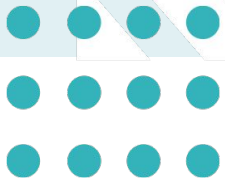
DataHelix - the data generator

Andrew Carr & Simon Laing

1st May 2019

The problem



The Problem

FCA FINANCIAL CONDUCT AUTHORITY

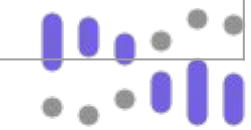
0101010101010101010101111100001101010111100001111000

DataHelix Generator

Solutions in the Market

Solutions on the Market

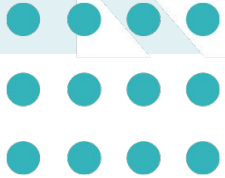
Tool name	Complex conditions	User functions to create complex types	Common FS types (RIC, ISIN, CUSIP, …)	Random mode	Profiler to auto generate profiler from real data	Violation mode	Ability to stream data
Mockaroo	Yes	No	No	Yes	No	No	No
Redgate (SQL) data generator	Yes	No	No	Yes (seeded)	No	No	No
generatedata.com	No	No	No	Yes	No	No	No
FINRA data generator	Yes	Yes	Yes	Yes	No	No	Yes



DataHelix
Generator

Data Helix

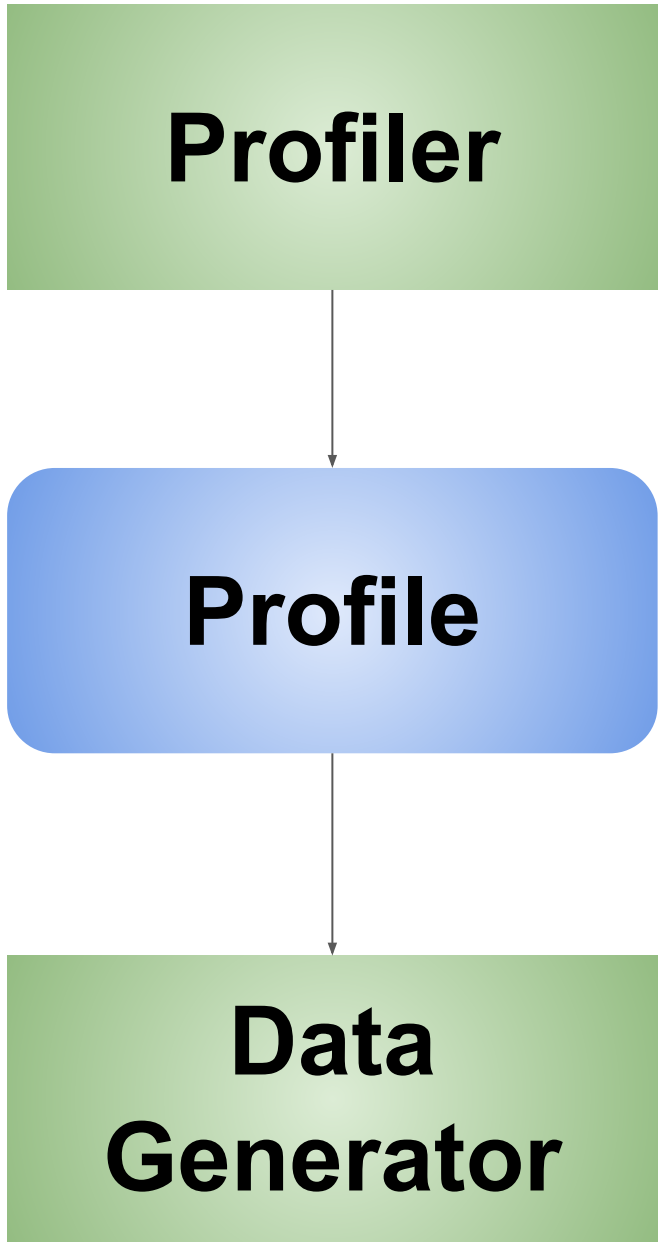
Open source Data Generator



Data Helix - USPs

- What is the Data Helix?
  - Generator (in beta)
  - Profiler (in development)

