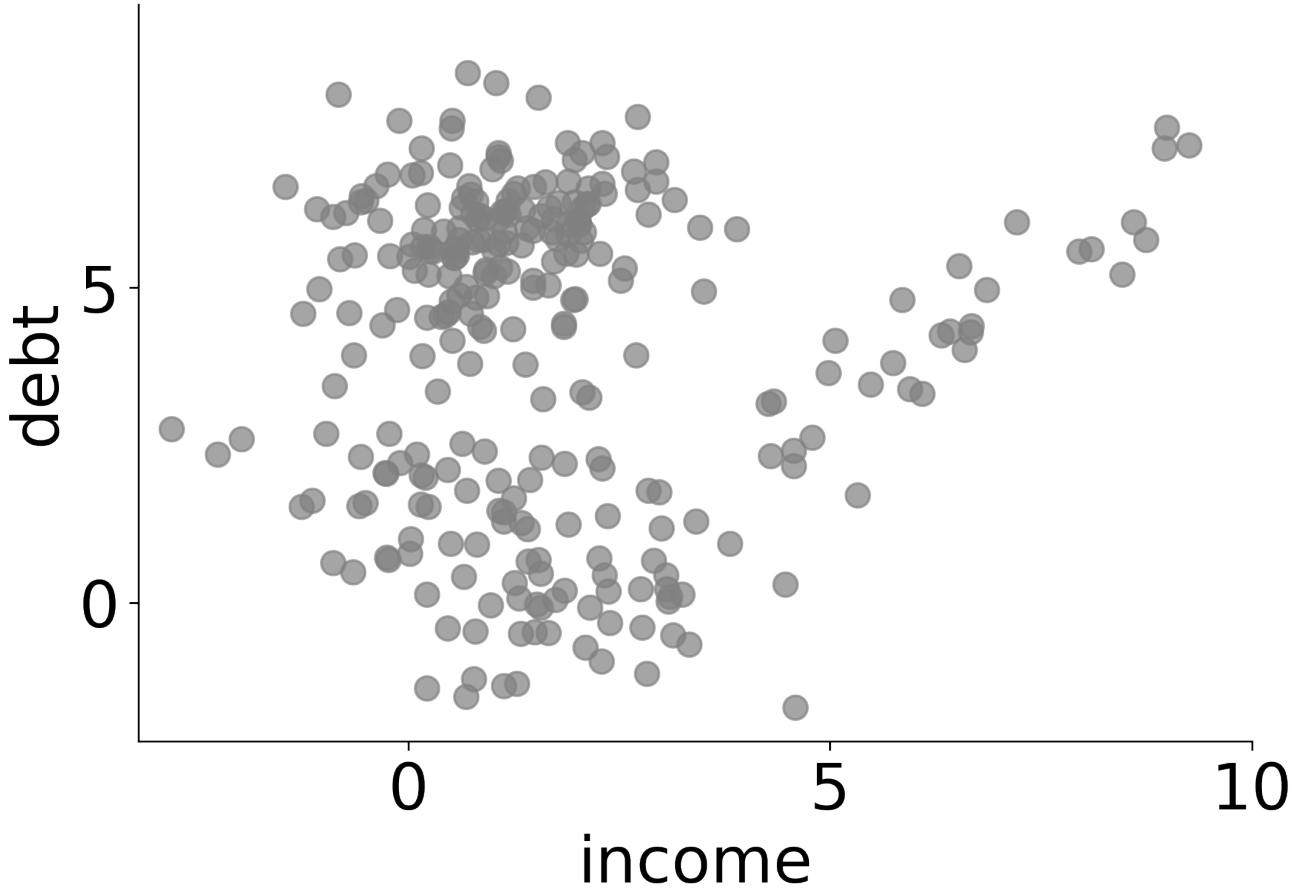
