QUEEN'S UNIVERSITY BELFAST

SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Data Analysis & Visualisation

CSC3062

BEng (CS & SE), MEng (CS & SE), BIT & CIT

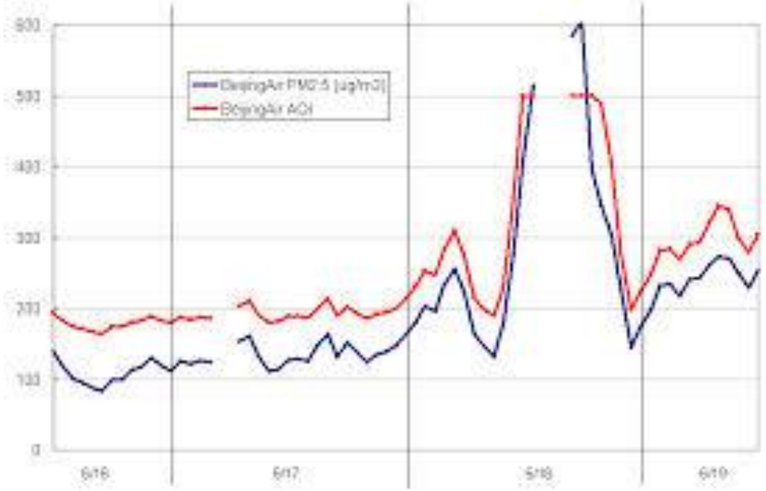
Dr Reza Rafiee

Semester 1 – 2019/2020

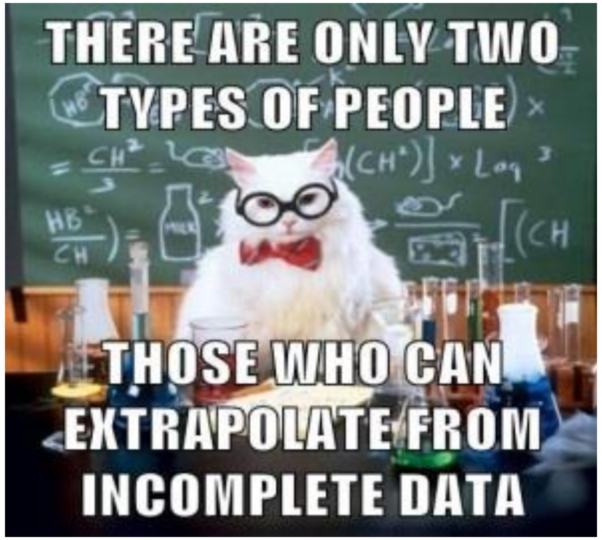
What to Do with Missing Data?

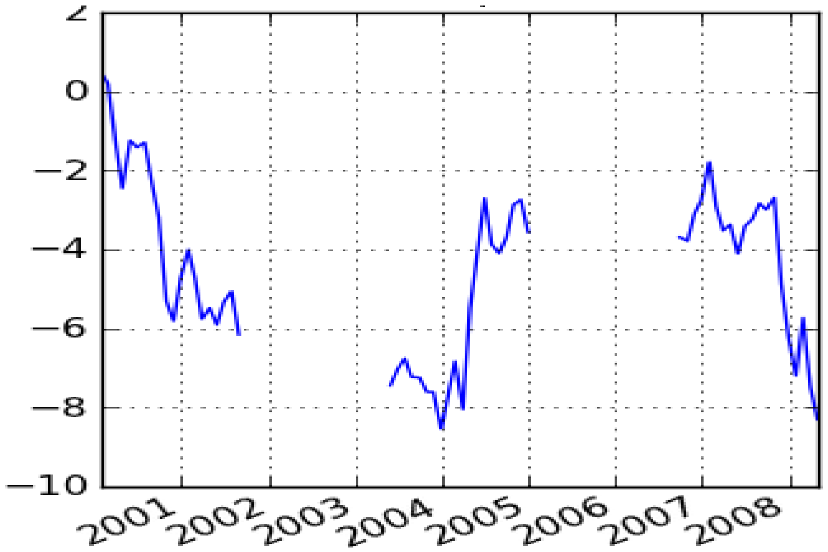
SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Missing data is everywhere



In almost any research you perform, or any data analysis, there is the potential for missing or incomplete data.

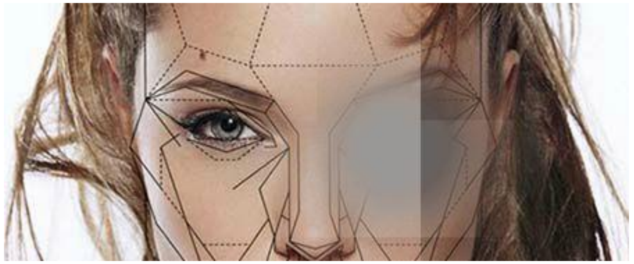




Missing data is everywhere

$$\begin{bmatrix}
1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & & & 0 & 0 & 1 & 0 & 0 & 1 & & ? & & 0 & 0 & 1 & 0 \\
1 & 1 & 0 & 0 & ? & & 1 & 0 & 1 & 0 & 0 & 0 & & & & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 1 & & & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & ? & & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & & & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1
\end{bmatrix}$$

$$\begin{array}{cc}
\begin{array}{ccc} 1 & \ldots & 103 \end{array} & \\
D = \begin{bmatrix} 0.9 & \cdots & ? \\ \vdots & \ddots & \vdots \\ ? & \cdots & 0.1 \end{bmatrix} & \begin{array}{c} 1 \\ \ldots \\ 17 \end{array}
\end{array}$$



