

HOW TO MAKE DECISIONS IN DATA VISUALIZATION



Alberto Cairo

Banco de Portugal, 2020

Time spent in a life of a Data Scientist

@datavizzdom

Gulrez

Perception



Time spent in a life of a Data Scientist

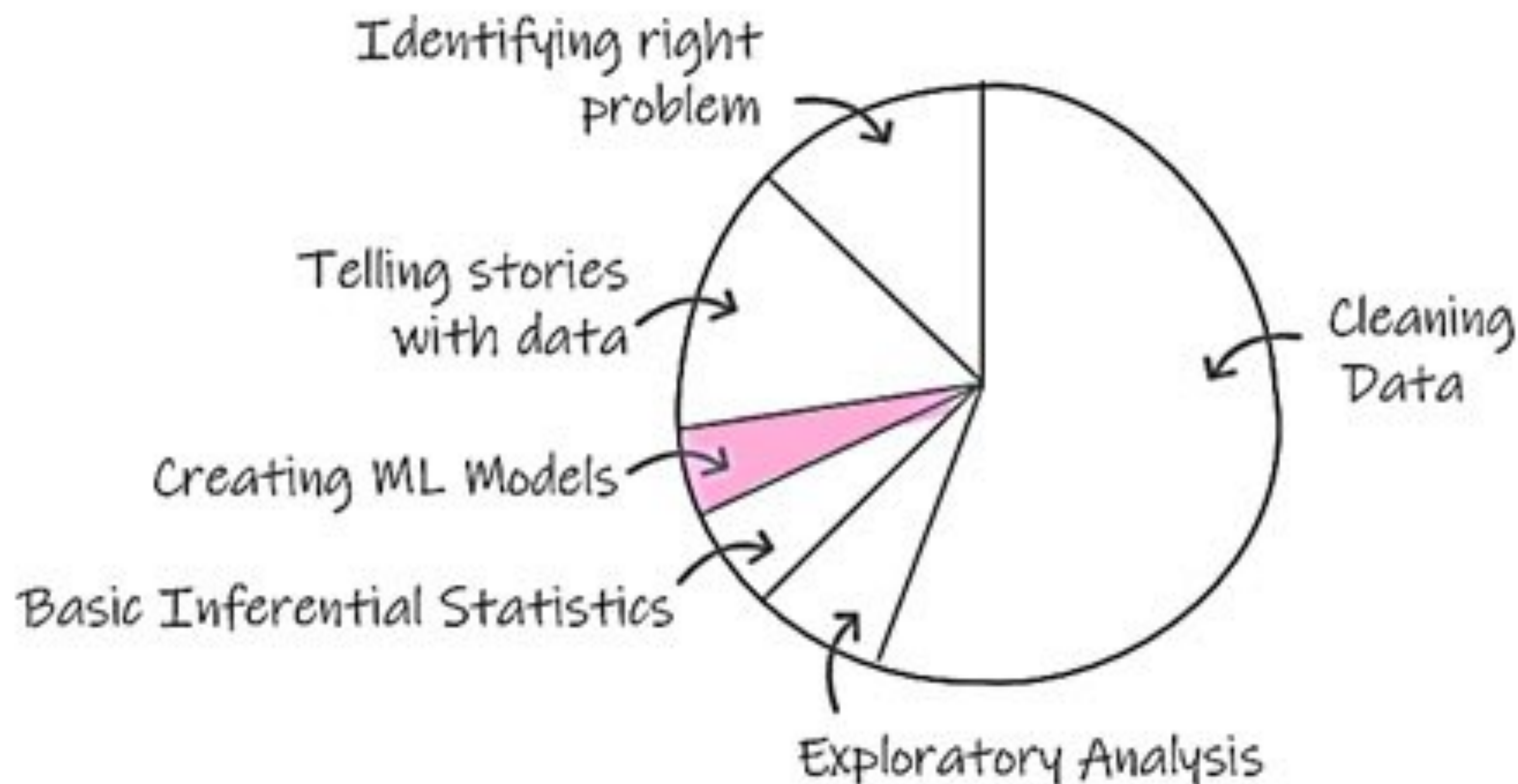
@datavizzdom

Gulrez

Perception



Reality



Time spent in a life of a Data Scientist

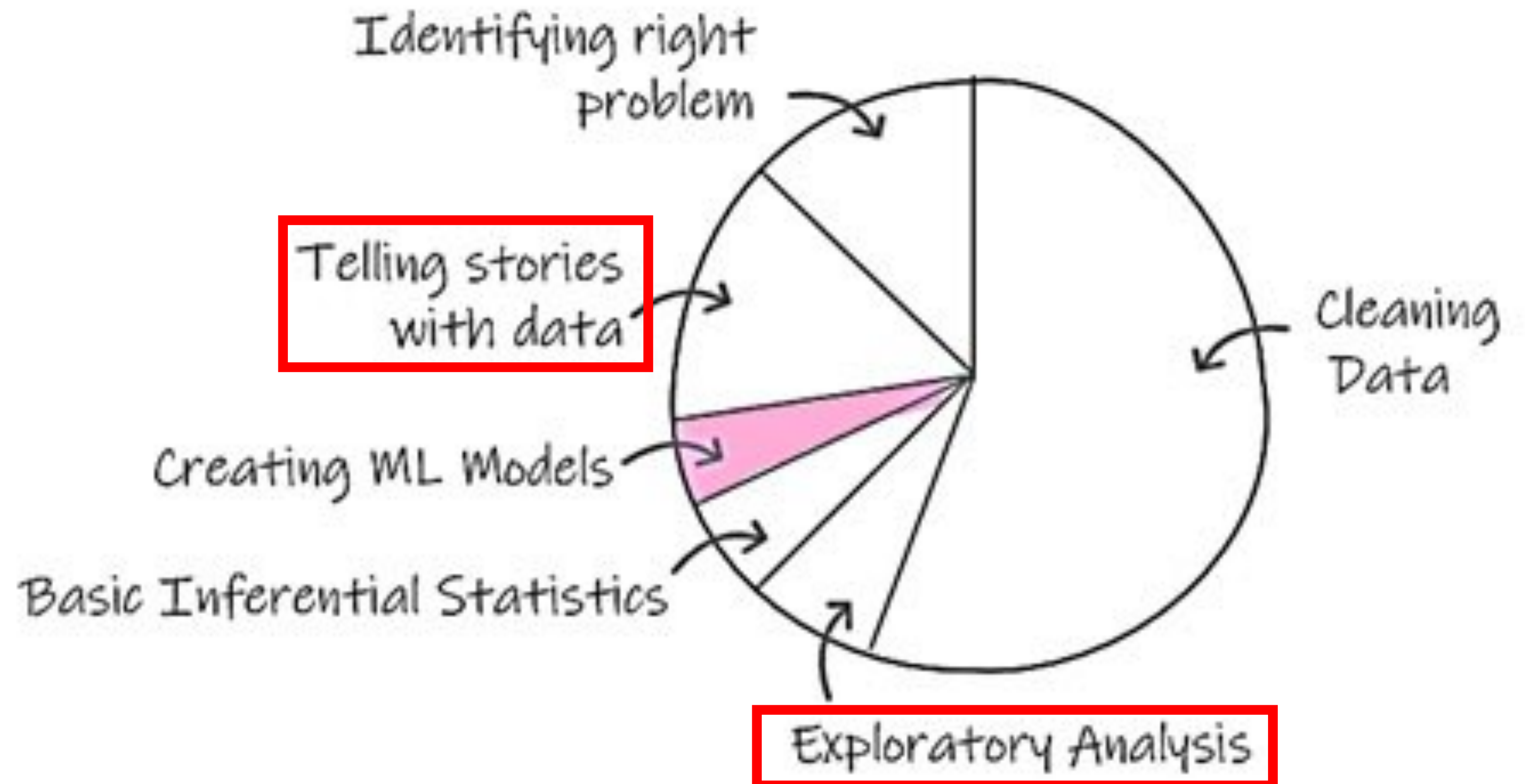
@datavizzdom

Gulrez

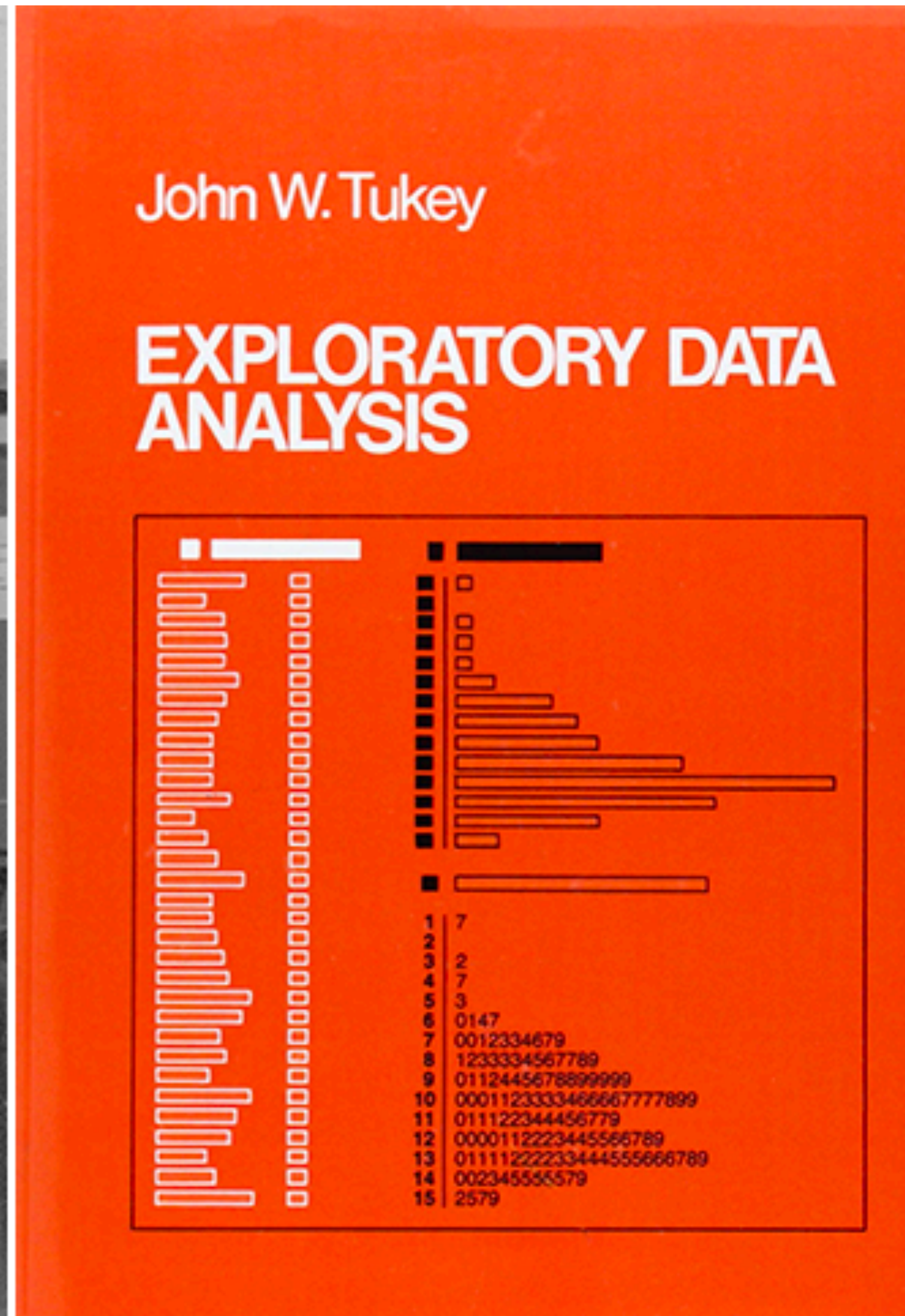
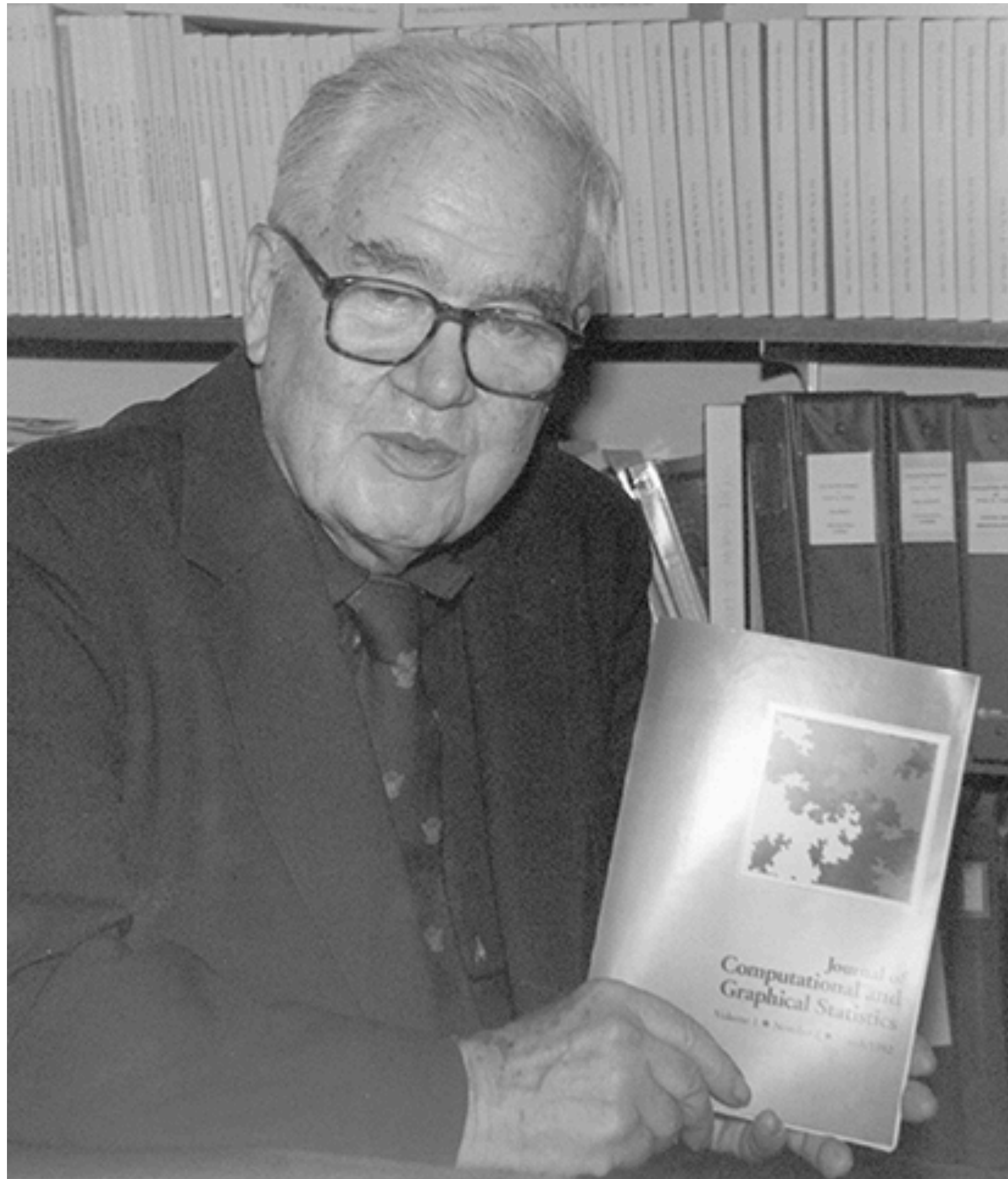
Perception



Reality



We are living in a Golden Age of visualization



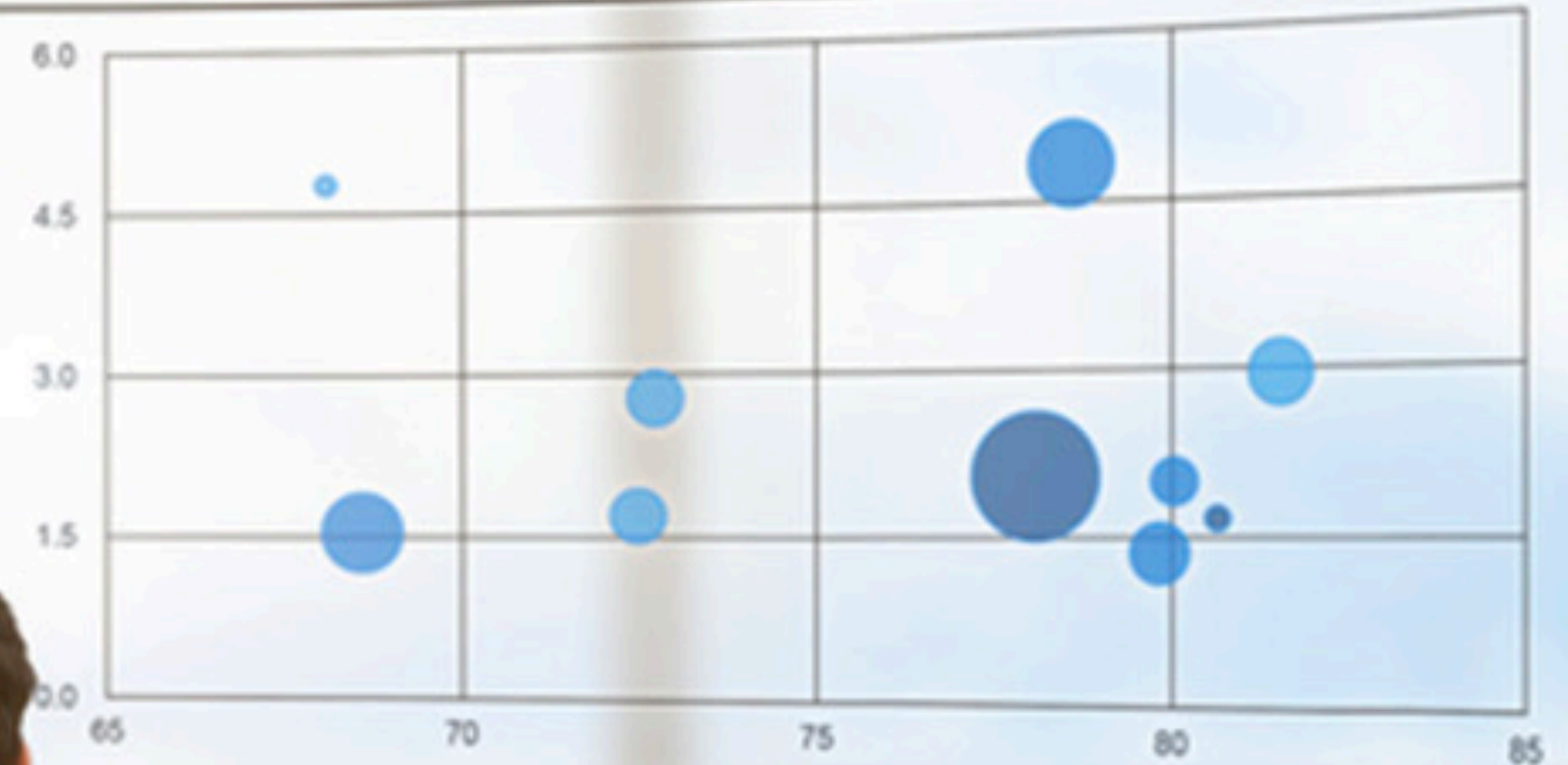
“The greatest value of a picture is when it forces us to notice what we never expected to see.”

John W. Tukey

We are living in a Golden Age of visualization



Products positioning



Sales per countries



Top 10 products



We are living in a Golden Age of visualization

Ed Hawkins's 'Warming stripes' (read more: <https://chezvoila.com/blog/warmingstripes/>)



We are living in a Golden Age of visualization

Q

Sections

The Washington Post

Democracy Dies in Darkness

Alberto Cairo To...

f

t

e

i

n

p

t

2.6k

Health

Why outbreaks like coronavirus spread exponentially, and how to “flatten the curve”

By Harry Stevens March 14, 2020

PLEASE NOTE

The Washington Post is providing this story for free so that all readers have access to this important information about the coronavirus. For more free stories, [sign up for our daily Coronavirus Updates newsletter](#).

<https://www.washingtonpost.com/graphics/2020/world/corona-simulator/>

We are living in a Golden Age of visualization

The Post's visual journalism, which involves staff throughout the newsroom, has attracted large audiences and contributed to record subscriber growth.

Six of the seven most visited stories in The Washington Post's history have been graphics, including the [coronavirus simulator](#) that became the most visited article in The Post's history, with more than three times as many visits as the second. It also includes this year's [Democratic candidate quiz](#), which set the record for converting readers to subscribers.

<https://www.washingtonpost.com/pr/2020/06/26/washington-post-expand-graphics-design-teams-with-14-new-positions/>

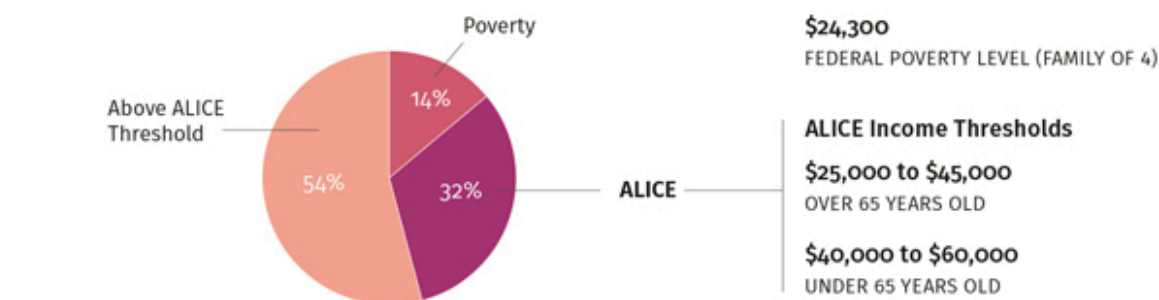
We are living in a Golden Age of visualization

FLORIDA: WHY ARE SO MANY SENIORS STRUGGLING?

Asset Limited, Income Constrained, Employed (ALICE) is a segment of the U.S. population who do not meet federal poverty levels but are struggling to make ends meet. In Florida, households 65 years and older saw the greatest increases below the ALICE threshold across all ethnic and racial groups; however, total households for those 25 years and under decreased. Florida also leads the nation with the greatest number of seniors

with many dependent on Social Security as their primary source of income. Housing is at an all-time high negatively affecting the overall health of many low-income Floridians. With formal care costs out of reach for most households, informal caregivers will continue to feel the financial burden of long term care. As more Floridians struggle to get by, will Florida have policies in place to address the monumental impacts of an aging population?

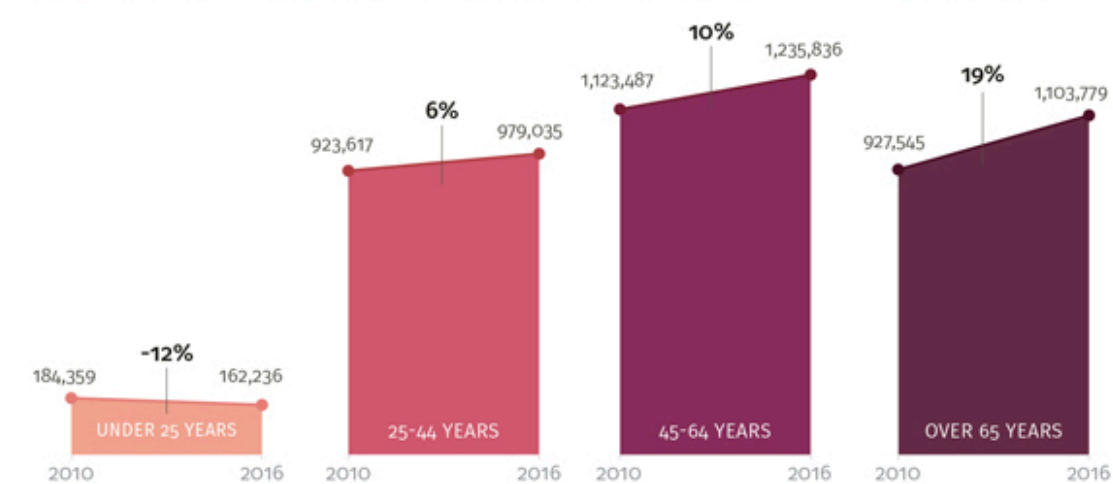
NEARLY 50% OF FLORIDIANS FIGHT TO SURVIVE



SOURCE: UNITED WAY 2016 ALICE REPORT

SENIORS, 45-64 YEAR OLDS INCREASINGLY STRUGGLE

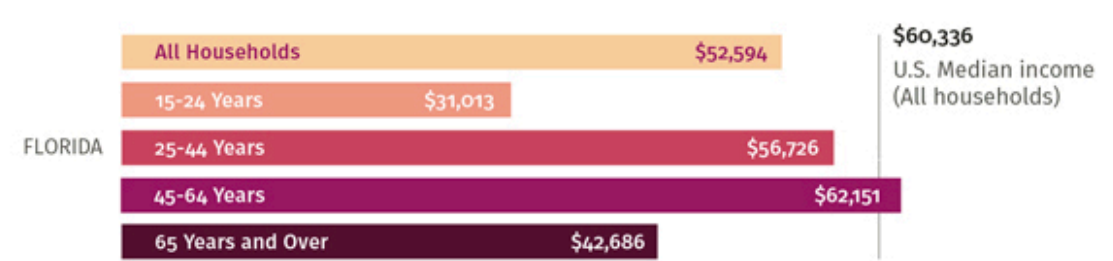
PERCENT CHANGE OF HOUSEHOLDS BY AGE BELOW THE ALICE THRESHOLD FROM 2010 TO 2016



SOURCE: UNITED WAY 2016 ALICE REPORT

MEDIAN INCOME BARELY COVERS THE BASICS

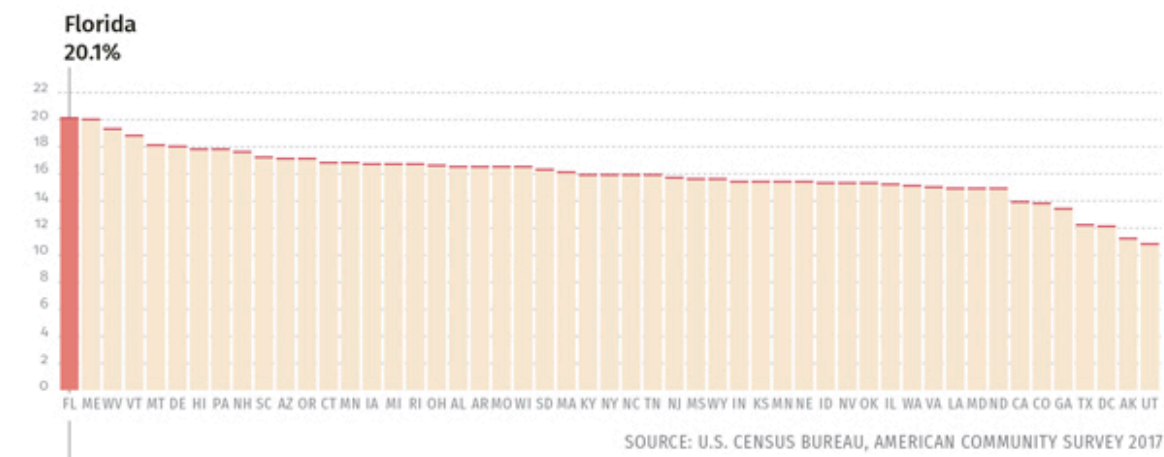
THE STRUGGLE TO AFFORD THE ESSENTIALS LEAVES LITTLE ROOM FOR UNEXPECTED COSTS



SOURCE: AMERICAN COMMUNITY SURVEY 2017

FLORIDA: THE NATION'S OLDEST STATE

MAINE, WEST VIRGINIA AND VERMONT FOLLOW WITH THE HIGHEST PERCENTAGE OF SENIORS (65+)



SOURCE: U.S. CENSUS BUREAU, AMERICAN COMMUNITY SURVEY 2017

PERCENT OF SENIORS BELOW THE ALICE THRESHOLD

HIGHEST PERCENTAGE OF ALICE HOUSEHOLDS 65+

Lafayette 49%
Glades 46%

LOWEST PERCENTAGE OF ALICE HOUSEHOLDS 65+

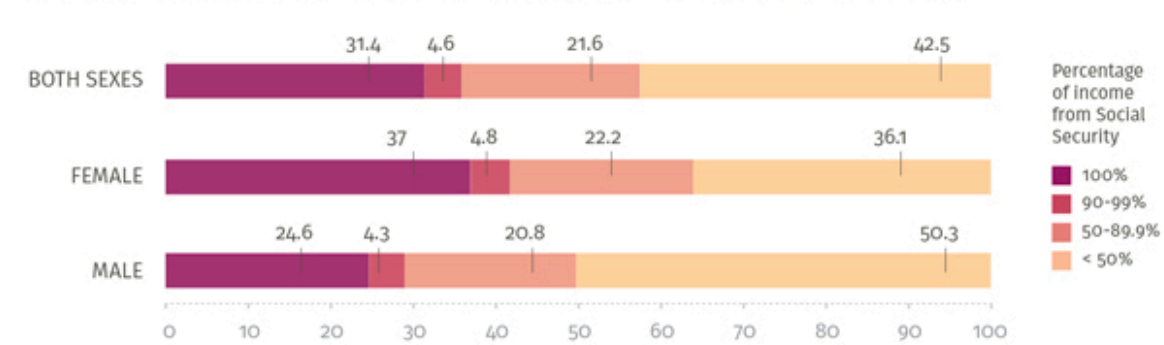
Wakulla 19%
Collier 20%
Leon 20%

20% OR LESS 50% OR MORE

SOURCE: UNITED WAY 2016 ALICE REPORT

SOCIAL SECURITY: A KEY INCOME SOURCE FOR SENIORS

RELiance ON SOCIAL SECURITY AS PERCENT OF TOTAL INCOME FOR PEOPLE 65+ YEARS



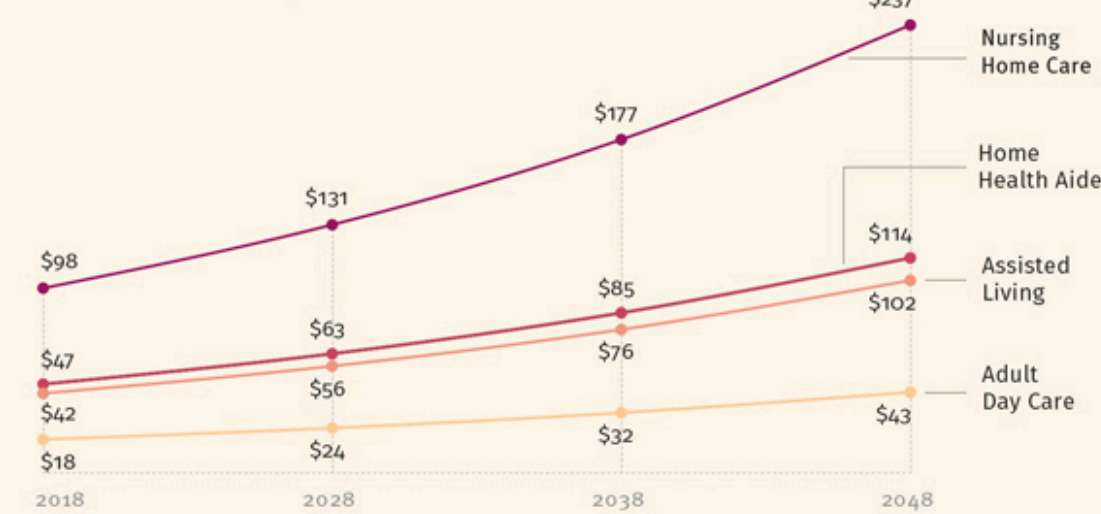
SOURCE: AARP PUBLIC POLICY INSTITUTE

LONG TERM CARE STRAIN ON THE HORIZON

Costs for longterm care ("custodial care") support in Florida will continue to increase at alarming rates forcing vulnerable populations to seek alternatives. Medicare can cover a fraction of costs and the bulk of the financial burden falls on individuals and families. Medicaid is only available for Americans with the lowest incomes with caveats. When it comes to long term care, ALICE households are forced further to the margins.

