URBAN INSTITUTE · ELEVATE · THE · DEBATE

Tuesday, August 6th, 2024

Flexible Synthetic Data Generation with library(tidysynthesis)

Aaron R. Williams

Summary

1. We created a better, more flexible synthesizer for generating FCS synthetic data.

Summary

1. We created a better, more flexible synthesizer for generating FCS synthetic data.
2. Synthetic data are only as good as the synthesizer.

Summary

1. We created a better, more flexible synthesizer for generating FCS synthetic data.
2. Synthetic data are only as good as the synthesizer.
3. Existing tools are limited and difficult to extend.

Summary

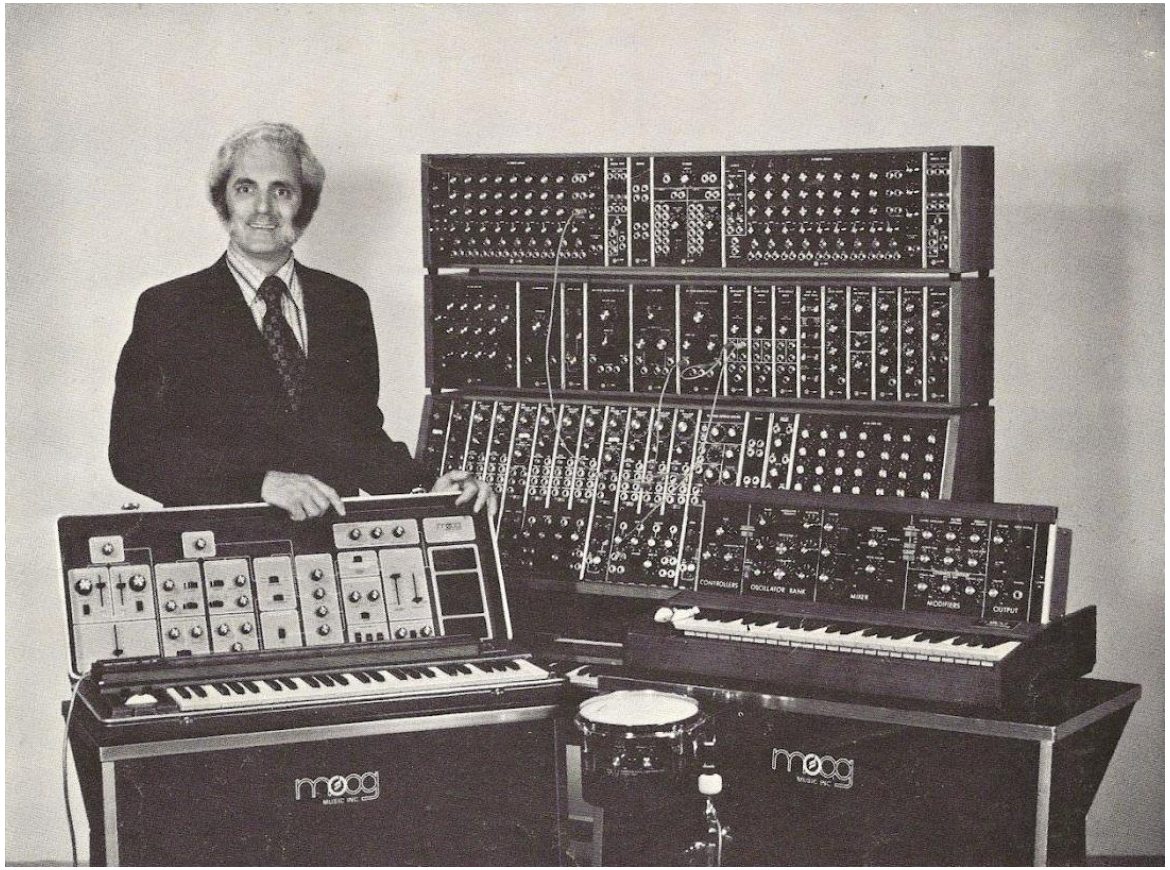
1. We created a better, more flexible synthesizer for generating FCS synthetic data.
2. Synthetic data are only as good as the synthesizer.
3. Existing tools are limited and difficult to extend.
4. library(tidysynthesis) is modular, extensible, and builds on library(tidymodels).

