Synthetic Data Generation

Enabling external researchers to use the Netherlands Cancer Registry, safely

Daan Knoors
d.knoors@iknl.nl

Enabling external researchers to use the NCR, safely

Show what data we have available

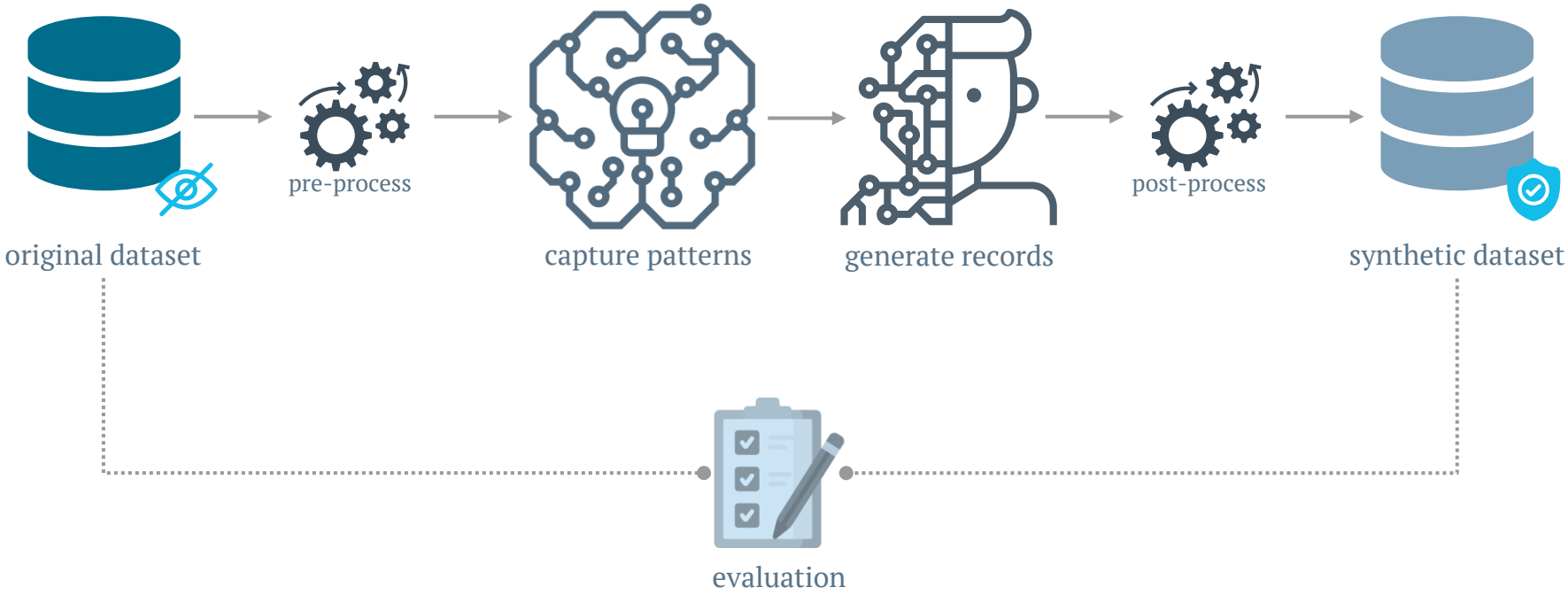
Enable anyone to work with cancer data

Anything developed on synthetic data works on real data as well

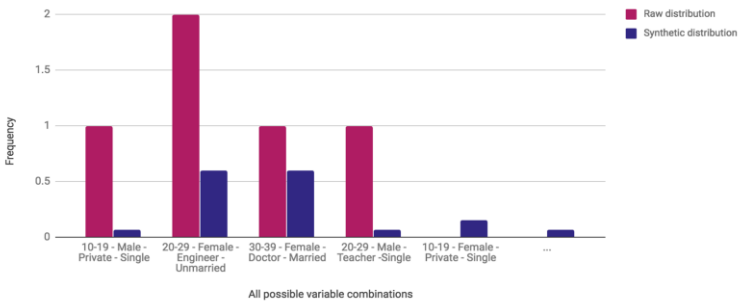
Findable Accessible Interoperable Reusable

