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

ICCT: Big Data Analytics

Research Areas

- In-Database Analytics, Big Data, No SQL (e.g. Graph Databases)
- Predictive Reasoning, Intelligent Systems
- Monitoring Applications, Data Stream Processing, Temporal Data, Index Structures
- Data Integration with Machine Learning, Schema Matching, Entity Resolution

Prof. Dr. Andreas Behrend

<andreas.behrend@th-koeln.de>

Leonard Traeger

Ph.D. Candidate in Information Systems *@UMBC*

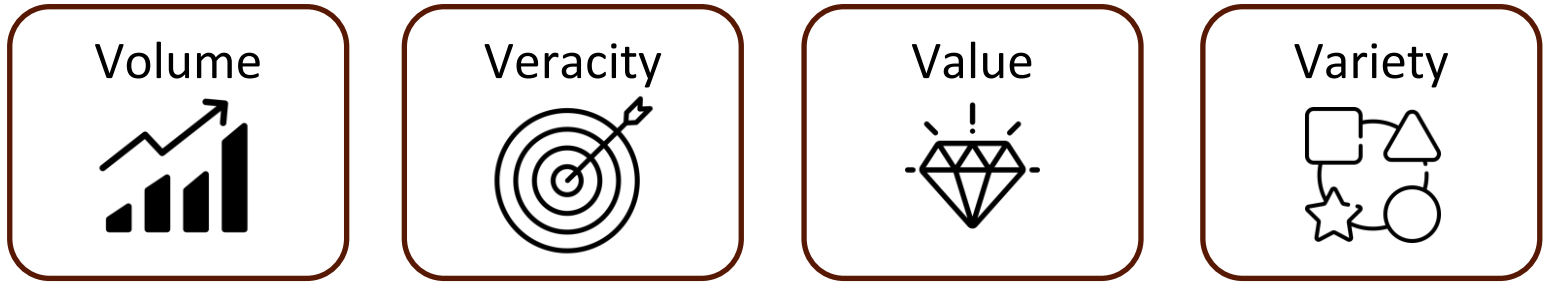
<leonard.traeger@th-koeln.de>

Campus Deutz, Betzdorfer Straße 2
50679 Köln

ZW-7-17/21 Computer Pools

ZW-7-17/21 Meeting Room

Problems with Big Data and Integration



