Mining and Forecasting of Big Time-series Data

Yasushi Sakurai (Kumamoto University)
Yasuko Matsubara (Kumamoto University)
Christos Faloutsos (Carnegie Mellon University)

Roadmap

- Motivation
- Similarity search, pattern discovery and summarization — **Part 1**
- **Non-linear modeling and forecasting** — **Part 2**
- Extension of time-series data: tensor analysis — **Part 3**

Part 2 Roadmap

Problem
- Why: "non-linear" modeling

Fundamentals
- Non-linear ("gray-box") models

Applications
- Epidemics
- Information diffusion
- (Online) competition vs.

Non-linear mining and forecasting

Q. What are "non-linear phenomena"?

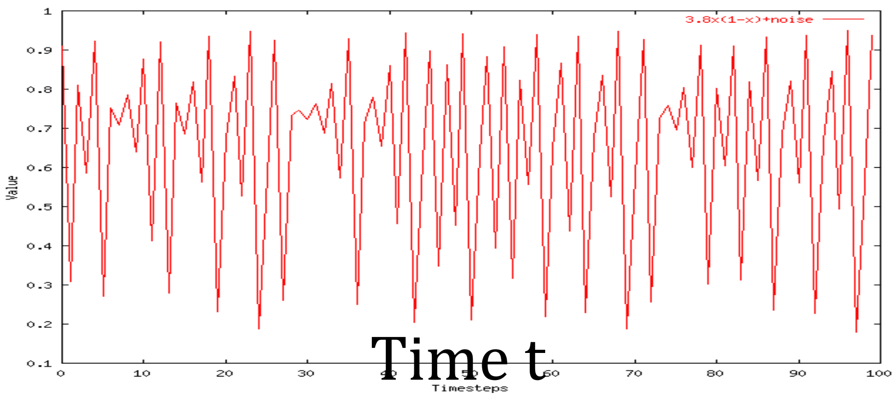
Example: logistic parabola

Models population of flies [R. May/1976]

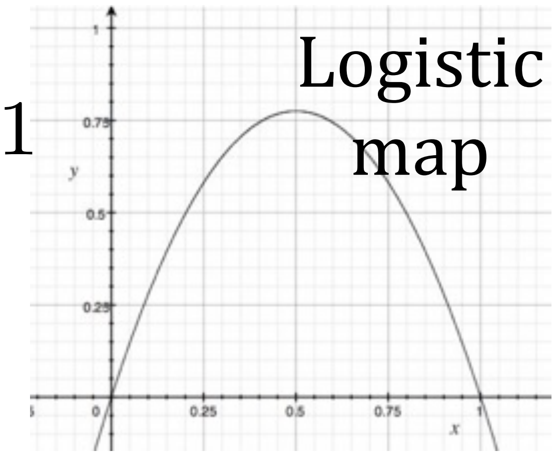
$$x_{t+1} = a x_t \cdot (1 - x_t)$$

Time-series plot

x_t



x_{t+1}



x_t

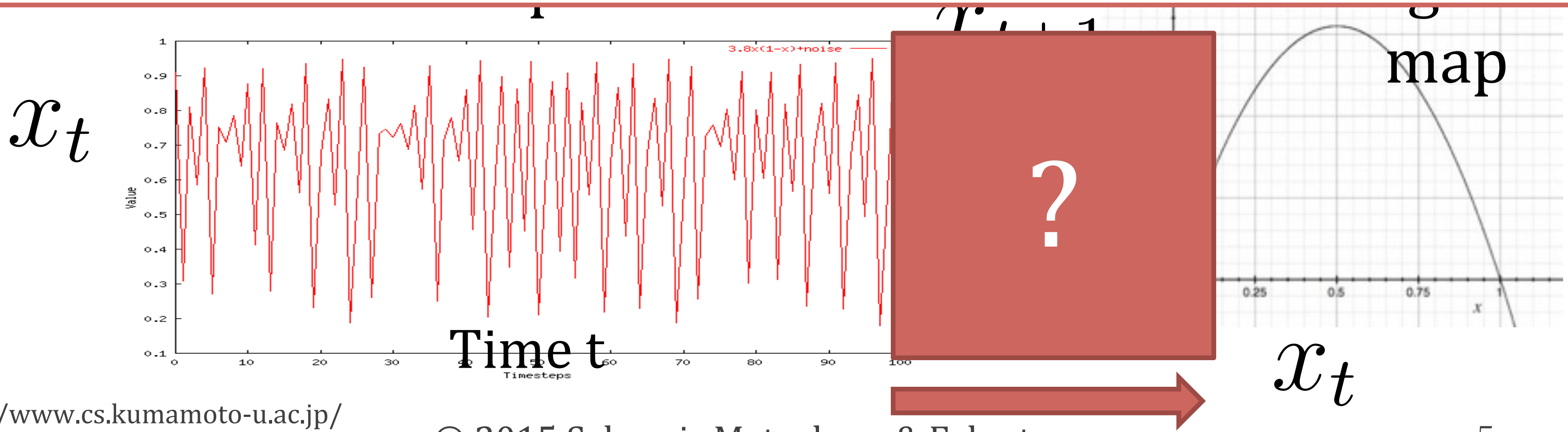
Non-linear mining and forecasting

Q. What are "non-linear phenomena"?

Problem:

Given: a time series x_t

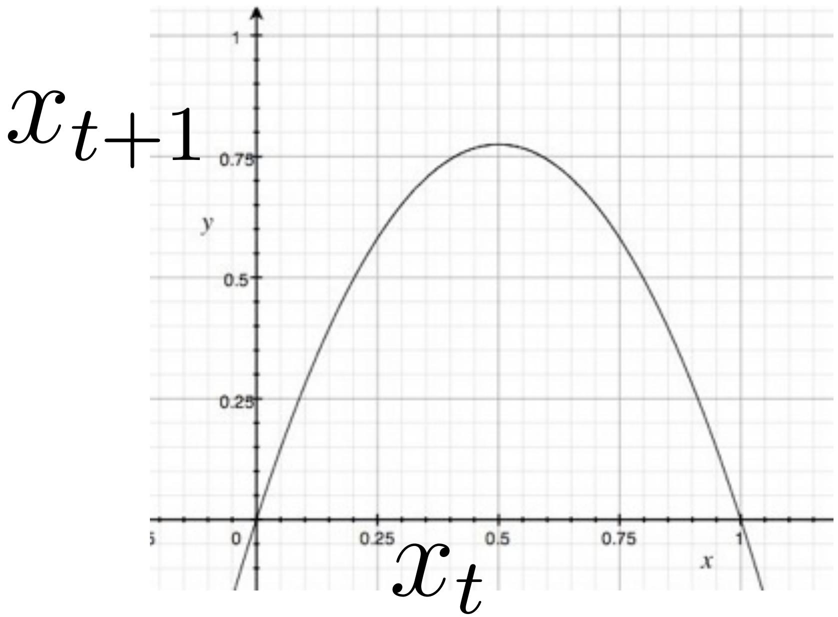
Predict: its future course, i.e., $x_{t+1}, x_{t+2}, \ldots$



How to forecast?

Solution 1

Linear equations, e.g., AR, ARIMA, …



Kumamoto U

CMU CS

How to forecast?

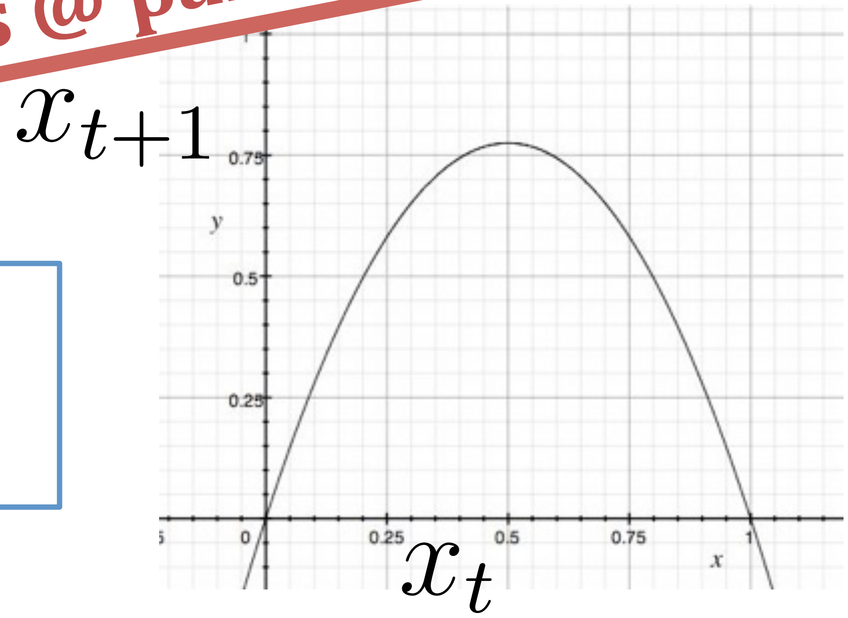
Solution 1

Linear equations, e.g., AR, ARIMA, …

Details @ part1

e.g., AR(1)

$$x_{t+1} = a x_t + \epsilon$$



Kumamoto U

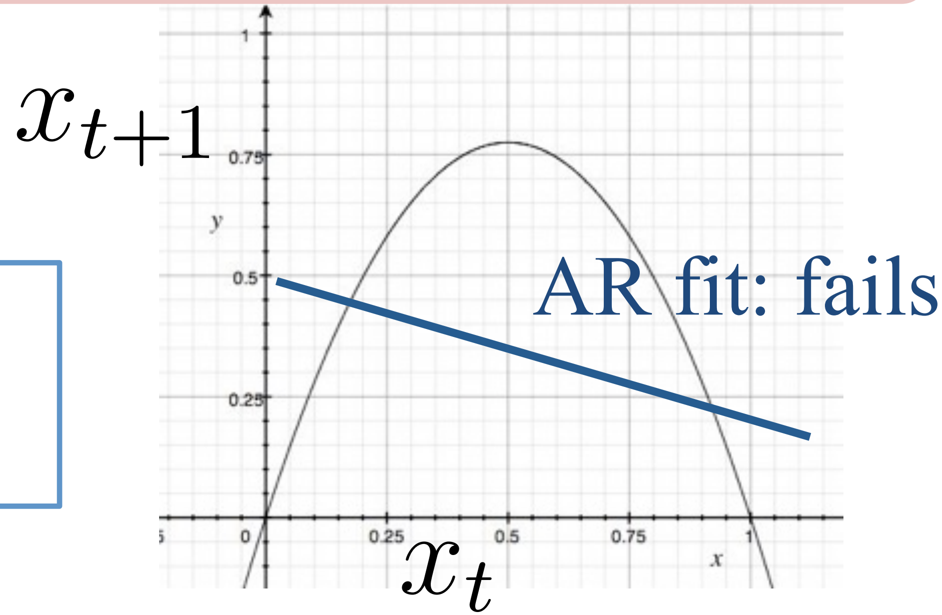
CMU CS

How to forecast?

Solution 1

Linear equations, e.g., AR, ARIMA, …

but: linearity assumption



e.g., AR(1)

$$x_{t+1} = ax_t + \epsilon$$

How to forecast?

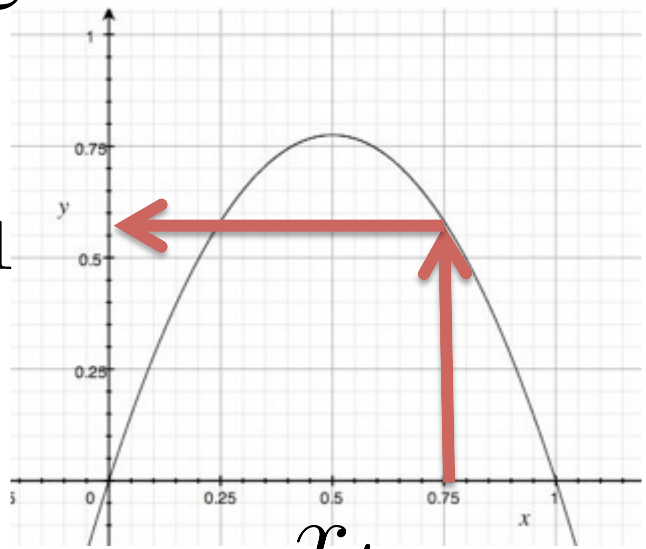
Solution 2

"Delayed Coordinate Embedding"

= Lag Plots [Sauer92]

- Based on k-nearest neighbor search

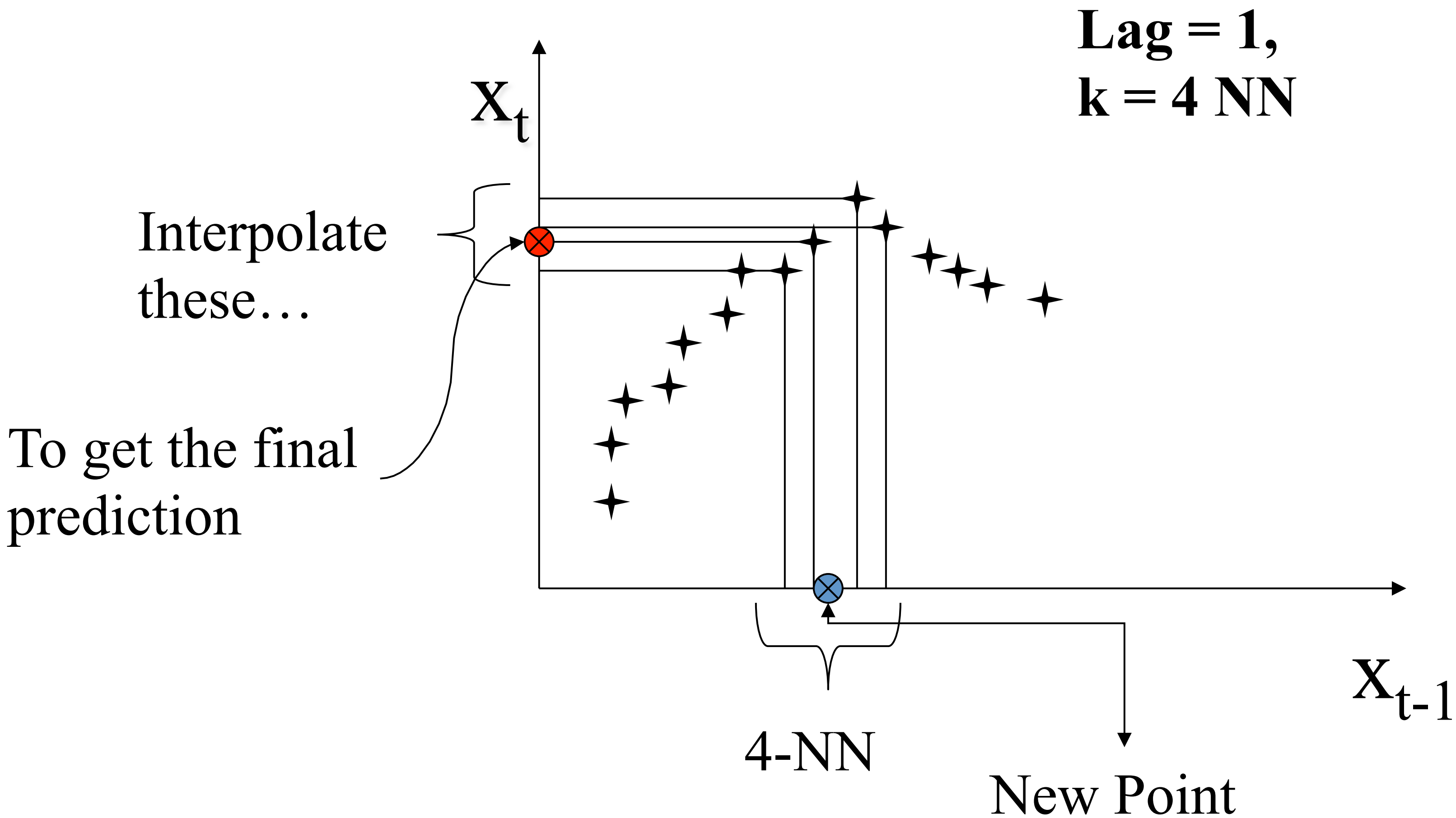
x_{t+1}



x_t

General Intuition (Lag Plot)

Solution 2



Kumamoto U

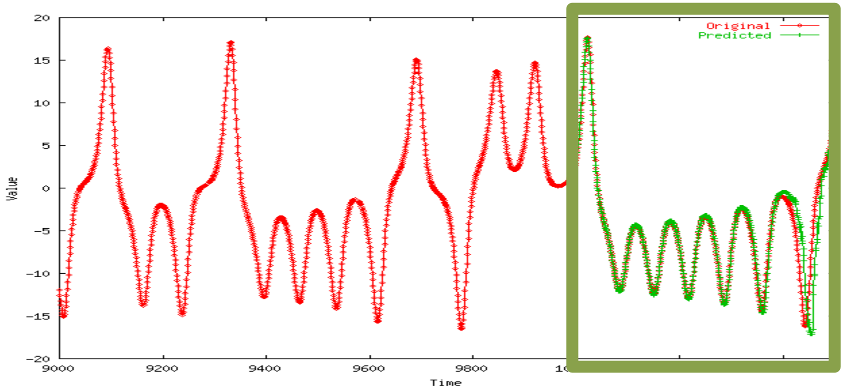
Forecasting results (Lag Plot)

CMU CS

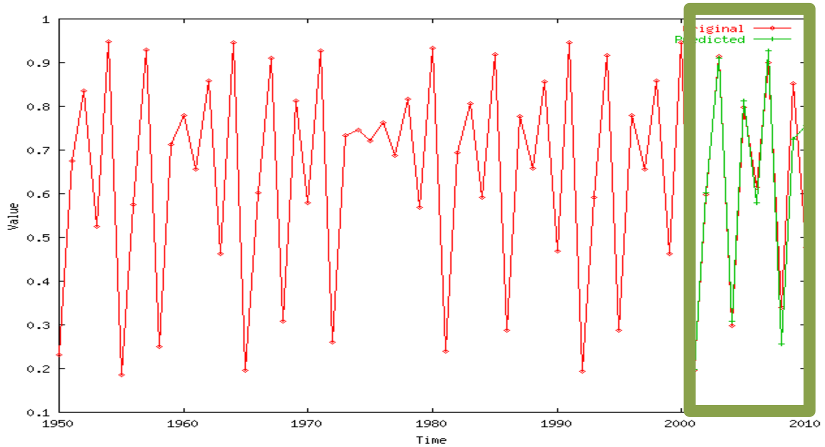
[Chakrabarti+ CIKM'02]

Solution 2

LORENZ

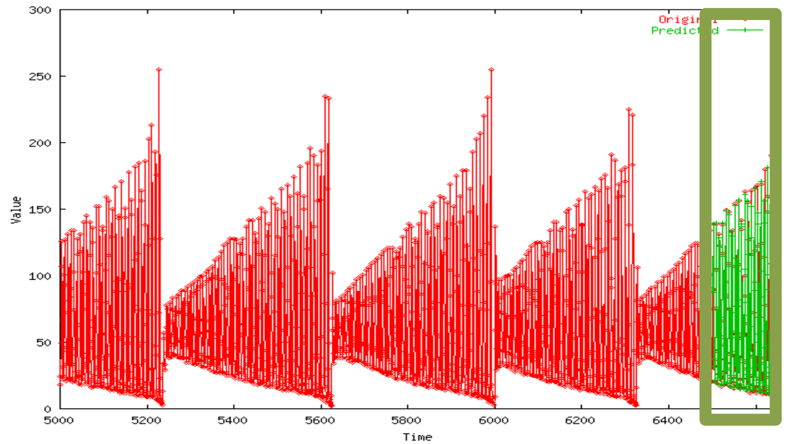


Logistic parabola



Laser

Forecast



Original x_t (red)

Forecasted $x_{t+1,..}$ (green)

http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/

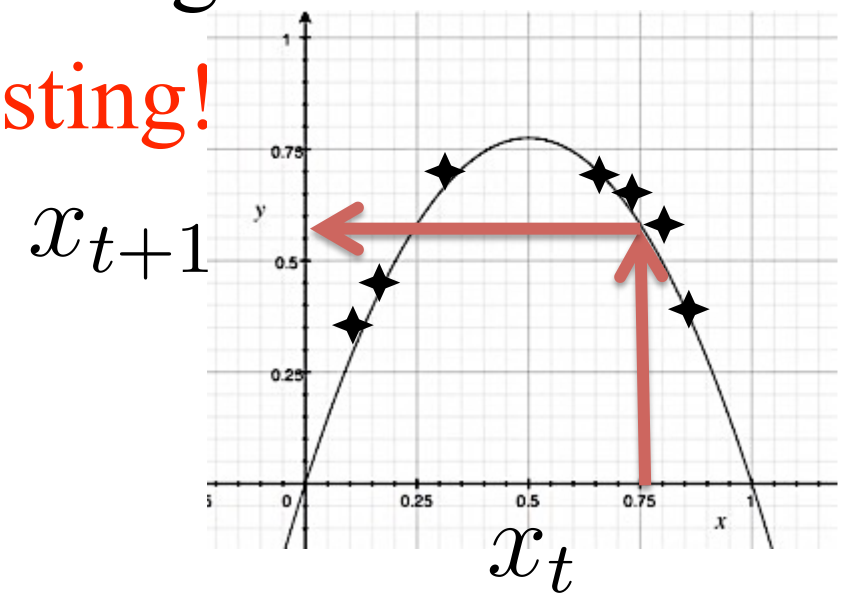
How to forecast?

Solution 2

"Delayed Coordinate Embedding"

= Lag Plots [Sauer92]

- Based on k-nearest neighbor search
- Non-linear Forecasting!



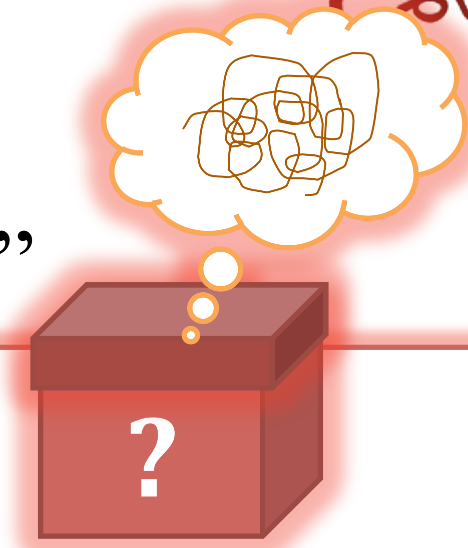
Kumamoto U

CMU CS

How to forecast?

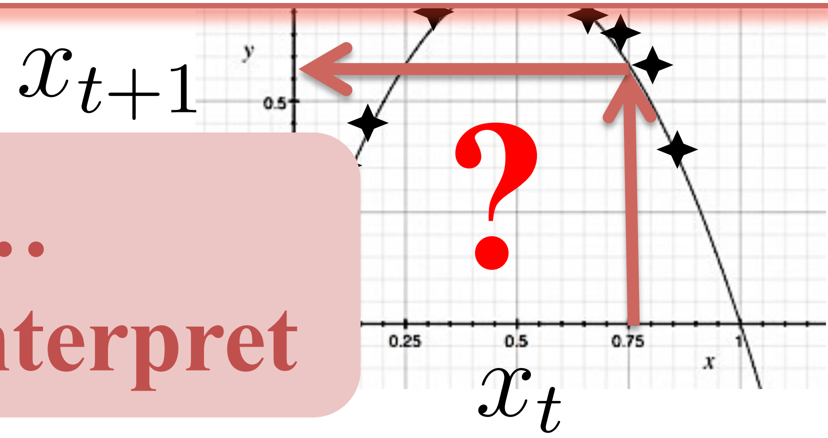
Solution 2

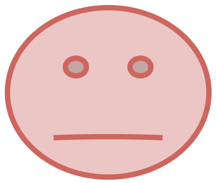
"Delayed Coordinate Embedding"



"Black-box" mining

(we don't know the equations)





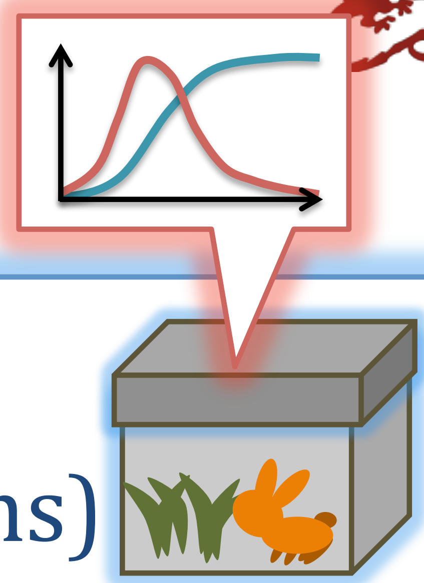
But, still,…
Hard to interpret

Kumamoto U

CMU CS

How to forecast?

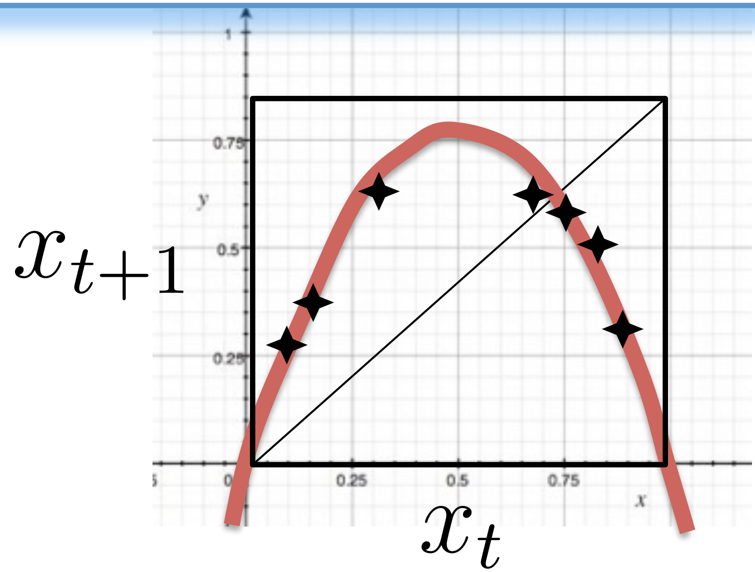
Solution 3



"Gray-box" mining
(if we know the equations)

Non-linear modeling!

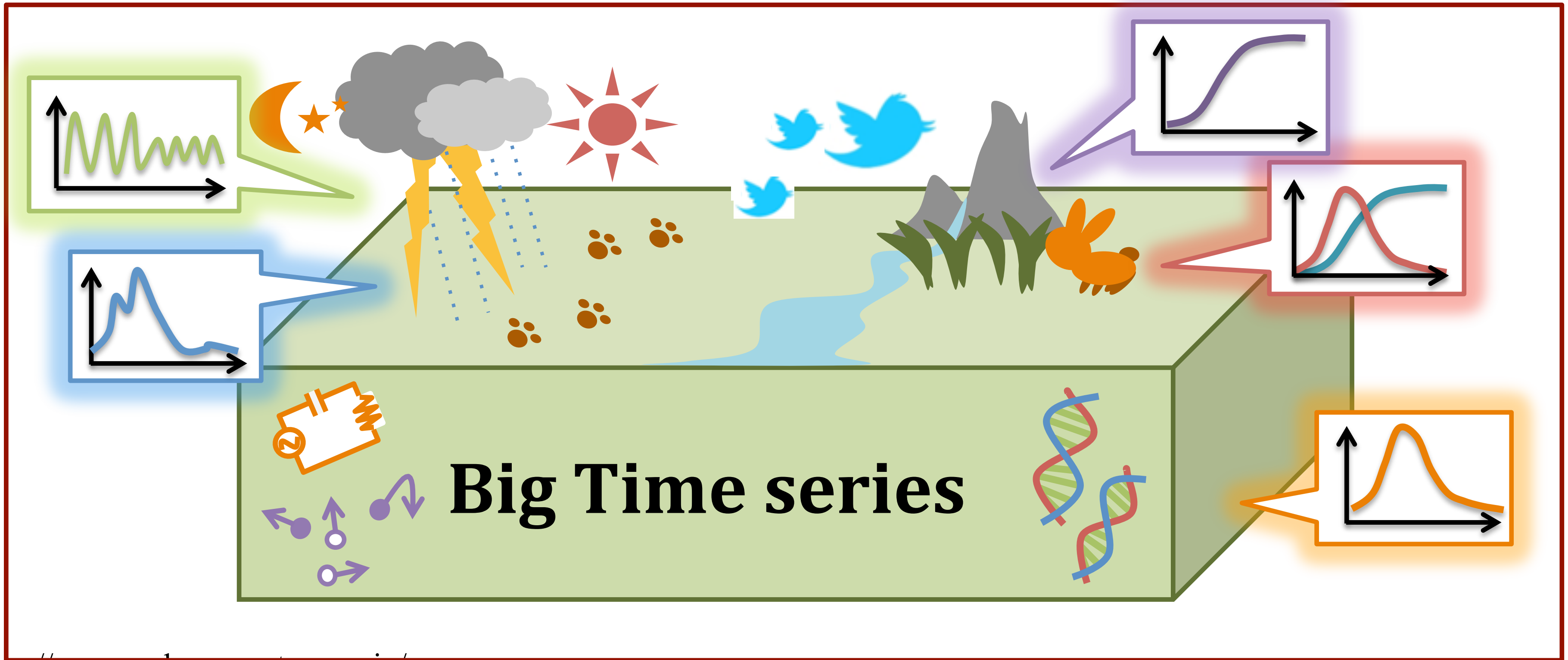
$$x_{t+1} = a x_t \cdot (1 - x_t)$$



How to forecast?

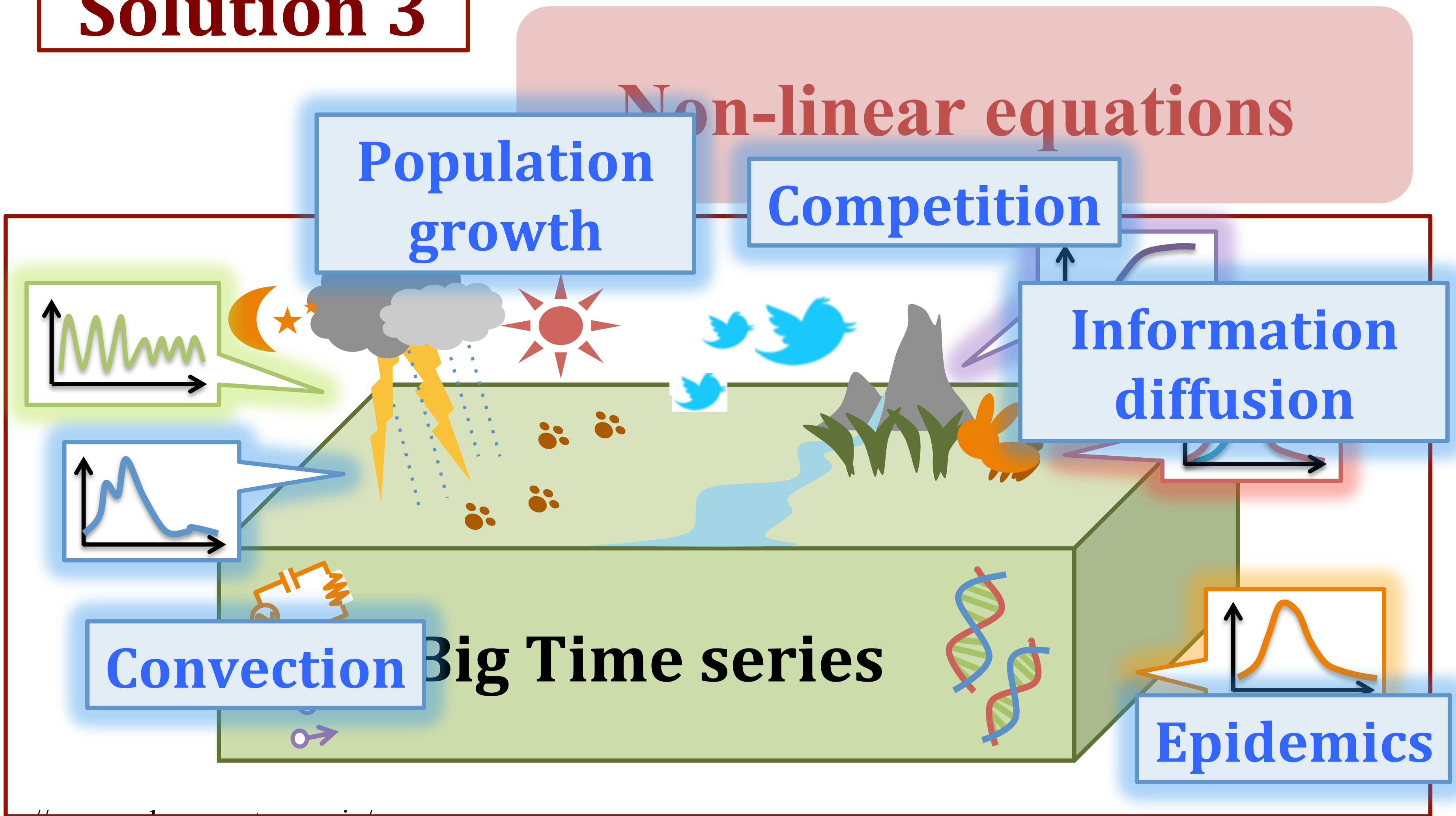
Solution 3

Non-linear equations



How to forecast?

Solution 3



Part 2 Roadmap

Problem

- ✓ Why: "non-linear" modeling

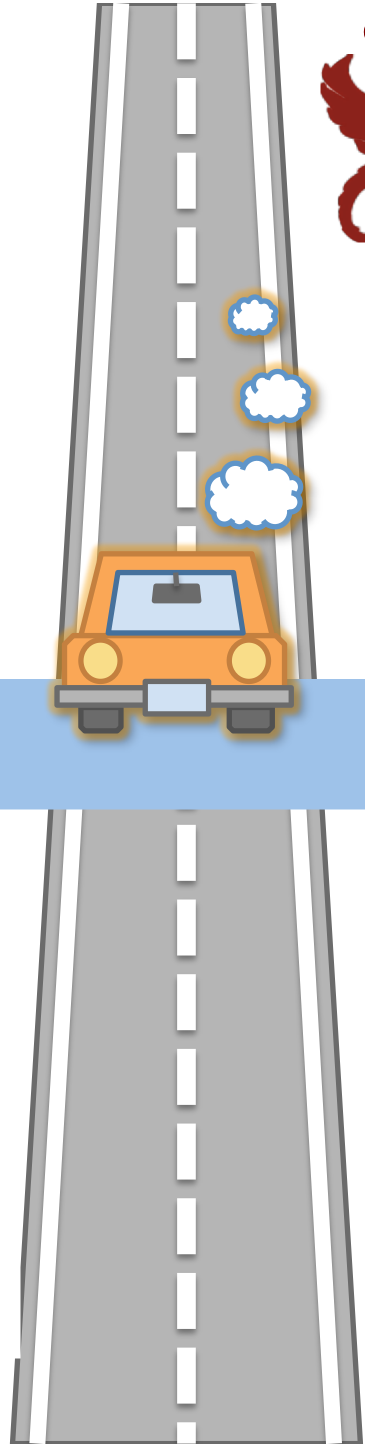
Fundamentals

- Non-linear (grey-box) models

Applications

- Epidemics
- Information diffusion
- (Online) competition

vs.



Kumamoto U

CMU CS

Part 2 Roadmap

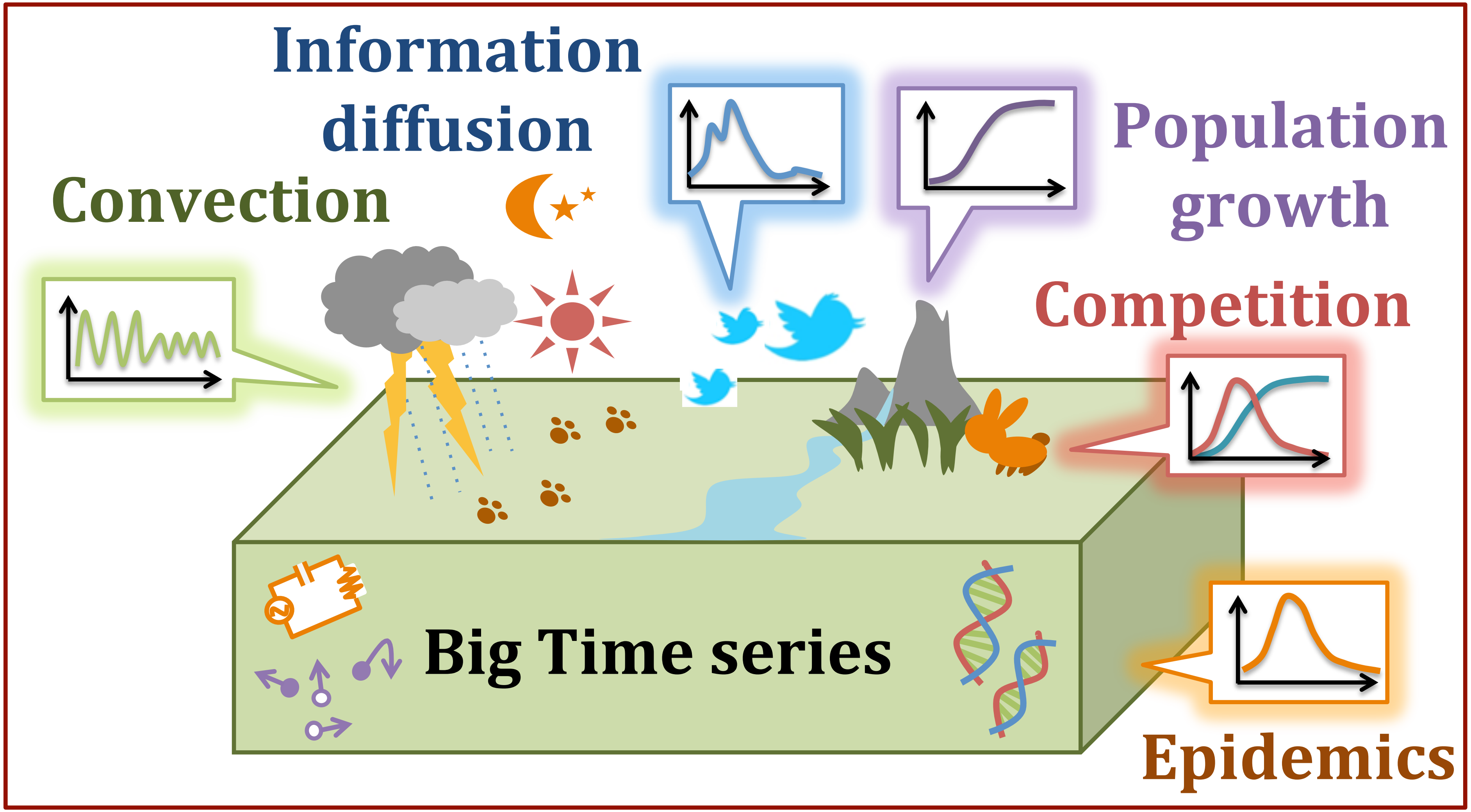
Problem

- ✓ Why: "non-linear" modeling

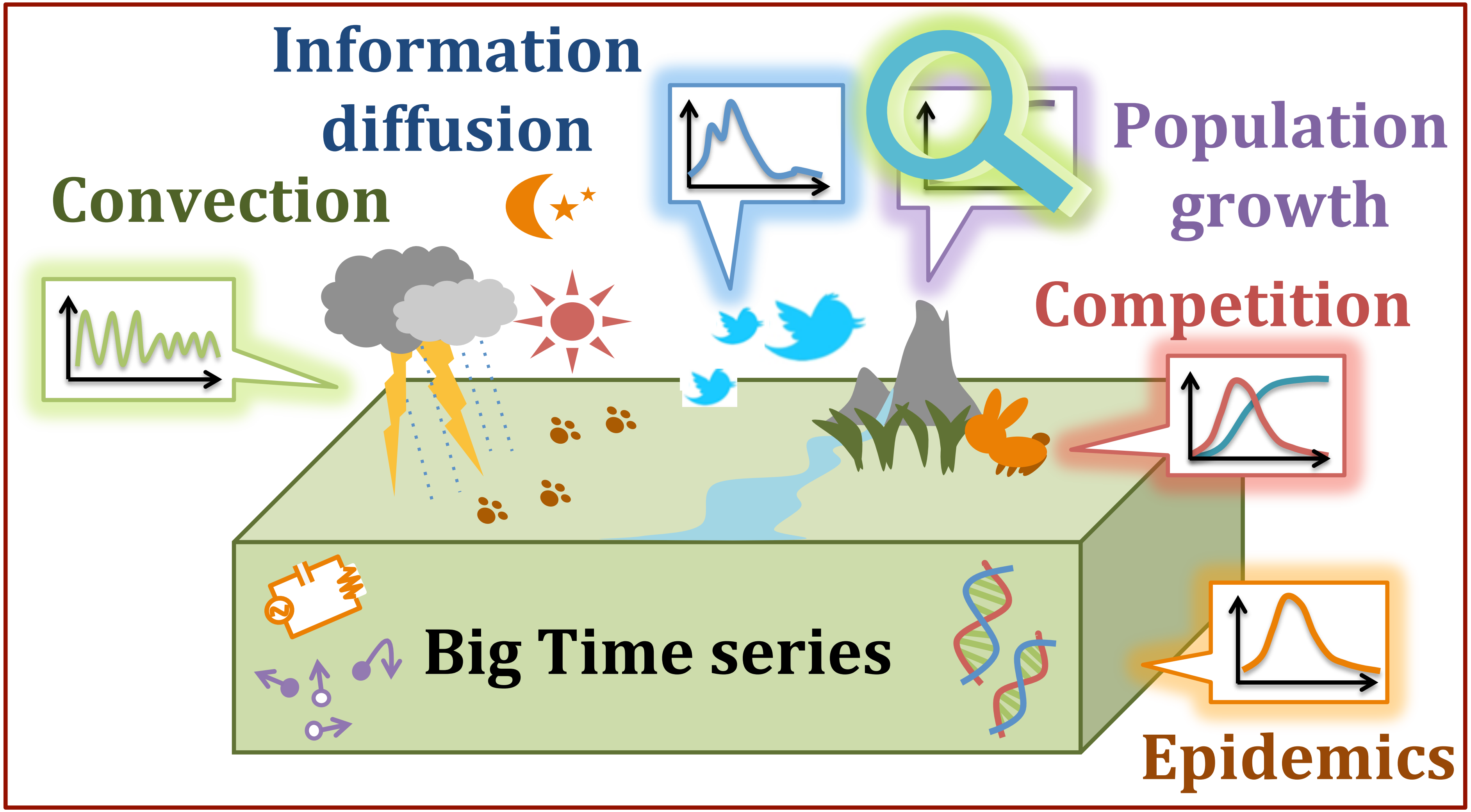
Fundamentals

- Non-linear (grey-box) models
  - Logistic function
  - Lotka-Volterra (prey-predator, competition)
  - SI, SIR models, etc.
  - Lorenz equations, etc.

Grey-box mining and non-linear equations



Grey-box mining and non-linear equations



Logistic function

So-called "Verhulst" model (=sigmoid, =Bass)

- Population expansion with limited resources

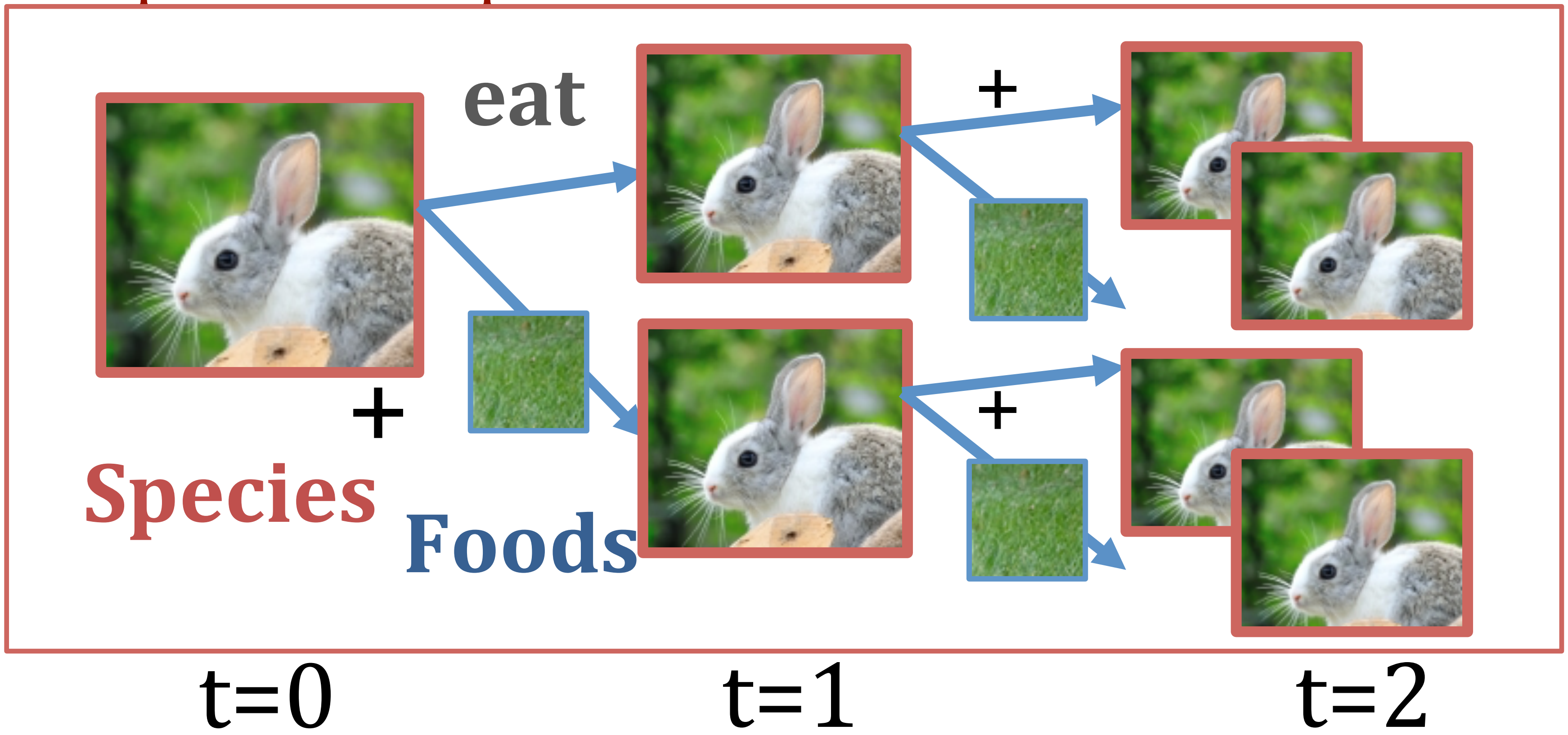


Image courtesy of amenic181 at FreeDigitalPhotos.net.

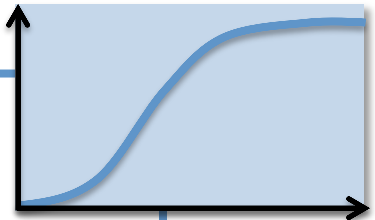
Logistic function

So-called "Verhulst" model (=sigmoid, =Bass)

- Population expansion with limited resources

P: Population size

$$\frac{dP}{dt} = rP\left(1 - \frac{P}{K}\right)$$



p – Initial condition (i.e., P(0) =p)

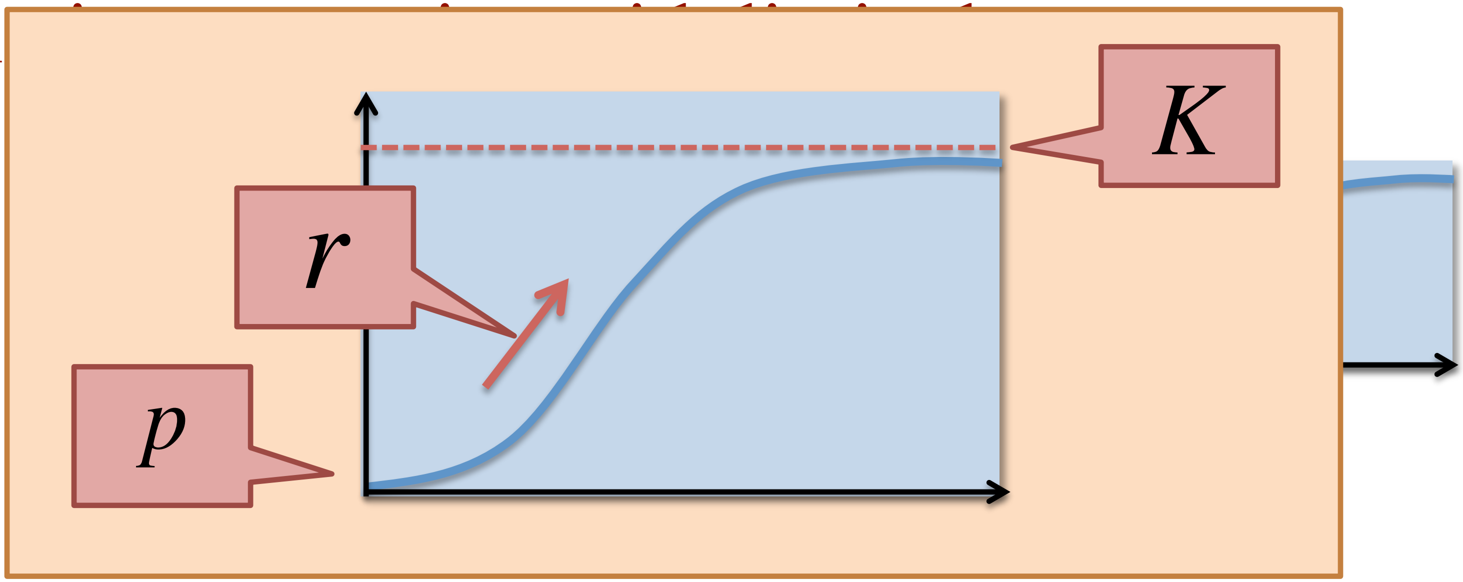
r – Growth rate, reproductively

K – Carrying capacity (=available resources)

Logistic function

So-called "Verhulst" model (=sigmoid, =Bass)

- Popul



p – Initial condition (i.e., P(0) =p)

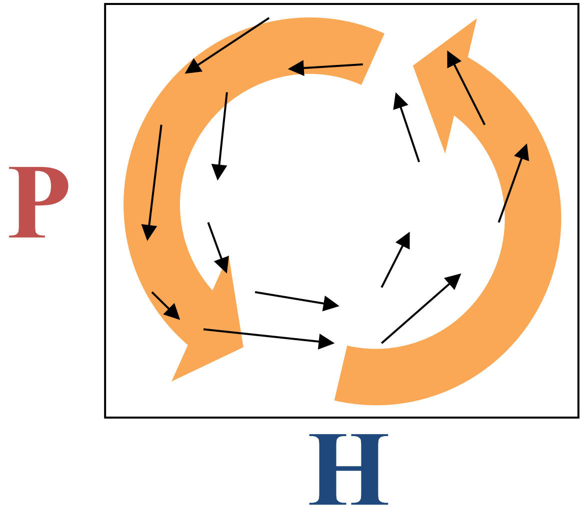
r – Growth rate, reproductively

K – Carrying capacity (=available resources)

Lotka-Volterra equations

So-called "prey-predator" model

Prey (H)



Predator (P)

- **H : count of prey (e.g., hare)**
- **P : count of predators (e.g., lynx)**

Image courtesy of Tina Phillips and amenic181 at FreeDigitalPhotos.net.

Lotka-Volterra equations

So-called "prey-predator" model

$$\frac{dH}{dt} = rH - aHP$$

$$\frac{dP}{dt} = bHP - mP$$

Prey (H)

Predator (P)

- **H : count of prey (e.g., hare)**
- **P : count of predators (e.g., lynx)**

Image courtesy of Tina Phillips and amenic181 at FreeDigitalPhotos.net.

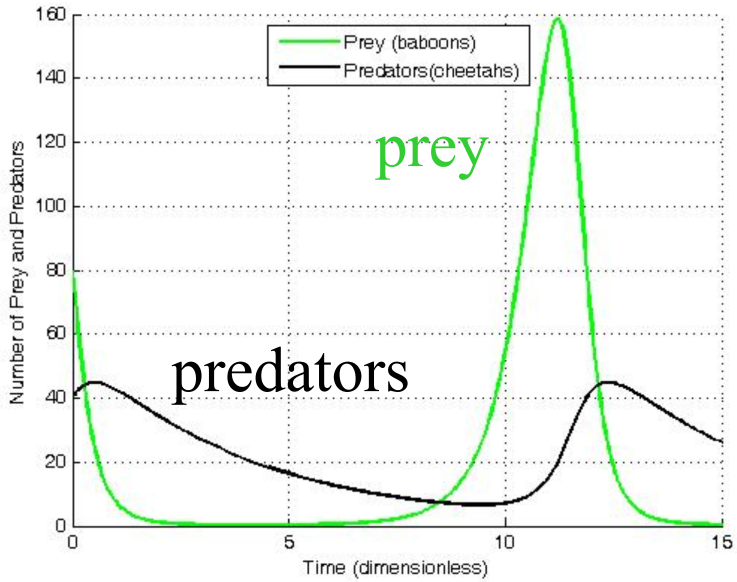
Kumamoto U

CMU CS

Solution to the Lotka-Volterra equations.

Frequency Plot

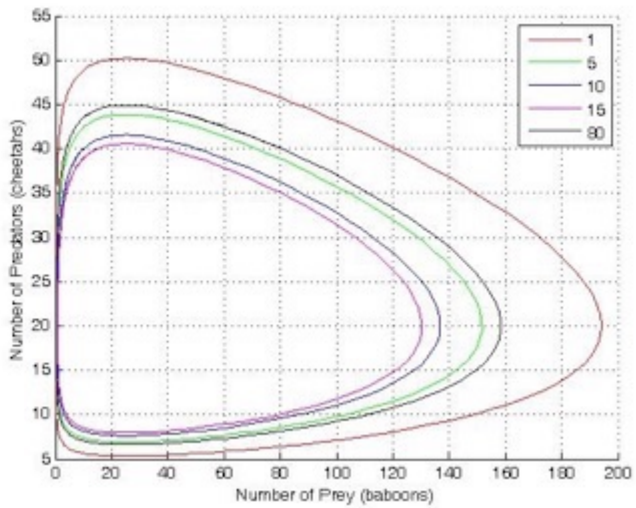
of prey/predators



time

Phase Space Plot

predators

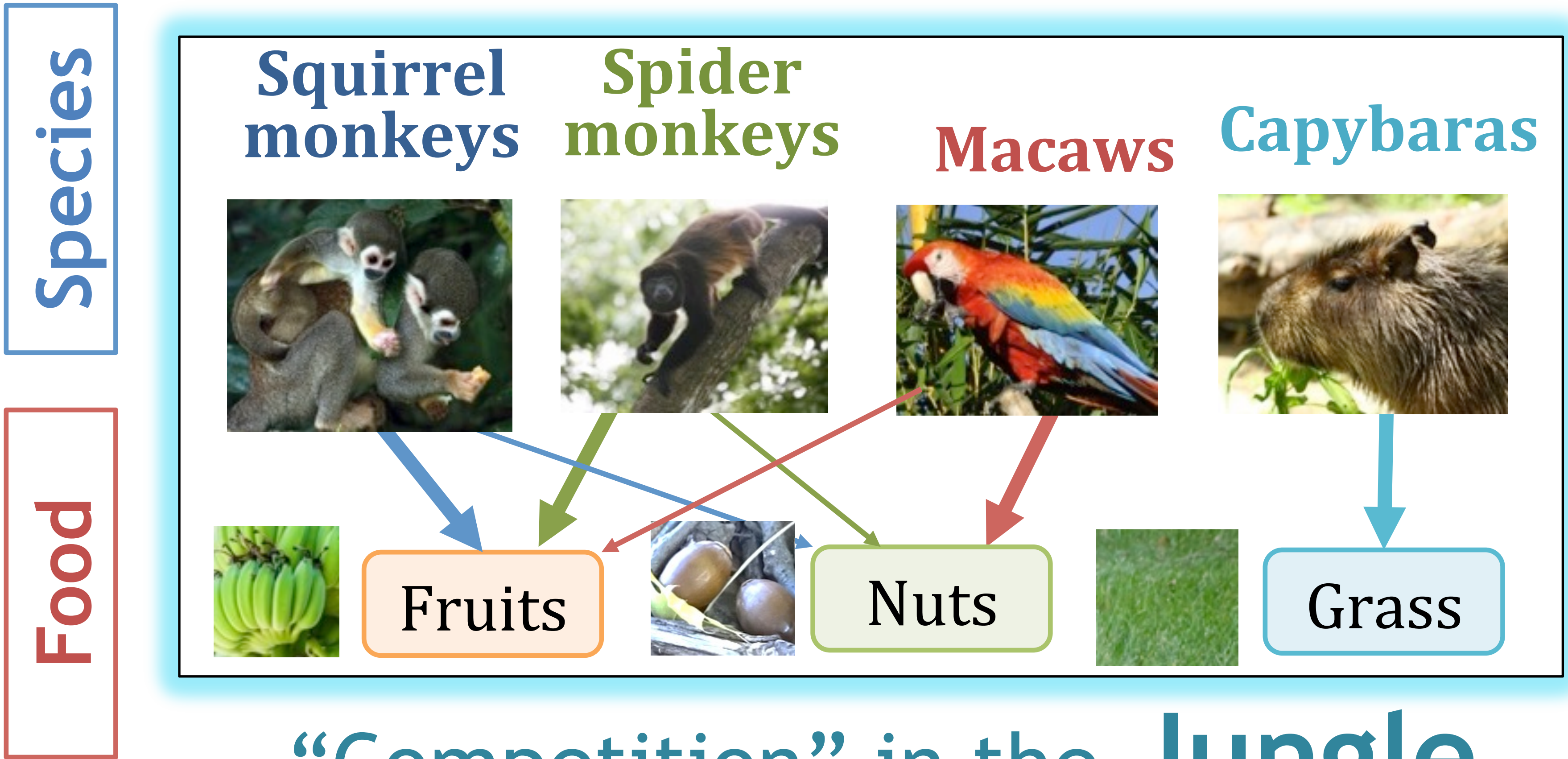


prey

From Wikipedia

Extension: "Competitive" Lotka-Volterra equations

Competition between multiple (d) species



"Competition" in the **Jungle**

Image courtesy of Tina Phillips and amenic181 at FreeDigitalPhotos.net.