Summary

1. We created a better, more flexible synthesizer for generating FCS synthetic data.
2. Synthetic data are only as good as the synthesizer.
3. Existing tools are limited and difficult to extend.
4. library(tidysynthesis) is modular, extensible, and builds on library(tidymodels).
5. This works well for generating select synthetic data.

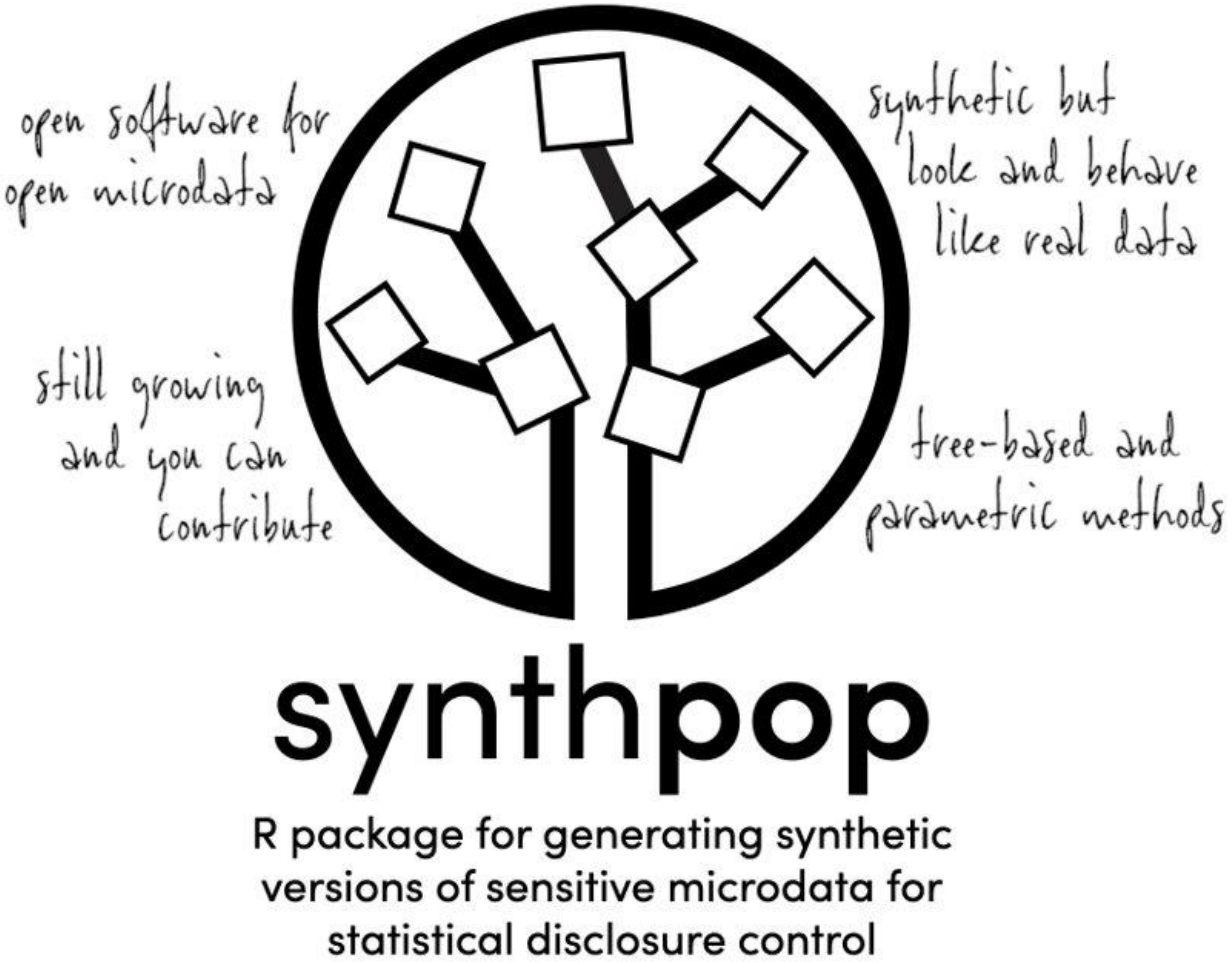
Build a Synthesizer

- Protects the confidentiality of individual information.
- Uses a fully conditional specification to model the joint multivariate distribution.



