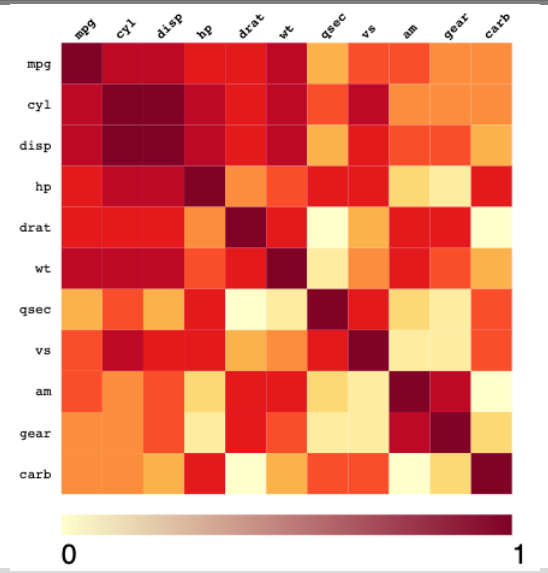
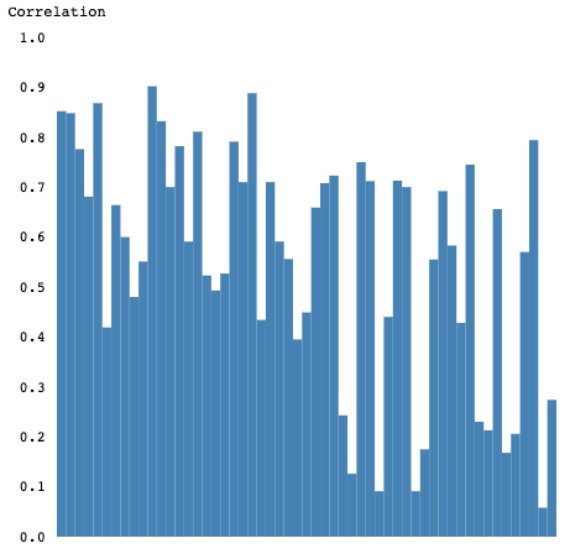
Karlsruhe Institute of Technology

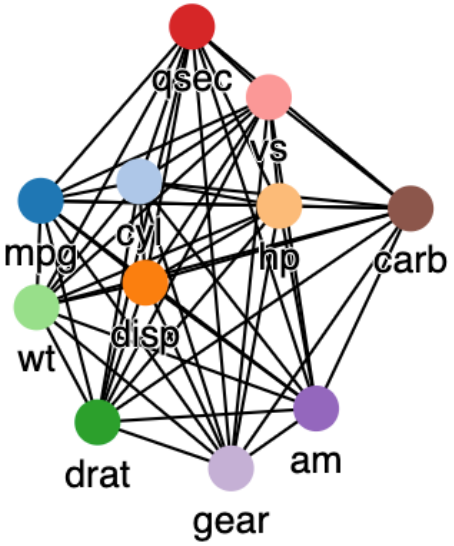
Interactive Visualization of Correlations in High-Dimensional Streams

Yimin Zhang | July 19, 2019

INSTITUTE FOR PROGRAM STRUCTURES AND DATA ORGANIZATION







Outline

- Motivation
- Challenges
- Related Work
- Interface
- Evaluation
- Summary

Motivation

- Correlation analysis aims at discovering and summarizing the relationship between the attributes of a data set
- Pearson correlation is in [-1,+1]
- An example on financial investments

Fund	U.S. Stock	European Stock	Pacific Stock	U.S. Bond	U.S. Money Market
U.S. Stock	1.00	.59	.33	.29	−.05
European Stock		1.00	.53	.22	−.13
Pacific Stock			1.00	.14	−.10
U.S. Bond				1.00	.14
U.S. Money Market					1.00

Fig: Correlations Among the Five Funds' Returns, Monthly Returns, from 1980 to 1998 by Katrina Simons [8]

Motivation

During the period 1987 – 2009:

- Iraq invasion in Kuwait/first war in Iraq
- Asian economic crisis
- Housing market boom
- Chinese economic growth
- Second war in Iraq
- Global financial crisis

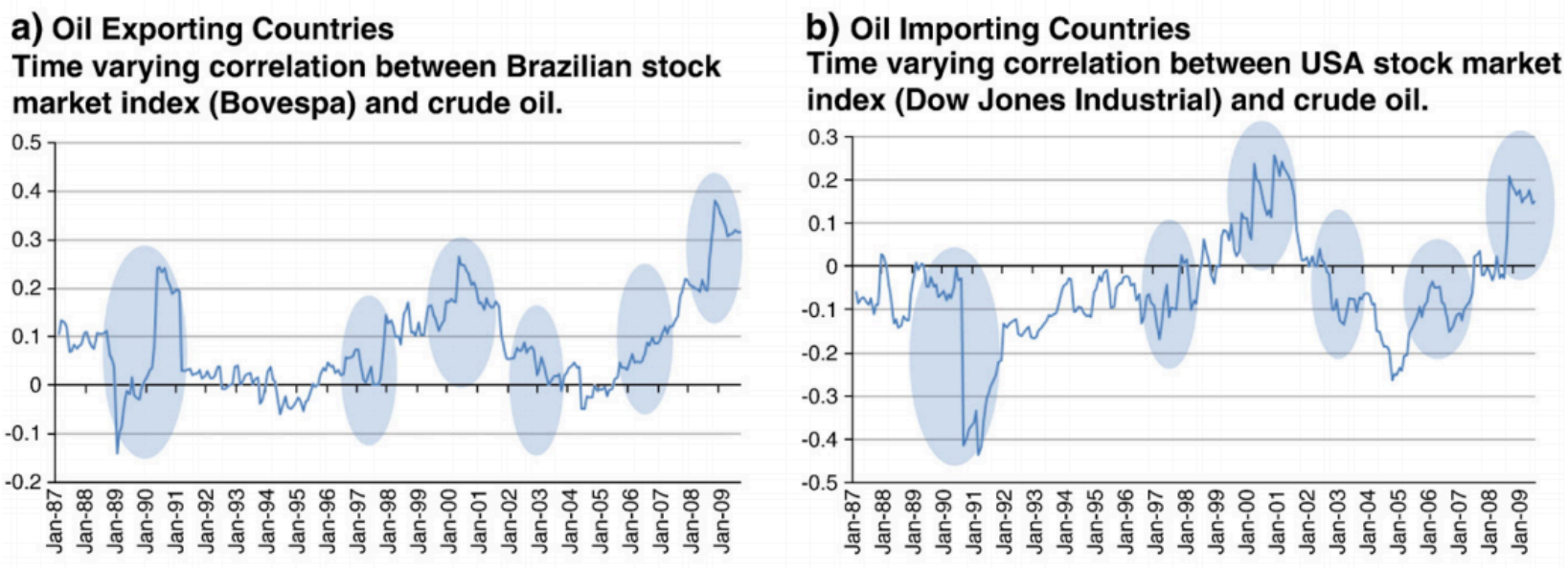


Fig: Dynamic correlation between stock market index and the crude oil price by George Filis et al. [3]

Challenges - 1. Streaming setting

- Data evolves over time
- Correlation analysis should not be static

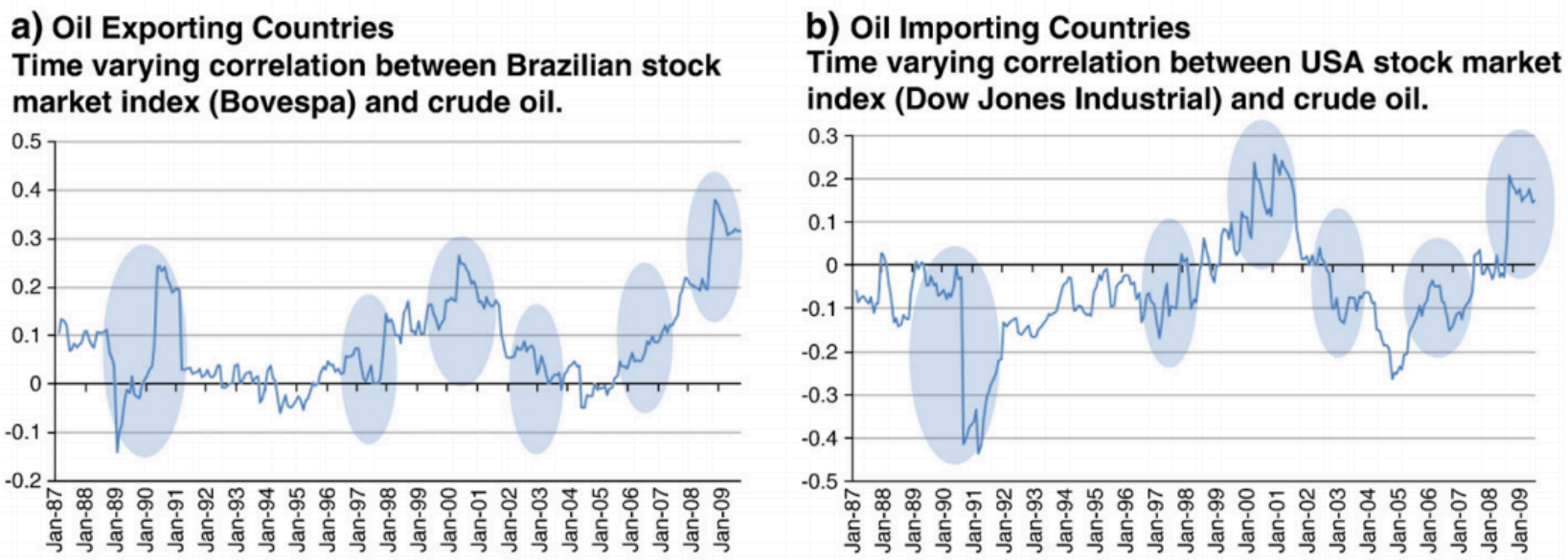


Fig: Dynamic correlation between stock market index and the crude oil price[3]

Challenges - 2.The high-dimensionality

- Compute the correlation between $\frac{n(n-1)}{2}$ pairs for the pairwise correlation analysis of any data steam with n components
- Difficult to extract actionable insights from the correlation analysis
- Impossible to understand the result of the correlation analysis

