Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

MODELLING LARGE SCALE DATASETS USING PARTITIONING-BASED PCA

ICIP 2019
IEEE International Conference on Image Processing

Salaheddin Alakkari and John Dingliana
AlakkarS@tcd.ie, John.Dingliana@scss.tcd.ie

School of Computer Science and Statistics, Trinity College Dublin

INTRODUCTION

- Principal Component Analysis (PCA) is one of the most well-known dimensionality reduction and feature extraction methods.
- What distinguishes PCA from other data representations, such as Discrete Cosine Transform (DCT) is the fact that PCA finds the optimal basis vectors, referred to as *eigenvectors*, for representing a set of samples with minimal reconstruction errors.
- However, PCA is considered a holistic data representation which implies an expensive computational cost of $\mathcal{O}(nd\min(n,d))$ considering n samples of dimensionality d.
- Such a computational cost is prohibitive for high dimensional data.
- In this study, we examine the performance of partitioning-based variants of the PCA algorithm.

PROPOSED METHODS

Cell-based PCA

- Divide data attributes into spatially uniform contiguous blocks that are called cells.
- PCA is then applied to attributes within each cell separately.
- This scheme is inspired by the block-wise DCT used in the JPEG compression standard.
- Unlike DCT, PCA enjoys better theoretical guarantees in terms of finding the basis vectors that optimally captures variability in the corresponding cell.
- Such data-dependent basis vectors achieve minimal reconstruction errors.

Band-based PCA

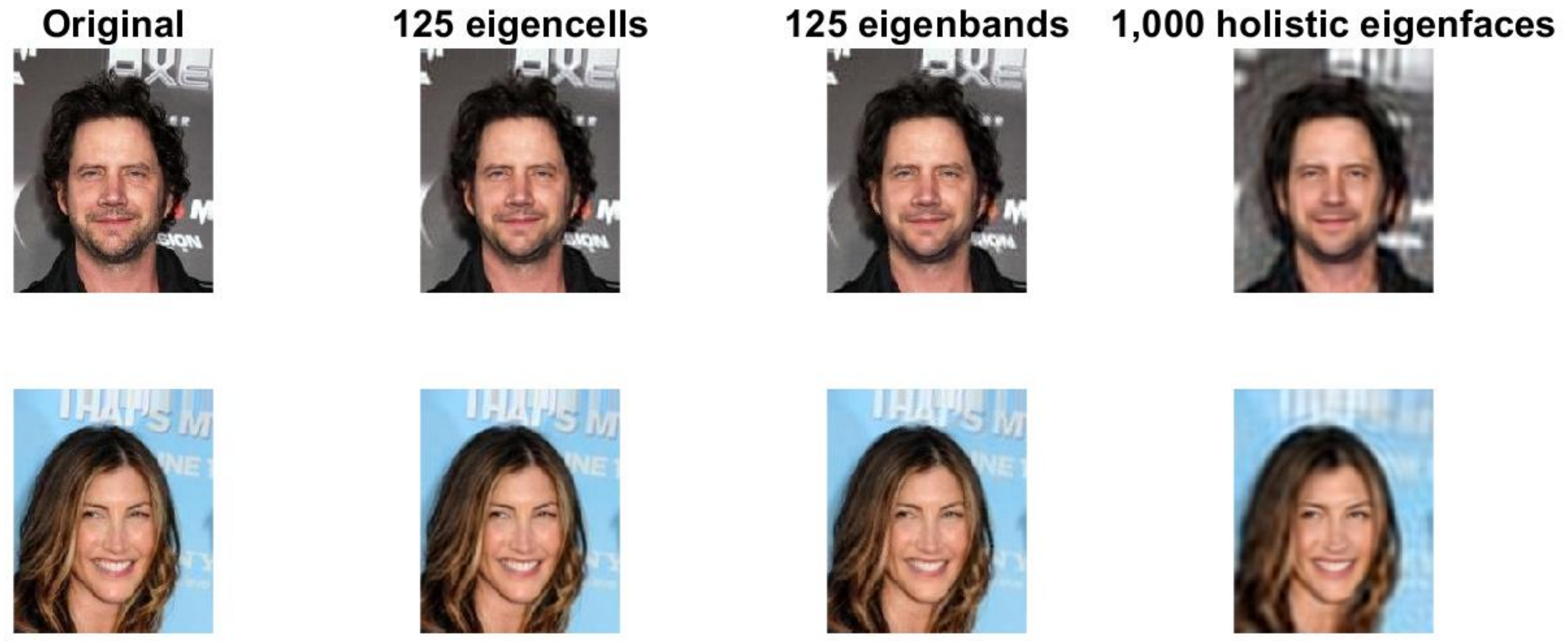
- Since standard PCA does not make any spatial assumption on data samples, we propose a more general model, band-based PCA, that partitions attributes based on their values distribution.
- Such a model can be employed to datasets that are not spatially localized.
- We show that partitioning based on the mean and variance achieves competitive results in comparison to cell-based PCA.

