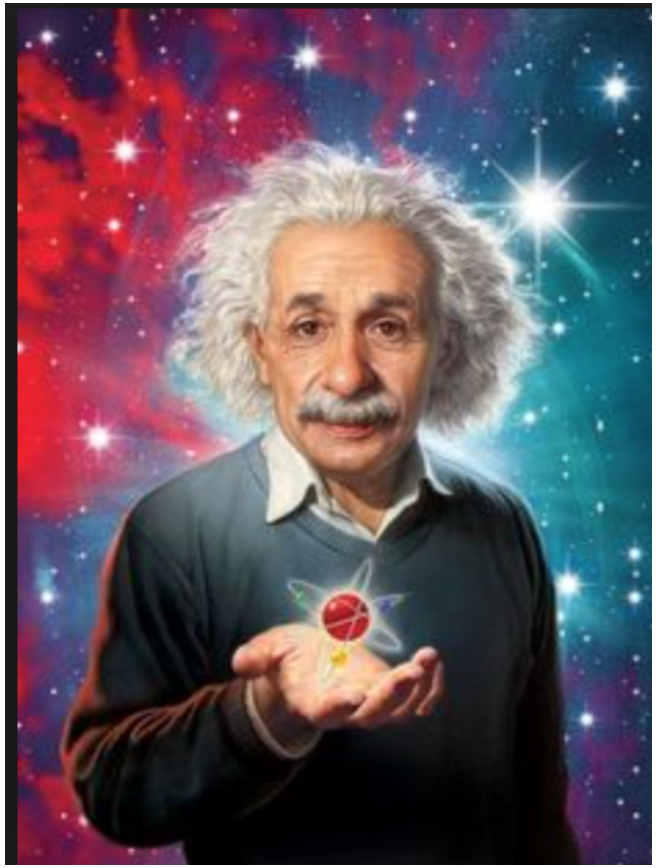
Conquering the Complexity of Time: Machine Learning for Big Time Series Data

Yan Liu

Computer Science Department
University of Southern California

Mini-Workshop on Theoretical Foundations of Cyber-Physical Systems

April 14, 2017

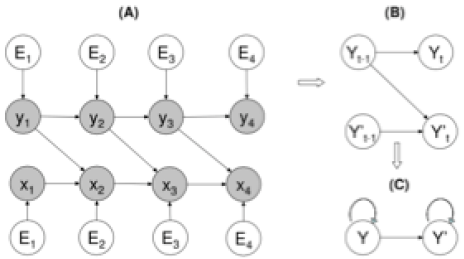


A human being is a part of a whole, called by us "universe", a part limited in time and space.

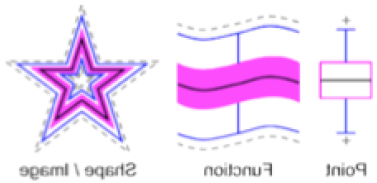
Research Synopsis: Machine Learning Models

Developing scalable and effective solutions by leveraging recent progresses across disciplines

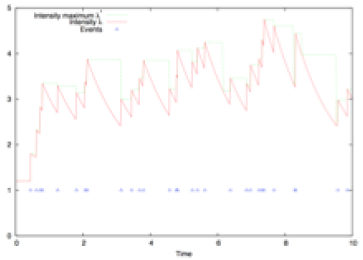
Granger graphical models



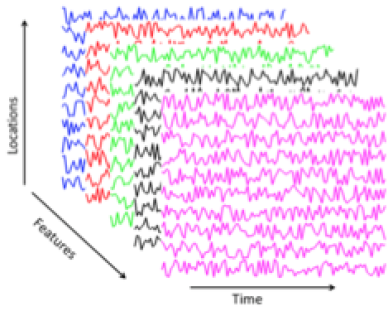
Functional data analysis



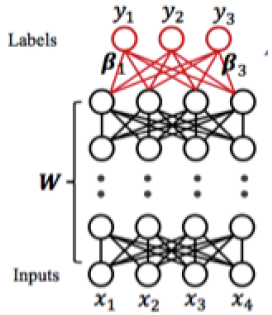
Point process model



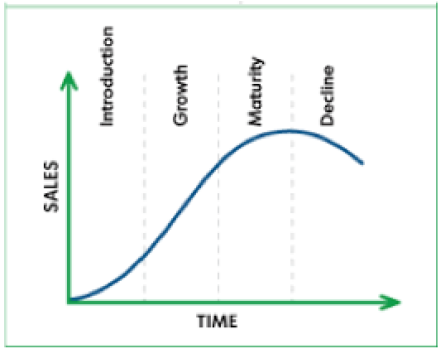
Low rank tensor analysis



Deep neural networks



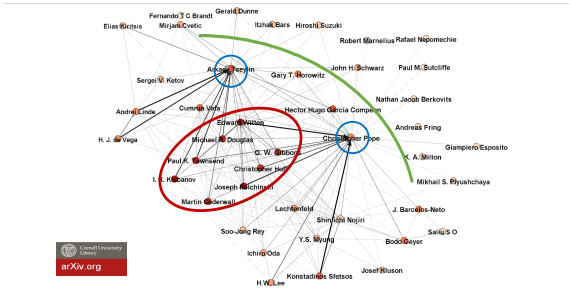
Life cycle model



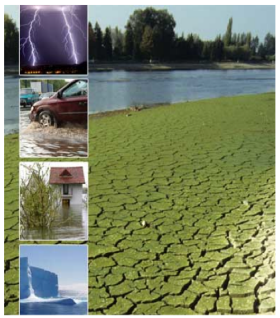
Research Synopsis: Applications

- Doctor AI - Healthcare Analytics

- Social Network Analysis



- Computational sustainability



Google Alpha Go Beats Human

After AlphaGo, what's next for AI?