Existing Functionality is Groundbreaking, but Limited

- Limited to a small set of methods
- Difficult to extend



R Packages

- **tidysynthesis:** Flexible tools for generating fully and partially synthetic data.
- **syntheval:** Utility and disclosure risk evaluation of synthetic data.

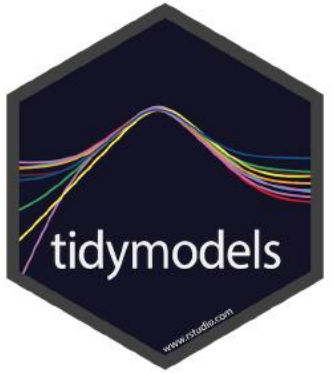
Our Approach

1. Embrace design philosophy from tidyverse and tidymodels 

2. Flexible
3. Modular
4. Extensible

library(tidymodels)

- *All the power of library(tidymodels) for data synthesis, concisely, with a few special tools.*



- Full predictive modeling toolkit

Flexibility

- Express different recipes, predictive models, and samplers for each variable
- Hyperparameter tuning
- Additional noise
- Mid-synthesis constraints
- Synthesize missing data
- Parallel computation with futures/furrr

Modular

- Everything is handled through objects with classes
  - Interchangeable objects that act like building blocks
- Robust testing suite
- Manage computation
  - Lazy evaluation and checks that catch errors before computation

Extensibility

- Ability for someone else to add the thing we haven't thought of

Demonstration

Synthetic data

Confidential data

select	species	island	sex	bill_length_mm	…
TRUE	Adelie	Torgersen	male	39.1	…
FALSE	Adelie	Torgersen	female	39.5	…

Synthetic data

select	species	island	sex	bill_length_mm	…
TRUE					
TRUE					


Synthetic data!