Two samples including missing data

	Sample 1	Sample 2
cg00583535	NA	0.317394283
cg18788664	1	0.192024985
cg08123444	0.532659205	0.867010408
cg17185060	0.774338632	0.70392815
cg04541368	0.079894678	0.659468157
cg25923609	0.109138594	0.600225461
cg06795768	0.04605561	0.870753578
cg19336198	0.713845623	0.707326444
cg05851505	NA	0.981375746
cg20912770	0.039837473	0.0646352
cg09190051	1	0.336904134
cg01986767	NA	NA
cg01561259	0.133410152	0.113869472
cg12373208	NA	0.04628476
cg24280645	0.163157983	0.088281769
cg00388871	0.239179168	0.308942014
cg09923107	0.091227524	0.121433558

Categories of missingness

- Failure in:
  - Responding to a question (in surveys)
  - Equipment (Sensors), Recording Mechanisms
  - Data entry
  - …

Missing at Random (MAR)

Missing Completely at Random (MCAR)

Missing Not at Random (MNAR)

The probability that a value is missing depends only on observed values.

the missingness cannot be predicted from any other variables or sets of variables

Categories of missingness

Missing at Random (MAR)

Missing Completely at Random (MCAR)

Missing Not at Random (MNAR)

The probability that a value is missing depends only on observed values.

- Lab tubes that broke
- Forms that got lost
- Interviewer forgot to ask

The missingness cannot be predicted from any other variables or sets of variables

Missing at random (MAR)

- Assumption in Missing At Random (MAR): The probability that a value is missing depends only on observed values. The missing data occurred randomly but **that the pattern of missing data can be predicted from the existing data.**

Two other types of missing data: When the missingness cannot be predicted from any other variables or sets of variables, we called **MCAR** (Missing Completely At Random). **MNAR**: missing NOT at random.

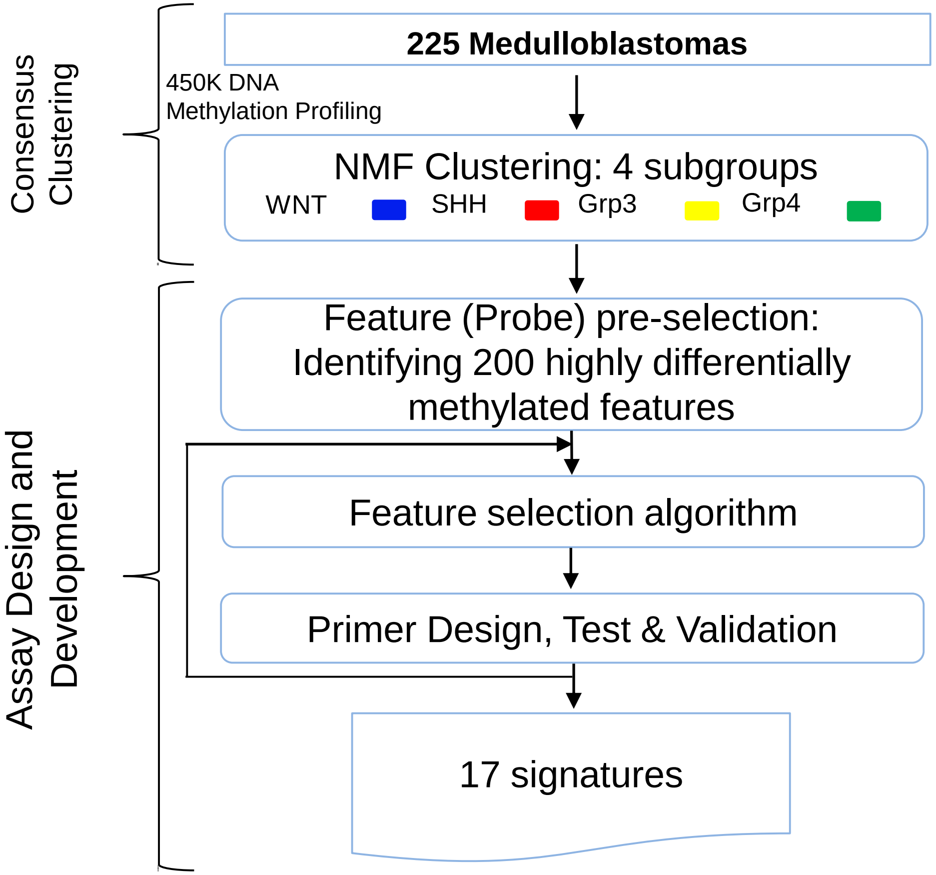
Schafer JL, Graham JW. Missing data: Our view of the state of the art. Psychological Methods. 2002; 7:147–177. [PubMed: 12090408]

