StackOverFlow

The Technology Trends

ECE 143 Project: Group-4

Motivation & Research Questions

- How did StackOverFlow **evolve** over the years?
- What are the **key tags** that constitute SOF and do they differ in QA statistics?
- What **technologies declined** and what have **emerged** over these years?
- What key terms constitute the **Question titles**?
- How are the main technology tags **correlated** with each other?

StackOverFlow - Data

- A public platform building the definitive collection of technical questions & answers through crowdsourcing
- Public API to extract information on questions, answers, users and badges

GCP BigQuery

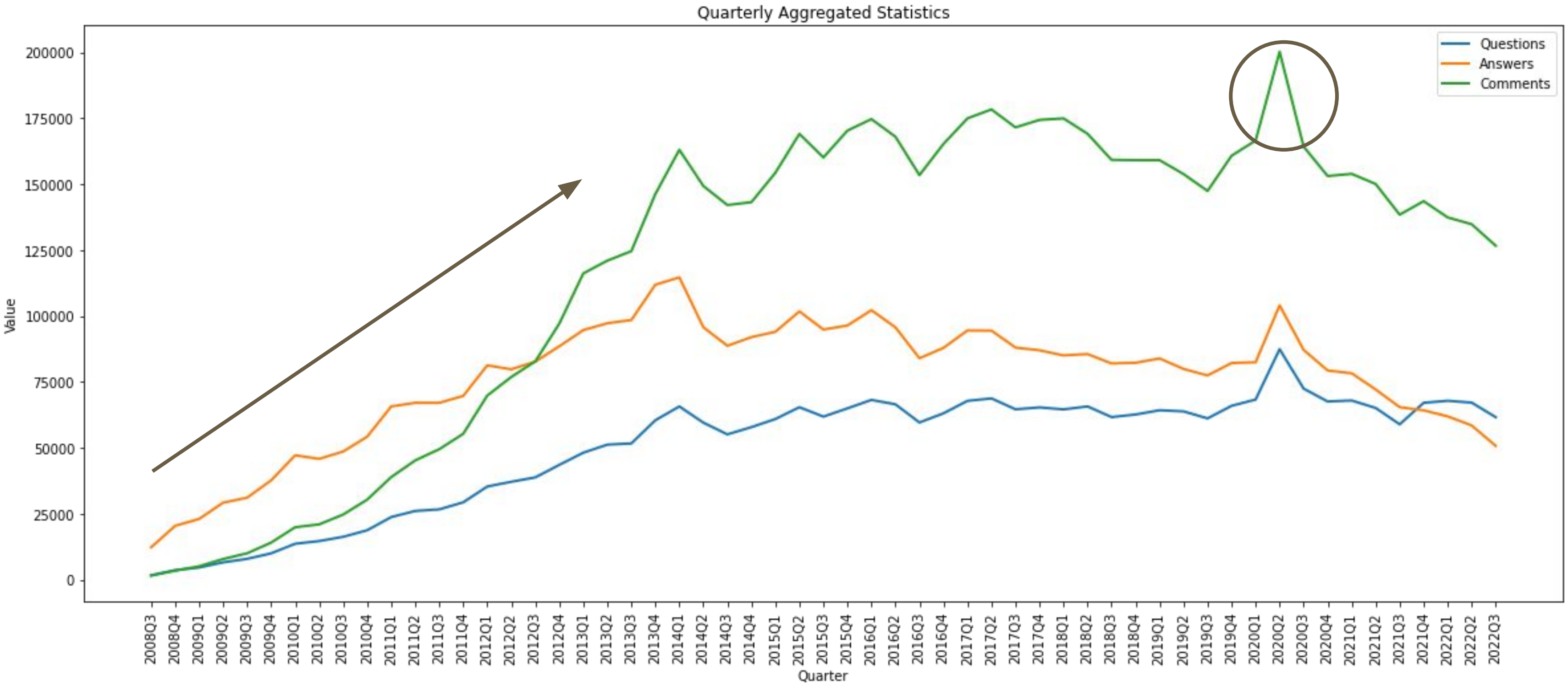
Data and Methodology

- Data from 2008 till date
- Filtered and stratified sampled for 45+ tags with total of 5.6M questions
- Tags are from the following main topics:

- Programming Languages
- Big Data
- Python Packages
- Cloud
- Data Science fields
- MLOps

