Niamh McCombe [1] Xuemei Ding [1] Girijesh Prasad [1] David P. Finn[2] Stephen Todd [3] Paula L. McClean [4] KongFatt Wong-Lin [1]

Predicting Feature Imputability in the Absence of Ground Truth

[1] Intelligent Systems Research Centre, Ulster University, Magee Campus, Derry ~Londonderry, Northern Ireland, UK; [2]Pharmacology and Therapeutics, School of Medicine, National University of Ireland Galway, Galway, Ireland [3]Altnagelvin Area Hospital, Western Health and Social Care Trust; [4]Northern Ireland Centre for Stratified Medicine, Biomedical Sciences Research Institute, Ulster University, Derry ~ Londonderry, Northern Ireland, UK.
This project is supported by the European Union's INTERREG VA Programme, managed by the Special EU Programmes Body (SEUPB)

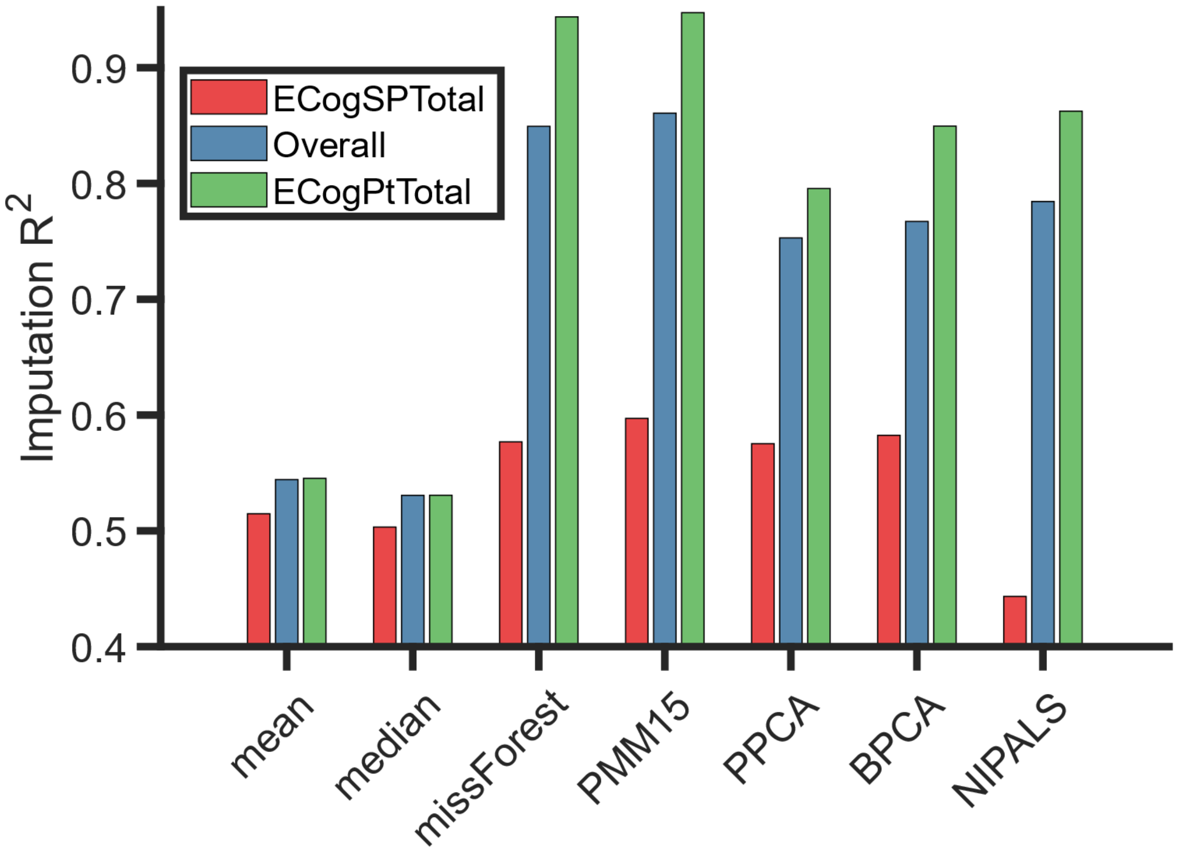
1st Workshop on the Art of Learning with Missing Values (Artemiss) hosted by the 37th International Conference on Machine Learning (ICML).

