Exploratory Data Analysis, Painlessly

X.Y. Han

Exploratory data analysis

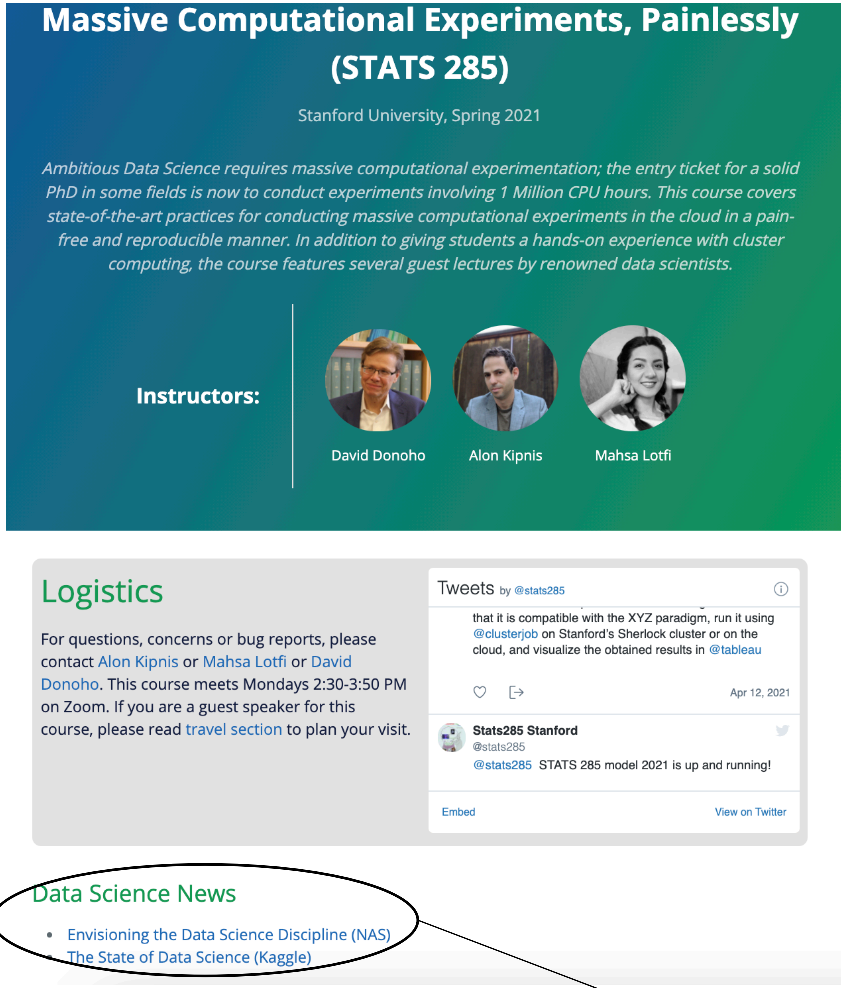
From Wikipedia, the free encyclopedia

In statistics, **exploratory data analysis** (**EDA**) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments.

A Data Science Story

You

Data scientist


Massive Computational Experiments, Painlessly (STATS 285)

Stanford University, Spring 2021

Ambitious Data Science requires massive computational experimentation; the entry ticket for a solid PhD in some fields is now to conduct experiments involving 1 Million CPU hours. This course covers state-of-the-art practices for conducting massive computational experiments in the cloud in a pain-free and reproducible manner. In addition to giving students a hands-on experience with cluster computing, the course features several guest lectures by renowned data scientists.

Instructors: David Donoho, Alon Kipnis, Mahsa Lotfi

Logistics

For questions, concerns or bug reports, please contact Alon Kipnis or Mahsa Lotfi or David Donoho. This course meets Mondays 2:30-3:50 PM on Zoom. If you are a guest speaker for this course, please read travel section to plan your visit.

Tweets by @stats285

that it is compatible with the XYZ paradigm, run it using @clusterjob on Stanford's Sherlock cluster or on the cloud, and visualize the obtained results in @tableau

Apr 12, 2021

Stats285 Stanford
@stats285
@stats285 STATS 285 model 2021 is up and running!

Embed View on Twitter

Data Science News

- Envisioning the Data Science Discipline (NAS)
- The State of Data Science (Kaggle)

Data Science News

- Envisioning the Data Science Discipline (NAS)
- The State of Data Science (Kaggle)

A Data Science Story

A Data Science Story

kaggle

State of Machine Learning and Data Science 2020

Enterprise Executive Summary Report

Overview

For the fourth year, Kaggle surveyed its community of data enthusiasts to share trends within a quickly growing field. **Based on responses from 20,036 Kaggle members, we've created this report focused on the 13% (2,675 respondents) who are currently employed as data scientists.**

We can see a clear picture of what is common in the community but also the diverse attributes of its members.

