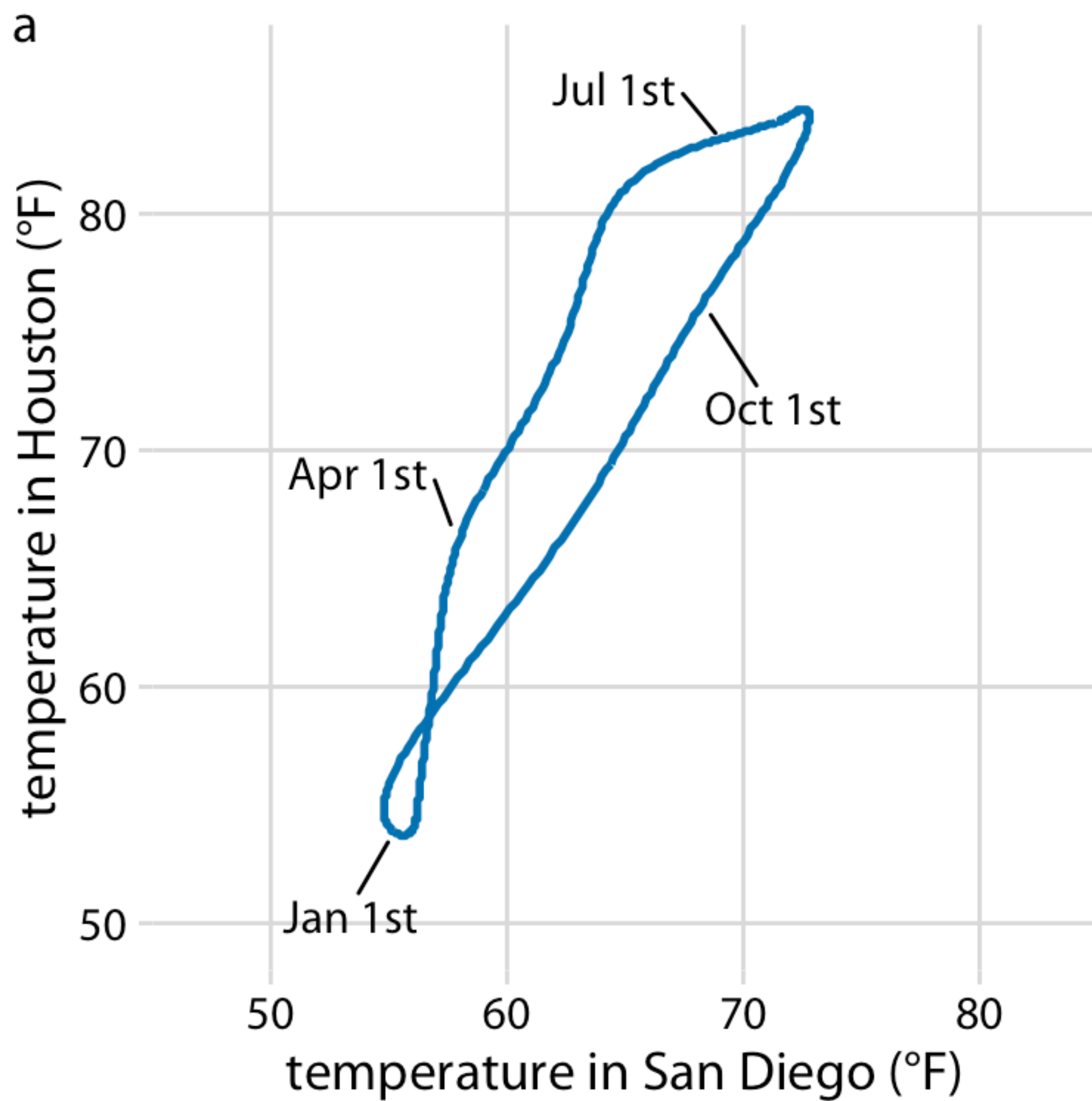
Emigration from the Northern Triangle* to United States,
by weather extremity, 2012-18

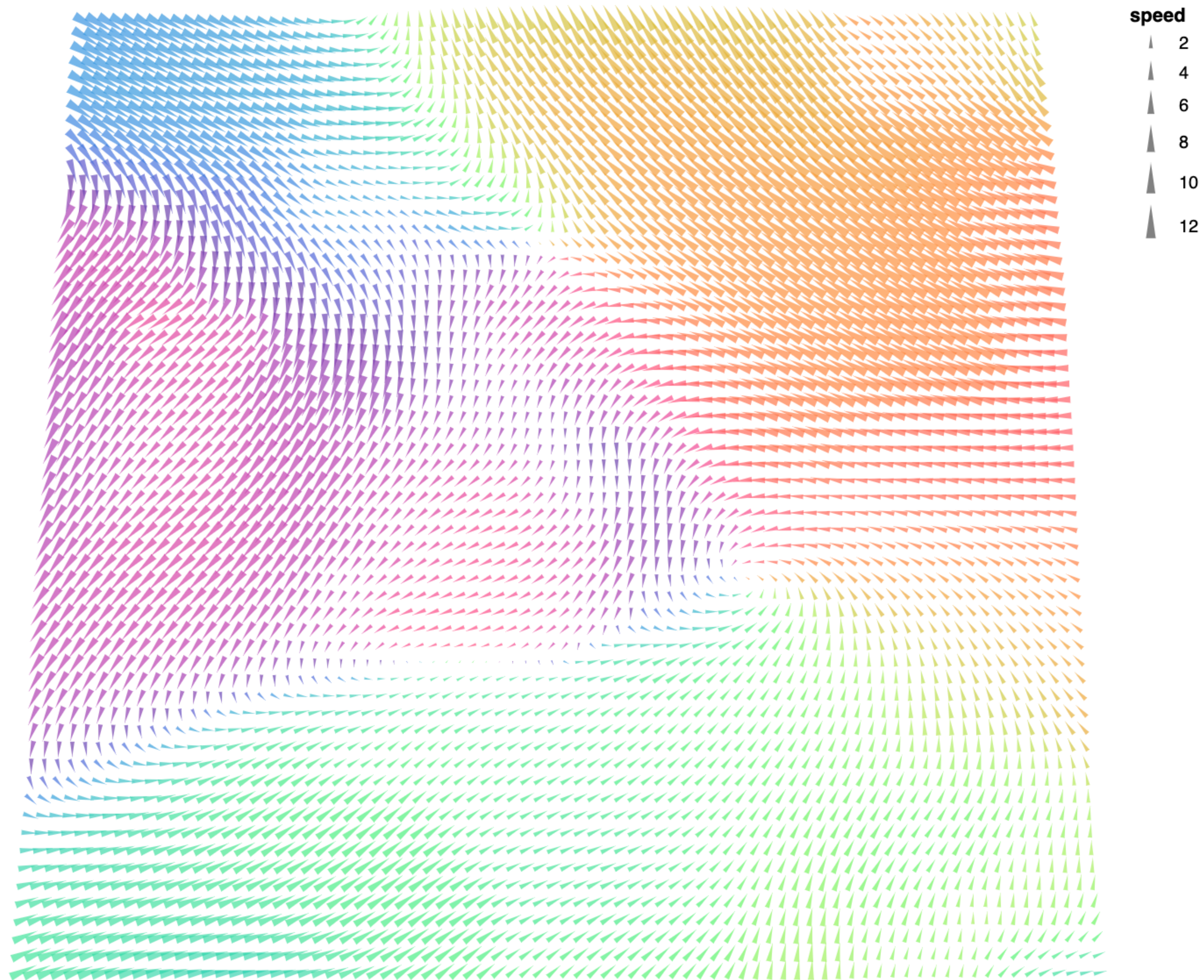


*El Salvador, Guatemala and Honduras [†]Using the Standardised Precipitation-Evapotranspiration Index three-month average

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







Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite. Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. — Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre. Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davout qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

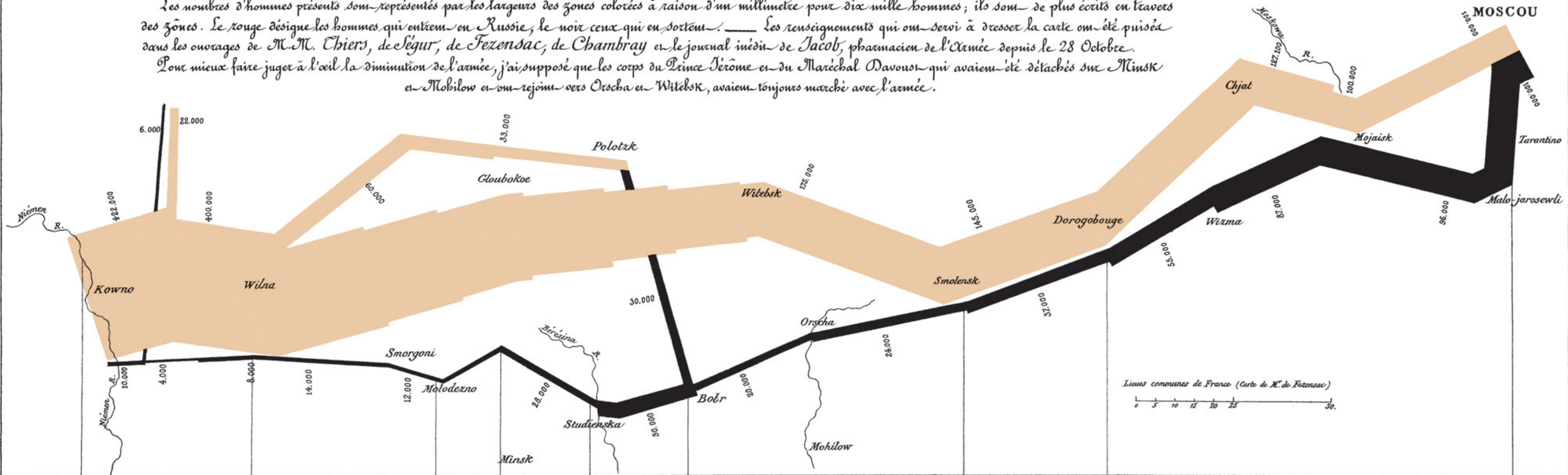
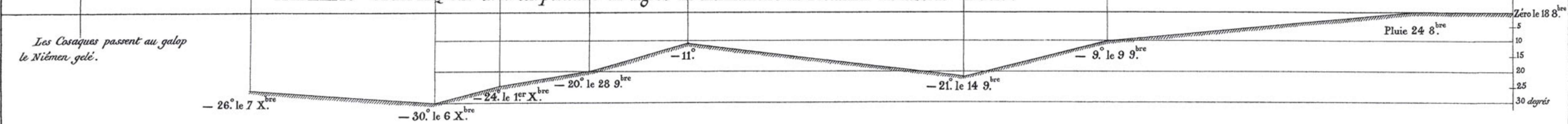
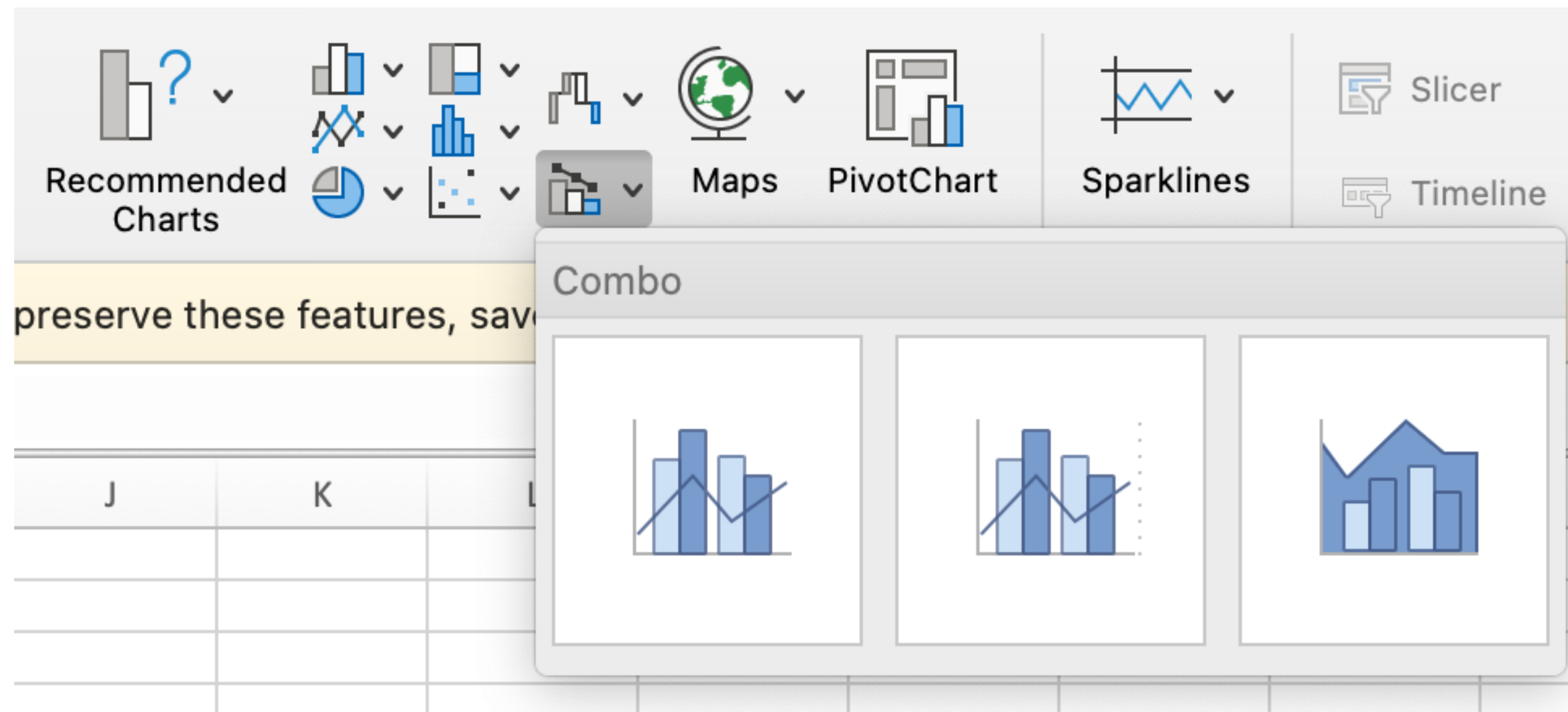
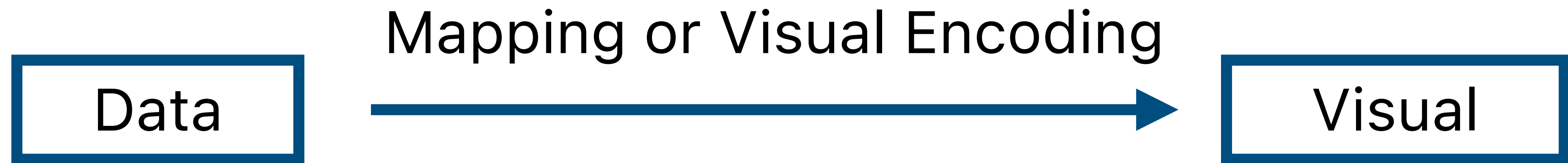


