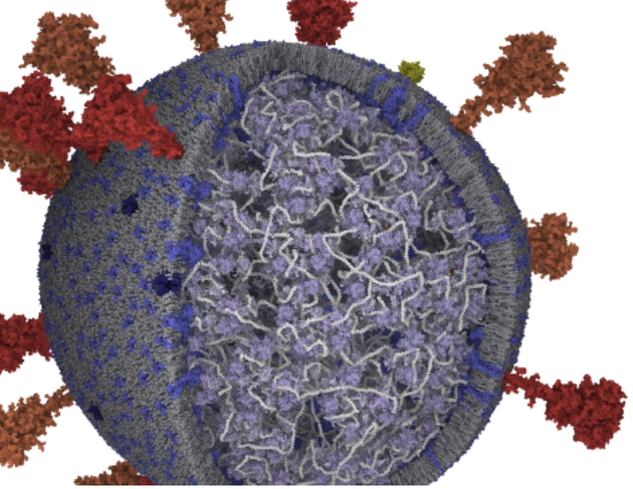
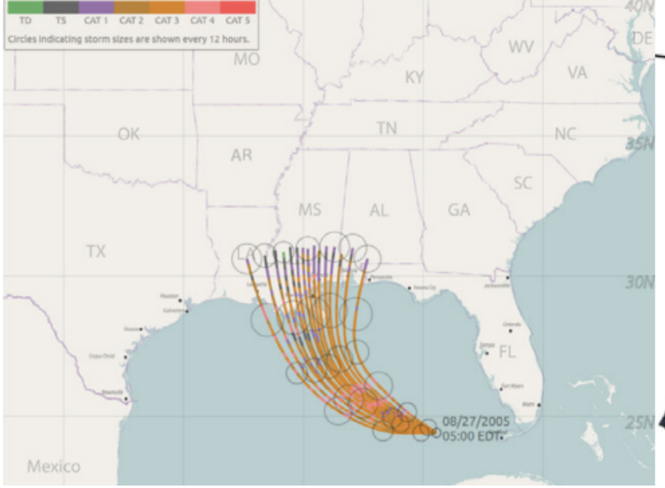
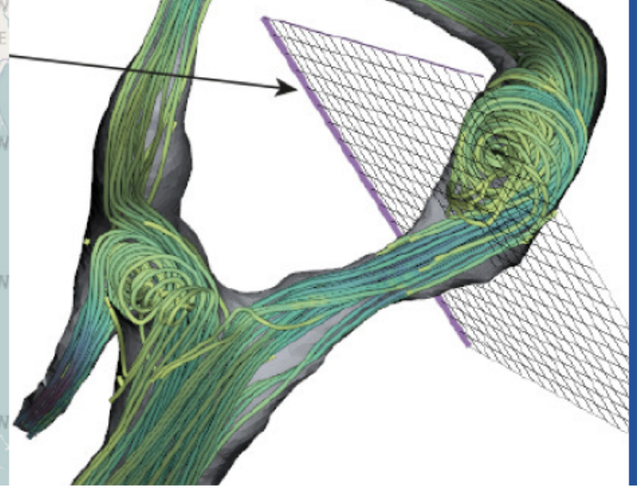
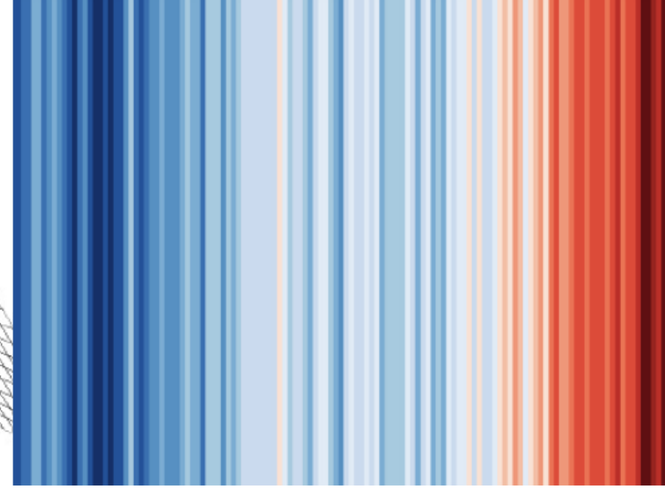



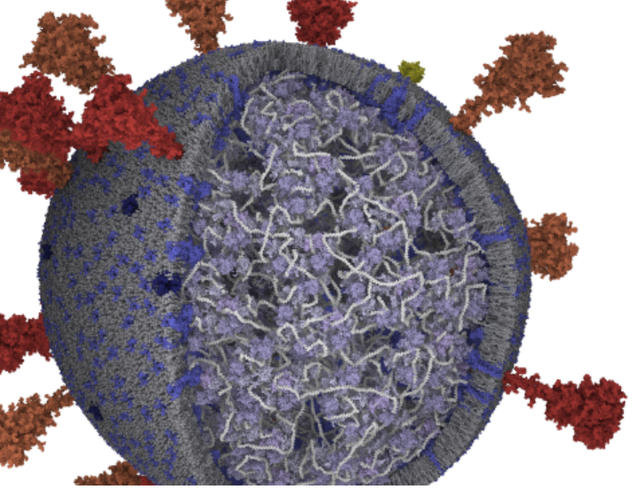
Data Visualization

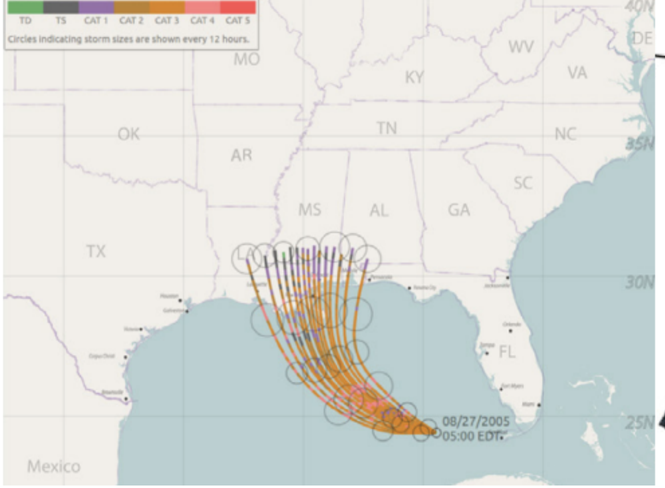
INF100 Vår 2025 | 08. april

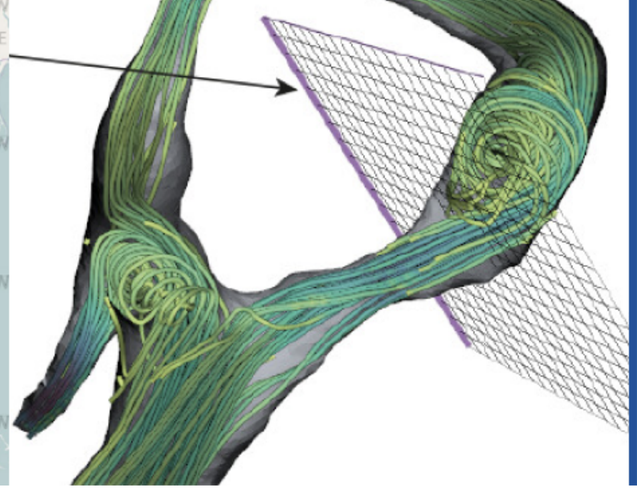
Amy Zhang
Institutt for Informatikk, Universitetet i Bergen
ke.zhang@uib.no

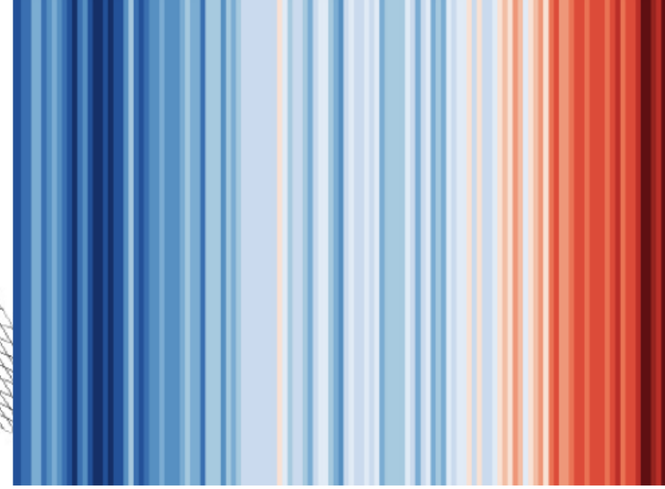
Today

- Why visualization?
- Core programming concepts reflected in visualization
  - Transformation
  - Abstraction: Data and Task
  - Encodings
- Demo: Drawing charts with Python libraries





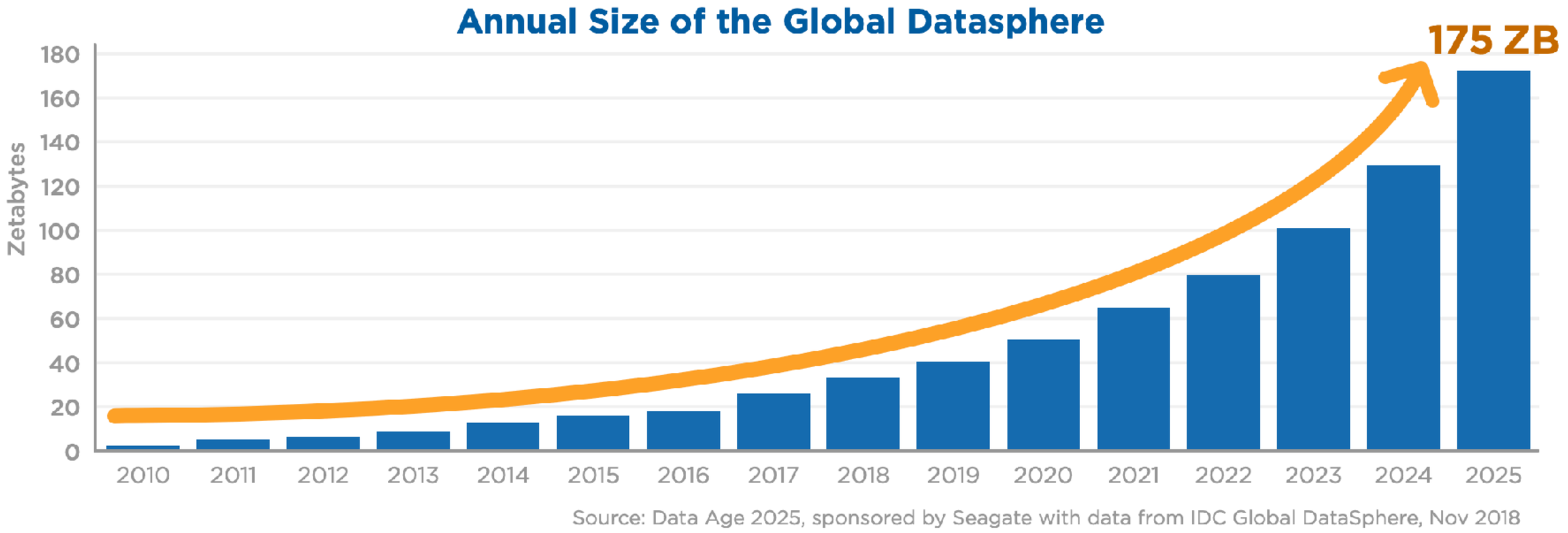




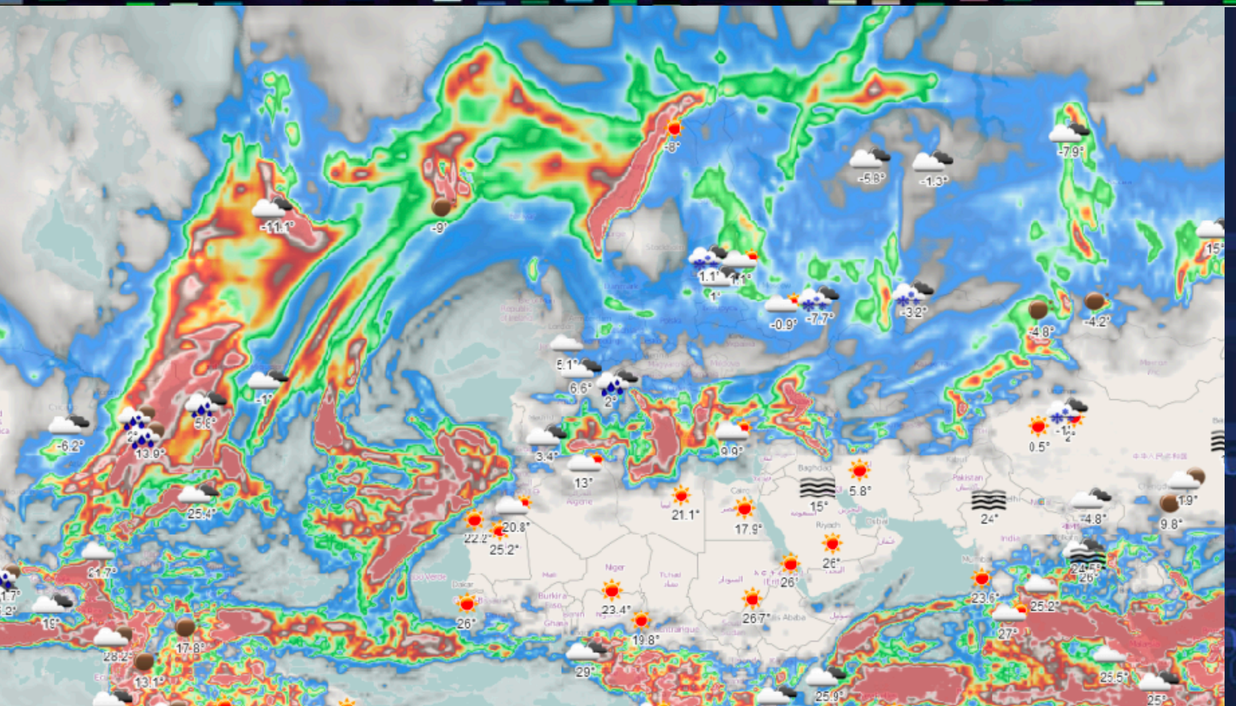
Data visualization:

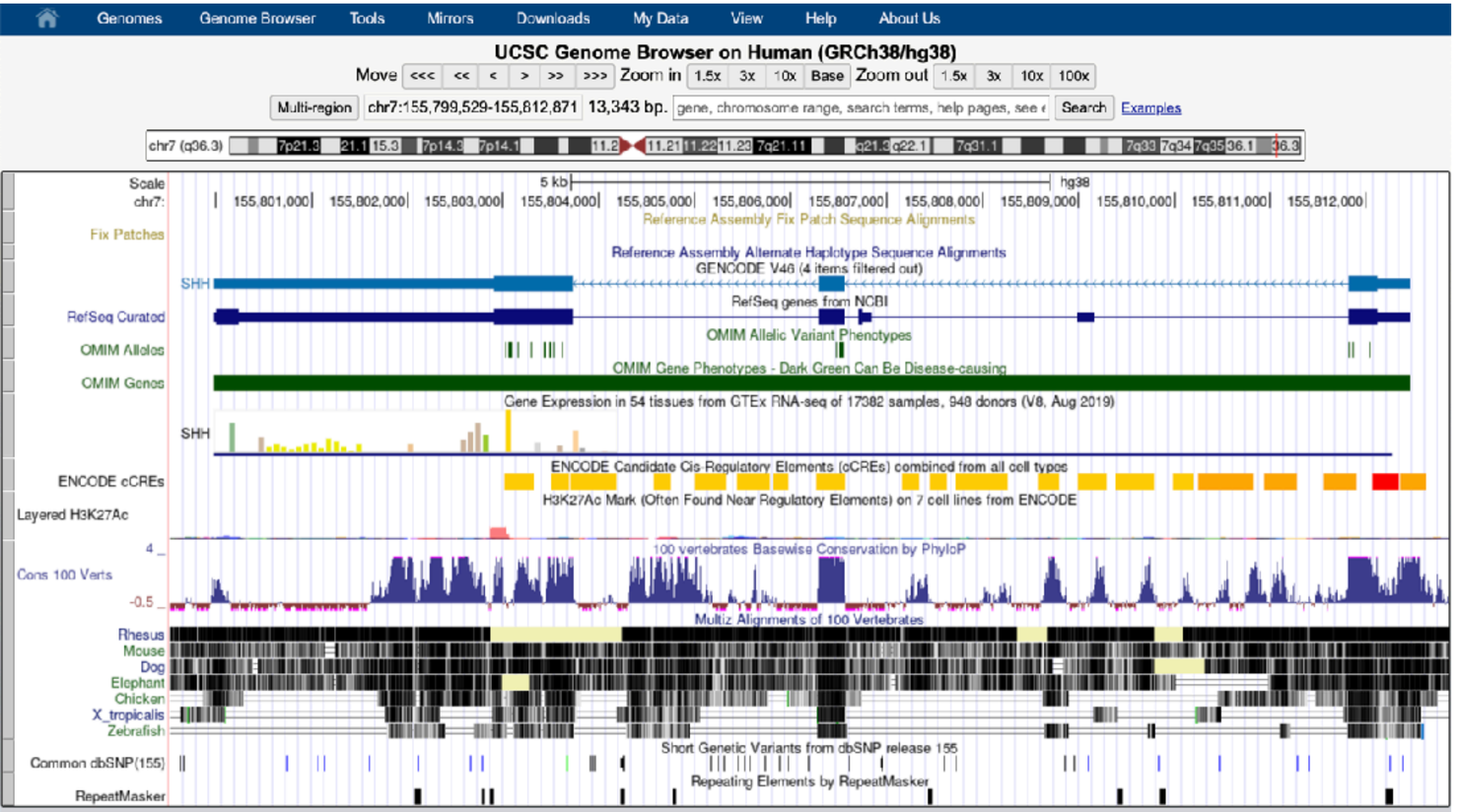
What is it and why do we do it?

More and more data…



https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf





Valerie Altounian, Filling the Gaps:
https://www.altounianillustration.com/portfolio/filling-the-gaps

What is visualization?

Ways to *define* visualization

"…**transformation** of the **symbolic** to the **geometric**."

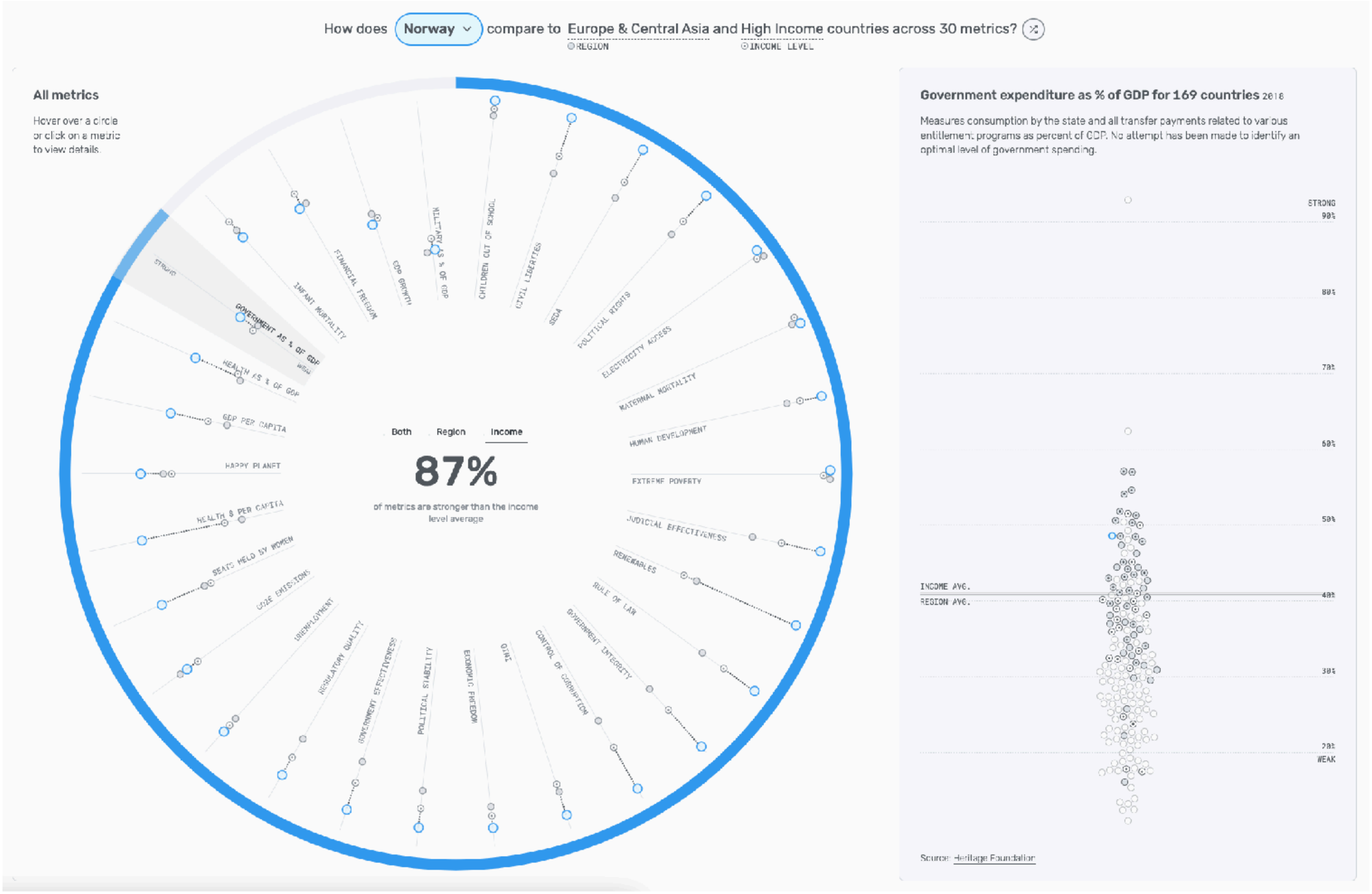
—*McCormick et al. 1987*

"A visualization is a **picture** that **helps someone** to **do something**."

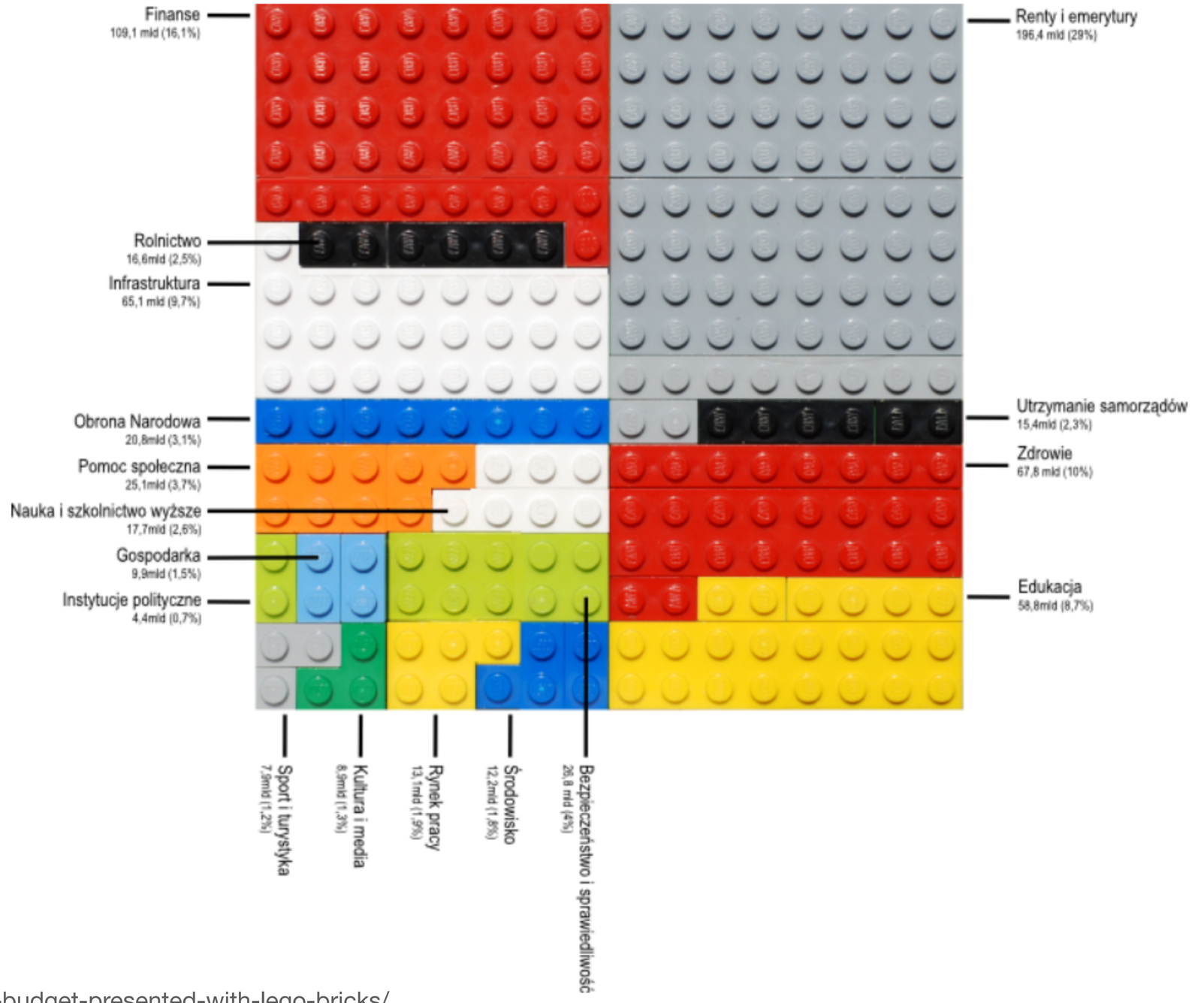
—*Michael Gleicher 2019*

"Computer-based visualization systems provide **visual representations** of **datasets** designed to help **people** carry out **tasks** more effectively."