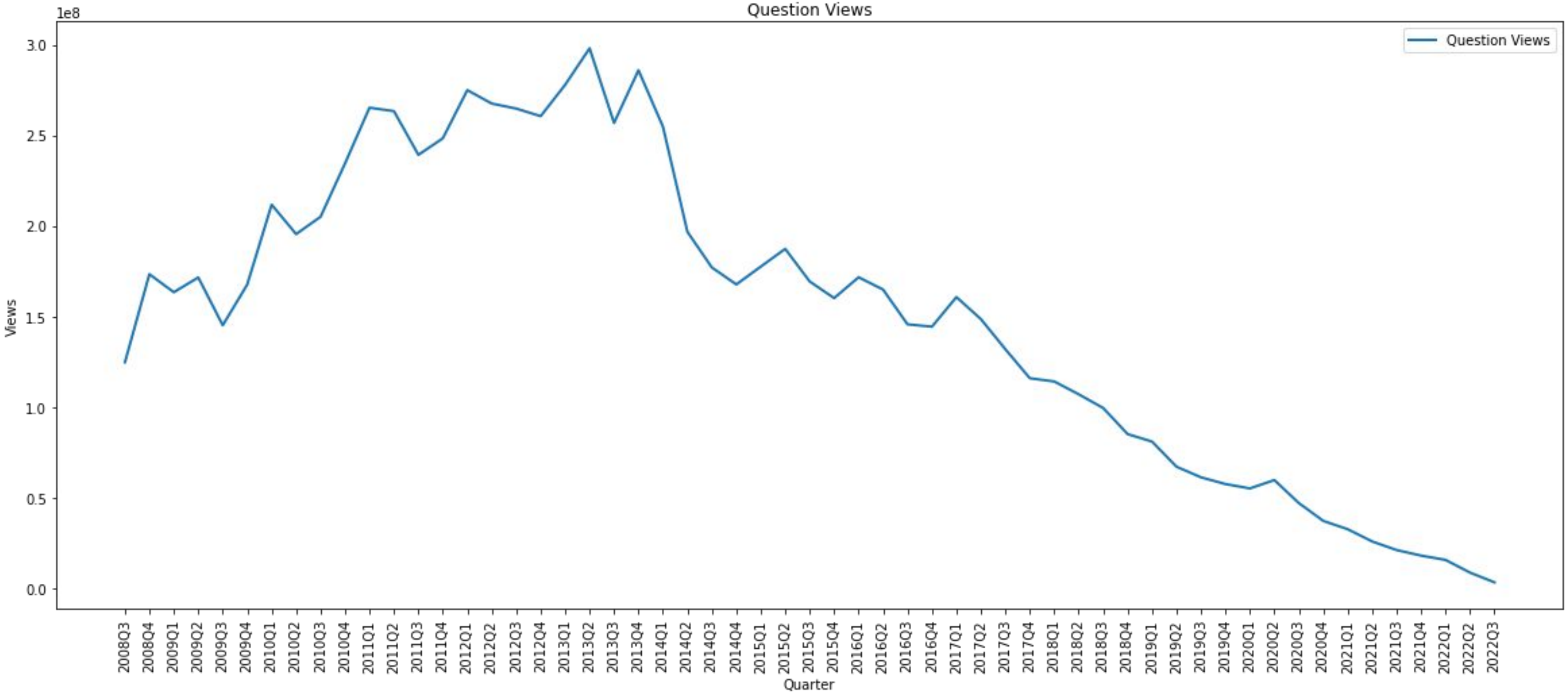
Overall Trends (1/2)

- Linear Growth from 2008-2013 (lot of new questions)
- Consistency from 2014-2019 (increased comments, decline in answers)
- COVID Peak (2020) - rise of online learning and WFH



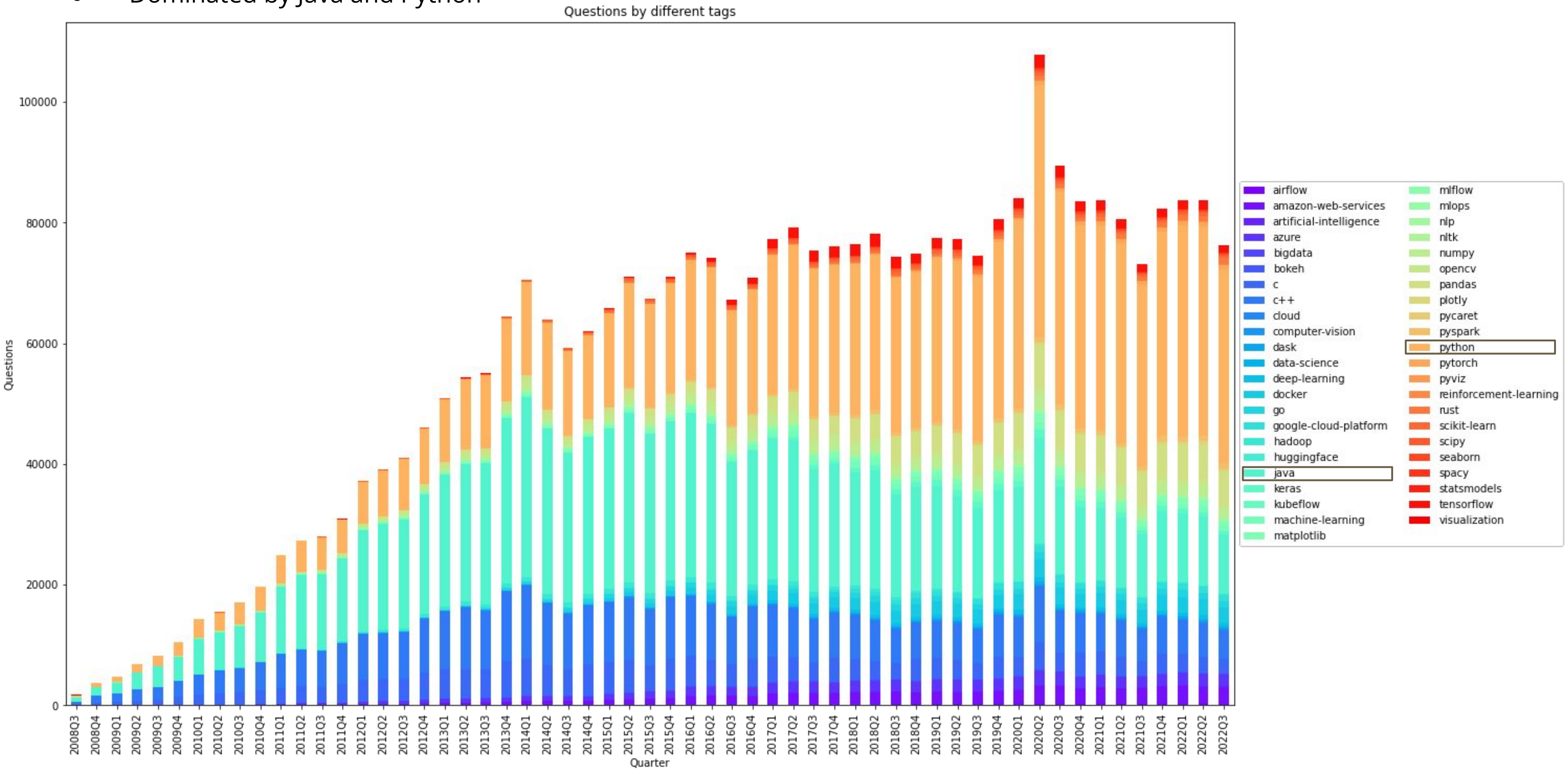
Overall Trends (2/2)

