Evaluating Data Linkage: Creating longitudinal synthetic data to provide a gold-standard linked dataset

Tom Dalton, Graham Kirby, Alan Dearle, Özgür Akgün

University of St Andrews

University of St Andrews

digitising SCOTLAND
understanding Scotland's people

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

wellcome

Background

- Digitising Scotland project
  - will transcribe vital event records 1855-1973
    - births
    - marriages
    - deaths
  - aim to link records to form family tree(s)
    - how do we evaluate our data linkage approach?

Why Synthetic Data?

- Inspired by real world hand-linked gold-standard data
    - Limited availability
    - Inherent errors
- Synthetic Data
    - Known truth gives a perfect gold-standard
    - Vary populations
        - Characteristics
        - Size
    - Many populations
    - Known level of corruption

Data Driven problems - what synthetic data do we need to evaluate the problems we solve?

Our approaches

- Organic Population Model
    - Event driven micro-simulation
    - *Tom Dalton, Victor Andrei*

- Verified Population Model
    - Time step driven micro-simulation

OPM – Overview

- Approach
  - Takes in a set of distributions defined by the user and a seed size
  - Sets up a population
  - Runs population for given time
  - Generates logging graphs
  - Outputs to desired format

OPM – Inputs

Genealogical controlling inputs are variable over time

Annotations

- female first name
- male first name
- surname
- occupation
- cause of death
- address

Seed

- seed age for males
- seed age for females

Birth

- children number of in cohab
- children number of in cohab then marriage
- children number of in marriage

- children number of in pregnancy

Partnering

- partnership characteristic
- partnership remarriage characteristic

- marriage age for males
- marriage age for females

- cohabitation age for males
- cohabitation age for females

- cohabitation to marriage time
- cohabitation length

Death

- death age at

Separation

- divorce age for male
- divorce age for female

- divorce instigated by gender
- divorce reason male
- divorce reason female

- divorce remarriage boolean
- remarriage time to

Genealogical complexity

- affair number of

- affair number of children
- affair with single or married

OPM – Inputs

- Age at death

0	36525																			
1600	2	2	2	3	7	4	3	5	20	21	35	63	115	139	143	143	149	94	20	20
1700	2	1	2	3	7	4	3	5	20	21	35	63	115	120	125	150	160	110	25	22

- Female age at marriage

5478	36525																
1600	6	166	222	190	150	114	82	24	24	15	7	1	1	1	1	1	1
1700	6	120	222	192	148	103	93	26	22	12	10	1	1	1	1	0	0

- Male age at marriage

5478	36525																
1600	6	137	214	192	161	122	91	28	28	14	6	1	1	1	1	1	1
1700	3	144	210	180	160	125	96	30	25	18	5	3	2	2	2	1	1

OPM – Approach

1. Set inputs
2. Choose start date
3. Choose seed population size
4. Decide ages of people in seed population

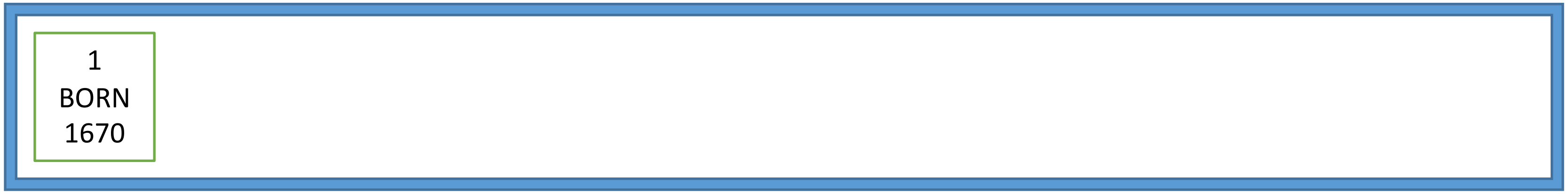
Head of queue

For each person in seed:

- Work out D.O.B.
- Make a birth event
- Insert into queue

OPM – Creating the seed

1. Set inputs
2. Choose start date
3. Choose seed population size
4. Decide ages of people in seed population



1
BORN
1670

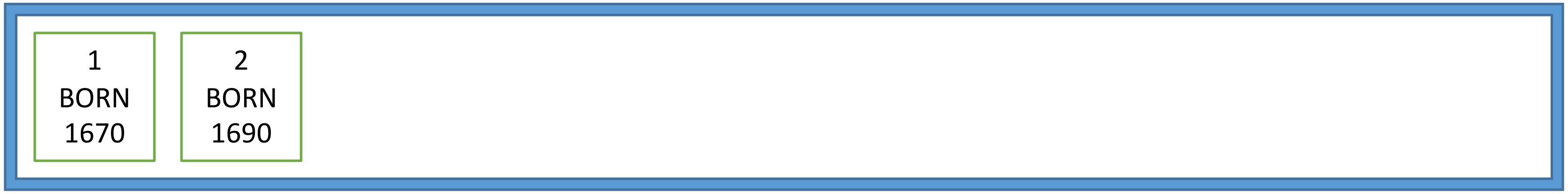
Head of
queue

For each person in seed:

- Work out D.O.B.
- Make a birth event
- Insert into queue

OPM – Creating the seed

1. Set inputs
2. Choose start date
3. Choose seed population size
4. Decide ages of people in seed population



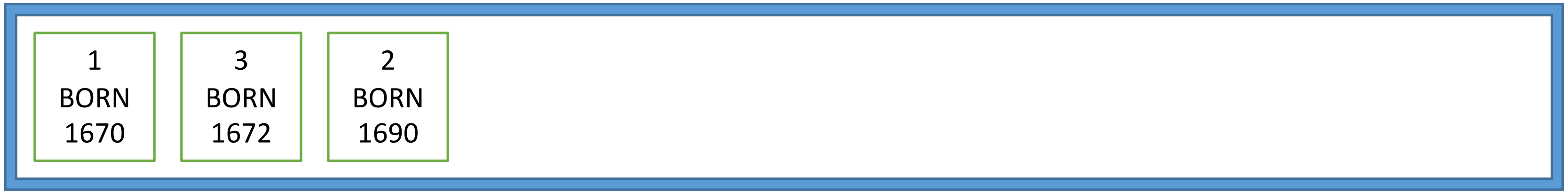
Head of queue

For each person in seed:

- Work out D.O.B.
- Make a birth event
- Insert into queue

OPM – Creating the seed

1. Set inputs
2. Choose start date
3. Choose seed population size
4. Decide ages of people in seed population



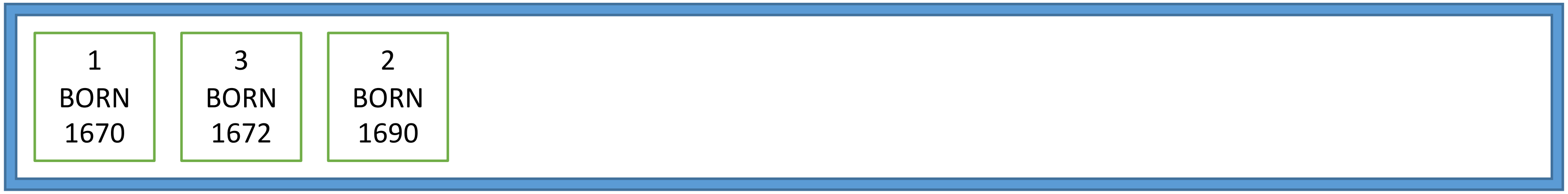
Head of queue

For each person in seed:

- Work out D.O.B.
- Make a birth event
- Insert into queue

OPM – Handling events

1. Take event from from of queue
2. Perform event
3. Create resultant events
4. Insert events into queue



1
BORN
1670

3
BORN
1672

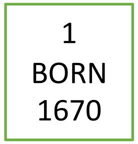
2
BORN
1690

Head of
queue

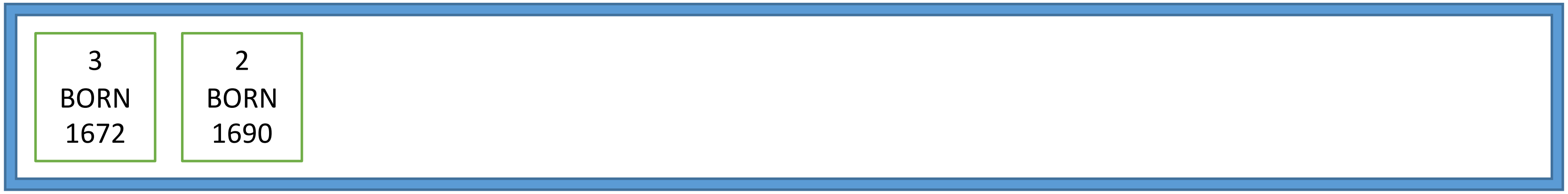
For BORN event:

- Create person
- Decide on first partnership characteristic
  - Set date
  - Insert
- Death
  - Set date
  - Insert

OPM – Handling events


1
BORN
1670

1. Take event from from of queue
2. Perform event
3. Create resultant events
4. Insert events into queue


3
BORN
1672

2
BORN
1690

Head of queue

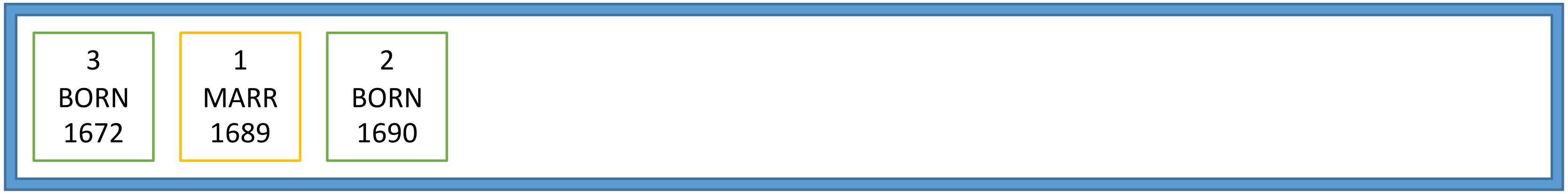
For BORN event:

- Create person
- Decide on first partnership characteristic
  - Set date
  - Insert
- Death
  - Set date
  - Insert

1
BORN
1670

OPM – Handling events

1. Take event from from of queue
2. Perform event
3. Create resultant events
4. Insert events into queue



3
BORN
1672

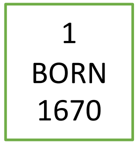
1
MARR
1689

2
BORN
1690

Head of
queue

For BORN event:

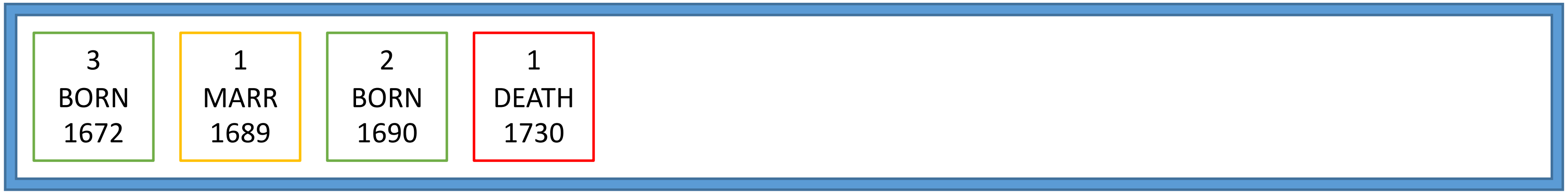
- Create person
- Decide on first partnership characteristic
  - Set date
  - Insert
- Death
  - Set date
  - Insert



1
BORN
1670

OPM – Handling events

1. Take event from from of queue
2. Perform event
3. Create resultant events
4. Insert events into queue



3
BORN
1672

1
MARR
1689

2
BORN
1690

1
DEATH
1730

Head of
queue

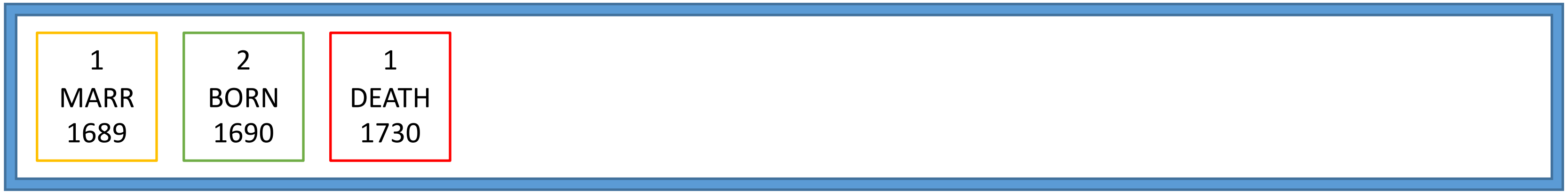
For BORN event:

- Create person
- Decide on first partnership characteristic
  - Set date
  - Insert
- Death
  - Set date
  - Insert

3
BORN
1672

OPM – Handling events

1. Take event from from of queue
2. Perform event
3. Create resultant events
4. Insert events into queue



1
MARR
1689

2
BORN
1690

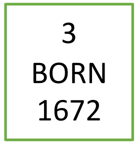
1
DEATH
1730

Head of
queue

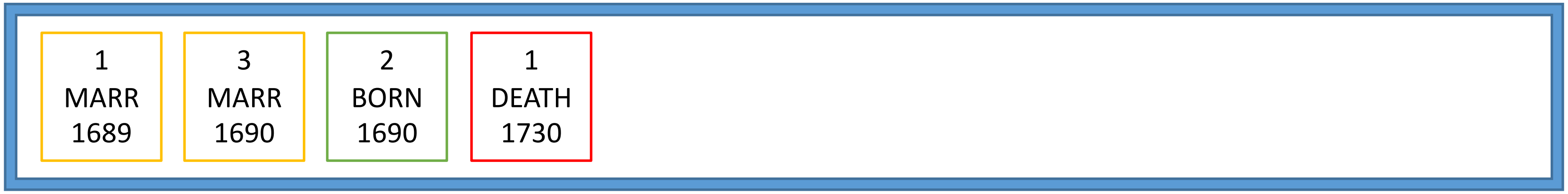
For BORN event:

- Create person
- Decide on first partnership characteristic
  - Set date
  - Insert
- Death
  - Set date
  - Insert

OPM – Handling events


3
BORN
1672

1. Take event from from of queue
2. Perform event
3. Create resultant events
4. Insert events into queue


1
MARR
1689

3
MARR
1690

2
BORN
1690

1
DEATH
1730

Head of queue

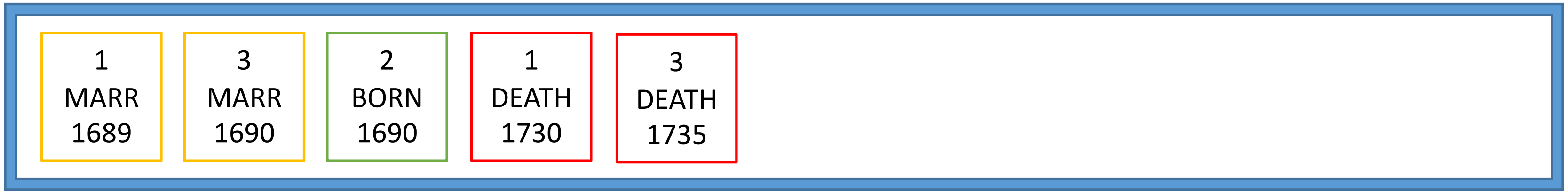
For BORN event:

- Create person
- Decide on first partnership characteristic
  - Set date
  - Insert
- Death
  - Set date
  - Insert

OPM – Handling events

3
BORN
1672

1. Take event from from of queue
2. Perform event
3. Create resultant events
4. Insert events into queue



1
MARR
1689

3
MARR
1690

2
BORN
1690

1
DEATH
1730

3
DEATH
1735

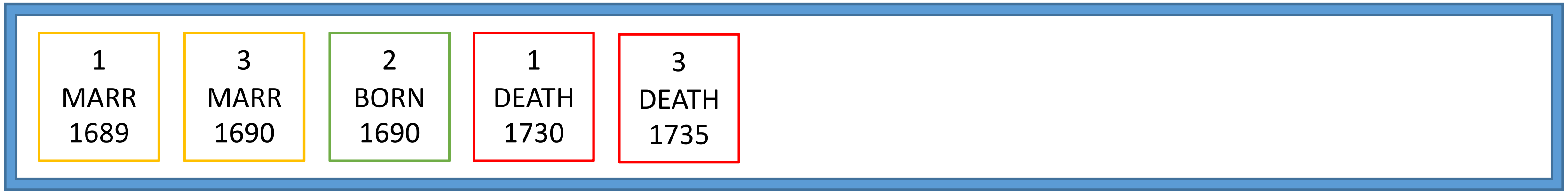
Head of
queue

For BORN event:

- Create person
- Decide on first partnership characteristic
  - Set date
  - Insert
- Death
  - Set date
  - Insert

OPM – Handling events

1. Take event from from of queue
2. Perform event
3. Create resultant events
4. Insert events into queue



1 MARR 1689

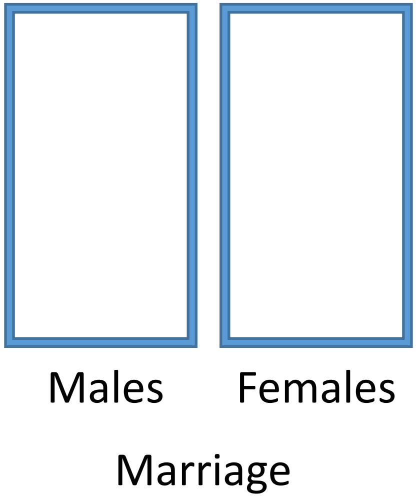
3 MARR 1690

2 BORN 1690

1 DEATH 1730

3 DEATH 1735

Head of queue



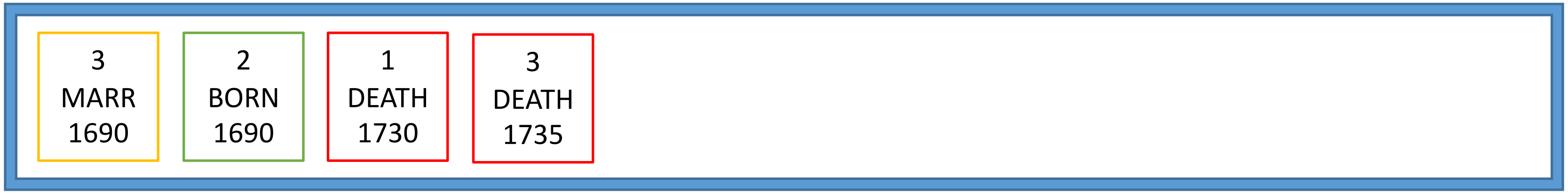
For MARRIAGE event:

- Add person to correct marriage pairing queue

OPM – Handling events

1
MARR
1689

1. Take event from from of queue
2. Perform event
3. Create resultant events
4. Insert events into queue



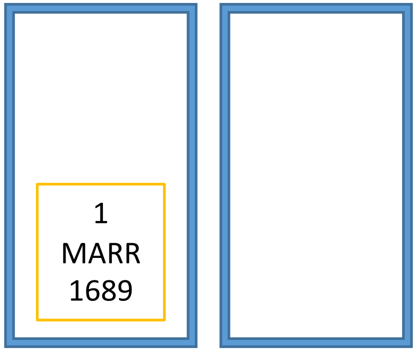
3
MARR
1690

2
BORN
1690

1
DEATH
1730

3
DEATH
1735

Head of
queue



1
MARR
1689

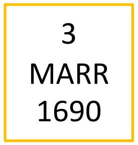
Males

Females

Marriage

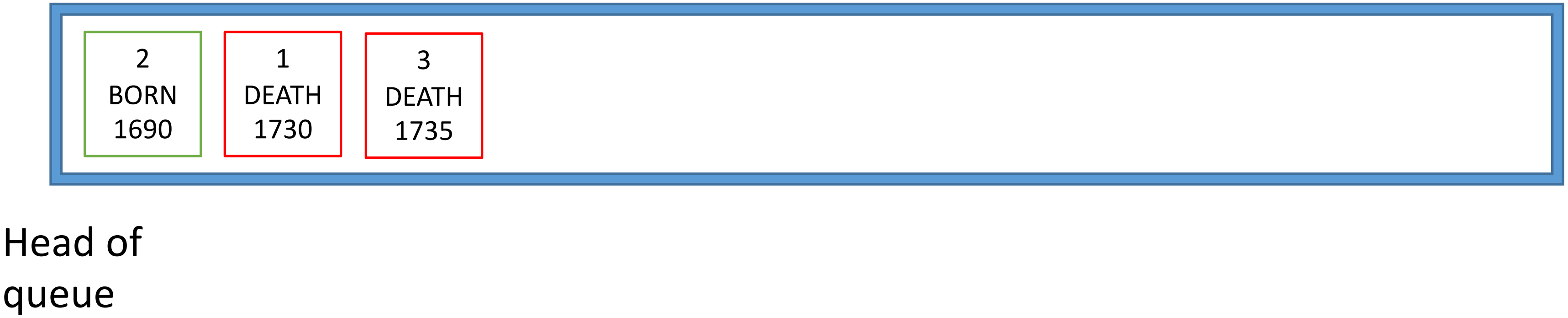
For MARRIAGE event:

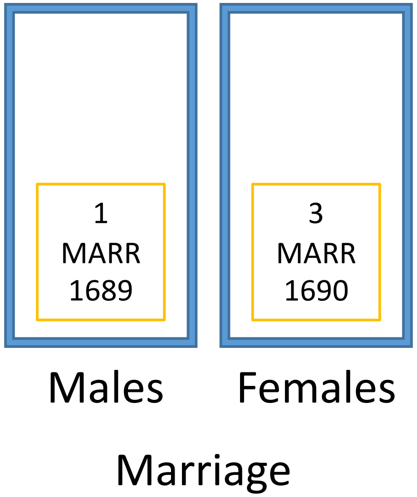
- Add person to correct marriage pairing queue



OPM – Handling events

1. Take event from from of queue
2. Perform event
3. Create resultant events
4. Insert events into queue



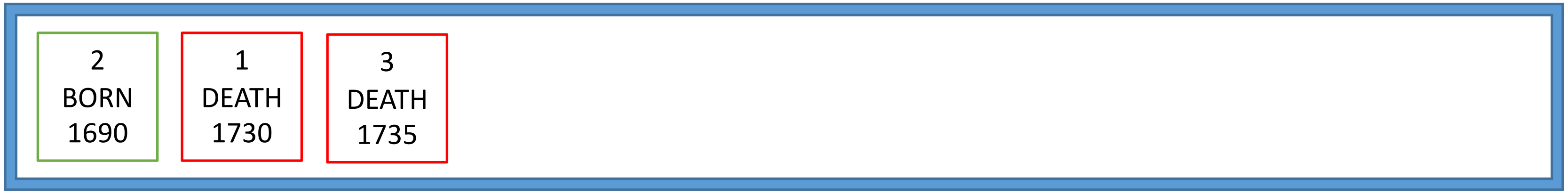


For MARRIAGE event:

- Add person to correct marriage pairing queue

OPM – Partnering

1. Once a year
2. Iterate over partnering queues
3. Partner together eligible individuals
4. Create resultant and insert events into queue

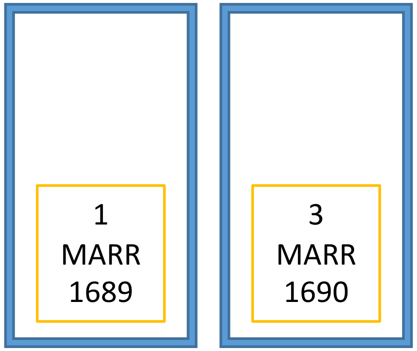


2 BORN 1690

1 DEATH 1730

3 DEATH 1735

Head of queue



1 MARR 1689

3 MARR 1690

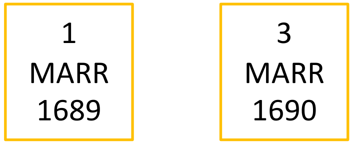
Males Females

Marriage

On Partnering of individuals:

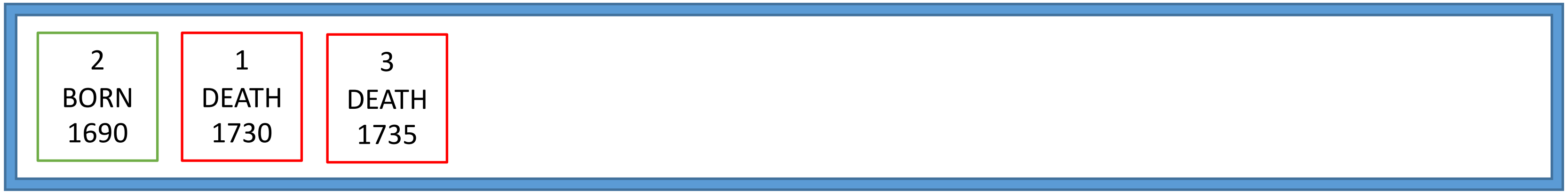
- Decide on end date
  - Insert end event
- Decide on first children
  - Insert BIRTH and BORN events

OPM – Partnering


1
MARR
1689

3
MARR
1690

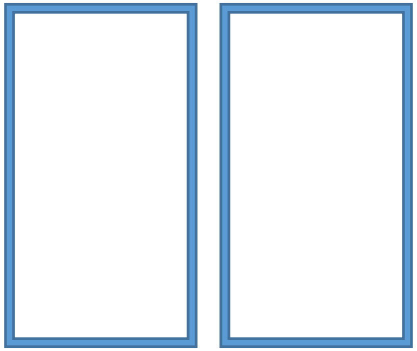
1. Once a year
2. Iterate over partnering queues
3. Partner together eligible individuals
4. Create resultant and insert events into queue


2
BORN
1690

1
DEATH
1730

3
DEATH
1735

Head of
queue

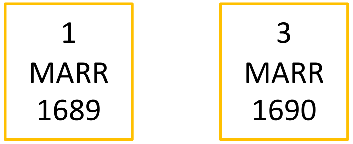

Males

Females

Marriage

On Partnering of individuals:

- Decide on end date
    - Insert end event
- Decide on children
    - Insert BIRTH and BORN events

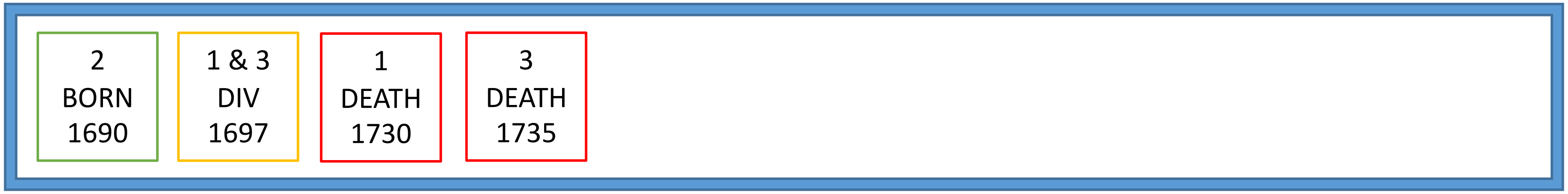


1 MARR 1689

3 MARR 1690

OPM – Partnering

1. Once a year
2. Iterate over partnering queues
3. Partner together eligible individuals
4. Create resultant and insert events into queue



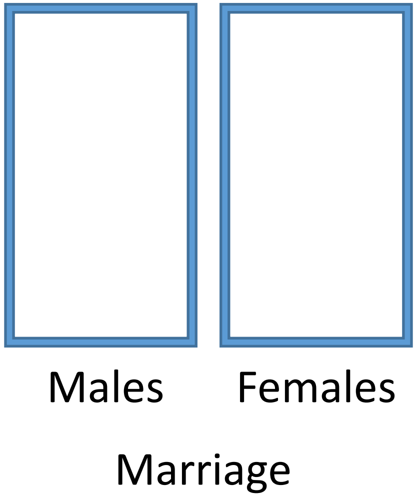
2 BORN 1690

1 & 3 DIV 1697

1 DEATH 1730

3 DEATH 1735

Head of queue



Males

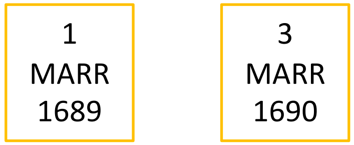
Females

Marriage

On Partnering of individuals:

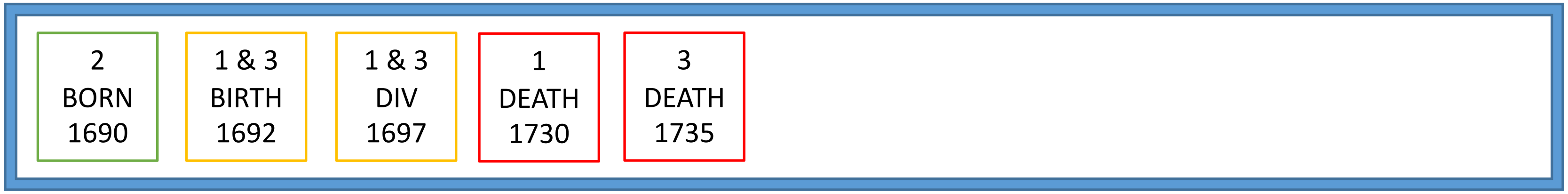
- Decide on end date
  - Insert end event
- Decide on children
  - Insert BIRTH and BORN events

OPM – Event Handling



1 MARR 1689

3 MARR 1690



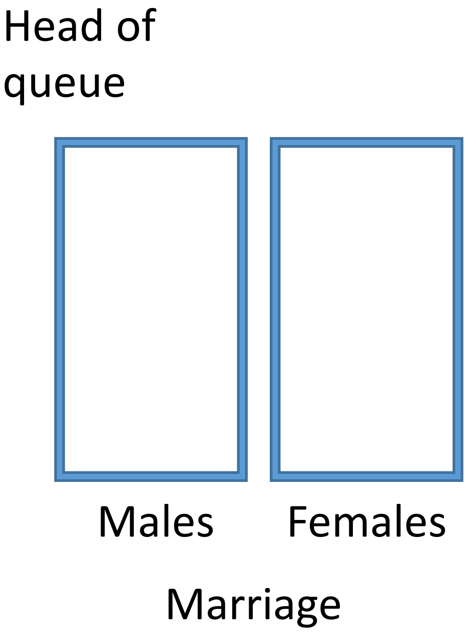
2 BORN 1690

1 & 3 BIRTH 1692

1 & 3 DIV 1697

1 DEATH 1730

3 DEATH 1735



Head of queue

Males

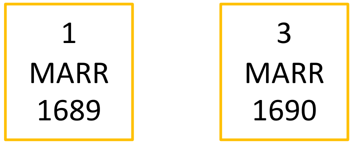
Females

Marriage

On Partnering of individuals:

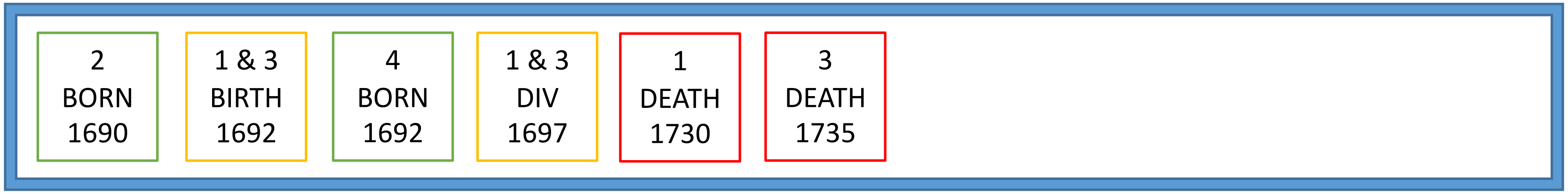
- Decide on end date
  - Insert end event
- Decide on children
  - Insert BIRTH and BORN events

OPM – Event Handling



1
MARR
1689

3
MARR
1690



2
BORN
1690

1 & 3
BIRTH
1692

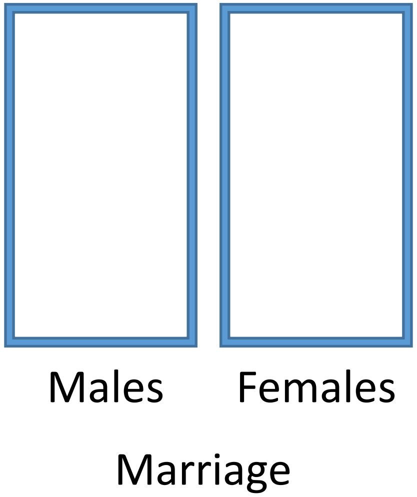
4
BORN
1692

1 & 3
DIV
1697

1
DEATH
1730

3
DEATH
1735

Head of
queue



Males

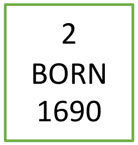
Females

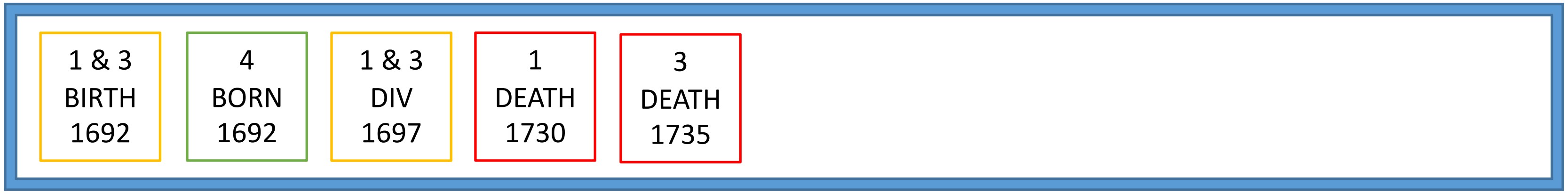
Marriage

For BORN event:

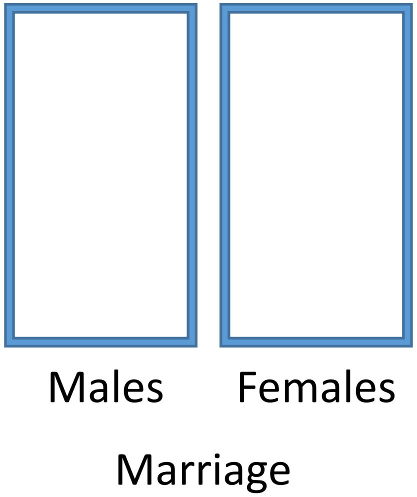
- Create person
- Decide on first partnership characteristic
    - Set date
    - Insert
- Death
    - Set date
    - Insert

OPM – Event Handling





Head of queue

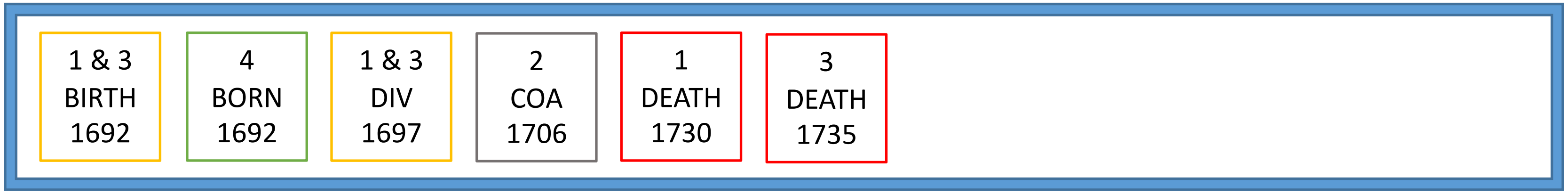


For BORN event:

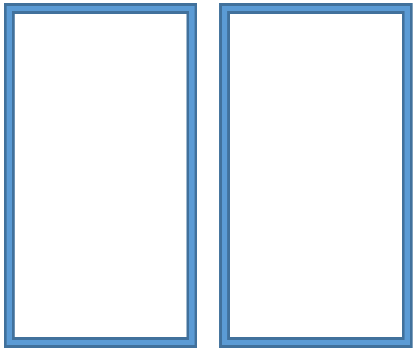
- Create person
- Decide on first partnership characteristic
  - Set date
  - Insert
- Death
  - Set date
  - Insert

OPM – Event Handling

2
BORN
1690



Head of queue



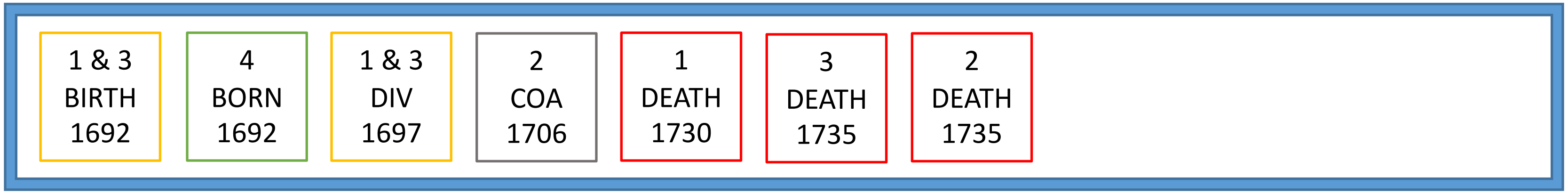
Marriage

For BORN event:

- Create person
- Decide on first partnership characteristic
  - Set date
  - Insert
- Death
  - Set date
  - Insert

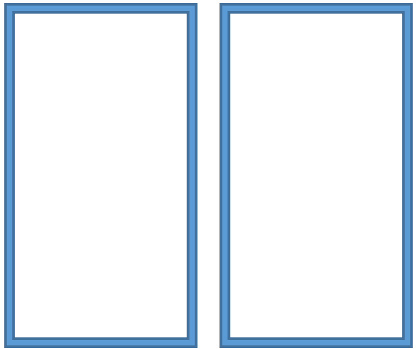
2
BORN
1690

OPM – Event Handling



1 & 3 BIRTH 1692	4 BORN 1692	1 & 3 DIV 1697	2 COA 1706	1 DEATH 1730	3 DEATH 1735	2 DEATH 1735

Head of queue



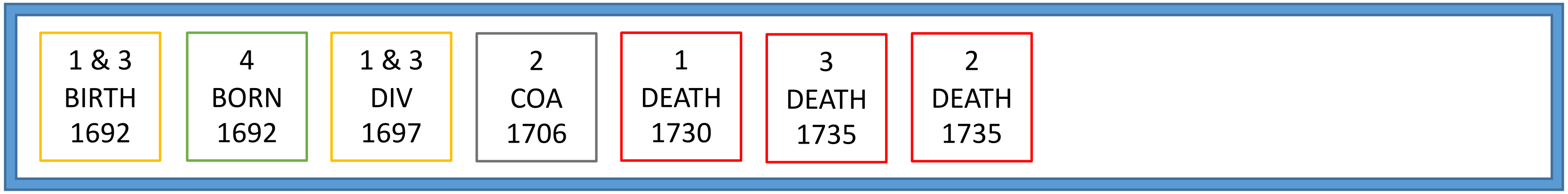
Males Females

Marriage

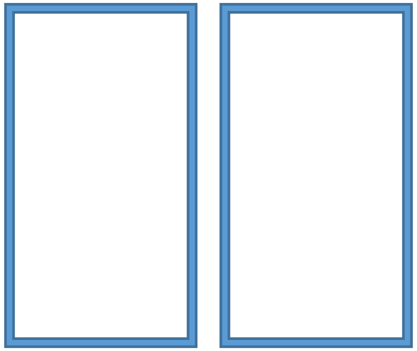
For BORN event:

- Create person
- Decide on first partnership characteristic
  - Set date
  - Insert
- Death
  - Set date
  - Insert

OPM – Event Handling



Head of queue

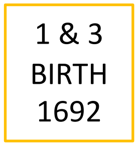


Males Females

Marriage

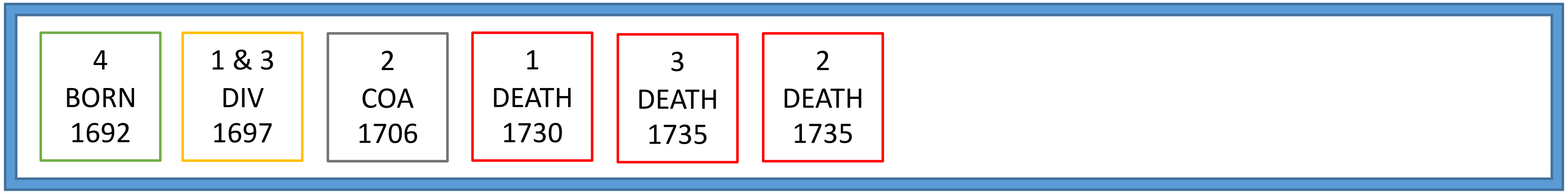
For BIRTH event:

- Decide if another birth
    - Set date
    - Insert BIRTH and BORN event



1 & 3
BIRTH
1692

OPM – Event Handling



4
BORN
1692

1 & 3
DIV
1697

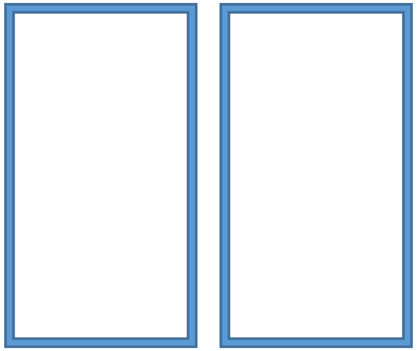
2
COA
1706

1
DEATH
1730

3
DEATH
1735

2
DEATH
1735

Head of queue



Males

Females

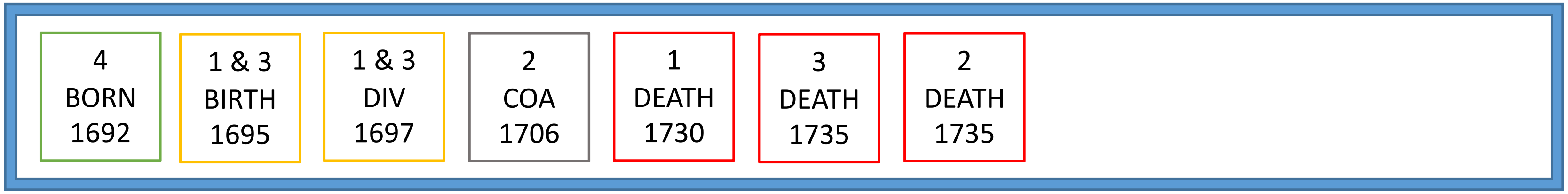
Marriage

For BIRTH event:

- Decide if another birth
  - Set date
  - Insert BIRTH and BORN event

OPM – Event Handling

1 & 3
BIRTH
1692



4
BORN
1692

1 & 3
BIRTH
1695

1 & 3
DIV
1697

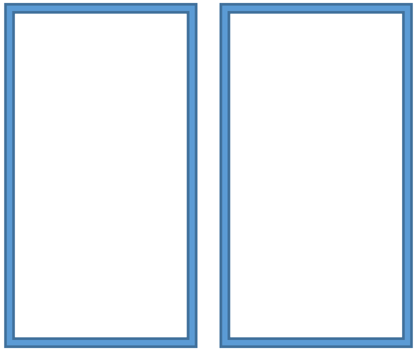
2
COA
1706

1
DEATH
1730

3
DEATH
1735

2
DEATH
1735

Head of
queue



Males Females

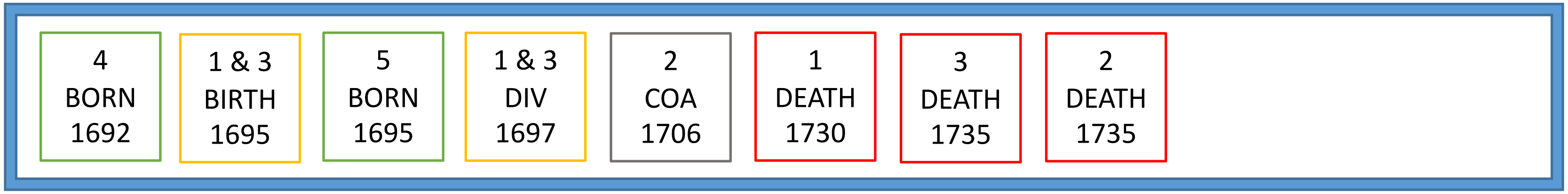
Marriage

For BIRTH event:

- Decide if another birth
  - Set date
  - Insert BIRTH and BORN event

1 & 3
BIRTH
1692

OPM – Event Handling



4
BORN
1692

1 & 3
BIRTH
1695

5
BORN
1695

1 & 3
DIV
1697

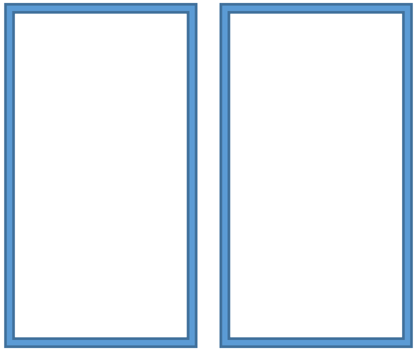
2
COA
1706

1
DEATH
1730

3
DEATH
1735

2
DEATH
1735

Head of queue



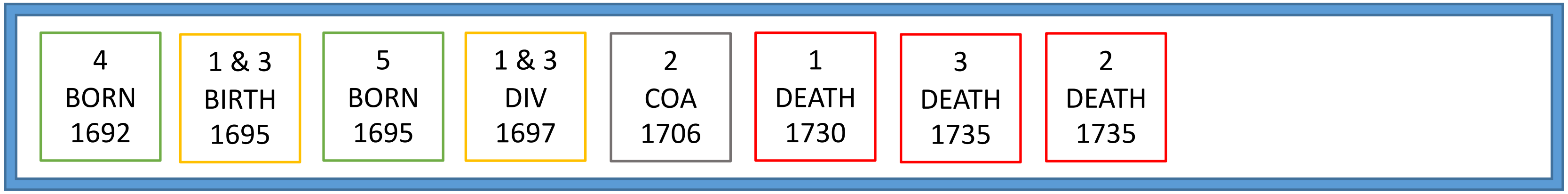
Males Females

Marriage

For BIRTH event:

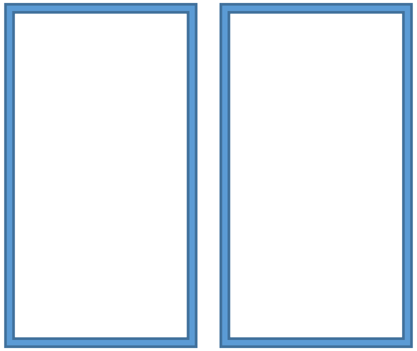
- Decide if another birth
  - Set date
  - Insert BIRTH and BORN event

OPM – Event Handling



4 BORN 1692	1 & 3 BIRTH 1695	5 BORN 1695	1 & 3 DIV 1697	2 COA 1706	1 DEATH 1730	3 DEATH 1735	2 DEATH 1735

Head of
queue



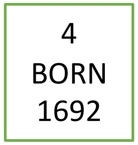
Males Females

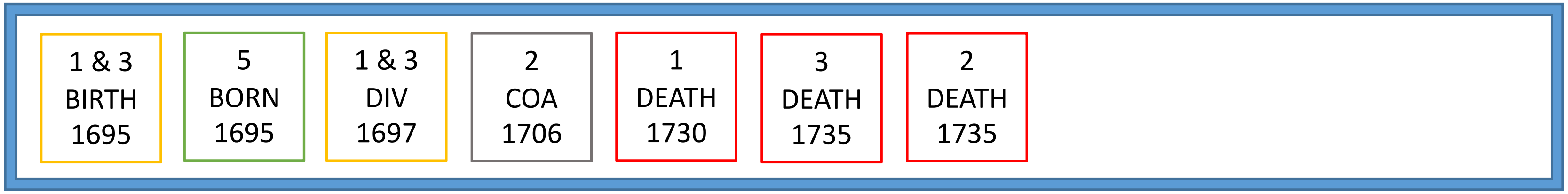
Marriage

For BORN event:

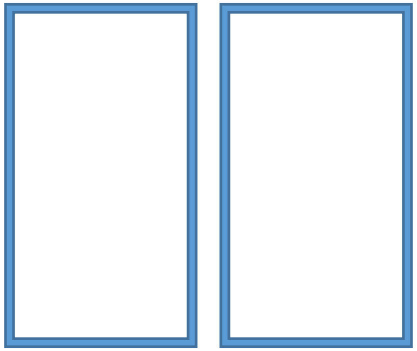
- Create person
- Decide on first partnership characteristic
  - Set date
  - Insert
- Death
  - Set date
  - Insert

OPM – Event Handling





Head of queue



Males Females

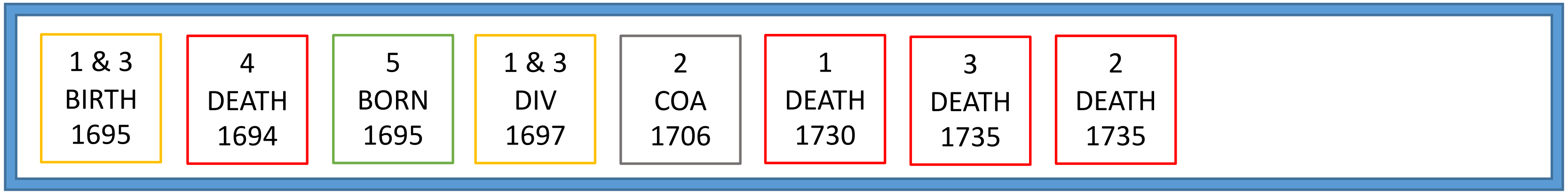
Marriage

For BORN event:

- Create person
- Decide on first partnership characteristic
    - Set date
    - Insert
- Death
    - Set date
    - Insert

OPM – Event Handling

4
BORN
1692



1 & 3 BIRTH 1695

4 DEATH 1694

5 BORN 1695

1 & 3 DIV 1697

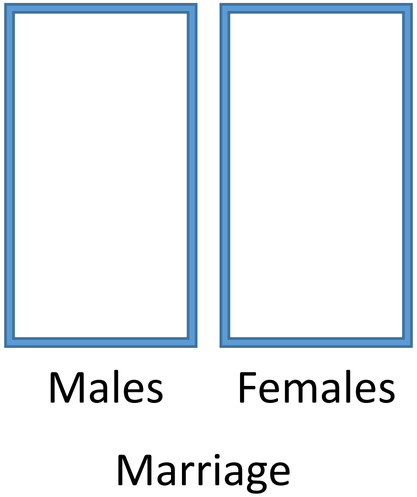
2 COA 1706

1 DEATH 1730

3 DEATH 1735

2 DEATH 1735

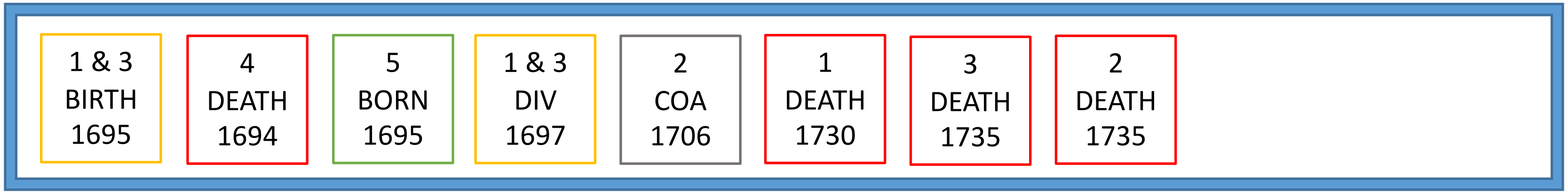
Head of queue



For BORN event:

- Create person
- Decide on first partnership characteristic
  - Set date
  - Insert
- Death
  - Set date
  - Insert

OPM – Event Handling



1 & 3 BIRTH 1695

4 DEATH 1694

5 BORN 1695

1 & 3 DIV 1697

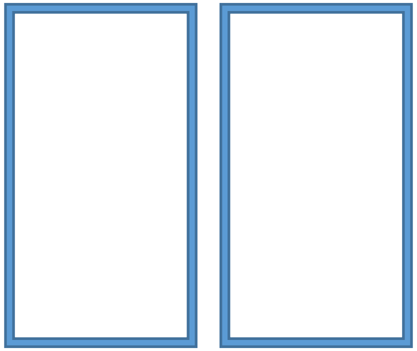
2 COA 1706

1 DEATH 1730

3 DEATH 1735

2 DEATH 1735

Head of queue



Males

Females

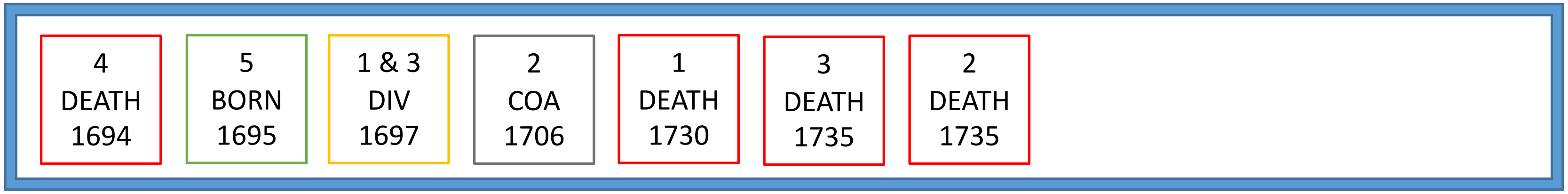
Marriage

For BIRTH event:

- Decide if another birth
  - Set date
  - Insert BIRTH and BORN event

1 & 3
BIRTH
1695

OPM – Event Handling


4
DEATH
1694

5
BORN
1695

1 & 3
DIV
1697

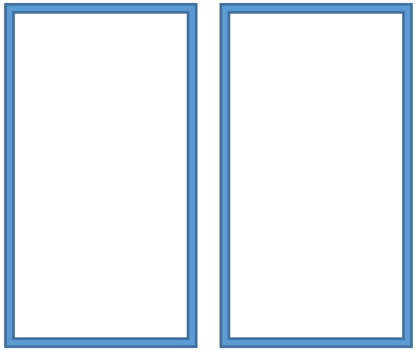
2
COA
1706

1
DEATH
1730

3
DEATH
1735

2
DEATH
1735

Head of
queue


Males

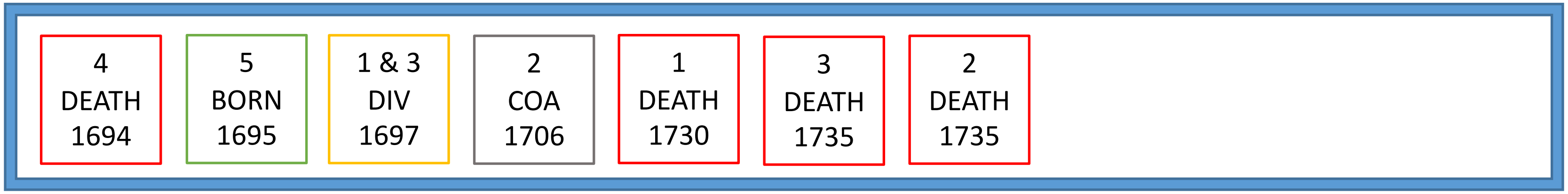
Females

Marriage

For BIRTH event:

- Decide if another birth
    - Set date
    - Insert BIRTH and BORN event

OPM – Event Handling



4
DEATH
1694

5
BORN
1695

1 & 3
DIV
1697

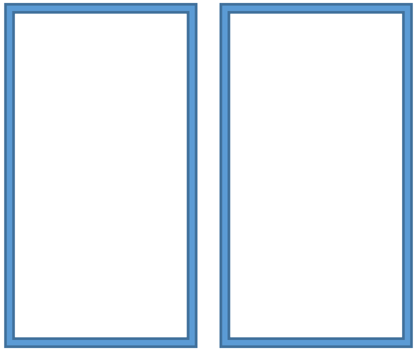
2
COA
1706

1
DEATH
1730

3
DEATH
1735

2
DEATH
1735

Head of queue



Males Females

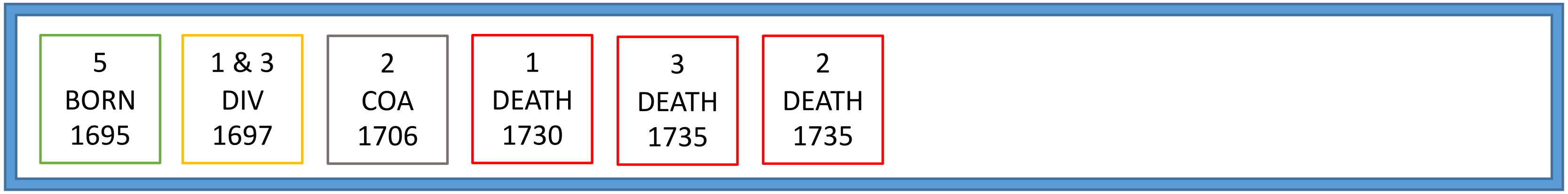
Marriage

For DEATH event:

- Remove

OPM – Event Handling

4
DEATH
1694



5
BORN
1695

1 & 3
DIV
1697

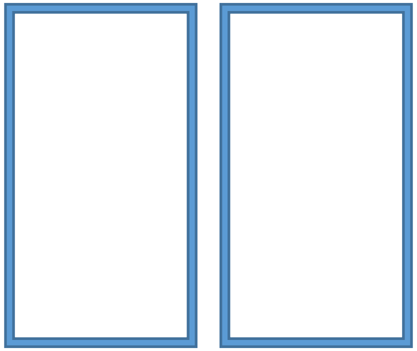
2
COA
1706

1
DEATH
1730

3
DEATH
1735

2
DEATH
1735

Head of queue



Males

Females

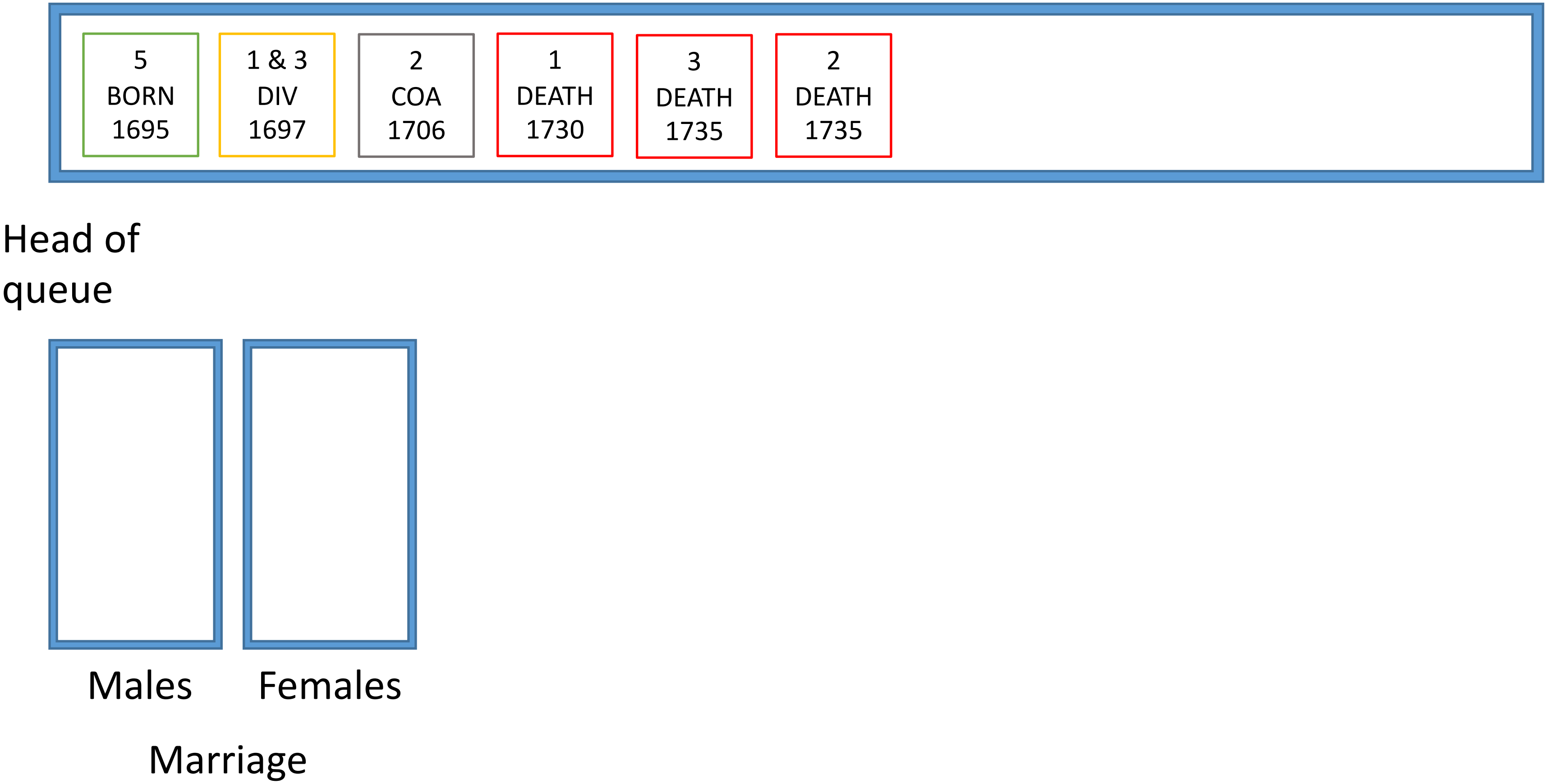
Marriage

For DEATH event:

- Remove

OPM – Event Handling

4
DEATH
1694



5
BORN
1695

1 & 3
DIV
1697

2
COA
1706

1
DEATH
1730

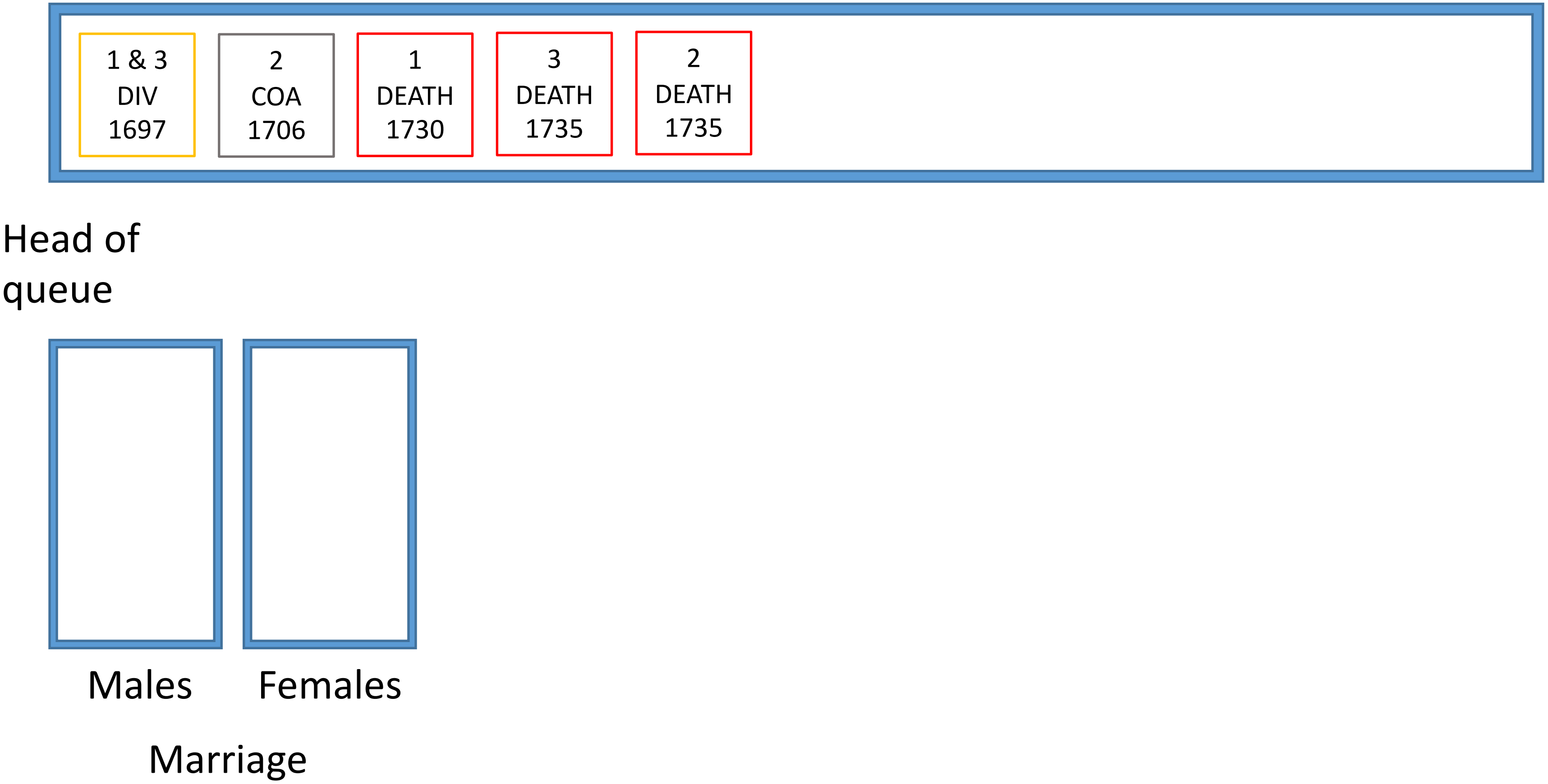
3
DEATH
1735

2
DEATH
1735

Head of
queue

OPM – Event Handling

5
BORN
1695



1 & 3
DIV
1697

2
COA
1706

1
DEATH
1730

3
DEATH
1735

2
DEATH
1735

Head of
queue

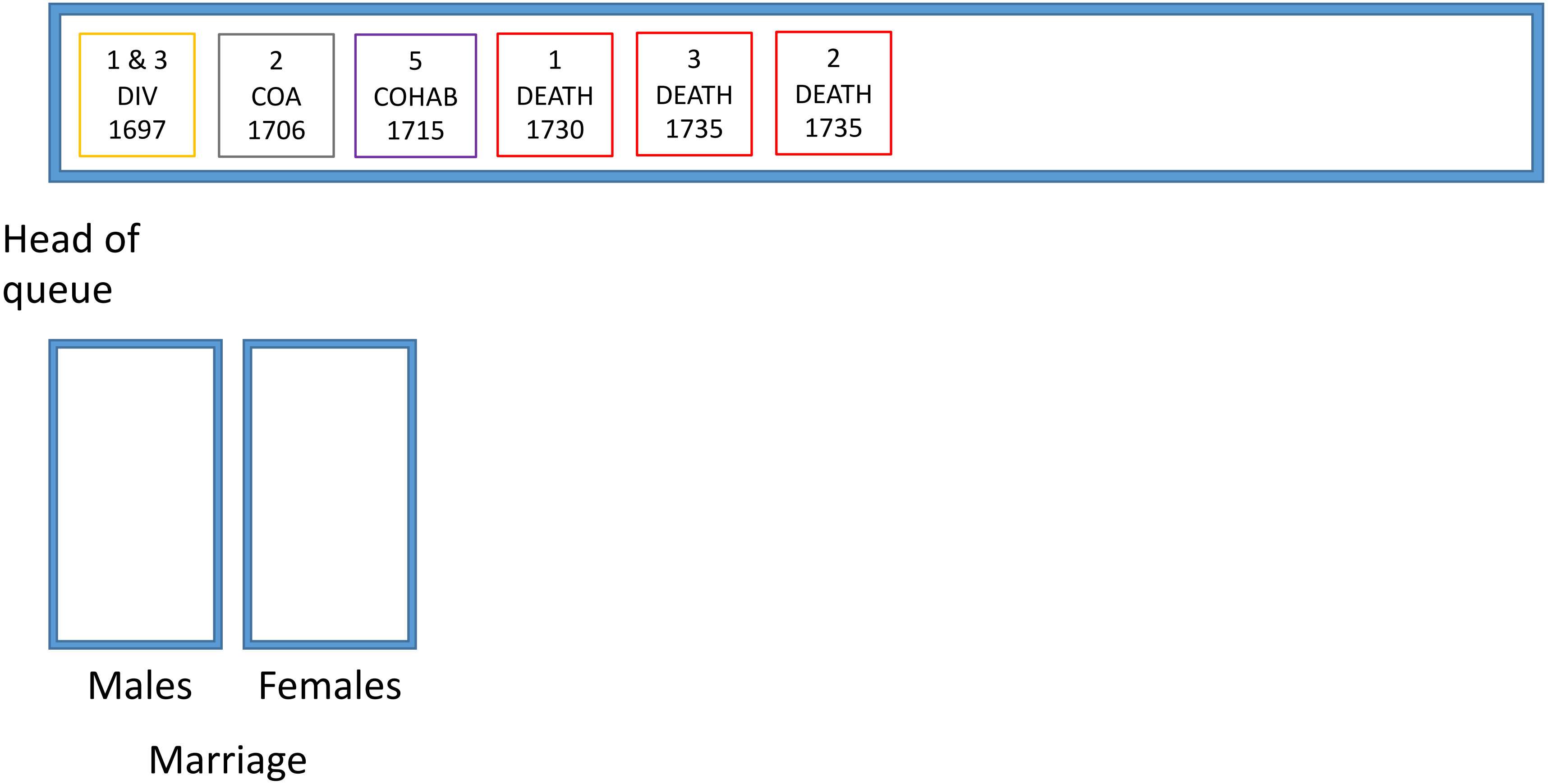
Males

Females

Marriage

OPM – Event Handling

5
BORN
1695



1 & 3
DIV
1697

2
COA
1706

5
COHAB
1715

1
DEATH
1730

3
DEATH
1735

2
DEATH
1735

Head of
queue

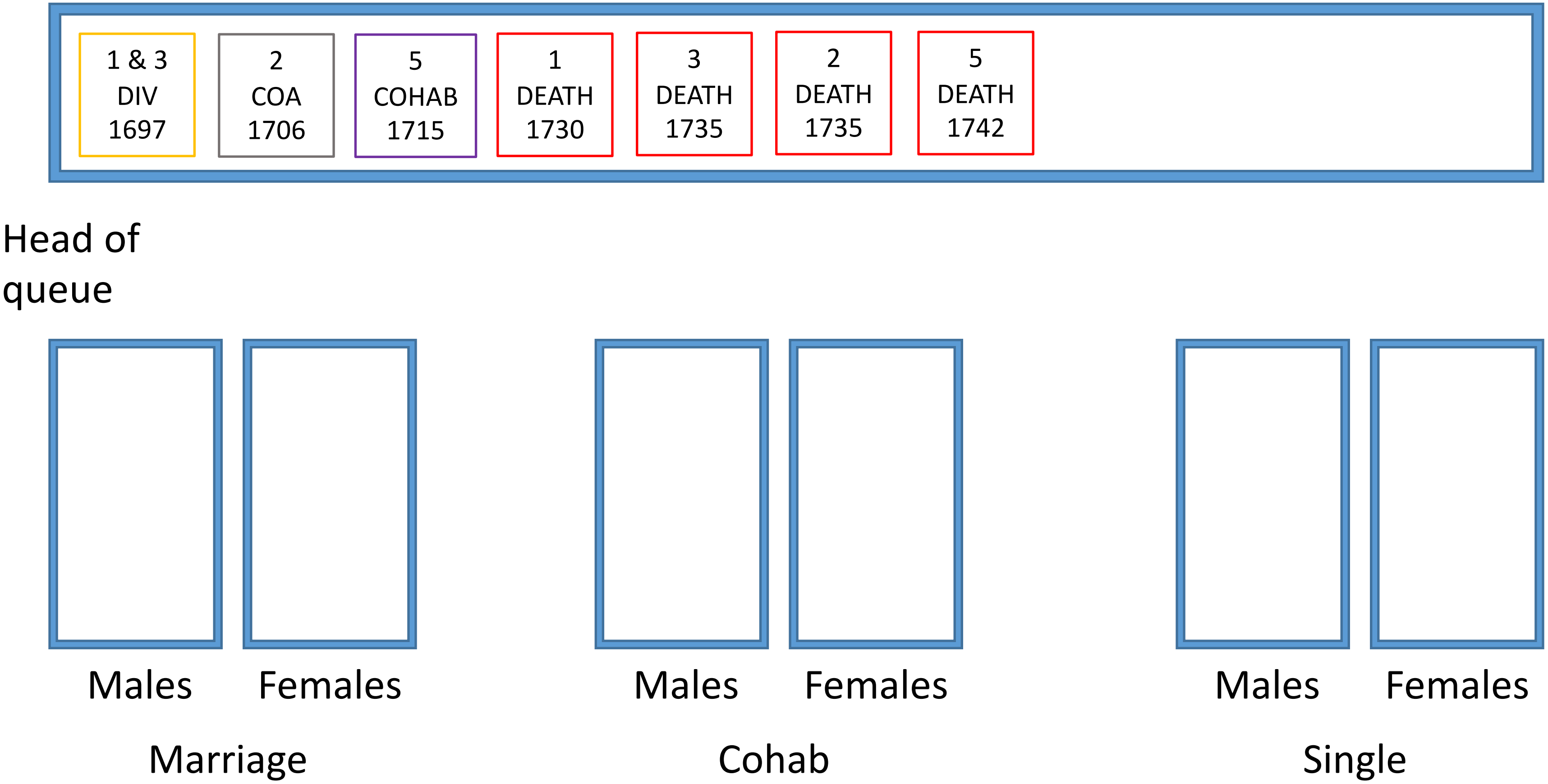
Males

Females

Marriage

OPM – Event Handling

5
BORN
1695



1 & 3
DIV
1697

2
COA
1706

5
COHAB
1715

1
DEATH
1730

3
DEATH
1735

2
DEATH
1735

5
DEATH
1742

Head of queue

Males Females

Marriage

Males Females

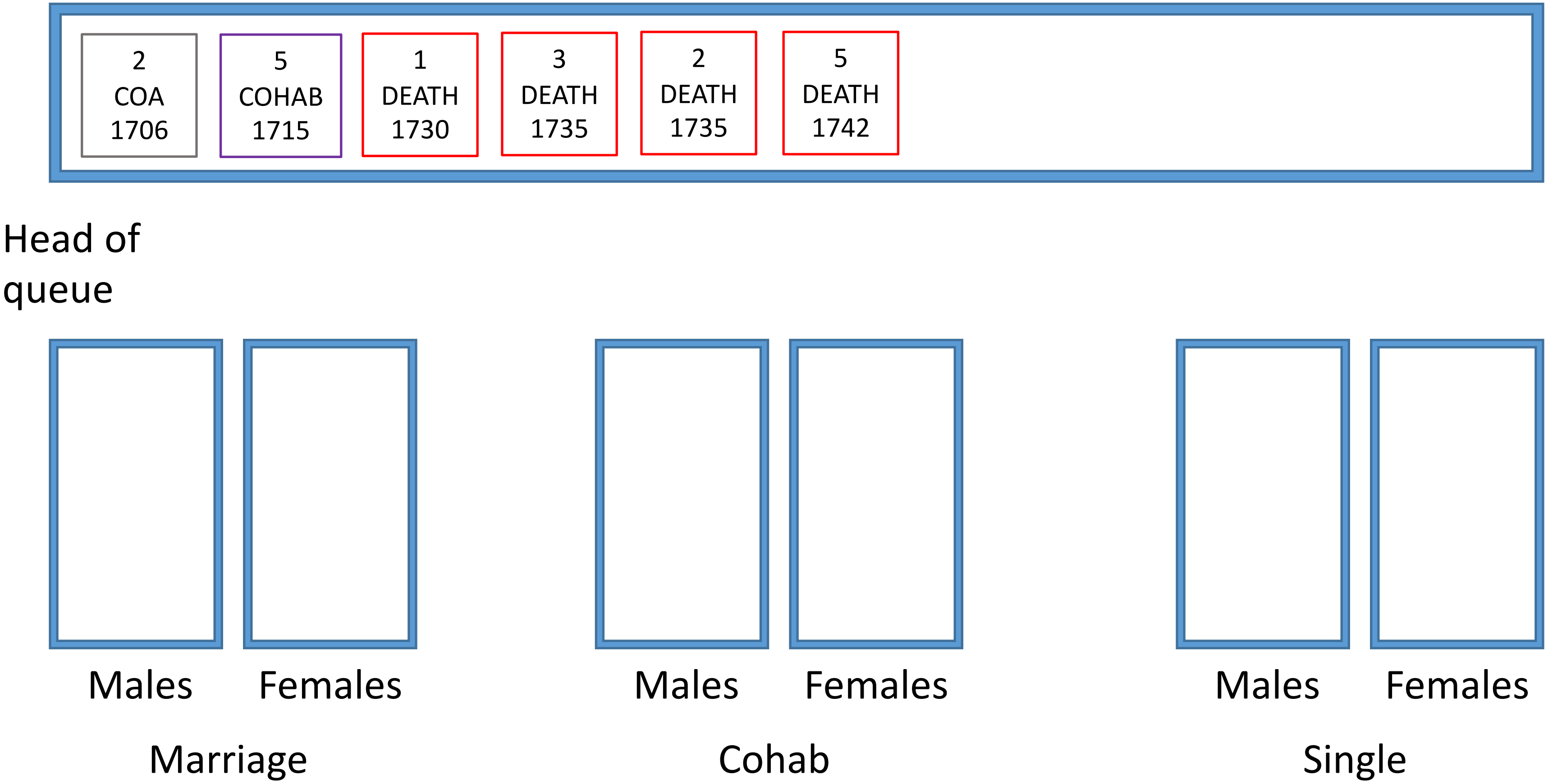
Cohab

Males Females

Single

OPM – Event Handling

1 & 3
DIV
1697



2
COA
1706

5
COHAB
1715

1
DEATH
1730

3
DEATH
1735

2
DEATH
1735

5
DEATH
1742

Head of queue

Males Females

Marriage

Males Females

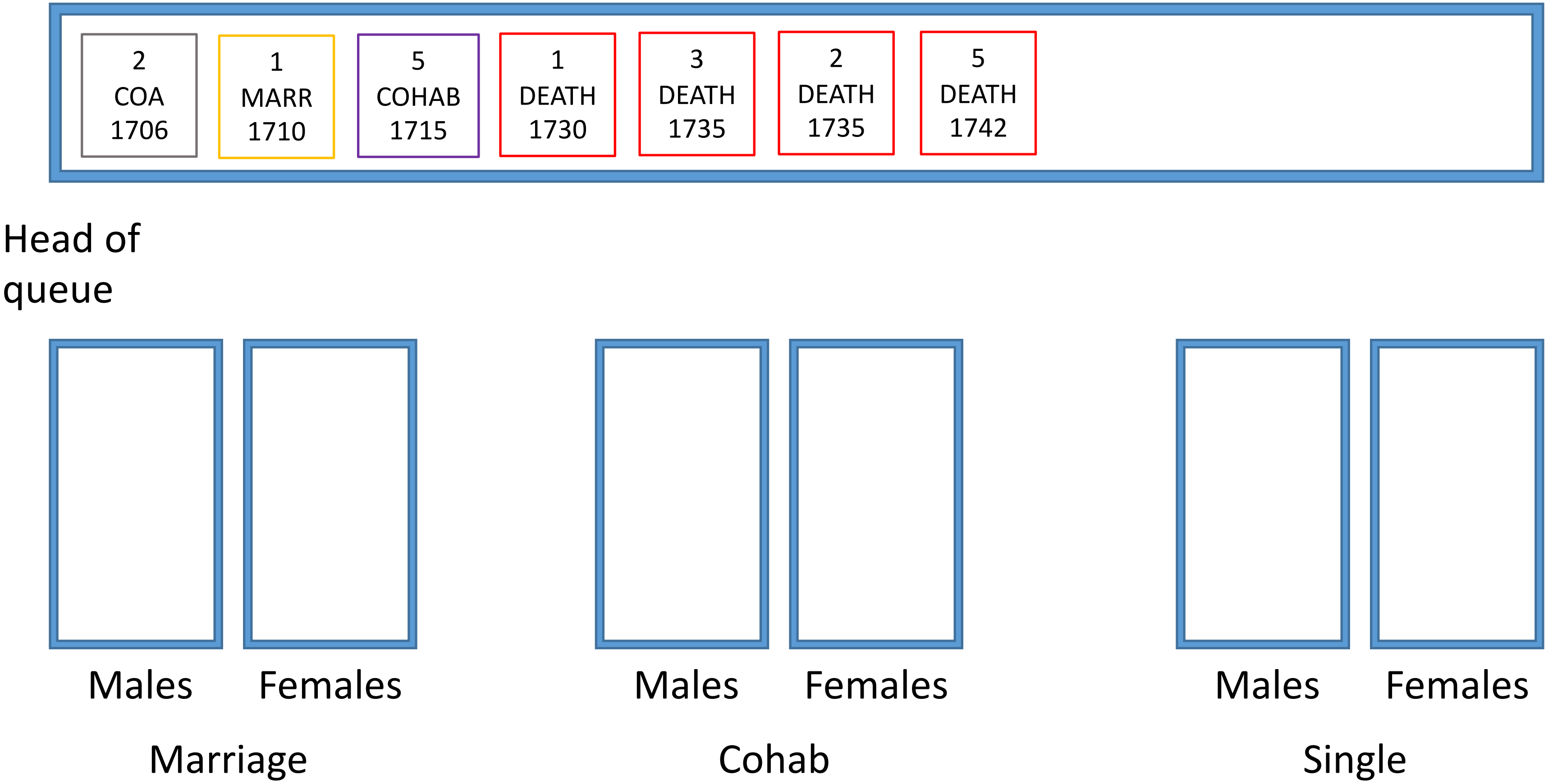
Cohab

Males Females

Single

OPM – Event Handling

1 & 3
DIV
1697



2
COA
1706

1
MARR
1710

5
COHAB
1715

1
DEATH
1730

3
DEATH
1735

2
DEATH
1735

5
DEATH
1742

Head of queue

Males Females

Marriage

Males Females

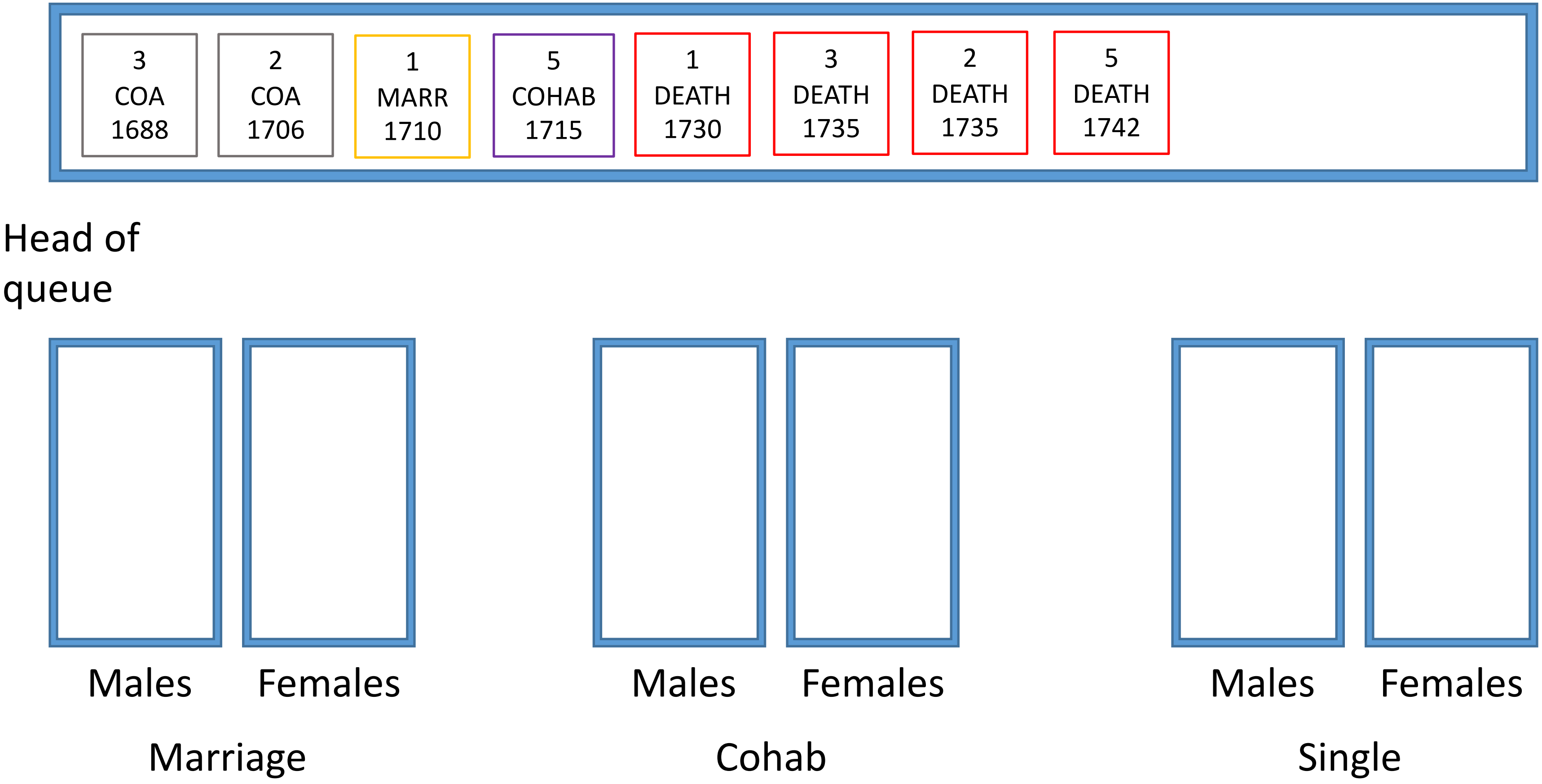
Cohab

Males Females

Single

OPM – Event Handling

1 & 3
DIV
1697



3
COA
1688

2
COA
1706

1
MARR
1710

5
COHAB
1715

1
DEATH
1730

3
DEATH
1735

2
DEATH
1735

5
DEATH
1742

Head of queue

Males Females

Marriage

Males Females

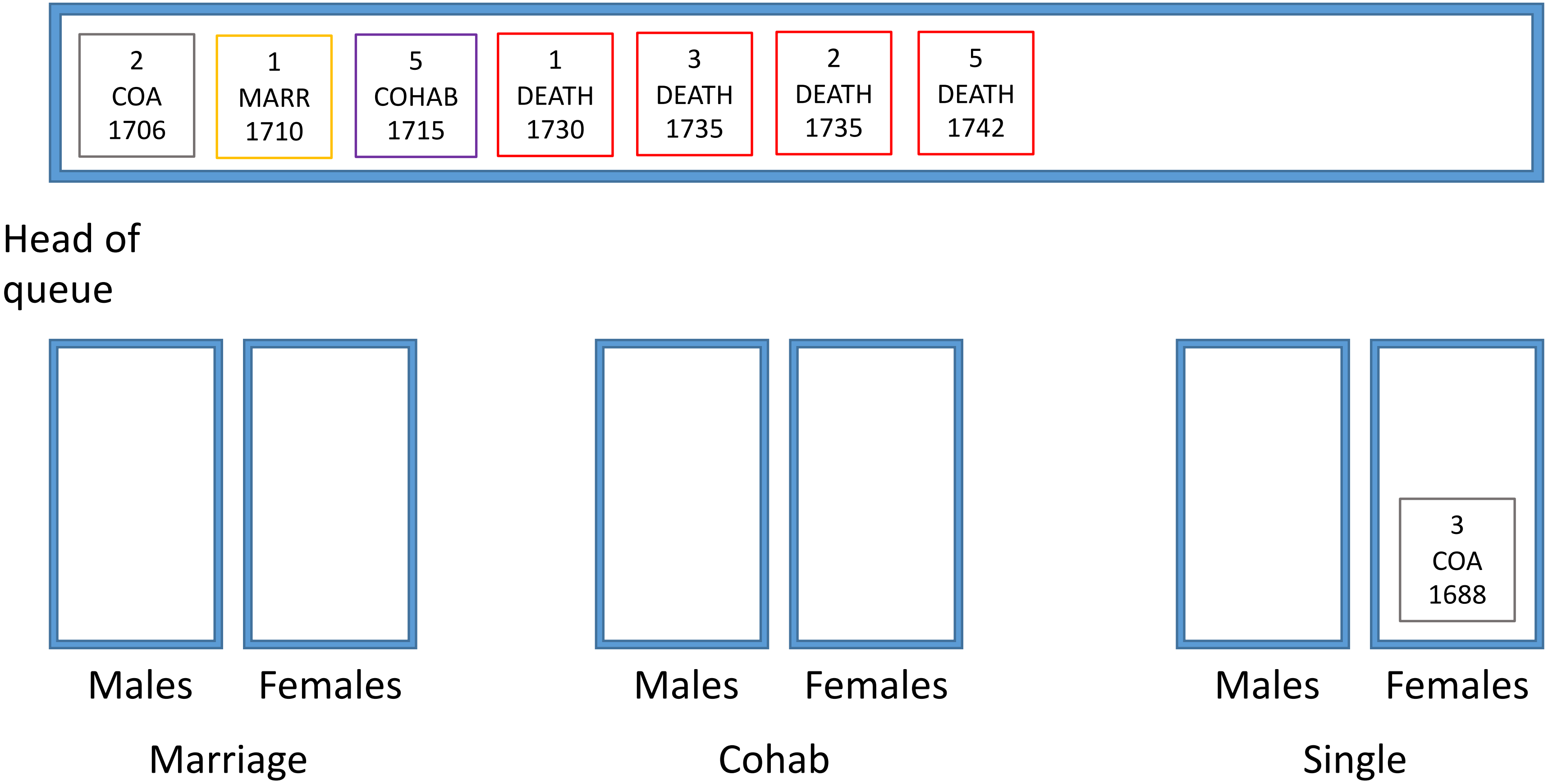
Cohab

Males Females

Single

OPM – Event Handling

3
COA
1688



2 COA 1706

1 MARR 1710

5 COHAB 1715

1 DEATH 1730

3 DEATH 1735

2 DEATH 1735

5 DEATH 1742

Head of queue

3 COA 1688

Males Females

Marriage

Males Females

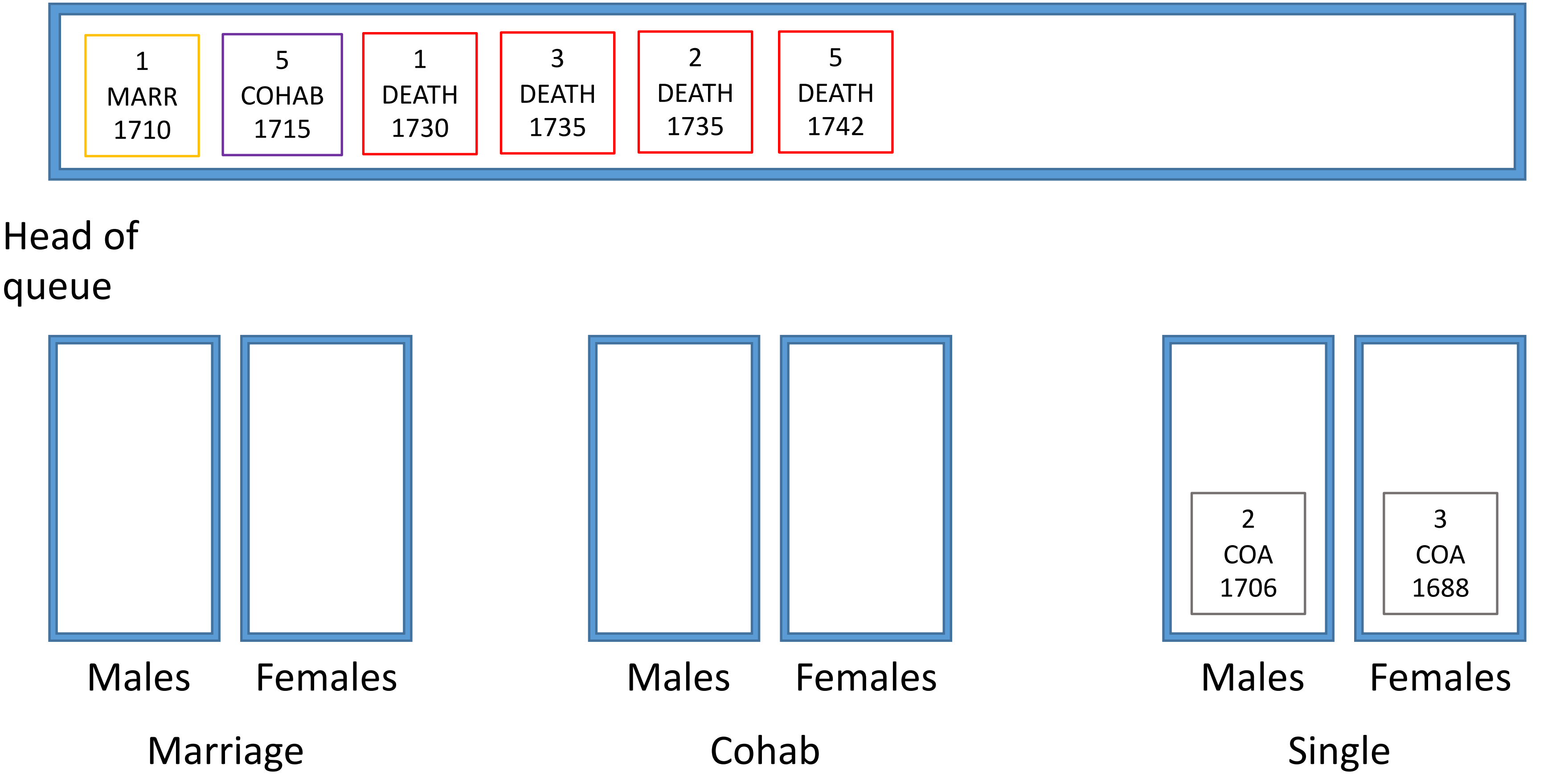
Cohab

Males Females

Single

OPM – Event Handling

2
COA
1706



1
MARR
1710

5
COHAB
1715

1
DEATH
1730

3
DEATH
1735

2
DEATH
1735

5
DEATH
1742

Head of
queue

2
COA
1706

3
COA
1688

Males Females

Marriage

Males Females

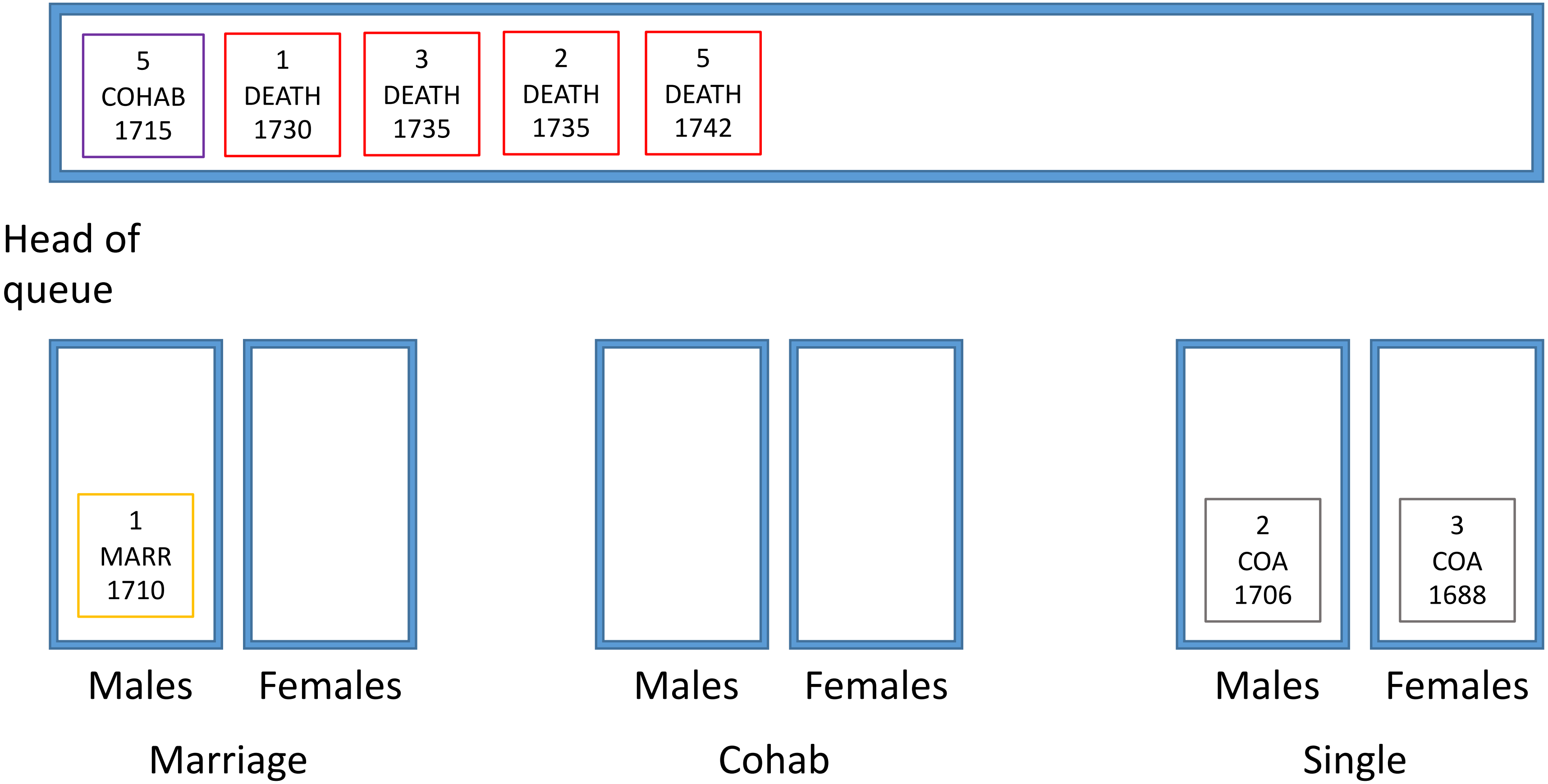
Cohab

Males Females

Single

OPM – Event Handling

1
MARR
1710



5
COHAB
1715

1
DEATH
1730

3
DEATH
1735

2
DEATH
1735

5
DEATH
1742

Head of queue

1
MARR
1710

2
COA
1706

3
COA
1688

Males Females

Marriage

Males Females

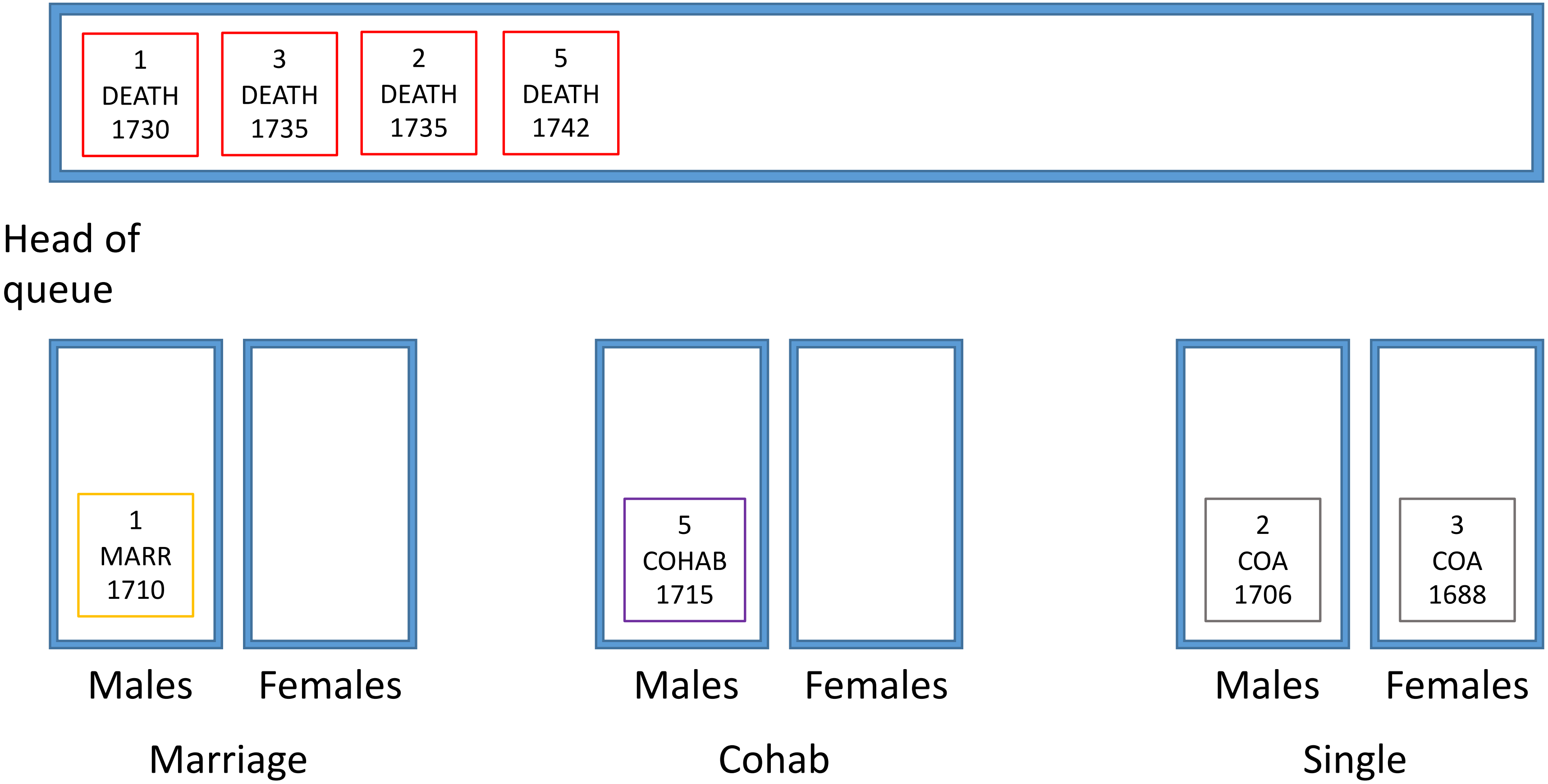
Cohab

Males Females

Single

OPM – Event Handling

5
COHAB
1715



1
DEATH
1730

3
DEATH
1735

2
DEATH
1735

5
DEATH
1742

Head of
queue

1