—*Tamara Munzner 2014*



[How do we compare? by Lindsey Poulter]



http://dataphys.org/list/poland-budget-presented-with-lego-bricks/

Why do visualization?

How many times does V occur in this string?

MTHIVLWYADCEQGHKILKMTWY
NARDCAIREQGHLVKMFPSTWYA
RNGFPSVCEILQGKMFPSNDRC
EQDIFPSGHLMFHKMVPSTWYA
CEQTWRN

How many times does V occur in this string?

MTHIVLWYADCEQGHKILKMTWY
NARDCAIREQGHLVKMFPSTWYA
RNGFPSVCEILQGKMFPSNDRC
EQDIFPSGHLMFHKMVPSTWYA
CEQTWRN

How many items have a Y value between **6.00 and 8.00***?*

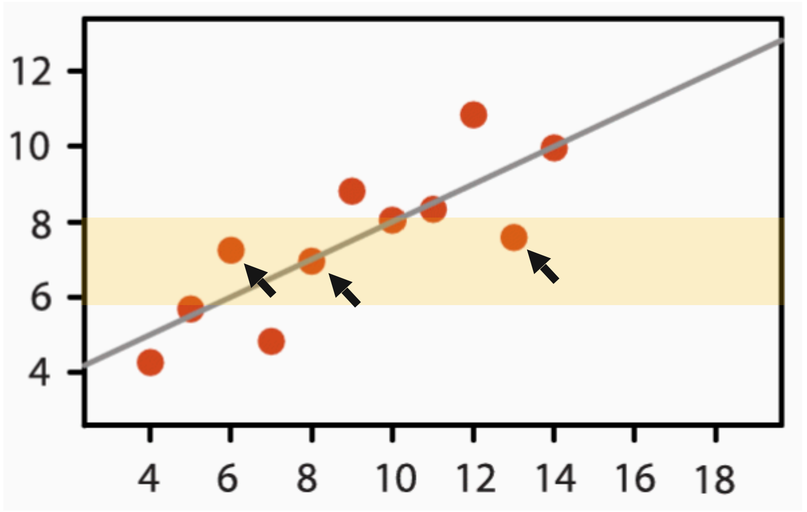
X	Y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

Munzner, T. (2014). Visualization analysis and design. AK Peters Visualization Series, CRC Press, Visualization Series. Chp 1
Values from *from set 1 of* *Anscombe's Quartet*

How many items have a Y value between **6.00 and 8.00***?*

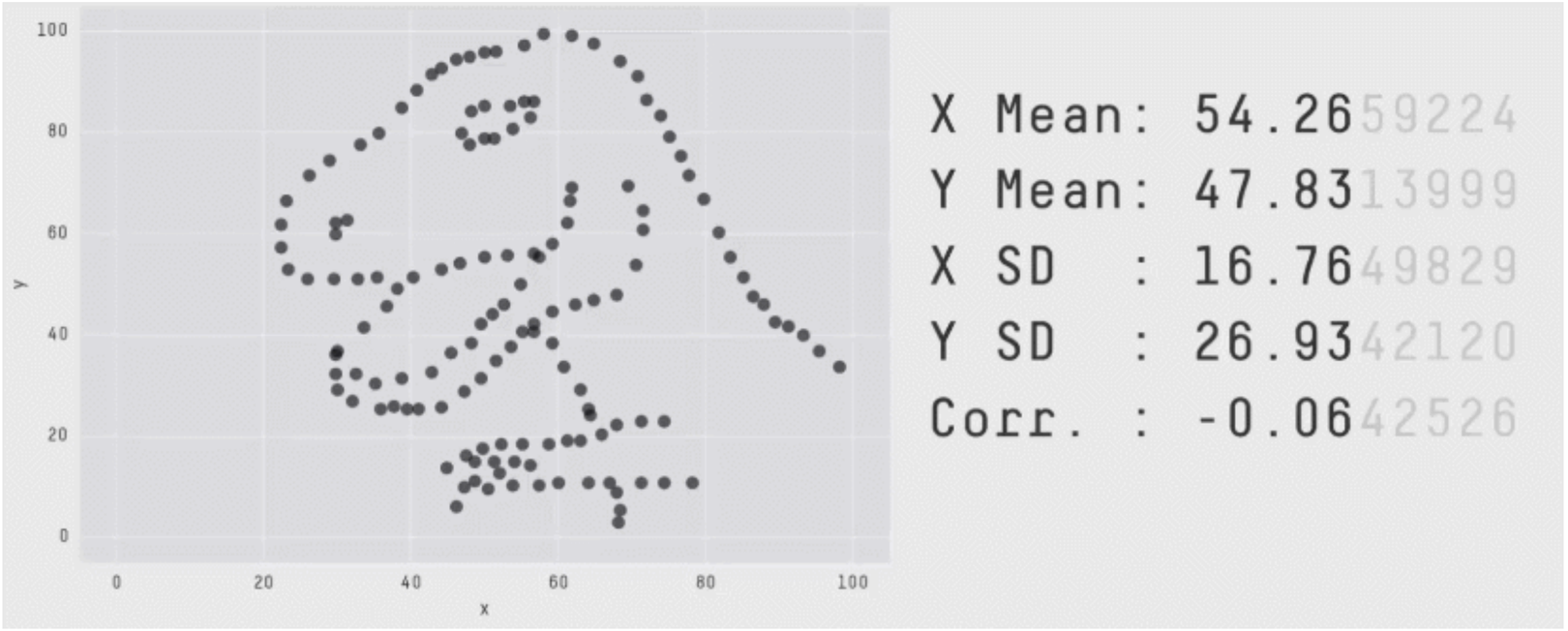
X	Y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

Replace **cognition** with **perception**



Munzner, T. (2014). Visualization analysis and design. AK Peters Visualization Series, CRC Press, Visualization Series. Chp 1
Values from *from set 1 of* [*Anscombe's Quartet*]

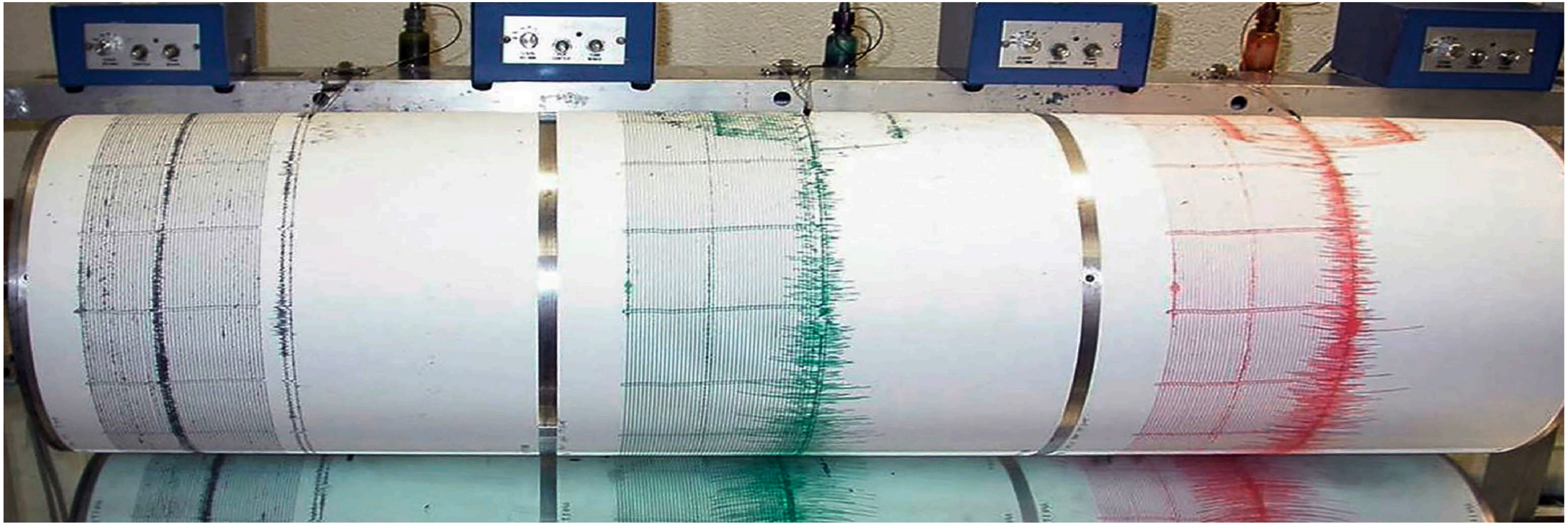
Summaries can lose information. **Details matter!**



Matejka, J., & Fitzmaurice, G. (2017). Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. In Proceedings of the 2017 CHI conference on human factors in computing systems (pp. 1290-1294).

To record information

(Show the data)



Seismograph, [Indiana University]

To record information

(Show the data)

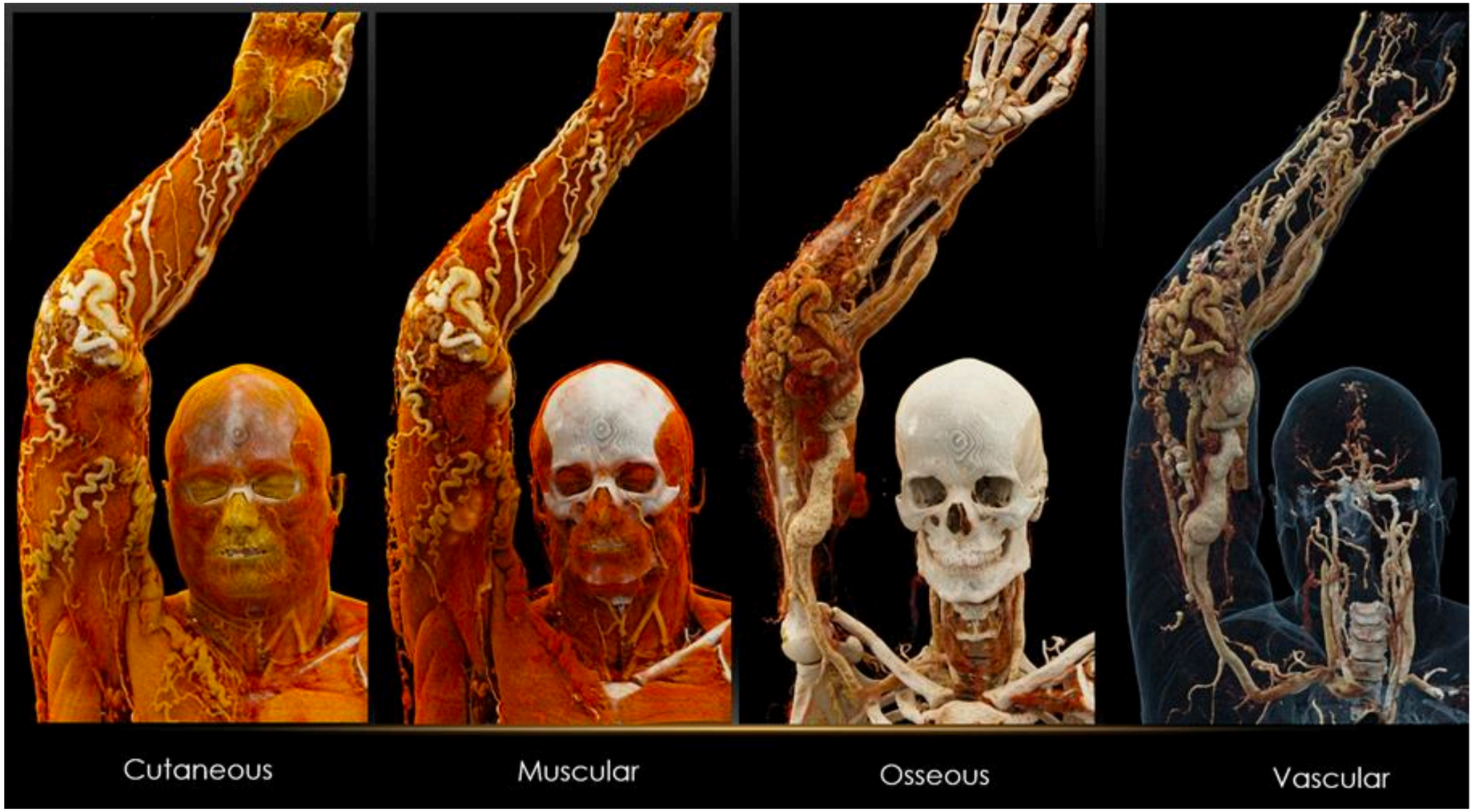
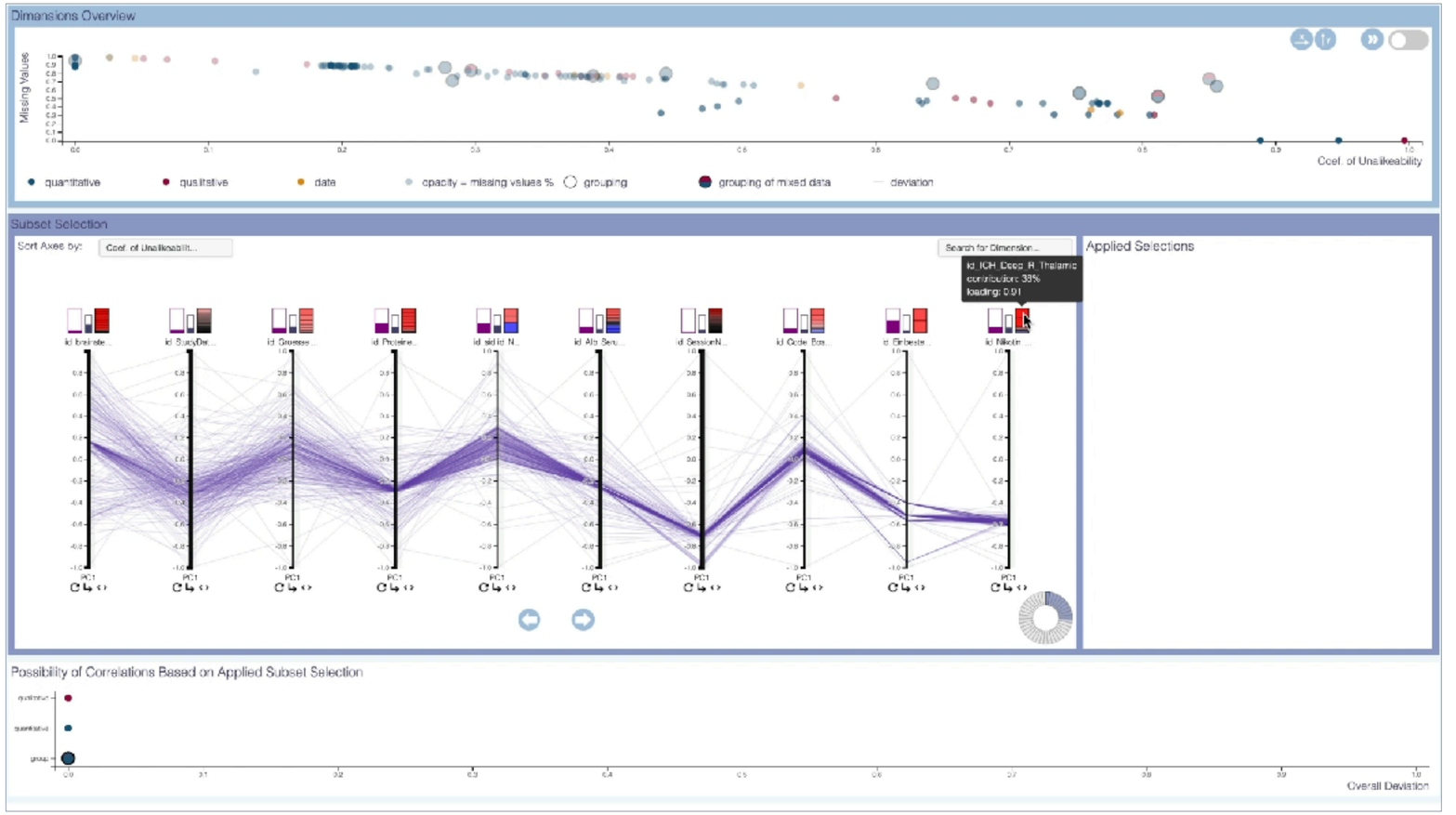


Image from Dr. Savvas Nicolaou

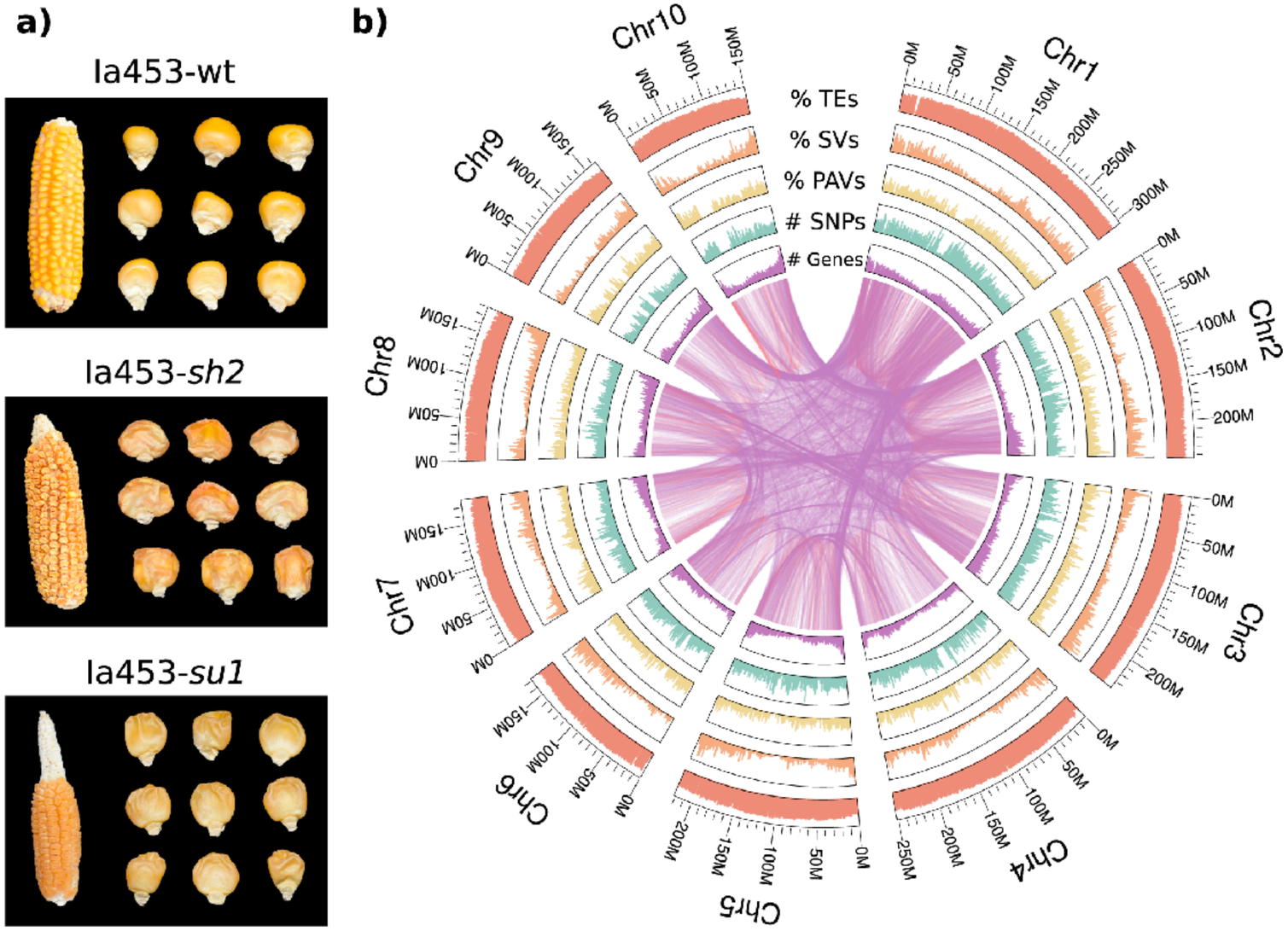
To augment human abilities, not replace



Garrison et al. 2021

Munzner, T. (2014). Visualization analysis and design. AK Peters Visualization Series, CRC Press, Visualization Series. Chp 1

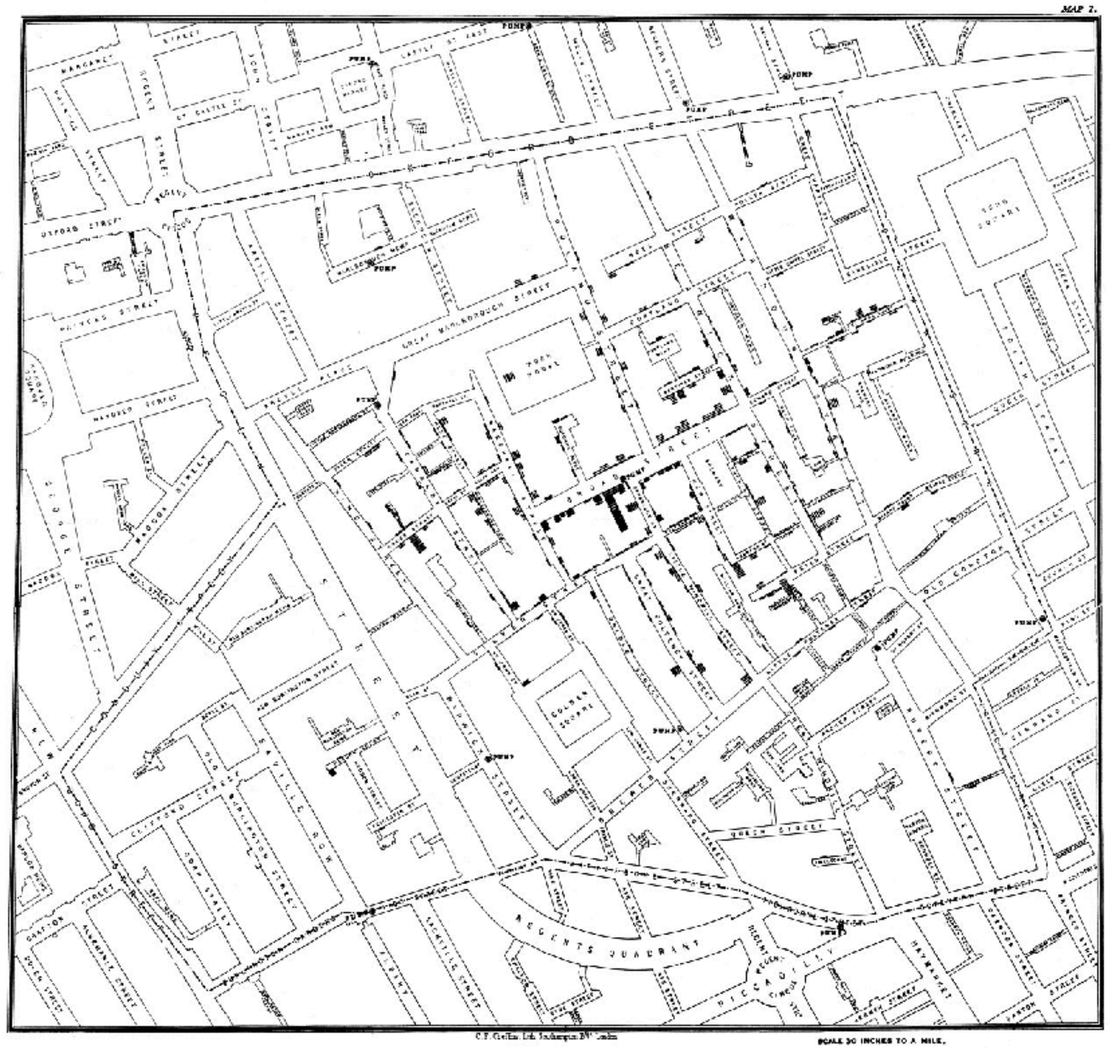
To ask questions



[Appearance of sweet corn in Ia453 isolines and Ia453-sh2 genomic features]