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

Microsoft
LinkedIn
LIEFERHELD
Lieferando
foodora
Google
YouTube
BAYER
MONSANTO

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

Example

Lieferando

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

LIEFERHELD

Schema *S2*

CUSTOMER

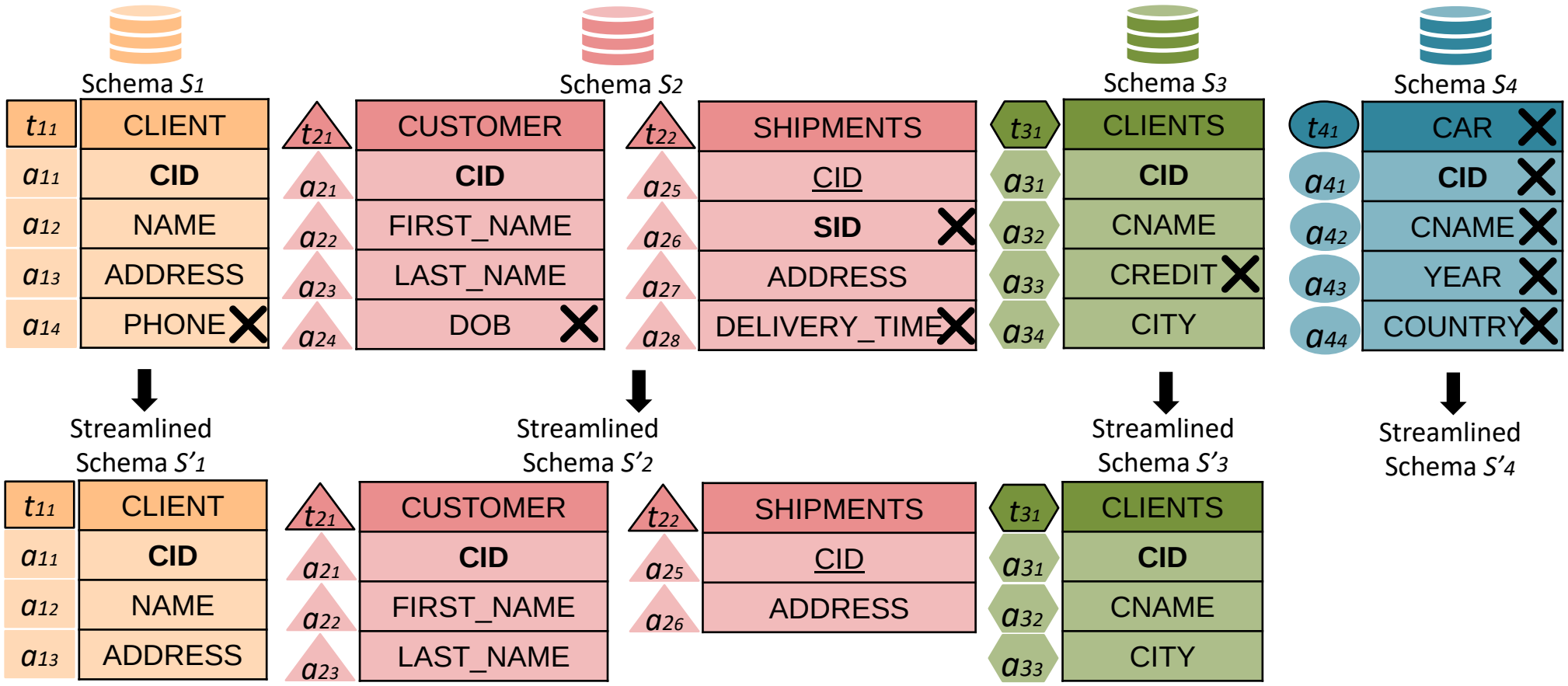
CID	FIRST_NAME	LAST_NAME	DATE_OF_BIRTH
1	Hannah	Sitz	01.01.2000
2	Edgar	Muster	01.01.2001
3	Mathias	Polster	01.01.2002

SHIPMENTS

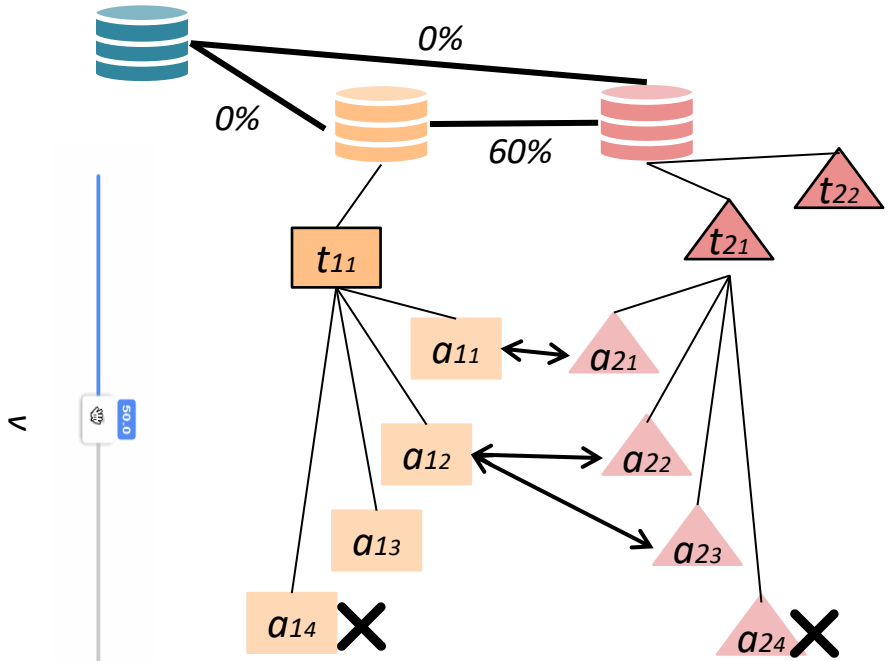
CID	SID	ADDRESS	DELIVERY_TIME
1	1	Betzdorfer Straße 2, Köln	01.02.2025
1	2	Betzdorfer Straße 2, Köln	02.02.2025
3	3	Betzdorfer Straße 4, Köln	01.01.2002

