Spatial Summarization of Image Collections

Diego A. Ballesteros Villamizar

ETH Zürich

February 22nd, 2016

Outline

1. Data and baseline model changes
2. Partial set size and scores
3. Changes in learning
4. Data distribution and sampling
5. Featurized models

Clustering

- A minor change on the mean-shift algorithm settings. "Orphan" photos are ignored, previously they were assigned to the closest cluster even if they were outside the bandwidth.
- This reduces the number of photos and paths in the dataset. E.g. For $N = 10$ the number of paths is reduced from 8090 to 6781 and the number of photos covered is 30655.

Markov model

- Previously the Markov model had two issues: The set order was not preserved in the evaluation data and the model could use "future" items in the path for the prediction.
- After changes, the model is strictly as it would be used in an interactive system.
- To suggest an item given a partial sequence S_k, only the last item s_k is used and the results are ordered according to the transition probabilities, i.e. $P(s_{k+1} = x \mid s_k)$, learned from the data.
- The heuristic Markov model, assigns 0 probability to items previously seen in the path, i.e. $\forall_{x \in S_k} P(s_{k+1} = x \mid s_k) = 0$.
- This makes the model slightly worse than before, but still better than the current sub-/super-modular model.

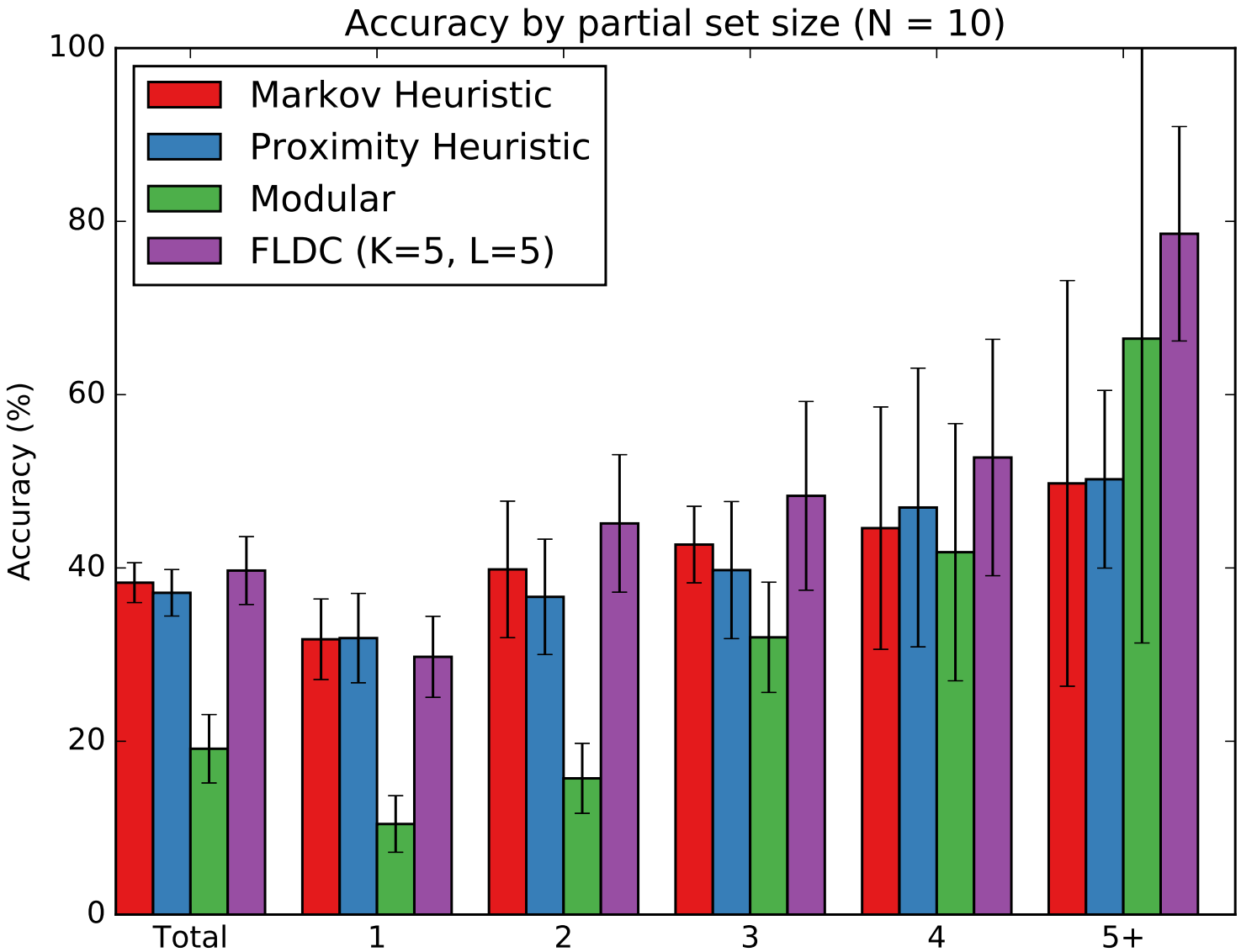
Proximity model

- Same as with the Markov model, the proximity model now only considers the items already seen in the path and only uses the distance from the item immediately before the one to predict in the sequence.
- This also decreases the accuracy score of the model.
- These changes allow the model to be used in a setting where a user is requesting item $k+1$ given a sequence S_k.