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.





Visualizing Data



Physical Data Types

int, float, string

Graphical Marks

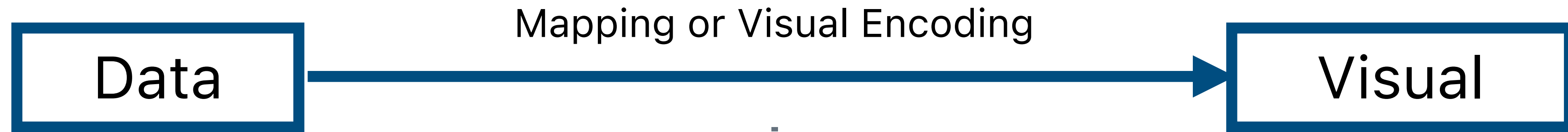
rect, line, point, area

Conceptual Data Types

temperature, location

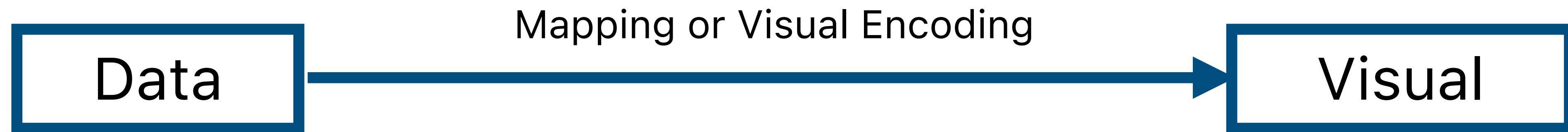
Visual Channels

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.

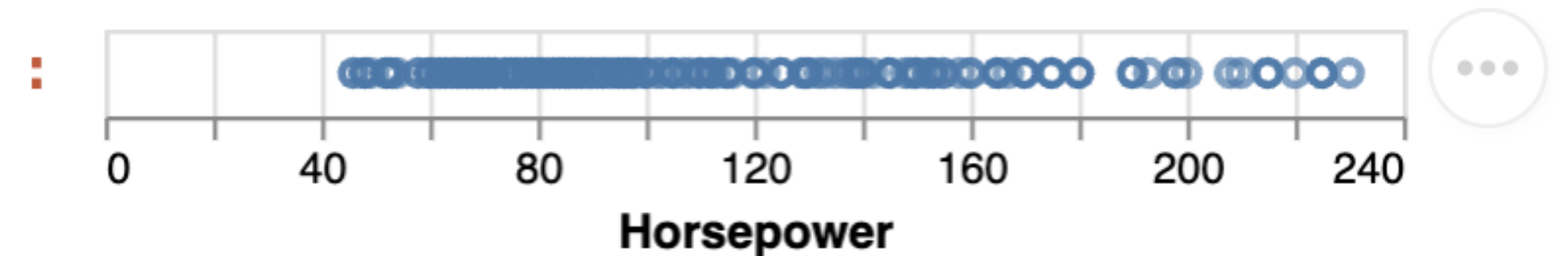


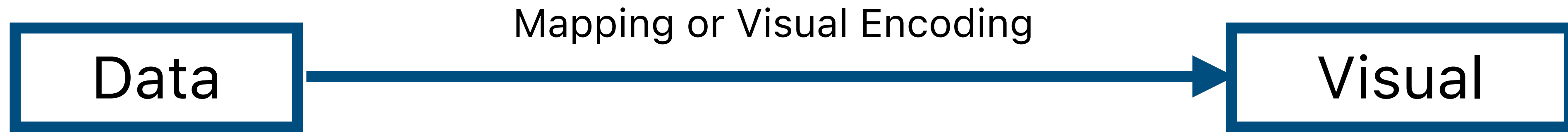
Expressiveness

Can't express the facts

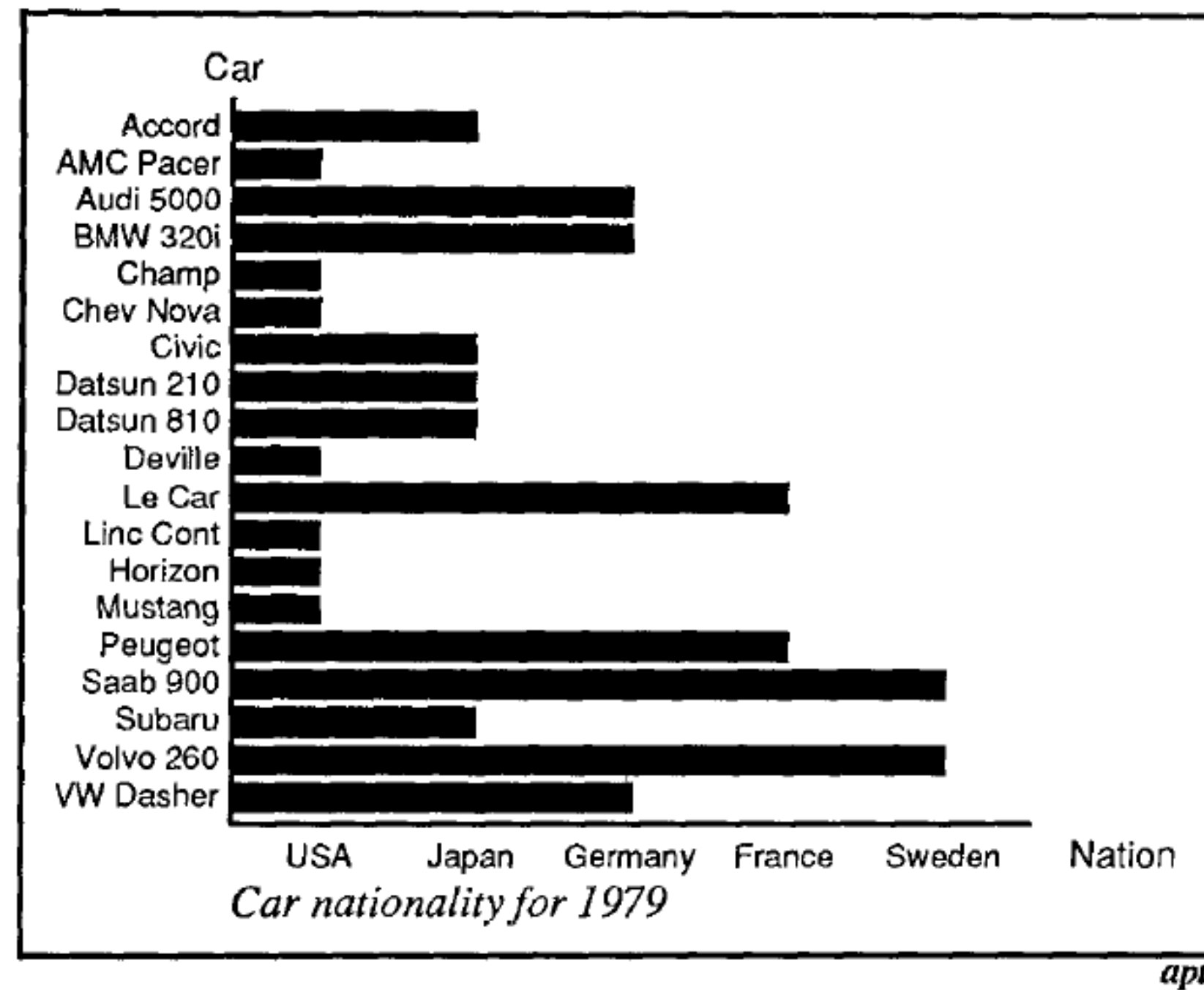
A dataset with many variables may be *inexpressive* in a single horizontal dot plot because multiple records are mapped to the same position.

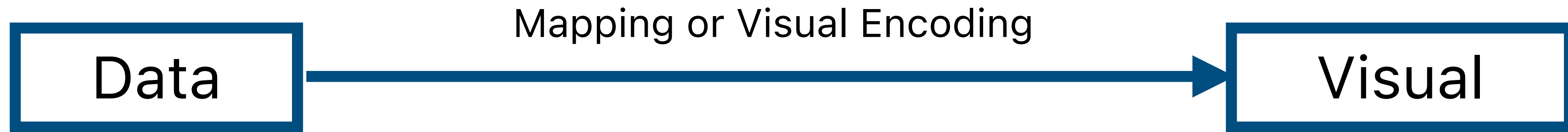
```
: alt.Chart(source).mark_point().encode(  
  x='Horsepower'  
)
```





