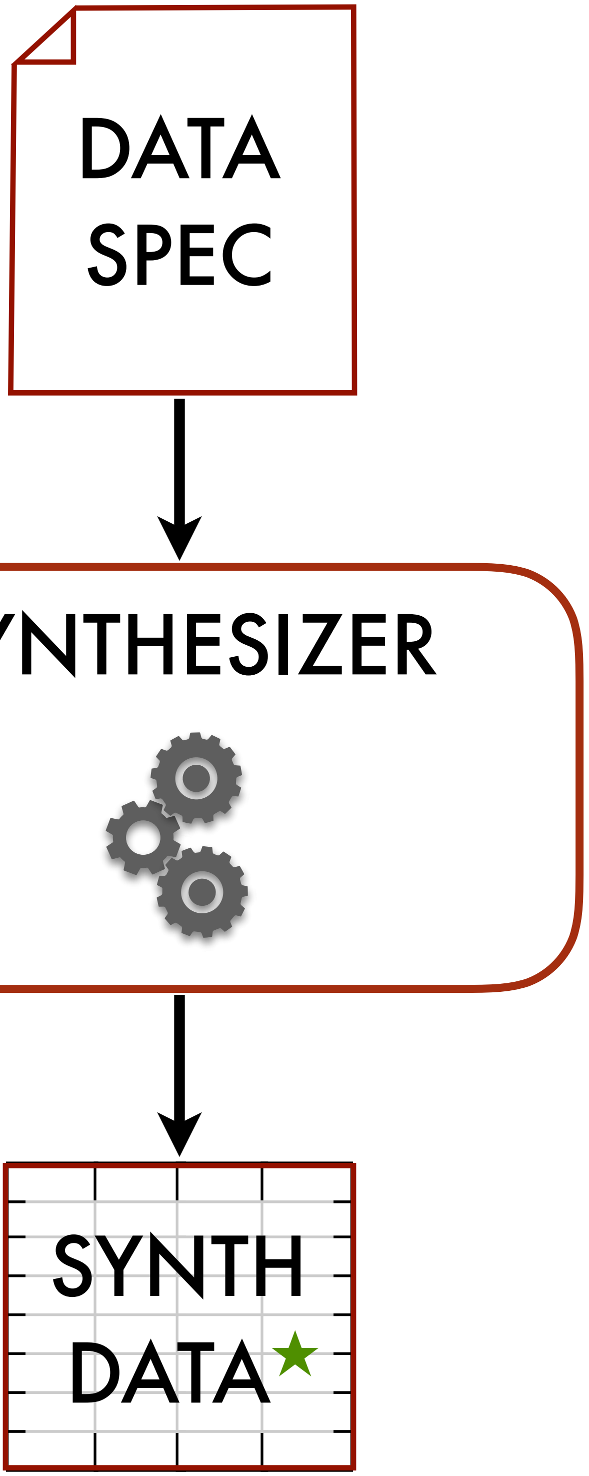
# CONSTRAINED DATA SYNTHESIS

EuroSciPy 2019 • Bilbao • 4th September 2019

**http://www.tdda.info/pdf/constrained-data-synthesis-euroscipy-2019.pdf**

