INODE - Intelligent Open Data Exploration

Kurt Stockinger
Zurich University of Applied Sciences
Switzerland

SNTA Workshop @ 30th International Symposium on High-Performance Parallel and Distributed Computing

June 21, 2021

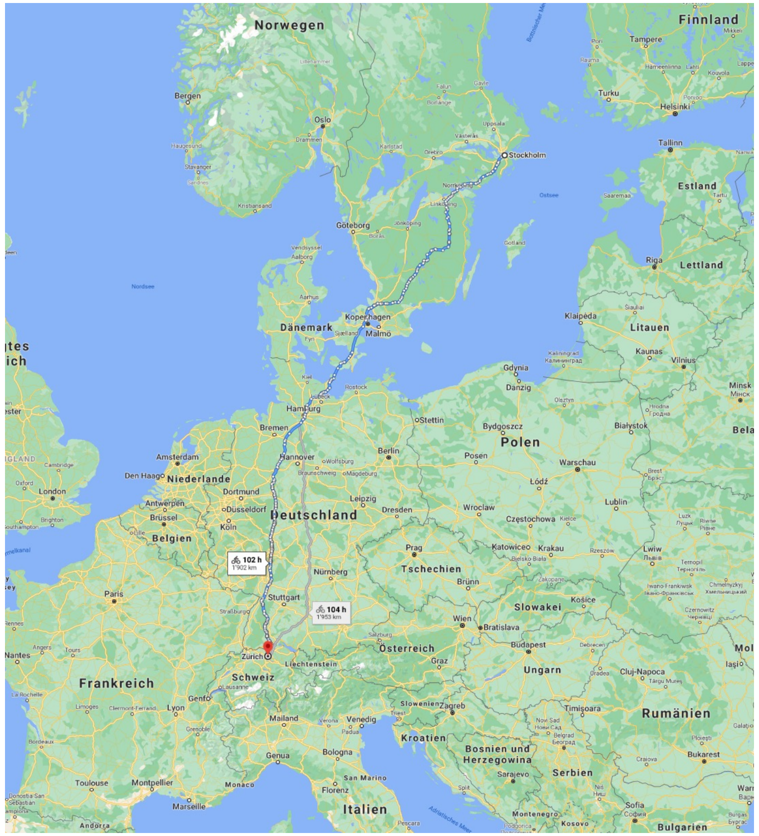
Zurich University of Applied Sciences ZHAW

Switzerland’s biggest fully-featured university of applied sciences

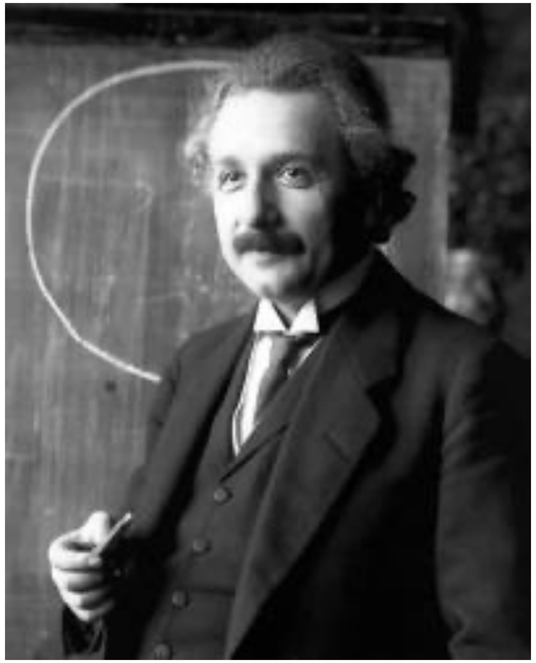
Zurich University of Applied Sciences ZHAW

Switzerland’s biggest fully-featured university of applied sciences

From Stockholm, Sweden to Zurich, Switzerland it’s 1,902 km and it takes 102 hours by bike



Our Most Famous Lecturer

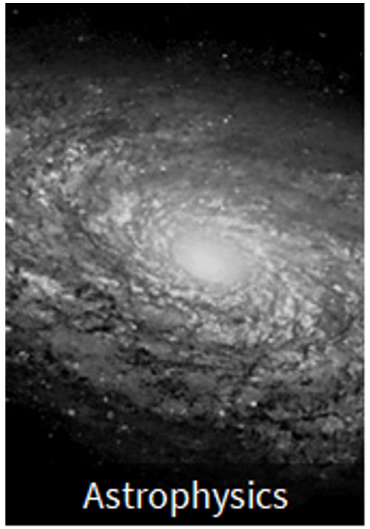


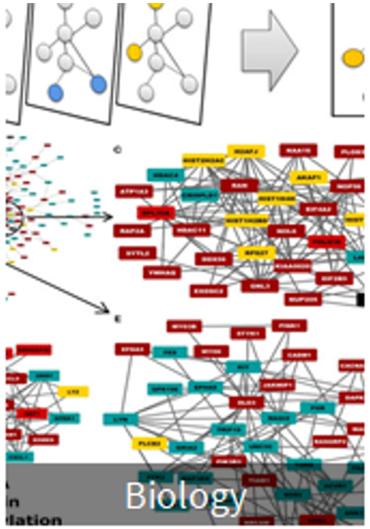
1901: Albert Einstein

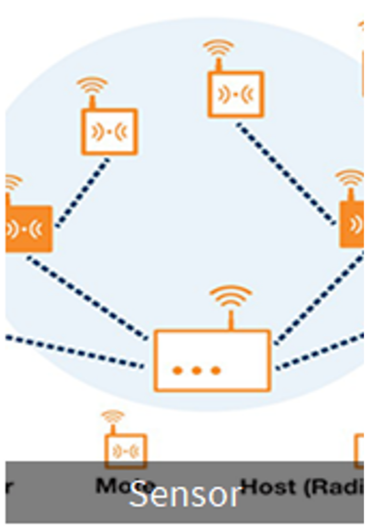
Outline

- The Data Promise and the Problem
- Need for Novel Tools to Explore Data
- Experience in Building Systems for Intelligent Data Exploration
  - Bio-SODA: Natural Language to SPARQL without Neural Networks
  - ValueNet: Natural Language to SQL with Neural Networks

The Data Promise







- Many different data sets are generated by users, systems and sensors
- Many processes are increasingly more data-driven
- Many aspects of our lives are in fact more data-driven

Data is the new oil … but we need the right tools to leverage it!

Source: Georgia Koutrika, Athena Research Center

Data-Intensive Use Cases

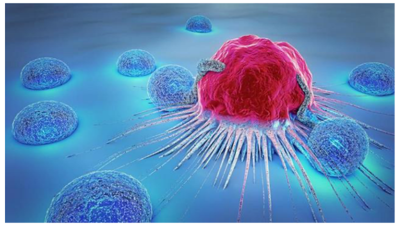
Need for novel approaches and tools to access and understand data

- **Astrophysics:**
  - Massive amounts of data about galaxies in relational databases to study star formations

  ▶ Understanding hundreds of database tables with thousands of attributes is very hard

- **Cancer Biomarker Research:**
  - Very complex data sets to allow integration of cancer biomarkers to study cancer types
  - ~50,000 human genes and ~3 million base pairs

  ▶ Datasets are very hard to analyze for domain experts



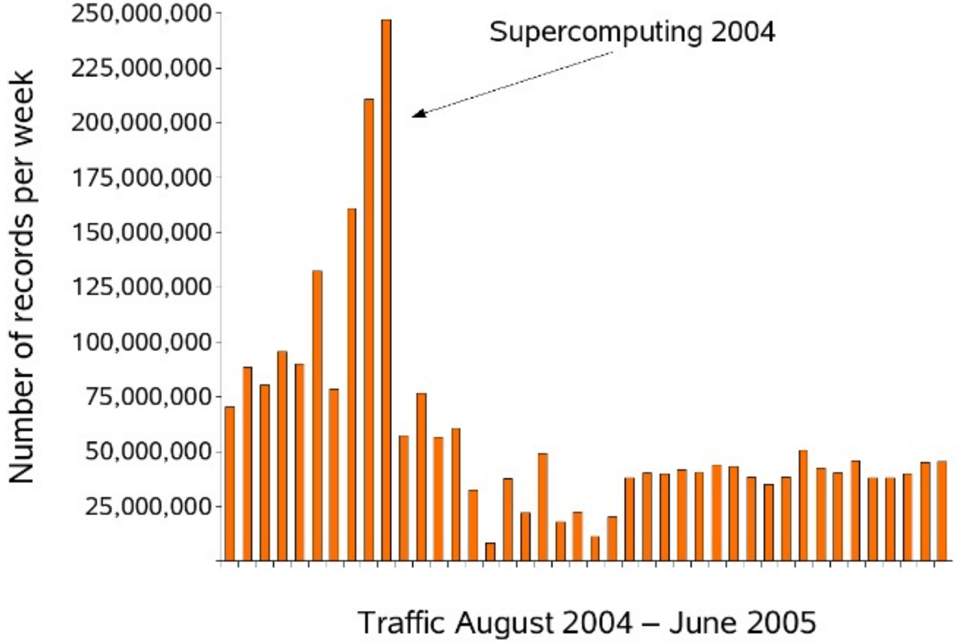
- **Research & Innovation Policy Making:**
  - Many heterogeneous database of EU projects

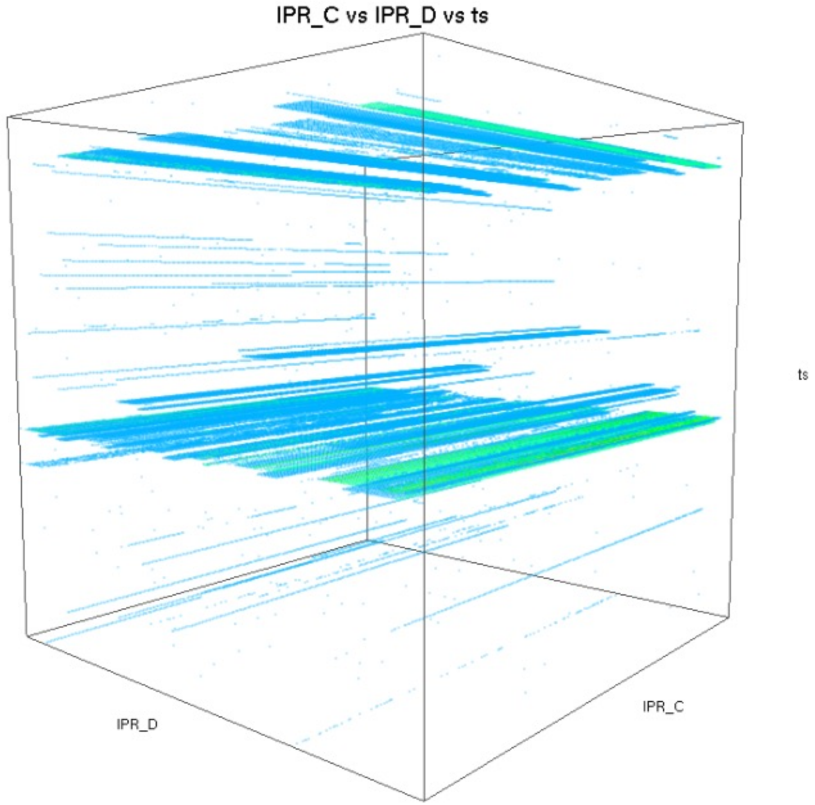
  ▶ Datasets are hard to analyze and understand for non-technical experts

European Commission | CORDIS EU research results

Why is this relevant for the distributed systems or network community?

