THE STATISTICS COMMUNICATION SECTION PRESENTS

DATA-VIZ

A WORKSHOP SERIES ON DATA VISUALIZATION

YAN HOLTZ

1 **CHOOSING THE RIGHT CHART FOR YOUR DATA**

November 24th: 11.00 – 13.00

CEDRIC SCHERER

2 **DATA VISUALIZATION IN R (BEGINNER)**

December 1st: 11.00 – 13.00

3 **DATA VISUALIZATION IN R (ADVANCED)**

December 8th: 11.00 – 13.00

REGISTER AT: https://bit.ly/3GFwgCj

ORGANIZED BY

VVSOR STATISTICS COMMUNICATION

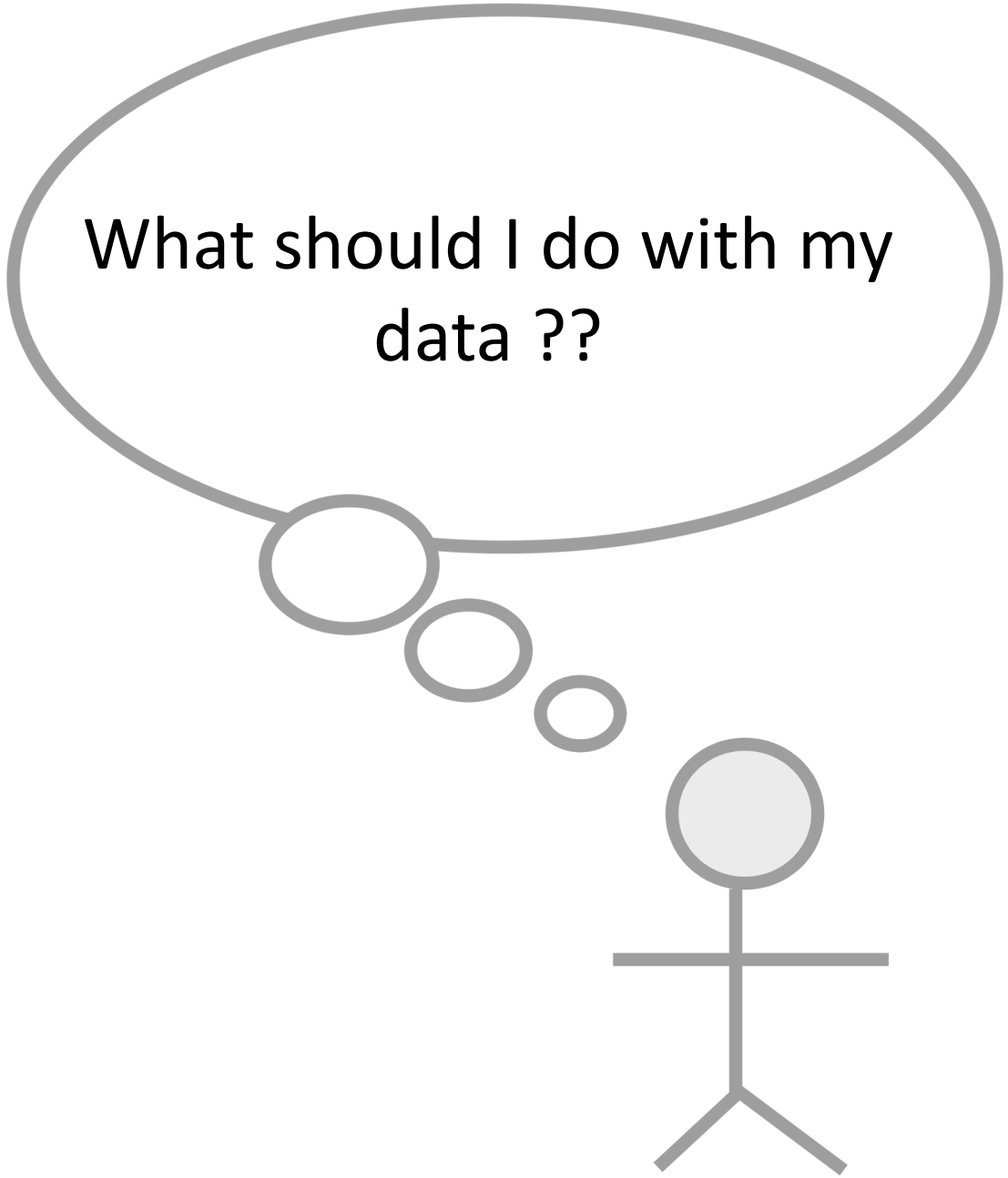
A project by Yan Holtz and Conor Healy

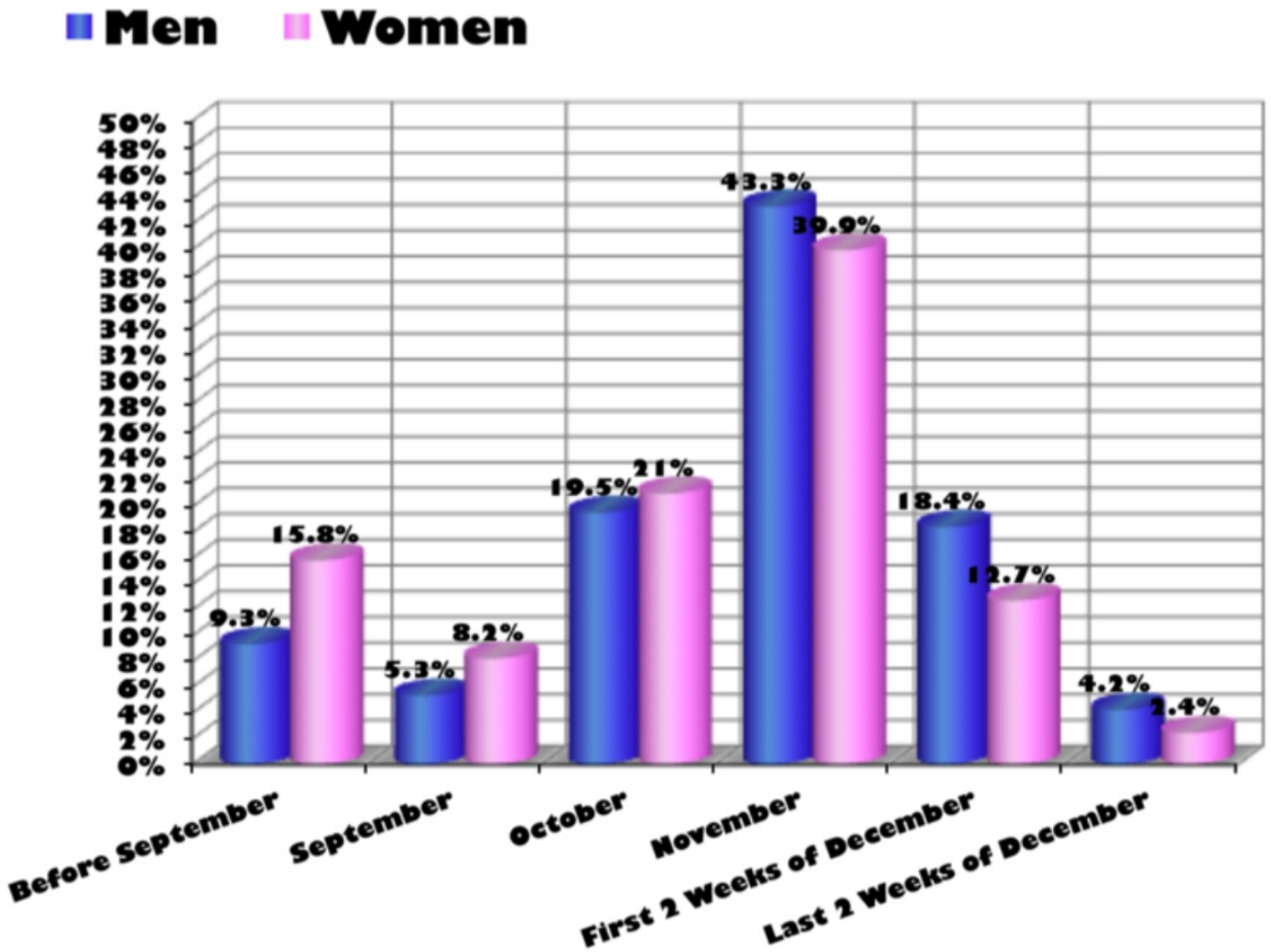
from Data to Viz

Choosing the right chart for your data

The Netherland & Remote | November 2021

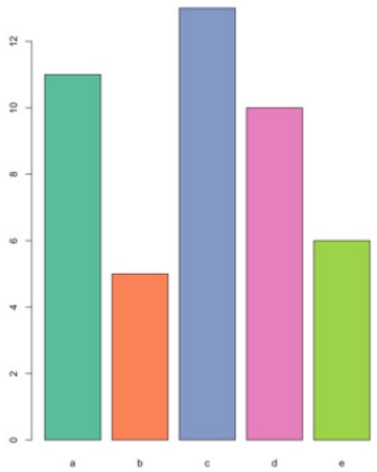
Id	feature 1
A	10
B	12
C	15
...	...





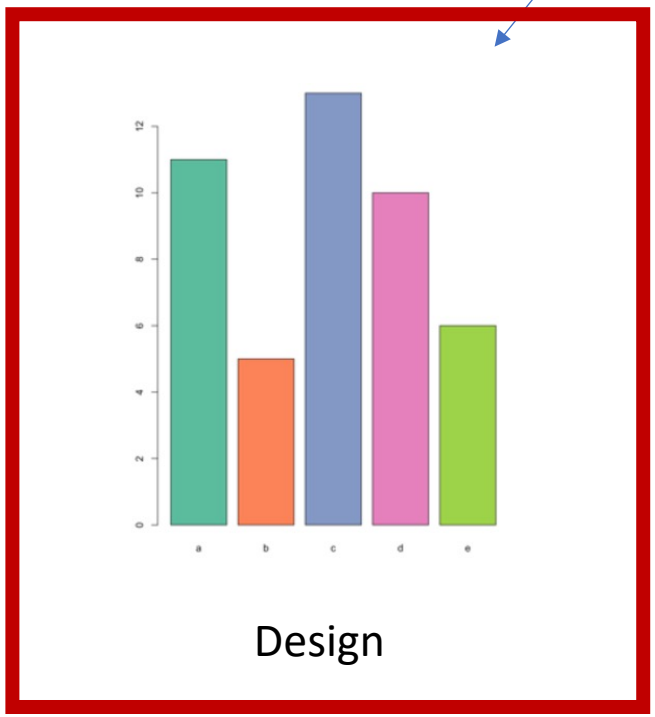
…
What about that?

Source: Storytelling with data by Cole Nussbaumer Knaflic



Design

Code



Design

Code

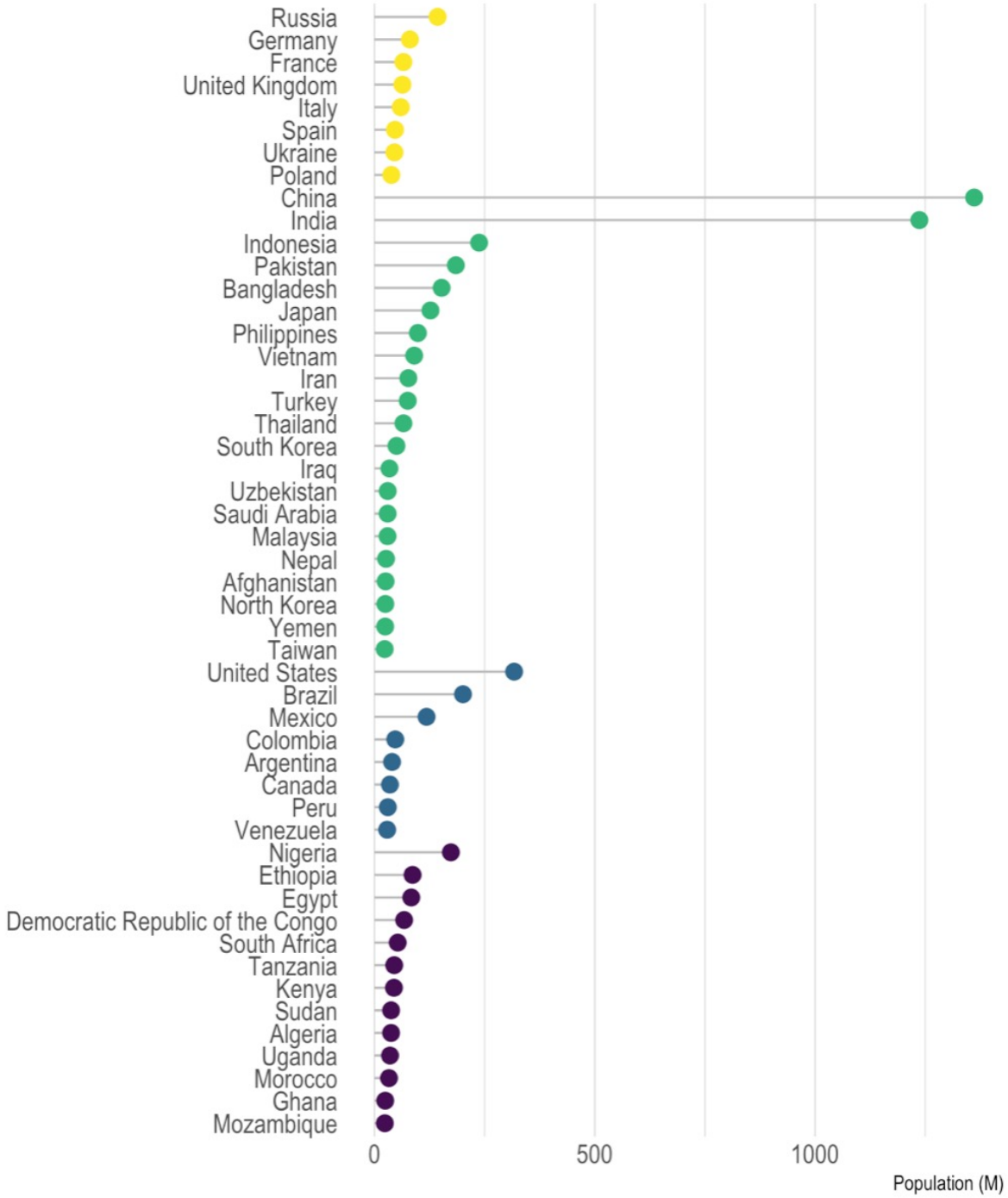
A chart, why?

World population distribution

	A	B	C	D
1	region	subregion	key	value
2	Asia	Southern Asia	Afghanistan	25500100
3	Europe	Northern Europe	Ã…land Islands	28502
4	Europe	Southern Europe	Albania	2821977
5	Africa	Northern Africa	Algeria	37900000
6	Oceania	Polynesia	American Samoa	55519
7	Europe	Southern Europe	Andorra	76246
8	Africa	Middle Africa	Angola	20609294
9	Americas	Caribbean	Anguilla	13452
10			Antarctica	-1
11	Americas	Caribbean	Antigua and Barbuda	86295
12	Americas	South America	Argentina	40117096
13	Asia	Western Asia	Armenia	3024100
14	Americas	Caribbean	Aruba	101484
15	Oceania	Australia and New Zealand	Australia	23254142
16	Europe	Western Europe	Austria	8501502
17	Asia	Western Asia	Azerbaijan	9235100
18	Americas	Caribbean	Bahamas	-1
19	Asia	Western Asia	Bahrain	1234571
20	Asia	Southern Asia	Bangladesh	152518015
21	Americas	Caribbean	Barbados	274200
22	Europe	Eastern Europe	Belarus	9465500
23	Europe	Western Europe	Belgium	11175653
24	Americas	Central America	Belize	312971
25	Africa	Western Africa	Benin	10323000
26	Americas	Northern America	Bermuda	64237
27	Asia	Southern Asia	Bhutan	740990
28	Americas	South America	Bolivia	10027254
29	Americas	Caribbean	Bonaire	-1
30	Europe	Southern Europe	Bosnia and Herzegovina	3791622
31	Africa	Southern Africa	Botswana	2024904
32			Bouvet Island	-1
33	Americas	South America	Brazil	201032714
34	Africa	Eastern Africa	British Indian Ocean Territory	-1
35	Americas	Caribbean	British Virgin Islands	29537

- Biggest country?
- Rank?
- Distribution by continent?

“Picture superiority effect”



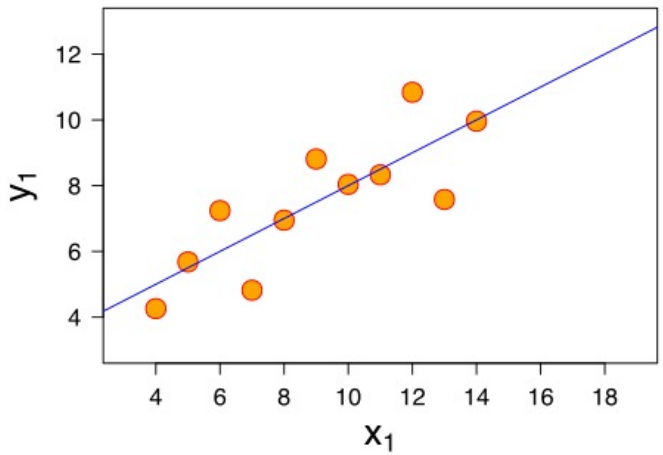
OK, but we have summary statistics!

