

Combining information
visualization theory and the
grammar of graphics to do and
teach modern data analysis

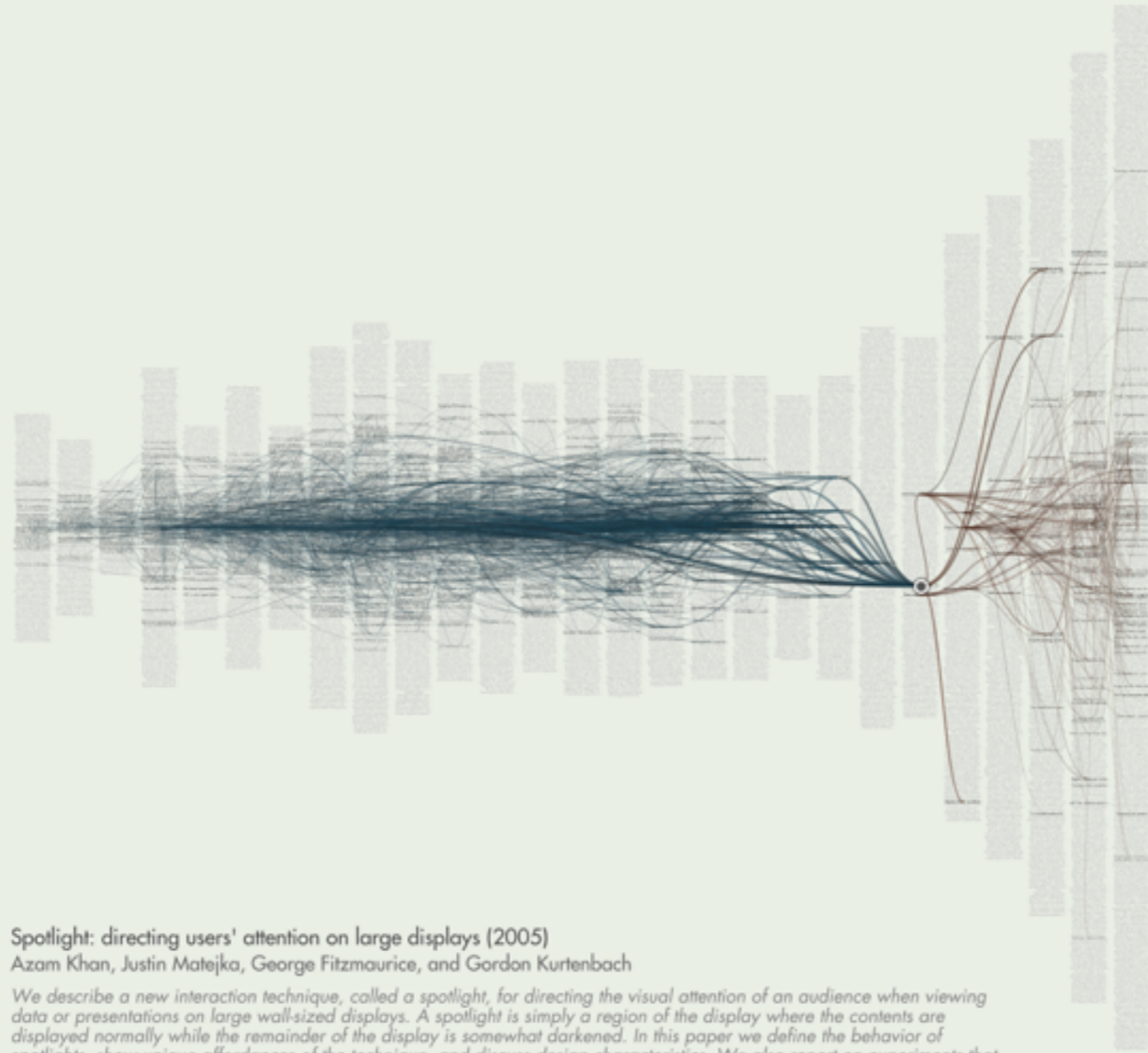
6th February 2018

Roger Beecham
www.roger-beecham.com

CITEOLOGY

3,502 CHI/UIST PAPERS AND THE 11,699 CITATIONS BETWEEN THEM

1982 1983 1985 1986 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 **2005** 2006 2007 2008 2009 2010



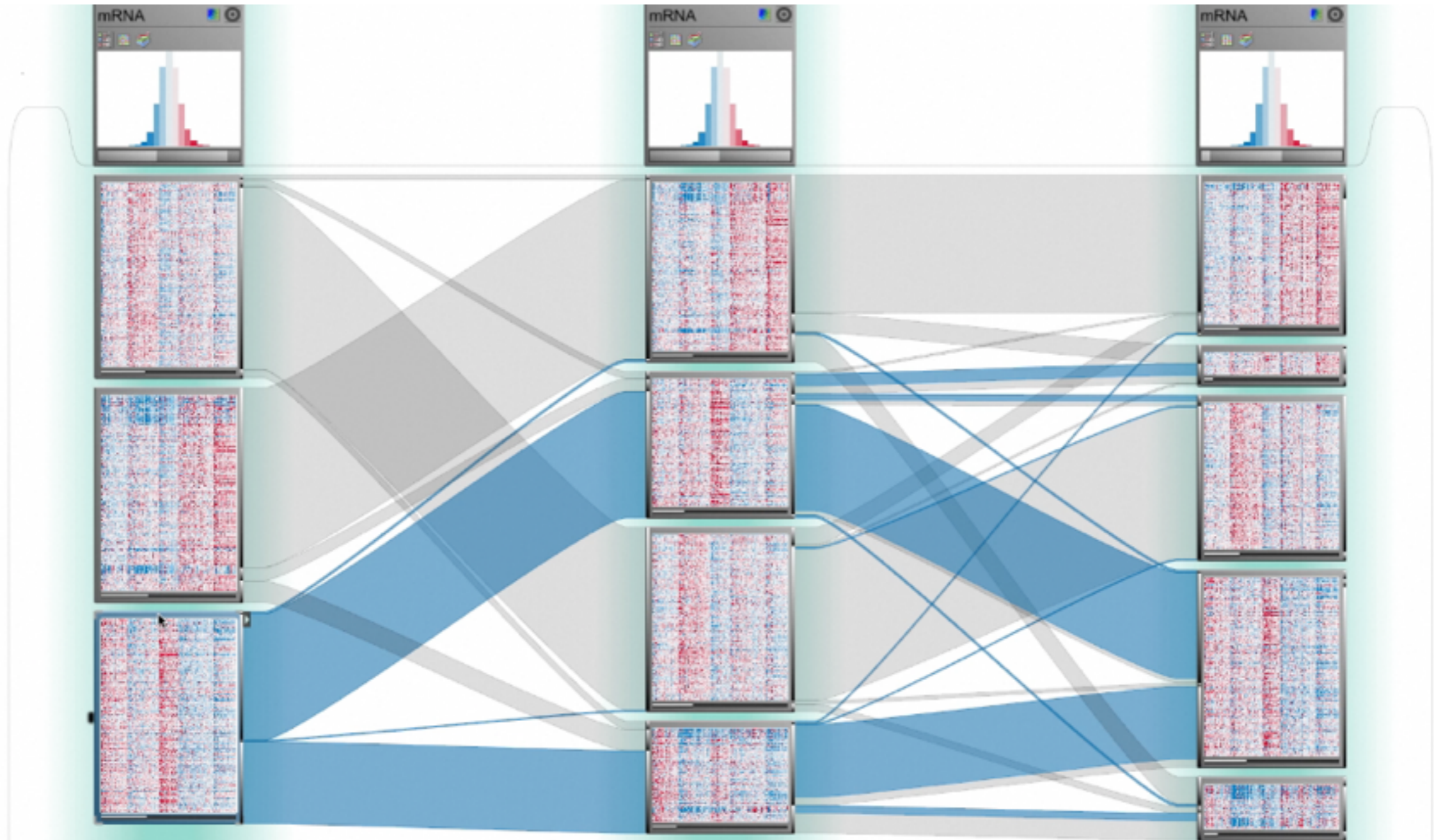
Spotlight: directing users' attention on large displays (2005)

Azam Khan, Justin Matejka, George Fitzmaurice, and Gordon Kurtenbach

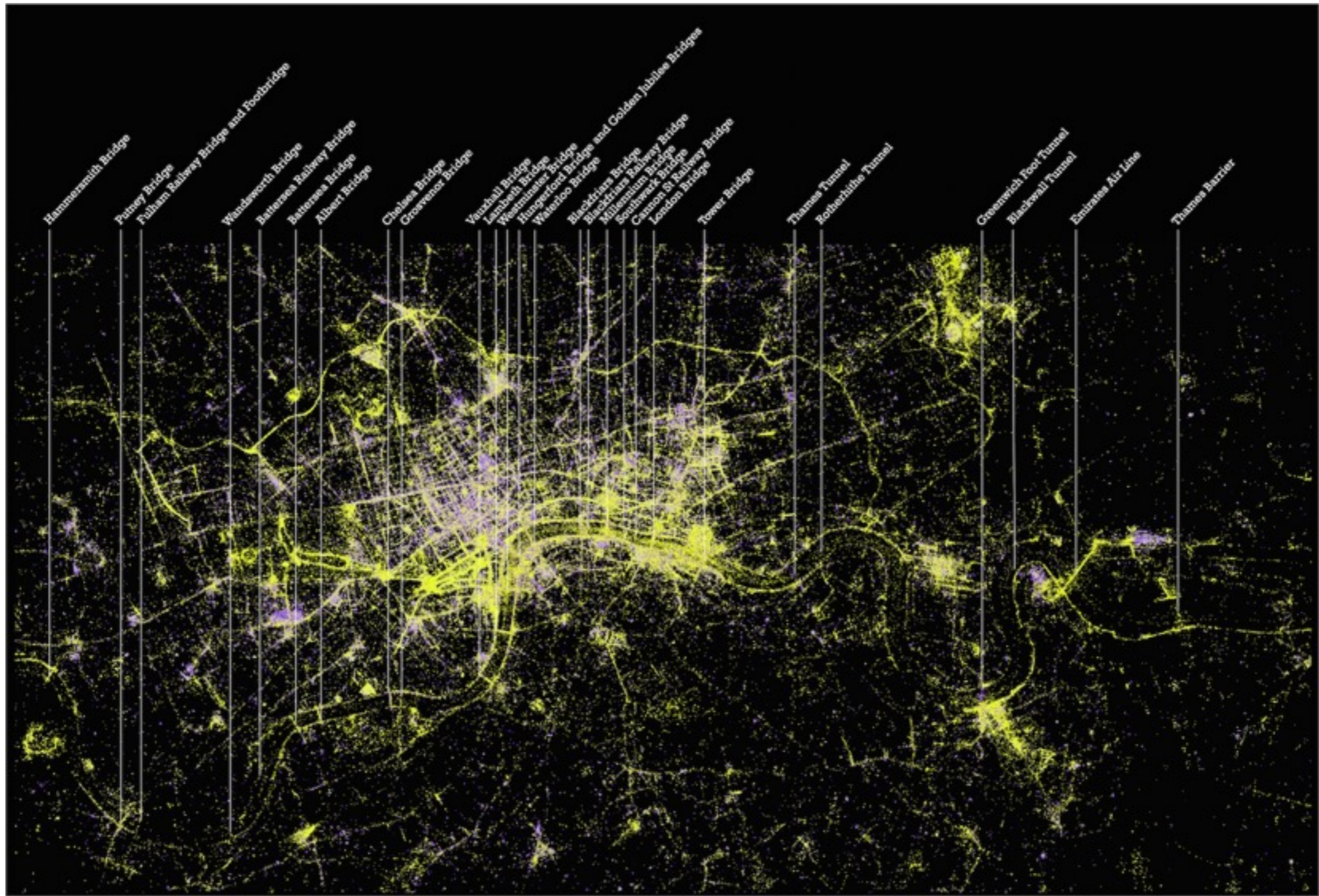
We describe a new interaction technique, called a spotlight, for directing the visual attention of an audience when viewing data or presentations on large wall-sized displays. A spotlight is simply a region of the display where the contents are displayed normally while the remainder of the display is somewhat darkened. In this paper we define the behavior of spotlights, show unique affordances of the technique, and discuss design characteristics. We also report on experiments that show the benefit of using the spotlight a large display and standard desktop configuration. Our results suggest that the...

www.autodeskresearch.com/projects/citeology

Justin Matejka, Tovi Grossman, George Mitzmaurice



Marc Streit, Alexander Lex, Samuel Gratzl, Hanspeter Pfister, Nils Gehlenbourg



Alex Kachkaev, Jo Wood

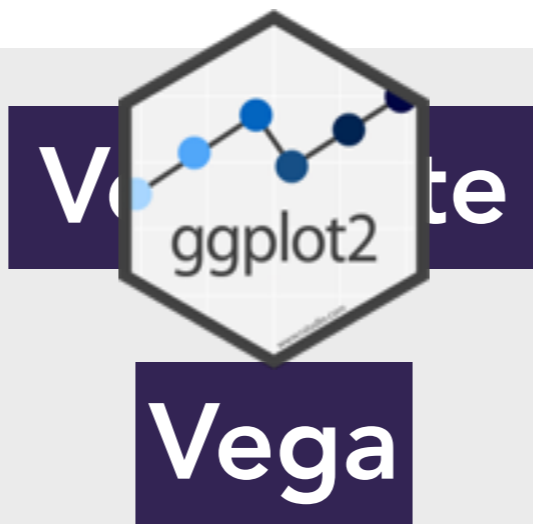




Vega-Lite

Vega





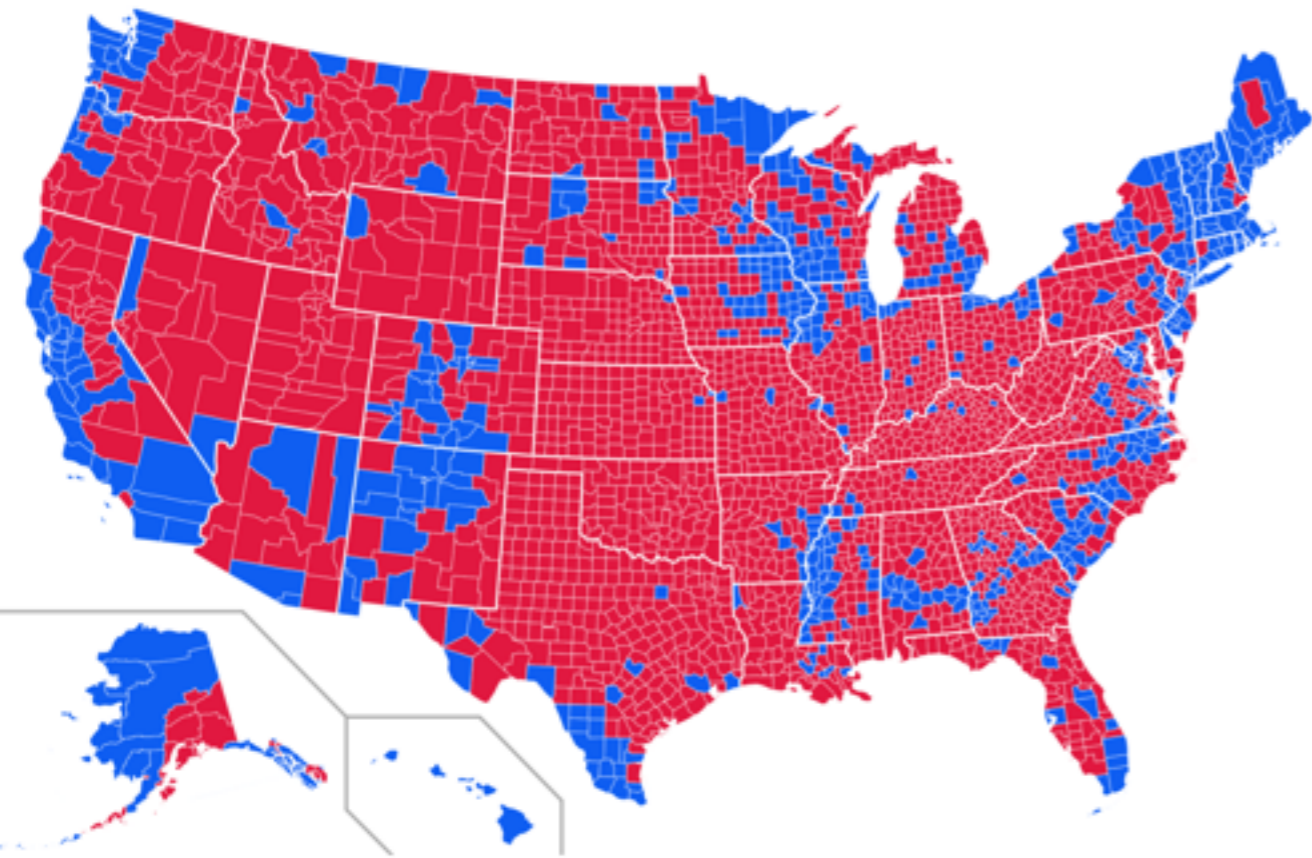


Data graphics visually display measured quantities by means of the combined use of points, lines, a coordinate system, numbers, symbols, words, shading, and color.

