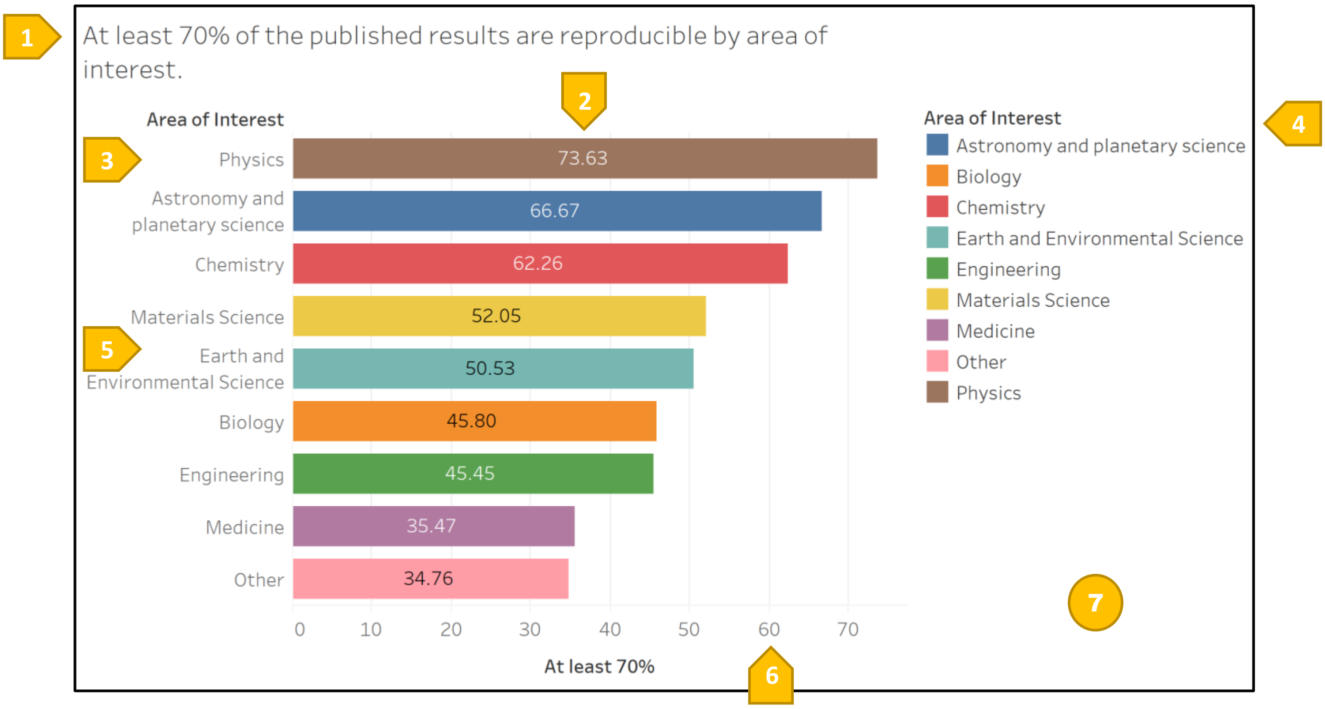
DataViz 5 **13 Feb 2020**

What Proportion of Published Results in your Field are Reproducible?

Teo Lip Peng Raymond (lippeng.teo.2019@mitb.smu.edu.sg)

Data Visualisation Link (Tableau Public) –
https://public.tableau.com/profile/raymondteo#!/vizhome/DataViz5_ReproducibilityUncertainty/Dashboard