Expressiveness





Expressiveness

Expresses facts not in the data

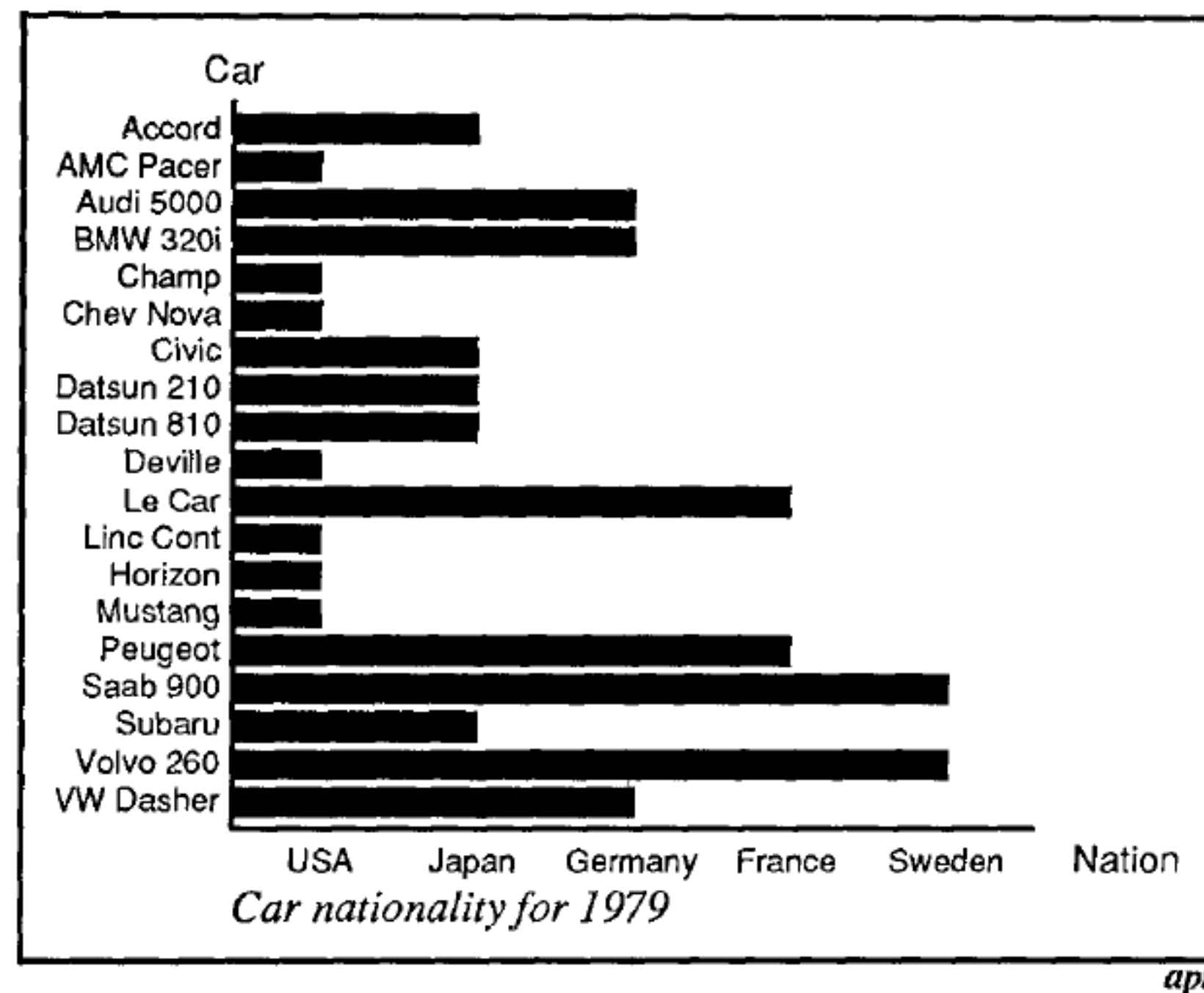
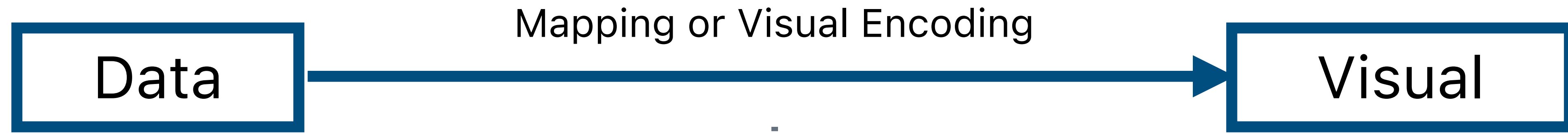
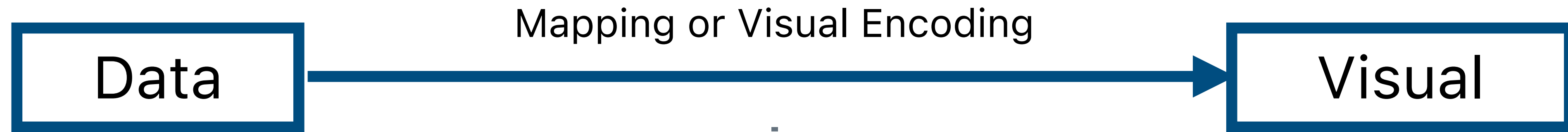


Fig. 11. Incorrect use of a bar chart for the *Nation* relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the *Nation* relation.



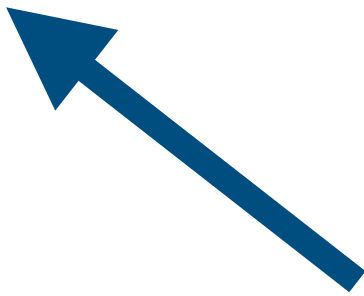
Expressiveness

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.



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.

Data Models

Conceptual Models vs. Data Models

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

	city	lat	lon	month	sunshine	rain
0	San Diego	32.715736	-117.161087	Mar	217	1.53
1	San Diego	32.715736	-117.161087	Apr	255	0.15
2	San Diego	32.715736	-117.161087	May	234	0.57
3	San Diego	32.715736	-117.161087	Jun	236	1.01
4	San Diego	32.715736	-117.161087	Jul	277	0.02
...	...	...	...	...	...	...
67	Miami	25.761681	-80.191788	Aug	263	8.88
68	Miami	25.761681	-80.191788	Sep	216	9.86
69	Miami	25.761681	-80.191788	Oct	215	6.33
70	Miami	25.761681	-80.191788	Nov	212	3.27
71	Miami	25.761681	-80.191788	Dec	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	\$					217	1.53
1	\$					255	0.15
2	\$					234	0.57
3	San Diego	32.715736	-117.161087	Apr		236	1.01
4						277	0.02
...	...	...	...	...	...	...	...
67	Miami	25.761681	-80.191788	Aug		263	8.88
68	Miami	25.761681	-80.191788	Sep		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

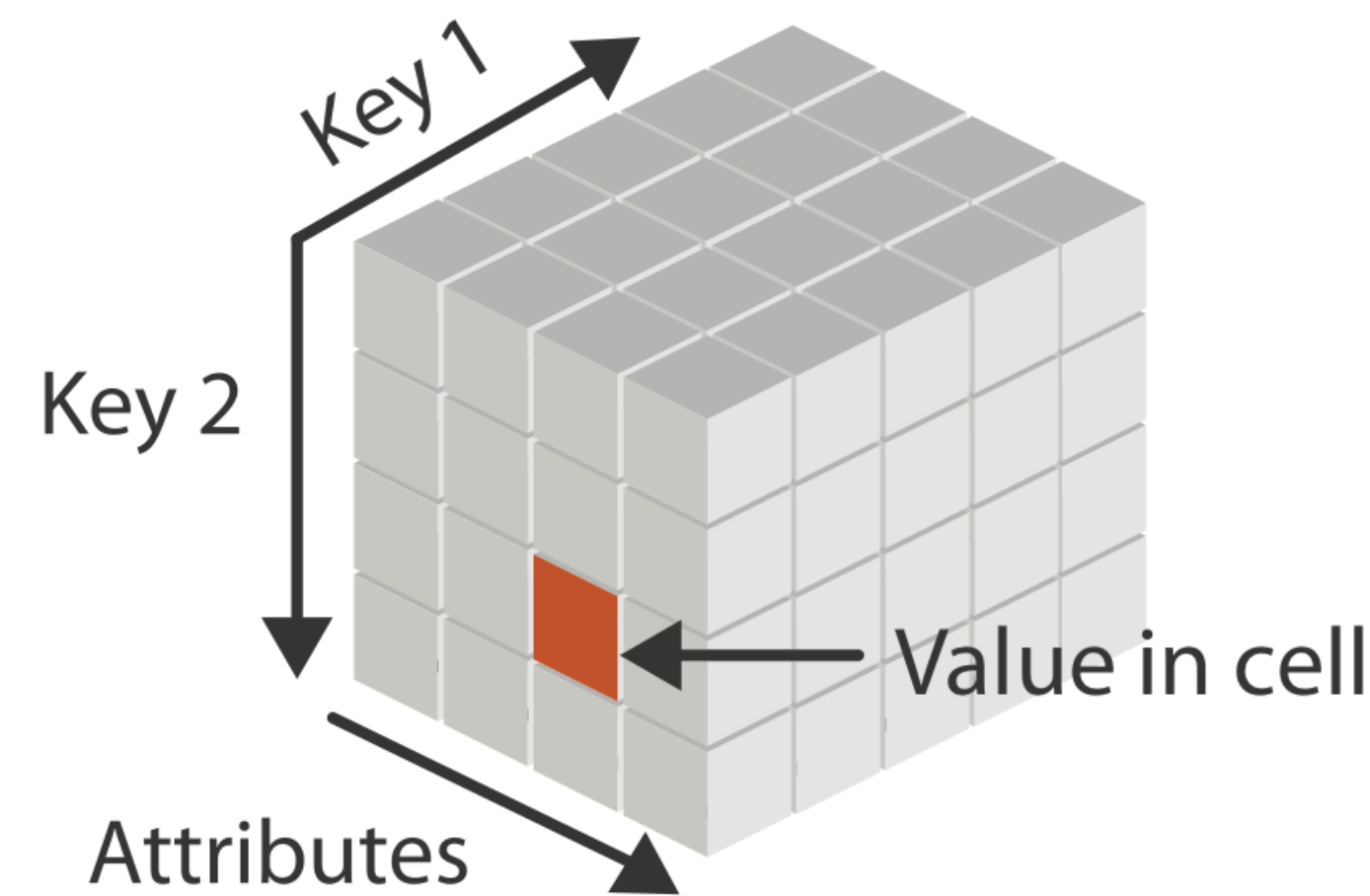
Data Model:
column contains numbers

Higher level of abstraction!

Dataset Types

1. Tabular

rows/records/items



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

columns/attributes/variables

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

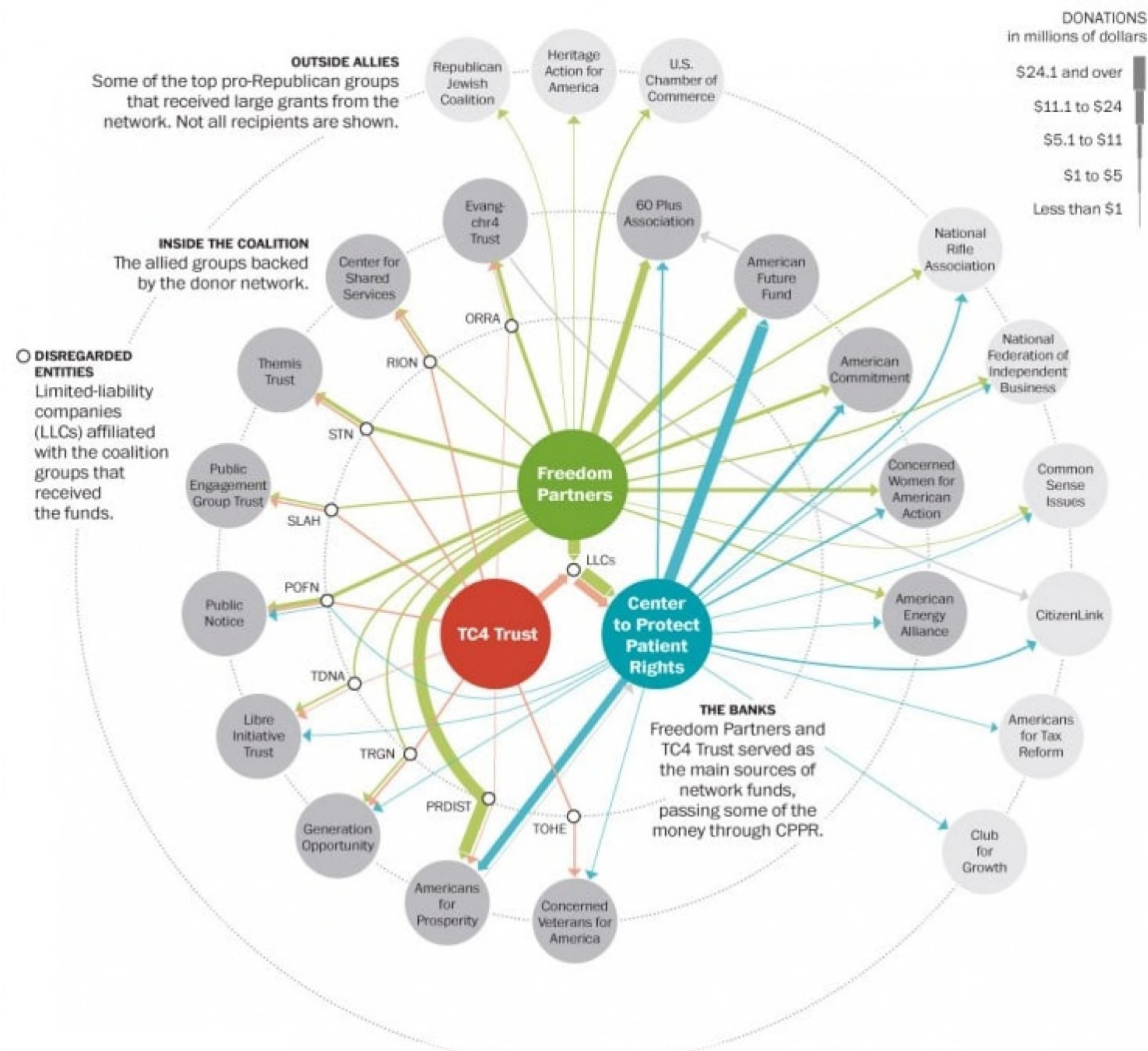
cell containing value

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.



