

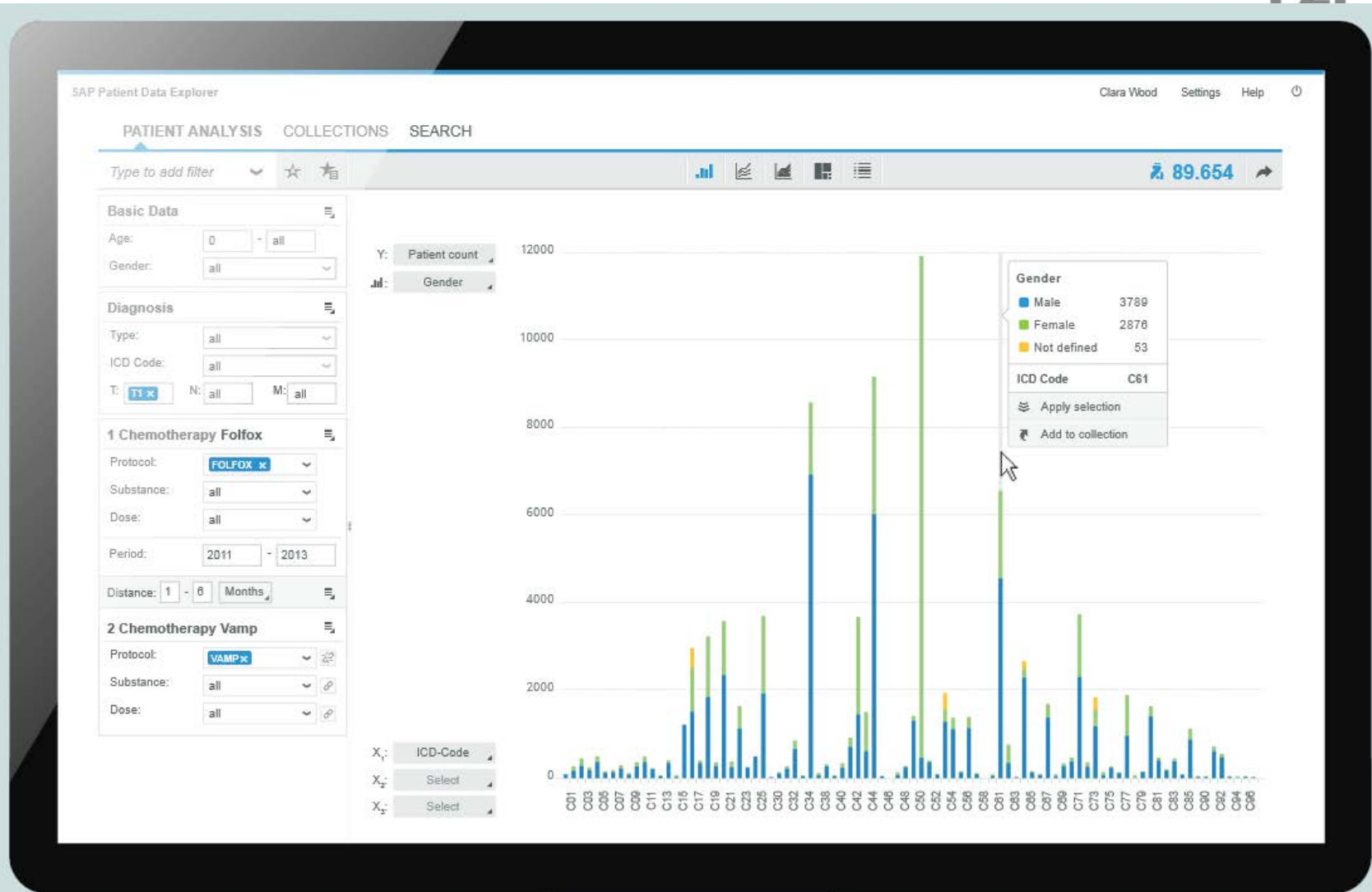
Towards a Semantic Clinical Data Warehouse: A Case Study of Discovering Similar Genes

2015-05-31, Know@LOD – ESWC

Benedikt Kämpgen, Horst Werner,
Radwan Deeb, and Christof Bornhövd



Patient Data Explorer



Patient Data Explorer (2)

SAP Patient Data Explorer Clara Wood Settings Help

PATIENT ANALYSIS **COLLECTIONS** SEARCH

My Lists (4)

C69 Konjunktiva
132 56

C41 Lung Cancer T=1
154 92 101

C00, C02.1
467

C11 Nasopharynx, T=2
263 70 35

C41 Prostate Cancer

Created: 27th October 2013
Updated: 2nd November 2013

DIABETES SUBSTANCE A ASPIRINE PENICILIN GLUTEN

Cancer, known medically as a malignant neoplasm, is a broad group of disease.

