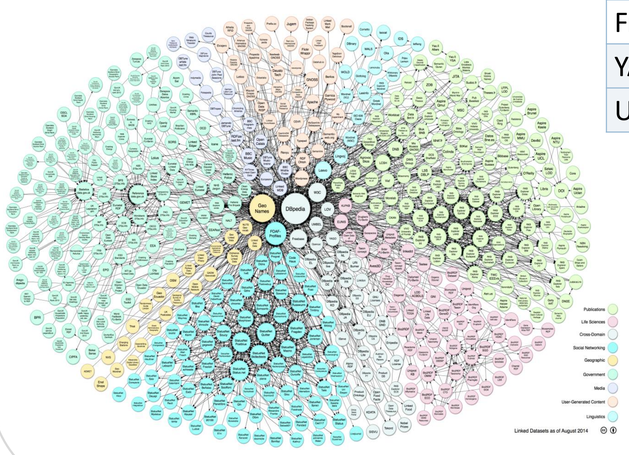
WebER: Resolving Entities in the Web

THE LOD CLOUD



KB examples	entities	entity types	properties
DBpedia(en)	4.58M	685	2,795
Freebase	46.3M	1.5K	4K
YAGO2	10M	350K	100
Uniprot	1.9B	123	112

Main characteristics of LOD:
- Important number of KBs (~ hundreds).
- Large number of entity types & properties (~ thousands).
- Massive volume of entities (~billions).

KBS OVERLAPPING

not identical descriptions, even if they have the same source

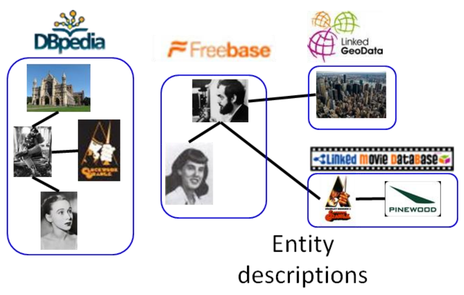
autonomous KBs

Highly and somehow similar descriptions of the same entity.
- 3-4 common tokens in central KBs.
- 1-2 common tokens in peripheral KBs.

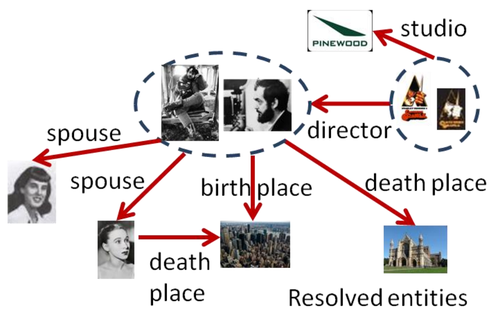
ENTITY RESOLUTION (ER): The problem of identifying descriptions of the same real-world entity