FORMAL CARE COSTS WILL CONTINUE TO OUTPACE U.S. INFLATION RATES

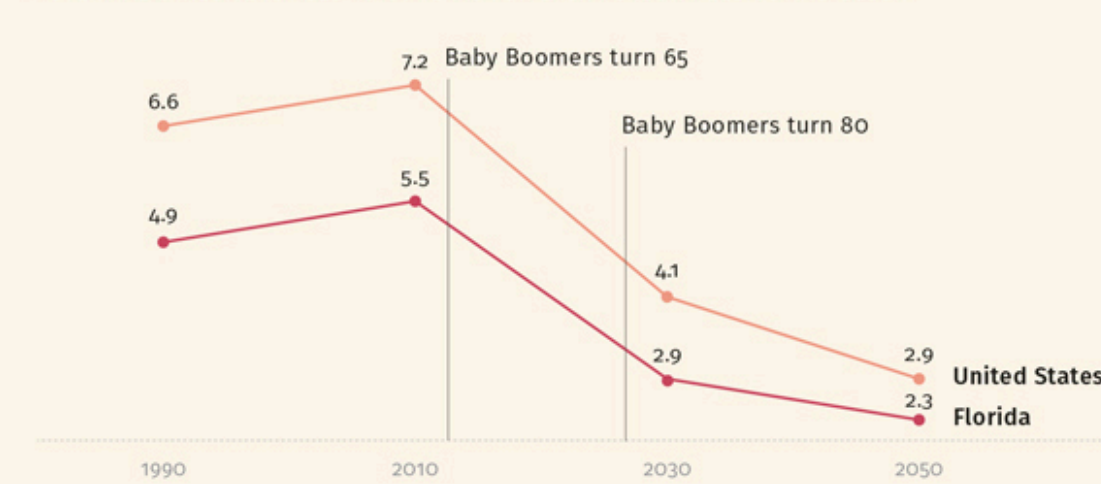
FLORIDA'S ANNUAL PROJECTED MEDIAN CARE COSTS IN THOUSANDS



NOTE: FUTURE YEARS CALCULATED ASSUMING AN ANNUAL 3% GROWTH RATE. HOME HEALTH CARE BASED ON 44 HOURS PER WEEK BY 52 WEEKS. ASSISTED LIVING BASED ON 12 MONTHS OF CARE, PRIVATE, ONE BEDROOM. ADULT DAY CARE BASED ON 5 DAYS PER WEEK FOR 52 WEEKS. NURSING HOME CARE BASED ON 365 DAYS OF CARE, SEMI-PRIVATE ROOM. SOURCE: GENWORTH COST OF CARE SURVEY

NUMBER OF FAMILY CAREGIVERS PROJECTED TO PLUNGE

POTENTIAL CAREGIVERS AGED 45-64 YEARS FOR EACH PERSON AGED 80 AND OLDER



SOURCE: AARP PUBLIC POLICY INSTITUTE

WHAT COULD FLORIDA DO TO HELP SENIORS?

ACCORDING TO AARP, COMPREHENSIVE PEOPLE-FOCUSED POLICIES COULD MAKE A BIG DIFFERENCE

- IMPROVE OVERALL LONG TERM SERVICES AND SUPPORT (LTSS):** Quality of life would increase, including the promise for seniors and adults with disabilities to afford housing.
- EXPAND HOME AND COMMUNITY-BASED SERVICES (HCBS):** More Floridians could avoid costly nursing homes and family caregivers would be able to receive assistance.
- EXPAND MEDICAID:** Non-elderly adults without dependents could be covered and the health coverage gap for nearly 400,000 Floridians would be reduced.

GRAPHIC: DEB PANG DAVIS / JMM622 INTRO TO DATA VISUALIZATION

ANYONE can learn to make data visualizations.

Example of my students' work:
<https://www.deb.is/>

How to teach or learn data visualization?

Visualization is a bit like writing: beyond some conventions and constraints regarding symbols, visual grammar, perception, and cognition, visualization **can't be based on “rules” that are set in stone.**

Instead, when designing visualizations, we need to be guided by **reasoned, justifiable choices.**

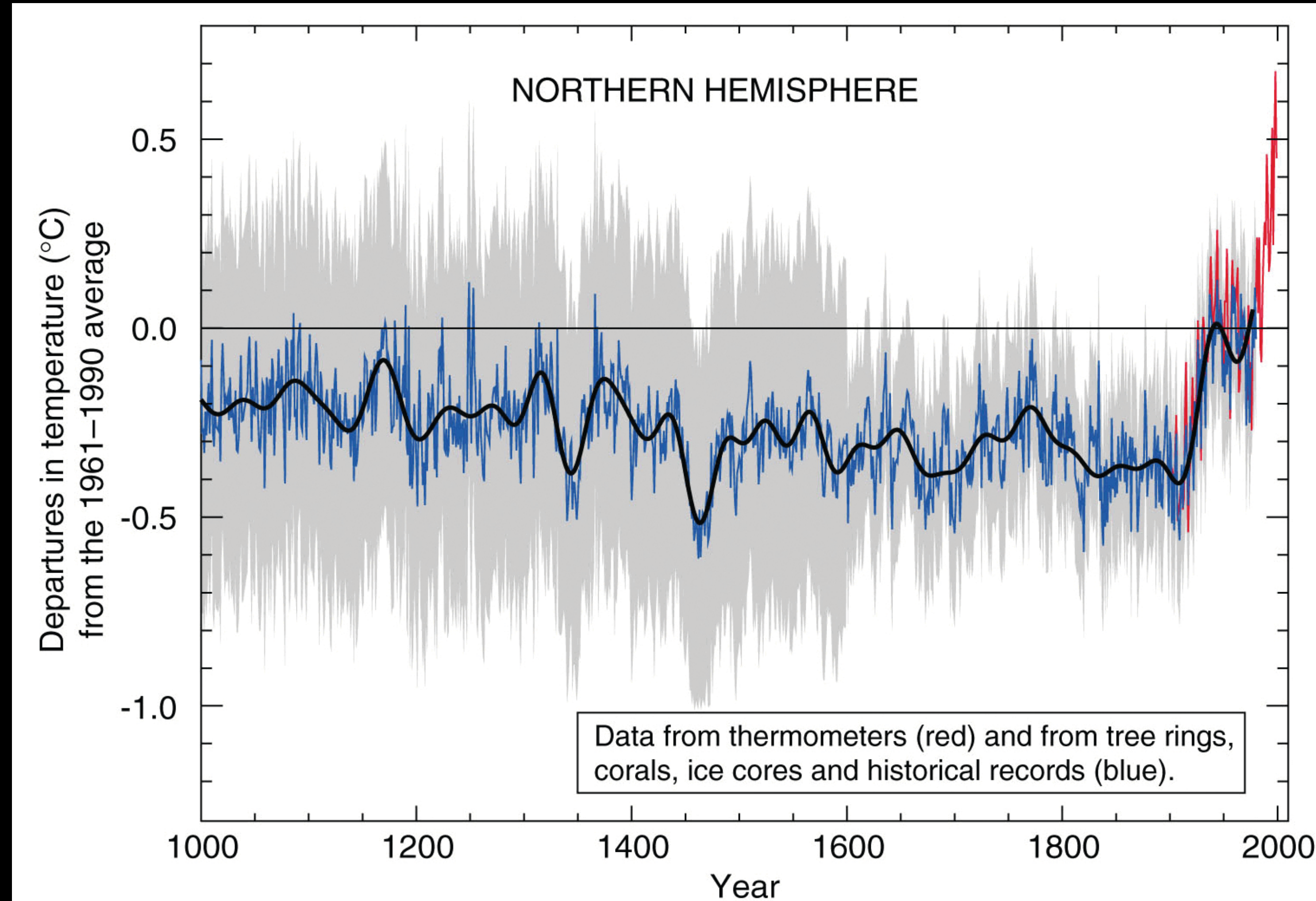


I. Why should my visualization exist?

What are tables useful for?

	A	B	C	D	E	F	G		A	B	C	D	E	F	G	H
1	YEAR	TEMP	YEAR	1 SIGMA	2 SIGMA			878	1876	-0.1891	1876	0.113228	0.226456	8.25297E-02	7.75207E-02	
2	1000	0.0659	1000	0.240346	0.480693	0.206137	0.123588	879	1877	-0.0140	1877	0.113228	0.226457	8.25299E-02	7.75209E-02	
3	1001	-0.1241	1001	0.240347	0.480694	0.206137	0.123589	880	1878	-0.0873	1878	0.113228	0.226457	8.25298E-02	7.75209E-02	
4	1002	-0.1208	1002	0.240346	0.480692	0.206136	0.123588	881	1879	-0.2959	1879	0.113229	0.226458	8.25302E-02	7.75212E-02	
5	1003	-0.1801	1003	0.240347	0.480694	0.206137	0.123589	882	1880	-0.2368	1880	0.113229	0.226457	8.25300E-02	7.75210E-02	
6	1004	-0.0711	1004	0.240347	0.480693	0.206137	0.123588	883	1881	-0.1977	1881	0.113229	0.226458	8.25302E-02	7.75212E-02	
7	1005	-0.1334	1005	0.240346	0.480692	0.206136	0.123588	884	1882	-0.2036	1882	0.113229	0.226457	8.25300E-02	7.75210E-02	
8	1006	-0.0644	1006	0.240346	0.480693	0.206137	0.123588	885	1883	-0.2489	1883	0.113228	0.226455	8.25293E-02	7.75204E-02	
9	1007	0.0042	1007	0.240347	0.480693	0.206137	0.123588	886	1884	-0.2125	1884	0.113229	0.226457	8.25301E-02	7.75211E-02	
10	1008	-0.1288	1008	0.240347	0.480693	0.206137	0.123588	887	1885	-0.1896	1885	0.113228	0.226457	8.25299E-02	7.75210E-02	
11	1009	-0.0296	1009	0.240347	0.480693	0.206137	0.123588	888	1886	-0.1084	1886	0.113228	0.226456	8.25298E-02	7.75208E-02	
12	1010	0.1187	1010	0.240347	0.480694	0.206137	0.123589	889	1887	-0.3265	1887	0.113228	0.226456	8.25296E-02	7.75206E-02	
13	1011	-0.1252	1011	0.240346	0.480692	0.206136	0.123588	890	1888	-0.1694	1888	0.113228	0.226457	8.25298E-02	7.75209E-02	
14	1012	-0.1634	1012	0.240347	0.480694	0.206137	0.123588	891	1889	-0.1339	1889	0.113228	0.226456	8.25298E-02	7.75208E-02	
15	1013	-0.0791	1013	0.240347	0.480693	0.206137	0.123588	892	1890	-0.3107	1890	0.113229	0.226457	8.25301E-02	7.75211E-02	
16	1014	-0.1120	1014	0.240347	0.480693	0.206137	0.123588	893	1891	-0.1754	1891	0.113229	0.226457	8.25300E-02	7.75210E-02	
17	1015	-0.1146	1015	0.240346	0.480692	0.206136	0.123588	894	1892	-0.3186	1892	0.113228	0.226456	8.25295E-02	7.75205E-02	
18	1016	-0.1206	1016	0.240346	0.480692	0.206136	0.123588	895	1893	-0.3236	1893	0.113228	0.226456	8.25297E-02	7.75207E-02	
19	1017	-0.0815	1017	0.240347	0.480693	0.206137	0.123588	896	1894	-0.1970	1894	0.113228	0.226456	8.25295E-02	7.75205E-02	
20	1018	-0.2031	1018	0.240346	0.480693	0.206137	0.123588	897	1895	-0.1578	1895	0.113228	0.226456	8.25297E-02	7.75207E-02	
21	1019	0.0305	1019	0.240347	0.480693	0.206137	0.123588	898	1896	-0.0804	1896	0.113228	0.226456	8.25298E-02	7.75208E-02	
22	1020	0.1098	1020	0.240347	0.480694	0.206137	0.123589	899	1897	-0.0537	1897	0.113228	0.226456	8.25298E-02	7.75208E-02	
23	1021	0.0244	1021	0.240347	0.480693	0.206137	0.123588	900	1898	-0.2195	1898	0.113229	0.226457	8.25301E-02	7.75211E-02	
24	1022	-0.0743	1022	0.240347	0.480693	0.206137	0.123588	901	1899	-0.3486	1899	0.113228	0.226456	8.25297E-02	7.75207E-02	
25	1023	-0.0323	1023	0.240347	0.480693	0.206137	0.123588	902	1900	-0.1253	1900	0.113229	0.226457	8.25300E-02	7.75210E-02	
26	1024	-0.0434	1024	0.240346	0.480693	0.206137	0.123588	903	1901	-0.1575	1901	0.113228	0.226456	8.25296E-02	7.75206E-02	

Visualization is about patterns, trends, the big picture



Michael E. Mann, Raymond S. Bradley, and Malcolm K. Hughes

Intergovernmental Panel on Climate Change (IPCC), Third Report, 2001



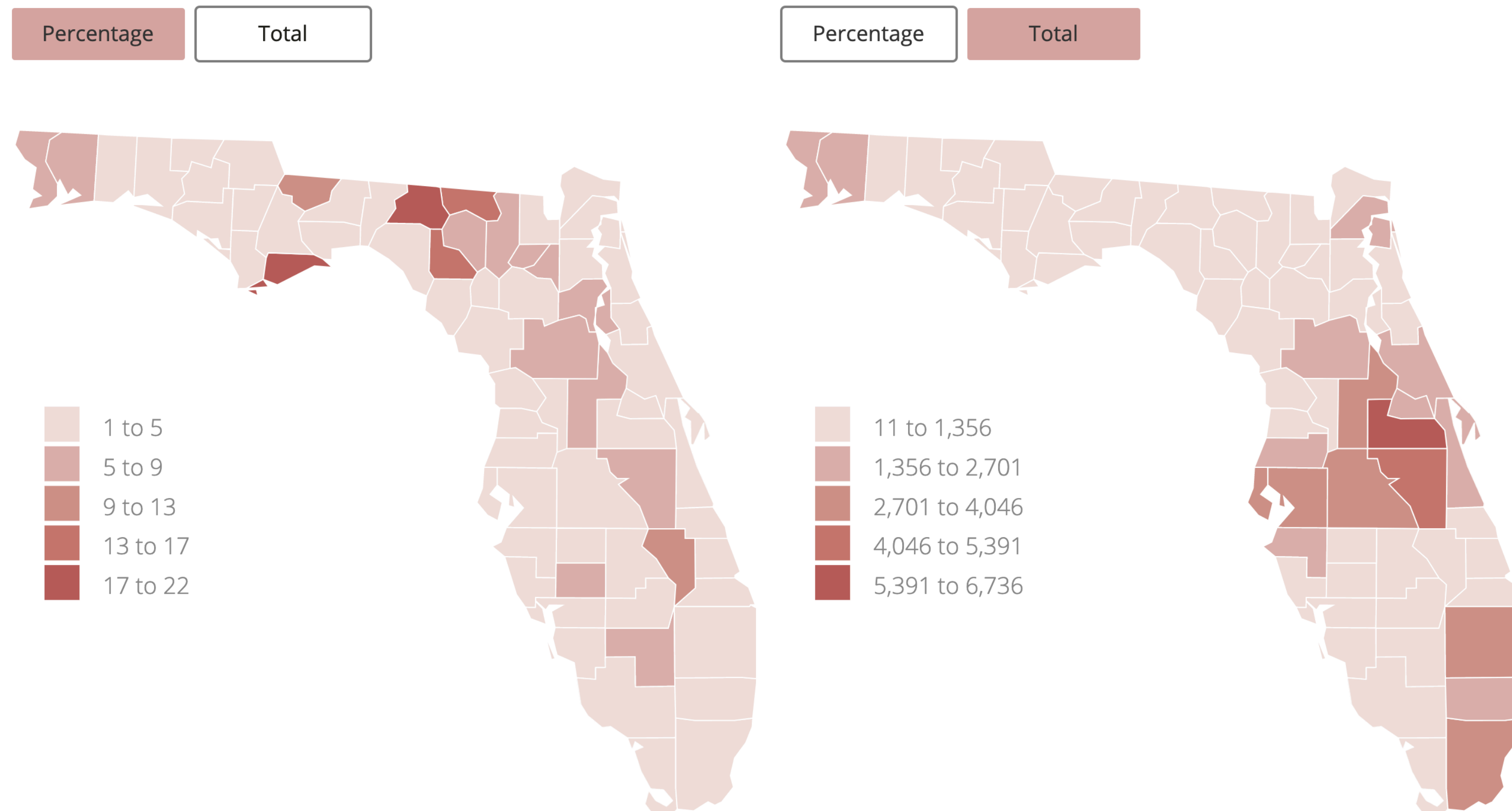
2. What to visualize?

AT SCHOOL

WITHOUT A ROOF

[http://
www.lmelgar.me/
without-a-roof/](http://www.lmelgar.me/without-a-roof/)

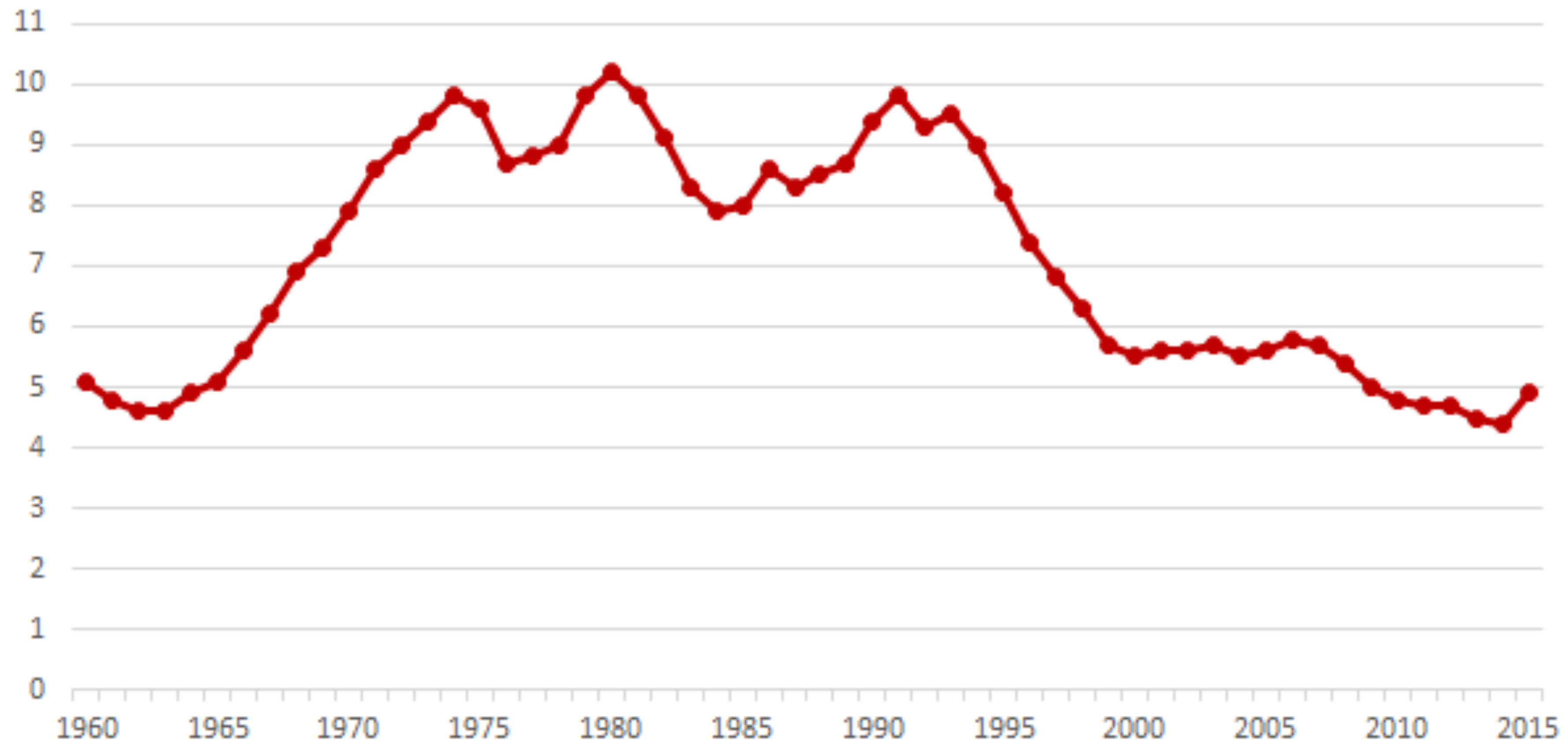
In Florida more than 71,000 students are homeless. During the last decade, this population rocketed as a result of the recession and how hard it has become for the poorest families to find affordable housing.





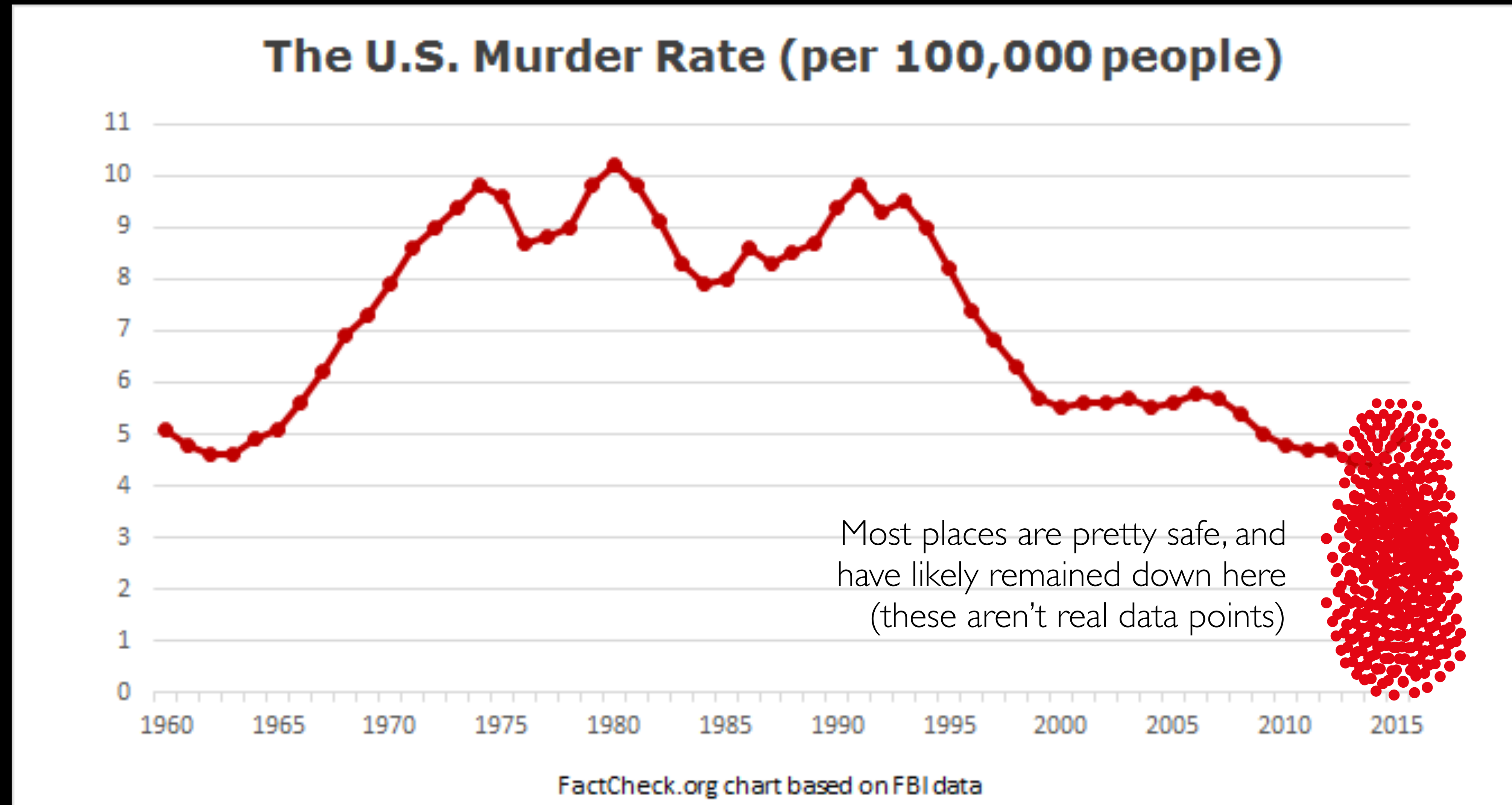
3. How much to visualize?

