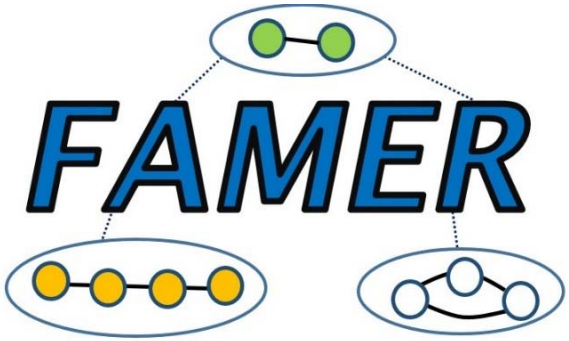
UNIVERSITÄT
LEIPZIG

ScaDS
DRESDEN LEIPZIG



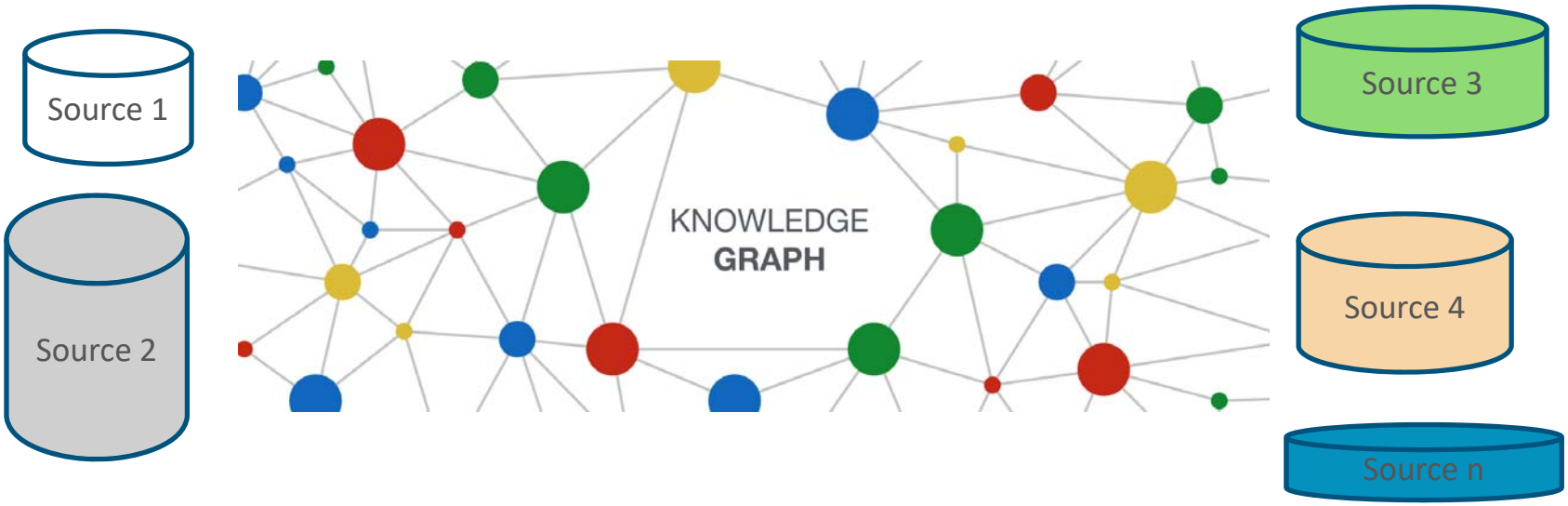
FAst Multi-source Entity Resolution System

Alieh Saeedi, Eric Peukert, Erhard Rahm

www.scads.de

MOTIVATION

- Physical data integration
  - Knowledge graph: Store data from **multiple sources** in a **graph-like** structure



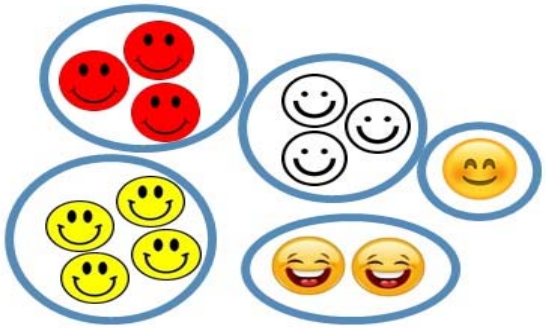
ScaDS DRESDEN LEIPZIG

MOTIVATION

- Automatic construction & maintenance of KG: **data quality**
- Challenges for data quality
  - Entity Resolution: The task of **identifying** and **linking** entities that refer to the **same** real-world entity

