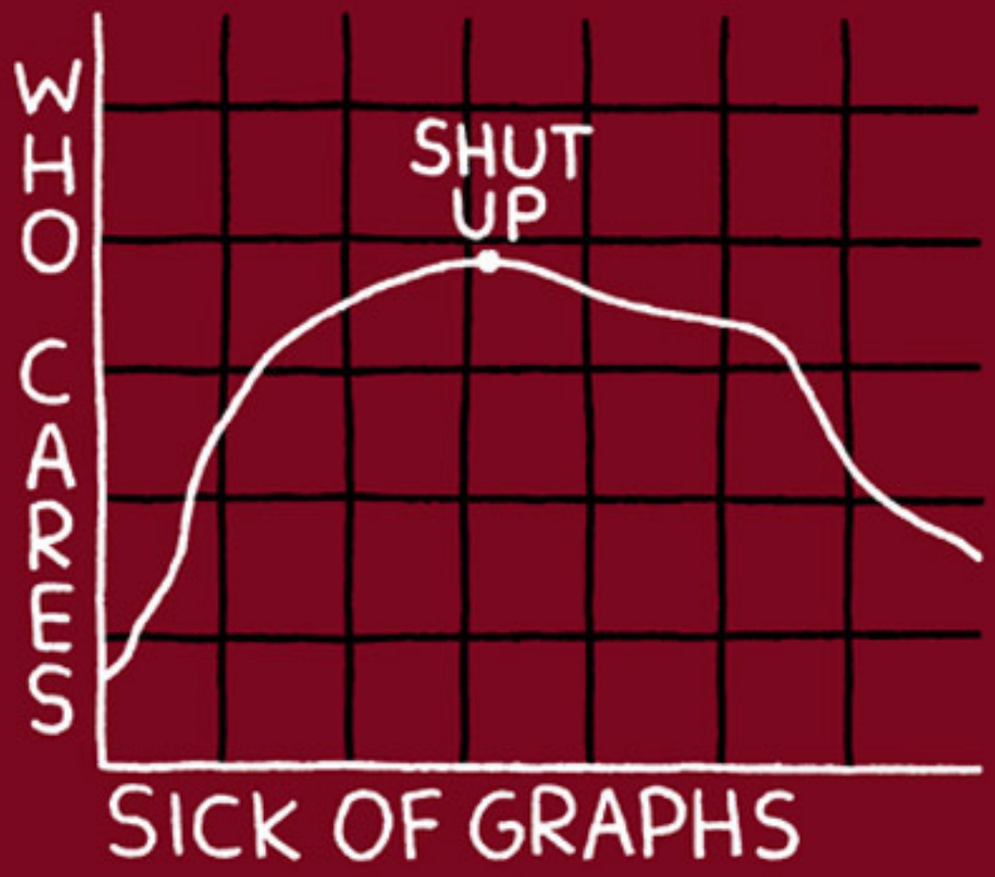
SMI610: Social Analytics & Visualisation



Mark Taylor

Who is this guy, anyway?

I'm "Senior Lecturer in Quantitative Methods" at Sheffield

- I research the relationship between culture and inequality
- participation, funding, workforces, etc

I draw a lot of graphs

What's happening this week?

Four days of new content

- data viz today
- text analysis tomorrow/Weds am
- social network analysis Weds pm/Thurs

One day of thinking about assessments

- (Friday, though you get moving earlier)

Some more logistics

Blackboard

- where docs will be posted
- where you'll submit the assessment

Slack

- it's likely quicker if you post queries there, as opposed to by email

Some acknowledgments

I very strongly recommend four books:

- R for Data Science, by Hadley Wickham
- Data Visualisation, by Kieran Healy
- Fundamentals of Data Visualisation, by Claus Wilke
- R Graphics Cookbook, by Winston Chang

The first three are available free online

- If you read them cover-to-cover, you'll recognise some material from today

What's happening today?

Introduce you to visualising data

- what do I mean by "visualisation"?
- what do I mean by "data"?
- introduction to ggplot2
- getting hold of data
- manipulating data so it can be visualised

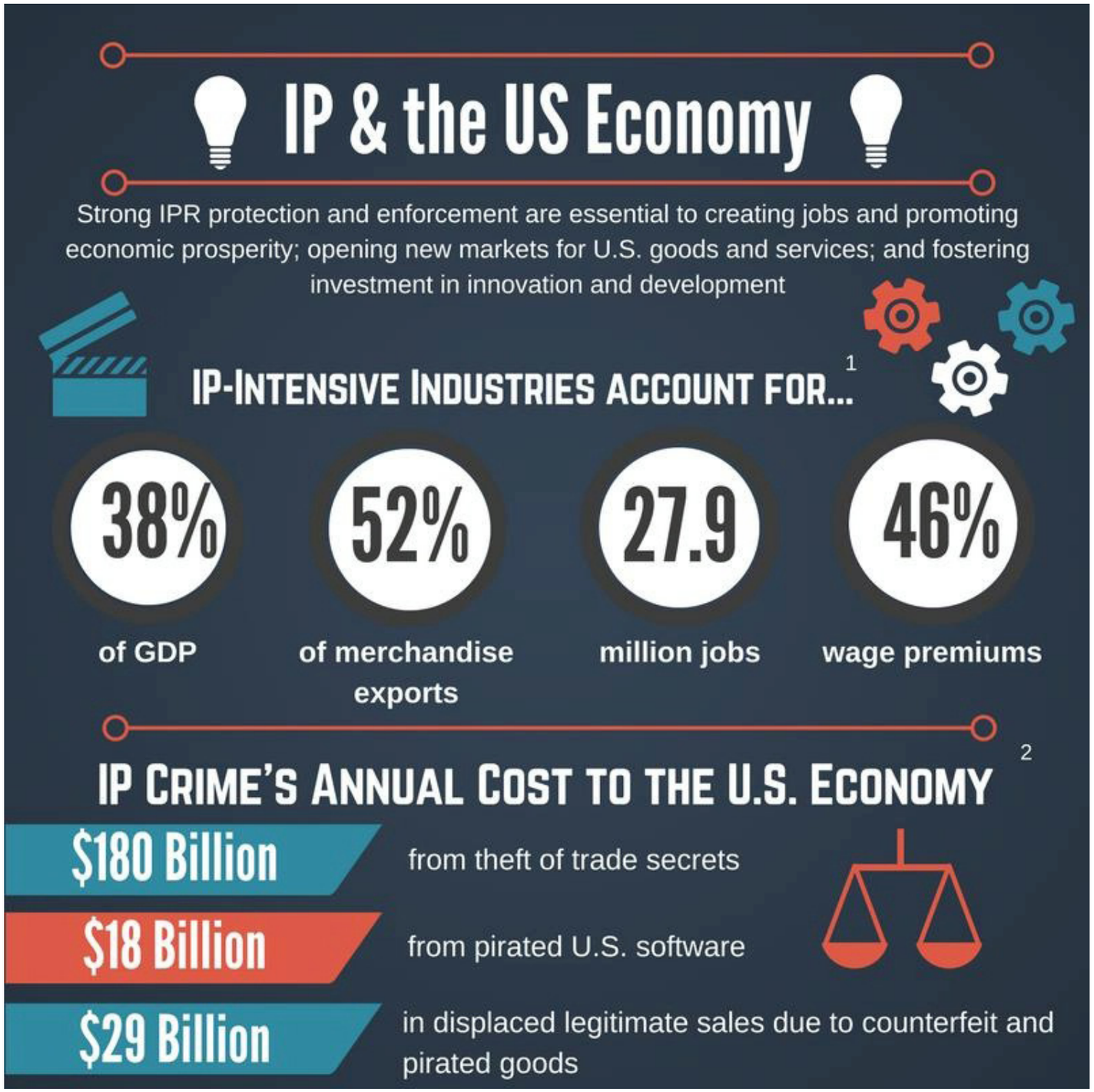
What's data visualisation?

"The techniques used to communicate data or information by encoding it as visual objects contained in graphics"

- (thanks, Wikipedia)

Anything that converts numbers into geometric objects (this is a term we'll be revisiting)

What am I excluding?



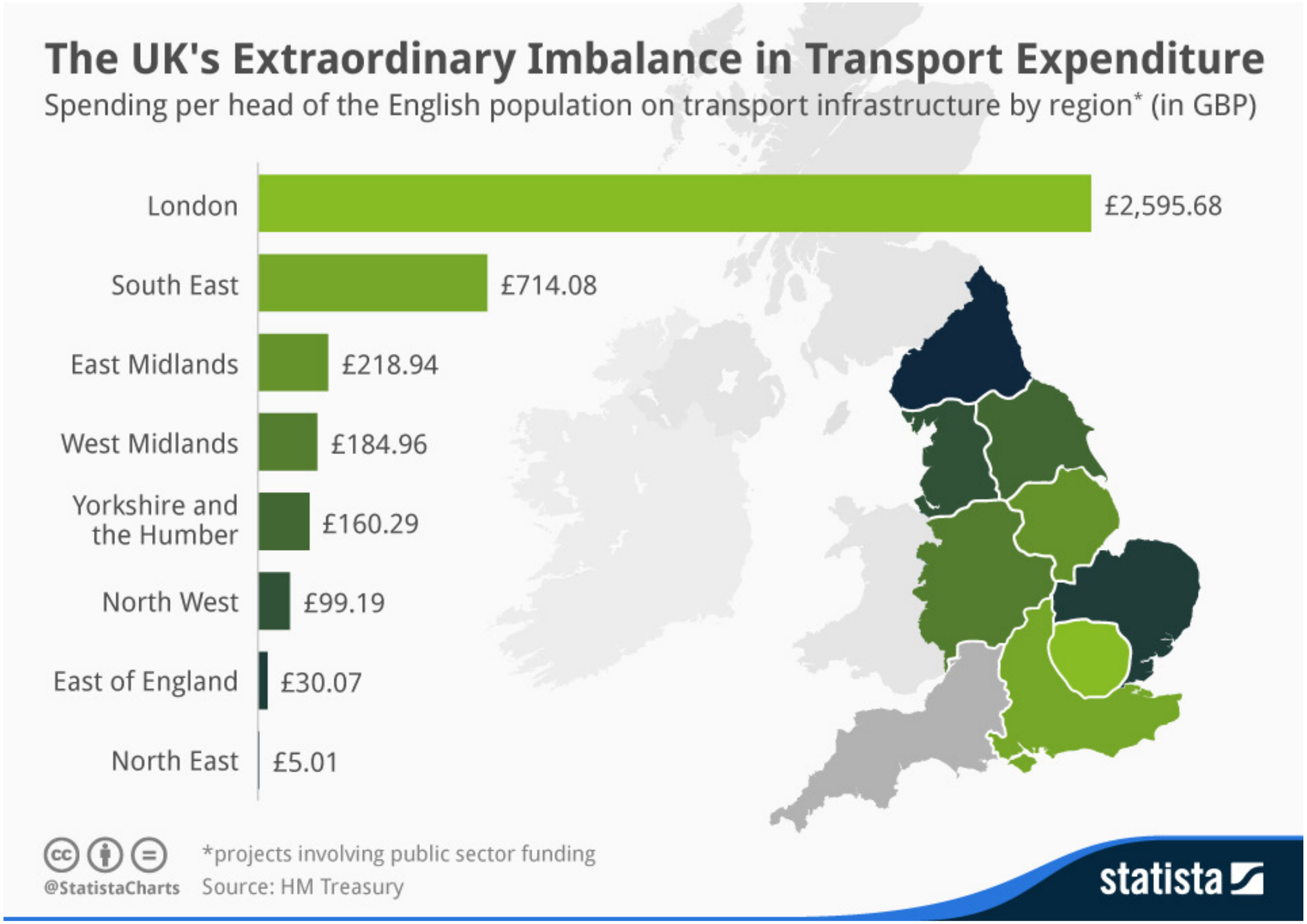
Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Let's look at some graphs

Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Sheffield Methods Institute.

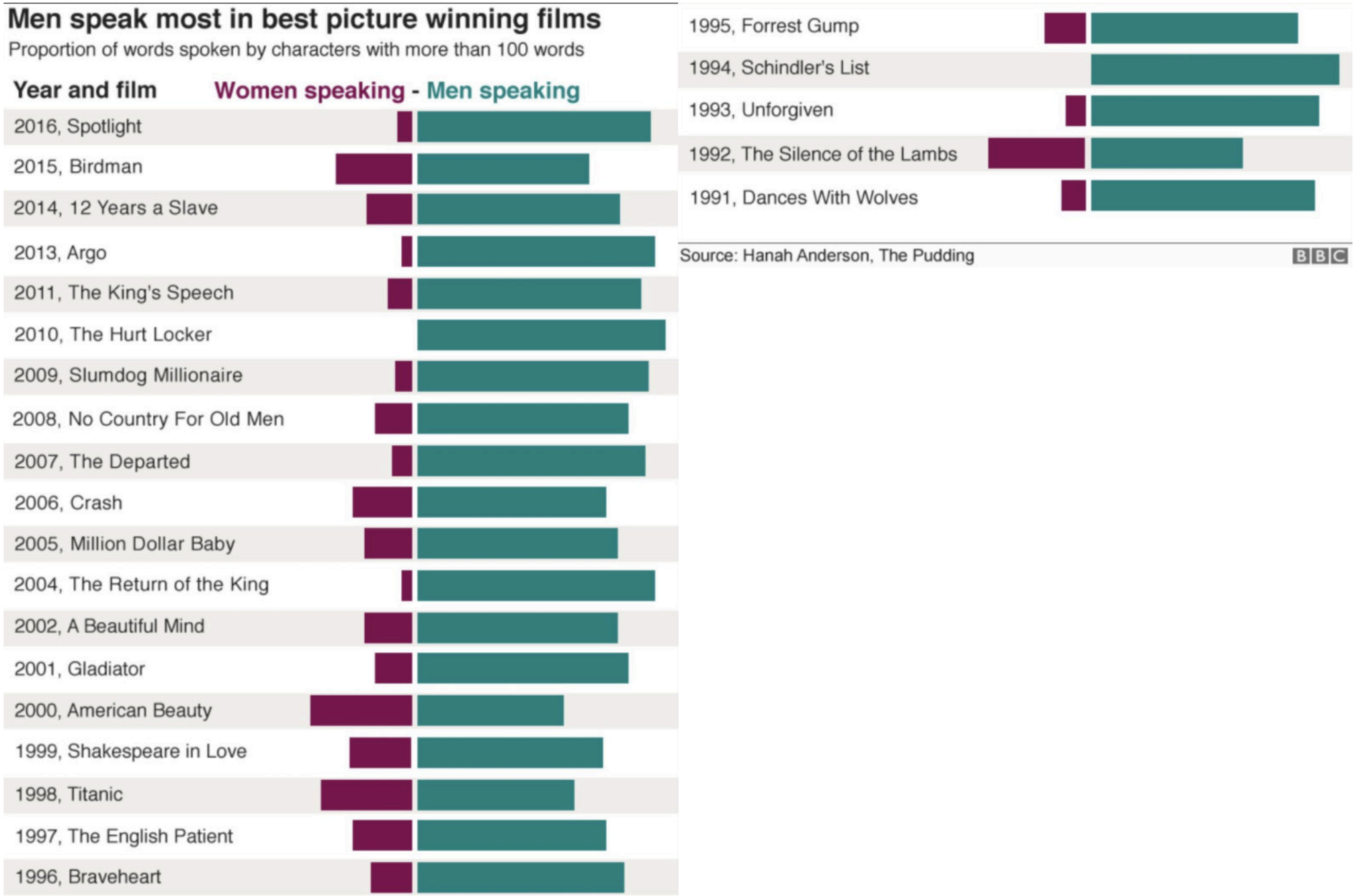
Some graphics



Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Sheffield Methods Institute.

Some graphics

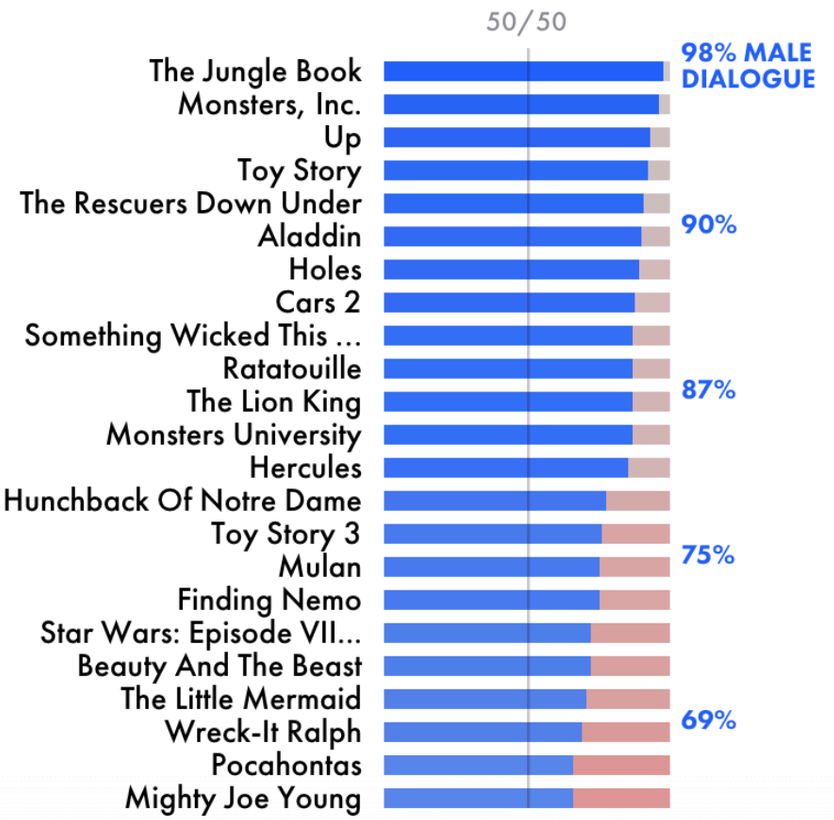


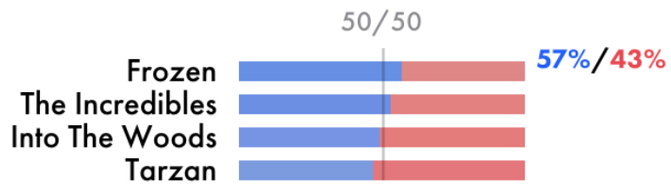
Some graphics

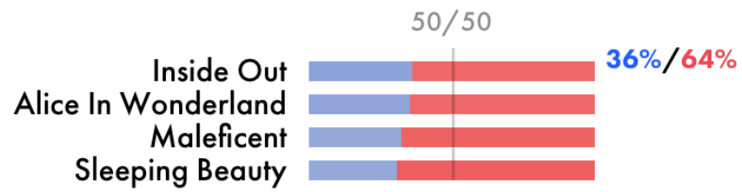
Screenplay Dialogue, Broken-down by Gender

2,000 Screenplays: Dialogue Broken-down by Gender

Only High-Grossing Films: Ranked in the Top 2,500 by US Box Office*







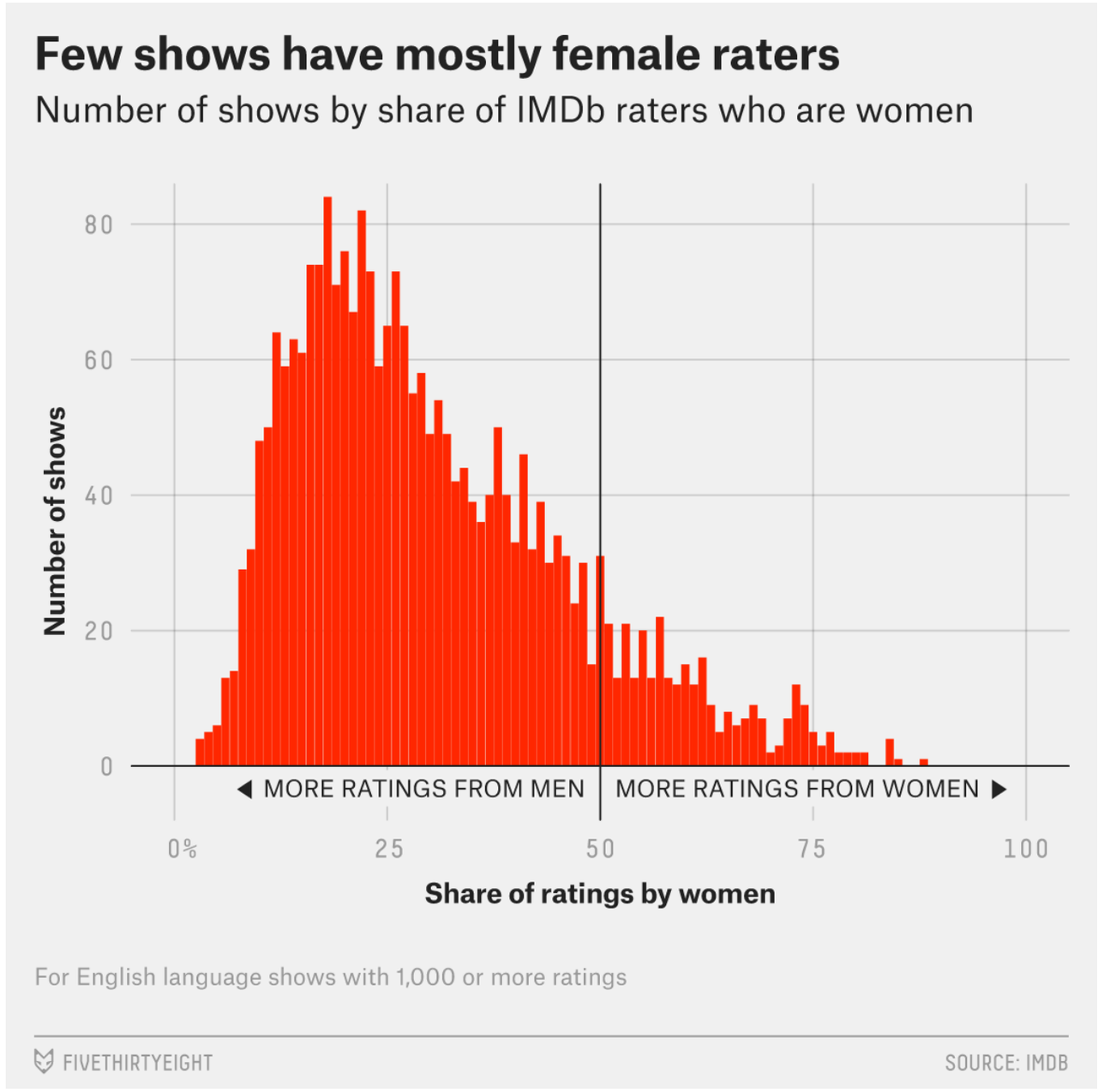
Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

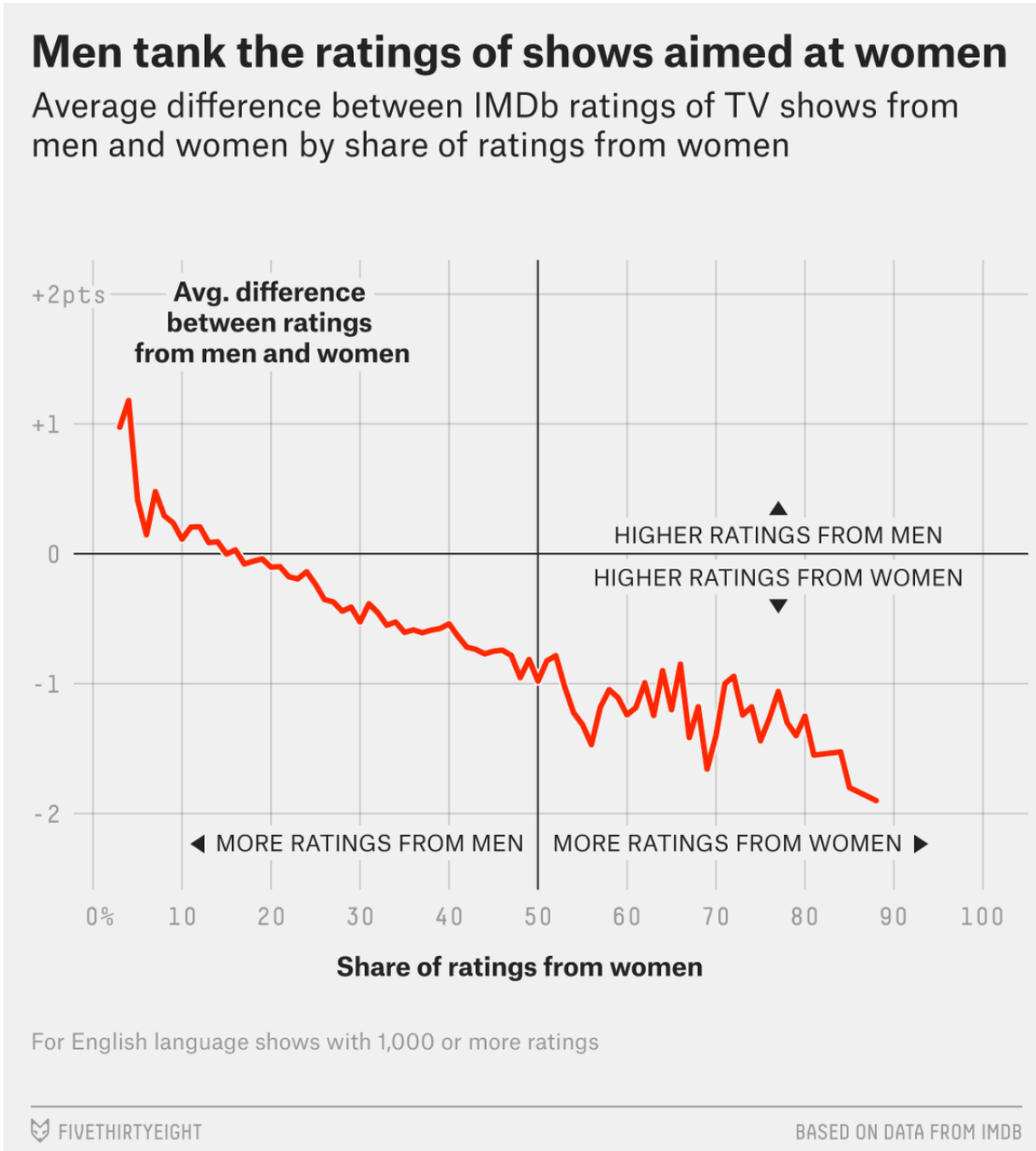
Some graphics

Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Sheffield Methods Institute.

