BigSem: Big Data Analytics for Semantic Data Tutorial

Introduction

Chelmis Charalampos, Bedirhan Gergin
University at Albany, SUNY

ISWC 2024

goo.gl/6Nrc1r

Organizers

- Charalampos Chelmis
  - Assistant Professor in the Department of Computer Science at the University at Albany, SUNY
  - Director of the Intelligent Big Data Analytics, Applications, and Systems (IDIAS) Lab.
  - Dr. Chelmis’ research interests comprise Network Science and Big Data analytics.

- Bedirhan Gergin
  - PhD Candidate specializing in Semantic Web and knowledge graphs.
  - Currently a Research Assistant at the IDIAS Lab (Intelligent Big Data Analytics, Applications, and Systems) at UAlbany.
  - Former Data Scientist intern @IBM.

Objectives

- By providing an overview of the state of the art in scalable, distributed analytics for semantic data, this tutorial aims to:
  - Raise awareness of the gap between the Semantic Web, Big Data analytics, and ML communities,
  - Help promote the synergy between these communities,
  - Encourage the discussion and exchange of ideas about this topic.

Tutorial Outline

I. Introduction, overview, setup instructions

II. Module 1: Libraries for analytics and ML in Python

- Numpy, Pandas, Scikit Learn

III. Module 2: Libraries for semantic data access

- RDFLib, SPARQLWrapper, Sparql-dataframe

IV. Module 3: Semantic data analytic engines and frameworks

- SANSA Stack, SparkKG-ML

Relevant Tutorials

- Scalable RDF Analytics with SANSA (ISWC 2020) [1].

- SANSA"s Leap of Faith: Scalable RDF and Heterogeneous Data Lakes (ISWC 2019) [2].

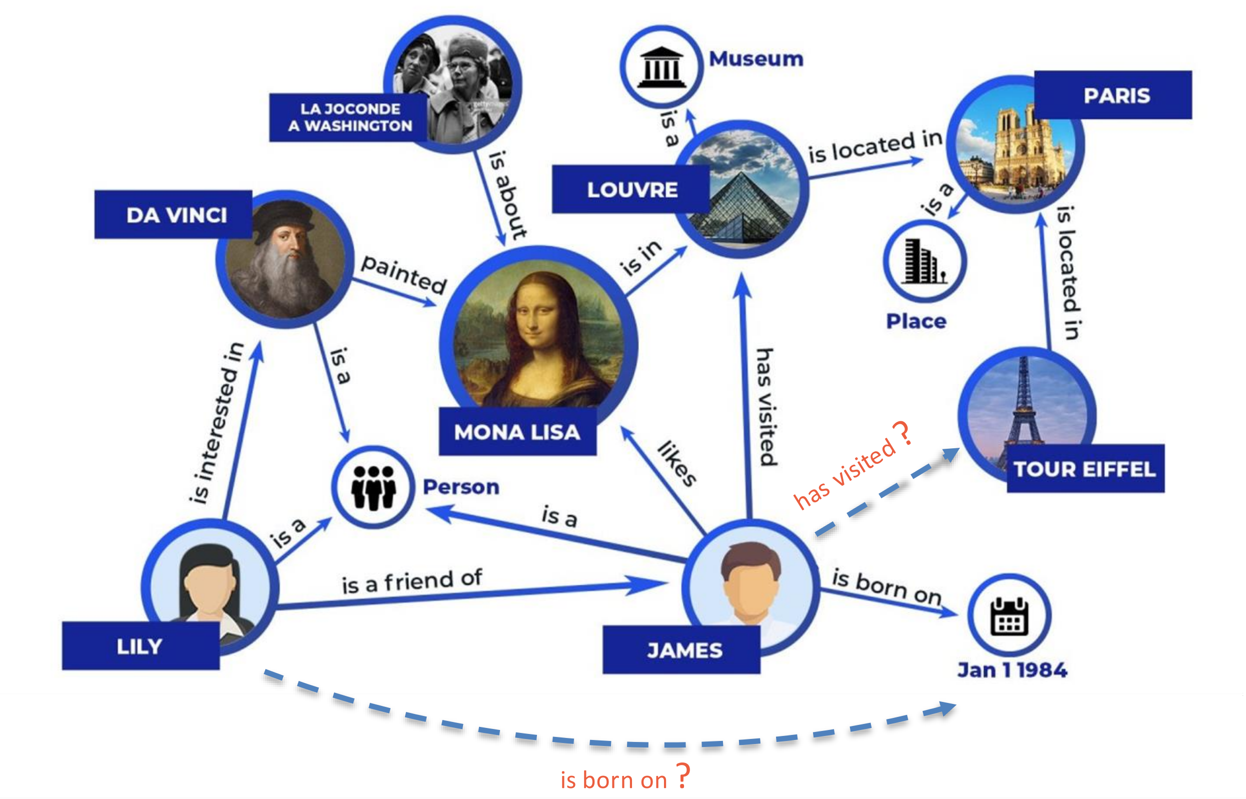
✓ Related to the "distributed analytics" session of this tutorial.