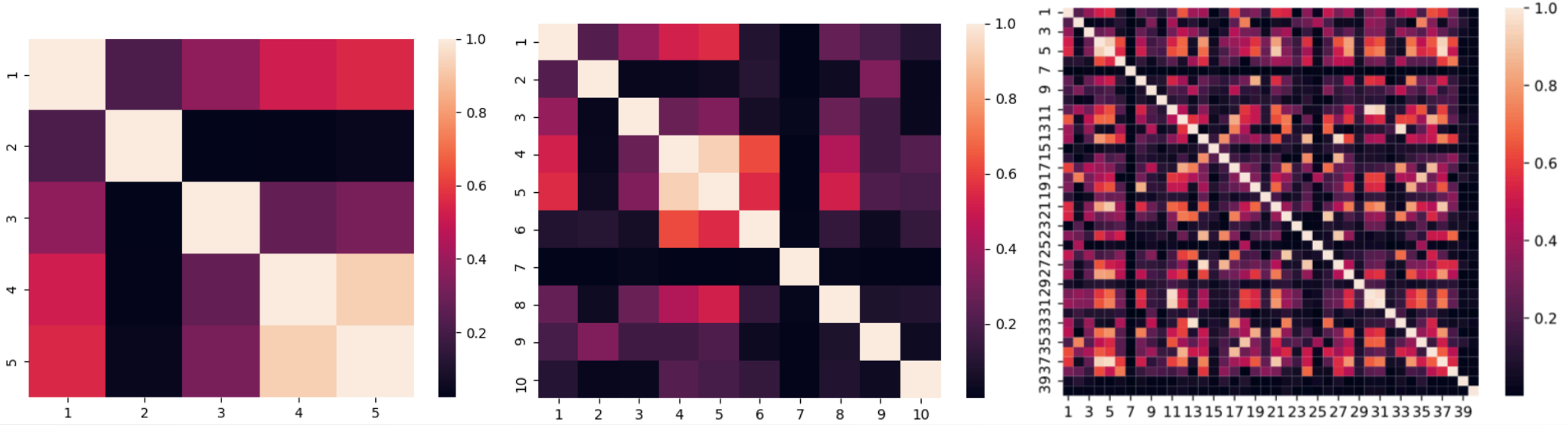


Fig: Correlation matrix of different number of attributes

Related Work – „FEXUM“ [2]

- Simultaneously visualize all feature correlations to the target and pairwise correlations using Force-Directed-Graph
- Nodes represent features and weighted edges represent distances
- Smaller distance between two features denotes a greater redundancy

Figure: „FEXUM“ by Louis Kirsch et al. [2]

Interactive Visualization of Correlations in High-Dimensional Streams

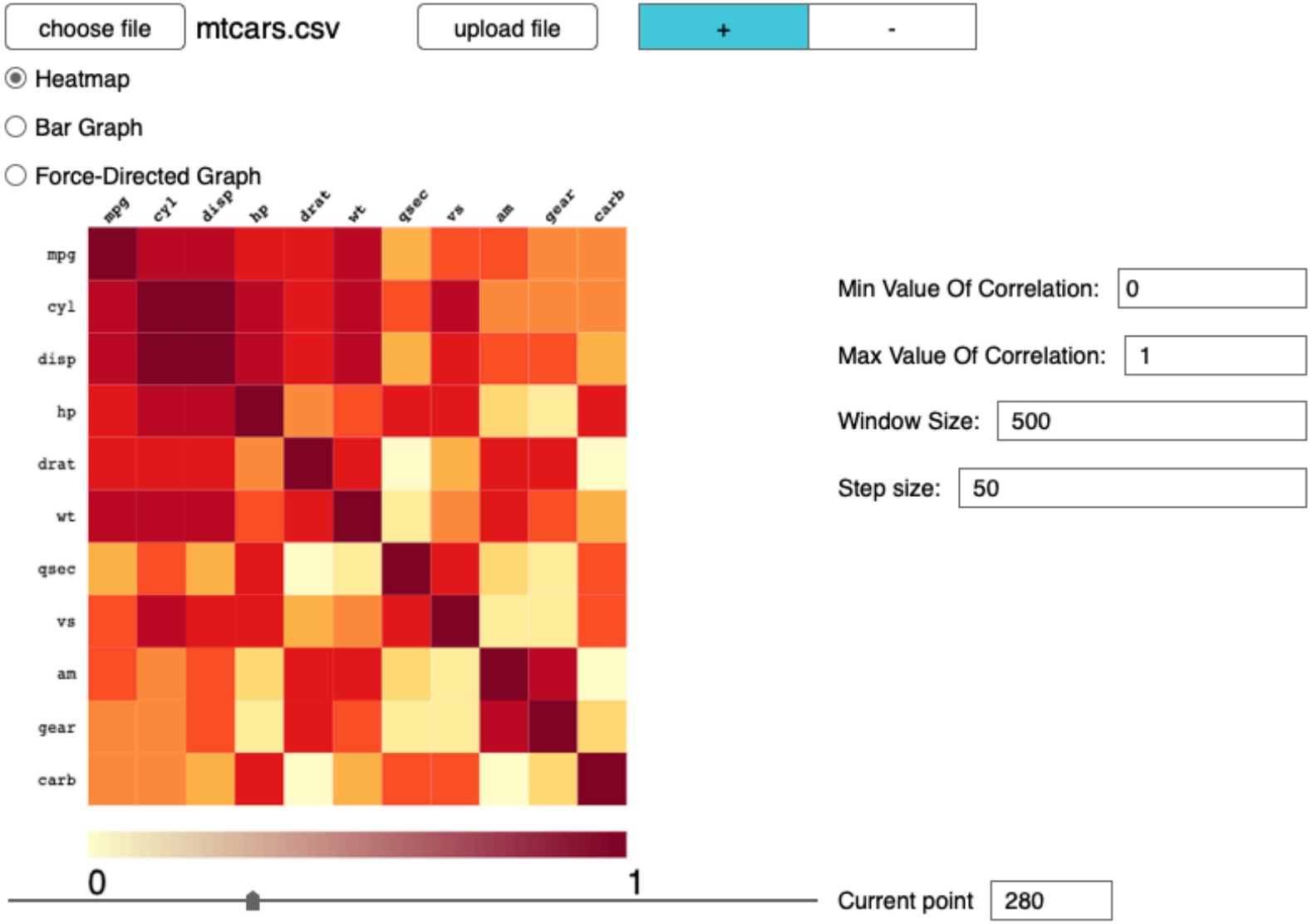


Fig: Mock-up of the interface

Interface – Visualization Methods

- Heatmap: variables with strong correlation (high values) are printed in dark colour and those with low correlation are in light colour

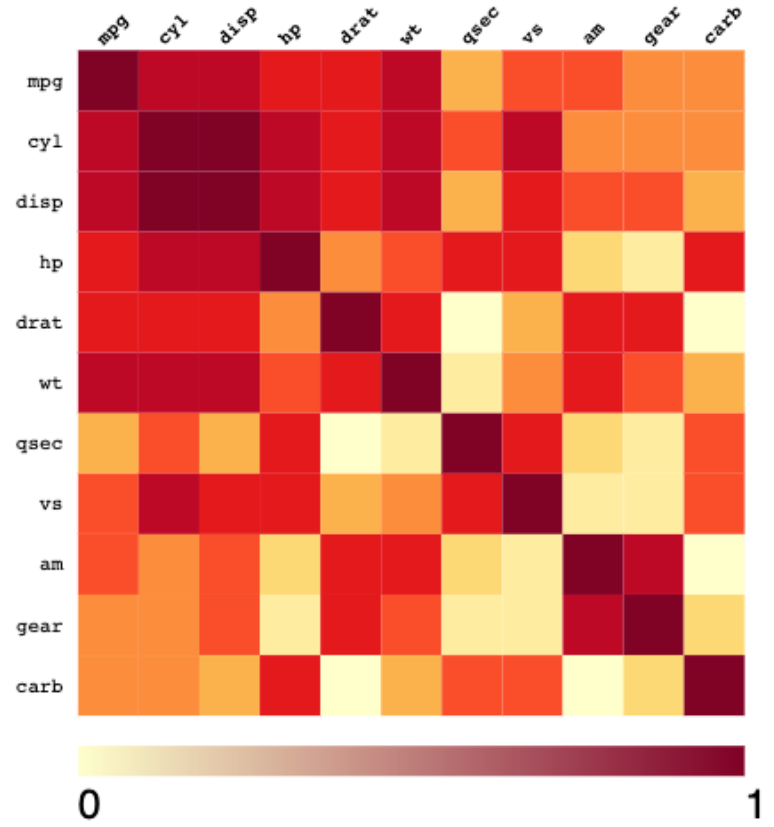


Fig: Heatmap of the Data Set Mtcars

Interface – Visualization Methods

- Bar Graph: presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent

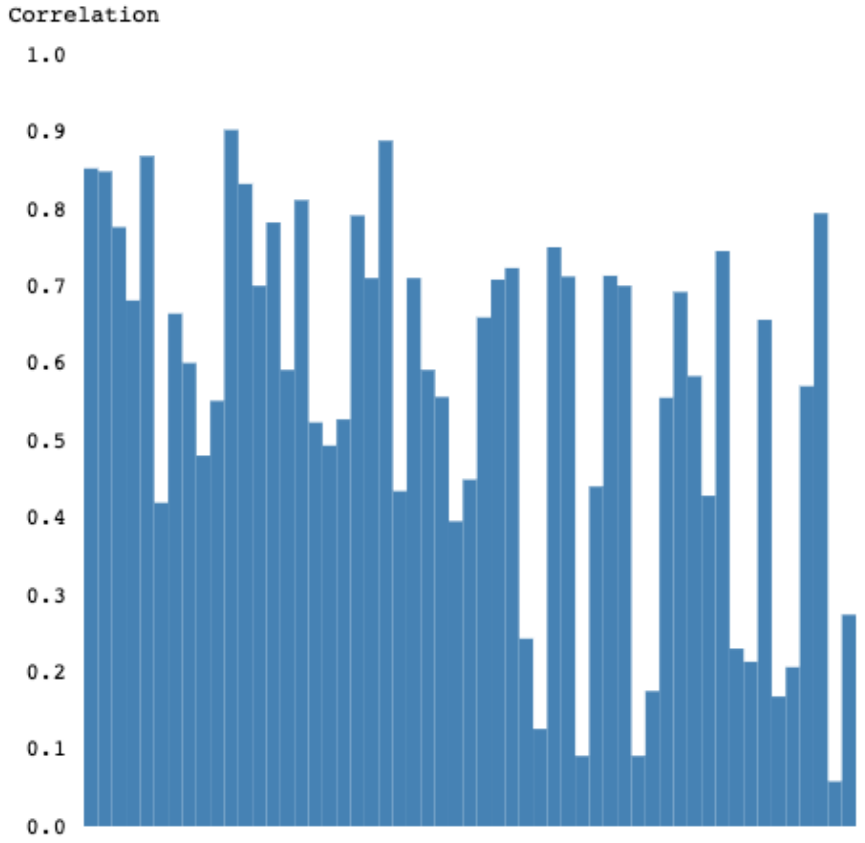


Fig: Bar Graph of the Data Set Mtcars

Interface – Visualization Methods

- Force-Directed-Graph
  - assigns forces among the set of edges and the set of nodes of a graph drawing
  - length of link between two nodes: $200 * \sqrt{(1 - x^2)}$