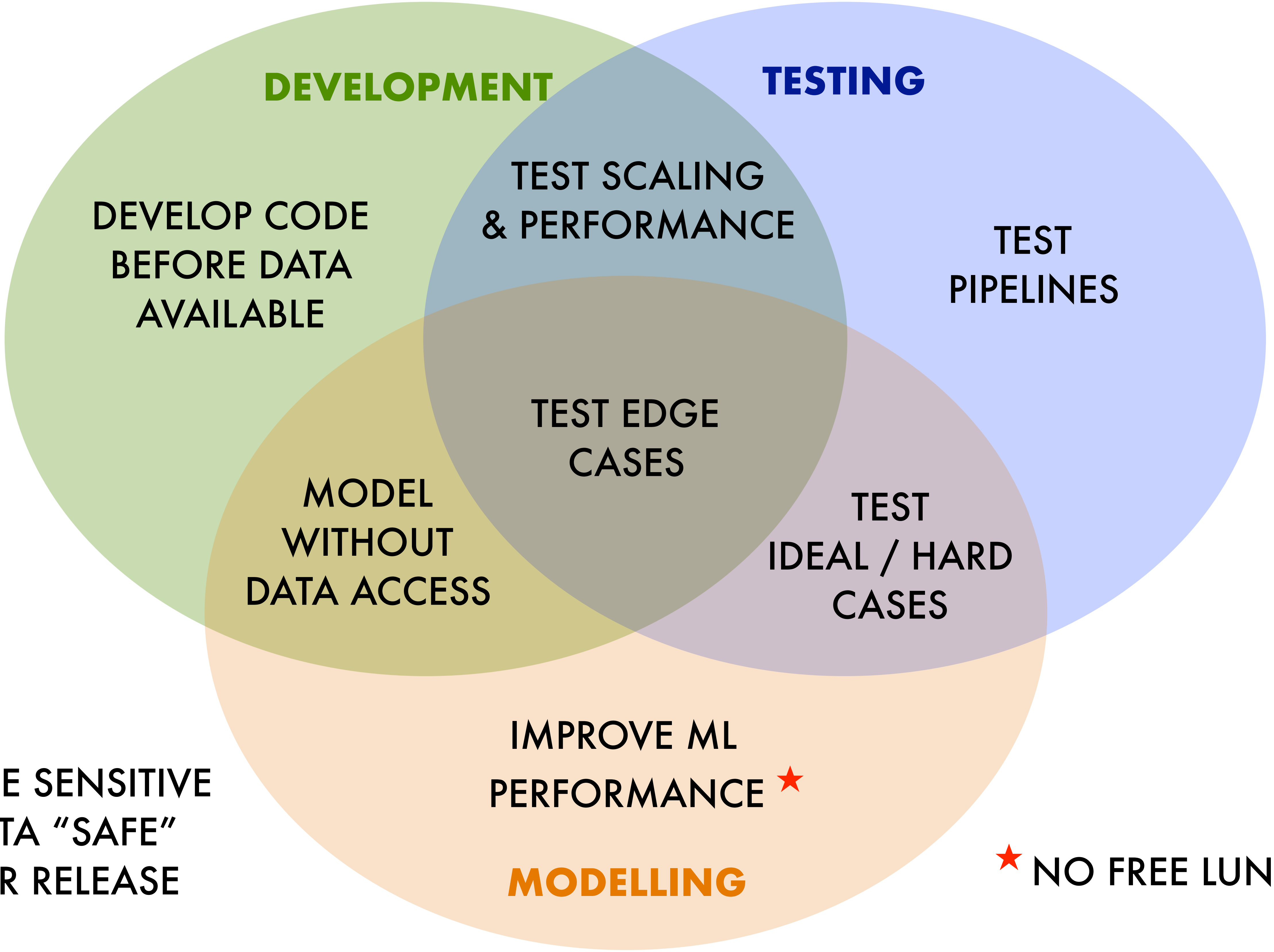
Nicholas J. Radcliffe
Stochastic Solutions Limited
& Department of Mathematics, University of Edinburgh
PyData Edinburgh

