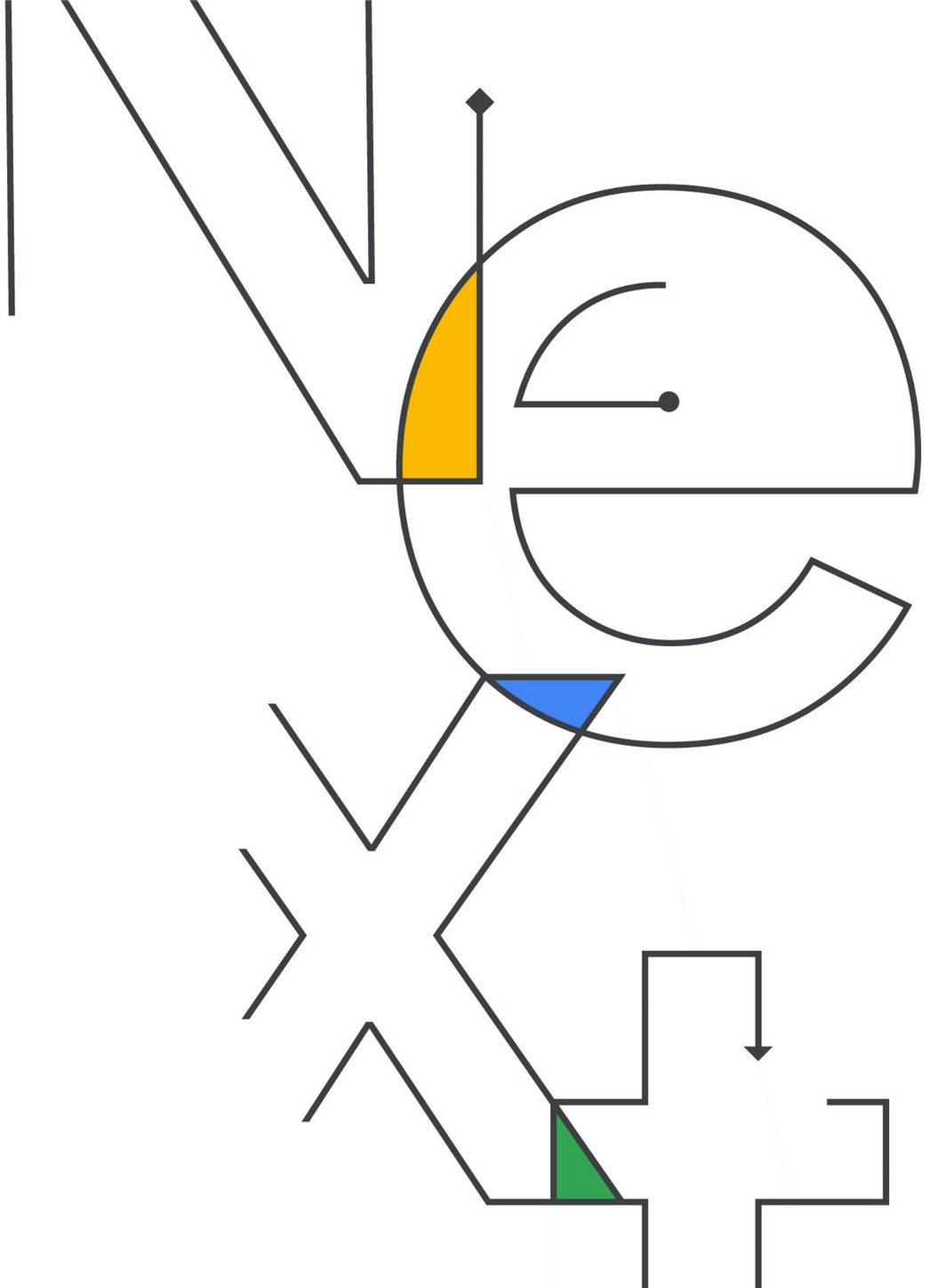
Google Cloud

Next '22

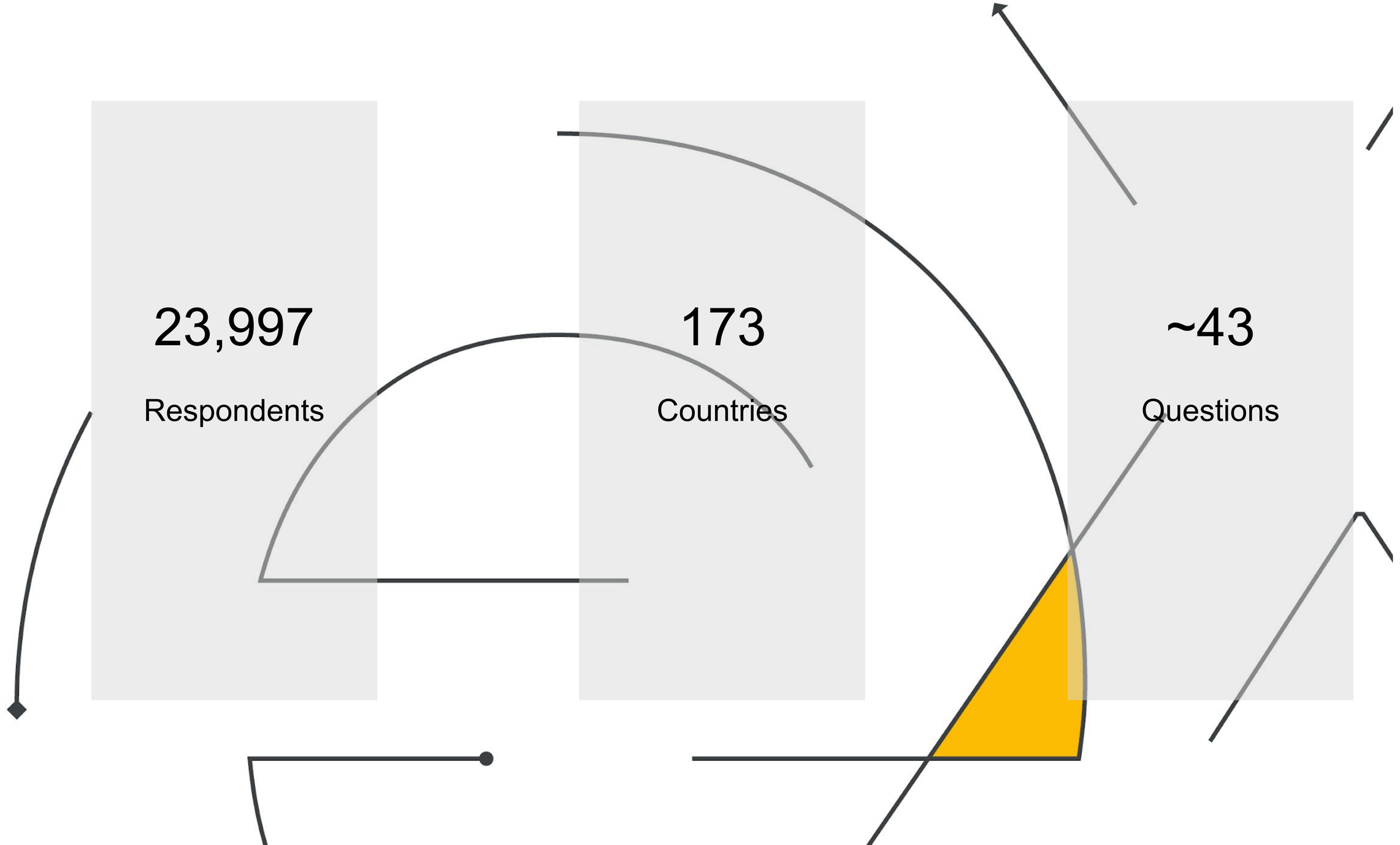
2022 Kaggle Data Science & ML Survey

Data Scientists' backgrounds, preferred technologies, and techniques

Oct/
11–13



In September 2022, Kaggle conducted its sixth annual industry-wide survey in an attempt to surface a truly comprehensive view of the state of data science and machine learning.