Synthetic data

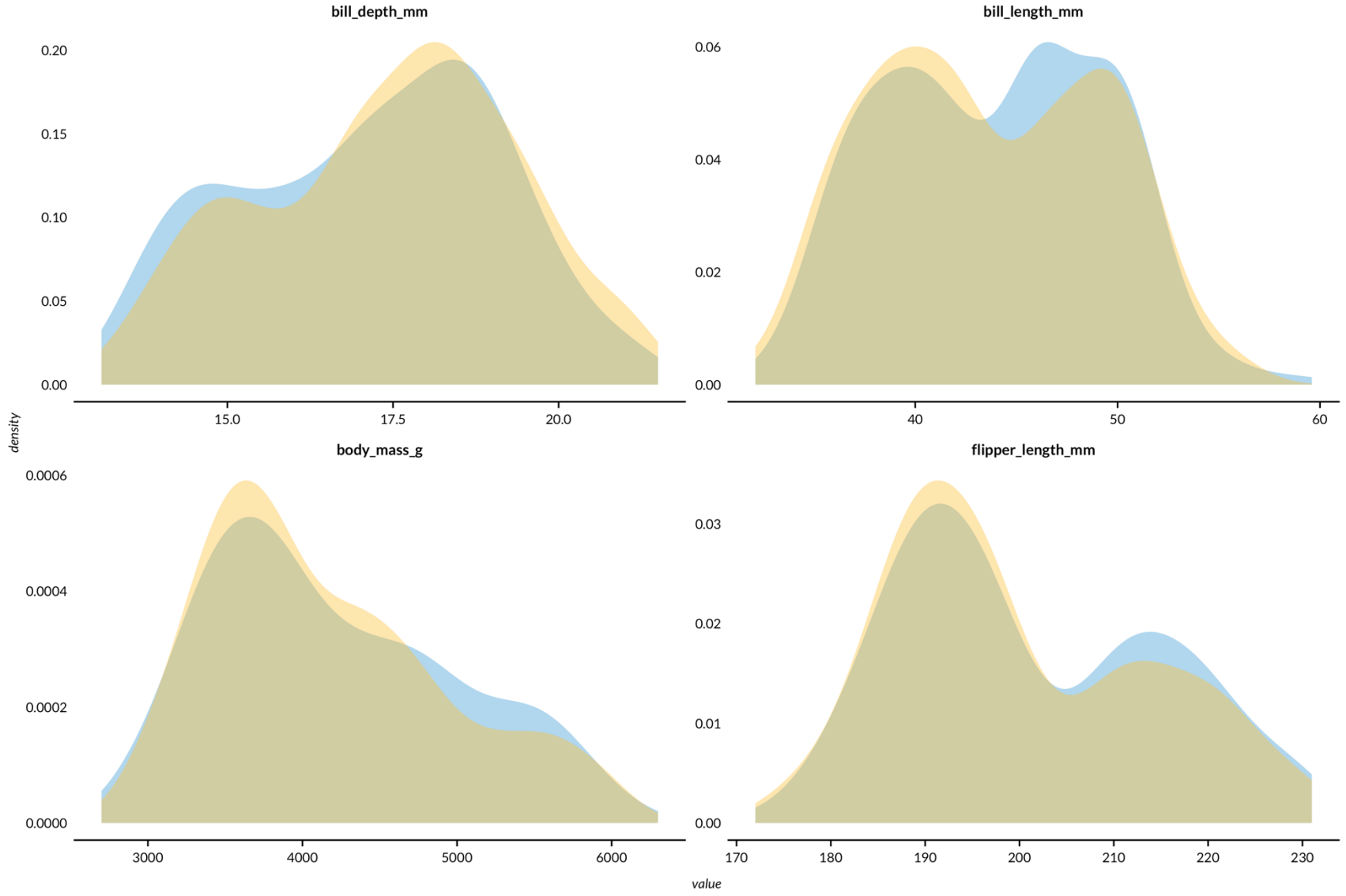
Confidential data

select	species	island	sex	bill_length_mm	…
TRUE	Adelie	Torgersen	male	39.1	…
FALSE	Adelie	Torgersen	female	39.5	…