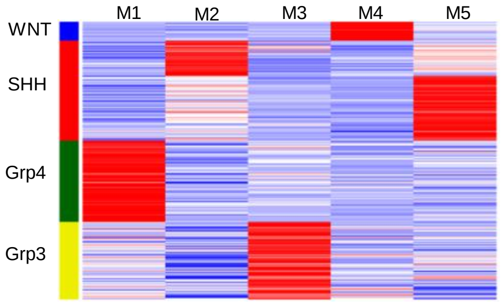
Package/library in R

- ‘Amelia’: Bootstrap + EM
- ‘mice’: Multivariate Imputation using Chained Equations
- ‘mi’: Multiple Imputation using an approximate Bayesian framework

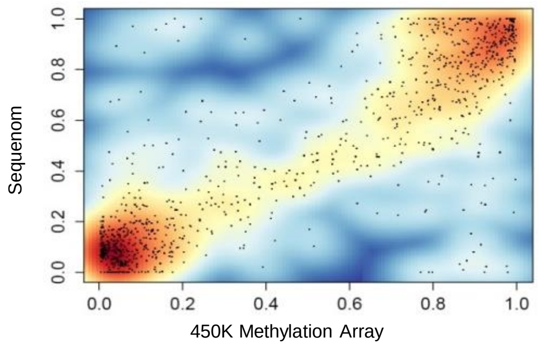
1) Diagnostics of the models
2) Provides graphics to visualize missing data patterns
3) Provides degree of sampling uncertainty
4) Applicable for both numerical and categorical data

Example: a biological assay development





Heatmap representation of metagene projection of most variably methylated probes in 450K methylation array data (samples, n=225).

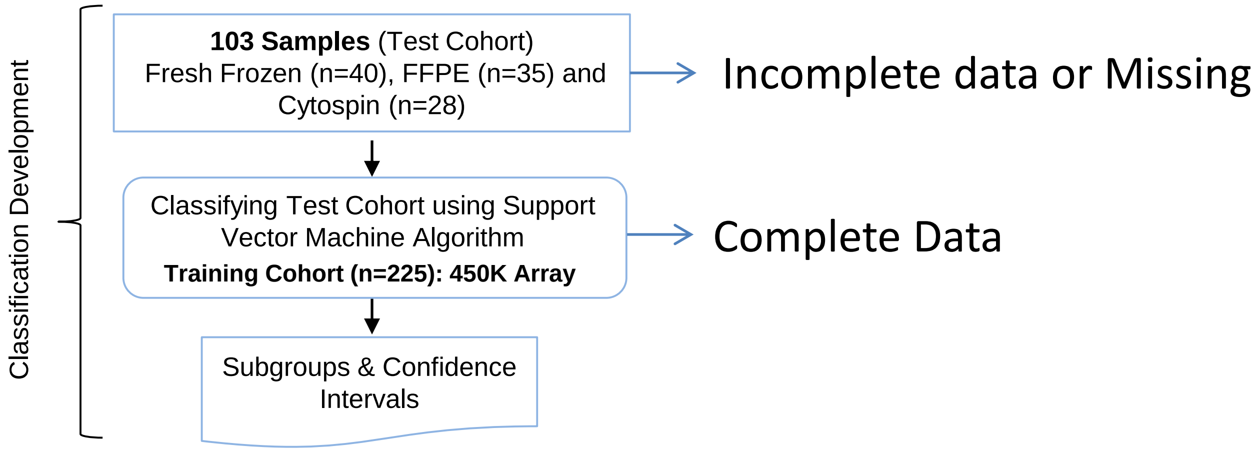


Correlation between Sequenom and 450K Methylation array's signatures (17 probes).