Network Traffic Analysis at Supercomputing in 2006





Stockinger, K., Bethel, E. W., Campbell, S., Dart, E., & Wu, K. (2006). Detecting Distributed Scans Using High-Performance Query-Driven Visualization. In *SC Conference*.

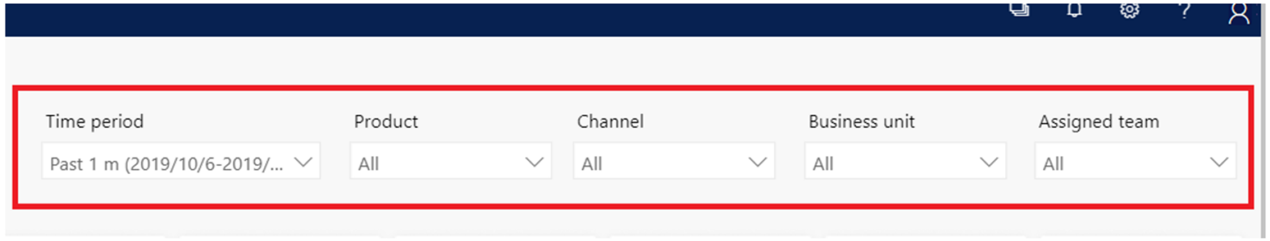
Network Traffic Analysis

- Very data-intensive
- Need smart way of analyzing data to avoid intrusions
- Problems are “similar” to other data-intensive disciplines

Limitations of Existing Data Exploration Tools

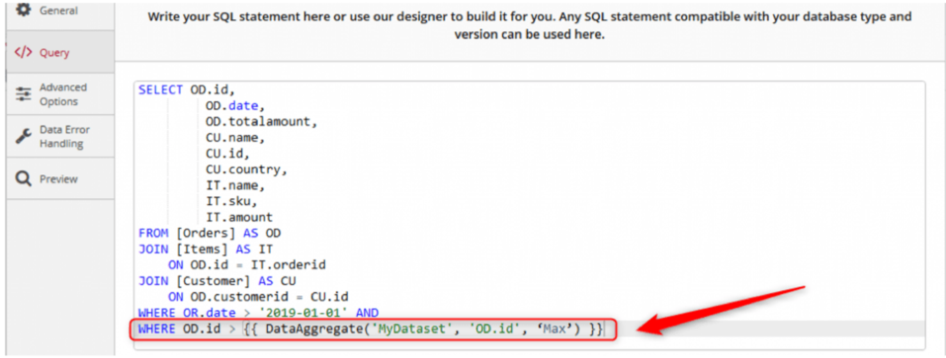
1. INPUT

Form-based interfaces



Time period	Product	Channel	Business unit	Assigned team
Past 1 m (2019/10/6-2019/...	All	All	All	All

Low-level query interfaces



Write your SQL statement here or use our designer to build it for you. Any SQL statement compatible with your database type and version can be used here.

```
SELECT OD.id,
         OD.date,
         OD.totalamount,
         CU.name,
         CU.id,
         CU.country,
         IT.name,
         IT.sku,
         IT.amount
FROM [Orders] AS OD
JOIN [Items] AS IT
    ON OD.id = IT.orderid
JOIN [Customer] AS CU
    ON OD.customerid = CU.id
WHERE OR.date > '2019-01-01' AND
WHERE OD.id > {{ DataAggregate('MyDataset', 'OD.id', 'Max') }}
```

- Limited Query Exploration Capabilities

- Knowledge of SQL (or SPARQL, etc)
- Knowledge of the database schema
- Well-formed information needs

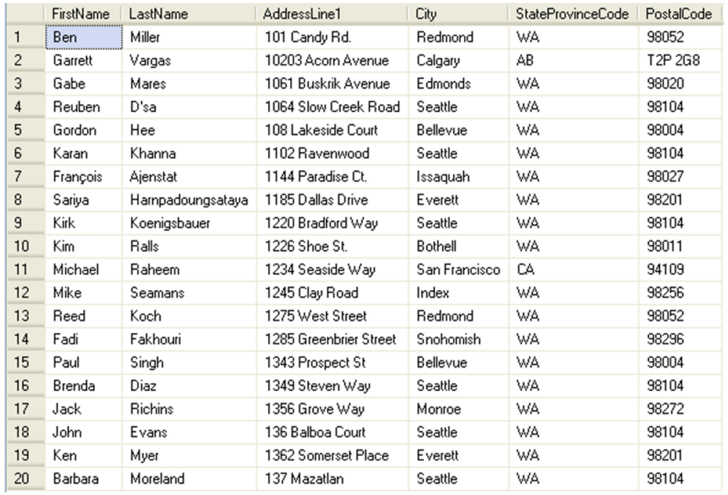
[1]SQL = Structured Query Language for relational databases
[2]SPARQL = SPARQL Protocol and RDF Query Language for graph databases

Source: Georgia Koutrika, Athena Research Center

Limitations of Existing Data Exploration Tools

2. OUTPUT

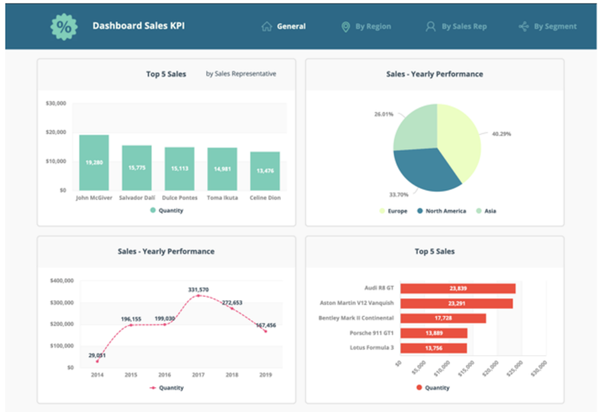
Tables



	FirstName	LastName	AddressLine1	City	StateProvinceCode	PostalCode
1	Ben	Miller	101 Candy Rd.	Redmond	WA	98052
2	Garrett	Vargas	10203 Acorn Avenue	Calgary	AB	T2P 2G8
3	Gabe	Mares	1061 Buskrik Avenue	Edmonds	WA	98020
4	Reuben	D'sa	1064 Slow Creek Road	Seattle	WA	98104
5	Gordon	Hee	108 Lakeside Court	Bellevue	WA	98004
6	Karan	Khanna	1102 Ravenwood	Seattle	WA	98104
7	François	Ajenstat	1144 Paradise Ct.	Issaquah	WA	98027
8	Sariya	Harnpadoungsataya	1185 Dallas Drive	Everett	WA	98201
9	Kirk	Koenigsbauer	1220 Bradford Way	Seattle	WA	98104
10	Kim	Ralls	1226 Shoe St.	Bothell	WA	98011
11	Michael	Raheem	1234 Seaside Way	San Francisco	CA	94109
12	Mike	Seamans	1245 Clay Road	Index	WA	98256
13	Reed	Koch	1275 West Street	Redmond	WA	98052
14	Fadi	Fakhouri	1285 Greenbrier Street	Snohomish	WA	98296
15	Paul	Singh	1343 Prospect St	Bellevue	WA	98004
16	Brenda	Diaz	1349 Steven Way	Seattle	WA	98104
17	Jack	Richins	1356 Grove Way	Monroe	WA	98272
18	John	Evans	136 Balboa Court	Seattle	WA	98104
19	Ken	Myer	1362 Somerset Place	Everett	WA	98201
20	Barbara	Moreland	137 Mazatlan	Seattle	WA	98104

- No interpretation of results
- No explanation of system choices/answers
- No clue how to proceed next

Reports/dashboards

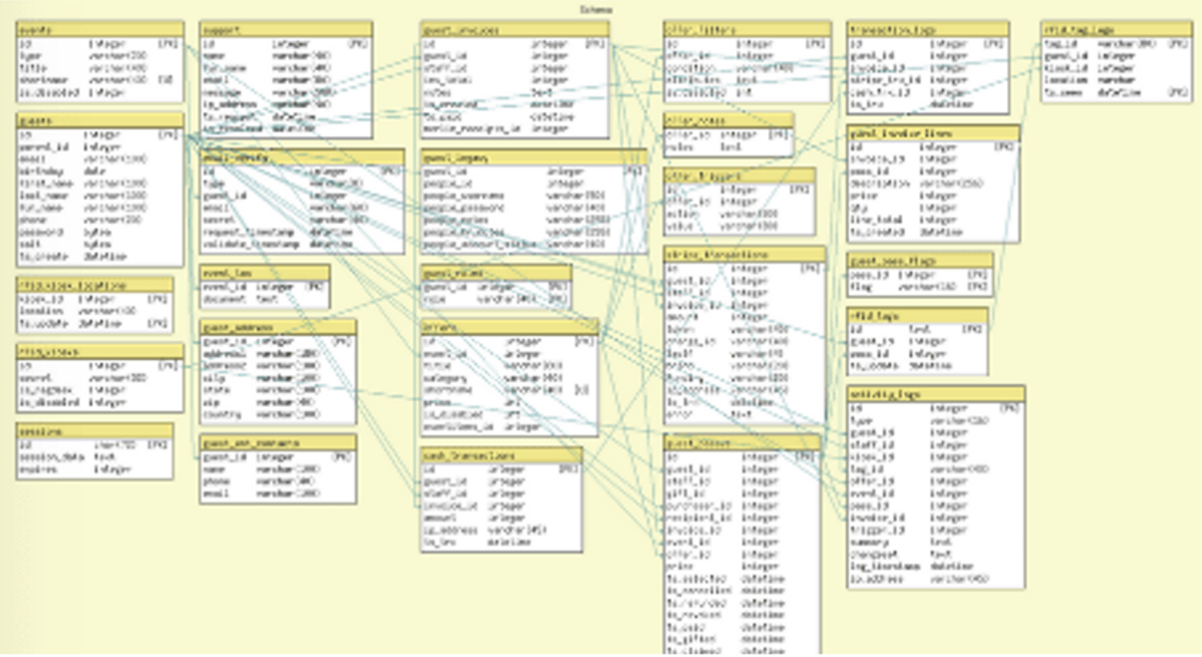


Source: Georgia Koutrika, Athena Research Center

Limitations of Existing Data Exploration Tools

3. DATA

Static, Known Schema



- Hard to find new related sources
- Hard to link and query new related sources

Source: Georgia Koutrika, Athena Research Center

Needs of Scientists and Business Analysts

Need for novel approaches and tools to access and understand data

Astrophysics

Make the interface to data as natural as possible to allow us formulate our scientific questions .