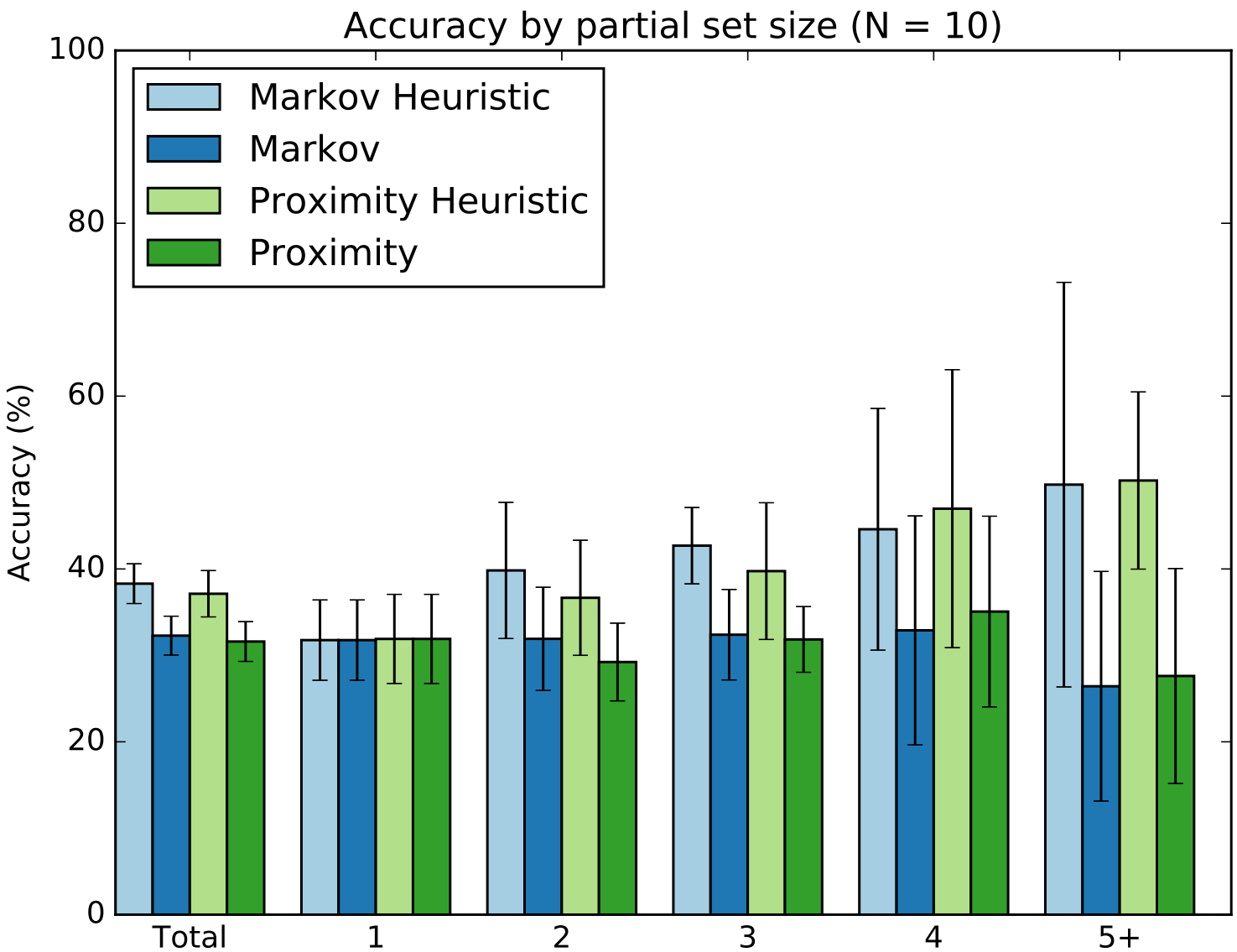
Outline

1. Data and baseline model changes
2. Partial set size and scores
3. Changes in learning
4. Data distribution and sampling
5. Featurized models