Googles DeepMind AI group unveils health care ambitions

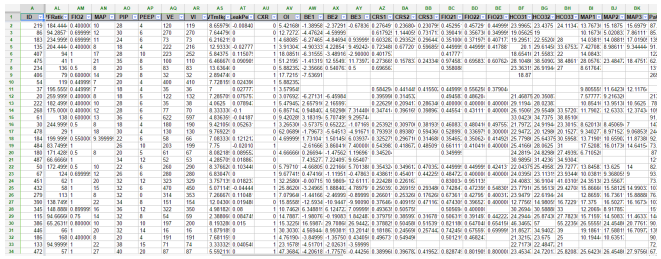
http://www.theverge.com/2016/3/14/11219258/google-deepmind-alphago-go-challenge-ai-future

Properties of Health Care Data

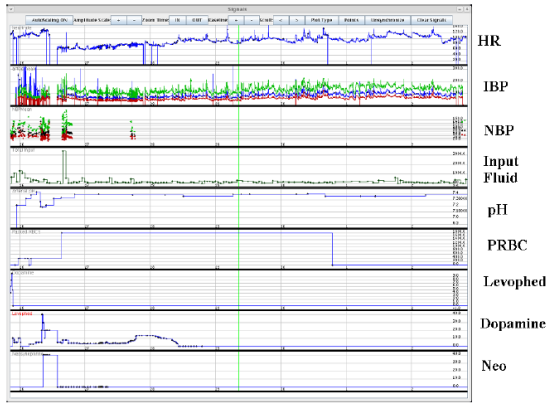
How are health care data different from the data from existing applications of deep learning?

- Privacy, privacy!
- Heterogeneity
- Lots lots of missing data
- Big small data
- *Worst of all: doctors do not believe anything they cannot understand no matter how cool and how deep they are!!*

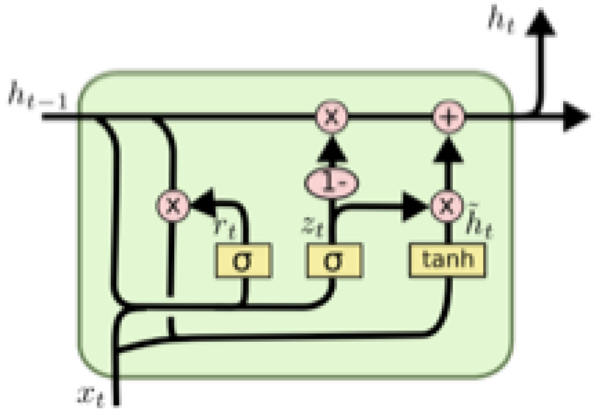
Example 1:



Example 2:



Recurrent Neural Networks for Time Series Data

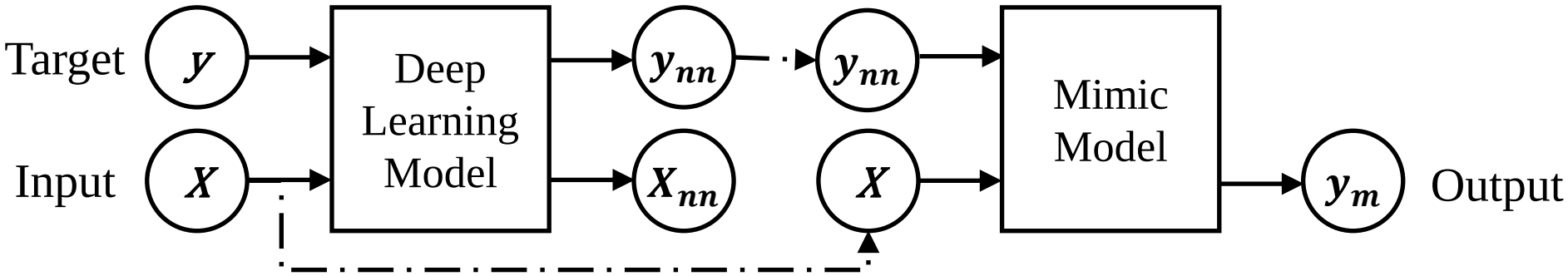


$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Our Contributions [KDD 2015, AMIA 2015, 2016, arXiv 2016, ICLR 2017]

- **Multi-modal Deep Neural Networks**
- **Handling Missing Data in DNN**
- **Big Small Data Solution via DNN**
- **Interpretation of DNN**