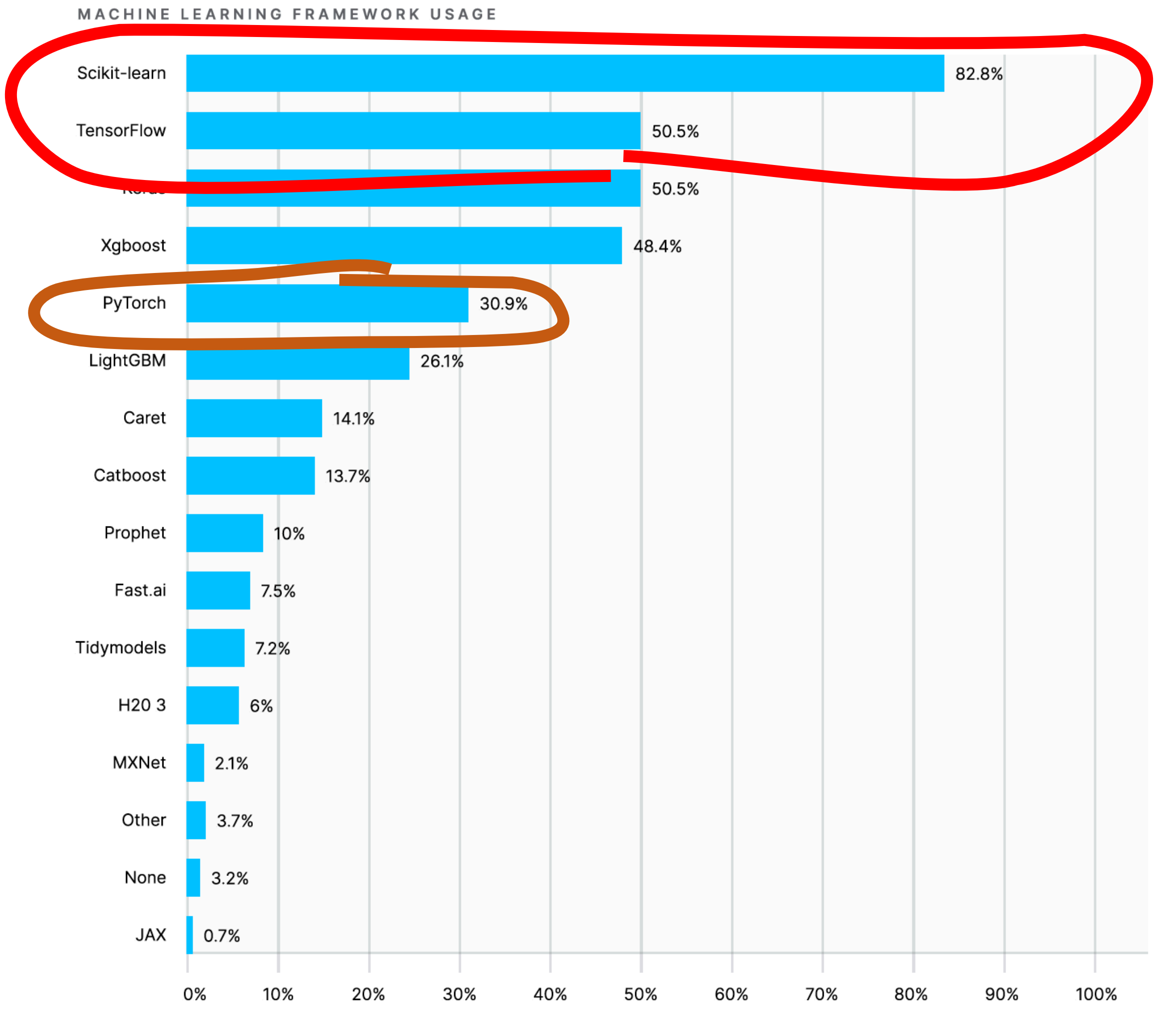
https://www.kaggle.com/kaggle-survey-2020

A Data Science Story

Q16

Which of the following machine learning frameworks do you use on a regular basis?

- [Scikit-learn]
- [TensorFlow]
- [Keras]
- [PyTorch]
- [Fast.ai]
- [MXNet]
- [Xgboost]
- [LightGBM]
- [CatBoost]
- [Prophet]
- [H2O 3]
- [Caret]
- [Tidymodels]
- [JAX]
- None
- Other



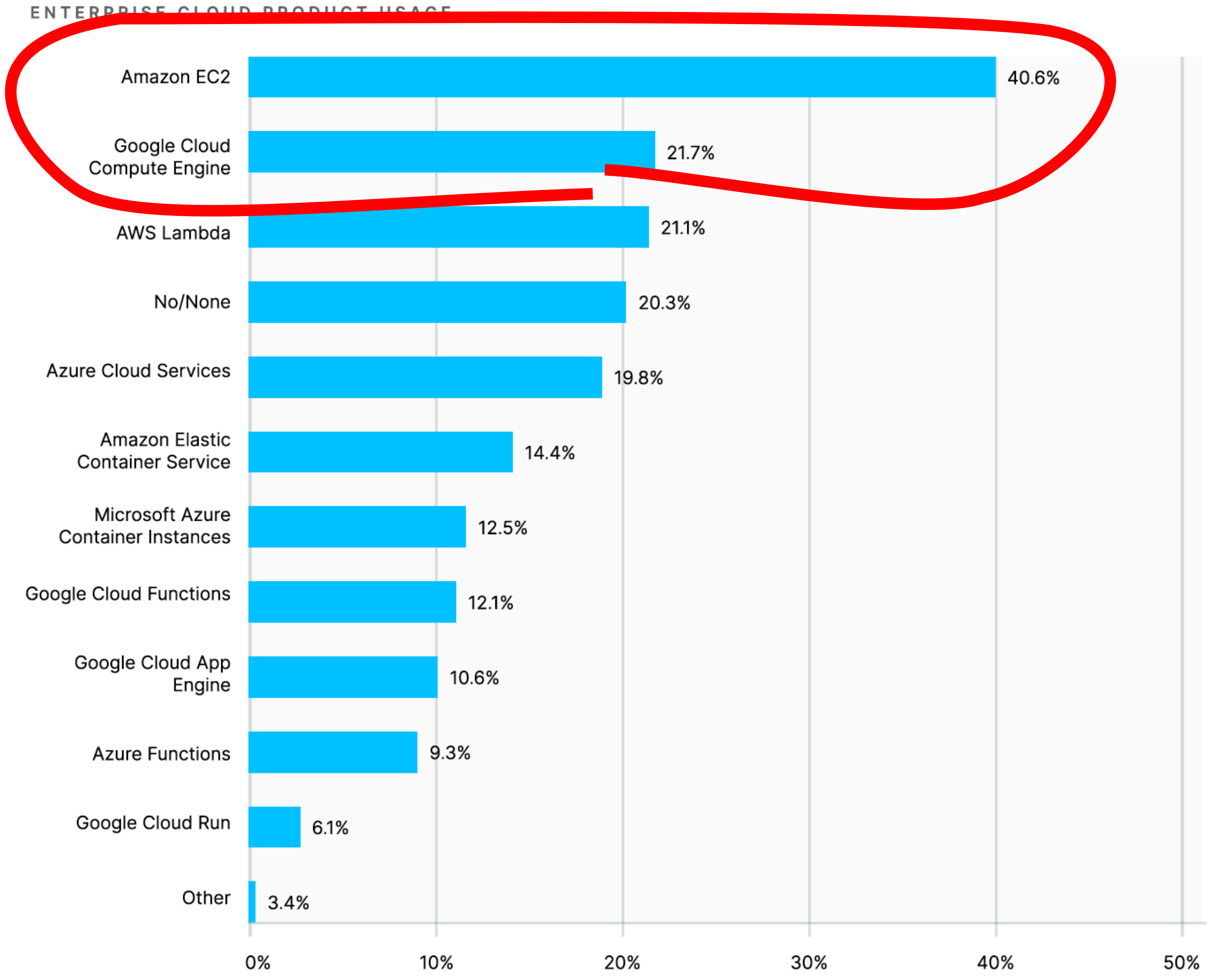
https://www.kaggle.com/c/kaggle-survey-2020/data

A Data Science Story

Q27-A

Do you use any of the following cloud computing products on a regular basis?