✓ These tutorials focused on scalable KG processing with SANSA.

[1] https://sansa-stack.net/iswc2020-tutorial/
[2] https://sansa-stack.net/iswc2019-tutorial/

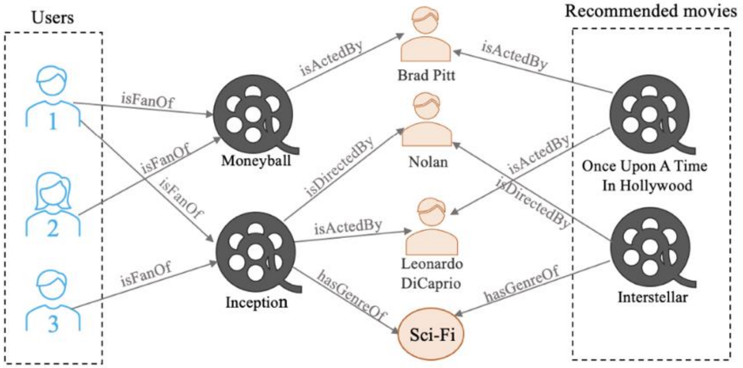
Motivation

- Knowledge graphs and Linked Open Data increased popularity.
- By analyzing KGs:
  - one can identify patterns, connections, and dependencies
  - infer new knowledge from given facts

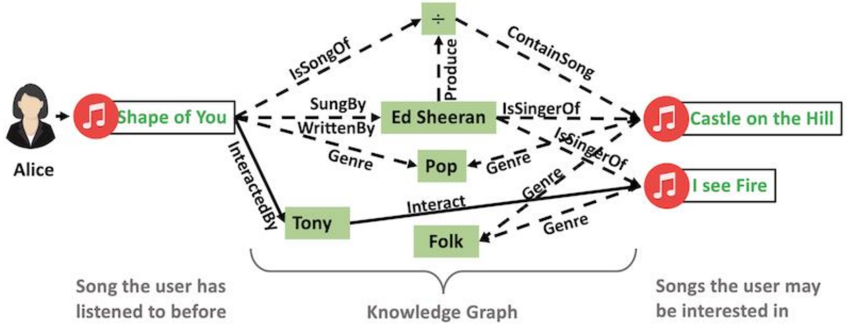


Motivation

- Useful in applications including question answering, recommendation systems, and expert systems.



[3]



[4]

[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.

