# An Introduction To Structured Missingness

Chris Harbron, Roche

25th April 2022 | For Public Use

# Structured Missingness

What Is Structured Missingness?

Structured Missingness as an Emerging Research Area - Grand Challenges

Characterising Structured Missingness

# Acknowledgements


**Alan Turing Institute**

Robin Mitra

Ben MacArthur

Chris Holmes

**Roche**

Niels Hagenbuch

Sarah McGough

Ryan Copping

**Structured Missingness Workshops Participants**

Athena Sheppard
Alejandro Frangi
Alisha Davies
Aditi Shenvi
Ana Basiri
Anower Hossain
Arun Sujenthiran
Brieuc Lehmann
David Leslie
Deepak Parashar
Eleni-Rosalina Andrinopoulou
Eda Ozyigit
Francisco Azuaje
Ghita Berrada

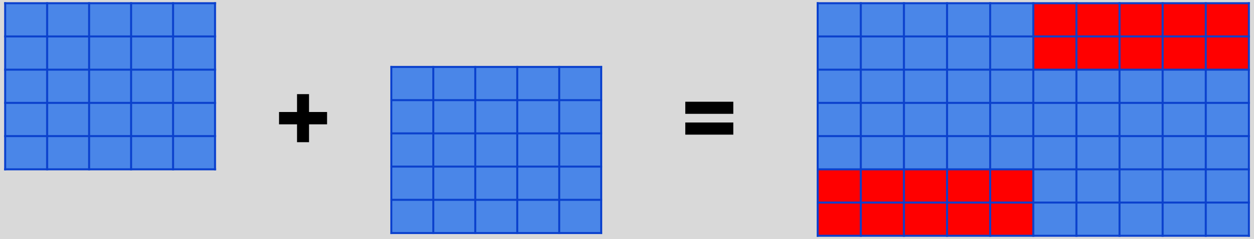
Ginestra Bianconi
Julia Brettschneider
Luis Santos
Marc De Kamps
Mark Gilthorpe
Michael Barnes
Maxine Mackintosh
Musa Abdulkareem
Nina Deliu
Jack Noonan
Nursen Aydin
Paolo Missier
Roy Ruddle
Rebecca Ward

Ruben Sanchez-Garcia
Ruby Chang
Sara Johansson Fernstad
Seppo Virtanen
Sorina Maciuca
Stefanie Biedermann
Stefanie Bienert
Stephen Gardiner
Thomas Burnett
Timothe Menard
Trevor Graham
Wenjuan Wang
Xijin Chen
Xuan Vinh Doan

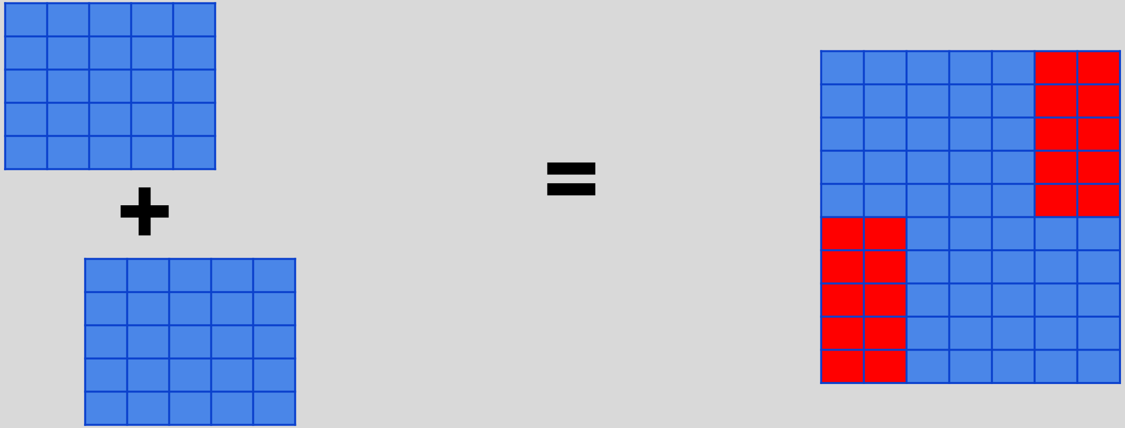
# What Is Structured Missingness?

Inevitable Consequence of Combining Multiple Datasets At Scale

**Adding Variables / Different Data Modalities**



**Adding Subjects**



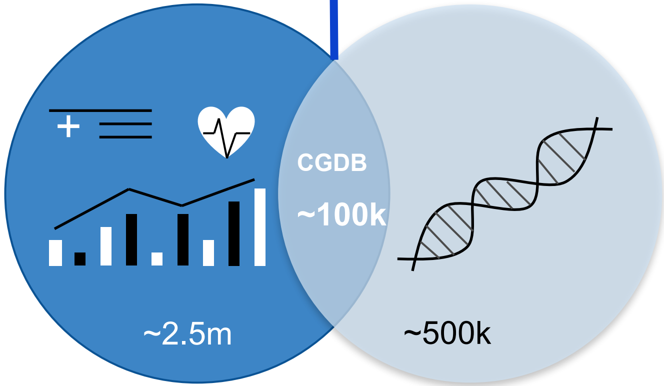
Many more subtle effects can also occur

Roche

# An Example of Structured Missingness - CGDB

The Clinico-Genomic Database links Flatiron electronic health records with Foundation Medicine (FMI) comprehensive genomic profiling for tens of thousands of cancer patients in the U.S.

