

# 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

**Non-Markovian SIR epidemic spreading model of COVID-19**  
Lasko Basnarkov<sup>1,2</sup>, Igor Tomovski<sup>2</sup>, Trifce Sandev<sup>2,3,4</sup>, Ljupco Kocarev<sup>1,2</sup>

**A SIR-type model describing the successive waves of COVID-19**  
Gustavo A Muñoz-Fernández<sup>1</sup>, Jesús M Seoane<sup>2</sup>, Juan B Seoane-Sepúlveda<sup>1</sup>

**Fuzzy-SIRD model: Forecasting COVID-19 deaths considering governments intervention**  
Amir Arslan Haghrah<sup>1</sup>, Sevraneh Ghaemi<sup>2</sup>, Mohammad Ali Badamchizadeh<sup>3</sup>

**Age-structured homogeneous epidemic system application to the MSEIR epidemic model**  
Hisashi Inaba<sup>1</sup>

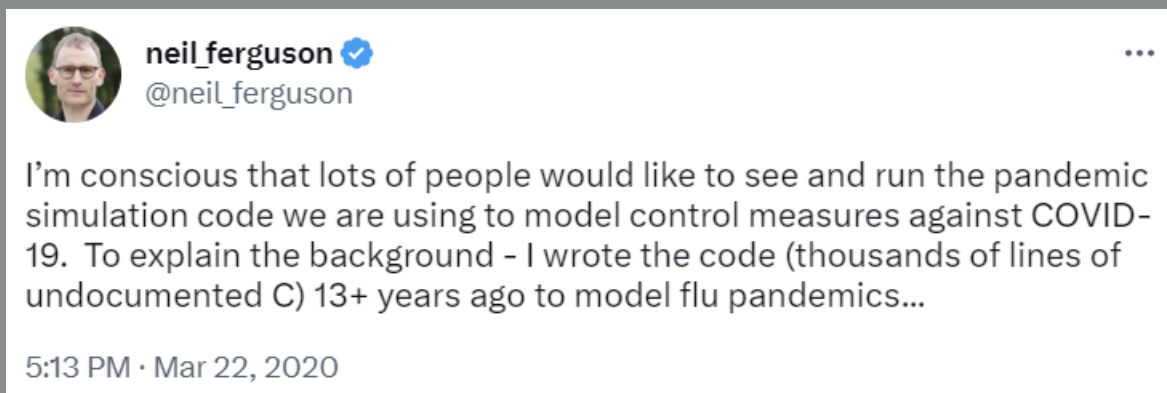
**Adaptive SIR model for propagation of SARS-CoV-2 in Brazil**  
I F F Dos Santos<sup>1</sup>, G M A Almeida<sup>1</sup>, F A B F de Moura<sup>1</sup>

**Modified SIR model for COVID-19 transmission dynamics: Simulation with case study of UK, US and India**  
Pranati Rakshit<sup>1</sup>, Soumen Kumar<sup>2</sup>, Samad Noeiaghdam<sup>3,4</sup>, Unai Fernandez-Gamiz<sup>5</sup>, Mohamed Altanji<sup>6</sup>, Shyam Sundar Santra<sup>7,8</sup>

**Infinite subharmonic bifurcation in an SEIR epidemic model**  
I B Schwartz, H L Smith

**Novel fractional order SIDARTHE mathematical model of COVID-19 pandemic**  
M Higazy<sup>1,2</sup>

- 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?**

#### Paper Text [2]

The total population is partitioned into eight stages of disease: S, susceptible (uninfected), I, infected (asymptomatic or pauci-symptomatic, infected, undetected); D, diagnosed (asymptomatic, infected, detected); A, ailing (symptomatic, infected, undetected); R, recognized (symptomatic, infected, detected); T, threatened (infected with life-threatening symptoms, detected); H, healed (recovered); E, extinct (dead). The interactions among