Cancer Biomarker Research

Allow users to query other annotation sources

Enable ontology-based data integration and access

Research and Innovation Policy Making

Enable powerful queries by even non-technical users, who are the majority in R&I use cases.

Source: Georgia Koutrika, Athena Research Center

Requirements/Challenges

1. **Data accessibility**:

   Enable more natural interfaces to data.

1. **User guidance**:
   Offer guidance to users to understand the data and formulate the right queries.

1. **Data discovery and linking**:

   Allow linking and combining data sets to generate rich information and insights.

Outline

- The Data Promise and the Problem
- Need for Novel Tools to Explore Data
- Experience in Building Systems for Intelligent Data Exploration
    - Bio-SODA: Natural Language to SPARQL without Neural Networks
    - ValueNet: Natural Language to SQL with Neural Networks

Natural Language Interfaces to Data: Building Data Systems with Academia and Industry

- SODA – Search Over Data Warehouse:
  - ("Future ZHAW employee" + Credit Suisse + ETH Zurich)
  - Accessing business data warehouses in natural language

- Bio-SODA:
  - (ZHAW + Swiss Institute of Bioinformatics)
  - Accessing bioinformatics databases in natural language

- NQuest - Natural Language Query Exploration System:
  - (ZHAW + Zurich Startup Veezoo)
  - Accessing databases and (partially) machine learning in natural language

- INODE – Intelligent Open Data Exploration System
  - (ZHAW + 8 partners in Europe)
  - Exploring structured and unstructured data in natural language

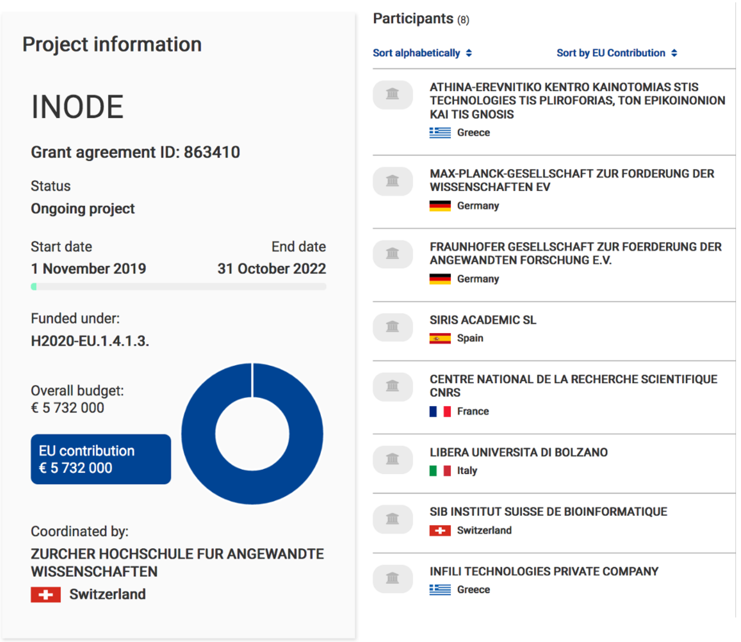
References are given after the conclusions

INODE – Intelligent Open Data Exploration

http://www.inode-project.eu/

- Users should **interact with data in a more dialectic** and intuitive way similar to a **dialog with a human**

- Services for exploration of open data sets that help users:
  - **Link** and leverage **multiple datasets**
  - Access and **search data using natural language**, using examples and using analytics
  - **Get guidance from the system** in understanding the data and formulating the right queries
  - Explore data and discover new insights through **visualizations**
  - Focus on **Astrophysics, Cancer Biomarker Research** and **Research & Innovation Policy Making**



Project information

INODE

Grant agreement ID: 863410

Status
Ongoing project

Start date	End date
1 November 2019	31 October 2022

Funded under:
H2020-EU.1.4.1.3.

Overall budget:
€ 5 732 000

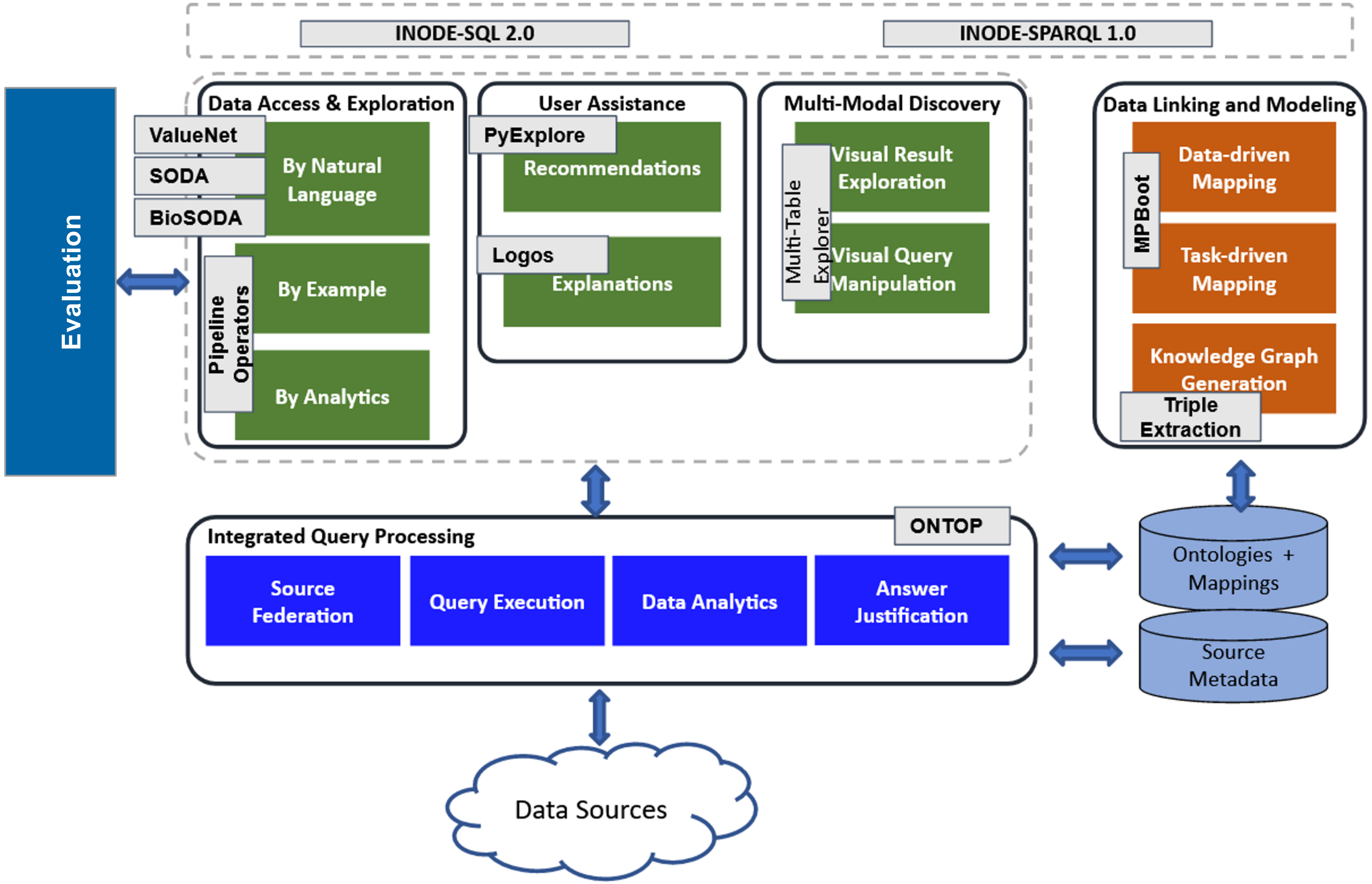
EU contribution
€ 5 732 000

Coordinated by:
ZURCHER HOCHSCHULE FUR ANGEWANDTE WISSENSCHAFTEN
Switzerland

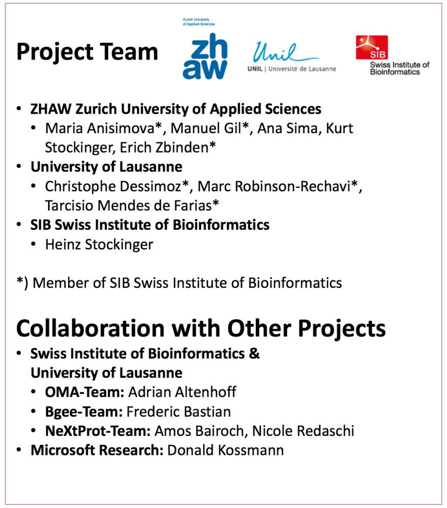
Participants (8)

- ATHINA-EREVNITIKO KENTRO KAINOTOMIAS STIS TECHNOLOGIES TIS PLIROFORIAS, TON EPIKOINONION KAI TIS GNOSIS — Greece
- MAX-PLANCK-GESELLSCHAFT ZUR FORDERUNG DER WISSENSCHAFTEN EV — Germany
- FRAUNHOFER GESELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHUNG E.V. — Germany
- SIRIS ACADEMIC SL — Spain
- CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE CNRS — France
- LIBERA UNIVERSITA DI BOLZANO — Italy
- SIB INSTITUT SUISSE DE BIOINFORMATIQUE — Switzerland
- INFILI TECHNOLOGIES PRIVATE COMPANY — Greece

INODE Architecture



Bio-SODA: Building a Natural Language-to-SPARQL System without Neural Networks



Project Team

- **ZHAW Zurich University of Applied Sciences**
    - Maria Anisimova*, Manuel Gil*, Ana Sima, Kurt Stockinger, Erich Zbinden*
- **University of Lausanne**
    - Christophe Dessimoz*, Marc Robinson-Rechavi*, Tarcisio Mendes de Farias*
- **SIB Swiss Institute of Bioinformatics**
    - Heinz Stockinger

*) Member of SIB Swiss Institute of Bioinformatics

Collaboration with Other Projects

- **Swiss Institute of Bioinformatics & University of Lausanne**
    - **OMA-Team:** Adrian Altenhoff
    - **Bgee-Team:** Frederic Bastian
    - **NeXtProt-Team:** Amos Bairoch, Nicole Redaschi
- **Microsoft Research:** Donald Kossmann

The Current Way of Querying Graph Databases in Bioinformatics

Assume that we have a graph database about drugs and diseases
A typical question could be:

- What are the drugs for diseases associated with the brca[1] genes?

