Temporal and relational machine learning for biostatistical and other scientific applications

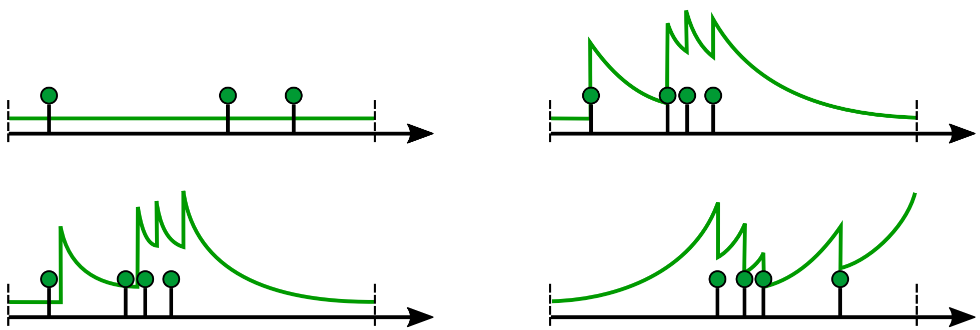
Austin R. Benson · Cornell University

Annual Conference of the International Society for Clinical Biostatistics

July 20, 2021

Joint work with
Junteng Jia
Cornell ⟶ Facebook

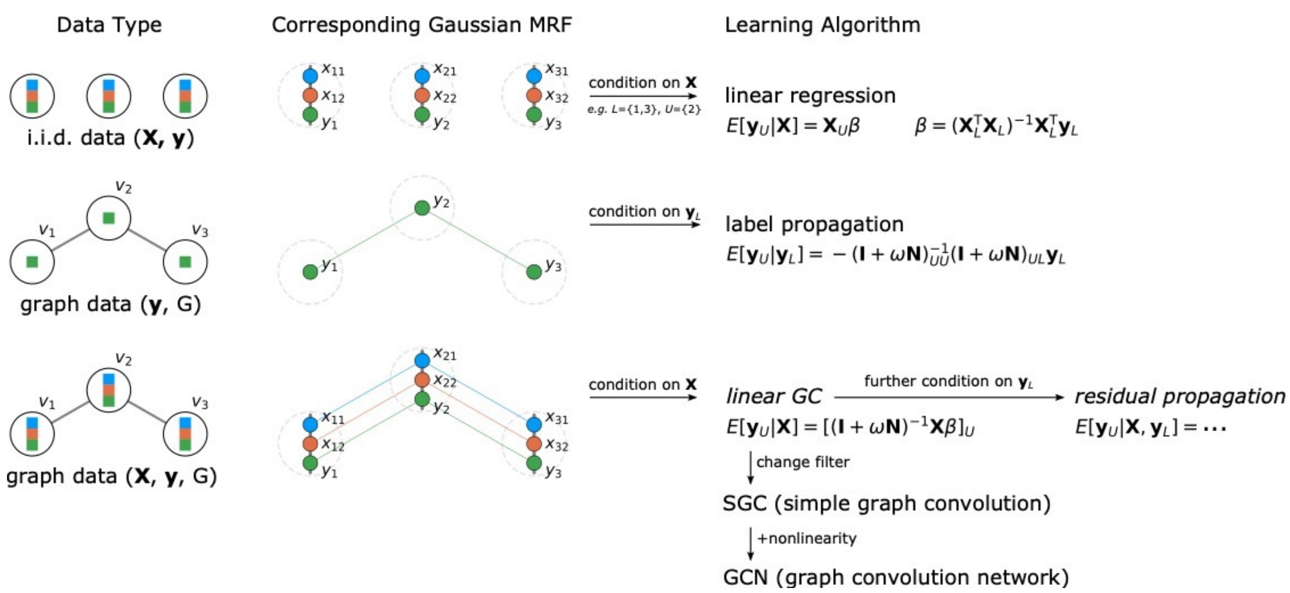
1. An existing simple model augmented with deep learning.



Predict reason for patient visit in ICU.

Neural Jump Stochastic Differential Equations, Jia and Benson, Neurips 2019.

2. A model for understanding an existing deep learning method.

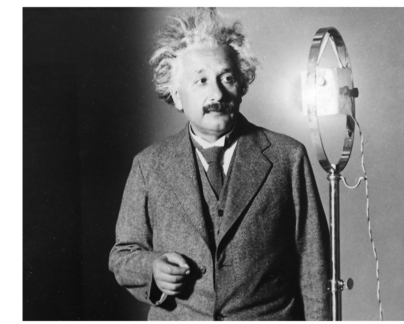


Predict air quality in regions of USA given nearby climate statistics.

A Unifying Generative Model for Graph Learning Algorithms, Jia and Benson, arXiv 2021.

Everything should be made as simple as possible,
but no simpler.

– Albert Einstein (paraphrased)



Many real-world systems evolve continuously over time but are interrupted by random events.

1. Patient health interrupted by clinical visits.
2. Disease progression interrupted by treatments.
3. Social network user interrupted by ads.

Key question

How can we simultaneously learn continuous and discrete dynamics?

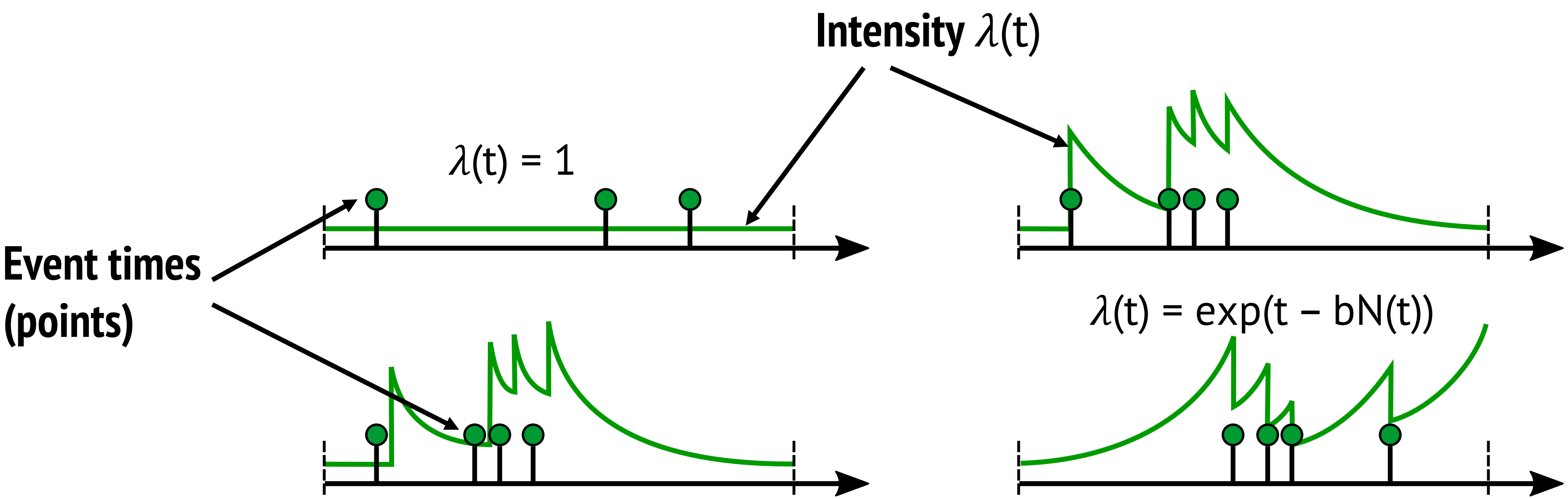
Data

$t_1, t_2, t_3, \ldots$ or $(t_1, k_1), (t_2, k_2), (t_3, k_3), \ldots$

Goals

1. Learn latent dynamics that generated the data.
2. Predict likelihood or label of future events.

Point processes are an established model for the dynamics of such systems.