MARR
1710

5
COHAB
1715

2
COA
1706

3
COA
1688

Males Females

Marriage

Males Females

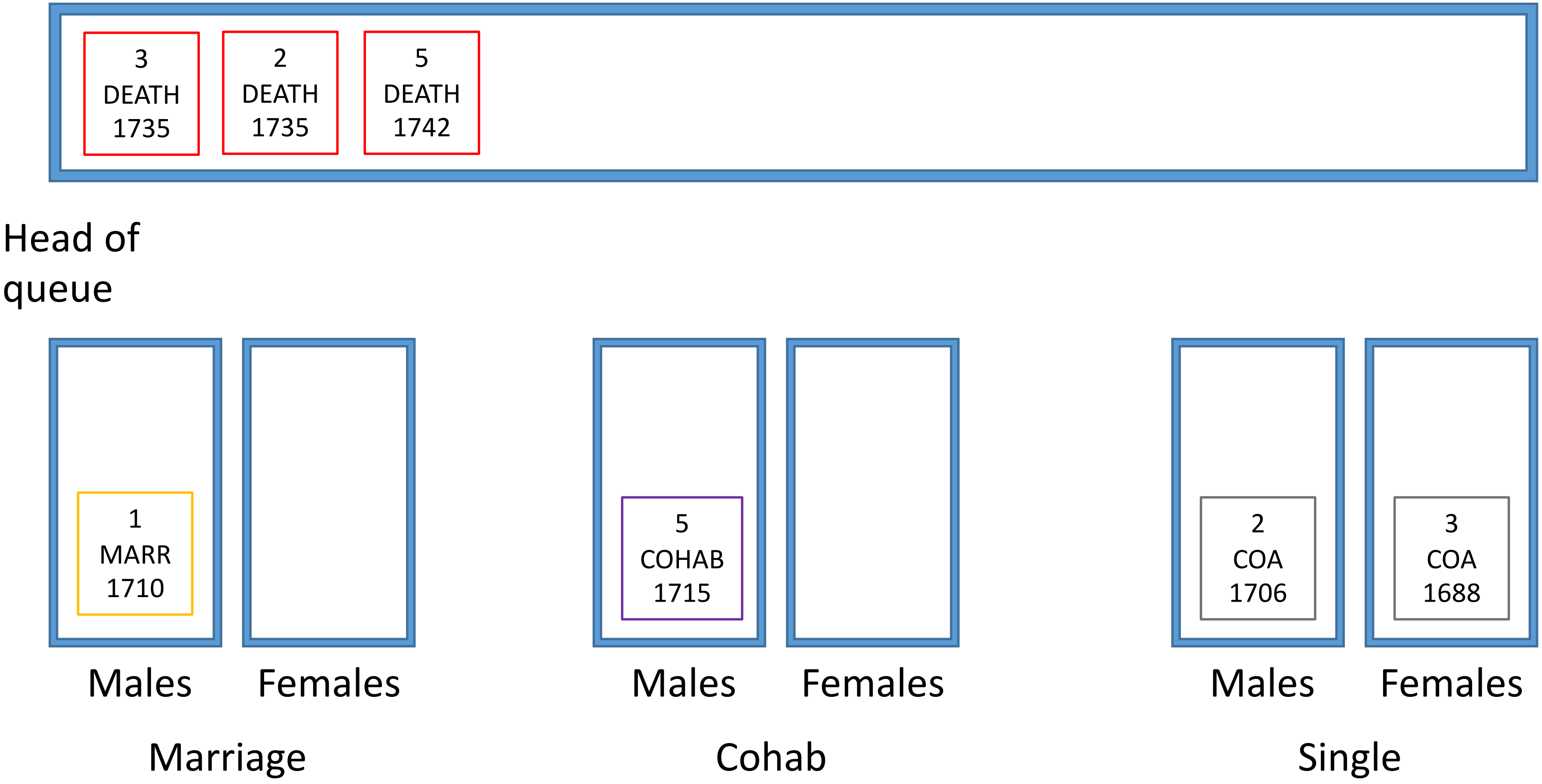
Cohab

Males Females

Single

OPM – Event Handling

1
DEATH
1730



3
DEATH
1735

2
DEATH
1735

5
DEATH
1742

Head of queue

1
MARR
1710

5
COHAB
1715

2
COA
1706

3
COA
1688

Males Females

Marriage

Males Females

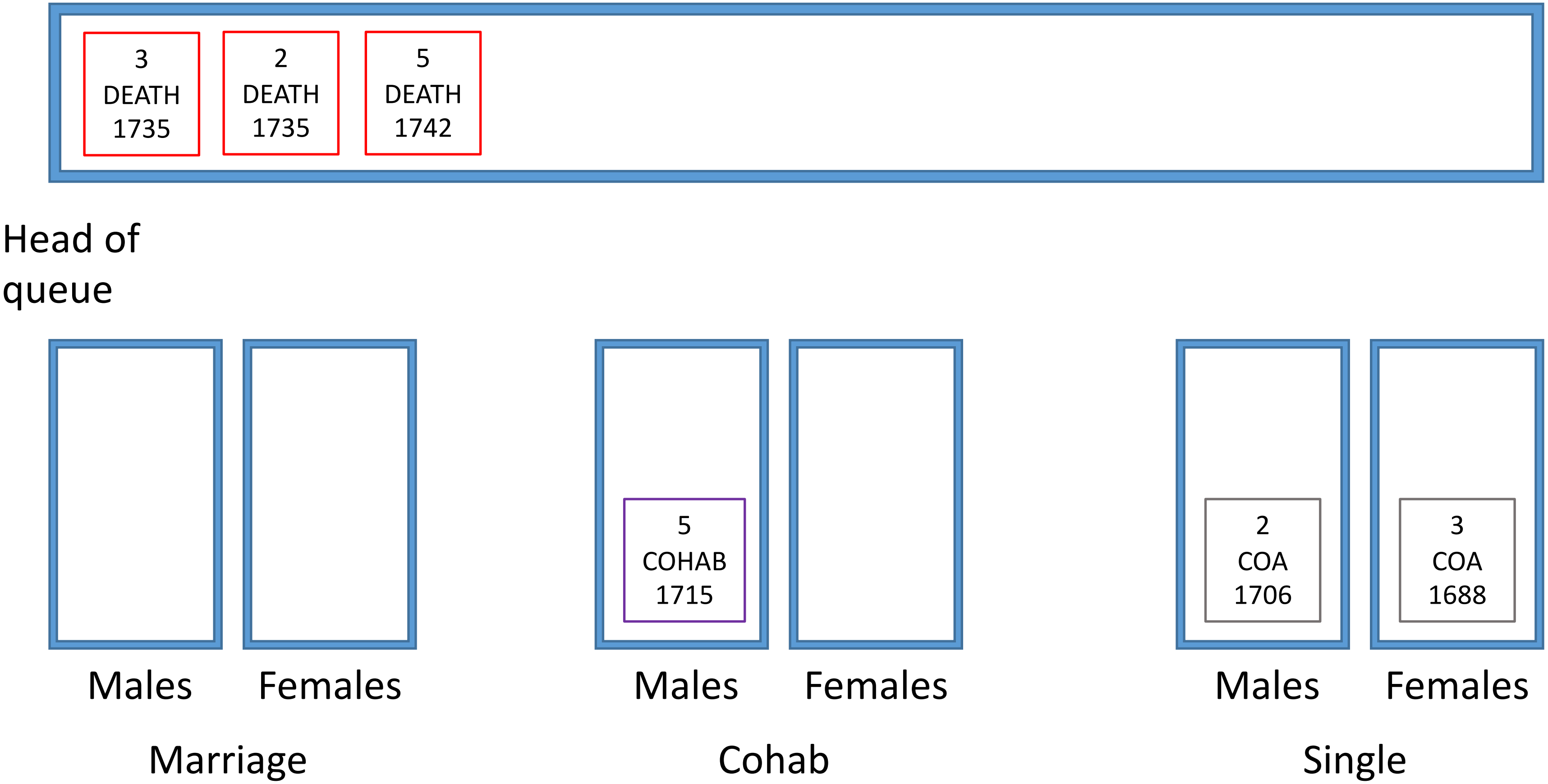
Cohab

Males Females

Single

OPM – Event Handling

1
DEATH
1730



3
DEATH
1735

2
DEATH
1735

5
DEATH
1742

Head of queue

5
COHAB
1715

2
COA
1706

3
COA
1688

Males Females

Marriage

Males Females

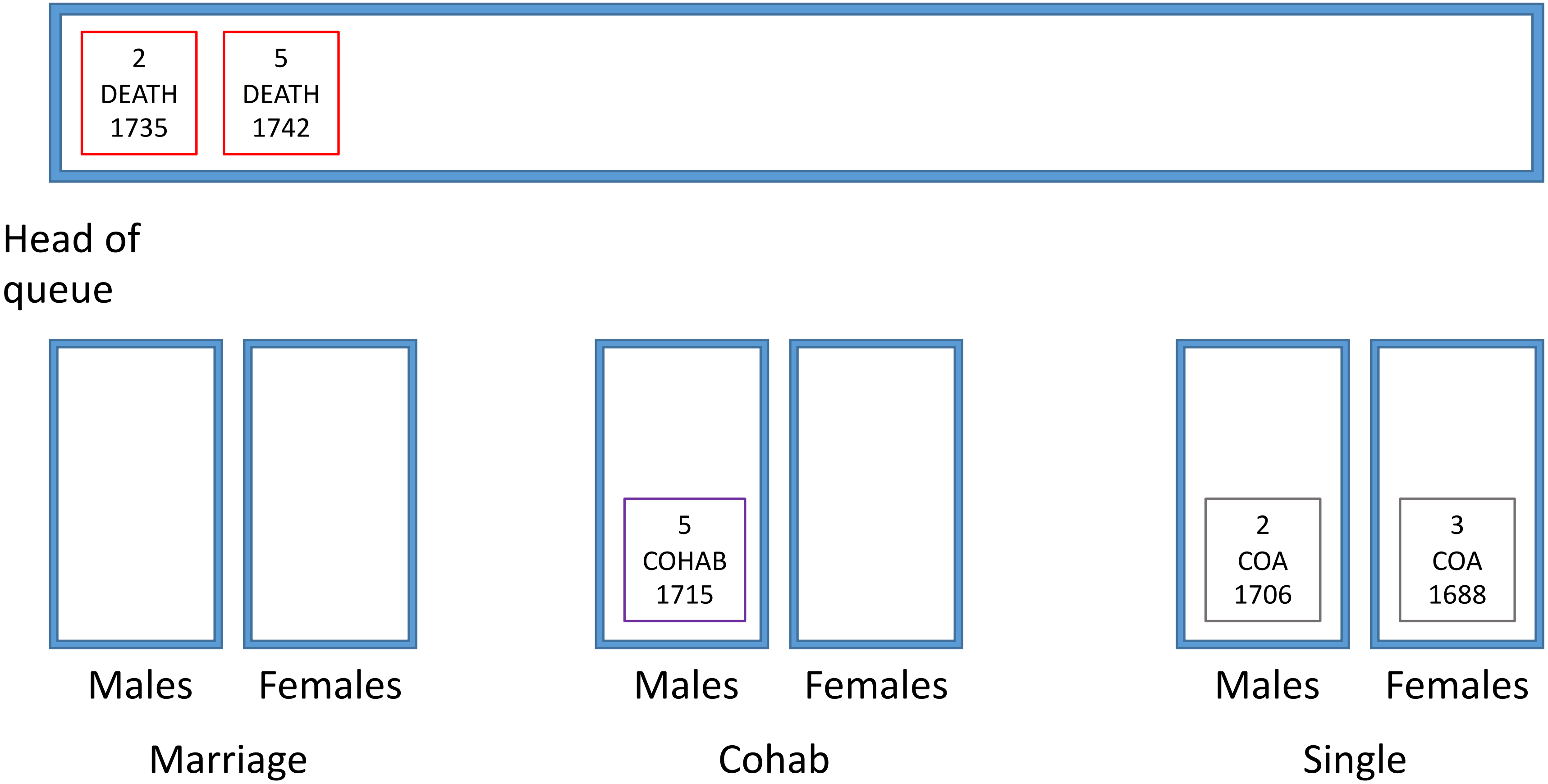
Cohab

Males Females

Single

OPM – Event Handling

3
DEATH
1735



2
DEATH
1735

5
DEATH
1742

Head of queue

5
COHAB
1715

2
COA
1706

3
COA
1688

Males Females

Marriage

Males Females

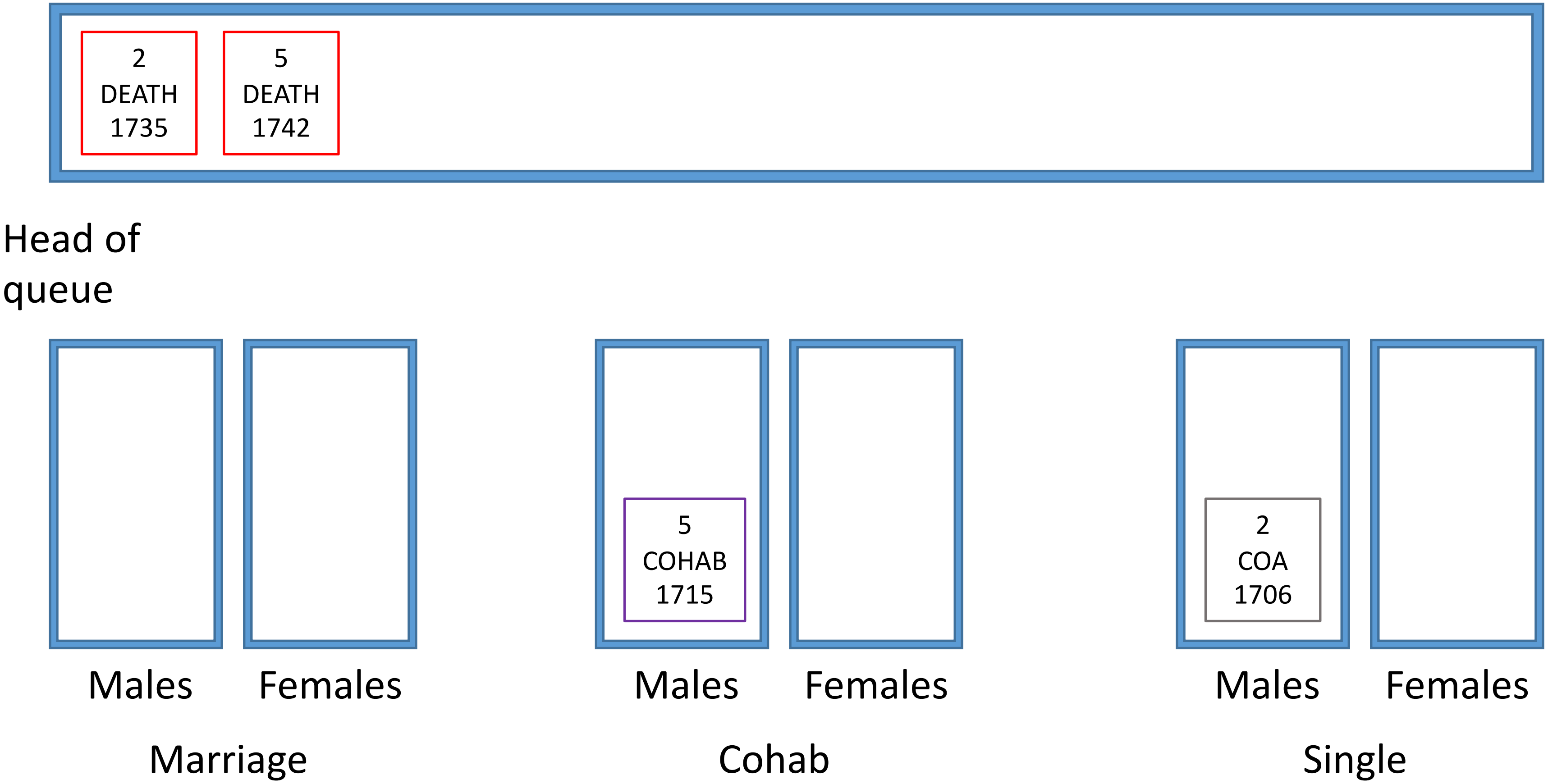
Cohab

Males Females

Single

OPM – Event Handling

3
DEATH
1735



2
DEATH
1735

5
DEATH
1742

Head of
queue

5
COHAB
1715

2
COA
1706

Males Females

Marriage

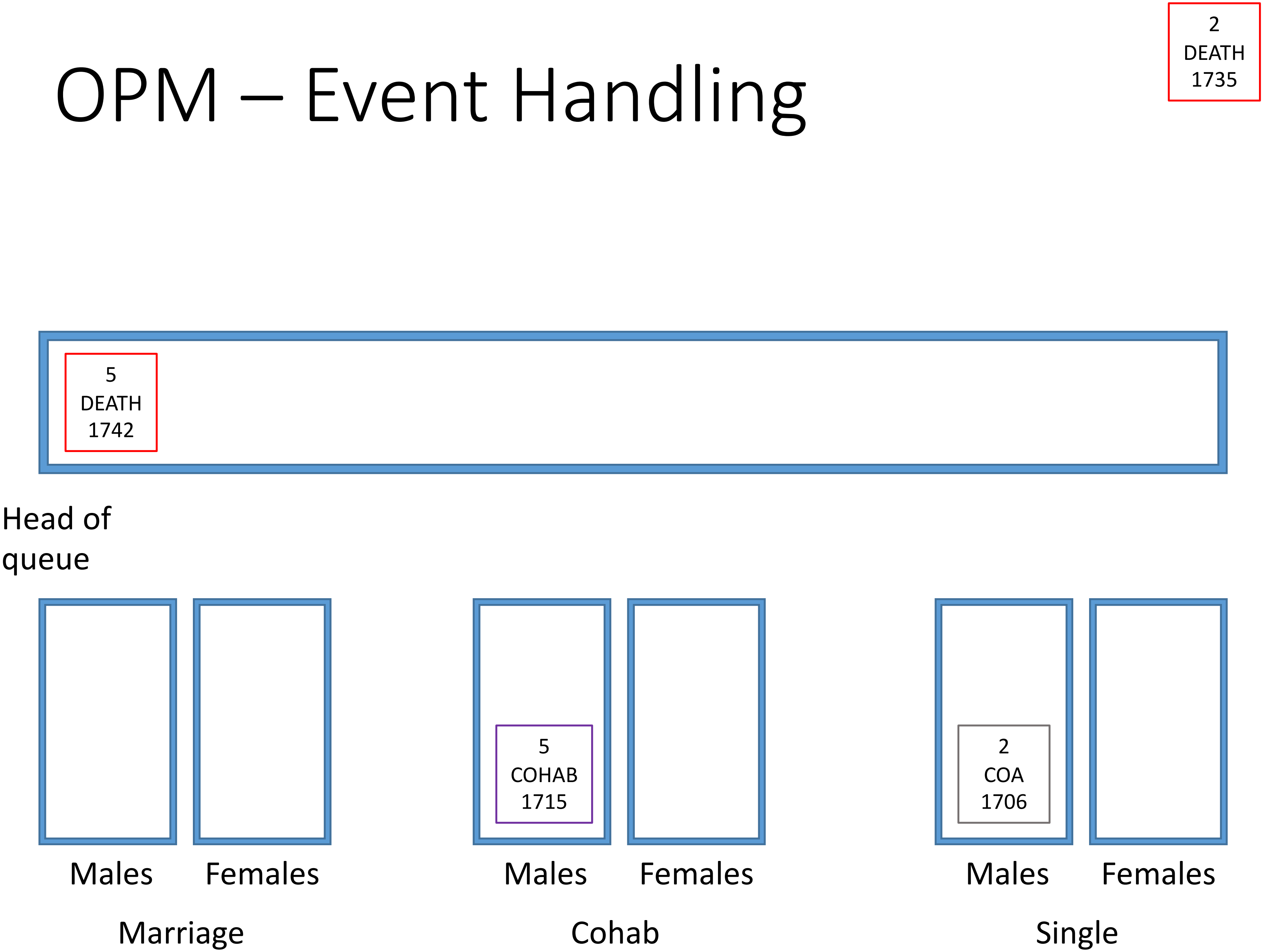
Males Females

Cohab

Males Females

Single

OPM – Event Handling



2
DEATH
1735

5
DEATH
1742

Head of queue

5
COHAB
1715

2
COA
1706

Males

Females

Marriage

Males

Females

Cohab

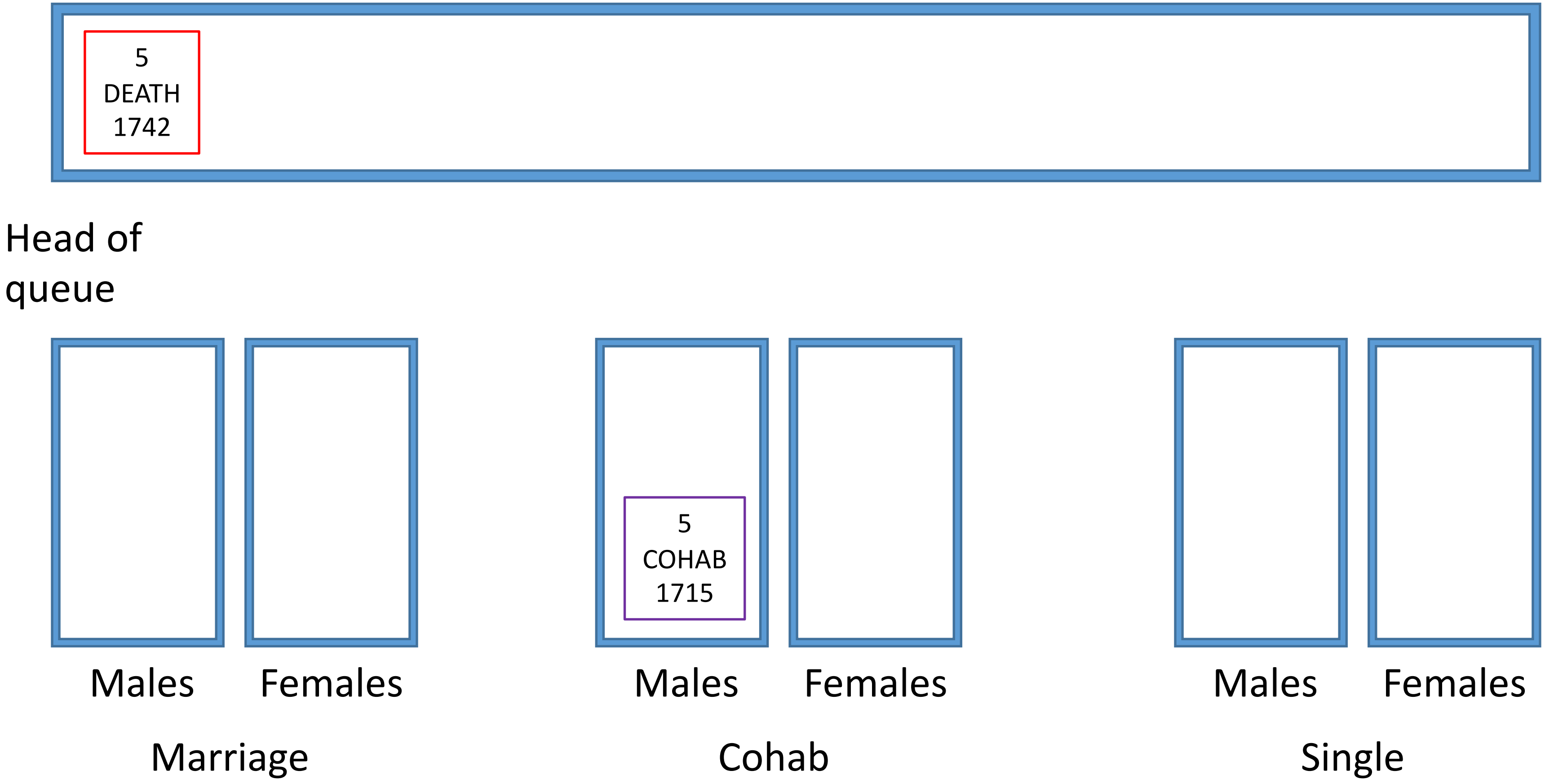
Males

Females

Single

OPM – Event Handling

2
DEATH
1735



5
DEATH
1742

Head of queue

5
COHAB
1715

Males Females

Marriage

Males Females

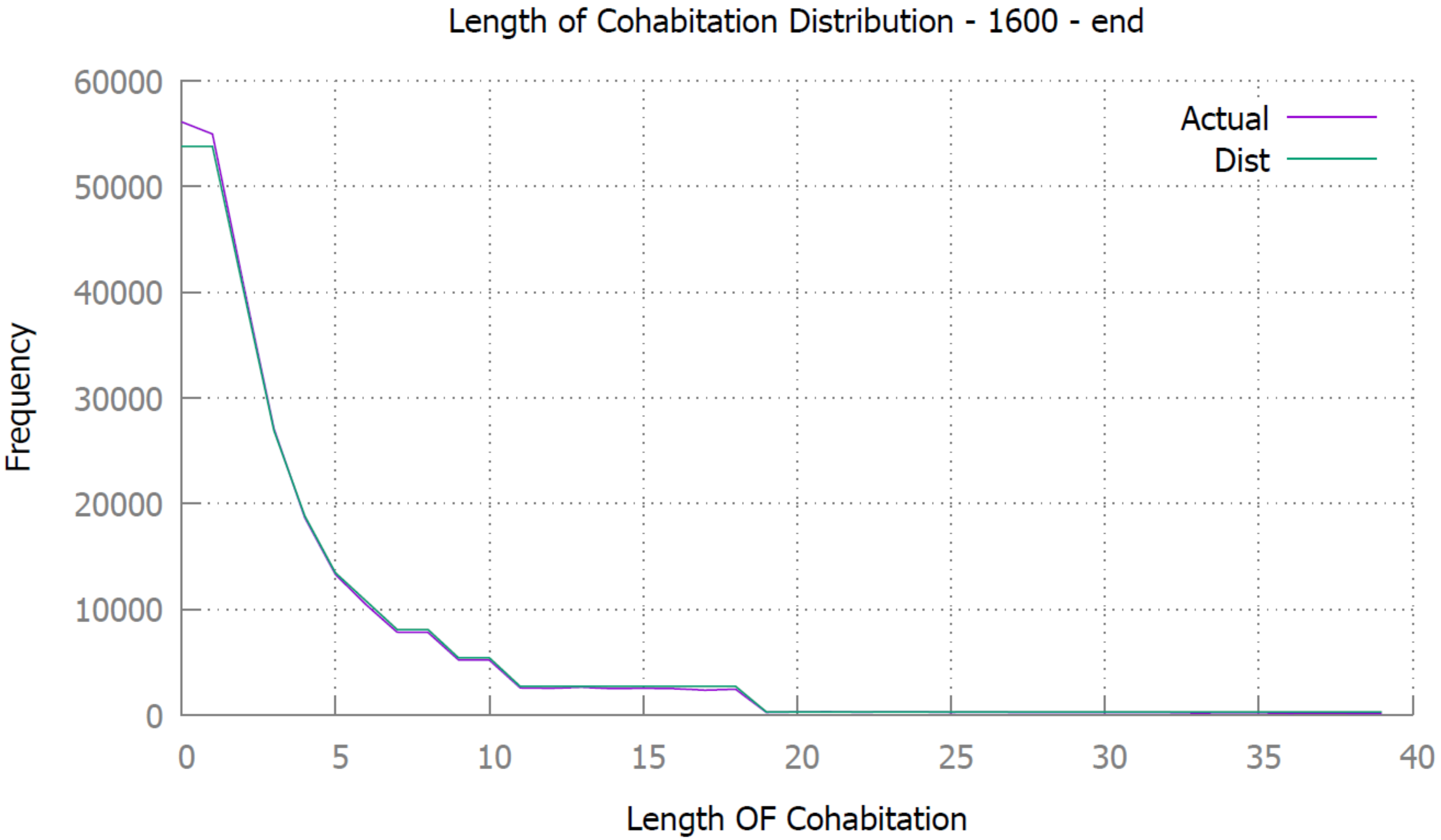
Cohab

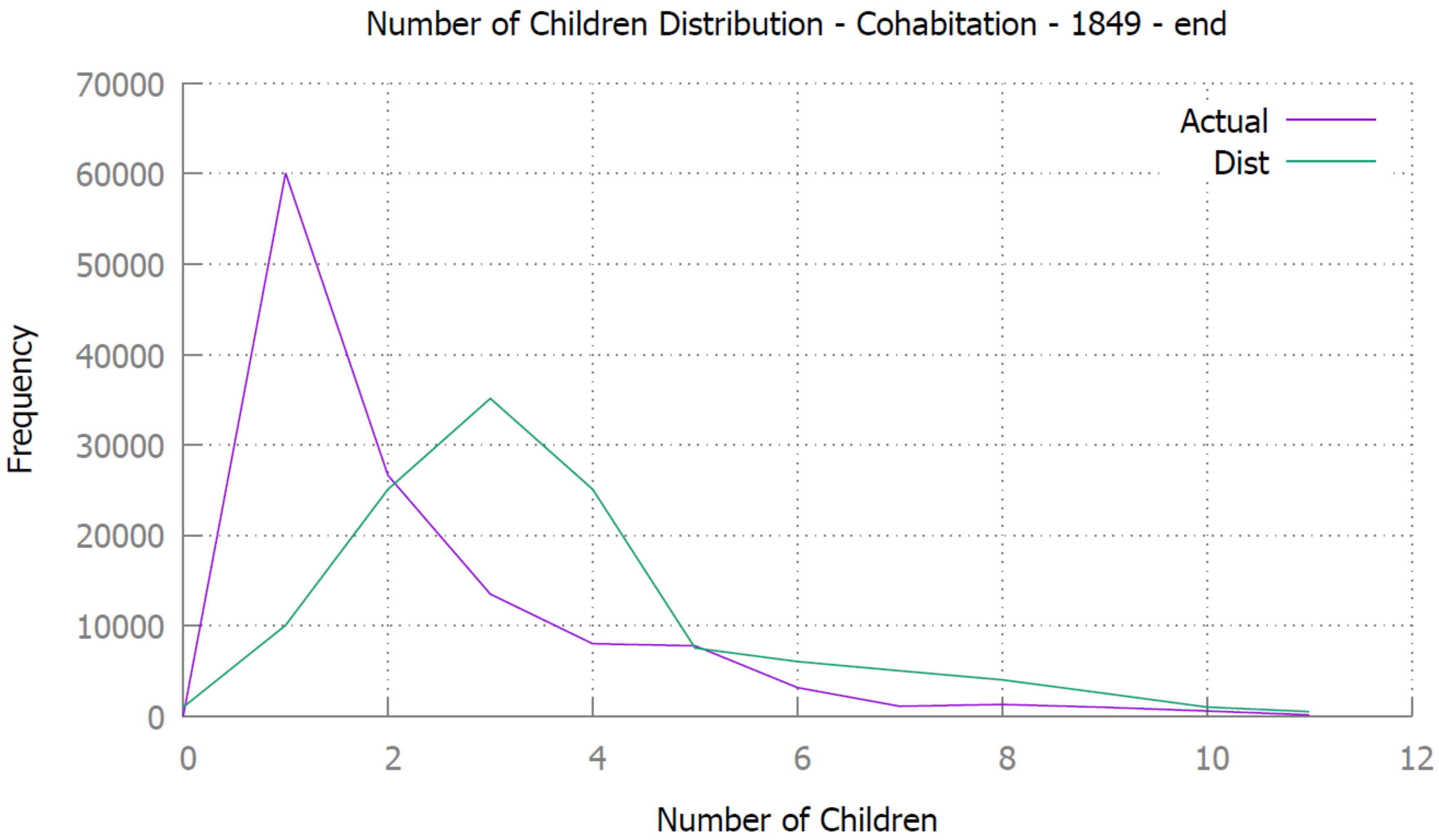
Males Females

Single

OPM - Problems

- Clashing of inputs
- Lack of expression in the model
  - Extraordinary Events
  - Quantification of inputs
- Verifying the generated population matched the desired inputs


Length of Cohabitation Distribution - 1600 - end
Frequency
Length OF Cohabitation
Actual
Dist
0
10000
20000
30000
40000
50000
60000
0
5
10
15
20
25
30
35
40


Number of Children Distribution - Cohabitation - 1849 - end
Actual
Dist
Frequency
70000
60000
50000
40000
30000
20000
10000
0
0
2
4
6
8
10
12
Number of Children

OPM - Problems

- Clashing of inputs
- Lack of expression in the model
  - Extraordinary Events
  - Quantification of inputs
- Verifying the generated population matched the desired inputs

OPM - Problems

- Clashing of inputs
- Lack of expression in the model
  - Extraordinary Events
  - Quantification of inputs
- Verifying the generated population matches the desired inputs

Verified Population Model

To produce a synthetic population

- A graph (tree structure) representing the true linkage of the population