2 sources:
Binary linking

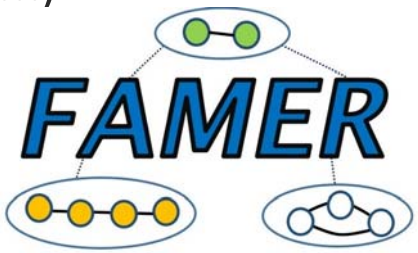
N sources:
Clustering



ScaDS DRESDEN LEIPZIG

FAMER

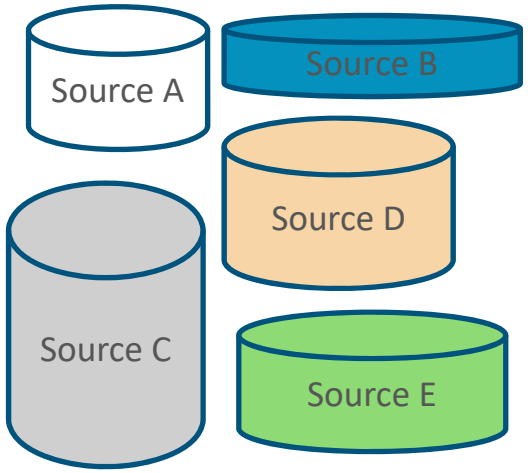
- **FAMER** (**FA**st **M**ulti-source **E**ntity **R**esolution system)
  - Scalable ER approaches for big data
    - Multiple data sources
    - Large volumes of data
  - Built on top of the distributed data flow framework Apache Flink and Gradoop
    - High scalability
    - Large amounts of data
    - Many machines