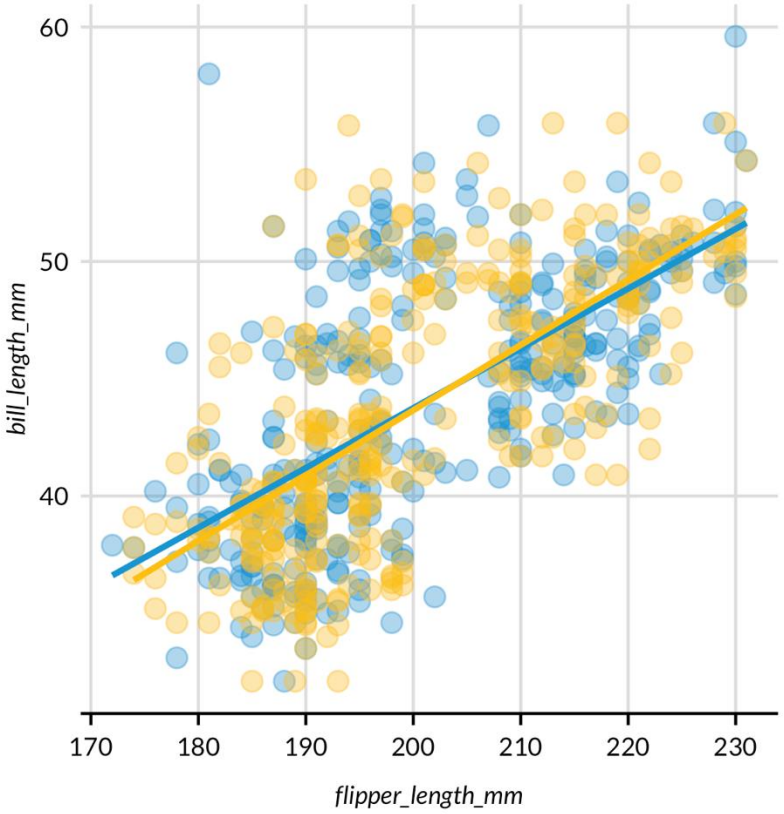
Synthetic data

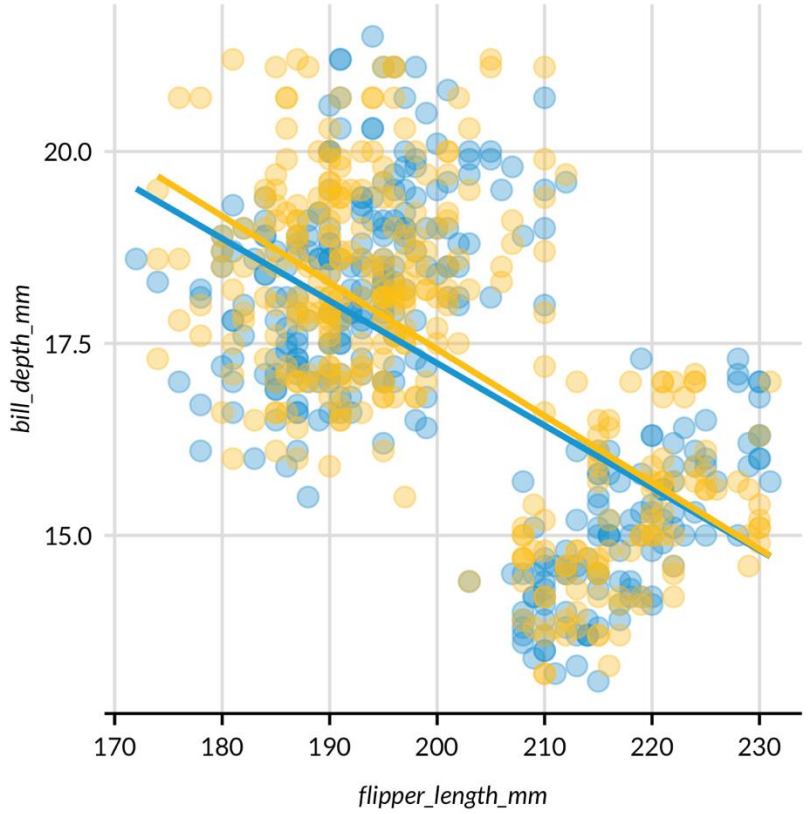
select	species	island	sex	bill_length_mm	…
TRUE	Chinstrap	Dream	male	48.4	…
TRUE	Gentoo	Biscoe	male	51.4	…