#### Equations [2]

**Methods**  
SIDARTHE mathematical model. The SIDARTHE dynamical system consists of eight ordinary differential equations, describing the evolution of the population in each stage over time:

$$\begin{aligned} \dot{S}(t) &= -S(t)(aI(t) + \beta D(t) + \gamma A(t) + \delta R(t)) & (1) \\ \dot{I}(t) &= S(t)(aI(t) + \beta D(t) + \gamma A(t) + \delta R(t)) - (\epsilon + \zeta + \lambda)I(t) & (2) \\ \dot{D}(t) &= \delta I(t) - (\eta + \rho)D(t) & (3) \\ \dot{A}(t) &= \zeta I(t) - (\theta + \mu + \kappa)A(t) & (4) \\ \dot{R}(t) &= \eta D(t) + \theta A(t) - (\nu + \tau)R(t) & (5) \\ \dot{T}(t) &= \mu A(t) + \nu R(t) - (\sigma + \epsilon)T(t) & (6) \\ \dot{H}(t) &= \lambda I(t) + \rho D(t) + \kappa A(t) + \tau R(t) + \sigma T(t) & (7) \end{aligned}$$

#### Datasets

date	state	positive	probableCases	negative	pending	totalTestResultsSource	totalTestResults	hospitalizedCurrently
2021-03-07	AK	5686.0				totalTestViral	1731638.0	33.0
2021-03-07	AL	499819.0	107742.0	1937711.0		totalTestHospitalViral	2323788.0	4848.0

#### Knowledge Graph [3]

```
graph TD
    S["S: susceptible (uninfected)"]
    I["I: infected (asymptomatic or pauci-symptomatic, infected, undetected)"]
    D["D: diagnosed (asymptomatic, infected, detected)"]
    A["A: ailing (symptomatic, infected, undetected)"]
    R["R: recognized (symptomatic, infected, detected)"]
    T["T: threatened (infected with life-threatening symptoms, detected)"]
    H["H: healed (recovered)"]
    E["E: extinct (dead)"]
    S --- I
    I --- D
    I --- A
    I --- R
    I --- T
    D --- H
    A --- H
    R --- H
    T --- H
```

#### Code

```
def SIDARTHE_model(y, t, alpha, beta, gamma, delta, epsilon, mu, zeta, lambda, eta, rho, theta, kappa, nu, xi, sigma, tau):
    S, I, D, A, R, T, H, E = y
    dSdt = -S*(alpha*I + beta*D + gamma*A + delta*R)
    dIdt = S*(alpha*I + beta*D + gamma*A + delta*R) - (epsilon + zeta + lambda)*I
    dDdt = delta*I - (eta + rho)*D
    dAdt = zeta*I - (theta + mu + kappa)*A
    dRdt = eta*D + theta*A - (nu + xi)*R
    dTdt = mu*A + nu*R - (sigma + epsilon)*T
    dHdt = lambda*I + rho*D + kappa*A + xi*R + sigma*T
    dEdt = tau*T
```

## Large Language Models to the Rescue!

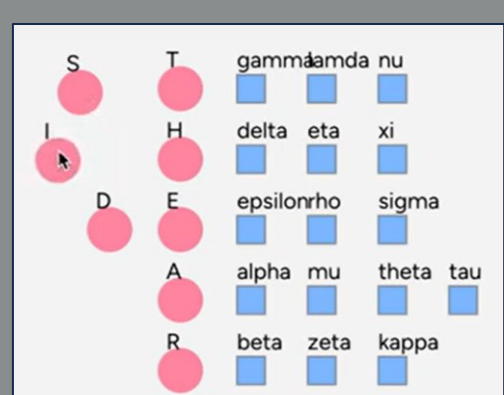
We have several variable names and descriptions as follows, each on a newline:  
[DESC]

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 id: 0000511 COVID-19 Infection nct: C171133
Equation Annotations	(lambda*I) - (rho*D) - (kappa*A) - (xi*R) - (sigma*T)

## 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.