Chains of Autoreplicative Random Forests for missing value imputation in high-dimensional datasets

Multi-Label Learning workshop, ECML/PKDD 2022
Grenoble, France

Ekaterina Antonenko and Jesse Read

September 19, 2022

Laboratoire d'informatique, École Polytechnique, IP Paris
DigitalentLab, MIA, Moteurs d'Intelligence Artificielle

Digitalent

Missing data

Why data is missing?

- Errors in sensors
- Human factor (reluctance to answer particular questions)
- Combining different studies
- ...

Missing data

Why data is missing?

- Errors in sensors
- Human factor (reluctance to answer particular questions)
- Combining different studies
- ...

Why to impute missing data?

- Most off-the-shelf statistical and machine learning methods cannot handle missing values
- Considering only instances with complete information can lead to a loss of necessary information and can yield a very poor or even empty dataset
- Missing data itself might be of interest

Types of missingness

- Missing Completely at Random (MCAR)

  The absence occurs entirely independently from feature values
- Missing at Random (MAR)

  The absence depends only on the observed feature values
- Missing Not at Random (MNAR)

  The absence depends on both observed and the unobserved feature values

Imputation methods

Methods		Limitations
• Mode / mean / median	$\Longrightarrow$	Poor performance

Imputation methods

Methods		Limitations
• Mode / mean / median	$\Longrightarrow$	Poor performance
• k Nearest Neighbors (kNN)	$\Longrightarrow$	Ok, but can we do better?

Imputation methods

Methods		Limitations
• Mode / mean / median	$\Longrightarrow$	Poor performance
• k Nearest Neighbors (kNN)	$\Longrightarrow$	Ok, but can we do better?
• Singular Value Decomposition (SVD)	$\Longrightarrow$	Ok, but can we do better?

Imputation methods

Methods		Limitations
• Mode / mean / median	$\Longrightarrow$	Poor performance
• k Nearest Neighbors (kNN)	$\Longrightarrow$	Ok, but can we do better?
• Singular Value Decomposition (SVD)	$\Longrightarrow$	Ok, but can we do better?
• Multiple Imputation by Chained Equations (MICE)	$\Longrightarrow$	Too slow for high-dimensional data

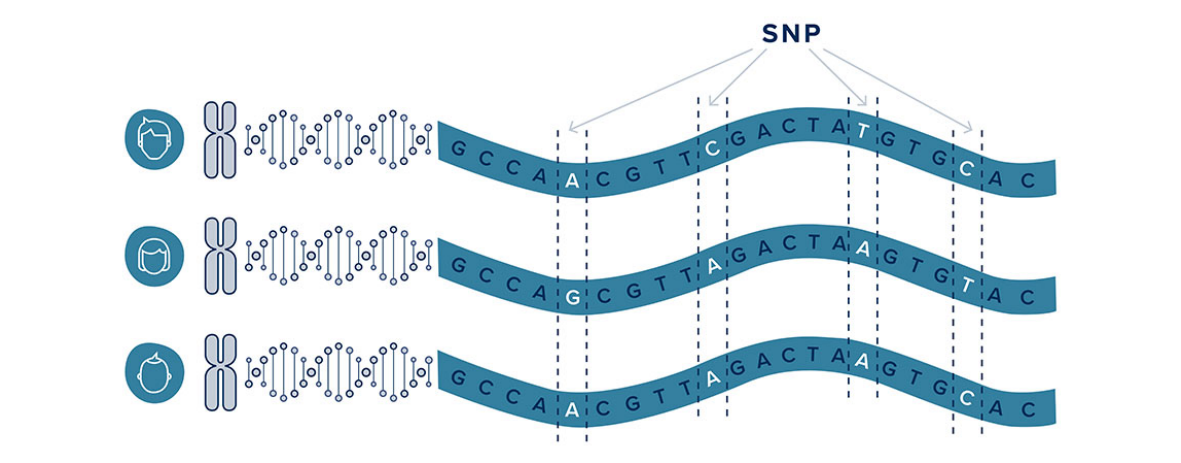
Imputation methods

Methods		Limitations
• Mode / mean / median	$\Longrightarrow$	Poor performance
• k Nearest Neighbors (kNN)	$\Longrightarrow$	Ok, but can we do better?
• Singular Value Decomposition (SVD)	$\Longrightarrow$	Ok, but can we do better?
• Multiple Imputation by Chained Equations (MICE)	$\Longrightarrow$	Too slow for high-dimensional data
• MissForest	$\Longrightarrow$	Too slow for high-dimensional data