The U.S. Murder Rate (per 100,000 people)

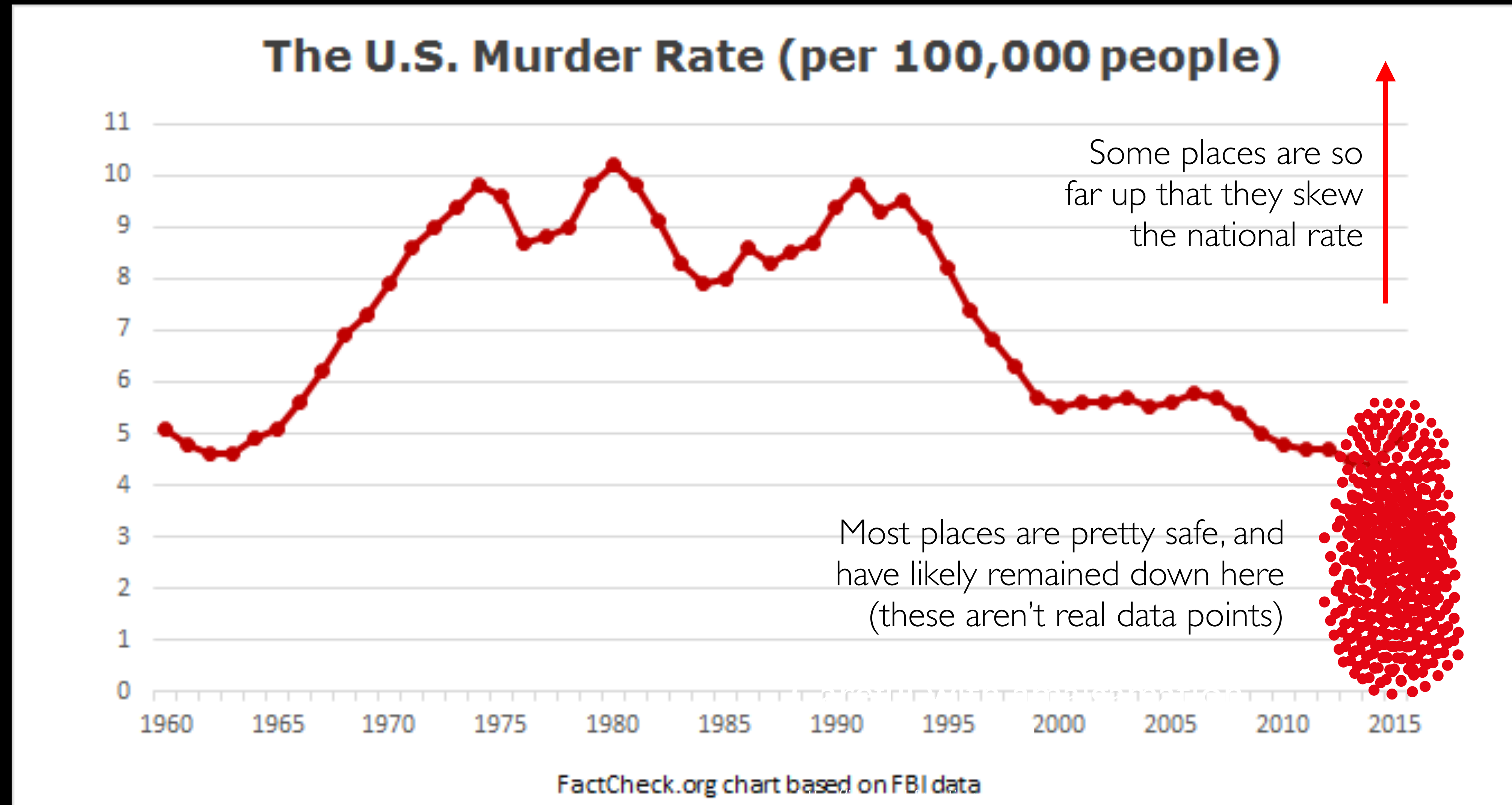


FactCheck.org chart based on FBI data

The danger of aggregating data too much,
and presenting just averages and other statistical summaries



The danger of aggregating data too much,
and presenting just averages and other statistical summaries





4. How to visualize it?

Figure 2 - Main nationalities of arriving migrants – 2016

Greece

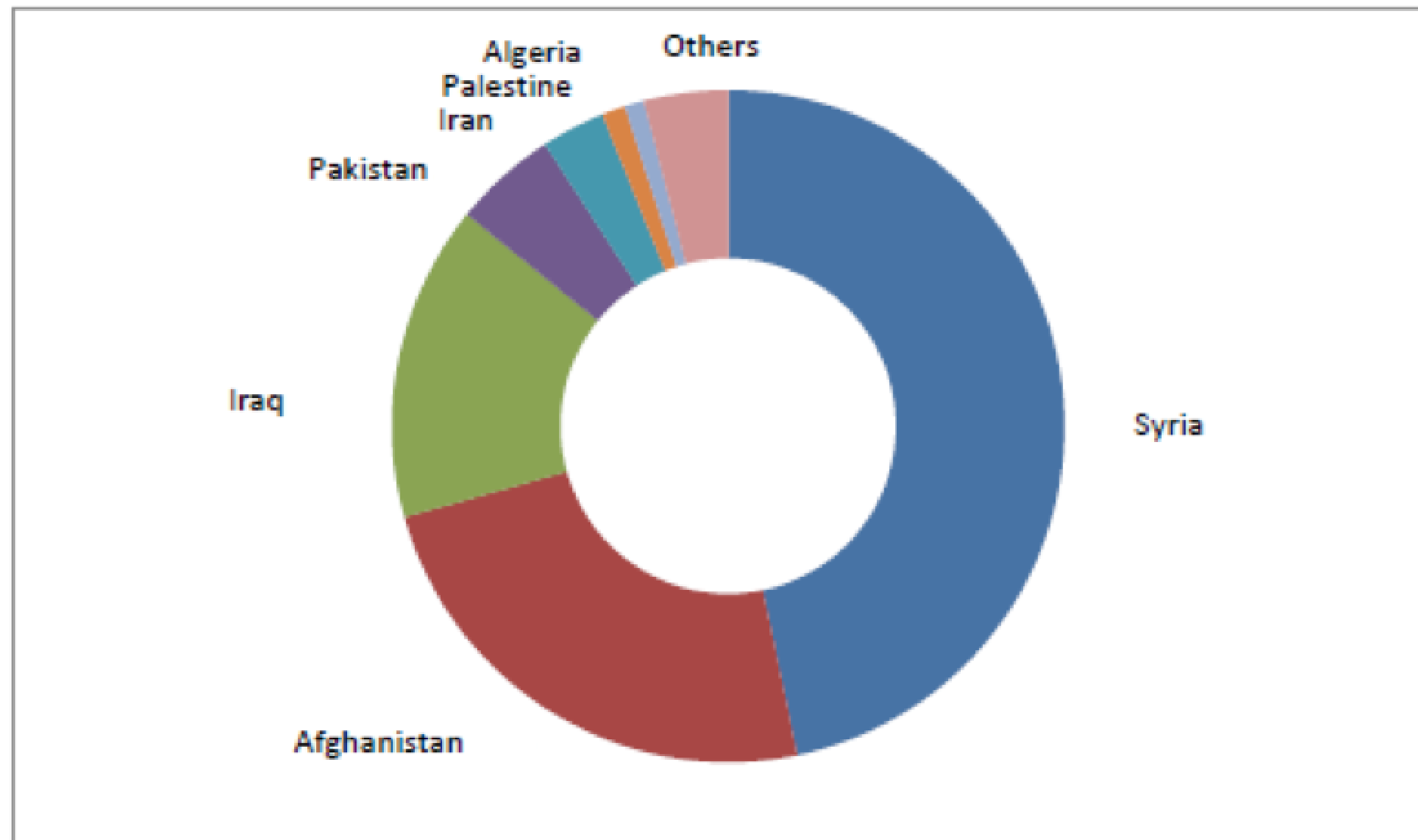
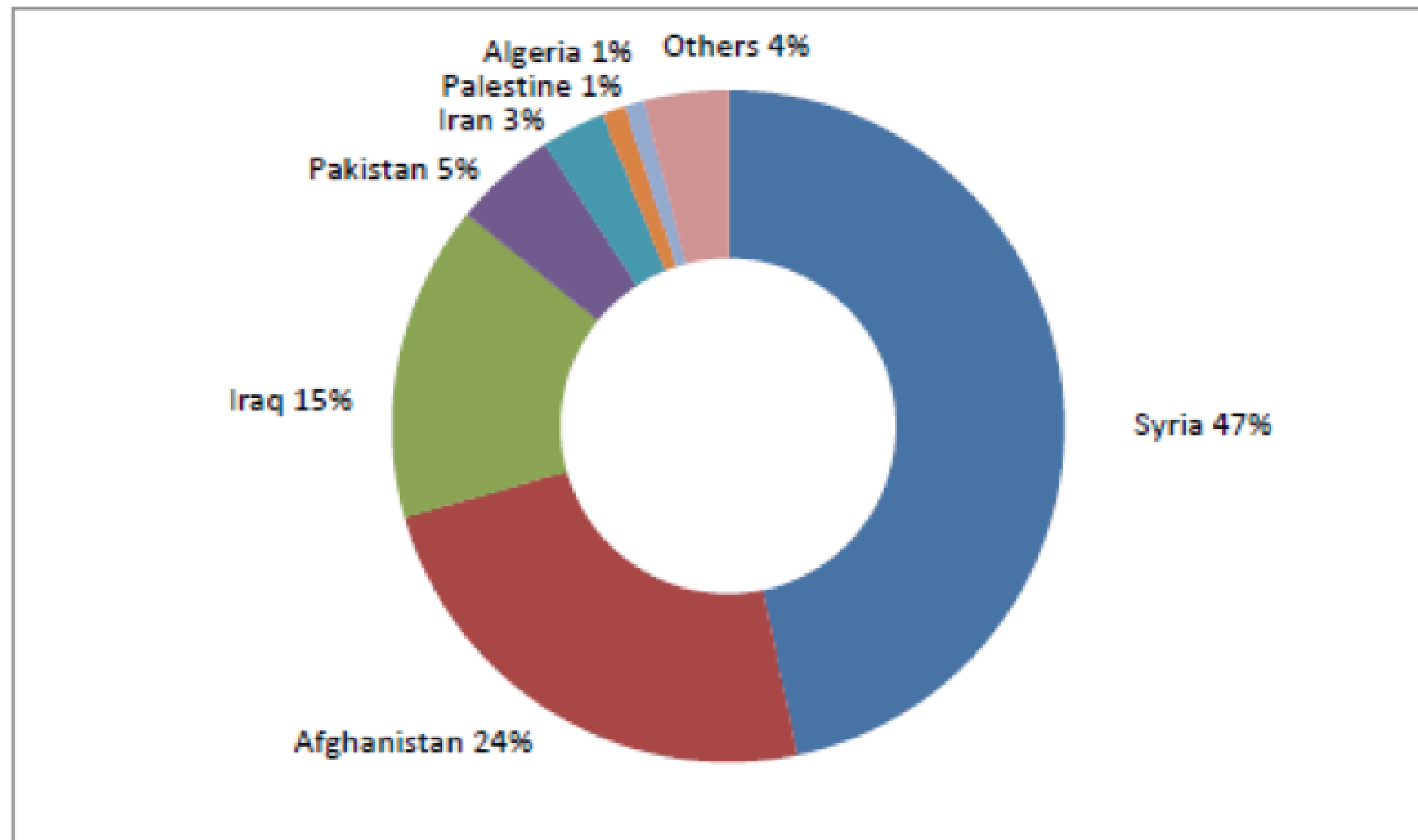
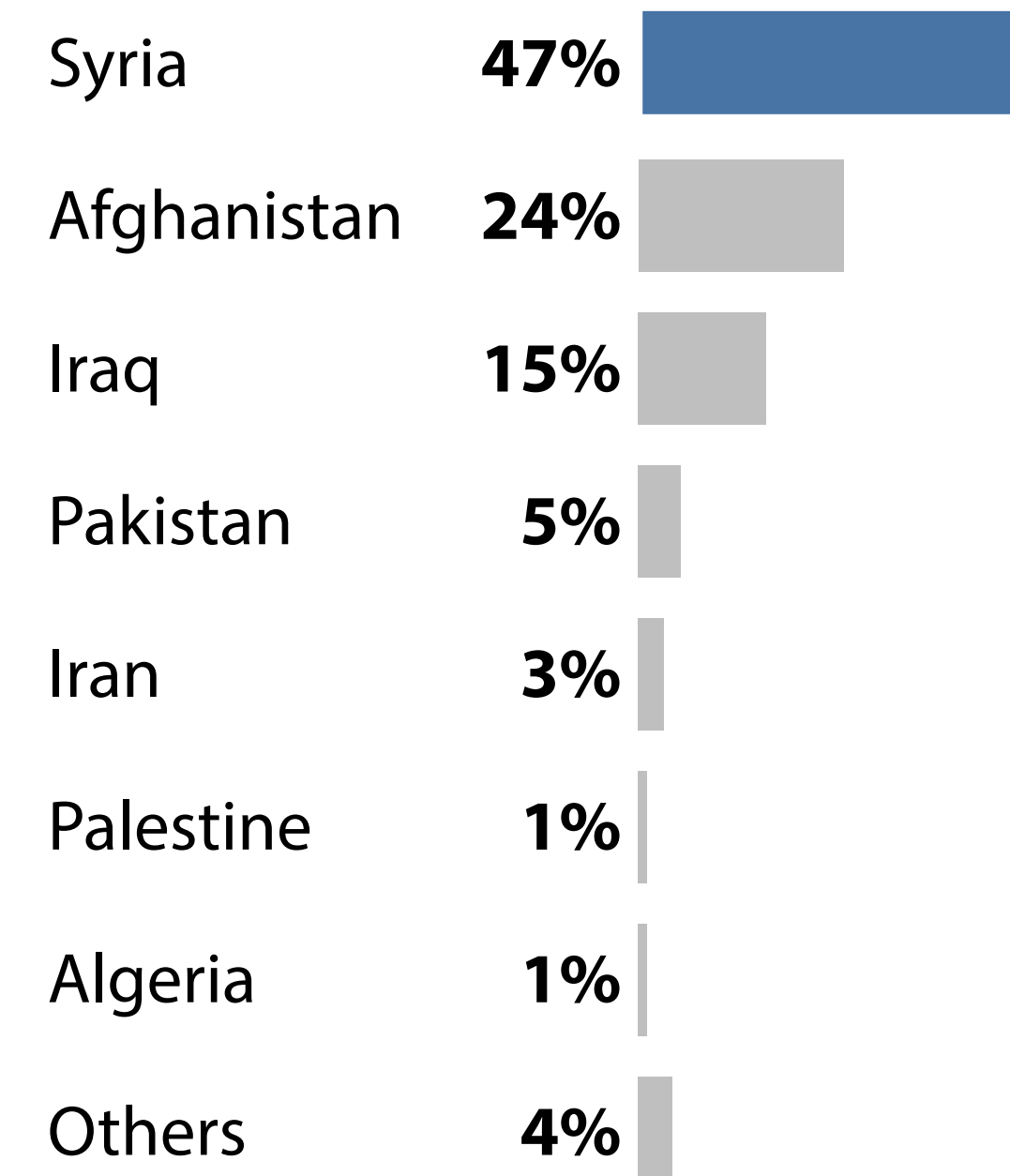
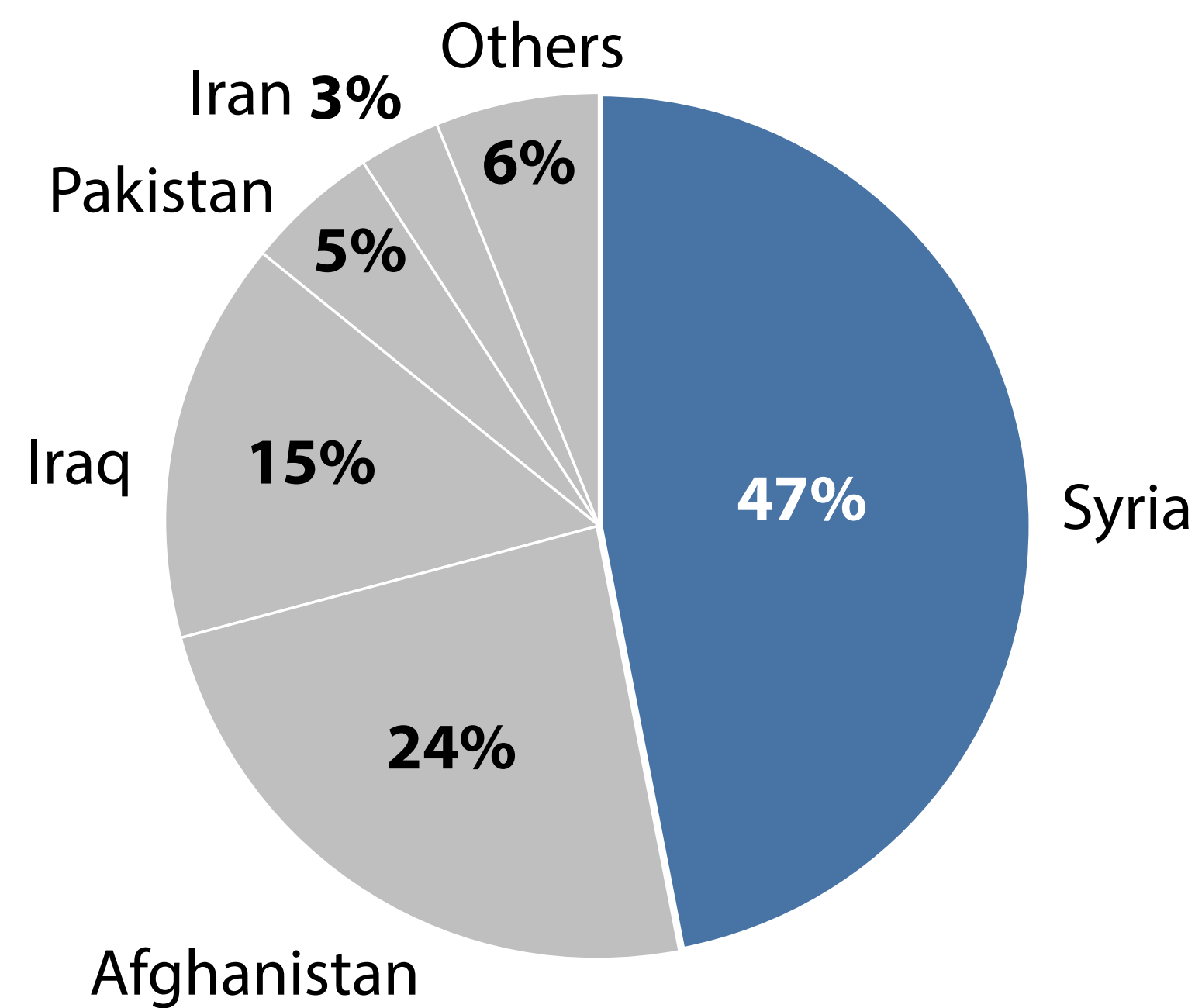
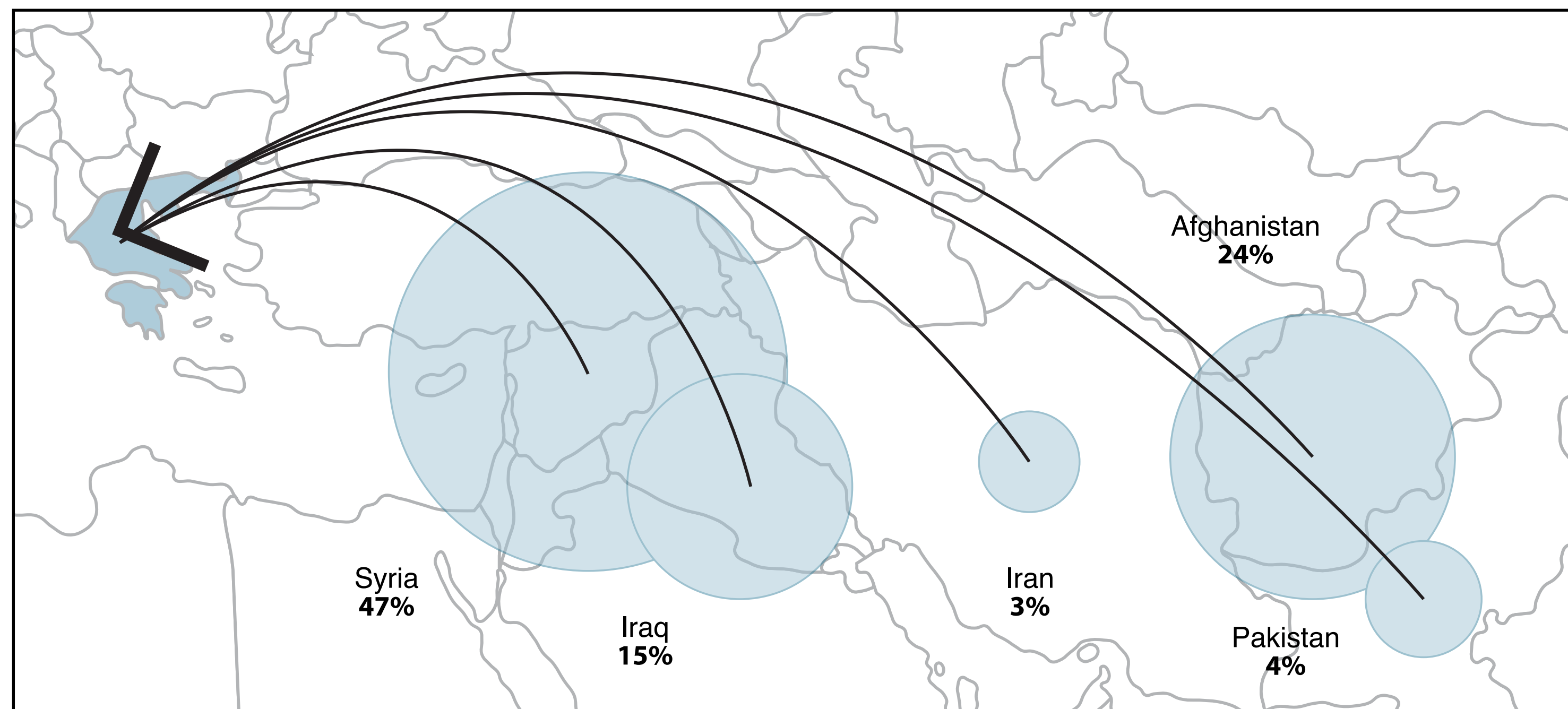


Figure 2 - Main nationalities of arriving migrants – 2016

Greece





The Data Visualisation Catalogue

About • Suggest • Shop • Resources

Search by Function

View by List



Deviation

Emphasize variations (+/-) from a fixed reference point. Typically the reference point is zero but it can also be a target or a long-term average. Can also be used to show sentiment (positive/negative/neutral).

Example FT uses
Trade surplus/deficit, climate change

Diverging bar
A simple standard bar chart that can handle both negative and positive magnitude values.

Diverging stacked bar
Perfect for presenting survey results which involve sentiment (eg. disagree/neutral/agree).

Spine chart
Splits a single value into 2 contrasting components (eg. Male/Female).

Surplus/deficit filled line
The shaded area of these charts allows a balance to be shown – either against a baseline or between two series.

Correlation

Show the relationship between two or more variables. Be mindful that unless you tell them otherwise, many readers will assume the relationship you show them to be causal (ie. one causes the other).

Example FT uses
Inflation & unemployment, income & life expectancy

Scatterplot
The standard way to show the relationship between two continuous variables, each of which has its own axis.

Line + Column
A good way of showing the relationship between an amount (columns) and a rate (line).

Connected scatterplot
Usually used to show how the relationship between 2 variables has changed over time.

Bubble
Like a scatterplot, but adds additional detail by storing the circles according to a third variable.

XY heatmap
A good way of showing the patterns between 2 categories of data, less good at showing fine differences in amounts.

Ranking

Use where an item's position in an ordered list is more important than its absolute or relative value. Don't be afraid to highlight the points of interest.

Example FT uses
Wealth, deprivation, league tables, constituency election results

Ordered bar
Standard bar charts display the ranks of values much more easily when sorted into order.

Ordered column
See above.

Ordered proportional symbol
Use when there are big variations between values and/or seeing fine differences between data is not so important.

Dot strip plot
Dots placed in order on a strip are a space-efficient method of laying out ranks across multiple categories.

Slope
Perfect for showing how ranks have changed over time or vary between categories.

Lollipop chart
Lollipops draw more attention to the data value than standard bar/columns and can also show rank and value effectively.

Distribution

Show values in a dataset and how often they occur. The shape (or 'skew') of a distribution can be a memorable way of highlighting the lack of uniformity or equity in the data.

Example FT uses
Income distribution, population, (age)sex distribution

Histogram
The standard way to show a statistical distribution – keep the gaps between columns small to highlight the 'shape' of the data.

Boxplot
Summarise multiple distributions by showing the median, quartiles and range of the data.

Violin plot
Similar to a box plot but more effective with complex distributions (data that cannot be summarised with simple averages).

Population pyramid
A standard way for showing the age and sex breakdown of a population distribution; effectively, back to back histograms.

Dot strip plot
Good for showing individual values in a distribution, can be a problem when there are many dots have the same value.

Dot plot
A simple way of showing the change or range (min/max) of data across multiple categories.

Barcode plot
Like dot strip plots, good for displaying all the data in a table; they work best when highlighting individual values.

Cumulative curve
A good way of showing how unequal a distribution is; y axis is always cumulative frequency, x axis is always a measure.

Change over Time

Give emphasis to changing trends. These can be short (or 'spike') movements or extended series (trending upwards or downwards). Choosing the correct time period is important to provide suitable context for the reader.

Example FT uses
Share price movements, economic time series

Line
The standard way to show a changing time series. If data are irregular, consider markers to represent data points.

Column
Columns work well for showing change over time – but usually best with only one series of data at a time.

Line + column
A good way of showing the relationship over time between an amount (columns) and a rate (line).

Stock price
Usually focused on day-to-day activity, these charts show opening/closing and high/low points of each day.

Slope
Good for showing changing data as long as the data can be simplified into 2 or 3 points without missing a key part of story.

Area chart
Use with care – these are good at showing changes to total, but seeing change in components can be very difficult.

Fan chart (projections)
Use to show the uncertainty in future projections – usually this grows the further forward to projection.

Connected scatterplot
A good way of showing changing data for two variables whenever there is a relatively clear pattern of progression.

Calendar heatmap
A great way of showing temporal patterns (daily, weekly, monthly) – at the expense of showing precision in quantity.

Priestley timeline
Great when date and duration are key elements of the story in the data.

Circle timeline
Good for showing discrete values of varying size across multiple categories (eg. earthquakes by continent).

Seismogram
Another alternative to the circle timeline for showing series where there are big variations in the data.

Part-to-whole

Show how a single entity can be broken down into its constituent elements, if the reader's interest is solely in the size of the components, consider a magnitude-type chart instead.

Example FT uses
Fiscal budgets, company structures, national election results

Stacked column
A simple way of showing the part-to-whole relationships but can be difficult to read with more than a few components.

Proportional stacked bar
A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Pie
A common way of showing part-to-whole data – but be aware that it's difficult to accurately compare the size of the segments.

Donut
Similar to a pie chart – but the centre can be a good way of making space to include more information about the data (eg. total).

Treemap
Use for hierarchical data where the size and proportion of data at the same time – as long as the data are not too complicated.

Voronoi
A way of turning points into areas – any point within each area is closer to the central point than any other centroid.

Sunburst
Another way of visualising hierarchical part-to-whole relationships – usually only with whole numbers (do not slice off an arm to represent a decimal).

Arc
A hemicircle, often used for visualising political results in parliaments.

Gridplot
Good for showing % information, they work best when used on whole numbers and work well in multiple layout form.

Venn
Generally only used for schematic representation.

Waterfall
Can be useful for showing part-to-whole relationships where some of the components are negative.

Magnitude

Show size comparisons. These can be relative (size being able to see larger/smaller) or absolute (need to see fine differences). Usually these show a 'counted' number (for example, barrels, dollars or people) rather than a calculated rate or per cent.

Example FT uses
Commodity production, market capitalisation

Column
The standard way to compare the size of things. Must always start at 0 on the axis.

Bar
See above. Good when the data are not time series and labels have long category names.

Paired column
As per standard column but allows for multiple series. Can become tricky to read with more than 2 series.

Paired bar
See above.

Proportional stacked bar
A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Proportional symbol
Use when there are big variations between values and/or seeing fine differences between data is not so important.

Isotype (pictogram)
Excellent solution in some instances – use only with whole numbers (do not slice off an arm to represent a decimal).

Lollipop chart
Lollipop charts draw more attention to the data value than standard bar/column – does not HAVE to start at zero (but preferable).

Radar chart
A space-efficient way of showing value of multiple variables – but make sure they are organised in a way that makes sense to reader.

Parallel coordinates
An alternative to radar charts – again, the arrangement of the variables is important. Usually benefits from highlighting values.

Spatial

Used only when precise locations or geographical patterns in data are more important to the reader than anything else.

Example FT uses
Locator maps, population density, natural resource locations, natural disaster risk/impact, catchment areas, variation in election results

Basic choropleth (categorical)
The standard approach for putting data on a map – should always be rates rather than values and use a sensible base geography.

Proportional symbol (count/magnitude)
Use for totals rather than rates – be wary that small differences in data will be hard to see.

Flow map
For showing unambiguous movement across a map.

Contour map
For showing areas of equal value on a map. Can use deviation colour schemes for showing +/- values.

Equalized cartogram
Converting each unit on a map to a regular and equally-sized shape – good for representing voting regions with equal value.

Scaled cartogram (value)
Stretching and shrinking a map so that each area is sized according to a particular value.

Dot density
Used to show the location of individual events/locations – make sure to annotate any patterns the reader should see.

Heat map
Grid-based data values mapped with an intensity colour scale. As choropleth map – but not trapped to an administrative unit.

Flow

Show the reader volumes or intensity of movement between two or more states or conditions. These might be logical sequences or geographical locations.

Example FT uses
Movement of funds, trade, migrants, lawsuits, information relationship graphs.

