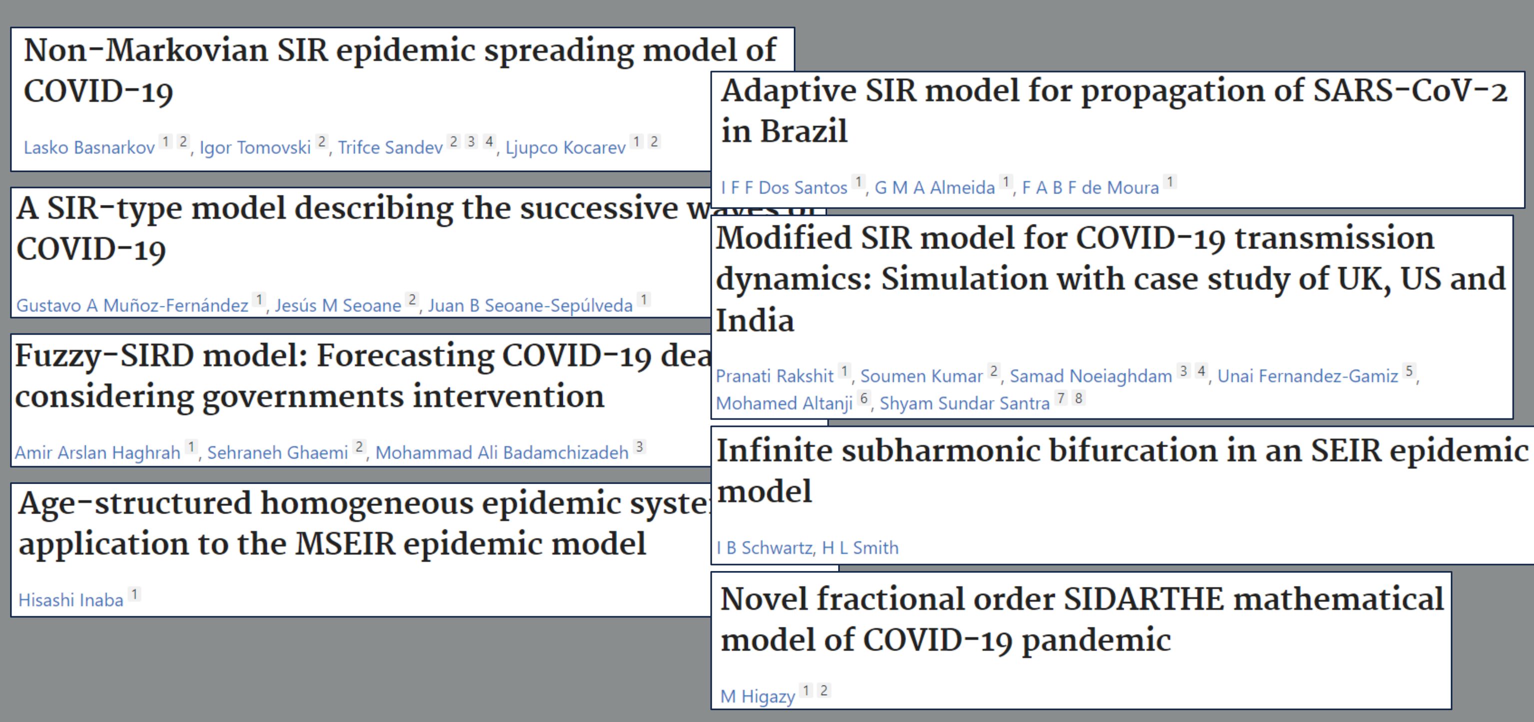
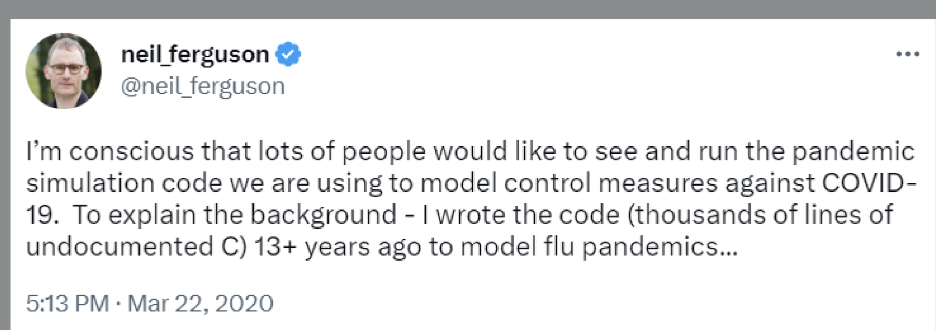
Automatically Extracting and Annotating Models From Scientific Publications and Code

