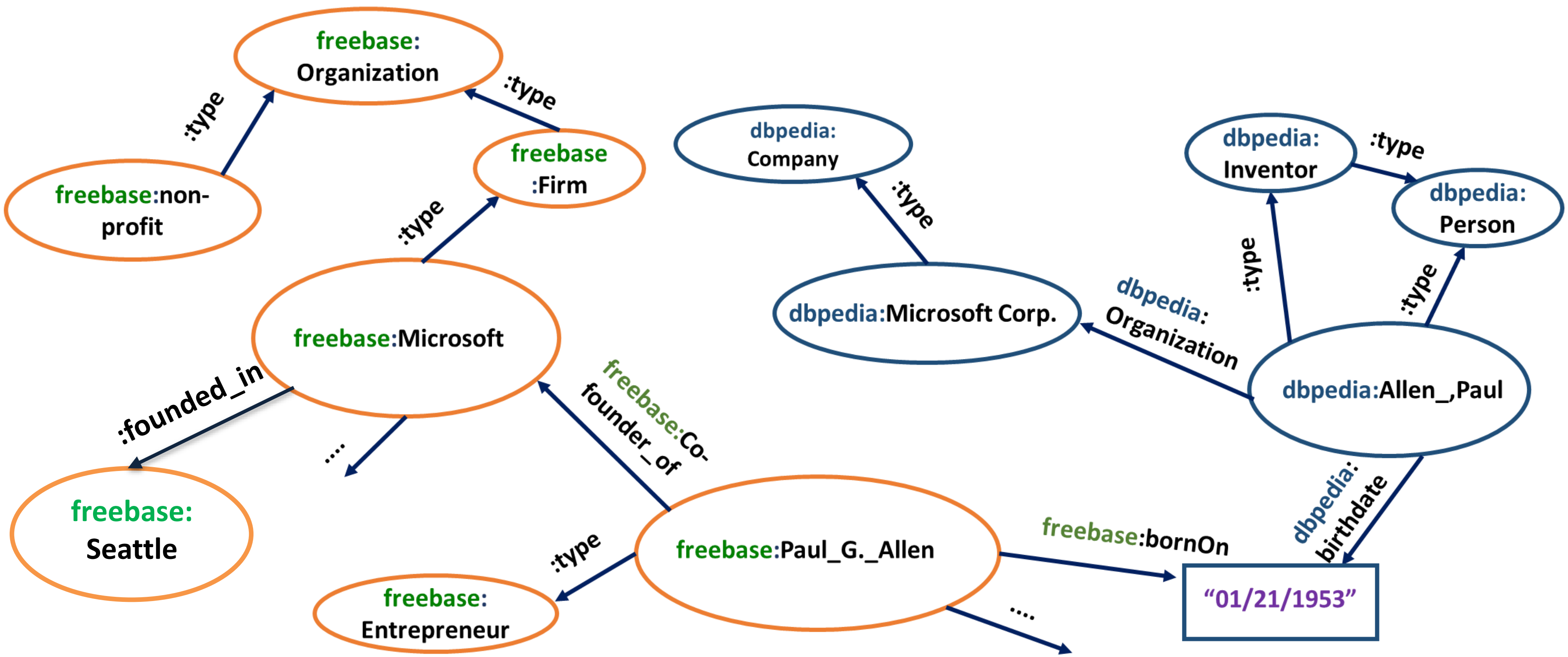
Knowledge Graph Completion

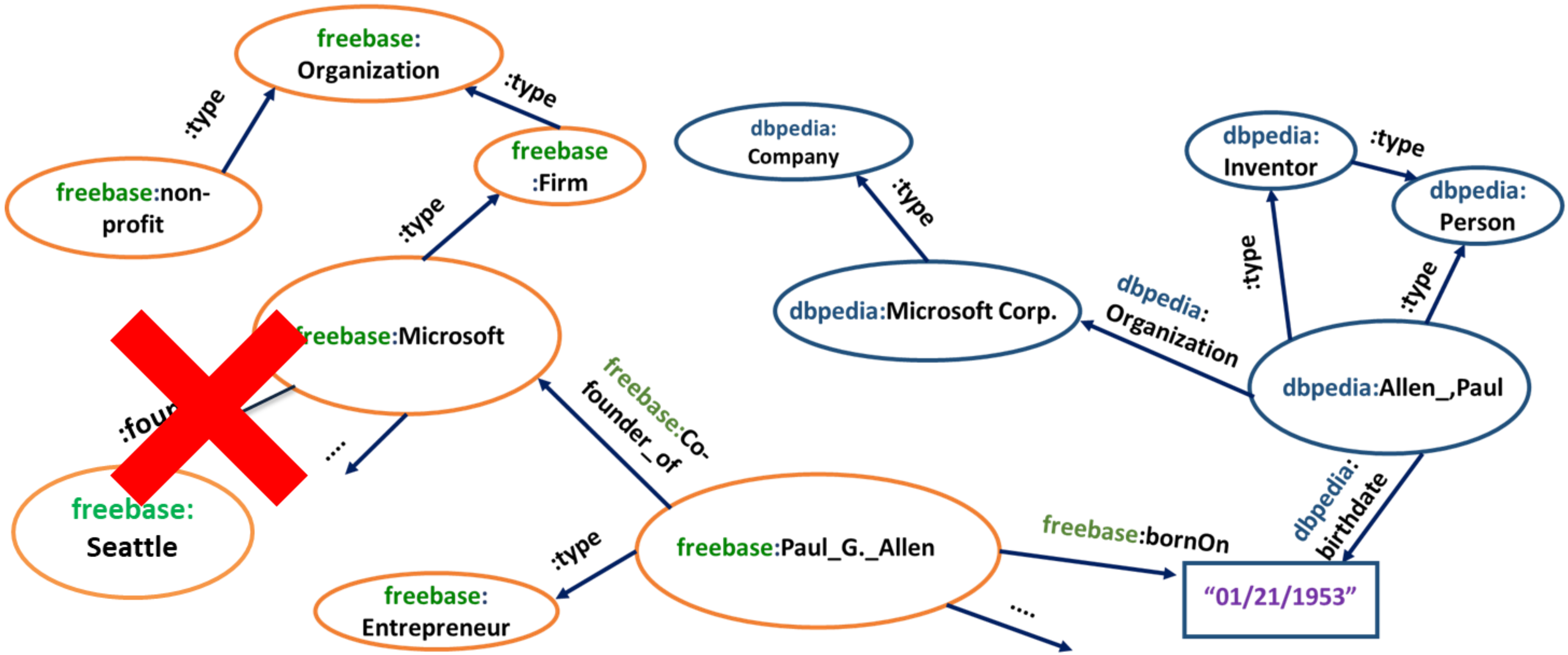
Introduction and motivation

We have our 'constructed' knowledge graph, now what?



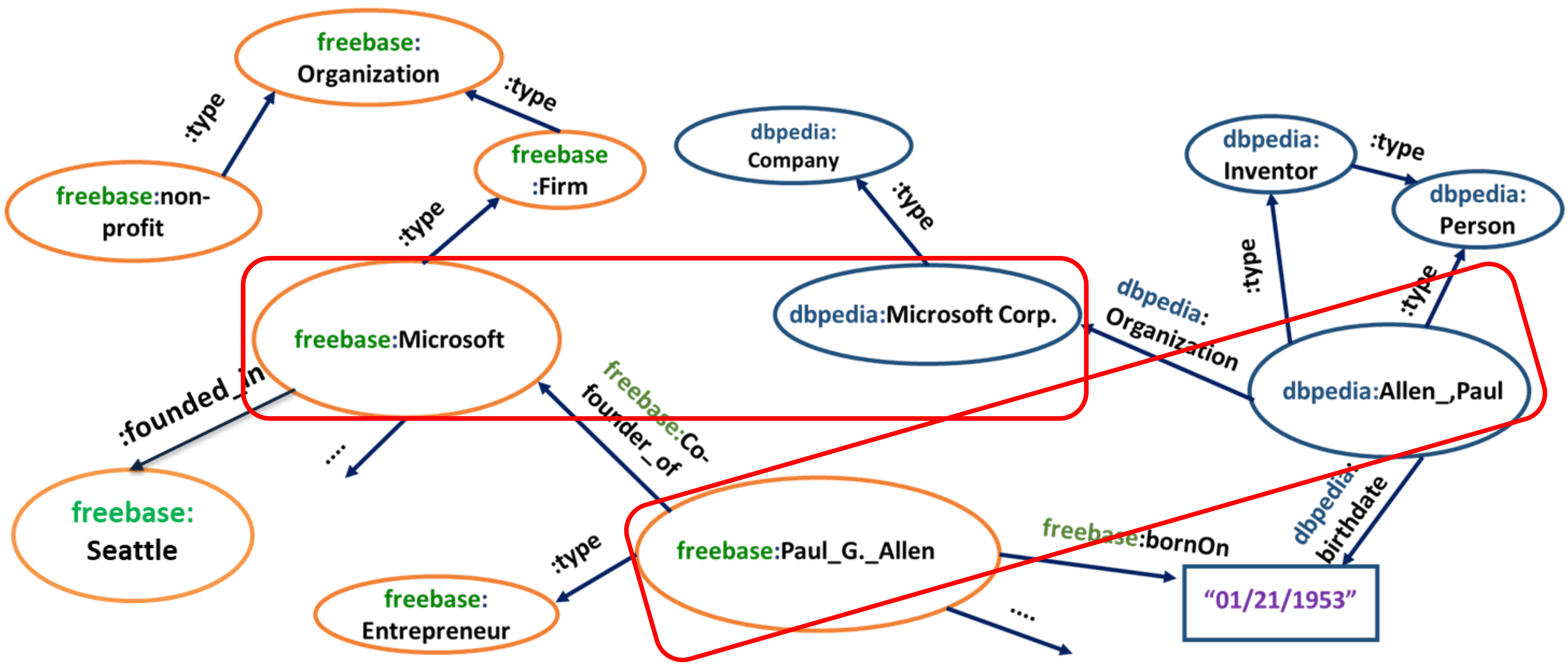
Introduction and motivation

Problem 1: Wrong/missing triples



Introduction and motivation

Problem 2: Many nodes refer to the same underlying entity



For Web extractions, noise is inevitable

- Thousands of web domains
- Many page formats
- Distracting & irrelevant content
- Purposeful obfuscation
- Poor grammar & spelling
- Tables