    x *is the correlation value*

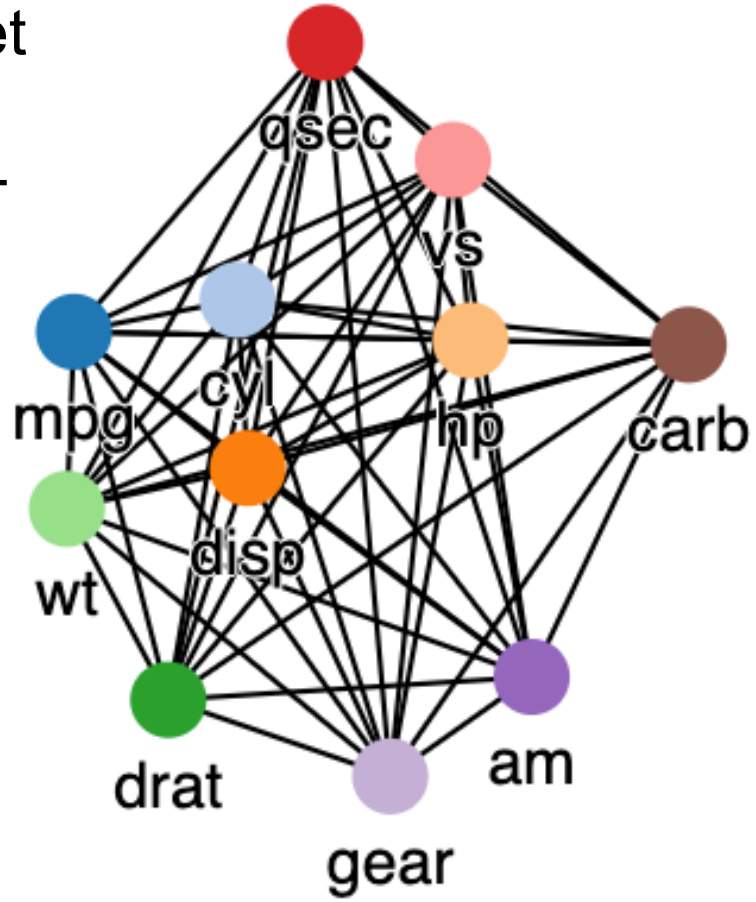
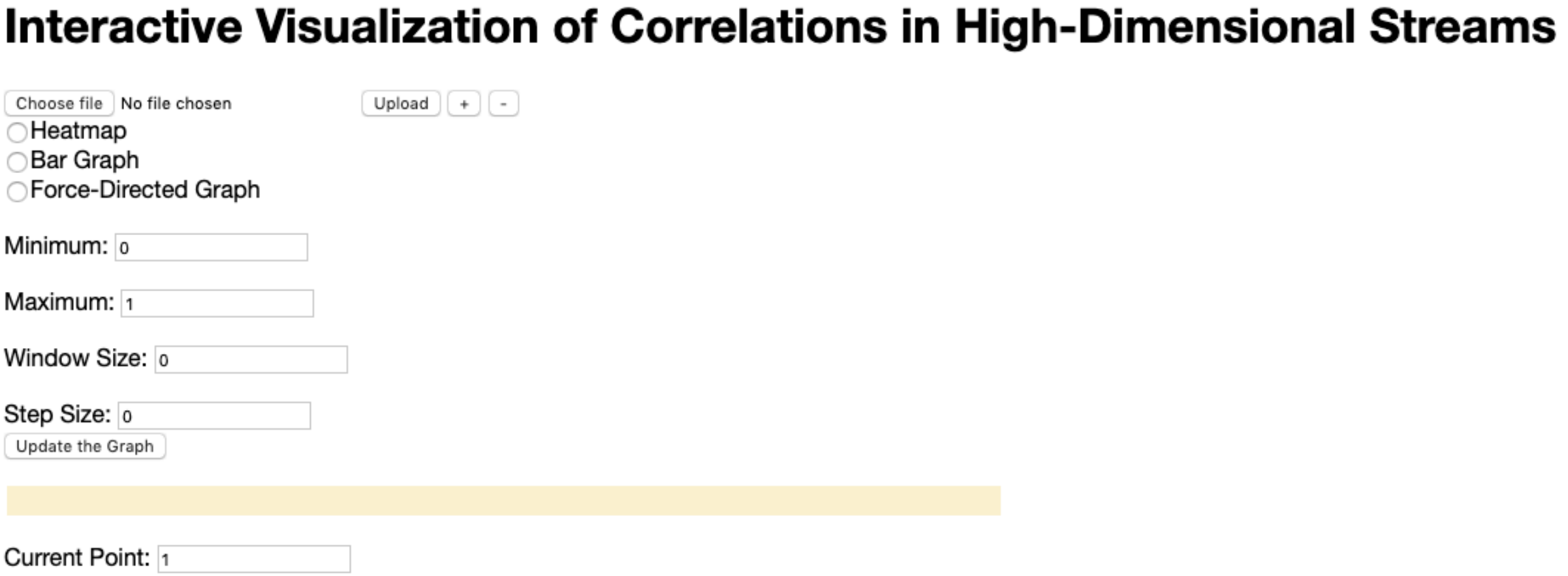


Fig: Force-Directed-Graph of the Data Set Mtcars

Interface

- Javascript (D3.js for producing dynamic, interactive data visualizations in web browsers)



Evaluation – Reasearch Questions

- What visualization method is the most appropriate to visualize correlation for various specific user information needs?
- What visualization method is the most suitable to visualize characteristics of a data set?
- What are the desirable features of a correlation monitoring interface?

Evaluation – Experimental Settings

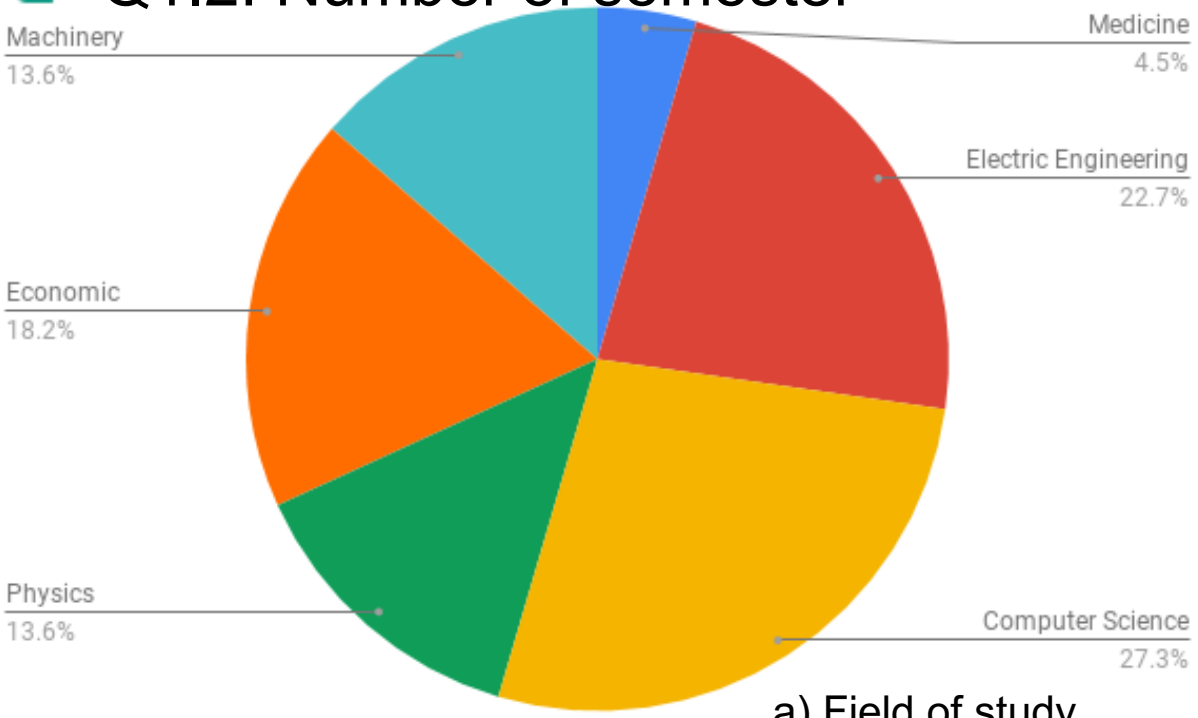
- A preliminary study (sample size might be too small)
- Large-scale study -> Future work
- Participants are assigned randomly to different conditions:
  - Condition A: Participants can only use the Heatmap
  - Condition B: Participants can only use the Bar Graph
  - Condition C: Participants can only use the Force-Directed-Graph
  - Condition D: Participants can use any visualization method they want, which are mentioned in Condition A, B and C
- Real-world data set at random: modifications on Data set Bioliq
  - Data Set 1 (DS1): 10 attributes within 2000 instances
  - Data Set 2 (DS2): 20 attributes within 2000 instances
  - Data Set 3 (DS3): 40 attributes within 2000 instances
- The window size is set to 200 and the step size is set 50 for all three data sets