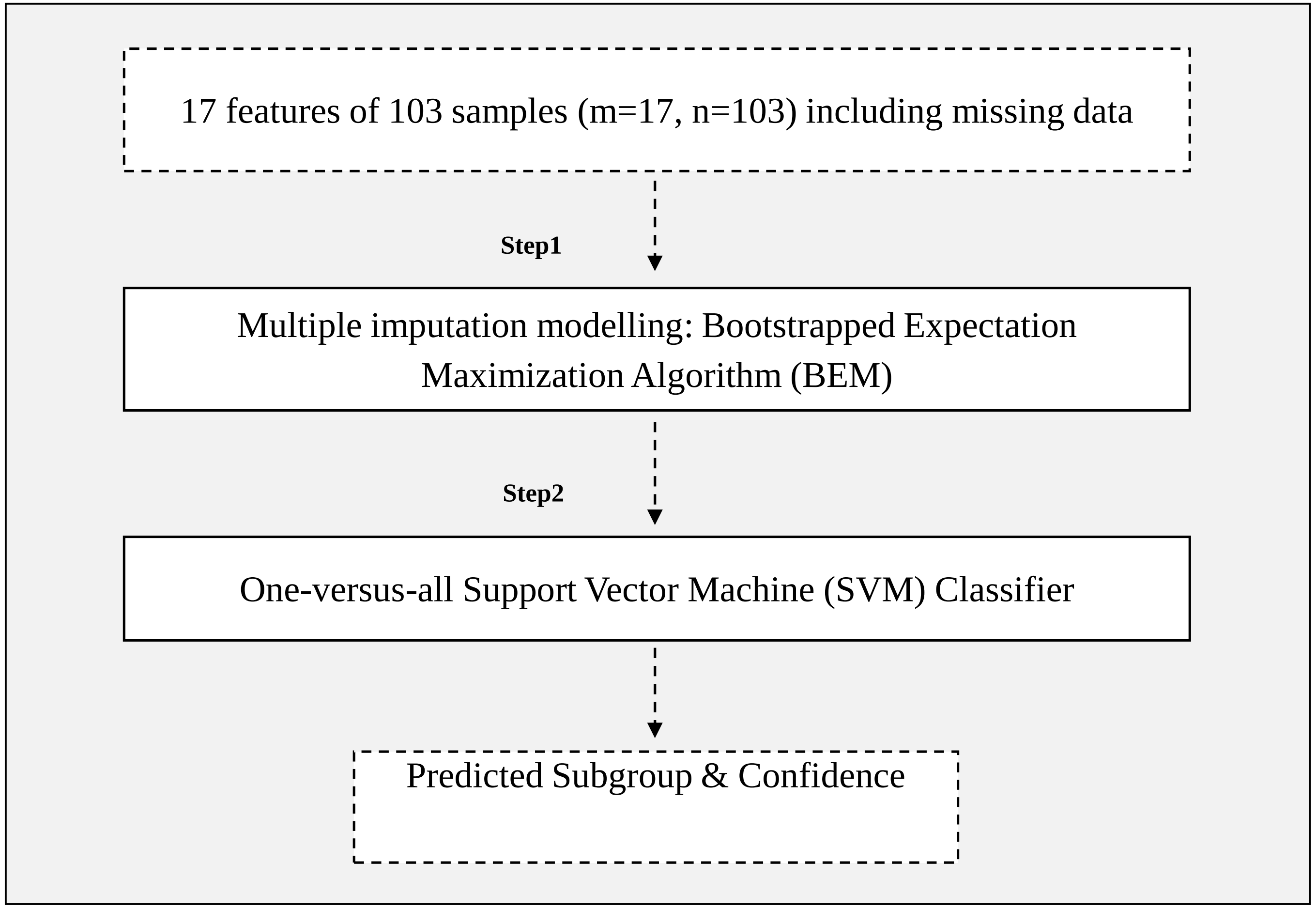
QUEEN'S UNIVERSITY BELFAST

SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Missing data in sequenom assay



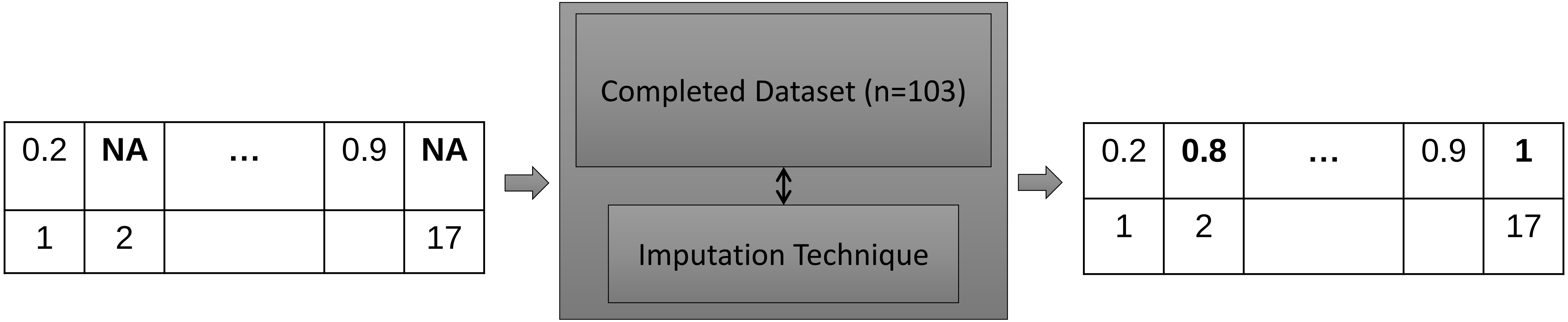
Example of addressing missing data



QUEEN'S UNIVERSITY BELFAST
SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Example of addressing missing data

Pre-processing step



Input csv file with missing

output csv file – completed data

The origin of missing

- Due to the failure during the quality control steps of a technique (i.e., bisulphite conversion, test PCR, and multiplex PCR stages), **missing values** appear in a number of probes for each individual sample. Therefore, the input dataset includes incomplete data for a number of probes.

How to visualise missing?

Heatmap representation of 103 samples including missing. Blue colour is showing missing for number of features.

