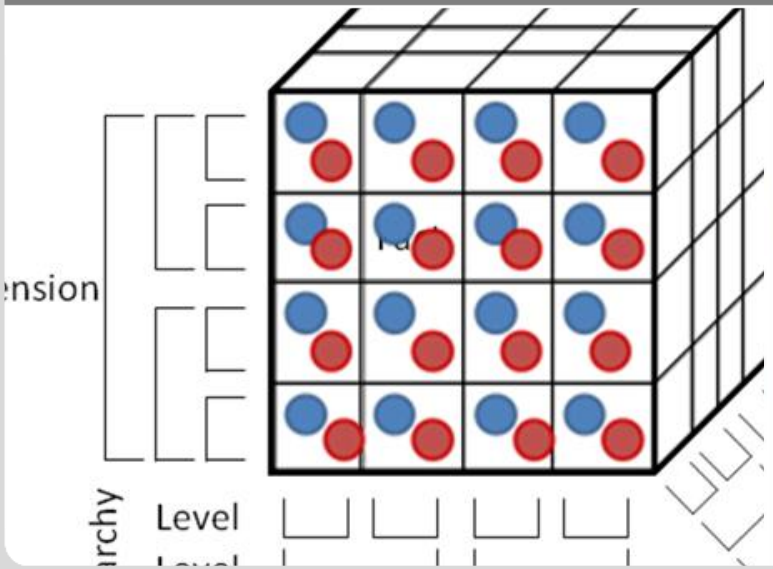
Karlsruhe Institute of Technology

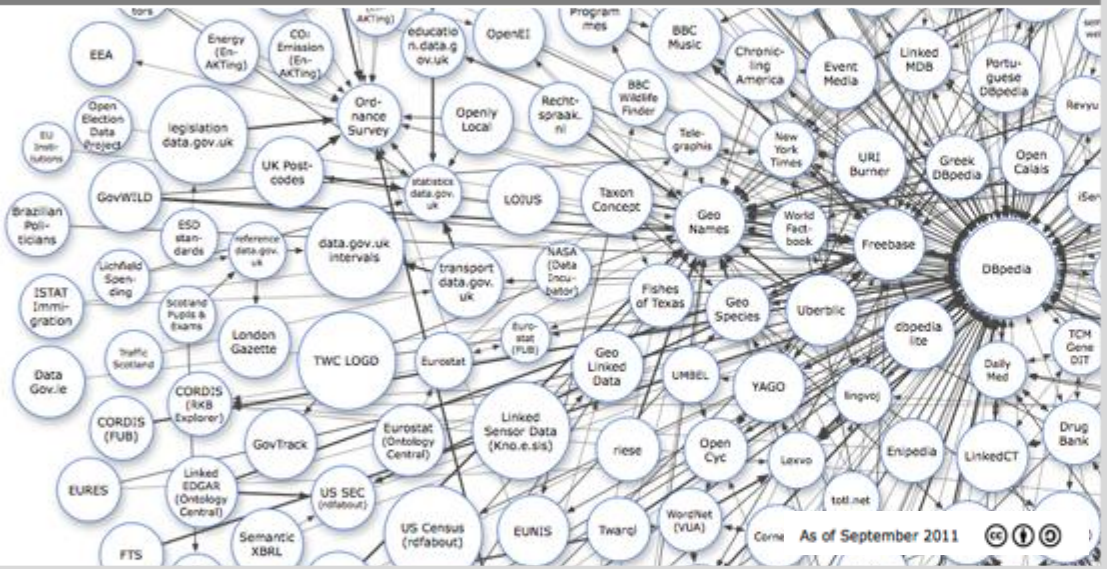
Computing the Similarity of Entities using HANA Graph and Linked Data

