From Messy Data to Medical Insights: Creating Knowledge Graphs for Drug Repurposing

Matilde Pato
matilde.pato@isel.pt
Instituto Superior de Engenharia de Lisboa (ISEL), IPL
IBEB & LASIGE, Faculdade de Ciências da Universidade de Lisboa
NOVA LINCS, NOVA School of Science and Technology
Lisbon, Portugal

Carolina Pereira
carolinadpereira18@gmail.com
Instituto Superior de Engenharia de Lisboa
Lisbon, Portugal

Nuno Datia
nuno.datia@isel.pt
Instituto Superior de Engenharia de Lisboa (ISEL), IPL
NOVA LINCS, NOVA School of Science and Technology
Monte da Caparica, Portugal

Abstract

Drug repurposing – finding new therapeutic applications for existing drugs – has emerged as a cost-effective strategy to address unmet medical needs. However, the fragmented nature of biomedical data presents significant challenges to researchers working in this domain. This tutorial introduces participants to advanced data wrangling techniques [6] for creating knowledge graphs (KGs) that can reveal hidden relationships between drugs, diseases, and biological mechanisms. Through a combination of conceptual explanation and hands-on exercises, participants will learn how to transform heterogeneous pharmaceutical data into structured knowledge representations that support novel drug discovery. The tutorial begins by exploring why KGs are particularly valuable for drug repurposing, showing how their ability to represent complex relationships makes them better than traditional tabular data structures. We then guide participants through a streamlined data wrangling pipeline designed specifically for biomedical data integration. Participants will learn practical techniques for addressing common challenges in this domain, including entity resolution across multiple data sources, handling medical terminology inconsistencies. The main part of the tutorial is a demonstration of our "MedJsonify" methodology, a Python-based framework that simplifies the process of wrangling and integrating data from diverse pharmaceutical sources. Using prepared datasets derived from public resources like DailyMed [1], Purple Book [4] and Orange Book [3], as well as ontologies from DrugBank [5], DO [7], and ChEBI [2] and Orphanet [8], participants will practice key data transformation steps and visualization techniques within the Neo4j graph database environment. Rather than attempting to build a complete system within the time constraints, we focus on the most critical steps in the KG creation process, providing participants with reusable code templates and visualization strategies they can apply to their own research or educational contexts. We emphasize how these techniques not only support scientific discovery but can also transform healthcare education by making complex pharmaceutical relationships more accessible and interpretable. The tutorial concludes with a discussion of how KG-based approaches can catalyse educational change in pharmaceutical sciences and healthcare informatics, allowing more intuitive exploration of drug-disease relationships and supporting more informed clinical decision-making. We present examples of how these techniques have been successfully applied in educational settings to improve students' understanding of complex pharmaceutical concepts and to train healthcare professionals in evidence-based prescribing practices. By the end of this tutorial, participants will have gained practical skills in biomedical data wrangling and KG creation, along with a deeper understanding of how these computational approaches can transform both drug discovery research and healthcare education. Participants will leave with access to sample datasets, code templates, and visualization examples that they can immediately apply to their own work, whether in research, industry, or educational contexts.

CCS Concepts

• Information systems → Hierarchical data models; **XPath**; **Ontologies**; **Extraction, transformation and loading**; **Data cleaning**; **• Applied computing → Bioinformatics**.

Keywords

Knowledge Graphs, Drug Repurposing, Data Wrangling, Healthcare Informatics, Graph Databases, Biomedical Data Integration, Neo4j, Named Entity Recognition

ACM Reference Format:
Matilde Pato, Carolina Pereira, and Nuno Datia. 2018. From Messy Data to Medical Insights: Creating Knowledge Graphs for Drug Repurposing. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Objectives

- Introduce participants to the fundamentals of **knowledge graph construction** for pharmaceutical data.
- Demonstrate practical techniques for **biomedical data wrangling**, focusing on entity resolution and relationship extraction.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

- Provide hands-on experience with **visualization and querying** of drug-disease relationships using Neo4j.
- Illustrate how knowledge graph-based approaches can transform **healthcare education and drug discovery research**.
- Participants are to be equipped with **reusable templates and methodologies** that they can apply to their own work.