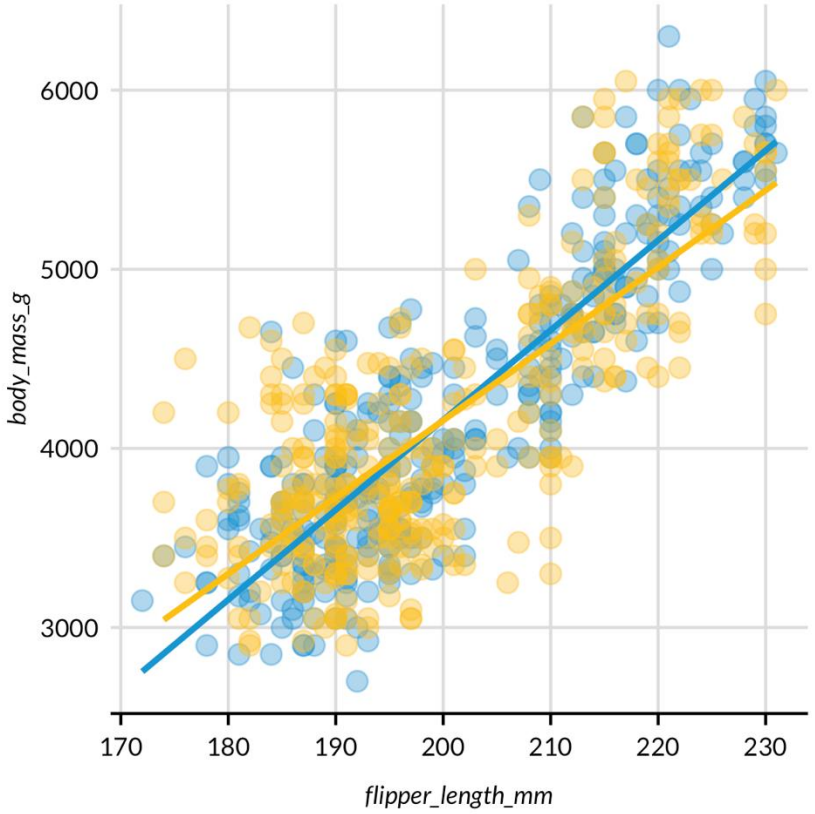
The Synthetic Data are Similar to the Confidential Data



The Synthetic Data are Similar to the Confidential Data



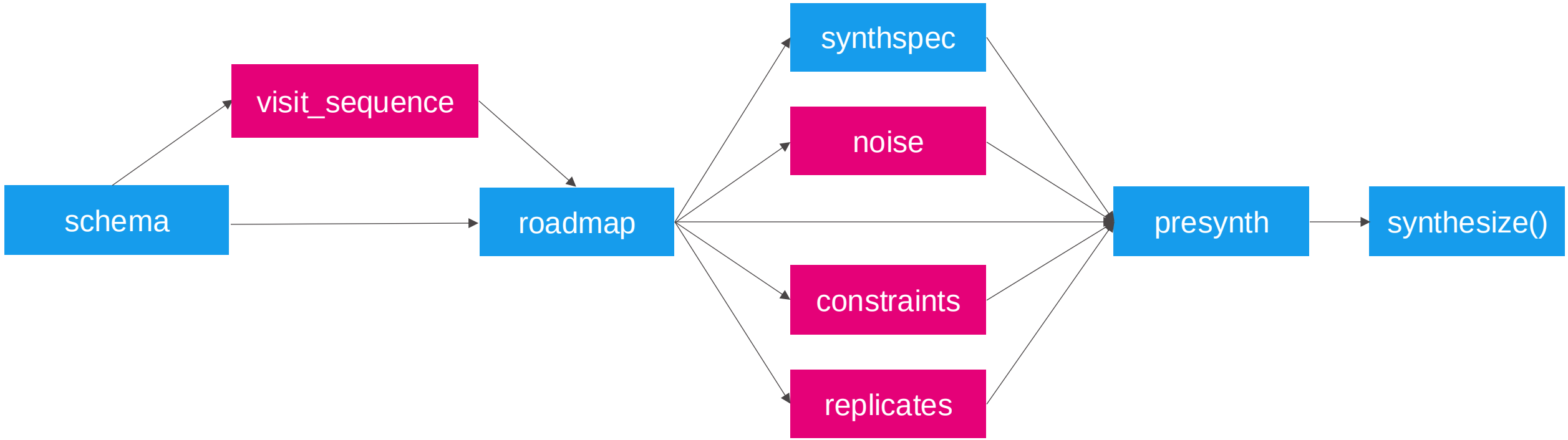




Summary

1. We created a better, more flexible synthesizer for generating FCS synthetic data.
2. Synthetic data are only as good as the synthesizer.
3. Existing tools are limited and difficult to extend.
4. library(tidysynthesis) is modular, extensible, and builds on library(tidymodels).
5. This works well for generating select synthetic data.

Workflow


synthspec
visit_sequence
noise
schema
roadmap
presynth
synthesize()
constraints
replicates

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:

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

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

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