HPC/HDPA/AI Applications in Data Analytics, Natural Language Processing and Information Retrieval

Maria Nisheva, Sofia University St. Kliment Ohridski

Decision support systems: use of big data and small data

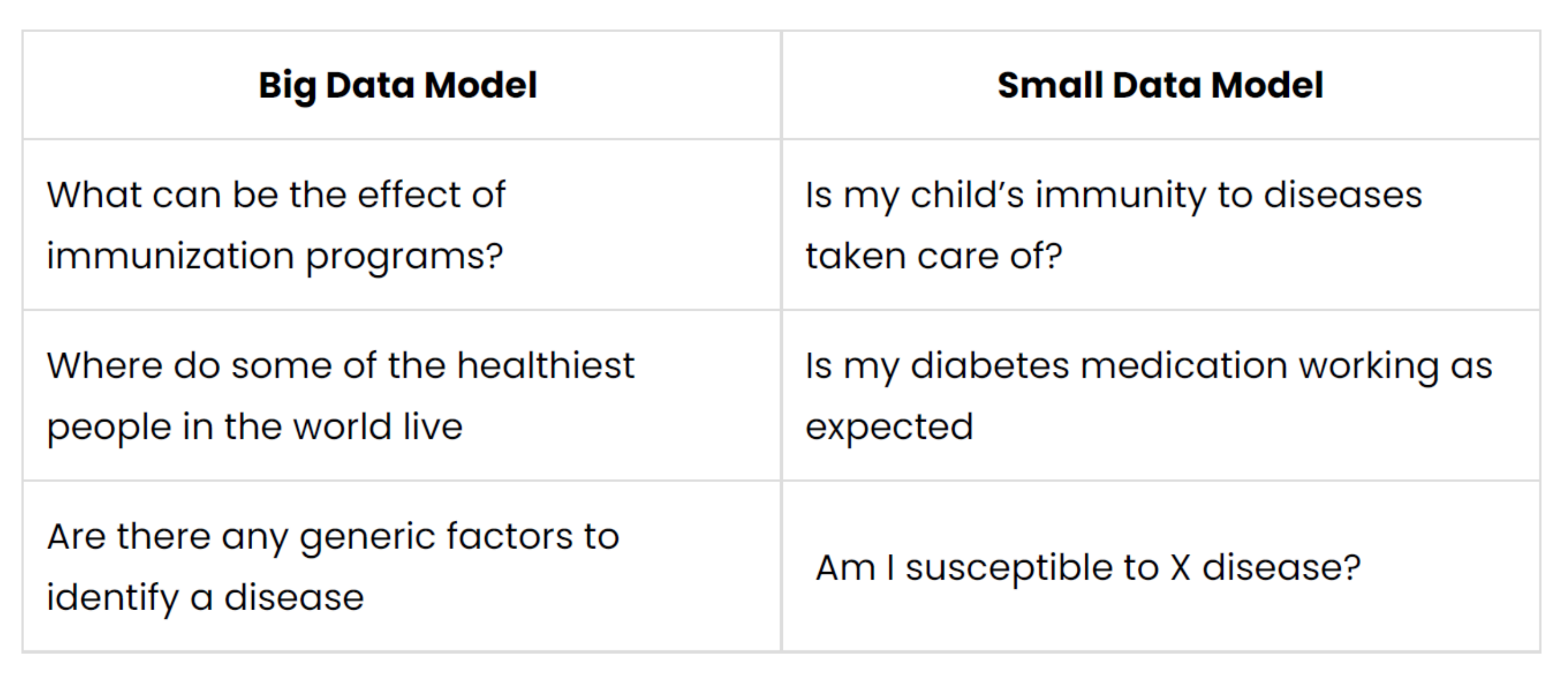
- DSSs: information systems that support business or organizational decision-making activities. DSSs serve the management, operations and planning levels of an organization and help people make decisions about problems that may be rapidly changing and not easily specified in advance
- Our recent experience: a DSS for dietary recommendations for type 2 diabetes mellitus

- Big Data: the term is used to describe the massive volume of both structured and unstructured data that is so large that it is difficult to process using traditional techniques
- Small Data: data that is ‘small’ enough for human comprehension. It is data in a volume and format that makes it accessible, informative and actionable

- Role of Big Data and Small Data in building DSSs:
  - “Big Data is all about *finding correlations*, but Small Data is all about *finding the causation*, the reason why”
  - big data is important is all cases of building medium-term and long-term policies and strategic decisions
  - small data refers to definite and specific attributes of datasets, which can be used to analyze the current situation in depth and to make adequate personalized decisions. Therefore, small data is best placed to support decision-making at the current time

➢*Example area: healthcare*

Clinicians favor small data over big data for healthcare assessments and prediction models.