Sankey
Shows changes in flows from one condition to at least one other; good for tracing the eventual outcome of a complex process.

Waterfall
Designed to show the sequencing of data through a flow process, typically budgets. Can include +/- components.

Chord
A complex but powerful diagram which can illustrate 2-way flows (and net flows) in a matrix.

Network
Used for showing the strength and inter-connectedness of relationships of varying types.

Visual vocabulary

Designing with data

There are so many ways to visualise data - how do we know which one to pick? Use the categories across the top to decide which data relationship is most important in your story, then look at the different types of chart within the category to form some initial ideas about what might work best. This list is not meant to be exhaustive, nor a wizard, but is a useful starting point for making informative and meaningful data visualisations.

FT graphic: Alan Smith; Chris Campbell; Jan Both; Li Feunice; Graham Parish; Billy Ehrenberg; Paul McCallum; Martin Stabe
Inspired by the Graphic Continuum by Jon Schwabish and Severino Ribeca

ft.com/vocabulary



<http://www.datavizcatalogue.com/>

<https://github.com/ft-interactive/chart-doctor/blob/master/visual-vocabulary/Visual-vocabulary.pdf>

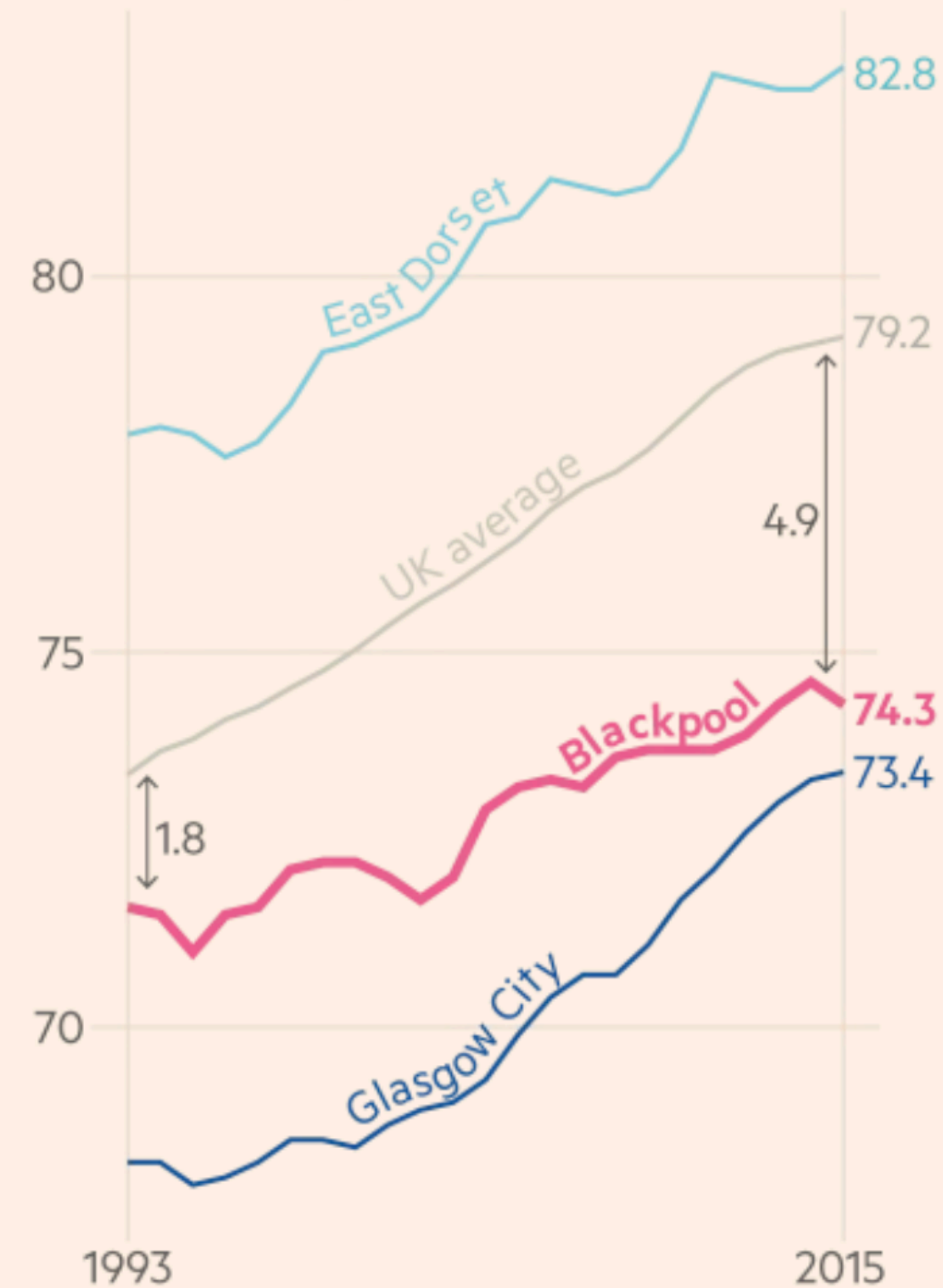


5. What words to add?

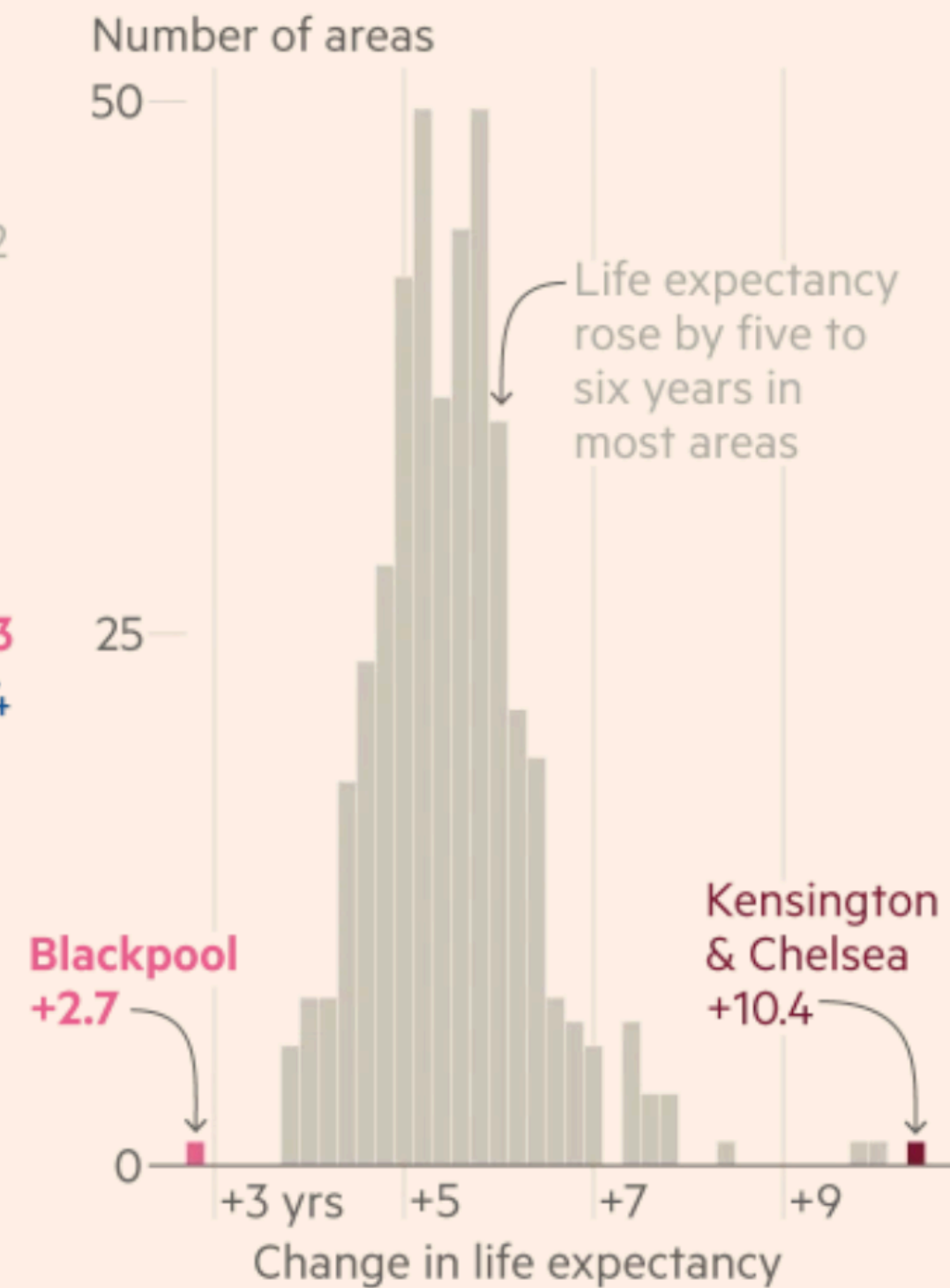
Visualization for communication isn't just about drawing things,
but also about verbally or textually explaining those things

Boys born in **Blackpool** can expect to live just 74 years — the second lowest in the UK, and up by just 2.7 years since 1993

Male life expectancy at birth in selected local authorities, 1993-2015



Distribution of change in male life expectancy at birth from 1993 to 2015, all UK local authorities



Source: ONS

Graphic by John Burn-Murdoch / @jburnmurdoch

© FT

“I and my colleagues here at the FT, we really do think one of the most valuable things we can do as data visualization practitioners is add this expert annotation layer.”

John Burn-Murdoch

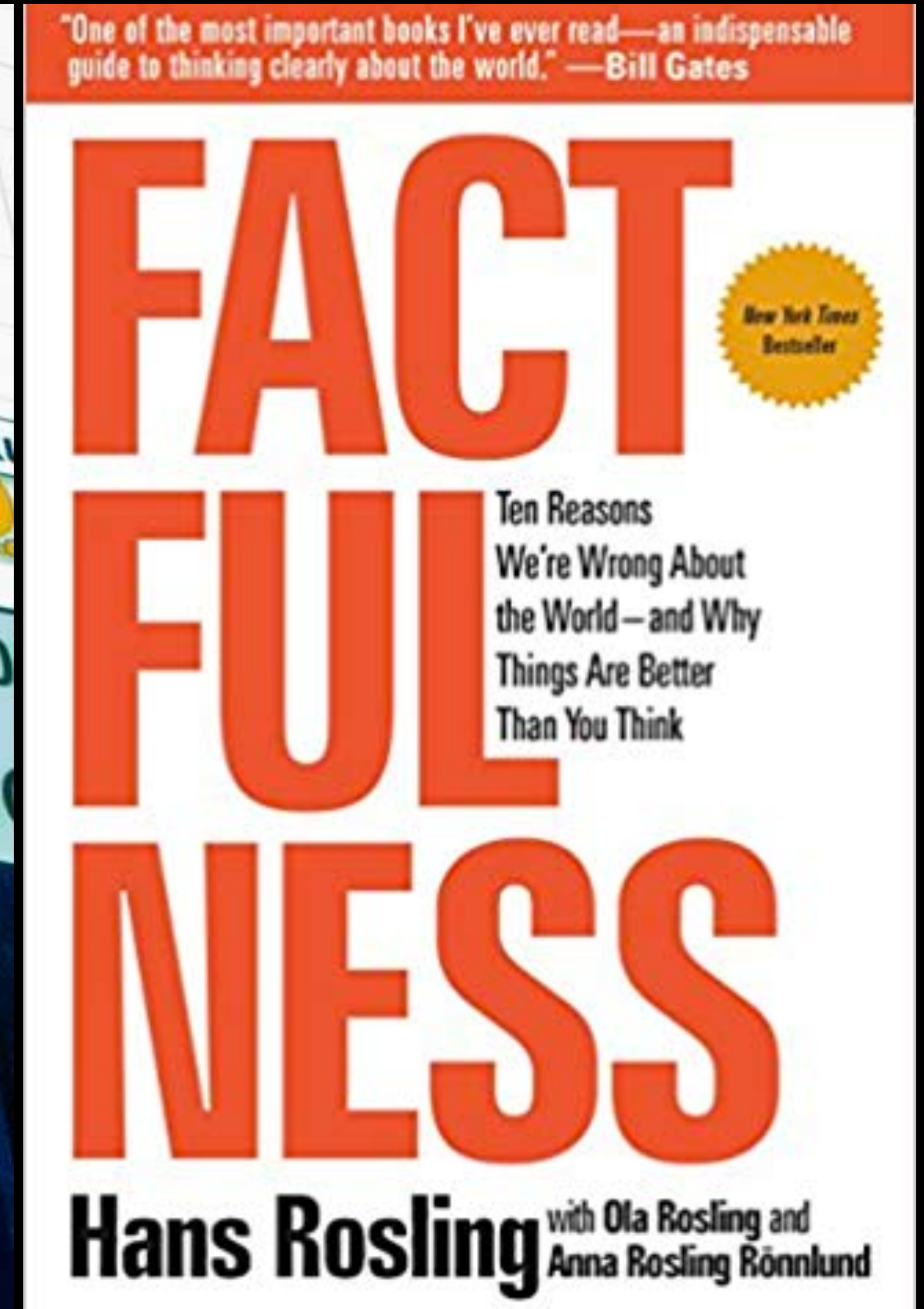
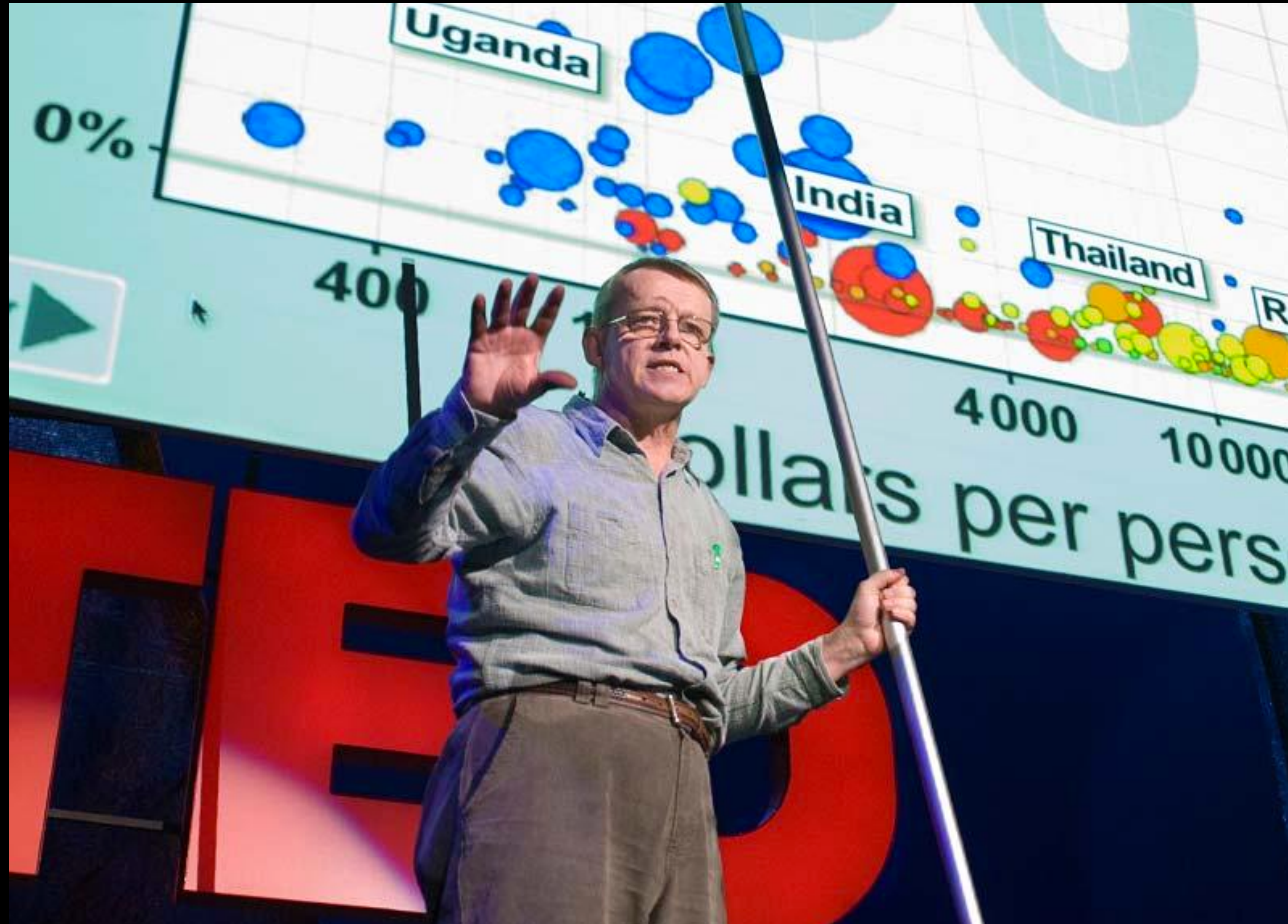
Financial Times

<https://policyviz.com/podcast/episode-155-john-burn-murdoch/>

“Design secrets behind the FT’s best charts of the year”

<https://www.ft.com/content/4743ce96-e4bf-11e7-97e2-916d4fbac0da>

Show **AND** tell



Hans Rosling, www.gapminder.org

Show **AND** tell

BBC FOUR



Hans Rosling, *The Joy of Stats*

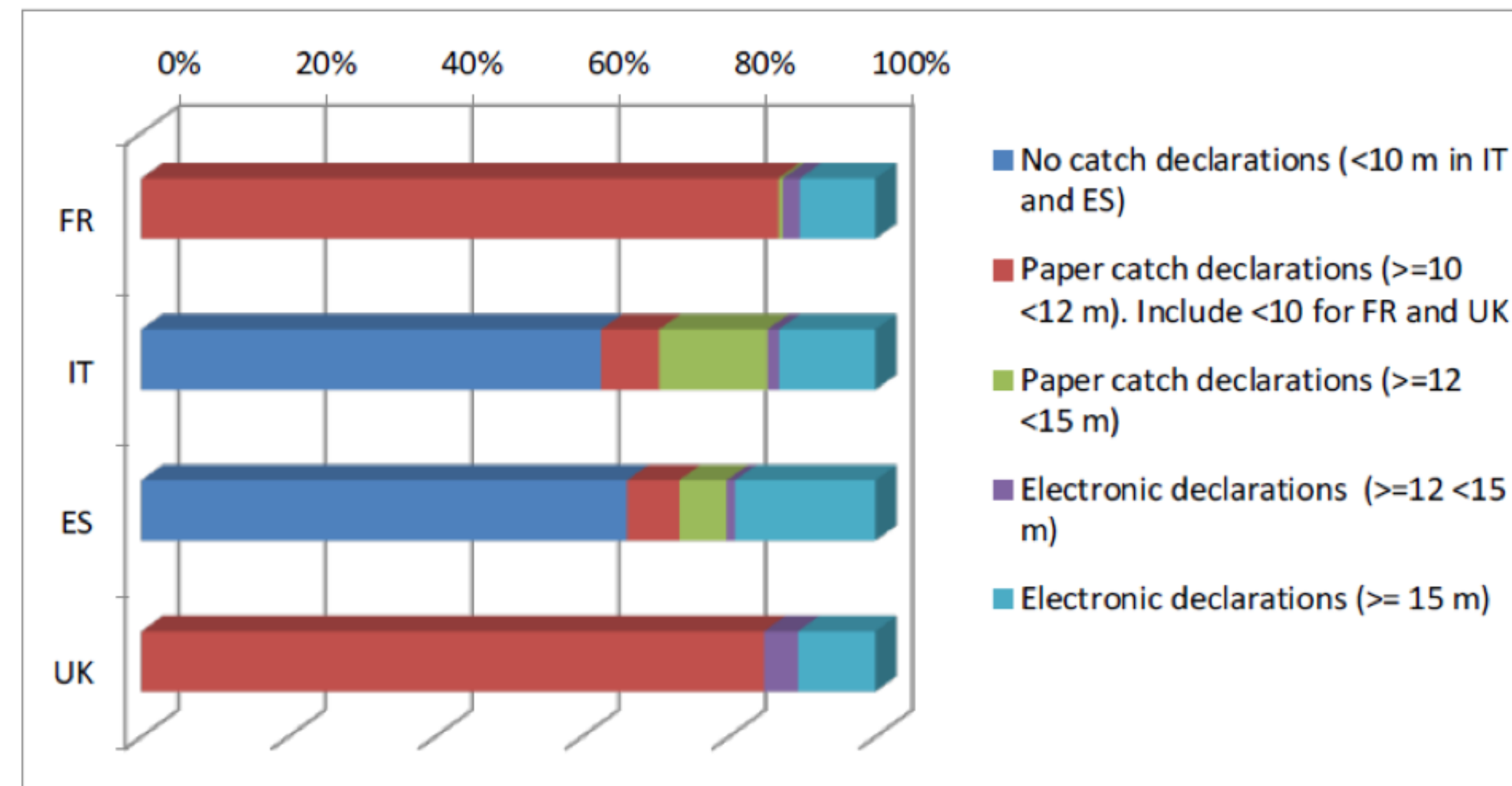


6. What visual design and style should I use?

Pay attention to visual design and style.
Polishing your design should never be an
afterthought. If after writing a paper, you polish
the language, you should do the same with charts.

(Also, don't trust software defaults!)

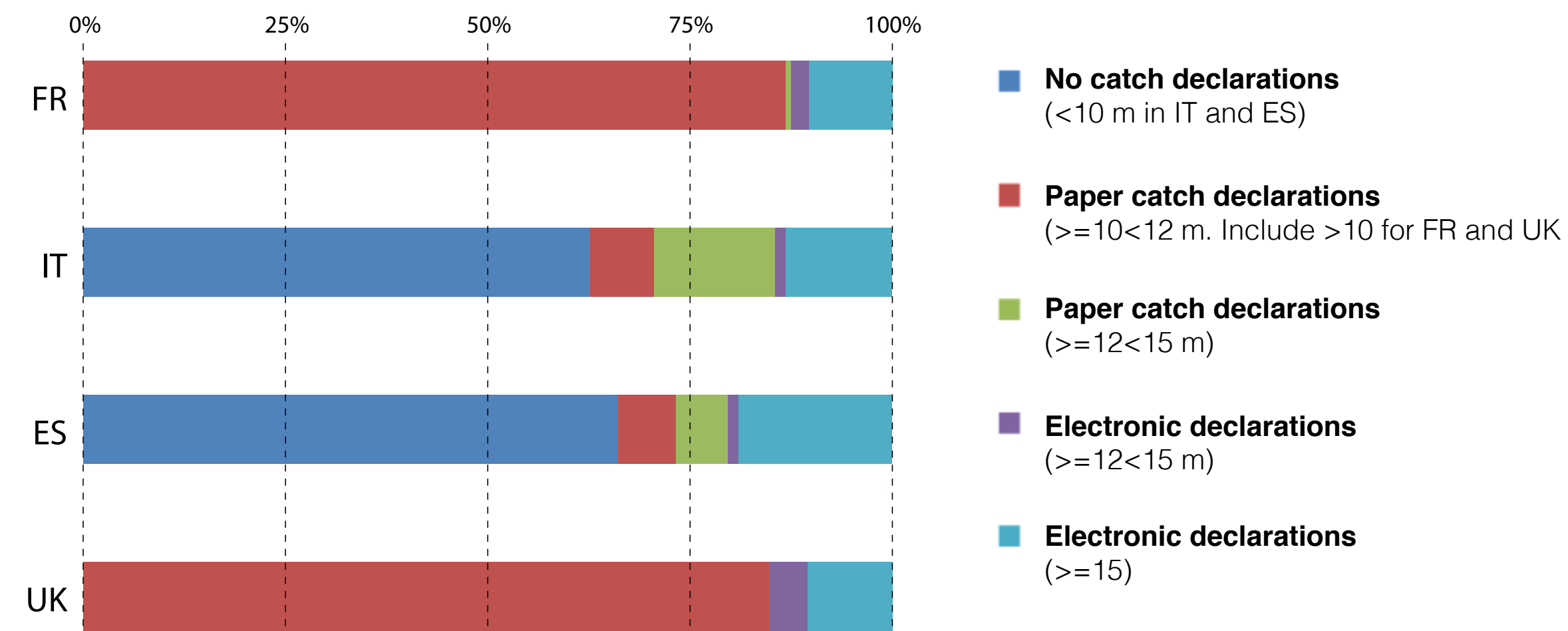
Figure 6 - Proportion of vessels that register catch and landing data, and data format



Pay attention to visual design and style.
Polishing your design should never be an
afterthought. If after writing a paper, you polish
the language, you should do the same with charts.

(Also, don't trust software defaults!)

Figure 6 - Proportion of vessels that register catch and landing data, and data format



Multi-scale Modeling and Assessment of Malaria Risk in Northern South America

Alimi, T. O.¹; Fuller, D. O.^{1,2} and Beier, J.C.^{1,3}

¹ Abess Center for Ecosystem Science and Policy; ² Department of Geography and Regional Studies; ³ Department of Epidemiology and Public Health, University of Miami

1. Introduction

The public health problem posed by malaria has made it a top priority for control efforts and the general consensus globally, is that its elimination is crucial for continued international development. Consequently, there is ongoing research in different regions including South America (SA) to better understand the disease dynamics with the intent that findings may establish scientific framework that would support the development of new intervention strategies for malaria elimination in areas with seasonal malaria. One of such investigations is undertaken by the International Centers of Excellence in Malaria Research (ICEMR) under a National Institutes of Health (NIH) grant.

While only about 3% of the global malaria burden is borne by SA¹, undertaking malaria research in the region is currently important because an estimated 23million people are still at risk² and approximately about 80% of clinical cases are found in Northern South America (NSA)³. A key factor limiting effective control is lack of data and uneven implementation of control measures, including use of bed-nets, sprays, early diagnosis, and treatment. As part of the ICEMR investigation, this project seeks to model the spatial patterns of malaria risk in NSA through vector distribution and land-use changes. Furthermore, I intend to investigate the perceptions of malaria risk in order to identify barriers to adoption and how they can be circumvented.

2. Significance

Spatial distribution of malaria risk is still perceived as broadly categorized by the WHO's traditional risk maps which are highly generalized, of low resolution and have broad categories with uncertain boundaries (see da Nunes-Silva et al. 2012). There is need for up- to-date high resolution risk maps which can aid malaria control efforts. Secondly, modeling distribution of principal malaria vectors and land use changes which may explain the observed distribution and risk are useful tools which would guide future management strategies. Finally, understanding the perceptions of at risk populations may help address barriers to adoption of interventions and influence policies. Overall, findings will empower NMCPs to achieve effective control and move them closer to elimination.

3. Specific Aims

- Specific Aim 1: Model the spatial patterns of malaria risk through vector distribution and land use changes
 - Hypothesis 1.1: GIS-based Multi-Criteria Evaluation (MCE) model can accurately predict spatial extent of malaria risk areas. **Objective:** Generate risk maps that represent risk of malaria transmission.
 - Hypothesis 1.2: The Maximum Entropy (Maxent) model can accurately depict actual and predict potential distribution of three *Anopheles* species. **Objective:** Model observed and potential spread of *An. albimanus*, *An. darlingi*, and *An. nuneztovari*.
 - Hypothesis 1.3: Land- use changes can explain the variations in predicted malaria risk. **Objective:** Characterize land use land cover (LULC) and investigate changes in areas of risk.
- Specific Aim 2: Investigate the perceptions of malaria risk in order to identify barriers to adoption and how they can be circumvented.
 - Hypothesis 2.1: Knowledge of perception of malaria risk can aid design of malaria control strategies. **Objective:** Obtain and analyze data on subjective perceptions of risk.
 - Hypothesis 2.2: Identification of barriers to adoption of malaria control interventions provide means of tackling them. **Objective:** Analyze data addressing perceived barriers and policy implications

*Only ongoing work on Hypothesis 1.1 in presented here

4. Materials and Methods

- **Study Area:** is NSA comprising of ten countries- Bolivia, Brazil, Colombia, Ecuador, French Guiana, Guyana, Panama, Peru, Suriname and Venezuela. These countries account for approximately 90% of clinical cases in the region hence, the choice as study area (Fig. 1).



Figure1: Map of study area

- **Research Approach:** Due to the complexity of malaria problem, I'm employing an interdisciplinary approach to address the problem (Fig. 2).