Pie chart: percentage of missing for each feature

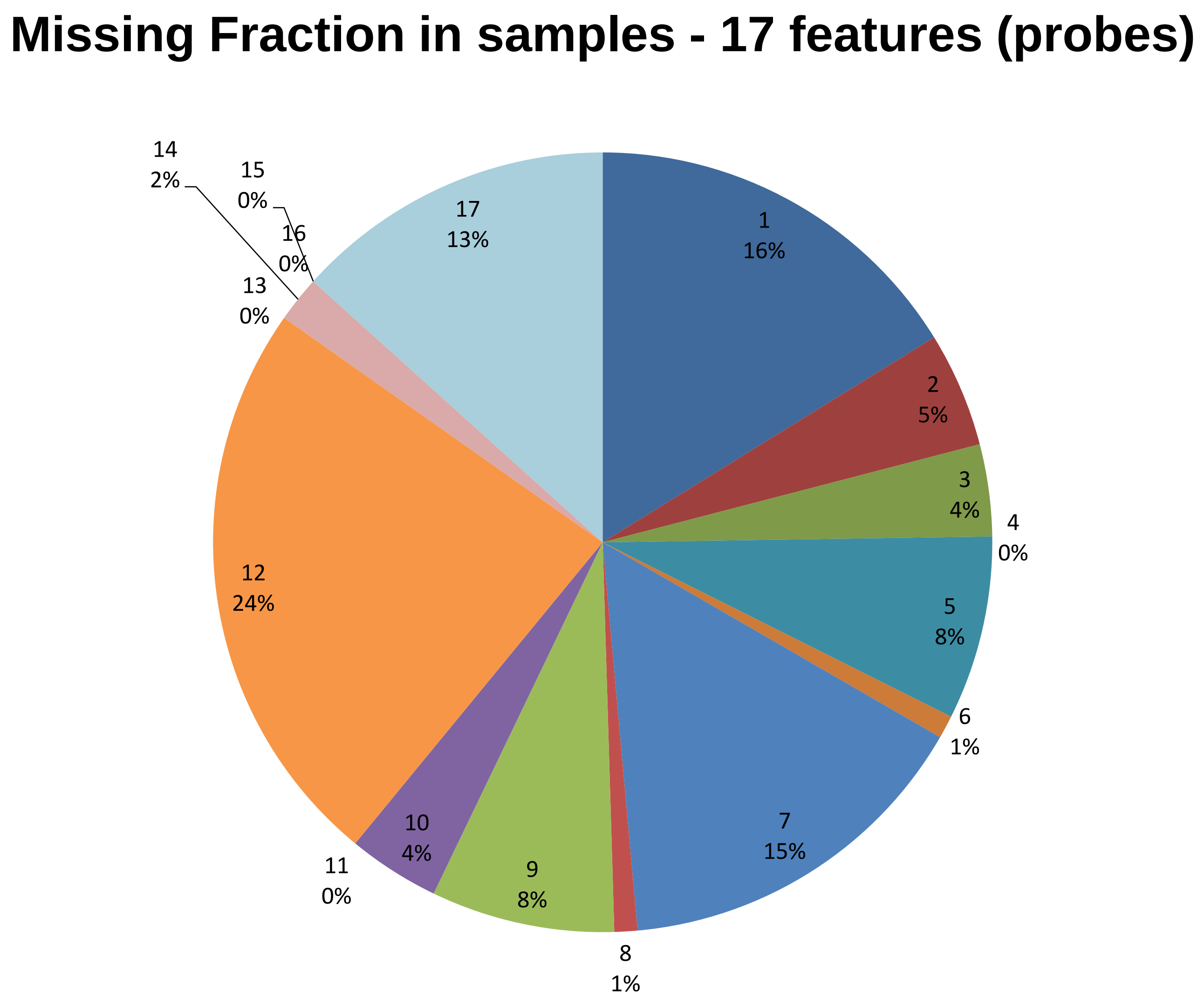
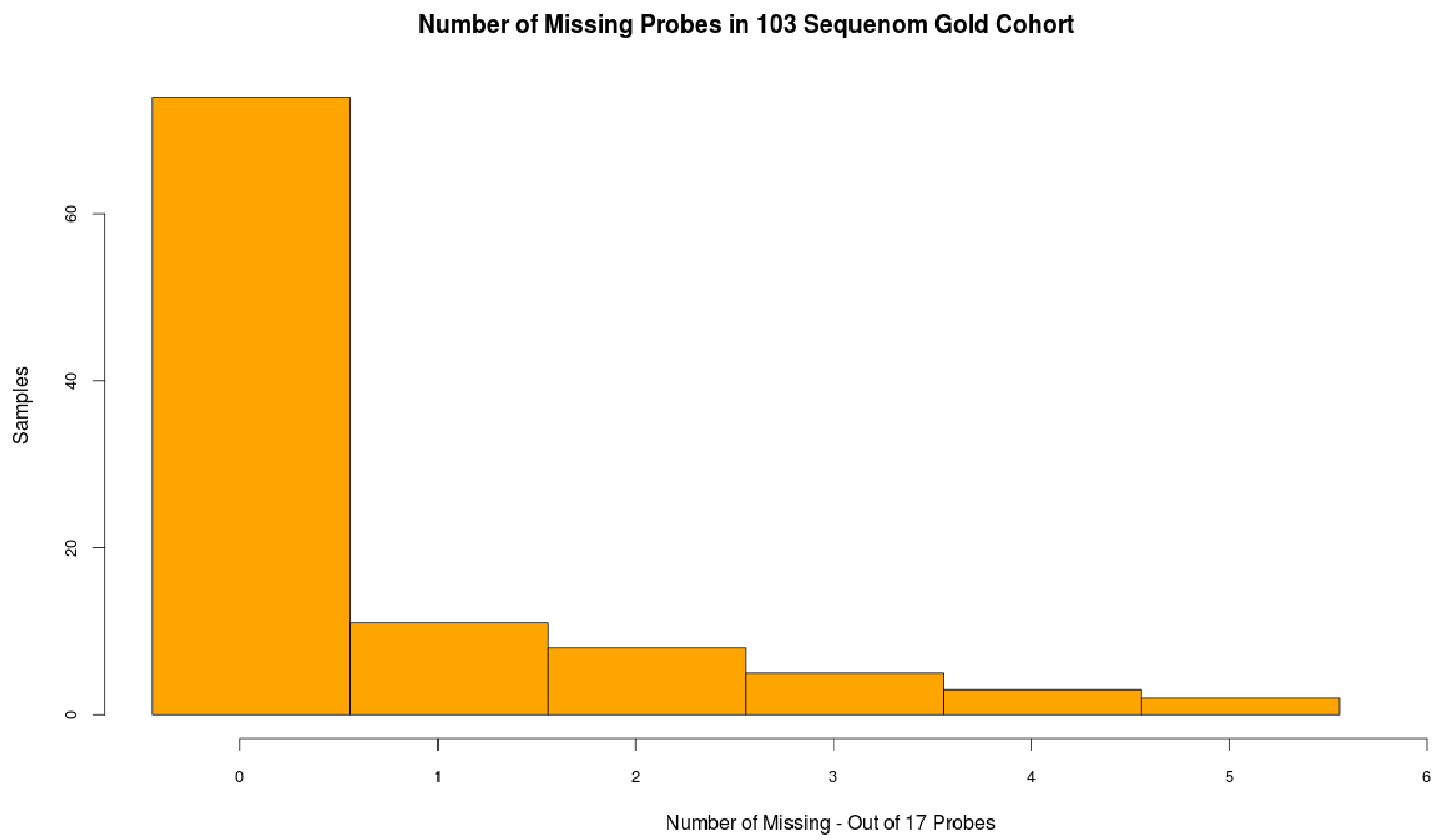