Explainable Artificial Intelligence: Mimic learning Framework



Main idea:

- Use Gradient Boosting Trees (GBT) to mimic the performance of deep learning models

Benefits:

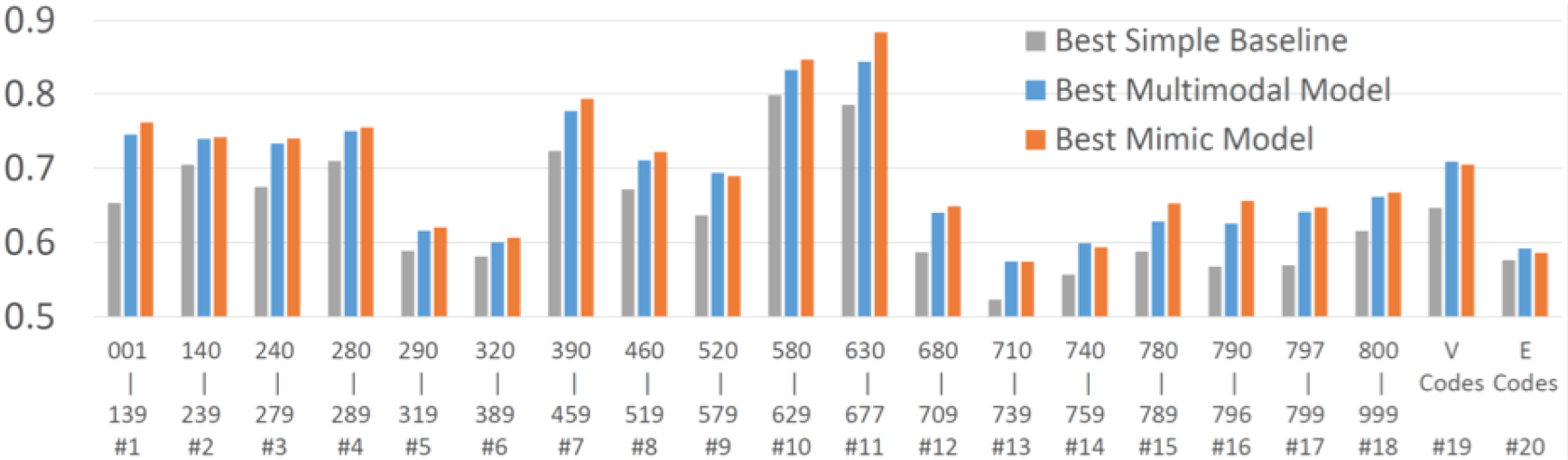
- Good performance from complex deep networks – Mimic model keeps it
- Easy overfitting in original decision tree methods – Mimic model avoids it
- Interpretations are hard to get in original models – Mimic model provides it

Experiment Result

Prediction on patients with respiratory disease

	Mortality		Ventilator Free Days	
	AUROC	AUPRC	AUROC	AUPRC
Simple Model	0.7196	0.4171	0.7592	0.8142
Deep Model	0.7813	**0.4874**	**0.7896**	**0.8397**
Mimic Model	**0.7898**	0.4766	0.7889	0.8324

AUROC across 20 ICD-9 diagnosis prediction tasks

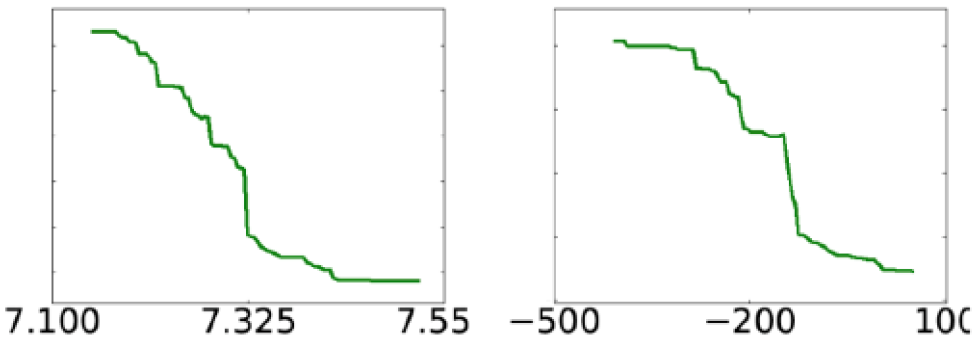


Experiment Result

Important Features for ICD-9 diagnosis prediction tasks

Blood System Diagnosis	• Hemoglobin • Platelet count • Red cell distribution width
Respiratory System Diagnosis	• Mean corpuscular hemoglobin concentration • Partial pressure of oxygen