Clinico-genomic database (CGDB)
>100 000 patients *(all cancer types)*




flatiron

Enhanced clinical RWD
+
Clinical outcomes

FOUNDATION MEDICINE

Comprehensive genomic data
+
Advanced genomic analysis

# An Example of Structured Missingness - CGDB

## Block Missingness From Measuring Different Genes

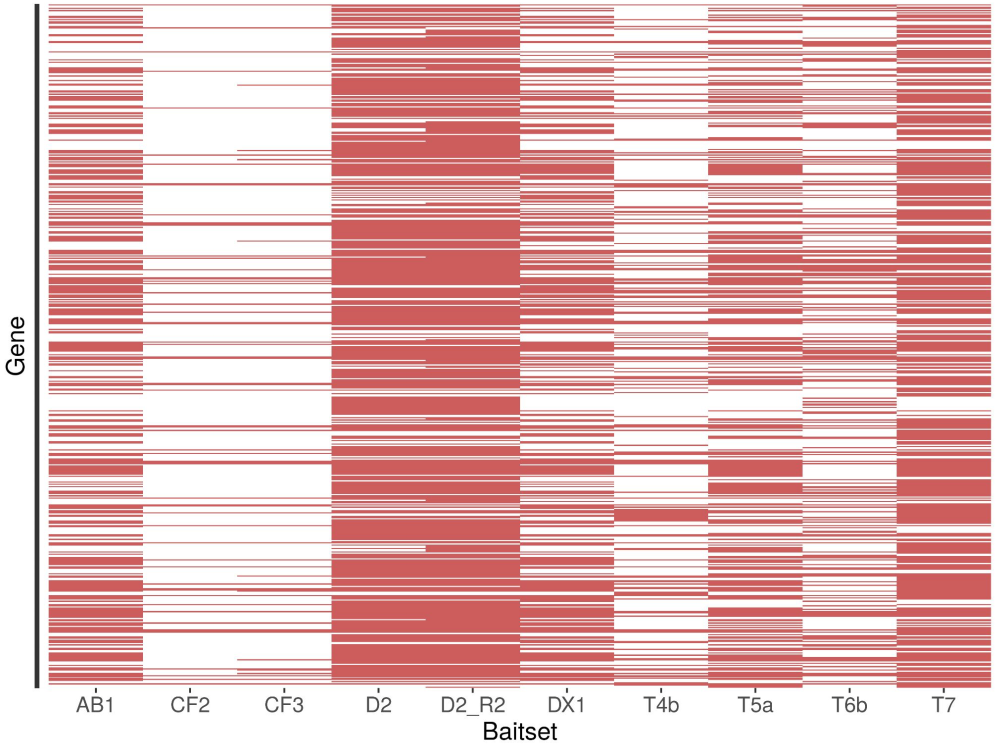
Each patient usually receives **1 test**.

Patients receive a variety of **different tests**
Tests are ordered to target specific treatment, prognosis, disease progression goals, and most importantly haematological vs solid tumours.
Tests also evolve over time
Tests can use different samples : solid tissue or liquid biopsy

Of **596 unique genes** measured in the CGDB, only **30** are measured across all tests.

Here, genes are **block missing** by test type.

**Measured** genes by test (baitset).



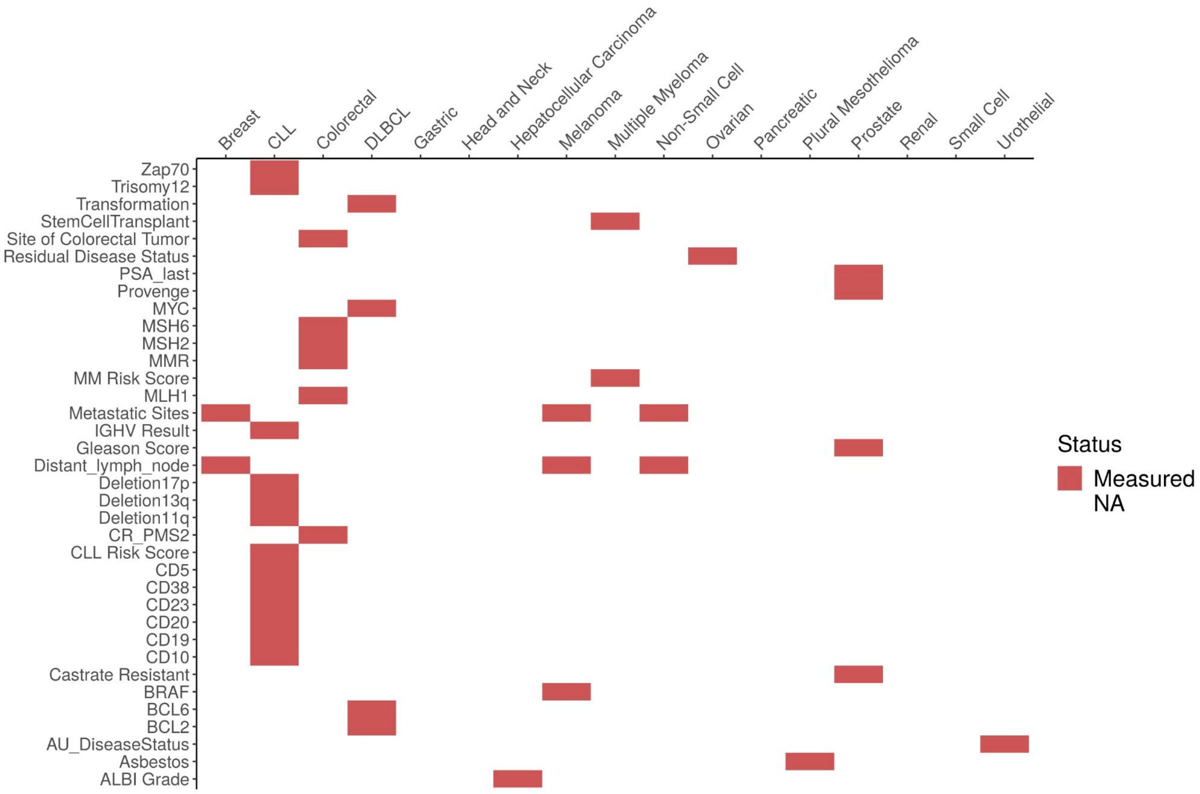
# An Example of Structured Missingness - CGDB