Illustration by histogram



QUEEN'S UNIVERSITY BELFAST

SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Illustration in tabular format

Fraction of missing for each individual feature

1	cg00583535	0.165048544
2	cg18788664	0.048543689
3	cg08123444	0.038834951
4	cg17185060	0.000000000
5	cg04541368	0.077669903
6	cg25923609	0.009708738
7	cg06795768	0.155339806
8	cg19336198	0.009708738
9	cg05851505	0.077669903
10	cg20912770	0.038834951
11	cg09190051	0.000000000
12	cg01986767	0.242718447
13	cg01561259	0.000000000
14	cg12373208	0.019417476
15	cg24280645	0.000000000
16	cg00388871	0.000000000
17	cg09923107	0.135922330

What to do with missing features?

$$D = \begin{array}{cccc} 1 & \dots & 103 & \\ \begin{bmatrix} 0.9 & \cdots & ? \\ \vdots & \ddots & \vdots \\ ? & \cdots & 0.1 \end{bmatrix} & & & \begin{array}{l} \text{Feature 1} \\ \dots \\ \text{Feature 17} \end{array} \end{array}$$

$$D = \{D_{obs}, D_{mis}\}$$

Missing feature is unobserved but do "exist" in *a specific metaphysical sense*

This means by repeating the technique (by which original data has been collected) on this sample it might be observed.

Multiple Imputation Modelling and Diagnostics

What is multiple imputation?

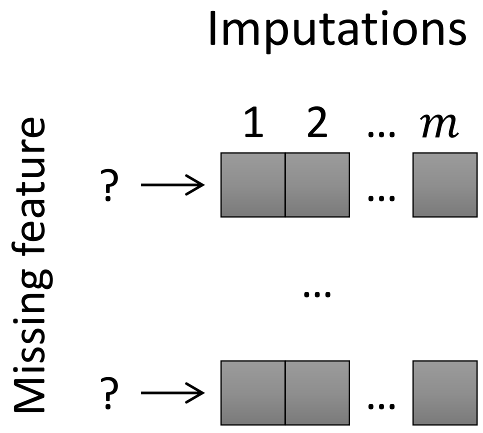
- This statistical technique (algorithm) takes the incomplete dataset (i.e., including missing data) and returns $\boldsymbol{m}$ imputed datasets with no missing values.