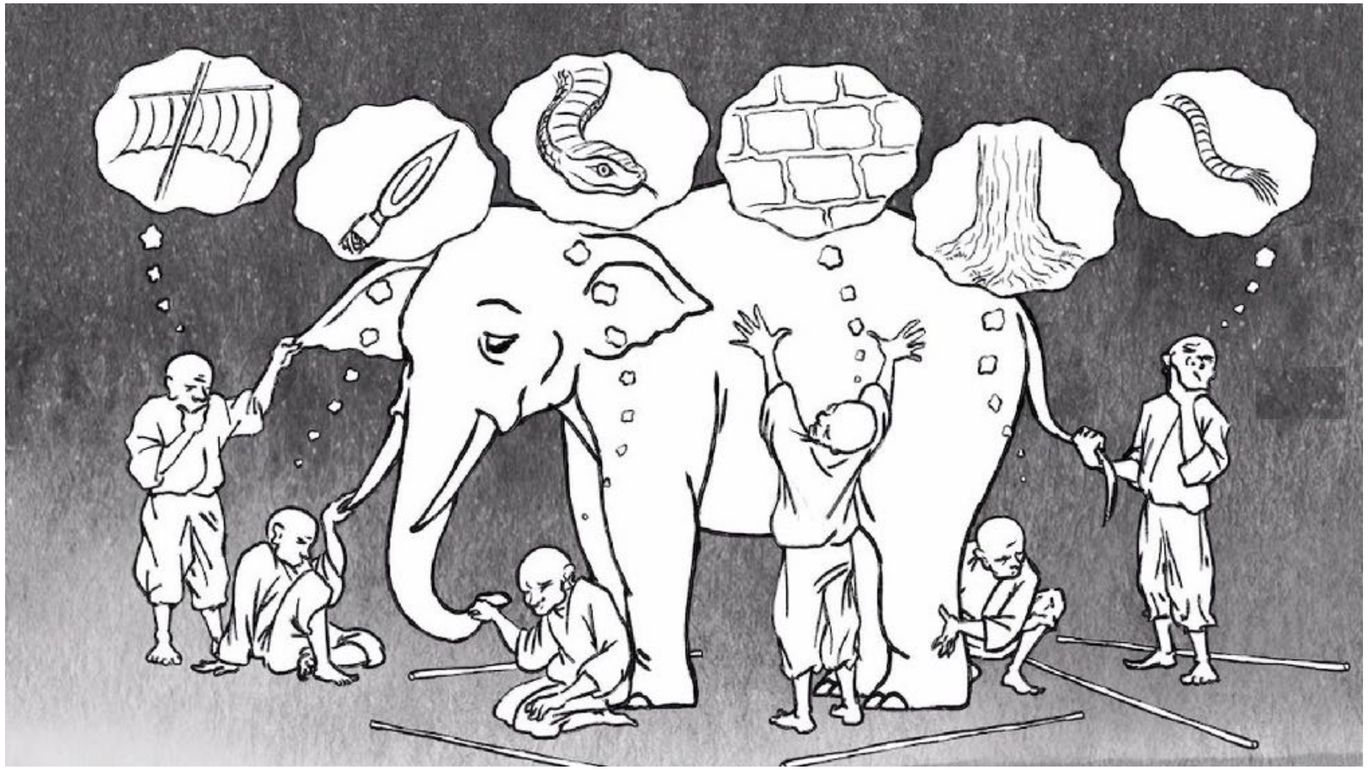
To confirm hypotheses

Snow's 1854 cholera map
[Wikimedia]



To explain differences

- **Rashomon Effect**
  - Different representations tell different stories
- Visualization can help us identify and understand reasons for these differences

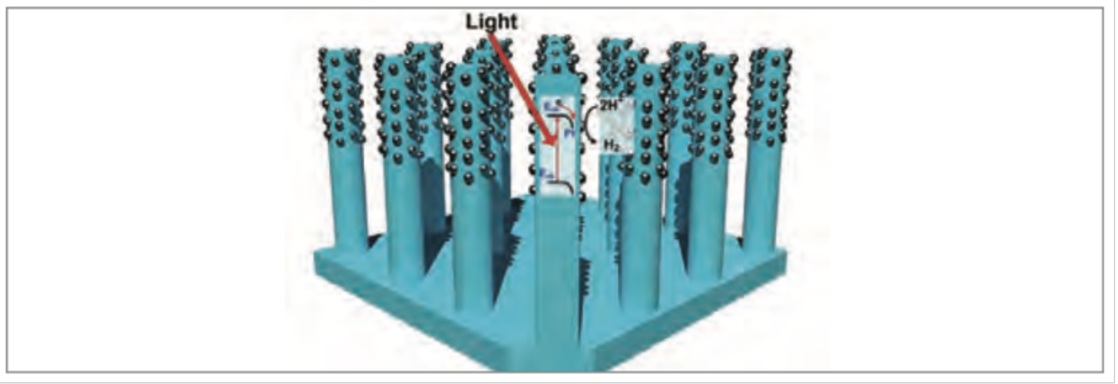


Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science, 16(3), 199-231.

To communicate information…

- High-quality visualizations are perceived as **“more interesting, clearly written, and more scientifically rigorous”** [1]
- Highly-cited papers tend to have **more diagrams** per page [2]

Before redesign:



After redesign:

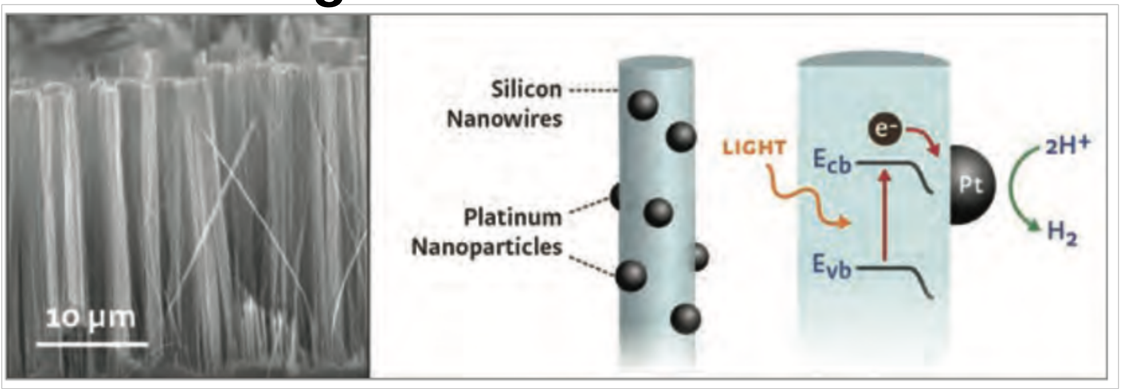
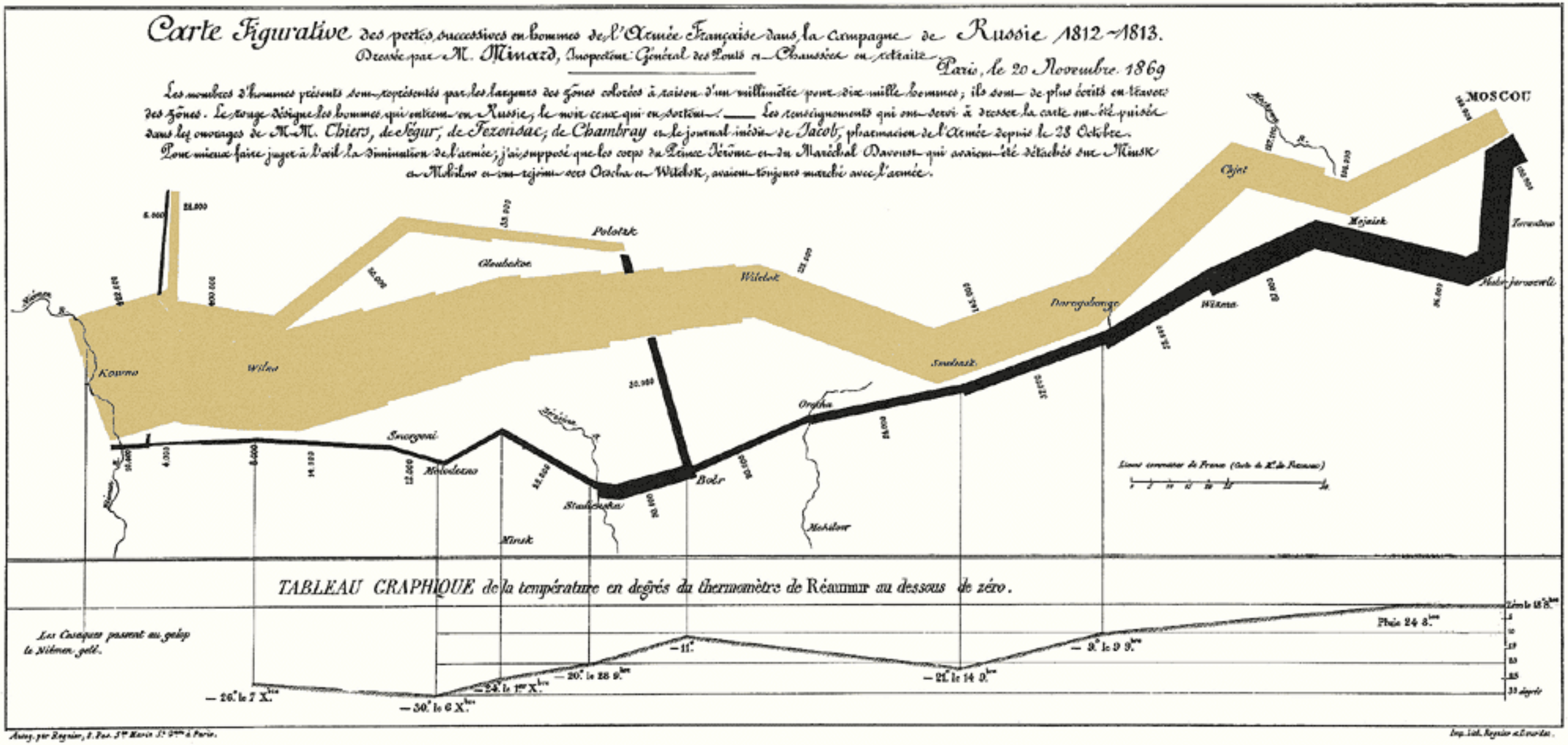


Image Credit: Jen Christiansen, Building Science Graphics (CRC Press 2023), p28

[1] K. Cheng et al., “Proving the Value of Visual Design in Scientific Communication,” Information Design Journal, Vol. 23, 2017.
[2] P. Lee, J. D. West, and B. Howe, “Viziometrics: Analyzing Visual Information in the Scientific Literature,” IEEE Transactions on Big Data Vol. 4, 2018.

To communicate information…



Figurative map of the successive losses in men of the French Army in the Russian campaign 1812–1813, by Charles Minard (1869)

…and connect


The New York Times
LOG IN
OPINION
GUEST ESSAY
1,374 DAYS
MY LIFE WITH LONG COVID

Making a visualization is easy.

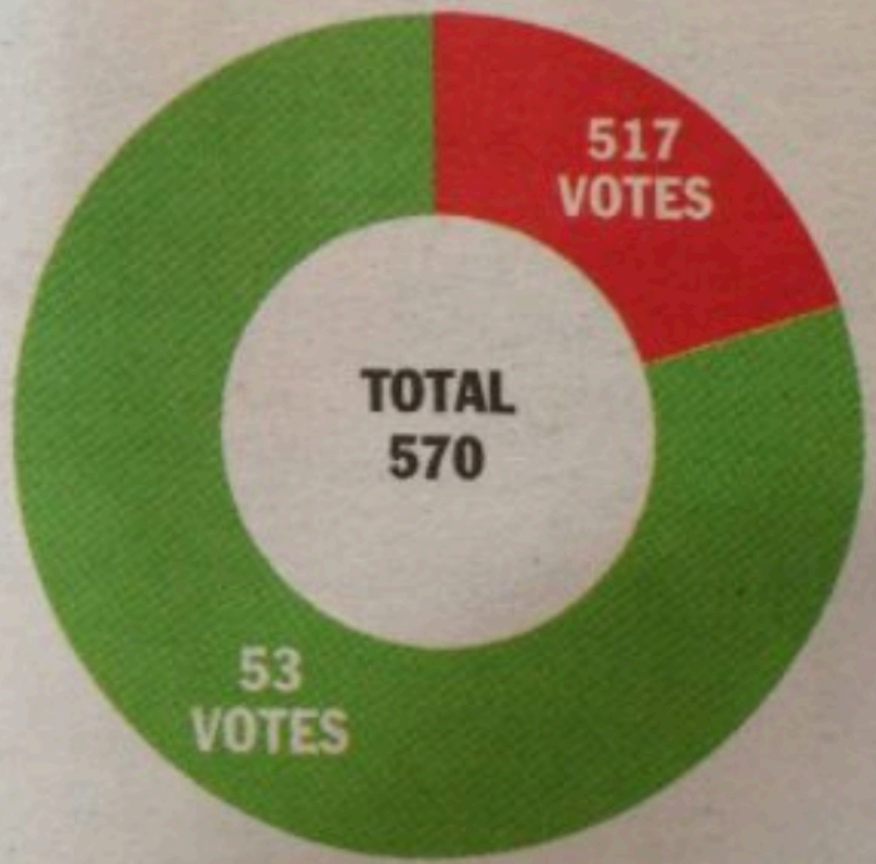
Making a *good* visualization is hard.

VOTELINE

SATURDAY'S RESULT

Can Julia Gillard win the next federal election?

10% YES | **90% NO**



TODAY'S QUESTION

Do you like the plan to remove a car lane of the Princes Bridge for sole use by cyclists?

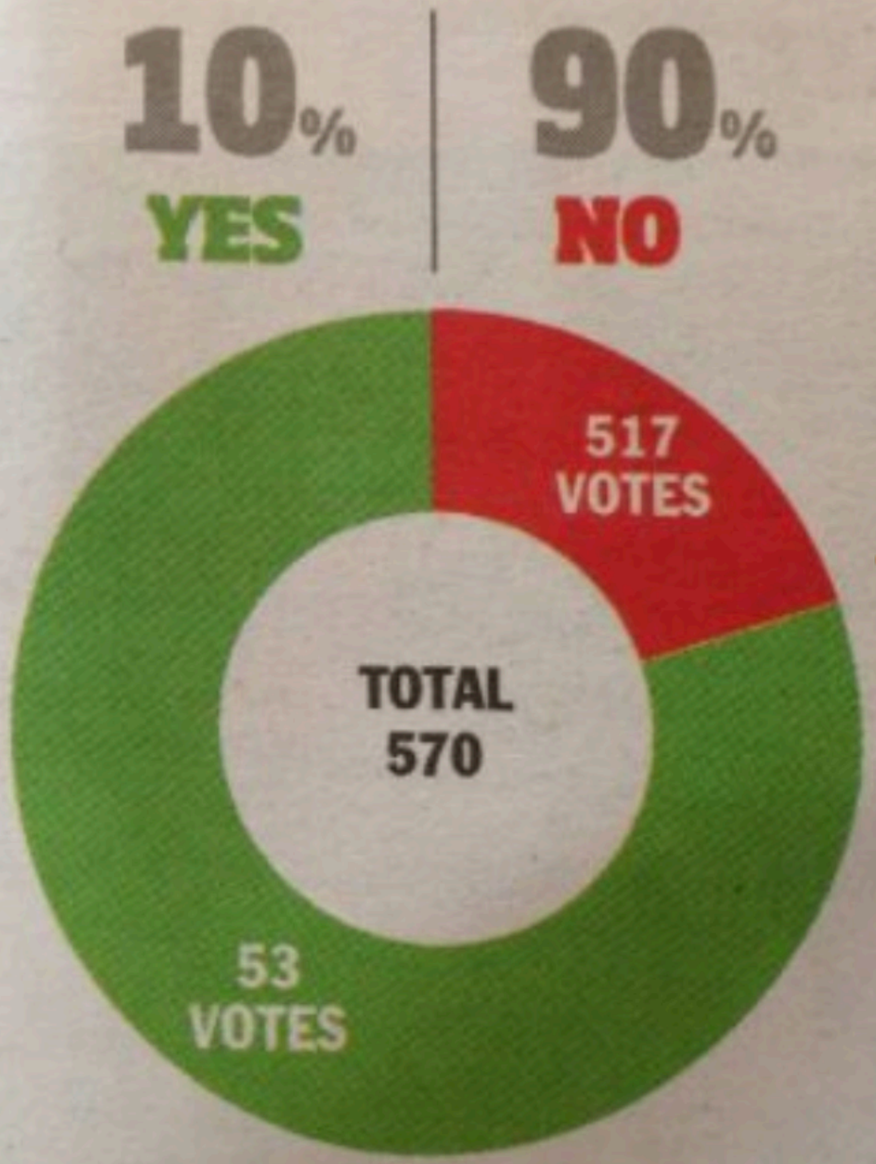
YES	1900 956 434
NO	1900 956 435

Calls cost 38.5c including GST. You can also have your say at heraldsun.com.au or facebook.com/heraldsun

VOTELINE

SATURDAY'S RESULT

Can Julia Gillard win the next federal election?



TODAY'S QUESTION

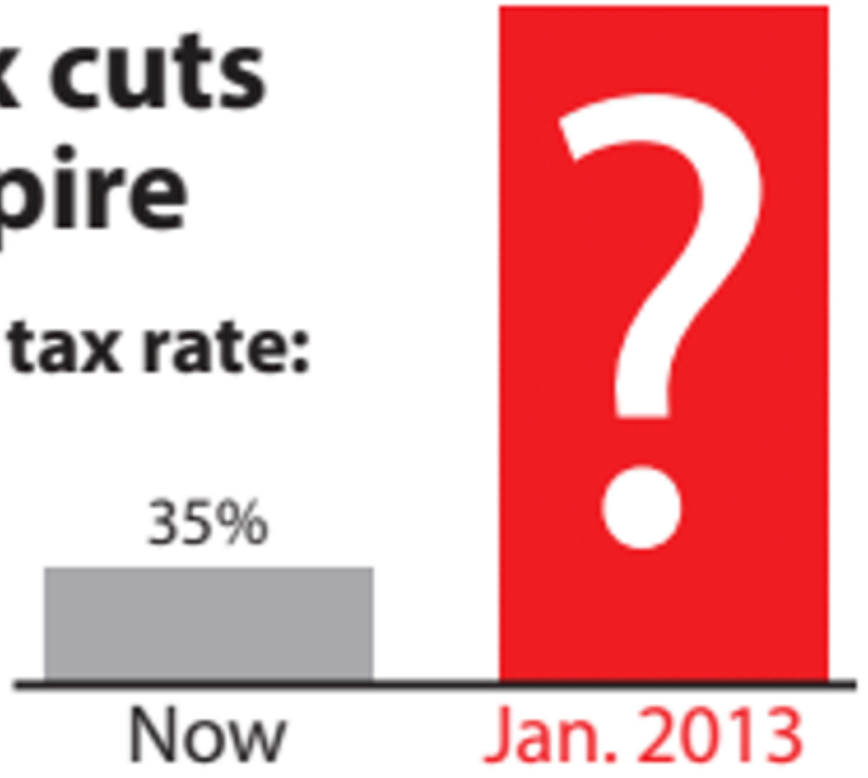
Do you like the plan to remove a car lane of the Princes Bridge for sole use by cyclists?