- The event records for the population

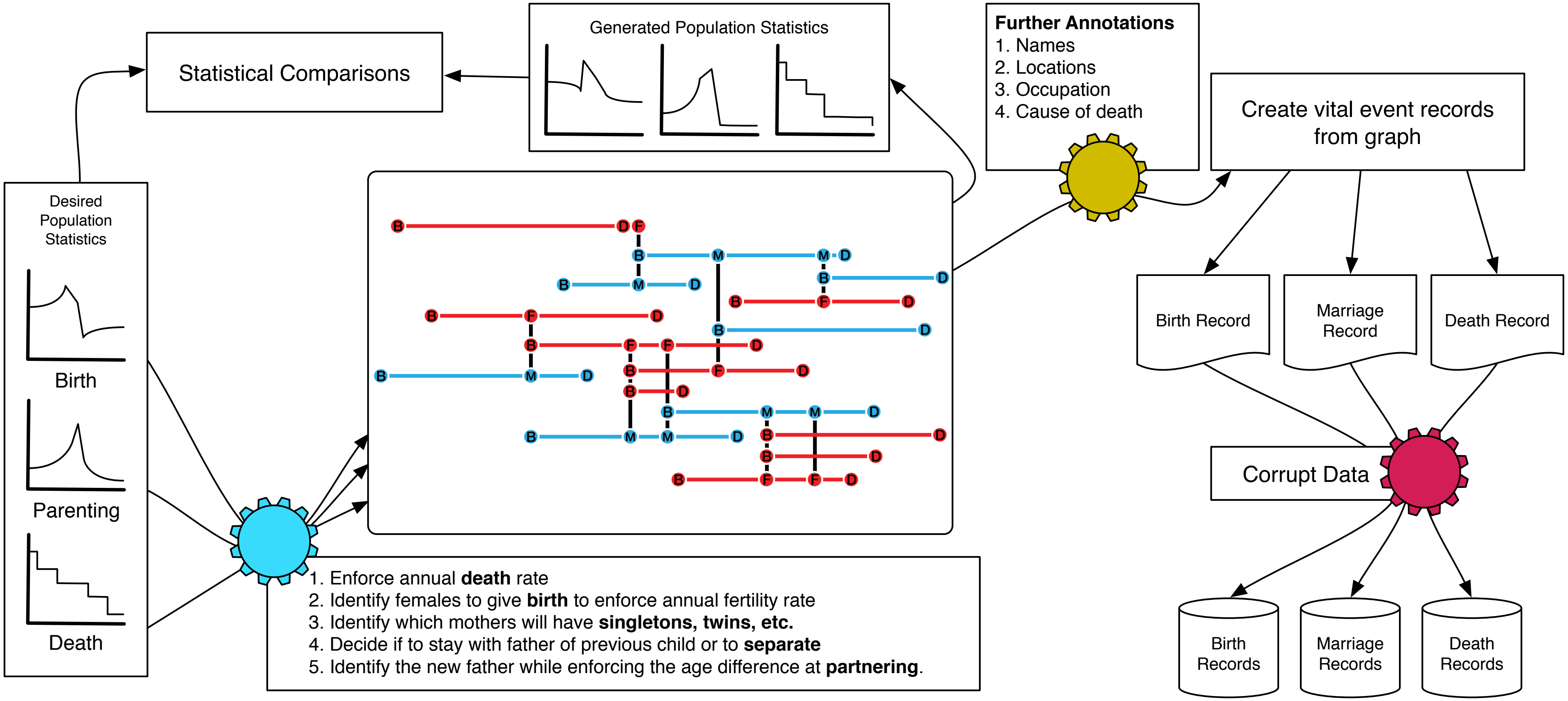
Based on a range of summative input statistics

- Ordered birth rates, death rates, parenting

Statistically verifiable

- against input statistics
- against secondary ‘unseen’ statistics
- ? ‘Turing test’

VPM – Overview



Further Annotations
1. Names
2. Locations
3. Occupation
4. Cause of death

1. Enforce annual **death** rate
2. Identify females to give **birth** to enforce annual fertility rate
3. Identify which mothers will have **singletons, twins, etc.**
4. Decide if to stay with father of previous child or to **separate**
5. Identify the new father while enforcing the age difference at **partnering**.

VPM – Overview

- Inputs
- Integrity and Initialisation
- Simulation approach
    - Simulation
    - Self-correction
- Validation
    - Kaplan Meier
    - ANOVA

VPM – Overview

- Inputs
- Integrity and Initialisation
- Simulation approach
    - Simulation
    - Self-correction
- Validation
    - Kaplan Meier
    - ANOVA

VPM – Inputs

Genealogical controlling inputs are variable over time

Annotations

- female first name
- male first name
- surname
- occupation
- cause of death
- address

Seed

- ~~seed age for males~~
- ~~seed age for females~~

Birth

- ~~children number of in cohab~~
- ~~children number of in cohab then marriage~~
- ~~children number of in marriage~~

- ordered birth rates

- children number of in pregnancy

Partnering

- ~~partnership characteristic~~
- ~~partnership remarriage characteristic~~

- ~~marriage age for males~~
- ~~marriage age for females~~

- ~~cohabitation age for males~~
- ~~cohabitation age for females~~

- ~~cohabitation to marriage time~~
- ~~cohabitation length~~

- age difference at partnering

Death

- ~~death age at~~

- lifetable

Separation

- ~~divorce age for male~~
- ~~divorce age for female~~

- ~~divorce instigated by gender~~
- ~~divorce reason male~~
- ~~divorce reason female~~

- ~~divorce remarriage boolean~~
- ~~remarriage time to~~

- separation following number of children in partnership

Genealogical complexity

- ~~affair number of~~

- ~~affair number of children~~
- ~~affair with single or married~~

VPM – Inputs

- Life tables
  - Age at death
  - Sudden changes in death rate

YEAR	1630
POPULATION	SCOTLAND
SOURCE	ONS
VAR	DEATH
FORM	RATE
GENDER	M
DATA	
0	0.012996
1	0.000945
2	0.000572
3	0.000532
4	0.000403
5	0.00038
6	0.000345
7	0.000237
8	0.000323
9	0.000293
10	0.000248
11	0.00037
12	0.000324

…

81	0.138126
