Reminder: Course participation is 14% of your grade.

Communicating data science through visualization

CSE481DS Data Science Capstone

Tim Althoff

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Due next week

- Midpoint presentation video
  - See template on website under deliverables
  - 10 min 0 sec max.
- Think of this as a draft of your final project presentation but without major results.
  - We expect that you have completed ca. 50% of the project.
  - We would like to see your data and some initial results
  - We are asking you to discuss two related papers
  - Provide a complete picture of your project even if certain key parts have not yet been implemented/analyzed/solved.
- We grade based on the quality, as well as the completion of sections described in the template.
- Reminder: Now is a good time to start planning for your final report writing as well.
  - Midpoint includes briefly highlighting two similar research papers. Start early!
- Reminder: Office hours are a great way to get early feedback and support!

Agenda

1. Visualization in data science
2. Human perception
3. Storytelling with data
4. Visualization design
5. Break + Prototyping
6. *Visualization for papers*
7. *Bad visualization*
8. *Visualization tools and resources*
9. Visualization Lab

Acknowledgements

Contents of this lecture are generously borrowed from:

- UW CSE 512 Data Visualization course slides by Jeff Heer and guest lecturers (Matt Conlen, Michael Correll)
- Tutorial by Marinka Zitnik from Harvard University
- CSE481DS materials by Jina Suh

Visualization in Data Science

What is the role of visualization in data science?

What is data science

Data contains value and knowledge

Data science **extracts knowledge from data**, seeks to discover new knowledge by answering question through data

What is visualization?

Transformation of the symbolic into the geometric

- McCormick et al. 1987

The use of computer-supported, interactive, visual representations of abstract data to **amplify cognition**

- Card, Mackinlay, and Shneiderman 1999

What does visualization do?

Graphics **reveal data**. Indeed graphics can be more precise and revealing than conventional statistical computations.

- Tufte 1983

One great virtue of good graphical representation is that it can serve to **display clearly and effectively a message** carried by quantities whose calculation or observation is far from simple.

- Tukey and Wilk 1965

Superpower of visualization

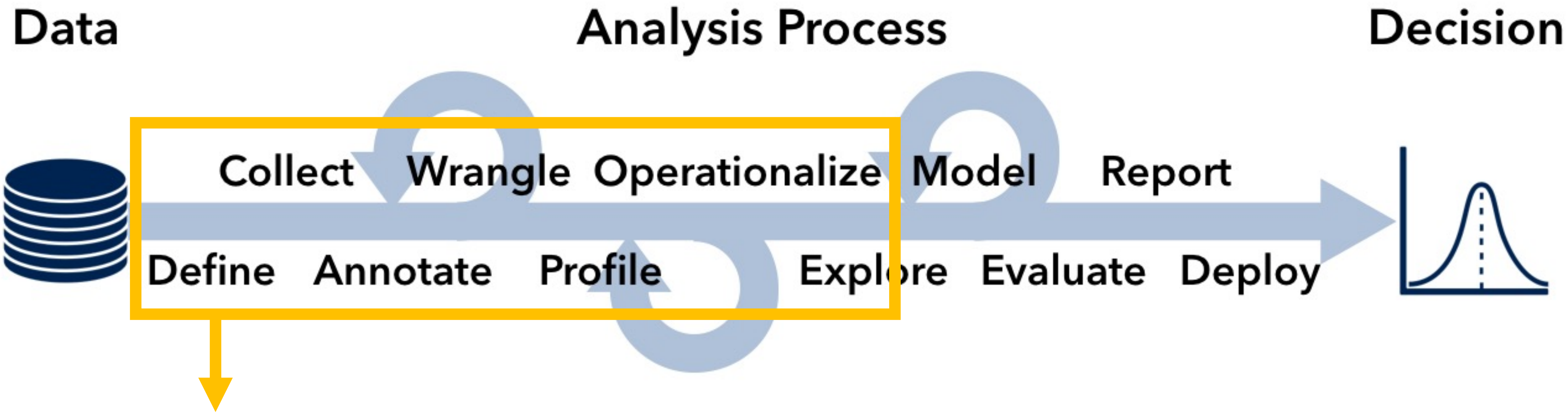
When applied effectively to promote data exploration, analysis, and insight, we will experience what Joseph Berkson called "**interocular traumatic impact**: a conclusion that hits us between the eyes."

- Cleveland 1993

Empower understanding of data and analysis processes

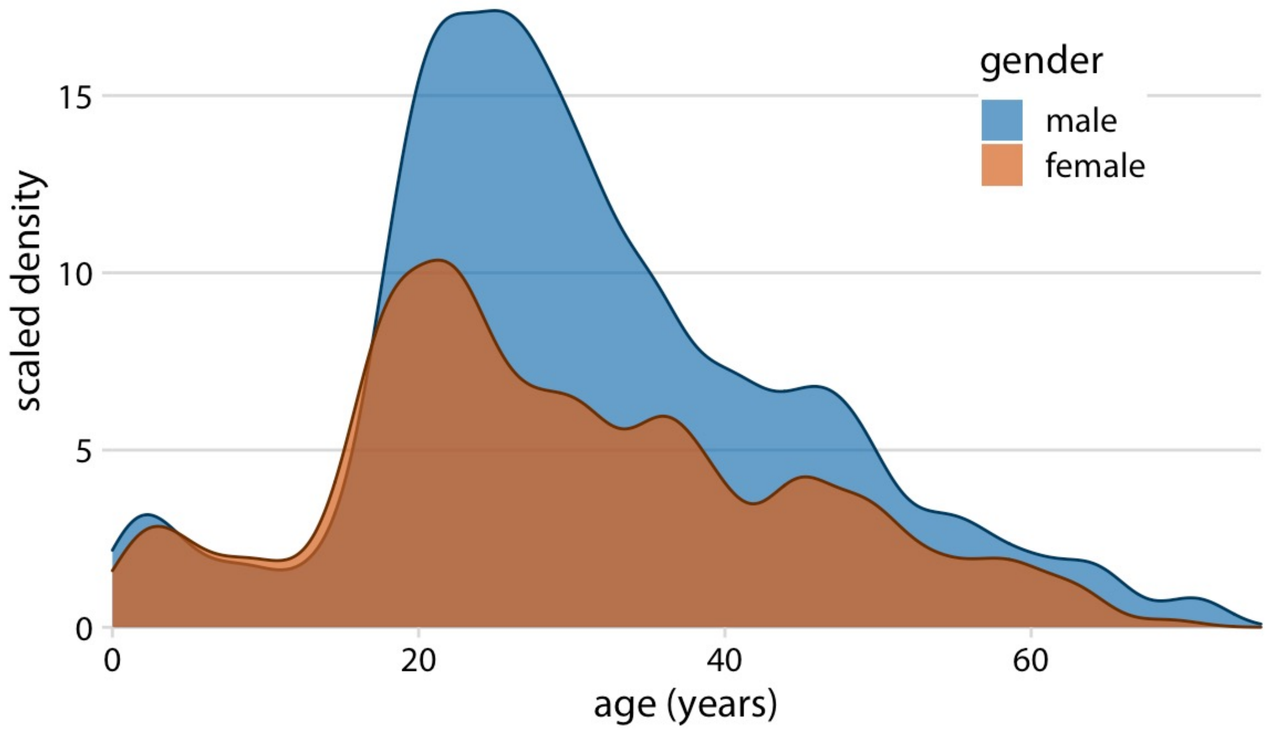
Visualization in data analysis process

Visualization in data analysis process



Stage 1:
Understanding data quality and research task at hand

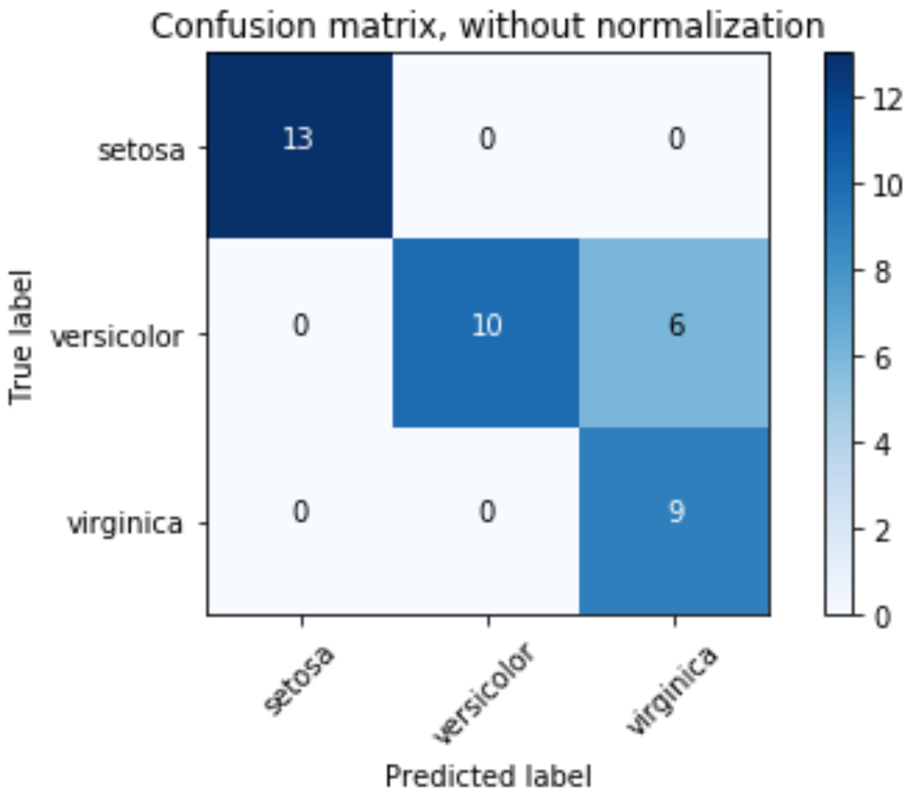
Collect: Do I have the right population?



https://clauswilke.com/dataviz/histograms-density-plots.html

- Less female than male
- Females are younger

Annotate: Are there disagreements?



https://medium.com/@rakeshrajpurohit/confusion-matrix-469248ed0397

- 84% accuracy (32/38)
- All errors isolated in versicolor

Wrangle

I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I'm lucky if I get to do any "analysis" at all.

- Anonymous Data Scientist

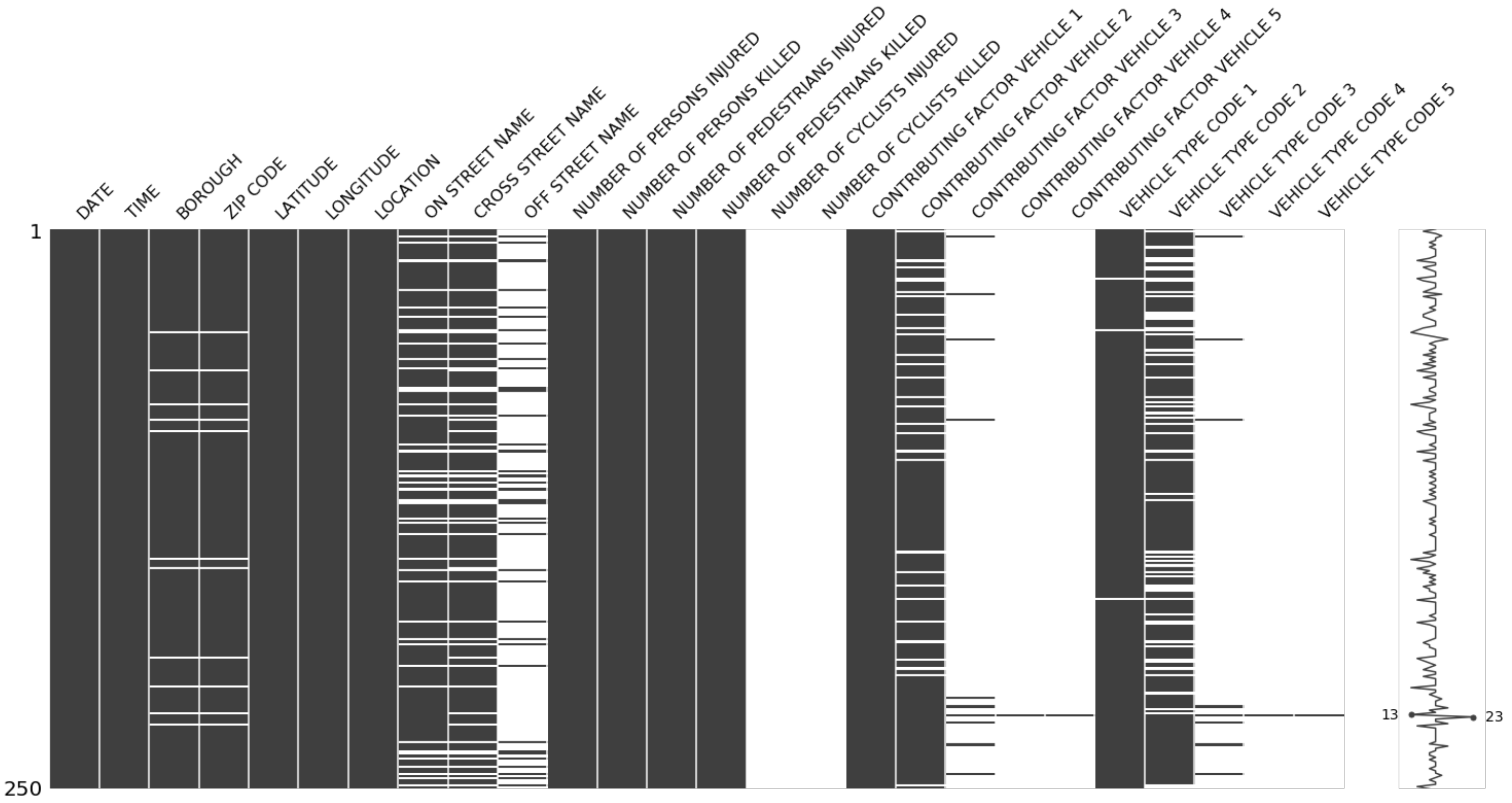
But wait… Visualizations can be my **superpower**

Wrangle

The first sign that a visualization is good is that it **shows you a problem in your data**… …every successful visualization that I've been involved with has had this stage where you realize, "Oh my God, this data is not what I thought it would be!" So already, you've discovered something.

- Martin Wattenberg

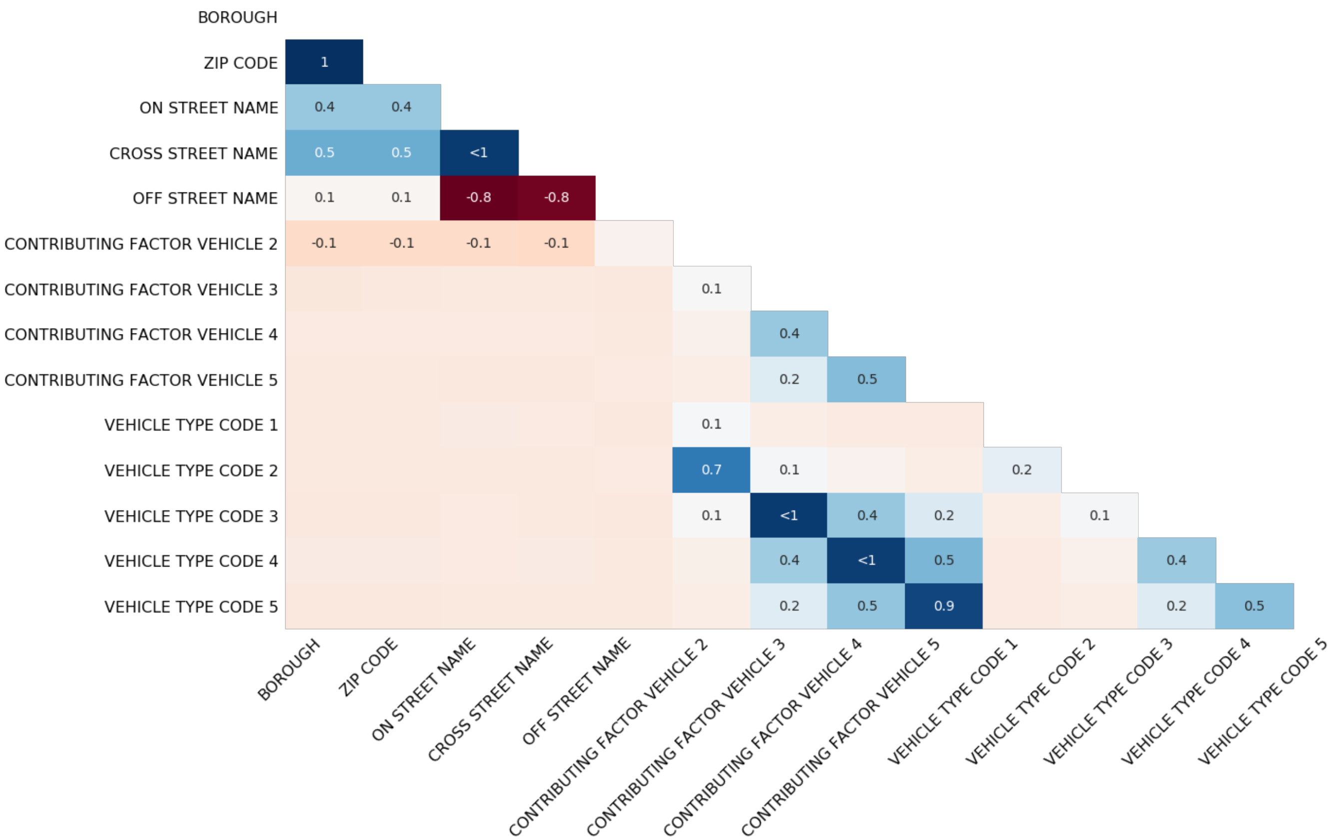
Wrangle: How messy is this dataset?



https://github.com/ResidentMario/missingno

- What feature can I live without?

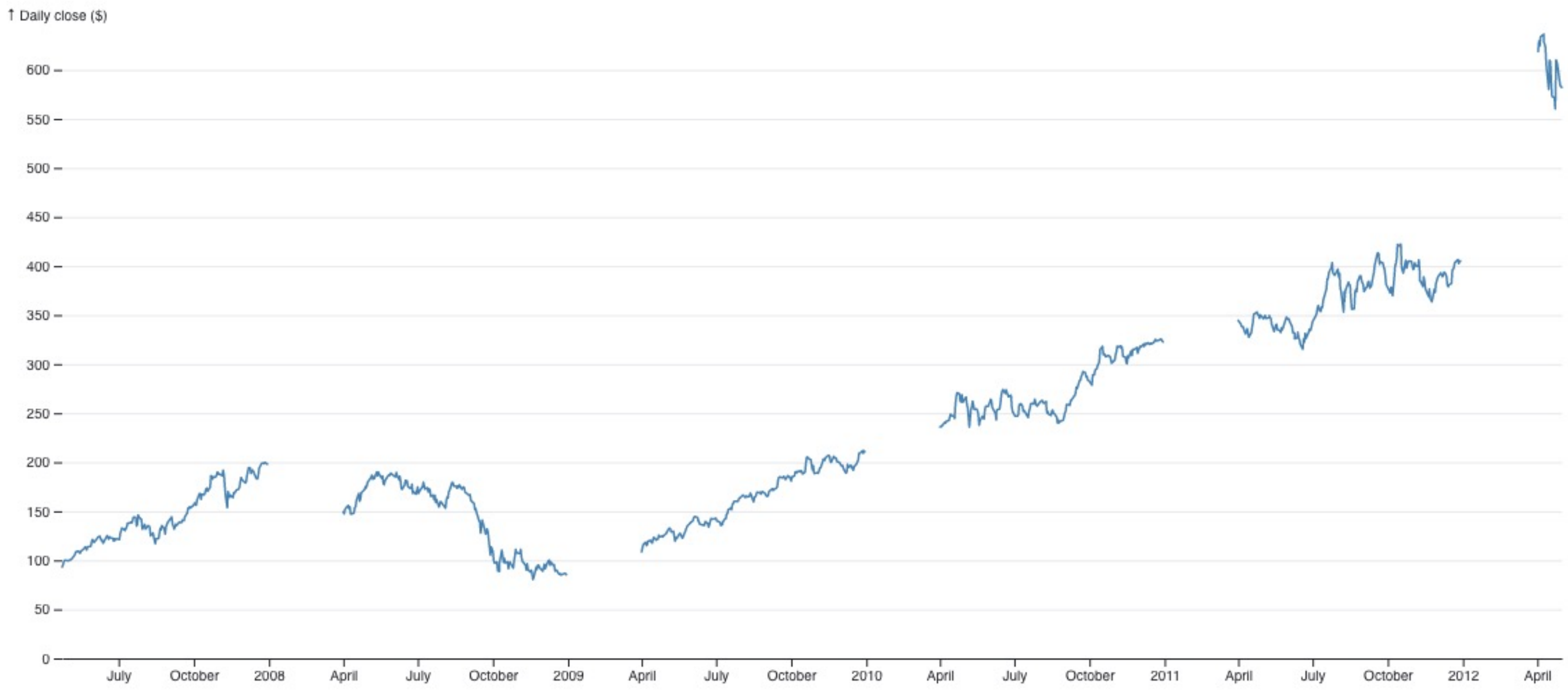
Wrangle: How messy is this dataset?



https://github.com/ResidentMario/missingno

- Which pairs can I live without?

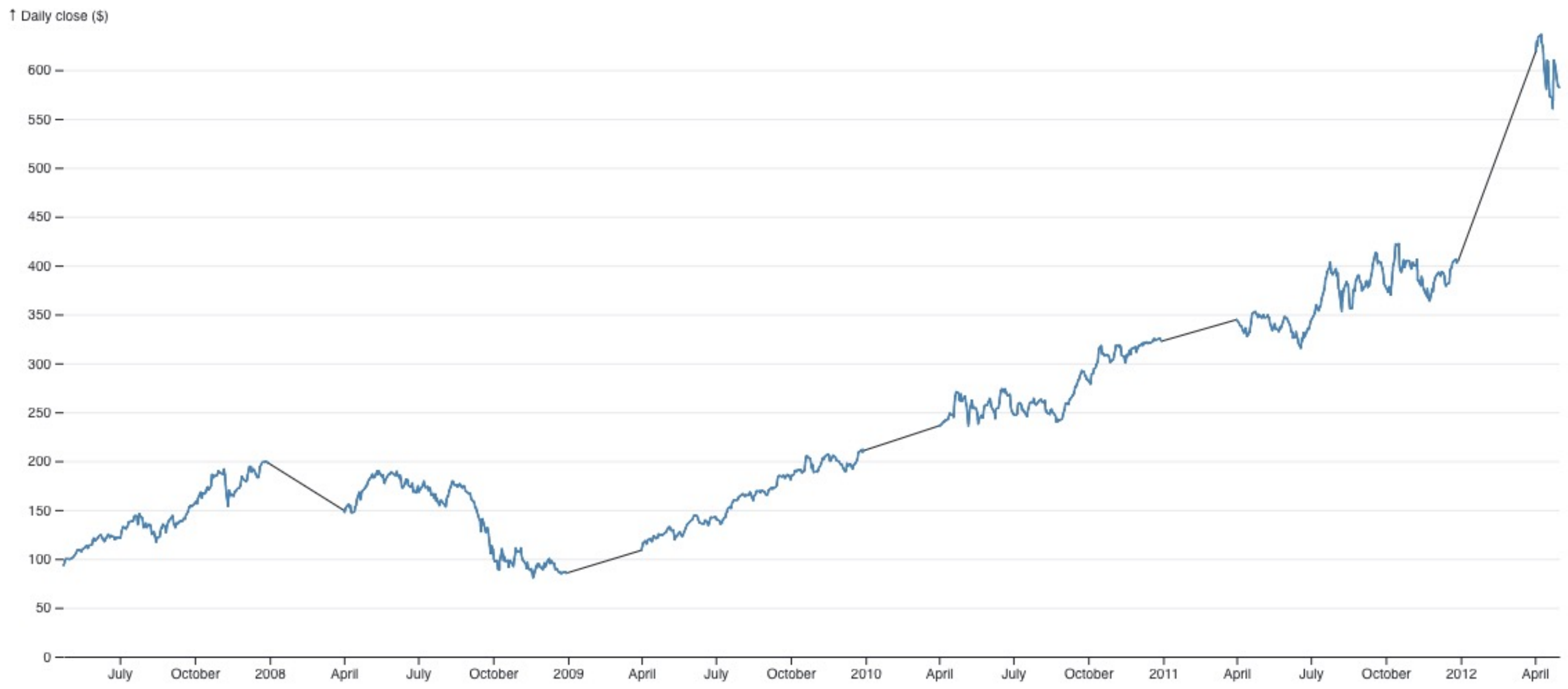
Wrangle: Do I impute or not?



https://observablehq.com/@d3/line-with-missing-data

- To impute or not to impute, that is the question

Wrangle: Do I impute or not?



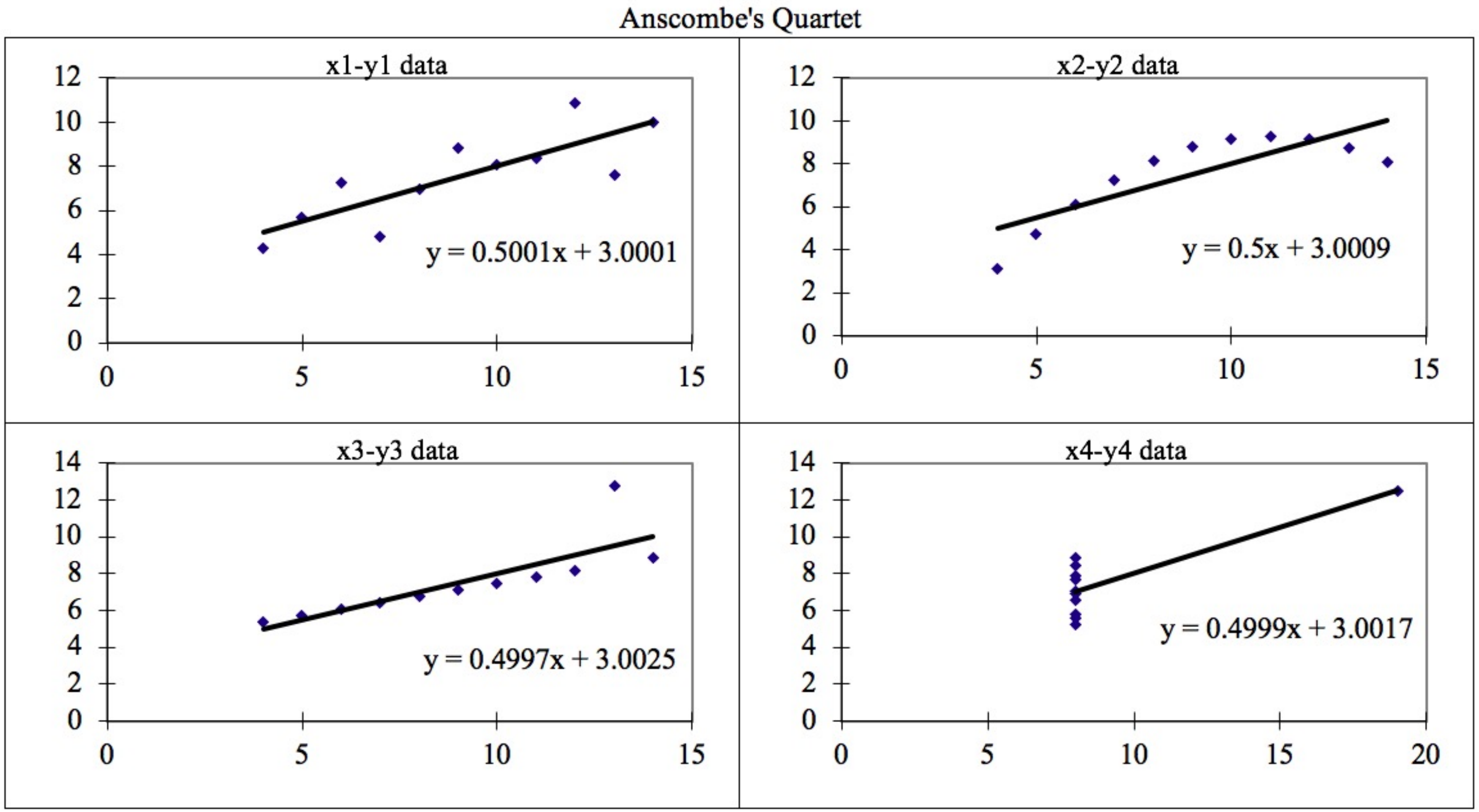
https://observablehq.com/@d3/line-with-missing-data

- To impute or not to impute, that is the question

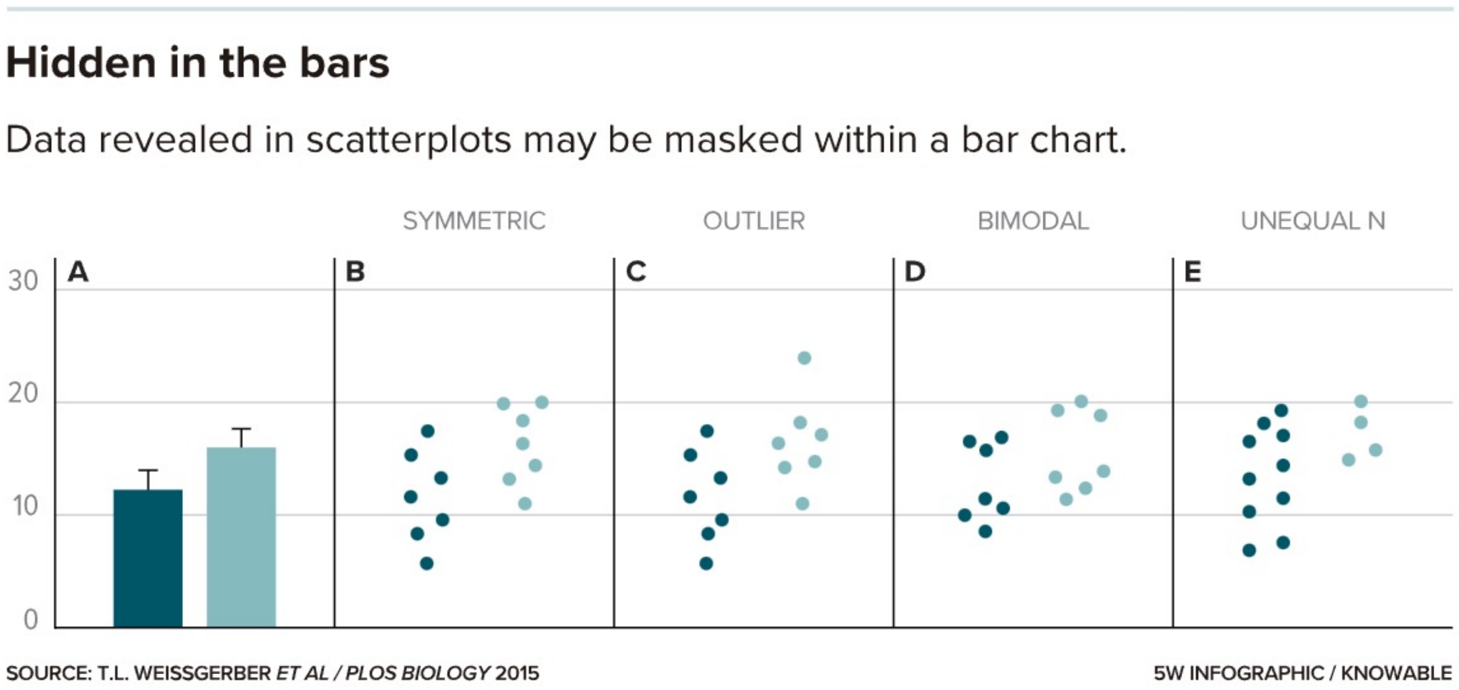
Profile: How is my data distributed?

