Data-Integration Web-Tool

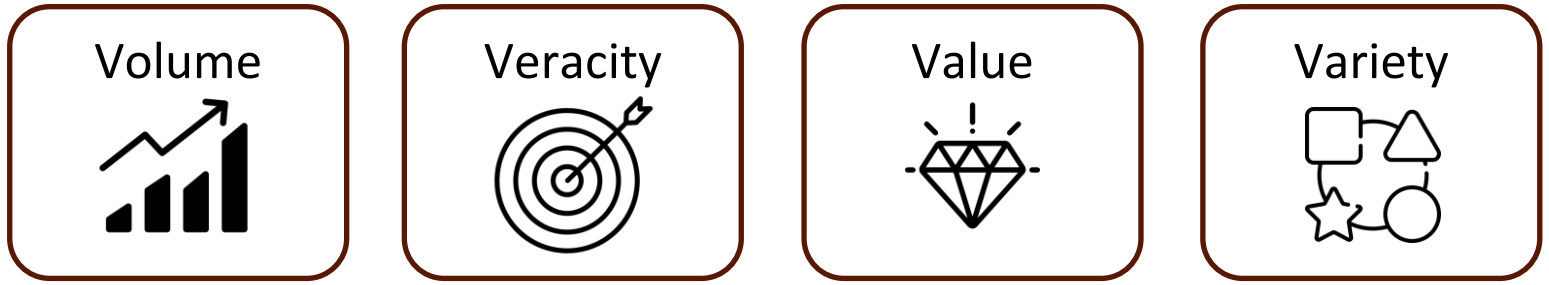
Praxisphase at Institute of Computer and Communication Technology (ICCT): Big Data Analytics

Leonard Traeger <*leonard.traeger@th-koeln.de*>
Ph.D. Candidate in Information Systems

Prof. Dr. Andreas Behrend <*andreas.behrend@th-koeln.de*>

Technology
Arts Sciences
TH Köln

Problems with Big Data and Integration



"Data Scientists spend more time looking for data than analysing it!" – Stonebraker (2018)

Data Integration critical cost factors for Mergers & Acquisitions (Ernst & Young 2023)

Example

Schema *S1*

CLIENT

CID	NAME	ADDRESS	PHONE
1	Leonard Traeger	Betzdorfer Straße 2, Köln, 50827	0157012345
2	Andreas Stock	Betzdorfer Straße 3, Köln, 50827	0157112345
3	Simon Haus	Betzdorfer Straße 4, Köln, 50827	0157212345

