CLEMI

CLustering to Evaluate Multiple Imputation

Anthony Chapman

Dr Steve Turner Dr Wei Pang Dr Lorna Aucott

Dept. of Applied Medical Sciences, University of Aberdeen
Dept. of Computing Science, University of Aberdeen
e-mail: r01ac14@abdn.ac.uk

Outline

1. Introduction
2. Imputation
3. Benchmark
4. Evaluating Imputation
5. Discussion

Content

1 Introduction

2 Imputation

3 Benchmark

4 Evaluating Imputation

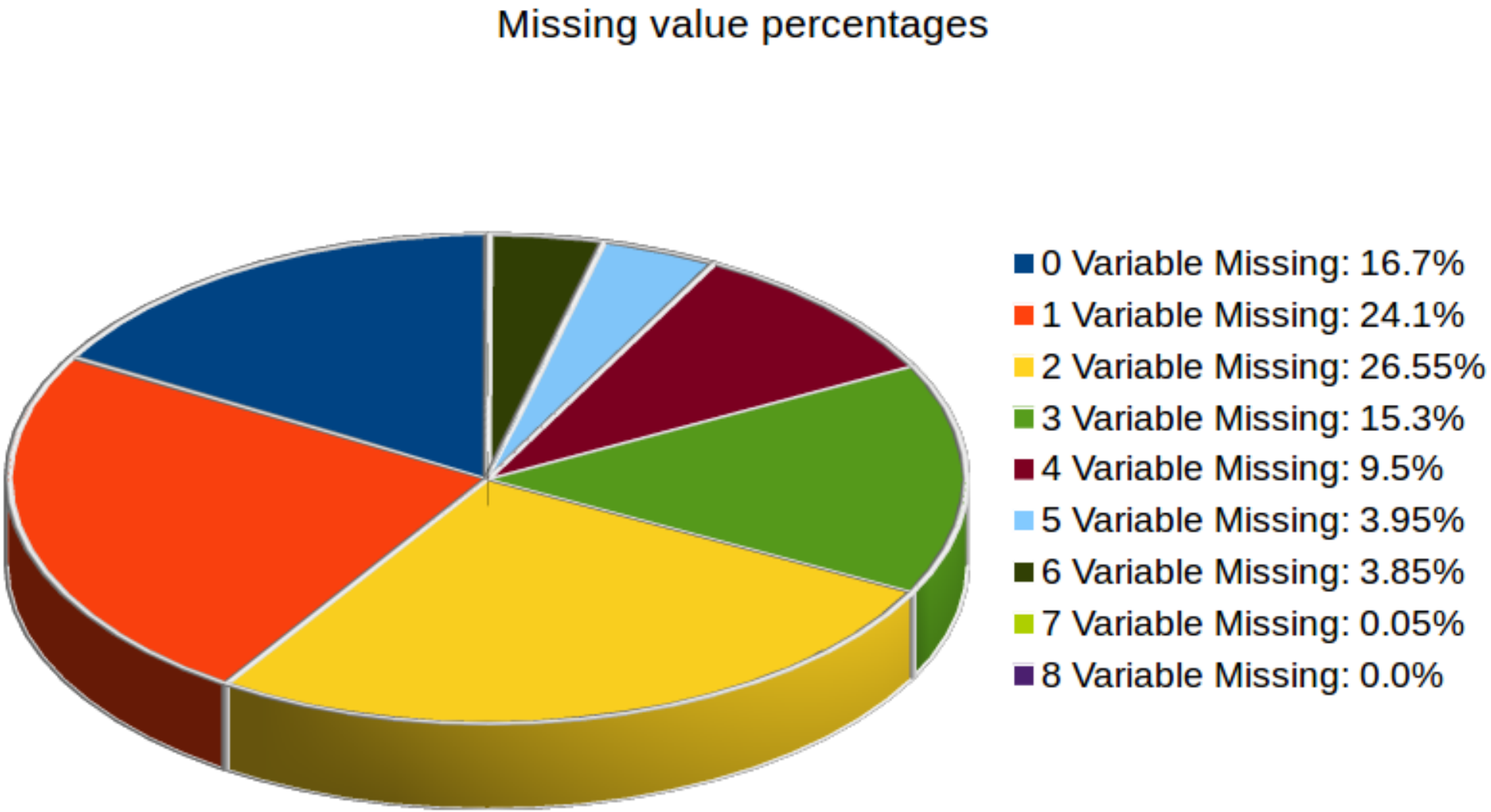
5 Discussion

Introduction

Motivation:

- Routinely acquired data has large amounts of missing data
- Most researchers carry out complete case analyses
- Need to use as much of the available data as possible.
- Need something every researcher can trust
- A way to evaluate imputation
- Must be user friendly to most researchers (no need for a computing degree)

Missing Values in Raw Data



Introduction

Motivation:

- Routinely acquired data has large amounts of missing data
- Most researchers carry out complete case analyses
- Need to use as much of the available data as possible.
- Need something every researcher can trust
- A way to evaluate imputation
- Must be user friendly to most researchers (no need for a computing degree)

Current Work

Imputation:

- Statistical Software (SPSS, R, StatSol)
- Mean Impuation, Multiple Imputation
- MICE: Multivariate Imputation by Chained Equations

Evaluation:

- No generalisation
- Nothing
- Zilch

Content

1 Introduction

2 Imputation

3 Benchmark

4 Evaluating Imputation

5 Discussion

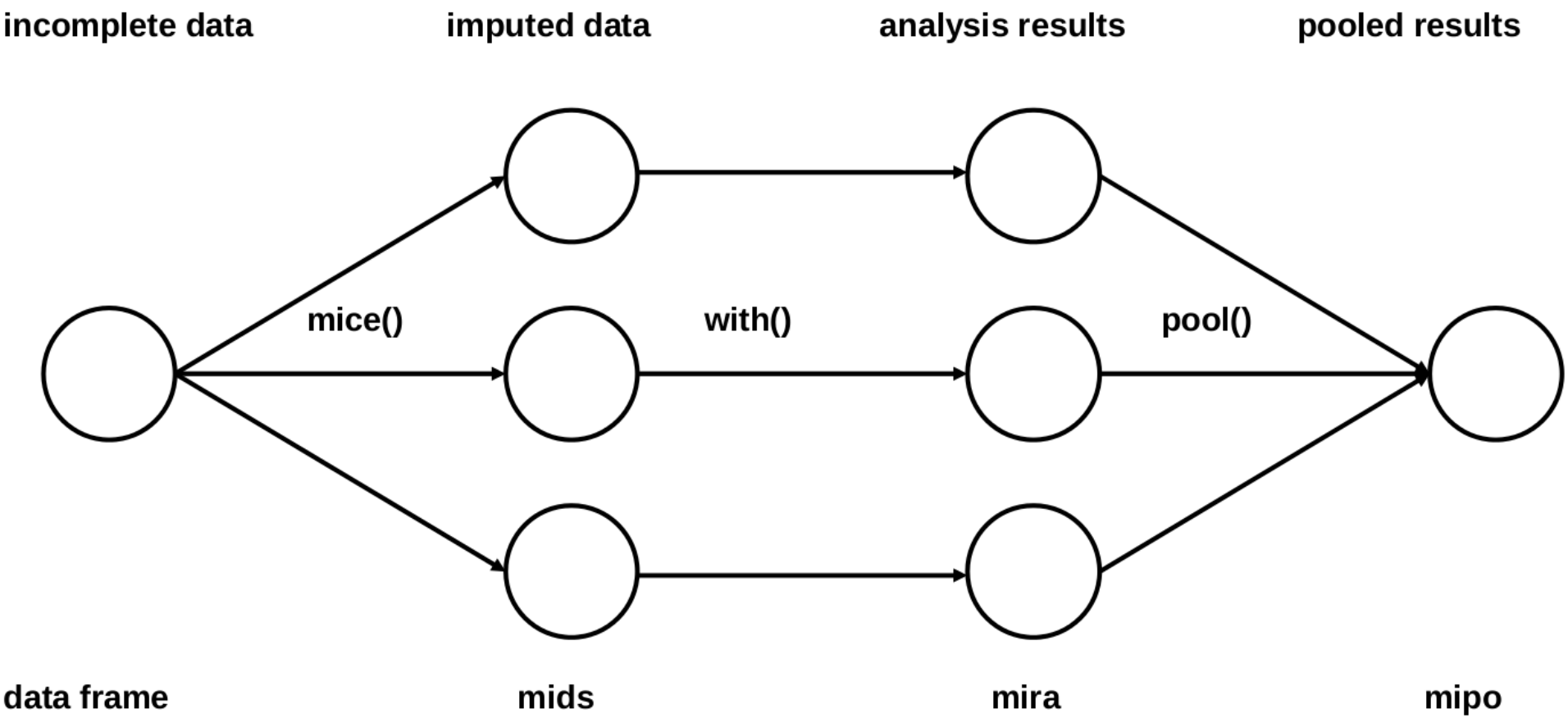
Imputation - MICE

MICE - Multivariate Imputation by Chained Equations

- Uses the whole dataset
- Preserved the relations in the data
- Can work with longitudinal data

Imputation ctn.