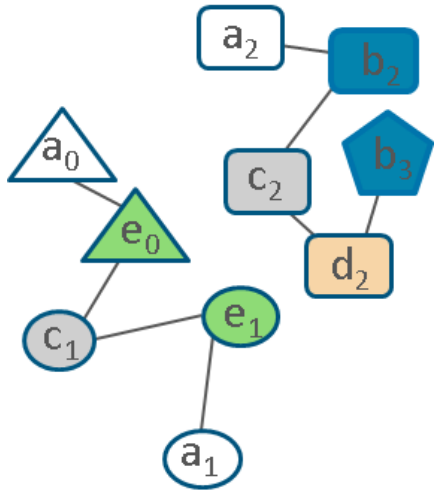
ScaDS DRESDEN LEIPZIG

FAMER OVERVIEW

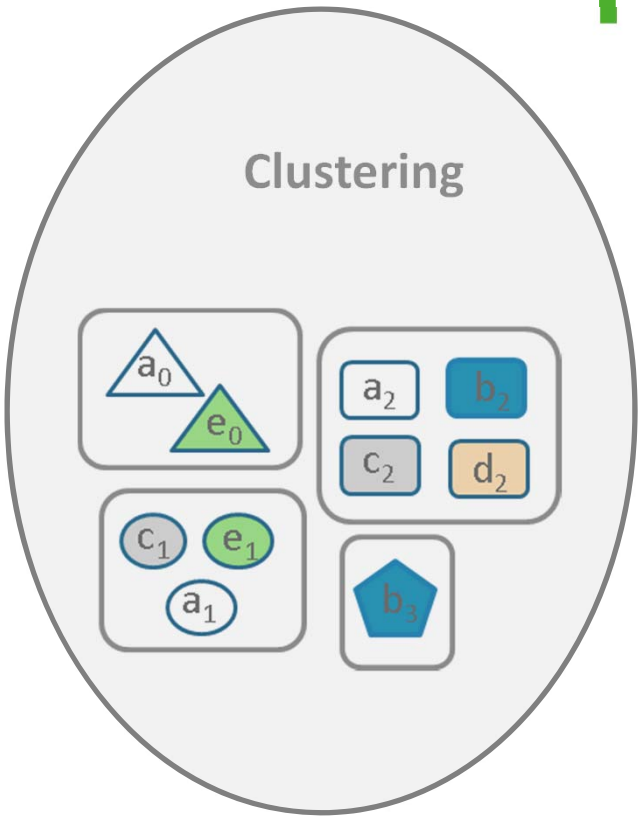
Input



Linking: Similarity Graph

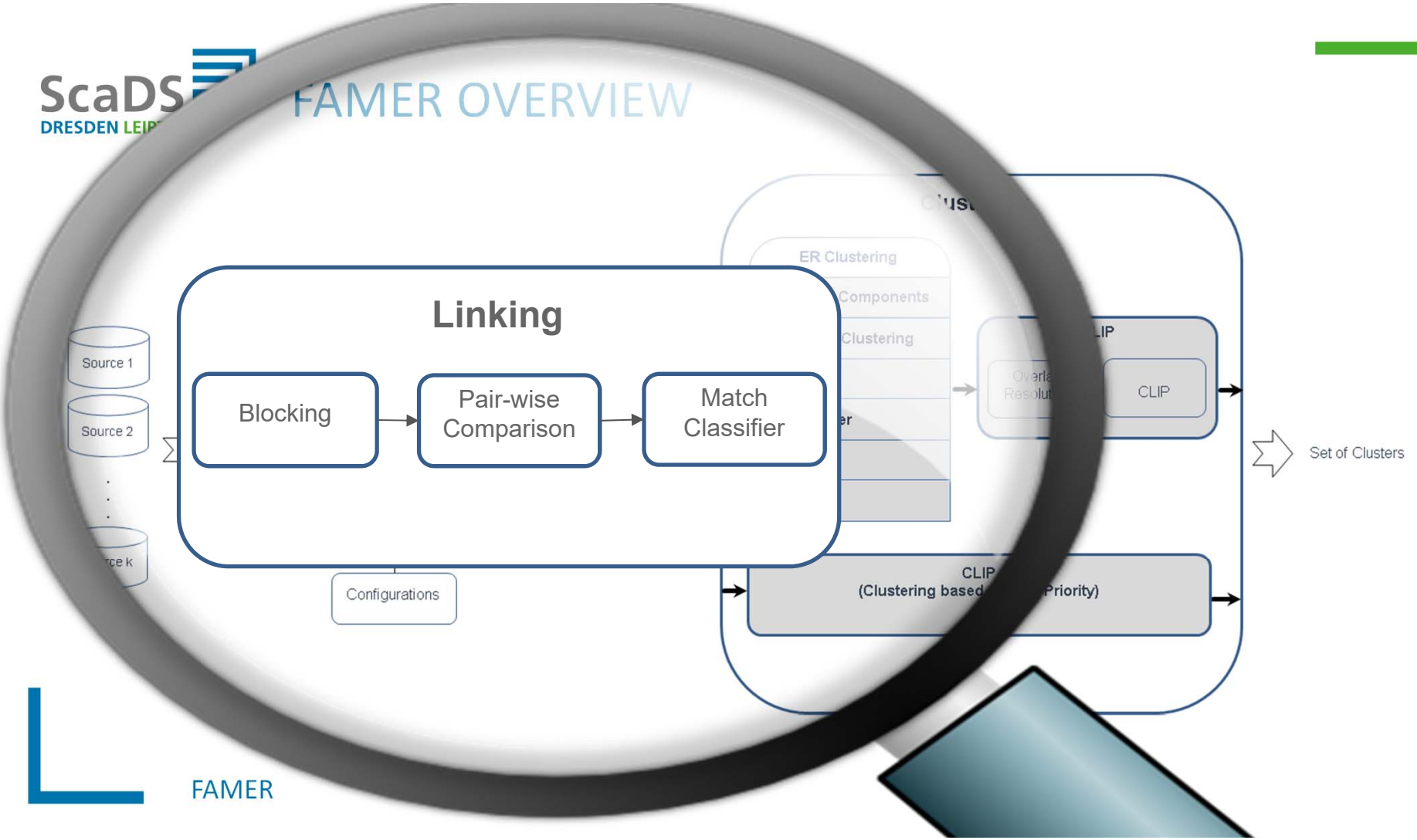


Clustering



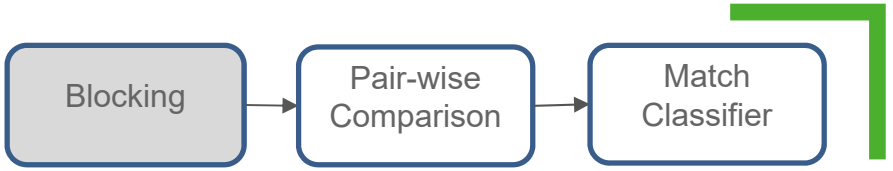
ScaDS DRESDEN LEIP

FAMER OVERVIEW



ScaDS DRESDEN LEIPZIG

FAMER LINKING

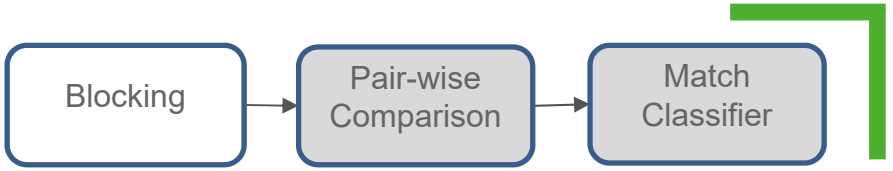


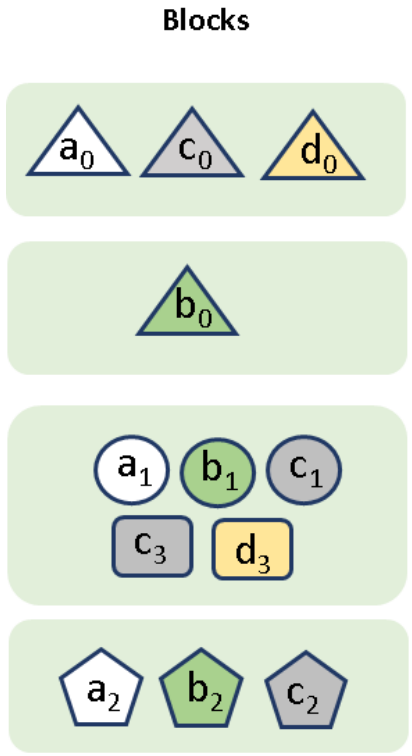
Id	Name	Surname	Suburb	Post code	SourceId
a_0	ge0rge	Walker	winston salem	271o6	Src A
b_0	George	Alker	winstom salem	27106	Src B
c_0	George	Walker	Winstons	27106	Src C
d_0	Geoahge	Waker	Winston	271oo	Src D
a_1	Bernie	Davis	pink hill	28572	Src A
b_1	Bernie	Daviis	Pinkeba	2787z	Src B
c_1	Bernii	Davs	pink hill	28571	Src C
a_2	Bertha	Summercille	Charlotte	28282	Src A
b_2	Bertha	Summeahville	Charlotte	2822	Src B
d_2	Brtha	Summerville	Charlotte	28222	Src D
c_3	Bereni	dan'lel	Pinkeba	27840	Src C
d_3	Bereni	Dasniel	Pinkeba	2788o	Src D