Big Data Model	Small Data Model
What can be the effect of immunization programs?	Is my child’s immunity to diseases taken care of?
Where do some of the healthiest people in the world live	Is my diabetes medication working as expected
Are there any generic factors to identify a disease	Am I susceptible to X disease?

- Data centric AI: the concept refers to *building AI systems with quality data*. The data centric AI approach is based on the idea to focus on ensuring that the data used clearly show what the developed AI system needs to learn
  - ➢ If until recently the dominant idea was to focus on improving the code, nowadays *it is more effective for a lot of applications to consider that* the quality of code is generally a solved problem and *the focus should be moved to finding approaches to improve the data*

- In particular, instead of working directly with a large amount of raw and noisy data, it is better to make at the beginning appropriate efforts to improve the consistency of the data and in this way to achieve a significant improvement in productivity
- Especially for big data applications, the common approach has been: “If the data is noisy, let’s just get a lot of data and the algorithm will average over it”. But the data centric approach assumes to try to develop tools that point on data inconsistencies and give an effective way to overcome most of them in order to get a truly high performing system

- Transfer learning: *a machine learning method where a model developed for a task is reused as the starting point for a model on a second task*
  - ➢ It is a popular approach in deep learning where pre-trained models are used as the starting point on computer vision and NLP tasks

- A transfer learning model needs less data as compared to a model built from scratch
- A transfer learning model needs less computation power as compared to a model built from scratch
- A transfer learning model requires less time because most of the heavy work is already done on the pre-trained model and only a relatively small part is done by the new model

Medico-Help: A virtual health assistant

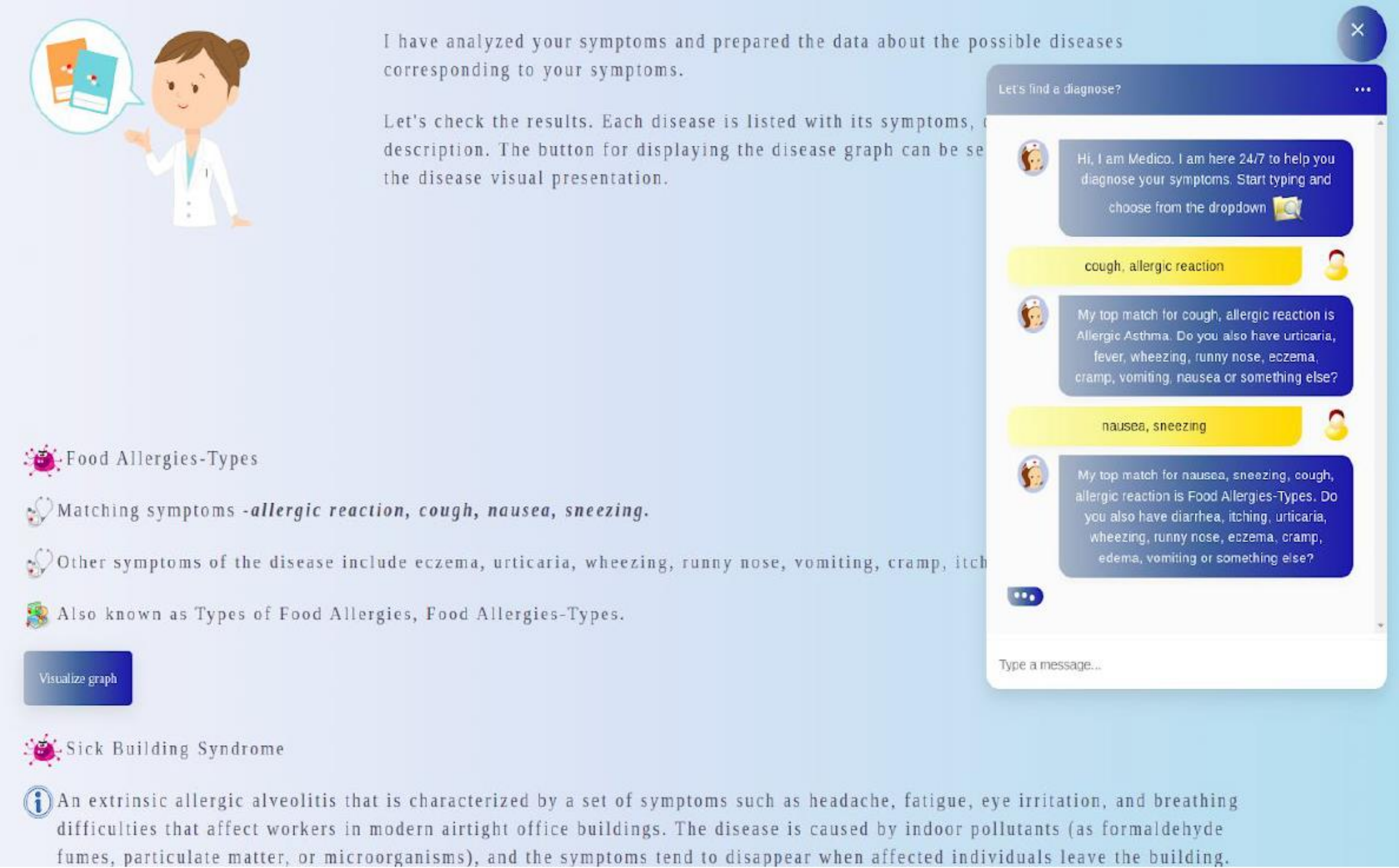
- A web-based expert system that functions as an intelligent chatbot, capable
  - ➢ to automatically collect data from trusted websites
  - ➢ to build and extend automatically a medical knowledge base and to search in it
  - ➢ to generate hypotheses for medical diagnoses based on symptoms