Journal of Statistical Software



Content

1 Introduction

2 Imputation

3 Benchmark

4 Evaluating Imputation

5 Discussion

Benchmarks

- Needed to see effects of imputation
- Needed for a controlled test
- They are suppose to represent the truth

Benchmarks

How to create a benchmark:

- Extract complete cases
- Analyse missingness in original dataset
- Create copies of the complete cases and apply missingness
- Every mini-me is a replica of the original dataset but from the benchmark.

Content

1 Introduction

2 Imputation

3 Benchmark

4 Evaluating Imputation

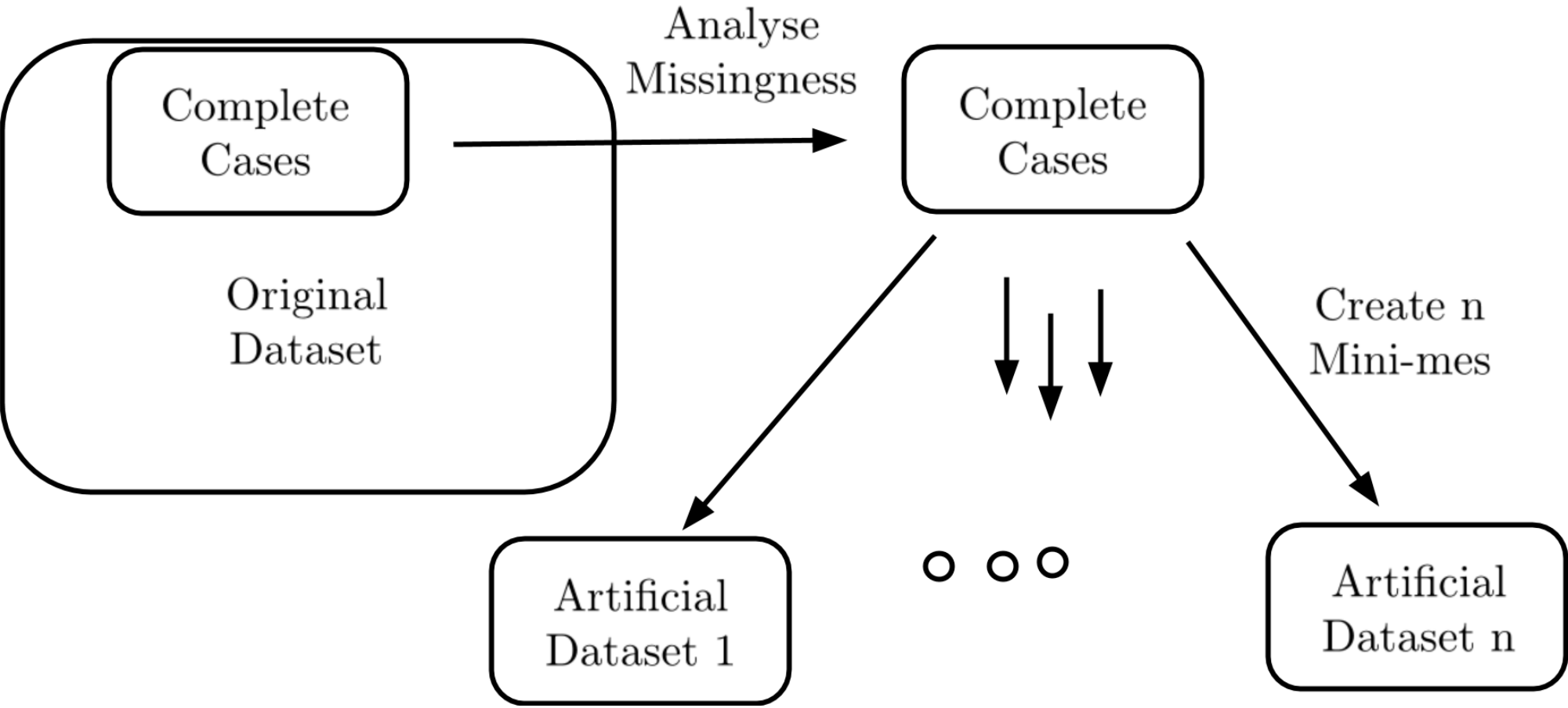
5 Discussion

Impute Artificially Incomplete Datasets

Apply MICE to all mini-me:

- Need to minimise uncertainty
- Apply to multiple datasets
- Exact same method on all datasets

