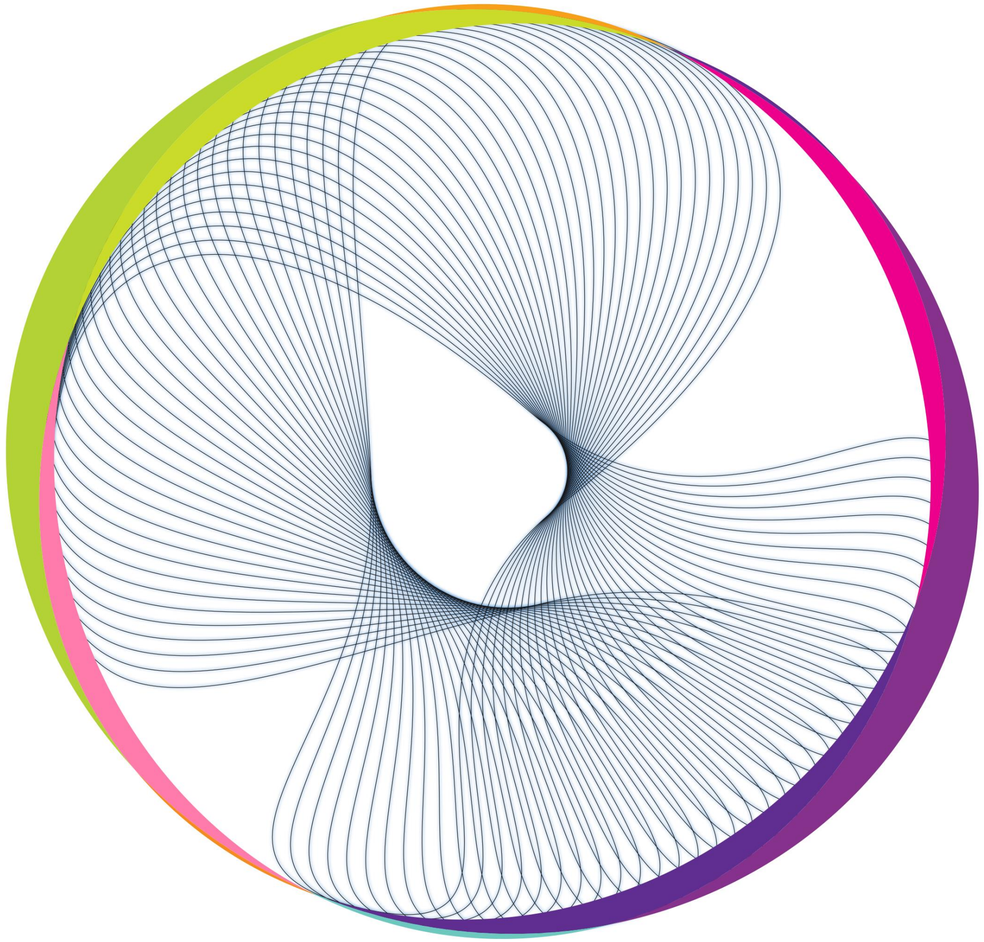
ROBERT KOCH INSTITUT



NLPCGen
Natural Language Processing For Cancer Genomics

Center for Artificial Intelligence in Public Health Research, Robert Koch Institute
Department of Mathematics and Computer Science, Freie Universität Berlin

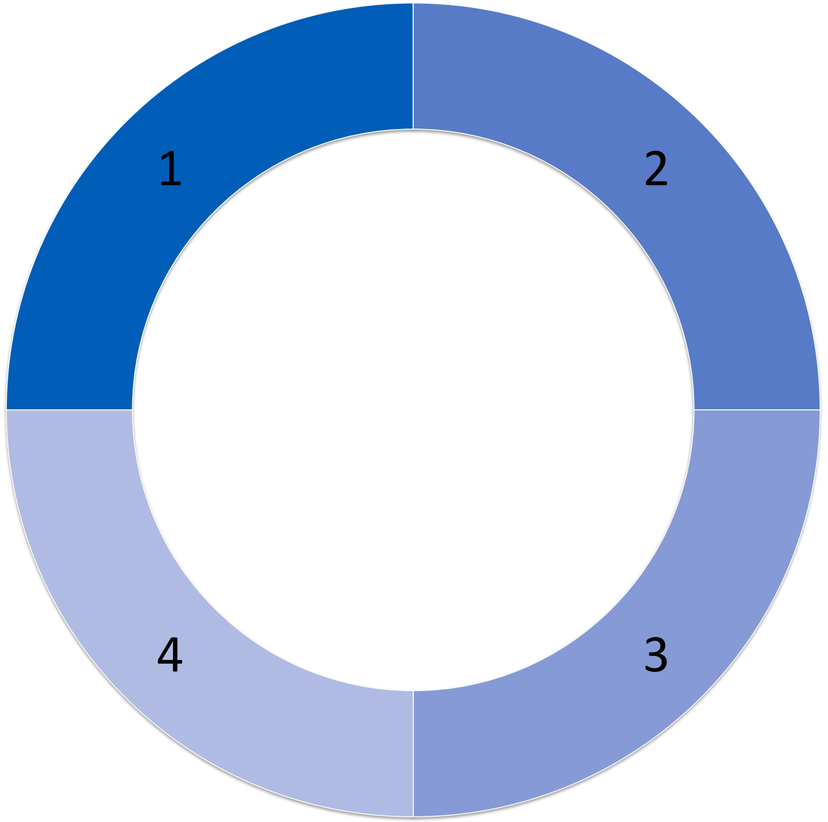
24.06.2024 | EU Health Policy Platform Thematic Network Webinar