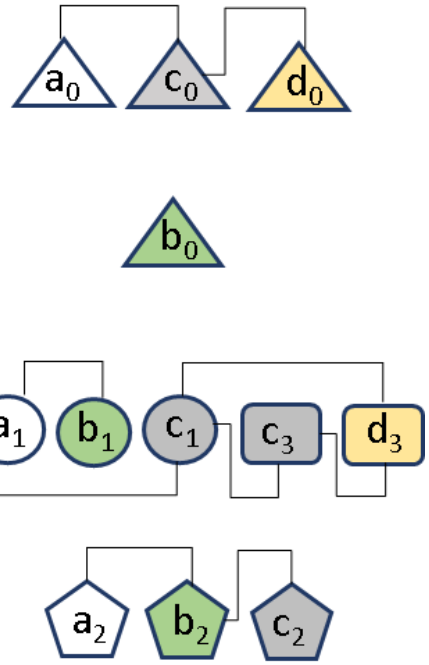
Id	key
a_0	wa
c_0	wa
d_0	wa
b_0	al
a_1	da
b_1	da
c_1	da
c_3	da
d_3	da
a_2	su
b_2	su
d_2	su

ScaDS DRESDEN LEIPZIG

FAMER LINKING

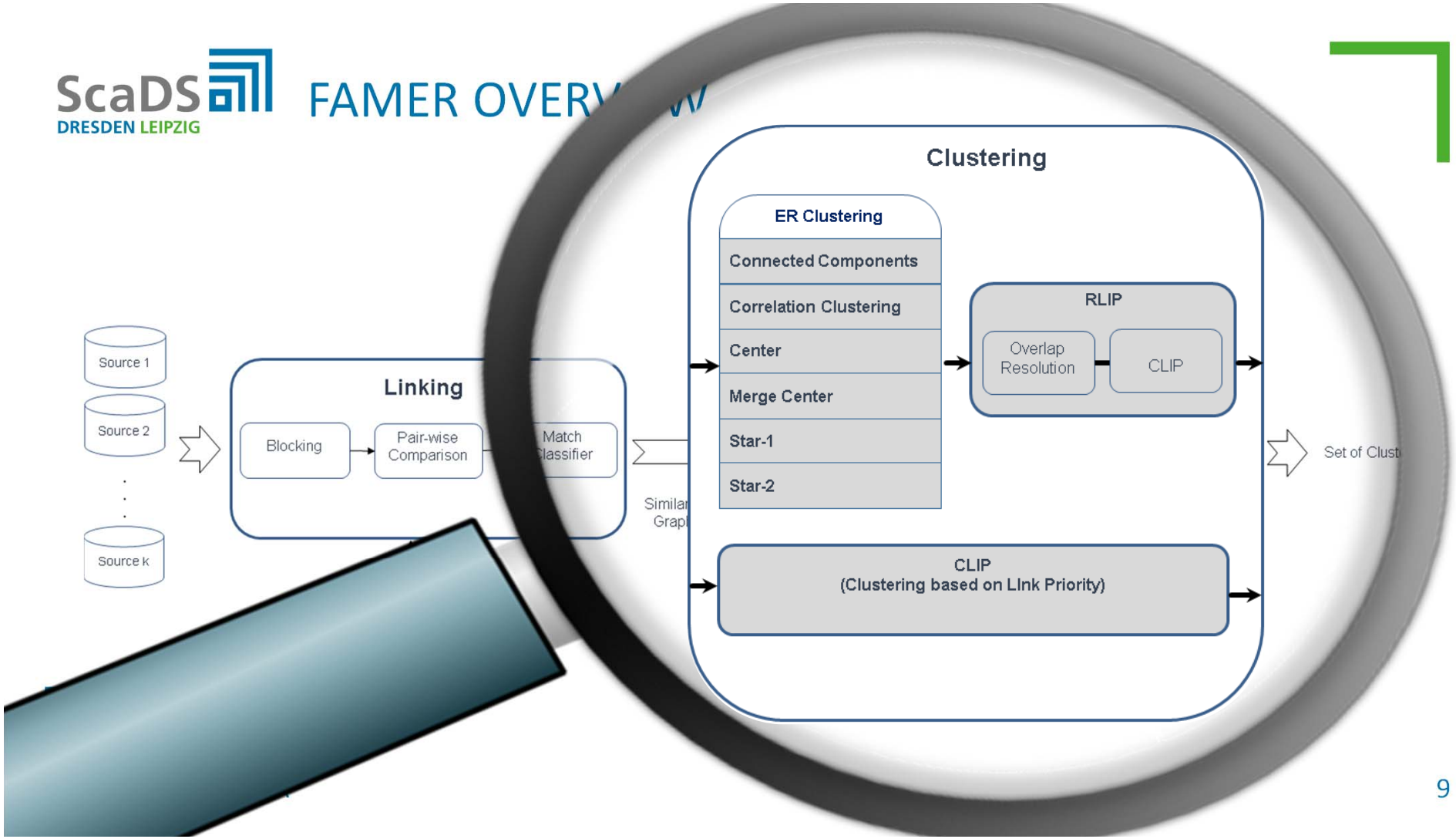




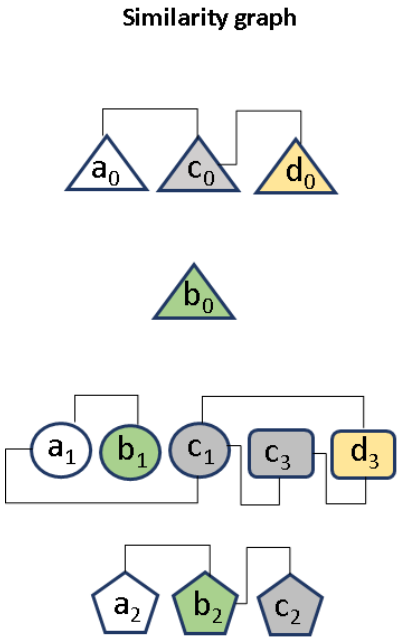


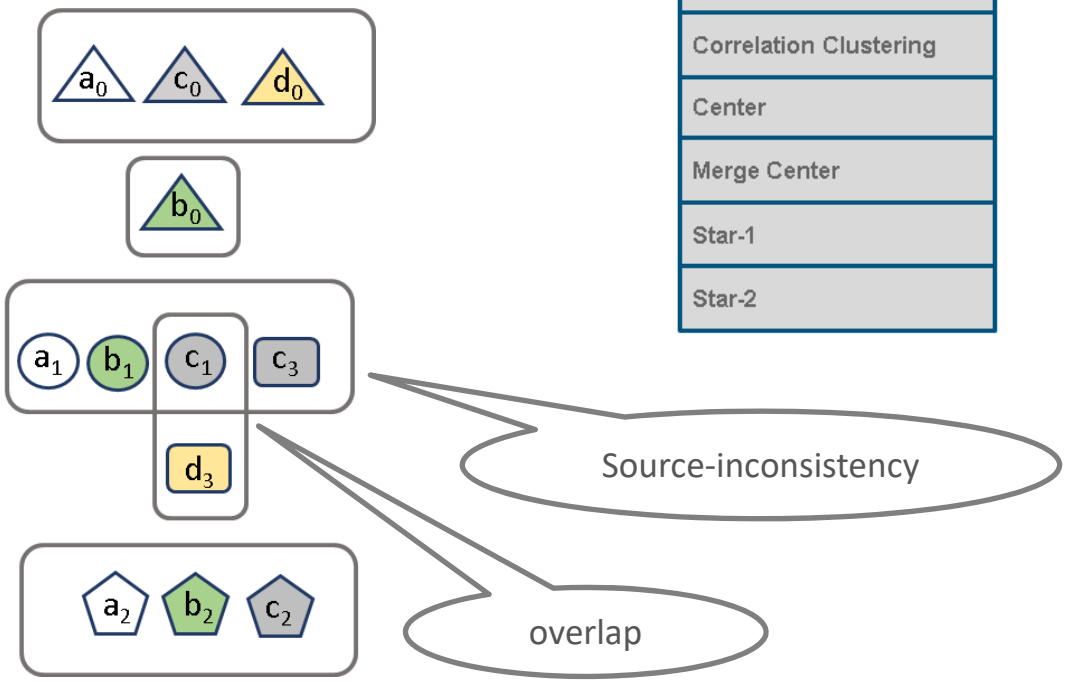
ScaDS DRESDEN LEIPZIG

FAMER OVERVIEW



FAMER CLUSTERING



