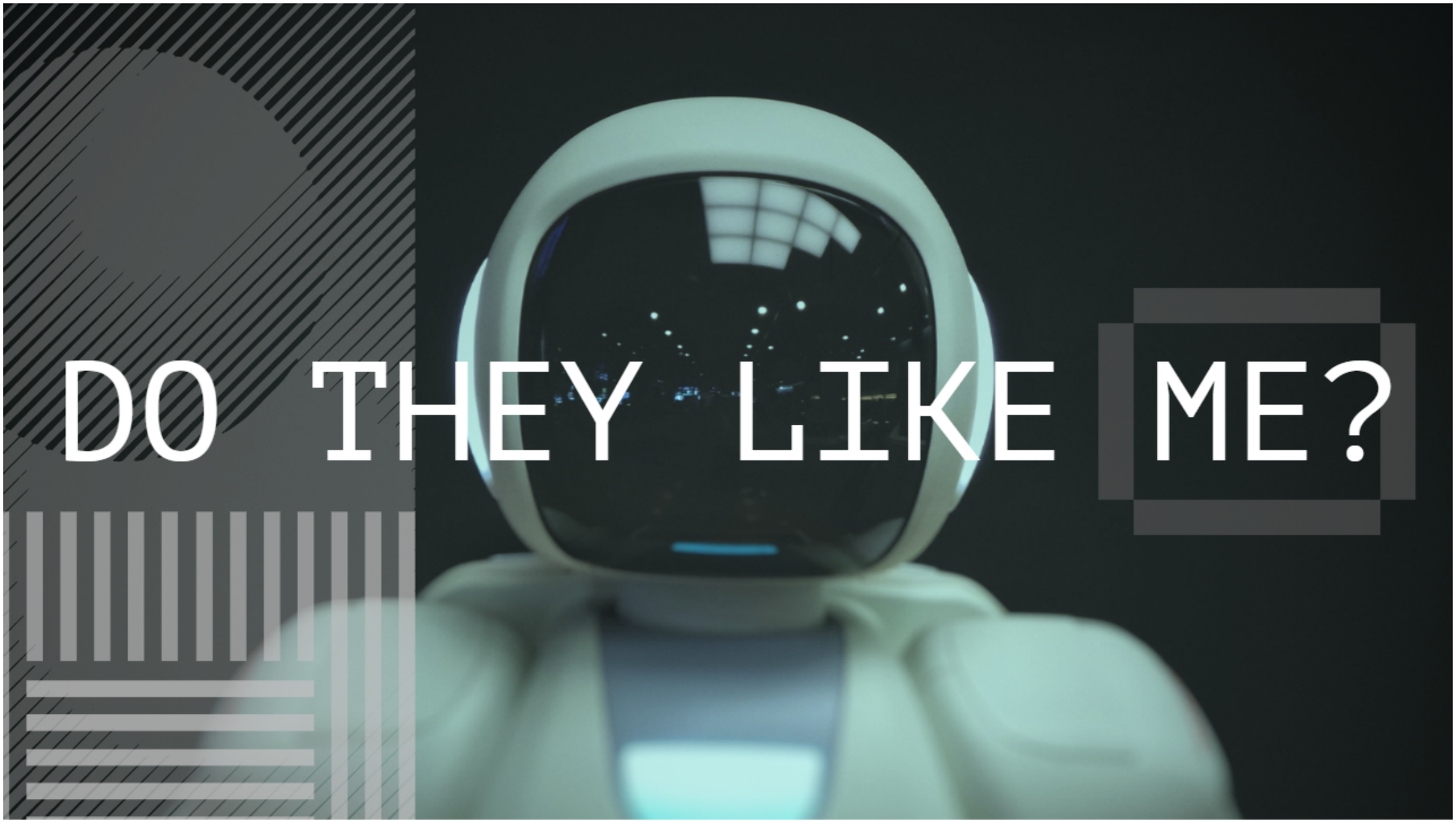
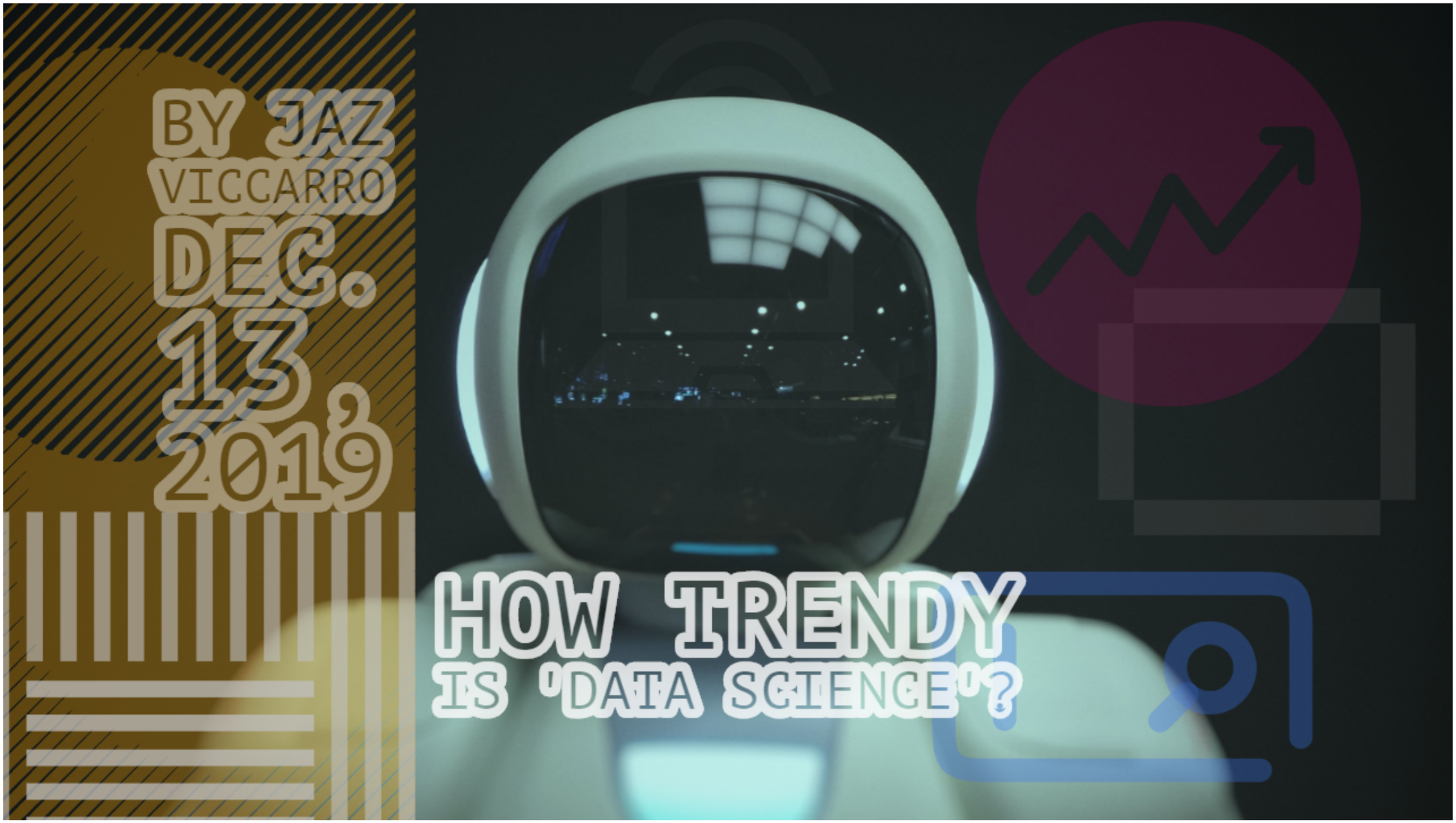
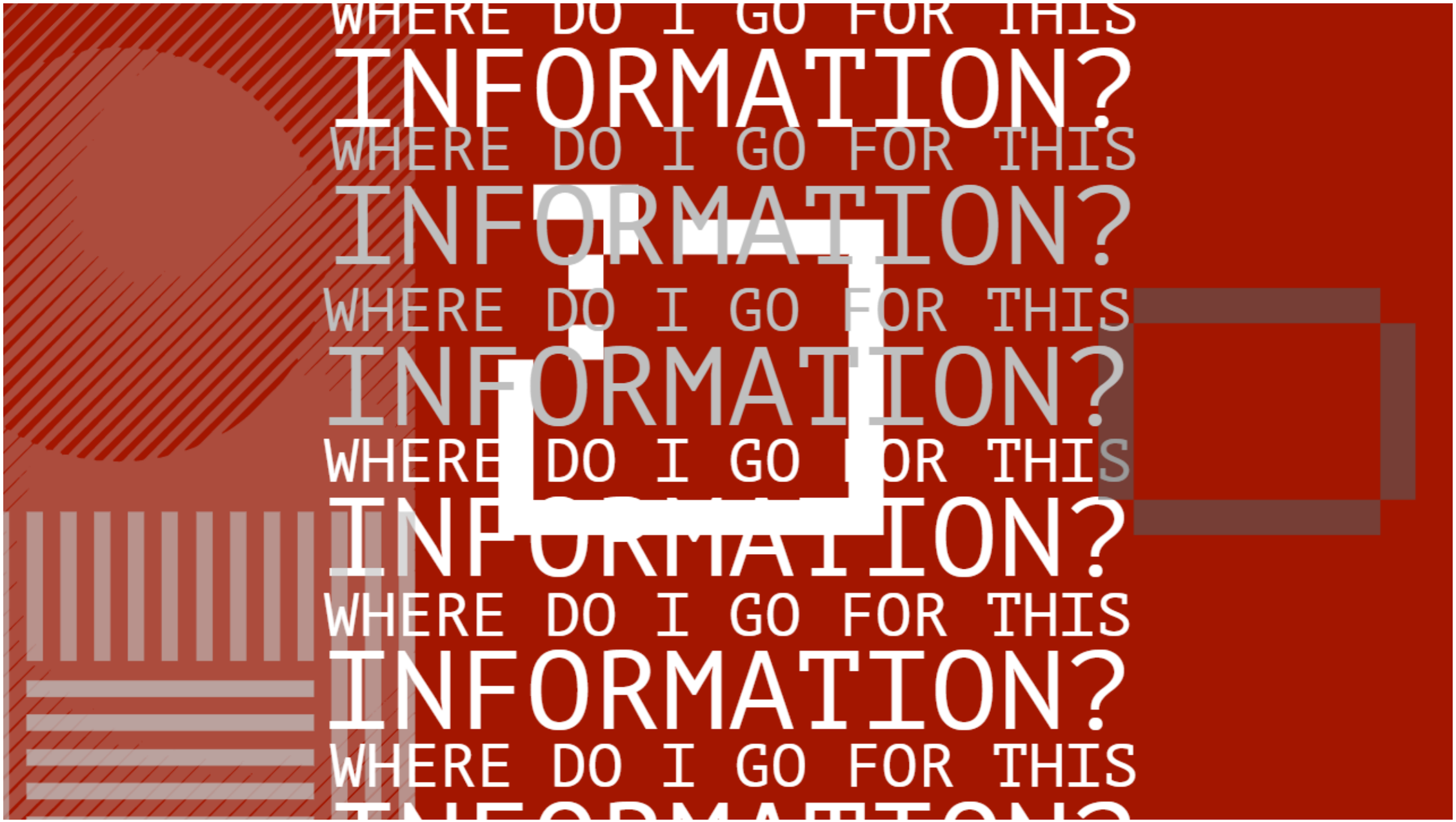