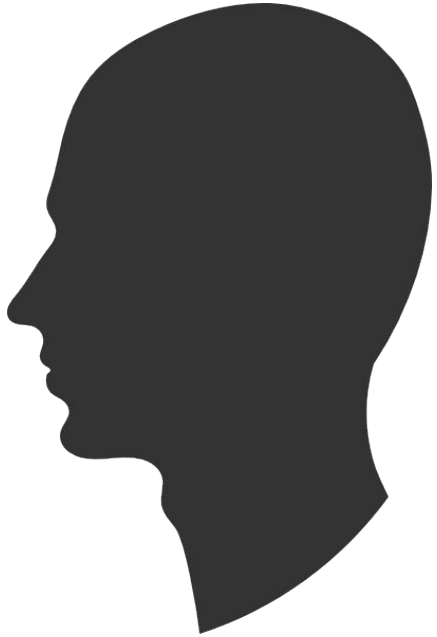
- Being considered as a contribution into FINOS



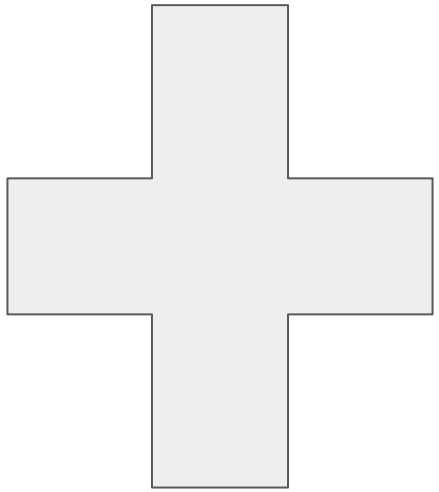
DataHelix Generator

Data Helix - A possible solution

Data Profile



Test data attributes



DataHelix
Generator

Data Helix - USPs

Tool name	Complex conditions	User functions to create complex types	Common FS types (RIC, ISIN, CUSIP, …)	Random mode	Profiler to auto generate profiler from real data	Violation mode	Ability to stream data
Data Helix	Yes	Yes	Yes	Yes	Yes	Yes	Yes

- Profiler → Production data → Data Profile using AI
- Data Profile from Production → review → clean Data Profile for generating Production like data
- Data Profile + Test Data attributes = random data | data violations

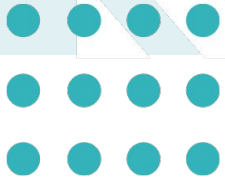
DataHelix
Generator

Data Profile

- Columns and types (integer, float, string, date)
- Complex types (ISINs, CUSIPs, RIC, …)
- Range for column
- Conditions for column (restrictions under certain conditions)