- Initial version of the knowledge base of Medico-Help: a small standardized ontology for human diseases, developed at the School of Medicine of the University of Maryland
- A module for automated collection of specialized data from trusted sources on the Internet: the role of such a source in the pilot version of Medico-Help is played by MedIndia

- The new data retrieved from the documents provided by MedIndia are analyzed and used to gradually enrich the domain knowledge base of Medico-Help. Information about new drugs and additional symptoms is also periodically added for this purpose

- The available version of the knowledge base is used to generate answers to the user questions, most often in the form of assumptions about diagnoses corresponding to the indicated symptoms as well as suggestions about possible treatment regimens

- Each diagnosis assumption includes information about the disease such as description, related symptoms, synonyms and drugs. The virtual assistant can also draw the user’s attention to possible other related symptoms that might be missed



I have analyzed your symptoms and prepared the data about the possible diseases corresponding to your symptoms.

Let's check the results. Each disease is listed with its symptoms, (description. The button for displaying the disease graph can be se the disease visual presentation.

Food Allergies-Types

Matching symptoms -***allergic reaction, cough, nausea, sneezing.***

Other symptoms of the disease include eczema, urticaria, wheezing, runny nose, vomiting, cramp, itch

Also known as Types of Food Allergies, Food Allergies-Types.

Sick Building Syndrome

An extrinsic allergic alveolitis that is characterized by a set of symptoms such as headache, fatigue, eye irritation, and breathing difficulties that affect workers in modern airtight office buildings. The disease is caused by indoor pollutants (as formaldehyde fumes, particulate matter, or microorganisms), and the symptoms tend to disappear when affected individuals leave the building.

Hi, I am Medico. I am here 24/7 to help you diagnose your symptoms. Start typing and choose from the dropdown

cough, allergic reaction

My top match for cough, allergic reaction is Allergic Asthma. Do you also have urticaria, fever, wheezing, runny nose, eczema, cramp, vomiting, nausea or something else?