Some graphics





- https://fivethirtyeight.com/features/men-are-sabotaging-the-online-reviews-of-tv-shows-aimed-at-women/

Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Some graphics

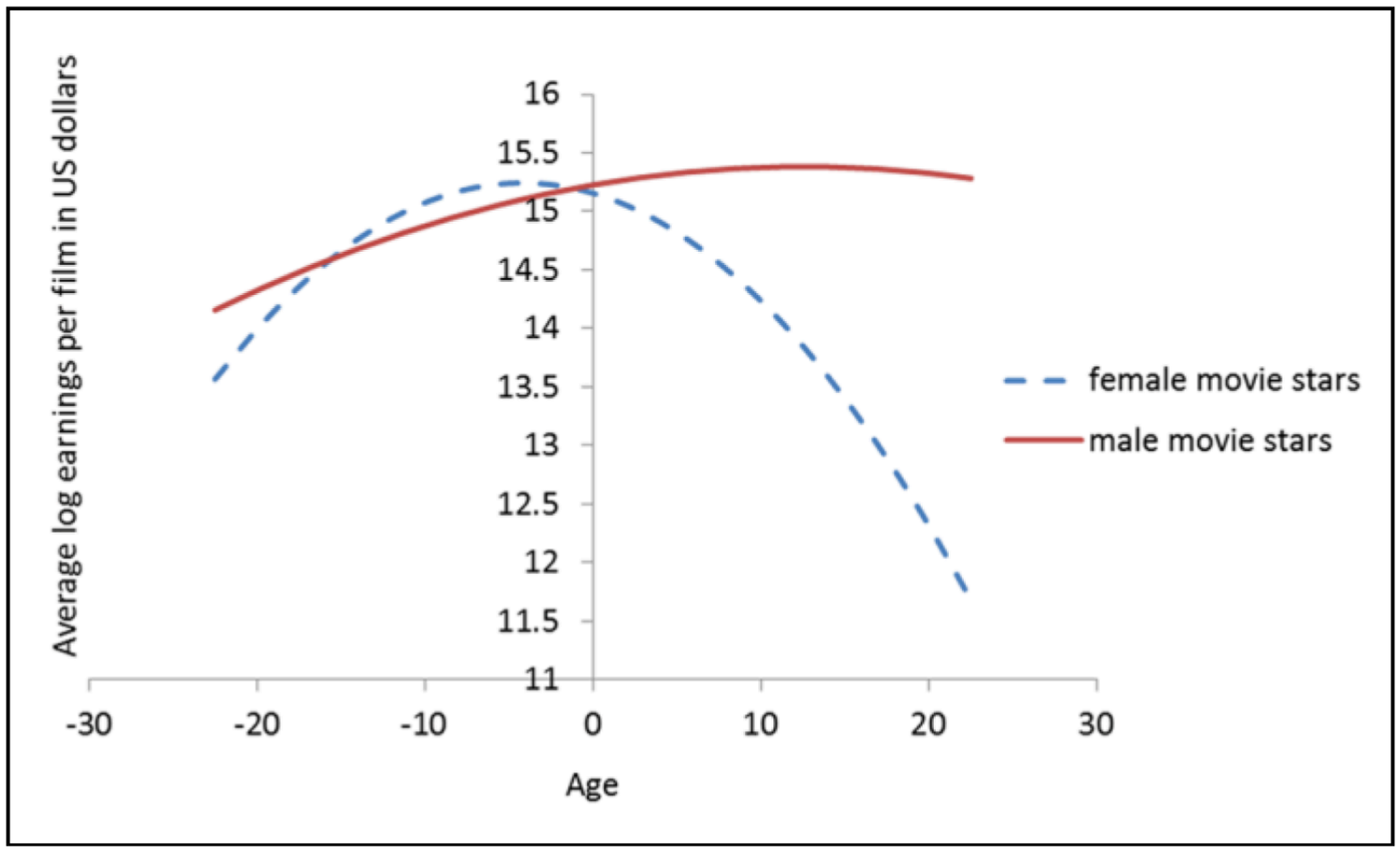
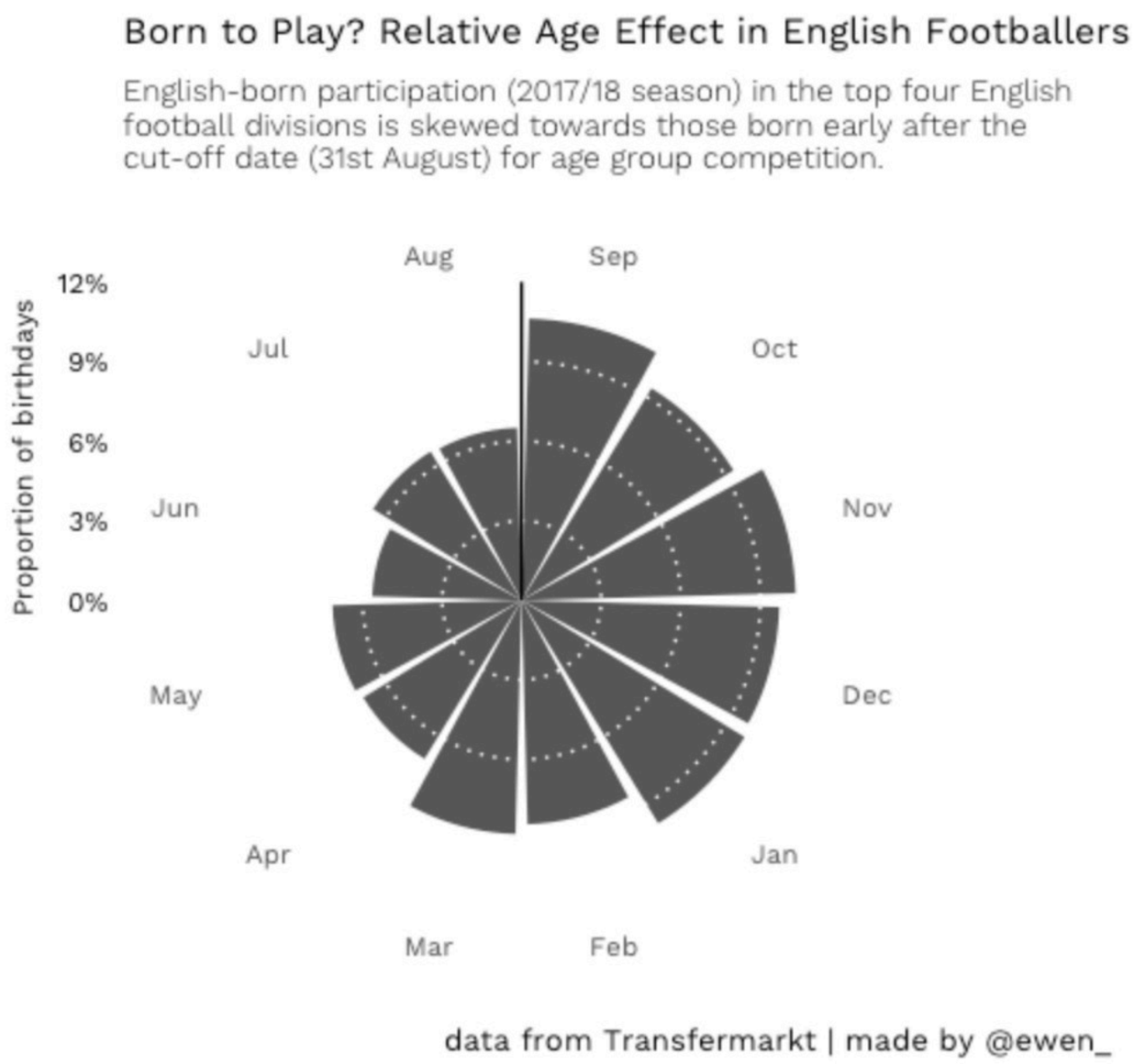


Figure 1. Interactive effect of gender and age^2 on average log earnings per film.

- De Pater, IE, Judge, TA, and Scott, BA (2014). Age, Gender, and Compensation: A Study of Hollywood Movie Stars. *Journal of Management Inquiry* 23(4).

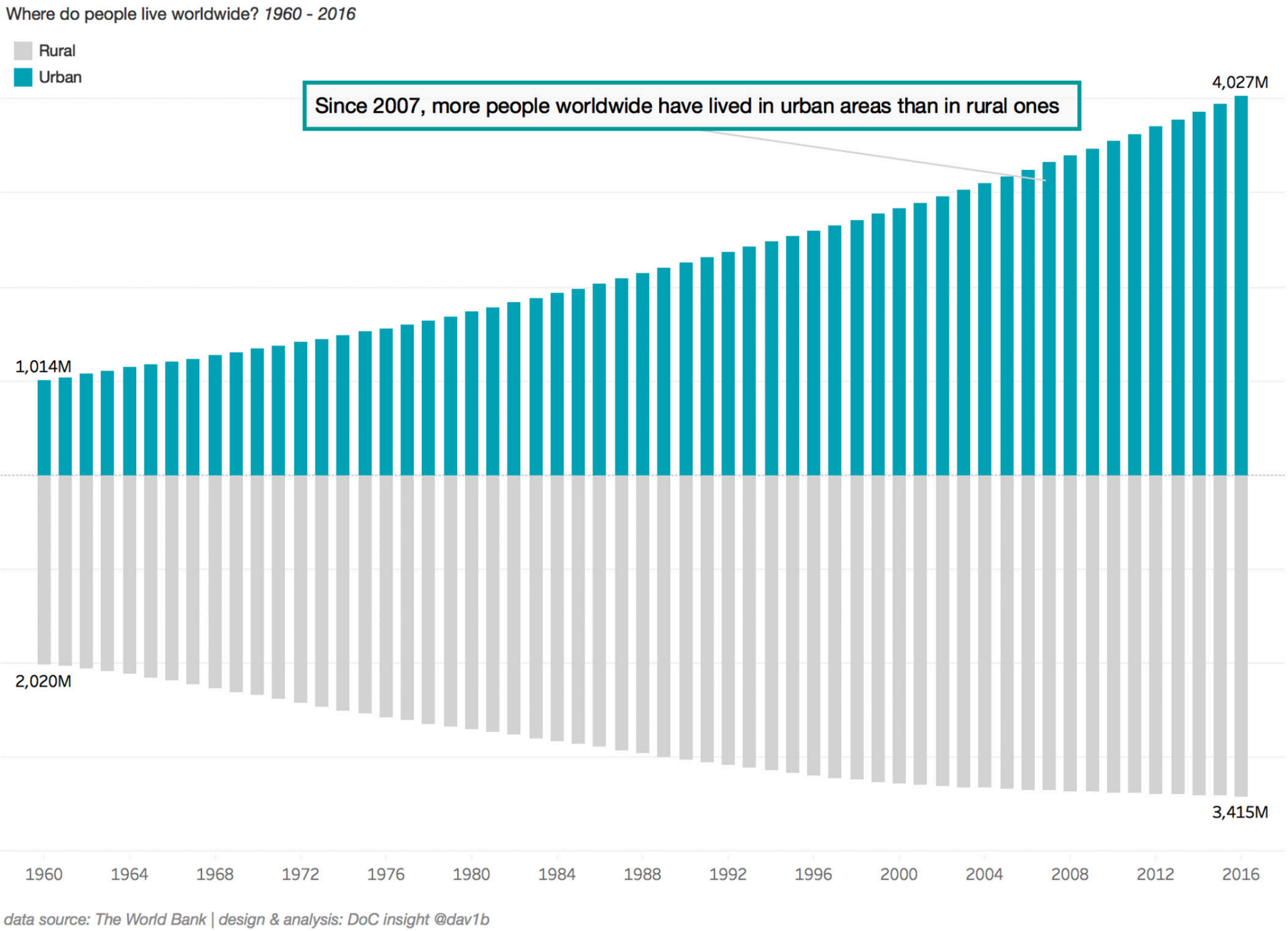
Some graphics



Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Sheffield Methods Institute.

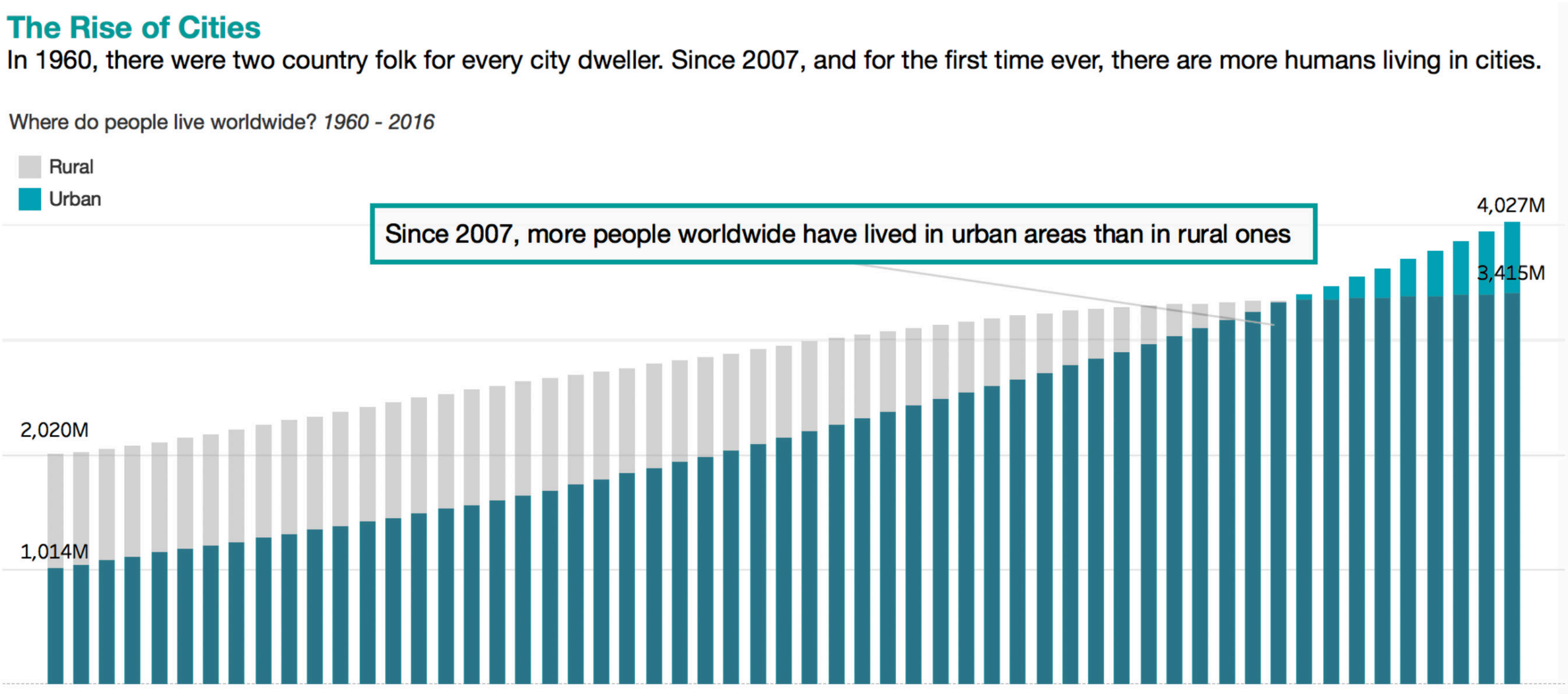
Some graphics



Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Sheffield Methods Institute.

Some graphics

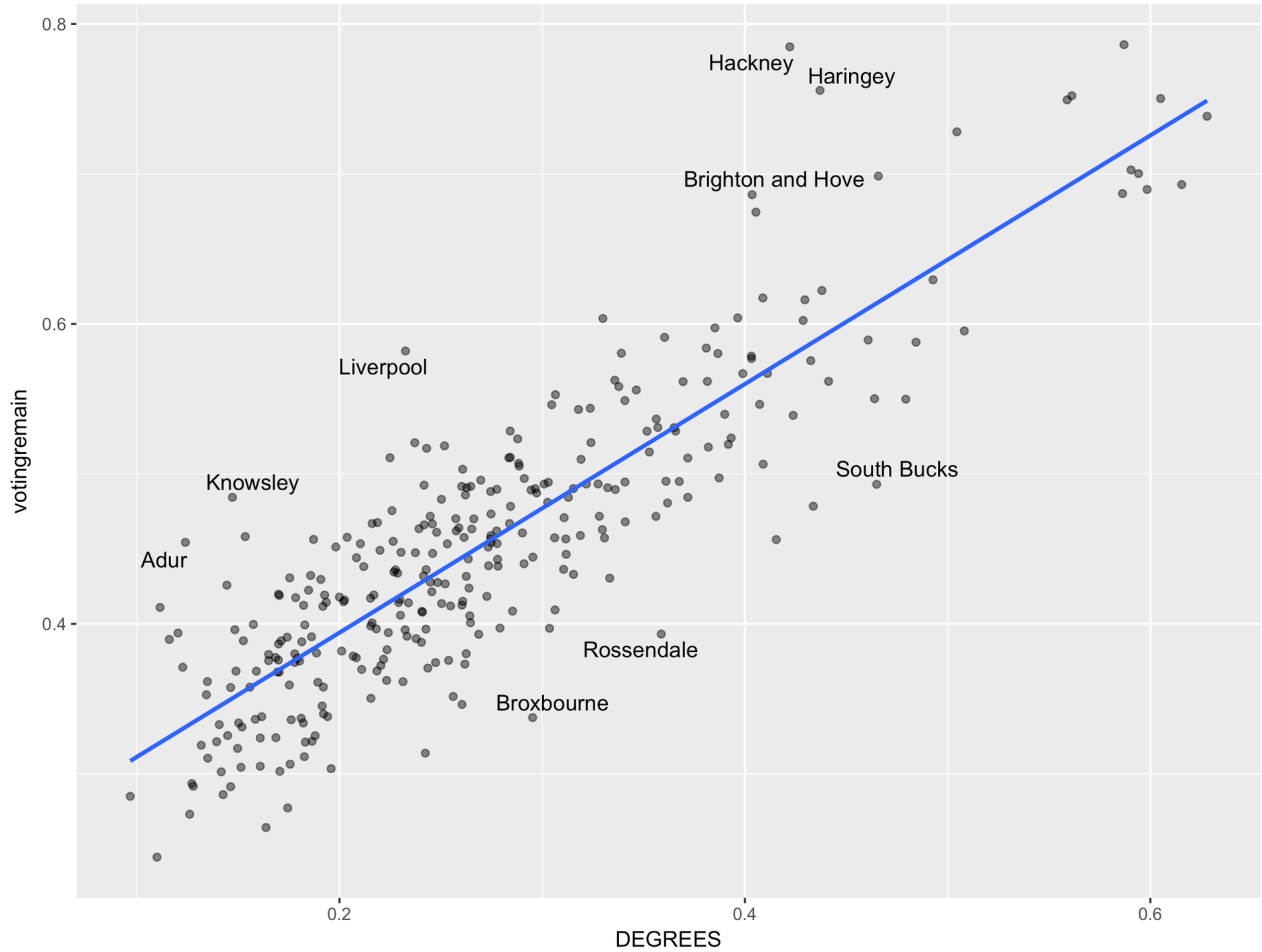


Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Social Analytics & Visualisation
Sheffield, 13/6/2022

Sheffield Methods Institute.

Some graphics



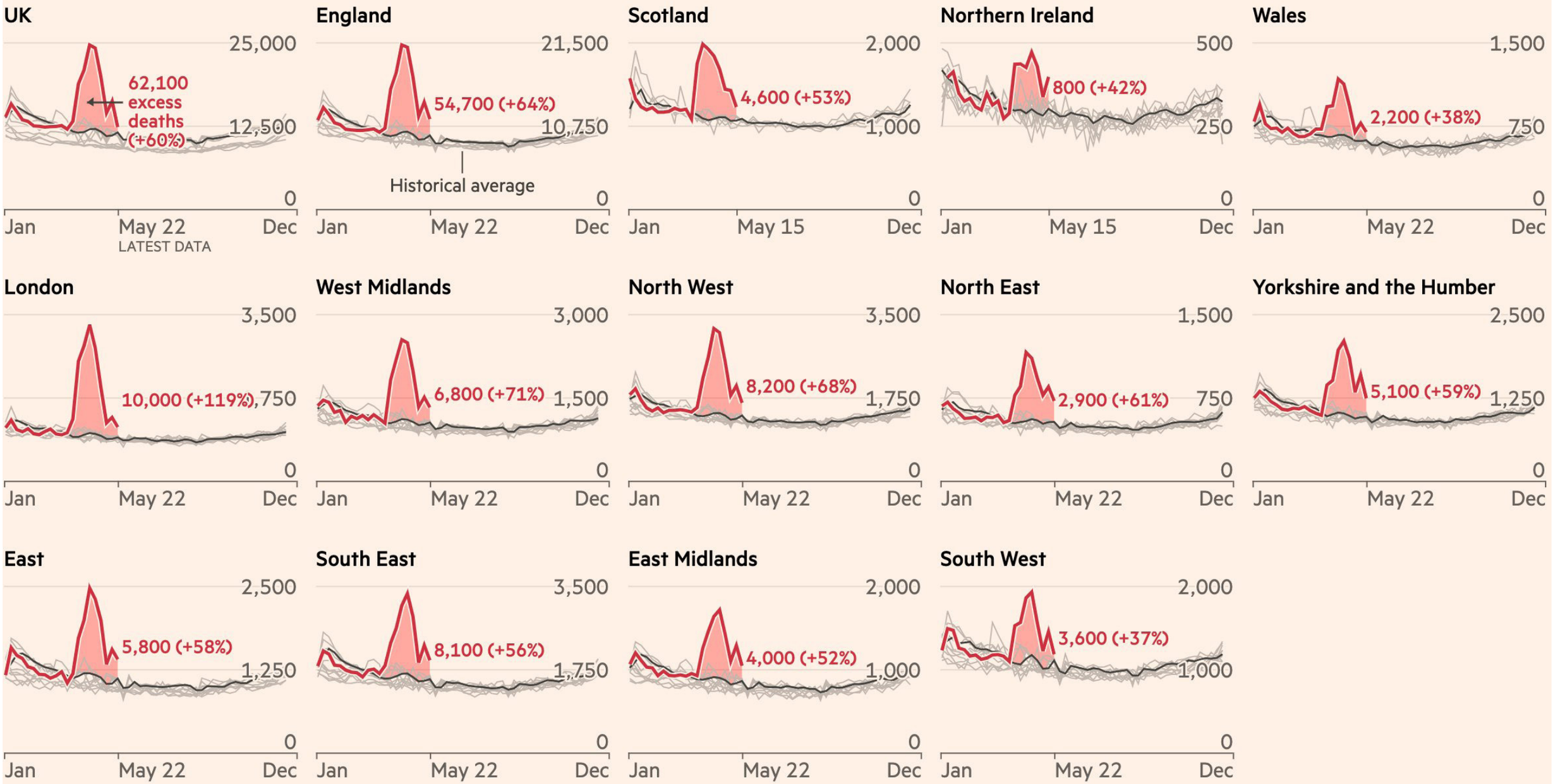
Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Social Analytics & Visualisation
Sheffield, 13/6/2022

Sheffield Methods Institute.

Every part of the UK has recorded high levels of excess deaths, with London the hardest hit

Number of deaths per week from all causes, 2020 vs recent years: Shading indicates total excess deaths during outbreak



Source: FT analysis of ONS mortality data. Data updated June 02
FT graphic: John Burn-Murdoch / @jburnmurdoch

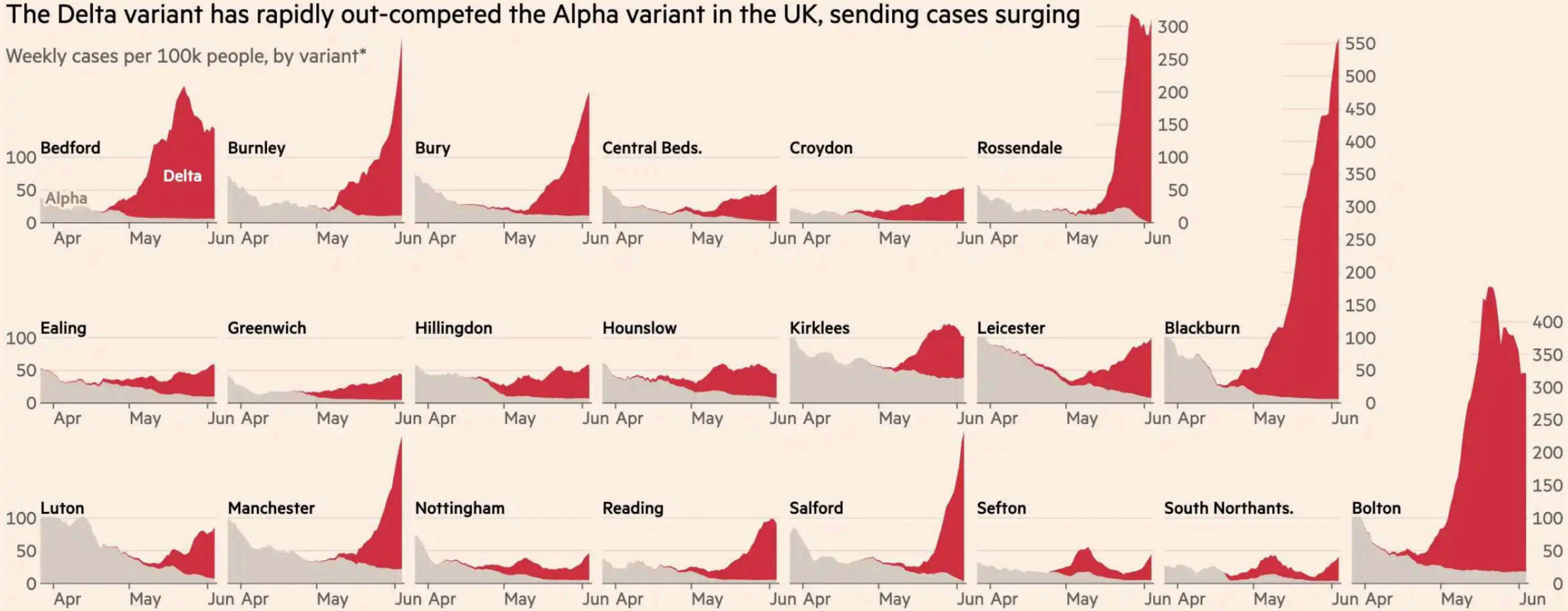
Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Social Analytics & Visualisation
Sheffield, 13/6/2022

Sheffield Methods Institute.

The Delta variant has rapidly out-competed the Alpha variant in the UK, sending cases surging

Weekly cases per 100k people, by variant*



*Based on applying proportions of sequenced samples to total cases
Sources: FT analysis of data from the Sanger Institute and UK government Covid-19 dashboard
© FT

Task

Please list as many types of graph as you can think of

- preferably by name – eg “x chart”, “y graph”, etc
- if you can think of different types of graph, but aren’t sure what they’re called, that’s also fine

What did you end up with?

Let's draw some graphs

Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Sheffield Methods Institute.

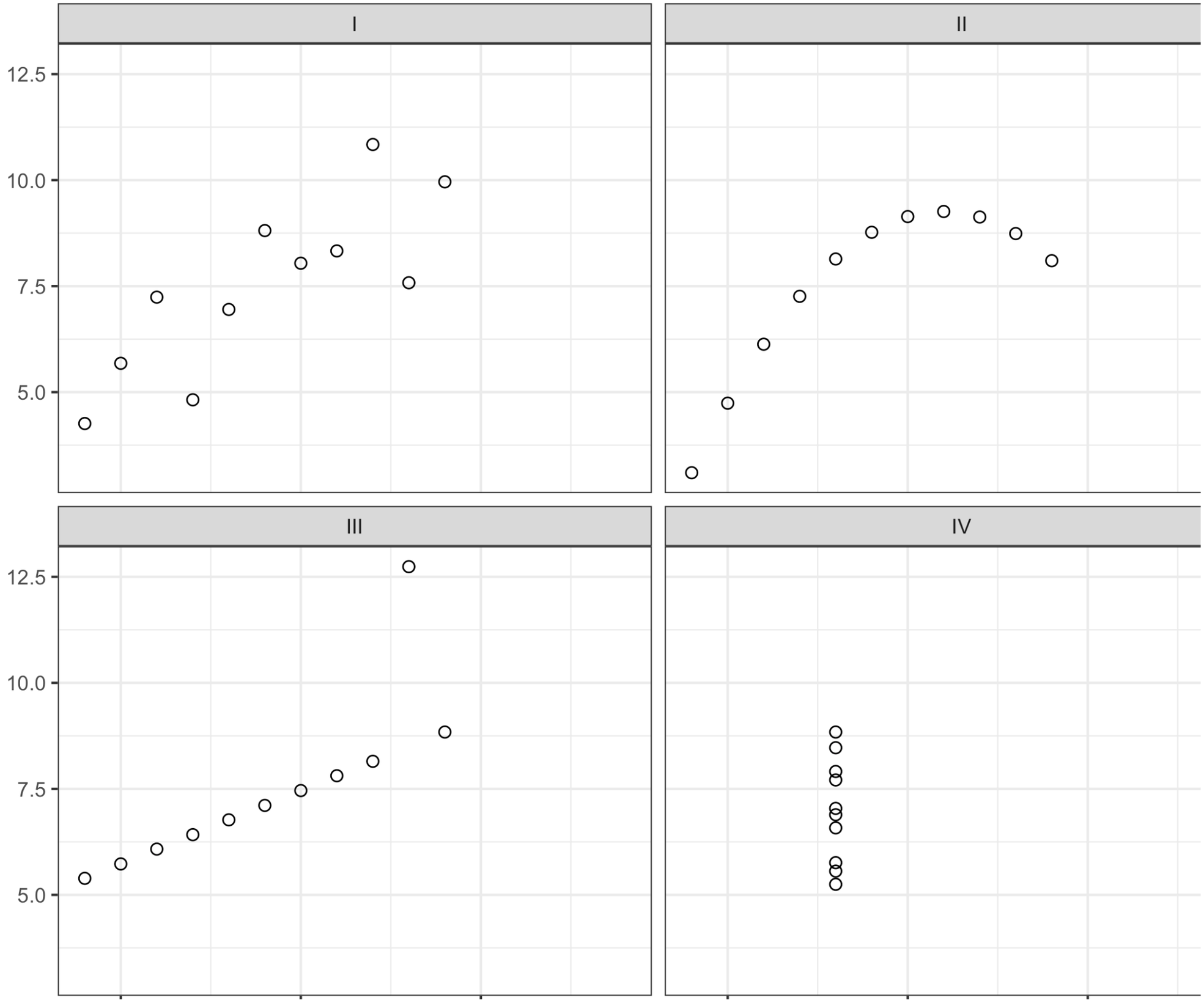
Let's take a step back

What's the point of visualising data?

- Tables exist
- Summary statistics exist
- There's only so many pages available in the world

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

N = 11
mean of X's = 9.0
mean of Y's = 7.5
equation of regression line: Y = 3 + 0.5X
standard error of estimate of slope = 0.118
t = 4.24
sum of squares $X - \bar{X} = 110.0$
regression sum of squares = 27.50
residual sum of squares of Y = 13.75
correlation coefficient = .82
$r^2 = .67$



Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

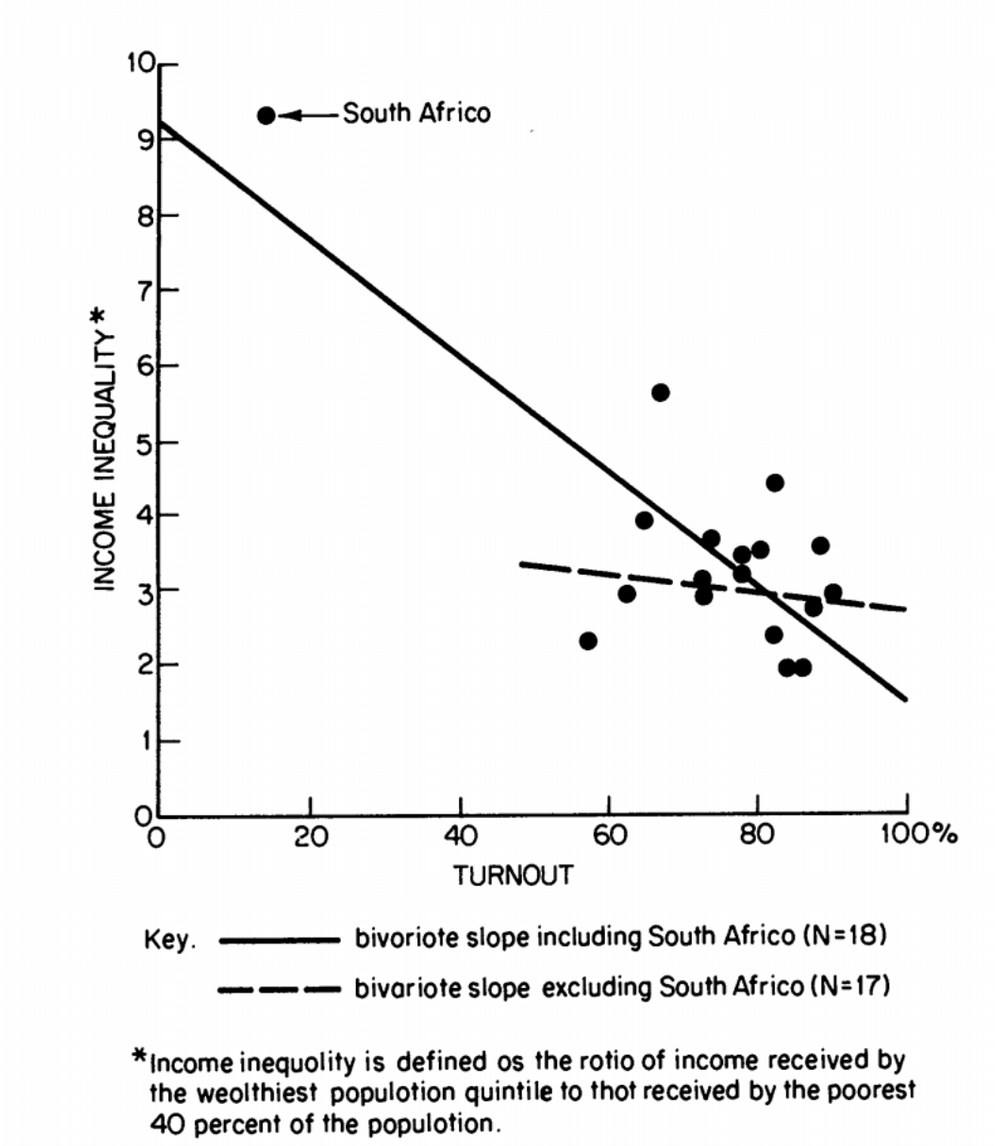
Income inequality & turnout

What do you think is the relationship between income inequality and turnout?

Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Sheffield Methods Institute.

Income inequality and turnout



Jackman, R. M. (1980). The impact of outliers on income inequality. American Sociological Review, 45, 344–347.

Let's pause

Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Sheffield Methods Institute.

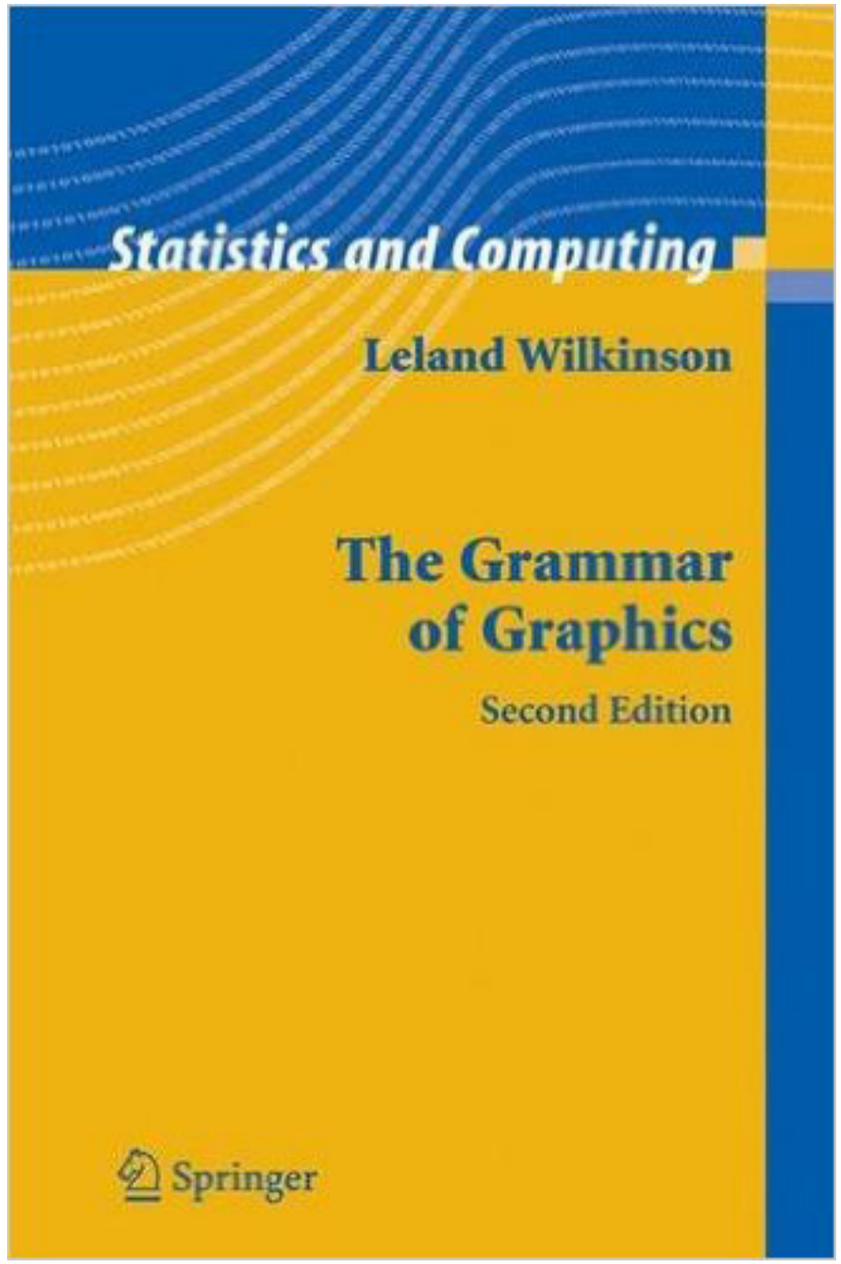
The grammar of graphics

Mainly associated with Leland Wilkinson

- and developed by Hadley Wickham

Why am I talking about it?

- Framework for thinking about any graphic
- Heavily used in statistical computing



The grammar of graphics

In the event I don't talk through this coming section in full, please read...

- Wickham, H. (2010) A layered grammar of graphics. Journal of Computational and Graphical Statistics 19(1): 3–28.
- Alternatively (or in addition!): ggplot2-book.org

Getting started

Data

- Might be a vector, might be a matrix (might be a point)

A	B	C	D
2	3	4	a
1	2	1	a
4	5	15	b
9	10	80	b

this example from Wickham, H. (2012) A Layered Grammar of Graphics. *Journal of Computational and Graphical Statistics* 19(1) 3–28.

Turning it into aesthetics

Data

- These reflect the things we want to see on the page

x	y	Shape
2	4	a
1	1	a
4	15	b
9	80	b

this example from Wickham, H. (2012) A Layered Grammar of Graphics. *Journal of Computational and Graphical Statistics* 19(1) 3–28.

Adding position

Data

- These reflect the things we want to see on the page

x	y	Shape
25	11	Circle
0	0	Circle
75	53	Square
200	300	Square

this example from Wickham, H. (2012) A Layered Grammar of Graphics. *Journal of Computational and Graphical Statistics* 19(1) 3–28.

Combining (3) graphic objects

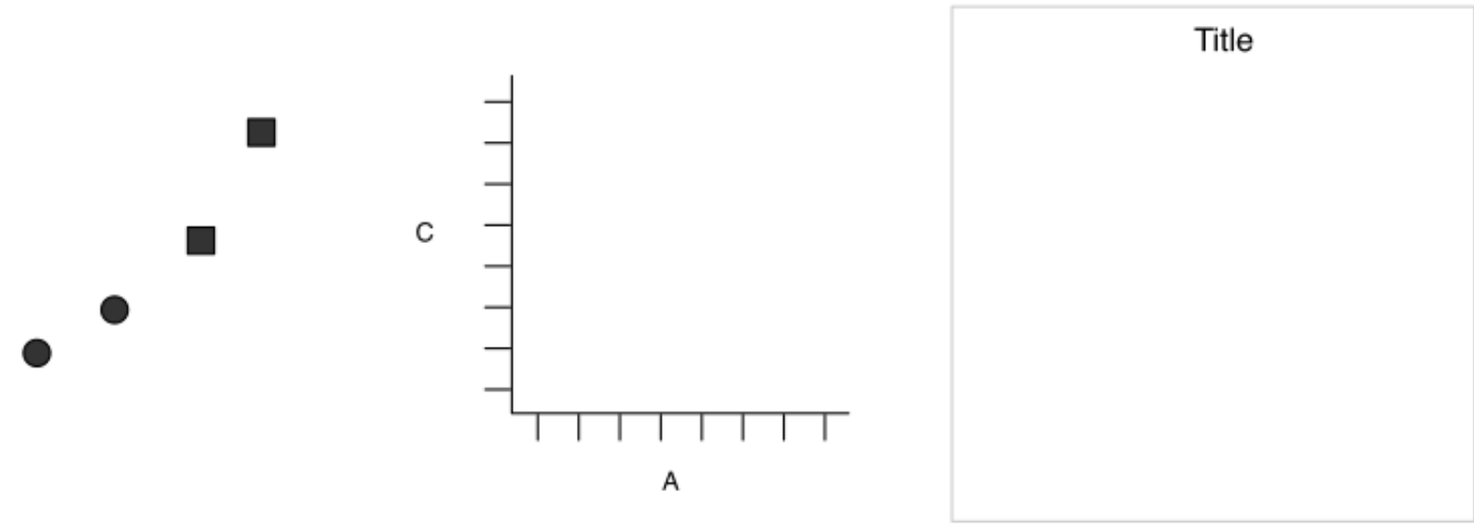


Figure 1. Graphics objects produced by (from left to right): geometric objects, scales and coordinate system, plot annotations.

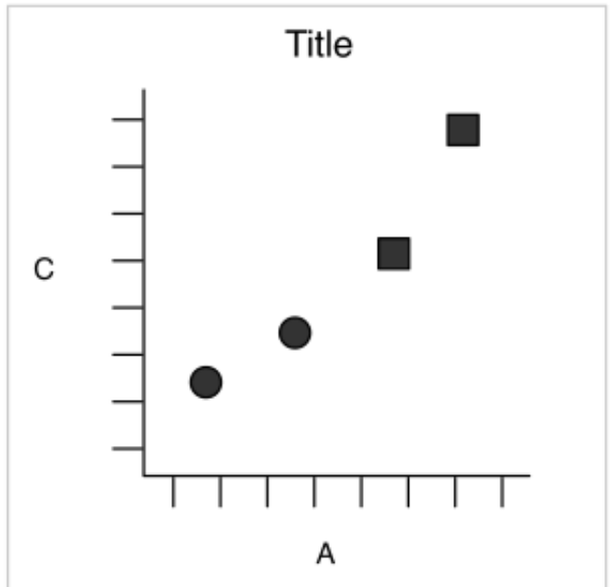


Figure 2. The final graphic, produced by combining the pieces in Figure 1.

this example from Wickham, H. (2012) A Layered Grammar of Graphics. *Journal of Computational and Graphical Statistics* 19(1) 3–28.

Adding facets

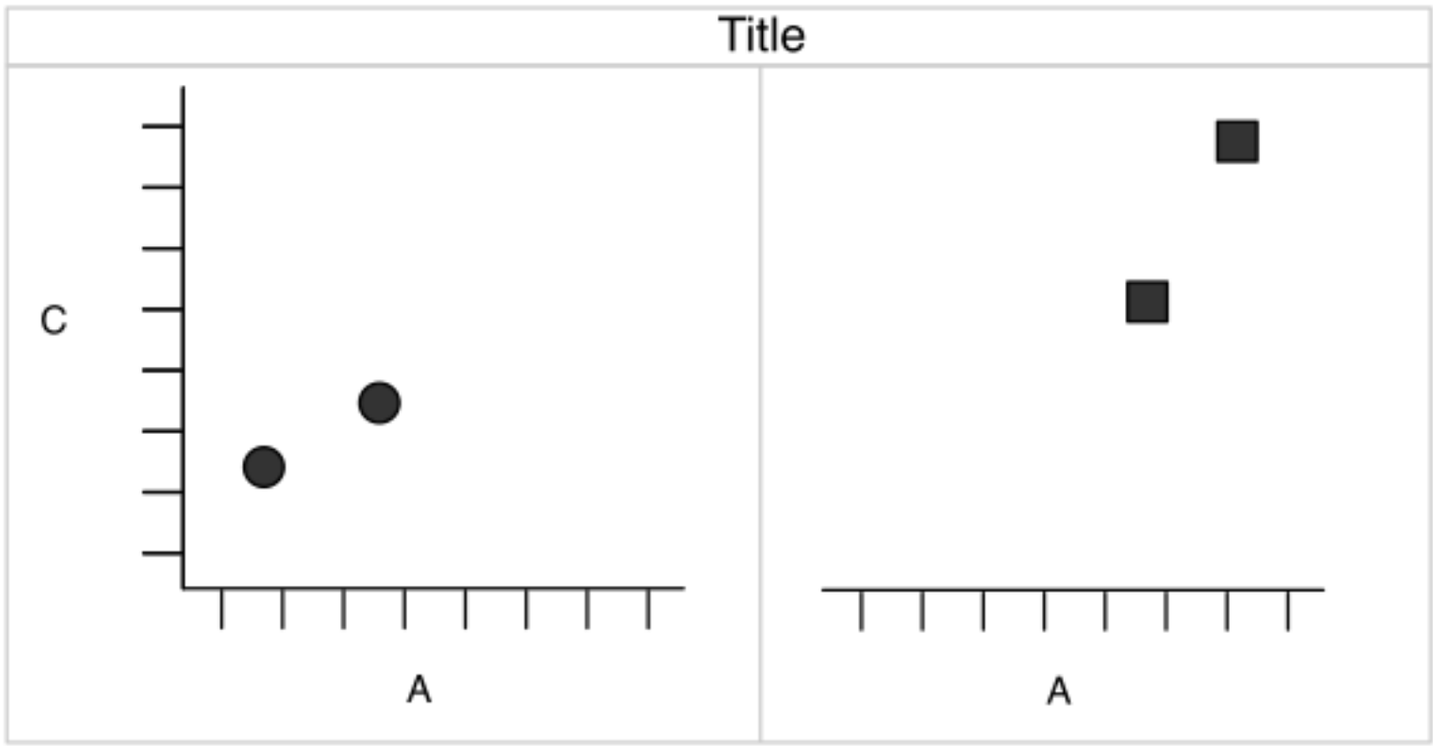


Figure 3. A more complicated plot, faceted by variable *D*. Here the faceting uses the same variable that is mapped to shape so that there is some redundancy in our visual representation. This allows us to easily see how the data have been broken into panels.

this example from Wickham, H. (2012) A Layered Grammar of Graphics. *Journal of Computational and Graphical Statistics* 19(1) 3–28.

What have we seen?

- Data
- Aesthetic mappings
- Scales
- Facets
- Statistical transformations
- Coordinates

The components of a plot

- Data
- Mappings from variables to aesthetics
- Layers
  - where a layer has one geometric object, one statistical transformation, one position adjustment (+ maybe one dataset and set of aesthetic mappings)
- A scale for each aesthetic mapping
- A coordinate system
- Facet specification

Layers

"One geometric object"

- Line
- Ribbon
- Point
- Letter
- Box
- Polygon
- etc

Layers

"One statistical transformation"

- Identity
- Bin
- Boxplot
- Smooth
- etc

Layers

One position adjustment

- (0,0)
- jitter
- dodge (side-by-side)

Scales

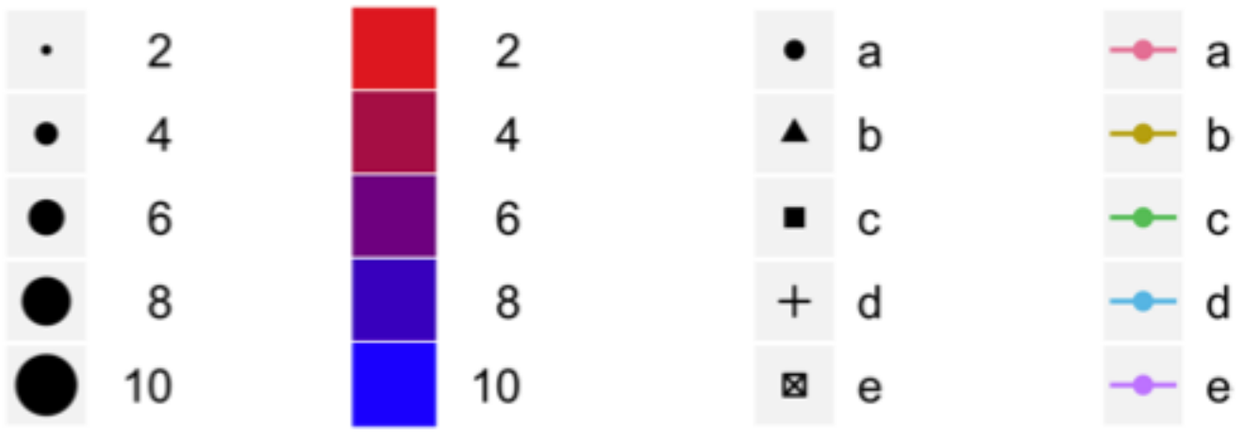


Figure 7. Examples of legends from four different scales. From left to right: continuous variable mapped to size and color, discrete variable mapped to shape and color. The legend automatically responds to the geoms used in the plot, from left to right: points, tiles, points, points and lines.

A coordinate system

Probably Cartesian

- (x, y)

but might not be

- log, semi-log scales
- polar
- geographic projections
- etc

Facet specification

Might be "don't facet"

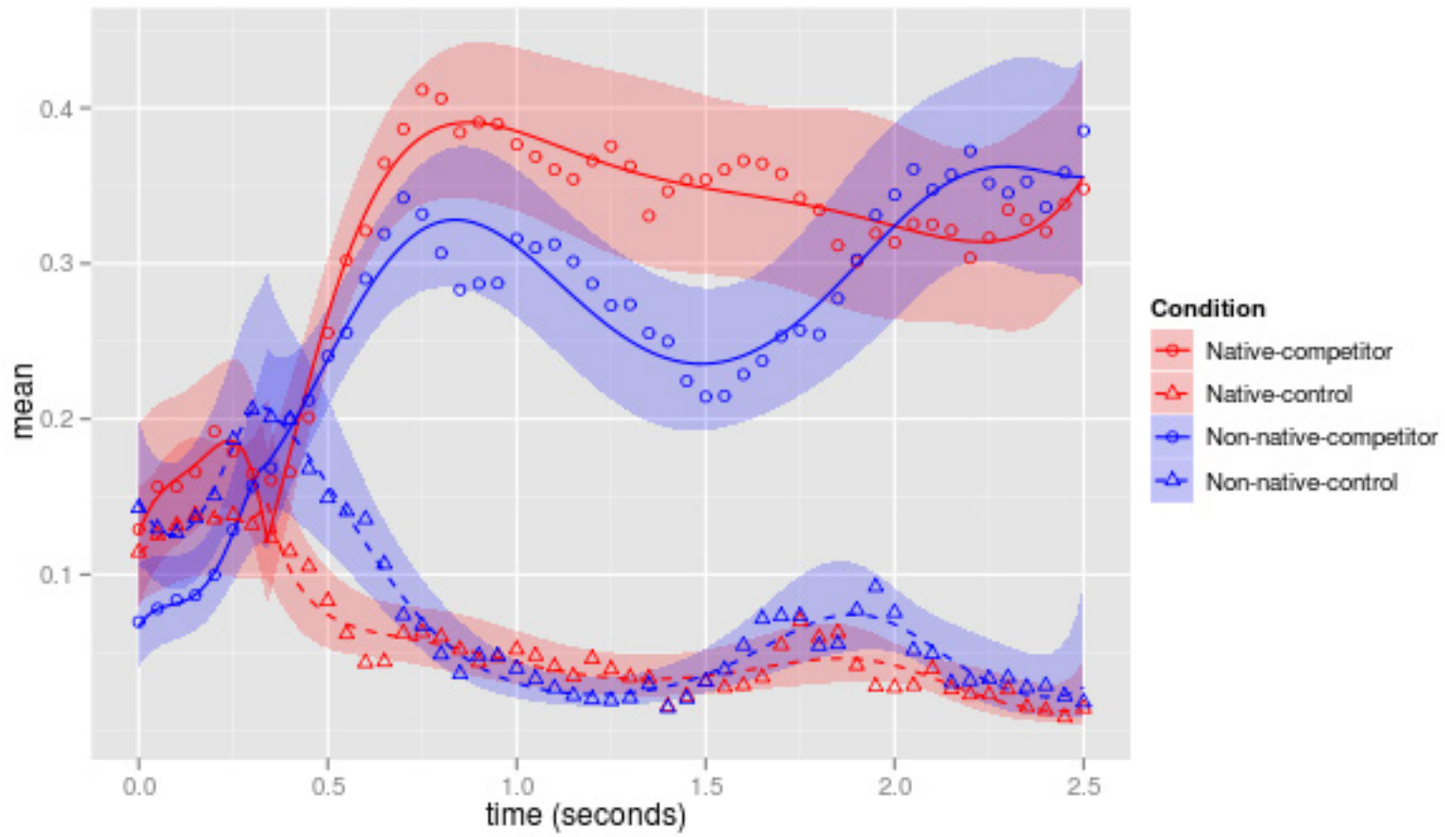
- but might include which variables should be used to split up the data
- and how they should be arranged

A *layered* grammar of graphics

Not limited to a single layer

- At least, you're likely to combine a geometric object, a scale and coordinate system, and a set of plot annotations
- But you can include a number of geometric objects, etc

A layered grammar of graphics

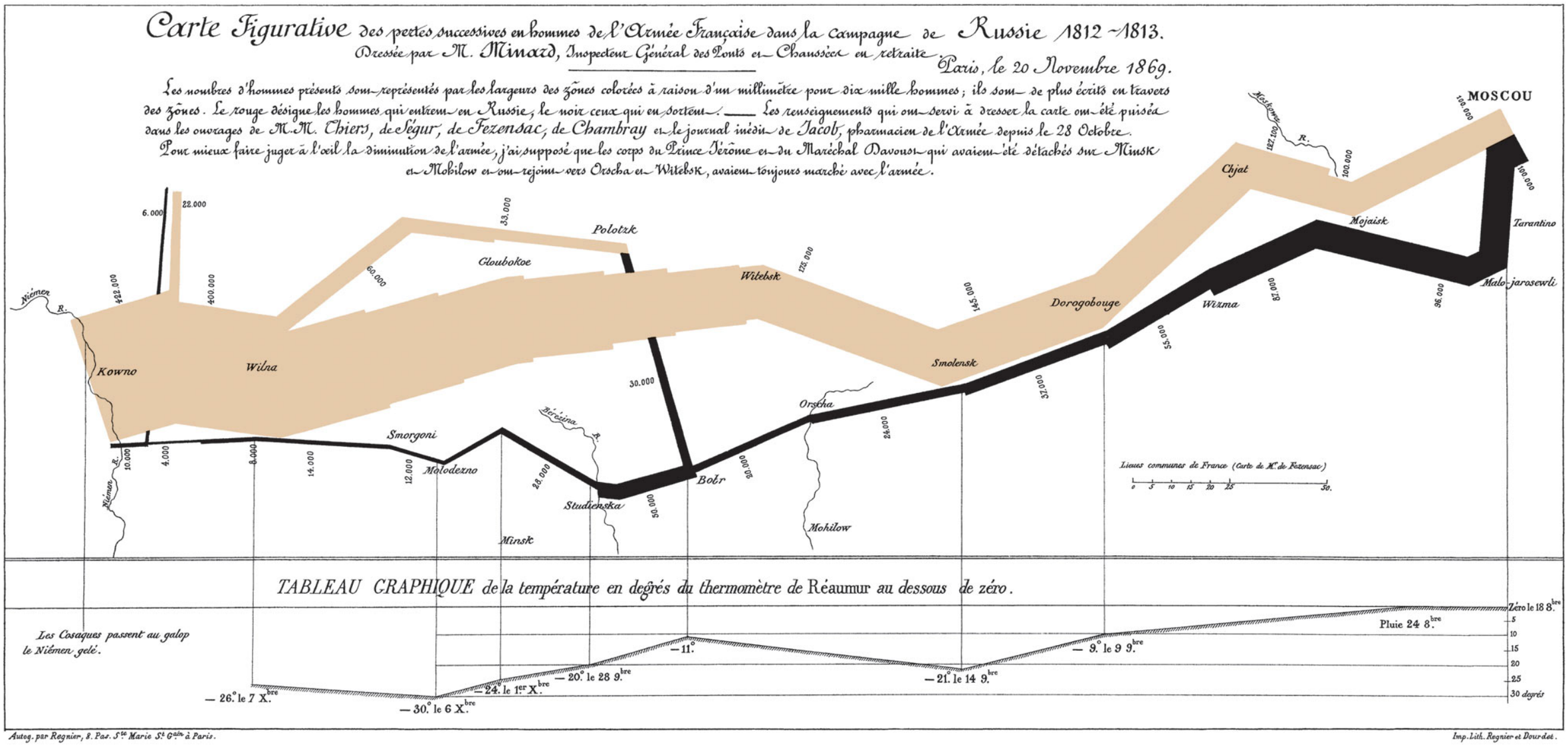


Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Social Analytics & Visualisation
Sheffield, 13/6/2022

Sheffield Methods Institute.

A layered grammar of graphics



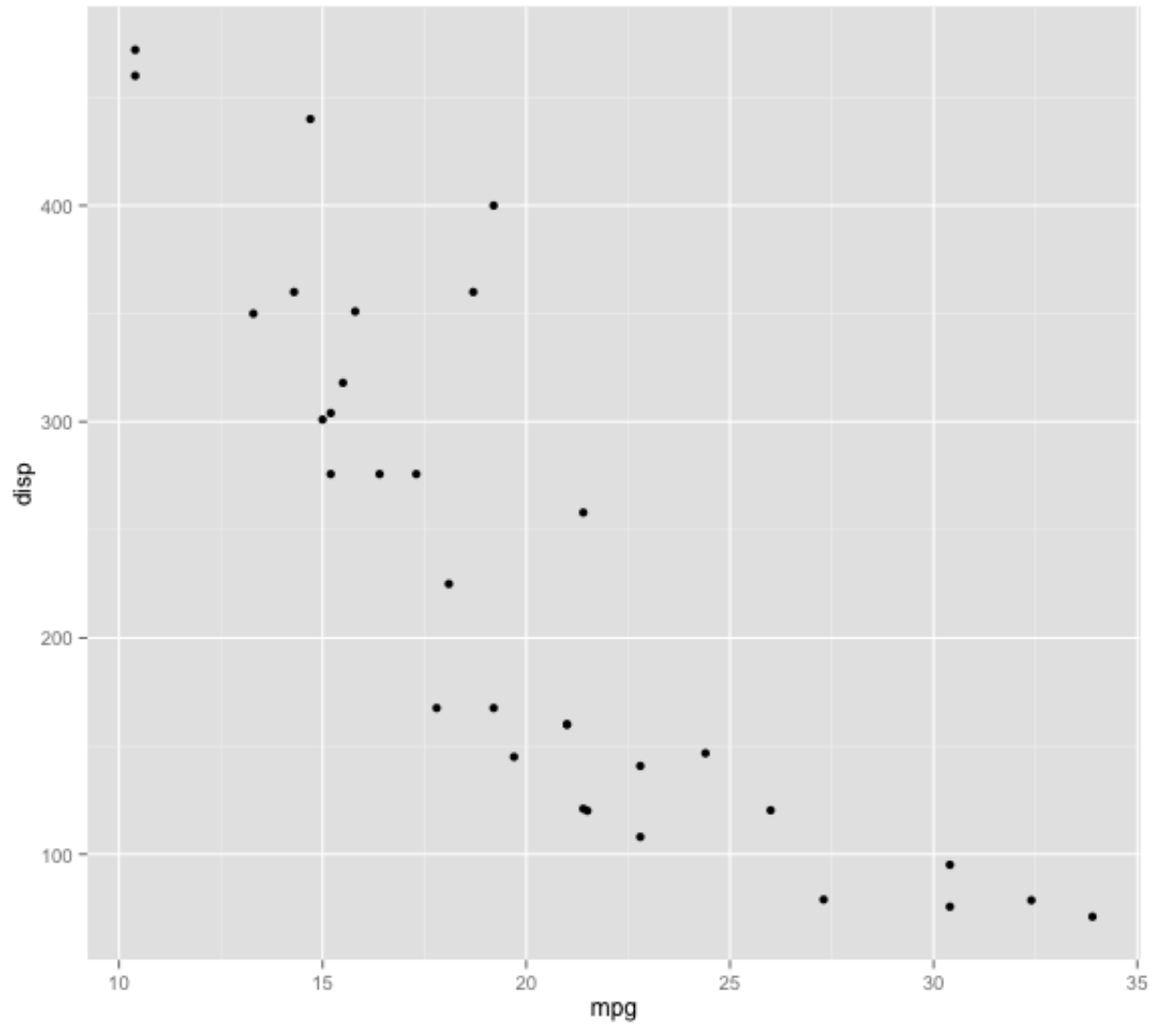
Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Social Analytics & Visualisation
Sheffield, 13/6/2022

Sheffield Methods Institute.

Exercise

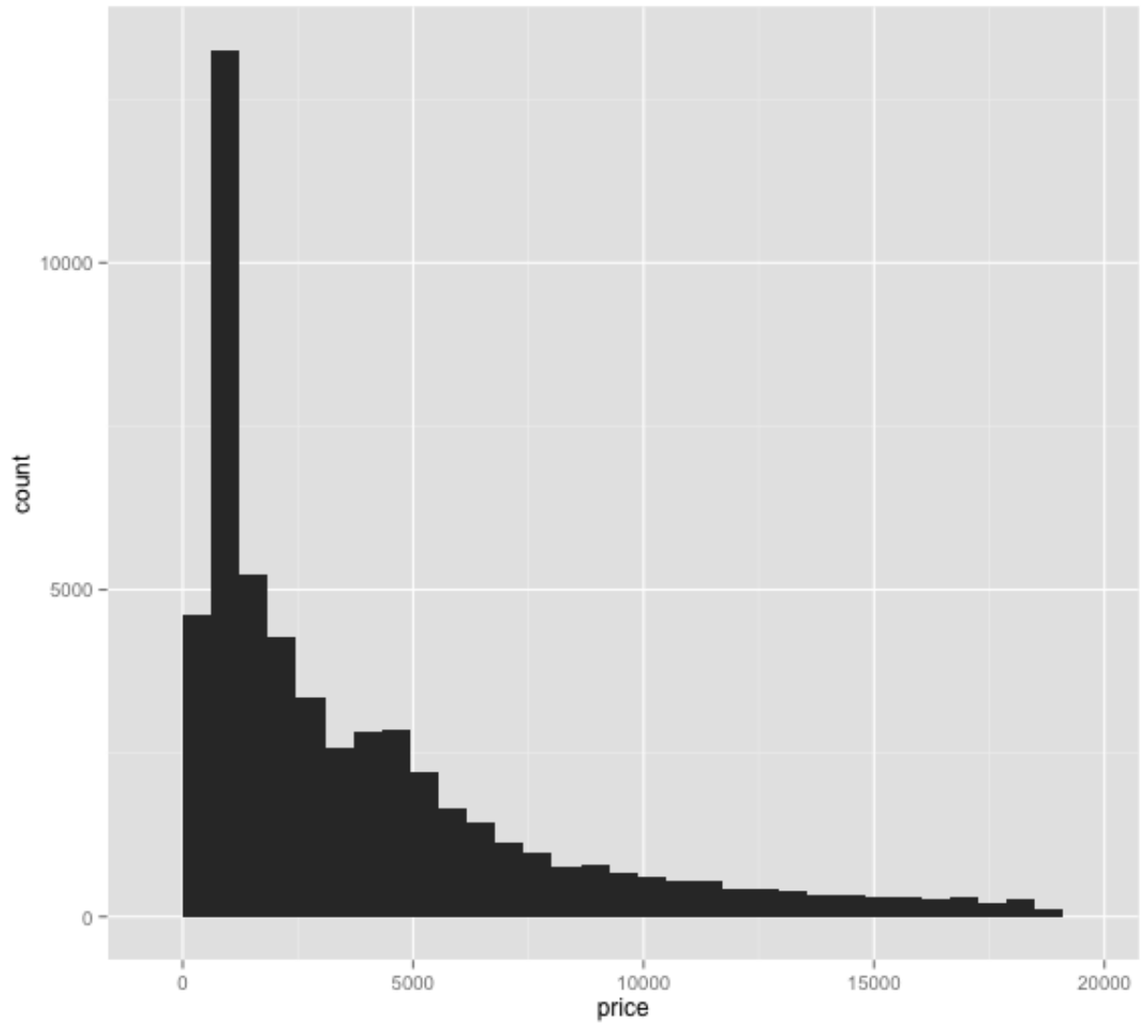
What are the geoms, stats, scales, and coordinates here?



Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Exercise

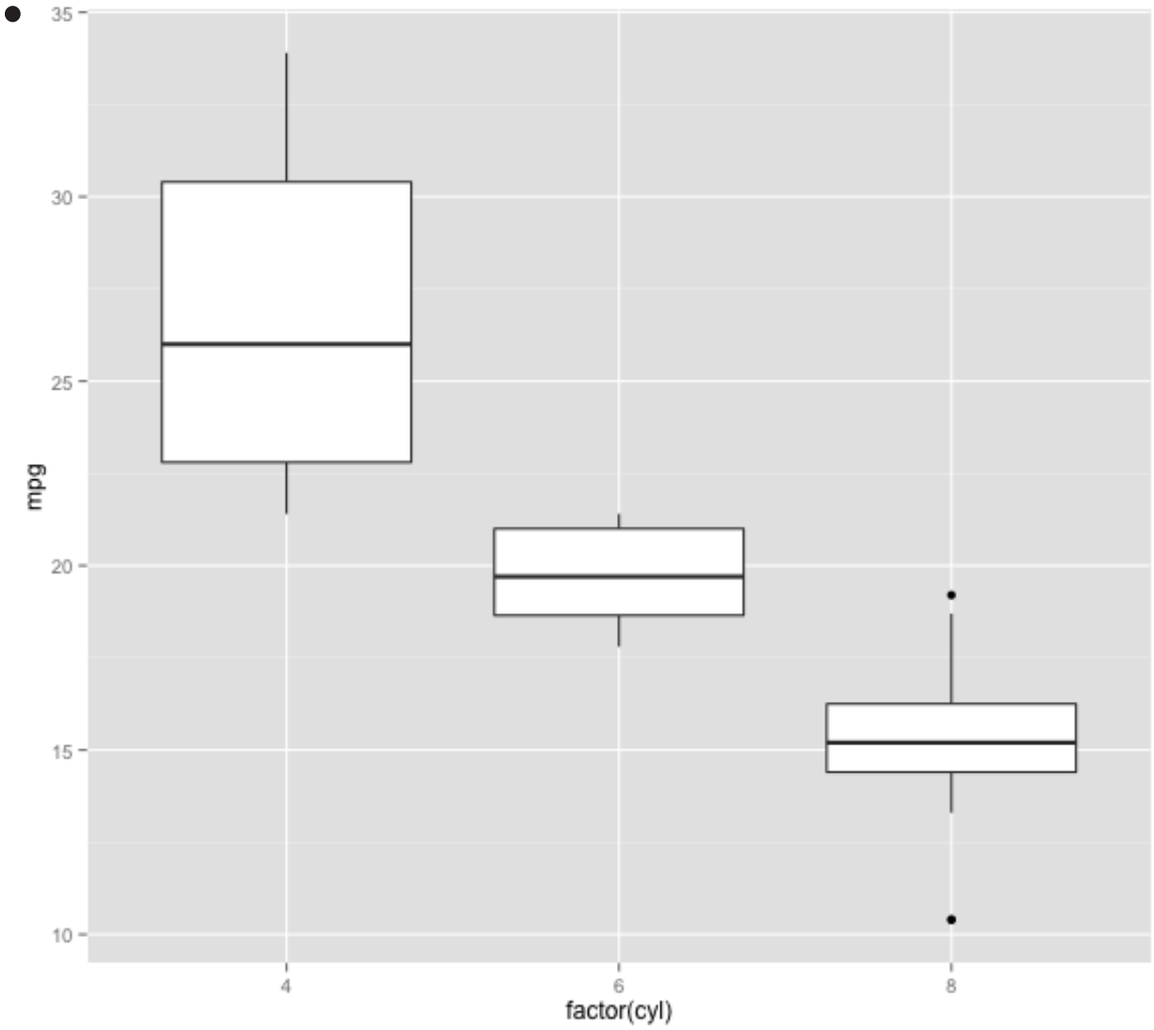
What are the geoms, stats, scales, and coordinates here?



Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

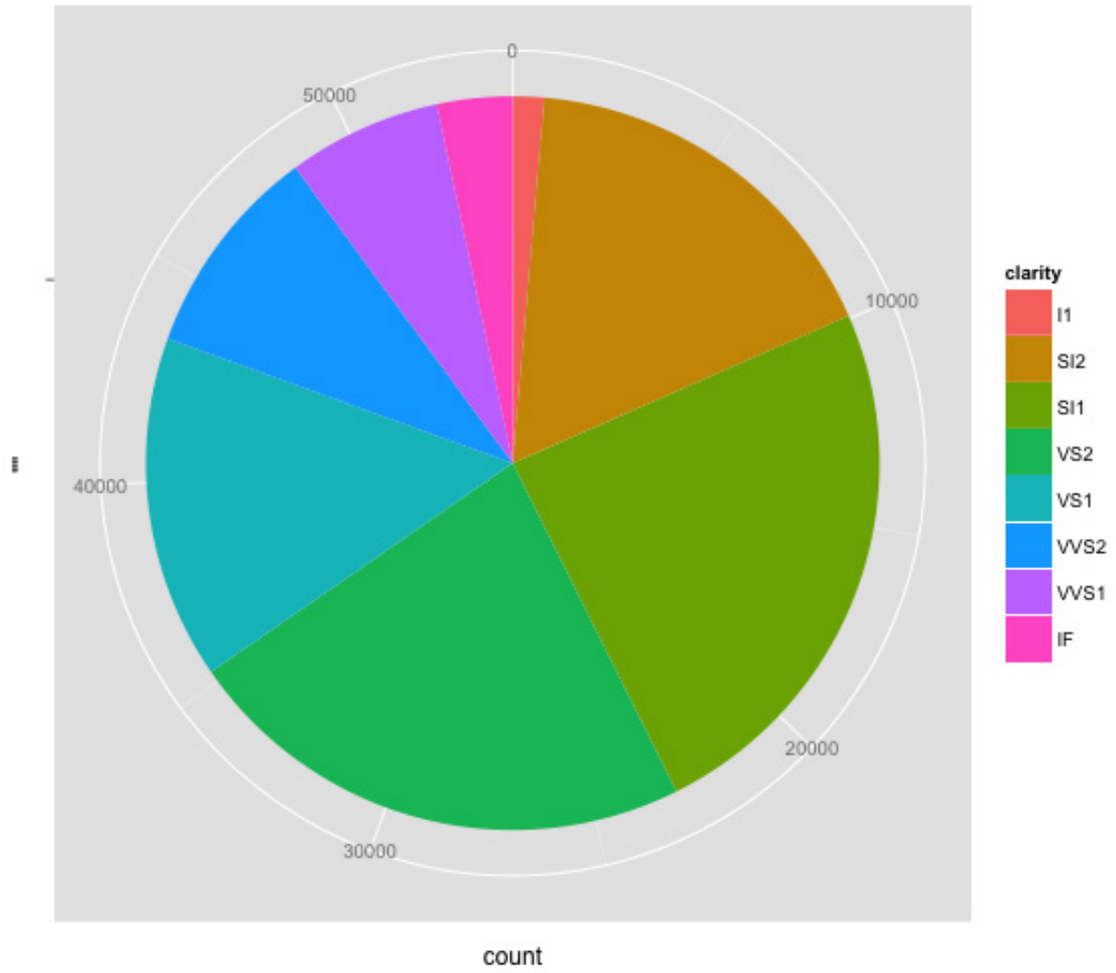
Exercise

What are the geoms, stats, scales, and coordinates here?



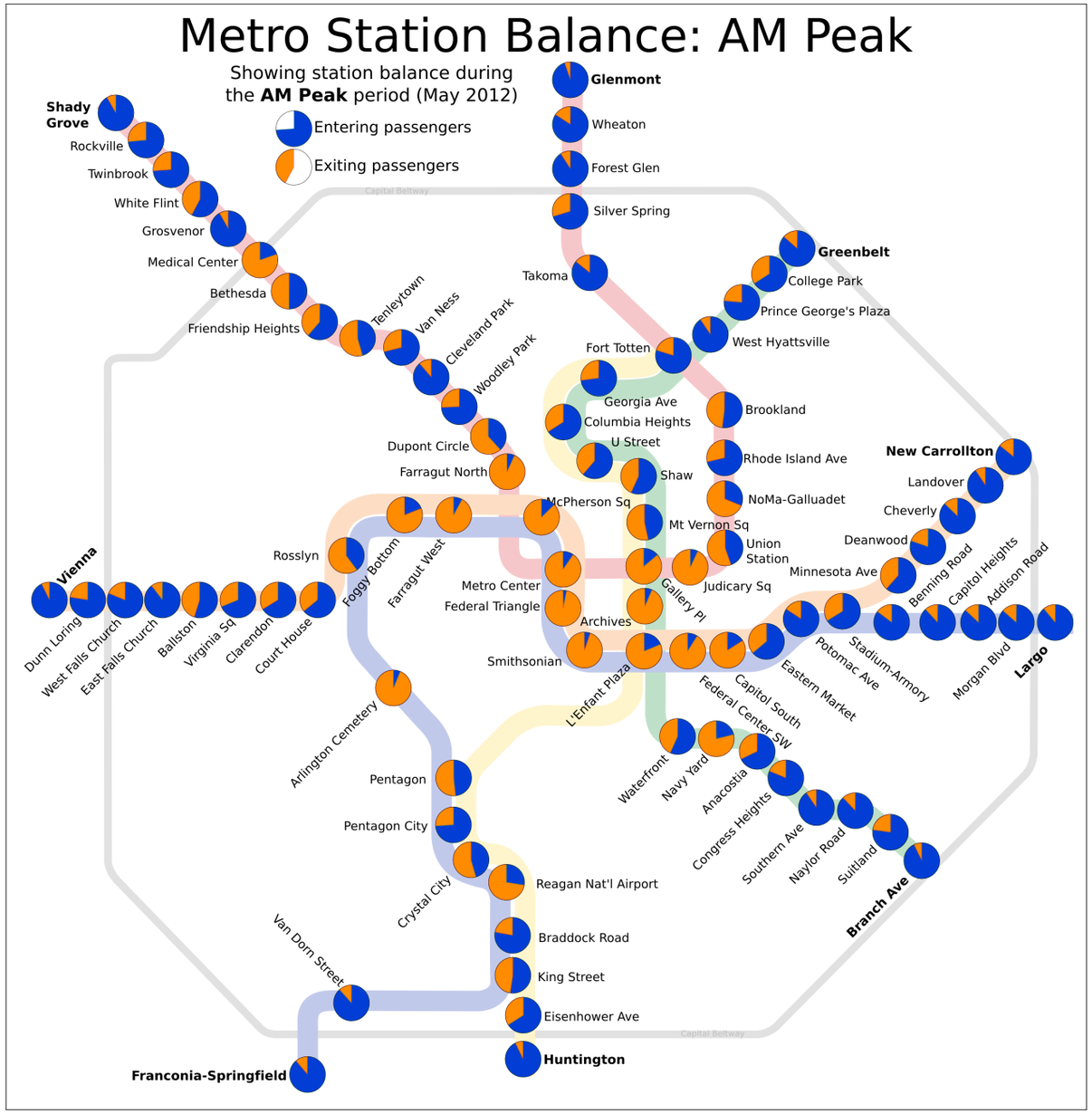
Exercise

What are the geoms, stats, scales, and coordinates here?



Exercise

What are the geoms, stats, scales, and coordinates here?



Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Social Analytics & Visualisation
Sheffield, 13/6/2022

Sheffield Methods Institute.

ggplot syntax

ggplot syntax is entirely built on different elements of the grammar of graphics

- commands can get really long
- and difficult to follow
- but if you can immerse yourself in the grammar of graphics, this will get more straightforward

ggplot2 syntax

```
ggplot(data = data) +
                aes(x = variable_1,
                    y = variable_2,
                    colour = variable_3...) +
    geom_something()
```

- or

```
ggplot(data) +
        aes(variable_1, variable_2,
                colour = variable_3) +
    geom_something()
```

But first...

What do we mean by "data"?

- Normally we mean individual-level records about a range of different characteristics (aka variables)
- Each row a case, each column a variable
- (cases don't have to be people: they can be organisations, countries, dogs, etc)

But first...

What do we mean by "variable"?

- Anything which can hold different values
  - Year
  - Revenue
  - Country
  - Whether someone won an Oscar
  - % of lines spoken by men
  - Number of words spoken by men
  - How left-wing

So we want to end up with a table

	% voting Republican	% White	% College educated	Governor
County A				
County B				
County C				
County D				
etc				

Then what?

This is where aesthetic mappings come in

- % voting Republican on the x-axis
- % White on the y-axis
- Governor in colour (or shape, or...)

And adding geometric objects

- points, bars, pies, histograms, density curves, lines, etc, etc

The major exception

Networks

- (for discussion later this week)

The most annoying bit

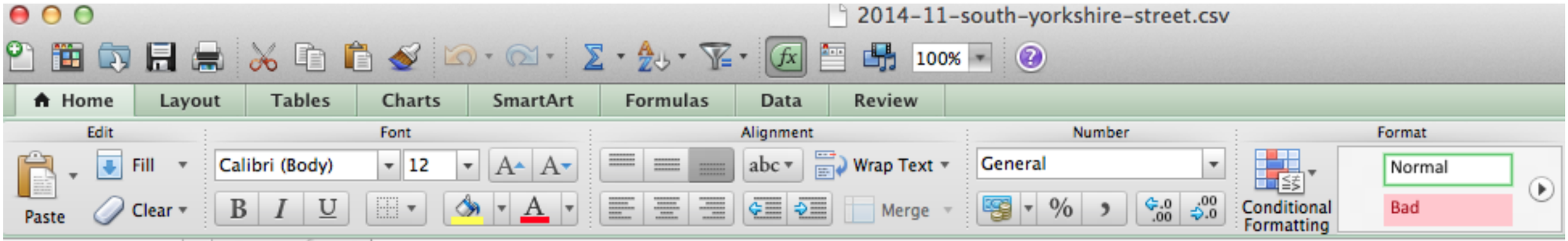
The standard ratio in data viz:

- 80% of time spent on data cleaning
- 20% of time spent on actually visualising

"Tidy datasets are all alike, but every messy dataset is messy in its own way"

Then what?

Ideally your data looks something like this



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Crime ID	Month	Reported by	Falls within	Longitude	Latitude	Location	LSOA code	LSOA name	Crime type	Last outcom	Context			
2		2014-11	South Yorksh	South Yorksh	-1.452703	53.597074	On or near C	E01007434	Barnsley 001	Anti-social behaviour					
3		2014-11	South Yorksh	South Yorksh	-1.450357	53.600838	On or near G	E01007434	Barnsley 001	Anti-social behaviour					
4		2014-11	South Yorksh	South Yorksh	-1.452703	53.597074	On or near C	E01007434	Barnsley 001	Anti-social behaviour					
5		2014-11	South Yorksh	South Yorksh	-1.450357	53.600838	On or near G	E01007434	Barnsley 001	Anti-social behaviour					
6		2014-11	South Yorksh	South Yorksh	-1.451664	53.599163	On or near Ja	E01007434	Barnsley 001	Anti-social behaviour					
7		2014-11	South Yorksh	South Yorksh	-1.458372	53.595643	On or near C	E01007434	Barnsley 001	Anti-social behaviour					
8	b2d0f58115c	2014-11	South Yorksh	South Yorksh	-1.445925	53.603505	On or near W	E01007434	Barnsley 001	Burglary	Under investigation				
9	603eee01a1	2014-11	South Yorksh	South Yorksh	-1.450844	53.60057	On or near Tl	E01007434	Barnsley 001	Criminal dam	Investigation complete; no suspect identified				
10	bb504e2d5e	2014-11	South Yorksh	South Yorksh	-1.454931	53.597713	On or near N	E01007434	Barnsley 001	Other theft	Investigation complete; no suspect identified				
11	afd749320a4	2014-11	South Yorksh	South Yorksh	-1.458007	53.598212	On or near H	E01007434	Barnsley 001	Shoplifting	Under investigation				
12	fea07ab23d3	2014-11	South Yorksh	South Yorksh	-1.449981	53.603074	On or near A	E01007434	Barnsley 001	Vehicle crime	Investigation complete; no suspect identified				
13	dcb0ec44901	2014-11	South Yorksh	South Yorksh	-1.445925	53.603505	On or near W	E01007434	Barnsley 001	Violence and	Unable to prosecute suspect				
14		2014-11	South Yorksh	South Yorksh	-1.46656	53.59817	On or near W	E01007435	Barnsley 001	Anti-social behaviour					
15		2014-11	South Yorksh	South Yorksh	-1.461821	53.596567	On or near D	E01007435	Barnsley 001	Anti-social behaviour					
16		2014-11	South Yorksh	South Yorksh	-1.458974	53.594693	On or near W	E01007435	Barnsley 001	Anti-social behaviour					
17		2014-11	South Yorksh	South Yorksh	-1.460654	53.598026	On or near M	E01007435	Barnsley 001	Anti-social behaviour					

But it might look like this

Country c	Country name	Series code	Series name	SCALE	Decimals	2001	2002	2003	
TUN	Tunisia	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	3.64	3.68	3.56	
TUR	Turkey	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	..	12.66	12.83	1
TUV	Tuvalu	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	..	..	..	
TZA	Tanzania	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	..	..	..	
UGA	Uganda	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	..	..	..	
UKR	Ukraine	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	..	..	..	
UMC	Upper middle income	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	..	..	..	
URY	Uruguay	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	..	..	..	
USA	United States	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	..	..	..	
UZB	Uzbekistan	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	..	..	..	
VCT	St. Vincent and the Grenadines	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	..	..	..	
VEN	Venezuela, RB	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	..	..	..	
VIR	Virgin Islands (U.S.)	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	..	..	..	
VNM	Vietnam	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	..	..	..	
VUT	Vanuatu	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	..	..	..	
WBG	West Bank and Gaza	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	4.34	4.32	4.09	
WLD	World	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	..	..	..	
WSM	Samoa	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	..	..	..	
YEM	Yemen, Rep.	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	..	2.88	2.00	
ZAF	South Africa	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	..	..	..	
ZAR	Congo, Dem. Rep.	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	..	..	..	
ZMB	Zambia	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	..	..	..	
ZWE	Zimbabwe	AG.LND.IRIG.AG.ZS	Agricultural land under irrigation (% of total ag. la	0	1	..	..	..	
ABW	Aruba	AG.YLD.CREL.KG	Cereal yield (kg per hectare)	0	0	..	..	..	
ADO	Andorra	AG.YLD.CREL.KG	Cereal yield (kg per hectare)	0	0	..	..	..	
AFG	Afghanistan	AG.YLD.CREL.KG	Cereal yield (kg per hectare)	0	0	1006.60	1669.70	1458.00	133
AGO	Angola	AG.YLD.CREL.KG	Cereal yield (kg per hectare)	0	0	623.70	639.70	668.40	49
ALB	Albania	AG.YLD.CREL.KG	Cereal yield (kg per hectare)	0	0	3030.70	3291.80	3185.70	346
ARE	United Arab Emirates	AG.YLD.CREL.KG	Cereal yield (kg per hectare)	0	0	3533.30	3411.70	2500.00	200
ARG	Argentina	AG.YLD.CREL.KG	Cereal yield (kg per hectare)	0	0	3206.40	3240.70	3673.00	366
ARM	Armenia	AG.YLD.CREL.KG	Cereal yield (kg per hectare)	0	0	1902.60	2229.00	1589.00	232
ASM	American Samoa	AG.YLD.CREL.KG	Cereal yield (kg per hectare)	0	0	..	..	..	

Or this

A1 | ï»¿"Data Source"

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ï»¿"Data Sou	World Development Indicators											
2													
3	Last Updatec	18/09/2015											
4													
5	Country Nam	Country Cod	Indicator Na	Indicator Coc	1961	1962	1963	1964	1965	1966	1967	1968	1969
6	Aruba	ABW	Scientific anc	IP.JRN.ARTC.SC									
7	Andorra	AND	Scientific anc	IP.JRN.ARTC.SC									
8	Afghanistan	AFG	Scientific anc	IP.JRN.ARTC.SC									
9	Angola	AGO	Scientific anc	IP.JRN.ARTC.SC									
10	Albania	ALB	Scientific anc	IP.JRN.ARTC.SC									
11	Arab World	ARB	Scientific anc	IP.JRN.ARTC.SC									
12	United Arab	ARE	Scientific anc	IP.JRN.ARTC.SC									
13	Argentina	ARG	Scientific anc	IP.JRN.ARTC.SC									
14	Armenia	ARM	Scientific anc	IP.JRN.ARTC.SC									
15	American Sa	ASM	Scientific anc	IP.JRN.ARTC.SC									
16	Antigua and	ATG	Scientific anc	IP.JRN.ARTC.SC									
17	Australia	AUS	Scientific anc	IP.JRN.ARTC.SC									
18	Austria	AUT	Scientific anc	IP.JRN.ARTC.SC									
19	Azerbaijan	AZE	Scientific anc	IP.JRN.ARTC.SC									
20	Burundi	BDI	Scientific anc	IP.JRN.ARTC.SC									
21	Belgium	BEL	Scientific anc	IP.JRN.ARTC.SC									
22	Benin	BEN	Scientific anc	IP.JRN.ARTC.SC									
23	Burkina Faso	BFA	Scientific anc	IP.JRN.ARTC.SC									
24	Bangladesh	BGD	Scientific anc	IP.JRN.ARTC.SC									
25	Bulgaria	BGR	Scientific anc	IP.JRN.ARTC.SC									
26	Bahrain	BHR	Scientific anc	IP.JRN.ARTC.SC									
27	Bahamas, Th	BHS	Scientific anc	IP.JRN.ARTC.SC									
28	Bosnia and H	BIH	Scientific anc	IP.JRN.ARTC.SC									

Solving problems

Sometimes: just use Excel (or similar)

- Redundant info at the top

Sometimes: filter

- Often there's too much data

Sometimes: there's more to it than that

- A variable across multiple columns, a column with multiple variables, some cases in multiple rows...

Ideally: get your head round tidyr & dplyr

Let's draw some more graphs

Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Sheffield Methods Institute.

Who's the audience for visualisation?

Often it's you

- ideally it's often other people as well, but you're not purely visualising data for other people's benefit
- looking at your data is a good way to get a sense of what's going on, and to catch any problems before they get too serious

How do you choose?

Sometimes it's informed by variables

- if you've got exactly two continuous variables and nothing else, you'll probably draw a scatterplot

Often you need to make a decision

- and this is fundamentally an editorial decision: you can't just blame the data and say you had no alternative

Some typologies

(non-exhaustive)

- FT Visual Vocabulary
- Claus Wilke's Directory of Visualisations
- The typology in the middle of Andy Kirk's book
- datavizcatalogue.com
- Andrew Abela's chart chooser

They're generally organised according to what the point of your graph is

Something I'm not doing

What makes a graphic any good?

- It's fun and easy to flag bad graphics
- If we were here for several days, I'd probably do that, and break down what works and doesn't work
- (This also isn't an aesthetics course)

But

Healy's "What makes bad figures bad?" section draws a helpful distinction

- Bad taste
- Bad data
- Bad perception

Let's draw even more graphs

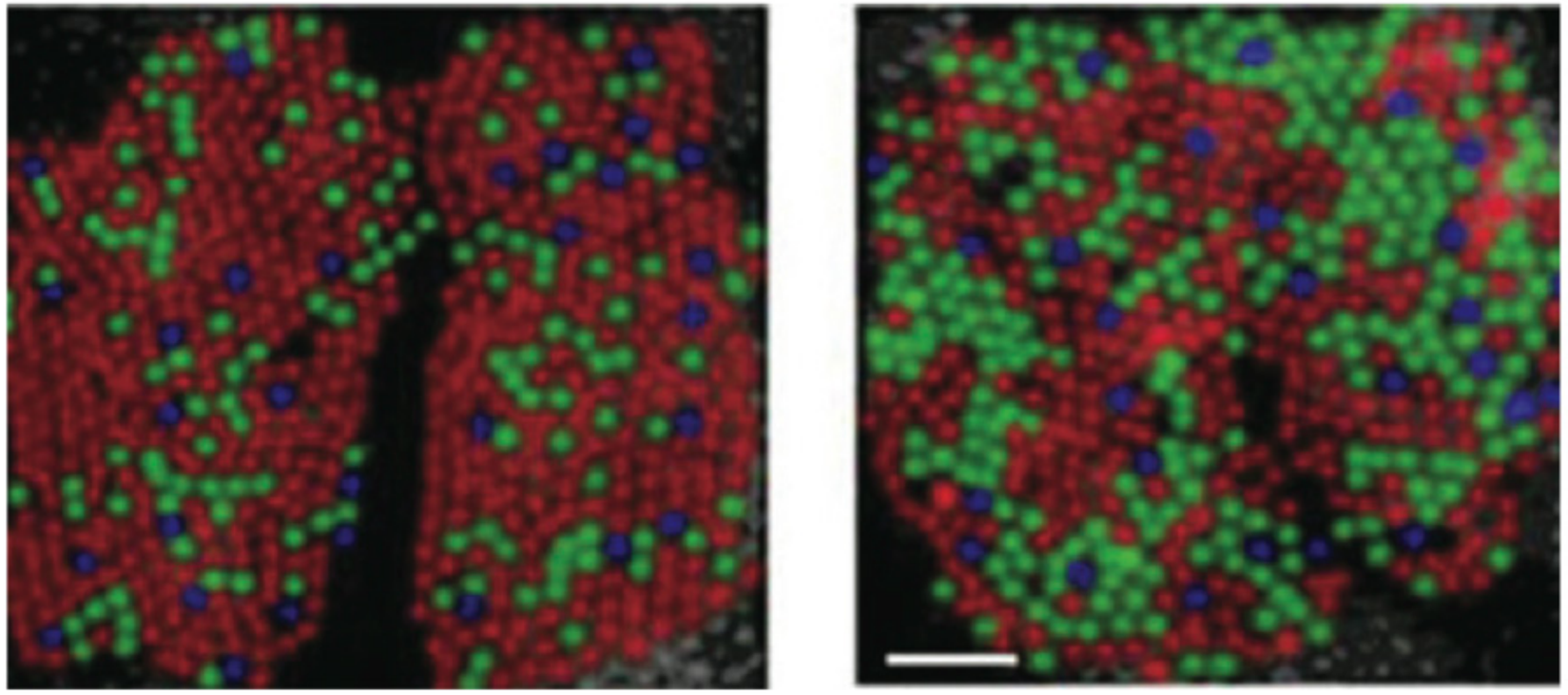
Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Sheffield Methods Institute.