Details

"Competitive" Lotka-Volterra equations

CMU CS

Competition between multiple (d) species

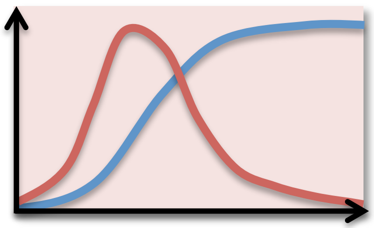
Population of species i — Population of j

$$\frac{dP_i}{dt} = r_i P_i \left(1 - \frac{\sum_{j=1}^{d} a_{ij} P_j}{K_i} \right)$$

$$(i = 1, \cdots, d)$$

a_{ij}: Interaction coefficient

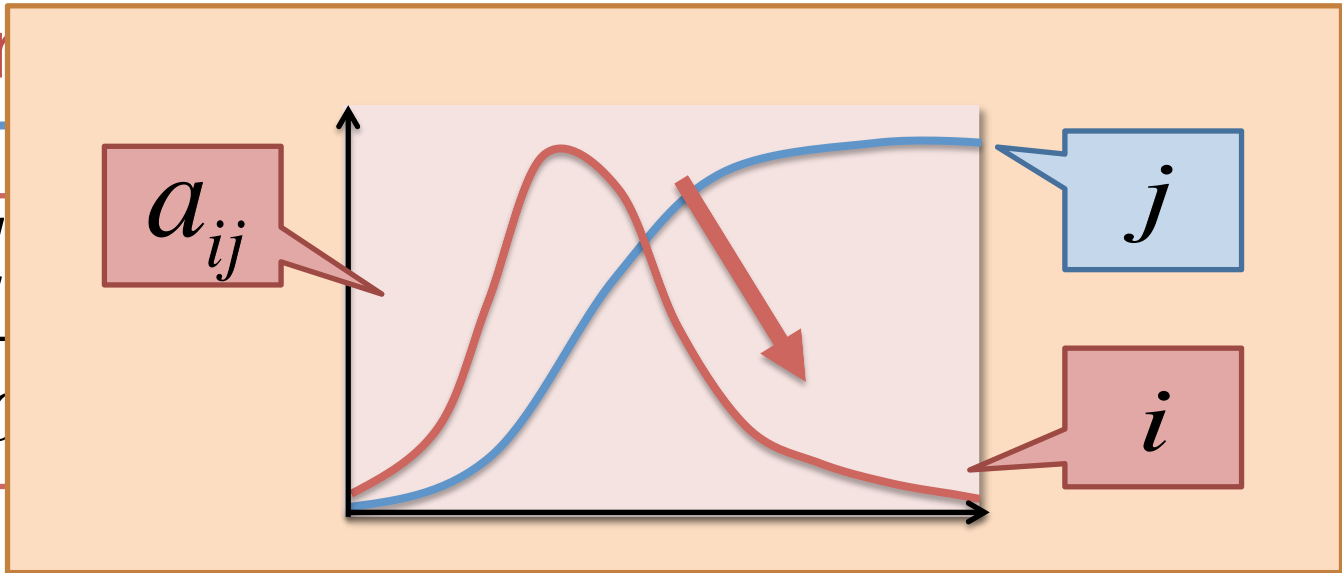
i.e., effect rate of species j on i



"Competitive" Lotka-Volterra equations

Competition between multiple (d) species

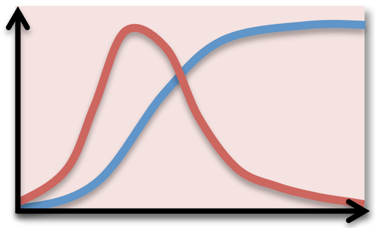
Popular



$(i = 1, \cdots, d)$

a_{ij} : Interaction coefficient

i.e., effect rate of species j on i



Details

"Competitive" Lotka-Volterra equations

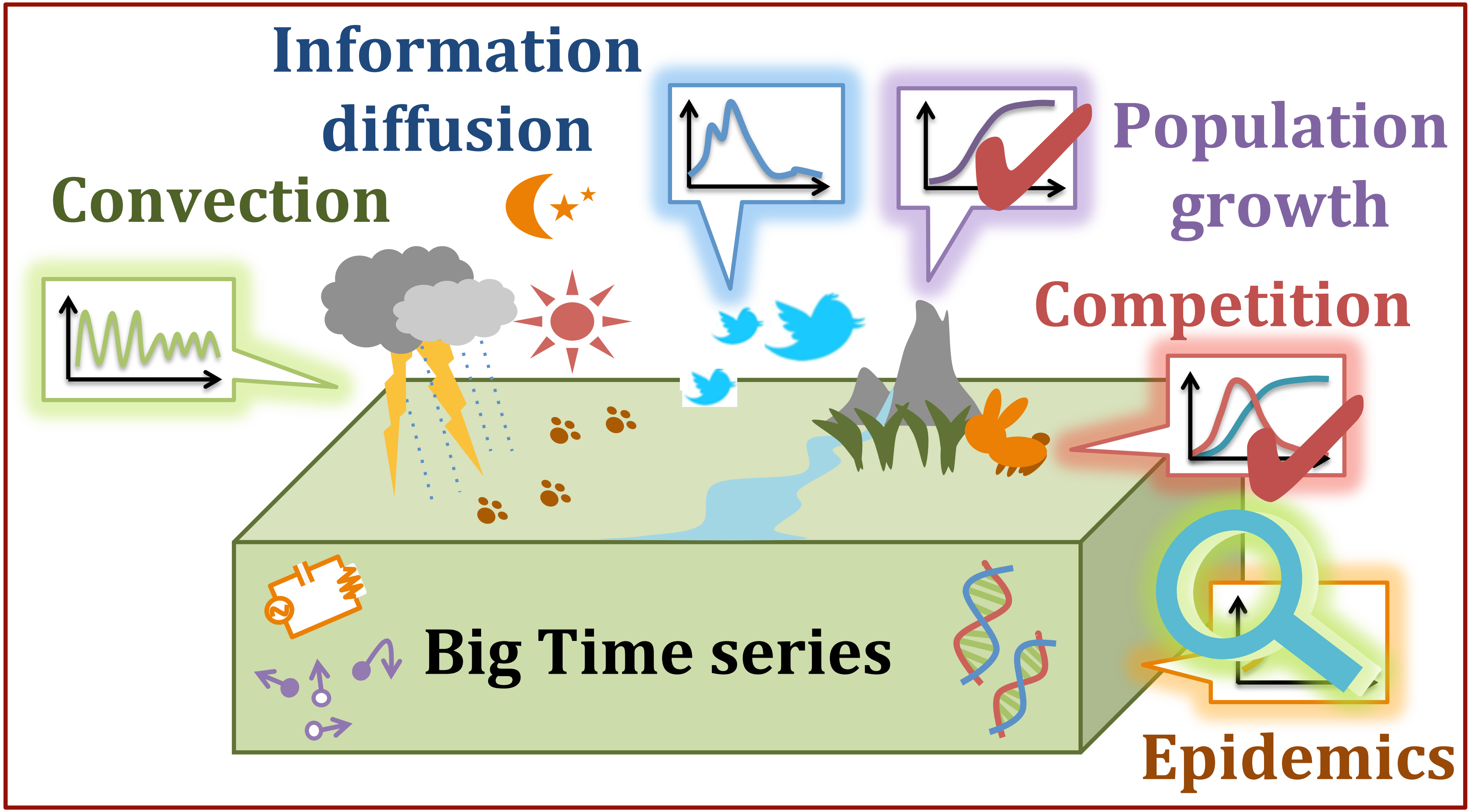
CMU CS

- Biological interaction
  - Table: Type of interaction

0 : no effect
- : detrimental
+ : beneficial

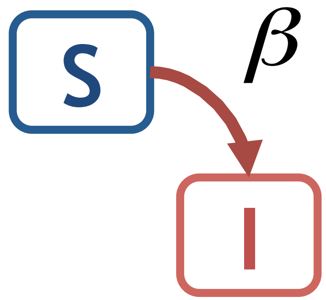
Species A \ Species B	+	0	-
+	Mutualism		
0	Commensalism	Neutralism	
-	Antagonism	Amensalism	Competition

Grey-box mining and non-linear equations



Epidemics: Susceptible-Infected (SI) model

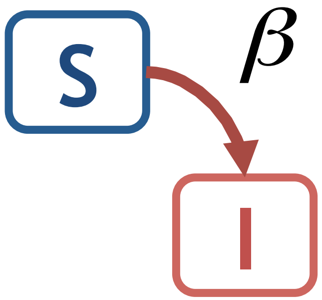
Each node is in one of two states



- S – Susceptible (healthy)
- I – Infected

Epidemics: Susceptible-Infected (SI) model

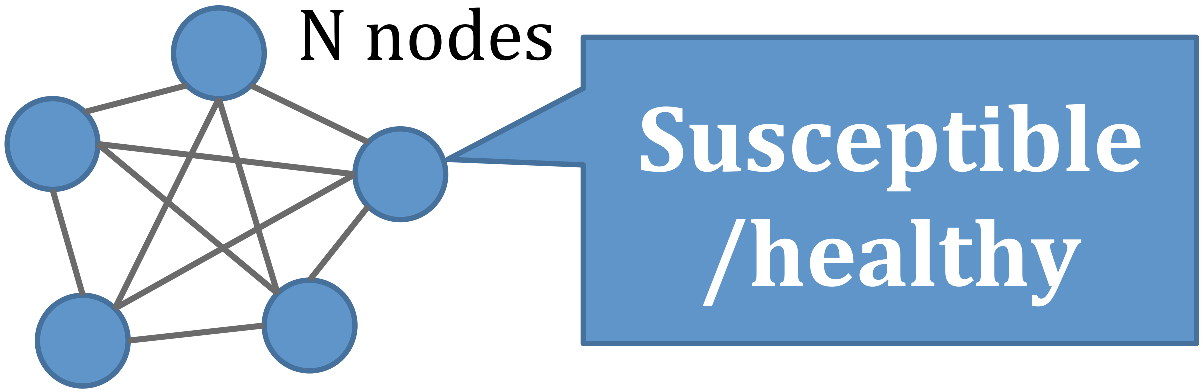
Each node is in one of two states



S – **Susceptible (healthy)**

I – **Infected**

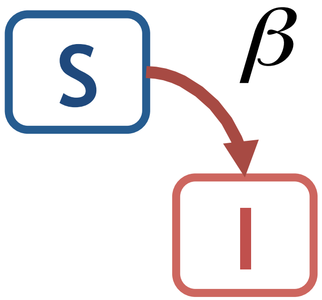
N nodes



Time t=0

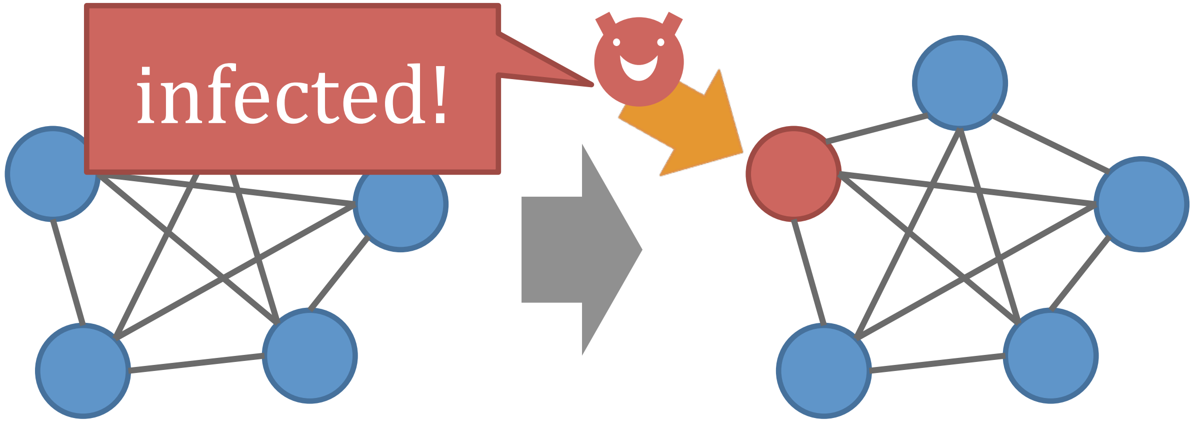
Epidemics: Susceptible-Infected (SI) model

Each node is in one of two states



S – Susceptible (healthy)

I – Infected



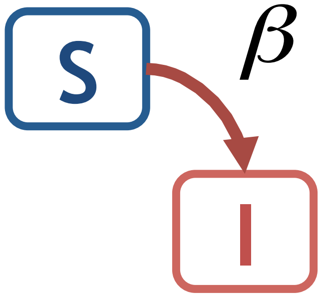
Time t=0

Time t=1

Epidemics: Susceptible-Infected (SI) model

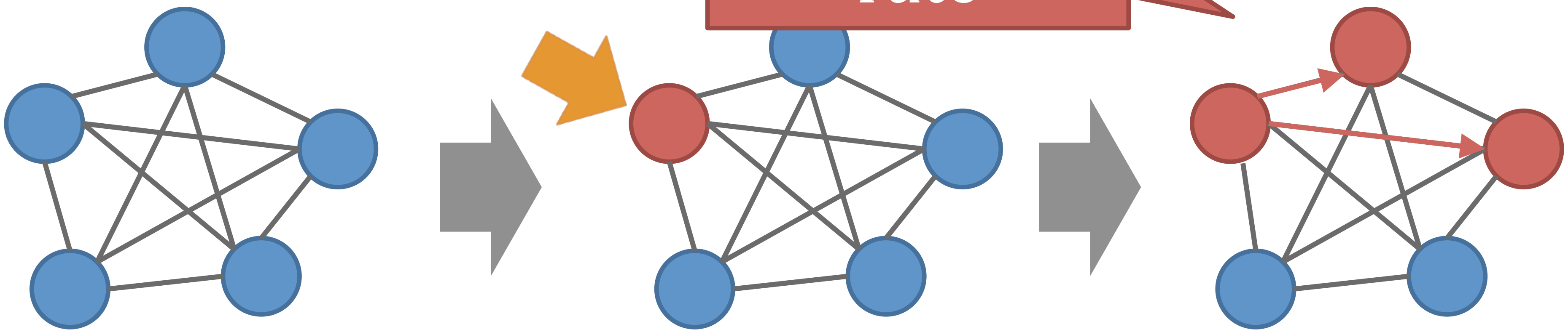
Each node is in one of two states

- S – **Susceptible (healthy)**
- I – **Infected**



Prob. β

β : infection rate



Time t=0

Time t=1

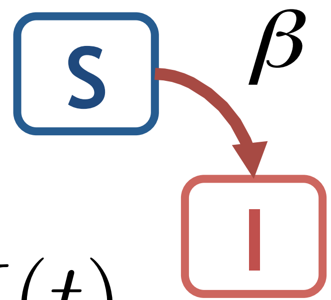
Time t=2

Kumamoto U

Epidemics: Susceptible-Infected (SI) model

CMU CS

Each node is in one of two states



$$\frac{dS}{dt} = -\beta S I$$

$$\frac{dI}{dt} = +\beta S I$$

$$N = S(t) + I(t)$$

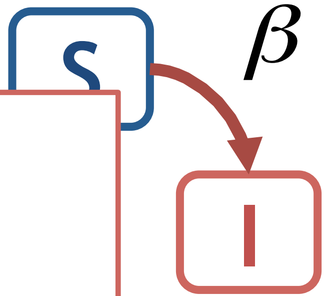
β : Infection strength

N : Population size

i.e., $$\frac{dI}{dt} = \beta(N - I)I$$

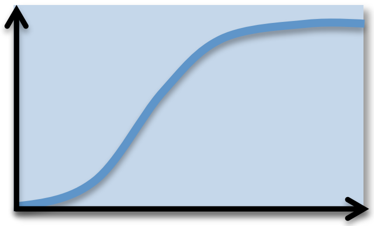
Epidemics: Susceptible-Infected (SI) model

Each node is in one of two states



Logistic function

$$\frac{dP}{dt} = rP(1 - \frac{P}{K})$$



SI model

$$\frac{dI}{dt} = \beta N \cdot I(1 - \frac{I}{N})$$

i.e., $$\frac{dI}{dt} = \beta(N - I)I$$

Kumamoto U

Susceptible-Infected-recovered (SIR) model

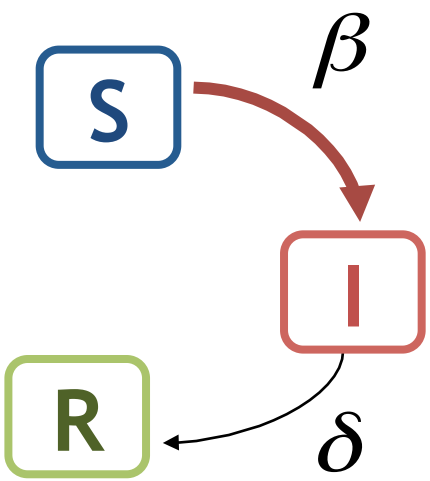
CMU CS

Recovered with immunity

- **S** – **Susceptible (healthy)**
- **I** – **Infected**
- **R** – **Recovered (immune)**

β : Infection rate

δ : Recovery rate



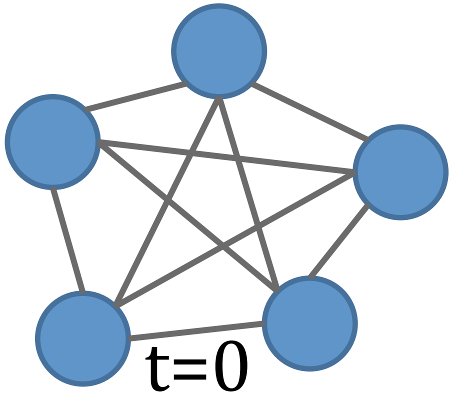
Details

Susceptible-Infected-recovered (SIR) model

CMU CS

Recovered with immunity

S I R



N nodes (healthy)

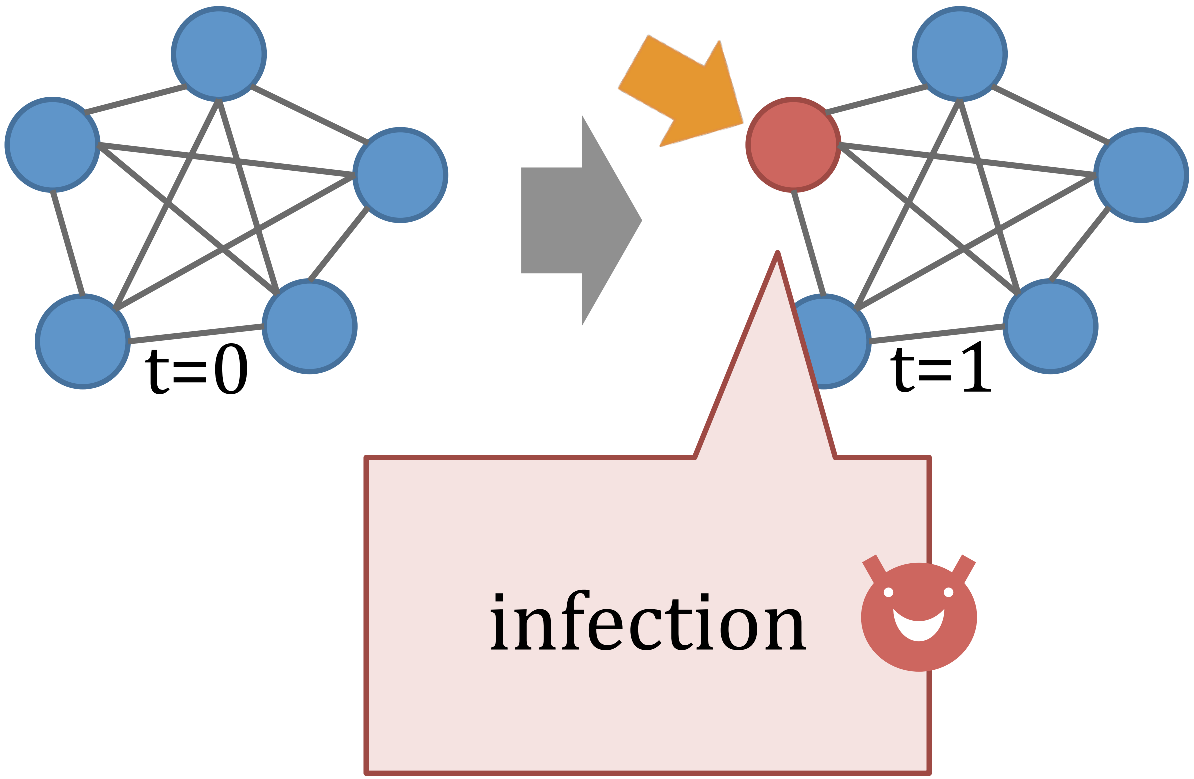
Details

Susceptible-Infected-recovered (SIR) model

CMU CS

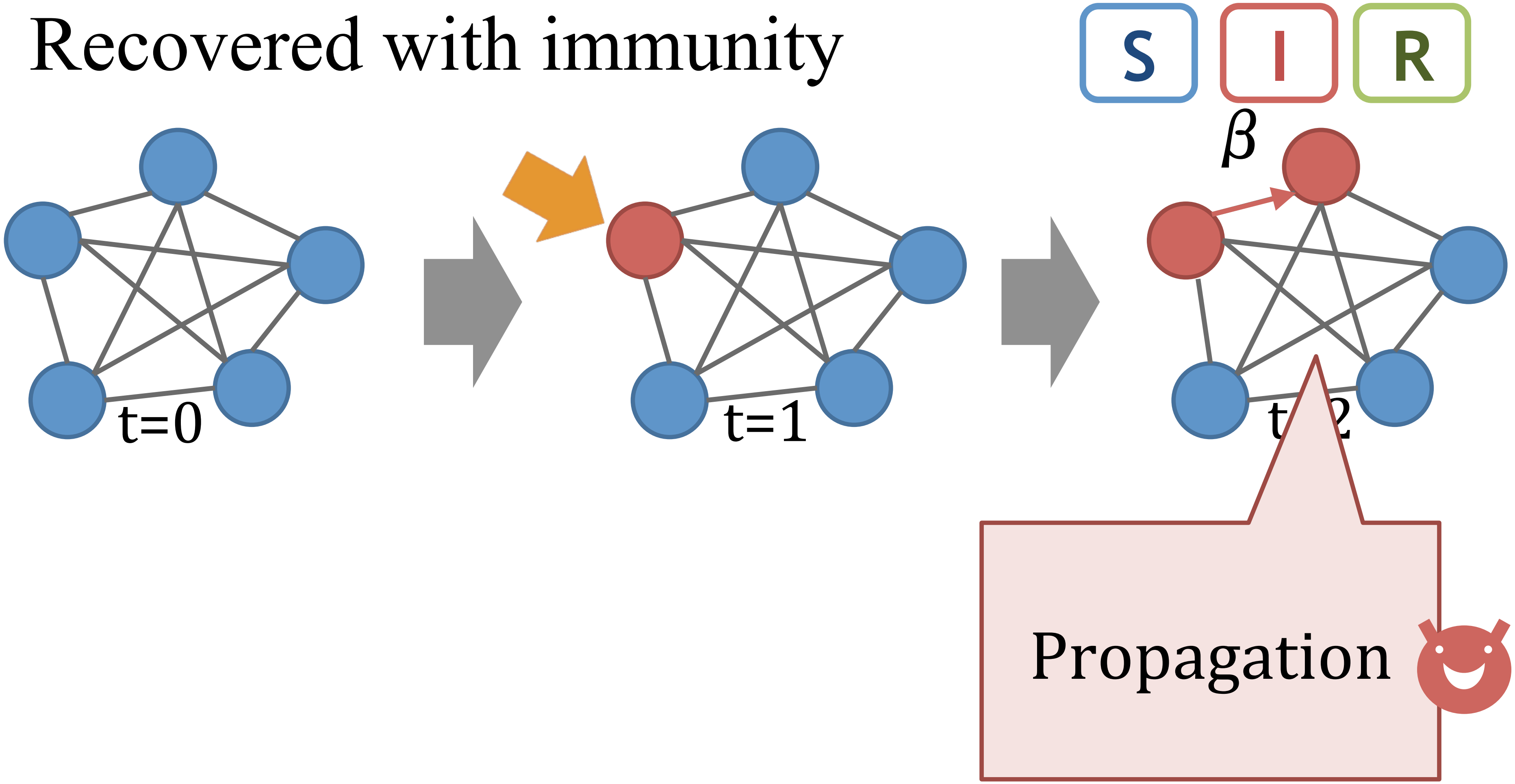
Recovered with immunity

S I R



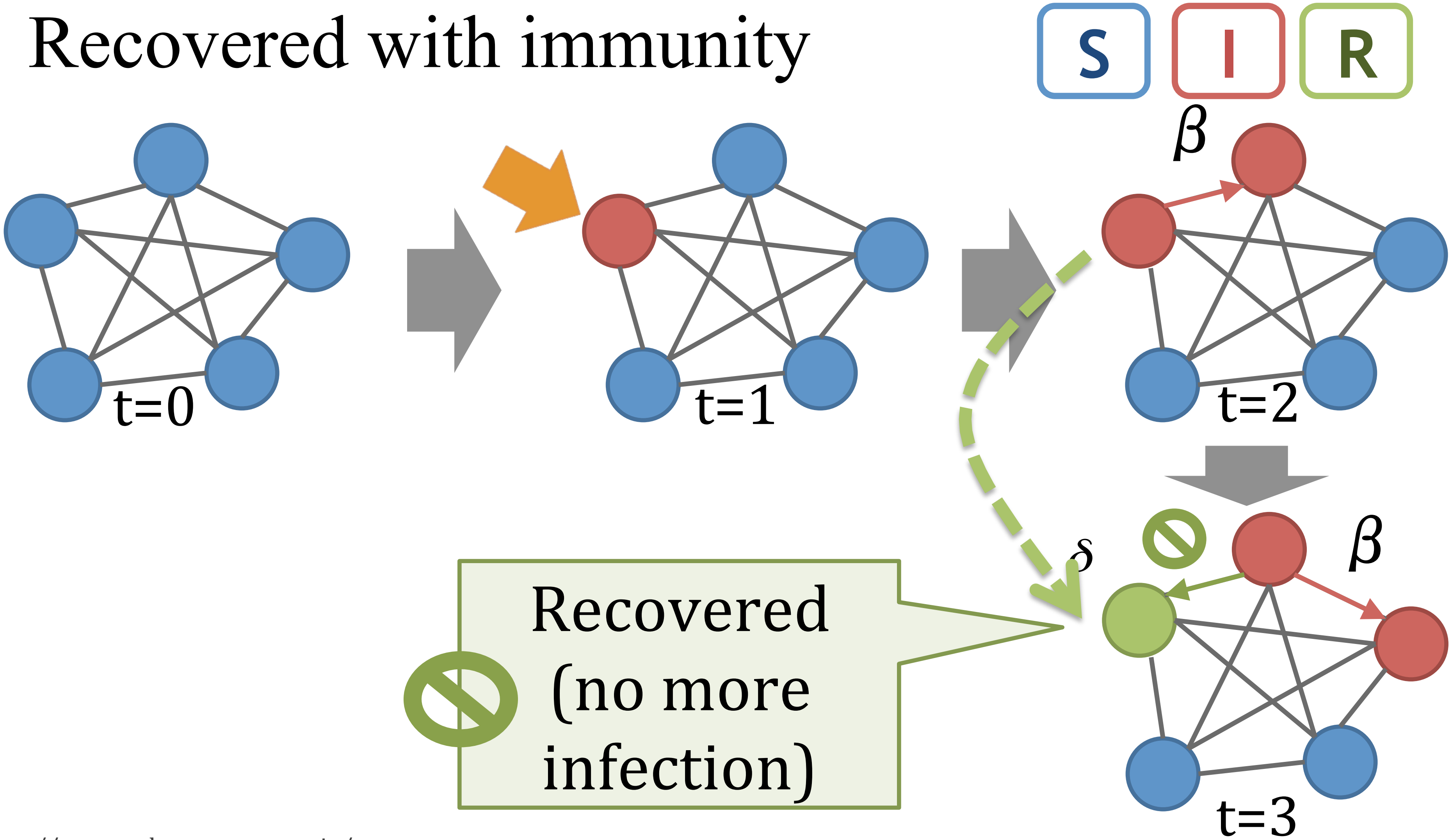
Susceptible-Infected-recovered (SIR) model

Recovered with immunity



Susceptible-Infected-recovered (SIR) model

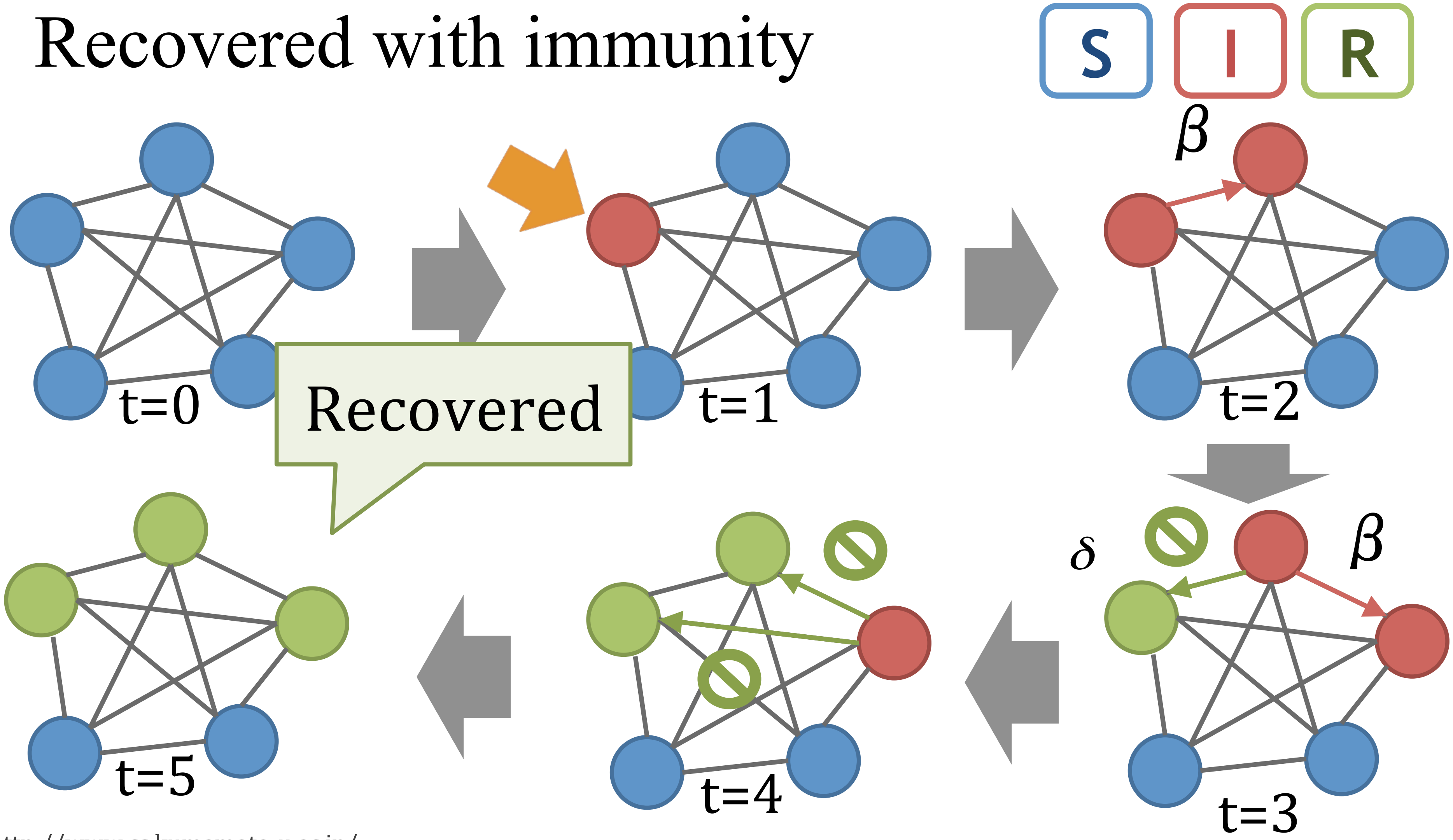
Recovered with immunity



Details

Susceptible-Infected-recovered (SIR) model

Recovered with immunity



Details

Susceptible-Infected-recovered (SIR) model

CMU CS

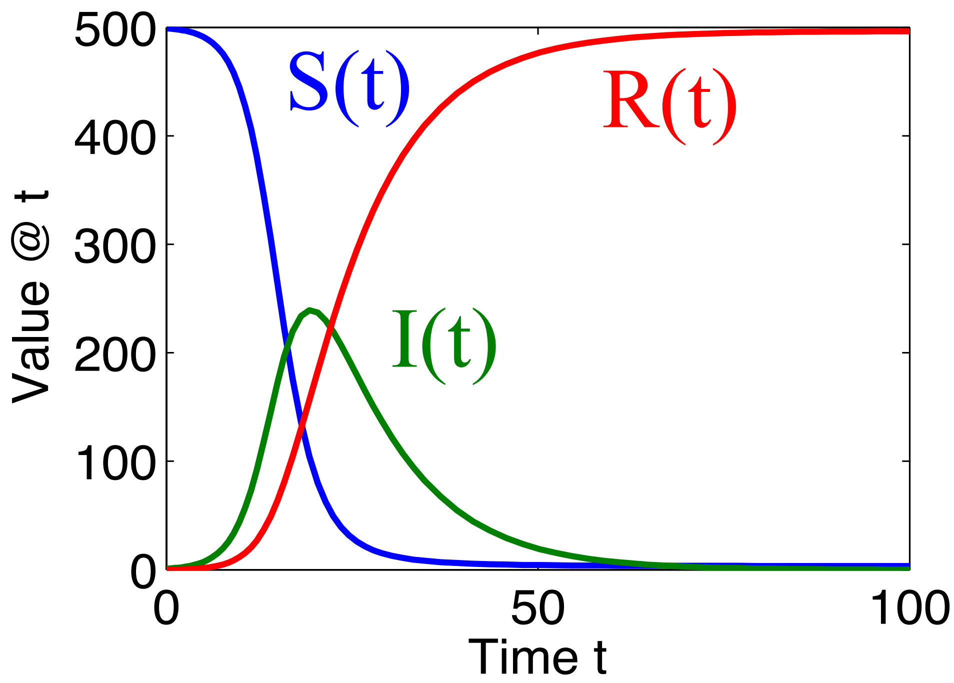
Recovered with immunity

$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$

$$\frac{dI}{dt} = -\frac{\beta SI}{N} - \delta I$$

$$\frac{dR}{dt} = \delta I$$

$$S(t) + I(t) + R(t) = N$$



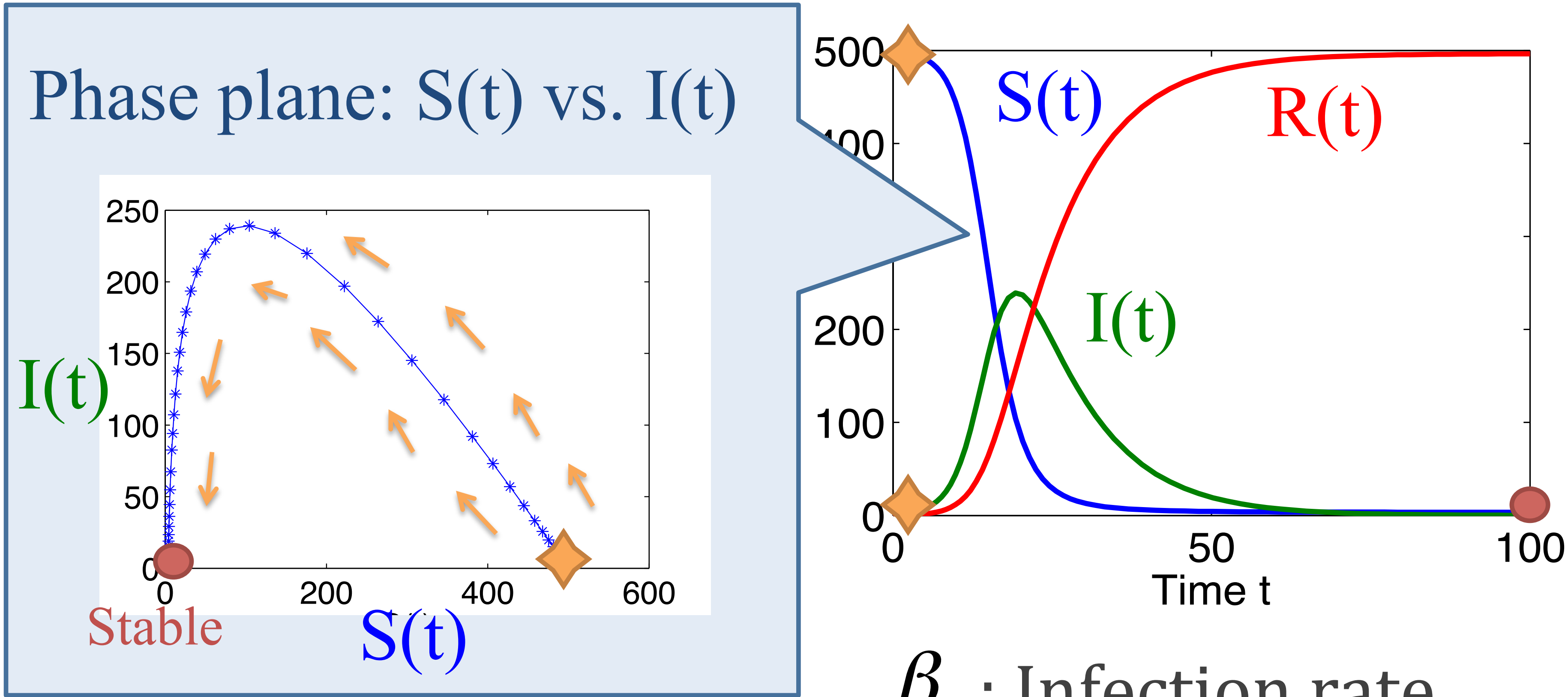
β : Infection rate

δ : Recovery rate

Susceptible-Infected-recovered (SIR) model

CMU CS

Recovered with immunity



β : Infection rate

δ : Recovery rate

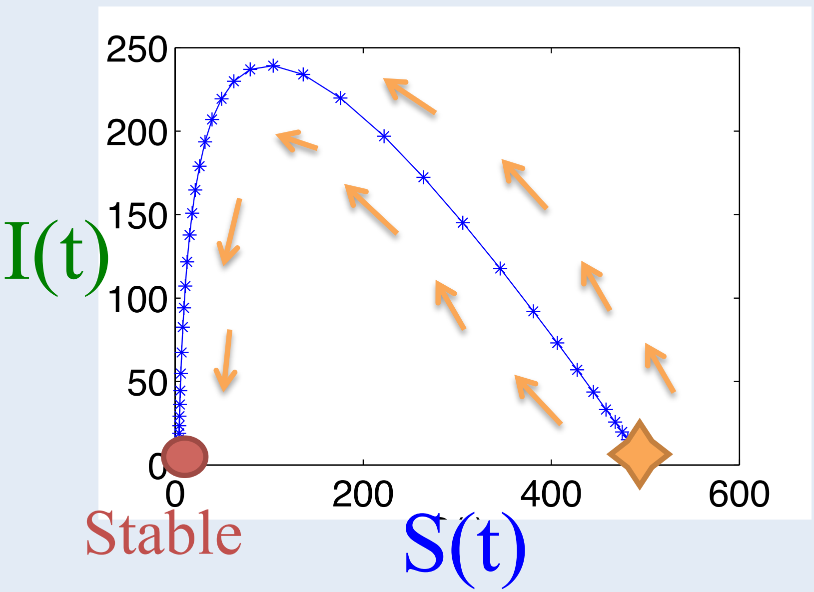
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/

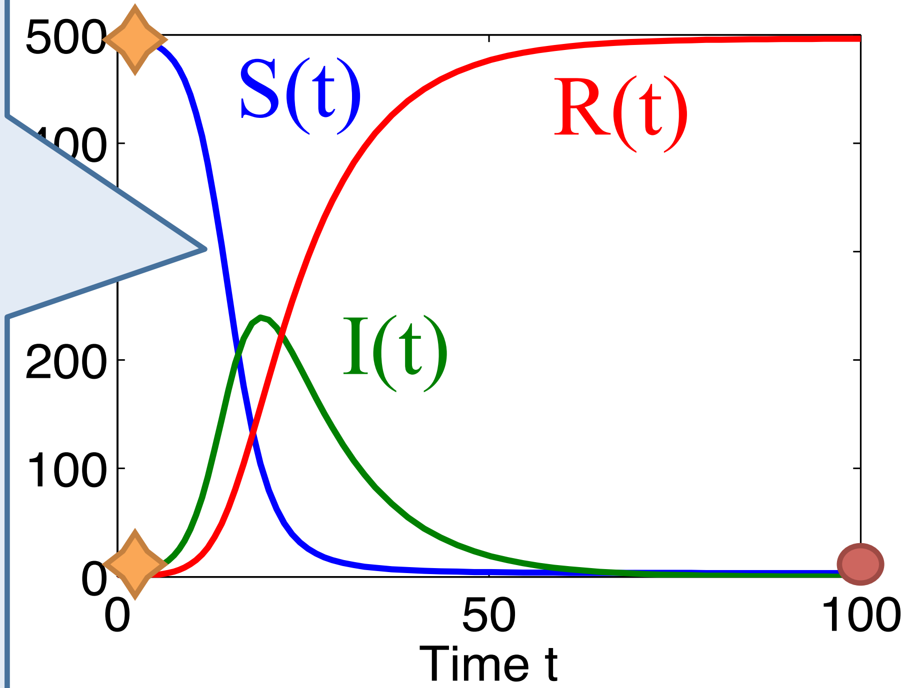
Susceptible-Infected-recovered (SIR) model

Details

Recovered with immunity

Phase plane: S(t) vs. I(t)





β : Infection rate

δ : Recovery rate

Other epidemic models

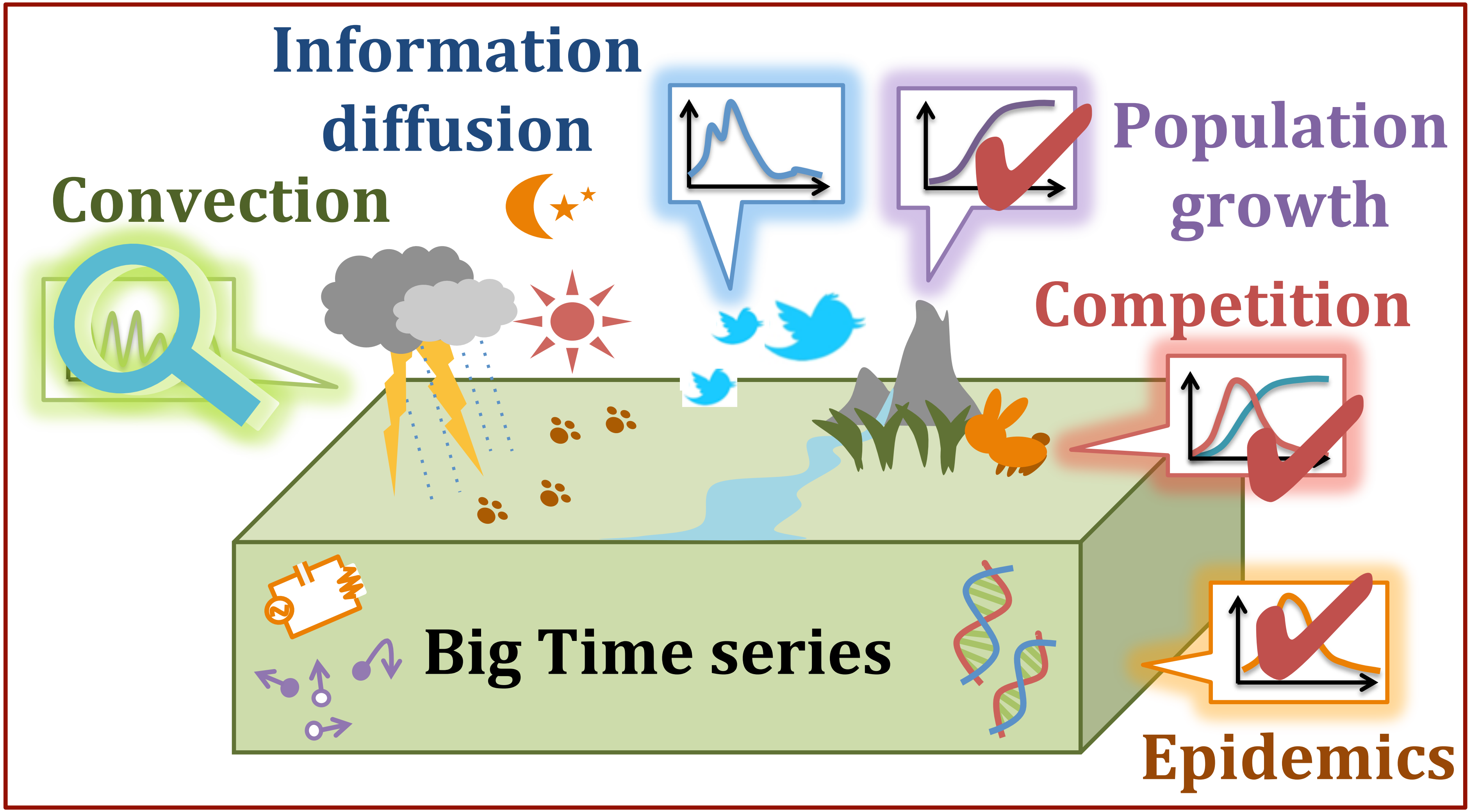
Other virus propagation models ("VPM")

- **SIS** : susceptible-infected-susceptible, flu-like
- **SIRS** : temporary immunity, like pertussis
- **SEIR** : mumps-like, with virus incubation
  (E = Exposed)
- **SEIR-birth/death**: with birth/death rate

Underlying contact-network

– 'who-can-infect-whom'

Grey-box mining and non-linear equations



Other non-linear models

LORENZ: eqs. for atmospheric convection

$$\frac{dx}{dt} = \sigma(y - x)$$

$$\frac{dy}{dt} = x(\rho - z) - y$$

$$\frac{dz}{dt} = xy - \beta z$$

- x: convective intensity
- y: temperature difference between ascending and descending currents
- z: difference in vertical temperature profile from linearity

Other non-linear models

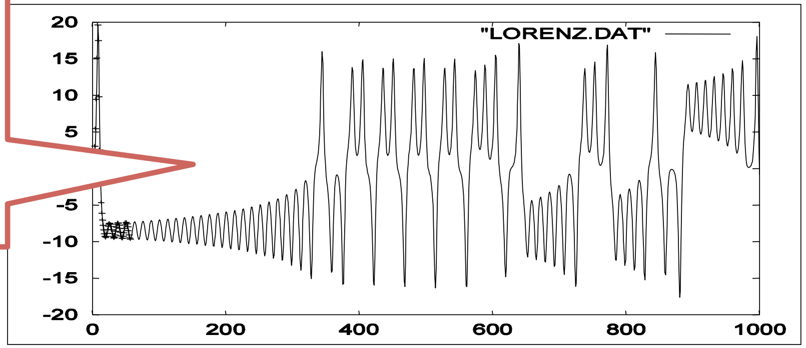
LORENZ: eqs. for atmospheric convection

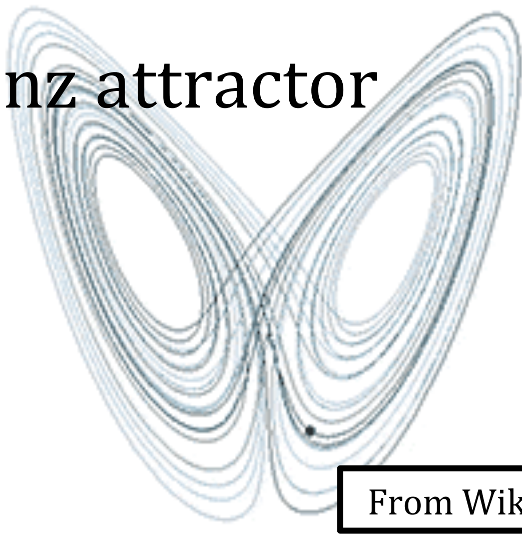
$$\frac{dx}{dt} =$$

$$\frac{dy}{dt} = x(\rho - z) - y$$

$$\frac{dz}{dt} = xy - \beta z$$

Butterfly effect (chaos)





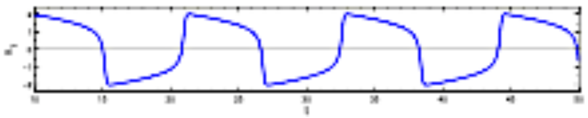
Lorenz attractor

From Wikipedia

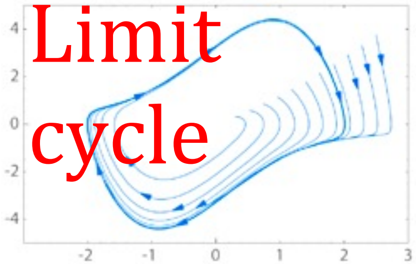
Other non-linear models

From Wikipedia

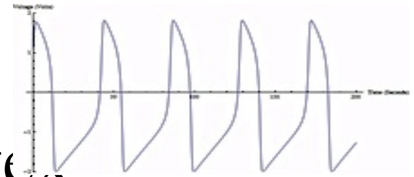
- Van del Pol oscillator
  - Electric circuits, heart-beats, neurons



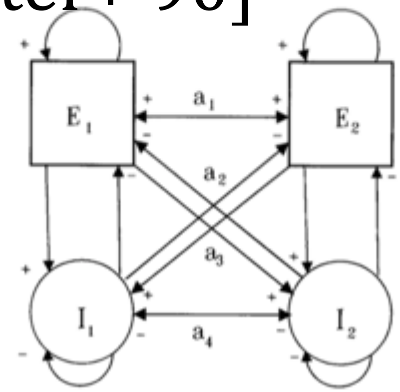
- FitzHugh-Nagumo model
  - An excitable system (e.g., a neuron)



- Excitatory-inhibitory (EI) model
  - Neuronal oscillations in the visual cortex
  - Epilepsy



[Schuster+ 90]



- …
- …

Kumamoto U

CMU CS

Part 2 Roadmap

Problem

- ✓ Why: "non-linear" modeling

Fundamentals

- ✓ Non-linear ("gray-box") models

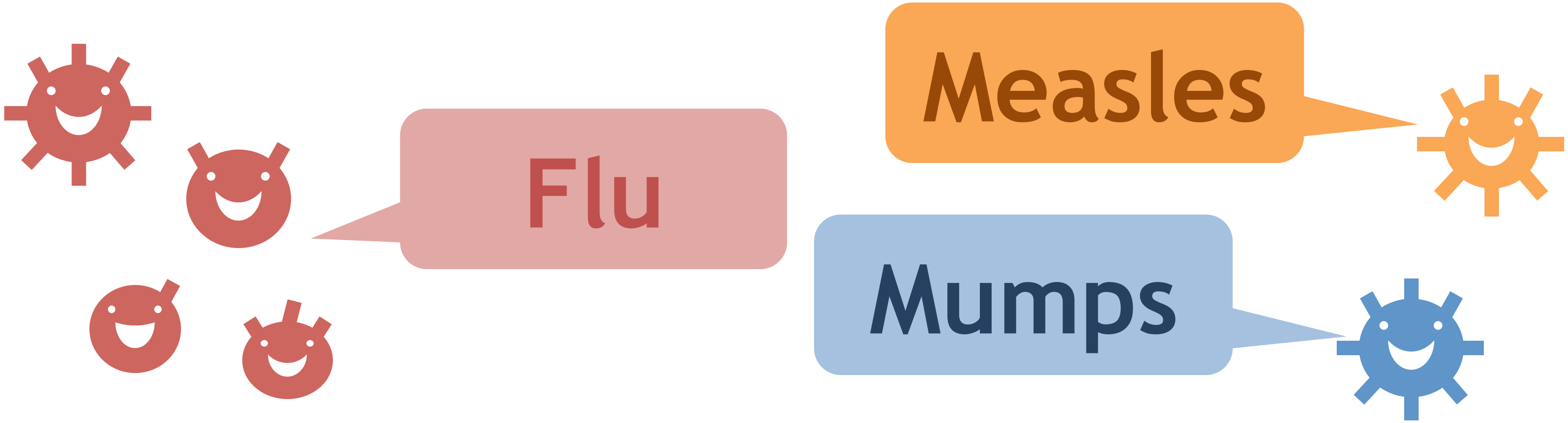
Applications

- Epidemics (skips, competition, "shocks")
- Information diffusion
- Online competition

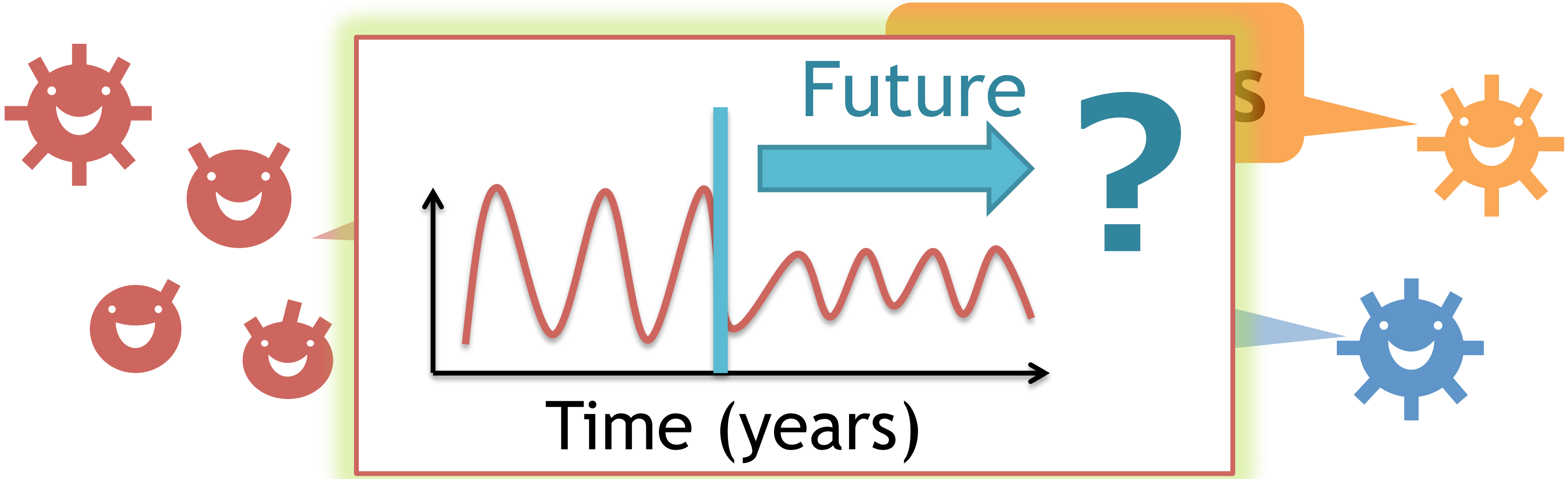
Kumamoto U

Mining and forecasting of co-evolving epidemics

CMU CS

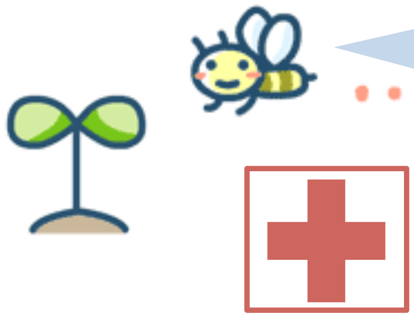


Mining and forecasting of co-evolving epidemics

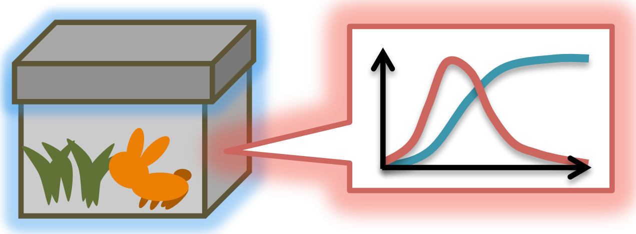


Q. Can we forecast future epidemics?

Epidemics - roadmap



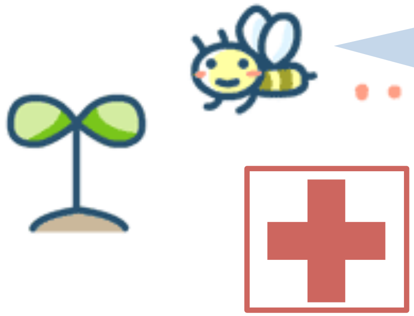
A. Non-linear (gray-box) modeling!



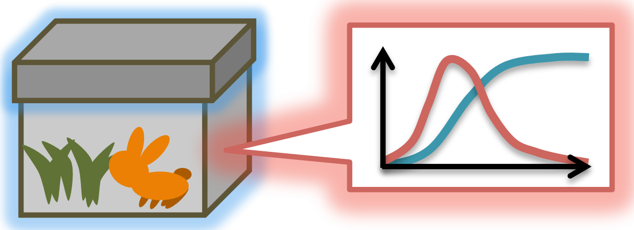
Solutions

- Outbreak vs. Skips [Stone+ Nature'07]
- Interaction between diseases [Rohani+ Nature'03]
- FUNNEL [Matsubara+ KDD'14]

Epidemics - roadmap



A. Non-linear (gray-box) modeling!



Solutions

- **Outbreak vs. Skips [Stone+ Nature'07]**
- Interaction between diseases [Rohani+ Nature'03]
- FUNNEL [Matsubara+ KDD'14]

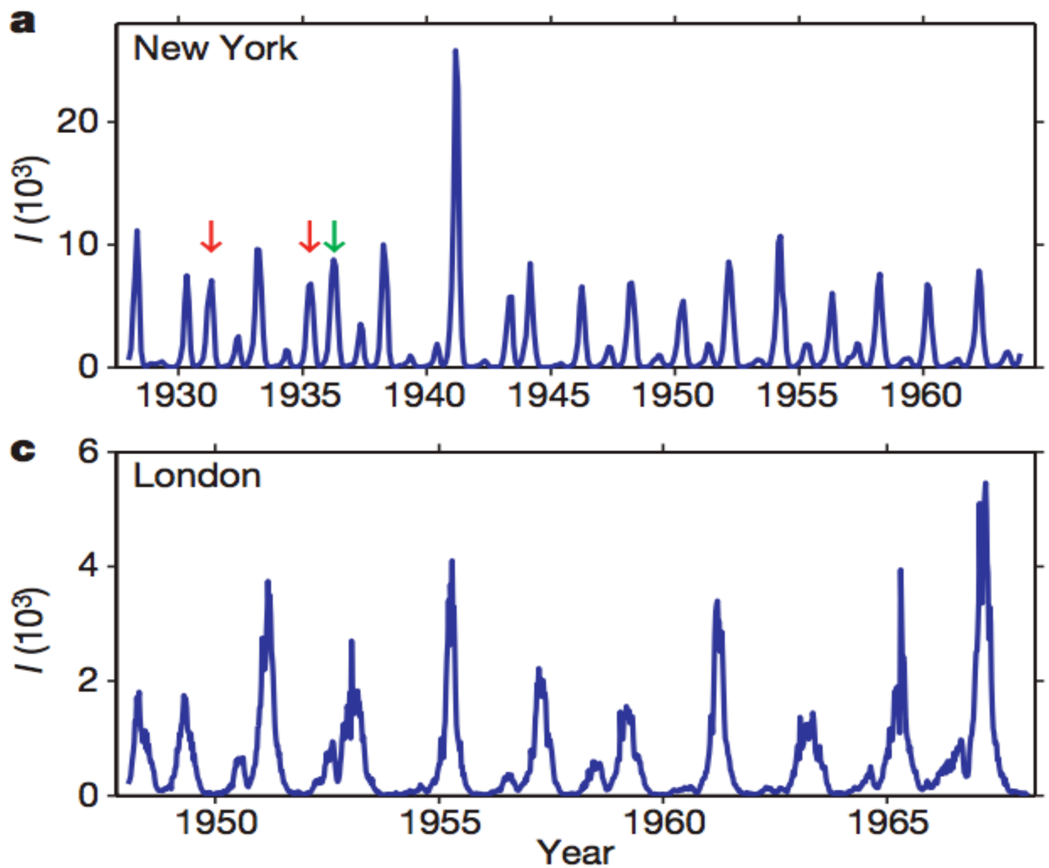
Recurrent epidemics: Outbreak or skip?

[Stone+ Nature'07]

- Time series of reported measles cases

New York

London



Kumamoto U

CMU CS

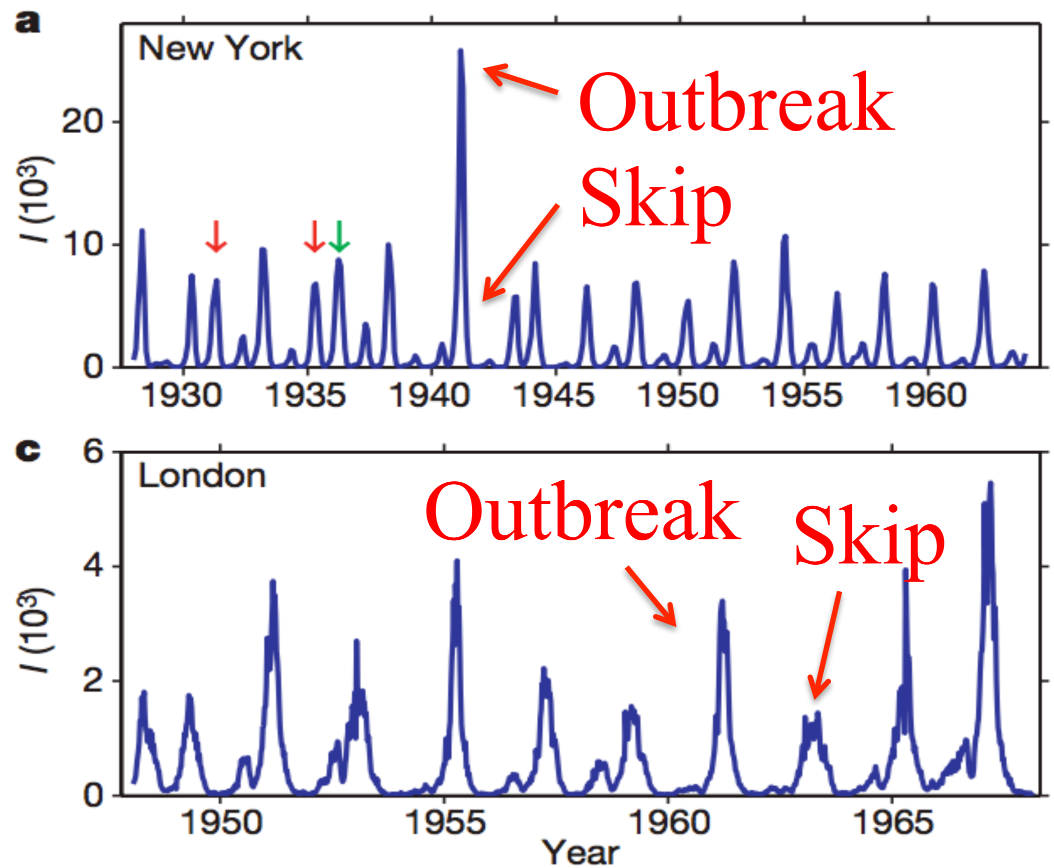
Recurrent epidemics: Outbreak or skip?

[Stone+ Nature'07]

- Time series of reported measles cases

New York

London



Kumamoto U

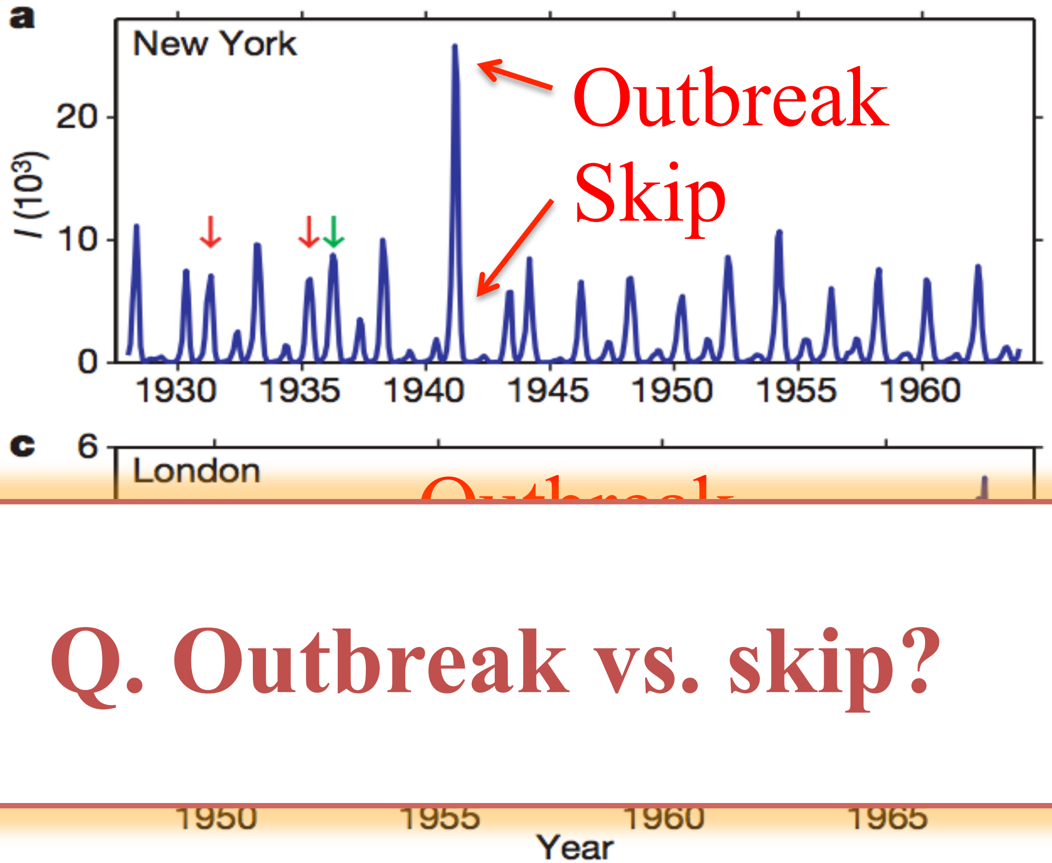
Recurrent epidemics: Outbreak or skip?

CMU CS

[Stone+ Nature'07]

- Time series of reported measles cases

New York



Q. Outbreak vs. skip?

Details

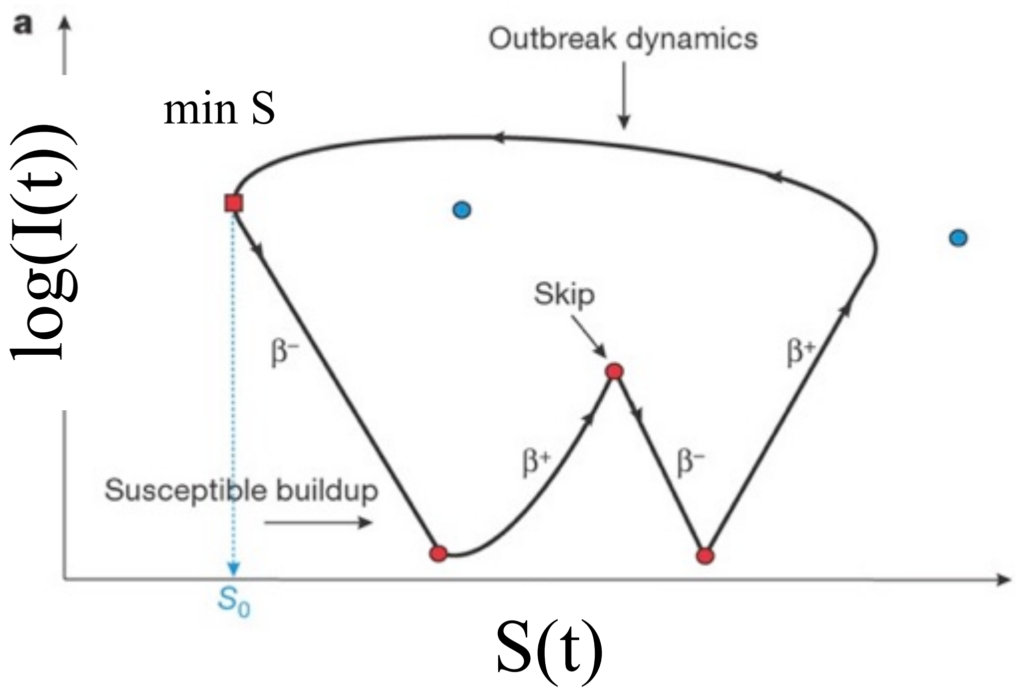
Recurrent epidemics: Outbreak or skip?

CMU CS

[Stone+ Nature'07]

- Conditions for predicting "outbreak vs. skip"
  - SIR model with high/low seasons

Phase plane diagram (S vs. log(I))



Contact rate

β+ : high season

β - : low season

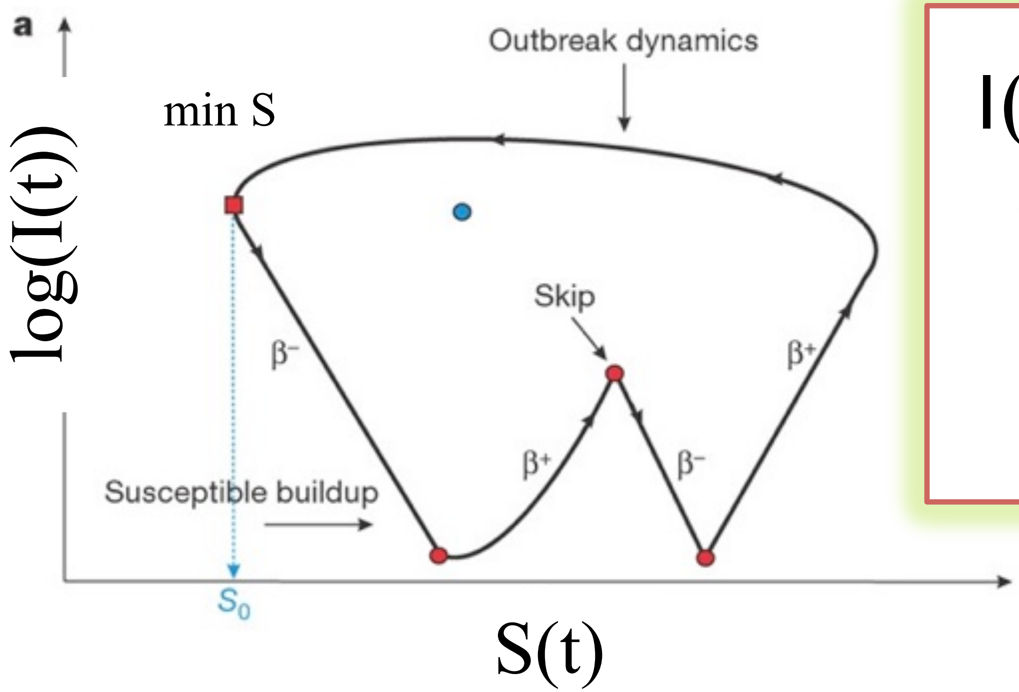
Recurrent epidemics: Outbreak or skip?

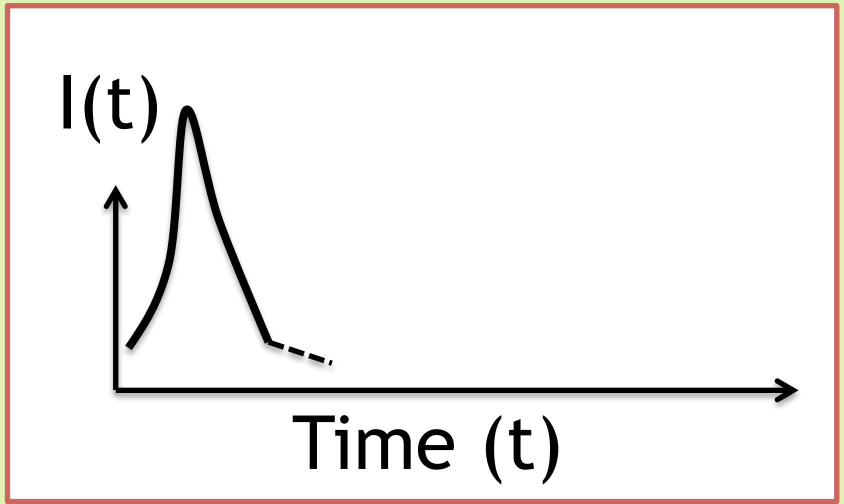
CMU CS

[Stone+ Nature'07]

- Conditions for predicting "outbreak vs. skip"
  - SIR model with high/low seasons

Phase plane diagram (S vs. log(I))





Details

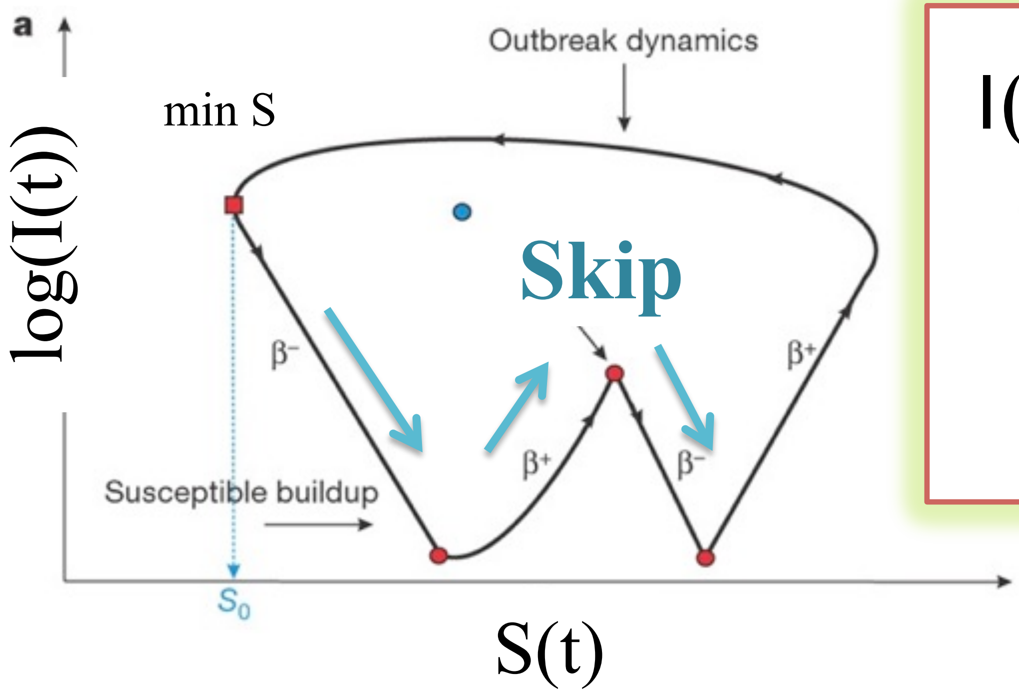
Recurrent epidemics: Outbreak or skip?

