Contrasting Diverse, Probabilistic, and Visualization-Aware Data Selection Methods for Visual Analytics

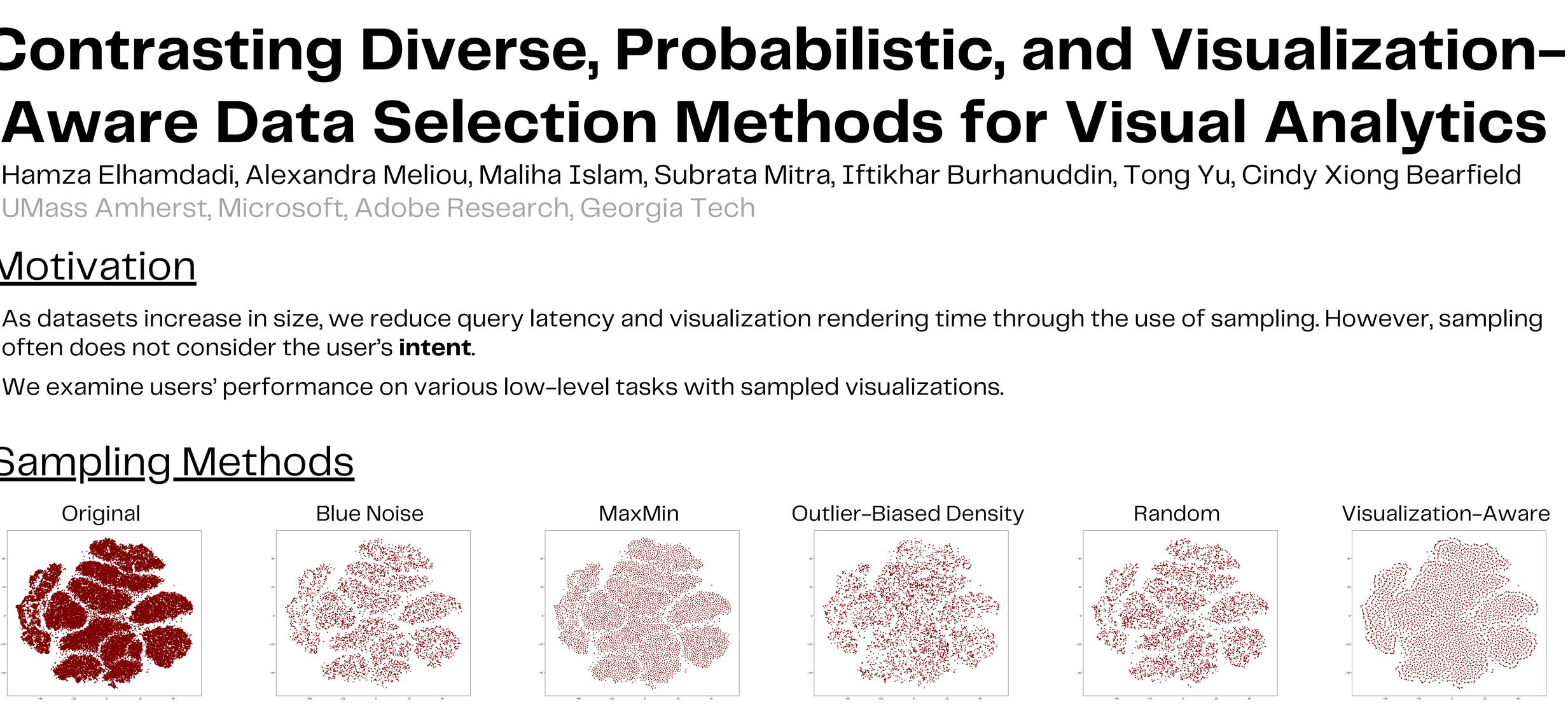
Hamza Elhamdadi, Alexandra Meliou, Maliha Islam, Subrata Mitra, Iftikhar Burhanuddin, Tong Yu, Cindy Xiong Bearfield
UMass Amherst, Microsoft, Adobe Research, Georgia Tech