To reach its potential, a **constructed** KG must be **completed or identified**

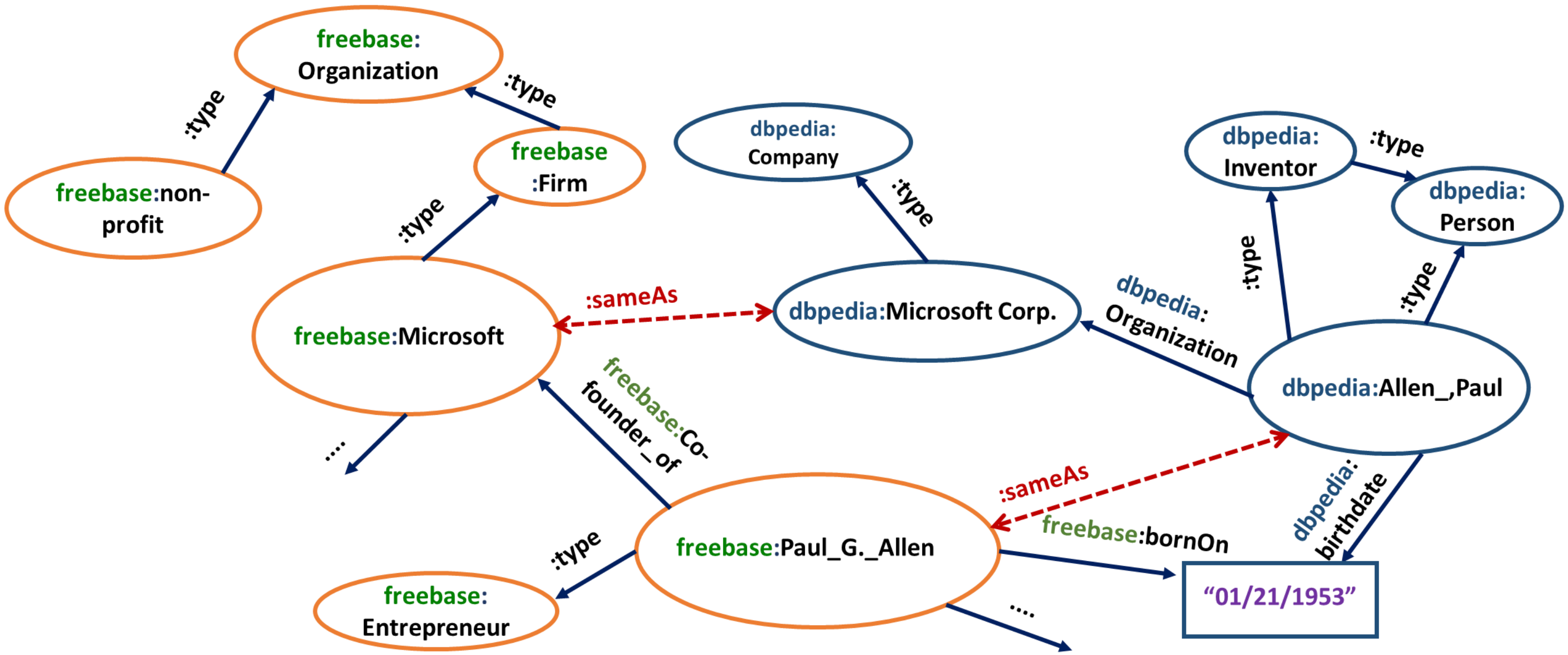
Noise Analysis

- Extractors found to offer a collective tradeoff between multiple dimensions
  - Noise is rarely ‘random’!

	Glossary	Regex	Landmark	CRF	NER
Easy to define	**4**	**2**	**4**	**4**	**4**
Site coverage	**All**	**All**	**Short Tail**	**All**	**All**
Precision	**2-3**	**3-4**	**4**	**2-3**	**3**
Recall	**3-4**	**2**	**1**	**2**	**1**

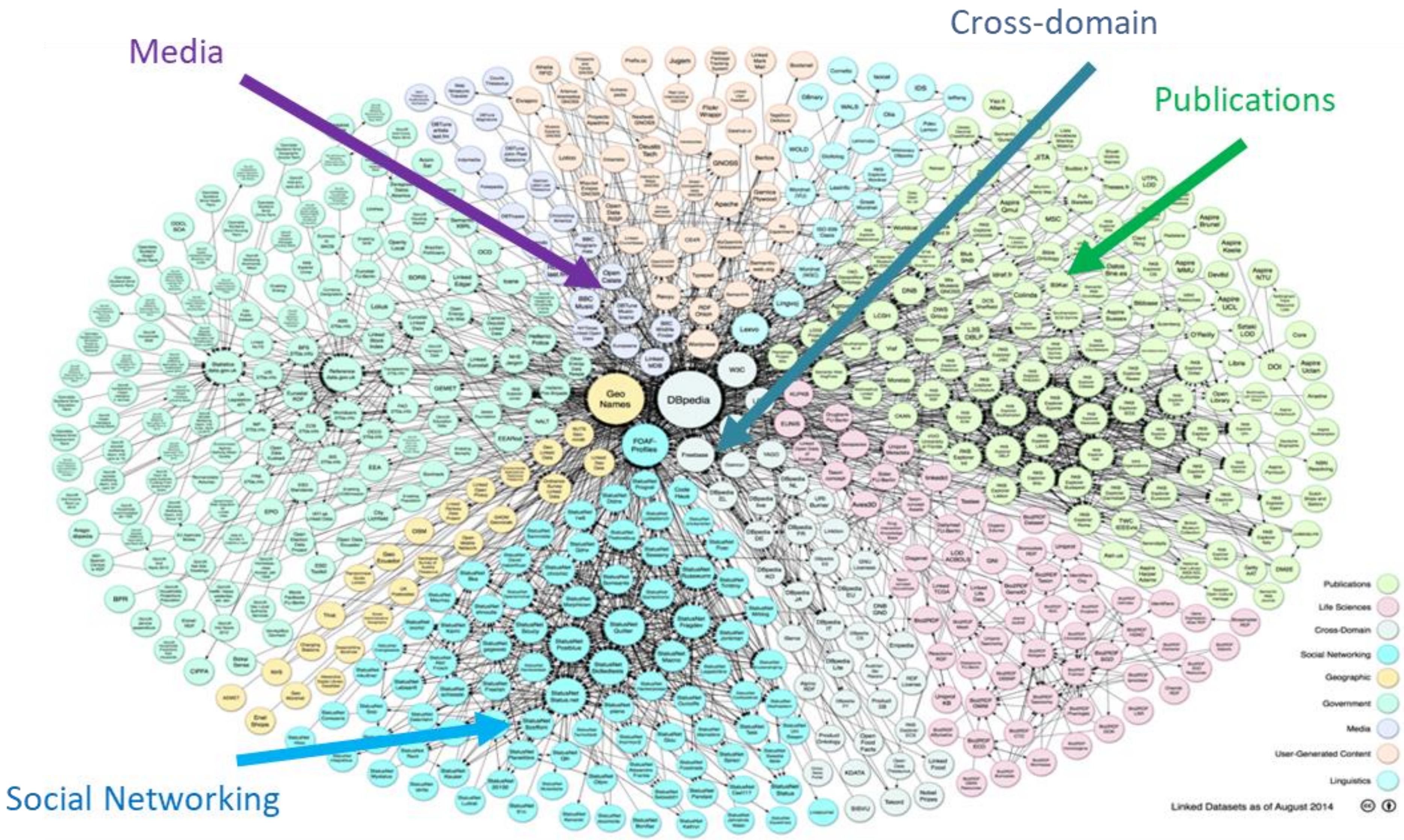
ENTITY RESOLUTION

Definitions and alternate names



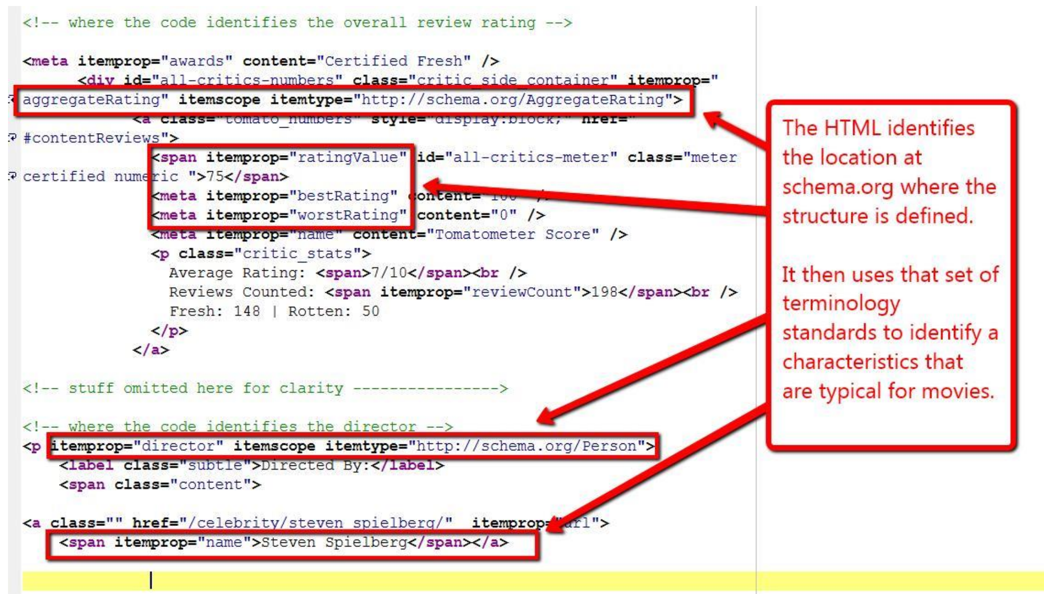
- Common sense:
  - Which entities refer to the same thing?
- Slightly more formal:
  - Which mentions (aka records, instances, nodes, surface strings…) refer to the same underlying entity?
- Rigorous mathematical/logical definition
  - Doesn't exist, or unknown! Just like other hard AI problems...
- Why try to solve the problem aka why is it a problem?

