URBAN INSTITUTE · ELEVATE · THE · DEBATE

Synthetic Data Generation with `library(tidysynthesis)`

Aaron R. Williams

Synthetic Data

Synthetic Data

Confidential data

Sex	Age	Wages	Tax
y_{11}	y_{12}	y_{13}	y_{14}
y_{21}	y_{22}	y_{23}	y_{24}

Goals

- Produce synthetic data file with the same record layout as administrative data that:
  - Protects the confidentiality of individual information
  - May be used for statistically valid analysis for certain research purposes
  - May be used as a “training dataset” to develop programs to run on confidential data or a formally private validation server

Estimating the multivariate distribution of the data

- Goal is to approximate the empirical multivariate distribution function for the data
- Joint multivariate probability distribution can be represented as the product of sequential, conditional probability distributions:

$$f\,(\mathrm{Y}_1, \mathrm{Y}_2, \ldots, \mathrm{Y}_k \mid \theta_1, \theta_2, \ldots, \theta_k) \;=$$

$$f_1(Y_1 \mid \theta_1) \;\cdot f_2(Y_2 \mid Y_1, \theta_2) \cdots\; f_{\mathrm{k}}(Y_k \mid Y_1, Y_2, \ldots, Y_{k-1}, \theta_k)$$

  - where Y_i the variables and θ_i are vectors of model parameters

Synthetic Data

Confidential data

Sex	Age	Wages	Tax
y_{11}	y_{12}	y_{13}	y_{14}
y_{21}	y_{22}	y_{23}	y_{24}

Partially synthetic data (R. Little - 1993)

Sex	Age	Wages	Tax
y_{11}	y_{12}	$\hat{y}_{13}$	$\hat{y}_{14}$
y_{21}	y_{22}	$\hat{y}_{23}$	$\hat{y}_{24}$

Synthetic Data

Confidential data

Sex	Age	Wages	Tax
y_{11}	y_{12}	y_{13}	y_{14}
y_{21}	y_{22}	y_{23}	y_{24}

Partially synthetic data (R. Little - 1993)

Sex	Age	Wages	Tax
y_{11}	y_{12}	$\hat{y}_{13}$	$\hat{y}_{14}$
y_{21}	y_{22}	$\hat{y}_{23}$	$\hat{y}_{24}$

Fully synthetic data (D. Rubin – 1993)

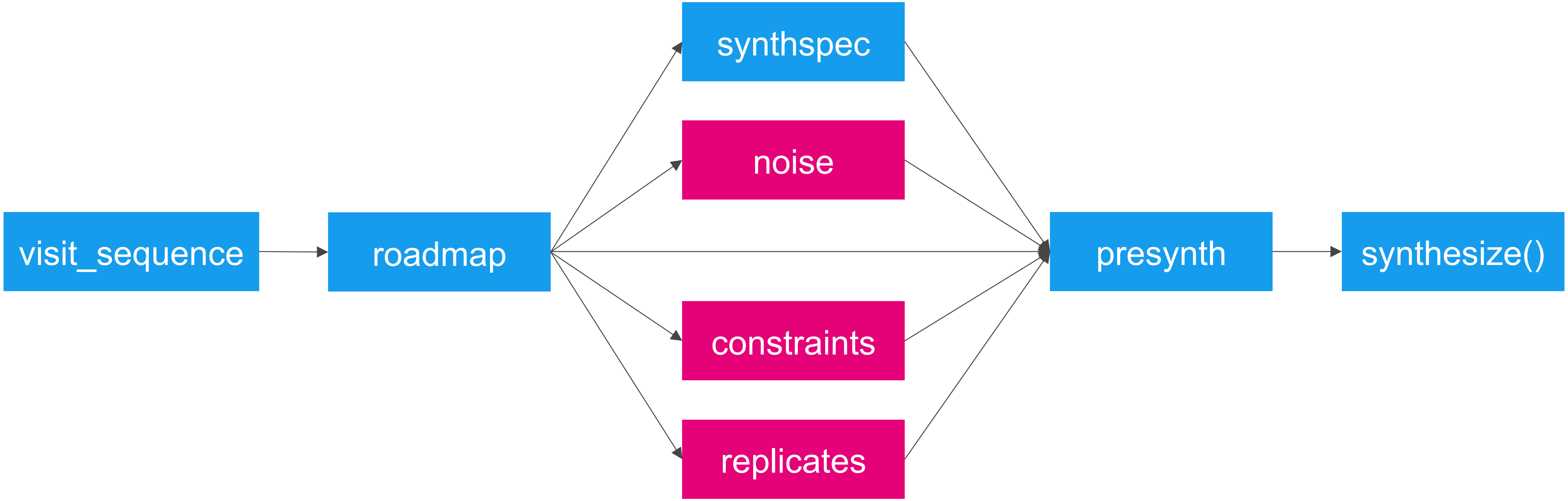
Sex	Age	Wages	Tax
$\hat{y}_{11}$	$\hat{y}_{12}$	$\hat{y}_{13}$	$\hat{y}_{14}$
$\hat{y}_{11}$	$\hat{y}_{12}$	$\hat{y}_{23}$	$\hat{y}_{24}$

```
library(tidysynthesis)
```

New features

- Feature and target engineering with `library(recipes)`
- Model metrics
- Mid-synthesis constraints
- Weighted data