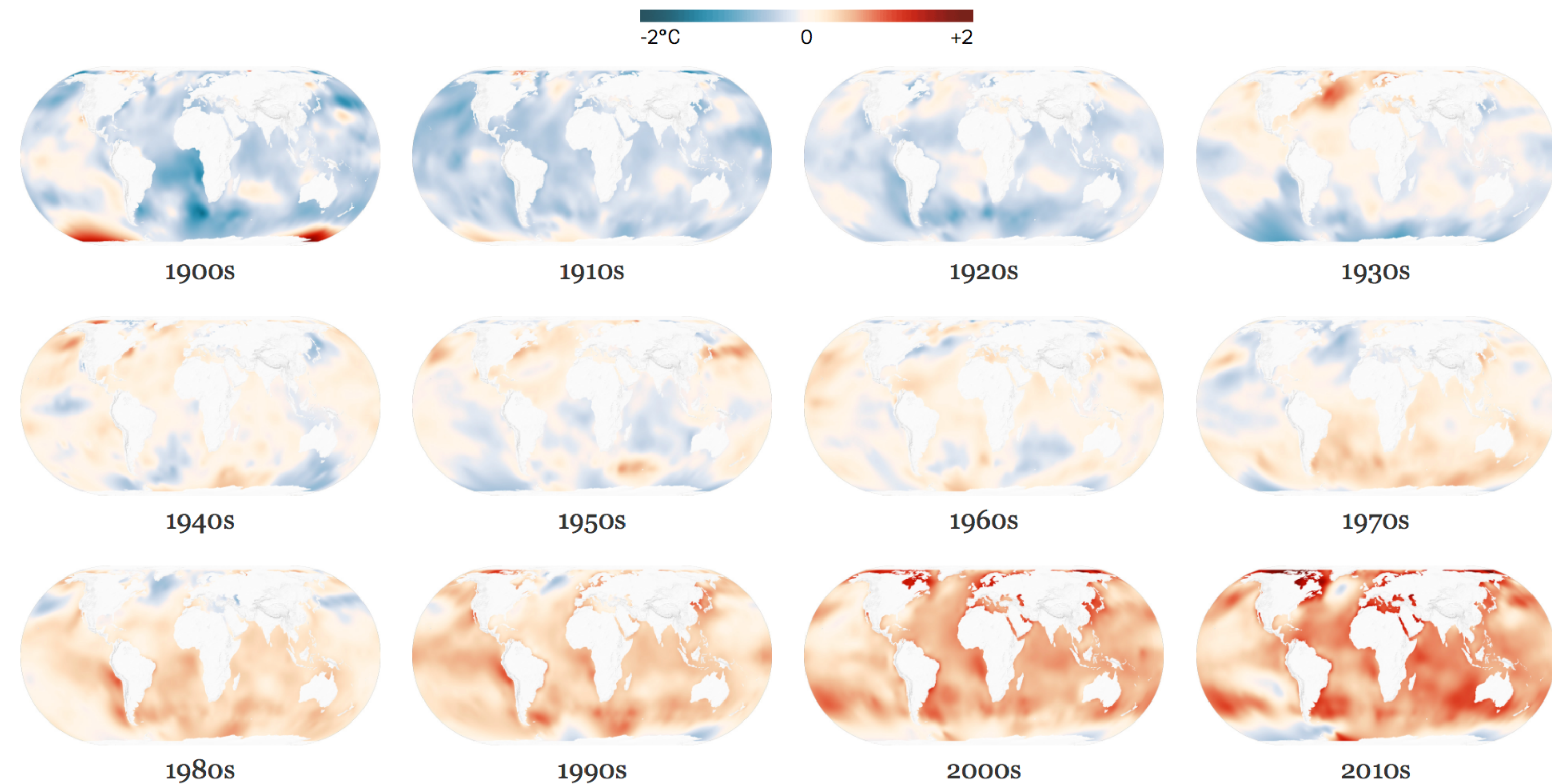
Dataset Types

1. Tabular:
collection of records
with named attributes

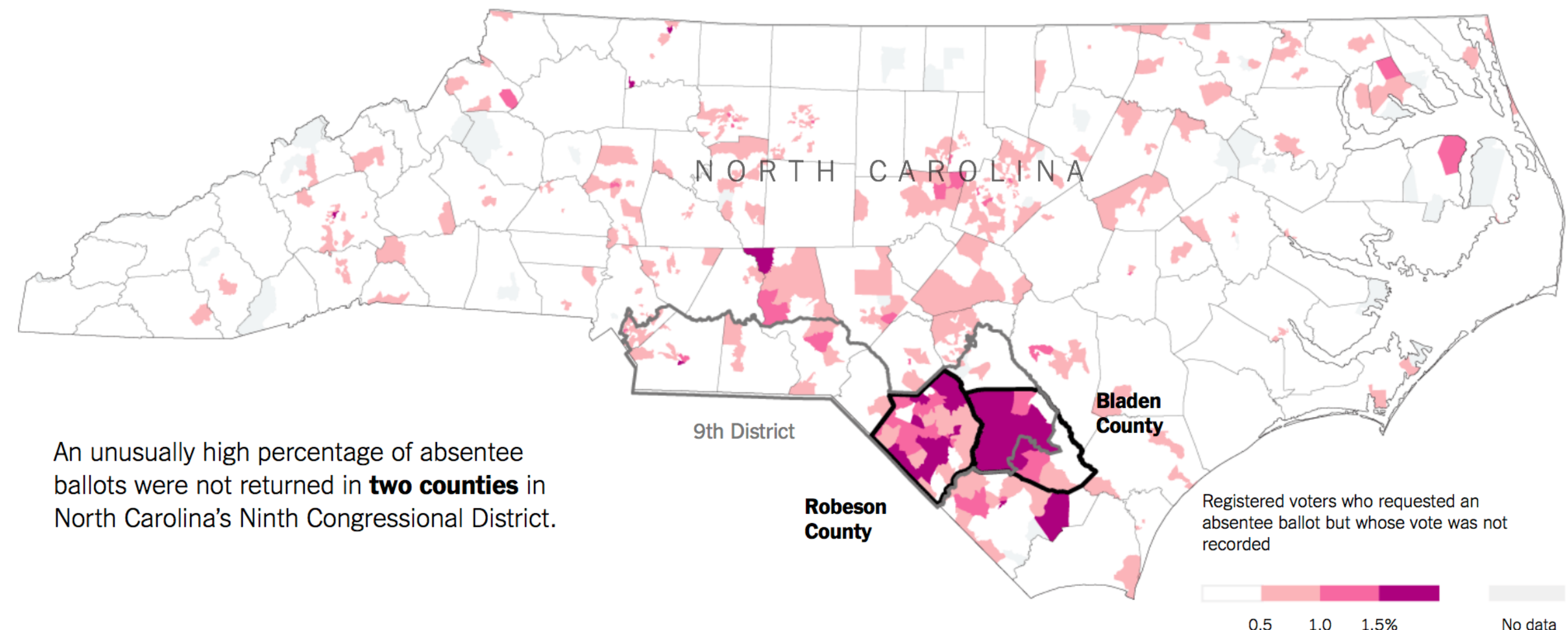
2. Network:
Nodes and links can also have
attributes (e.g., size of nodes,
thickness/directionality of links).

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>



An unusually high percentage of absentee ballots were not returned in **two counties** in North Carolina's Ninth Congressional District.

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

Attribute / Data Types

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-interval, or Q-ratio)

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	4211
21	1850	45	0	2	341254
22	1850	50	0	1	321242

Think on your own for 1 minute

Activity: U.S. Census

What are the types of these attributes?

(N, O, Q-interval, or Q-ratio)

People Count: # of people in group

Year: 1850 – 2000 (every decade)

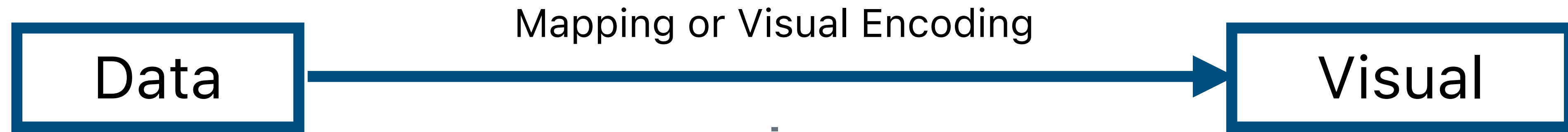
Age: 0 – 90+

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			11
19	1850	40			85
20	1850	45			11
21	1850	45	0	2	341254
22	1850	50	0	1	321343

tryclassbuzz.com:
census

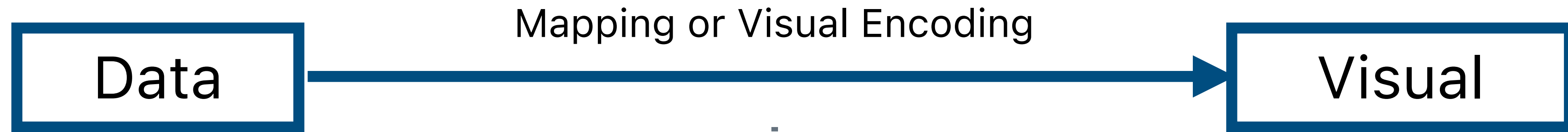


Expressiveness

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.*

A dashed vertical line connects the "Data" box to a box below. A blue arrow points from this box to the underlined text in the "Expressiveness" section.

Data models give us a way of talking about what the facts are.



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.*

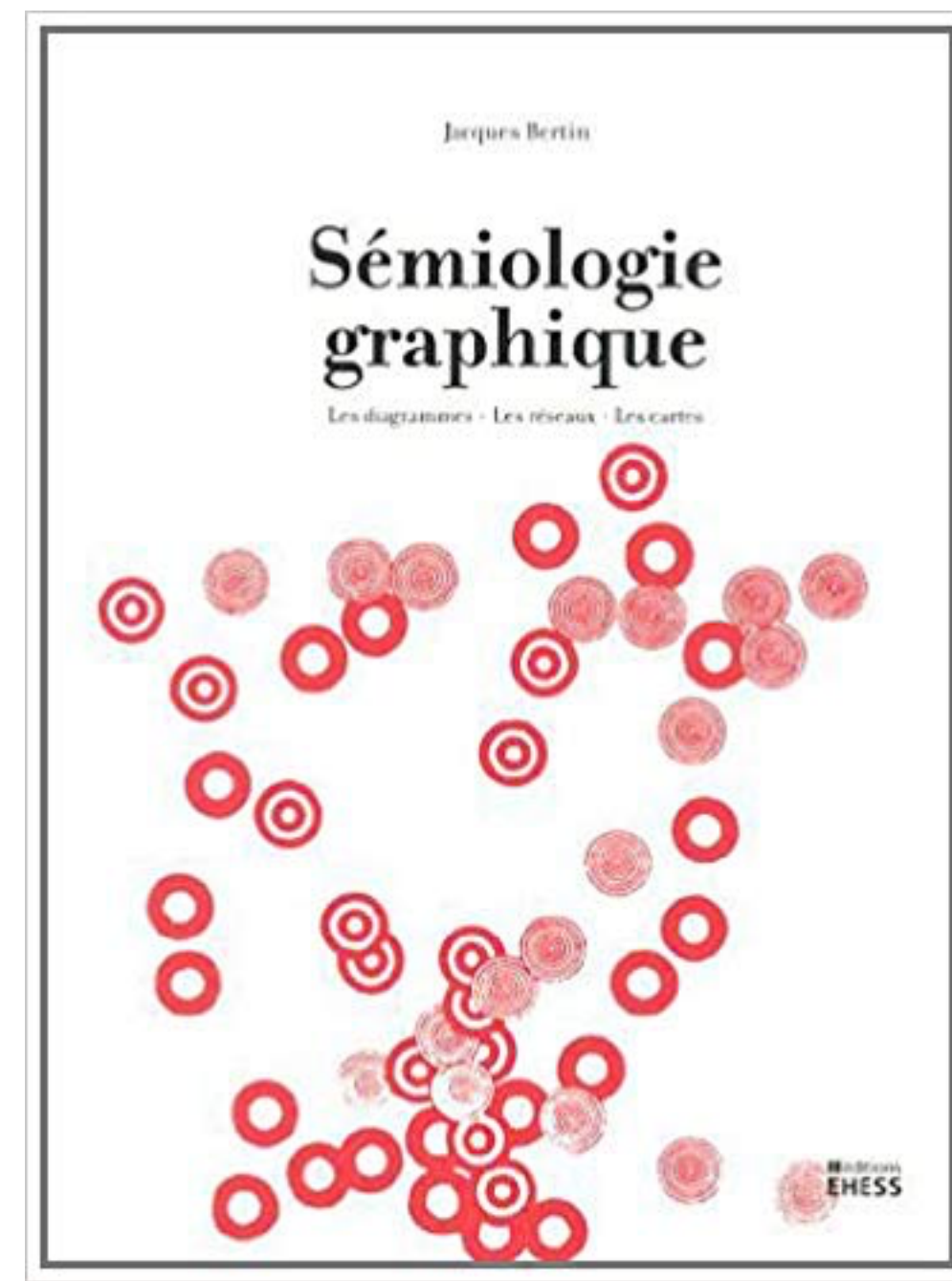
Effectiveness

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.