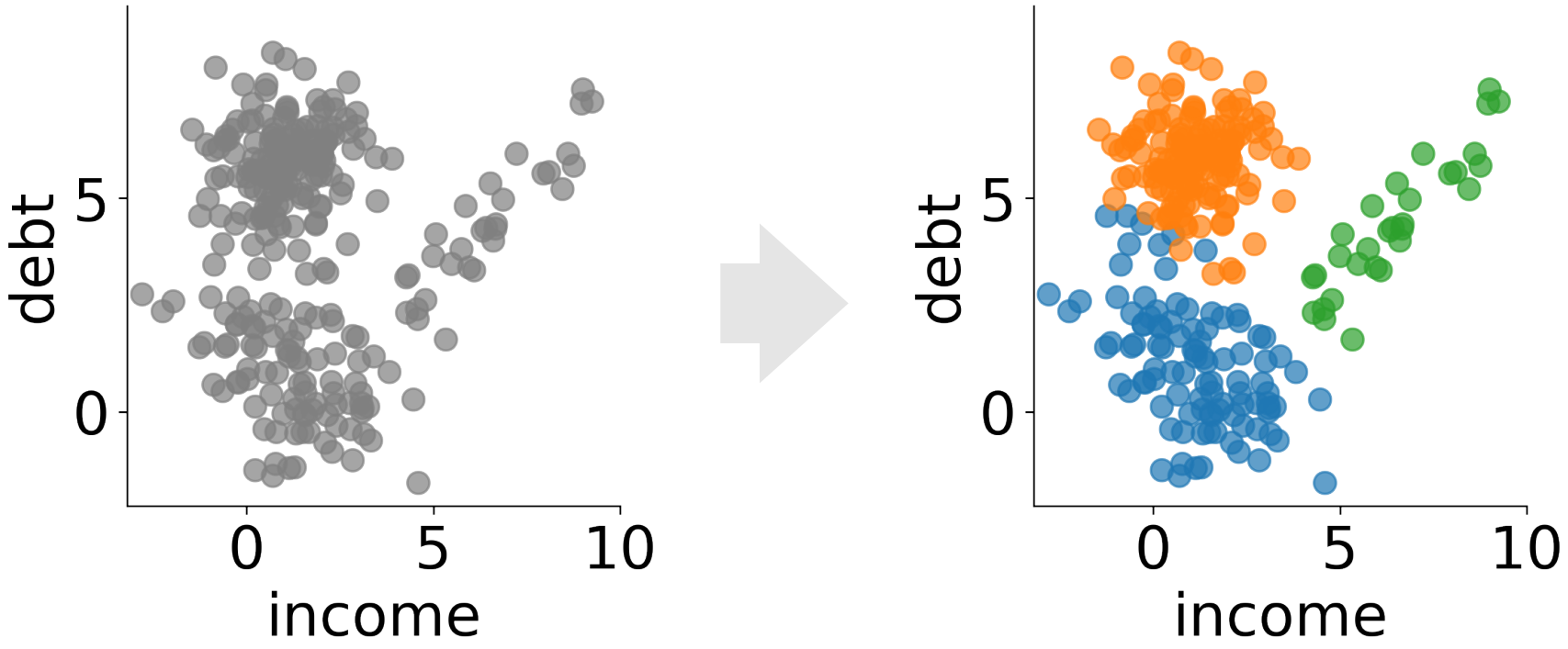
Probabilistic clustering

