# BALANCING CLUSTERS TO REDUCE RESPONSE TIME VARIABILITY IN LARGE SCALE IMAGE SEARCH

Romain Tavenard (IRISA / Univ. Rennes I)
Hervé Jégou (INRIA Rennes)
Laurent Amsaleg (IRISA / CNRS)

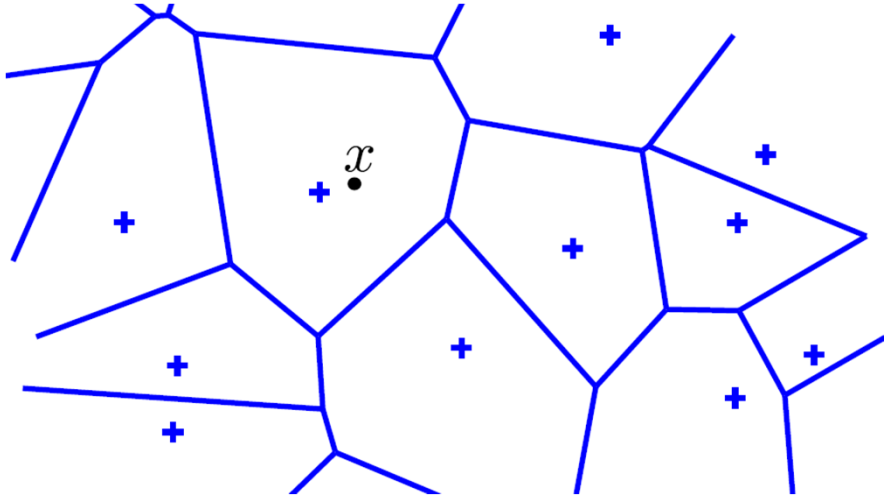
# Image retrieval & vector space

- Image ~ (set of) numerical feature(s)
- Visual similarity ~ feature space proximity

Image retrieval turned into a *k*-Nearest Neighbour (*k*-NN) problem in a vector space

# *k*-NN & quantization

- Many approximate *k*-NN algorithms rely on first quantizing the features using *k*-means-like algos
  - IVFADC(+R) [Jégou *et al.*, PAMI 2010]
  - Hierarchical *k*-means
  - *etc.*
- Quantization used to build a short-list of candidates
- (Costly) search performed on this short-list



# Impact of the cluster balance

- Average search complexity:

$$C \propto \sum_{i=1}^{k} {p_i}^2$$

- Optimal complexity (in terms of mean & variance) is achieved for balanced clusters

# Imbalance factor

- [Jégou *et al.*, IJCV 2010] defined imbalance factor as :