How features are evaluated in our model



x-axis: input value; y-axis: mortality risk

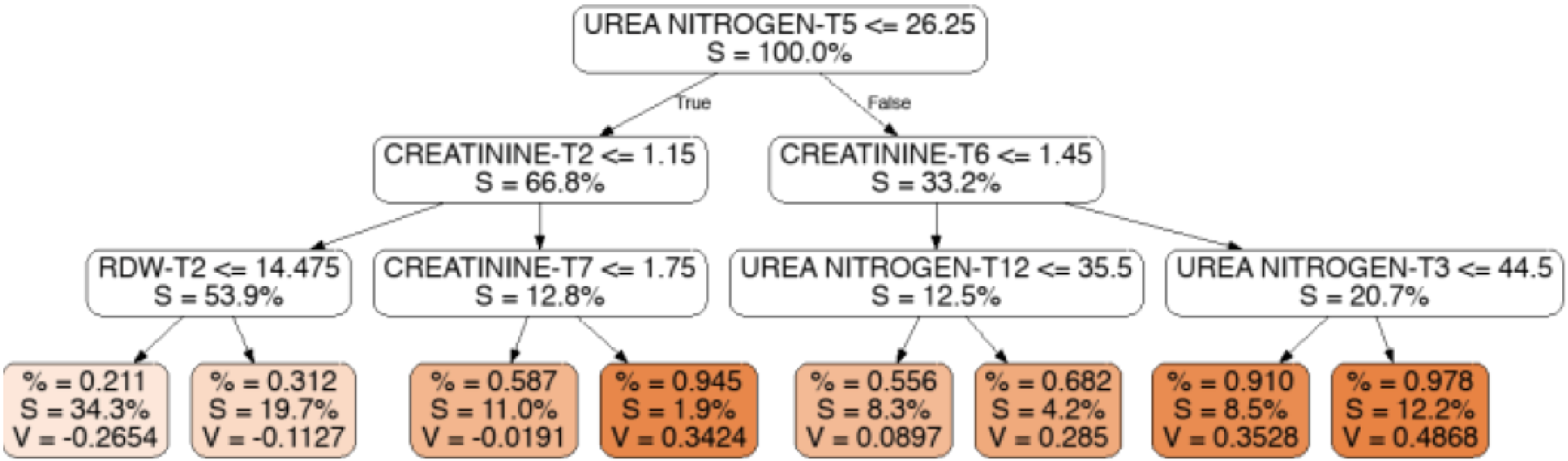
Left: PH value in blood

- A very narrow normal range around 7.35-7.45.

Right: change of PaO2/FiO2 ratio

- Normal range: 400-500 mmHg;
- < 200: necessary for the diagnosis of respiratory distress syndrome.

Experiment Result



% and color: class distribution; S: # of samples; V: prediction value

Genitourinary system diagnosis

- Urea nitrogen
- Creatinine

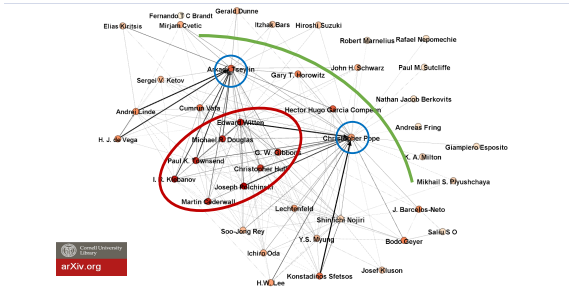
Ventilator free days

- Lung injury score
- Oxygenation index
- Change of PaO2/FiO2 ratio

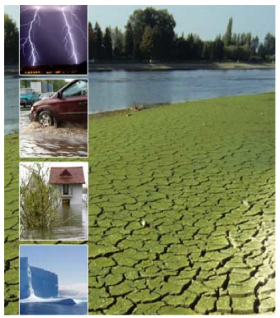
Research Synopsis: Applications

- Doctor AI - Healthcare Analytics

- Social Network Analysis

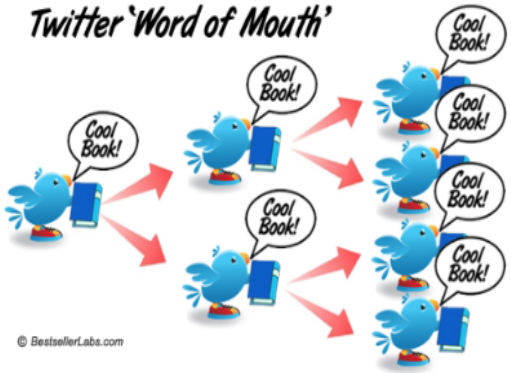


- Computational sustainability



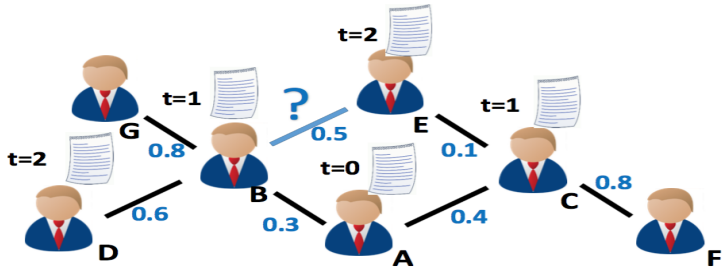
Diffusion Analysis

- Adoption of innovation: new treatment, new technology
- Marketing: word of mouth effect, viral marketing
- Public opinion surveillance: detection of rumors
- Media analysis: modeling news dynamics



