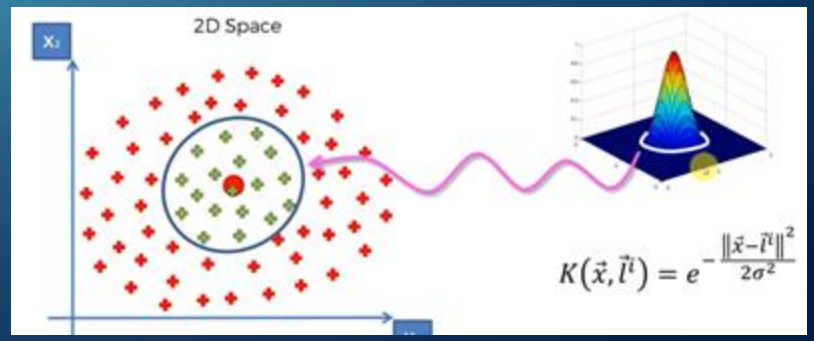
NYSTRÖM-OPTIMIZED KERNEL K-MEANS CLUSTERING

A COMPARISON OF CLUSTERING METHODS

LUKE FLECKER AND TANUSH KALLEM

INTRODUCTION TO CLUSTERING

- **Clustering:** Fundamental unsupervised machine learning task
- **Goal:** Divide data into clusters based on similarity
- **K-means:** Simple and efficient, but struggles with non-linear data
- **Kernel K-means:** Uses RBF kernels for non-linear patterns
  - Computationally intensive for large datasets



THE NYSTRÖM APPROXIMATION

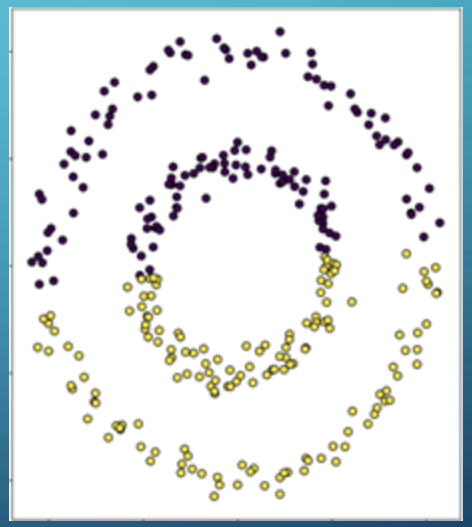
- **Problem:** Full kernel matrix calculation is expensive
- **Solution:** Nyström approximation
  - Approximates the kernel matrix by using a subset of data points
  - Reduces computational cost
  - Improves scalability of kernel methods for clustering non-linear data

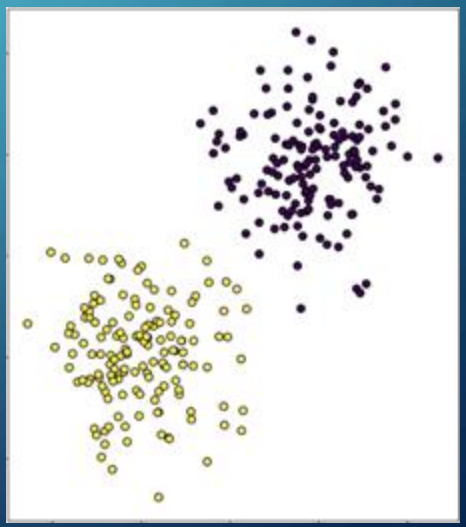
NYSTRÖM KERNEL K-MEANS METHOD

- **Nyström Approximation:** Approximates full kernel matrix (K) using a subset of data points(m)
- $K_{approx} = K_{nm}K_{mm}^{-1}K_M^T$
- K_nm: Kernel matrix between full dataset and sampled points
- K_mm: Kernel matrix of sampled points
- **Implementation** Samples m=50 points uniformly at random
- Efficient solution for complex data

