Graphes de Connaissances : Cas d'Usages en Biologie et Santé

Gayo DIALLO

Team AHeaD, BPH INSERM U1219
ISPED - Université de Bordeaux
Gayo.Diallo@u-Bordeaux.fr



Instituts thématiques

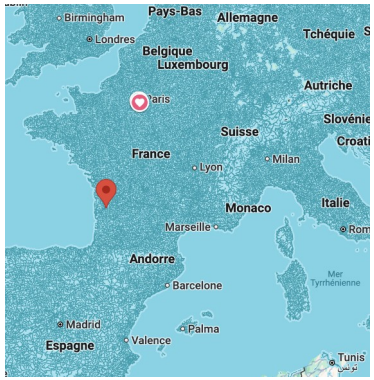
Inserm
Institut national de la santé et de la recherche médicale

BORDEAUX POPULATION HEALTH Research Center - U1219

AHeaD

université de BORDEAUX

Univ. Bordeaux – BPH - ISPED

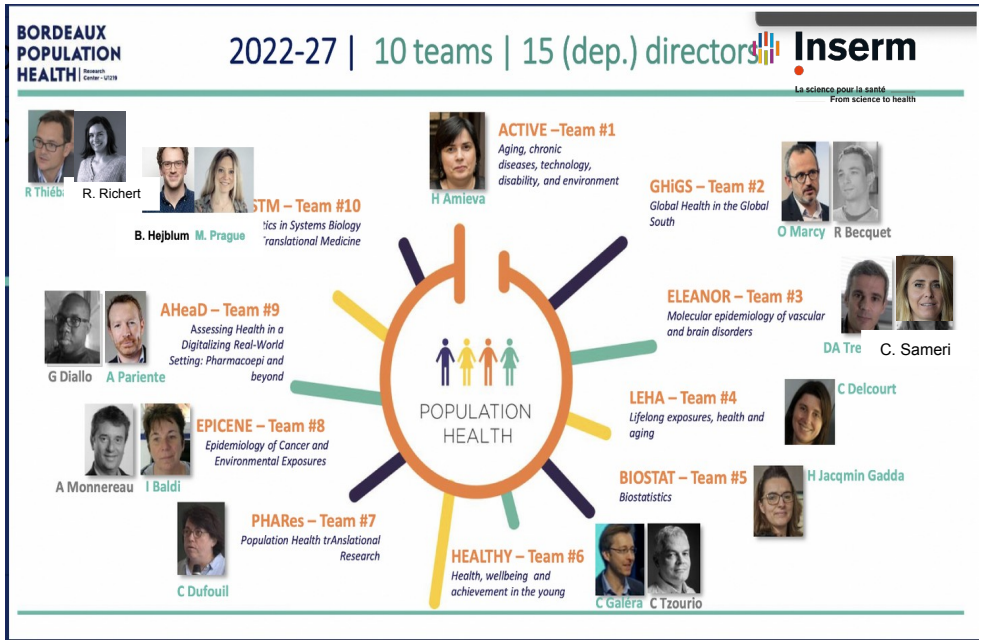




57.000 students
3200 academics
13% internat. students
2000+ PhD students
200 joint PhD theses
7 graduate programs



https://www.u-bordeaux.fr/en



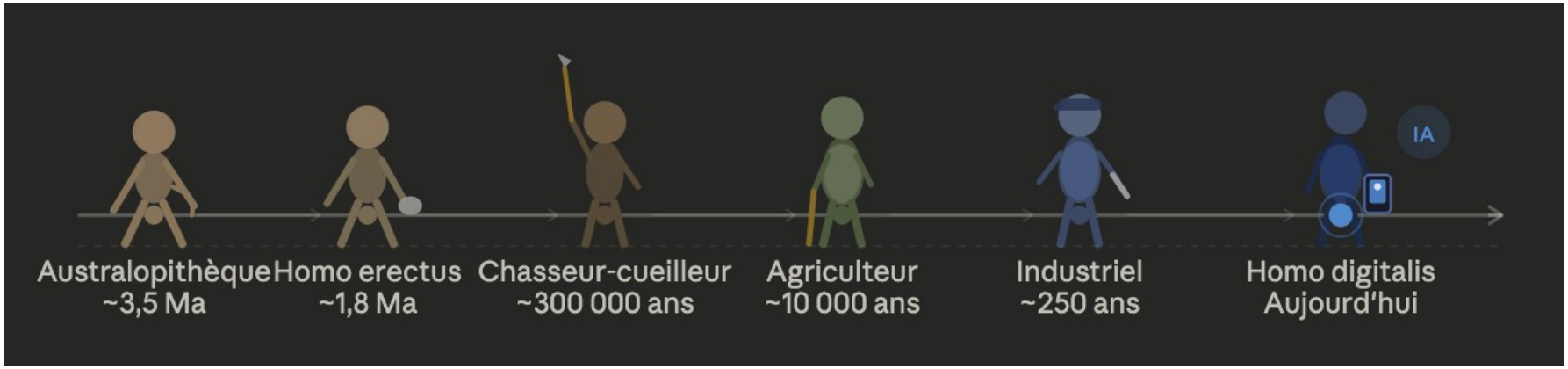
https://www.bph.center



Digital Public Health Graduate Program Bordeaux

https://www.isped.u-bordeaux.fr

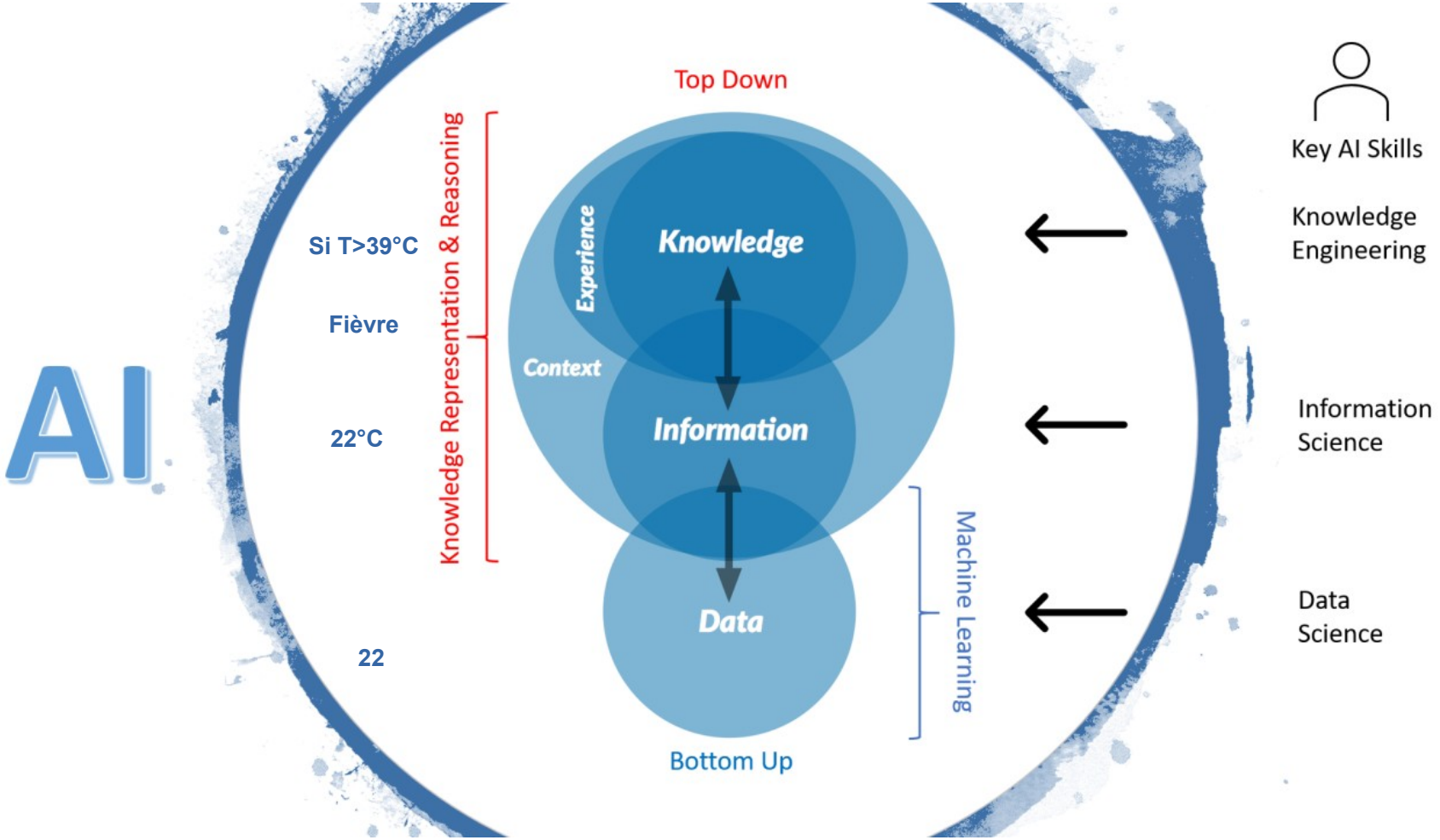
Evolution of the …Human Intelligence



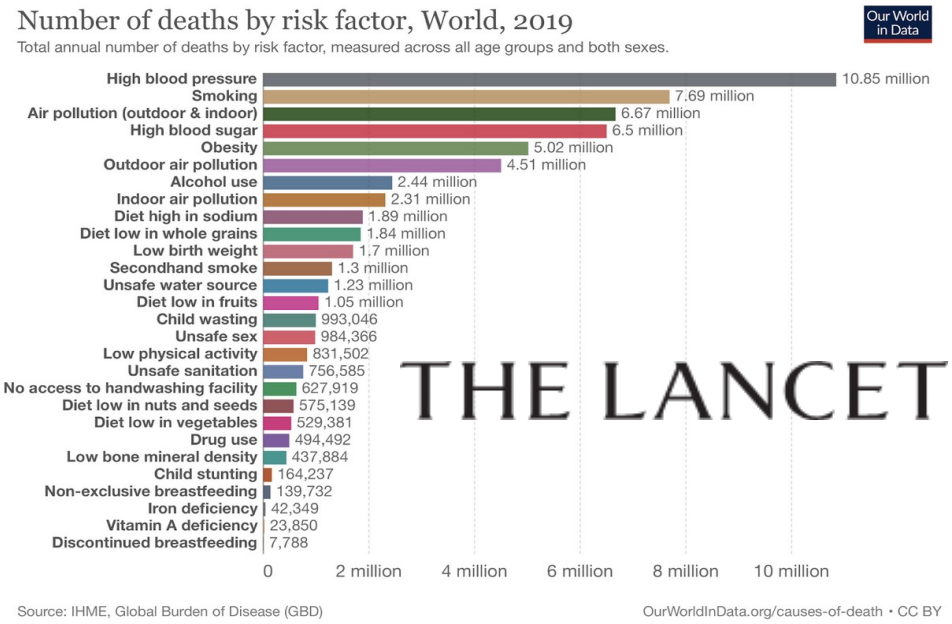
Australopithecus (~3.5 million years ago) — very stooped posture, distant ancestor
Homo erectus (~1.8 Ma) — semi-upright, holding a chipped stone tool
Hunter-gatherer (~300,000 years ago) — fully upright, armed with a spear
Farmer (~10,000 years) — farming stick, rudimentary clothing
Industrial worker (~250 years) — worker's cap, turnkey
Homo digitalis (today) — glued to their smartphone, wearing an AI badge

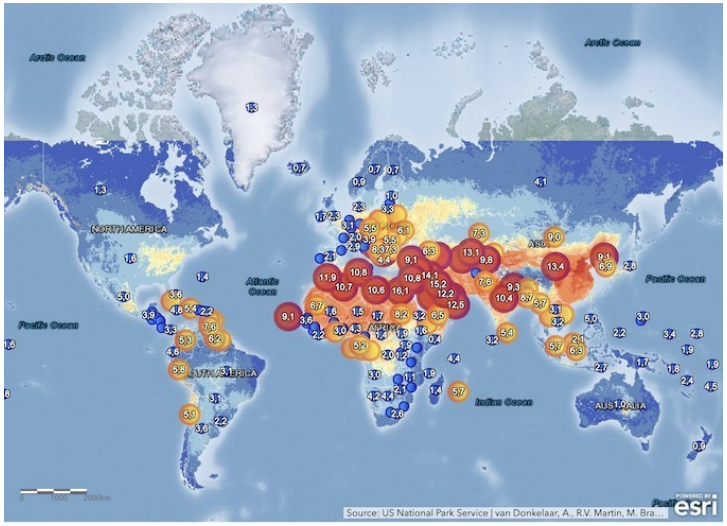
AI: Data, Information, Knowledge, ...Wisdom

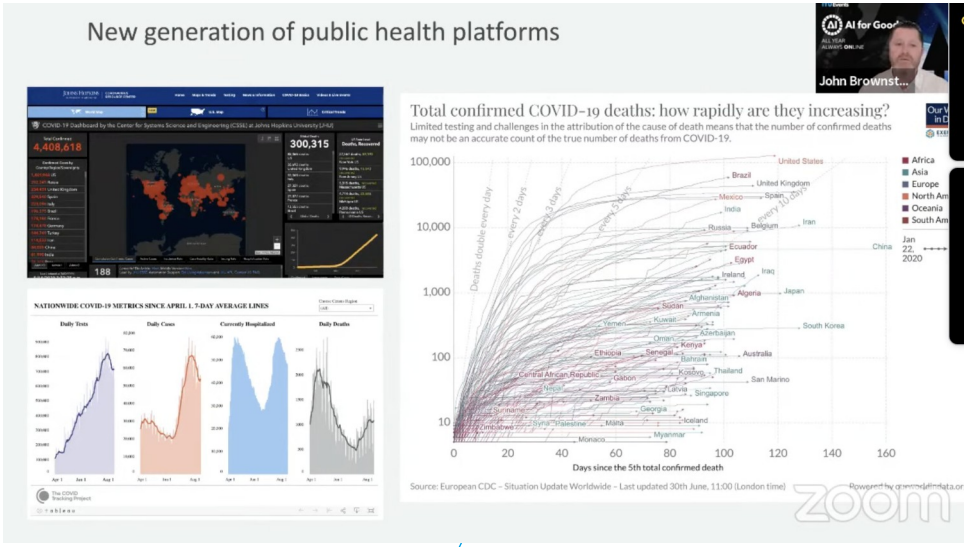
Le modèle DIKW: Données, Information, Connaissances, ...Sagesse