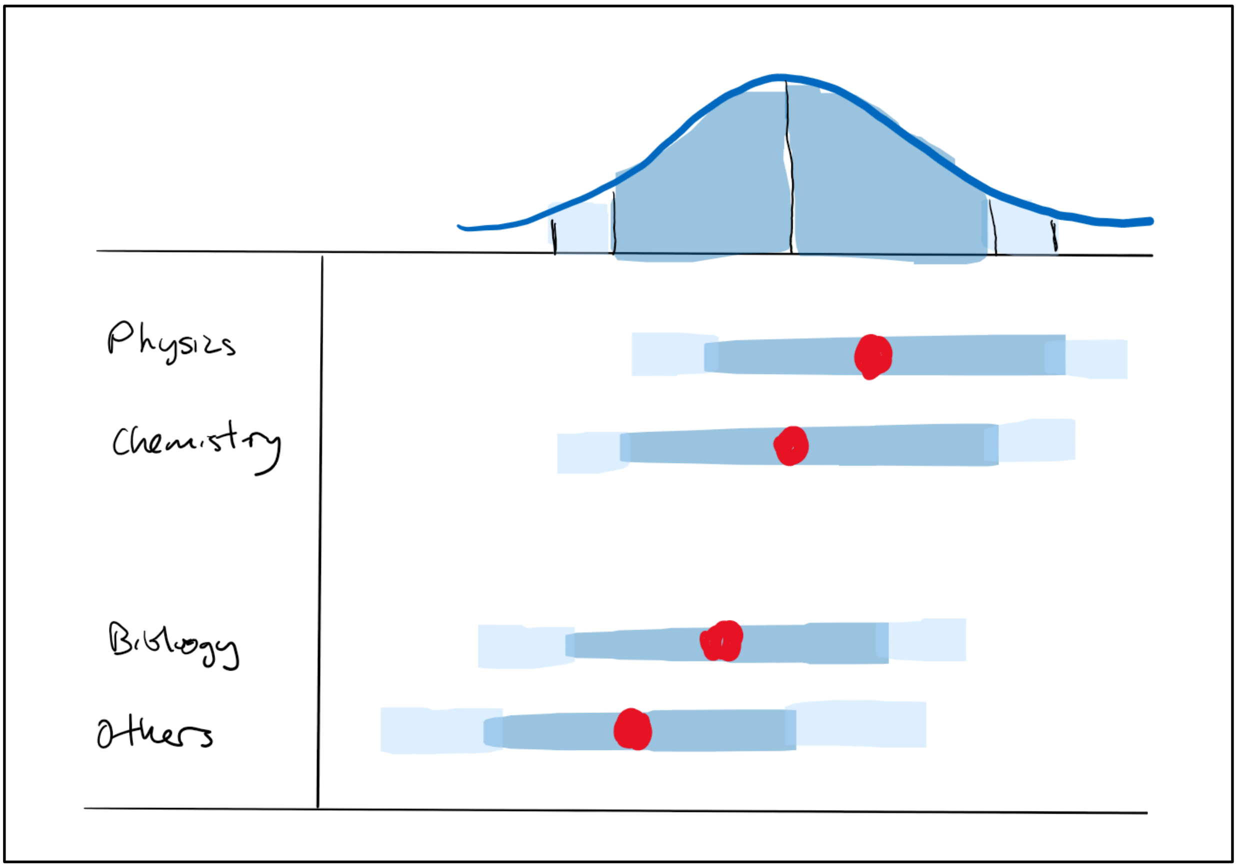
1. Critiques and Suggestions for Current Visualisations



SN	Critique	Suggestion
Clarity		
1.	Graph is referring to the question on "Proportion of published results" by computing the percentage of respondents who selected 70% and higher, grouped by their area of interests. However, the 70% is an arbitrary number which does not reflect statistical measures and why it was chosen in the first place. This manner of visualisation also ignored the proportion of respondents who choose less than 70%.	Explore using statistical values to depict the survey results.
2.	The simple percentage does not reflect the number of respondents for each area of interests, which could be a small sample size where each response will carry a higher weightage.	Explore using statistical values to depict the survey results.
3.	It is redundant to colour the bars distinctly to represent different area of interests as the bars are easily well defined by the axis labels. Similarly as such, the legends are redundant.	Use single colour to depict the bars, or use similar colour gradations to depict the percentage ranges.

4.	Legend is sorted alphabetically and not following the nested sort order of the data, which makes it difficult to cross-reference the bars and legend.	Sort legend following the same manner that the [Are of Interests] field is nested sorted by.
Aesthetics		
SN	Critique	Suggestion
5.	Generally, clear use of fonts, font sizes and layout with most important messages in the top left quadrant. Text are not truncated.	Follow and format to ensure so.
6.	Good axis marks in tens and grid lines to facilitate easy readings and context of bar lengths.	Follow and format to ensure so.
7.	Not efficient use of space, with much white spaces in the top right and bottom right segments.	Remove the legend. Optimise use of space.

