20th ACM SIGKDD Conference
August 24-27, 2014, New York City
ACM SIGKDD 2014

FUNNEL: Automatic Mining of Spatially Coevolving Epidemics

Kumamoto University

Yasuko Matsubara
Kumamoto University
yasuko@cs.kumamoto-u.ac.jp

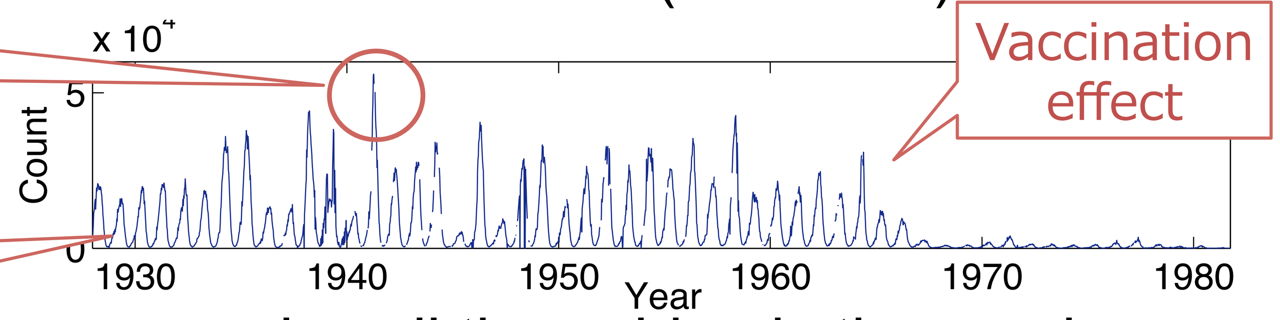
Yasushi Sakurai
Kumamoto University
yasushi@cs.kumamoto-u.ac.jp

Willem G. van Panhuis
University of Pittsburgh
wav10@pitt.edu

Christos Faloutsos
Carnegie Mellon University
christos@cs.cmu.edu

Motivation - Given: large set of epidemiological data

e.g., Measles cases from 1928 to 1982 (50 states)



Goal: statistically summarize all the epidemic time-series

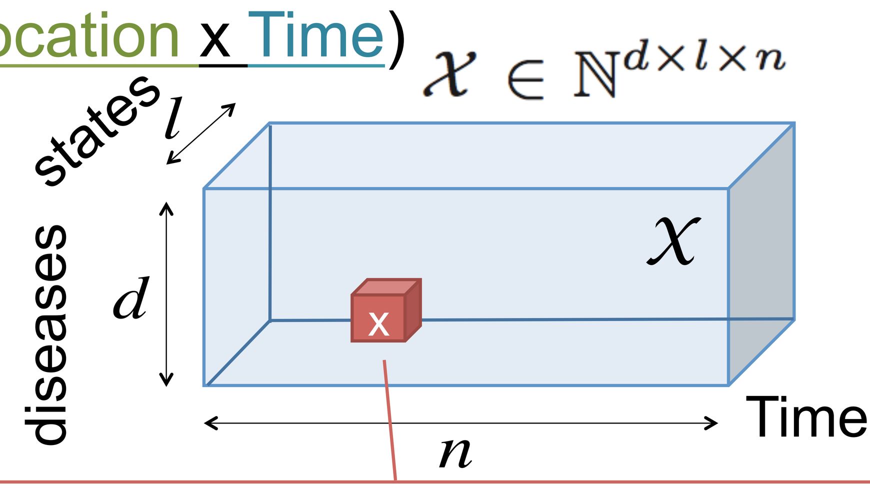
Data description – Project Tycho

- 56 contagious diseases for U.S. states
- from 1888 to the present (>125 years)

3rd order tensor (diseases x location x Time) $\mathcal{X} \in \mathbb{N}^{d \times l \times n}$

