

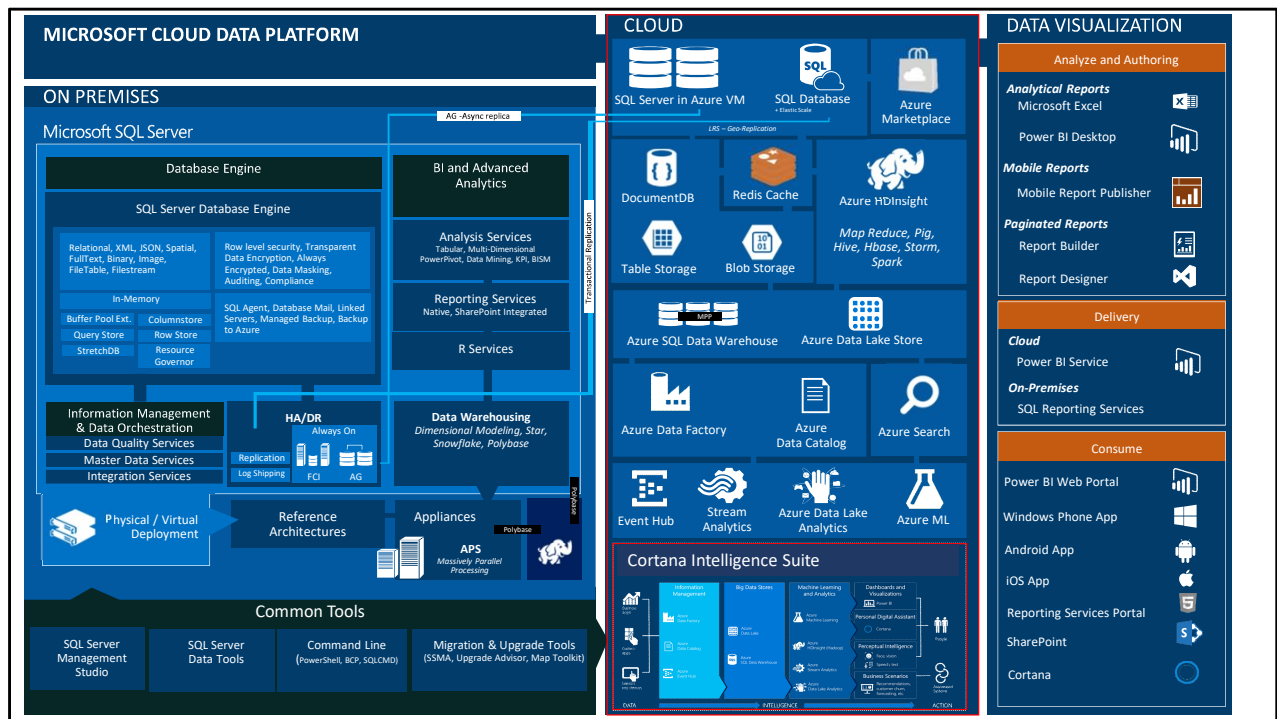


Microsoft

Data Synthesis in R

St. Louis R User Group

Kay Apperson, PhD
Azure Advanced Analytics & AI
kay.apperson@microsoft.com

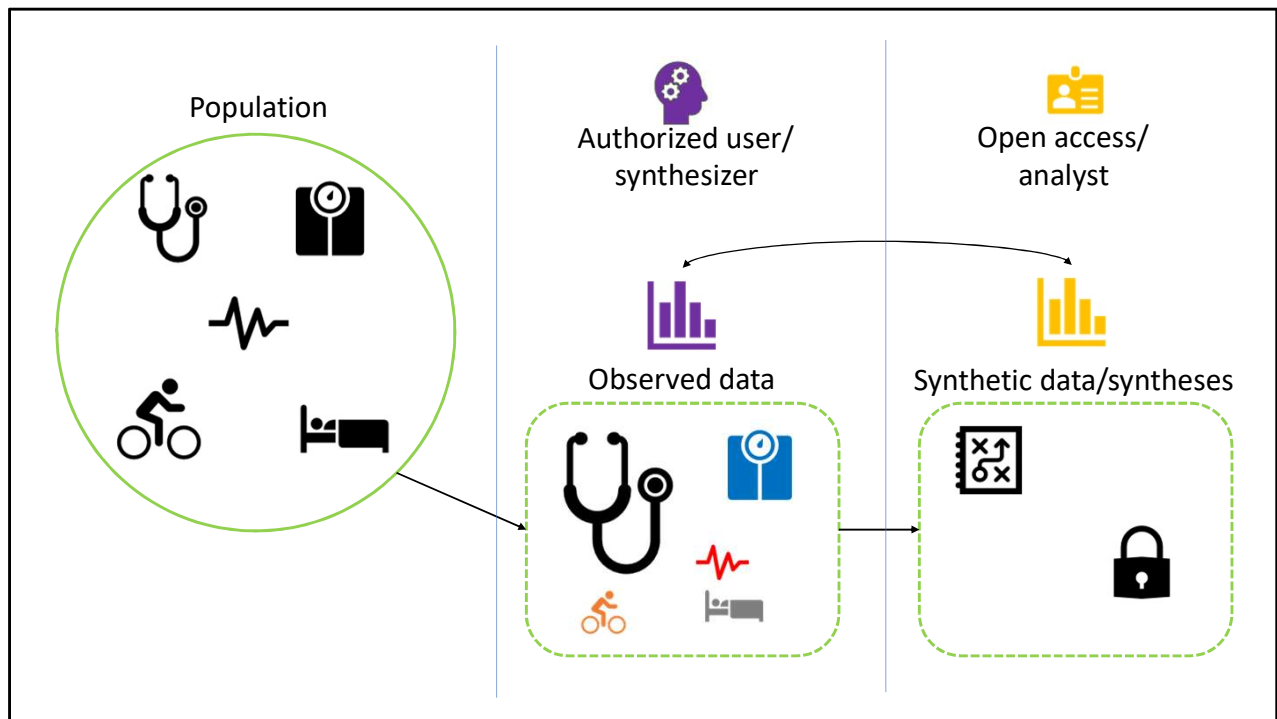


These are some of the technologies in Kay's life.

synthpop


[https://cran.r-project.org/
web/packages/synthpop/
vignettes/synthpop.pdf](https://cran.r-project.org/web/packages/synthpop/vignettes/synthpop.pdf)





The basic idea of synthetic data is to replace some or all of the observed values by sampling from appropriate probability distributions so that the essential statistical features of the original data are preserved.

