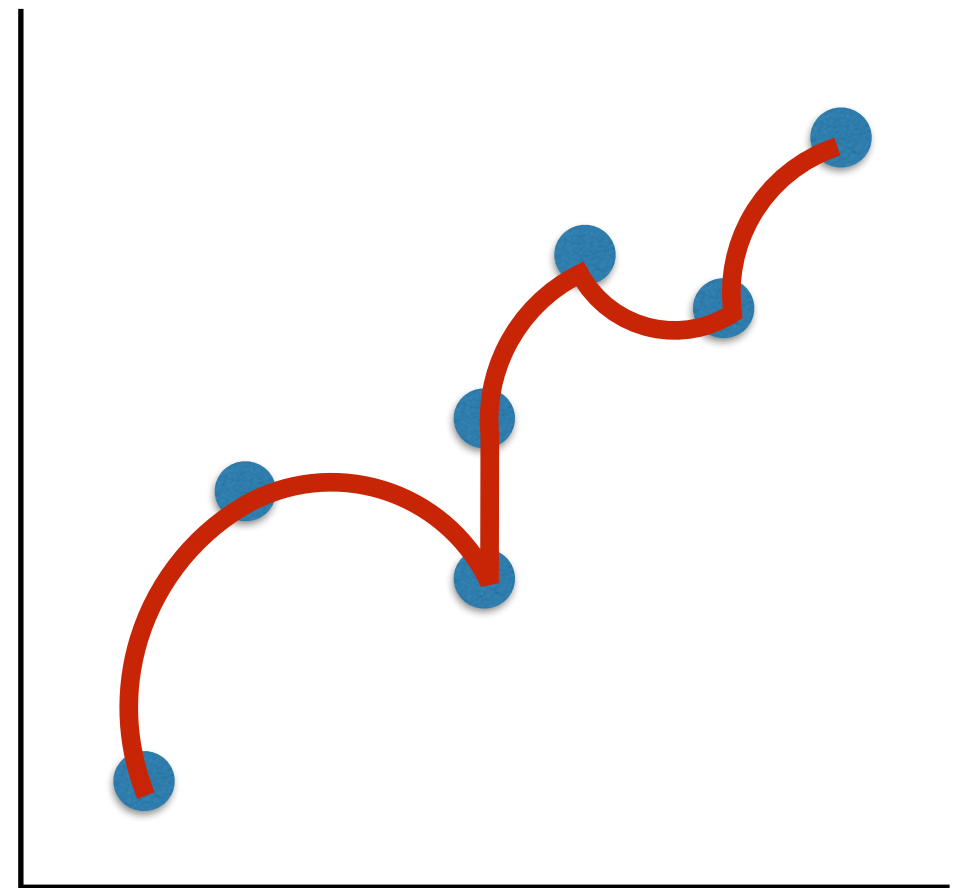
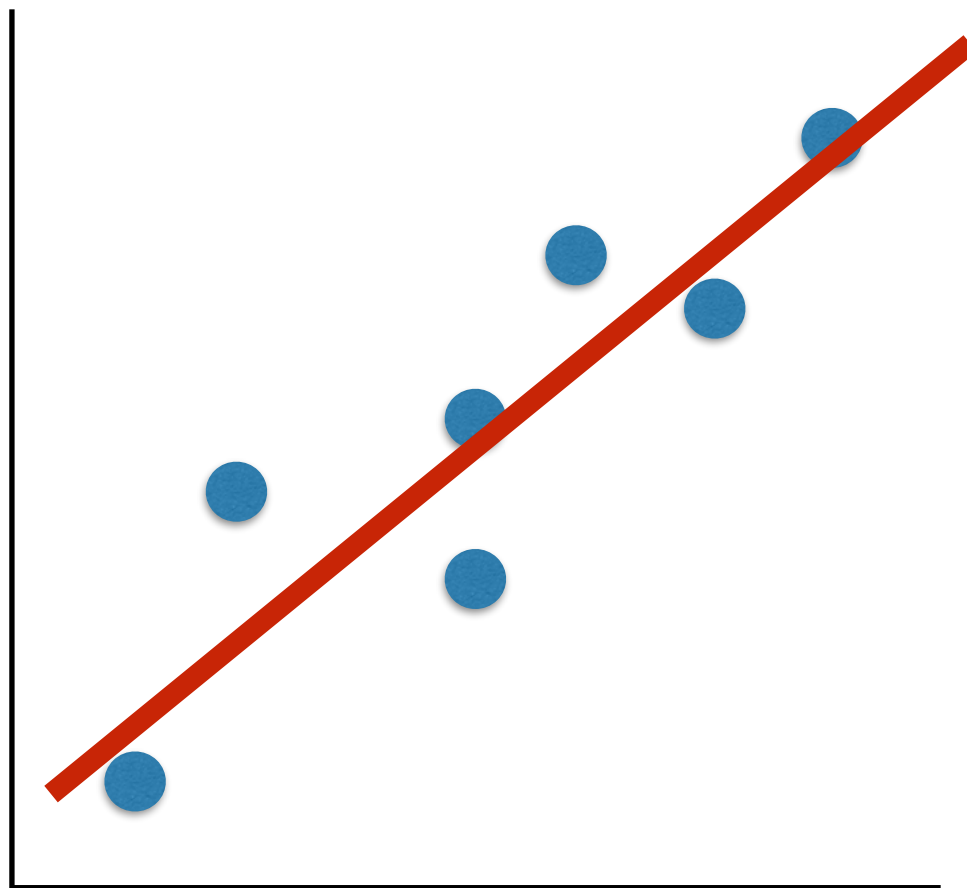


HUDK 4050: CORE METHODS IN EDM

Assignment 3

- Purpose of clustering
- Part II - missing data issues
- Part III - mosaic plots



Which is more “accurate”?