Block missingness from combining cancer specific information across dozens of cancer types

Each cancer type collects cancer-specific information, such as the **Gleason Score** for Prostate Cancer patients or **Stem Cell Transplant** for DLBCL patients

Here, variables are **“block missing”** by cancer type.

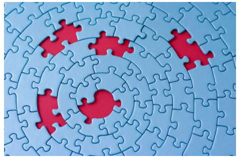
Is imputation appropriate when the missing value **doesn’t exist** or have any meaning?



# Why Is Structured Missingness Worth Considering?

It's **inevitable** & **ubiquitous** if we are combining datasets at scale

Many analytics and machine learning methodologies require **complete data**



The structure may present additional **challenges** or additional **opportunities** compared to unstructured missing data

The missingness may contain information - i.e. be **informative**

The missingness may highlight **limitations** of the data e.g. underrepresentation

# Grand Challenges For Structured Missingness

- Defining and Characterising Structured Missingness
- Exploring SM Geometry and Visualisation
- Prediction
- Inference and Estimation
- Causality
- The Role of Imputation
- Design Considerations
- Benchmarking And Evaluation
- Ethical Implications

# Characterising Structured Missingness

## Multiple Dimensions

| | | |
|---|---|---|
| 1 | Relationship of missingness patterns to values | • **MR** : Missingness occurs independently<br>• **MO** : Missingness related to values of other variables<br>• **MV** : Missingness related to value of variable |
| 2 | Nature of relationship of missingness patterns to values | • **D** : Deterministic<br>• **P** : Probabilistic |
| 3 | Relationship of missingness patterns to missingness patterns in other variables | • **U** : Unstructured<br>• **SS** : Strong Structure - (Deterministic)<br>• **WS** : Weak Structure - (Probabilistic) |
| 4 | Sub-characterisation by different patterns or structures of missingness | • e.g.<br>• **(B)** : Block Missing<br>• **(S)** : Sequentially Missing |
| 5 | Does a missing value exists but is unobserved, or no value exists | • **E** : Value exists but was not observed<br>• **N** : Value doesn't exist for logical/biological reasons |

# Characterising Structured Missingness

**Relationship of Missingness to Other Missingness in Data**

*Relationship of Missingness To Values In Data*

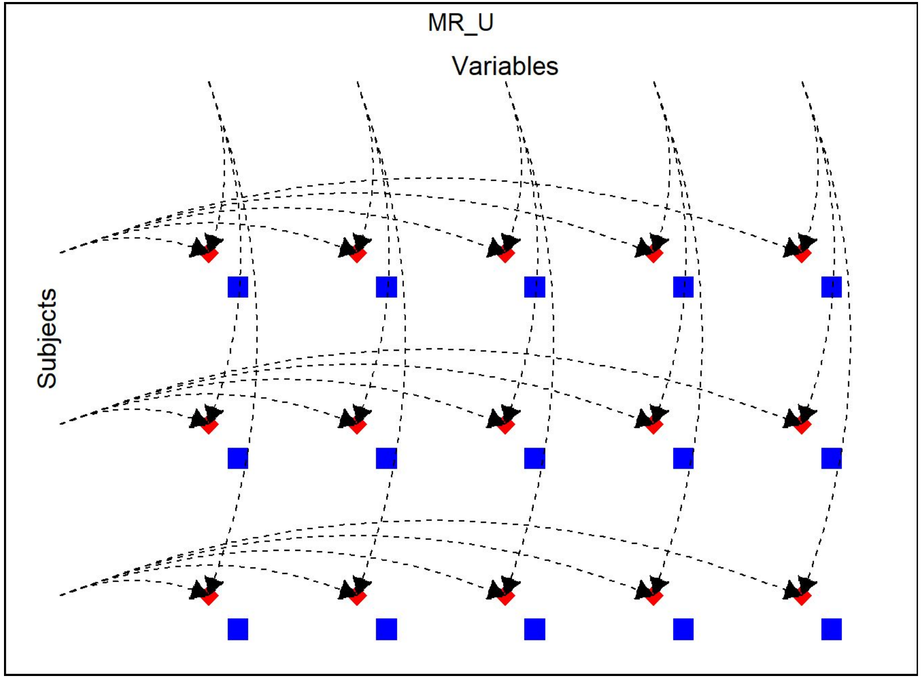
| | Unstructured | Structured: Strong | Structured: Weak |
|---|---|---|---|
| **MR** | MR-U | MR-SS<br>MR-SS(B) MR-SS(S) | MR-WS<br>MR-WS(B) MR-WS(S) |
| **MO - Prob** | MOP-U | MOP-SS<br>MOP-SS(B) MOP-SS(S) | MOP-WS<br>MOP-WS(B) MOP-WS(S) |
| **MO - Det** | MOD-U | MOD-SS<br>MOD-SS(B) MOD-SS(S) | MOD-WS<br>MOD-WS(B) MOD-WS(S) |
| **MV - Prob** | MVP-U | MVP-SS<br>MVP-SS(B) MVP-SS(S) | MVP-WS<br>MVP-WS(B) MVP-WS(S) |
| **MV - Det** | MVD-U | MVD-SS<br>MVD-SS(B) MVD-SS(S) | MVD-WS<br>MVD-WS(B) MVD-WS(S) |