RESULTS

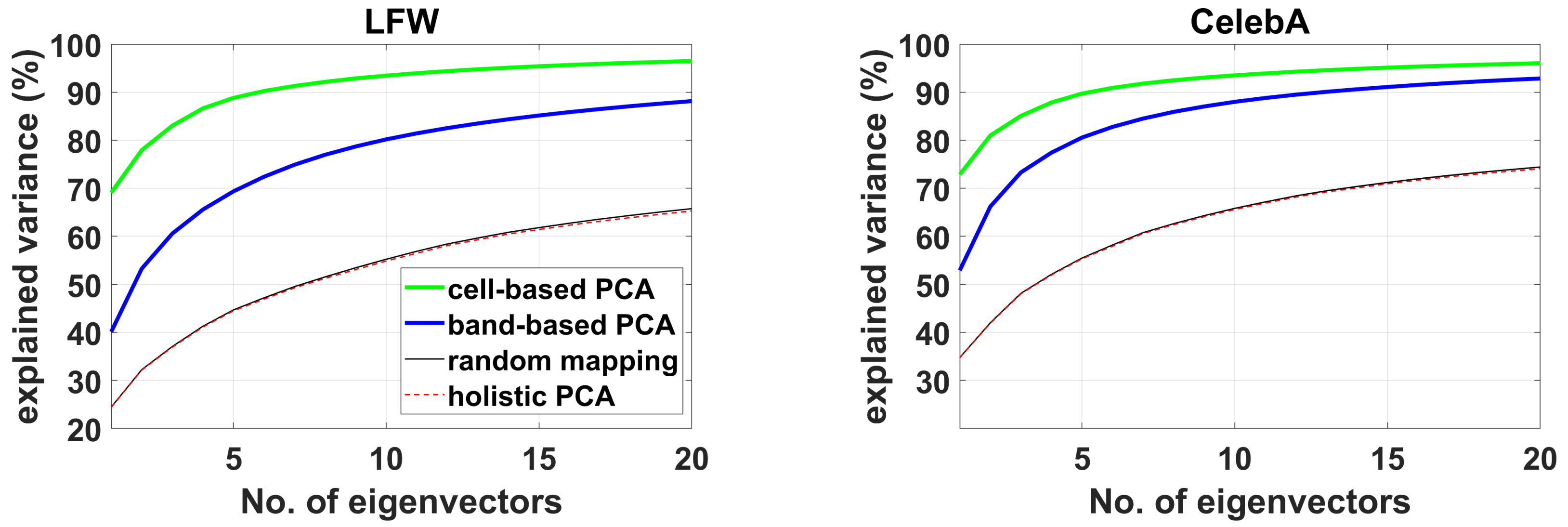
- We run experiments on two large-scale face datasets, namely, Labelled Faces in the Wild (LFW) and CelebFaces Attributes (CelebA).
- Computing the holistic PCA for such large datasets is infeasible.
- We hence approximate the significant 1,000 holistic eigenvectors of both datasets using the streaming PCA approach proposed by Arora et al. which took 21 hours for LFW and 4.6 days for CelebA.
- Both cell-based PCA and band-based PCA achieve remarkably better reconstruction results compared to the holistic approach in terms of SSIM, MSE and explained variance, not to mention their speed-up.
- The cell-based PCA gives slightly better results over band-based PCA.
- Randomly mapping attributes to different bands gives the baseline performance in terms of explained variance which is very similar to the holistic model.