- Linear Growth from 2008-2013 (lots of new questions and views)
- Consistent decline in in views for questions from 2014 - Mostly due to reduced novelty, duplication, linking to old questions

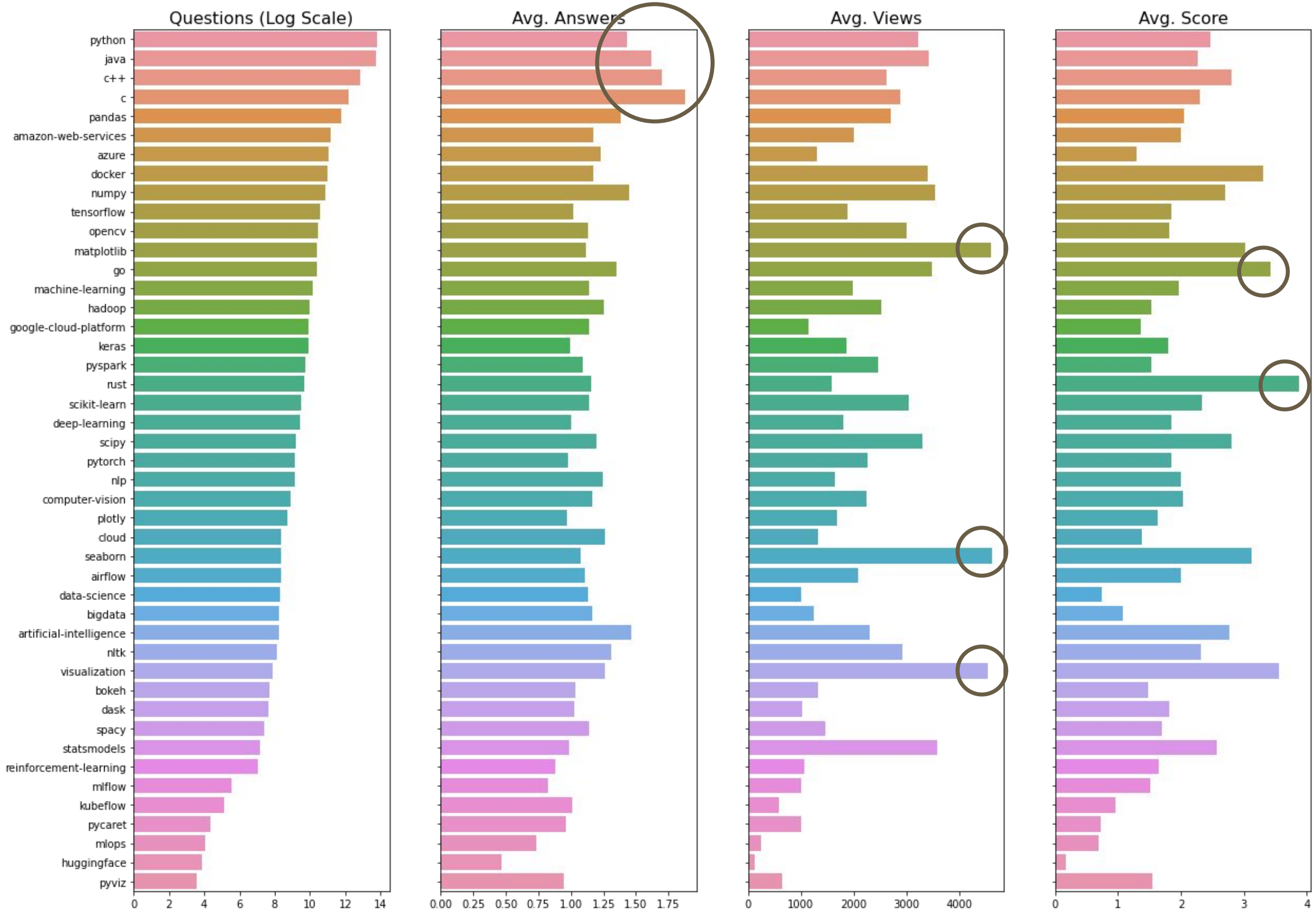


Trend by different tags

- Dominated by Java and Python



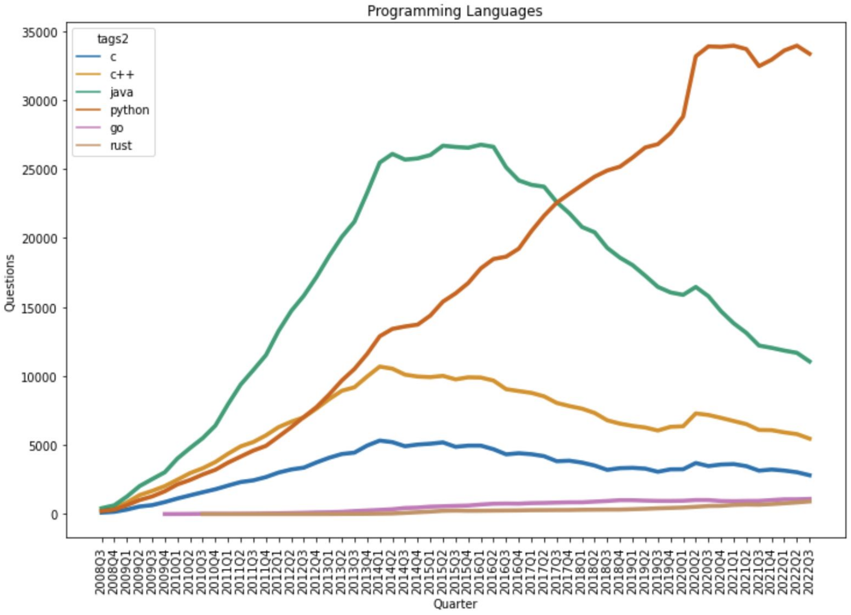
Statistics across different tags