- Amazon EC2
- AWS Lambda
- Amazon Elastic Container Service
- Azure Cloud Services
- Microsoft Azure Container Instances
- Azure Functions
- Google Cloud Compute Engine
- Google Cloud Functions
- Google Cloud Run
- Google Cloud App Engine
- No / None
- Other



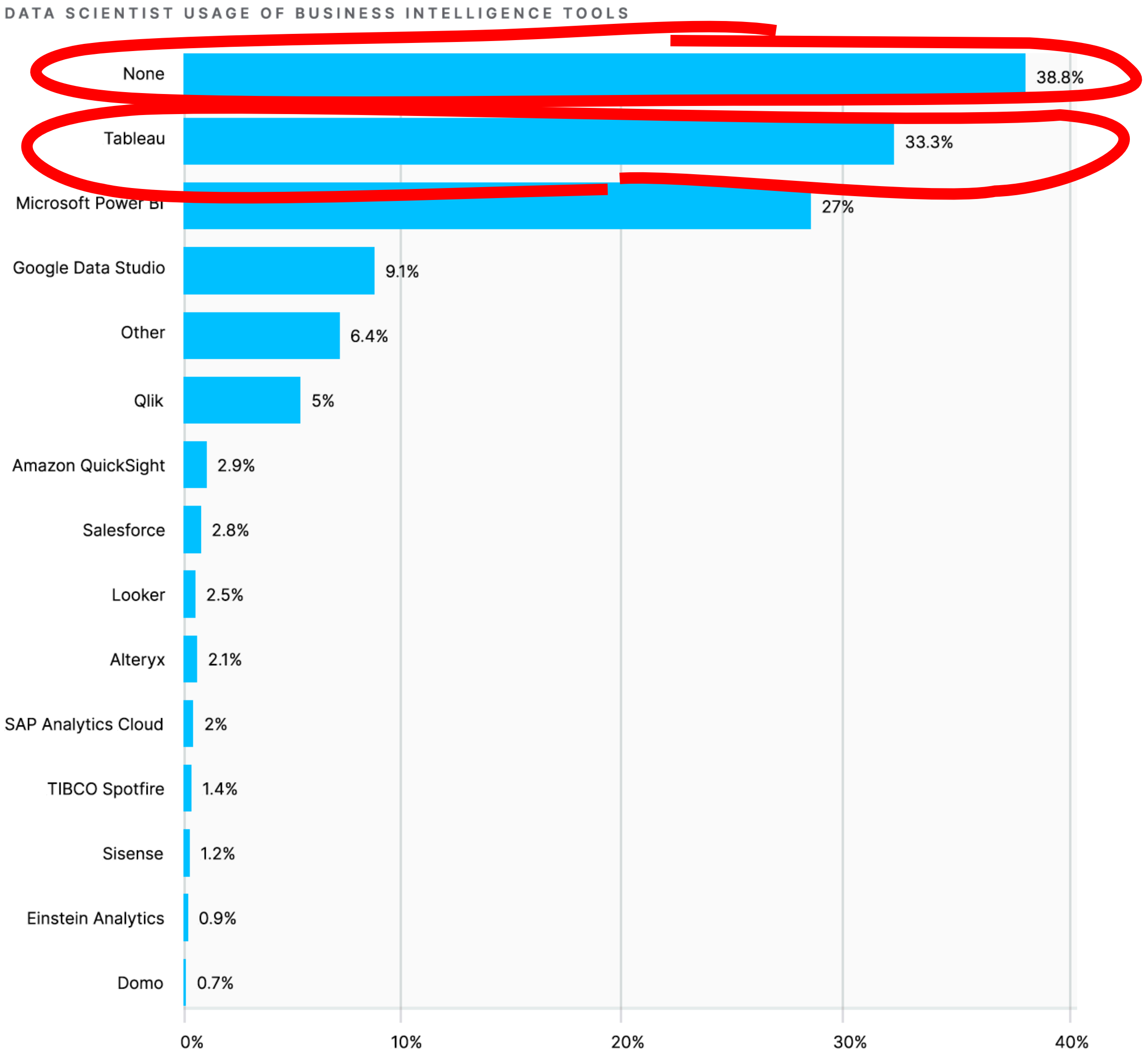
https://www.kaggle.com/c/kaggle-survey-2020/data

A Data Science Story

Q32

Which of the following business intelligence tools do you use most often?[6]

- » Amazon QuickSight
- » Microsoft Power BI
- » Google Data Studio
- » Looker
- » Tableau
- » Salesforce
- » Einstein Analytics
- » Qlik
- » Domo
- » TIBCO Spotfire
- » Alteryx
- » Sisense
- » SAP Analytics Cloud
- » None
- » Other