Imputation methods

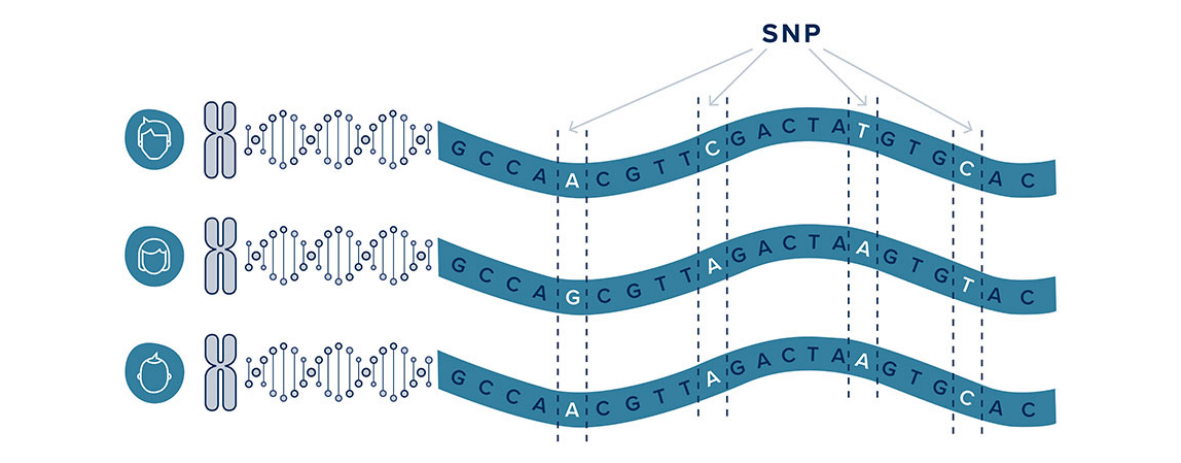
Methods		Limitations
• Mode / mean / median	$\Longrightarrow$	Poor performance
• k Nearest Neighbors (kNN)	$\Longrightarrow$	Ok, but can we do better?
• Singular Value Decomposition (SVD)	$\Longrightarrow$	Ok, but can we do better?
• Multiple Imputation by Chained Equations (MICE)	$\Longrightarrow$	Too slow for high-dimensional data
• MissForest	$\Longrightarrow$	Too slow for high-dimensional data
• Denoising Autoencoders (SCDA)	$\Longrightarrow$	Require complete data for training

Example: Single Nucleotide Polymorphisms (SNP)



- Categorical: 0 (dominant-dominant), 1 (dominant-mutant), 2 (mutant-mutant)
- High-dimensional ($10^5 - 10^6$) and low-sampled ($10^2 - 10^3$)
- Ordering is important
- Missing values occur due to external mechanisms $\implies$ MCAR

Example: Single Nucleotide Polymorphisms (SNP)