Benedikt Kämpgen, Christof Bornhoevd, Horst Werner

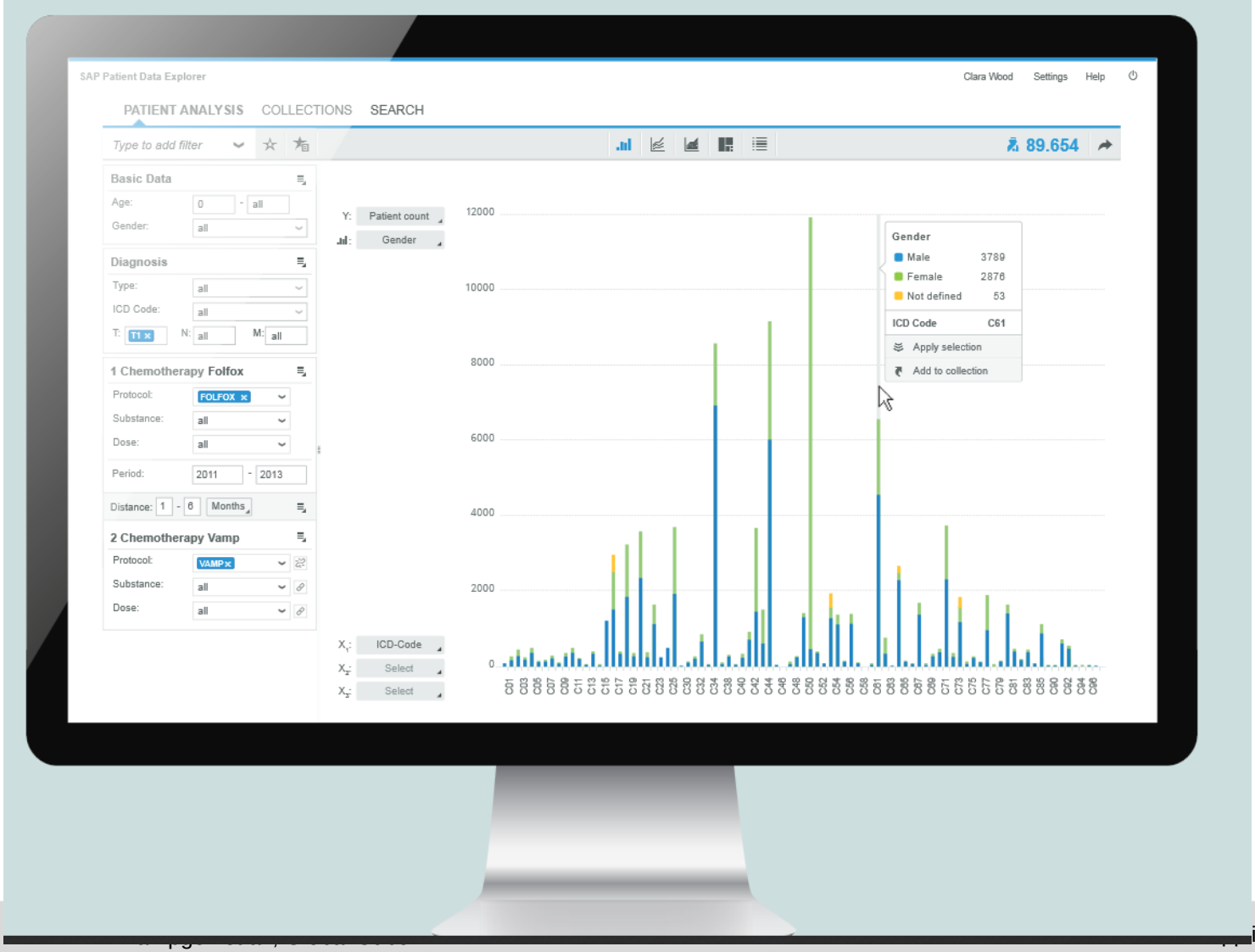
SAP

Institute of Applied Informatics and Formal Description Methods (AIFB)