Network Inference

Network Inference: inferring the latent diffusion network from observed cascades

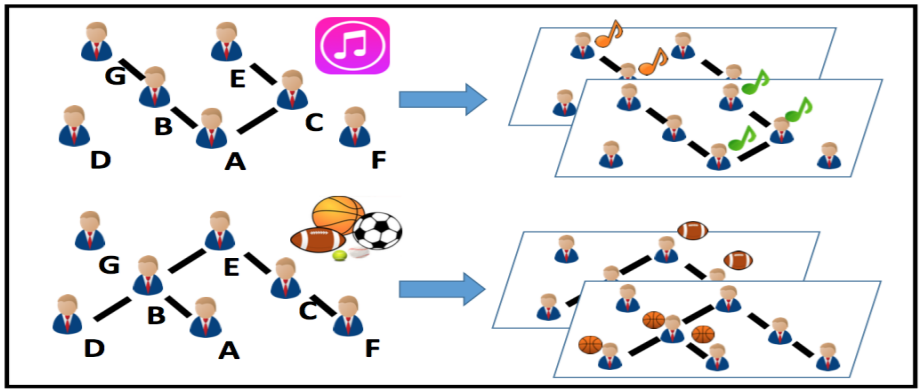


Our contribution [ICML 2015, NIPS 2016, WSDM submission]

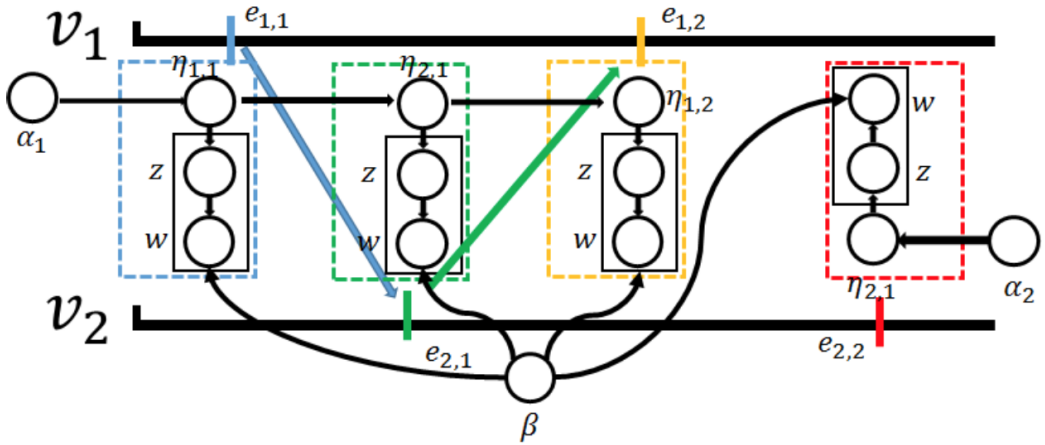
Network inference with **incomplete** cascades

Network inference with **content** information

Joint inference of **multiple** networks



Hawkes-Topic Models: Joint Inference of Diffusion Networks and Topics



1. Generate all the events and the event times via the Multivariate Hawkes Process
2. For each topic k: draw $\beta_k \sim Dir(\alpha)$.
3. For each event e of node v:
   1. If e is a spontaneous event: $\eta_e \sim N(\alpha_v, \sigma^2 I)$. Otherwise $\eta_e \sim N(\eta_{\text{parent[e]}}, \sigma^2 I)$.
   2. For each word n: $z_{e,n} \sim \text{Discrete}(\pi(\eta_e)), w_{e,n} \sim \text{Discrete}(\beta_{z_{e,n}})$.