AI: Data, Information, Knowledge, ...Wisdom

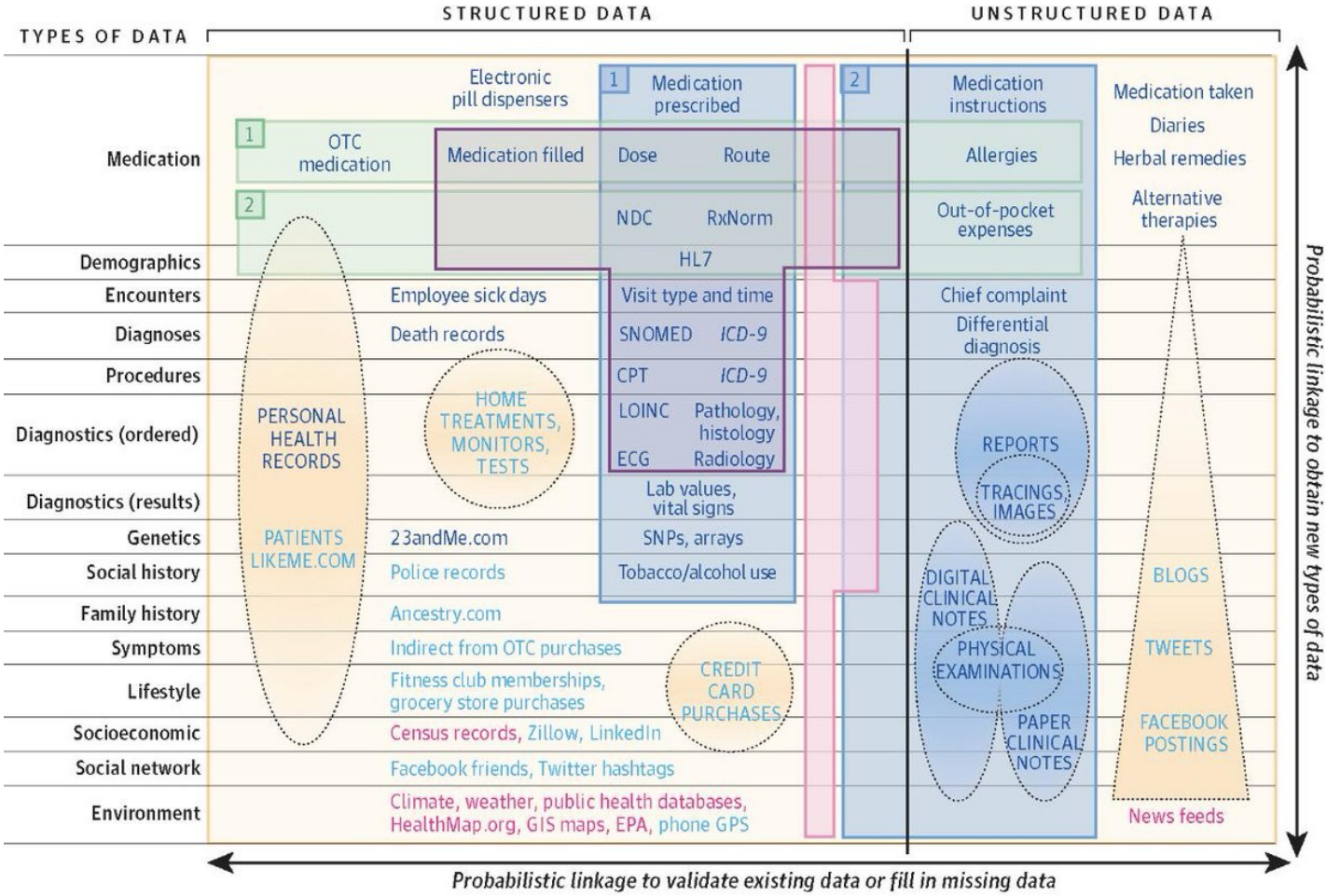






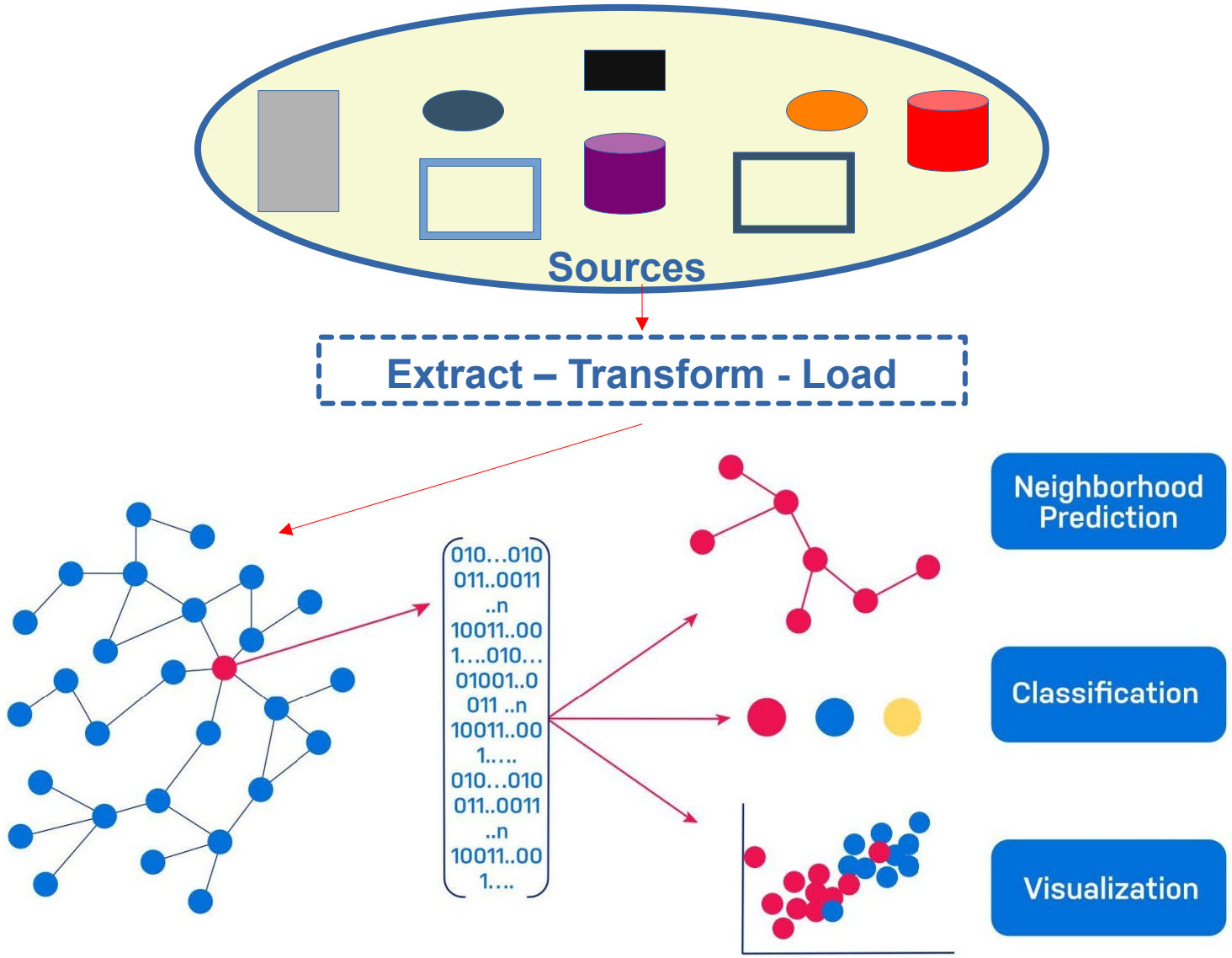
IA: Data, Information, Knowledge, ...Wisdom

- **The tapestry of potentially high-value information sources that may be linked to an individual for use in health care**



Source: Weber et al. Finding the mi
big biomedical data. JAMA 2014;311

Overall Workflow for Applying Knowledge Graphs



Agenda

- Drug Discovery: Repositioning of Molecules
- Repurposing non-pharmacological interventions
- Environmental Health and Exposure Characterization

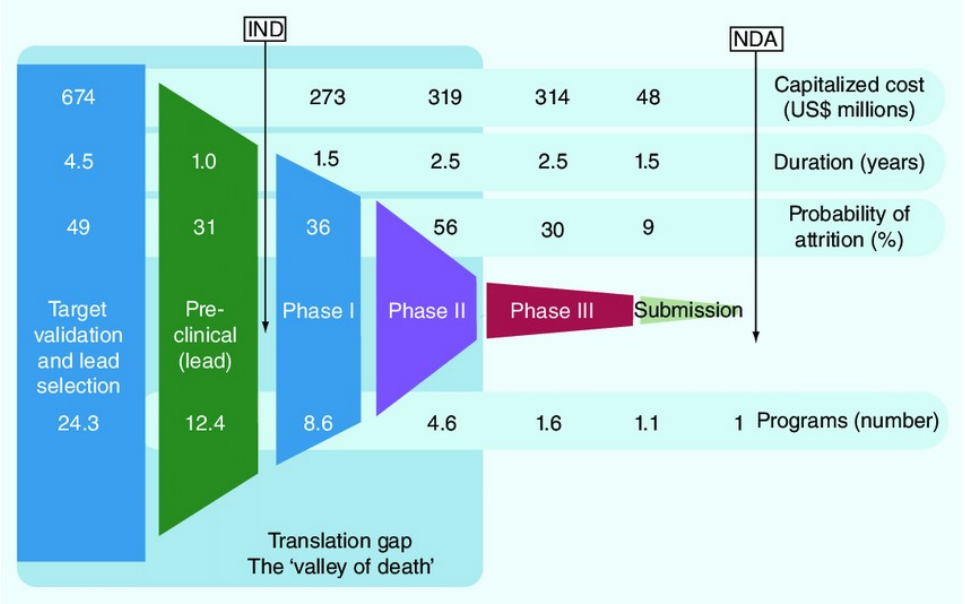
Agenda

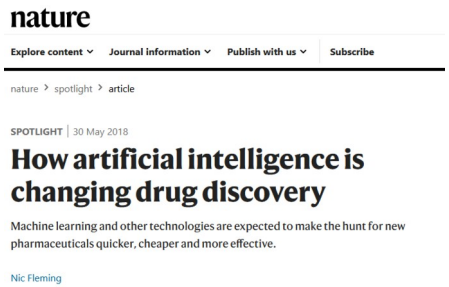
- Drug Discovery: Repositioning of Molecules
- Repurposing non-pharmacological interventions
- Environmental Health and Exposure Characterization

Drug Discovery and AI

- Drug discovery is a costly process both financially and in terms of time
- Bringing a new drug to market takes 10 to 15 years or more ~2 billion € for a treatment
- 3 phases for the process and 1 phase for post-market surveillance
- Drug repositioning (DR) is a promising alternative

1. Finding new uses for existing drugs
2. With Covid-19, we observe that this avenue has been exploited
3. Emblematic example: Sildenafil (anti hypertension) repositioned against sexual disorders (Viagra)




nature

Explore content · Journal information · Publish with us · Subscribe

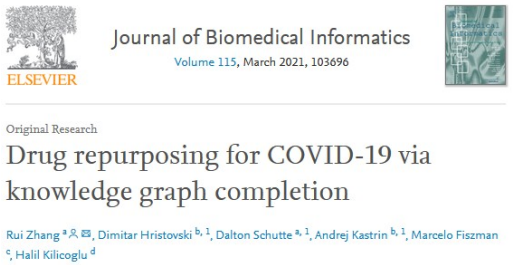
nature › spotlight › article

SPOTLIGHT | 30 May 2018

How artificial intelligence is changing drug discovery

Machine learning and other technologies are expected to make the hunt for new pharmaceuticals quicker, cheaper and more effective.

Nic Fleming


ELSEVIER

Journal of Biomedical Informatics

Volume 115, March 2021, 103696

Original Research

Drug repurposing for COVID-19 via knowledge graph completion

Rui Zhang [a], Dimitar Hristovski [b, 1], Dalton Schutte [a, 1], Andrej Kastrin [b, 1], Marcelo Fiszman [c], Halil Kilicoglu [d]

Drug Discovery/Repositioning and AI

PubMed®

(drug repositioning OR Drug repurposing) and (AI OR Artificial Intelligence) Search

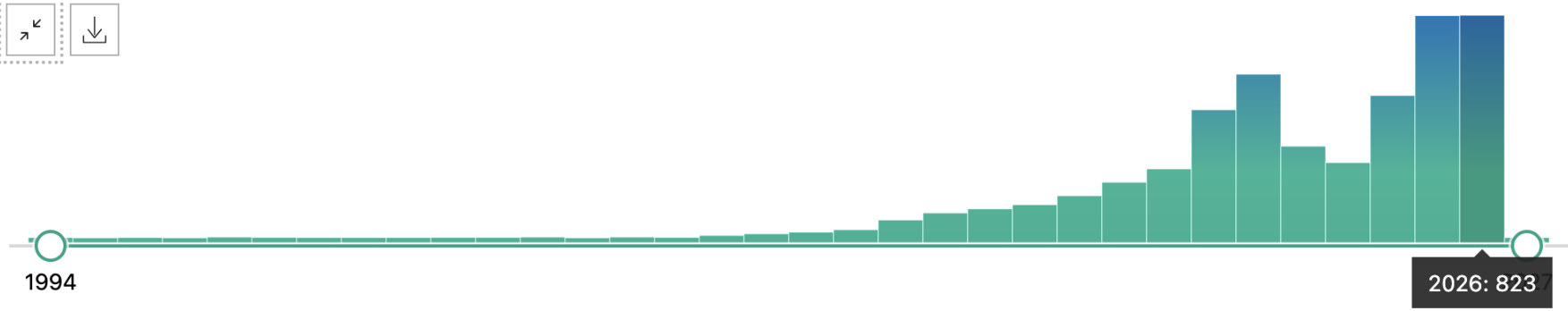
Advanced Create alert Create RSS User Guide

Save Email Send to Sort by: Best match Display options

RESULTS BY YEAR

4,468 results

Page 1 of 447



MY CUSTOM FILTERS

Edit custom filters

PUBLICATION DATE

- 1 year
- 5 years

1 COVID-19: Advances in diagnostic tools, treatment strategies, and vaccine development.

Cite

Sreepadmanabh M, Sahu AK, Chande A.

J Biosci. 2020;45(1):148. doi: 10.1007/s12038-020-00114-6.

PMID: 33410425 **Free PMC article.** Review.

Importantly, we furnish the reader with a multidisciplinary perspective on how conventional methods like serology and RT-PCR, as well as cutting-edge technologies like CRISPR/Cas and **artificial**

Drug Repositioning (DR) and AI

MIT News
ON CAMPUS AND AROUND THE WORLD
SUBSCRIBE

Machine learning uncovers potential new TB drugs

Computational method for screening drug compounds can help predict which ones will work best against tuberculosis or other diseases.

Anne Trafton | MIT News Office
October 15, 2020



Cell
Volume 180, Issue 4, 20 February 2020, Pages 688-702.e13

CellPress

Article

A Deep Learning Approach to Antibiotic Discovery

Jonathan M. Stokes [1, 2, 3], Kevin Yang [3, 4, 10], Kyle Swanson [3, 4, 10], Wengong Jin [3, 4], Andres Cubillos-Ruiz [1, 2, 5], Nina M. Donghia [1, 5], Craig R. MacNair [6], Shawn French [6], Lindsey A. Carfrae [6], Zohar Bloom-Ackermann [2, 7], Victoria M. Tran [2], Anush Chiappino-Pepe [5, 7], Ahmed H. Badran [2], Ian W. Andrews [1, 2, 5], Emma J. Chory [1, 2], George M. Church [5, 7, 8], Eric D. Brown [6], Tommi S. Jaakkola [3, 4] ... James J. Collins [1, 2, 5, 8, 9, 11]

Halicine

Drug repositioning: identifying and developing new uses for existing drugs

Ted T. Ashburn & Karl B. Thor

Nature Reviews Drug Discovery **3**, 673–683 (2004) | Cite this article

Thalidomide

Leverage: Symbolic AI – Ontology - Knowledge Graph

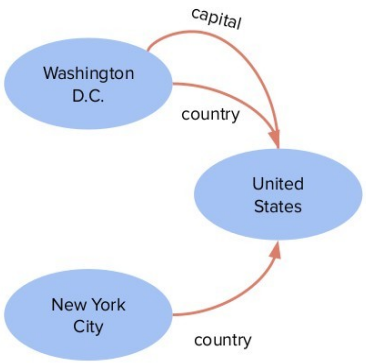
Why Embeddings?

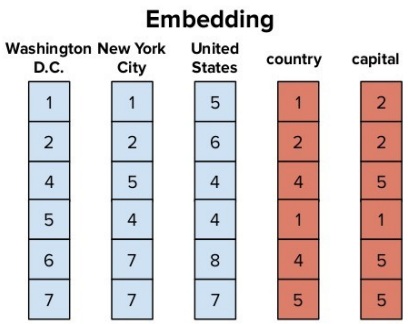
Knowledge graphs are symbolic and discrete
Difficult for machines to process directly.
Embeddings capture semantic relationships
Similar entities/relations have closer vectors.
Enables machine learning
Use embeddings as features for tasks like link prediction, node classification, etc.

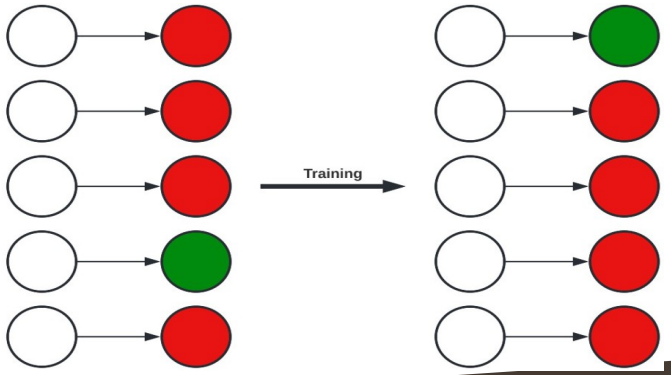
Models

- **TransE** (2013)
- **TransH** (2014)
- **DistMult** (2014)
- **ComplEx** (2016)
- **ConvE** (2018)
- …

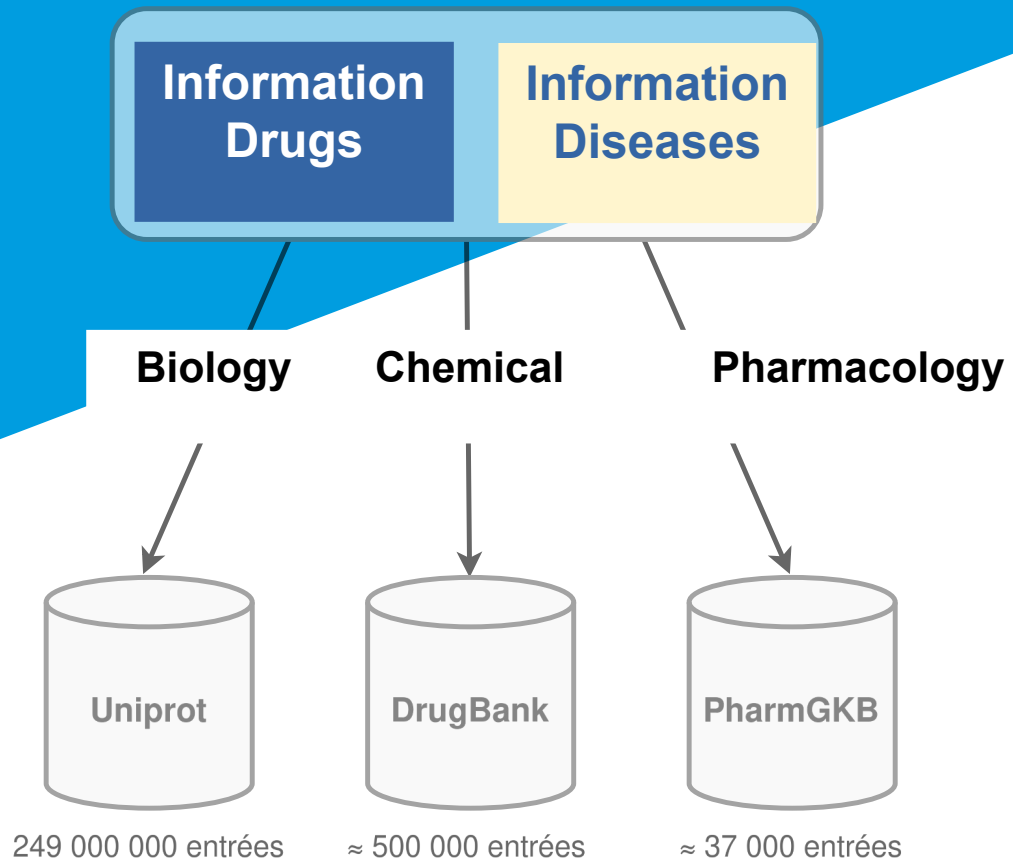
Federico Bianchi et al. /







Knowledge Graphs-based approach to DR



- **Heterogenous Information**

- **Multiple Knowledge Bases**
- **Collecting and storage of data**

Consortium, T. U. (2022), UniProt: the Universal Protein Knowledge base in 2023.
Knox, C., et al., (2010), Drugbank 3.0: a comprehensive resource for 'omics' research on drugs.
Whirl-Carrillo, M., et al., (2021), An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine.

Oregano: turning Link Prediction into Drug- Target relations detection

Main Hypothesis

The more we integrate drug related information, better we will be able to discover relevant link between drugs and targets

==> Design a holistic knowledge graph about drugs and related information, including natural products

==> Plant: ~ 30% of marketed drug

Repurposing Non-pharmacological Interventions for Alzheimer's Diseases through Link Prediction on Biomedical Literature

Yongkang Xiao, Yu Hou, Huixue Zhou, Gayo Diallo, Marcelo Fiszman, Julian Wolfson, Halil Kilicoglu, You Chen, Chang Su, Hua Xu, William G Mantyh, Rui Zhang

PMID: 37292731 PMCID: PMC10246059 DOI: 10.1101/2023.05.15.23290002

Home > Database Systems for Advanced Applications > Conference paper

Neuro-Symbolic XAI: Application to Drug Repurposing for Rare Diseases

Conference paper | First Online: 08 April 2022

pp 539–543 | Cite this conference paper



Database Systems for Advanced Applications

(DASFAA 2022)

www.nature.com/scientificdata

scientific data

Check for updates

OPEN

DATA DESCRIPTOR

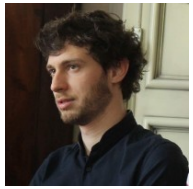
The OREGANO knowledge graph for computational drug repurposing

Marina Boudin, Gayo Diallo, Martin Drancé & Fleur Mougin

Drug repositioning is a faster and more affordable solution than traditional drug discovery approaches. From this perspective, computational drug repositioning using knowledge graphs is a very promising direction. Knowledge graphs constructed from drug data and information can be used to generate hypotheses (molecule/drug - target links) through link prediction using machine learning algorithms. However, it remains rare to have a holistically constructed knowledge graph using the broadest possible features and drug characteristics, which is freely available to the community. The OREGANO knowledge graph aims at filling this gap. The purpose of this paper is to present the OREGANO knowledge graph, which includes natural compounds related data. The graph was developed from scratch by retrieving data directly from the knowledge sources to be integrated. We therefore designed the expected graph model and proposed a method for merging nodes between the different knowledge sources, and finally, the data were cleaned. The knowledge graph, as well as the source codes for the ETL process, are openly available on the GitHub of the OREGANO project (https://gitub.u-bordeaux.fr/erias/oregano).