# of cases in 1931, …

Time	disease	loc	cases
04-01	measles	PA	4740
04-01	measles	NY	5310
04-01	rubella	CA	1923
…	…	…	…



Element x: # of cases (weekly)
e.g., 'measles', 'PA', 'April 1-7, 1931', '4740'

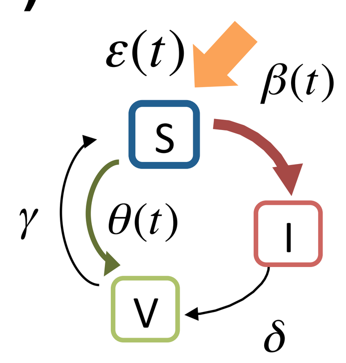
Observations - Properties of real epidemic data

- P1 yearly periodicity (e.g., flu peaks in the winter)
- P2 disease reduction effects (e.g., vaccination)
- P3 area specificity and sensitivity (e.g., correlation)
- P4 external shock events (e.g., outbreaks)
- P5 mistakes, incorrect values (e.g., typos)

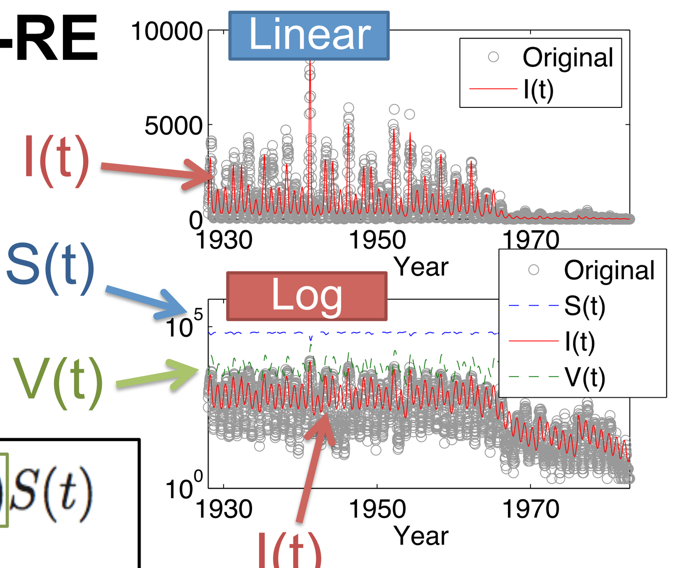
Proposed model: FUNNEL

(a) With a single epidemic: FUNNEL-RE

e.g., measles in NY



People of 3 classes
- S(t) : susceptible