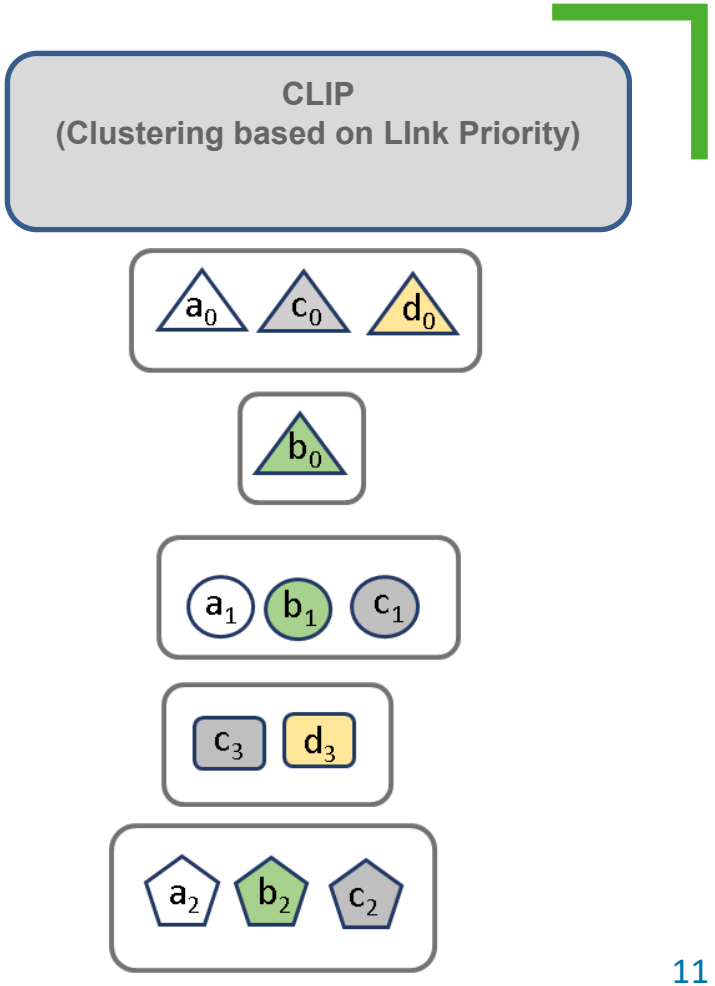


ER Clustering

- Connected Components
- Correlation Clustering
- Center
- Merge Center
- Star-1
- Star-2

FAMER CLUSTERING

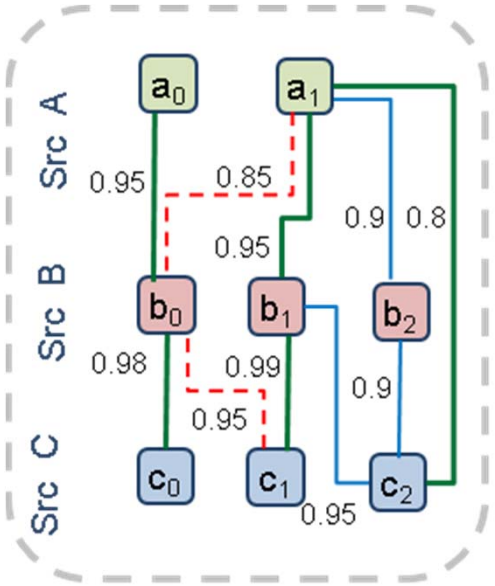
- Prioritize links based on
  - Link strength
    - Strong, Normal, Weak
  - Link degree
  - Similarity value
- produces
  - Source-consistent clusters
  - No overlap