Evaluation – Statics

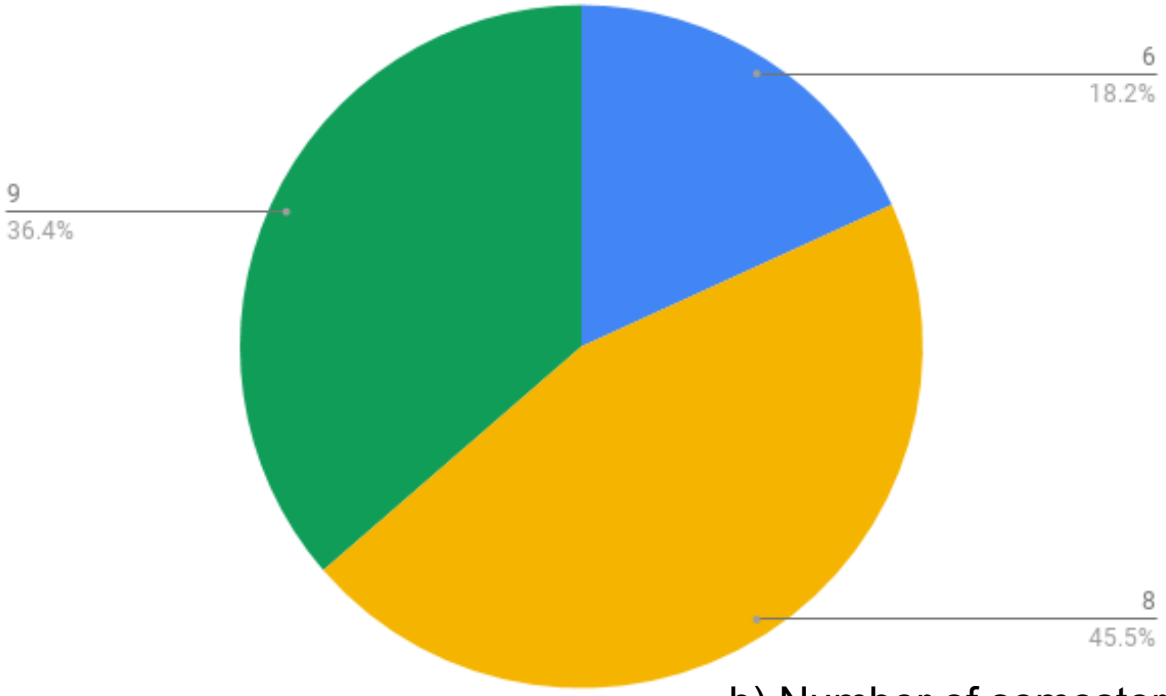
1.Basic Information:

22 participants from KIT, studying different majors

- Q1.1: Field of study
- Q1.2: Number of semester



a) Field of study

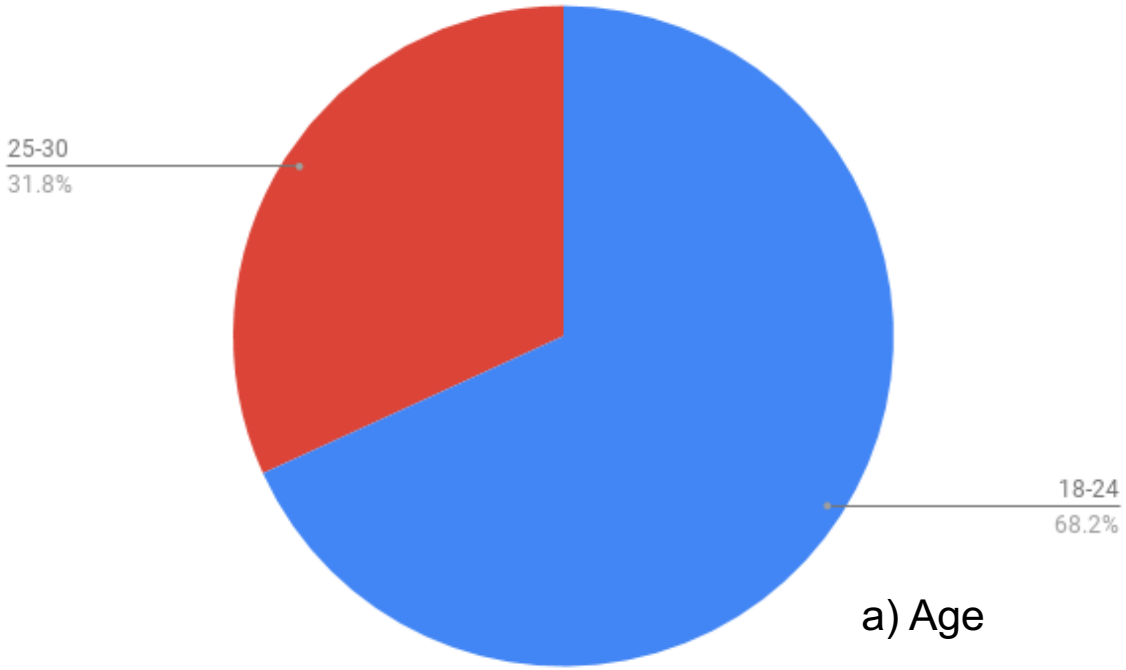


b) Number of semester

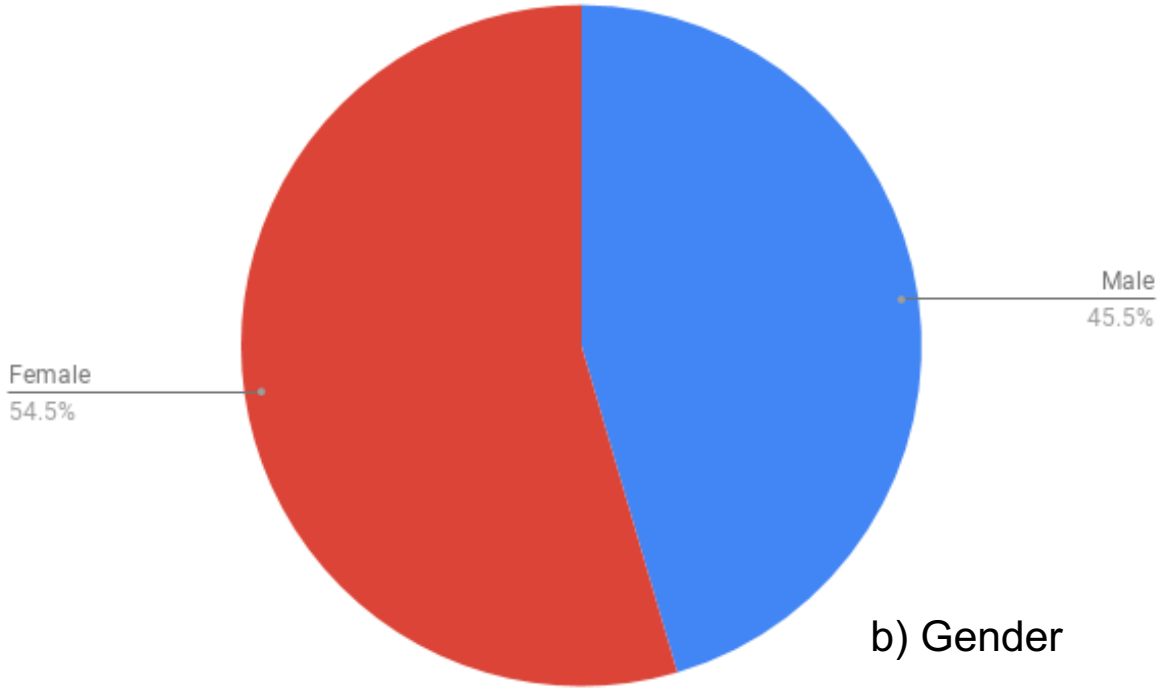
Evaluation – Statics

1.Basic Information:

- Q1.3: Age: 18-24 25-30 More than 30
- Q1.4: Gender: M F Do not wish to answer



a) Age

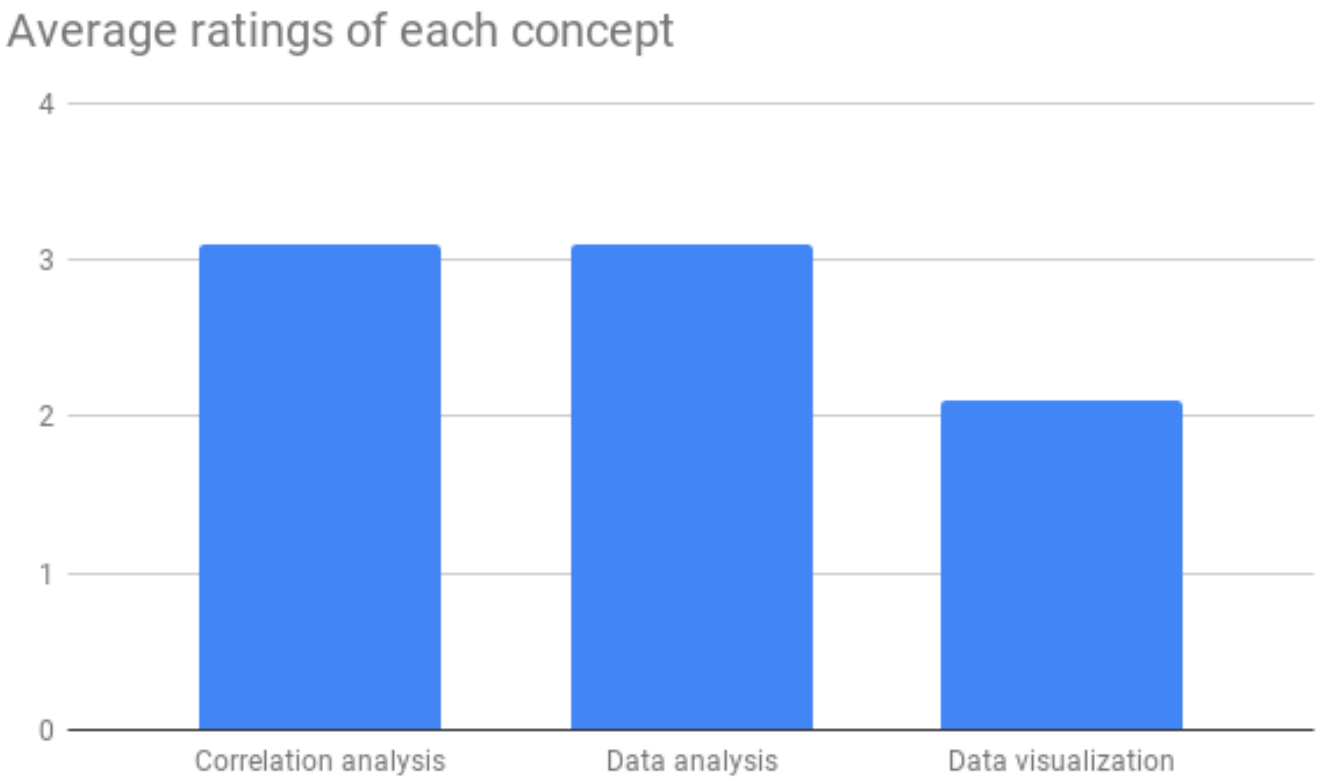


b) Gender

Evaluation – Statics

1.Basic Information:

- Q1.5: How familiar are you with the following concepts?
  - Using LIKERT scale (1-7)
  - 1: unfamiliar
  - 7: very familiar



Evaluation

- Q2.1: How many attributes are available in this data set?
- Q2.2: How many pairs of attributes are available in this data set?

Whatever method the user uses, the user can get the correct answer.

Evaluation

Q2.3: What is the correlation value between *Attribute A* and *Attribute B* at *Timestamp T* (with precision +- 0.1) ?

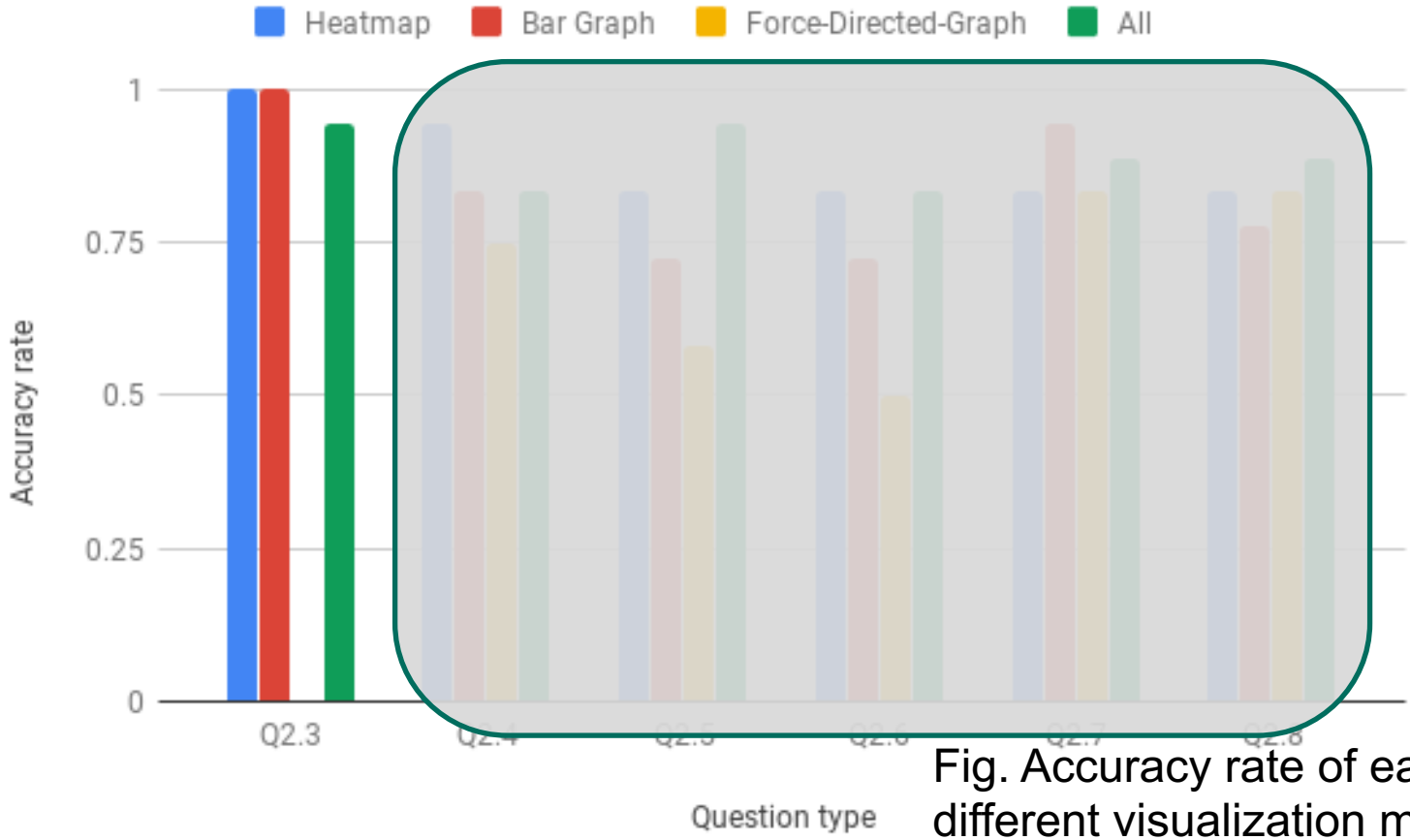


Fig. Accuracy rate of each question type using different visualization methods

Evaluation

Q2.4: Which pair of attributes has the biggest correlation at *Timestamp T* ?

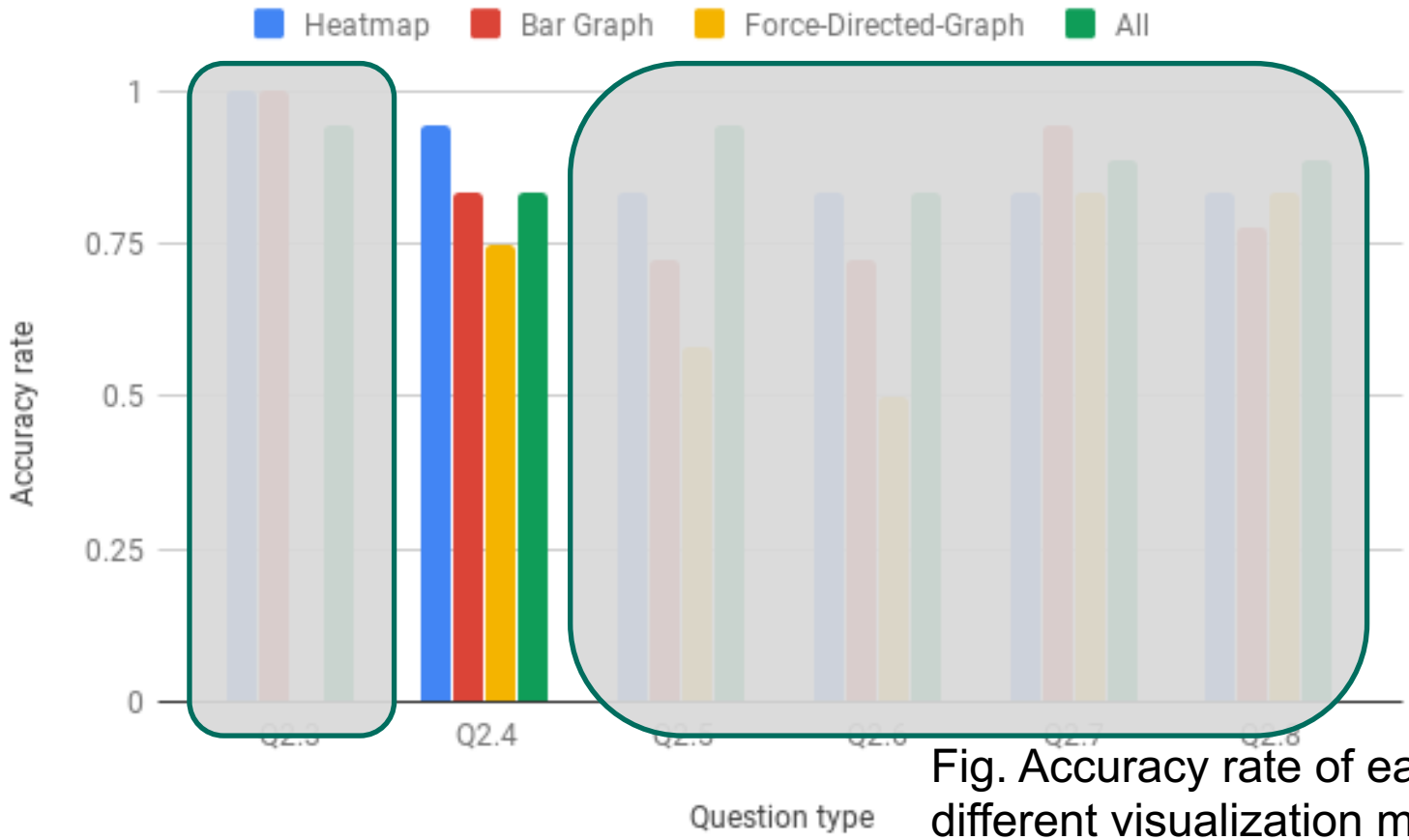


Fig. Accuracy rate of each question type using different visualization methods

Evaluation

Q2.5: Which pair of attributes has the smallest correlation at *Timestamp T* ?