# GOAL



★*"any" quantity; conforming*

# WHY SYNTHESIZE DATA?



**DEVELOPMENT**

**TESTING**

**MODELLING**

DEVELOP CODE
BEFORE DATA
AVAILABLE

TEST SCALING
& PERFORMANCE

TEST
PIPELINES

TEST EDGE
CASES

MODEL
WITHOUT
DATA ACCESS

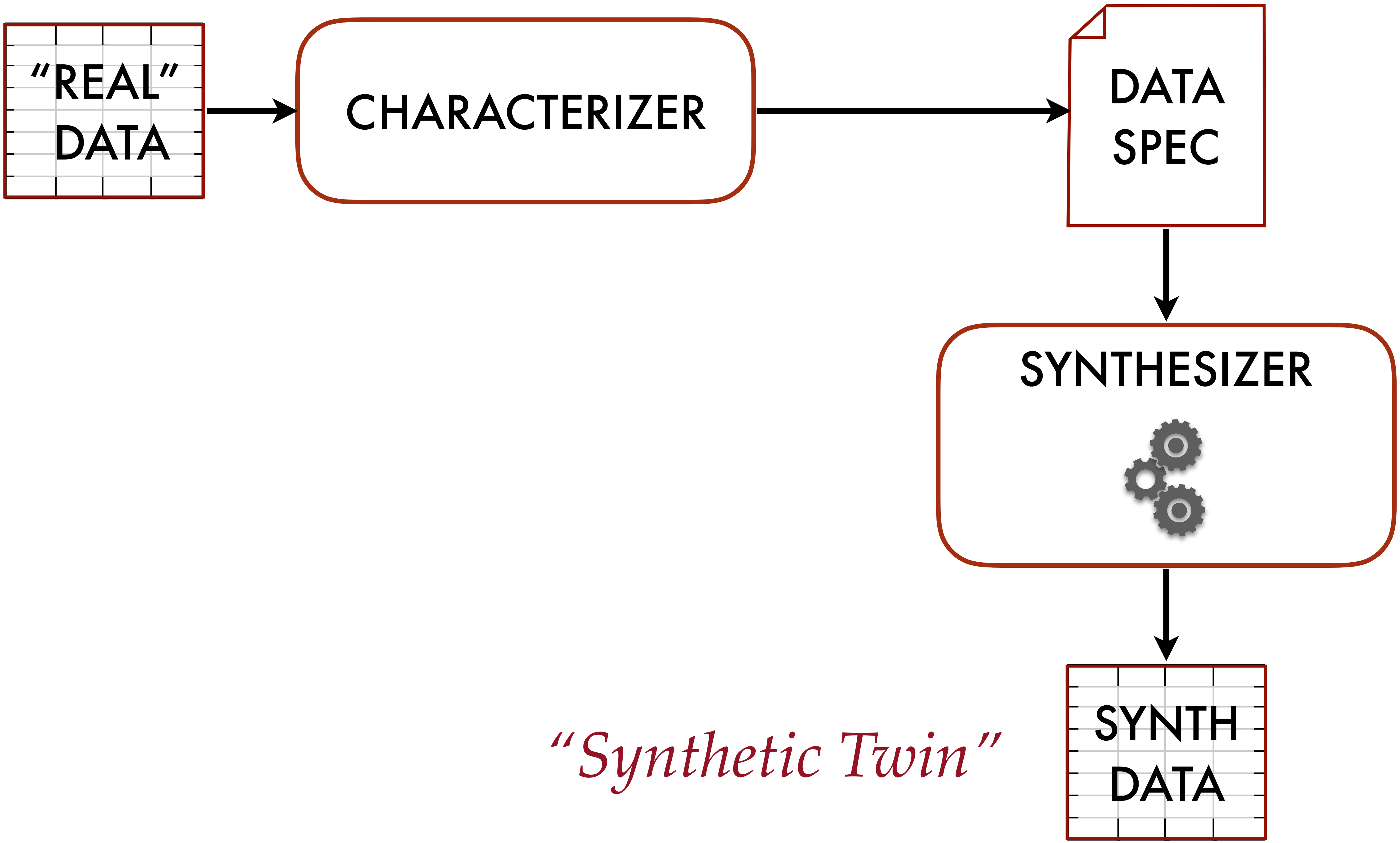
TEST
IDEAL / HARD
CASES

IMPROVE ML
PERFORMANCE ★

MAKE SENSITIVE
DATA "SAFE"
FOR RELEASE

★ NO FREE LUNCH

# SPECIAL CASE



“REAL” DATA

CHARACTERIZER

DATA SPEC

SYNTHESIZER

SYNTH DATA

*“Synthetic Twin”*

*but*

*but*

*But what*
***exactly***
*do we want to be*
*"the same"*★
*about our synthetic twin?*