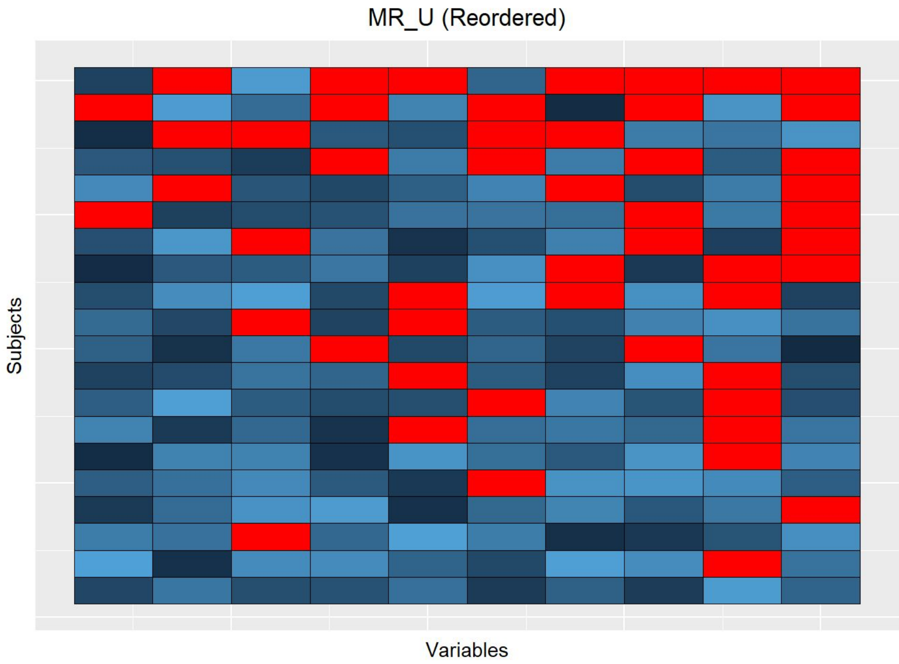
# Characterising Structured Missingness

## MR_U : Unstructured Missing Randomly

Simplest Case : $P(M_{ij} = 1) = k \quad \forall i, j$

General Case : $P(M_{ij} = 1) = f(s(i), v(j))$

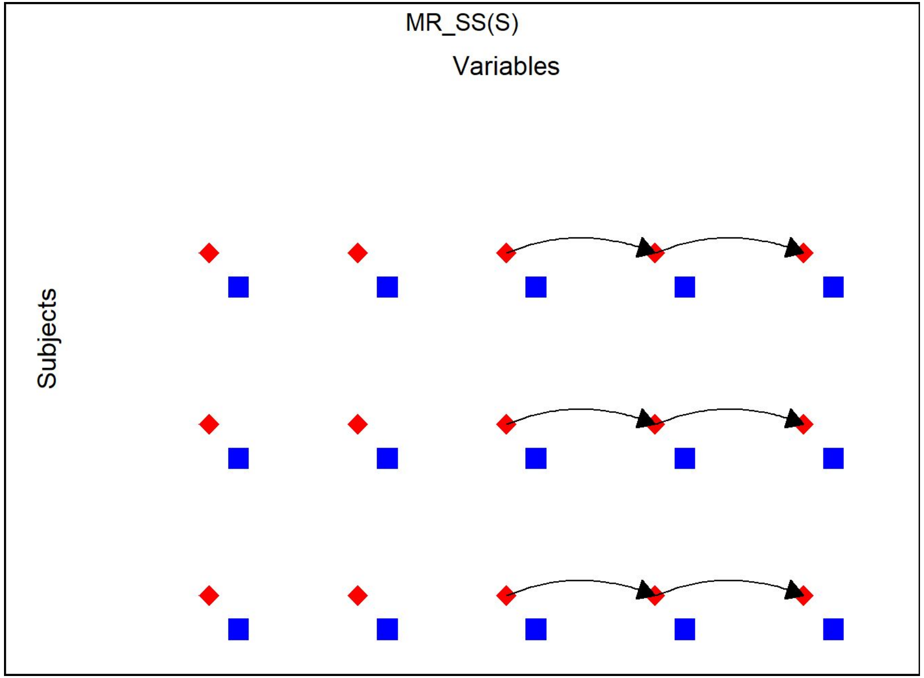


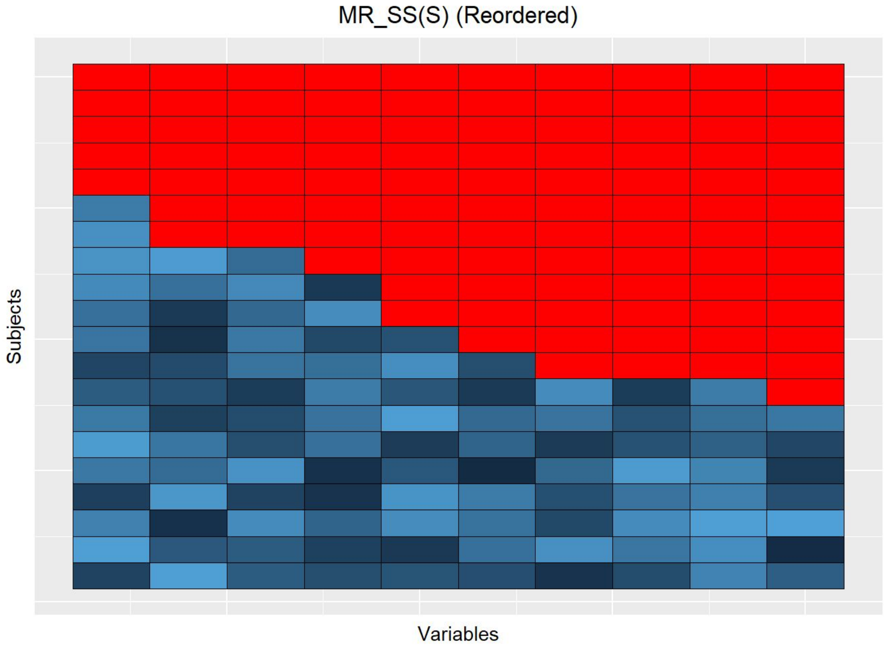