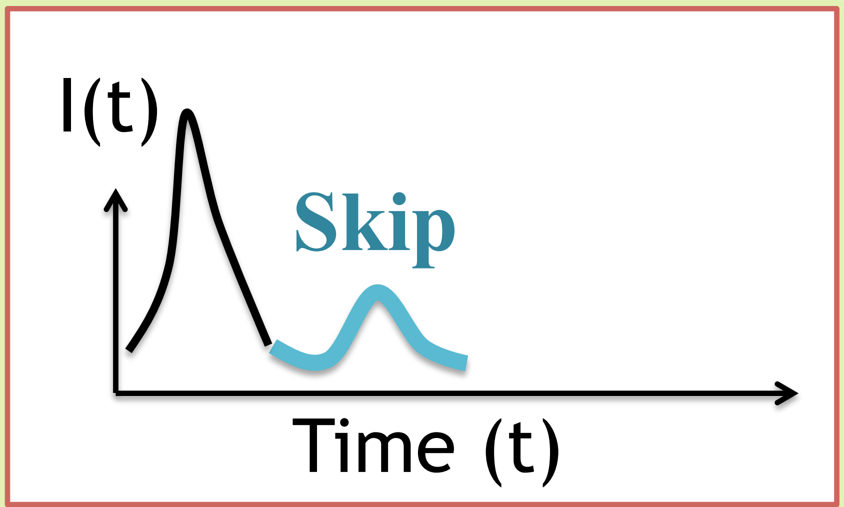
CMU CS

[Stone+ Nature'07]

- Conditions for predicting "outbreak vs. skip"
  - SIR model with high/low seasons

Phase plane diagram (S vs. log(I))



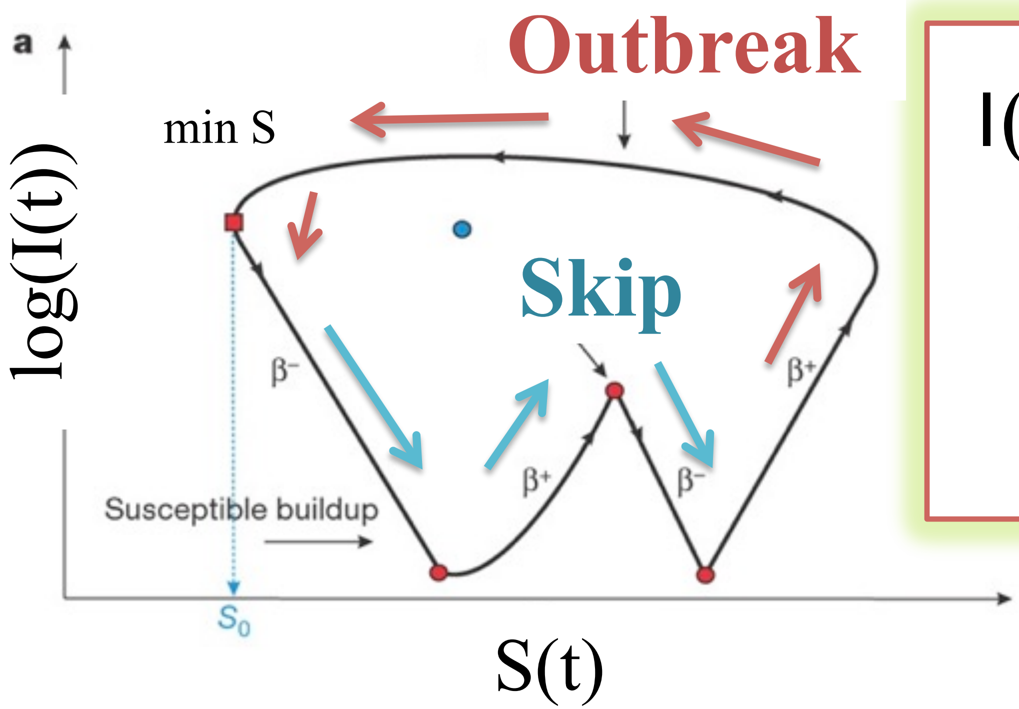


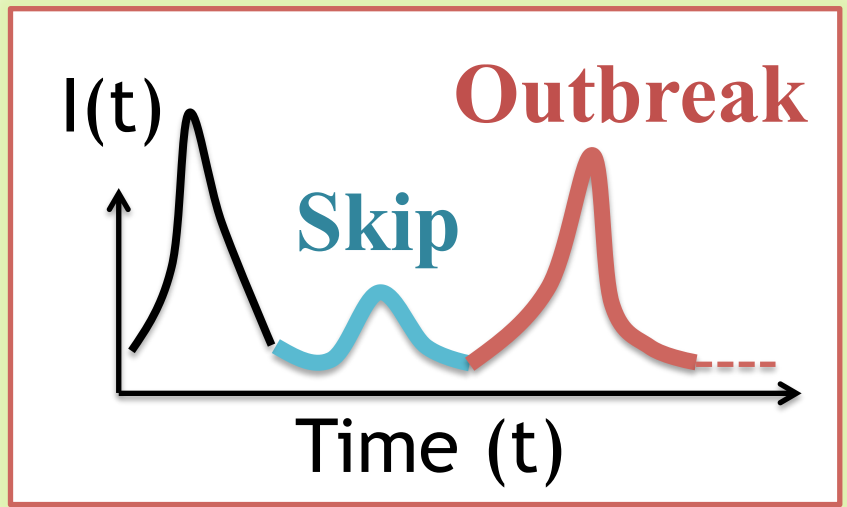
Recurrent epidemics: Outbreak or skip?

[Stone+ Nature'07]

- Conditions for predicting "outbreak vs. skip"
  - SIR model with high/low seasons

Phase plane diagram (S vs. log(I))





Details

Recurrent epidemics: Outbreak or skip?

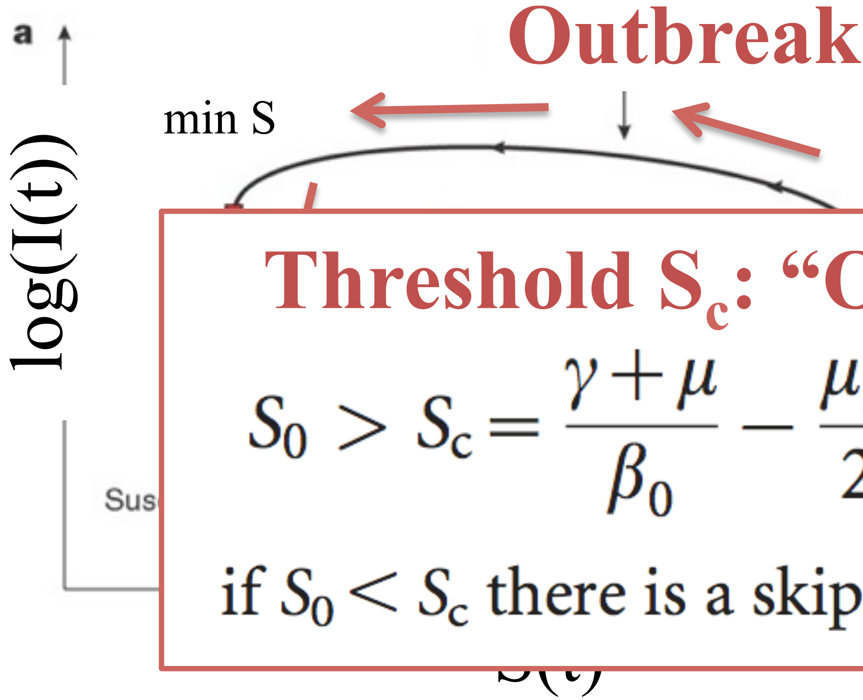
CMU CS

[Stone+ Nature'07]

- Conditions for predicting "outbreak vs. skip"
  - SIR model with high/low seasons

Phase plane diagram (S vs. log(I))

ϒ: recover rate
μ: birth/death rate
β_0:infection rate
χ: time period

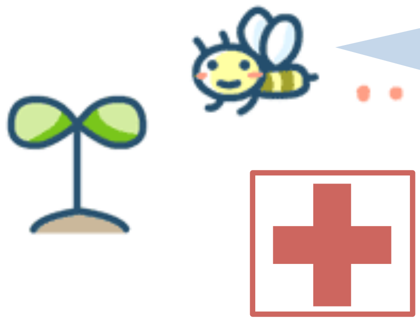


Threshold S_c: "Outbreak vs. Skip"

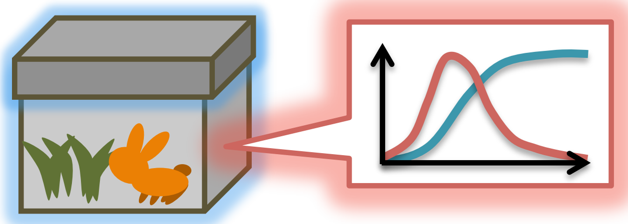
$$S_0 > S_c = \frac{\gamma+\mu}{\beta_0} - \frac{\mu\chi}{2} \quad \Rightarrow \quad \text{epidemic}$$

if $S_0 < S_c$ there is a skip in the following year.

Epidemics - roadmap



A. Non-linear (gray-box) modeling!



Solutions

- Outbreak vs. Skips [Stone+ Nature'07]
- **Interaction between diseases [Rohani+ Nature'03]**
- FUNNEL [Matsubara+ KDD'14]

Ecological interference between fatal diseases

Q. Any relationship (i.e., interaction)

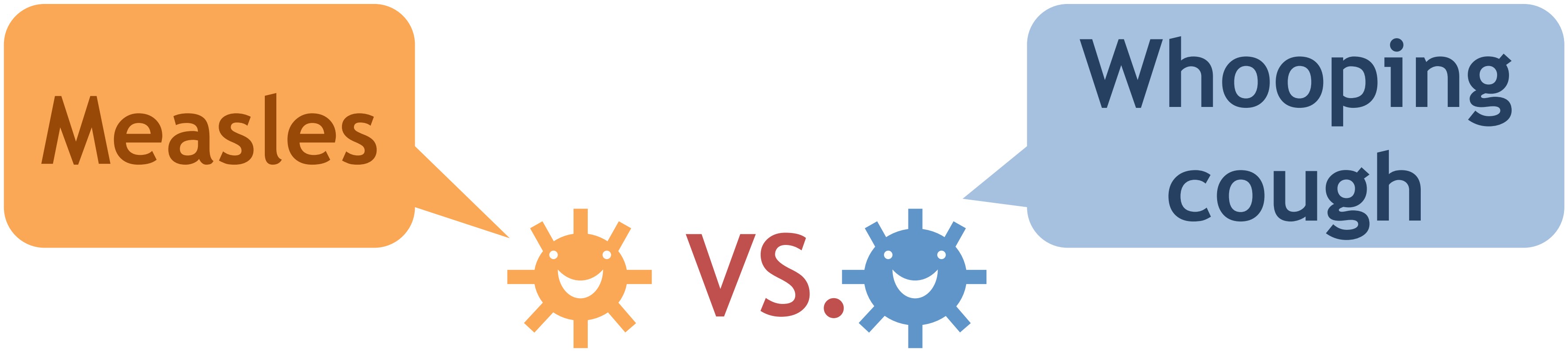
between two different diseases

(e.g., measles vs. whooping cough)?

Ecological interference between fatal diseases

Q. Any relationship (i.e., interaction)
between two different diseases
(e.g., measles vs. whooping cough)?

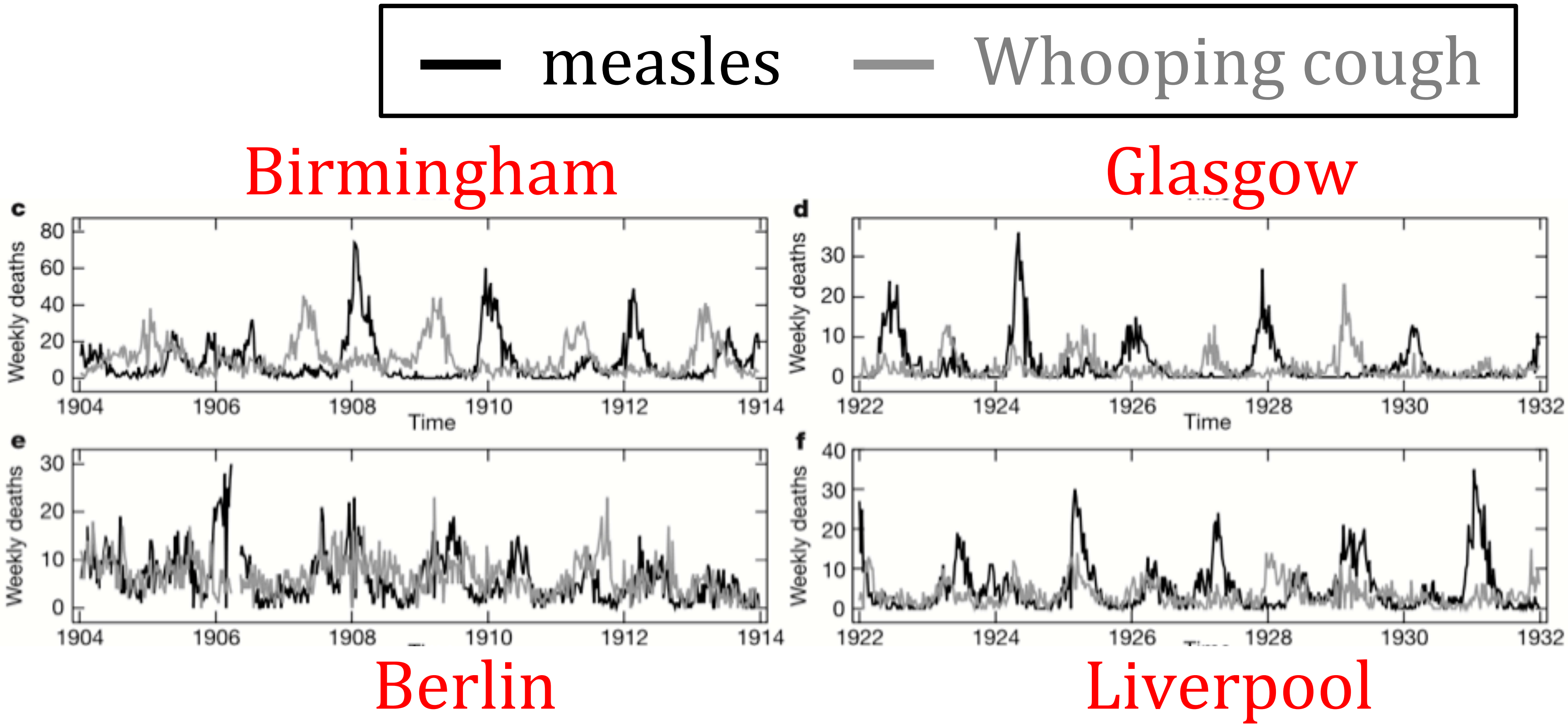
A. Yes. There are "competing" diseases!



Ecological interference between fatal diseases

[Rohani+ Nature'03]

Weekly case fatality reports for two diseases



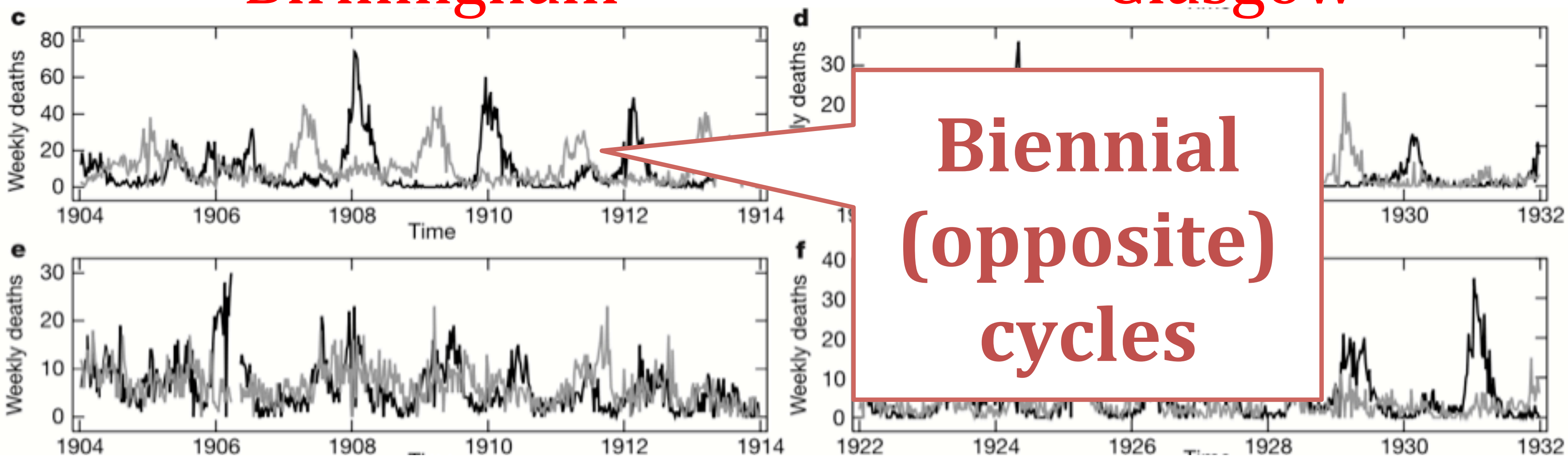
Ecological interference between fatal diseases

[Rohani+ Nature'03]

Weekly case fatality reports for two diseases

— measles — Whooping cough

Birmingham Glasgow

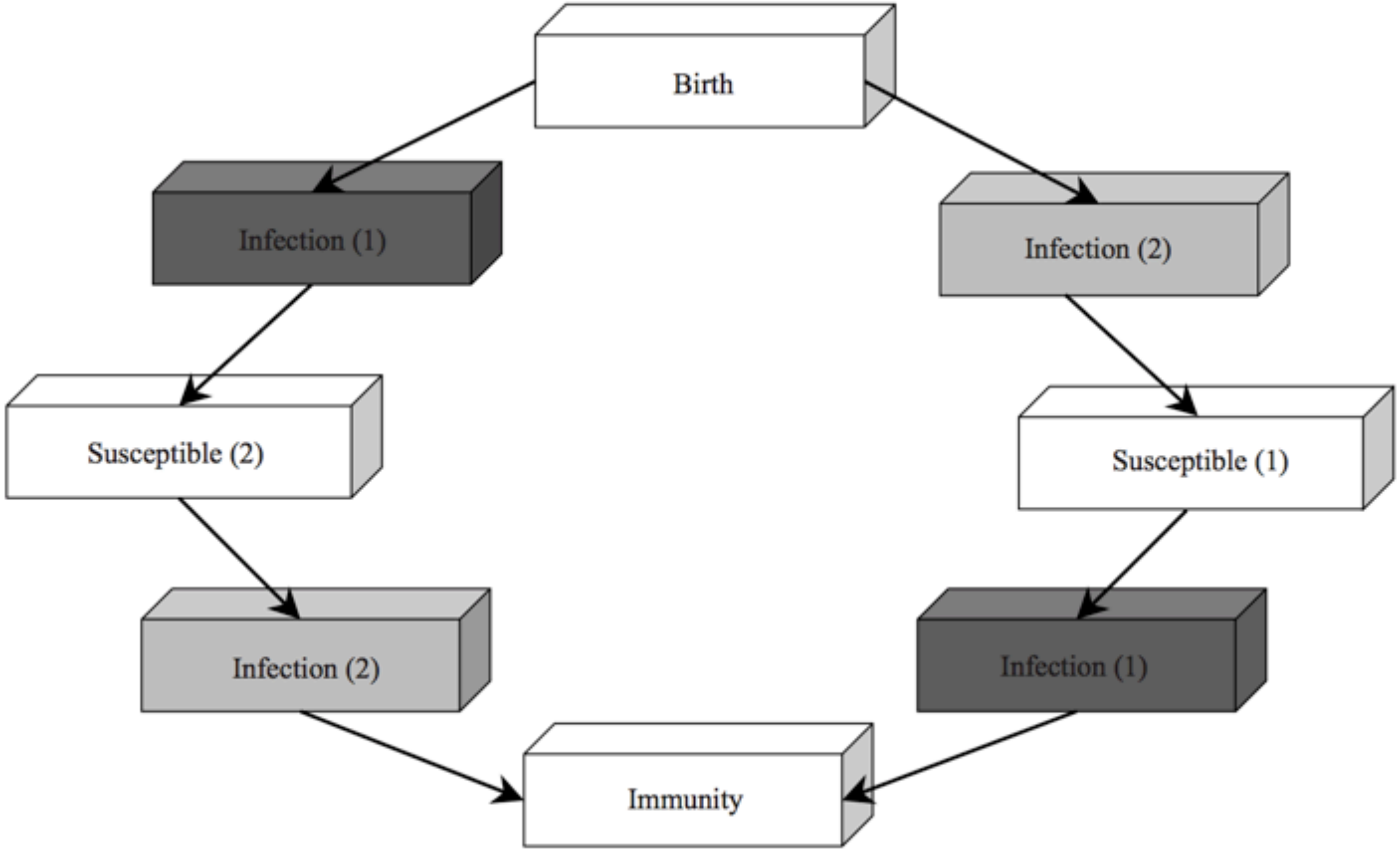


Berlin Liverpool

Details

Ecological interference between fatal diseases

Extension of SIR model [Rohani+'98]

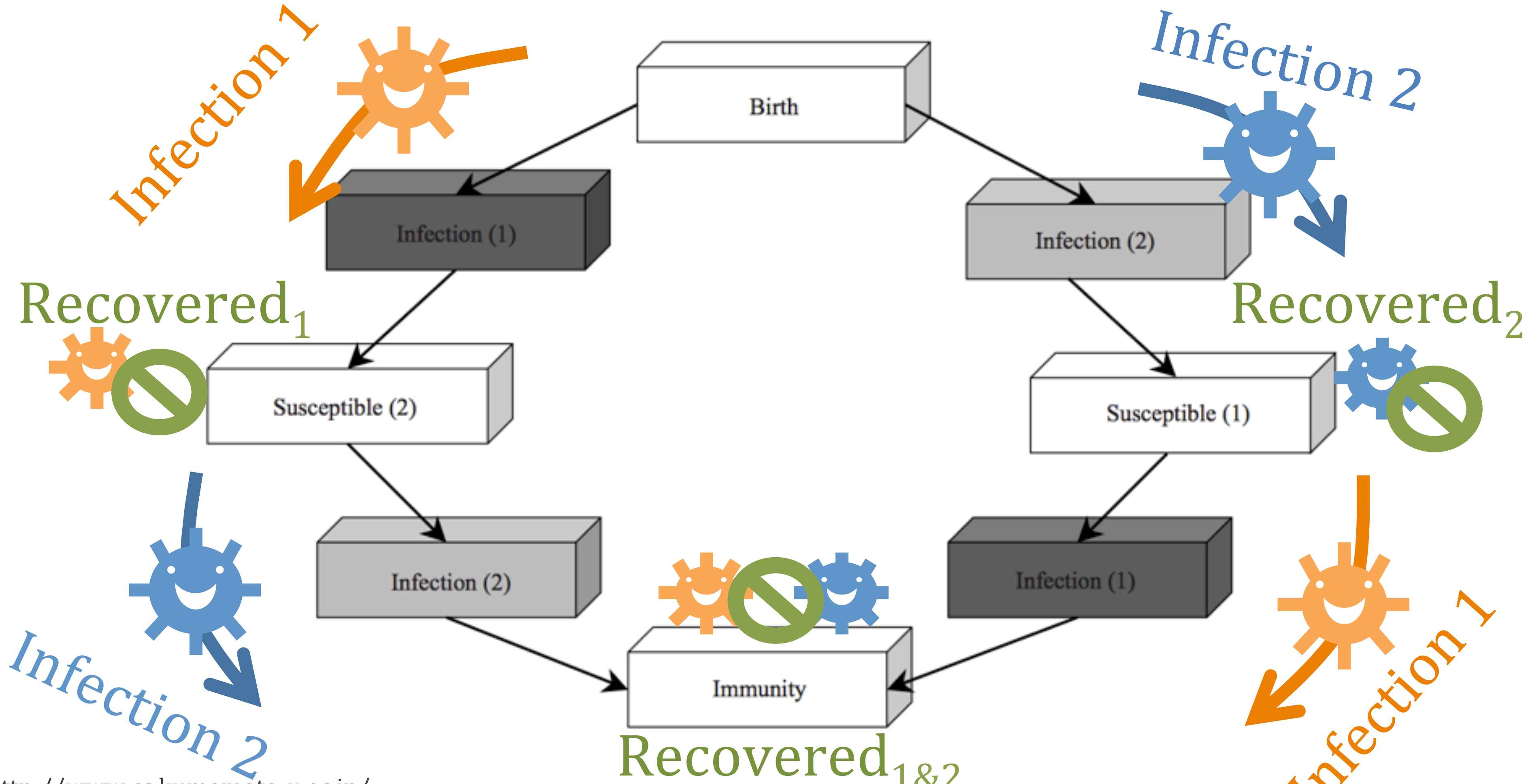


Details

Ecological interference between fatal diseases

CMU CS

Extension of SIR model [Rohani+'98]



http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/

Details

Ecological interference between fatal diseases

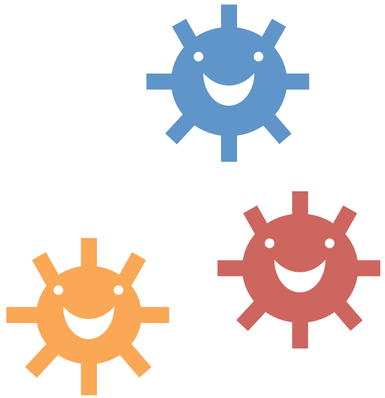
CMU CS

Equations for 3 disease model

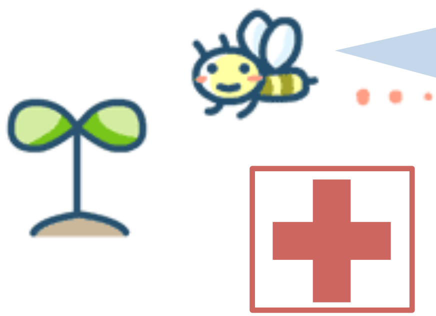
[Rohani+ Nature'03]

$$\begin{aligned}
\frac{dS_{SSS}}{dt} &= \nu N(1-p) - \mu S_{SSS} \\
& - \frac{\beta_1(t)S_{SSS}}{N}(I_{IRR} + I_{IRT} + I_{ITR} + I_{ITT}) \\
& - \frac{\beta_2(t)S_{SSS}}{N}(I_{RIR} + I_{RIT} + I_{TIR} + I_{TIT}) \\
& - \frac{\beta_3(t)S_{SSS}}{N}(I_{RRI} + I_{RTI} + I_{TRI} + I_{TTI}) \\
\frac{dI_{ITT}}{dt} &= \frac{\beta_1(t)S_{SSS}}{N}(I_{IRR} + I_{IRT} + I_{ITR} + I_{ITT}) \\
& - (\mu + \gamma_1)I_{ITT} \\
\frac{dI_{IRT}}{dt} &= \frac{\beta_1(t)S_{SRS}}{N}(I_{IRR} + I_{IRT} + I_{ITR} + I_{ITT}) \\
& - (\mu + \gamma_1)I_{IRT}
\end{aligned}$$

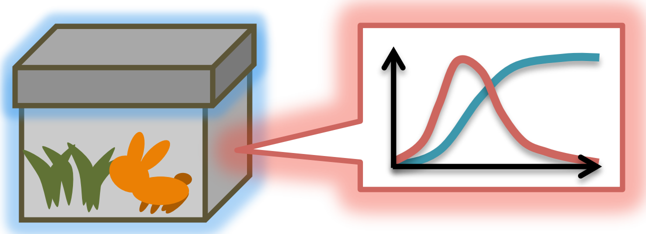
...



Epidemics - roadmap



Non-linear (gray-box) modeling!



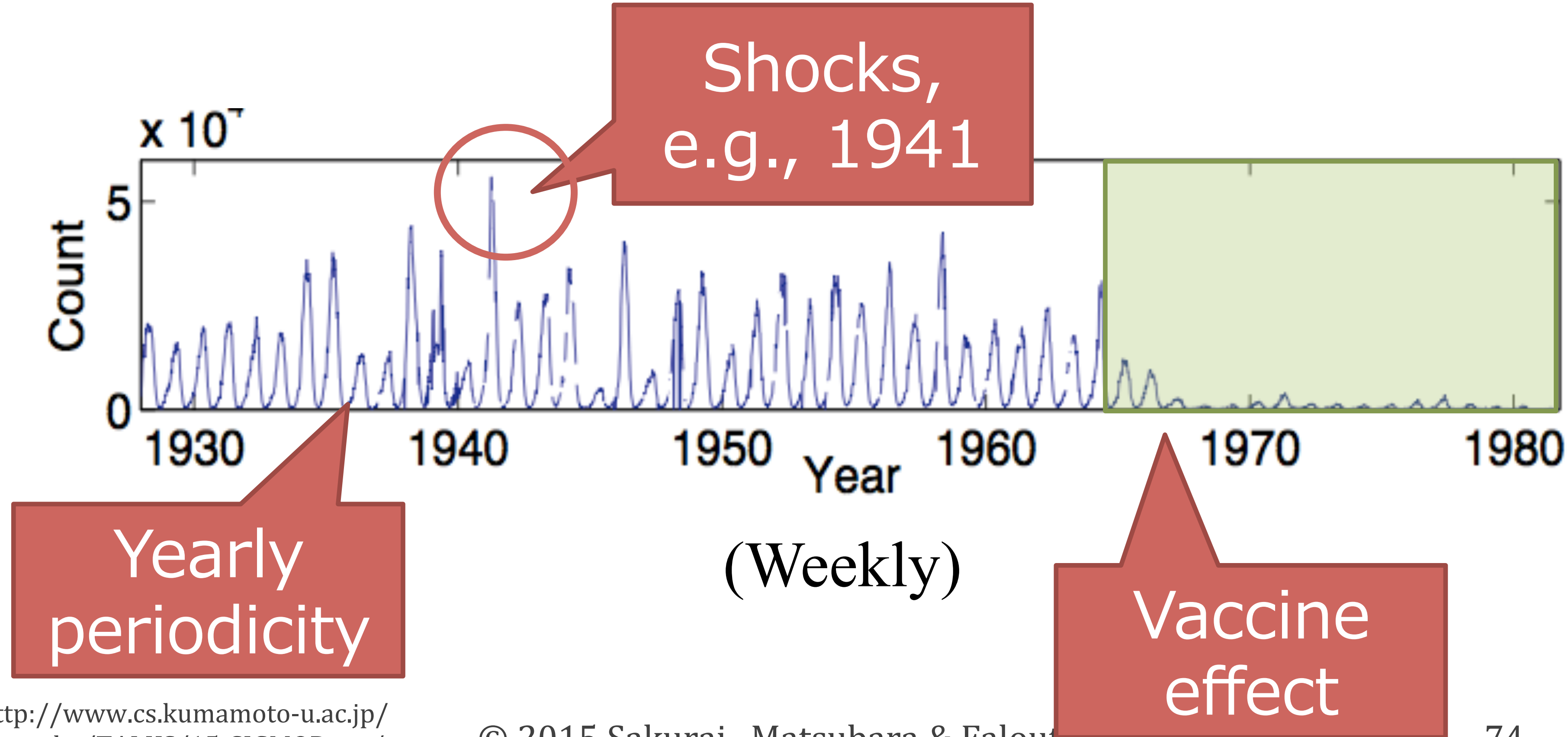
Solutions

- E1. Outbreak vs. Skips [Stone+ Nature'07]
- E2. Interaction between diseases [Rohani+ Nature'03]
- **E3. FUNNEL [Matsubara+ KDD'14]**

FUNNEL with a single epidemic

[Matsubara+ KDD'14]

e.g., Measles cases in the U.S.



(Weekly)

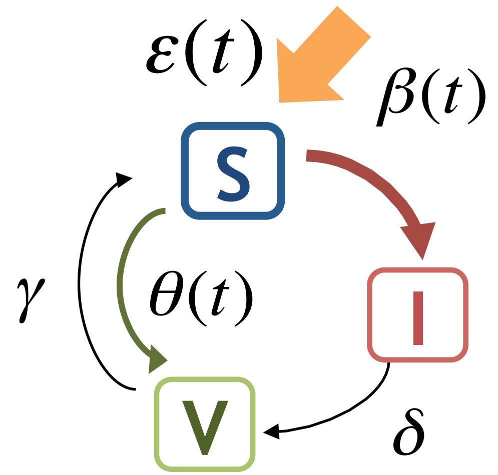
FUNNEL with a single epidemic

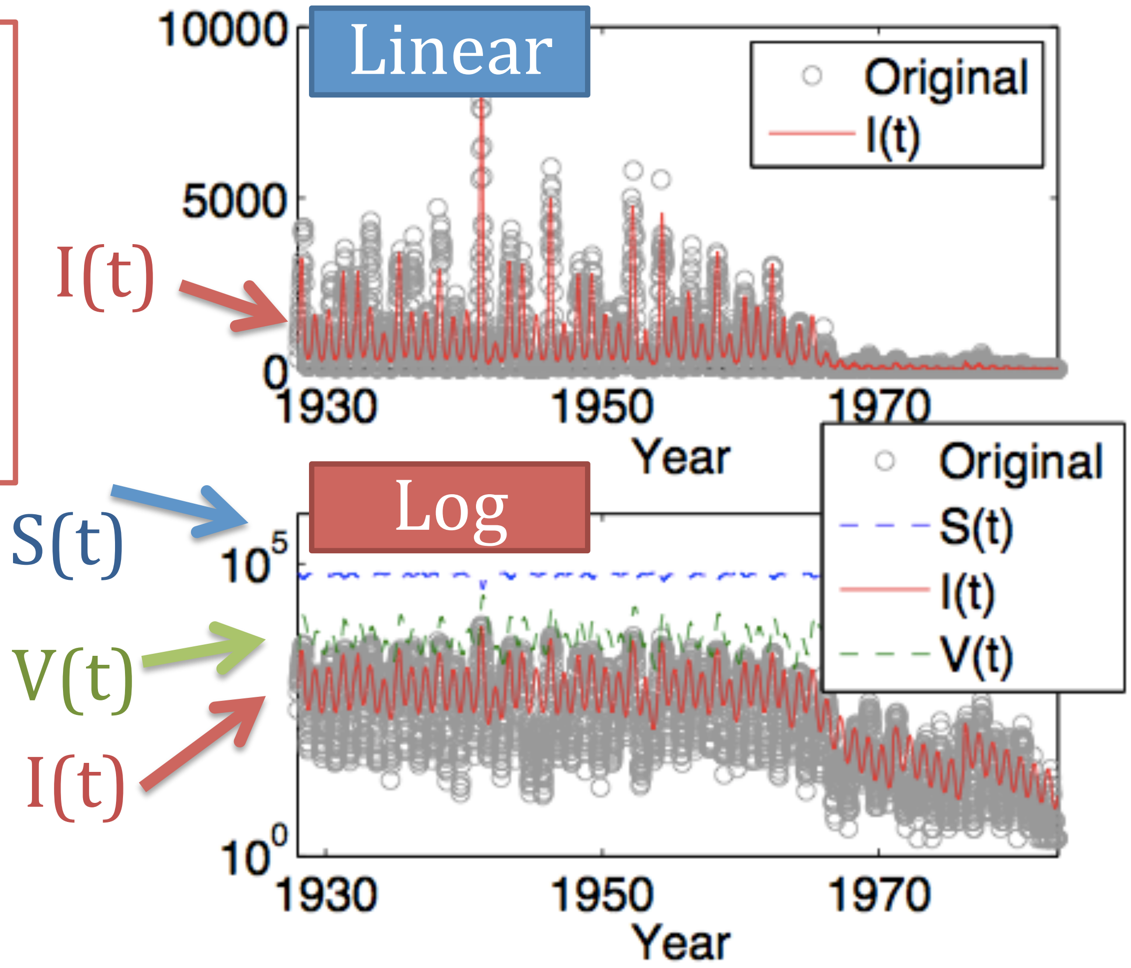
[Matsubara+ KDD'14]

With a single epidemic: Funnel-RE

People of 3 classes

- **S** : Susceptible
- **I** : Infected
- **V** : Vigilant/ vaccinated





Kumamoto U

FUNNEL with a single epidemic

[Matsubara+ KDD'14]

CMU CS

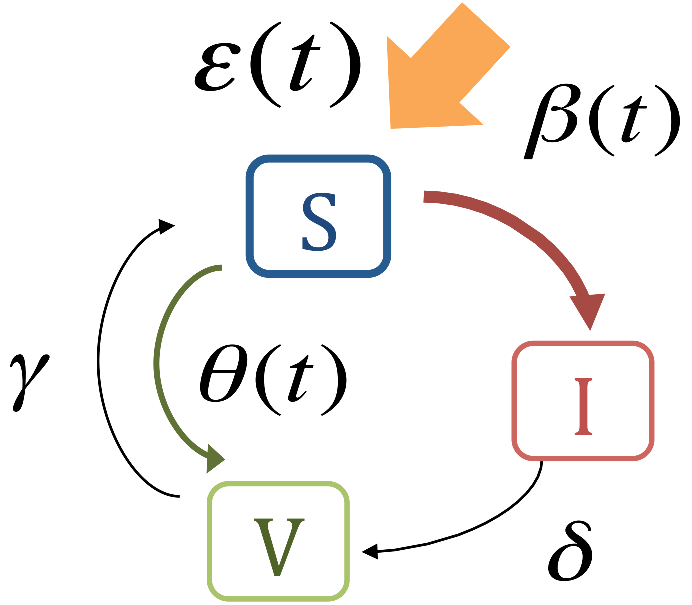
With a single epidemic: Funnel-RE

$$
\begin{aligned}
S(t+1) &= S(t) - \beta(t)\epsilon(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t) \\
I(t+1) &= I(t) + \beta(t)\epsilon(t)S(t)I(t) - \delta I(t) \\
V(t+1) &= V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t) \qquad (3)
\end{aligned}
$$

S(t) : susceptible

I (t) : Infected

V(t) : Vigilant /Vaccinated



Kumamoto U

FUNNEL with a single epidemic

[Matsubara+ KDD'14]

CMU CS

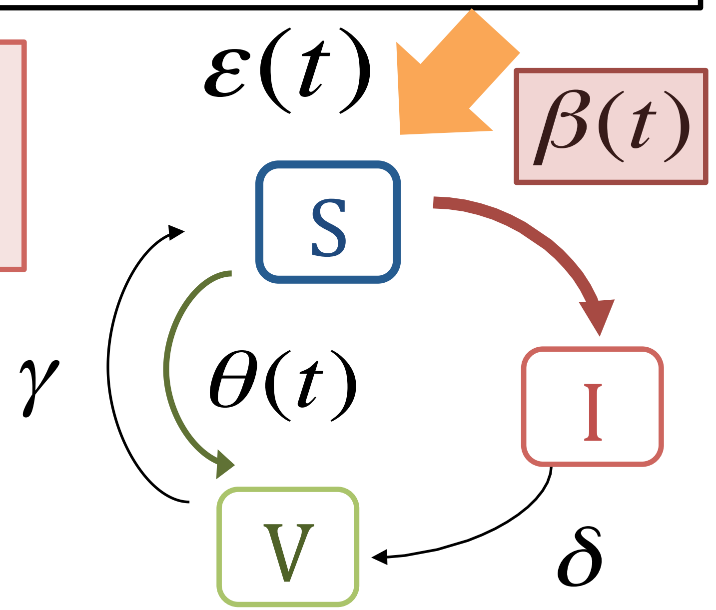
With a single epidemic: Funnel-RE

$$\begin{aligned} S(t+1) &= S(t) - \beta(t)\epsilon(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t) \\ I(t+1) &= I(t) + \beta(t)\epsilon(t)S(t)I(t) - \delta I(t) \\ V(t+1) &= V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t) \qquad (3) \end{aligned}$$

$\beta(t)$: strength of infection (yearly periodic func)

$$\beta(t) = \beta_0 \cdot \left(1 + P_a \cdot cos\left(\frac{2\pi}{P_p}(t + P_s)\right)\right)$$

$$P_p = 52$$



FUNNEL with a single epidemic

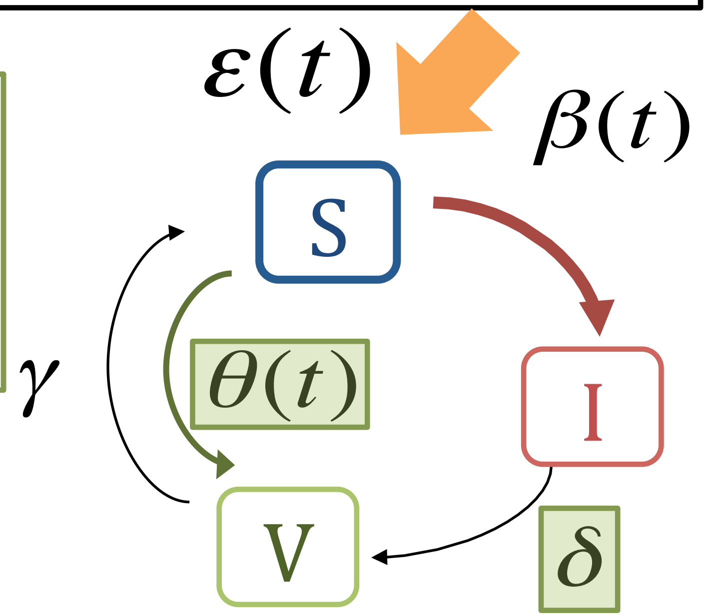
[Matsubara+ KDD'14]

With a single epidemic: Funnel-RE

$$
\begin{aligned}
S(t+1) &= S(t) - \beta(t)\epsilon(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t) \\
I(t+1) &= I(t) + \beta(t)\epsilon(t)S(t)I(t) - \delta I(t) \\
V(t+1) &= V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t) \qquad (3)
\end{aligned}
$$

δ : healing rate
$\theta(t)$: disease reduction effect

$$
\theta(t) = \begin{cases} 0 & (t < t_\theta) \\ \theta_0 & (t \geq t_\theta) \end{cases}
$$



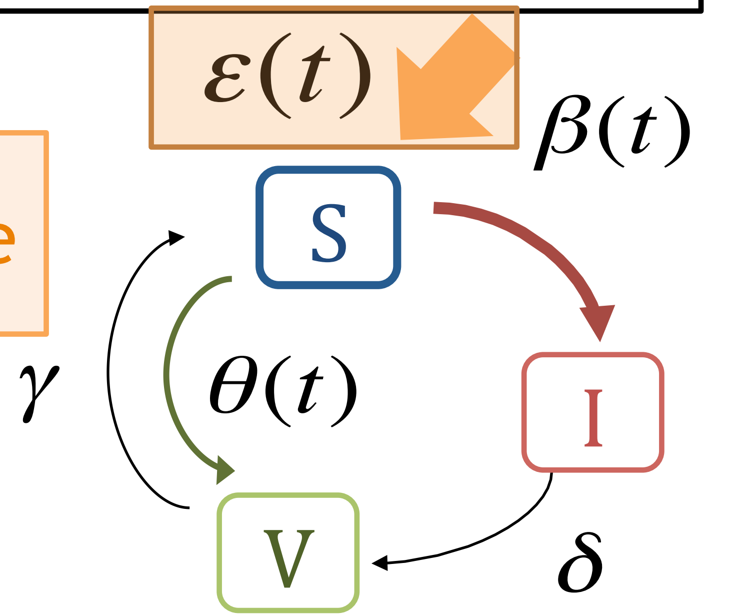
FUNNEL with a single epidemic

[Matsubara+ KDD'14]

With a single epidemic: Funnel-RE

$$
\begin{aligned}
S(t+1) &= S(t) - \beta(t)\epsilon(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t) \\
I(t+1) &= I(t) + \beta(t)\epsilon(t)S(t)I(t) - \delta I(t) \\
V(t+1) &= V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t) \qquad (3)
\end{aligned}
$$

$\varepsilon(t)$: temporal susceptible rate



FUNNEL with a single epidemic

[Matsubara+ KDD'14]

Kumamoto U

CMU CS

With a single epidemic: Funnel-RE

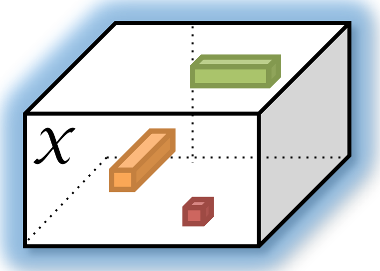
$S(t+1) = S(t) - \beta(t)\epsilon(t)S($ ... $(t) - \theta(t)S(t)$

$I(t+1) = I(t)$ ... $I(t) - \delta I(t)$

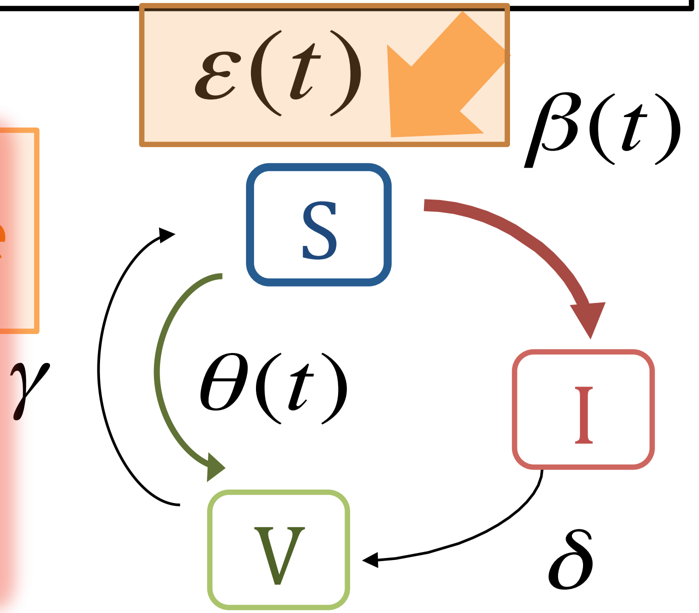
$V(t+1)$ ... $I(t) - \gamma V(t) + \theta(t)S(t)$ (3)

FUNNEL: Details @ part3

$\varepsilon(t)$: tem ... e



+ tensor analysis



Kumamoto U

CMU CS

Part 2 Roadmap

Problem

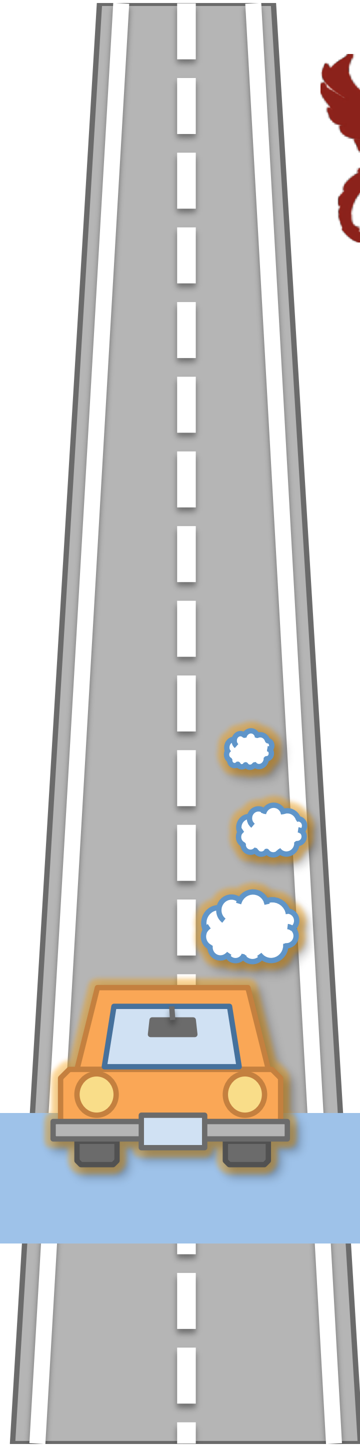
- ✓ Why: "non-linear" modeling

Fundamentals

- ✓ Non-linear (grey-box) models

Applications

- ✓ Epidemics
- Information diffusion
- Online competition



Kumamoto U

Information diffusion in social networks

CMU CS

Kumamoto U

Information diffusion in social networks

CMU CS

Q. How news/rumors spread in social media?

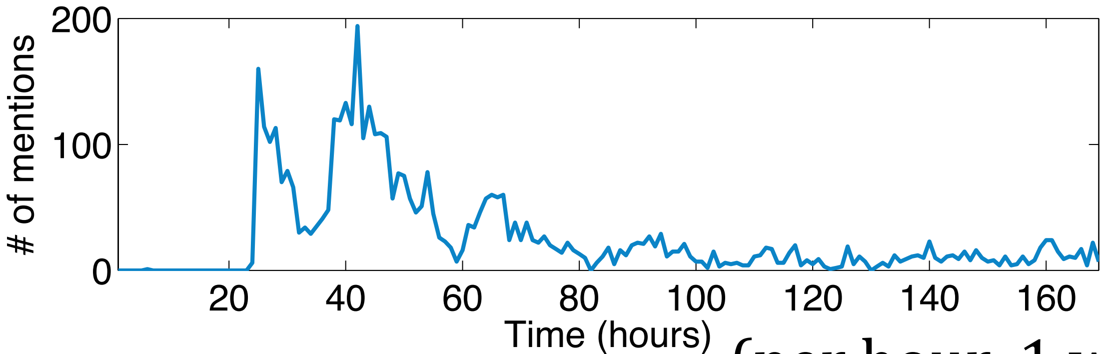
News spread in social media

MemeTracker [Leskovec+ KDD'09]

MemeTracker

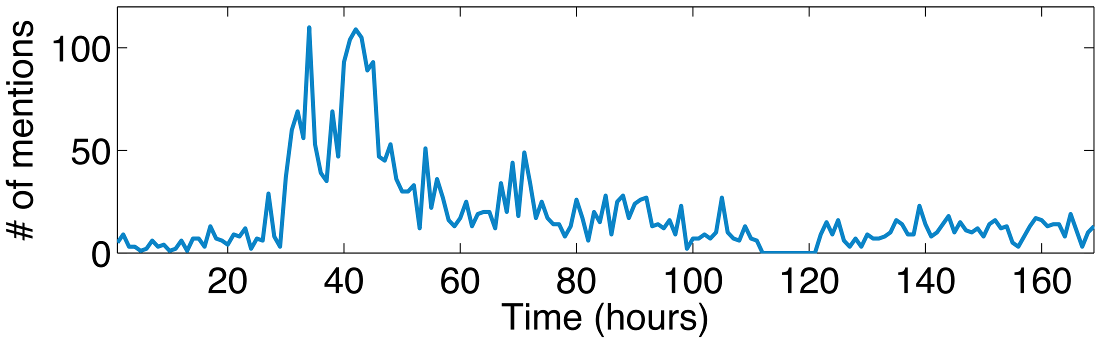
- Short phrases sourced from U.S. politics in 2008

"you can put lipstick on a pig" (# of mentions in blogs)



(per hour, 1 week)

"yes we can"



Kumamoto U

CMU CS

News spread in social media

MemeTracker [Leskovec+ KDD’09]

MemeTracker

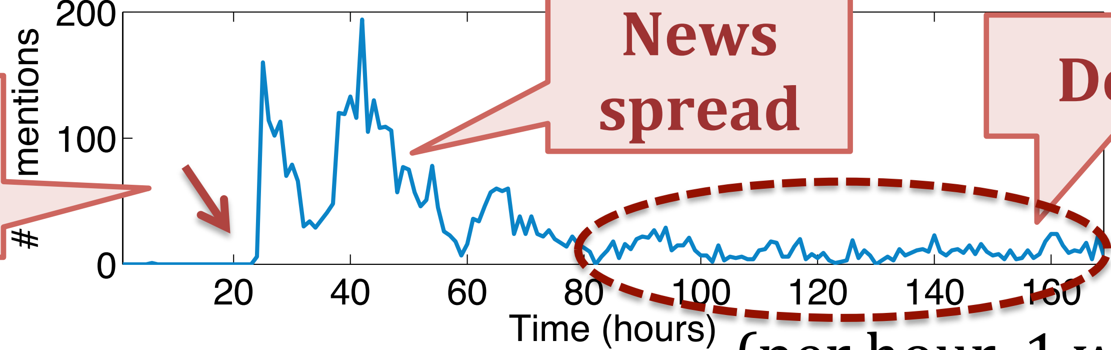
- Short phrases sourced from U.S. politics in 2008

“you can put lipstick on a pig” (# of mentions in blogs)

Breaking news

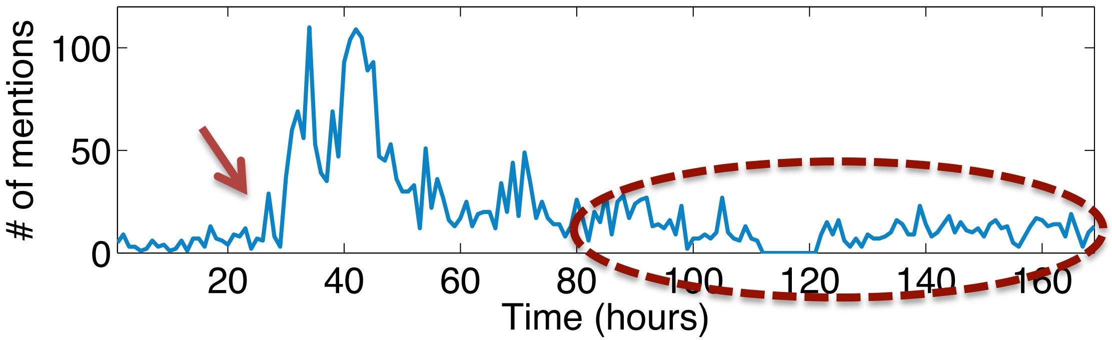
News spread

Decay



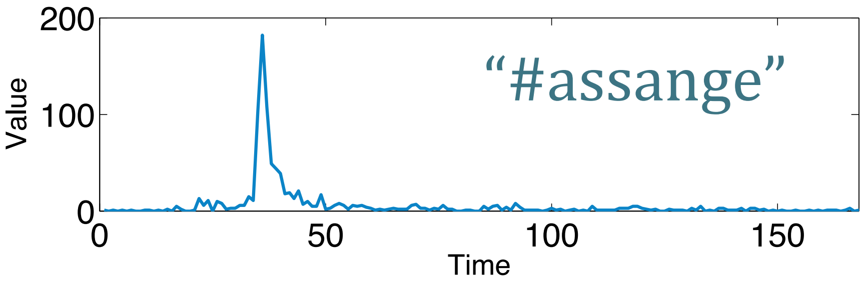
(per hour, 1 week)

“yes we can”

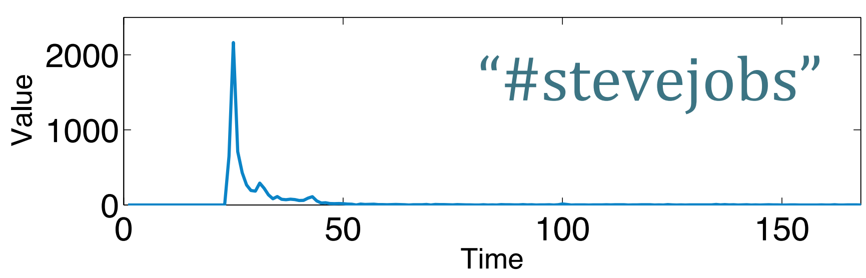


News spread in social media

- Twitter (# of hashtags per hour)



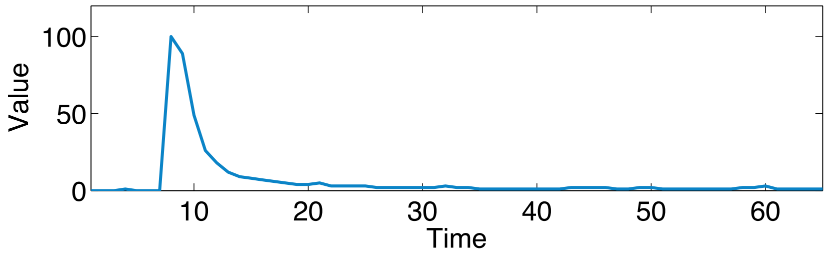
(per hour, 1week)



(per hour, 1 week)

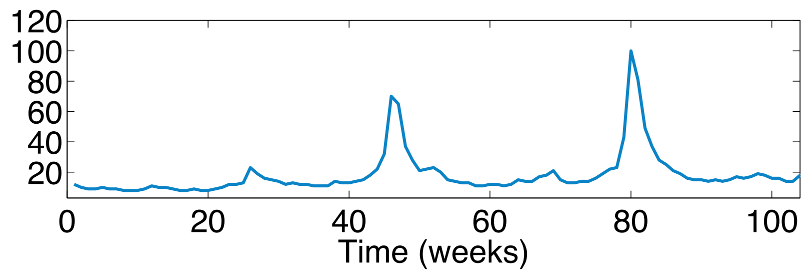
- Google trend (# of queries per week)

"tsunami" (in 2005)



(per week, 1 year)

"harry potter" (2010 - 2011)



(per week, 2 years)

Kumamoto U

News spread in social media

CMU CS

Q. How many patterns are there?

- Four classes on YouTube, etc.