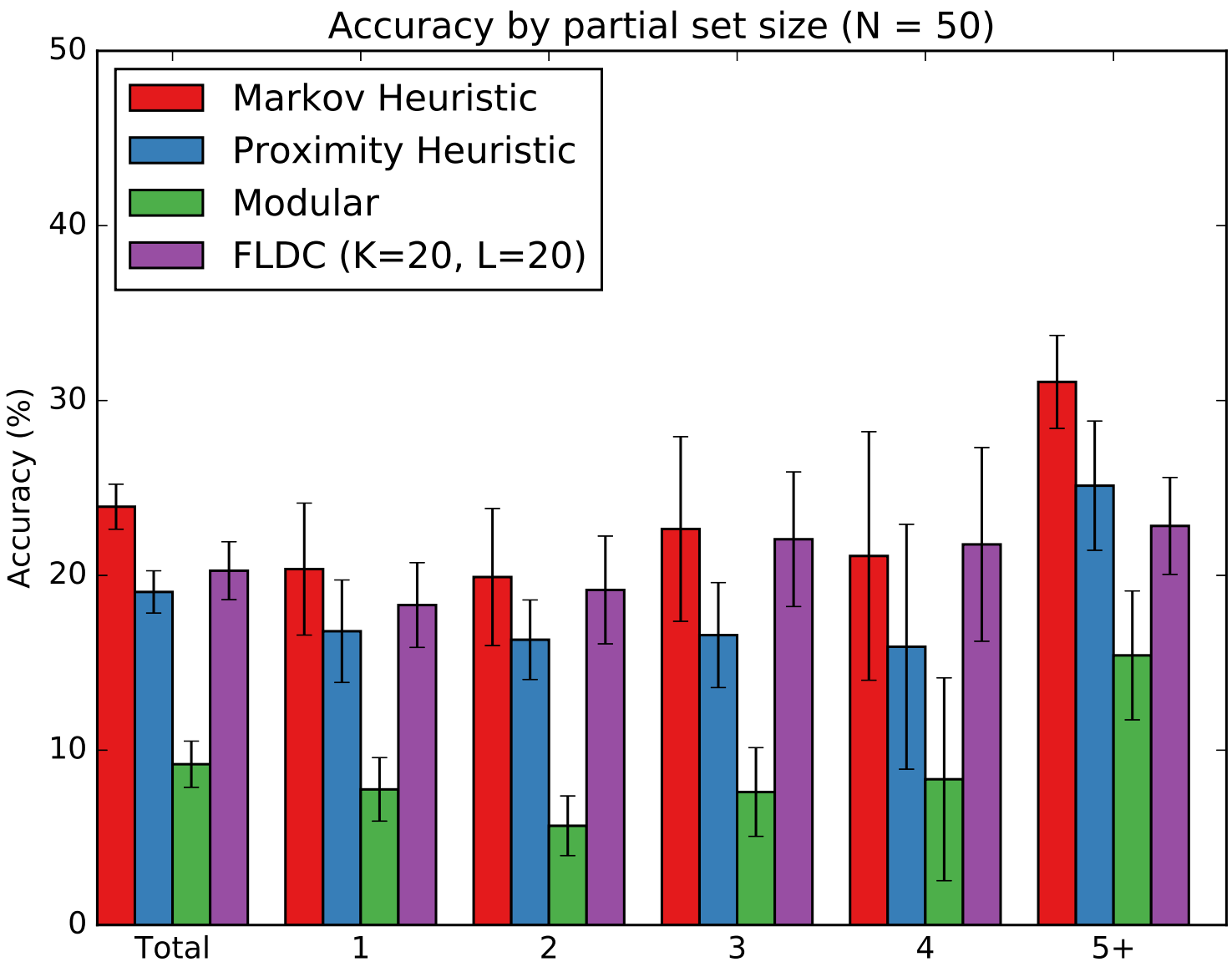
Partial set size effect on score for $N = 10$