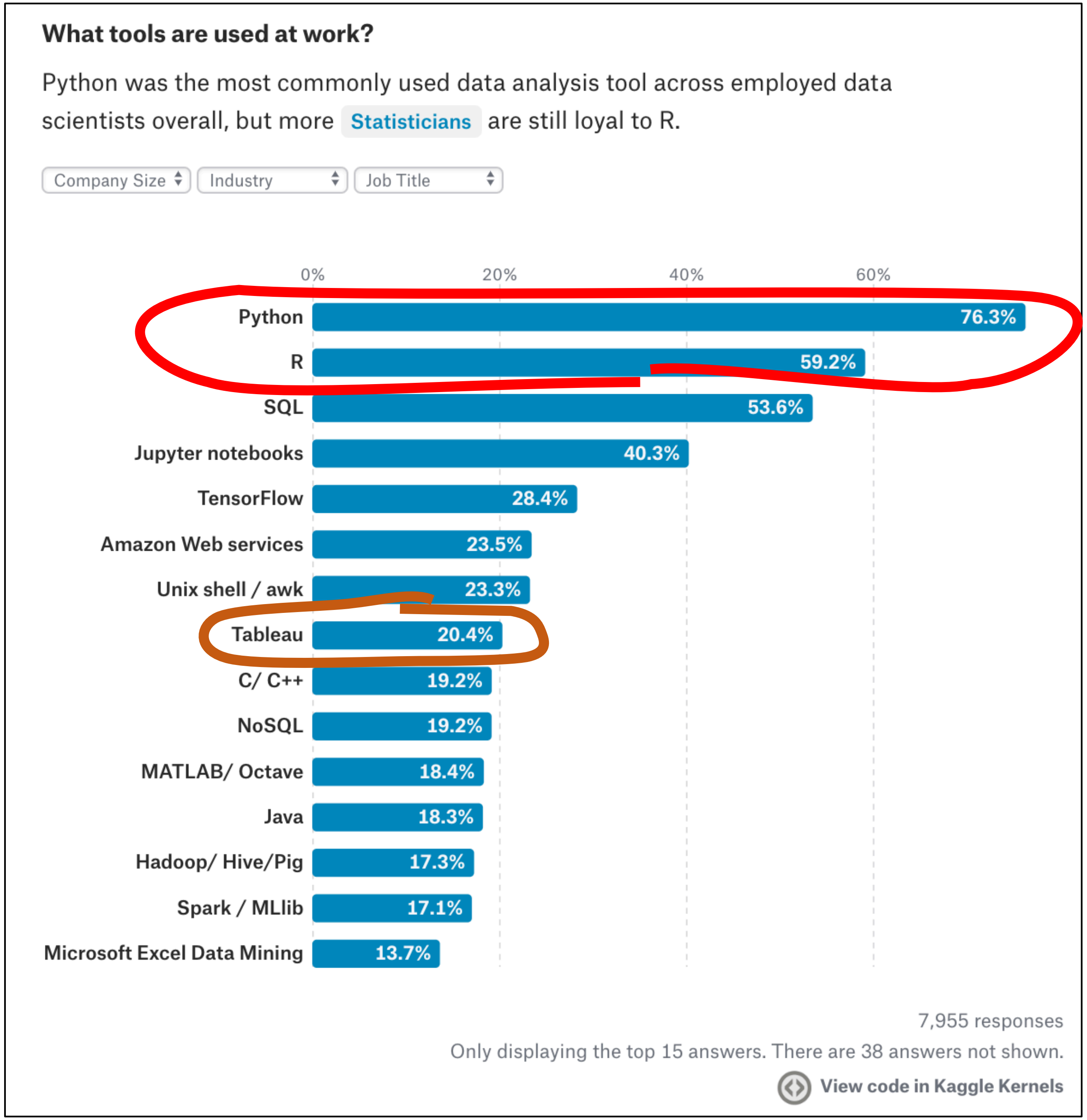


Answers are all very specific....