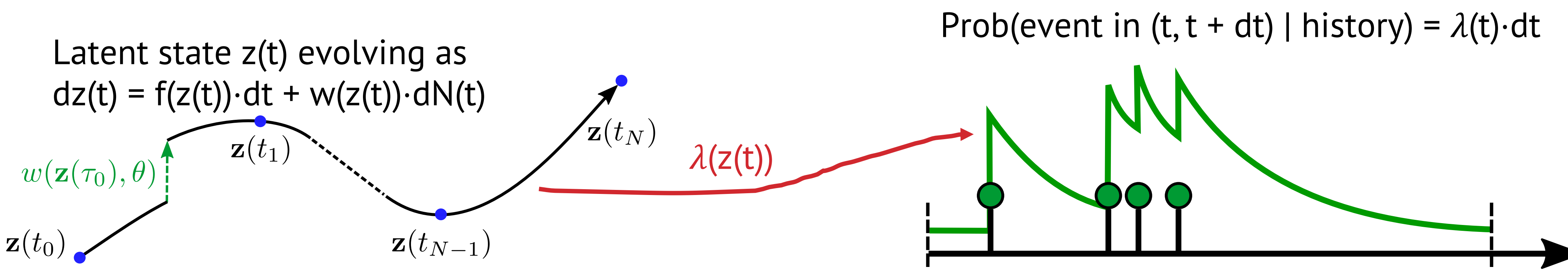


Prob(event in $(t, t + dt)$ | history) = $\lambda(t) \cdot dt$

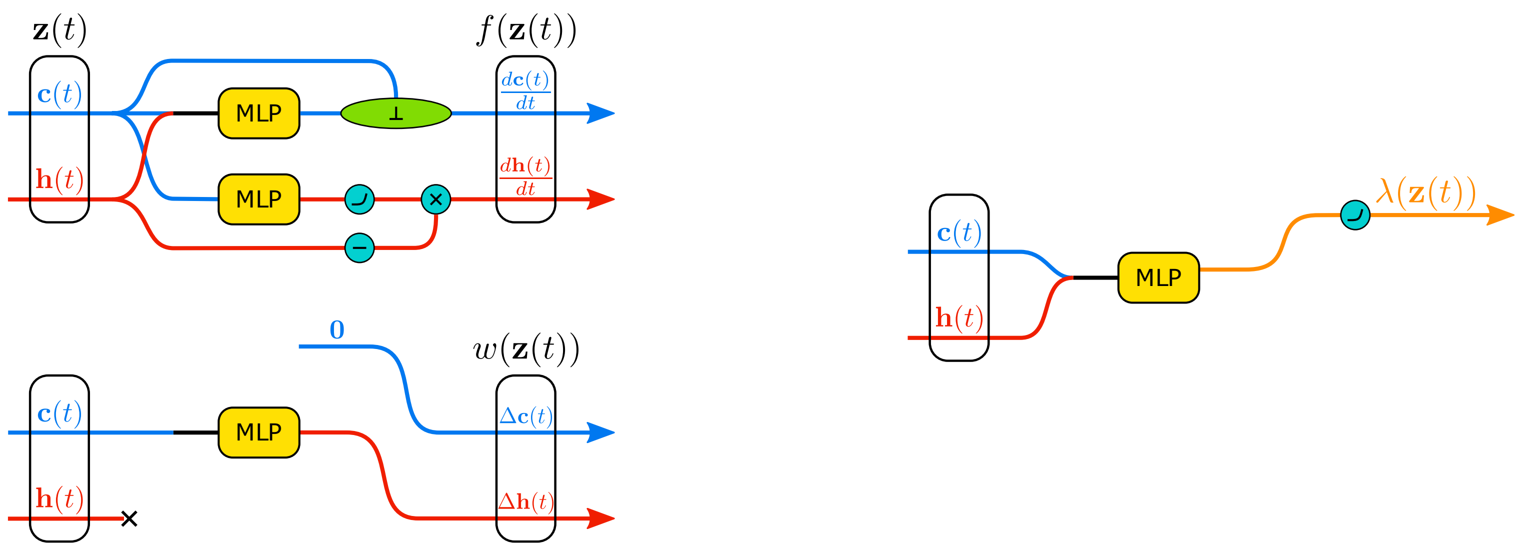
Captures self-exciting, self-inhibiting, delays, background, …

Problem. The functional form of the process must be specified ahead of time.

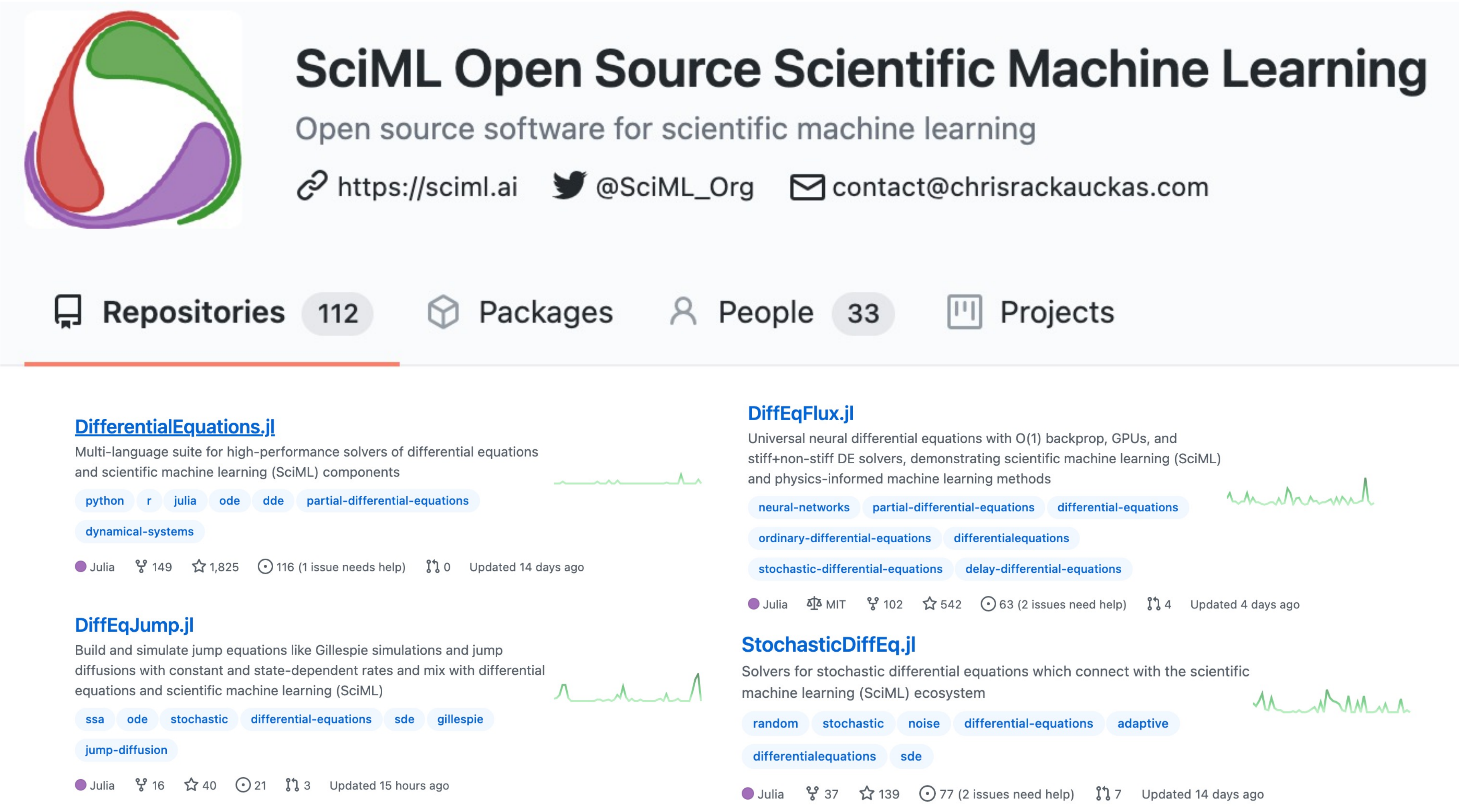
We model the dynamics generally with a latent state.