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

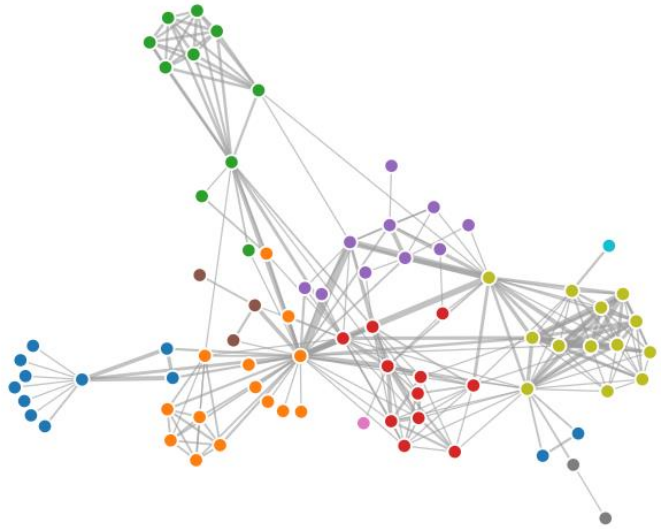
Scoping Example



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

3. Linkages

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

Praxisphase

Thema: Entwicklung einer Web-Anwendung zur Visualisierung von Relationalen Datenbanken Schemata als Graph mit Verlinkungen für die Datenintegration.

Aufgabenstellung: Im Rahmen dieser Projektarbeit soll eine moderne Web-Anwendung zur Visualisierung von Verlinkungen zwischen Graphen entwickelt werden. Hierfür sind zunächst relationale Datenbanken als Graph mit den Knoten und Kanten zwischen Schema, Tabelle, und Attribut zu visualisieren. Durch die Konfiguration von Parametern der ML stützenden Verlinkungsmethoden sind dann Kanten zwischen den Knoten der unterschiedlichen Datenbanken anzuzeigen.

Rahmenbedingungen:

- Entwicklung einer Web-Anwendung zur Visualisierung von relationalen Schemata als interaktive Graphen mittels D3.js o.ä. Bibliothek (Beispiel: https://observablehq.com/@d3/force-directed-graph/2).
- HTML, CSS und JavaScript Kenntnisse und Vorerfahrungen in modernen Frontend Frameworks wie React, Vue.js, Angular o.ä. wünschenswert.
- Die darzustellenden Knoten und Kanten werden samt ML generierten Werten als Flat-File (.csv) bereitgestellt, sodass eine aufwendige Back-End Anbindung zunächst entfällt. Perspektivisch können
  - Vektordatenbanken als Back-End angebunden werden.
  - Eingebettetes Aufrufen von ML Python-Skripten oder Algorithmen in HTML implementiert werden.
  - Benchmark Visualisierungen der ML-Modelle (Precision, Recall, F1, AUCs) erstellet werden.

Angebot

- Wissen zu Datenintegration.
- Teilnahme am wissenschaftlichen Arbeiten im Data Science und Big Data Management Bereich.
- Einblick in praktische Datenintegrationslösungen mittels KI.
- Selbstbestimmte Arbeit.
- Flexible Absprachen.

- Keine Bezahlung.