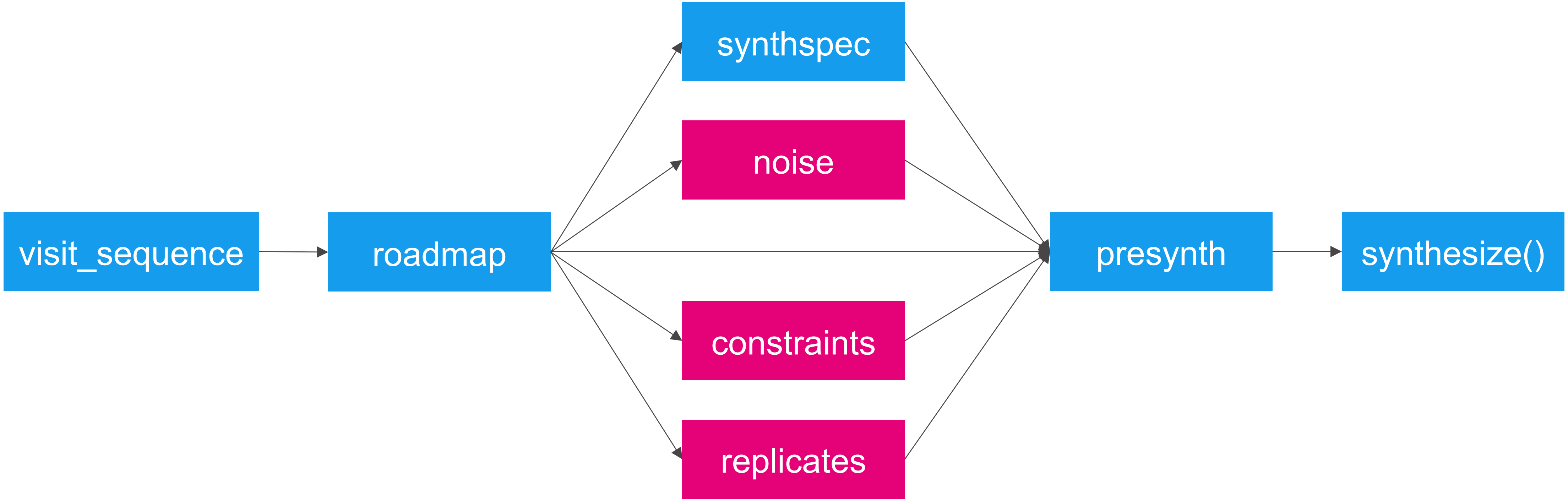
Workflow


visit_sequence
roadmap
synthspec
noise
constraints
replicates
presynth
synthesize()

`visit_sequence()` and `roadmap()`

- Apply rules for determining a synthesis order
- Add starting data
- Add confidential data

Workflow


visit_sequence
roadmap
synthspec
noise
constraints
replicates
presynth
synthesize()

synthspec()

- Add model specifications
  - Feature and target engineering
  - Algorithms
  - Sampling methods
- Methods can vary from variable-to-variable

`noise()`

- Add noise to predictions beyond prediction error

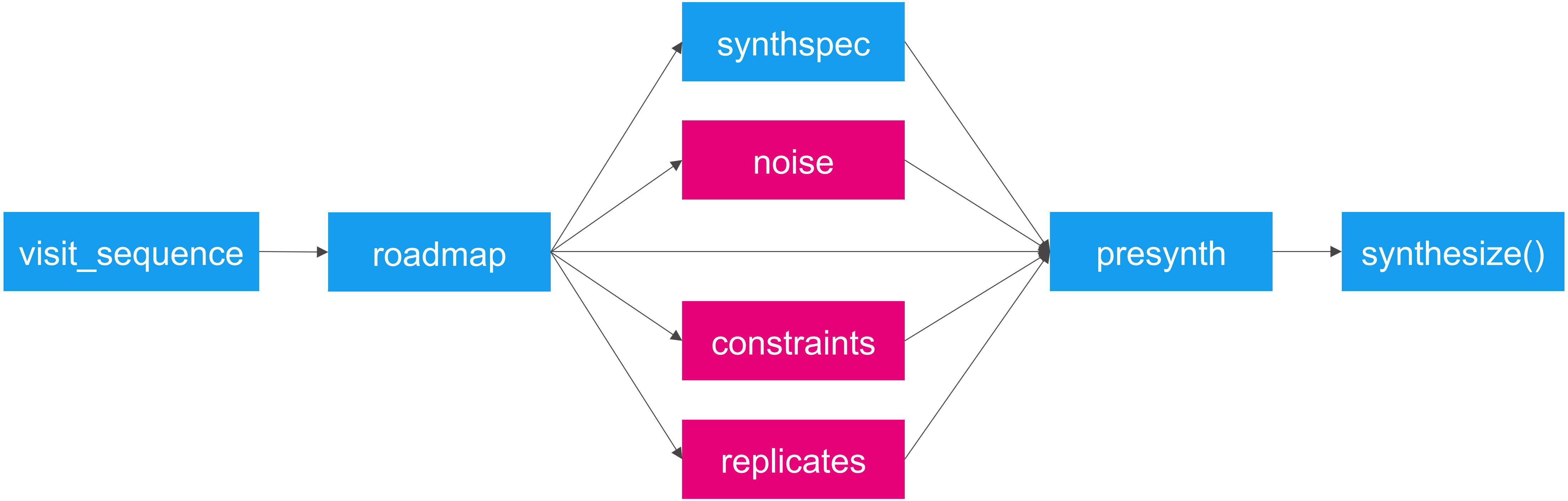
`constraints()`

- Implement univariate and multivariate constraints
- Remedial measures:
  - Exclusions
  - Hard bounding
  - Z-bounding

`replicates()`

- Create multiple synthetic data sets
- Run the synthesis process in parallel

Workflow


visit_sequence
roadmap
synthspec
noise
constraints
replicates
presynth
synthesize()