23,997
Respondents
173
Countries
~43
Questions

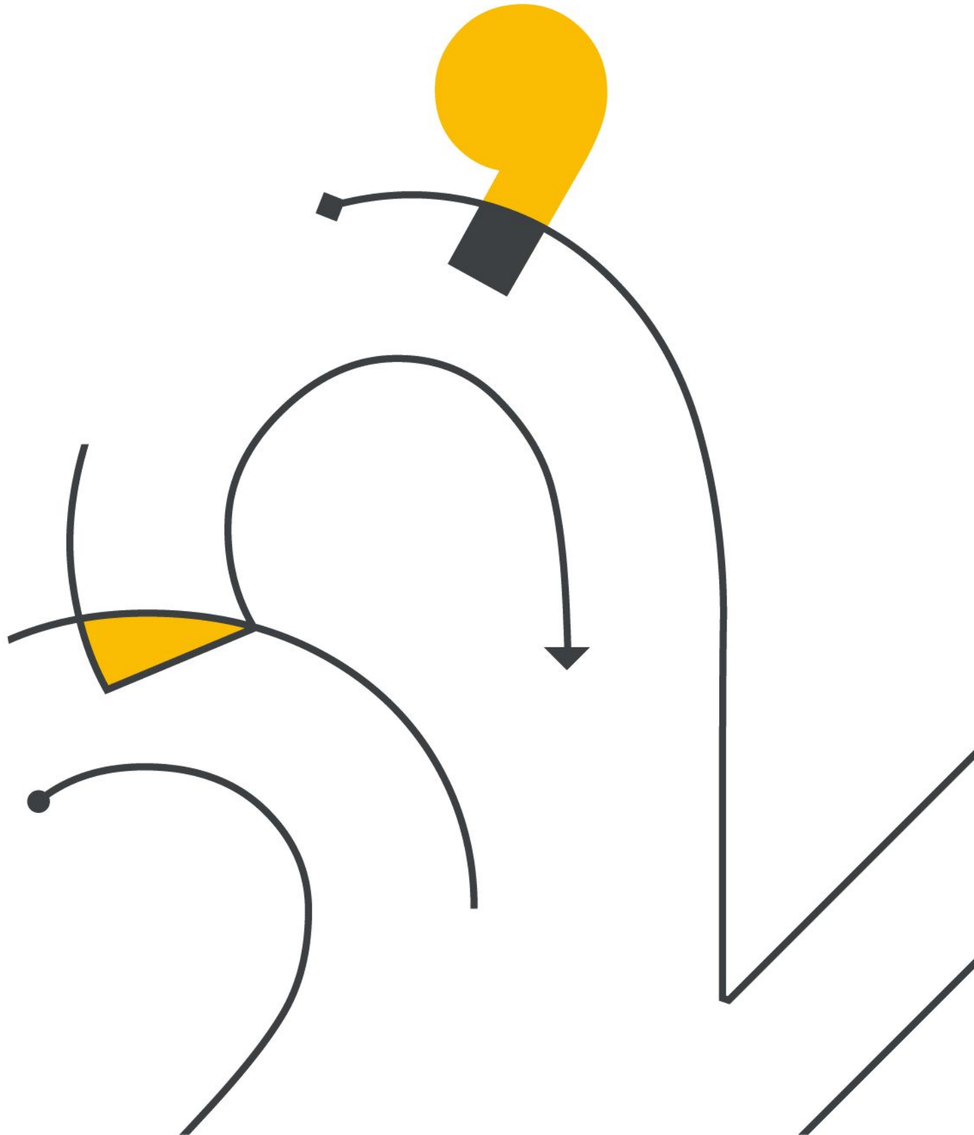
Meet Kaggle

Kaggle is the world’s largest data science community with powerful tools and resources to help you achieve your data science learning goals.

- > 10 Million Data Scientists
- 300+ Machine Learning Competitions
- 170k+ Public Datasets
- 750k+ Public Notebooks

Download the full survey results at:

kaggle.com/kaggle-survey-2022



Today's Presentation:

Working Data Scientists

01 Demographics

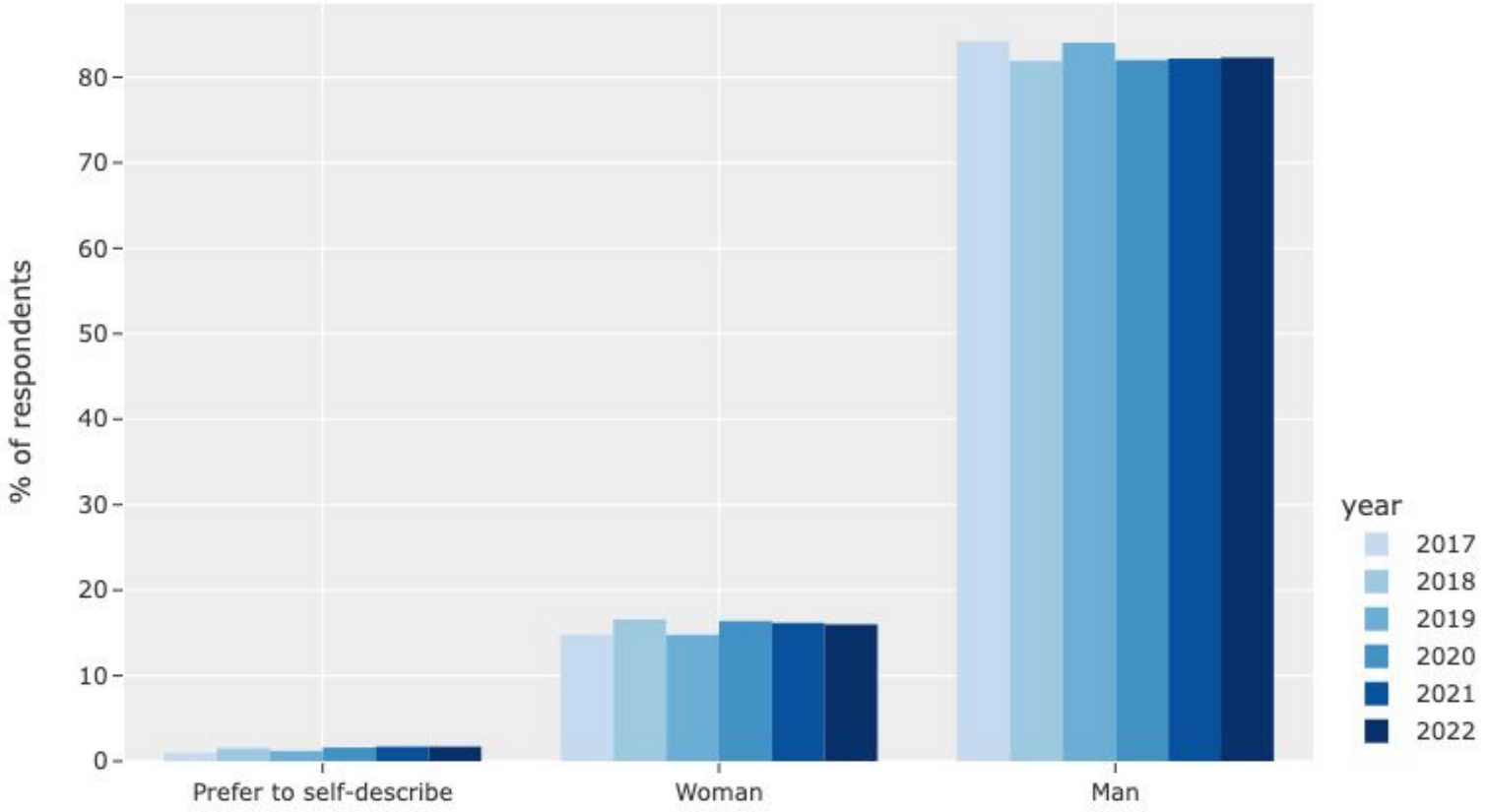
02 Programming

03 Machine Learning

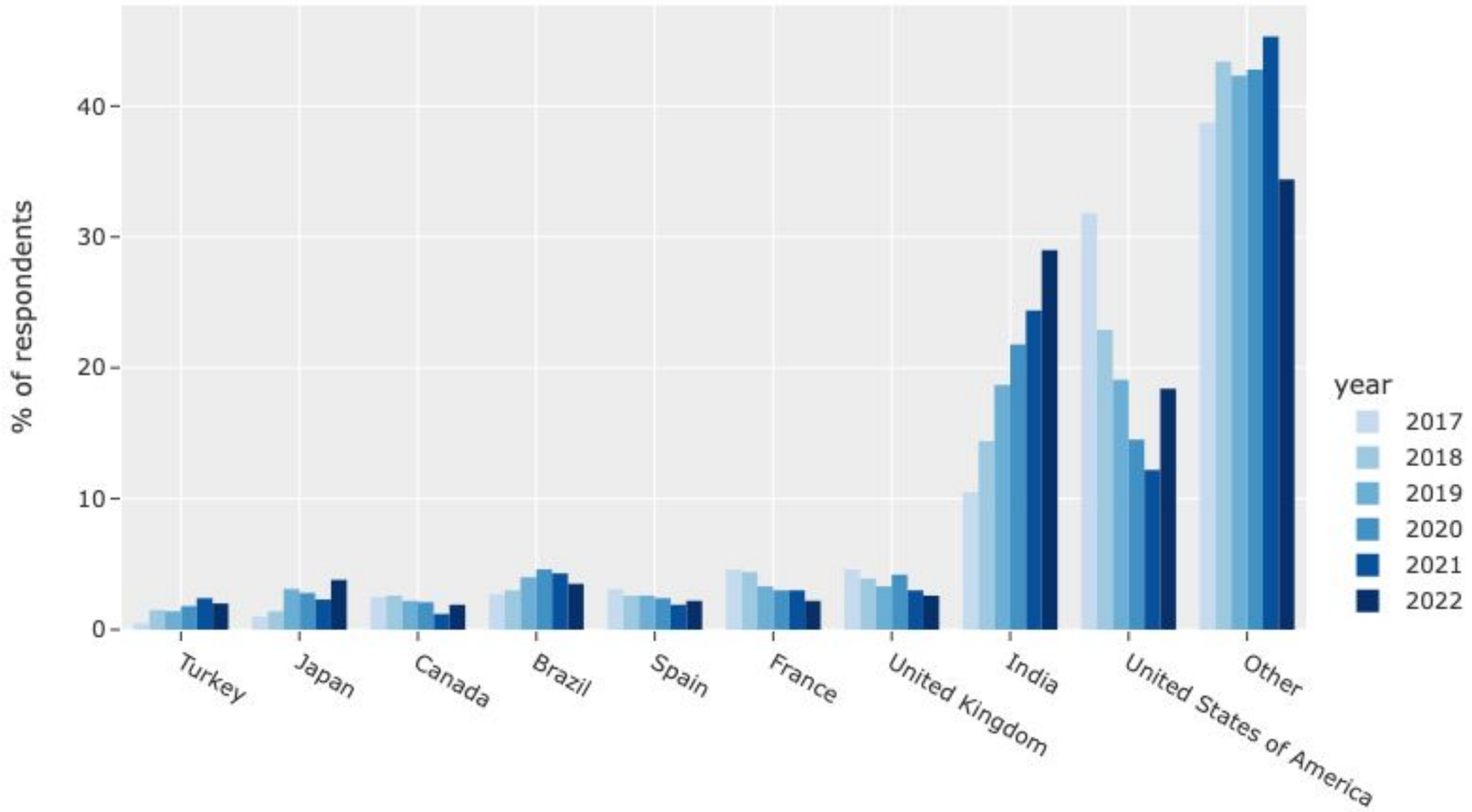
04 Cloud Computing

Demographics

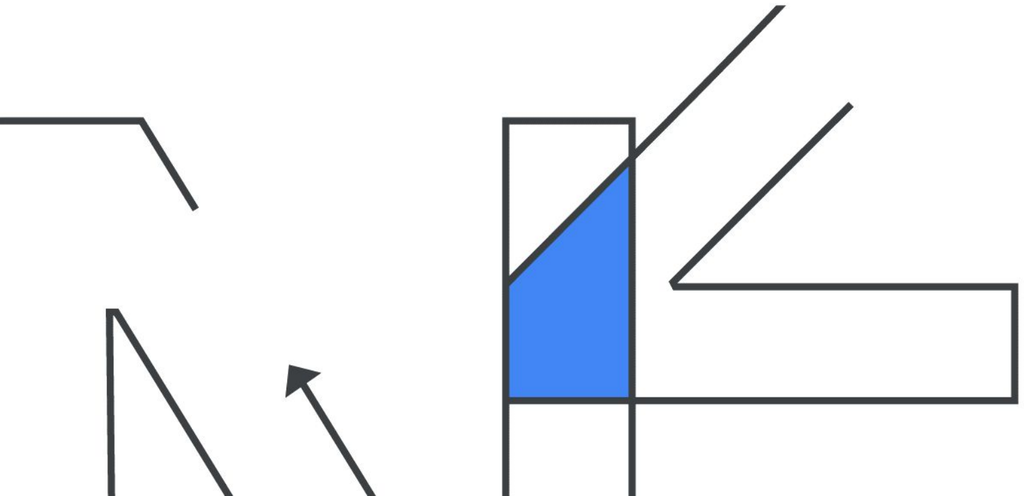
The data science industry remains highly gender imbalanced



An increasing number of data scientists are living and working in India and Japan