Image Models

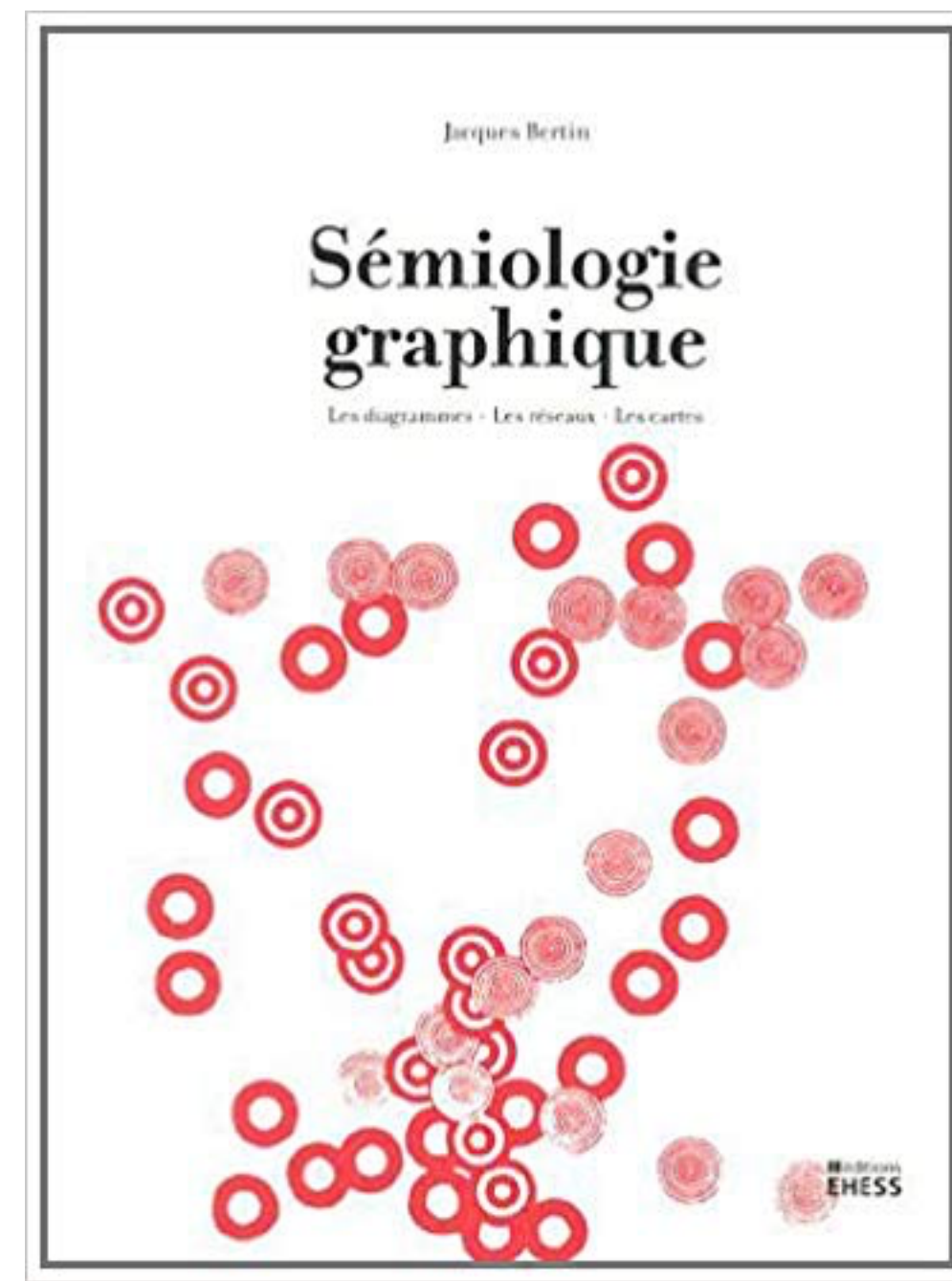
The Semiology of Graphics (1967)



Jacques Bertin (1918 – 2010)
French cartographer

The **Semiology** of Graphics (1967)

Study of signs and how cultures use them.

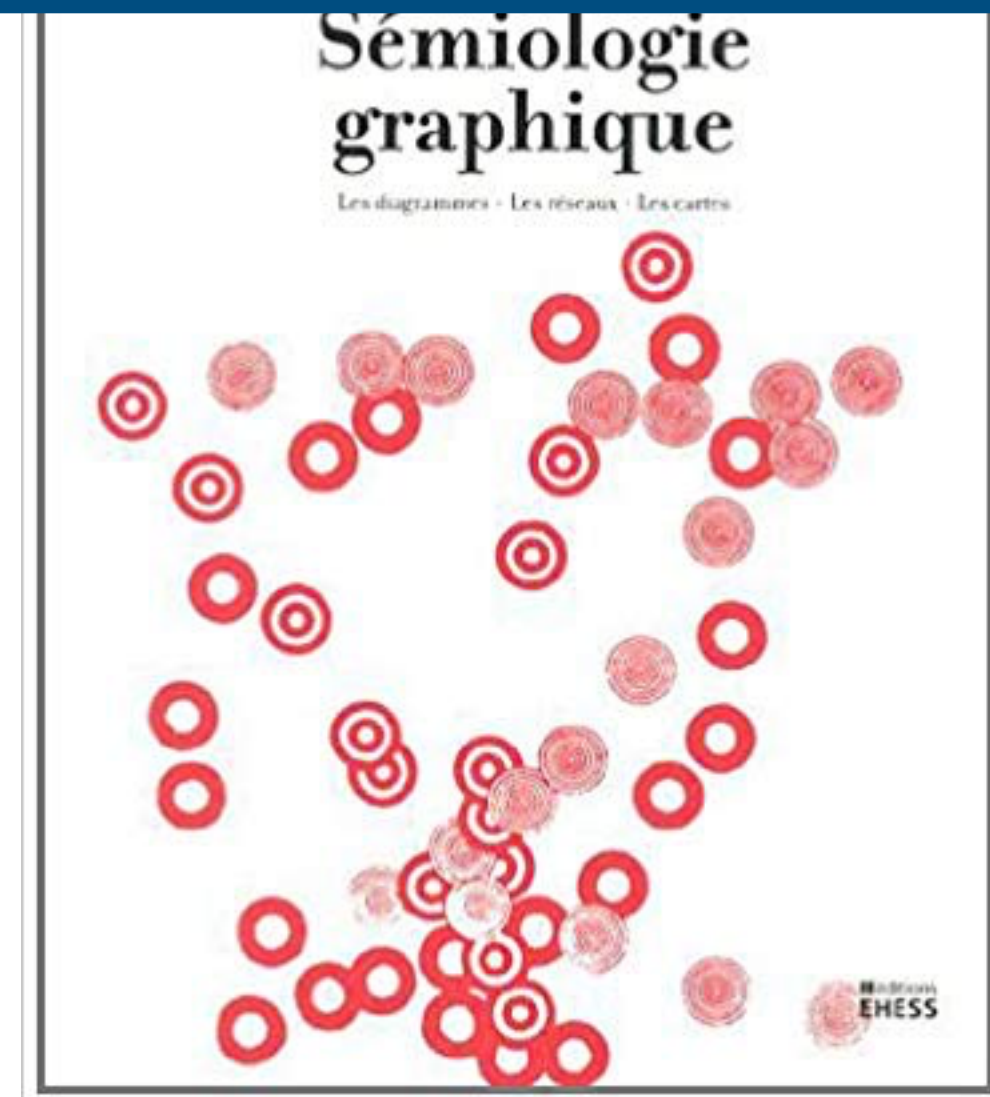


Jacques Bertin (1918 – 2010)
French cartographer

The **Semiology** of Graphics (1967)

Study of signs and how cultures use them.

Anything that stands for something other than itself.

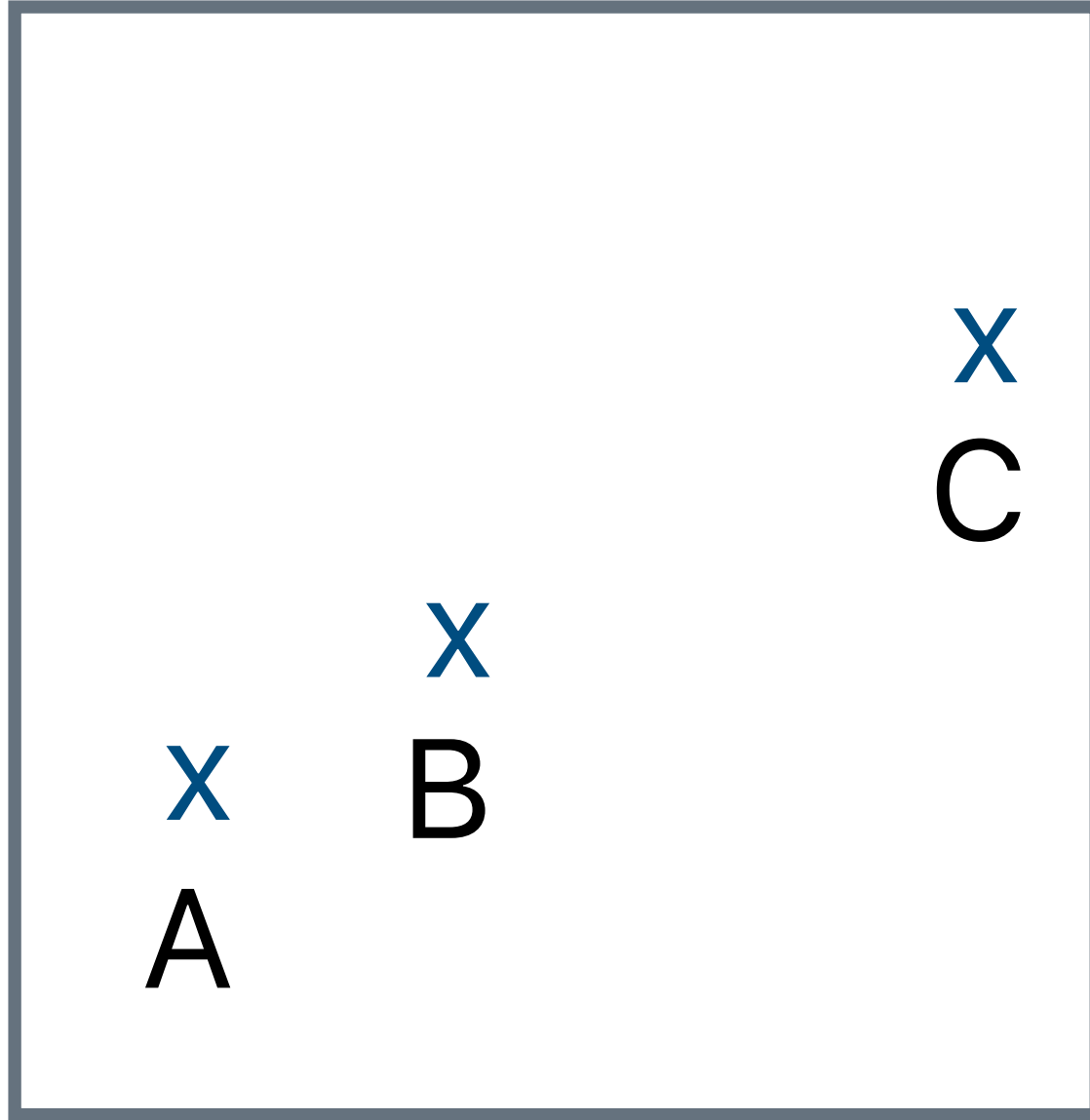


"Metal painted red"?

or

"Hit the brakes!"

Jacques Bertin (1918 – 2010)
French cartographer



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.