Experiment Results: EventRegistry

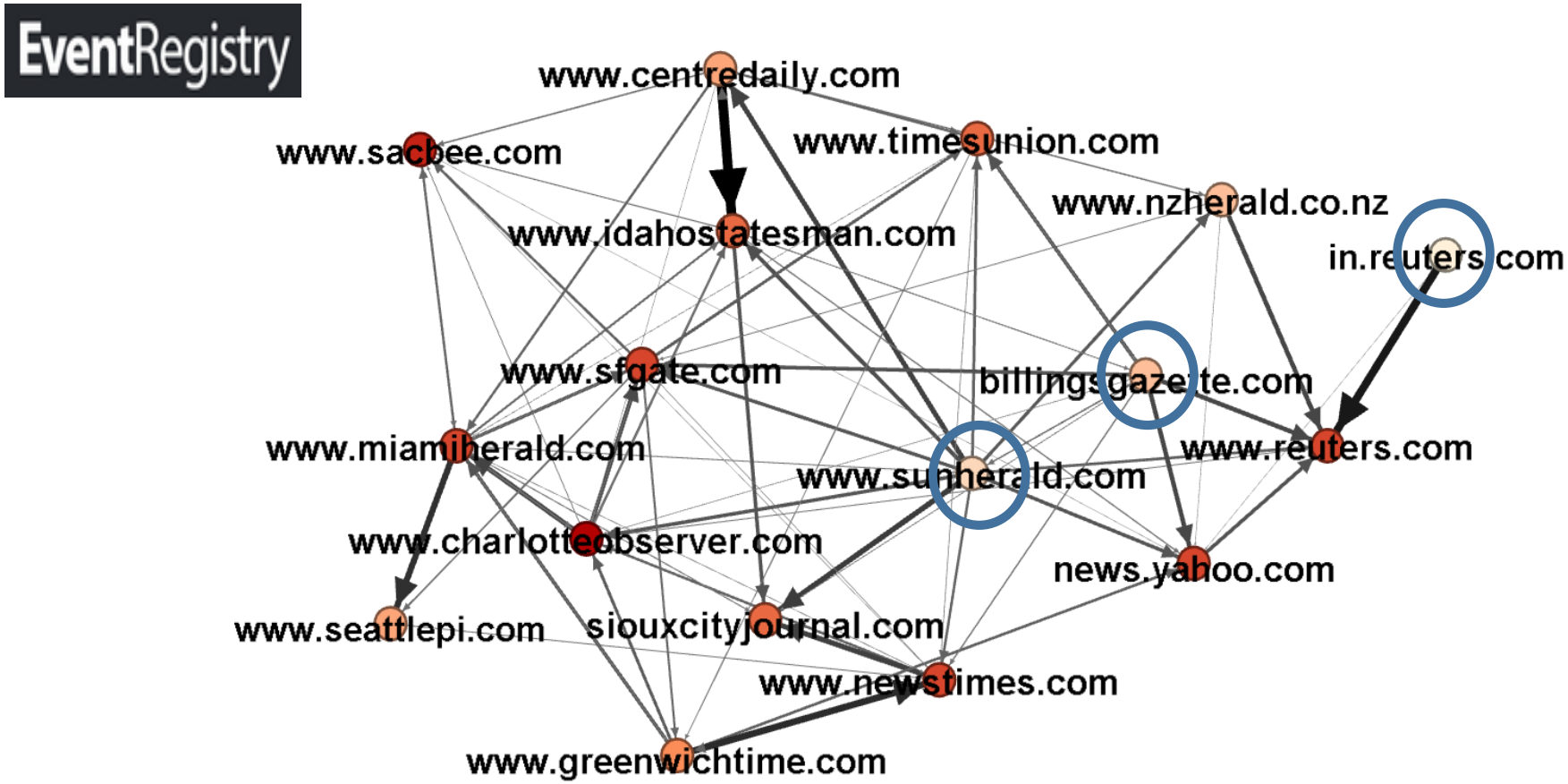
Network Inference accuracy: 10% improvement

	Hawkes	Hawkes-LDA	Hawkes-CTM	HTM
Component 1	0.622	0.669	0.673	**0.697**
Component 2	0.670	0.704	0.716	**0.730**
Component 3	0.666	0.665	0.669	**0.700**

Topic modeling accuracy:

	LDA	CTM	HTM
Component 1	-42945	-42458	**-42325**
Component 2	-22558	-22181	-22164
Component 3	-17574	-17574	-17571

Experiment Results: EventRegistry

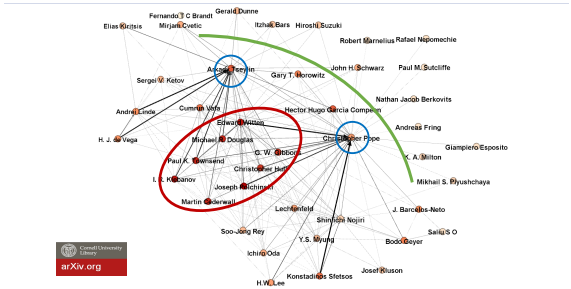


Early bird report agency: sunherald.com, miamiherald.com and in.reuters.com
News gathering and re-distribution agency: www.reuters.com

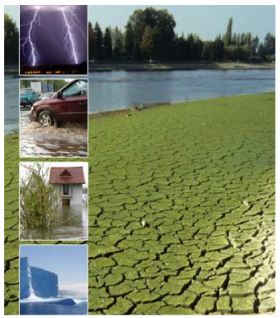
Research Synopsis: Applications

- Doctor AI - Healthcare Analytics

- Social Network Analysis



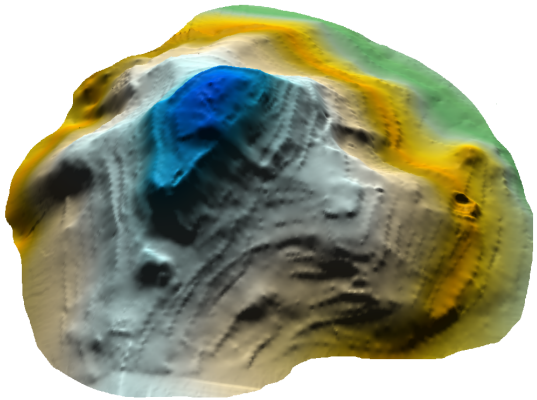
- Computational sustainability



Spatiotemporal Data Analysis [NIPS 2104 Spotlight Presentation, ICML 2015, ICML 2016]