All of the functions are just parameterized by simple neural networks.
[Based on *Neural Ordinary Differential Equations* by Chen et al., 2018.]

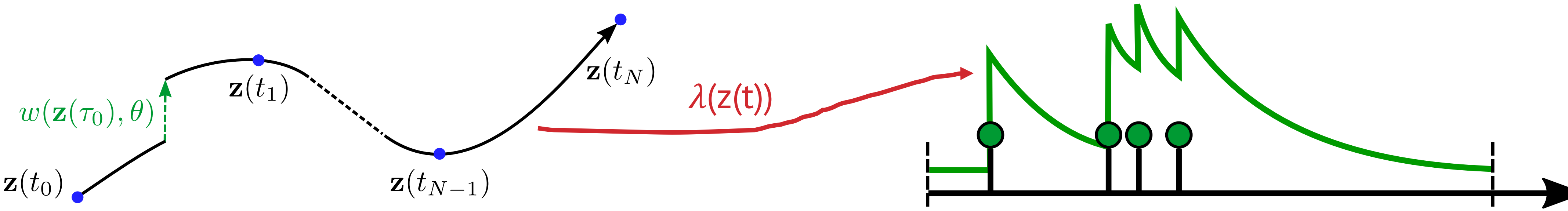


Robust automatic differentiation software now makes learning these models pretty easy.

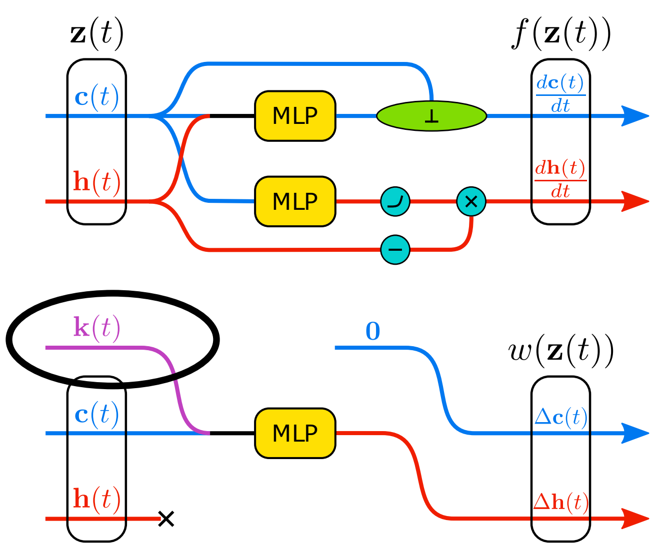


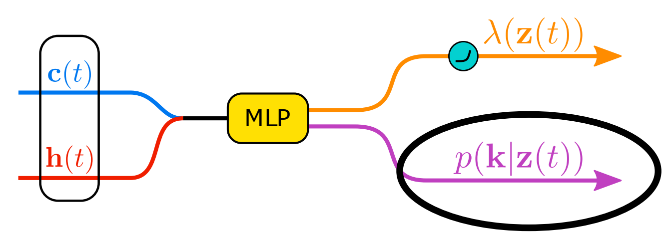
Application. Predicting reasons for ICU visits.

- 650 de-identified ICU patients tracked 2001–2007.
- Data for time of visits and reason for visit (75 total reasons).



When a patient arrives at the ICU, given their history, what is the reason *k* for the visit?

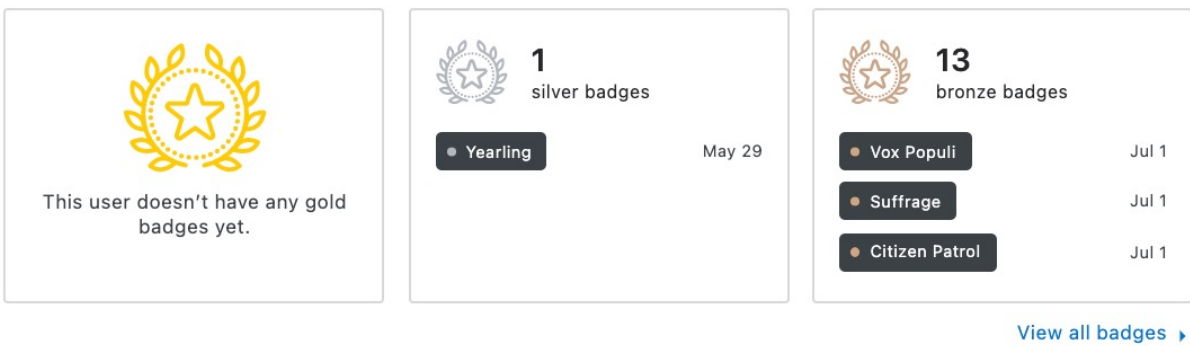


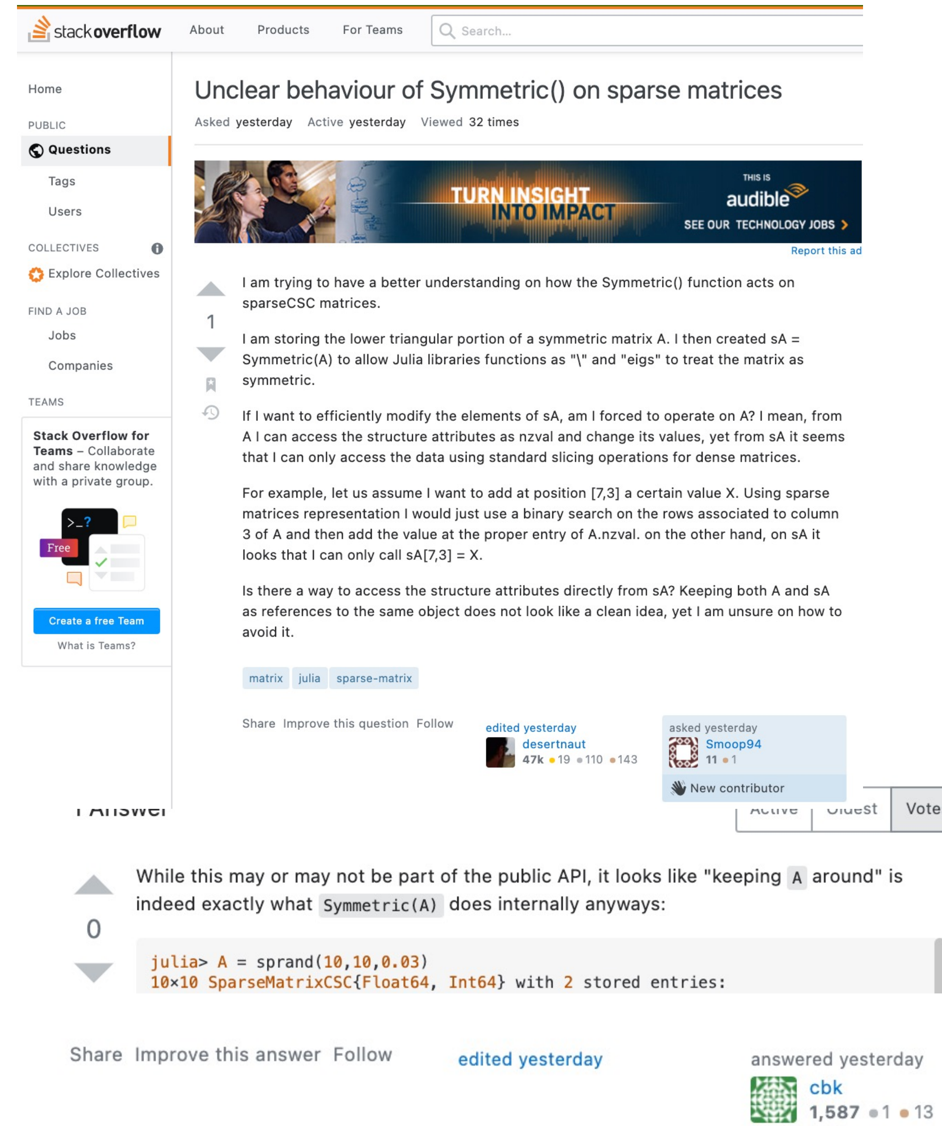


This achieve ~80% accuracy.