Background and Scope

- **Reduce the burden** by avoiding risk factors, implementing evidence-based prevention strategies, early detection via screening, appropriate treatment and care of cancer patients.
- **Screening** aims to identify individuals with findings suggestive of a specific cancer or pre-cancer before they have developed symptoms.
- Natural Language Processing (NLP) techniques can find patterns, structures, and subtleties within the language of DNA and already help in genome annotation.
- **Aim**: Make use of NLP techniques for cancer genomics.

Motivation

1. Address global cancer challenge (EU strategic priority).
2. Integrate and interpret diverse genomic data and other metadata.
3. Leverage the proven potential of NLP as part of Artificial Intelligence (AI).
4. Facilitate downstream tasks.



From Global Challenge to Bite-sized Problem

- Special focus on Cancer-causing Viruses.
- Use NLP to extract information from genomes, literature, clinical notes, and databases.
- Identify genomic hotspots linked to cancer development, progression, and treatment response.

Seven cancer-causing viruses

There are currently seven viruses known that can cause cancer, which are technically called "oncogenic viruses."

This virus	Can cause this cancer
Human papillomavirus	Cervical carcinoma
Epstein-Barr virus	Hodgkin lymphomas
Human T-lymphotropic virus	Adult T-cell leukemia
Kaposi's sarcoma-associated herpes virus	Kaposi's sarcoma
Merkel cell polyoma virus	Merkel cell carcinoma
Hepatitis B virus	Hepatocellular carcinoma
Hepatitis C virus	Hepatocellular carcinoma

Table: The Conversation, CC-BY-ND • Source: American Society for Microbiology

Challenges of NLP Techniques for Cancer Genomics

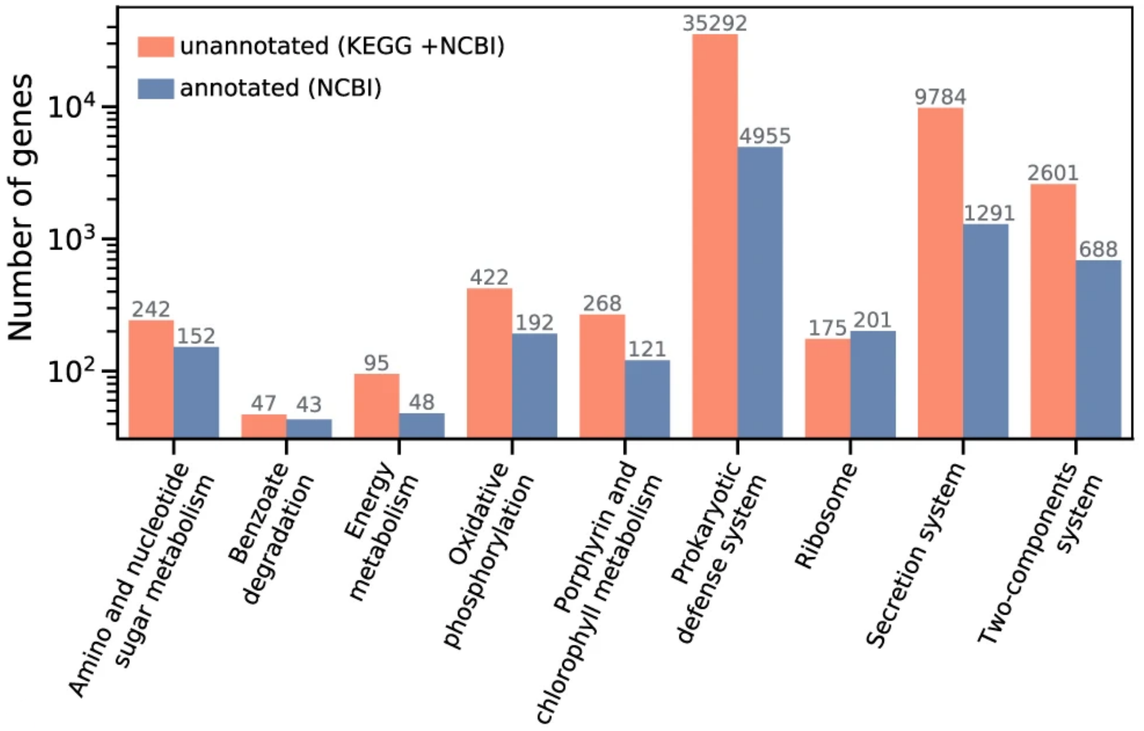
- High data volume from Next Generation Sequencing technologies.
- Handling unstructured and structured data.
- Dealing with Genomic Variants and Mutations.
- Capturing Genomic Context and Functional Annotations.
- Integration with existing Systems and large Databases.
- Ethical and Privacy Concerns with Genomic data.