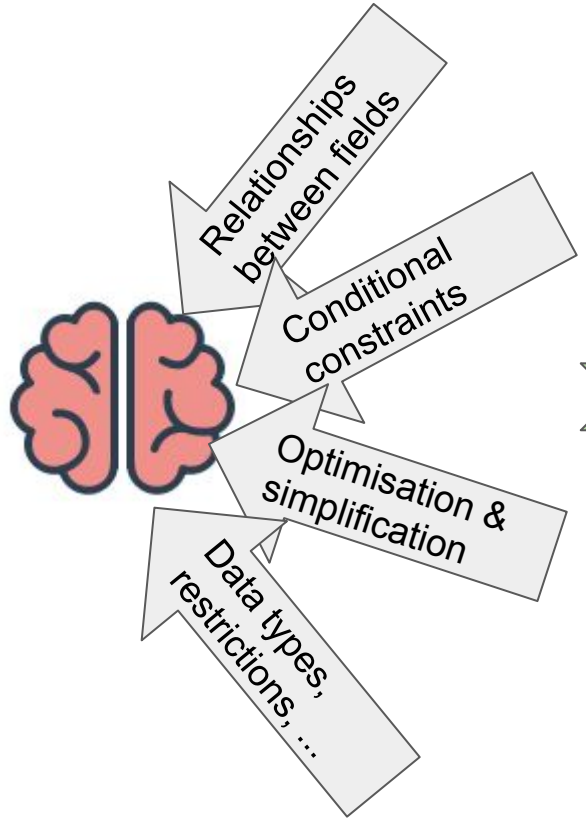
DataHelix
Generator

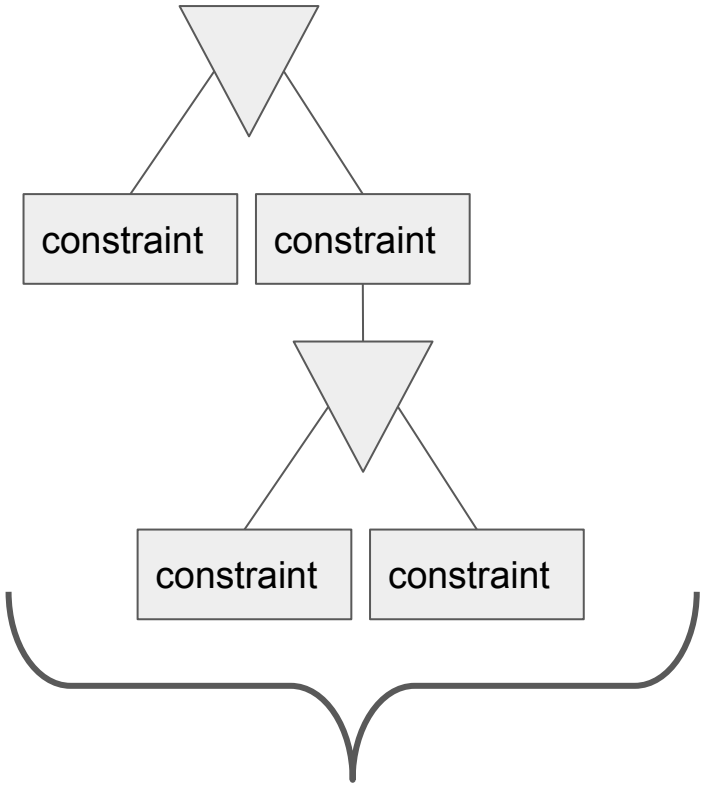
Data Generator
What can it do?



Conditions

Profile:
- Fields
- Constraints
- Restrictions
- Relationships





Boolean logic

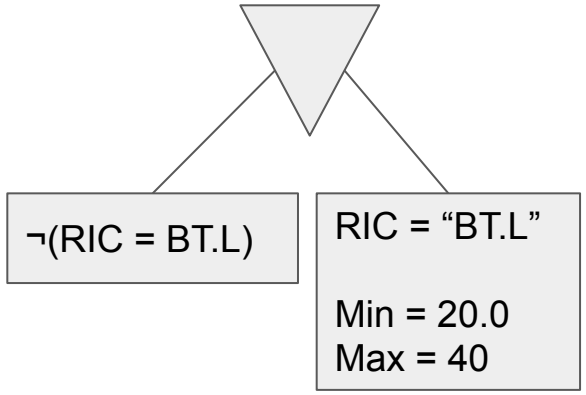
Provides guarantees about data production

