Knowledge Science for AI-based biomedical and clinical applications

Catia Pesquita
LASIGE, University of Lisbon
@CPesquita clpesquita@fc.ul.pt
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.

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

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

Mitigating bias in machine learning for medicine

1. Data collection & data preparation
   - Report patient characteristics from training data
   - Create datasets with diverse patient cohorts
   - Evaluate error rates across different patient cohorts

2. Model development
   - Adopt techniques for de-biasing

3. Model evaluation
   - Inspect decision logic through explainability
   - Compare against prior knowledge
   - Use interpretable models

4. Deployment (post-authorization)
   - Monitor post-authorization data (e.g., patient characteristics)
   - Monitor post-authorization use (e.g., performance)

Feedback loop

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.

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

<table>
<thead>
<tr>
<th>Term 1</th>
<th>Term 2</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gingiva</td>
<td>Gum</td>
<td>0.98</td>
</tr>
</tbody>
</table>
Domain Knowledge Context is key for trust

<table>
<thead>
<tr>
<th>Term 1</th>
<th>Term 2</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gingiva</td>
<td>Gum</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Anatomical Part

Chemical Substance

Gingiva

Gum

Similarity: 0.98
User Context

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

Protein A and Protein B are very similar because they perform the same molecular function!

Protein A and Protein B are not similar at all since they are involved in unrelated diseases...

Biochemist

Clinician
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.
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

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

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
- Align multiple ontologies
- Scalability

Ensure rich semantic integration
- Related but not equal domains
- Complex relations involving more than one ontology

Provide high quality alignments
- 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
  - Pairwise ontology alignment
  - Holistic ontology alignment of multiple ontologies

- Ensure rich semantic integration
  - Simple equivalence ontology alignment
  - Complex ontology alignment

- Provide high quality alignments
  - User validation
  - 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.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Total Mappings</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPA</td>
<td>554547</td>
</tr>
<tr>
<td>CPA+anchoring</td>
<td>442649</td>
</tr>
<tr>
<td>CIA+anchoring</td>
<td>417131</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Load</th>
<th>Match</th>
<th>Total</th>
<th>Mappings</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPA</td>
<td>11:47</td>
<td>19:51</td>
<td>31:37</td>
<td>554547</td>
<td>378</td>
</tr>
<tr>
<td>Anchoring</td>
<td>11:47</td>
<td>01:59</td>
<td>13:46</td>
<td>427300</td>
<td>378</td>
</tr>
<tr>
<td>CPA</td>
<td>02:25</td>
<td>07:42</td>
<td>10:07</td>
<td>219021</td>
<td>117</td>
</tr>
<tr>
<td>CIA</td>
<td>01:05</td>
<td>01:05</td>
<td>02:10</td>
<td>193503</td>
<td>24</td>
</tr>
</tbody>
</table>

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?

Step 1

HP:0001650 aortic stenosis → FMA:3734 aorta

Filter unmapped source classes
Remove mapped words from class labels.

Step 2

HP:0001650 (...) stenosis → PATO:0001847 constricted

Selection of best scoring mappings

<table>
<thead>
<tr>
<th>Ontology sets</th>
<th>Correct (New)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP-CL-PATO</td>
<td>47.1% (17.6%)</td>
</tr>
<tr>
<td>MP-GO-PATO</td>
<td>86.9% (22.3%)</td>
</tr>
<tr>
<td>MP-NBO-PATO</td>
<td>70.4% (20.7%)</td>
</tr>
<tr>
<td>MP-UBERON-PATO</td>
<td>83.5% (24.7%)</td>
</tr>
<tr>
<td>WBP-GO-PATO</td>
<td>44.9% (33.3%)</td>
</tr>
<tr>
<td>HP-FMA-PATO</td>
<td>81.5% (44.1%)</td>
</tr>
</tbody>
</table>

Supporting contextualized ontology alignment validation

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

Clinical, biological & molecular exams in hospital

Building a precise personalised medicine system empowered by trusted and explainable Artificial Intelligence (AI)

T-Cell

Renal Cancer Tumour

Selecting the best treatment for every patient

Predicts Suggests Explains

Informs

Predicting responses of kidney cancer to targeted therapies

Enriching histological images with data and meta data about patients

Predicting responses to immunotherapies that prompt the immune system to fight against cancer

Knowledge Graph, Data Lake and Artificial Intelligence based on public omics, histological data and biological knowledge

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...
How diverse is my dataset in terms of patient features?

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

- Patient
- Ex-smoker
- Cancer
- Antineoplastic Agent
- Renal cell carcinoma, somatic
- Tyrosine kinase activity
- Sunitinib
- MET
- Diabetic

How **similar** are my negative and positive cases?
The KG can be used to contextualize the patient and the drug recommendation

How does the prediction match current scientific knowledge?
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.
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
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

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.