Application. Predicting user behavior in social networks.

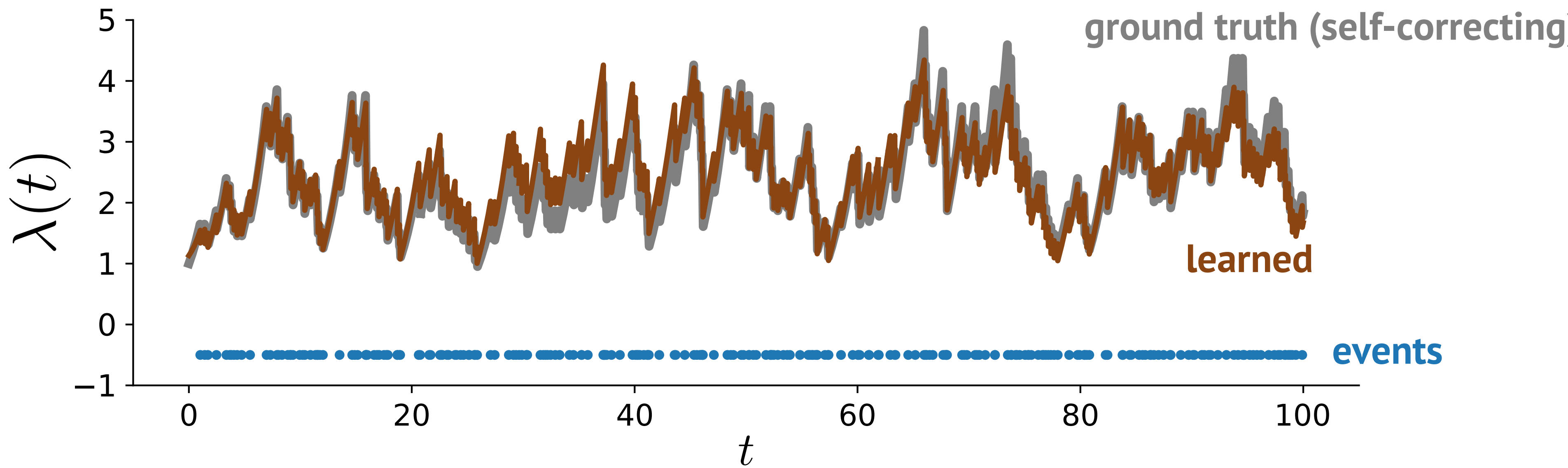
- 6,663 users of Stack Overflow tracked over 2 years.
- Users earn badges for certain activities.
- Data on when badges were earned (22 types).





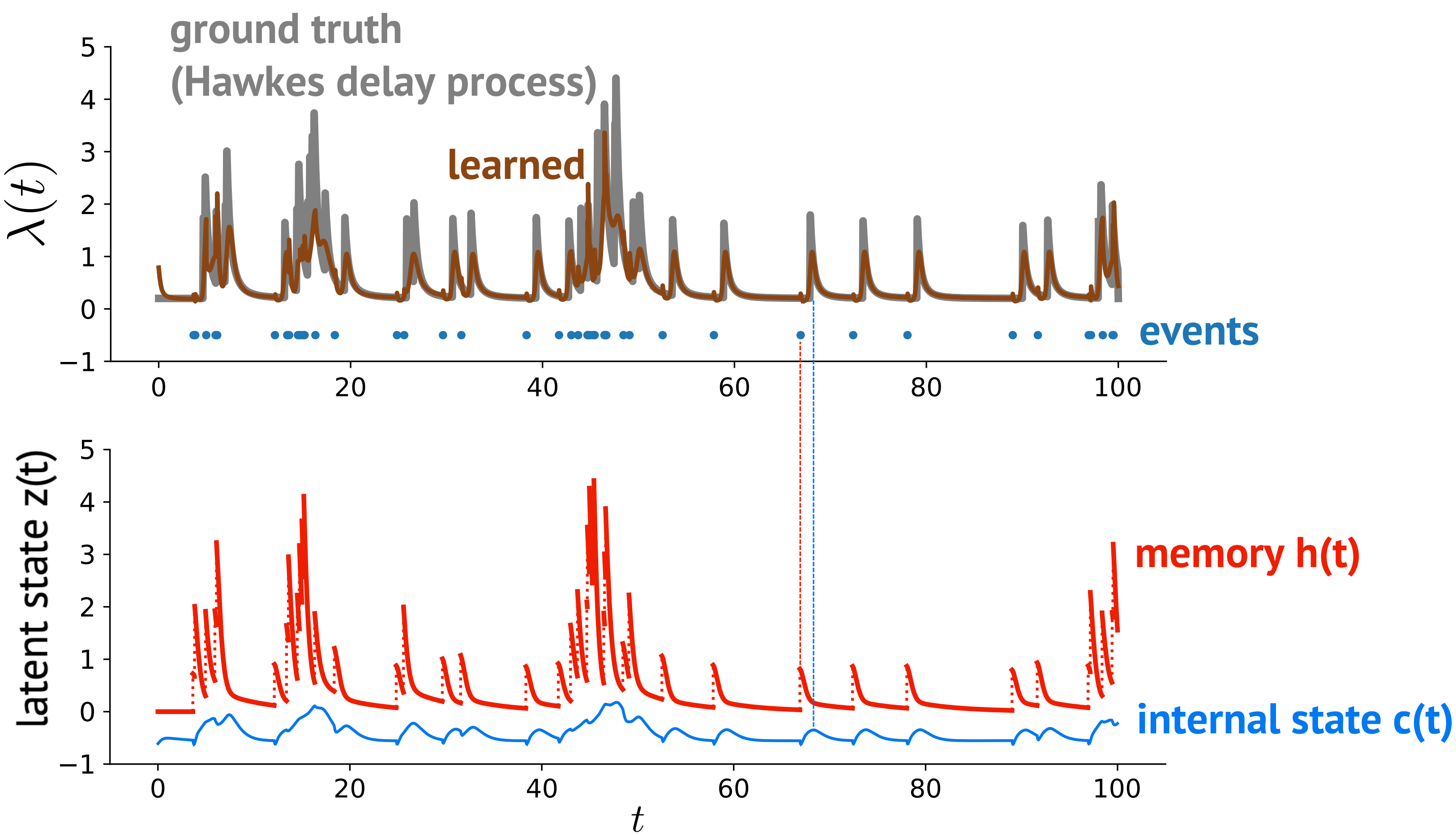
Given a user's history, which badge will they earn next?
We get ~47% accuracy.

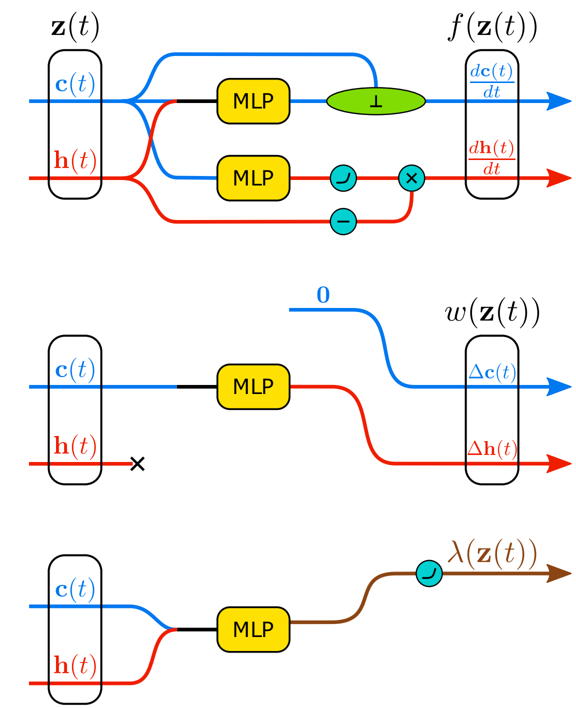
We can capture true dynamics in synthetic data.