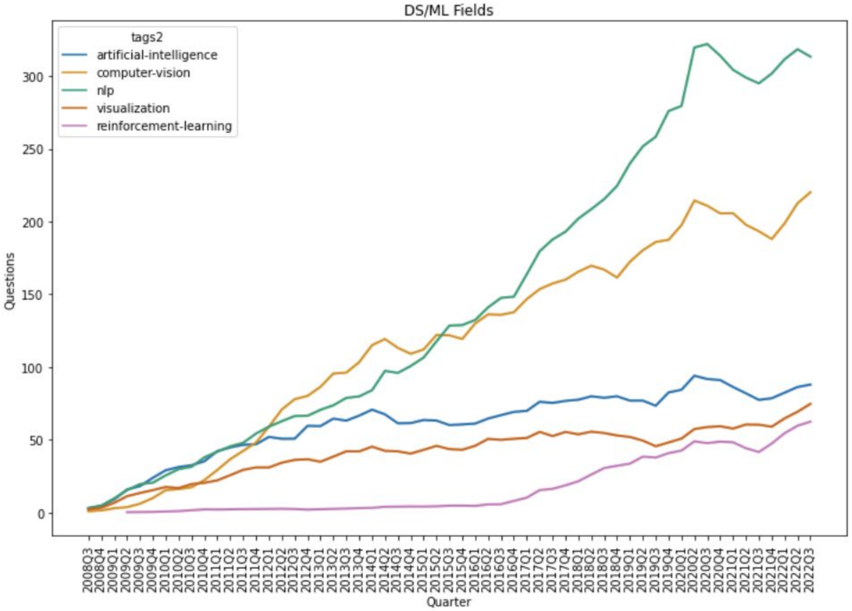


- **Avg. answers: Java, C++** (traditional languages like Java, C++ as they are present from long time)
- **Avg. Views: Visualization** (indicating less diversity and repeated problems for users)
- **Avg. score: New tools like Docker, Go**(indicating evolution of usage)

Growth of different technologies (1/3)