Synthpop's aim is only to make inferences to the results that would have been obtained by the gold standard analysis, with the expectation that the analyst will run final models on the observed data.



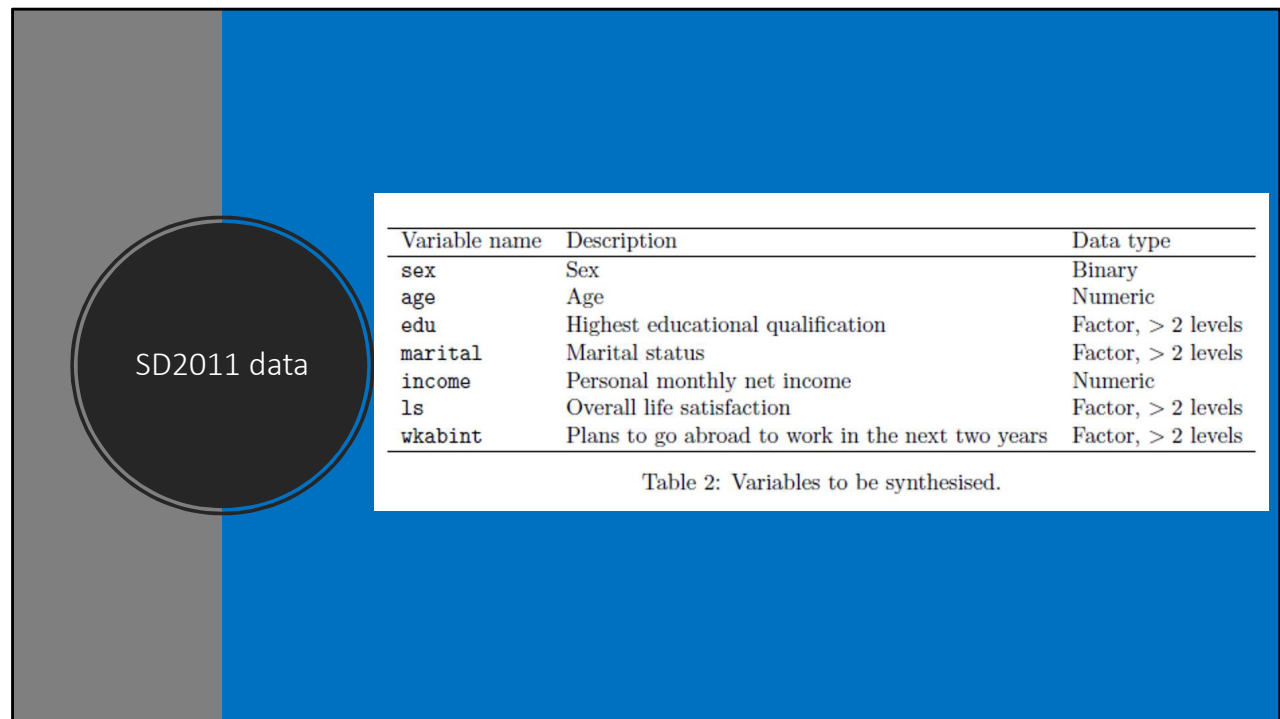
synthpop
syn()

```
library(synthpop)
# Synthesized data set, sds.
my.seed <- 17914709
sds.default <- syn(ods, seed = my.seed)
sds.default

# sds.default$syn is the dataframe we're looking for.
head(sds.default$syn)
str(sds.default$syn)

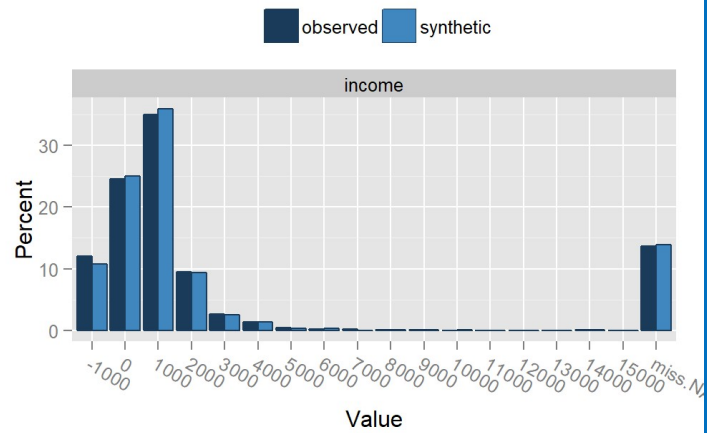
names(sds.default)

sdsdefault <- sds.default$syn
```