Generating Synthetic Data

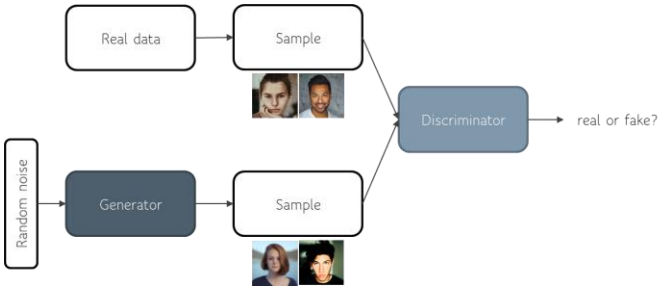
Synthetic Data Generation



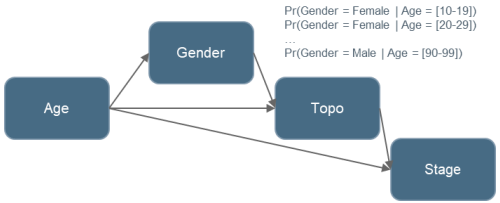
Methods for generating synthetic data



Probability Tables



Generative Adversarial Networks



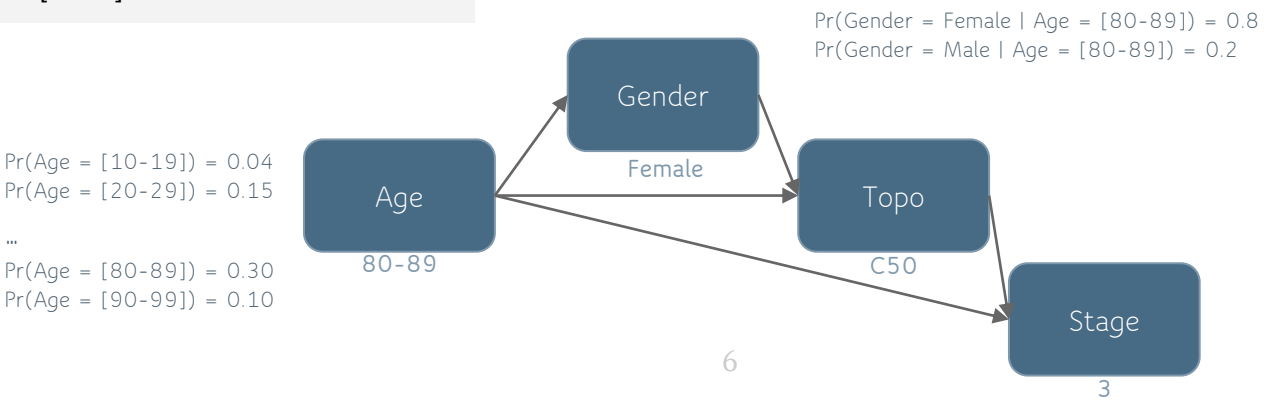
Bayesian Networks

PrivBayes (Zhang et al, 2017)

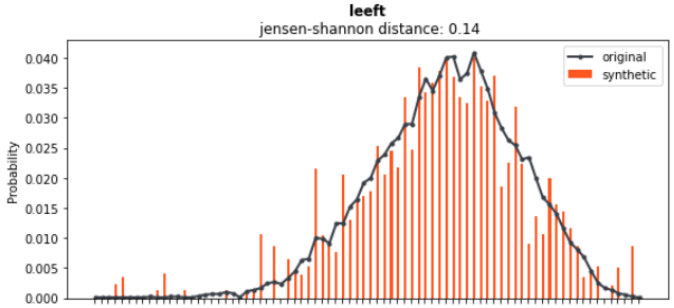
Gender	Age	Topo	Stage
Female	[70-79]	C55	2
Female	[70-79]	C50	3
Female	[80-89]	C50	1
Male	[60-69]	C50	1
Male	[60-69]	C61	3