2 Relevance

This tutorial directly addresses the conference theme "Computer Science: a Catalyst for Educational Change" by demonstrating how computational approaches to data integration can transform healthcare education. Knowledge graphs make complex biomedical relationships more intuitive and accessible, enabling students, researchers, and clinicians to develop deeper insights into drug mechanisms and therapeutic opportunities.

The tutorial showcases how computer science techniques – specifically **data wrangling**, **graph theory**, and **information visualization** – can overcome the limitations of traditional educational resources in pharmaceutical sciences. By making invisible relationships between drugs, diseases, and biological processes visible and interactive, these approaches foster more drawing and effective learning experiences.

Beyond education, the tutorial also demonstrates how these same techniques support scientific discovery and clinical decision-making, illustrating the broader impact of computer science in addressing real-world healthcare challenges.

3 Details of organiser(s) and presenter(s):

Matilde Pato. (PhD in Biomedical Engineering) is an Adjunct Professor at the Lisbon School of Engineering (ISEL) and a Researcher at both the Institute of Biophysics and Biomedical Engineering (IBEB) and LASIGE research laboratory. Her teaching portfolio includes courses in Information Systems, Big Data Engineering, and Large Scale Machine Learning, providing her with comprehensive expertise in data management and analytics techniques essential for this tutorial. Pato has supervised and co-supervised numerous master's degree works in Computer Science and is co-author of dozens of papers in prestigious journals and conference proceedings. Her research extends beyond dataset development to include the implementation of recommender systems in bioinformatics, with a particular focus on healthcare applications. She has a strong track record of participation in European and national R&D projects in collaboration with industry partners, bringing practical, real-world perspectives to her academic work. Her experience in knowledge dissemination includes co-organizing a Lecture-Style Tutorial at KDD 2021 entitled "Creating Recommender Systems Datasets in Scientific Fields" and chairing a national conference. Pato also leads a discussion forum where teachers and researchers engage with contemporary computer engineering topics. Her current research focuses on developing recommendation datasets and systems for the health field, making her uniquely qualified to lead this tutorial on knowledge graph creation for drug repurposing. This tutorial represents a natural extension of Pato's ongoing work in biomedical data integration and her commitment to educational innovation in healthcare informatics. Her interdisciplinary background bridges the technical aspects of computer science with practical applications in biomedical research, enabling her to effectively communicate complex concepts to diverse audiences.

Carolina Pereira. (Undergraduate in Computer Science) is a final-year student in Computer Science and Engineering at Lisbon School of Engineering (ISEL). Throughout her academic journey, she has shown a growing interest in the application of data science across various sectors, having contributed to research projects focused on biomedical data integration and the curation of datasets for clinical and scientific purposes. Her academic background includes coursework in Data Engineering, Information Systems, and Data Integration, where she developed strong technical skills in the analysis and processing of large volumes of data. She has practical experience with tools and technologies such as Python, SQL, R, Apache Airflow, and relational database management systems, applied in both academic and research settings. She is driven by the desire to contribute to computational solutions that enhance efficiency, accessibility, and quality through the responsible and innovative use of data. She is particularly interested in methodologies for integrating and harmonizing heterogeneous data and how these can support, in particular, biomedical research and clinical decision-making.

Nuno Datia. (PhD in Informatics) is an Associate Professor at ISEL and a researcher at NOVA LINCS (Laboratory of Computer Science and Informatics). He coordinates a Master's programme in Informatics and Computer Engineering and leads the "Learning and Knowledge Modelling" group. Prof. Datia has supervised and co-supervised numerous undergraduate and master theses in Computer Science and Computer Engineering. He is co-author of several dozen publications in national and international journals and conference proceedings, and an active member in European and national R&D projects, with an active role in data management and machine learning related tasks. Within the U!REKA European University Alliance, he is part of the core team and leads Task 5.3, which focuses on co-creating urban solutions with students. He has also co-organised and participated in several Blended Intensive Programmes (BIPs) and Collaborative Online International Learning (COIL) initiatives in partnership with European and African institutions. His research interests include machine learning, spatio-temporal analysis, visualisation and data analytics.