Y = 3 + 0.5x
Cor = 0.8

Mean(x) = 9
Var(x) = 11

Mean(Y) = 7.5
Var(Y) = 4.1

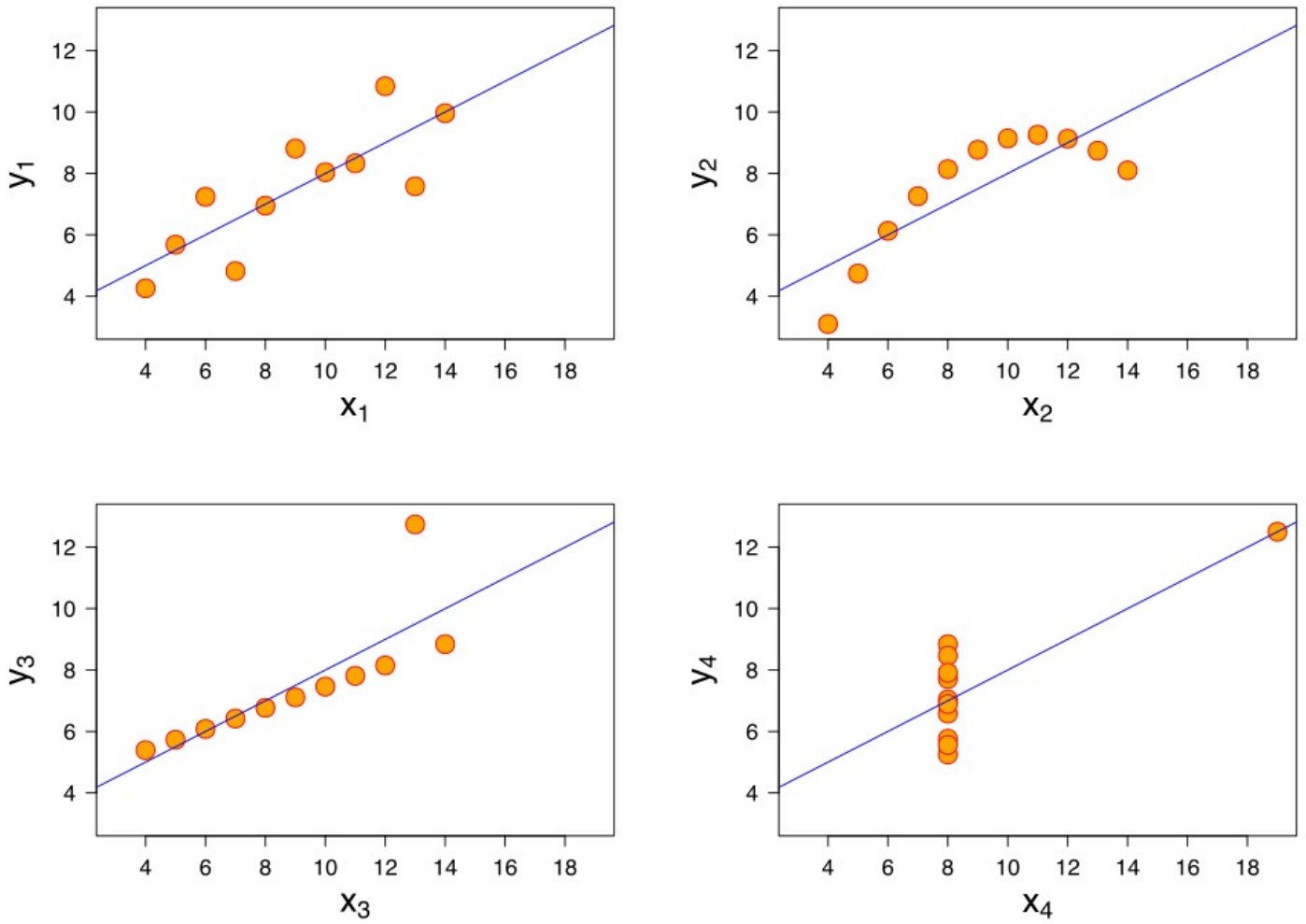
Anscombe et al. 1973



Y = 3 + 0.5x
Cor = 0.8

Mean(x) = 9
Var(x) = 11

Mean(Y) = 7.5
Var(Y) = 4.1



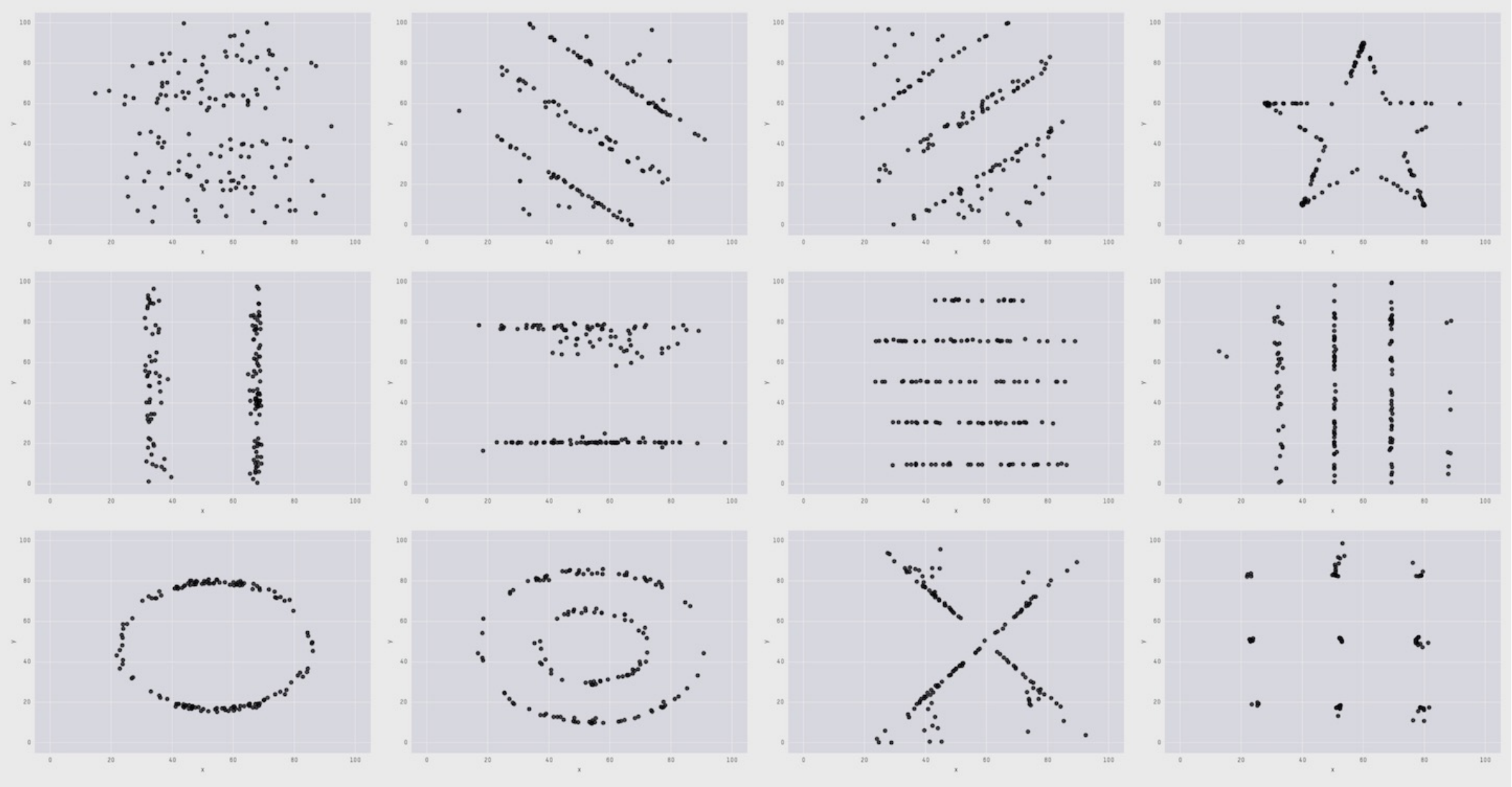
Y = 3 + 0.5x
Cor = 0.8

Mean(x) = 9
Var(x) = 11

Mean(Y) = 7.5
Var(Y) = 4.1

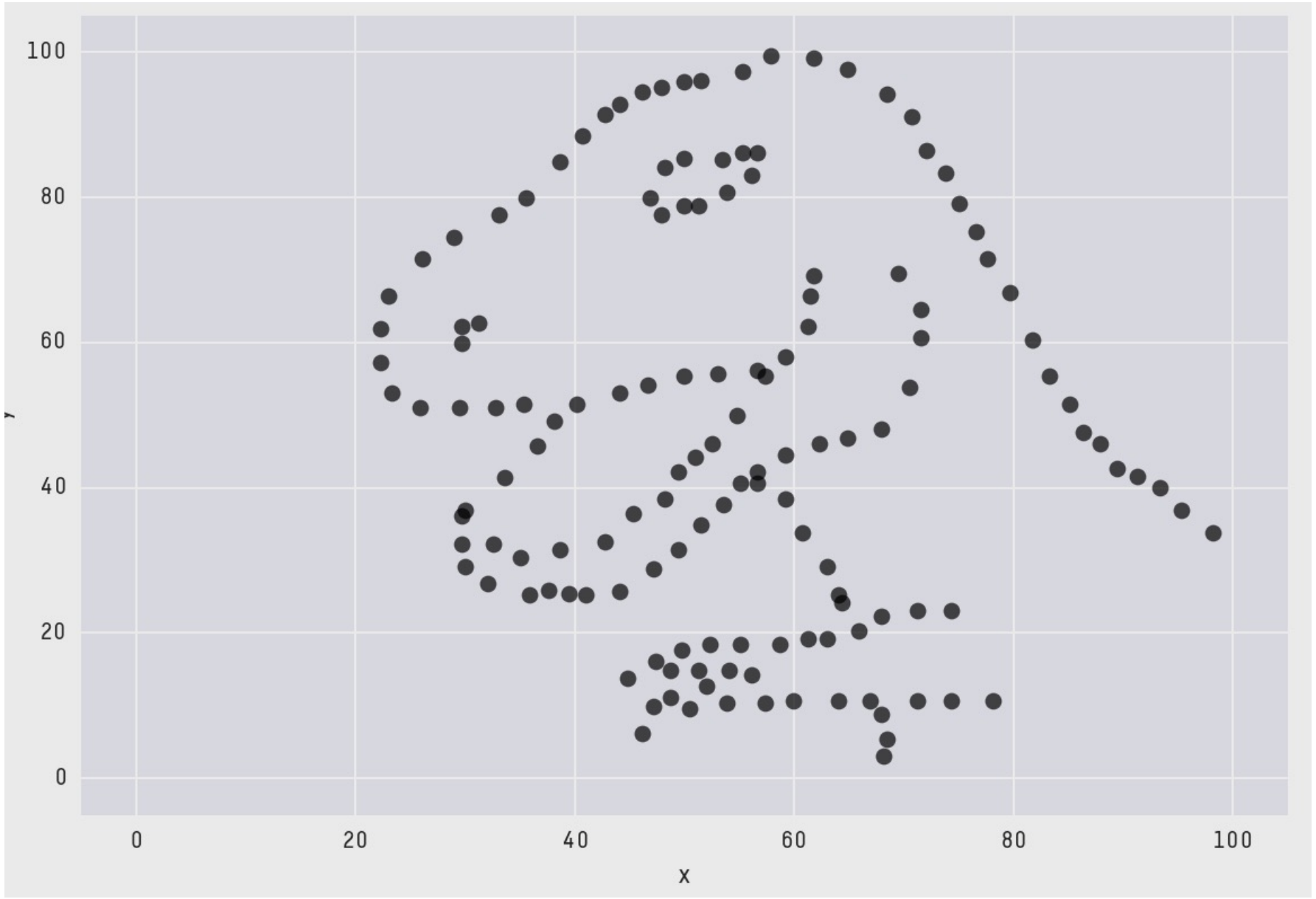
Anscombe's quartet

Anscombe et al. 1973



The Datasaurus Dozen

Cairo , 2017
Matejka & Fitzmaurice, 2017



The Datasaurus Dozen

Cairo , 2017
Matejka & Fitzmaurice, 2017

Looking for a chart ?

What you **can** do

What you **should** do

Caveats to avoid

How to build it

WHAT YOU CAN DO

A classification of chart types based on data input format

Scatterplot

Scatterplot

2d density chart

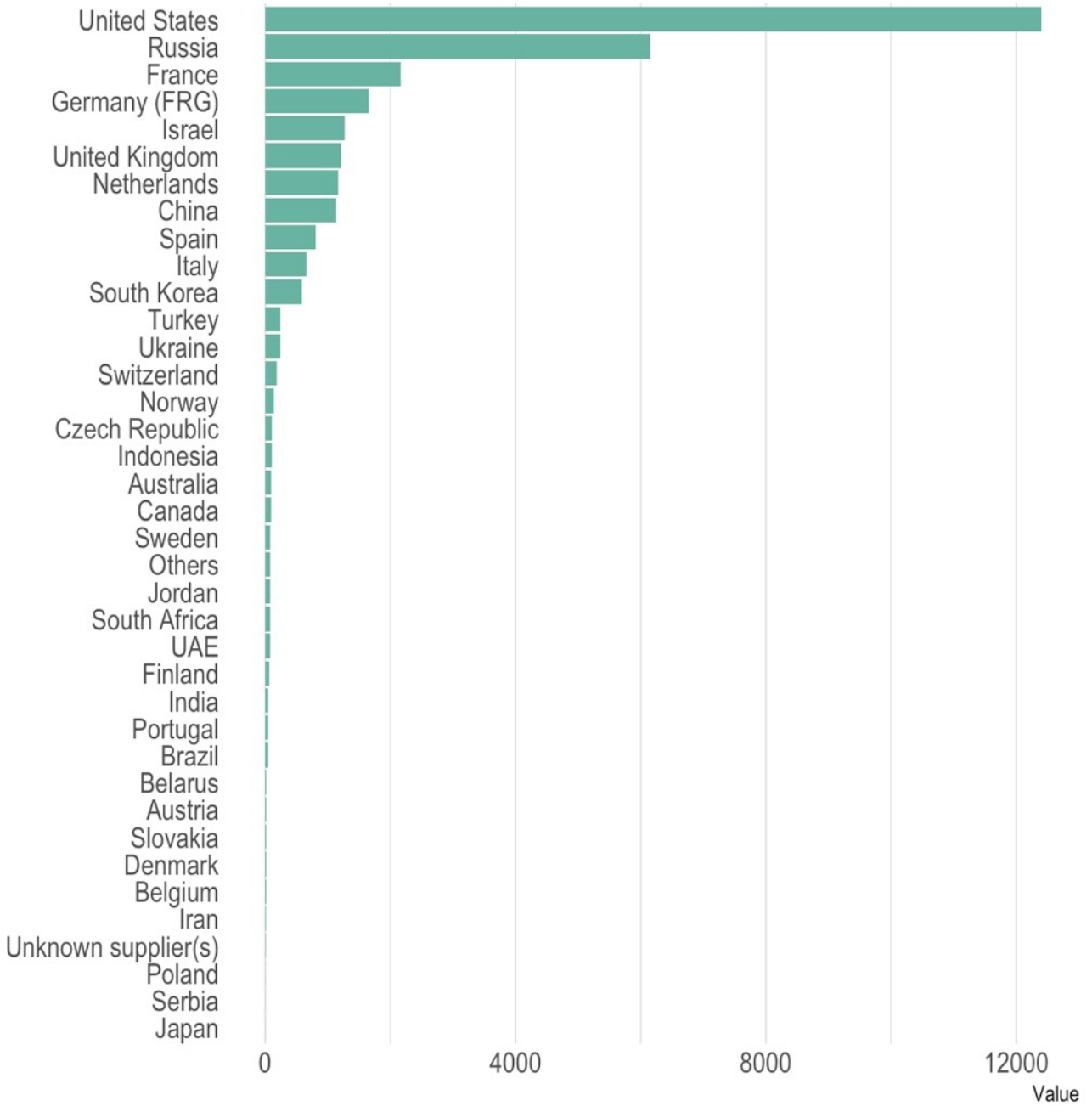
Who sells more weapons ?

Country	Value
United States	12394
Russia	6148
Germany (FRG)	1653
France	2162
United Kingdom	1214
China	1131

?

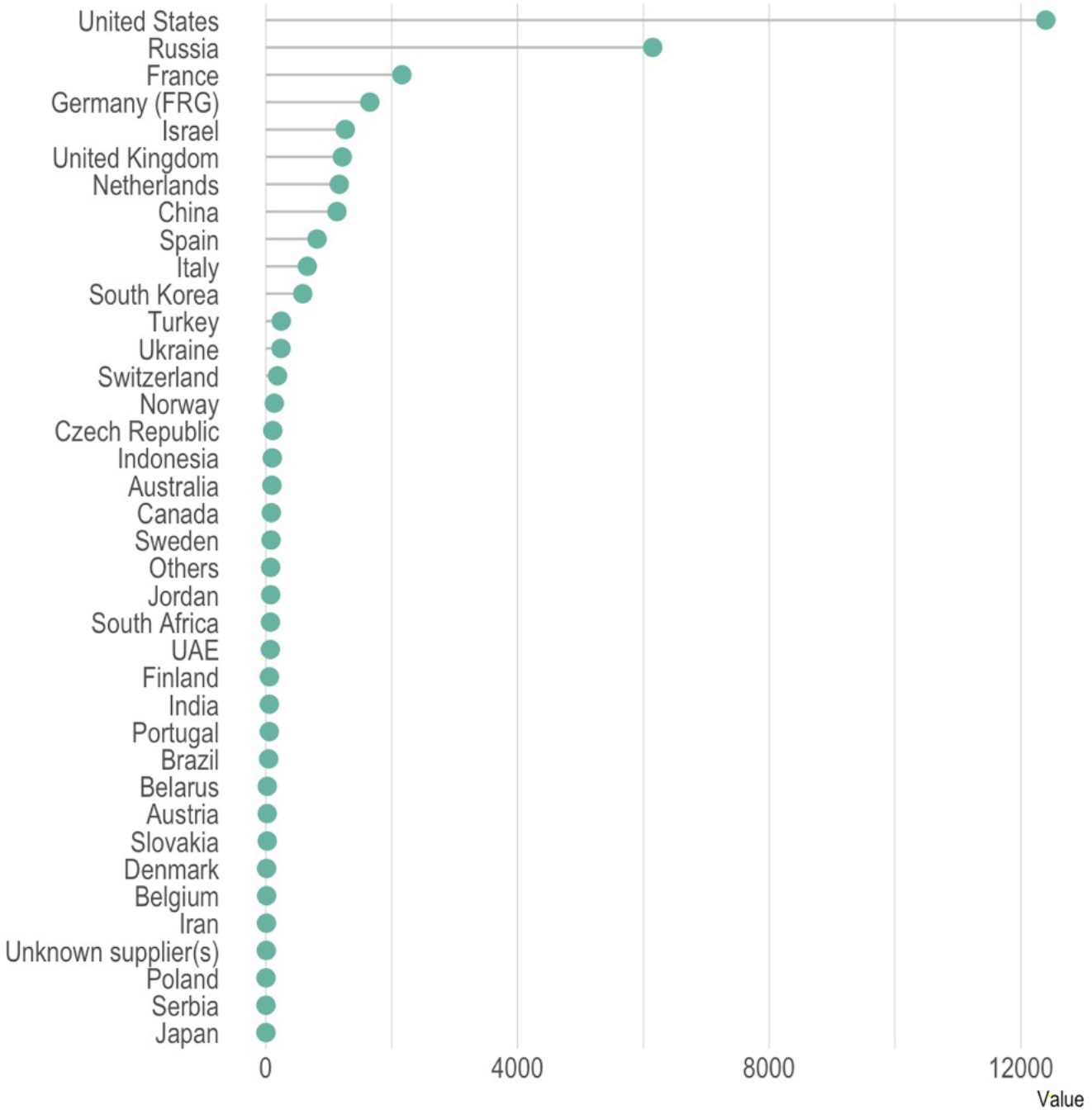
Who sells more weapons ?

Country	Value
United States	12394
Russia	6148
Germany (FRG)	1653
France	2162
United Kingdom	1214
China	1131



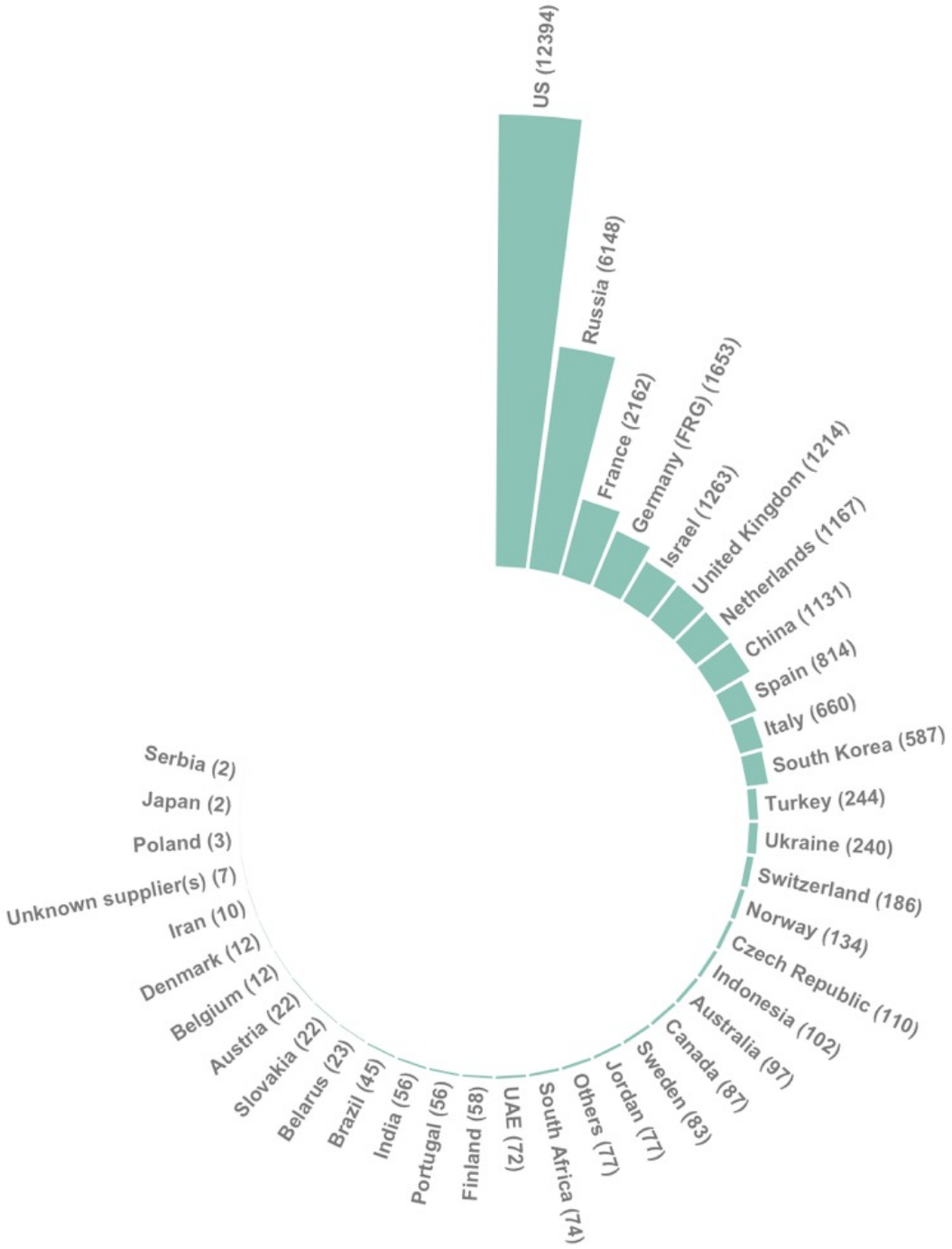
Who sells more weapons ?

Country	Value
United States	12394
Russia	6148
Germany (FRG)	1653
France	2162
United Kingdom	1214
China	1131



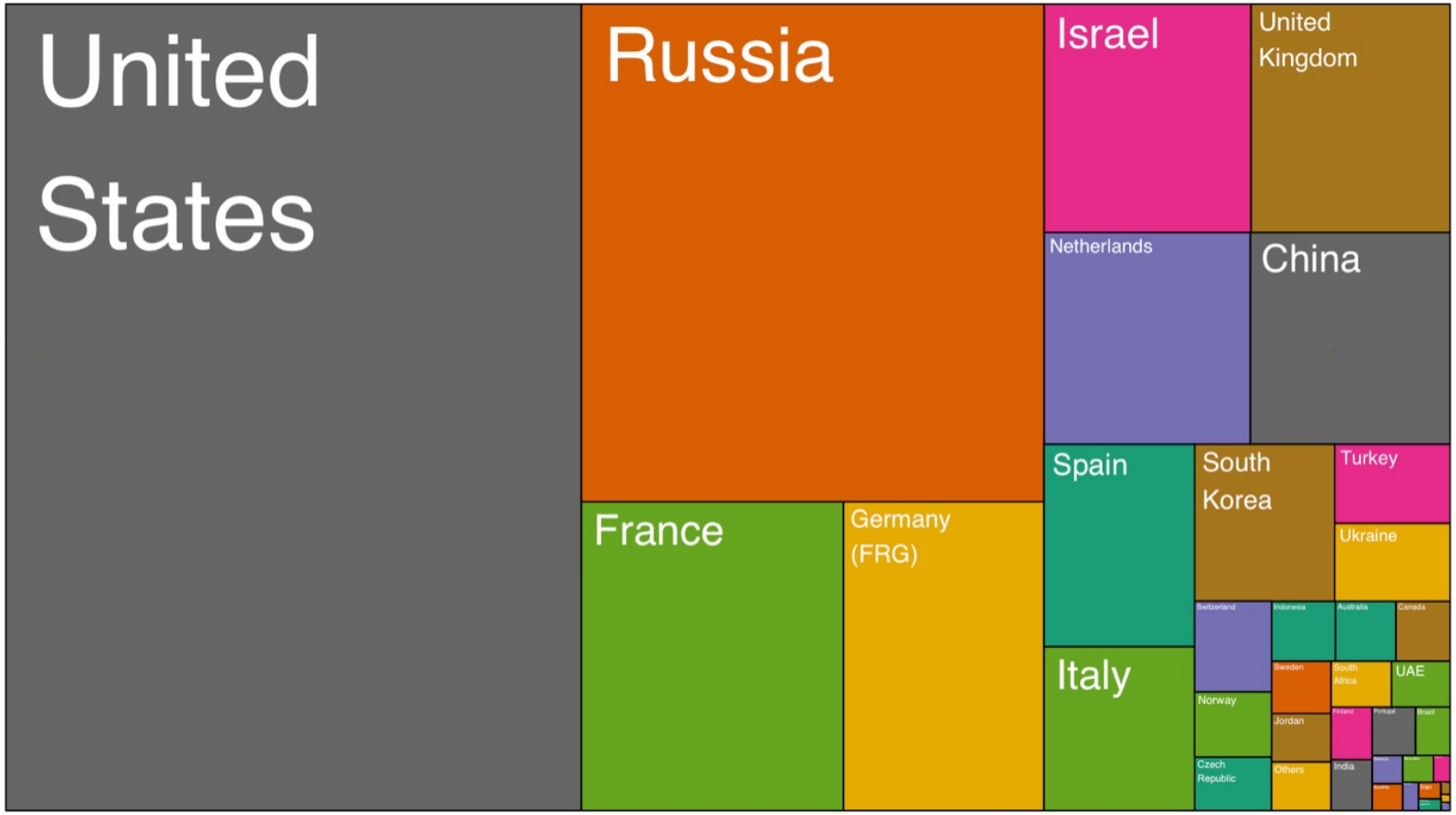
Who sells more weapons ?

Country	Value
United States	12394
Russia	6148
Germany (FRG)	1653
France	2162
United Kingdom	1214
China	1131



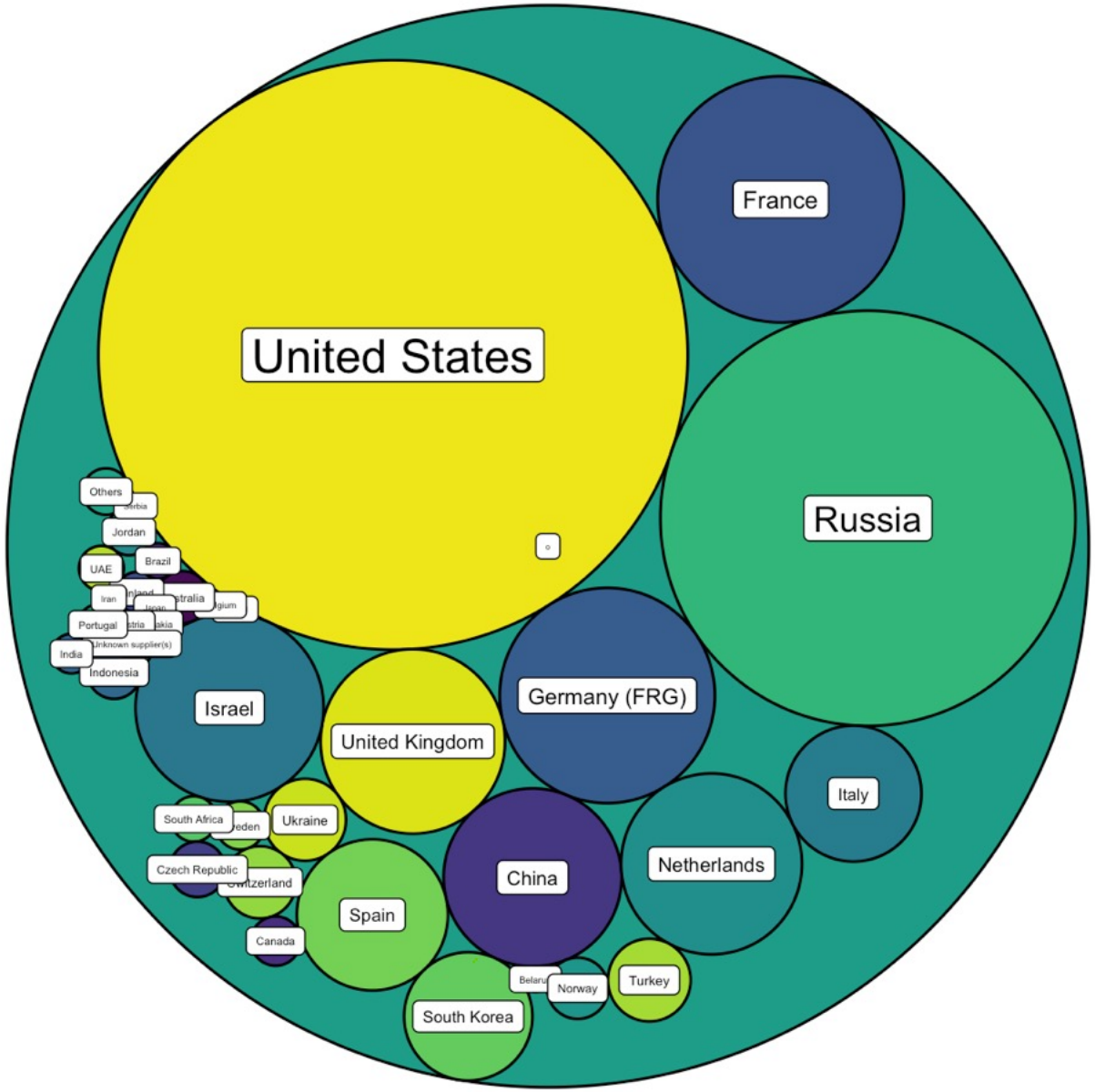
Who sells more weapons ?