2. Proposed Design

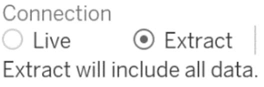
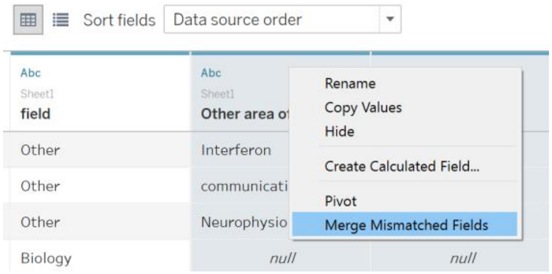
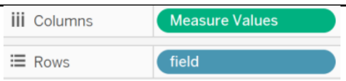


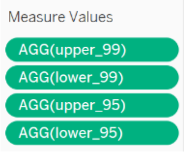
3. Data Visualisation Steps

<table>
<tr><th colspan="2">Data Preparation</th><th></th></tr>
<tr><th>SN</th><th>Area</th><th>Action</th></tr>
<tr><td>1.</td><td>Extract relevant columns of data source.</td><td>Extract only columns A, V, CM to DO.
1,576 records total.</td></tr>
<tr><td>2.</td><td>Rename fields to match</td><td><table><tr><th>Original Header</th><th>Renamed Header</th></tr><tr><td>'In your opinion, what proportion of published results in your field are reproducible? i.e. the results of a given study could be replicated exactly</td><td>response</td></tr></table></td></tr>
</table>

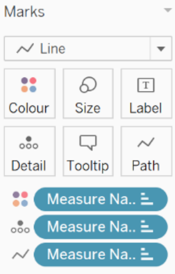
graph text.	or reproduced in multiple similar experimental systems with variations of experimental settings such as materials and experimental model)	
	Which of the following best describes your area of interest?	field

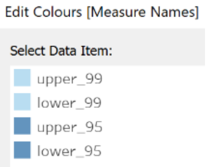
Tableau Works

SN	Area	Action
1.	Import Excel Worksheet and create an extract of the data.	Select an "Extract" of the Connection to facilitate uploading as Tableau Public does not support external files.  Connection ○ Live ● Extract Extract will include all data.
2.	Select all columns after [field] and Merge Mismatched Fields to create a new [subfield] for drilling down of the main [field]. Create new Worksheet "ErrorBars".	 Sort fields Data source order Abc Sheet1 field Abc Sheet1 Other area of Other Other Other Biology Interferon communicati Neurophysio null null Rename Copy Values Hide Create Calculated Field... Pivot Merge Mismatched Fields
3.	[response] - Change Data Type to Number (decimal) Convert to Measure	
4.	Create new Calculated Field [mean]	SUM([response])/SUM([Number of Records])
5.	Create new Calculated Field [zvalue_95]	1.959964
6.	Create new Calculated Field [zvalue_99]	2.575829
7.	Create new Calculated Field [ci_95]	[zvalue_95]*(STDEV([response])/sqrt(COUNT([Number of Records])))
8.	Create new Calculated Field [ci_99]	[zvalue_99]*(STDEV([response])/sqrt(COUNT([Number of Records]))
9..	Create new Calculated Field [lower_95]	[mean]-[ci_95]
10.	Create new Calculated Field [lower_99]	[mean]-[ci_99]
11.	Create new Calculated Field [upper_95]	[mean]+[ci_95]
12.	Create new Calculated Field [upper_99]	[mean]+[ci_99]
13.	Drag [Measure Values] and [field] to the Columns and Rows shelves	 Columns: Measure Values Rows: field

14.	Remove unwanted Measure Values, leaving these.	
		Measure Values
		AGG(upper_99)
		AGG(lower_99)
		AGG(upper_95)
		AGG(lower_95)

15. Change Marks to Lines.
Ctrl+Drag [Measure Names] to Path card.
Ctrl+Drag [Measure Names] to Colour card. Edit Colours

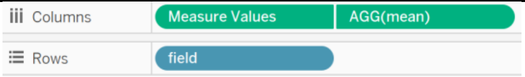



Edit Colours [Measure Names]