Panel



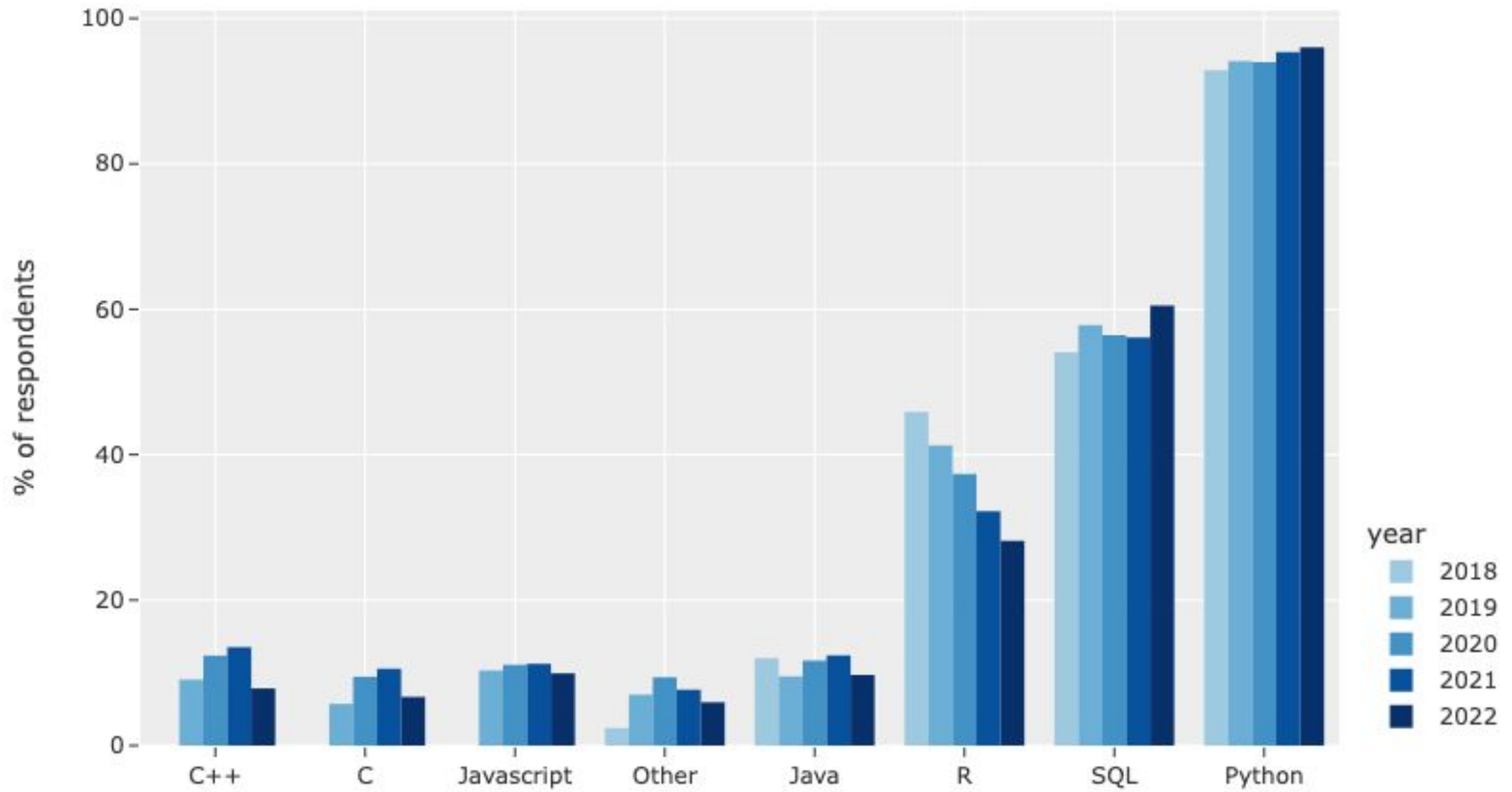
Questions

1. Do you have any insights on the growth of data science as a career in specific geographic regions?
2. Any unique dynamics or initiatives that accelerate growth?