Applications: A Web of Linked 'Data'



Applications: Schema.org

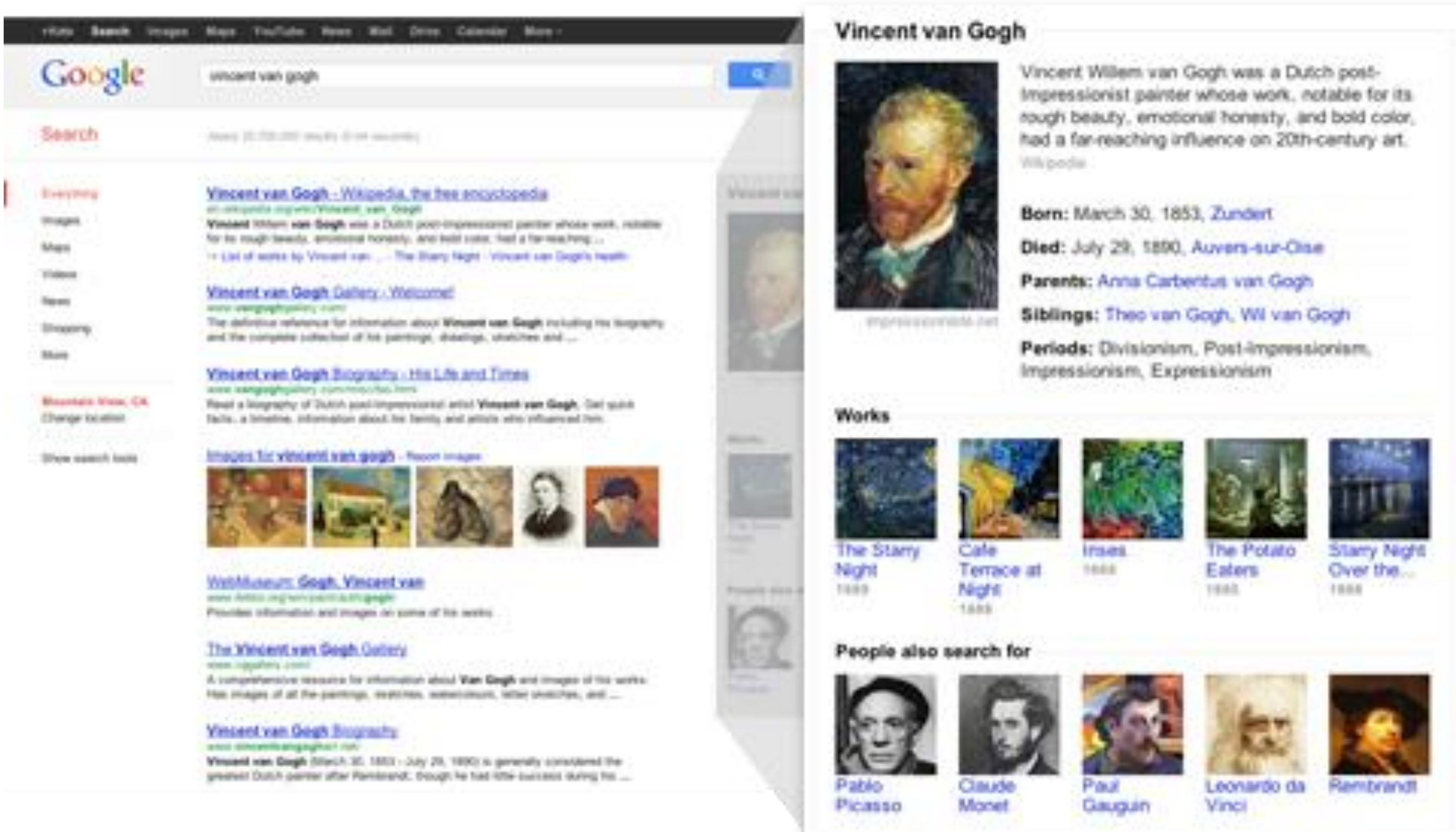
- **Schema.org** is an RDF ontology from which triples (with Web-dereferencable URIs) can be embedded in HTML pages



YAHOO! bing Yandex Google

http://schema.org/

Applications: Google Knowledge Graph



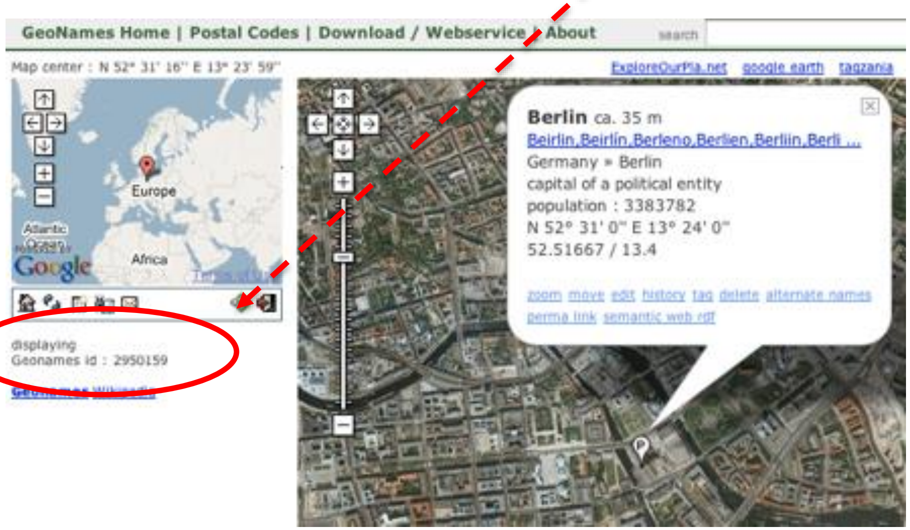
https://developers.google.com/knowledge-graph/

SUB-COMMUNITIES

Entity Linking/Canonicalization

- Berlin, California, the former name of Genevra, California
- Berlin, Connecticut
  - Berlin (Amtrak station), rail station in Berlin, Connecticut
- Berlin, Georgia
- Berlin, Illinois
- Berlin, Indiana, extinct town
- Berlin, Kansas
- Berlin, Kentucky
- Berlin, Maryland
- Berlin, Massachusetts
- Berlin, Michigan (disambiguation)
- Berlin, Nebraska, a former name of Otoe, Nebraska

- Name of an entity (such as a city or location) not enough to resolve ambiguity
- Use Geonames knowledge base to canonicalize entity using machine learning and text features



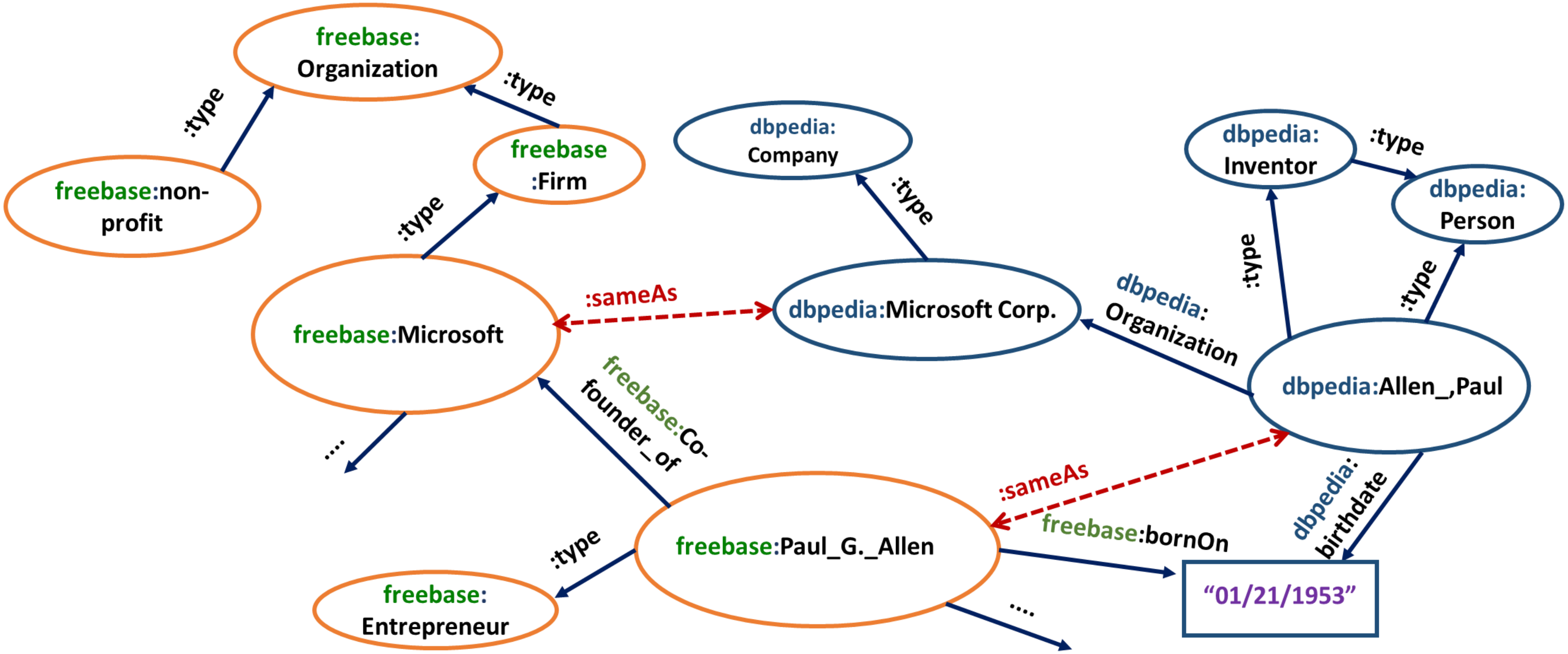
Co-reference Resolution

⊟ coreference (discourse)