DO THEY LIKE ME?


BY JAZ VICCARRO
DEC. 13 2019
HOW TRENDY
IS 'DATA SCIENCE'?


CAN GOOGLE TELL US HOW
INTERESTING DATA SCIENCE IS?


WHERE DO I GO FOR THIS
INFORMATION?
WHERE DO I GO FOR THIS
INFORMATION?
WHERE DO I GO FOR THIS
INFORMATION?
WHERE DO I GO FOR THIS
INFORMATION?
WHERE DO I GO FOR THIS
INFORMATION?
WHERE DO I GO FOR THIS

OVERVIEW OF DATA

Datasets -Interest Over Time

Search Terms:

- Data Scientist
- Data Science
- Both Categories

Google Trends

United States

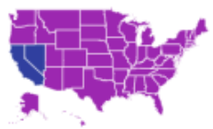
Explore what the world is searching

Enter a search term or a topic

Or start with an example

HIDE

Taylor Swift Kim Kardashian



Interest by subregion, Past 7 days, United States

World Cup

Interest by region, Past 7 days, Worldwide

Football American football

Interest by subregion, 2004 - present, United States

Showing 1-3 of 6 examples

WE WOULD
LIKE TO KNOW...

HAS INTEREST IN DATA SCIENCE
INCREASED OVER TIME?
WELL--YES!

Out[1]: The raw code for this IPython notebook is by default hidden for easier reading. To toggle on/off the raw code, click here.

Hypothesis:

Has interest in Data Science increased or decreased?

Has interest in Data Science increased or decreased since the release of Harvard Business Review article, naming 'Data Scientist' the sexiest job in the 21st Century?

ACCORDING TO GOOGLE, INTEREST LEVEL IS...

"INTEREST OVER TIME NUMBERS REPRESENT SEARCH INTEREST RELATIVE TO THE HIGHEST POINT ON THE CHART FOR THE GIVEN REGION AND TIME. A VALUE OF 100 IS THE PEAK POPULARITY FOR THE TERM. A VALUE OF 50 MEANS THAT THE TERM IS HALF AS POPULAR. A SCORE OF 0 MEANS THERE WAS NOT ENOUGH DATA FOR THIS TERM."