25.02.2025

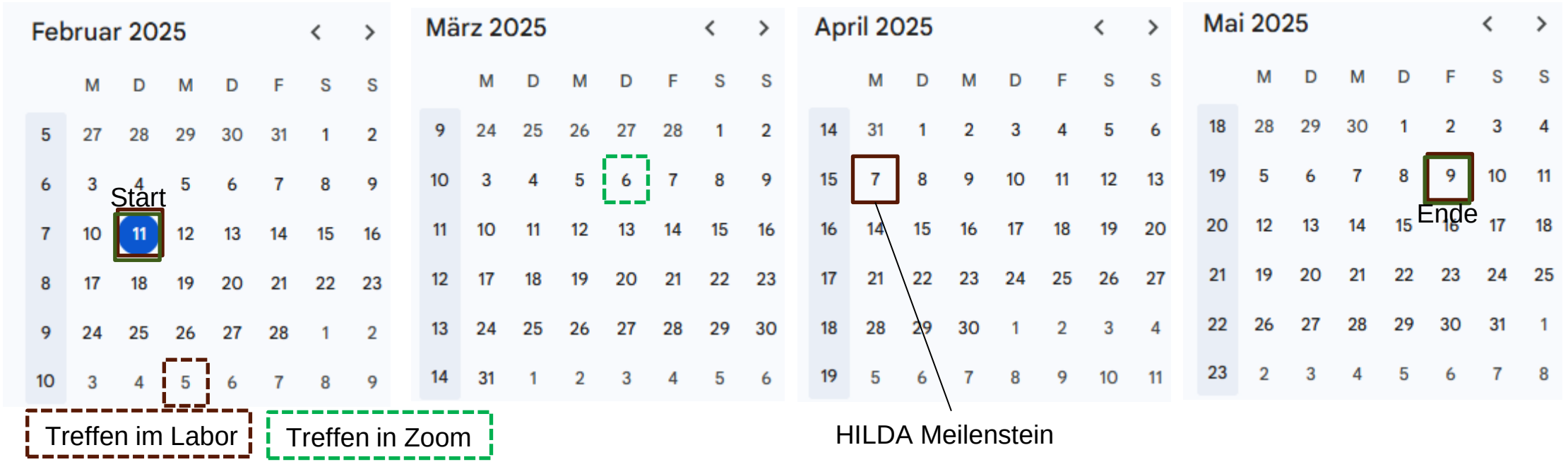
- GitHub anlegen

Planung

Selbstständige Planung und Erarbeitung eines Projekts

Dauer: 3 Monate je 40 Stunden (VZÄ)

Sem.	Module						Summe SWS
7	Praxisphase		Bachelorarbeit und Kolloquium				0
6			IT-Projekt-Management 4 SWS	8 Wahlmodule aus verschiedenen Gebieten der Technischen Informatik 32 SWS insgesamt			12
5	Präsentation und Kommunikation 3 SWS	Systementwurfs-praktikum 4 SWS	Fachübergreifende Kompetenzen u. Soft Skills 1 / 4 SWS				23
4	IT-Sicherheit 4 SWS	Software-Praktikum 4 SWS	Betriebsysteme und Verteilte Systeme 2 / 4 SWS	Betriebswirtschaft und Recht 4 SWS			24
3	Graph. Oberflächen und Interaktion 4 SWS	Software Engineering 4 SWS	Betriebsysteme und Verteilte Systeme 1 / 4 SWS	Netze und Protokolle 4 SWS	Datenbanken 4 SWS	Signalverarbeitung 4 SWS	24
2	Praktische Informatik 2 4 SWS	Algorithmen und Datenstrukturen 4 SWS	Grundl. der System-programmierung 4 SWS	Formale Sprachen-Automatentheorie 4 SWS	Mathematik 2 8 SWS		24
1	Praktische Informatik 1 4 SWS	Programmier-praktikum 4 SWS	Digitaltechnik 4 SWS	Elektrotechnische-Grundlagen für die TI / 4 SWS	Mathematik 1 8 SWS		24



Technology Arts Sciences TH Köln

Absprachen

Nächstes Treffen
06.03.2025
11:00 Uhr

Tools

- Kommunikation: **Zoom**
- Cloud Drive und Datenstruktur: Sciebo, GDrive, **GitHub**,…
- Entwicklung: **VisualStudio**, **GitHub**, …