Select Data Item:

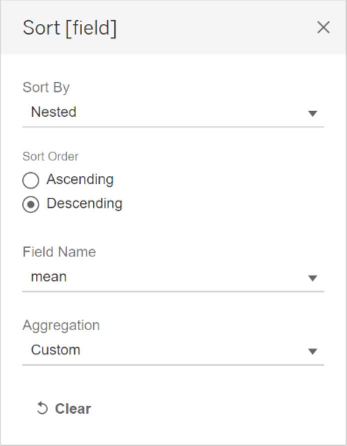
- upper_99
- lower_99
- upper_95
- lower_95

16. Drag [mean] to Columns shelf.
Select Dual Axis and Synchronise Axis.
Change type to Circle.


Columns: Measure Values, AGG(mean)

Rows: field

17. Nested sort [field] using [mean] in Descending order.


Sort [field]

Sort By: Nested

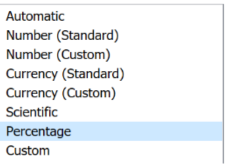
Sort Order
○ Ascending
● Descending

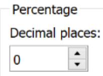
Field Name: mean

Aggregation: Custom

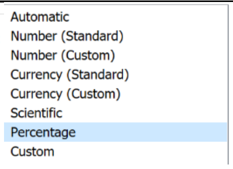
18. Adjust size of line and circle to match display.

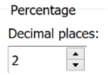
19. Format axis scale to percentage with 0 decimal places.
Format fields to percentage with 2 decimal places.




Percentage

Decimal places: 0

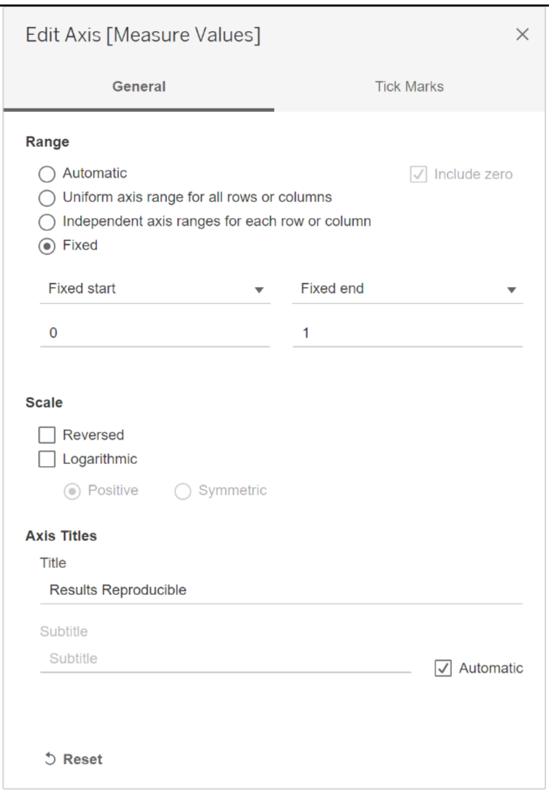



Percentage

Decimal places: 2

20. Analysis > Totals > Show Column Grand Totals
Analysis > Totals > Column Totals to Top

21. Fixed x-axis range to start at 0 and end at 1.
Rename x-axis to “Results Reproducible”.



Edit Axis [Measure Values] ×

General | Tick Marks

Range

○ Automatic ☑ Include zero
○ Uniform axis range for all rows or columns
○ Independent axis ranges for each row or column
● Fixed

Fixed start: 0
Fixed end: 1

Scale

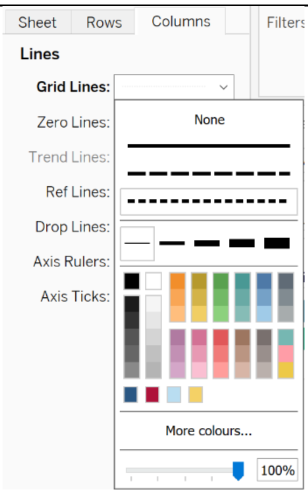
□ Reversed
□ Logarithmic
● Positive ○ Symmetric