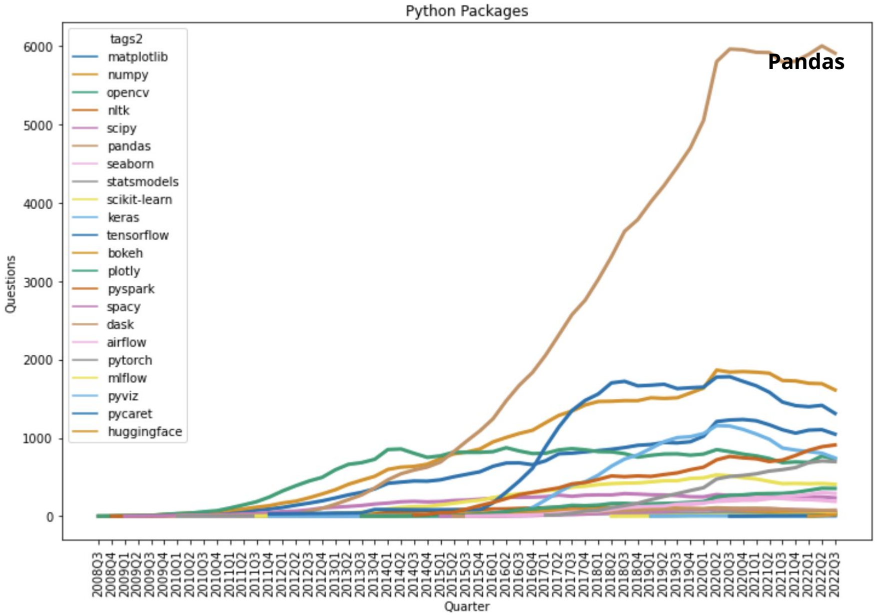
- Rise of Python, decline of Java from 2016, Go and Rust has stable base
- **NLP** has top growth, followed by **CV**
- **RL** rising of late

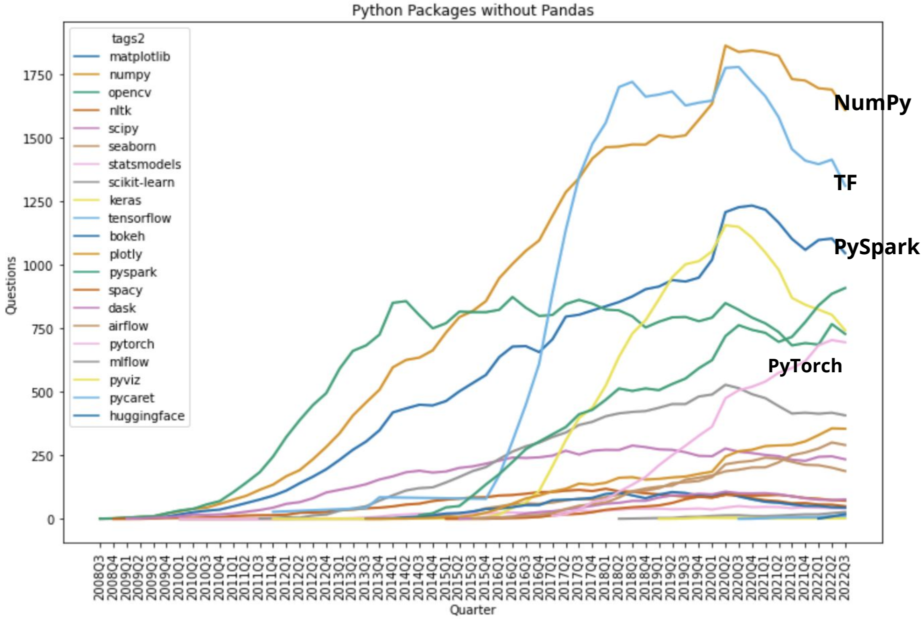




Growth of different technologies (2/3)

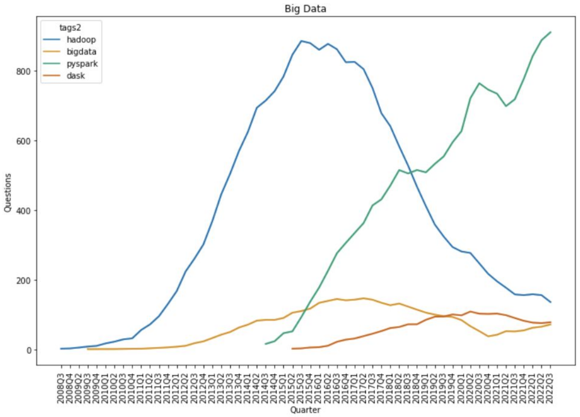
- **Pandas** dominating the Python toolkit
- **TF/Keras** Exp. growth from 2016, and declining since 2020, replaced by **PyTorch**
- Steady growth of **PySpark** since 2015
- **OpenCV**: Fastest growth since 2011, stagnated from 2014