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Profile: How is my data distributed?



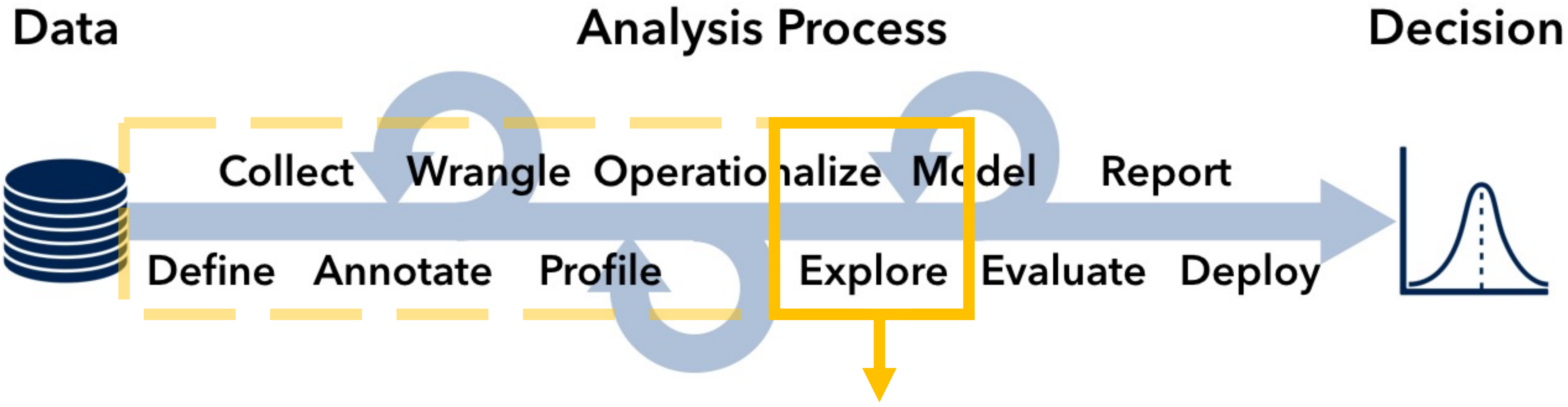
Profile: How is my data distributed?



- Check my assumptions

https://knowablemagazine.org/article/mind/2019/science-data-visualization

Visualization in data analysis process



Stage 2:
Exploring the data

Explore

Open-ended ⟷ Specific

Data quality
Univariate summaries
Check assumptions
Distributions

Relationships among variables
Correlations
Breakdowns

Checking different models
Hypothesis testing

Visual exploration process

Pick a question

Construct visualizations

Inspect the answer

Identify new questions

Repeat

Visual analysis journal

Write down your question

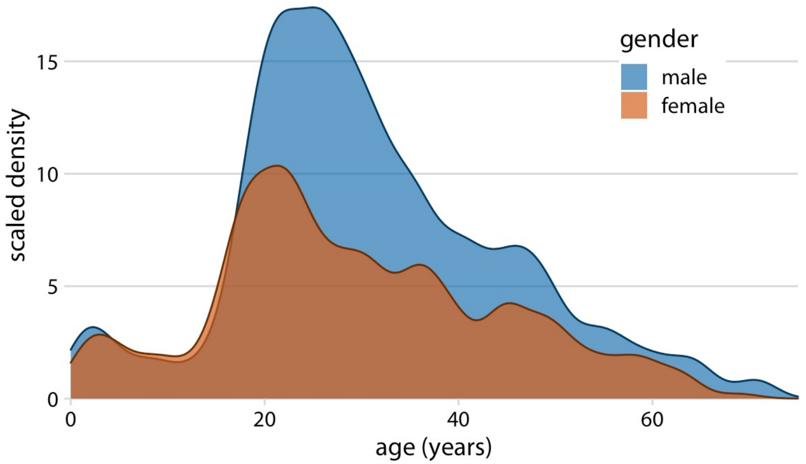
Generate the visualization

Summarize your insight

Identify next steps or question

Document the how

Are genders balanced in my dataset?



Nope. There are less female than male. Females are slightly younger.

Need to collect more female data.

Visual exploration tips

Avoid premature fixation on perfection!

It's expected that exploratory data visualizations are not perfect.

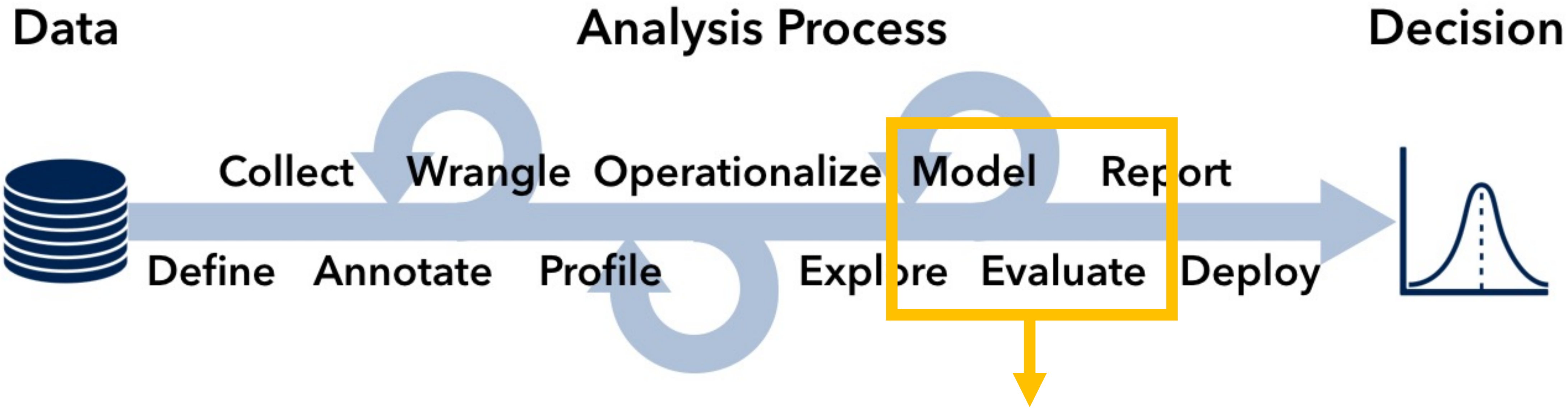
Show data variation, not design variation

Your viz may not be perfect, but does it do a decent job?

Iterate quickly

Choose the right tool for the right job

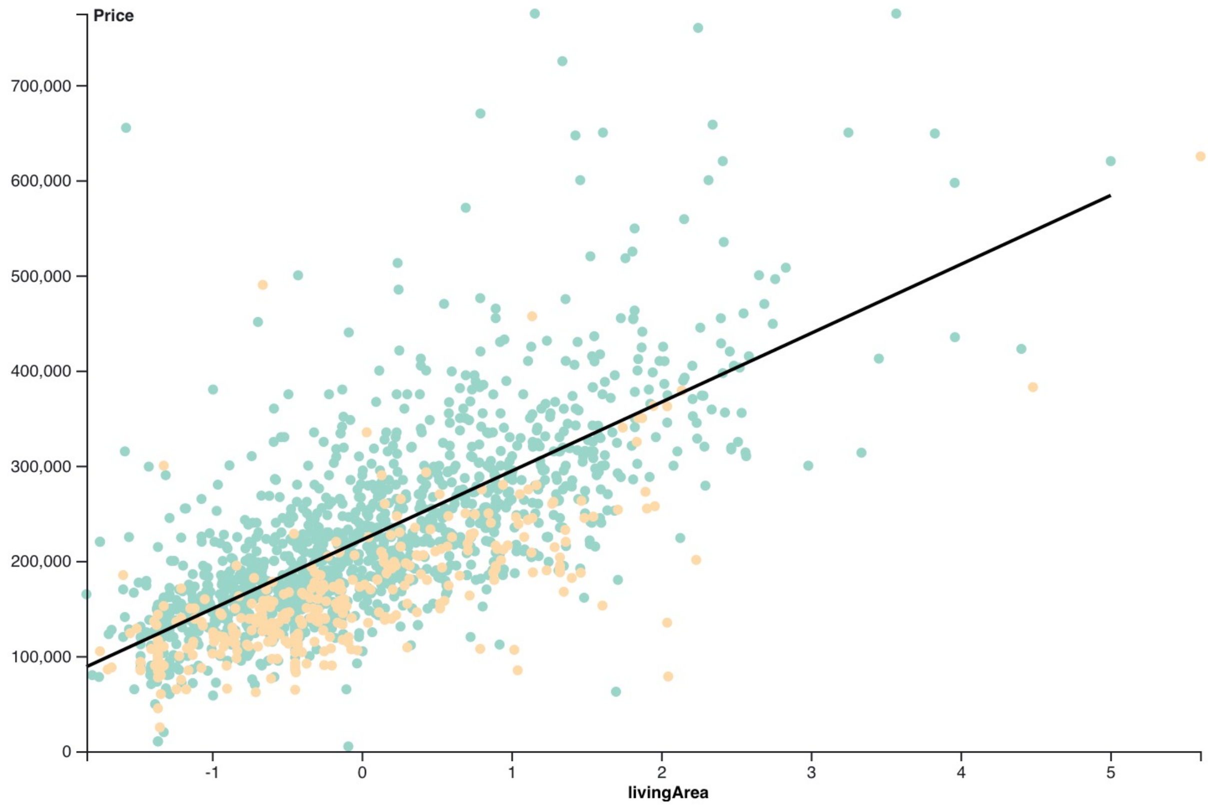
Visualization in data analysis process



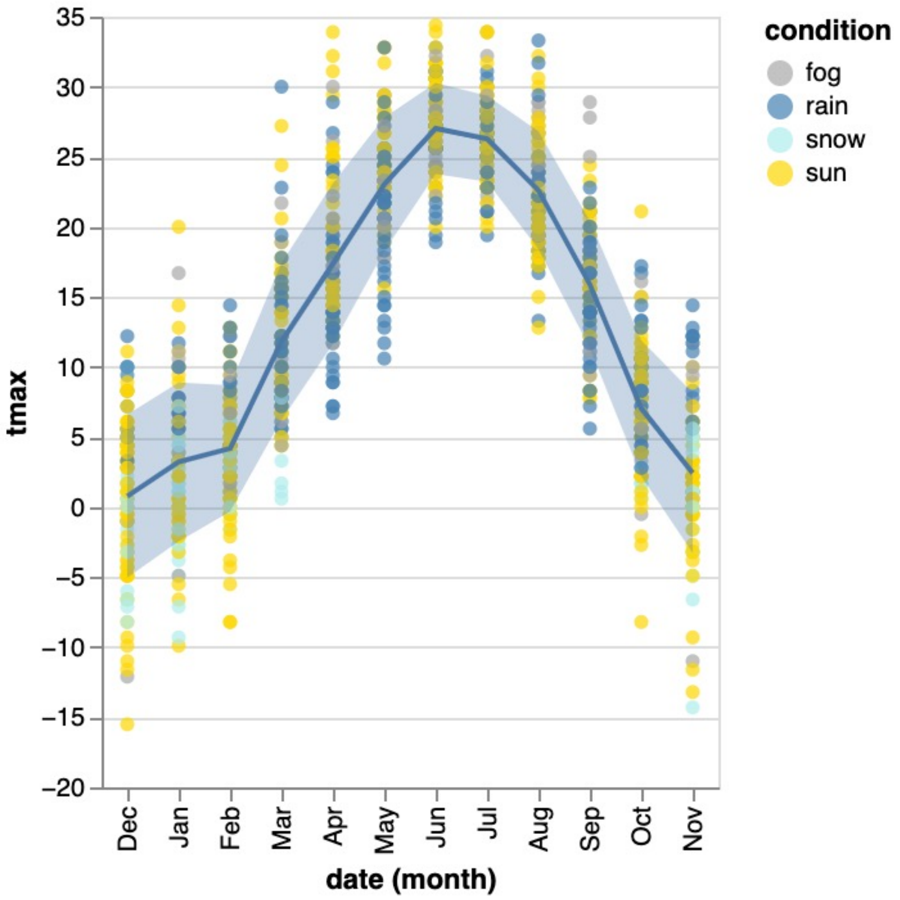
Stage 3:

Modeling and evaluating relationships and insights

Model

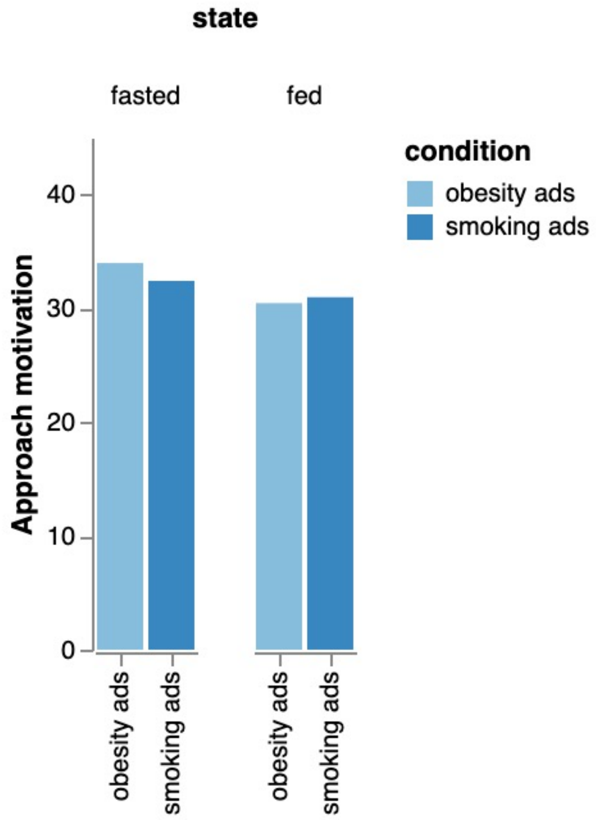


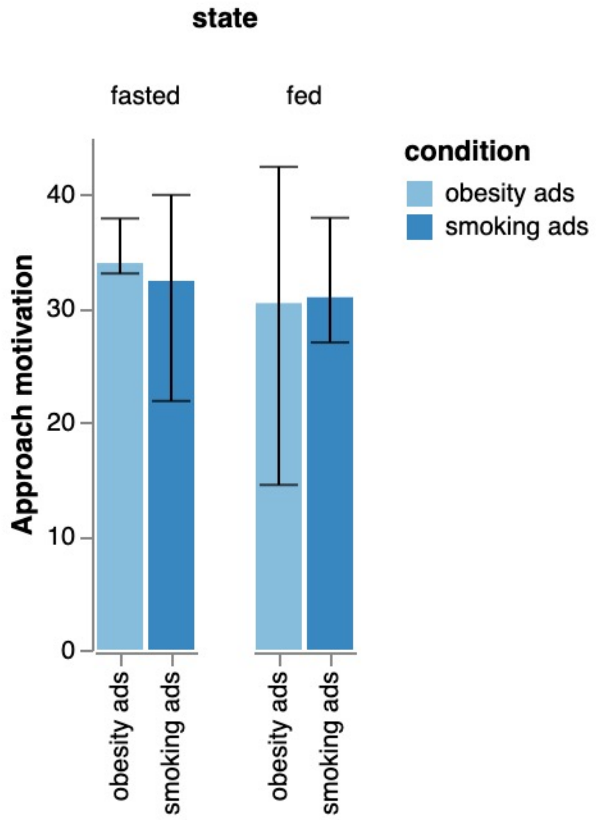
Price
700,000
600,000
500,000
400,000
300,000
200,000
100,000
0
-1
0
1
2
3
4
5
livingArea



35
30
25
20
15
10
5
0
−5
−10
−15
−20
tmax
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
date (month)
condition
fog
rain
snow
sun

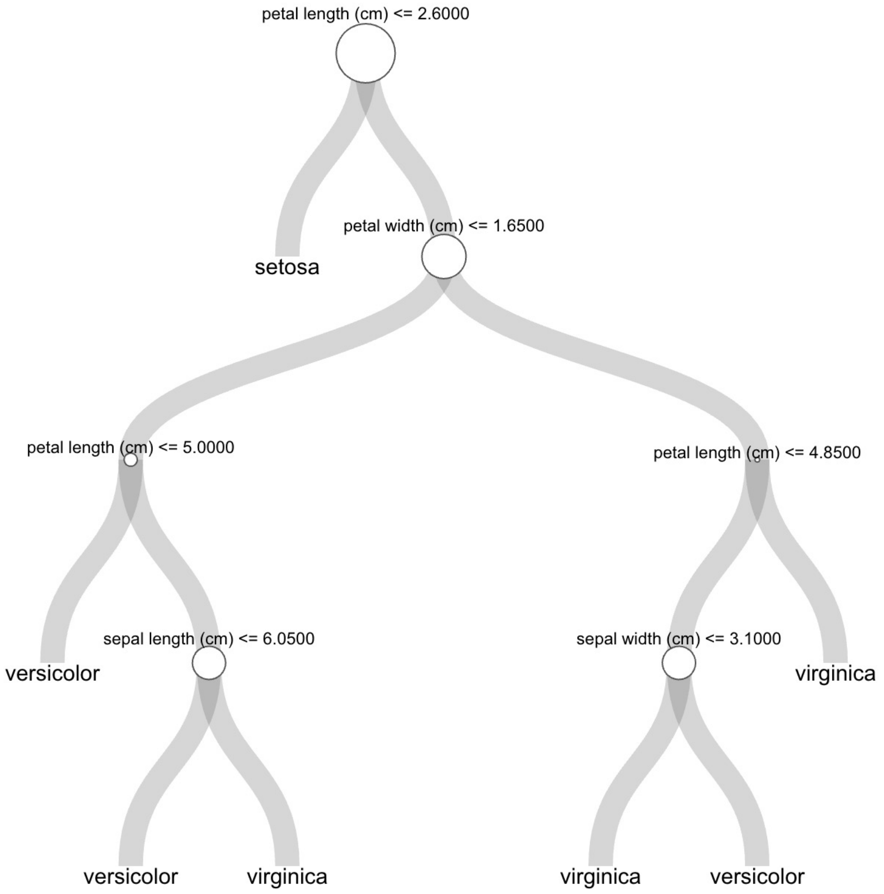
Evaluate


state
fasted
fed
condition
obesity ads
smoking ads
Approach motivation
0
10
20
30
40
obesity ads
smoking ads
obesity ads
smoking ads

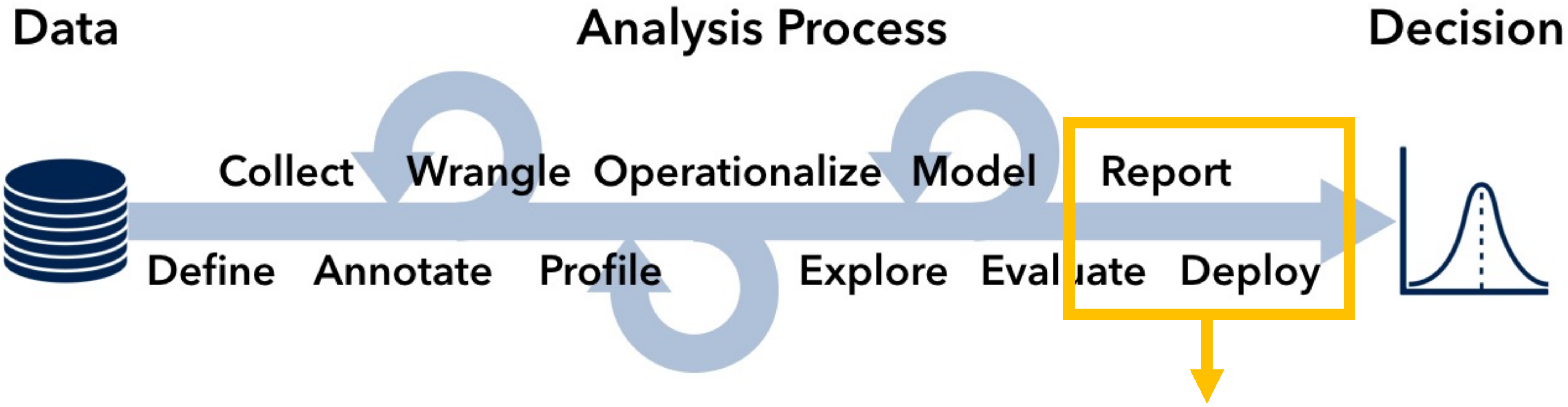

state
fasted
fed
condition
obesity ads
smoking ads
Approach motivation
0
10
20
30
40
obesity ads
smoking ads
obesity ads
smoking ads

Model & Evaluate


```
data = ({"id": "0", "children": [{"id": "1", "impurity": "0.0", "samples":
"39", "value": "[39.  0.  0.]", "class": "0", "self": "0"}, {"id": "2",
"children": [{"id": "3", "children": [{"id": "4", "impurity": "0.0",
"samples": "33", "value": "[ 0. 33.  0.]", "class": "1", "self": "1"},
{"id": "5", "children": [{"id": "6", "impurity": "0.0", "samples": "1",
"value": "[0. 1. 0.]", "class": "1", "self": "1"}, {"id": "7", "impurity":
"0.0", "samples": "3", "value": "[0. 0. 3.]", "class": "2", "self": "2"}],
"name": "sepal length (cm) <= 6.0500", "impurity": "0.375", "samples":
"4"}], "name": "petal length (cm) <= 5.0000", "impurity":
"0.1490138787436085", "samples": "37"}, {"id": "8", "children": [{"id":
"9", "children": [{"id": "10", "impurity": "0.0", "samples": "3", "value":
"[0. 0. 3.]", "class": "2", "self": "2"}, {"id": "11", "impurity": "0.0",
"samples": "1", "value": "[0. 1. 0.]", "class": "1", "self": "1"}],
"name": "sepal width (cm) <= 3.1000", "impurity": "0.375", "samples":
"4"}, {"id": "12", "impurity": "0.0", "samples": "32", "value": "[ 0.  0.
32.]", "class": "2", "self": "2"}], "name": "petal length (cm) <= 4.8500",
"impurity": "0.054012345679012363", "samples": "36"}], "name": "petal
width (cm) <= 1.6500", "impurity": "0.4991555638956652", "samples":
"73"}], "name": "petal length (cm) <= 2.6000", "impurity":
"0.6659757653061225", "samples": "112"})
```

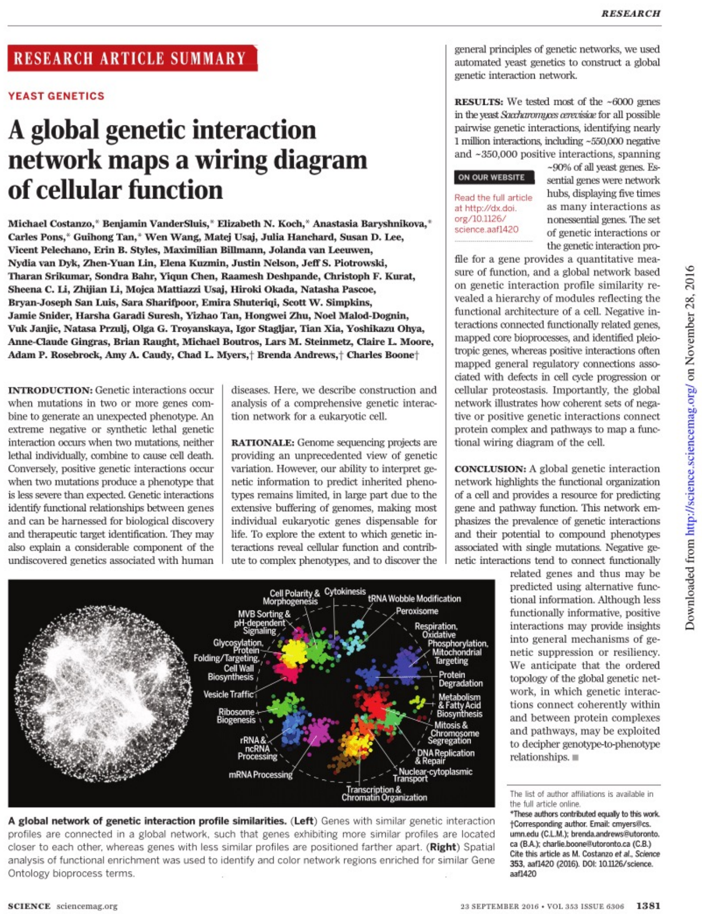



Visualization in data analysis process



Stage 4:
Sharing the insights

Report



RESEARCH

RESEARCH ARTICLE SUMMARY

YEAST GENETICS

A global genetic interaction network maps a wiring diagram of cellular function

Michael Costanzo,* Benjamin VanderSluis,* Elizabeth N. Koch,* Anastasia Baryshnikova,* Carles Pons,* Guihong Tan,* Wen Wang, Matej Usaj, Julia Hanchard, Susan D. Lee, Vicent Pelechano, Erin B. Styles, Maximilian Billmann, Jolanda van Leeuwen, Nydia van Dyk, Zhen-Yuan Lin, Elena Kuzmin, Justin Nelson, Jeff S. Piotrowski, Tharan Srikumar, Sondra Bahr, Yiqun Chen, Raamesh Deshpande, Christoph F. Kurat, Sheena C. Li, Zhijian Li, Mojca Mattiazzi Usaj, Hiroki Okada, Natasha Pascoe, Bryan-Joseph San Luis, Sara Sharifpoor, Emira Shuteriqi, Scott W. Simpkins, Jamie Snider, Harsha Garadi Suresh, Yizhao Tan, Hongwei Zhu, Noel Malod-Dognin, Vuk Janjic, Natasa Przulj, Olga G. Troyanskaya, Igor Stagljar, Tian Xia, Yoshikazu Ohya, Anne-Claude Gingras, Brian Raught, Michael Boutros, Lars M. Steinmetz, Claire L. Moore, Adam P. Rosebrock, Amy A. Caudy, Chad L. Myers,† Brenda Andrews,† Charles Boone†

INTRODUCTION: Genetic interactions occur when mutations in two or more genes combine to generate an unexpected phenotype. An extreme negative or synthetic lethal genetic interaction occurs when two mutations, neither lethal individually, combine to cause cell death. Conversely, positive genetic interactions occur when two mutations produce a phenotype that is less severe than expected. Genetic interactions identify functional relationships between genes and can be harnessed for biological discovery and therapeutic target identification. They may also explain a considerable component of the undiscovered genetics associated with human diseases. Here, we describe construction and analysis of a comprehensive genetic interaction network for a eukaryotic cell.

RATIONALE: Genome sequencing projects are providing an unprecedented view of genetic variation. However, our ability to interpret genetic information to predict inherited phenotypes remains limited, in large part due to the extensive buffering of genomes, making most individual eukaryotic genes dispensable for life. To explore the extent to which genetic interactions reveal cellular function and contribute to complex phenotypes, and to discover the general principles of genetic networks, we used automated yeast genetics to construct a global genetic interaction network.