  [Crane et al. PNAS'08]

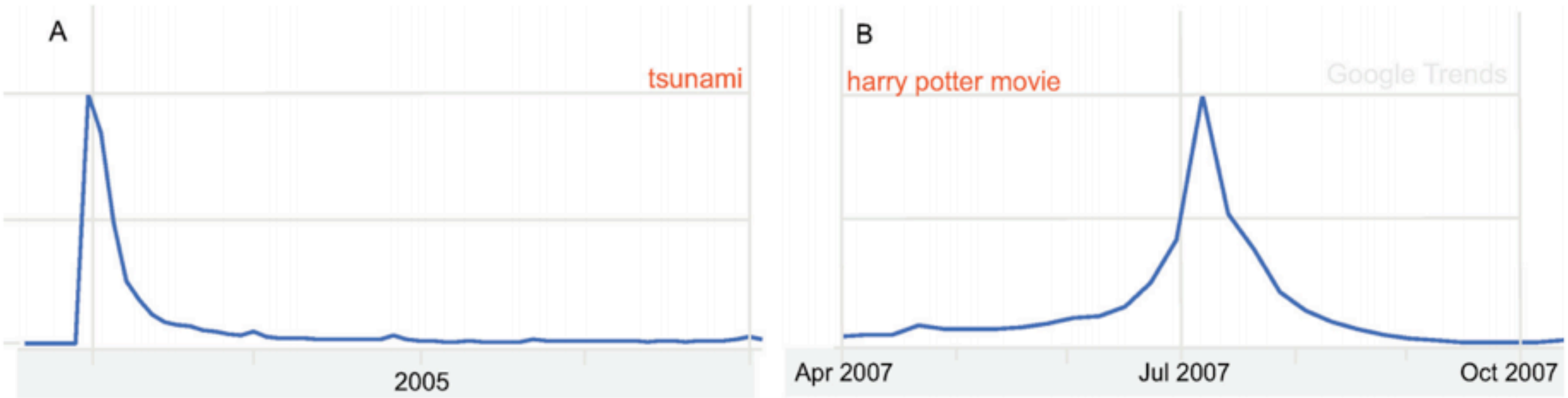
- Six classes on Social media

  [Yang et al. WSDM'11]

News spread in social media

[Crane et al. PNAS'08]

- The volume of Google searches

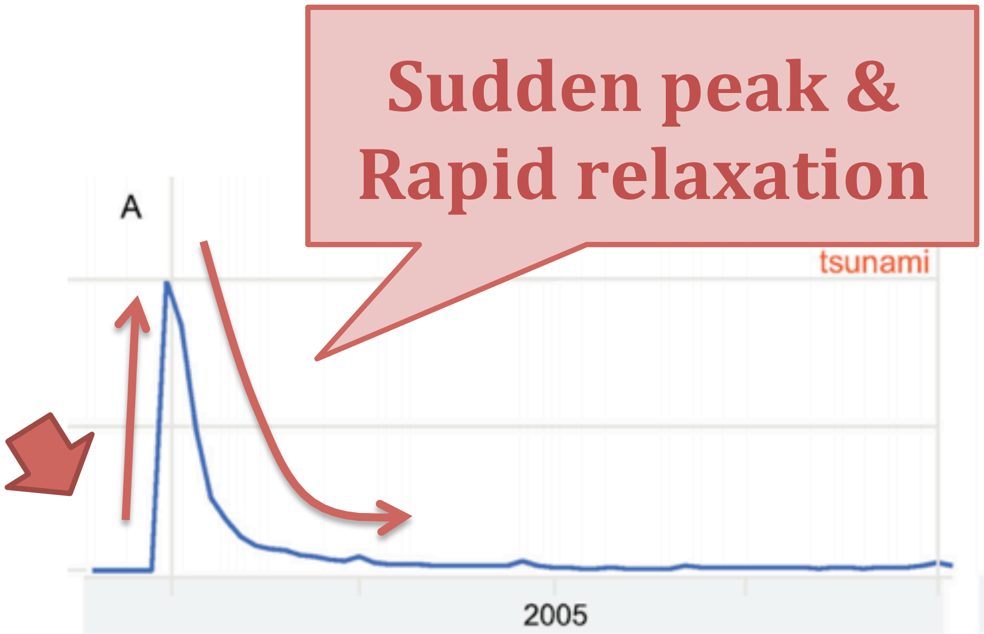


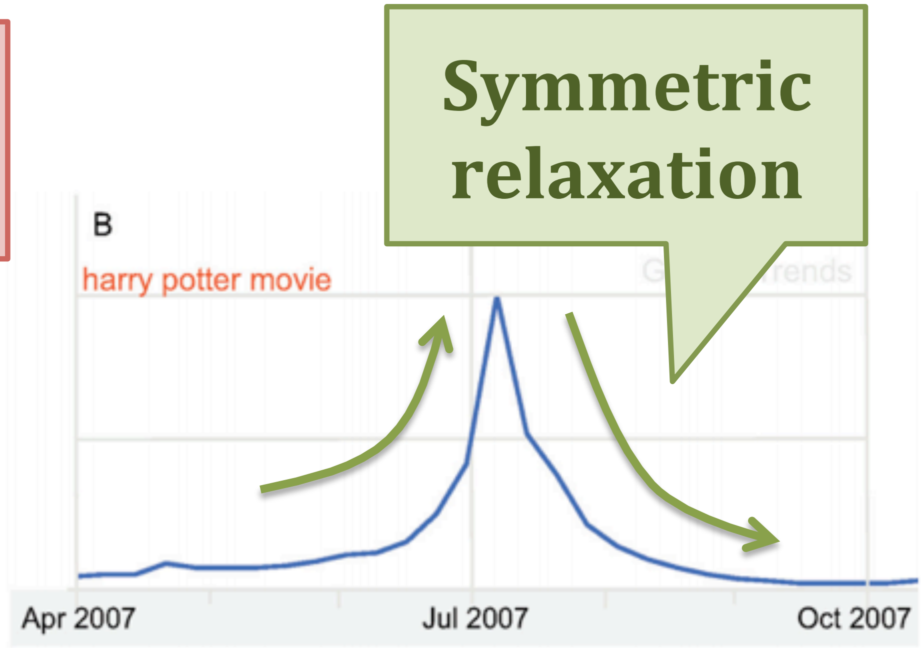
"Tsunami" "Harry potter movie"

News spread in social media

[Crane et al. PNAS'08]

- The volume of Google searches





"Tsunami"
(Exogenous)

"Harry potter movie"
(Endogenous)

News spread in social media

[Crane et al. PNAS'08]

- Based on self-excited Hawkes Poisson process*

$$\frac{dB(t)}{dt} = S(t) + \sum_{i, t_i \leq t} \mu_i \cdot \phi(t - t_i)$$

*[Hawkes+ 1974]

Kumamoto U

News spread in social media

CMU CS

[Crane et al. PNAS'08]

- Based on self-excited Hawkes Poisson process*

$$\frac{dB(t)}{dt} = S(t) + \sum_{i,t_i \leq t} \mu_i \cdot \phi(t - t_i)$$

Rate of spread of infection/ propagation	Exogenous /External source	# of Potential viewers	Decaying virus/news strength

*[Hawkes+ 1974]

News spread in social media

[Crane et al. PNAS'08]

- Based on self-excited Hawkes Poisson process*

$$\frac{dB(t)}{dt} = S(t) + \sum_{i, t_i \le t} \mu_i \cdot \phi(t - t_i)$$

of ... tial ... vers

Decaying virus/news strength (Power law)

$$\phi(t) \sim \frac{1}{t^{1+\theta}} \quad (0 < \theta < 1)$$

propagation

*[Hawkes+ 1974]

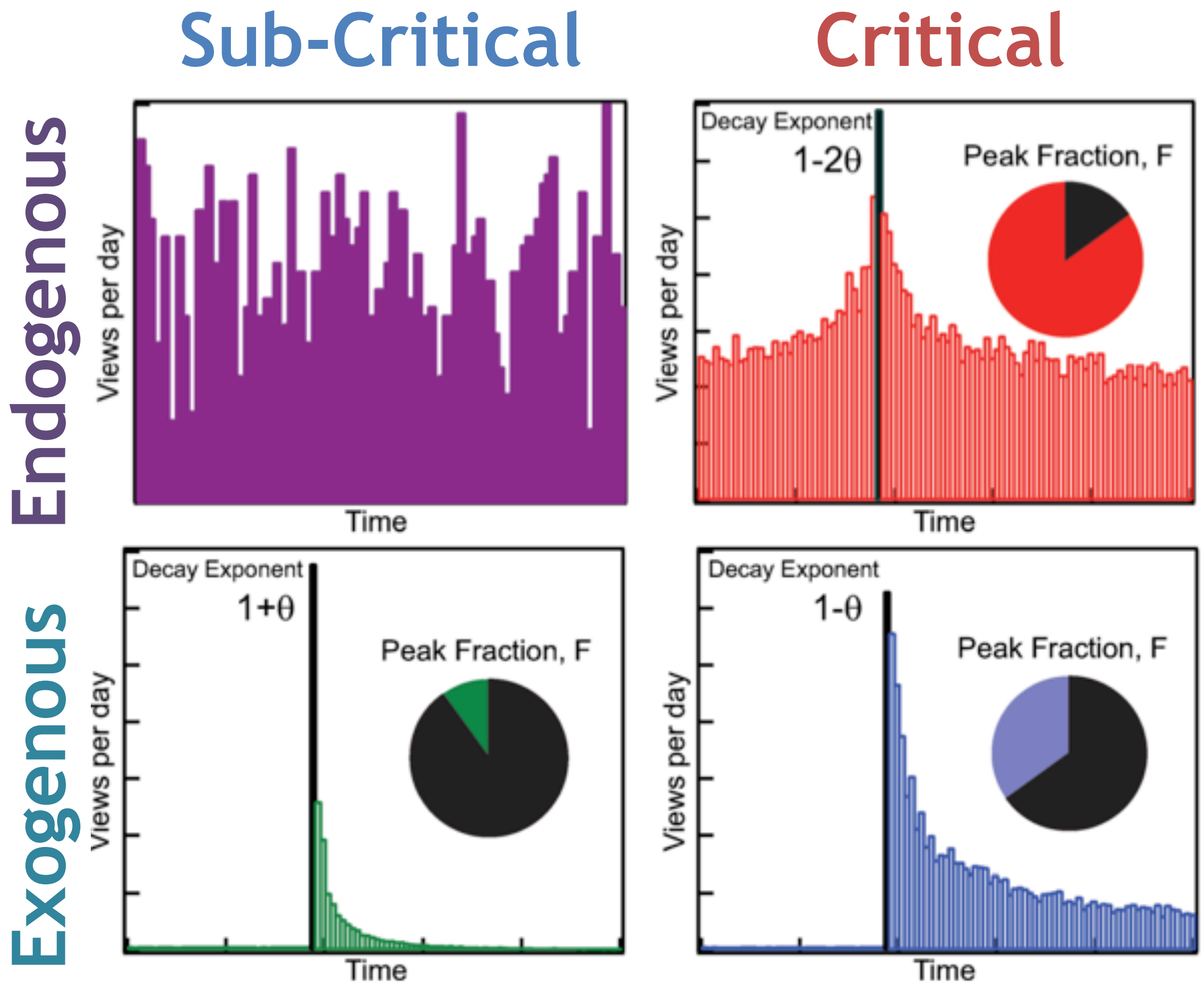
Kumamoto U

News spread in social media

CMU CS

[Crane et al. PNAS'08]

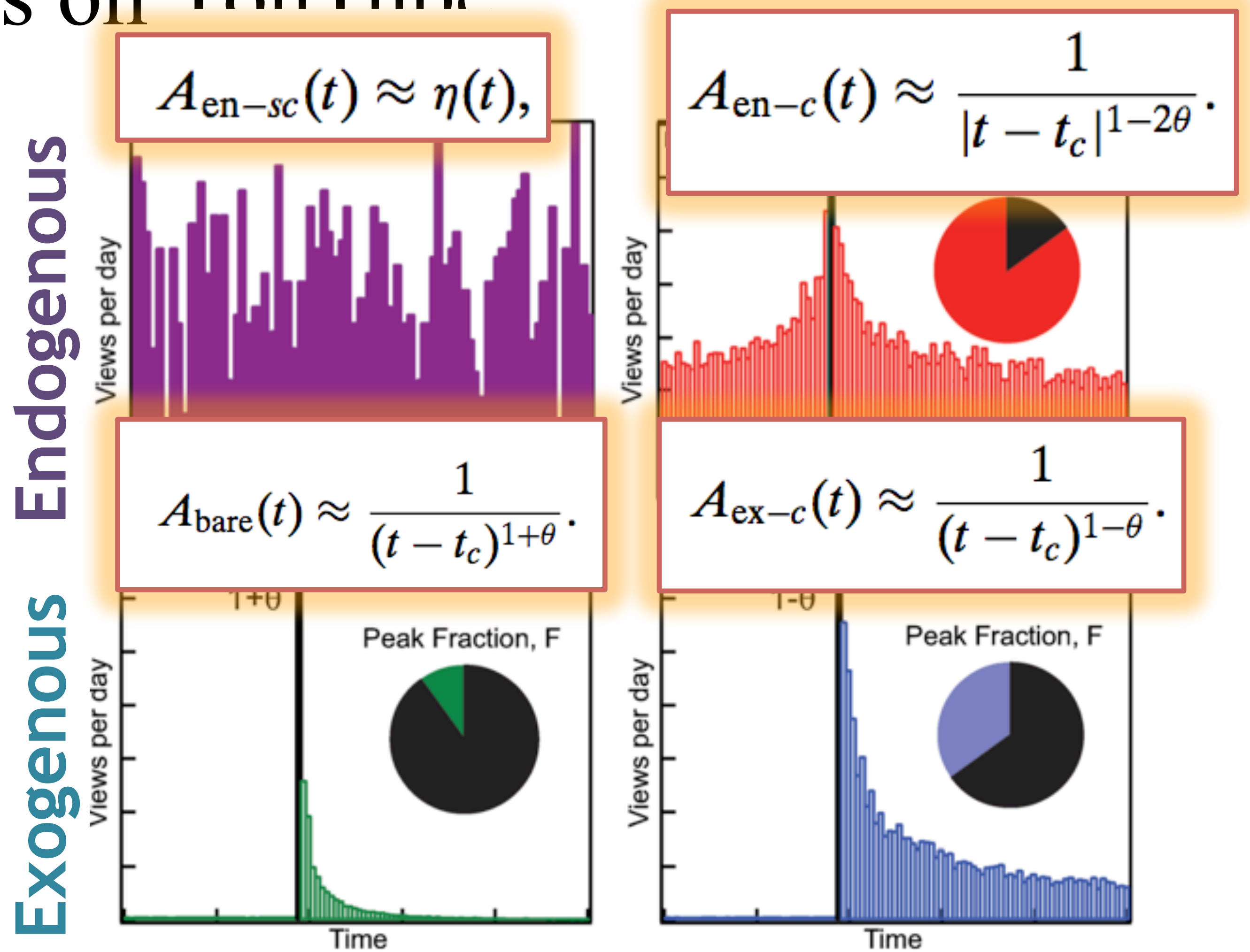
- Four classes on YouTube



News spread in social media

[Crane et al. PNAS'08]

- Four classes on YouTube



$$A_{en-sc}(t) \approx \eta(t),$$

$$A_{en-c}(t) \approx \frac{1}{|t - t_c|^{1-2\theta}}.$$

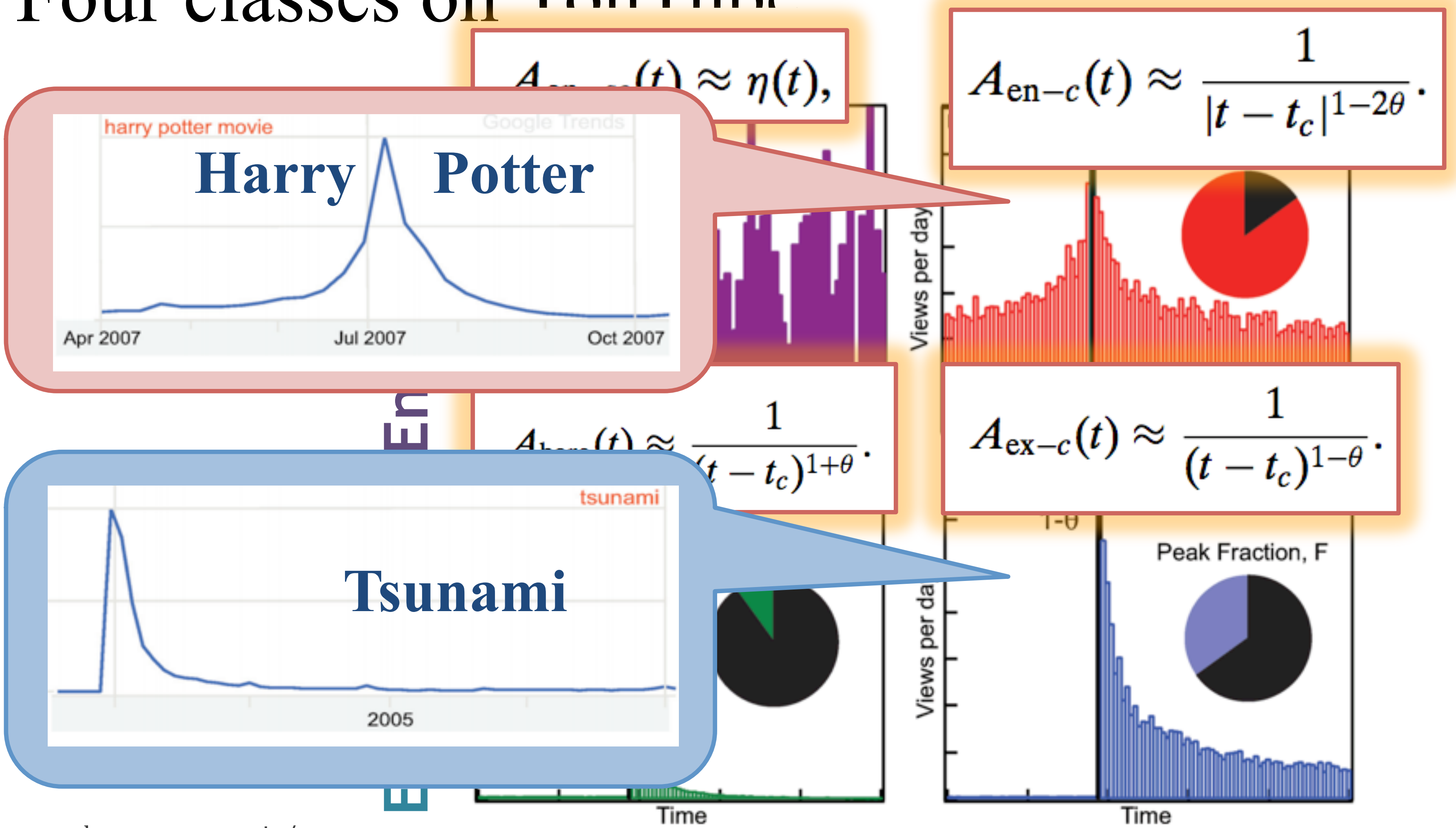
$$A_{bare}(t) \approx \frac{1}{(t - t_c)^{1+\theta}}.$$

$$A_{ex-c}(t) \approx \frac{1}{(t - t_c)^{1-\theta}}.$$

News spread in social media

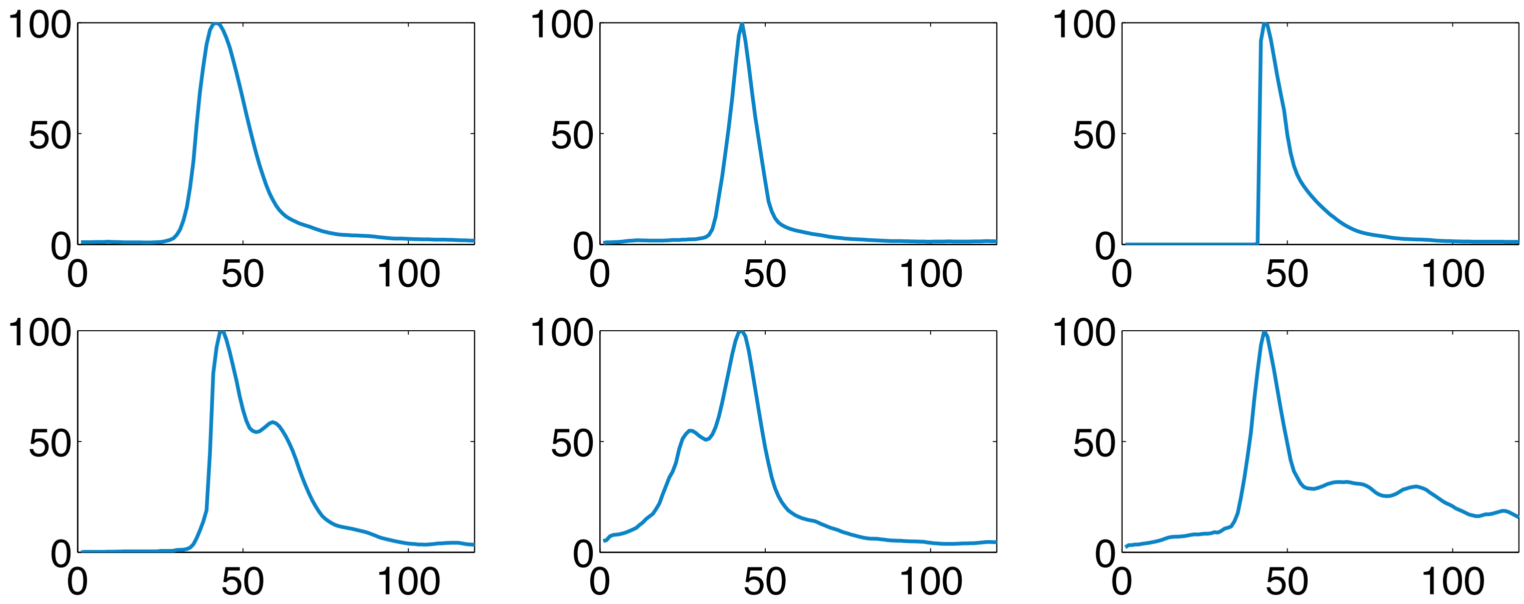
[Crane et al. PNAS'08]

- Four classes on YouTube



News spread in social media

- Six classes of information diffusion patterns on social media [Yang et al. WSDM'11]

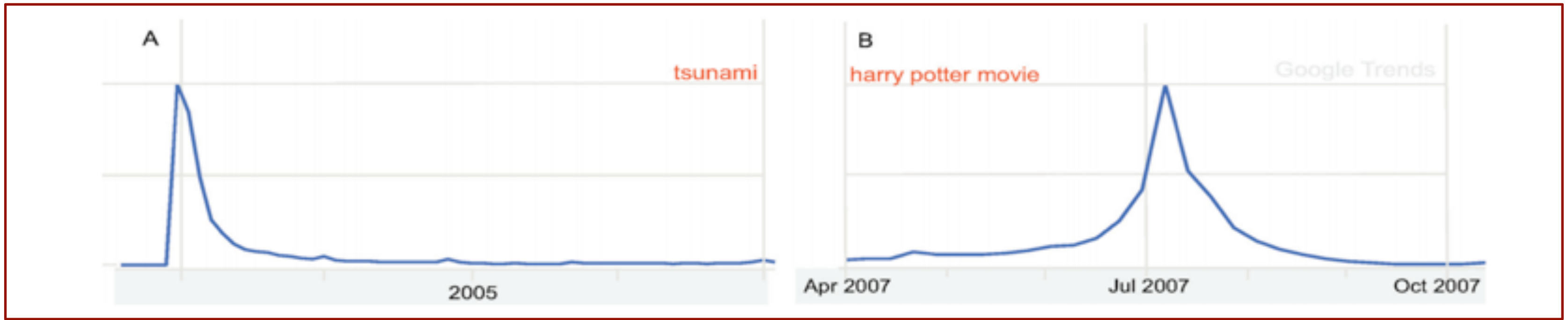


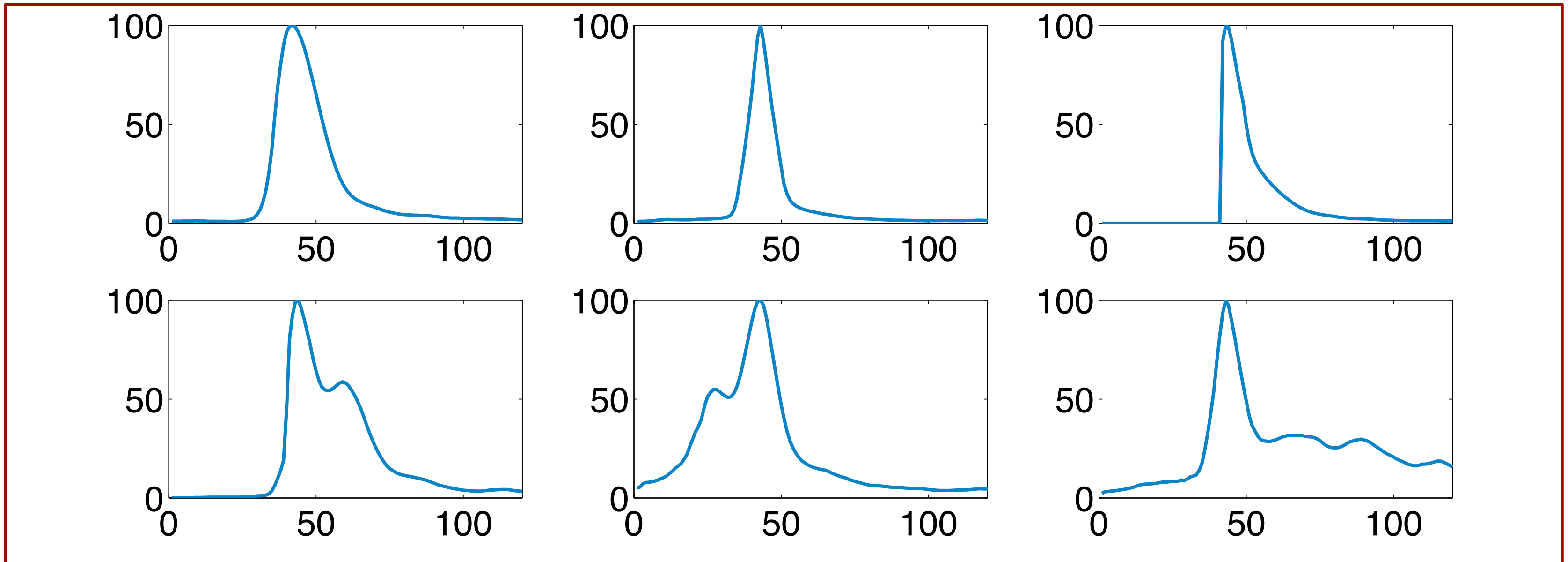
Kumamoto U

News spread in social media

CMU CS

Q. How many patterns are there, after all?





Kumamoto U

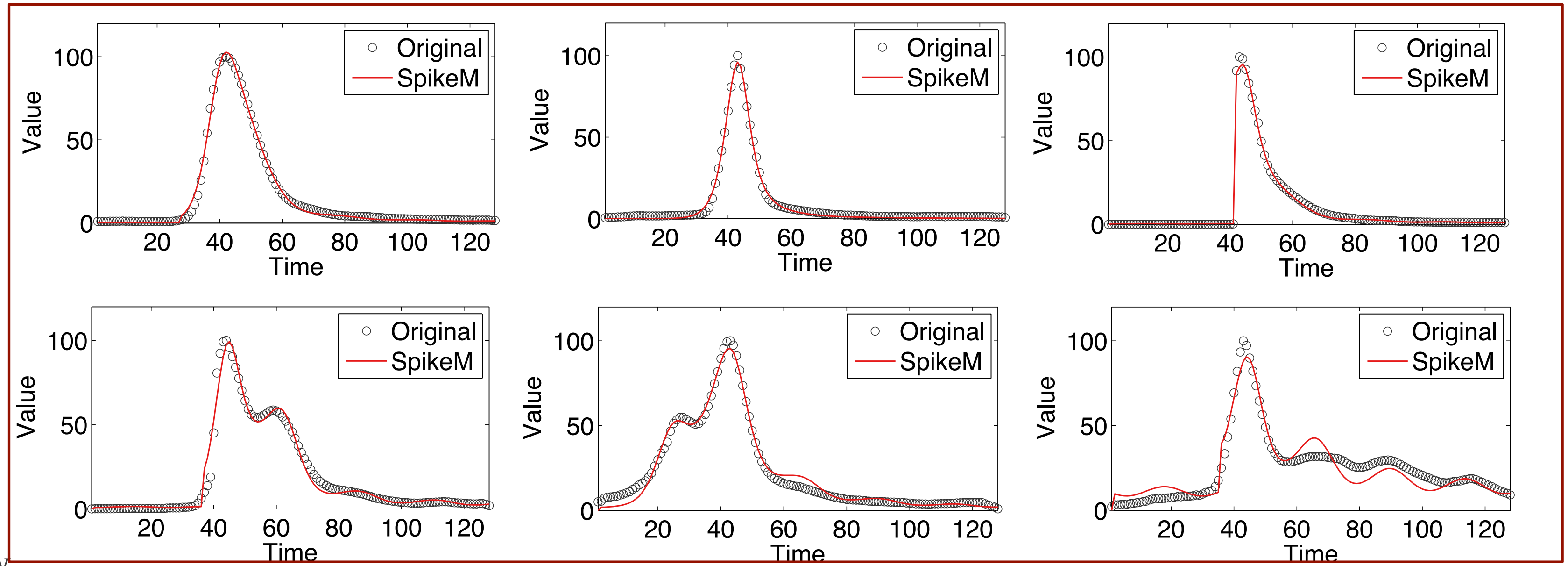
News spread in social media

CMU CS

A. Our answer is "ONE"!

A single non-linear model !

"SpikeM"



[Matsubara+ KDD'12]

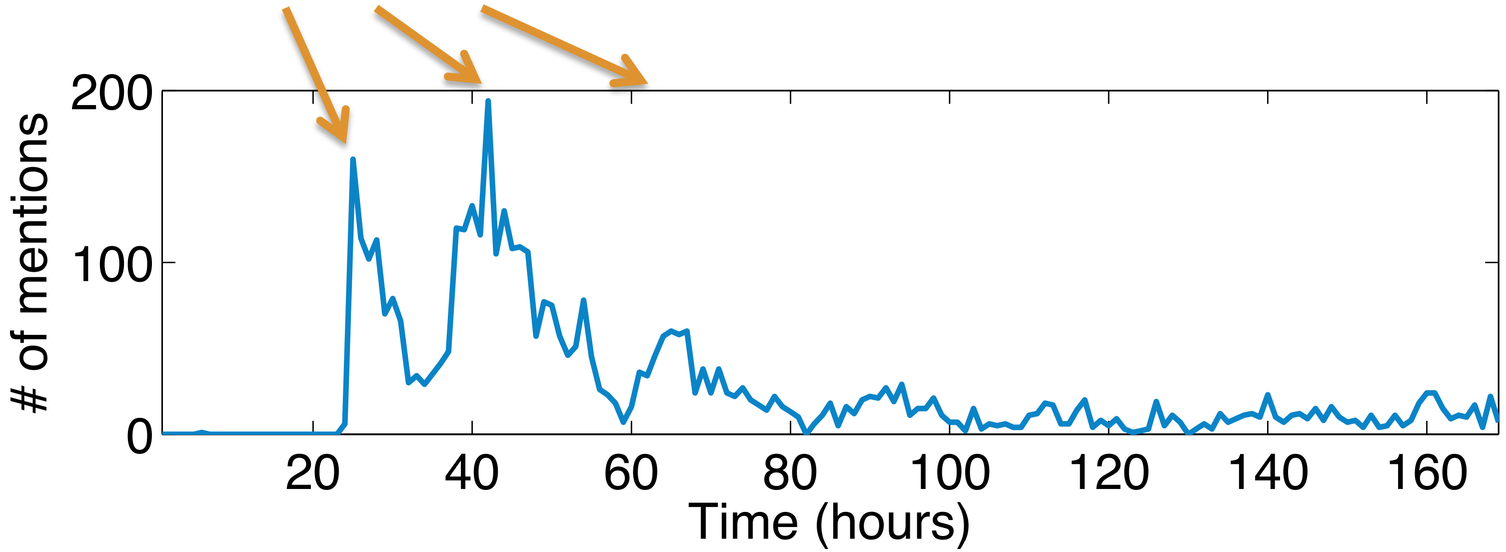
Rise and Fall Patterns of Information Diffusion: Model and Implications

Yasuko Matsubara (Kyoto University),
Yasushi Sakurai (NTT),
B. Aditya Prakash (CMU),
Lei Li (UCB), Christos Faloutsos (CMU)

Rise and fall patterns in social media

SpikeM captures 3 properties of real spike

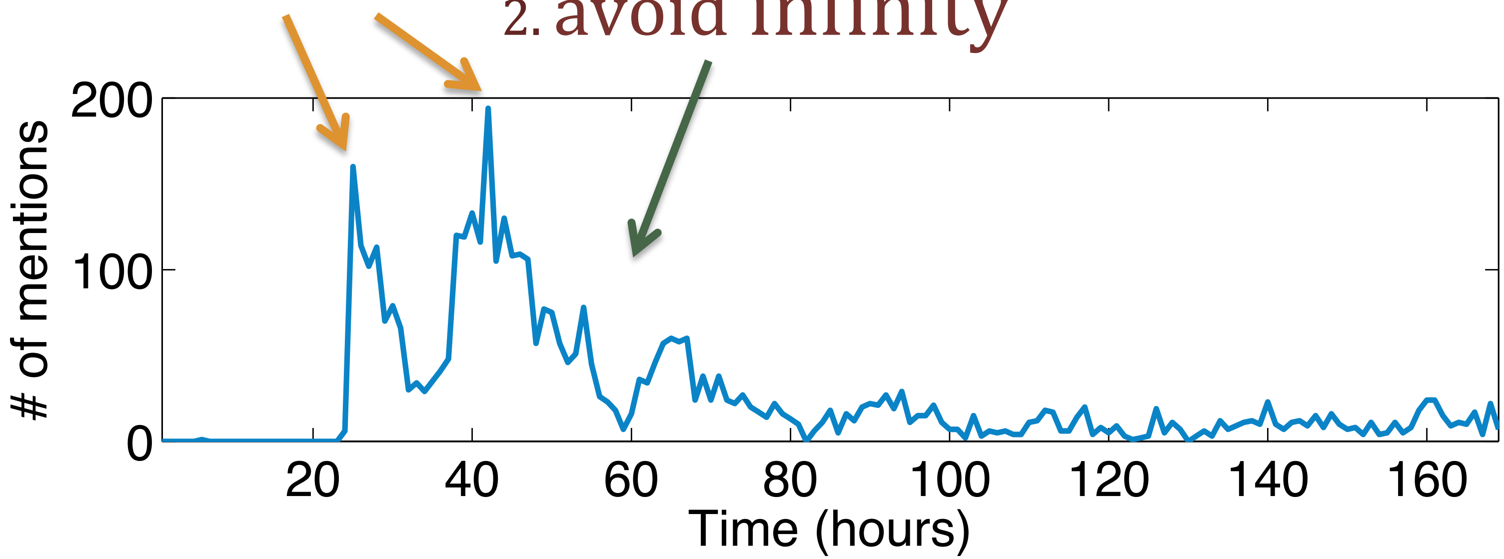
1. periodicities



Rise and fall patterns in social media

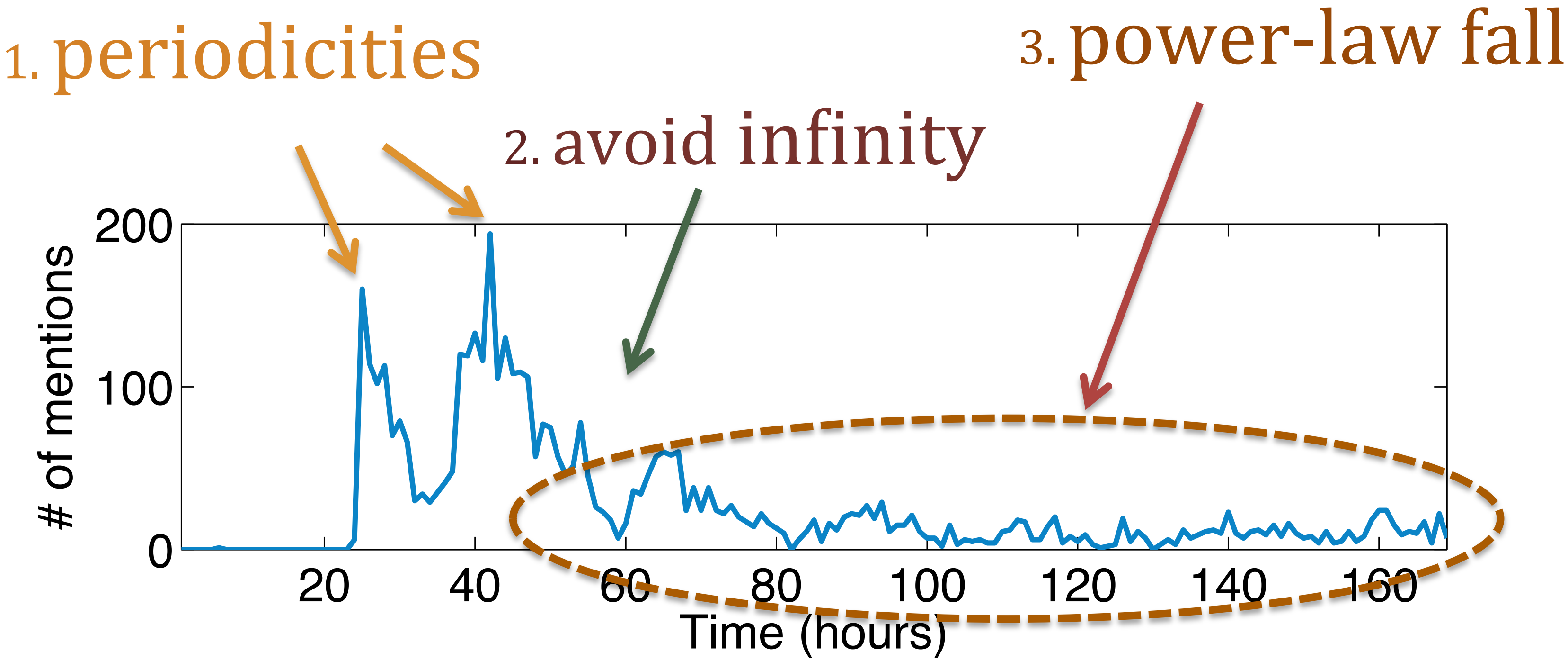
SpikeM captures 3 properties of real spike

1. periodicities
2. avoid infinity



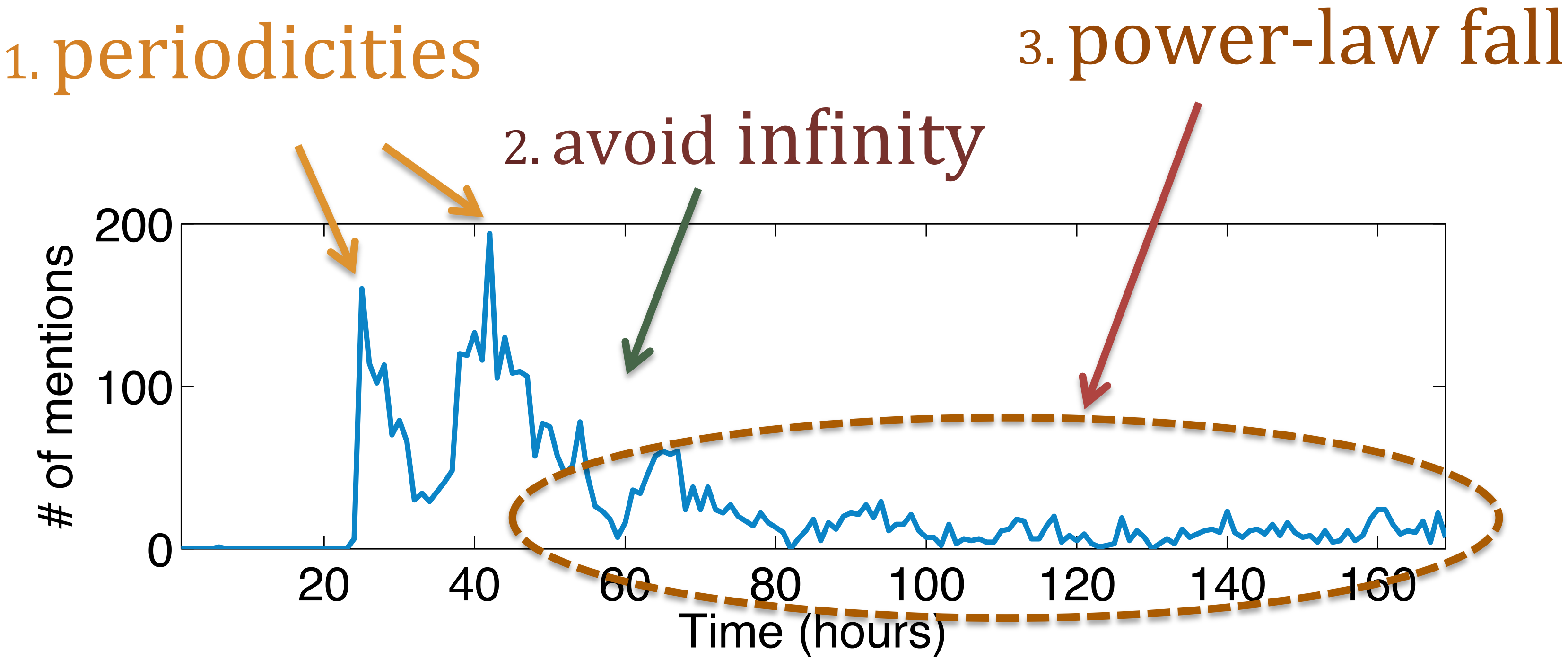
Rise and fall patterns in social media

SpikeM captures 3 properties of real spike



Rise and fall patterns in social media

SpikeM captures 3 properties of real spike

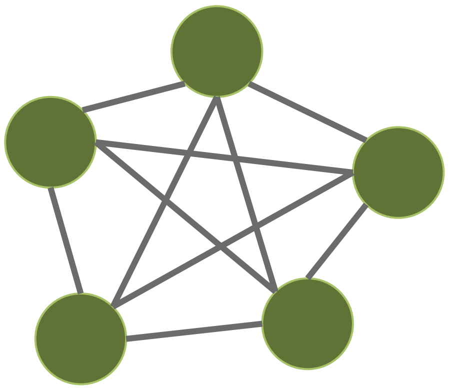


SpikeM can capture behavior of real spikes using few parameters

Details

Main idea (details)

- 1. **Un**-informed bloggers (clique of N bloggers/nodes)



Time n=0

Nodes (bloggers) consist of two states

U – **U**n-informed of rumor

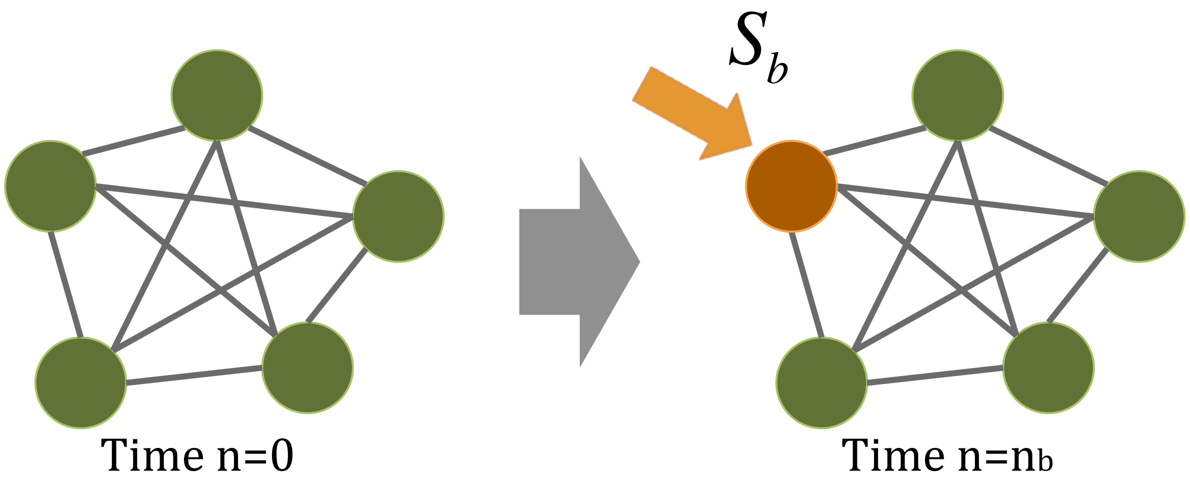
B – informed, and **B**logged about rumor

Details

CMU CS

Main idea (details)

- 1. **Un**-informed bloggers (clique of N bloggers/nodes)
- 2. External shock at time n_b (e.g, breaking news)



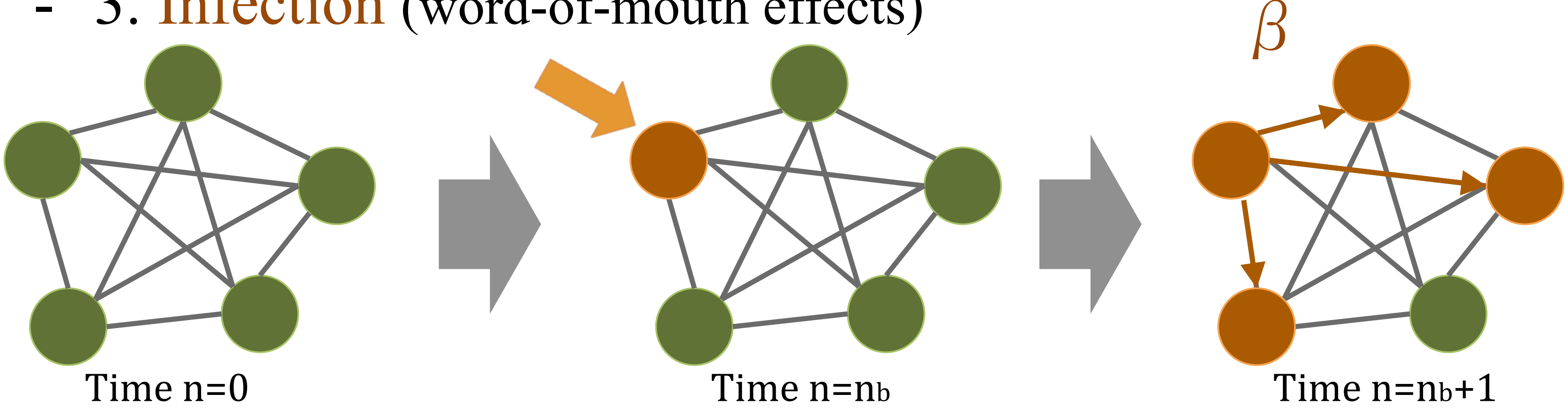
External shock

- Event happened at time n_b
- S_b bloggers are informed, blog about news

Details

Main idea (details)

- 1. **Un**-informed bloggers (clique of N bloggers/nodes)
- 2. External shock at time n_b (e.g, breaking news)
- 3. Infection (word-of-mouth effects)



Infectiveness of a blog-post

β – Strength of infection (quality of news)

$f(n)$ – Decay function (how infective a blog posting is)

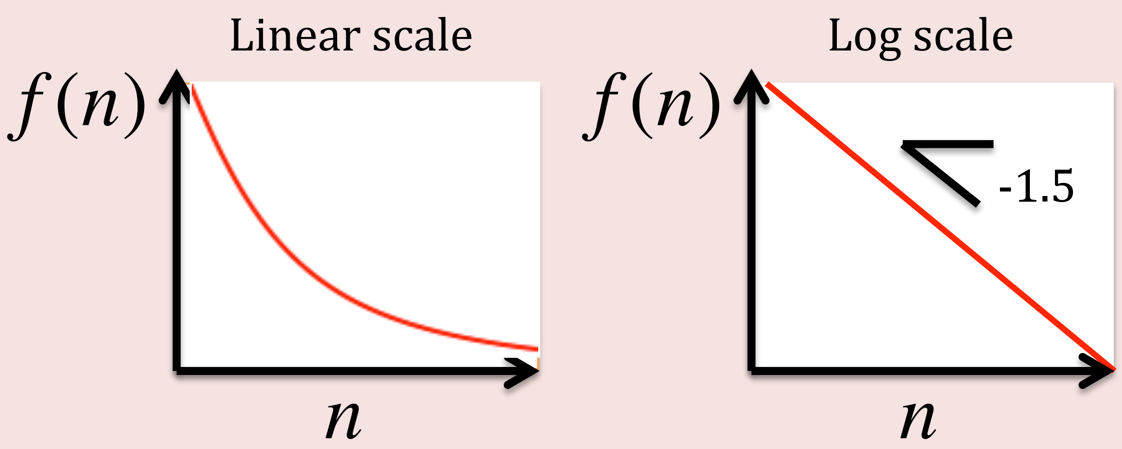
Main idea (details)

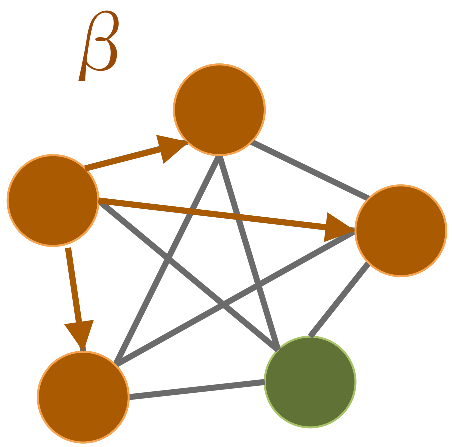
CMU CS

- 1. **Un**-informed bloggers (clique of N bloggers/nodes)

news)

Decay function: $f(n) = \beta * n^{-1.5}$





Time n=nb+1

Infectiveness of a blog-post

β – Strength of infection (quality of news)

$f(n)$ – Decay function (how infective a blog posting is)

Details

SpikeM-base (details)

CMU CS

Equations of SpikeM (base)

$$\Delta B(n+1) = U(n) \cdot \sum_{t=n_b}^{n} (\Delta B(t) + S(t)) \cdot f(n+1-t) + \varepsilon$$

Blogged

$$U(n+1) = U(n) - \Delta B(n+1)$$

Un-informed

N	– Total population of available bloggers
β	– Strength of infection/news
n_b, S_b	– External shock S_b at birth (time n_b)
ε	– Background noise

Details

SpikeM - periodicity

Full equation of SpikeM

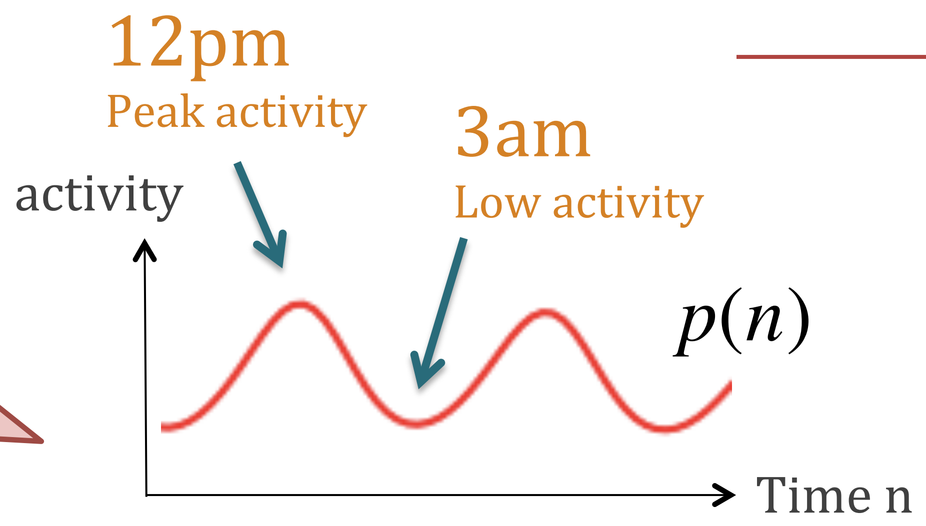
$$\Delta B(n+1) = p(n+1) \cdot \left[U(n) \cdot \sum_{t=n_b}^{n} (\Delta B(t) + S(t)) \cdot f(n+1-t) + \varepsilon \right]$$

Blogged — Periodicity

$$U(n+1) = U(n) - \Delta B(n+1)$$

Un-informed

Bloggers change their activity over time
(e.g., daily, weekly, yearly)



Details

CMU CS

Model fitting (Details)

- SpikeM consists of 7 parameters

$$\theta = \{N, \beta, n_b, S_b, \varepsilon, P_a, P_s\}$$

Learning parameters

- Given a real time sequence

$$X = \{X(1), \ldots X(n), \ldots, X(n_d)\}$$

- Minimize the error

(Levenberg-Marquardt (LM) fitting)

$$D(X, \theta) = \sum_{n=1}^{n_d} (X(n) - \Delta B(n))^2$$

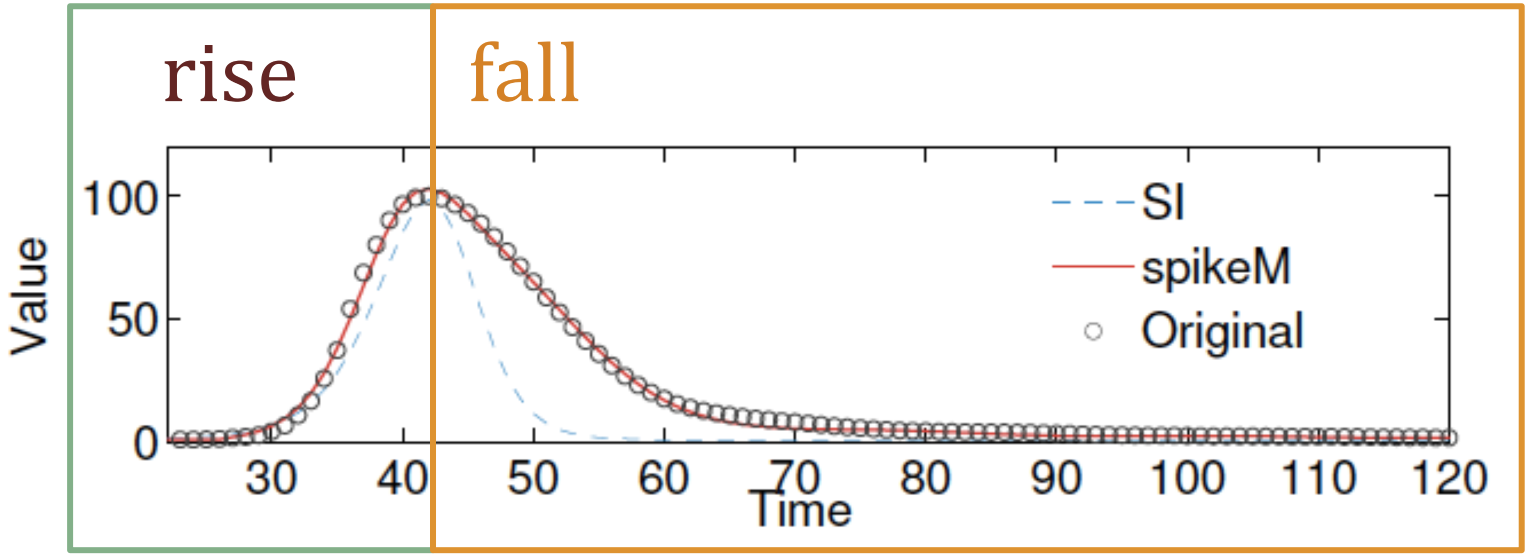
Details

Analysis

CMU CS

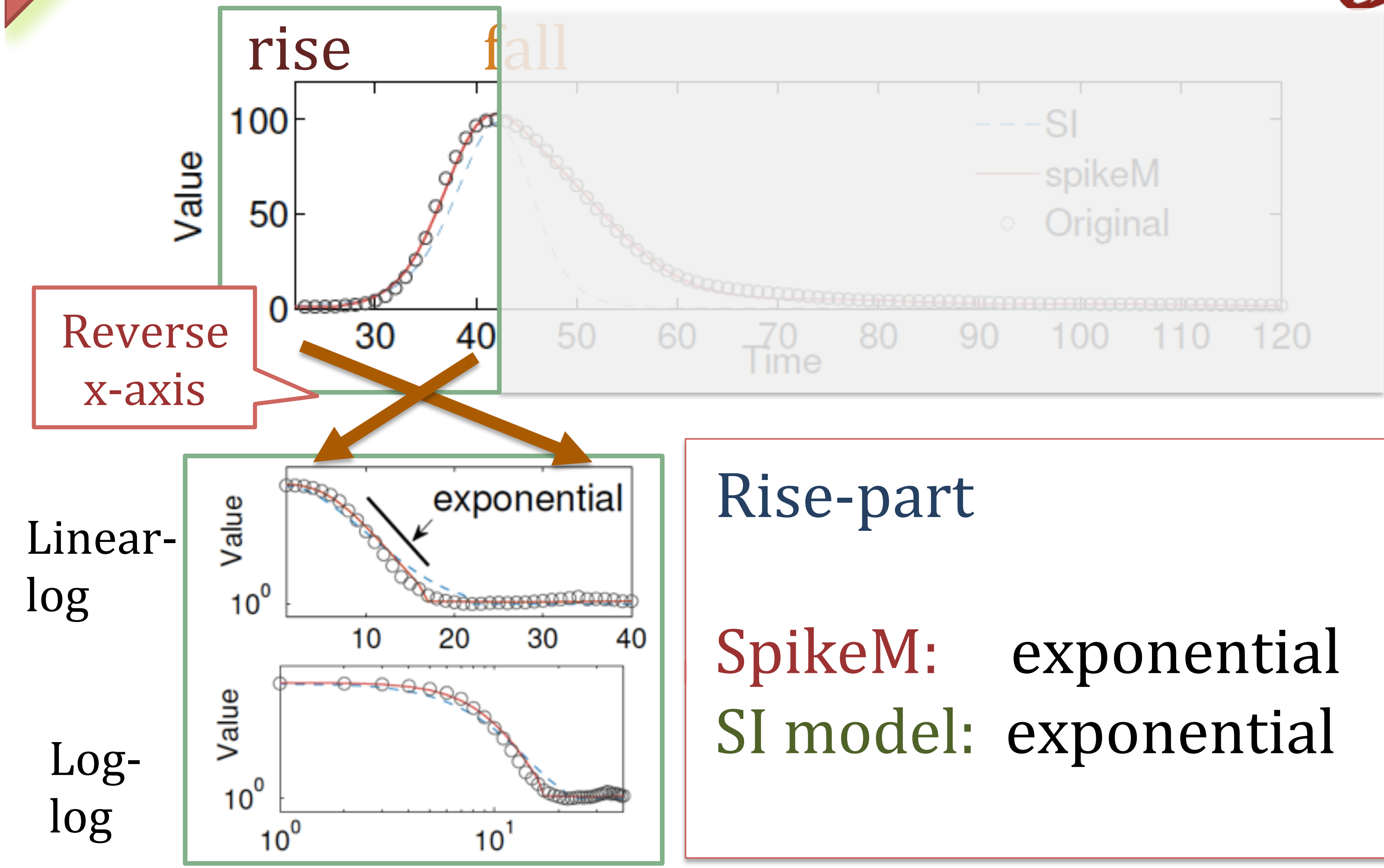
SpikeM matches reality

exponential rise and power-raw fall



SpikeM vs. **SI model** (susceptible infected model)

Analysis

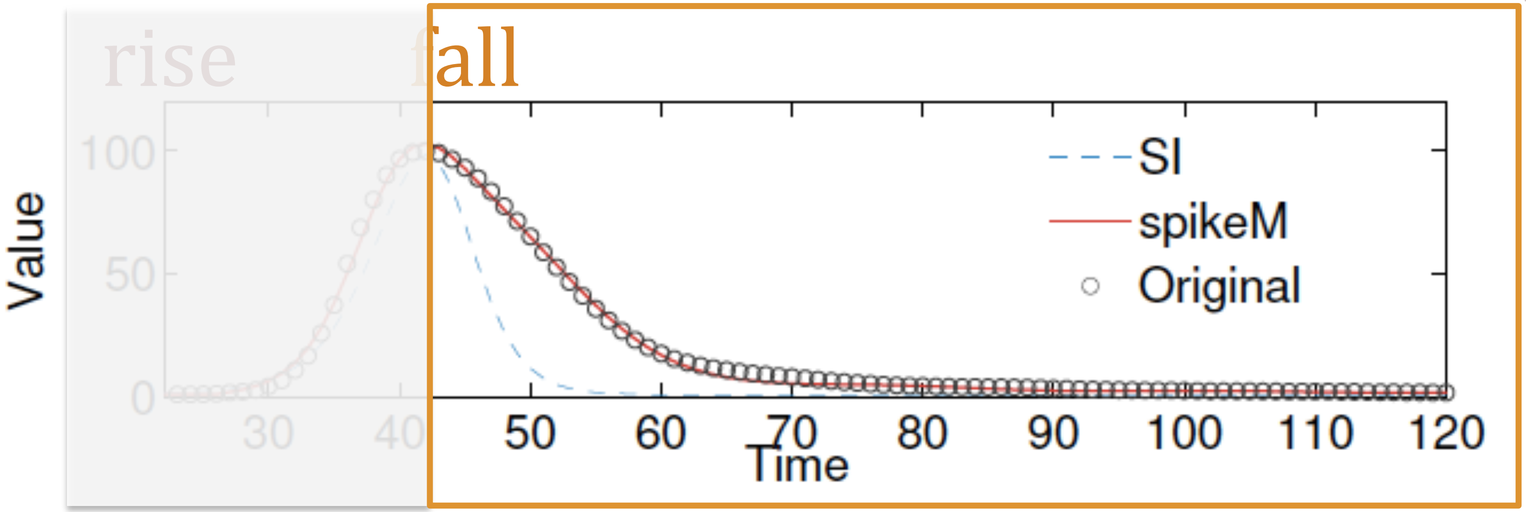


Rise-part

SpikeM: exponential

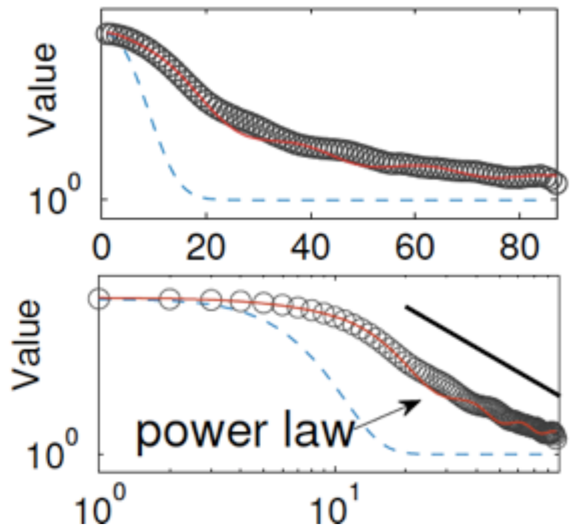
SI model: exponential

Analysis



Fall-part
SpikeM: power law
SI model: exponential

SpikeM matches reality

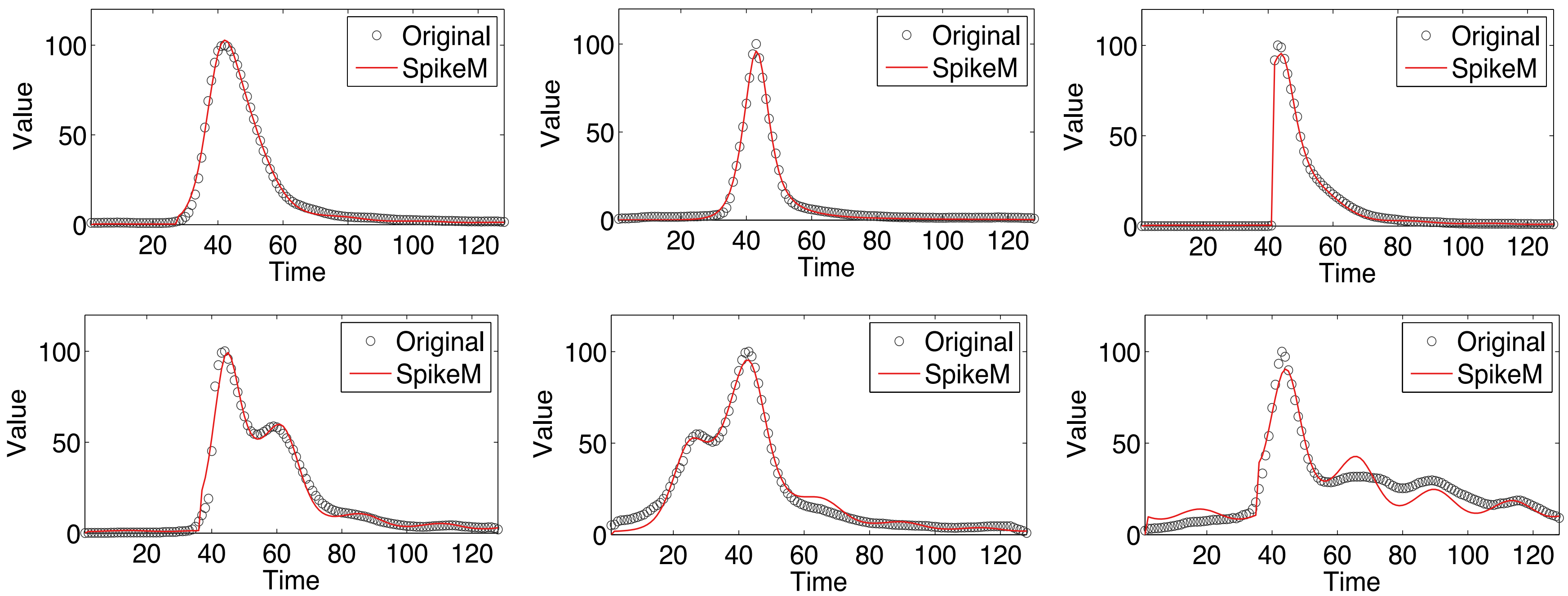


Linear-log

Log-log

Q1-1 Explaining K-SC clusters

–Six patterns of K-SC [Yang et al. WSDM’11]



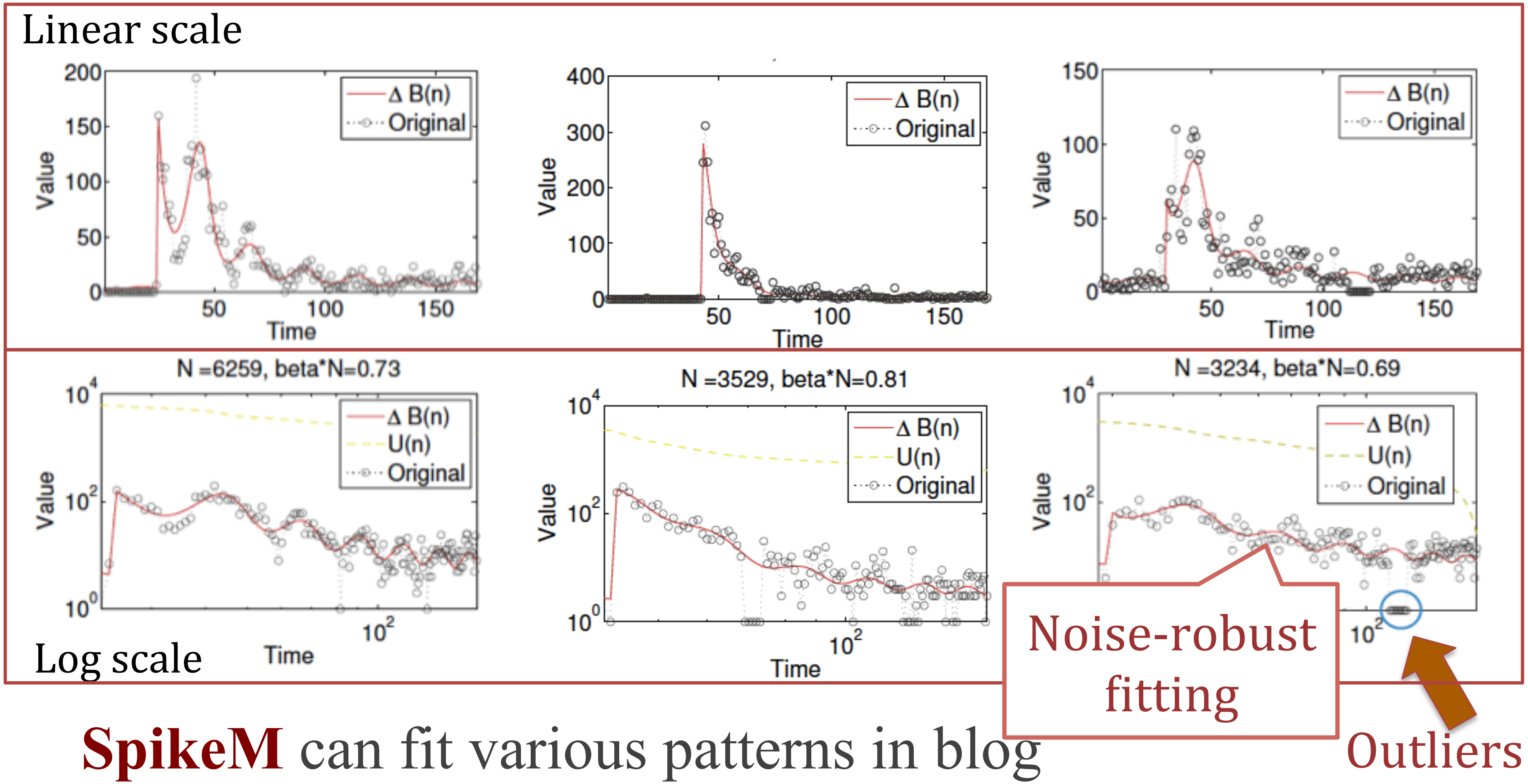
- **SpikeM** can generate all patterns in K-SC

Kumamoto U

Q1-2 Matching MemeTracker patterns

CMU CS

MemeTracker (memes in blogs) [Leskovec et al. KDD'09]



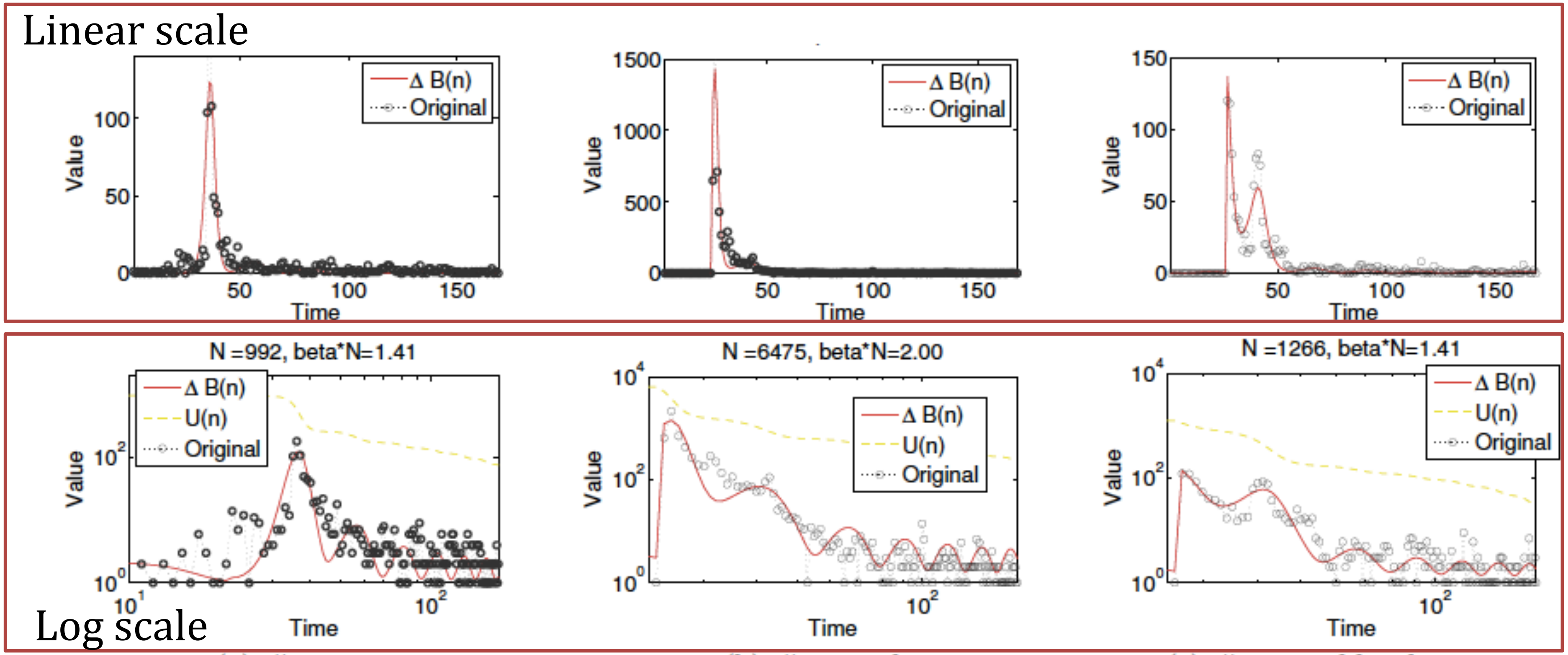
SpikeM can fit various patterns in blog

Q1-3 Matching Twitter data

Kumamoto U

CMU CS

Twitter data (hashtags)