4 Tutorial description

Target audience

This tutorial is designed for participants with intermediate programming skills and an interest in healthcare informatics, bioinformatics, or pharmaceutical research. The ideal participant would have:

(1) Basic familiarity with Python programming.
(2) Some experience with data processing or analysis.
(3) Interest in healthcare applications of computer science.

While the tutorial focuses on drug-related data, the techniques presented are applicable to many domains involving complex relational data. We welcome participants from diverse backgrounds, including computer science, healthcare, and life sciences.

Maximum number of participants

15 (to ensure adequate support during hands-on exercises)

Equipment requirements

(1) Participants should bring laptops with Python installed.
(2) Pre-tutorial instructions will be provided for installing Neo4j Desktop and required Python libraries.
(3) Sample datasets will be made available for download prior to the tutorial.

Schedule

00:00-00:15 (15 min)	Introduction to KGs for drug repurposing Challenges in biomedical data integration Advantages of graph-based representations Overview of public pharmaceutical data sources
00:15-00:45 (30 min)	Key techniques for biomedical data wrangling NER for drugs and diseases Relationship extraction from unstructured text Demonstration of the MedJsonify methodology
00:45-01:15 (30 min)	Hands-on: Creating and visualizing a simple drug-disease graph Loading prepared datasets into Neo4j Writing basic Cypher queries to explore relationships Visualization techniques for revealing patterns Identifying potential drug repurposing candidates
01:15-01:30 (15 min)	Q&A and resources for continued learning Access to code templates and datasets Recommended tools and libraries Community resources and further reading

Acknowledgments

This work is supported by UID/04516/NOVA Laboratory for Computer Science and Informatics (NOVA LINCS), UID/BIO/00645/IBEB, UIDB/00408/LASIGE and UIDP/00408/LASIGE with the financial support of FCT.IP.

References

[1] Bethesda (MD) 2024. DailyMed, U.S. National Library of Medicine. Available from https://dailymed.nlm.nih.gov/dailymed/.
[2] Kirill Degtyarenko, Paula de Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. 2007. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research* 36, suppl_1 (10 2007), D344–D350. https://doi.org/10.1093/nar/gkm791
[3] FDA. 2025. Orange Book, Approved Drug Products With Therapeutic Equivalence Evaluations. Available from https://www.accessdata.fda.gov/scripts/cder/ob/default.cfm.
[4] Food FDA and Drug Administration. 2024. Purple Book, Database of Licensed Biological Products. Available from https://purplebooksearch.fda.gov/.
[5] Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Chin, Seth A Strawbridge, et al. 2023. DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Research* 52, D1 (11 2023), D1265–D1275. https://doi.org/10.1093/nar/gkad976
[6] Tye Rattenbury, Joseph M Hellerstein, Jeffrey Heer, Sean Kandel, and Connor Carreras. 2017. *Principles of data wrangling: Practical techniques for data preparation.* " O'Reilly Media, Inc.".
[7] Lynn M Schriml, James B Munro, Mike Schor, Dustin Olley, Carrie McCracken, Victor Felix, J Allen Baron, Rebecca Jackson, Susan M Bello, Cynthia Bearer, Richard Lichenstein, Katharine Bisordi, Nicole Campion Dialo, Michelle Giglio, and Carol Greene. 2021. The Human Disease Ontology 2022 update. *Nucleic Acids Research* 50, D1 (11 2021), D1255–D1261. https://doi.org/10.1093/nar/gkab1063
[8] SS Weinreich, R Mangon, JJ Sikkens, M E en Teeuw, and MC Cornel. 2008. Orphanet: a European database for rare diseases. *Nederlands tijdschrift voor geneeskunde* 152, 9 (March 2008), 518—519.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009