All (347) **Patients (154)** Clinical Trials (92) SOPs (101) Save to collections

	LAST NAME	FIRST NAME	DATE AT DIAGNOSIS	GENDER	FIRST DIAGNOSIS
<input type="checkbox"/>	Smith	John	31 JAN 1991	male	C12
<input checked="" type="checkbox"/>	Harrison	Kate	04 MAR 1986	female	C25
<input type="checkbox"/>	Steel	George	11 AUG 1979	male	C01
<input type="checkbox"/>	Holmes	Bruce	27 OCT 1961	male	C65
<input type="checkbox"/>	Brookstone	Rachel	06 MAY 1952	female	C12
<input type="checkbox"/>	Flynt	Alison	15 MAR 1952	female	C14
<input checked="" type="checkbox"/>	Evans	Jack	14 OCT 1951	male	C12
<input type="checkbox"/>	Baker	Magret	07 APR 1948	female	C02
<input type="checkbox"/>	Turner	James	19 SEP 1947	male	C32
<input type="checkbox"/>	Silver	Tyler	08 JUN 1946	male	C12
<input type="checkbox"/>	Lane	Lance	12 MAR 1946	male	C23

Oh, yeah?

SAP Patient Data Explorer Clara Wood Settings Help

PATIENT ANALYSIS COLLECTIONS SEARCH

My Lists (4)

C69 Konjunktiva
132 56

C41 Lung Cancer T=1
154 92 101

C00, C02.1
467

C11 Nasopharynx, T=2
263 70 35

C41 Prostate Cancer

Created: 27th October 2013
Updated: 2nd November 2013

DIABETES SUBSTANCE A ASPIRINE PENICILIN GLUTEN

Cancer, known medically as a malignant neoplasm, is a broad group of disease.

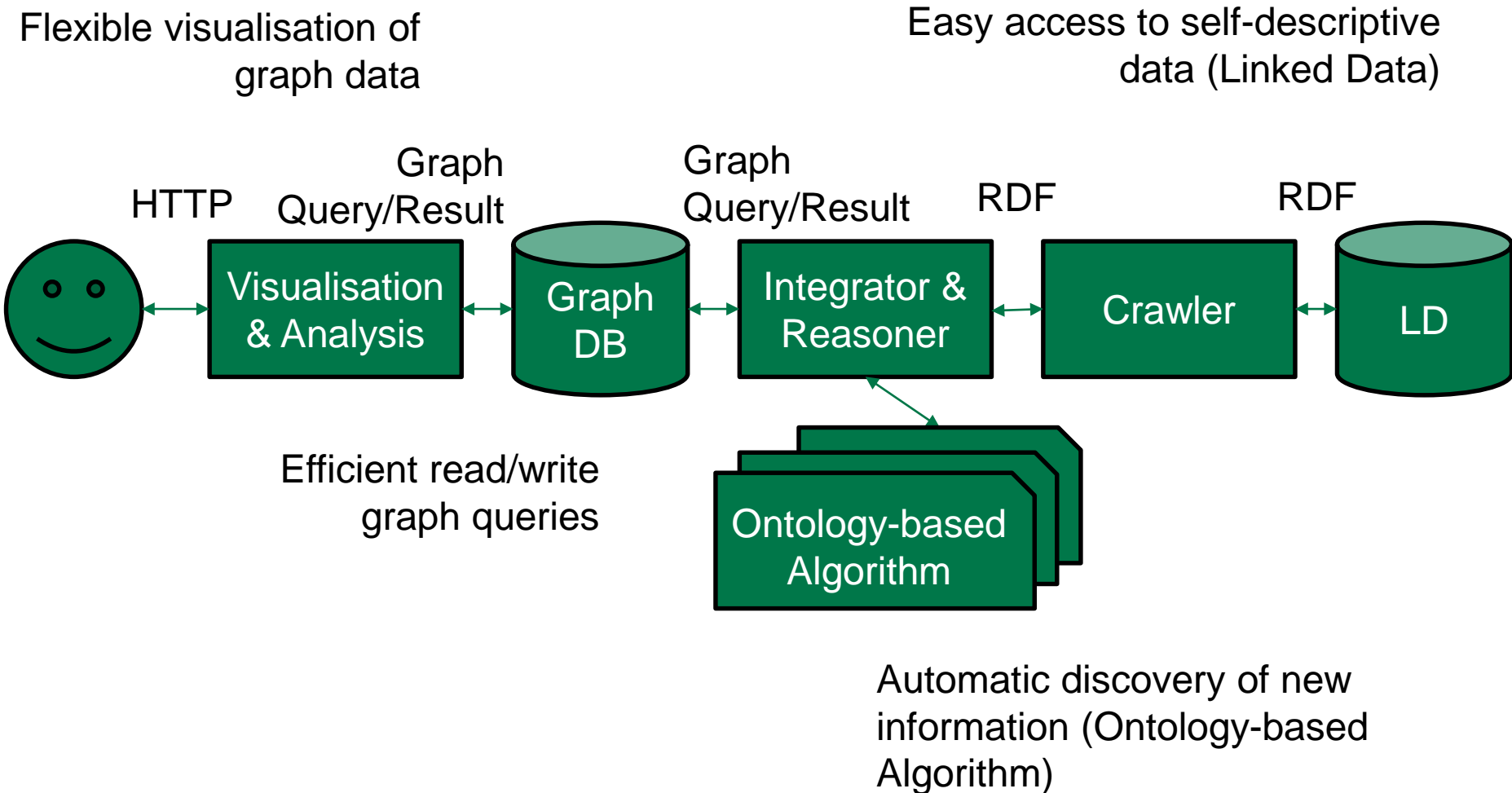
All (347) **Patients (154)** Clinical Trials (92) SOPs (101) Save to collections