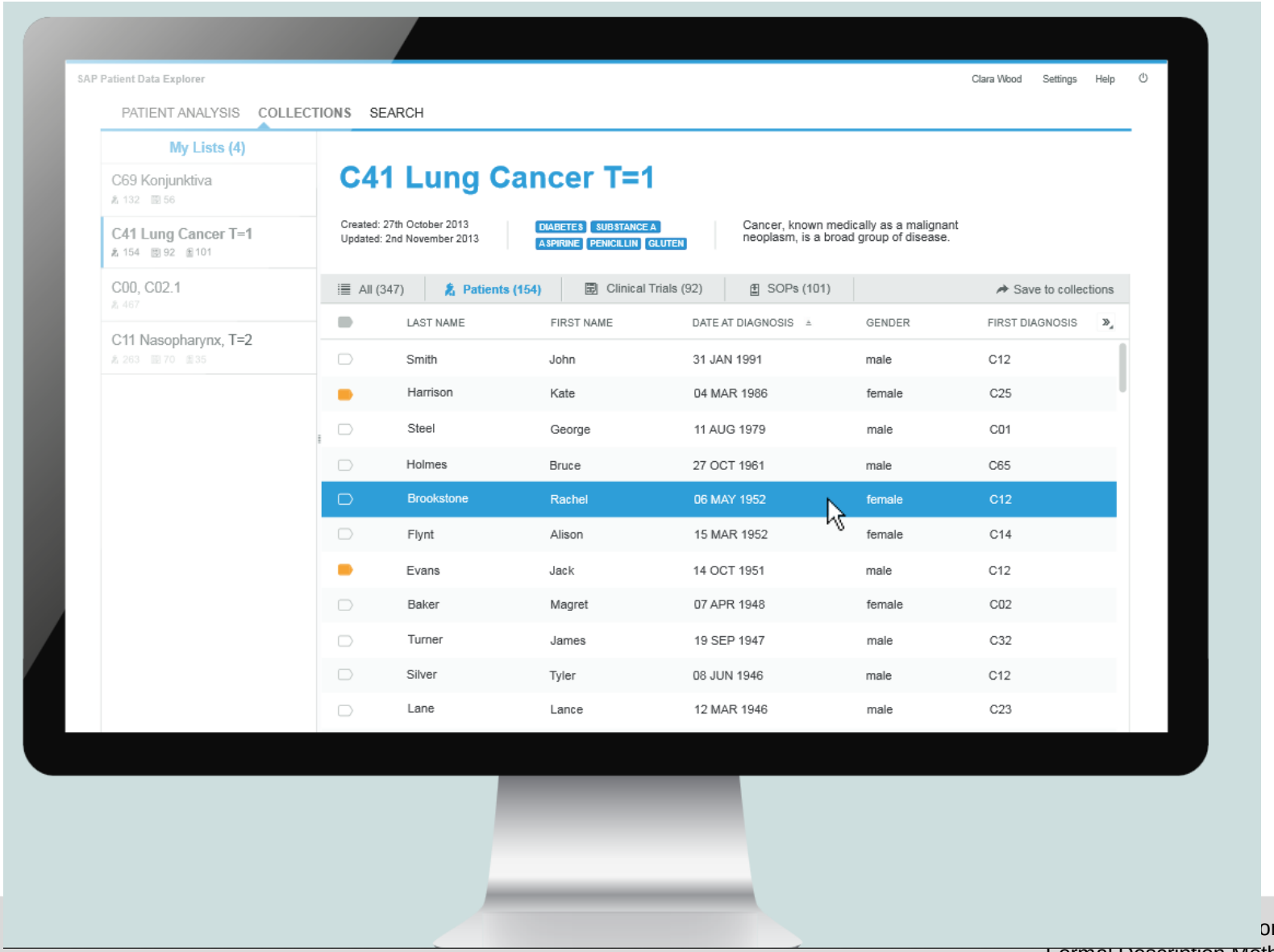




Motivation


SAP Patient Data Explorer
Clara Wood
Settings
Help
PATIENT ANALYSIS
COLLECTIONS
SEARCH
Type to add filter
89.654
Basic Data
Age:
Gender:
Diagnosis
Type:
ICD Code:
T:
N:
M:
1 Chemotherapy Folfox
Protocol:
Substance:
Dose:
Period:
Distance:
Months
2 Chemotherapy Vamp
Protocol:
Substance:
Dose:
Y:
Patient count
Gender
X1:
ICD-Code
X2:
Select
X3:
Select
12000
10000
8000
6000
4000
2000
0
Gender
Male 3789
Female 2876
Not defined 53
ICD Code C61
Apply selection
Add to collection

Motivation (2)



Motivation (3)

- More and more data sources become available that are potentially interesting for a physician:
- New data sources: Patienten - LinkedCT, MeSH, UMLS - Studien
- One needs more information about patients.
- Semantic Clinical Data Warehouse
- Another example is the increasing importance of genotype information
- Another example are logics based ontologies that may confirm or disagree with female prostate cancer
- Prostate cancer confirmed by, not confirmed by...