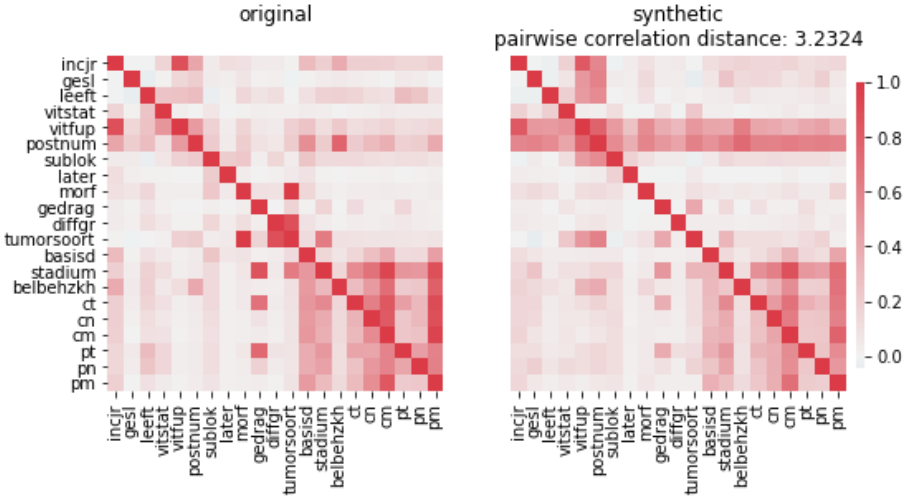
Learn a Bayesian Network from the data, where nodes are linked with high mutual information.

Generate records by sampling values from the conditional distributions in the network.