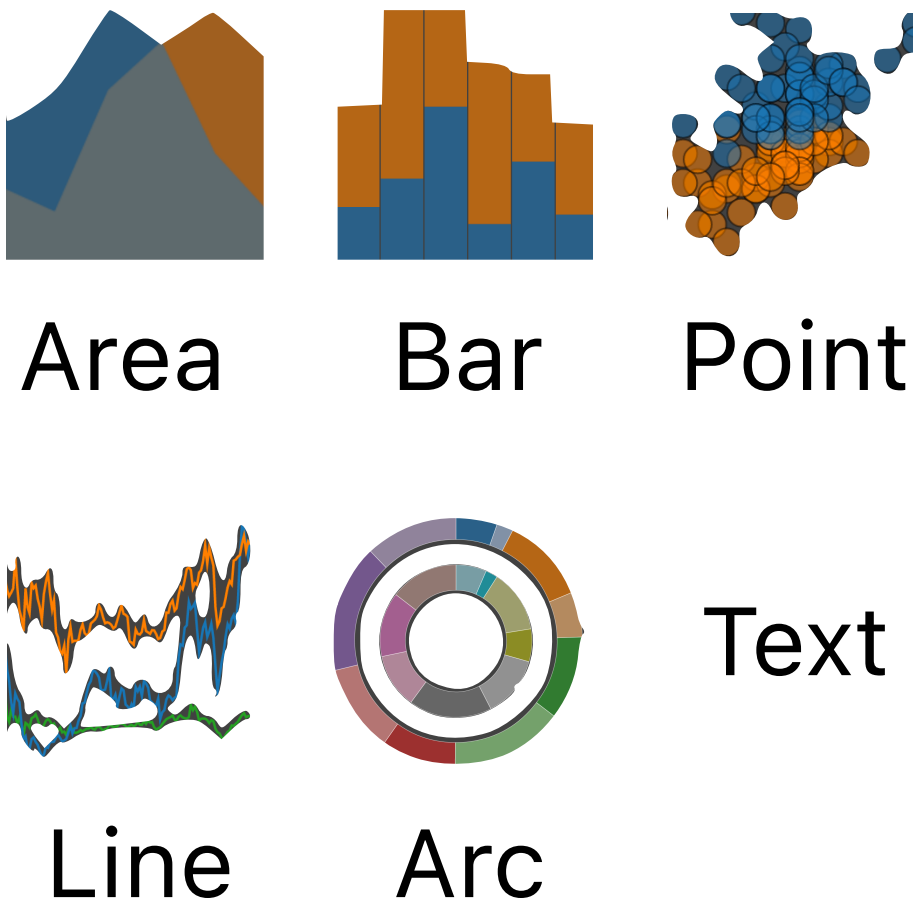
Used to encode data values as characteristics of marks.

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

LES VARIABLES DE L'IMAGE											
			POINTS			LIGNES			ZONES		
XY	2 DIMENSIONS DU PLAN										
Z	TAILLE										
	VALEUR										
LES VARIABLES DE SÉPARATION DES IMAGES											
	GRAIN										
	COULEUR										
	ORIENTATION										
	FORME										

Marks

Basic graphical elements that represent data items.



Channels: Expressiveness Types and Effectiveness Ranks

➔ **Magnitude** Channels: **Ordered** Attributes

Position on common scale



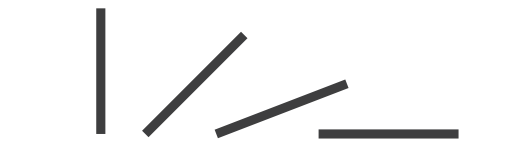
Position on unaligned scale



Length (1D size)



Tilt/angle



Area (2D size)



Depth (3D position)



Color luminance



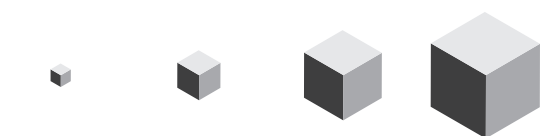
Color saturation



Curvature



Volume (3D size)



Same

Same

Most
Effectiveness
Least

➔ **Identity** Channels: **Categorical** Attributes

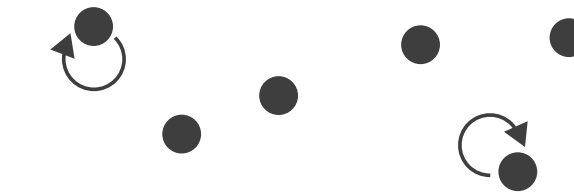
Spatial region



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



➔ **Identity** Channels: **Categorical** Attributes

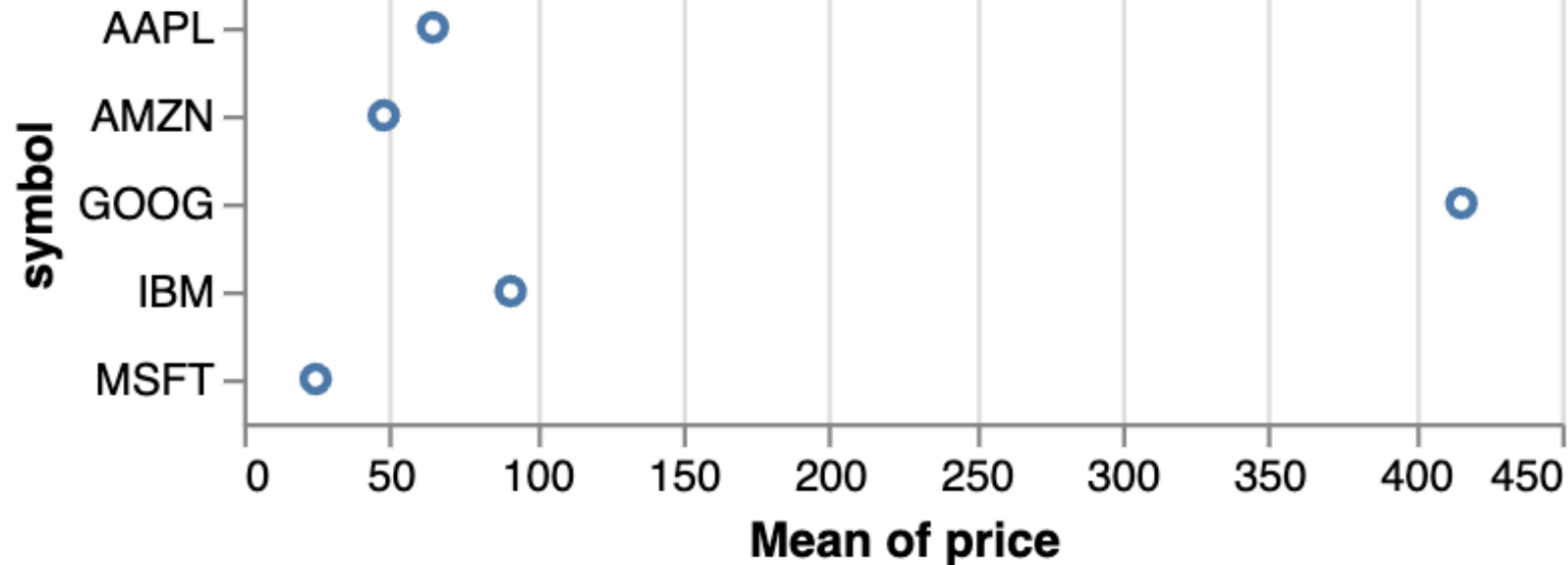
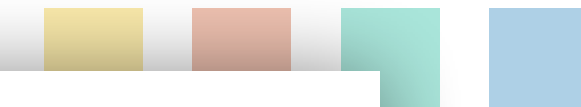
Spatial region



Position on unaligned scale



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)



➔ **Identity** Channels: **Categorical** Attributes

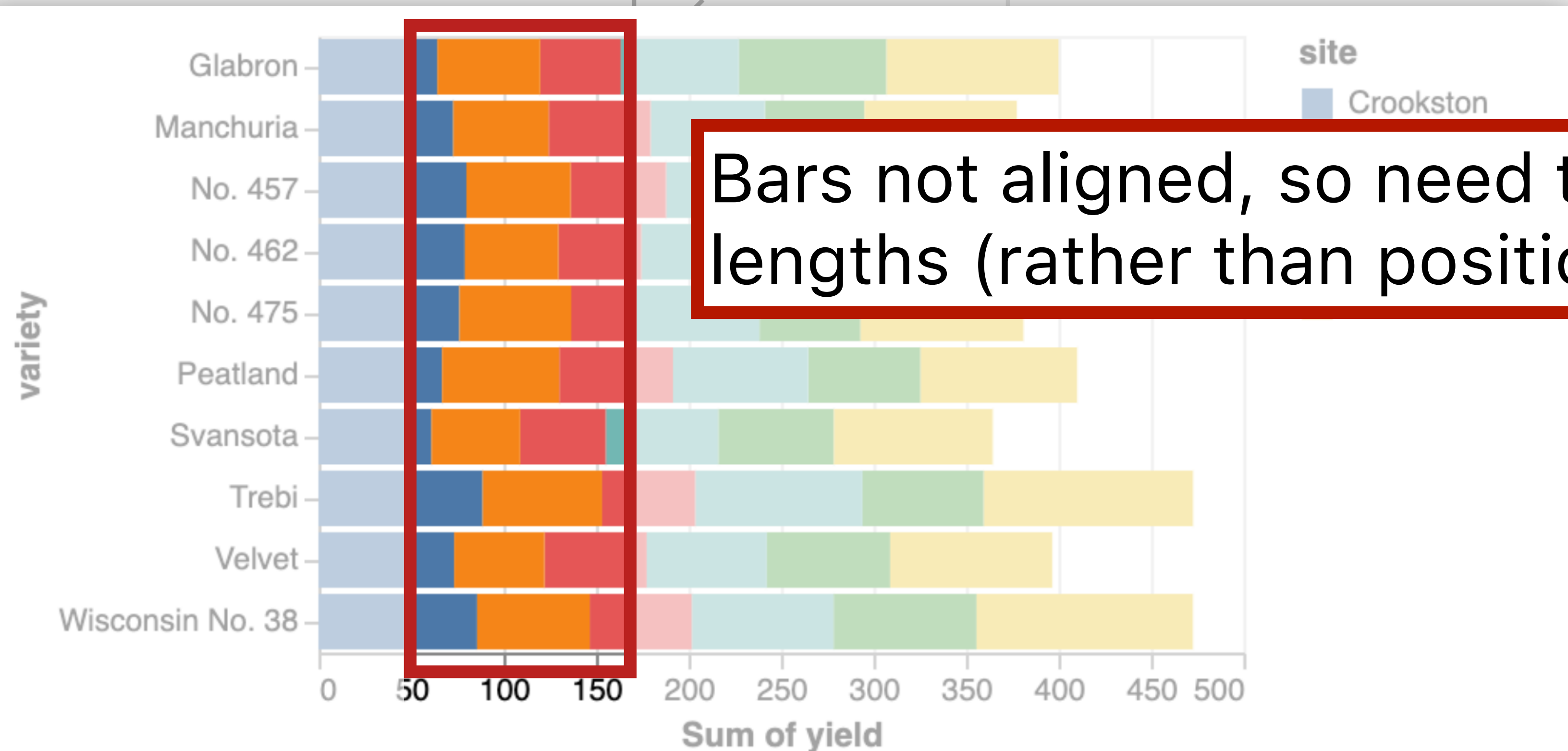
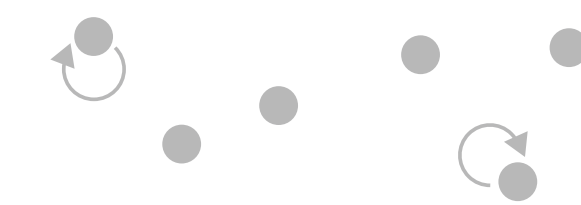
Spatial region



Color hue



Motion



Visualization
sign (2014).