RESULTS: We tested most of the ~6000 genes in the yeast *Saccharomyces cerevisiae* for all possible pairwise genetic interactions, identifying nearly 1 million interactions, including ~550,000 negative and ~350,000 positive interactions, spanning ~90% of all yeast genes. Essential genes were network hubs, displaying five times as many interactions as nonessential genes. The set of genetic interactions or the genetic interaction profile for a gene provides a quantitative measure of function, and a global network based on genetic interaction profile similarity revealed a hierarchy of modules reflecting the functional architecture of a cell. Negative interactions connected functionally related genes, mapped core bioprocesses, and identified pleiotropic genes, whereas positive interactions often mapped general regulatory connections associated with defects in cell cycle progression or cellular proteostasis. Importantly, the global network illustrates how coherent sets of negative or positive genetic interactions connect protein complex and pathways to map a functional wiring diagram of the cell.

ON OUR WEBSITE

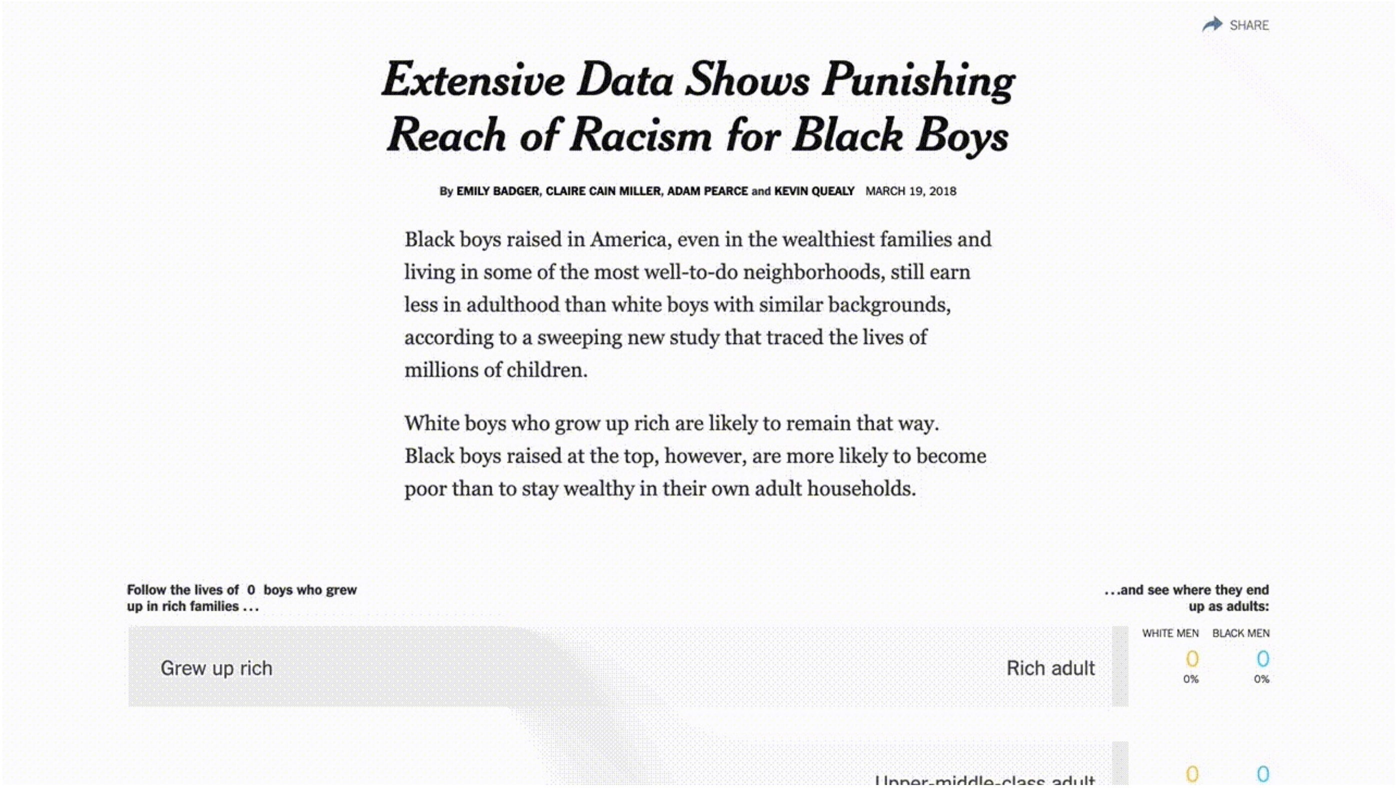
Read the full article at http://dx.doi.org/10.1126/science.aaf1420

CONCLUSION: A global genetic interaction network highlights the functional organization of a cell and provides a resource for predicting gene and pathway function. This network emphasizes the prevalence of genetic interactions and their potential to compound phenotypes associated with single mutations. Negative genetic interactions tend to connect functionally related genes and thus may be predicted using alternative functional information. Although less functionally informative, positive interactions may provide insights into general mechanisms of genetic suppression or resiliency. We anticipate that the ordered topology of the global genetic network, in which genetic interactions connect coherently within and between protein complexes and pathways, may be exploited to decipher genotype-to-phenotype relationships. ■

A global network of genetic interaction profile similarities. (**Left**) Genes with similar genetic interaction profiles are connected in a global network, such that genes exhibiting more similar profiles are located closer to each other, whereas genes with less similar profiles are positioned farther apart. (**Right**) Spatial analysis of functional enrichment was used to identify and color network regions enriched for similar Gene Ontology bioprocess terms.

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. Email: cmyers@cs.umn.edu (C.L.M.); brenda.andrews@utoronto.ca (B.A.); charlie.boone@utoronto.ca (C.B.)
Cite this article as M. Costanzo *et al.*, *Science* 353, aaf1420 (2016). DOI: 10.1126/science.aaf1420

SCIENCE sciencemag.org 23 SEPTEMBER 2016 • VOL 353 ISSUE 6306 1381



SHARE

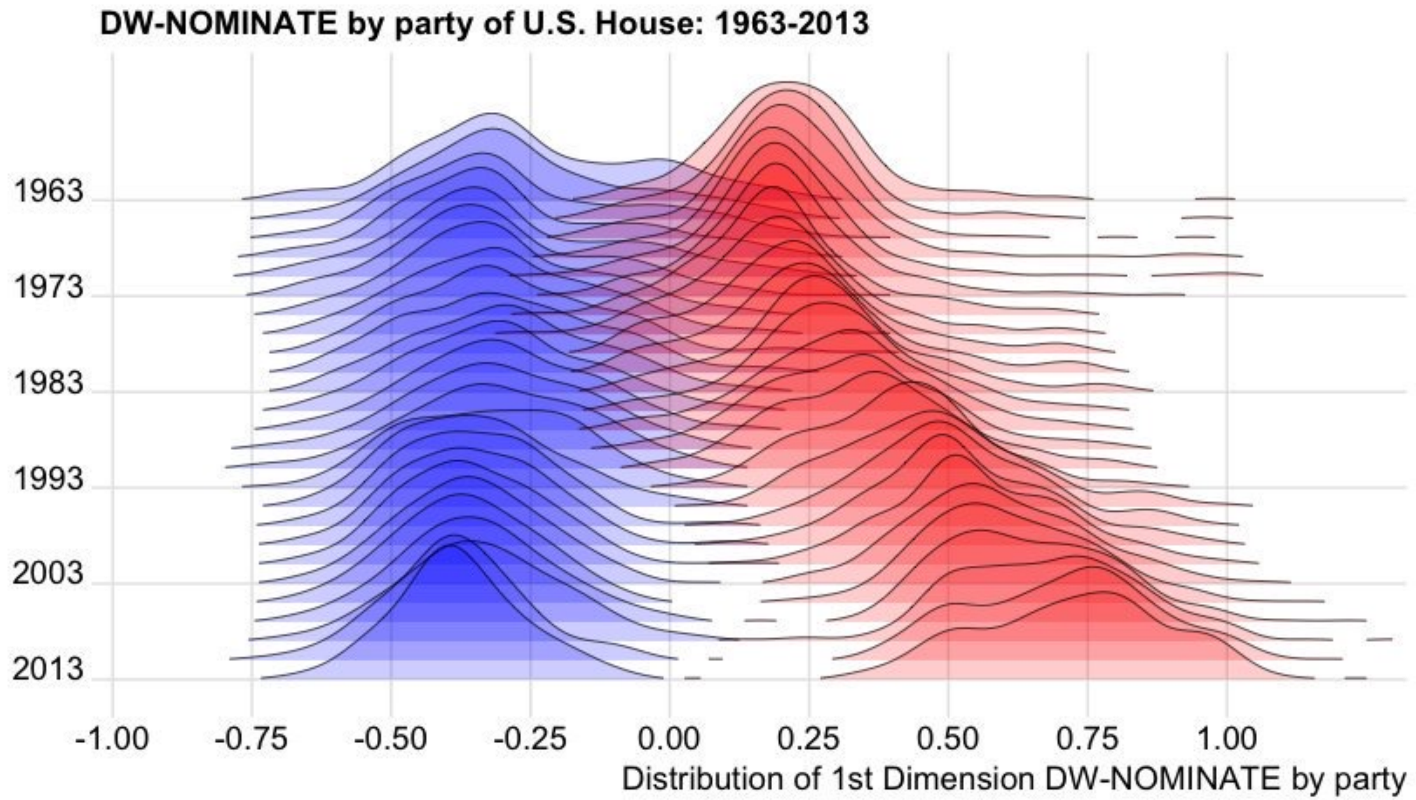
Extensive Data Shows Punishing Reach of Racism for Black Boys

By EMILY BADGER, CLAIRE CAIN MILLER, ADAM PEARCE and KEVIN QUEALY MARCH 19, 2018

Black boys raised in America, even in the wealthiest families and living in some of the most well-to-do neighborhoods, still earn less in adulthood than white boys with similar backgrounds, according to a sweeping new study that traced the lives of millions of children.

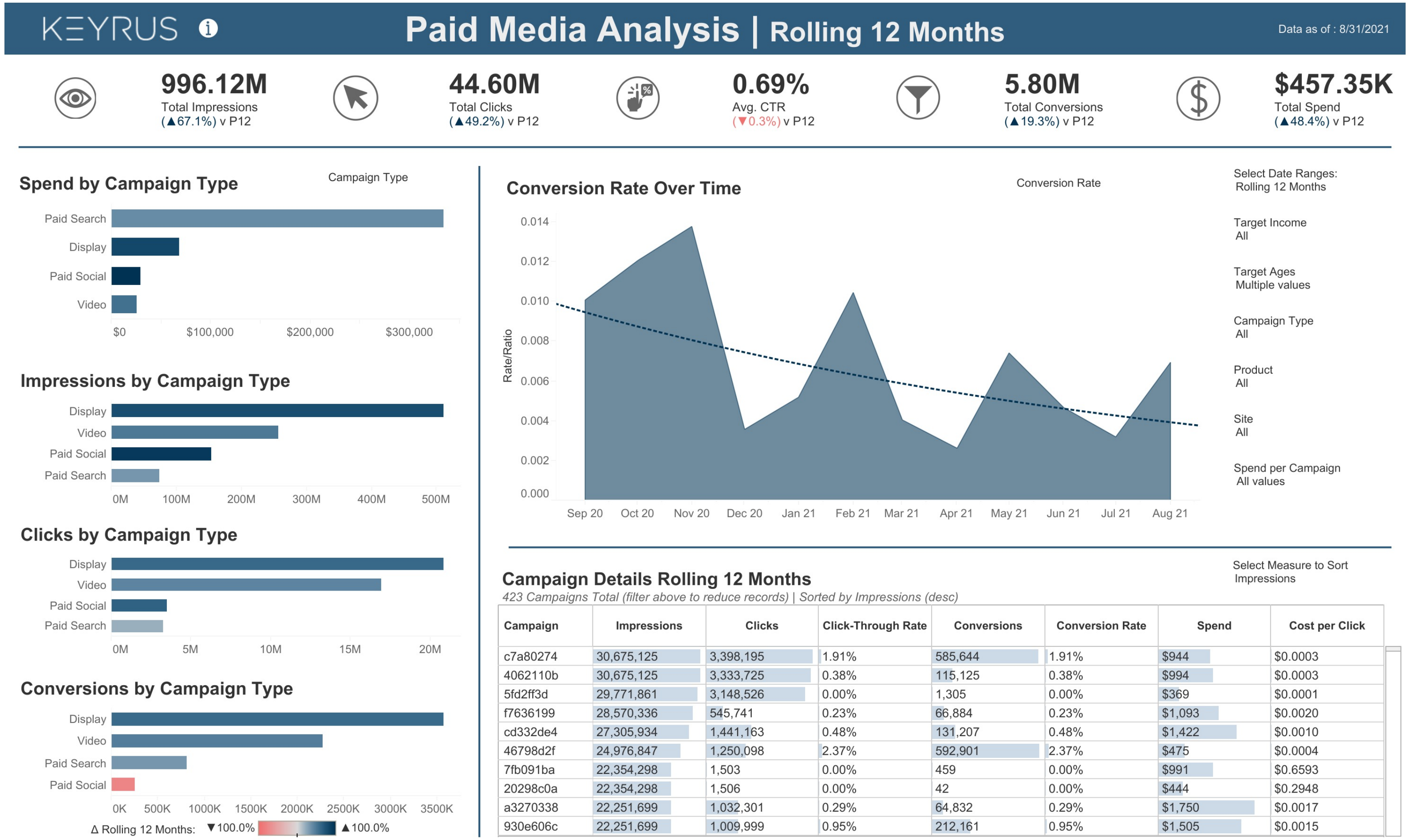
White boys who grow up rich are likely to remain that way. Black boys raised at the top, however, are more likely to become poor than to stay wealthy in their own adult households.

Deploy: Is my distribution shifting?



https://rpubs.com/paul4forest/movingdistribution

Deploy: Dashboard



Campaign	Impressions	Clicks	Click-Through Rate	Conversions	Conversion Rate	Spend	Cost per Click
c7a80274	30,675,125	3,398,195	1.91%	585,644	1.91%	$944	$0.0003
4062110b	30,675,125	3,333,725	0.38%	115,125	0.38%	$994	$0.0003
5fd2ff3d	29,771,861	3,148,526	0.00%	1,305	0.00%	$369	$0.0001
f7636199	28,570,336	545,741	0.23%	66,884	0.23%	$1,093	$0.0020
cd332de4	27,305,934	1,441,163	0.48%	131,207	0.48%	$1,422	$0.0010
46798d2f	24,976,847	1,250,098	2.37%	592,901	2.37%	$475	$0.0004
7fb091ba	22,354,298	1,503	0.00%	459	0.00%	$991	$0.6593
20298c0a	22,354,298	1,506	0.00%	42	0.00%	$444	$0.2948
a3270338	22,251,699	1,032,301	0.29%	64,832	0.29%	$1,750	$0.0017
930e606c	22,251,699	1,009,999	0.95%	212,161	0.95%	$1,505	$0.0015

https://public.tableau.com/app/profile/keyrus/viz/PaidMediaAnalysisKeyrus/PaidMediaAnalysisOverview

Visualization in data analysis process

Remember: Data visualization is critical to your analysis process

Human Perception

How do humans see data?

Perceptual grammar

Why should we be interested in visualization?

Because the human visual system is a **pattern seeker** of enormous power and subtlety. The eye and the visual cortex of the brain form a massively parallel processor that provides the highest bandwidth channel into human cognitive centers. At higher levels of processing, **perception and cognition are closely interrelated** which is the reason why the words "understanding" and "seeing" are synonymous.

- Ware 1998

Figure 1. Adapted from Nakayama et al. 1989

Perceptual grammar

The more general point is that **when data is presented in certain ways, the patterns can be readily perceived**. We can think of a "grammar" of perception and this grammar of perception can be translated directly into a rules for displaying information.

If we can understand this **perceptual grammar,** then we can present our data in such a way that the important and informative patterns stand out. If we disobey the rules, our data will be incomprehensible or misleading.

- Ware 1998

Figure 1. Adapted from Nakayama et al. 1989

How can we leverage our perception?

Signal detection

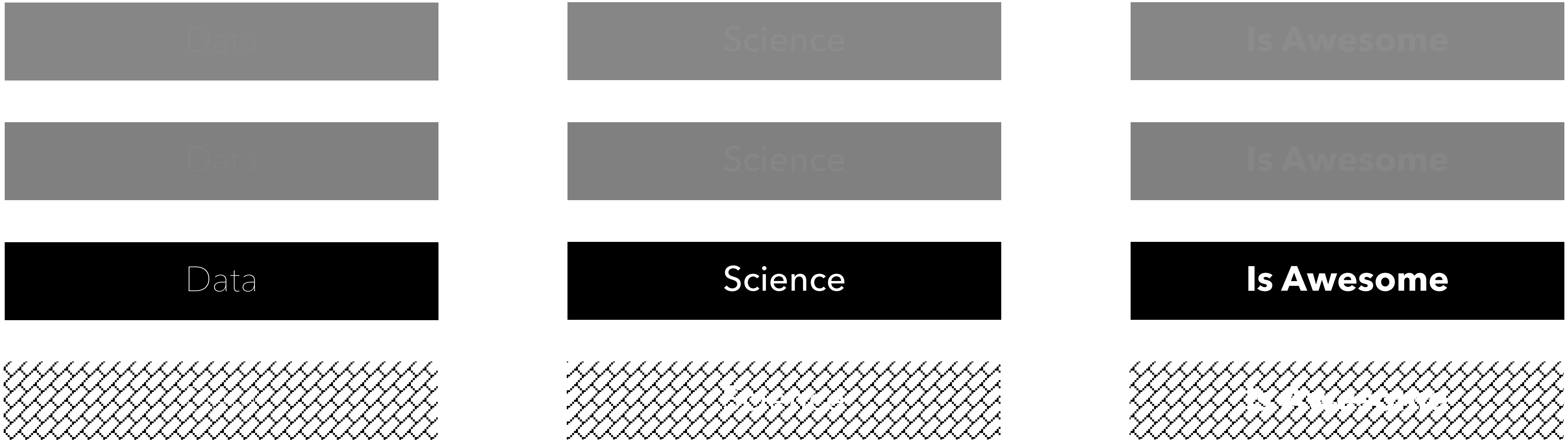
Magnitude estimation

Pre-attentive processing

Distinctive colors

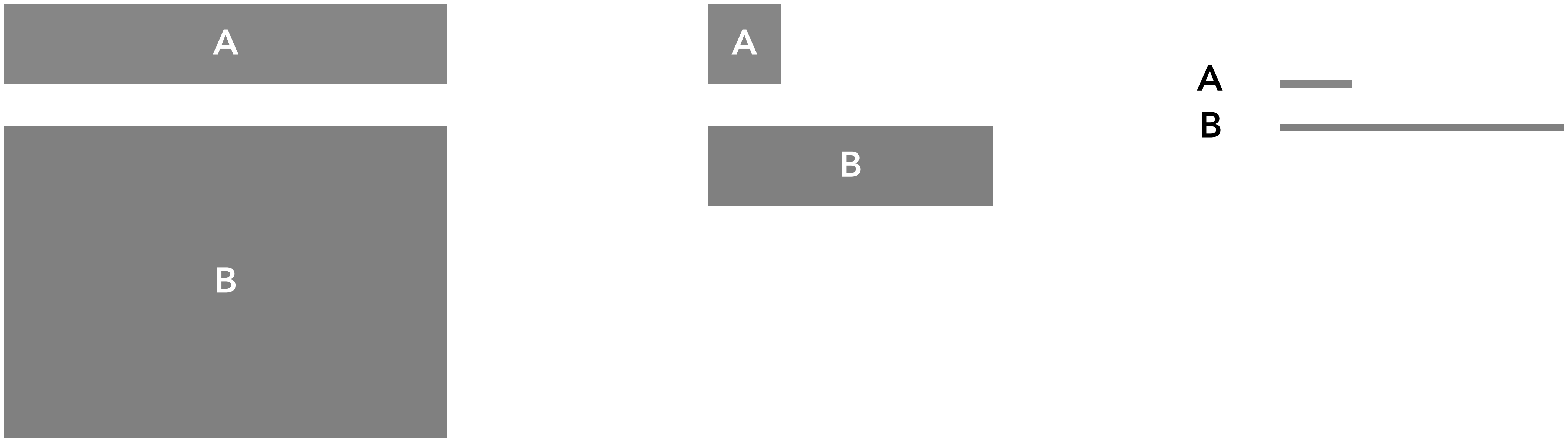
Signal detection

Can you read the text?



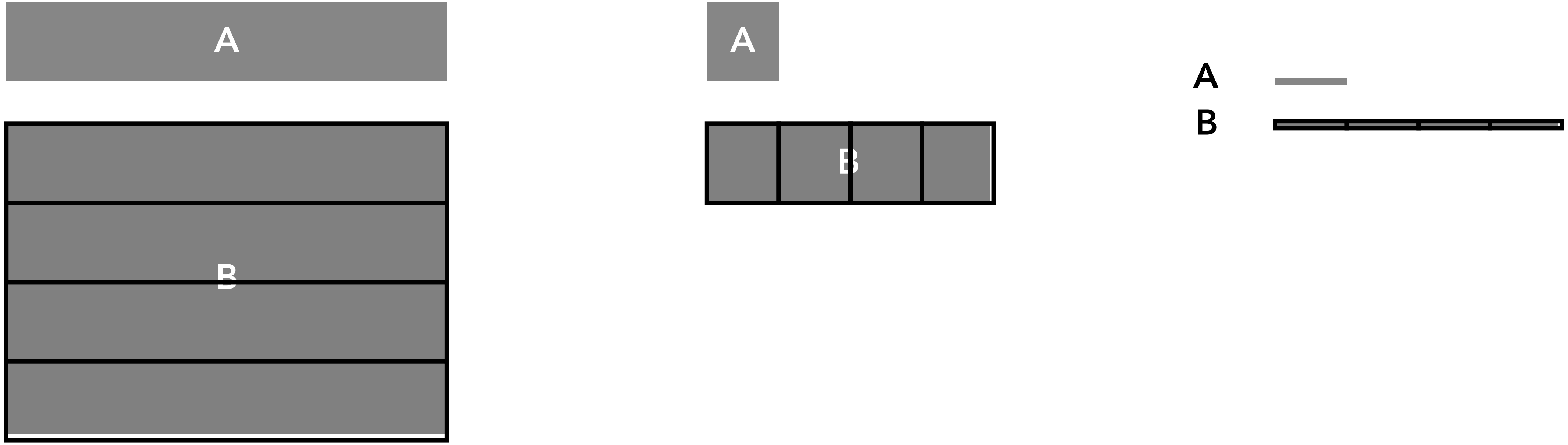
Magnitude estimation

How many A's in B?



Magnitude estimation

How many A's in B?



Encoding

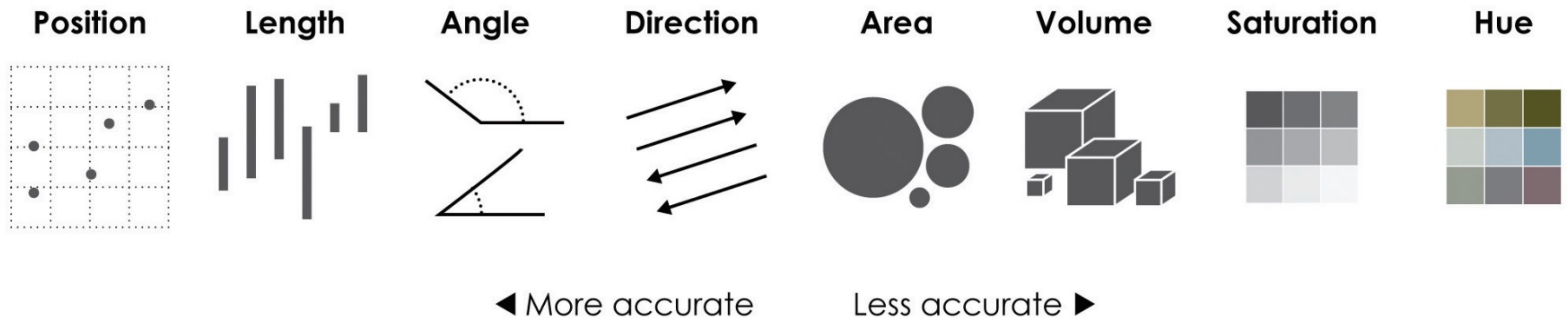
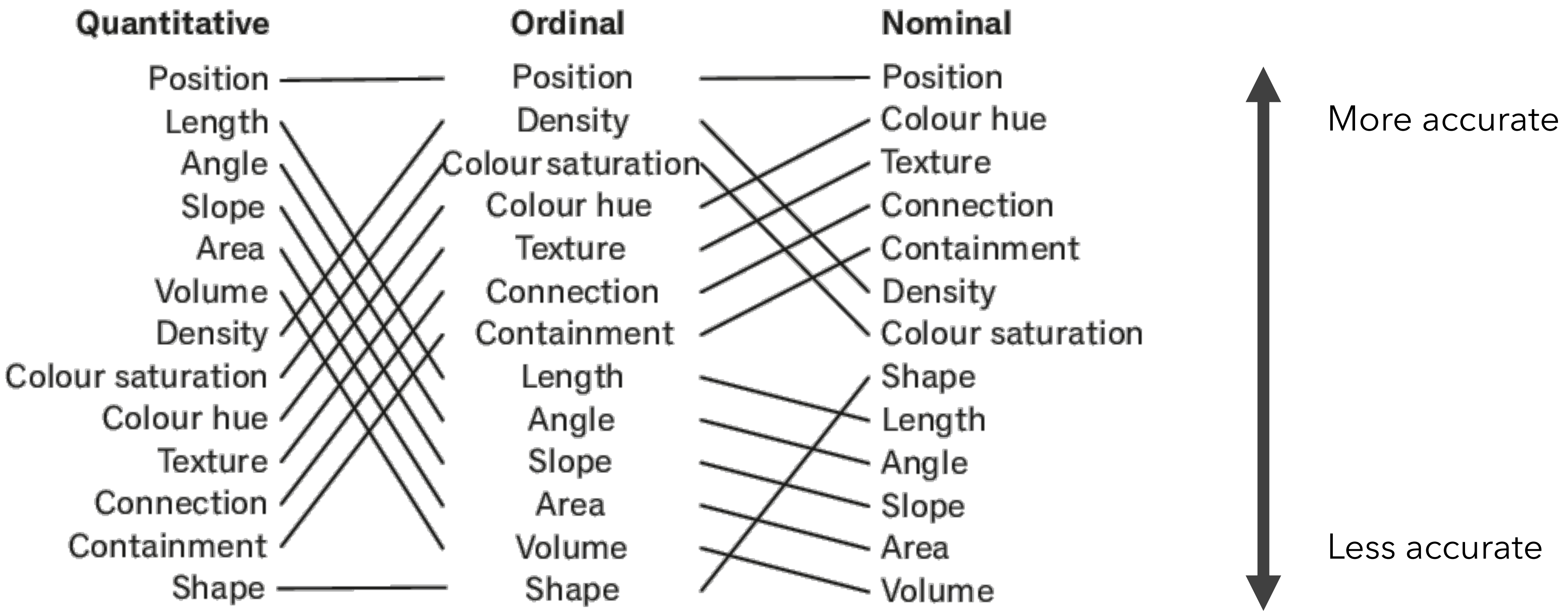


FIGURE 3-12 *Visual cues ranked by Cleveland and McGill*

Data Points, Nathan Yau

Task to find the best encoding



Ranking of visual variables by data type. Mackinlay 1986

Multiple encodings

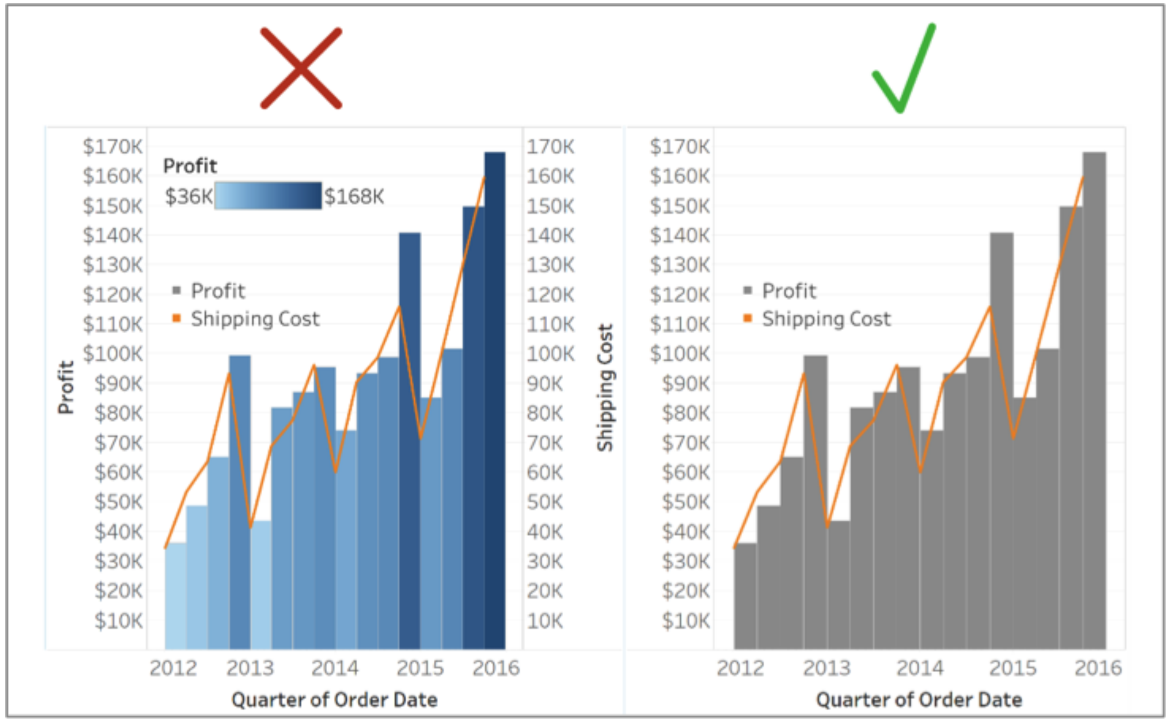
Redundant encoding can be beneficial

Color Shape Redundant

https://visualthinking.psych.northwestern.edu/projects/redundantencoding.html

Multiple encodings

Redundant encoding can be beneficial, except when it's not



https://www.teknionusa.com/blog/the-10-commandments-of-visual-analytics-in-tableau

Pre-attentive processing

Subconscious accumulation of information from the environment

All information is pre-attentively processed

Brain filters and processes what's important

Salient or relevant information is selected and analyzed by conscious (attentive) processing

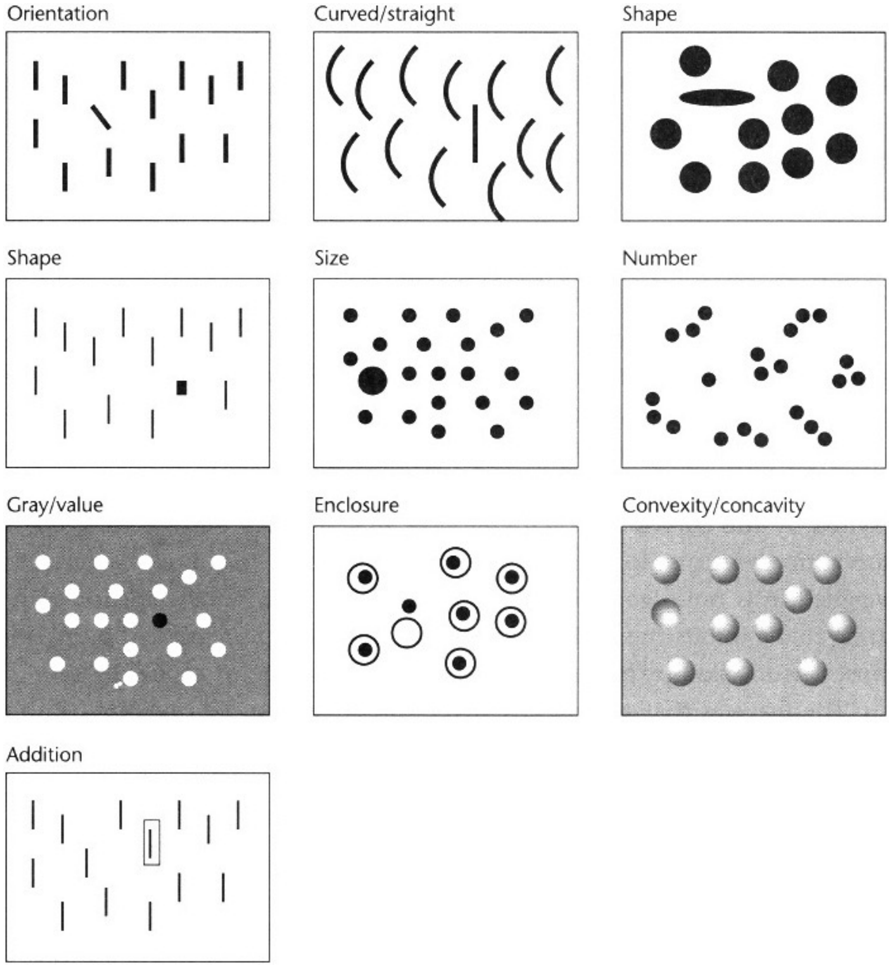
Pre-attentive features

Form – line orientation, line length, line width, line collinearity, size, curvature, spatial grouping, added marks, luminosity.