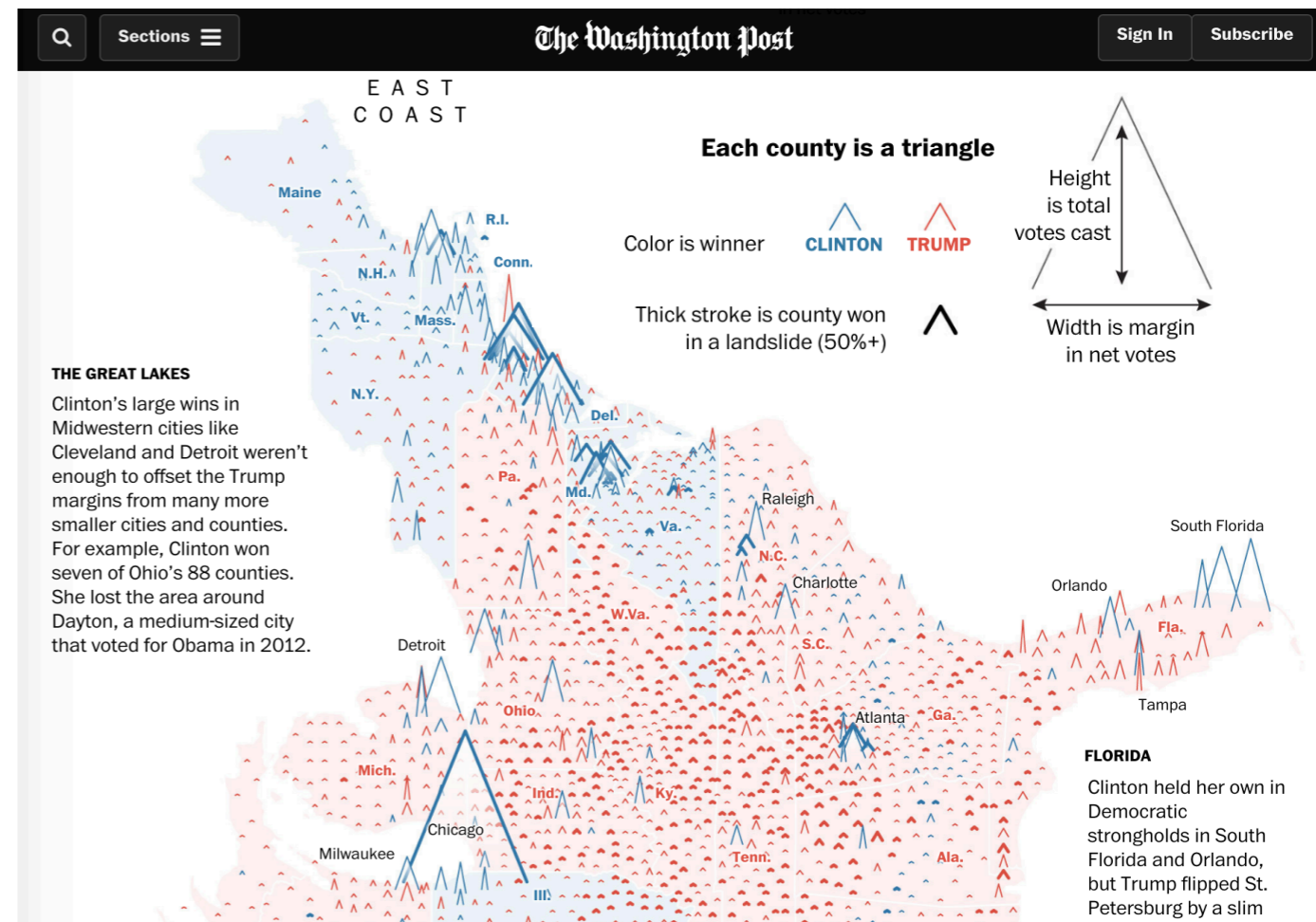
Tufte, 1983

Effective data graphics should

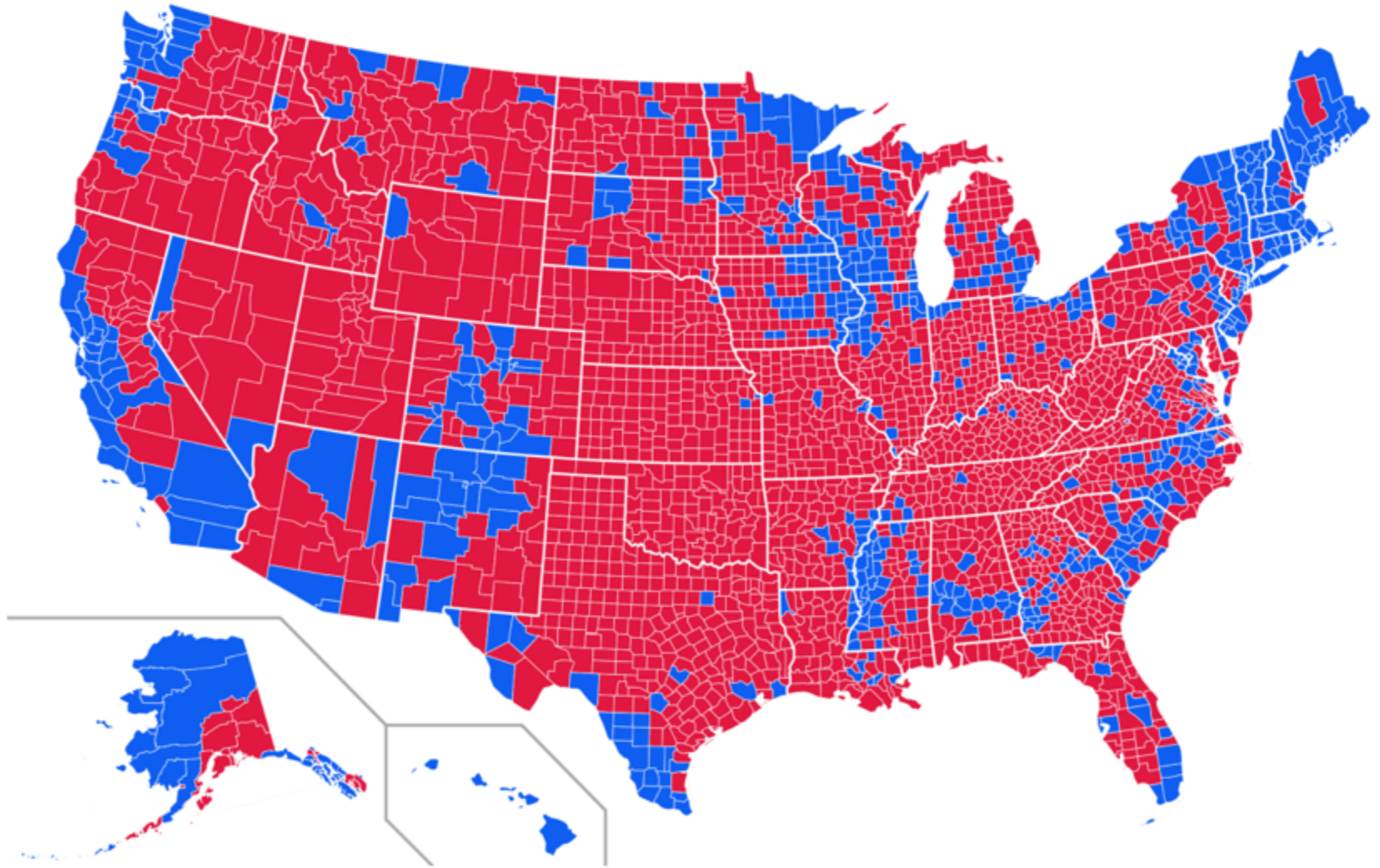
1. Show the data
2. Induce the viewer to think about the substance of the data rather than about graphic design
3. Avoid distorting what the data have to say
4. Present many numbers in a small space
5. Make large data sets coherent
6. Encourage the eye to compare different pieces of data
7. Reveal the data at several levels of detail from a broad overview to a fine structure



Natalie Schmidt, on Medium



Lazaro Gamio and Dan Keating, Washington Post



Natalie Schmidt, on Medium



Hue



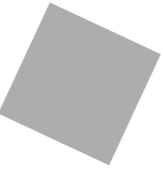
Saturation



Brightness



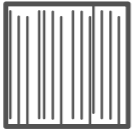
Shape



Orientation



Arrangement



Texture



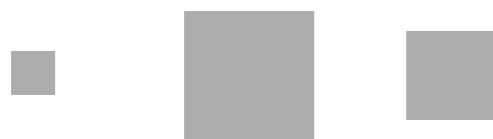
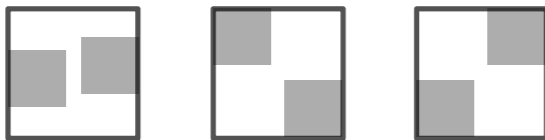
Size