Motivation (4)

The Female Prostate

Milan Zaviačič and Richard J. Ablin
+ Author Affiliations

Correspondence to: Richard J. Ablin, Ph.D., Innapharma, Inc., 10 Mountainview Road, Suite 301, Upper Saddle River, NJ 07458.

Contrary to the statement by Borchert et al. *(1)* that "Women have no prostate ... ," women do have a prostate, the presence of which has clinical significance for the female and for our understanding of the expression of prostate- specific antigen (PSA) in women and its possible implications.

In 1672 the anatomist Regnier de Graaf described and illustrated a set of glands and ducts surrounding the female urethra that he called the female prostate. Subsequently, in 1880, Alexander Skene redirected attention to this structure, particularly to two paraurethral ducts (Skene's ducts) therein, and emphasized their importance in infection of the female genitalia.

Problem

- How to access?
  - Different data formats.
  - Semi-structured data.
  - …
- How to integrate and query?
  - No explicit links between entities.
  - Ad-hoc queries.
  - Graph-analysis queries.
  - …

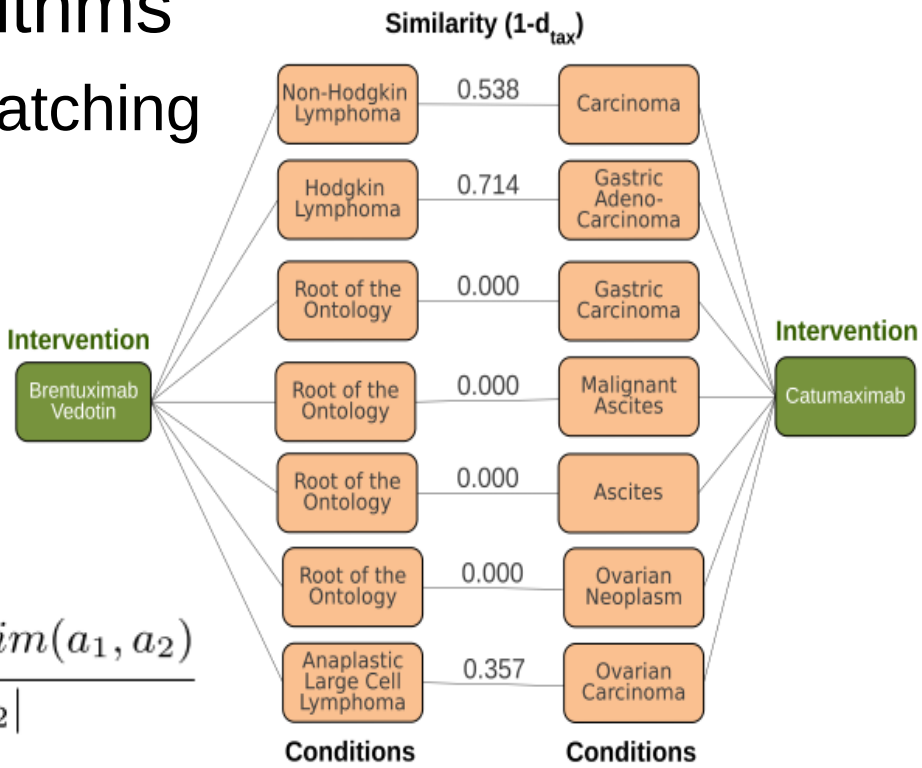
Scenario

- Similarity between entities such as genes, drugs, diseases.
- Why important?
  - Similarity as building block for pattern mining for relationships between entities.
  - E.g., drug relationships depend on similarity of the genes their target.
  - …

Palma, G., Vidal, M.-E., Haag, E., Raschid, L., & Thor, A. (2007). Measuring Relatedness Between Scientific Entities in Annotation Datasets. *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics - BCB'13*, 367–376. doi:10.1145/2506583.2506651

Scenario – AnnSim

- General data sources
  - E.g., genes, annotations, ontology
- Graph-analysis algorithms
  - Min-Weight Perfect Matching
  - …