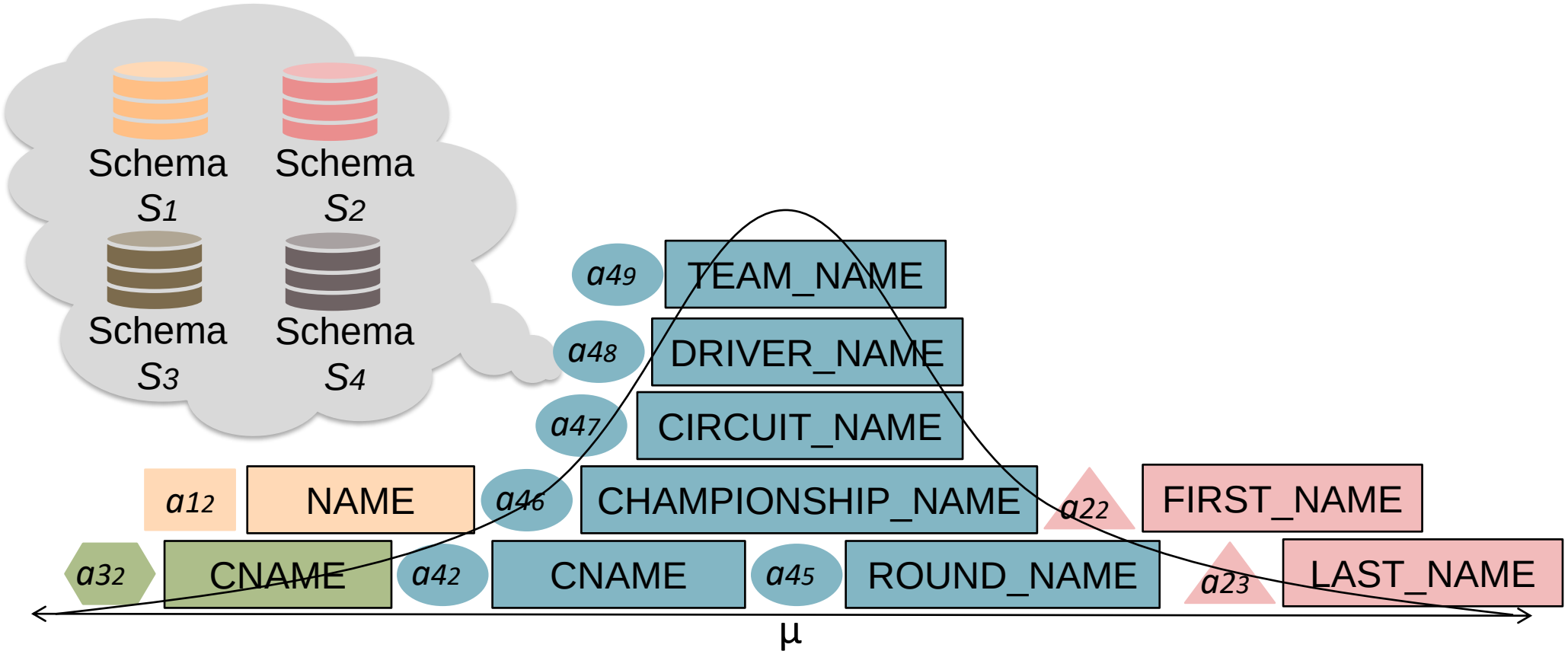
Agile Arbeitsweise

- Fortschritte werden fortlaufend präsentiert und Weiterentwicklungen als Arbeitspakete gemeinsam besprochen.
- Regelmäßigkeit der Treffen: 1-2 wöchentlich Montags/**Donnerstag Zooms**, Vor Ort Meilensteine
- *ToDo*: GitHub Tickets, TrelloBoard (Ready, Doing, OnHold/DependantOn, Done), taiga.io