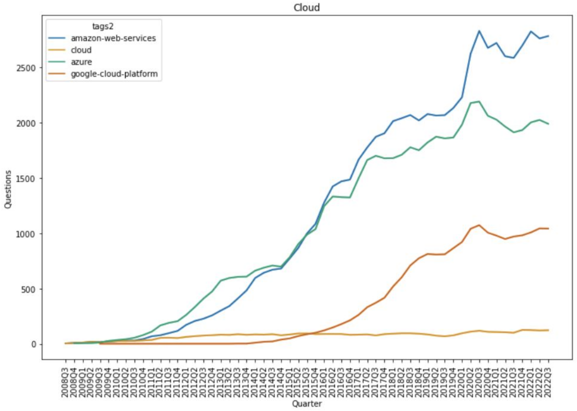


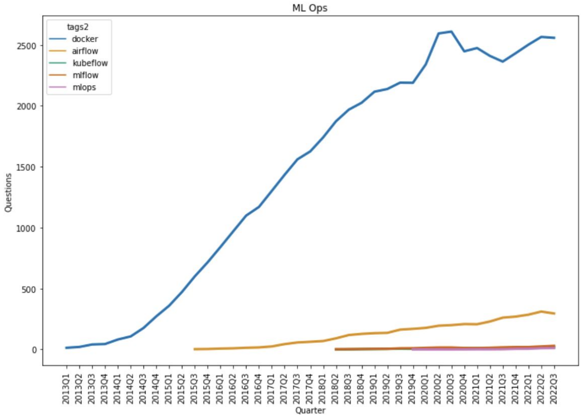


Growth of different technologies (3/3)

- BigData: **Hadoop** rose and fell exponentially, steady rise of **PySpark**
- **AWS** and **Azure** are early birds, **GCP** catching up
- **Docker** indispensable to ML Ops