Focus



Location

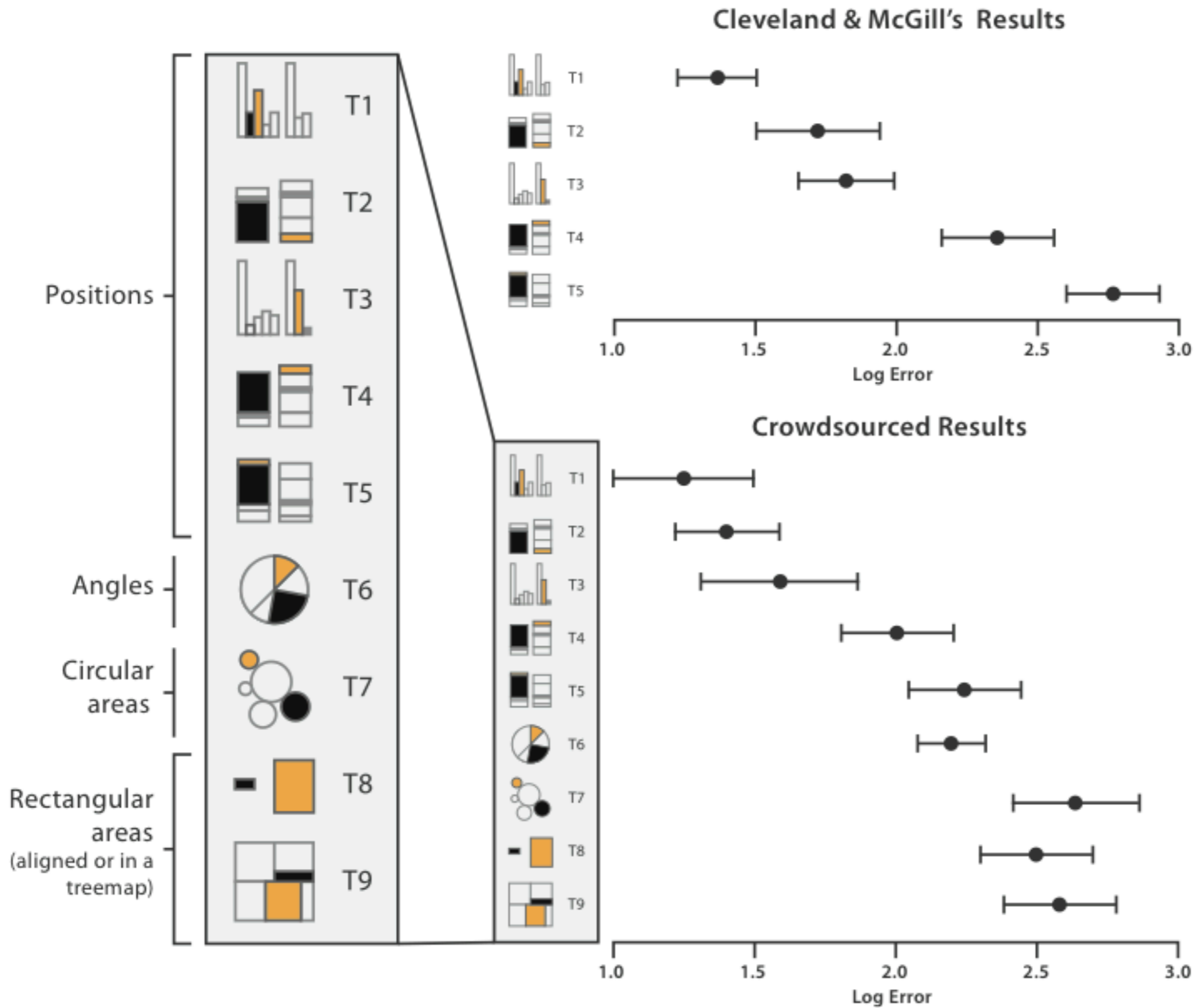


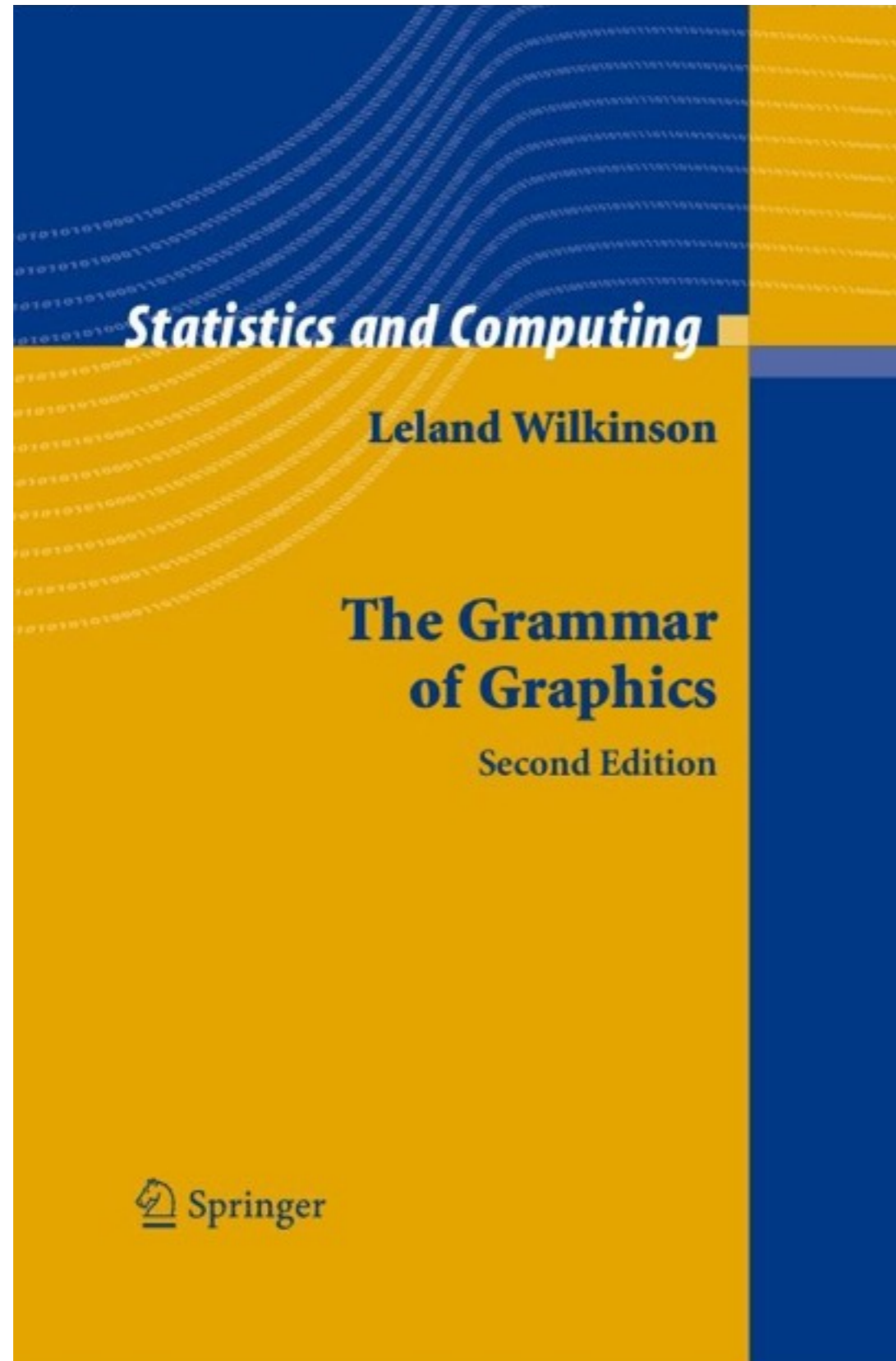
Selective: Change in this visual variable alone is enough to allow a symbol to be selected from a group.

Associative: Symbols that are alike in all other ways can be grouped according to change in this visual variable.

Quantitative: A numerical reading is obtainable from changes in this visual variable.

Order: Changes in this variable perceived as ordered





Data

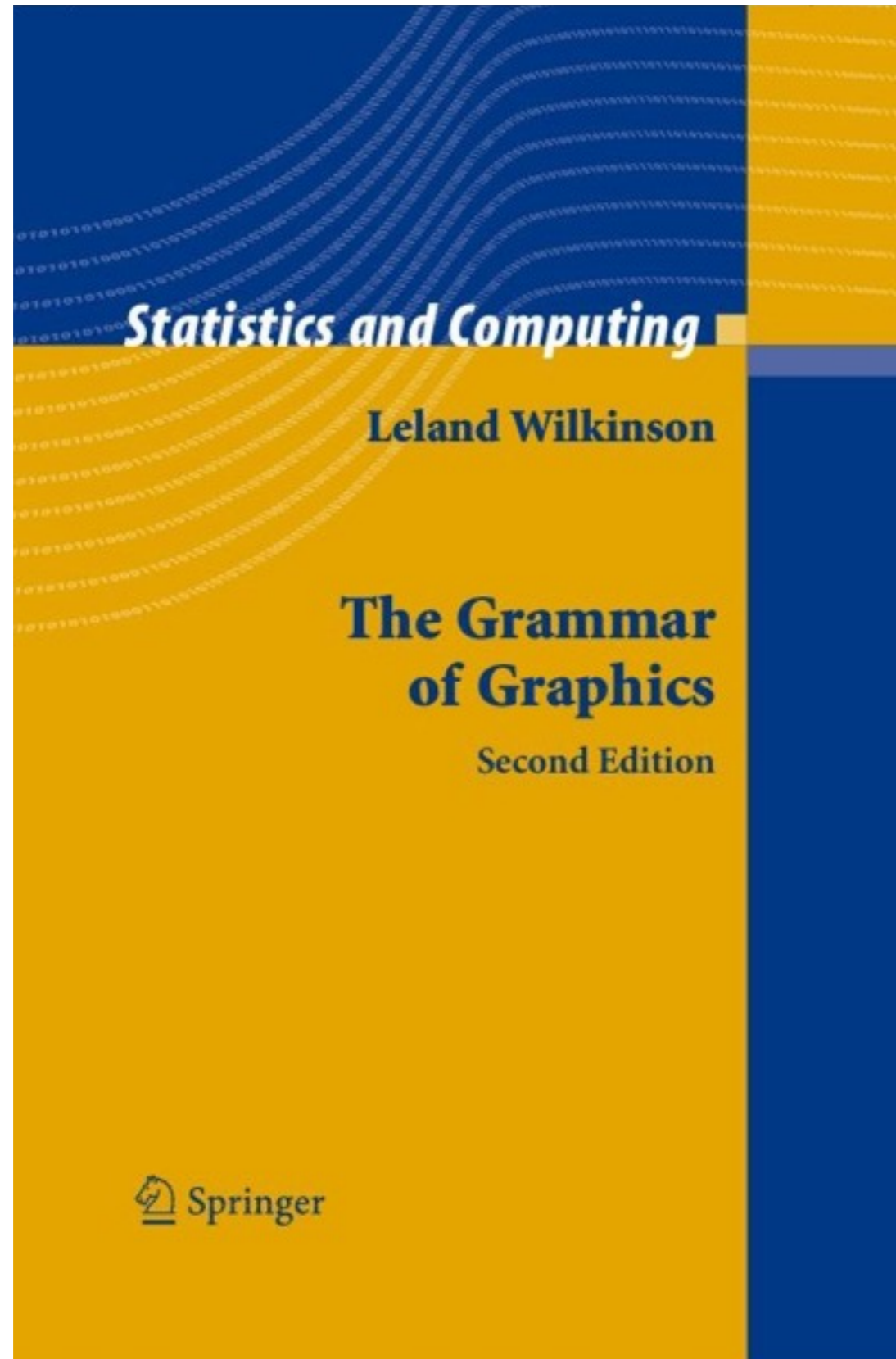
Transformation

Element

Scale

Guide

Coord



Vega-Lite





Data	variables you want to represent
Aesthetics	mapping of data to visual channels
Geom	shapes to represent data (point, line, bar)
Facets	split on a (nominal/ordinal) variable to generate small multiples
Statistics	aggregates using statistical models
Coordinates	plotting space you are using
Themes	non-data ink: design with a particular visual fonts, colours and other design elements.

+ informed defaults



I am a...

Find information by subject

Our work

Political parties, campaigning & donations

Elections & referendums

Promoting voter registration

Performance standards

Electoral fraud

2011 Census

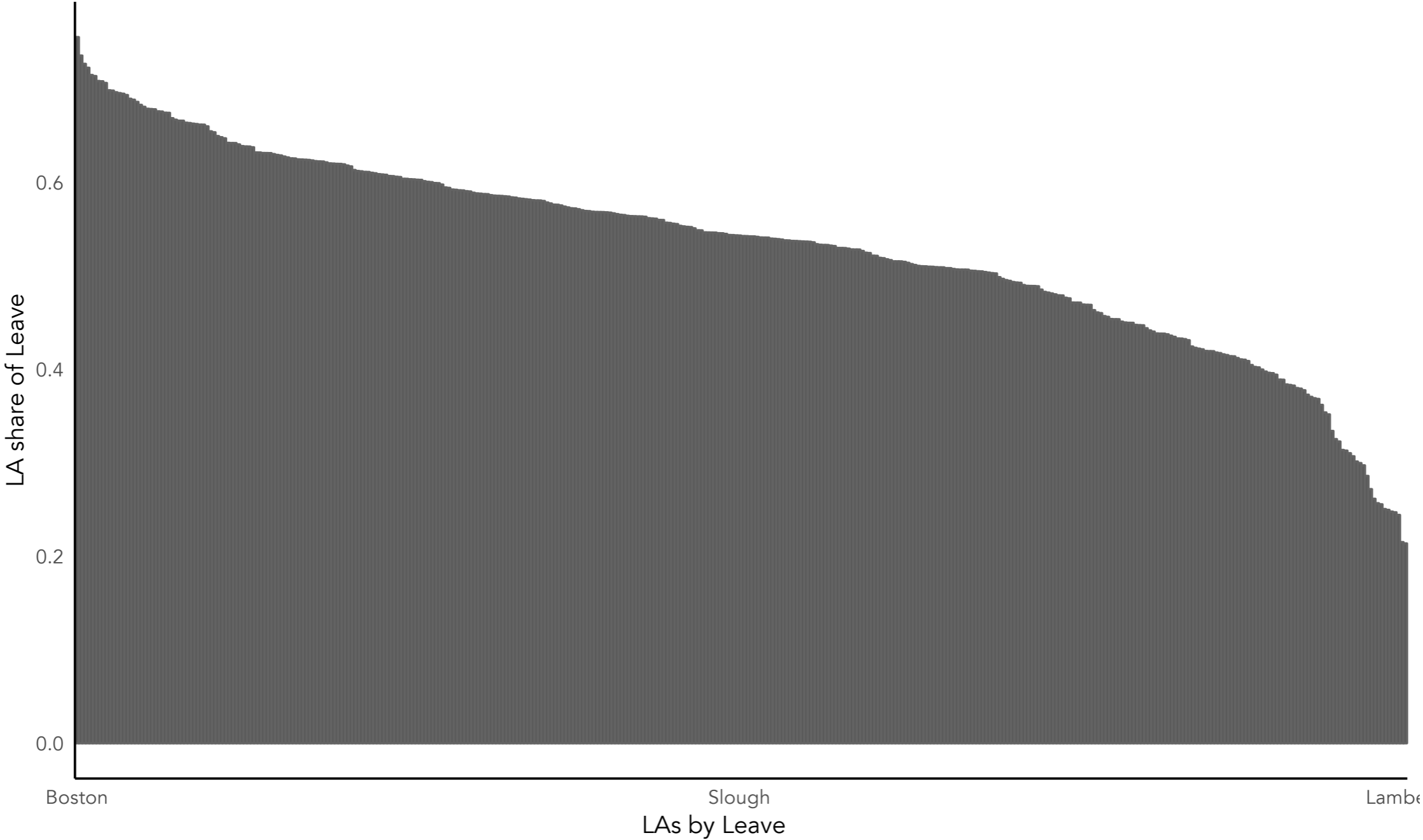
Census statistics help paint a picture of the nation and how we live. They provide a detailed snapshot of the population and its characteristics, and underpin funding allocation to provide public services. The population of England & Wales on Census Day, 27 March 2011, was 56,075,912.