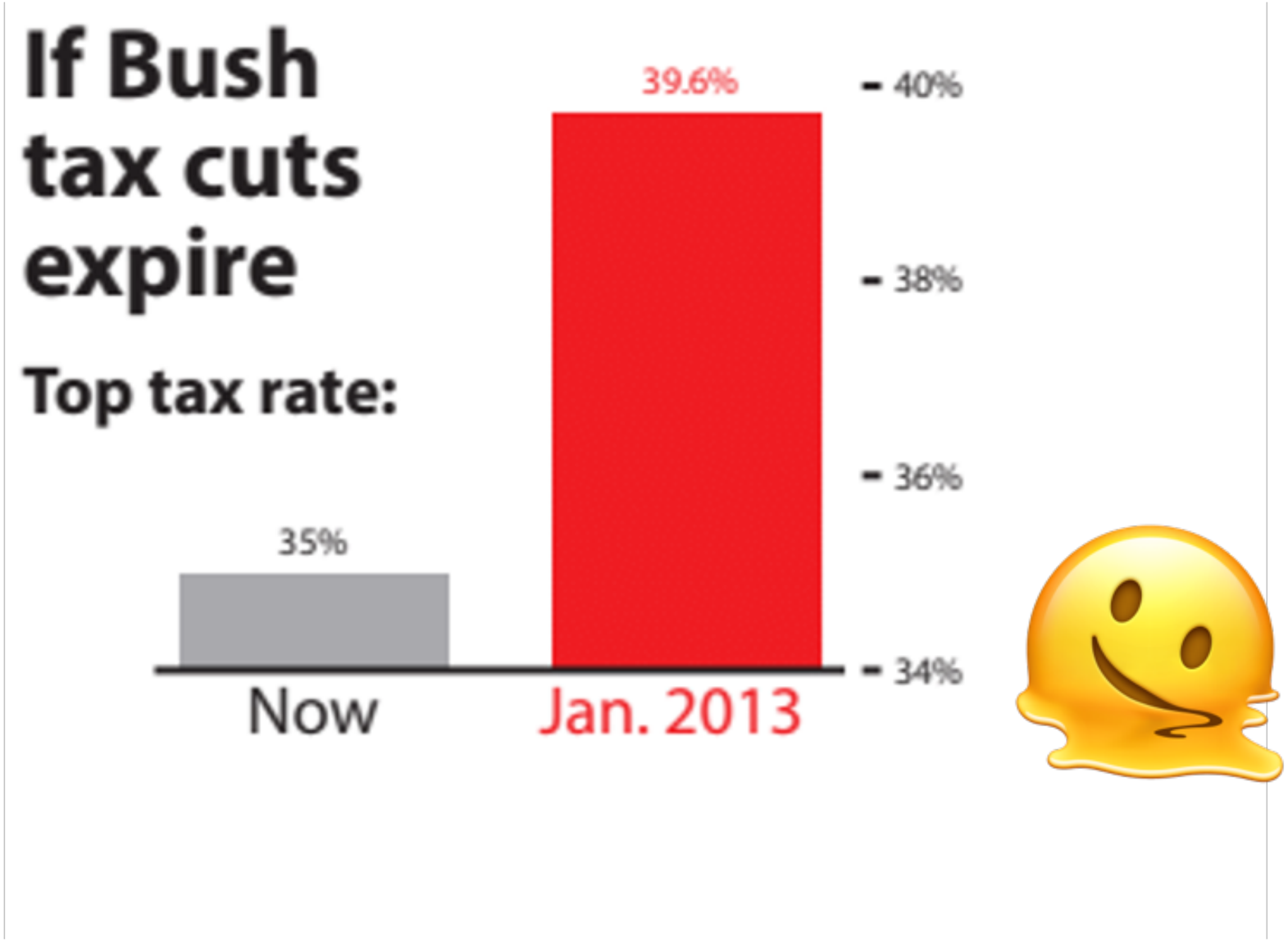
YES	1900 956 434
NO	1900 956 435

Calls cost 38.5c including GST. You can also have your say at heraldsun.com.au or facebook.com/heraldsun

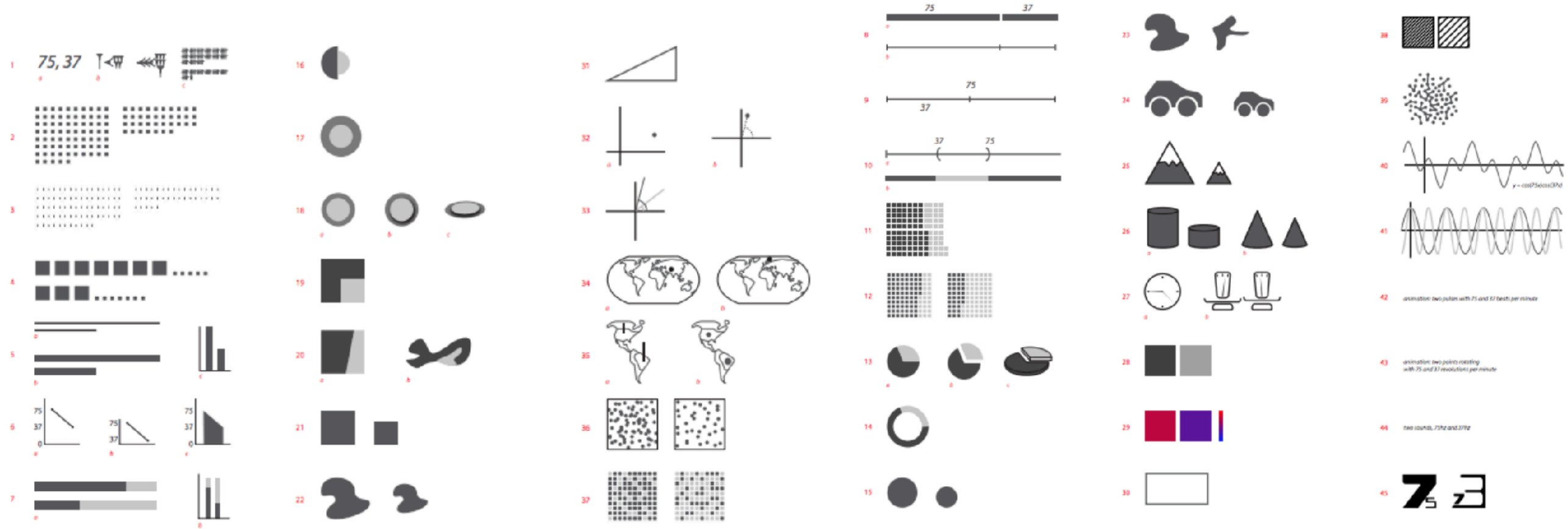
If Bush tax cuts expire

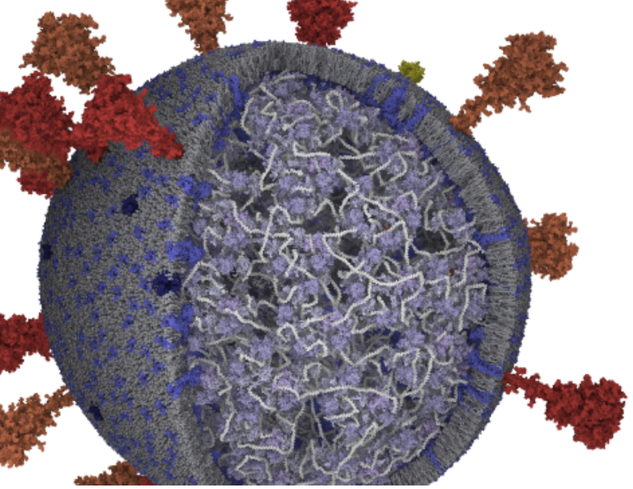
Top tax rate:

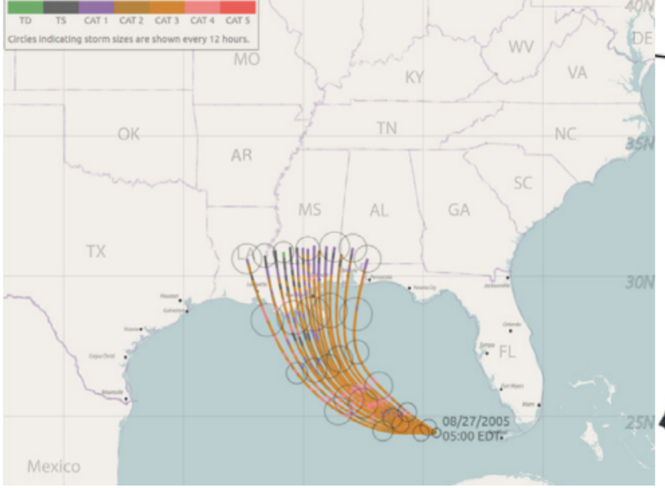


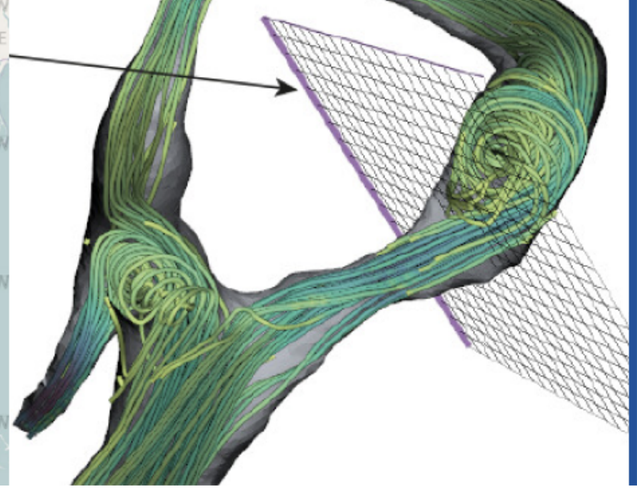

If Bush tax cuts expire
Top tax rate:
39.6%
35%
40%
38%
36%
34%
Now
Jan. 2013

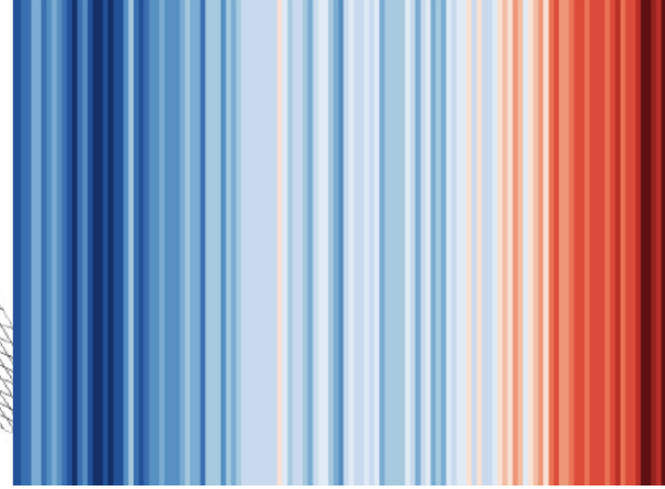
45 ways to communicate 75 and 37





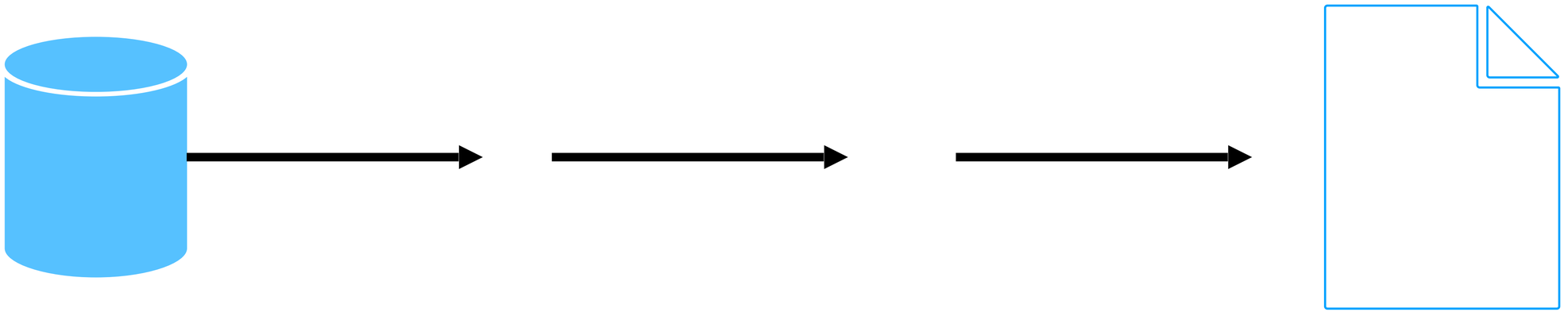




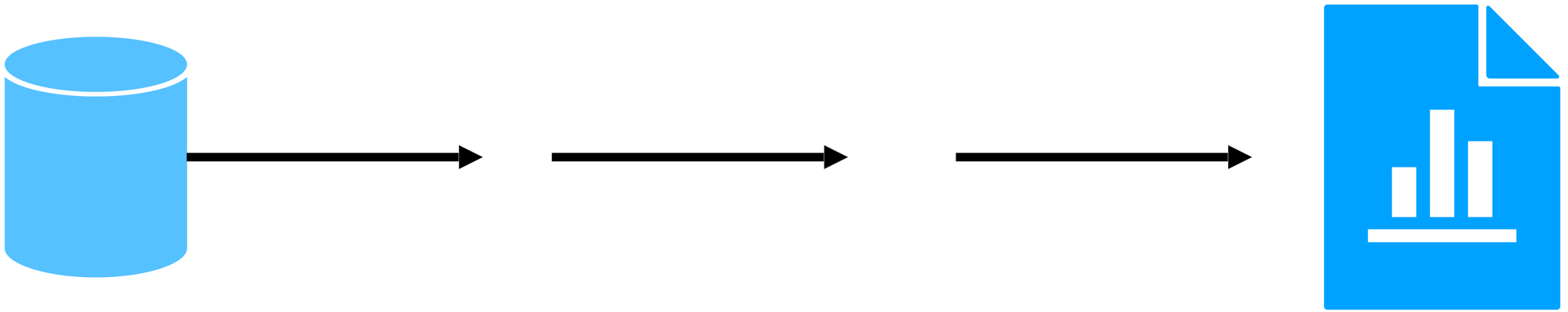


Programming concepts in visualization

Data Transformation



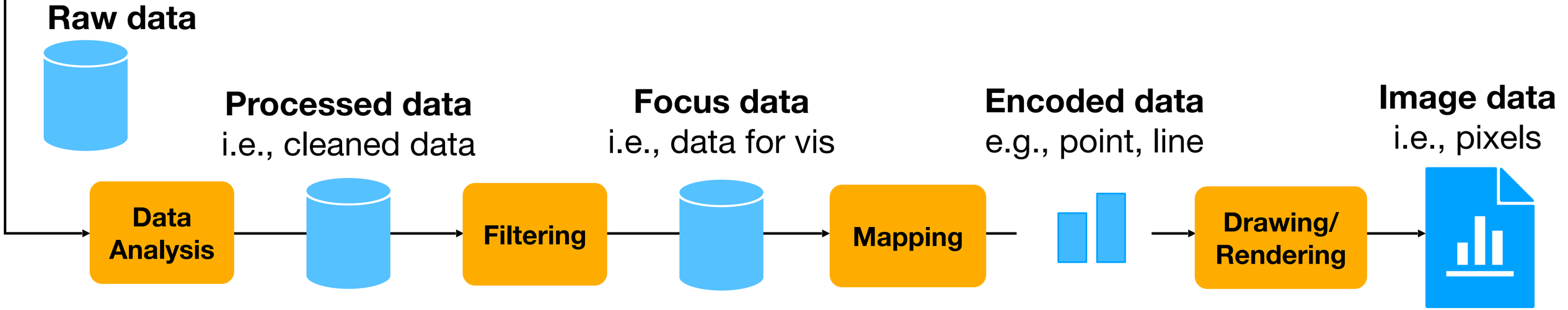
Data Transformation



Data are key to the visualization pipeline

Domain
Metadata, semantics, conventions

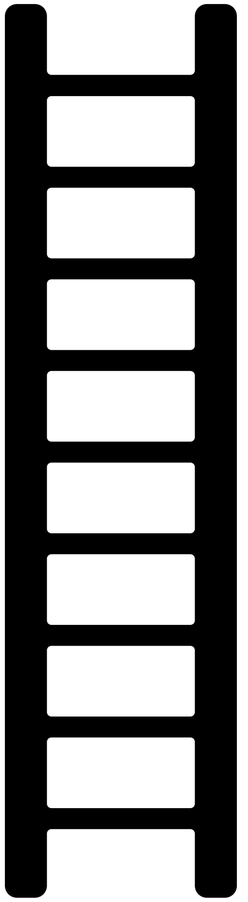
Task
Questions, goals, assumptions

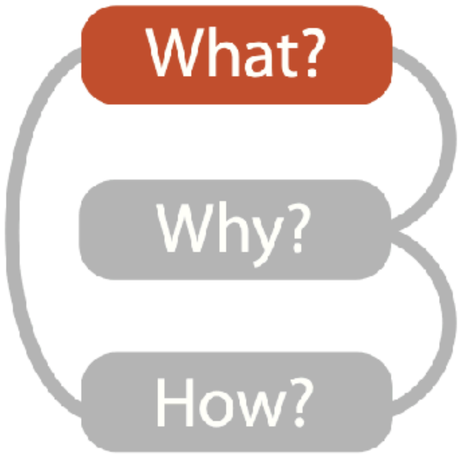


https://infovis-wiki.net/wiki/Visualization_Pipeline

Abstraction helps us think about data

- Common vocabulary for talking about visualization
- Allows for general discussion about specific problems
- Transfer ideas between domains





Data Abstraction

What type of data are you using?

Data type: Item

- Individual, discrete entity
  - e.g., Dog 🐶 or Coffee shop ☕
- Represented as…
  - Rows in data table
  - Nodes in network

row

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

Data type: Attribute

(variables, dimensions)

- Specific, measurable/observable property
  - e.g., Dog breed, weight, fur color, etc. 🐕 🐩 🐕‍🦺
  - e.g., Coffee shop rating, whether offer baked goods, etc.

Attributes can be…

- Quantitative/continuous 
- Categorical
  - Ordered (ordinal) 
  - Unordered (nominal) 

Attributes can be…

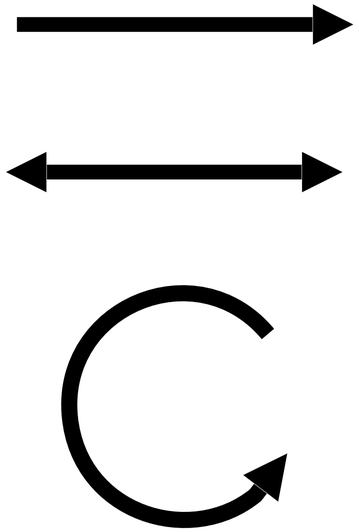
- Quantitative/continuous
- Categorical
  - Ordered (ordinal)
  - Unordered (nominal)

For item “dog”, what attributes might fall under each of these attribute types?

- Breed = Nominal (unordered cat.)
- Breed sizes = Ordinal (ordered cat.)
- Weight = Quantitative

Attributes can be…

- Sequential
- Diverging
- Cyclic



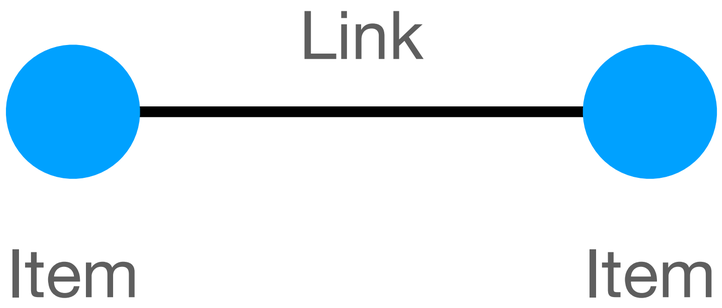
Attributes can be…

- **Temporal attribute:** anything related to time
- Time can be **complicated -** hierarchical, multiscale

https://cfmm.uwo.ca/research/cardiac.html

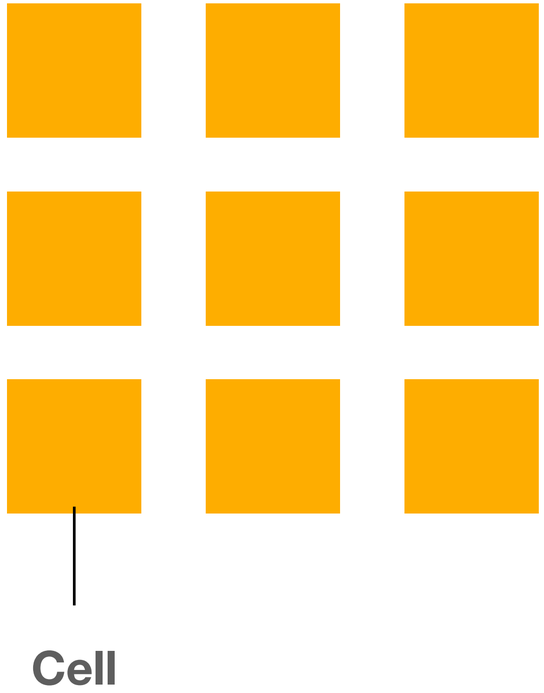
Other data types

Link



Position

Grid (of positions)



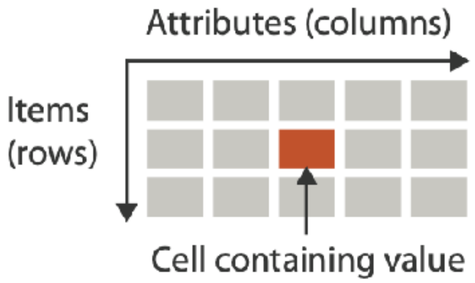
Data → Datasets

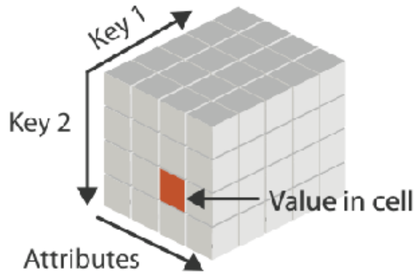
How are data types combined into a larger structure?

Where do the data “live”? What **“space”** do they live in?

Tables

- Cells indexed by **items and attributes**
  - Rows = items
  - Columns = attributes
  - Value = (item, attribute)
- Flat or multi-dimensional

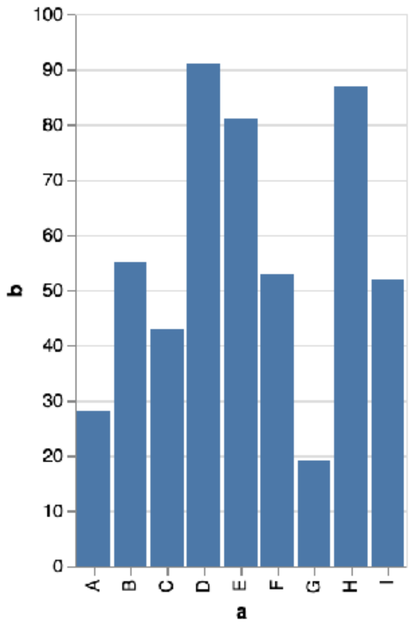




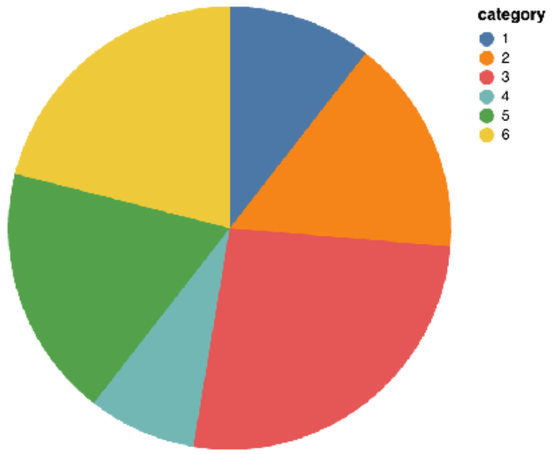
A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	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

Tables - Examples

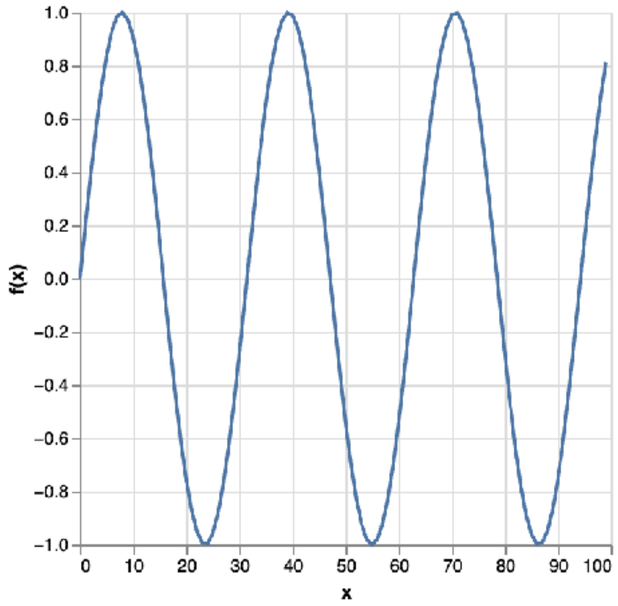
Bar chart



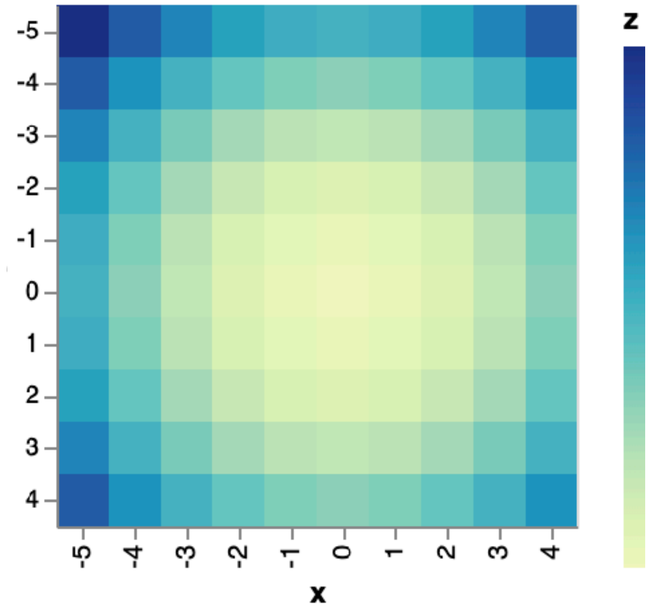
Pie chart



Line chart



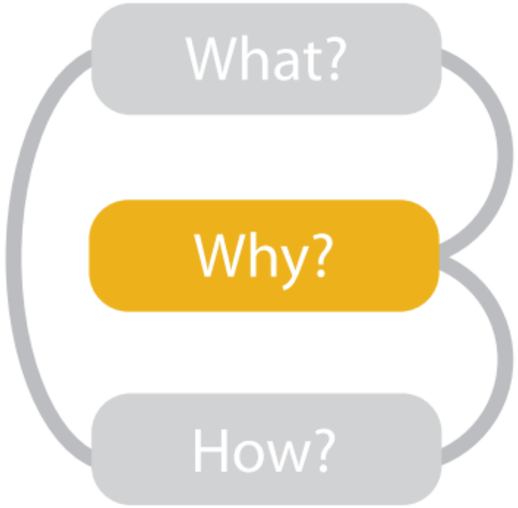
Heatmap



https://altair-viz.github.io/gallery/simple_bar_chart.html
https://altair-viz.github.io/gallery/pie_chart.html
https://altair-viz.github.io/gallery/simple_line_chart.html
https://altair-viz.github.io/gallery/simple_heatmap.html