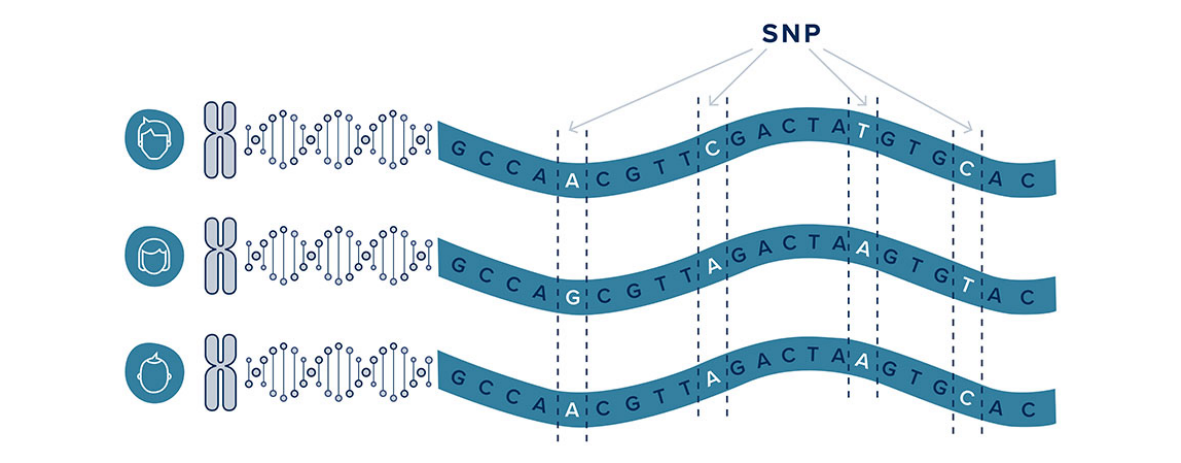


- Categorical: 0 (dominant-dominant), 1 (dominant-mutant), 2 (mutant-mutant)
- High-dimensional ($10^5 - 10^6$) and low-sampled ($10^2 - 10^3$)
- Ordering is important
- Missing values occur due to external mechanisms $\implies$ MCAR

Methods:

- reference-based (state-of-the-art for human data)
- reference-free (when reference panels are not available).

Example: Single Nucleotide Polymorphisms (SNP)



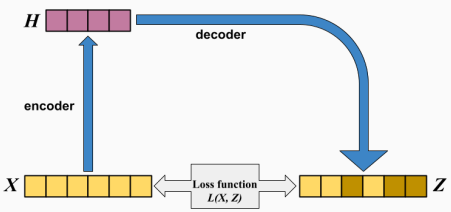
- Categorical: 0 (dominant-dominant), 1 (dominant-mutant), 2 (mutant-mutant)
- High-dimensional ($10^5 - 10^6$) and low-sampled ($10^2 - 10^3$)
- Ordering is important
- Missing values occur due to external mechanisms $\implies$ MCAR

Methods:

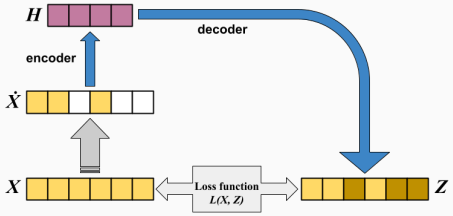
- reference-based (state-of-the-art for human data)
- reference-free (when reference panels are not available).

Other possible examples: gene expression arrays, classification problems in astronomy, tool development for finance data, and weather prediction.

Autoencoders $\Longrightarrow$ Autoreplicative Random Forests

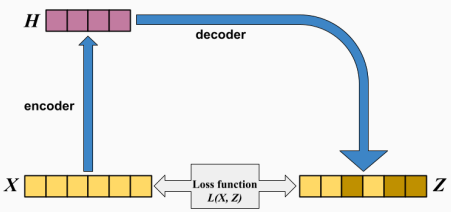


Autoencoder

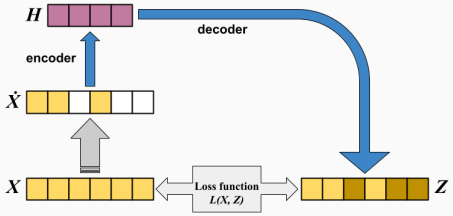


Denoising Autoencoder

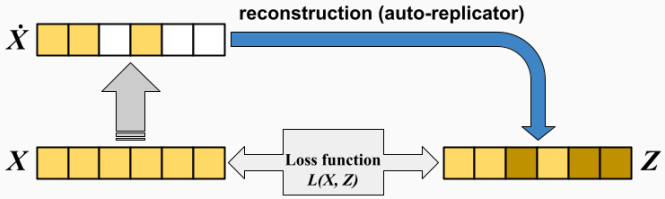
Autoencoders $\Longrightarrow$ Autoreplicative Random Forests