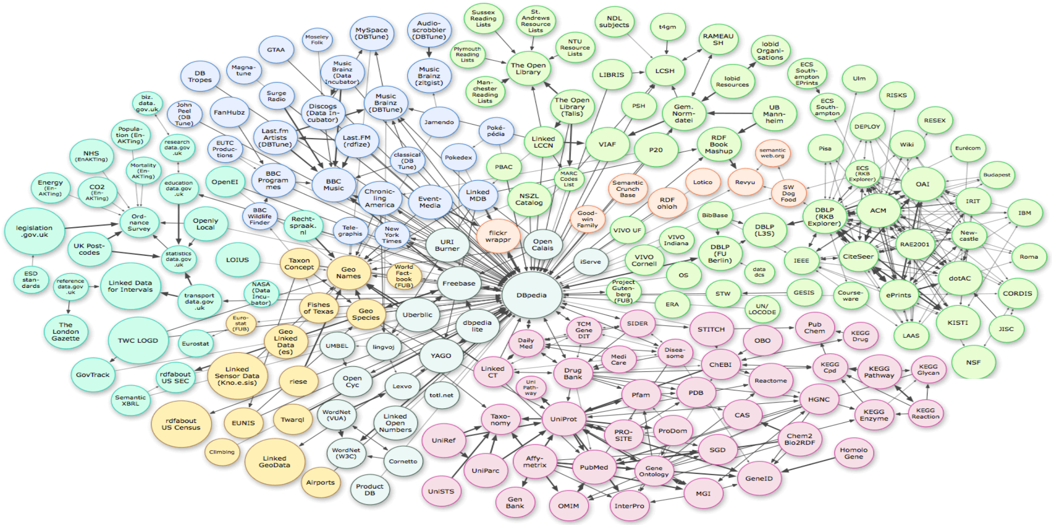
Motivation

- As the popularity of KGs expands, so does their size.

  Dbpedia (over 850 million facts from 111 different language editions of Wikipedia)

  Yago (2 billion type-consistent triples for 64 million entities)

  TweetKB (billions of tweet-related information spanning more than 9 years)



Motivation

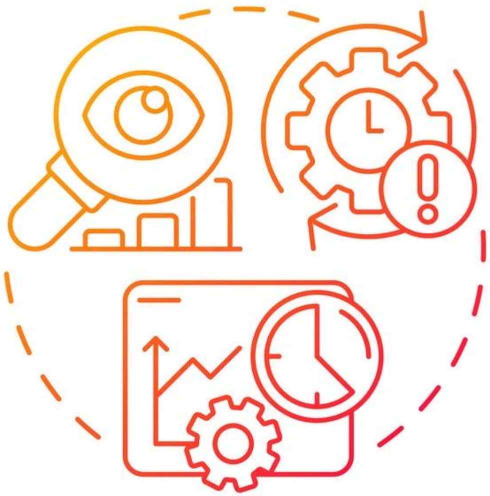
- Impossible to process within the main memory of a single computer.

  Can't load data into main memory,

  Even if you do, your memory is already consumed,

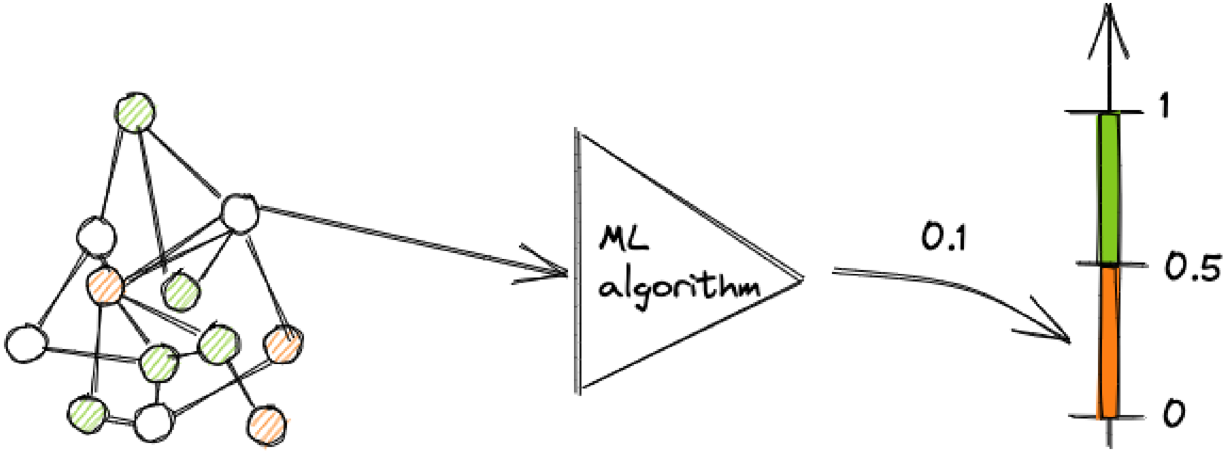
  Can’t obtain stats or do ML on it,

  Then it takes hours..