Impact of heuristic methods



Partial set size effect on score for $N = 50$



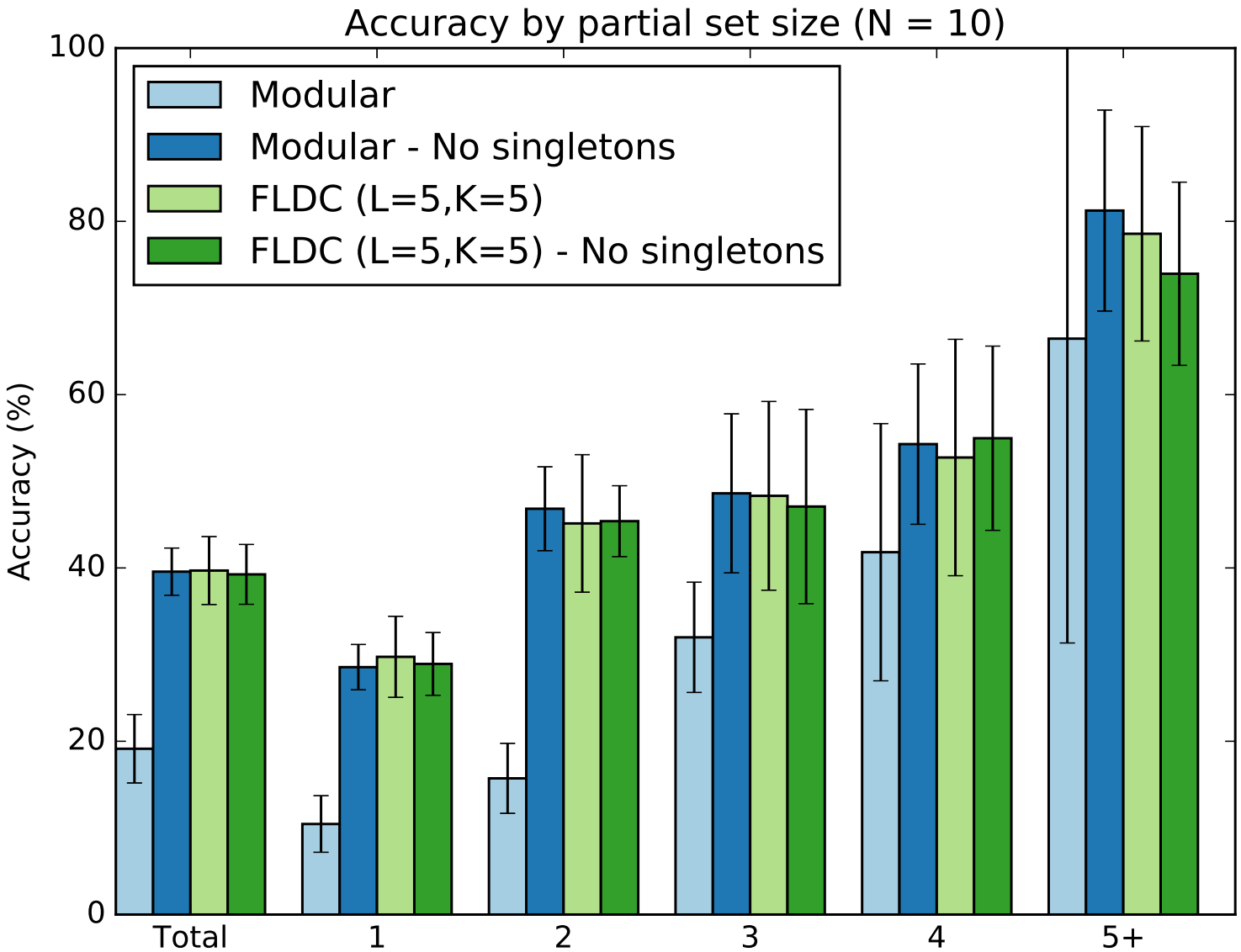
Outline

1. Data and baseline model changes
2. Partial set size and scores
3. Changes in learning
4. Data distribution and sampling
5. Featurized models

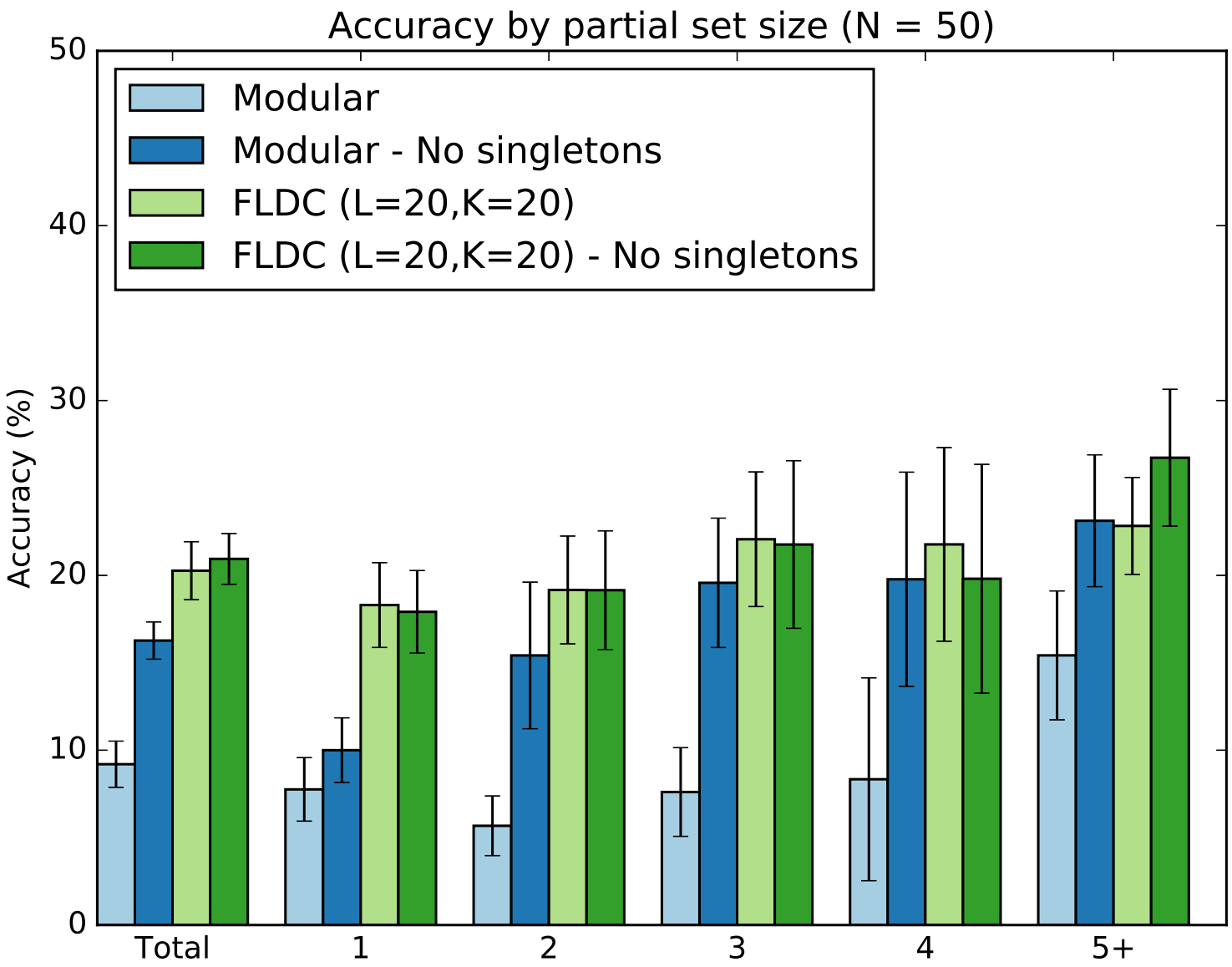
Learning from elements with at least 2 elements

- A lot of the paths are singletons, which are not relevant to the evaluation and all methods except the modular one.
- Removing all singletons changes the dataset from 6781 to 1242 for $N = 10$, and from 12614 to 2832 for $N = 50$.
- The expectation is that singletons don't add relevant information when dealing with sets of 2 or more items.
- Markov and proximity models are not affected by this change, they already learn only from non-singleton paths.