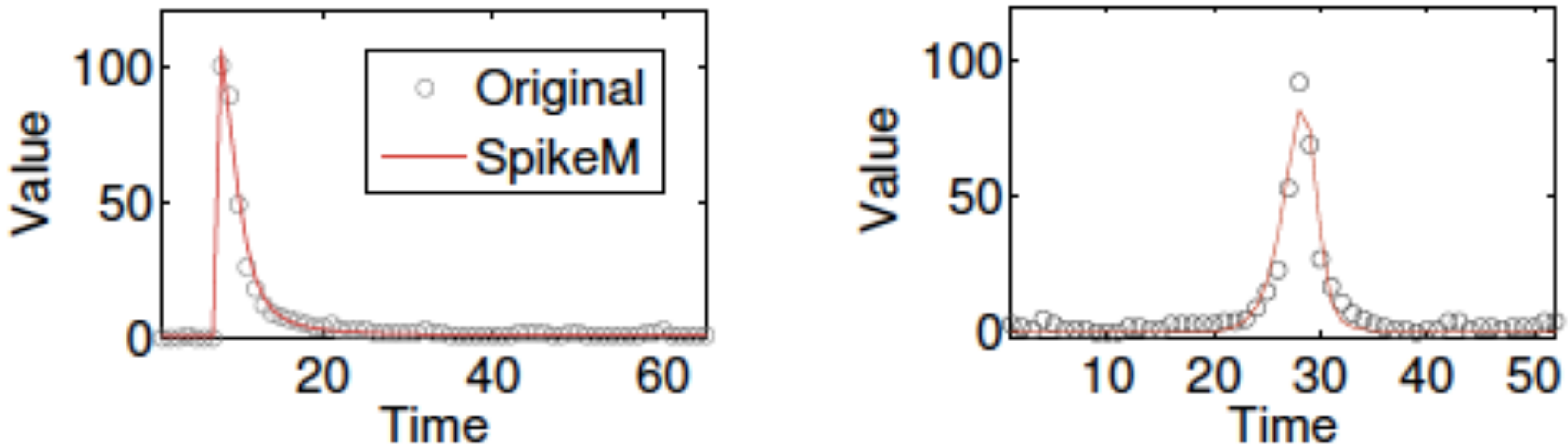


(a) #assange (b) #stevejobs (c) #arresteddevelopment

It can generate various patterns in social media

Q1-4 Matching Google trend data

Volume of searches for queries on Google

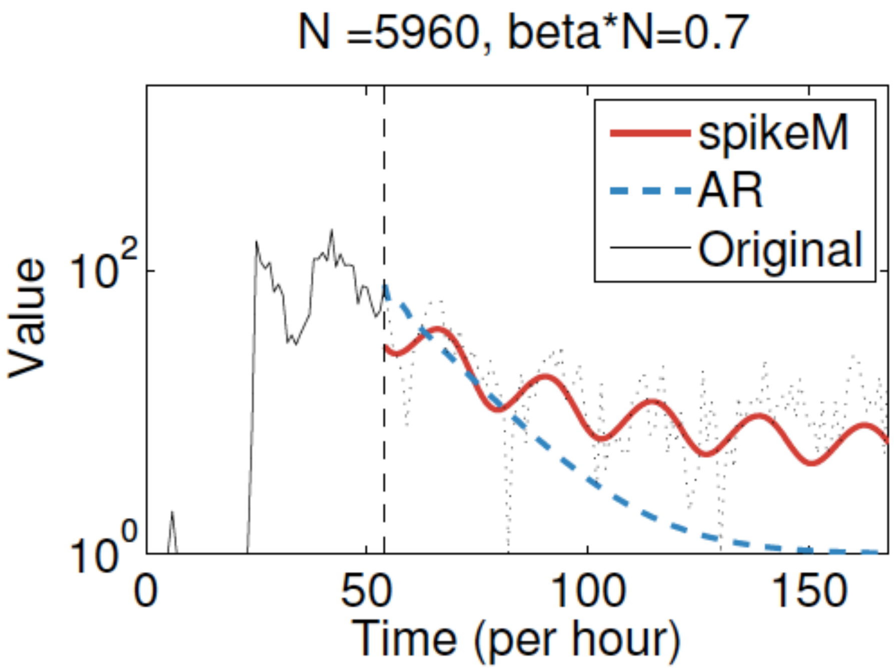


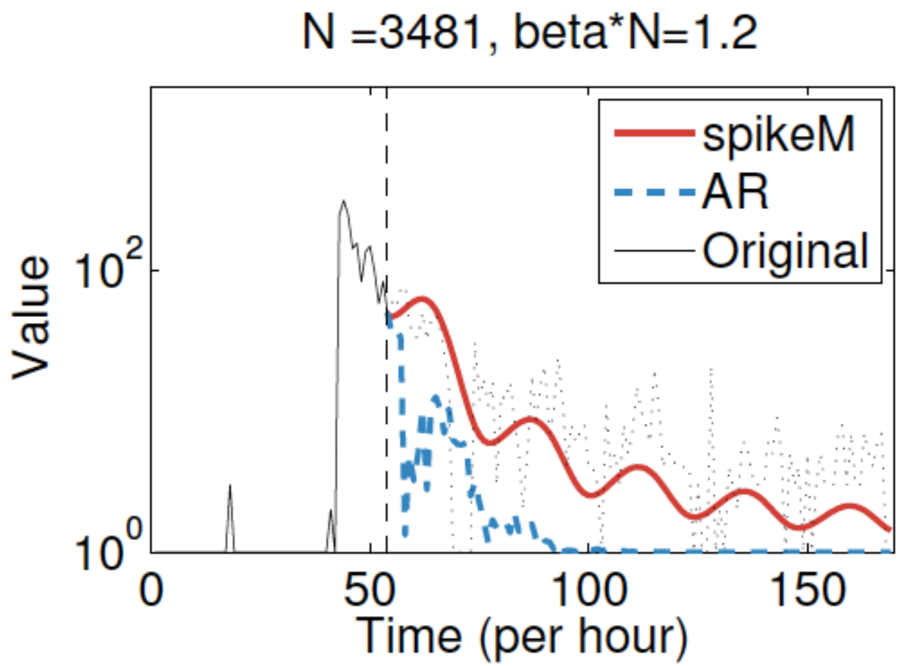
(a) "tsunami" (2005) (b) "Harry Potter" (2007)

SpikeM can capture various patterns

Q2 Tail-part forecasts

- Given a first part of the spike
  - forecast the tail part





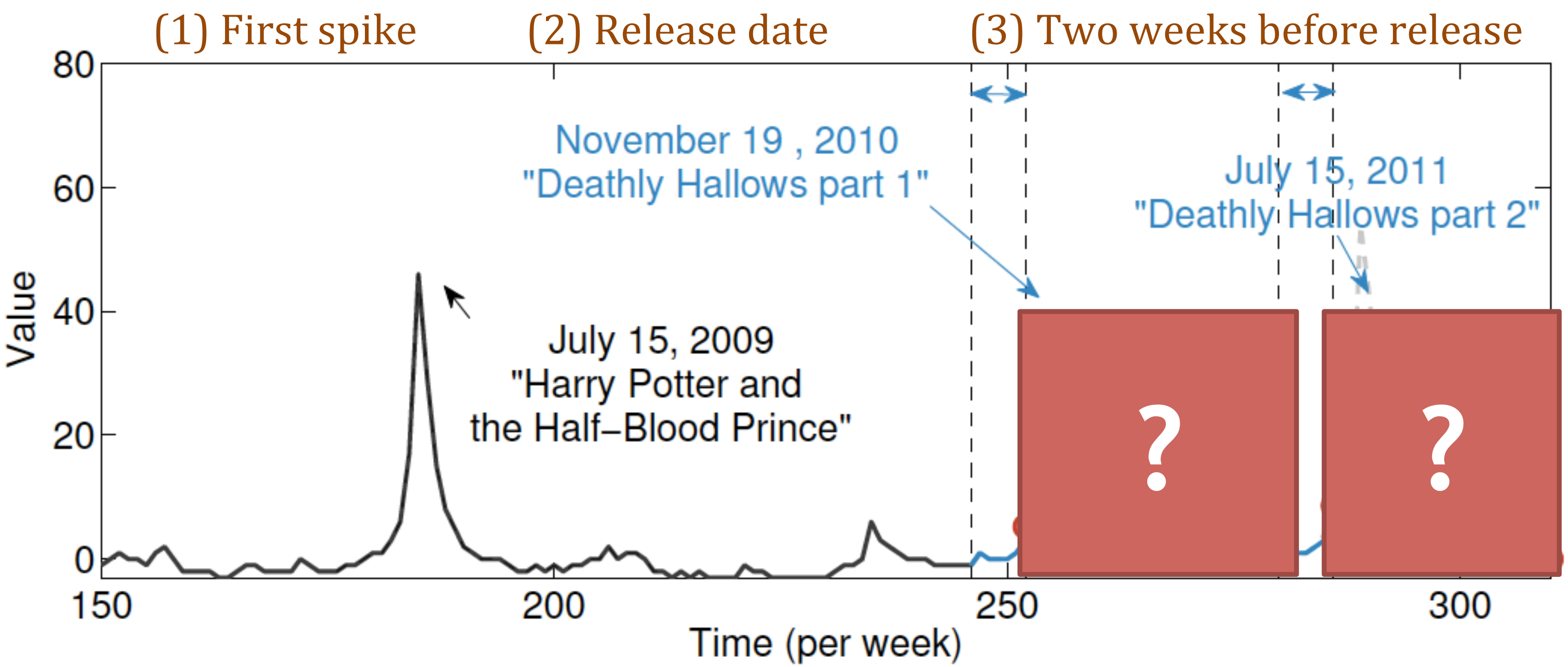
SpikeM can capture tail part (AR: fail)

Application

A1. "What-if" forecasting

CMU CS

Forecast not only tail-part, but also **rise-part**!



e.g., given (1) first spike,

(2) release date of two sequel movies

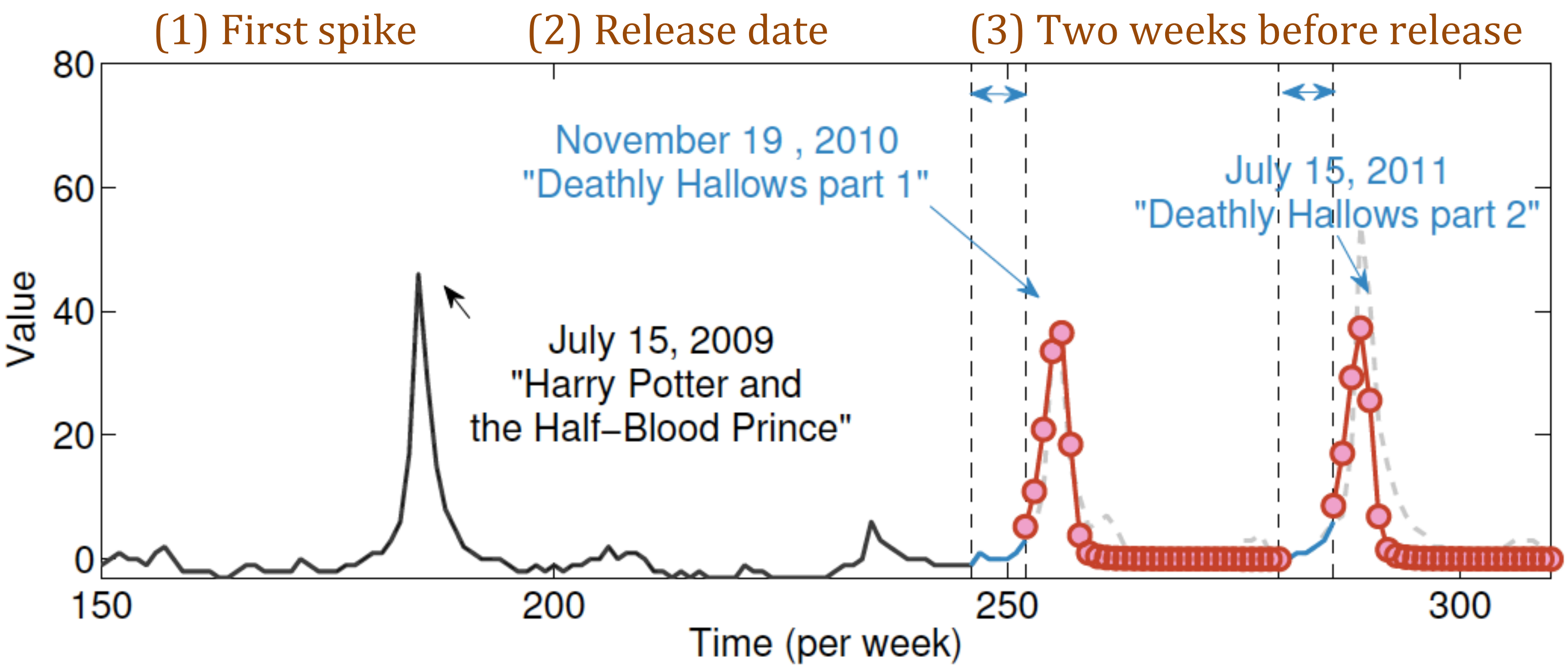
(3) access volume before the release date

A1. "What-if" forecasting

Application

CMU CS

Forecast not only tail-part, but also **rise-part**!



SpikeM can forecast upcoming spikes!

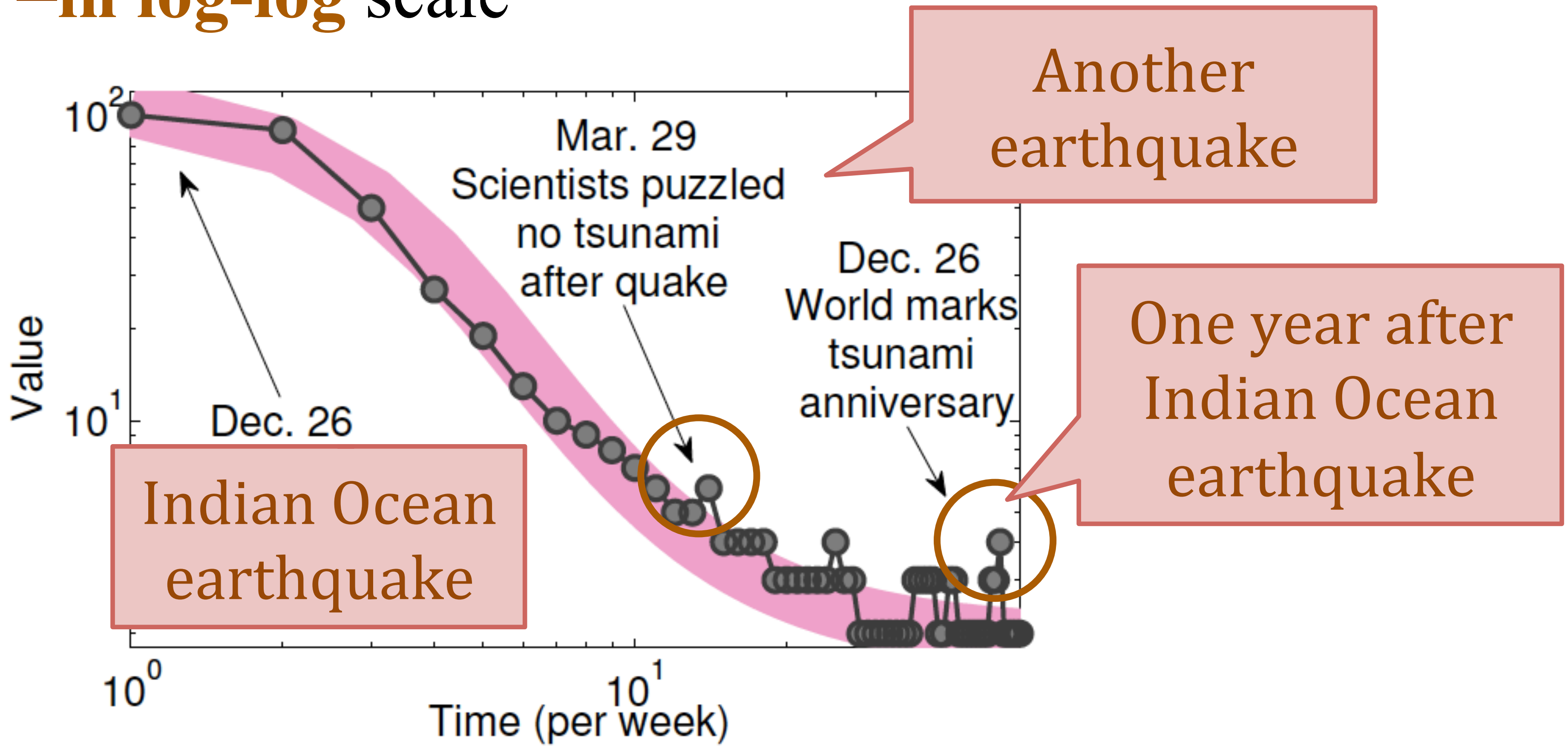
A2. Outlier detection

Application

CMU CS

–Fitting result of "tsunami (Google trend)"

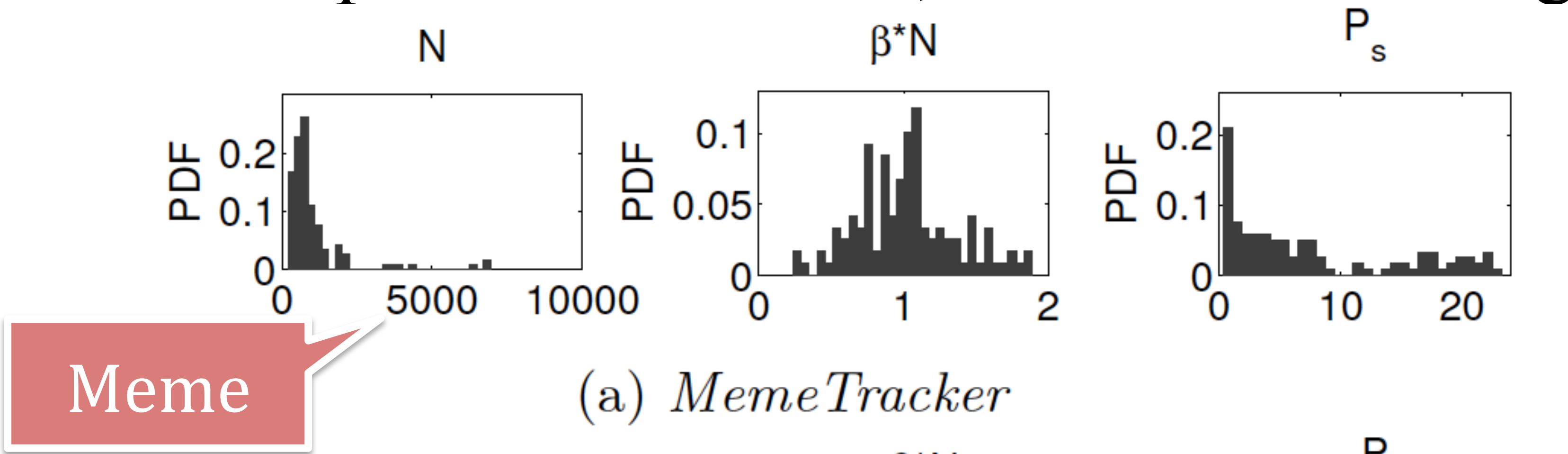
–**in log-log scale**



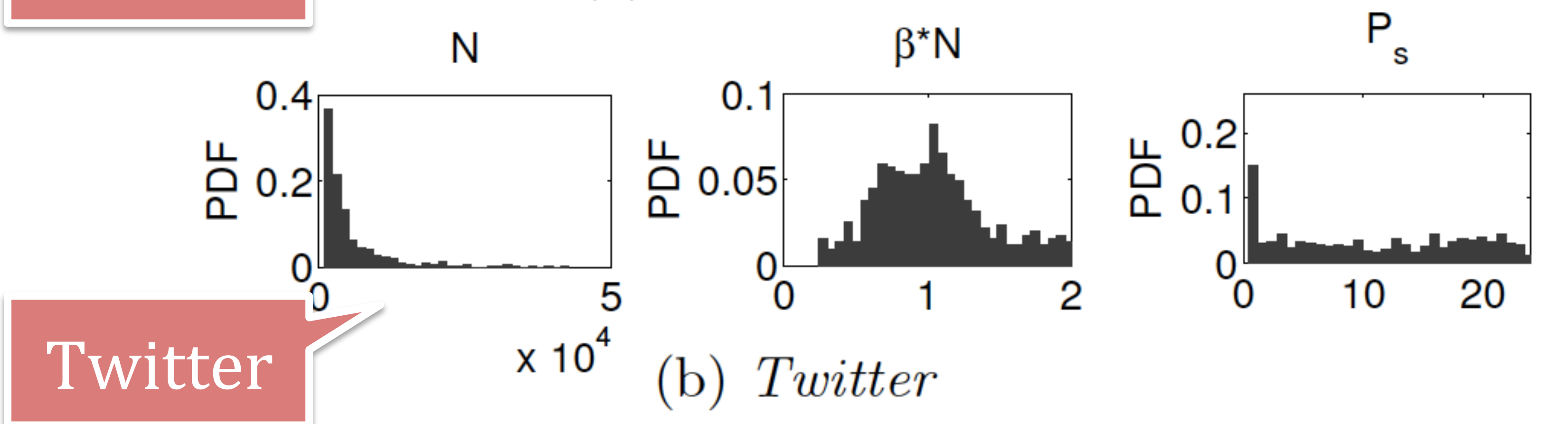
A3. Reverse engineering

SpikeM provide an intuitive explanation

PDF of parameters over 1,000 memes/hashtags



(a) *MemeTracker*



(b) *Twitter*

http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/

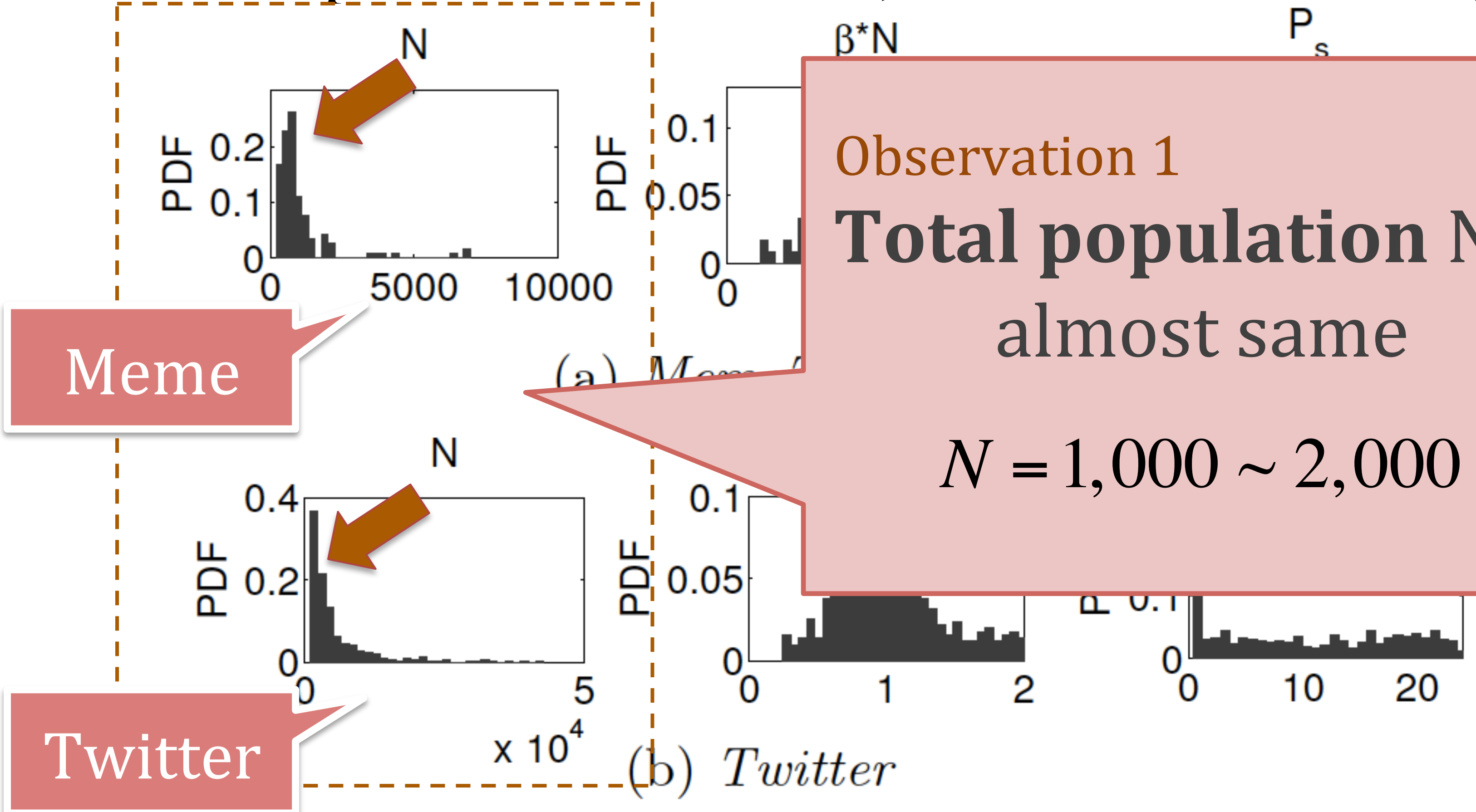
A3. Reverse engineering

Application

CMU CS

SpikeM provide an intuitive explanation

PDF of parameters over 1,000 memes/hashtags



Observation 1

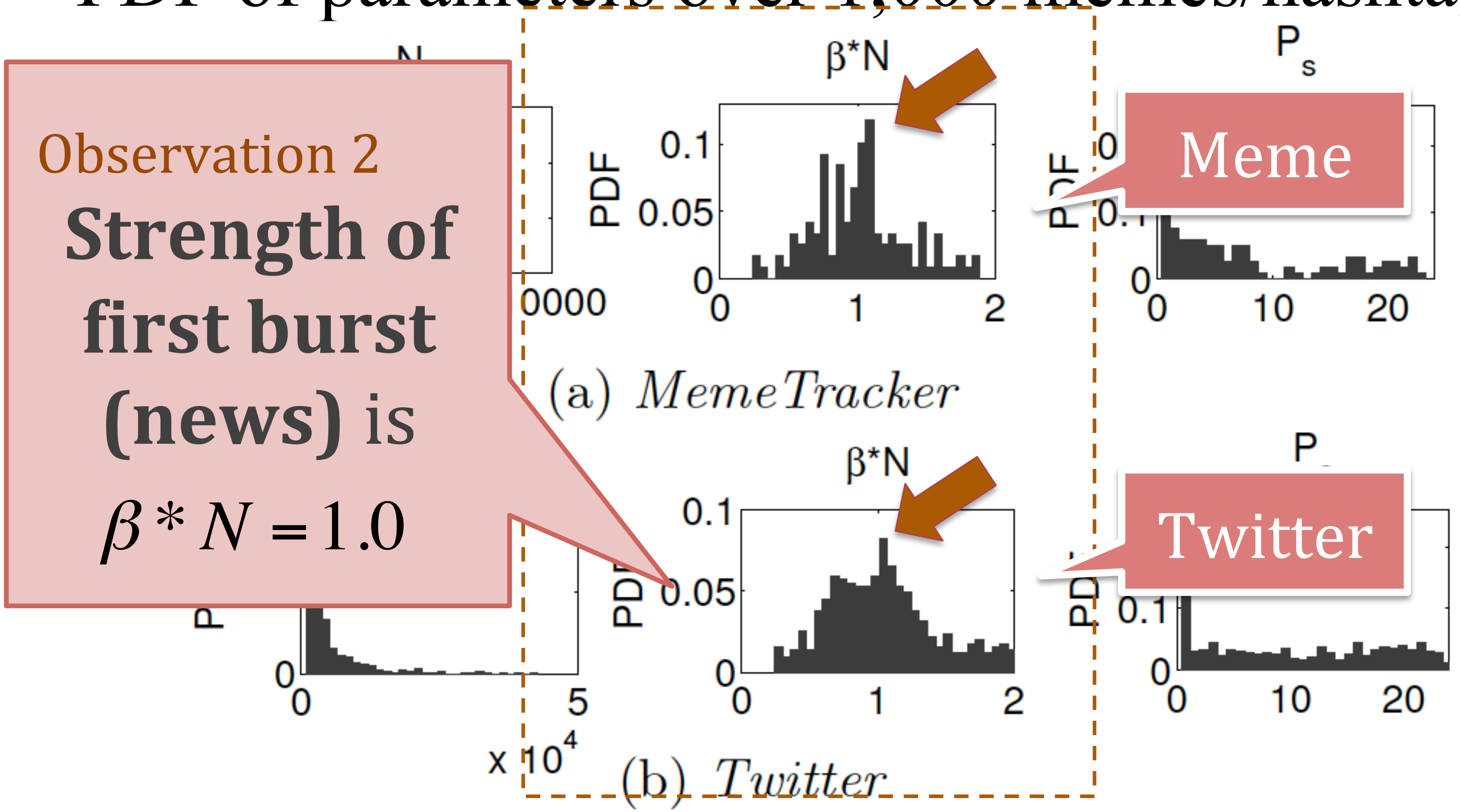
Total population N is almost same

$$N = 1{,}000 \sim 2{,}000$$

A3. Reverse engineering

SpikeM provide an intuitive explanation

PDF of parameters over 1,000 memes/hashtags



Application

A3. Reverse engineering

CMU CS

SpikeM provide an intuitive explanation

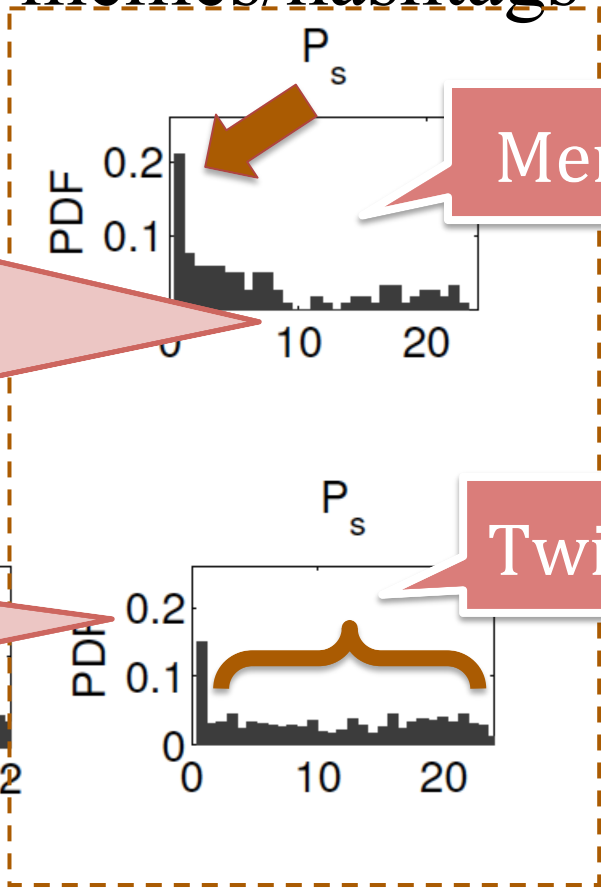
0 memes/hashtags

Observation 3

Daily periodicity

with phase shift $P_s = 0$

Every meme has the same periodicity without lag



(Twitter)

Daily periodicity with

more spread in P_s

(i.e., Multiple time zone)

Kumamoto U

CMU CS

Part 2 Roadmap

Problem

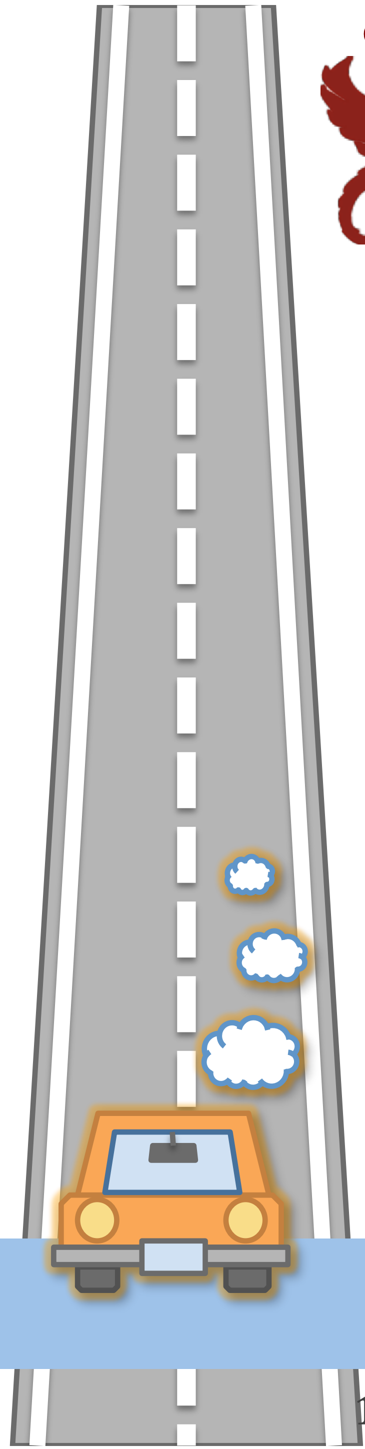
- ✓ Why: “non-linear” modeling

Fundamentals

- ✓ Non-linear (grey-box) models

Applications

- ✓ Epidemics
- ✓ Information diffusion vs.
- – Online competition



Online competition in social networks

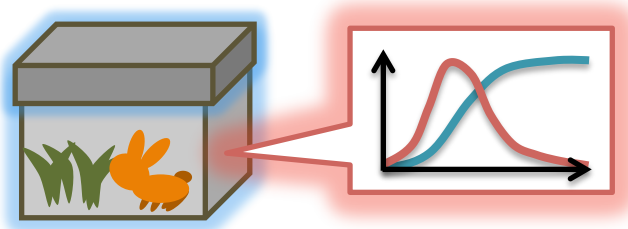
Online competition in social networks

Q. How can we describe "virtual competition"?

Online competition - roadmap

A. Non-linear (gray-box) modeling!

Solutions



- Winner-Takes-All [Prakash+ WWW'12]
- Co-existence of the two viruses [Beutel+ KDD'12]
- The Web as a Jungle [Matsubara+ WWW'15]

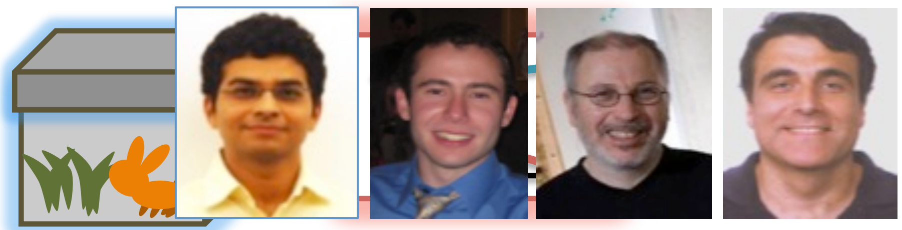
Kumamoto U

Online competition - roadmap

CMU CS

A. Non-linear (gray-box) modeling!

Solutions



- **Winner-Takes-All [Prakash+ WWW'12]**
- Co-existence of the two viruses [Beutel+ KDD'12]
- The Web as a Jungle [Matsubara+ WWW'15]

Competing contagions

[Prakash+ WWW'12]

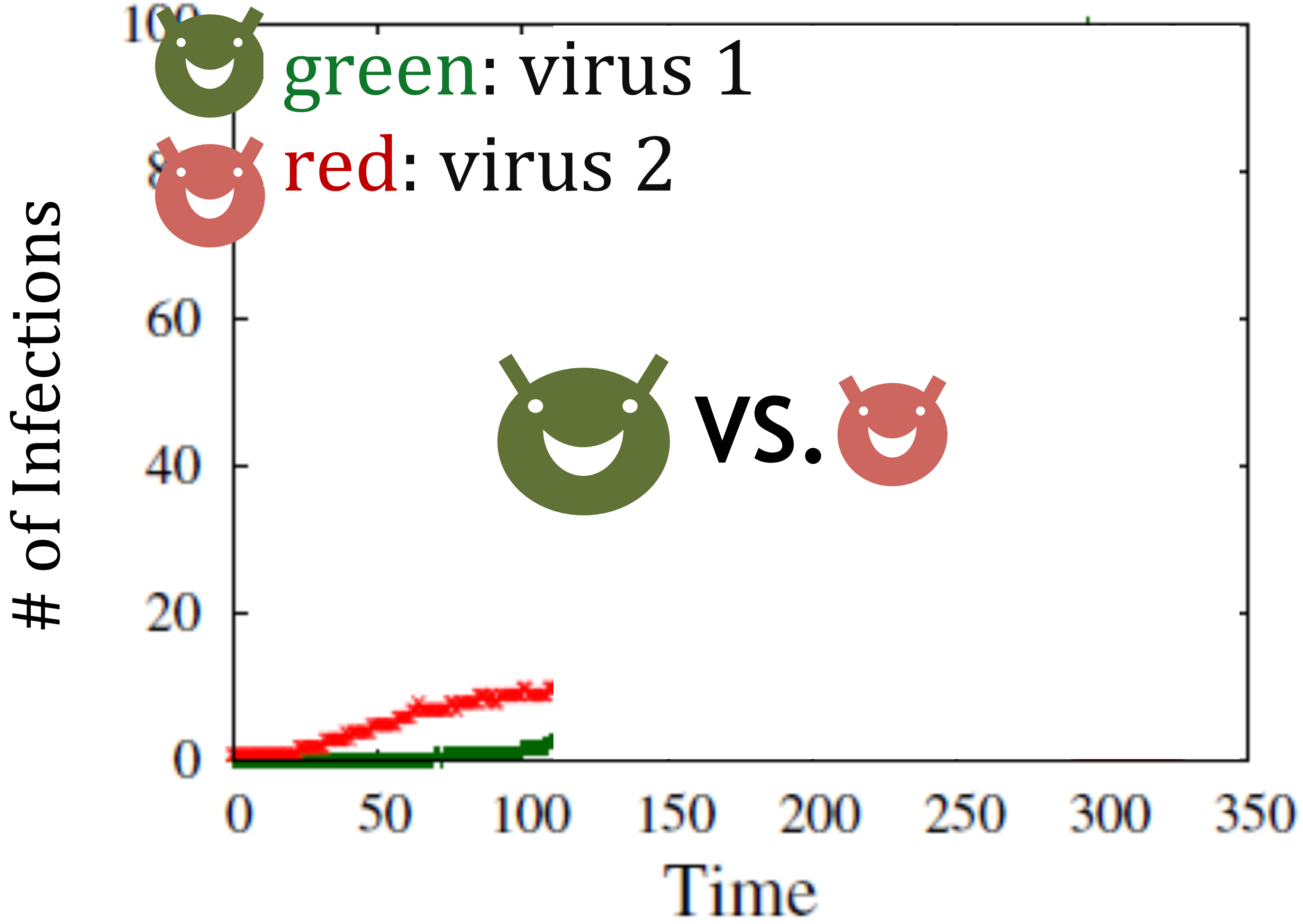
Contagions: viruses, online activities

iPhone v **Android** **Blu-ray** v **HD-DVD**

Q. What happen when two viruses compete?

Competing contagions

[Prakash+ WWW'12]



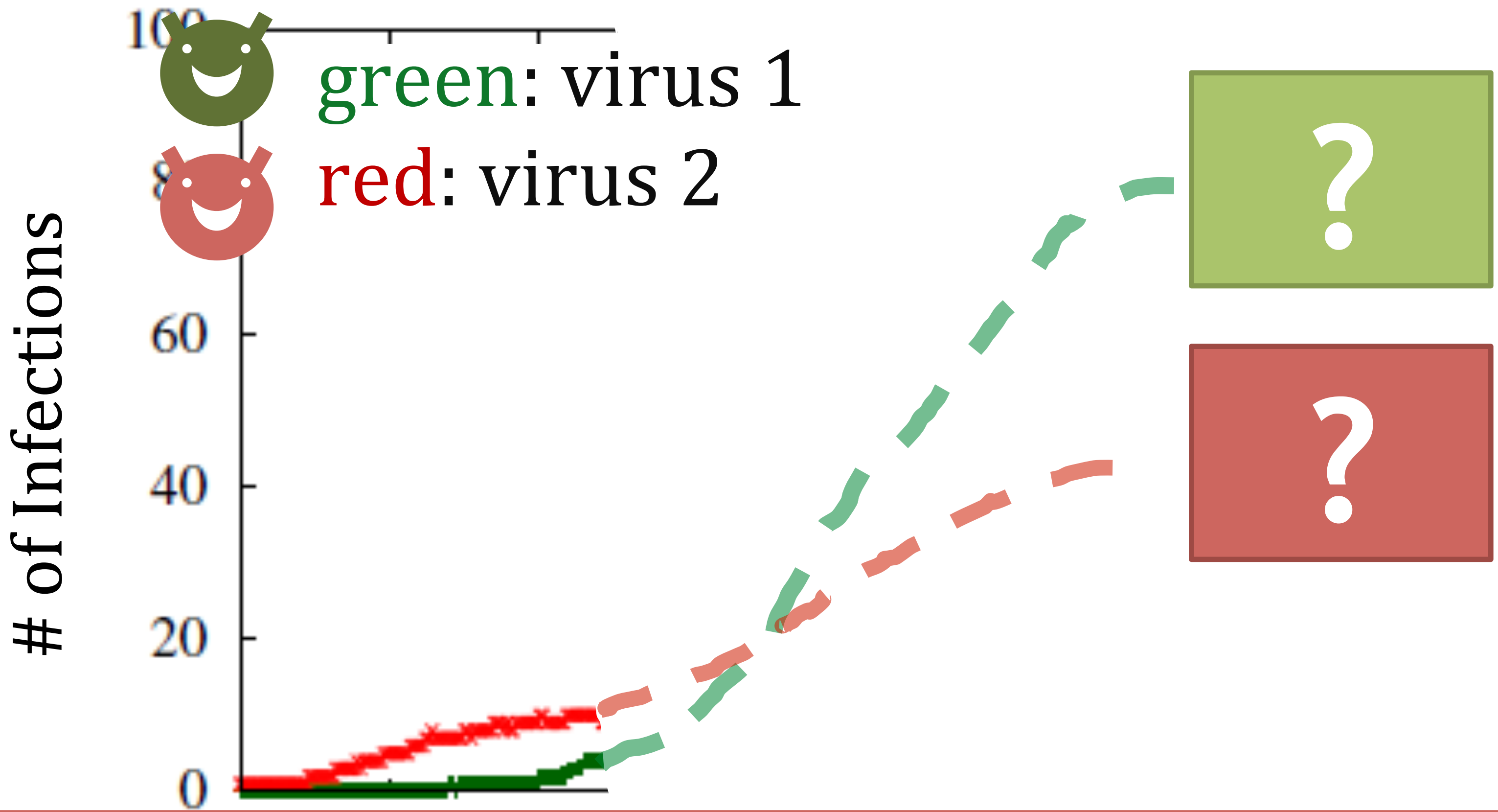
ASSUME: Virus 1 is stronger than Virus 2

Kumamoto U

CMU CS

Competing contagions

[Prakash+ WWW'12]



Q: What happens in the end?

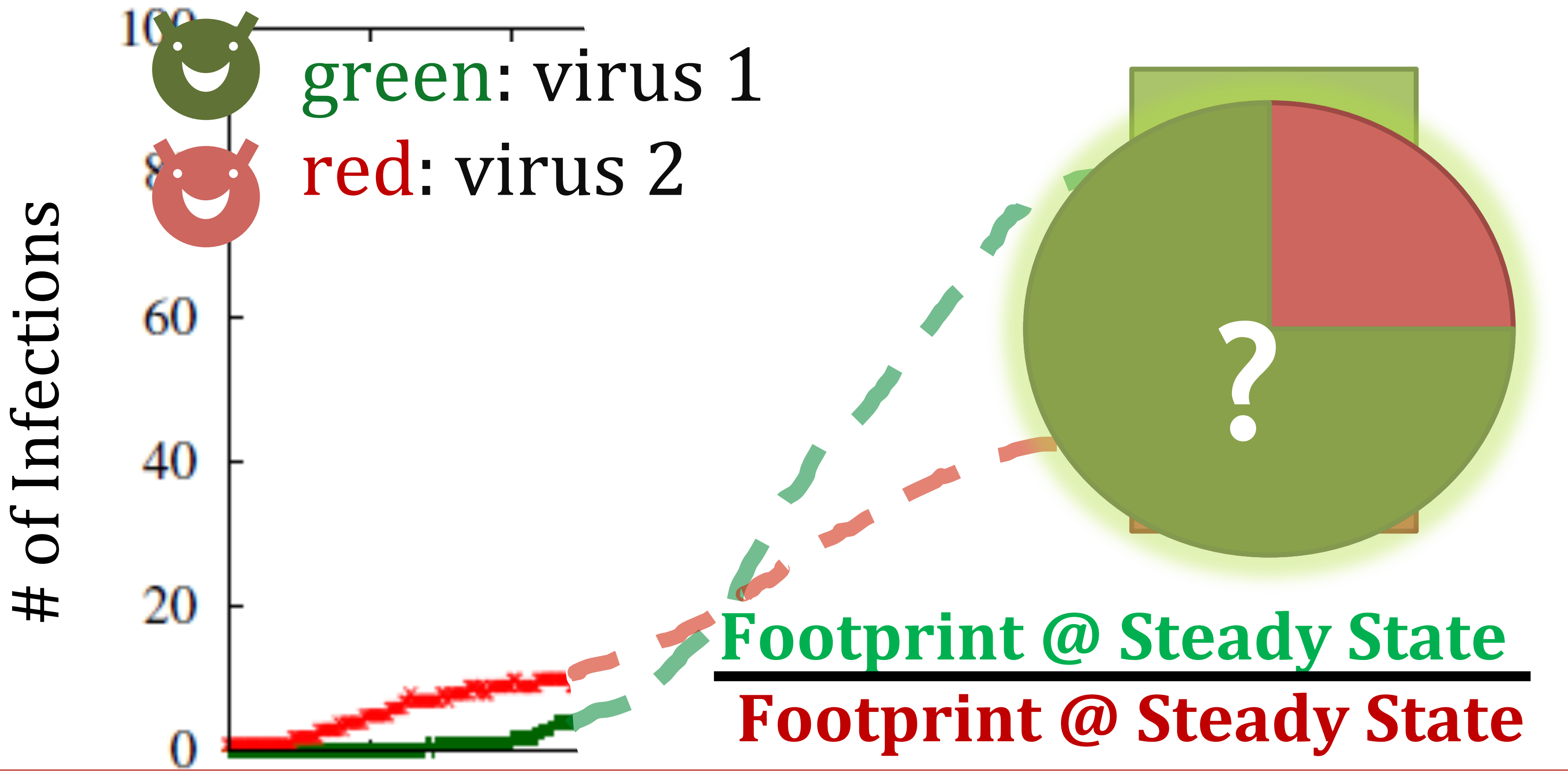
ASSUME: Virus 1 is stronger than virus 2

Kumamoto U

CMU CS

Competing contagions

[Prakash+ WWW'12]



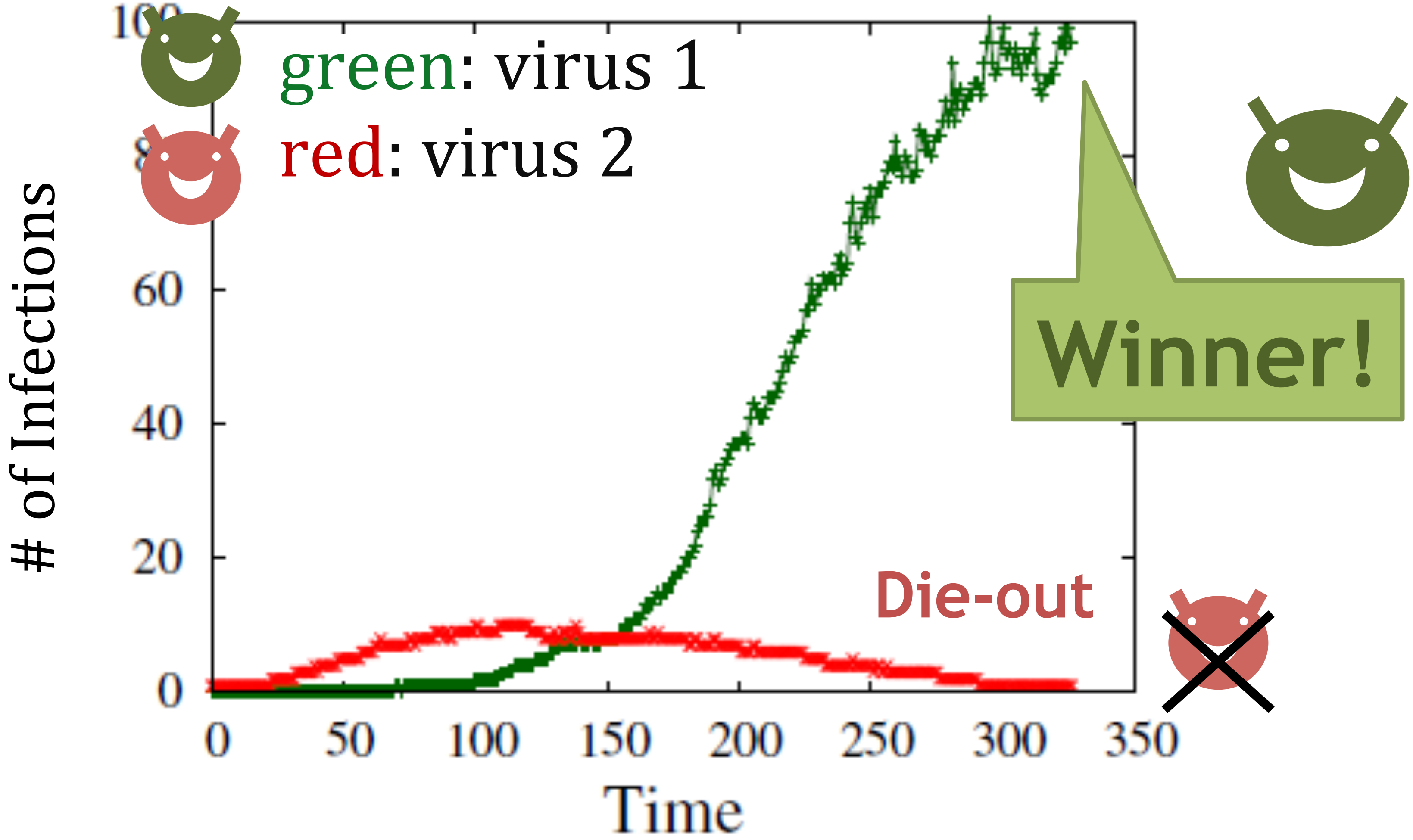
$$\frac{\text{Footprint @ Steady State}}{\text{Footprint @ Steady State}}$$

Q: What happens in the end?

ASSUME: Virus 1 is stronger than Virus 2

Answer: Winner-Takes-All!

[Prakash+ WWW'12]

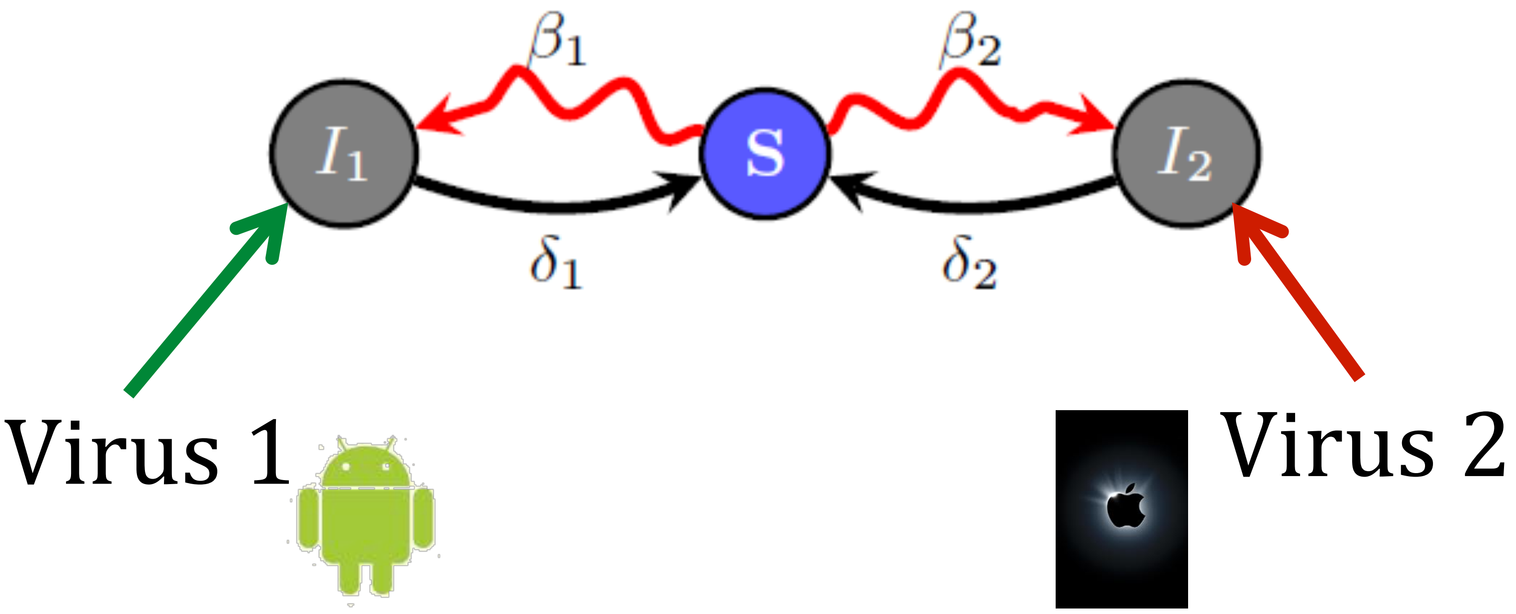


ASSUME: Virus 1 is stronger than Virus 2

A simple model

[Prakash+ WWW'12]

- Modified flu-like (SIS) model
- Mutual Immunity ("pick one of the two")
- Susceptible-Infected1-Infected2-Susceptible



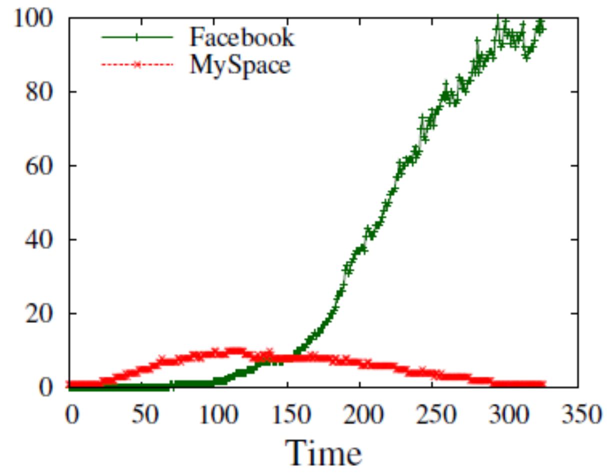
Kumamoto U

Result: Winner-Takes-All

CMU CS

[Prakash+ WWW'12]

Given this model,
and ***any graph***,
the weaker virus always
dies-out, *completely*

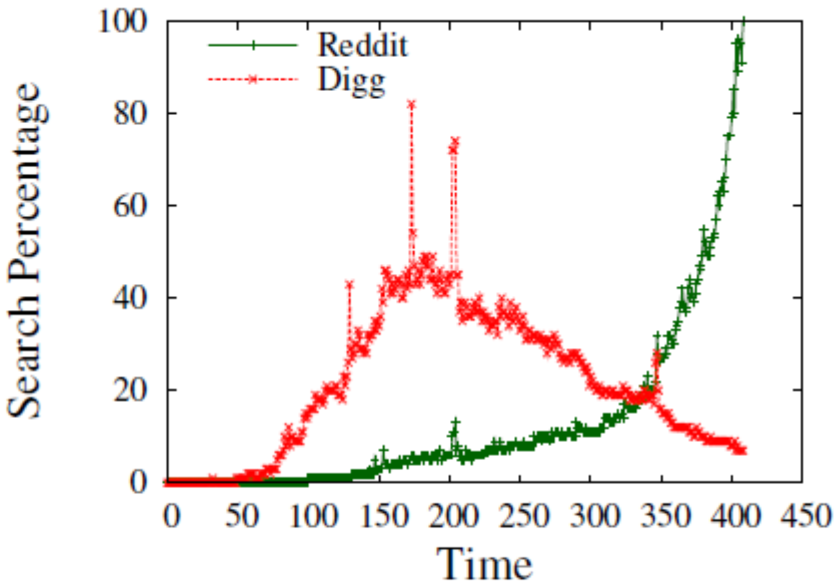


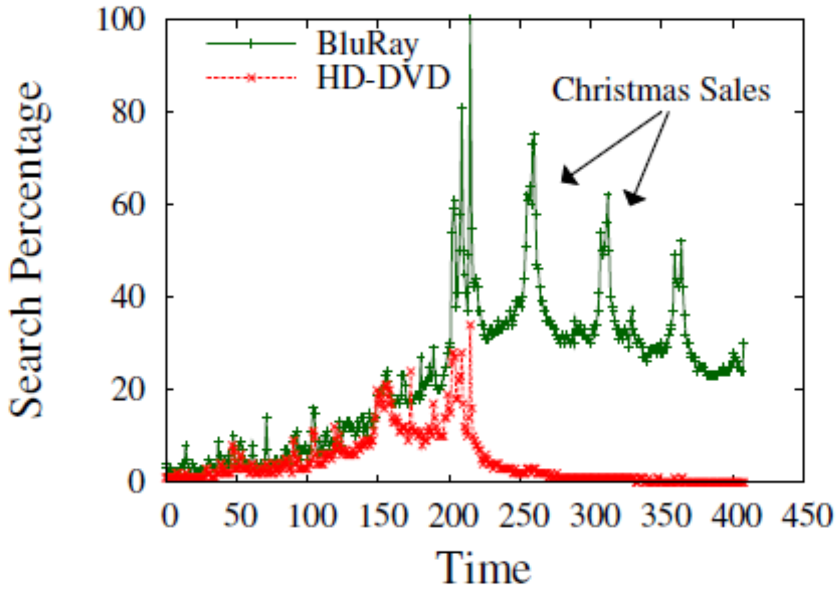
1. The stronger survives only if it is above threshold
2. Virus 1 is stronger than Virus 2, if:
   strength(Virus 1) > strength(Virus 2)
3. Strength(Virus) = $\lambda \beta / \delta$ → same as before!

Real Examples of “WTA”

[Prakash+ WWW’12]

[Google Search Trends data]



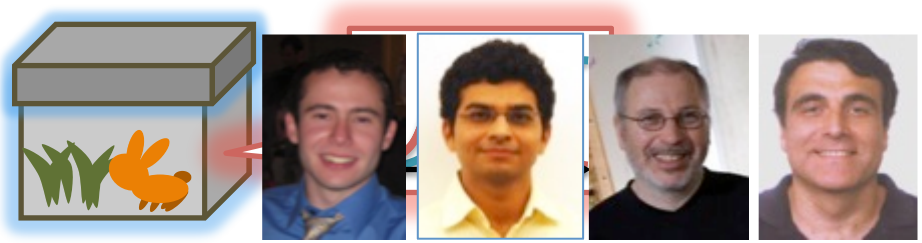


Reddit v **Digg**

Blu-Ray v **HD-DVD**

Online competition in social networks

A. Non-linear (gray-box) modeling!



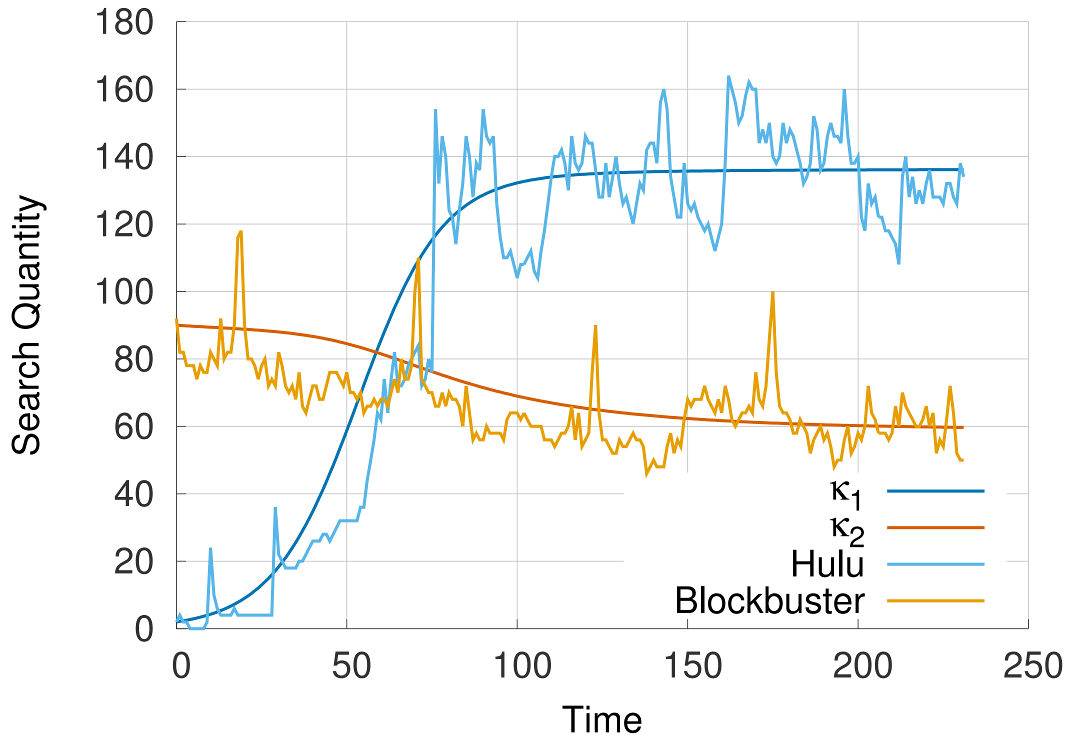
Solutions

- Winner-Takes-All [Prakash+ WWW'12]
- **Co-existence of the two viruses [Beutel+ KDD'12]**
- The Web as a Jungle [Matsubara+ WWW'15]

Interacting Viruses: Can Both Survive?

Real example of "co-existence"

[Google Search Trends data]



Hulu v **Blockbuster**

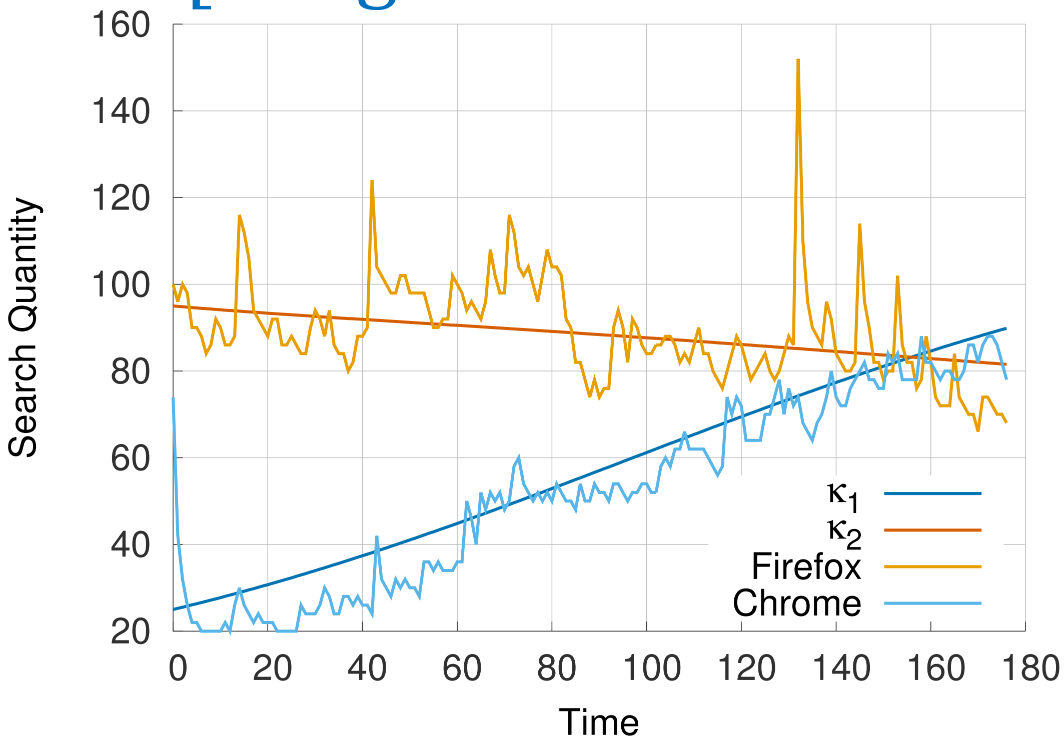
Kumamoto U

Interacting Viruses: Can Both Survive?