Wikinews interviews President of the International Brotherhood of Magicians Wednesday , October 9 , 2013 October is National Magic Month in the United States . Wikinews spoke with William Evans , president of the International Brotherhood of Magicians , about the current state of magic and what its future looks like in the world of entertainment . For how long have you been involved in performing / studying magic ? Over 50 years . I am 61 now so I really started when I was about 10

Entity Resolution (what we'll be covering)

- Itself has many sub-communities and approaches
- Because of flexible representations (compared to databases or strict models like OWL), KG-ER systems tend to be community-agnostic

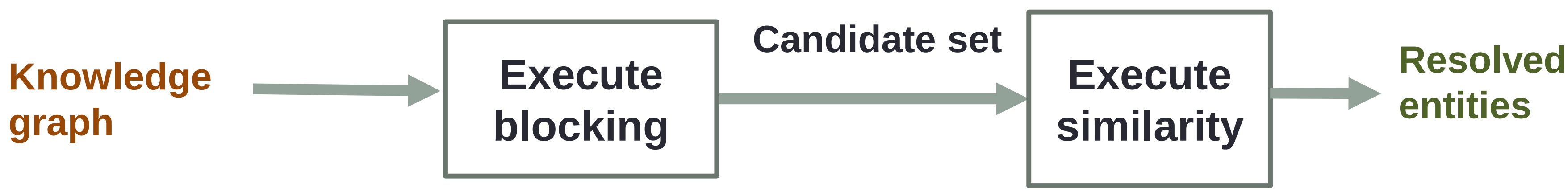


STANDARD ER ARCHITECTURE

Entity Resolution is fundamentally non-linear

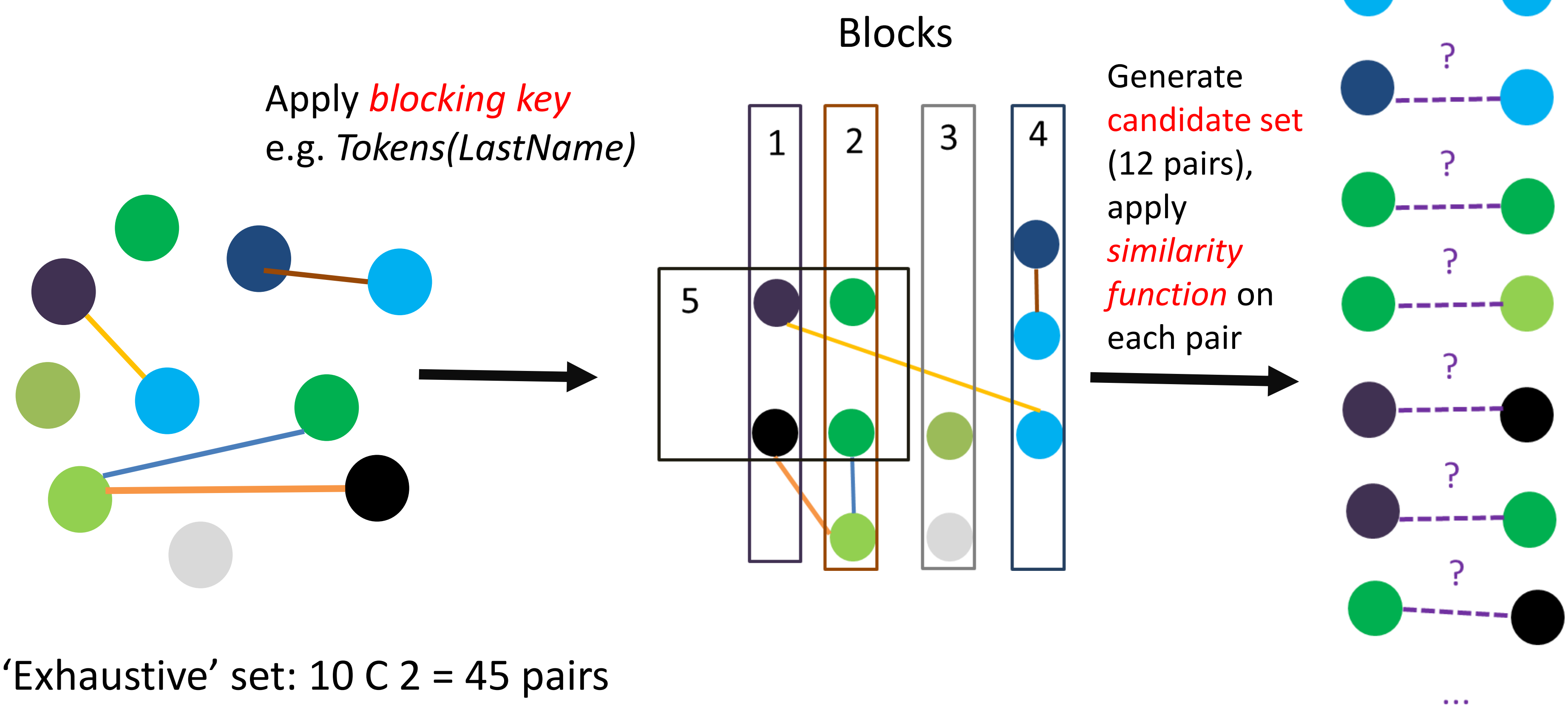
- Theoretically quadratic in the number of nodes, even if ‘resolution rule’ was known
- In practice, number of ‘duplicates’ tends to grow linearly, and duplicates overlap in non-trivial ways
- How to devise efficient algorithms?

50 years of research has agreed on a two-step solutions



Blocking

- Key idea is to use a cheap heuristic that efficiently clusters approximately similar entities into (possibly overlapping) blocks



Aside: some blocks have skewed size...

- Property of real-world data (zipf distribution, power laws...)
- How to address data skew?
  - Apply blocking methods with guarantees
  - May lose some recall in the process

Example

Sorted Neighborhood aka merge-purge:
--use blocking key as 'sorting' key
--slide a window of constant size (w) over sorted nodes
--only pairs of nodes within window are paired, added to candidate set

ID	First Name	Last Name	Zip	BKV
1	Cathy	Ransom	77111	CR7
2	Catherine	Ridley	77093	CR7
3	Cathy	Ridley	77093	CR7
4	John	Rogers	78751	JR7
5	J.	Rogers	78732	JR7
6	John	Ridley	77093	JR7
7	John	Ridley Sr.	77093	JRS7

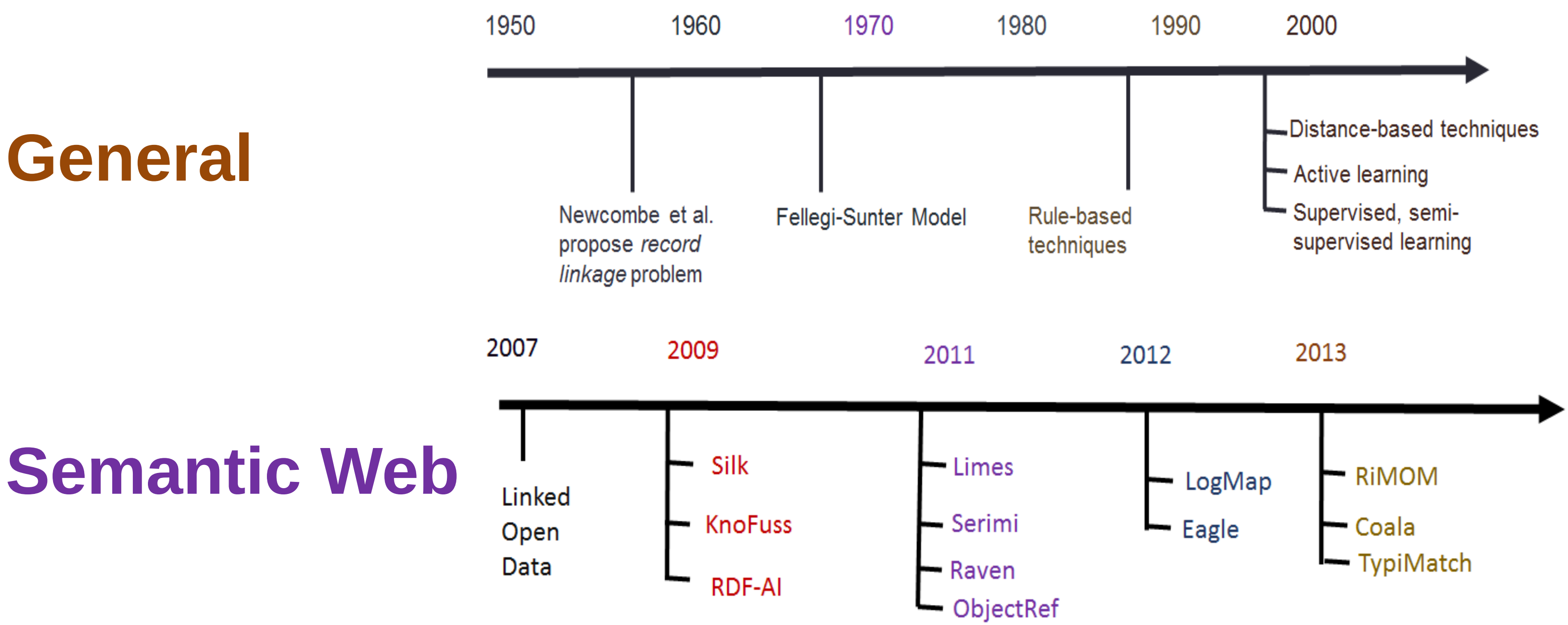
Final Candidate Set ($w = 3$):