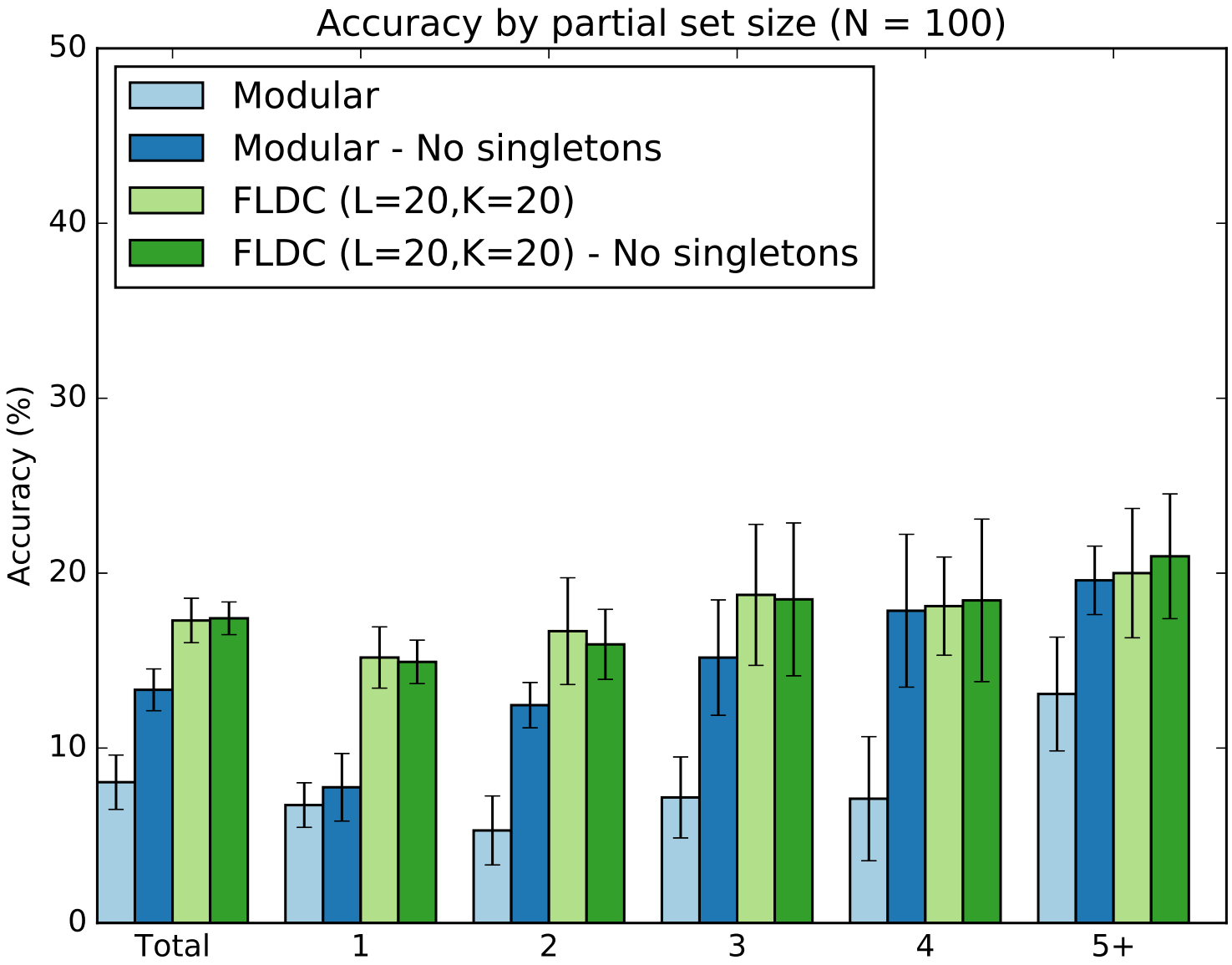
Results comparison $N = 10$



Results comparison $N = 50$



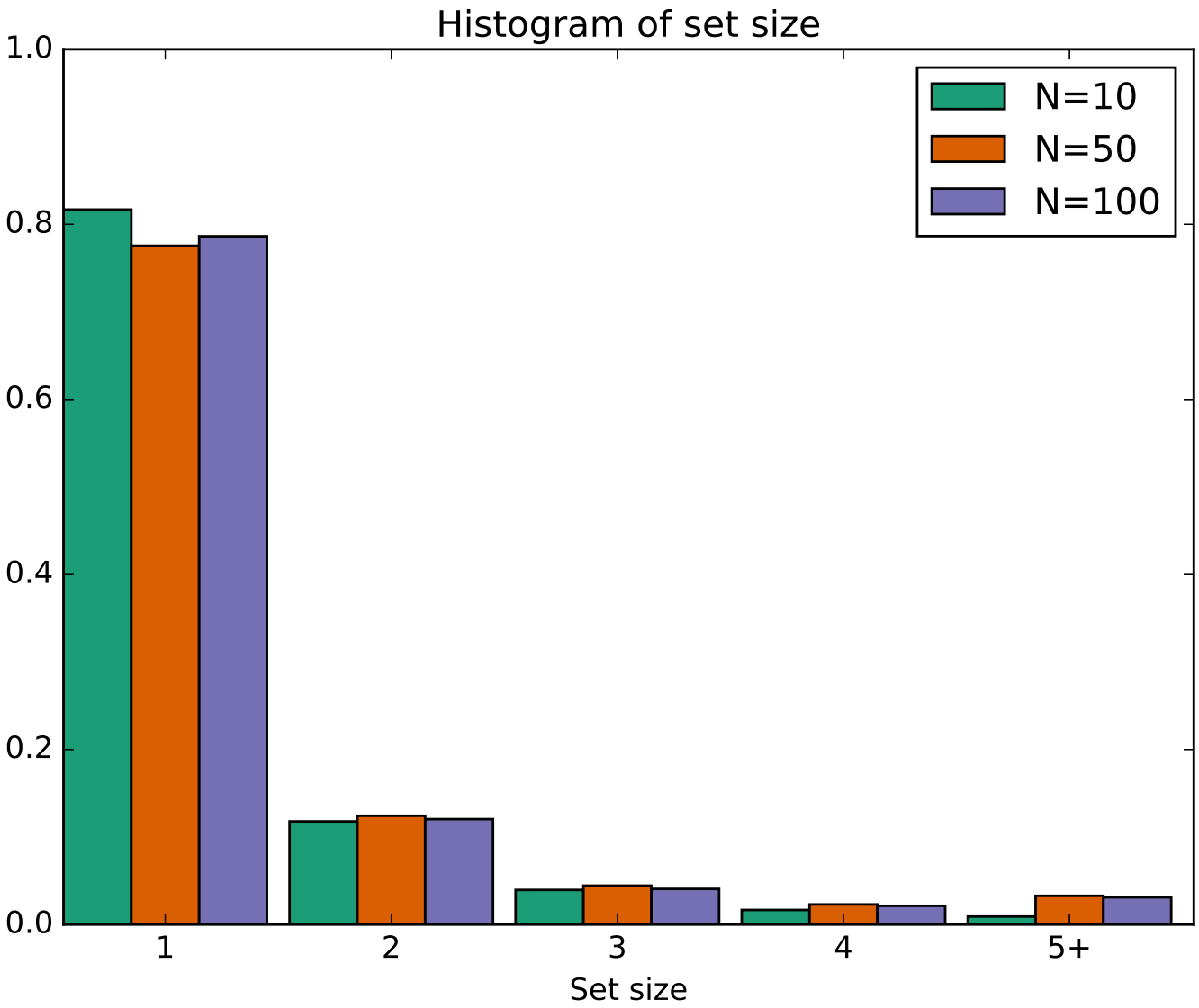
Results comparison $N = 100$



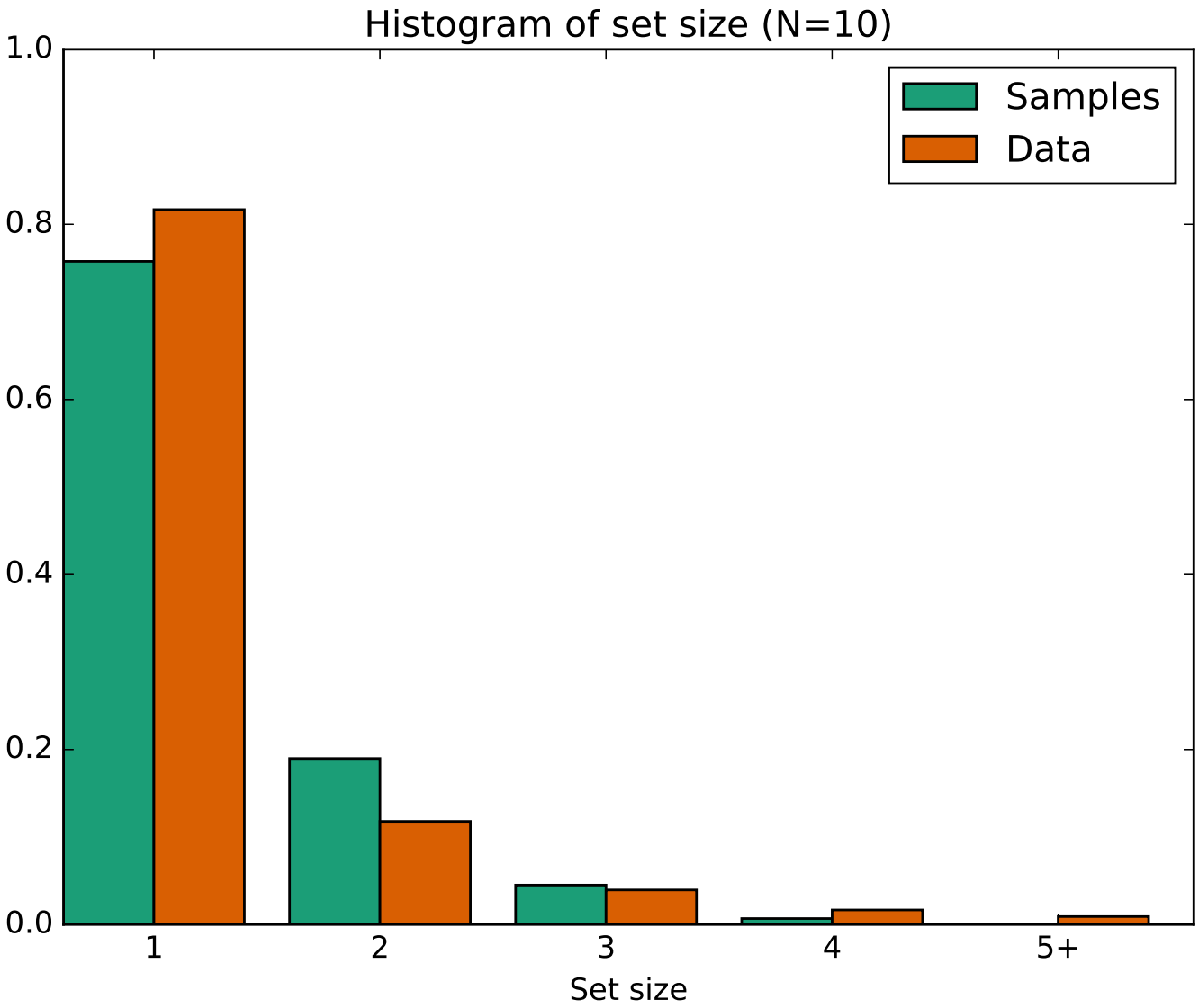
Outline

1. Data and baseline model changes
2. Partial set size and scores
3. Changes in learning
4. Data distribution and sampling