Schema *S2*

CUSTOMER

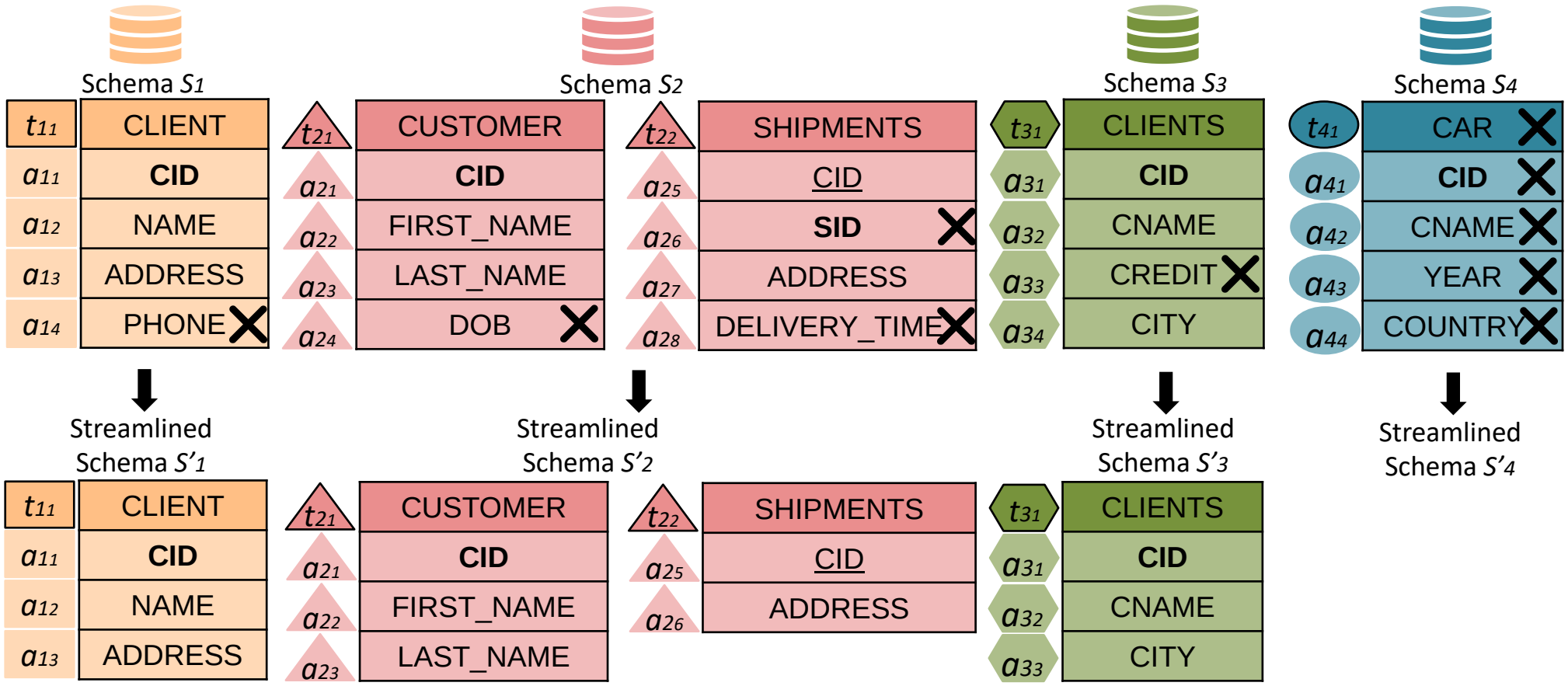
CID	FIRST_NAME	LAST_NAME	DATE_OF_BIRTH
1	Hannah	Sitz	01.01.2000
2	Edgar	Muster	01.01.2001
3	Mathias	Polster	01.01.2002

SHIPMENTS

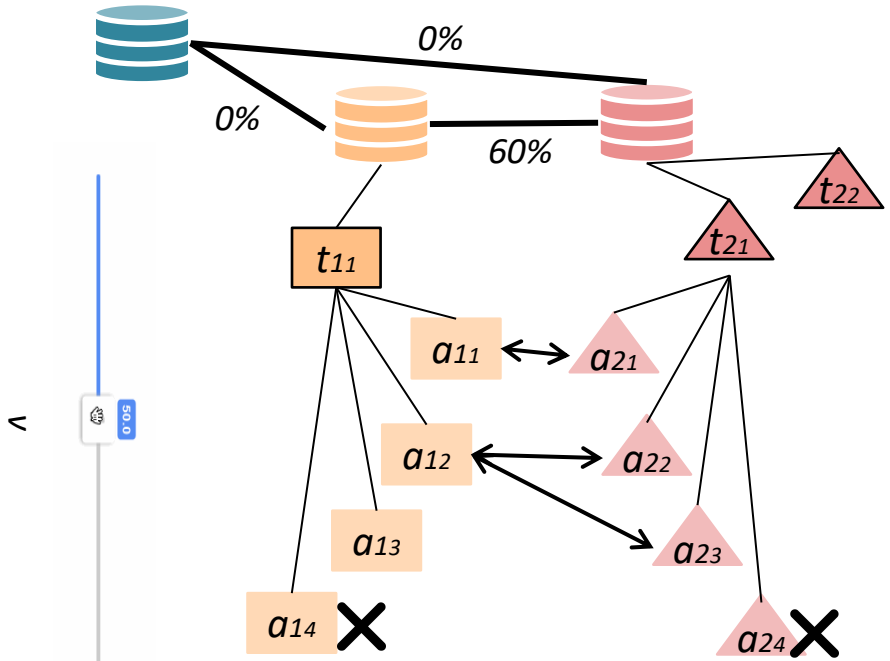
CID	SID	ADDRESS	DELIVERY_TIME
1	1	Betzdorfer Straße 2, Köln	01.02.2025
1	2	Betzdorfer Straße 2, Köln	02.02.2025
3	3	Betzdorfer Straße 4, Köln	01.01.2002