{(1,2), (2,3), (1,3), (2,4), (3,4), (3,5), (4,5), (4,6), (5,6), (5,7), (6,7)}

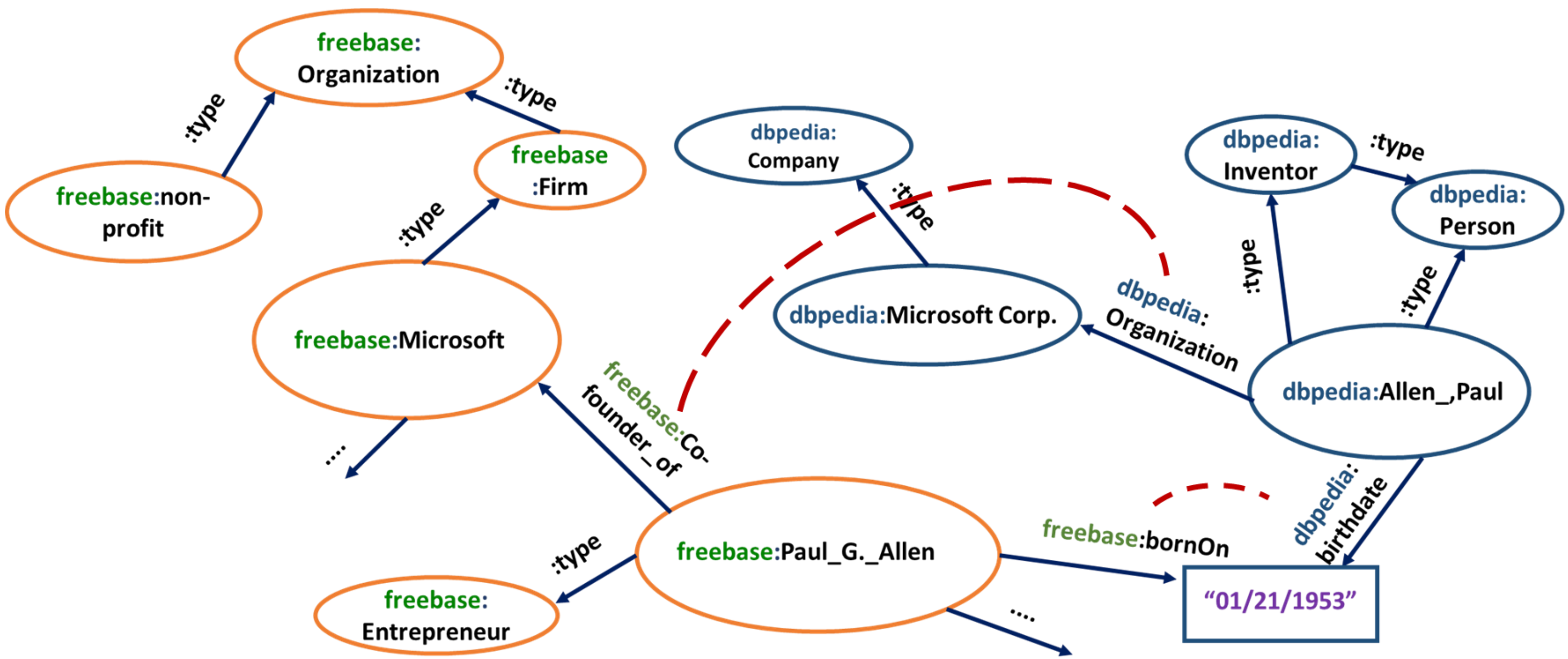
Other methods: block purging, canopies...

Similarity/link specification

- Over 50 years of research on what makes for a good 'similarity' function
- Current approach: apply 'typical' machine learning workflow to candidate set
- Important to remember that features are extracted from 'mention pairs'...leads to non-trivial alignment issues
  - Some form of schema-matching almost always attempted in practical systems
  - Some (but not much) work on so-called schema-free similarity



Aside: why schema matching?



Feature engineering

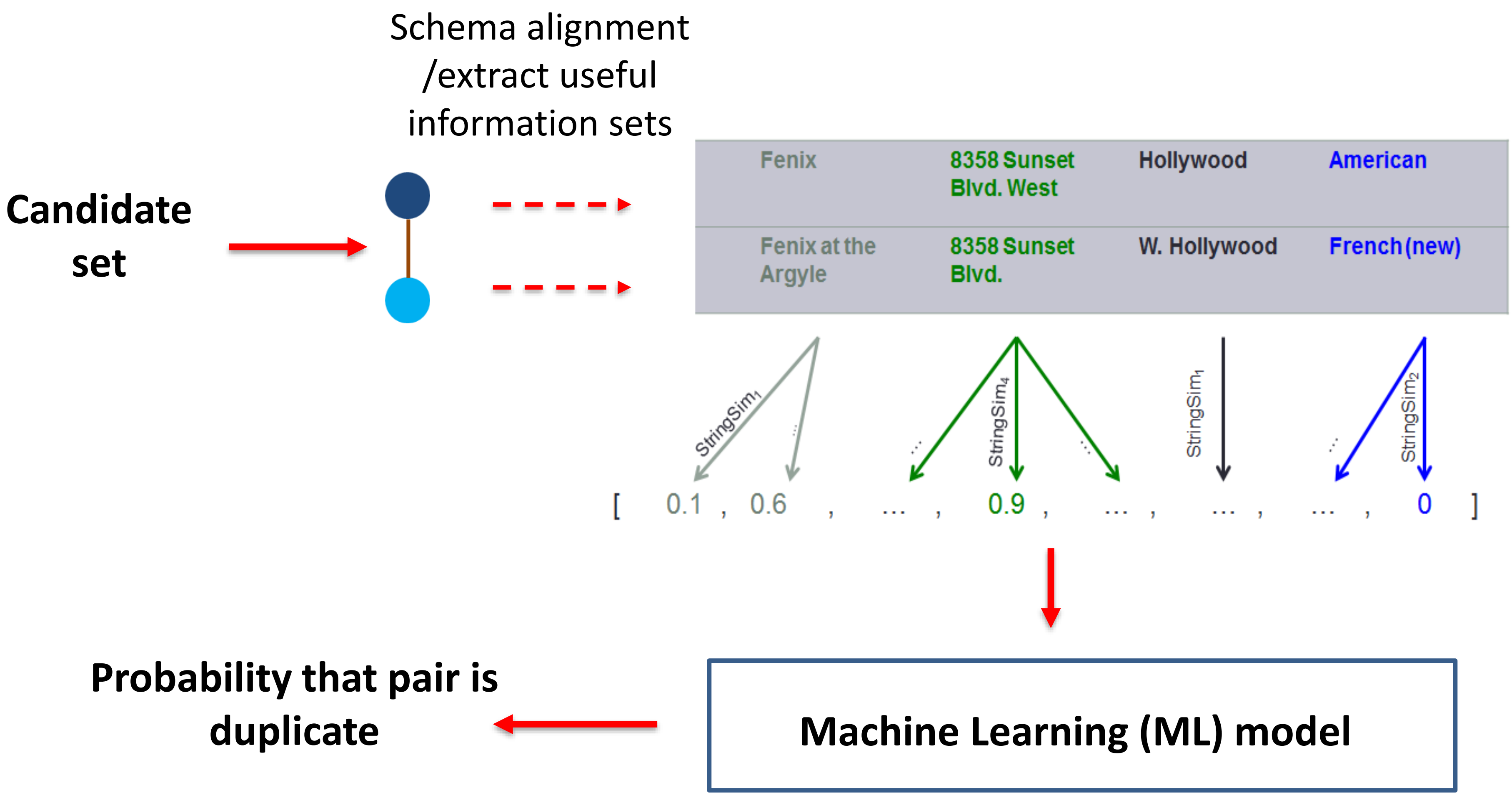
Library of features

- Character based
  - Edit Distance
  - Affine Gap
  - Smith-Waterman
  - Jaro
  - Q-gram
- Token based
  - Monge Elkan
  - TF-IDF
    - Soft
    - Q-gram
  - Jaccard
- Phonetic based
  - Soundex
  - NYSIIS
  - ONCA
  - Metaphone
  - Double Metaphone
- …

Open question: how much can representation learning contribute to Entity Resolution?

Similarity: putting it together

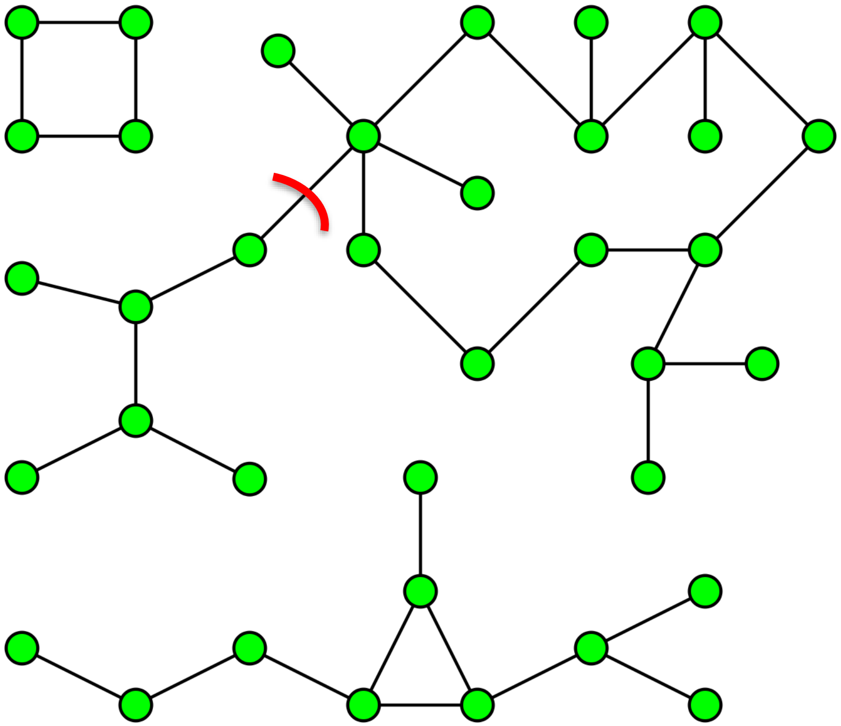
- ML model can be supervised, semi-supervised or unsupervised