Motivation

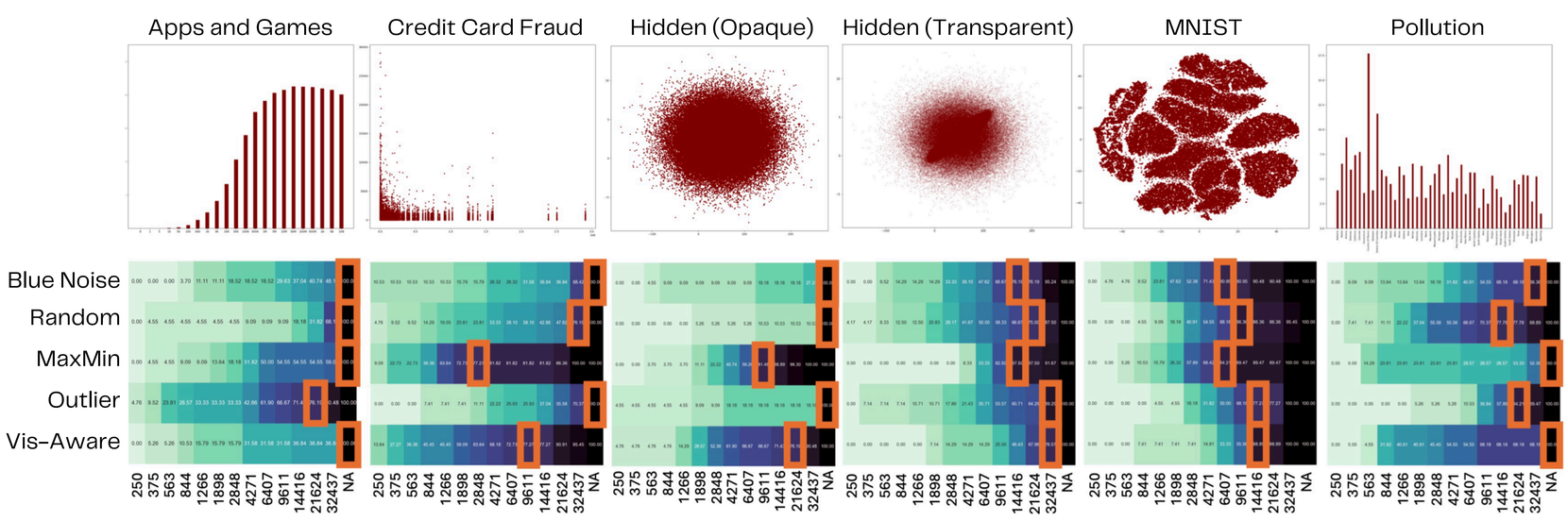
As datasets increase in size, we reduce query latency and visualization rendering time through the use of sampling. However, sampling often does not consider the user's **intent**.

We examine users' performance on various low-level tasks with sampled visualizations.

Sampling Methods



We consider five sampling methods: two diverse (blue noise, maxmin), two probabilistic (outlier, random) and one visualization-aware



Experiment 1 (Subjective Feature Capture)

For each sampling method, we asked participants to select the smallest sample size that captures important subjective features of five datasets.

Experiment 1 Results

For datasets visualized as scatter plots, MaxMin captured these features at much smaller sample sizes than other sampling methods.

For datasets visualized using bar charts, all sampling methods required relatively large sample sizes.

Experiment 2 (Task Performance)

Participants completed ten low-level tasks with sampled visualizations. Each task is depicted with the corresponding dataset used for that task.

A subset of participants (control) completed the tasks using the original dataset. A vertical line (and grey-shaded region) is used to indicate the mean and standard deviation of the error/percentage.

For the **find anomalies**, **characterize distribution**, **filter**, and **order** tasks, a **higher value** indicates better performance. For all **other tasks**, a **lower value** indicates better performance.

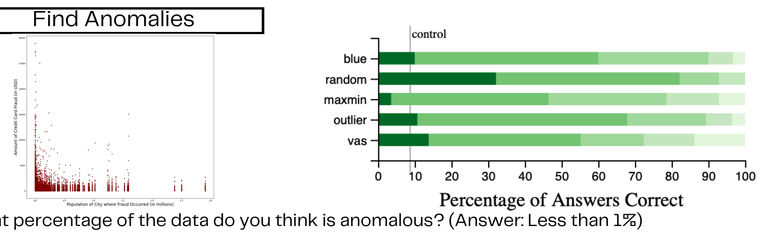
Experiment 2 Results

For most tasks, participant performance using visualizations sampled by **maxmin** and **visualization-aware** sampling was similar to the control (excluding **retrieve value** for both methods, and **find clusters**, **compute derived value**, and **determine range** for visualization-aware).