- Answering the question would require the following SPARQL[2] query:

SPARQL query:

```
SELECT DISTINCT ?diseases ?diseases_label ?drugs ?drugs_label ?genes ?genes_label WHERE {
        ?drugs <http://www.w3.org/2000/01/rdf-schema#label> ?drugs_label.
        ?diseases <http://www.w3.org/2000/01/rdf-schema#label> ?diseases_label.
        ?drugs a <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/drugs>.
        ?diseases a <http://www4.wiwiss.fu-berlin.de/diseasome/resource/diseasome/diseases>.
        ?drugs <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/possibleDiseaseTarget> ?diseases.
        ?diseases <http://www4.wiwiss.fu-berlin.de/diseasome/resource/diseasome/associatedGene> ?genes.
        ?genes <http://www.w3.org/2000/01/rdf-schema#label> ?genes_label.
        FILTER (contains(lcase(str(?genes_label)), "brca"))
        }
```

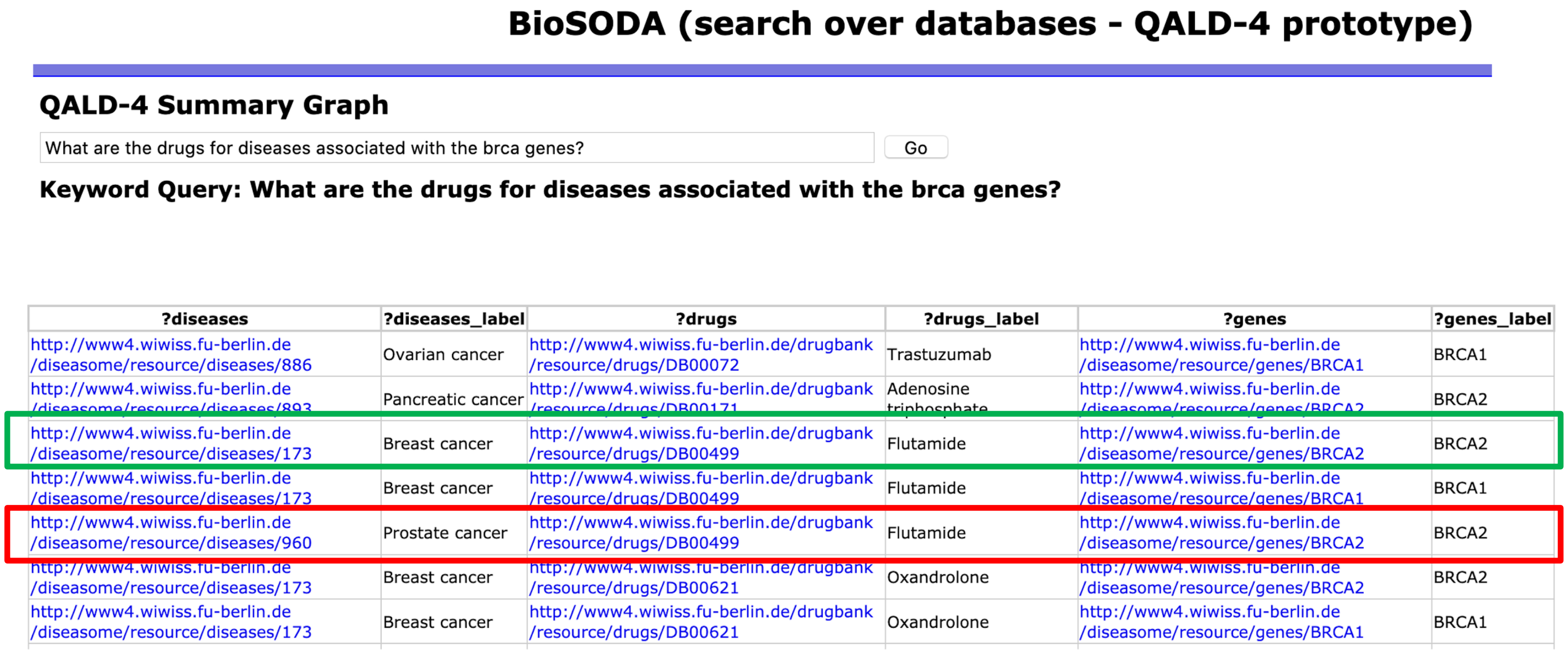
[1] brca refers to breast cancer
[2] SPARQL = SPARQL Protocol and RDF Query Language for graph databases

INODE

The Bio-SODA Way of Querying Graph Databases

Question

Answer



BioSODA (search over databases - QALD-4 prototype)

QALD-4 Summary Graph

What are the drugs for diseases associated with the brca genes? Go

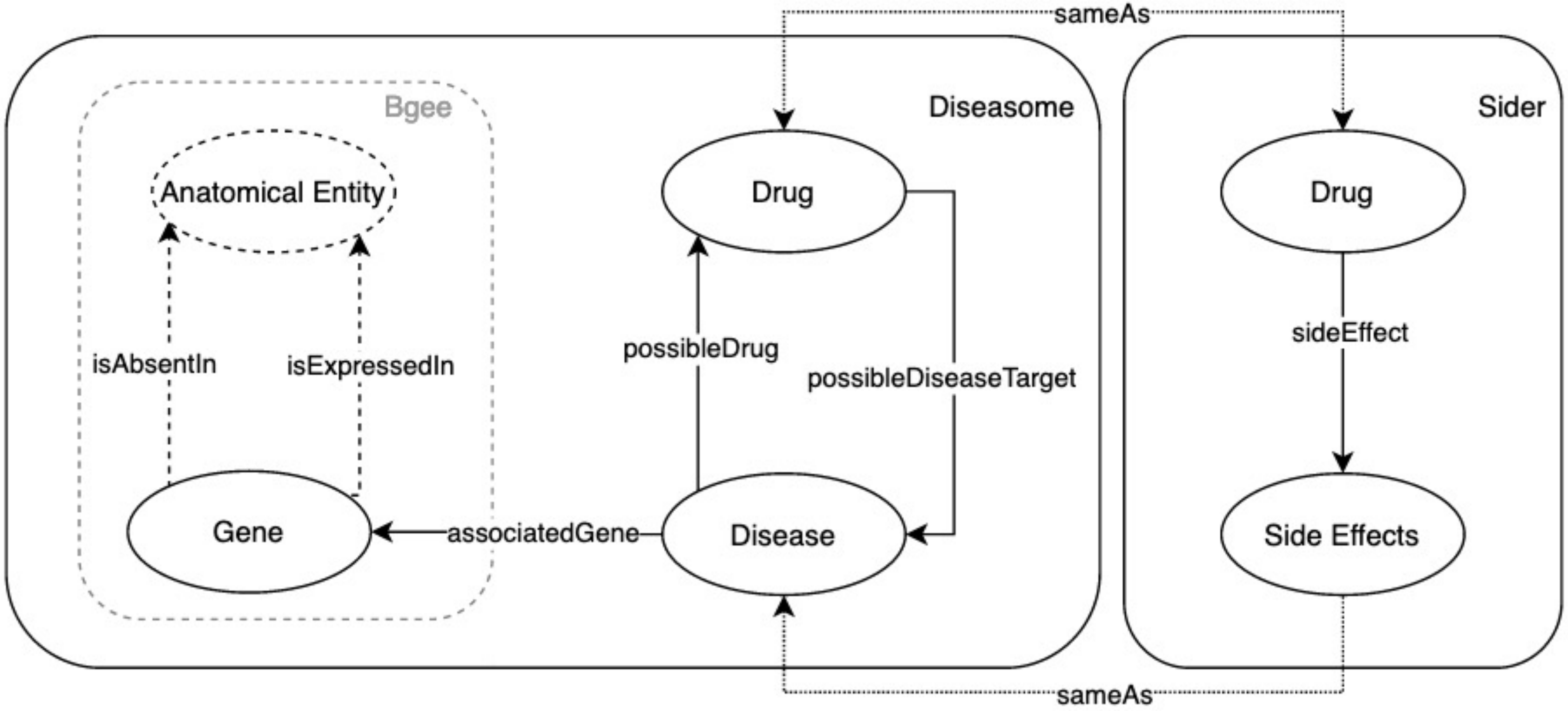
Keyword Query: What are the drugs for diseases associated with the brca genes?

?diseases	?diseases_label	?drugs	?drugs_label	?genes	?genes_label
http://www4.wiwiss.fu-berlin.de /diseasome/resource/diseases/886	Ovarian cancer	http://www4.wiwiss.fu-berlin.de/drugbank /resource/drugs/DB00072	Trastuzumab	http://www4.wiwiss.fu-berlin.de /diseasome/resource/genes/BRCA1	BRCA1
http://www4.wiwiss.fu-berlin.de /diseasome/resource/diseases/893	Pancreatic cancer	http://www4.wiwiss.fu-berlin.de/drugbank /resource/drugs/DB00171	Adenosine triphosphate	http://www4.wiwiss.fu-berlin.de /diseasome/resource/genes/BRCA2	BRCA2
http://www4.wiwiss.fu-berlin.de /diseasome/resource/diseases/173	Breast cancer	http://www4.wiwiss.fu-berlin.de/drugbank /resource/drugs/DB00499	Flutamide	http://www4.wiwiss.fu-berlin.de /diseasome/resource/genes/BRCA2	BRCA2
http://www4.wiwiss.fu-berlin.de /diseasome/resource/diseases/173	Breast cancer	http://www4.wiwiss.fu-berlin.de/drugbank /resource/drugs/DB00499	Flutamide	http://www4.wiwiss.fu-berlin.de /diseasome/resource/genes/BRCA1	BRCA1
http://www4.wiwiss.fu-berlin.de /diseasome/resource/diseases/960	Prostate cancer	http://www4.wiwiss.fu-berlin.de/drugbank /resource/drugs/DB00499	Flutamide	http://www4.wiwiss.fu-berlin.de /diseasome/resource/genes/BRCA2	BRCA2
http://www4.wiwiss.fu-berlin.de /diseasome/resource/diseases/173	Breast cancer	http://www4.wiwiss.fu-berlin.de/drugbank /resource/drugs/DB00621	Oxandrolone	http://www4.wiwiss.fu-berlin.de /diseasome/resource/genes/BRCA2	BRCA2
http://www4.wiwiss.fu-berlin.de /diseasome/resource/diseases/173	Breast cancer	http://www4.wiwiss.fu-berlin.de/drugbank /resource/drugs/DB00621	Oxandrolone	http://www4.wiwiss.fu-berlin.de /diseasome/resource/genes/BRCA1	BRCA1

[1] QALD-4: Benchmark for Question Answering over Linked Data

.Sima, A. C., de Farias, T. M., Anisimova, M., Dessimoz, C., Robinson-Rechavi, M., Zbinden, E., & Stockinger, K. (2021). Bio-SODA: Enabling Natural Language Question Answering over Knowledge Graphs without Training Data., International Conference on Scientific and Statistical Database Management (SSDBM), 2021