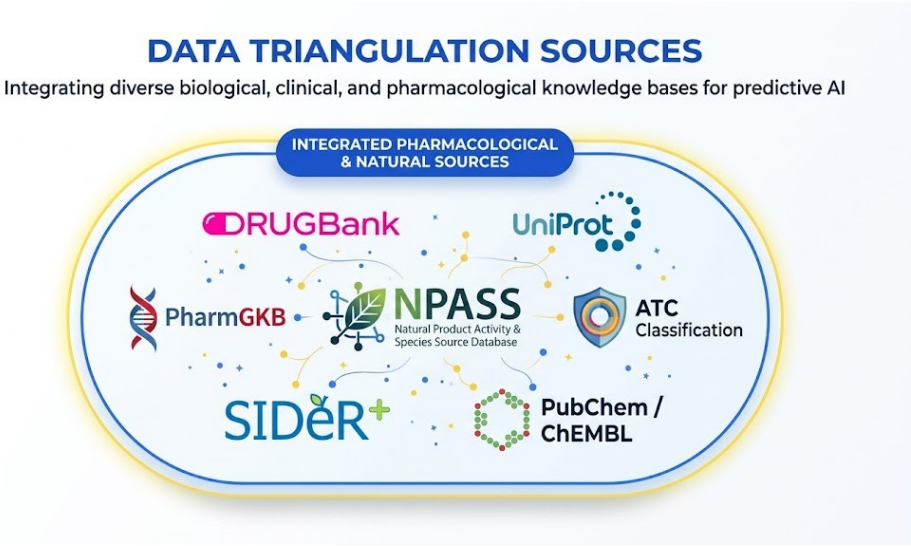


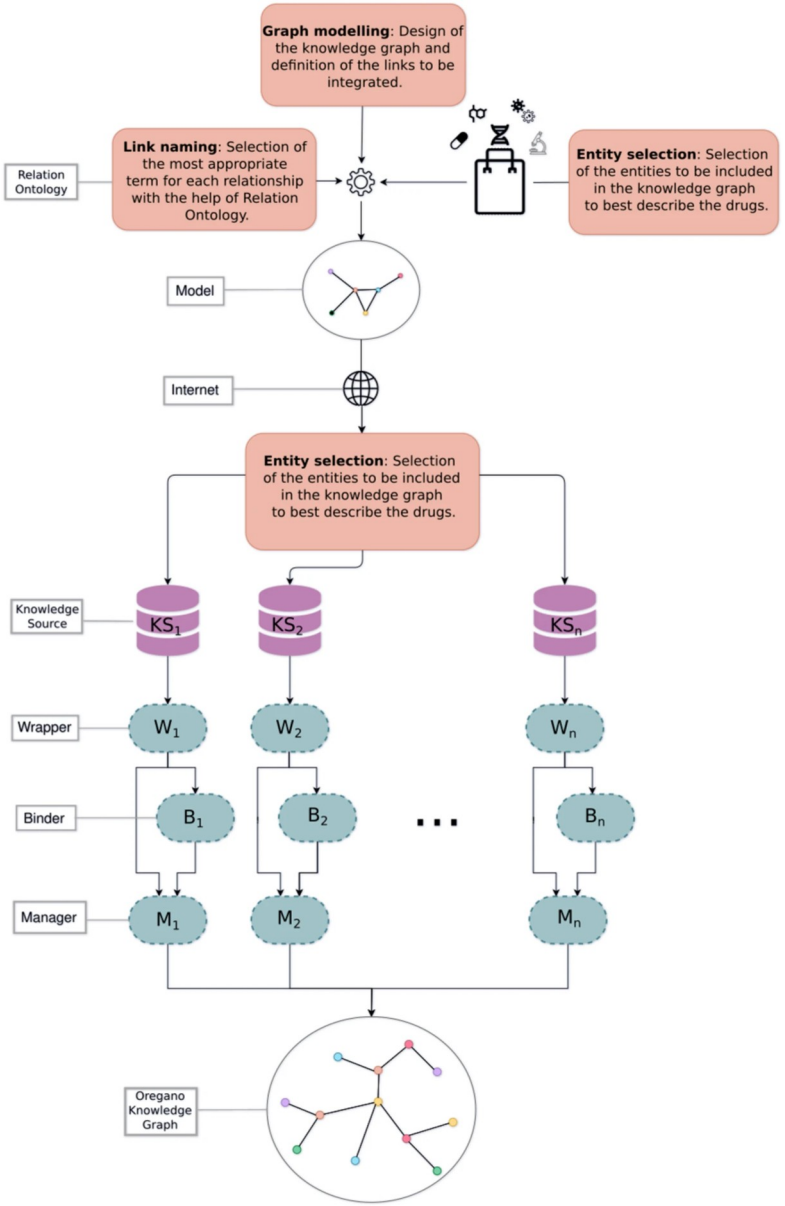
Marina Boudin



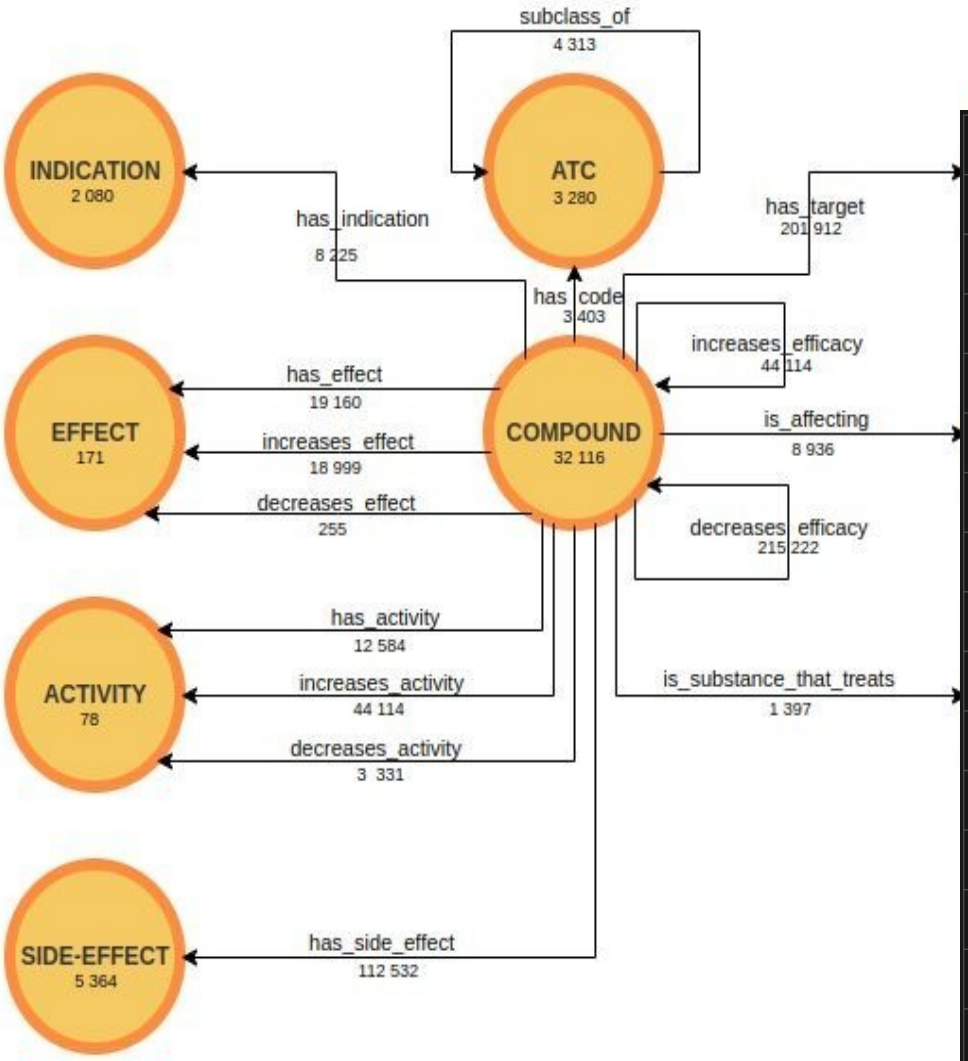
Martin Drancé

Oregano: Datasets and ETL Workflow





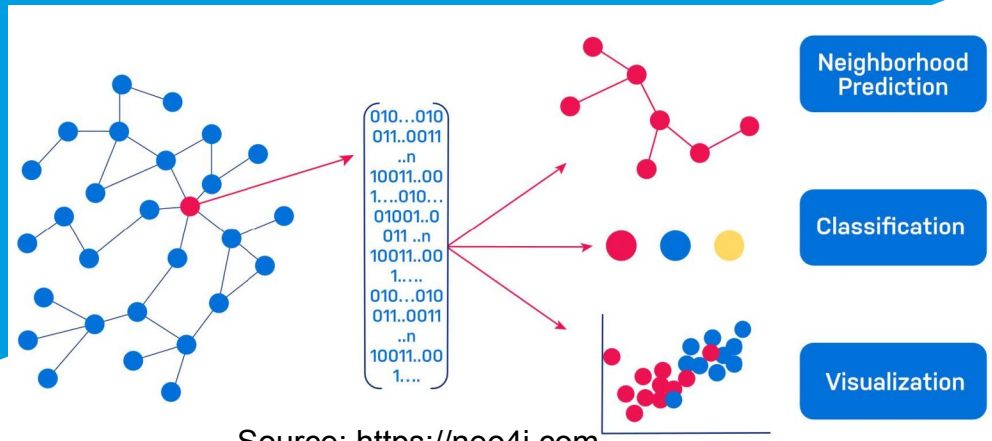
Oregano: The resulting Knowledge Graph



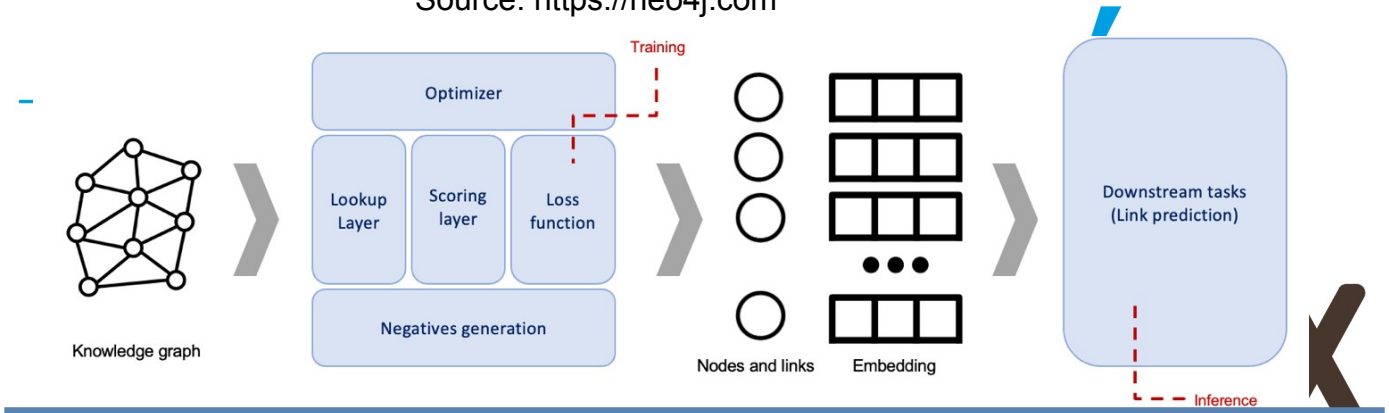
	Version 1	Version 2	Version 3
Compound	42856	32116	46138
including Natural Compound	30926	22676	22704
Target	122757	14602	23411
Gene	8774	13363	19191
Disease	9011	8934	10478
ATC code	1005	3280	3280
Effect	153	171	0
Activity	74	78	0
Indication	2154	2080	2080
Side-effect	5556	5364	5364
Pathway	297	2128	2128
Phenotype	11210	6854	6756
Biological processes	0	0	12280
Reaction	0	0	14973
Total	203847	88970	146079

https://gitub.u-bordeaux.fr/erias/oregano

Graph Embeddings: From graph to vector and prediction



Source: https://neo4j.com



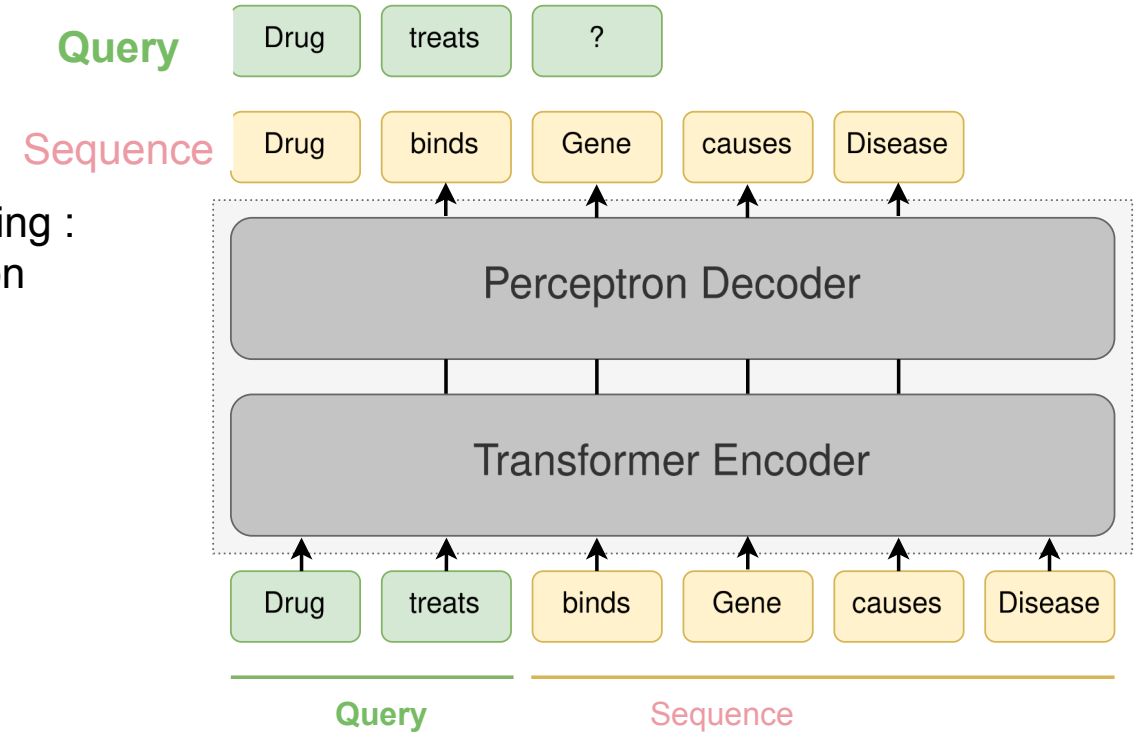
Lookup layer (simply assigns an embedding to each node and edge type of the graph)
Scoring layer interacting with a **loss function**
Negatives generation
Optimization to train embeddings, which are used by the above scoring function to predict new links (the downstream task)

Explainability: Justification of generated hypotheses

How to provide justification for the Provided results?

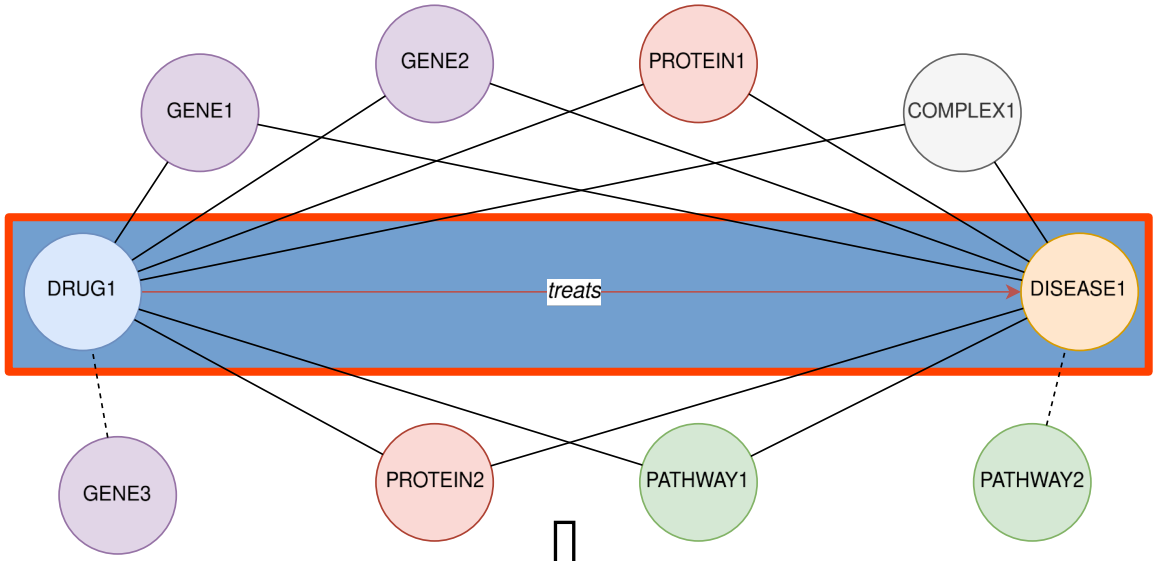
Triplets Classification

- Works in a similar way to text processing :
  - ➔ Token by token sequence generation
- Faster than reinforcement learning,
- can be parallelized



Bai, Y., et al., (2022), SQUIRE : A sequence-to-sequence framework for multi-hop knowledge graph reasoning.