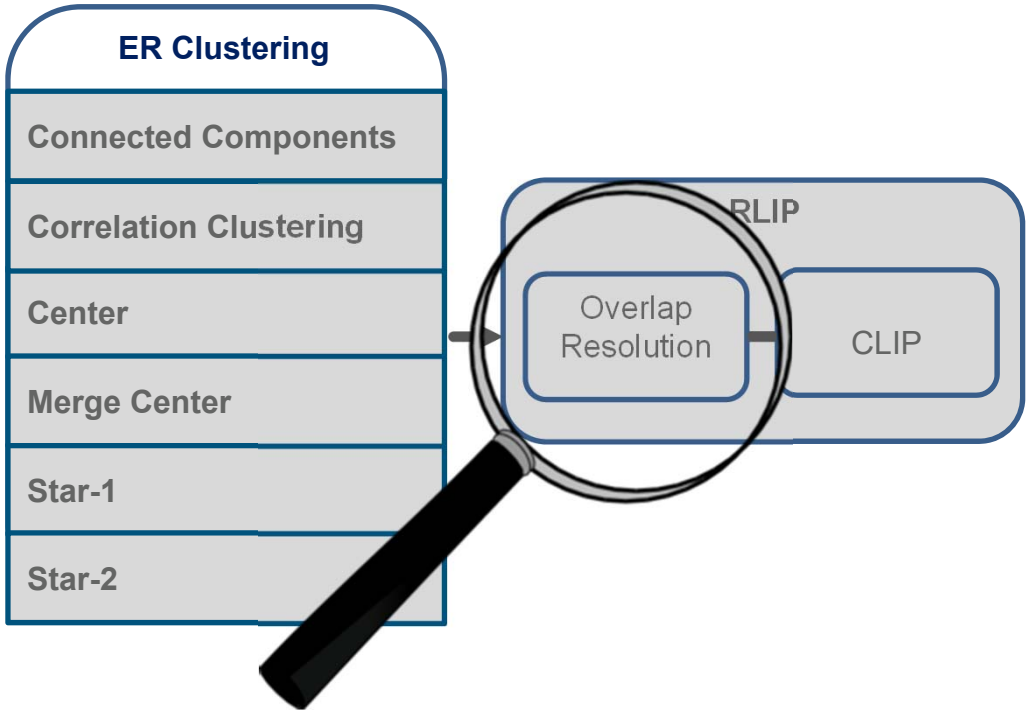
ScaDS DRESDEN LEIPZIG

LINK STRENGTH

– Link Strength
  - **Strong**
  - **Normal**
  - **Weak**

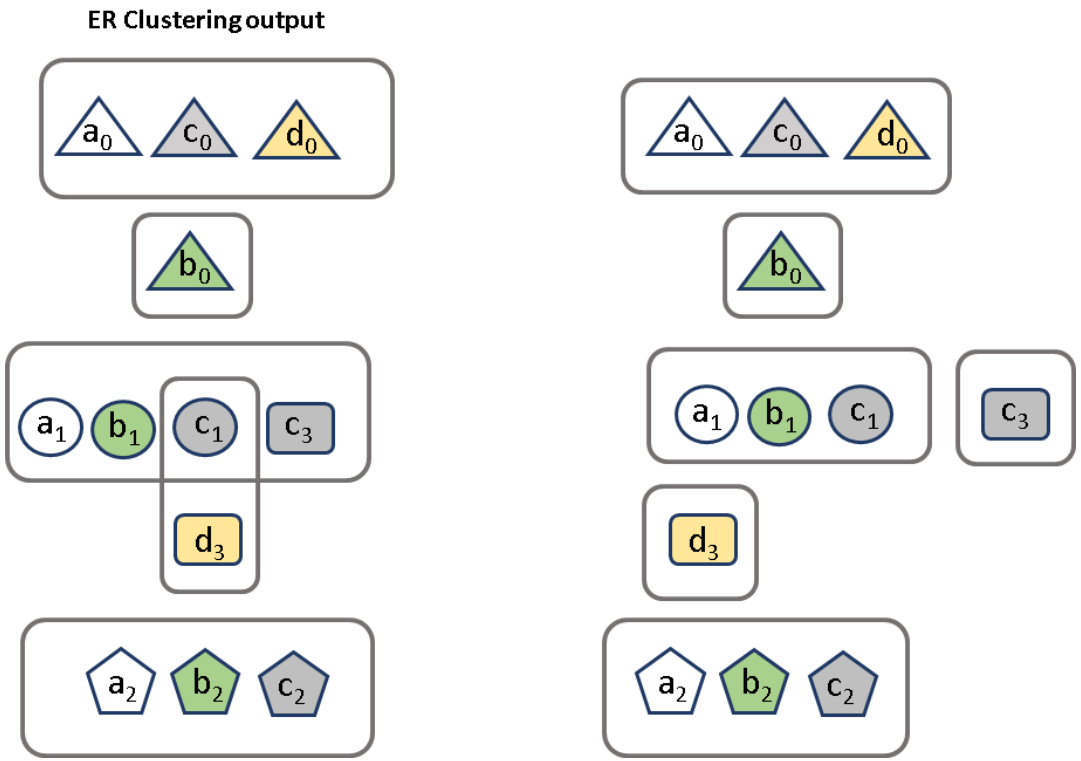


FAMER CLUSTERING



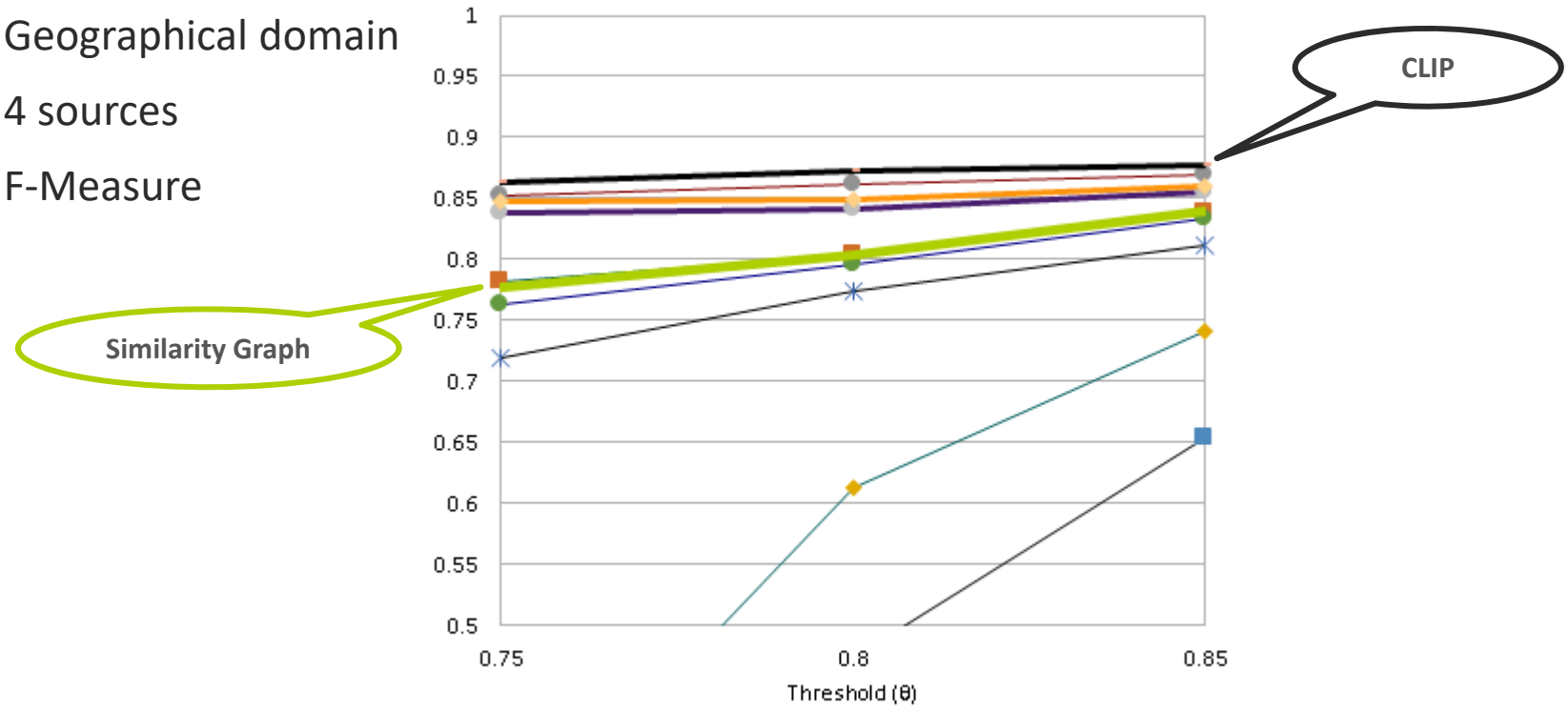
ScaDS DRESDEN LEIPZIG

FAMER CLUSTER REPAIR



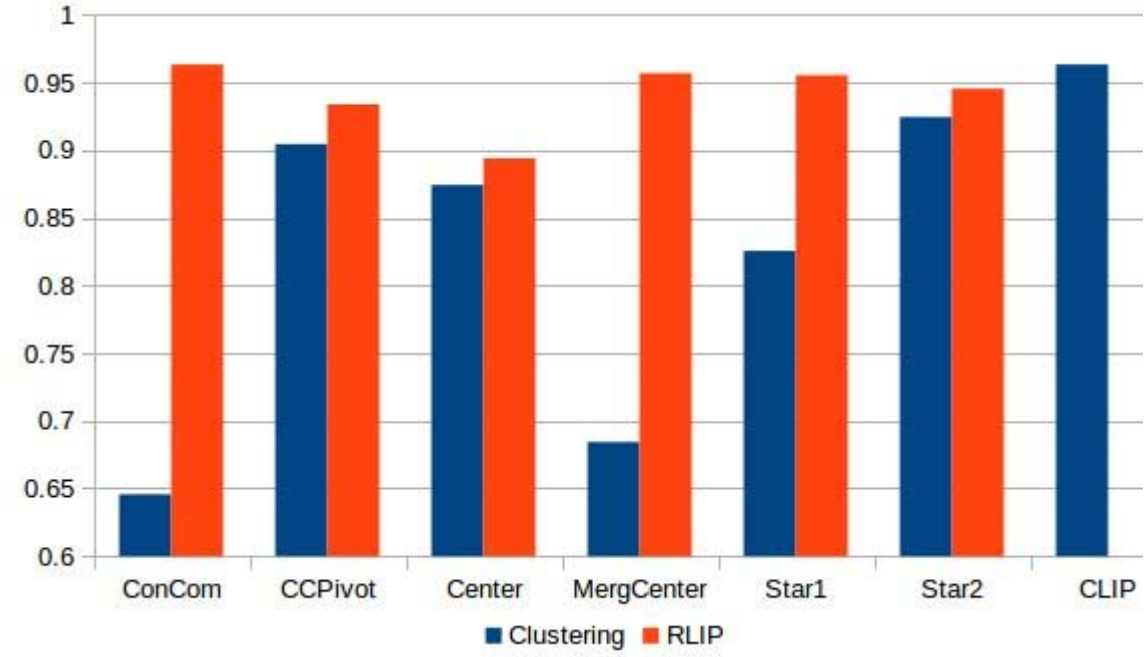
CLIP QUALITY

- Geographical domain
- 4 sources
- F-Measure



RLIP QUALITY

- Average F-Measure



ScaDS DRESDEN LEIPZIG

RUNTIMES

- 5 sources
- 5,000,000 entities
- Flink cluster of 16 workers
- CLIP
  - 69s (1 min) for 16 workers
  - 10 sources/
    10 million entities: 228 s

	#workers		
clustering	4	8	16
ConCom	51	57	55
CCPiv	1530	10008	688
Center	390	208	117
Merge Center	640	349	194
Star-1	288	149	85
Star-2	214	124	67
CLIP	**190**	**101**	**69**

ScaDS DRESDEN LEIPZIG

SUMMER SCHOOL 2018

- **Hackathon**: June 30th- July 1st 2018
- **Summer School**: July 2nd- July 6th 2018
- **Location:** University of Leipzig
- Bonus: LSWT attendants obtain early registration rates

www.scads.de/en/summerschool-2018

ScaDS DRESDEN LEIPZIG

SUMMER SCHOOL 2018

- Topics & Highlights
  - Deep Learning
  - Data Mining in Stream Data
  - Machine Learning for the Sciences
  - Apache SystemML
  - Secure Cloud Databases
  - Visual Analytics for Big Data
  - Graph Analytics on Spark
  - Web-scale information extraction
  - Big Data Integration
  - Big Data and HPC

www.scads.de/en/summerschool-2018