OUTPUT REPRESENTATION AND HANDLING

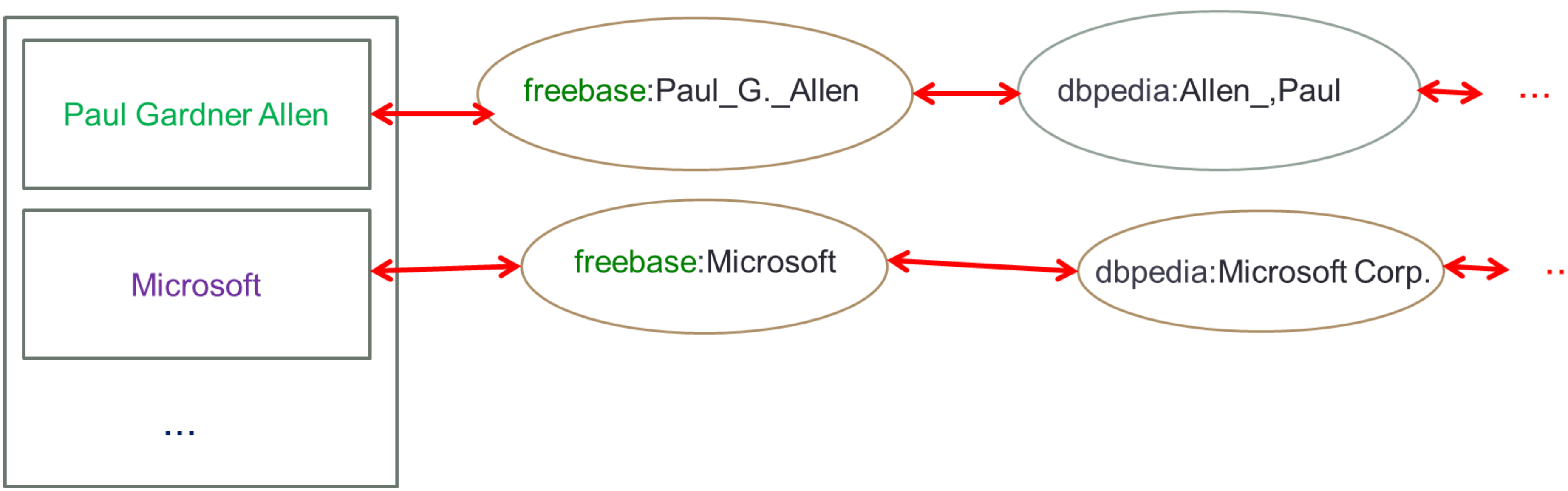
From links to clusters

- For perfect links, transitive closure/connected components works
- With imperfect links, effect can be severe
  - One weak link is all it takes to form a giant component
  - Not uncommon in the real world
- More robust clustering methods have to be applied
  - Community detection literature
  - Spectral clustering
  - Many more!
- Some recent work has proposed to explore ER as a micro-clustering problem



From (possibly noisy) clusters to…???

- Surprisingly under-studied problem!
- Should the entities be fused into a single entity? How?
  - Entity linking has a conceptually elegant solution to this problem…
  - …but how to deal with NIL clusters?
- Semantic Web approach
  - Represent individual links as KG triples and leave it at that
  - Entity Name Systems for advanced search/reasoning



BEYOND ENTITY RESOLUTION

By itself, generic ER is unlikely to be enough to sufficiently boost KG quality

- Other things explored in the literature:
  - Domain knowledge
    - Collective ER methods have tried to exploit these systematically
  - Multi-type Entity Resolution
    - Extremely useful for knowledge graphs, lots more work to be done
  - Entity Resolution+Ontologies+IE Confidences:
    - Probabilistic Graphical Models like Probabilistic Soft Logic
  - Knowledge graph embeddings
    - Useful for link prediction and triples classification
    - Recall the Microsoft-founded_in-Seattle example earlier

Knowledge graph embeddings/representation learning

- Useful for link prediction/missing relationships/triples classification
- Not clear if it is really better than PSL on noisy KGs
- Not clear how to combine KGEs with domain engineering

Model	#Parameters	# Operations (Time complexity)
Unstructured (Bordes et al. 2012; 2014)	$O(N_e m)$	$O(N_t)$
SE (Bordes et al. 2011)	$O(N_e m + 2N_r n^2)(m = n)$	$O(2m^2 N_t)$
SME(linear) (Bordes et al. 2012; 2014)	$O(N_e m + N_r n + 4mk + 4k)(m = n)$	$O(4mkN_t)$
SME (bilinear) (Bordes et al. 2012; 2014)	$O(N_e m + N_r n + 4mks + 4k)(m = n)$	$O(4mksN_t)$
LFM (Jenatton et al. 2012; Sutskever et al. 2009)	$O(N_e m + N_r n^2)(m = n)$	$O((m^2 + m)N_t)$
SLM (Socher et al. 2013)	$O(N_e m + N_r(2k + 2nk))(m = n)$	$O((2mk + k)N_t)$
NTN (Socher et al. 2013)	$O(N_e m + N_r(n^2 s + 2ns + 2s))(m = n)$	$O(((m^2 + m)s + 2mk + k)N_t)$
TransE (Bordes et al. 2013)	$O(N_e m + N_r n)(m = n)$	$O(N_t)$
TransH (Wang et al. 2014)	$O(N_e m + 2N_r n)(m = n)$	$O(2mN_t)$
TransR (Lin et al. 2015)	$O(N_e m + N_r(m + 1)n)$	$O(2mnN_t)$
CTransR (Lin et al. 2015)	$O(N_e m + N_r(m + d)n)$	$O(2mnN_t)$

Concluding notes

- Entity Resolution (ER) is a **hard problem for machines**, may be AI complete
  - It's 'easy' for us because we're so good at it
  - Not clear what will achieve the next breakthrough in ER
- Essential to attempt a solution if KGs are semi-automatically constructed from Web data
  - Quality doesn't have to be perfect, as we showed earlier with KG search
- Wealth of solutions but can be broken down into standard components
  - Blocking, to make ER efficient
  - Similarity, to make ER automatic/adaptive
- Many open questions, especially in relation to new ML models
- More broadly, lots of opportunities for KG completion

Bibliography

- Alani, H., Kim, S., Millard, D. E., Weal, M. J., Hall, W., Lewis, P. H., & Shadbolt, N. R. (2003). Automatic ontology-based knowledge extraction from web documents. *Intelligent Systems, IEEE*, *18*(1), 14-21.
- Allemang, D., & Hendler, J. (2011). *Semantic web for the working ontologist: effective model*
- Angles, R., & Gutierrez, C. (2005). Querying RDF data from a graph database perspective. In *The Semantic Web: Research and Applications*(pp. 346-360). Springer Berlin Heidelberg.
- Araujo, S., Hidders, J., Schwabe, D., & De Vries, A. P. (2011). Serimi-resource description similarity, rdf instance matching and interlinking. *arXiv preprint arXiv:1107.1104*.
- Arenas, M., Díaz, G., Fokoue, A., Kementsietsidis, A., & Srinivas, K. (2014). A principled approach to bridging the gap between graph data and their schemas. *Proceedings of the VLDB Endowment*, *7*(8), 601-612.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... & Harris, M. A. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics*, *25*(1), 25-29.
- Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). *Dbpedia: A nucleus for a web of open data* (pp. 722-735). Springer Berlin Heidelberg.
- Baxter, J. (1998). Theoretical models of learning to learn. In Learning to learn (pp. 71-94). Springer US.
- Baxter, R., Christen, P., & Churches, T. (2003, August). A comparison of fast blocking methods for record linkage. In ACM SIGKDD (Vol. 3, pp. 25-27).
- Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463). New York: ACM press.
- Bellahsene, Z., Bonifati, A., & Rahm, E. (2011). *Schema matching and mapping* (Vol. 20). Heidelberg (DE): Springer.
- Benjelloun, O., Garcia-Molina, H., Gong, H., Kawai, H., Larson, T. E., Menestrina, D., & Thavisomboon, S. (2007, June). D-swoosh: A family of algorithms for generic, distributed entity resolution. In *Distributed Computing Systems, 2007. ICDCS'07. 27th International Conference on* (pp. 37-37). IEEE.
- Benjelloun, O., Garcia-Molina, H., Menestrina, D., Su, Q., Whang, S. E., & Widom, J. (2009). Swoosh: a generic approach to entity resolution. The VLDB Journal—The International Journal on Very Large Data Bases, 18(1), 255-276.
- Bhattacharya, I., & Getoor, L. (2006, April). A Latent Dirichlet Model for Unsupervised Entity Resolution. In *SDM* (Vol. 5, No. 7, p. 59).
- Bilenko, M., & Mooney, R. J. (2003, August). Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 39-48). ACM.
- Bilenko, M., Kamath, B., & Mooney, R. J. (2006, December). Adaptive blocking: Learning to scale up record linkage. In *Data Mining, 2006. ICDM'06. Sixth International Conference on* (pp. 87-96). IEEE.
- Bilke, A., & Naumann, F. (2005, April). Schema matching using duplicates. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on* (pp. 69-80). IEEE.
- Bizer, C. (2009). The emerging web of linked data. *Intelligent Systems, IEEE*, *24*(5), 87-92.
- Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. Semantic Services, Interoperability and Web Applications: Emerging Concepts, 205-227.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993-1022.
- Borthakur, D. (2007). The hadoop distributed file system: Architecture and design. *Hadoop Project Website*, *11*(2007), 21.
- Bouquet, P., & Molinari, A. (2013). A global entity name system (ens) for data ecosystems. *Proceedings of the VLDB Endowment*, *6*(11), 1182-1183.