# Characterising Structured Missingness

## MR_SS(S) : Sequential Strong Structure Missing Randomly

$$M_{ij} = max(M_{ik:k<j}, P_{ij})$$

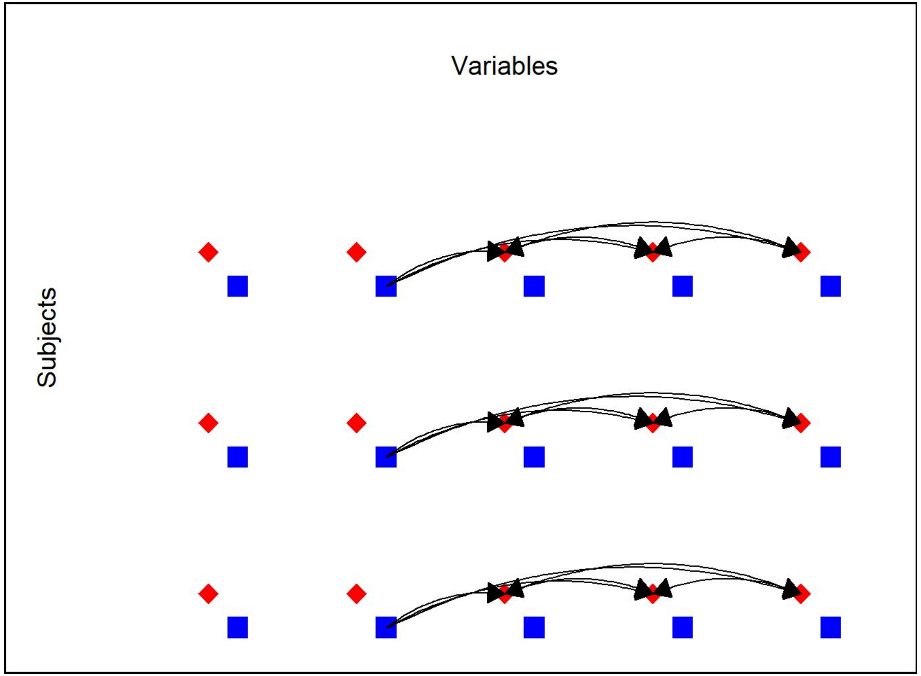


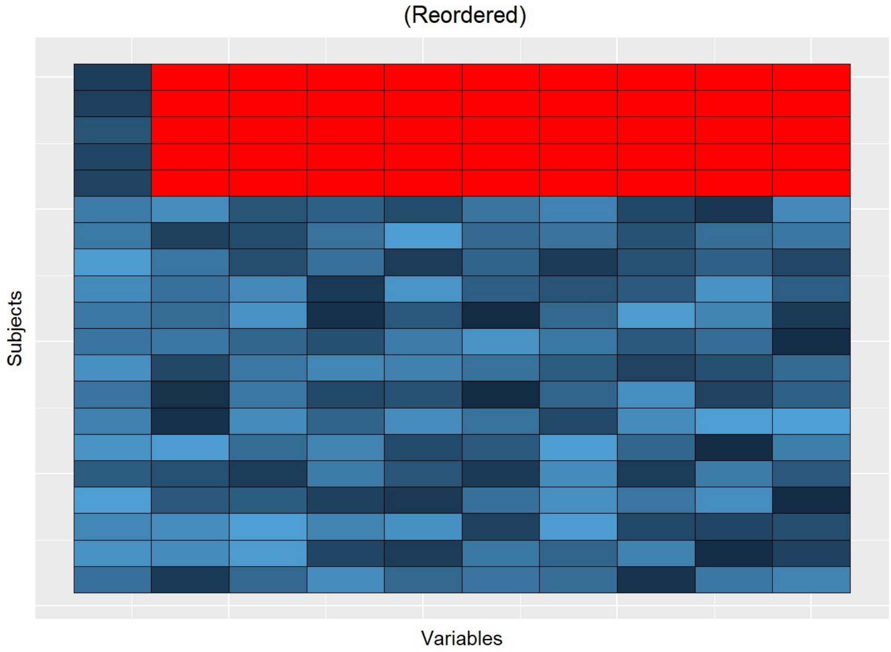



# Characterising Structured Missingness

## MOD_SS(B) :
## Strong Block Structure Missing Deterministically Based On Other Variables