DataHelix Generator

Examples

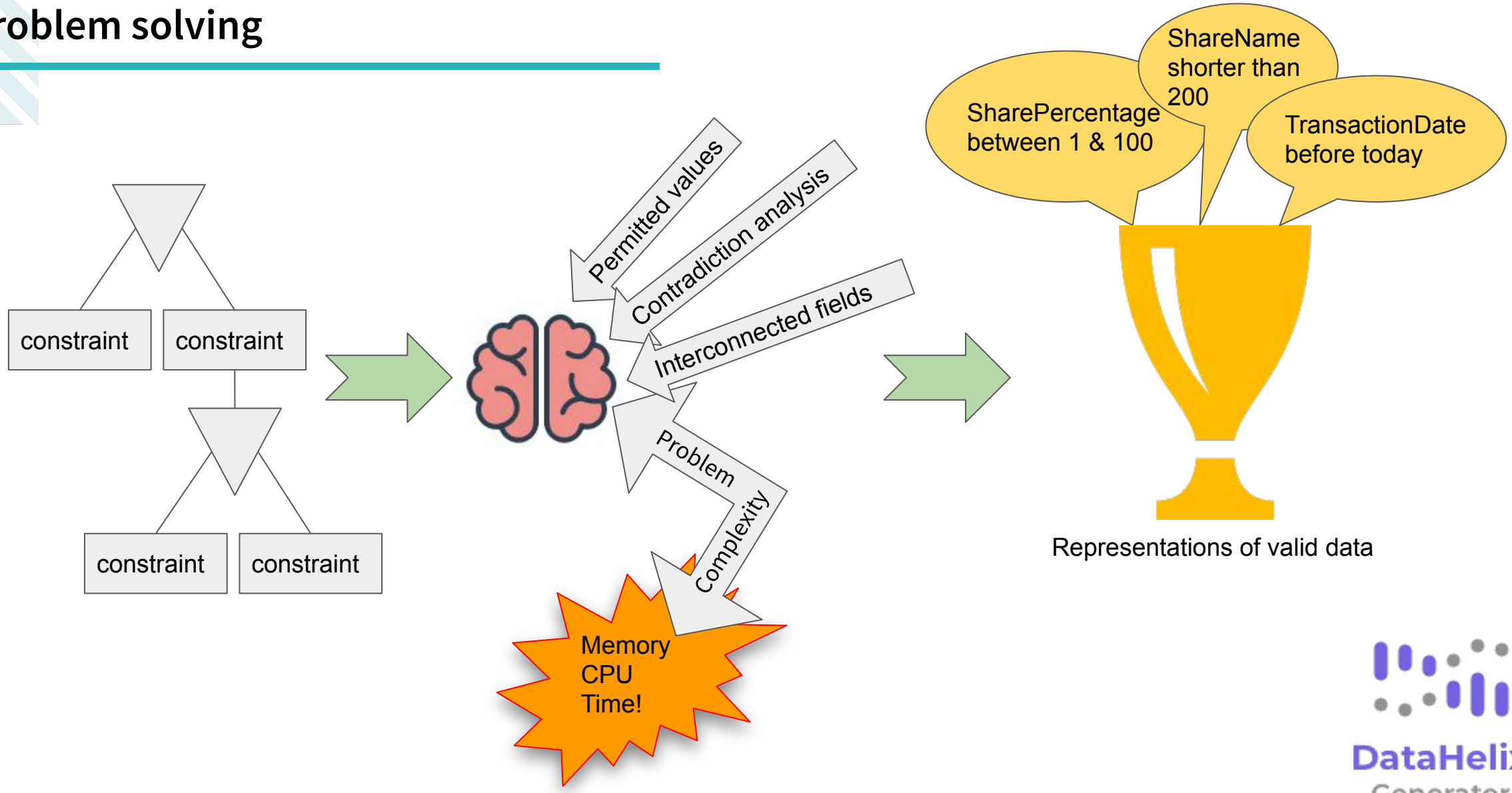
- If RIC_Code column = “BT.L” then Price Column Min = 20.0, Max = 40

```
{
    "if": { "field": "ric_code", "is": "equalTo", "value": "BT.L" },
    "then": { "allOf": [
        { "field": "min", "is": "equalTo", "value": 20.0 },
        { "field": "max", "is": "equalTo", "value": 40 } ]
    }
}
```



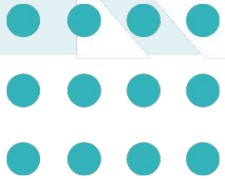
DataHelix
Generator

Problem solving



DataHelix Generator

Profiler
What can it do?



Profiler

- Uses traditional stats and AI to magically guess data profile
- Rules, types, correlations between columns
- Creating Profile from Production data → quick and easy
- Built on Apache Spark to scale

DataHelix
Generator

Summary

- Data Helix is ready for use
  - Data Generator (in beta)
  - Profiler (in development - not needed)
- Being considered for FINOS contribution

- Visit - https://github.com/scottlogic/datahelix/
- Or chat to Simon, Colin or myself
- Thank you for your time and questions

DataHelix
Generator