Eliot Huang
Harley Mueller
Xu Lee
INFO 3300
Project 1, Written Summary

1. Getting the Data

We retrieved our data from StackOverflow's own database that contains information about its users, posts, tags, upvotes, etc. In order to filter out the data that we needed, tags containing certain programming languages throughout the course of StackOverflow's lifetime, Harley used SQL to maneuver through the data and download the files we needed as CSVs. We used SQL to process the data to find the tags over time for the most popular languages in 2009 and 2016 respectively. The link to the data dump is the following:
https://data.stackexchange.com/stackoverflow/query/new

In the first graph, the variables are year on the y axis, and number of questions on the y axis. The variable are year, month, month number, and count. This gave us the number of questions that were asked per month so that the first graph could be made from the data. The query entered was the following:

```
SELECT
    YEAR(CreationDate) as Year,
    MONTH(CreationDate) as Month,
    COUNT(CreationDate) as Count
FROM Posts
GROUP BY MONTH(CreationDate), YEAR(CreationDate)
ORDER BY YEAR(CreationDate) ASC, MONTH(CreationDate) ASC;
```

In order to get the data for the two big graphs, the following queries were entered with tag_name as the different tags. Harley typed this query for every tag that she was searching and compiled them into a csv file. We had to selectively choose a subset of the huge data, so we decided on the top 9 tags for 2009 and 2016.

```
SELECT
    YEAR(CreationDate) as Year,
    MONTH(CreationDate) as Month,
    COUNT(CreationDate) as Count
FROM Posts
WHERE Tags LIKE '<tag_name>'
GROUP BY MONTH(CreationDate), YEAR(CreationDate)
ORDER BY YEAR(CreationDate) ASC, MONTH(CreationDate) ASC;
```

The query that was used to get the top tags in 2009 and 2016 was the following:
Source: http://data.stackexchange.com/wordpress/query/edit/417509

```
select
       num.TagName as Tag,
       row_number() over (order by rate.Rate desc) as YearRank,
       rate.Rate as QuestionsInYear,
       num.Num as QuestionsTotal

from

(select count(PostId) as Rate, TagName
from
  Tags, PostTags, Posts
where Tags.Id = PostTags.TagId and Posts.Id = PostId
and Posts.CreationDate < '2009-12-31'
and Posts.CreationDate > '2009-01-01'
group by TagName) as rate

INNER JOIN

(select count(PostId) as Num, TagName
from
  Tags, PostTags, Posts
where Tags.Id = PostTags.TagId and Posts.Id = PostId
group by TagName
having count(PostId) > 800)
as num ON rate.TagName = num.TagName
order by rate.rate desc
;
```

2. Scales and Transformations Used

The scales that were used throughout were linear scales on the y axis and time scales on the x axis. In order to map from the data to the visual elements, we had to make sure that the dates were in the correct format. They all were formatted through timeFormat in order to standardize them to use scaleTime. The data all had to be scaled, and we chose to scale all the small multiples to the same value so that they could be easily compared to. We were not prioritizing clarity in the graphs because the point of the small multiples is not to pick out a particular point, but is to compare and contrast overall trends.

An ordinal scale was also used for the legends for the two stacked area graphs.

The orange color we chose was StackOverflow's orange color. We used the complement color blue to highlight the text explaining what we were graphing and key annotated points. On the stacked area line graphs, we used automatic colors given by d3.

3. The Story

The story that these graphs show is the trends and the ebb and flow of tags on StackOverflow since its inception in 2009. In the first graph, we could have graphed the total number of questions on StackOverflow, but with that, you couldn't tell that the website peaked in popularity in 2014. You also would not be able to see the periodicity in some of the graphs like Java and Python that we explain below.

Some interesting points to note is that, when looking at the small multiple graph of Java questions, StackOverflow questions seem to vary heavily depending on the season, with approximately two peaks detectable as far back as 2011. We suspect that this is because Java is a common introductory language to learn for college students, who would be asking more questions during the fall and spring semesters. This would also explain why the amount of Java questions dip back down towards December and at the start of summer. Similar trends can be found in Python, another introductory programming language that is becoming increasingly popular to teach.

Another interesting point is that there are many tags popular in 2009 that have practically faded out by 2016. Tags such as “iPhone”, “asp.net”, and “.net” used to be in the top 9, but now don't even land in the top 30. A fun fact is that although iPhone was the 9th most popular tag in 2009, it was *never* tagged in 2016. You can also use the small multiples to see more impressive trending tags such as “android”, which had very few tags in 2008, immediately gain popularity around 2011. These small multiple graphs allow the viewer to see how technologies trend rapidly in and out of favor in the tech community.