MSE and SSIM of cell-based PCA (green entries) and band-based PCA (blue entries) compared to the holistic approach.

	LFW			CelebA		
Partition-size	25× 25 × 3	1,875	Holistic	25× 25 × 3	1,875	Holistic
# of eigenvectors per part	125	125	1,000	125	125	1,000
MSE	**0.00027**	0.00062	0.0029	**0.0007**	0.0011	0.003
SSIM	**0.9425**	0.89	0.71	**0.89**	0.85	0.72



Two samples from CelebA reconstructed using 125 cell and band eigenvectors compared to 1,000 holistic eigenvectors.



The percentage of variance maintained by the first 20 eigenvectors computed using each scheme.

COMPUTATIONAL COMPLEXITY

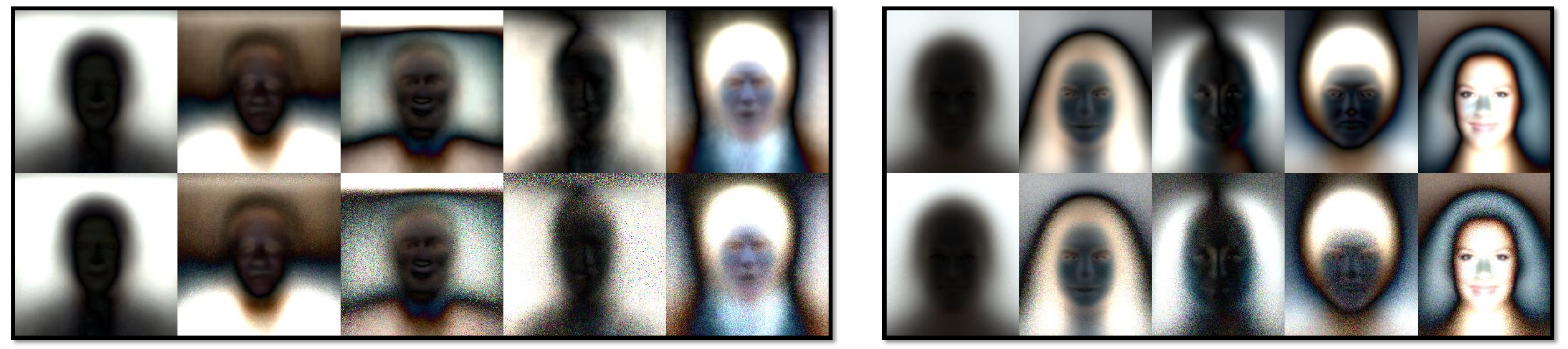
- Since the approach is embarrassingly parallel, the computational cost can be reduced down to $\mathcal{O}\left(n(d/p)^2\right)$ flops instead of $\mathcal{O}\left(nd^2\right)$ for the standard approach given n samples of dimensionality d and p uniform partitions.
- Streaming PCA methods, while reducing computational cost of the standard PCA, have quadratic complexity dependence on the number of eigenvectors k expressed as $\mathcal{O}\left(k^2nd\right)$.

Average run-time per cell (in seconds) using CPU and GPU implementations. This reflects performance in parallel settings.

	LFW			CelebA		
Cell-size	5× 5 × 3	10× 10 × 3	25× 25 × 3	5× 5 × 3	10× 10 × 3	25× 25 × 3
CPU run-time/cell	**0.0427**	**0.16**	7.87	**0.44**	1.92	63.3
GPU run-time/cell	0.3273	0.295	**3.19**	0.9	**1.2845**	**4.39**

ANALOGY TO THE HOLISTIC APPROACH

The solution produced by the baseline model of band-based PCA when assigning attributes randomly shows high resemblance to the holistic eigenvectors. This suggests that such baseline model may be a more practical alternative to state-of-the-art streaming PCA algorithms.



First five eigenvectors of LFW (left) and CelebA (right) resulted from streaming PCA (top) and random mapping (bottom).

CONCLUSION

Applying PCA in the partitioning mode is

- **embarrassingly parallel,**
- **generalizable to non-spatially localized data,**
- **and significantly better in reconstruction quality.**

This research has been conducted with the financial support of Science Foundation Ireland (SFI) under Grant Numbers 13/IA/1895 and 13/RC/2106.

ADAPT

SFI Science Foundation Ireland