Out[4]:

<bound method DataFrame.info of

	time	data_scientist	data_science	computer	food
0	2012-01	8	12	100	58
1	2012-02	8	10	93	58
2	2012-03	8	9	90	60
3	2012-04	7	10	88	60
4	2012-05	8	9	90	61
5	2012-06	8	6	89	61
6	2012-07	8	6	94	62
7	2012-08	11	12	95	60
8	2012-09	18	17	96	62
9	2012-10	12	12	88	58
10	2012-11	9	12	91	60
11	2012-12	10	10	91	58
12	2013-01	12	14	92	59
13	2013-02	12	12	88	59
14	2013-03	15	11	87	60
15	2013-04	23	14	83	60
16	2013-05	19	14	80	60
17	2013-06	15	11	81	61
18	2013-07	14	10	82	59
19	2013-08	19	15	87	59
20	2013-09	24	23	89	58
21	2013-10	18	19	84	60
22	2013-11	15	18	85	61
23	2013-12	15	16	86	57
24	2014-01	18	21	88	59
25	2014-02	23	22	85	58
26	2014-03	19	21	81	60
27	2014-04	22	20	78	59
28	2014-05	23	21	76	60
29	2014-06	24	17	76	58
..	...	...	...	...	...
66	2017-07	66	51	60	82
67	2017-08	67	65	64	79
68	2017-09	77	71	67	77
69	2017-10	71	72	64	80
70	2017-11	65	65	68	78
71	2017-12	57	55	65	79
72	2018-01	75	71	65	78
73	2018-02	72	67	62	78
74	2018-03	80	68	62	82
75	2018-04	76	75	61	85

```
76  2018-05              82             68        58    85
77  2018-06              77             62        55    86
78  2018-07              81             61        56    91
79  2018-08              90             74        62    89
80  2018-09              90             86        62    85
81  2018-10              87             83        59    83
82  2018-11              78             74        60    84
83  2018-12              65             65        59    87
84  2019-01              92             86        60    88
85  2019-02              89             83        59    86
86  2019-03              84             82        56    90
87  2019-04              88             84        54    91
88  2019-05              90             77        53    93
89  2019-06              80             76        52    99
90  2019-07              84             72        51   100
91  2019-08             100             90        57    96
92  2019-09              98            100        59    90
93  2019-10              92             94        56    91
94  2019-11              80             83        57    90
95  2019-12              90             87        62    83

[96 rows x 5 columns]>
```

Out[5]:

	time	data_scientist	data_science	computer	food
0	2012-01	8	12	100	58
1	2012-02	8	10	93	58
2	2012-03	8	9	90	60
3	2012-04	7	10	88	60
4	2012-05	8	9	90	61

Before Let's change our data types: make the week date into something more standard and make the level of interest into integers

Out[6]: `dtype('O')`

```
Out[7]: 0   2012-01-01
        1   2012-02-01
        2   2012-03-01
        3   2012-04-01
        4   2012-05-01
        Name: time, dtype: datetime64[ns]
```

When we first gathered and looked at our data,

- we had interest level broken down by month only...

Procedure We Used:

We needed to change our data into a different type, and, add some columns, manipulate the data, visualize and test results.

```
        - python, numpy, scipy and pandas:
        - filtering, grouping, sorting
        - aggregations, descriptive stats
        - tests for statistical analysis
    - feature design:
        - year (2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019)
        - month (numerical; 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12),
        - month (name: Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sept, Oct, Nov, Dec
        - weekday (numerical: 0, 1, 2, 3, 4, 5, 6)
        - weekday (name: Mon, Tues, Wed, Thurs, Fri, Sat, Sun)
        - quarter (financial: 1, 2, 3, 4)
     - visualize discrete features (new columns) as a y-axis variables:
         - Matplotlib and Seaborn libraries
             - Histograms, Bar charts, Line Charts
```

Average Interest Level by Day of the Week:

- "Best Day" for Data Science: Thursday (Avg. 44.85)

- "Best Day" for Data Scientist: Monday (Avg. 49.23)

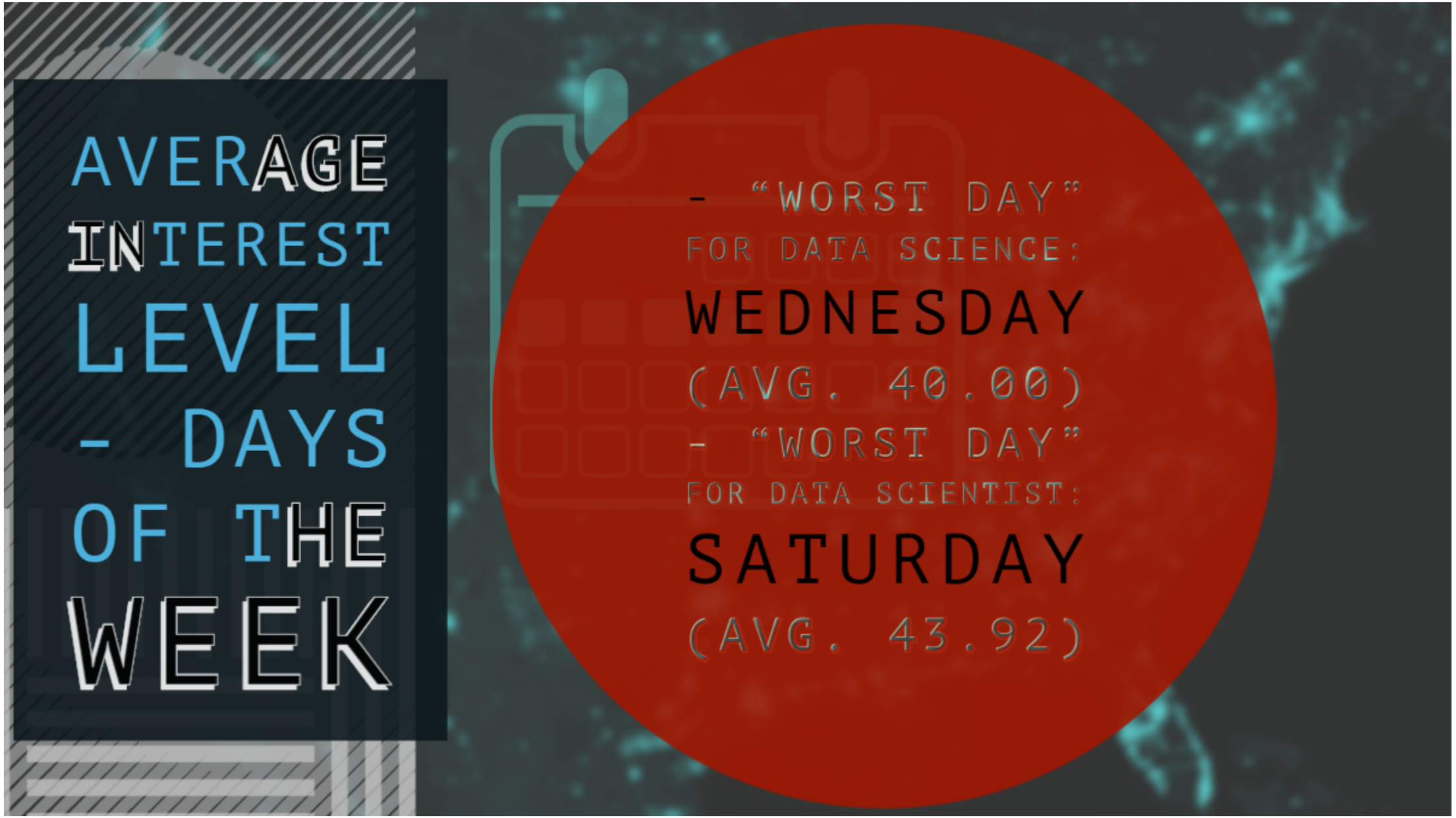
- "Worst Day" for Data Science: Wednesday (Avg. 40.00)