https://www.kaggle.com/c/kaggle-survey-2020/data

A Data Science Story

A Data Science Story

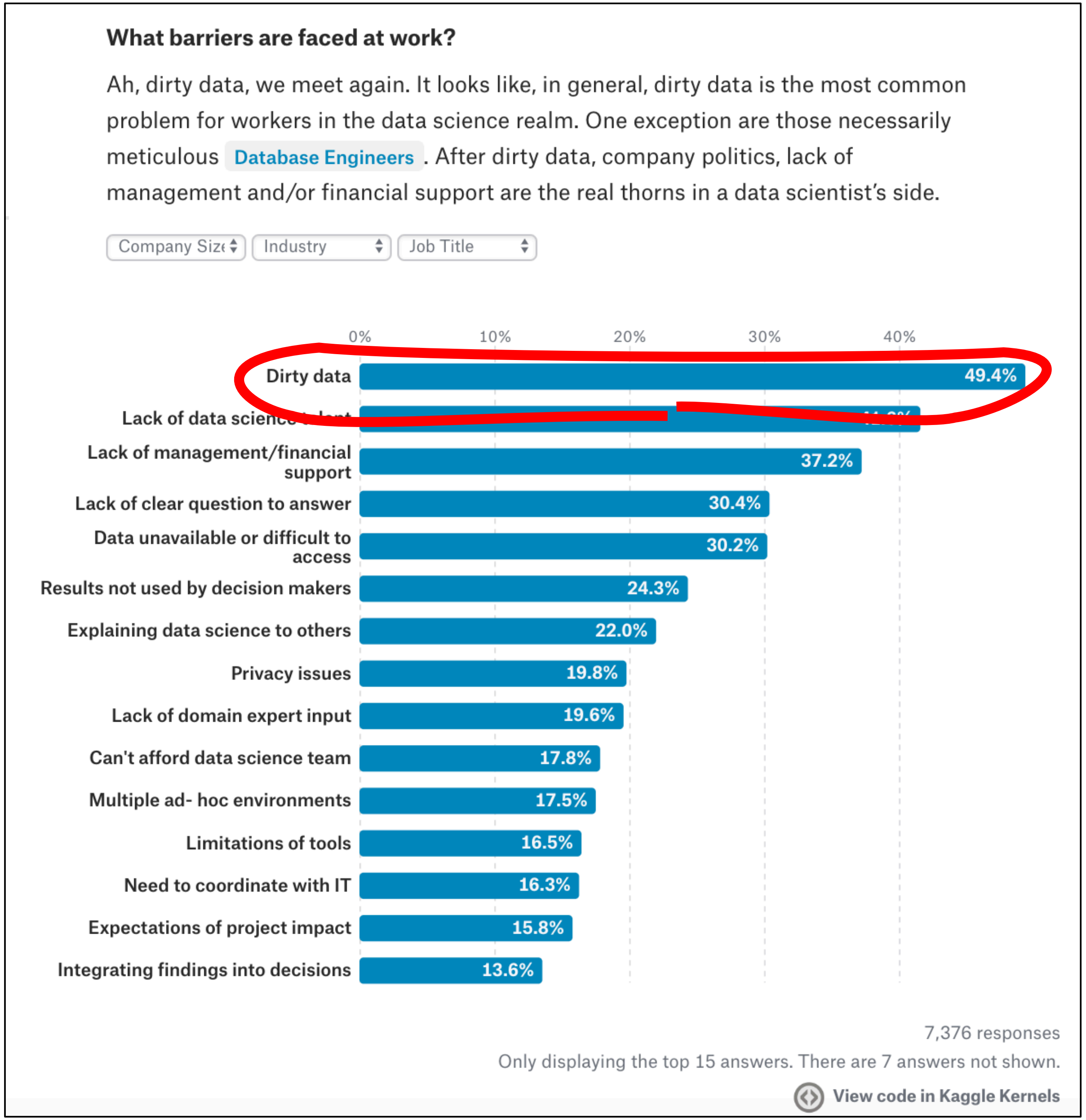


A Data Science Story

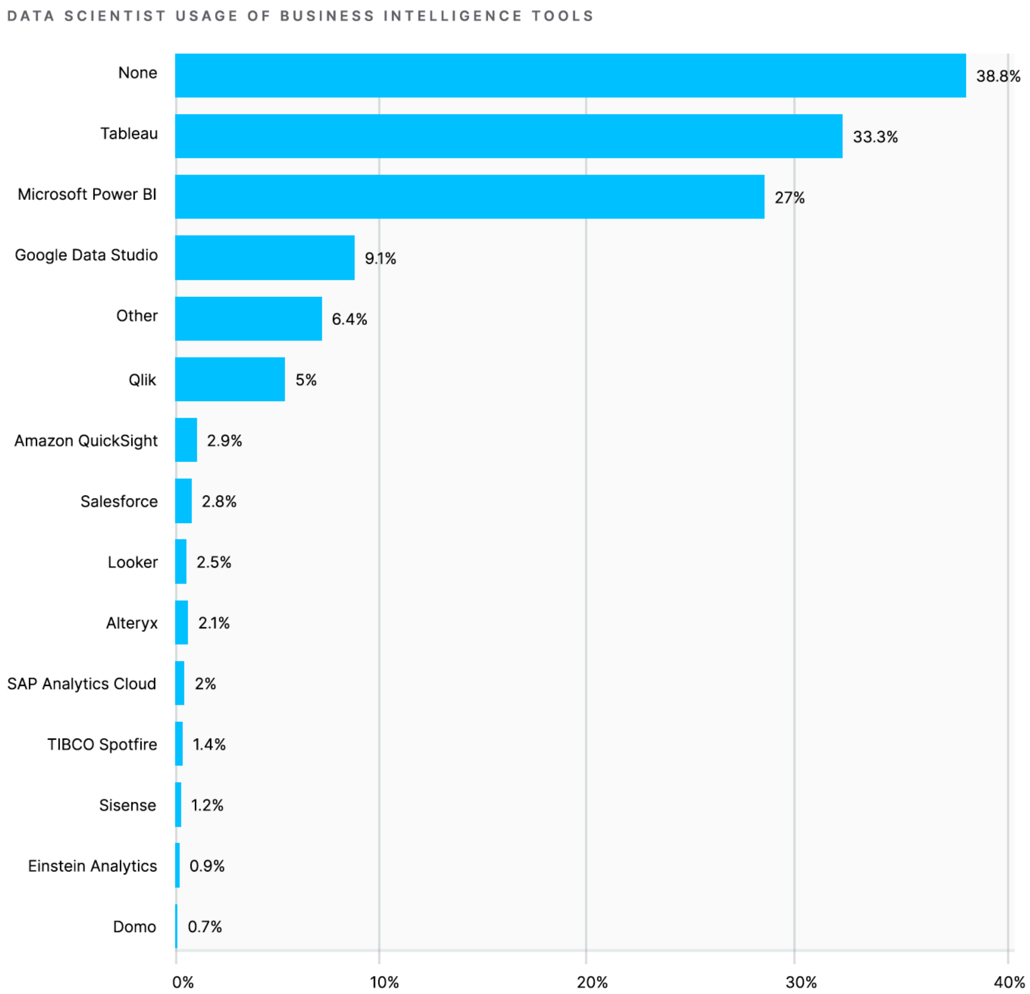
A Data Science Story

Can I generate these myself?

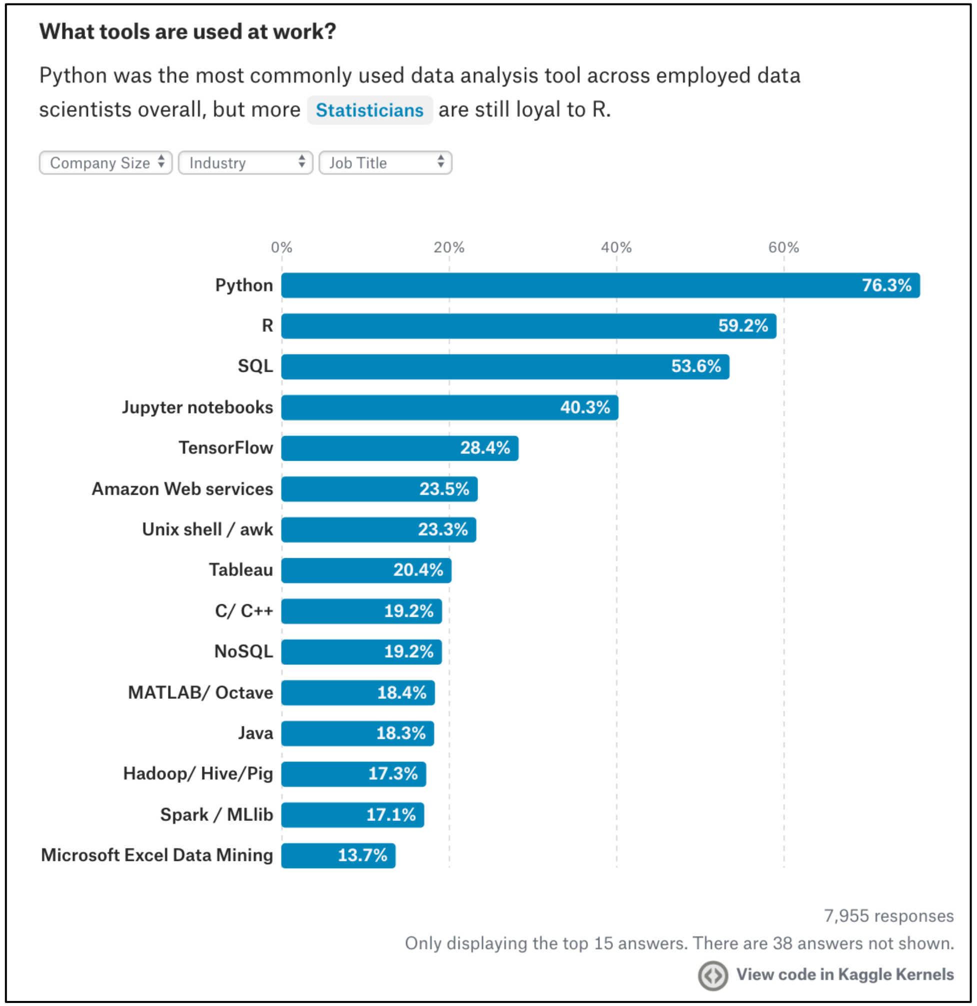
Painlessly?