Explainability: Overall Approach



Neighbors Selection

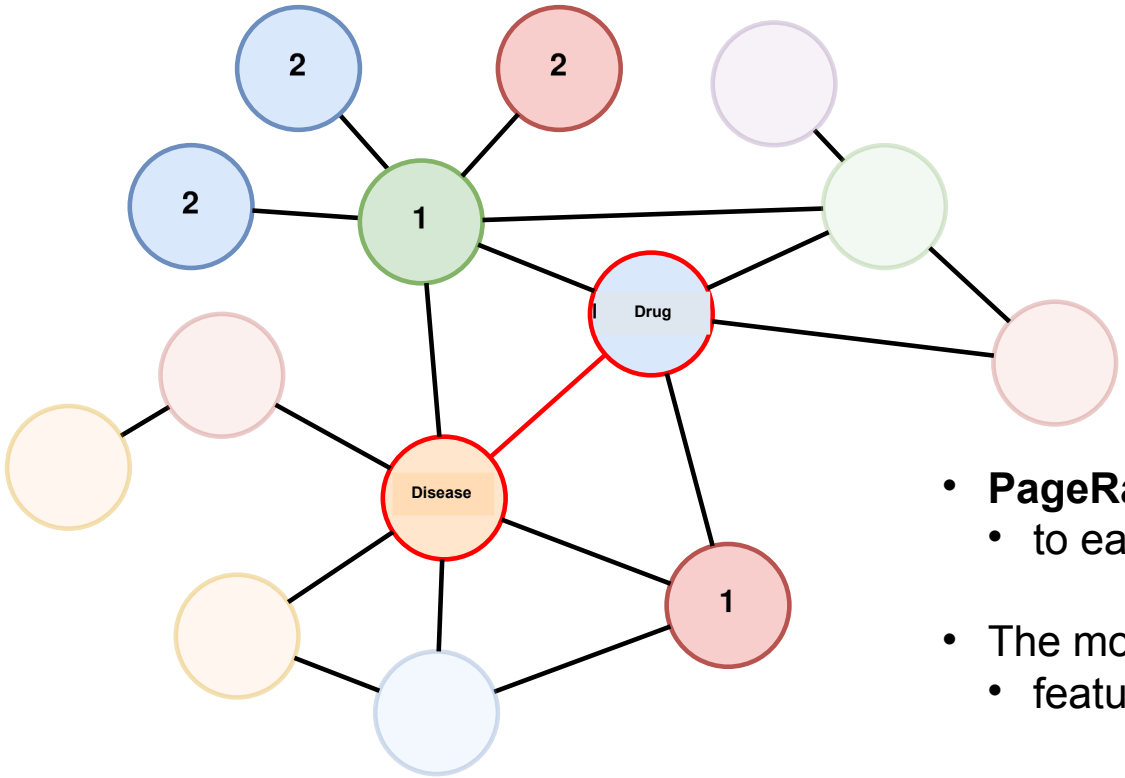
Pairs \ Features	GENE	PROTEIN	COMPLEX	PATHWAY	ANATOMY
DRUG1-DISEASE1	2	2	1	1	0
...	...	...	...	...	...

Random Forest based learning

1. Selecting positive triplets
2. Generating negative triplets
3. Selecting common neighbors
4. Features = Number of common
   1. neighbors for each type
5. Training the Random Forest

Explainability: Overall Approach

PageRank algorithm for features generation



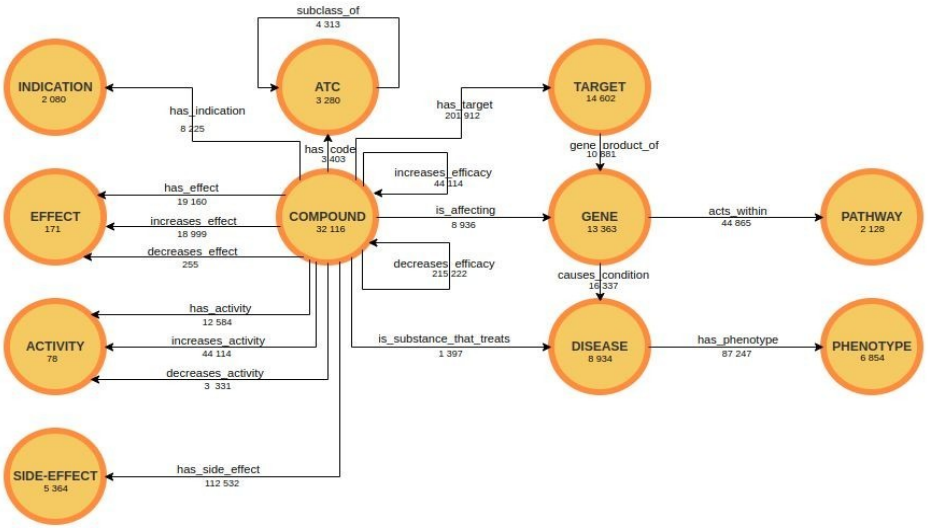
- **PageRank** allocates an importance score
  - to each node
- The most important nodes are ketp as
  - features

Page, L., Brin, et al., (1999), The PageRank Citation Ranking: Bringing order to the web.

Explainability: Evaluation

- **Precision :** The model's ability to distinguish between true and false triplets
- **Recall :** Ability of the model to find all true triplets

Oregano

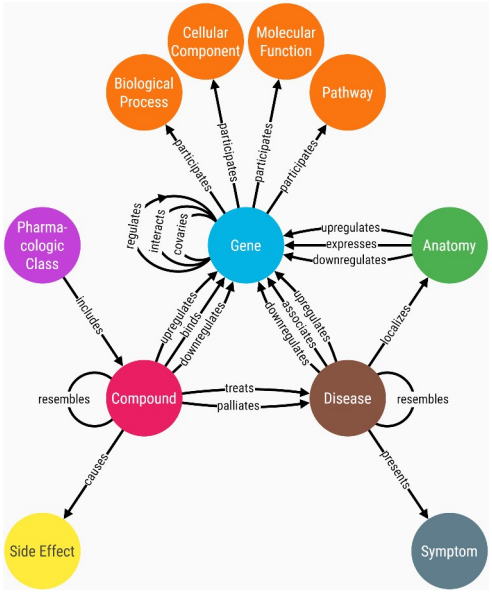


BioKG

A comprehensive large scale biomedical knowledge graph for AI powered data driven biomedical research

Yuan Zhang, Xin Sui, Feng Pan, Kaixian Yu, Keqiao Li, View ORCID ProfileShubo Tian, Arslan Erdengasileng, Qing Han, Wanjing Wang, Jianan Wang, Jian Wang, Donghu Sun, Henry Chung, Jun Zhou, Eric Zhou, Ben Lee, Peili Zhang, View ORCID ProfileXing Qiu, Tingting Zhao, Jinfeng Zhang doi: https://doi.org/10.1101/2023.10.13.562216

Hetionet

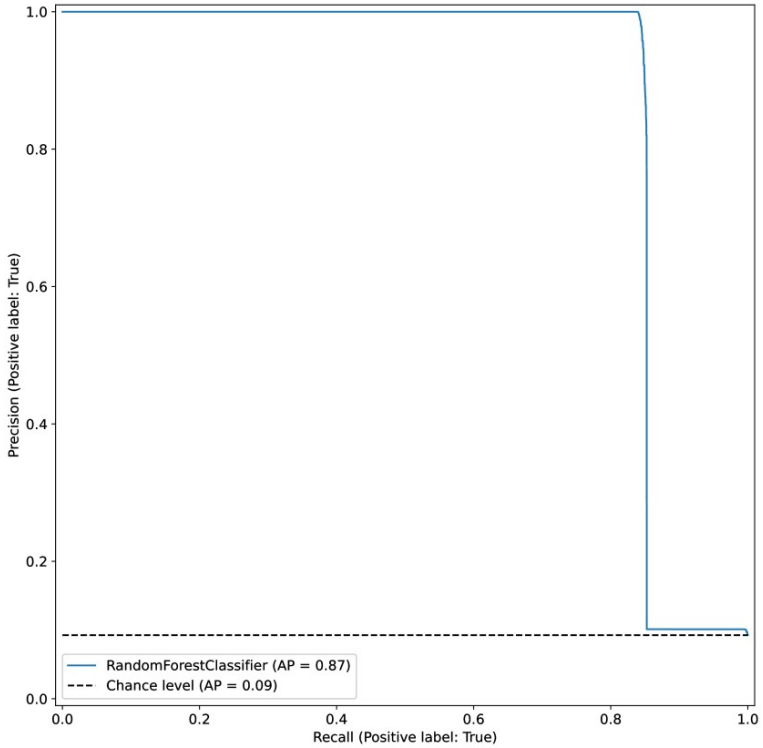


Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, Green A, Khankhanian P, Baranzini SE. ELife (2017) DOI: 10.7554/elife.26726

Explainability: Evaluation

- **Precision :** The model's ability to distinguish between true and false triplets
- **Recall :** Ability of the model to find all true triplets

  - **<u>Precision / Recall curve on BioKG :</u>**



- Precision on Oregano : 0,832
- Precision on Hetionet : 0,923
- Precision on BioKG : 0,999

Explainability: the ALS use case

Amyotrophic lateral sclerosis

- Collaboration with **Pr. Cyril Goizet**



- 12 Drugs proposed for ALS
  - ➔ ALS : 1600 new cases each year
  - ➔ No any curative treatment
- Common results on Oregano and BioKG
- Metformine on phase 2 clinical trial

Results for ALS

Drugs	Publications	Diseases
Metformine Rosiglitazone	2 2	Diabetes
Fluorouracile Oxaliplatine Épirubicine Bévacizumab	1 0 0 0	Cancer
Atorvastatine Simvastatine	2 2	Cholesterol
Paliperidone	2	Schizophrenia
Clopidogrel	0	Accidents cardiovasculaires
Nicotine	0	Désintoxication tabagique
Aspirine	0	Pain, Fever

Explainability: the ALS use case

- Nighbors used by the model to link ALS to **Metformine** and **Rosiglitazone**



Take Home Message

- **Highlights**

  - Improving of triple classification methods for the task of DR
  - using only the neighborhood information of nodes in a KG
  - The first approach to test the performance of an RF classifier
  - on these KGs for DR, without any external information or
  - KG embedding methods
  - High performance compared to SOTA, high Accuracy and Speed
  - Fully Transparent and Explainable results

- **Limitations and Future Work**

  - Handling sparse KGs ==> uninformative explanations
  - Taking into account the availability of new knowledge
  - Deep clinical evaluation of the results

    https://gitub.u-bordeaux.fr/erias/oregano

  - More information:

Agenda

- Drug Discovery: Repositioning of Molecules
- Repurposing non-pharmacological interventions
- Environmental Health and Exposure Characterization

Repositioning non-Pharmacological Interventions: Alzheimer Disease Prevention

scientific reports

Check for updates

OPEN

Repurposing non-pharmacological interventions for Alzheimer's disease through link prediction on biomedical literature

Yongkang Xiao[1,13], Yu Hou[2,13], Huixue Zhou[1], Gayo Diallo[3], Marcelo Fiszman[4,5], Julian Wolfson[6], Li Zhou[7], Halil Kilicoglu[8], You Chen[9], Chang Su[10], Hua Xu[11], William G. Mantyh[12] & Rui Zhang[2]✉

Non-pharmaceutical interventions (NPI) have great potential to improve cognitive function but limited investigation to discover NPI repurposing for Alzheimer's Disease (AD). This is the first study to develop an innovative framework to extract and represent NPI information from biomedical literature in a knowledge graph (KG), and train link prediction models to repurpose novel NPIs for AD prevention. We constructed a comprehensive KG, called ADInt, by extracting NPI information from biomedical literature. We used the previously-created SuppKG and NPI lexicon to identify NPI entities. Four KG embedding models (i.e., TransE, RotatE, DistMult and ComplEX) and two novel graph convolutional network models (i.e., R-GCN and CompGCN) were trained and compared to learn the representation of ADInt. Models were evaluated and compared on two test sets (time slice and clinical trial ground truth) and the best performing model was used to predict novel NPIs for AD. Discovery patterns were applied to generate mechanistic pathways for high scoring candidates. The ADInt has 162,212 nodes and 1,017,284 edges. R-GCN performed best in time slice (MR = 5.2054, Hits@10 = 0.8496) and clinical trial ground truth (MR = 3.4996, Hits@10 = 0.9192) test sets. After evaluation by domain experts, 10 novel dietary supplements and 10 complementary and integrative health were proposed from the score table calculated by R-GCN. Among proposed novel NPIs, we found plausible mechanistic pathways for photodynamic therapy and Choerospondias axillaris to prevent AD, and validated psychotherapy and manual therapy techniques using real-world data analysis. The proposed framework shows potential for discovering new NPIs for AD prevention and understanding their mechanistic pathways.

The Alzheimer Disease Prevention Use Case

REPURPOSING NON-PHARMACOLOGICAL INTERVENTIONS FOR ALZHEIMER'S (ADInt)

Framework: Literature-Based Discovery, Graph Neural Networks & In-Silico Triangulation

1. ADInt GRAPH CONSTRUCTION

Data Engineering & Pipelines

- **Sources:** SemMedDB (PubMed triples) & SuppKG (UMLS-based DS ontology)
- **Quality & Preprocessing Pipeline:**
  - Generic concept pruning
  - Log-likelihood (G^2) filtering
  - Fact verification via fine-tuned PubMedBERT (F1 = 0.854)
- **Final KG:** 162,212 entities across 1,017,284 semantic triples

2. KGE & GNN PREDICTIVE ARCHITECTURE

Machine Learning & Benchmarking

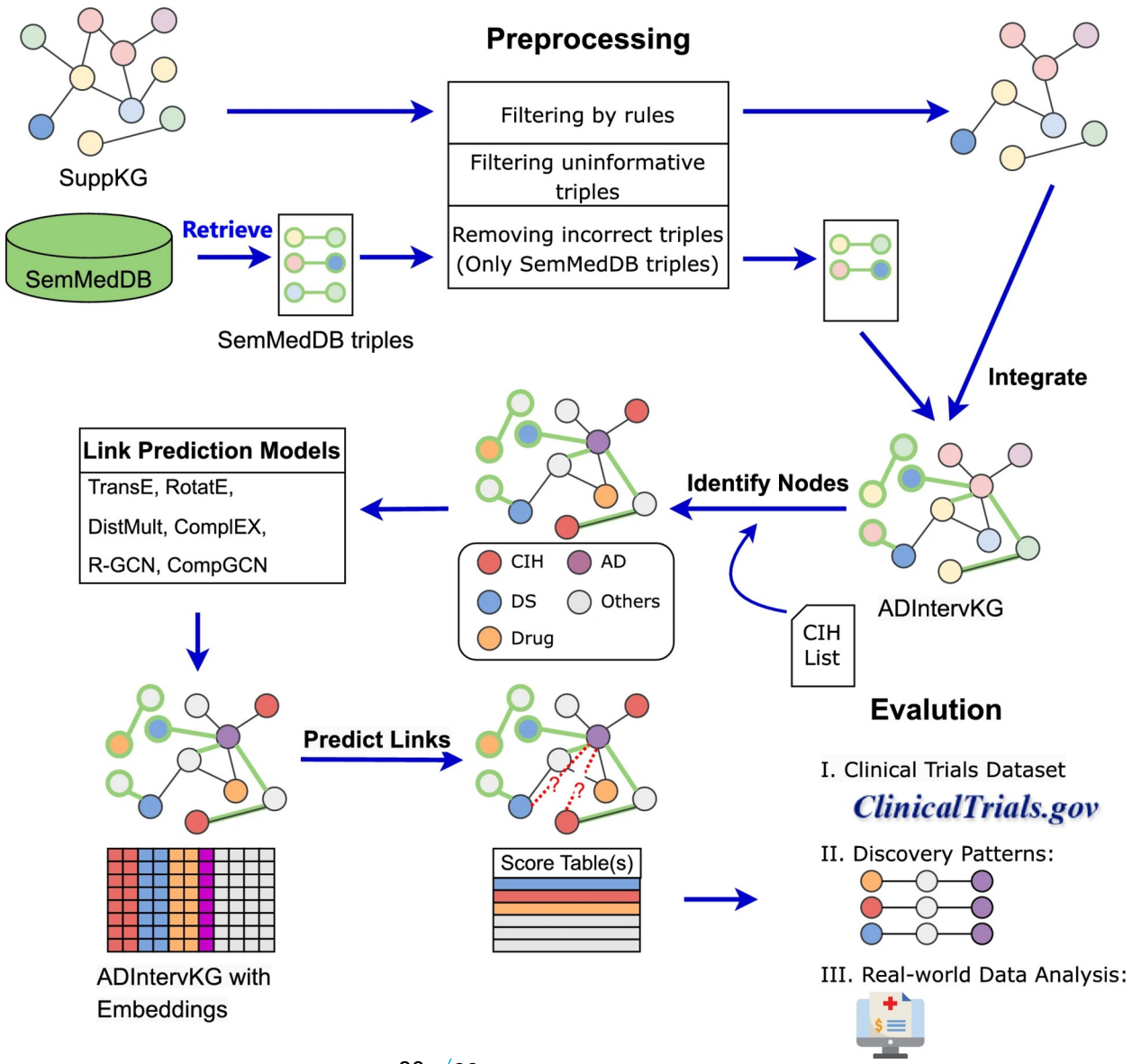
- **Benchmarked Models:**
  - **Distance/Matching:** TransE, RotatE, DistMult, ComplEx
  - **Relational GNNs:** R-GCN, CompGCN
- **Multi-Perspective Evaluation:**
  - Time-slicing split (chronological cutoff)
  - Benchmark against registered clinical trials (ClinicalTrials.gov ground truth)

PROBLEM & RATIONALE

Clinical Context & Need

- **Alzheimer's Disease (AD)** lacks curative pharmacological options; drug trials suffer high attrition.
- **Non-Pharmacological Interventions (NPIs):** Dietary Supplements (DS) and Complementary & Integrative Health (CIH) offer promising **neuroprotective potential** with **minimal side effects.**
- **Challenge:** Manual hypothesis exploration across massive biomedical literature is a major bottleneck.

The Alzheimer Disease Prevention Use Case



SemMedDB

Repository of semantic triples extracted from PubMed abstracts and titles using the SemRep program

SuppKG

Comprises 56,635 nodes and 595,222 directed edges, including 2928 DS-specific nodes and 164,738 edges

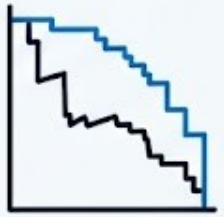
The Alzheimer Disease Prevention Use Case

MULTI-LEVEL VALIDATION: PREDICTIVE METRICS, DISCOVERED CANDIDATES & RWD

Superiority of Relational Graph Convolutional Networks & Real-World Clinical Confirmation

Model Performance: R-GCN Leads Across Benchmarks

- **Time-Slicing Benchmark:** Hits@10 = **0.8496** | Mean Rank = **5.205** | AUROC = **0.74**
- **Clinical Trials Ground Truth:** Hits@10 = **0.9192** | Mean Rank = **3.500** | AUROC = **0.79**
- **Key Insight:** Multi-relational neighborhood aggregation via R-GCN significantly outperforms translational and bilinear embedding models.