CMU CS

Real example of "co-existence"

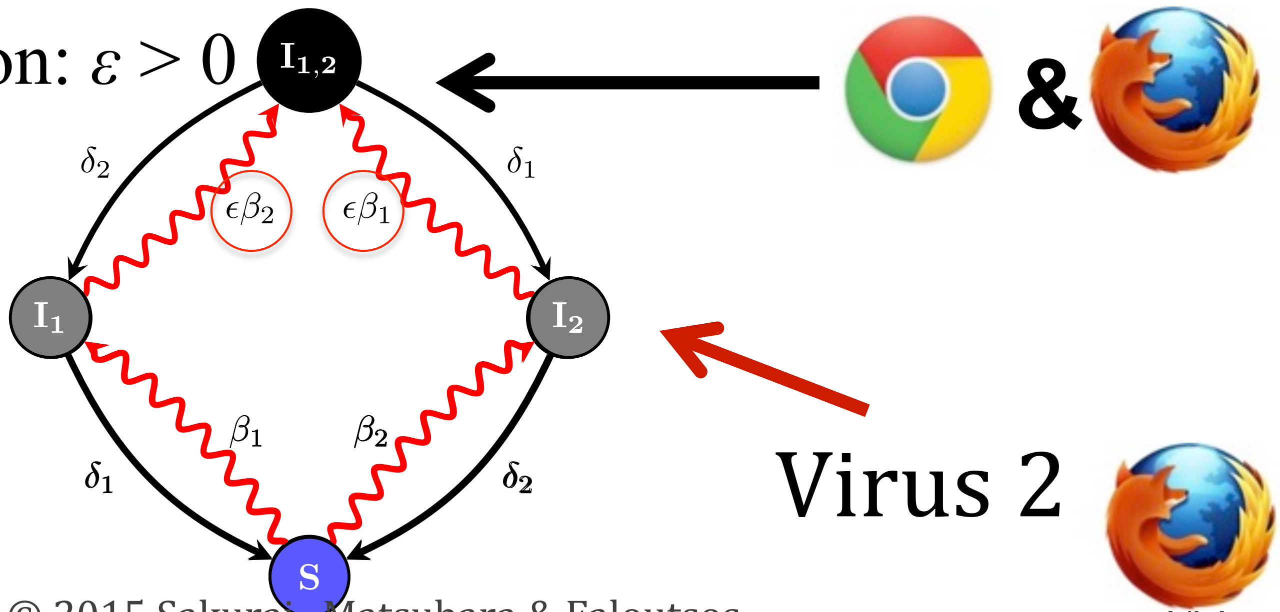
[Google Search Trends data]



Chrome v **Firefox**

A simple model: $SI_{1|2}S$

- Modified flu-like (SIS)
- Susceptible-Infected$_{1 \text{ or } 2}$-Susceptible
- Interaction Factor ε
  - Full Mutual Immunity: $\varepsilon = 0$
  - Partial Mutual Immunity (competition): $\varepsilon < 0$
  - Cooperation: $\varepsilon > 0$



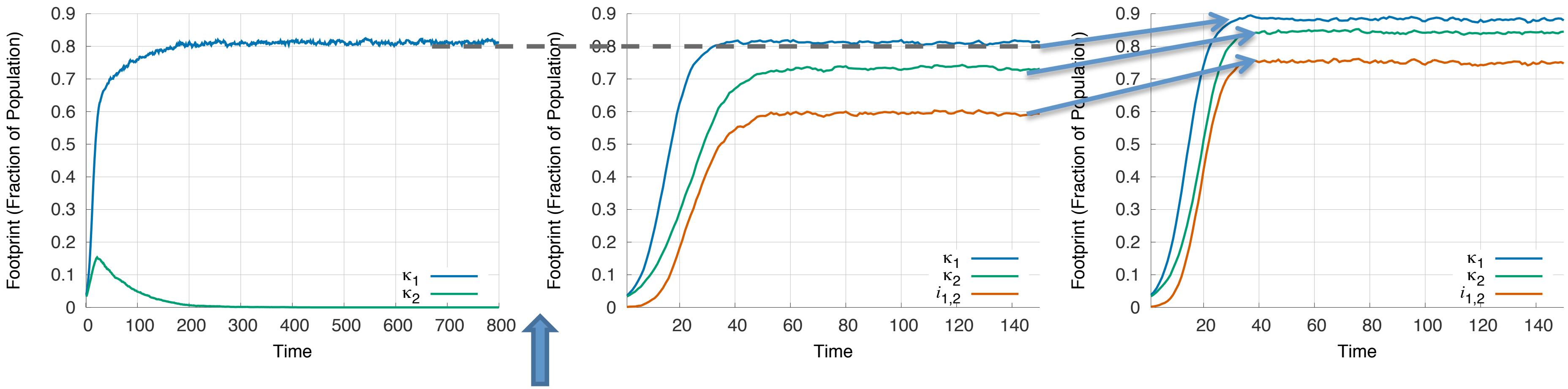
Virus 1

Kumamoto U

CMU CS

Question: What happens in the end?

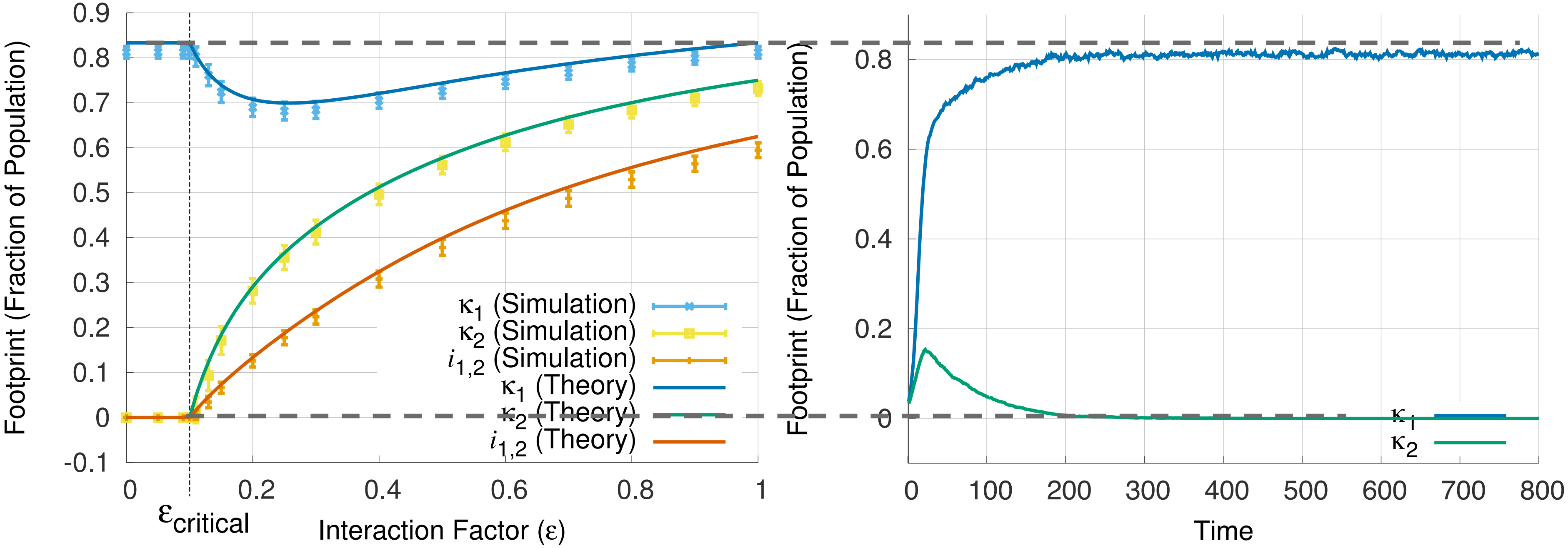
$\varepsilon = 0$	$\varepsilon = 1$	$\varepsilon = 2$
Winner takes all	Co-exist independently	Viruses cooperate



What about for $0 < \varepsilon < 1$?
Is there a point at which both viruses can *co-exist*?

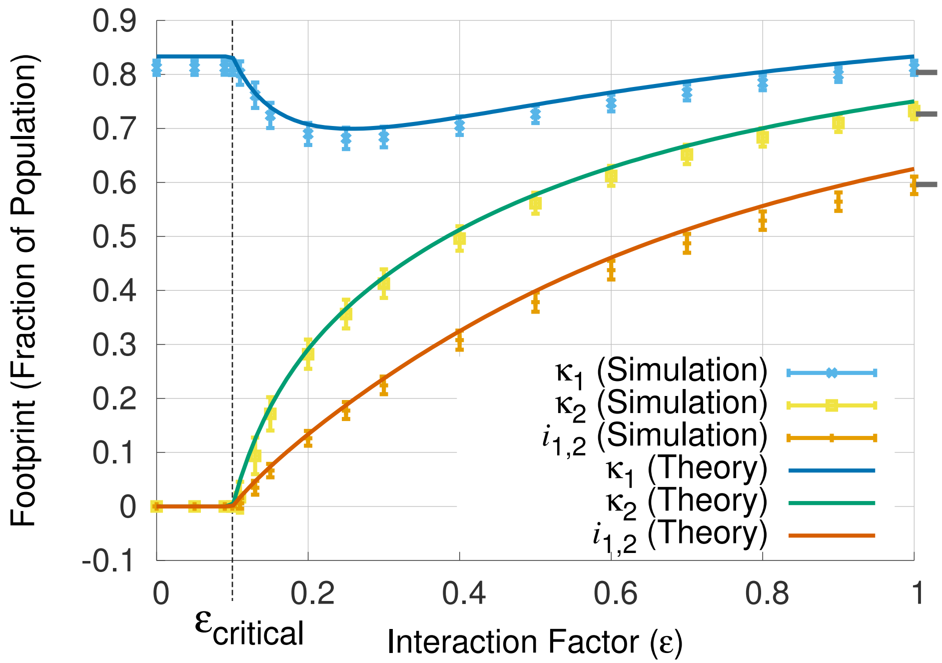
ASSUME: Virus 1 is stronger than Virus 2

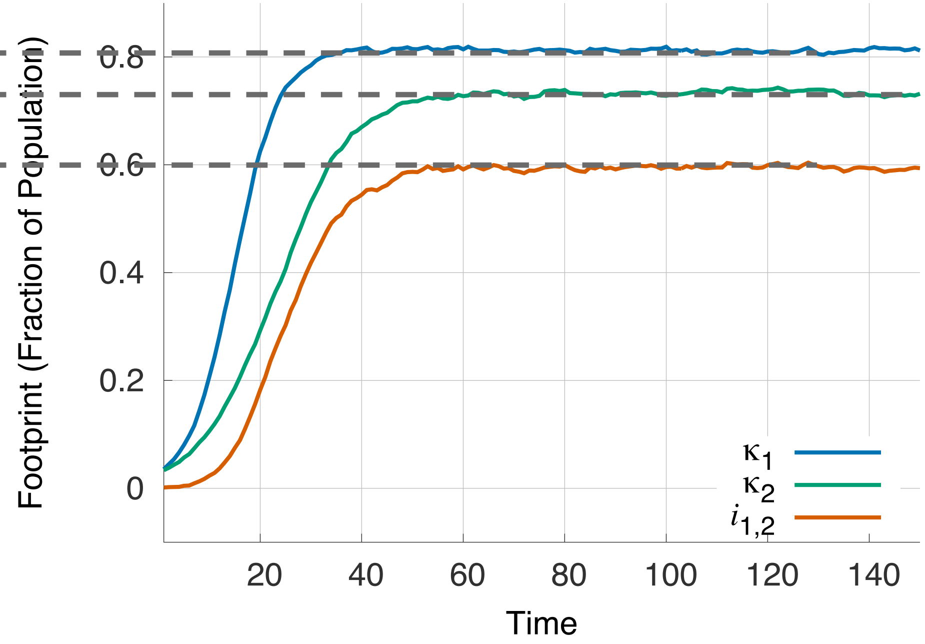
Answer: Yes! There is a phase transition



ASSUME: Virus 1 is stronger than Virus 2

Answer: Yes!
There is a phase transition



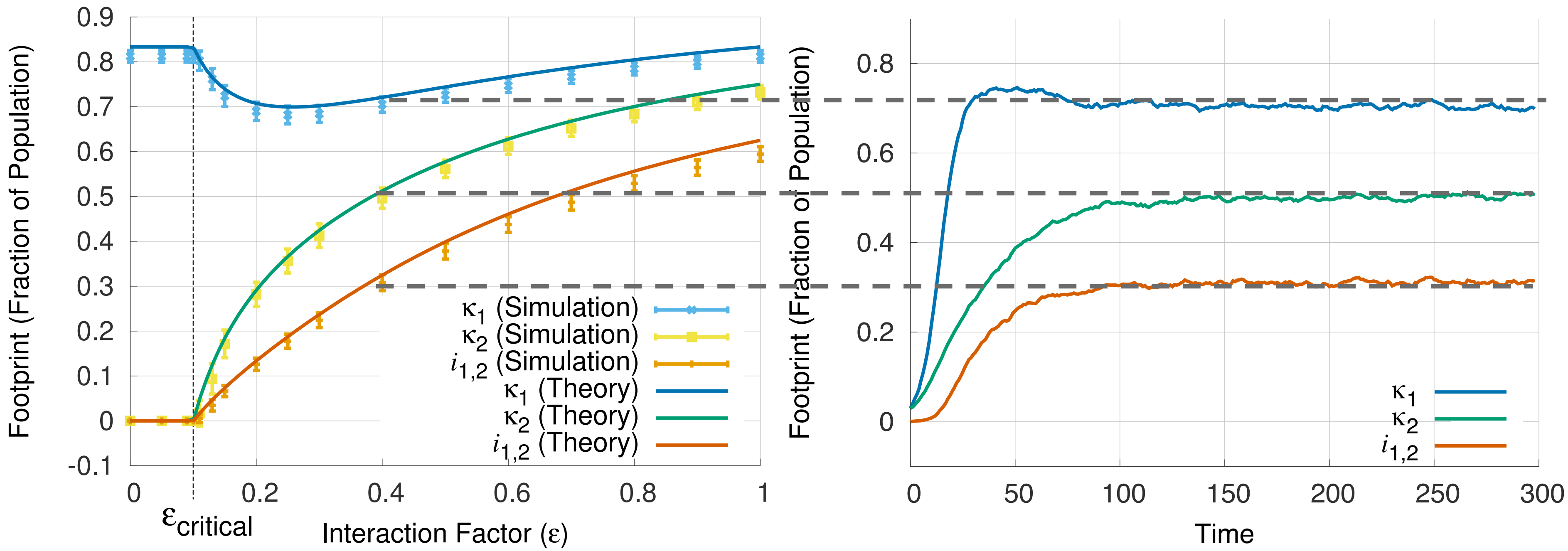


ASSUME: Virus 1 is stronger than Virus 2

Kumamoto U

CMU CS

Answer: Yes! There is a phase transition



ASSUME: Virus 1 is stronger than Virus 2

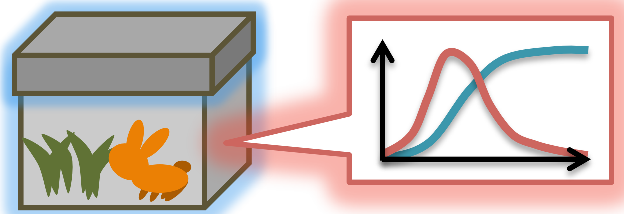
Result: Viruses can Co-exist

Given this model and a fully connected graph, there exists an $\varepsilon_{critical}$ such that for $\varepsilon \geq \varepsilon_{critical}$, there is a fixed point where both viruses survive.

1. The stronger survives only if it is above threshold
2. Virus 1 is stronger than Virus 2, if:
   strength(Virus 1) > strength(Virus 2)
3. Strength(Virus) $\sigma = N \beta / \delta$

Online competition in social networks

A. Non-linear (gray-box) modeling!



Solutions

- Winner-Takes-All [Prakash+ WWW'12]
- Co-existence of the two viruses [Beutel+ KDD'12]
- **The Web as a Jungle [Matsubara+ WWW'15]**

[Matsubara+ WWW'15]

The Web as a Jungle: Non-Linear Dynamical Systems for Co-evolving Online Activities

Yasuko Matsubara (Kumamoto University)
Yasushi Sakurai (Kumamoto University)
Christos Faloutsos (CMU)

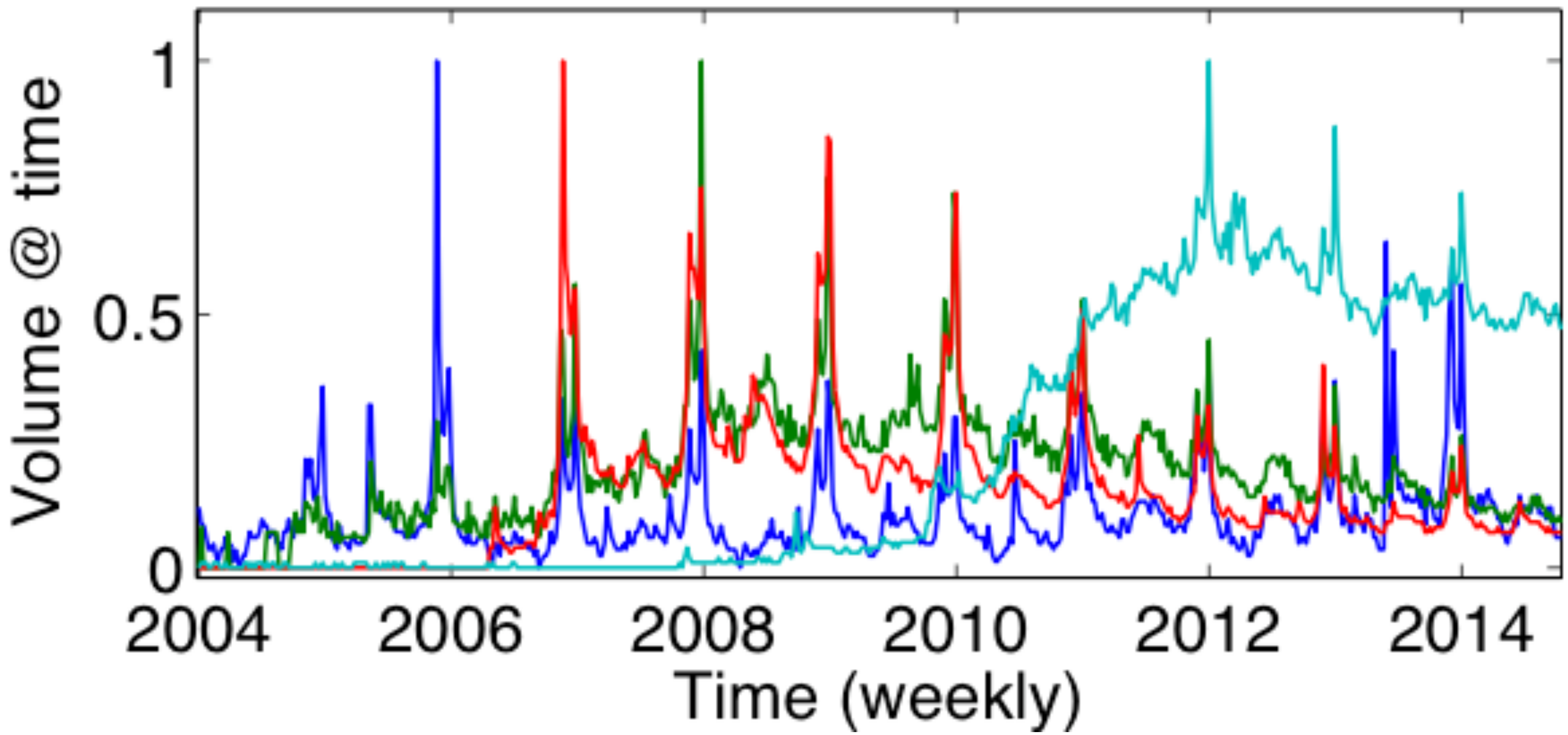
Kumamoto U

Given: online user activities

CMU CS

e.g., *Google* search volumes for

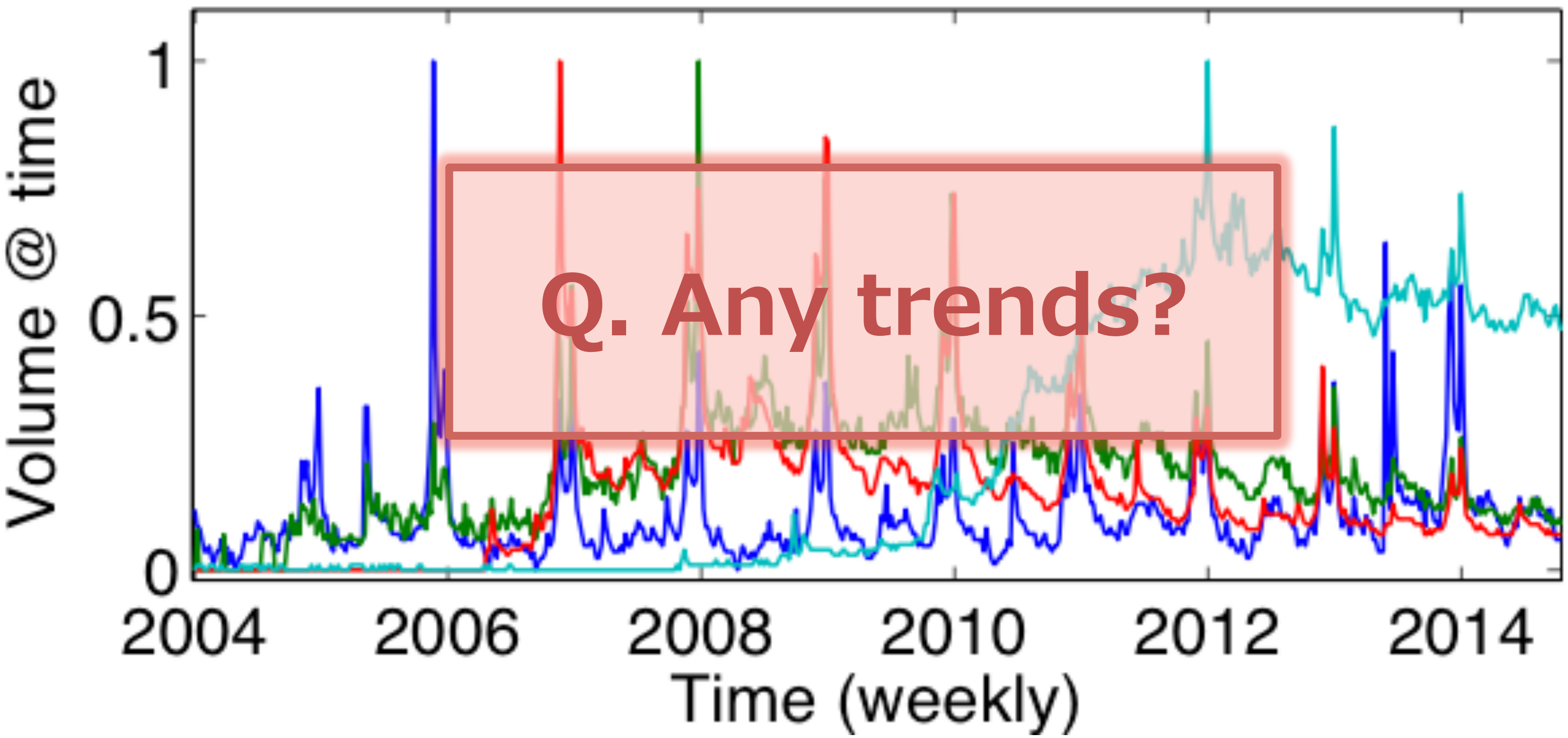
Xbox, PlayStation, Wii, Android



Given: online user activities

e.g., *Google* search volumes for

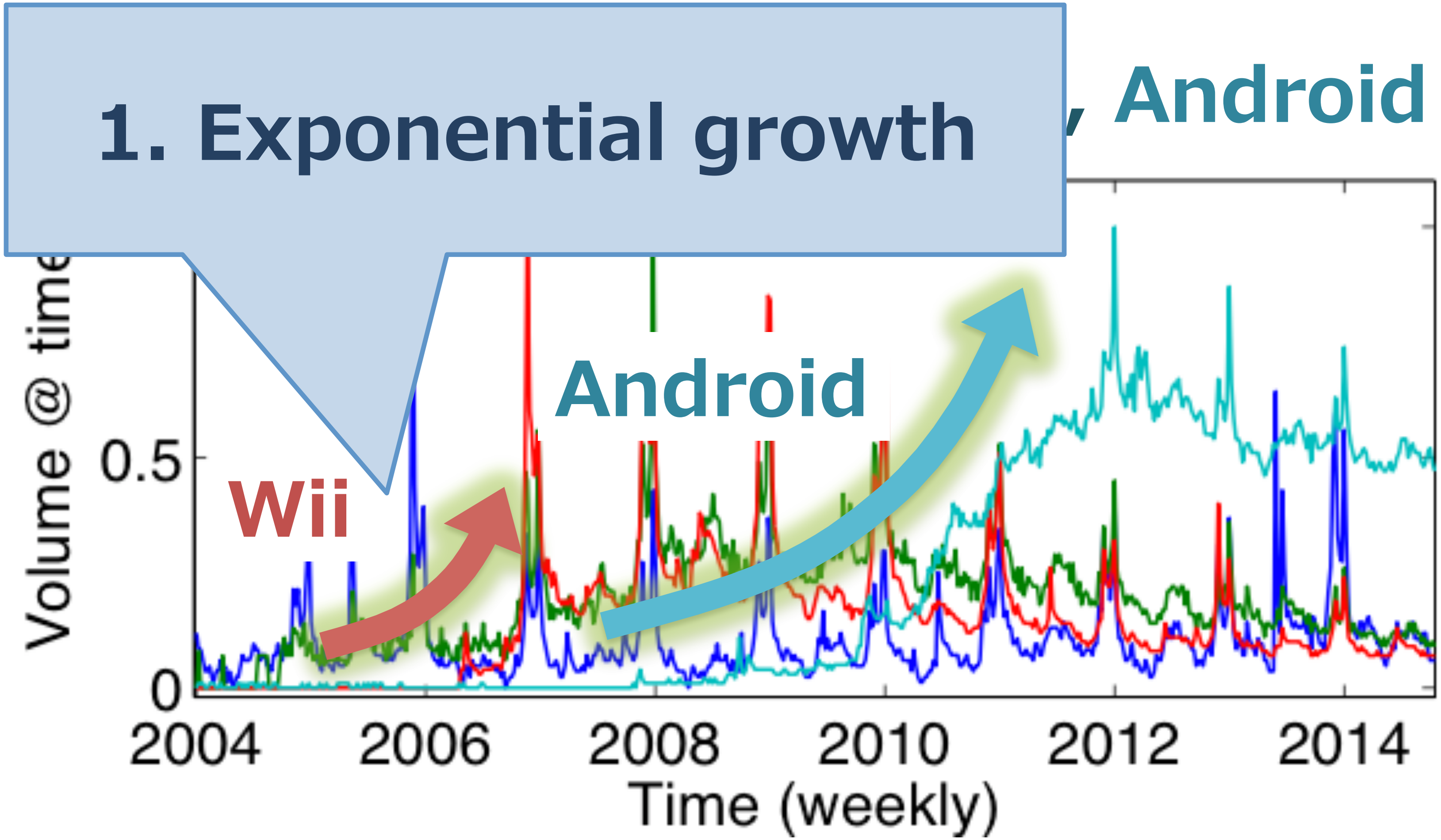
Xbox, PlayStation, Wii, Android



Given: online user activities

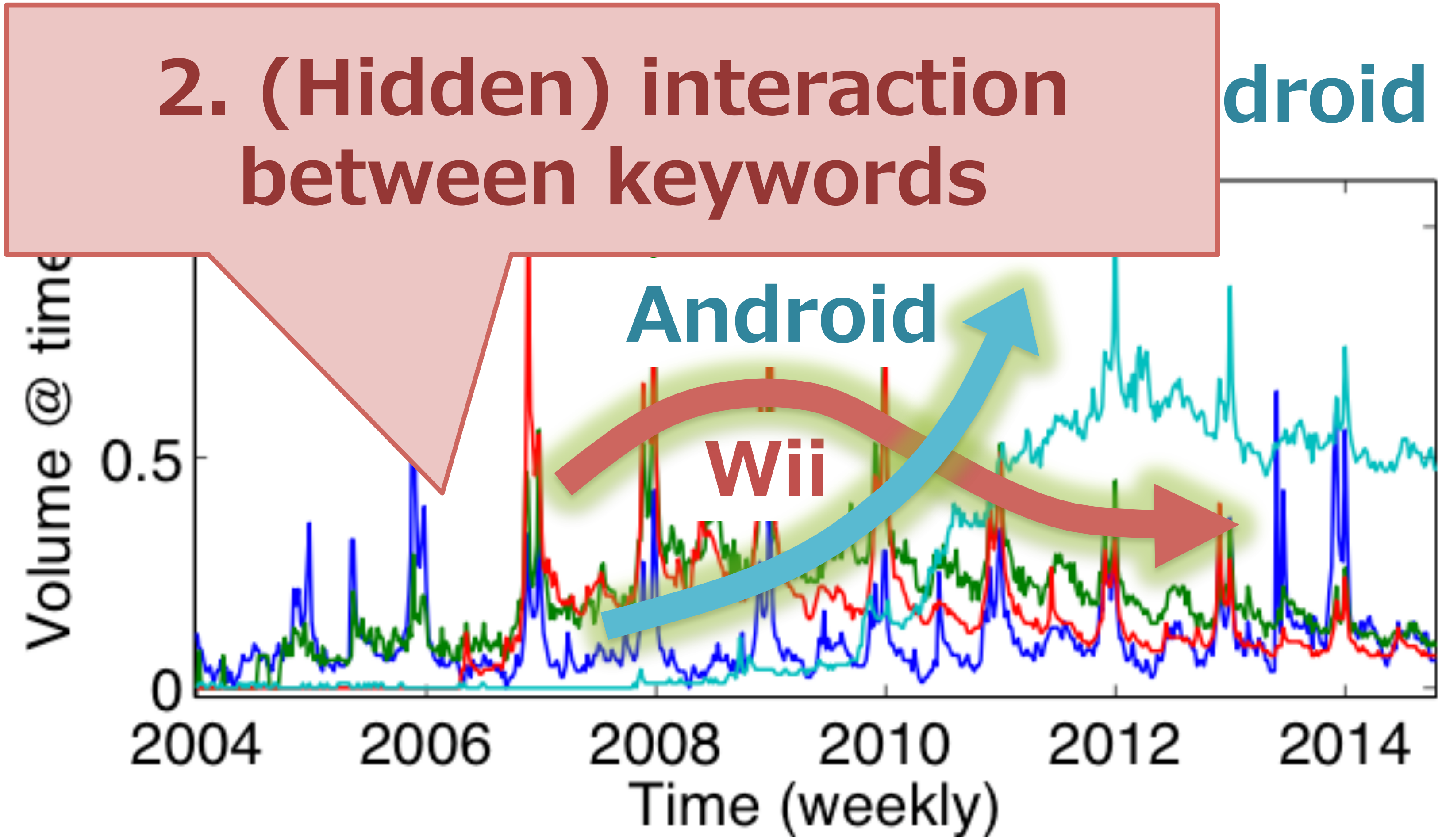
e.g., *Google* search volumes for

, Android



Given: online user activities

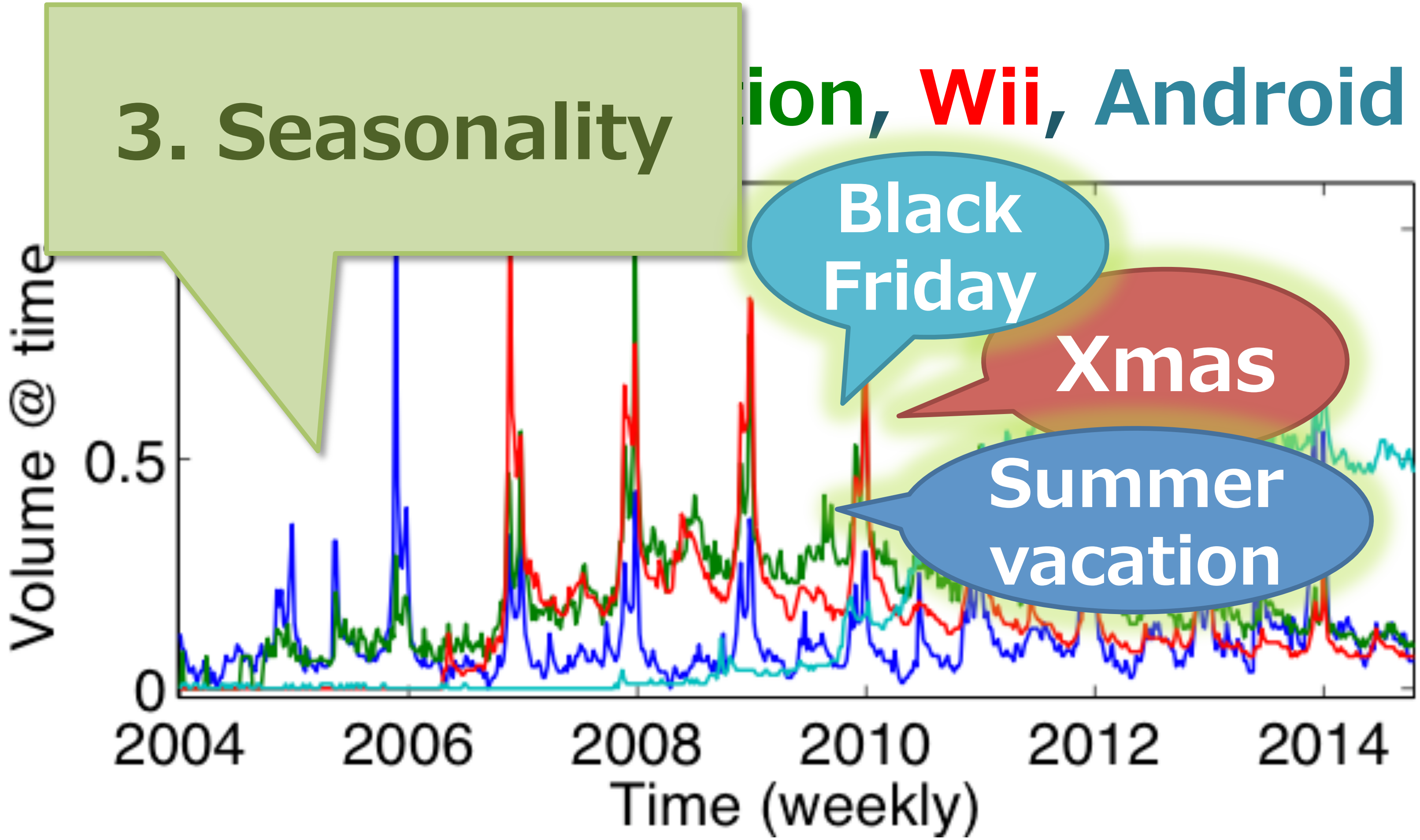
e.g., *Google* search volumes for



Given: online user activities

e.g., *Google* search volumes for

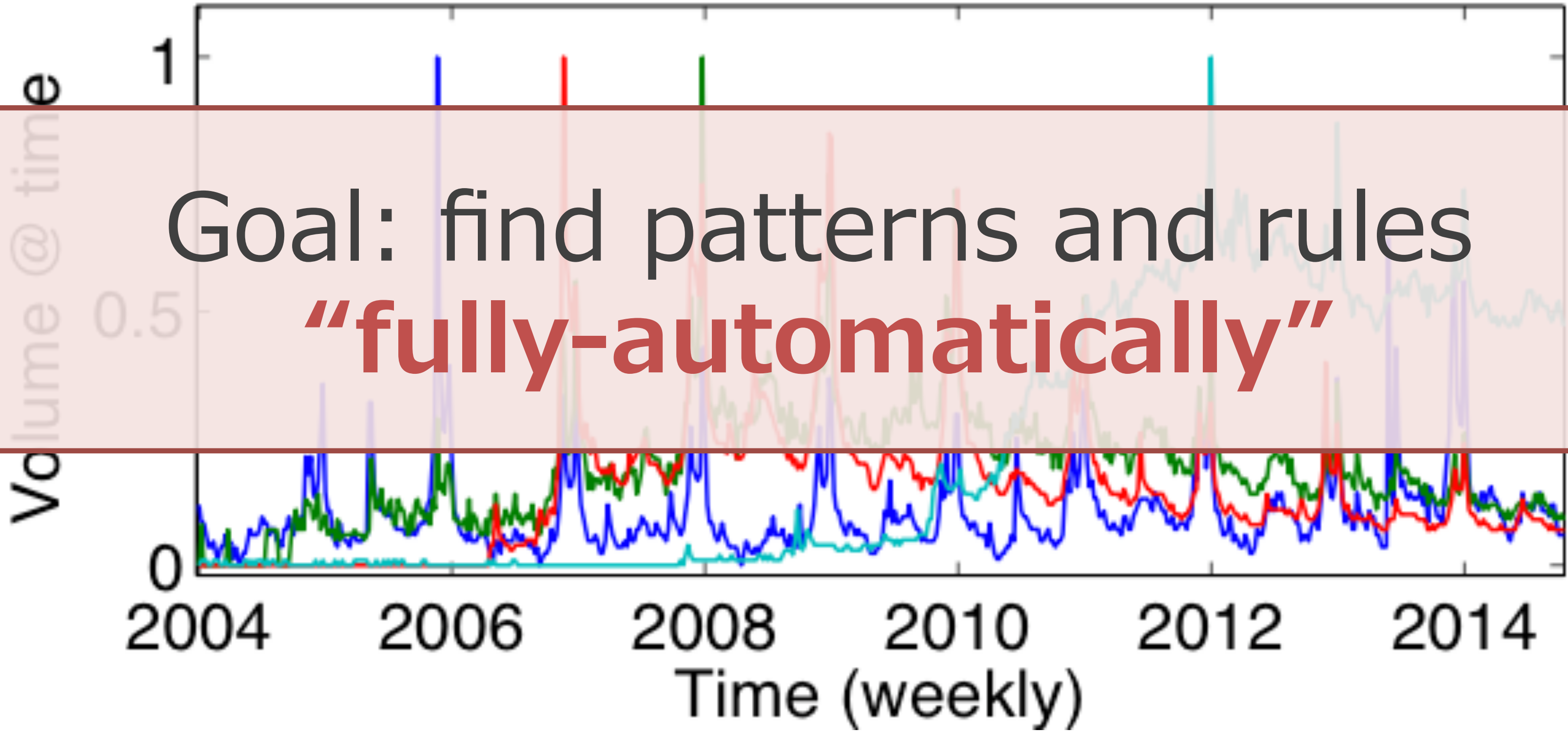
…ion, Wii, Android



Given: online user activities

e.g., *Google* search volumes for

Xbox, PlayStation, Wii, Android



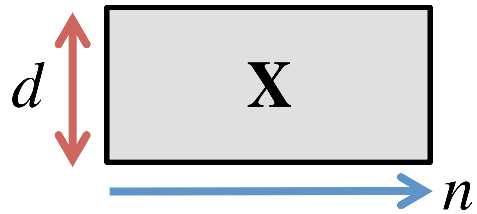
Goal: find patterns and rules

"fully-automatically"

Problem definition

Given: Co-evolving online activities

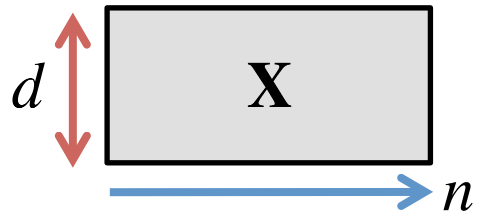
X (activity x time)



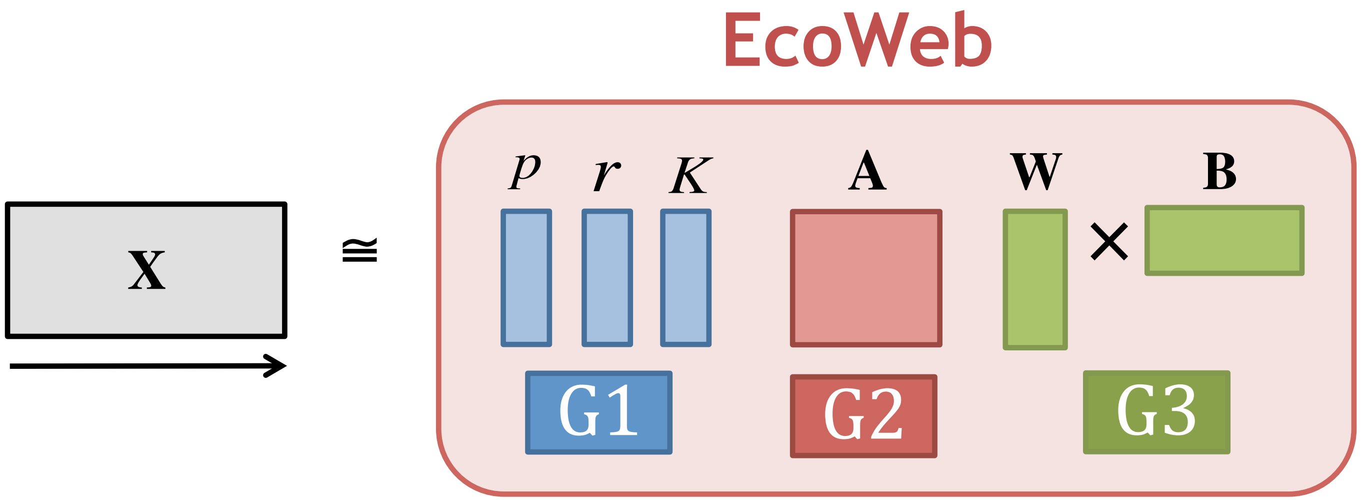
Problem definition

Given: Co-evolving online activities

X (activity x time)



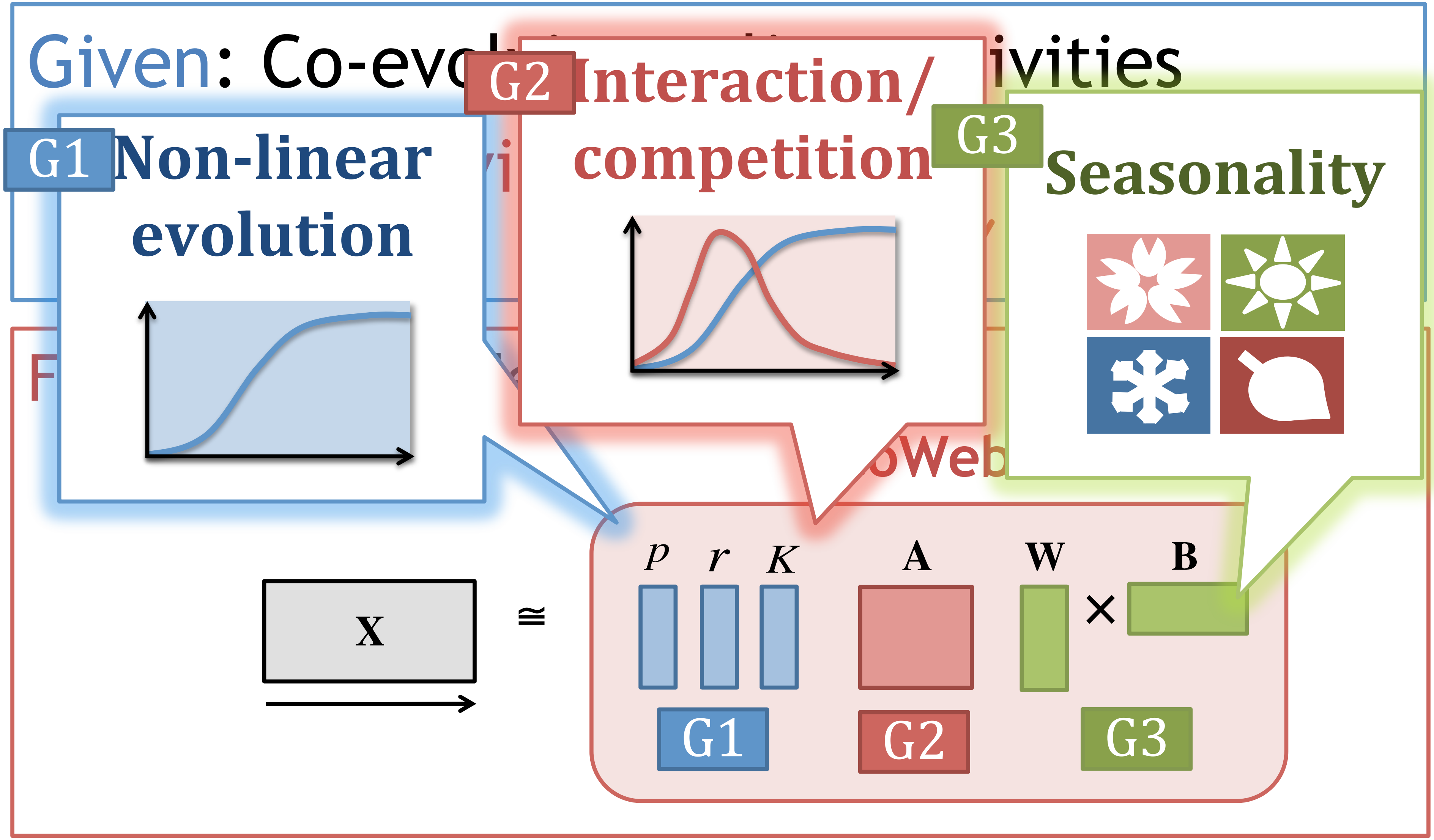
Find: Compact description of X



Kumamoto U

CMU CS

Problem definition



Problem definition

Given: Co… …s

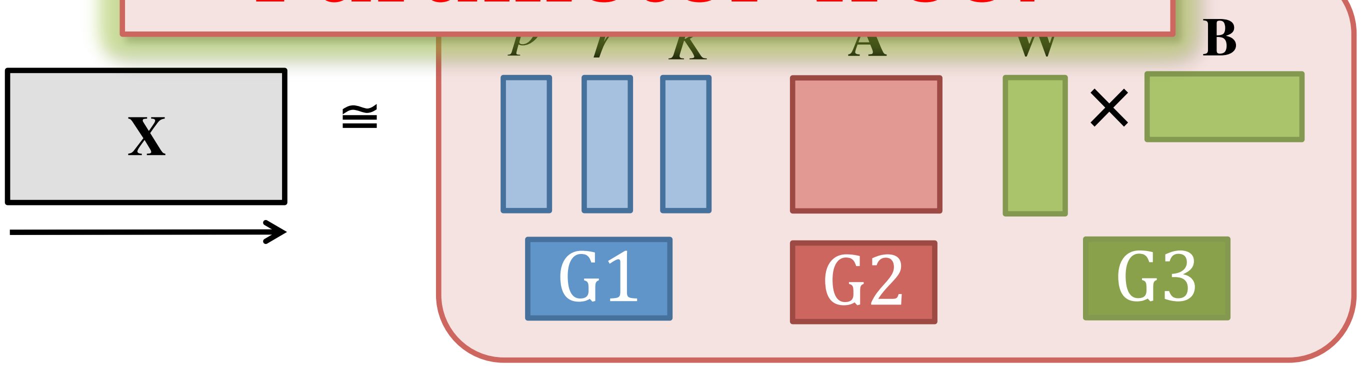
X (a…

n

Find: Comp…

NO magic numbers !

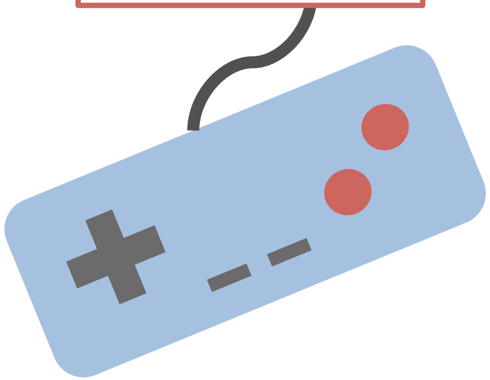
Parameter-free!



Modeling power of EcoWeb

Questions

Q1



Q2

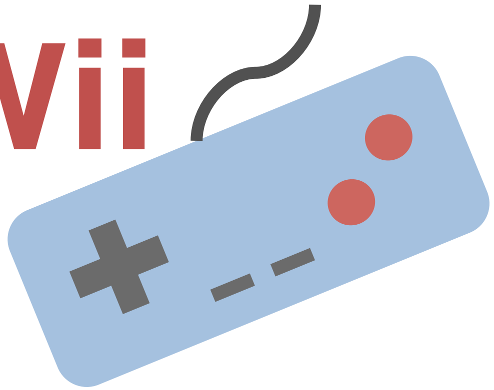
Q3

amazon

Modeling power of EcoWeb

Q1 (games)

Who is the competitor?

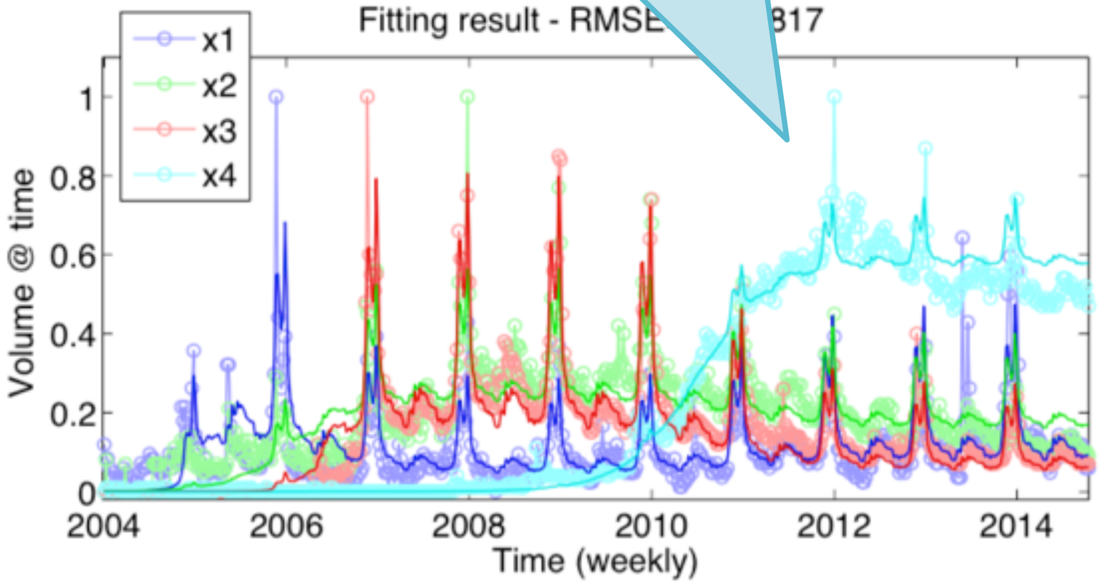
Wii  vs. ?

Modeling power of EcoWeb

Kumamoto U

CMU CS

A. Android!

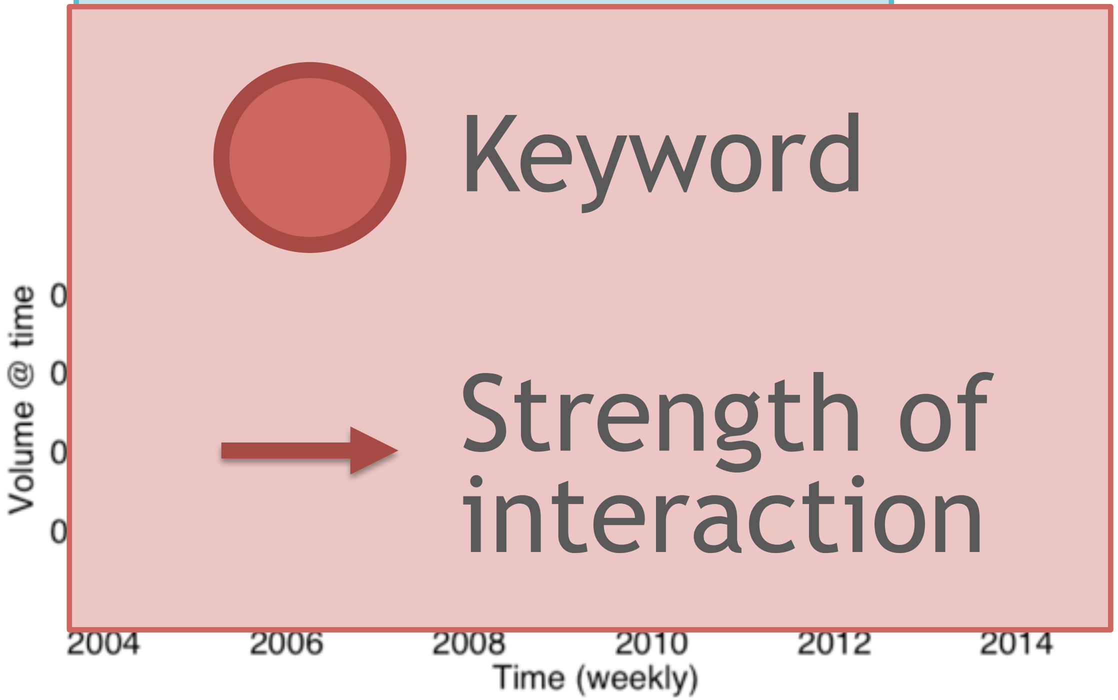


EcoWeb-Fit

Interaction network (latent)

Modeling power of EcoWeb

A. Android!



EcoWeb-Fit

Interaction network (latent)

Kumamoto U

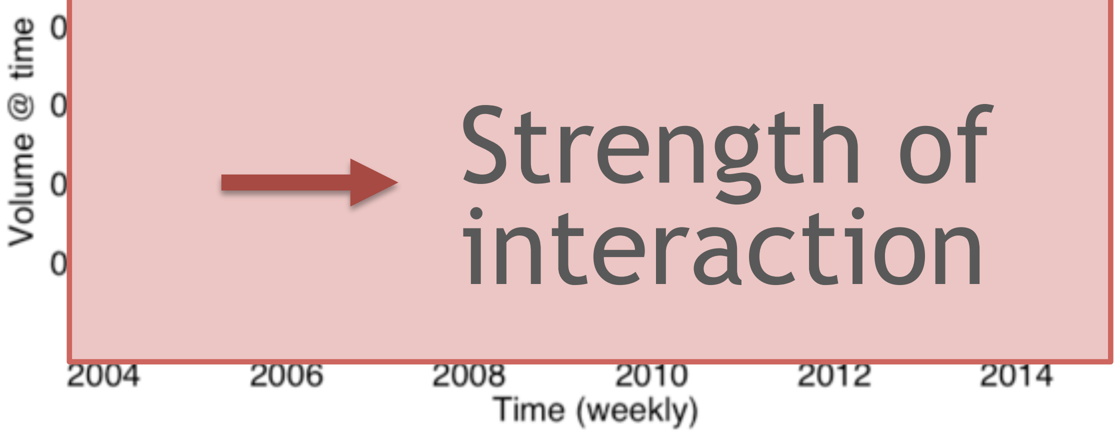
Modeling power of EcoWeb

CMU CS

A. Android!

Keyword

Strength of interaction



EcoWeb-Fit

Interaction network (latent)

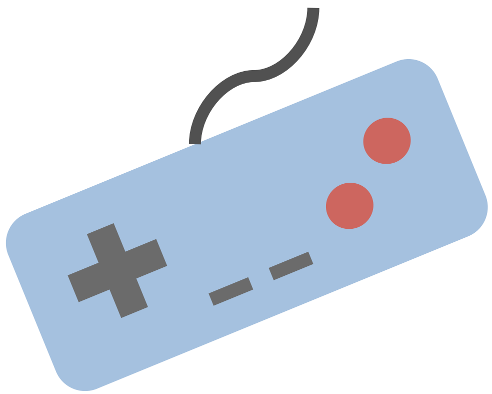
Kumamoto U

Modeling power of EcoWeb

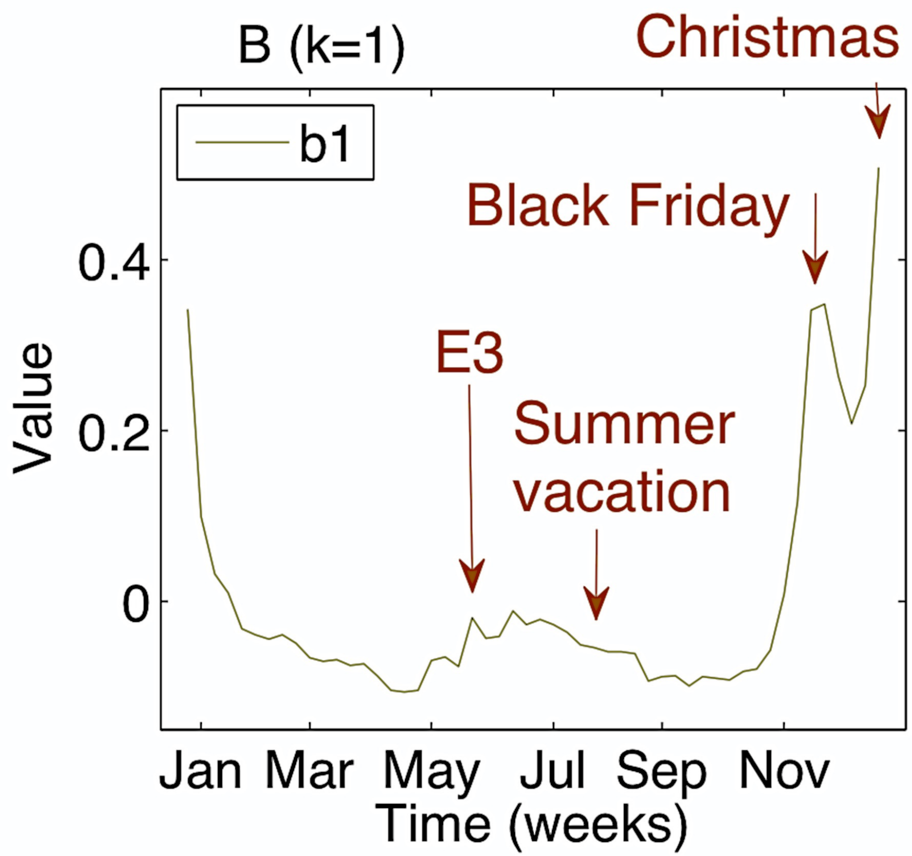
CMU CS

Q1 (games)

Any seasonal events?



Modeling power of EcoWeb



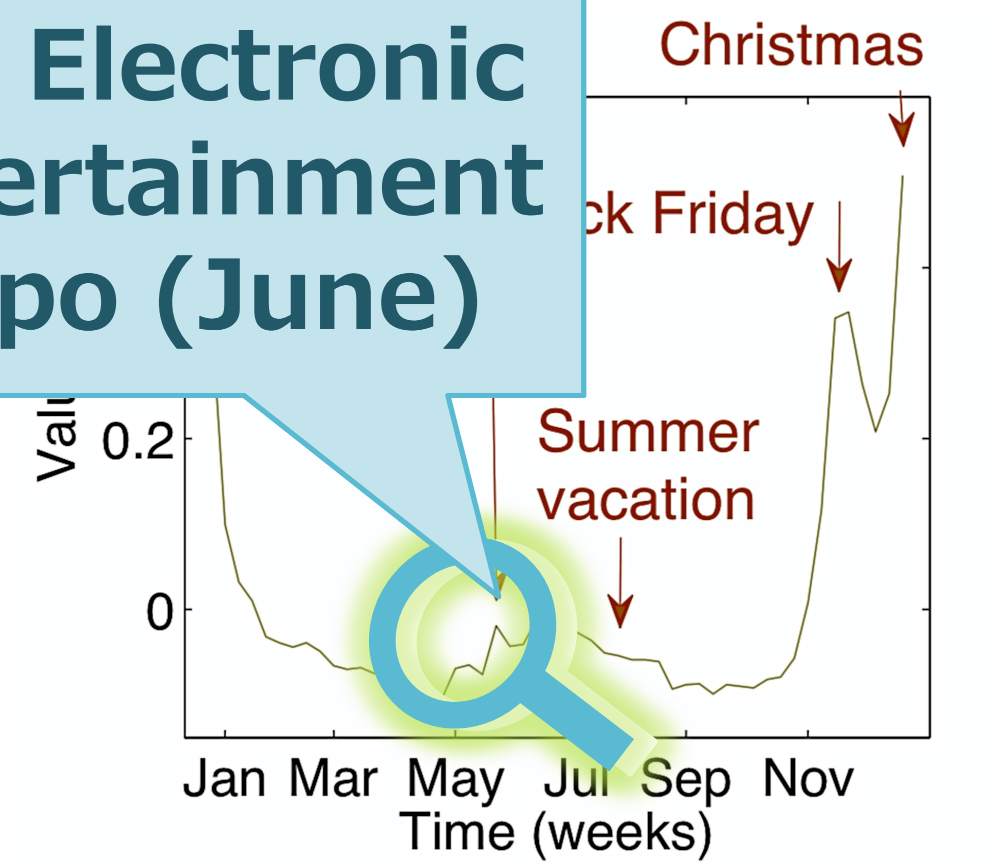
EcoWeb: seasonal component

Kumamoto U

Modeling power of EcoWeb

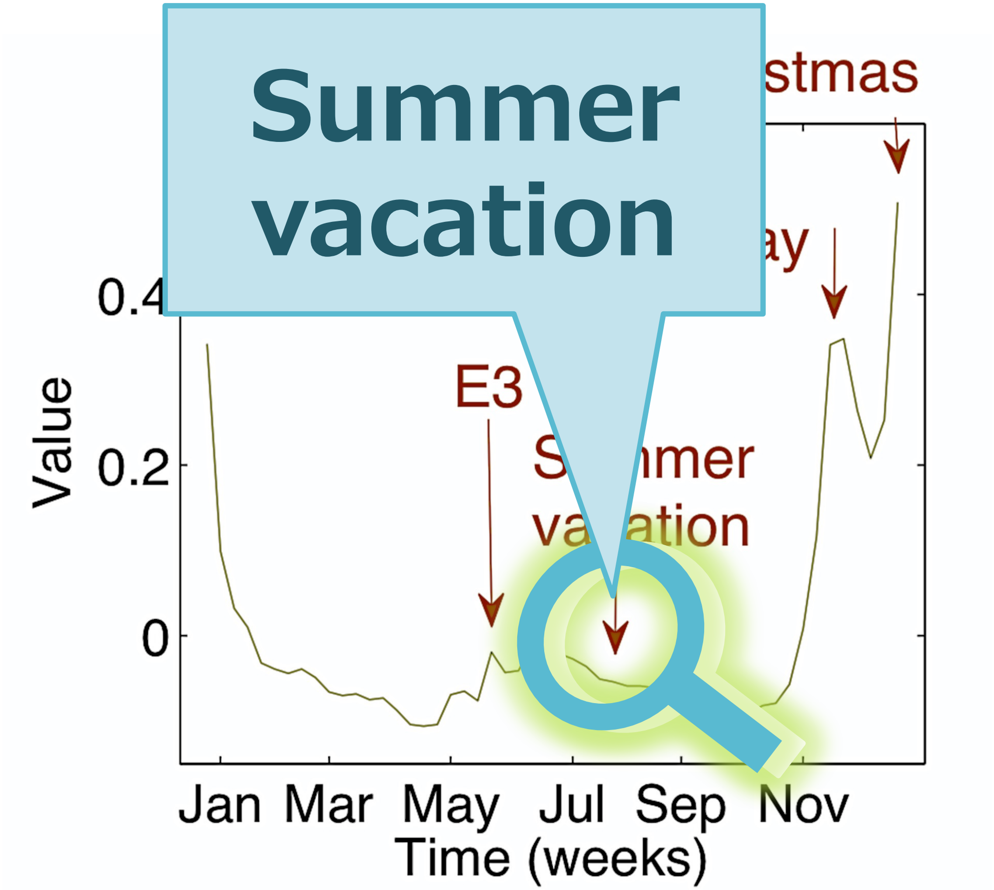
CMU CS

E3: Electronic Entertainment Expo (June)



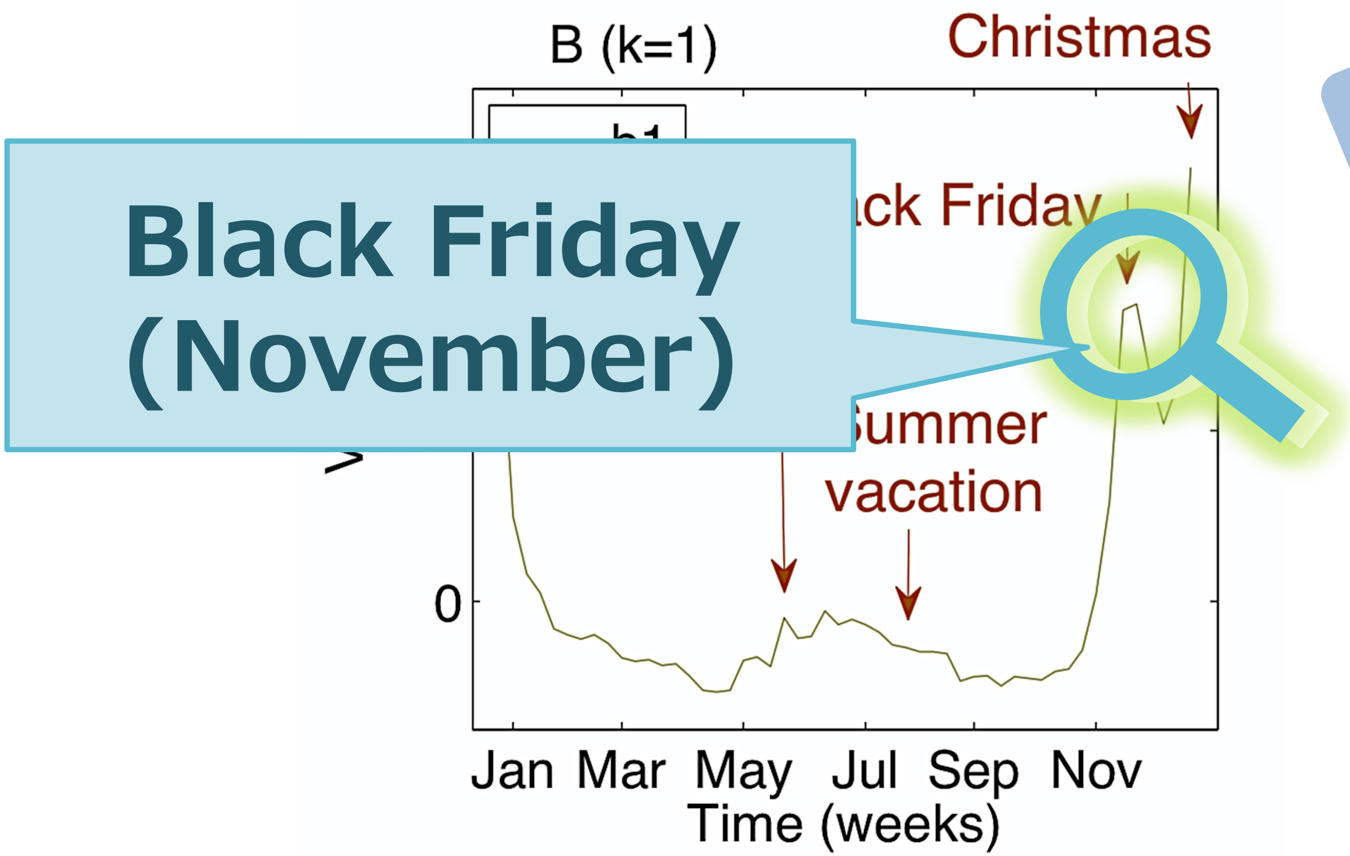
EcoWeb: seasonal component

Modeling power of EcoWeb



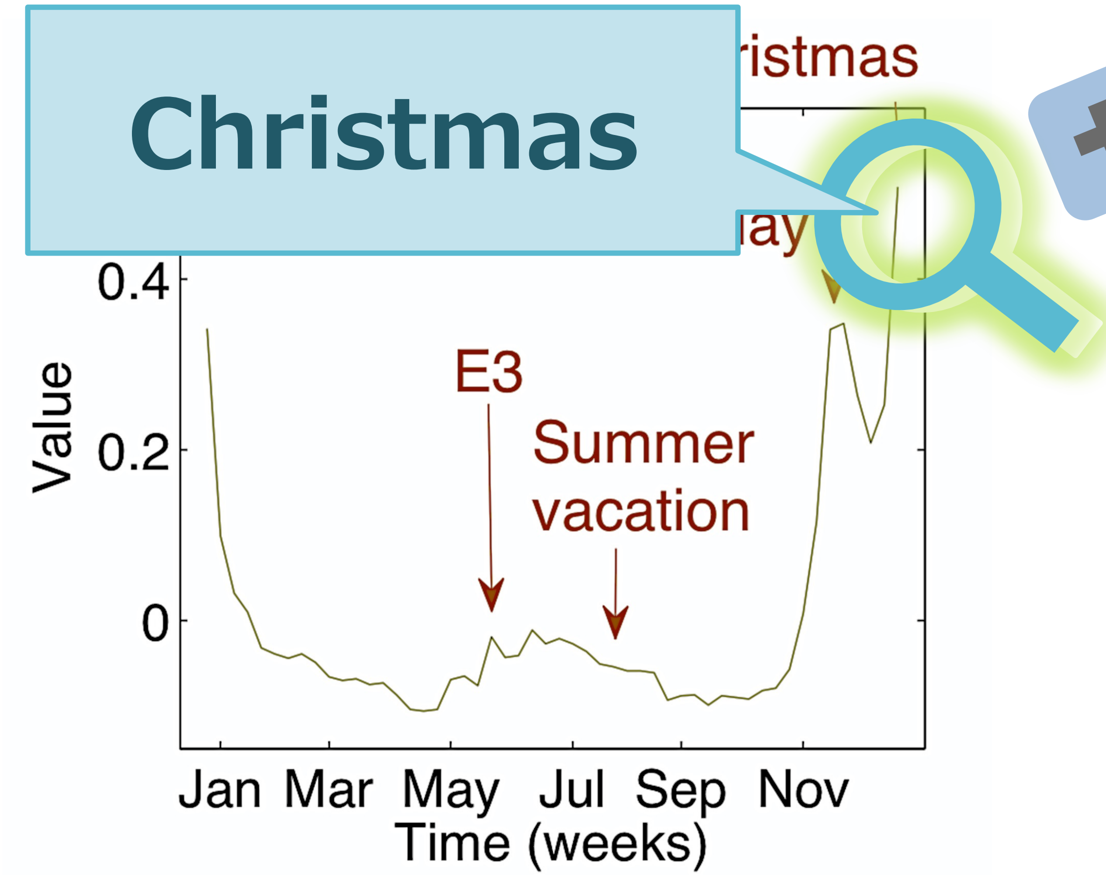
EcoWeb: seasonal component

Modeling power of EcoWeb



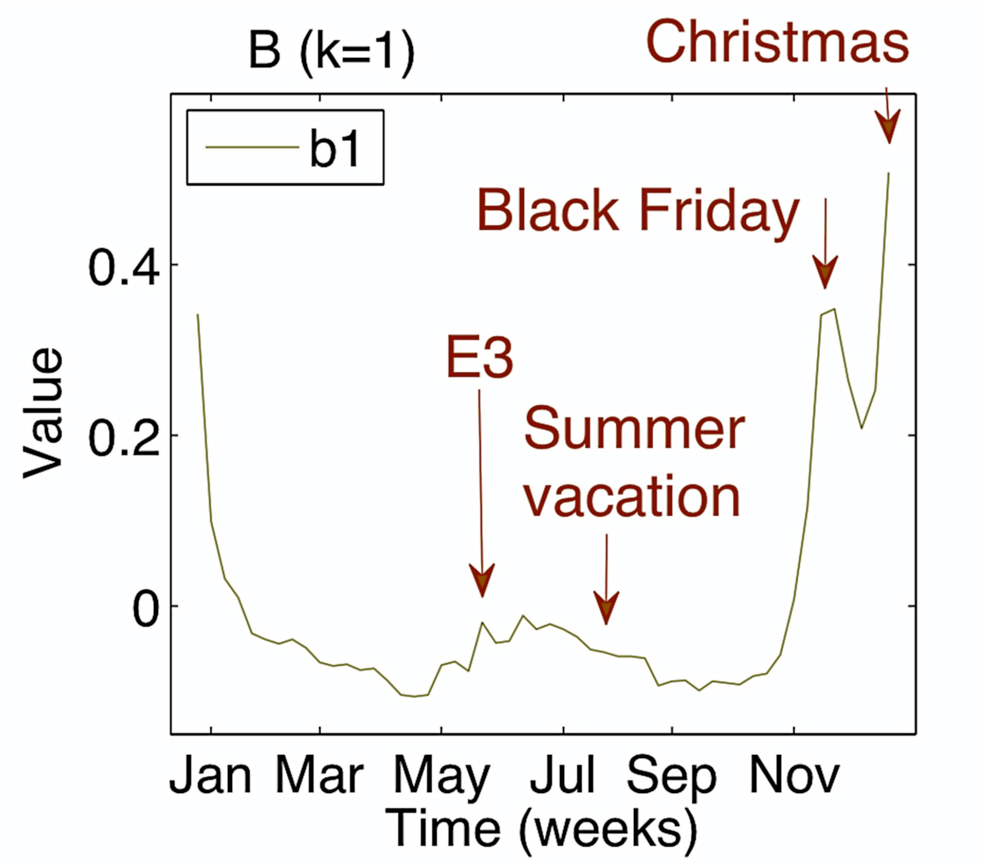
EcoWeb: seasonal component

Modeling power of EcoWeb



EcoWeb: seasonal component

Modeling power of EcoWeb



EcoWeb: seasonal component

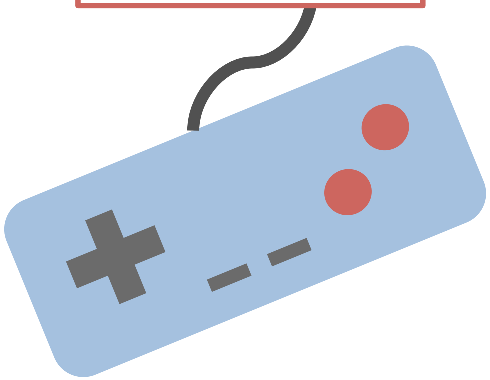
Kumamoto U

Modeling power of EcoWeb

CMU CS

Questions

Q1



Q2

Q3

amazon

Kumamoto U

Modeling power of EcoWeb

CMU CS

Q2 (apparels)

Who is the competitor?