nausea, sneezing

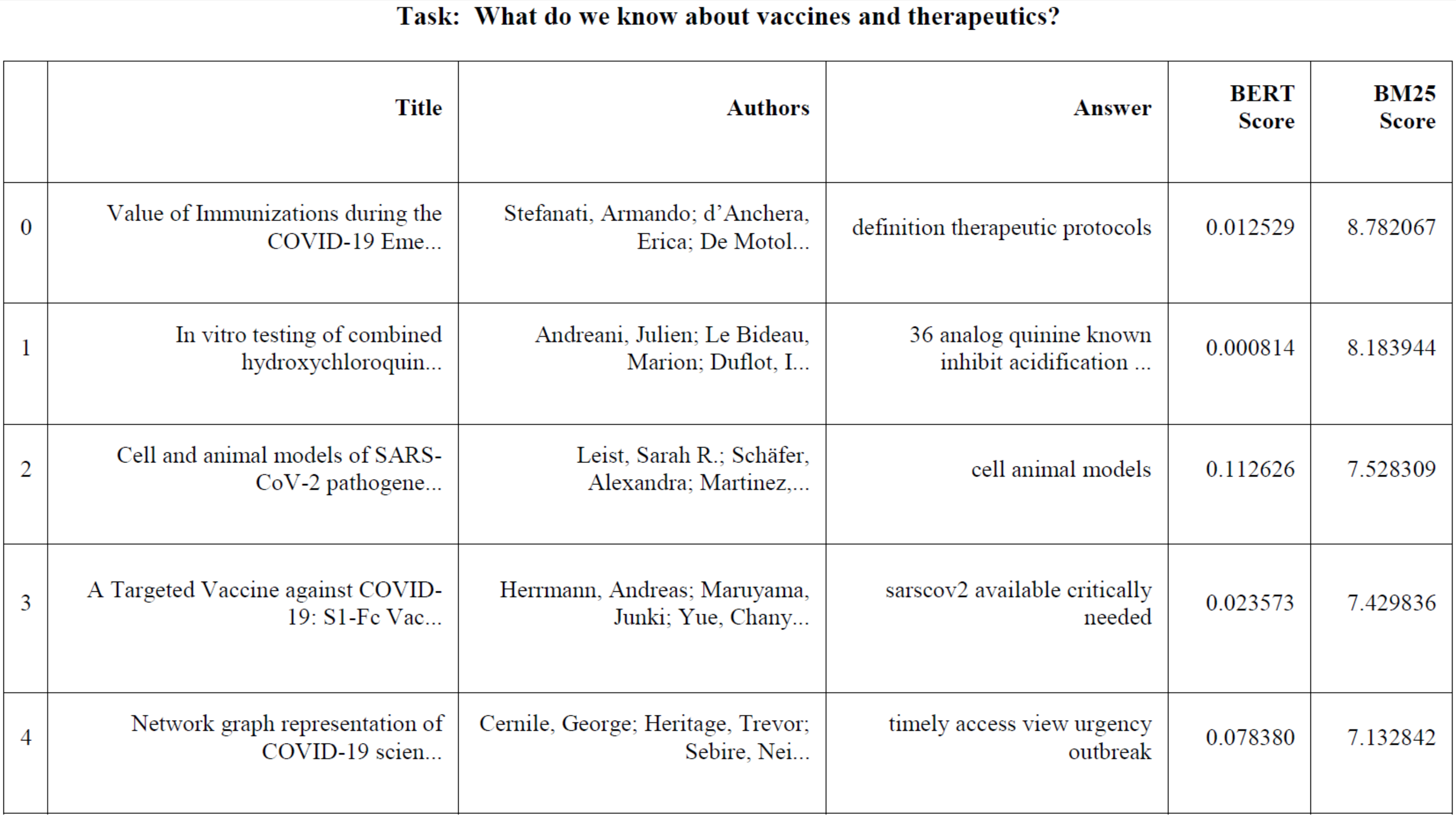
My top match for nausea, sneezing, cough, allergic reaction is Food Allergies-Types. Do you also have diarrhea, itching, urticaria, wheezing, runny nose, eczema, cramp, edema, vomiting or something else?

Intelligent system for answering specialized questions about COVID-19

- The development of the system was motivated by the popular COVID-19 Open Research Dataset Challenge of Kaggle
- CORD-19: a large and growing collection of publications and preprints on Covid-19 and previous coronaviruses such as SARS and MERS. It integrates papers and preprints from several sources, collected by Semantic Scholar

- The process of developing the system
  - preliminary preparation of the data (recognition of the language of each of the available papers and selection of those in English; tokenization of the abstracts and texts of the selected papers) – results in the actual working version of the dataset, the content of which is used to generate the answers to the user questions

- The process of generating answers to the user question
  - ➢ determining the rank of each paper in the dataset relative to the user question (Okapi BM25 best matching ranking algorithm) and selecting the first five of them
  - ➢ using the BERT Large Uncased Whole World Masking model, pre-trained and fine-tuned on the Stanford Question Answering Dataset. The texts of the selected papers and the user question are submitted to it. As a result of the execution of BERT, the generated answers to the user question are returned



Task: What do we know about vaccines and therapeutics?

	Title	Authors	Answer	BERT Score	BM25 Score
0	Value of Immunizations during the COVID-19 Eme...	Stefanati, Armando; d'Anchera, Erica; De Motol...	definition therapeutic protocols	0.012529	8.782067
1	In vitro testing of combined hydroxychloroquin...	Andreani, Julien; Le Bideau, Marion; Duflot, I...	36 analog quinine known inhibit acidification ...	0.000814	8.183944
2	Cell and animal models of SARS-CoV-2 pathogene...	Leist, Sarah R.; Schäfer, Alexandra; Martinez,...	cell animal models	0.112626	7.528309
3	A Targeted Vaccine against COVID-19: S1-Fc Vac...	Herrmann, Andreas; Maruyama, Junki; Yue, Chany...	sarscov2 available critically needed	0.023573	7.429836
4	Network graph representation of COVID-19 scien...	Cernile, George; Heritage, Trevor; Sebire, Nei...	timely access view urgency outbreak	0.078380	7.132842

- Analysis of the obtained experimental results
  - ➢ the system is relatively good at generating answers to specific questions
  - ➢ it would be useful to enrich the dataset with which the system works with other types of documents related to COVID-19 (technical reports, messages from governmental institutions and public organizations, etc.)
  - ➢ although a domain-specific corpus of data was used to create the system, the approach developed is quite general and can be applied in other areas

Conclusion

The application of data centric AI techniques would contribute to the rapid creation of intelligent software systems with great impact on large target groups, providing personalized services and reliable content.

It is a promising direction of effective collaboration with industry and public institutions.

EURO

Thank you for your attention!

For more information and contacts:

marian@fmi.uni-sofia.bg