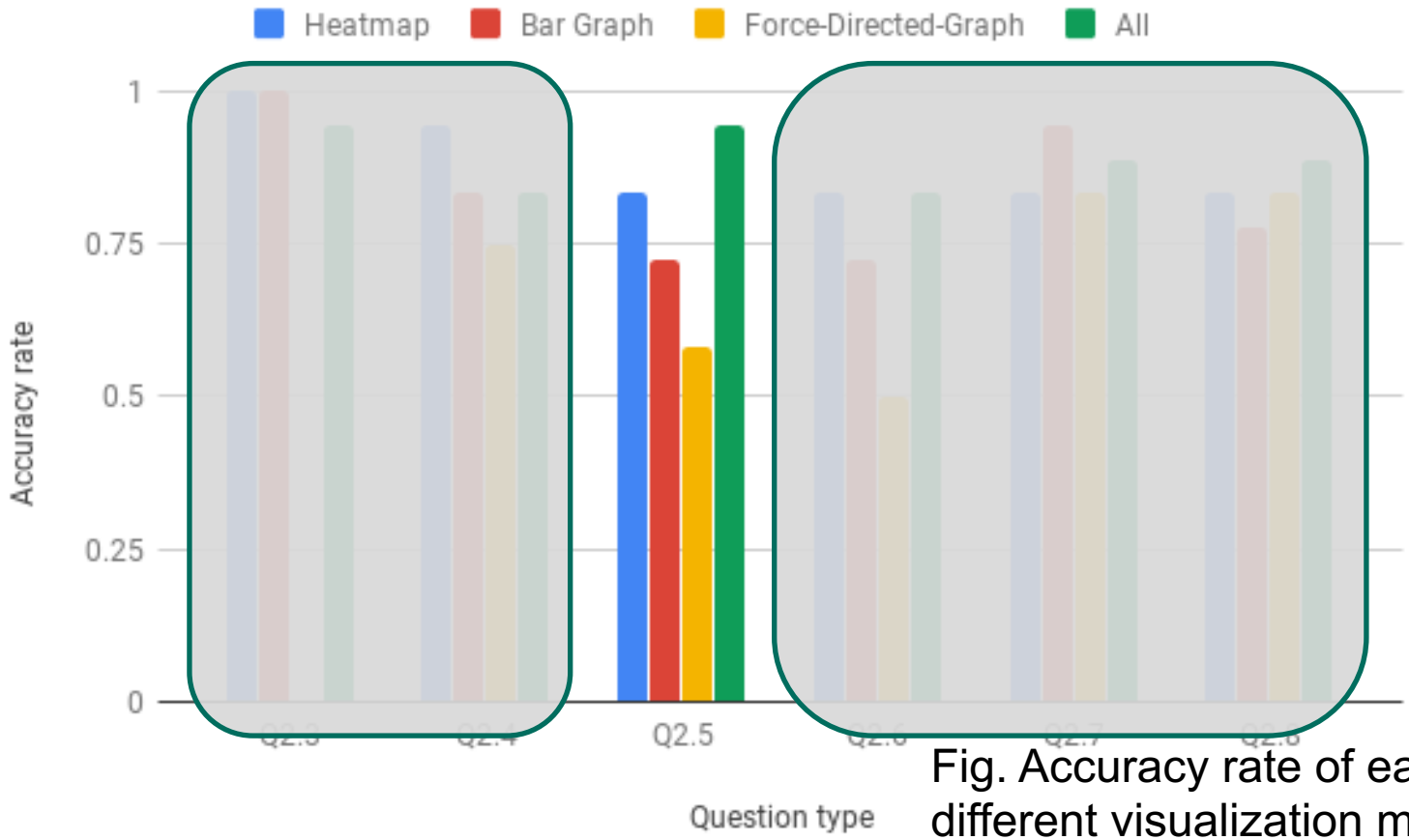


Fig. Accuracy rate of each question type using different visualization methods

Evaluation

Q2.6: The following statement is true or false: “The correlation value between *Attribute A* and *Attribute B* remains the same at *Timestamp T1* and at *Timestamp T2*”?

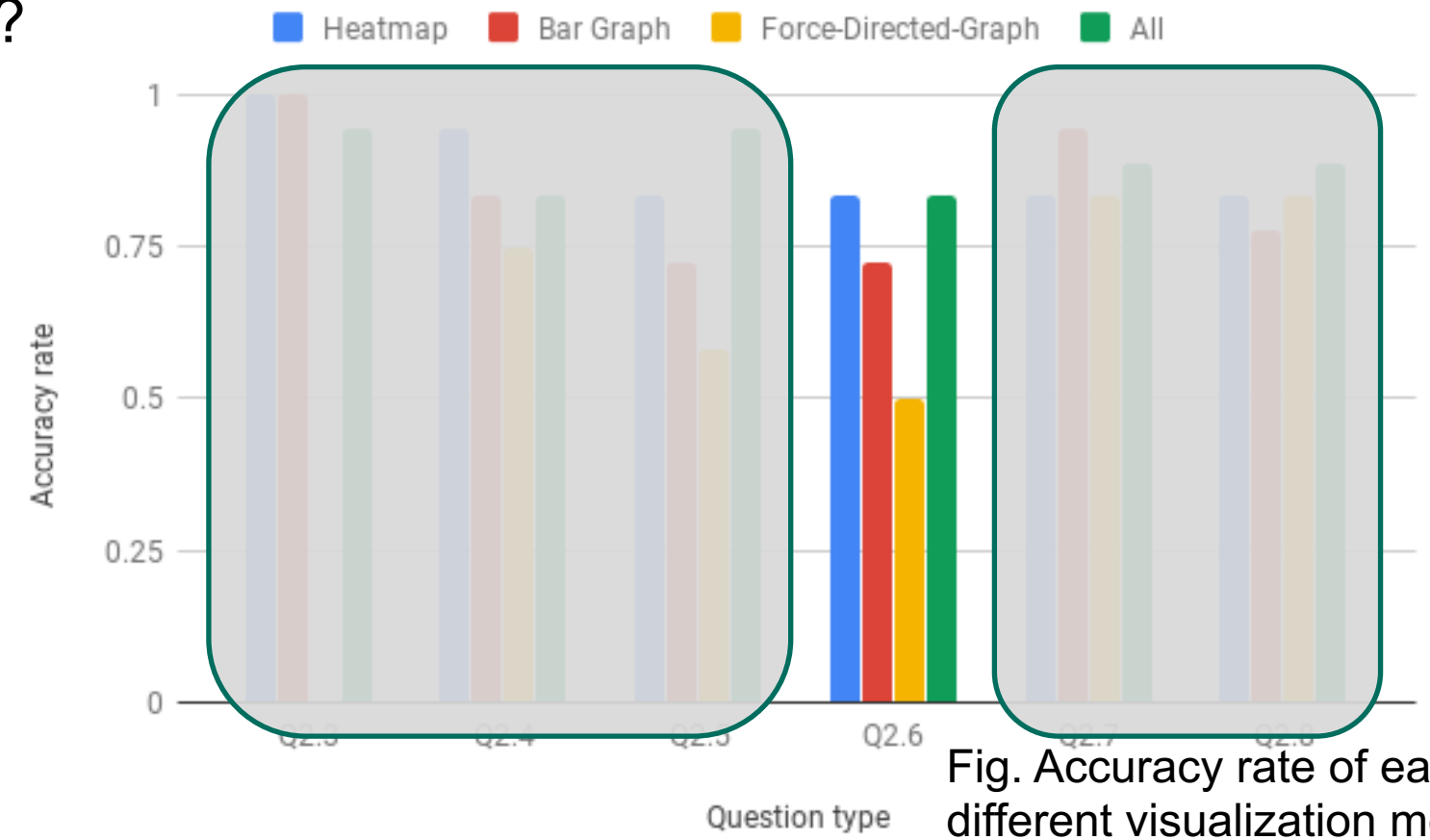


Fig. Accuracy rate of each question type using different visualization methods

Evaluation

Q2.7: Which pair(s) of attributes has/have a correlation value that is not smaller than *X* at *Timestamp T*?

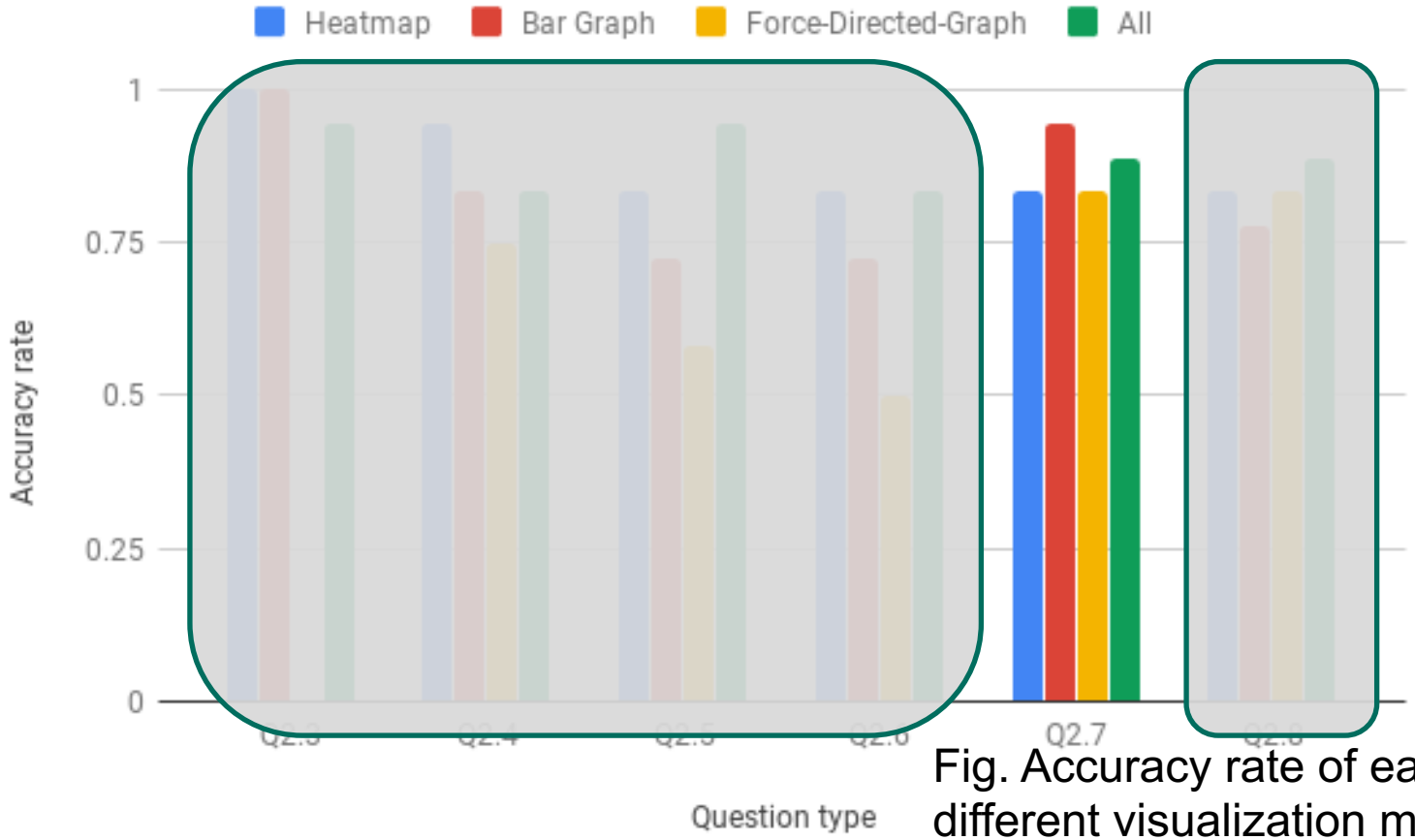


Fig. Accuracy rate of each question type using different visualization methods

Evaluation

Q2.8: Which pair(s) of attributes has/have a correlation value that is not bigger than *X* at *Timestamp T*?