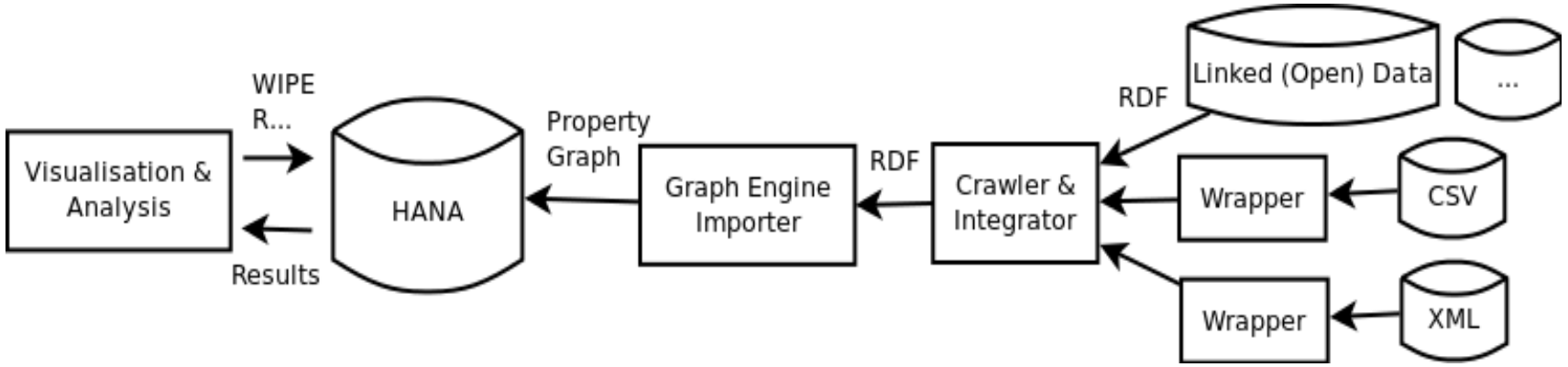


$$AnnSim(c_1, c_2) = \frac{2 \cdot \sum_{(a_1,a_2) \in WEr} sim(a_1, a_2)}{|A_1| + |A_2|}$$

Related Work

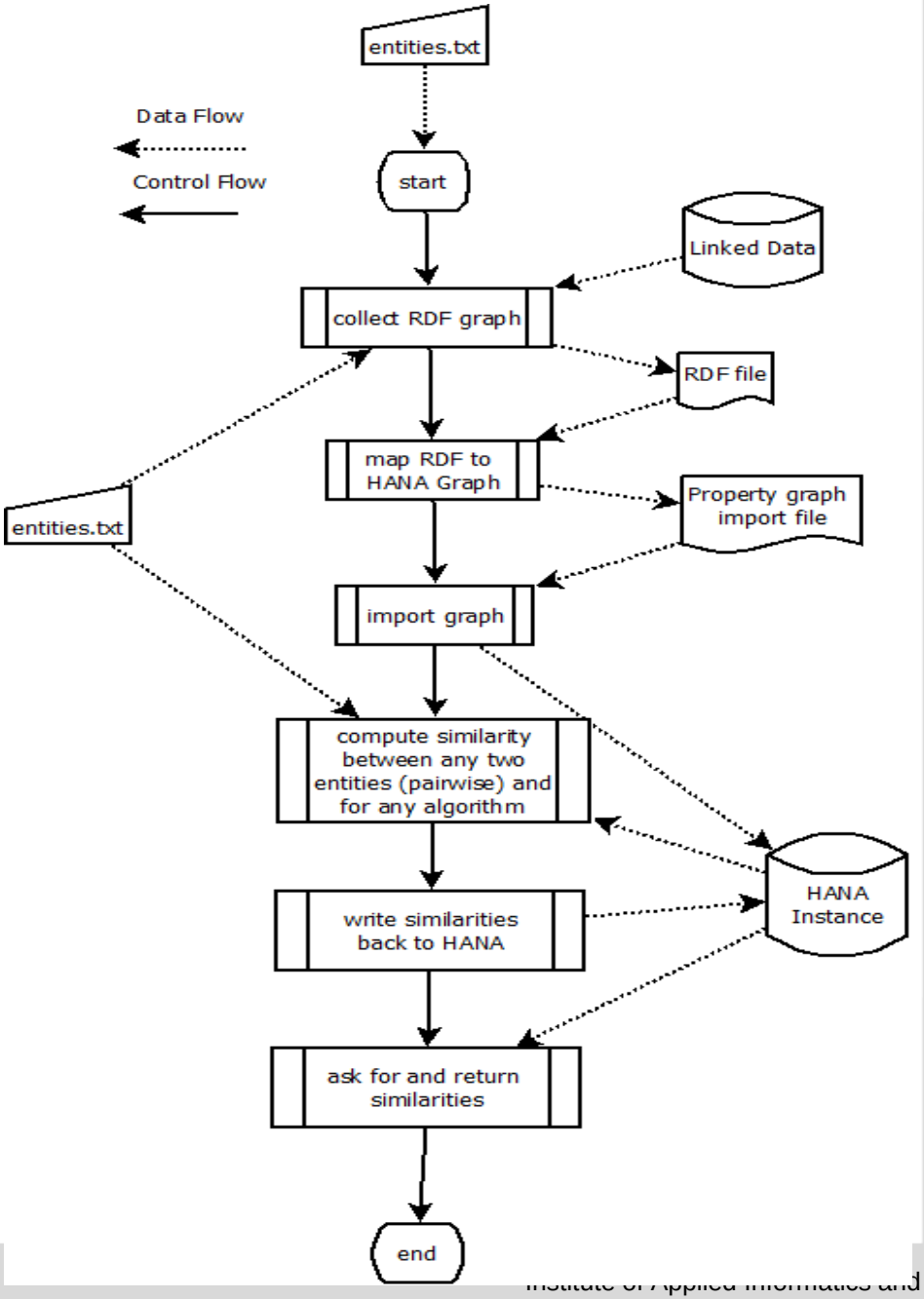
- Palma et al.
  - Hard-coded data sources
  - All pure C++