- 9% average error in learned dynamics.
- 24% average error with a recurrent neural network (no dynamics).

The latent state can capture complex processes.





- 17% error in learned dynamics.
- 20% error with a recurrent neural network (no dynamics).
- 10% error if given the functional form.

Recap

1. We started with a well-established technique (temporal point processes) and replaced pre-specified functions with general neural networks.

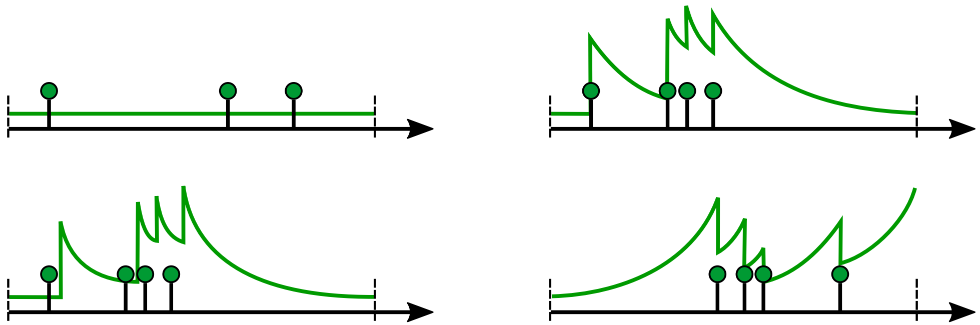
2. This leads to good empirical performance and out-performs general-purpose DL tools.

3. Also get benefits of original technique, such as sampling, simulation.

4. The neural network part also requires some modeling!

5. Robust high-level software can make learning parameters easy.

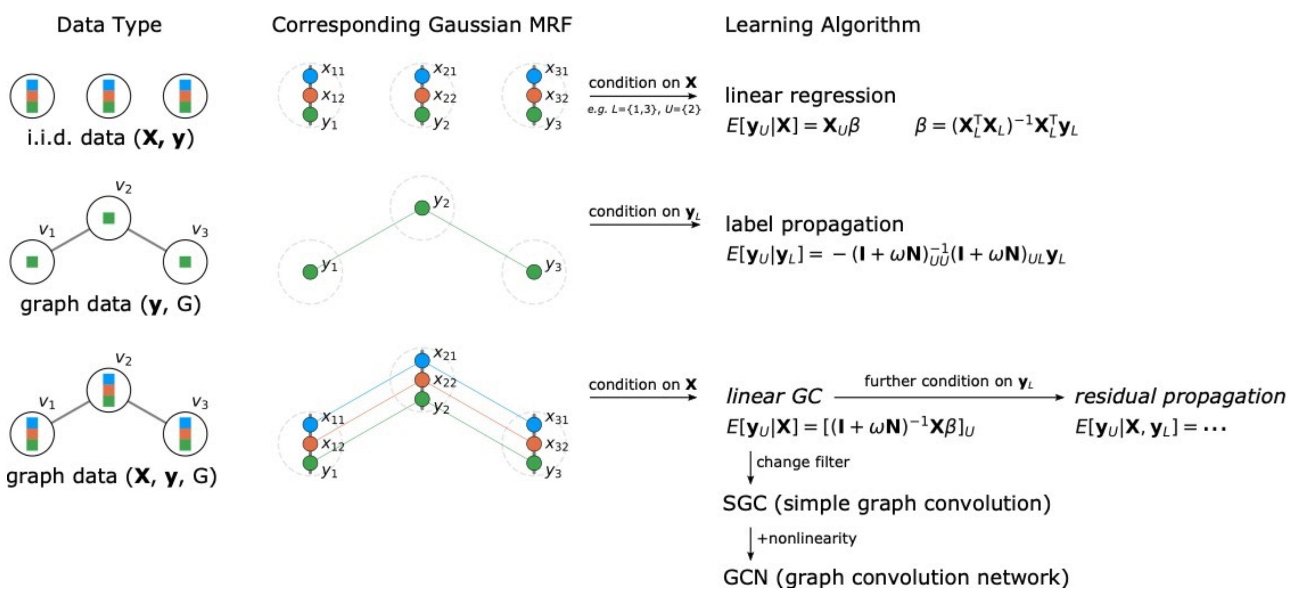
1. An existing simple model augmented with deep learning.



Predict reason for patient visit in ICU.

Neural Jump Stochastic Differential Equations, Jia and Benson, Neurips 2019.

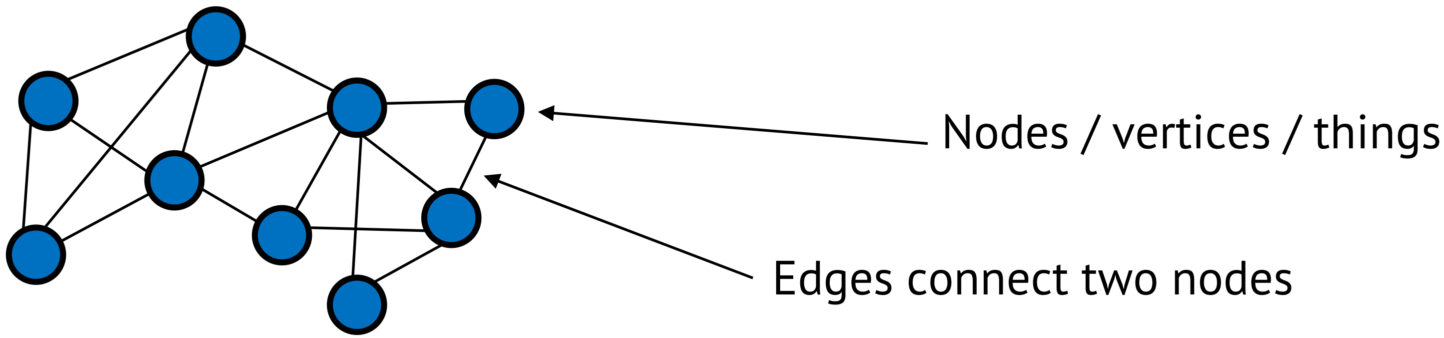
2. A model for understanding an existing deep learning method.



Predict air quality in regions of USA given nearby climate statistics.

A Unifying Generative Model for Graph Learning Algorithms, Jia and Benson, arXiv 2021.

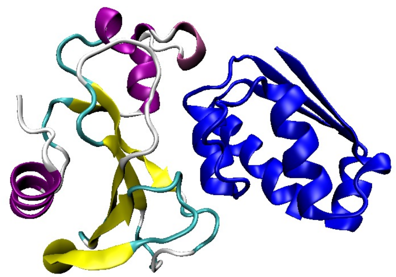
Graphs aka networks are a common abstraction.