The Graph Data Model



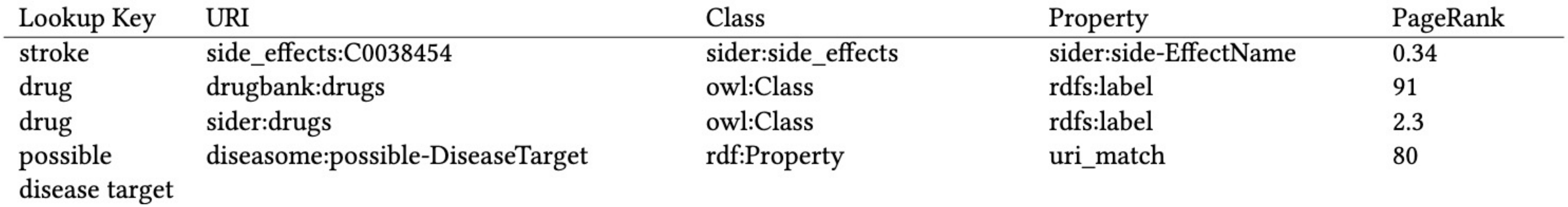
Intuition for Translating from Natural Language to SPARQL

- We can use basic concepts of information retrieval to search the search:

  - Build inverted index on all data stored in the database
  - Use the inverted index to find matches between the query and the database
  - Use the graph structure to find connections between data

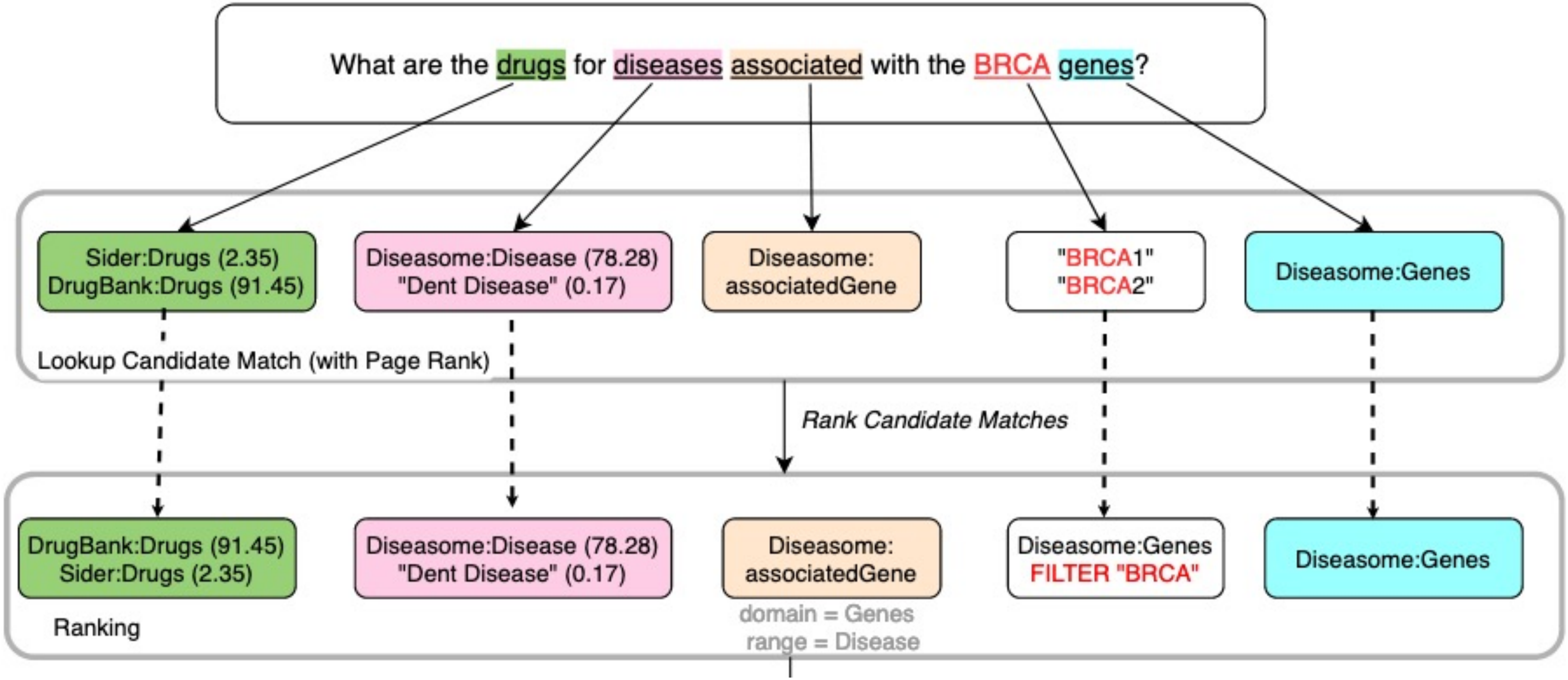
- Essentially the translation task is a pattern matching problem – no learning required

Example of an Inverted Index

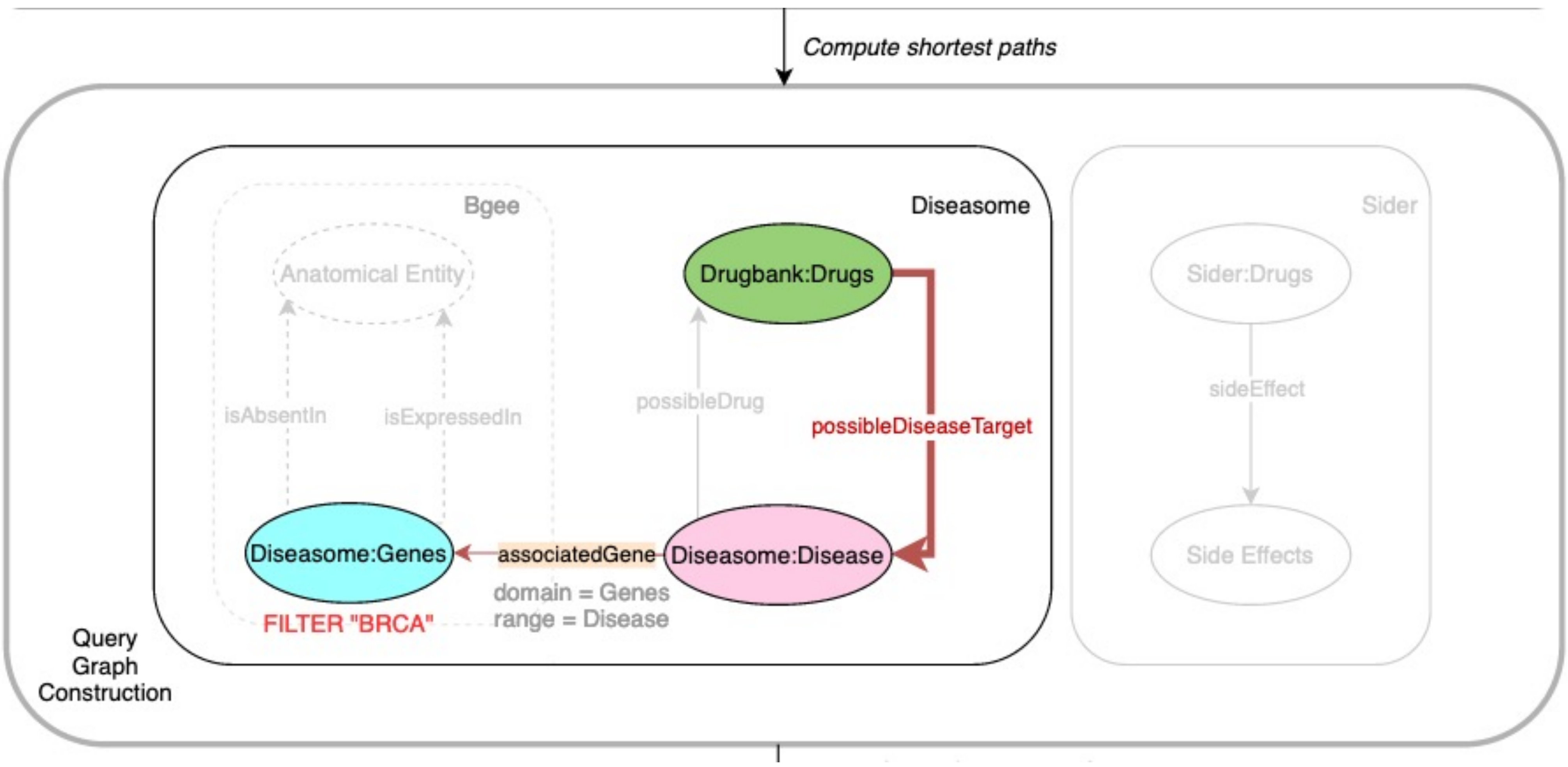


Lookup Key	URI	Class	Property	PageRank
stroke	side_effects:C0038454	sider:side_effects	sider:side-EffectName	0.34
drug	drugbank:drugs	owl:Class	rdfs:label	91
drug	sider:drugs	owl:Class	rdfs:label	2.3
possible disease target	diseasome:possible-DiseaseTarget	rdf:Property	uri_match	80