Perception of colour

Two kinds of light receptors in the retina

- rods and cones
  - where rods are only really used in low light
  - while cones aren't evenly distributed



Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

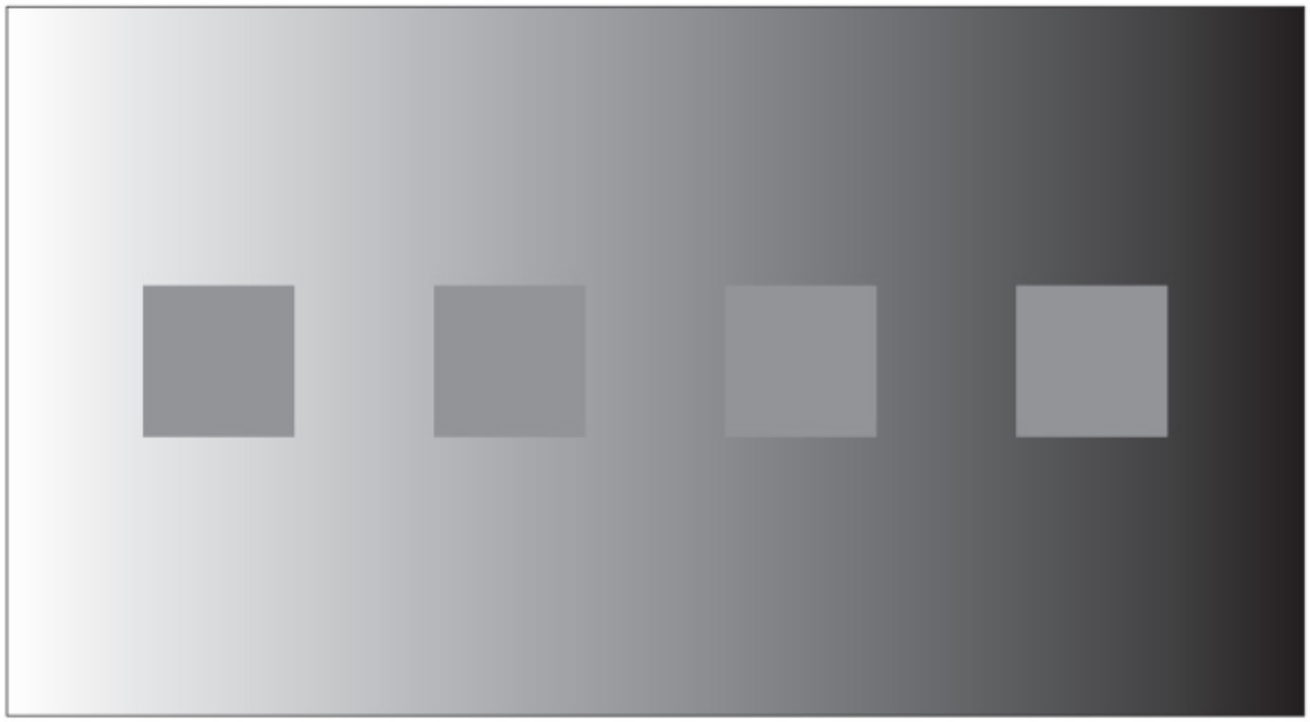
Rods and cones

Cones are sensitive to one of short-, medium-, and long-wavelengths; there are fewer long-wavelength sensitive cones

- The upshot of this is blue seems darker, and yellow seems lighter
- This is hard to read

- So's this

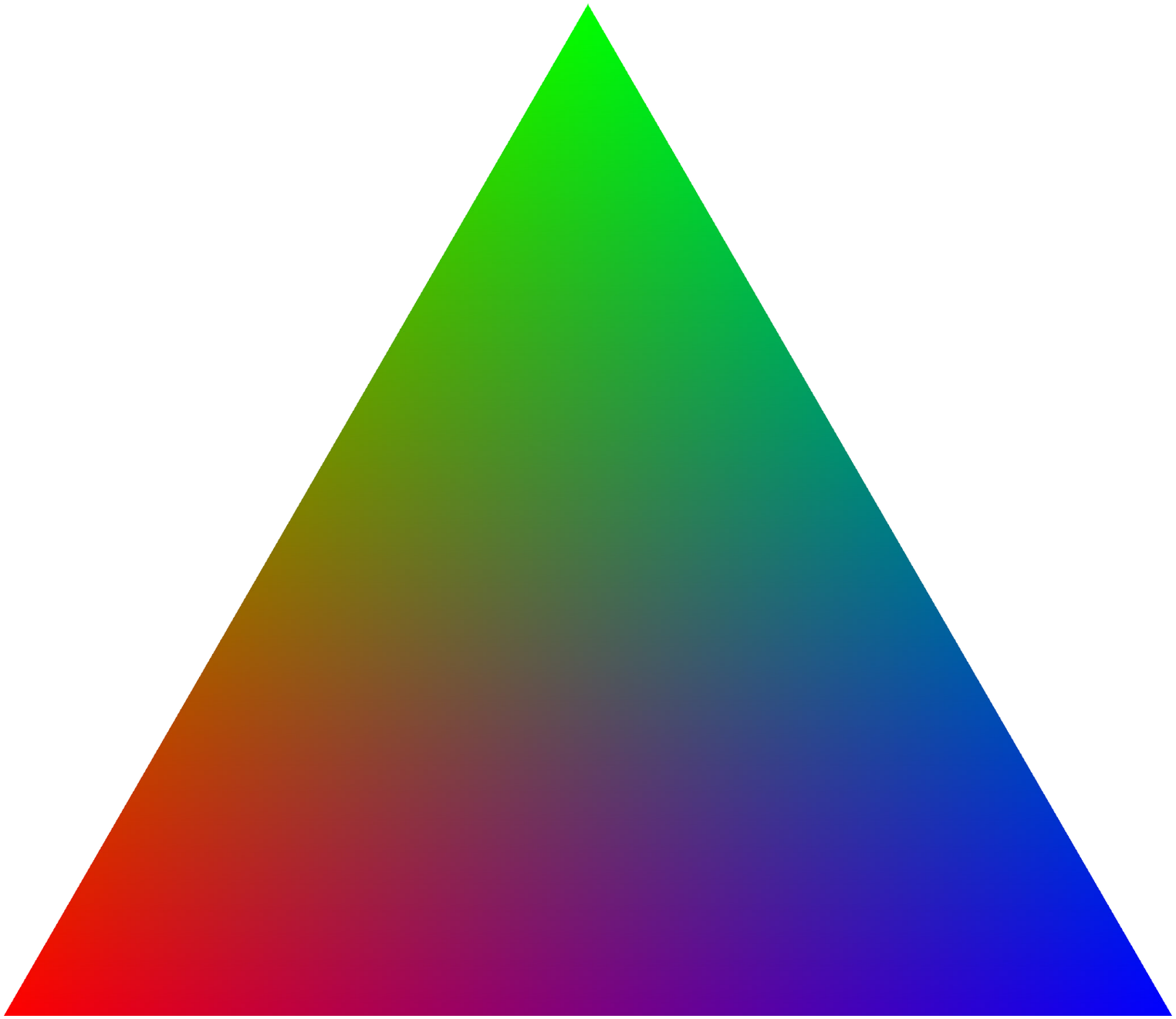
A bit more perception



Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Sheffield Methods Institute.

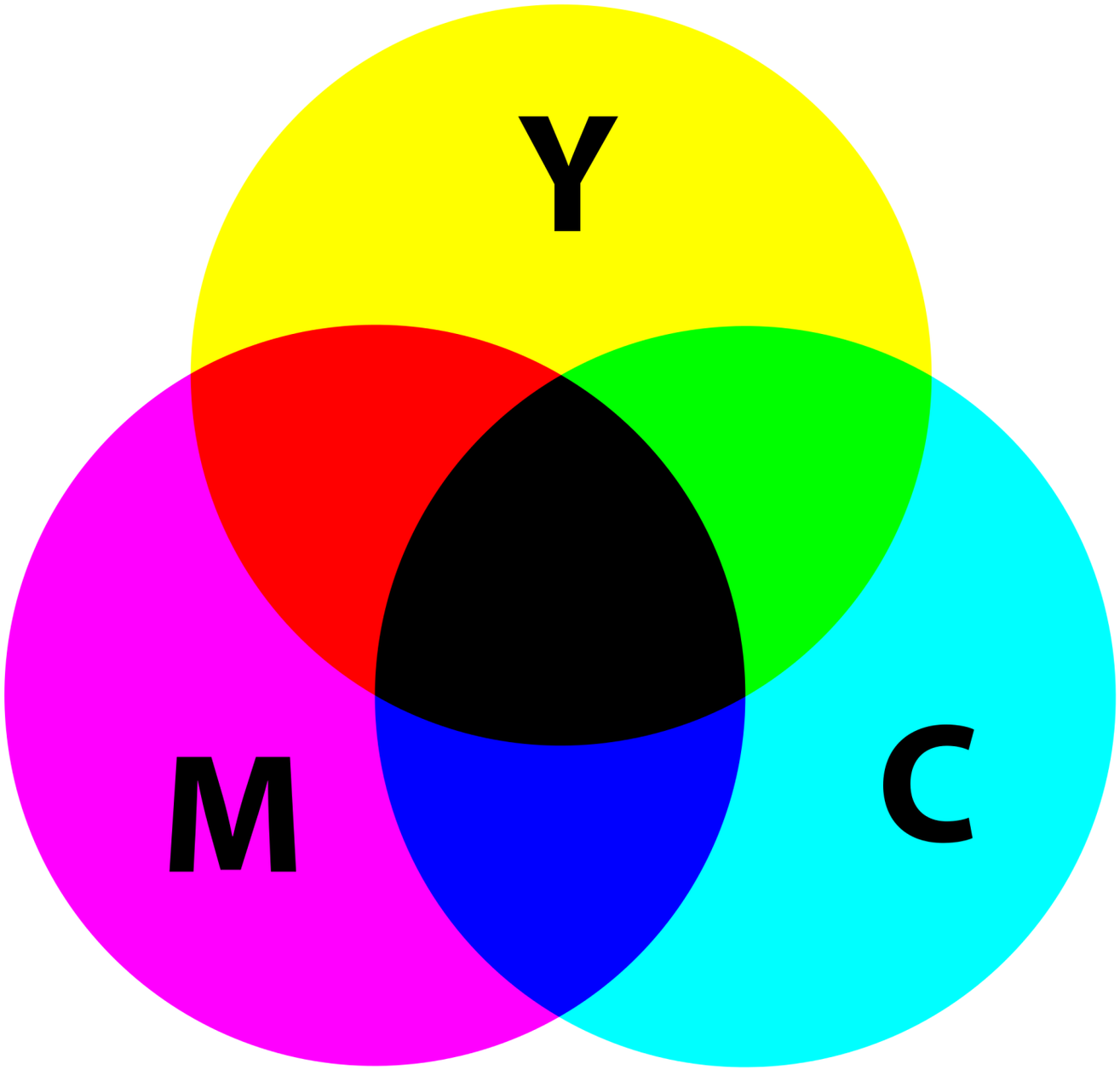
Colour spaces



Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

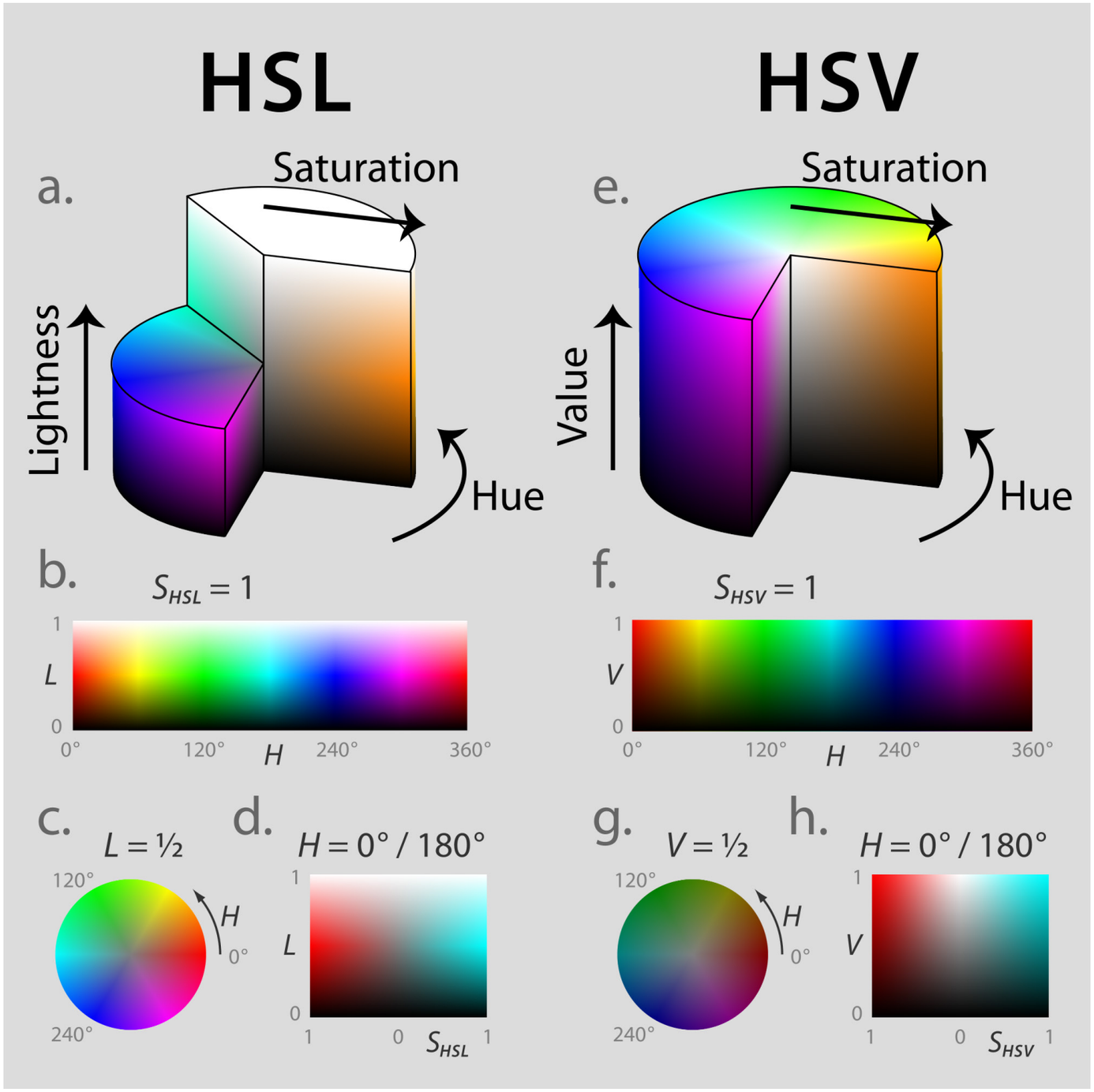
Sheffield Methods Institute.

Colour spaces



Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

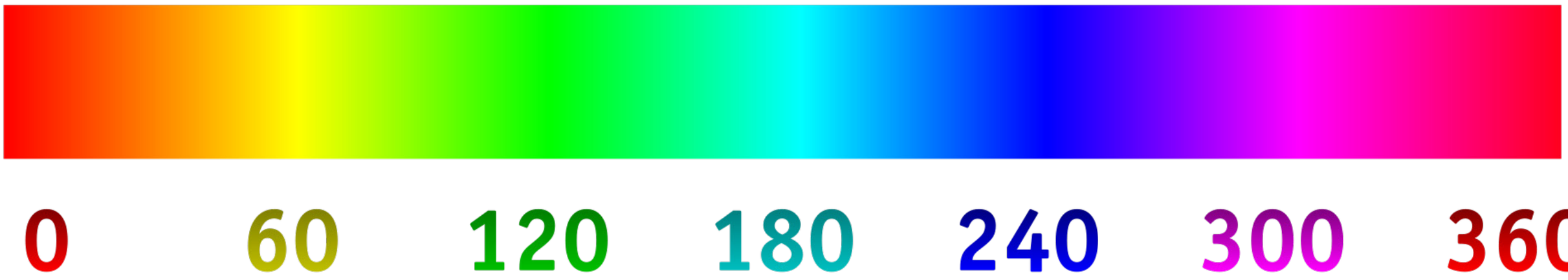
Colour spaces



Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Sheffield Methods Institute.

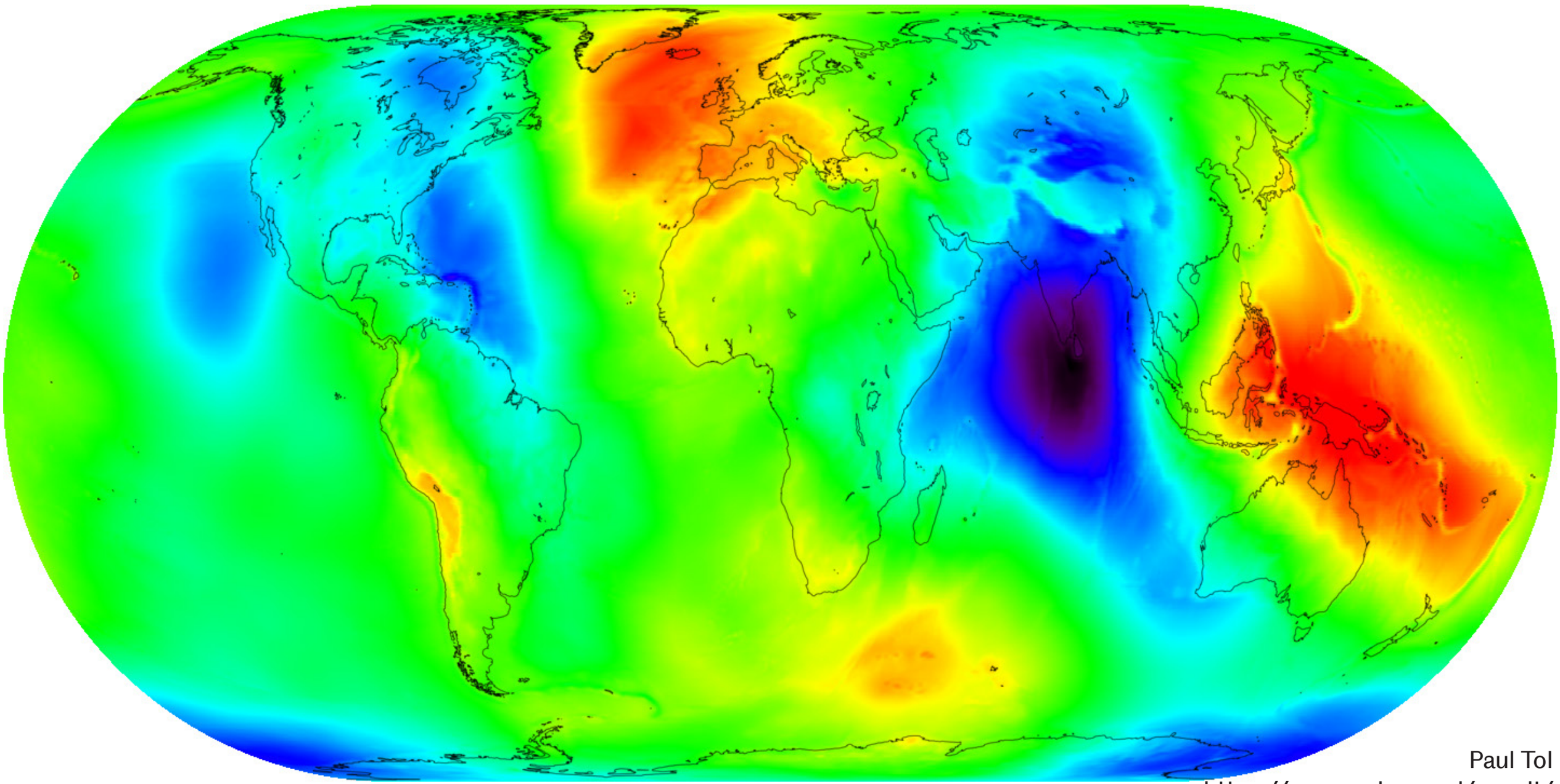
Colour spaces



Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Sheffield Methods Institute.

Colour spaces

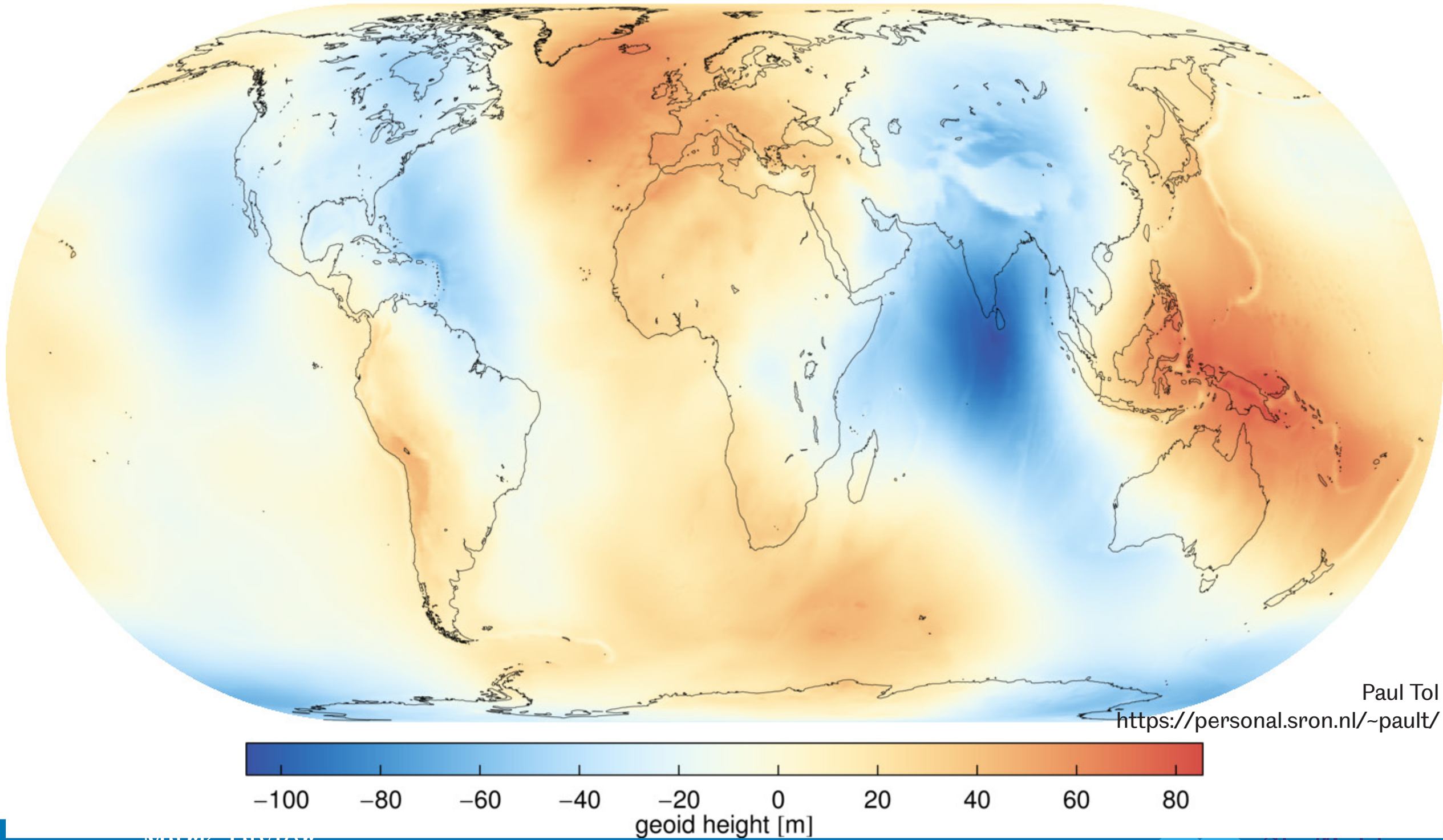


Paul Tol
https://personal.sron.nl/~pault/

Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Sheffield Methods Institute.

Colour spaces



Paul Tol
https://personal.sron.nl/~pault/

Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Sheffield Methods Institute.

What's the problem here?

Remember cones from before

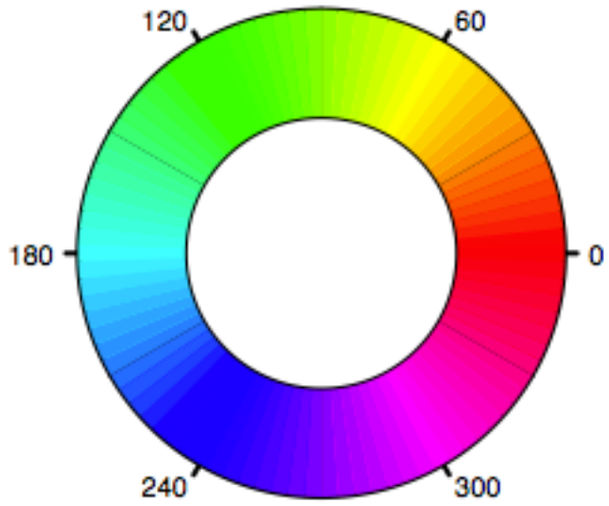
- In a rainbow palette, you see stripes
- Reflection of peaks and troughs in perceived brightness

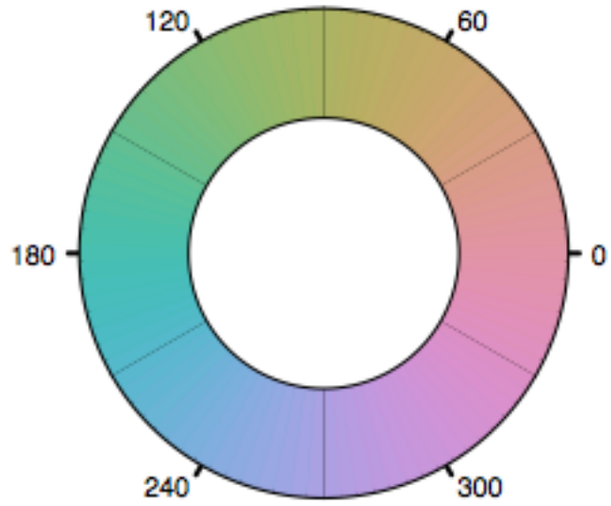
Alternative colour spaces

I recommend HCL

- Hue
  - dominant wavelength (or “colour” I guess)
- Chroma
  - colourfulness (distance from grey)
- Luminance
  - brightness (amount of grey)
- Doesn’t include saturation as different wavelengths’ perceived brightness varies

Why HCL?





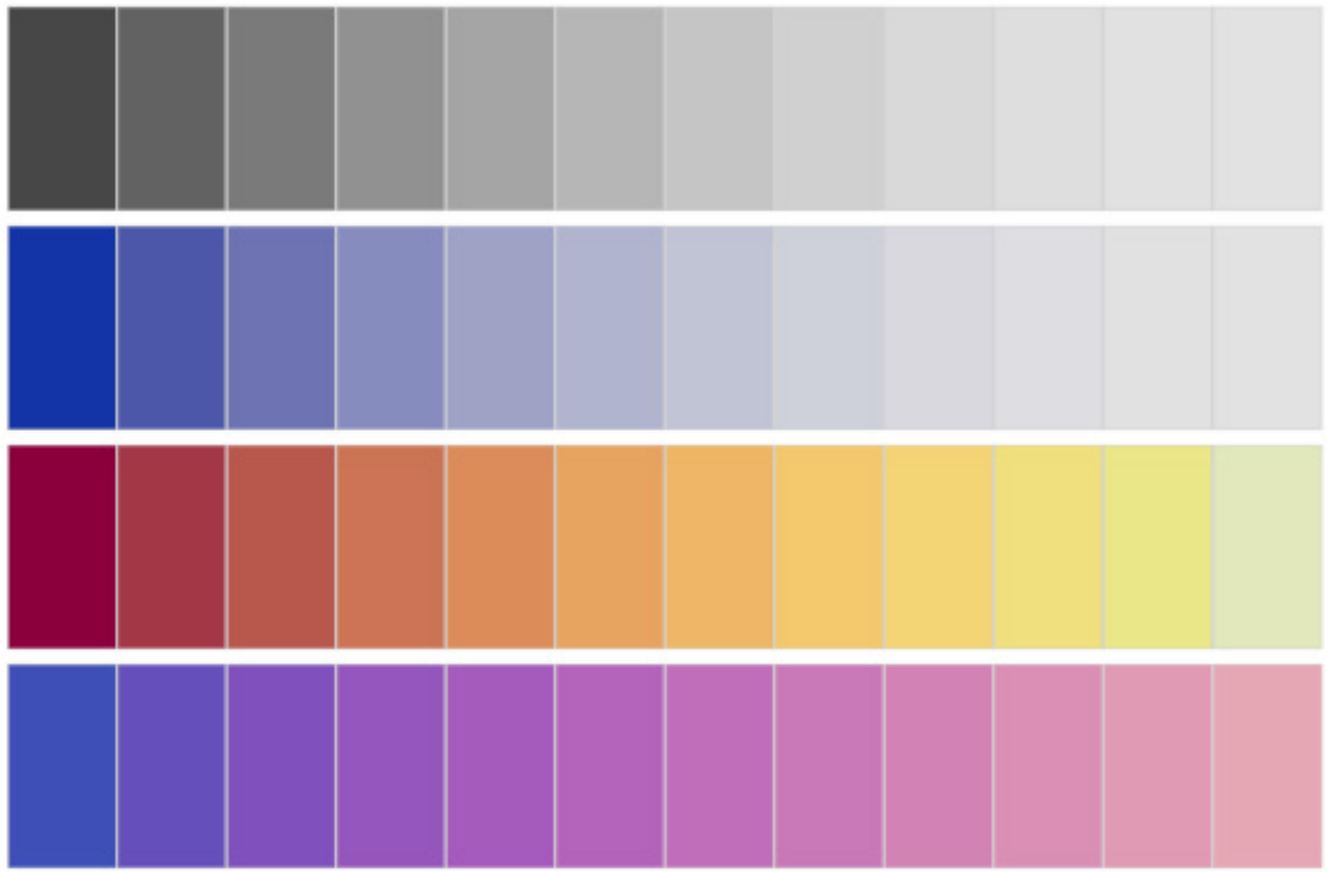
Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

S M I
Sheffield Methods Institute.

Why HCL?

Easier to vary one on its own

- Or vary two, or all three
- Less “lumpy” effect

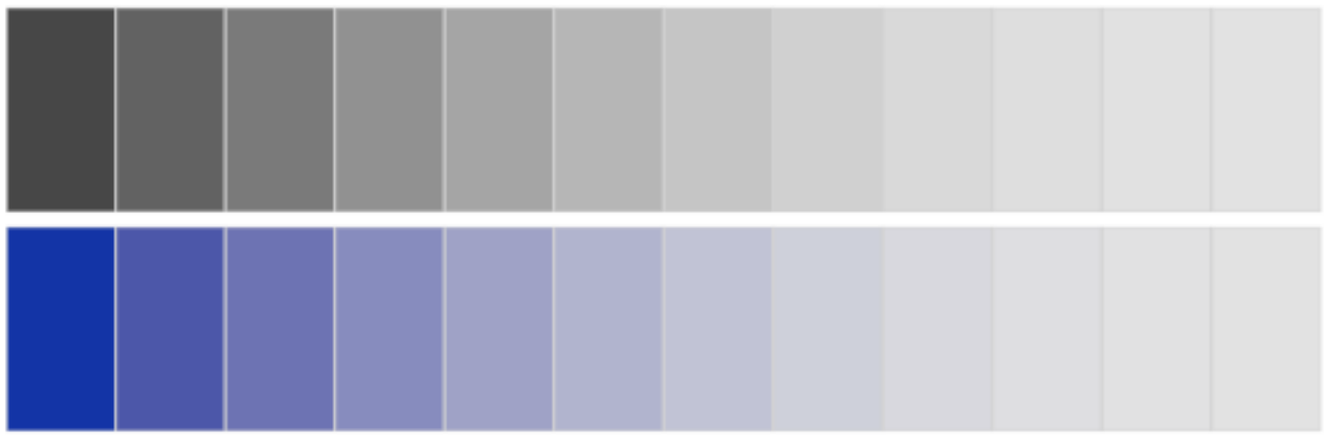


Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

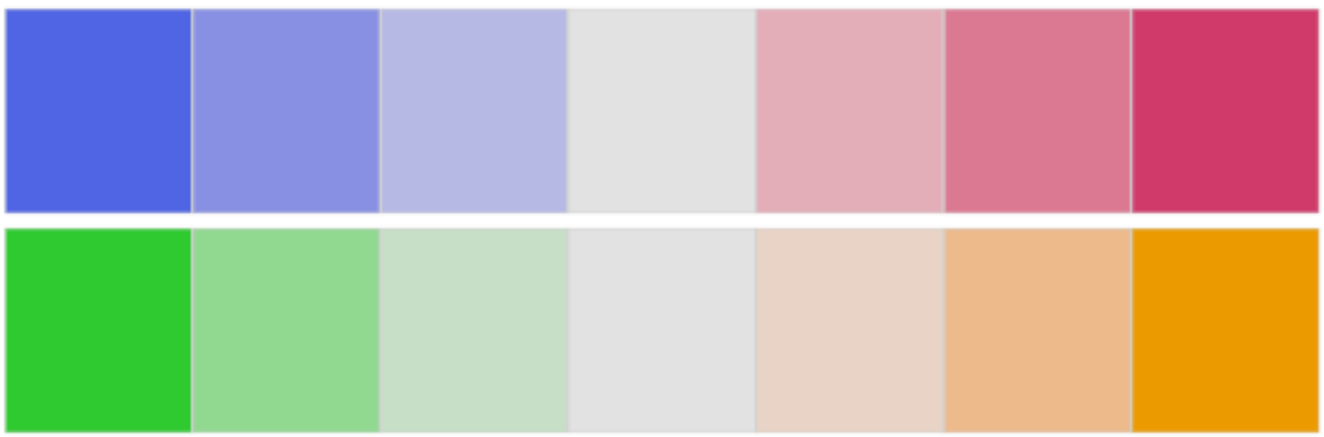
Sheffield Methods Institute.