Society
nodes are people
edges are friendships

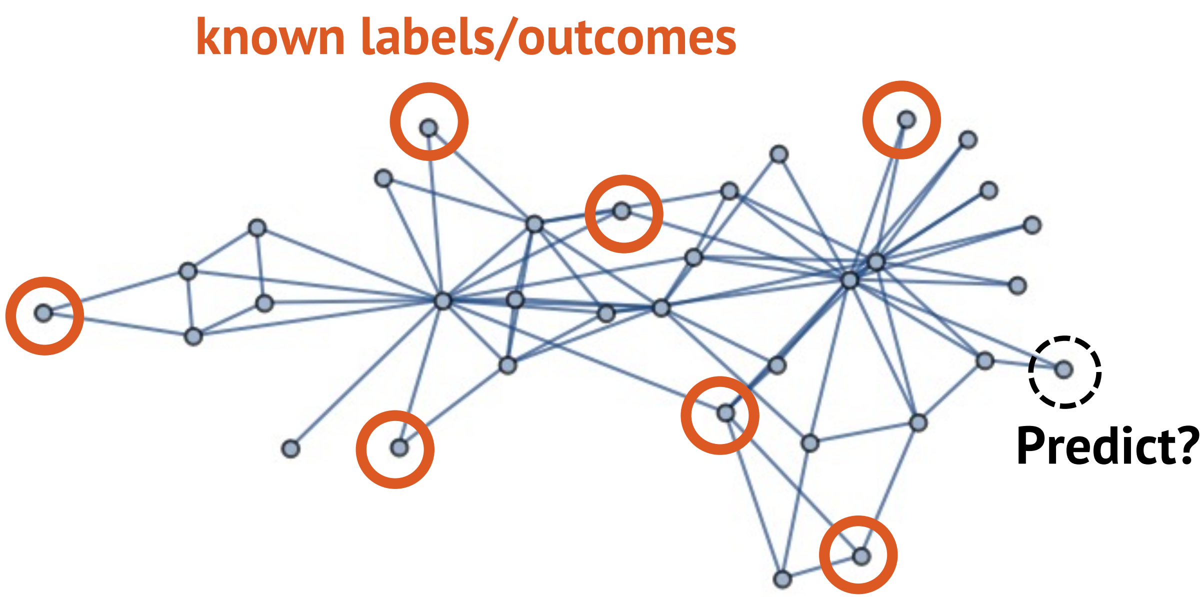
Finance
nodes are accounts
edges are transactions

Drug interactions
nodes are drugs
edge are co-usage by patients



Cell biology
nodes are proteins
edge are interactions

We often want to predict/estimate/construct/forecast attributes/labels/outcomes/clusters on nodes.



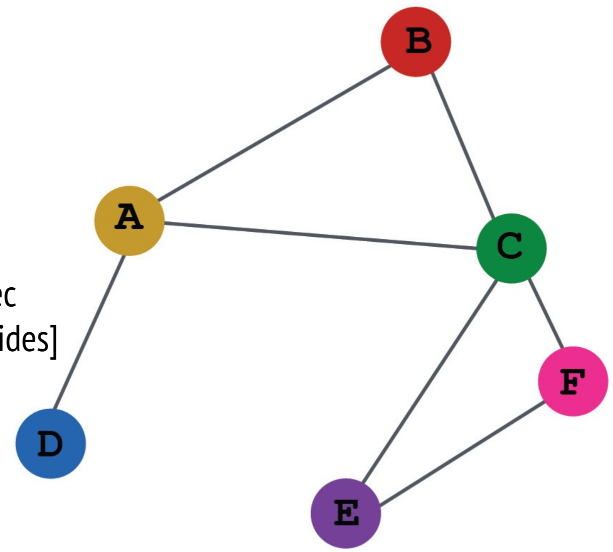
- Gender in social networks
  [Peel 17; Altenburger–Ugander 18]
- Bad actors in financial transaction graphs
  [Weber+ 18, 19; Pareja+ 20]
- Protein function in PPI networks
  [Hamilton+ 17]

- Might have additional info on nodes (features)
  user interests, transaction history, gene sets in proteins

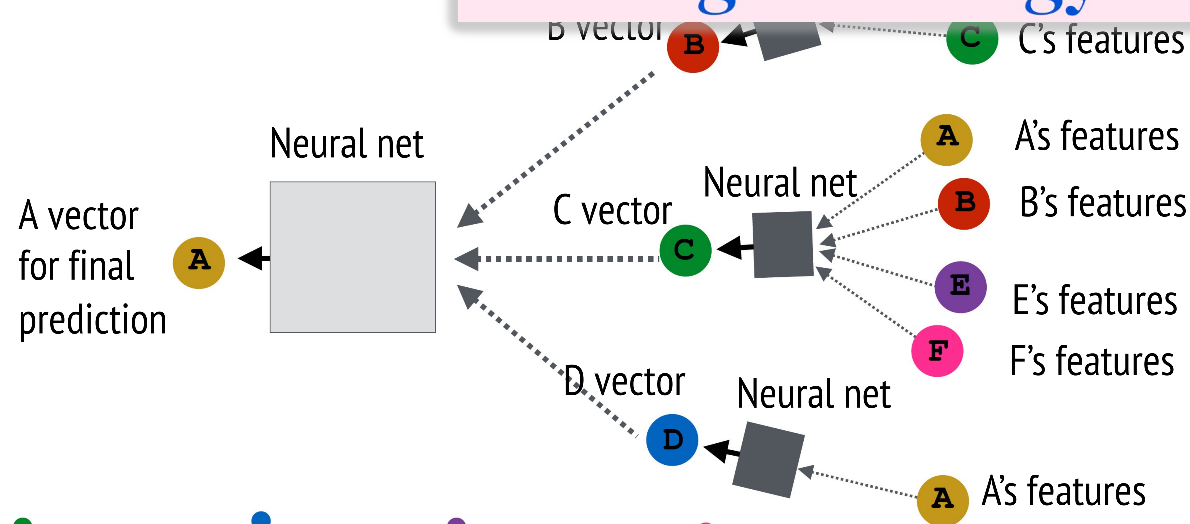
- Graph-based semi-supervised learning, clustering, node prediction, relational learning, collective classification, community detection, ...

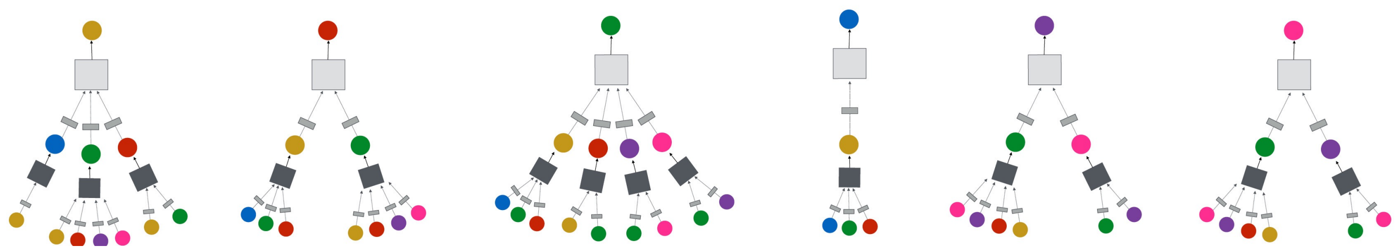
Graph neural networks (GNNs) are … method for this problem, based on …


BLOG POST
RESEARCH
30 NOV 2020
AlphaFold: a solution to a 50-year-old grand challenge in biology



[From Leskovec 224W 2021 slides]





- **BIG** optimization problem trained with labeled nodes and automatic differentiation.
- **DIFFICULT** to implement, interpret, and scale to large datasets.

Is there a statistical model that gives rise to GNNs?

Do we need the complexity of the neural network parts?

Get Started Updates Datasets Leaderboards Paper Team Github

Leaderboard for ogbn-products

The classification accuracy on the test and validation sets. The higher, the better.