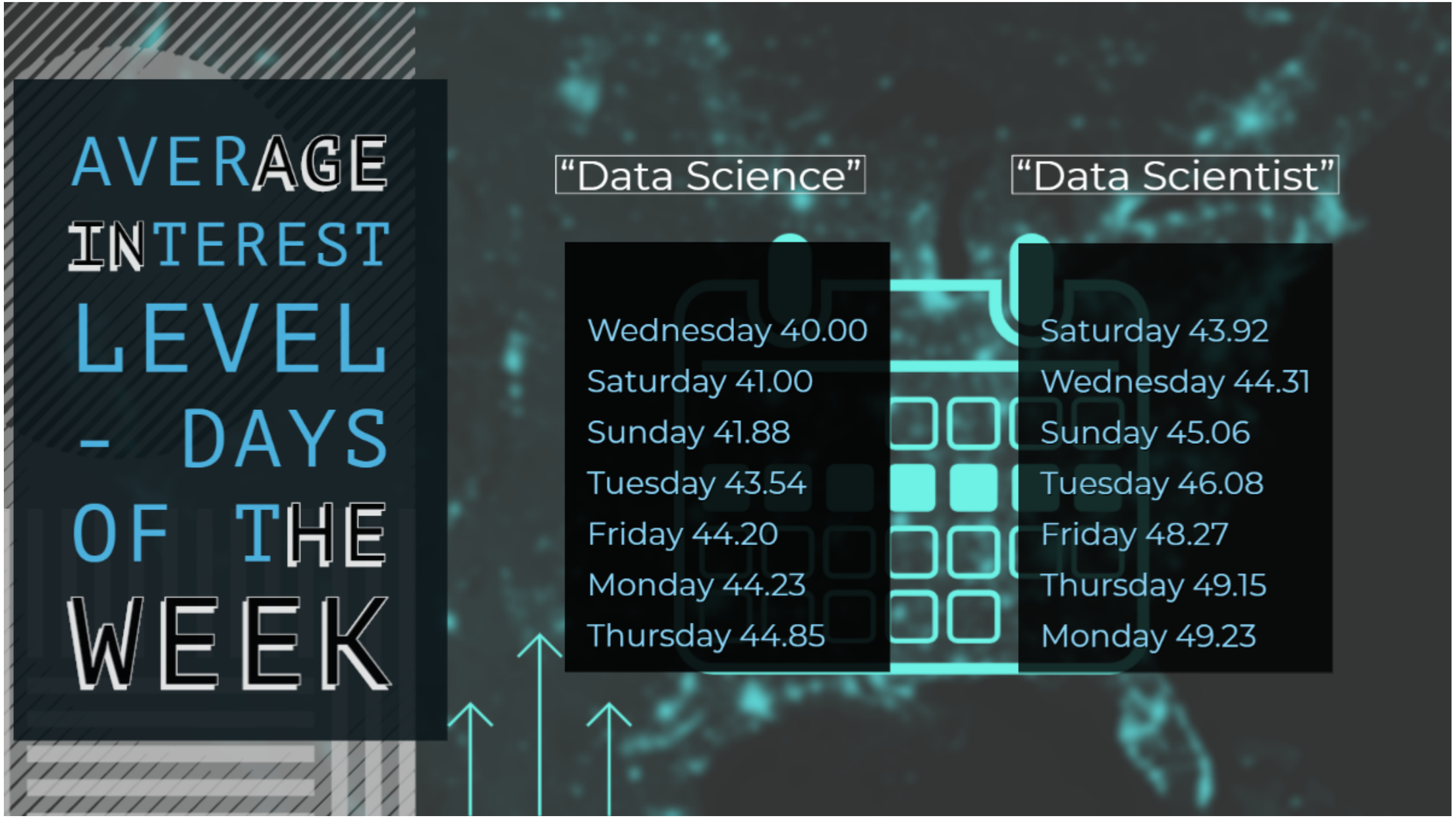
- "Worst Day" for Data Scientist: Saturday (Avg. 43.92)

AVERAGE INTEREST LEVEL - DAYS OF THE WEEK

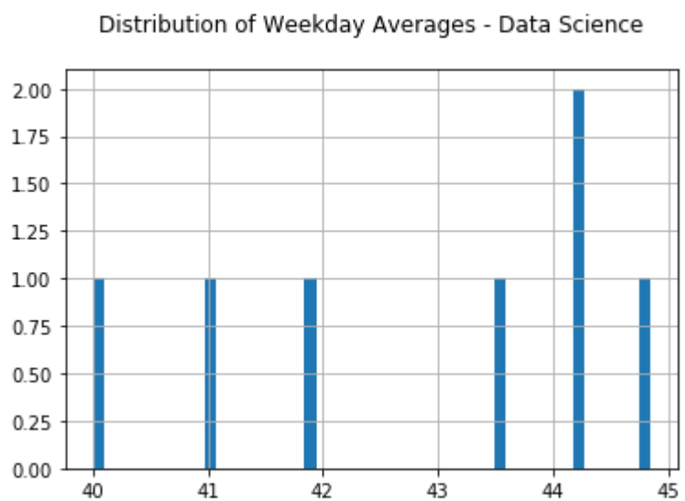
"BEST DAY" FOR DATA SCIENCE:
- **THURSDAY** (AVG. 44.85)

"BEST DAY" FOR DATA SCIENTIST:
- **MONDAY** (AVG. 49.23)

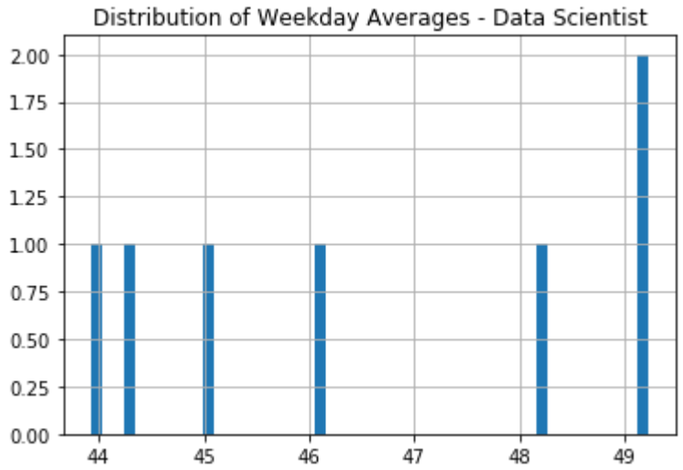

AVERAGE INTEREST LEVEL - DAYS OF THE WEEK
- "WORST DAY" FOR DATA SCIENCE:
WEDNESDAY
(AVG. 40.00)
- "WORST DAY" FOR DATA SCIENTIST:
SATURDAY
(AVG. 43.92)