$\boldsymbol{m}$ is a user-selected parameter

Multiple imputation

- Each missing feature is imputed (filled in) with a set of $m > 1$ plausible values which reflect the uncertainty about the missing feature.

$$\begin{array}{r c} & \begin{array}{ccc} 1 & \dots & 103 \end{array} \\ \begin{array}{r} \text{Feature 1} \\ \dots \\ \text{Feature 17} \end{array} & \begin{bmatrix} 0.9 & \cdots & ? \\ \vdots & \ddots & \vdots \\ ? & \cdots & 0.1 \end{bmatrix} \end{array}$$

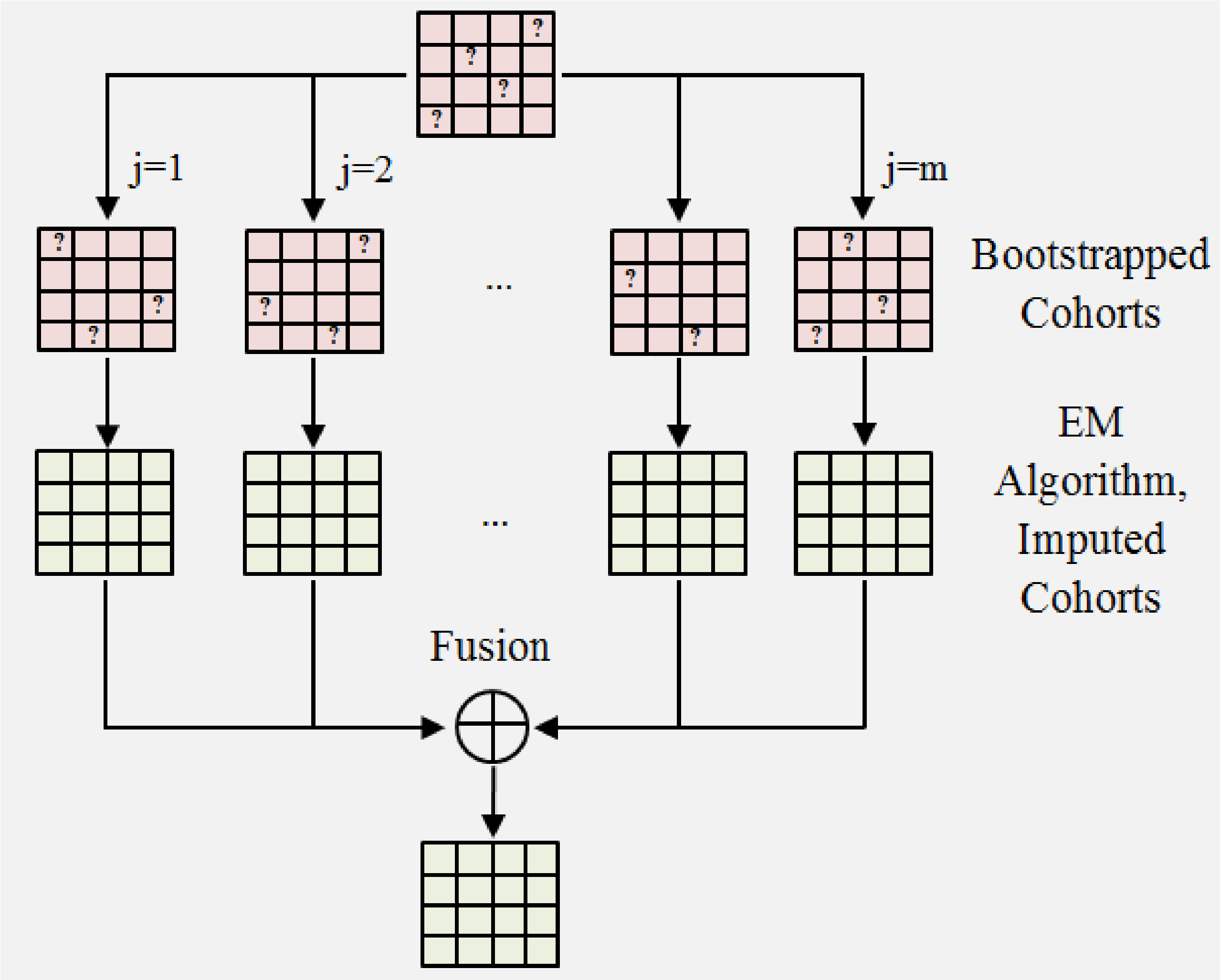


Multiple imputation modelling techniques

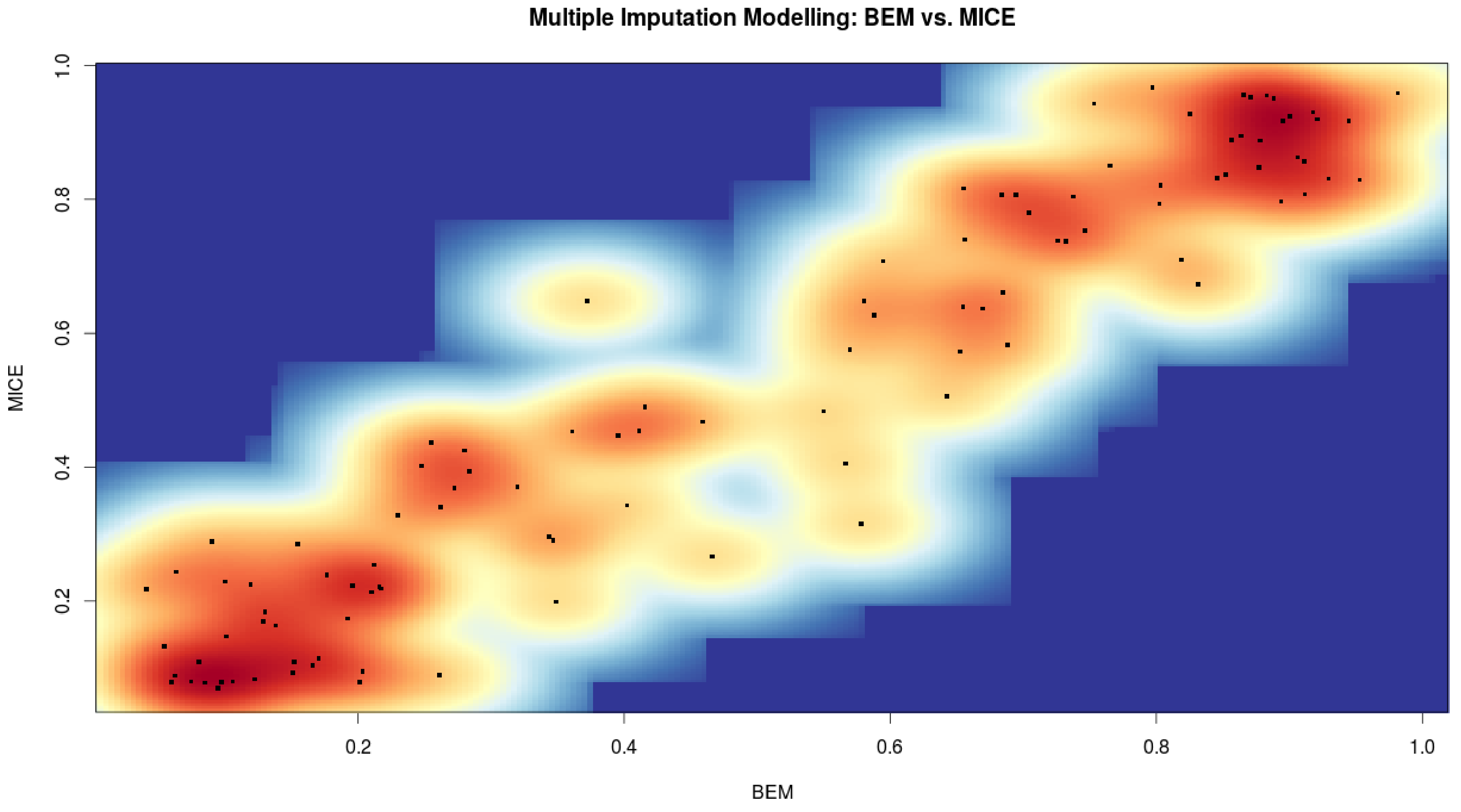
- Multivariate Imputation by Chained Equations (MICE)
- Bootstrapped Expectation-Maximisation (BEM)
- Multiple Imputation using an approximate Bayesian framework (MI)

Multiple imputation modelling techniques

- Bootstrapped Expectation-Maximisation (BEM)



Visualising BEM vs. MICE using scatter plot



Impact of a method (BEM or MICE)

Using different multiple imputation methods may affect the final results (e.g., classification results)

BEM

Predicted Subgroup		Reference subgroup: WNT	SHH	Grp 3	Grp 4
	WNT	22	0	0	0
	SHH	0	23	0	0
	Grp 3	0	0	23	0
	Grp 4	0	0	0	28
	NC+	2	4	1	0
	Total	24	27	24	28

MICE

Predicted Subgroup		Reference subgroup: WNT	SHH	Grp 3	Grp 4
	WNT	22	0	0	0