	LAST NAME	FIRST NAME	DATE AT DIAGNOSIS	GENDER	FIRST DIAGNOSIS
<input type="checkbox"/>	Smith	John	31 JAN 1991	male	C12
<input checked="" type="checkbox"/>	Harrison	Kate	04 MAR 1986	female	C25
<input type="checkbox"/>	Steel	George	11 AUG 1979	male	C01
<input type="checkbox"/>	Holmes	Bruce	27 OCT 1961	male	C65
<input type="checkbox"/>	Brookstone	Rachel	06 MAY 1952	female	C12
<input type="checkbox"/>	Flynt	Alison	15 MAR 1952	female	C14
<input checked="" type="checkbox"/>	Evans	Jack	14 OCT 1951	male	C12
<input type="checkbox"/>	Baker	Magret	07 APR 1948	female	C02
<input type="checkbox"/>	Turner	James	19 SEP 1947	male	C32
<input type="checkbox"/>	Silver	Tyler	08 JUN 1946	male	C12
<input type="checkbox"/>	Lane	Lance	12 MAR 1946	male	C23

Problems

- ETL + RDBMS need adaptation if additional background information to be considered (e.g., PubMed references)
- No automatic interpretation and knowledge discovery (e.g., possibly wrong information)
- Analysis results not written back to data warehouse (e.g., for provenance tracking and information sharing)

Semantic Clinical Data Warehouse



Application: HANA Linked Data AnnSim (HLA)

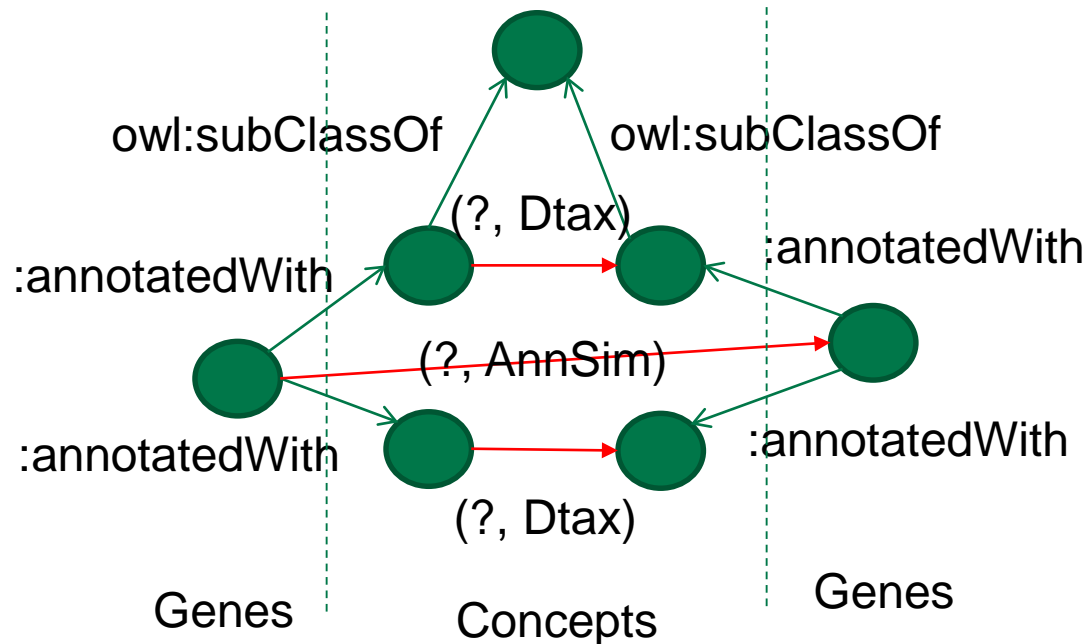
- Use Case

- Discovering similar genes of plant *Arabidopsis Thaliana*
- Similarity is an important basis for other relationships