★ NO FREE LUNCH

# WHAT PROPERTIES DO WE WANT FROM SYNTHETIC (TWIN) DATA?

*“Same” shape of data (names, types, values, structure, ranges)*

*Similar univariate distributions*

*Similar patterns of missing values, outliers, duplicates etc.*

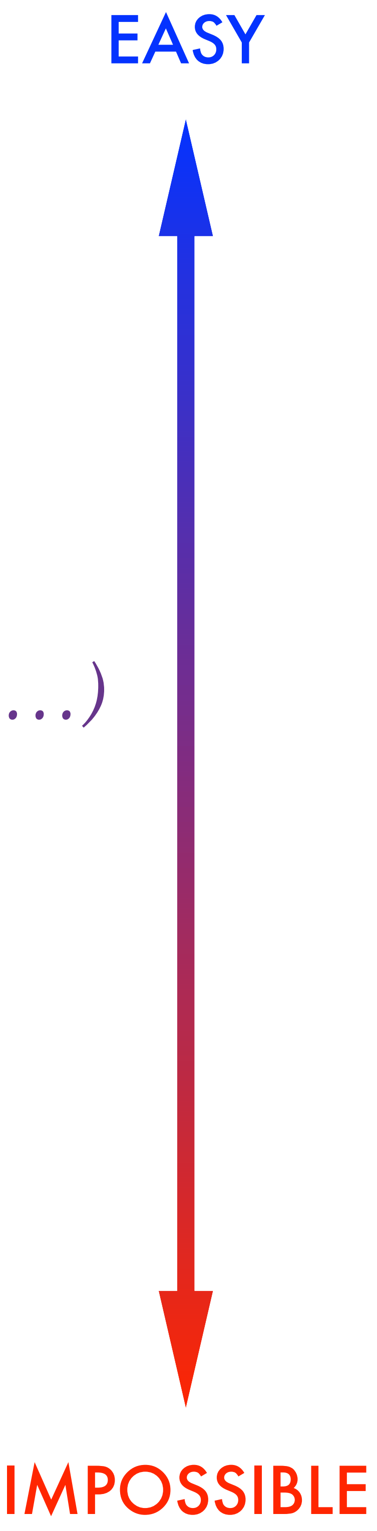
*Similar string patterns (categoricals, structured text, unstructured…)*