Color – hue, intensity

Motion – flicker, direction of motion

Spatial position – 2d position, stereoscopic depth, convex/concave shape from shading



Information Visualization. Ware 1999

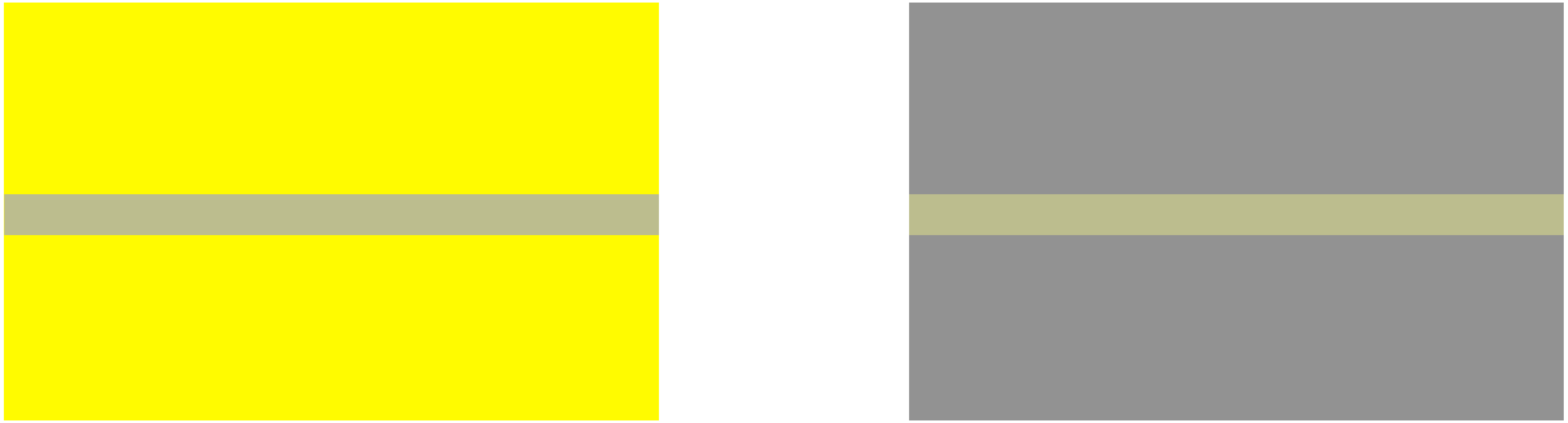
Effective use of color

In order to use color effectively it is necessary to recognize that it **deceives continually**.

- Josef Albers, Interaction of Color

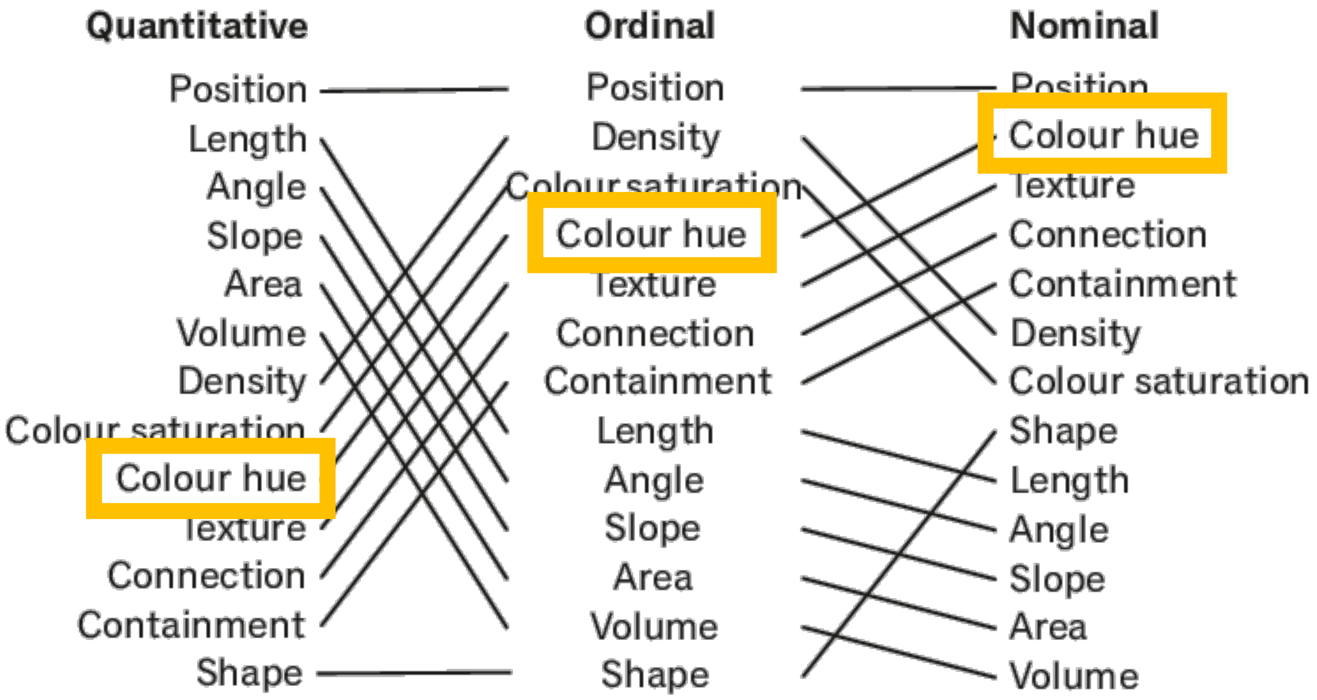
Effective use of color

Are the lines in the middle of the two boxes the same color?



Color

Best for **nominal variables** (categorical, binary)



Visually distinct colors

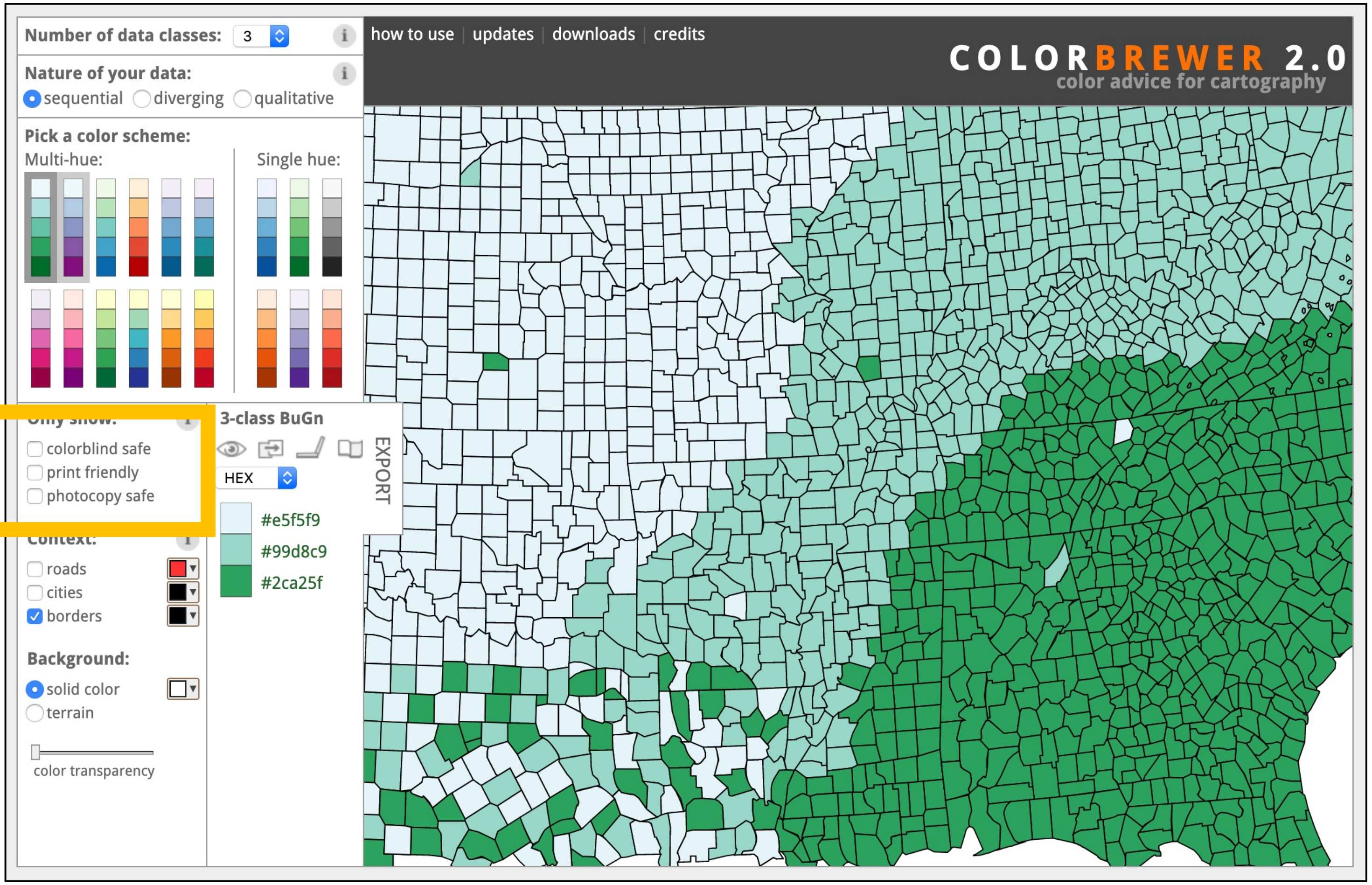
Color Name Distance										Salience	Name
0.00	1.00	1.00	1.00	0.96	1.00	1.00	0.99	1.00	0.19	.47	**blue** 65.3%
1.00	**0.00**	1.00	0.98	1.00	1.00	1.00	1.00	0.97	1.00	.87	**orange** 92.2%
1.00	1.00	**0.00**	1.00	1.00	1.00	1.00	1.00	0.70	0.99	.70	**green** 81.3%
1.00	0.98	1.00	**0.00**	1.00	0.96	0.99	1.00	1.00	1.00	.64	**red** 79.3%
0.96	1.00	1.00	1.00	**0.00**	0.95	0.83	0.98	1.00	0.97	.43	**purple** 52.5%
1.00	1.00	1.00	0.96	0.95	**0.00**	0.99	0.96	0.96	1.00	.47	**brown** 60.5%
1.00	1.00	1.00	0.99	0.83	0.99	**0.00**	1.00	1.00	1.00	.47	**pink** 60.3%
0.99	1.00	1.00	1.00	0.98	0.96	1.00	**0.00**	1.00	0.99	.74	**grey** 83.7%
1.00	0.97	0.70	1.00	1.00	0.96	1.00	1.00	**0.00**	1.00	.11	**yellow** 20.1%
0.19	1.00	0.99	1.00	0.97	1.00	1.00	0.99	1.00	**0.00**	.25	**blue** 27.2%
Tableau-10								*Average*	*0.96*	*.52*	

Heer & Stone, 2012

Brewer palettes

Color combinations selected for cartography

Don't forget about **colorblindness** and **black/white printing**



http://colorbrewer2.org

Storytelling with Data

How do we tell stories with visualization?

Exploratory analysis

Understand and get familiar with your data and generate lots of information

Mining!

Explanatory analysis

Learning more about what you found to communicate what you found and tell a story about it

Steps to storytelling with data

Think about the context

Who are you telling the story to?

Craft the narrative

How are you telling the story?

Design appropriate visualizations

What are you telling the story with?

Context matters….a lot

Who?

What?

How?

Context matters….a lot

Who?

Who is your audience? Can you be very specific?

What's your relationship with your audience? Do they know you well enough to understand your assumptions? Do you have credibility?

Context matters….a lot

What?

What do you want your audience to know or do? What action do you want them to take?

Context matters….a lot

How?

What data do you have to make your case? How will you present your data?

How will you communicate to your audience? What affordances do you have? How much control do you have?

Example context: executive pitch

Who is my audience?

Executives and program directors who approved funding for research internship program.

What does success look like?

Funded research under the program was a success and provided tangible impact to the product. They should continue funding the program.

How would I do this?

Illustrate the number of publications, product features that were shipped, successful career paths of the interns in the program.

Example context: public

Large-scale physical activity data reveal worldwide activity inequality

Tim Althoff[1], Rok Sosič[1], Jennifer L. Hicks[2], Abby C. King[3,4], Scott L. Delp[2,5] & Jure Leskovec[1,6]

Who is my audience?

- General public audience

What does success look like?

- As an individual, I can see where I and my country stand in worldwide physical activity inequality data

How would I do this?

- Interactive visualization where, given a country and my daily average step count, display the step distribution

Example context: public

- Country = China
- Daily steps = 3500

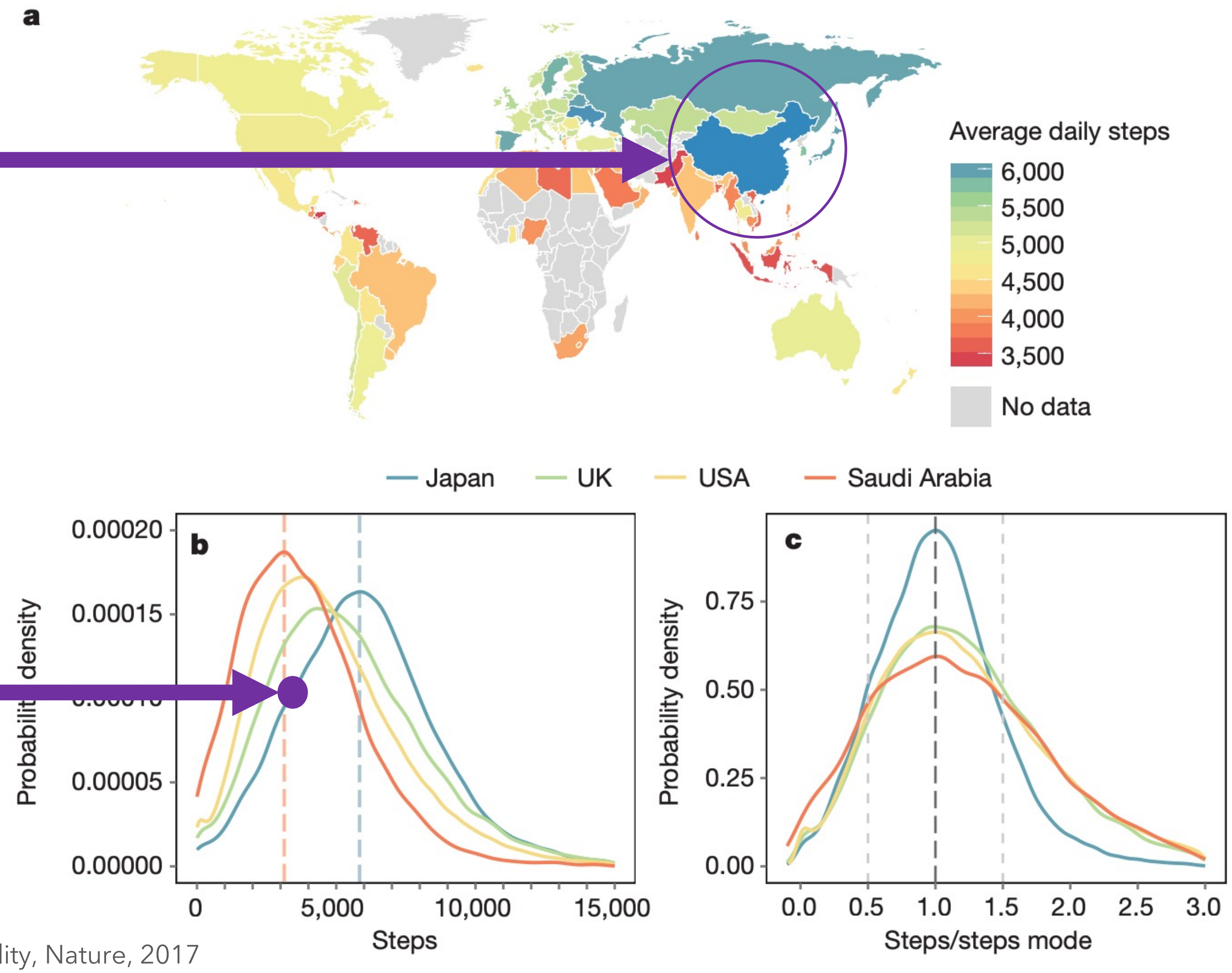


Figure 1 | Smartphone data from over 68 million days of activity by 717,527 individuals reveal variability in physical activity across the world.

Althoff et al., Large-scale physical activity data reveal worldwide activity inequality, Nature, 2017

Steps to storytelling with data

Think about the context

Who are you telling the story to?

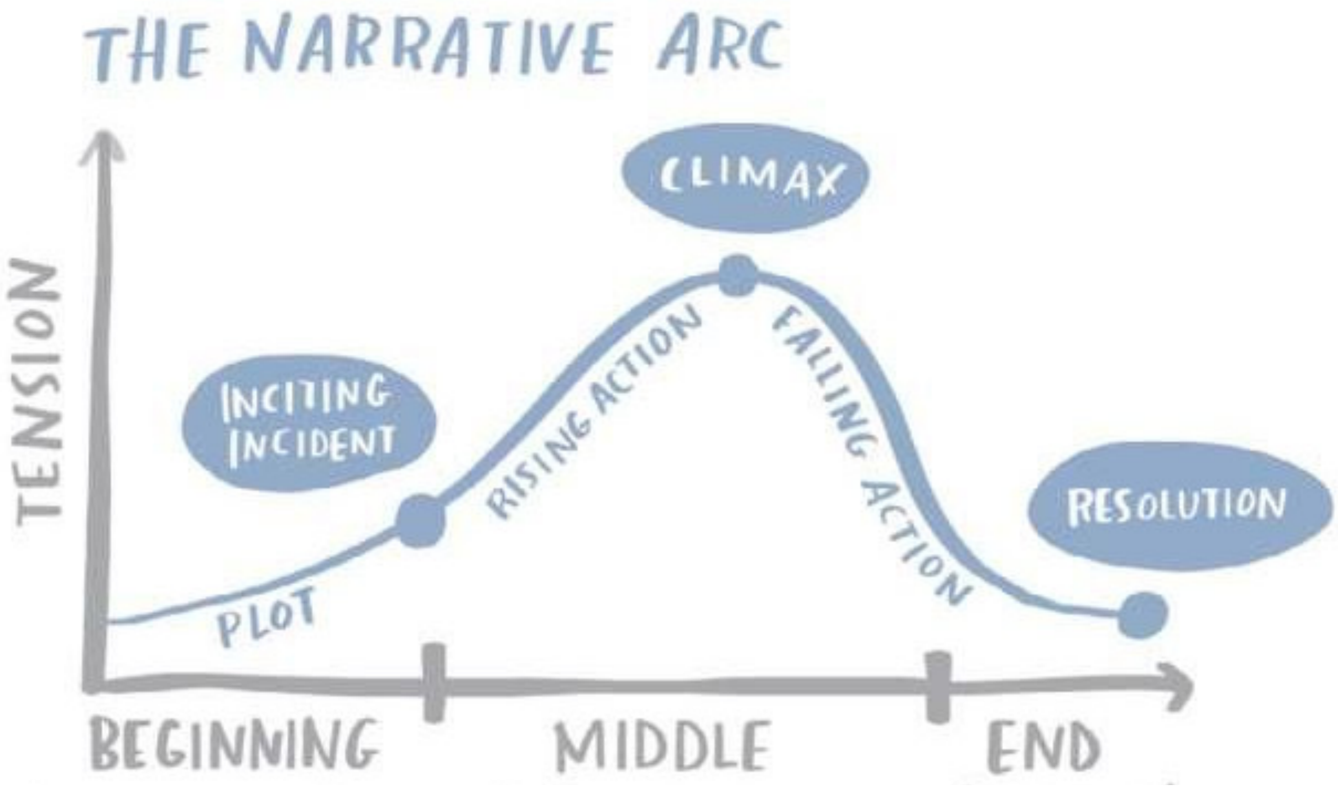
Craft the narrative

How are you telling the story?

Design appropriate visualizations

What are you telling the story with?

Constructing the narrative



https://www.storytellingwithdata.com/

The beginning: set the stage

Setting: when and where does the story take place?

Main character: who is driving the action?

Imbalance: Why is it necessary, what has changed?

Balance: What do you want to see happen?

Solution: How will you bring about the changes?

The middle: show the data

Provide evidence through data

Incorporate external context or comparisons

Provide examples to illustrate the issue

Articulate what would happen if no action was taken

Discuss potential mitigations or solutions and benefits

Remind them they are in unique position to drive action

The ending: call to action

Tie it back to the beginning

Recap problems and resulting need for action

Reiterate sense of urgency

Key takeaways and action items

Steps to storytelling with data

Think about the context

Who are you telling the story to?

Craft the narrative

How are you telling the story?

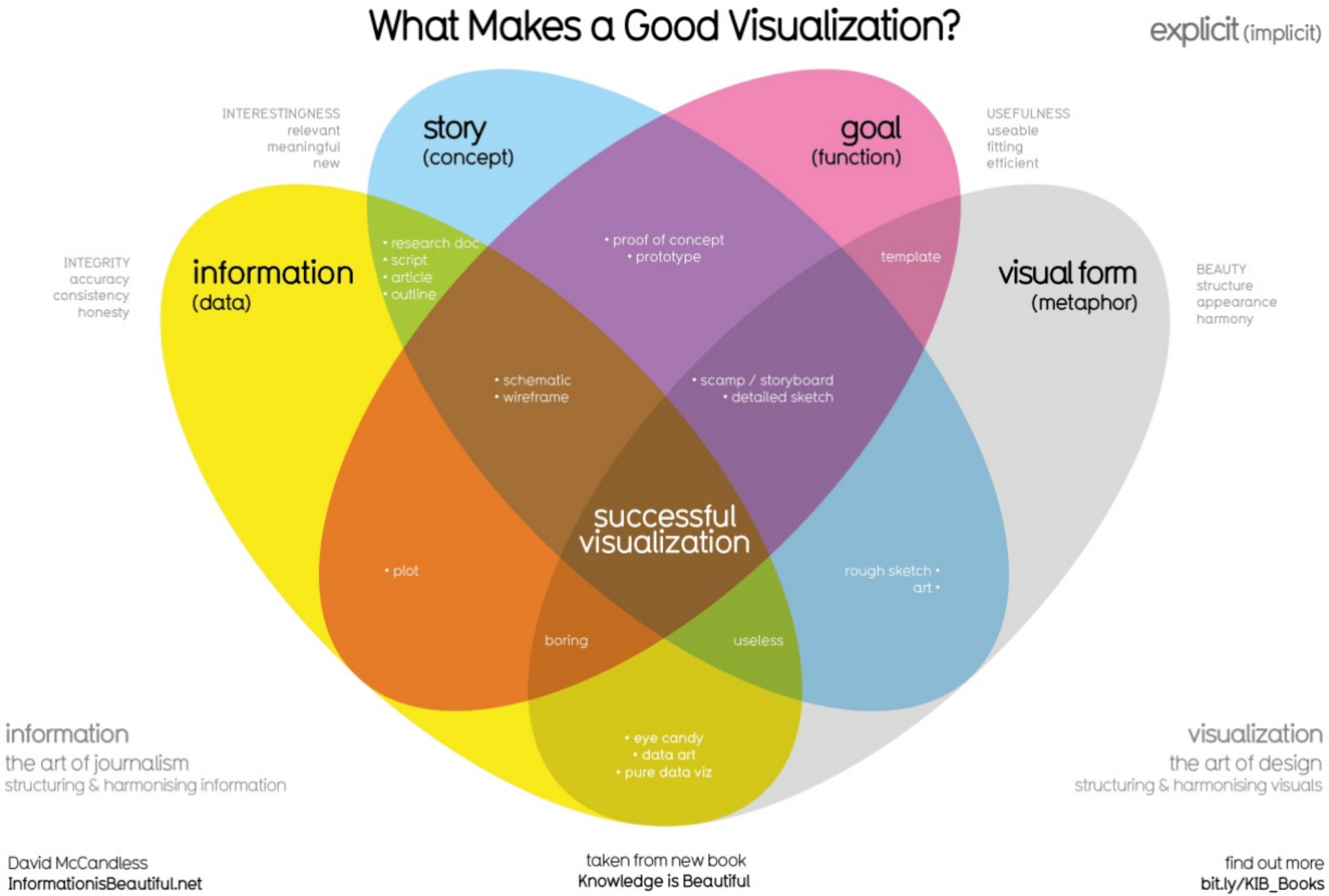
Design appropriate visualizations

What are you telling the story with?

Visualization Design

How do we design "good" visualizations?

What makes a good visualization?