Brexit data: share of leave vote by Local Authority

Demographics data: skills levels, occupation and diversity by Local Authority

LAs ordered by share of
Leave

LAs ordered by share of Leave



LAs ordered by share of Leave

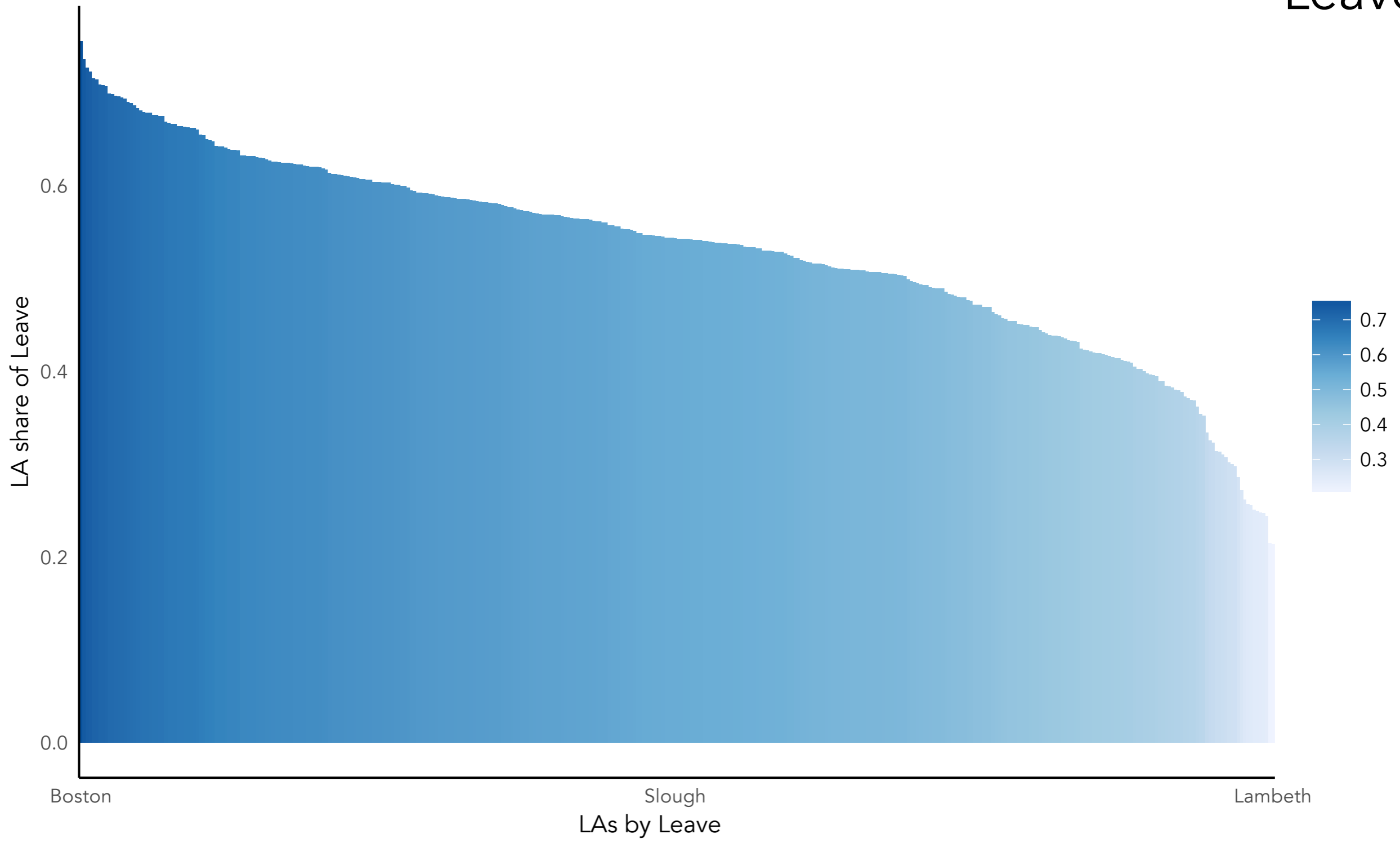


Summary of mappings - specification

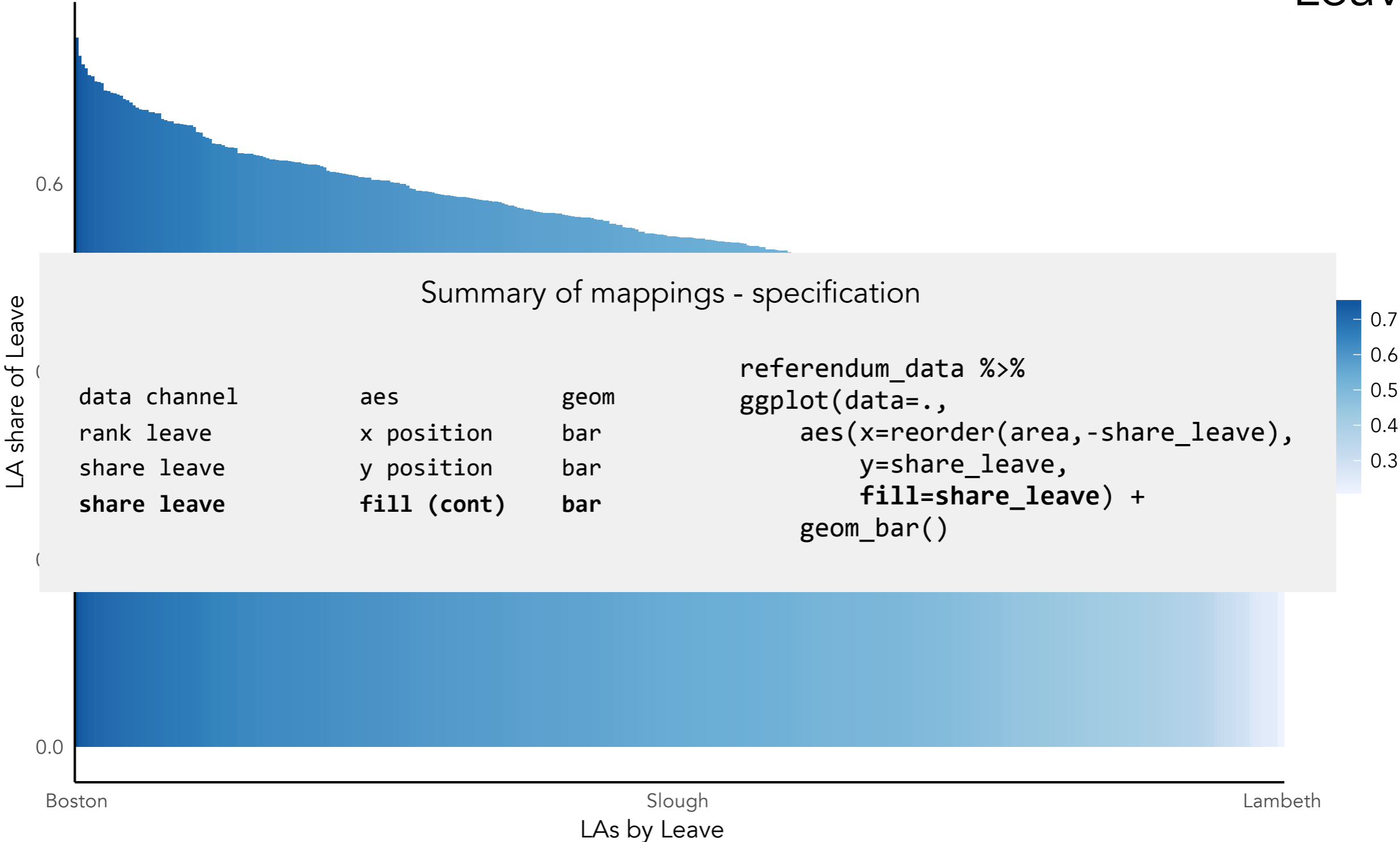
data channel	aes	geom	referendum_data %>% ggplot(data=.,
rank leave	x position	bar	aes(x=reorder(area, -share_leave),
share leave	y position	bar	y=share_leave) + geom_bar()



LAs ordered by share of Leave



LAs ordered by share of Leave

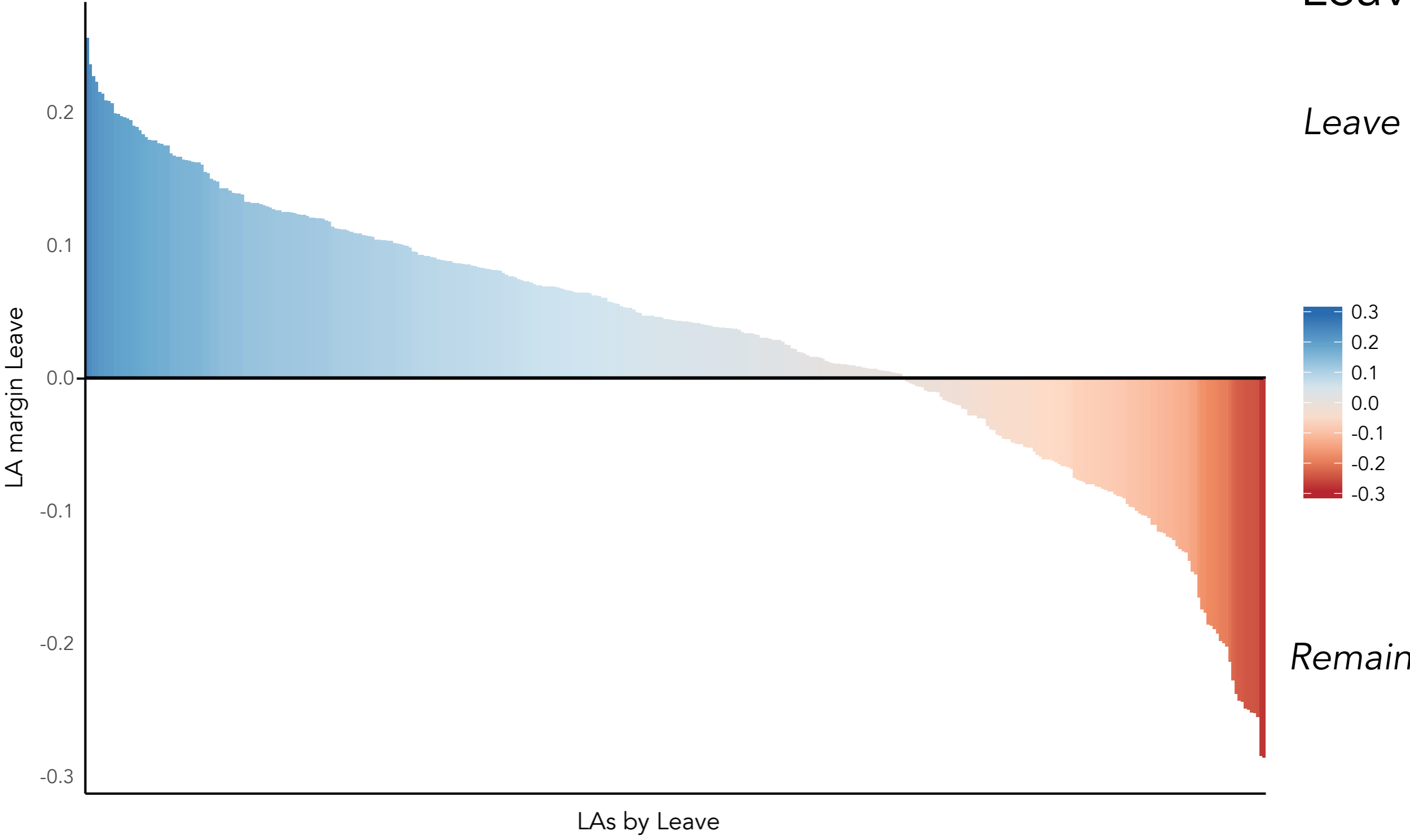


LAs ordered by share of
Leave

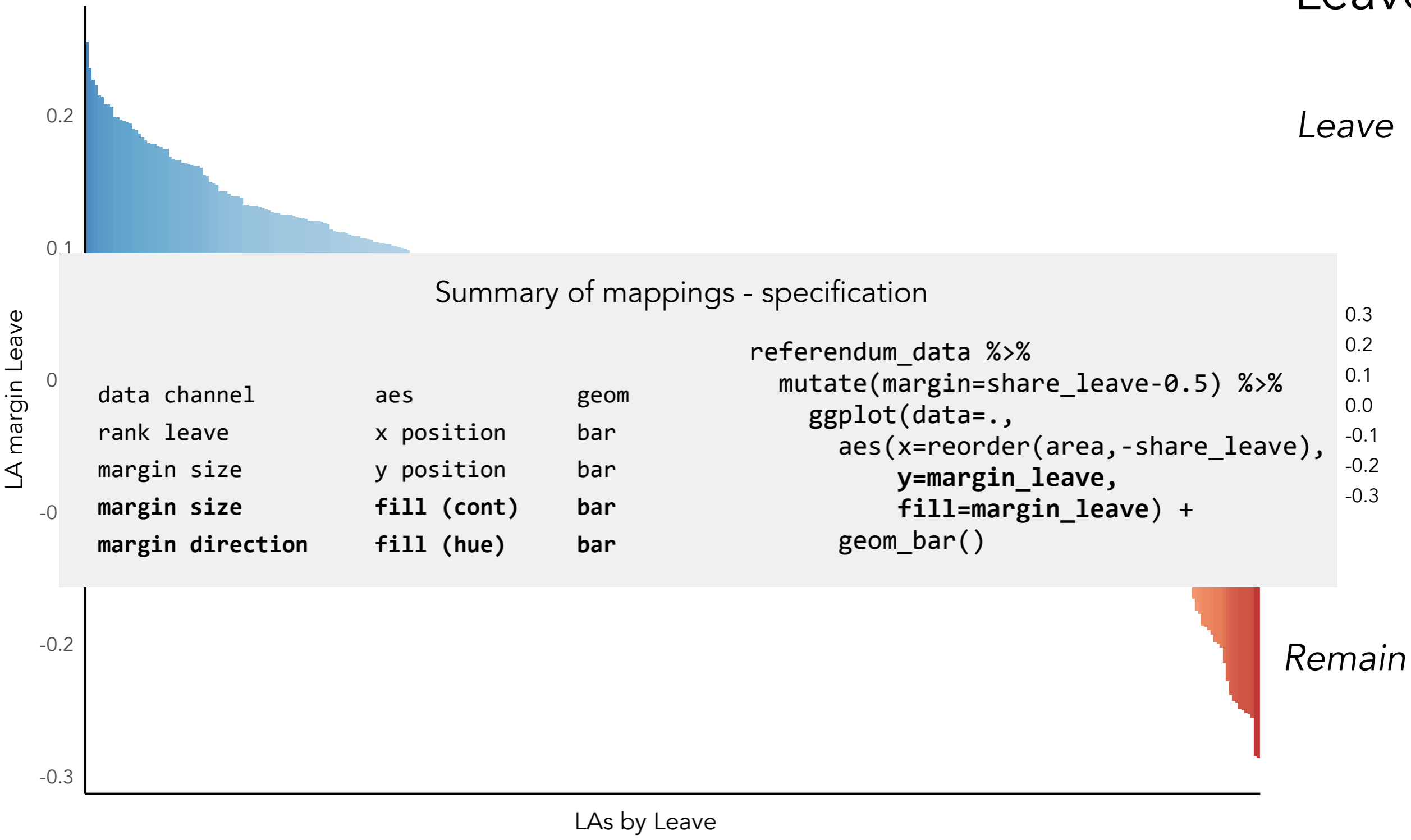
Leave

Remain

LAs ordered by share of Leave



LAs ordered by share of Leave



Leave

Summary of mappings - specification

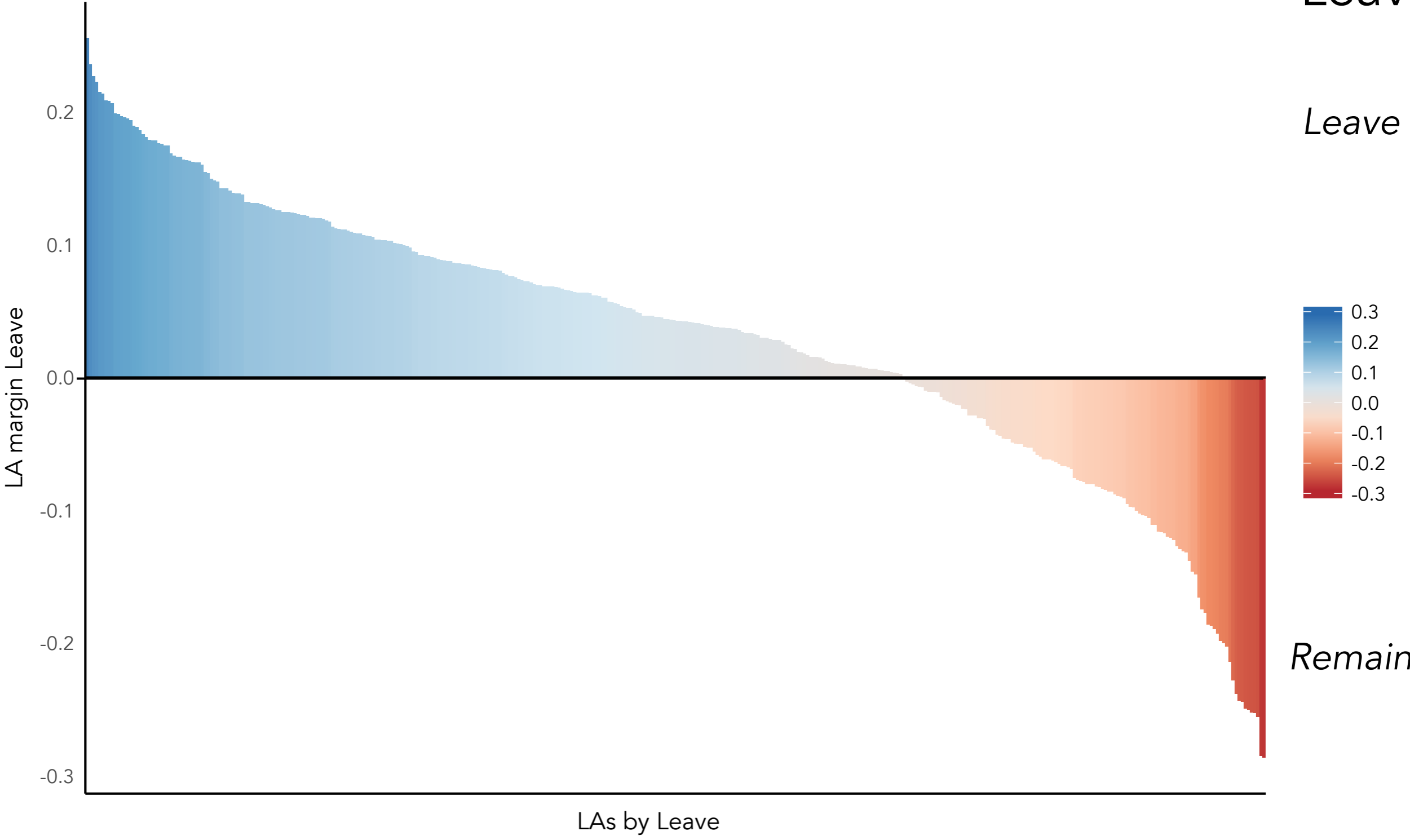
data channel	aes	geom
rank leave	x position	bar
margin size	y position	bar
margin size	fill (cont)	bar
margin direction	fill (hue)	bar

```
referendum_data %>%
  mutate(margin=share_leave-0.5) %>%
  ggplot(data=.,
    aes(x=reorder(area, -share_leave),
      y=margin_leave,
      fill=margin_leave) +
    geom_bar()
```