Autoencoder

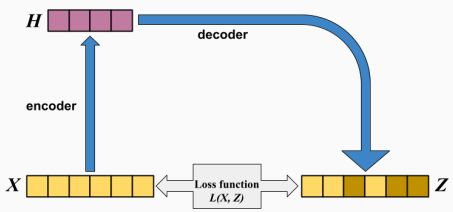


Denoising Autoencoder

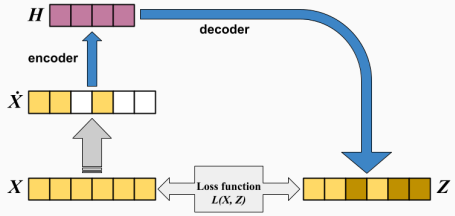


Autoreplicator

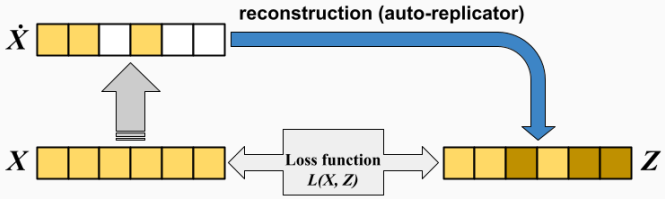
Autoencoders $\Longrightarrow$ Autoreplicative Random Forests



Autoencoder



Denoising Autoencoder

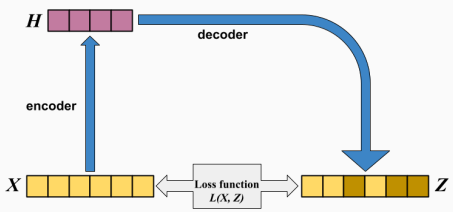


Autoreplicator

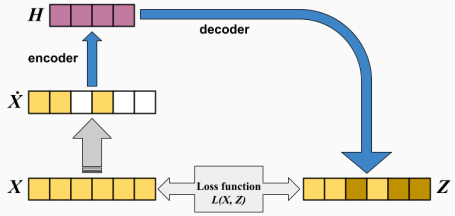
Why to use multi-label methods?

- fewer parameters than neural networks (good for low-sampled data)
- no need for hidden layers

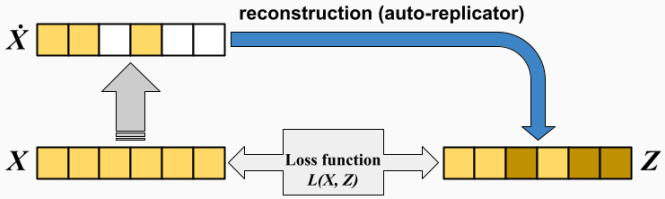
Autoencoders $\Longrightarrow$ Autoreplicative Random Forests



Autoencoder



Denoising Autoencoder



Autoreplicator

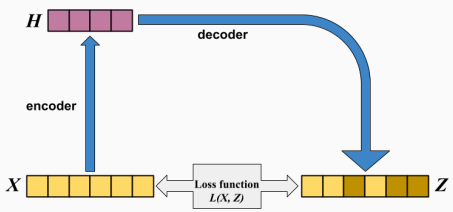
Why to use multi-label methods?

- fewer parameters than neural networks (good for low-sampled data)
- no need for hidden layers

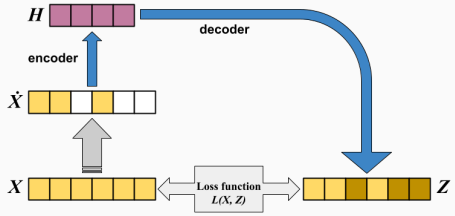
Which methods?

- Decision Trees, Random Forests
- Classifier Chains, Multilabel k Nearest Neighbours, Random k-Labelsets, Conditional Dependency Networks, etc.

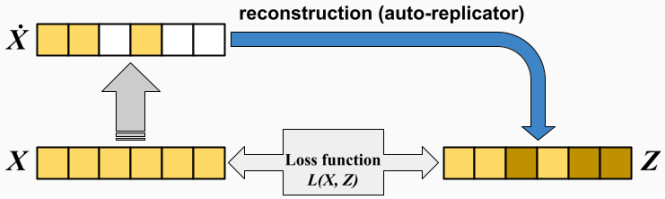
Autoencoders $\Longrightarrow$ Autoreplicative Random Forests



Autoencoder



Denoising Autoencoder



Autoreplicator

Why to use multi-label methods?