Two key principles in designing spatial-temporal models

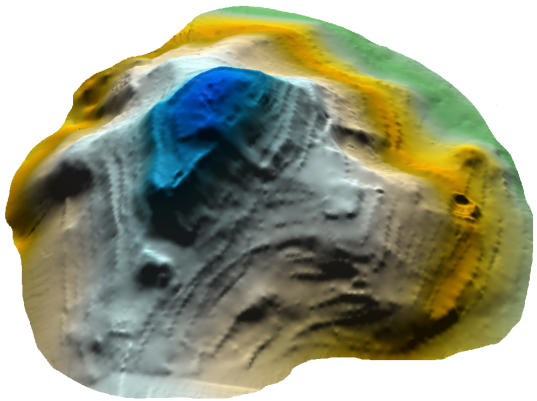
Local smoothness: features in the same neighborhood share similar value



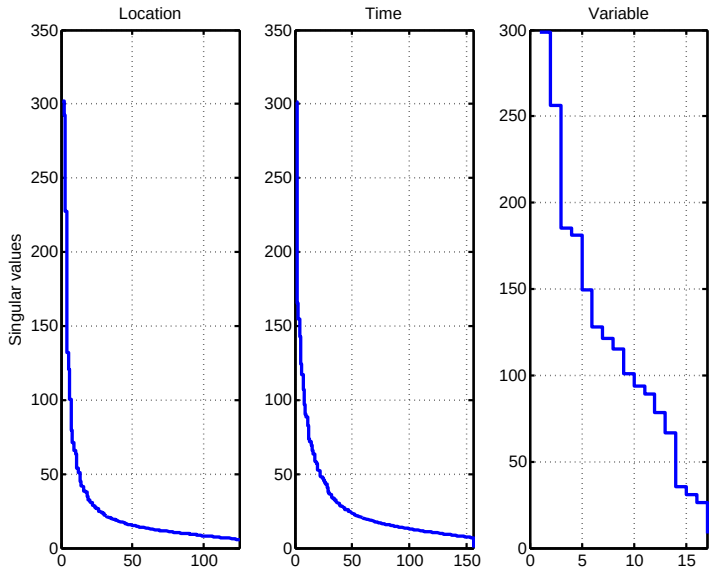
Spatiotemporal Data Analysis [NIPS 2104 Spotlight Presentation, ICML 2015, ICML 2016]

Two key principles in designing spatial-temporal models

Local smoothness: features in the same neighborhood share similar value

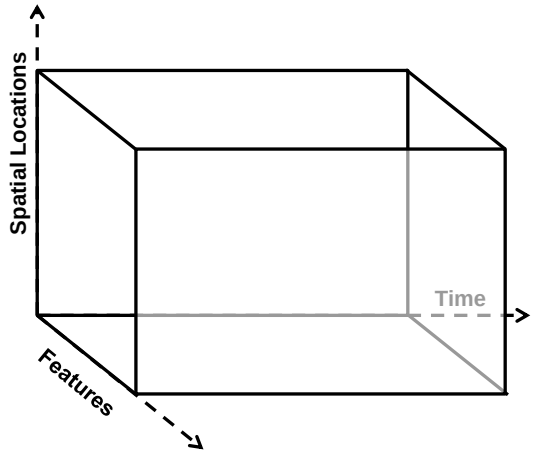


Global latent structure: the data lie on a lower dimensional latent structure



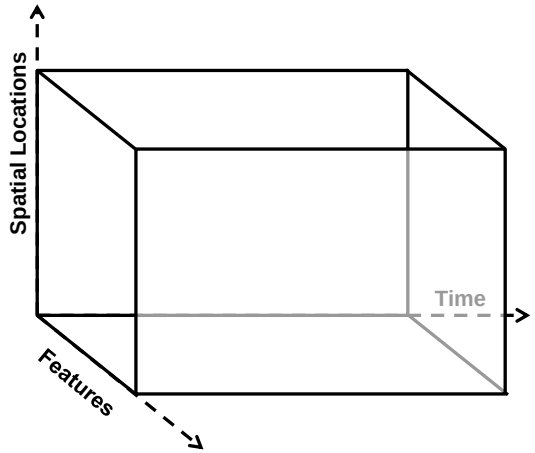
Tensor Representation for Spatial Temporal Data

Spatial temporal data can naturally be represented as tensors



Tensor Representation for Spatial Temporal Data

Spatial temporal data can naturally be represented as tensors



Simple solutions based on tensor formulation:

- Local smoothness can be achieved by Laplacian regularizer
- Global latent structures can be achieved by low-rank constraint

Application I: Cokriging

Task description Jointly predicting the value of multiple features at unknown locations by taking advantage of the observations from known locations.

Application I: Cokriging

Task description Jointly predicting the value of multiple features at unknown locations by taking advantage of the observations from known locations.