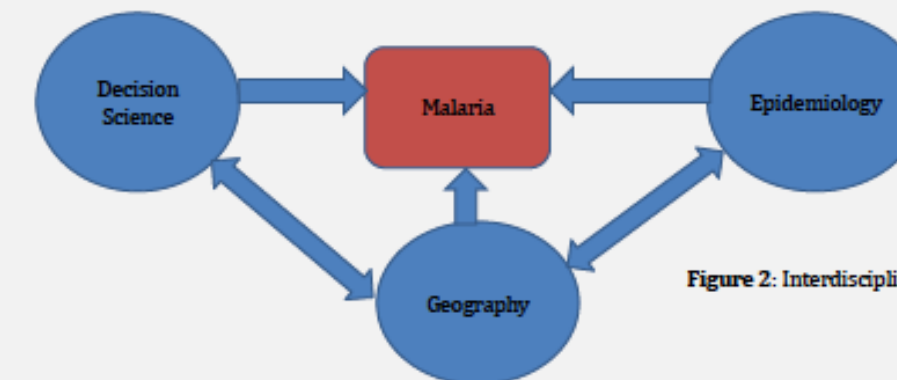


Figure 2: Interdisciplinary approach

- **Materials:** Raster data layers of environmental, climatic and anthropogenic parameters from satellite imageries, weather monitoring stations, global land cover and population data were collected from Worldclim, Digital Charts of the World, Globcover and Landsat. Vector data was collected from field sampling by our collaborators and the Walter Reed Biosystematics Unit. Sociological data would be collected through questionnaires to be administered in one of the study area. Other data will be collected as needed.
- **Procedure:** To test hypothesis 1.1, raster data of parameters that influence mosquito distribution (rivers, wetlands, urban areas, roads, population and elevation) were combined using a Multi-Criteria Evaluation in Idrisi GIS package. This produced a map of potential exposure to malaria vectors which is used as a proxy for risk of malaria transmission. All the data layers were gridded at 1km spatial resolution. A set of distance layers had been created for discrete factors using standard GIS operations. All factors were subsequently standardized into a continuous common numeric range on a byte 0-255 probability scale using a fuzzy function based on knowledge of mosquito interaction with the factor. Weights were generated for each factor based on the importance of the factor to malaria transmission by expert opinions and then assigned using Analytical Hierarchy Process. The risk maps produced were validated statistically using data on *An. darlingi* distribution and malaria case data from some parts of the study area. See preliminary results (Fig. 3,4,5)

5. Preliminary Results

- Areas of high to moderate risk corresponded with locations of some of the anophelines collected.

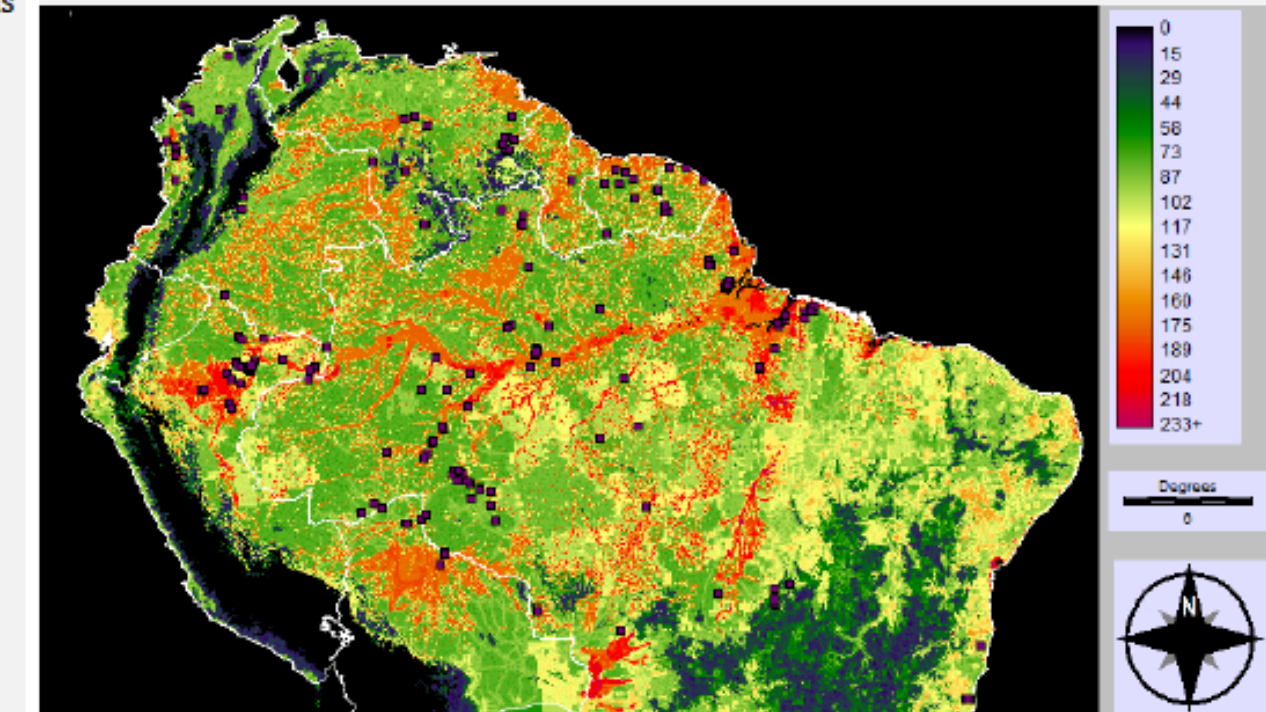


Figure 3: Potential risk of exposure to malaria vectors across NSA (0 indicate little or no risk while 233 indicate high risk)

- Risk scores for mosquito occurrence points were significantly higher than those generated randomly (Fig. 4).

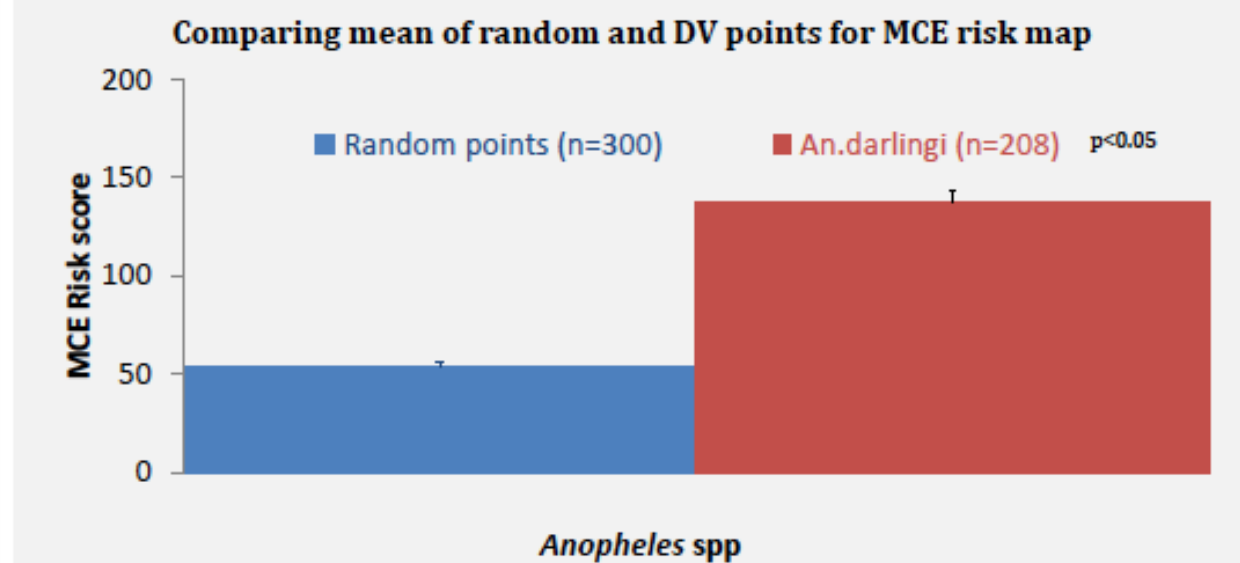


Figure 4: Plot showing the MCE risk values for randomly sampled points and for occurrence points of a DV, *An. darlingi*

6. Conclusion

Findings from preliminary results suggest that the MCE approach is a viable method to modeling spatial risk. The high resolution risk map produced aligned well with sampled vector points and may therefore be used to plan control of malaria vectors. Further analysis is planned to generate and validate risk maps with actual measures of malaria transmission, results of which could be used to plan containment of future outbreaks.

References

1. WHO. (2007). MALARIA ELIMINATION: A field manual for low and moderate endemic countries
2. PAHO (2012) PAHO Honors 2012 Malaria Champions of the Americas. Available: http://new.paho.org/hq/index.php?option=com_content&view=article&id=7429&Itemid=39639
3. Gusmao R. (1999) Overview of malaria control in the Americas. Parasitologia 41:355-60.
4. Da Silva-Nunes, M., Moreno, M., Conn, J.E., Gamboa, D., Abeles, S., Vinetz, J.M., and Ferreira, M.U. (2012) Amazonian malaria: Asymptomatic human reservoirs, diagnostic challenges, environmentally driven changes in mosquito vector populations, and the mandate for sustainable control strategies. *Acta Tropica* 121 (3): 281-29

Multi-scale Modeling and Assessment of Malaria Risk in Northern South America

Alimi, T. O.¹; Fuller, D. O.^{1,2} and Beier, J.C.^{1,3}

INTRODUCTION

Malaria as a public health problem has become a priority for control efforts worldwide. The global consensus is that its elimination is crucial for continual development. Ongoing research projects in different regions, including South America (SA), try to improve our understanding of the disease dynamics. Their goal is to establish a new framework that would lead to new intervention strategies for malaria elimination in areas where the disease is seasonal. One of such investigations is undertaken by the International Centers of Excellence in Malaria Research (ICEMR) under a National Institutes of Health grant.

While only about 3% of the global malaria burden is borne by SA, undertaking malaria research in the region is currently important because an estimated 23 million people are still at risk and approximately about 80% of clinical cases are found in **Northern South America (NSA)**. A key factor limiting effective control is lack of data and uneven implementation of control measures, including use of bednets, sprays, early diagnosis, and treatment. As part of the ICEMR investigation, this project seeks to model the spatial patterns of malaria risk in NSA through vector distribution and land-use changes. Furthermore, I intend to investigate the perceptions of malaria risk in order to identify barriers to adoption and how they can be circumvented.

SIGNIFICANCE

Spatial distribution of malaria risk is still perceived as broadly categorized by the WHO's traditional risk maps which are highly generalized, of low resolution and have broad categories with uncertain boundaries (see da Nunes-Silva et al. 2012). There is need for up-to-date high resolution risk maps which can aid malaria control efforts. Secondly, modeling distribution of principal malaria vectors and land use changes which may explain the observed distribution and risk are useful tools which would guide future management strategies. Finally, understanding the perceptions of at risk populations may help address barriers to adoption of interventions and influence policies. Overall, findings will empower NMCPs to achieve effective control and move them closer to elimination.

AIMS

Specific Aim 1: Model the spatial patterns of malaria risk through vector distribution and land use changes

- **Hypothesis 1.1:** GIS-based Multi-Criteria Evaluation (MCE) model can accurately predict spatial extent of malaria risk areas. **Objective:** Generate risk maps that represent risk of malaria transmission
- **Hypothesis 1.2:** The Maximum Entropy (Maxent) model can accurately depict actual and predict potential distribution of three Anopheles species. **Objective:** Model observed and potential spread of *An. albimanus*, *An. darlingi*, and *An. nuneztovari*.
- **Hypothesis 1.3:** Land-use changes can explain the variations in predicted malaria risk. **Objective:** Characterize land use land cover (LULC) and investigate changes in areas of risk

Specific Aim 2: Investigate the perceptions of malaria risk in order to identify barriers to adoption and how they can be circumvented.

- **Hypothesis 2.1:** Knowledge of perception of malaria risk can aid design of malaria control strategies. **Objective:** Obtain and analyze data on subjective perceptions of risk.
- **Hypothesis 2.2:** Identification of barriers to adoption of malaria control interventions provide means of tackling them. **Objective:** Analyze data addressing perceived barriers and policy implications

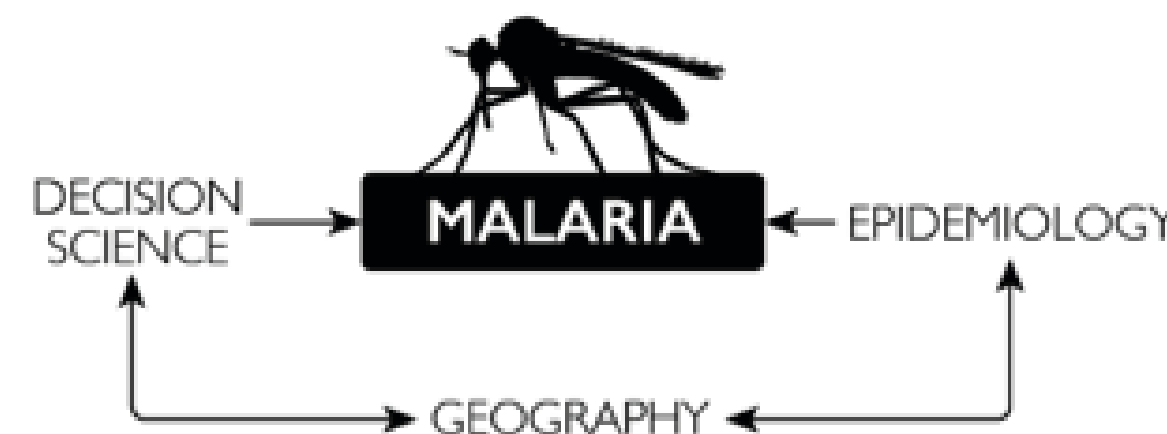
*Only ongoing work on Hypothesis 1.1 is presented here

MATERIALS AND METHODS

NSA comprising of ten countries - Bolivia, Brazil, Colombia, Ecuador, French Guiana, Guyana, Panama, Peru, Suriname and Venezuela. These countries account for approximately 90% of clinical cases in the region



Research approach: Due to the complexity of malaria problem, I'm employing an interdisciplinary approach to address the problem.

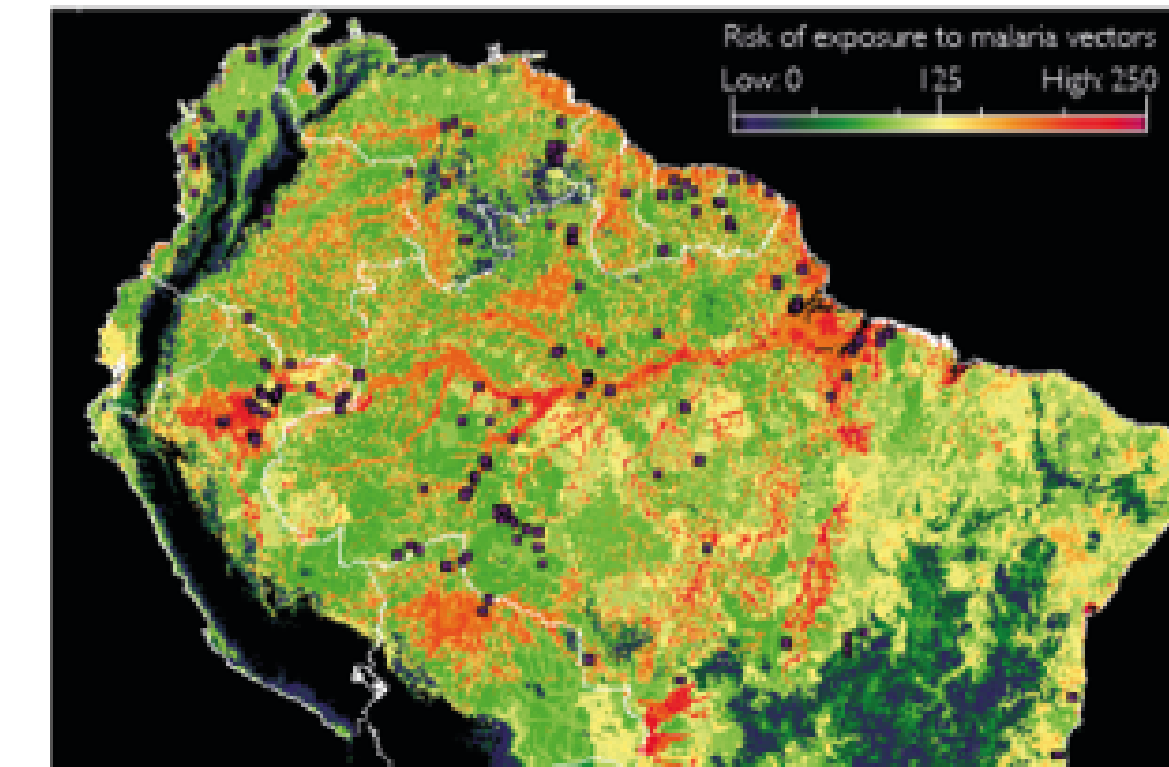


Materials: Raster data layers of environmental, climatic and anthropogenic parameters from satellite imageries, weather monitoring stations, global land cover and population data were collected from Worldclim, Digital Charts of the World, Globcover and Landsat. Vector data was collected from field sampling by our collaborators and the Walter Reed Biosystematics Unit. Sociological data would be collected through questionnaires to be administered in one of the study area. Other data will be collected as needed.

Procedure: To test hypothesis 1.1, raster data of parameters that influence mosquito distribution (rivers, wetlands, urban areas, roads, population and elevation) were combined using a Multi-Criteria Evaluation in Idrisi GIS package. This produced a map of potential exposure to malaria vectors which is used as a proxy for risk of malaria transmission. All the data layers were gridded at 1 km spatial resolution. A set of distance layers had been created for discrete factors using standard GIS operations. All factors were subsequently standardized into a continuous common numeric range on a byte 0-255 probability scale using a fuzzy function based on knowledge of mosquito interaction with the factor. Weights were generated for each factor based on the importance of the factor to malaria transmission by expert opinions and then assigned using Analytical Hierarchy Process. The risk maps produced were validated statistically using data on *An. darlingi* distribution and malaria case data from some parts of the study area. See preliminary results

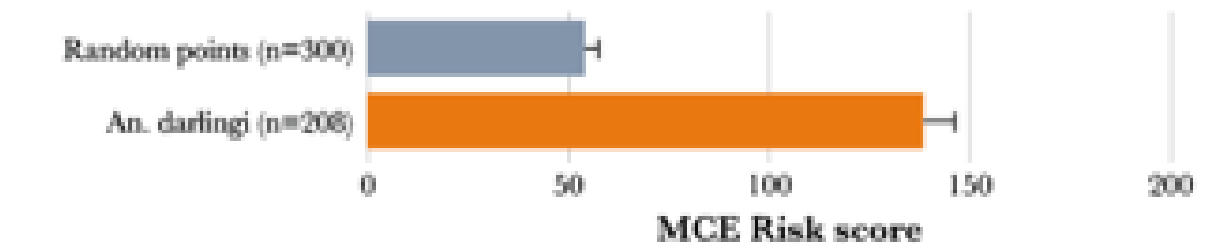
RESULTS

Areas of high to moderate risk corresponded with locations of some of the anophelines collected.



Risk scores for mosquito occurrence points were significantly higher than those generated randomly.

Comparing mean of random and DV points for MCE risk map - $p < 0.05$



CONCLUSION

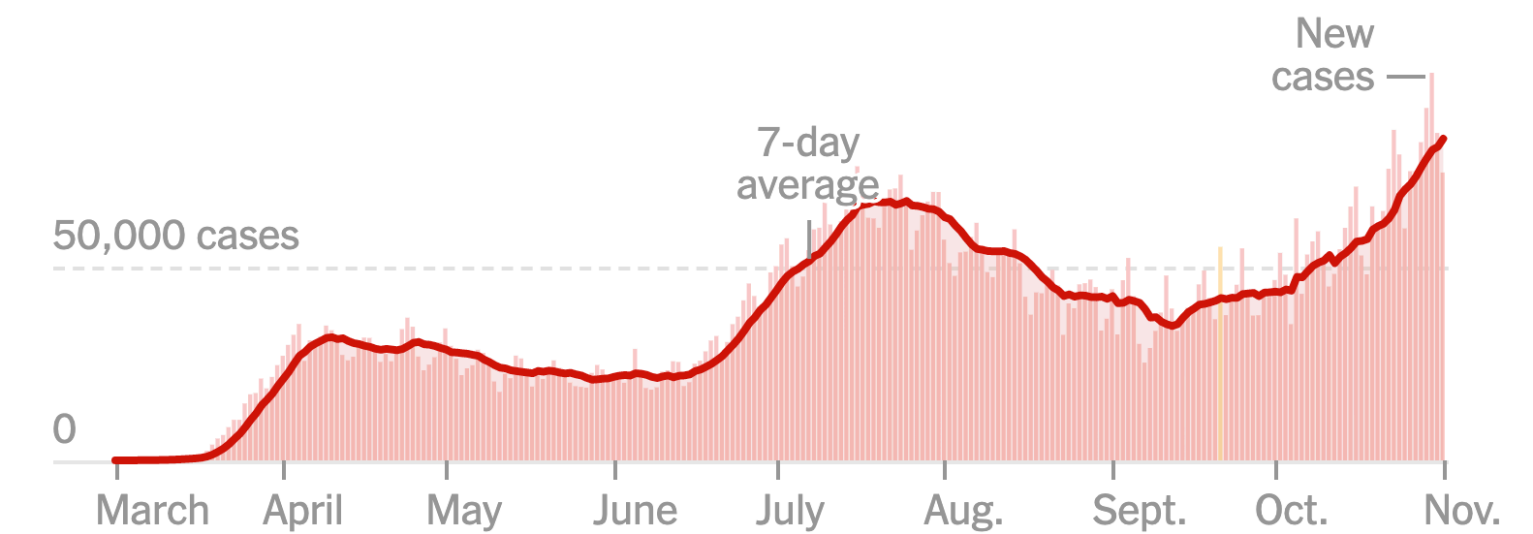
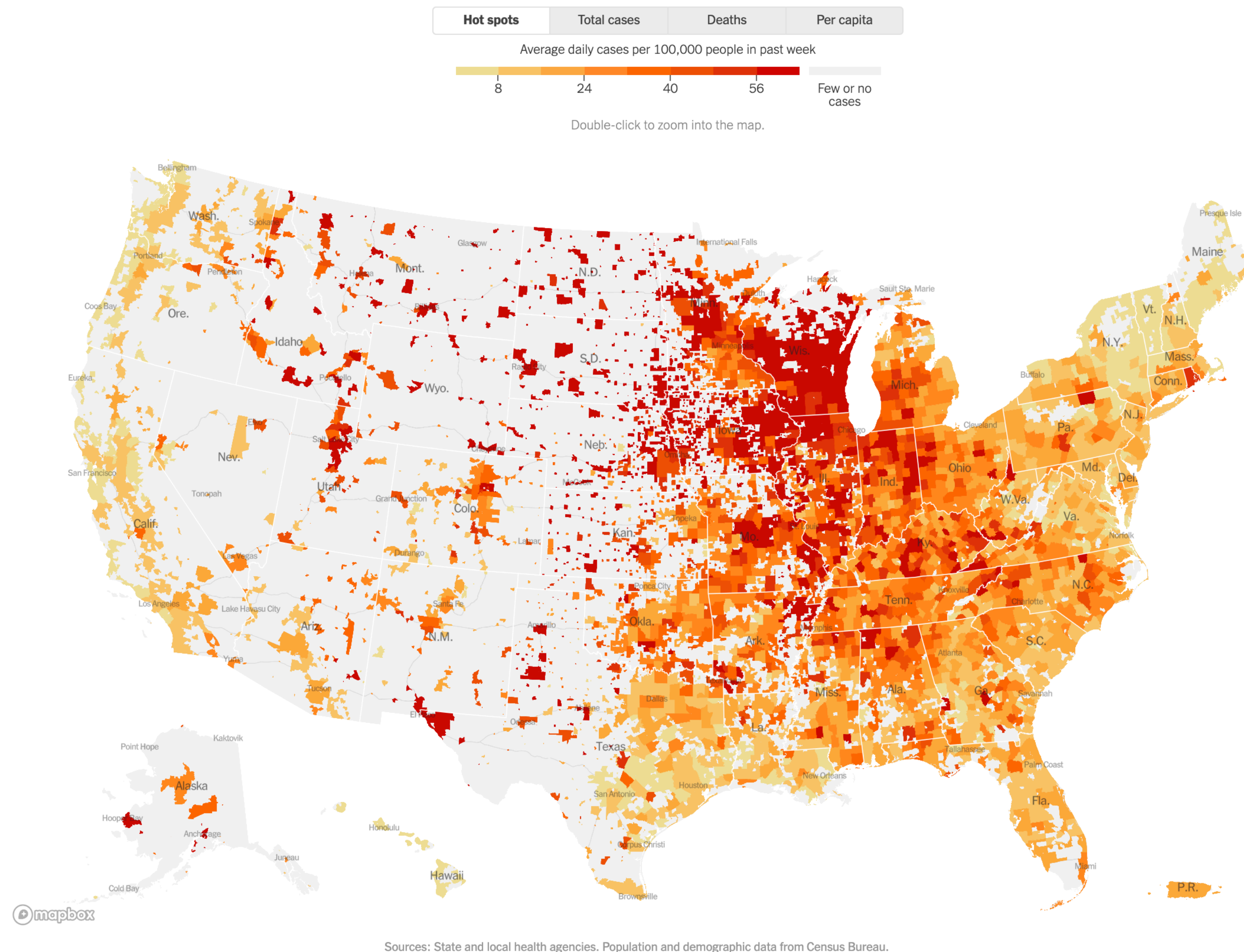
Findings from preliminary results suggest that the MCE approach is a viable method to modeling spatial risk. The high resolution risk map produced aligned well with sample vector points and may therefore be used to plan control of malaria vectors. Further analysis is planned to generate and validate risk maps with actual measures of malaria transmission, results of which could be used to plan containment of future outbreaks.

References

1. WHO. (2007). MALARIA ELIMINATION: A field manual for low and moderate endemic countries
2. PAHO. (2012). PAHO Honors 2012 Malaria Champions of the Americas. Available: http://www.paho.org/hq/index.php?option=com_content&view=article&id=7429&Itemid=39639
3. Guimao R. (1999) Overview of malaria control in the Americas. Parasitologia 41:355-60.
4. Da Silva-Nunes, M., Morono, M., Conn, J.E., Gamboa, D., Abeles, S., Vintz, J.M., and Ferreira, M.U. (2012) Amazonian malaria: Asymptomatic human reservoirs, diagnostic challenges, environmentally driven changes in mosquito vector populations, and the mandate for sustainable control strategies. Acta Tropica 121 (3): 281-29

Standard visualizations

Appropriate for graphics we use all the time



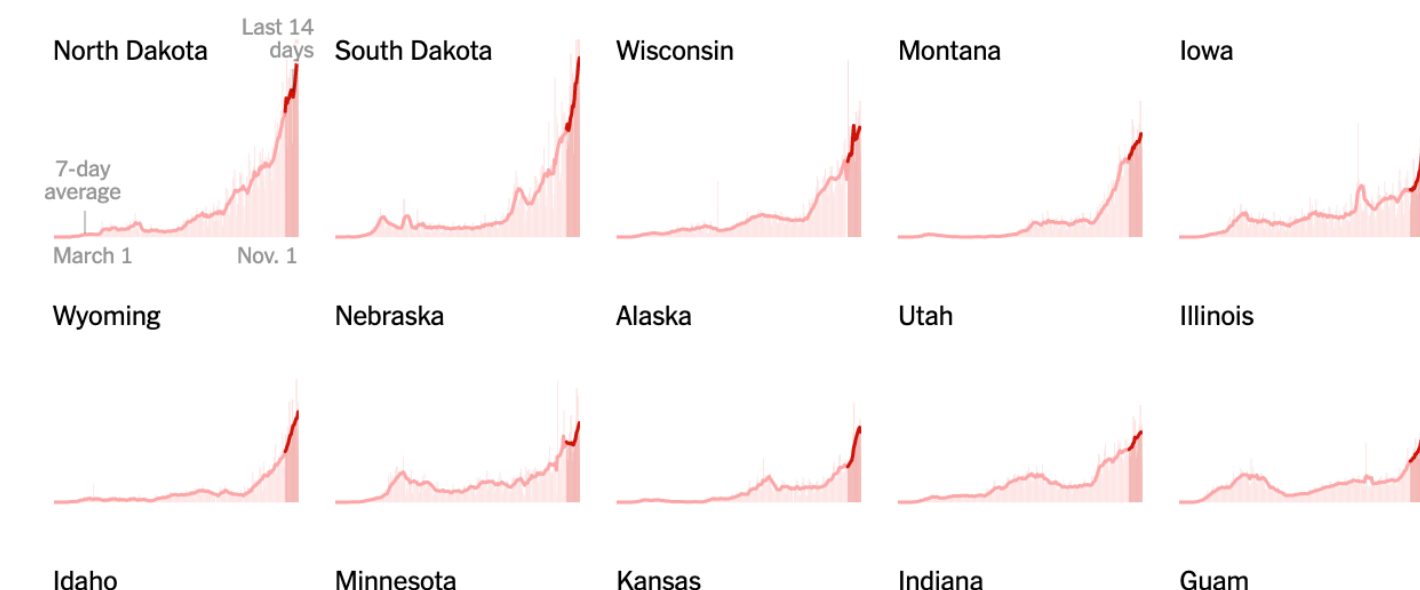
	TOTAL REPORTED	ON NOV. 1	14-DAY CHANGE
Cases	9.2 million+	74,113	+45% →
Deaths	230,937	427	+17% →

Day with data reporting anomaly.

Includes confirmed and probable cases where available. 14-day change trends use 7-day averages.