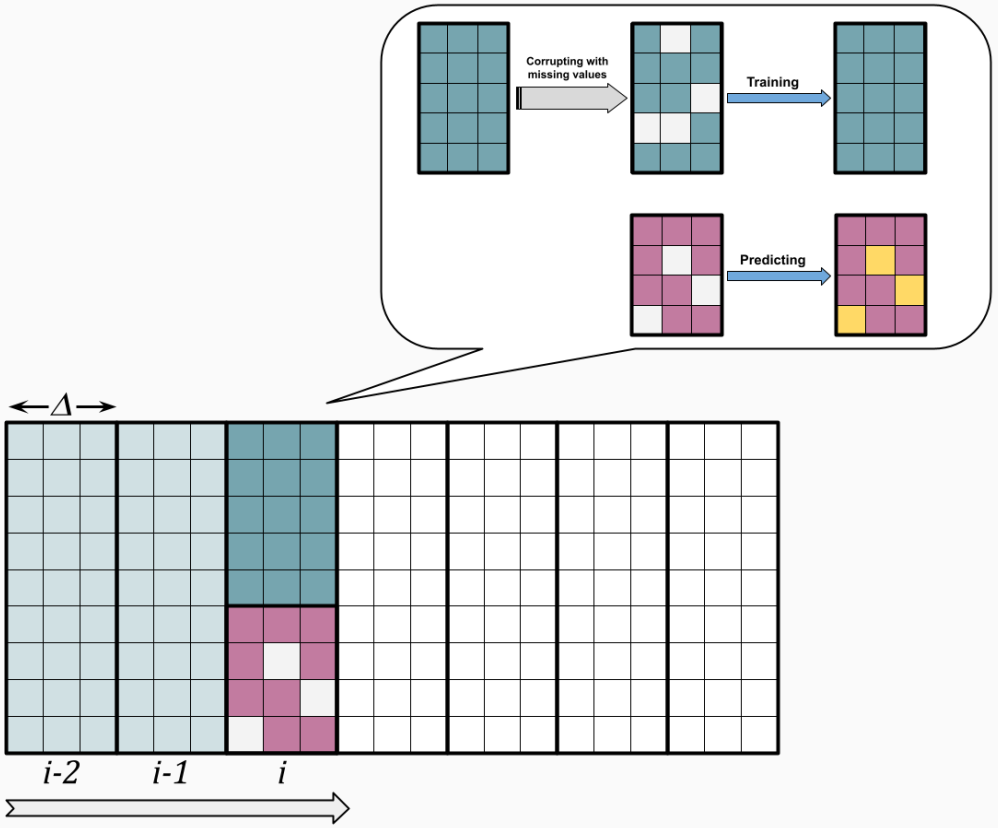
- fewer parameters than neural networks (good for low-sampled data)
- no need for hidden layers

Which methods?

- Decision Trees, Random Forests
- Classifier Chains, Multilabel k Nearest Neighbours, Random k-Labelsets, Conditional Dependency Networks, etc.

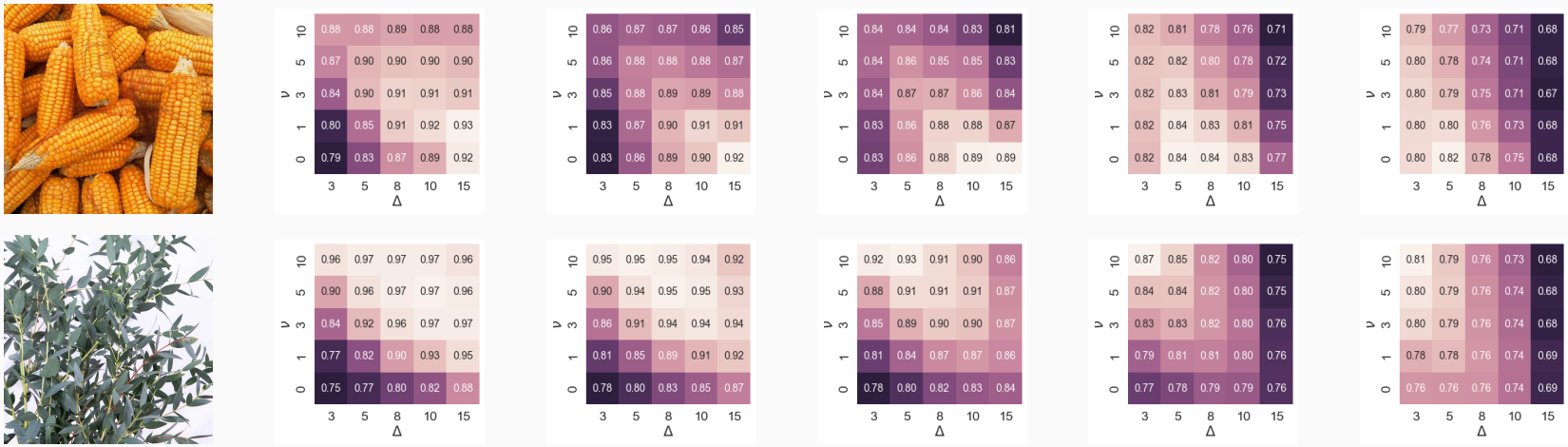
BUT... Still need complete data for training