0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3

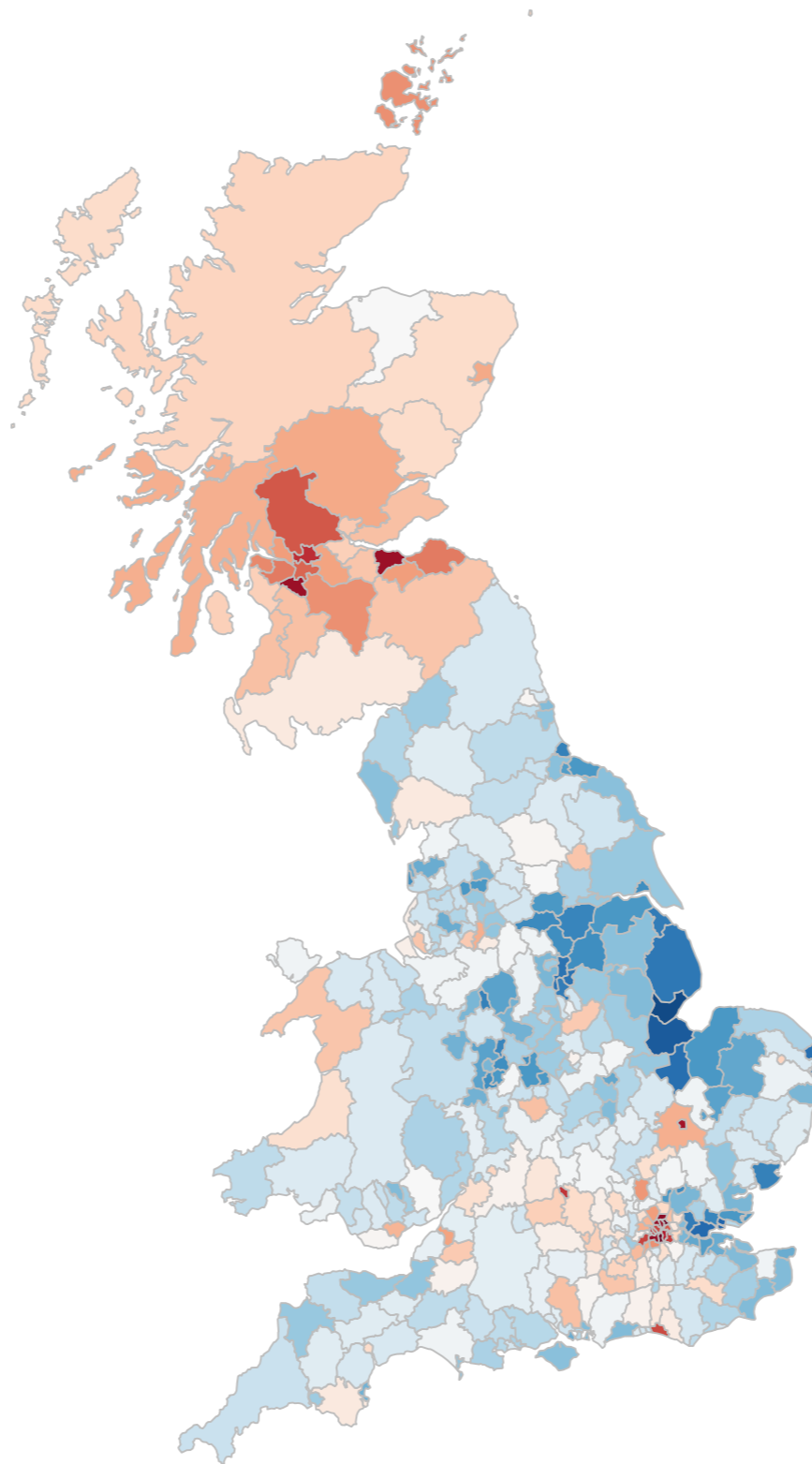
Remain

LAs ordered by share of Leave

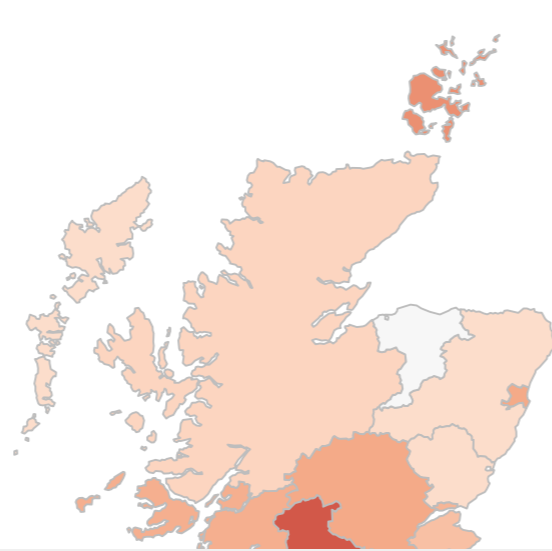


LAs ordered by geospatial
position

LAs ordered by geospatial
position

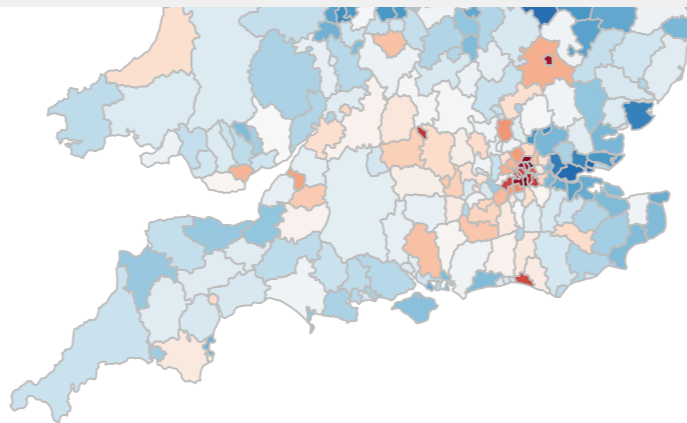


LAs ordered by geospatial position



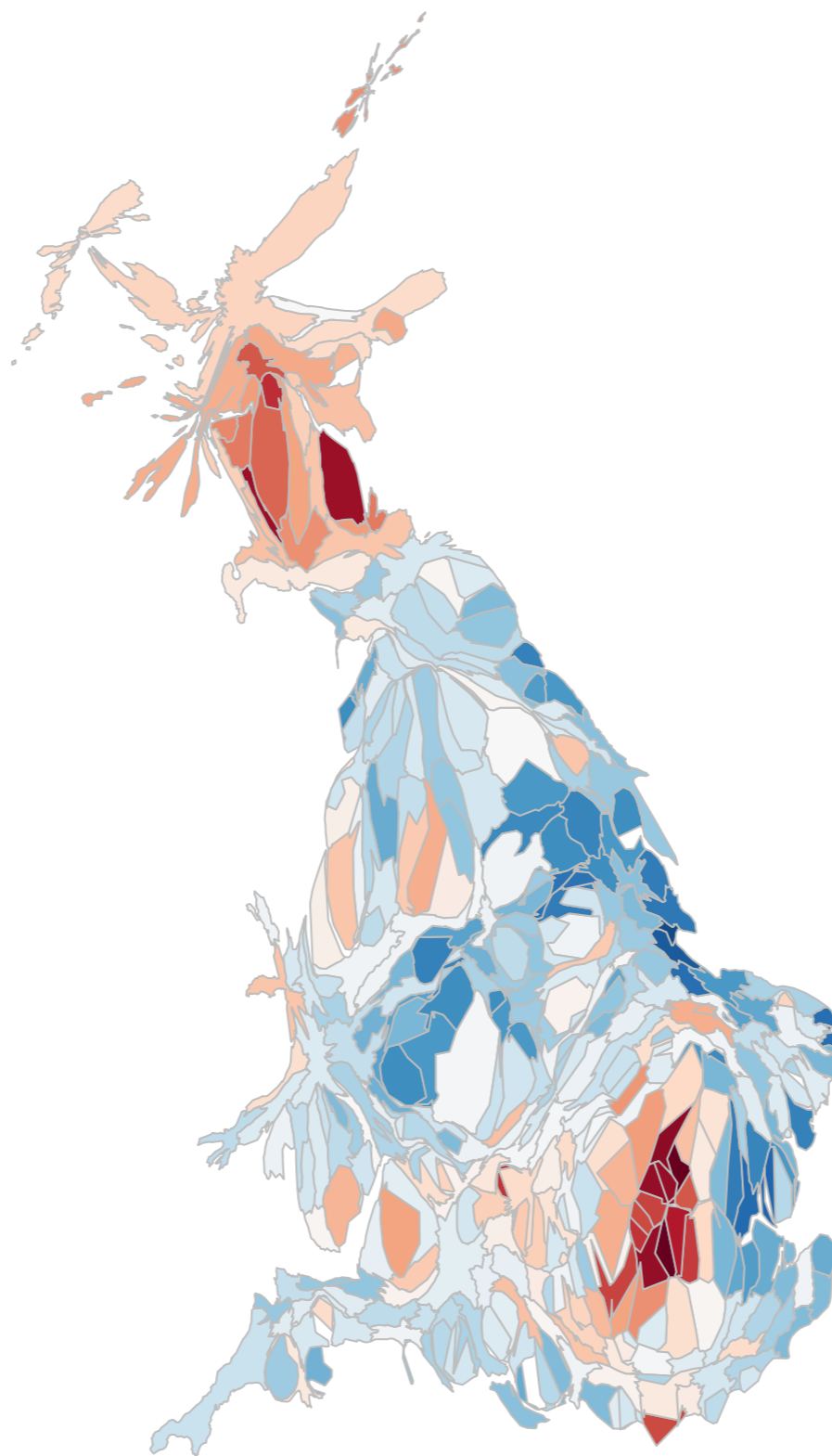
Summary of mappings - specification

data channel	aes	geom	referendum_data %>% ggplot(data=.,
la position	x,y position	poly	aes(x=easting,
margin size	fill (cont)	poly	y=northing,
margin direction	fill (hue)	poly	fill=share_leave) +
			geom_polygon()

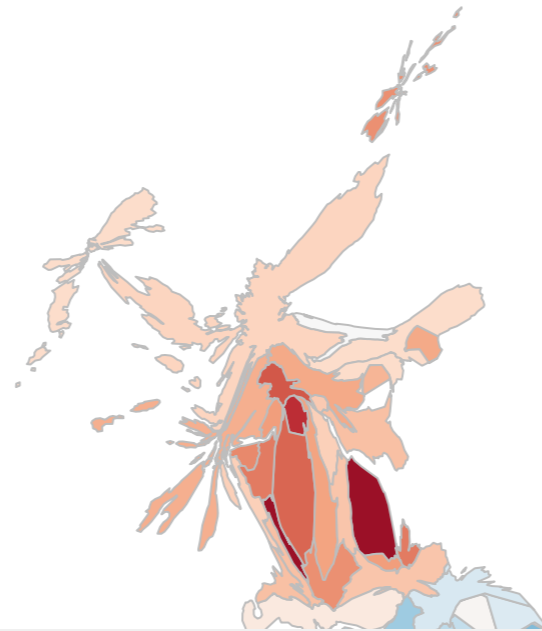


LAs ordered by geospatial
position

LAs ordered by geospatial
position

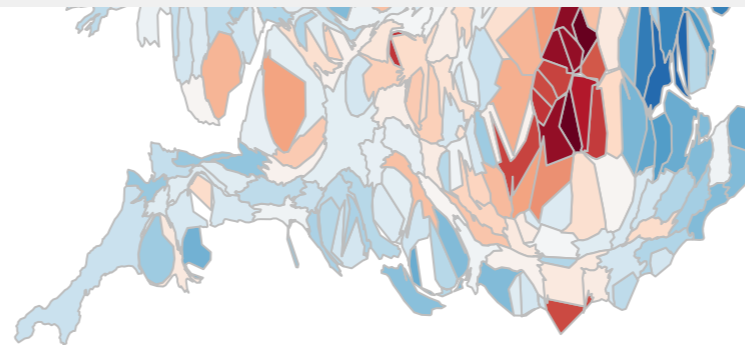


LAs ordered by geospatial position



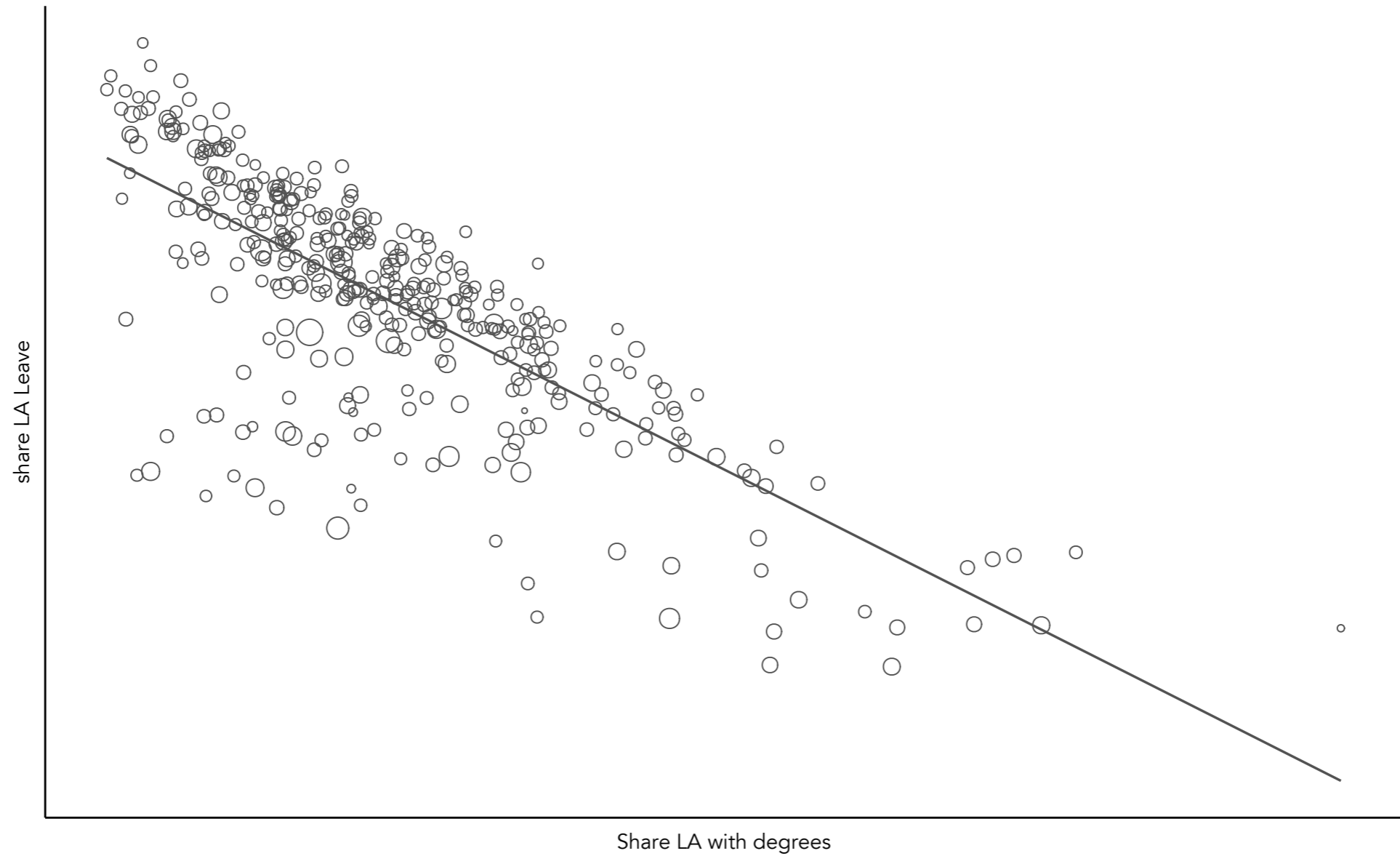
Summary of mappings - specification

data channel	aes	geom	referendum_data %>% ggplot(data=.,
la position	x,y position	poly	aes(x=easting, y=northing,
la area	size	poly	fill=share_leave, size=area) +
margin size	fill (cont)	poly	geom_polygon()
margin direction	fill (hue)	poly	