Types of colour palettes

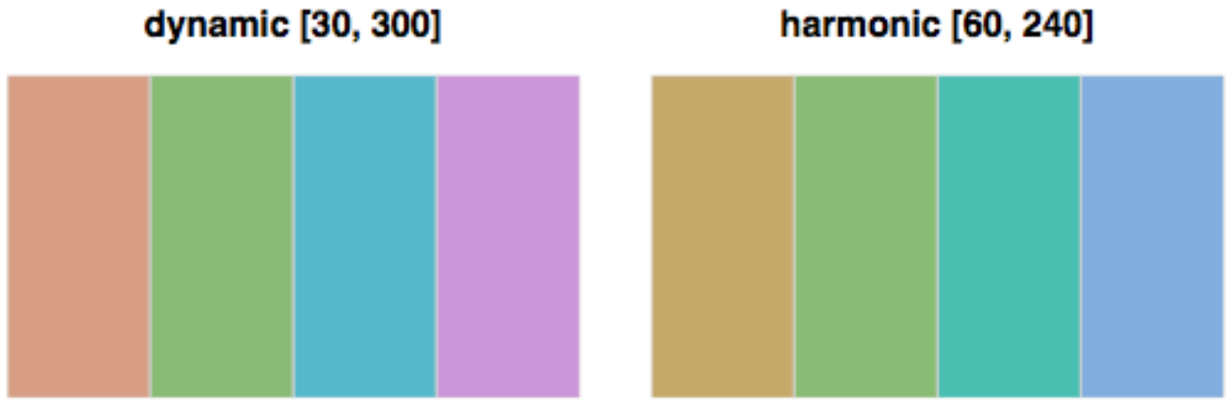
Sequential



Diverging

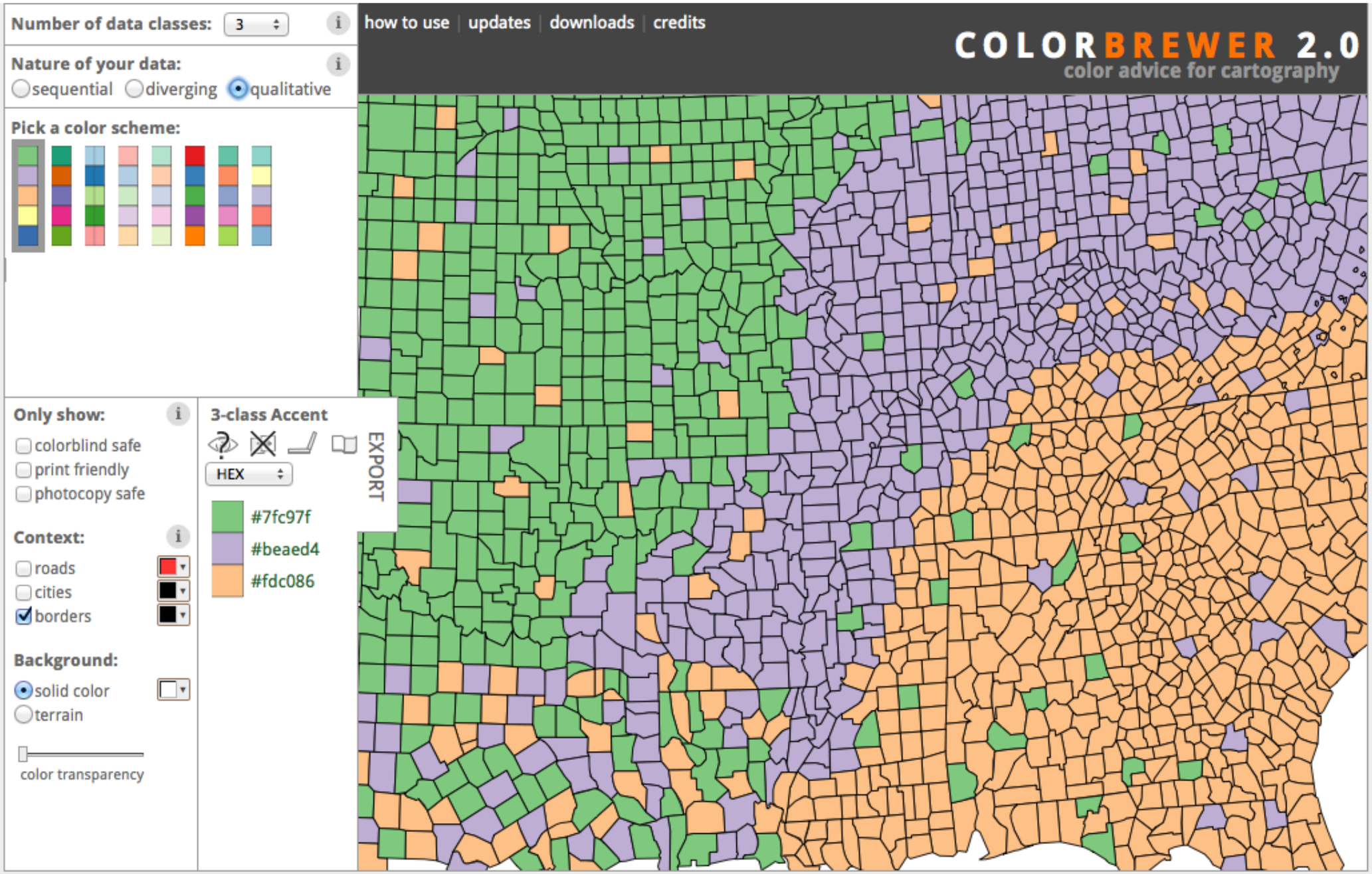


Qualitative



Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

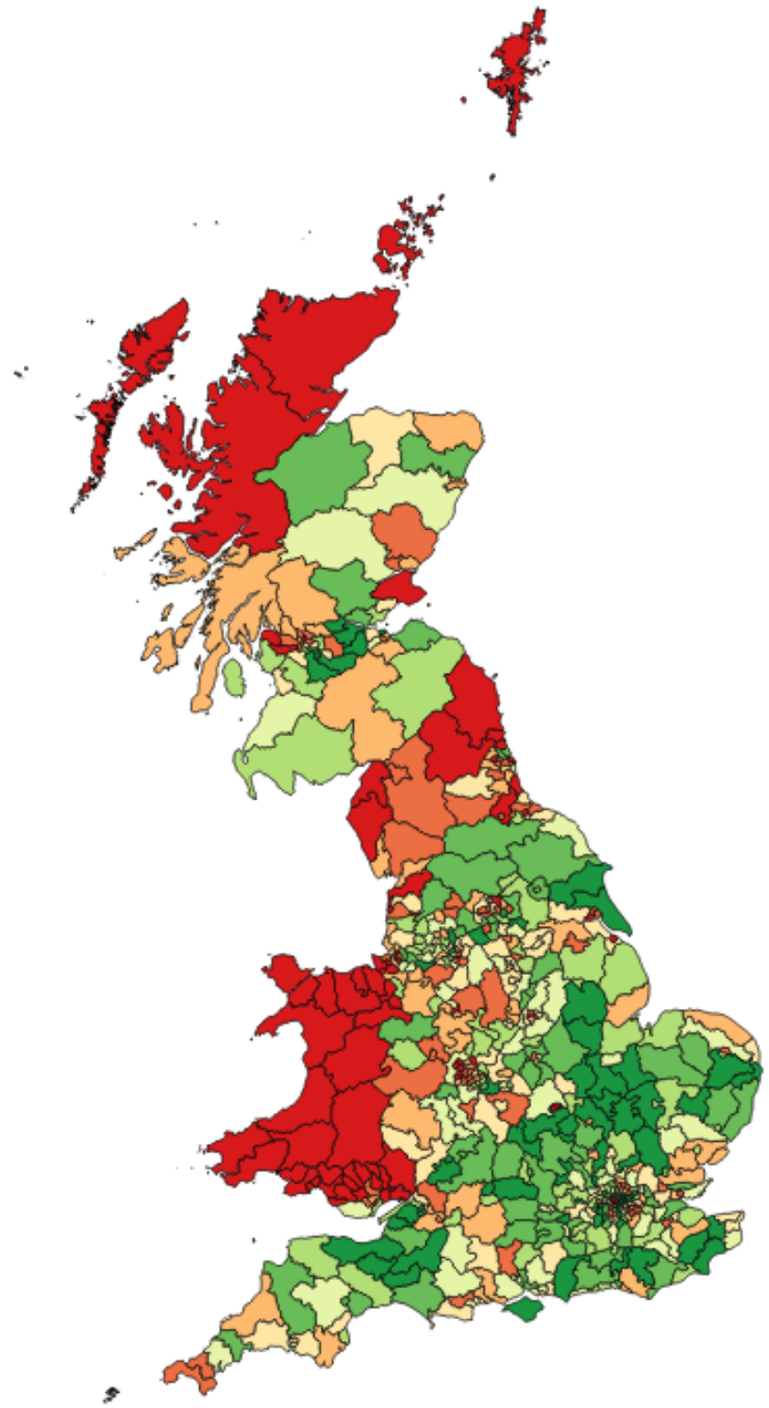
Recommendations



Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Social Analytics & Visualisation
Sheffield, 13/6/2022

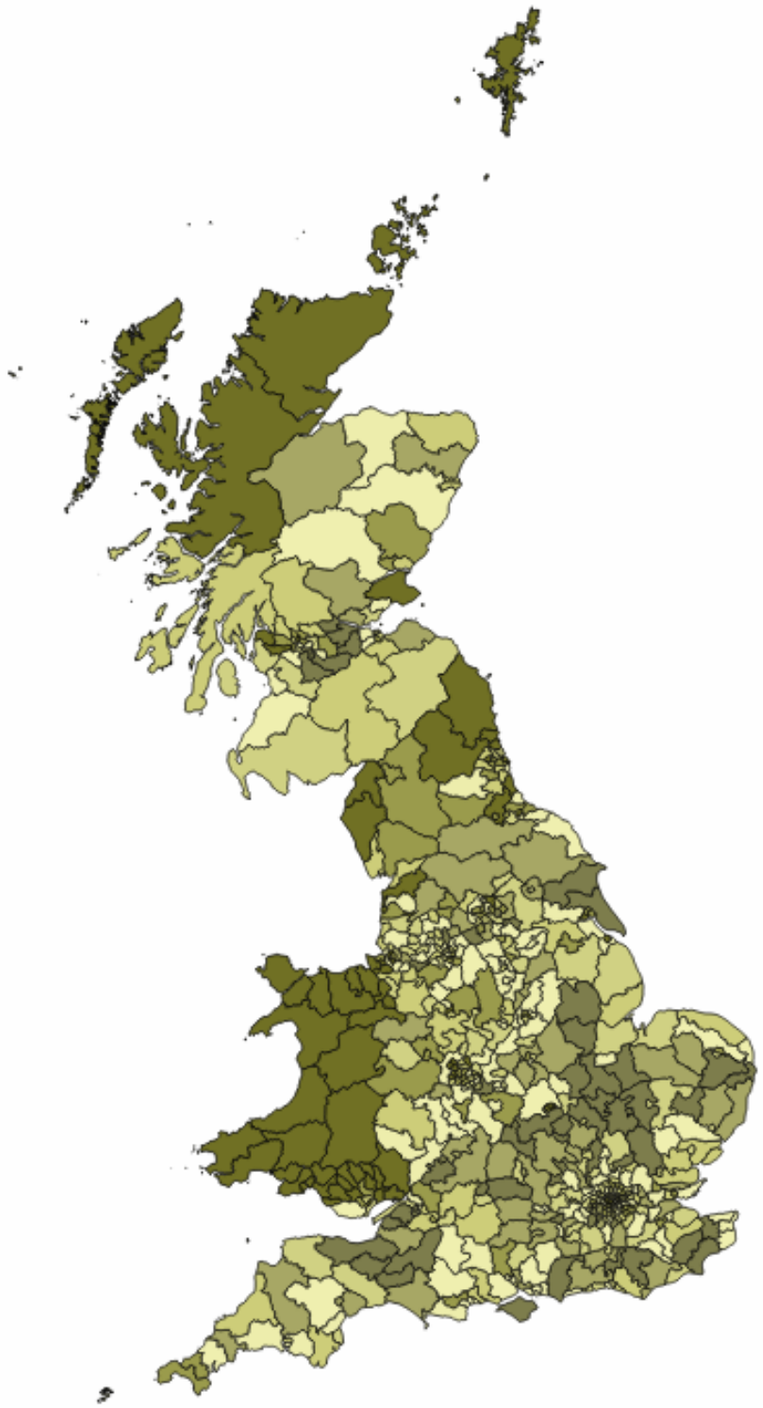
Sheffield Methods Institute.



Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Social Analytics & Visualisation
Sheffield, 13/6/2022

Sheffield Methods Institute.

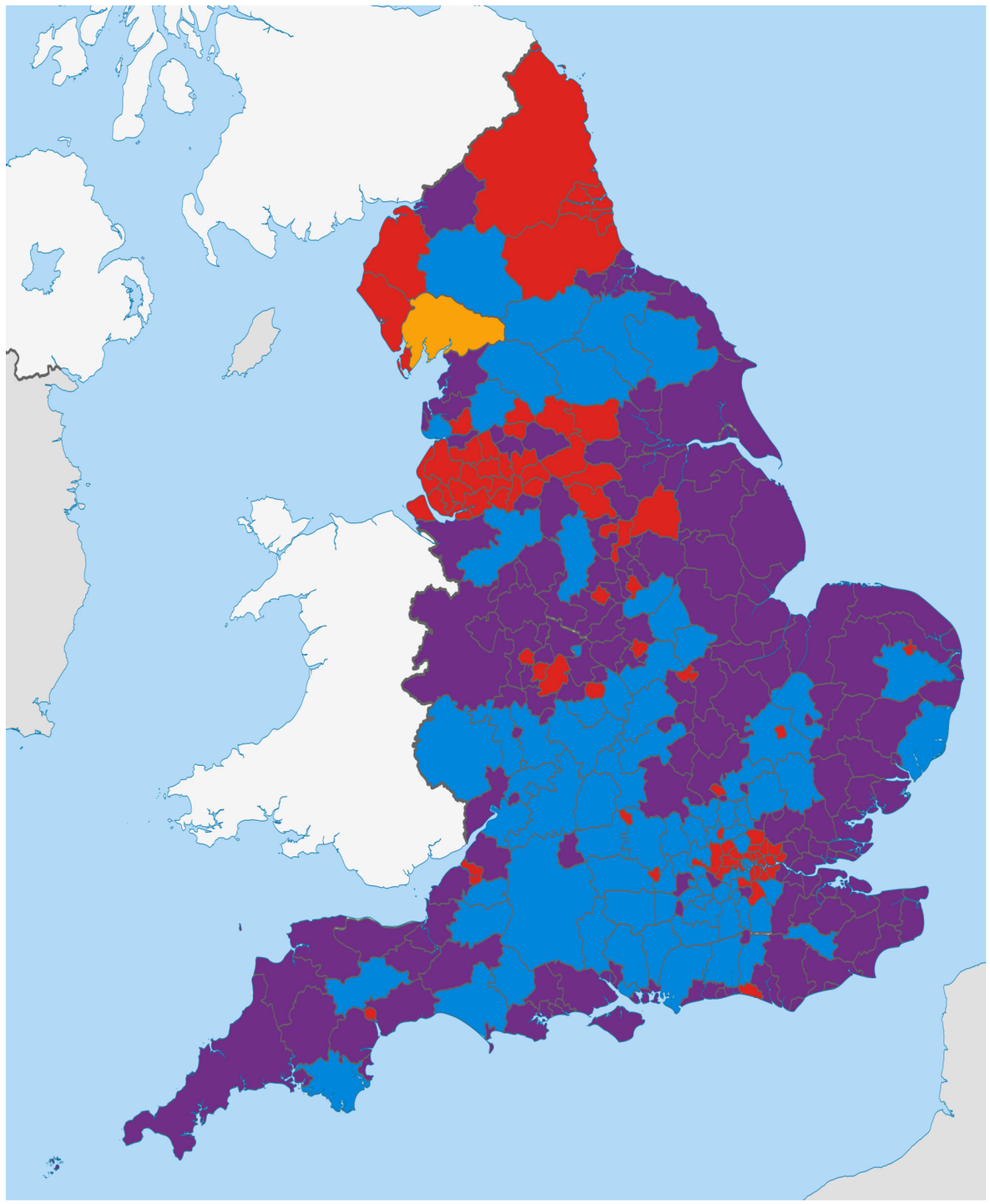


Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Social Analytics & Visualisation
Sheffield, 13/6/2022

Sheffield Methods Institute.

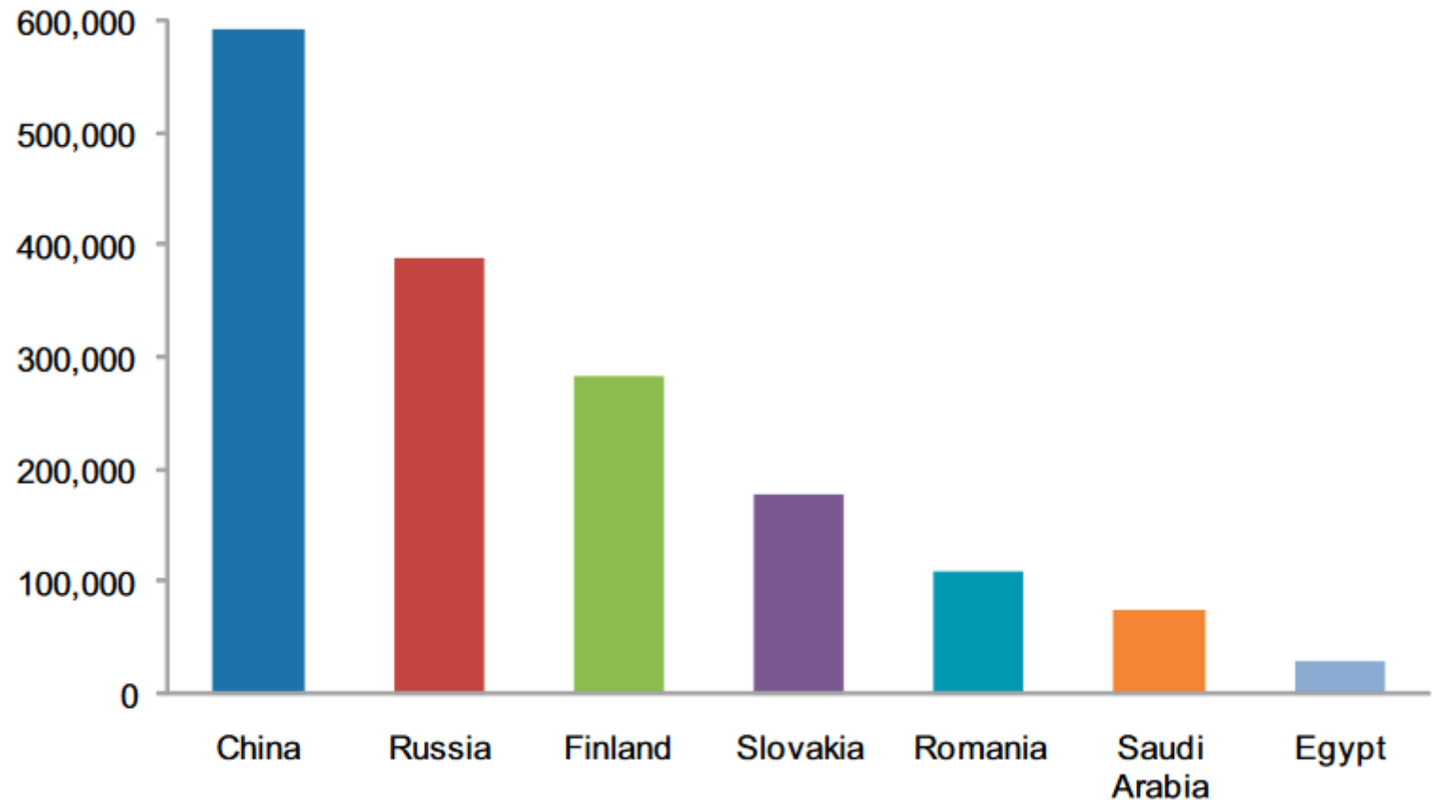
For example



Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

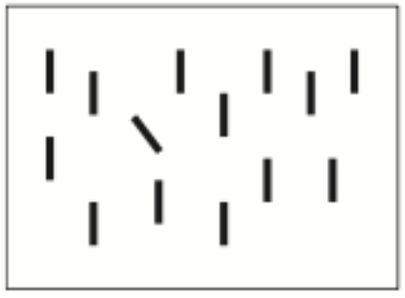
Sheffield Methods Institute.

For example

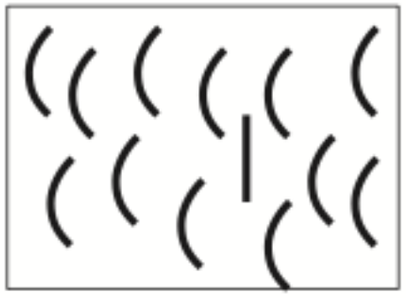

600,000
500,000
400,000
300,000
200,000
100,000
0
China
Russia
Finland
Slovakia
Romania
Saudi Arabia
Egypt

Shapes (briefly)

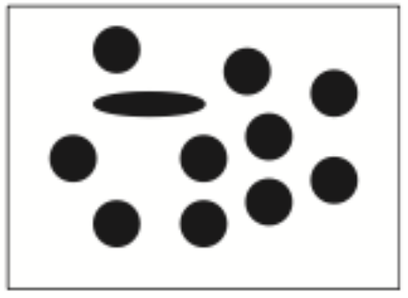
Orientation



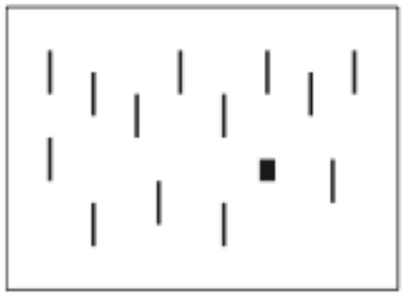
Curved/straight



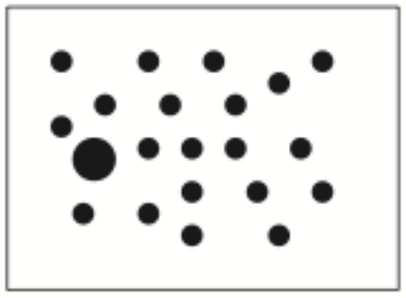
Shape



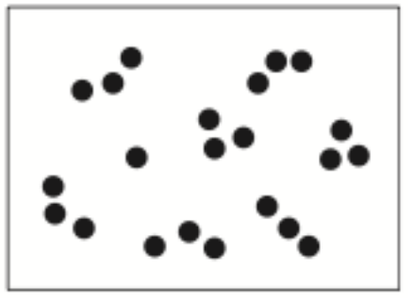
Shape



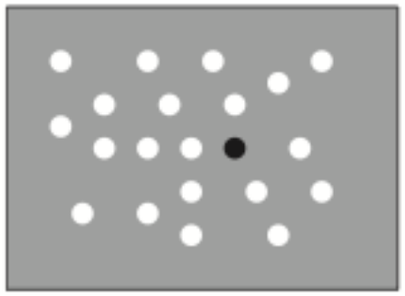
Size



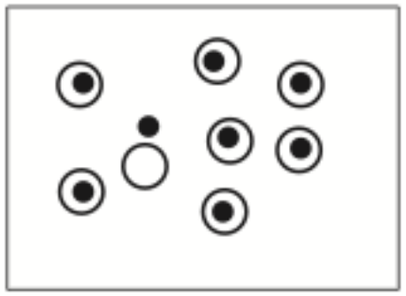
Number



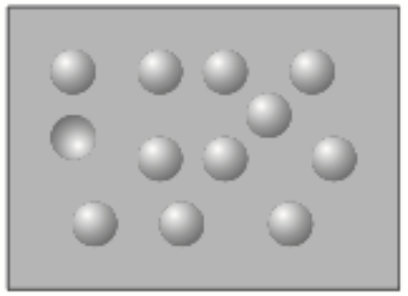
Gray/value



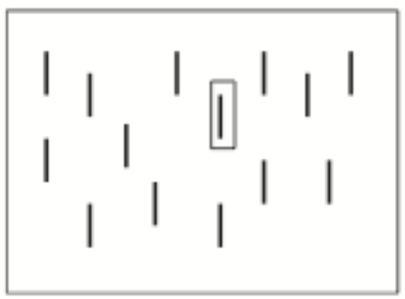
Enclosure



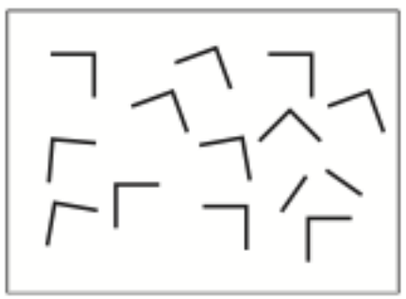
Convexity/concavity



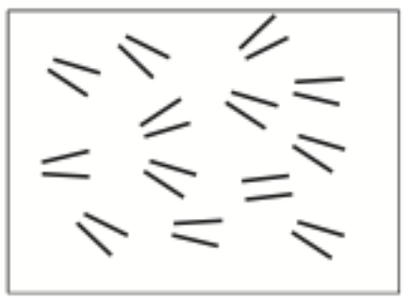
Addition



Juncture



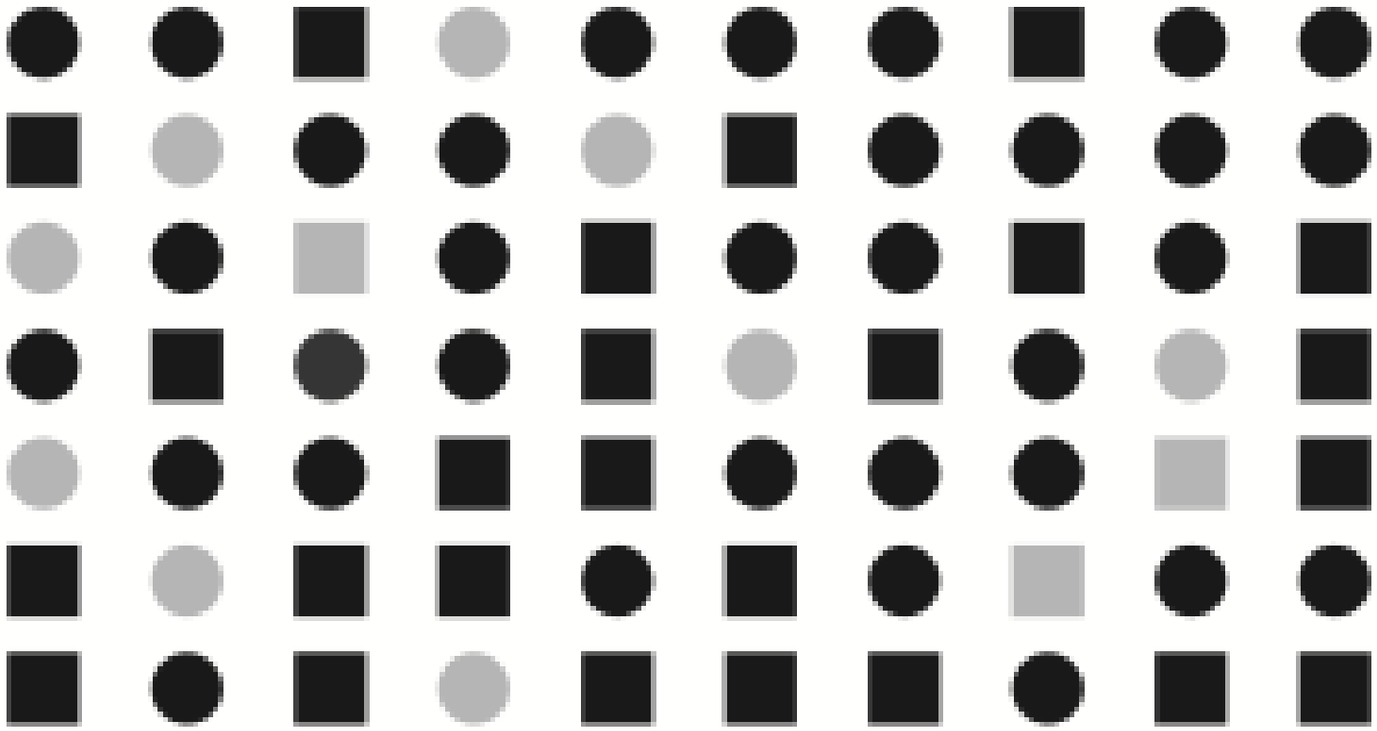
Parallelism



Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Shapes (briefly)

Resist the temptation to plot all the dimensions you can

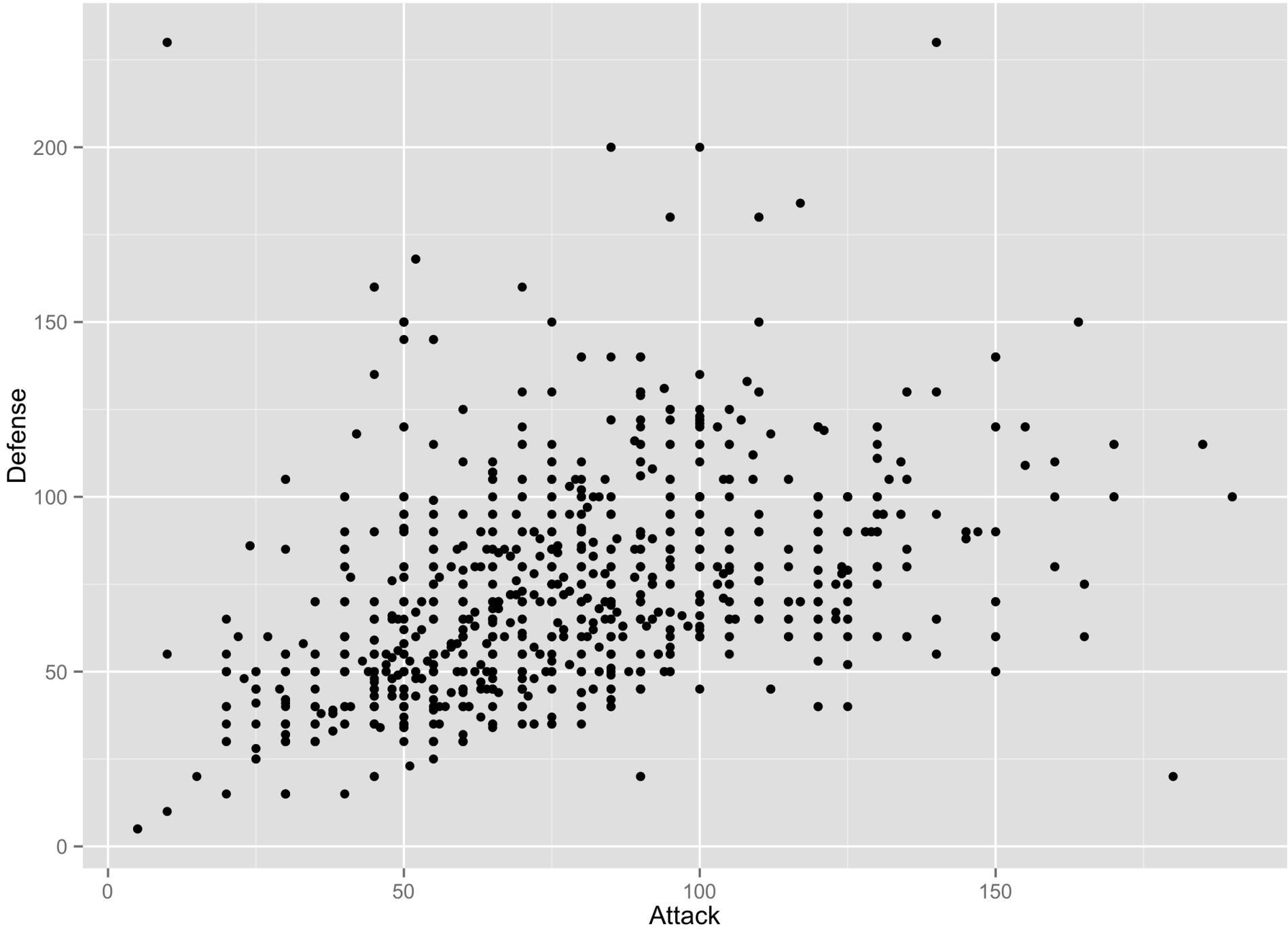


Let's colour in some graphs

Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Sheffield Methods Institute.