Country	Value
United States	12394
Russia	6148
Germany (FRG)	1653
France	2162
United Kingdom	1214
China	1131



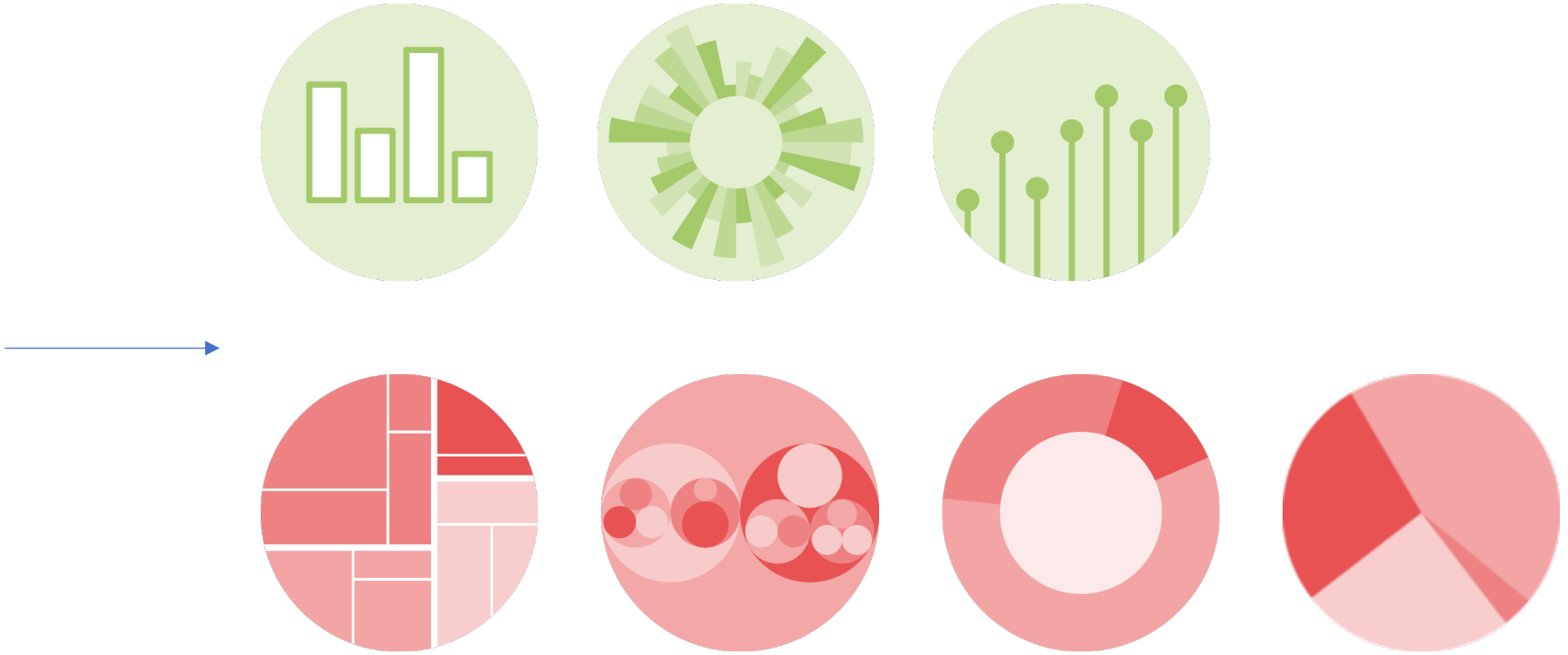
Who sells more weapons ?

Country	Value
United States	12394
Russia	6148
Germany (FRG)	1653
France	2162
United Kingdom	1214
China	1131



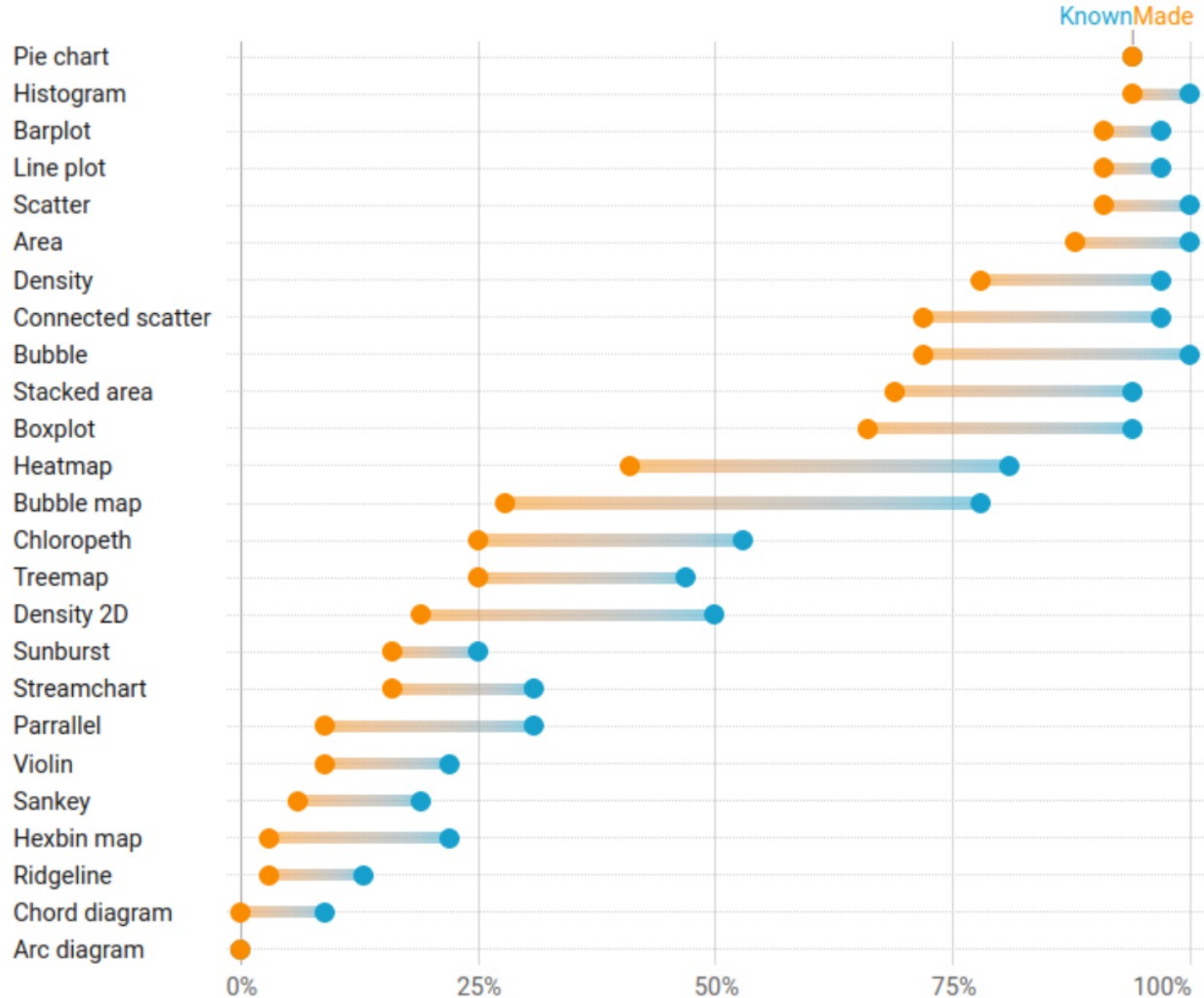
Who sells more weapons ?

Country	Value
United States	12394
Russia	6148
Germany (FRG)	1653
France	2162
United Kingdom	1214
China	1131



Charts you made and charts you know 18/12

Dataviz training session 18/12



Who sells more weapons ?

Country	Value
United States	12394
Russia	6148
Germany (FRG)	1653
France	2162
United Kingdom	1214
China	1131

1 observation per group

1 categorical variable

1 numerical variable

Perception of probability

text	value
Improbable	33
Almost Certainly	98
Likely	60
Almost Certainly	98
Unlikely	10
Probably Not	25
About Even	50
Probably	75

Perception of probability

text	value
Improbable	33
Almost Certainly	98
Likely	60
Almost Certainly	98
Unlikely	10
Probably Not	25
About Even	50
Probably	75

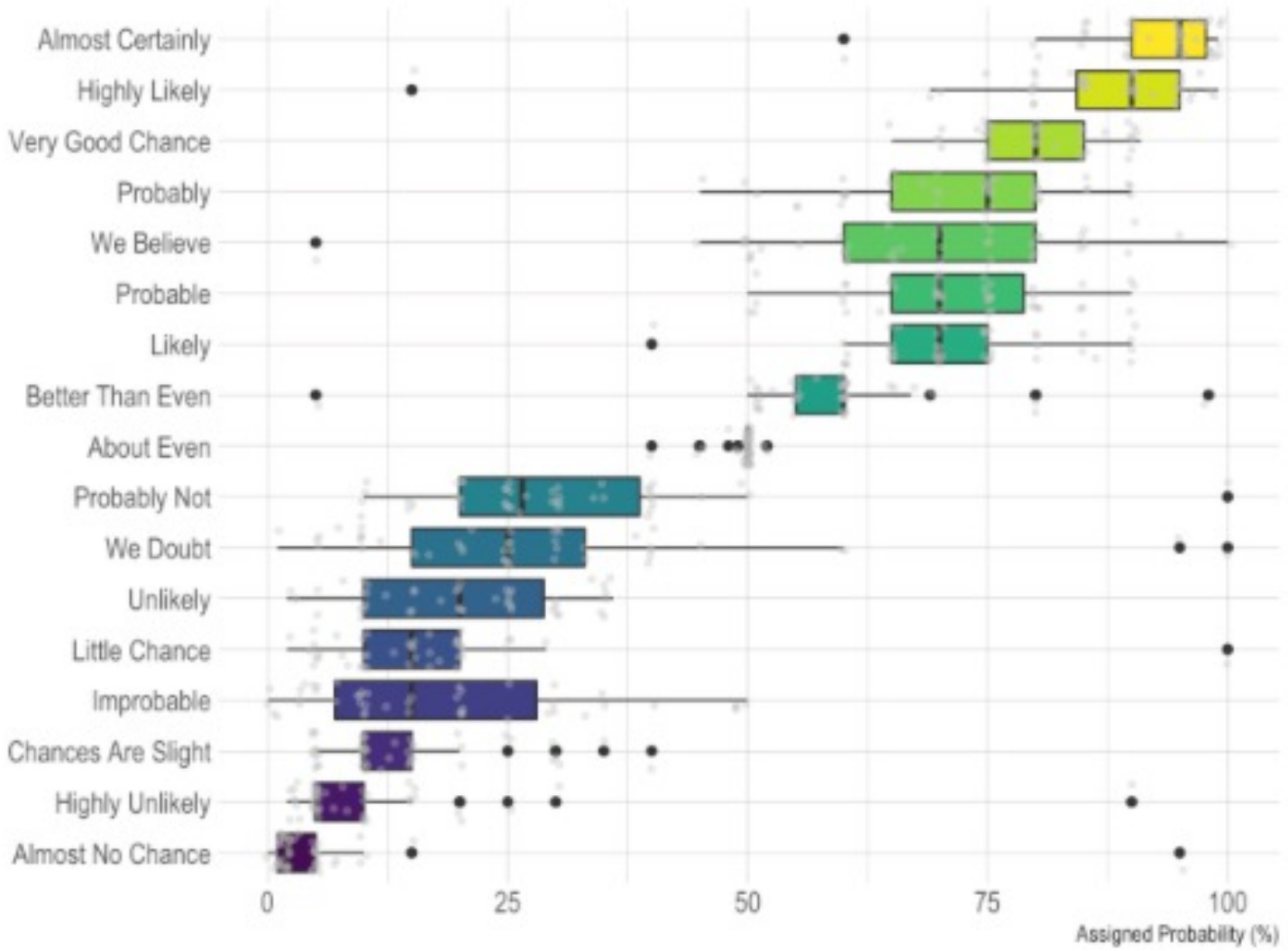
Several observations per group

1 categorical variable

1 numeric variable

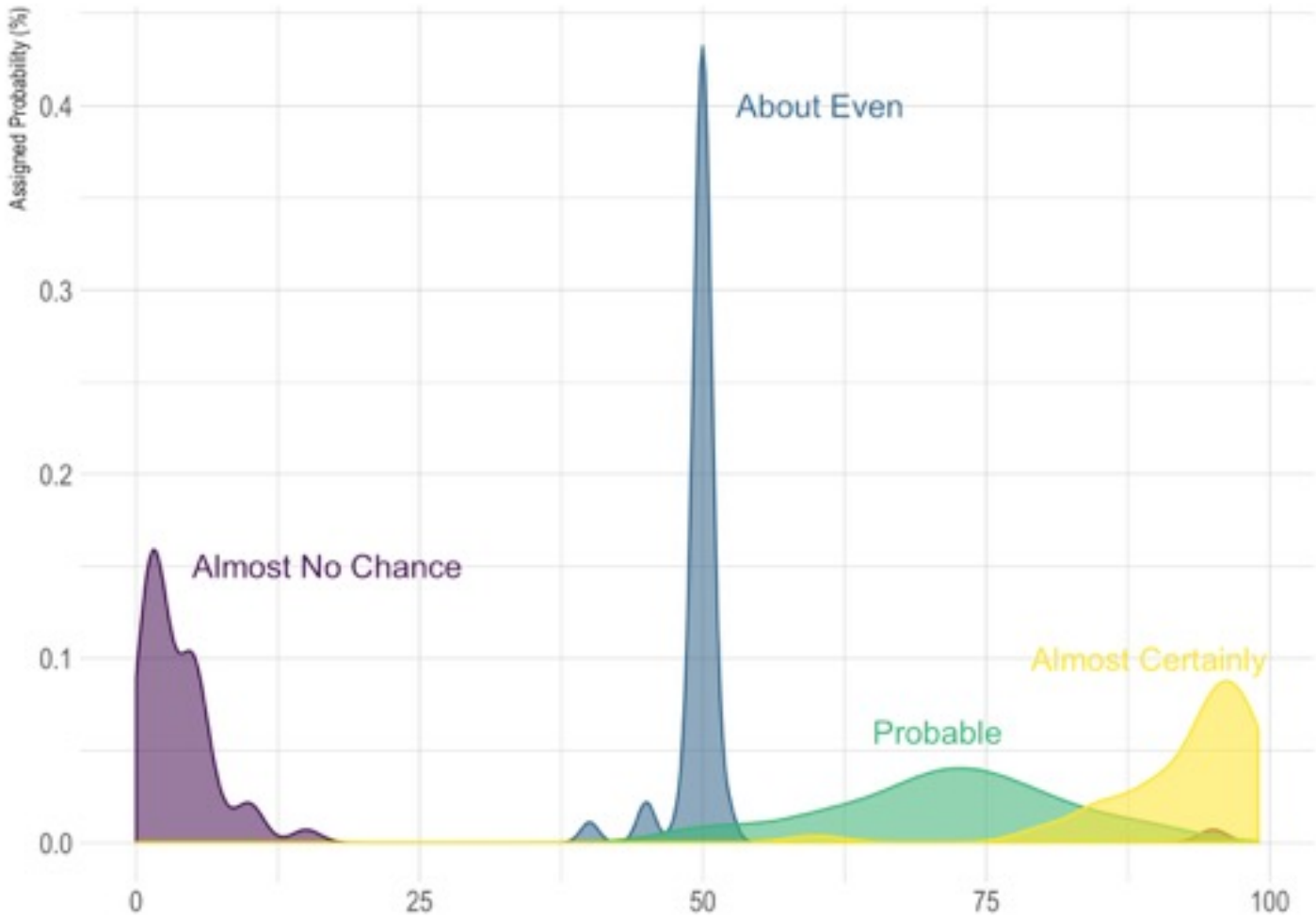
Perception of probability

text	value
Improbable	33
Almost Certainly	98
Likely	60
Almost Certainly	98
Unlikely	10
Probably Not	25
About Even	50
Probably	75



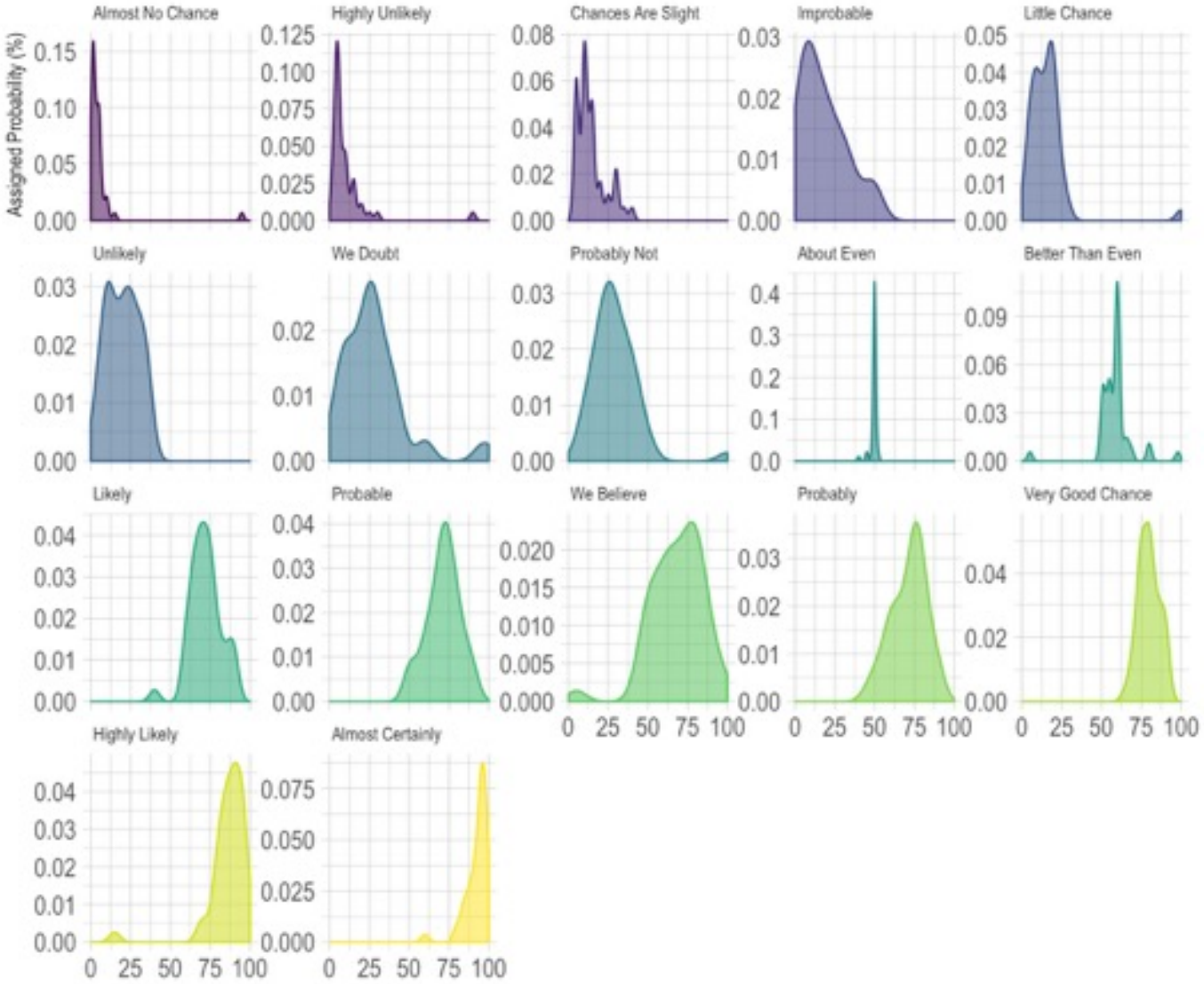
Perception of probability

text	value
Improbable	33
Almost Certainly	98
Likely	60
Almost Certainly	98
Unlikely	10
Probably Not	25
About Even	50
Probably	75



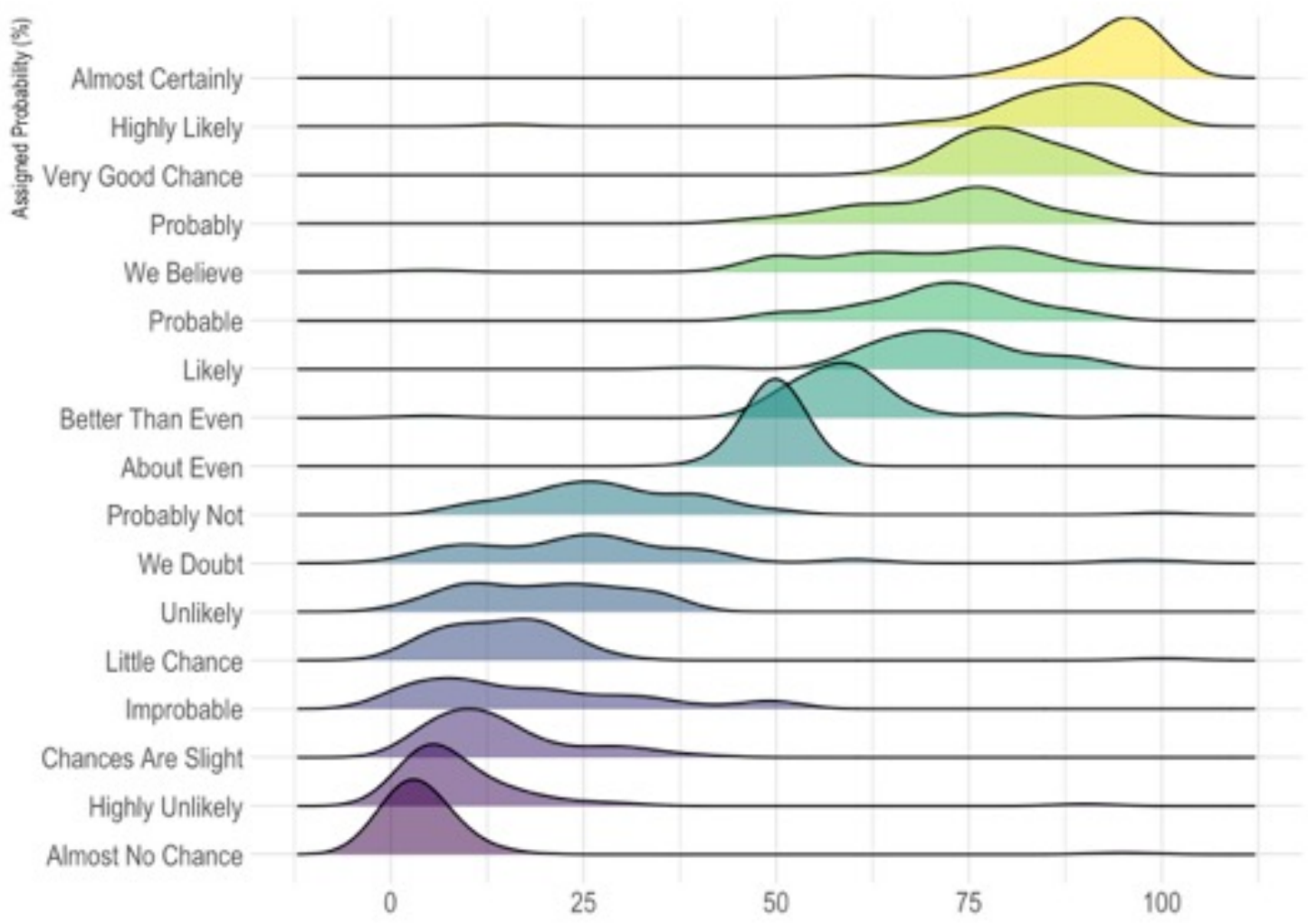
Perception of probability

text	value
Improbable	33
Almost Certainly	98
Likely	60
Almost Certainly	98
Unlikely	10
Probably Not	25
About Even	50
Probably	75



Perception of probability

text	value
Improbable	33
Almost Certainly	98
Likely	60
Almost Certainly	98
Unlikely	10
Probably Not	25
About Even	50
Probably	75



So…

- Knowing the **possibilities** is the **first step** in chart choice
- Not easy to know **all** the **chart types**
- Hard to figure out options **from a dataset**

→ Let’s build a decision tree

from Data to Viz

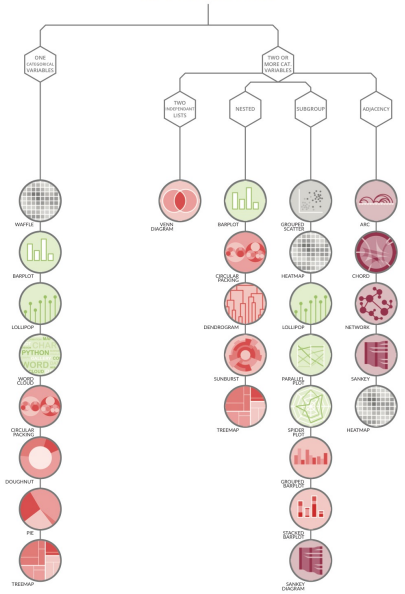
'From Data to Viz' is a classification of chart types based on input data format. It will help you find the perfect chart in three simple steps :

1. Identify what type of data you have.
2. Go to the corresponding decision tree and follow it down to a set of possible charts.
3. Choose the chart from the set that will suit your data and your needs best.