Clustering

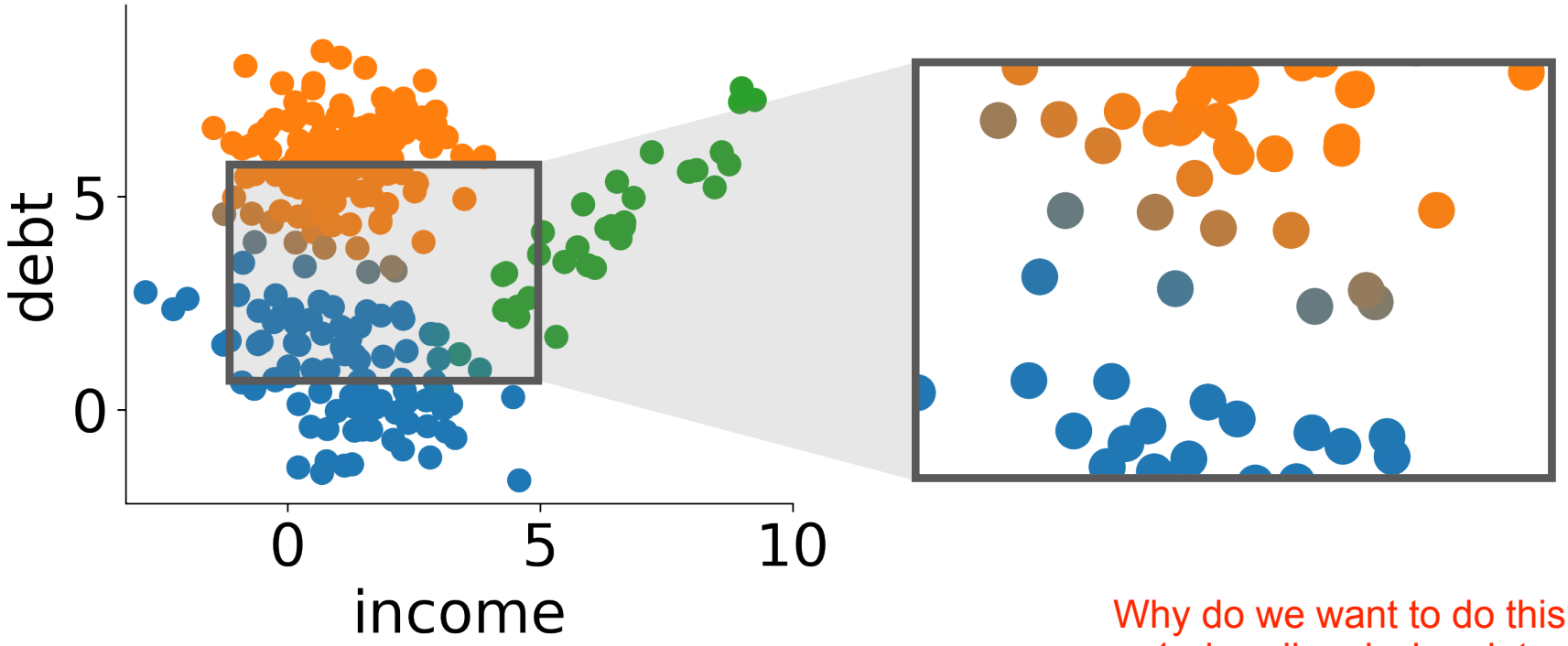


Hard clustering



So hard clustering is when we 'force' it onto them.
For each data point, we assign to it a colour

Soft clustering



Soft clustering is
we assign probability
distributions over the cluster,
instead of assigning each point
a cluster idx.