82	0.153255
83	0.170838
84	0.180342
85	0.197232
86	0.197111
87	0.223026
88	0.237387
89	0.237154
90	0.266047
91	0.293669
92	0.289936
93	0.267021
94	0.34
95	0.406818
96	0.415323
97	0.397727
98	0.371429
99	0.532258
100+	0.909091

VPM – Inputs

- Ordered Birth Table
  - Fertility rate (TFR and ASFR)
  - Age of females at birth and partnering
  - Controls family size – paired with **separation**

YEAR	1980				
POPULATION	ENGWALES				
SOURCE	ONS				
VAR	BIRTH				
TYPE	ORDERED				
FORM	RATE				
LABELS	0	1	2	3	4+
DATA					
15	0.003	0	0	0	0
16	0.01067	0.00033	0	0	0
17-19	0.0386209	0.006538	0.0015411	0	0
20-24	0.069174	0.020412	0.018144	0.004536	0.001134
25-29	0.04008	0.02672	0.044088	0.016032	0.00668
30-34	0.011442424	0.010012121	0.030751515	0.012872727	0.005721212
35-39	0.0022	0.00308	0.00946	0.00462	0.00264
40-49	0.000264	0.000312	0.000864	0.000528	0.000432

VPM - Inputs

- Multiple births in pregnancy
  - Twinning

YEAR	2013			
POPULATION	ENGWALES			
SOURCE	ONS			
VAR	MULTIPLE_BIRTH			
FORM	RATE			
LABELS	1	2	3	4
DATA				
15-19	0.994061	0.00587	0.000069	0
20-24	0.991185	0.008714	0.000101	0
25-29	0.987437	0.012378	0.00018	0.000005
30-34	0.982819	0.016912	0.000268	0
35-39	0.977418	0.022068	0.000495	0.000018
40-44	0.972353	0.027117	0.00053	0
45-49	0.906608	0.089022	0.004369	0

VPM – Inputs

- Partnering
  - Age difference at partnering
  - Male age at partnering

POPULATION	ENGWALES							
SOURCE	ONS							
VAR	PARTNERING							
TYPE	FEMALE_AGES_ON_ROWS							
FORM	PROPORTIONS							
LABELS	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-100
DATA								
15-19	0.1868	0.5580	0.1784	0.0502	0.0173	0.0058	0.0021	0.0015
20-24	0.0211	0.4409	0.3663	0.1140	0.0373	0.0133	0.0045	0.0026
25-29	0.0048	0.1247	0.4497	0.2677	0.1026	0.0318	0.0118	0.0068
30-34	0.0030	0.0567	0.2149	0.3662	0.2124	0.0910	0.0366	0.0192
35-39	0.0024	0.0325	0.1214	0.2248	0.2983	0.1846	0.0841	0.0518
40-44	0.0016	0.0185	0.0749	0.1340	0.2111	0.2622	0.1745	0.1232
45-49	0.0004	0.0125	0.0600	0.1009	0.1459	0.1784	0.2459	0.2559

VPM – Inputs

- Separation following number of children in partnership
  - Family size
  - A genealogy focused way of modelling separation

YEAR	1981
POPULATION	ENGWALES
SOURCE	ONS
VAR	SEPARATION
FORM	RATE
DATA	
1	0.003222
2	0.003425984
3	0.001090183
4	0.000281235
5+	7.27E-05

VPM – Overview

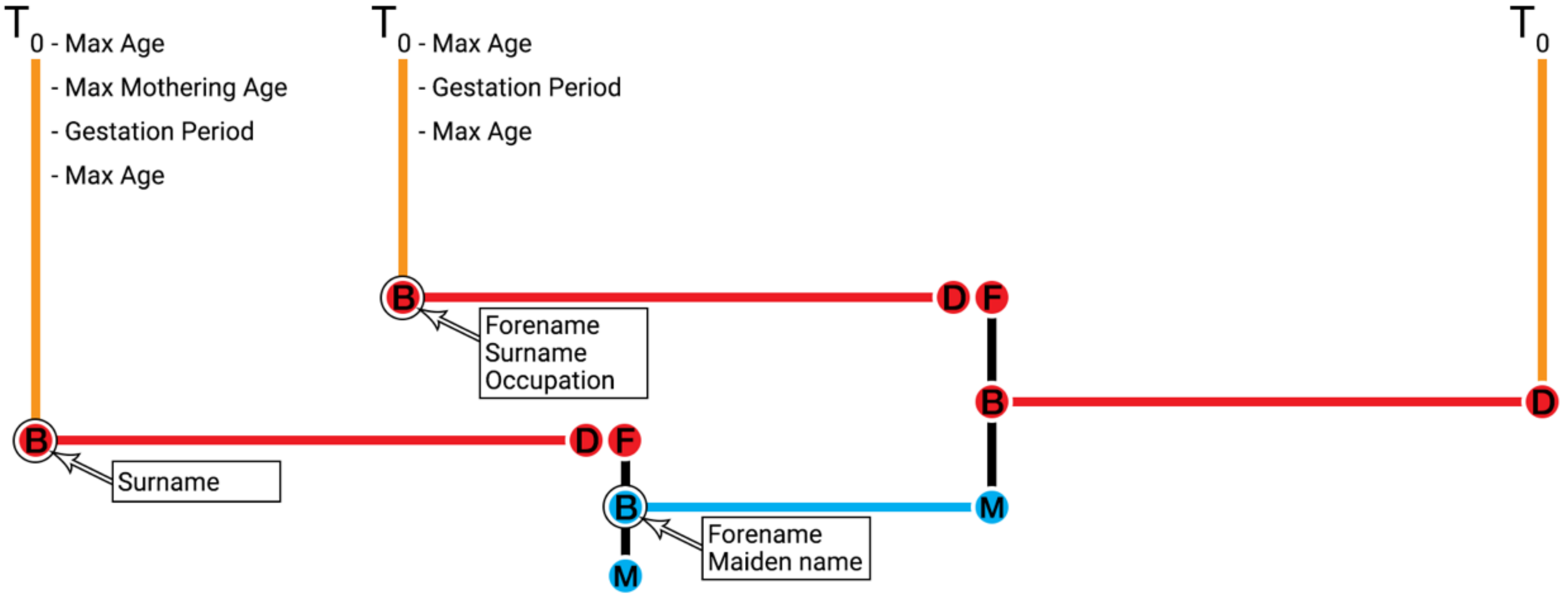
- Inputs
- Integrity and Initialisation
- Simulation approach
  - Simulation
  - Self-correction
- Validation
  - Kaplan Meier
  - ANOVA

VPM – Integrity

How far back from our ‘start date’?