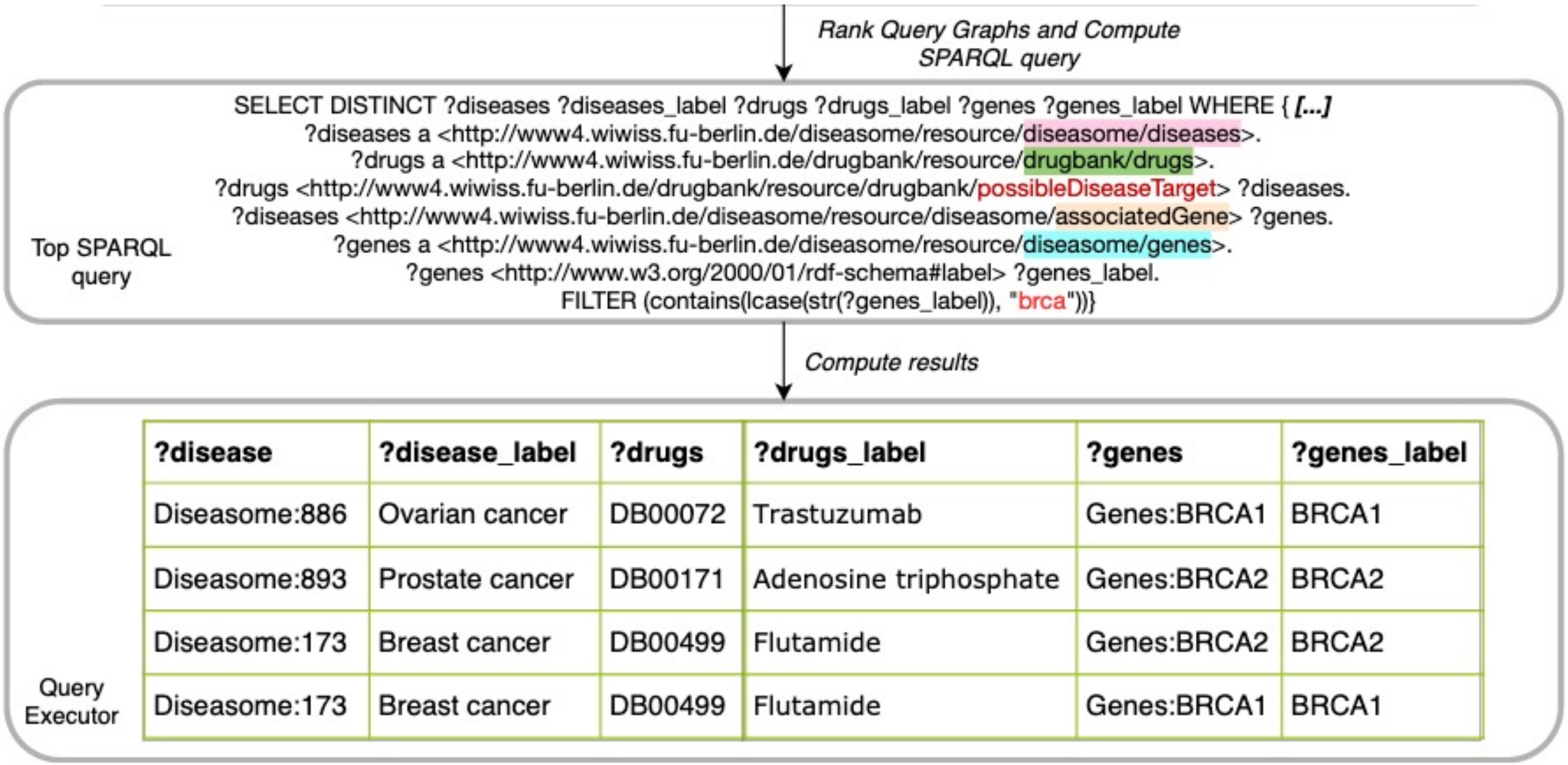
The Bio-SODA Approach #1



The Bio-SODA Approach #2



The Bio-SODA Approach #3



Rank Query Graphs and Compute SPARQL query

Top SPARQL query

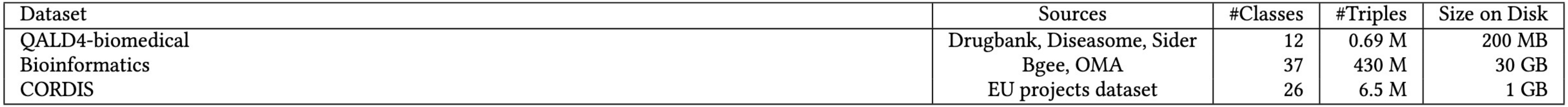
```
SELECT DISTINCT ?diseases ?diseases_label ?drugs ?drugs_label ?genes ?genes_label WHERE { [...]
?diseases a <http://www4.wiwiss.fu-berlin.de/diseasome/resource/diseasome/diseases>.
?drugs a <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/drugs>.
?drugs <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/possibleDiseaseTarget> ?diseases.
?diseases <http://www4.wiwiss.fu-berlin.de/diseasome/resource/diseasome/associatedGene> ?genes.
?genes a <http://www4.wiwiss.fu-berlin.de/diseasome/resource/diseasome/genes>.
?genes <http://www.w3.org/2000/01/rdf-schema#label> ?genes_label.
FILTER (contains(lcase(str(?genes_label)), "brca"))}
```

Compute results

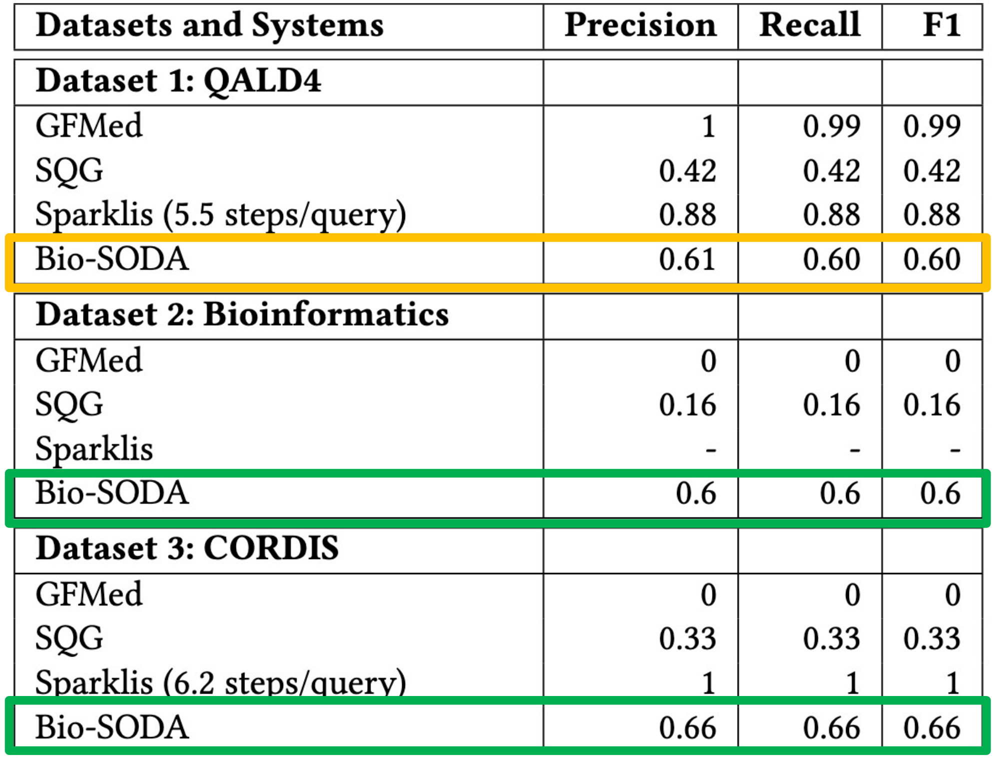
Query Executor

?disease	?disease_label	?drugs	?drugs_label	?genes	?genes_label
Diseasome:886	Ovarian cancer	DB00072	Trastuzumab	Genes:BRCA1	BRCA1
Diseasome:893	Prostate cancer	DB00171	Adenosine triphosphate	Genes:BRCA2	BRCA2
Diseasome:173	Breast cancer	DB00499	Flutamide	Genes:BRCA2	BRCA2
Diseasome:173	Breast cancer	DB00499	Flutamide	Genes:BRCA1	BRCA1

Evaluation of Bio-SODA for Question Answering



Dataset	Sources	#Classes	#Triples	Size on Disk
QALD4-biomedical	Drugbank, Diseasome, Sider	12	0.69 M	200 MB
Bioinformatics	Bgee, OMA	37	430 M	30 GB
CORDIS	EU projects dataset	26	6.5 M	1 GB



Datasets and Systems	Precision	Recall	F1
Dataset 1: QALD4			
GFMed	1	0.99	0.99
SQG	0.42	0.42	0.42
Sparklis (5.5 steps/query)	0.88	0.88	0.88
Bio-SODA	0.61	0.60	0.60
Dataset 2: Bioinformatics			
GFMed	0	0	0
SQG	0.16	0.16	0.16
Sparklis	-	-	-
Bio-SODA	0.6	0.6	0.6
Dataset 3: CORDIS			
GFMed	0	0	0
SQG	0.33	0.33	0.33
Sparklis (6.2 steps/query)	1	1	1
Bio-SODA	0.66	0.66	0.66

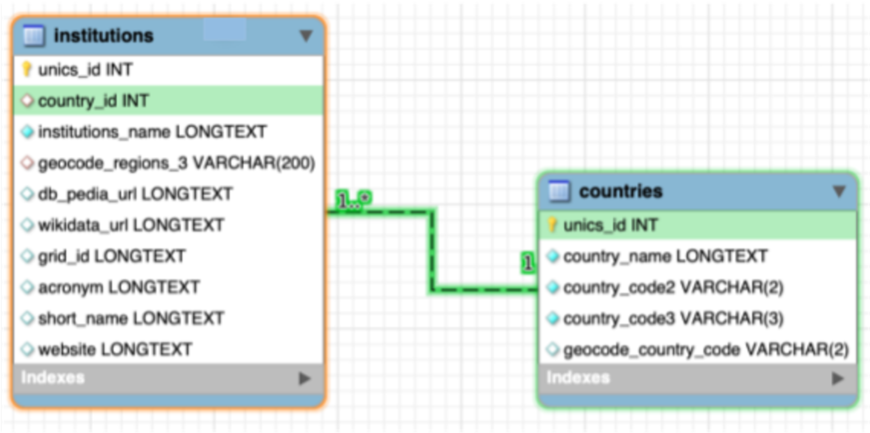
Bio-SODA significantly outperforms state of the art systems for large and complex datasets

ValueNet: Building a Natural Language-to-SQL System with Neural Networks

Querying a Relational Database in Natural Language

Question:
Find all of the institutions located in Italy.

Schema:



institutions
- unics_id INT
- country_id INT
- institutions_name LONGTEXT
- geocode_regions_3 VARCHAR(200)
- db_pedia_url LONGTEXT
- wikidata_url LONGTEXT
- grid_id LONGTEXT
- acronym LONGTEXT
- short_name LONGTEXT
- website LONGTEXT

Indexes

countries
- unics_id INT
- country_name LONGTEXT
- country_code2 VARCHAR(2)
- country_code3 VARCHAR(3)
- geocode_country_code VARCHAR(2)