- Implementation uses

- SAP HANA Graph
- AnnSim (Palma, '13)



Application: Setup

	HLA (Java)	AnnSim 1.0 (C/C++)
Objects	20 genes (1-aaap)	20 genes (1-aaap)
Hardware		
Client Workstation	Ubuntu 14.04 VM on W7 Intel Core i5-3360M CPU 2.80GHz, 16 GB RAM	Ubuntu 14.04 VM on W7 Intel Core i5-3360M CPU 2.80GHz, 16 GB RAM
HANA Instance	SUSE Linux Enterprise Server 11.1 500 GB RAM, 80 cores	-
Data	[1,2,3] (2014)	[1,2,3] (2013)
Triples	7,337,447	-
Size of data	537 MB	2.80 MB
Vertices	601,519	39,209
Edges	1,658,322	74,123

[1] ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/TAIR10_functional_descriptions

[2] <http://purl.obolibrary.org/obo/go.owl>

[3] ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/ATH_GO_GOSLIM.txt

Disclaimer: A direct comparison between implementations is not possible.

Application: Lessons Learned

- Correct Computation of Similarities?

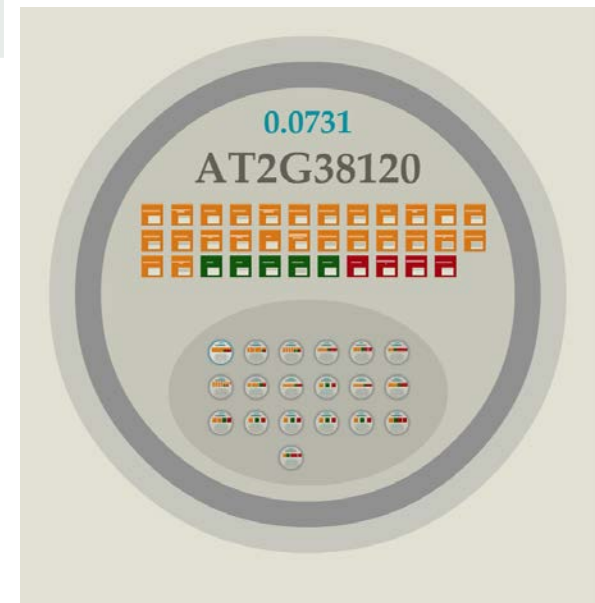
	Mean Squared Error
HLA – AnnSim 1.0	0.09
HLA – SeqSim	0.19
AnnSim 1.0 – SeqSim	0.36

- Efficient Computation of Similarities?

Approach	Load Graph	Compute AnnSim	Read Queries	Write Queries
HLA	370s	2,667s	230s	2,202s
AnnSim 1.0	0s	408s	-	-