Let’s See…



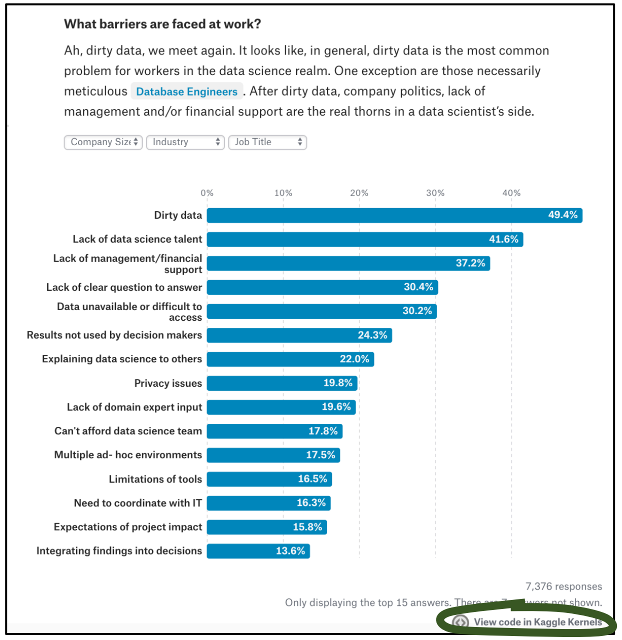
[2020]



[2017]

Obstacles in code-based data analysis

- Reading/cleaning/aggregating data
- Learning/deciphering syntax

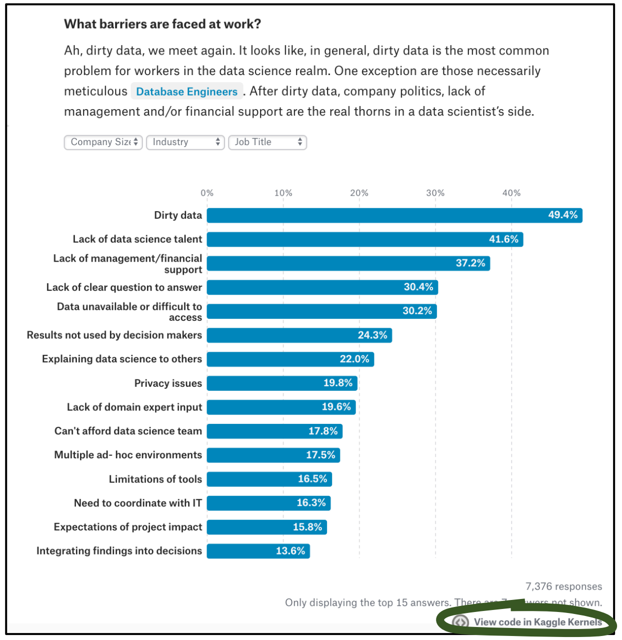



```
chooseMultiple = function(question, filteredData = cleanData){

  filteredData %>%
    # Remove any rows where the respondent didn't answer the question
    filter(!UQ(sym(question)) == "") %>%
    # Remove all columns except question
    select(question) %>%
    # Add a column with the initial number of respondents to question
    mutate(totalCount = n()) %>%
    # Split multiple answers apart at the comma, but ignore commas inside parentheses
    mutate(selections = strsplit(as.character(UQ(sym(question))),
                                 '\\([^)]+,(*SKIP)(*FAIL)|,\\s*', perl = TRUE)) %>%
    # Split answers are now nested, need to unnest them
    unnest(selections) %>%
    # Group by the selected responses to the question
    group_by(selections) %>%
   # Count how many respondents selected each option
    summarise(totalCount = max(totalCount),
              count = n()) %>%
    # Calculate what percent of respondents selected each option
    mutate(percent = (count / totalCount) * 100) %>%
    # Arrange the counts in descending order
    arrange(desc(count))

}
```

```
# Filter the data
filterBarriers <- workLife %>%
  # Remove blank responses on employment question
  filter(!EmploymentStatus == "") %>%
  # Keep only entries that indicated that they use code to analyze data at work
  filter(CodeWriter == "Yes") %>%
  # Keep only entries that included one of the above "employed" statuses
  filter(grepl(paste(employed, collapse = "|"), EmploymentStatus))

# Using the filtered data, run chooseMultiple() function
chooseMultiple("WorkChallengesSelect", filterBarriers)
```

Obstacles in code-based data analysis

- Reading/cleaning/aggregating data
- Learning/deciphering syntax
- Code and deliverable mismatch



selections <chr>
Dirty data
Lack of data science talent in the organization
Company politics / Lack of management/financial support for a data science team
The lack of a clear question to be answering or a clear direction to go in with the available data
Unavailability of/difficult access to data
Data Science results not used by business decision makers
Explaining data science to others
Privacy issues
Lack of significant domain expert input
Organization is small and cannot afford a data science team

1-10 of 22 rows | 1-1 of 4 columns Previous 1 2 3 Next

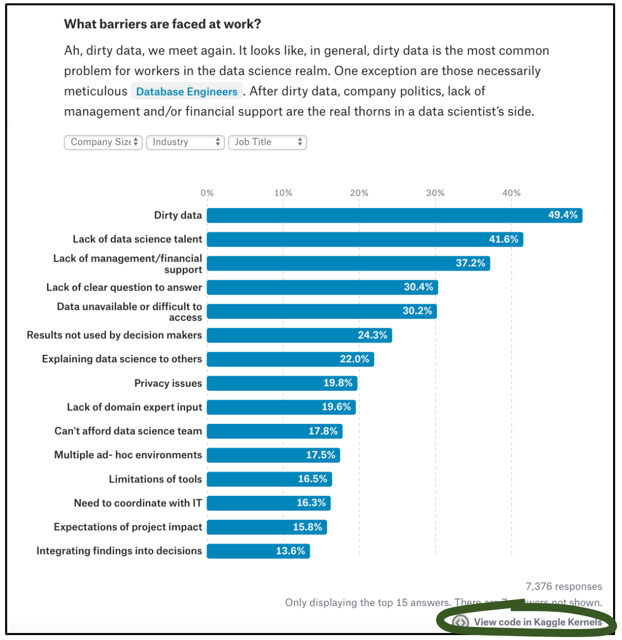
totalCount <dbl>	co... <int>	percent <dbl>
7376	3641	49.362798
7376	3067	41.580803
7376	2746	37.228850
7376	2242	30.395879
7376	2230	30.233189
7376	1796	24.349241
7376	1622	21.990239
7376	1460	19.793926
7376	1444	19.577007
7376	1316	17.841649