5. Featurized models

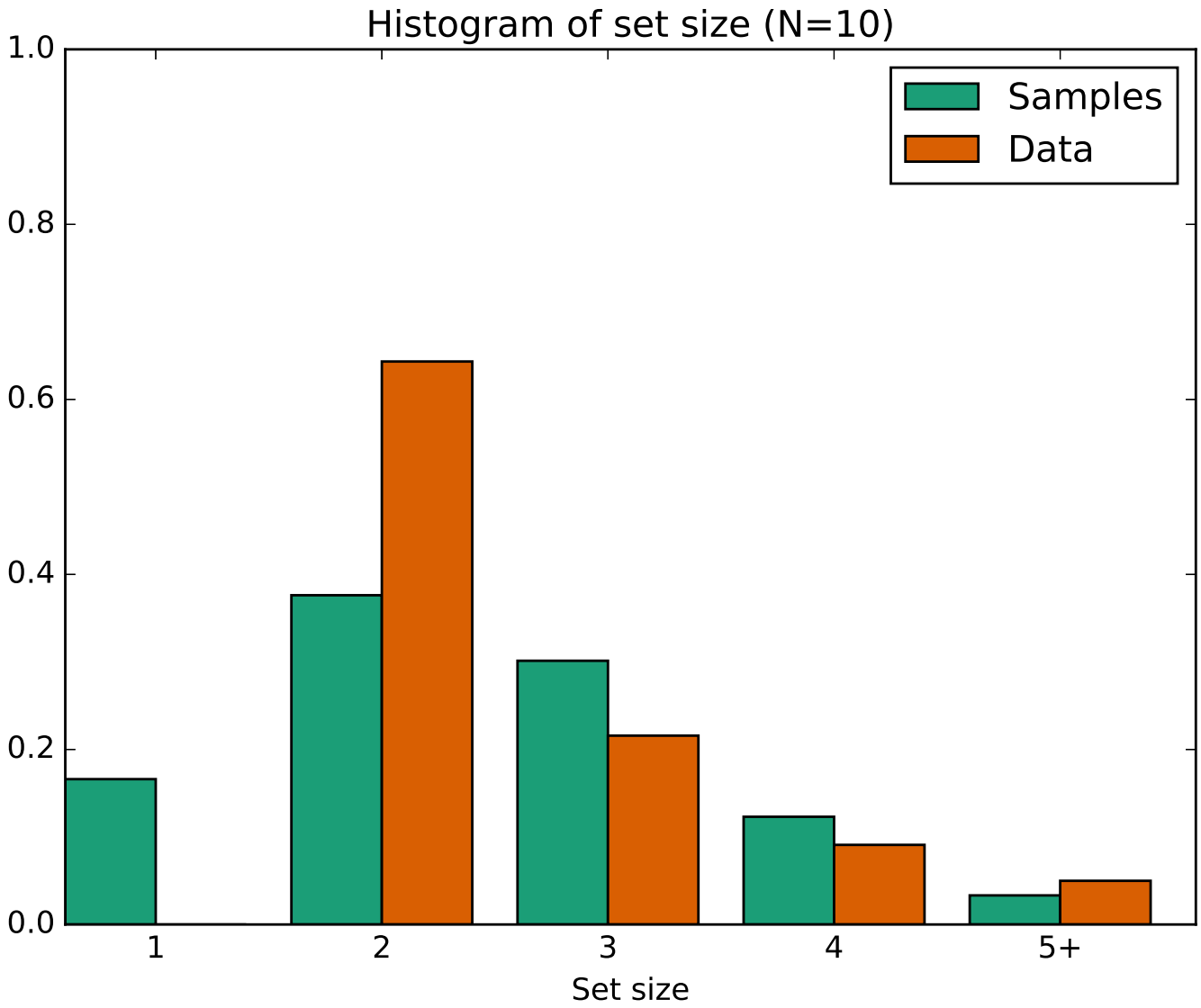
Set size distribution in data



Set size distribution after sampling



Set size distribution after sampling - no singletons



Outline

1. Data and baseline model changes
2. Partial set size and scores
3. Changes in learning
4. Data distribution and sampling
5. Featurized models

Dataset creation

- First remove "duplicates", i.e. photos with the exact same location.
- Filter photos that do not form a path of more than one item, i.e. when a user takes a single photo in a day.
- Sample 10k photos from the resulting set.
- Form a path dataset from the sample and remove singleton paths.
- The result has 1936 paths and comprises 9141 photos.