Dataset availability

- **Static** (offline): entire dataset is available all at once
  - e.g., Hawks and Penguin
- **Dynamic** stream (online): data update
  - **Add** new/**delete** old
  - **Update** existing item values

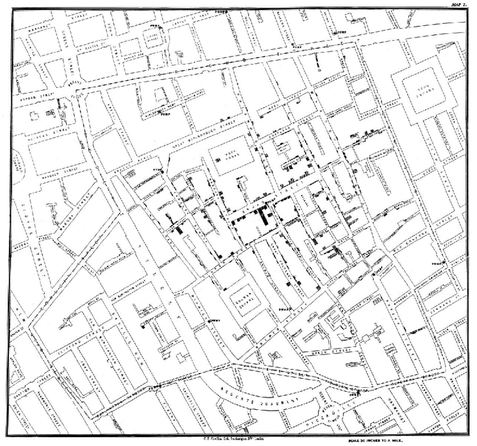


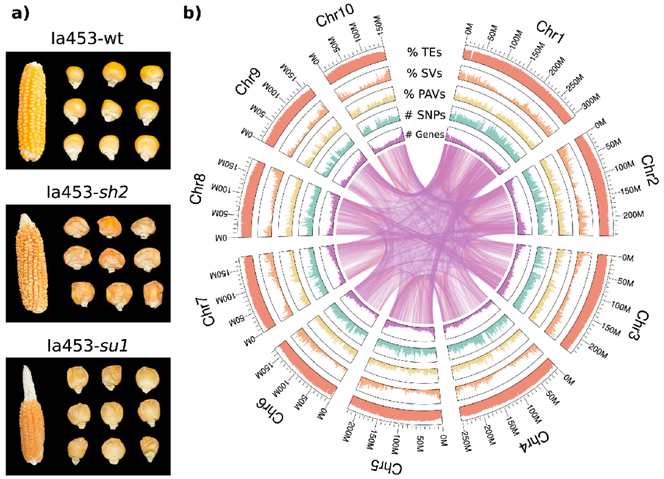
Task Abstraction

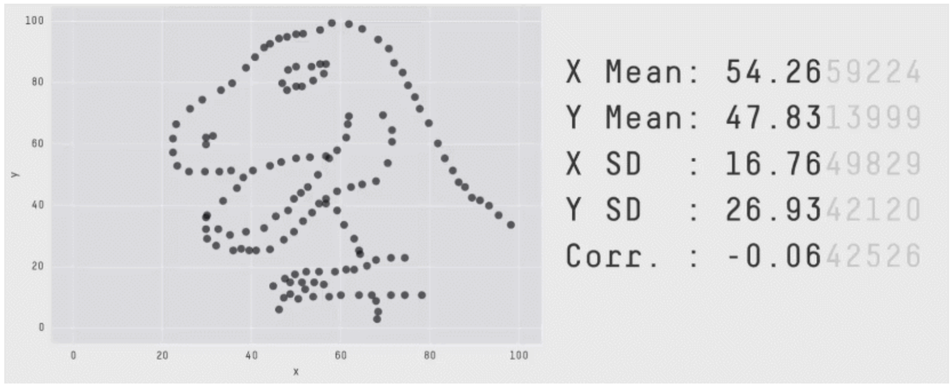
Why are we looking at the data?

Typical visualization tasks

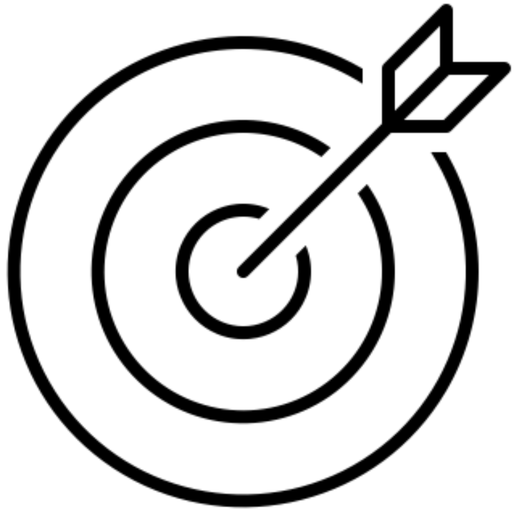
- **Explore**
  - Find the unknown or unexpected
  - Form hypotheses
- **Analyze**
  - Confirm/reject hypotheses
  - Drill-down for details
- **Present**
  - Convey information
  - Share results
  - Share a story/create empathy







Tasks are a combo of actions + targets

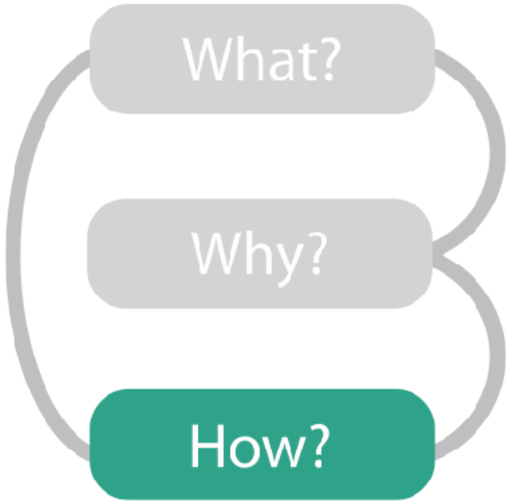


Action:
Goal

+

Target:
What part(s) of data is your goal focusing on?

- Discover distribution
- Compare trends
- Browse outliers
- Explore topology



Encodings

How do you show the data,
based on the data type and tasks?

How to choose the “right” visual encodings?

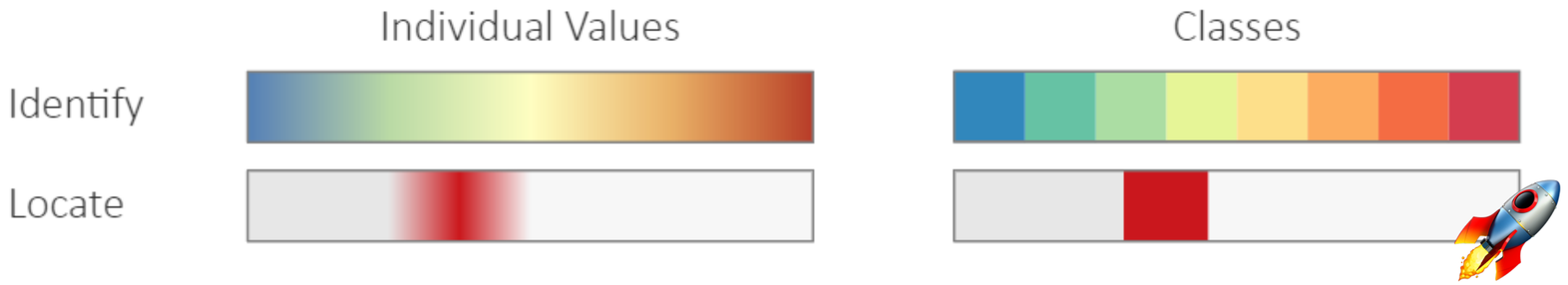
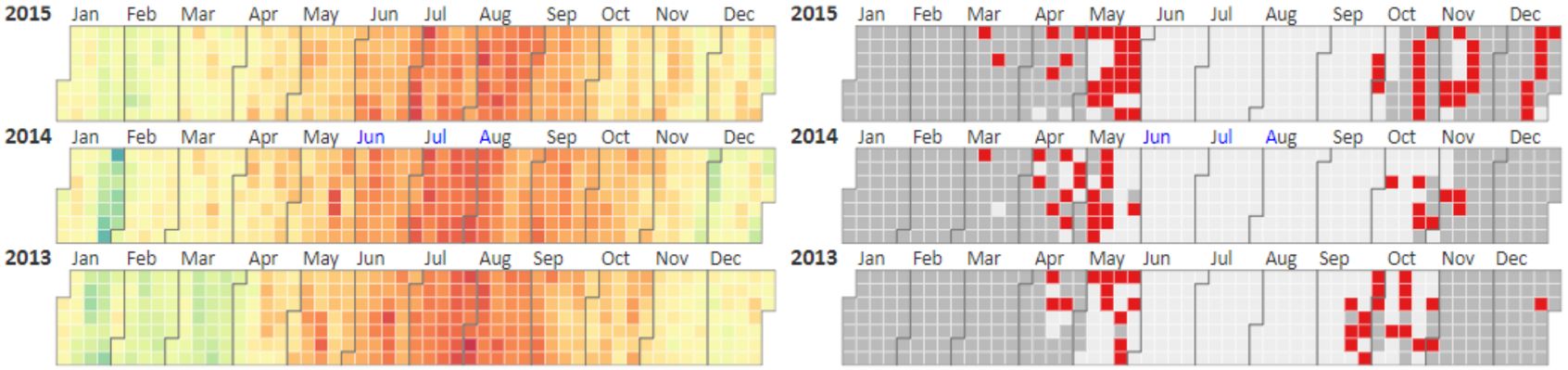


Figure 3.4 Color maps for identifying and locating values and classes.

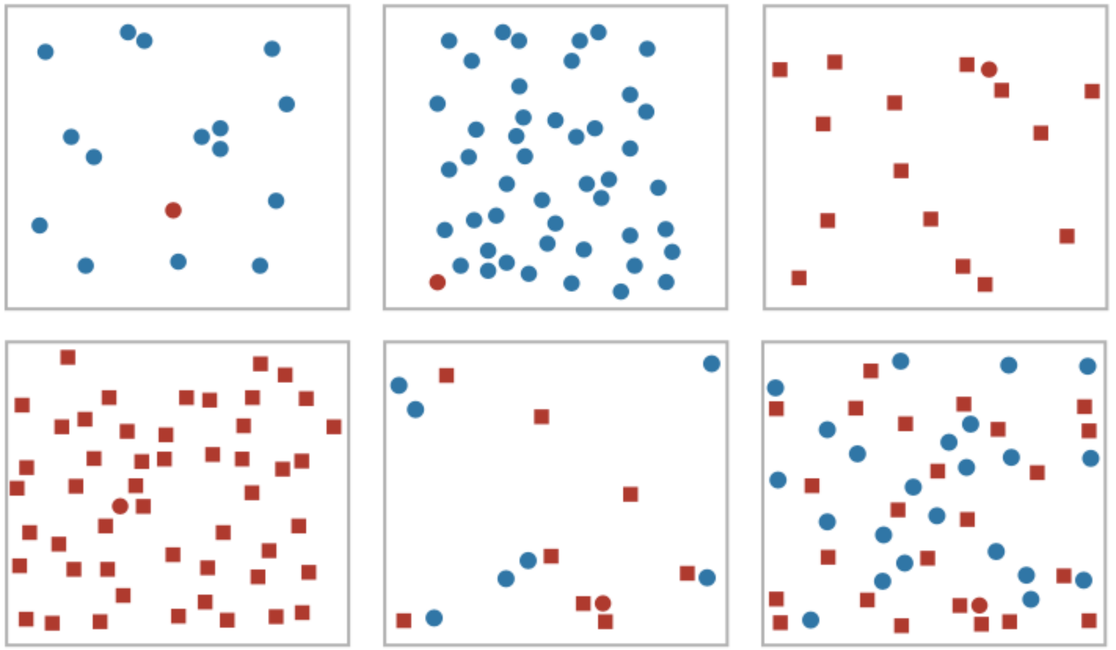


(a) Color coding for identification tasks. (b) Color coding for location tasks.

Figure 3.5 Applying the color maps from Figure 3.4 to temperature data. Adapted from bl.ocks.org/mbostock/4063318.

Preattentive processing

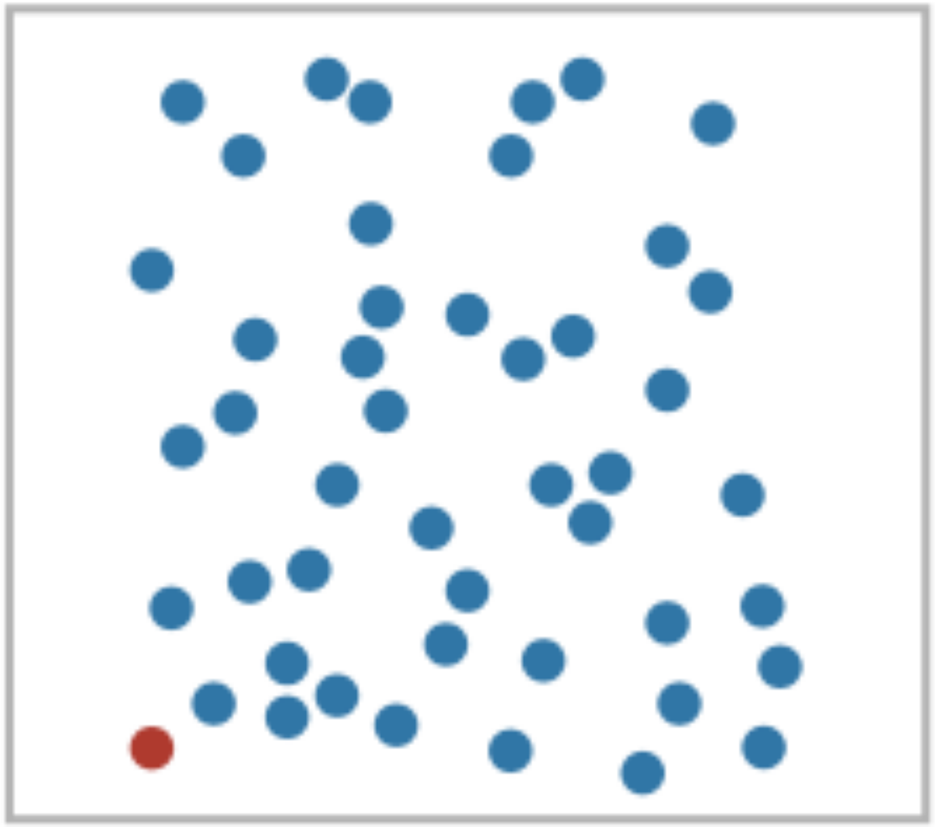
- **Very fast**: < 200 ms
- Contrast between features most important
- Make object “**pop out**”
- Requires attention, despite name



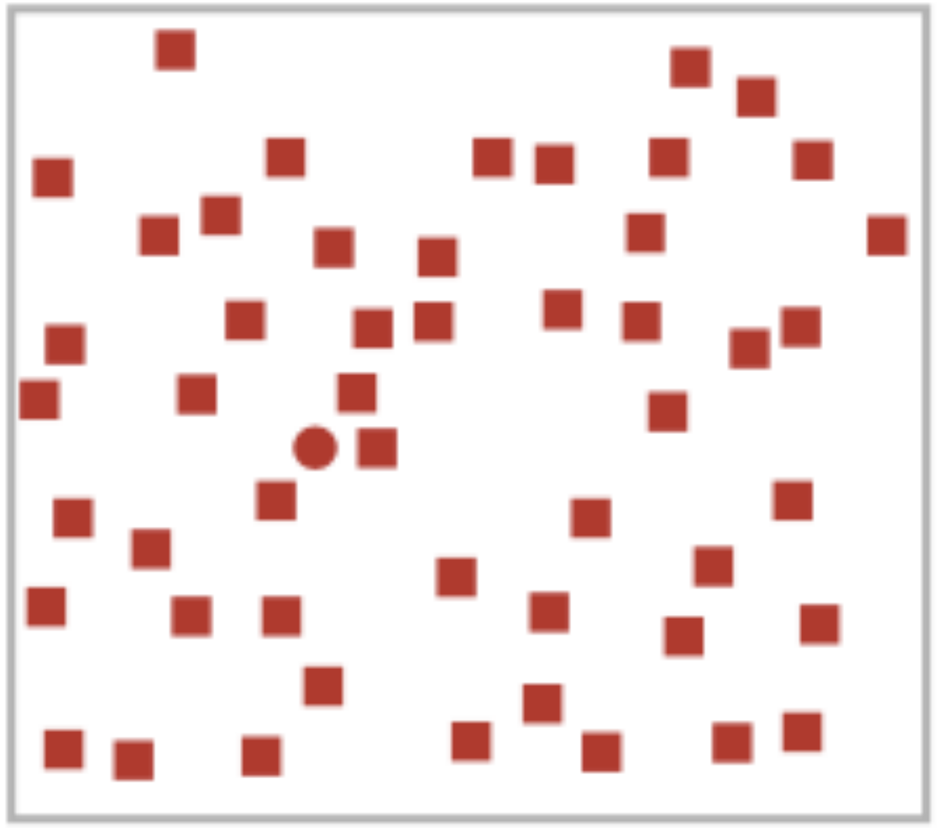
VAD Chp 5

Wolfe JM, Utochkin IS. What is a preattentive feature?. Current opinion in psychology. 2019 Oct 1;29:19-26.

Find the red dot



Find the red dot



How many V's?

MTHIVLWYADCEQGHKILKMTWYN
ARDCAIREQGHLVKMFPSTWYARN
GFPSVCEILQGKMFPSNDRCEQDIFP
SGHLMFHKMVPSTWYACEQTWRN

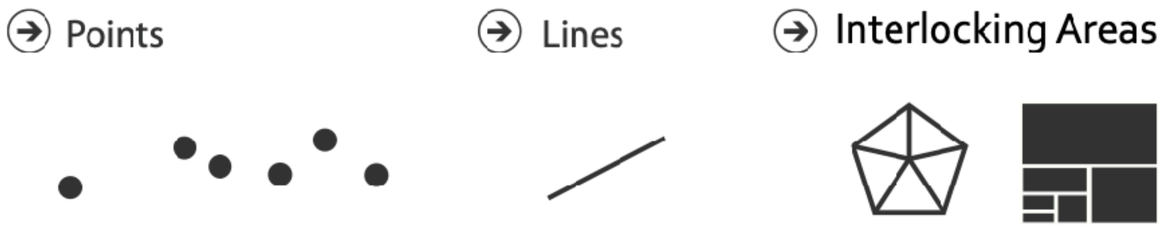
Preattentive Features

- **Form**
  - Line orientation
  - Line length
  - Line width
  - Line collinearity
  - Size
  - Curvature
  - Spatial grouping
  - Added marks
  - Numerosity
- **Color**
  - Hue & Intensity
- **Motion**
  - Flicker
  - Direction of motion
- **Spatial position**
  - 2D position
  - Stereoscopic depth
  - Convex/concave from shading

Marks

- Graphical elements in an image
- Classified according to number of spatial dimensions

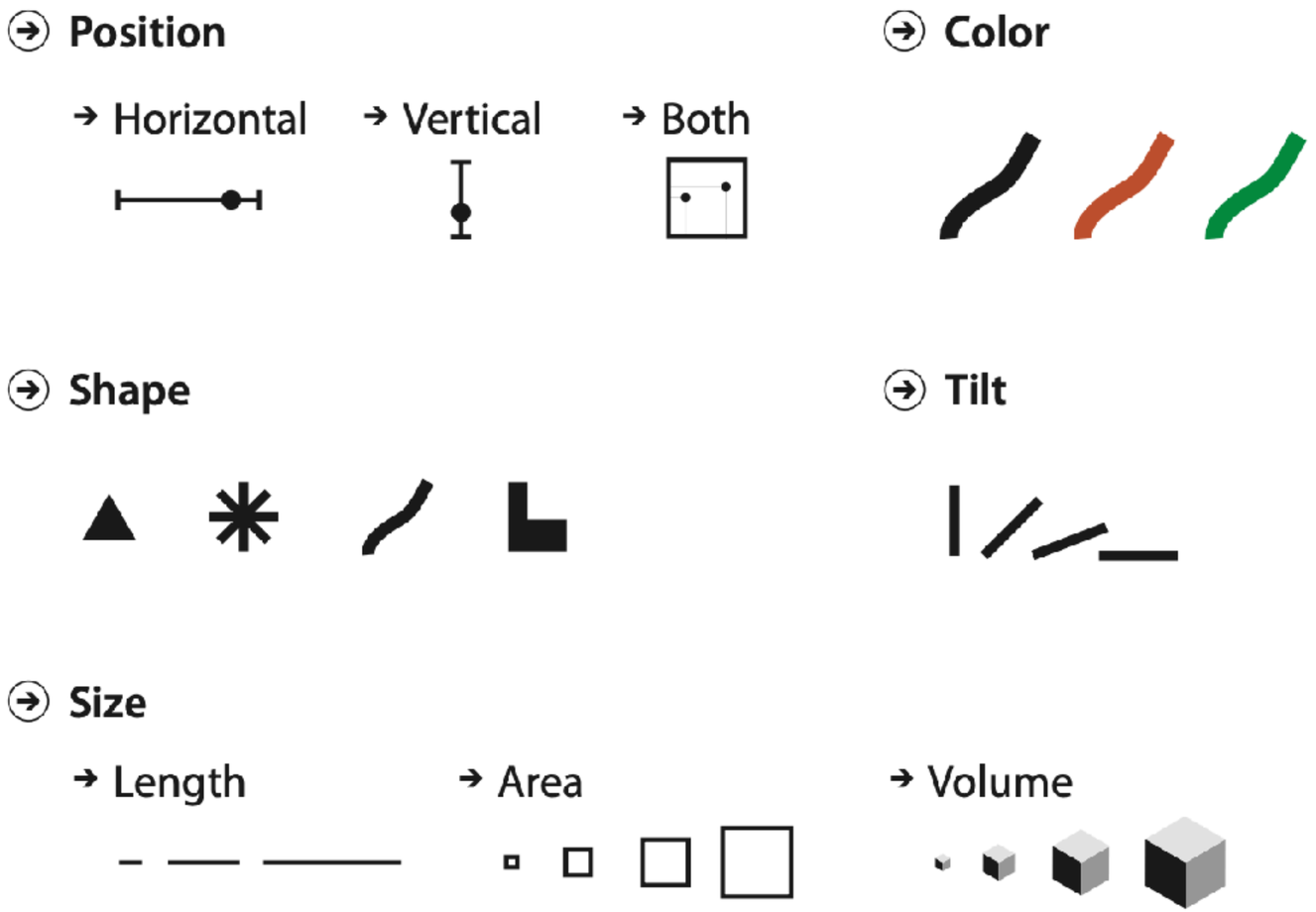
Marks as items/nodes



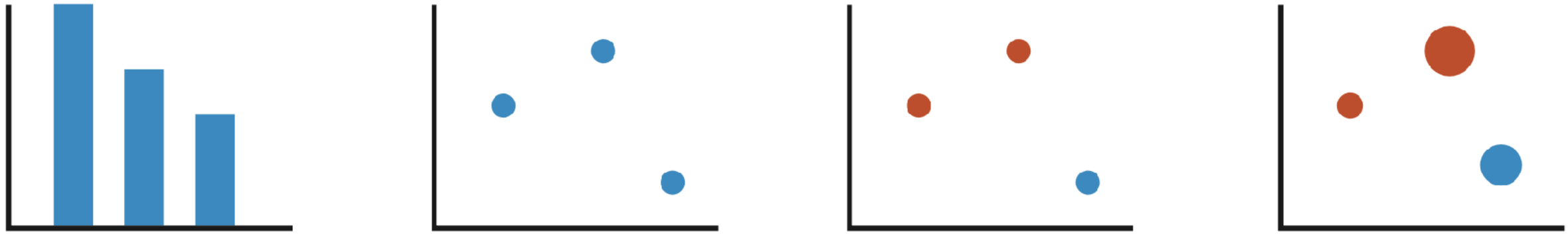
Marks as links

Channels

- Channels control mark appearance
- Are independent of the mark (geometric primitive)

