TOWARDS TRUSTWORTHY DATA

First International Workshop on Knowledge Science

March 31, 2022

Eindhoven Artificial Intelligence Systems Institute

Knowledge Science for AI-based biomedical and clinical applications

Catia Pesquita

LASIGE, University of Lisbon

@CPesquita clpesquita@fc.ul.pt

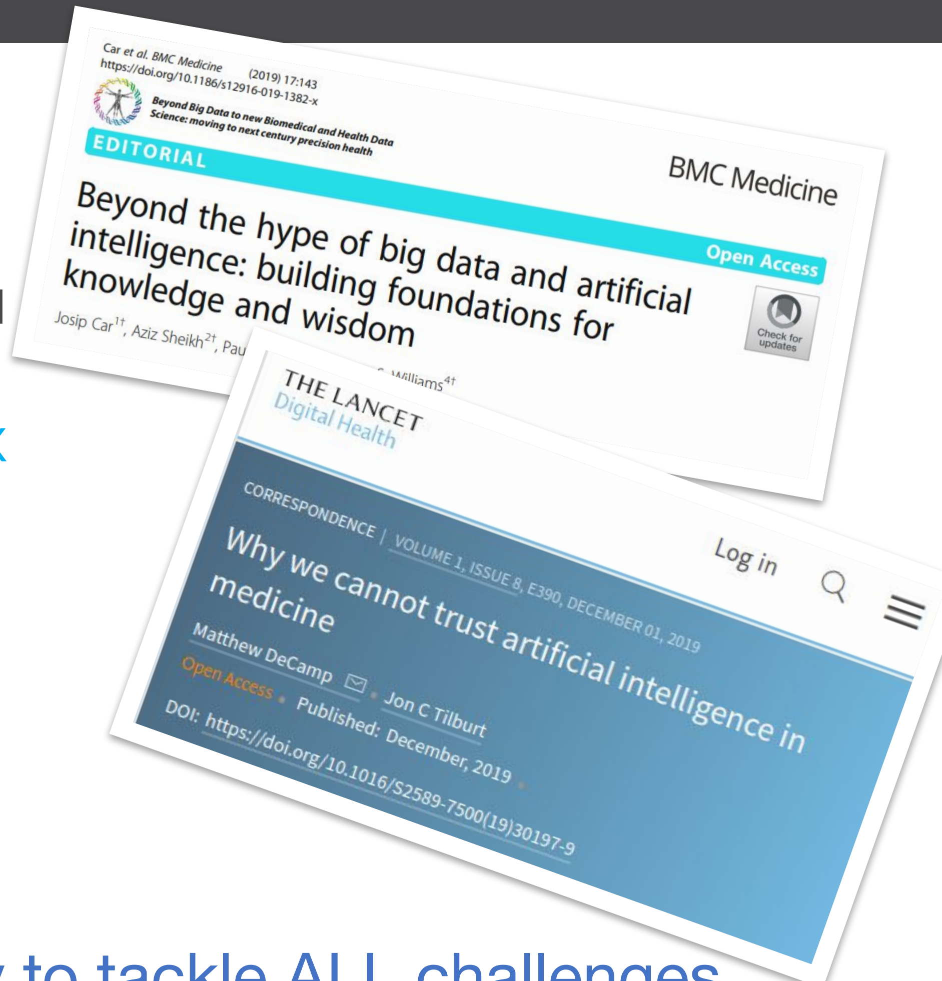
LASIGE driven by excellence

Ciências ULisboa

FCT Fundação para a Ciência e a Tecnologia

Why biomedical AI needs Knowledge Science

1. Large amounts of data in non-standard formats which need to be converted, interpreted, and merged into readable formats.
2. Heterogeneous and complex data which current ML/AI approaches are processing without context
3. Lack of explainability hinders trust


Car et al. BMC Medicine (2019) 17:143
https://doi.org/10.1186/s12916-019-1382-x

Beyond Big Data to new Biomedical and Health Data Science: moving to next century precision health

EDITORIAL

BMC Medicine

Open Access

Beyond the hype of big data and artificial intelligence: building foundations for knowledge and wisdom

Check for updates

Josip Car[1†], Aziz Sheikh[2†], Pau ... Williams[4†]

THE LANCET Digital Health

Log in

CORRESPONDENCE | VOLUME 1, ISSUE 8, E390, DECEMBER 01, 2019

Why we cannot trust artificial intelligence in medicine

Matthew DeCamp • Jon C Tilburt

Open Access • Published: December, 2019 •

DOI: https://doi.org/10.1016/S2589-7500(19)30197-9

Knowledge Science is key to tackle ALL challenges

Trust in AI

- the user successfully comprehends how the model arrives at an outcome

- the model’s outcomes/workings match the user’s prior knowledge

Jacovi, Alon, et al. "Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI." *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021.

Trust in AI

- the user successfully comprehends how the model arrives at an outcome represent a model’s processes

- the model’s outcomes/workings match the user’s prior knowledge evaluation

What happens when data is complex, multi-domain, heterogeneous, incomplete, ambiguous?

Trust in AI for biomedical and clinical applications

- the user successfully comprehends how the model arrives at an outcome represent a model’s processes
- the model’s outcomes/workings match the user’s prior knowledge evaluation

Trust in AI for biomedical and clinical applications

- the user successfully comprehends how the model arrives at an outcome □ represent a model's inputs, outputs and processes

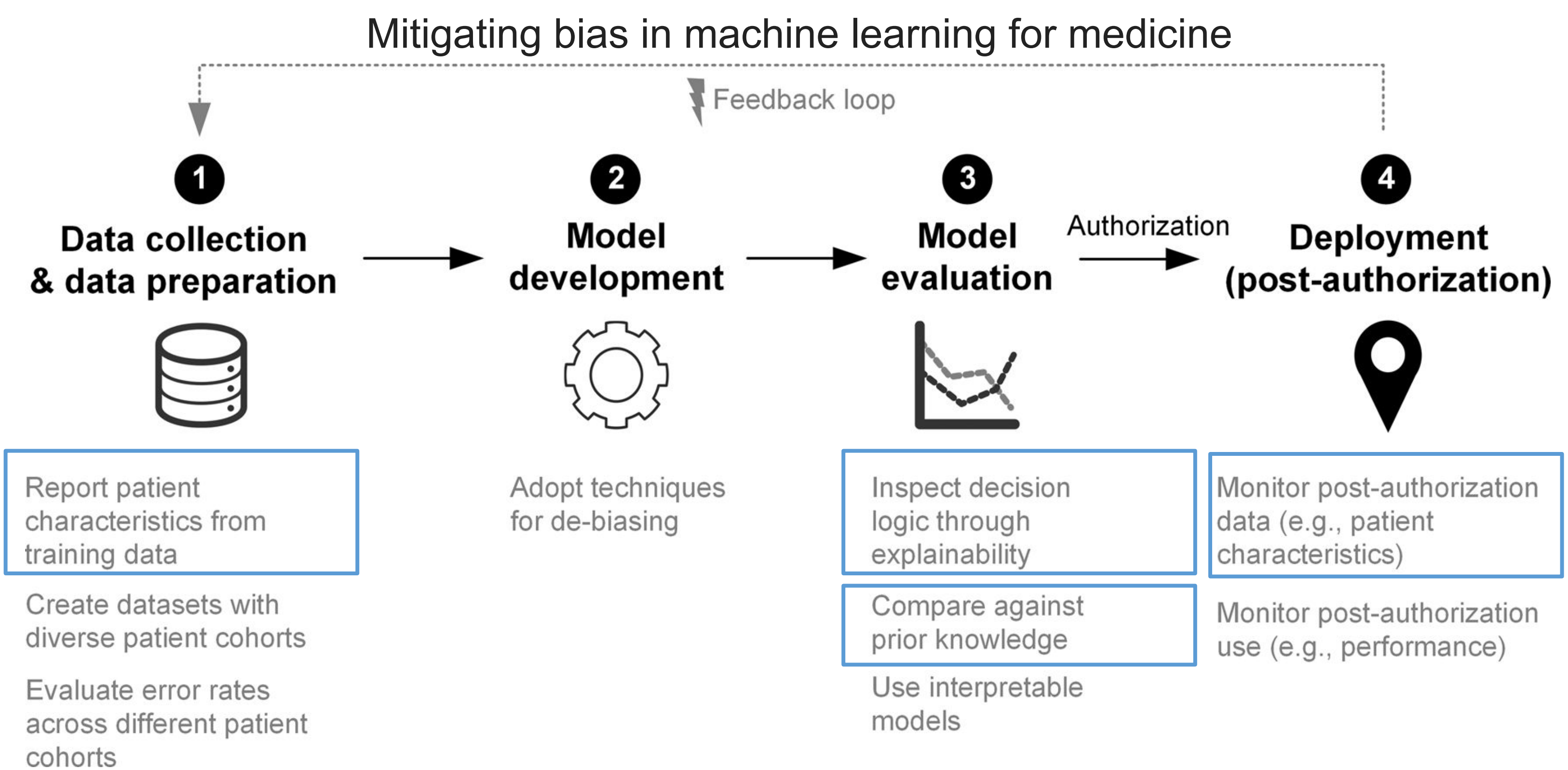
- the model's outcomes/workings match the user's prior knowledge □ evaluation

Trust in AI for biomedical and clinical applications

- the user successfully comprehends how the model arrives at an outcome represent a model’s inputs, outputs and processes

- the model’s outcomes/workings match the user’s prior knowledge represent the scientific context

Knowledge Science can help mitigate bias in biomedical AI



10.1038/s43856-021-00028-w

Knowledge Science for Trust in Biomedical AI

Assessing trustworthiness requires data, domain and user context

Data context: represent data provenance and transformations/processing

Domain/background knowledge: represent the scientific context of the data and application