TOP DISCOVERED NPI CANDIDATES (Relation: PREVENTS)

In-Silico Candidate Prioritization

Dietary Supplements (DS)

- **Desmodii herba** (Confidence Score: 0.976) #1
- **Choerospondias axillaris** (Nepali Hog Plum - antioxidant) #2
- **Glucomannan** (Prebiotic gut-brain axis modulator) #3
- **Lutein** (Carotenoid linked to lower neuro-pathology)

Complementary & Integrative Health (CIH)

- **Mindfulness relaxation** (Confidence Score: 1.000) #1
- **Interpersonal psychothera** (Score: 0.406) #2
- **Photodynamic therapy (PDT)** (Score: 0.379) #3

MECHANISTIC PATHWAYS & REAL-WORLD EVIDENCE (RWD)

Explainability & Clinical Translation

Closed Discovery Biological Pathways

- **Choerospondias axillaris:** Downregulates pro-inflammatory cytokines TNF & IL-6.
- **Photodynamic Therapy (PDT):** Modulates neuro-immune cascades.

EHR Cohort Validation (10,844 MCI Patients at UMN)

- **Psychotherapy:** Slower conversion rate to ADRD (HR = 0.79, 95% CI: 0.58–1.07; $p < 0.001$).
- **Manual Therapy Techniques:** Reduced cognitive decline trend (HR = 0.81, 95% CI: 0.53–1.23).



The Alzheimer Disease Use Case

TAKE-HOME MESSAGES: KNOWLEDGE GRAPHS & GNNs SHAPING PREVENTIVE MEDICINE

Bridging Unstructured Literature, Predictive AI, and Mechanistic Interpretability

Scope Expansion

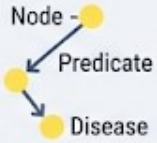
1. EXPANDING BEYOND DRUGS



- First proof of concept demonstrating Literature-Based Discovery (LBD) scales to non-pharmacological interventions (NPIs).
- Extends predictive knowledge graphs from synthetic small molecules to nutrition, dietary supplements, and lifestyle modalities.
- Unlocks systematic screening for complex, multifactorial neurodegenerative conditions.

Neuro-Symbolic Synergy

2. GNNs + EXPLAINABLE PATTERNS



- Relational Graph Convolutional Networks (R-GCN) resolve combinatorial bottlenecks to prioritize candidate interventions.
- Closed-discovery semantic path traversal restores mechanistic transparency.
- Converts opaque vector representations into actionable, biologically plausible hypotheses validated by clinical experts.

Translation Blueprint

3. THE TRIANGULATION BLUEPRINT



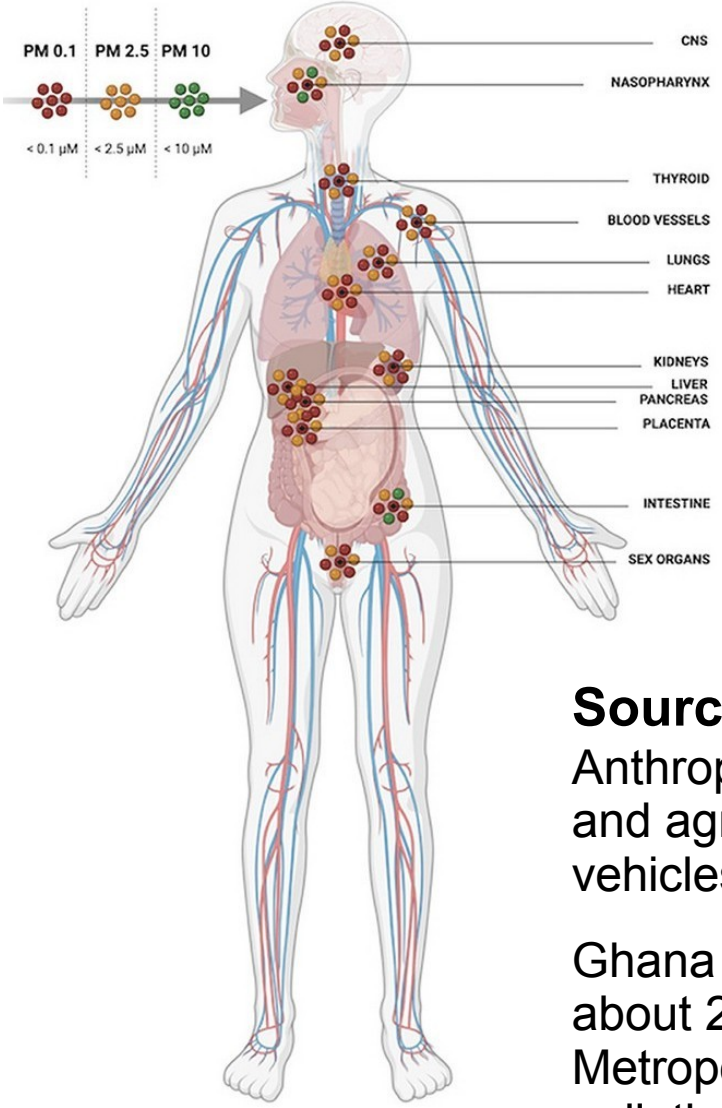
- Combines NLP fact-filtering, inductive graph learning, and real-world clinical cohort validation.
- Establishes a reproducible methodology to derisk and accelerate translational trials in public health.



Agenda

- Drug Discovery: Repositioning of Molecules
- Repurposing non-pharmacological interventions
- Environmental Health and Exposure Characterization

Evidence from the literature



Air Polluants	Diseases
Particulate matter (PM10 ,PM2.5)	Stroke, asthma, bronchitis, heart attacks, lung diseases, cancer and premature death from heart ailments, impairment in brain development in children
Nitrogen Oxides (NO and NO2)	Bronchitis asthma, heart disease
Black Carbon (BC)	Chronic obstructive pulmonary diseases, bronchitis, asthma, hypertension, cancer
Sulphur dioxide	Respiratory infection, eye irritation, asthma
Ozone	Chest pain, coughing, chronic obstructive pulmonary disease

Sources of PM2.5

Anthropogenic activities such as open solid (household and agricultural) waste burning and emissions from vehicles

Ghana Environmental Protection Agency estimates that about 2800 deaths annually in the Greater Accra Metropolitan Area (GAMA) can be attributed to PM2.5 pollution

WHO References About Pollutants (Air, Cov)

World Health Organization

Seuils de référence OMS recommandés en 2021 par rapport à ceux figurant dans les lignes directrices sur la qualité de l'air de 2005

Polluant	Durée retenue pour le calcul des moyennes	Seuils de référence OMS 2005 (ref) Concentrations	Seuils de référence OMS 2021 (ref) Concentrations
$PM_{2.5}$ (µg/m³)	Année	10	5
	24 heures [a]	25	15
PM_{10} (µg/m³)	Année	20	15
	24 heures [a]	50	45
NO_2 (µg/m³)	Année	40	10
	24 heures [a]	--	25
O_3 (µg/m³)	Pic saisonnier [b]	--	60
	8 heures [a]	100	100
SO_2 (µg/m³)	24 heures [a]	20	40
CO (mg/m³)	24 heures [a]	--	4

Sources :

WHO 2006, Air quality guidelines: Global update 2005

WHO 2021 Air quality guidelines: Global update 2021.

Informations :

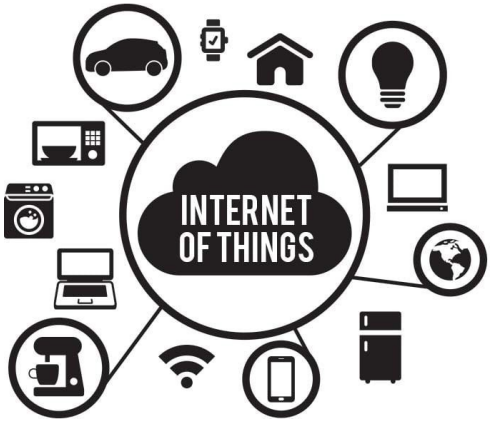
µg = microgramme

[a] 99ème percentile (3 à 4 jours de dépassement par an).

[b] Moyenne de la concentration moyenne journalière maximale d'O_3 sur 8 heures au cours des six mois consécutifs où la concentration moyenne d'O_3 a été la plus élevée.

Source: Site Santé Publique France

Heterogenous data as an opportunity to rely on AI-based approaches to solve health issues



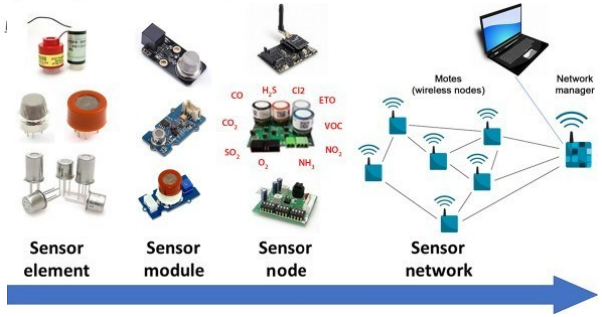






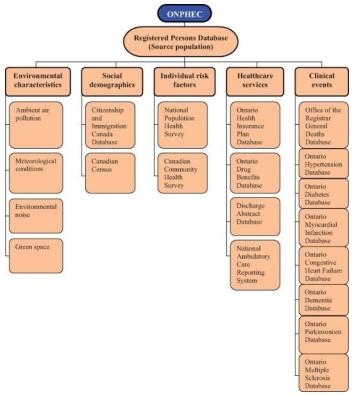






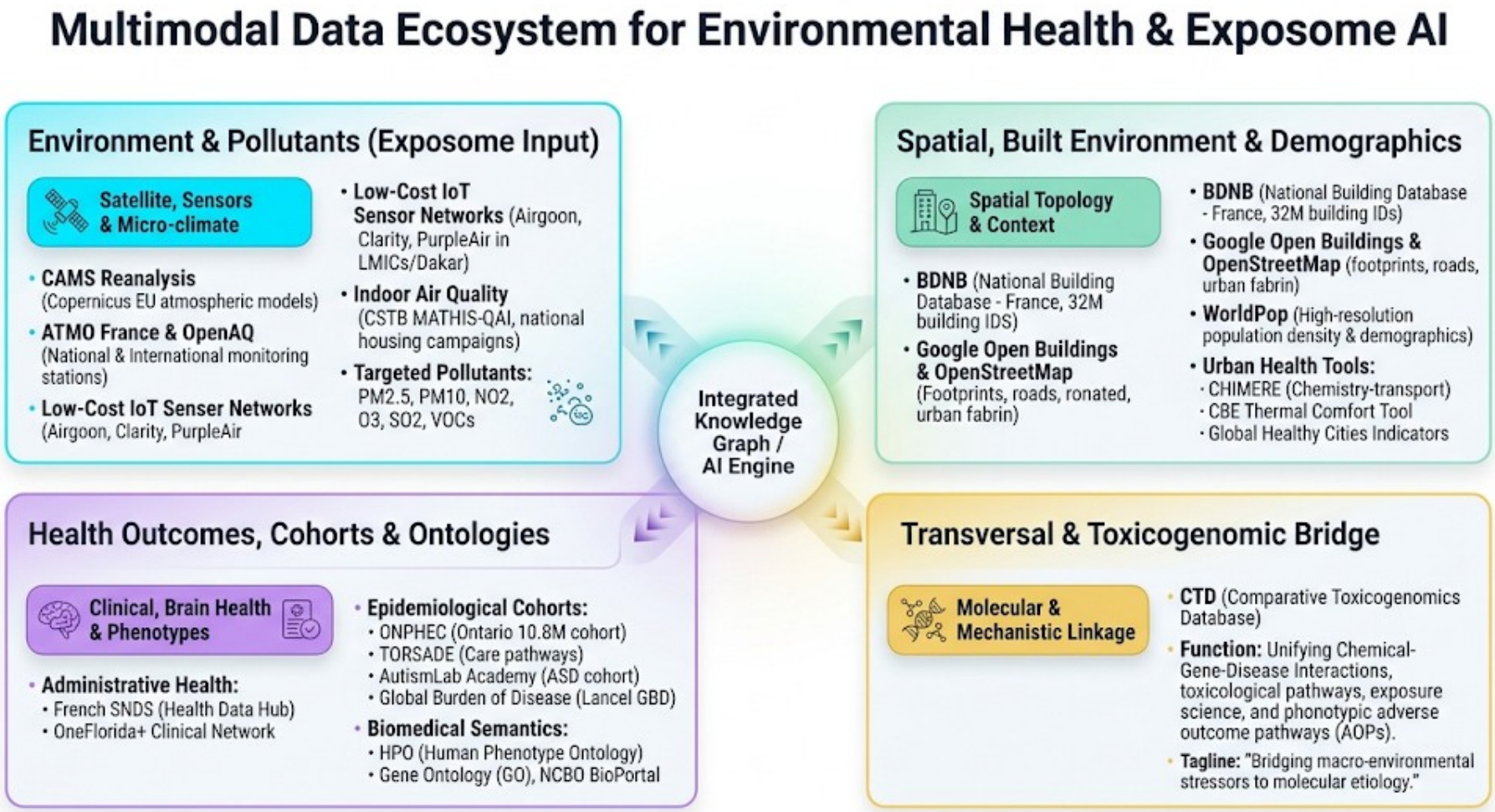
Source:[10.1016/B978-0-12-814827-3.00012-8](https://doi.org/10.1016/B978-0-12-814827-3.00012-8)







Heterogenous data as an opportunity to rely on AI-based approaches



PM2.5 Data From a Low-Cost Sensor

DateTime	Average	Université de Labé A
15/11/2023 13:50	29,3	21,29
15/11/2023 14:00		23,95
15/11/2023 14:10		18,377
15/11/2023 14:20		21,468
15/11/2023 14:30		17,33
15/11/2023 14:40		17,846
15/11/2023 14:50		18,205
15/11/2023 15:00		22,734
15/11/2023 15:10		22,434
15/11/2023 15:20		20,448
15/11/2023 15:30		20,232
15/11/2023 15:40		20,062
15/11/2023 15:50		18,538
15/11/2023 16:00		18,206
15/11/2023 16:10		18,164
15/11/2023 16:20		21,565
15/11/2023 16:30		23,926
15/11/2023 16:40		22,442
15/11/2023 16:50		21,352
15/11/2023 17:00		21,722
15/11/2023 17:10		21,98
16/11/2023 17:50		35,994
16/11/2023 18:00		33,328
16/11/2023 18:10		37,546
16/11/2023 18:20		42,198
16/11/2023 18:30		46,522
16/11/2023 18:40		38,308
16/11/2023 18:50		37,034
16/11/2023 19:00		40,662
16/11/2023 19:10		40,904
16/11/2023 19:20		45,24
16/11/2023 19:30		47,012
16/11/2023 19:40		52,04
16/11/2023 19:50		47,476
16/11/2023 20:00		47,58
16/11/2023 20:10		58,7
16/11/2023 20:20		55,066
16/11/2023 20:30		60,494
16/11/2023 20:40		80,834
16/11/2023 20:50		60,222
16/11/2023 21:00		49,946
16/11/2023 21:10		46,478
16/11/2023 21:20		46,24
16/11/2023 21:30	29,3	45,464



Various topics addressed on air pollution & health using AI

CEO Clinical and Experimental Otorhinolaryngology Vol. 15, No. 2: 168-176, May 2022
https://doi.org/10.21053/ceo.2021.01536
pISSN 1976-8710 eISSN 2005-0720

Original Article

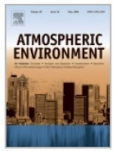
Machine Learning Models for I Occurrence of Respiratory Disease and Air-Pollution Fac

Yunseo Ku[1,*] · Soon Bin Kwon[2,*] · Jeong-Hwa Yoon[3] · Seog-Kyun

[1]Department of Biomedical Engineering, Chungnam National University College of Medicin
Columbia University, New York, NY, USA; [3]Institute of Health Policy and Management, Medic
Seoul; [4]Department of Otorhinolaryngology-Head and Neck Surgery, Chung-Ang Univ

ELSEVIER

Atmospheric Environment
Volume 40, Issue 16, May 2006, Pages 2845-2851



An integrated neural network model for PM10 forecasting

Patricio Perez, Jorge Reyes

Air pollution and hospital admissions for cardiorespiratory diseases in Iran: artificial neural network versus conditional logistic regression

M. Shakerkhatibi, I. Dianat, M. Asghari Jafarabadi, R. Azak & A. Kousha

International Journal of Environmental Science and Technology 12, 3433–3442 (2015)
article

Graph Learning Techniques Using Structured Data for IoT Air Pollution Monitoring Platforms

Publisher: IEEE Cite This PDF

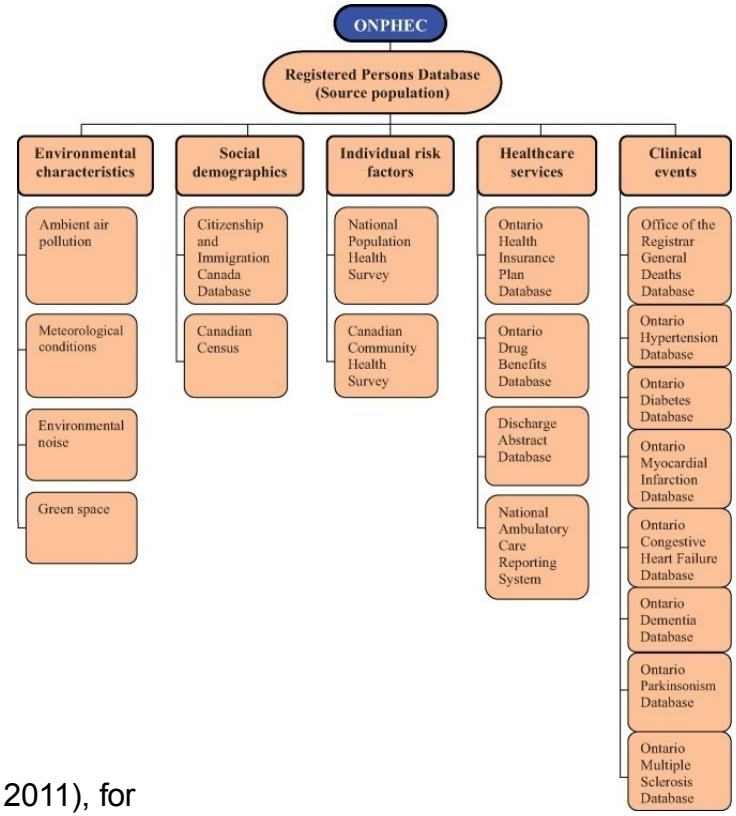
Pau Ferrer-Cid ; Jose M. Barcelo-Ordinas ; Jorge Garcia-Vidal All Authors

The ONtario Population Health and Environment Cohort (ONPHEC)

A large, retrospective cohort in the province of Ontario Canada (population of 10.8 million in 1996), created in 2014

Inclusion criteria: Canadian-born population of Ontario who were 35 years or older in 1996 (~4.9 million)

Objective: to investigate the independent and combined effects of environmental stressors (such as air pollution, traffic-related noise) on the incidence of chronic diseases and their interactions with 'healthy' environmental factors (e.g. green areas)



Specific objectives

- to determine the effects of environmental exposures on the long-term survival of adults with selected chronic diseases;
- to quantify the burden of disease from adverse environmental factors;
- to evaluate the impacts of changes in environmental exposure on disease risk such as moving to less polluted neighbourhoods or vice versa;
- to identify population subgroups who are most susceptible to the effects and who are disproportionately exposed to environmental stressors

Examples of studies and findings:

- Among 62 012 adults (population-based health surveys between 1996 and 2011), for every **10-µg/m3 increase in exposure to PM2.5**, the incidence of **diabetes increased by 11% [95% confidence interval (CI): 2-21%**)
- every **10-µg/m3 increase** in exposure to **PM2.5** was associated with a **22% increase** in risk of non-accidental deaths, **43% increase in deaths from ischaemic heart disease** and **64% increase in AMI-related deaths**

Chen H, Kwong JC, Copes R, Villeneuve PJ, Goldberg MS, Ally SL, Weichenthal S, van Donkelaar A, Jerrett M, Martin RV, Brook JR, Kopp A, Burnett RT. Cohort Profile: The ONtario Population Health and Environment Cohort (ONPHEC). Int J Epidemiol. 2017 Apr 1;46(2):405-405j. doi: 10.1093/ije/dyw030.