Chains of Autoreplicative Random Forests



- One window of size Δ = training part with complete data + testing part with missing values
- Chain of windows: on each step, stacking ν windows with already imputed values as additional features
- Ensemble of chains: one forward chain, one backward chain, several random chains

Gridsearch for parameters Δ and ν



Lighter color / higher accuracy

Δ: as expected, bigger fraction of missing values $\rightarrow$ smaller size of window $\Longrightarrow$ can be estimated theoretically, no need for search

ν: may be different

Accuracy

	0.01	0.05	0.1	0.2	0.3	0.01	0.05	0.1	0.2	0.3
	Maize					Eucalyptus				
ChARF	$\Delta = 15^*$ $\nu = 1^*$	$\Delta = 15^*$ $\nu = 1^*$	$\Delta = 10^*$ $\nu = 1^*$	$\Delta = 5^*$ $\nu = 1^*$	$\Delta = 5^*$ $\nu = 1^*$	$\Delta = 10^*$ $\nu = 5^*$	$\Delta = 5^*$ $\nu = 10^*$	$\Delta = 5^*$ $\nu = 10^*$	$\Delta = 3^*$ $\nu = 10^*$	$\Delta = 3^*$ $\nu = 10^*$
	0.952	**0.935**	**0.916**	**0.882**	**0.845**	**0.970**	**0.950**	**0.926**	**0.866**	0.810
kNN (5/10)	0.803	0.802	0.801	0.798	0.794	0.851	0.849	0.847	0.843	**0.839**
mode	0.727	0.727	0.726	0.727	0.726	0.725	0.732	0.731	0.730	0.729
SVD (50/500)	0.647	0.648	0.645	0.643	0.636	0.788	0.788	0.788	0.785	0.780
MICE	–	–	–	–	–	–	–	–	–	–
missForest	0.662	0.650	0.622	0.593	0.580	0.684	0.673	0.626	0.564	0.521

	Colorado Beetle					Arabica Coffee				
ChARF	$\Delta = 10^*$ $\nu = 1^*$	$\Delta = 10^*$ $\nu = 1^*$	$\Delta = 5^*$ $\nu = 1^*$	$\Delta = 5^*$ $\nu = 1^*$	$\Delta = 3^*$ $\nu = 1^*$	$\Delta = 15^*$ $\nu = 3^*$	$\Delta = 10^*$ $\nu = 3^*$	$\Delta = 5^*$ $\nu = 5^*$	$\Delta = 3^*$ $\nu = 10^*$	$\Delta = 3^*$ $\nu = 3^*$
	0.835	**0.824**	**0.818**	**0.805**	**0.792**	**0.897**	**0.886**	**0.878**	**0.866**	0.854
kNN (50/10)	0.765	0.763	0.765	0.765	0.764	0.867	0.866	0.866	0.865	**0.864**
mode	0.761	0.760	0.762	0.761	0.761	0.807	0.804	0.805	0.805	0.804
SVD (50/100)	0.740	0.737	0.737	0.735	0.734	0.693	0.694	0.696	0.692	0.690
MICE	–	–	–	–	–	0.757	0.741	0.724	0.689	0.664
missForest	0.352	0.349	0.361	0.326	0.335	0.497	0.480	0.533	0.541	0.586