Probabilistic

$$p(\text{cluster idx} \mid x)$$

instead of

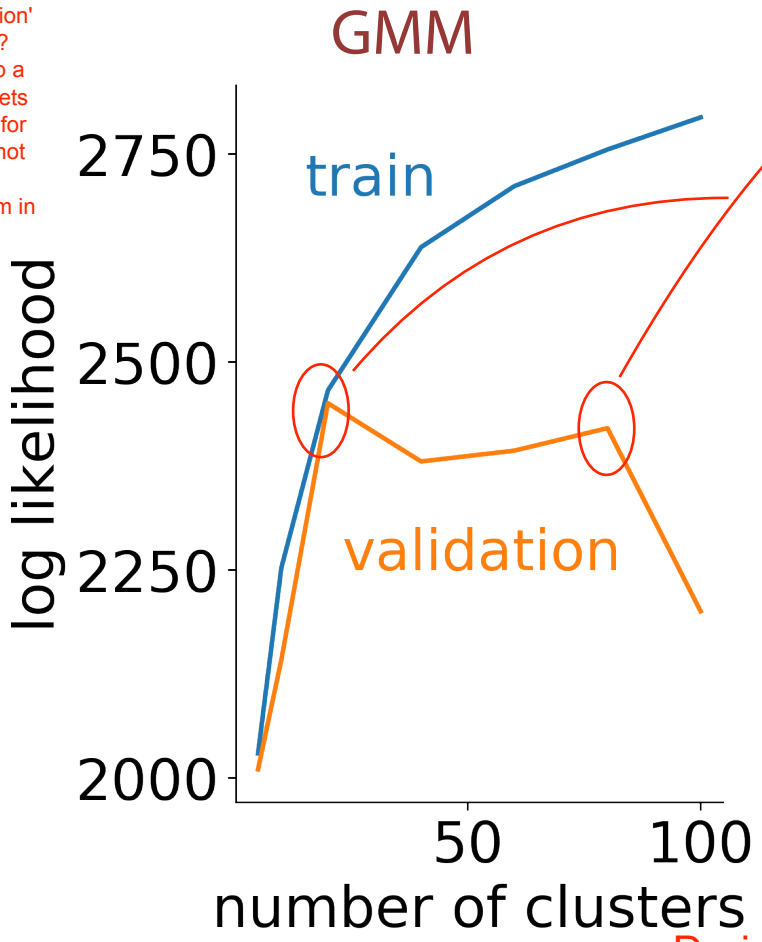
$$\text{cluster idx} = f(x)$$

Why do we want to do this?
1. handle missing data naturally
2. tune hyperparameters.

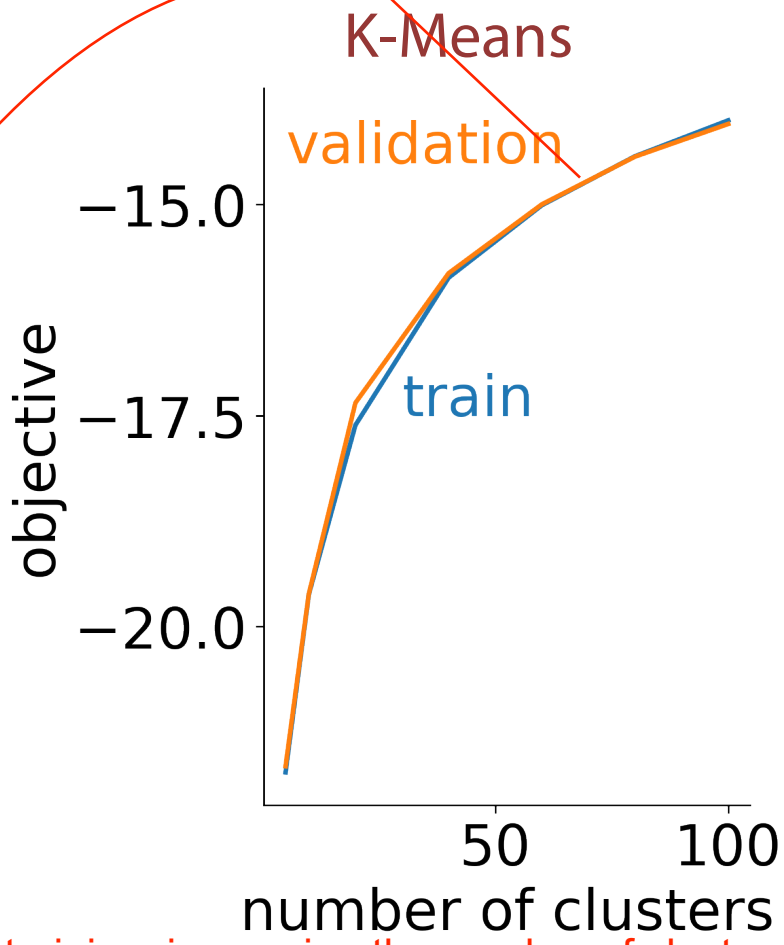
Hyperparameter tuning

We compare how hyperparameter tuning is different for soft (GMM) and hard (K-means) clustering.