$$\gamma = \frac{C}{C_{opt}} = k \sum_{i=1}^{k} {p_i}^2$$

- Ratio between a given clustering's search complexity and the optimal one

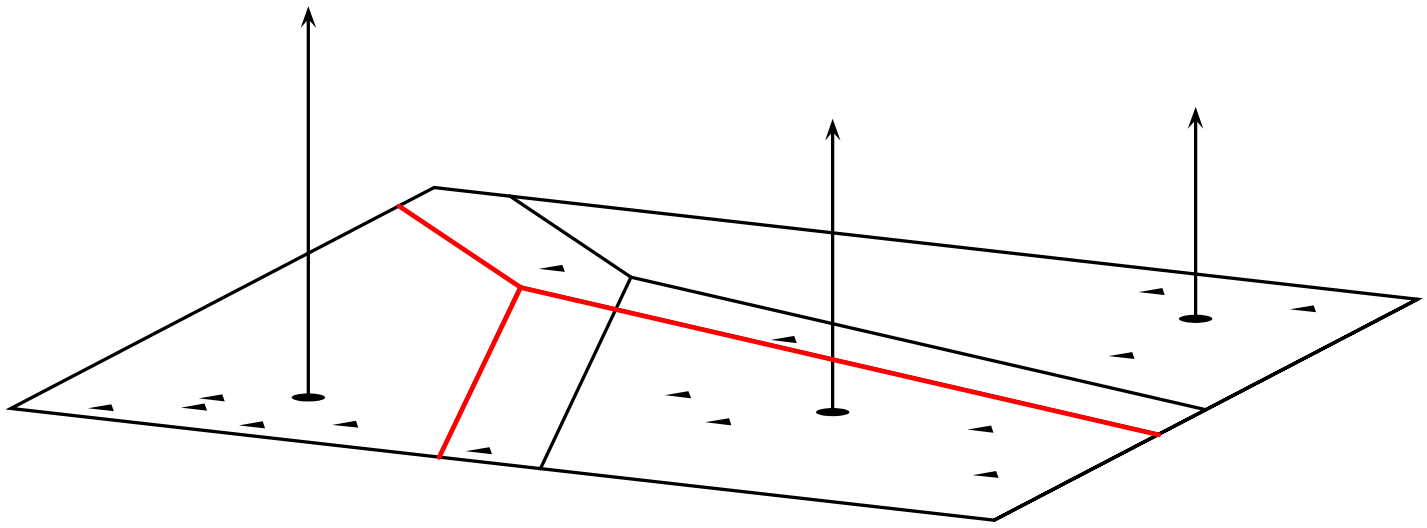
# Measured imbalance factors

- Databases of 1M features extracted from flickr images

| Feature | Dimension | $\gamma$ ($k$=256) | $\gamma$ ($k$=1024) |
|---|---|---|---|
| SIFT | 128 | 1.08 | 1.09 |
| BOF | 1,000 | 1.65 | 1.93 |
| GIST | 960 | 1.72 | 3.75 |
| VLAD | 8,192 | 5.41 | 6.23 |

# Proposed balancing strategy

- Before:
  - |A| = 7
  - |B| = 5
  - |C| = 3
- After:
  - |A| = 5
  - |B| = 5
  - |C| = 5