JCPenny

vs.

KOHL'S

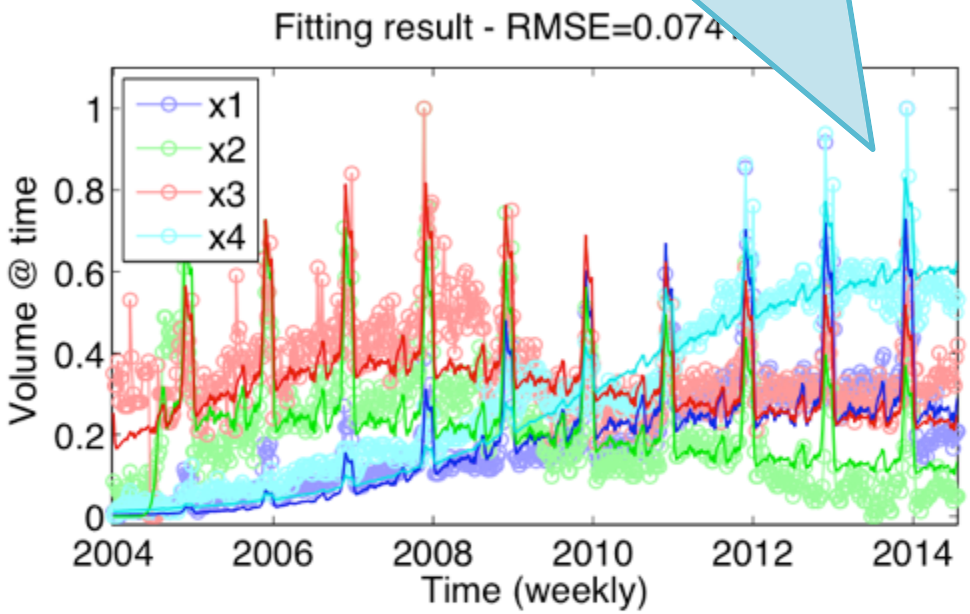
FOREVER 21

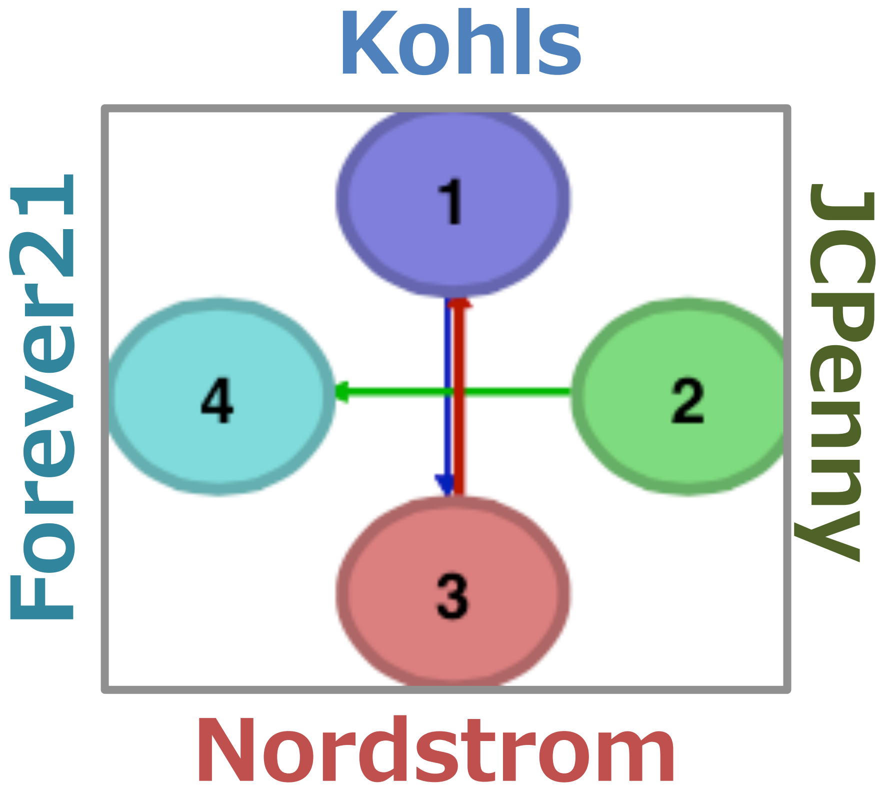
?

NORDSTROM

Modeling power of EcoWeb

A2. Forever21!



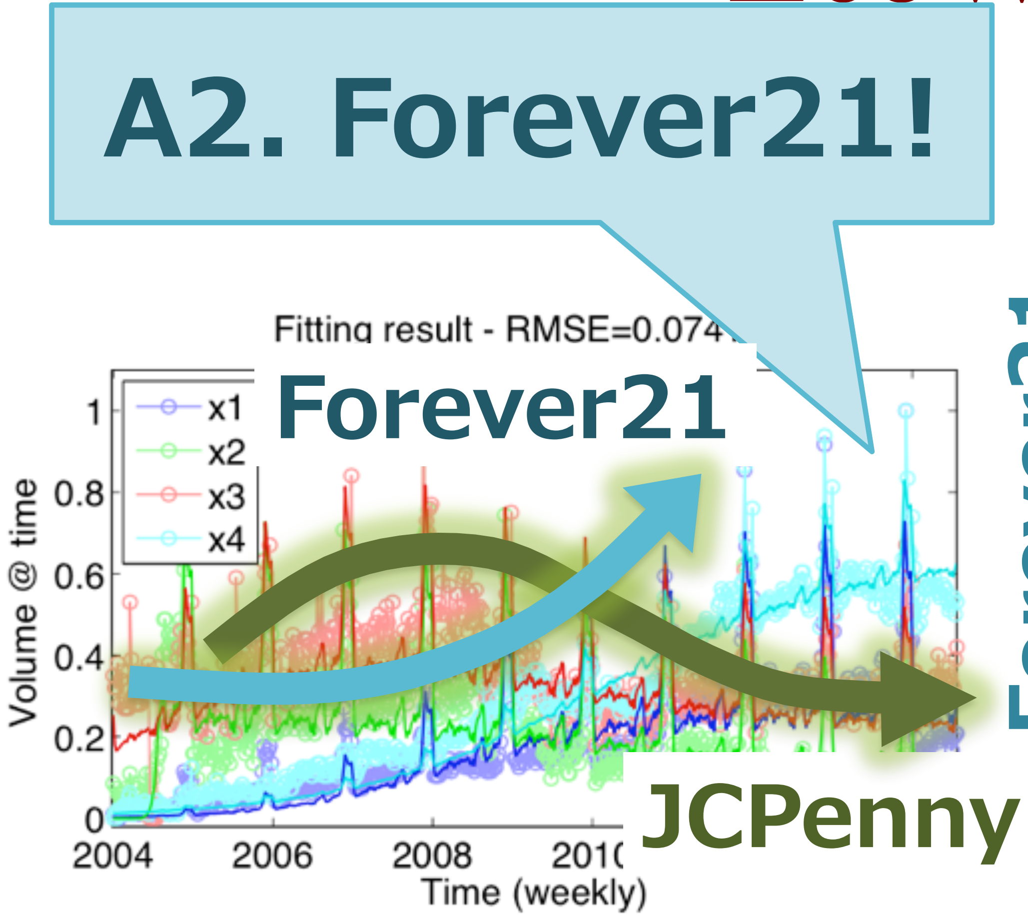


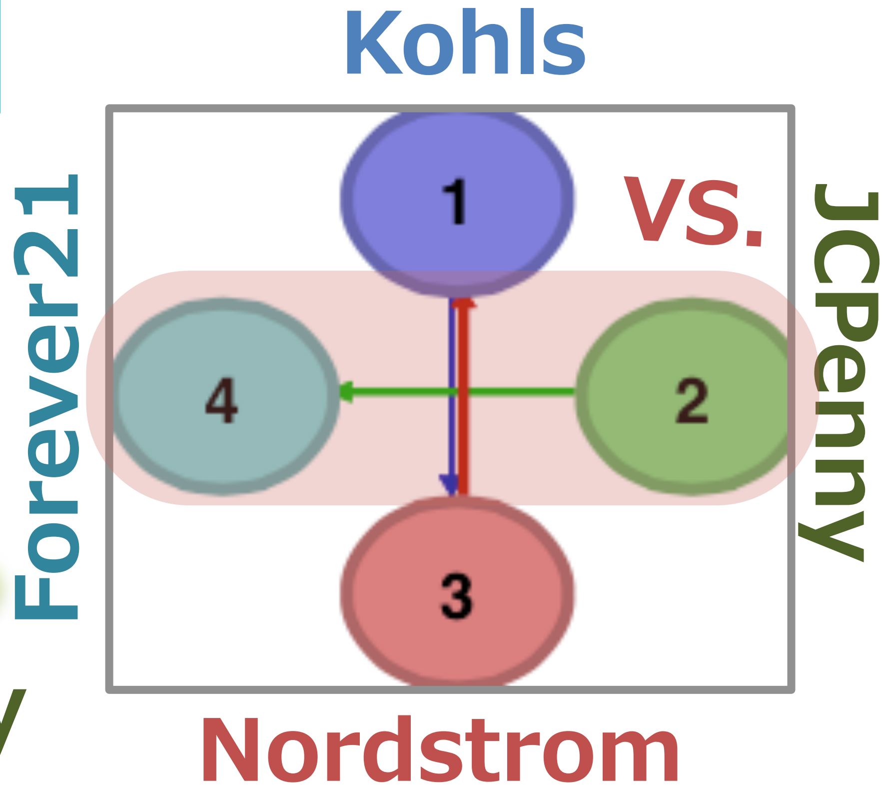
EcoWeb: Interaction network

Kumamoto U

Modeling power of EcoWeb

CMU CS





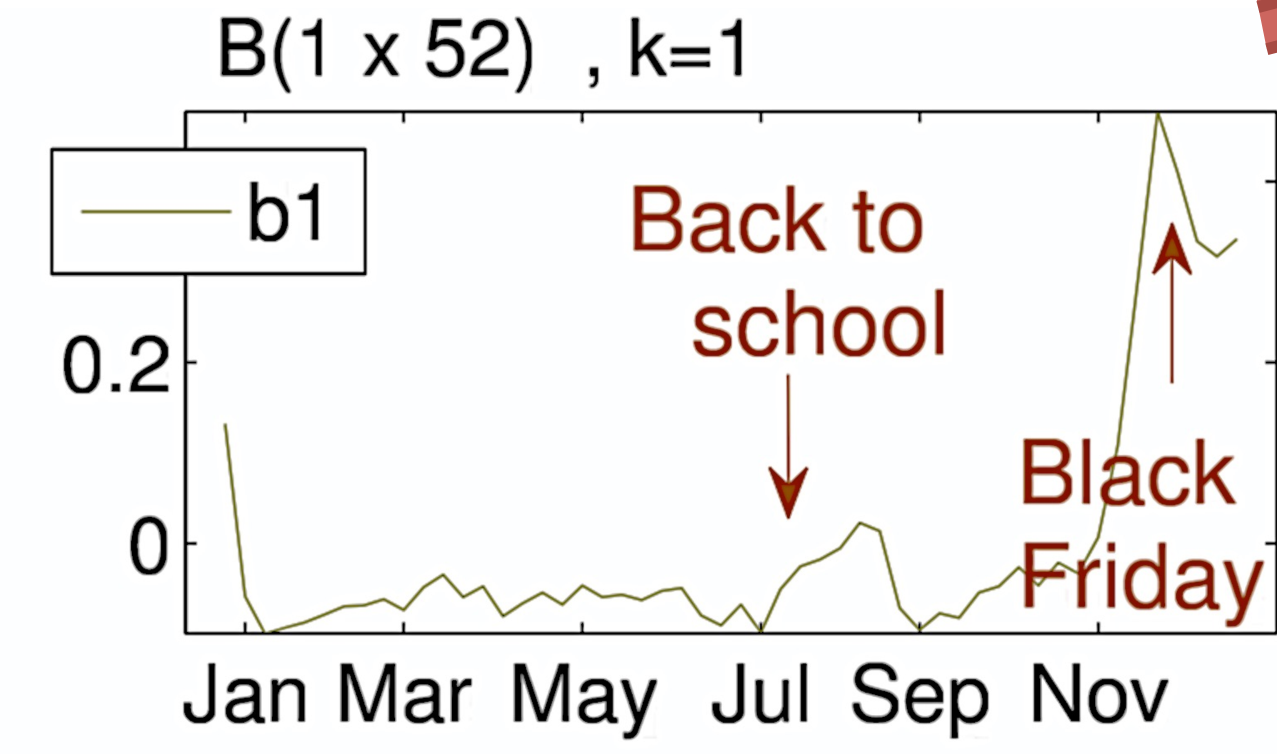
EcoWeb: Interaction network

Modeling power of EcoWeb

Q2 (apparels)

Any seasonal events?

Modeling power of EcoWeb



EcoWeb: seasonal component

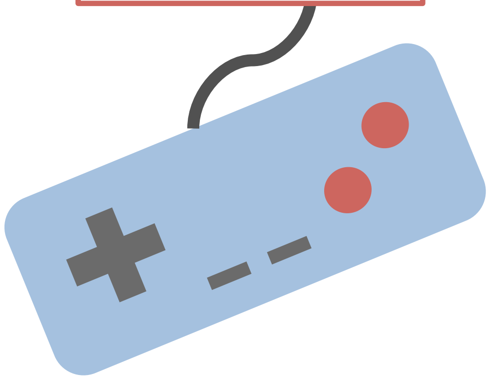
Kumamoto U

Modeling power of EcoWeb

CMU CS

Questions

Q1



Q2

Q3

amazon

Kumamoto U

Modeling power of EcoWeb

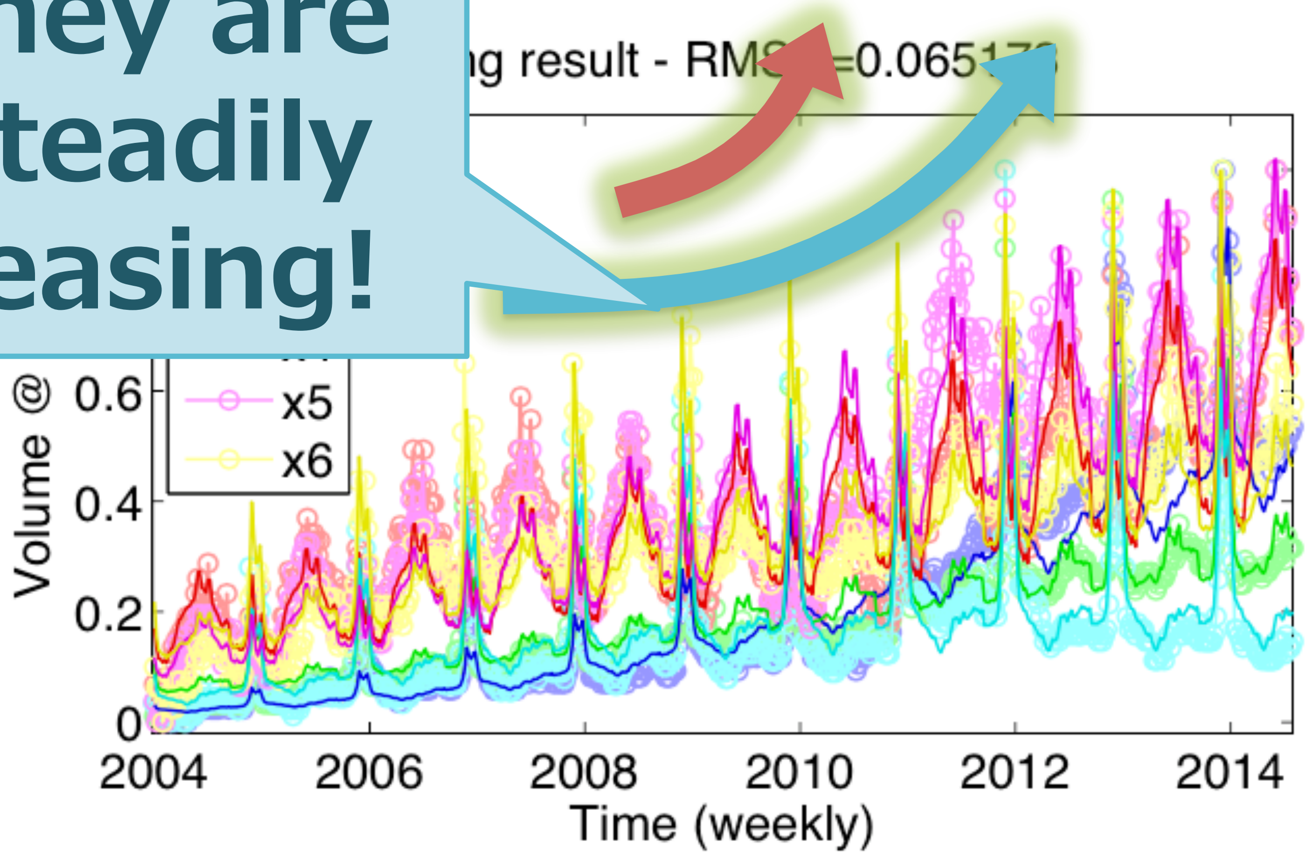
CMU CS

Q3 (retails)

Any patterns/trends?

Modeling power of EcoWeb

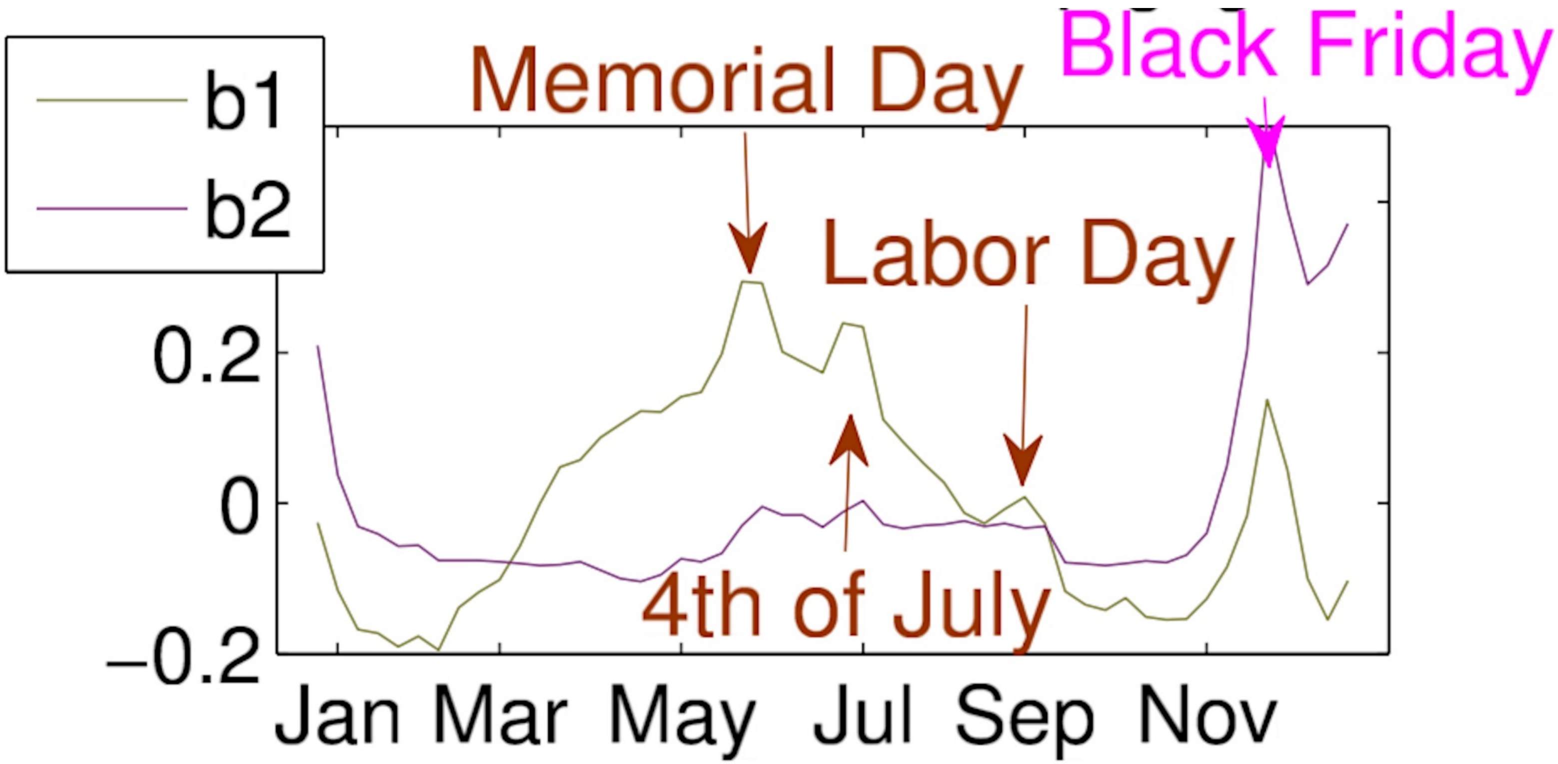
A. They are all steadily increasing!



Amazon, Walmart, Home Depot, Best buy, …

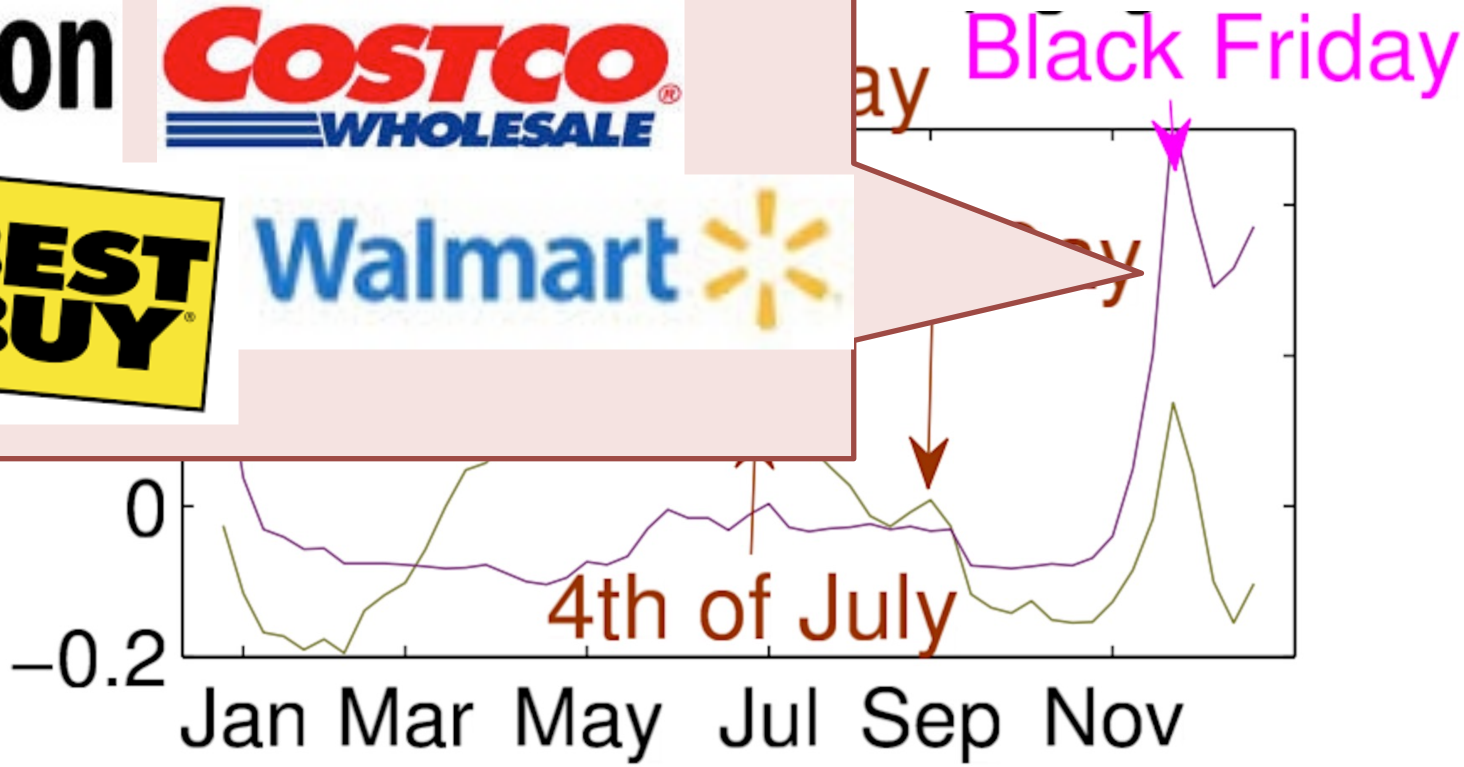
Modeling power of EcoWeb

2 seasonal components



Modeling power of EcoWeb

Black Friday sale

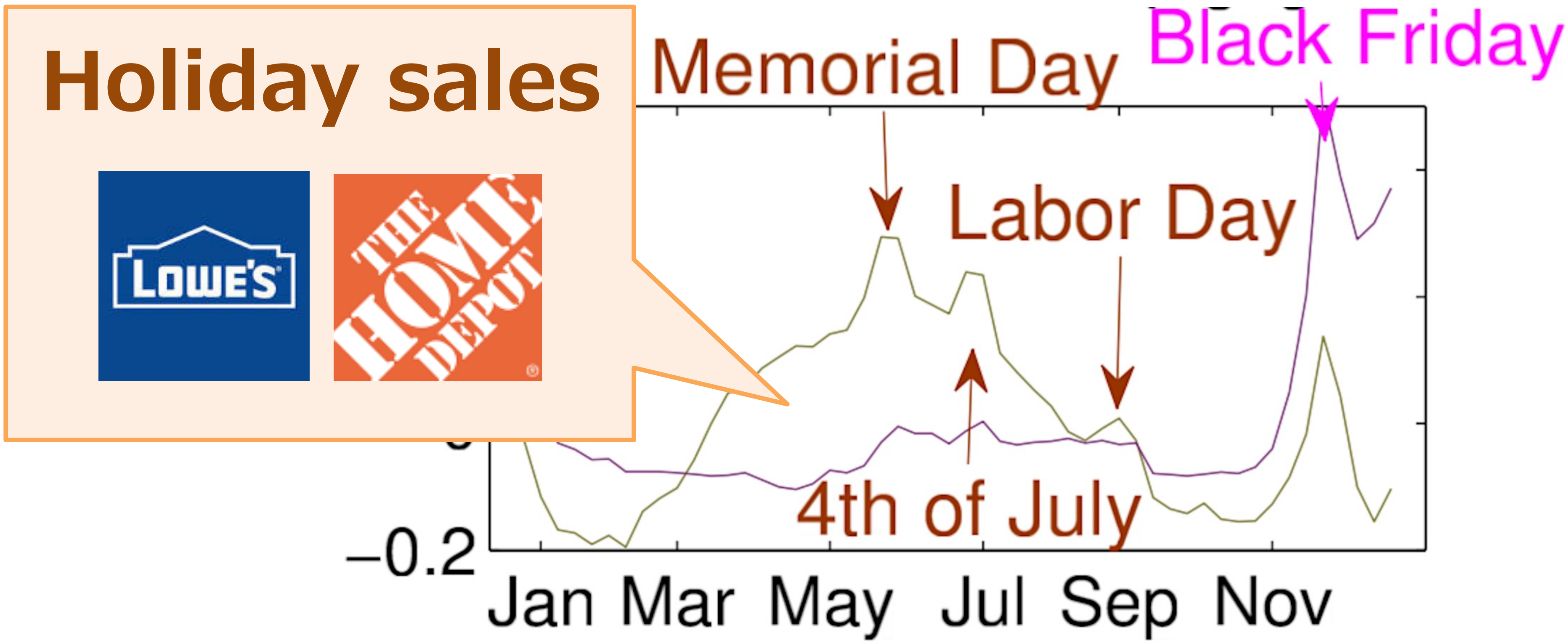


Kumamoto U

CMU CS

Modeling power of EcoWeb

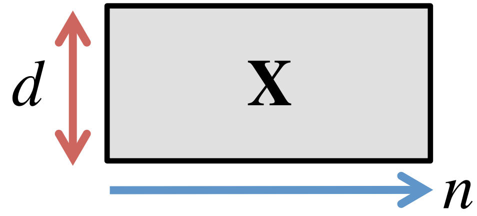
2 seasonal components



Problem definition

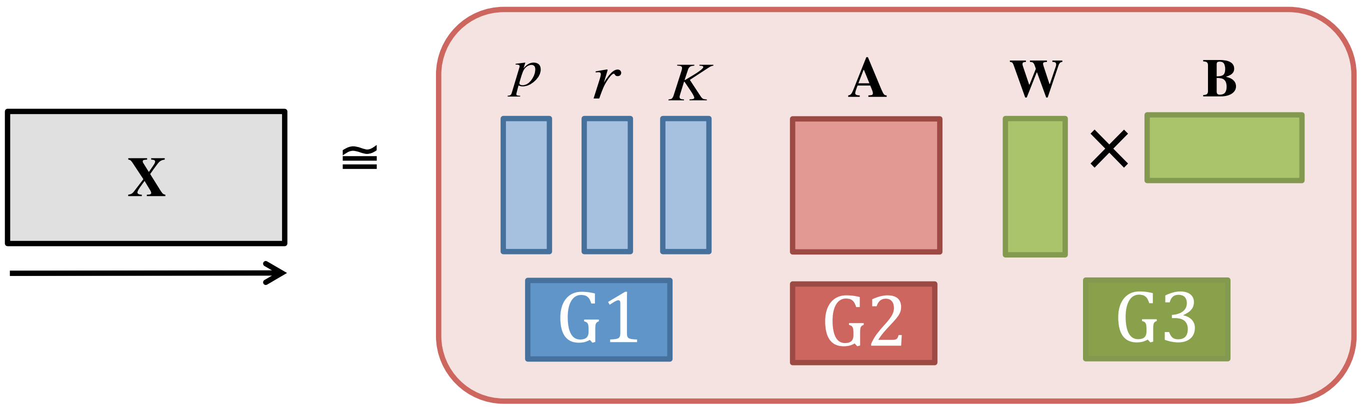
Given: Co-evolving online activities

X (activity x time)



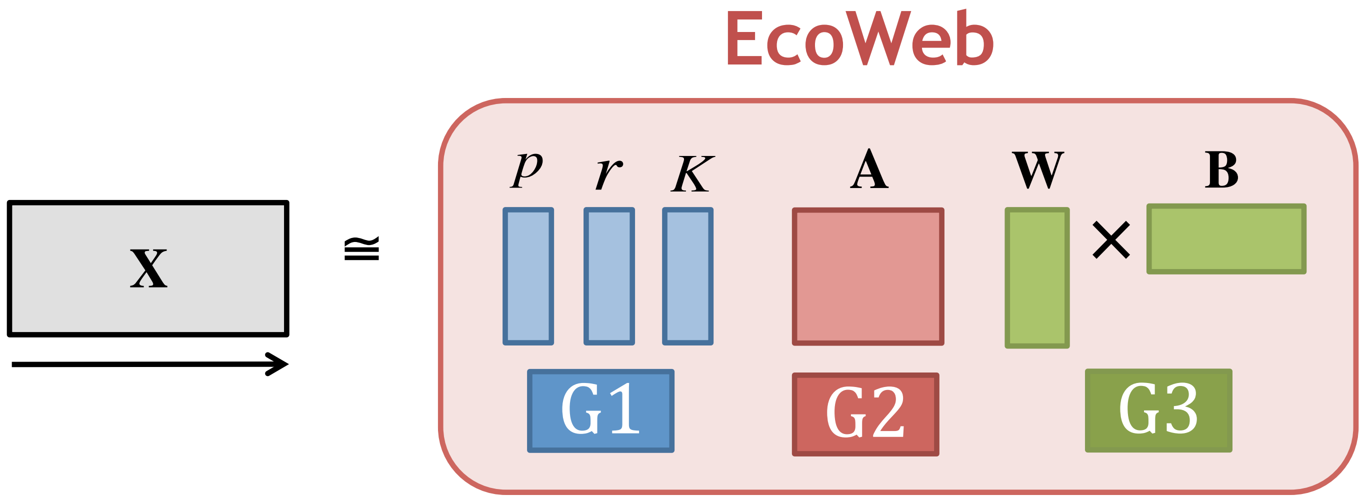
Find: Compact description of X

EcoWeb



EcoWeb: Main idea

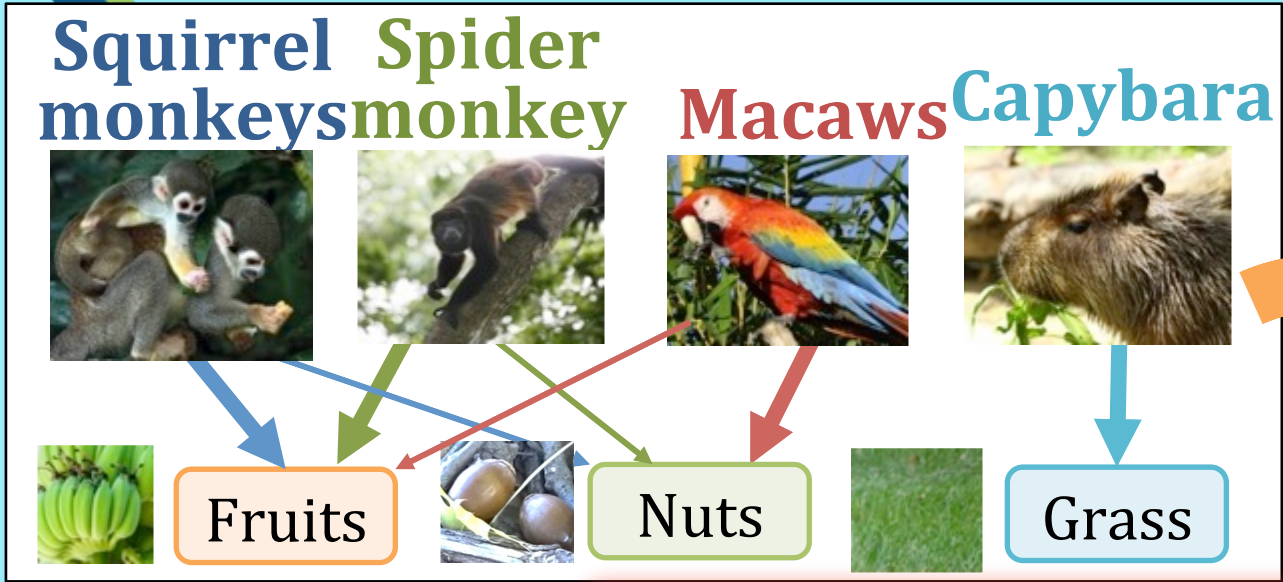
Q. How can we describe the evolutions of X ?



A. The Web as a jungle!

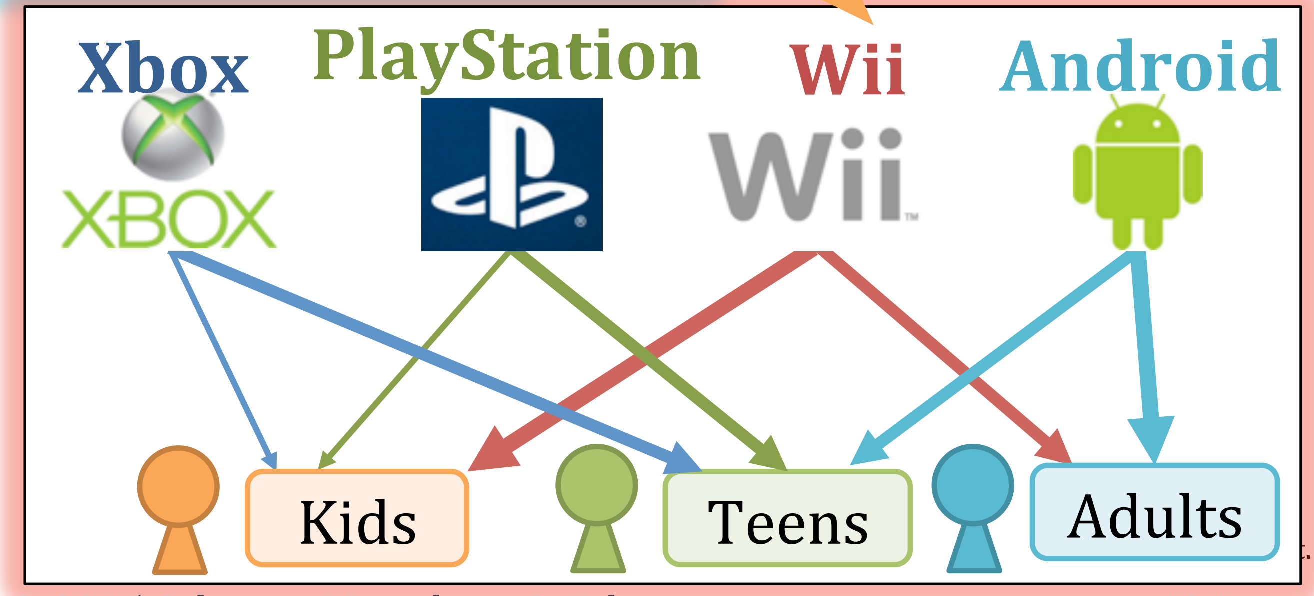
- "Virtual species" living on the Web
- Interacting with other species (activities)

The Web as a jungle



Ecosystem in the **Jungle**

Ecosystem on the **Web**



Ecosystem on the Web

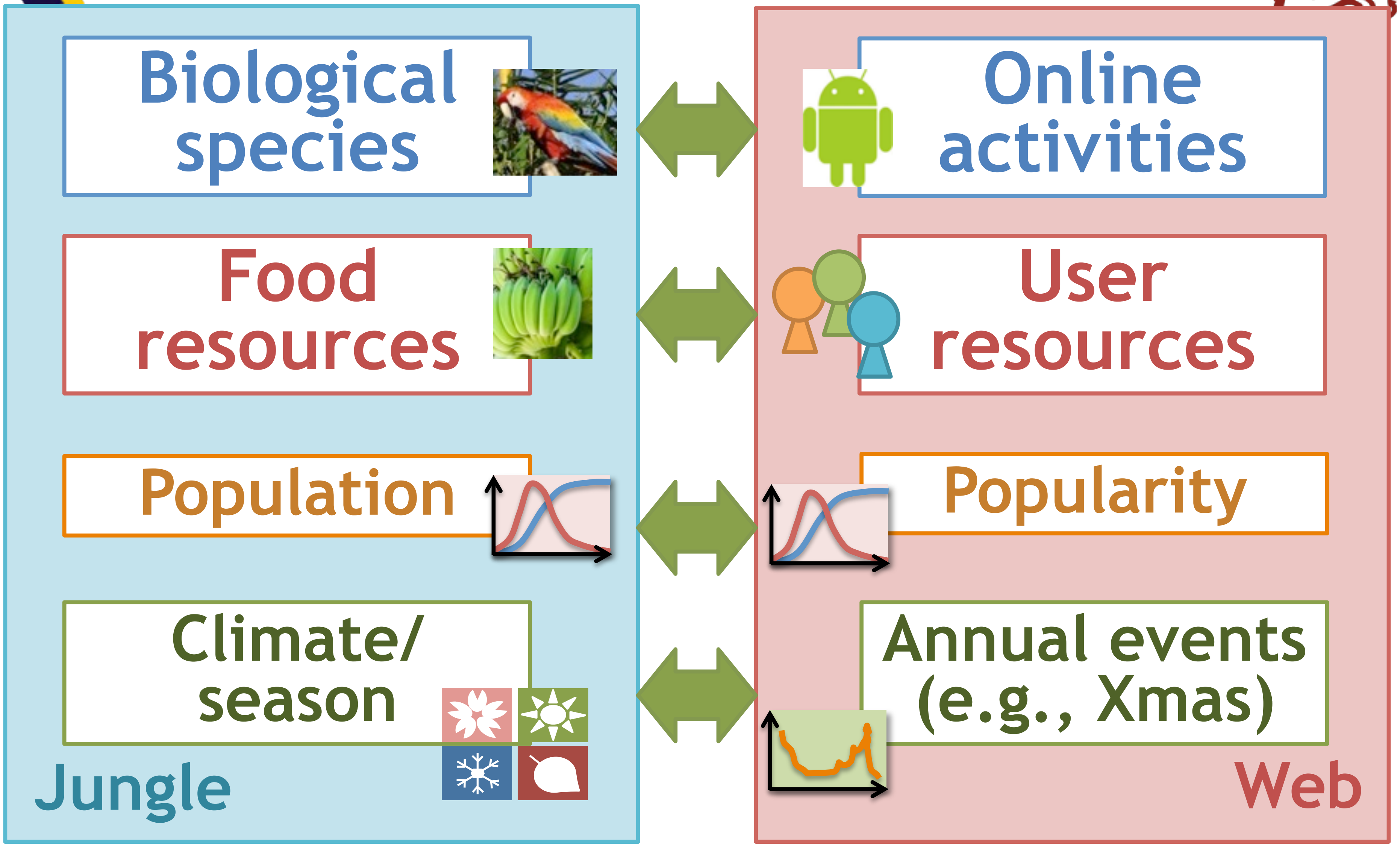
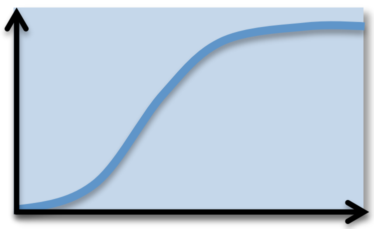


Image courtesy of xura, criminalatt, David Castillo Dominici, happykanppy at FreeDigitalPhotos.net.

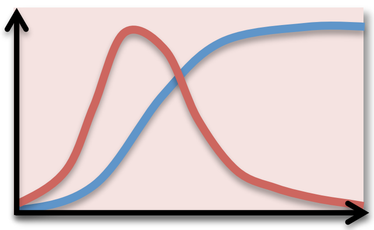
EcoWeb: Main idea

Q. How can we describe the evolutions of X ?

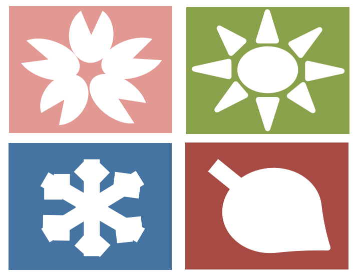
Non-linear evolution



Interaction/ competition



Seasonality

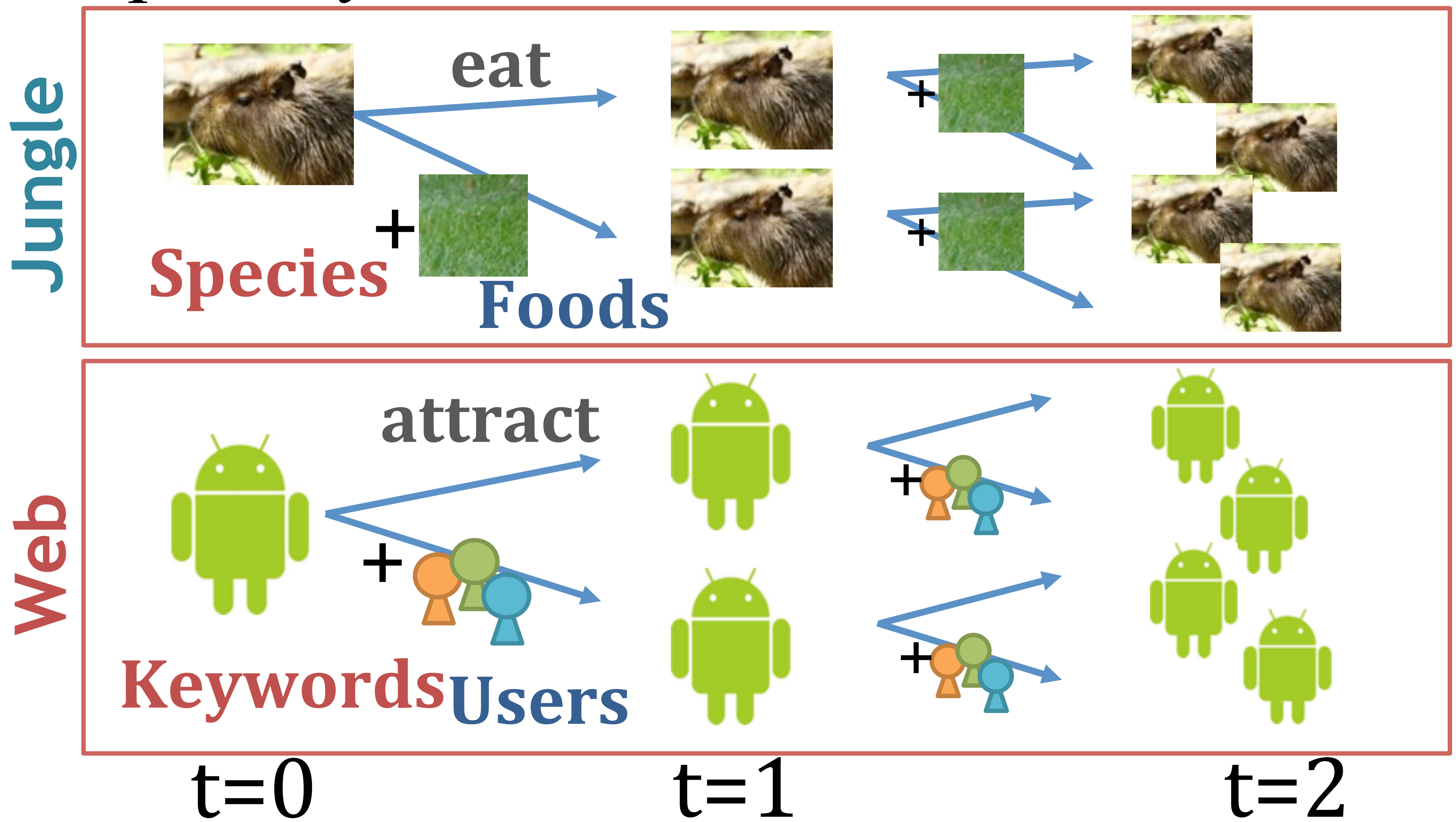


A. Web as a jungle!

G1 G2 G3

G1: EcoWeb-individual

Popularity size increases over time

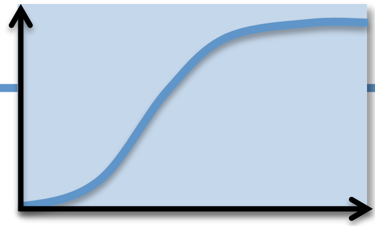


Details

G1: EcoWeb-individual

CMU CS

Non-linear evolution of a single keyword



Popularity size

$$P(t+1) = P(t)\left[1 + r\left(1 - \frac{P(t)}{K}\right)\right],$$

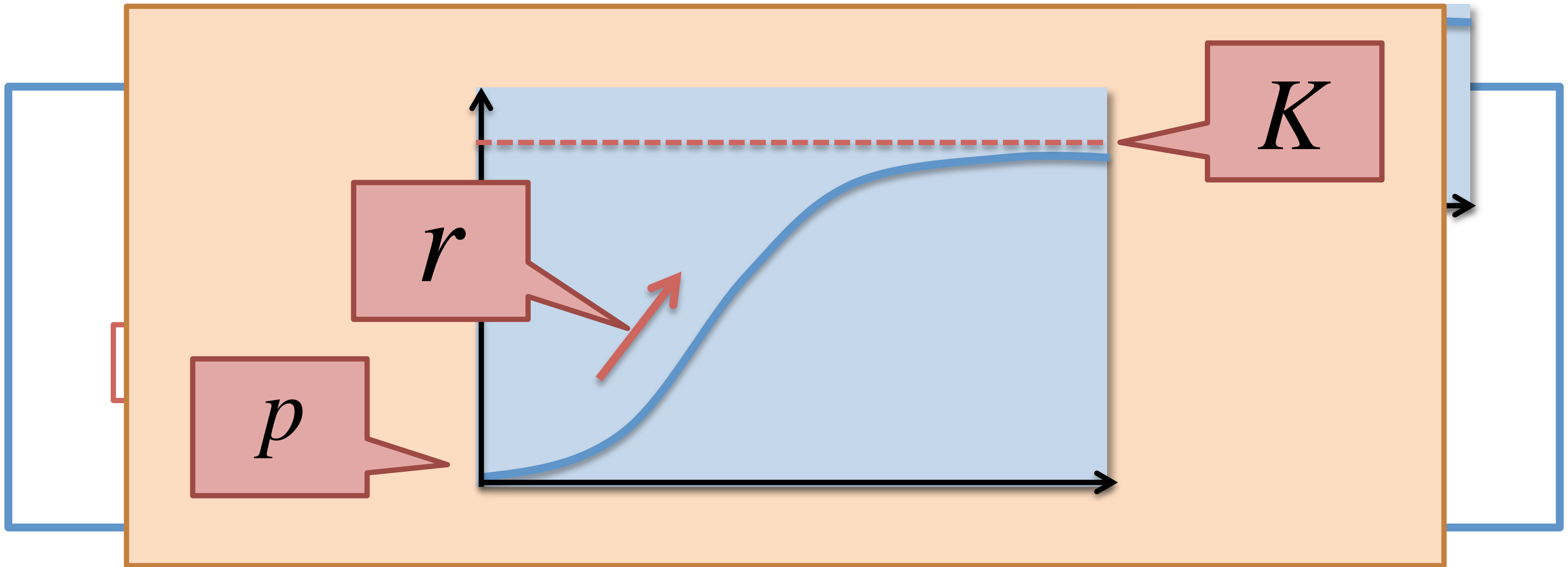
- p – Initial condition (i.e., P(0) =p)
- r – Growth rate, attractiveness
- K – Carrying capacity (=available user resources)

Details

G1: EcoWeb-individual

CMU CS

Non-linear evolution of a single keyword



- p – Initial condition (i.e., $P(0) = p$)
- r – Growth rate, attractiveness
- K – Carrying capacity (=available user resources)

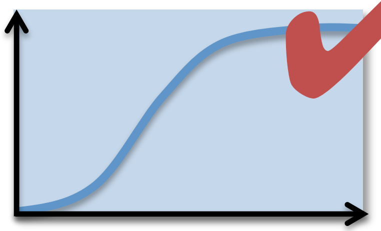
Kumamoto U

EcoWeb: Main idea

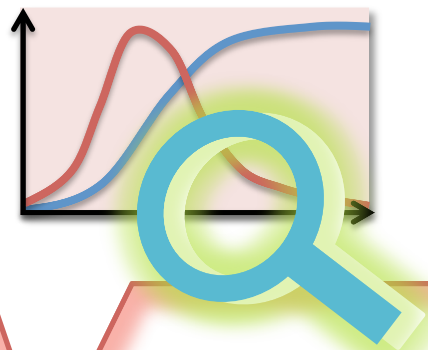
CMU CS

Q. How can we describe the evolutions of X ?

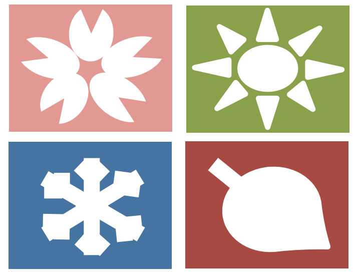
Non-linear evolution



Interaction/ competition



Seasonality

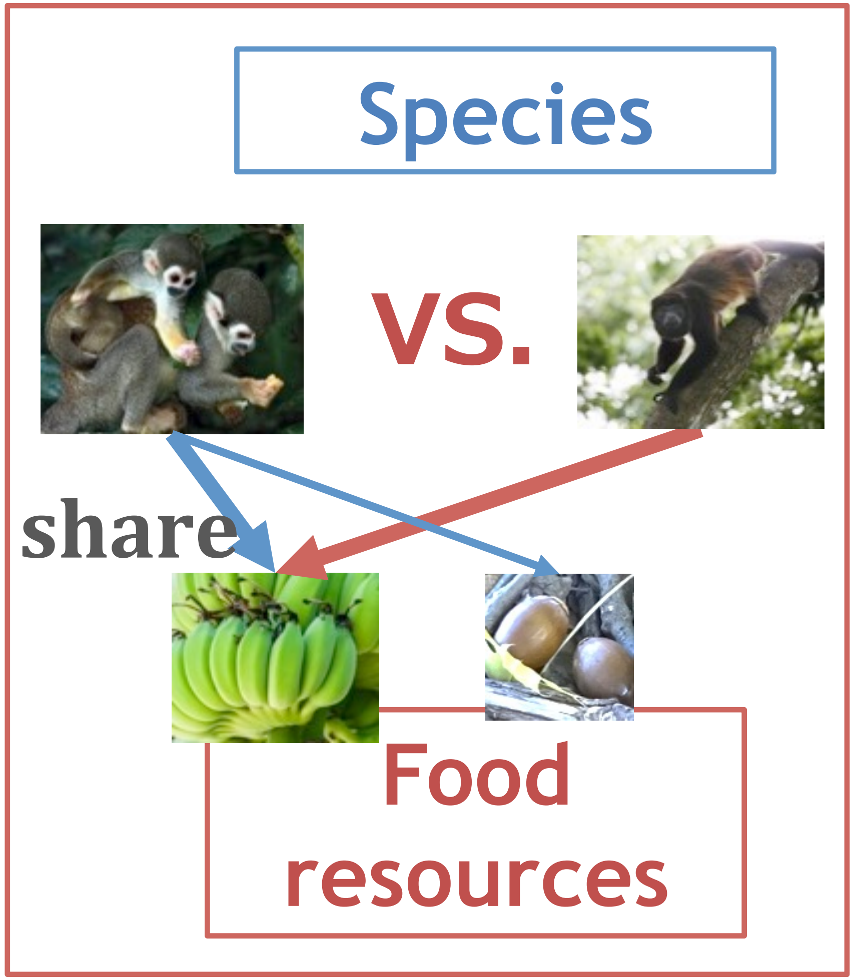


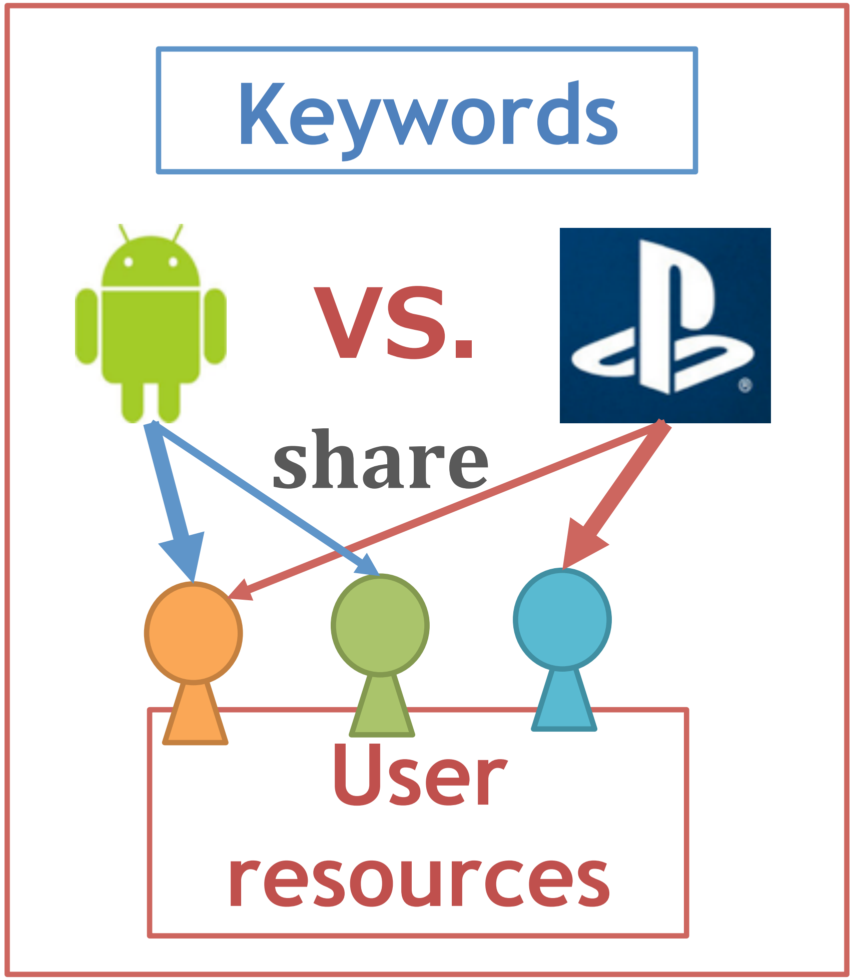
A. Web as a jungle!