- I (t) : Infected
- V(t) : Vigilant

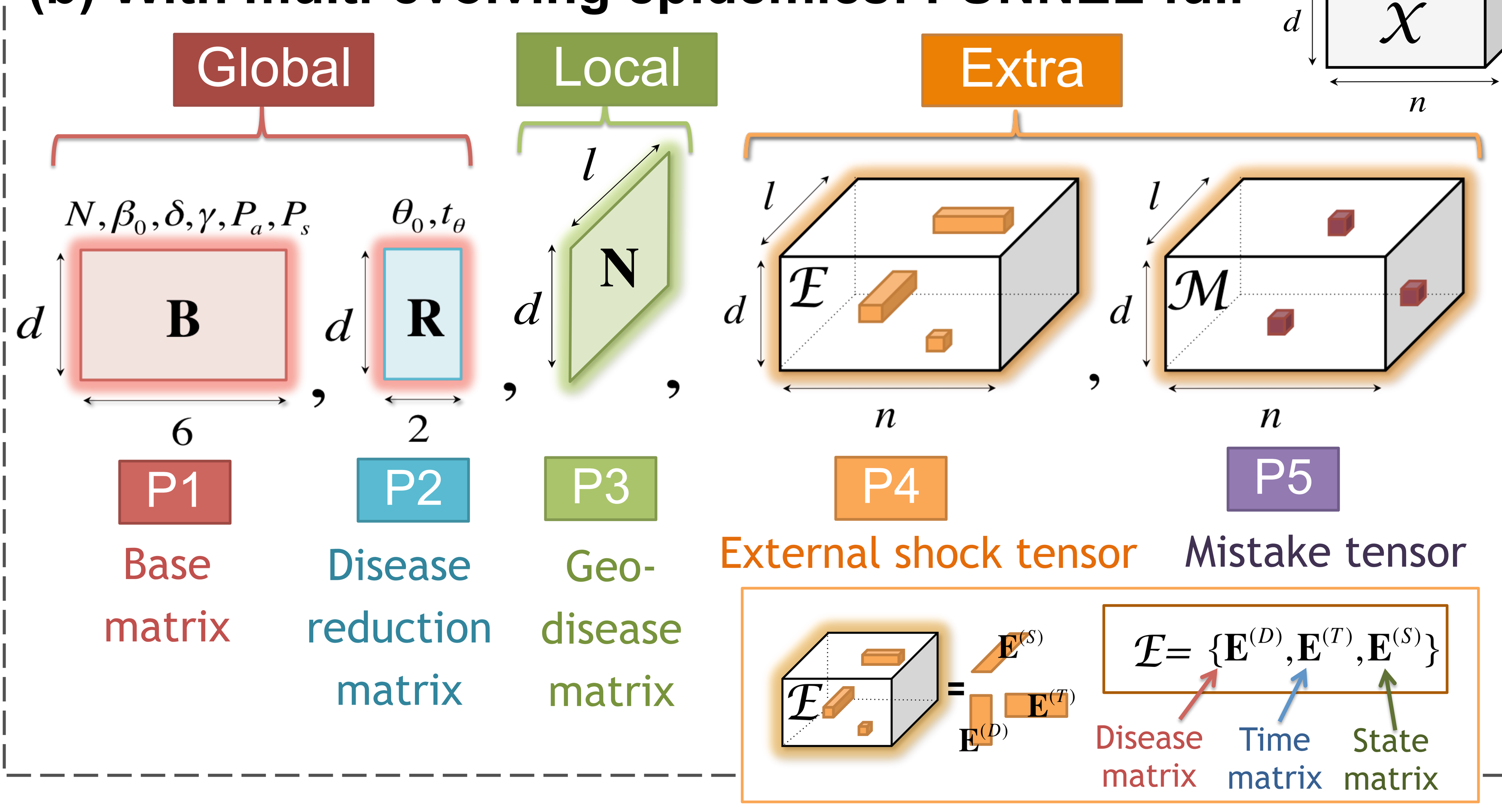


$$S(t+1) = S(t) - \beta(t)\epsilon(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t)$$
$$I(t+1) = I(t) + \beta(t)\epsilon(t)S(t)I(t) - \delta I(t)$$
$$V(t+1) = V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t) \quad (3)$$

- $\beta(t)$: strength of infection (yearly cycle)
- δ : healing rate γ : forgetting rate
- $\theta(t)$: disease reduction effect
- $\varepsilon(t)$: temporal susceptible rate

$$\beta(t) = \beta_0 \cdot \left(1 + P_a \cdot cos\left(\frac{2\pi}{P_p}(t + P_s)\right)\right)$$
$$P_p = 52$$
$$\theta(t) = \begin{cases} 0 & (t < t_\theta) \\ \theta_0 & (t \geq t_\theta) \end{cases}$$