Feature definition

- Ten reference points corresponding to the top 10 cluster centers calculated using mean-shift clustering.
- For each item, the feature vector contains the distances to each cluster center.
- $\mathbf{X} \in \mathbb{R}^{9141 \times 10}$.

Models

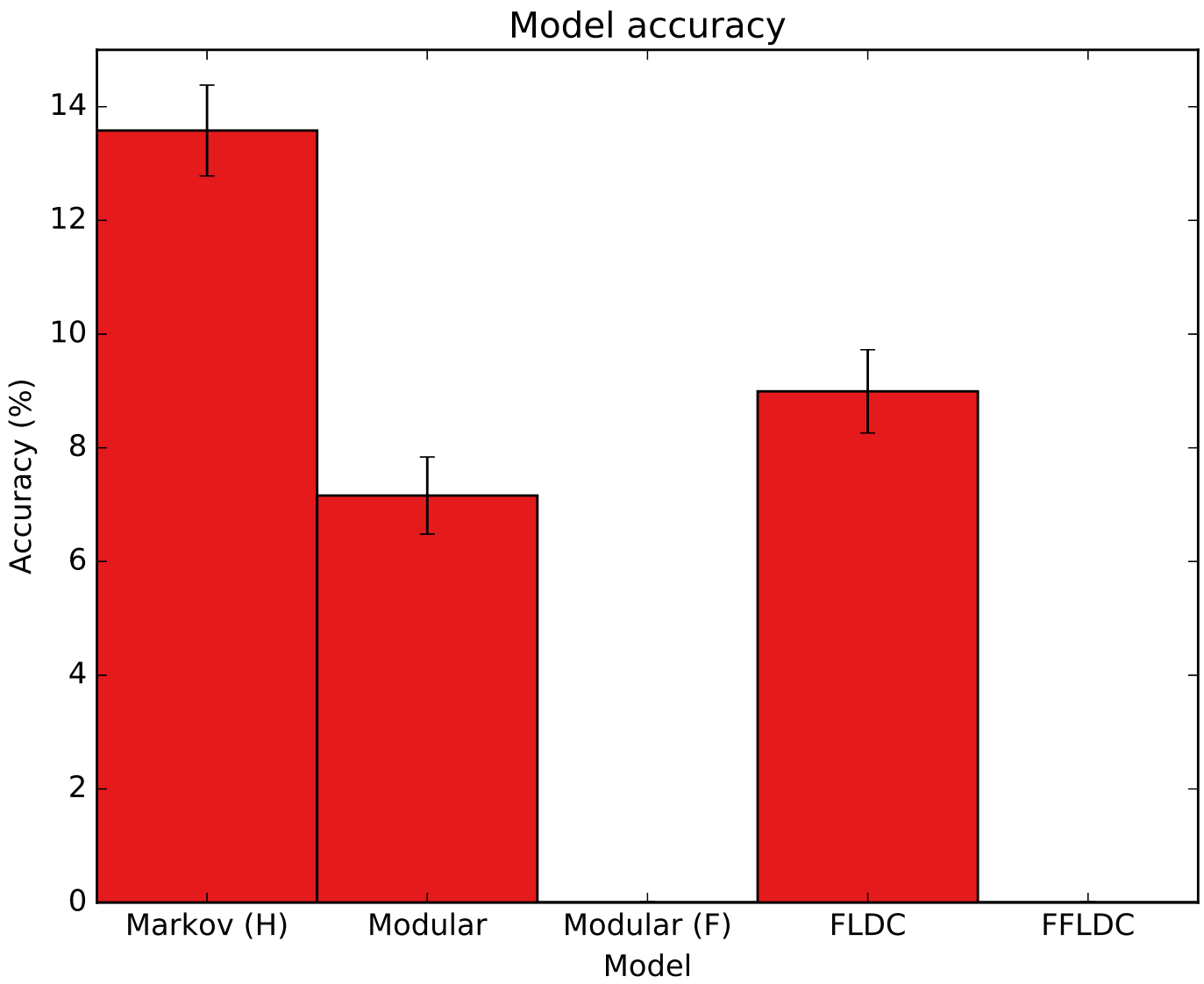
- A baseline proximity model that proposes an item closest to the previous item in the sequence.
- Modular model that using frequencies computes the utility vector $\mathbf{u} \in \mathbb{R}^{10}$.
- FLDC model with features, dimension parameters $L = 10, K = 10$.
- The results are:

Model	MRR (%)
Proximity	18.94 ± 3.28
Modular	0.11 ± 0.04
FFLDC (L=10, K=10)	0.10 ± 0.04
FFLDC (L=20, K=20)	0.07 ± 0.03

Bad model?

- Another possible dataset to test the featurized model is the dataset of all mean-shift clusters without filtering the top N.
- Feature vector is the location of each cluster center.
- Dataset has 6431 paths with more than one element, out of 27707, and 2685 items.

Evaluation


Model accuracy
Accuracy (%)
0
2
4
6
8
10
12
14
Markov (H)
Modular
Modular (F)
FLDC
FFLDC
Model