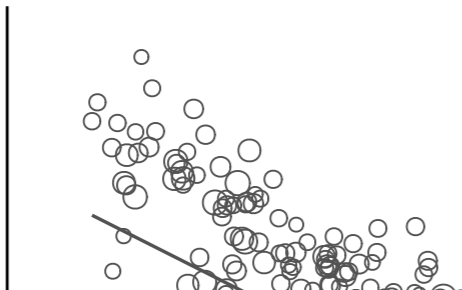


Leave vote by degree-level education

Leave vote by degree-level education



Leave vote by degree-level education

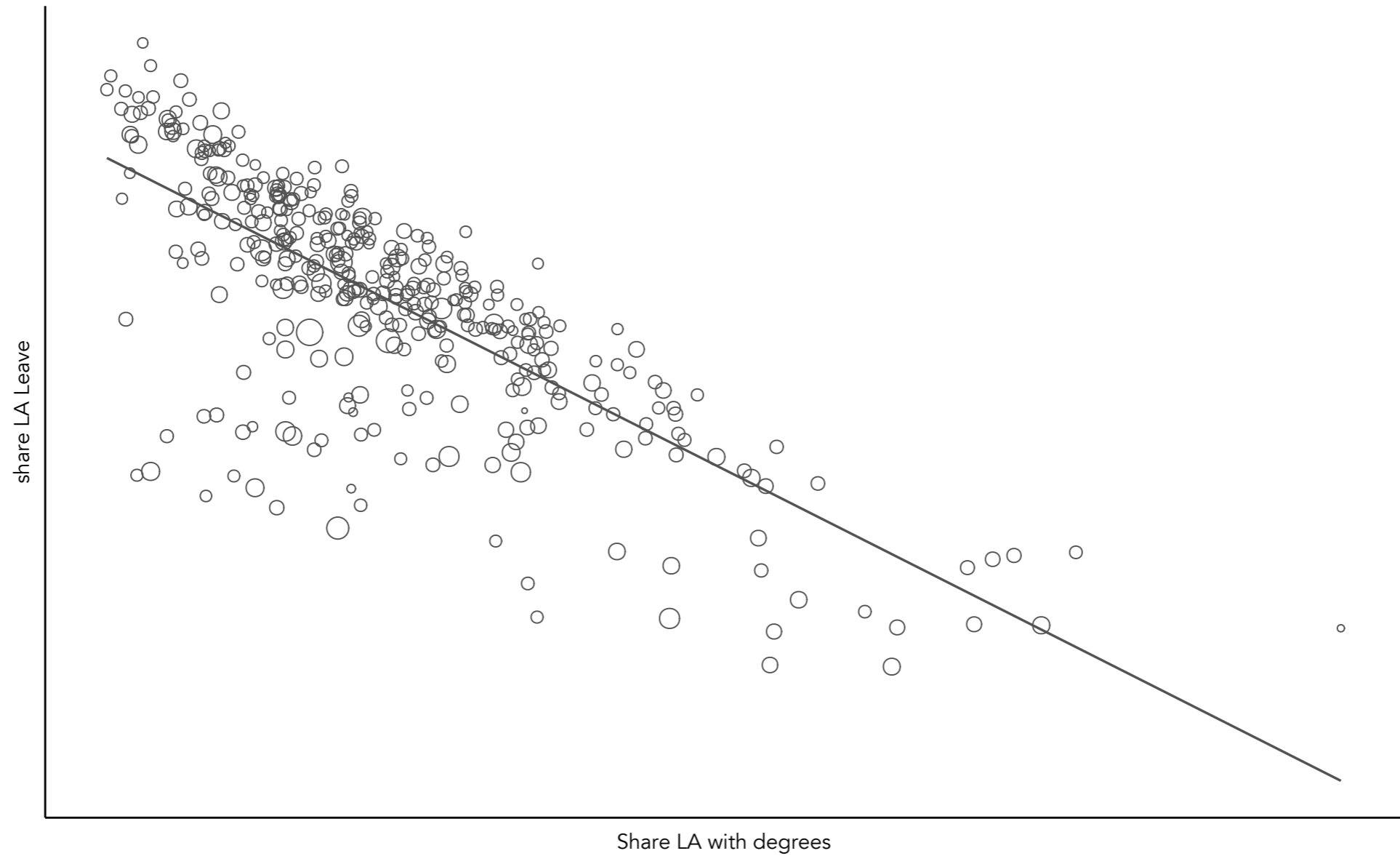


Summary of mappings - specification

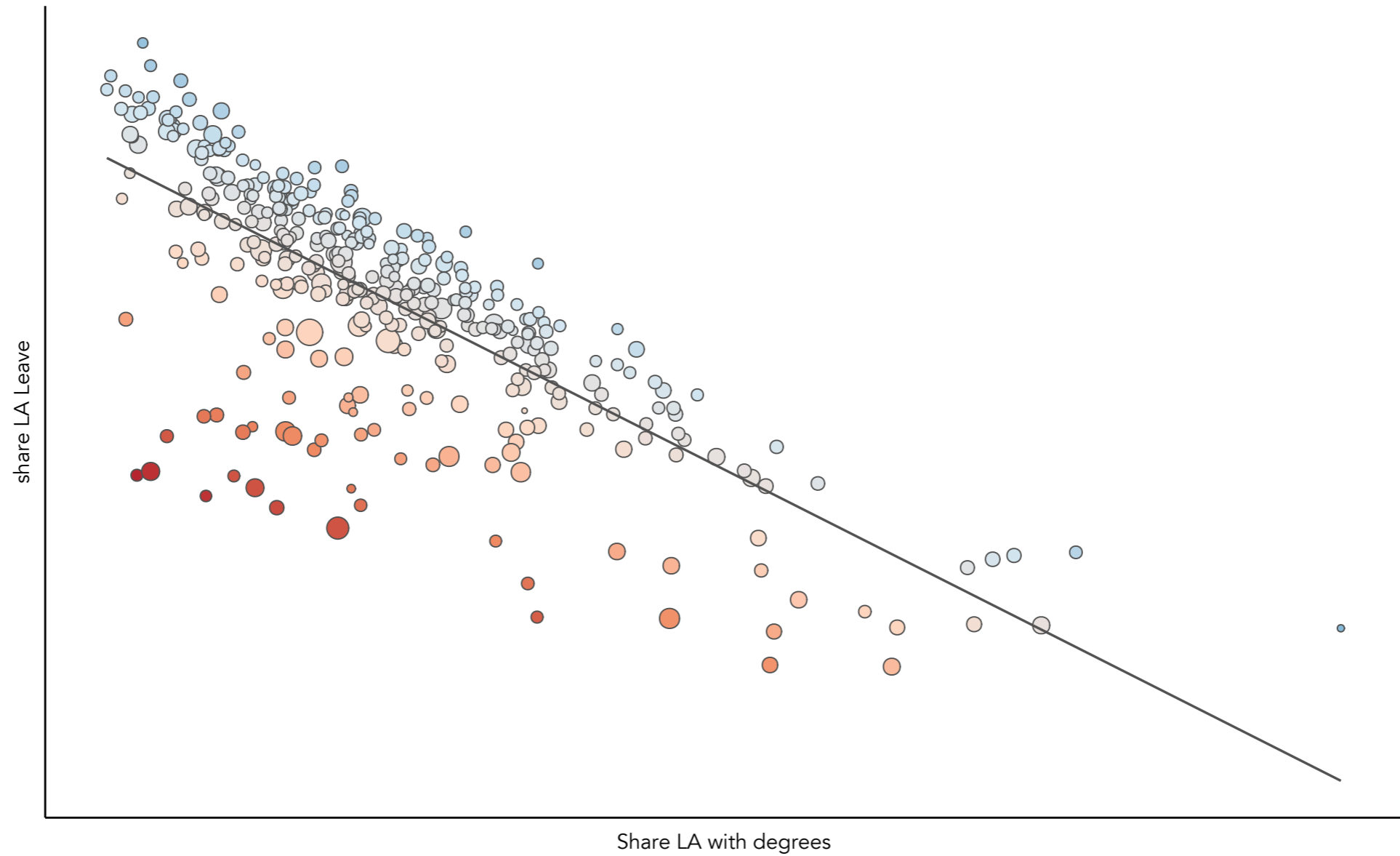
data channel	aes	geom	referendum_data %>% ggplot(data=.,
share leave	x position	point	aes(x=share_leave,
share degrees	y position	point	y=degree_educated,
pop size	size (area)	point	size=electorate) + geom_point()