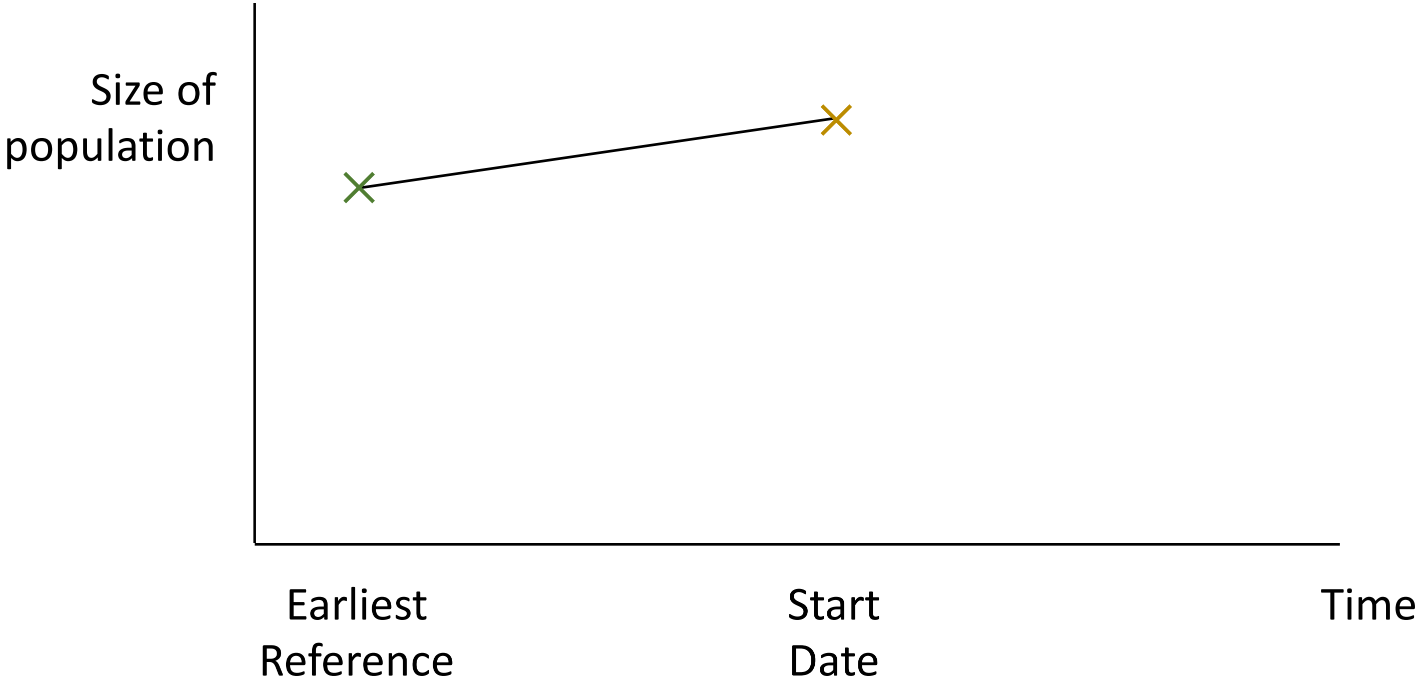
- Integrity
- Dependent on desired records

For a death certificate:

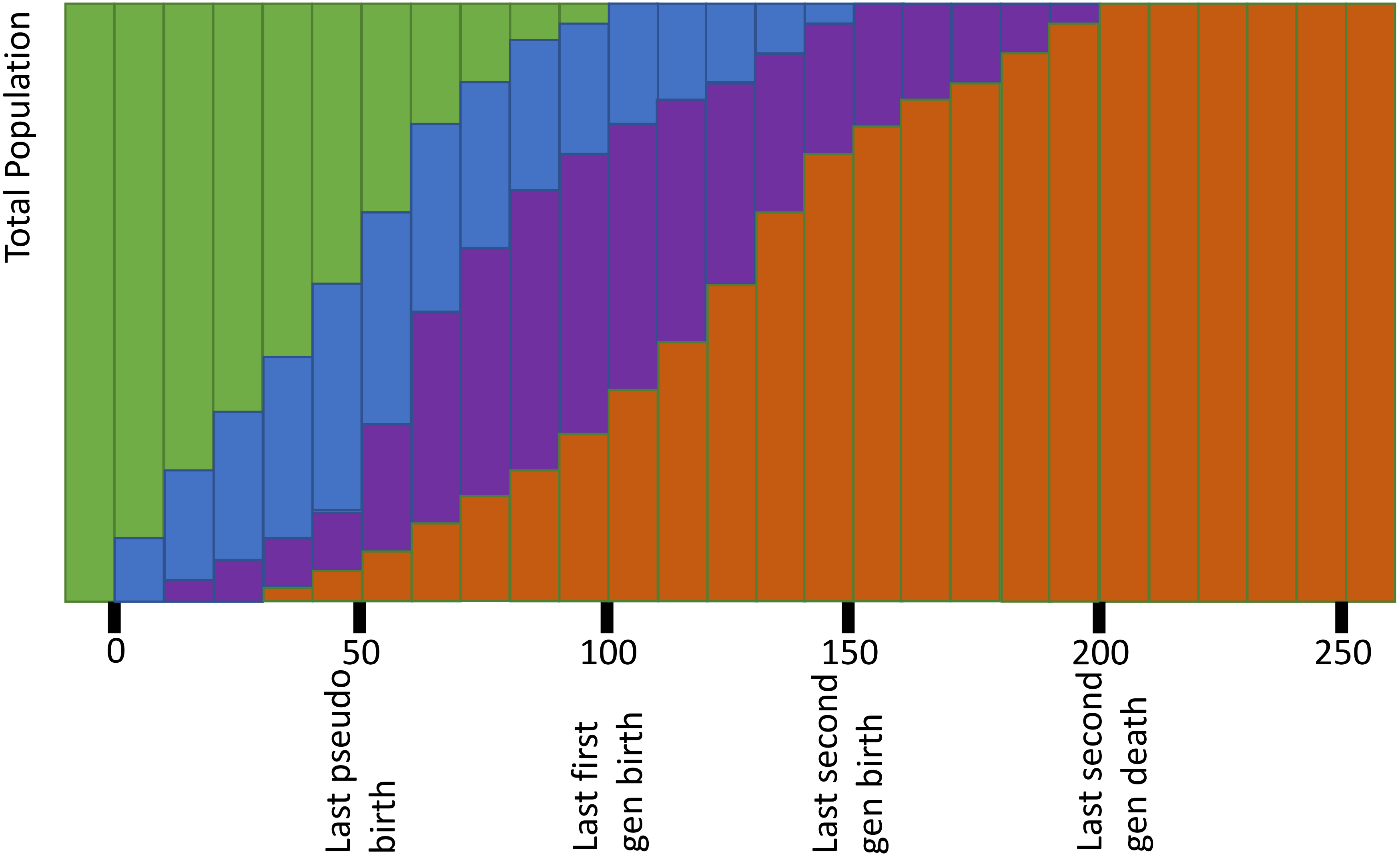


VPM – Initialisation

- Information known
  - Start Date
  - Desired initial population size
  - Earliest reference
  - Pre-model BR and DR



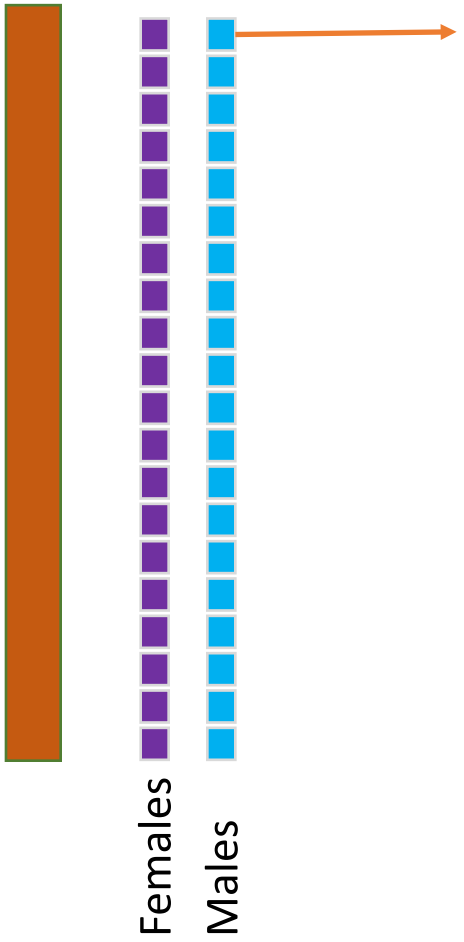
VPM – Initialisation



VPM – Overview

- Inputs
- Integrity and Initialisation
- Simulation approach
  - Simulation
  - Self-correction
- Validation
  - Kaplan Meier
  - ANOVA

VPM - Death



Death

Males aged 0

Population size	= S	= 250
nMx	= $_0M_1$	= 0.0130
Number to die	= $S \times nMx$	= 3.25

Select 3 males to die

Self correction

- Rounding errors

VPM - Birth

Males
Females

Birth

Females aged 20 with 2 children

Population size	= S	= 5000
nMx	= $_{20(2)}M_1$	= 0.069
Number to birth	= $S \times nMx$	= 345

Select 345 females to give birth

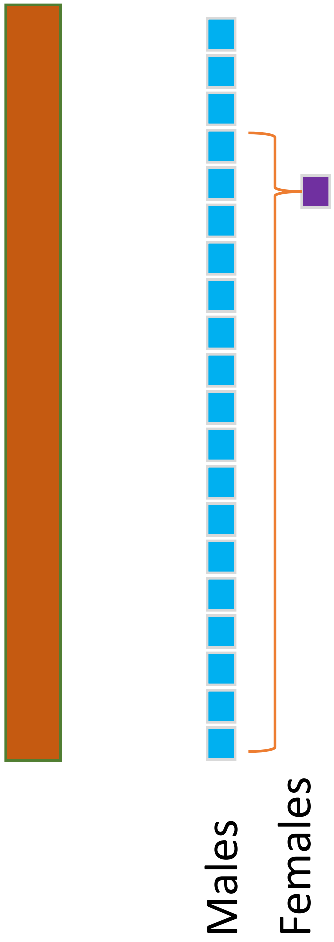
Separation

We have 345 females where they have had 2 children in a partnership

Population size	= S	= 345
nMx	= $_{2C}\mathrm{M}_{1C}$	= 0.0034
Number to sep.	= $S \times nMx$	= 1.173

Select 1 female to separate

VPM - Partnering



Partnering

We have 1 female selected to be a mother in need of a partner

We also have the females of other birth orders

- Total of 350 mothers

	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-100
20-24	0.021	0.441	0.366	0.114	0.037	0.013	0.004	0.003
Exact	7.388	154.321	128.194	39.888	13.068	4.666	1.562	0.913
Chosen	7	154	128	40	13	5	2	1

Self correction

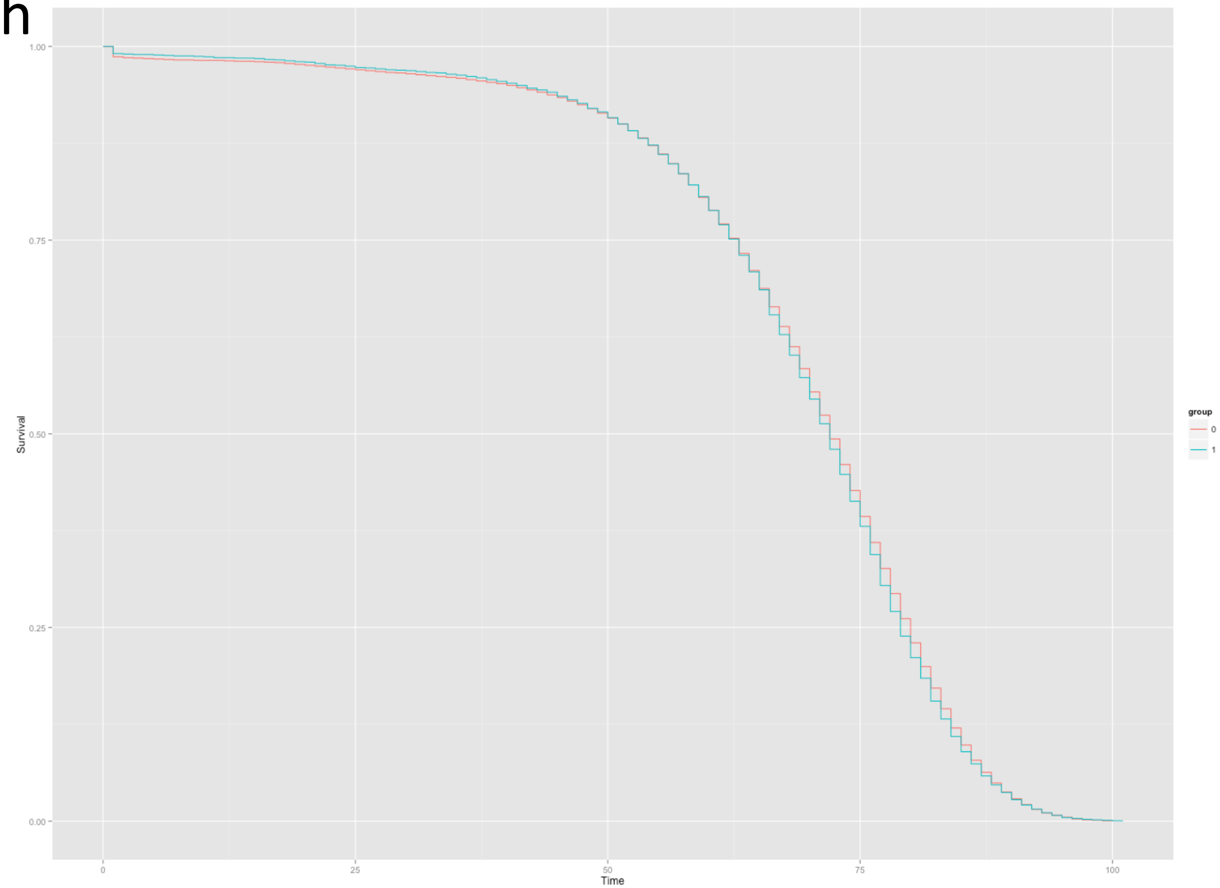
- Rounding errors
- Insufficient people

VPM – Overview

- Inputs
- Integrity and Initialisation
- Simulation approach
    - Simulation
    - Self-correction
- Validation
    - Kaplan Meier
    - ANOVA

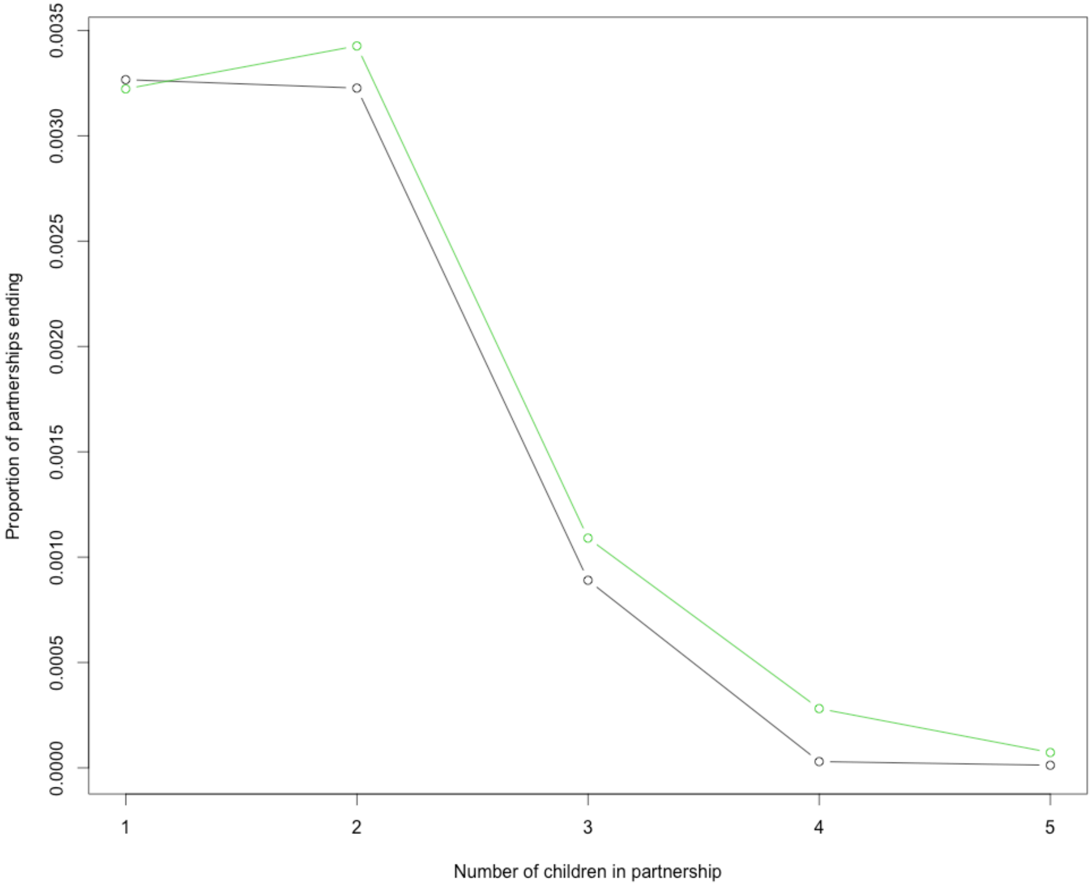
VPM – Statistical Verification

- Kaplan Meier Analysis
  - Ordered birth
  - Death
  - Separation



VPM – Statistical Verification

- ANOVA
  - Partnering
  - Multiple births



VPM – Evaluation

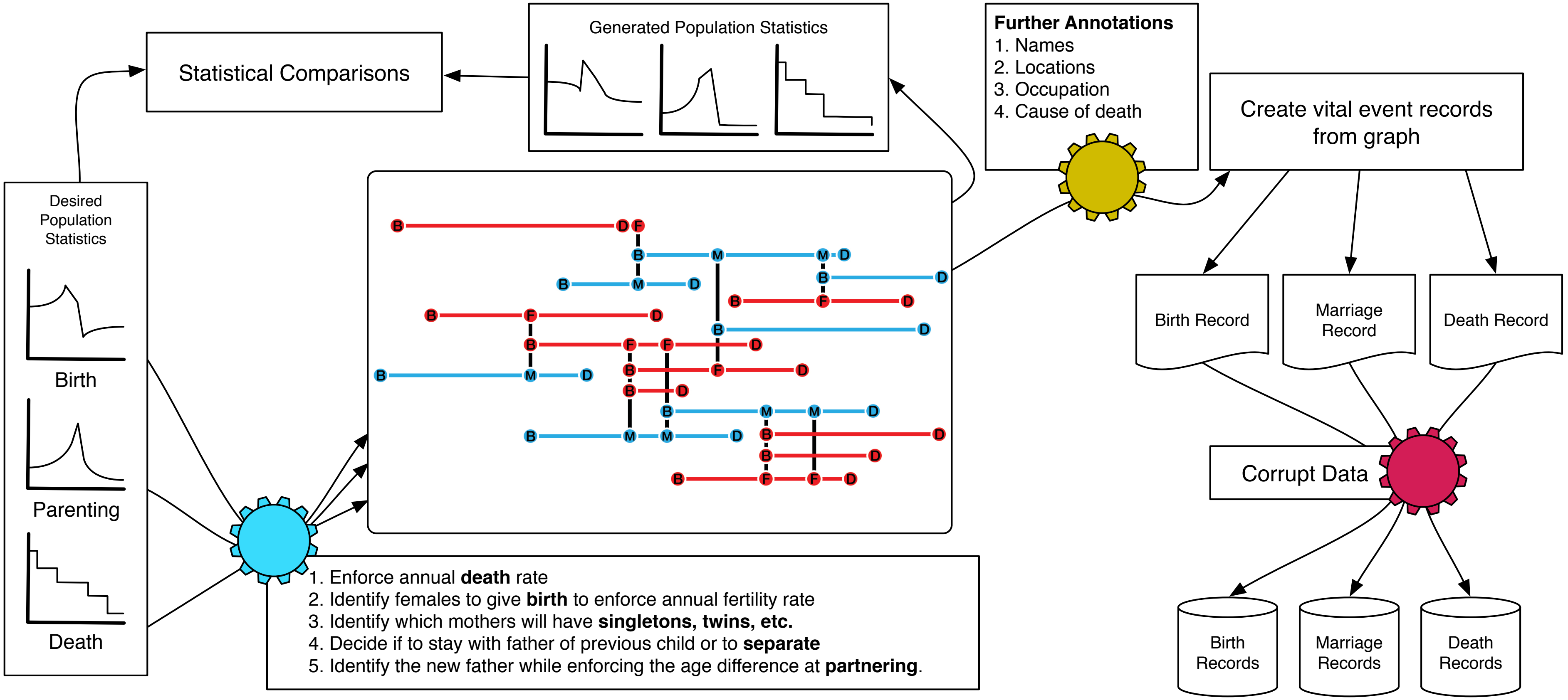
- Infinite number of possible input combination
- How to test?
    - Characteristics
    - Input generation
    - Objective correctness measure

- Generalising to different domains

VPM – Overview

- Inputs
- Integrity and Initialisation
- Simulation approach
    - Simulation
    - Self-correction
- Validation
    - Kaplan Meier
    - ANOVA

VPM – Overview



Statistical Comparisons

Generated Population Statistics

Further Annotations
1. Names
2. Locations
3. Occupation
4. Cause of death

Create vital event records from graph

Desired Population Statistics

Birth

Parenting

Death

1. Enforce annual **death** rate
2. Identify females to give **birth** to enforce annual fertility rate
3. Identify which mothers will have **singletons, twins, etc.**
4. Decide if to stay with father of previous child or to **separate**
5. Identify the new father while enforcing the age difference at **partnering**.

Birth Record

Marriage Record

Death Record

Corrupt Data

Birth Records

Marriage Records

Death Records

Future work and Other Uses

Creating synthetic data sets in privacy sensitive environments

- Data safe havens

Opportunities to explore supervised learning approaches to linkage based on synthetic population topologies

Questions?

Tom Dalton – tsd4@st-andrews.ac.uk