Potential of NLP Techniques for Cancer Genomics

- Proven potential as part of AI techniques.
- Human assistance with literature review, clinical decision support, drug discovery, etc.
- It might extract gene names, protein names, drug names, or specific types of cancer (entity extraction).
- It might identify that a particular gene is associated with a specific type of cancer (relation extraction).
- Already in use to identify unknown (unannotated) genes.

NLP Example to Identify Unknown or Unannotated Genes

- Gene Function Analysis from Bioinformatics.
- Known (annotated) genes play a key role in finding the causes and treatment for many diseases.
- Develop NLP techniques using deep learning and specialized algorithms.



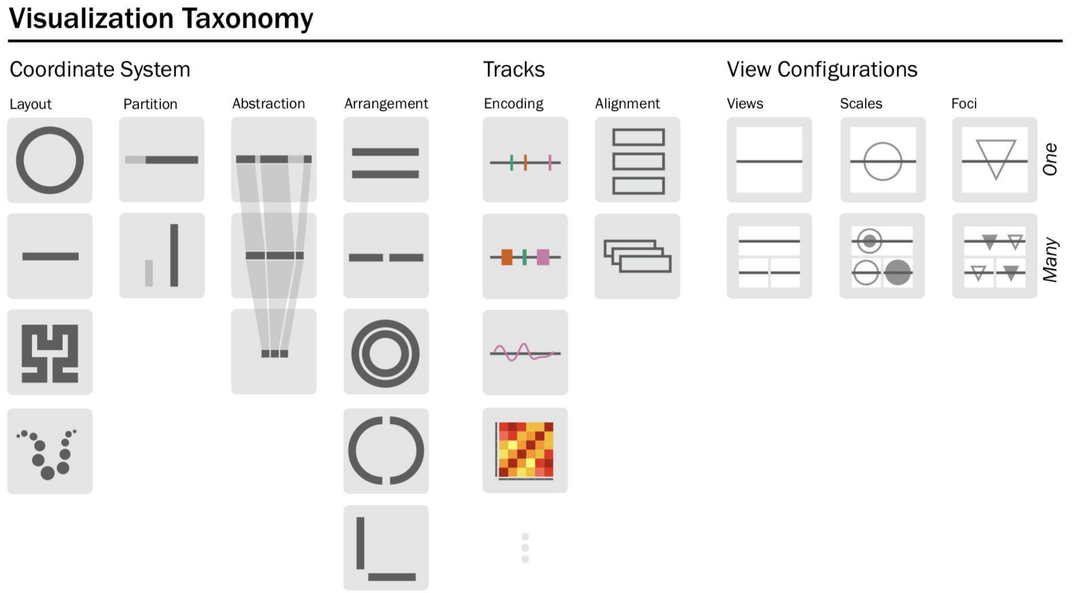
Adapted from Miller, D., Stern, A., & Burstein, D. (2022). Deciphering microbial gene function using natural language processing. *Nature Communications*, *13*(1), 5731.

Necessary Efforts

- A genome contains genes, it is important to understand the “semantics” to decipher it.
- Make use of NLP to capture "gene semantics" from genomic data for context understanding.
- Context dependence for broader comprehension in cancer genomics.
- Enrich genomic data by recognizing gene symbols and full gene names in cancer literature and reports.
- Integrative solutions powered by Visualization to identify actionable insights and facilitate screening.

Expected Outcomes

- Improved understanding of cancer genomics and molecular mechanisms.
- Discovery of genetic variations, mutations, biomarkers for specific cancers.
- Development of screening strategies.
- Enhanced interpretation and integration of diverse genomic data and metadata.
- Increase the adoption of AI in healthcare.



Nusrat, S., Harbig, T., & Gehlenborg, N. (2019, June). Tasks, techniques, and tools for genomic data visualization. In *Computer Graphics Forum* (Vol. 38, No. 3, pp. 781-805).

Potential Limitations and Biases of AI-based Solutions

- Large gap between unstructured and structured data for creating dictionaries in cancer genomics.
- Over- and under-representativeness of certain genes in biomedical texts, literature, and clinical notes.
- Underrepresented populations in databases, later used for machine learning.

https://digital-strategy.ec.europa.eu/en/policies/1-million-genomes

Call for Collaborative Network and Research

- Tackle challenges openly and together with bioinformaticians, oncologists, clinicians, etc.
- Foster collaborations and ensure clinical utility and applicability of the research.
- Share our Visualization, AI, and NLP expertise.
- Draft an action-plan to improve research in the field.

Conclusion

- Foster interdisciplinary collaborations and develop AI/NLP methodologies.
- Transform AI potential to bring about positive change to society by promoting a screening future.
- Significant impact on cancer diagnosis, treatment, patient outcomes.
- Advancing cancer genomics research, early detection, personalized treatment.

Thank you for your attention.

NLPCGen