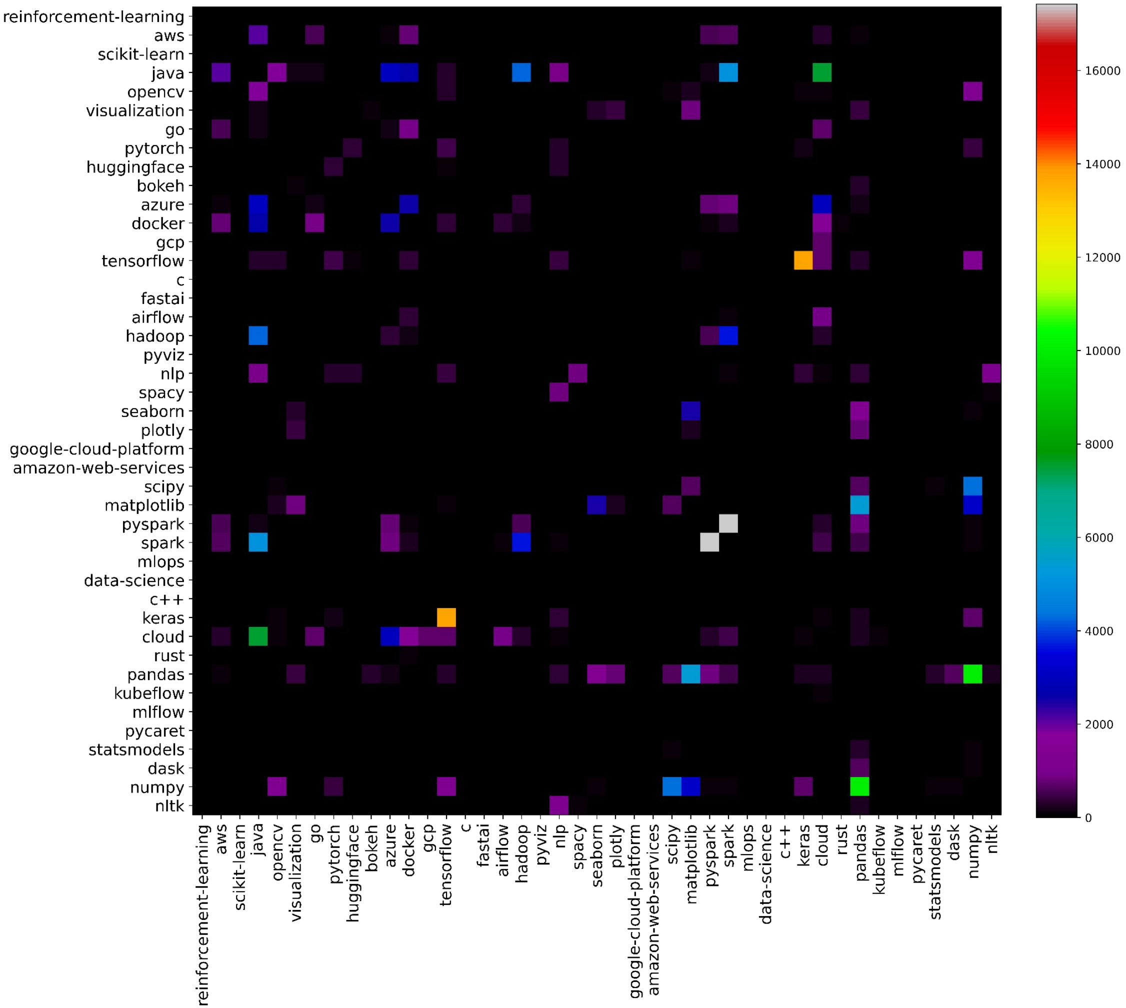






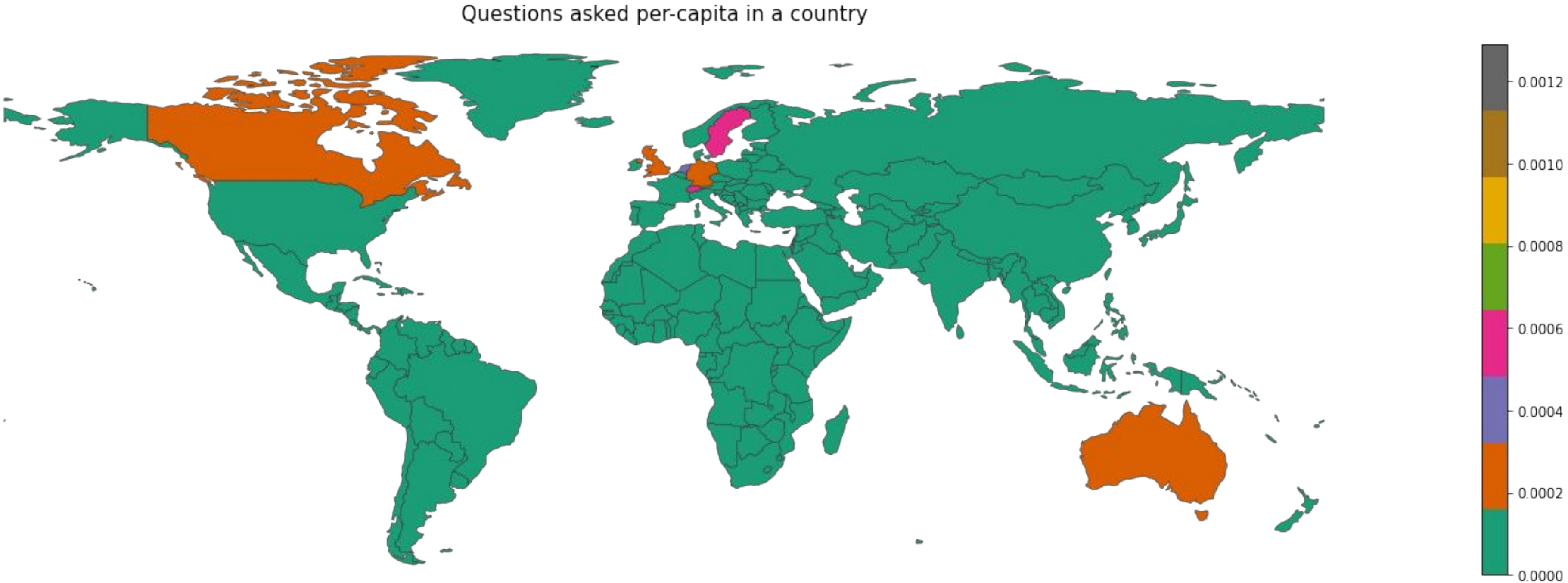
Co-occurrences

- **Java** - has high co-occurrence with cloud computing tech. (like **azure** and **spark**), also significantly used for **nlp**
- **Cloud** - apart from java, has high co-occurrence with **go** and **airflow**
- **Docker** - used along with java, **azure** and **aws**