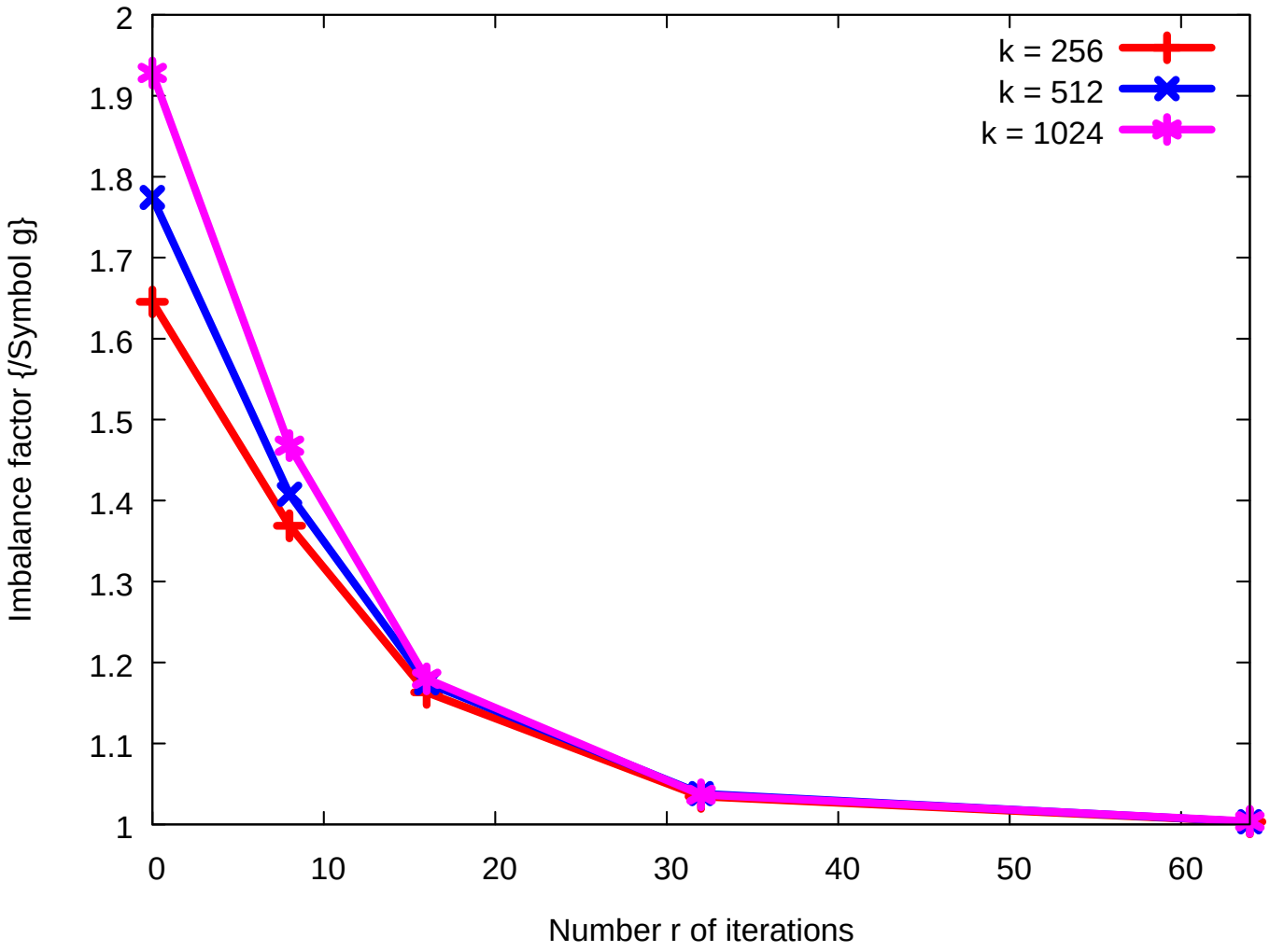
# Proposed balancing strategy - 2

- Iterative estimation of centroid elevation
- Resulting boundaries are parallel to the original ones
- Parameters :
  - Number *r* of iterations
  - Balancing power of each iteration (high balancing power = faster balancing but higher possible distortion on boundaries)

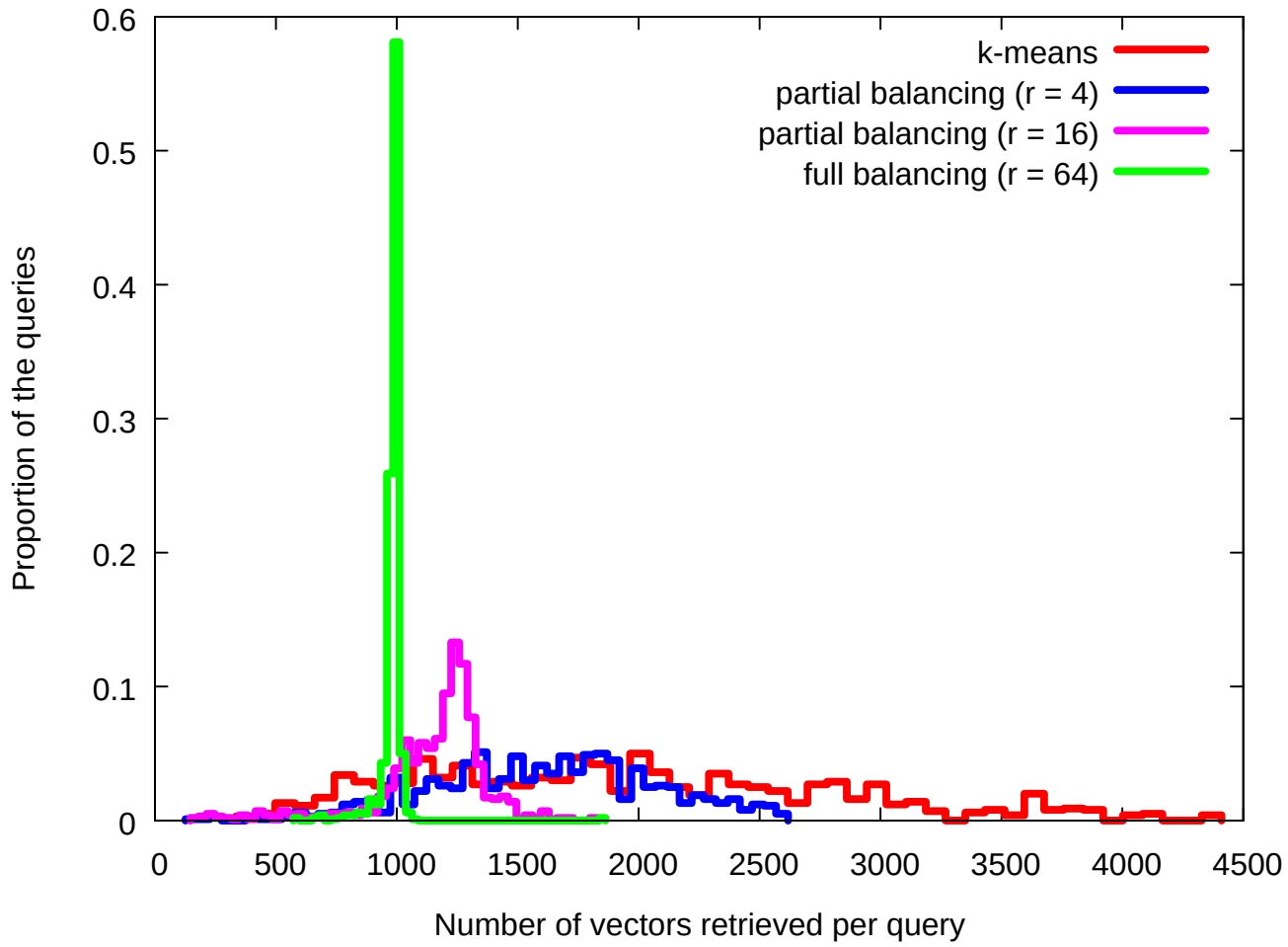
# Experiments

- BOF ($d$=1,000) features
- Database : 1M images from flickr
- Queries : 1k images from flickr