Programming

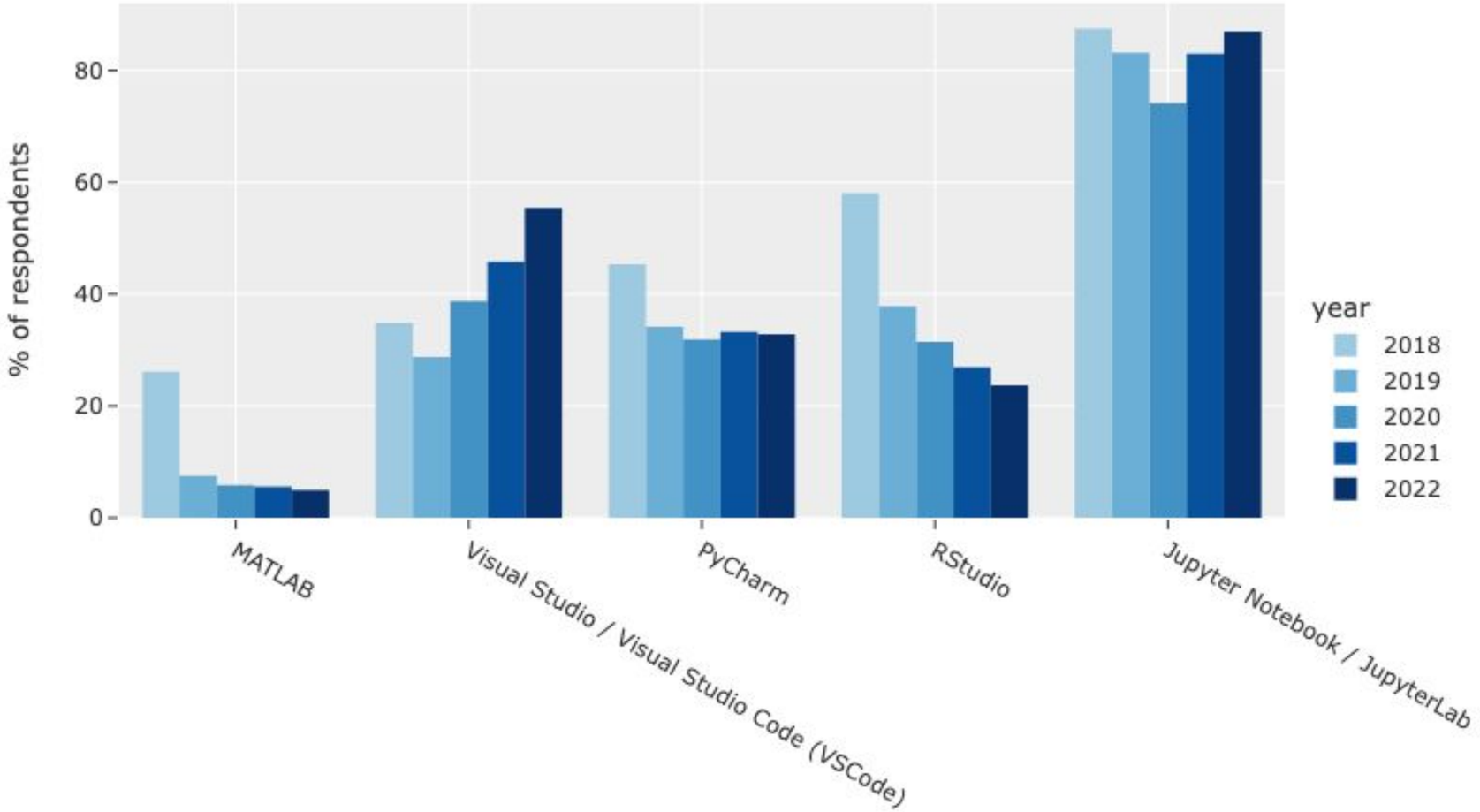
Kaggle DS & ML Survey 2022

Python and SQL remain the two most common programming skills for data scientists

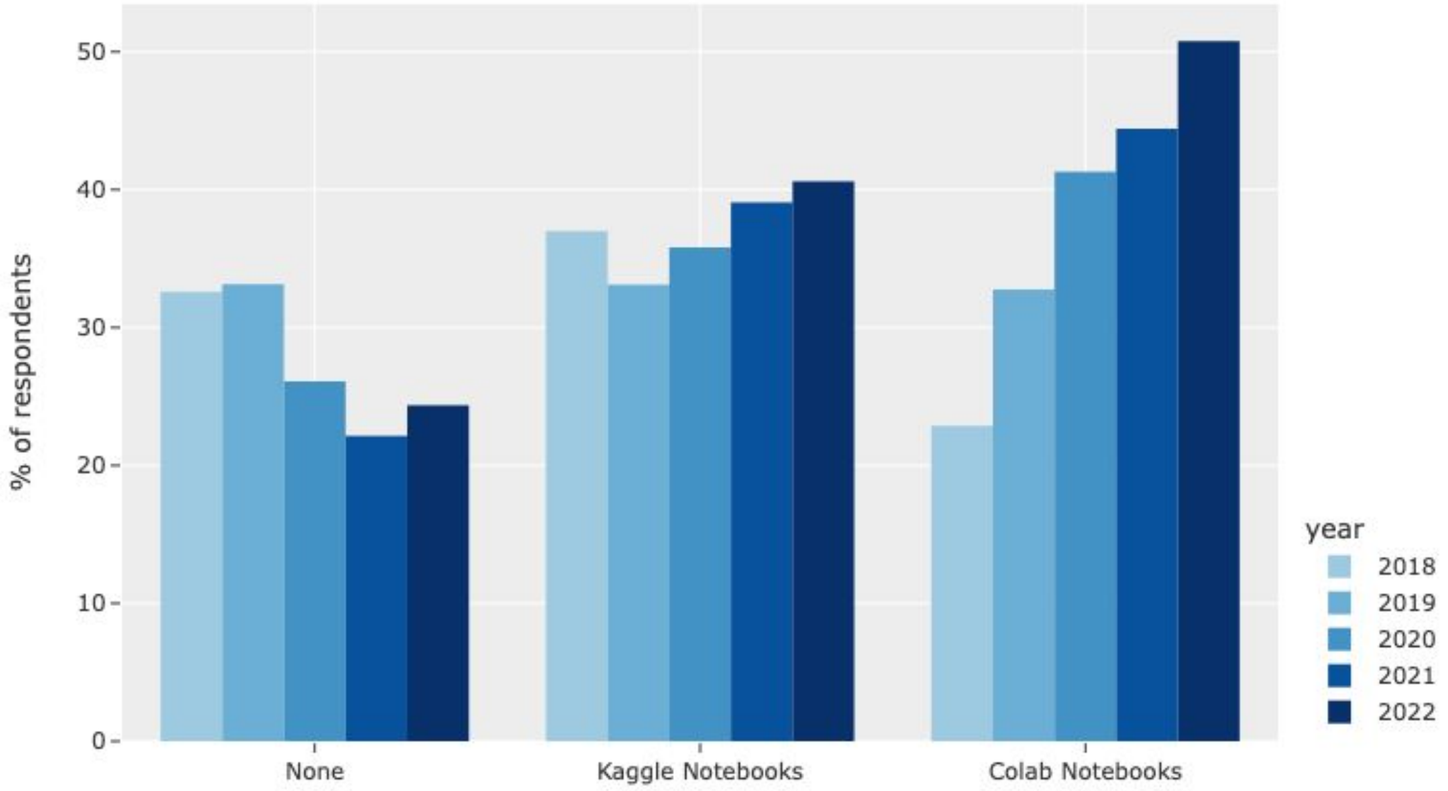


Kaggle DS & ML Survey 2022

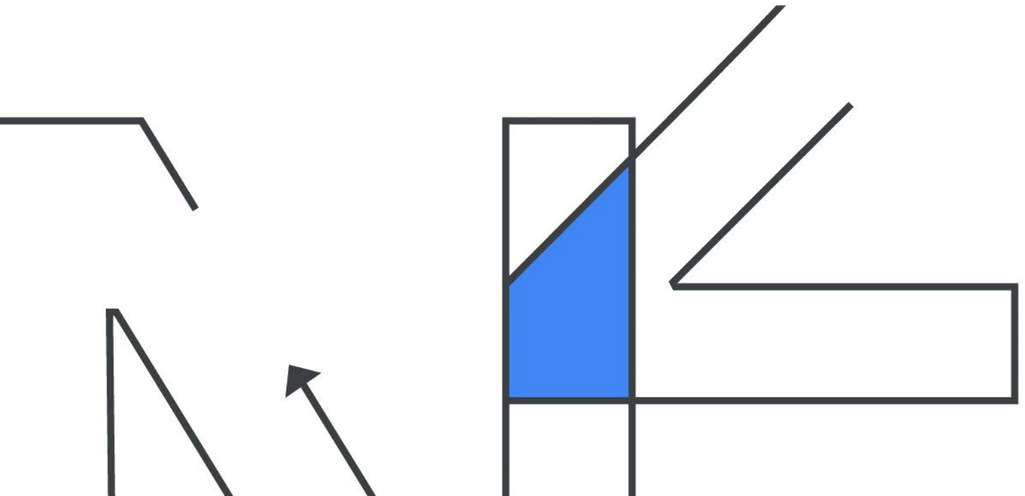
VSCode is now used by over 50% of working data scientists



