Big Data Analytics for Semantic Data
BigSem Tutorial

Discussion and Conclusion

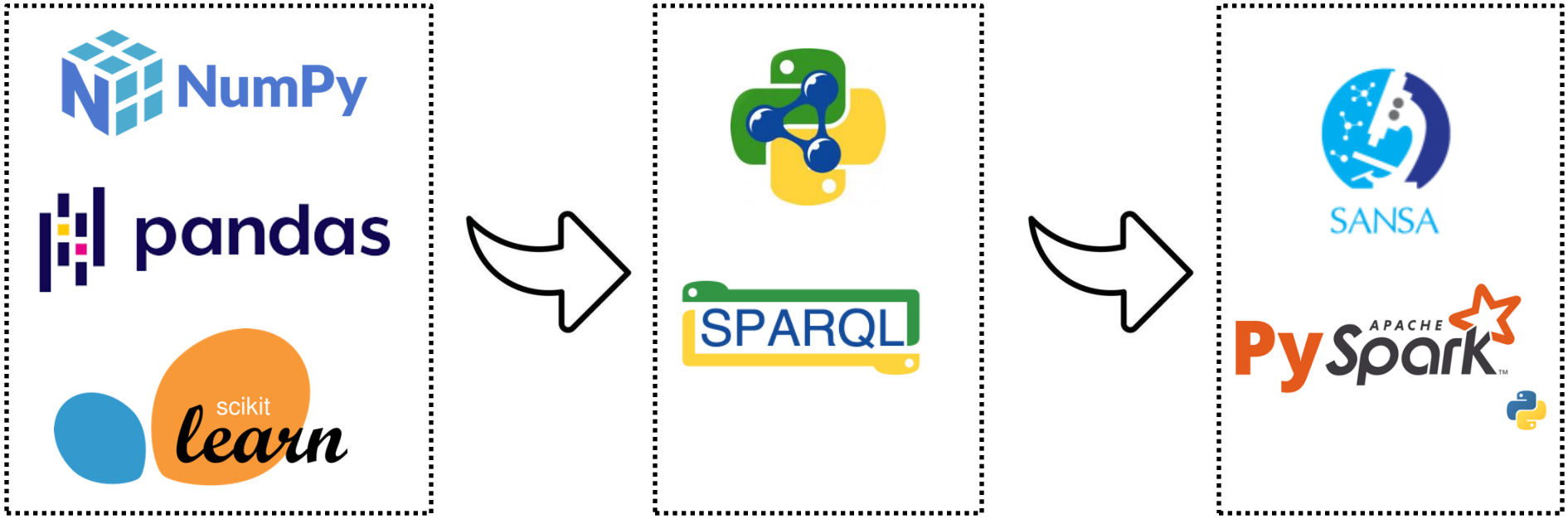
Chelmis Charalampos, Bedirhan Gergin
University at Albany, SUNY

ISWC 2024

goo.gl/6Nrc1r

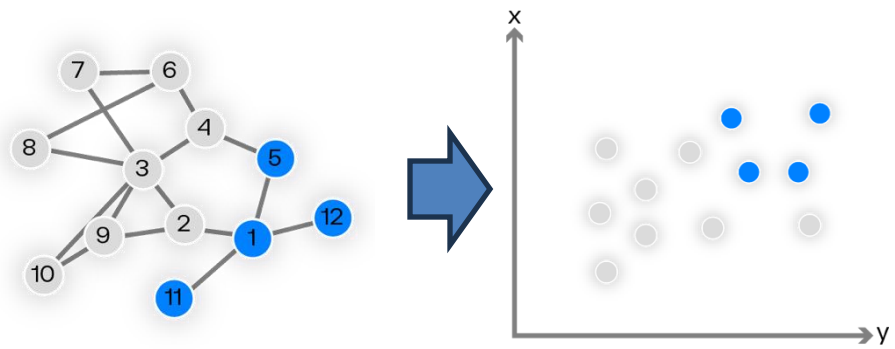
Recap

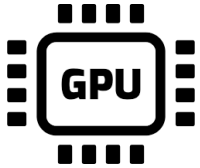
- What did we cover?
  - Libraries for analytics and ML in Python (Numpy, Pandas, Scikit Learn)
  - Libraries for semantic data access (RDFLib, SPARQLWrapper, Sparql-dataframe)
  - Semantic data analytic engines and frameworks (SANSA Stack, SparkKG-ML)



Recap

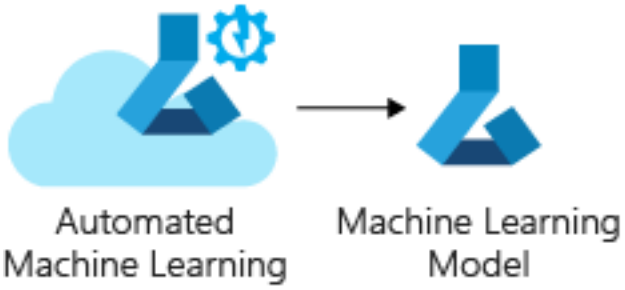
- What didn't we cover?
  - KG embeddings (supported in SparkKG-ML but not in SANSA)
  - Semantic operations in SANSA: e.g inference
  - Single/Multi-machine large-scale processing using GPUs
  - Commercial products: e.g. stardog