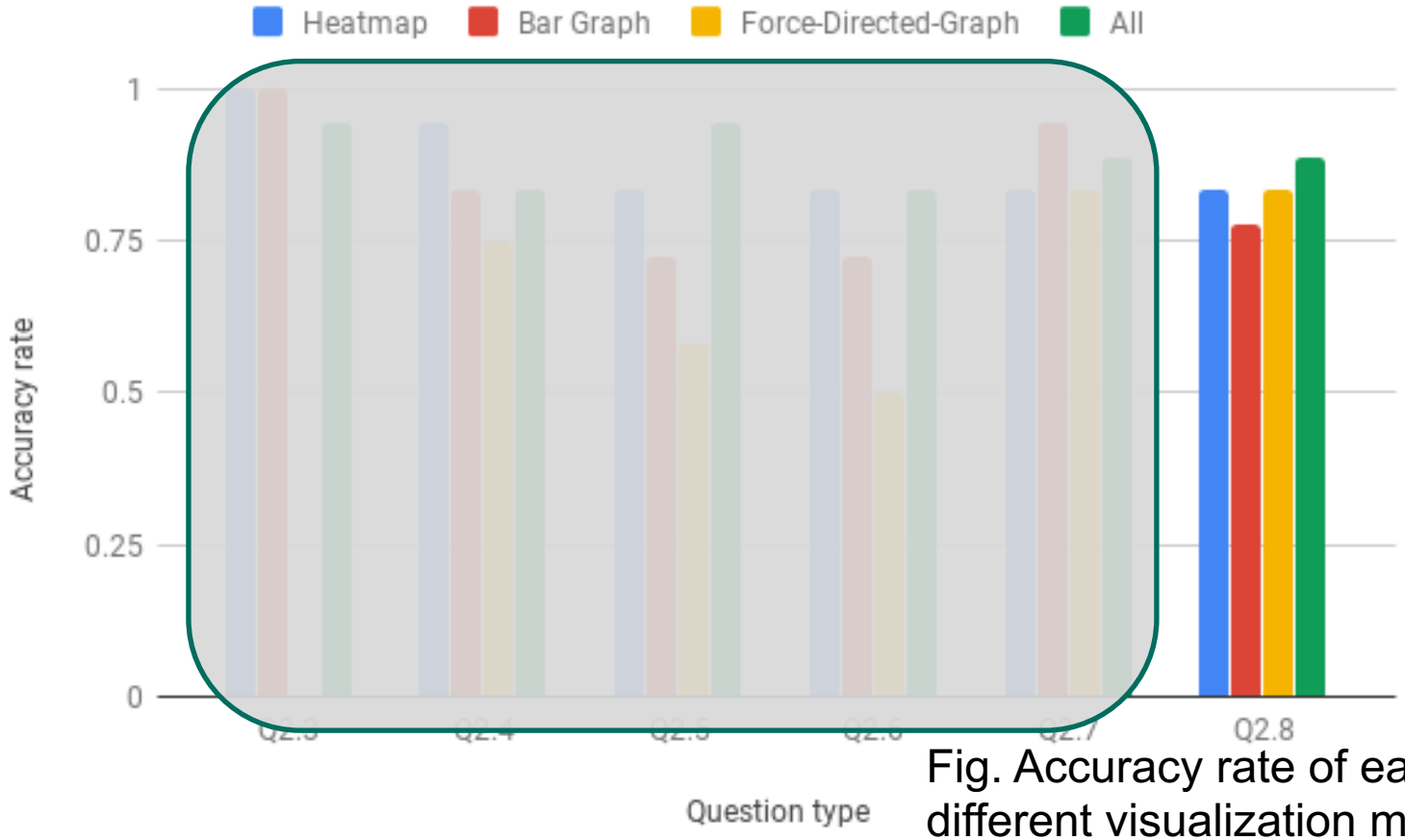


Fig. Accuracy rate of each question type using different visualization methods

Evaluation

Accuracy rate by using Data Set 1 is always the biggest and reaches over 75%

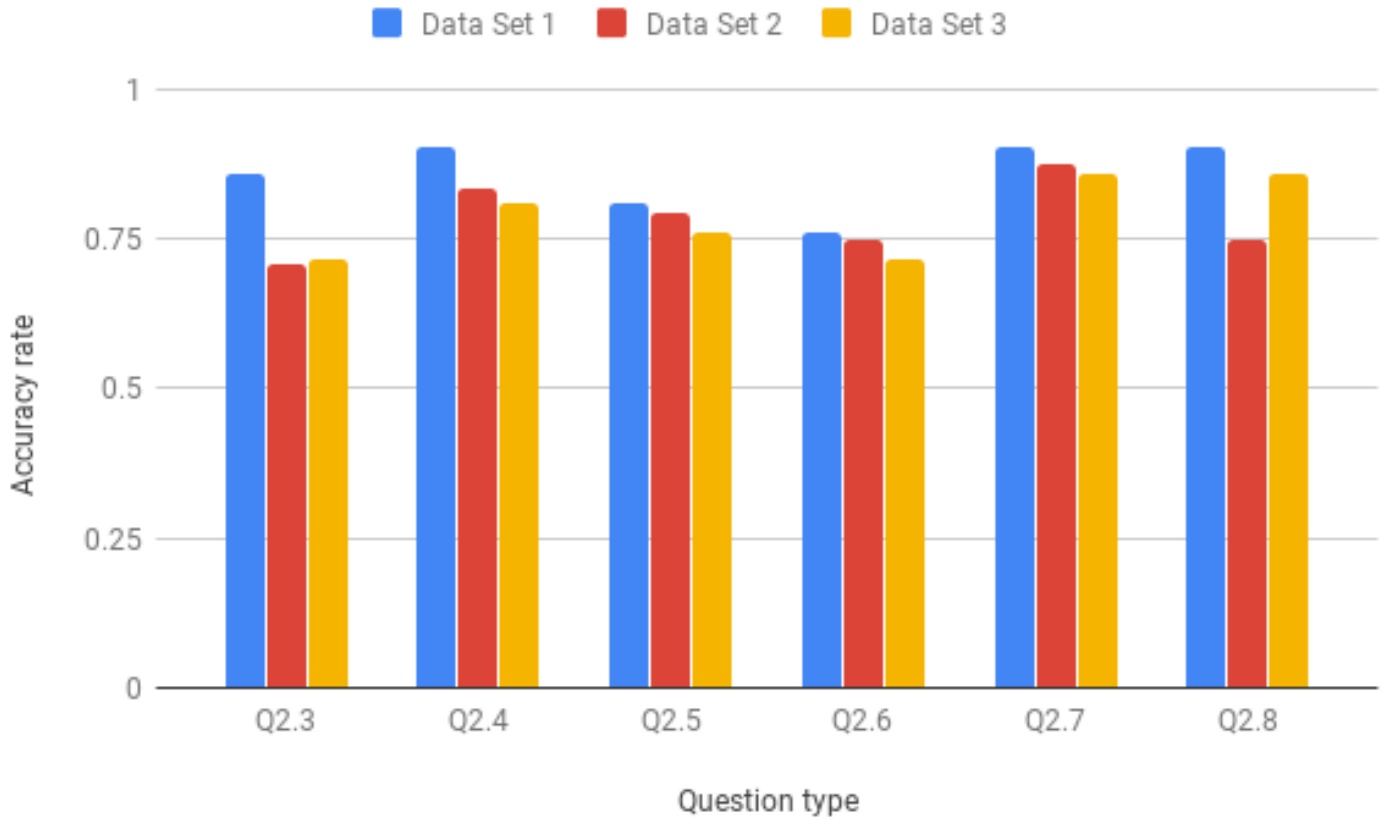
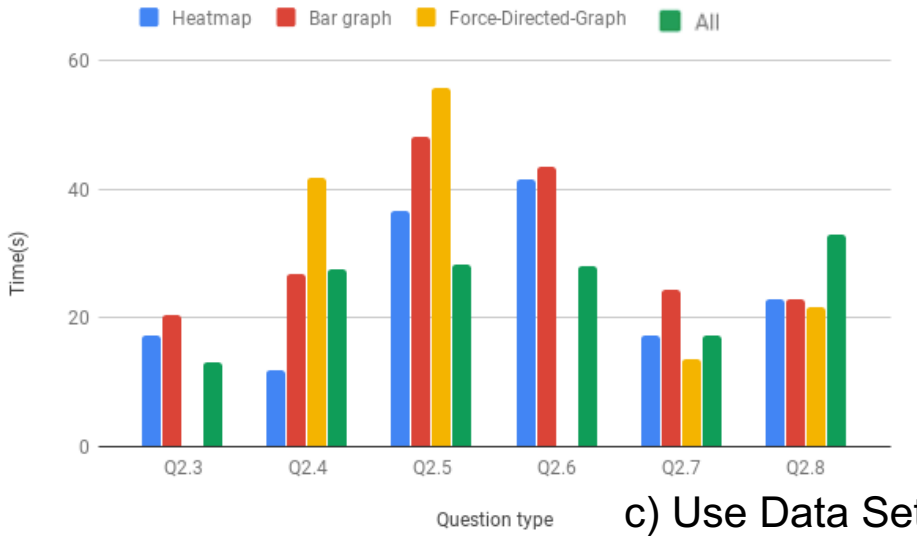