(b) With multi-evolving epidemics: FUNNEL-full



Optimization algorithm: FUNNEL-Fit

Idea (1) Model description cost

Q. How can we find externals and mistakes??

A. Minimize coding cost!

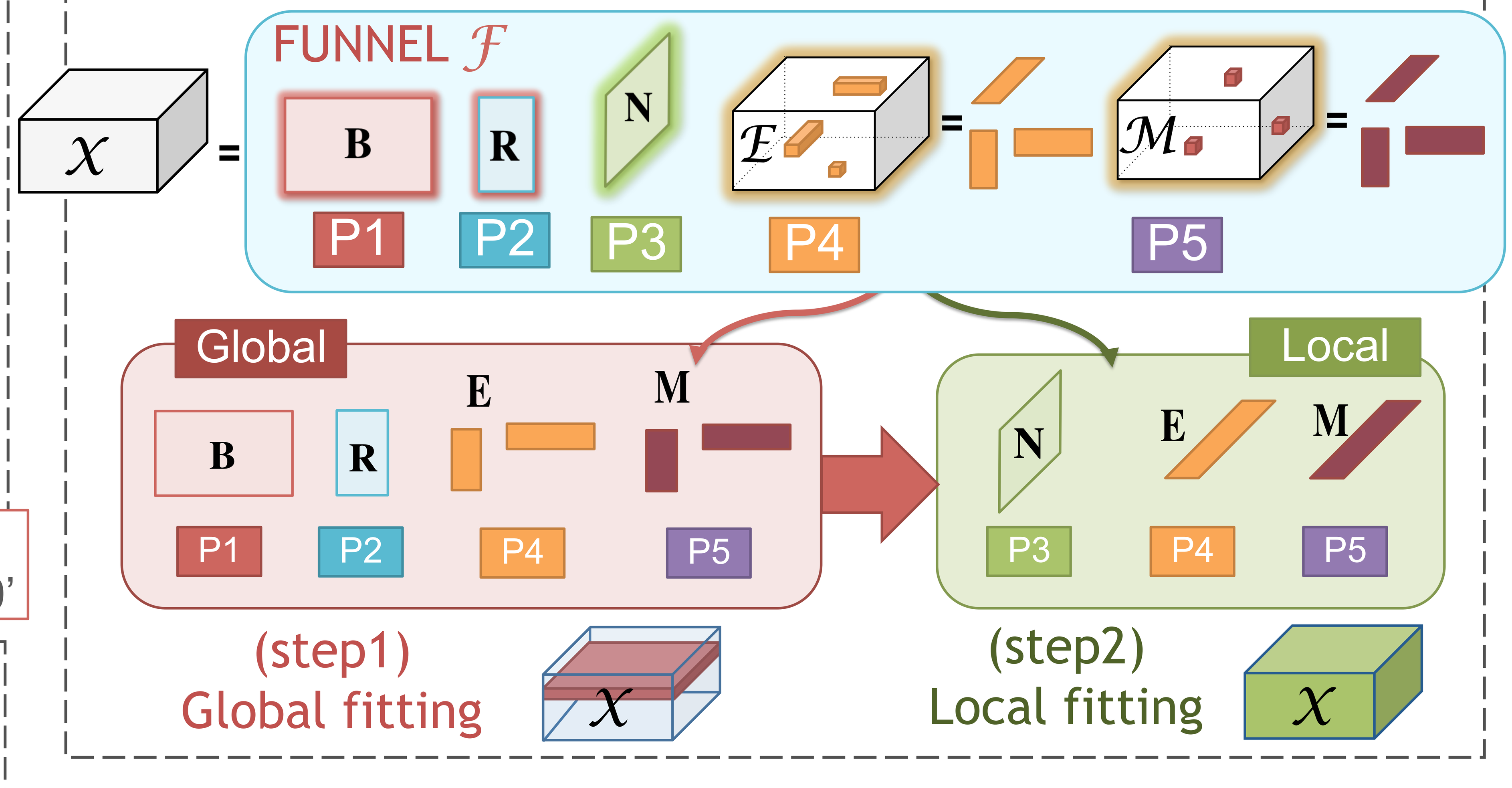
Dimensions of $\mathcal{X}$

$$Cost_T(\mathcal{X}; \mathcal{F}) = \log^*(d) + \log^*(l) + \log^*(n)$$
$$+ Cost_M(\mathbf{B}) + Cost_M(\mathbf{R}) + Cost_M(\mathbf{N})$$
$$+ Cost_M(\mathcal{E}) + Cost_M(\mathcal{M}) + Cost_C(\mathcal{X}|\mathcal{F})$$