Motivation

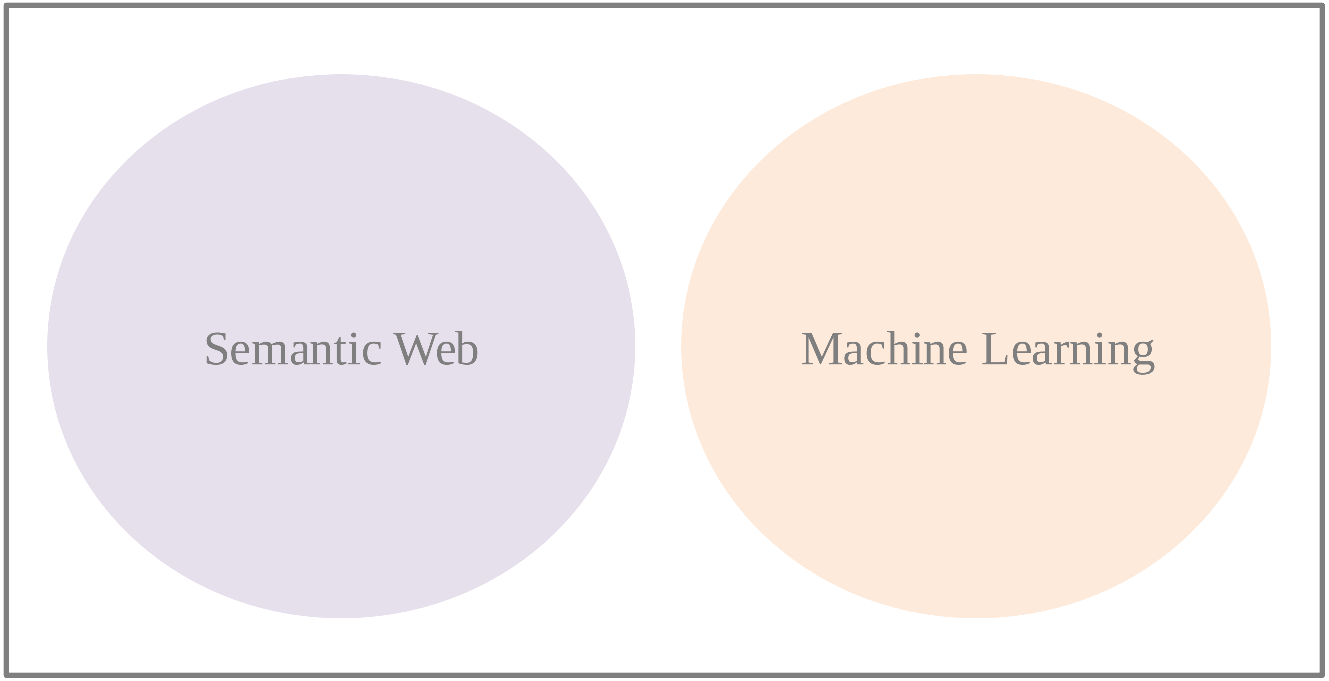
- Need for scalable data mining/analytics and ML over KGs.



[5] Sourced from https://blog.ml6.eu/how-are-knowledge-graphs-and-machine-learning-related-ff6f5c1760b5

