We would like to thank all of the reviewers for giving detailed and helpful feedback. We will incorporate the suggestions in the revised paper. We are glad that all the reviewers found the problem of data deletion interesting and important to study. In fact, just a few weeks ago, the Dashboard Act proposed in the U.S. Senate stipulates that large consumer companies need to enable users to delete their data, making this a timely and significant challenge to the ML community.

Response to Reviewer 1: Thank you for your suggestions on making our exposition more accessible. In the revision, we will add a concrete example with a figure to illustrate the steps of the quantized k-means algorithm. The Appendix will contain additional background information on clustering to make the paper more self-contained. We will also integrate the four design principles of efficient deletion more tightly into the rest of the paper. In particular, we will discuss how the quantized k-means and divide-and-conquer k-means implement these principles. These edits will make the paper more accessible to the broad ML community. Thanks also for pointing out the typos; we will correct these.

Response to Reviewer 3: Thank you for your very helpful suggestions. We focus on clustering because it's widely-used in applied ML, including on the UK Biobank dataset [1], and it's a good illustration of our deletion definition and design principles. We will add more real-world context and references for this point. For example, the cited work by Galinsky et al. makes uses of k-means clustering: "In this study, analyses of 113,851 UK Biobank samples showed that population structure in the UK is dominated by five principal components (PCs) spanning six clusters: Northern Ireland, Scotland, northern England, southern England, and two Welsh clusters."

You bring up an excellent point concerning the conceptual differences between deleting a single data point and deleting all the data pertaining to a single user. As you have pointed out, in production software stacks, such as at large internet companies, user data may exist in many databases, in many models, and in many relational formats (even between users and platforms). In production systems, it may indeed be the case that a deletion request from a single user's data may require the deletion of multiple points from multiple databases and multiple models in the system/datacenter. This is a promising direction for future research in data management for AI systems. We have clarified this distinction in our manuscript, as well as highlighted this as another topic for future work. Furthermore, even determining what pieces of data should get deleted in response to a user request is an interesting question.

Furthermore, thanks for your suggestion that we might compare the runtime of our algorithms to the standard k-means, which runs in time $O(nkTd)$. Both of our proposed methods run in fractional power time in n (i.e. $\sqrt{n}$). We will state this comparison explicitly in our paper.

Specific Response to Reviewer 4: Thank you for your constructive feedback. We will add a centralized section on related works in the main text; currently the discussion of relevant works are dispersed in a few places (e.g. in the discussion of the distinction from differential privacy). We agree that supervised learning is an important direction of future work for data deletion. We had some brief comments on efficient deletion for linear regression in the original submission. We will flesh this out as an illustration of how efficient deletion could work in supervised setting.

References

[1] Kevin J Galinsky, Po-Ru Loh, Swapan Mallick, Nick J Patterson, and Alkes L Price. Population structure of uk biobank and ancient eurasians reveals adaptation at genes influencing blood pressure. *The American Journal of Human Genetics*, 99(5):1130–1139, 2016.