WEB-SCALE ER: Large-scale, multi-type, and cross-domain ER

WHY ER IS AN INTERESTING PROBLEM

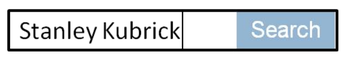


Entity Resolution

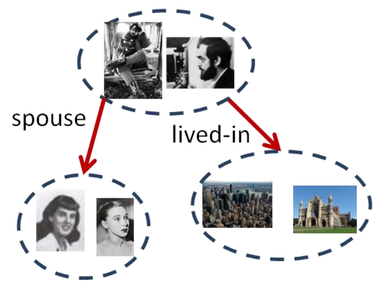


Entity-centric search

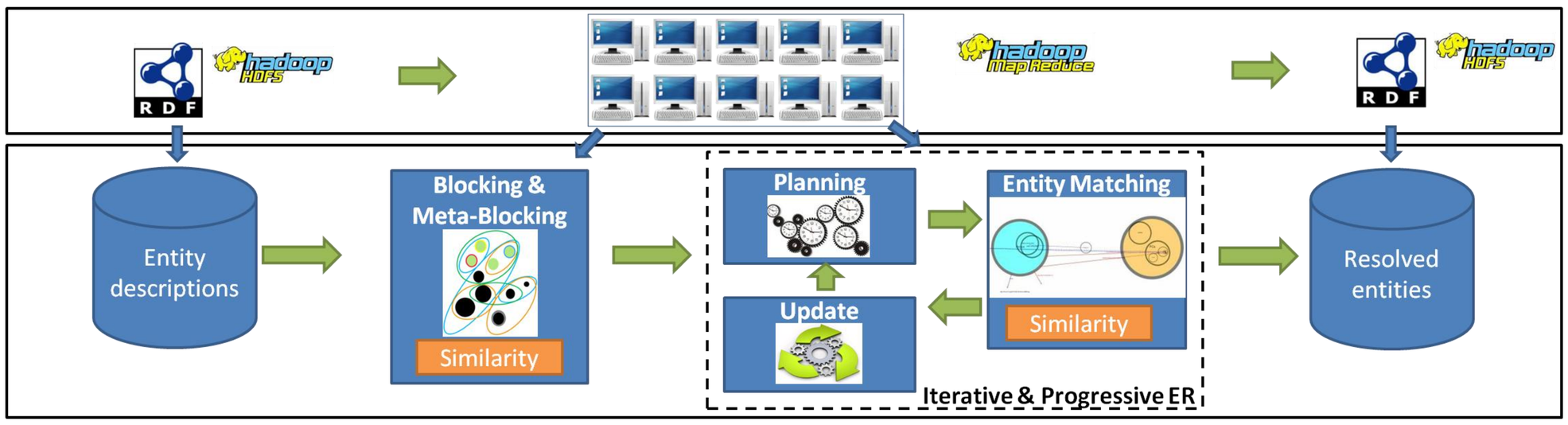
entity results



Entity-centric recommendation



THE WORKFLOW OF ER



BLOCKING

Avoid the quadratic complexity of ER:
- Split descriptions into blocks.
- Suggest the comparison of descriptions placed in a common block.

Examples: token blocking, PPJoin+, and Locality-sensitive Hashing (LSH).

[Christophides et al., 2015, Efthymiou et al., IEEE BigData 2015a]

META-BLOCKING

Blocking post-processing:
- Prune the redundant comparisons generated by blocking.
- Prune the least promising comparisons generated by blocking.

[Efthymiou et al., IEEE BigData 2015b]

ITERATIVE & PROGRESSIVE ER

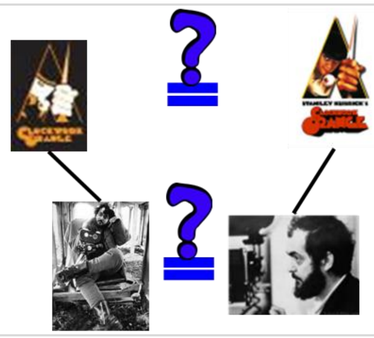
Iterative ER: identify new matches based on previous results and similarity propagation.

Progressive ER → Optimization: maximize benefit (number of matches|type of matches) for a given cost (#comparisons|disk access).
- Planning: select which pairs will be compared next and in what order.
- Update: propagate the results of matching, such that a new planning phase will promote the comparison of pairs influenced by the previous matches.

Exploit the parallel processing power of a computer cluster via Hadoop MapReduce

SIMILARITY

There is no single ideal similarity measure to identify all matches.
Indications for matches can be provided by similarity in:
- Content, i.e., the values of the descriptions (e.g., Jaccard)
- Structure, i.e., the attributes of the descriptions (e.g., SimRank)
- Neighborhoods, i.e., related descriptions (e.g., LINDA)



Example: Exploit the similarity of the descriptions of "A Clockwork Orange", to identify if the descriptions of Stanley Kubrick match.

[Christophides et al., 2015]

More details, source code and datasets available at: http://www.csd.uoc.gr/~vefthym/minoanER/

V. Christophides, V. Efthymiou, K. Stefanidis. Entity Resolution in the Web of Data. Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool, August 2015
V. Efthymiou, K. Stefanidis, V. Christophides. Big Data Entity Resolution: From Highly to Somehow Similar Entity Descriptions in the Web. In IEEE Big Data, 2015a
V. Efthymiou, G. Papadakis, G. Papastefanatos, K. Stefanidis, T. Palpanas. Parallel Meta-blocking: Realizing Scalable Entity Resolution over Large, Heterogeneous Data. In IEEE Big Data, 2015b

VASSILIS CHRISTOPHIDES[1,2], KOSTAS STEFANIDIS[3], VASILIS EFTHYMIOU[1,3]
[1]University of Crete, Computer Science Department, Heraklion, Greece.
[2]INRIA, Paris-Rocquencourt, France.
[3]Foundation for Research and Technology - Hellas (FORTH), Institute of Computer Science, Heraklion, Greece.