User context: different users will trust based on different expectations

Data Context

Augment explanations with the data creation and processing context

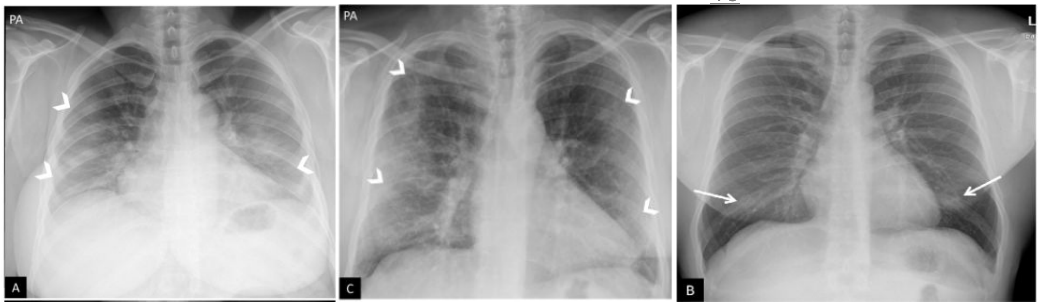
- Provide a rich contextual semantic layer to the underlying data using domain ontologies and knowledge graphs.
- Preserve uncertainty and highlight potential ambiguity and incompleteness at the data level

Data Context is key for trust

25% of the works that developed ML approaches to diagnose COVID-19 in adults based on chest X-rays and CT scans used pediatric (ages 1-5) pneumonia images as control.

10.1016/j.cell.2018.02.010

COVID-19



10.1016/j.rxeng.2020.11.001

Roberts, Michael, et al. "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans." *Nature Machine Intelligence* 3.3 (2021): 199-217.

Domain Knowledge Context

Contextualize an explanation within existing knowledge

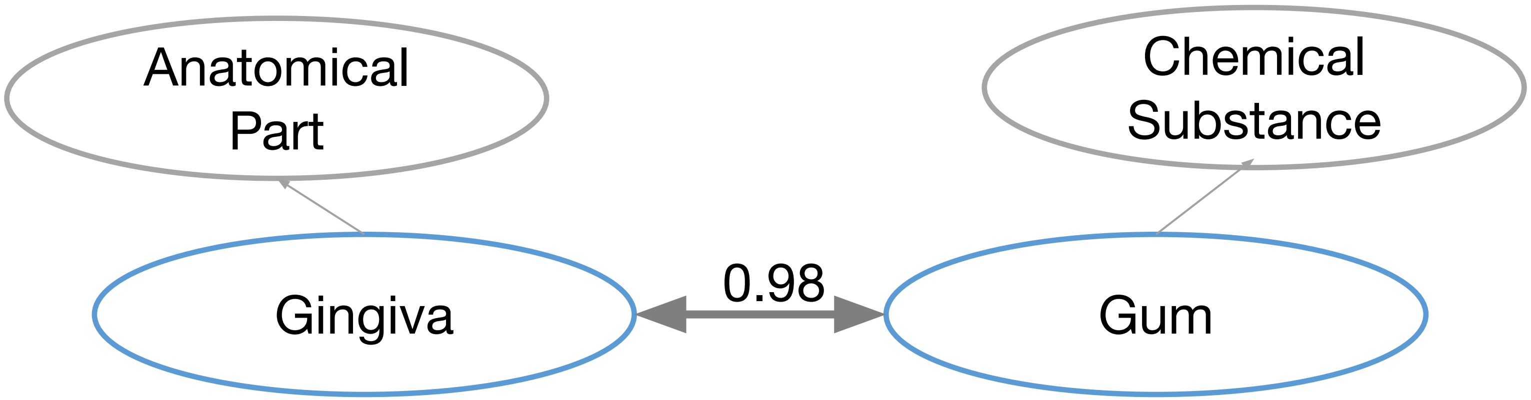
- Include prior knowledge through links to ontologies
- Enrich the contextual semantic layer with links and relations across domains of knowledge

Domain Knowledge Context is key for trust

Term 1	Term 2	Similarity
Gingiva	Gum	0.98

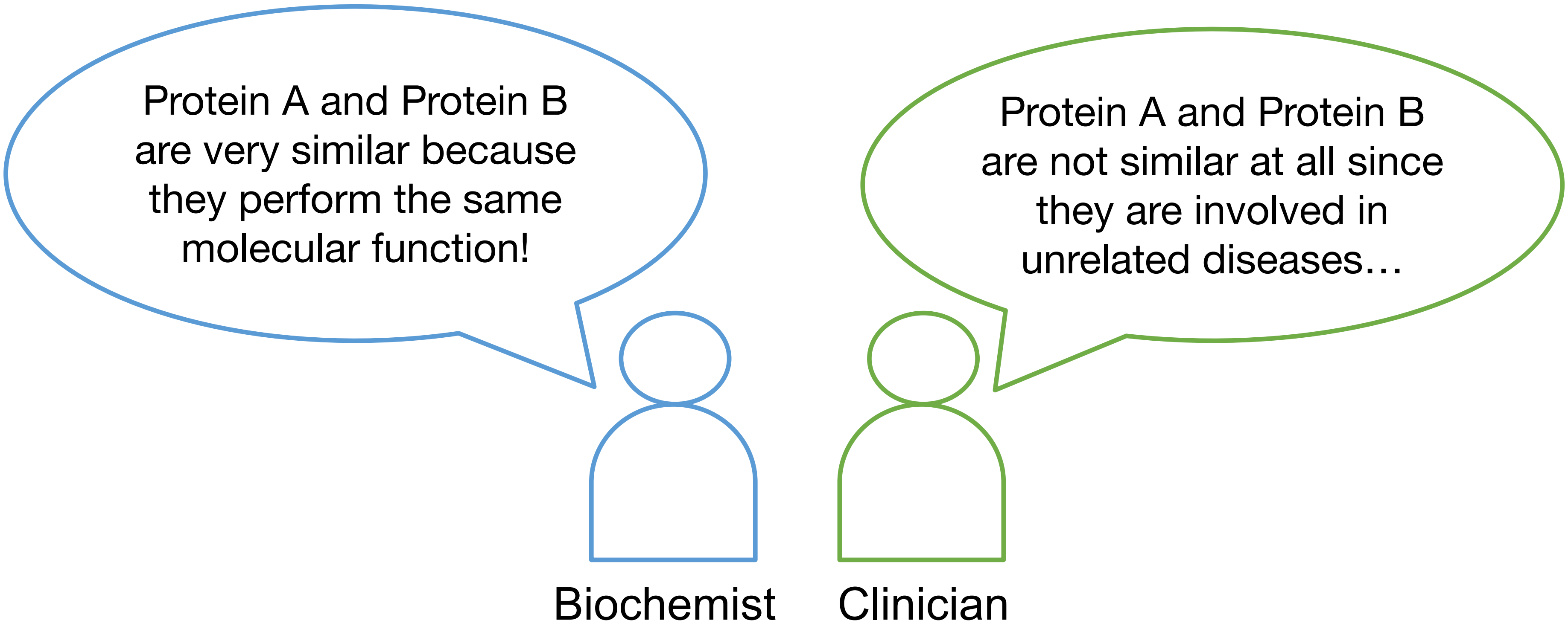
Domain Knowledge Context is key for trust