*Remove sensitive / PII data*

*Similar relational structure between tables*

*Provably / definitely remove sensitive / PII data*

*“Same” (similar) multi-variate distributions (all orders)* ★



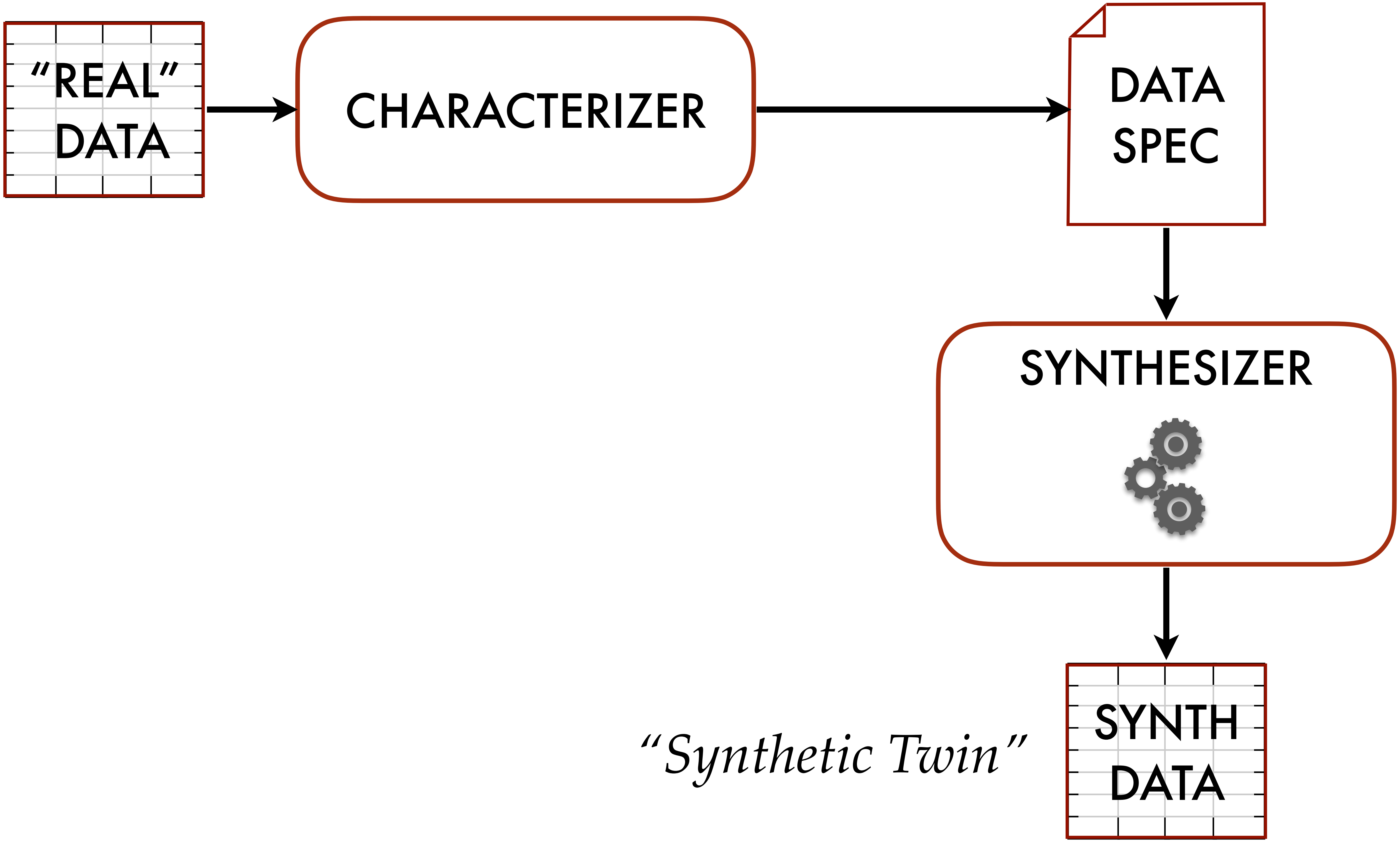
# SOME APPROACHES

- *Heuristic ("just write some code")*
- *Anonymize/resample/disguise/remix*
- *Generate*
  - *e.g. Generative Adversarial Networks*
    - *e.g. Synthpop, VAEGAN*
  - *Constraint-based Generation like here, currently in Miró*