QUESTION: How to measure 'validation' log likelihood ? I guess we keep a sample of datasets to be used only for validation, and not to train. Then we put them in to get the validationlog likelihood.



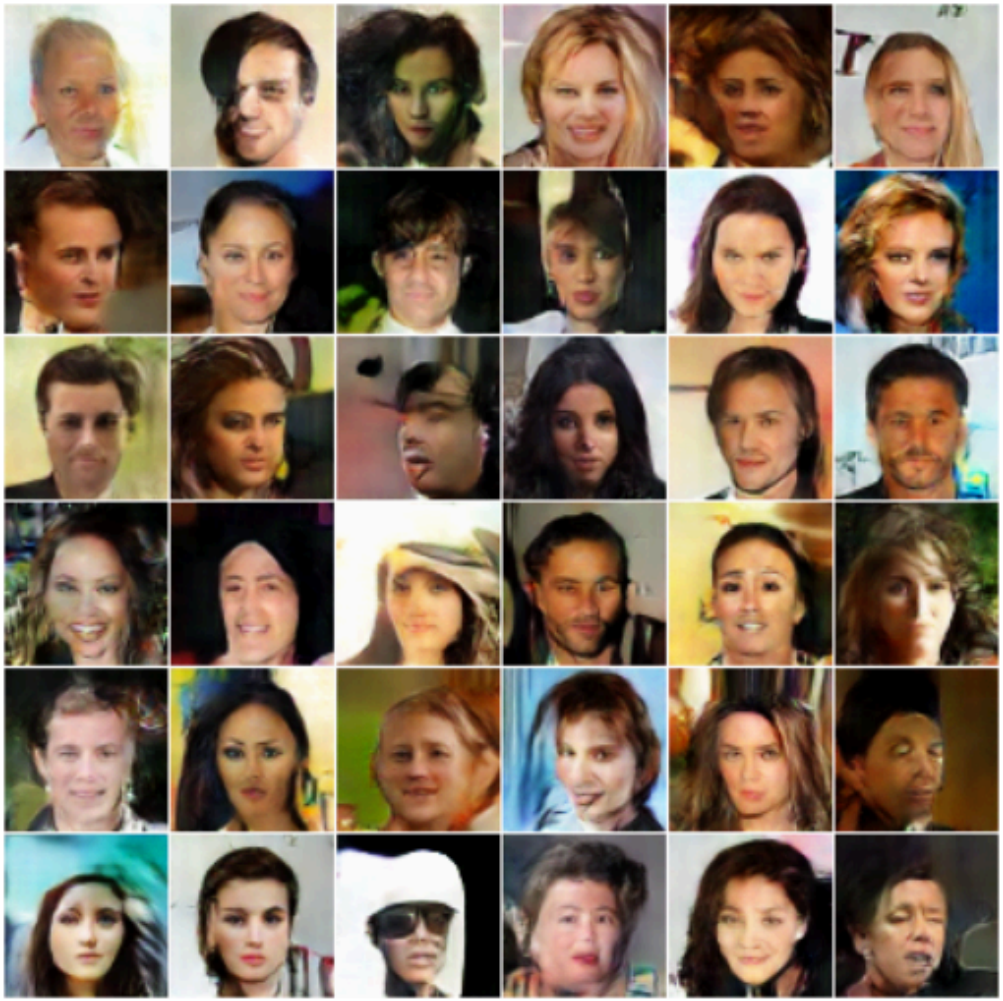
For K-means, there isn't a way to understand validation loss meaningfully like GMM.



So we can vary the number of clusters there are.
Training performance is the log likelihood. The higher, the better.

During training, increasing the number of clusters will naturally make log likelihood higher. This is because eventually, we will reach a point where every data point has its own 'cluster', thus being very exact. Of course, this isn't meaningful and generalizing.

Generating new data points



Junbo Zhao, https://arxiv.org/pdf/1609.03126.pdf

Summary

Want to cluster data in a soft way

- Allows to tune hyper parameters
- Generative model of the data