Approach


Linked (Open) Data
...
RDF
WIPE
R...
Property
Graph
RDF
Visualisation &
Analysis
HANA
Graph Engine
Importer
Crawler &
Integrator
Wrapper
CSV
Wrapper
XML
Results

HANA-LD-AnnSim (HLA)

- Input: List of entities
- Output: List of similarities



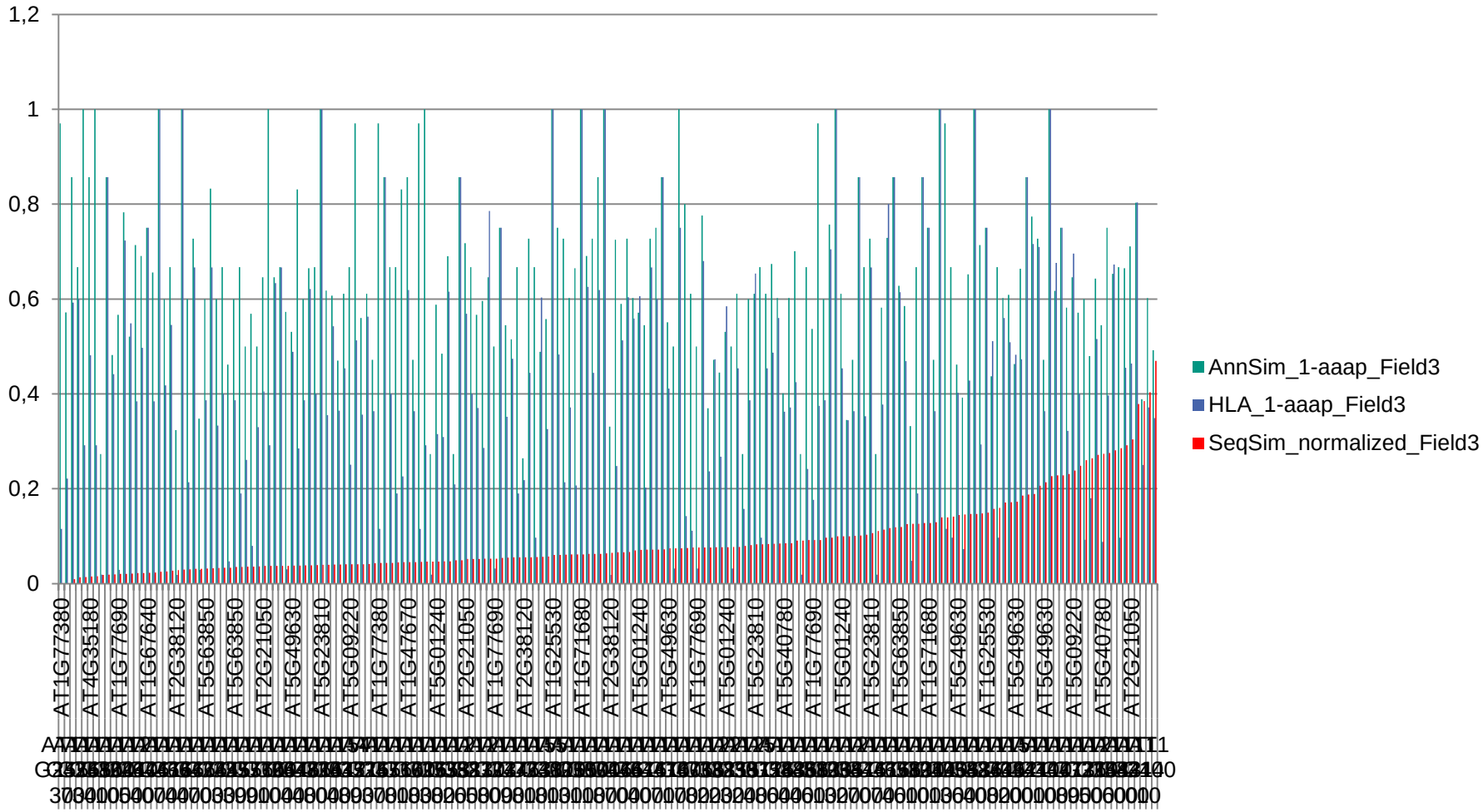
Benefits of HANA Graph

Benefits of Linked Data

Evaluation

- Correctness
- Performance
- Flexibility

Correctness



Correctness (2)

- Mean Squared Error ($\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2.$)
  - HLA – AnnSim 0.09325386
  - AnnSim – SeqSim 0.356347132
  - HLA – SeqSim 0.18890245
- Differences explained
  - AnnSim/HLA only using annotation information.
  - SeqSim Gold Standard based on DNA sequence.
  - HLA uses newer Gene Ontology than AnnSim.

MSE Wikipedia definition

Performance – Setup

	HLA	AnnSim
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 XXX	-
Triples	7,337,447	-
Size of data	537 MB	2.79 MB + 670 B
Vertices	601,519	39,209
Edges	1,658,322	74,123
Genes	20 (1-aaap)	20 (1-aaap)

Performance – Results

	HLA	AnnSim
Collect RDF graph	*641s	-
Map RDF to HANA	565s	-
Import graph	60s	-
Compute similarities	8,212s	408s
-> GEM Ask Queries	?	-
-> GEM Update Queries	?	-
-> Get Distances	?	-
-> Program Logic	?	408s

*Estimation with 6.7 Mbps

http://techcrunch.com/2012/08/09/akamai-global-average-broadband-speeds-up-by-25-u-s-up-29-to-6-7-mbps/