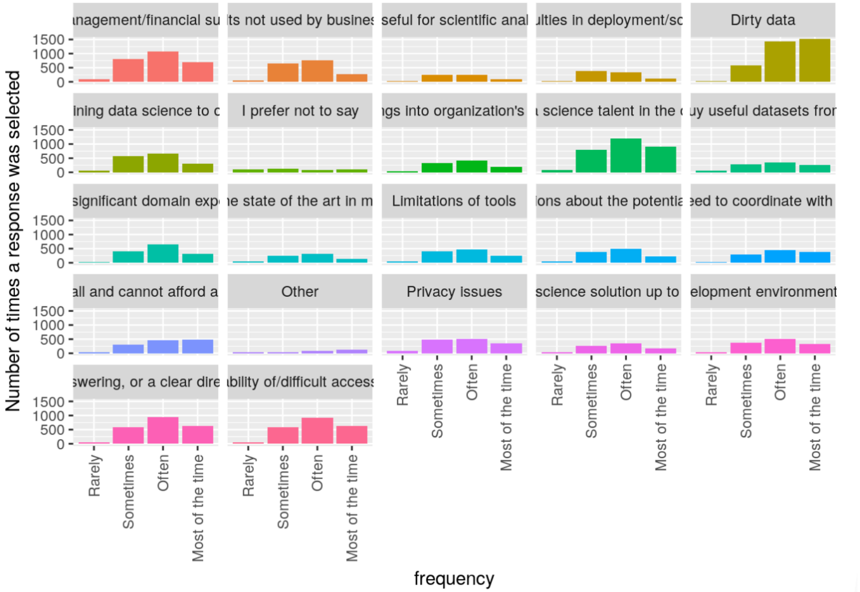
1-10 of 22 rows | 2-4 of 4 columns Previous 1 2 3 Next

Obstacles in code-based data analysis

- Reading/cleaning/aggregating data
- Learning/deciphering syntax
- Code and deliverable mismatch
- Formatting output

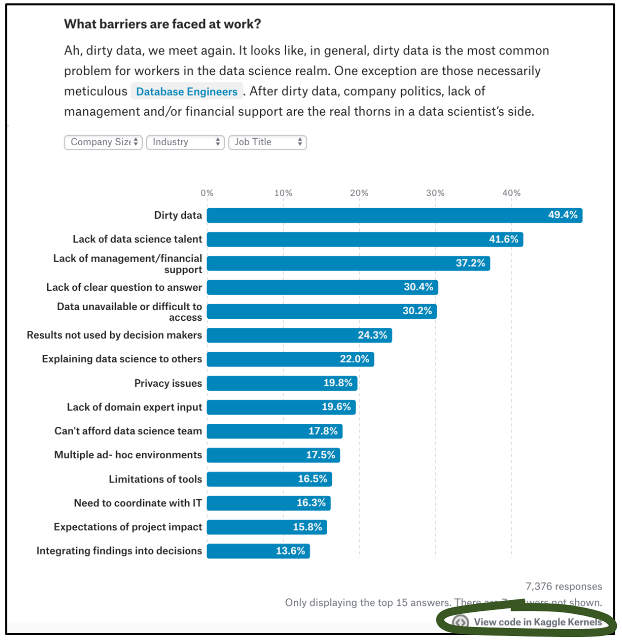
```
# Plot
ggplot(challengeNamesChar, aes(x = frequency, y = count, fill = response.y)) +
  geom_bar(stat = "identity") +
  facet_wrap(~response.y) +
  ylab("Number of times a response was selected") +
  theme(legend.position="none") +
  theme(axis.text.x = element_text(angle = 90,
                                   vjust = 0.5,
                                   hjust = 1))

# Since the names are often too long to be displayed well in this figure, print
levels(as.factor(challengeNamesChar$response.y))
```





Obstacles in code-based data analysis

- Reading/cleaning/aggregating data
- Learning/deciphering syntax
- Code and deliverable mismatch
- Formatting output