Graphs-based air quality data analysis and possible impacts on health

Context

A study of U. Waterloo in Mongolia

- UNIVERSITY OF WATERLOO
-

A real-time monitoring platform
An early warning tool



Use of Internet of things (IOT) and big data to quantify outcomes and health impacts of air pollution

Dr. Plinio Morita plinio.morita@uwaterloo.ca

Shahan Salim shahan.salim@uwaterloo.ca

NSERC CRSNG · UBILAB · UNIVERSITY OF WATERLOO | FACULTY OF HEALTH

Data Sources for Prediction System

Hospital
- ICD 10 Codes
- Time of visit
- Location
- Age
- Gender

ER Visit
- ICD 10 Codes
- Time of visit
- Location
- Age
- Gender

Ambulance
- ICD 10 Codes
- Time of call
- Pickup location
- Age
- Gender

Education
- School name
- Location
- Date
- # of students absent
- Gender

Air Quality
- PM 2.5
- PM 10

UNIVERSITY OF WATERLOO

The PANORAMA Use Case

Outdoor Air Quality and Health Impact: The PANORAMA Knowledge Graph Based Approach

Nareesa Karmali, Abdougafarou Mamam, and Gayo Diallo(✉)

Team AHeaD, BPH Inserm 1219, Univ. Bordeaux, 33000 Bordeaux, France
{Nareesa.Karmali,Abdougafarou.Mamam,Gayo.Diallo}@u-bordeaux.fr

Abstract. Current air quality monitoring systems struggle to integrate heterogeneous historical data into AI algorithms that could provide deeper insights into severe pollution impacts. We present PANORAMA, a knowledge graph approach that predicts links between pollutant exposure and health outcomes by integrating diverse air quality and health datasets. Using an Extract, Transform, and Load process, we incorporated discrete data into a knowledge graph and evaluated its inferential capabilities through embedding models. Our case study in France's Gironde department yielded promising results (MR: 11.0, MRR: 0.403, Hits@1/3/10: 0.326/0.447/0.499) using the Rotate model, demonstrating the potential of semantic structuring and knowledge graph technology for environmental public health applications.

Keywords: Knowledge Graph · Links Prediction · Environmental Health

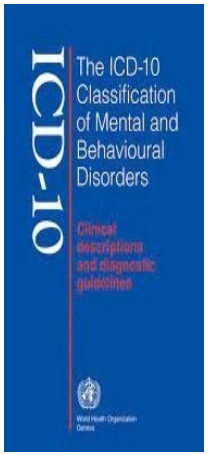
Karmali, N., Mamam, A., Diallo, G. (2026). Outdoor Air Quality and Health Impact: The PANORAMA Knowledge Graph Based Approach. In: Nguyen, N.T., et al. Computational Collective Intelligence. ICCCI 2025. Lecture Notes in Computer Science(), vol 16139. Springer, Cham. https://doi.org/10.1007/978-3-032-09321-9_3

The PANORAMA Use Case

Dataset





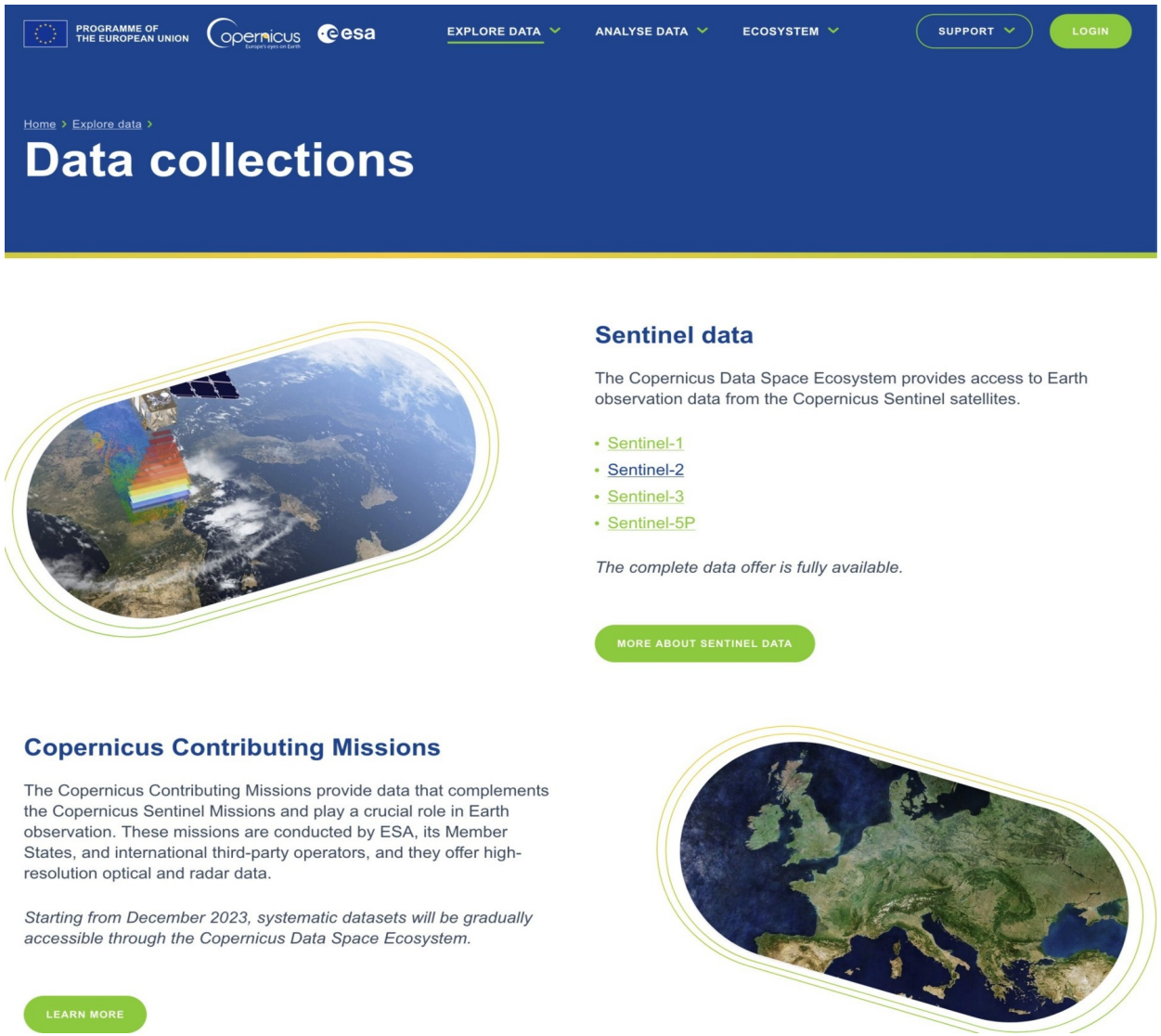








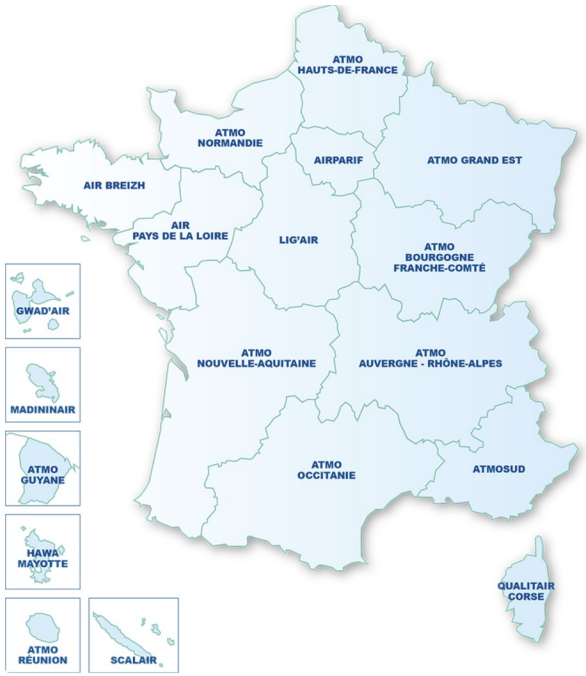
Copernicus Data and Information Access Services (DIAS)



https://atmosphere.copernicus.eu/data

Ground Sensors Measurement: ATMO example

Fédération des Associations agréées de surveillance de la qualité de l'air



https://www.atmo-france.org/article/laasqa-de-votre-region

votre parten'air Atmo NOUVELLE-AQUITAINE

Votre observatoire de la qualité de l'air en Nouvelle-Aquitaine

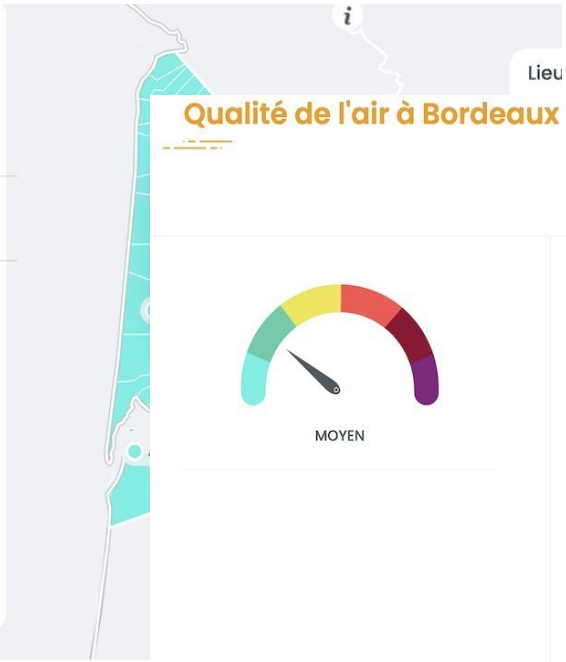
Les indices de la qualité de l'air dans votre région

Carte mise à jour le 15/11/2023 - 10:50

AUJOURD'HUI | 16 NOV

commune ou code postal

- Indisponible
- Bon
- Moyen
- Dégradé
- Mauvais
- Très mauvais
- Extrêmement mauvais
- Événement



Lieu

Qualité de l'air à Bordeaux

MOYEN

Données mises à jour le 15/11/2023 - 10:50

Particules fines PM2,5	Particules PM10
MOYEN	BON

Ozone O_3	Dioxyde d'azote NO_2
MOYEN	MOYEN

Dioxyde de soufre SO_2

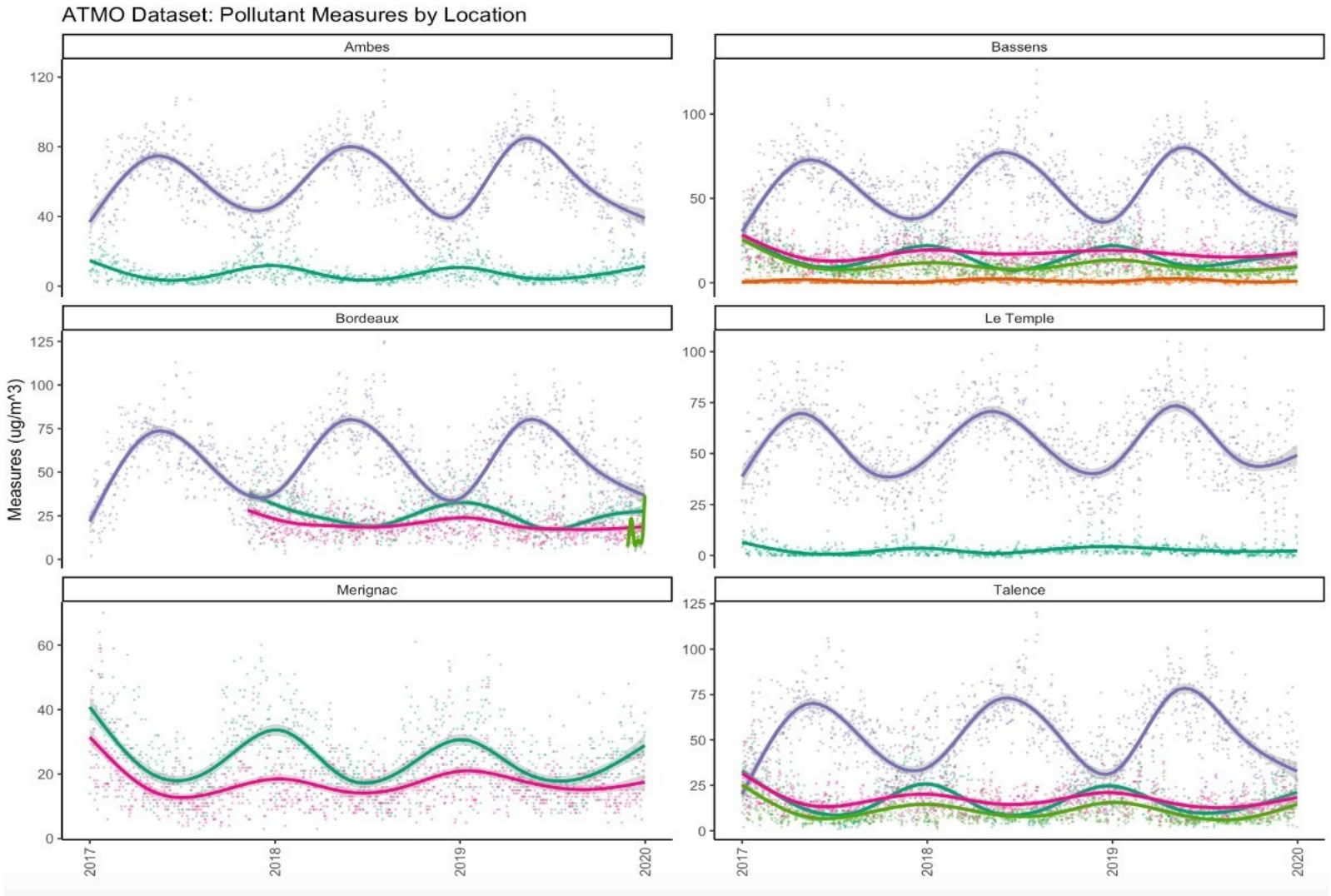
BON

Légende

Bon | Moyen | Dégradé | Mauvais | Très mauvais | Extrêmement mauvais | Événement | Indisponible

https://nouvelleaquitaine.org://www.atmo-

Ground Sensors Measurement: ATMO example



Source: M Thibault, X. Wan (Web-based data TU, PHDS)

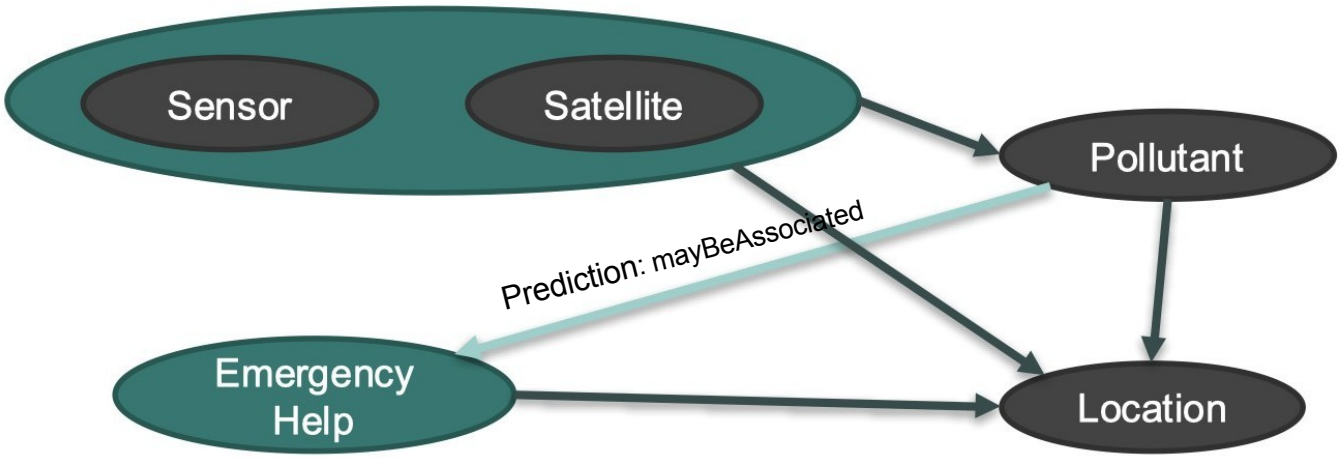
Health Data: SAMU Calls

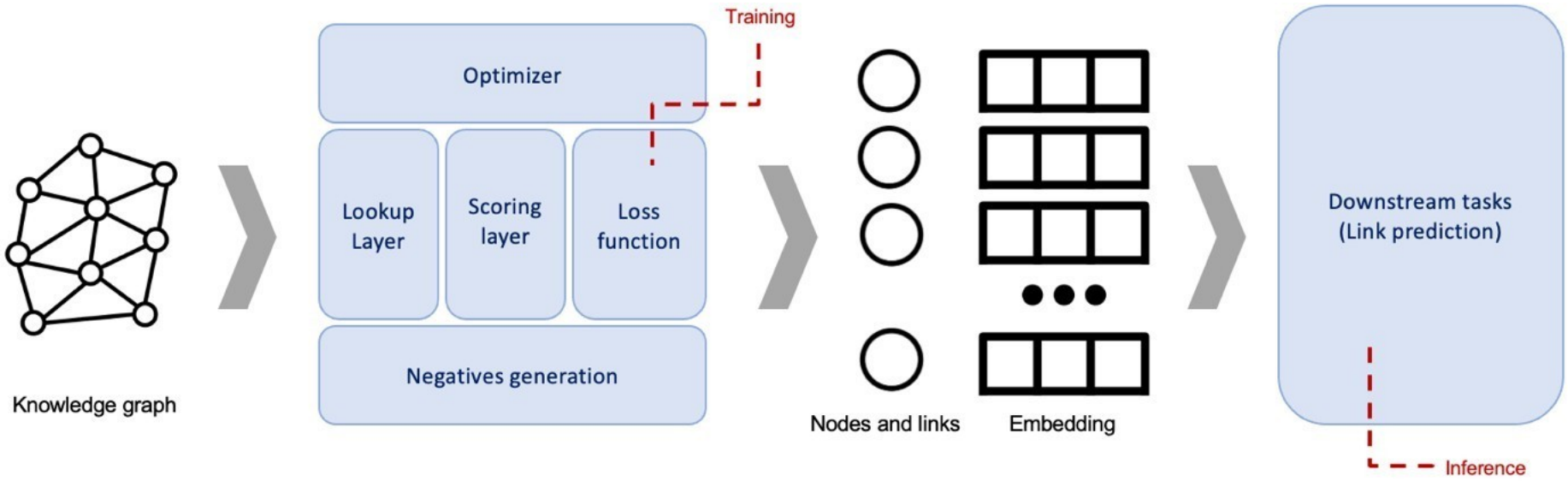


date	main_complaint	INSEE_area	age	sexe	diagnostic_short	diagnostic
2017-01-01	autre malade	XXXXXXXXX	41	masculin	K92.0	K92.0 HEMATEMES
2017-01-01	autre malade	XXXXXXXXX	1.5	fiminin	A03.01	A03.01 FIEVRE | ET
2017-01-01	autre malade	XXXXXXXXX	19	masculin	D02	D02 DOULEUR EST(
2017-01-01	autre malade	XXXXXXXXX	90	fiminin	R78.01	R78.01 BRONCHITE