	Wheat					Coffea Canephora				
ChARF	$\Delta = 8^*$ $\nu = 10^*$	$\Delta = 5^*$ $\nu = 10^*$	$\Delta = 5^*$ $\nu = 10^*$	$\Delta = 3^*$ $\nu = 10^*$	$\Delta = 3^*$ $\nu = 10^*$	$\Delta = 10^*$ $\nu = 1^*$	$\Delta = 10^*$ $\nu = 1^*$	$\Delta = 5^*$ $\nu = 1^*$	$\Delta = 5^*$ $\nu = 1^*$	$\Delta = 3^*$ $\nu = 1^*$
	0.821	0.808	0.795	0.777	0.762	**0.799**	**0.781**	**0.761**	0.731	0.717
kNN (10/10)	**0.823**	**0.819**	**0.818**	**0.815**	**0.811**	0.737	0.739	0.737	**0.734**	**0.731**
mode	0.729	0.727	0.729	0.729	0.727	0.691	0.693	0.692	0.692	0.691
SVD (200/50)	0.622	0.618	0.609	0.600	0.594	0.456	0.453	0.450	0.449	0.450
MICE	0.641	0.635	0.621	0.585	0.545	–	–	–	–	–
missForest	0.614	0.736	0.746	0.756	0.755	0.377	0.449	0.442	0.395	0.383

- MICE: run with 10 neighbors for each feature, still worked only for smaller data
- MissForest: run for first 100 features only
- SCDA: not taken into comparison (no complete data for training)

Time complexity

In theory:

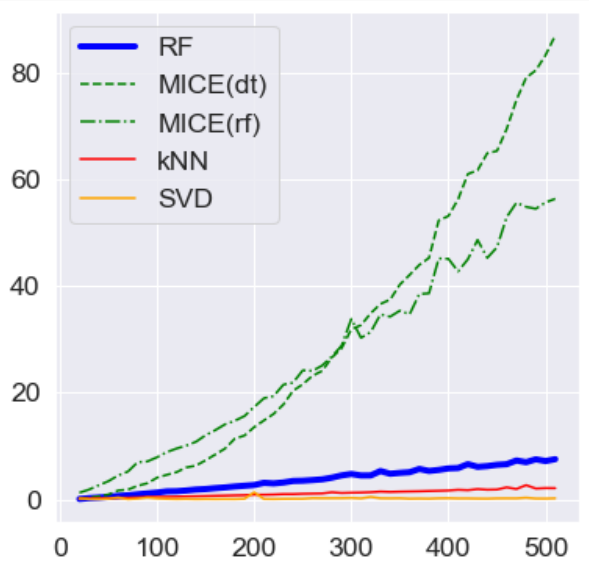
kNN	SVD	MICE(dt)	MICE(rf)	**ChARF**
$\mathcal{O}(p)$	$\mathcal{O}(p)$	$\mathcal{O}(p^2)$	$\mathcal{O}(p\sqrt{p})$	$\mathcal{O}(\frac{p}{\Delta} \cdot \Delta)$

Time complexity

In theory:

kNN	SVD	MICE(dt)	MICE(rf)	**ChARF**
$\mathcal{O}(p)$	$\mathcal{O}(p)$	$\mathcal{O}(p^2)$	$\mathcal{O}(p\sqrt{p})$	$\mathcal{O}(\frac{p}{\Delta} \cdot \Delta)$

In practice:



Eucalyptus dataset, first 10, 20, ..., 500 features

Conclusions

- Unusual and effective usage of multi-label methods, e.g. Random Forests:
  - autoreplication
  - missing value imputation
  - denoising
  - outlier detection

Conclusions

- Unusual and effective usage of multi-label methods, e.g. Random Forests:
  - autoreplication
  - missing value imputation
  - denoising
  - outlier detection
- ChARF is an effective method for missing value imputation in high-dimensional data
  - lower time complexity $\implies$ works for high-dimensional datasets
  - no need for complete data

Future work

- Studies for MAR and MNAR scenarios
- Regression (e.g. gene expression)
- Iterative approach (see MICE/missForest) + Autoreplicative Random Forests
- Probabilistic interpretation

Thank you!