- Cao, Y., Chen, Z., Zhu, J., Yue, P., Lin, C. Y., & Yu, Y. (2011, July). Leveraging unlabeled data to scale blocking for record linkage. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence* (Vol. 22, No. 3, p. 2211).
- Carr, R. D., Doddi, S., Konjevod, G., & Marathe, M. V. (2000, January). On the red-blue set cover problem. In *SODA* (pp. 345-353).
- Carroll, J. J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., & Wilkinson, K. (2004, May). Jena: implementing the semantic web recommendations. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters* (pp. 74-83). ACM.
- Chakrabarti, K., Chaudhuri, S., Cheng, T., & Xin, D. (2012, August). A framework for robust discovery of entity synonyms. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1384-1392). ACM.
- Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3), 27.
- Chaudhuri, S., Ganti, V., & Motwani, R. (2005, April). Robust identification of fuzzy duplicates. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on* (pp. 865-876). IEEE.
- Chen, P. P. S. (1976). The entity-relationship model—toward a unified view of data. *ACM Transactions on Database Systems (TODS)*, *1*(1), 9-36.
- Christen, P. (2008, August). Febrl-: an open source data cleaning, deduplication and record linkage system with a graphical user interface. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1065-1068). ACM.
- Christen, P. (2008, August). Automatic record linkage using seeded nearest neighbour and support vector machine classification. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 151-159). ACM.
- Christen, P. (2012). *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media.
- Christen, P. (2012). A survey of indexing techniques for scalable record linkage and deduplication. *Knowledge and Data Engineering, IEEE Transactions on*,*24*(9), 1537-1555.
- Chvatal, V. (1979). A greedy heuristic for the set-covering problem. Mathematics of operations research, 4(3), 233-235.
- Codd, E. F. (1970). A relational model of data for large shared data banks.*Communications of the ACM*, *13*(6), 377-387.
- Cohen, W. W. (2000). Data integration using similarity joins and a word-based information representation language. *ACM Transactions on Information Systems (TOIS)*, *18*(3), 288-321.
- Cohen, W. W., Kautz, H., & McAllester, D. (2000, August). Hardening soft information sources. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 255-259). ACM.
- Cucerzan, S. (2007, June). Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *EMNLP-CoNLL* (Vol. 7, pp. 708-716).
- Das Sarma, A., He, Y., & Chaudhuri, S. (2014). Clusterjoin: a similarity joins framework using map-reduce. *Proceedings of the VLDB Endowment*, *7*(12), 1059-1070.
- Datar, M., Immorlica, N., Indyk, P., & Mirrokni, V. S. (2004, June). Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry* (pp. 253-262). ACM.
- Date, C. J., & Darwen, H. (1993). *A guide to the SQL Standard: a user's guide to the standard relational language SQL* (Vol. 55822). Addison-Wesley Longman.
- Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38.
- Dietze, S., Sanchez-Alonso, S., Ebner, H., Qing Yu, H., Giordano, D., Marenzi, I., & Pereira Nunes, B. (2013). Interlinking educational resources and the web of data: A survey of challenges and approaches. *Program*, *47*(1), 60-91.
- Doan, A., Halevy, A., & Ives, Z. (2012). *Principles of data integration*. Elsevier.
- Dong, X. L., & Srivastava, D. (2013, April). Big data integration. In *Data Engineering (ICDE), 2013 IEEE 29th International Conference on* (pp. 1245-1248). IEEE.
- Dredze, M., McNamee, P., Rao, D., Gerber, A., & Finin, T. (2010, August). Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 277-285). Association for Computational Linguistics.

- Duan, S., Fokoue, A., Hassanzadeh, O., Kementsietsidis, A., Srinivas, K., & Ward, M. J. (2012). Instance-based matching of large ontologies using locality-sensitive hashing. In *The Semantic Web–ISWC 2012* (pp. 49-64). Springer Berlin Heidelberg.
- Elfeky, M. G., Verykios, V. S., & Elmagarmid, A. K. (2002). TAILOR: A record linkage toolbox. In *Data Engineering, 2002. Proceedings. 18th International Conference on* (pp. 17-28). IEEE.
- Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. Knowledge and Data Engineering, IEEE Transactions on, 19(1), 1-16.
- Euzenat, J., & Shvaiko, P. (2007). *Ontology matching* (Vol. 333). Heidelberg: Springer.
- Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183-1210.
- Ferrara, A., Lorusso, D., Montanelli, S., & Varese, G. (2008, October). Towards a benchmark for instance matching. In The 7th International Semantic Web Conference (p. 37).
- Ferrara, A., Montanelli, S., Noessner, J., & Stuckenschmidt, H. (2011). Benchmarking matching applications on the semantic web. In The Semantic Web: Research and Applications (pp. 108-122). Springer Berlin Heidelberg.
- Ferrara, A., Nikolov, A., Noessner, J., & Scharffe, F. (2013). Evaluation of instance matching tools: The experience of OAEI. *Web semantics: Science, services and agents on the World Wide Web*, *21*, 49-60.
- Ferrara, A., Nikolov, A., & Scharffe, F. (2013). Data linking for the semantic web. Semantic Web: Ontology and Knowledge Base Enabled Tools, Services, and Applications, 169.
- Getoor, L., & Machanavajjhala, A. (2013, August). Entity resolution for big data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1527-1527). ACM.
- Gropp, W., Lusk, E., Doss, N., & Skjellum, A. (1996). A high-performance, portable implementation of the MPI message passing interface standard.*Parallel computing*, *22*(6), 789-828.
- Gusfield, D., & Irving, R. W. (1989). *The stable marriage problem: structure and algorithms*. MIT press.
- Halevy, A., Rajaraman, A., & Ordille, J. (2006, September). Data integration: the teenage years. In Proceedings of the 32nd international conference on Very large data bases (pp. 9-16). VLDB Endowment.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10-18.
- Han, S. N., Lee, G. M., & Crespi, N. (2014). Semantic context-aware service composition for building automation system. Industrial Informatics, IEEE Transactions on, 10(1), 752-761.
- Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36.
- Hernández, M. A., & Stolfo, S. J. (1995, June). The merge/purge problem for large databases. In ACM SIGMOD Record (Vol. 24, No. 2, pp. 127-138). ACM.
- Hernández, M. A., & Stolfo, S. J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. Data mining and knowledge discovery, 2(1), 9-37.
- Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. Neural computation, 18(7), 1527-1554.
- Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification.
- Hu, W., Qu, Y. Z., & Sun, X. Z. (2011). Bootstrapping object coreferencing on the semantic web. *Journal of Computer Science and Technology*, *26*(4), 663-675.
- Isele, R., Jentzsch, A., & Bizer, C. (2011, June). Efficient Multidimensional Blocking for Link Discovery without losing Recall. In *WebDB*.
- Jaffri, A., Glaser, H., & Millard, I. (2008). Uri disambiguation in the context of linked data.
- Jean-Mary, Y. R., Shironoshita, E. P., & Kabuka, M. R. (2010). Asmov: Results for oaei 2010. *Ontology Matching*, *126*.
- Jeffery, S. R., Franklin, M. J., & Halevy, A. Y. (2008, June). Pay-as-you-go user feedback for dataspace systems. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* (pp. 847-860). ACM.
- Jiménez-Ruiz, E., & Grau, B. C. (2011). Logmap: Logic-based and scalable ontology matching. In *The Semantic Web–ISWC 2011* (pp. 273-288). Springer Berlin Heidelberg.
- Joachims, T. (1999). Making large scale SVM learning practical. Universität Dortmund.

- Kejriwal, M., & Miranker, D. P. (2013, December). An unsupervised algorithm for learning blocking schemes. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on* (pp. 340-349). IEEE.
- Kejriwal, M., & Miranker, D. P. (2014). A two-step blocking scheme learner for scalable link discovery. *Ontology Matching*, *49*.
- Kejriwal, M., & Miranker, D. P. (2015). A DNF Blocking Scheme Learner for Heterogeneous Datasets. *arXiv preprint arXiv:1501.01694*.
- Kejriwal, M., & Miranker, D. P. (2015). Semi-supervised Instance Matching Using Boosted Classifiers. In *The Semantic Web. Latest Advances and New Domains* (pp. 388-402). Springer International Publishing.
- Kejriwal, M., & Miranker, D. P. (2015). An Unsupervised Instance Matcher for Schema-free RDF Data. *Web Semantics: Science, Services and Agents on the World Wide Web*.
- Kejriwal, M., & Miranker, D. P (2015). Sorted Neighborhood for Schema-free RDF Data.
- Kim, H. S., & Lee, D. (2010, March). HARRA: fast iterative hashed record linkage for large-scale data collections. In *Proceedings of the 13th International Conference on Extending Database Technology* (pp. 525-536). ACM.
- Kirsten, T., Kolb, L., Hartung, M., Groß, A., Köpcke, H., & Rahm, E. (2010). Data partitioning for parallel entity matching. *arXiv preprint arXiv:1006.5309*.
- Klyne, G., & Carroll, J. J. (2006). Resource description framework (RDF): Concepts and abstract syntax.
- Kolb, L., Thor, A., & Rahm, E. (2012). Multi-pass sorted neighborhood blocking with MapReduce. Computer Science-Research and Development, 27(1), 45-63.
- Kolb, L., Thor, A., & Rahm, E. (2012). Dedoop: efficient deduplication with Hadoop. *Proceedings of the VLDB Endowment*, *5*(12), 1878-1881.
- Köpcke, H., Thor, A., & Rahm, E. (2010). Evaluation of entity resolution approaches on real-world match problems. *Proceedings of the VLDB Endowment*, *3*(1-2), 484-493.
- Köpcke, H., & Rahm, E. (2010). Frameworks for entity matching: A comparison. *Data & Knowledge Engineering*, *69*(2), 197-210.
- Lenzerini, M. (2002, June). Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 233-246). ACM.
- Leonardi, E., Abel, F., Heckmann, D., Herder, E., Hidders, J., & Houben, G. J. (2010). *A flexible rule-based method for interlinking, integrating, and enriching user data* (pp. 322-336). Springer Berlin Heidelberg.
- Li, J., Tang, J., Li, Y., & Luo, Q. (2009). Rimom: A dynamic multistrategy ontology alignment framework. *Knowledge and Data Engineering, IEEE Transactions on*, *21*(8), 1218-1232.
- Ma, Y. (2014, June). Effective Instance Matching for Heterogeneous Structured Data.
- Ma, Y., Tran, T., & Bicer, V. (2013, April). Typifier: Inferring the type semantics of structured data. In *Data Engineering (ICDE), 2013 IEEE 29th International Conference on* (pp. 206-217). IEEE.
- Ma, K., & Yang, B. (2015, September). Parallel NoSQL Entity Resolution Approach with Mapreduce. In *Intelligent Networking and Collaborative Systems (INCOS), 2015 International Conference on* (pp. 384-389). IEEE.
- McCallum, A., Nigam, K., & Ungar, L. H. (2000, August). Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 169-178). ACM.
- McCarthy, J. F., & Lehnert, W. G. (1995). Using decision trees for coreference resolution. arXiv preprint cmp-lg/9505043.
- McGuinness, D. L., & Van Harmelen, F. (2004). OWL web ontology language overview. *W3C recommendation*, *10*(10), 2004.
- Menestrina, D., Whang, S. E., & Garcia-Molina, H. (2010). Evaluating entity resolution results. *Proceedings of the VLDB Endowment*, *3*(1-2), 208-219.
- Moro, A., Raganato, A., & Navigli, R. (2014). Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, *2*, 231-244.
- Mesnil, G., Dauphin, Y., Glorot, X., Rifai, S., Bengio, Y., Goodfellow, I. J., ... & Vincent, P. (2012). Unsupervised and Transfer Learning Challenge: a Deep Learning Approach. ICML Unsupervised and Transfer Learning, 27, 97-110.
- Metwally, A., & Faloutsos, C. (2012). V-smart-join: A scalable mapreduce framework for all-pair similarity joins of multisets and vectors. *Proceedings of the VLDB Endowment*, *5*(8), 704-715.
- Michelson, M., & Knoblock, C. A. (2006, July). Learning blocking schemes for record linkage. In *Proceedings of the National Conference on Artificial Intelligence* (Vol. 21, No. 1, p. 440). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

- Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*,*5*(1), 32-38.
- Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, A. P. (1959). Automatic Linkage of Vital Records Computers can be used to extract" follow-up" statistics of families from files of routine records. *Science*, *130*(3381), 954-959.
- Ngomo, A. C. N. (2011). A time-efficient hybrid approach to link discovery. *Ontology Matching*, 1.
- Ngomo, A. C. N., & Auer, S. (2011). Limes-a time-efficient approach for large-scale link discovery on the web of data. *integration*, *15*, 3.
- Ngomo, A. C. N., Lehmann, J., Auer, S., & Höffner, K. (2011, October). Raven–active learning of link specifications. In *Proceedings of the Sixth International Workshop on Ontology Matching* (pp. 25-37).
- Ngomo, A. C. N., & Lyko, K. (2012). Eagle: Efficient active learning of link specifications using genetic programming. In *The Semantic Web: Research and Applications* (pp. 149-163). Springer Berlin Heidelberg.
- Ngomo, A. C. N., & Lyko, K. (2013, October). Unsupervised learning of link specifications: deterministic vs. non-deterministic. In *OM* (pp. 25-36).
- Ngomo, A. C. N., Lyko, K., & Christen, V. (2013). Coala–correlation-aware active learning of link specifications. In *The Semantic Web: Semantics and Big Data* (pp. 442-456). Springer Berlin Heidelberg.
- Nikolov, A., Uren, V., Motta, E., & De Roeck, A. (2008). Integration of semantically annotated data by the KnoFuss architecture. In *Knowledge Engineering: Practice and Patterns* (pp. 265-274). Springer Berlin Heidelberg.
- Nikolov, A., Uren, V., Motta, E., & De Roeck, A. (2009). Towards Data Fusion in a multi-ontology Environment.
- Niu, X., Rong, S., Wang, H., & Yu, Y. (2012, October). An effective rule miner for instance matching in a web of data. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 1085-1094). ACM.
- Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, *22*(10), 1345-1359.
- Papadakis, G., Demartini, G., Fankhauser, P., & Kärger, P. (2010, November). The missing links: Discovering hidden same-as links among a billion of triples. In Proceedings of the 12th International Conference on Information Integration and Web-based Applications & Services (pp. 453-460). ACM.
- Papadakis, G., Ioannou, E., Palpanas, T., Niederée, C., & Nejdl, W. (2013). A blocking framework for entity resolution in highly heterogeneous information spaces. *Knowledge and Data Engineering, IEEE Transactions on*, *25*(12), 2665-2682.
- Peleg, D. (2007). Approximation algorithms for the label-cover max and red-blue set cover problems. Journal of Discrete Algorithms, 5(1), 55-64.
- Puhlmann, S., Weis, M., & Naumann, F. (2006). XML duplicate detection using sorted neighborhoods. In Advances in Database Technology-EDBT 2006 (pp. 773-791). Springer Berlin Heidelberg.
- Pujara, J., Miao, H., Getoor, L., & Cohen, W. (2013). Knowledge graph identification. In The Semantic Web–ISWC 2013 (pp. 542-557). Springer Berlin Heidelberg.
- Quilitz, B., & Leser, U. (2008). *Querying distributed RDF data sources with SPARQL* (pp. 524-538). Springer Berlin Heidelberg.
- Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *the VLDB Journal*, *10*(4), 334-350.
- Raimond, Y., Sutton, C., & Sandler, M. B. (2008). Automatic Interlinking of Music Datasets on the Semantic Web. *LDOW*, *369*.
- Ravikumar, P., & Cohen, W. W. (2004, July). A hierarchical graphical model for record linkage. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence* (pp. 454-461). AUAI Press.
- Raz, R., & Safra, S. (1997, May). A sub-constant error-probability low-degree test, and a sub-constant error-probability PCP characterization of NP. In Proceedings of the twenty-ninth annual ACM symposium on Theory of computing (pp. 475-484). ACM.
- Rong, S., Niu, X., Xiang, E. W., Wang, H., Yang, Q., & Yu, Y. (2012). A machine learning approach for instance matching based on similarity metrics. In *The Semantic Web–ISWC 2012* (pp. 460-475). Springer Berlin Heidelberg.

- Sadosky, P., Shrivastava, A., Price, M., & Steorts, R. C. (2015). Blocking Methods Applied to Casualty Records from the Syrian Conflict. *arXiv preprint arXiv:1510.07714*.
- Salton, G., & McGill, M. J. (1986). Introduction to modern information retrieval.
- Sahoo, S. S., Halb, W., Hellmann, S., Idehen, K., Thibodeau Jr, T., Auer, S., ... & Ezzat, A. (2009). A survey of current approaches for mapping of relational databases to RDF. *W3C RDB2RDF Incubator Group Report*, 113-130.
- Scharffe, F., Liu, Y., & Zhou, C. (2009). Rdf-ai: an architecture for rdf datasets matching, fusion and interlink. In *Proc. IJCAI 2009 workshop on Identity, reference, and knowledge representation (IR-KR), Pasadena (CA US)*.
- Schmachtenberg, M., Bizer, C., & Paulheim, H. (2014). Adoption of the linked data best practices in different topical domains. In *The Semantic Web–ISWC 2014* (pp. 245-260). Springer International Publishing.
- Schroeder, B., & Gibson, G. (2010). A large-scale study of failures in high-performance computing systems. *Dependable and Secure Computing, IEEE Transactions on*, *7*(4), 337-350.
- Sequeda, J. F., & Miranker, D. P. (2013). Ultrawrap: SPARQL execution on relational data. *Web Semantics: Science, Services and Agents on the World Wide Web*, *22*, 19-39.
- Settles, B. (2010). Active learning literature survey. *University of Wisconsin, Madison*, *52*(55-66), 11.
- Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H., & Hall, W. (2012). Linked open government data: Lessons from data. gov. uk. *IEEE Intelligent Systems*, *27*(3), 16-24.
- Song, D., & Heflin, J. (2011). Automatically generating data linkages using a domain-independent candidate selection approach. In *The Semantic Web–ISWC 2011* (pp. 649-664). Springer Berlin Heidelberg.
- Soru, T., & Ngomo, A. C. N. (2014, September). A comparison of supervised learning classifiers for link discovery. In *Proceedings of the 10th International Conference on Semantic Systems* (pp. 41-44). ACM.
- Stephenson, C. (1980). The methodology of historical census record linkage: a users guide to the Soundex. Journal of Family History, 5(1), 112-115.
- Stoilos, G., Simou, N., Stamou, G., & Kollias, S. (2006). Uncertainty and the semantic web. Intelligent Systems, IEEE, 21(5), 84-87.
- Suchanek, F. M., Abiteboul, S., & Senellart, P. (2011). Paris: Probabilistic alignment of relations, instances, and schema. *Proceedings of the VLDB Endowment*, *5*(3), 157-168.
- Tian, A., Kejriwal, M., & Miranker, D. P. (2014, June). Schema matching over relations, attributes, and data values. In *Proceedings of the 26th International Conference on Scientific and Statistical Database Management*(p. 28). ACM.
- Vernica, R., Carey, M. J., & Li, C. (2010, June). Efficient parallel set-similarity joins using MapReduce. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (pp. 495-506). ACM.
- Volz, J., Bizer, C., Gaedke, M., & Kobilarov, G. (2009). Silk-A Link Discovery Framework for the Web of Data. *LDOW*, *538*.
- Whang, S. E., Menestrina, D., Koutrika, G., Theobald, M., & Garcia-Molina, H. (2009, June). Entity resolution with iterative blocking. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*(pp. 219-232). ACM.
- White, T. (2012). *Hadoop: The definitive guide*. " O'Reilly Media, Inc.".
- Winkler, W. E. (1993). Improved decision rules in the Fellegi-Sunter model of record linkage.
- Winkler, W. E. (1999). The state of record linkage and current research problems. In *Statistical Research Division, US Census Bureau*.
- Winkler, W. E. (2002). *Methods for record linkage and bayesian networks*. Technical report, Statistical Research Division, US Census Bureau, Washington, DC.
- Yan, S., Lee, D., Kan, M. Y., & Giles, L. C. (2007, June). Adaptive sorted neighborhood methods for efficient record linkage. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (pp. 185-194). ACM.
- Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010, June). Spark: cluster computing with working sets. In *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing* (Vol. 10, p. 10).
- Zhai, C., & Lafferty, J. (2001, September). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 334-342). ACM.
- Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning.*Synthesis lectures on artificial intelligence and machine learning*, *3*(1), 1-130.