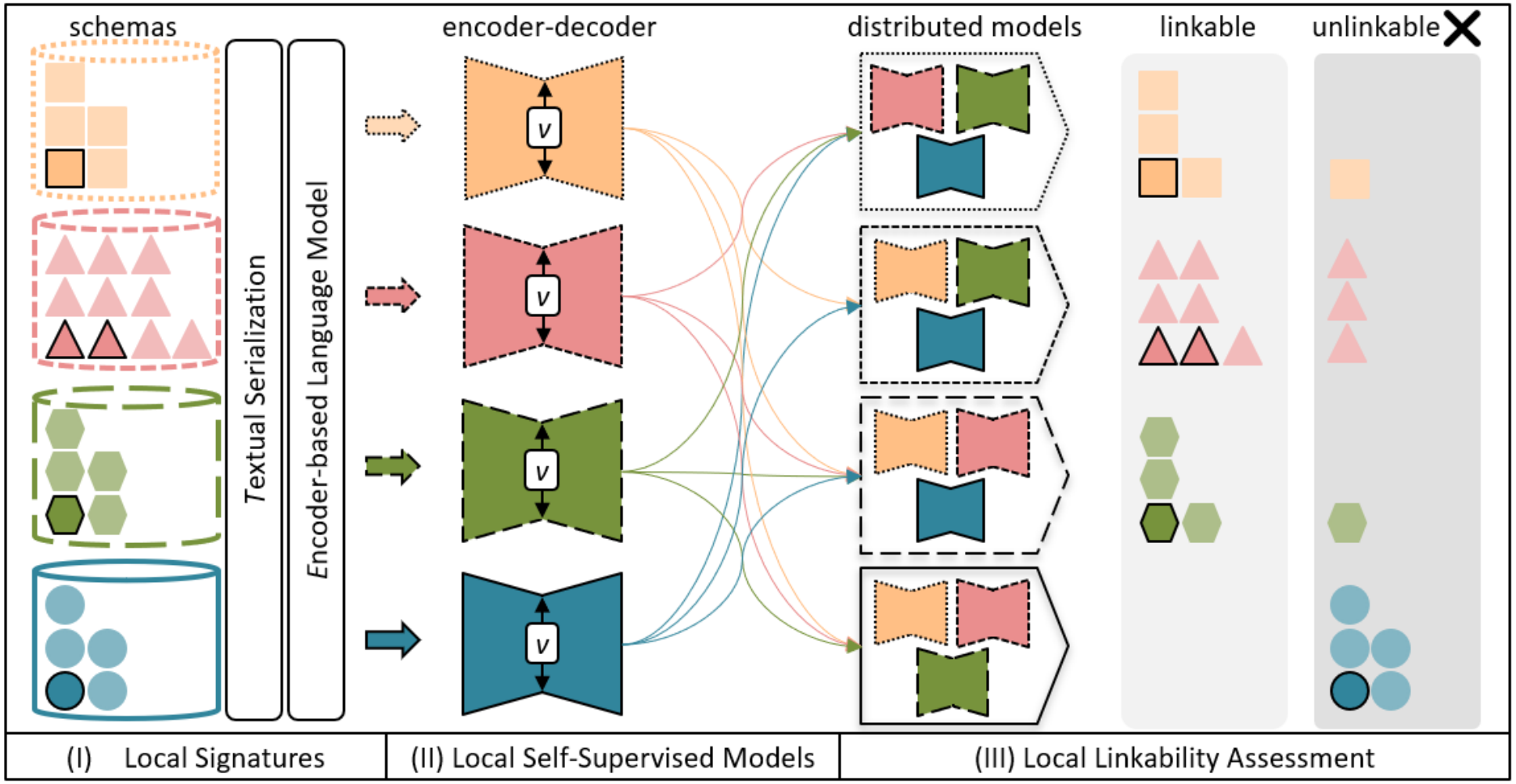
Programmierframework

- HTML, CSS, JavaScript, TypeScript, **React**, Vue.js, …
- **D3** Graph Kompatibilität
- **GitHub Pages** Kompatibilität