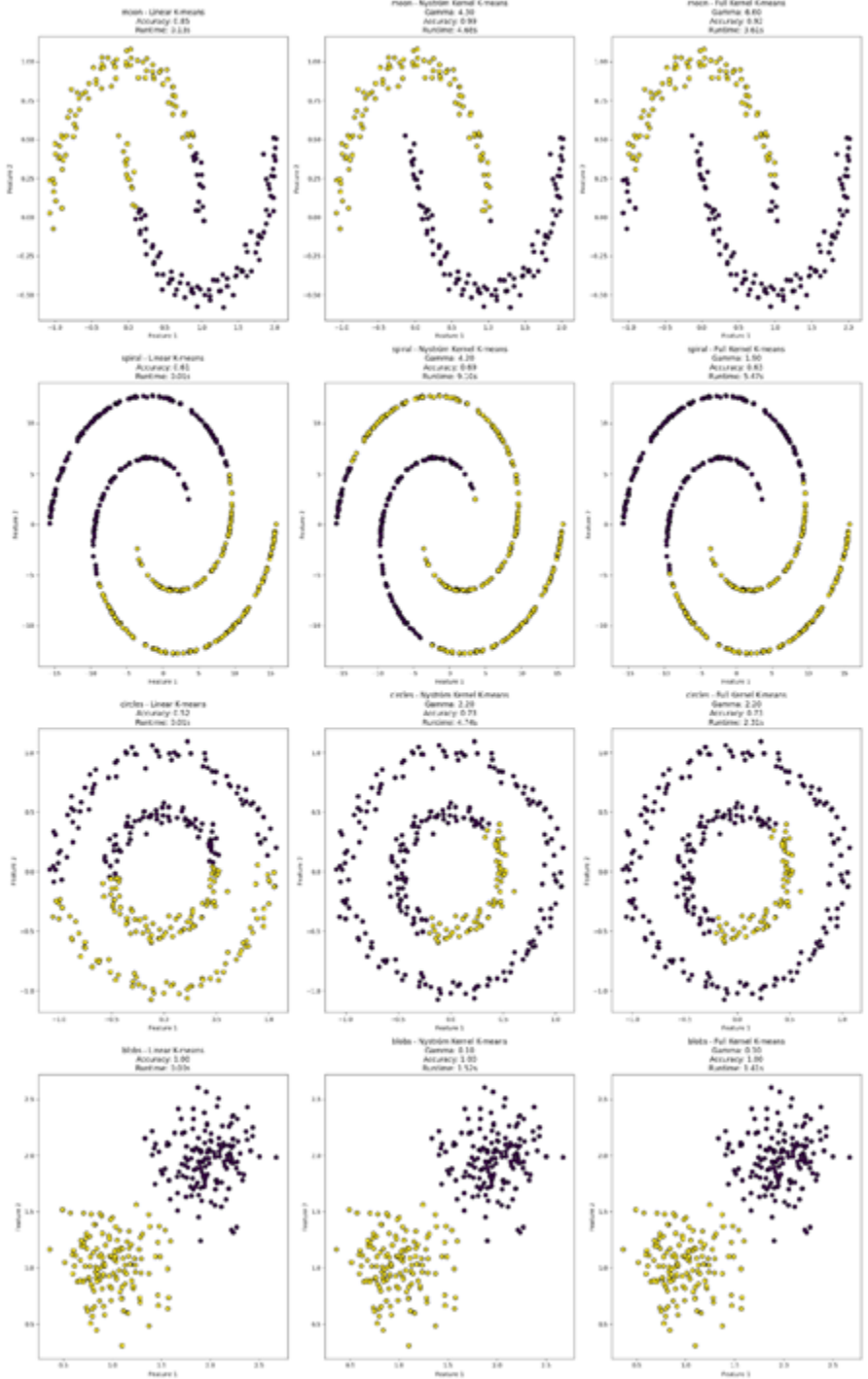
DATASETS

- **Moon:** Interleaving crescent shapes
  - Non-linearly separable
- **Spiral:** Spiral arrangement
  - Non-linearly separable
- **Circles:** Concentric circles
  - Non-linearly separable
- **Blobs:** Gaussian-distributed clusters





RESULTS



RESULTS

- **Linear K-means:**
  - Fast, but poor on non-linear data (Spiral, Circles)
  - Good on linearly separable data (Blobs)
- **Nyström Kernel K-means:**
  - Enhanced accuracy for non-linear data
  - Best accuracy on Moon dataset
  - Runtime slower than expected
- **Full Kernel K-means**
  - Consistently competitive accuracy
  - Runtime comparable to Nyström

Dataset	Method	Accuracy	Runtime(s)	Optimal γ
Moon	Linear K-means	0.85	0.13	-
	Nyström Kernel K-means	0.99	4.68	4.3
	Full Kernel K-means	0.92	3.61	6.6
Spiral	Linear K-means	0.61	0.01	-
	Nyström Kernel K-means	0.69	9.10	4.2
	Full Kernel K-means	0.68	5.47	1.7
Circles	Linear K-means	0.52	0.01	-
	Nyström Kernel K-means	0.73	4.74	2.2
	Full Kernel K-means	1.00	2.31	2.2
Blobs	Linear K-means	1.00	0.00	-
	Nyström Kernel K-means	1.00	1.52	0.1
	Full Kernel K-means	1.00	1.41	0.1

CONCLUSION & FUTURE WORK

- **Nyström Kernel K-means**: a good alternative for large datasets
  - Outperformed Full Kernel K-means on some datasets
- Can be improved with a coarser gamma search
- Try changing m
- Explore a coarser search for the gamma parameter
- Test on larger datasets
- Test Nyström kernel k-means on real-world applications