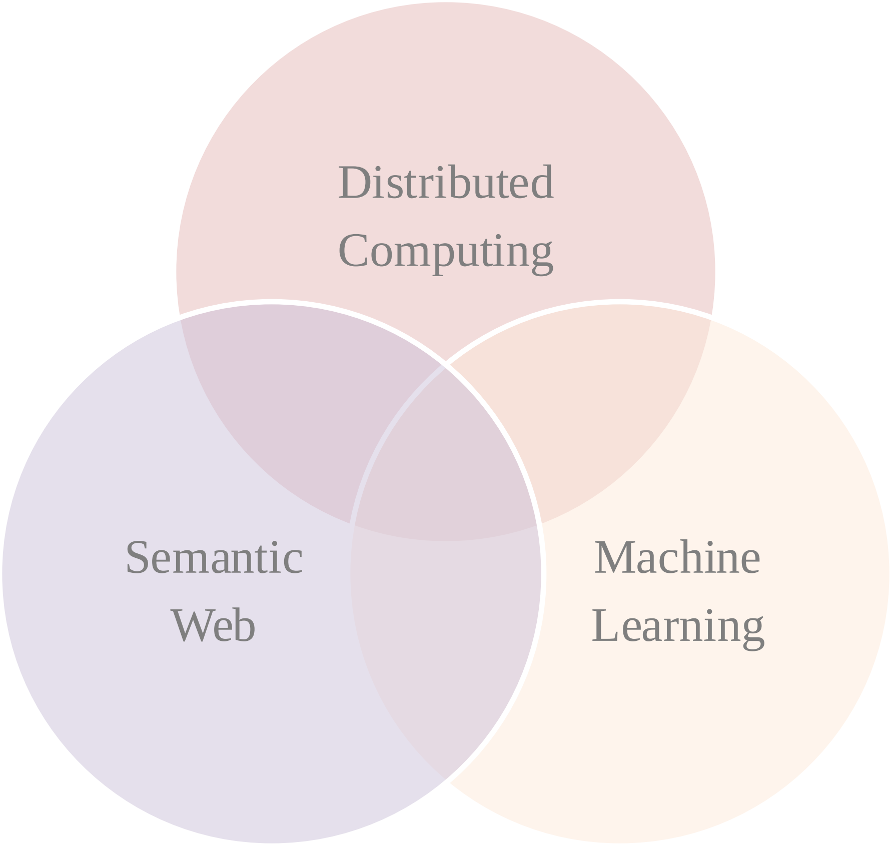
Motivation

- Disconnect between Semantic Web and Machine Learning community.



Motivation

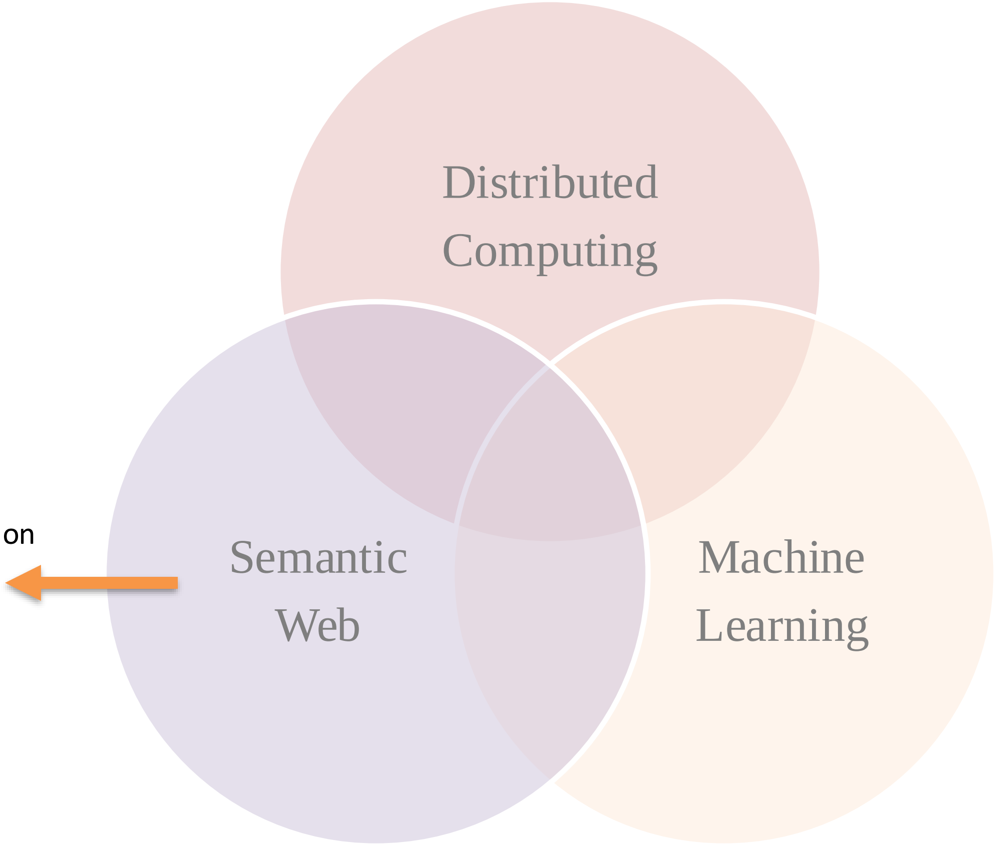
- Bridge the Semantic Web, Distributed Computing and Machine Learning communities.