Performance Explained

	HLA	AnnSim
Ask for all annotating concepts of entity	Given (0s)	WIPE Traversal Query
Ask for all ancestors of a concept	Depth first search through graph	WIPE Traversal Query
Ask for depth of all concepts	Topological sort of vertices in graph	WIPE Query for root vertices; breadth first from root vertices.
Ask for lowest-common-ancestor of two concepts	Compare ancestors and their depths.	Compare ancestors and their depths.
Shortest path from lca to concept.	Depth first search through graph from lca.	Take difference between depth of concept and lca.*
Shortest path from root to concept.	Depth first search from root.	Depth of concept + 1.*
Compute d_tax	Program logic formula.	Program logic formula.
Compute AnnSim	Program logic formula.	Program logic formula.

*heuristic

Flexibility

- We would probably need to change the path from an entity to its concepts (we could also make it more general)
  - Can we show that with anything annotated with something? (drugs were annotated by experts; we could translate them to RDF). Anything linked with NCIT terms?

Lessons Learned

Institute of Applied Informatics and
Formal Description Methods (AIFB)

Conclusions & Open Work

- Evaluating AnnSim on other entity types.
- Improving AnnSim to better similarity.
  - Not only certain concept annotations.
  - Not only concept annotations.
- Extending AnnSim to other relationship types.