$$M_{ij} = f(X_{ik}) \quad \forall j \in S, k \notin S$$

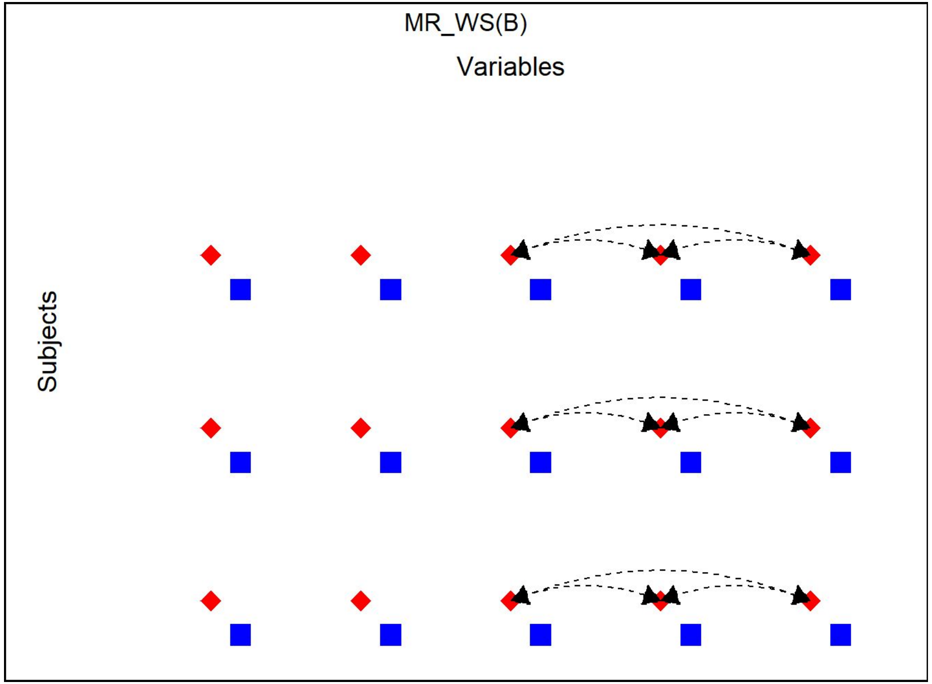


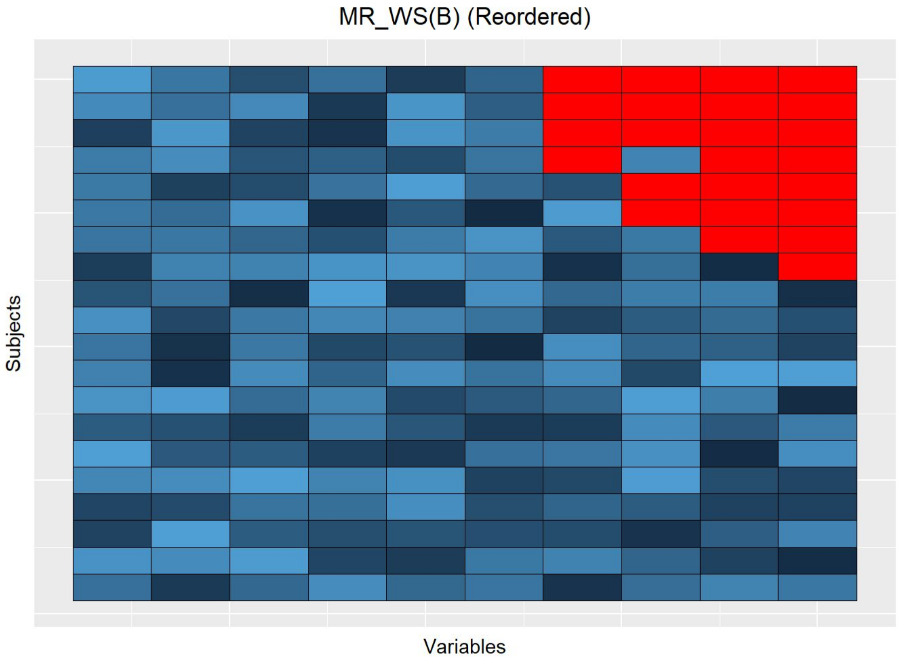



# Characterising Structured Missingness

## MR_WS(B) : Weak Block Structure Missing Randomly

$$P(M_{ij}=1)=f(M_{i,-j})$$

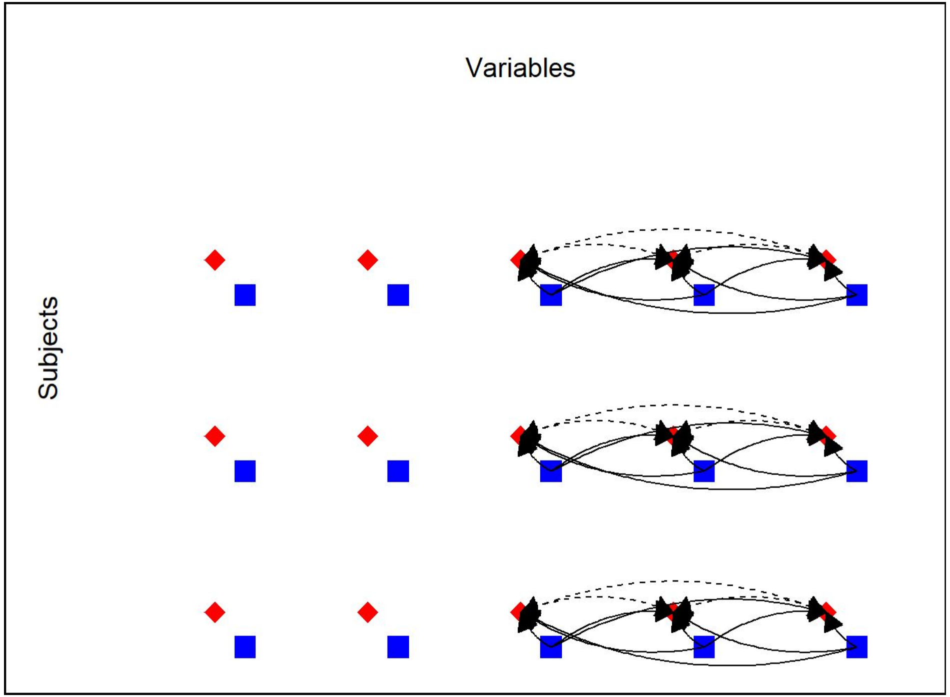


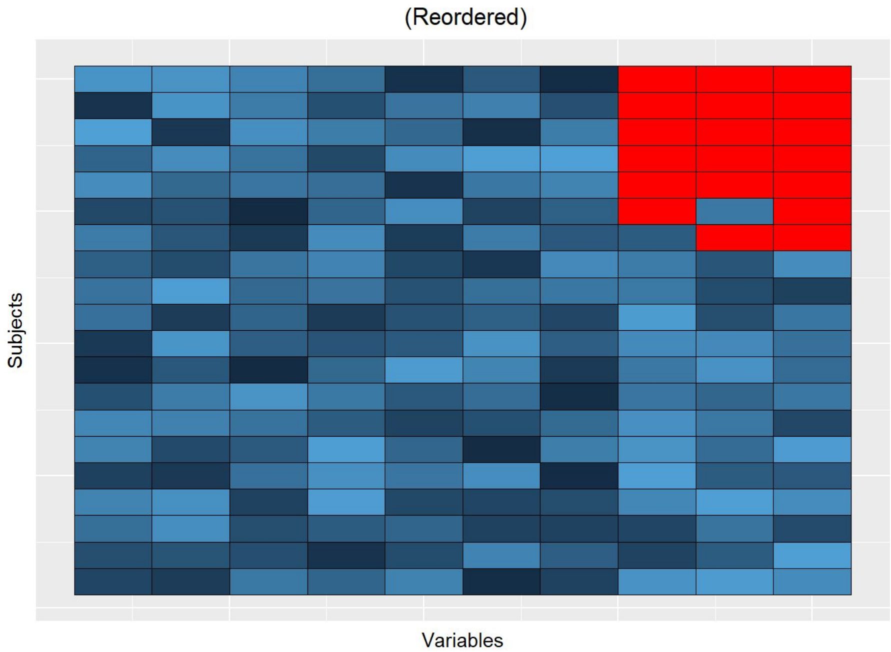



# Characterising Structured Missingness

## MVD_WS(B) : Weak Block Structure Deterministically On Variables' Values

$$P(M_{ij} = 1) = f(T_{ij}) \quad j \in S$$



# Where Next?


Roche-Turing Partnership Projects in SM

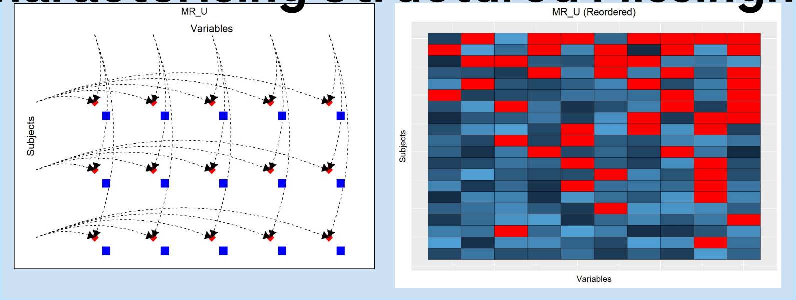
Publications on Grand Challenges in & Characterising SM