Colab notebooks are the most popular cloud-based Jupyter notebook environment



Panel

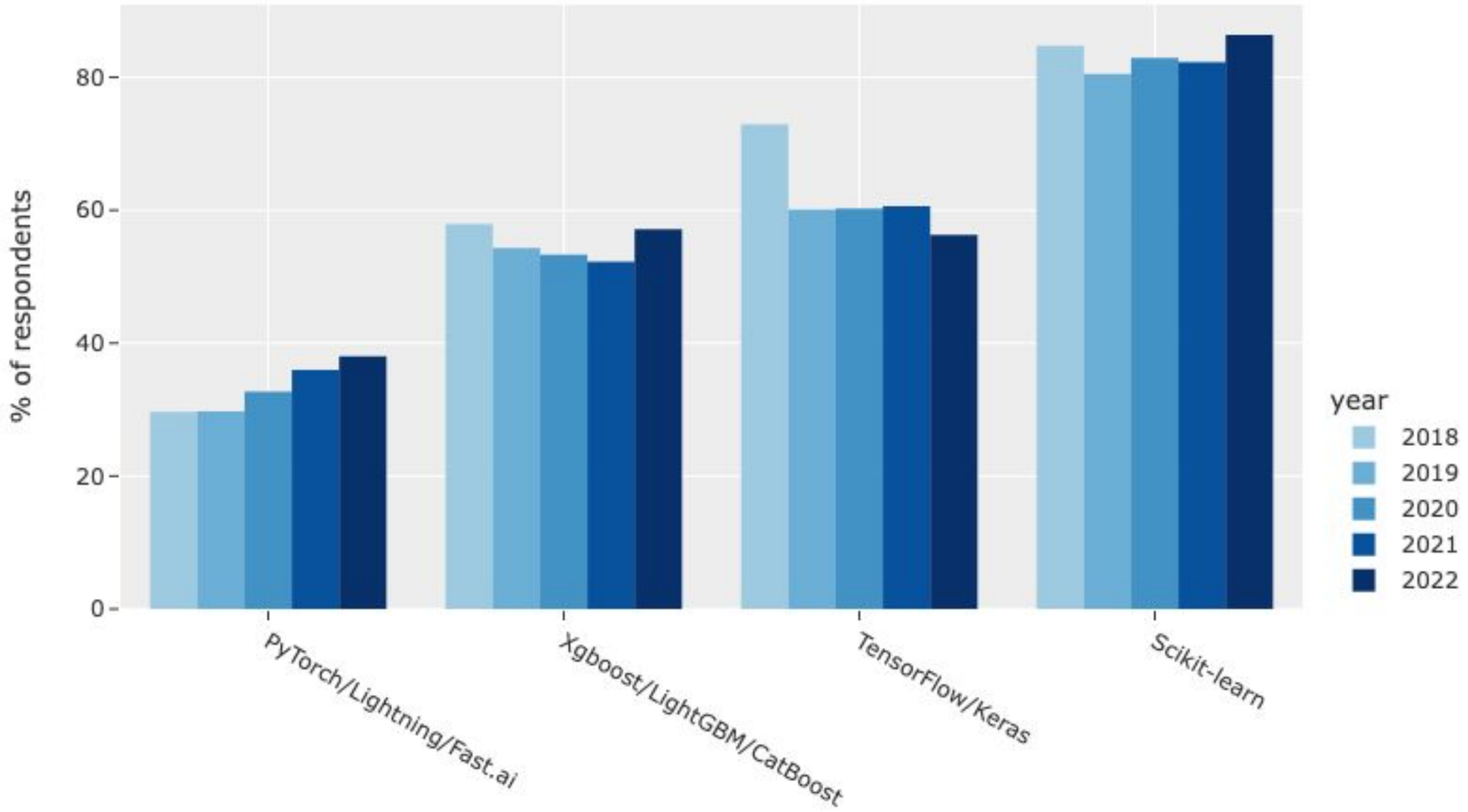


Questions

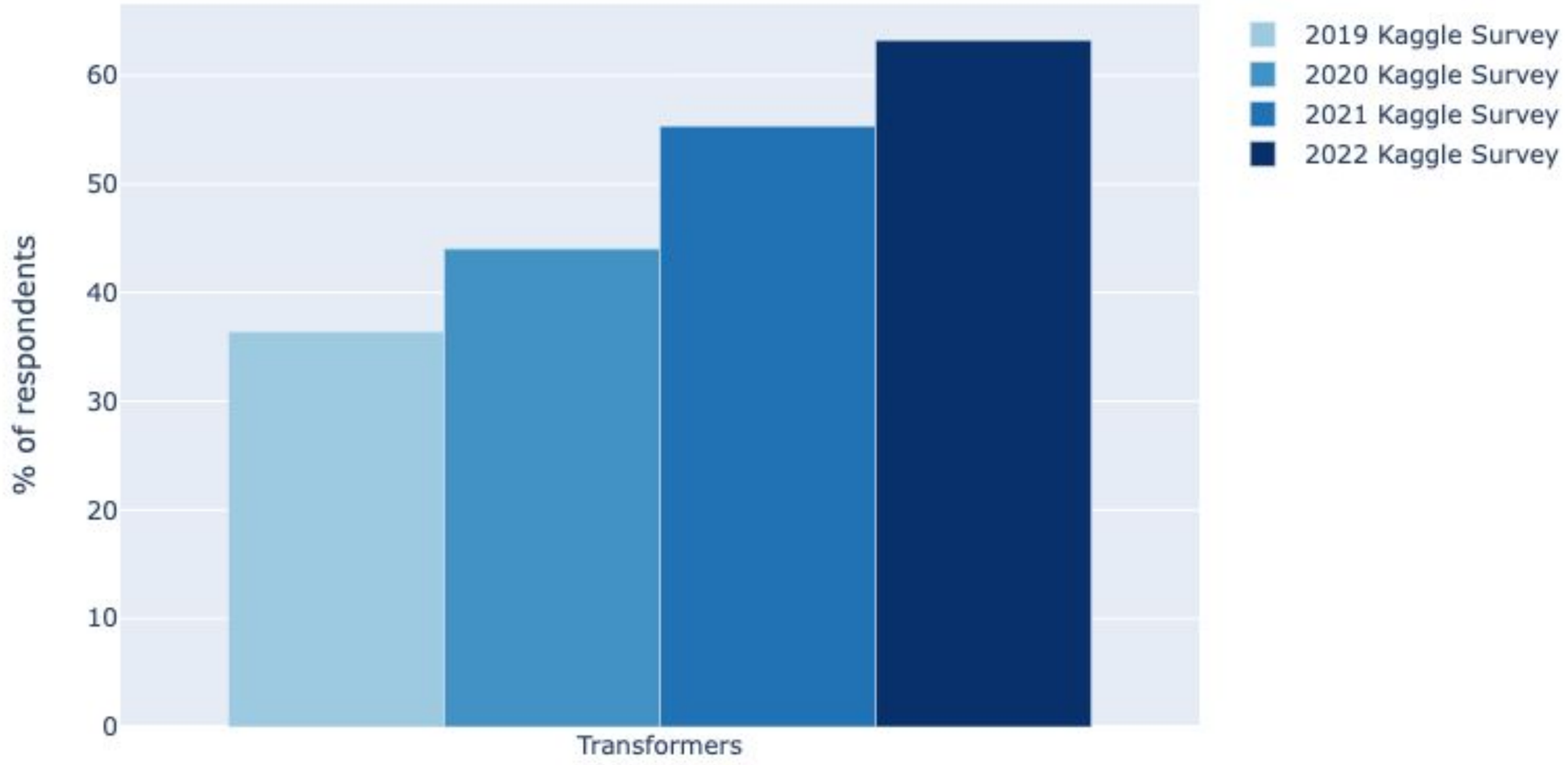
1. Does the shift toward VSCode and Jupyter Notebooks reflect a trend towards choosing IDEs that have the option of being hosted within a web browser? What do you think drives people’s choices of IDE?
2. Why would users be shifting away from desktop apps?