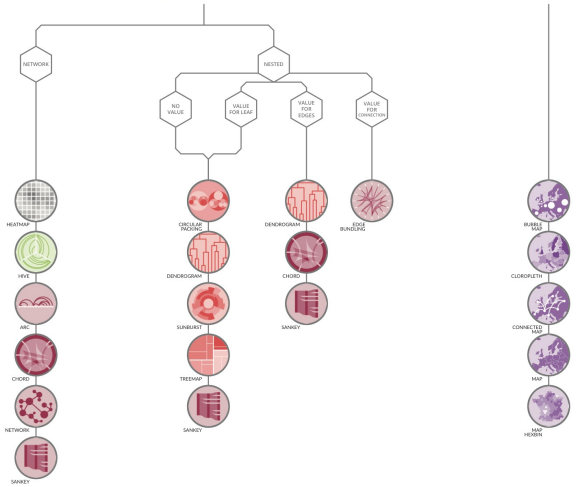
Dataviz is a world with endless possibilities and this project does not claim to be exhaustive. However it should provide you with a good starting point. For an interactive version and much more, visit:

data-to-viz.com

CATEGORIC

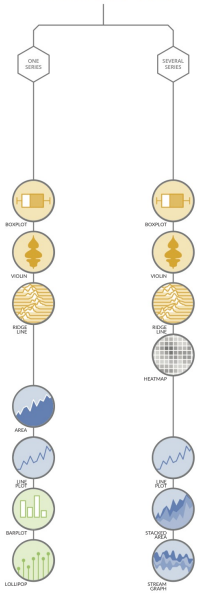


RELATIONAL



MAP

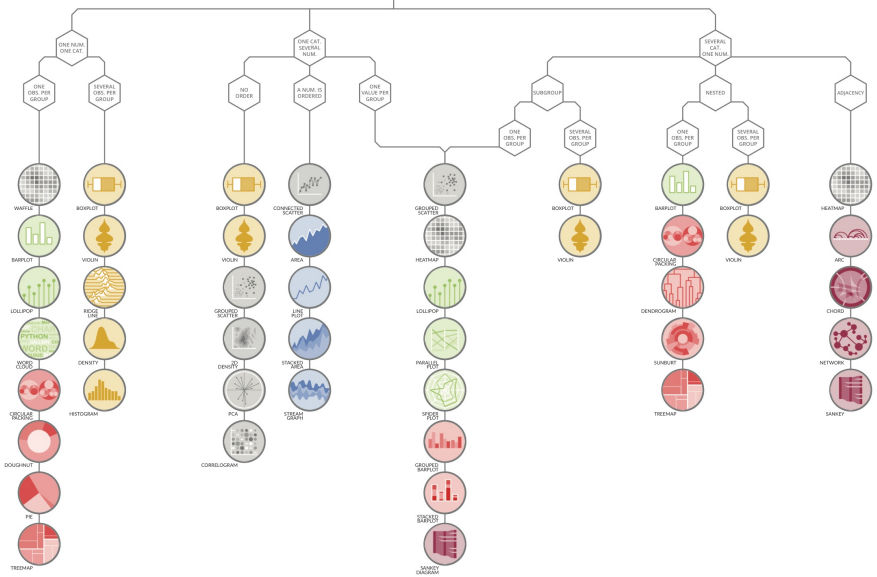
TIME SERIES



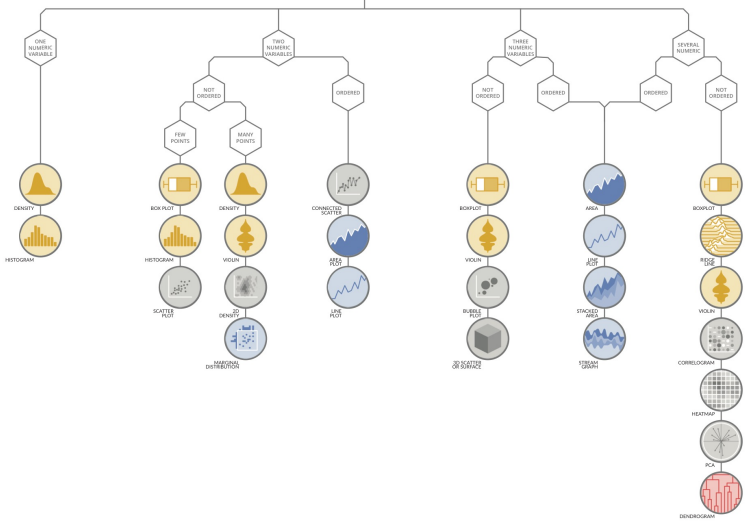
WHAT DO YOU WANT TO SHOW ?

- Distribution
- Correlation
- Ranking
- Part of a whole
- Evolution
- Maps
- Flow

CATEGORIC AND NUMERIC



NUMERIC



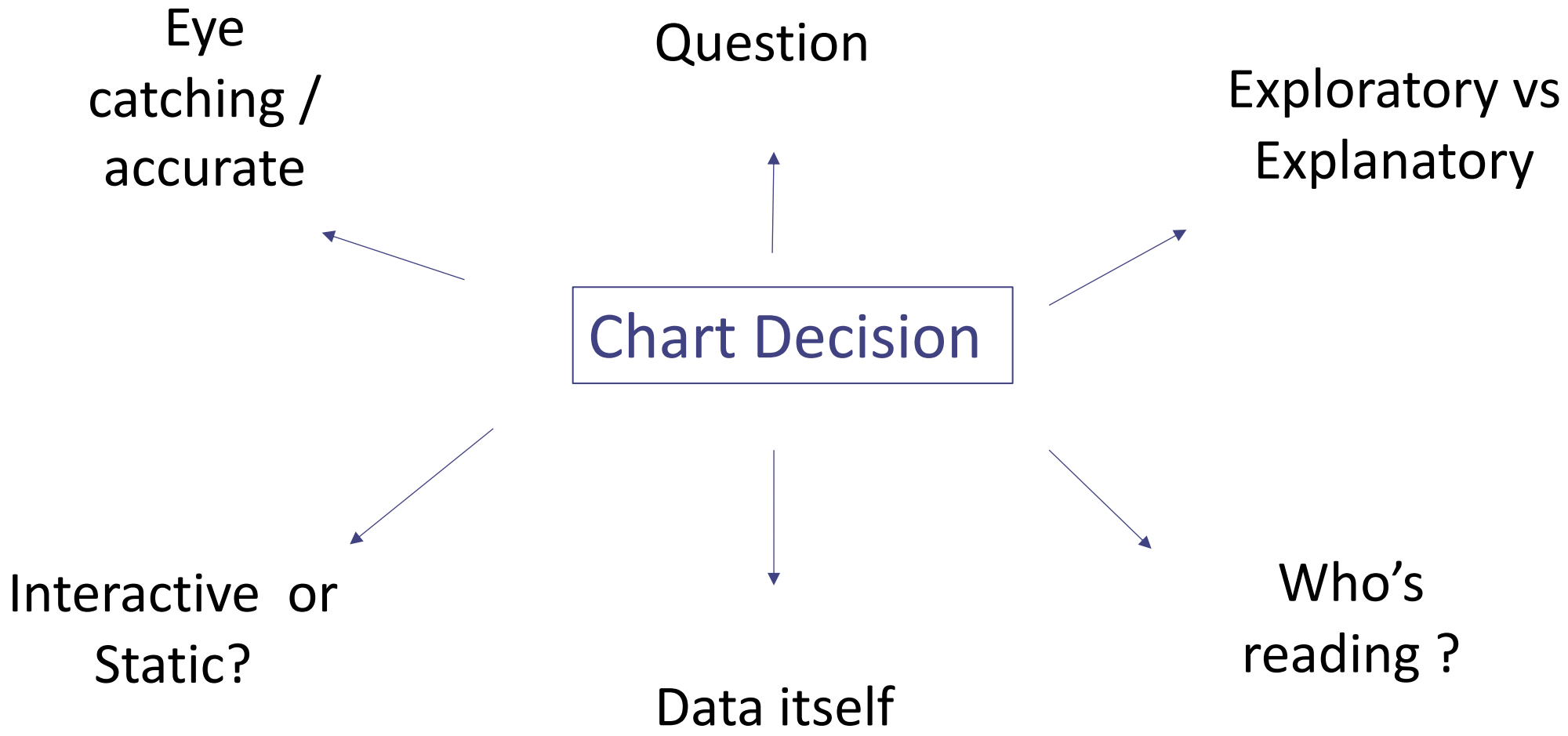
Data-to-viz.com

There are no limits in dataviz!

https://xeno.graphics/

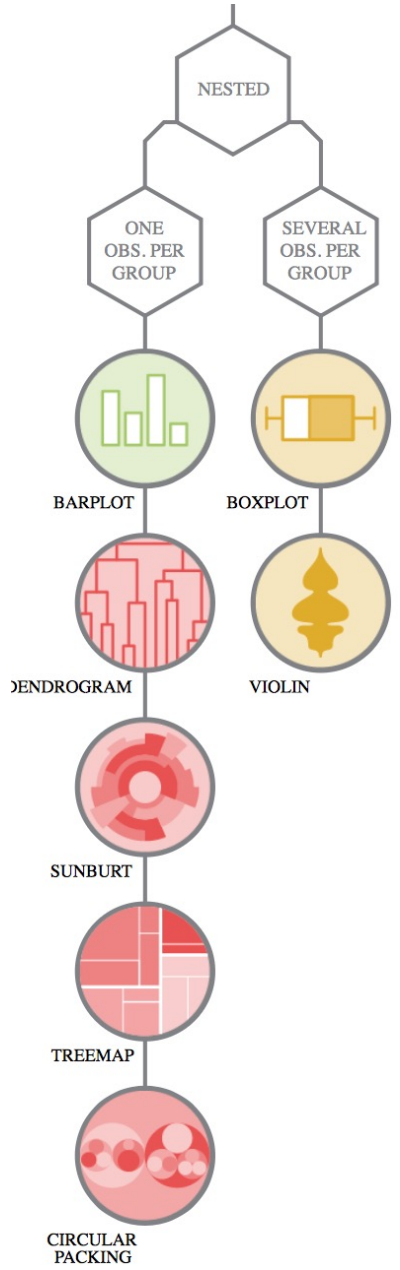
WHAT YOU SHOULD DO

About 20 examples of storytelling with data


Eye catching / accurate
Question
Exploratory vs Explanatory
Chart Decision
Interactive or Static?
Data itself
Who's reading ?

WHAT DO YOU WANT TO SHOW ?

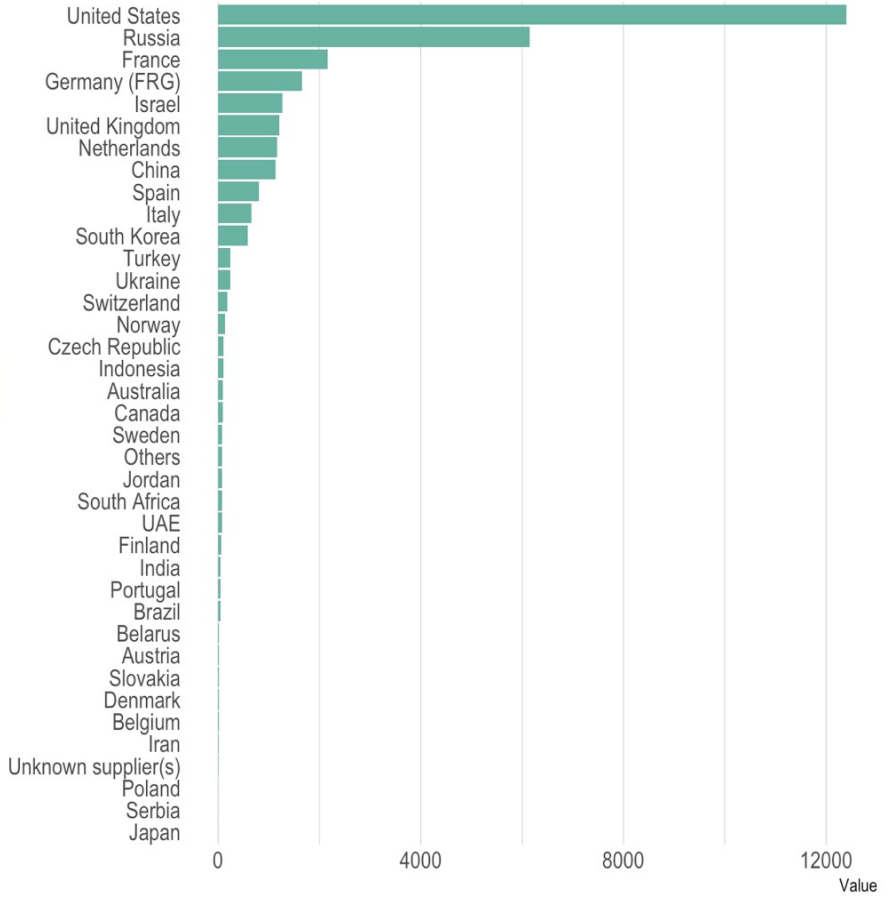
- Distribution
- Correlation
- Ranking
- Part of a whole
- Evolution
- Maps
- Flow



Who sells more weapons ?

Country	Value
United States	12394
Russia	6148
Germany (FRG)	1653
France	2162
United Kingdom	1214
China	1131

Barplot



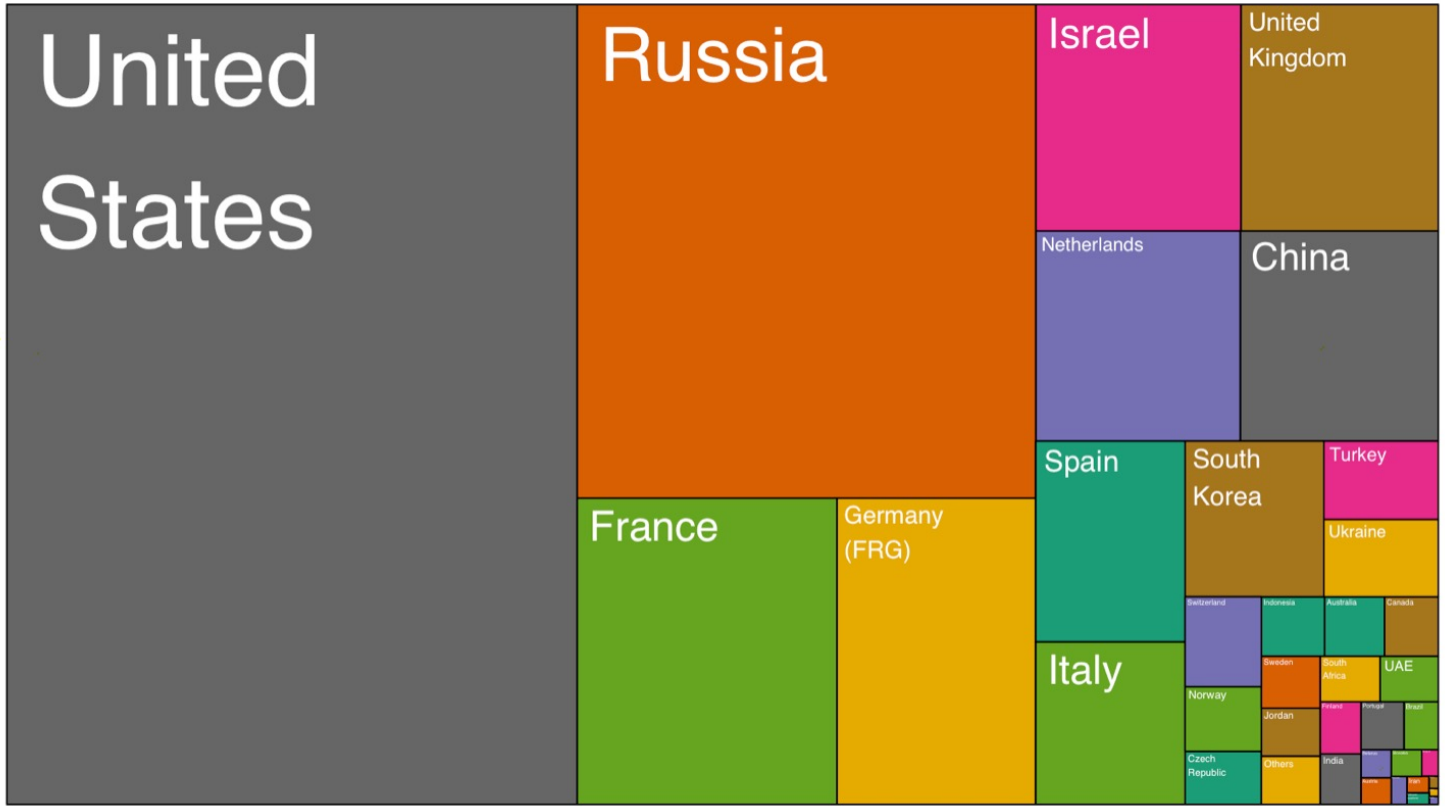
South Korea > Spain ?

% of total sold by the US ?

Who sells more weapons ?

Country	Value
United States	12394
Russia	6148
Germany (FRG)	1653
France	2162
United Kingdom	1214
China	1131

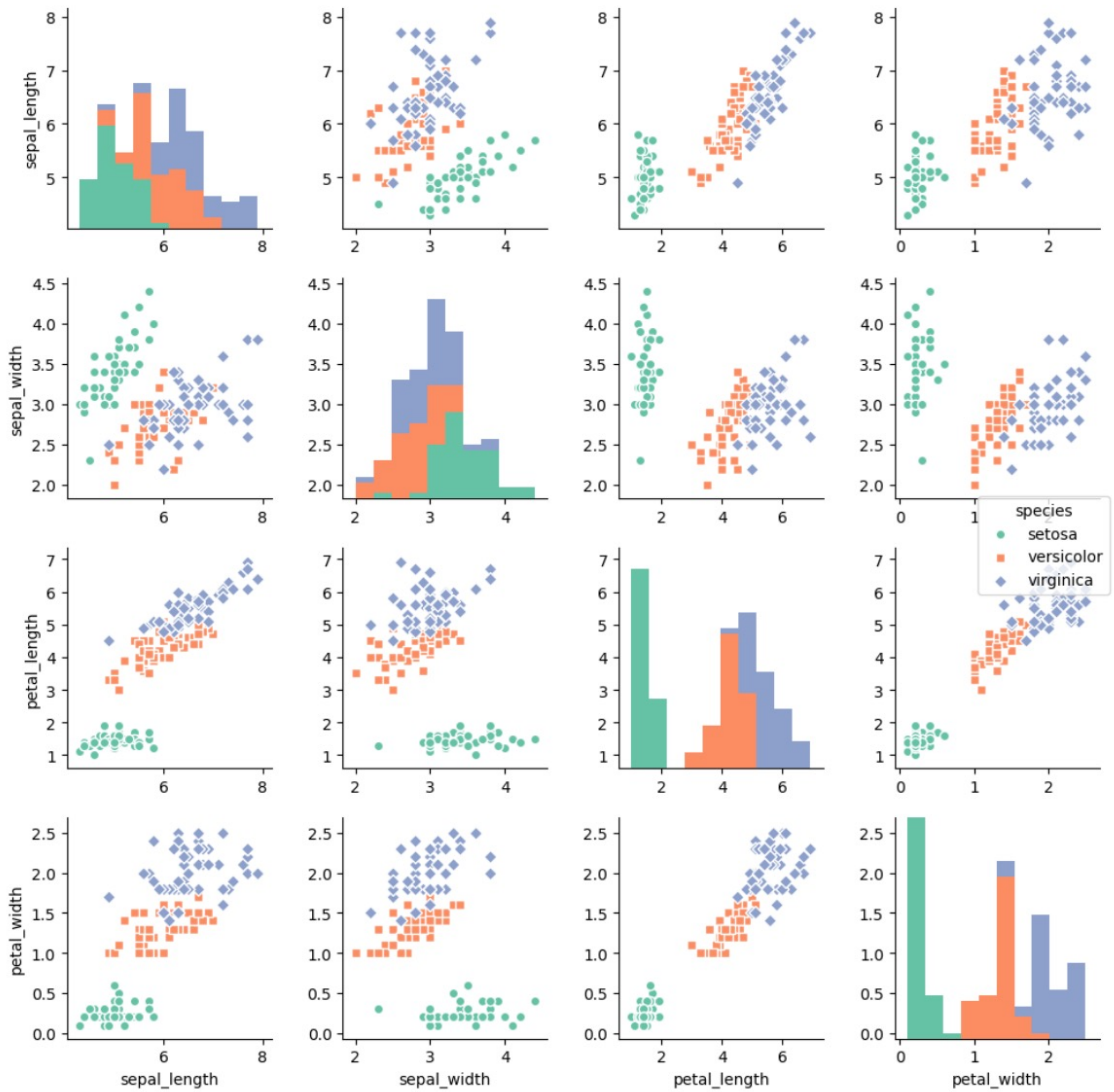
Treemap



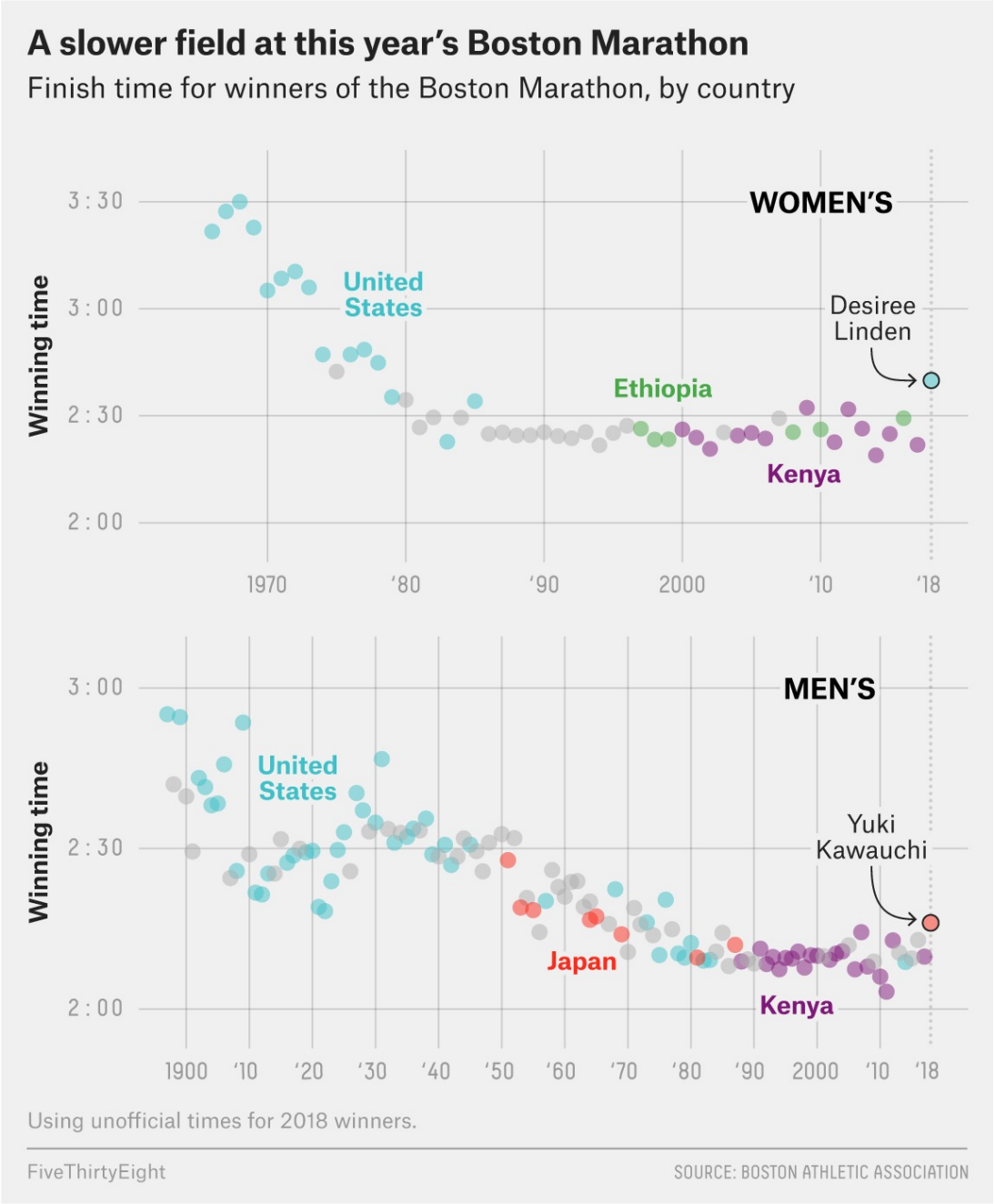
South Korea > Spain ?