Axis Titles

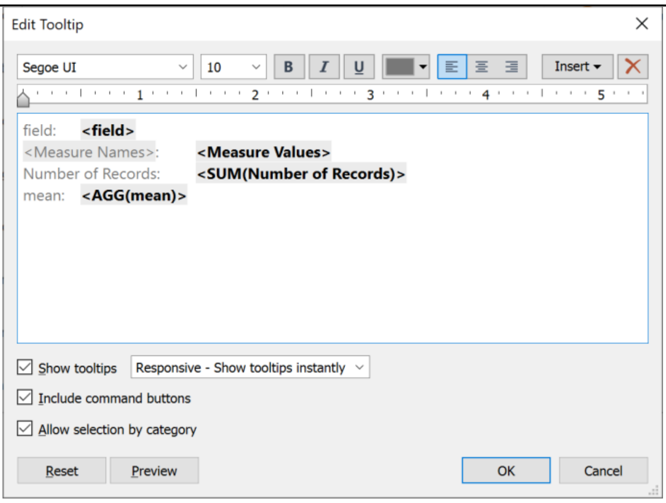
Title: Results Reproducible

Subtitle: Subtitle ☑ Automatic

22. Format Columns Grid Lines.



23. Edit Tooltips to include [Number of records].



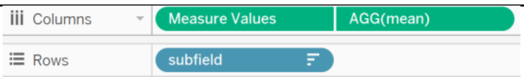
Edit Tooltip ×

Segoe UI | 10 | B | I | U | Insert ▾

field: <field>
<Measure Names>: <Measure Values>
Number of Records: <SUM(Number of Records)>
mean: <AGG(mean)>

☑ Show tooltips | Responsive - Show tooltips instantly
☑ Include command buttons
☑ Allow selection by category

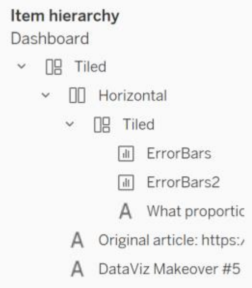
24. Duplicate “ErrorBars” to new Worksheet “ErrorBars2”.



Columns: Measure Values | AGG(mean)
Rows: subfield

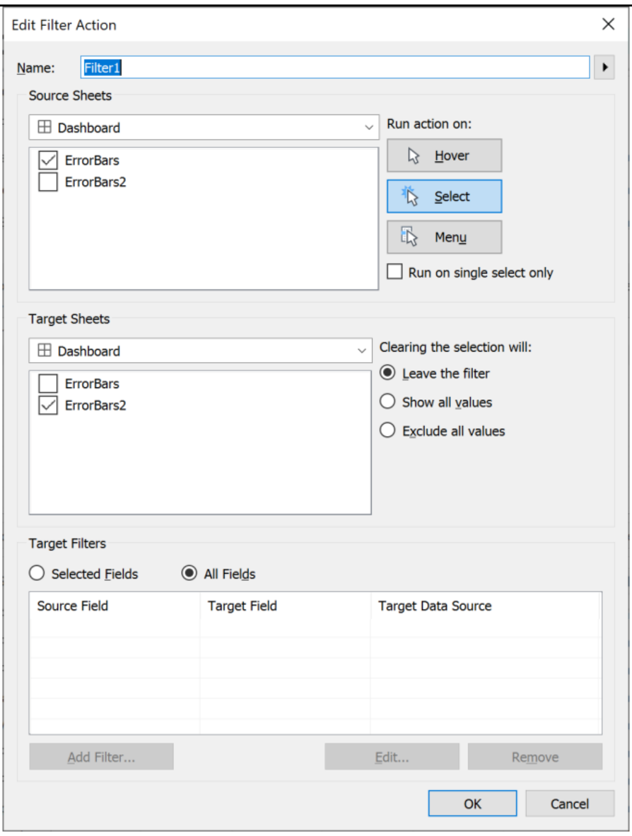
Replace Rows shelf with measure [subfield]

25. Create new Dashboard and layout accordingly.



Item hierarchy
Dashboard
- Tiled
  - Horizontal
    - Tiled
      - ErrorBars
      - ErrorBars2
      - What proportic
  - Original article: https:/
  - DataViz Makeover #5