What Makes a Good Visualization?
explicit (implicit)
INTERESTINGNESS
relevant
meaningful
new
story
(concept)
goal
(function)
USEFULNESS
useable
fitting
efficient
INTEGRITY
accuracy
consistency
honesty
information
(data)
• research doc
• script
• article
• outline
• proof of concept
• prototype
template
visual form
(metaphor)
BEAUTY
structure
appearance
harmony
• schematic
• wireframe
• scamp / storyboard
• detailed sketch
successful visualization
• plot
rough sketch •
art •
boring
useless
• eye candy
• data art
• pure data viz
information
the art of journalism
structuring & harmonising information
visualization
the art of design
structuring & harmonising visuals
David McCandless
InformationisBeautiful.net
taken from new book
Knowledge is Beautiful
find out more
bit.ly/KIB_Books

Critique by redesign

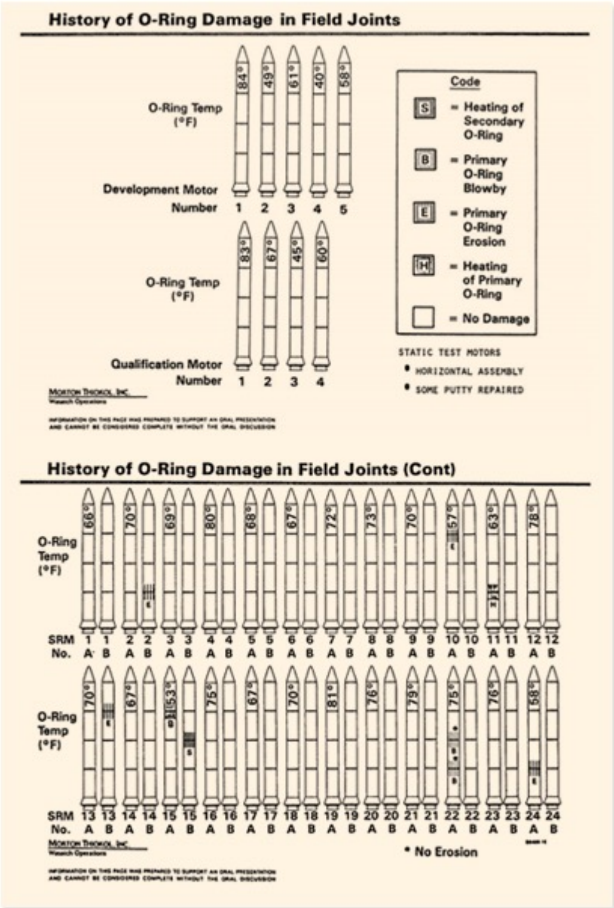
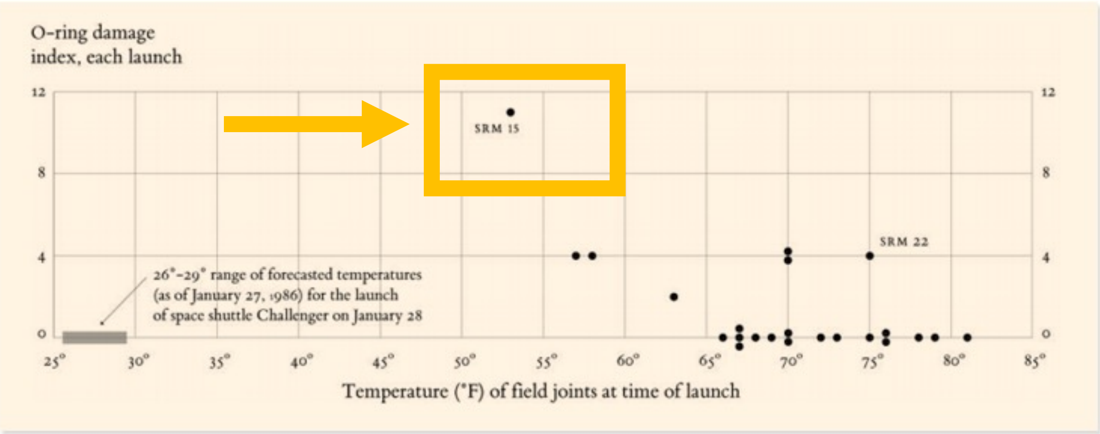


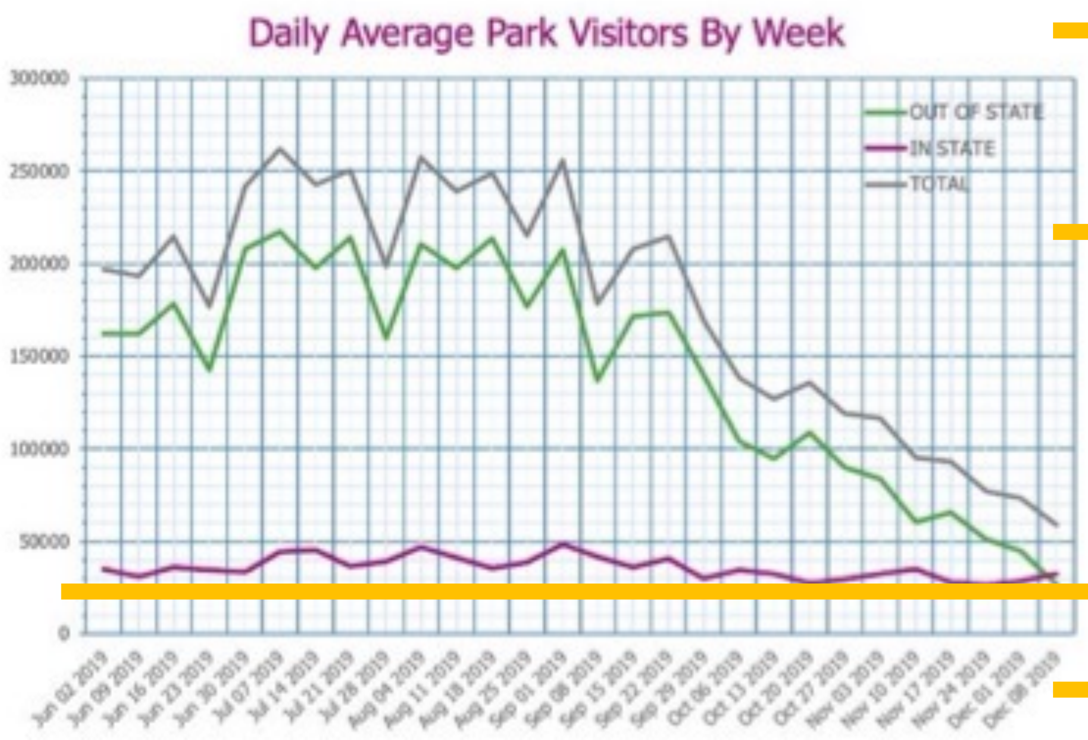
Chart shown to the presidential commission investigation on the Space Shuttle Challenger in 1986. The chart shows the history of O-Ring failures



Edward Tufte's redesign of the same chart showing O-Ring failures.

https://medium.com/@hint_fm/design-and-redesign-4ab77206cf9

Identify and eliminate clutter



Weekly? Daily?

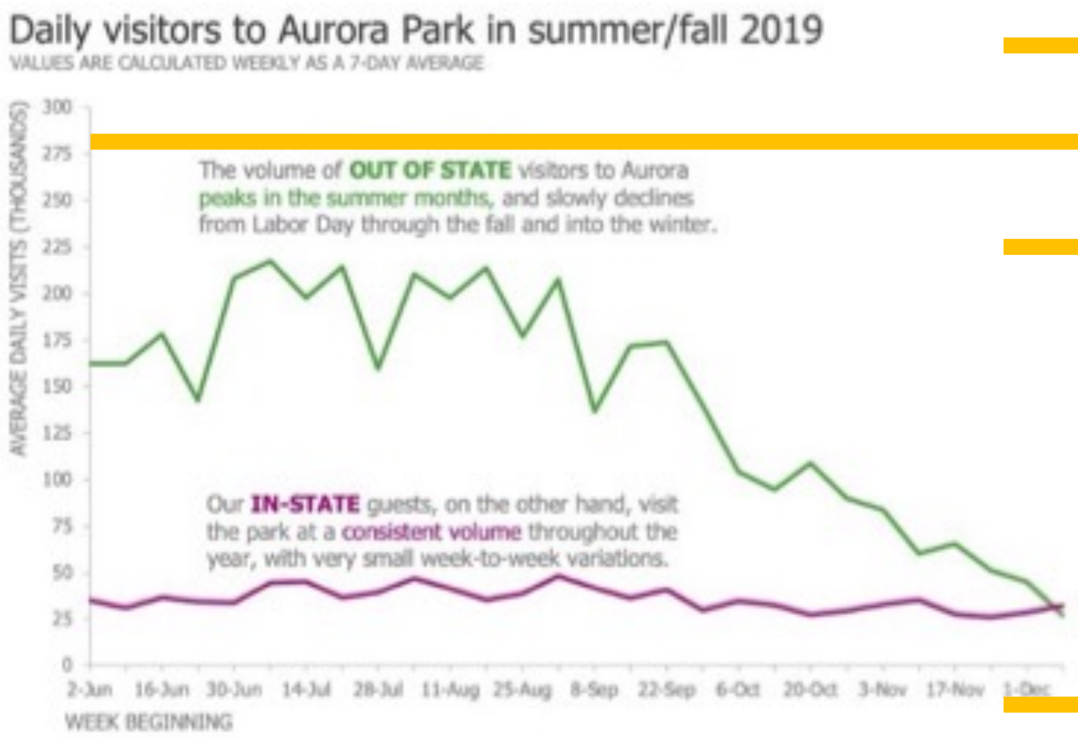
Background noise

No axis label

A lot of dates

https://www.storytellingwithdata.com/

Identify and eliminate clutter



More explanation

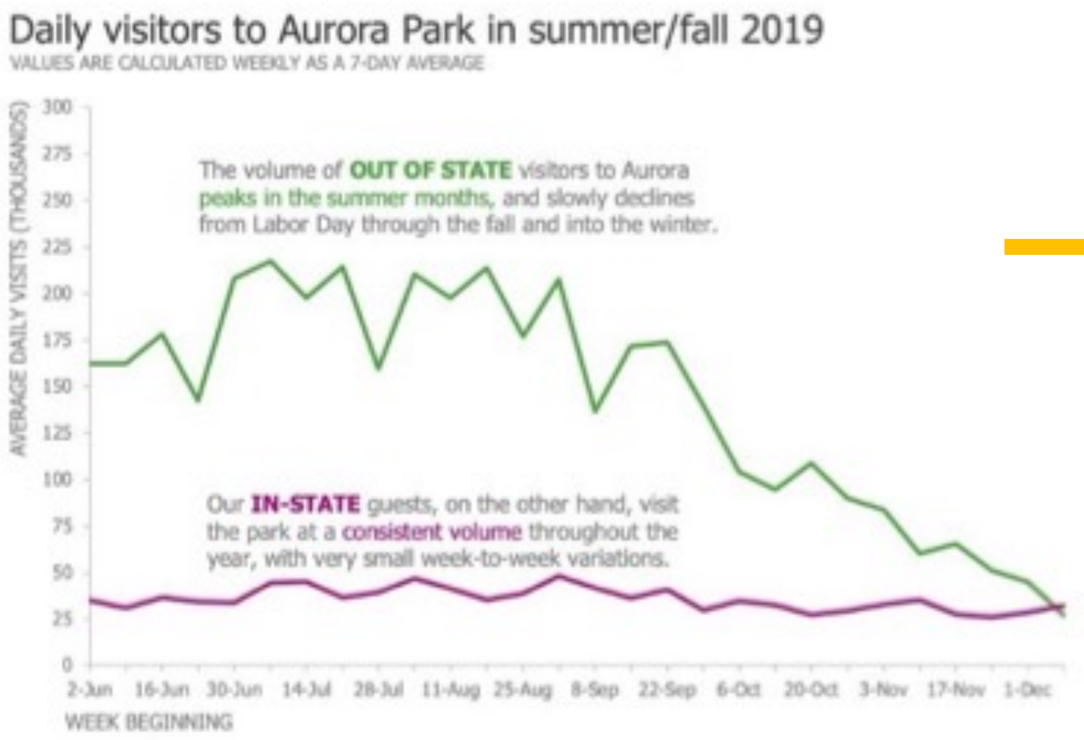
Label + new units

Removed background

Fewer dates + label

https://www.storytellingwithdata.com/

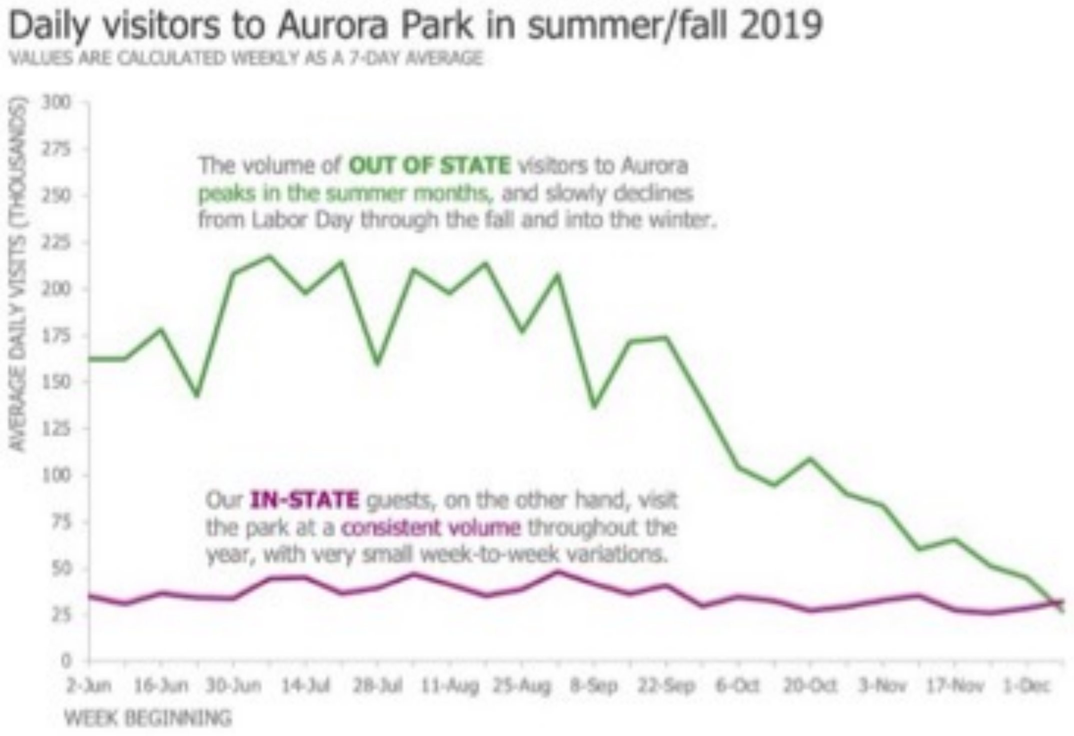
Identify and eliminate clutter



Inline legend + summary

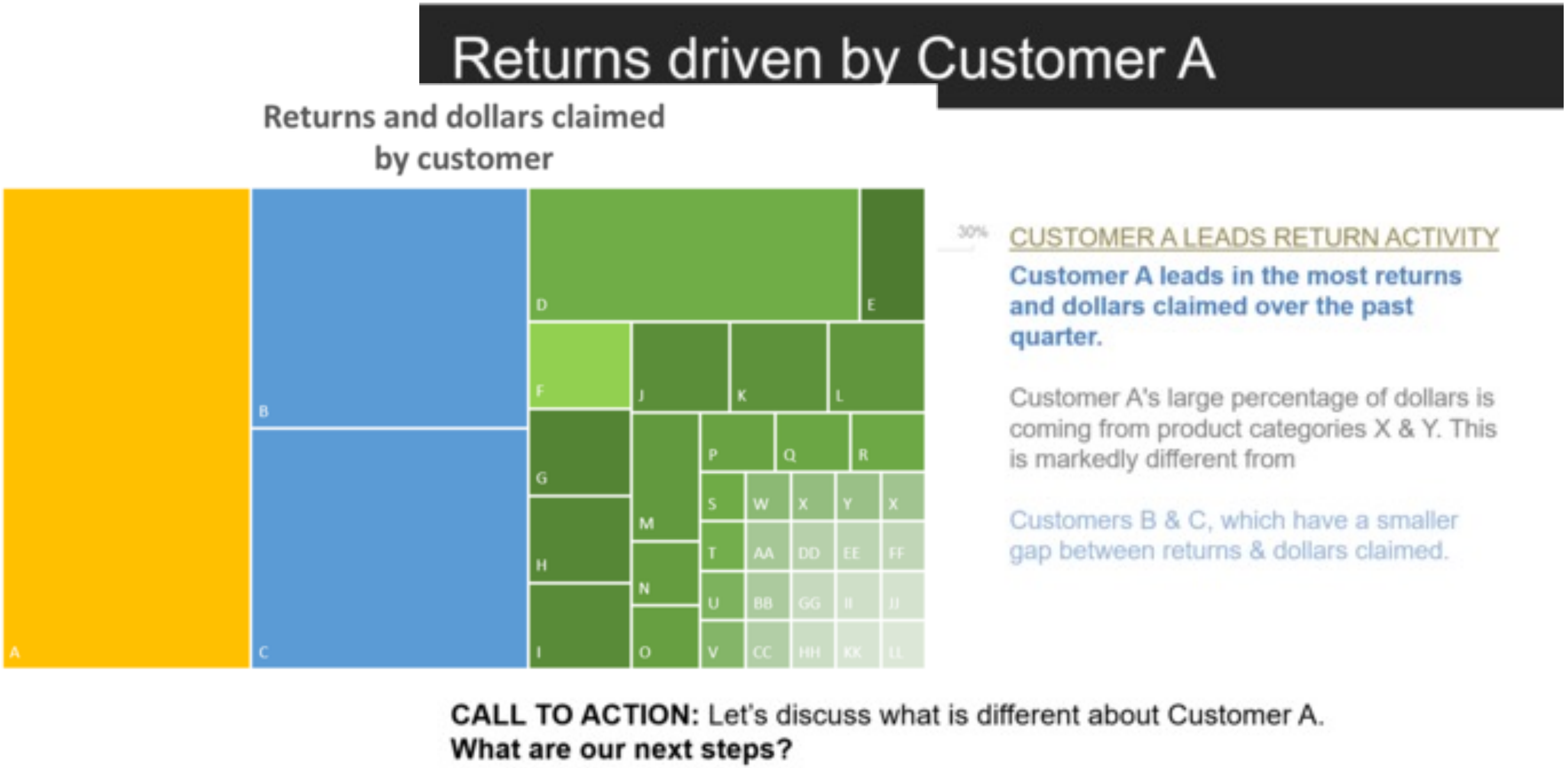
https://www.storytellingwithdata.com/

Identify and eliminate clutter



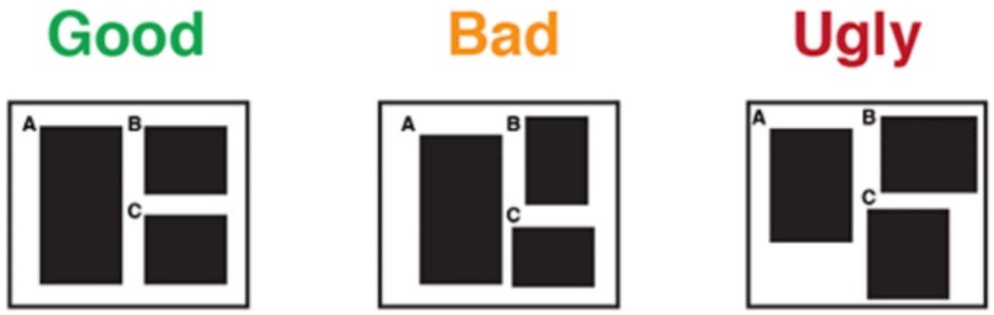
https://www.storytellingwithdata.com/

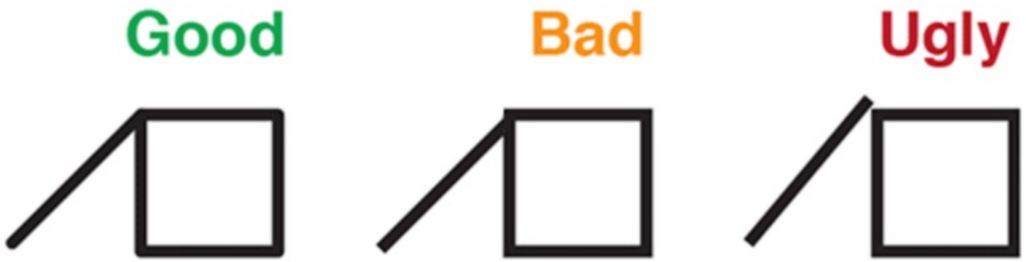
Use visualizations that are easy to read



https://www.storytellingwithdata.com/

Keep consistent order and alignment





Keep consistent color

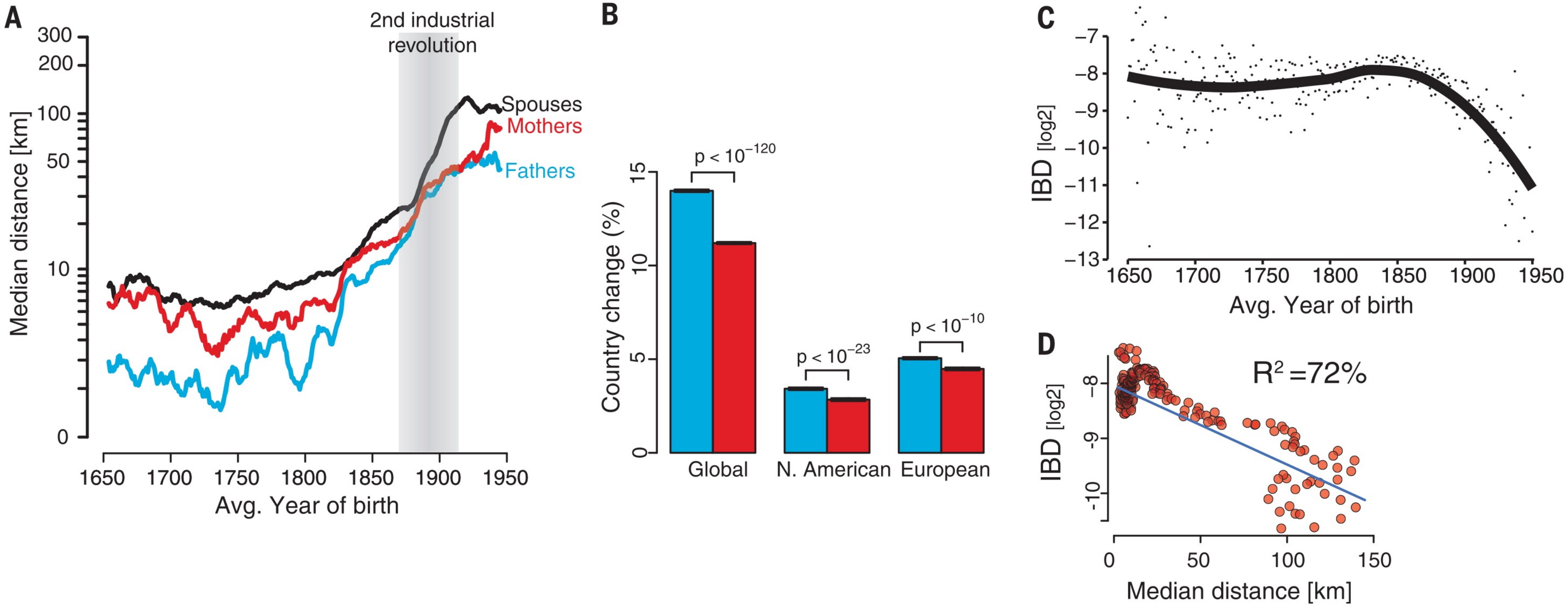
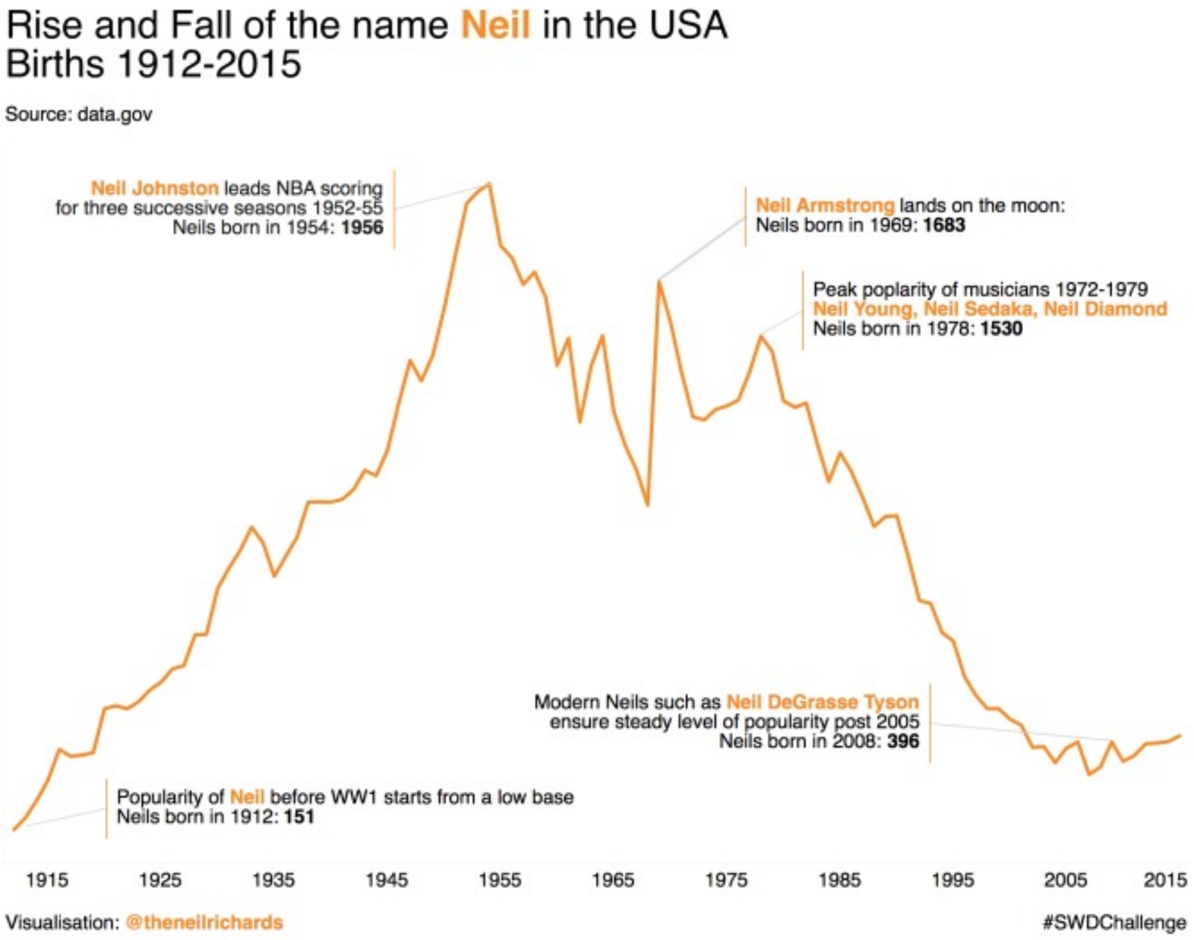


Fig. 4. Analysis of familial dispersion. (**A**) Median distance [$\log_{10}(x + 1)$] of father-offspring places of birth (cyan), mother-offspring (red), and marital radius (black) as a function of time (average year of birth). (**B**) Rate of change in the country of birth for father-offspring (cyan) or mother-offspring (red) stratified by major geographic areas. (**C**) Average IBD ($\log_2$) between couples as a function of average year of birth. Individual dots represent the measured average per year; the black line denotes the smooth trend using locally weighted regression. (**D**) IBD of couples as a function of marital radius. Each dot represents a year between 1650 to 1950. The blue line denotes the best linear regression line in log-log space.

Kaplanis et al., Quantitative analysis of population-scale family trees with millions of relatives, Science, 2018.