% of total sold by the US ?

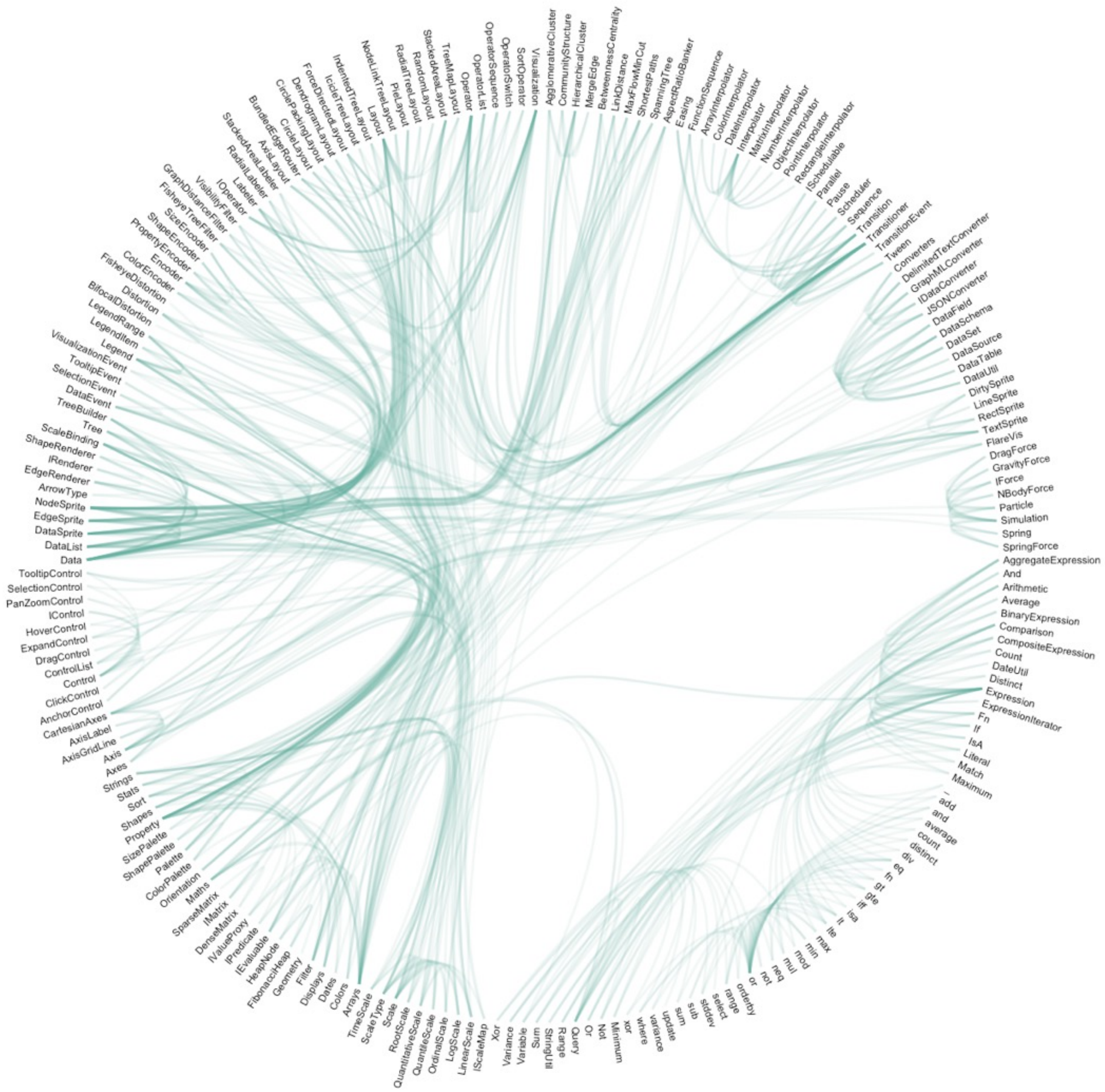
Question – **Explo/Expla** – Reader – Data – Interactivity – Eye catching



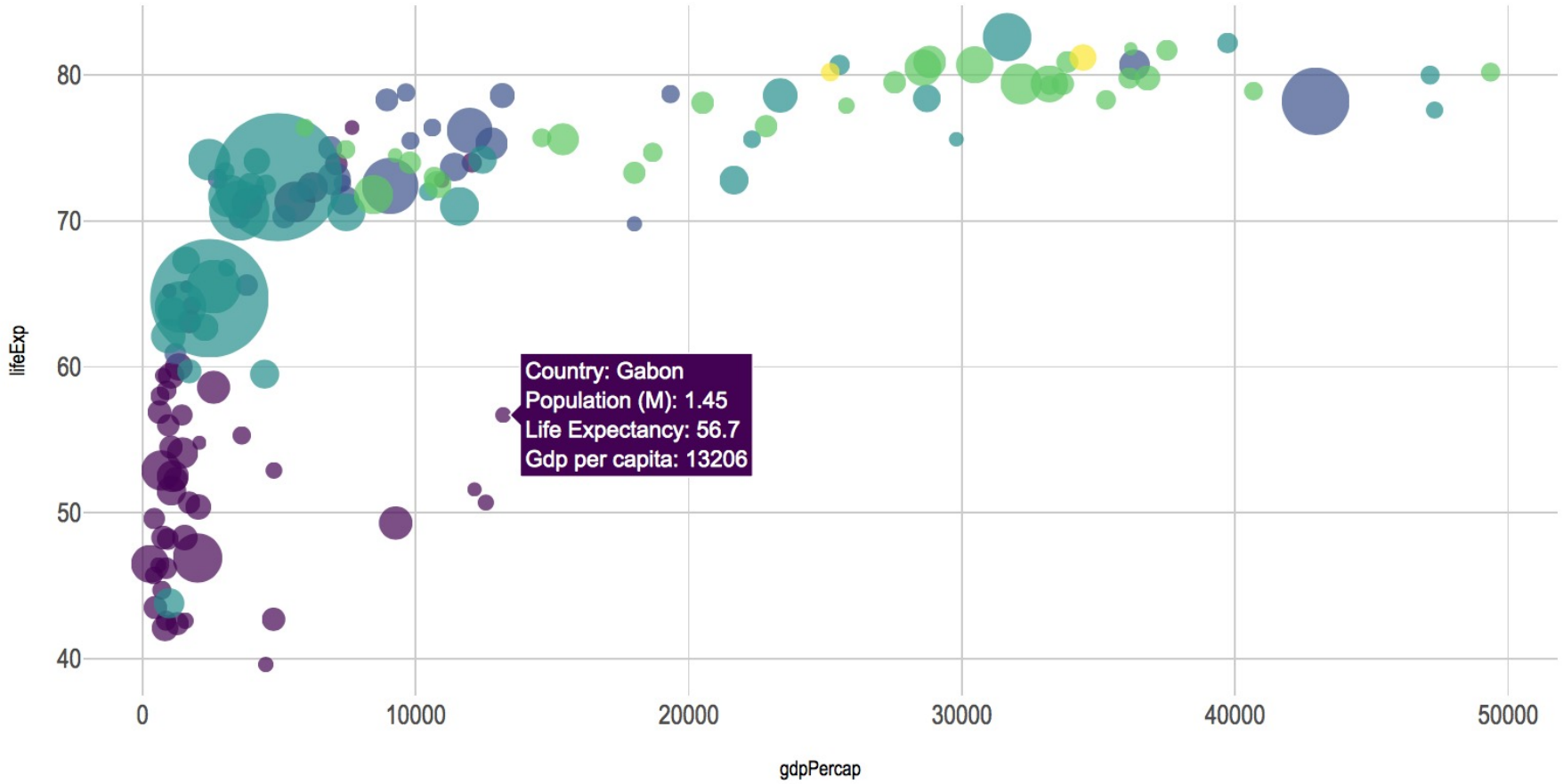
- Communicate result
- Show something specific
- Tells a story



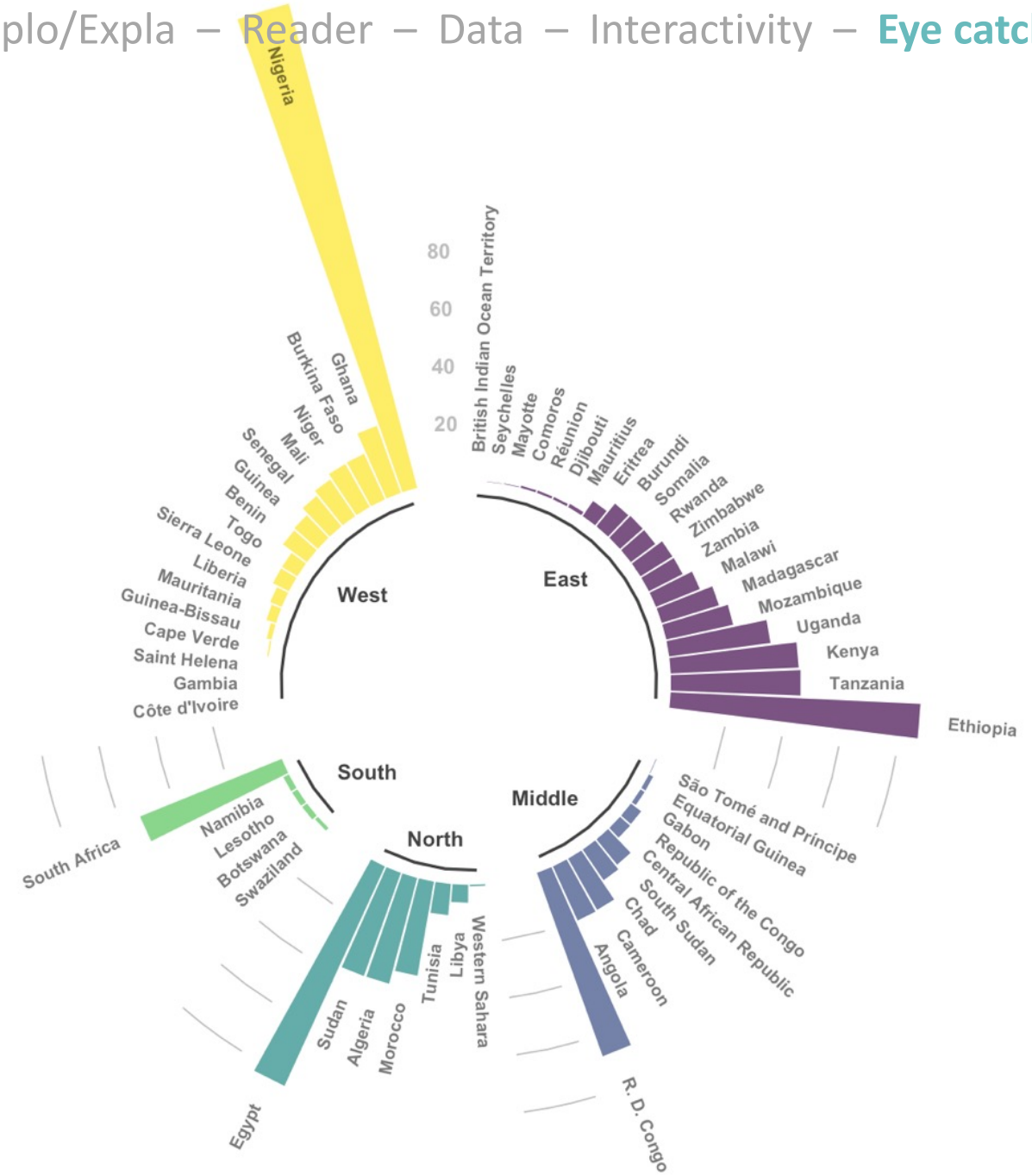
[Five thirty height]


Visualization
AgglomerativeCluster
CommunityStructure
HierarchicalCluster
MergeEdge
BetweennessCentrality
LinkDistance
MaxFlowMinCut
ShortestPaths
SpanningTree
AspectRatioBanker
Easing
FunctionSequence
ArrayInterpolator
ColorInterpolator
DateInterpolator
Interpolator
MatrixInterpolator
NumberInterpolator
ObjectInterpolator
PointInterpolator
RectangleInterpolator
ISchedulable
Parallel
Pause
Scheduler
Sequence
Transition
Transitioner
TransitionEvent
Tween
Converters
DelimitedTextConverter
GraphMLConverter
IDataConverter
JSONConverter
DataField
DataSchema
DataSet
DataSource
DataTable
DataUtil
DirtySprite
LineSprite
RectSprite
TextSprite
FlareVis
DragForce
GravityForce
IForce
NBodyForce
Particle
Simulation
Spring
SpringForce
AggregateExpression
And
Arithmetic
Average
BinaryExpression
Comparison
CompositeExpression
Count
DateUtil
Distinct
Expression
ExpressionIterator
Fn
If
IsA
Literal
Match
Maximum
add
and
average
count
distinct
div
eq
fn
gt
gte
iff
isa
lt
lte
max
min
mod
mul
neq
not
or
orderby
range
select
stddev
sub
sum
update
variance
where
xor
Minimum
Not
Or
Query
Range
StringUtil
Sum
Variable
Variance
Xor
IScaleMap
LinearScale
LogScale
OrdinalScale
QuantileScale
QuantitativeScale
RootScale
Scale
ScaleType
TimeScale
Arrays
Colors
Dates
Displays
Filter
Geometry
FibonacciHeap
HeapNode
IEvaluable
IPredicate
IValueProxy
DenseMatrix
IMatrix
SparseMatrix
Maths
Orientation
ColorPalette
Palette
ShapePalette
SizePalette
Property
Shapes
Sort
Stats
Strings
Axes
Axis
AxisGridLine
AxisLabel
CartesianAxes
AnchorControl
ClickControl
Control
ControlList
DragControl
ExpandControl
HoverControl
IControl
PanZoomControl
SelectionControl
TooltipControl
Data
DataList
DataSprite
EdgeSprite
NodeSprite
ArrowType
EdgeRenderer
IRenderer
ShapeRenderer
ScaleBinding
Tree
TreeBuilder
DataEvent
SelectionEvent
TooltipEvent
VisualizationEvent
Legend
LegendItem
LegendRange
BifocalDistortion
Distortion
FisheyeDistortion
ColorEncoder
Encoder
PropertyEncoder
ShapeEncoder
SizeEncoder
FisheyeTreeFilter
GraphDistanceFilter
VisibilityFilter
IOperator
Labeler
RadialLabeler
StackedAreaLabeler
AxisLayout
BundledEdgeRouter
CircleLayout
CirclePackingLayout
DendrogramLayout
ForceDirectedLayout
IcicleTreeLayout
IndentedTreeLayout
Layout
NodeLinkTreeLayout
PieLayout
RadialTreeLayout
RandomLayout
StackedAreaLayout
TreeMapLayout
Operator
OperatorList
OperatorSequence
OperatorSwitch
SortOperator

Question – Explo/Expla – Reader – Data – **Interactivity** – Eye catching

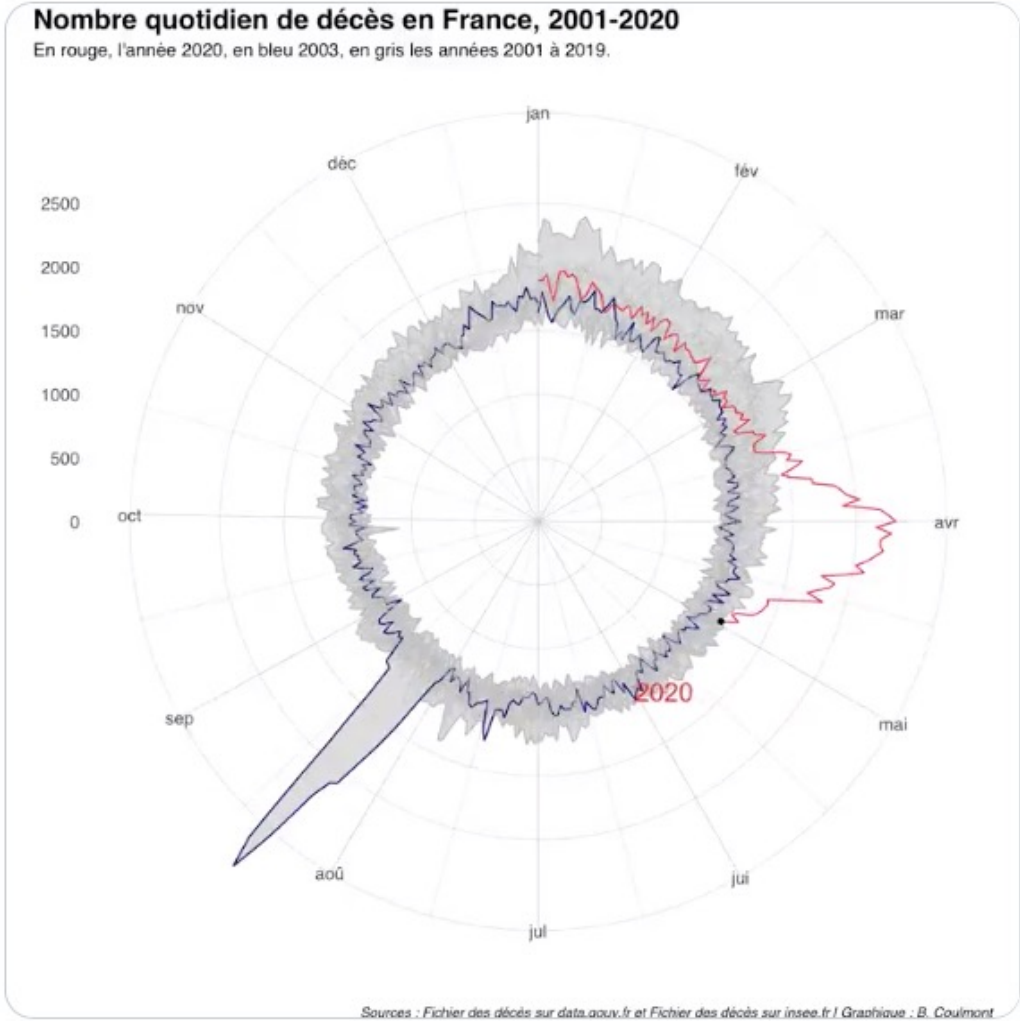


Question – Explo/Expla – Reader – Data – Interactivity – **Eye catching**



coulmont
@coulmont

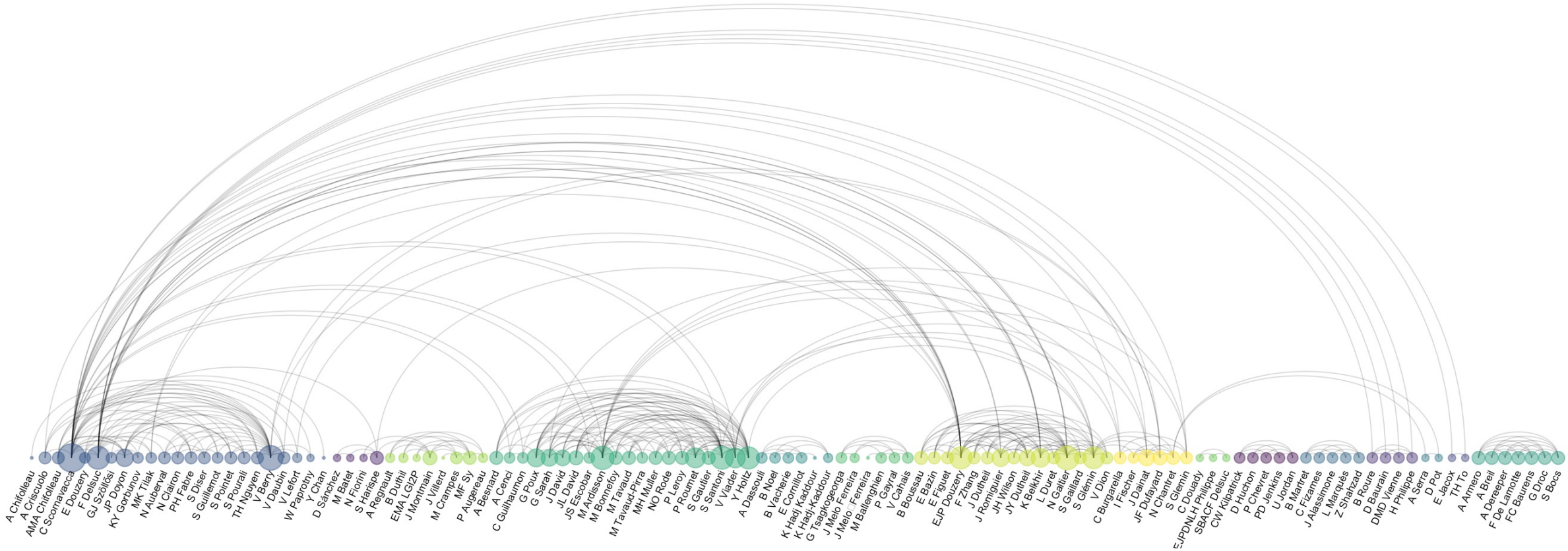
Nombre quotidien de décès en France, 2001-2020, en version animée, coordonnées polaires.