Missing Data Imputation Experiment on Dementia Data

- Artificial missing values introduced into CFA (cognitive and functional assessment) variables in ADNI (Alzheimer's Disease Neuroimaging Initiative) open source data.
- Missingness pattern introduced as observed in local memory clinic data: $P_{miss} = 0.48 \mp 0.06MMSE$ **(48% missing)**
- Dataset: 8 CFAs, Gender, Age, and Class Variable CD-RSB (Disease severity)
- Commonly used imputation methods and several PCA based methods tested. Imputed values of individual CFA features regressed against ground truth.

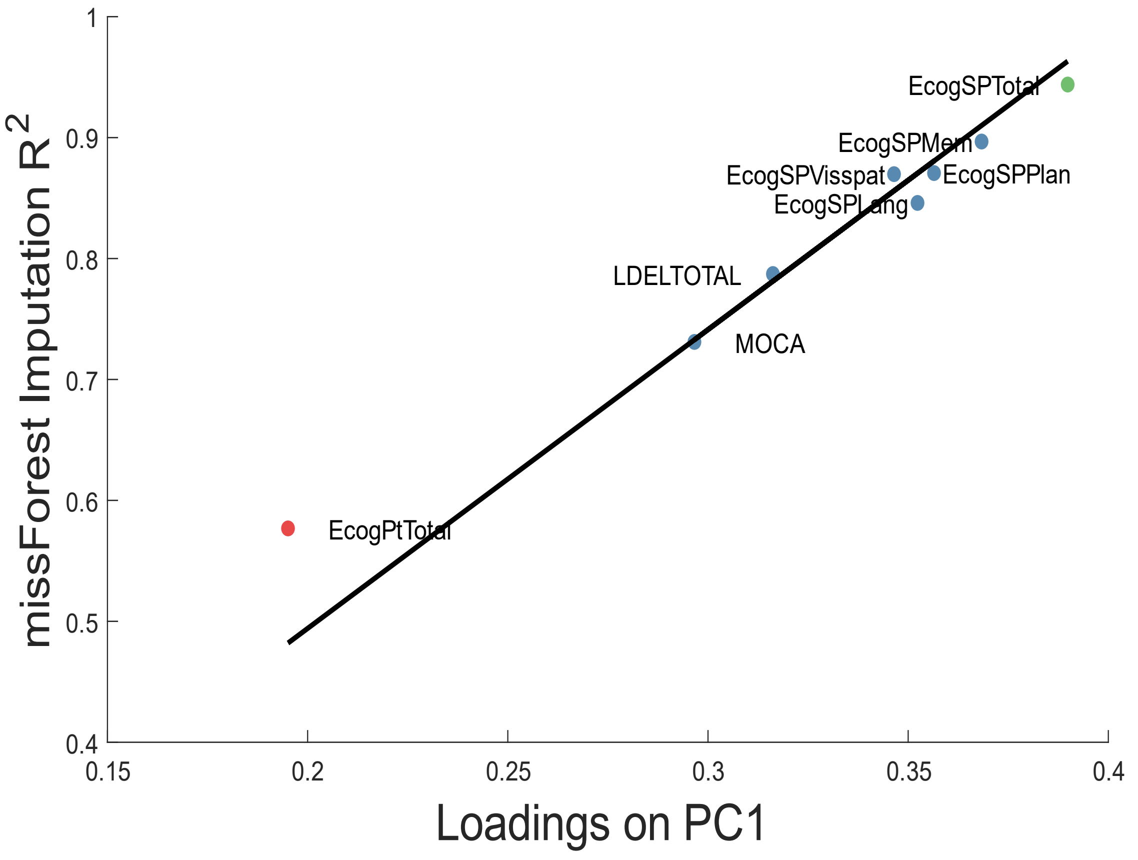
Motivation

- High degree of missingness in clinical data for dementia necessitates careful imputation.
- Little work considers imputability of different data features