Re-Cap

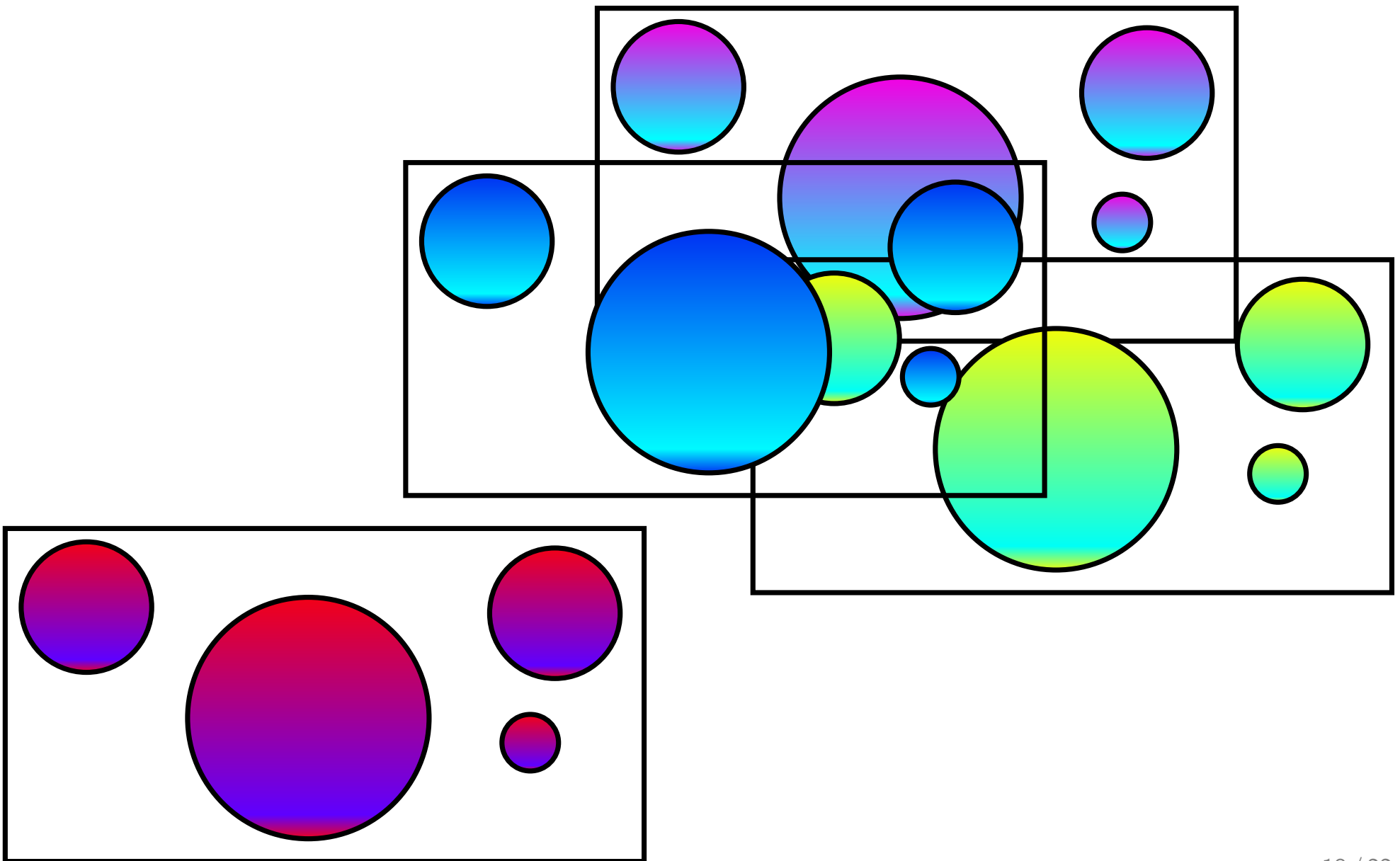

Complete Cases
Original Dataset
Analyse Missingness
Complete Cases
Create n Mini-mes
Artificial Dataset 1
Artificial Dataset n

Evaluation

Clustering:

- Group objects into a sets with similar objects
- Unsupervised
- Good for higher dimensional data (Big Data)

Evaluation ctn.



Evaluation ctn.

Clustering:

- Get relevant clustering characteristics
- Compare imputed datasets to benchmark to see the effects
- Mean imputation as a reference

Content

1 Introduction

2 Imputation

3 Benchmark

4 Evaluating Imputation

5 Discussion

Discussion & Conclusion

Limitations

- Output is subjective
- Some may over-interpret the results
- What if the complete subset is too small

Outcomes

- Optimised number of ignored records
- Compare different imputation methods
- Optimize imputation features

To Consider

- Use a modelling to verify the outcome
- Use any imputation method

Thanks & Questions

Thanks for your attention!
Question & Comments