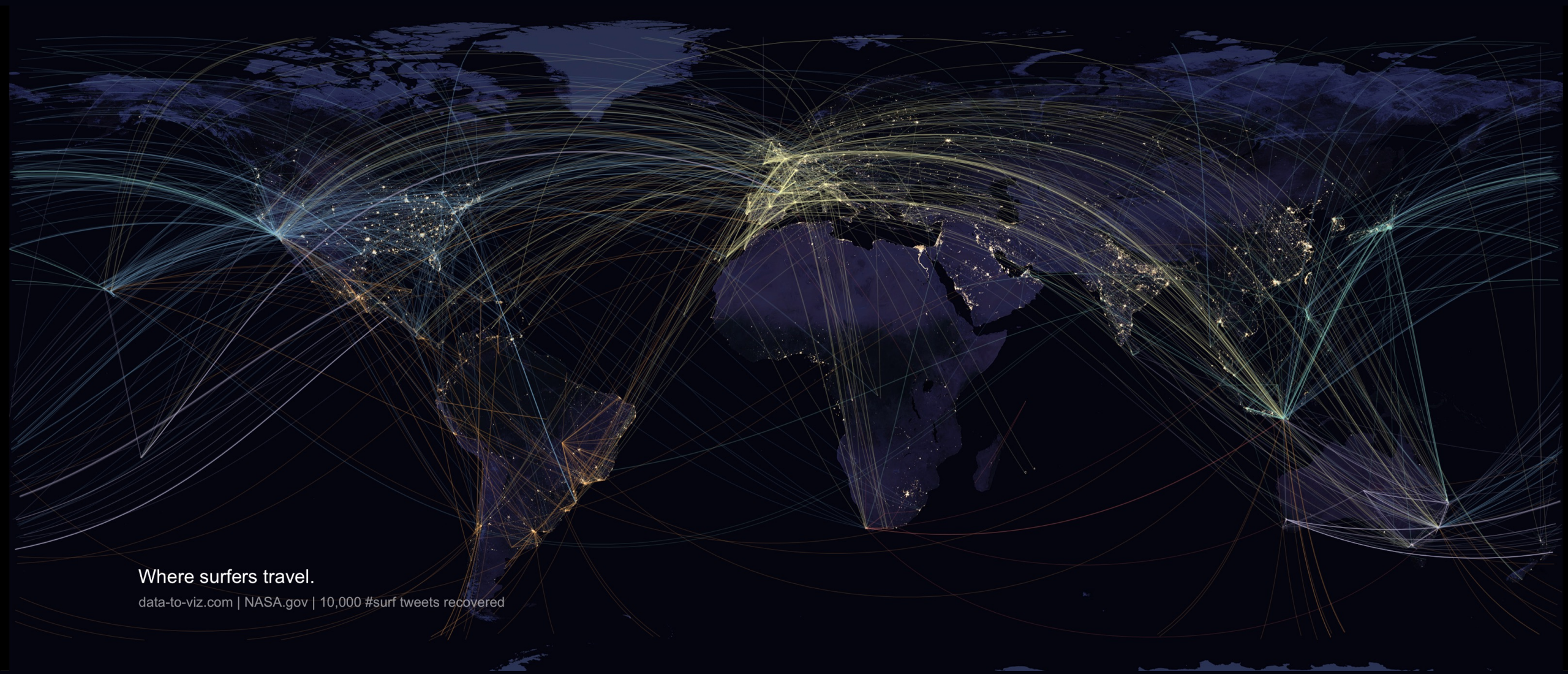
12:43 PM · 2 déc. 2020 · Twitter Web App

6,5 k Retweets **1 k** Tweets cités **16,1 k** J'aime

[Link]



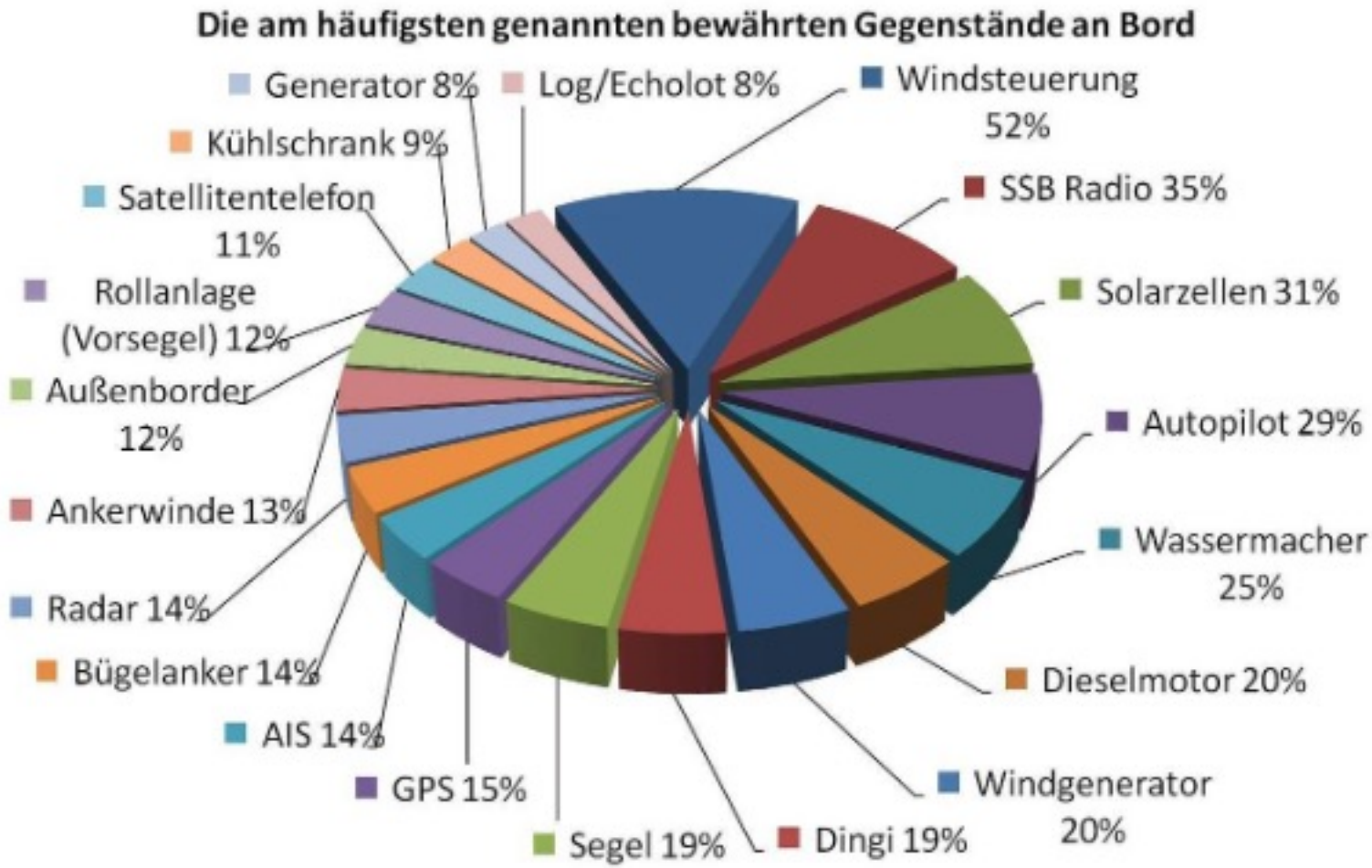
Co-authorship network of a researcher


Where surfers travel.
data-to-viz.com | NASA.gov | 10,000 #surf tweets recovered

WHAT YOU SHOULD NOT DO

A gallery of most common caveats

What's wrong with this chart?



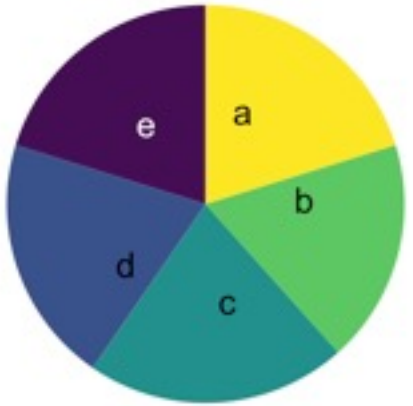
Source: WTF Visualizations

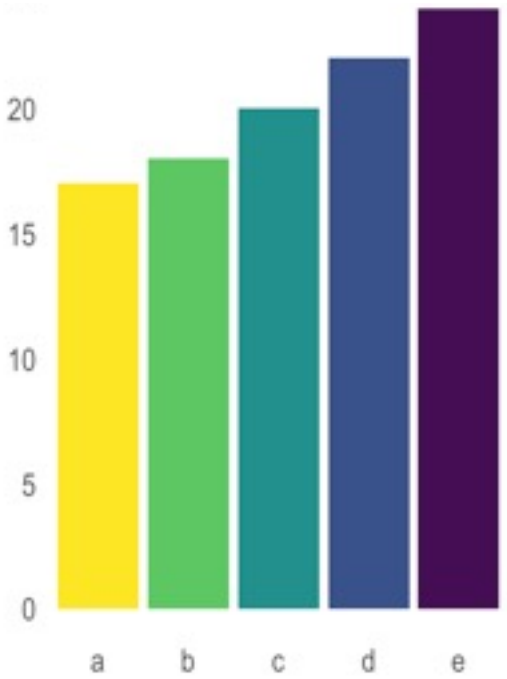
What's wrong with pie chart?

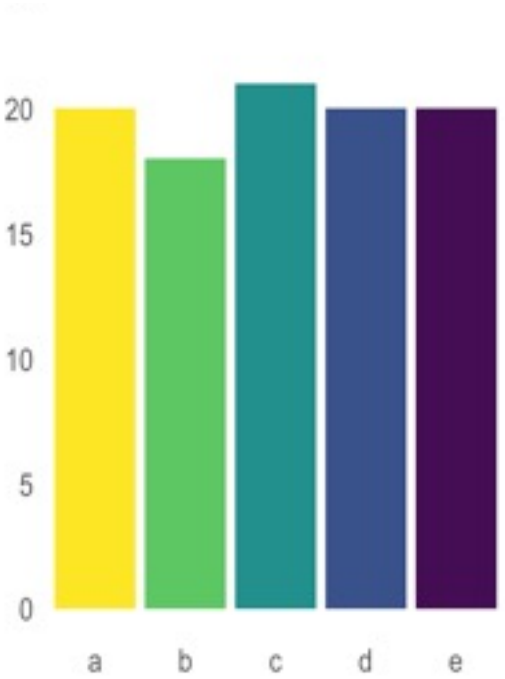
What can you see?

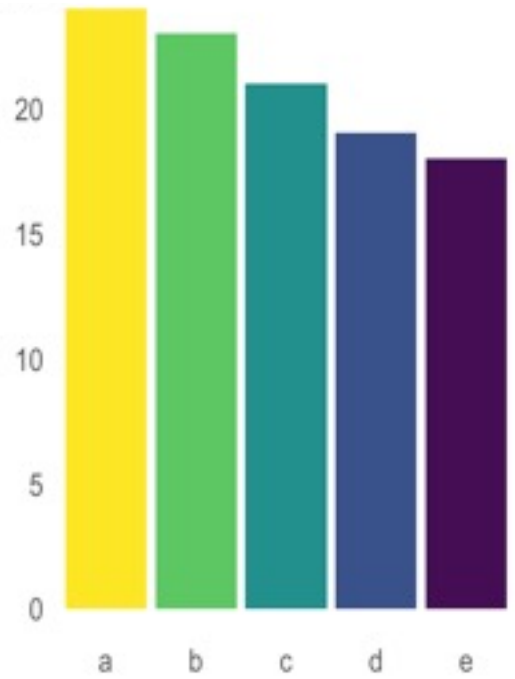
What's wrong with pie chart?

It is hard to distinguish angles

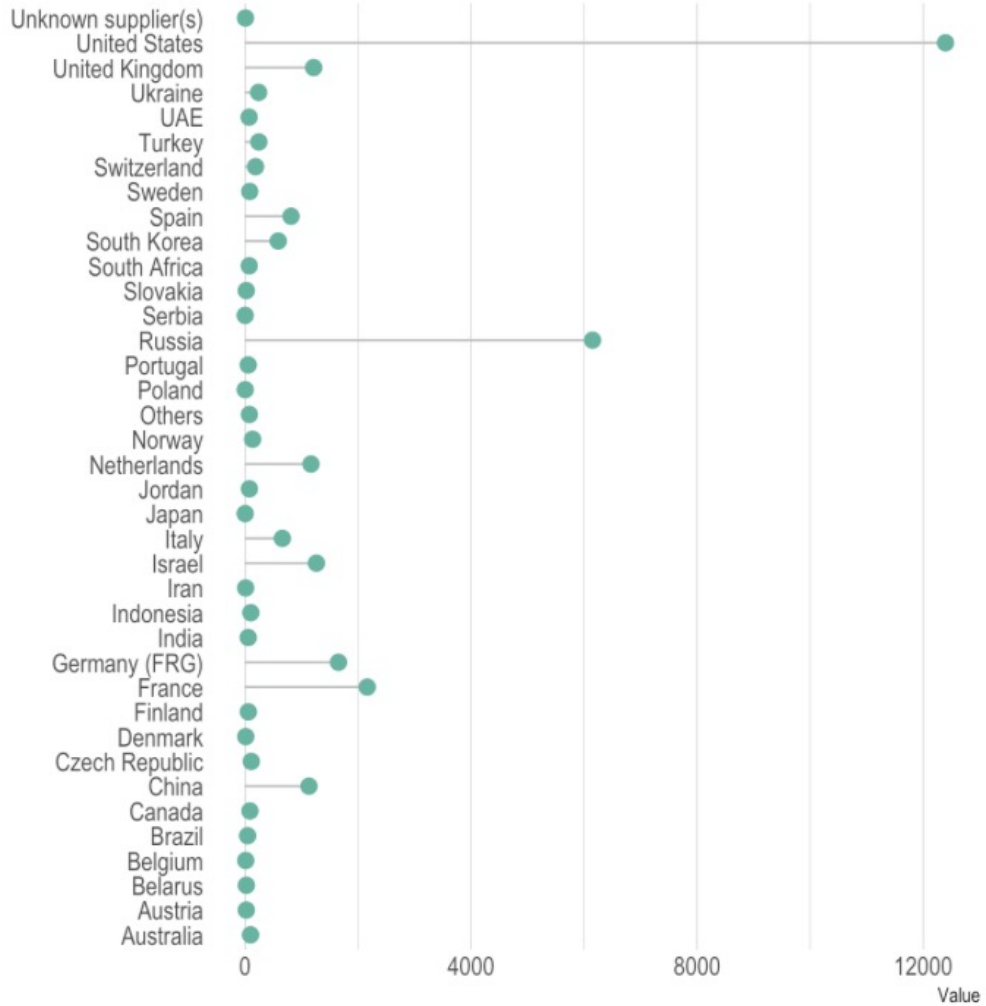




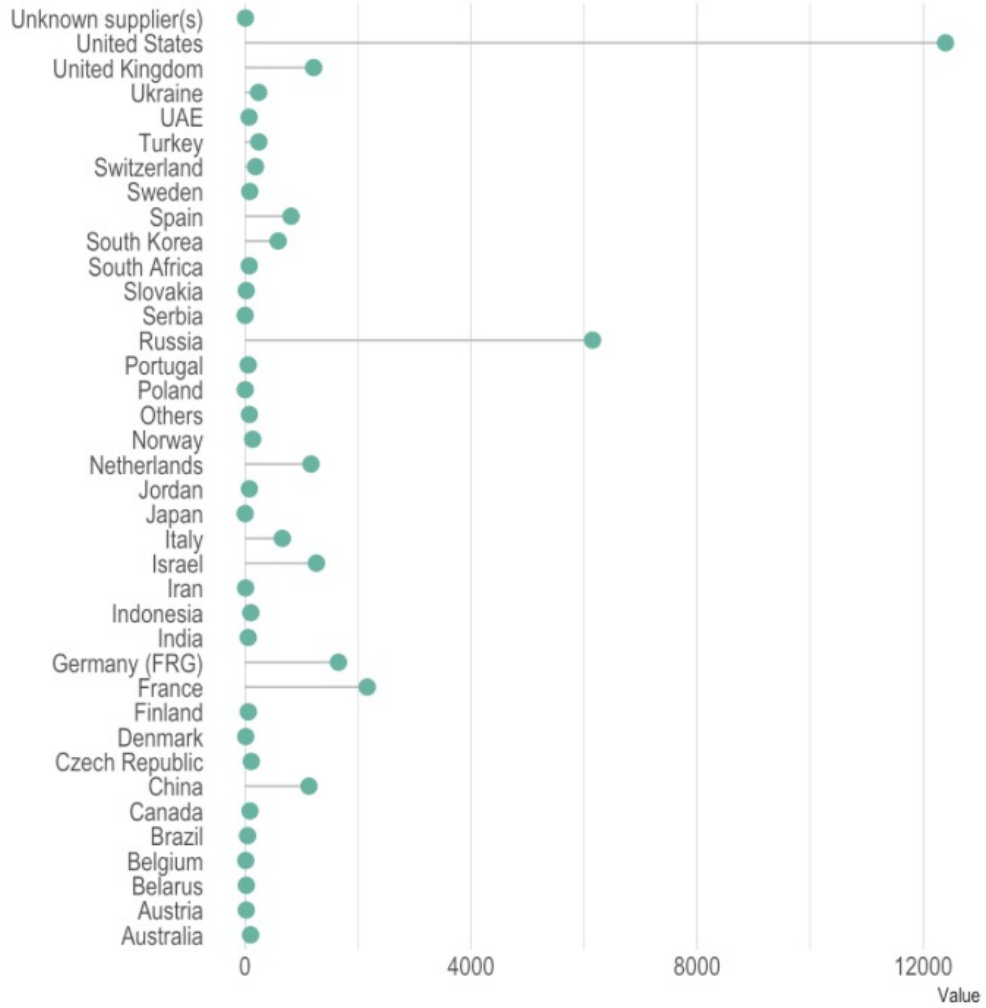


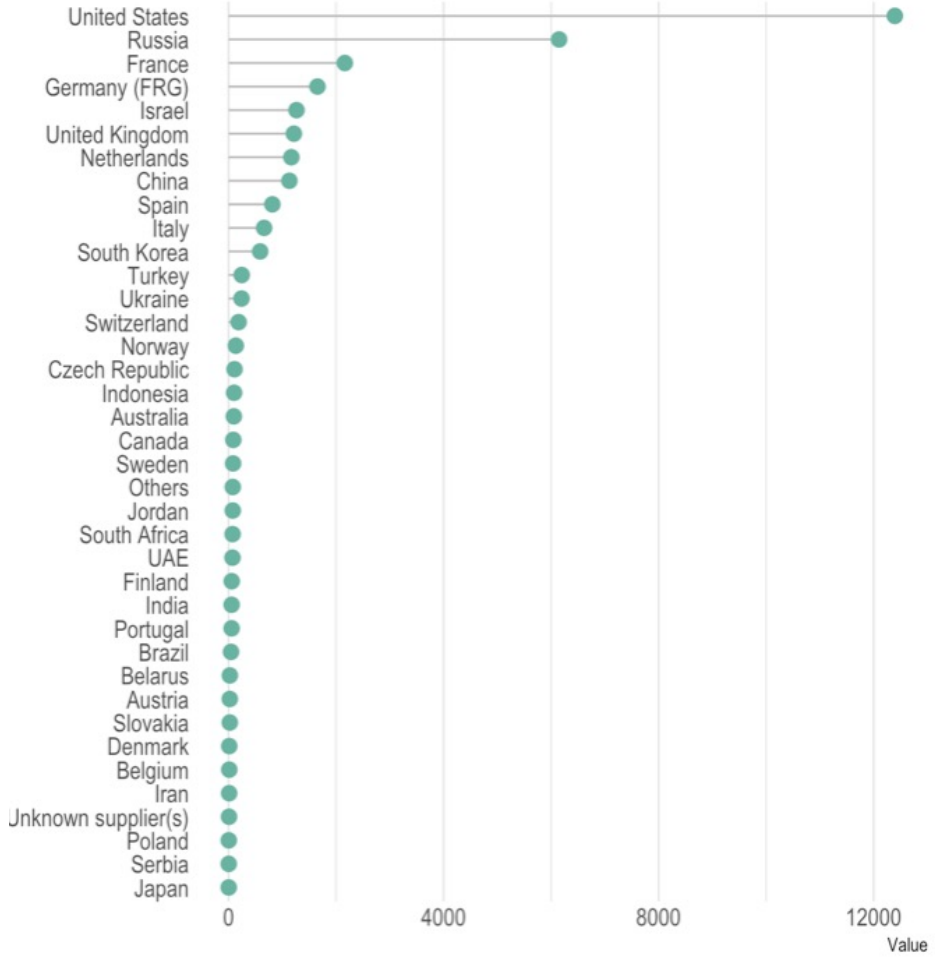


What could be better here?