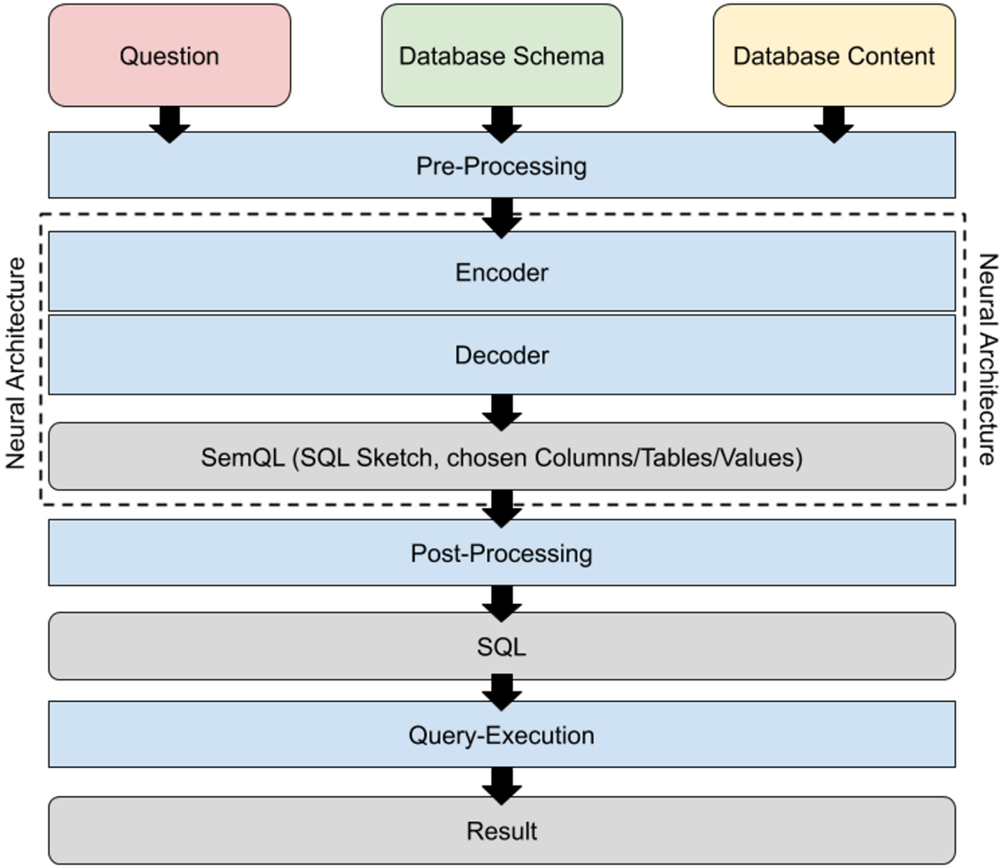
Indexes

Query:

```
SELECT T1.institutions_name
FROM institutions AS T1
      JOIN countries AS T2 ON T1.country_id = T2.unics_id
WHERE T2.country_name = 'Italy'
```

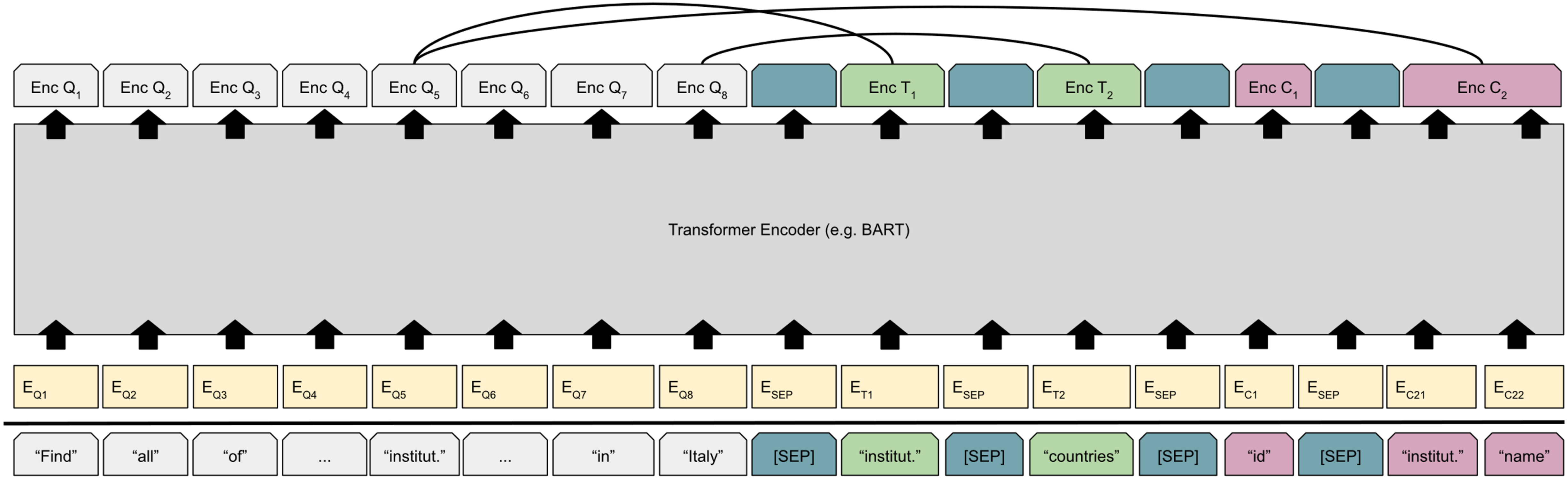
ValueNet: A Transformer-Based Neural Network Architecture

- **Generate SQL** given a natural language question – end to end
- At its core a **neural network** – consisting of an encoder / decoder architecture
- Generates an **intermediate language – SemQL** – which abstracts technical details
- SemQL is **deterministically transformed** to SQL, or any other query language (e.g. SPARQL)
- Uses state of the art **pre-trained transformers** to understand the natural language question.

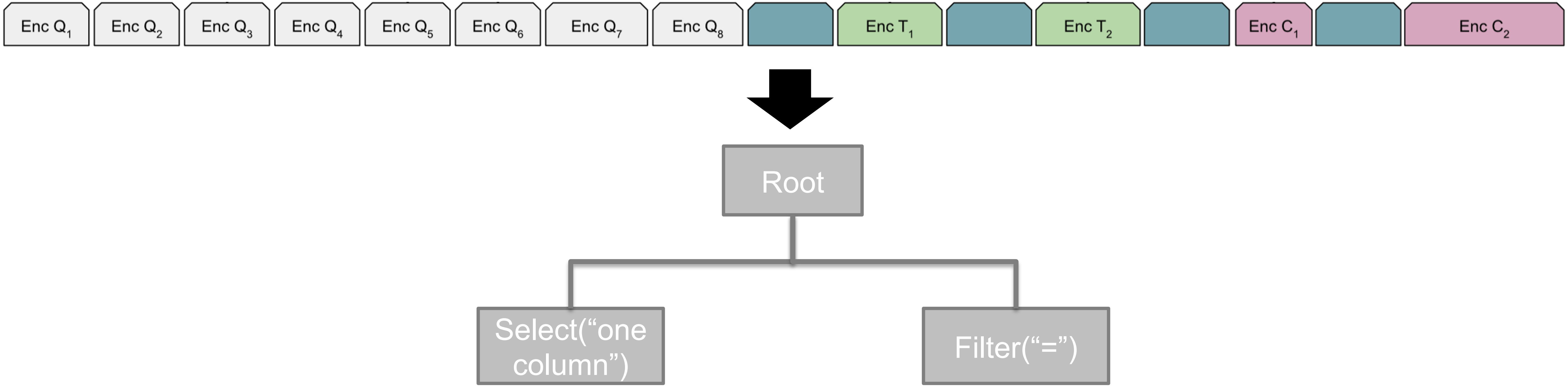


Brunner, U., & Stockinger, K. (2021). ValueNet: A Neural Text-to-SQL Architecture Incorporating Values. International Conference on Data Engineering (ICDE), Chania, Greece, 19-22 April 2021.