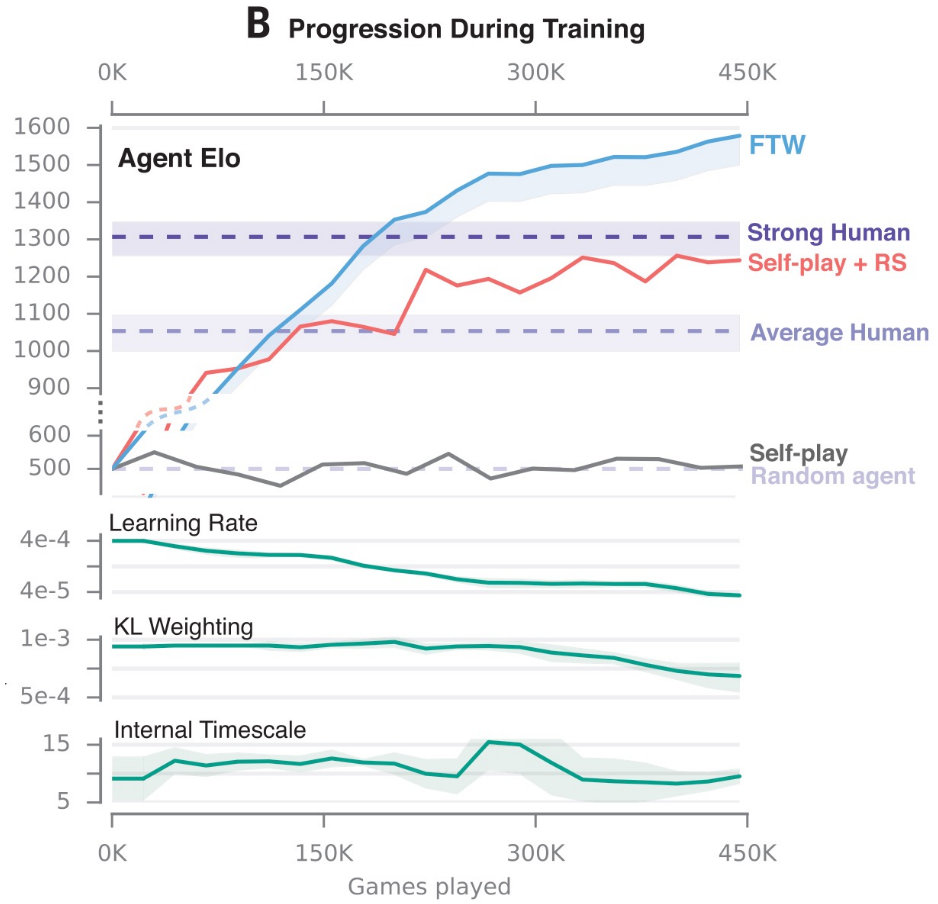
Contextualize your data



https://questionsindataviz.com/2018/01/06/is-white-space-always-your-friend/

→ Notable events

Contextualize your data



Comparison to humans

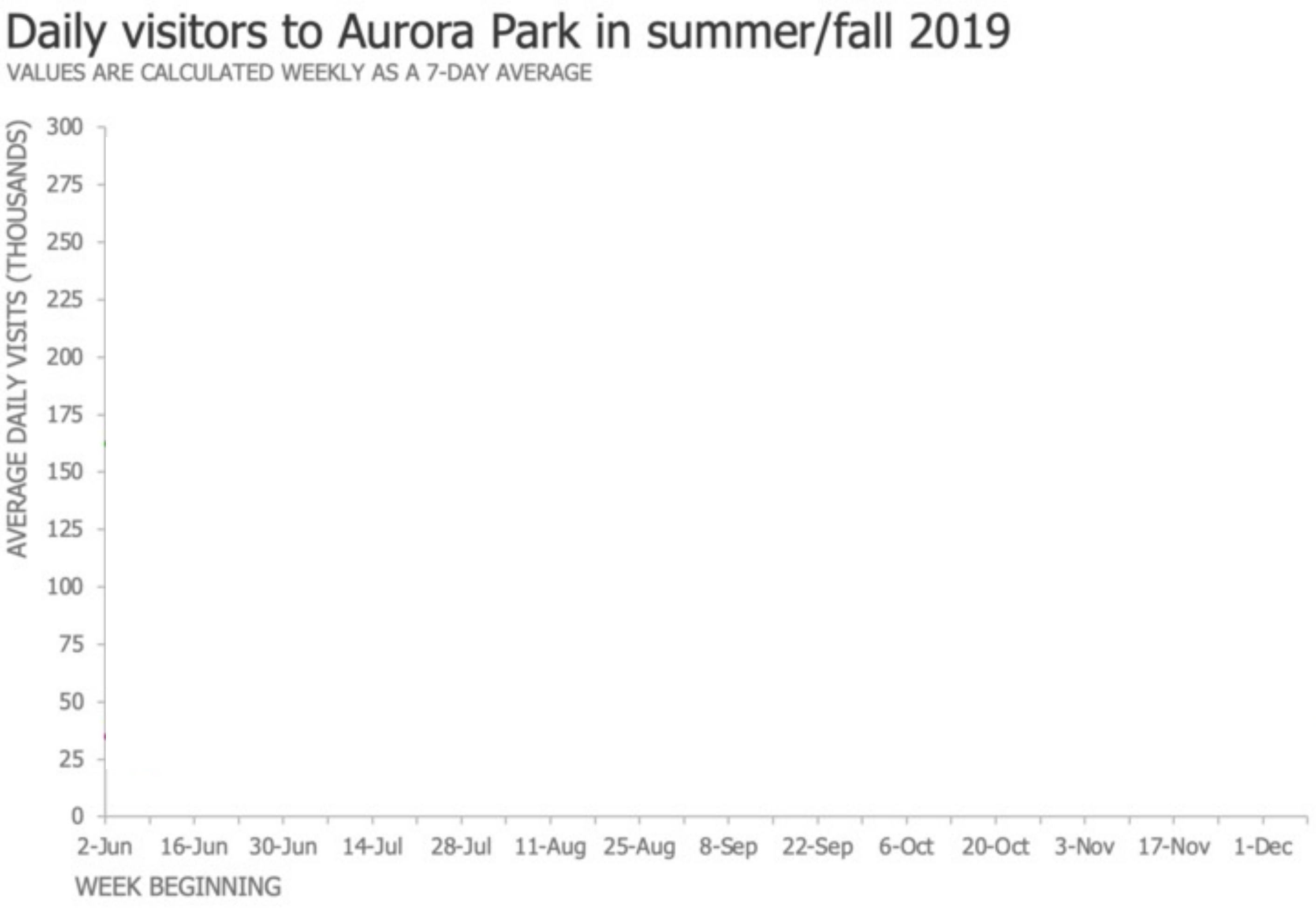
Jaderberg et al., Human-level performance in 3D multiplayer games with population-based reinforcement learning, Science, 2019.

Draw attention….really draw attention

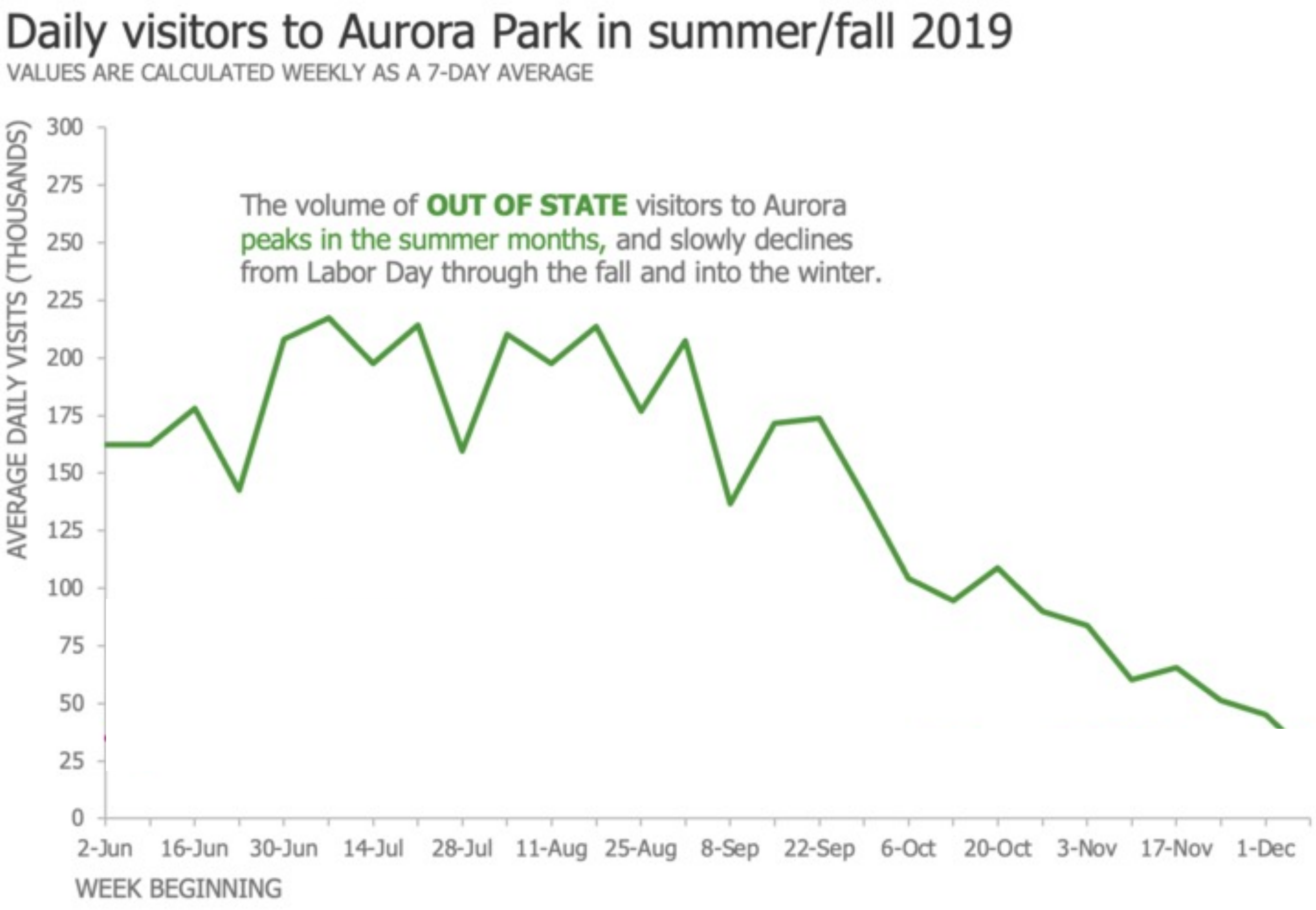
Daily visitors to Aurora Park in summer/fall 2019

VALUES ARE CALCULATED WEEKLY AS A 7-DAY AVERAGE

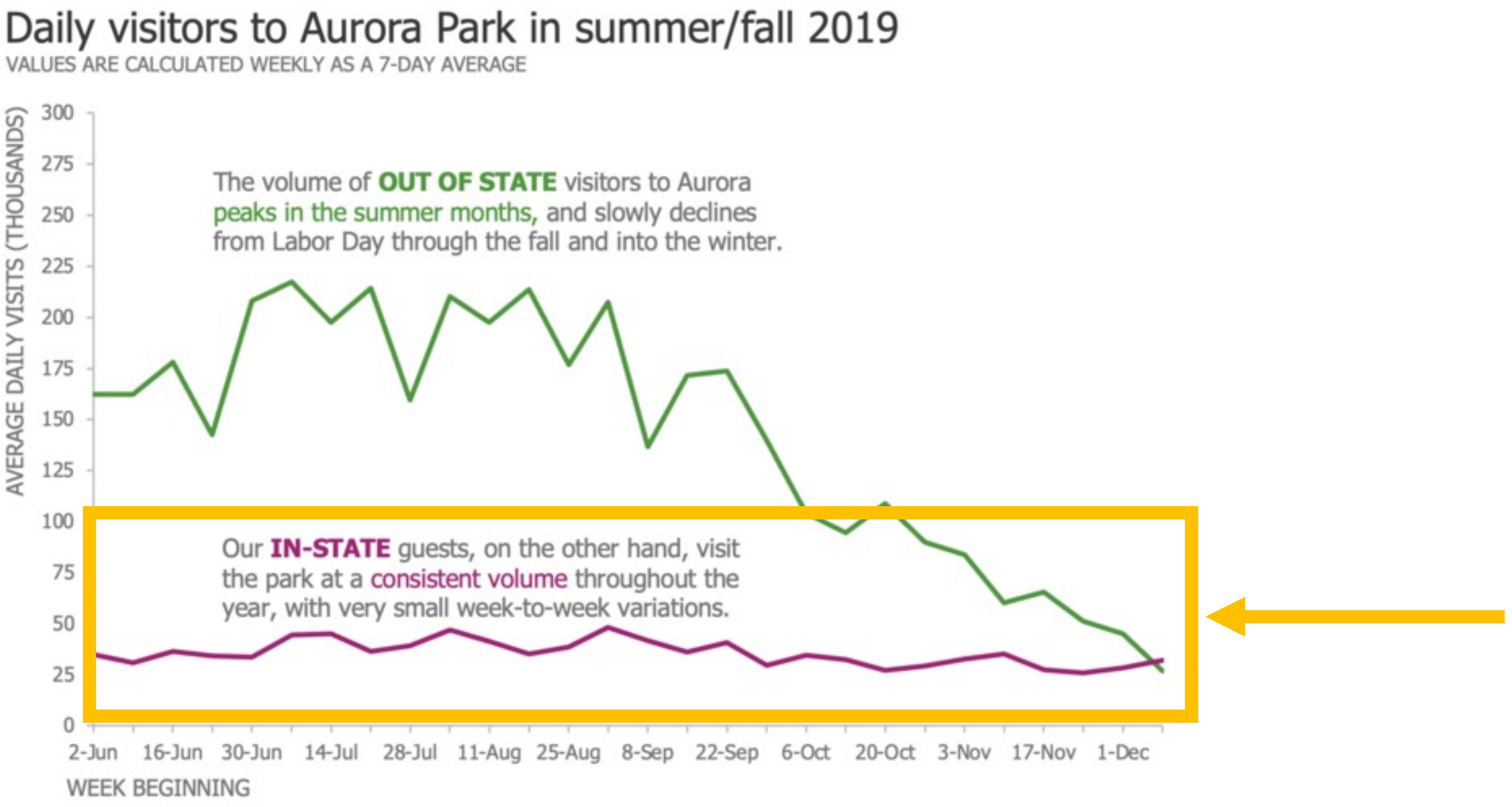
Draw attention….really draw attention



Draw attention….really draw attention



Draw attention....really draw attention

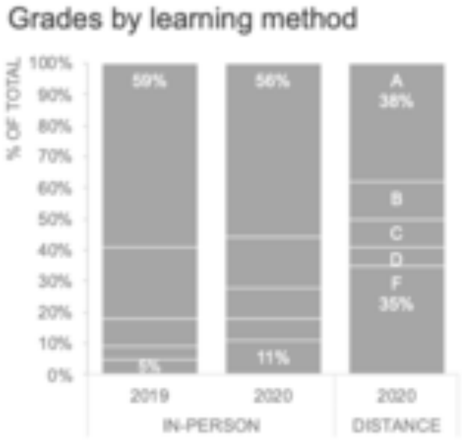


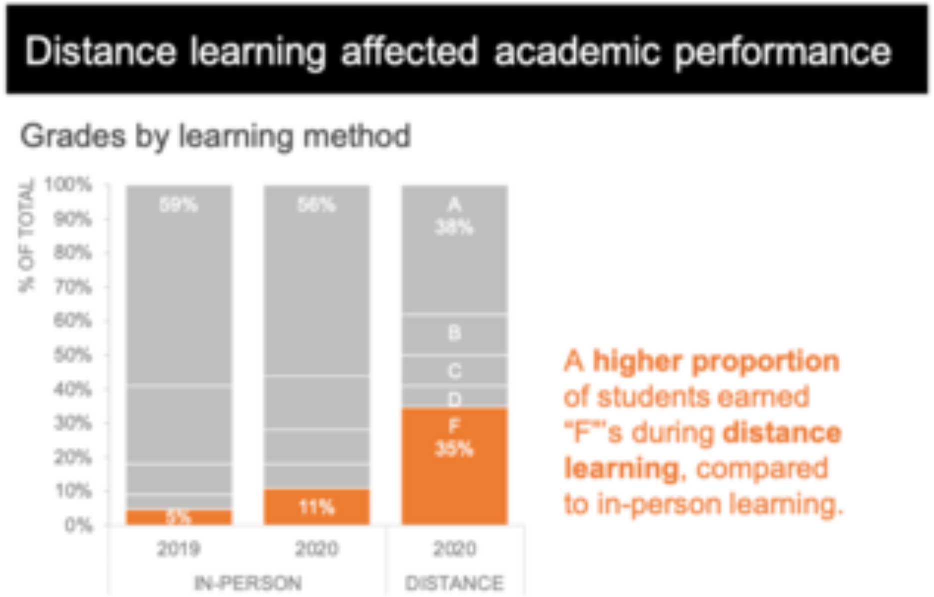
Design and redesign

Grades by learning method

GRADE EARNED	IN-PERSON 2019	IN-PERSON 2020	DISTANCE 2020
A	59%	56%	38%
B	23%	16%	12%
C	9%	10%	9%
D	4%	7%	6%
F	5%	11%	35%
TOTAL	100%	100%	100%

DATA SOURCE: The Times Record/Roane County Reporter | Feb 18, 2021
Compares high school student grade distribution for the second 9-week period of the term





Data source: *The Times Record/Roane County Reporter* | Feb 18, 2021
Compares high school student grade distribution for the second 9-week period of the term

https://www.storytellingwithdata.com/

Takeaways

Four essential components of a good visualization (information, story, goal, and visual form)

Design, redesign, and critique by redesign

Eliminate unnecessary clutter and noise

Use visualizations that are easy to read

Keep consistent order and alignment

Be thoughtful about the use of color

Contextualize your data

Draw attention to the key points about your data

5 Minute Break

Prototyping

Storytelling with your data

Goal: Design a key visualization for your project!

NOTE: Pretend you have all the data you want

Ideate (5 minutes) – within project group

Brainstorm key points you want to communicate from your analysis

Design (10 minutes) – individually

Set the context (audience, context, key point)

Design a visualization and its variations

Critique (5 minutes) – individually

Find someone from another group to swap your designs with

Pretend you're the target audience

Evaluate their design and provide redesign ideas

Storytelling with your data

Discuss how it went:

- Who was your audience?
- What point were you trying to make?
- What worked and didn't work?

What challenges did you encounter in your design?

What compromises did you have to make?

Did your audience "get" your design? why or why not?

What redesign recommendations did you give/receive?

Visualization for Papers

How do you create effective figures for scientific papers?

Why do figures matter?

Figures are often the first part of research papers examined by editors and your peers

Informative and well-designed figures:

- Convey facts, ideas, and relationships far more clearly and concisely than text
- Provide a means for discovering/quantifying patterns, trends, and comparisons
- Help the audience better understand the objective and results of your research

Why are figures difficult to design?

It doesn't come with you to explain it

It's the first thing people look at with zero context/background

There's no animation or interactivity

Design space is limited

Different types of visual structure

Interdisciplinary journal papers:

- Nature, Science, PNAS, etc.
- The focus is on new **scientific insights** and demonstrating the importance of those insights to advance science

Core CS conference papers:

- KDD, WebConf, NeurIPS, ICML, ICLR, AAAI, etc.
- The focus is on the development of **new methods** and their evaluation and comparison on benchmark datasets

Interdisciplinary journal papers

Figure 1: Dataset, approach and key result

Impress your audience!

Figure 2: Key result, detailed and unpacked

Figure 3: Orthogonal evidence supporting results

Figure 4: Orthogonal evidence supporting results

Supplementary Figures: Methodological contributions, algorithms, robustness analyses

Core CS conference papers

Figure 1: Key methodological contribution

Focus on most important information

Impress your audience!

Is your method/system the fastest, the largest, the most accurate?

What is the hard problem that your method solves?

What makes your method different from related work?

Figure 2-3: Overview and algorithmic details

Inputs + Data transformation + Outputs

Show details about data transformations:

Graph convolutions, neural architectures, etc.

Figure 4+: Results

Impress your audience

Abstract

Supervised learning on molecules has incredible potential to be useful in chemistry, drug discovery, and materials science. Luckily, several promising and closely related neural network models invariant to molecular symmetries have already been described in the literature. These models learn a message passing algorithm and aggregation procedure to compute a function of their entire input graph. At this point, the next step is to find a particularly effective variant of this general approach and apply it to chemical prediction benchmarks until we either solve them or reach the limits of the approach. In this paper, we reformulate existing models into a single common framework we call Message Passing Neural Networks (MPNNs) and explore additional novel variations within this framework. Using MPNNs we demonstrate state of the art results on an important molecular property prediction benchmark; these results are strong enough that we believe future work should focus on datasets with larger molecules or more accurate ground truth labels.

Gilmer et al., Neural Message Passing for Quantum Chemistry, ICML, 2017.

Brag about the speed

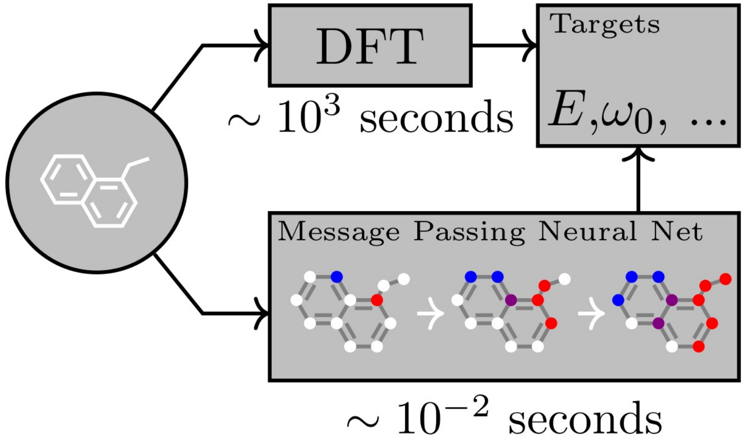


Figure 1. A Message Passi… …ural Network predicts quantum properties of an org… …e by modeling a computationally exp…

"Our method is so fast! Our paper should be published at ICML!"

Abstract

Large cascades can develop in online social networks as people share information with one another. Though simple reshare cascades have been studied extensively, the full range of cascading behaviors on social media is much more diverse. Here we study how *diffusion protocols*, or the social exchanges that enable information transmission, affect cascade growth, analogous to the way communication protocols define how information is transmitted from one point to another. Studying 98 of the largest information cascades on Facebook, we find a wide range of diffusion protocols – from cascading reshares of images, which use a simple protocol of tapping a single button for propagation, to the ALS Ice Bucket Challenge, whose diffusion protocol involved individuals creating and posting a video, and then nominating specific others to do the same. We find recurring classes of diffusion protocols, and identify two key counterbalancing factors in the construction of these protocols, with implications for a cascade's growth: the effort required to participate in the cascade, and the social cost of staying on the sidelines. Protocols requiring greater individual effort slow down a cascade's propagation, while those imposing a greater social cost of not participating increase the cascade's adoption likelihood. The predictability of transmission also varies with protocol. But regardless of mechanism, the cascades in our analysis all have a similar reproduction number (≈1.8), meaning that lower rates of exposure can be offset with higher per-exposure rates of adoption. Last, we show how a cascade's structure can not only differentiate these protocols, but also be modeled through branching processes. Together, these findings provide a framework for understanding how a wide variety of information cascades can achieve substantial adoption across a network.

Cheng et al., Do Diffusion Protocols Govern Cascade Growth?, ICWSM, 2018.

Brag about the data size

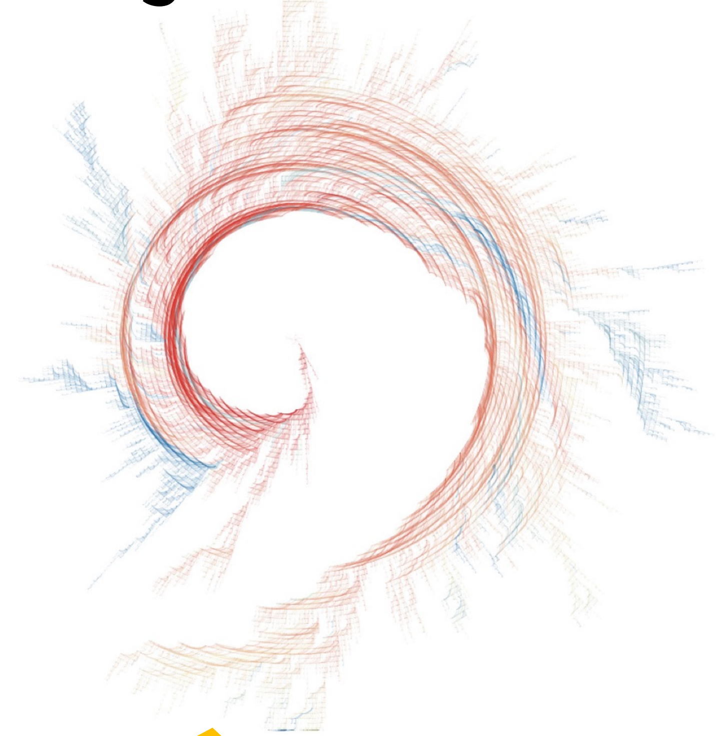


Figure 1: The d… n tree of a cascade with a volunteer diffusion p… individuals posted music from an

"Cascades can be so large! Despite that, we know how to study them! Our paper should be published at ICWSM!"

ABSTRACT

Cascades of information-sharing are a primary mechanism by which content reaches its audience on social media, and an active line of research has studied how such cascades, which form as content is reshared from person to person, develop and subside. In this paper, we perform a large-scale analysis of cascades on Facebook over significantly longer time scales, and find that a more complex picture emerges, in which many large cascades *recur*, exhibiting multiple bursts of popularity with periods of quiescence in between. We characterize recurrence by measuring the time elapsed between bursts, their overlap and proximity in the social network, and the diversity in the demographics of individuals participating in each peak. We discover that content virality, as revealed by its initial popularity, is a main driver of recurrence, with the availability of multiple copies of that content helping to spark new bursts. Still, beyond a certain popularity of content, the rate of recurrence drops as cascades start exhausting the population of interested individuals. We reproduce these observed patterns in a simple model of content recurrence simulated on a real social network. Using only characteristics of a cascade's initial burst, we demonstrate strong performance in predicting whether it will recur in the future.

Keywords: Cascade prediction; content recurrence; information diffusion; memes; virality.

Cheng et al., Do Cascades Recur?, WWW, 2016.

Answer the question in title

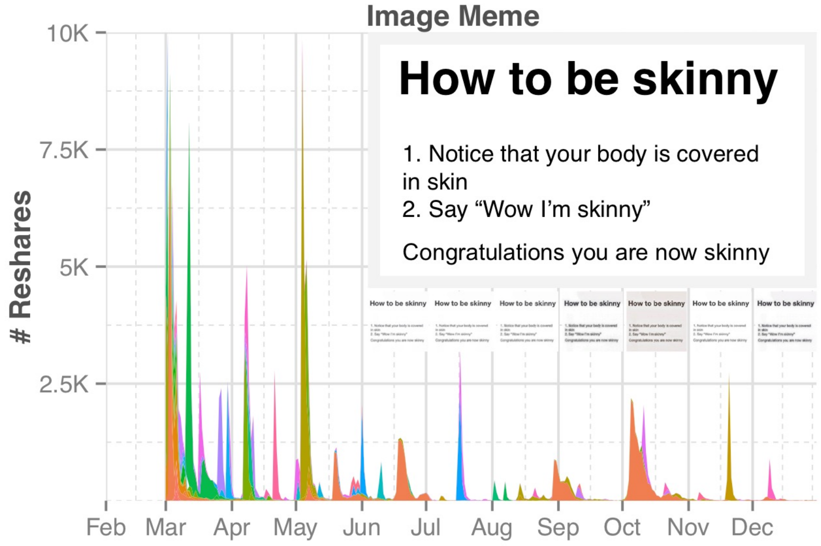


Figure 1: An example of a image meme that has recurred, or resurfaced in popularity multiple times, sometimes as a continuation of the same copy, and sometimes as a new copy of the same meme (example copies are shown as thumbnails). This recurrence appears as multiple peaks in the plot of reshares as a function of time.

"Cascades can be so complex! Despite that, we know how to study them! Our paper should be published at WWW!"

ABSTRACT

Deep learning models for graphs have achieved strong performance for the task of node classification. Despite their proliferation, currently there is no study of their robustness to adversarial attacks. Yet, in domains where they are likely to be used, e.g. the web, adversaries are common. Can deep learning models for graphs be easily fooled? In this work, we introduce the first study of adversarial attacks on attributed graphs, specifically focusing on models exploiting ideas of graph convolutions. In addition to attacks at test time, we tackle the more challenging class of poisoning/causative attacks, which focus on the training phase of a machine learning model. We generate adversarial perturbations targeting the *node's features* and the *graph structure*, thus, taking the dependencies between instances in account. Moreover, we ensure that the perturbations remain *unnoticeable* by preserving important data characteristics. To cope with the underlying discrete domain we propose an efficient algorithm NETTACK exploiting incremental computations. Our experimental study shows that accuracy of node classification significantly drops even when performing only few perturbations. Even more, our attacks are transferable: the learned attacks generalize to other state-of-the-art node classification models and unsupervised approaches, and likewise are successful even when only limited knowledge about the graph is given.

Make a statement about the problem

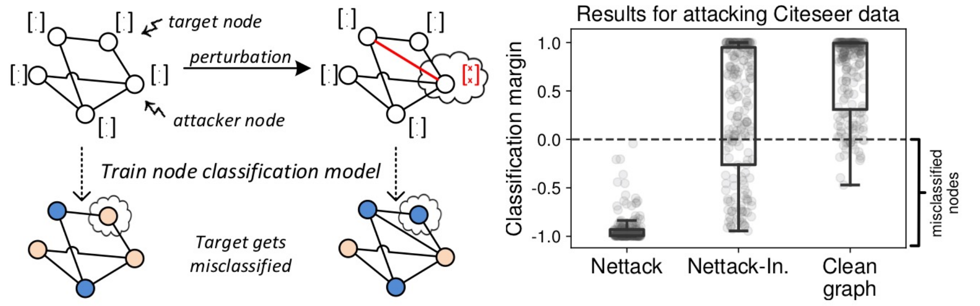


Figure 1: Small perturbations of the graph structure and node features lead to misclassification of the target.

"Yes, graph-based models for deep learning can be easily fooled. Here we show how devastating attacks can be."

Zugner et al., Adversarial Attacks on Neural Networks for Graph Data, KDD, 2018. (Best paper award)

Practical guidelines

Sketch low-fidelity prototypes of your visualization

Understand visual hierarchy, prioritize information, group/categorize

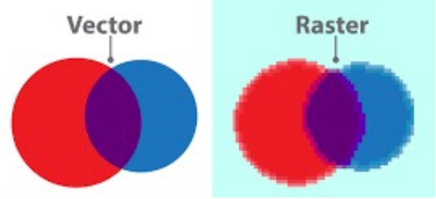
Save raw data and results to a tsv/csv/binary file

Your figures will need **multiple rounds of editing**

Read in the data and design figures

You may need multiple tools to draw a figure

Practical guidelines



Save figures as PDF or other vector formats

Raster images:

- Can't be dramatically resized (pixilation, distortion issues)
- When saved, they cannot be reopened and edited!