Term 1	Term 2	Similarity
Gingiva	Gum	0.98



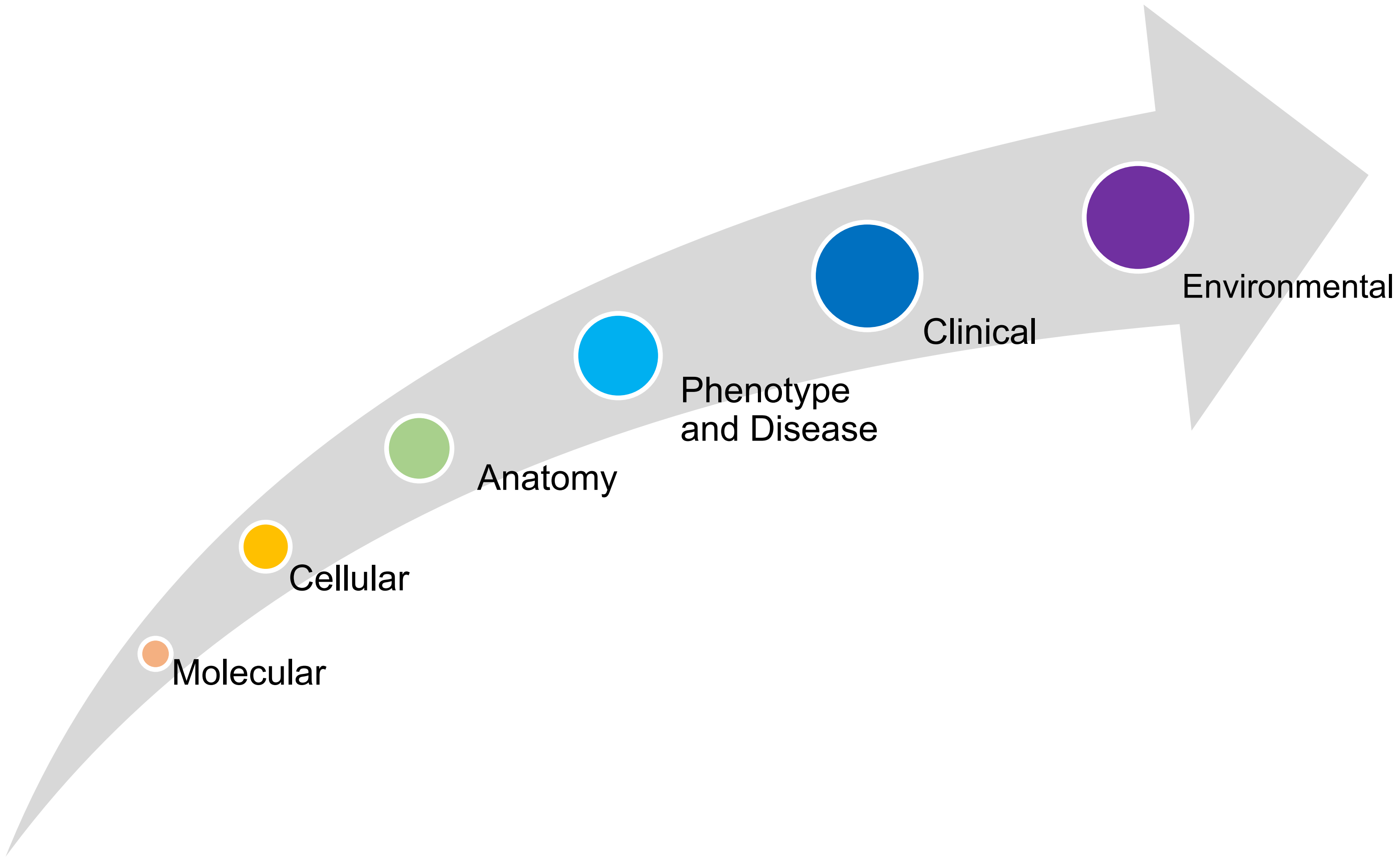
User Context

Trusting an AI outcome depends on the user context: task, prior knowledge, expectation, etc.



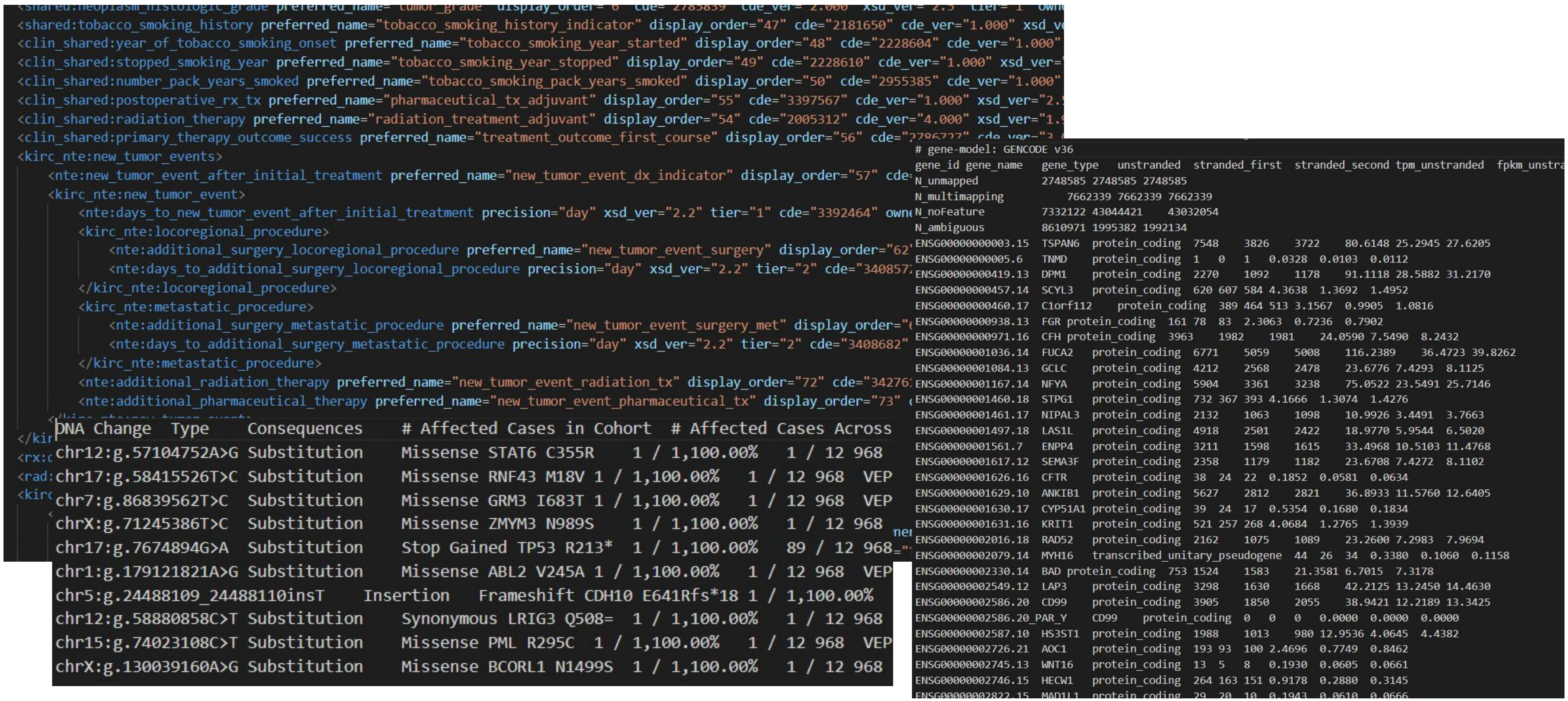
Ontologies, Knowledge Graphs and FAIR data principles can support data, domain and user context.

NextGen Biomedical AI requires integration of complex and diverse data



NextGen Biomedical AI requires integration of complex and diverse data

- 100s of very large files per patient covering genome sequence, mutations, transcriptome, clinical, etc.




```
<shared:tobacco_smoking_history preferred_name="tobacco_smoking_history_indicator" display_order="47" cde="2181650" cde_ver="1.000" xsd_ve
<clin_shared:year_of_tobacco_smoking_onset preferred_name="tobacco_smoking_year_started" display_order="48" cde="2228604" cde_ver="1.000"
<clin_shared:stopped_smoking_year preferred_name="tobacco_smoking_year_stopped" display_order="49" cde="2228610" cde_ver="1.000" xsd_ver=
<clin_shared:number_pack_years_smoked preferred_name="tobacco_smoking_pack_years_smoked" display_order="50" cde="2955385" cde_ver="1.000"
<clin_shared:postoperative_rx_tx preferred_name="pharmaceutical_tx_adjuvant" display_order="55" cde="3397567" cde_ver="1.000" xsd_ver="2.
<clin_shared:radiation_therapy preferred_name="radiation_treatment_adjuvant" display_order="54" cde="2005312" cde_ver="4.000" xsd_ver="1.
<clin_shared:primary_therapy_outcome_success preferred_name="treatment_outcome_first_course" display_order="56" cde="
<kirc_nte:new_tumor_events>
    <nte:new_tumor_event_after_initial_treatment preferred_name="new_tumor_event_dx_indicator" display_order="57" cde=
    <kirc_nte:new_tumor_event>
        <nte:days_to_new_tumor_event_after_initial_treatment precision="day" xsd_ver="2.2" tier="1" cde="3392464" own
        <kirc_nte:locoregional_procedure>
            <nte:additional_surgery_locoregional_procedure preferred_name="new_tumor_event_surgery" display_order="62
            <nte:days_to_additional_surgery_locoregional_procedure precision="day" xsd_ver="2.2" tier="2" cde="340857
        </kirc_nte:locoregional_procedure>
        <kirc_nte:metastatic_procedure>
            <nte:additional_surgery_metastatic_procedure preferred_name="new_tumor_event_surgery_met" display_order="
            <nte:days_to_additional_surgery_metastatic_procedure precision="day" xsd_ver="2.2" tier="2" cde="3408682"
        </kirc_nte:metastatic_procedure>
        <nte:additional_radiation_therapy preferred_name="new_tumor_event_radiation_tx" display_order="72" cde="34276
        <nte:additional_pharmaceutical_therapy preferred_name="new_tumor_event_pharmaceutical_tx" display_order="73"
```


```
DNA Change   Type          Consequences                 # Affected Cases in Cohort   # Affected Cases Across
chr12:g.57104752A>G Substitution   Missense STAT6 C355R    1 / 1,100.00%    1 / 12 968
chr17:g.58415526T>C Substitution   Missense RNF43 M18V 1 / 1,100.00%    1 / 12 968   VEP
chr7:g.86839562T>C  Substitution   Missense GRM3 I683T 1 / 1,100.00%    1 / 12 968   VEP
chrX:g.71245386T>C  Substitution   Missense ZMYM3 N989S    1 / 1,100.00%    1 / 12 968
chr17:g.7674894G>A  Substitution   Stop Gained TP53 R213*  1 / 1,100.00%    89 / 12 968
chr1:g.179121821A>G Substitution   Missense ABL2 V245A 1 / 1,100.00%    1 / 12 968   VEP
chr5:g.24488109_24488110insT    Insertion   Frameshift CDH10 E641Rfs*18 1 / 1,100.00%
chr12:g.58880858C>T Substitution   Synonymous LRIG3 Q508=  1 / 1,100.00%    1 / 12 968
chr15:g.74023108C>T Substitution   Missense PML R295C  1 / 1,100.00%    1 / 12 968   VEP
chrX:g.130039160A>G Substitution   Missense BCORL1 N1499S  1 / 1,100.00%    1 / 12 968
```

```
# gene-model: GENCODE v36
gene_id gene_name   gene_type   unstranded  stranded_first  stranded_second tpm_unstranded  fpkm_unstra
N_unmapped      2748585 2748585 2748585
N_multimapping      7662339 7662339 7662339
N_noFeature     7332122 43044421    43032054
N_ambiguous     8610971 1995382 1992134
ENSG00000000003.15  TSPAN6  protein_coding  7548    3826    3722    80.6148 25.2945 27.6205
ENSG00000000005.6   TNMD    protein_coding  1   0   1   0.0328  0.0103  0.0112
ENSG00000000419.13  DPM1    protein_coding  2270    1092    1178    91.1118 28.5882 31.2170
ENSG00000000457.14  SCYL3   protein_coding  620 607 584 4.3638  1.3692  1.4952
ENSG00000000460.17  C1orf112    protein_coding  389 464 513 3.1567  0.9905  1.0816
ENSG00000000938.13  FGR protein_coding  161 78  83  2.3063  0.7236  0.7902
ENSG00000000971.16  CFH protein_coding  3963    1982    1981    24.0590 7.5490  8.2432
ENSG00000001036.14  FUCA2   protein_coding  6771    5059    5008    116.2389    36.4723 39.8262
ENSG00000001084.13  GCLC    protein_coding  4212    2568    2478    23.6776 7.4293  8.1125
ENSG00000001167.14  NFYA    protein_coding  5904    3361    3238    75.0522 23.5491 25.7146
ENSG00000001460.18  STPG1   protein_coding  732 367 393 4.1666  1.3074  1.4276
ENSG00000001461.17  NIPAL3  protein_coding  2132    1063    1098    10.9926 3.4491  3.7663
ENSG00000001497.18  LAS1L   protein_coding  4918    2501    2422    18.9770 5.9544  6.5020
ENSG00000001561.7   ENPP4   protein_coding  3211    1598    1615    33.4968 10.5103 11.4768
ENSG00000001617.12  SEMA3F  protein_coding  2358    1179    1182    23.6708 7.4272  8.1102
ENSG00000001626.16  CFTR    protein_coding  38  24  22  0.1852  0.0581  0.0634
ENSG00000001629.10  ANKIB1  protein_coding  5627    2812    2821    36.8933 11.5760 12.6405
ENSG00000001630.17  CYP51A1 protein_coding  39  24  17  0.5354  0.1680  0.1834
ENSG00000001631.16  KRIT1   protein_coding  521 257 268 4.0684  1.2765  1.3939
ENSG00000002016.18  RAD52   protein_coding  2162    1075    1089    23.2600 7.2983  7.9694
ENSG00000002079.14  MYH16   transcribed_unitary_pseudogene  44  26  34  0.3380  0.1060  0.1158
ENSG00000002330.14  BAD protein_coding  753 1524    1583    21.3581 6.7015  7.3178
ENSG00000002549.12  LAP3    protein_coding  3298    1630    1668    42.2125 13.2450 14.4630
ENSG00000002586.20  CD99    protein_coding  3905    1850    2055    38.9421 12.2189 13.3425
ENSG00000002586.20_PAR_Y    CD99    protein_coding  0   0   0   0.0000  0.0000  0.0000
ENSG00000002587.10  HS3ST1  protein_coding  1988    1013    980 12.9536 4.0645  4.4382
ENSG00000002726.21  AOC1    protein_coding  193 93  100 2.4696  0.7749  0.8462
ENSG00000002745.13  WNT16   protein_coding  13  5   8   0.1930  0.0605  0.0661
ENSG00000002746.15  HECW1   protein_coding  264 163 151 0.9178  0.2880  0.3145
```