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



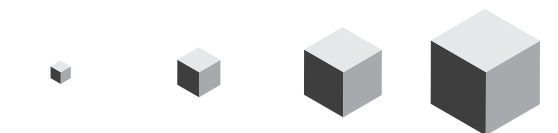
Color saturation



Curvature



Volume (3D size)



➔ **Identity Channels: Categorical** Attributes

Most effective to least effective

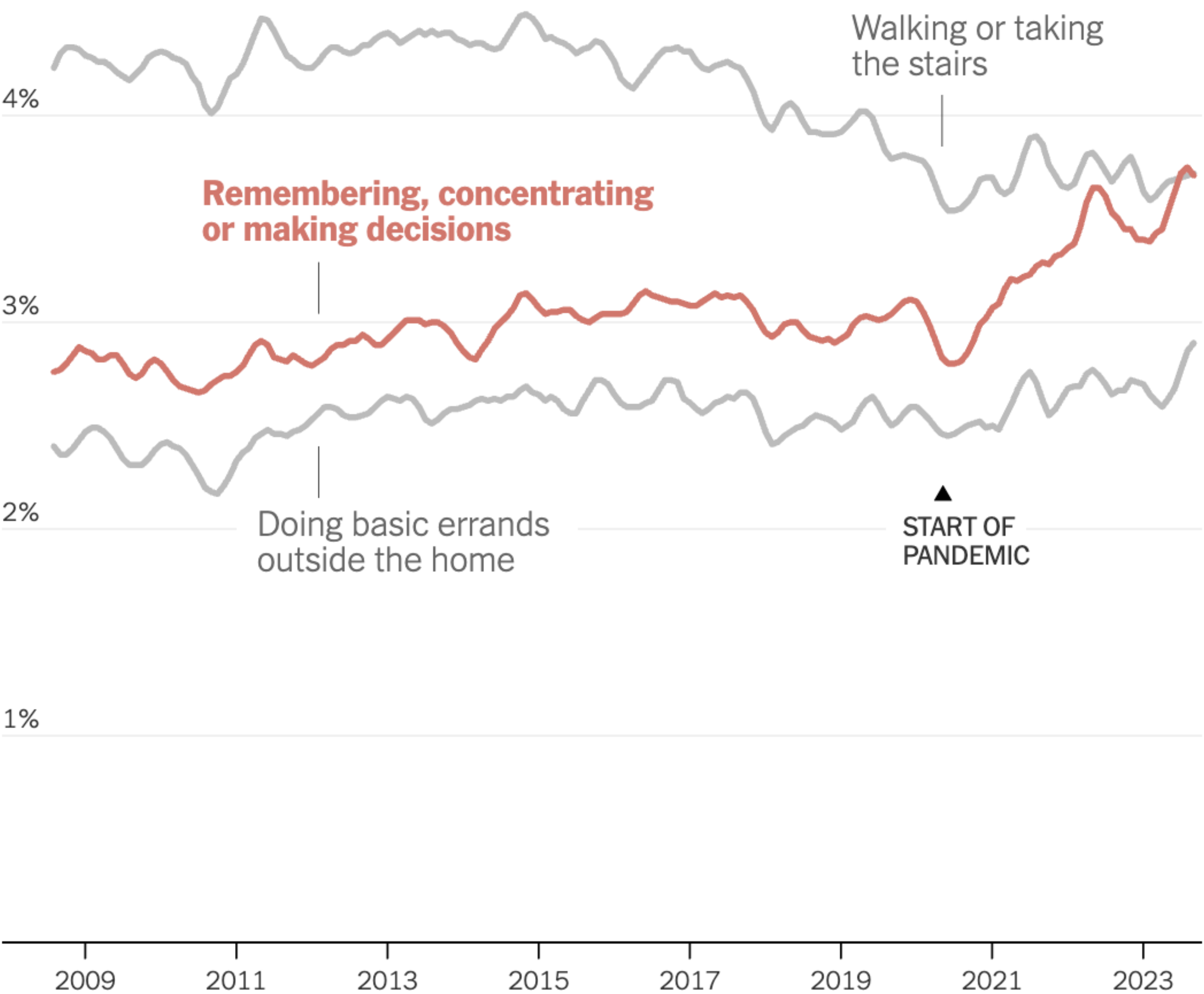
Top of scale = easiest for people to make accurate comparisons

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

Name that ~~chart!~~

Visual Encoding!

Percent of working-age people who said they had “serious difficulty” with ...



Mark: line

X-axis: date (Q-interval)

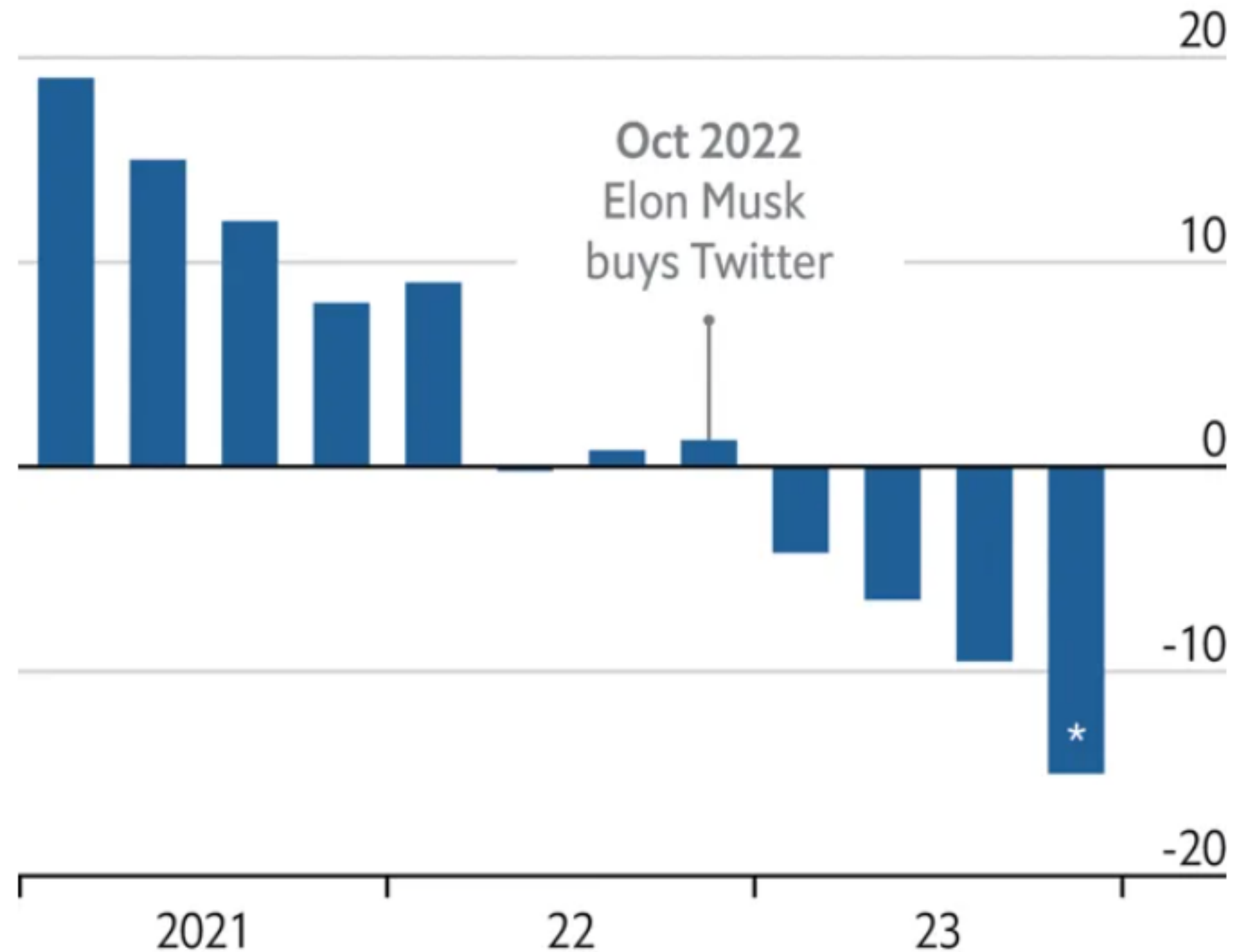
Y-axis: percent (Q-ratio)

What about color?

Drop off

Estimated monthly active Twitter/X users

% change on a year earlier



Mark: bar

X-axis: date (Q-interval)

Y-axis: percent (Q-ratio)

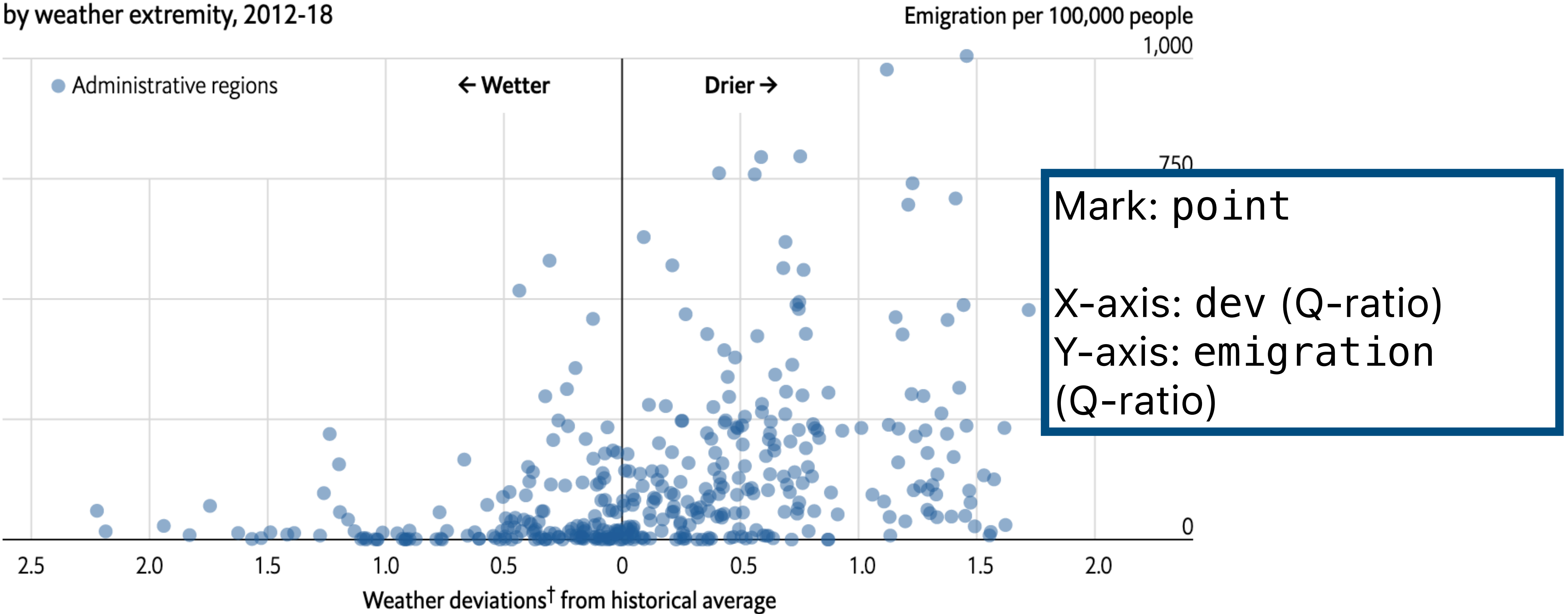
*To December 5th

Source: Sensor Tower

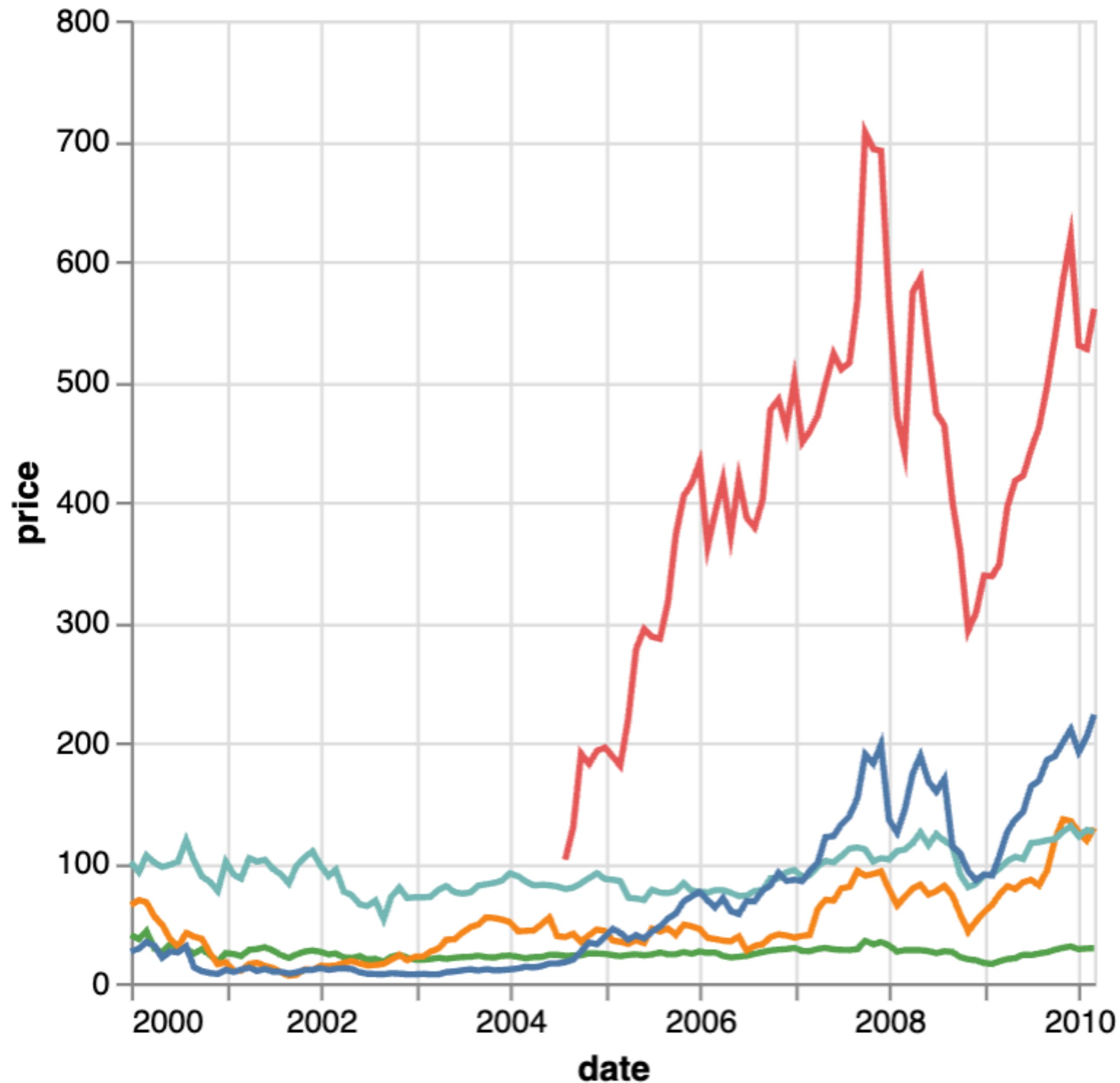


Spotting a trend

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



symbol

AAPL
AMZN
GOOG
IBM
MSFT

Mark: line

X-axis: date (Q-interval)