Machine Learning

Scikit-learn is the most popular ML framework while PyTorch has been growing steadily year-over-year



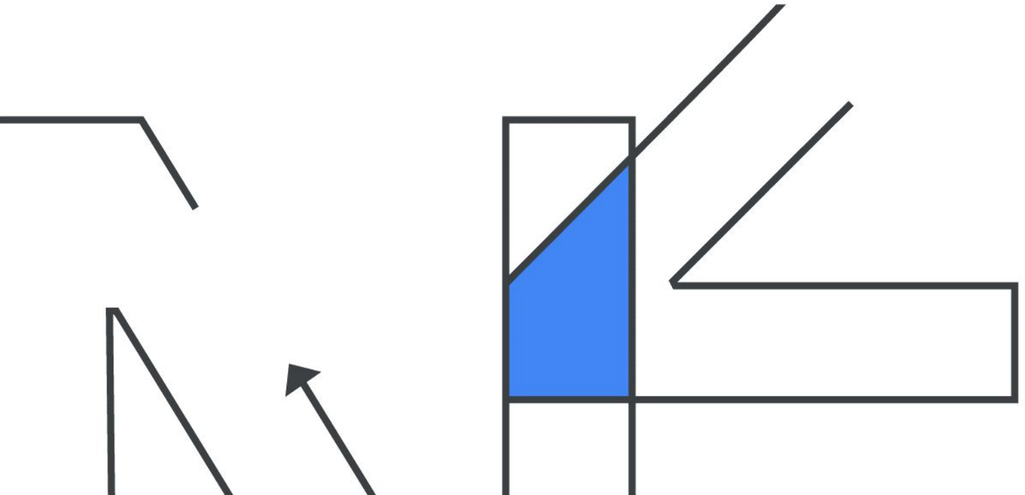
Transformer architectures are becoming more popular for deep learning models (both image and text data)



Panel

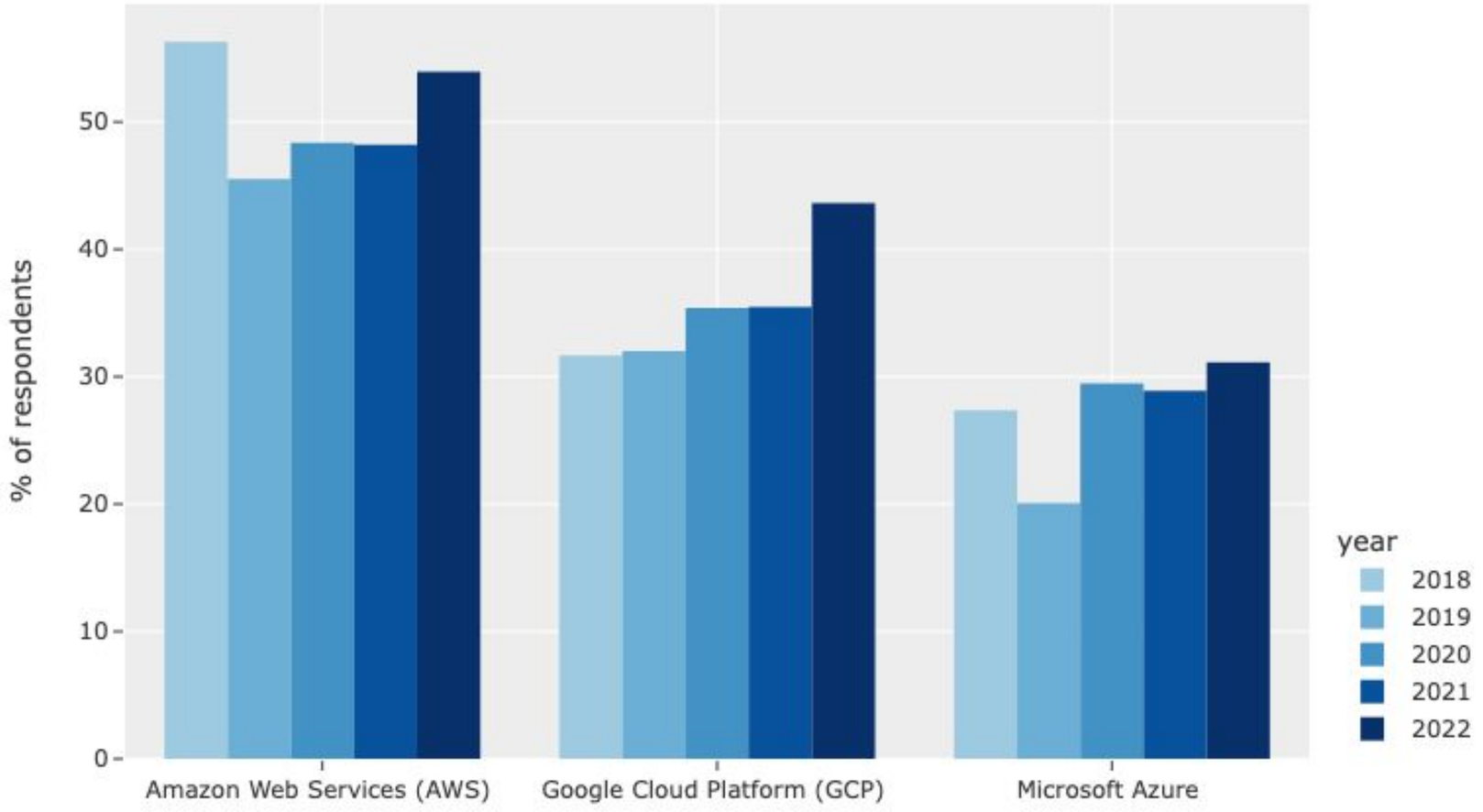
Questions

1. Do you suppose the popularity of scikit-learn is attributable to its ability to cover so many use cases?
2. Can you speak to the differences in which frameworks are best used for which applications?
3. How fundamental is tabular data in business? Do you see a clear winner in the boosted trees vs. tabular NNs space? Why are boosted trees dominant on Kaggle?