26. Create new Dashboard Filter Action to show subfield details of selected field.



Edit Filter Action

Name: Filter1

Source Sheets

Dashboard

☑ ErrorBars
☐ ErrorBars2

Run action on: Hover / **Select** / Menu

☐ Run on single select only

Target Sheets

Dashboard

☐ ErrorBars
☑ ErrorBars2

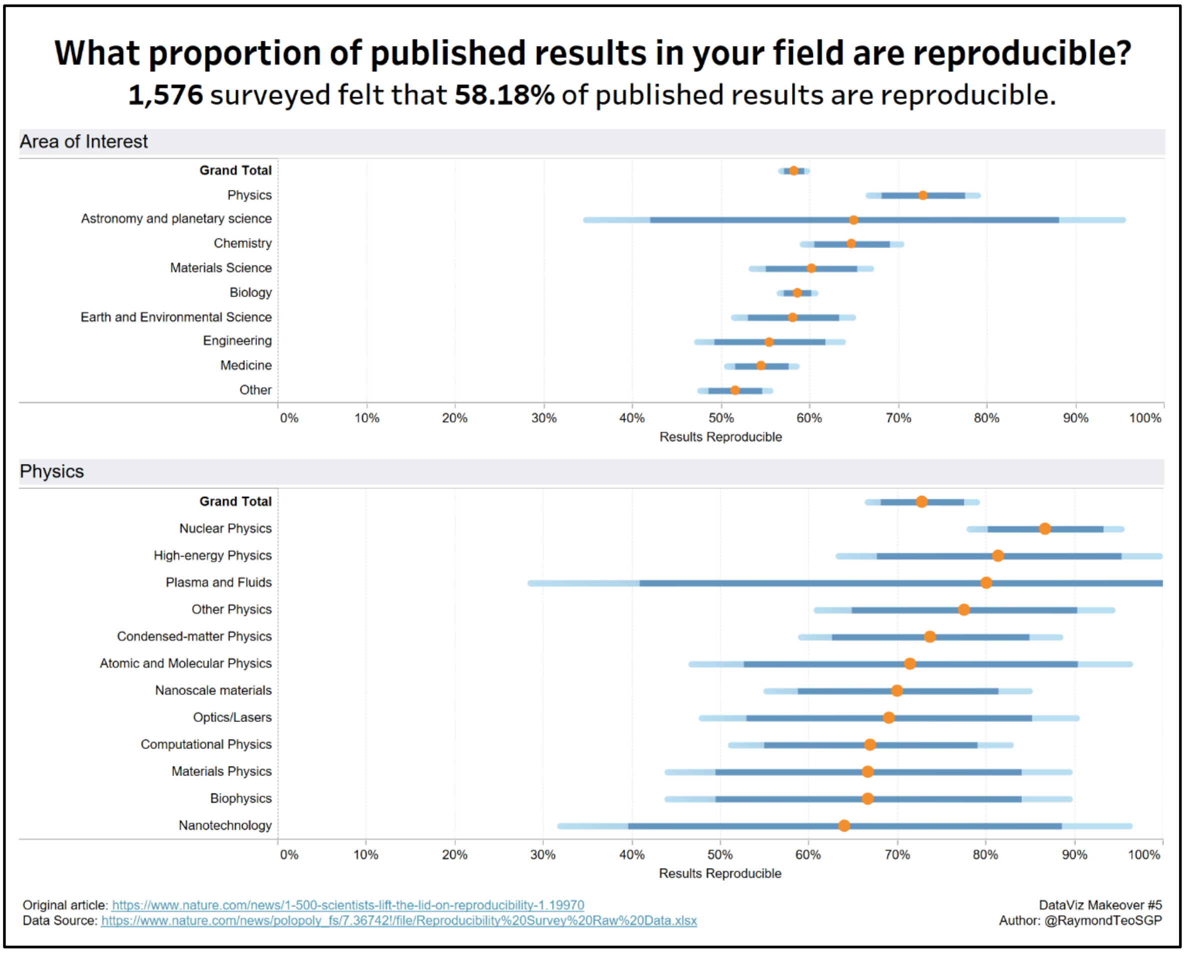
Clearing the selection will:
● Leave the filter
○ Show all values
○ Exclude all values