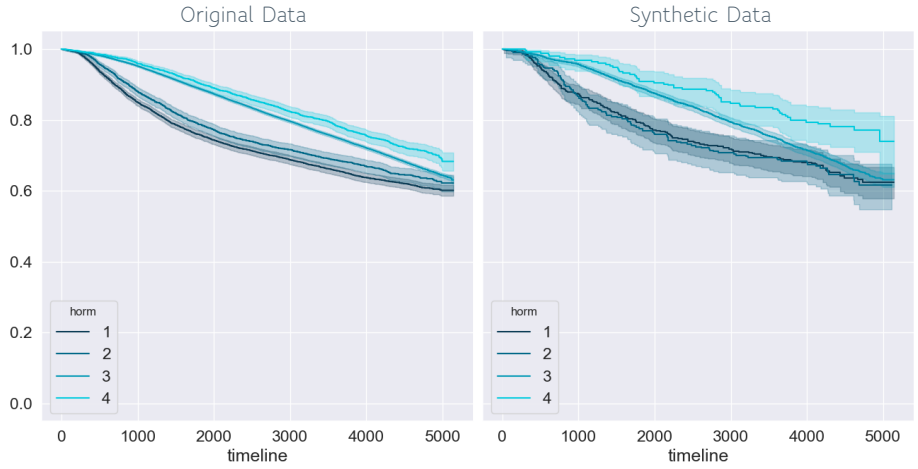


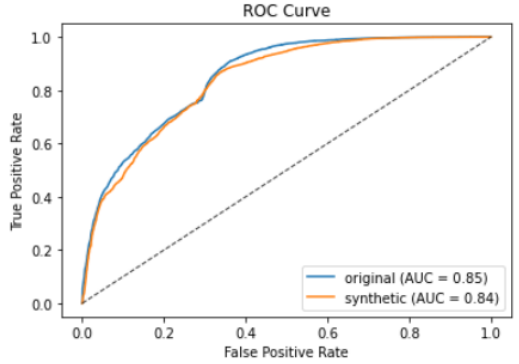
Evaluating Statistical Fidelity





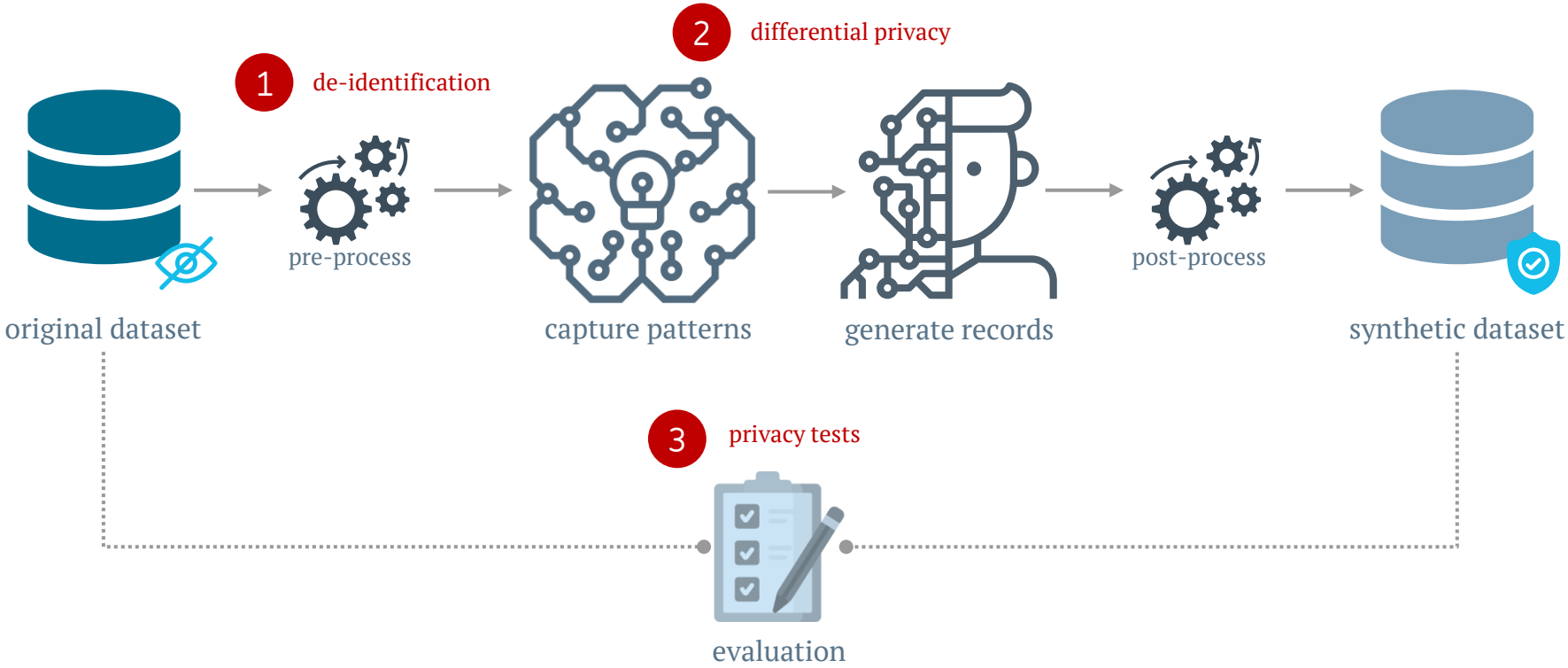
Survival Analysis





Private Synthetic Data Generation

3 steps to protect privacy



Netherlands comprehensive cancer organisation

synthetic data

Show what data we have available

Enable anyone to work with cancer data

Anything developed on synthetic data works on real data as well

Thanks

Daan Knoors
d.knoors@iknl.nl