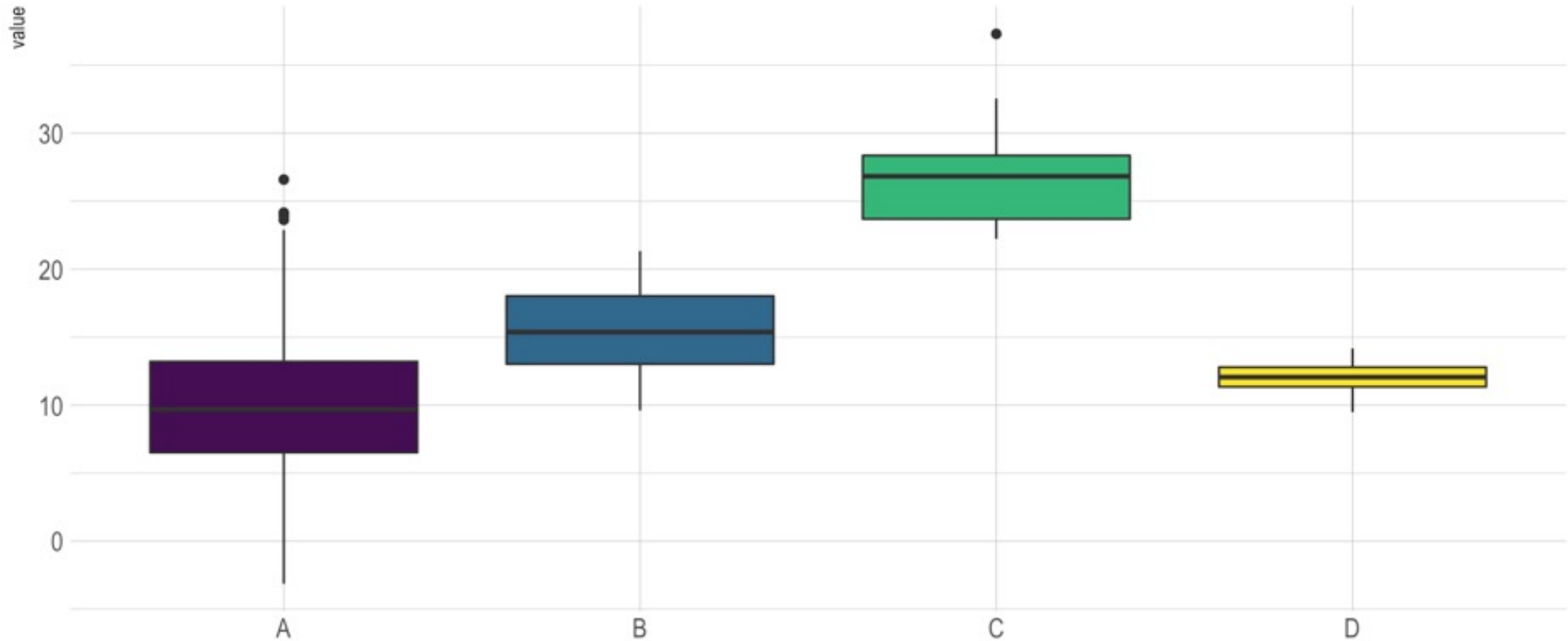


Order your data

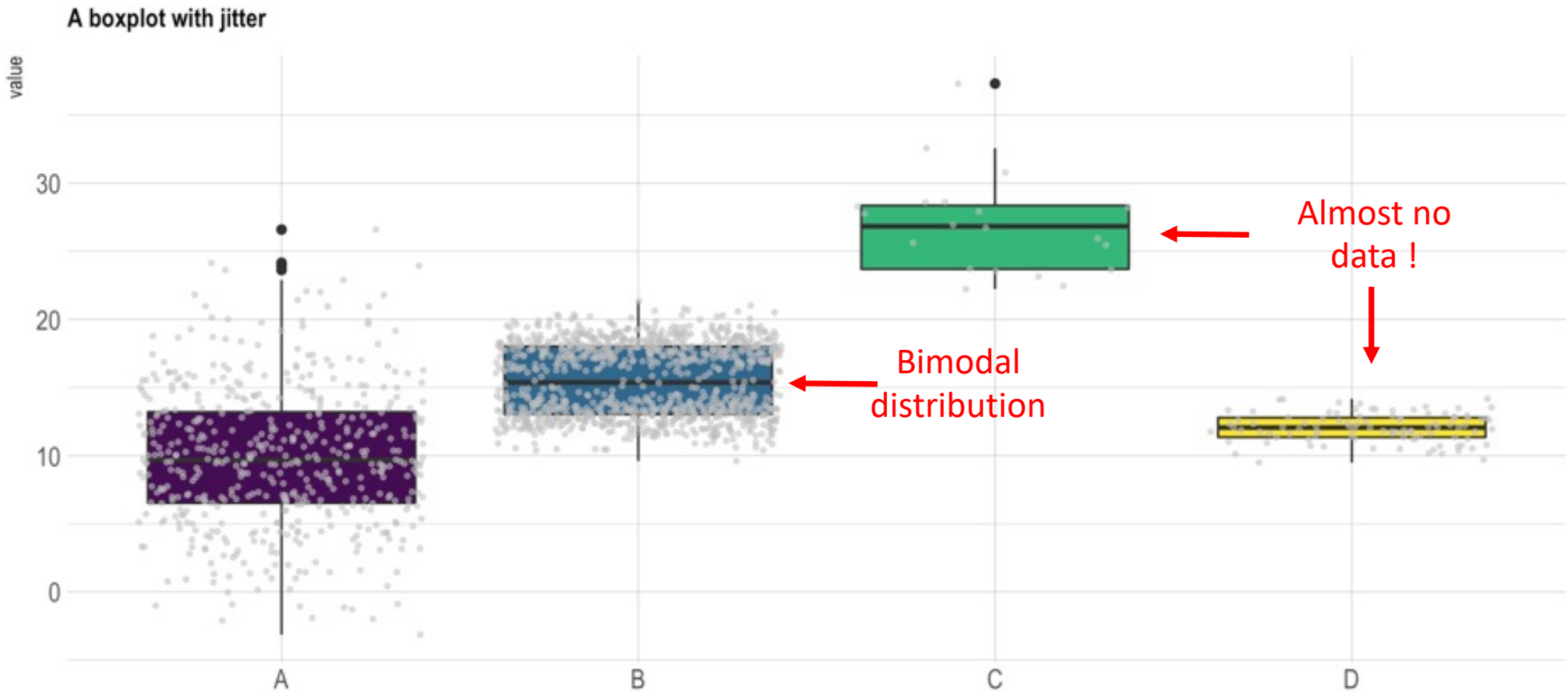




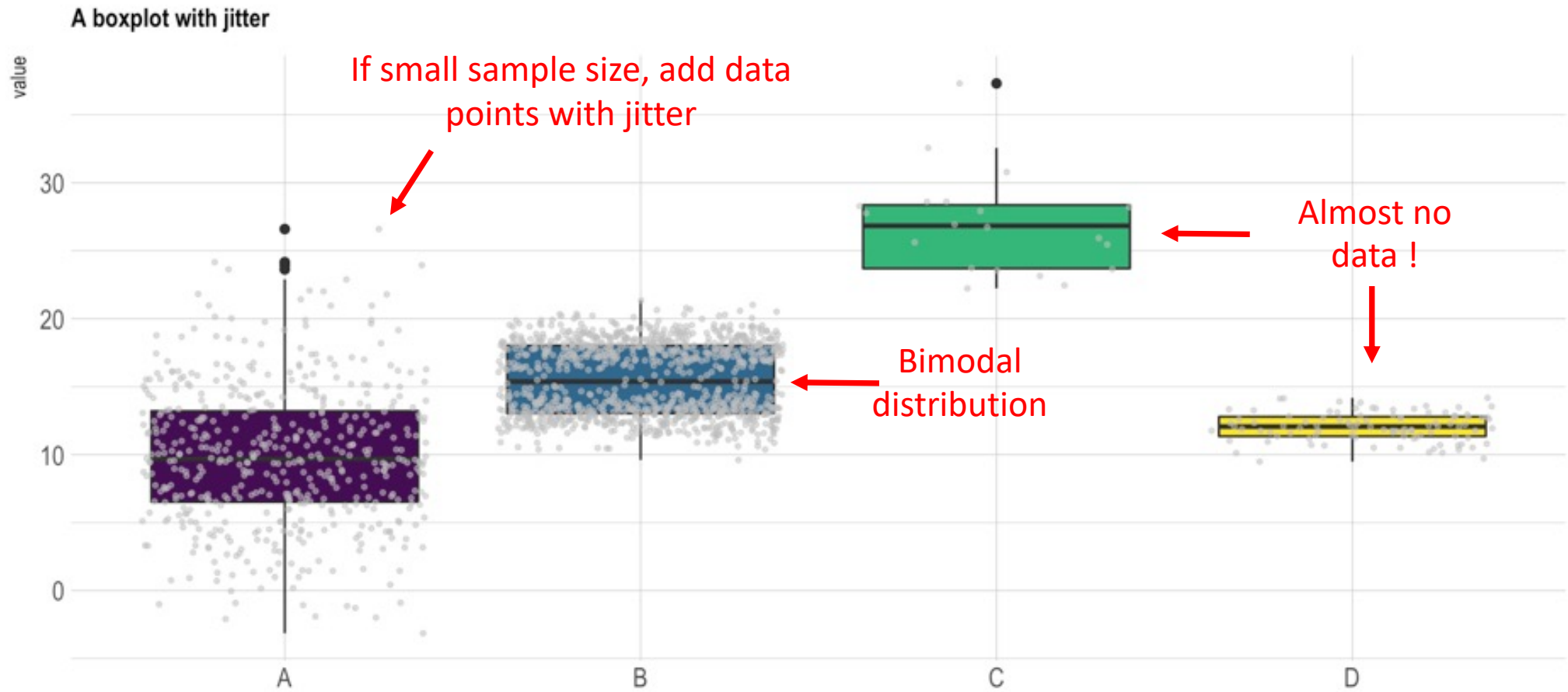
Anything wrong here?



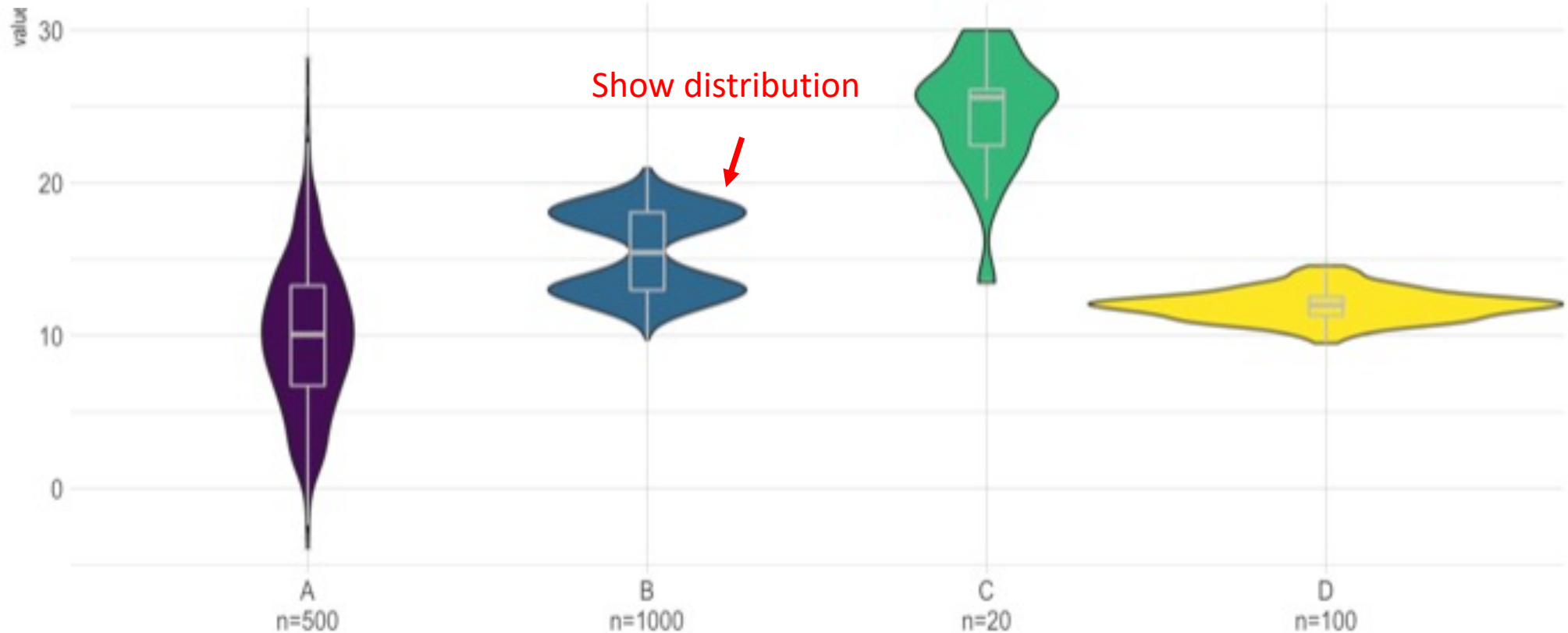
Boxplot = hide information



Boxplot = hide information



If big sample size, use violin plot

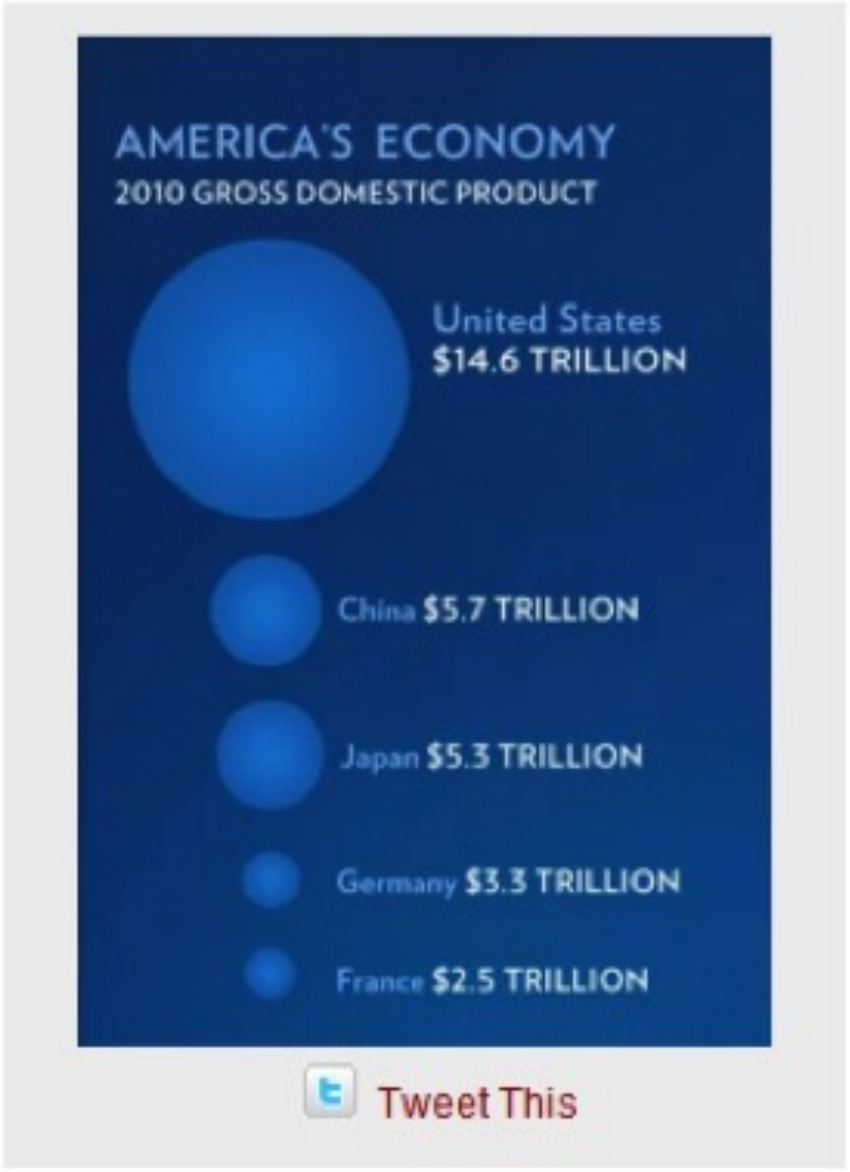


Show distribution

Show sample size


AMERICA'S ECONOMY
2010 GROSS DOMESTIC PRODUCT
United States
$14.6 TRILLION
China $5.7 TRILLION
Japan $5.3 TRILLION
Germany $3.3 TRILLION
France $2.5 TRILLION

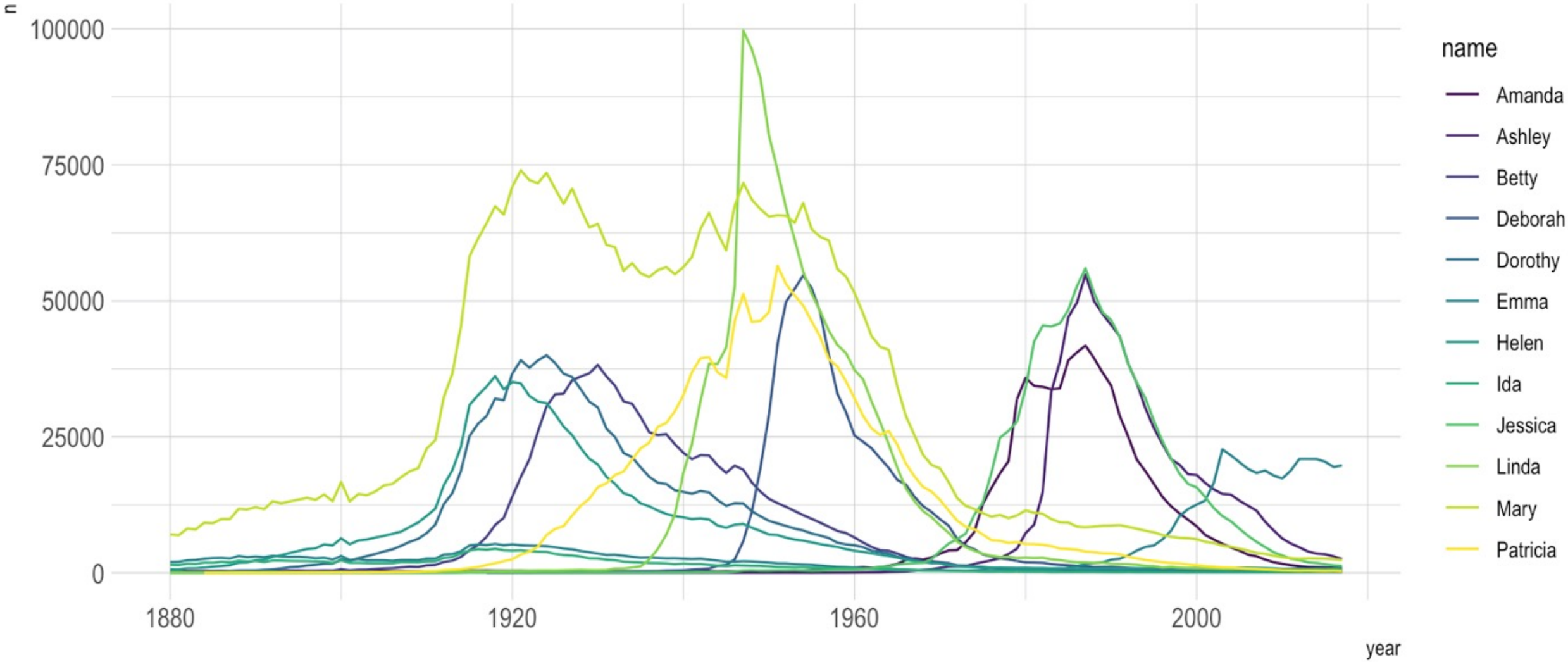
Size = radius



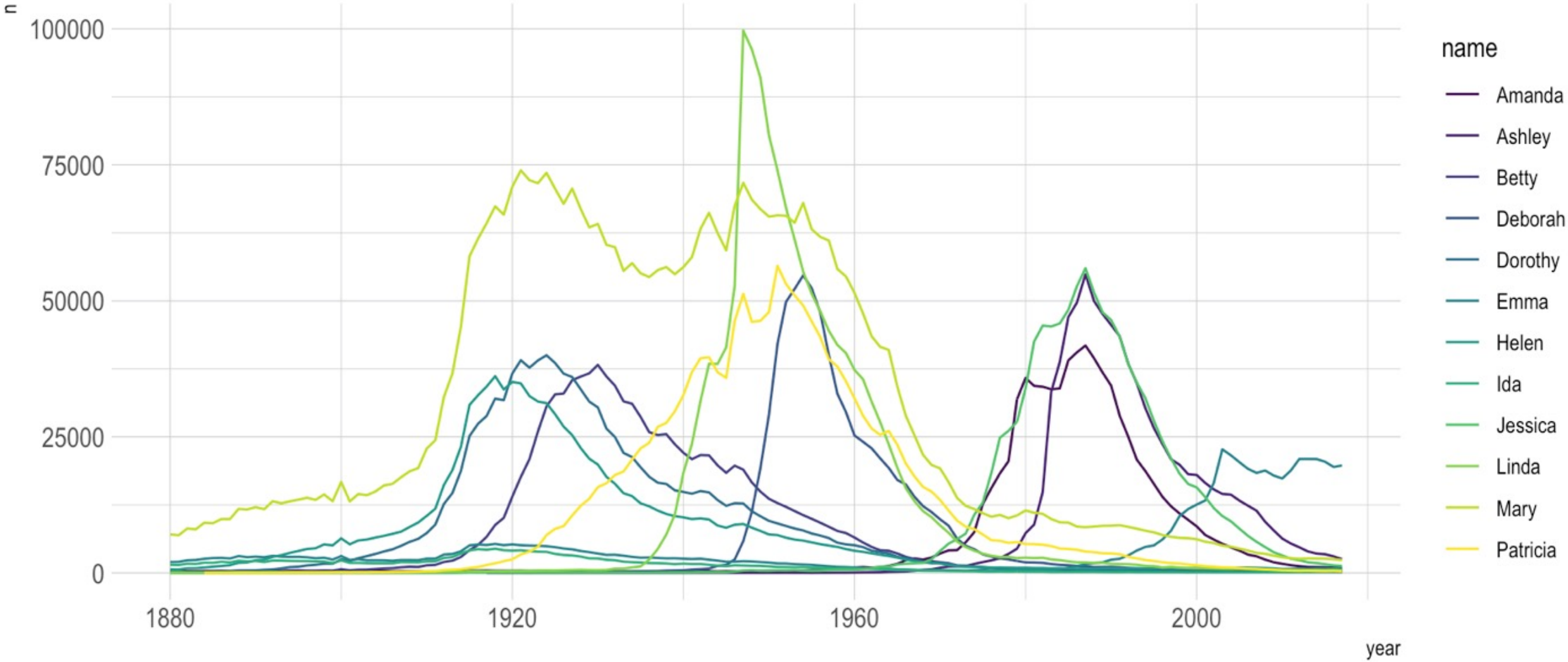
Size = area

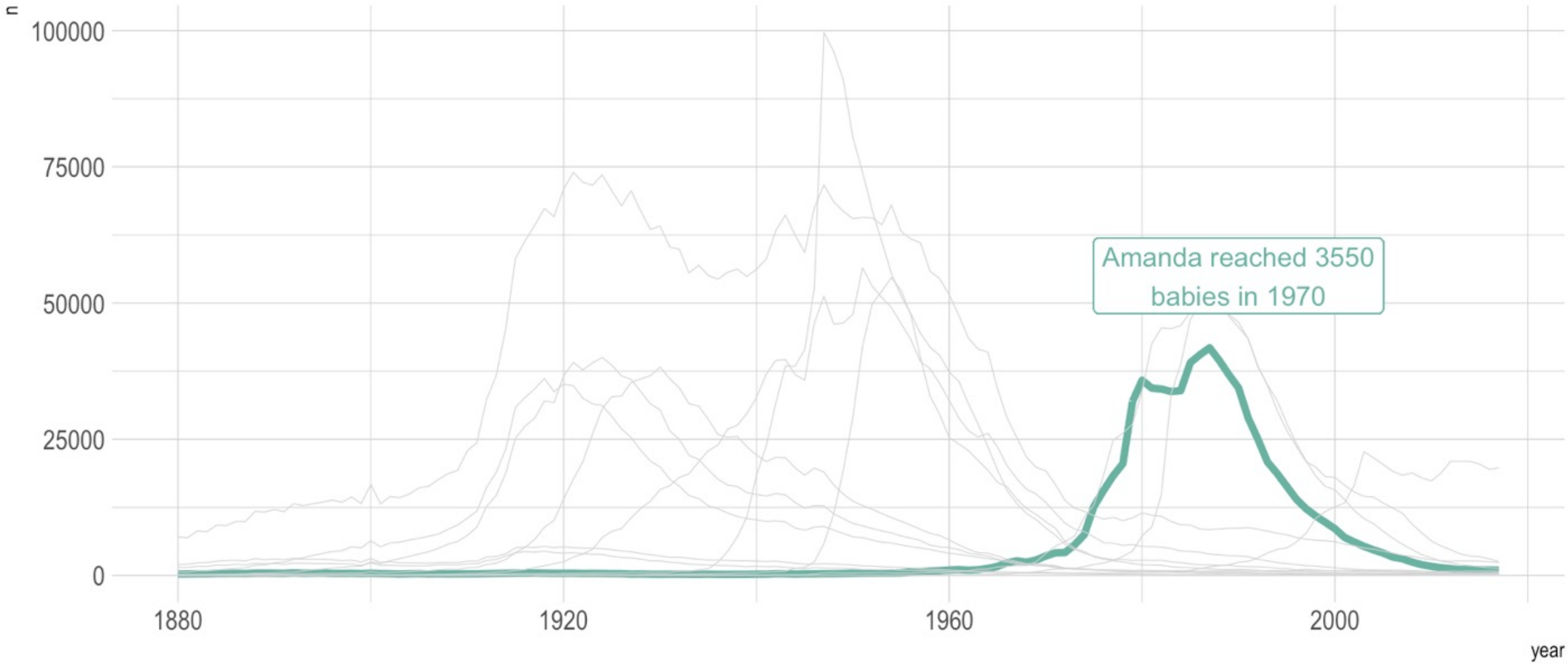
Source: [Fast Fedora blog]

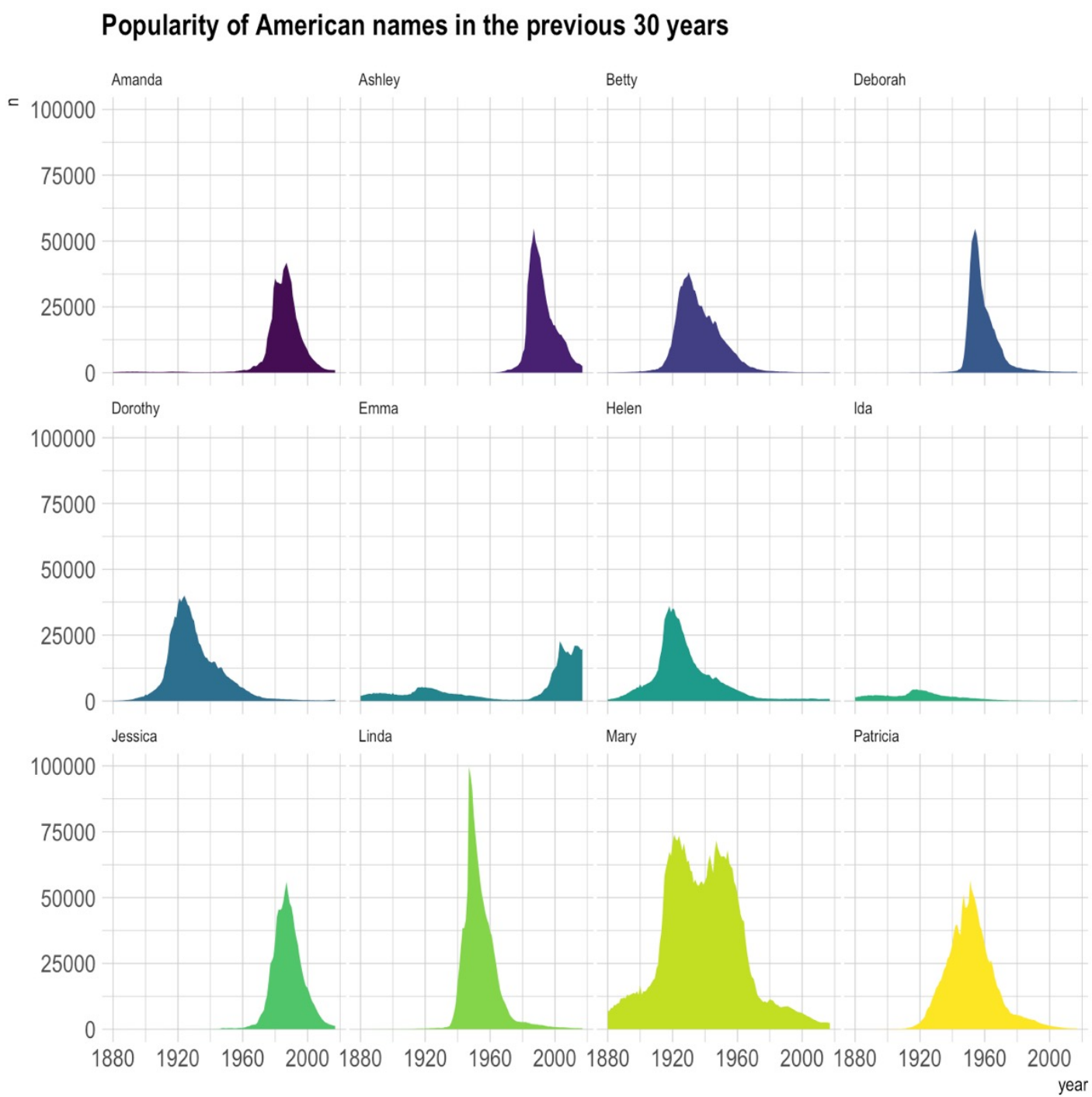
A spaghetti chart of baby names popularity