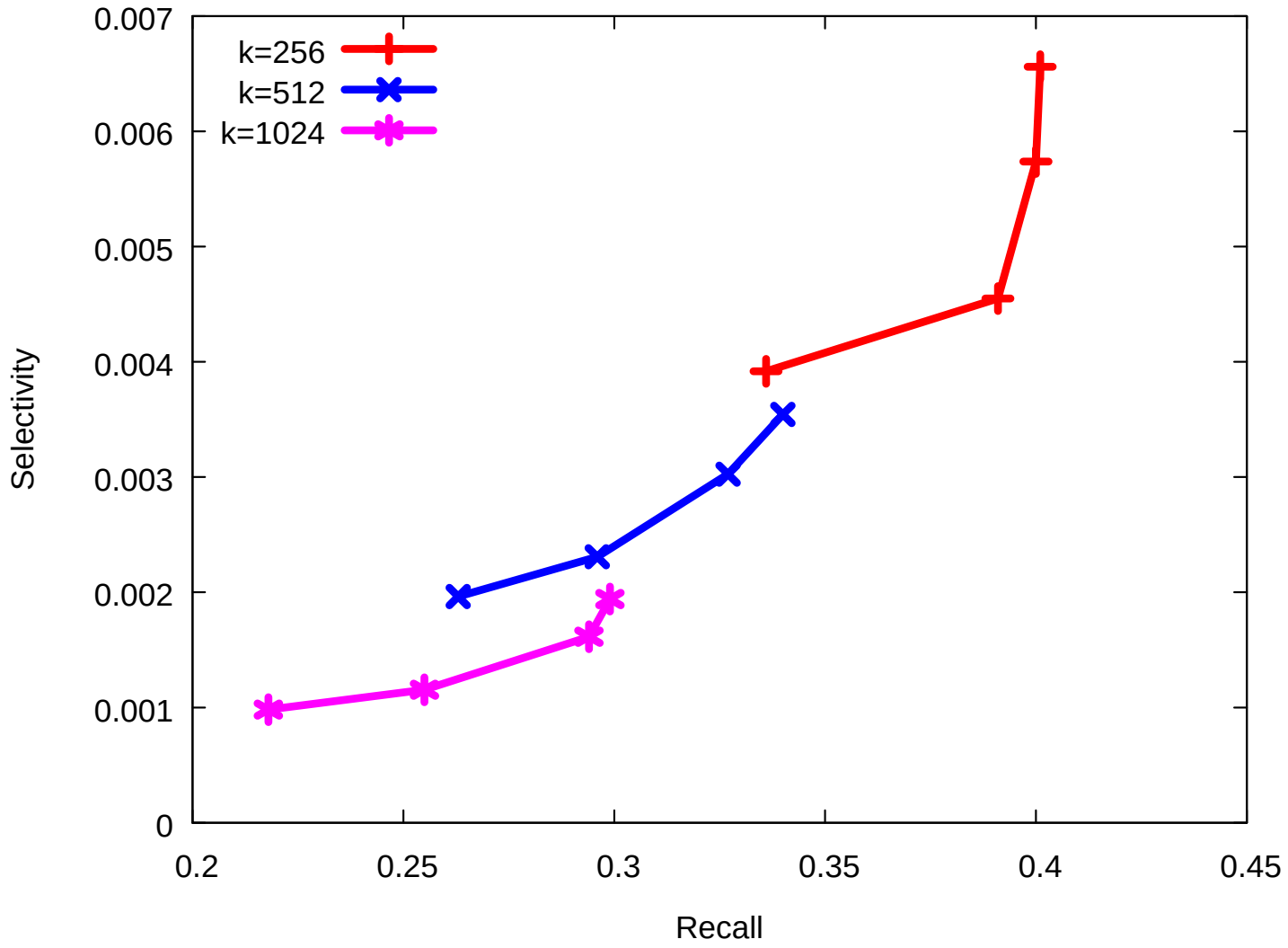
# Experiments



# Experiments



# Experiments



# Conclusion

- Issue:
  - Unbalanced clusters imply higher search complexity
- Proposition:
  - an algorithm to balance clusters generated by *k*-means
- Remarks:
  - especially crucial for high-dimensional features
  - be careful of perfect balancing