Motivation

- Bridge the Semantic Web, Distributed Computing and Machine Learning communities.

-Richer Data Representation
-Linked Data
-Interoperability



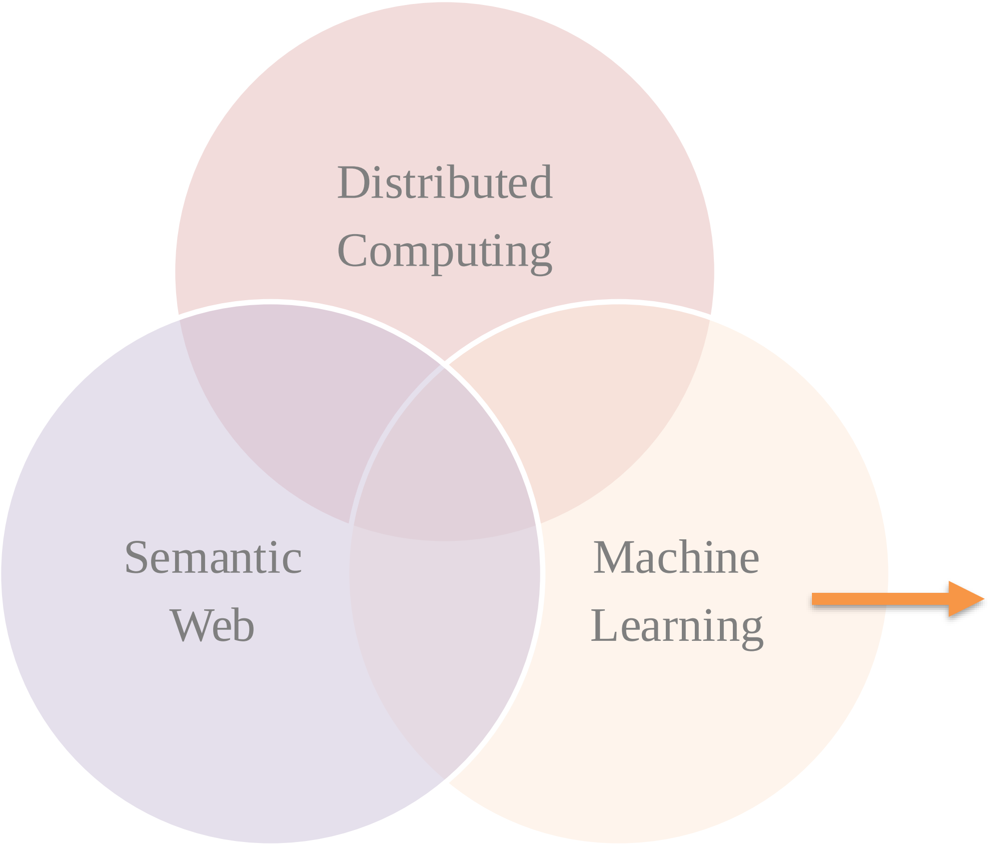
Distributed Computing

Semantic Web

Machine Learning

Motivation

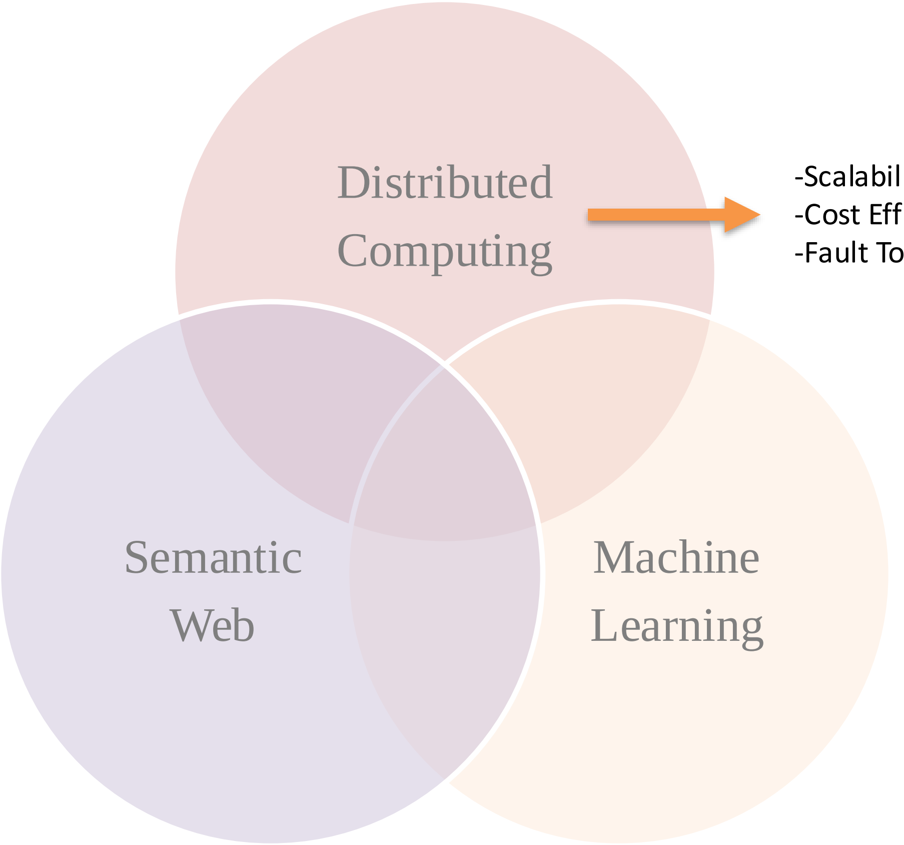