What happens when data is complex, multi-domain, heterogeneous, incomplete, ambiguous?

One ontology is not enough

Systems Biology, Systems Medicine and Personalized Medicine require holistic representations

Level	Ontology	Scope
Molecular	ChEBI	Small biomolecules
	ATC	Active ingredients of drugs
	GO	Protein function
Cellular	CL	Cell types
	CLO	Cell lines
Anatomy	UBERON	General anatomy
	FMA	Human anatomy
Phenotype and Disease	HPO	Human phenotypes
	DO	Human diseases
	ICD	Human diseases
Clinical	UMLS	Biomedical and clinical aspects
	SNOMED-CT	Clinical aspects
	LOINC	Laboratory findings
Environmental	ENVO	Environmental factors
Large-scope	KEGG	Biological systems
Research	SBO	Systems biology
	OBI	Biomedical investigation

JD Ferreira, DC Teixeira, C Pesquita. Biomedical Ontologies: Coverage, Access and Use. 2020 Systems Medicine Integrative, Qualitative and Computational Approaches, Academic Press, Elsevier

What do we need to create holistic representations?

Cover multiple domains	Ensure rich semantic integration	Provide high quality alignments
Align multiple ontologies Scalability	Related but not equal domains Complex relations involving more than one ontology	Support human interaction Visualizing the context of a mapping Balancing cognitive overload and informativeness

MC Silva, D Faria, and C Pesquita. Integrating knowledge graphs for explainable artificial intelligence in biomedicine. Ontology Matching workshop 2021

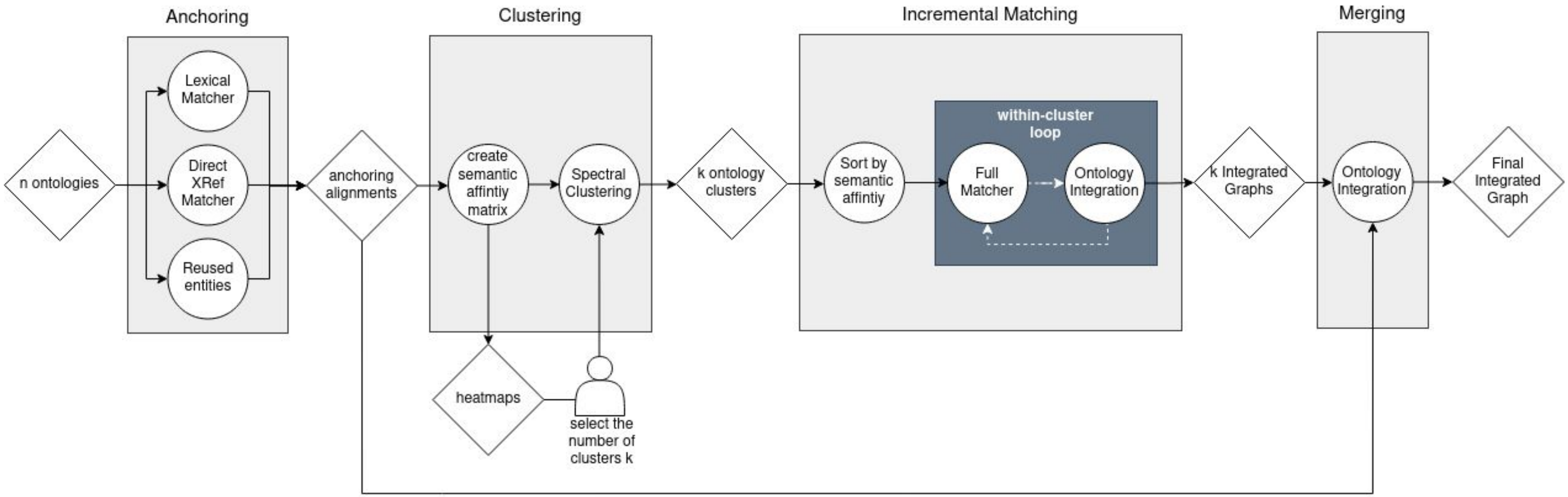
Rethinking biomedical ontology alignment

Cover multiple domains	Ensure rich semantic integration	Provide high quality alignments
Pairwise ontology alignment	Simple equivalence ontology alignment	User validation
↓	↓	↓
Holistic ontology alignment of multiple ontologies	**Complex ontology alignment**	**Human-in-the-loop Interactive Alignment**

MC Silva, D Faria, and C Pesquita. Integrating knowledge graphs for explainable artificial intelligence in biomedicine. Ontology Matching workshop 2021

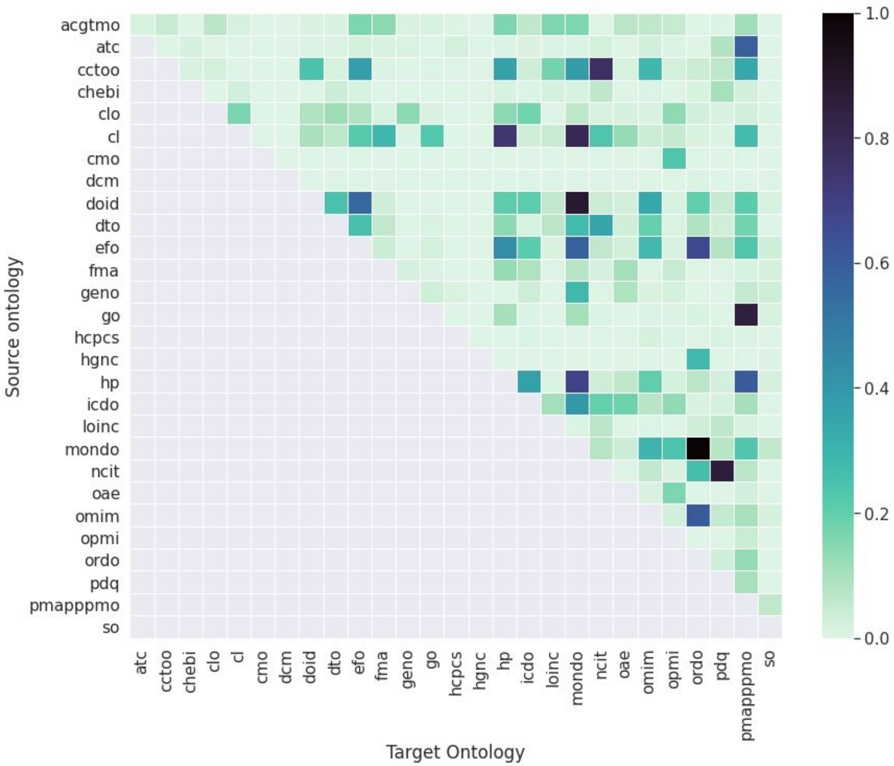
Holistic Matching with clustering and incremental matching

Can we ensure the scalability of matching tasks with many ontologies?