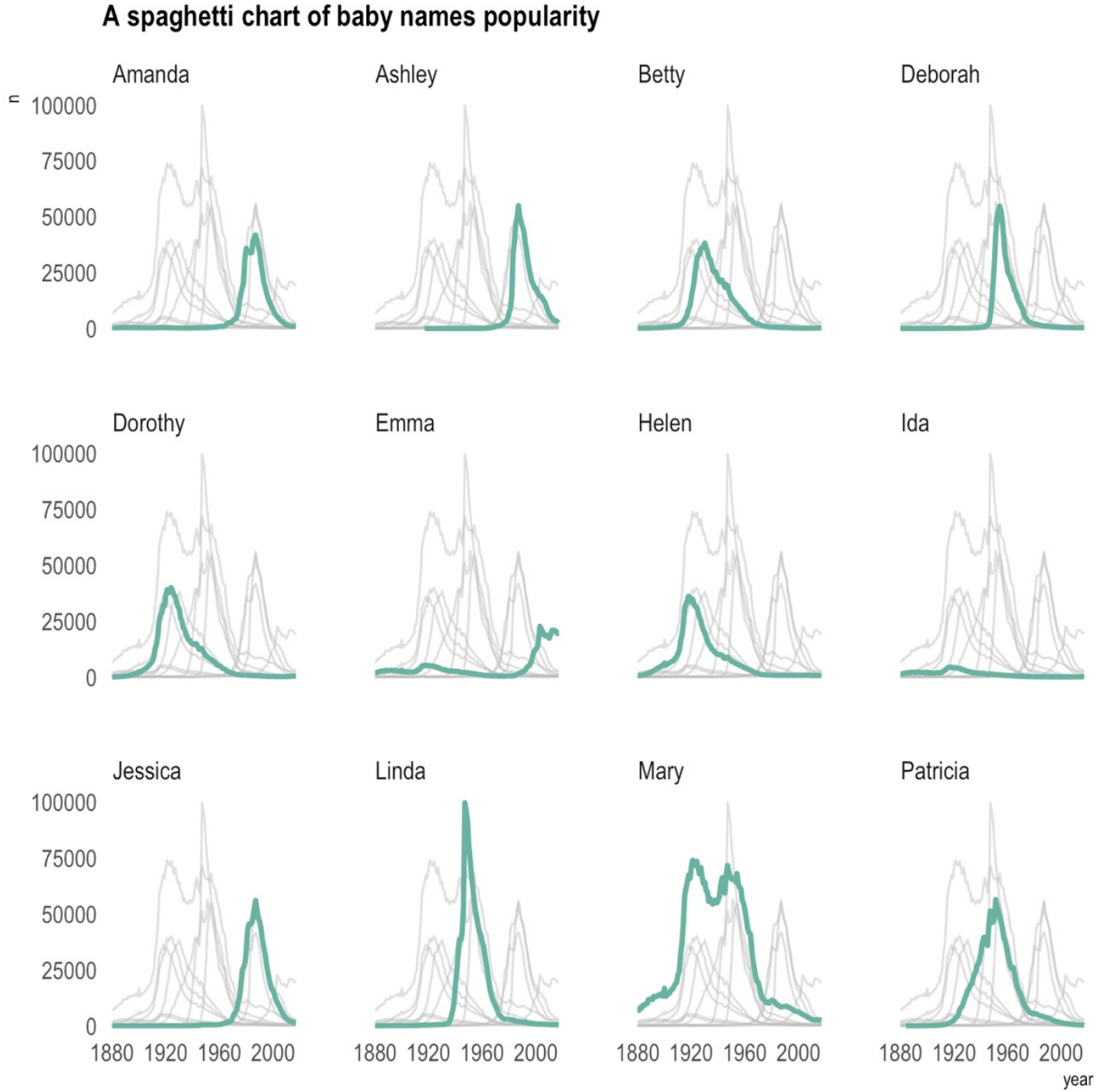


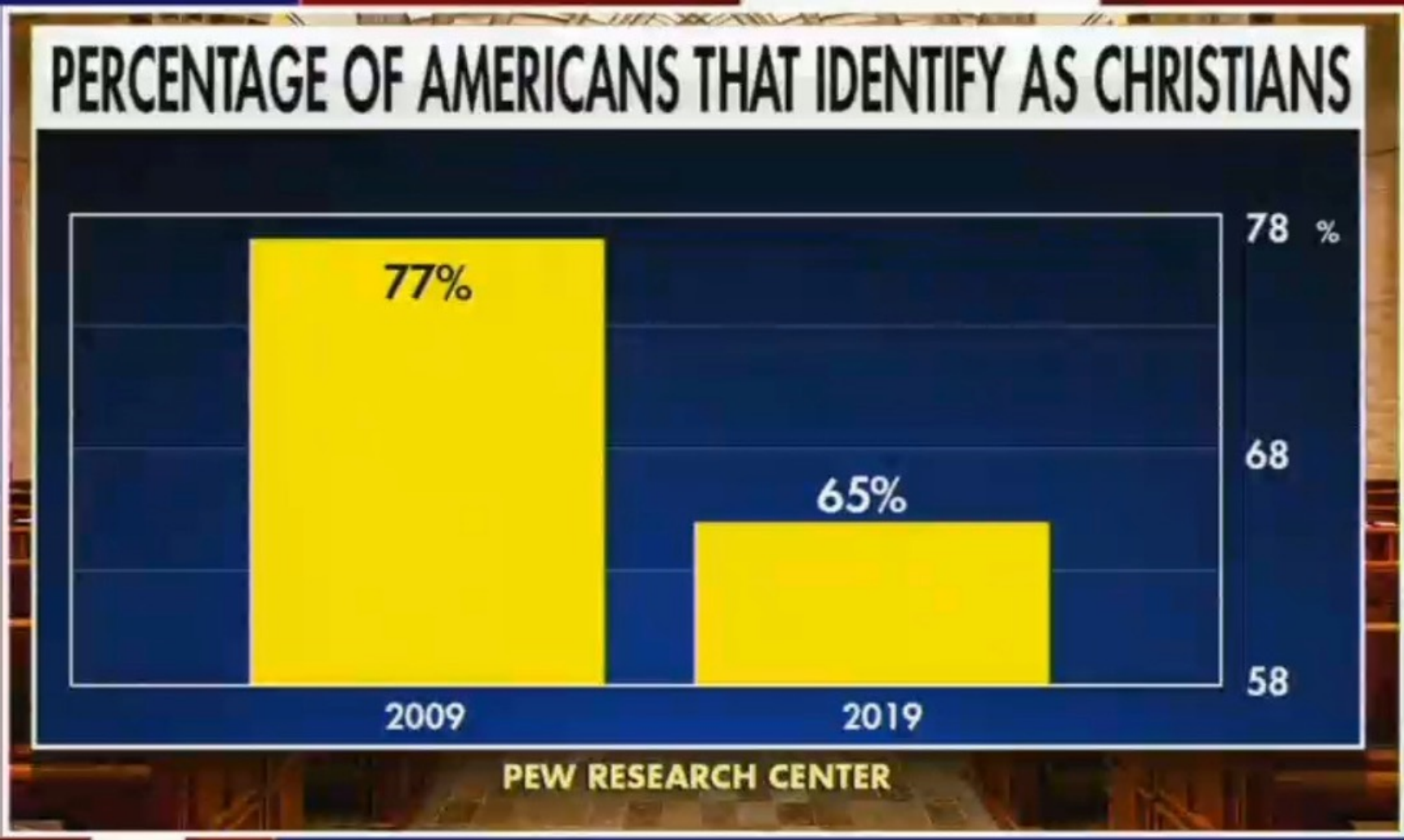
A spaghetti chart of baby names popularity




Popularity of American names in the previous 30 years
n
100000
75000
50000
25000
0
1880
1920
1960
2000
year
Amanda reached 3550
babies in 1970


Popularity of American names in the previous 30 years
Amanda
Ashley
Betty
Deborah
Dorothy
Emma
Helen
Ida
Jessica
Linda
Mary
Patricia
n
100000
75000
50000
25000
0
1880
1920
1960
2000
year


A spaghetti chart of baby names popularity
Amanda
Ashley
Betty
Deborah
Dorothy
Emma
Helen
Ida
Jessica
Linda
Mary
Patricia
n
100000
75000
50000
25000
0
1880
1920
1960
2000
year



FOX NEWS
8:00 ET

WE USED TO BE AN ENTHUSIASTICALLY CHRISTIAN NATION

TUCKER CARLSON ★TONIGHT★

Data to Viz

A collection of dataviz caveats

Data-to-viz.com/caveats

Order your data

When displaying the value of several entities, ordering them makes the graph much more insightful.

To cut or not to cut?

Cutting the Y-axis is one of the most controversial practice in data viz. See why.

The spaghetti chart

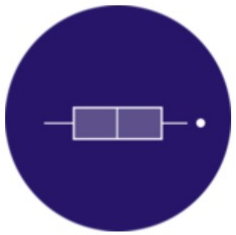
A line graph with too many lines becomes unreadable: it is called a spaghetti graph.

Pie chart

The human eye is bad at reading angles. See how to replace the most criticized chart ever.

Play with histogram bin size

Always try different bin sizes when you build a histogram, it can lead to different insights.



Do boxplots hide information?

Boxplots are a great way to summarize a distribution but hide the sample size and their distribution.

The problem with error bars

Barplots with error bars must be used with great care. See why and how to replace them.

Too many distributions.

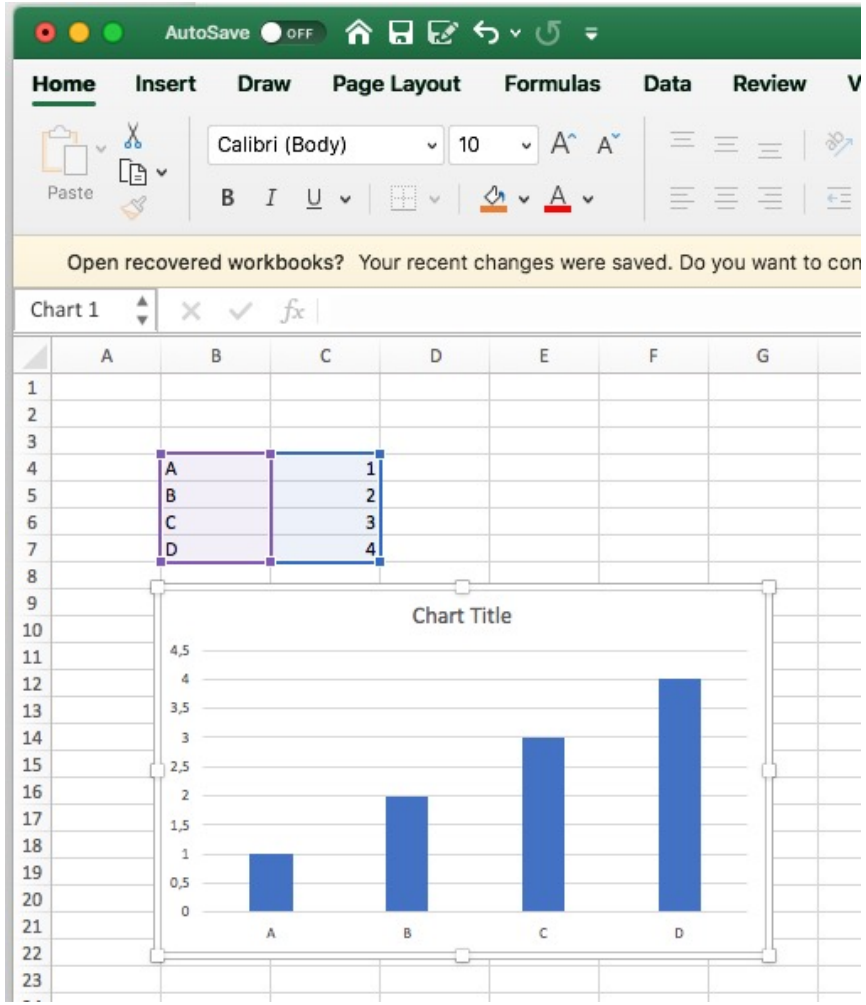
If you need to compare the distributions of many variables, don't clutter your graphic.

HOW TO DO IT

The R and Python graph galleries


Design the idea
Build it
time
20 %
80 %


Easy
Hard
Limited
Flexible



Excel

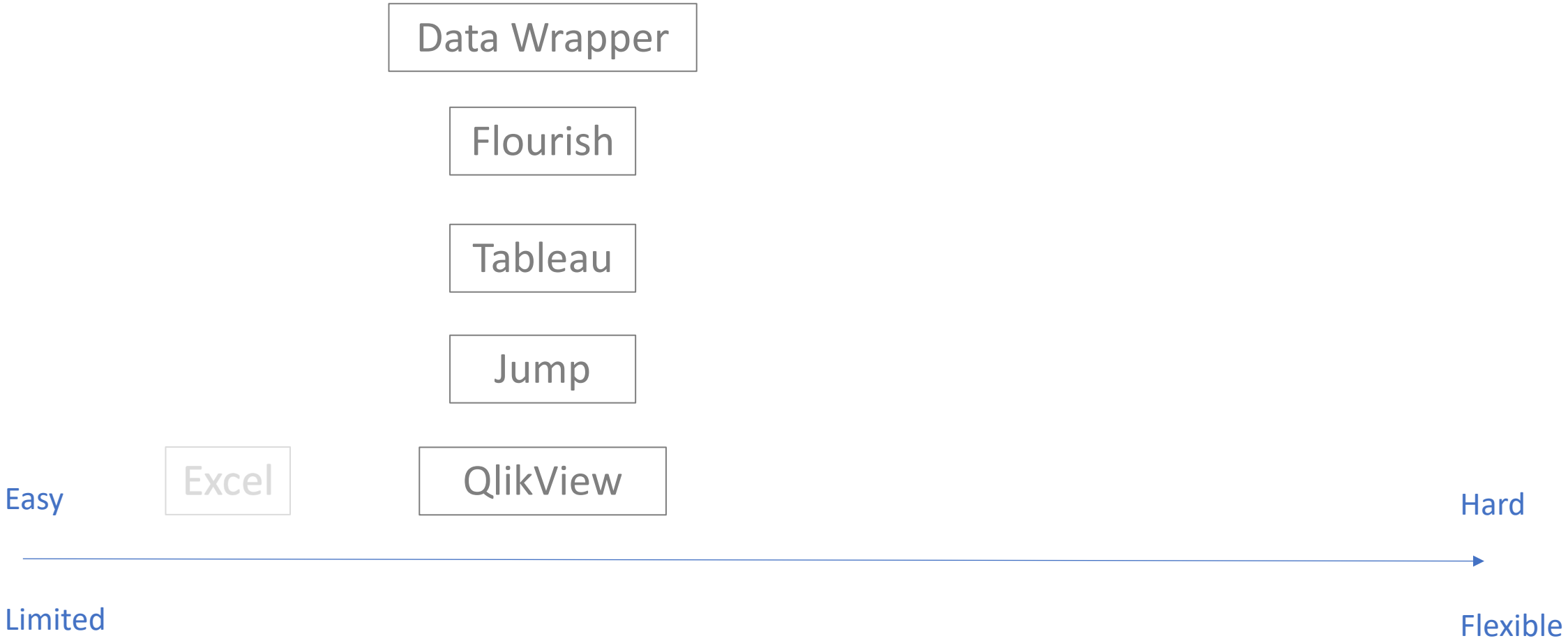
Easy

Hard

Limited

Flexible

Data Wrapper demo



[R graph gallery demo]

Data Wrapper

Flourish

Tableau

Jump

Excel

QlikView

R

Python

SAS

MatLab

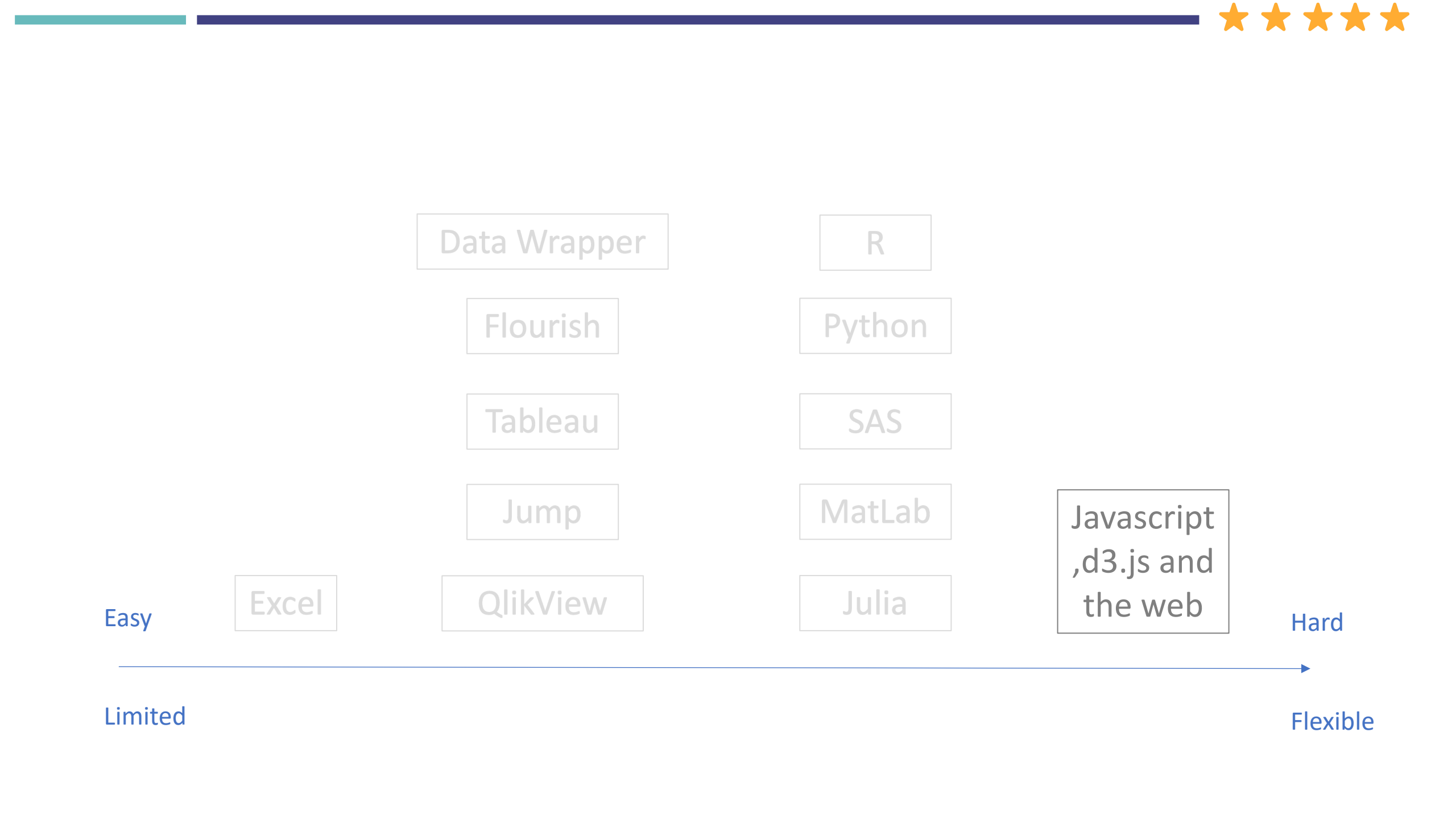
Julia

Easy

Hard

Limited

Flexible

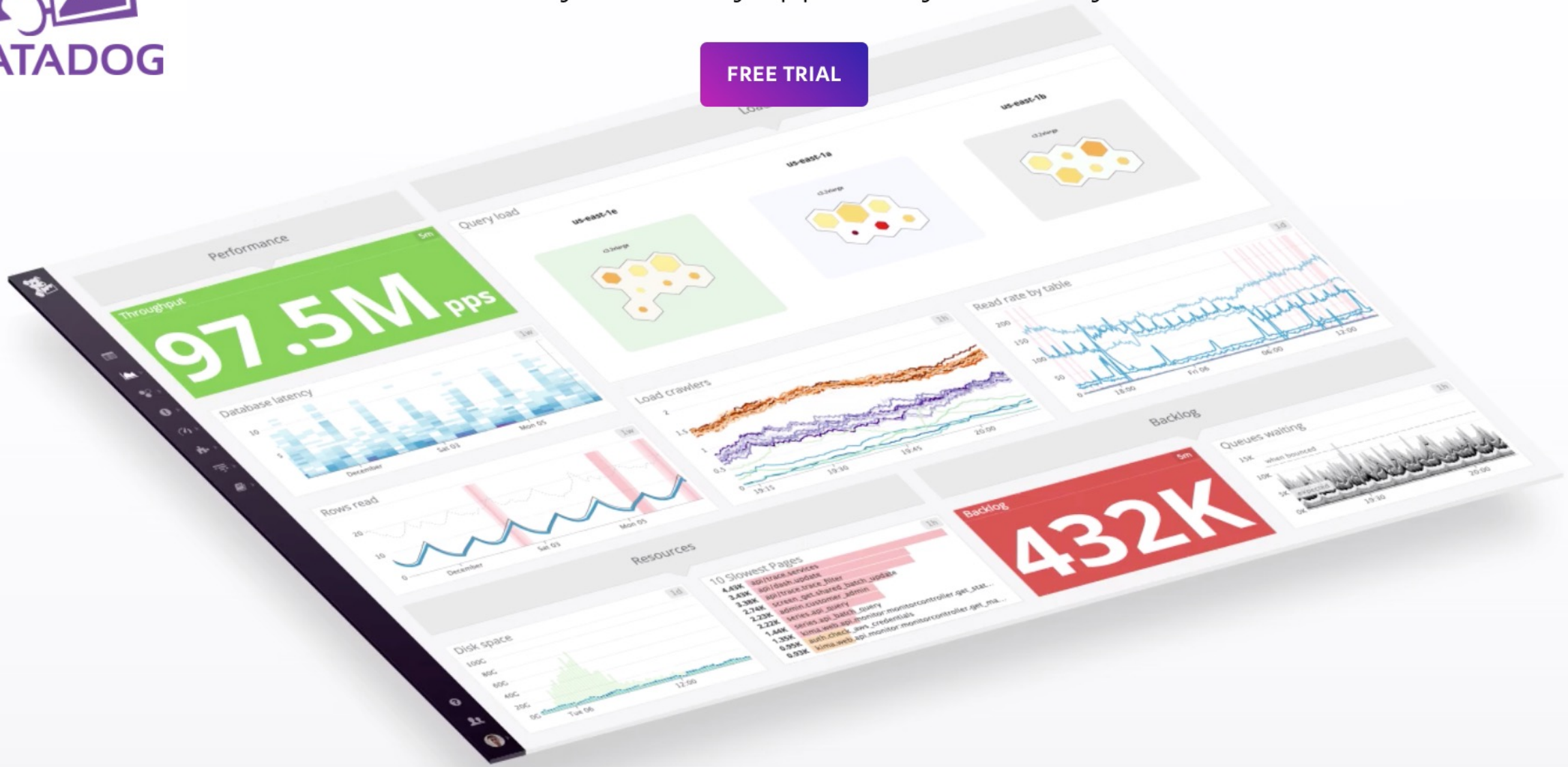

Data Wrapper
R
Flourish
Python
Tableau
SAS
Jump
MatLab
Javascript ,d3.js and the web
Excel
QlikView
Julia
Easy
Hard
Limited
Flexible

DATADOG

Modern monitoring & analytics

See inside any stack, any app, at any scale, anywhere.

FREE TRIAL



R-graph-gallery.com

Python-graph-gallery.com

D3-graph-gallery.com

KANTAR
Information is Beautiful
Awards

Data-to-viz.com

@R_Graph_Gallery

Yan.holtz.data@gmail.com

github.com/holtzy/Talk

www.yan-holtz.com