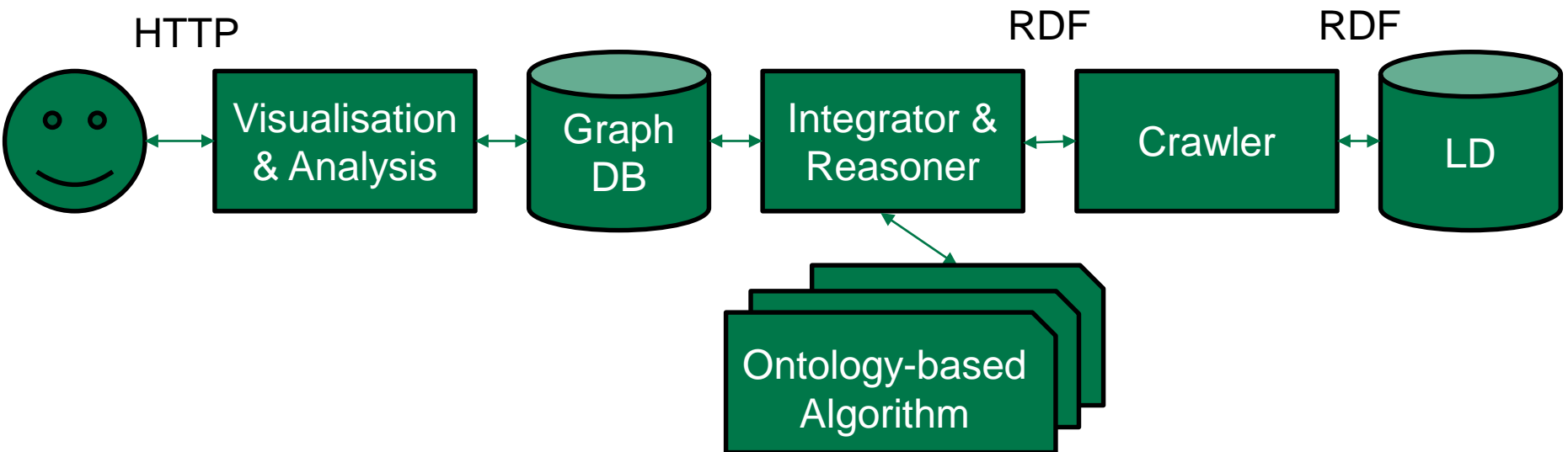
<http://people.aifb.kit.edu/bka/hla/>

- Flexible Visualisation of Similarities?



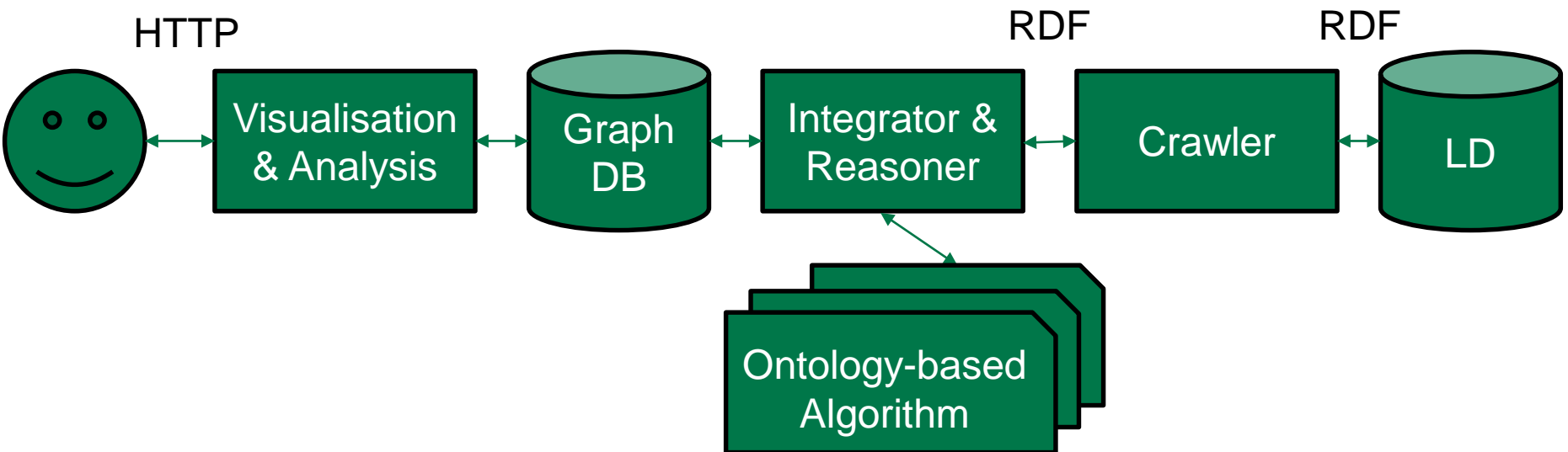
Conclusion

- A Semantic Clinical Data Warehouse should allow
 - Better results by providing more links (e.g., Bio2RDF)
 - Similarities between any objects (Pesquita, '09)
 - New algorithms (Callahan, '12)



Thanks!

- A Semantic Clinical Data Warehouse should allow
 - Better results by providing more links (e.g., Bio2RDF)
 - Similarities between any objects (Pesquita, '09)
 - New algorithms (Callahan, '12)



References

Palma, G., Vidal, M.E., Haag, E., Raschid, L., Thor, A.: Measuring Relatedness Between Scientific Entities in Annotation Datasets. In: International Conference on Bioinformatics, Computational Biology and Biomedical Informatics (2013)

Callahan, A., Dumontier, M.: Evaluating Scientific Hypotheses Using the SPARQL Inferencing Notation. In: The Semantic Web: Research and Applications (2012)

Pesquita, C., Faria, D., Falco, A.O., Lord, P., Couto, F.M.: Semantic Similarity in Biomedical Ontologies. PLOS Computational Biology 5(7) (2009)