Where new cases are **higher** and **staying high**

States where new cases are higher had a daily average of at least 15 new cases per 100,000 people over the past week. Charts show daily cases per capita and are on the same scale. Tap a state to see detailed map page.



<https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html>

Fully customized style:

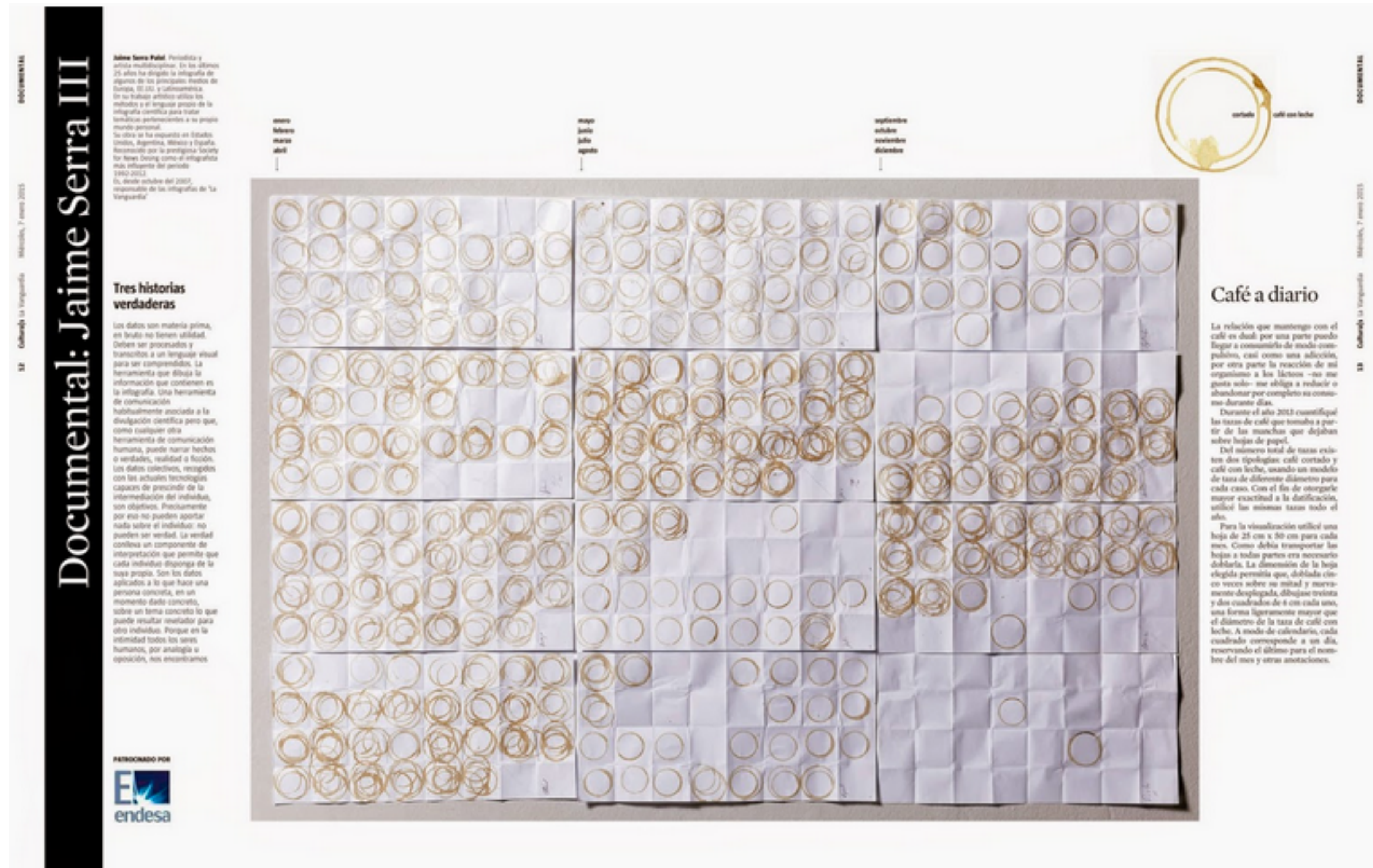
Appropriate for one-time use when we want to provoke curiosity, surprise —or simply a smile



<https://jaimeserra-archivos.blogspot.com/>

Fully customized style:

Appropriate for one-time use when we want to provoke curiosity, surprise —or simply a smile



Why does all this matter?

The purpose of visualization isn't visualization per se. The purpose of visualization is to help people **make sense of the world** through a combination of visuals and words.



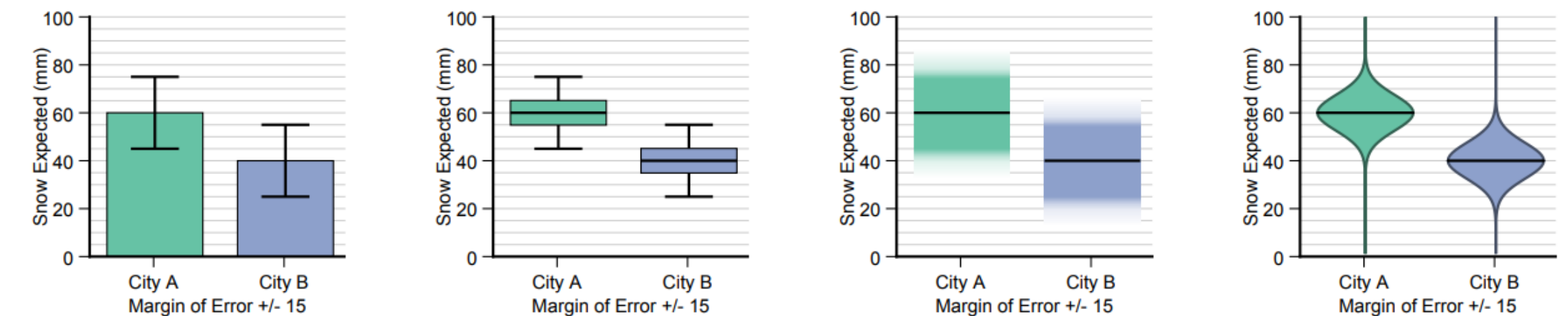
Where to go from here?
(reading recommendations)

Disclosing limitations and uncertainty

Uncertainty and graphicacy
How should statisticians, journalists, and designers reveal uncertainty in graphics for public consumption?

Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error

Michael Correll *Student Member, IEEE*, and Michael Gleicher *Member, IEEE*



(a) **Bar chart** with error bars: the height of the bars encodes the sample mean, and the whiskers encode a 95% t-confidence interval.

(b) **Modified box plot**: The whiskers are the 95% t-confidence interval, the box is a 50% t-confidence interval.

(c) **Gradient plot**: the transparency of the colored region corresponds to the cumulative density function of a t-distribution.

(d) **Violin plot**: the width of the colored region corresponds to the probability density function of a t-distribution.

<https://ec.europa.eu/eurostat/cros/powerfromstatistics/OR/PfS-OutlookReport-Cairo.pdf>

<https://graphics.cs.wisc.edu/Papers/2014/CGI4/Preprint.pdf>

Collection of papers about visualizing uncertainty:

<https://www.dropbox.com/sh/jk4ginxyai6ylqu/AABvqdyTlhJtyFN9nKNHyX9Ba?dl=0>

Articles and materials

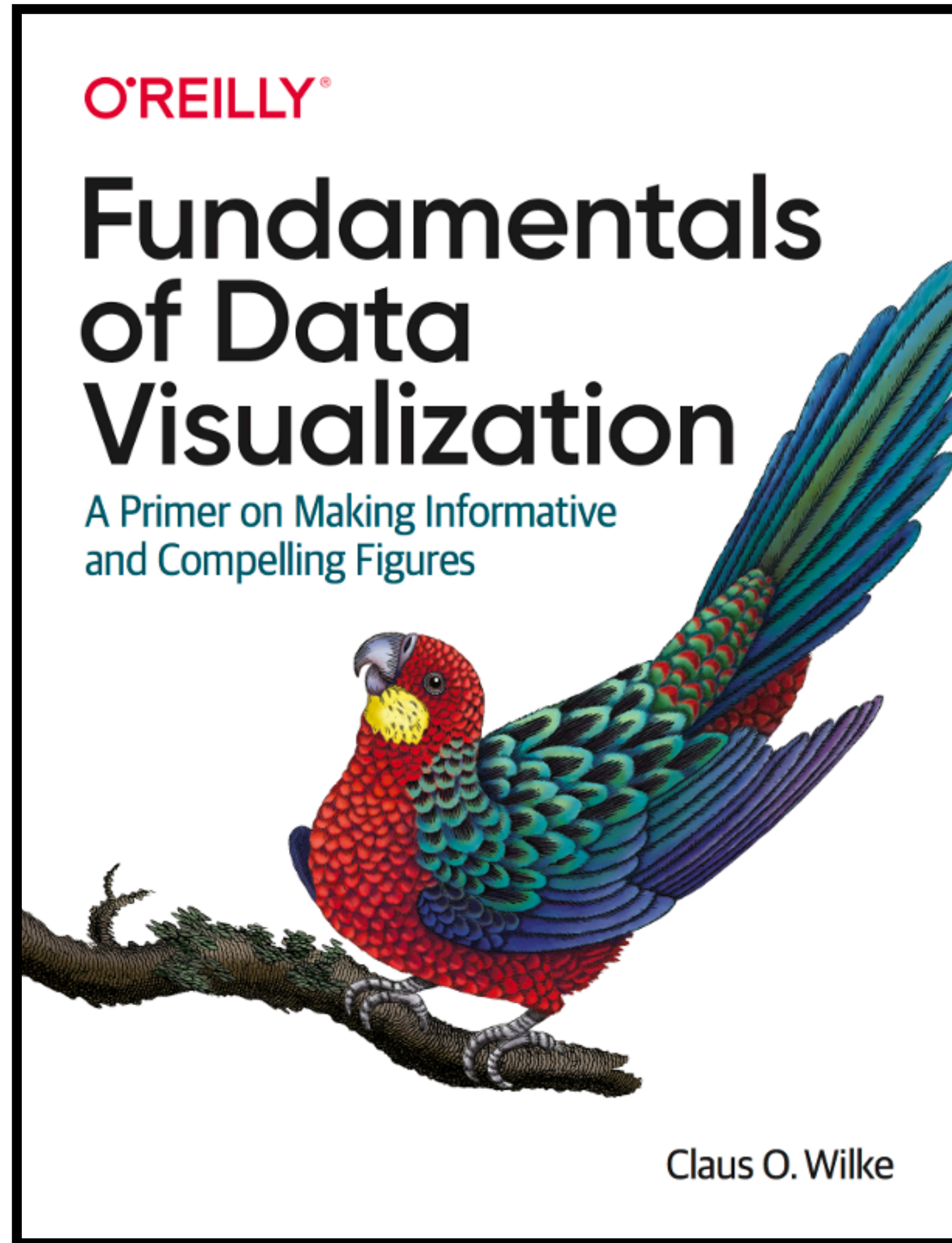
The Financial Times series of articles about data visualization

[https://www.dropbox.com/sh/9hbdxqsoel4n3zy/
AABhIk59ISo4dUb765DbLxjHa?dl=0](https://www.dropbox.com/sh/9hbdxqsoel4n3zy/AABhIk59ISo4dUb765DbLxjHa?dl=0)

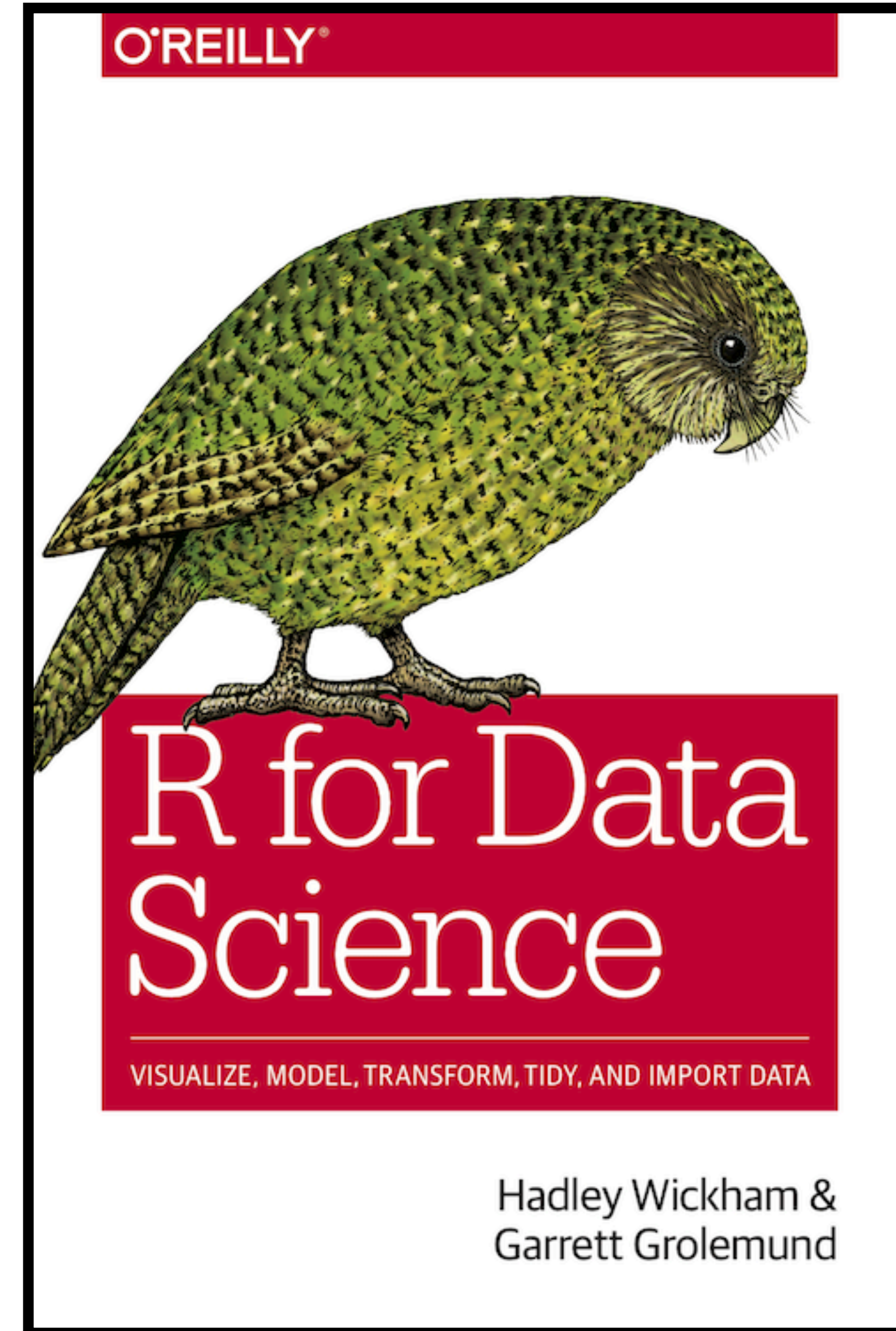
Various articles about data visualization

[https://www.dropbox.com/sh/elp6l2msxsawrkq/AABga-
vZ_0McK30_zETHEtWla?dl=0](https://www.dropbox.com/sh/elp6l2msxsawrkq/AABgavZ_0McK30_zETHEtWla?dl=0)

Books available online

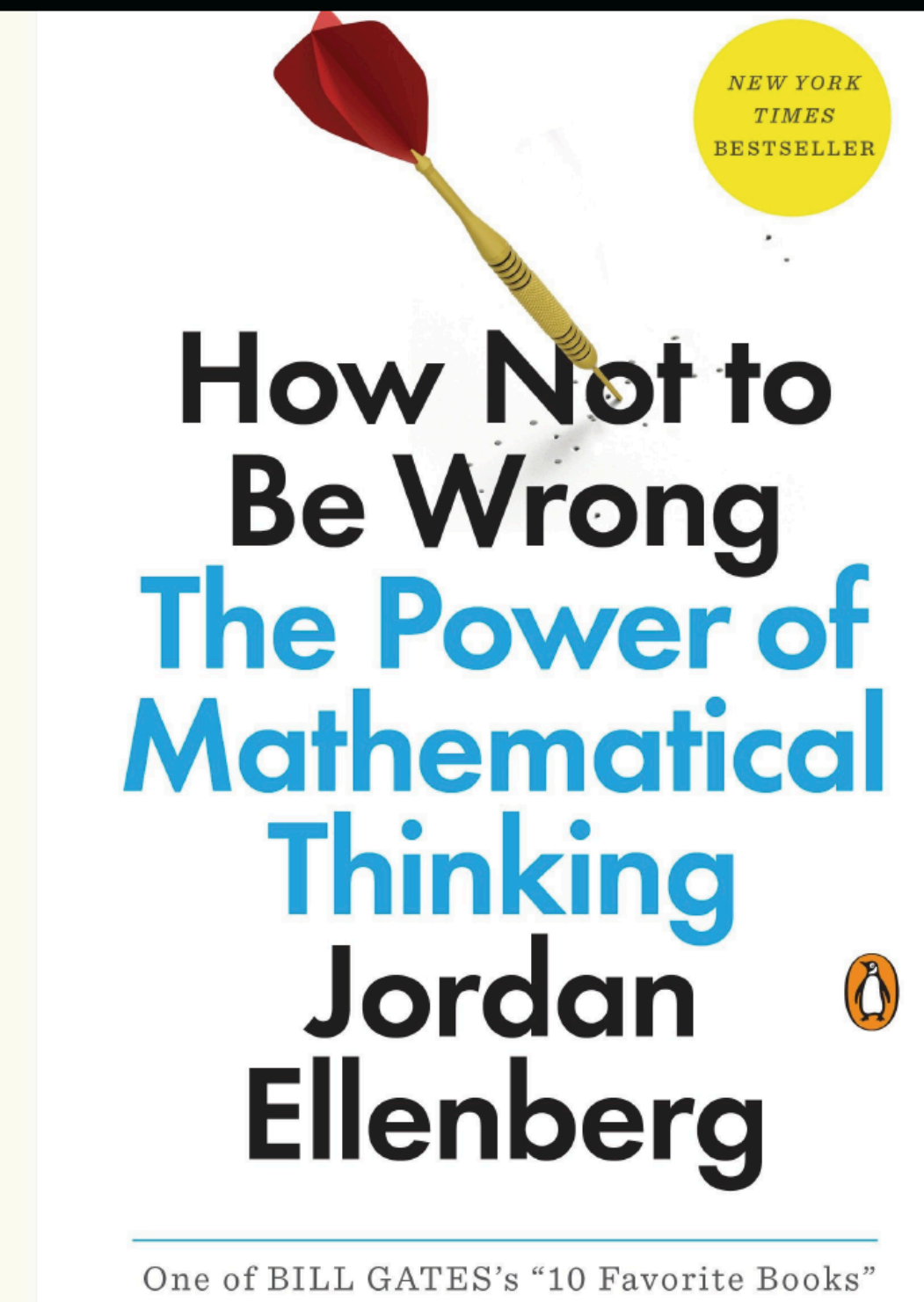
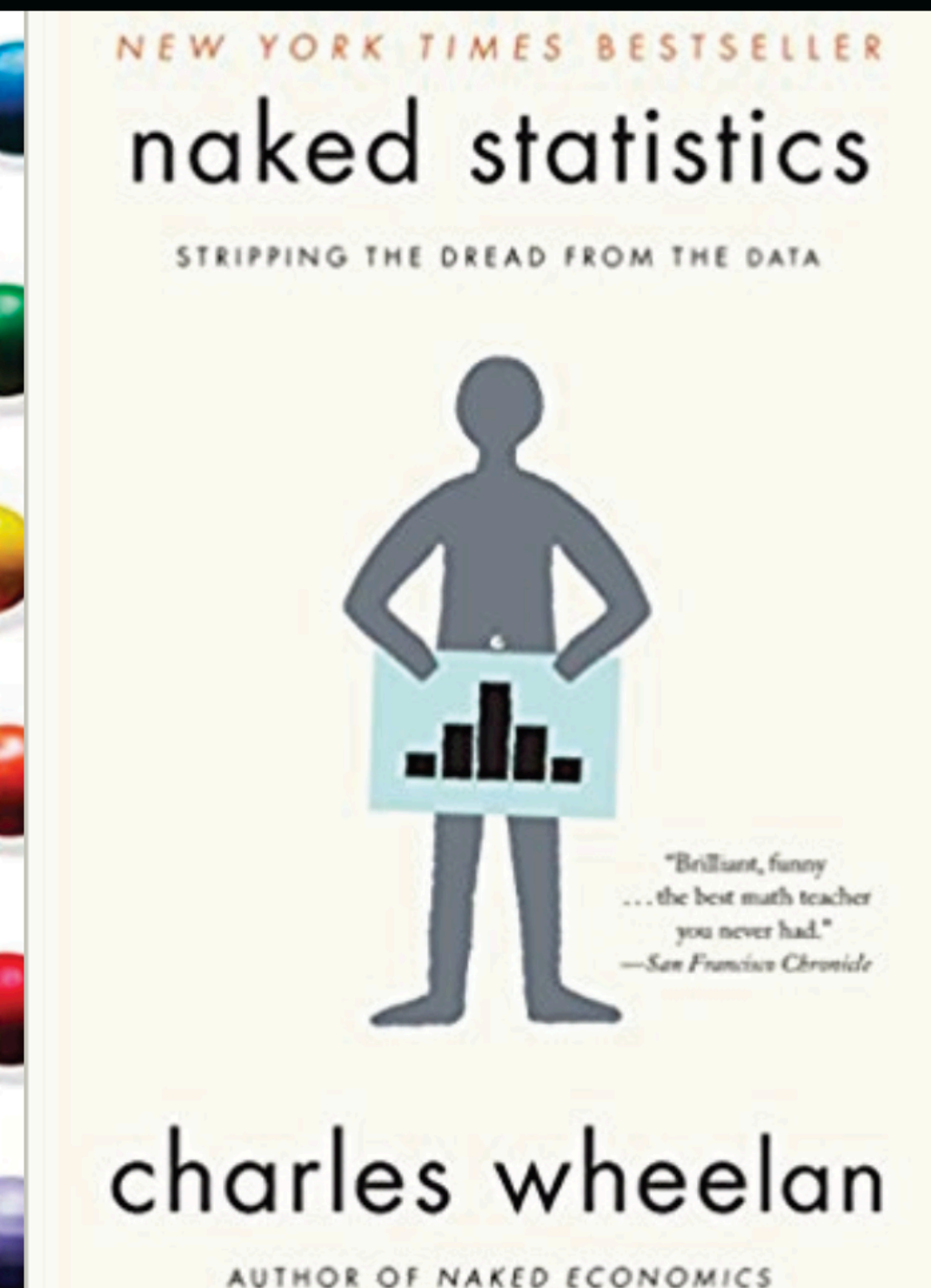
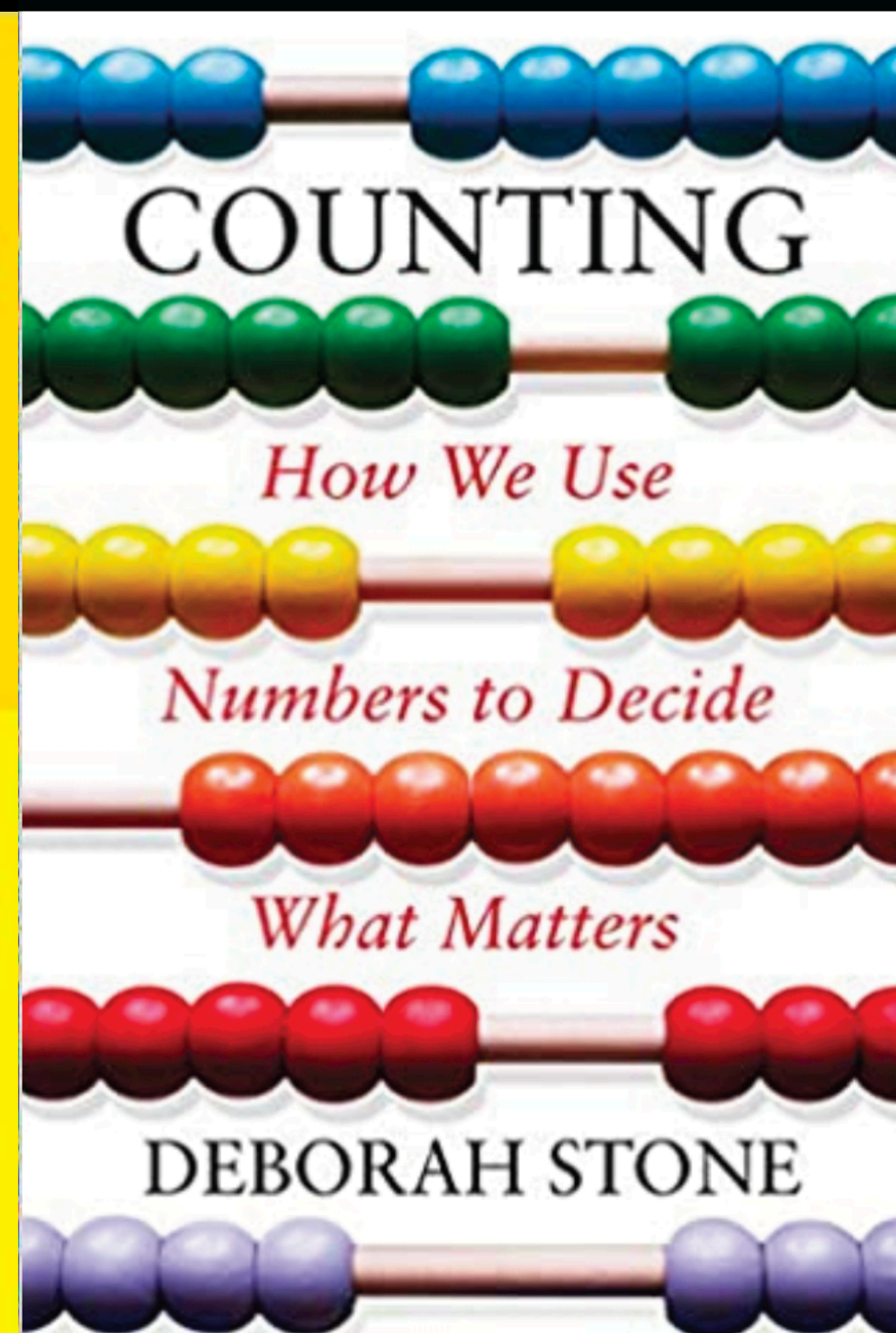
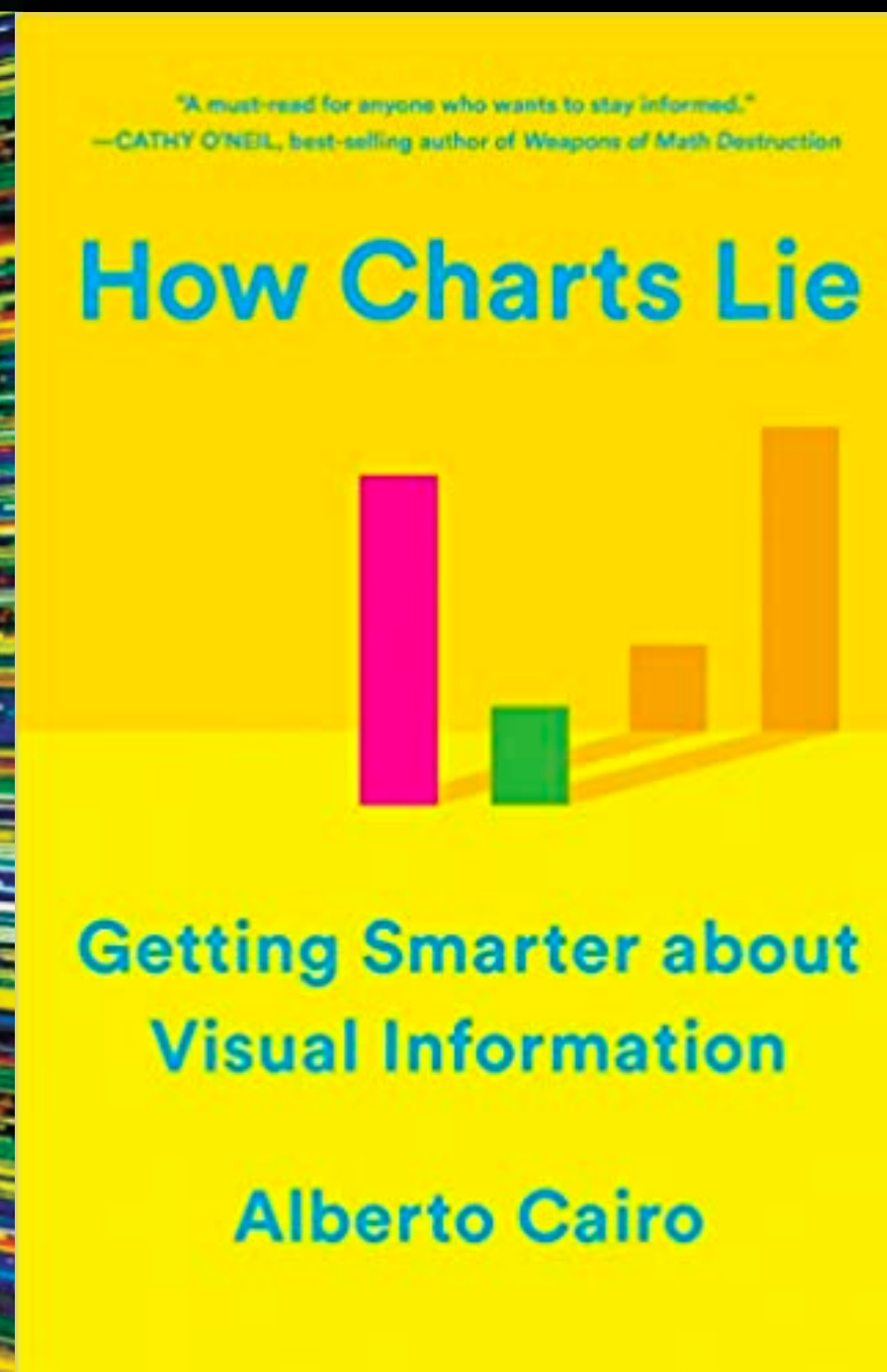
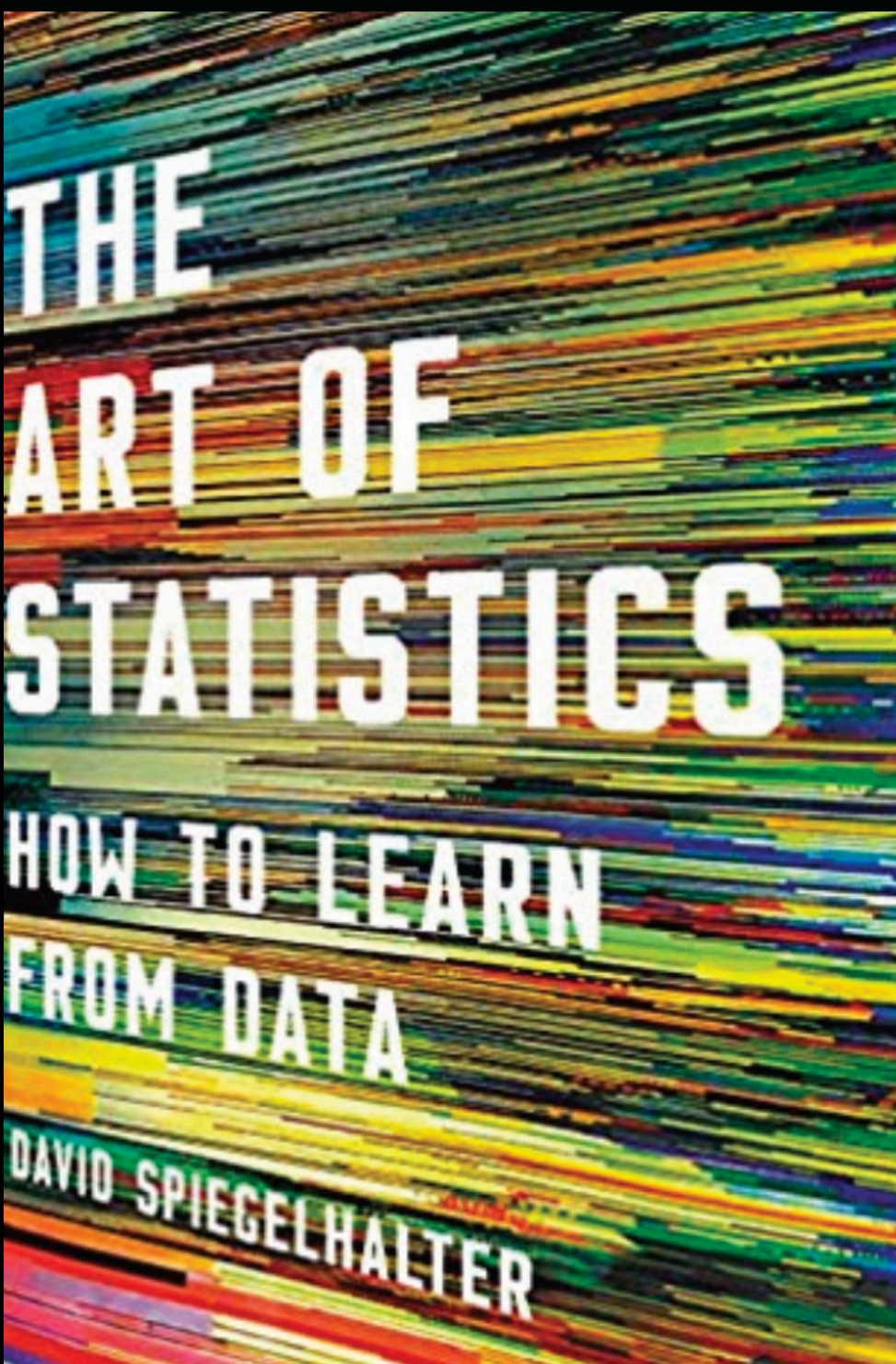