Y-axis: price (Q-ratio)

Color: symbol (N)

Notice the parallel with
plotly express syntax!

```
px.line(  
    stocks_df,  
    x='date',  
    y='price',  
    color='symbol',  
)
```

Actual win percentage

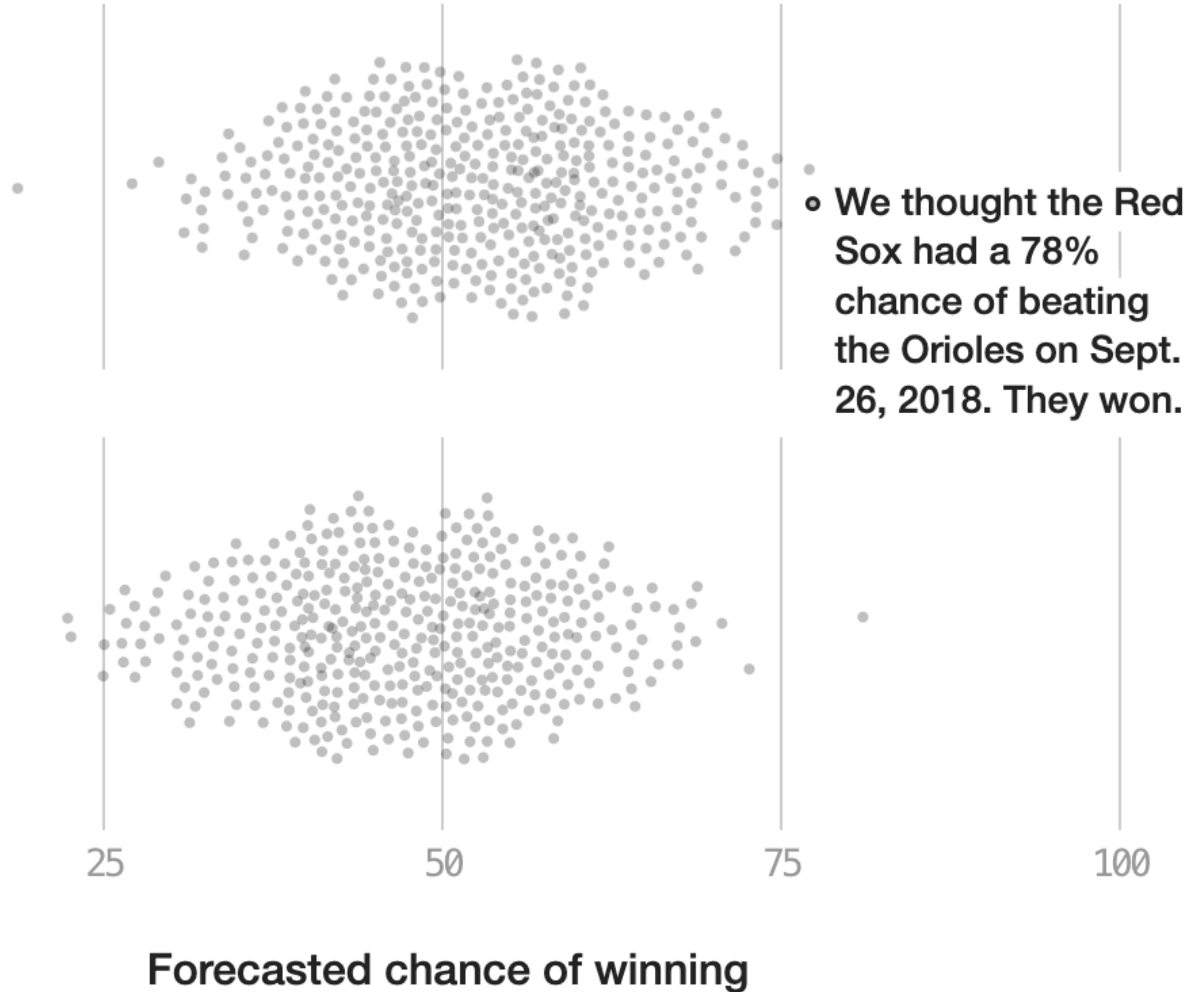
Team won
100%

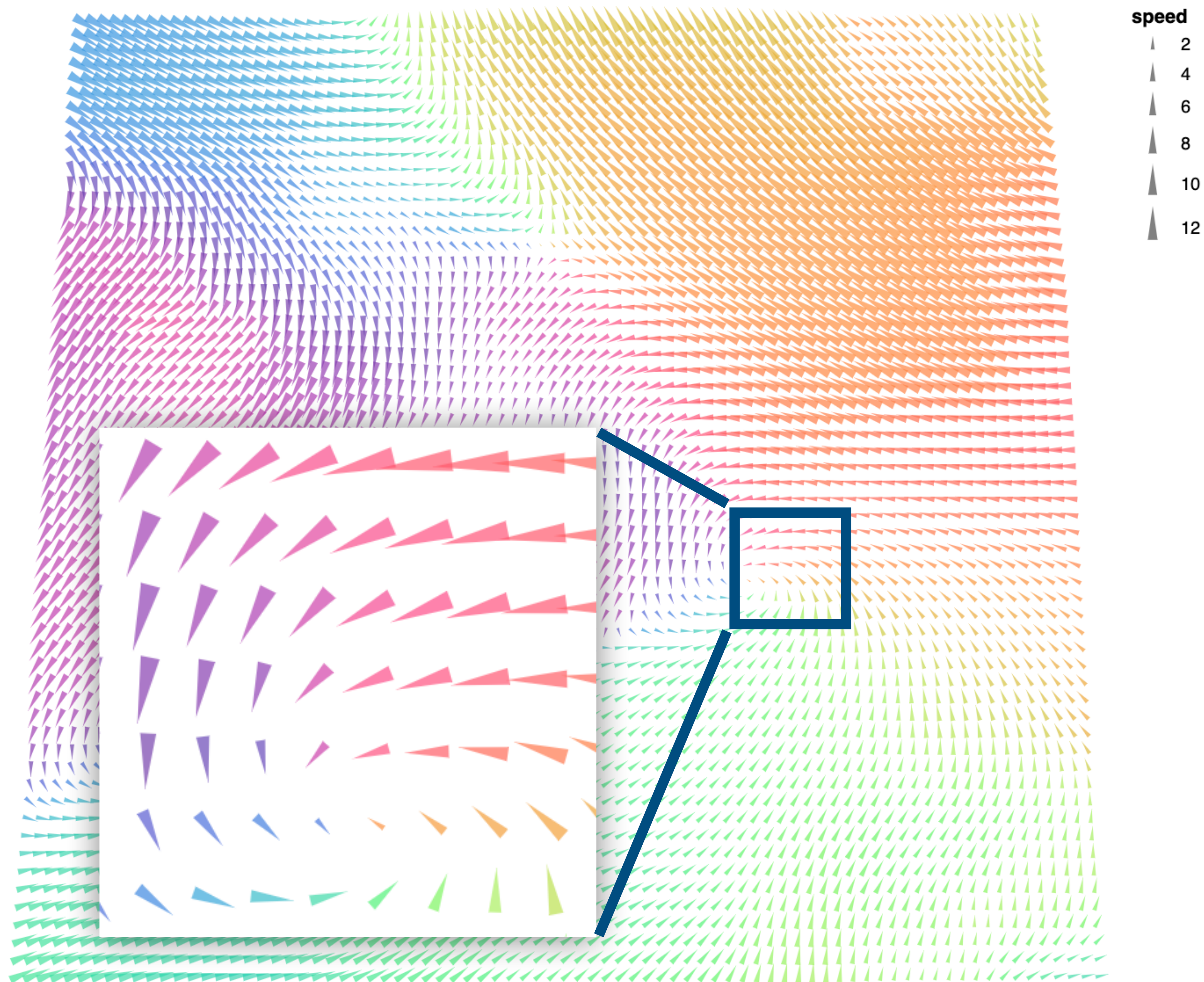
Team lost
0%

Mark: point

X-axis: chance (Q-ratio)

Y-axis: ?? (nothing!)





tryclassbuzz.com:
wind

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite. Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Chiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

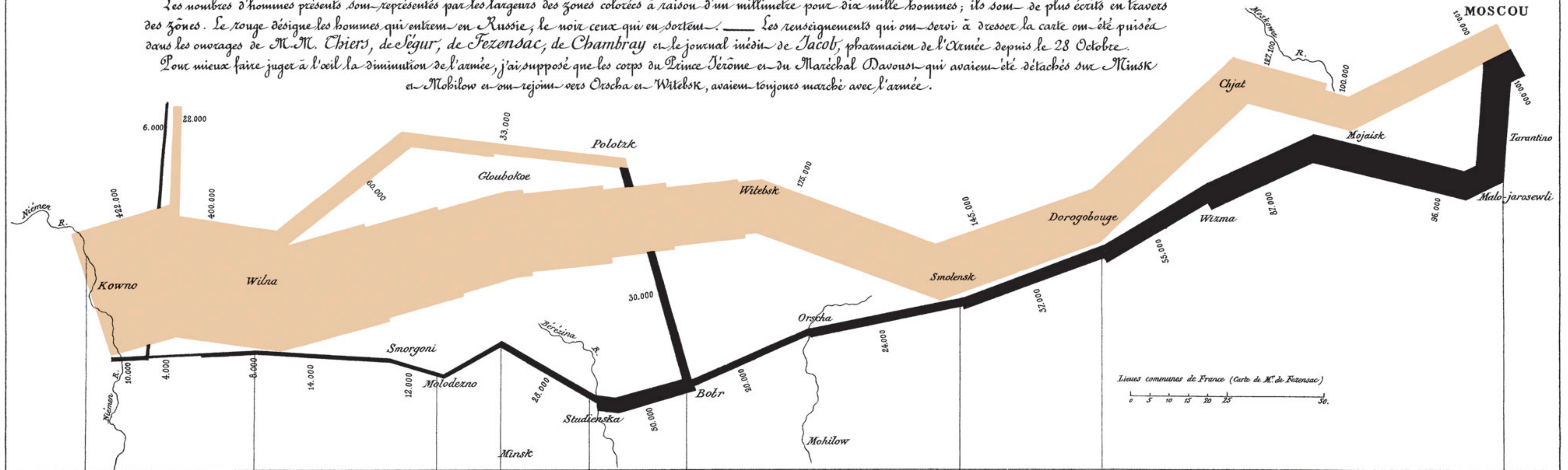
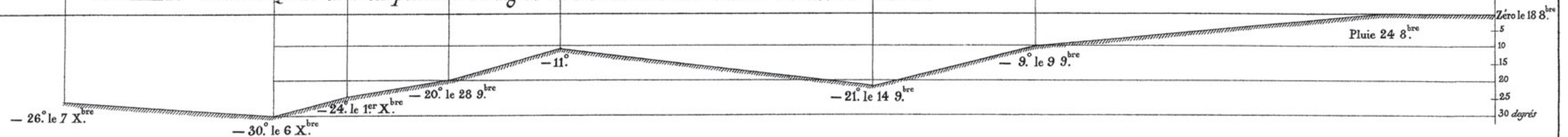


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Les Cosaques passent au galop le Niemen gelé.



Autog. par Regnier, 8. Par. 5^{te} Marie 5^{te} G^{de} à Paris.

Imp. Lith. Regnier et Dourdet.

tryclassbuzz.com:
minard

Next time: Visual Encoding & Design