Markos Markakis, Chunwei Liu, Peter Baile Chen, Michael Cafarella

"Standing on the Shoulders of Giants" is Becoming Harder





- Dozens of new models in publications each year. Difficult to:
  - Remain well-informed as a researcher.
  - Educate the general public.
  - Constructively build upon existing work.
- Models are not like long-lasting software products:
  - Partially described by papers, results and other artifacts.
  - Code might not even exist.
  - Definitely not implemented with extensibility in mind.
- Early days of the COVID-19 pandemic: top-level policy decisions based on old undocumented code.

DARPA's Automating Scientific Knowledge Extraction and Modeling (ASKEM) project [1]

- Develop "tools will enable experts to maintain, reuse, and adapt large collections of heterogeneous data, knowledge and models".

Entity Annotation is a Basic Building Block

Code Self-Documentation is Often Lacking

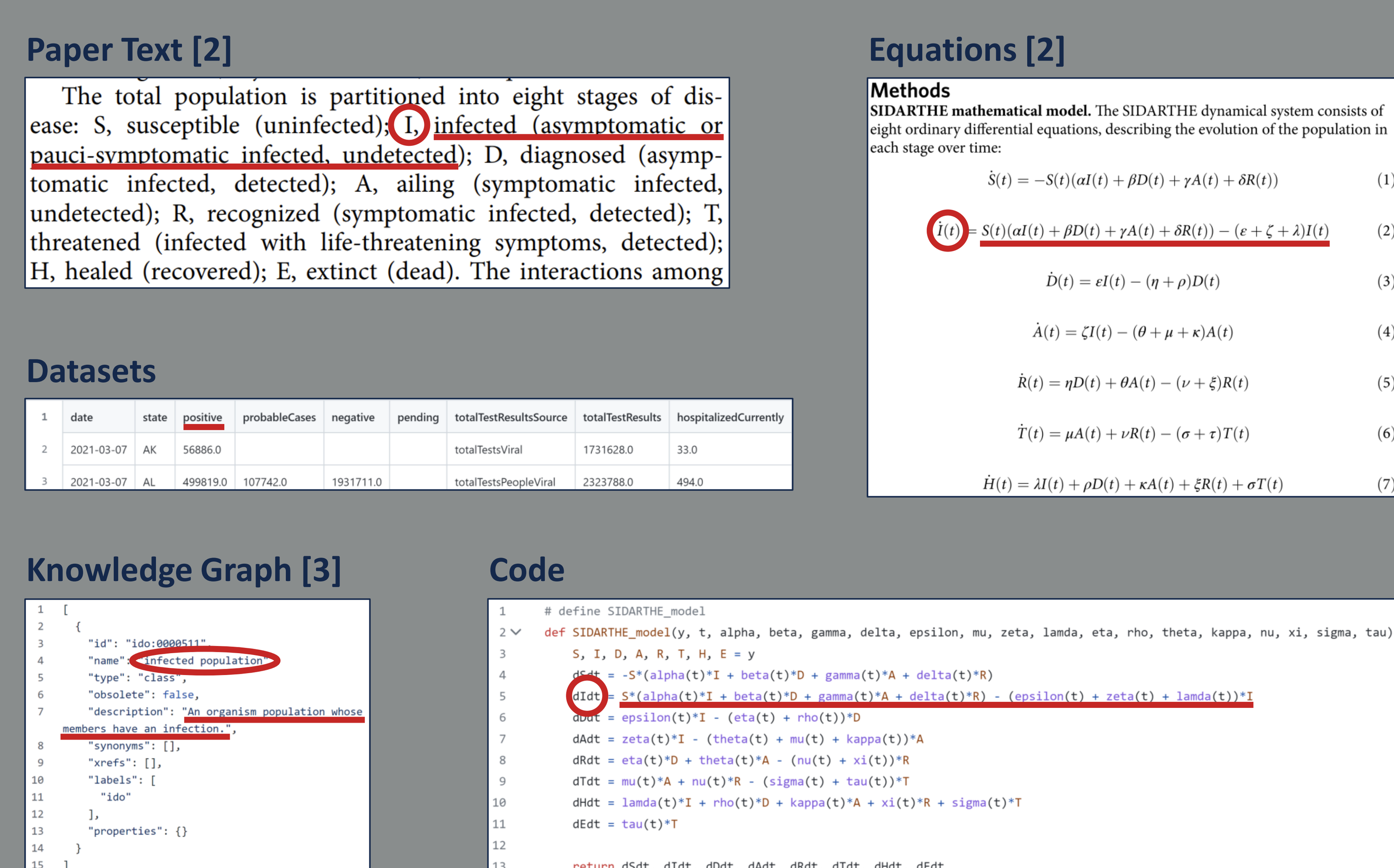
- In order to reuse/extend a model, scientists must understand it.
- Annotations in the code itself might be insufficient.
- But the knowledge exists in the original model description.
- **Task 1: Can we annotate code elements with their descriptions from text/equations?**