You've seen this before

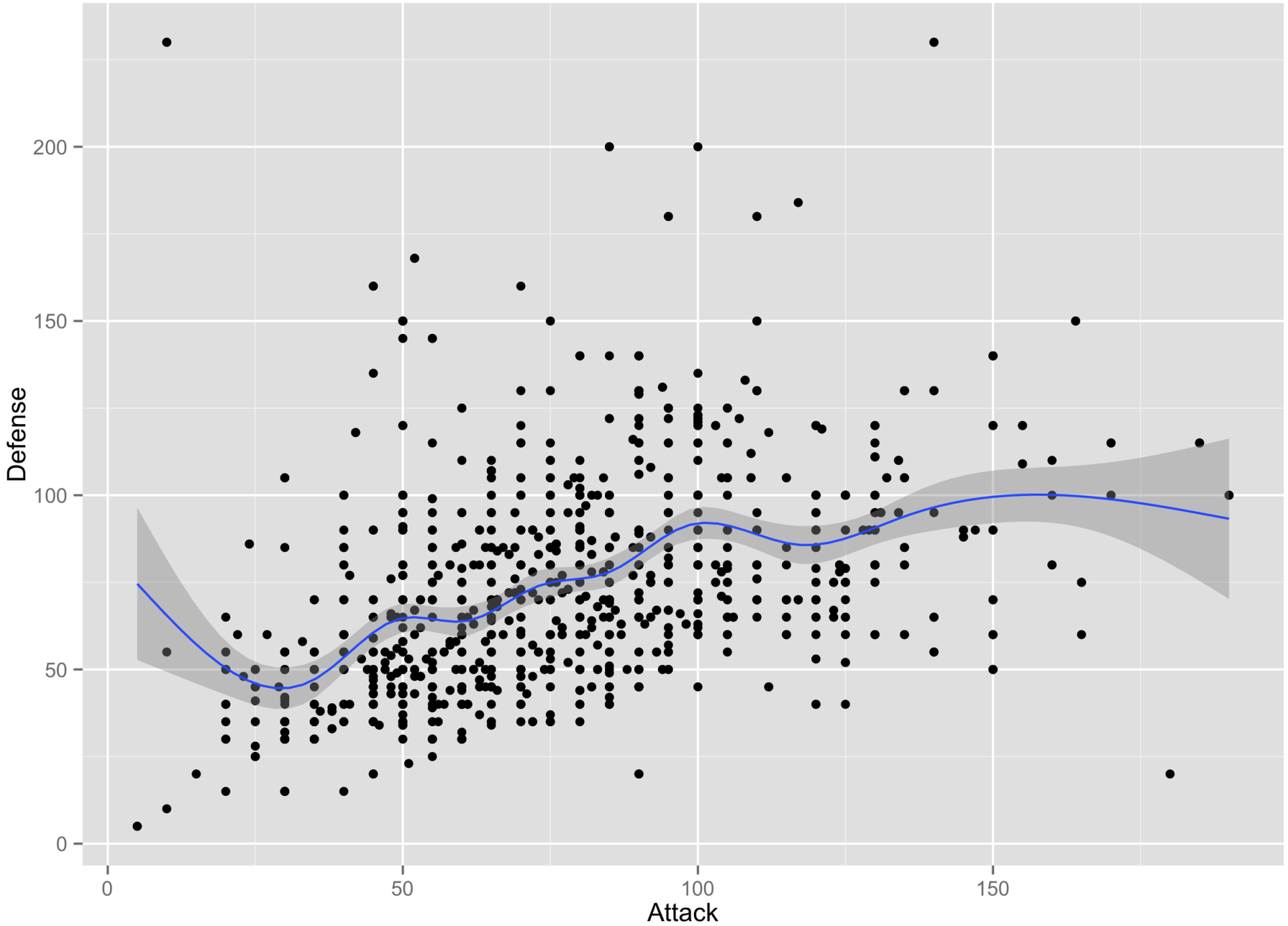


Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Social Analytics & Visualisation
Sheffield, 13/6/2022

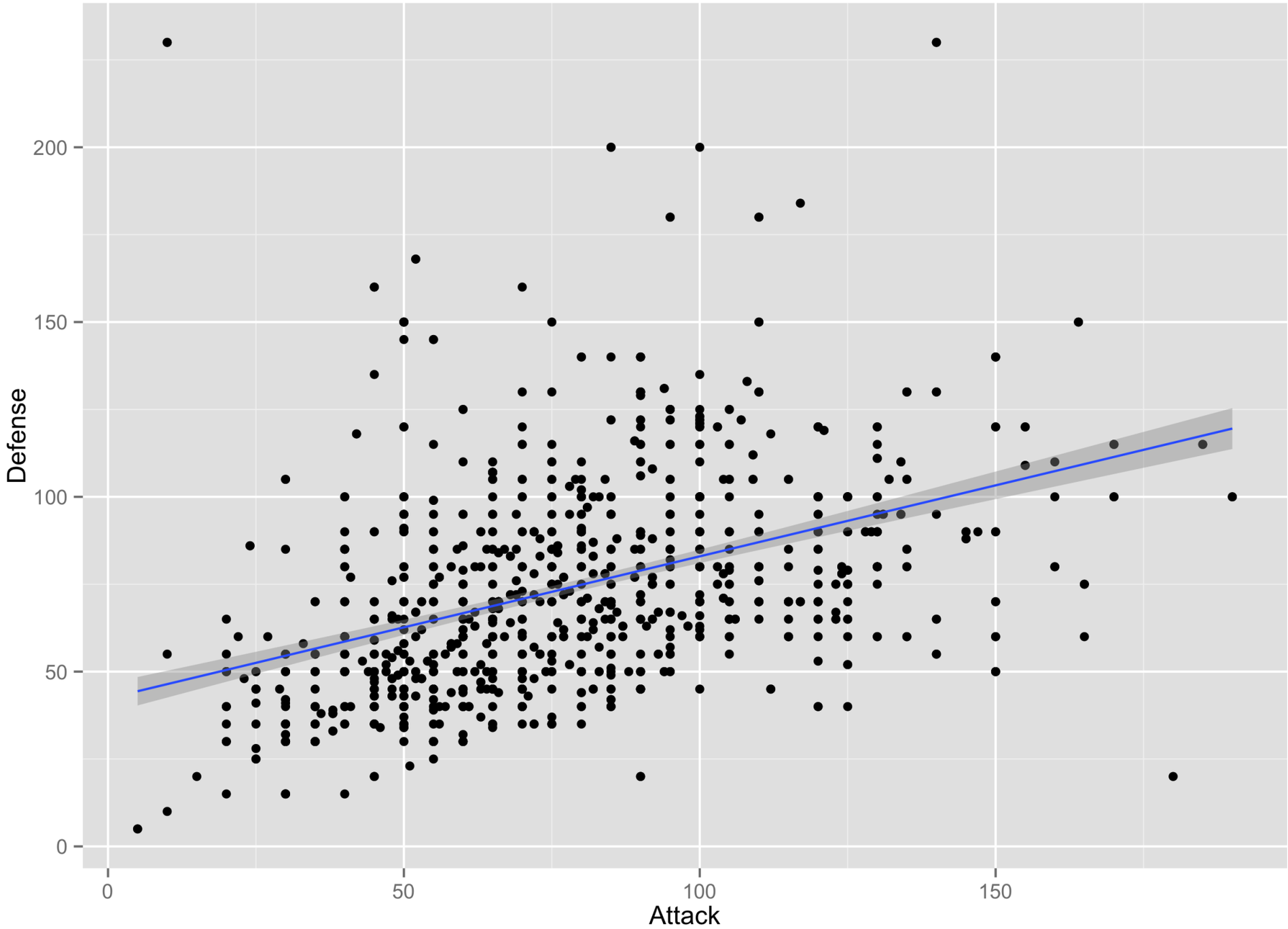
Sheffield Methods Institute.

You've seen this before



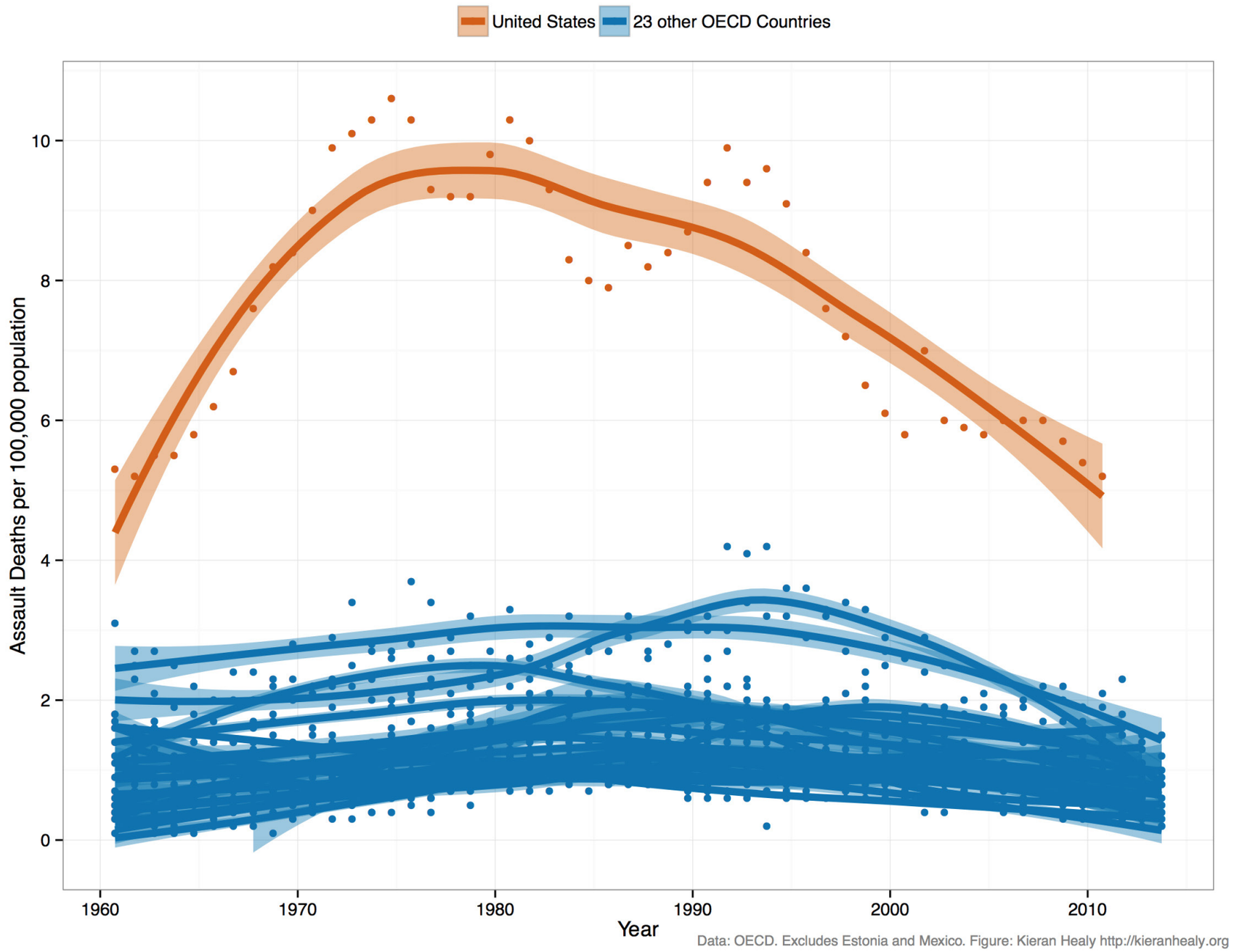
Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

You've seen this before



Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Utility of LOESS curves in viz



Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Social Analytics & Visualisation
Sheffield, 13/6/2022

Sheffield Methods Institute.

Models in viz

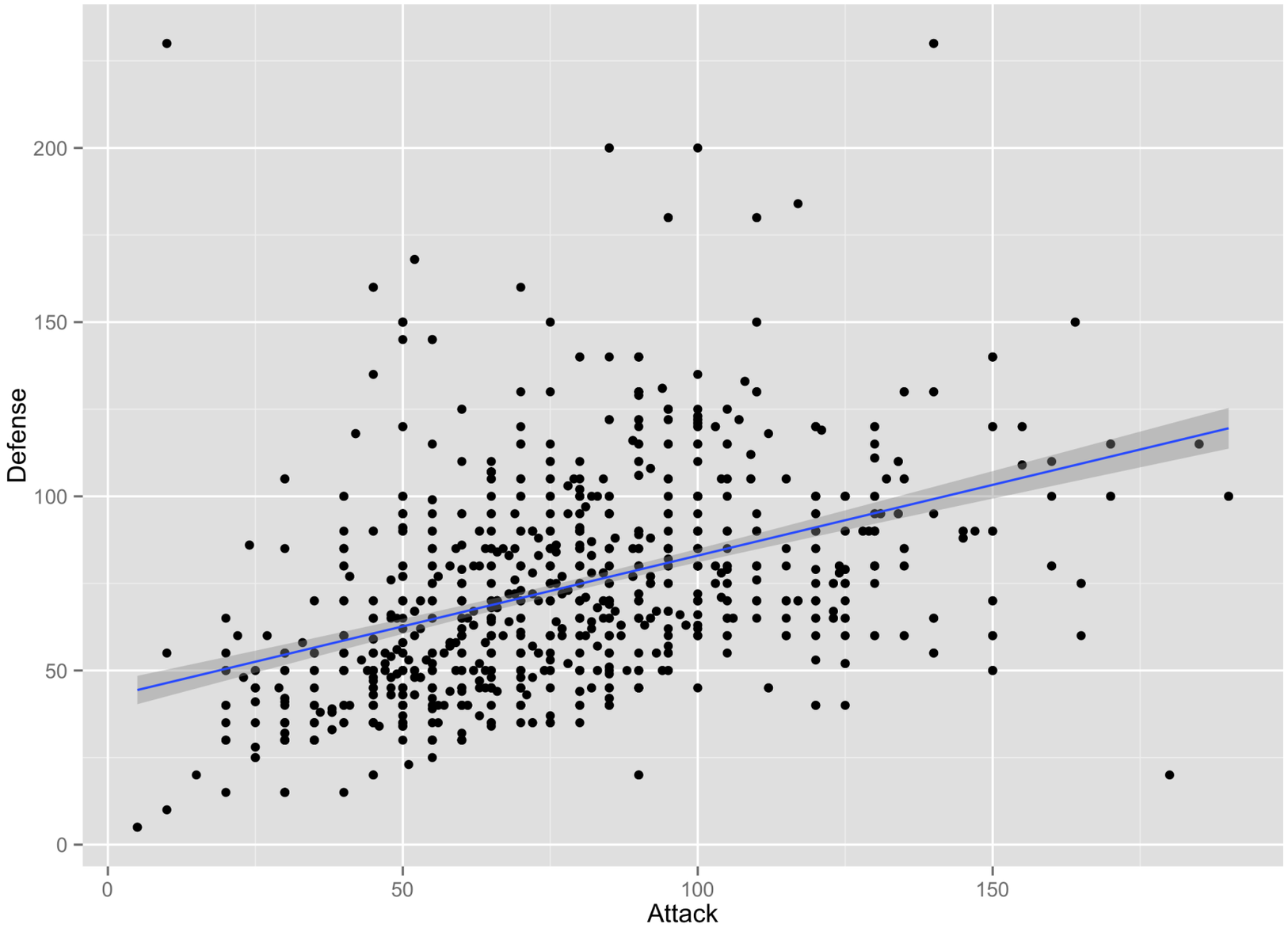
The model can be the finding itself

- as in the LOESS curves above

or viz can be used to *assess* the model

Visualising linear regression

Let's revisit the linear model from before



$$y_i = \alpha + \beta x_i + \varepsilon$$

Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Some regression assumptions

- The x and y variables have a linear relationship
- Residuals (ε) are normally distributed
- Homoskedasticity
- Average error = 0

If these all hold, then α and β are BLUE (Best Linear Unbiased Estimate)

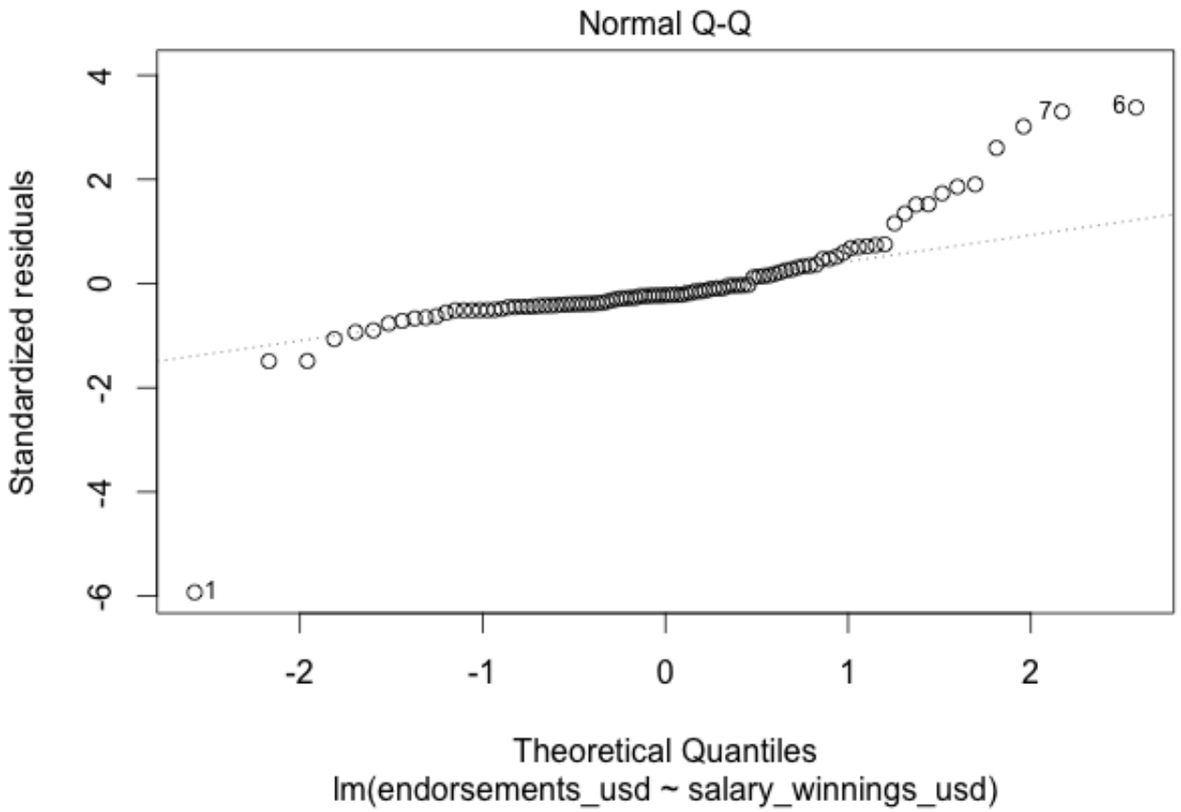
- What's the relevance for visualisation?

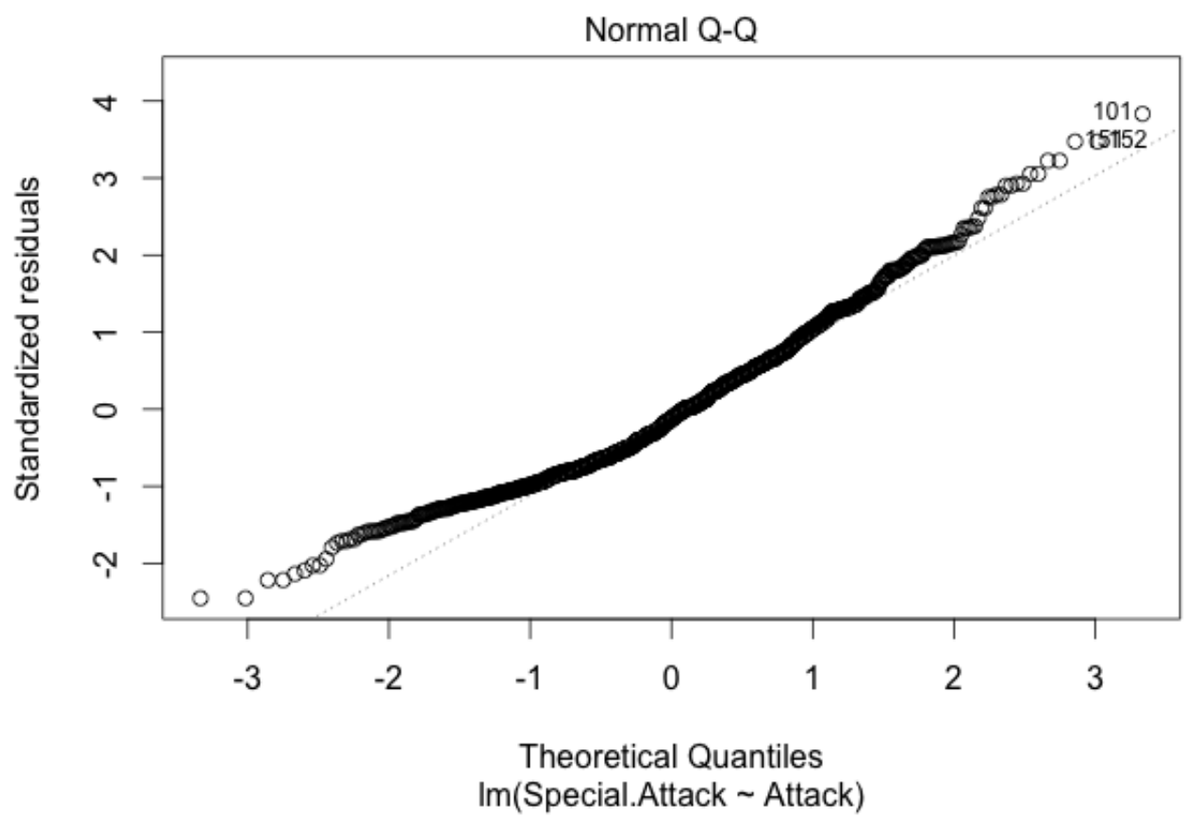
quantile-quantile plot

Compares a theoretical distribution to an empirical distribution

- If the assumptions are (perfectly) met, you get a straight line at a 45° angle;
- assesses normality of error terms