Target Filters

○ Selected Fields ● All Fields

Source Field	Target Field	Target Data Source

4. Final Data Visualisation Output



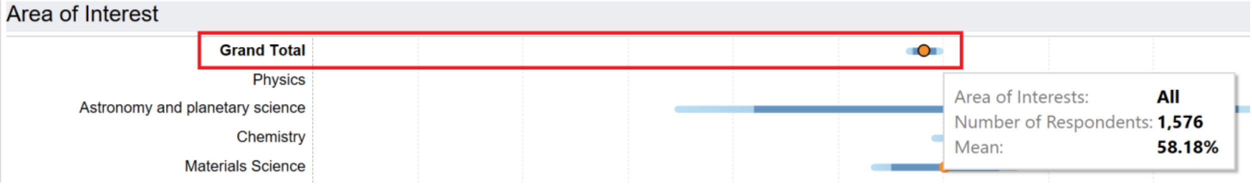
Original article: https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970
Data Source: https://www.nature.com/news/polopoly_fs/7.36742!/file/Reproducibility%20Survey%20Raw%20Data.xlsx

DataViz Makeover #5
Author: @RaymondTeoSGP

Insights

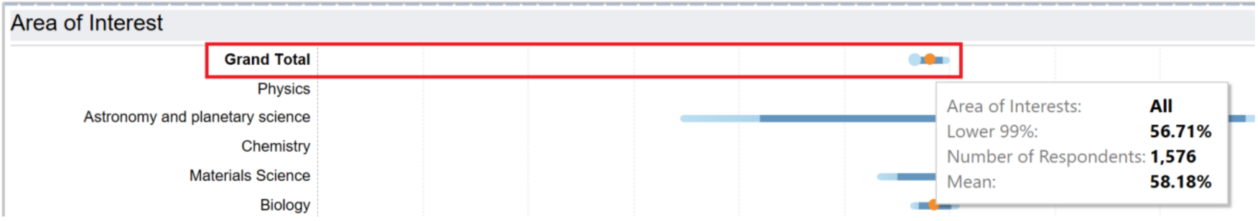
SN Insight

1. 1,576 were surveyed. They are confident that 58.18% of the published results are reproducible, at a small standard error.



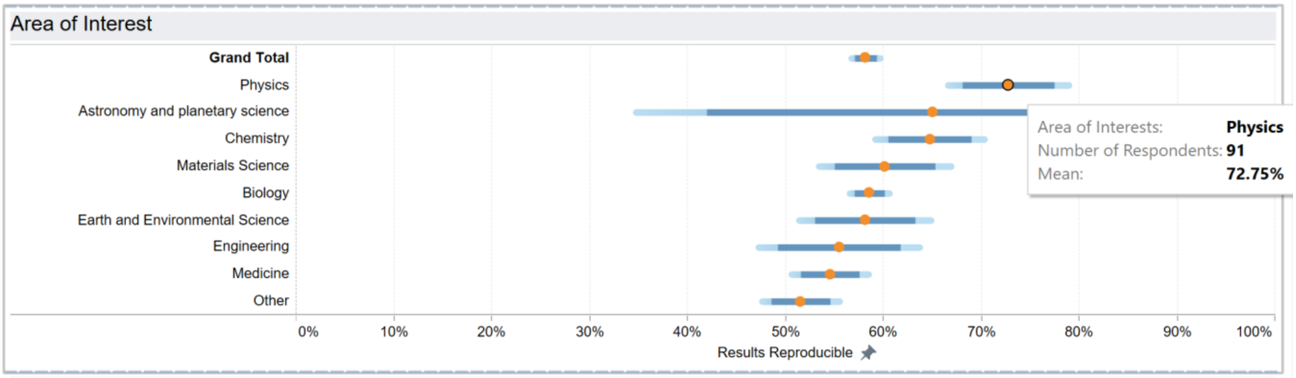
Area of Interests: All
Number of Respondents: 1,576
Mean: 58.18%

The lower 99% confidence interval is 56.71%, which is greater than 50%. This shows that it is of significance that the results are reproducible (greater than 50%).