*Mostly multivariate correlation*

*Mostly structure and per-field properties (currently)*

# SPECIAL CASE



"REAL" DATA

CHARACTERIZER

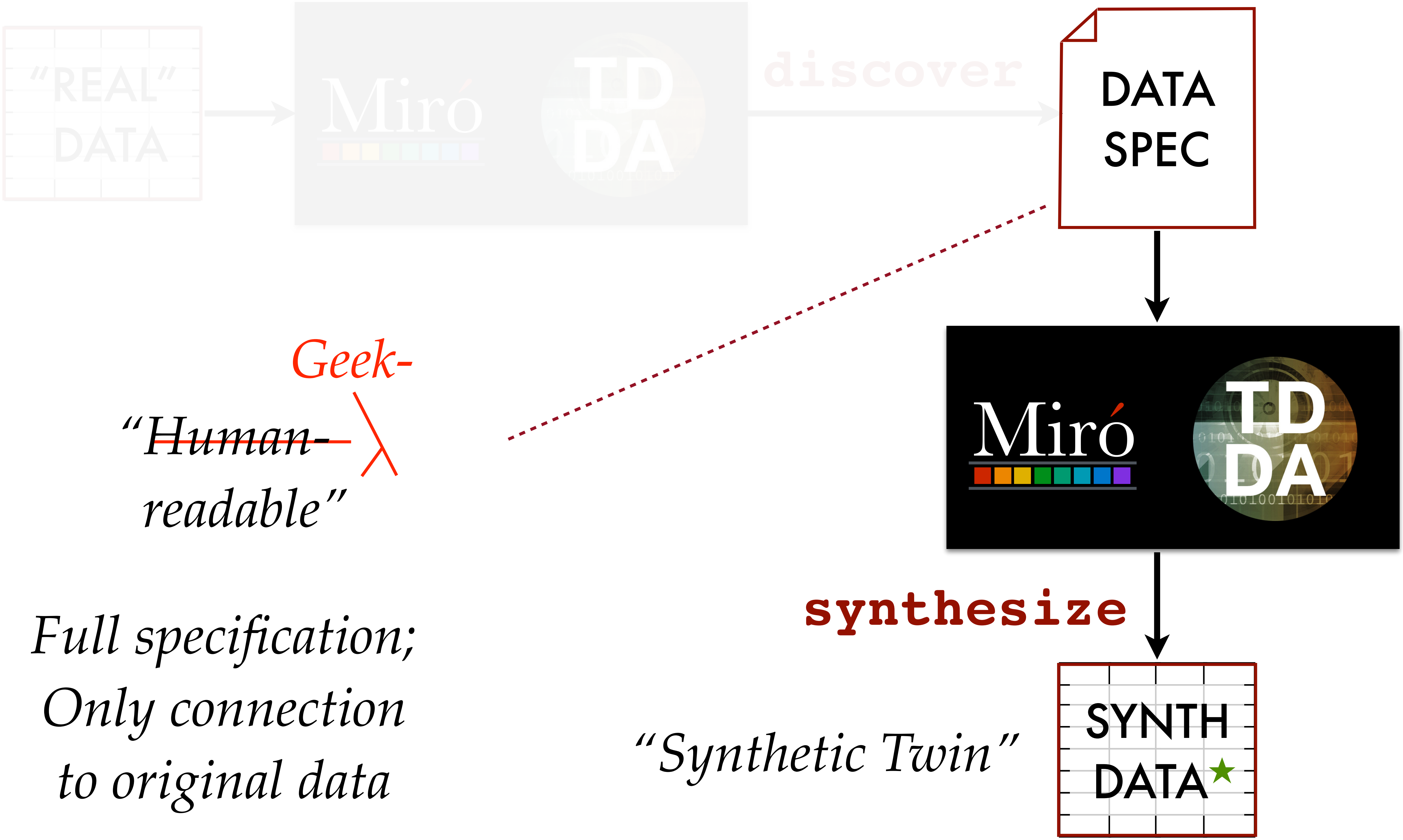
DATA SPEC

SYNTHESIZER

SYNTH DATA

*"Synthetic Twin"*

# CONSTRAINT-BASED DATA SYNTHESIS



*Geek-*

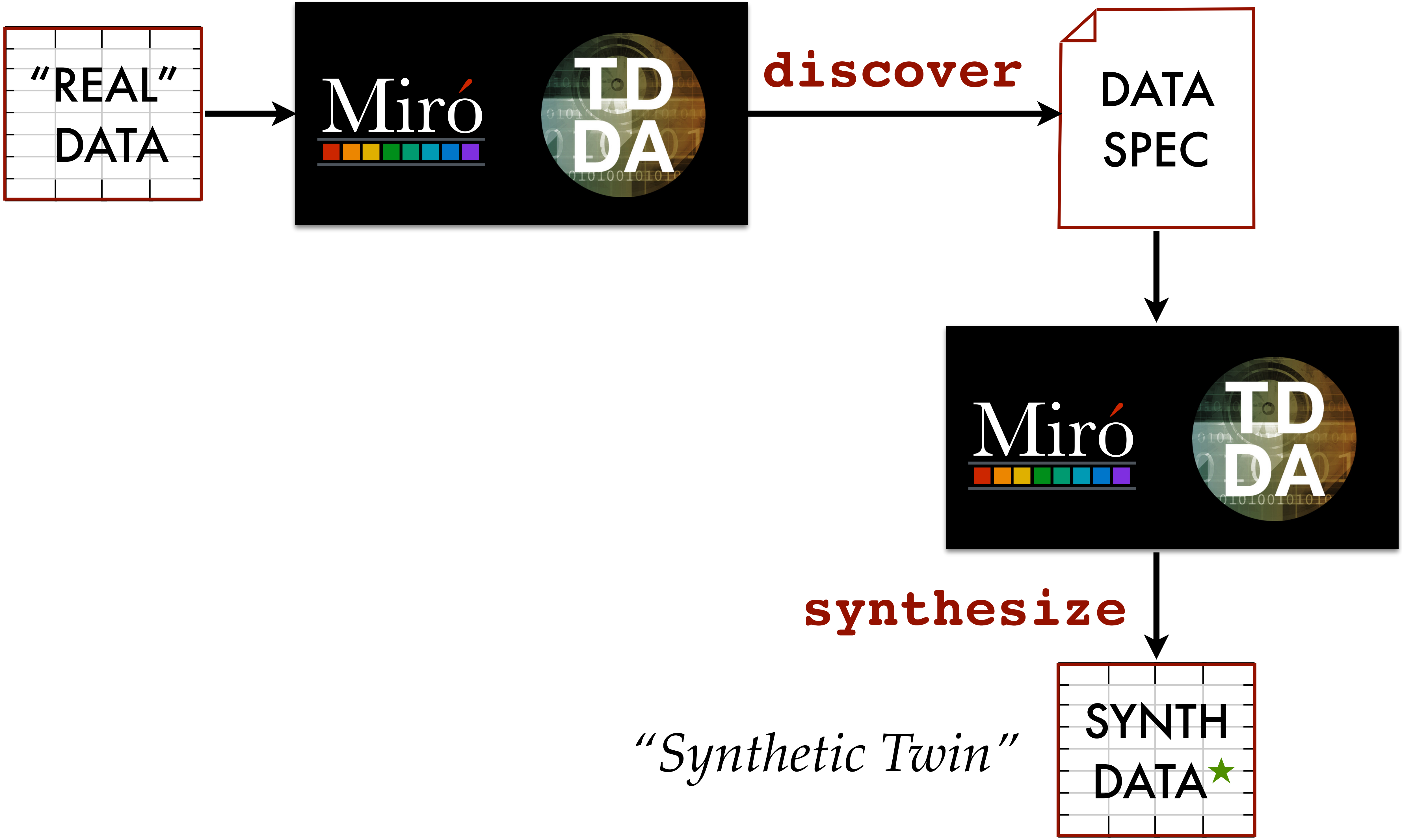
*"~~Human-~~ readable"*

*Full specification;*
*Only connection*
*to original data*

*"Synthetic Twin"*

DEMO

# CONSTRAINT-BASED DATA SYNTHESIS



"REAL" DATA

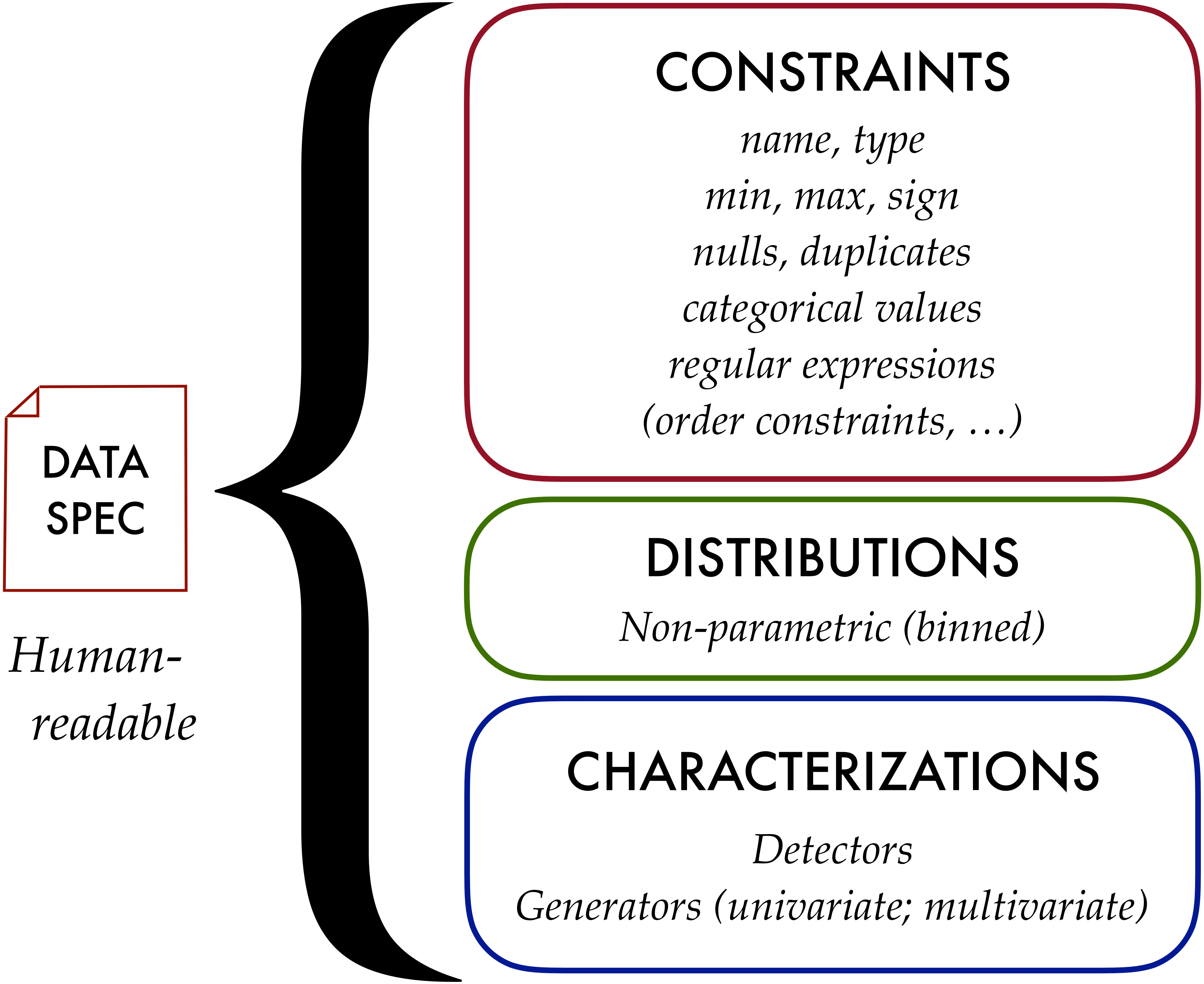
discover

DATA SPEC

synthesize