Performance similar to the control indicates that the sampling method is perceptually similar to the original dataset's visualization. This is beneficial when the original visualization is likely to produce accurate responses (e.g., **maxmin and visualization-aware sampling perform well for the 'find extremum', 'filter', and 'order' task**).

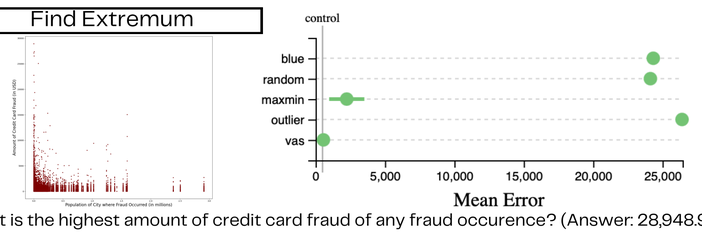
However, for some tasks, perceptual similarity to original dataset causes participants to perform poorly (**e.g., maxmin and visualization-aware perform poorly for the 'find correlation' task**).

Find Anomalies



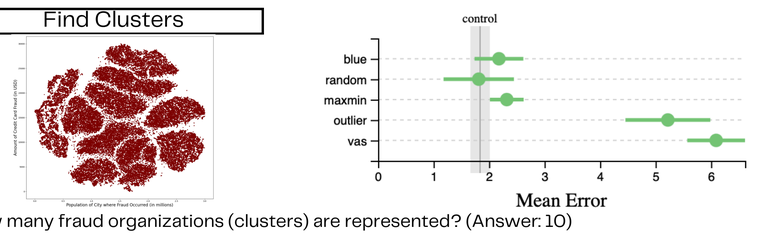
What percentage of the data do you think is anomalous? (Answer: Less than 1%)

Find Extremum



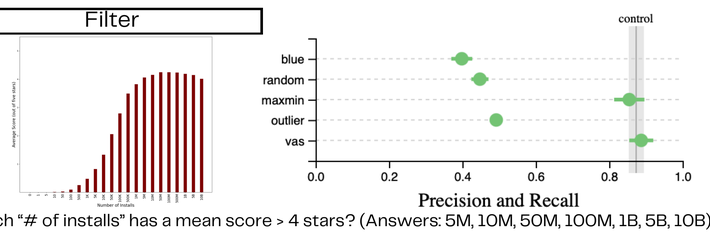
What is the highest amount of credit card fraud of any fraud occurence? (Answer: 28,948.9)

Find Clusters



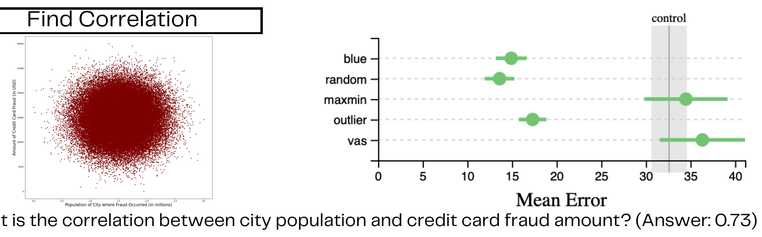
How many fraud organizations (clusters) are represented? (Answer: 10)

Filter



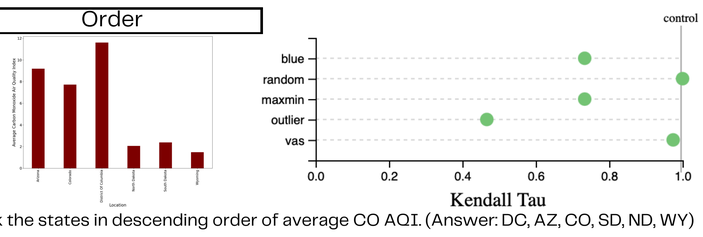
Which "# of installs" has a mean score > 4 stars? (Answers: 5M, 10M, 50M, 100M, 1B, 5B, 10B)