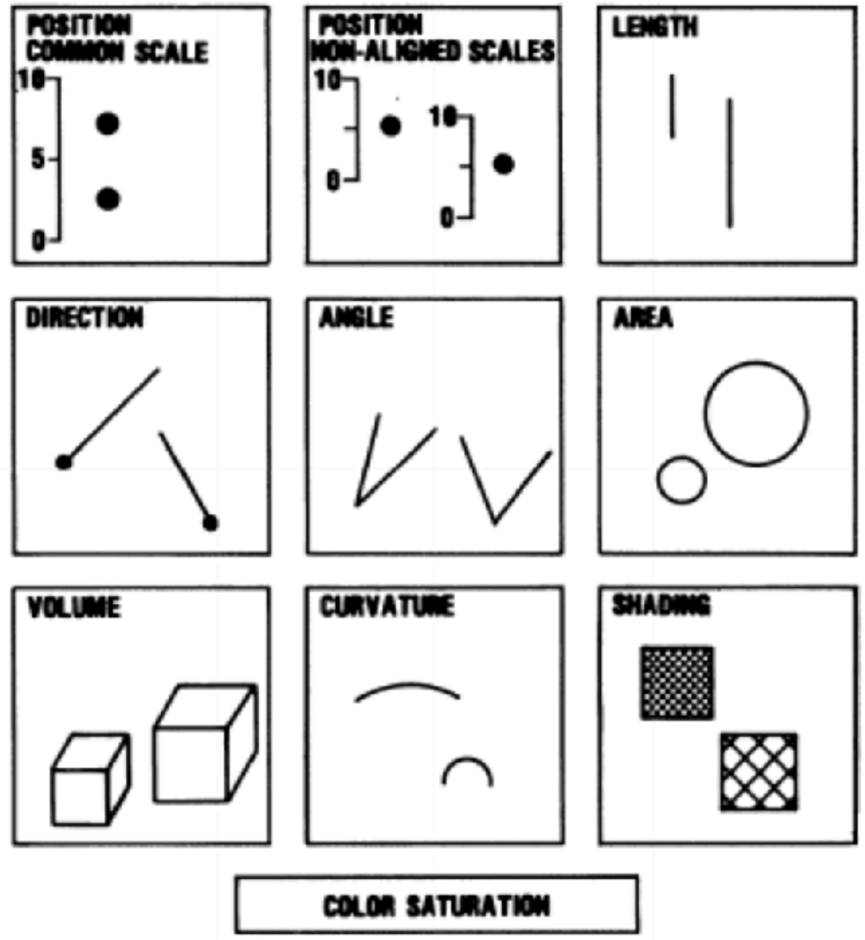


Applying marks and channels



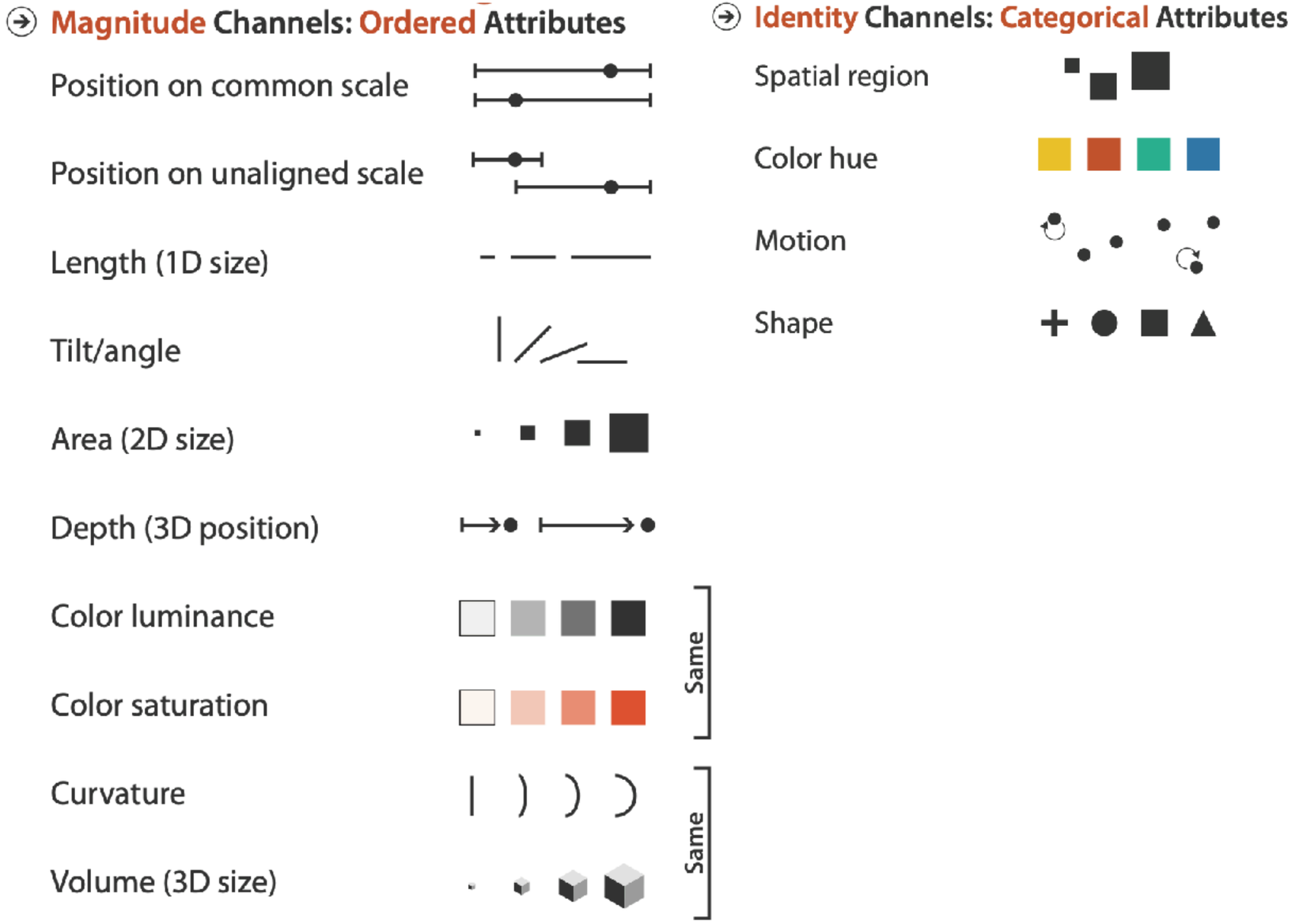
Cleveland & McGill 1984

- Foundational article in Visualization
- Elementary perceptual tasks that viewer performs to extract values of real variables shown on most charts



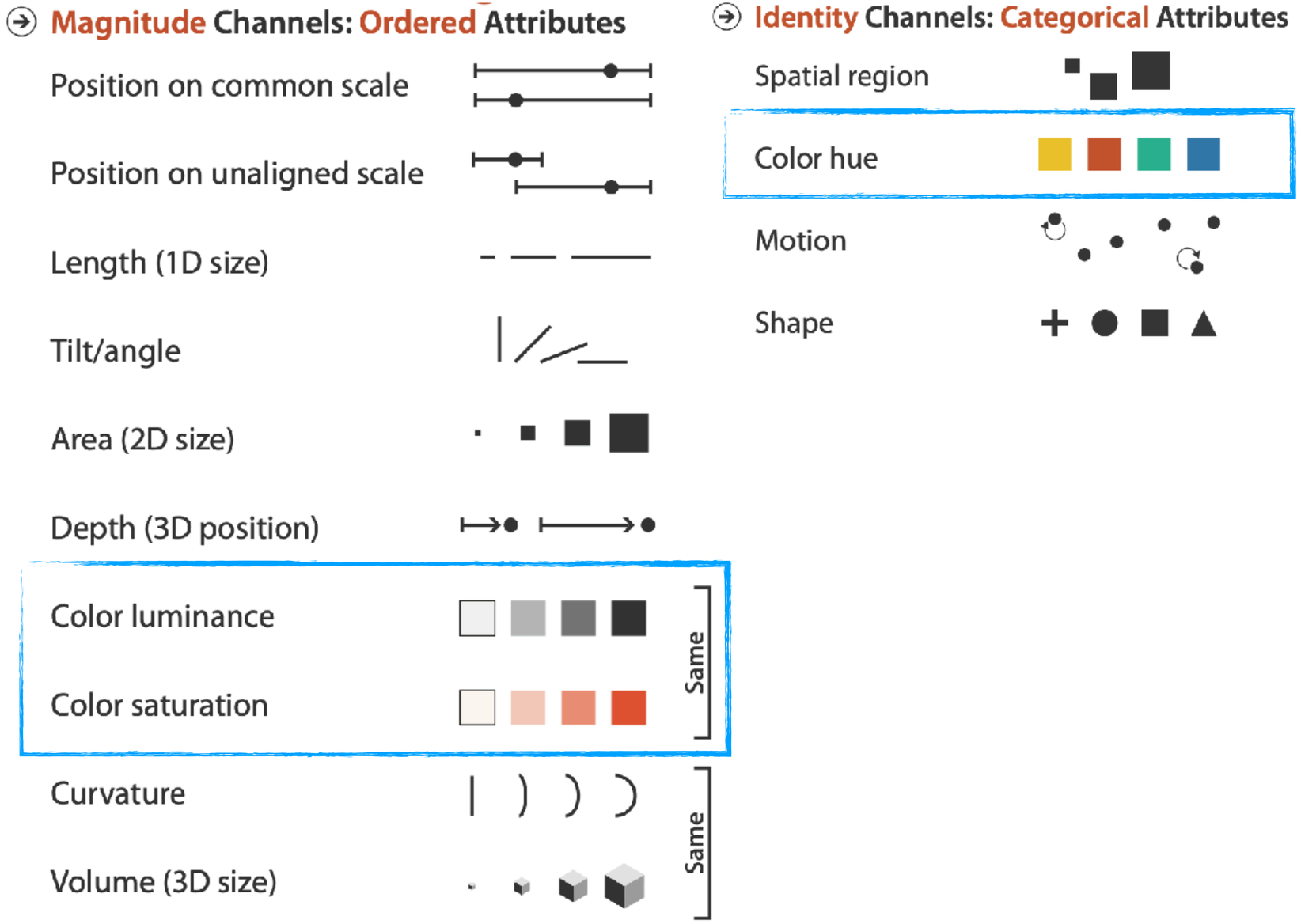
Channel Rankings

- **Expressiveness**
  - Encoding describes information in the dataset thoroughly
- **Effectiveness**
  - Encode the most important attributes with the most perceptually effective channels



Cleveland & McGill 1984, Heer & Bostock 2010, Munzner 2014

Channel Rankings

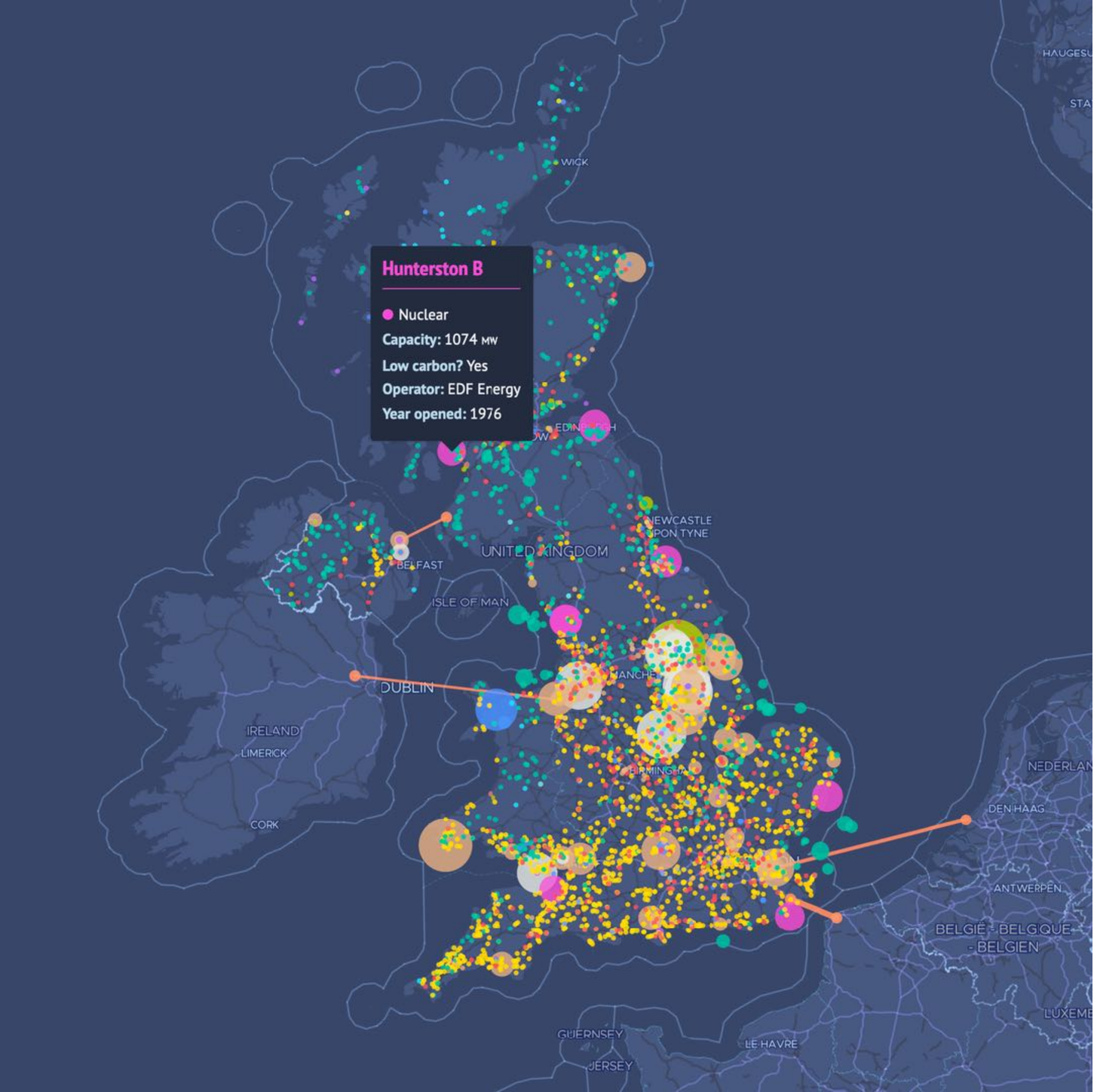


Cleveland & McGill 1984, Heer & Bostock 2010, Munzner 2014

Categorical Encoding

This is a discrete gradient – good for binned quant data, or ordered

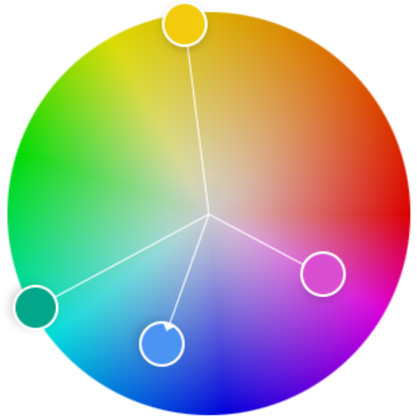
https://matplotlib.org/stable/users/explain/colors/colormaps.html#colormaps



How The UK Transformed Its Electricity Supply In Just A Decade
by Carbon Brief

Information is Beautiful 2023
Award Short list

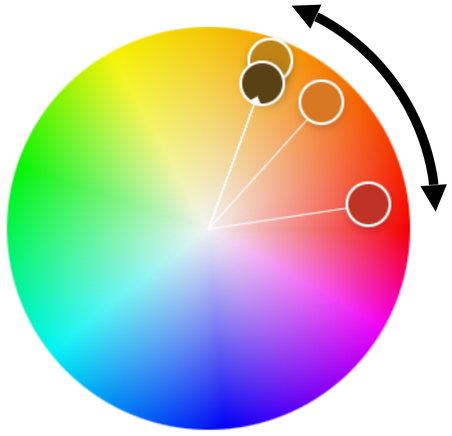
link



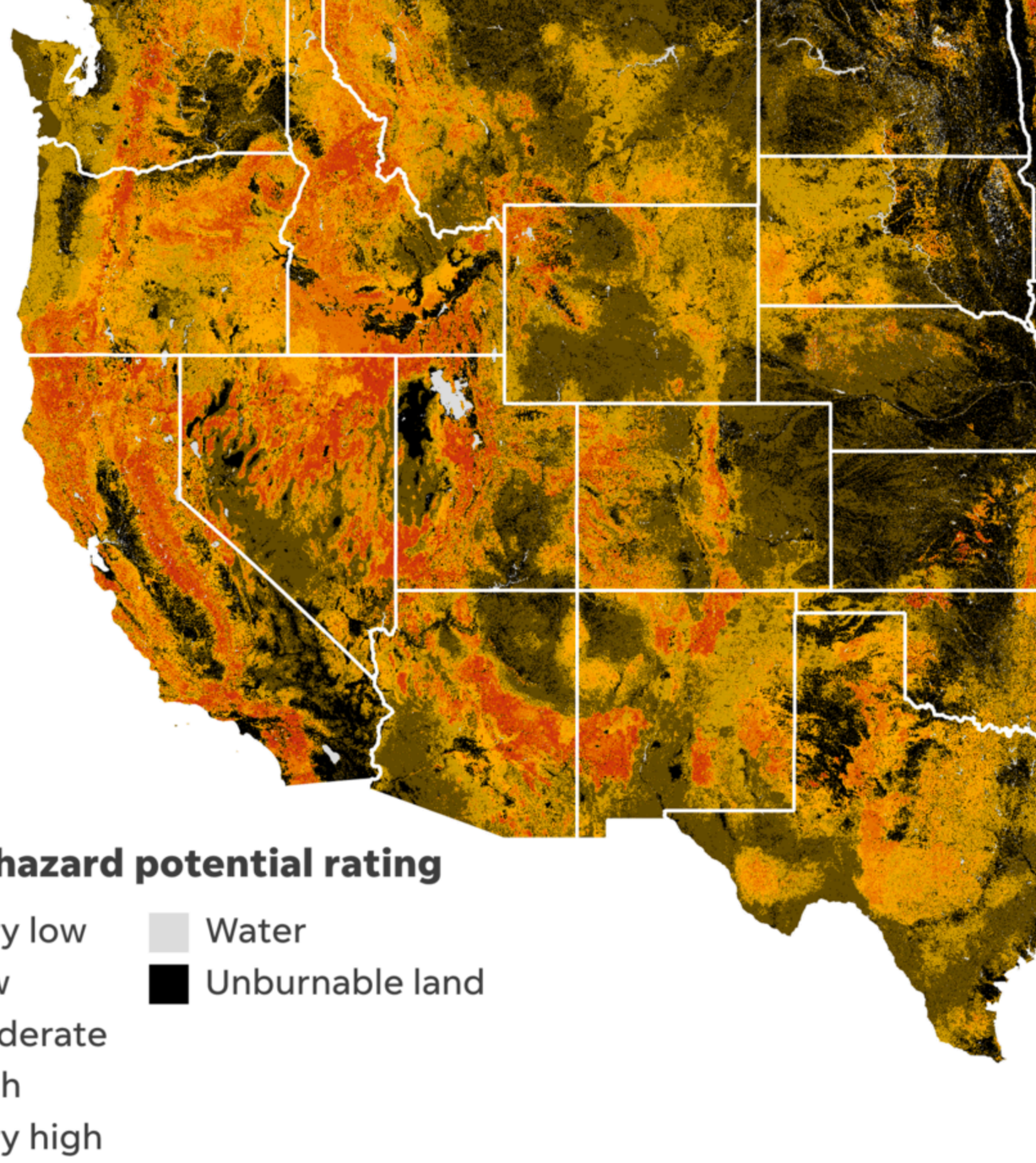
Hue to encode
nominal attribute

Ahead Of The Fire
By USA today

[link]

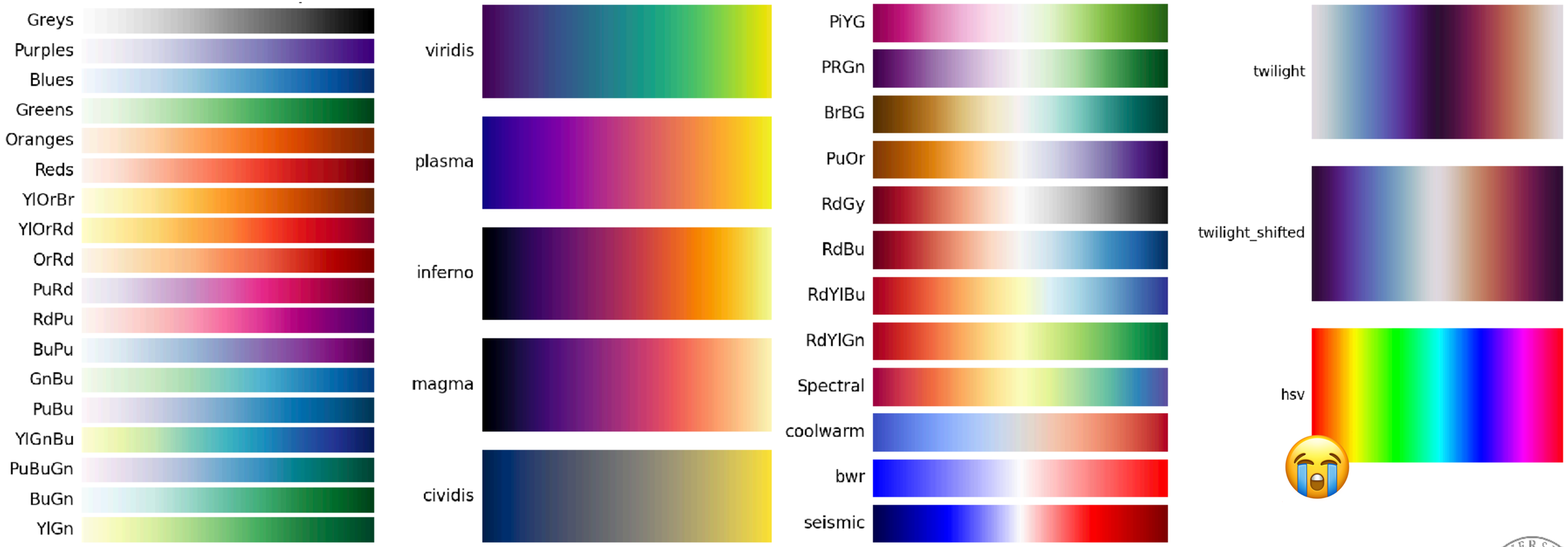


Sequential hue
(primarily) to encode
ordinal attribute

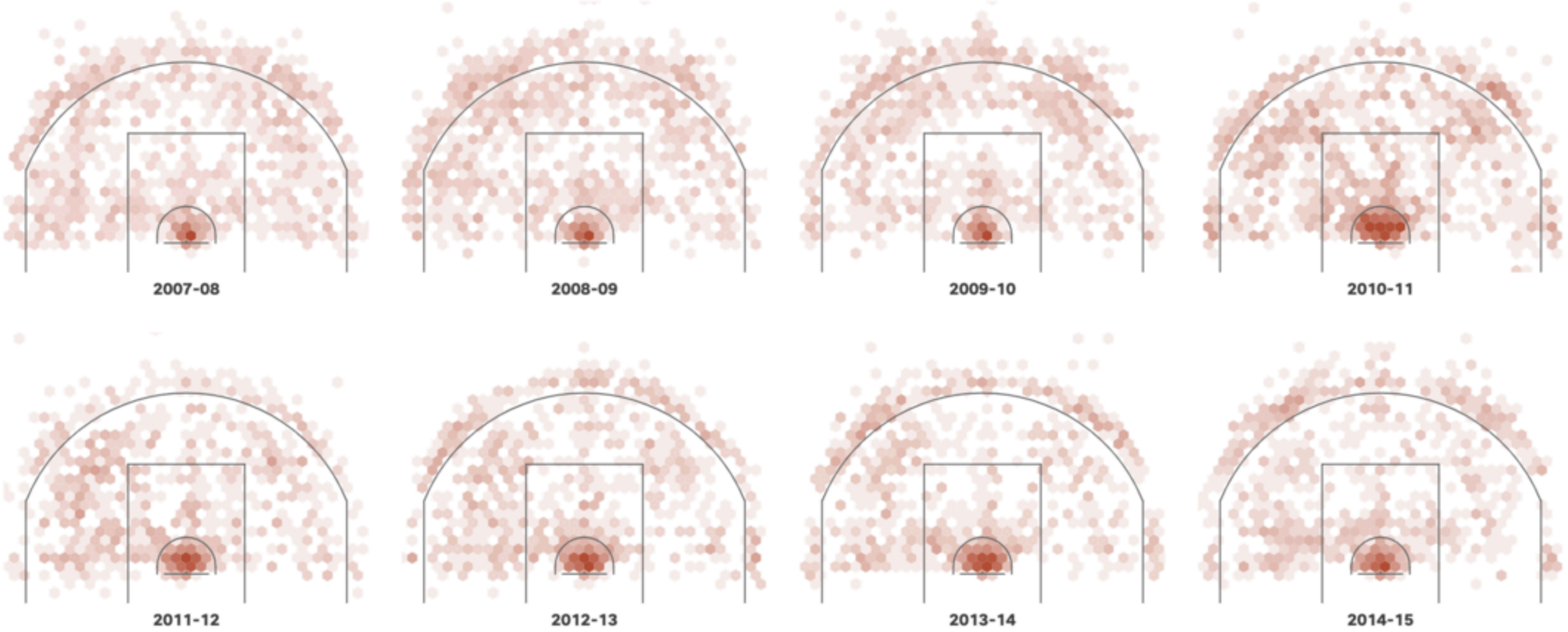


Quantitative Encoding

- Gradients for sequential, diverging, cyclic



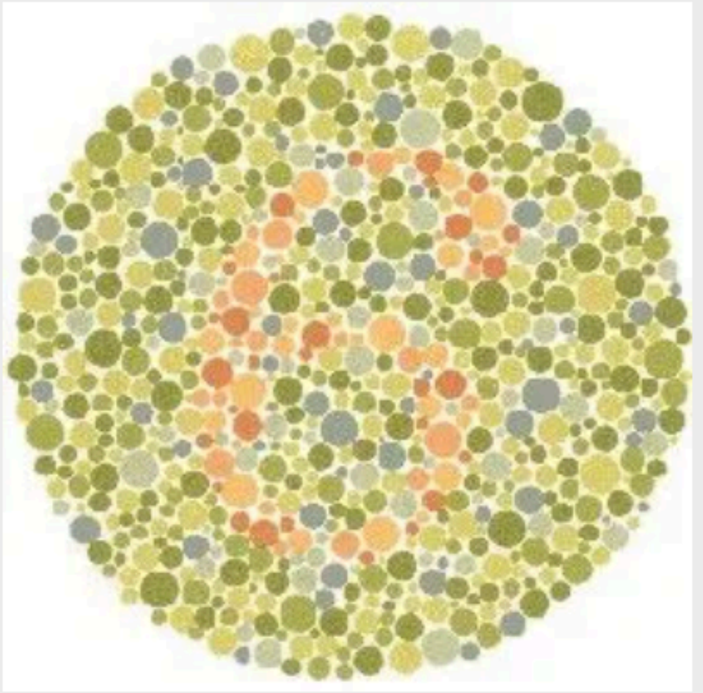
https://matplotlib.org/stable/users/explain/colors/colormaps.html#colormaps