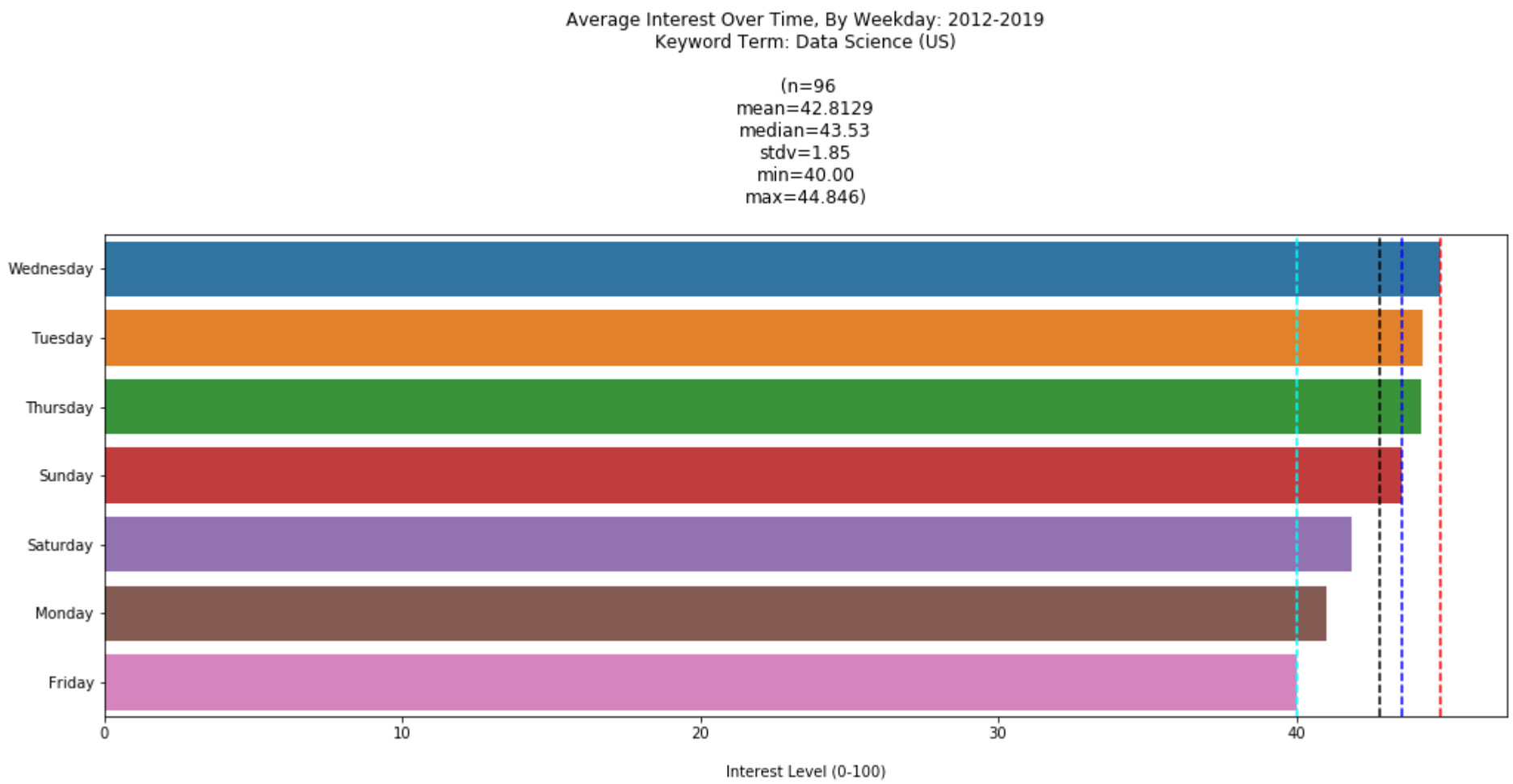

AVERAGE INTEREST LEVEL - DAYS OF THE WEEK
"Data Science"
Wednesday 40.00
Saturday 41.00
Sunday 41.88
Tuesday 43.54
Friday 44.20
Monday 44.23
Thursday 44.85
"Data Scientist"
Saturday 43.92
Wednesday 44.31
Sunday 45.06
Tuesday 46.08
Friday 48.27
Thursday 49.15
Monday 49.23