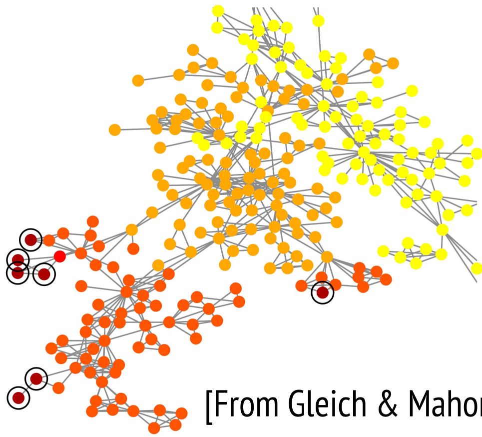
Package: >=1.1.1

Rank	Method	Test Accuracy	Validation Accuracy	Contact	References	#Params	Hardware	Date
1	**MLP + C&S**	0.8418 ± 0.0007	0.9147 ± 0.0009	Horace He (Cornell)	Paper, Code	96,247	GeForce RTX 2080 (11GB GPU)	Oct 27, 2020
2	**Linear + C&S**	0.8301 ± 0.0001	0.9134 ± 0.0001	Horace He (Cornell)	Paper, Code	10,763	GeForce RTX 2080 (11GB GPU)	Oct 27, 2020
3	**UniMP**	0.8256 ± 0.0031	0.9308 ± 0.0017	Yunsheng Shi (PGL team)	Paper, Code	1,475,605	Tesla V100 (32GB)	Sep 8, 2020
4	**Plain Linear + C&S**	0.8254 ± 0.0003	0.9103 ± 0.0001	Horace He (Cornell)	Paper, Code	4,747	GeForce RTX 2080 (11GB GPU)	Oct 27, 2020
5	**DeeperGCN+FLAG**	0.8193 ± 0.0031	0.9221 ± 0.0037	Kezhi Kong	Paper, Code	253,743	NVIDIA Tesla V100 (32GB GPU)	Oct 20, 2020

Combining Label Propagation and Simple Models Out-performs Graph Neural Networks.
Q. Huang et al., ICLR 2021.

Label propagation (LP) is a classical approach that is simple and is based on inference in a statistical model.

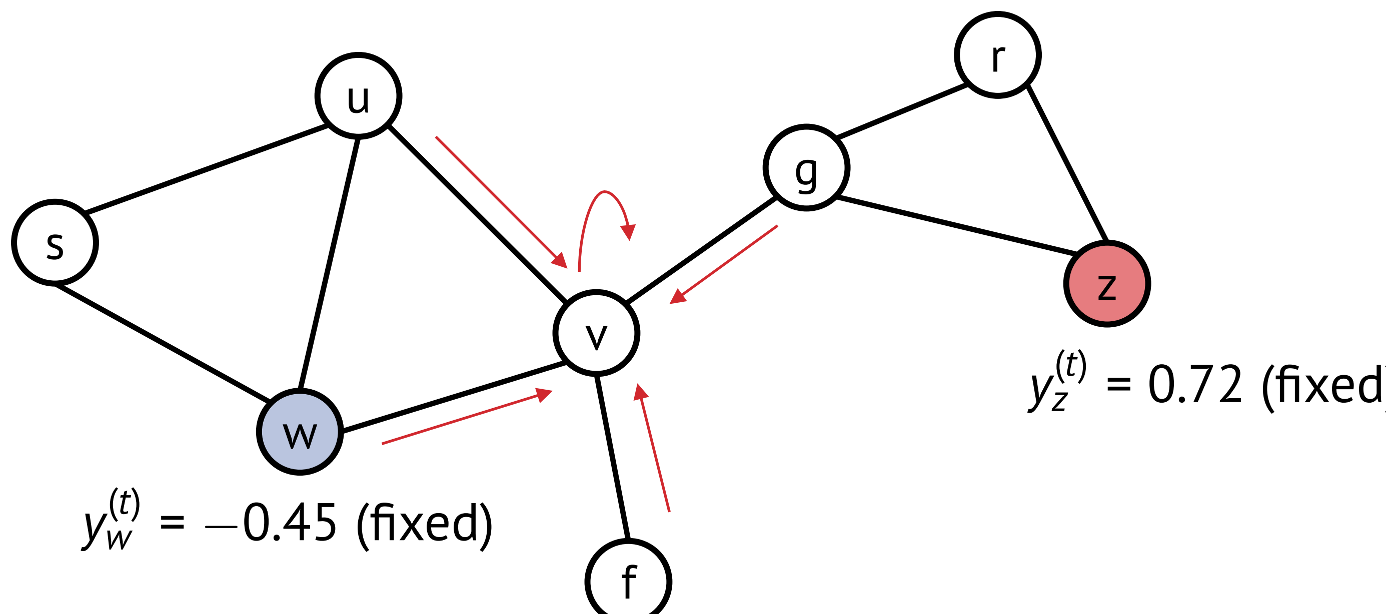
[Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions, Zhu, Ghahramani, and Lafferty, 2003]



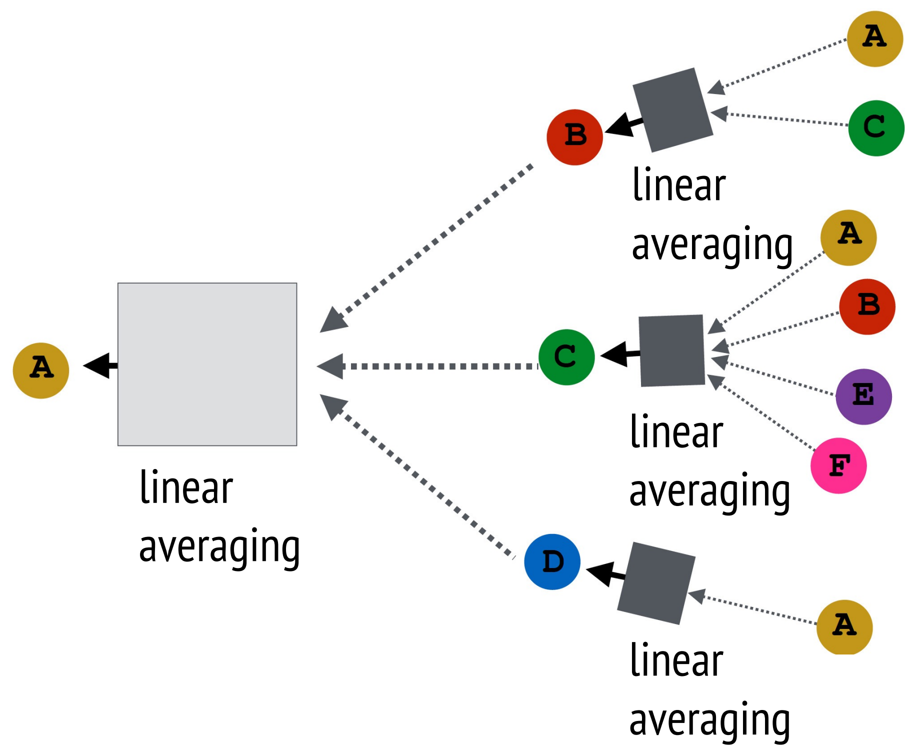
[From Gleich & Mahoney 2015]

- **Idea** connected nodes have similar labels.
- Works because of homophily [McPherson+ 01] aka assortativity [Newman 02]
- Doesn't use additional info/features, though
- fast and interpretable

- LP algorithm is just neighbor averaging.
- At convergence, everyone is roughly the average over their neighbors $\longrightarrow$ smooth!



Our hypothesis was that GNNs are smoothing or averaging the features, similar to LP.



Linear graph convolution (LGC).

1. Run LP on each feature $\longrightarrow$ smoothed features.
2. Ordinary least squares on these preprocessed, smoothed features.

We generalized the LP statistical generative model to a model for graphs with node features and labels.

Random real-valued attribute vectors $\mathbf{a}_\mathrm{u} = [\boldsymbol{x}_\mathrm{u}; y_\mathrm{u}]$ on each node u.