Vector images (e.g., PDF, EPS, AI, SVG):

- Remain editable!
- You can open them in Illustrator and edit text or any other element within the graphic
- Can be converted to a raster image but not vice-versa
- plt.savefig('myfig.pdf')

Only use raster format for web, Github repo, etc.

Bad Visualization

How do people misuse visualizations?

Superpower of visualization

When applied effectively to promote data exploration, analysis, and insight, we will experience what Joseph Berkson called "**interocular traumatic impact**: a conclusion that hits us between the eyes."

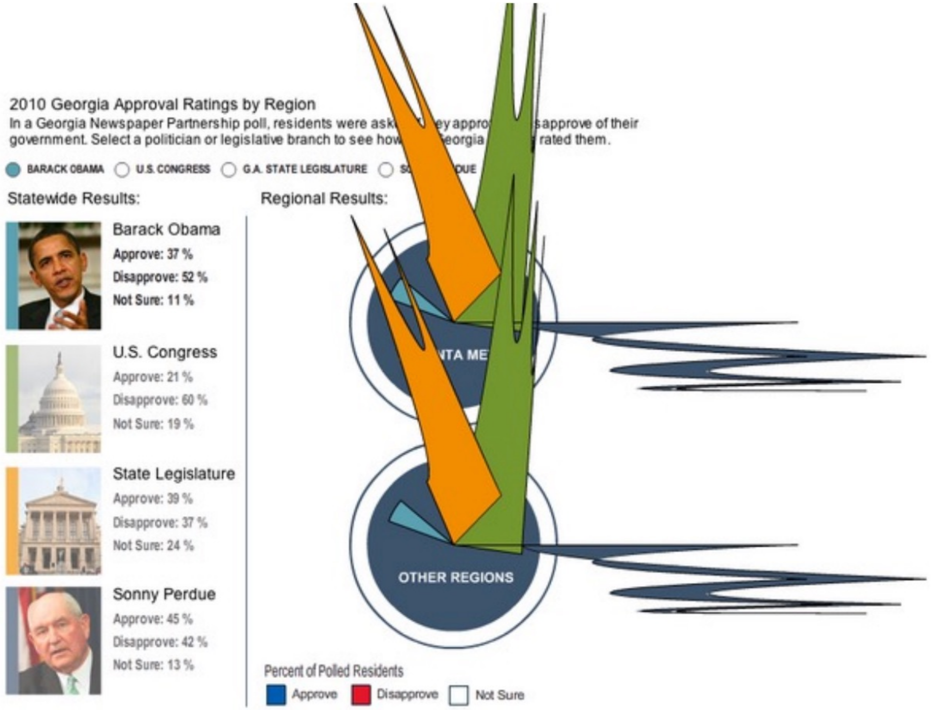
- Cleveland 1993

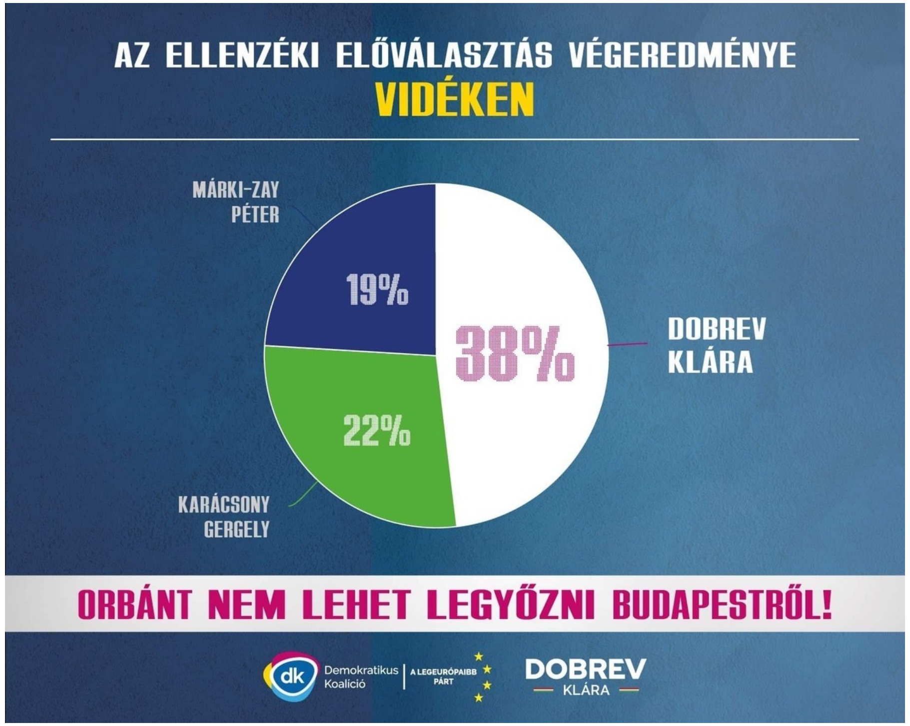
Empower understanding of data and analysis processes

Thou shall not create bad visualizations

Incorrect visualizations

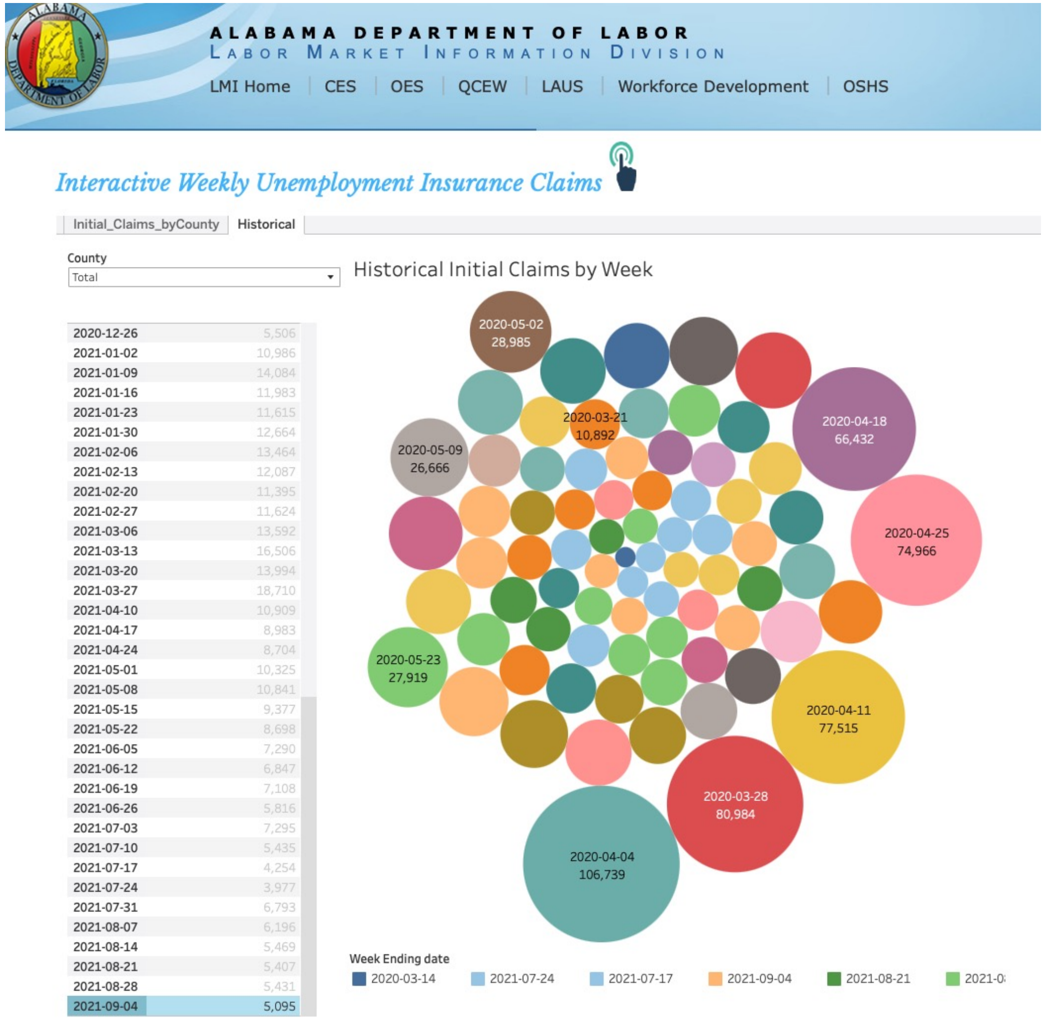
Bugs bugs bugs





Illegible visualizations

Plenty more at https://viz.wtf/



Deceptive visualizations

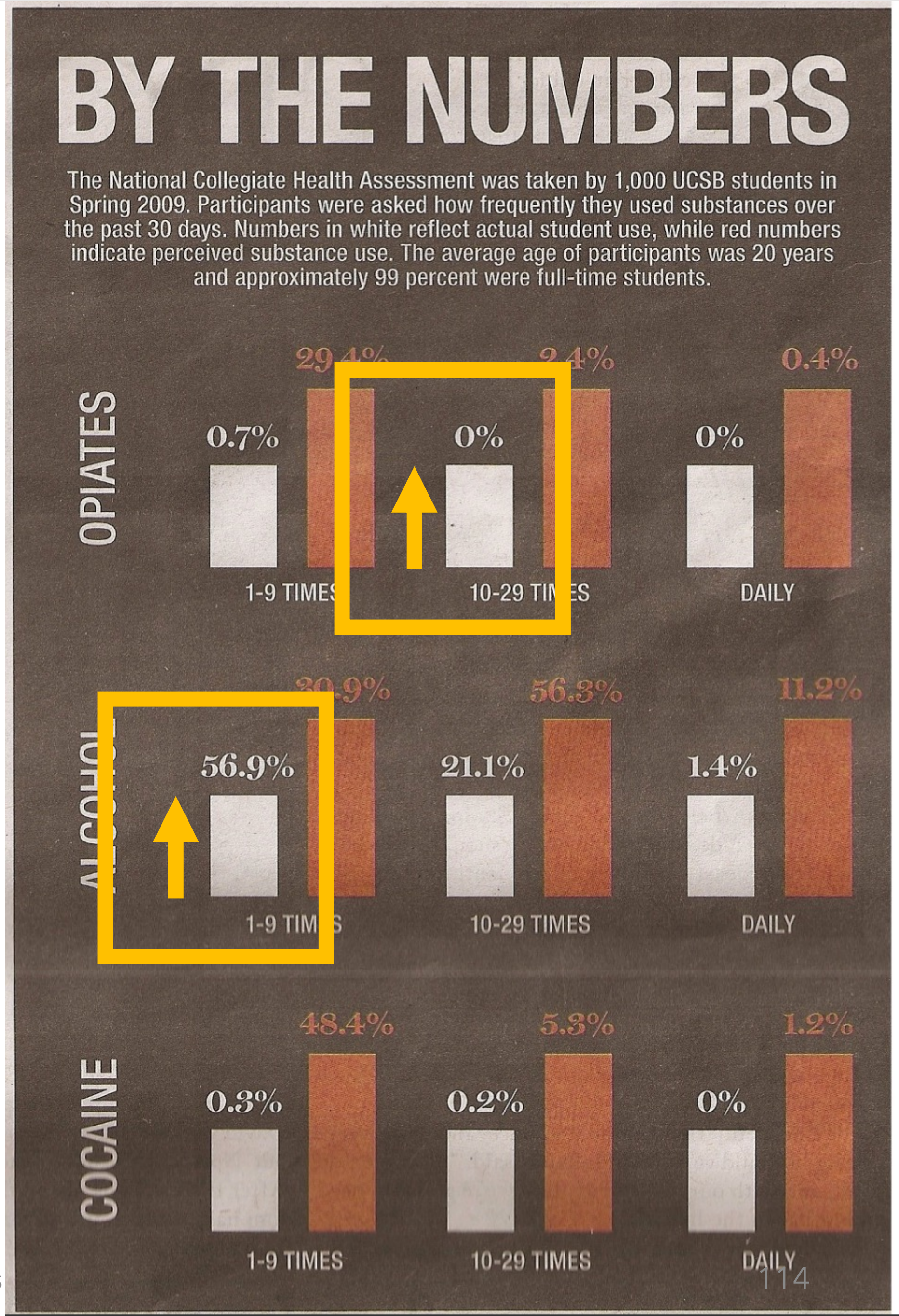
Lie factor

Scale manipulation

Convention manipulation

Lie factor

The size of the effect shown in the graphic should correspond to the size of the effect in the data



Lie factor

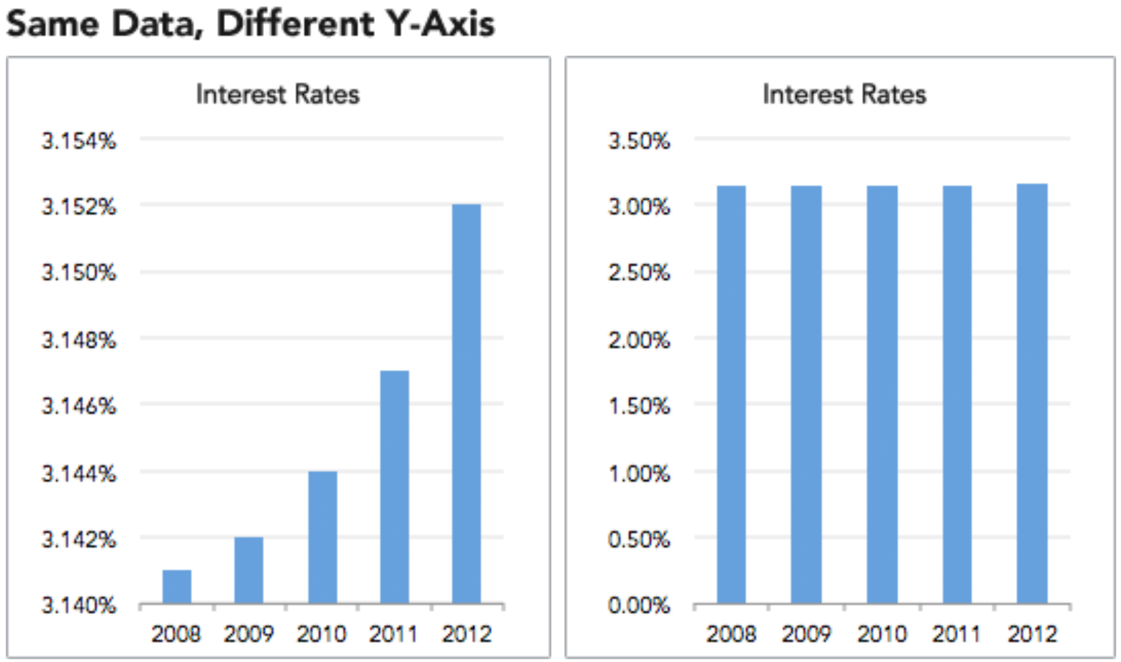

UNEMPLOYMENT RATE
UNDER PRESIDENT OBAMA
10%
9.5%
9%
8.5%
8%
9.0%
8.9%
8.8%
9.1%
9.2%
9.1%
9.1%
9.1%
8.6%
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV

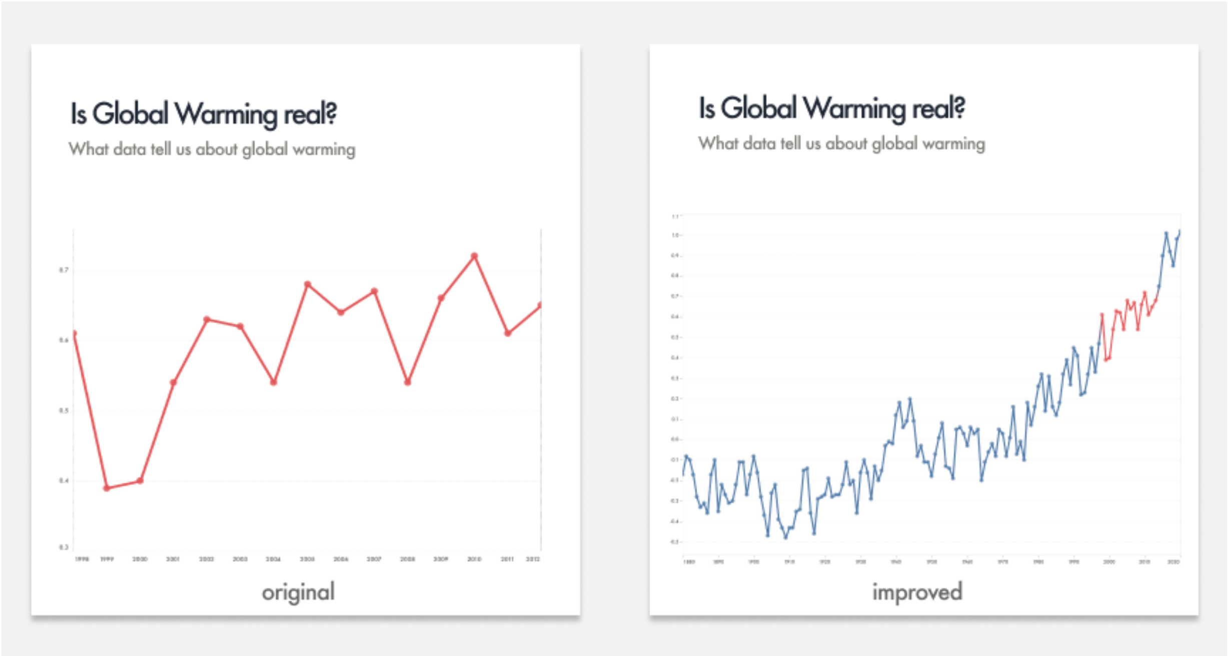
Lie factor


UNEMPLOYMENT RATE
UNDER PRESIDENT OBAMA
10%
9.5%
9%
8.5%
8%
9.0%
8.9%
8.8%
9.0%
9.1%
9.2%
9.1%
9.1%
9.1%
9.0%
8.6%
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV

Scale manipulation

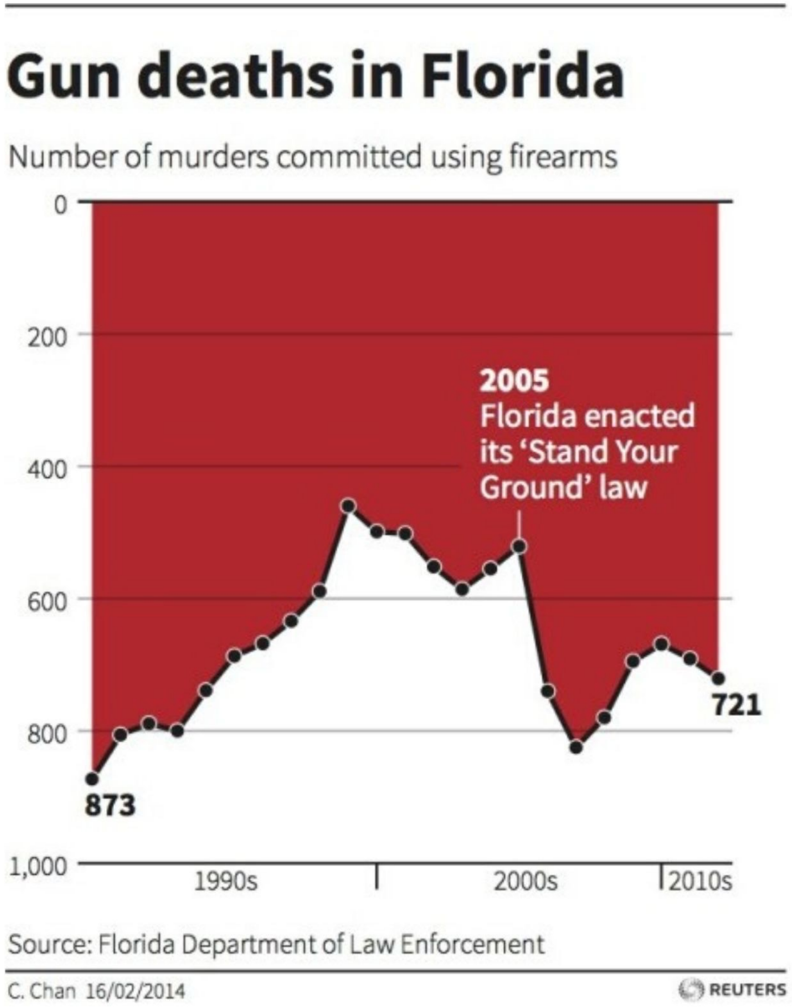
Changing with the scales of your chart to minimize, magnify, or invert the change in the data



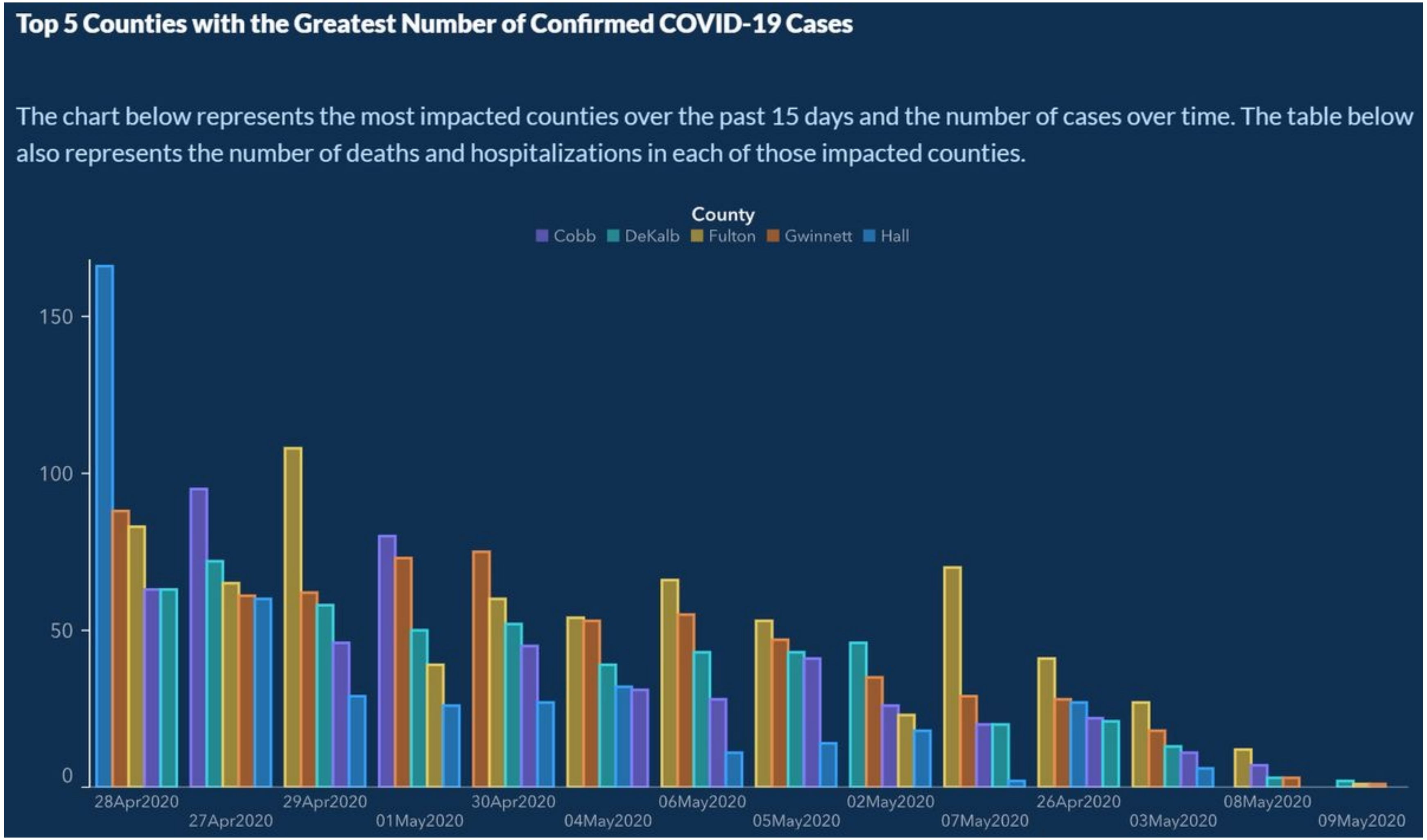


Convention manipulation

Breaking away from norms



Convention manipulation



Visualization Tools

Tools, software, and frameworks

Adobe Illustrator
  Adobe Creative Cloud
LaTeXiT
  chachatelier.fr/latexit
Matplotlib
  matplotlib.org
Seaborn
  seaborn.pydata.org
Bokeh
  bokeh.pydata.org
D3.js
  d3js.org
GeoPandas
  geopandas.org

Google Charts
  developers.google.com/chart
Circos
  circos.ca
gnuplut
  gnuplot.info
TikZ
  texample.net/tikz
Plotly
  plot.ly/python
missingno
  github.com/ResidentMario/missingno
billboard.js
  naver.github.io/billboard.js

Squaire.js
  wsj.github.io/squaire
Tableau
  tableau.com/
Vega
  vega.github.io/vega/
Vega-lite
  vega.github.io/vega-lite/
Altair
  altair-viz.github.io/

Adobe illustrator and alternatives

Where to get on campus:

For purchase: https://itconnect.uw.edu/wares/uware/adobe-creative-cloud/

Use for Free: UW Library https://www.lib.washington.edu/media/software

Free alternatives:

Inkscape, https://inkscape.org

GIMP, https://www.gimp.org

Boxy-SVG, https://boxy-svg.com

Leverage UX prototyping tools

- Adobe suite is a powerful prototyping tool, but it has a high learning curve and can be difficult to collaborate on with others
- There are online collaborative UX prototyping tools that can be used to prototype wireframes and flows
  - Figma: **free** for students and educators; can be exported as PDF, SVG, PNG

Convert JavaScript vis to figure

Three steps:

1) Use a JS library from two slide ago and generate a visualization

2) Generate a PDF file from HTML:
   - stackoverflow.com/questions/18191893/generate-pdf-from-html-in-div-using-javascript

3) Open the PDF in Illustrator and make further edits:
   - Change colors
   - Add labels and annotations
   - Add new visual elements, e.g., insets, logos
   - Combine with other graphics to get a multi-panel figure

Tools for network & relational data

- Gephi, gephi.org
- Graphviz, graphviz.org
- NetworkX, networkx.github.io
- JSNetworkX, jsnetworkx.org
- igraph, igraph.org/python
- sigma.js, sigmajs.org
- Cytoscape, cytoscape.org
- Hive plots, hiveplot.com

Where to get ideas for figures?

Papers published in last issues of Nature, Science, PNAS, Nature Methods, Nature Biotech, etc.

- No need to read the papers, just look at figures!

Martin Krzywinski, mkweb.bcgsc.ca

- Inventor of several popular visualization tools
- Designed many Nature, Science, etc. covers

www.d3-graph-gallery.com

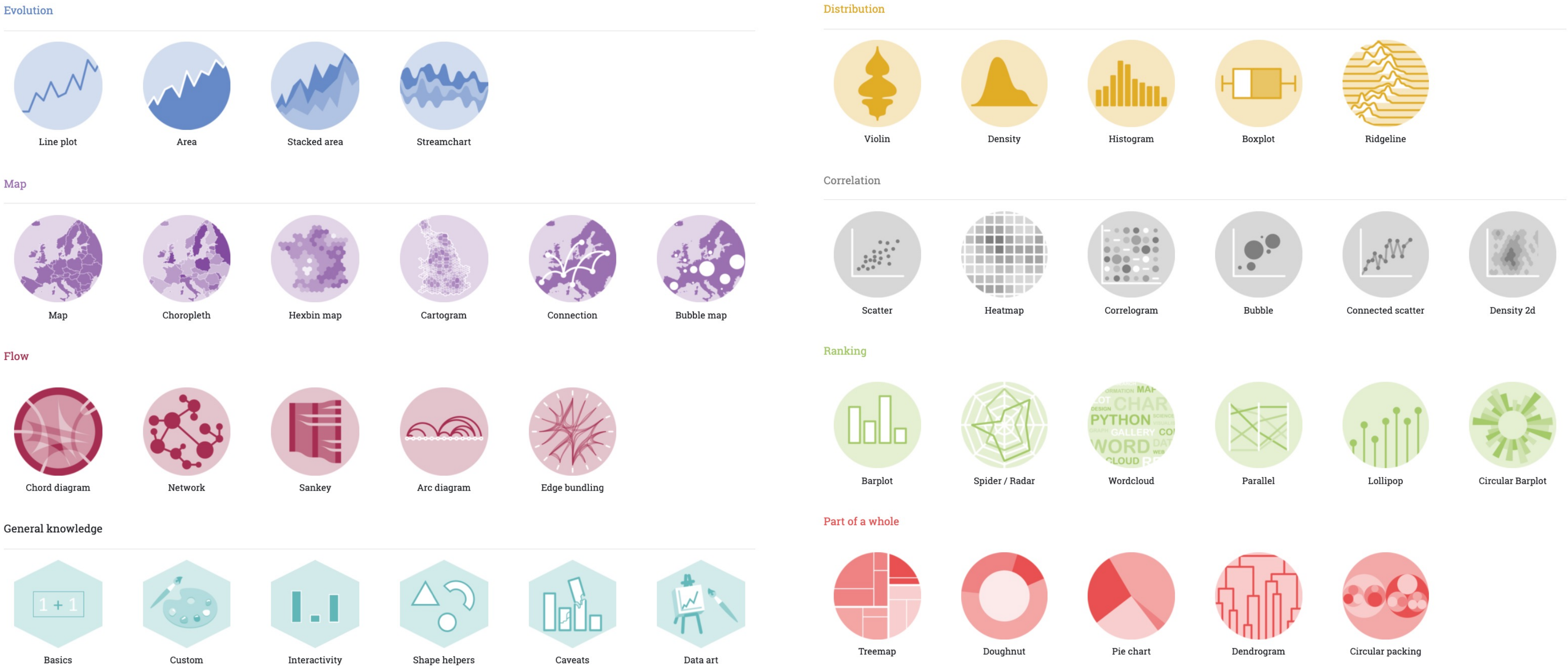
- Gallery with hundreds of chart, graphs, geo, part-of-whole
- Reproducible & editable source code!

developers.google.com/chart/interactive/docs/gallery

- Over 30 chart types, including many non-standard ones
- Tutorials and source code for every chart type!