Share LA with degrees

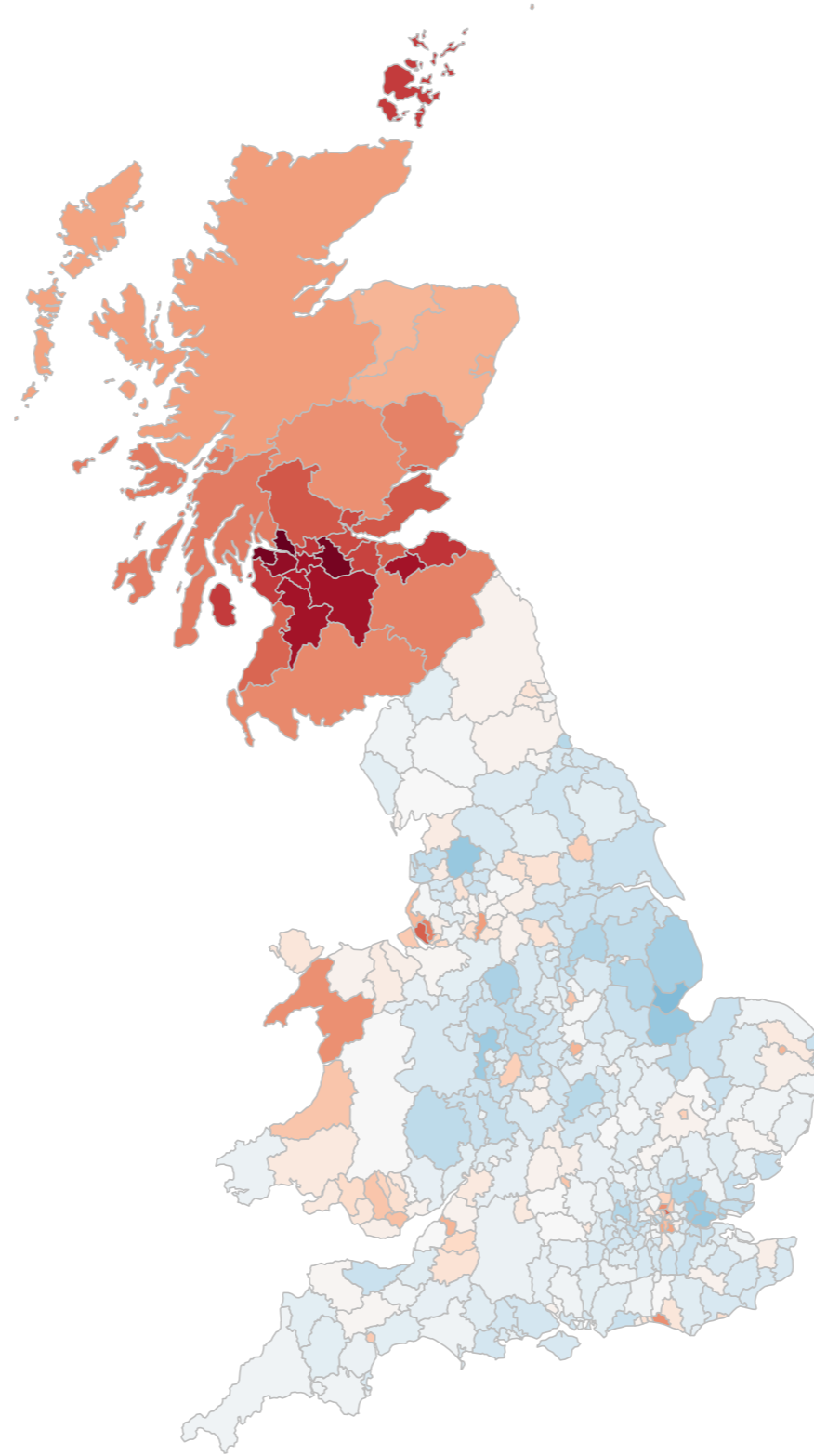
Leave vote by degree-level education



Leave vote by degree-level education



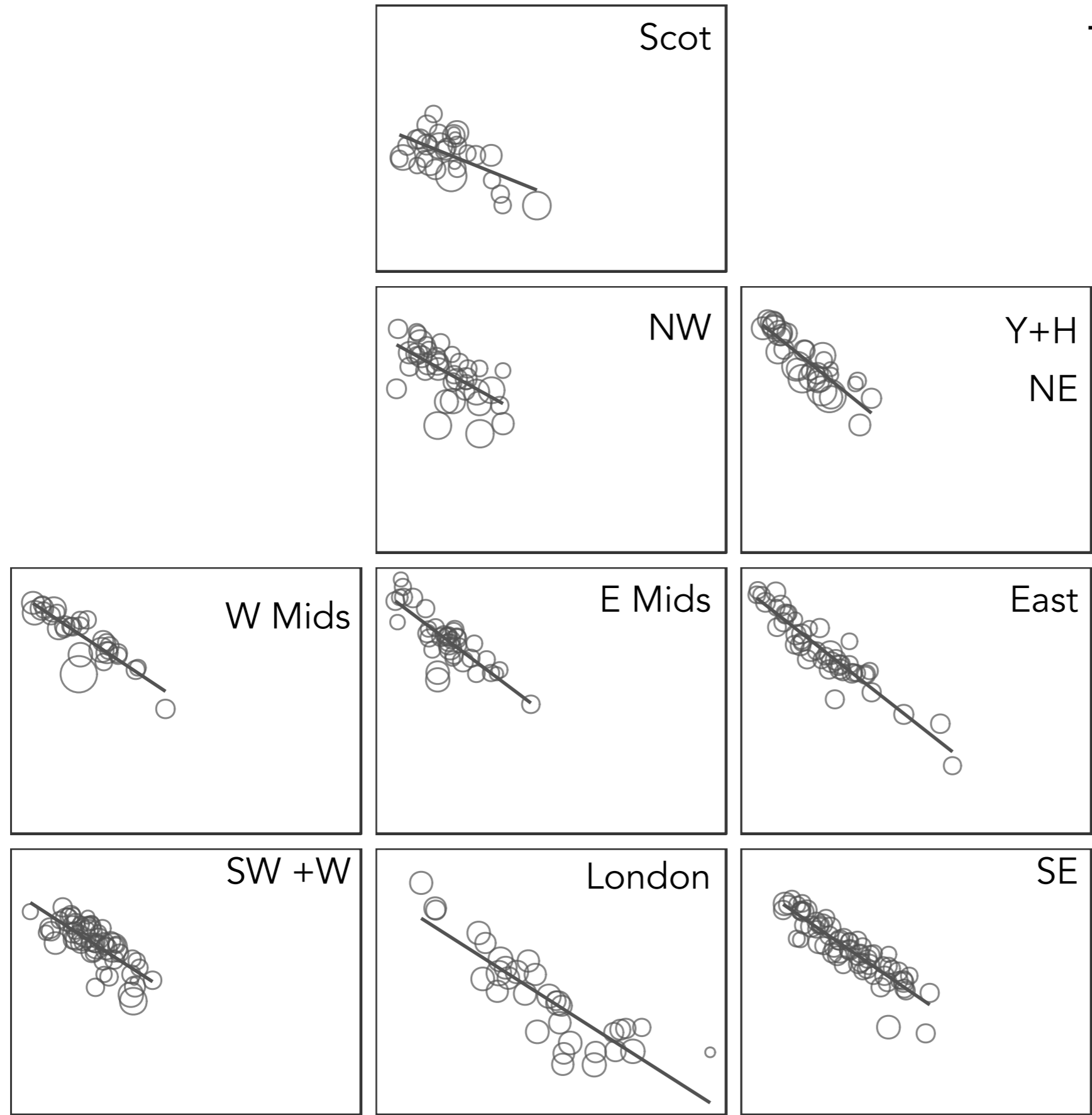
Leave vote by degree-level education



Leave vote by degree-level education
faceted by region

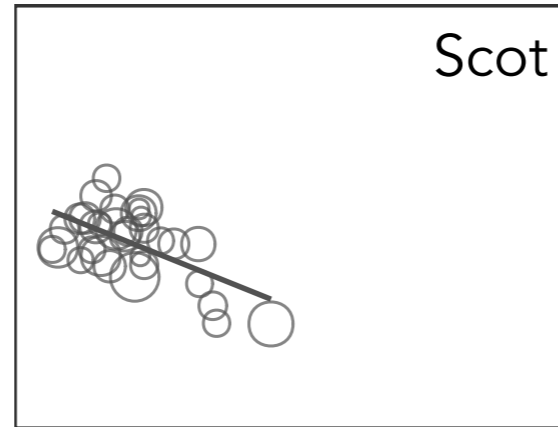
Leave vote by degree-level education

faceted by region



Leave vote by degree-level education

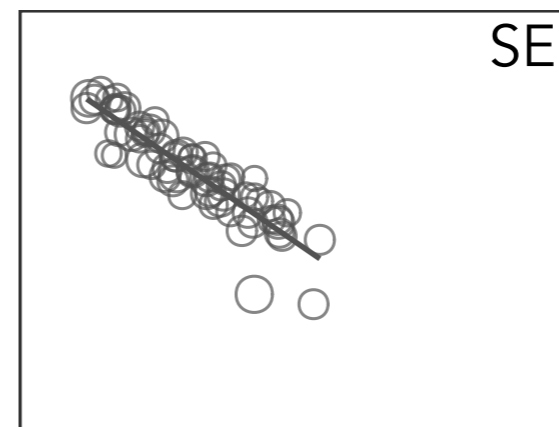
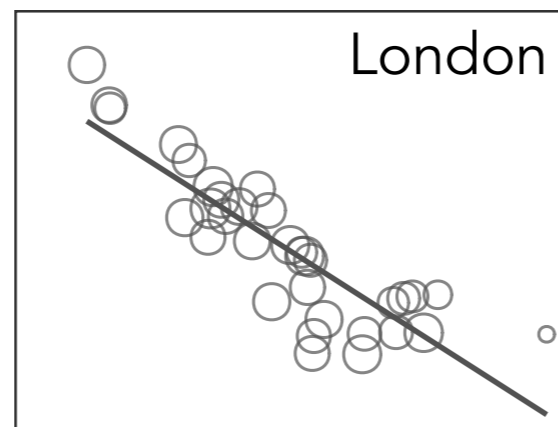
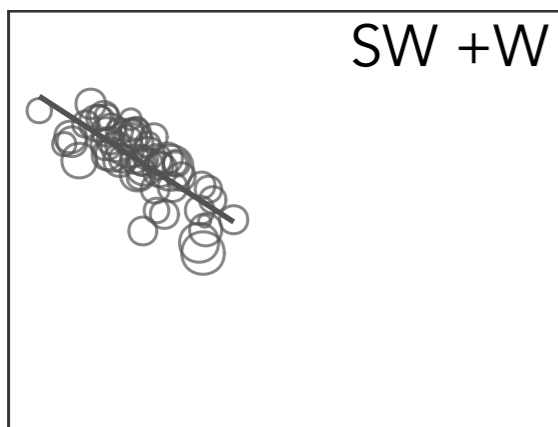
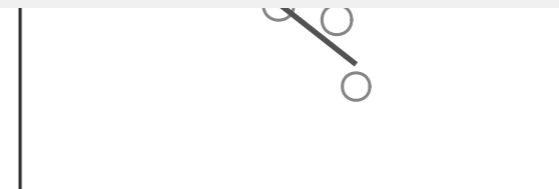
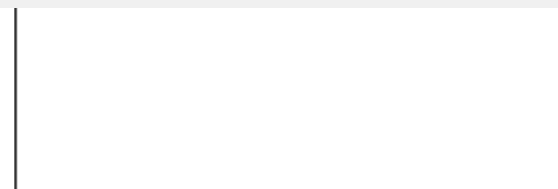
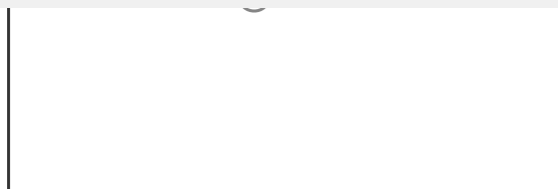
faceted by region

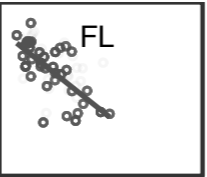
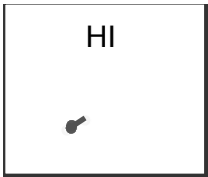
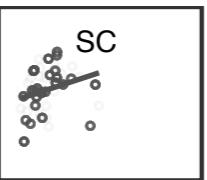
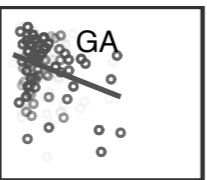
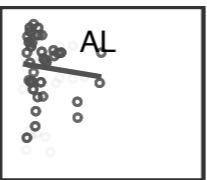
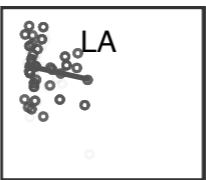
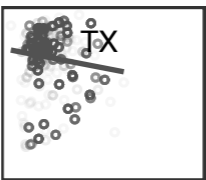
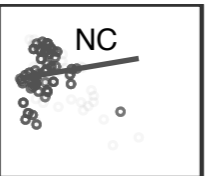
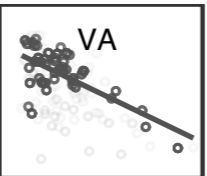
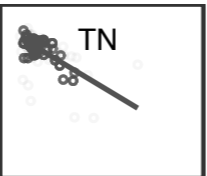
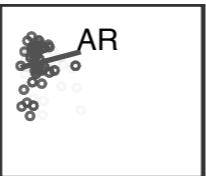
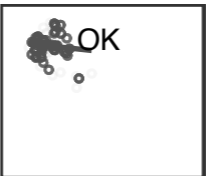
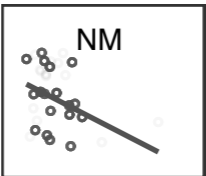
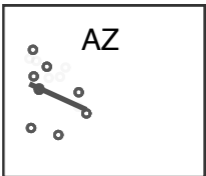
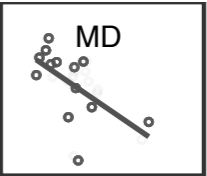
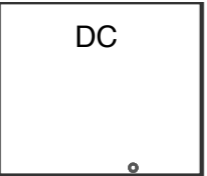
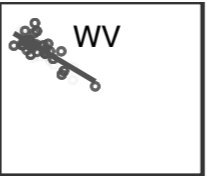
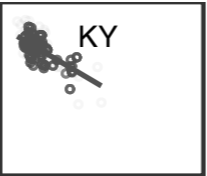
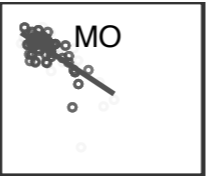
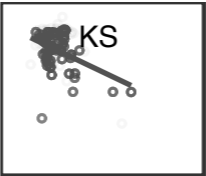
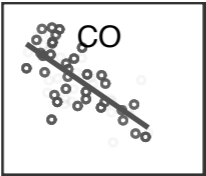
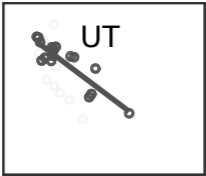
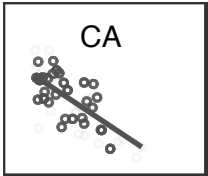
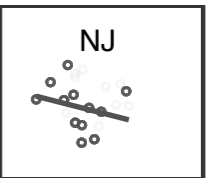
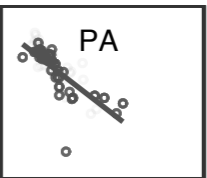
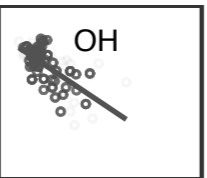
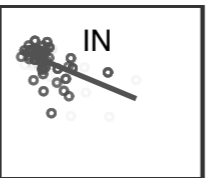
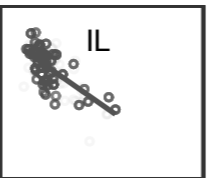
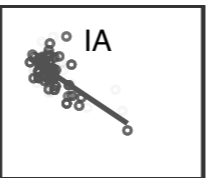
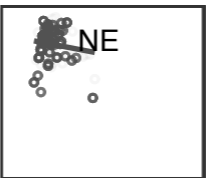
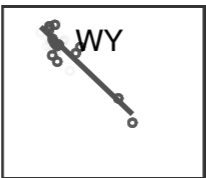
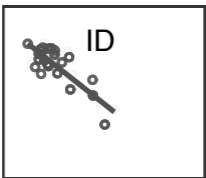
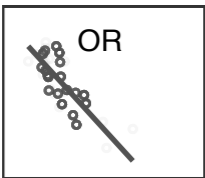
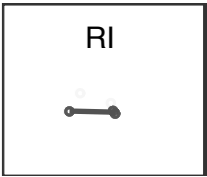
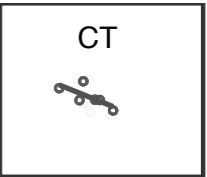
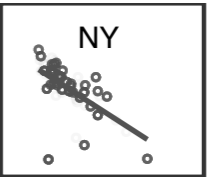
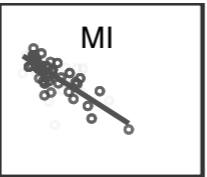
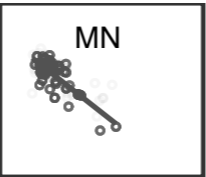
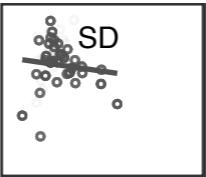
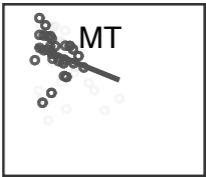
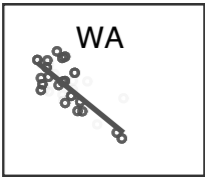
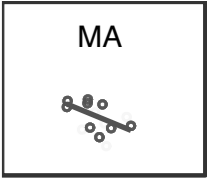
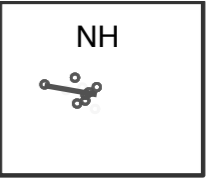
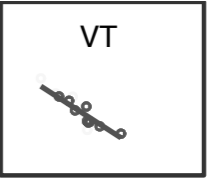
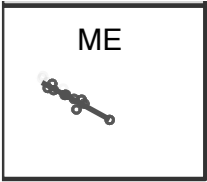


Summary of mappings - specification

data channel	aes	geom
share leave	x position	point
share degrees	y position	point
pop size	size (area)	point
region	plot position	point

```
referendum_data %>%  
ggplot(data=.,  
       aes(x=share_leave,  
           y=degree_educated,  
           size=electorate)) +  
facet_grid(smwgX~smwgY) +  
geom_point()
```





Combining information visualization theory and the grammar of graphics to do and **teach modern data analysis**

```
sp_gb <- as(data_gb, 'Spatial')
# Construct a new data frame, using the cartogram() function, passing as a
# parameter into the function the variable to which polygons are to be sized:
# Electorate (number of voters in LA). This may take a while.
sp_gb <- cartogram(sp_gb, "Electorate", itermax=10)

# Generate a choropleth using the same specification used in the conventional map,
# but supplying the cartogram SpatialDataFrame to tm_shape:
# e.g. tm_shape(sp_gb) +
# etc.
```