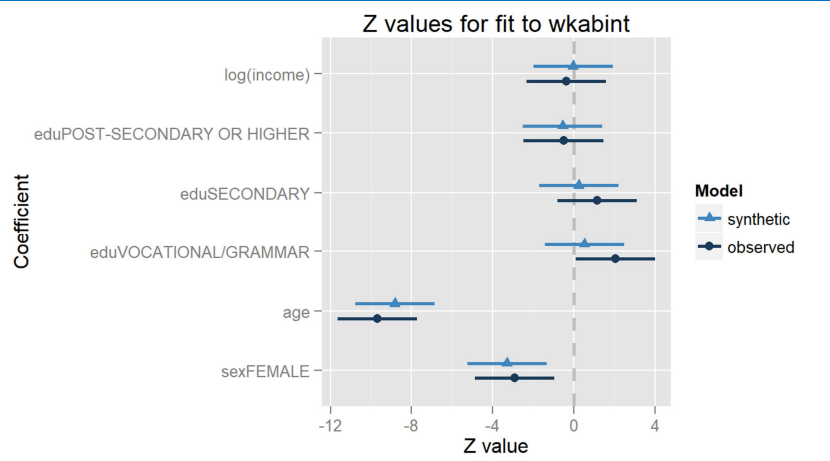


Data from Poland about life satisfactory (ls).

compare()



compare()



Different
methods
accept
different
data types

Method	Description	Data type
<i>Non-parametric</i>		
<code>ctree</code> , <code>cart</code>	Classification and regression trees	Any
<code>surv.ctree</code>	Classification and regression trees	Duration
<i>Parametric</i>		
<code>norm</code>	Normal linear regression	Numeric
<code>normrank</code> *	Normal linear regression preserving the marginal distribution	Numeric
<code>logreg</code> *	Logistic regression	Binary
<code>polyreg</code> *	Polytomous logistic regression	Factor, > 2 levels
<code>polr</code> *	Ordered polytomous logistic regression	Ordered factor, > 2 levels
<code>pmm</code>	Predictive mean matching	Numeric
<i>Other</i>		
<code>sample</code>	Random sample from the observed data	Any
<code>passive</code>	Function of other synthesised data	Any

Table 1: Built-in synthesising methods. * Indicates default parametric methods.

Tools

1

Visual Studio 2015
Community
Update 3.

[Here.](#)

2

R Tools for Visual
Studio (RTVS).

[Here.](#)

3

R Server
in SQL Server 2016
Developer Edition.

[Here.](#)

R Studio IDE is an option for the authoring tool.
R open source is an option for the R engine.

1

Note that synthpop is under continual development.

2

Future versions will include

- appropriate procedures for synthesizing multiple event data
- conducting stratified synthesis
- replacing only selected cases from selected variables

3

The ultimate aim of synthpop

- to provide a comprehensive, flexible and easy to use tool for generating bespoke synthetic data that can be safely released to interested data users.

Concluding Remarks

From both original and synthetic data, we conclude that men are more likely to declare intention to work abroad as are those who are young. The fact that the results from synthetic data can have a similar pattern to the results from the real data is encouraging for further developments of synthetic data tools.



© 2016 Microsoft Corporation. All rights reserved.