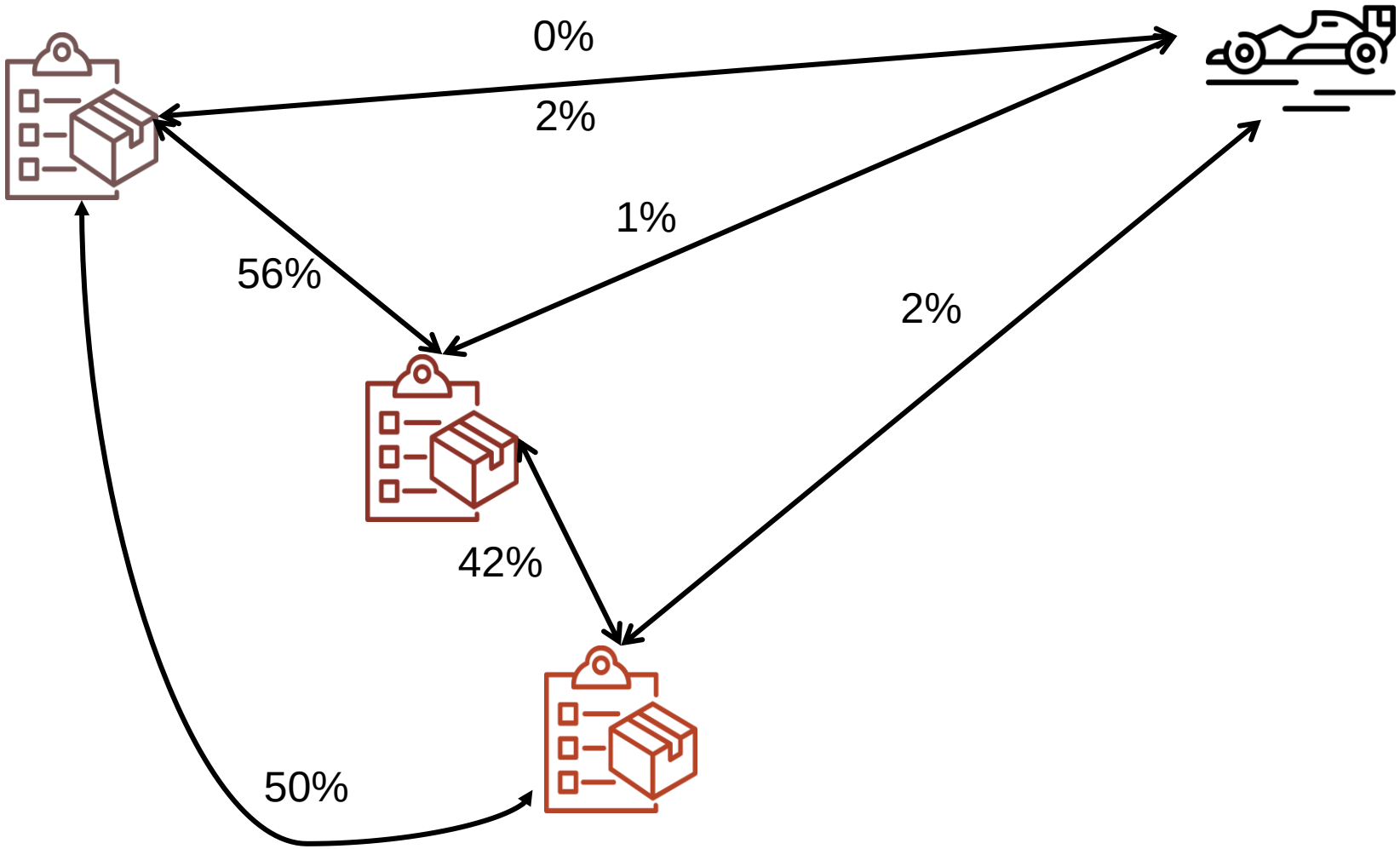
A problem in Scoping



Collaborative Scoping



Linkable Agreement Ratio at g=70%



OC-ORACLE
OC-MYSQL
OC-SAP
FORMULA