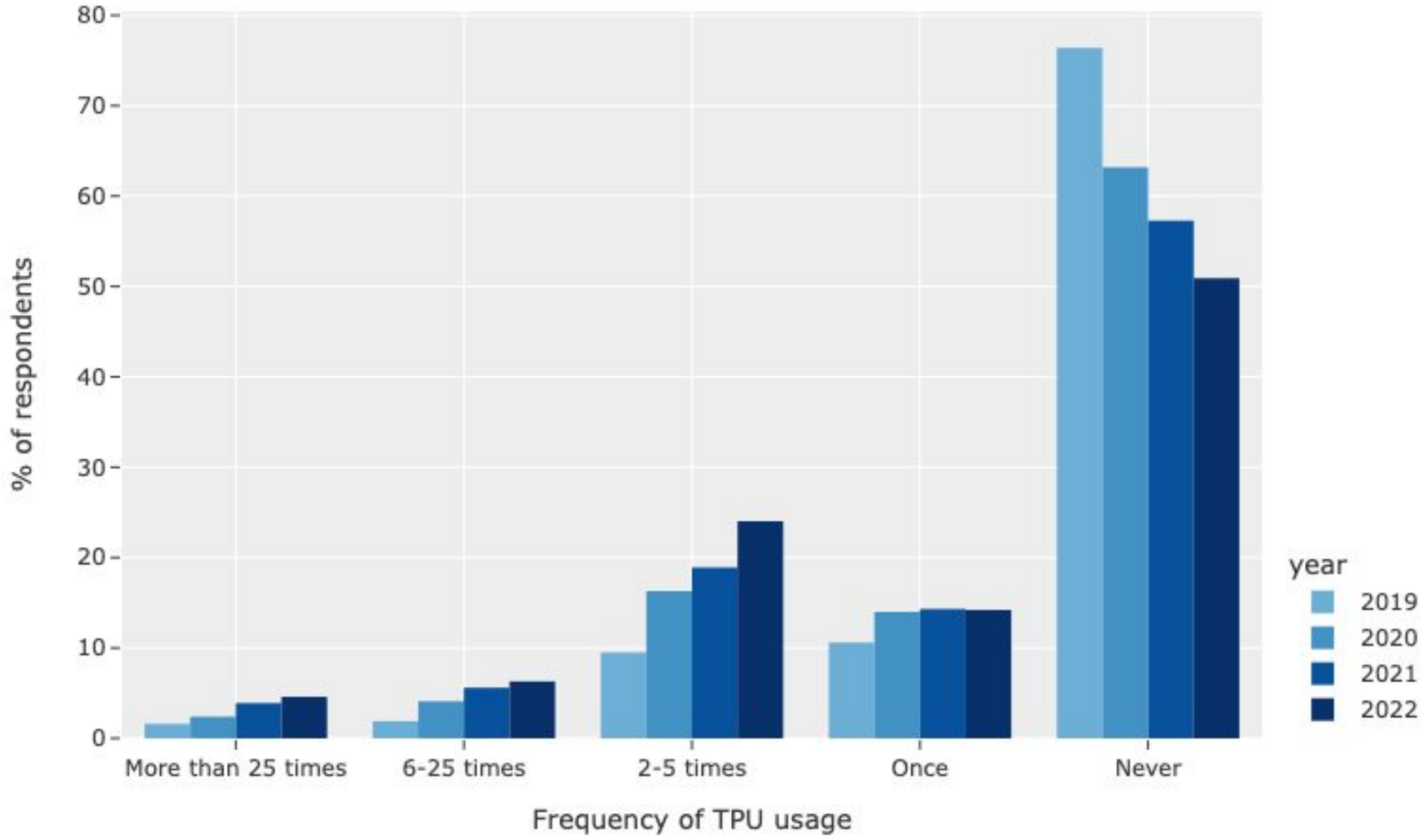


Cloud Computing

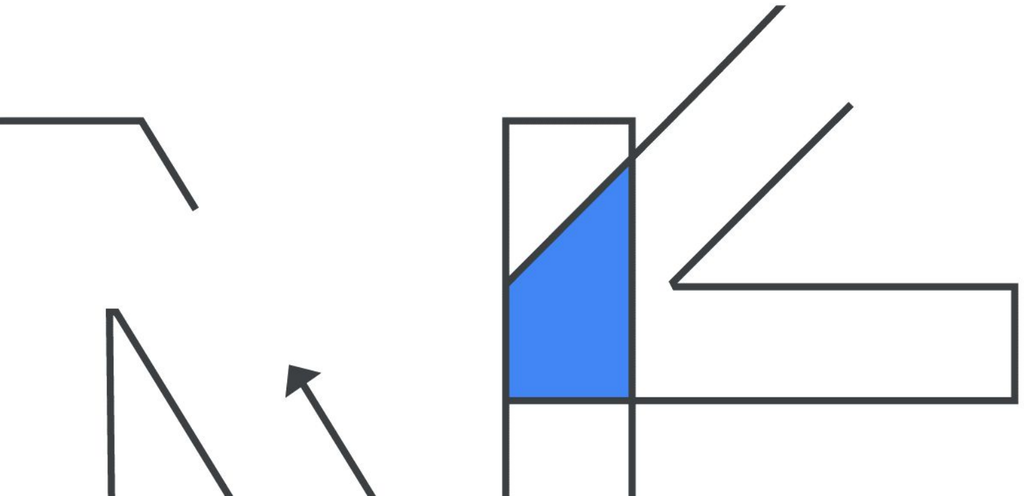
All major cloud computing providers saw strong year over year growth in 2022



Specialized hardware like Tensor Processing Units (TPUs) is gaining initial traction with Kaggle data scientists



Panel



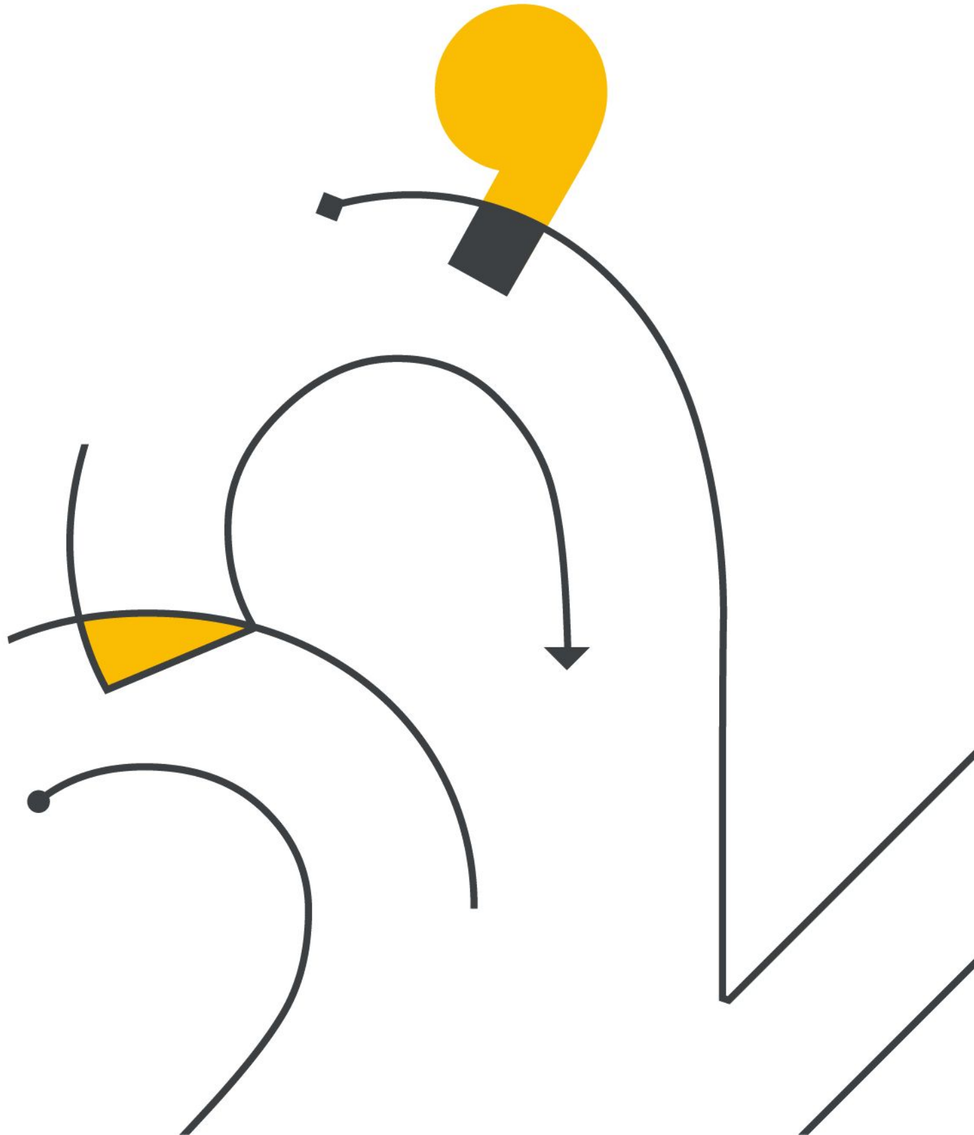
Questions

1. Can you share how users make choices in selecting between various cloud providers?

2. What do you think is driving the growth of accelerators? Are there projects better suited for these more specialized processors?

Download the full survey results at:

kaggle.com/kaggle-survey-2022



Thank you

Google Cloud
Next '22