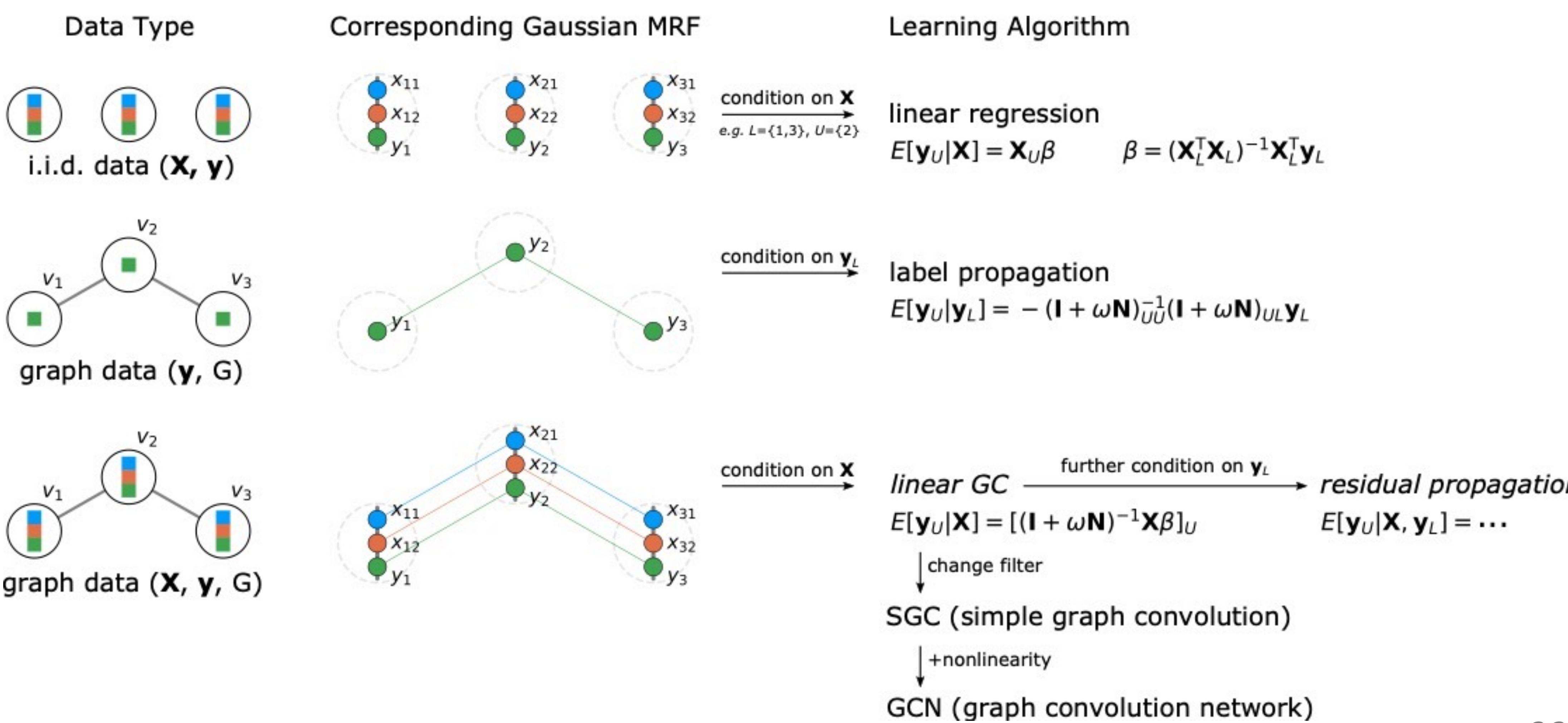
$$\phi(\boldsymbol{A}|\boldsymbol{H},\boldsymbol{h}) = \frac{1}{2}\sum_{u=1}^{n} \boldsymbol{a}_u^\mathsf{T}\boldsymbol{H}\boldsymbol{a}_u + \frac{1}{2}\sum_{i=1}^{p+1} h_i \boldsymbol{A}_i^\mathsf{T}\boldsymbol{N}\boldsymbol{A}_i, \quad \boldsymbol{H} \in \mathbb{R}^{(p+1)\times(p+1)} \text{ spd}, \quad \mathbf{0} \leq \boldsymbol{h} \in \mathbb{R}^{(p+1)}$$

$$\rho(\boldsymbol{A} = \boldsymbol{A}|\boldsymbol{H},\boldsymbol{h}) = \frac{e^{-\phi(\boldsymbol{A}|\boldsymbol{H},\boldsymbol{h})}}{\int d\boldsymbol{A}'\; e^{-\phi(\boldsymbol{A}'|\boldsymbol{H},\boldsymbol{h})}}$$

Key ingredients

1. The label is correlated with the features.
2. Connected nodes are more likely to have the same label.
3. Connected nodes are more likely to have similar attributes.

We developed a random model for attributes on nodes, where statistical inference leads to GNN/LP algorithms.



```julia
function LGC_params(S, X, y, L; α=0.9, num_iters=10)
    X_smooth = copy(X)
    for _ in 1:num_iters
        X_smooth = (1 - α) * X + α * S * X_smooth
    end
    return X_smooth, X_smooth[L, :] \ y[L]
end

function residual_prop(S, y, ȳ, U; α=0.9, num_iters=10)
    r = y - ȳ
    r[U] = 0
    for _ in 1:num_iters
        z = S * r
        r[U] = α * z[U]
    end
    return r
end

function LGC_RP_prediction(
    S,  # normalized adjacency D^{-1/2} A D^{-1/2}
    X,  # n x d feature matrix for n nodes
    U,  # indices of unlabeled nodes
    L   # indices of labeled nodes
    y,  # n x 1 label vector (zero on y[U])
    )
    X_smooth, β̂ = LGC_params(S, X, y, L)
    ȳ = X_smooth * β̂
    r = residual_prop(S, y, ȳ, L)
    return ȳ[U] + c[U]
end
```

Dataset	Outcome	LP	LR	LGC	GCN	LGC/RP
Climate	landT	0.89	0.81	0.81	**0.91**	0.90
	precipitation	0.89	0.59	0.61	0.79	**0.89**
	pm2.5	0.96	0.21	0.27	0.78	**0.96**
U.S. elections	education	0.31	**0.71**	**0.71**	0.47	**0.71**
	unemployment	0.47	0.34	0.39	0.45	**0.54**
	election	0.52	0.42	0.49	0.52	**0.64**

Recap

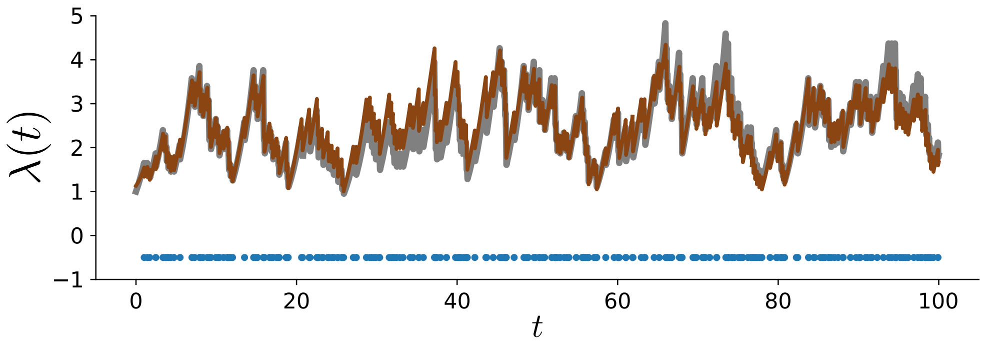
1. We started with a recent DL technique (GNNs) and simplified it to a linear algorithm based on a statistical generative model.

2. This leads to good empirical performance and out-performs general-purpose DL tools.

3. Also get benefits of model, such as sampling, simulation.

Temporal and relational machine learning for biostatistical and other scientific applications

THANKS!
Austin R. Benson
http://cs.cornell.edu/~arb
@austinbenson
arb@cs.cornell.edu



Supported by ARO MURI, ARO Award W911NF19- 1-0057, NSF Award DMS-1830274, JP Morgan Chase & Co, and Cornell University.