Silva, M.C., Faria, D. & Pesquita, C. (2022). Matching Multiple Ontologies to Build a Knowledge Graph for Personalized Medicine. *ESWC2022 (accepted)*

Holistic Matching (CIA) runs in half the time, ensures high recall within similar domains and high precision across domains



Strategy	Total Mappings
GPA	554547
CPA+anchoring	442649
CIA+anchoring	417131

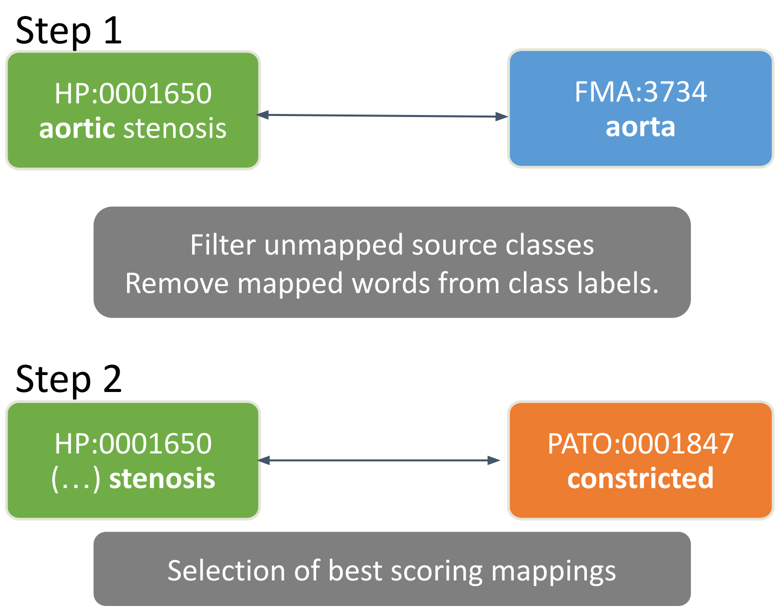
	Runtime (hh:mm)			Alignment	
Strategy	Load	Match	Total	Mappings	Tasks
GPA	11:47	19:51	31:37	554547	378
Anchoring	11:47	01:59	13:46	427300	378
CPA	02:25	07:42	10:07	219021	117
CIA	01:05	01:05	02:10	193503	24

GPA: global pairwise alignment. CPA: within-cluster pairwise alignment. CIA: within cluster incremental alignment.

Silva, M.C., Faria, D. & Pesquita, C. (2022). Matching Multiple Ontologies to Build a Knowledge Graph for Personalized Medicine. *ESWC2022 (accepted)*

Compound Matching for ontology triples

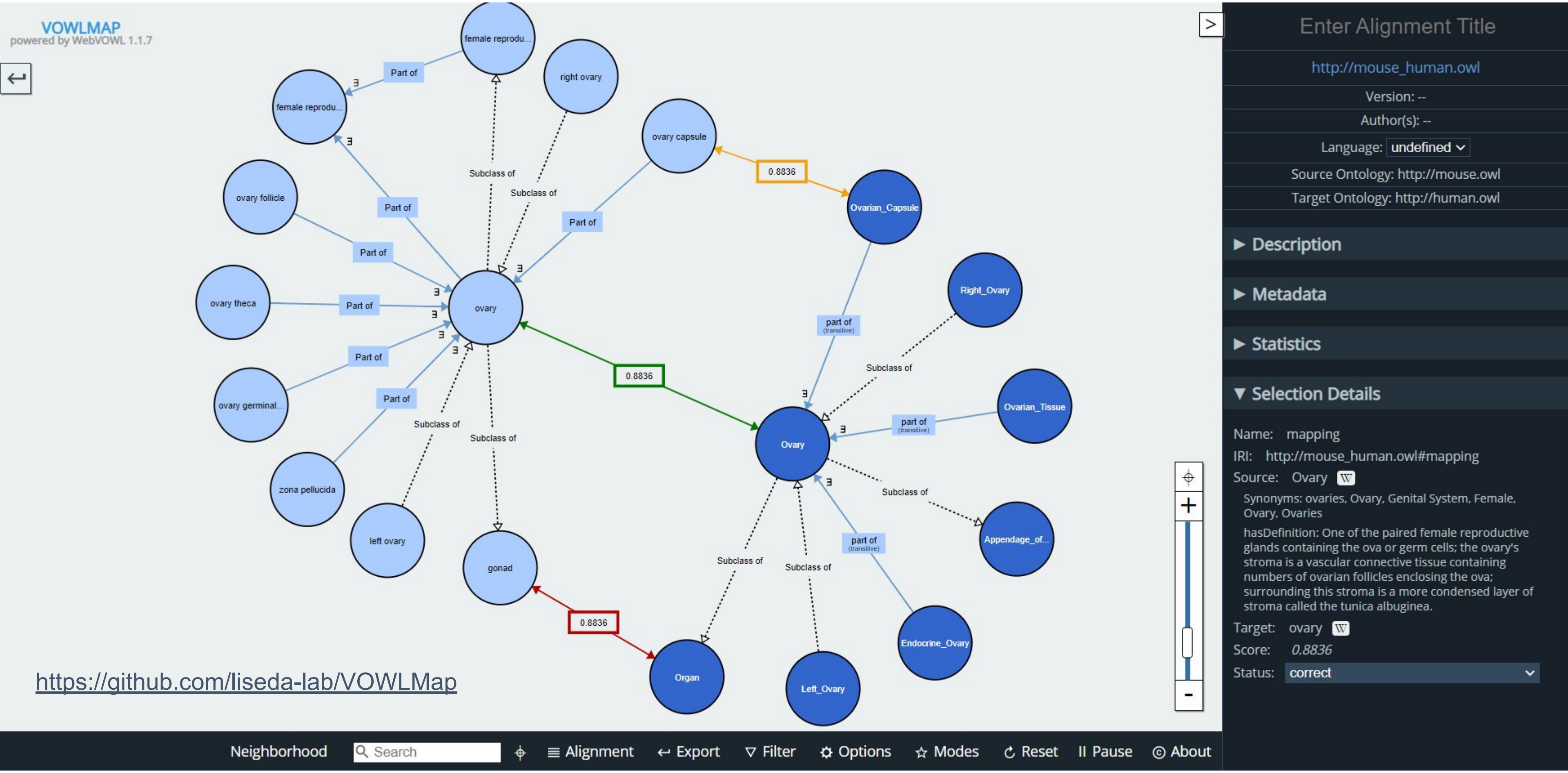
Can we find mappings involving multiple ontologies using lexical approaches and search space pruning?



Ontology sets	Correct (New)
MP-CL-PATO	47.1% (17.6%)
MP-GO-PATO	86.9% (22.3%)
MP-NBO-PATO	70.4% (20.7%)
MP-UBERON-PATO	83.5% (24.7%)
WBP-GO-PATO	44.9% (33.3%)
HP-FMA-PATO	81.5% (44.1%)

Oliveira, D., & Pesquita, C. (2018). Improving the interoperability of biomedical ontologies with compound alignments. *J. Biomedical Semantics*.

Supporting contextualized ontology alignment validation

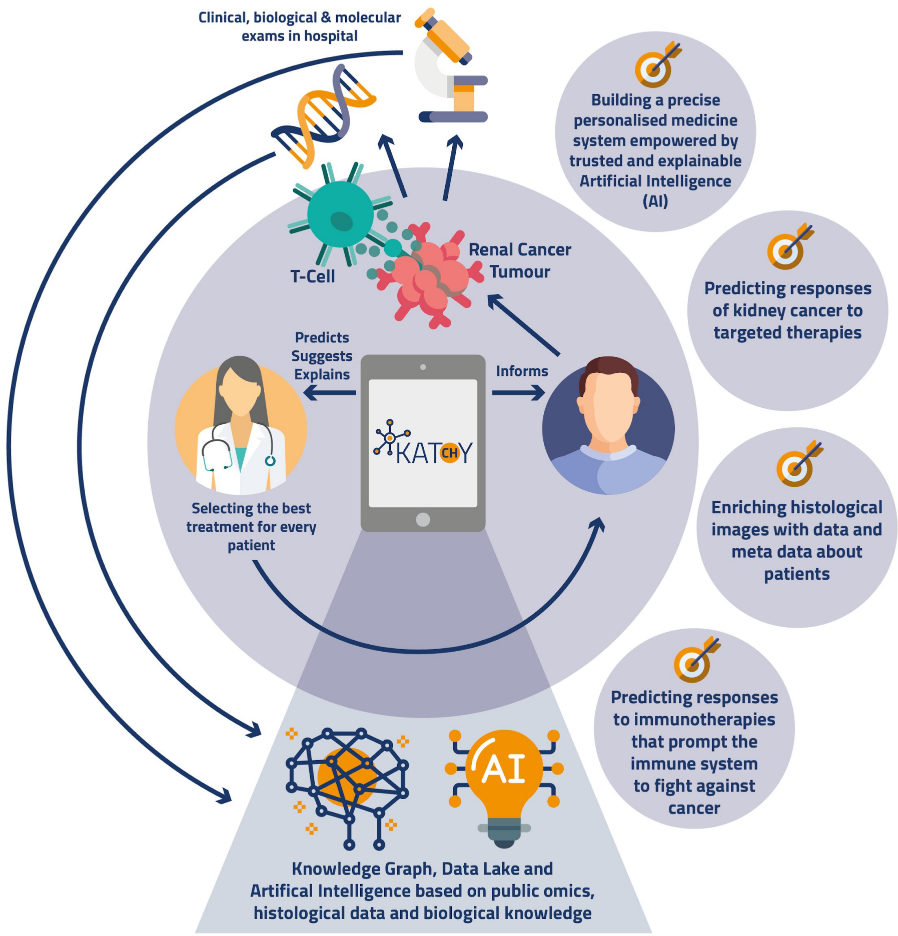


https://github.com/liseda-lab/VOWLMap

Guerreiro, A., Faria, D. & Pesquita, C. (2021) VOWLMap: Graph-based Ontology Alignment Visualization and Editing. VOILA workshop (ISWC 2021)