<https://serialmentor.com/dataviz/>

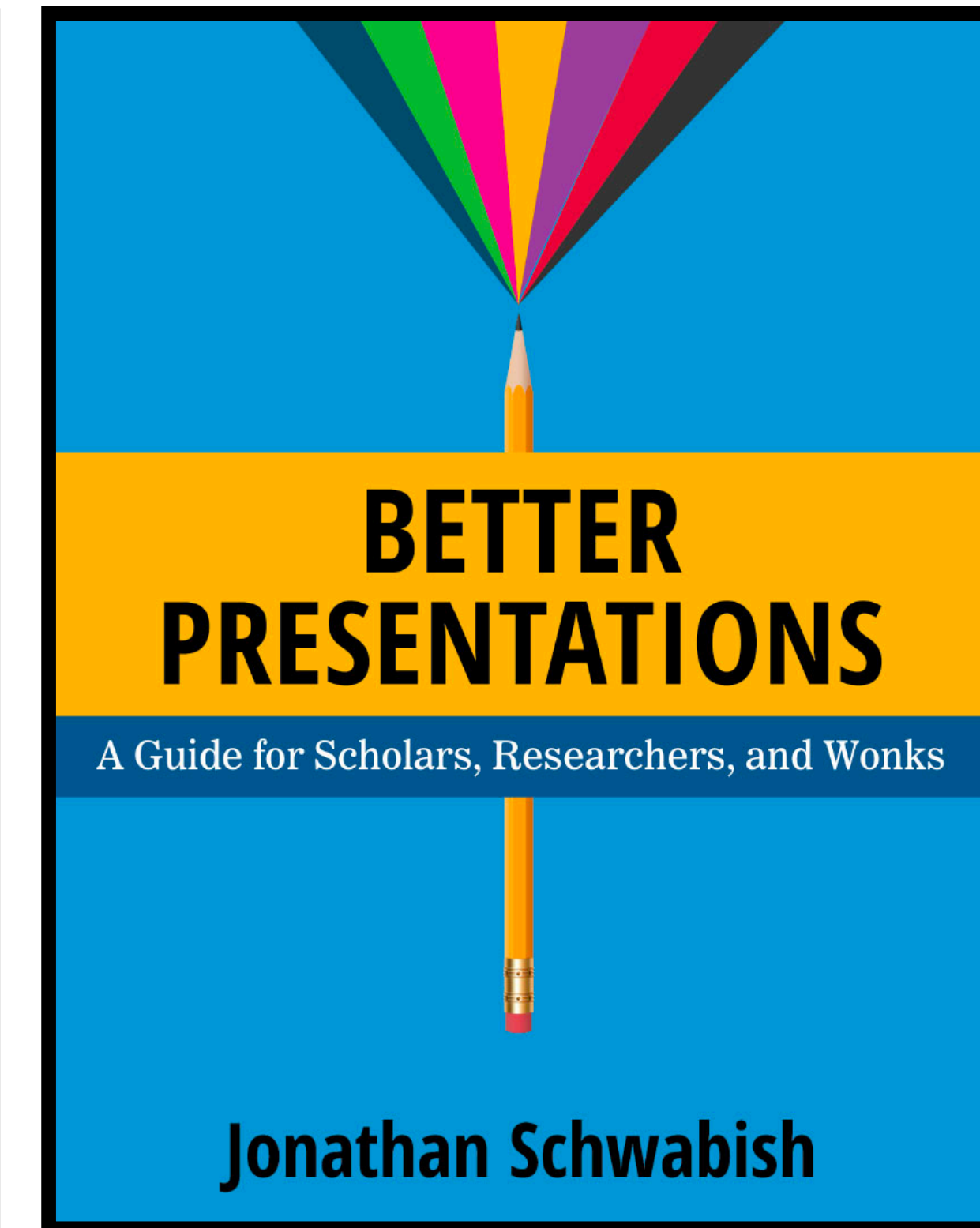
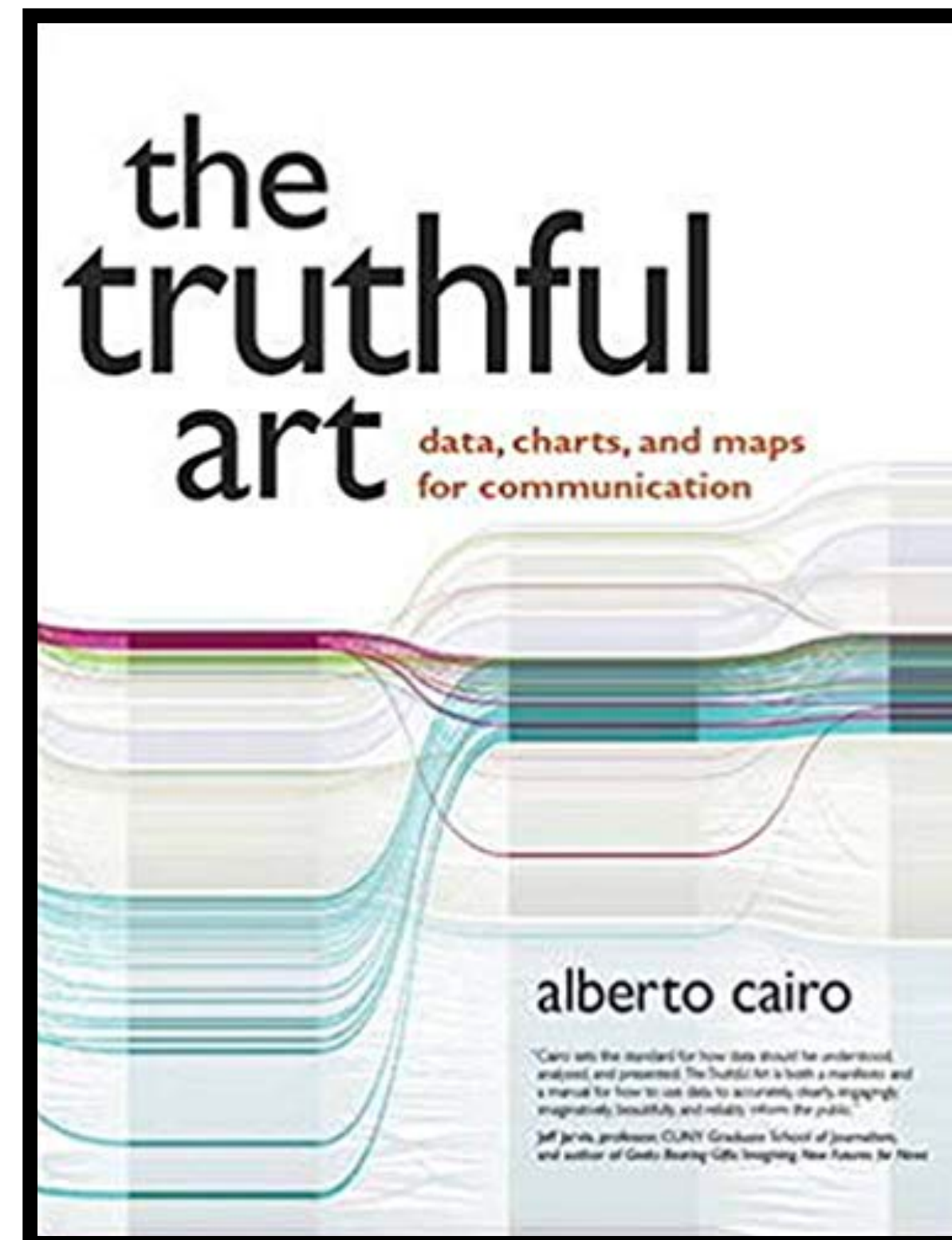
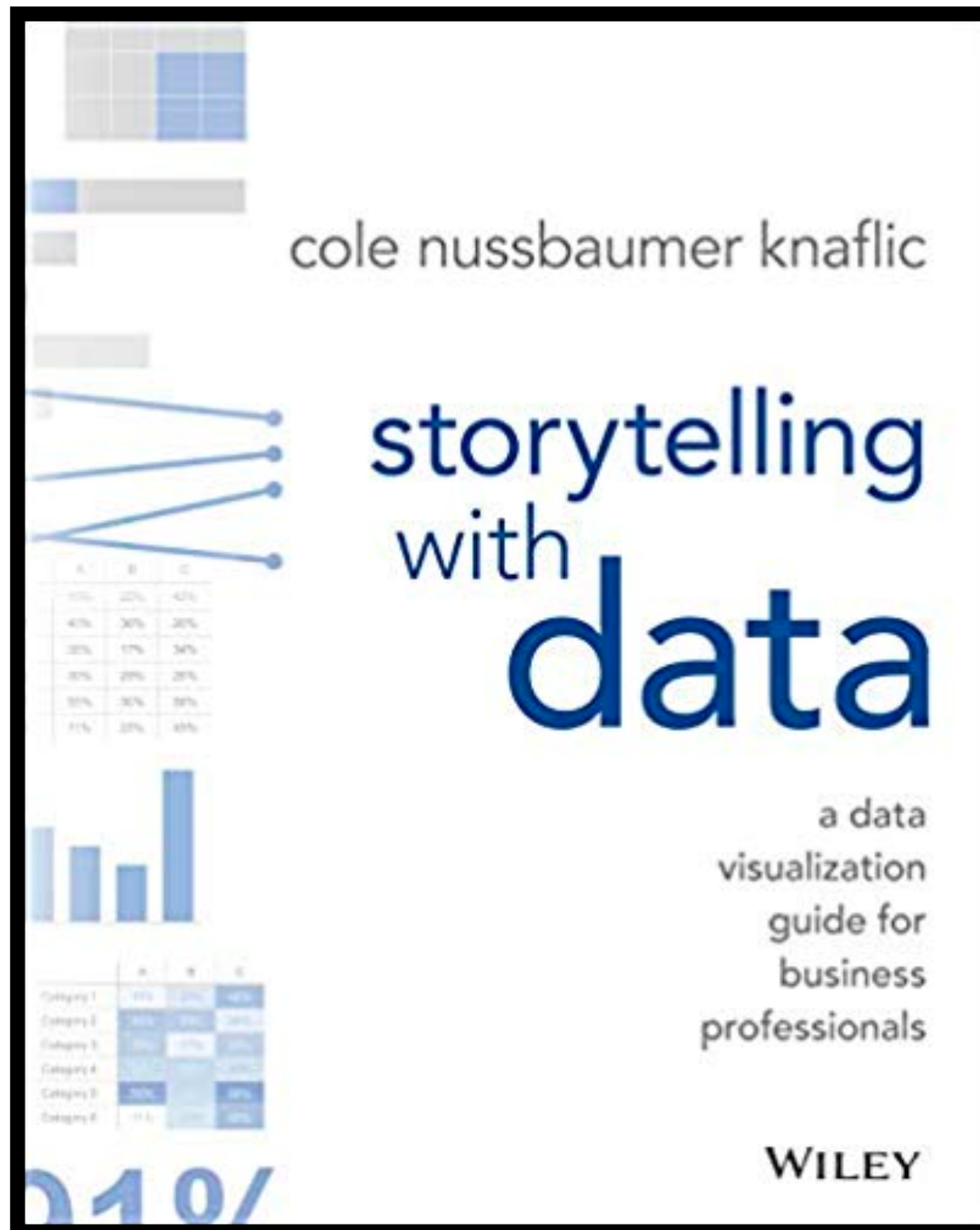


<https://r4ds.had.co.nz/>

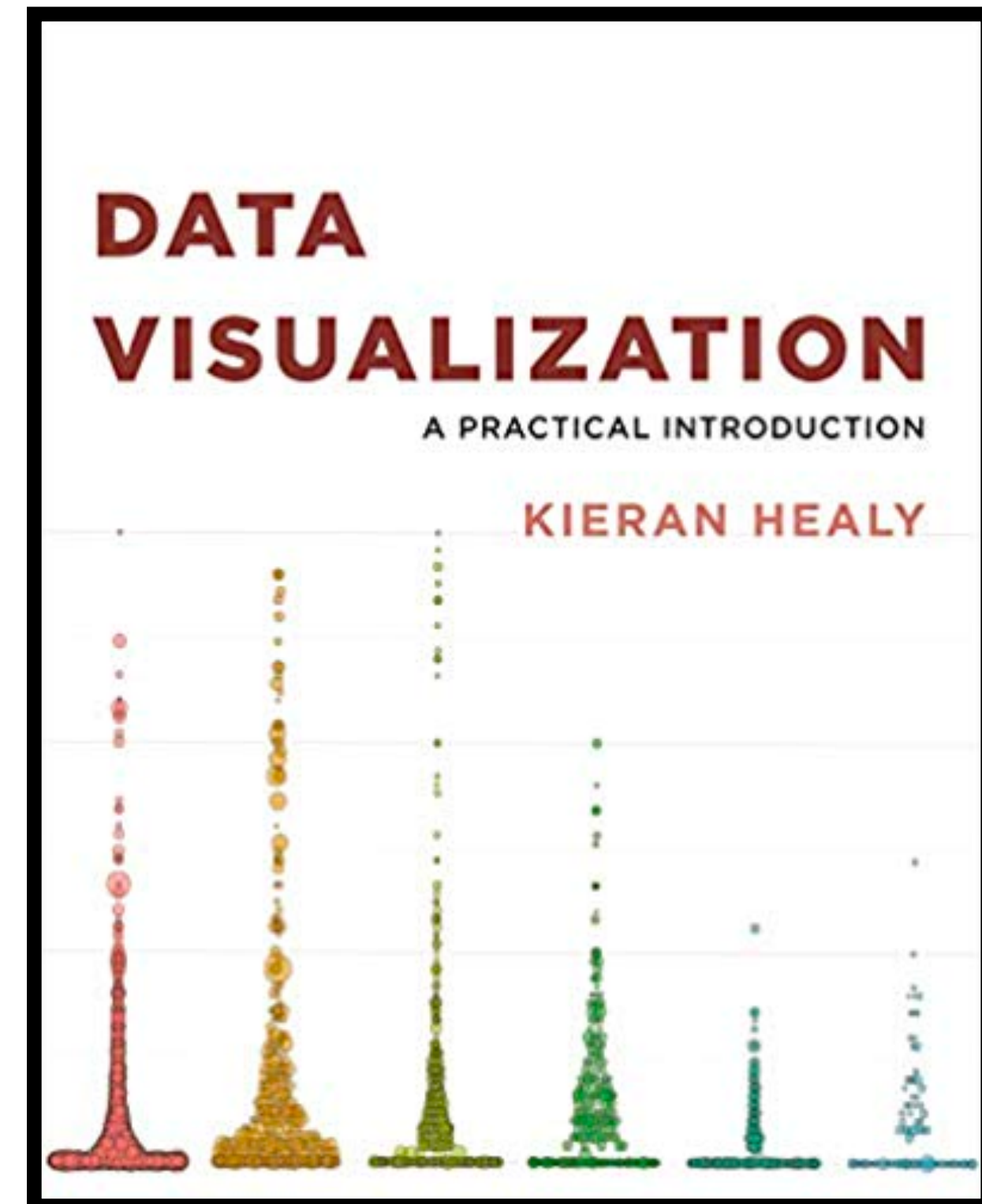
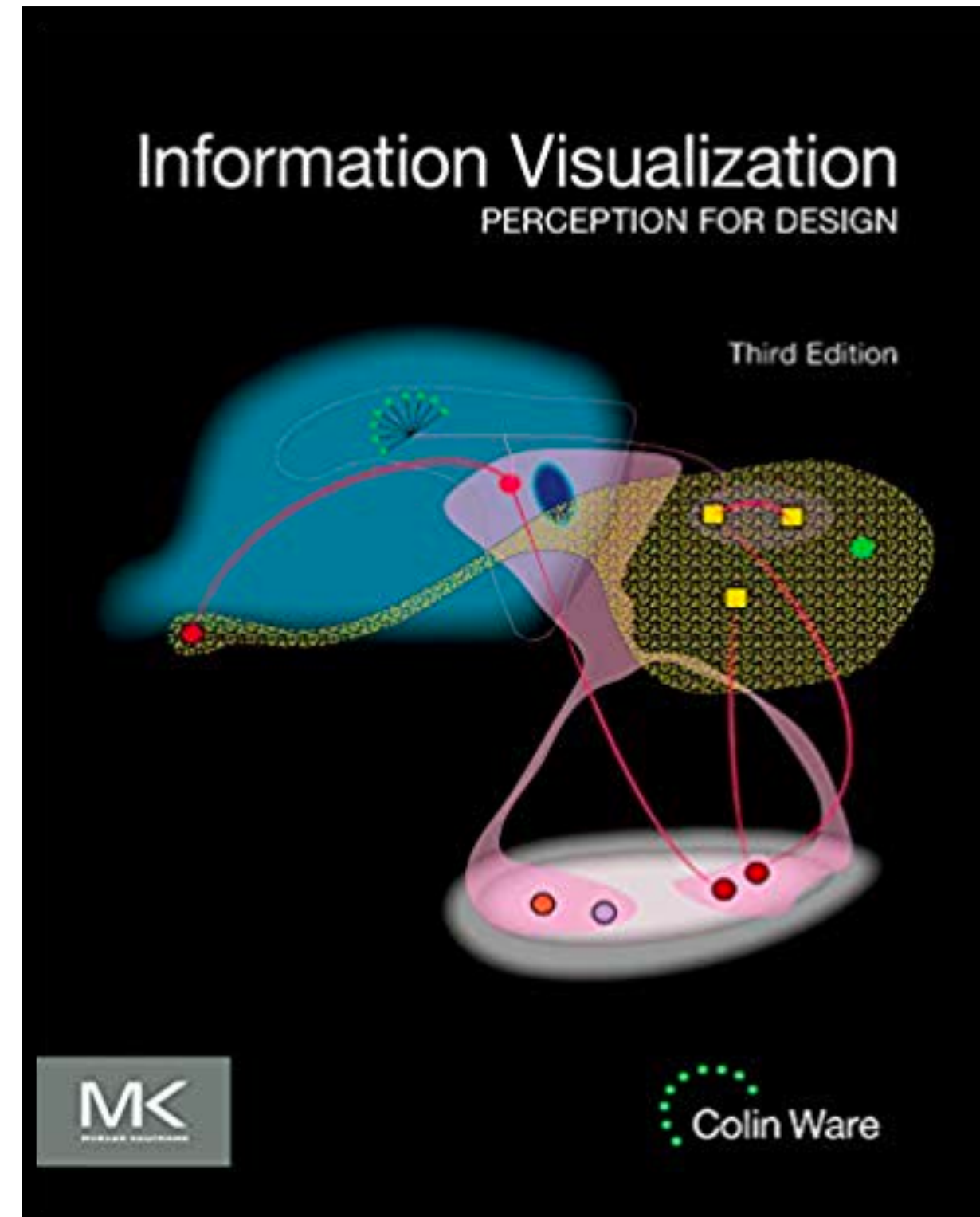
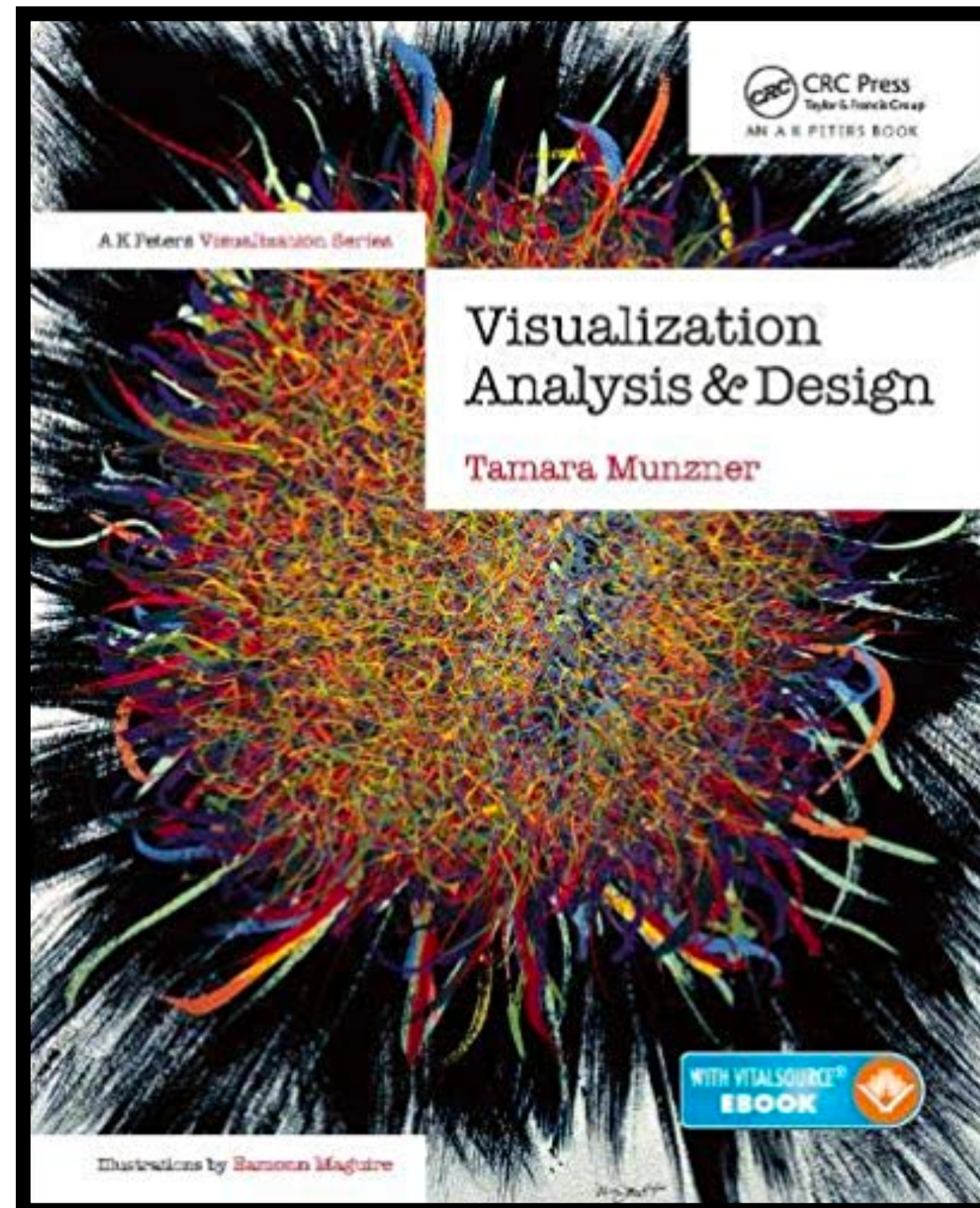
Popular science books about statistics and visualization