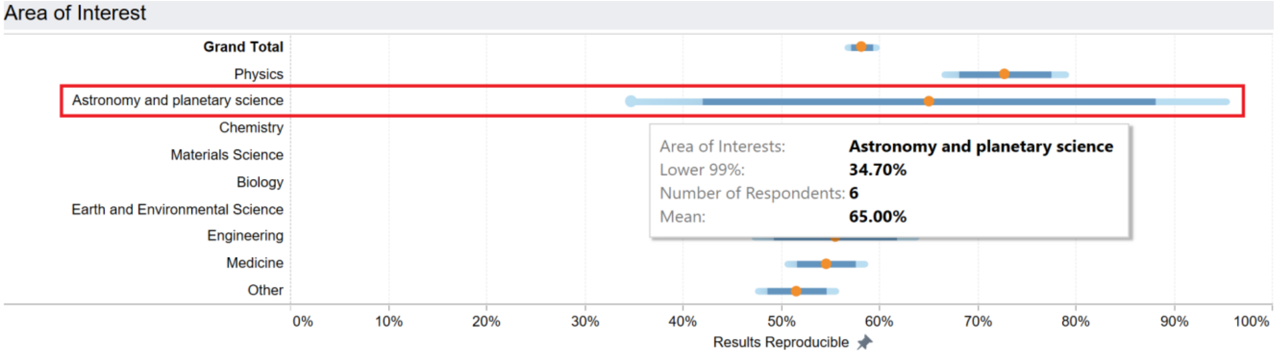
Area of Interests: All
Lower 99%: 56.71%
Number of Respondents: 1,576
Mean: 58.18%

2. Physics has the highest mean that 56.71% of the published results are reproducible, followed by the rest as ranked in descending order of the mean.



Area of Interests: Physics
Number of Respondents: 91
Mean: 72.75%

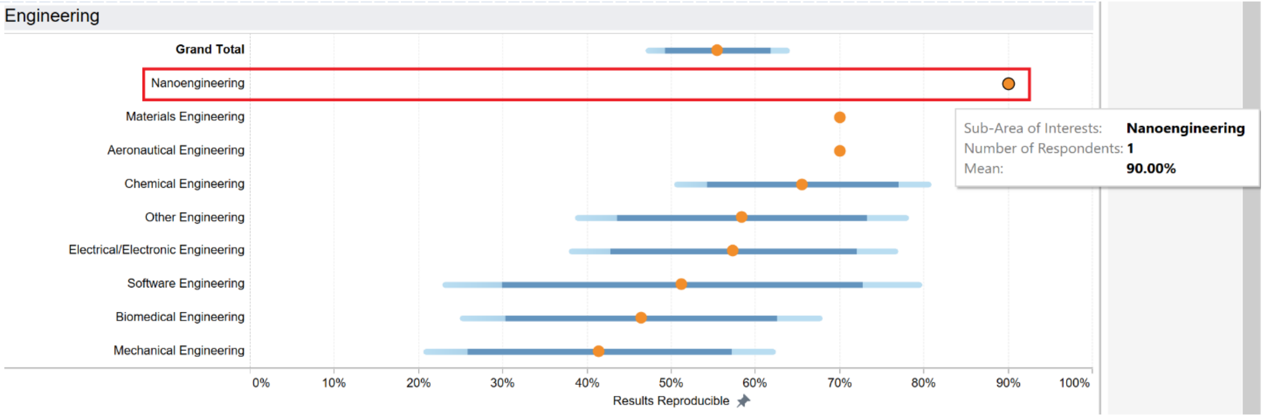
However, Astronomy and planetary science has a small sample size of 6 respondents only. As a result, the confidence interval is very wide. It should probably be combined with the Other category rather than be analysed on it's own.



Area of Interests: Astronomy and planetary science
Lower 99%: 34.70%
Number of Respondents: 6
Mean: 65.00%

3. When clicking any area of interests, the dashboard action will drill down to the sub – area of interests to display their means and error bars below. This is to allow looking at a lower level of details to understand the composition better.

However, at the sub – area of interests level, most of the sample sizes are small and should not be analysed using the normal distribution. Some of the sub – area of interests has only 1 respondent due to the niche area.



Sub-Area of Interests: Nanoengineering
Number of Respondents: 1
Mean: 90.00%

4. Am not able to reproduce the probability distribution curve currently as it is not a standard chart in Tableau and requires R extensions.