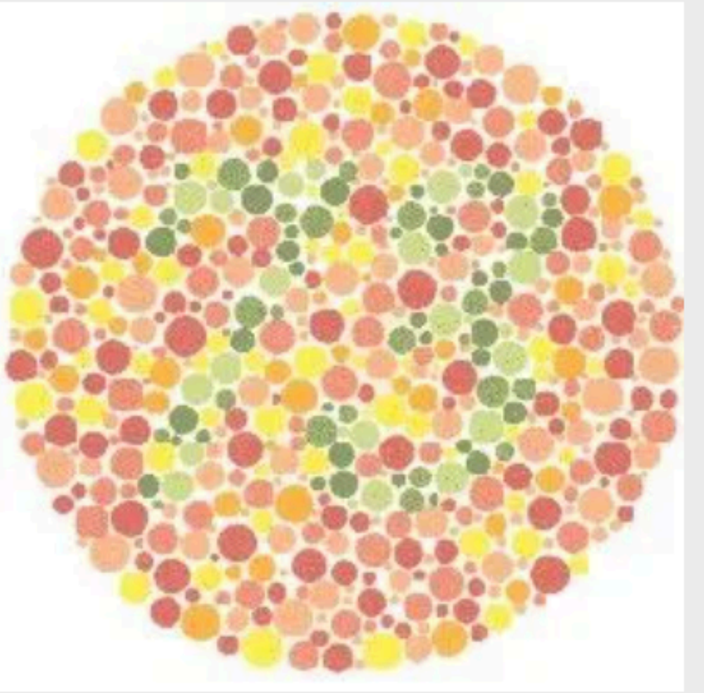
LeBron James has captured the scoring title. We visualized every shot.
by USA Today

link

Luminance/saturation to encode **numerical** (continuous) attribute



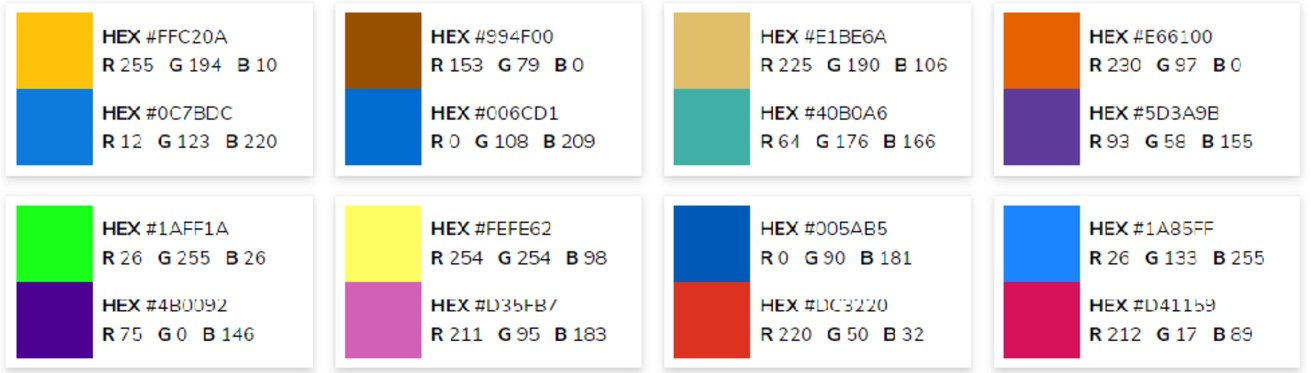
Enter The Number



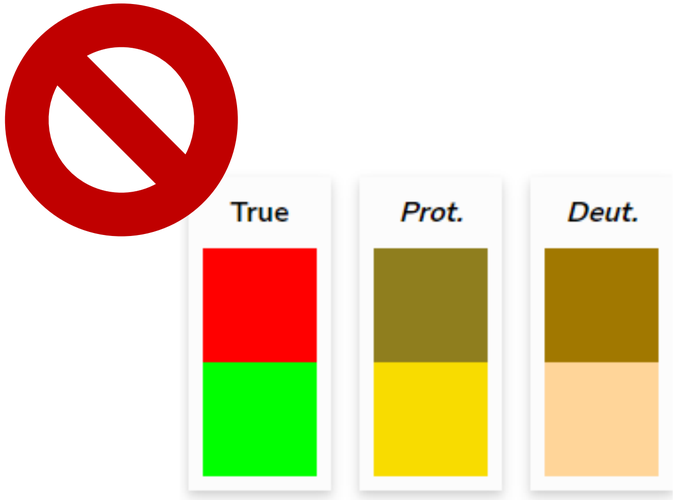
Enter The Number

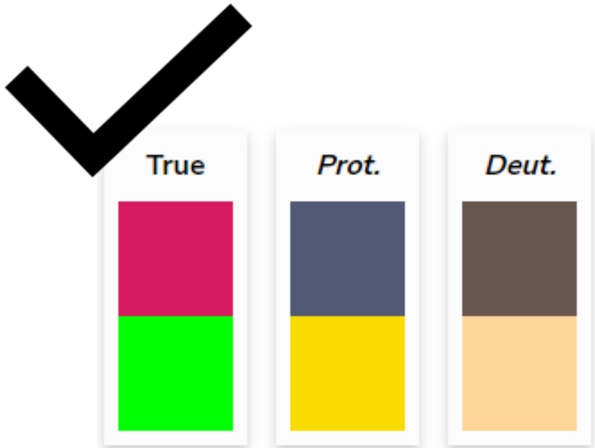
Accessible Colors

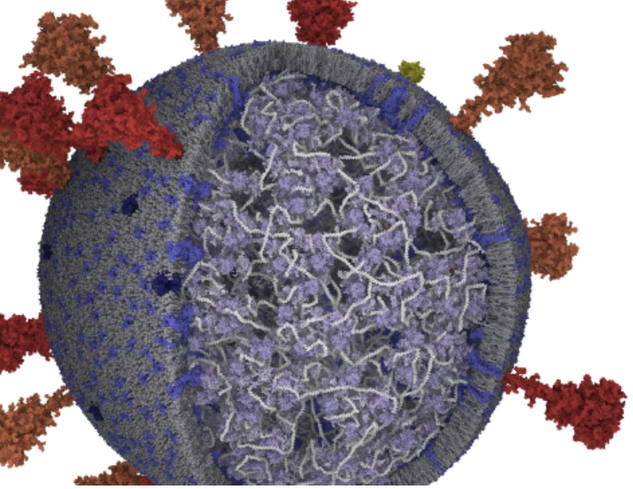
- Working color combinations (for colorblindness)

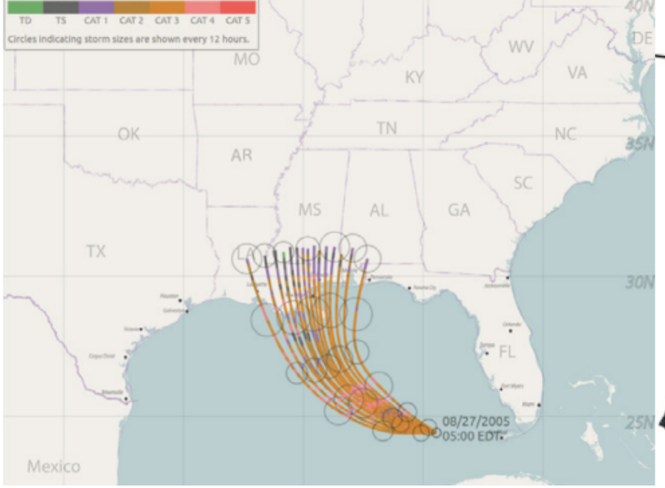


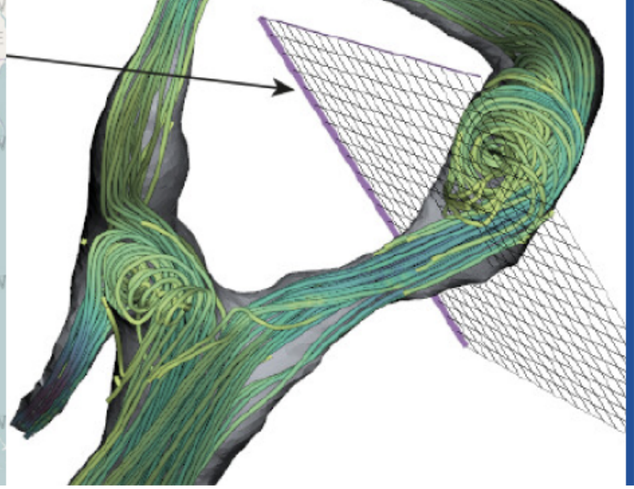
- Color combinations to avoid, e.g.:
  - Red-green
  - Purple-blue

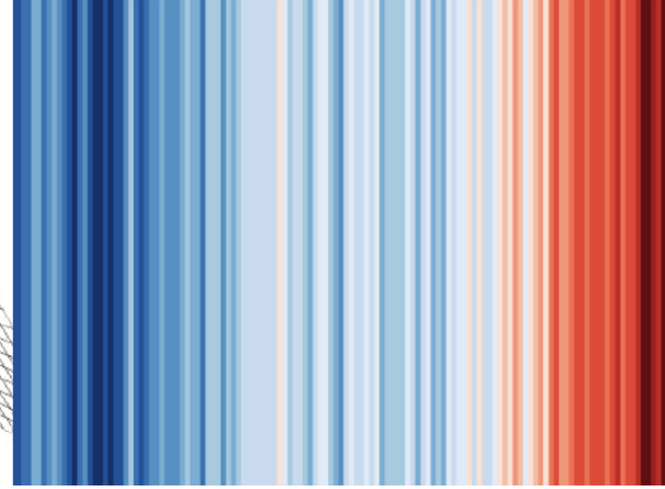






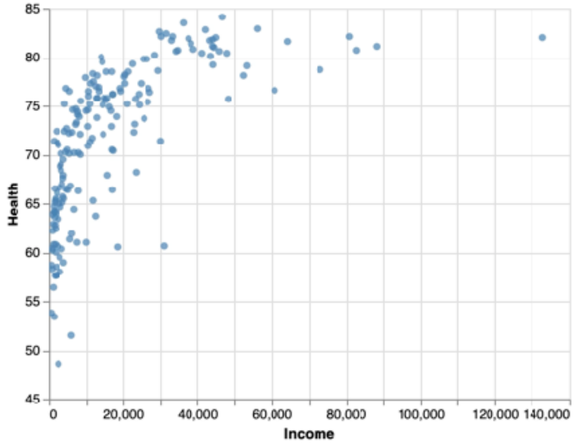


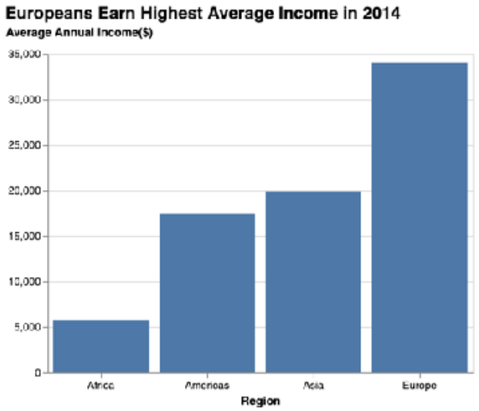


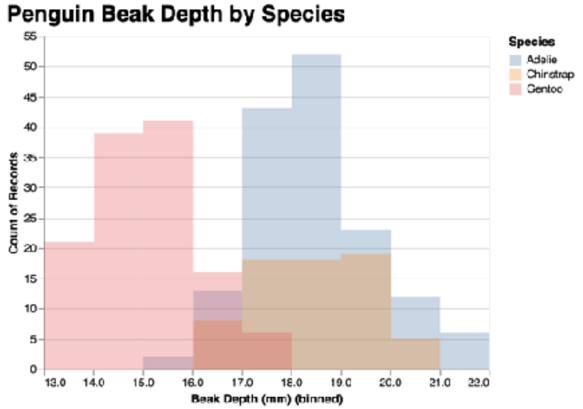


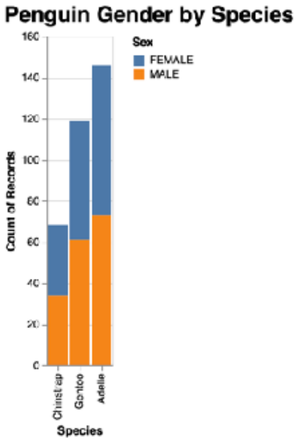
Applying visual grammar to tabular data

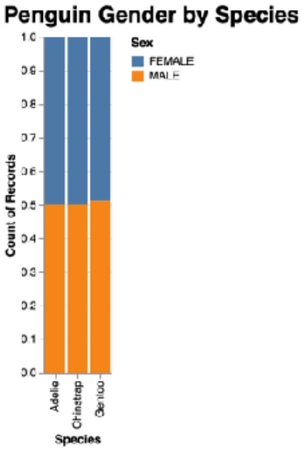
Different combinations of marks and channels lead to many possibilities…

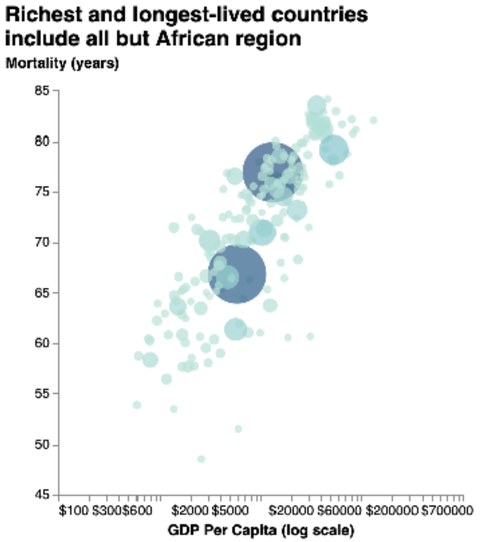


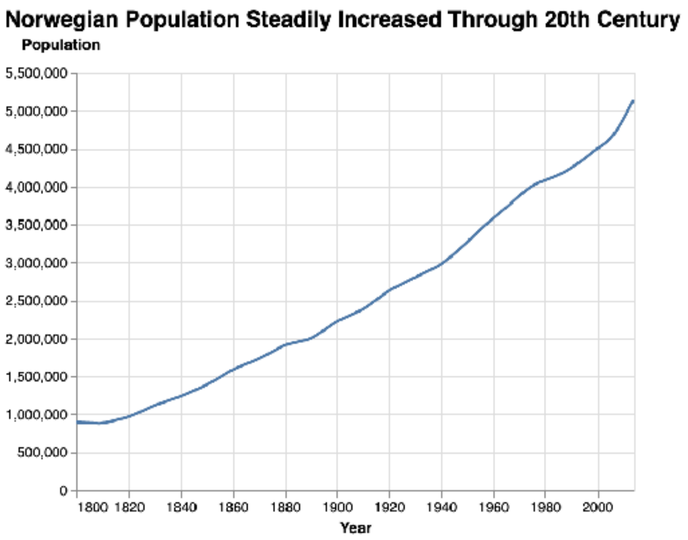


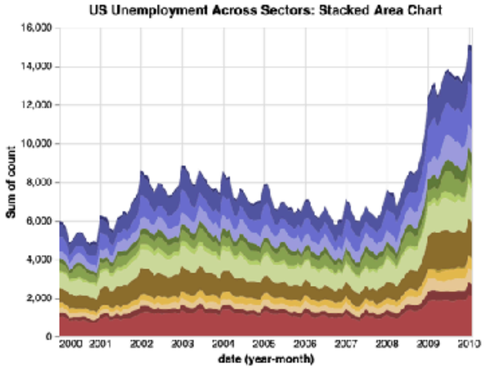


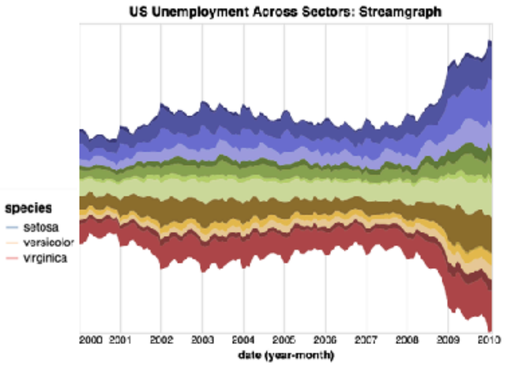


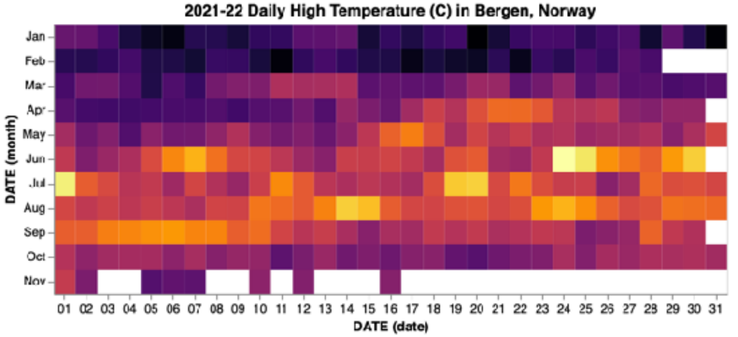


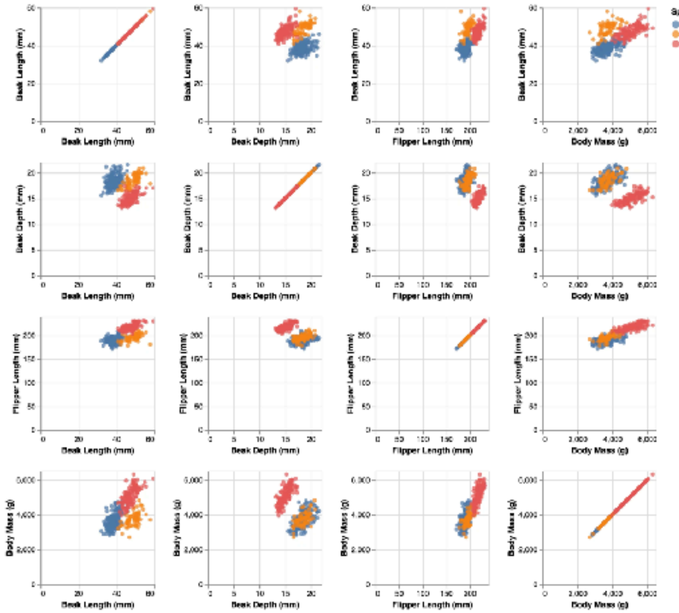


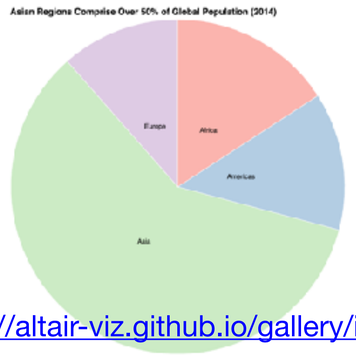


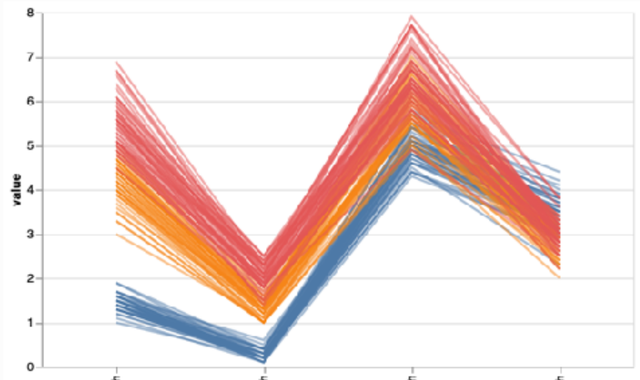




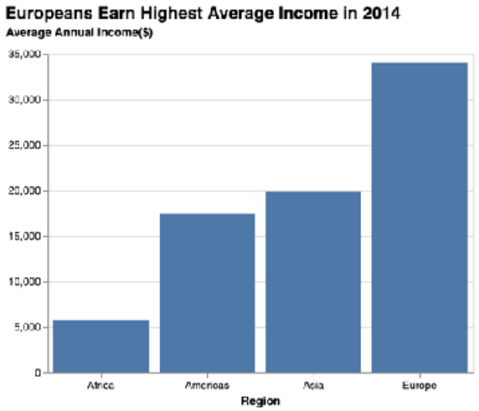


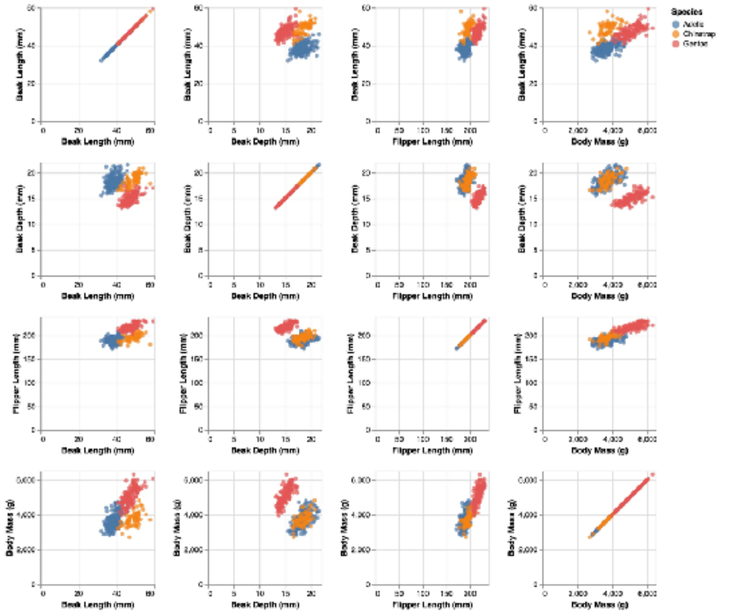






https://altair-viz.github.io/gallery/index.html

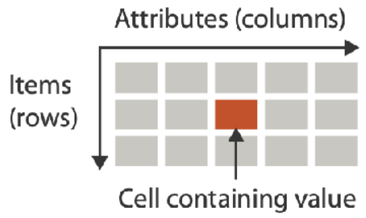


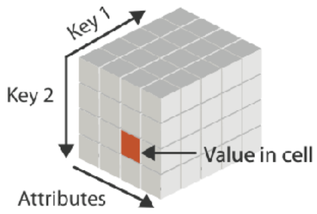


https://altair-viz.github.io/gallery/index.html

Recall: Tables

- Cells indexed by items and attributes
  - Rows = items
  - Columns = attributes
  - Value = (item, attribute)
- Flat or multi-dimensional