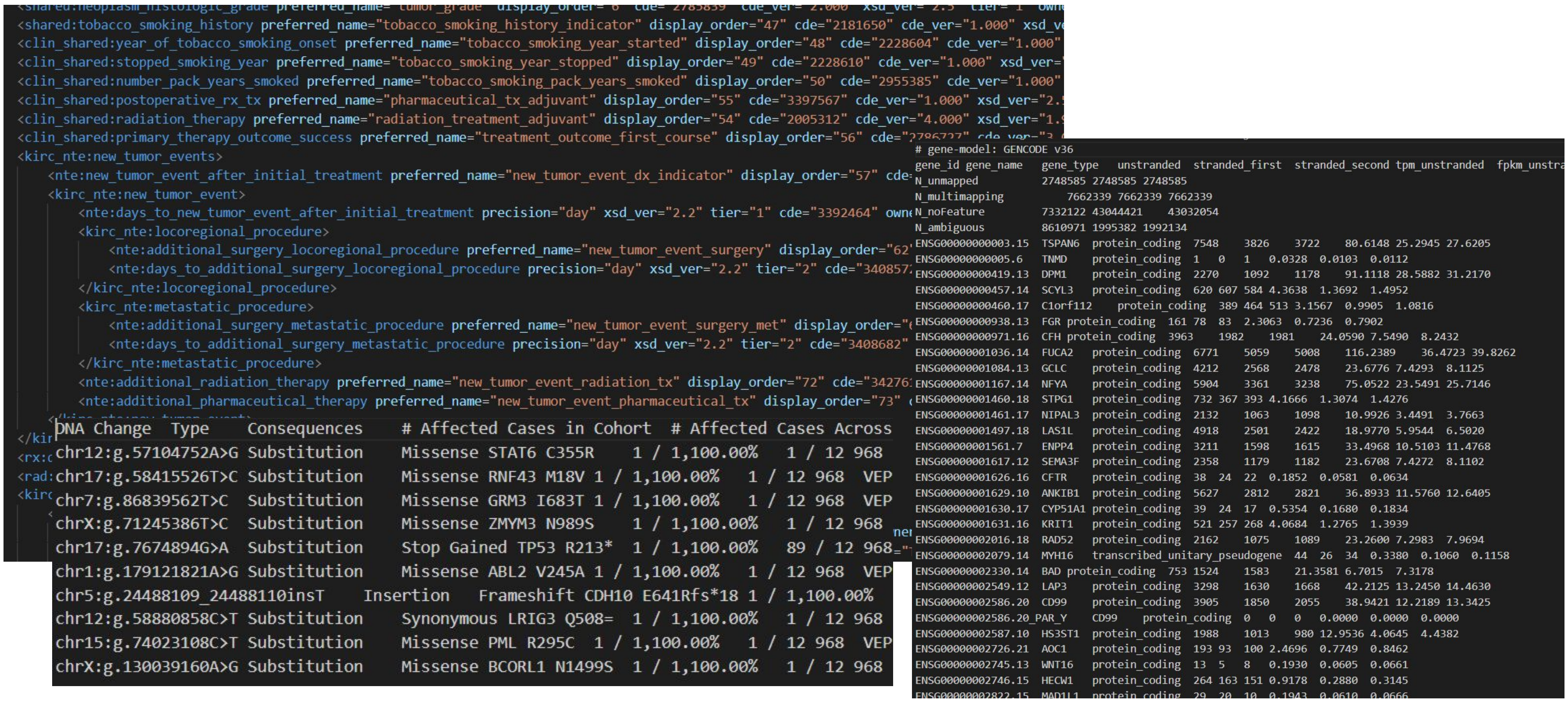
Trustworthy biomedical AI requires trustworthy data.