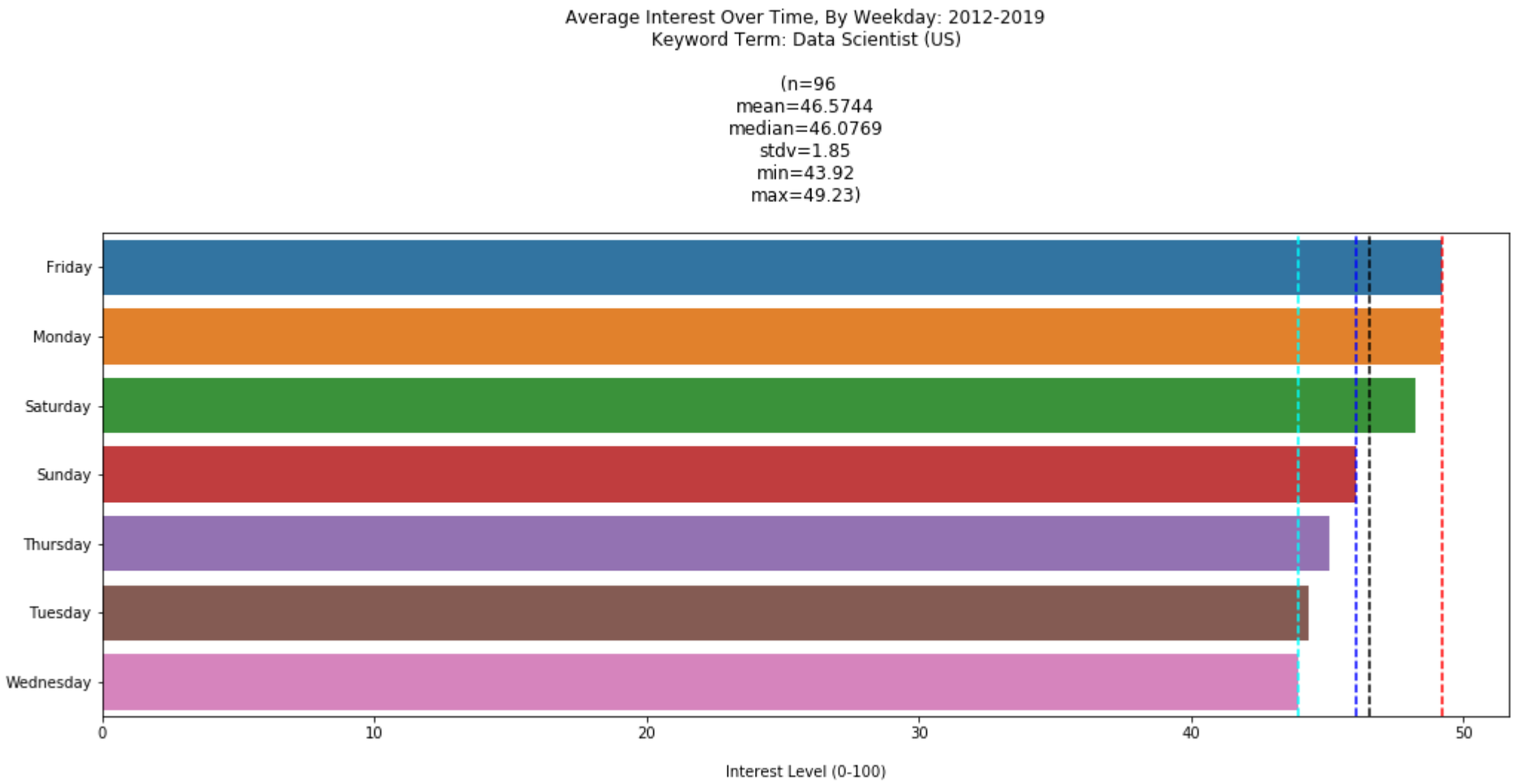
<Figure size 1080x1152 with 0 Axes>

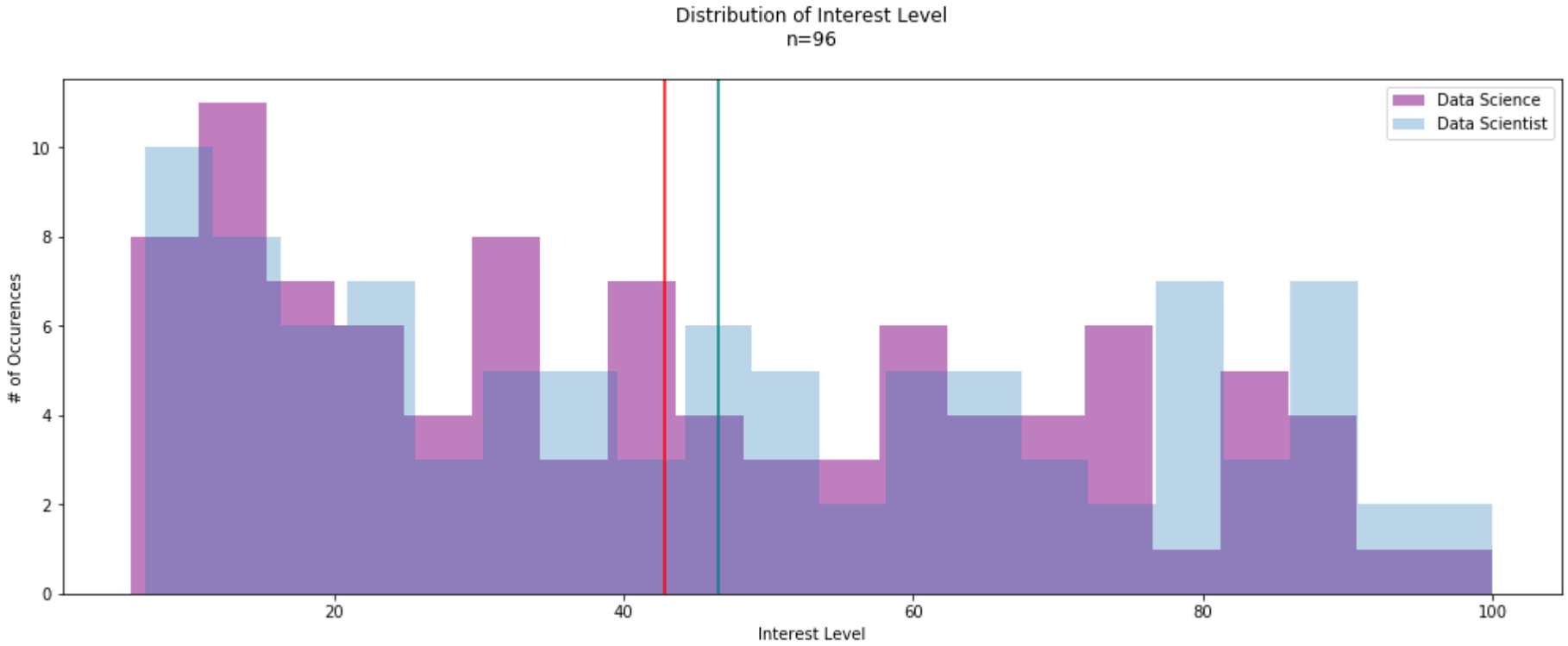


<Figure size 1440x1440 with 0 Axes>




Average Interest Over Time, By Weekday: 2012-2019
Keyword Term: Data Science (US)
(n=96
mean=42.8129
median=43.53
stdv=1.85
min=40.00
max=44.846)
Wednesday
Tuesday
Thursday
Sunday
Saturday
Monday
Friday
0
10
20
30
40
Interest Level (0-100)

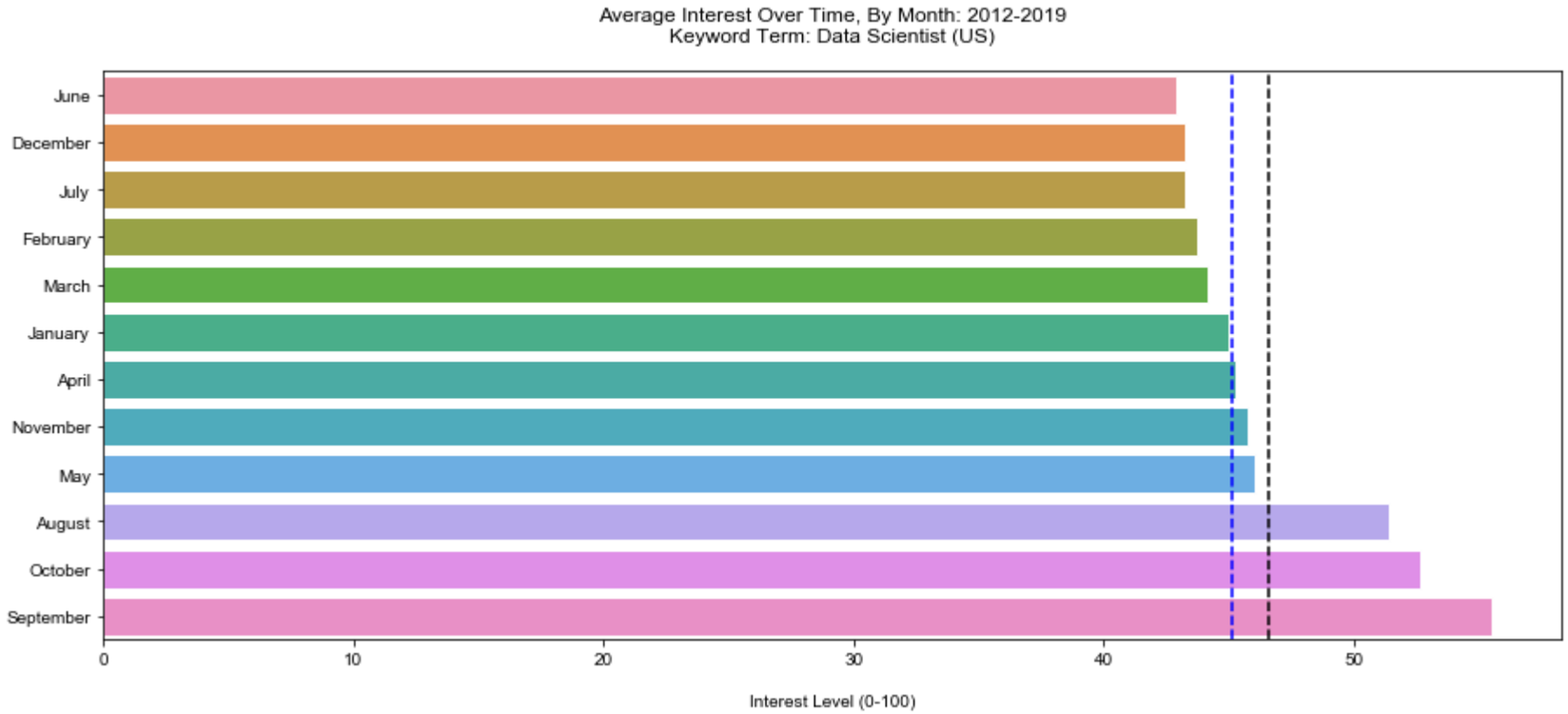

Average Interest Over Time, By Weekday: 2012-2019
Keyword Term: Data Scientist (US)
(n=96
mean=46.5744
median=46.0769
stdv=1.85
min=43.92
max=49.23)
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
0
10
20
30
40
50
Interest Level (0-100)