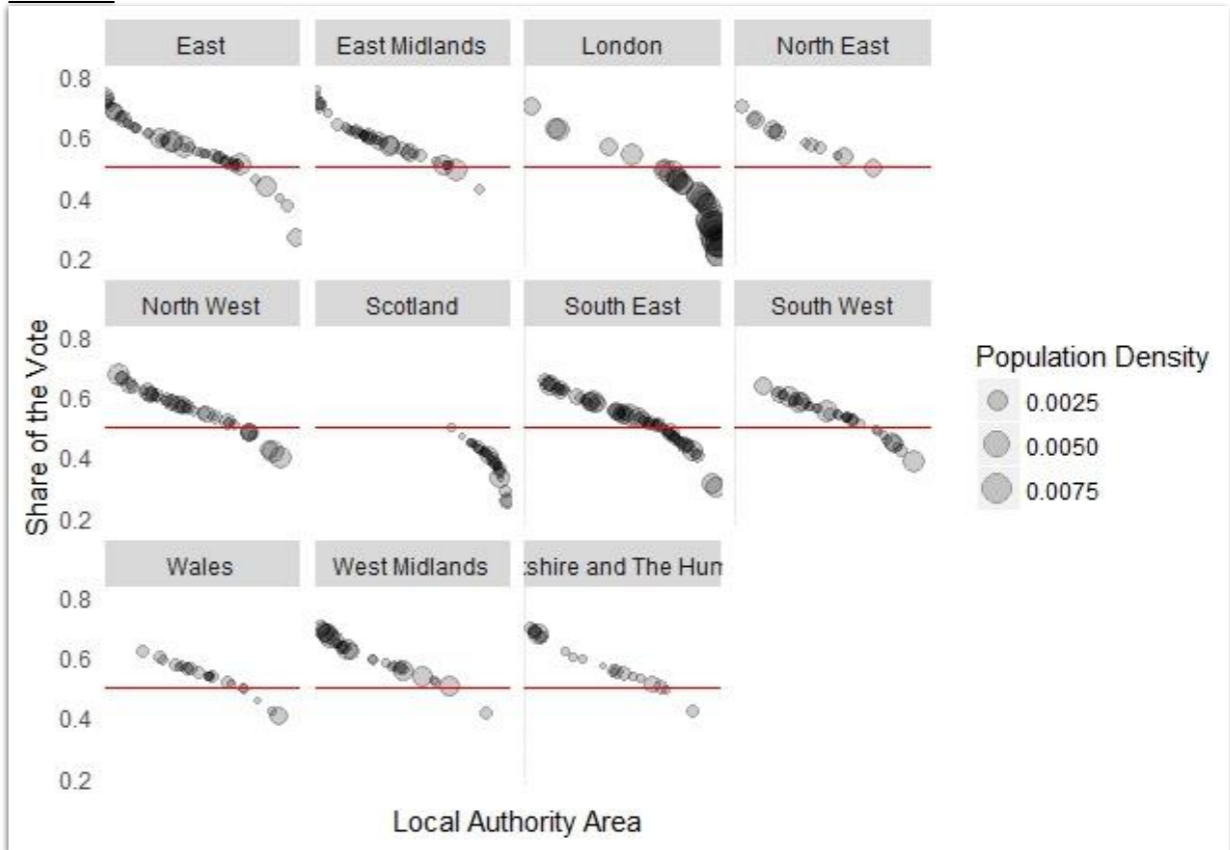
Figure 4: Choropleths displaying the Leave:Remain vote by UK Local Authority.

Table of Contents

- Introduction
- What is R?
- Why R?
 - Free, open-source with an active community
 - R supports modern data analysis workflows
- Task 1. Get familiar with the R and RStudio environment
 - Open RStudio
 - Type some console commands
 - Install some packages
 - Read in some data
- Task 2. Perform some exploratory data analysis
 - Do some data wrangling
 - Create some summary graphics
- Task 3. Perform some exploratory spatial data analysis
 - Load in some spatial (and other attribute) data
 - Explore spatial variation in the Leave:Remain vote
- Task 4. Data challenge
- Task 5. Coding exercises (optional)
- Further reading
 - R for data analysis
 - New Statistics and the reproducibility crisis (non-essential)

Task 4. Data challenge

Despite seeing various iterations of these maps in the weeks after the referendum, the very obvious contrast between most



Term 1 GIS MSc students

Combining information visualization theory and the grammar of graphics to do and teach modern data analysis



JOURNAL OF SPATIAL INFORMATION SCIENCE
Number N (YYYY), pp. xx–yy

doi:10.5311/JOSIS.YYYY.II.NNN

RESEARCH ARTICLE

Locally-varying explanations behind the United Kingdom's vote to leave the European Union

Roger Beecham¹, Aidan Slingsby², and Chris Brunsdon³

¹University of Leeds, UK

²City, University of London, UK

³Maynooth University, Republic of Ireland

Received: M, D, Y; returned: M, D, Y; revised: M, D, Y; accepted: M, D, Y.

Abstract: Explanations behind area-based (Local Authority-level) voting preference in the 2016 referendum on membership of the European Union are explored using aggregate-level data. Developing local models, special attention is paid to whether variables explain the vote equally well across the country. Variables describing the post-industrial and economic 'successfulness' of Local Authorities most strongly discriminate variation in the vote. To a lesser extent this is the case for variables linked to 'metropolitan' and 'big city' contexts, which assist the Remain vote, those that distinguish more traditional and 'nativist' values, assisting Leave, and those loosely describing material outcomes, again reinforcing Leave. Whilst variables describing economic competitiveness co-vary with voting preference equally well across the country, the importance of secondary variables – those distinguishing metropolitan settings, values and outcomes – does vary by region. For certain variables and in certain areas, the direction of effect on voting preference reverses. For example, whilst levels of European Union migration mostly assist the Remain vote, in parts of the country the opposite effect is observed.

Keywords: European Union; referendum; multi-level modelling; geographically-weighted statistics; LASSO; area-based analysis.

variables distinguishing LAs that are within London and Scotland. The line through the regression coefficients in Figure 4 and their transparency is determined by 95% confidence intervals calculated via a bootstrap.

The model created under this LASSO procedure identified six variables. *Degree-educated* contributes the largest coefficient effect. Holding the other variables constant, a one percent point increase in the *degree-educated* population decreases the leave vote by 0.9 percent points. The fact that Scotland is selected by the LASSO procedure is instructive: there is something fundamentally different about Scotland, not accounted for completely by census variables, that lowers preference for Leave (by 16% points after controlling for demographics). The effect of the *EU-born* variable is counter to that expected. In Figure 1 the variable appears negatively correlated with Leave and we speculate might represent economic opportunity and relative diversity. After controlling for variation in other demographic characteristics, the model suggests an increase in the *EU-born* population in fact *increases* the Leave vote. Notice, however, the large confidence interval around this coefficient. Given the resampling procedure used to generate our bootstrap, this interval indicates that the effect of *EU-born* is likely to vary across LAs.

4.3 Region-specific explanations implied by local models

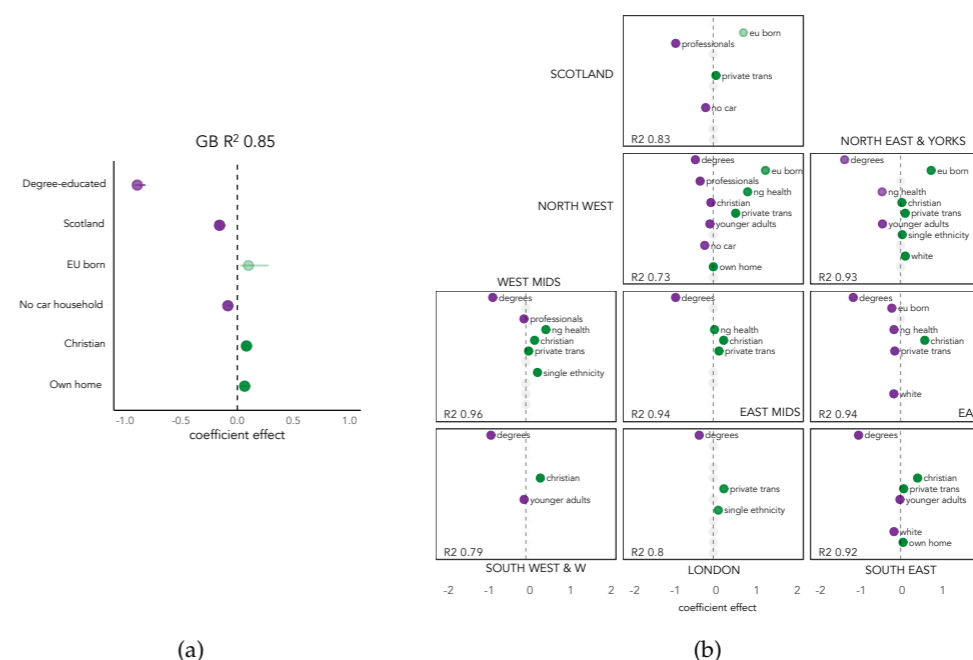
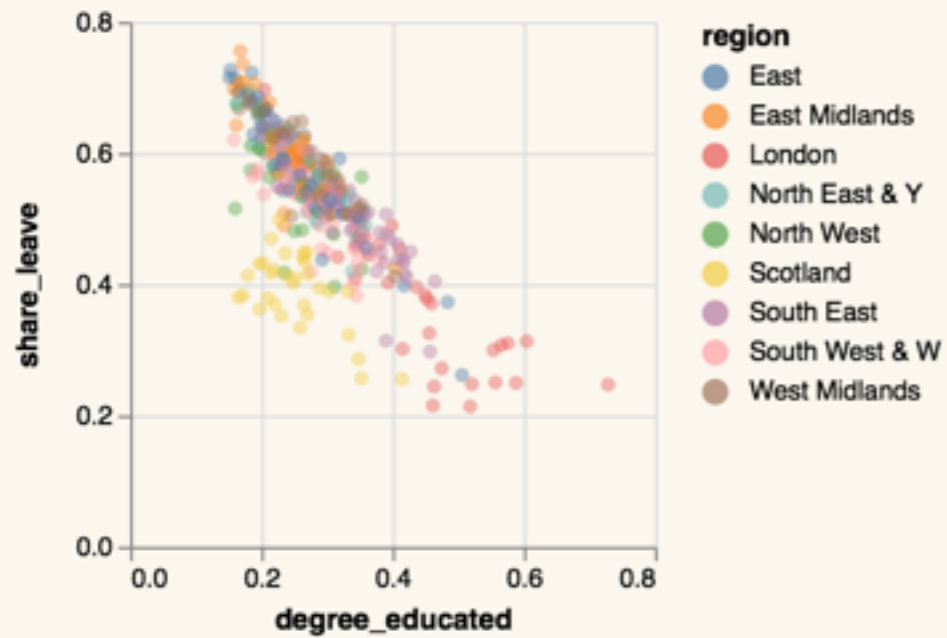


Figure 4: Coefficients for multivariate models fit to data for GB (4a) and super-regions (4b) and annotated with adjusted R^2 . Positive coefficients are green, negative purple and colour lightness varies according to a 95% Confidence Interval calculated via a bootstrap. Note that the GB model was specified with additional dummy variables for Scotland and London.

Vega-Lite

Data	static or data source
Transform	filter, aggregation, binning
Mark	point, line, bar, polygon
Encoding	mapping between data and mark properties
Scale	functions that map data values to visual values
Guides	axes and legends

Vega-Lite



Leave against degree-educated

```
"data": {"url": "data/data_gb.csv"},  
"mark": {"type": "point", "filled": true},  
"encoding": {  
  "x": {"field": "degree_educated", "type": "quantitative"},  
  "y": {"field": "share_leave", "type": "quantitative"},  
  "color": {"field": "region", "type": "nominal"}  
}
```

Plot grammar

Vega-Lite



```
"data": {"url": "data/data_gb.csv"},  
"repeat": {"column": ["degree_educated", "not_good_health",  
"private_transport_to_work"]},  
"spec": {  
  "mark": {"type": "point", "filled": true},  
  "encoding": {  
    "x": {"field": {"repeat": "column"}, "type": "quantitative"},  
    "y": {"field": "share_leave", "type": "quantitative"},  
    "color": {"field": "region", "type": "nominal"}  
  }  
}
```

selections

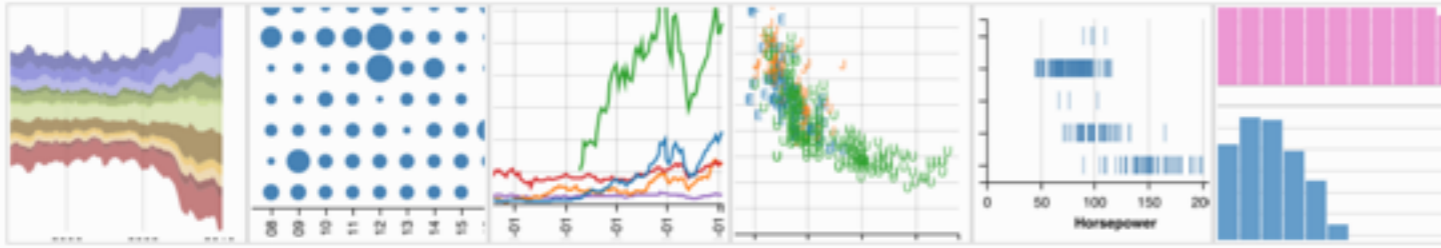
map user input (e.g. mouse moves)

into data queries

which drive conditional encodings, filter data points etc.

```
"data": {"url": "data/data_gb.csv"},
"mark": {"type": "point", "filled": true},
"selection" : {"picked": {"type": "single", "on": "mouseover"}},
"encoding": {
  "x": {"field": "degree_educated", "type": "quantitative"},
  "y": {"field": "share_leave", "type": "quantitative"},
  "color": {
    "condition":
      {"selection": "picked", "field": "region", "type": "nominal"}, "value": "grey"}
  }
}
```


Declarative Visualization in Python

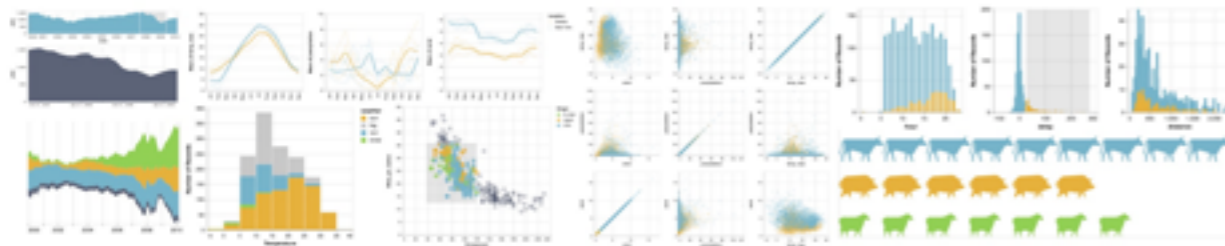


Altair is a declarative statistical visualization library for Python, based on [Vega-Lite](#).

With Altair, you can spend more time understanding your data and its meaning. Altair's API is simple, friendly and consistent and built on top of the powerful [Vega-Lite](#) visualization grammar. This elegant simplicity produces beautiful and effective visualizations with a minimal amount of code.

github.com/altair-viz/altair

elm-vega



Declarative visualization for Elm

This library allows you to create Vega-Lite specifications in Elm providing a pure functional interface for visualization construction.

The library does not generate graphical output directly, but instead it allows you to create a JSON sent to the Vega-Lite runtime to create the output. This is therefore a 'pure' Elm package without dependencies.

github.com/gicentre/elm-vega



Teaching materials github.com/rogerbeecham/intro-visual-data-analysis/

Paper and code github.com/rogerbeecham/brexit-analysis/

OBSERVABLE