Find Correlation



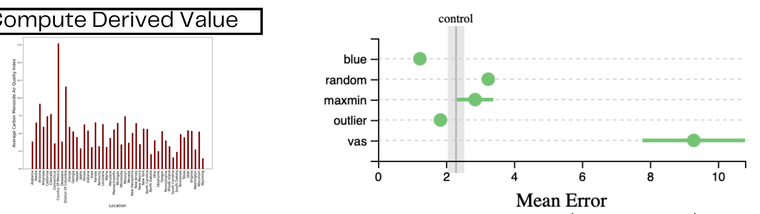
What is the correlation between city population and credit card fraud amount? (Answer: 0.73)

Order



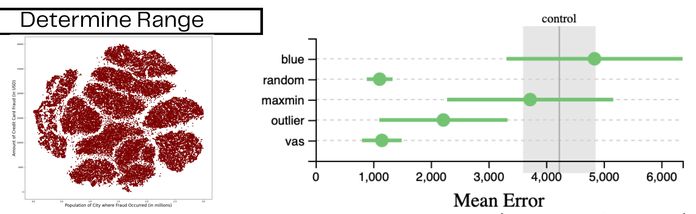
Rank the states in descending order of average CO AQI. (Answer: DC, AZ, CO, SD, ND, WY)

Compute Derived Value



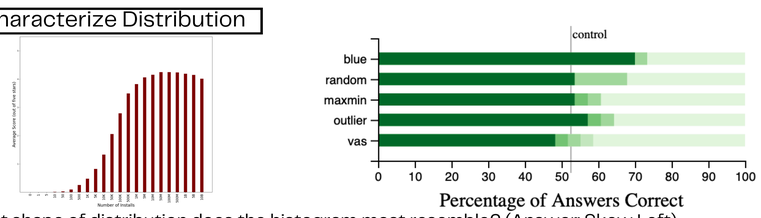
What is the average CO Air Quality Index in the entire U.S. and Mexico? (Answer: 5.997)

Determine Range



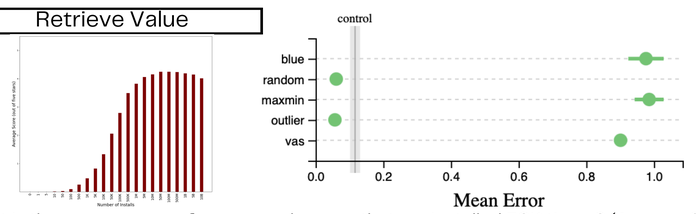
For cities with < 0.5m people, what is the range of fraud amount? (Answer: 6,969-23,631)

Characterize Distribution



What shape of distribution does the histogram most resemble? (Answer: Skew Left)

Retrieve Value



What is the average score for apps and games that are installed 50K times? (Answer: 2.06)

references

[1] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In IEEE Symposium on Information Visualization, 2005. INFOVIS 2005., pp. 111–117. IEEE, 2005. 1, 2
[2] C¸. Demiralp, P. J. Haas, S. Parthasarathy, and T. Pedapati. Foresight: Rapid data exploration through guideposts. arXiv preprint arXiv:1709.10513, 2017. 1
[3] M. Drosou, H. Jagadish, E. Pitoura, and J. Stoyanovich. Diversity in big data: A review. Big data, 5(2):73–84, 2017. 1
[4] Y. Park, M. Cafarella, and B. Mozafari. Visualization-aware sampling for very large databases. In 2016 IEEE 32nd International Conference on Data Engineering (ICDE), pp. 755–766. IEEE, 2016. 1
[5] R. D. Portugal and B. F. Svaiter. Weber-fechner law and the optimality of the logarithmic scale. Minds and Machines, 21:73–81, 2011. 1
[6] S. Xiang, X. Ye, J. Xia, J. Wu, Y. Chen, and S. Liu. Interactive correction of mislabeled training data. In 2019 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 57–68. IEEE, 2019. 1
[7] D.-M. Yan, J.-W. Guo, B. Wang, X.-P. Zhang, and P. Wonka. A survey of blue-noise sampling and its applications. Journal of Computer Science and Technology, 30(3):439–452, 2015. 1
[8] J. Yuan, S. Xiang, J. Xia, L. Yu, and S. Liu. Evaluation of sampling methods for scatterplots. IEEE Transactions on Visualization and Computer Graphics, 27(2):1720–1730, 2020. 2

Hamza Elhamdadi
Email: helhamdadi@umass.edu
Website: https://hamza-elhamdadi.github.io/

On the Job Market!

Contact Me!