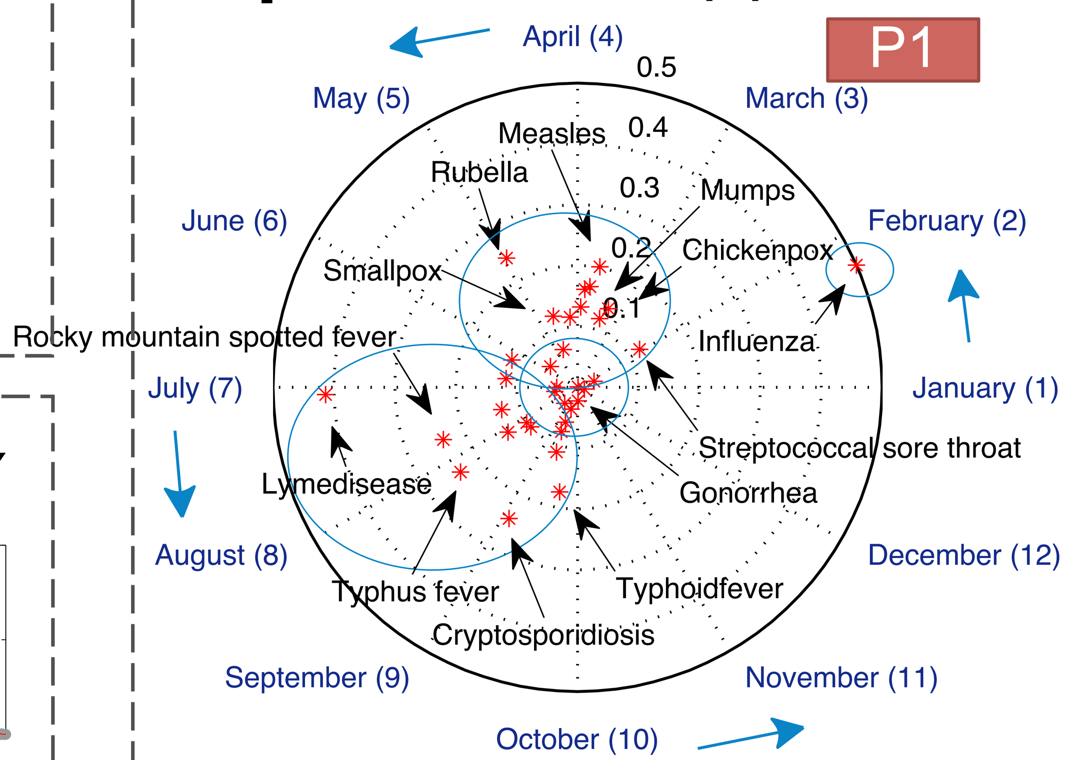
Model description cost of $\mathcal{F}$

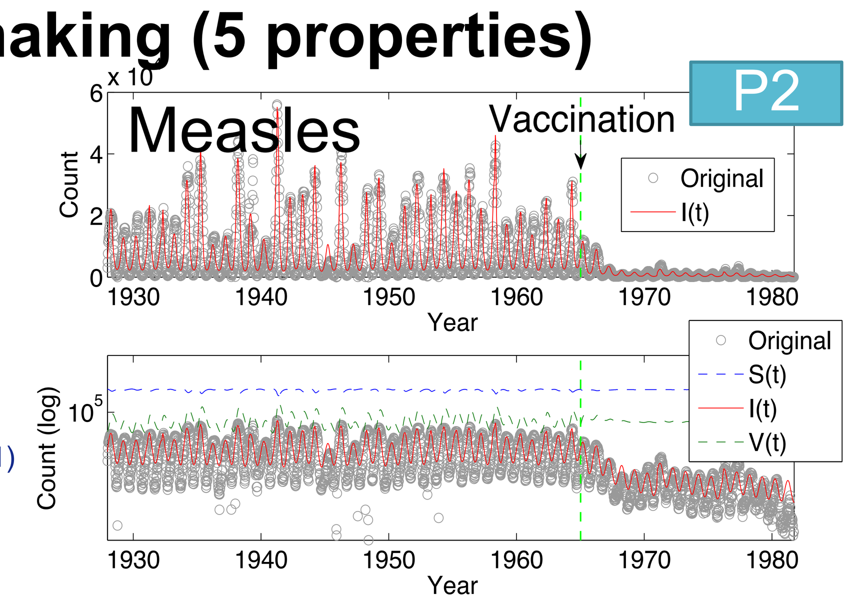
Coding cost of $\mathcal{X}$ given $\mathcal{F}$