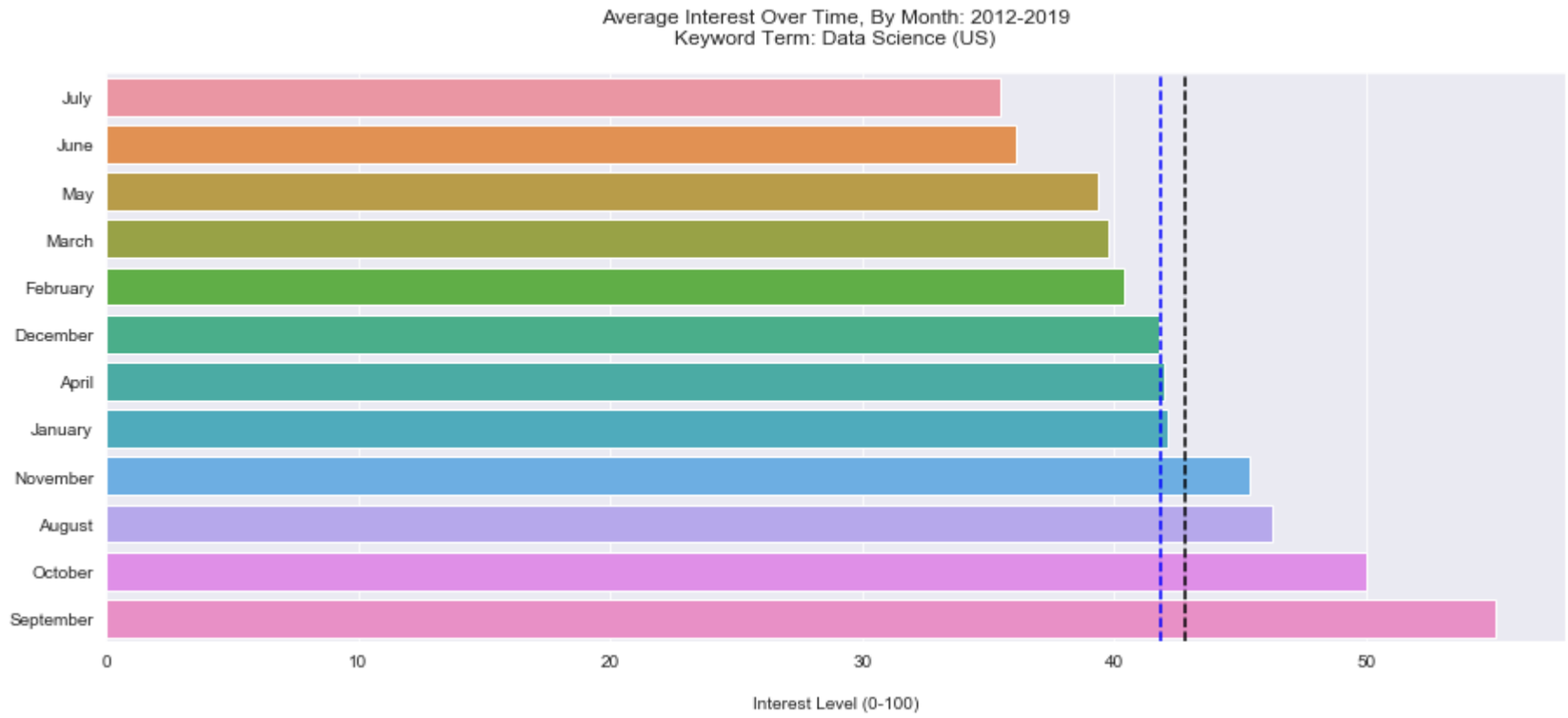


Unfortunately, Our data is very not normal...

And, if we wanna test our hypothesis, we'll need more data!

But, let's see some more charts to highlight relationships. Let's look at Averages By Month--


Average Interest Over Time, By Month: 2012-2019
Keyword Term: Data Scientist (US)
June
December
July
February
March
January
April
November
May
August
October
September
0
10
20
30
40
50
Interest Level (0-100)



The bar graph shows a few things:

1. Our month with the highest average level of interest is September in both Keyword categories;

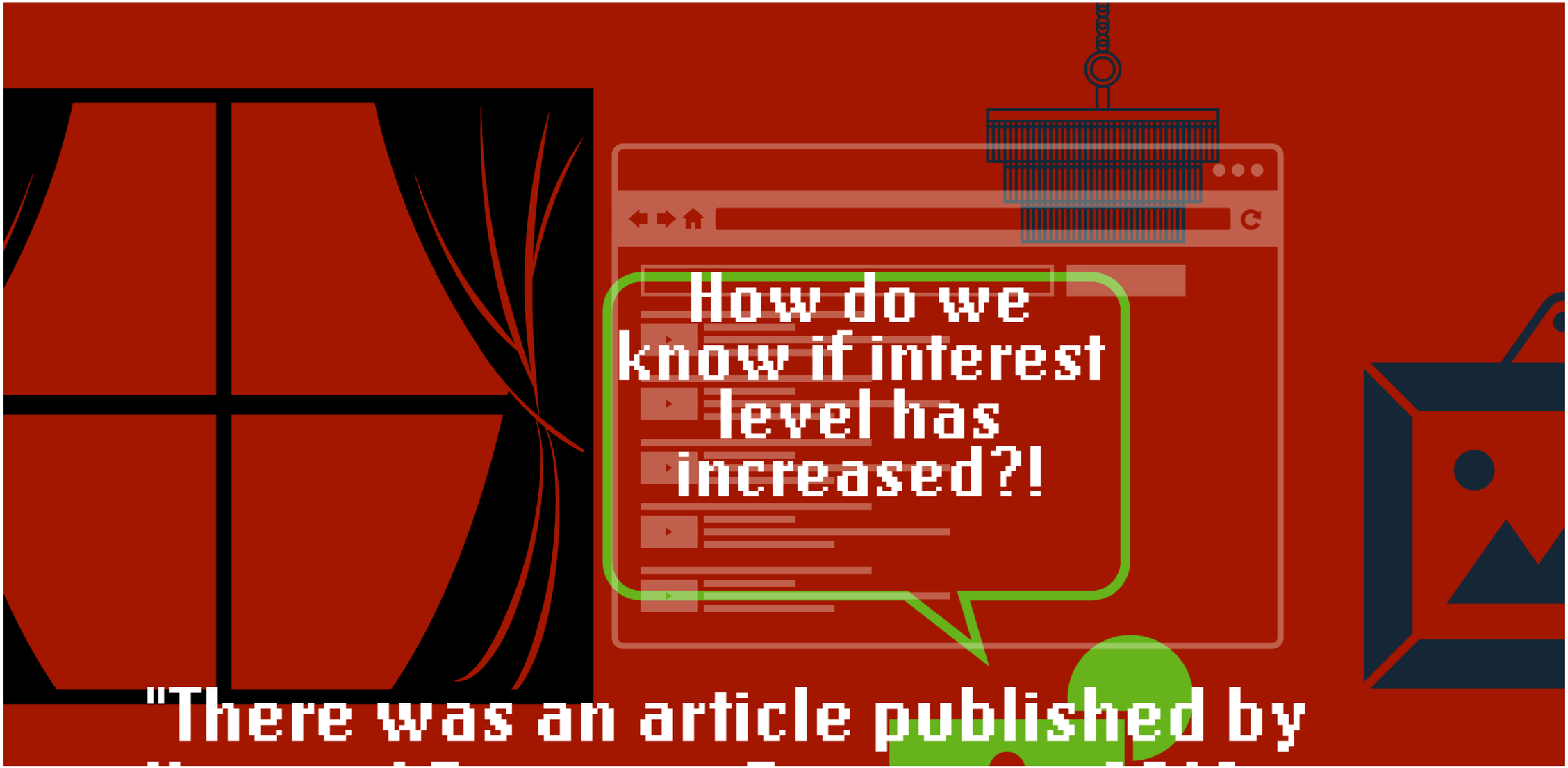
2. Our Month with the lowest average level of interest is July in both Keyword categories

3. In our control group, June was also the worst month for 'Computer'; January the worst for 'Love'

Here we'll see the trend increase by year, validating, visually that interest level in Data Science has increased over time...