- Bridge the Semantic Web, Distributed Computing and Machine Learning communities.



-Predictive Capabilities
-Data Mining and Discovery
-Improved Decision-Making

Motivation

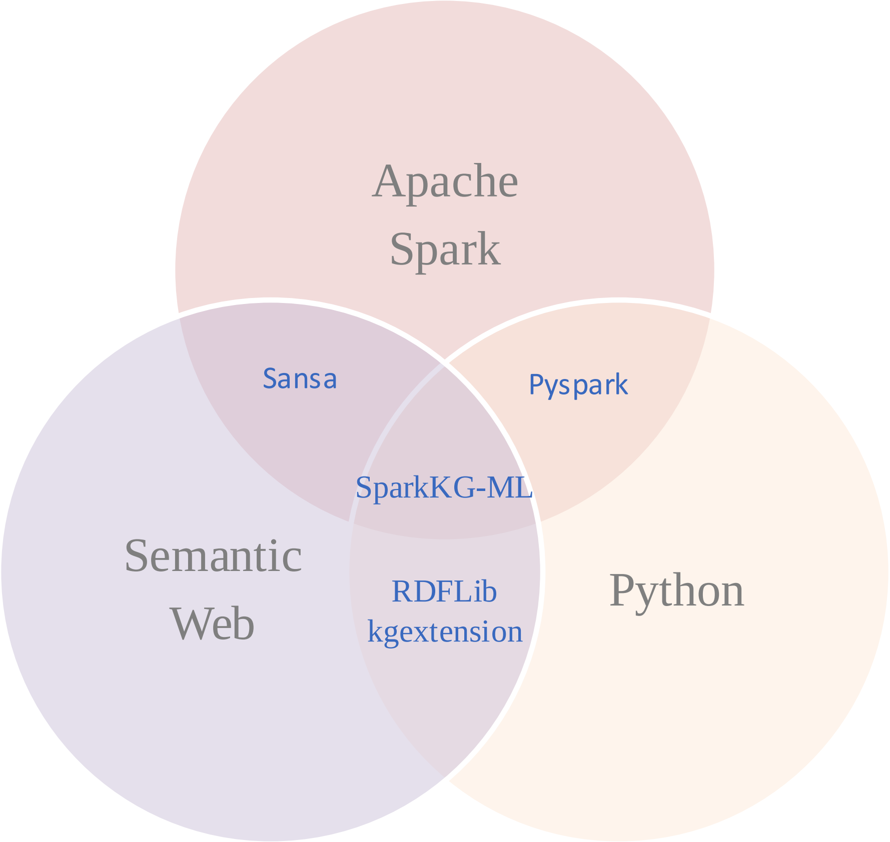
- Bridge the Semantic Web, Distributed Computing and Machine Learning communities.