SYNTH DATA★

*"Synthetic Twin"*



DATA SPEC

*Human-readable*

## CONSTRAINTS

*name, type*

*min, max, sign*

*nulls, duplicates*

*categorical values*

*regular expressions*

*(order constraints, …)*

## DISTRIBUTIONS

*Non-parametric (binned)*

## CHARACTERIZATIONS

*Detectors*

*Generators (univariate; multivariate)*

# AVAILABILITY

## NOW: TDDA LIBRARY

- *Constraint discovery*
  - *inc. Rexpy (regular expression generation)*
- *Data Verification*
- *Reference Testing for Data Science*
- *(Alpha) Automatic Test Generation*

```
pip install tdda

git clone https://github.com/tdda/tdda.git
```

# AVAILABILITY

## “SOON”: XERPY

- *String generation from regular expressions*

*I hereby commit…*

# AVAILABILITY

## POSSIBLY SOMETIME

- *Miró: All the data synthesis stuff*
- *… and everything else*

njr@StochasticSolutions.com

http://tdda.info

https://github.com/tdda

#tdda*

@tdda0 @njr0

*Correct interpretation: Zero*

*Error of interpretation: Letter "Oh"*

**http://www.tdda.info/pdf/constrained-data-synthesis-euroscipy-2019.pdf**

* *tweet (DM) us email address for invitation Or email me.*