Graph-based learning and prediction

Approach





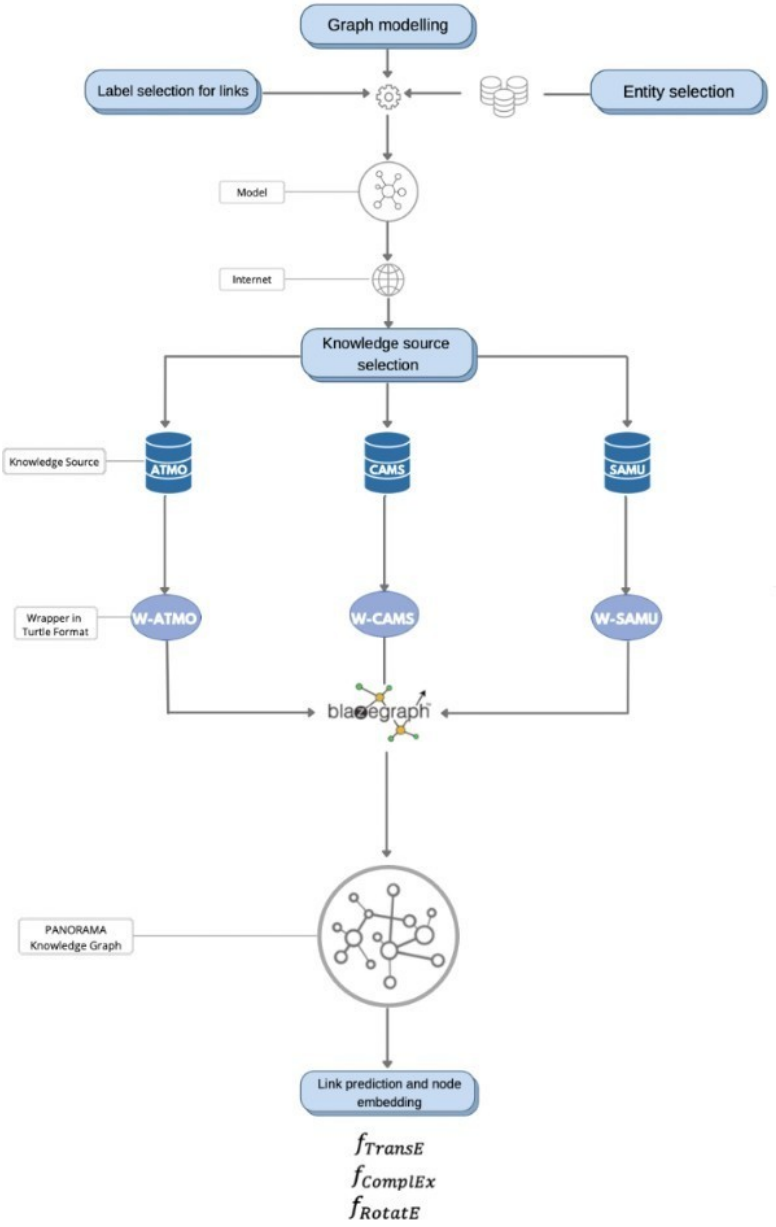
Panorama: Graph-based Learning and Prediction

Question/Design

Extract & Transform

Load

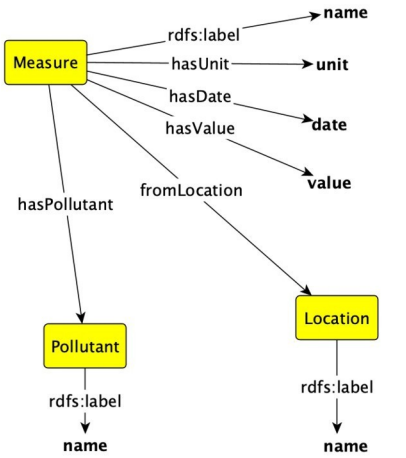
Embeddings and Analysis



Panorama: Graph-based Learning and Prediction

Atmo – Ground sensor air quality data in Gironde, France

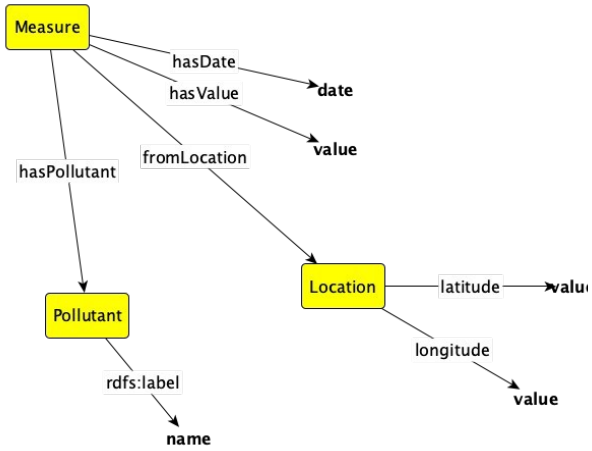
ATMO



Variables: measure, unit, pollutant, date, location, value

Copernicus – European Union satellite air quality data

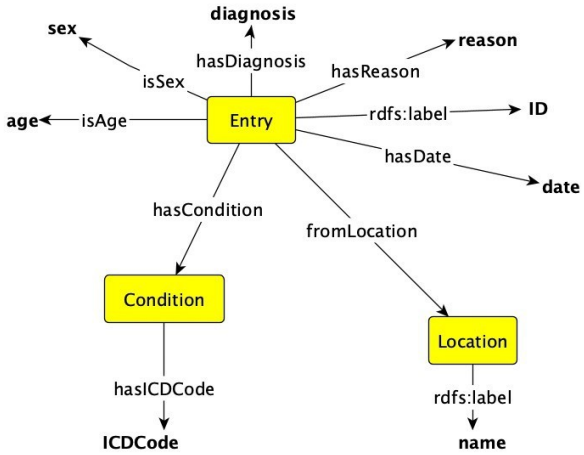
CAMS



Variables: latitude, longitude, pollutant, date, value

SAMU Urgences de France - first emergency help data in Gironde, France

SUDF



Variables: date, sex, age, ID, location, diagnosis, suspected diagnosis (ICD Code and longer diagnosis), reason

Panorama: Graph-based Learning and Prediction

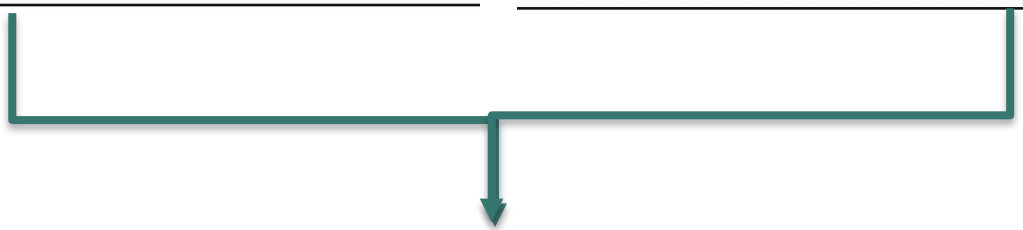
ENTITY NORMALIZATION

Normalizing locations from Atmo Nouvelle-Aquitaine to INSEE vocabulary

Atmo Sensor Locations	INSEE Commune	INSEE Code
Bordeaux-Gautier	Bordeaux	33063.0
Merignac	Merignac	33281.0
Talence	Talence	33522.0

Normalizing locations from CAMS to INSEE vocabulary

CAMS Coordinates	INSEE Commune	INSEE Code
44.8°N, -0.6°W	Bordeaux	33063.0
44.8°N, -0.7°W	Merignac	33281.0
44.8°N, -0.5°W	Talence	33522.0



Assigning INSEE communes to pre-defined INSEE uniform resource identifiers

INSEE Commune	INSEE URI
Bordeaux	<http://id.insee.fr/geo/commune/acebd21b-98a1-4a37-b584-d5956f719e7f>
Merignac	<http://id.insee.fr/geo/commune/aa63f8a3-49a2-41e3-a039-36dc639be4a1>
Talence	<http://id.insee.fr/geo/commune/1941bf43-b35d-48d0-8bc4-fd9cdeb312c4>

Table 3. INSEE communes mapped to CAMS, Atmo, and SUDF data were assigned to INSEE URIs available from the INSEE RDF space's pre-defined vocabulary.

Panorama: Graph-based Learning and Prediction

ETL EXTRACT

The integration process follows the Extract, Transform and Load (ETL)principle:

1. **Extract the data from knowledge sources**
2. Transform it RDF language via Turtle syntax
3. Load into Blazegraph graph database

```
# Create the measureConcept and entryConcept node
entryConcept = URIRef("http://panorama.org/Entry")
measureConcept = URIRef("http://panorama.org/Measure")
```

Creating a Uniform Resource Identifier (URI): a string of characters that uniquely identifies each tuple

```
hashtable['Bordeaux'] = 'acebd21b-98a1-4a37-b584-d5956f719e7f'
hashtable['Talence'] = "1941bf43-b35d-48d0-8bc4-fd9cdeb312c4"
hashtable['Merignac'] = "aa63f8a3-49a2-41e3-a039-36dc639be4a1"
hashtable['Le Temple'] = "4670f669-1fcb-4964-ac37-0be600d41f7d"
hashtable['Bassens'] = "661de373-48aa-4a22-b2ac-a24475ba2dbf"
hashtable['Ambes'] = "9518d867-9851-4ec0-a26a-ca07cbd793eb"
```

Hashing: assigning a fixed sized integer that identifies a particular location according to INSEE RDF Vocabulary

```
location = URIRef("http://panorama.org/Location/"+hashtable[row["INSEE_area"]])
icd = URIRef("http://purl.bioontology.org/ontology/ICD10/"+Literal(row["diagnostic_short"]))
```

Assigning nodes to namespaces: According to INSEE and ICD Vocabulary

```
g.add((entry, RDF.type, entryConcept))
g.add((entry, pano.fromLocation, location))
```

Adding nodes and triples: assigning nodes and integrating interactions

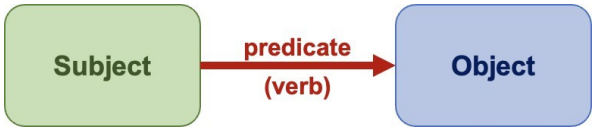
Panorama: Graph-based Learning and Prediction

ETL – TRANSFORM

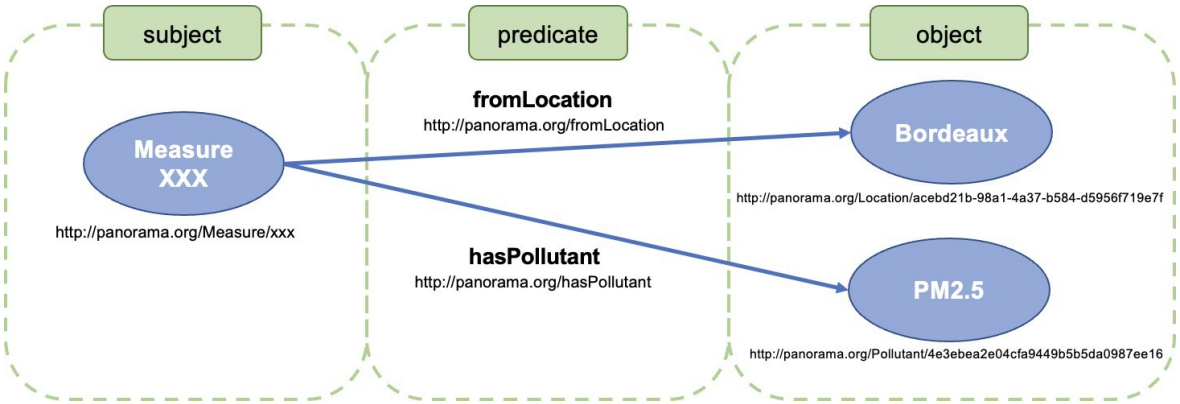
Station	Pollutant	Measure	Unit	Date	Value
Grand Parc	dioxyde d'azote (NO2)	Dioxyde d'azote	µg/m3	2020-01-01	16
Grand Parc	dioxyde d'azote (NO2)	Dioxyde d'azote	µg/m3	2020-01-02	23
Grand Parc	dioxyde d'azote (NO2)	Dioxyde d'azote	µg/m3	2020-01-03	26
Grand Parc	dioxyde d'azote (NO2)	Dioxyde d'azote	µg/m3	2020-01-04	16
Grand Parc	dioxyde d'azote (NO2)	Dioxyde d'azote	µg/m3	2020-01-05	30
Grand Parc	dioxyde d'azote (NO2)	Dioxyde d'azote	µg/m3	2020-01-06	24
Grand Parc	dioxyde d'azote (NO2)	Dioxyde d'azote	µg/m3	2020-01-07	32



rdflib 6.1.1



1. Extract the data from knowledge sources
2. **Transform it into RDF language via Turtle syntax**
3. Load into Blazegraph graph database



Panorama: Graph-based Learning and Prediction

1. Wrappers extract the data from knowledge sources
2. Transform it into RDF language via Turtle syntax
3. **Load into Blazegraph graph database**

Blazegraph is an open source Triplestore and graph database for storing and querying linked data

Supports RDF/SPARQL APIs

SPARQL: to query information from data mapped to RDF

ETL – LOAD

```
<http://panorama.org/Entry/c40467cd-5f2e-4135-9eab-f5b557d60734> a pano:Entry ;
    rdfs:label "4317165" ;
    pano:fromLocation <http://panorama.org/Location/acebd21b-98a1-4a37-b584-d5956f719e7f> ;
    pano:hasDate "2018-10-08"^^xsd:date ;
    pano:hasDiagnosis "R50.9 FIEVRE SP (SFUM)" ;
    pano:hasICD <http://purl.bioontology.org/ontology/ICD10/R50.9> ;
    pano:hasReason "autre malade" ;
    pano:isAge "84.0" ;
    pano:isSex "masculin" .
```

Outgoing Links

<http://panorama.org/fromLocation>	<http://panorama.org/Location/acebd21b-98a1-4a37-b584-d5956f719e7f>
<http://panorama.org/hasICD>	<http://purl.bioontology.org/ontology/ICD10/R50.9>
rdf:type	<http://panorama.org/Entry>

No Incoming Links

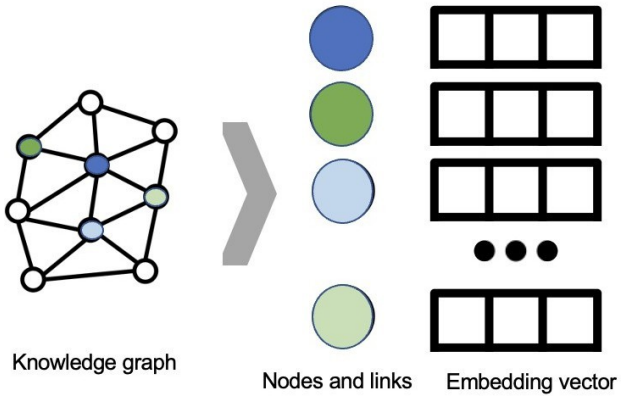
Attributes

<http://panorama.org/hasDate>	2018-10-08
<http://panorama.org/hasDiagnosis>	R50.9 FIEVRE SP (SFUM)
<http://panorama.org/hasReason>	autre malade
<http://panorama.org/isAge>	84.0
<http://panorama.org/isSex>	masculin
rdfs:label	4317165

Panorama: Graph-based Learning and Prediction

Maps entities and relations from a KG into a **low-dimensional and continuous vector space**, on which other operations can be conducted

Goal: encode nodes so that **similarities in the original graph network are approximated by similarities in the embedding space**



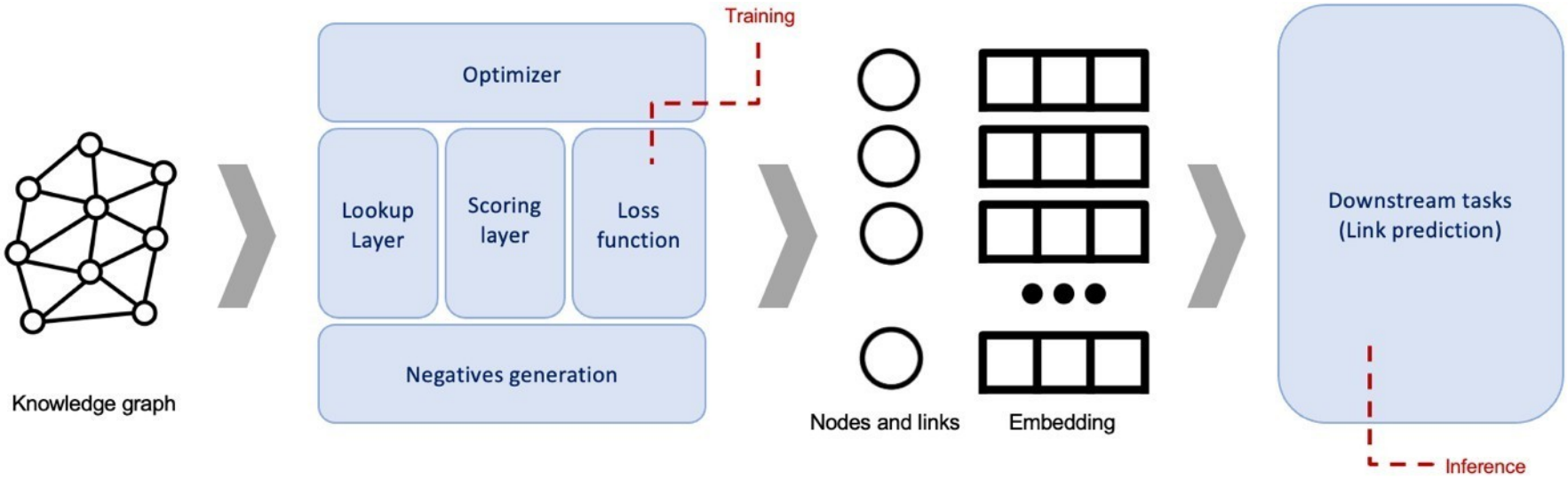
MACHINE LEARNING TASKS & INDUCTIVE REASONING

$$sim(u, v) \approx z_v^T z_u$$

Similarity in the original network

Similarity of the embedding

Panorama: Knowledge Graph Embeddings



Lookup layer (simply assigns an embedding to each node and edge type of the graph)
Scoring layer interacting with a **loss function Negatives generation**
Optimization to train embeddings, which are used by the above scoring function to predict new links (the downstream task)