Explainable AI for Personalized Oncology

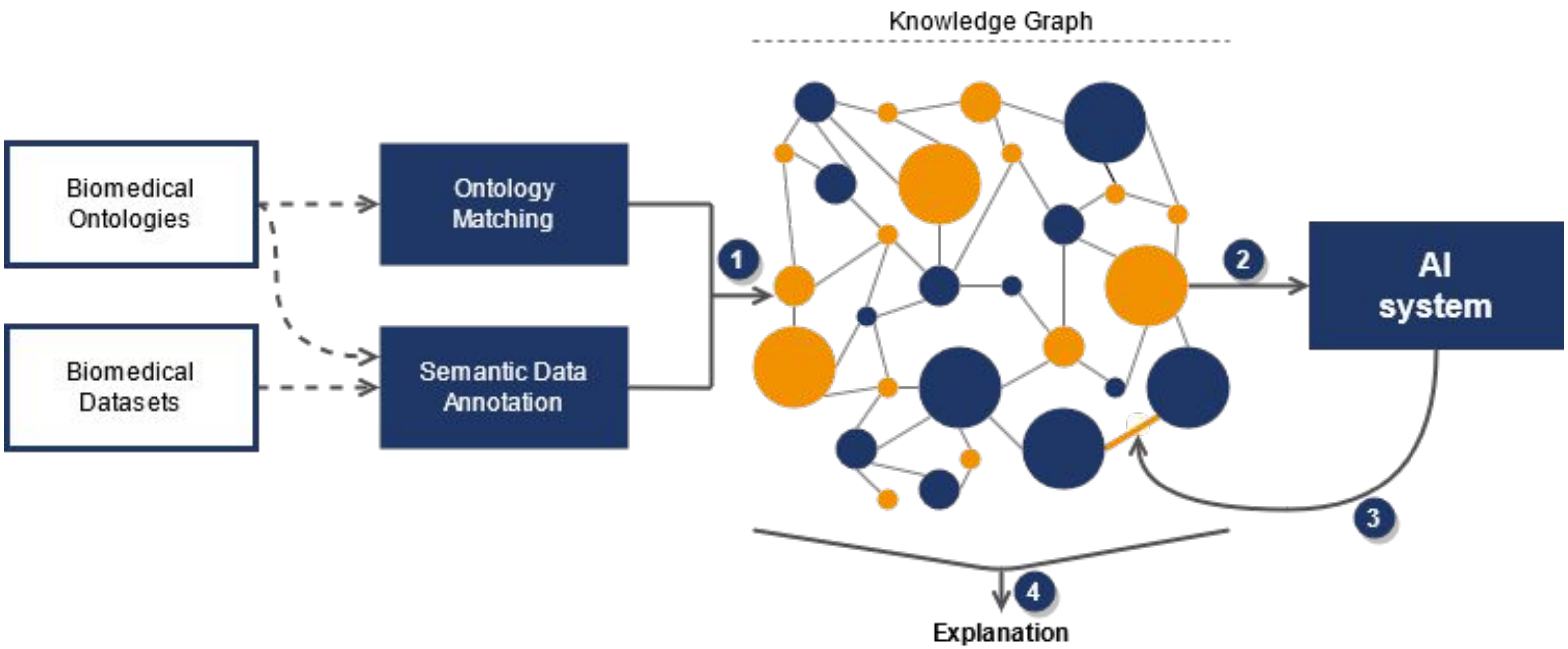


katy-project.eu

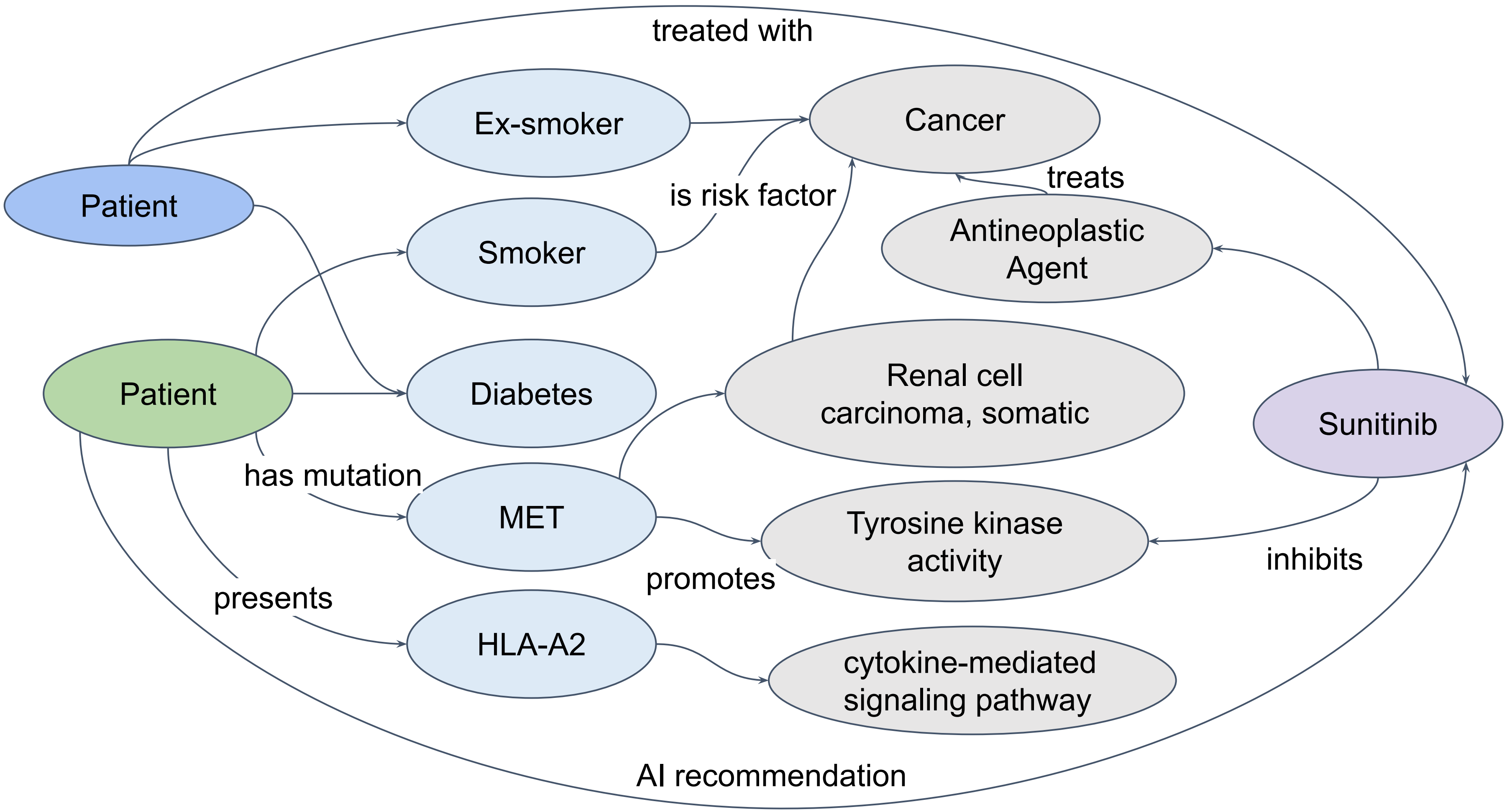
Is this Data Trustworthy?