	SHH	0	22	0	0
	Grp 3	0	1	23	0
	Grp 4	0	0	0	28
	NC	2	5	0	0
	Total	24	28	23	28

Summarising the performance of a classification algorithm using a "confusion matrix". A matrix (table) shows the discrepancy between predicted and reference subgroup.

+NC: Non-classifiable

Efficiency of Multiple Imputation

- Efficiency of an estimate based on m imputation is approximately:

$$(1 + \frac{\gamma}{m})^{-1}$$

Where γ is the fraction of missing information for the quality being estimated.

1) Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys . J. Wiley & Sons, New York.
2) Schafer, Joseph L. and Maren K. Olsen. 1998. \Multiple imputation for multivariate missing-data problems: A data analyst's perspective." Multivariate Behavioral Research 33(4):545-571.

Efficiency of m imputations for 17 probes

Feature #	missing fraction	Efficiency of m imputation per feature	Average of efficiency (12 feature)
1	0.165048544	0.991815118	**0.995782732**
2	0.048543689	0.997578693	m=20
3	0.038834951	0.998062016	
4	0	-	
5	0.077669903	0.996131528	
6	0.009708738	0.999514799	
7	0.155339806	0.992292871	
8	0.009708738	0.999514799	
9	0.077669903	0.996131528	
10	0.038834951	0.998062016	
11	0	-	
12	0.242718447	0.988009592	
13	0	-	
14	0.019417476	0.999030068	
15	0	-	
16	0	-	
17	0.13592233	0.993249759	