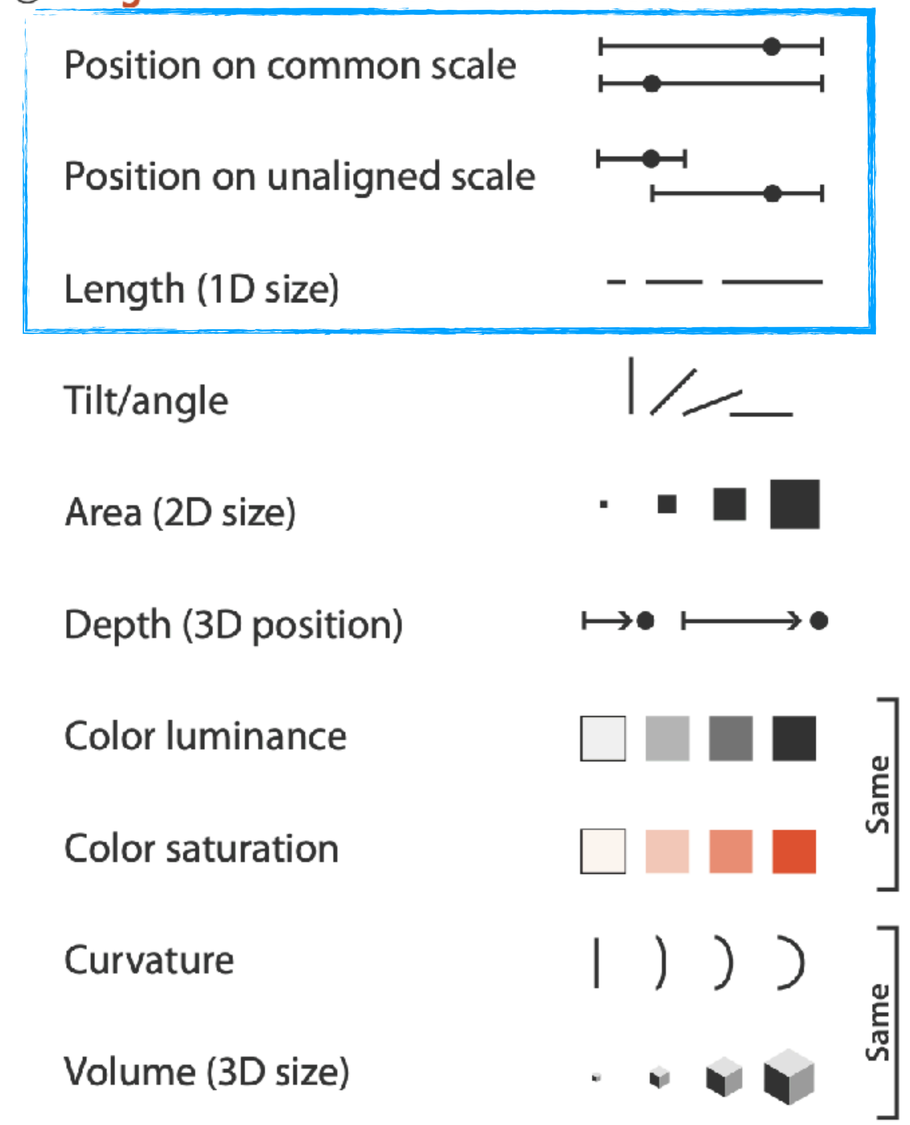




Key attribute (unique for each row)

Date	Activity Type	Title	Distance	Calories	Time	Avg HR	Max HR	Aerobic TE	Avg Run Cadence
2024-02-05 07:48:51	Running	Bergen Running	1.45	70	00:11:33	120	143	1.7	154
2024-02-04 12:18:25	Running	Bergen Running	6.50	377	00:48:43	157	184	3.9	162
2024-02-04 08:49:49	Running	Bergen - W03D7-Long Run	14.20	724	01:45:04	141	158	3.9	163
2024-02-02 16:28:39	Running	Bergen Running	4.34	247	00:34:04	149	165	3.0	155
2024-01-30 16:00:56	Running	Bergen - W02D7-Long Run	10.66	585	01:31:38	144	163	3.3	158
2024-01-26 17:36:22	Running	Bergen - W02D4-Threshold Run	4.26	195	00:25:23	144	177	3.2	169
2024-01-24 18:52:55	Running	Bergen - W02D2-Easy Run	3.85	228	00:30:27	154	184	3.2	157
2024-01-21 17:14:41	Running	Bergen - W01D7-Long Run	9.89	491	01:30:02	131	153	3.0	153
2024-01-16 15:36:08	Running	Bergen - W01D2-Easy Run	3.77	210	00:30:03	150	161	2.7	164
2023-12-31 07:08:20	Running	Dewitt - Base	9.74	527	01:06:03	153	169	4.1	170
2023-12-29 08:35:36	Running	Dewitt - Sprint	7.13	361	00:46:35	152	178	3.7	161
2023-12-27 06:46:57	Running	Dewitt - Base	5.25	279	00:34:28	150	184	3.6	171
2023-12-24 09:43:31	Running	Dewitt - Base	9.29	502	00:59:05	160	178	4.8	172
2023-12-22 07:25:44	Running	Dewitt - Base	5.34	283	00:34:06	152	169	3.8	171
2023-12-21 10:34:43	Running	Dewitt - Base	4.92	255	00:30:17	158	180	3.9	173
2023-12-17 10:09:35	Running	Bergen - Base	7.71	403	00:49:20	155	183	4.4	171
2023-12-10 10:11:34	Running	Bergen - Base	7.83	433	00:49:38	165	193	5.0	173
2023-12-09 11:07:49	Running	Bergen - Recovery	4.57	255	00:31:08	157	173	3.6	171
2023-11-05 08:28:47	Running	Bergen - Base	7.36	383	00:46:04	152	170	4.1	174
2023-11-01 17:38:58	Running	Bergen - Recovery	3.87	197	00:26:48	141	158	3.0	168

Magnitude Channels: Ordered Attributes



Identity Channels: Categorical Attributes

Arranging tabular data

- **Express Values**

- **Separate, Order, Align Regions**
  - Separate
  - Order
  - Align

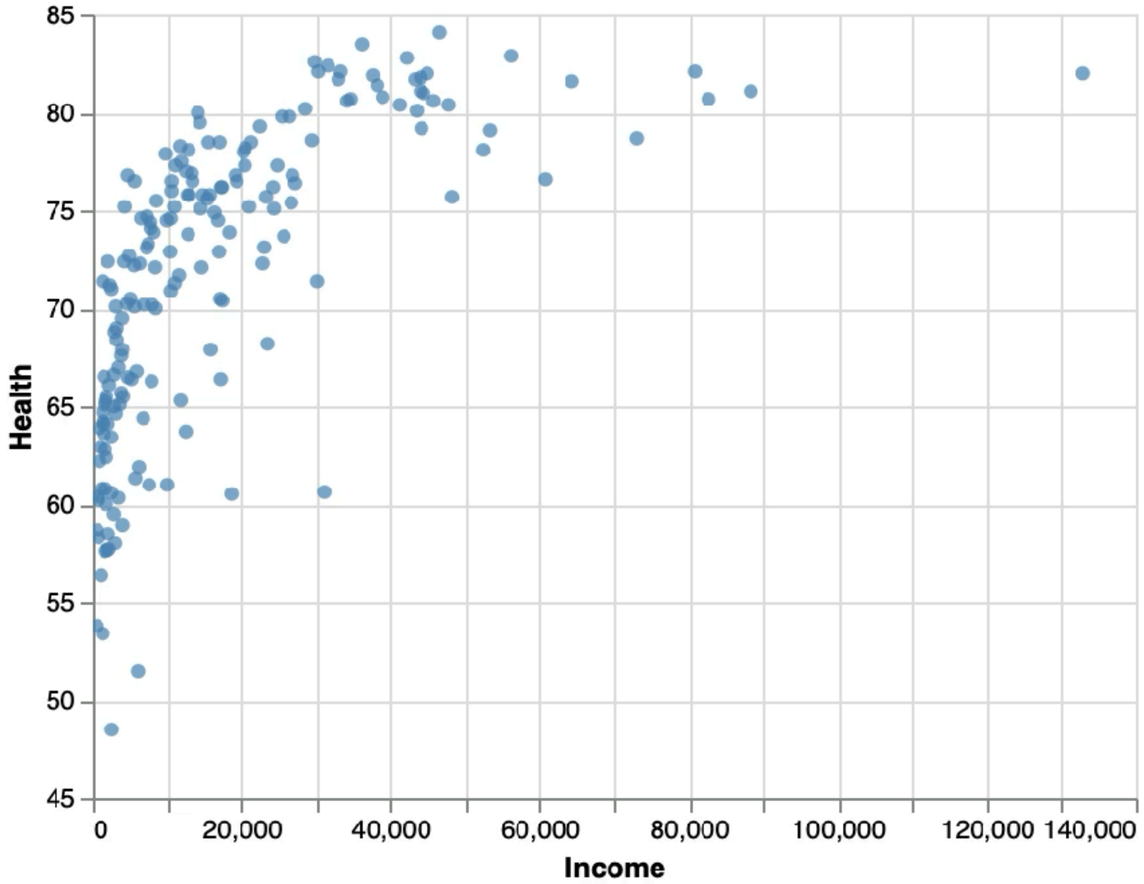
- **Axis Orientation**
  - Rectilinear
  - Parallel
  - Radial

- **Layout Density**
  - Dense
  - Space-Filling

Express values

Encode with no keys, only values!

Scatterplot



Data Source: GapMinder

Data: 2 quantitative values

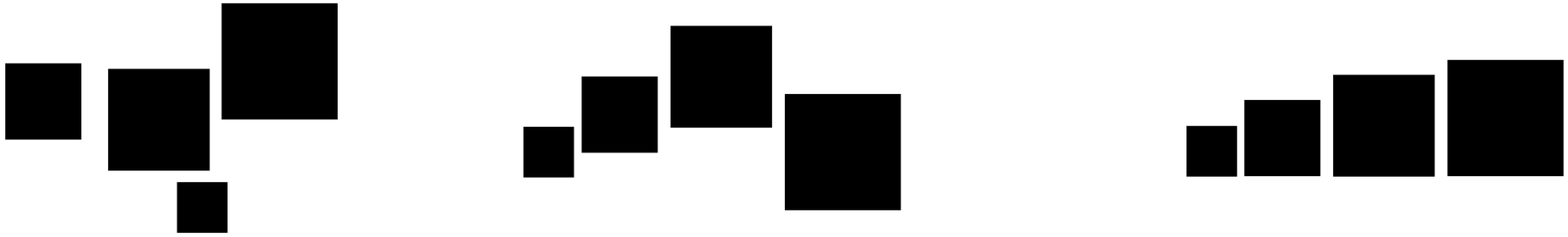
Mark: circles or points

Encoding Channels:

- horizontal position (X-axis)
- vertical position (Y-axis)

Tasks: overview, see distribution, find correlations, identify outliers

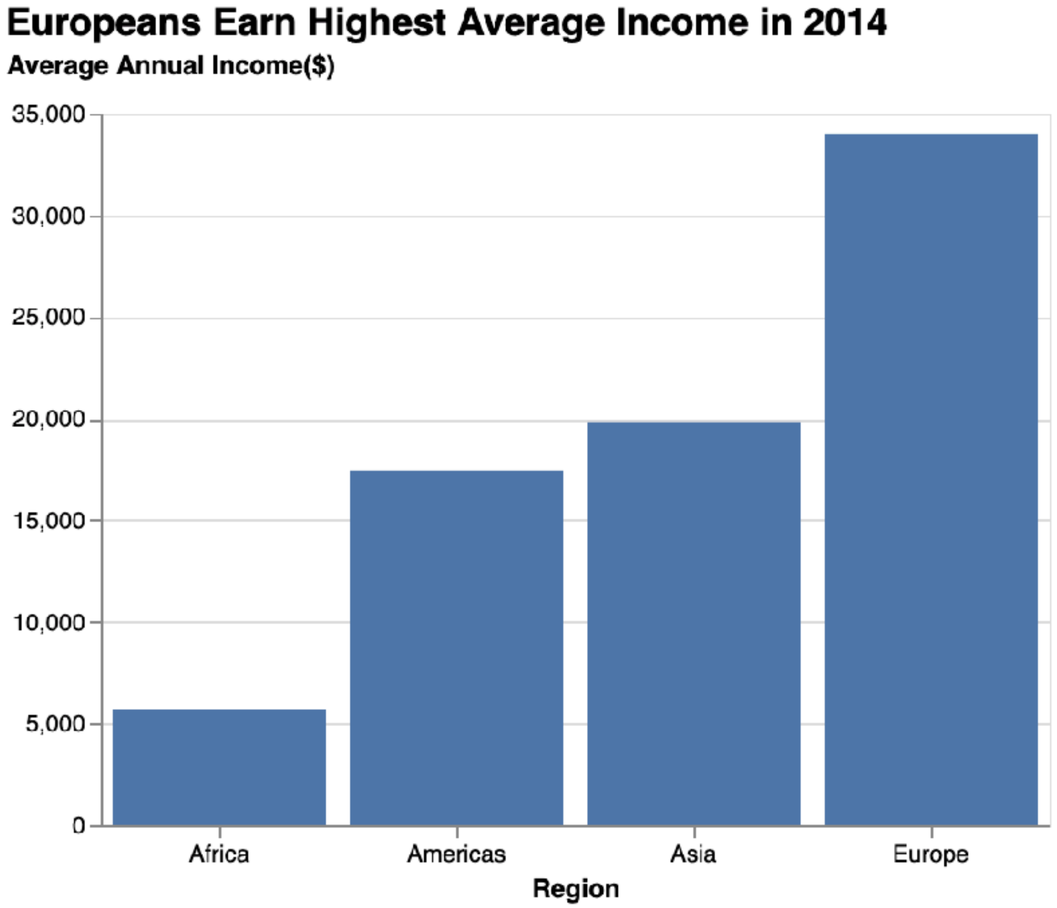
T. Munzner, *Visualization Analysis and Design (CRC Press, 2014), Chp 7*



Separate, Order, and Align

Encode with one attribute as key

Bar Chart



Data: 1 categorical attribute, 1 quantitative value

Mark: line (or bar)

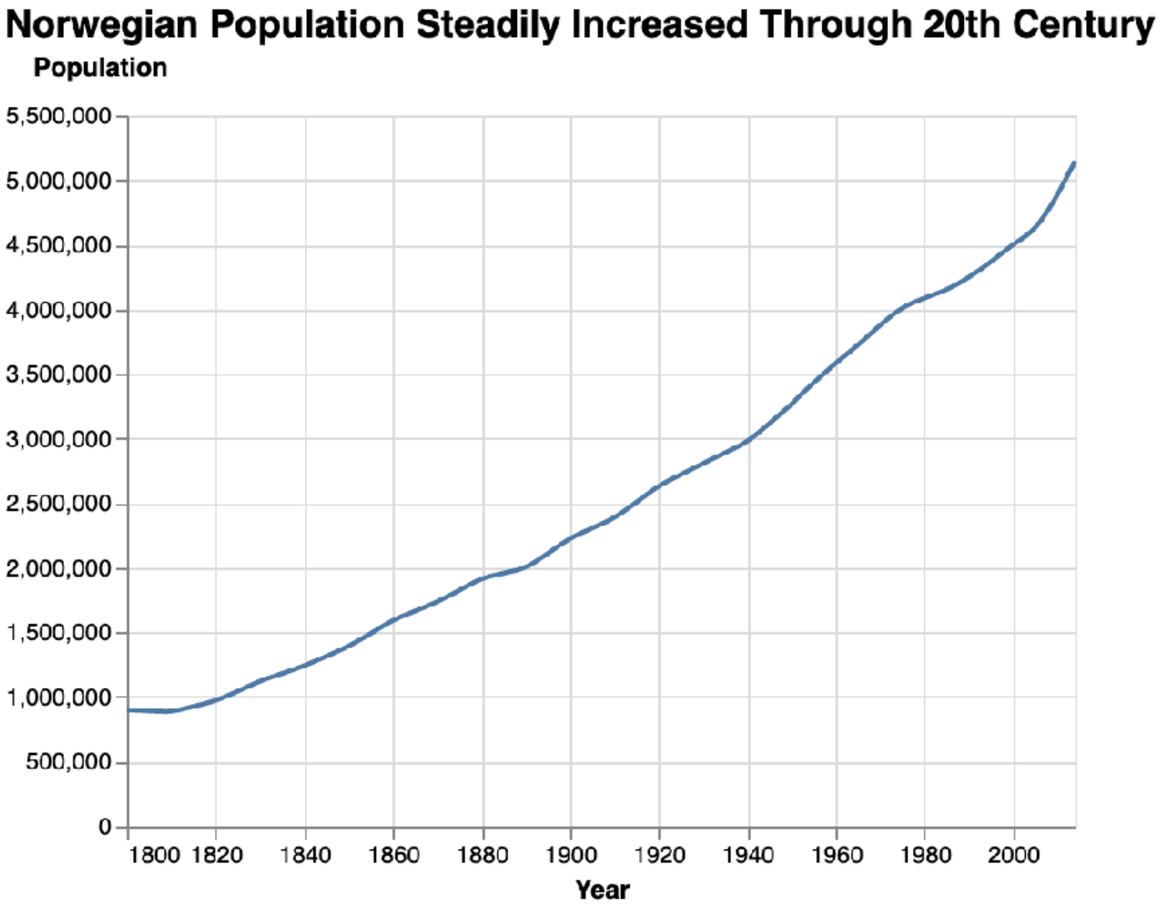
Encoding Channels:

- **length** to express quantity (count)
- **spatial region** per mark
  (horizontal separation, vertical alignment)
- **ordering** by attribute

Tasks: Compare magnitudes, lookup values

Data Source: GapMinder

Line Chart



Data Source: GapMinder

Data: 1 ordered attribute, 1 quantitative value

Mark: points connected by lines
(lines show relationship between items)

Channels:

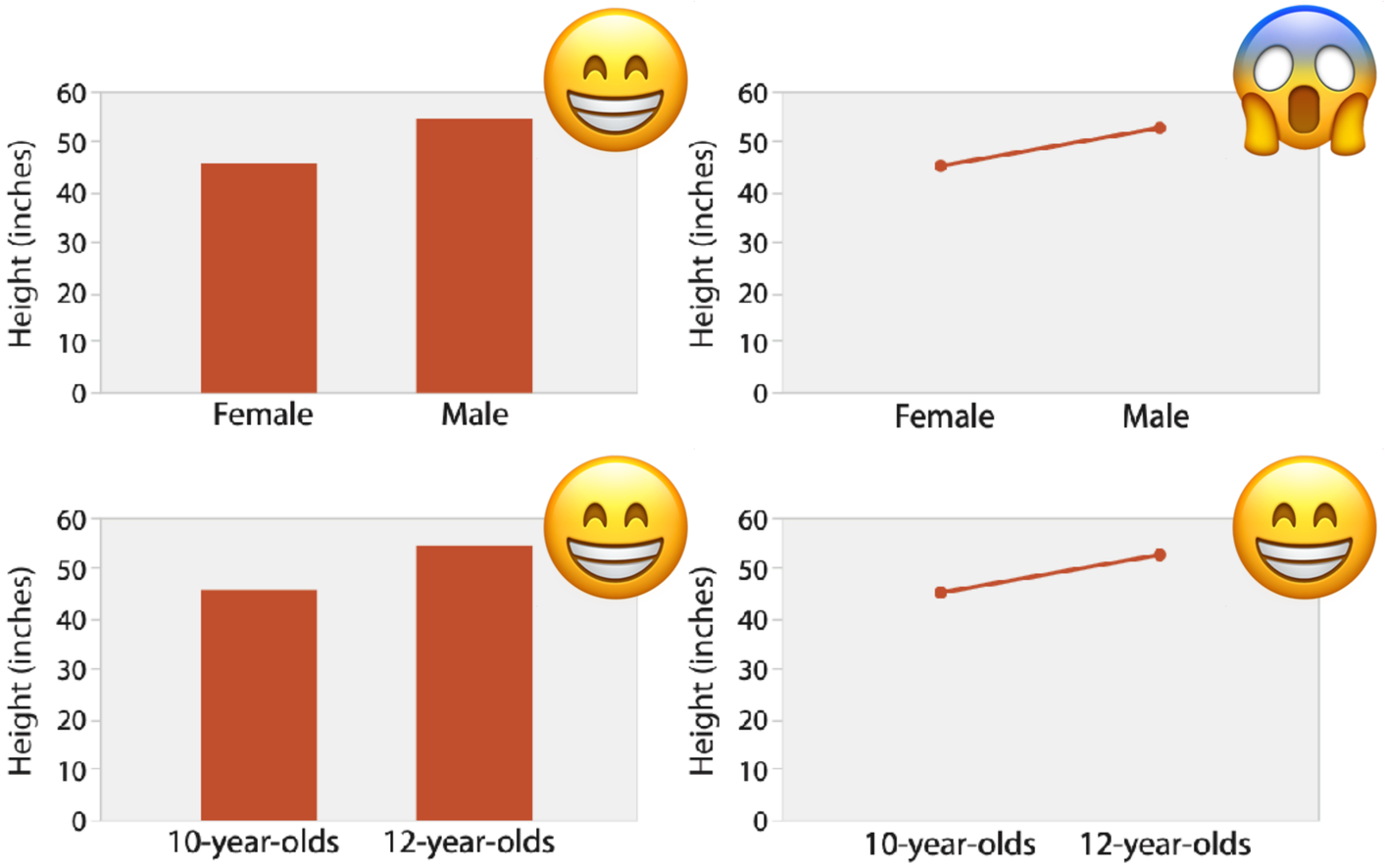
- horizontal position (X-axis)
- vertical position (Y-axis)

Task: find trends (relationship from one item to next)

T. Munzner, *Visualization Analysis and Design (CRC Press, 2014), Chp 7*

Should I use a Bar or Line Chart?

- Depends on your attribute
  - bar charts if **categorical attribute**
  - line charts if **ordered or quantitative attribute**



after [Bars and Lines: A Study of Graphic Communication. Zacks and Tversky. Memory and Cognition 27:6 (1999), 1073–1079.]