Where to get ideas for figures?

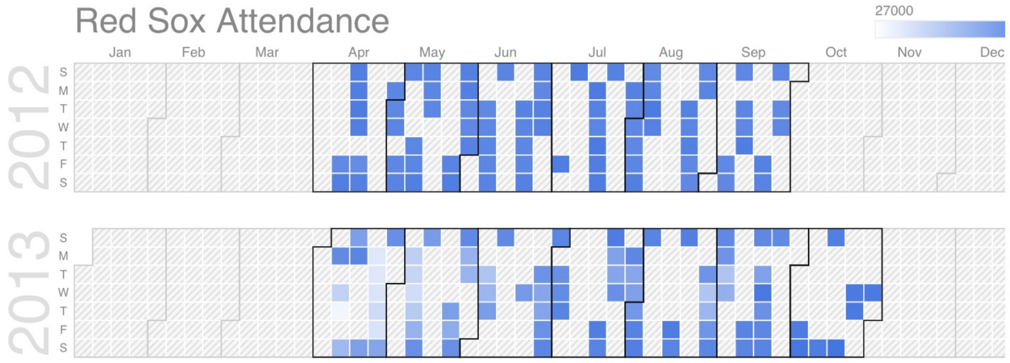
www.d3-graph-gallery.com
Many non-standard, but highly effective chart types. **Source code!**

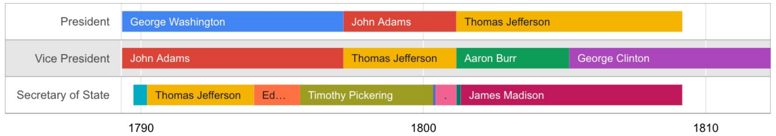


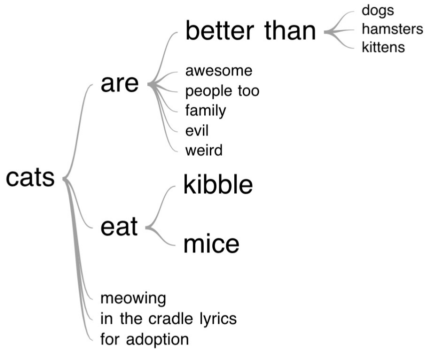
Where to get ideas for figures?

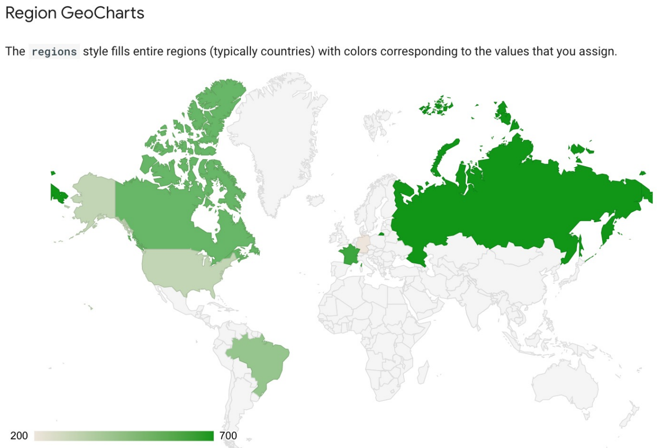
https://developers.google.com/chart with **source code!**

Chart Types
- Chart Gallery
- Annotation Charts
- Area Charts
- Bar Charts
- Bubble Charts
- Calendar Charts
- Candlestick Charts
- Column Charts
- Combo Charts
- Diff Charts
- Donut Charts
- Gantt Charts
- Gauge Charts
- GeoCharts
- Histograms
- Intervals
- Line Charts
- Maps
- Org Charts
- Pie Charts
- Sankey Diagrams
- Scatter Charts
- Stepped Area Charts
- Table Charts
- Timelines
- Tree Map Charts
- Trendlines
- Waterfall Charts
- Word Trees









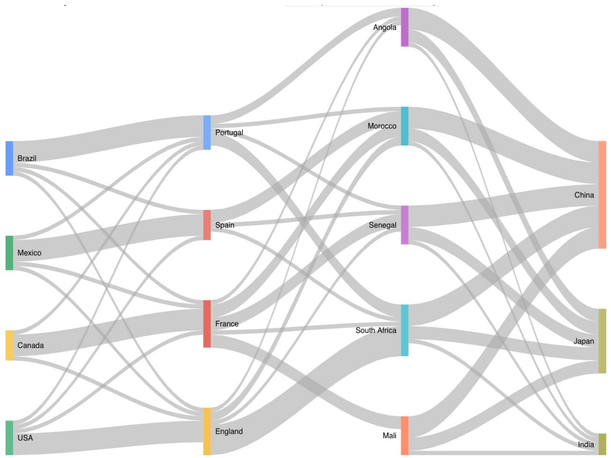


Chart guide

https://www.storytellingwithdata.com/chart-guide

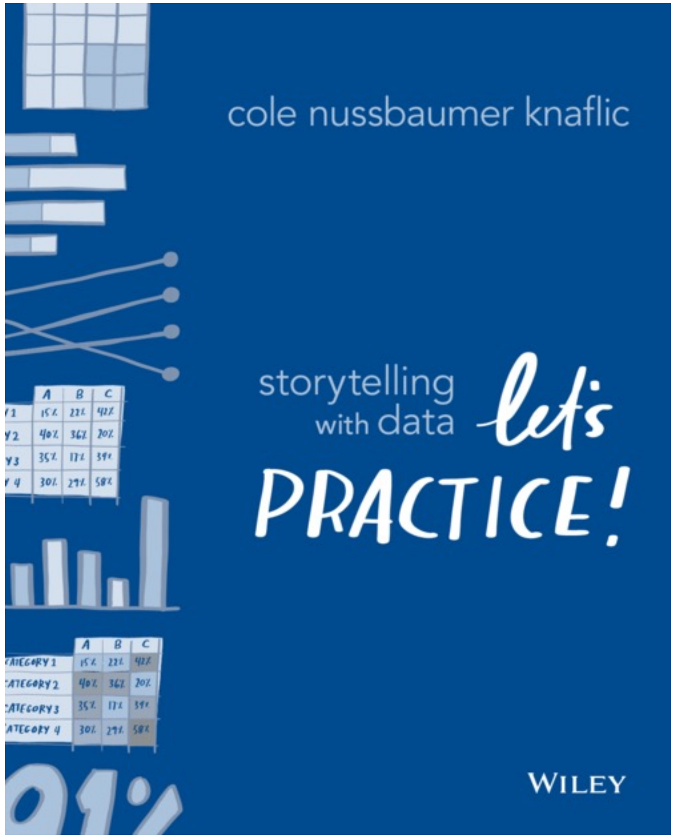
At *storytelling with data,* we encounter a ton of different graphs. Through our work, we've both learned strategies for effective application and identified common pitfalls (including some things to avoid!). In this guide, we share the good and the bad of commonly used charts and graphs for data communications.

Simply click on a graph below to learn more.

WHAT IS A LINE GRAPH? | WHAT IS A BAR CHART? | WHAT IS AN AREA GRAPH?

Practice visualization redesign

https://www.storytellingwithdata.com/letspractice/downloads



CHAPTER 3: identify & eliminate clutter

3.1: which Gestalt principles are in play? | data | see book for solution

3.2: how can we tie words to the graph? | data | solution

solutions in other tools:

Datawrapper | Flourish | Google Data Studio | PowerBI | Tableau

3.3: harness the power of alignment & white space | data | solution

3.4: declutter! | data | solution

3.5: which Gestalt principles are in play? | data

3.6: find an effective visual | see book

3.7: create alignment and use white space | data

3.8: declutter! | data

3.9: declutter (again!) | data

3.10: declutter (one more time!) | data

CHAPTER 4: focus attention

4.1: where are your eyes drawn? | see book for exercise & solution

Data visualization interactive notebooks

https://github.com/uwdata/visualization-curriculum

Table of Contents

1. Introduction to Vega-Lite / Altair
   Jupyter Book | Jupyter | Colab | Nextjournal | Observable | Deepnote
2. Data Types, Graphical Marks, and Visual Encoding Channels
   Jupyter Book | Jupyter | Colab | Nextjournal | Observable | Deepnote
3. Data Transformation
   Jupyter Book | Jupyter | Colab | Nextjournal | Observable | Deepnote
4. Scales, Axes, and Legends
   Jupyter Book | Jupyter | Colab | Nextjournal | Observable | Deepnote
5. Multi-View Composition
   Jupyter Book | Jupyter | Colab | Nextjournal | Observable | Deepnote
6. Interaction
   Jupyter Book | Jupyter | Colab | Nextjournal | Observable | Deepnote
7. Cartographic Visualization
   Jupyter Book | Jupyter | Colab | Nextjournal | Observable | Deepnote

Seaborn tutorial

https://bit.ly/cse481ds-seaborn-tutorial

Other resources

UW CSE 512 course materials:

https://courses.cs.washington.edu/courses/cse512/

Collaborative visualization tools:

https://observablehq.com/

Interactive visualization publications:

https://distill.pub/journal/

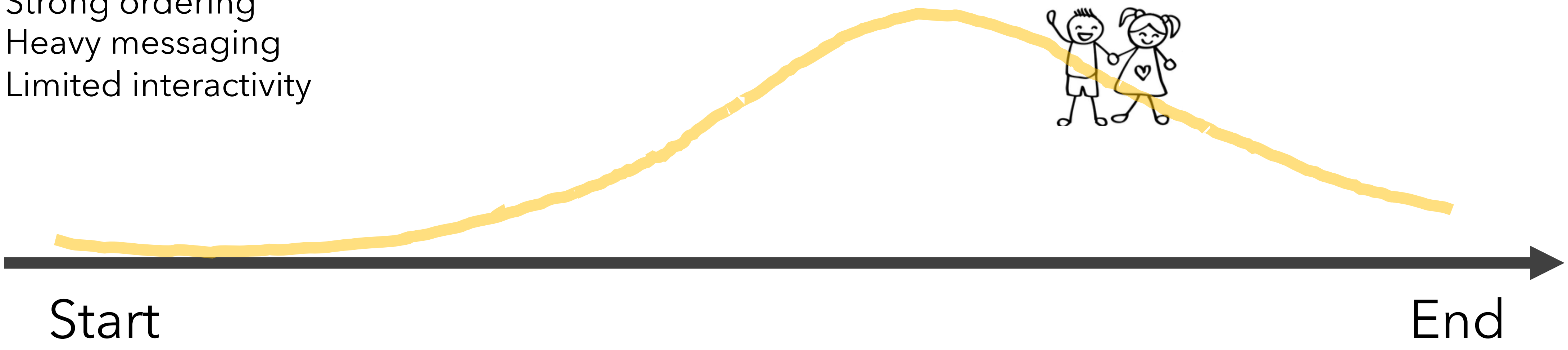
Extra

134

Narrative structure

Author-driven narratives

Strong ordering
Heavy messaging
Limited interactivity

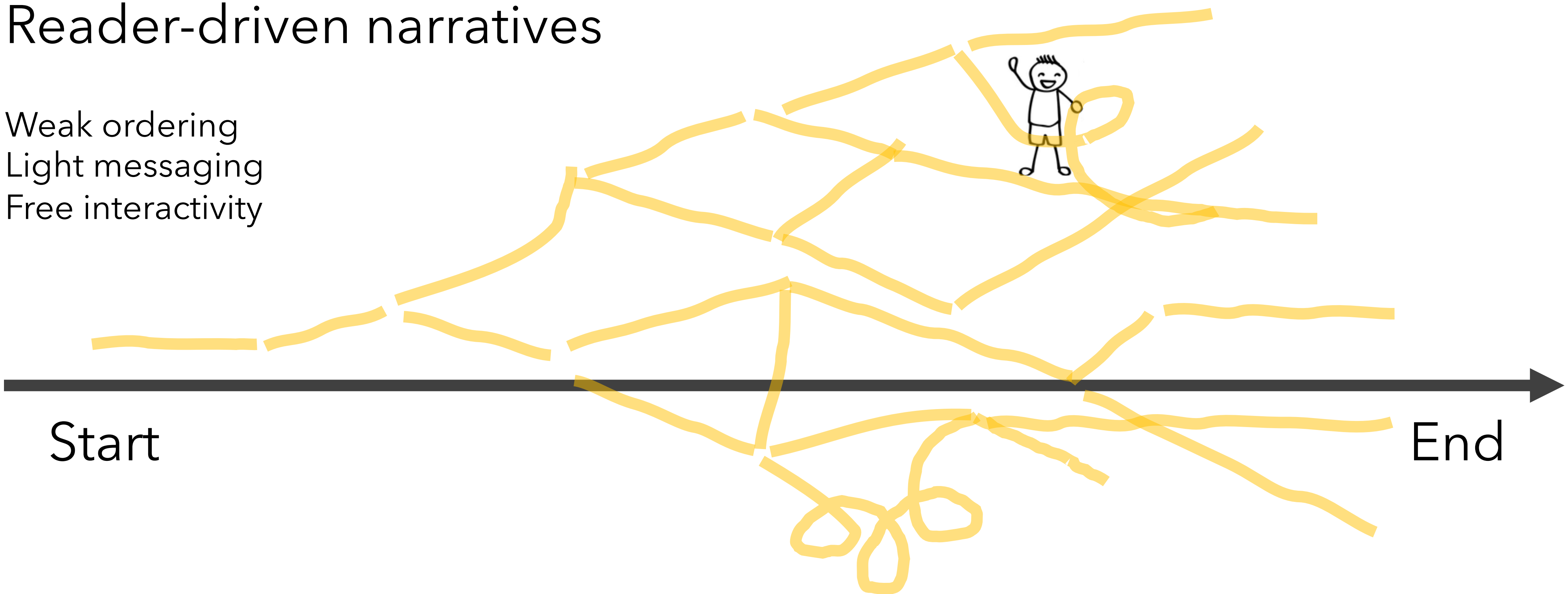


Segel & Heer, 2010

Narrative structure

Reader-driven narratives

Weak ordering
Light messaging
Free interactivity



Segel & Heer, 2010

Narrative structure

Author-driven	Reader-driven
Strong linear ordering Heavy messaging Limited interactivity	Weak ordering Light messaging Free interactivity
Tell stories Need for clarity and speed	Ask questions Explore and find
Most books	Choose your own adventure!

Segel & Heer, 2010

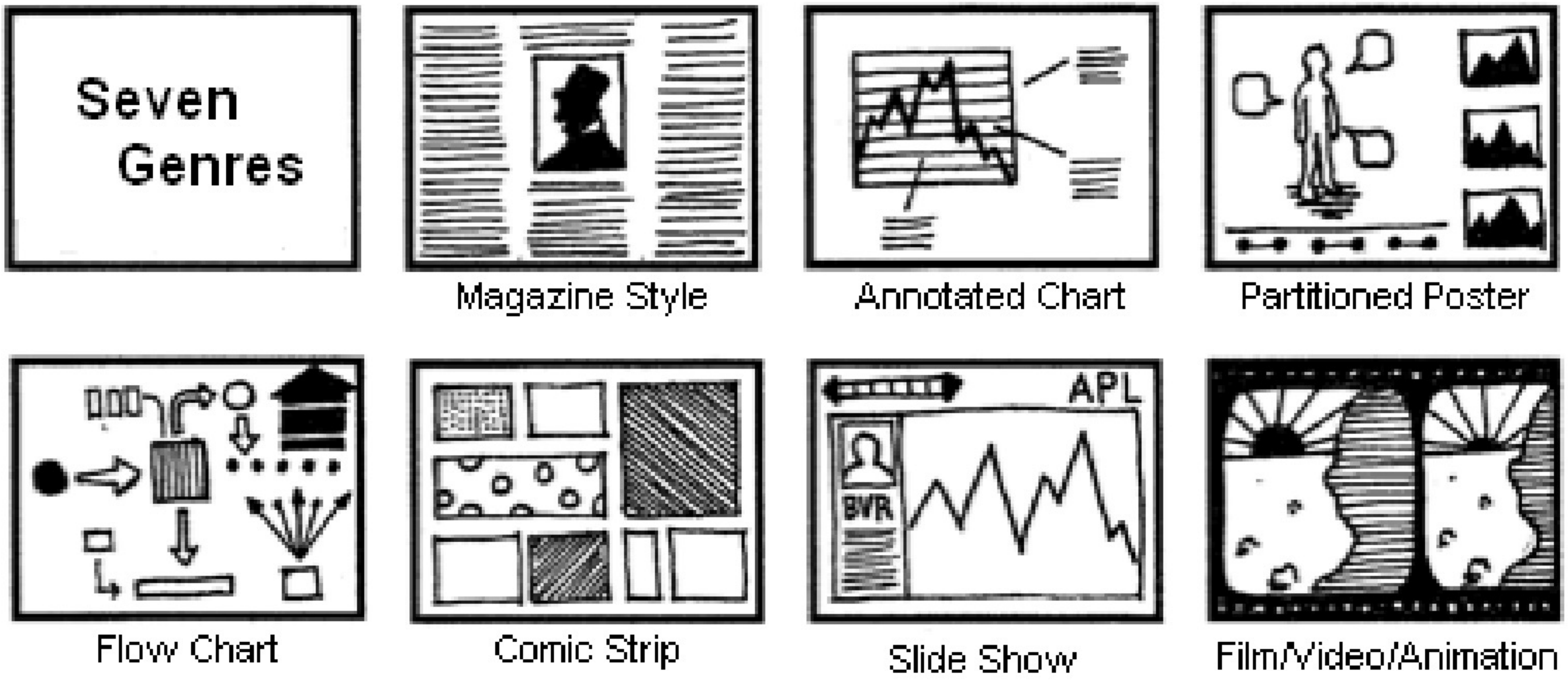
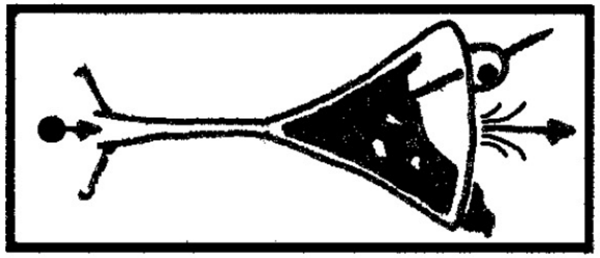


Fig. 8. Genres of Narrative Visualization.

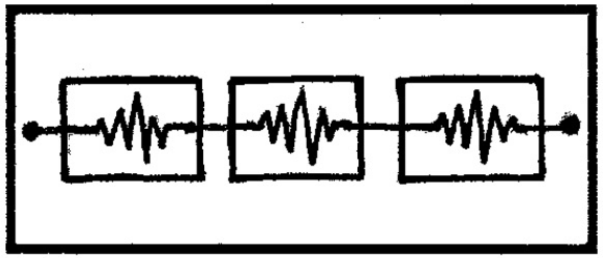
Segel & Heer, 2010

A little bit of both

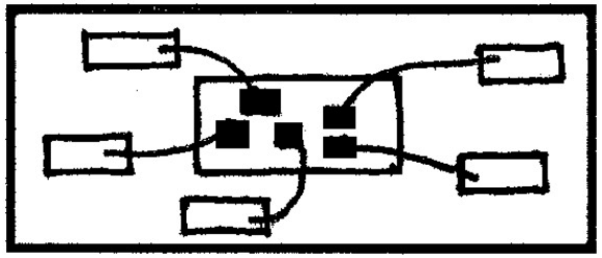
Author-driven ⟷ Reader-driven



Martini glass



Interactive slideshow

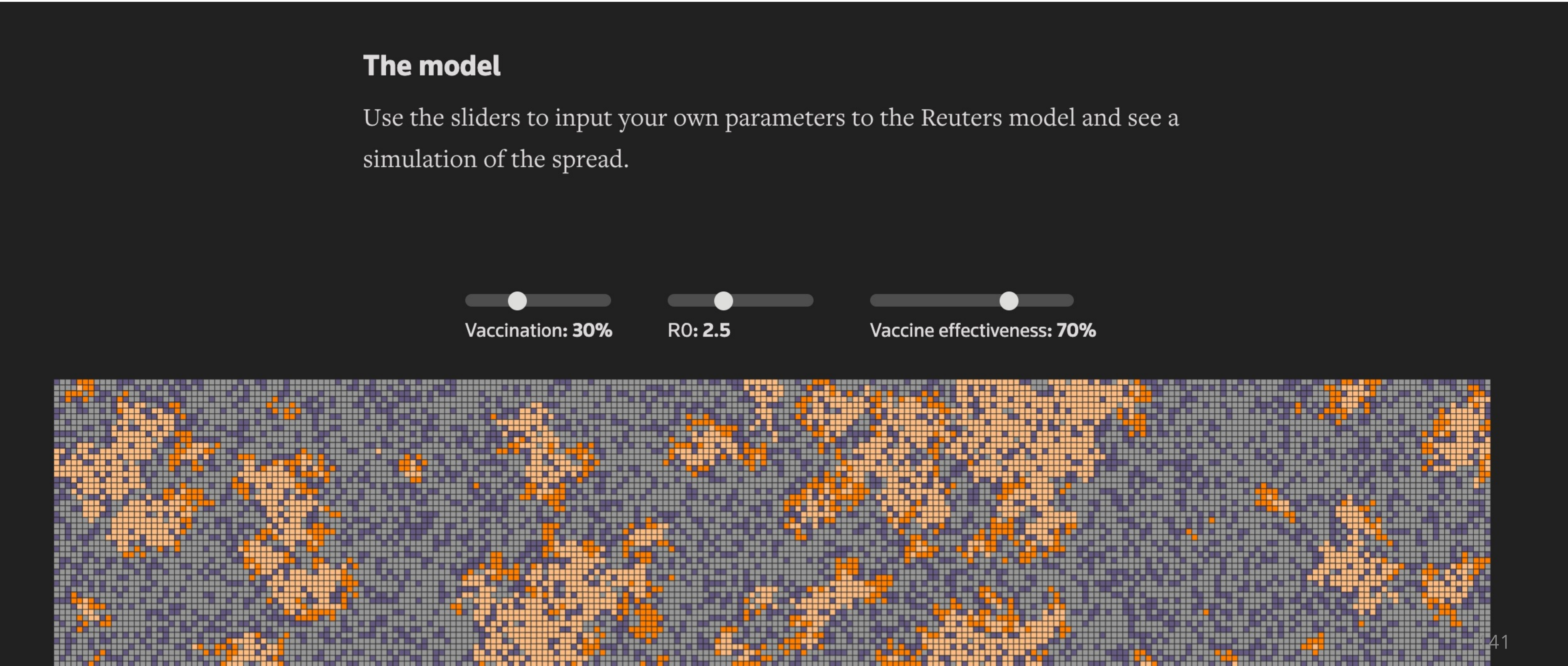


Drill-down story

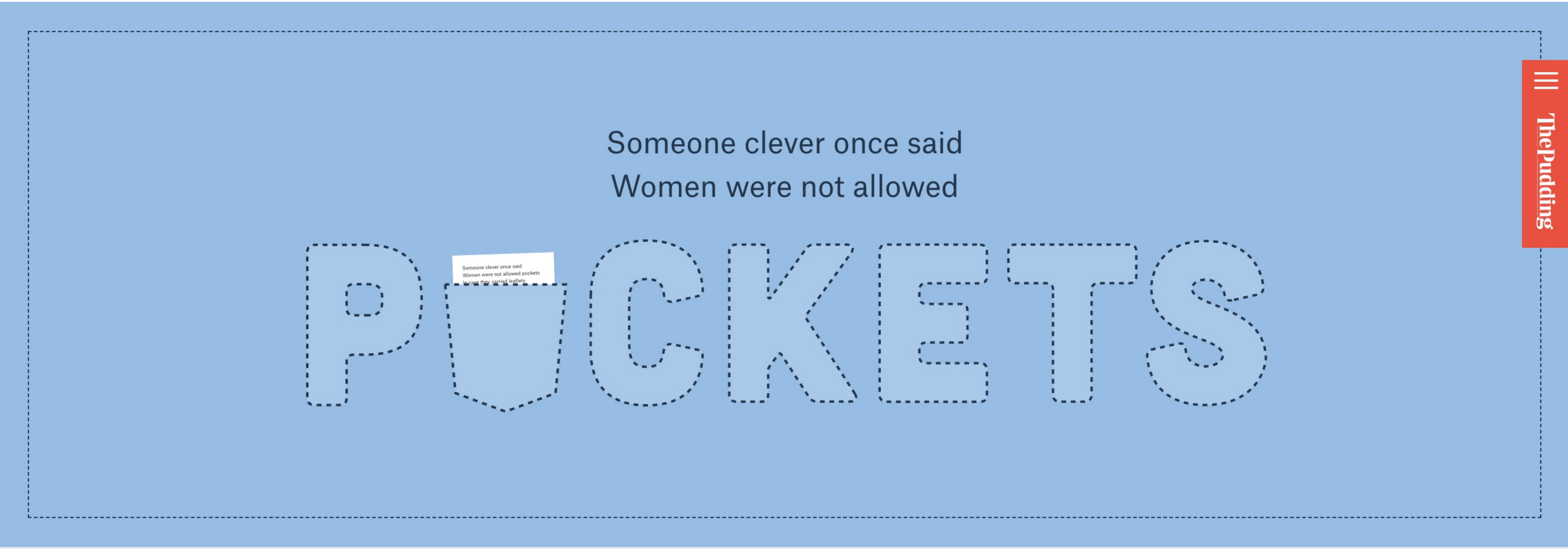
Segel & Heer, 2010

Martini glass

Martini glass



Scroller



https://pudding.cool/2018/08/pockets/

By **Jan Diehm** & **Amber Thomas**

August 2018

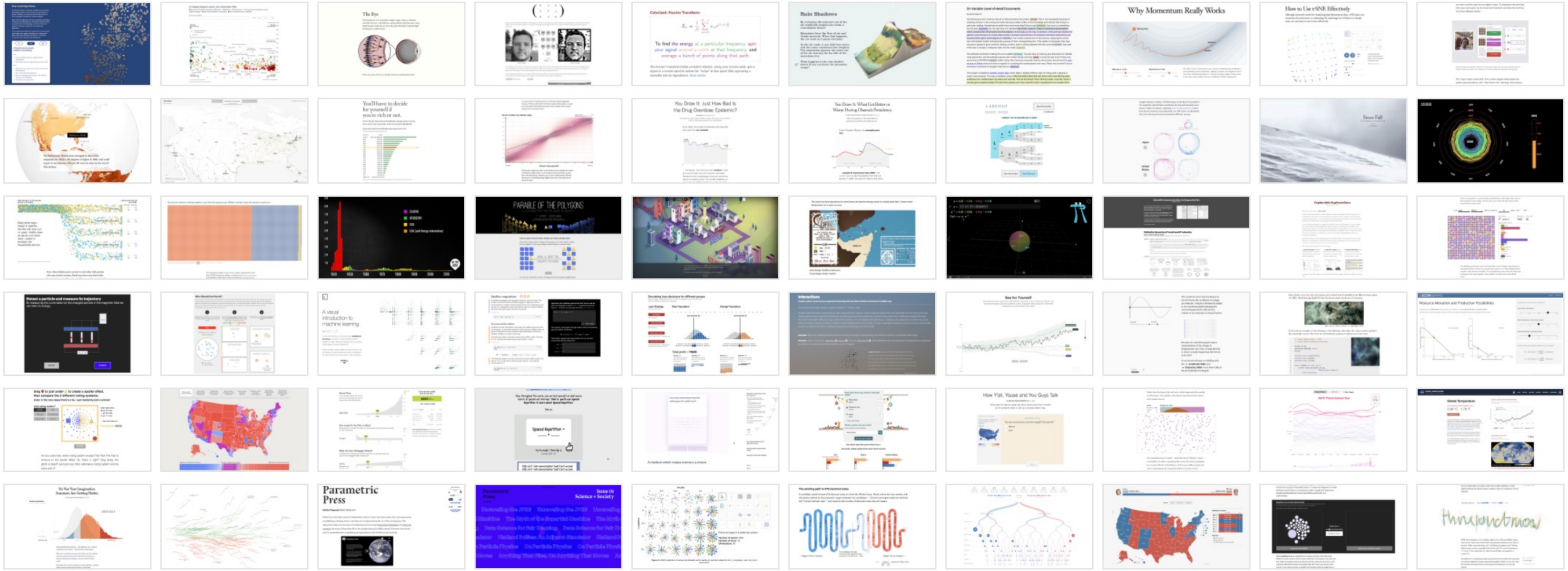
Slideshow

https://fivethirtyeight.com/features/gun-deaths/

Interactive articles

Communicating with Interactive Articles

Examining the design of interactive articles by synthesizing theory from disciplines such as education, journalism, and visualization.



https://distill.pub/2020/communicating-with-interactive-articles/

Thank you for your feedback!

https://bit.ly/cse481ds-feedback