1. Welche Schema Elemente sind relevant und welche irrelevant für eine integrative Sicht?
2. Welcher SQL-Ausdruck liefert alle Kunden mit Adressen?

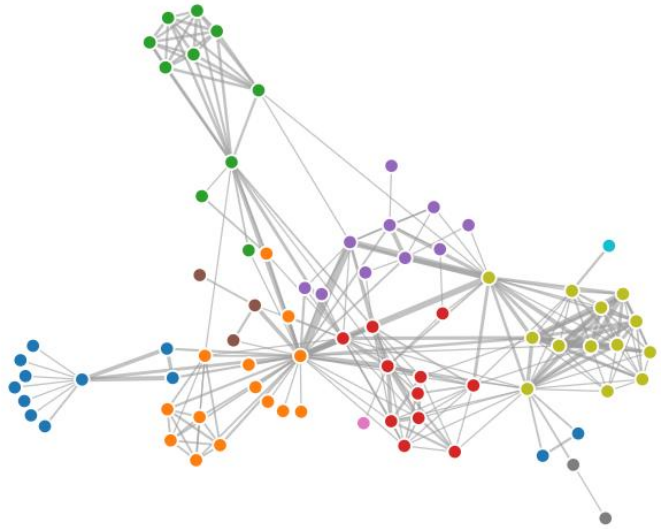
Scoping Example



Technical Hints



Schemas as a Graph



https://observablehq.com/@d3/force-directed-graph/2
or alternative implemented in web-based page with
HTML, CSS, JavaScript, TypeScript, …, React, Vue.js, Angular

Given:

1. Hierarchical Schema Elements

2. Scoping

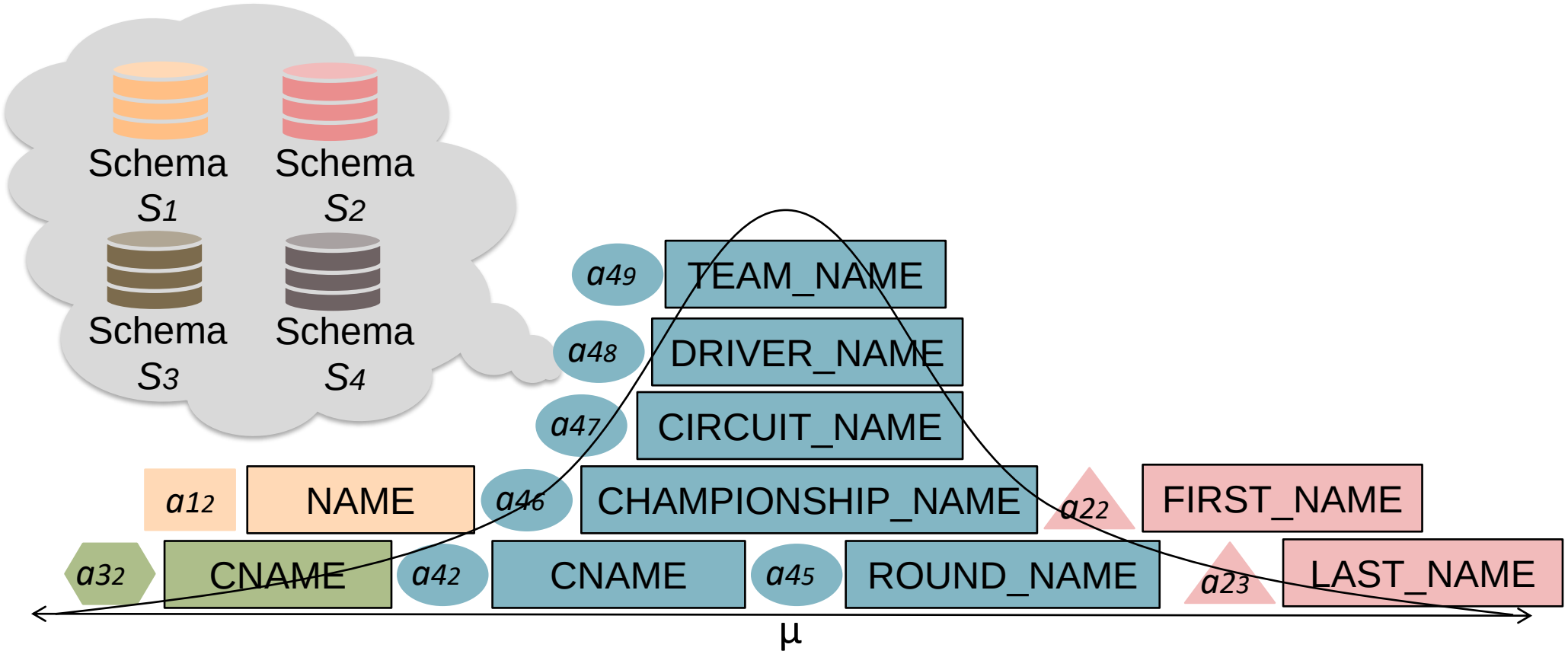
✕ % v

3. Linkages

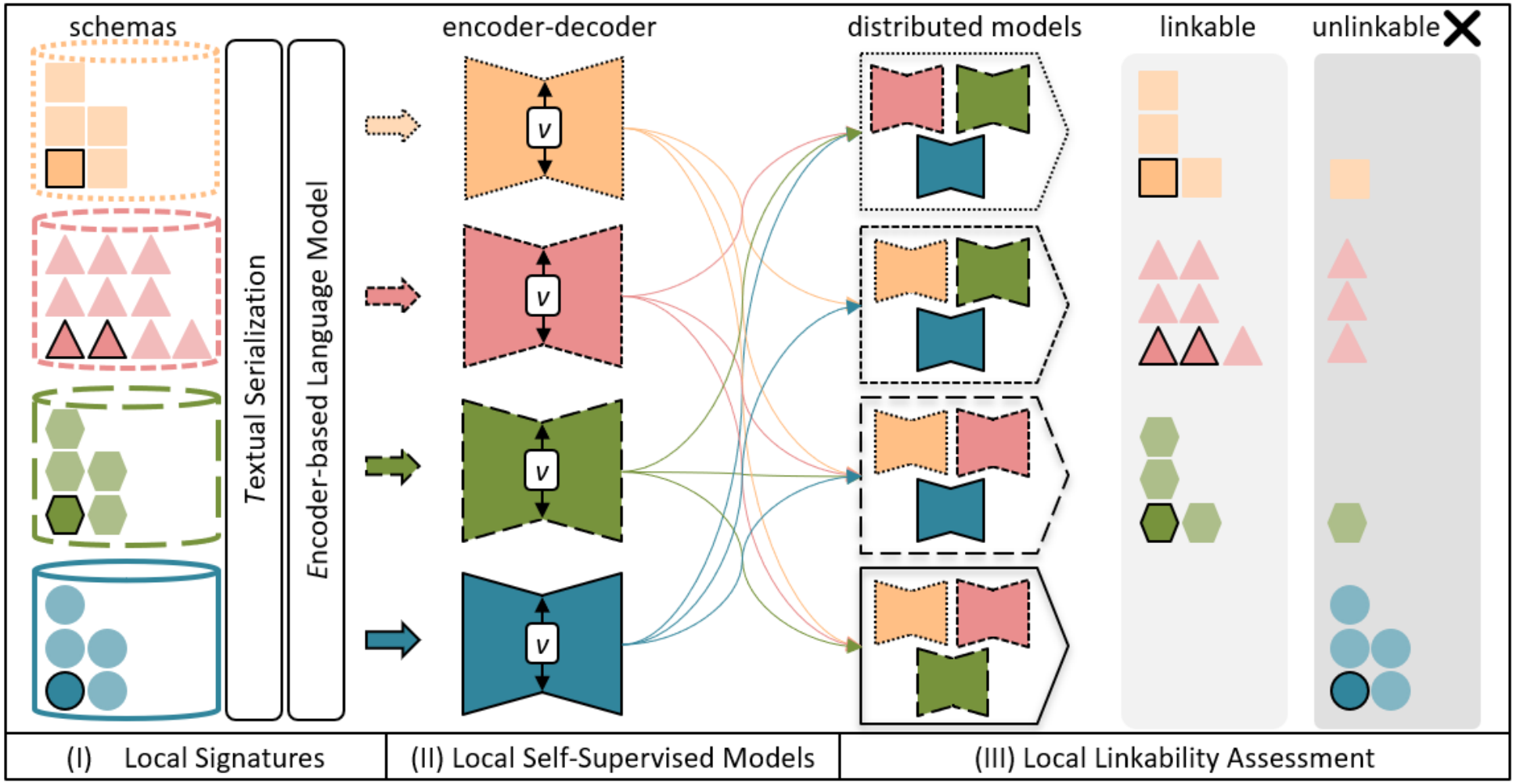
⟷

Webpage Hosting: https://pages.github.com/ (free) or alternative
reference project: https://github.com/leotraeg/Inteplato

A problem in Scoping

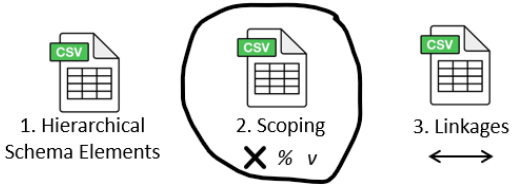


Collaborative Scoping



Datasets

Given:



1. OC3FO_schema_elements_dataset
2. OC3FO_collaborative_scoping
3. OC3_linkages

1. Schema Elements

- Visualize the schemas as-is

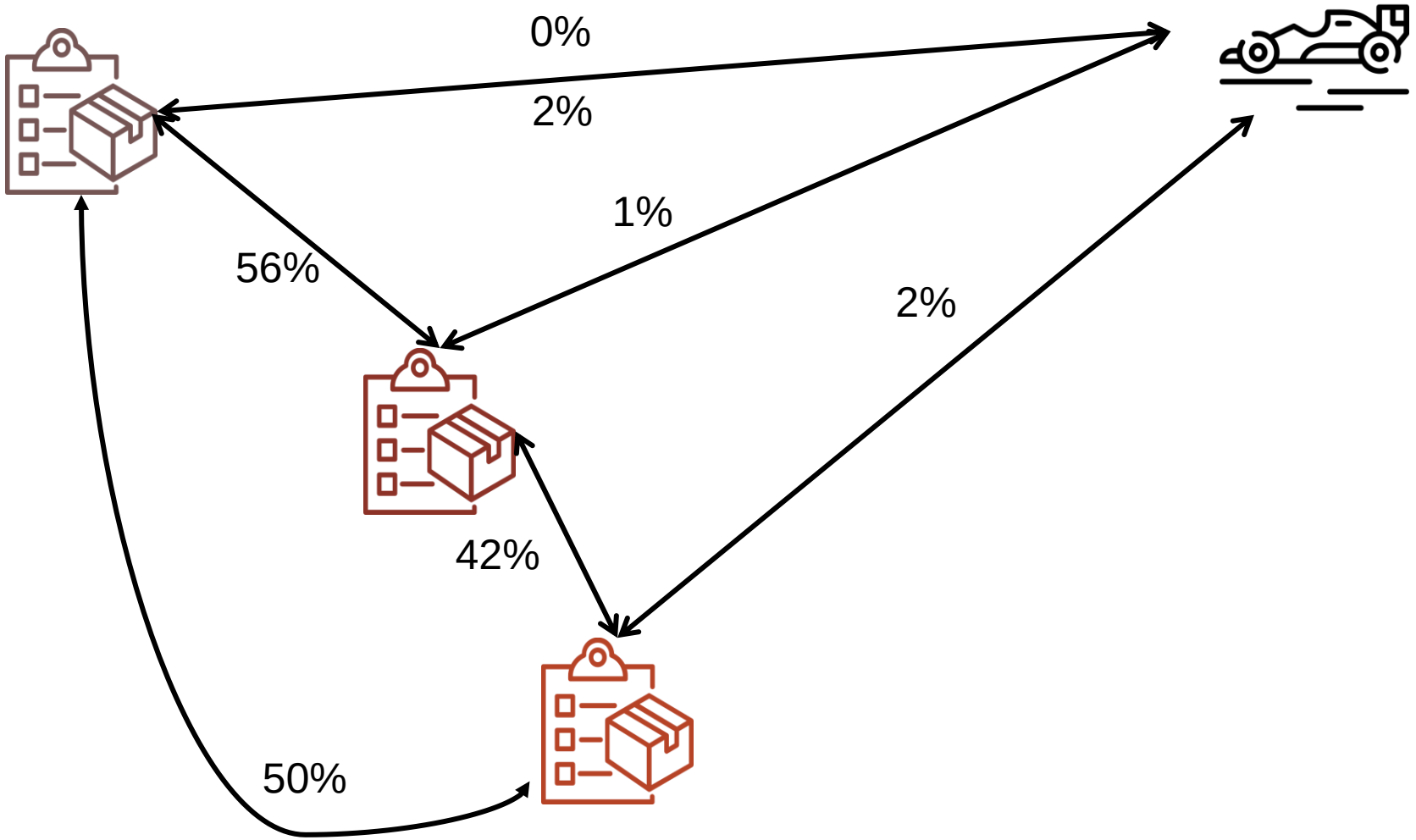
2. Collaborative Scoping

- ML-based computation on **predict_linkability** of schema elements in relationship to **v**

True	False
	✕

```
Data columns (total 22 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   id                   28413 non-null  object
 1   type                 28413 non-null  object
 2   parent_id            28413 non-null  object
 3   schema               28413 non-null  object
 4   name                 28413 non-null  object
 5   parent_name          25047 non-null  object
 6   datatype             25047 non-null  object
 7   constraints          8019 non-null   object
 8   text_sequence        28413 non-null  object
 9   label_linkability    28413 non-null  bool
 10  OC-ORACLE            28413 non-null  float32
 11  OC-MYSQL             28413 non-null  float32
 12  OC-SAP               28413 non-null  float32
 13  FORMULA              28413 non-null  float32
 14  OC-ORACLE_agree      28413 non-null  int64
 15  OC-MYSQL_agree       28413 non-null  int64
 16  OC-SAP_agree         28413 non-null  int64
 17  FORMULA_agree        28413 non-null  int64
 18  overall_agreement    28413 non-null  int64
 19  predict_linkability  28413 non-null  bool
 20  confusion            28413 non-null  object
 21  v                    28413 non-null  float64
dtypes: bool(2), float32(4), float64(1), int64(5), object(10)
memory usage: 4.0+ MB
```

Linkable Agreement Ratio at v=70%



OC-ORACLE
OC-MYSQL
OC-SAP
FORMULA