Different Papers, Different Terms

- Terminology might not be standardized across works, making model comparison harder.
- **Task 2: Can we map model terms to a single source, like a Domain Knowledge Graph (DKG)?**

Models Without Data are Unusable

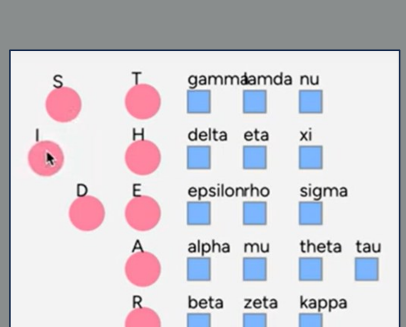
- To evaluate a model, data must be provided for each variable.
- The data schema might not match the variable definitions.
- **Task 3: Can we find the most appropriate data for each variable?**



Large Language Models to the Rescue!

```
We have several variable names and descriptions as follows, each on a newline:
[DESC]

We also have a list of column names as follows:
[DATASET]

From the dataset columns stated above, print up to two that match the following variable name: "[TARGET]", each one a newline.
```

```
Here is a text description of a model:

[TEXT]

Represent this model as a Petri net and print the places on one line.
```



Name	I	
ID	v1	
Type	variable	
DOI	https://doi.org/10.1038/s41591-020-0883-7	
File	41591_2020_Article_883.pdf	
Text Annotations	Infected (asymptomatic or pauci-symptomatic infected, undetected)	
DKG Annotations	infected population	ido:0000511
	COVID-19 Infection	ncit:C171133
Equation Annotations	{\dot{A}}(t)=\zeta I(t)-(\theta+\mu+\kappa)A(t)	I
	{\hat{D}}(t)=\varepsilon I(t)-(\eta+\rho)D(t)	I

Close Enough for Humans? Close Enough for GPT-3 [4]!

- Even though terminology for the same variable might differ across sources, the terms used are usually semantically similar enough for a human.
- Models like GPT-3 are also able to pick up on this similarity!
- After appropriate prompt engineering, we can use GPT-3 for the three tasks above.

Giving Downstream Users an Editable Model

- Our API can extract a graph description of a model from code.
- It can then annotate each variable with text descriptions, equations, datasets and/or DKG terms, whenever available.
- Downstream ASKEM teams can then visualize this model.
- Users can leverage the associated annotations to understand the model and evolve it as needed.

[1] Joshua Elliott. Automating Scientific Knowledge Extraction and Modeling (ASKEM), 2022. https://www.darpa.mil/program/automating-scientific-knowledge-extraction-and-modeling.
[2] Giordano G, Blanchini F, Bruno R, Colaneri P, Di Filippo A, Di Matteo A, Colaneri M. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. Nat Med. 2020 Jun;26(6):855-860. doi: 10.1038/s41591-020-0883-7. Epub 2020 Apr 22. PMID: 32322102; PMCID: PMC7175834.
[3] INDRA Lab in the Harvard Program in Therapeutic Science (HiTS). MIRA DKG Source Code. https://github.com/indralab/mira.
[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33:1877–1901, 2020.