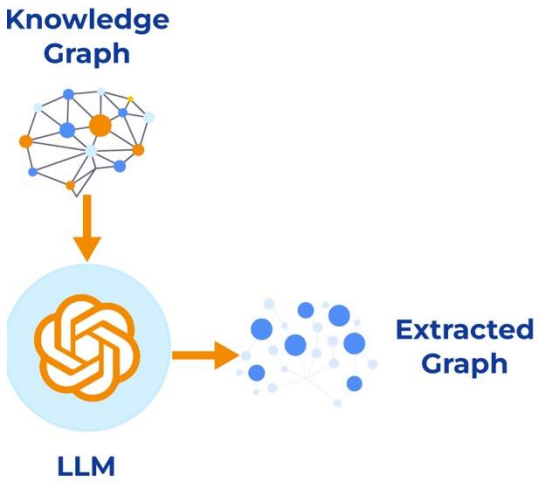




Potential Future Directions

- Addressing gaps in SANSA/SparkKG-ML
- AutoML (automating/simplifying ML pipelines end to end)
- Distributed + GPUs
- Temporal KGs
- LLMs





Conclusion

- IDIAS Lab Webpage: http://www.cs.albany.edu/~cchelmis/ideaslab.html
- IDIAS Lab Gitbhub: https://github.com/IDIASLab

- Let's make a 

!

References

- [1] https://sansa-stack.net/iswc2020-tutorial/
- [2] https://sansa-stack.net/iswc2019-tutorial/
- [3] Taken from https://adasci.org/knowledge-graphs/
- [4] Xiang Wang, Dingxian Wang, Canran Xu, Xiangnan He, Yixin Cao, and Tat-Seng Chua. 2019. Explainable reasoning over knowledge graphs for recommendation.
- [5] Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/
- [6] Sourced from https://blog.ml6.eu/how-are-knowledge-graphs-and-machine-learning-related-ff6f5c1760b5
- [7] Hassanzadeh, O., Consens, M.P.: Linked movie data base. In: LDOW (2009), https://api.semanticscholar.org/CorpusID:16810971
- [8] Chelmis, C., Gergin, B.: A Knowledge Graph for Semantic-Driven Healthiness Evaluation of Online Recipes. https://doi.org/10.7910/DVN/99PNJ5 (2022).
- [9] Lehmann, J., Sejdiu, G., Bühmann, L., Westphal, P., Stadler, C., Ermilov, I., Bin, S., Chakraborty, N., Saleem, M., Ngomo, A.C.N., et al.: Distributed semantic analytics using the sansa stack. In: International Semantic Web Conference. pp. 147–155. Springer (2017)
- [10] Carsten Felix Draschner, Claus Stadler, Farshad Bakhshandegan Moghaddam, Jens Lehmann, and Hajira Jabeen. 2021. DistRDF2ML - Scalable Distributed In-Memory Machine Learning Pipelines for RDF Knowledge Graphs. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM '21). Association for Computing Machinery, New York, NY, USA, 4465–4474. https://doi.org/10.1145/3459637.3481999
- [11] https://github.com/IDIASLab/SparkKG-ML

Additional References

- AnzoGraph DB. https://cambridgesemantics.com/anzograph/
- Pandas. https://pandas.pydata.org/
- RDFLib. https://rdflib.readthedocs.io/en/stable/
- scikit-learn. https://scikit-learn.org/stable/
- Stardog: The Enterprise Knowledge Graph Platform. https://www.stardog.com/
- Banane, M., Belangour, A.: Rdfspark: a new solution for querying massive RDF data using spark. International Journal of Engineering & Technology 8 (2019)
- Du, J.H., Wang, H.F., Ni, Y., Yu, Y.: Hadooprdf: A scalable semantic data analytical engine. In: International conference on intelligent computing. pp. 633–641. Springer (2012)
- Hassanzadeh, O., Consens, M.P.: Linked movie data base. In: LDOW (2009), https://api.semanticscholar.org/CorpusID:16810971
- Saeed, M.R., Chelmis, C., Prasanna, V.K.: Asqfor: Automatic sparql query formulation for the non-expert. AI Communications 31(1), 19–32 (2018)
- Saeed, M.R., Chelmis, C., Prasanna, V.K.: Extracting entity-specific substructures for rdf graph embeddings. Semantic Web 10(6), 1087–1108 (2019)
- Saeed, M.R., Chelmis, C., Prasanna, V.K., et al.: Thou shalt asqfor and shalt receive the semantic answer. In: IJCAI. pp. 4264–4265 (2016)
- Schätzle, A., Przyjaciel-Zablocki, M., Skilevic, S., Lausen, G.: S2rdf: Rdf querying with sparql on spark. Proceedings of the VLDB Endowment 9(10), 804–815 (2016)