-Scalability
-Cost Efficiency
-Fault Tolerance

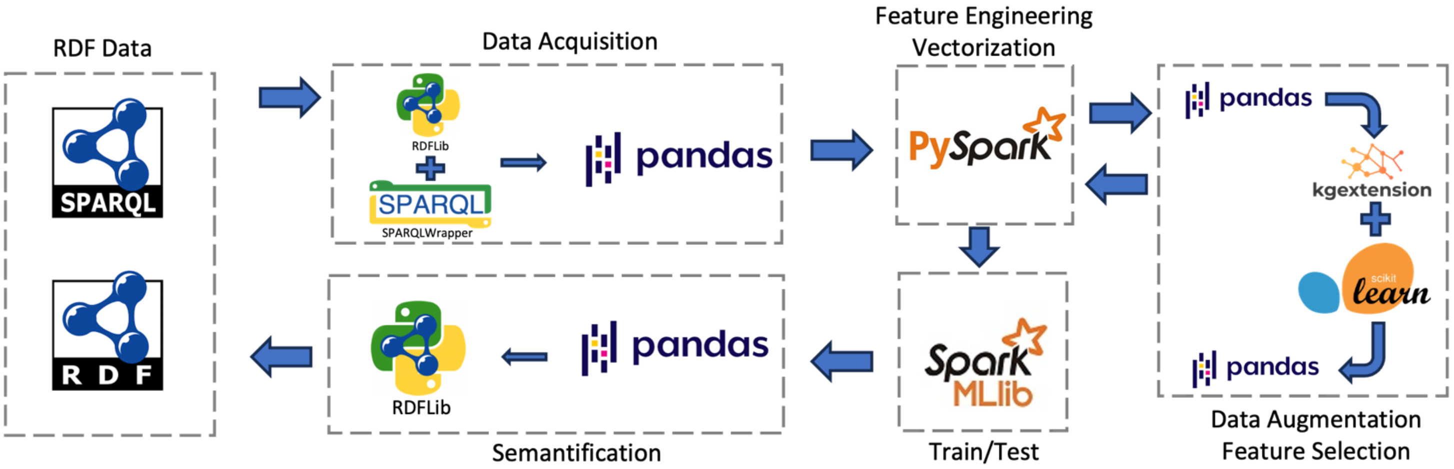
Motivation

- Bridge the Semantic Web, Distributed Computing and Machine Learning communities.



Motivation

- Current system:
  - Change from one platform to another
  - So many steps
  - No parallel processing



- Goal is to leverage advantages of all in a single framework (Python).

Setup Instructions

- Please refer to the setup_directions.md file on GitHub for detailed setup instructions.

- First, you will need a Python environment with PySpark.
  - For an easier and smoother setup, we recommend creating the environment on Google Colab.
- Second, set up the SANSA Stack on Databricks.

Data Access

- The dataset can be accessed from the original links provided on the Tutorial's GitHub website or from the datasets folder.

- Linked Movie Database [6] for SANSA hands-on.
- RecipeKG [7] for SparkKG-ML hand-son.

[6] Hassanzadeh, O., Consens, M.P.: Linked movie data base. In: LDOW (2009), https: //api.semanticscholar.org/CorpusID:16810971
[7] Chelmis, C., Gergin, B.: A Knowledge Graph for Semantic-Driven Healthiness Evaluation of Online Recipes. https://doi.org/10.7910/DVN/99PNJ5 (2022).