Misalignment of goals

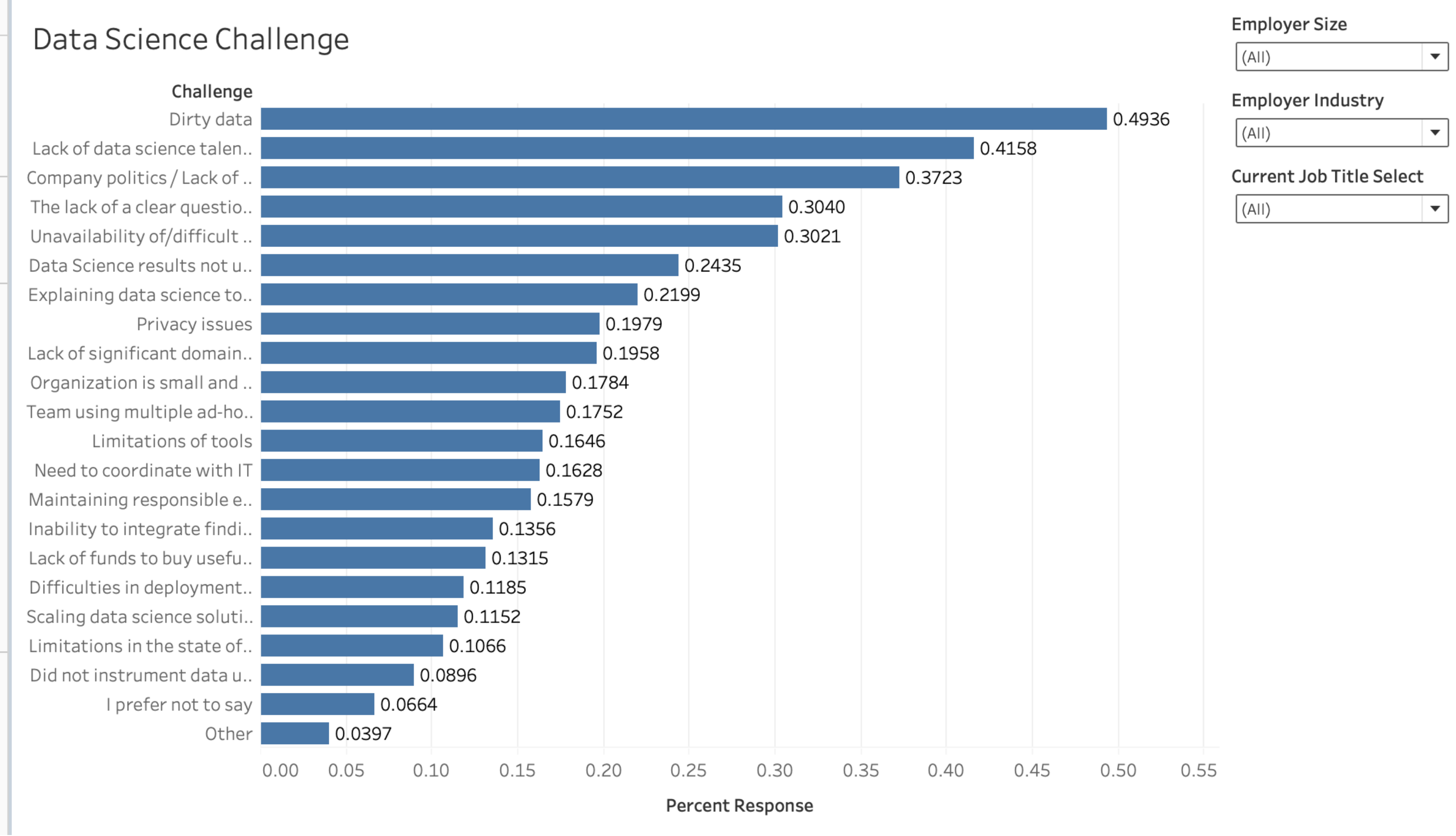
Data analysis for dissemination

- **Flexible** framework for implementing complex calculations
- **Concise** representations like scripts and functions
- Facilitates the efficient **communication** of insights.

Data analysis for exploration

- **Fast** manipulation of data
- **Intuitive** interface
- Facilitates the efficient **discovery** of insights.

Demo 1: Data Science Challenges



https://stats285.github.io/

RESEARCH ARTICLE

Prevalence of neural collapse during the terminal phase of deep learning training

Vardan Papyan, X. Y. Han, and David L. Donoho

+ See all authors and affiliations

PNAS October 6, 2020 117 (40) 24652-24663; first published September 21, 2020;
https://doi.org/10.1073/pnas.2015509117

Demo: Visualizing Research

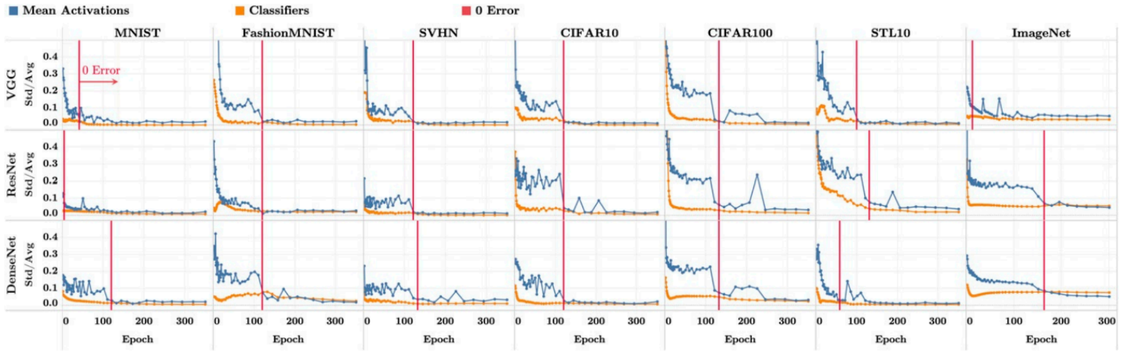


Fig. 2. **Train class means become equinorm.** The formatting and technical details are as described in Section 3. In each array cell, the vertical axis shows the coefficient of variation of the centered class-mean norms as well as the network classifiers norms. In particular, the blue lines show $\mathrm{Std}_c(\|\mu_c - \mu_G\|_2)/\mathrm{Avg}_c(\|\mu_c - \mu_G\|_2)$ where $\{\mu_c\}$ are the class means of the last-layer activations of the training data and μ_G is the corresponding train global mean; the orange lines show $\mathrm{Std}_c(\|\boldsymbol{w}_c\|_2)/\mathrm{Avg}_c(\|\boldsymbol{w}_c\|_2)$ where $\boldsymbol{w}_c$ is the last-layer classifier of the cth class. As training progresses, the coefficients of variation of both class means and classifiers decrease.

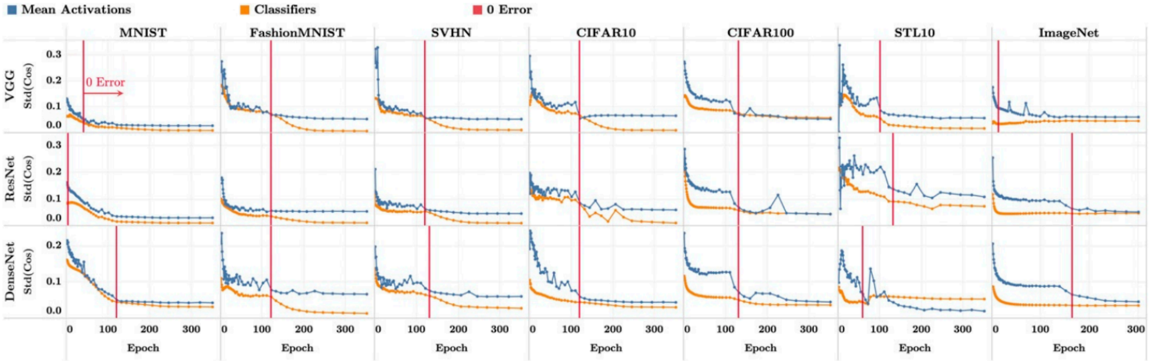


Fig. 3. **Classifiers and train class means approach equiangularity.** The formatting and technical details are as described in Section 3. In each array cell, the vertical axis shows the SD of the cosines between pairs of centered class means and classifiers across all distinct pairs of classes c and c'. Mathematically, denote $\cos_\mu(c, c') = \langle \mu_c - \mu_G, \mu_{c'} - \mu_G \rangle / (\|\mu_c - \mu_G\|_2 \|\mu_{c'} - \mu_G\|_2$ and $\cos_w(c, c') = \langle \boldsymbol{w}_c, \boldsymbol{w}_{c'} \rangle / (\|\boldsymbol{w}_c\|_2 \|\boldsymbol{w}_{c'}\|_2)$ where $\{\boldsymbol{w}_c\}_{c=1}^C$, $\{\mu_c\}_{c=1}^C$, and μ_G are as in Fig. 2. We measure $\mathrm{Std}_{c,c' \neq c}(\cos_\mu(c, c'))$ (blue) and $\mathrm{Std}_{c,c' \neq c}(\cos_w(c, c'))$ (orange). As training progresses, the SDs of the cosines approach zero, indicating equiangularity.

https://purl.stanford.edu/ng812mz4543