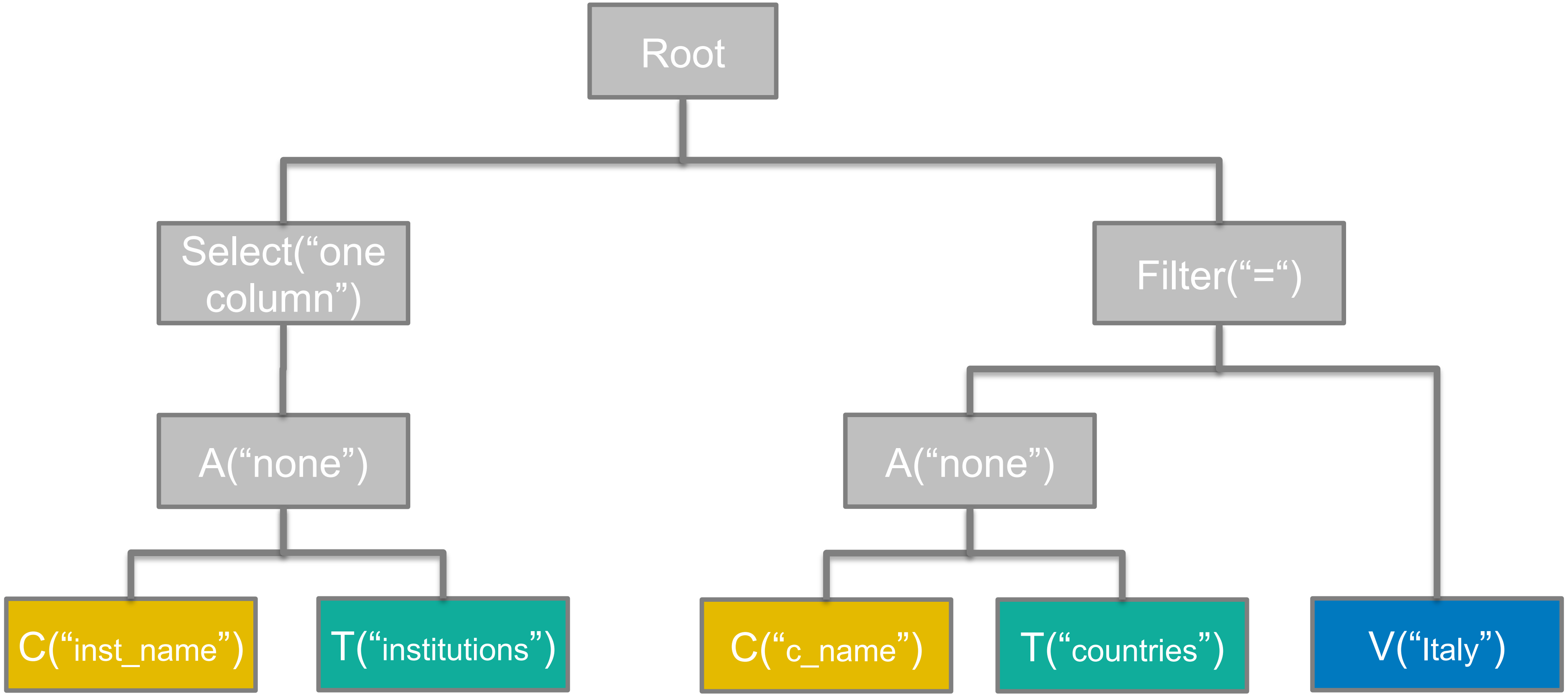
Encoding of Question & Database Schema



Decoder Recurrent Neural Network: Decoding a Query Step by Step #1

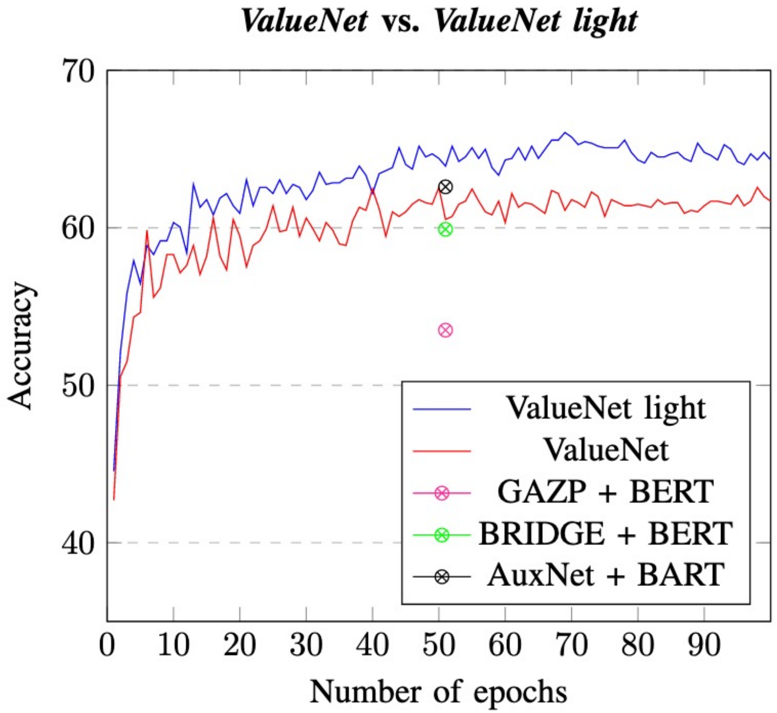


Decoder Recurrent Neural Network: Decoding a Query Step by Step #2



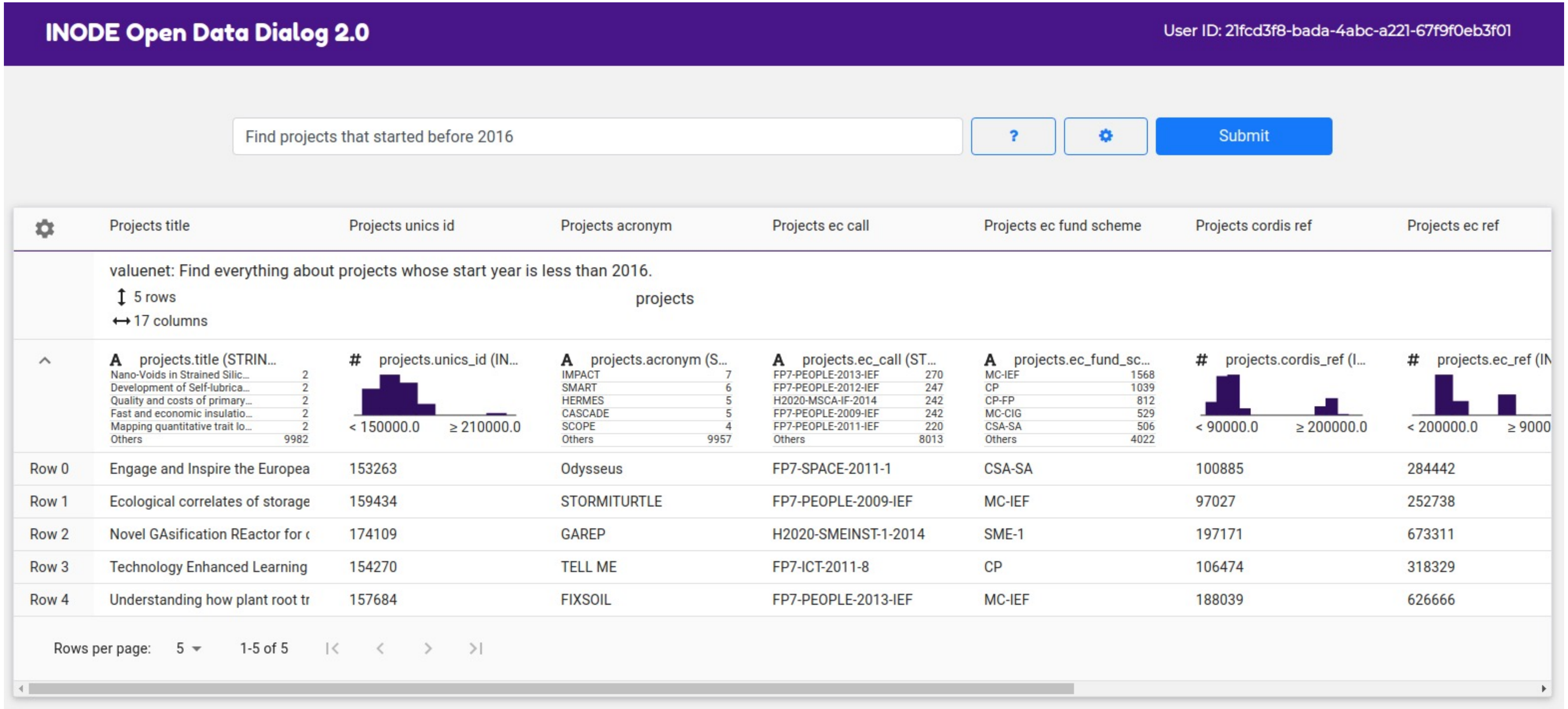
Evaluation of ValueNet for Question Answering

- Spider dataset: 200 publicly available databases with 10,181 natural language / SQL pairs
- Training set: 8,659 queries
- Validation set: 1,304 queries
- No access to test set



ValueNet outperformed initial state of the art systems

INODE in Action – A Natural Language Dialog with System Feedback



INODE Open Data Dialog 2.0

User ID: 21fcd3f8-bada-4abc-a221-67f9f0eb3f01

Find projects that started before 2016

⚙	Projects title	Projects unics id	Projects acronym	Projects ec call	Projects ec fund scheme	Projects cordis ref	Projects ec ref
	valuenet: Find everything about projects whose start year is less than 2016. ↕ 5 rows ↔ 17 columns		projects				
^	A projects.title (STRIN… Nano-Voids in Strained Silic… 2 Development of Self-lubrica… 2 Quality and costs of primary… 2 Fast and economic insulatio… 2 Mapping quantitative trait lo… 2 Others 9982	# projects.unics_id (IN… < 150000.0 ≥ 210000.0	A projects.acronym (S… IMPACT 7 SMART 6 HERMES 5 CASCADE 5 SCOPE 4 Others 9957	A projects.ec_call (ST… FP7-PEOPLE-2013-IEF 270 FP7-PEOPLE-2012-IEF 247 H2020-MSCA-IF-2014 242 FP7-PEOPLE-2009-IEF 242 FP7-PEOPLE-2011-IEF 220 Others 8013	A projects.ec_fund_sc… MC-IEF 1568 CP 1039 CP-FP 812 MC-CIG 529 CSA-SA 506 Others 4022	# projects.cordis_ref (I… < 90000.0 ≥ 200000.0	# projects.ec_ref (IN < 200000.0 ≥ 9000
Row 0	Engage and Inspire the Europea	153263	Odysseus	FP7-SPACE-2011-1	CSA-SA	100885	284442
Row 1	Ecological correlates of storage	159434	STORMITURTLE	FP7-PEOPLE-2009-IEF	MC-IEF	97027	252738
Row 2	Novel GAsification REactor for (	174109	GAREP	H2020-SMEINST-1-2014	SME-1	197171	673311
Row 3	Technology Enhanced Learning	154270	TELL ME	FP7-ICT-2011-8	CP	106474	318329
Row 4	Understanding how plant root tr	157684	FIXSOIL	FP7-PEOPLE-2013-IEF	MC-IEF	188039	626666

Rows per page: 5 ▾ 1-5 of 5 |< < > >|

ValueNet at INODE Demo by Kate Kosten, Zurich University of Applied Sciences

Conclusions and Further Information

- Building intelligent systems is not only fun but also enables access to data for a wide range of (non)-technical users

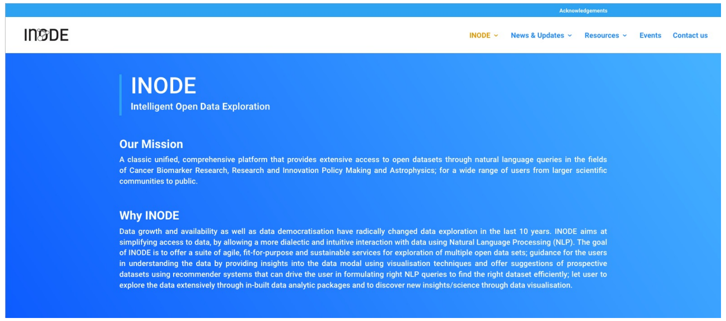
- We understand data faster and can also use it faster to generate scientific results or business value

- Further information:

  - http://www.inode-project.eu/

  - https://www.linkedin.com/in/project-inode/

  - Amer-Yahia, S., Koutrika, G., Bastian, F., Belmpas, T., Braschler, M., Brunner, U., ... & Stockinger, K. (2021). INODE: Building an End-to-End Data Exploration System in Practice [Extended Vision]. https://arxiv.org/abs/2104.04194



INODE

Intelligent Open Data Exploration

Our Mission

A classic unified, comprehensive platform that provides extensive access to open datasets through natural language queries in the fields of Cancer Biomarker Research, Research and Innovation Policy Making and Astrophysics; for a wide range of users from larger scientific communities to public.

Why INODE

Data growth and availability as well as data democratisation have radically changed data exploration in the last 10 years. INODE aims at simplifying access to data, by allowing a more dialectic and intuitive interaction with data using Natural Language Processing (NLP). The goal of INODE is to offer a suite of agile, fit-for-purpose and sustainable services for exploration of multiple open data sets; guidance for the users in understanding the data by providing insights into the data modal using visualisation techniques and offer suggestions of prospective datasets using recommender systems that can drive the user in formulating right NLP queries to find the right dataset efficiently; let user to explore the data extensively through in-built data analytic packages and to discover new insights/science through data visualisation.

References

- Amer-Yahia, S., Koutrika, G., Bastian, F., Belmpas, T., Braschler, M., Brunner, U., ... & Stockinger, K. (2021). INODE: Building an End-to-End Data Exploration System in Practice [Extended Vision]. *arXiv preprint arXiv:2104.04194*.
- Sima, A. C., de Farias, T. M., Anisimova, M., Dessimoz, C., Robinson-Rechavi, M., Zbinden, E., & Stockinger, K. (2021). Bio-SODA: Enabling Natural Language Question Answering over Knowledge Graphs without Training Data. *Scientific and Statistical Database Management Systems (SSDBM),* Tampa, Florida, USA, July *2021*
- Brunner, U., & Stockinger, K. (2021). ValueNet: a natural language-to-SQL system that learns from database information. In *International Conference on Data Engineering (ICDE), Chania, Greece, April 2021*.
- Liang, S., Stockinger, K., de Farias, T. M., Anisimova, M., & Gil, M. (2021). Querying knowledge graphs in natural language. *Journal of Big Data*, *8*(1), 1-23.
- Affolter, K., Stockinger, K., & Bernstein, A. (2019). A comparative survey of recent natural language interfaces for databases. *The VLDB Journal*, *28*(5), 793-819.
- Blunschi, L., Jossen, C., Kossmann, D., Mori, M., & Stockinger, K. (2012). SODA: Generating SQL for business users. *Proceedings of the VLDB Endowment*, *5*(10), 932-943.