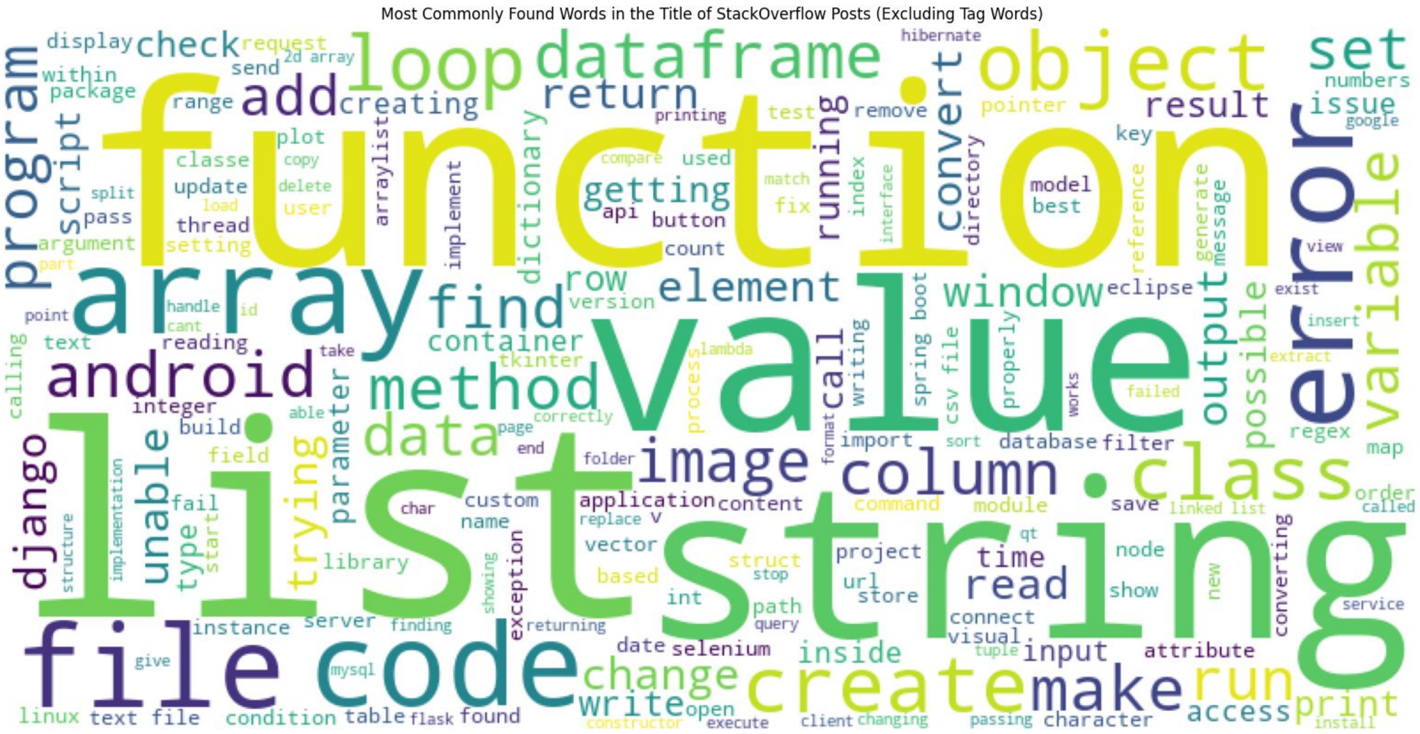


Questions by country (per capita)

- **Switzerland and Sweden**: Very high
- **Australia, Canada, Germany, UK**: high
- India, US: High Population driving this



Word Presence in Post Titles (excluding main tags)

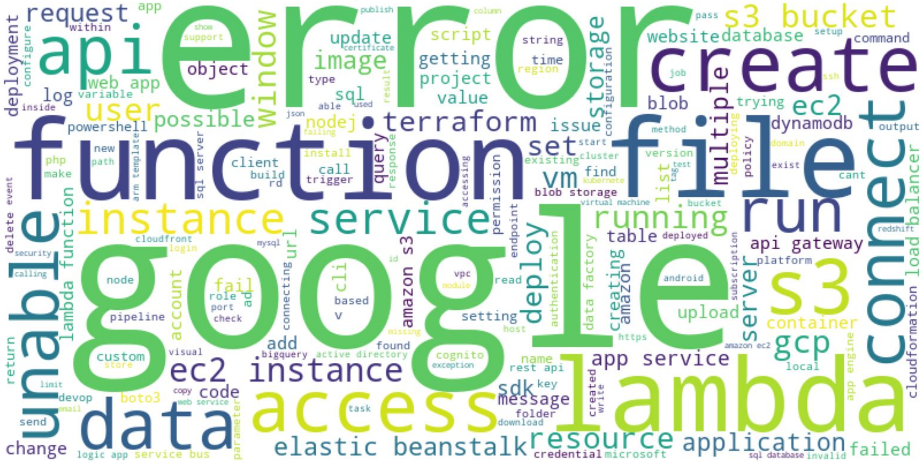


Most Commonly Found Words in the Title of StackOverflow Posts (Excluding Tag Words)

Word Presence in Post Titles (contd.)

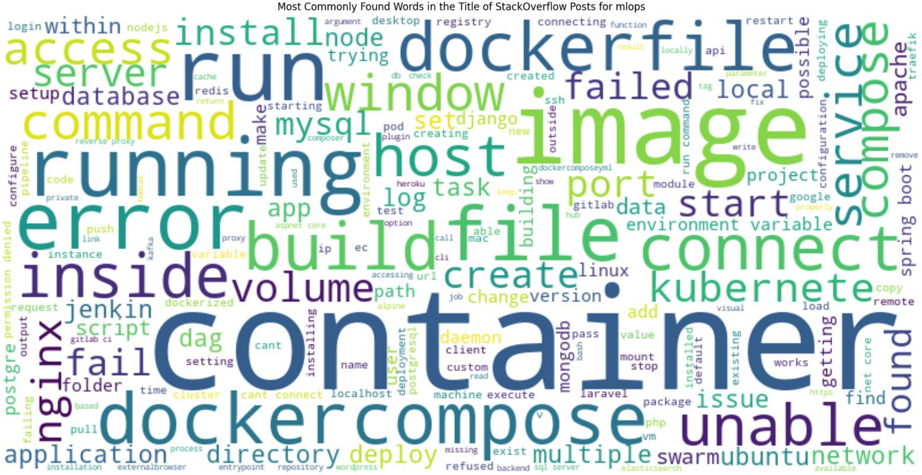
Cloud Computing

Most Commonly Found Words in the Title of StackOverflow Posts for cloud



ML Ops

Most Commonly Found Words in the Title of StackOverflow Posts for mlops



Thank You