As we saw above, the charts are pretty, and we see increase over time, which answers our first question. Without normality, our analysis was more focused on the *time* variable. However, this did not require much science...we can't test this dataset!

Next Hypothesis!--Let's See if we can answer our second question--


How do we know if interest level has increased?!
"There was an article published by

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

From the October 2012 Issue

Summary Save Share 17 Comment Print $8.95 Buy Copies

Oh no! looks like we need more years!

 # Our last dataset only had 96 observations

To make this hypothesis more testable, we'll need more observations (years) from Google Trends

Time for some science...! :)

Let's complete our analysis and experimentation in Jupyter Notebook.

Adding in data from 2004, we now have 192 observations!

For clarity, we'll stick with the Keyword "Data Scientist" to centralize our analysis within the context of career/job title vs. the subject itself.

Let's make two groups:

Group A will be before the Harvard Business Review Article (n=106)

"All interest levels, by month, for 'Data Scientist' from Jan 1. 2004 until Oct 1. 2012"

Group B will be after the Harvard Business Review Article (n=86)

"All interest levels, by month, for 'Data Scientist' from Nov 1. 2012 until Dec 1. 2019"

Can we see if there's any difference in the numbers--

Here are our stats for Group A:

count 106.000000
mean 2.792453
std 2.728045
min 0.000000
25% 1.000000
50% 2.000000
75% 3.000000
max 17.000000

Here are our stats for Group B:

count 86.000000
mean 50.441860
std 25.564039
min 10.000000
25% 27.500000
50% 47.500000
75% 75.500000
max 100.000000

Above we can see the difference in mean in the two groups. Let's see how this is visualized in a box plot:

Our box plots show us two things:

1) in the first boxplot we see a lot of outliers!

- This could be due to the fact that as our x-axis (time) increases, we see a increase in interest levels

2) In our second box plot, we see no outliers.

- Perhaps this means there's not only an increase in interest level, but also more years where interest levels were higher than the average seen before the HBR article in 2012.

Let's test for normality across groups--

First let's plot histograms for group's a and b--

The distributions of both groups look not normal.

So let's confirm this by printing out some descriptive statistics.

Again, these results are not in line with normality-- perhaps we can use further methods to test for normality--

Given that our test confirms an extremely low level of significance (p=6.21e-16), we're rejecting the null hypothesis.

we should reconsider answering our original question...

It's very likely that the observations seen in the experiment are not reflective of the population.

But to be sure, let's perform on last test to confirm our results further

We can further test this result by using the Shapiro-Wilk Test for Normality:

Results

A **Shapiro-Wilk** test confirms that unfortunately, if we were to use the raw Google Trends data to answer our original questions, the likelihood of our observations reflecting in truth in the general population, that our data is showing a statistically significant increase in trend is <1%.

p-value = 2.212...e-11

Conclusion/Discussion:

- Based on this research, it's more clear that relying on raw data from Google Trends is not a scientifically-backed decision-maker

- Motivations
  - validation
  - Socio-economic
  - Bio-cognitive[1] *(demonstrates how thoughts and their biological expression coemerge within a cultural history)*
    - Imposter Syndrome
  - Philosophical - Metaphysics
    - "What's real and what's not?"
    - How differentiated is the human from it's machine/the internet?
  - Scientific/Technological
    - Quantization[2]
    - Geographic Information intelligence
    - Stationarity of Time-Series Data

Conclusion/Discussion Cont'd:

- Opportunities for further research
  - Would like to add a map of the country
  - more "time buckets"
- Biases
  - Search Engine Market Share
  - Radical Anonymity
- Shifts in Perspectives
  - As communication and spread of information has increased... ...our reliance on the internet to debunk life's mysteries should be examined on a interpersonal level.

Discussion:

We've discovered that, using our hypothesis as a guide: has interest in Data Science increased or decreased since the release of the Harvard Business Review article in Oct. 2012--that our hypothesis must be rejected. Since the difference between Group A--before the HBR article, and Group B--afer the HBR article has been tested, our p-value indicates that the observations seen in these Groups are not statistically significant enough to substantiate the use of Google Trends data to validate their use.

Since we are only looking at Google Trends data for the search term "Data Science" only, there may be the bias that we are not collecting enough data on the entire picture that should be captured if we are to truly attempt investigating the status quo of the general public's stance on issues and topics. Additionally, this experiment may lead to the bias of non-representative sample size given that the data only reflects the search volume from the United States and only from Google data, which only represents a specific fraction of the internet search engine market share.

It may be useful to tie in other variables like other keywords or other points within the time-series that might help normalize the sample data where we could then perform other tests.

The experimenter found the results slightly telling, as it's hard to stomach knowing that so much of our reality is based on our perception of Google, the capabilities of the search bar, and thinking, perhaps in error, genuine wishes or the reflection of it can be seen just by observing whether or not it's searched for on Google.

Sources:

Raw Data - Interest Level over Time:

"Google Trends Data - Search Term: Data Scientist, January 1, 2004 - December 1, 2019; Interest over Time (https://trends.google.com/trends/explore?date=2004-01-01%202019-12-01&geo=US&q=data%20scientist (https://trends.google.com/trends/explore?date=2004-01-01%202019-12-01&geo=US&q=data%20scientist))

"Google Trends Data - Search Term: Data Science, January 1, 2004 - December 1, 2019; Interest over Time (https://trends.google.com/trends/explore?date=2004-01-01%202019-12-01&geo=US&q=data%20science (https://trends.google.com/trends/explore?date=2004-01-01%202019-12-01&geo=US&q=data%20science))

"Google Trends Data - Search Term: Data Science vs. Data Scientist, January 1, 2004 - December 1, 2019; Interest over Time" (https://trends.google.com/trends/explore?date=2004-01-01%202019-12-01&geo=US&q=data%20science,data%20scientist (https://trends.google.com/trends/explore?date=2004-01-01%202019-12-01&geo=US&q=data%20science,data%20scientist))

"[1]Biocognitive https://www.biocognitive.com/index.php/philosophy/page.html (https://www.biocognitive.com/index.php/philosophy/page.html))

"[2]Quantization" https://www.sciencedirect.com/topics/engineering/quantisation (https://www.sciencedirect.com/topics/engineering/quantisation)

GIS data:

"Google Trends Data - Search Term: Data Scientist, January 1, 2004 - December 1, 2019; Interest by SubRegion https://trends.google.com/trends/explore?date=2004-01-01%202019-12-01&geo=US&q=data%20scientist (https://trends.google.com/trends/explore?date=2004-01-01%202019-12-01&geo=US&q=data%20scientist)

"Google Trends Data - Search Term: Data Science, January 1, 2004 - December 1, 2019; Interest by SubRegion https://trends.google.com/trends/explore?date=2004-01-01%202019-12-01&geo=US&q=data%20science (https://trends.google.com/trends/explore?date=2004-01-01%202019-12-01&geo=US&q=data%20science)