Given the complete data $\mathcal{X} \in \mathbb{R}^{P \times T \times M}$ and partial observations at a subset of locations indexed by $\Omega \subset \{1, \ldots, P\}$, we need to estimate $\mathcal{W} \in \mathbb{R}^{P \times T \times M}$ so that $\mathcal{W}_\Omega = \mathcal{X}_\Omega$.

Application I: Cokriging

Task description Jointly predicting the value of multiple features at unknown locations by taking advantage of the observations from known locations.

Given the complete data $\mathcal{X} \in \mathbb{R}^{P \times T \times M}$ and partial observations at a subset of locations indexed by $\Omega \subset \{1, \ldots, P\}$, we need to estimate $\mathcal{W} \in \mathbb{R}^{P \times T \times M}$ so that $\mathcal{W}_\Omega = \mathcal{X}_\Omega$.

Our formulation

$$\widehat{\mathcal{W}} = \underset{\mathcal{W}}{\operatorname{argmin}} \left\{ \overbrace{\|\mathcal{W}_\Omega - \mathcal{X}_\Omega\|_F^2}^{\text{Loss term}} + \overbrace{\mu \sum_{m=1}^{M} \operatorname{tr}(\mathcal{W}_{:,:,m}^\top L \mathcal{W}_{:,:,m})}^{\text{Local consistency}} \right\}$$

$$\text{subject to} \quad \underbrace{\operatorname{rank}(\mathcal{W}) \leq \rho}_{\text{Global consistency}}.$$

Experiments on Climate Datasets

Cokriging

Dataset	Tensor-F	Tensor-O	ADMM	Simple	MTGP
USHCN	0.7594	**0.7210**	0.8051	0.8760	1.0007
CCDS	0.5555	**0.4532**	0.8292	0.7634	1.0296

Forecasting

Dataset	Tensor-F	Tensor-O	Tucker	ADMM	OrthoNL	Trace	MTL_{l1}	MTL_{dirt}
USHCN	0.9171	0.9069	**0.8975**	0.9227	0.9175	0.9273	0.9528	0.9735
CCDS	0.8810	**0.8325**	0.9438	0.8448	0.8555	0.8632	0.9105	1.0950

Experiments on Climate Datasets

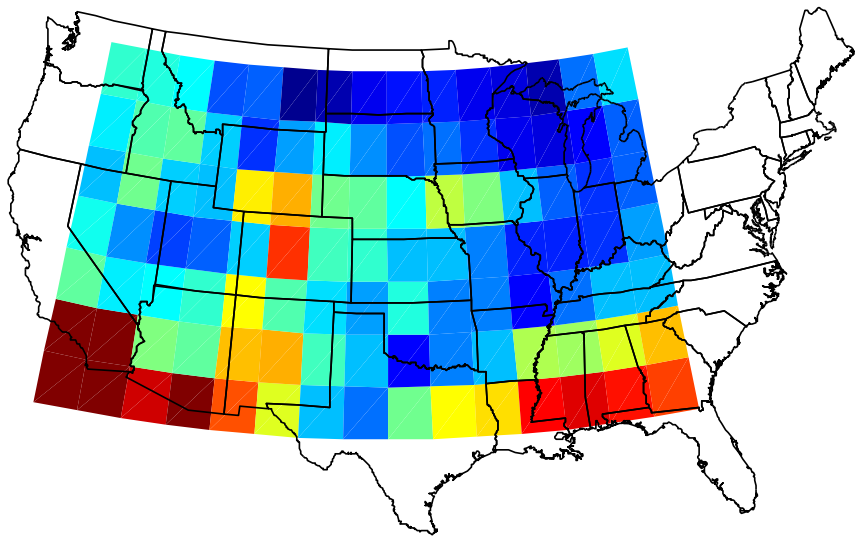
Cokriging

Dataset	Tensor-F	Tensor-O	ADMM	Simple	MTGP
USHCN	0.7594	**0.7210**	0.8051	0.8760	1.0007
CCDS	0.5555	**0.4532**	0.8292	0.7634	1.0296

Forecasting

Dataset	Tensor-F	Tensor-O	Tucker	ADMM	OrthoNL	Trace	MTL_{l1}	MTL_{dir}
USHCN	0.9171	0.9069	**0.8975**	0.9227	0.9175	0.9273	0.9528	0.9735
CCDS	0.8810	**0.8325**	0.9438	0.8448	0.8555	0.8632	0.9105	1.0950

Map of most predictive regions analyzed by the greedy algorithm using CCDS dataset.



Experiments on Scalability

Running time (in seconds) for cokriging and forecasting

	Cokriging			Forecasting		
Dataset	USHCN	CCDS	FSQ	USHCN	CCDS	FSQ
Tensor-O	93.03	16.98	91.51	75.47	21.38	37.70
ADMM	791.25	320.77	720.40	235.73	45.62	33.83
MTGP	∞					

Other Development

- Online prediction (ICML 2015)
- Memory efficient prediction (ICML 2016)
- Connections to Gaussian Processes (Upcoming)

Thank you!

For more information: USC Melady Group

`http://www-bcf.usc.edu/~liu32/melady.html`