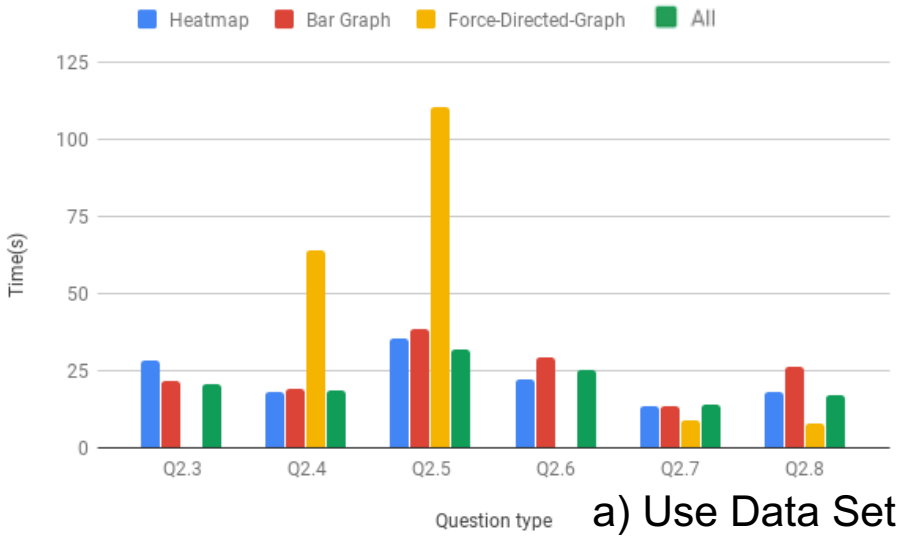
Fig. Accuracy rate of each question type using different data sets

Evaluation

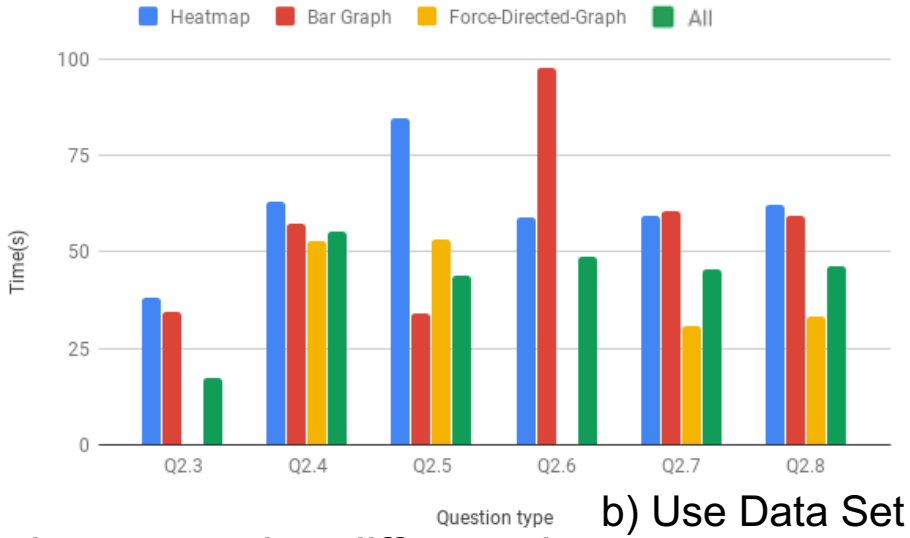
- Both the average time of using Heatmap and Bar Graph is close in most cases, whatever data set the participants use
- Q2.3 & Q2.6 cannot be answered by using Force-Directed-Graph



c) Use Data Set 3



a) Use Data Set 1

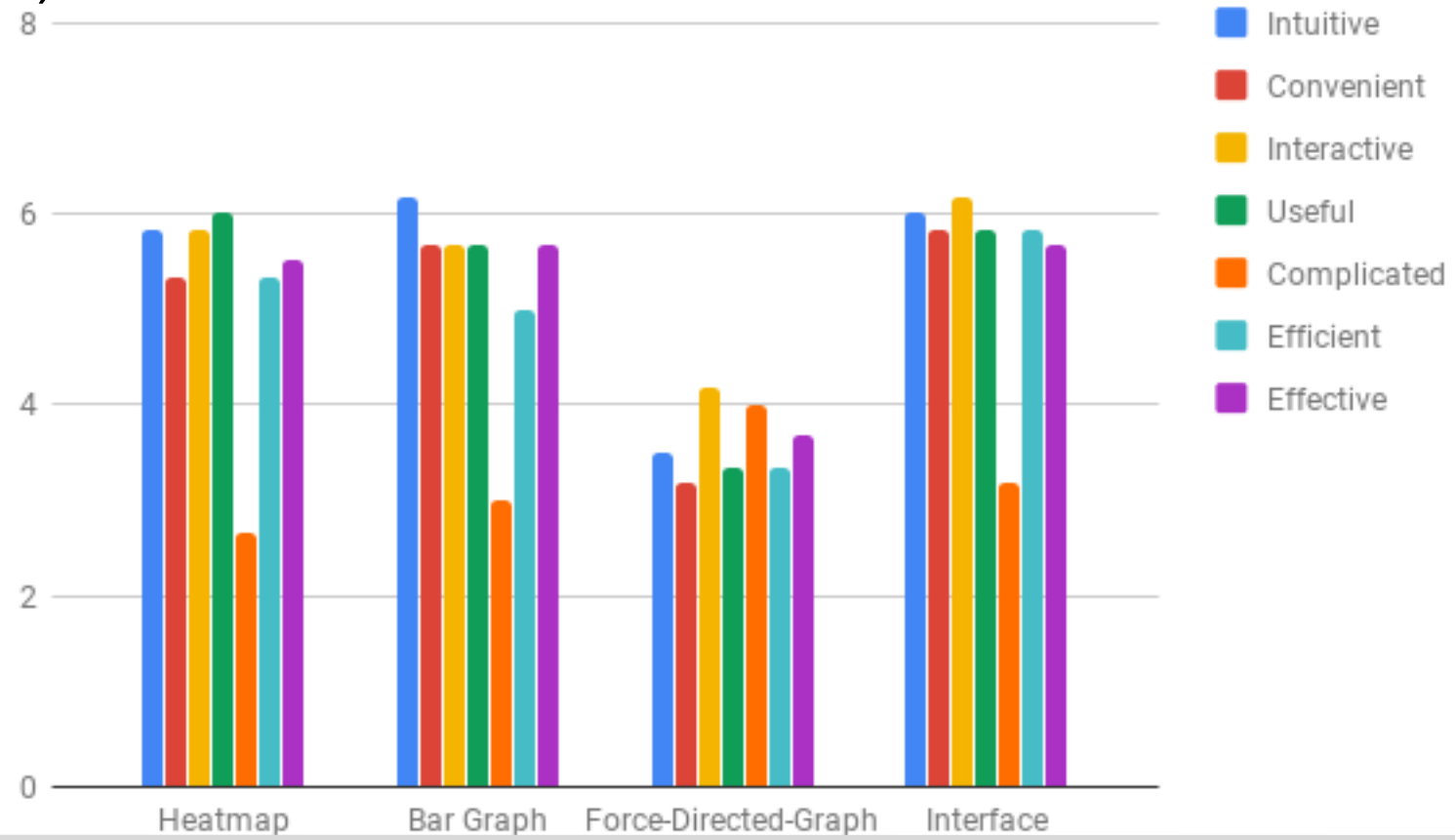


b) Use Data Set 2

Fig: Average time of participants finishing each question type using different data sets

Evaluation – Feedback

- Please write down the visualization method you have used and rate for it:
  - Using LIKERT scale (1-7)
  - 1: unfamiliar
  - 7: very familiar



Evaluation – Verbatim

Hard to read exact value

Only show related links when point to one node

Having different graphs for different uses

Easy to see the smallest and biggest

Convient and fast

Intuitively show the correlation of data, which is effective for further data processing. In addition, it meets many demands for data processing.

More convient functions/buttons should be added

Reflect a tendency of the data correlation

Summary

- Correlation analysis is one of the fundamental task of Data Mining
- Streaming setting and high-dimensionality are challenges of data visualization
- Interface using three visualization methods for visualizing correlation in high-dimensional streams
- Using Force-Directed-Graph is difficult to get the exact/probable value
- Heatmap is intuitive for finding biggest/smallest value correctly

References

- [1] Albuqerqe, G., Eisemann, M., Lehmann, D. J., Theisel, H., and Magnor, M. Quality-Based Visualization Matrices. In Proceedings of the Vision, Modeling and Visualization (2009), 341–350.
- [2] B,L.K.,Riekenbrauck,N.,Thevessen,D.,Pappik,M.,Stebner,A.,Kunze,J.,Meiss- ner, A., Shekar, A. K., and Emmanuel, M. Machine Learning and Knowledge Dis- covery in Databases. 404–408.
- [3] Filis, G., Degiannakis, S., and Floros, C. Dynamic correlation between stock mar- ket and oil prices: The case of oil-importing and oil-exporting countries. International Review of Financial Analysis 20, 3 (2011), 152 – 164.
- [4] Kelley, W. M.; Donnelly, R. A. The humongous book of statistics problems. New York (2009).
- [5] Kobourov, S. G. Spring embedders and force-directed graph drawing algorithms. eprint arXiv:1201.3011 (2012).
- [6] Murray, S. Interactive data visualization for the web, an introduction to designing with d3. 272.
- [7] Rodgers, J. L.; Nicewander, W. A. Thirteen ways to look at the correlation coefficient. The American Statistician (1988).
- [8] Simons, K., et al. Should us investors invest overseas? New England Economic Review (1999), 29–40.
- [9] Wilkinson,Leland;Friendly,M.Thehistoryoftheclusterheatmap.TheAmerican Statistician (2009).