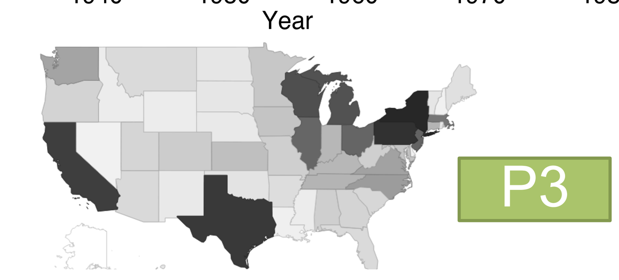
Idea (2) Multi-layer optimization

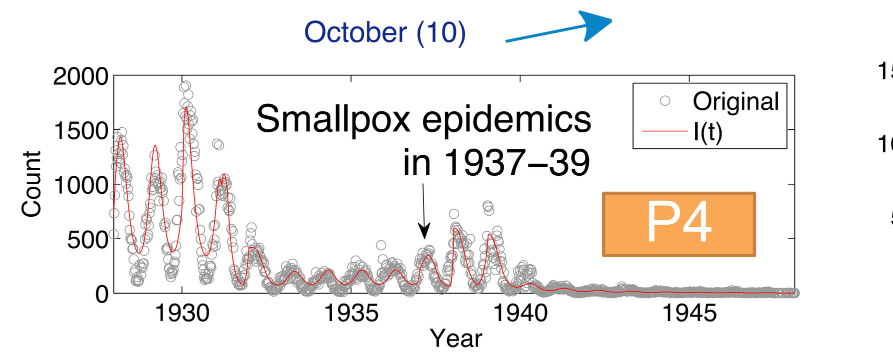


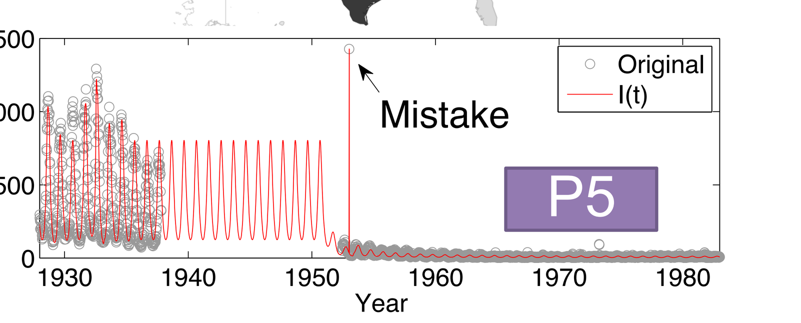
Experiments - (a) Sense-making (5 properties)



