Scalable Discovery of Complementary Datasets in Data Lakes
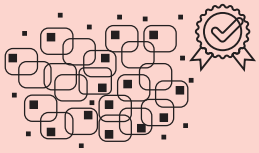
Andra Ionescu, Marios Fragkoulis, Asterios katsifodimos
{a.ionescu-3, m.fragkoulis, a.katsifodimos}@tudelft.nl

Complementarity

Two datasets are complementary if , when combined, they increase their data value.
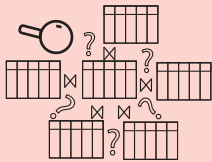
Attr#1	Attr#2	Attr#3	Attr#4	Attr#5	Attr#6

Schema complement
- Increase the number of features for ML
- Enhance an entity with more attributes

Entity complement
- Increase the number of data points

Row complement
- Partial union
- Fill in missing values

Column complement
- Partial join
- Fill in missing values

Challenges

Large number of datasets

Ranking results

Finding meaningful joins

Work in progress
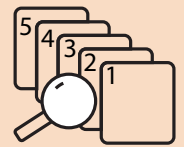
Quantify complementarity

Exploring ranking functions

User interactions as context

Exploring algs for union and join

TU**Delft**

Delta: Delft's Data Management Group