Panorama: Knowledge Graph Embeddings

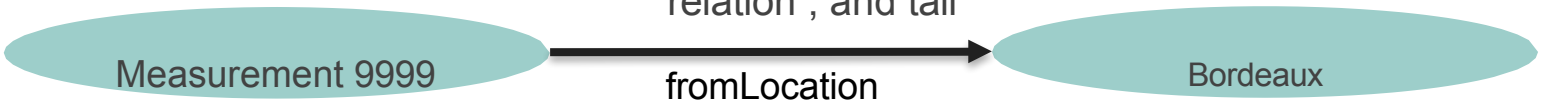
NEGATIVES GENERATION

OWA: all external triples are unknown (underfits)

CWA: all external triples are negative (overfits)

LCWA: KG is locally complete; corrupted triples serve as negatives

Negative triples generated by **triples corruption** of a known positive triple made of head , relation , and tail



Corrupt heads	Measurement 1234; fromLocation; Bordeaux
Corrupt relations	Measurement 9999; hasPollutant; Bordeaux
Corrupt tails	Measurement 9999; fromLocation; Merignac

Panorama: Knowledge Graph Embeddings

LINK PREDICTION

Main task used to evaluate KGEMs

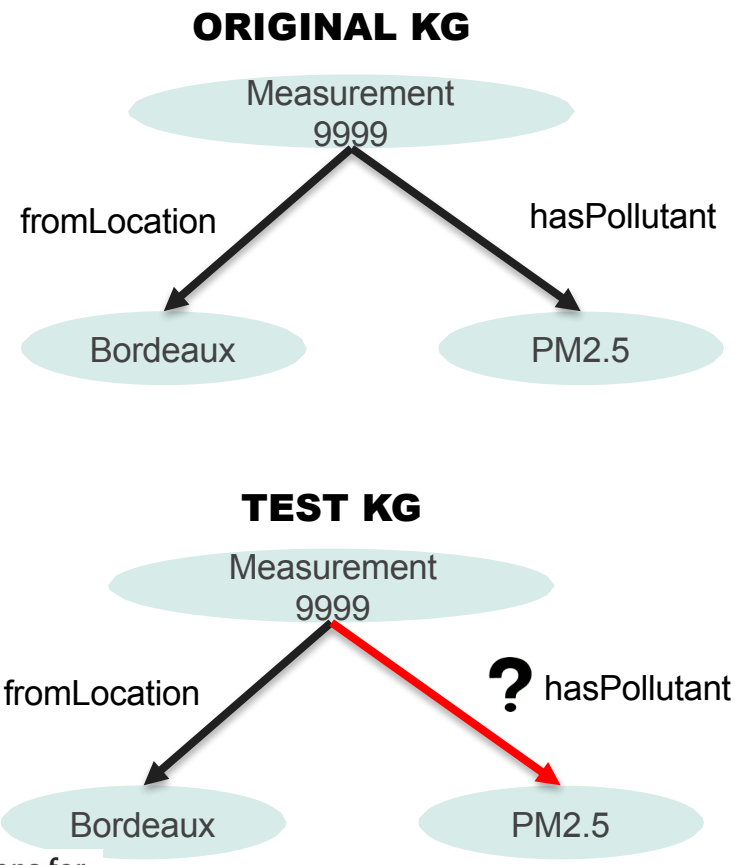
- Remove random set of links and aim to predict them
- Employs a **learning-to-rank problem**: how well positive triples rank against negatives.

Uses information retrieval metrics such as:

Mean Rank: the arithmetic mean over all individual ranks

Mean Reciprocal Rank: takes the reciprocal of the rank to guarantee robustness against outliers

Hits@K: fraction of positive facts that appear in the first k triples of the sorted rank list



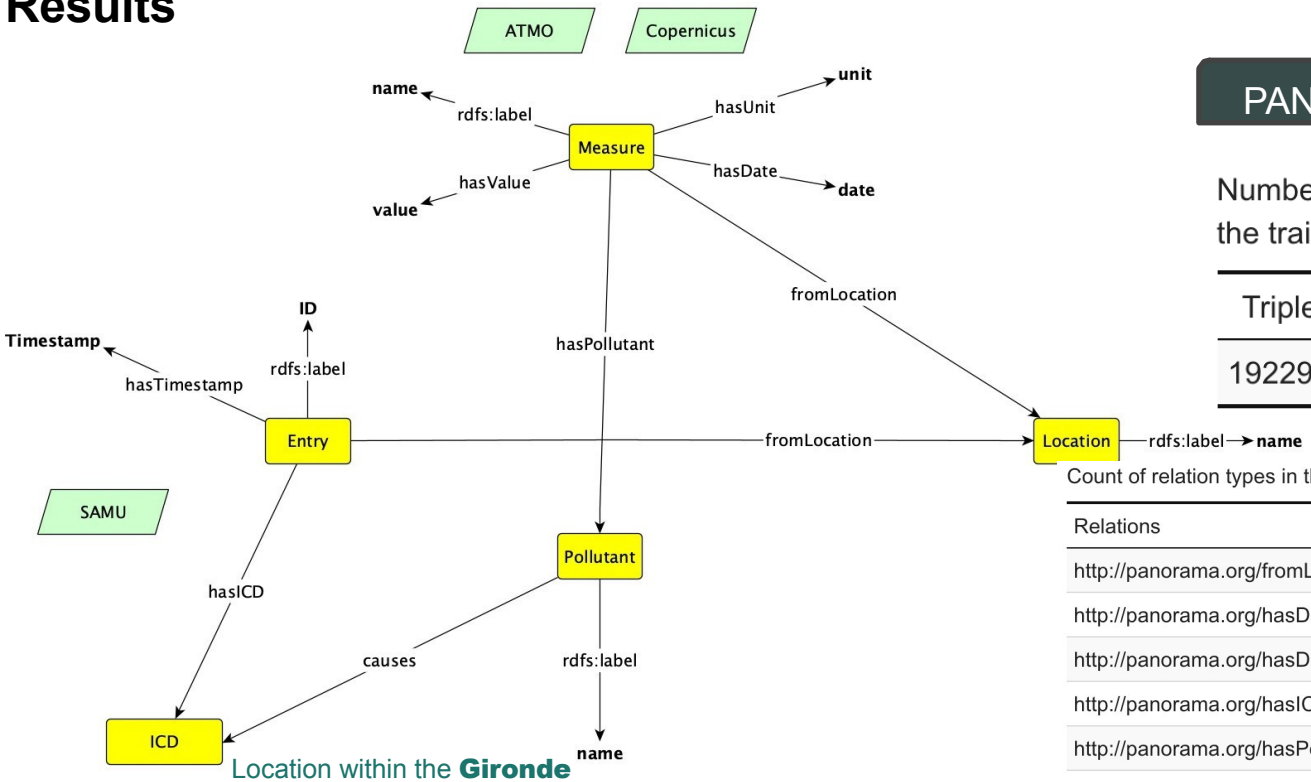
Repurposing Non-pharmacological Interventions for Alzheimer's Diseases through Link Prediction on Biomedical Literature

Yongkang Xiao, Yu Hou, Huixue Zhou, Gayo Diallo, Marcelo Fiszman, Julian Wolfson, Halil Kilicoglu, You Chen, Chang Su, Hua Xu, William G Mantyh, Rui Zhang

PMID: 37292731 PMCID: PMC10246059 DOI: 10.1101/2023.05.15.23290002

Panorama: Results

Results



Location within the **Gironde department** and dates between **January 1, 2017, and December 31, 2019.**

Two pre-defined acquired from RDF Vocabulary: rdf:type, and rdfs:label, instances of rdf:Property.

11 custom relations were defined:
hasValue, hasDate, hasUnit, hasPollutant, fromLocation, hasReason,
isSex, isAge, hasCondition, hasICDCode, hasDiagnosis.

PANORAMA Knowledge Graph

Number of triples, entities, relations, and the training and testing split

Triples	|E|	|R|	Training	Testing
1922998	443332	13	1634548	288450

Count of relation types in the PANORAMA knowledge graph

Relations	Count
http://panorama.org/fromLocation	238508
http://panorama.org/hasDate	238509
http://panorama.org/hasDiagnosis	171714
http://panorama.org/hasICD	171714
http://panorama.org/hasPollutant	66795
http://panorama.org/hasReason	170925
http://panorama.org/hasUnit	20805
http://panorama.org/hasValue	64297
http://panorama.org/isAge	170903
http://panorama.org/isSex	171329
http://panorama.org/mayHaveParticipant	2150
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	240672
http://www.w3.org/2000/01/rdf-schema#label	194676

Note:
Table 4. The number of each relation type present in the PANORAMA knowledge graph.

Panorama: Some Results

Hyper-parameters: embedding size: 50, batch size: 128

The **RotateE** model provides the best performance for the prediction

PyKEEN

Models	Epoch	Losses	Median Rank	Inverse median rank	Hit@		
					1	3	10
TransE	10	0.13472619404537617	68.000000	0.014706	0.147499	0.316142	0.400464
	50	0.0666156818259324	87.500000	0.011429	0.226147	0.284425	0.324136
	100	0.05804811778001313	30.000000	0.033333	0.205003	0.305054	0.417225
RotatE	10	0.14112802440598887	7.000000	0.142857	0.338577	0.465962	0.511862
	50	0.047476657357450186	16.000000	0.062500	0.288809	0.404848	0.483755
	100	0.04220049773200764	23.000000	0.043478	0.258381	0.402527	0.473182
ComplEx	10	0.22571184140252692	33423.000000	0.000030	0.000000	0.000000	0.001031
	50	0.11748231785320405	862.500000	0.001159	0.257091	0.364105	0.424703
	100	0.11884046556193759	389.000000	0.002571	0.243167	0.383187	0.465704

PubMed®

> J Cardiovasc Electrophysiol. 2019 Apr;30(4):487-492. doi: 10.1111/jce.13829. Epub 2019 Jan 6.

Ambient fine particulate matter (PM2.5) exposure is associated with idiopathic ventricular premature complexes burden: A cohort study with consecutive Holter recordings

Tsung-Ying Tsai [1], Li-Wei Lo [1 2], Shin-Huei Liu [1], Wen-Han Cheng [1], Yu-Hui Chou [1], Wei-Lun Lin [1 2], Yamada Shinya [1], Yenn-Jiang Lin [1 2], Shih-Lin Chang [1 2], Yu-Feng Hu [1 2], Fa-Po Chung [1 2], Jo-Nan Liao [1 2], Tze-Fan Chao [1 2], Ta-Chuan Tuan [1 2], Shih-Ann Chen [1 2]

Particulate Matter

(D052638)

mayBeAssociated

CATSHL syndrome

Ventricular Premature Complexes

Protein Aggregation Pathological

Glaucoma

Copper-Overload Cirrhosis

Skin Abnormalities

Gastritis Atrophic

Osteogenesis Imperfecta, Type IV

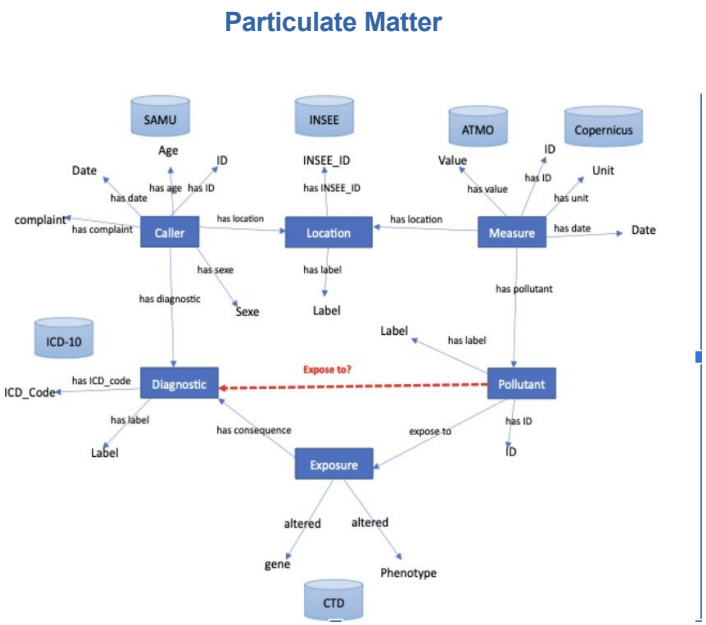
Thrombocythemia, Essential

Leukemia, Myeloid

Limitations of the study: cordinate precision disparities between data sources (e.g., CAMS versus Atmo sensors; lacks information on indoor pollutant exposure and exposure duration

Panorama: Some Results

Prediction of tail: head= Particulate Matter (D052638) ; relation= *MayHaveParticipant*



CATSHL syndrom

Ventricular Premature Complexes

Protein Aggregation Pathological

Glaucoma

Copper-Overload Cirrhosis

Skin Abnormalities

Gastritis Atrophic

Osteogenesis Imperfecta, Type IV

Thrombocythemia, Essential

Leukemia, Myeloid

Limitations of the Study

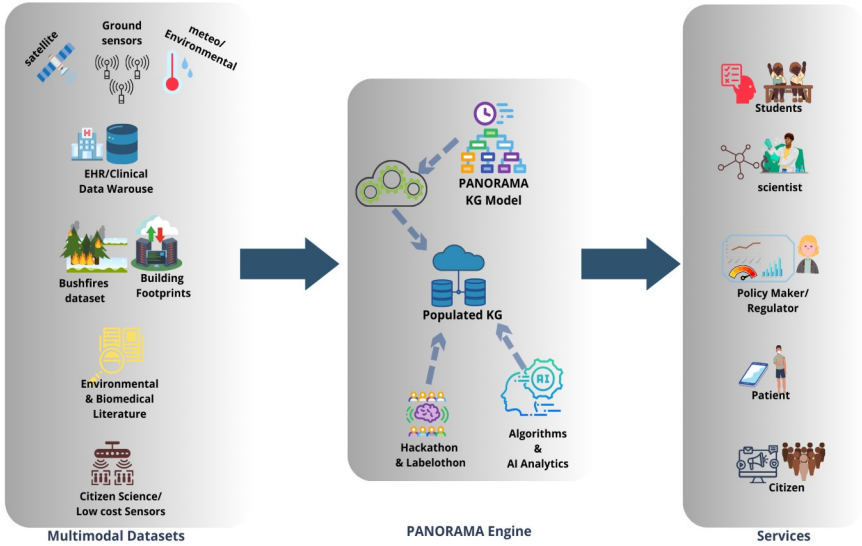
- The location is used as a proxy
  - The integrated data related to pollutant have different granularity: CAMS Vs SUDF
  - Missing data points due to sensors failure
- Ambiguity among datasets
  - Health data coded differently (CISP2 Vs ICD10) A need of improving the mappings
  -
- Limited evaluated embeddings and link prediction models

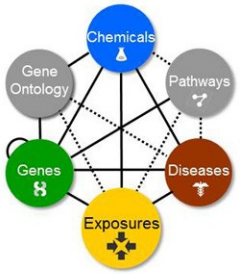
Ongoing Work: PhD Thesis Elisa Drouot (CSTB)

Health Data + Buildings Footprints

+ Indoor Environmental Quality, Temporal GNN, French Cohort

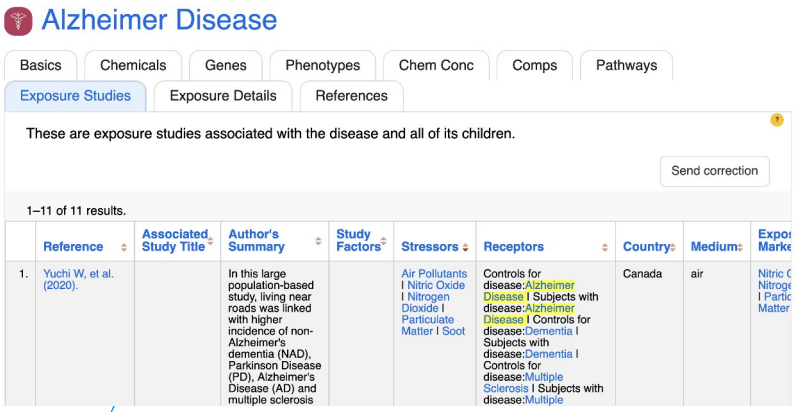
==> End-user services





ctd

https://ctdbase.org/



Conclusion

There are opportunities to exploit AI techniques/methods to tackle issues related to Public Health and Impact of Climate Change

These approaches come with some limitations

- Bias/Data representativity
- Impact on environment of learning algo

Taking benefits from advancements from other areas

- Knowledge representation & Learning over graphs, Multihop Reasoning, etc.

Big Thank You to my teams and colleagues





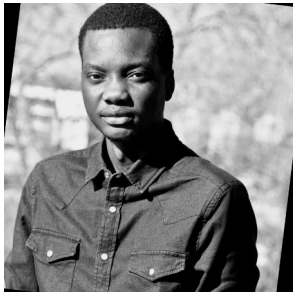


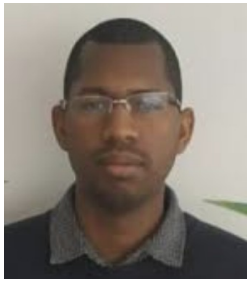












Graphes de Connaissances : Cas d'Usages en Biologie et Santé

Gayo DIALLO

Team AHeaD, BPH INSERM U1219
ISPED - Université de Bordeaux
Gayo.Diallo@u-Bordeaux.fr



Instituts thématiques

Inserm
Institut national de la santé et de la recherche médicale

BORDEAUX POPULATION HEALTH Research Center - U1219

AHeaD

université de BORDEAUX