qq plots – compare

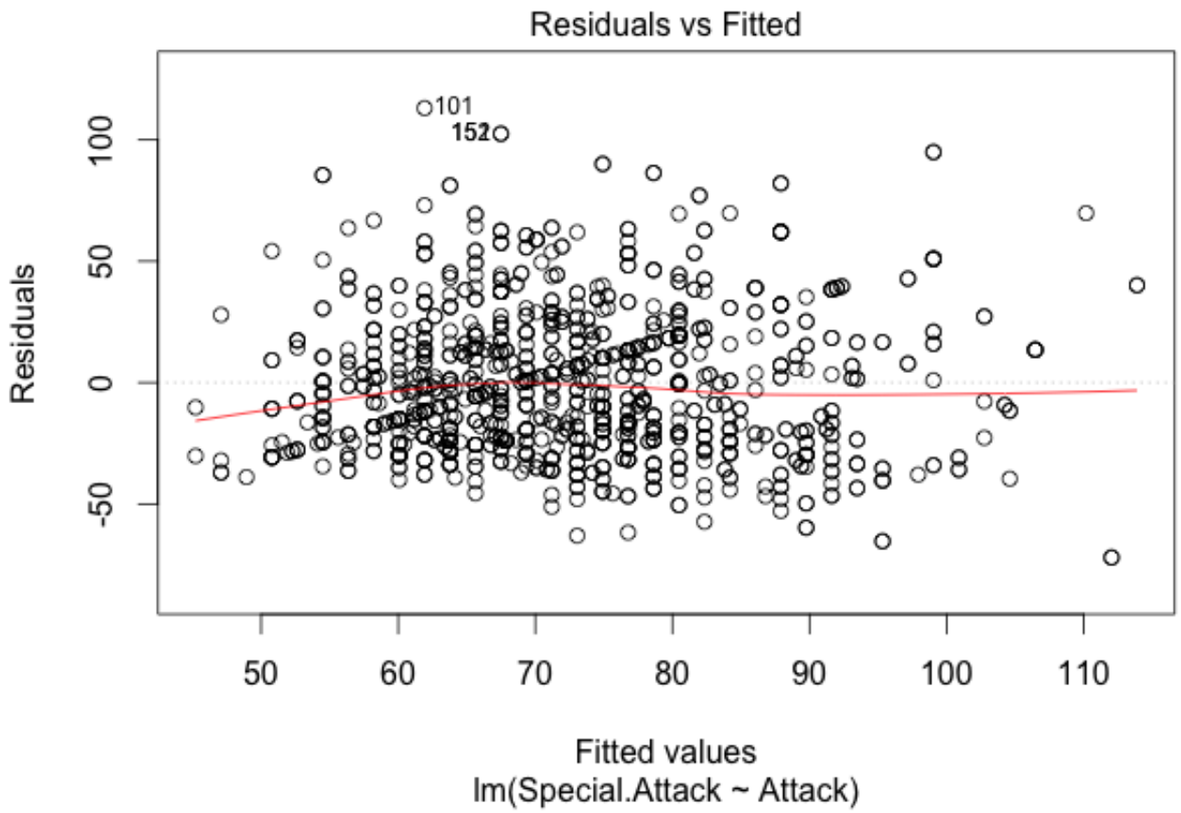


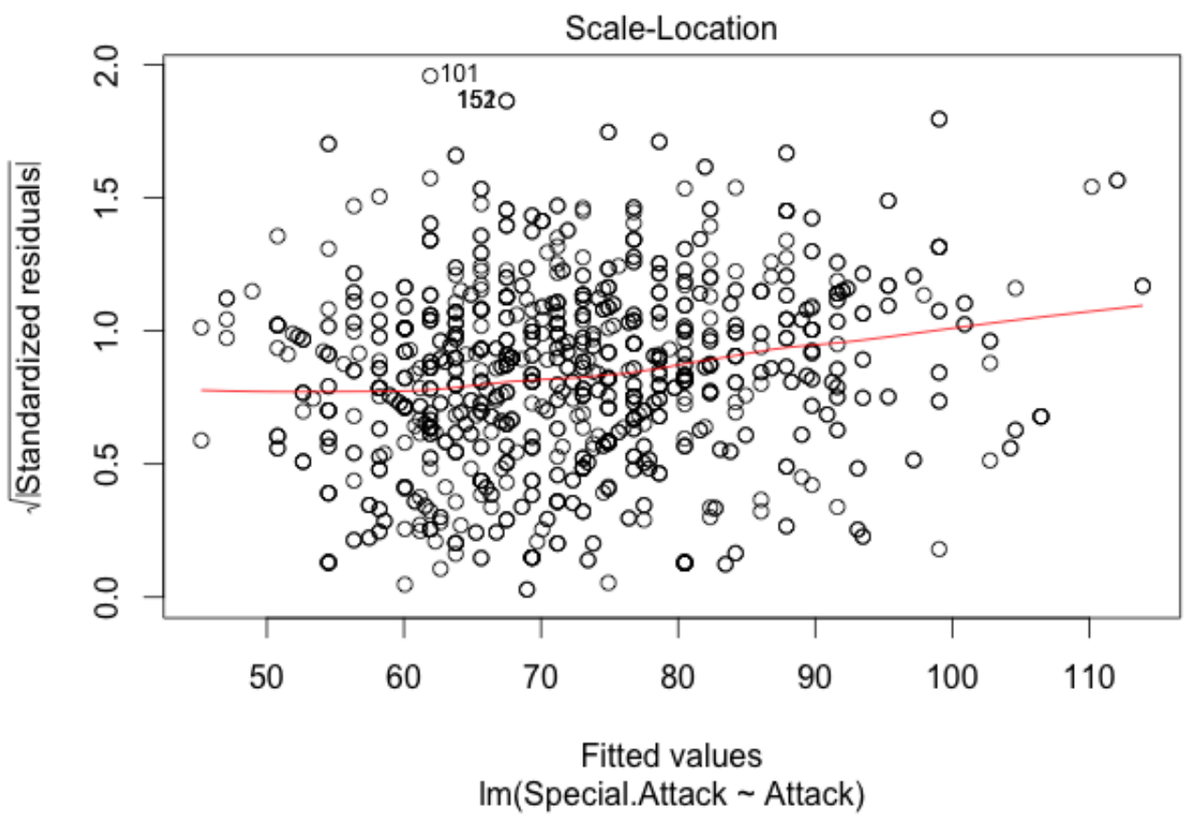


Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Heteroskedasticity

Observe distributions of error terms by x

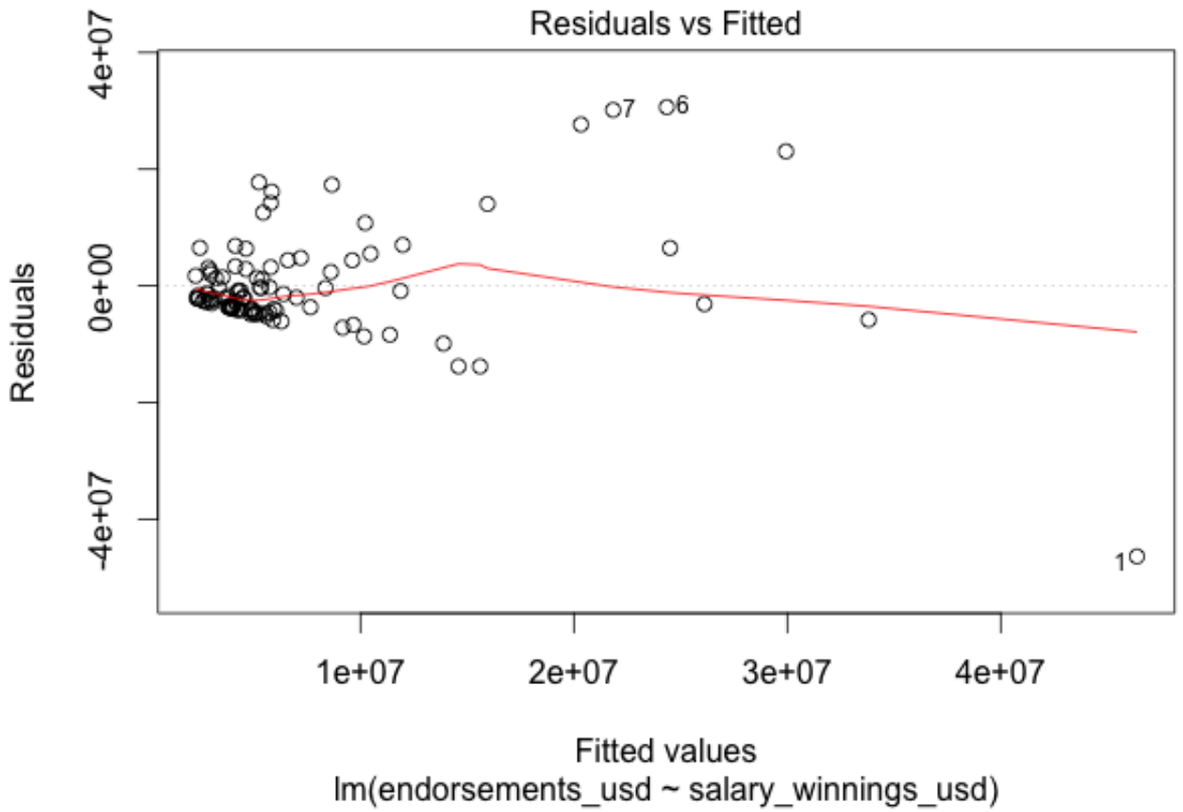


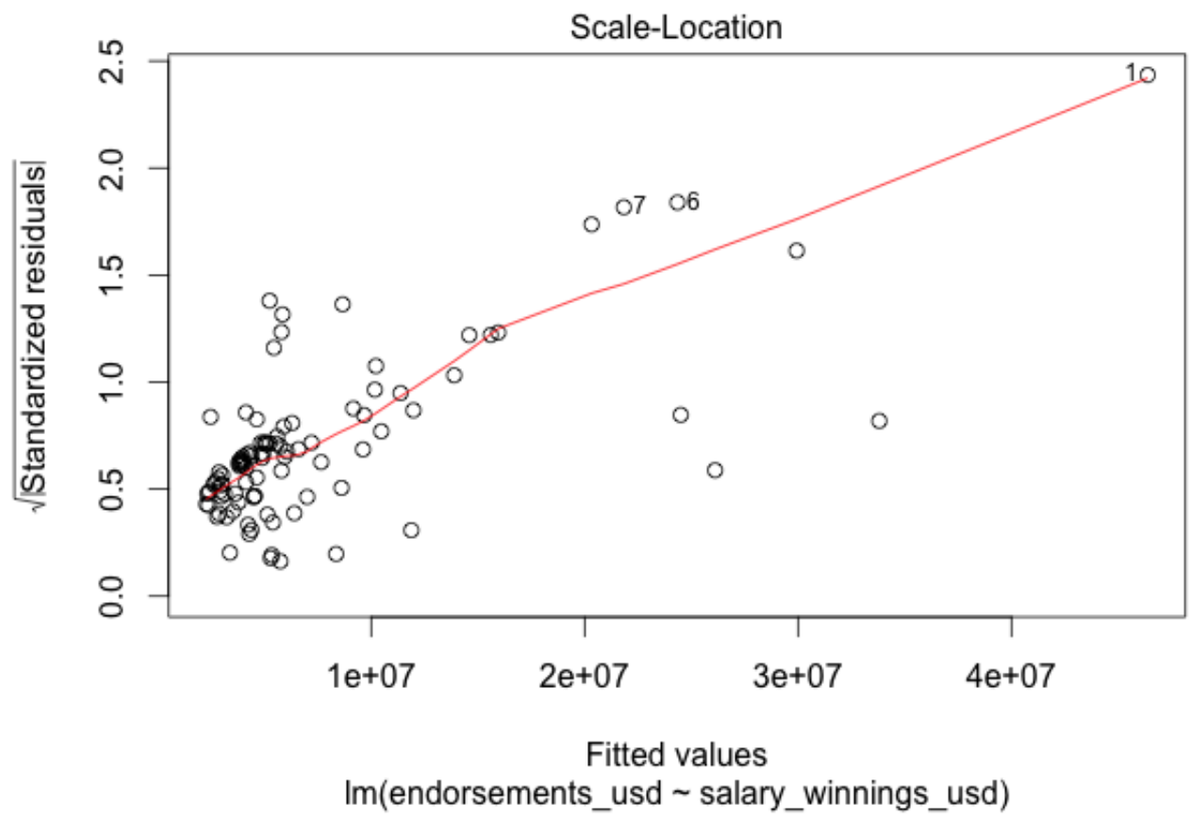


Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Sheffield Methods Institute.

Heteroskedasticity



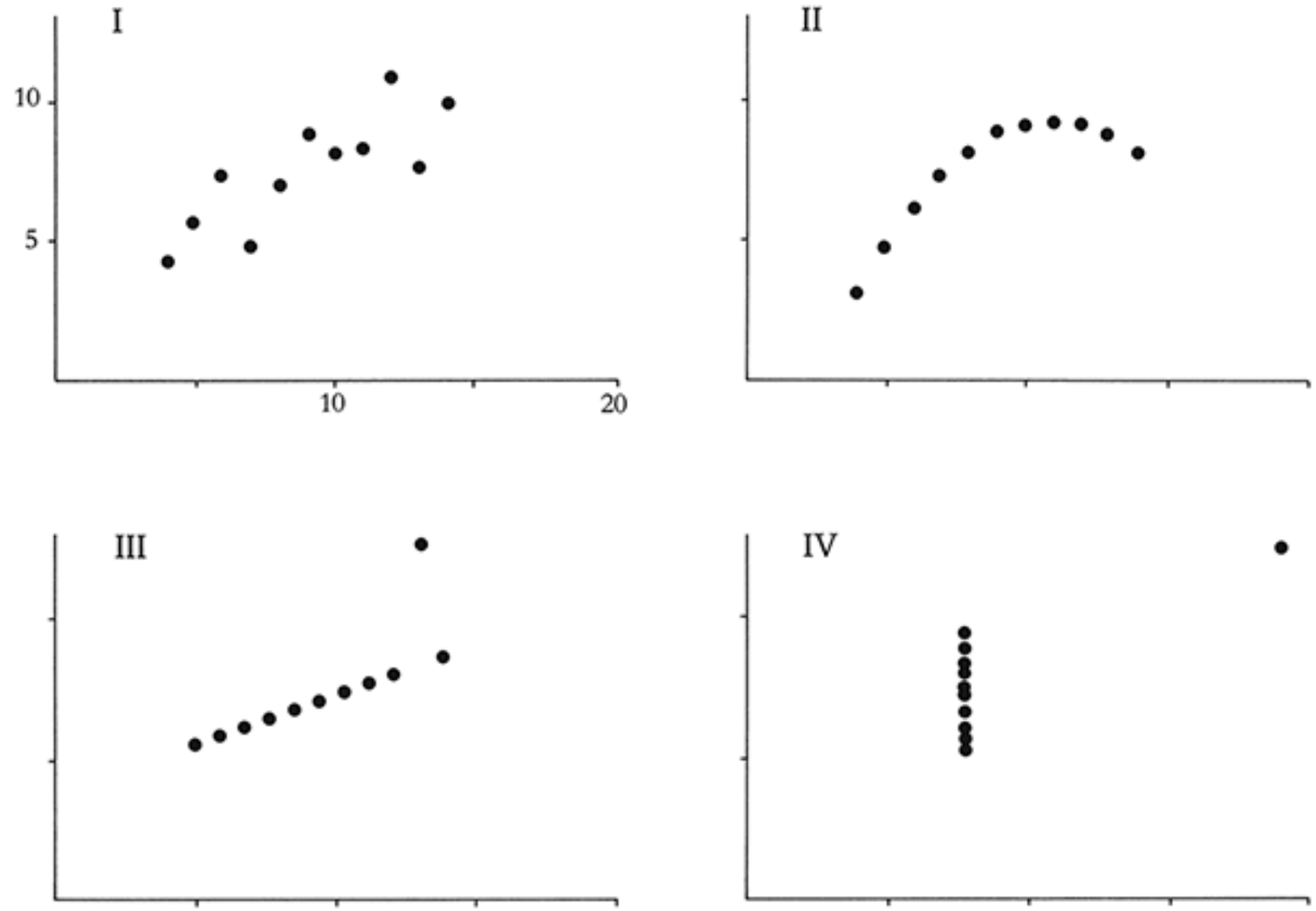


Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Sheffield Methods Institute.

Outliers

Recall Anscombe's quartet



I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

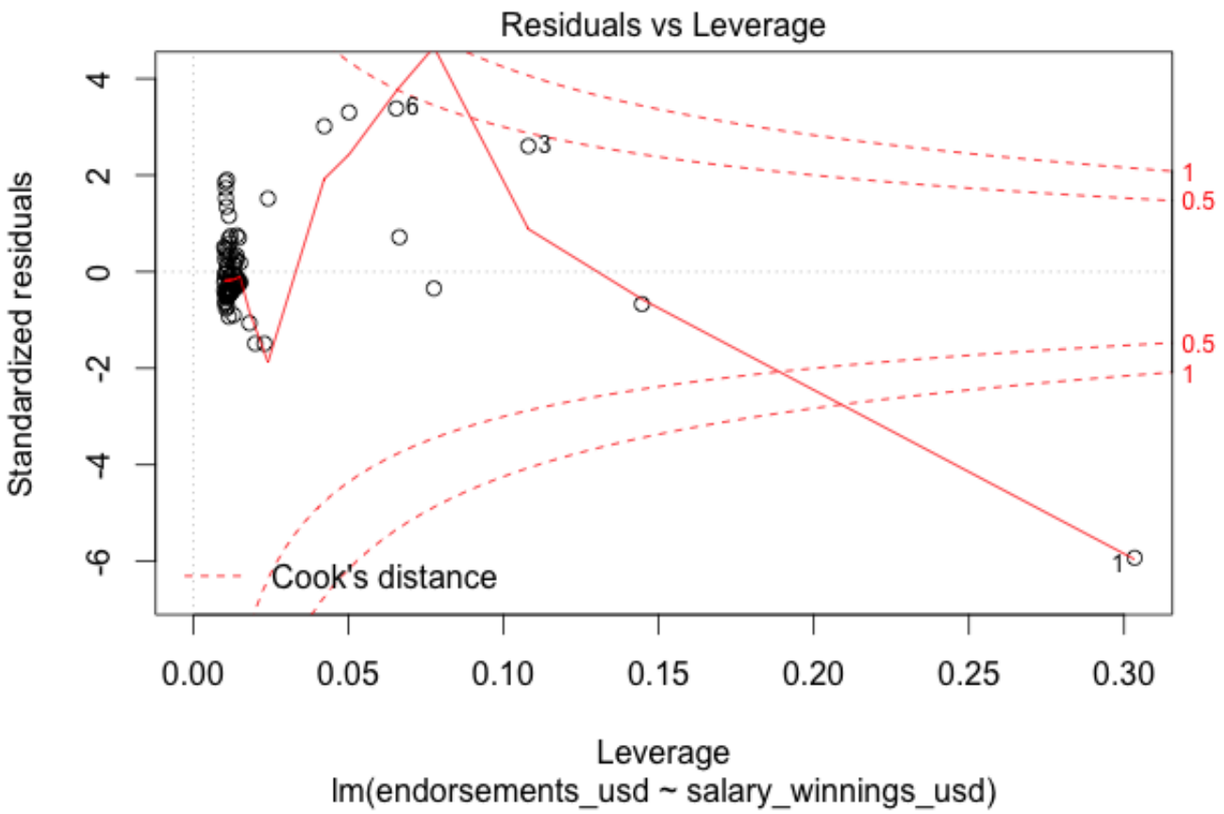
N = 11
mean of X's = 9.0
mean of Y's = 7.5
equation of regression line: Y = 3 + 0.5X
standard error of estimate of slope = 0.118
t = 4.24
sum of squares $X - \bar{X} = 110.0$
regression sum of squares = 27.50
residual sum of squares of Y = 13.75
correlation coefficient = .82
$r^2 = .67$

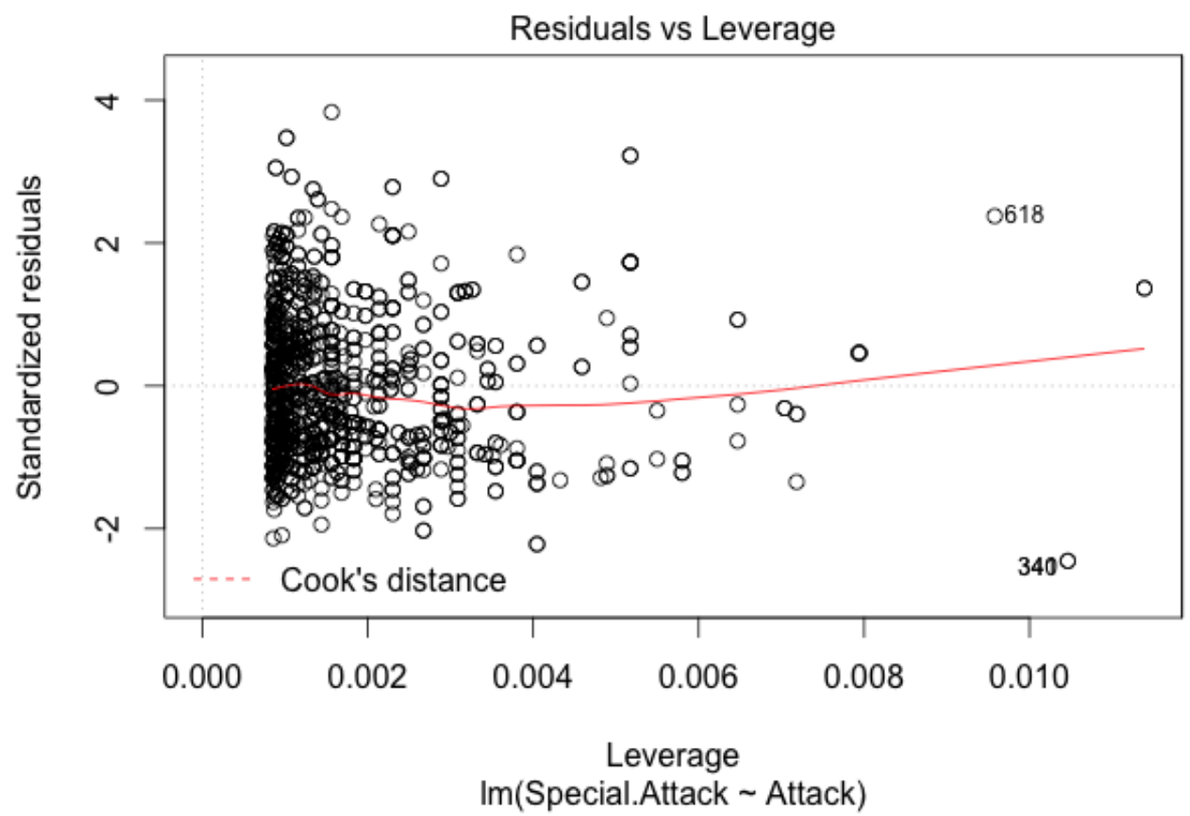
How to spot outliers?

Easy if small sample or outliers obvious

- Anscombe presents examples of both of these cases

Outliers

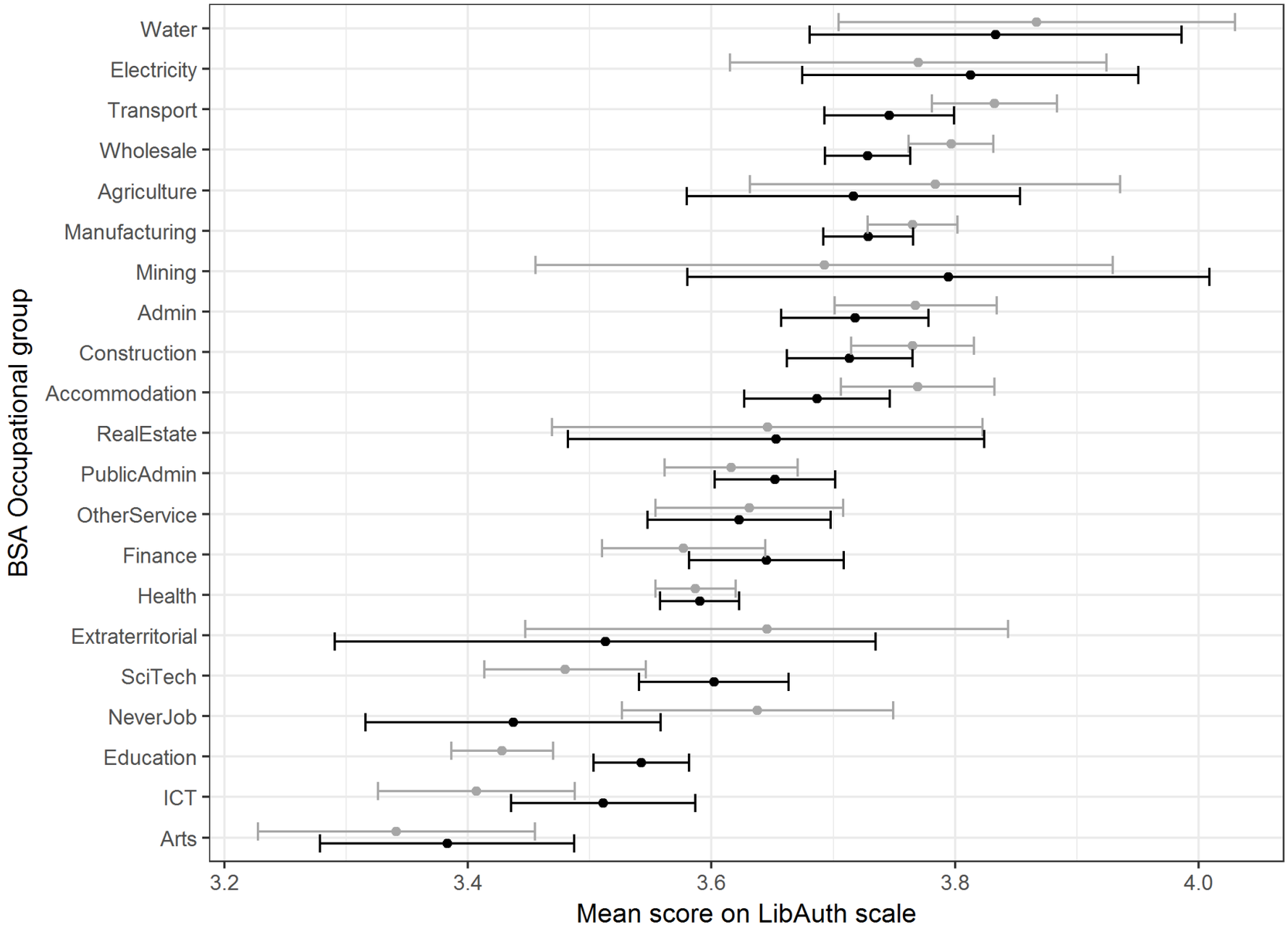




Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Sheffield Methods Institute.

Some graphics

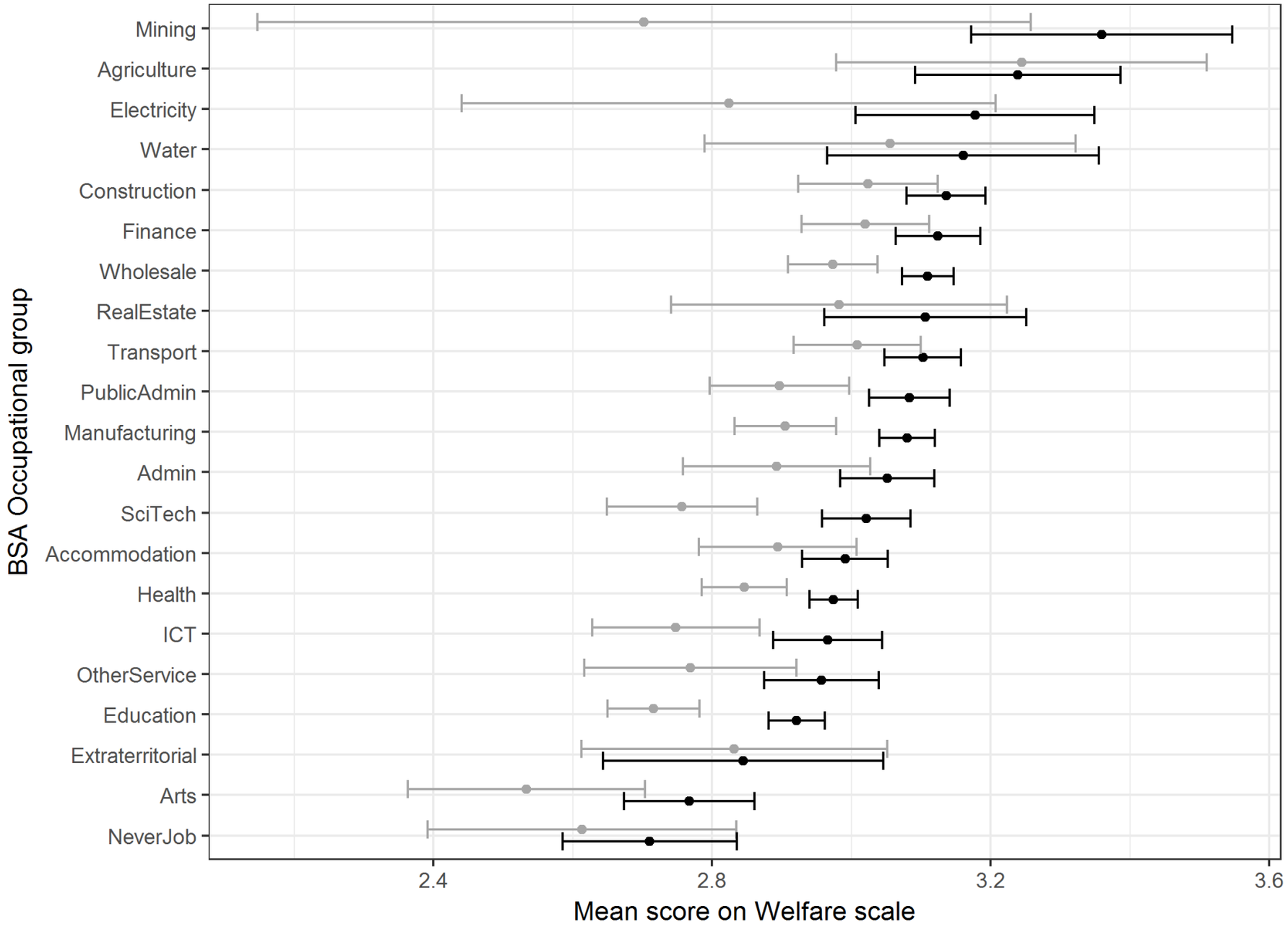


Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Social Analytics & Visualisation
Sheffield, 13/6/2022

Sheffield Methods Institute.

Some graphics



Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Social Analytics & Visualisation
Sheffield, 13/6/2022

Sheffield Methods Institute.

Next steps

Writing in RMarkdown

- will address a lot of the problems of working in Word, tweaking models, having to paste in updated tables, figures, etc
- (it'll cause some new problems, but I think it's better this way round)
- I'd recommend this both for writing documents and presentations

Some things we didn't do today

Some key graphs we didn't draw:

- maps
  - (vaguely) consistent with size in real life
  - cartograms
  - hexmaps (etc)
- networks
  - person-to-person (etc): one-mode
  - person-to-organisation (etc): two-mode
  - we're doing this later this week

Some things we didn't do today

Absolutely loads of R packages

- sf
- (gg)plotly
- igraph
- gganimate
- ggforce
- ggnetworkmap (etc)

Literally any other software

Some things we didn't do today

Reshaping data

- or any heavy-duty data cleaning
- for reshaping, have a look at pivot_longer() and pivot_wider()

Some things we didn't do today

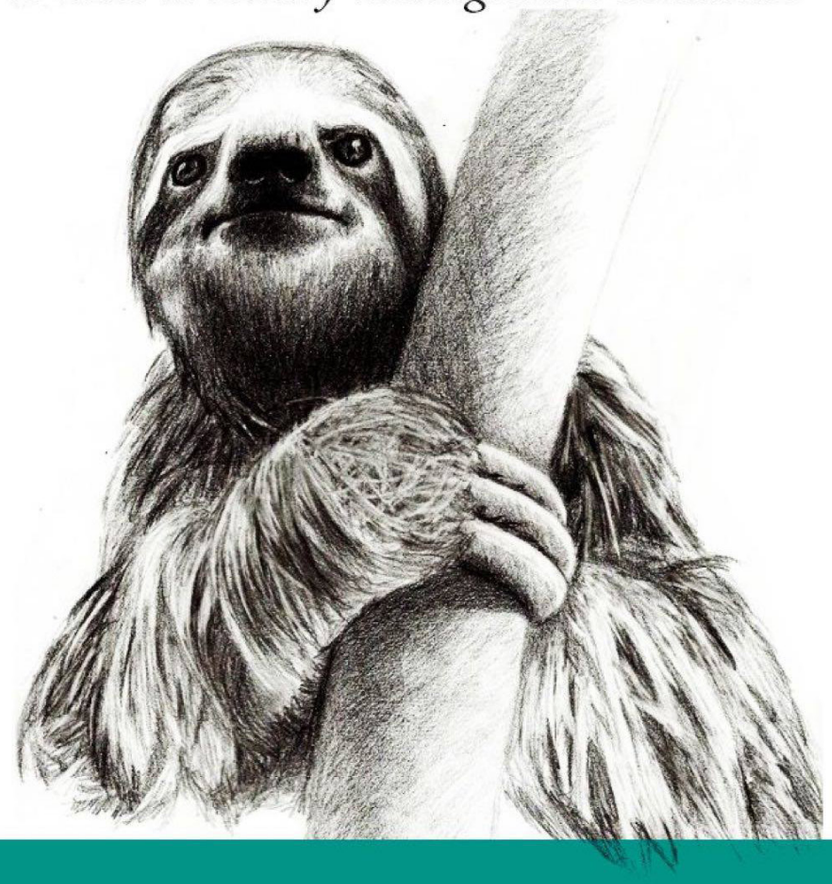
Much in the way of aesthetics

- why would you choose one graph over another?
- maybe (critically) read Tufte's "The Visual Display of Quantitative Information"

Anything on interpretation, intelligibility, or audiences more generally

- what kinds of data viz do people like?
- please read Helen Kennedy's work on this

Cutting corners to meet arbitrary management deadlines



Essential

Copying and Pasting from Stack Overflow

O'REILLY®

The Practical Developer
@ThePracticalDev

Mark Taylor
m.r.taylor@sheffield.ac.uk
@markrt

Sheffield Methods Institute.

What next?

Think about visualising everything

- any time you get new data
- any time you learn a new technique

The point is practice

- you'll spend most of your time debugging
- this is fine
- you'll end up with cool stuff

That's it!