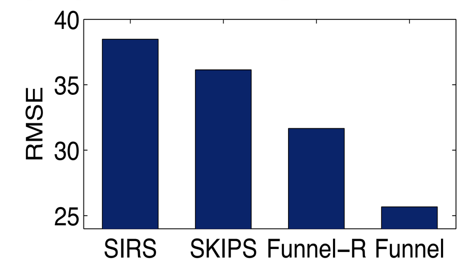


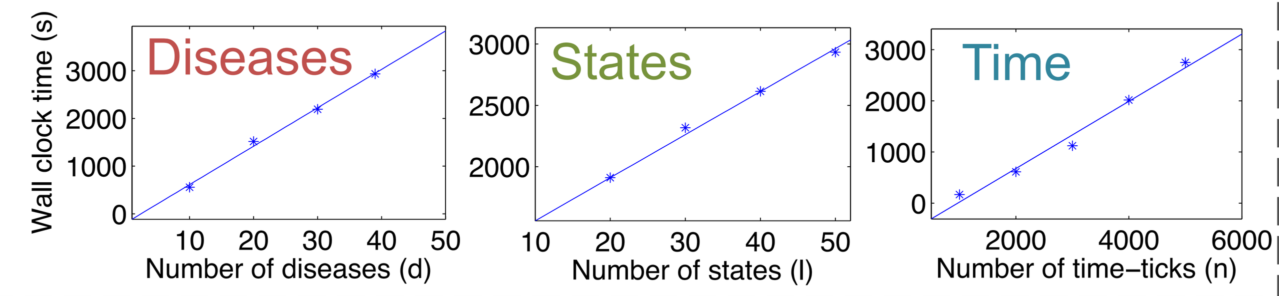




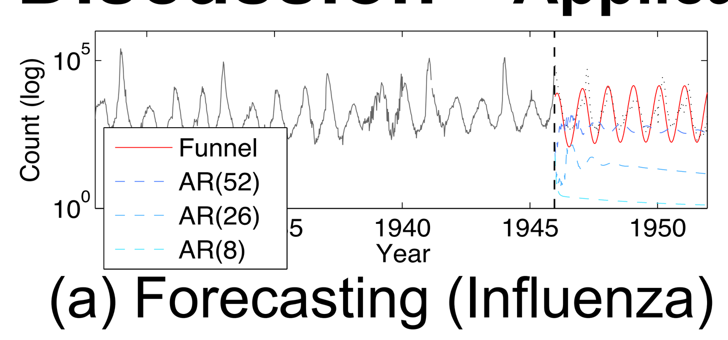


(b) Fitting accuracy (c) Scalability – (linear with data size)

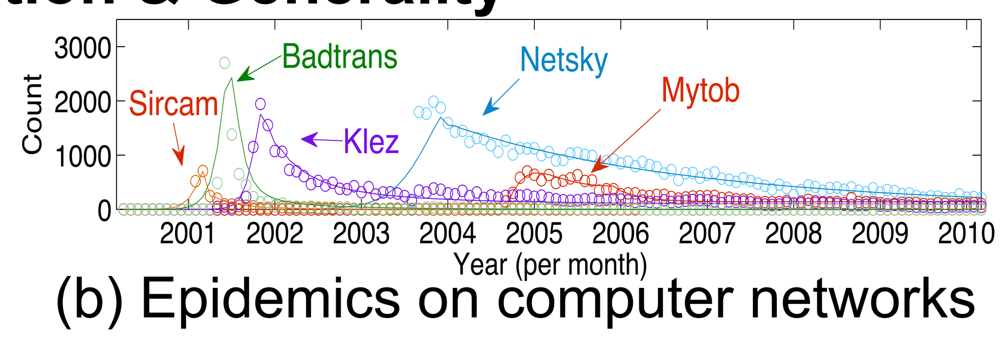




Discussion - Application & Generality



(a) Forecasting (Influenza)



(b) Epidemics on computer networks

Conclusions – FUNNEL has following advantages:

- General & Sense-making: it captures all essential aspects (P1-P5)
- Fully-automatic: it needs no training set
- Scalable: it scales linearly with the input size

Data: http://www.tycho.pitt.edu/

Code: http://www.cs.kumamoto-u.ac.jp/~yasuko/software.html