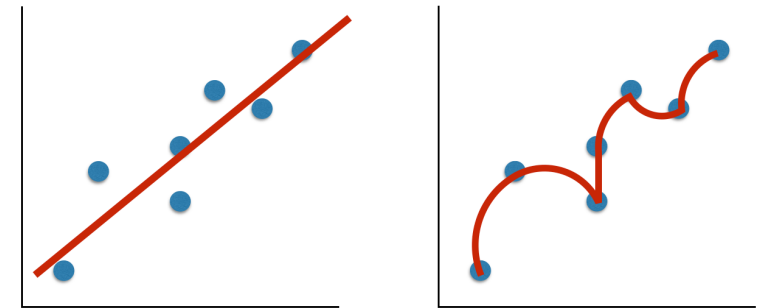
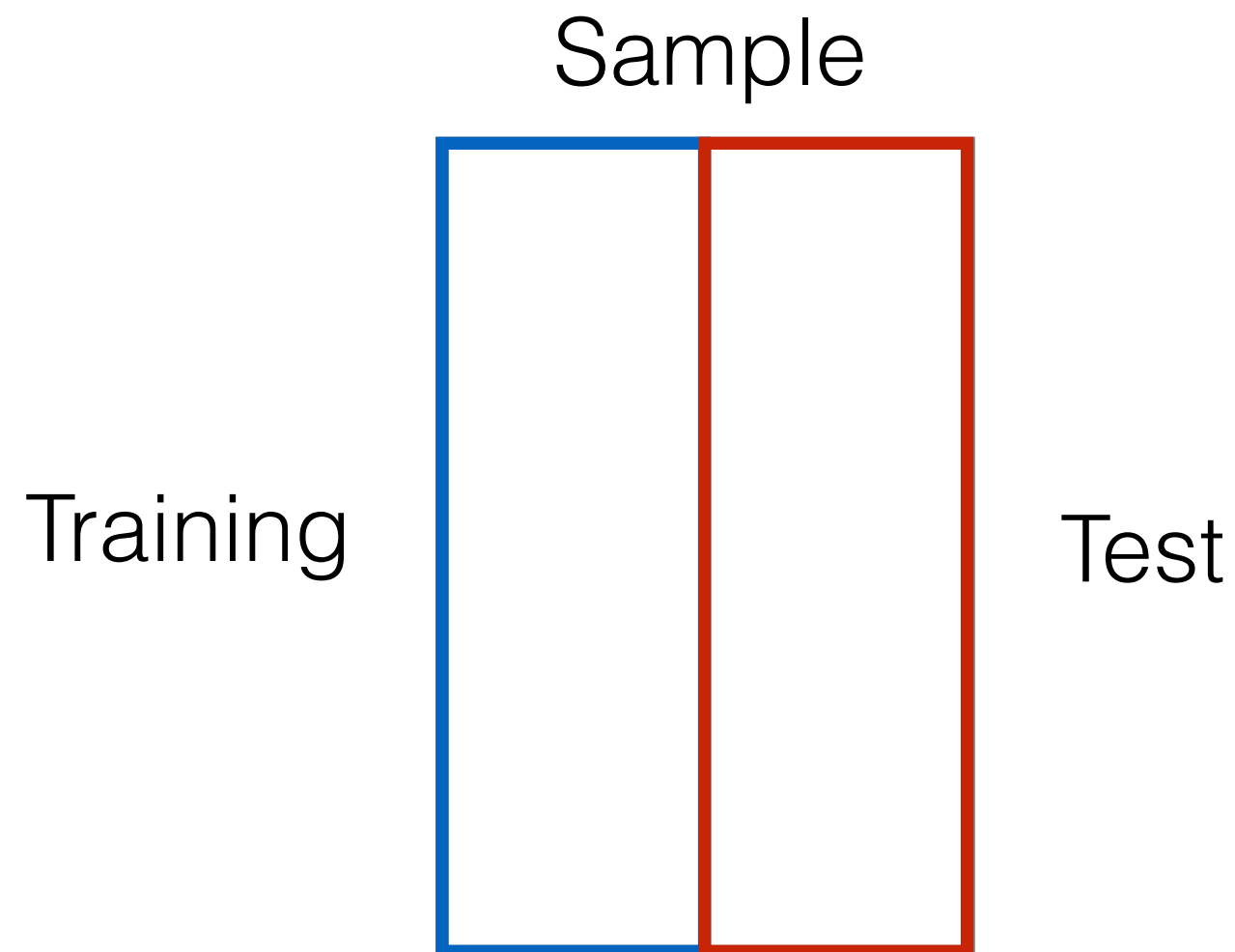
Which is more “useful”?

How can we tell?

Cross Validation

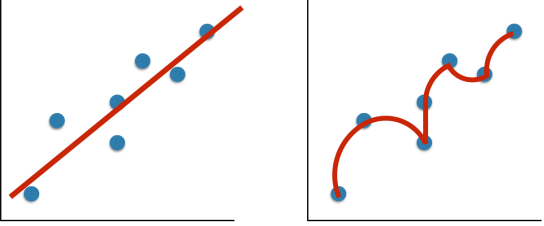
- Estimate how accurately a predictive model will perform in practice
- Give an insight on how the model will generalize to an independent dataset

Hold-out Validation



Problem: very dependent on which data are in each group

K-Fold Cross Validation

Sample			
Test 1	Training 1	5	2
Test 2	Training 2	4	2
Test 3	Training 3	3	1
Test 4	Training 4	5	4
Test 5	Training 5	4	2
		<hr/>	<hr/>
		4.2	2.2

Calculate how accurate we are in each “fold”
and average the answer