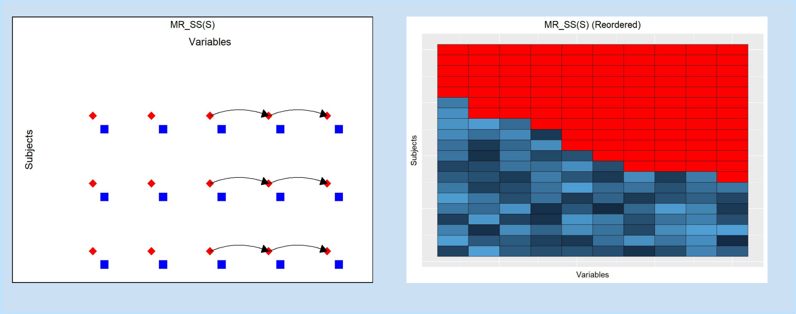
Continue to Build an SM Community

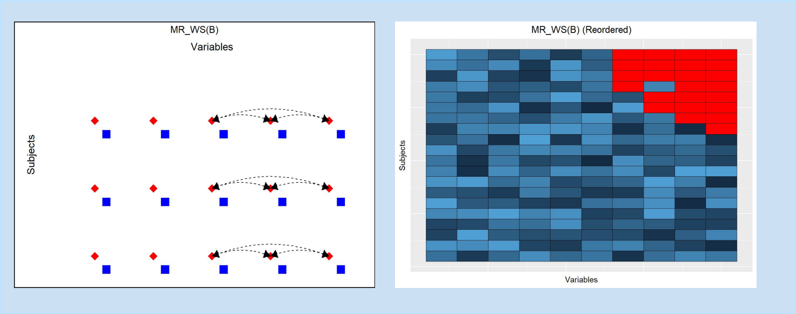
Slack Channels & Future Events

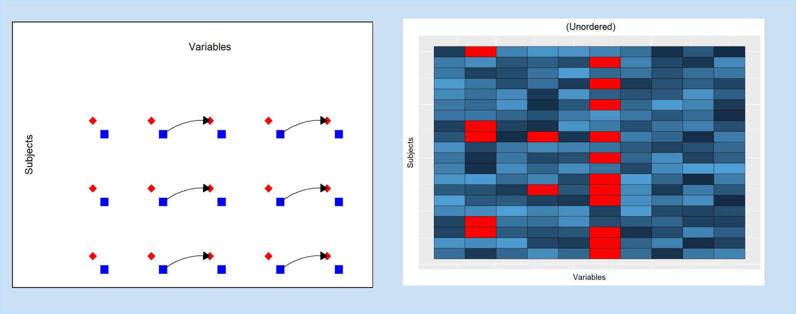
Doing now what patients need next

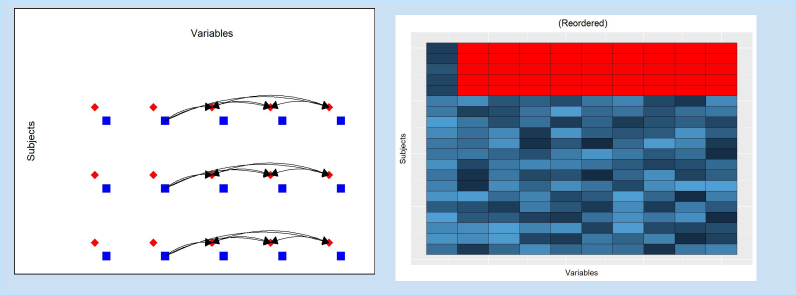
# Characterising Structured Missingness

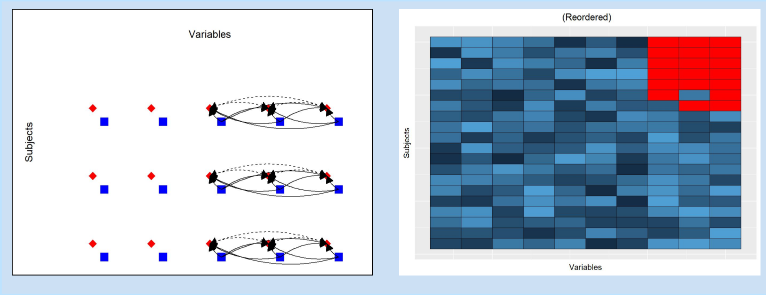


# Initial Project Theme

## Structured Missingness using CGDB as Motivation

Missing Data is a ubiquitous challenge across healthcare data, which compromises our ability to learn from data. This issue is exacerbated by structure in the missing values.

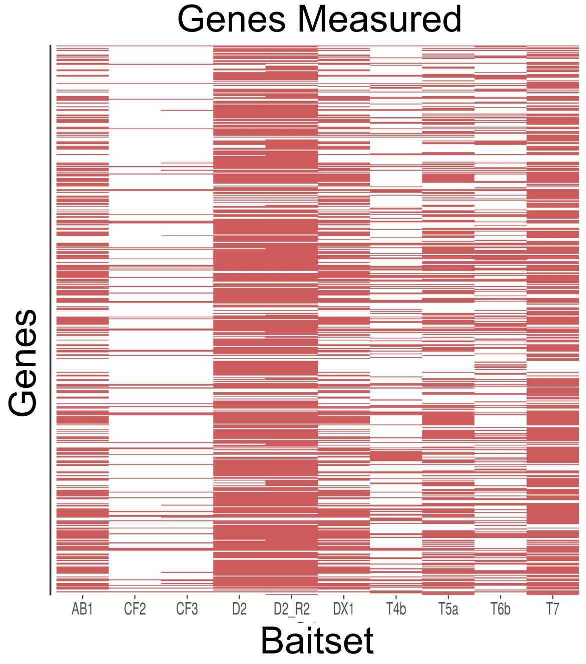
To make the most of data resources we need new methods to handle structured missingness, tailored to the particular challenges of healthcare data.

Clinico-genomic database (CGDB)
~ 80 000 patients
*(all cancer types)*

flatiron FOUNDATION MEDICINE

Data Model

| Enhanced clinical RWD | Comprehensive genomic data |
|---|---|
| + | + |
| Clinical outcomes | Advanced genomic analysis |




596 genes measured
30 measured across all baitsets

This is an essential building block which will be required across many different AA healthcare applications