Building a KG for Explainable AI for Personalized Oncology

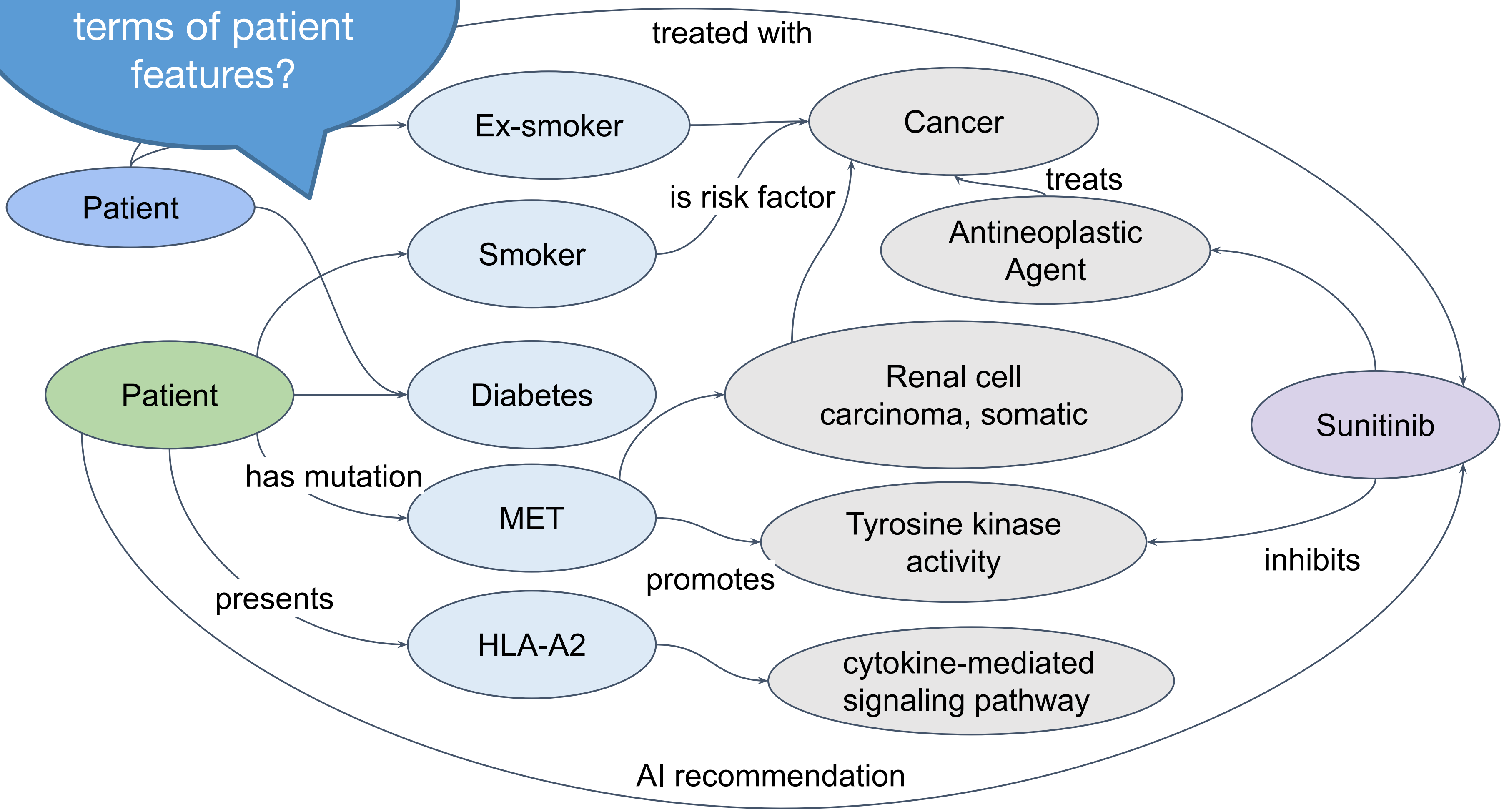


The KG can be used to contextualize the patient and the drug recommendation

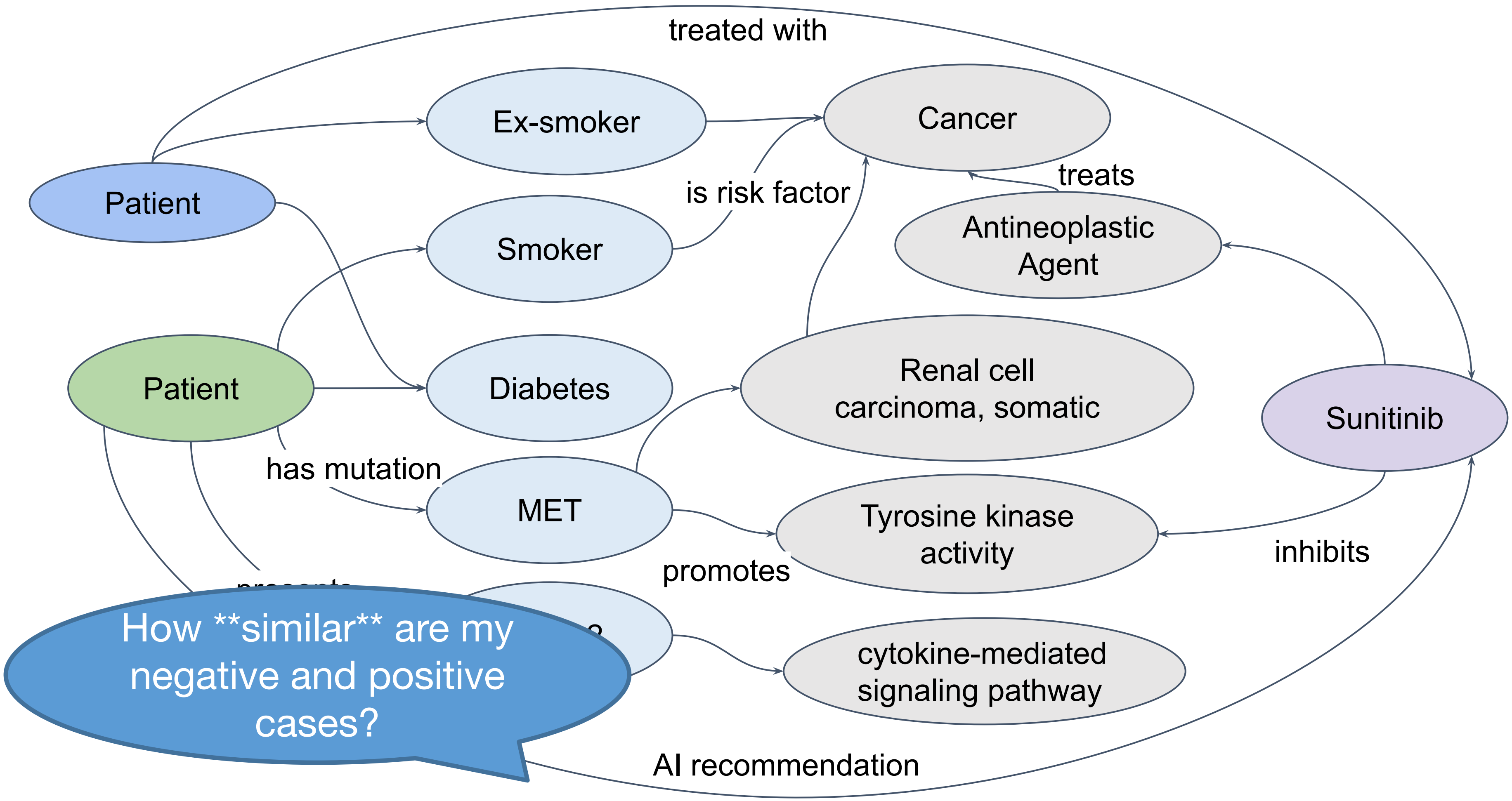


The KG can be used to contextualize the ... the drug recommendation

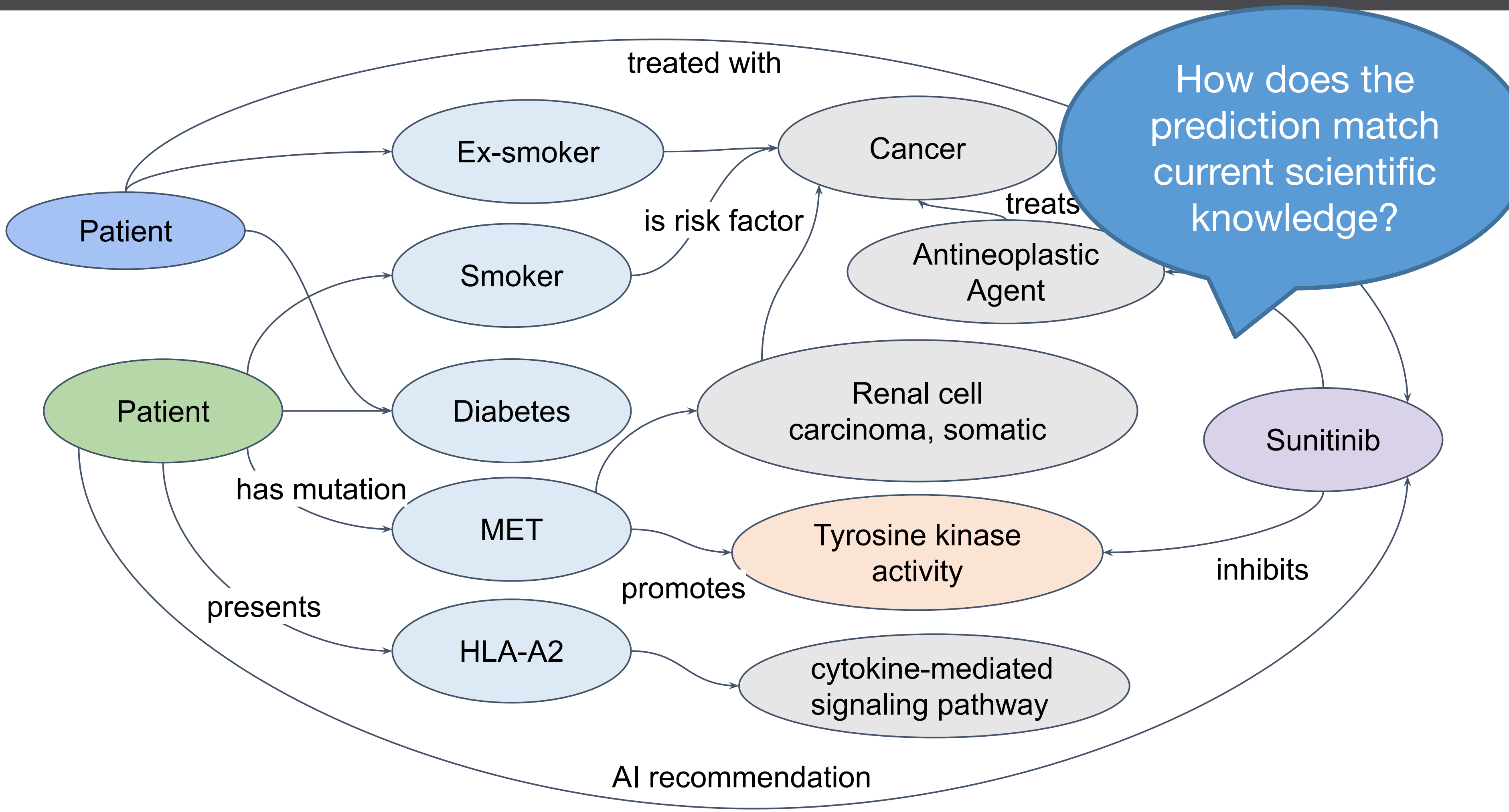
How diverse is my dataset in terms of patient features?



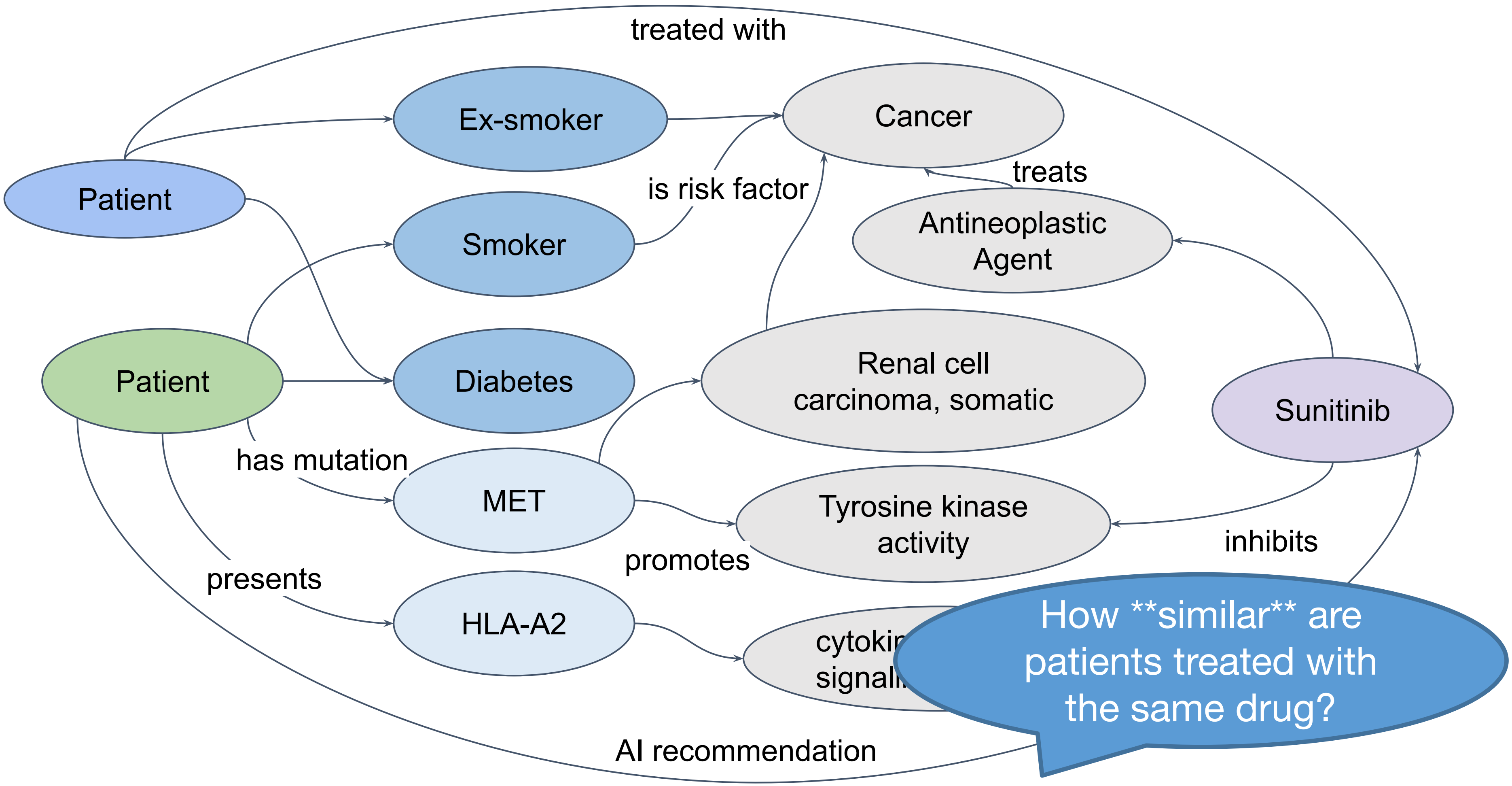
The KG can be used to contextualize the patient and the drug recommendation



The KG can be used to contextualize the patient and the drug recommendation



The KG can be used to contextualize the patient and the drug recommendation



The great barrier to AI adoption in healthcare and biomedical research is lack of trust.

Knowledge Science is the answer.

Photo by [National Cancer Institute](#) on [Unsplash](#)

Acknowledgements

Daniel Faria, LASIGE/Biodata.pt, Portugal

Isabel Cruz, U. Illinois, USA

Ernesto Jimenez-Ruiz, City, U. of London, UK

Daniela Oliveira, Novartis

Patrick Lambrix, U. Linkoping, Sweden

Valentina Ivanova, RISE, Sweden

Booma S. Balasubramani, Microsoft

and many others

Past and present students:

Rita Sousa

Marta Silva

Susana Nunes

Ana Guerreiro

Patrícia Eugénio

Filipa Serrano

Beatriz Lima

Carlota Cardoso

and many others

https://liseda-lab.github.io/
https://github.com/liseda-lab/
https://github.com/AgreementMakerLight

@CPesquita
clpesquita@fc.ul.pt

KATCHY
Clinical Knowledge for Humanity

This work was supported by FCT through the LASIGEResearch Unit (UIDB/00408/2020 and UIDP/00408/2020). It was also partially supported by the KATY project which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017453.