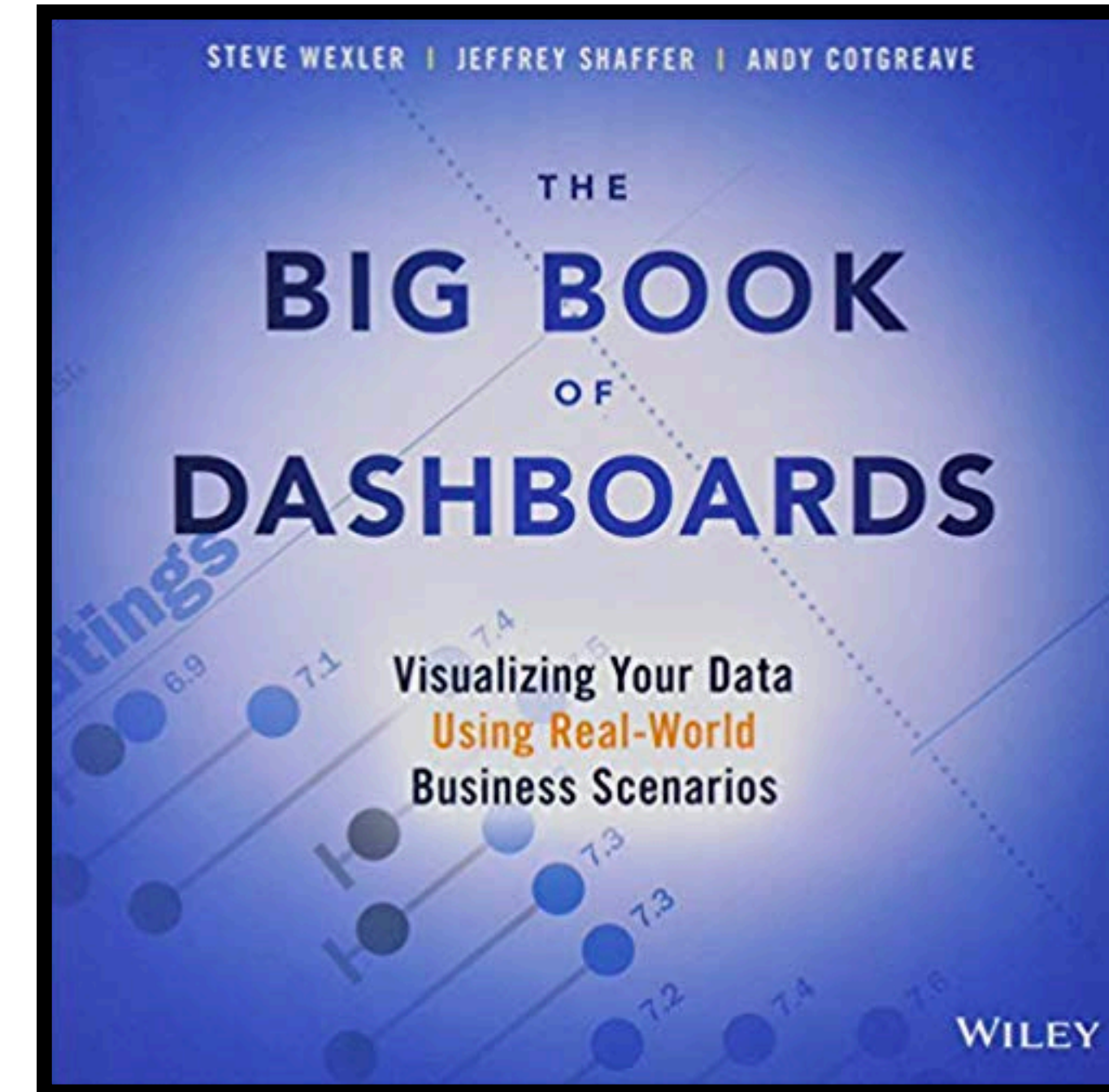
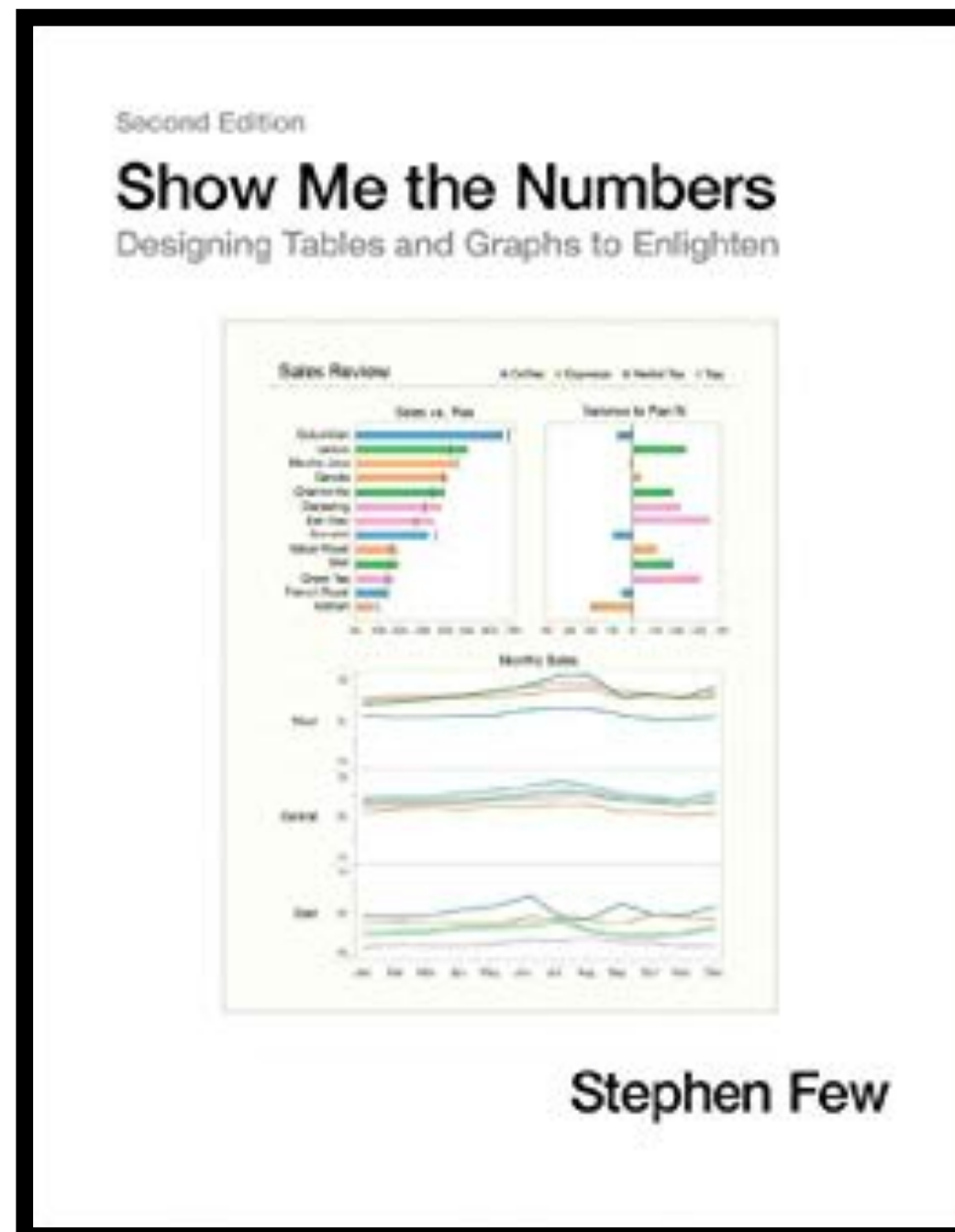
Fundamentals of visualization for communication



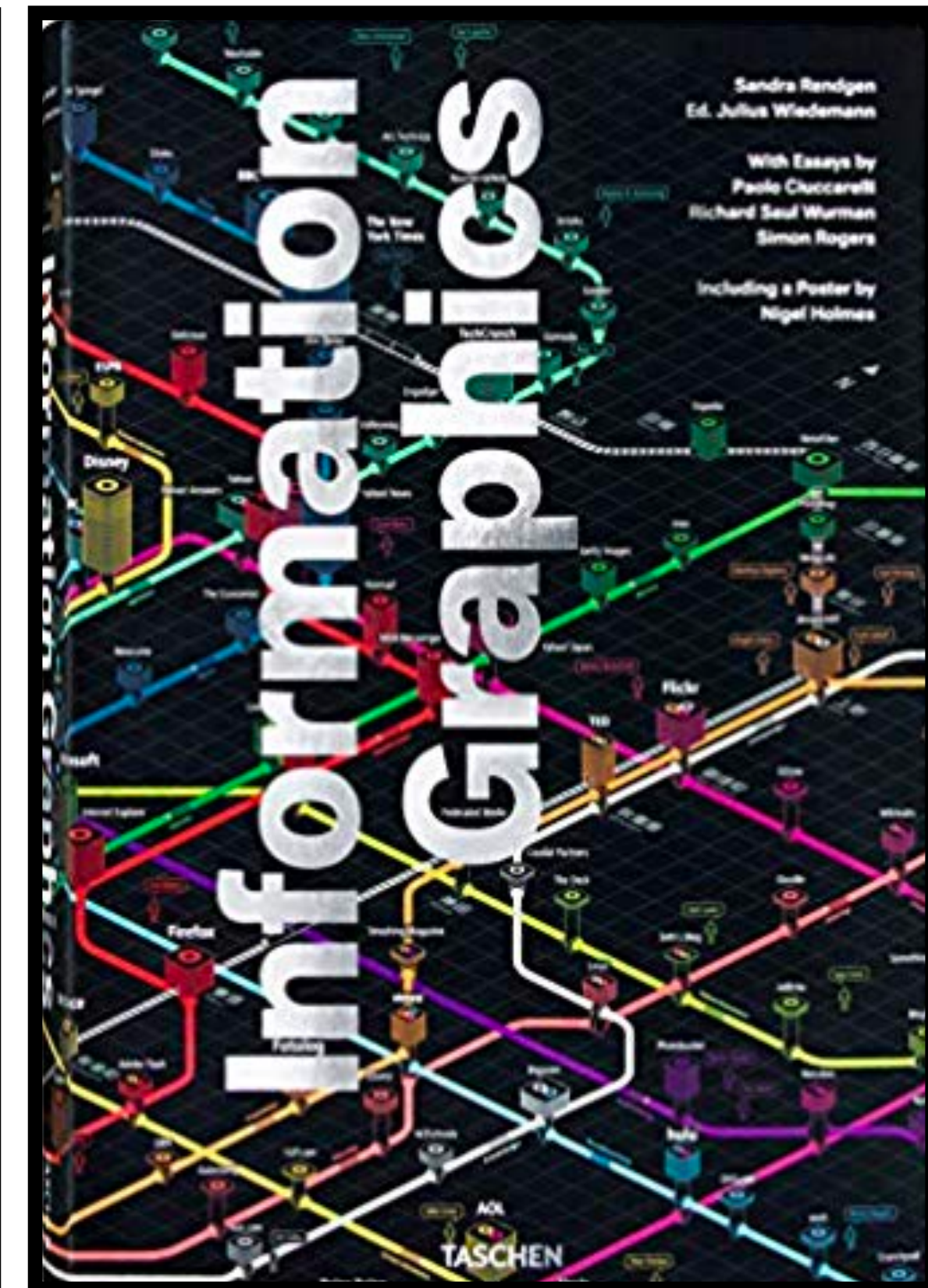
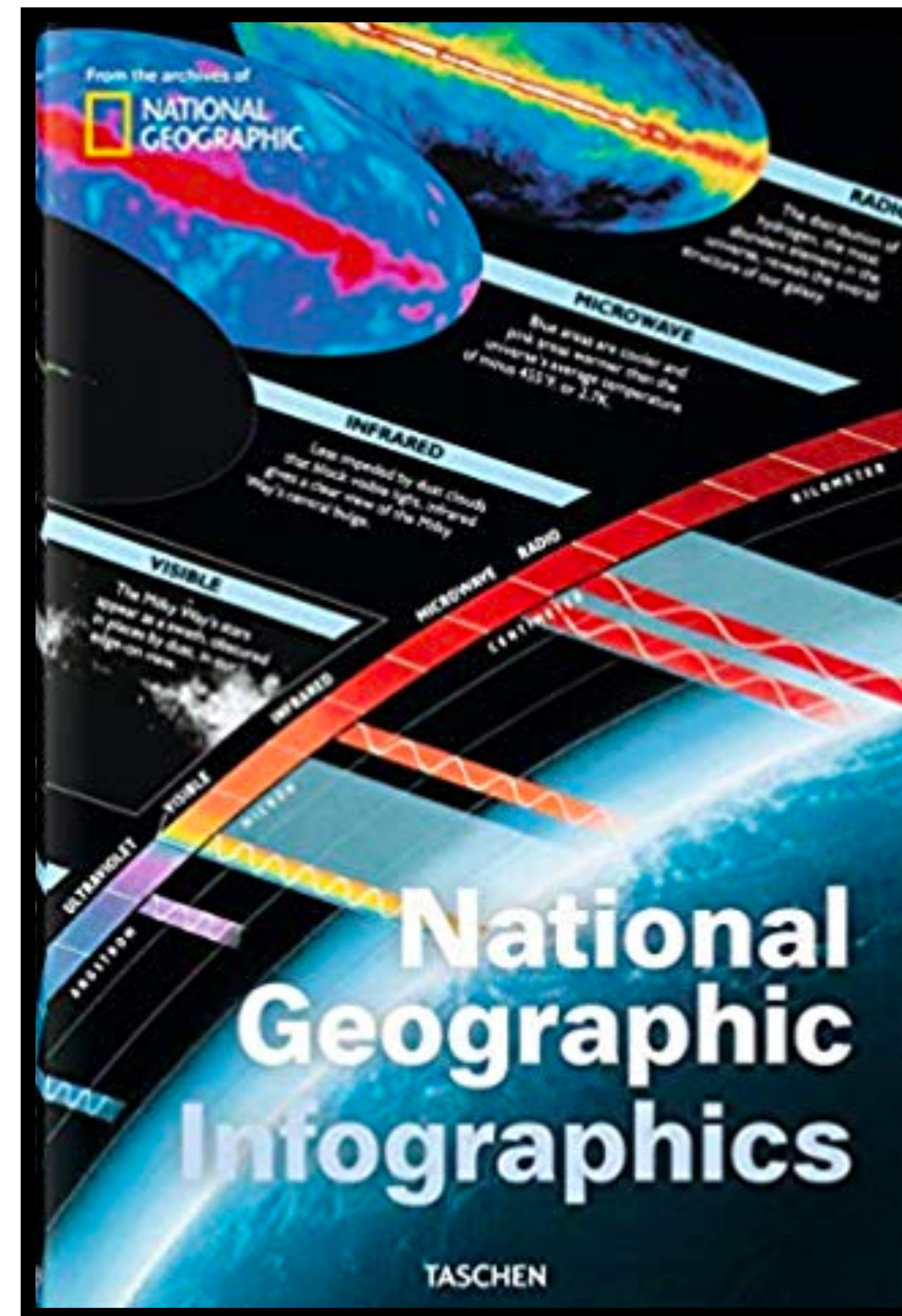
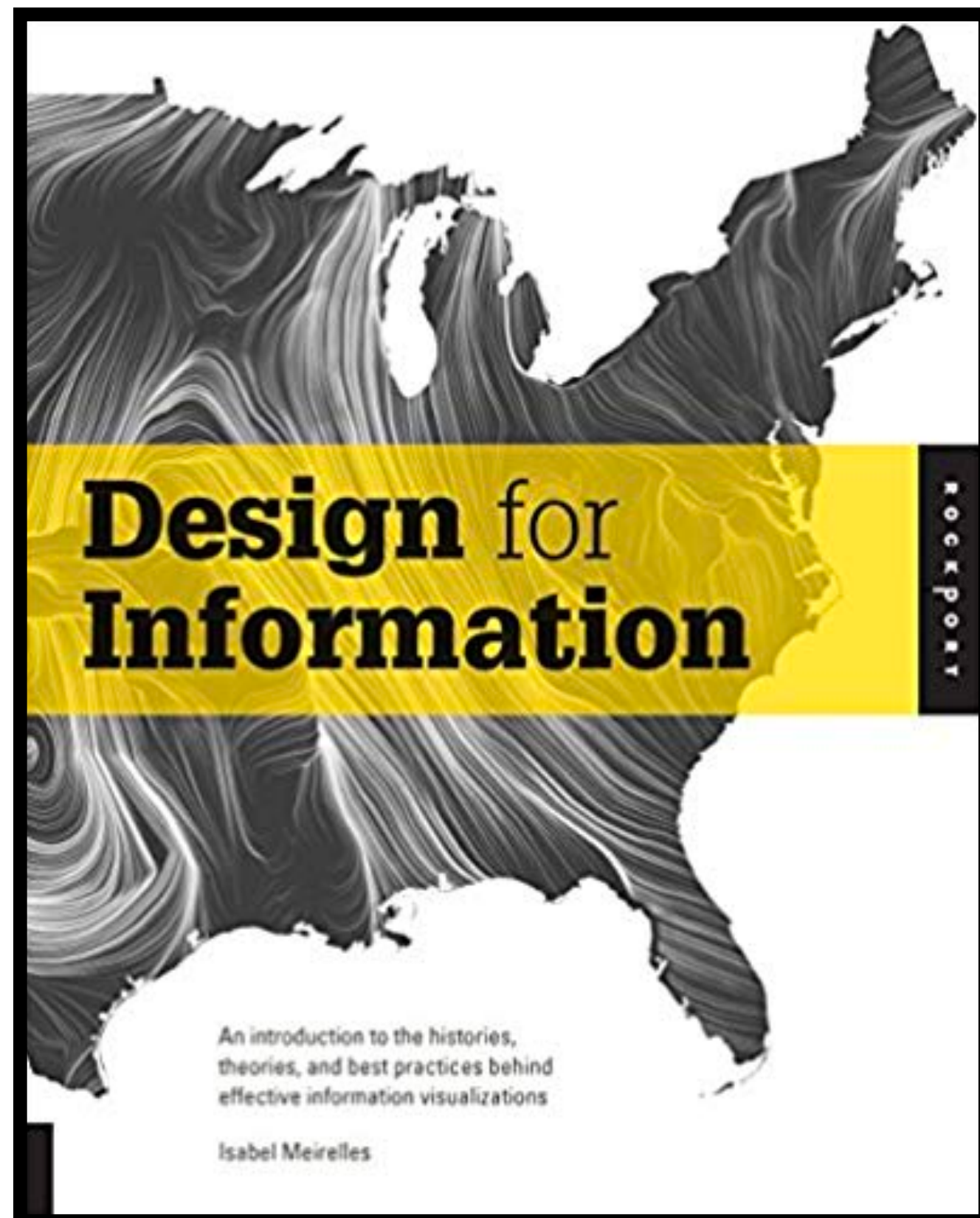
Exploratory and scientific visualization



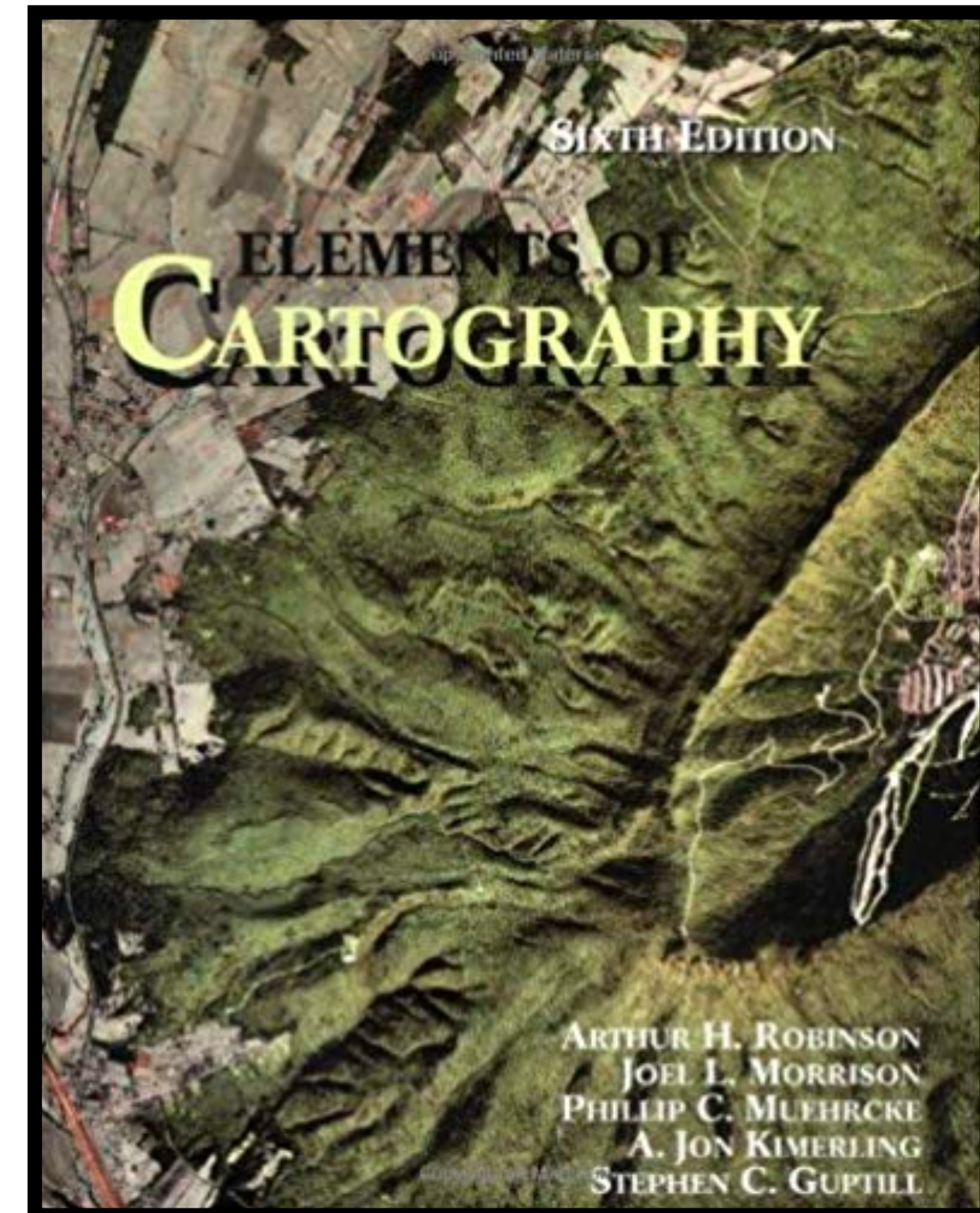
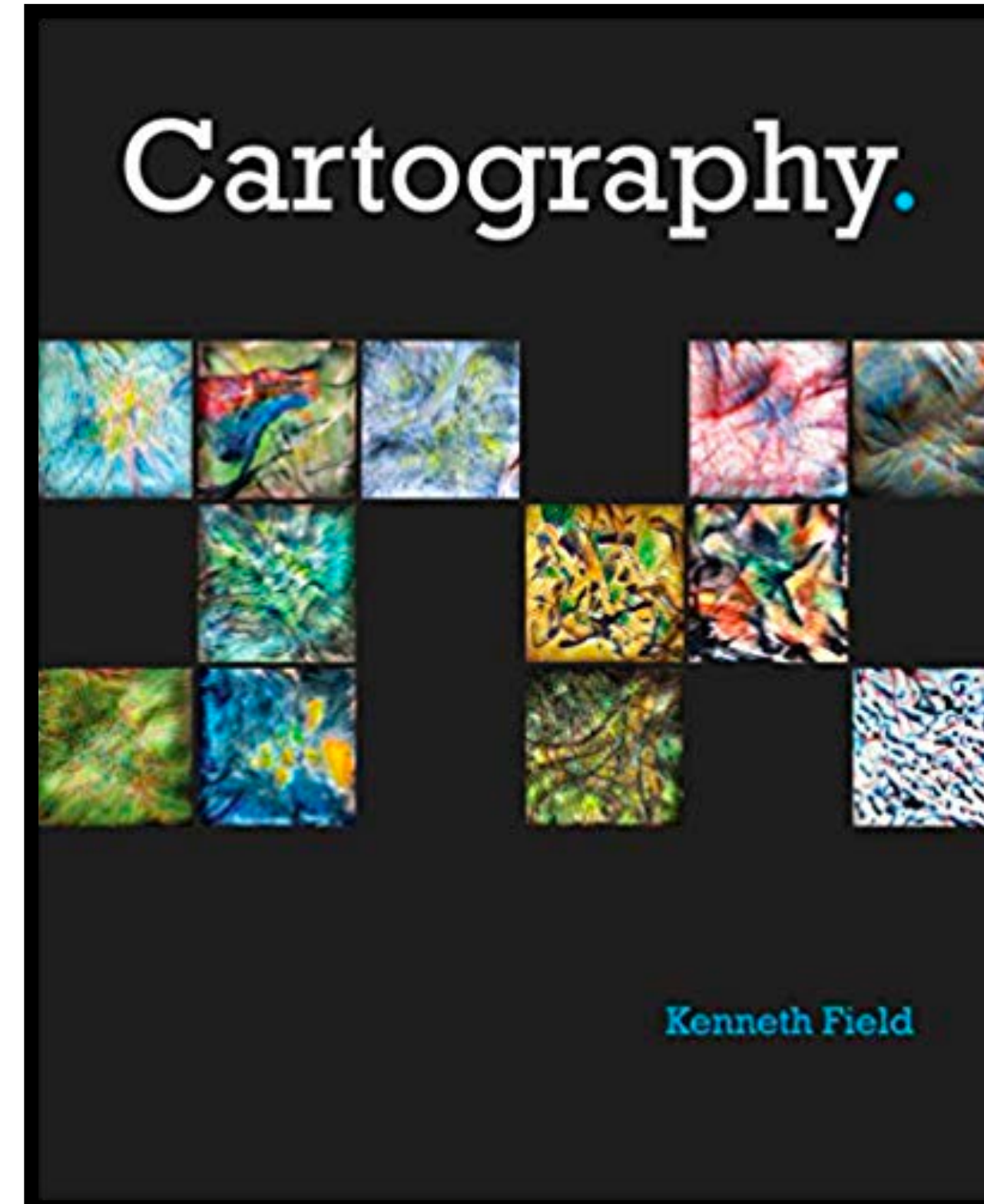
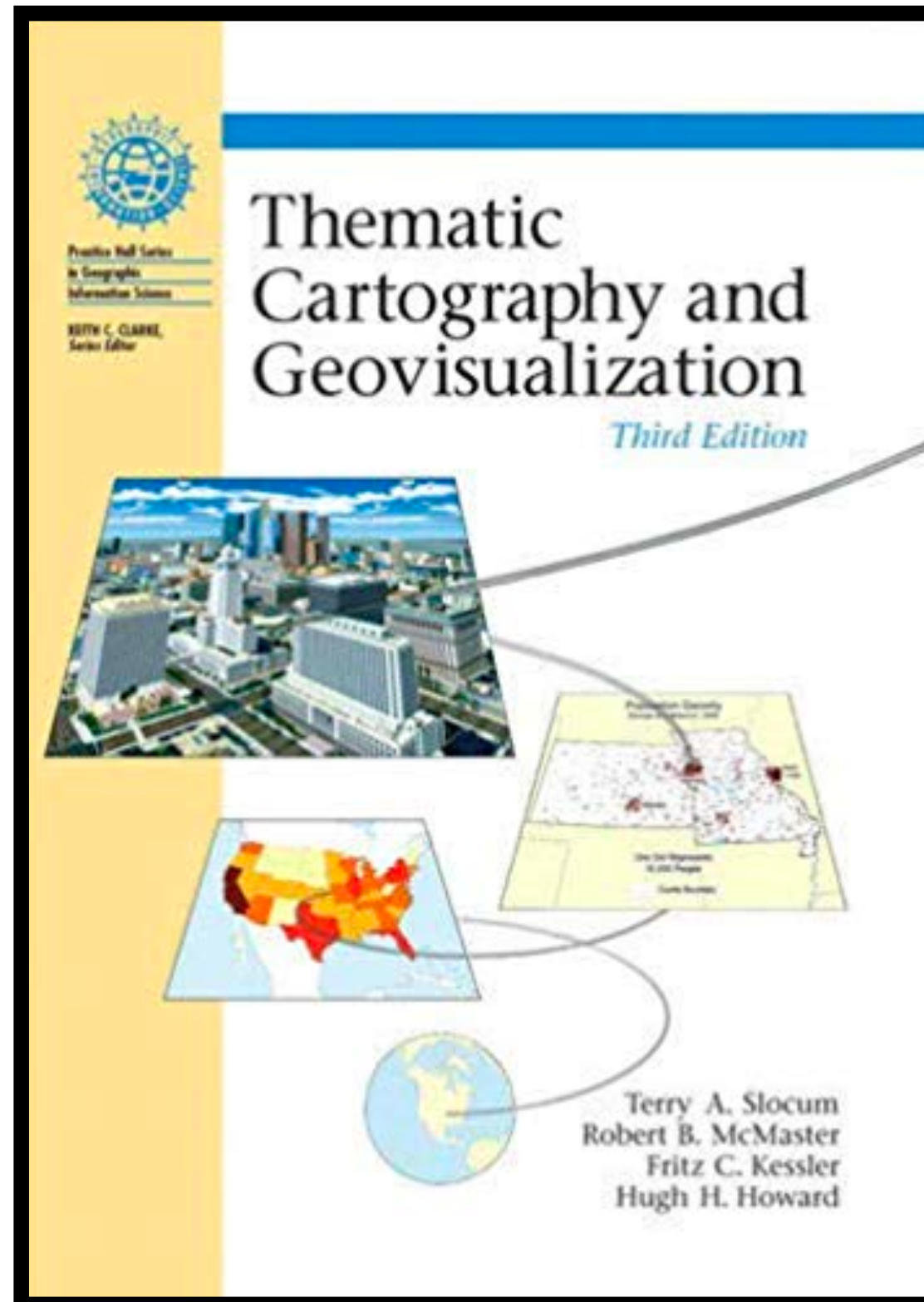
Business analytics and dashboard design



Inspirational books



Cartography





The End.

www.thefunctionalart.com , www.albertocairo.com , alberto.cairo@gmail.com