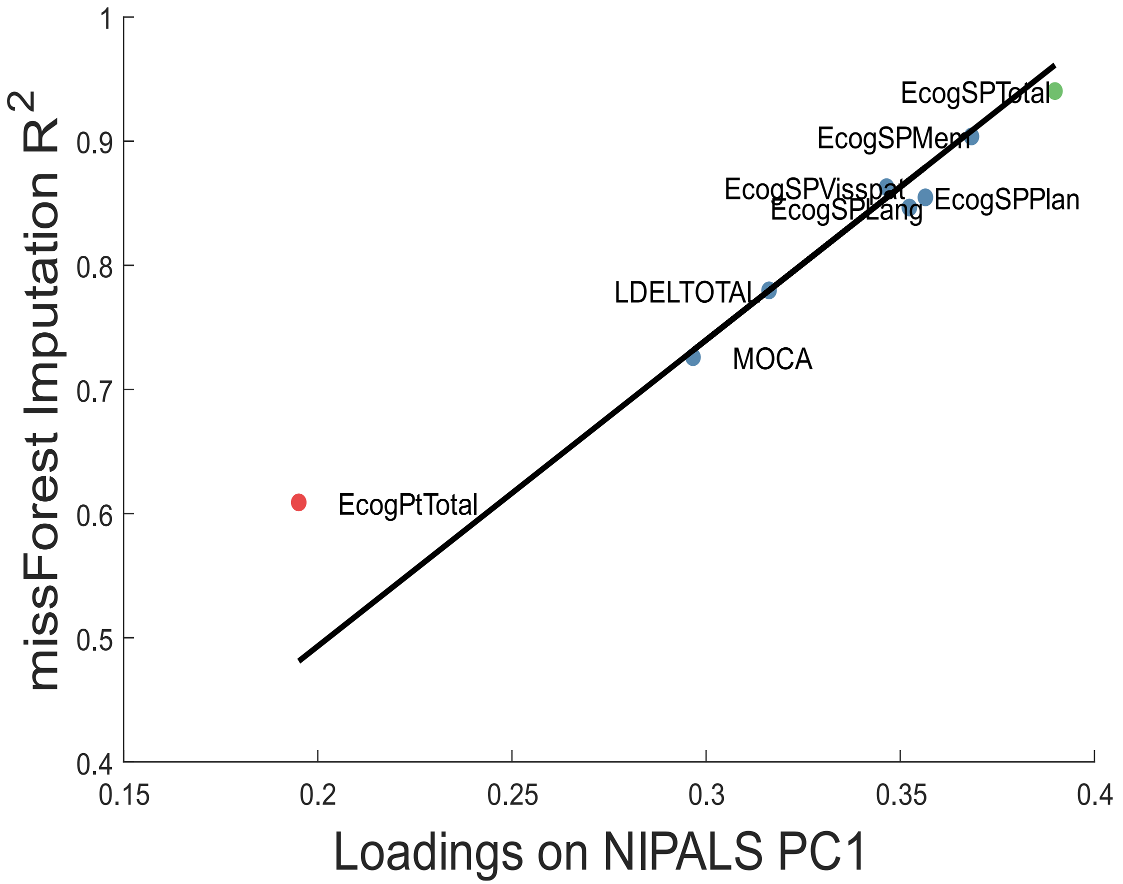


PCA loadings and missForest Feature Imputability

VARIABLE	PC1	PC2	PC3	R^2
CDR-SB	0.322	0.012	0.304	n/a
Gender	0.0719	-0.679	0.195	n/a
Age	0.079	-0.693	-0.303	n/a
EcogSPTotal	0.390	0.071	-0.194	0.862
EcogSPMem	0.368	0.045	-0.068	0.821
LDELTOTAL	-0.316	0.017	-0.296	0.775
EcogSPLang	0.352	0.0350	-0.148	0.763
MOCA	-0.297	0.144	-0.177	0.682
EcogSPPlan	0.356	0.103	-0.285	0.797
EcogSPVisspat	0.346	0.123	-0.306	0.791
EcogPtTotal	0.1959	0.0590	0.648	0.443



Predicting Feature Imputability in the Absence of Ground Truth



- **Summary:**
  - Feature imputation accuracy can be predicted even where missingness is very high
  - Informing further analysis of imputed datasets
  - **Potential groundwork for new missing data strategies.**
- Implications
  - Should less imputable features be omitted from ongoing analysis?
    - Orthogonality considerations
- Further work
  - Explore different PC structures and datasets
  - Experiment with workflows which explicitly consider feature imputability