G1 G2 G3

G2: EcoWeb-interaction

Interaction between multiple keywords





G2: EcoWeb-interaction

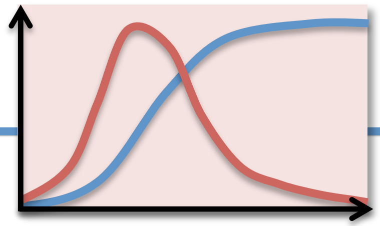
Details

CMU CS

Interaction between multiple keywords

Popularity of keyword i   Popularity of j

$$P_i(t+1) = P_i(t)\left[1 + r_i\left(1 - \frac{\sum_{j=1}^{d} a_{ij} P_j(t)}{K_i}\right)\right], \quad (i = 1, \cdots, d), \quad (3)$$



a_{ij} – Interaction coefficient

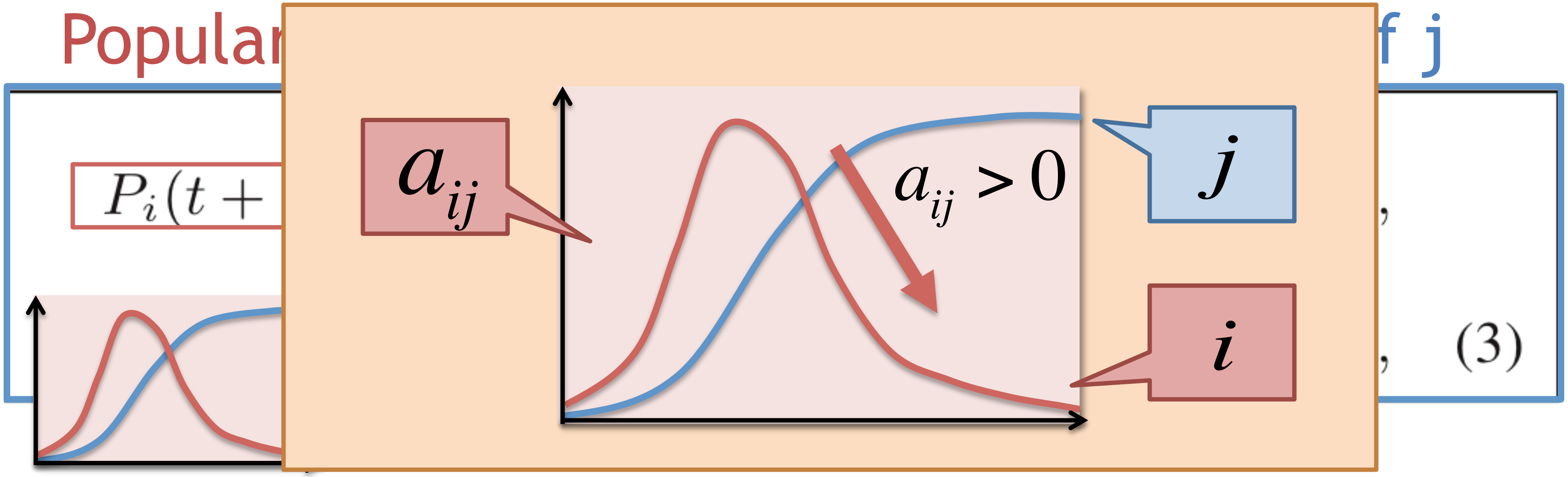
– i.e., effect rate of keyword j on i

G2: EcoWeb-interaction

Details

Interaction between multiple keywords

Popular ... f j

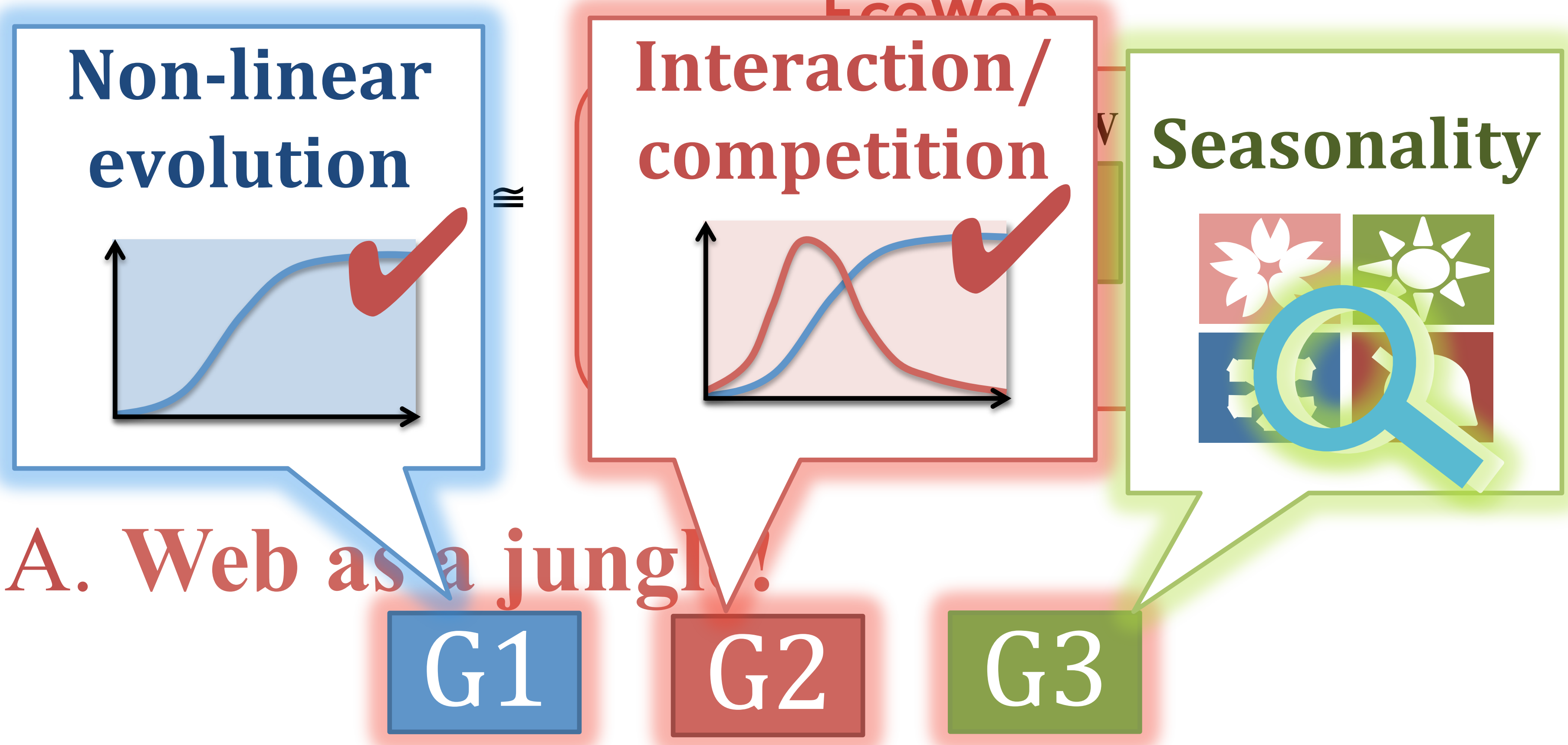


a_{ij} – Interaction coefficient

– i.e., effect rate of keyword j on i

EcoWeb: Main idea

Q. How can we describe the evolutions of X ?



A. Web as a jungle!

G3: EcoWeb-seasonality

"Hidden" seasonal activities

Season/ Climate

Seasonal events

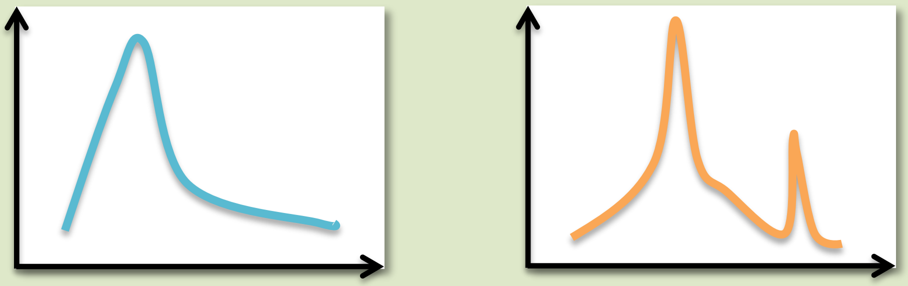
Kumamoto U

CMU CS

G3: EcoWeb-seasonality

“Hidden” seasonal activities

Users change their behavior according to seasonal events!



Climate

events

Details

G3: EcoWeb-seasonality

CMU CS

"Hidden" seasonal activities

Estimated volume of keyword i

$$C_i(t) = P_i(t)\,[1 + e_i(t)] \qquad (i = 1, \cdots, d),$$

$$e_i(t) \simeq f(i, t|\mathbf{W}, \mathbf{B}) = \sum_{j=1}^{k} w_{ij} b_j(\tau) \qquad (\tau = [t \mod n_p])$$

Seasonal activities of i

W – Participation (weight) matrix

B – Seasonality matrix

Details

G3: EcoWeb-seasonality

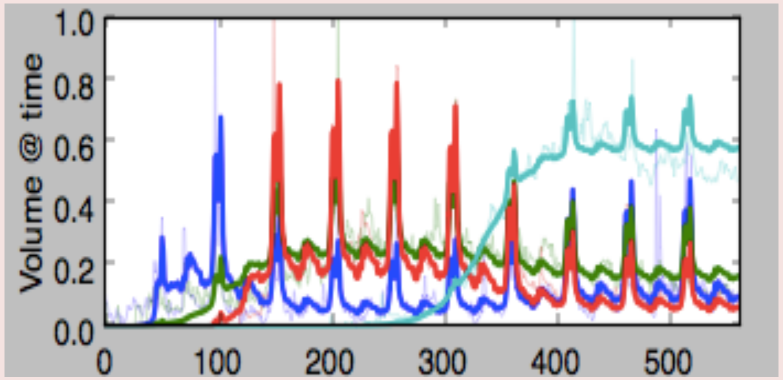
CMU CS

"Hidden" seasonal activities

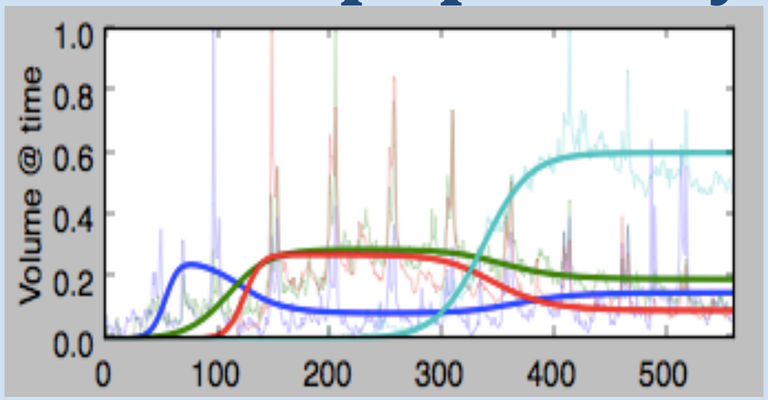
Estimated volume of keyword i

$$C_i(t) = P_i(t)\,[1 + e_i(t)] \qquad (i = 1, \cdots, d),$$

C: volume



P: latent popularity



Details

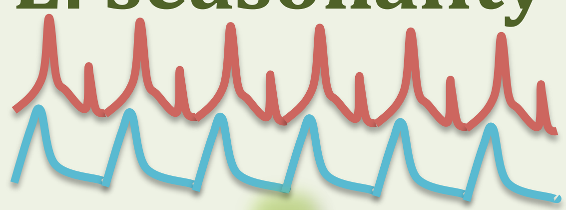
G3: EcoWeb-seasonality

"Hidden" seasonal activities

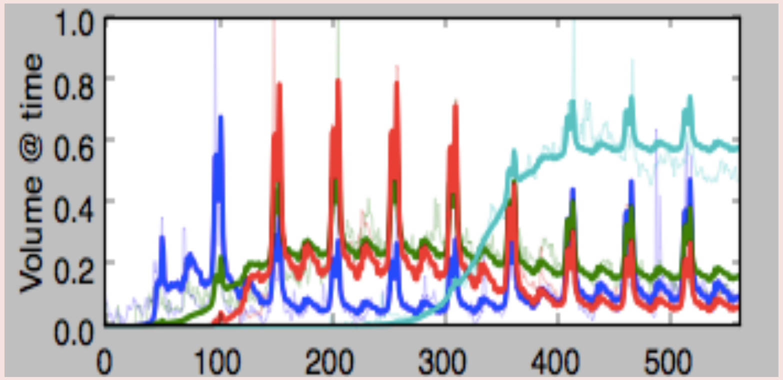
Estimated volume of keyword i

$$C_i(t) = P_i(t)\,[1 + e_i(t)]$$

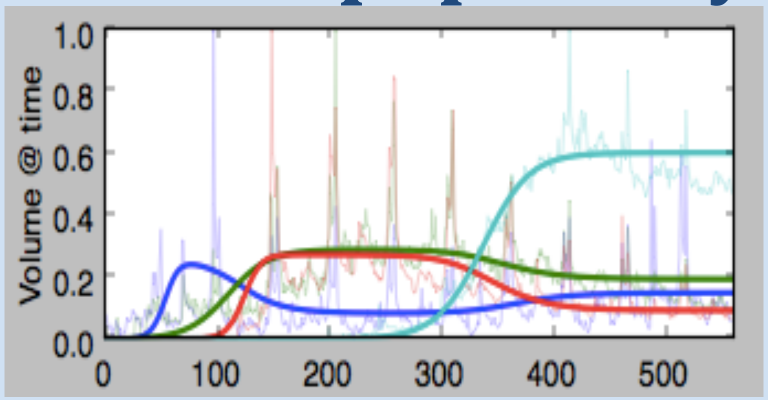
E: seasonality



C: volume



P: latent popularity



Details

CMU CS

G3: EcoWeb-seasonality

"Hidden" seasonal activities

Estimated volume of keyword i

$$C_i(t) = P_i(t)\,[1 + e_i(t)] \qquad (i = 1, \cdots, d),$$

$$e_i(t) \simeq f(i, t|\mathbf{W}, \mathbf{B}) = \sum_{j=1}^{k} w_{ij} b_j(\tau) \qquad (\tau = [t \mod n_p])$$

Seasonal activities of keyword i

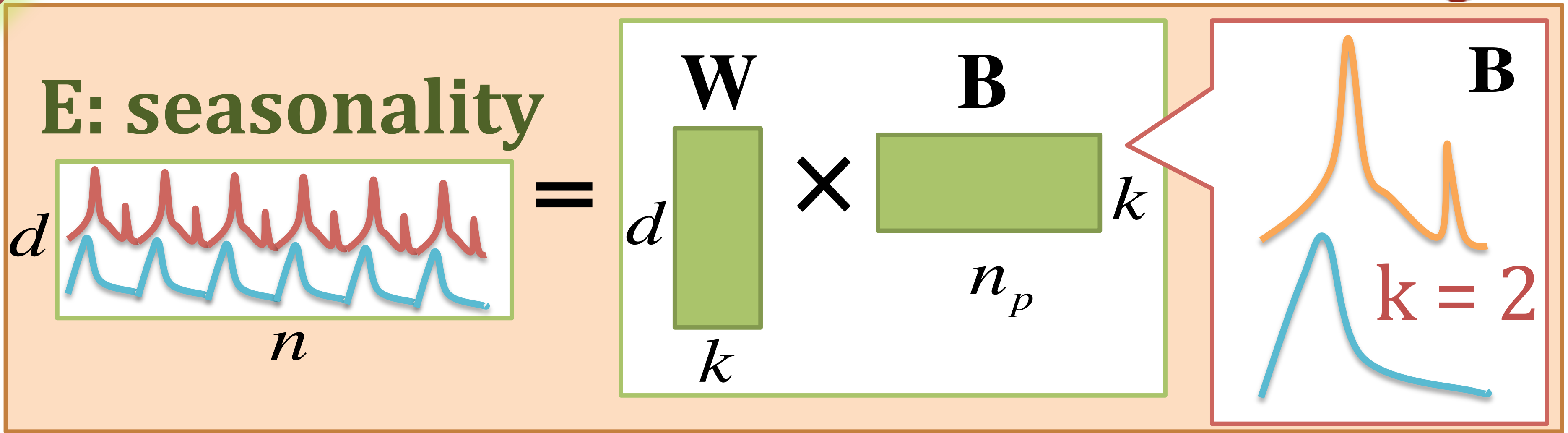
W – Participation (weight) matrix

B – Seasonality matrix

Details

G3: EcoWeb-seasonality

CMU CS



$$e_i(t) \simeq f(i,t|\mathbf{W},\mathbf{B}) = \sum_{j=1} w_{ij} b_j(\tau) \quad (\tau = [t \mod n_p])$$

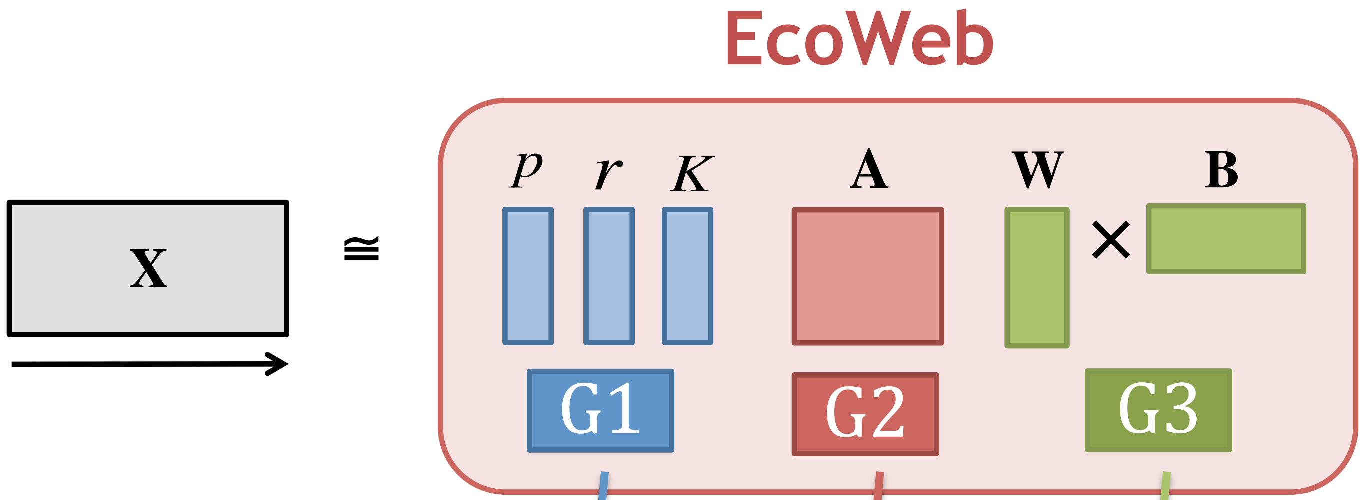
Seasonal activities of keyword i

W – Participation (weight) matrix

B – Seasonality matrix

EcoWeb: Main idea

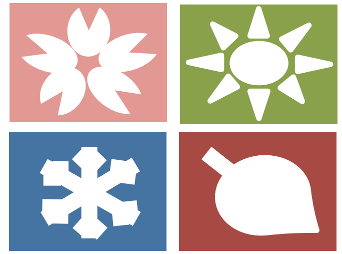
Q. How can we describe the evolutions of X ?



Full parameters

$$S = \{p, r, K, \mathbf{A}, \mathbf{W}, \mathbf{B}\}$$

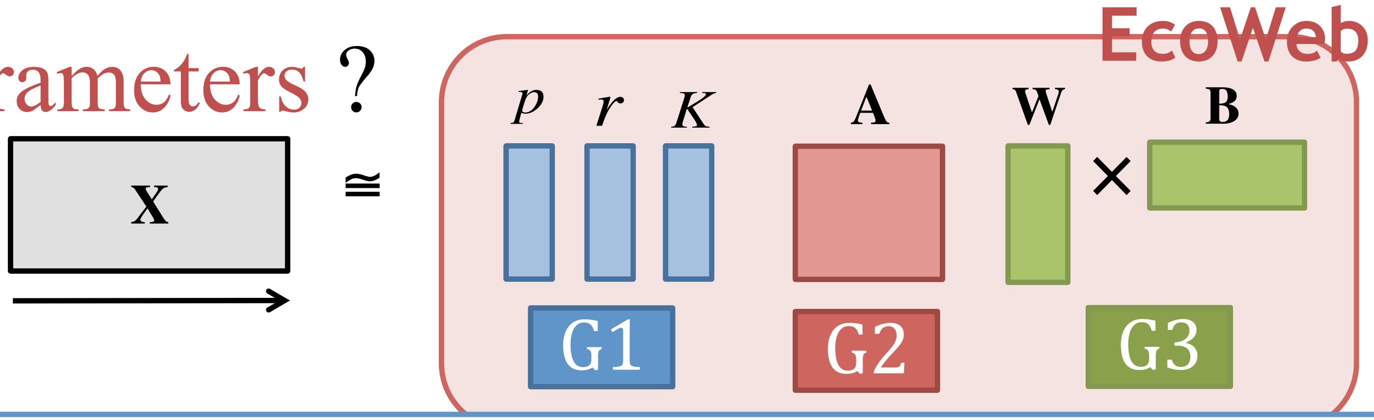
Algorithms



Q1. How can we automatically find "seasonal components" ?

Idea (1) : Seasonal component analysis

Q2. How can we efficiently estimate full-parameters ?



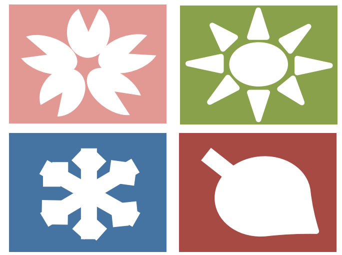
Idea (2): Multi-step fitting

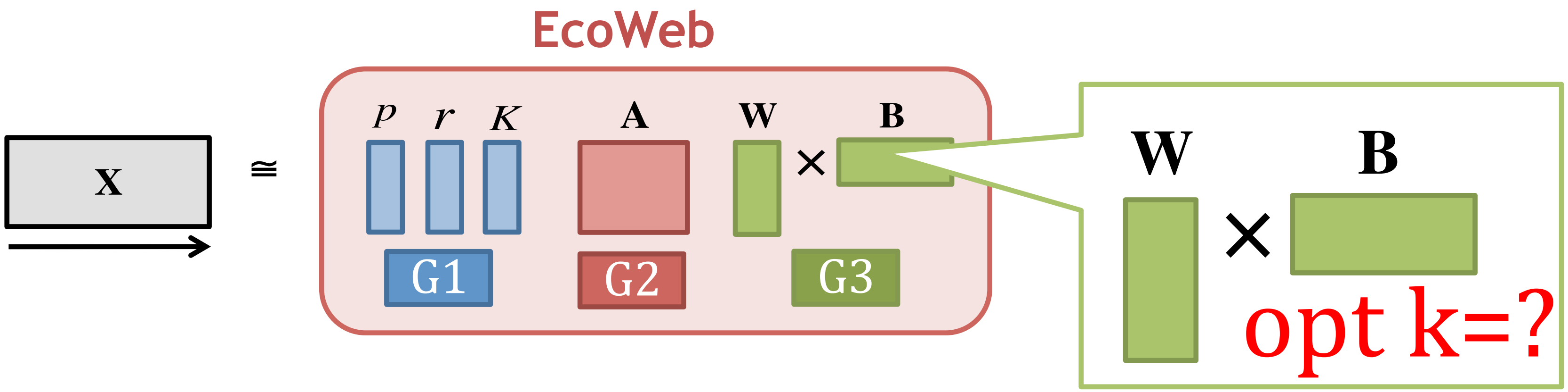
Kumamoto U

CMU CS

Idea (1): Seasonal component analysis

Q1. How can we automatically find "k-seasonal components" ?





Idea (1) :

a. Seasonal component detection
b. Automatic component analysis

Kumamoto U

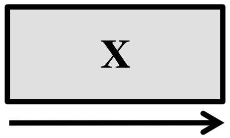
CMU CS

Idea (1): Seasonal component analysis

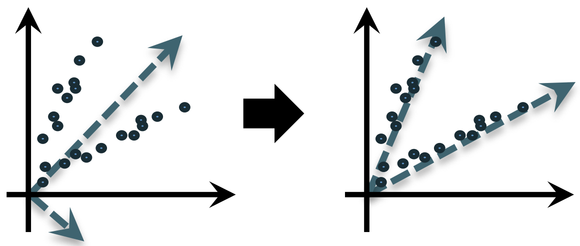
Q1. How can we automaticall...

find "... ...mponents" ?

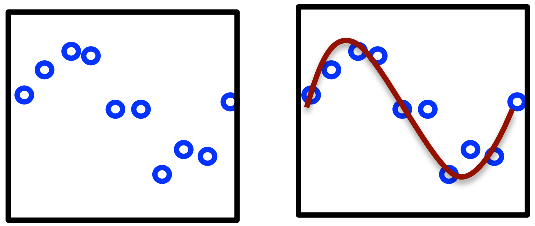
Details @ part1



ICA



MDL



Idea (1) :

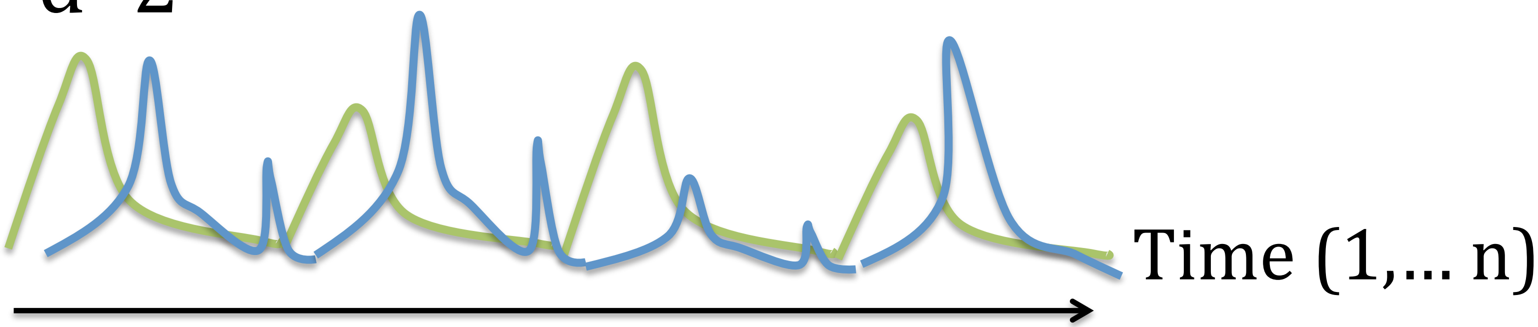
a. Seasonal component detection — **ICA**

b. Automatic component analysis — **MDL**

Idea (1): Seasonal component analysis

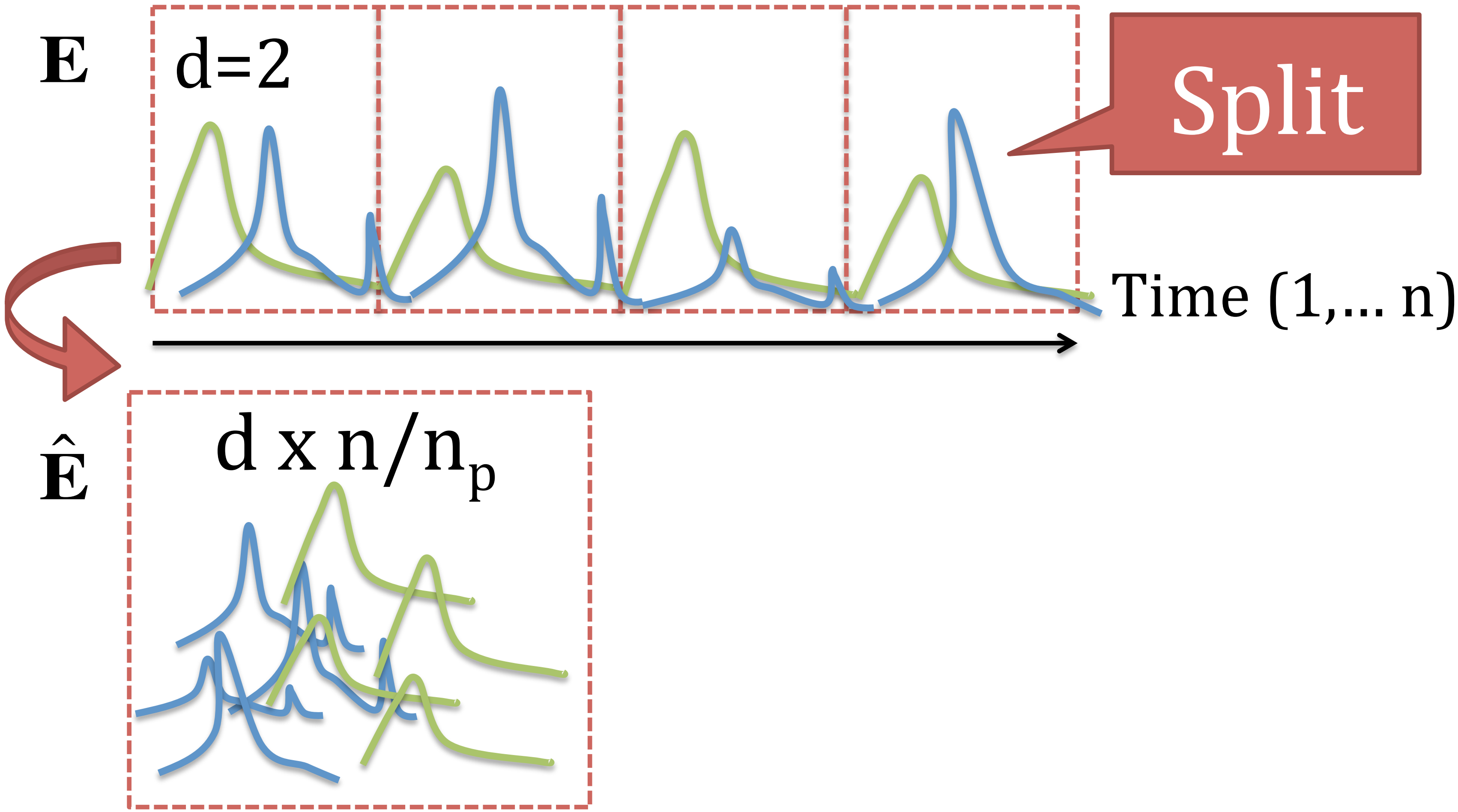
Idea(1-a)　Seasonal component detection

E　d=2



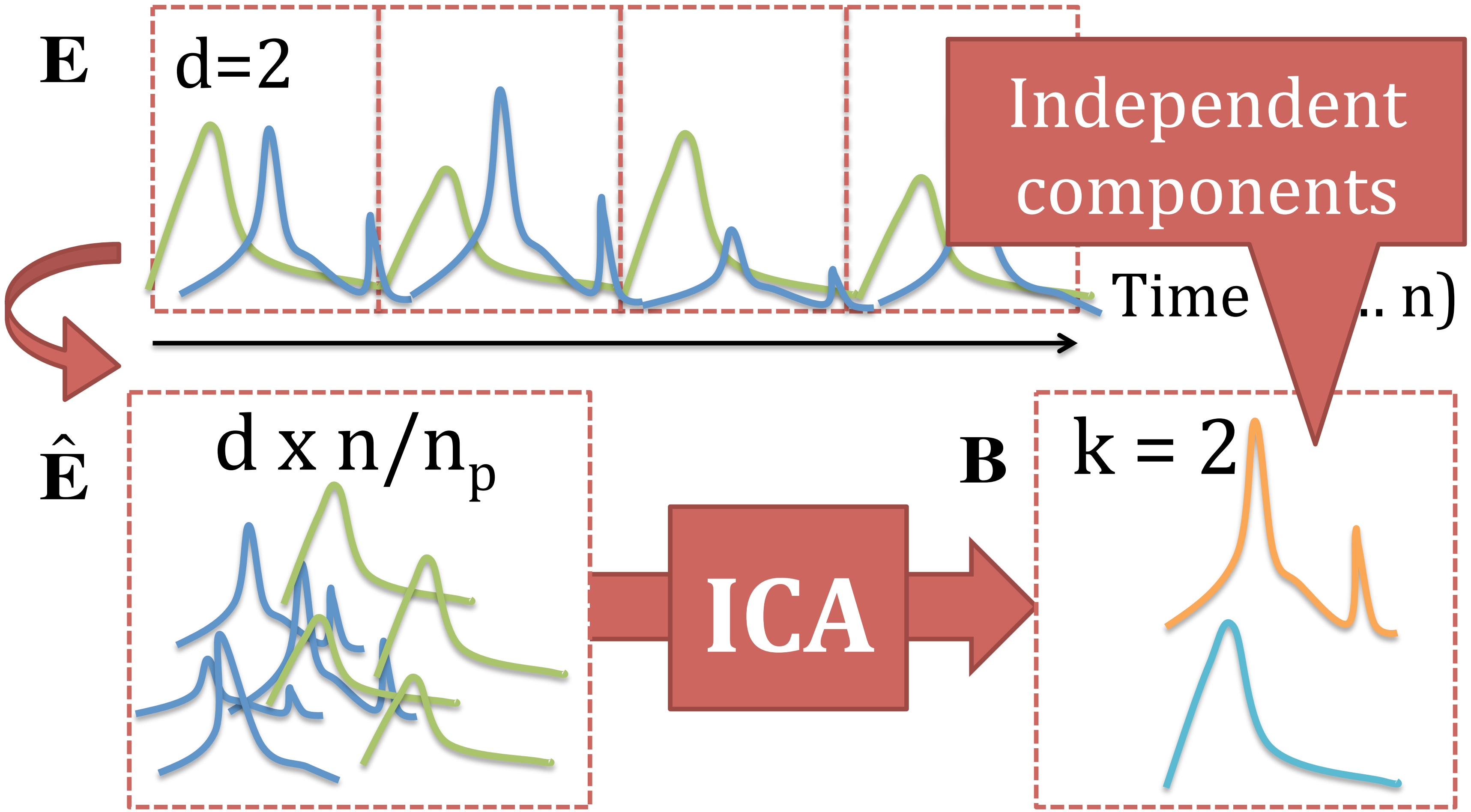
Idea (1): Seasonal component analysis

Idea(1-a) Seasonal component detection



Idea (1): Seasonal component analysis

Idea(1-a) Seasonal component detection

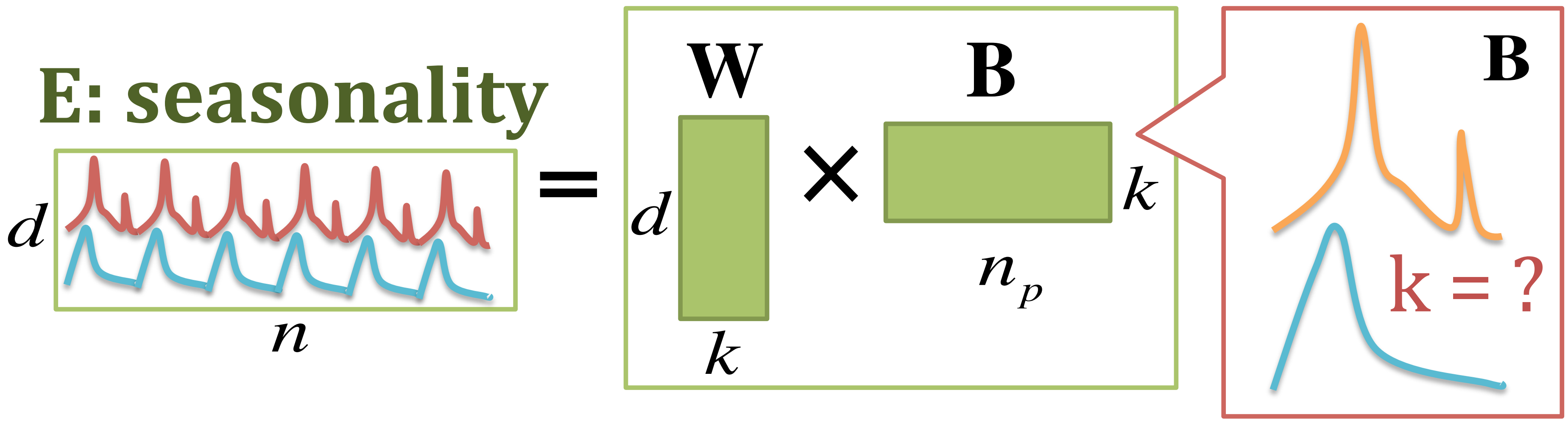


Idea (1): Seasonal component analysis

Idea(1-b) Automatic component analysis

Find optimal number k ($1 \le k \le d$)

d: dimension



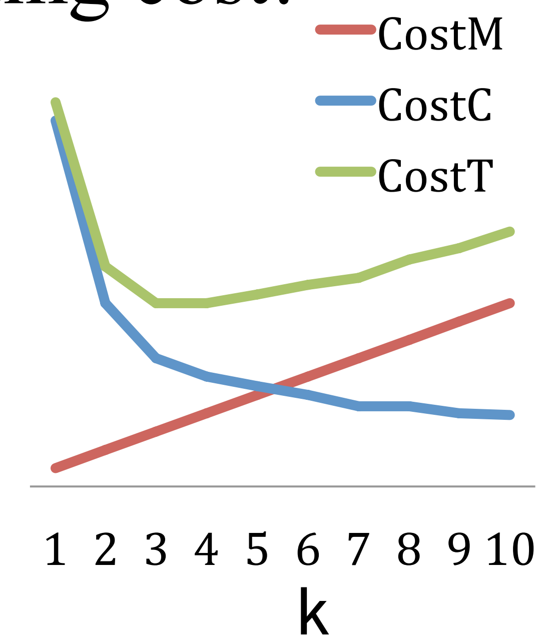
opt k=?

Idea (1): Seasonal component analysis

Idea(1-b) MDL -> Minimize encoding cost!

$$\min \left(\mathrm{Cost}_M(S) + \mathrm{Cost}_c(X \mid S) \right)$$

Model cost — Coding cost



Good compression ⟲ Good description

Kumamoto U

CMU CS

Idea (1): Seasonal component analysis

Idea(1-b) MDL -> Minimize encoding cost!

CostM

CostC

$$Cost_T(X;\mathcal{S}) = \log^*(d) + \log^*(n) + Cost_M(\boldsymbol{p}, \boldsymbol{r}, \boldsymbol{K})$$
$$+Cost_M(\mathbf{A}) + Cost_M(k, \mathbf{W}, \mathbf{B}) + Cost_C(X|\mathcal{S})$$

$$k_{opt} = \arg\min_k Cost_T(X;\mathcal{S})$$

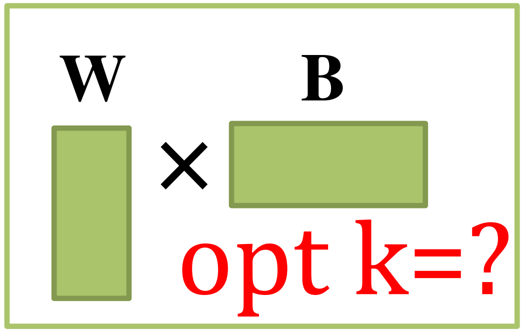
Good compression ⟲ Good description

Idea (1): Seasonal component analysis

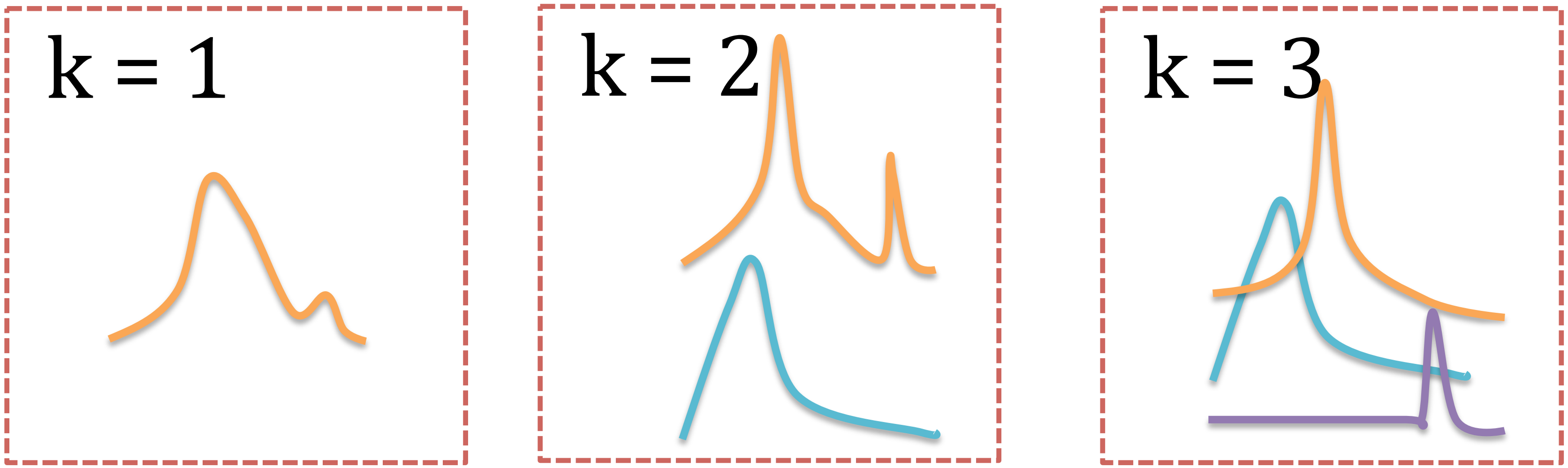
Idea(1-b)   Automatic component analysis

Find optimal number k (1≤k≤d)

d: dimension



B



Cost(1) = $$ Cost(2) = $ Cost(3) = $$$

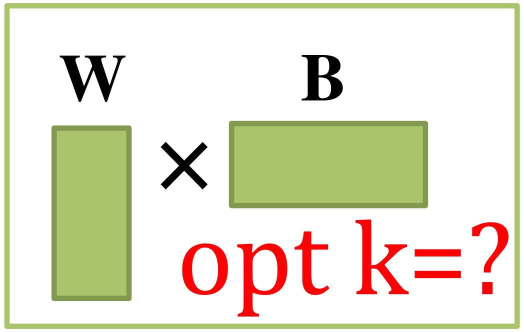
Kumamoto U

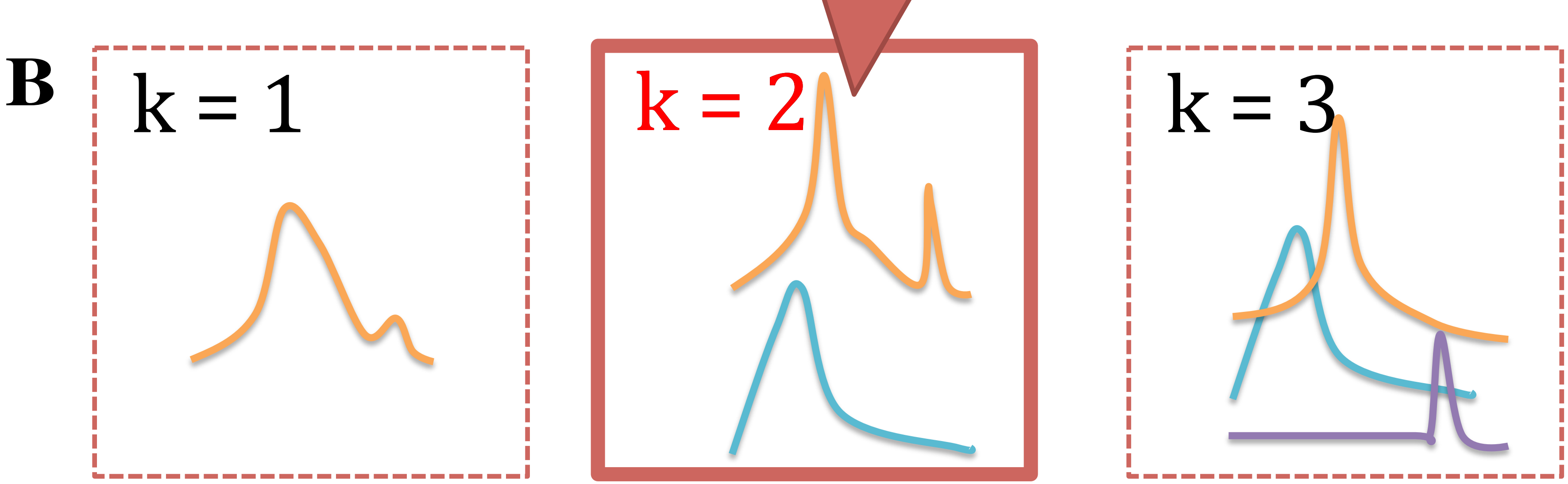
CMU CS

Idea (1): Seasonal component analysis

Idea(1-b)   Automatic component analysis

Find optimal number k ($1 \leq k \leq d$)

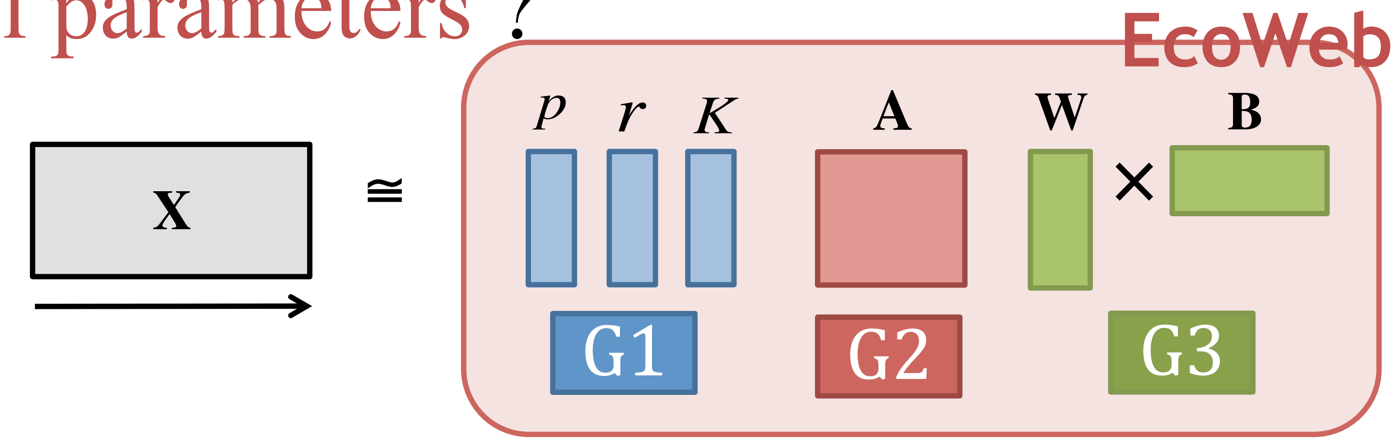




Cost(1) = $$ Cost(2) = $ Cost(3) = $$$

Idea (2): EcoWeb-Fit

Q2. How can we efficiently estimate model parameters ?



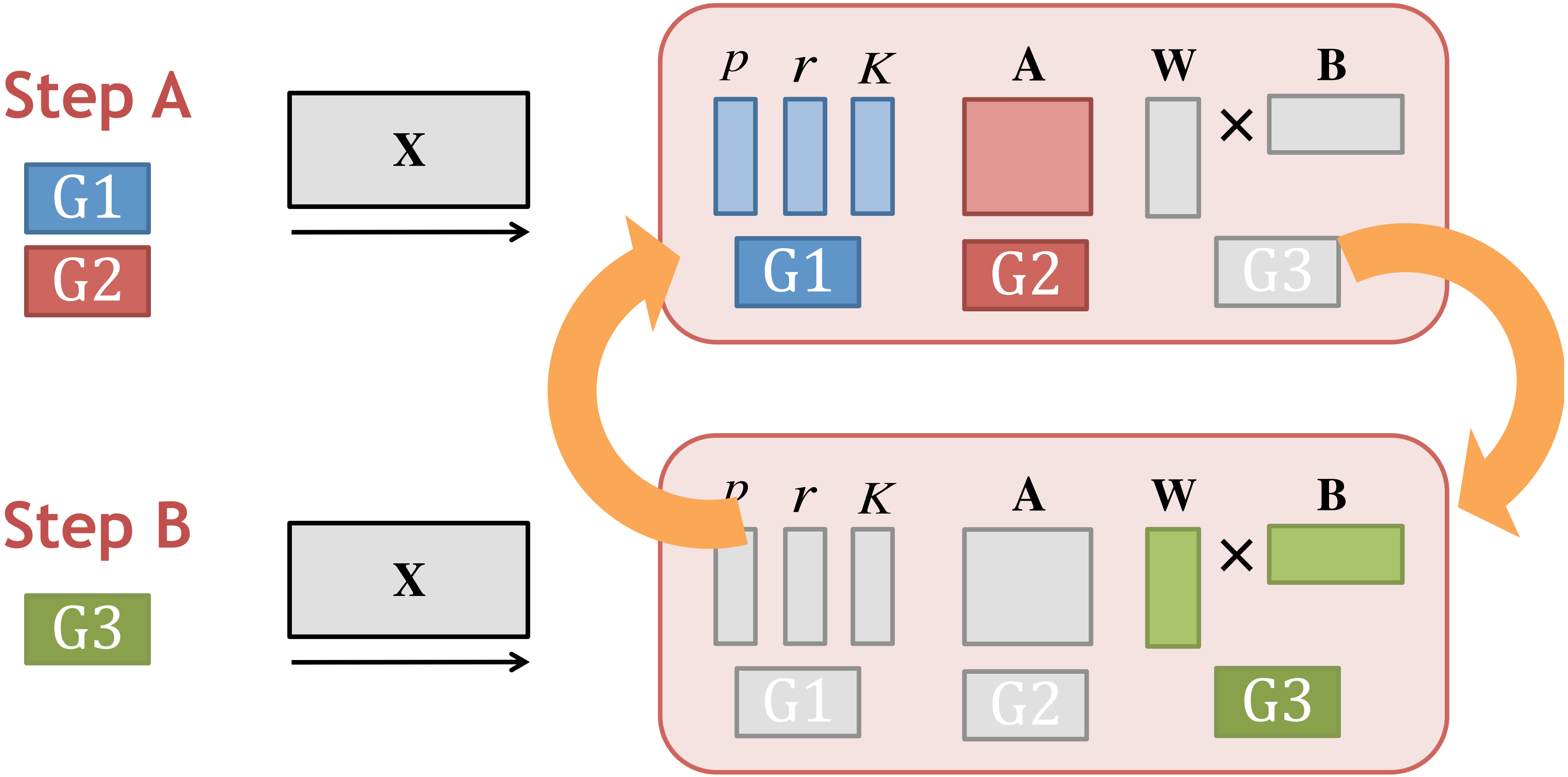
Idea (2): Multi-step fitting

a. **StepFit** (sub)

b. **EcoWeb-Fit** (full)

Idea (2): EcoWeb-Fit

(2-a). StepFit: Update parameters *alternately*

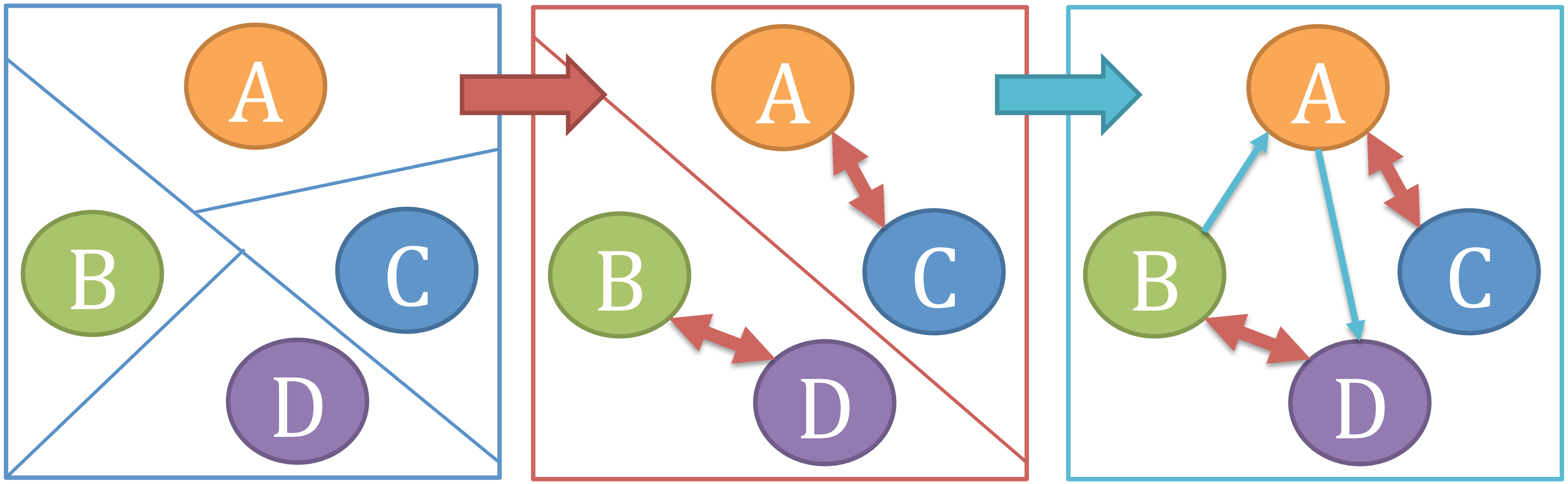


Idea (2): EcoWeb-Fit

(2-b). EcoWeb-Fit: full algorithm

e.g., 4 keywords: A B C D

1. Individual-Fit 2. Pair-Fit 3. Full-Fit



EcoWeb-Fit updates parameters, separately

Experiments

We answer the following questions…

Q1. Effectiveness

How successful is it in spotting patterns?

Q2. Accuracy

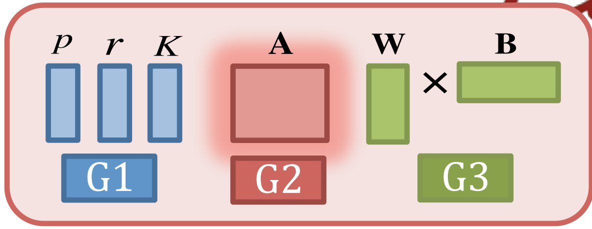
How well does it match the data?

Q3. Scalability

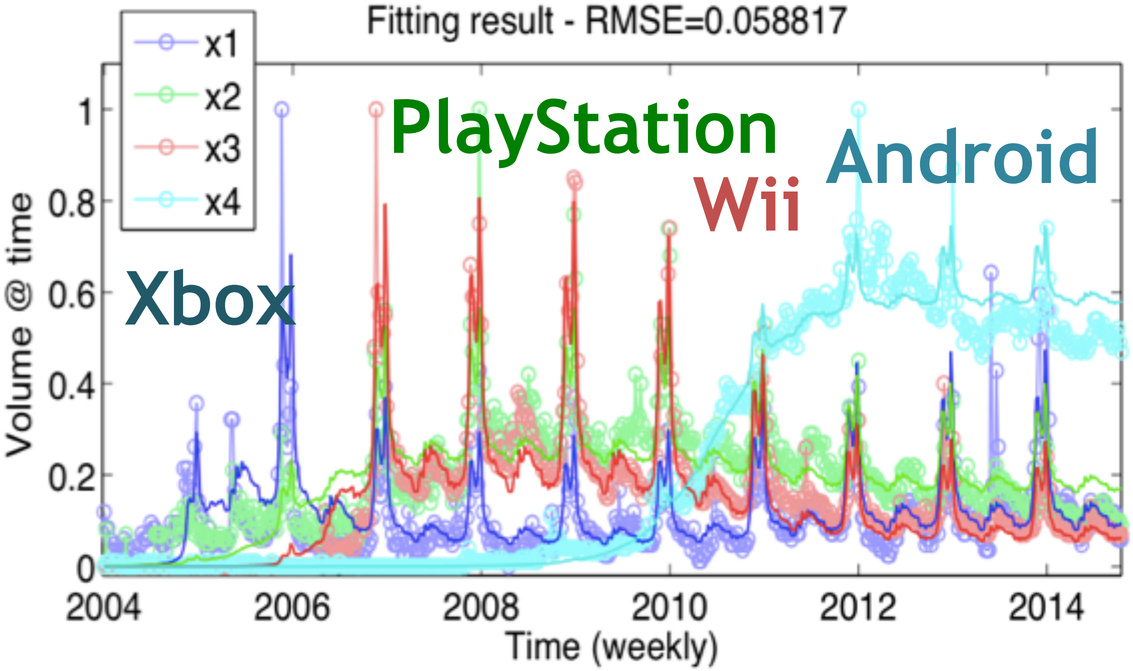
How does it scale in terms of computational time?

Q1. Effectiveness

(#1) Video games

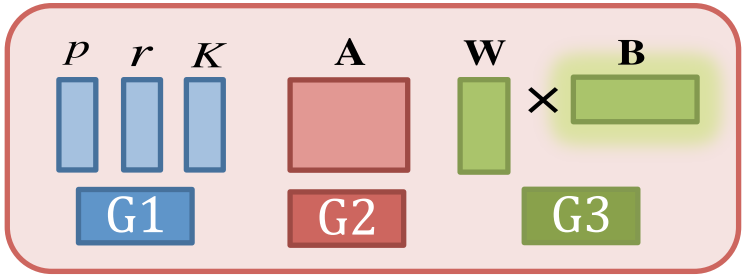


Interactions between keywords

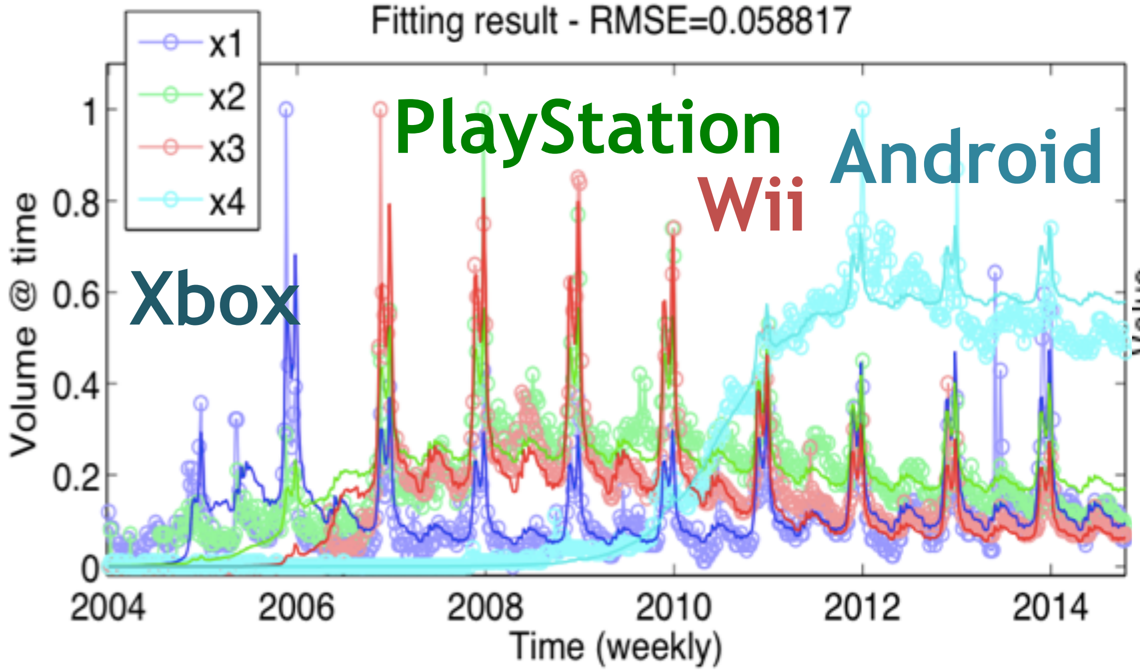


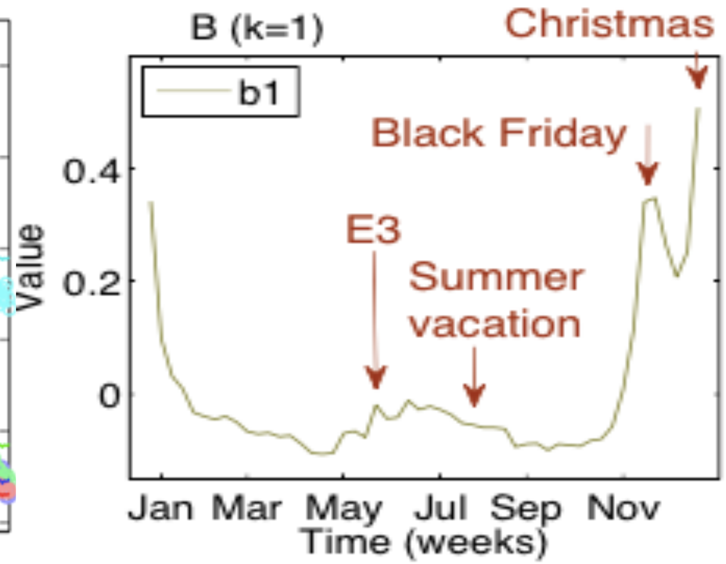
Q1. Effectiveness

(#1) Video games



Seasonality

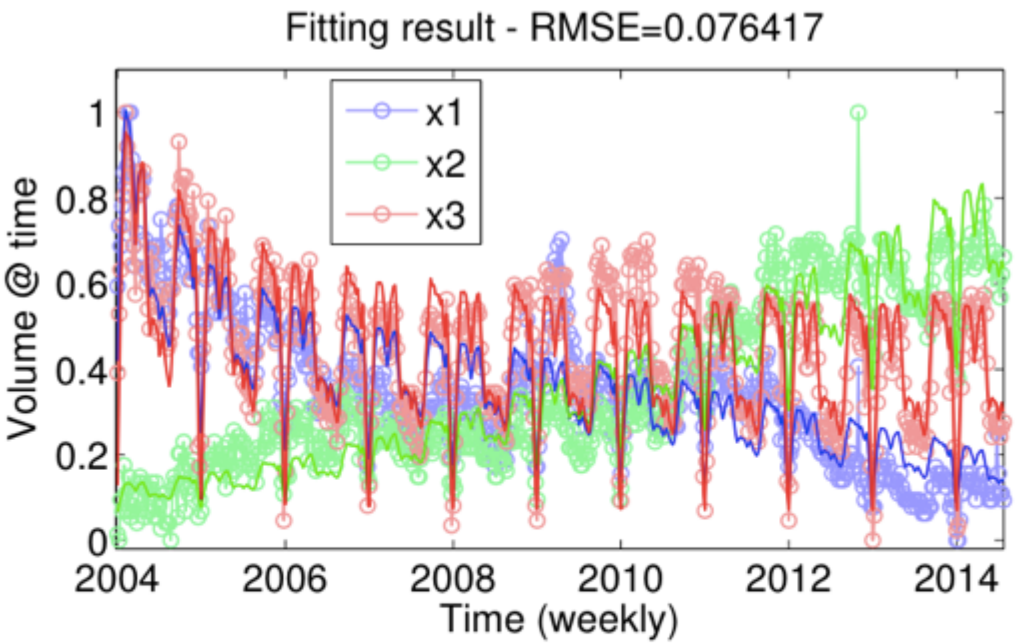




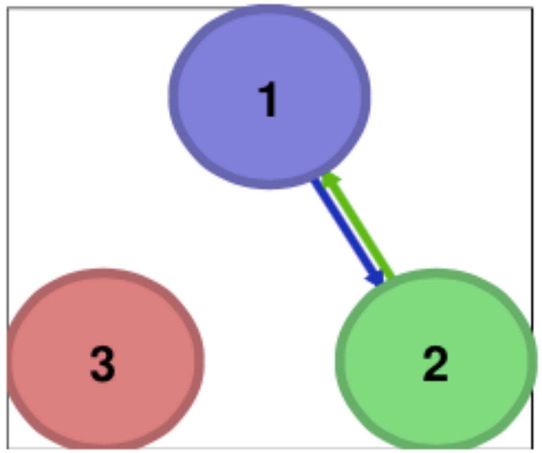
Q1. Effectiveness

(#2) Programming language

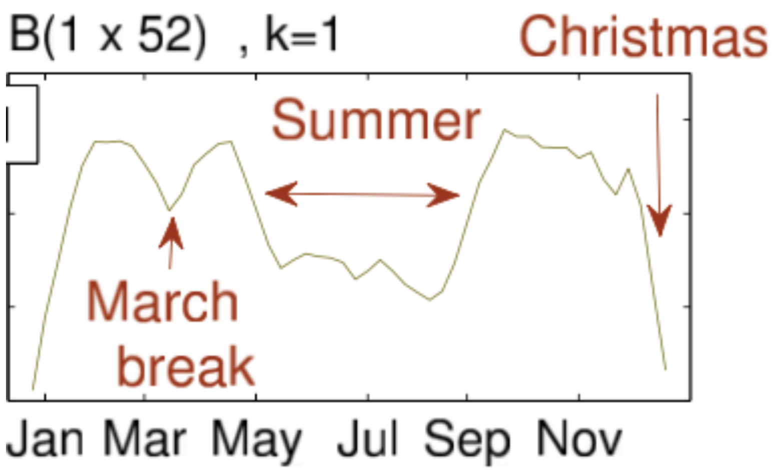
C , R , MATLAB



Interactions



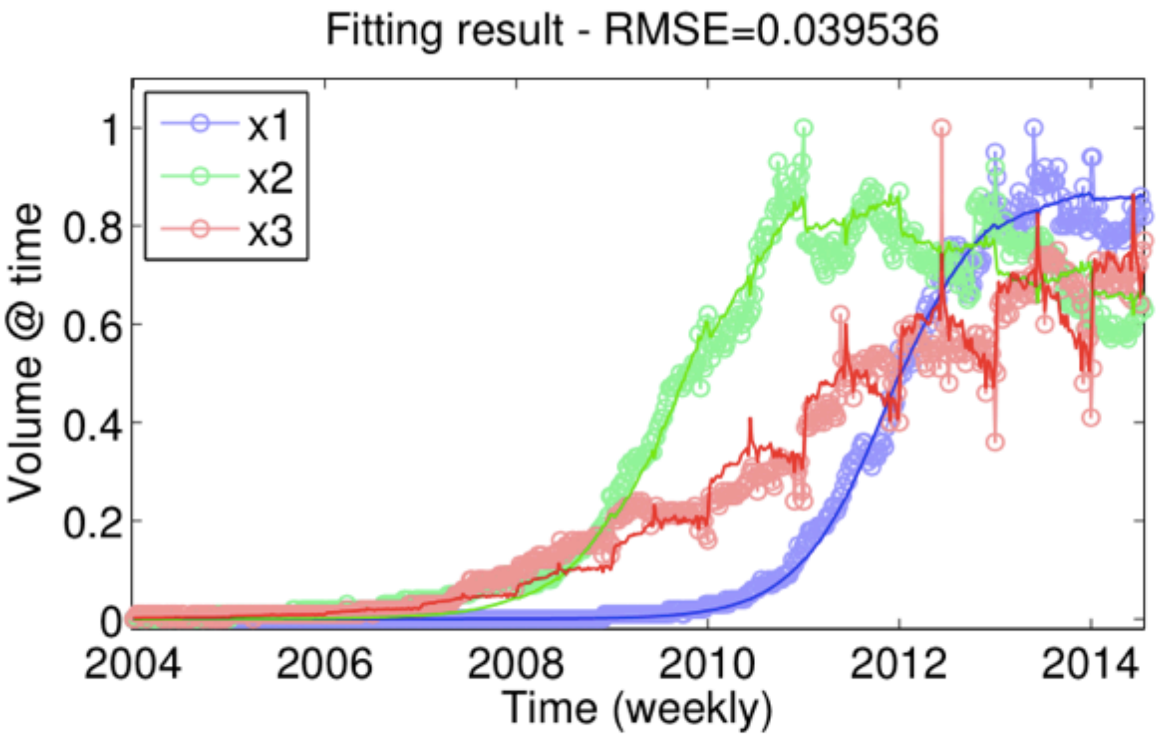
Seasonality



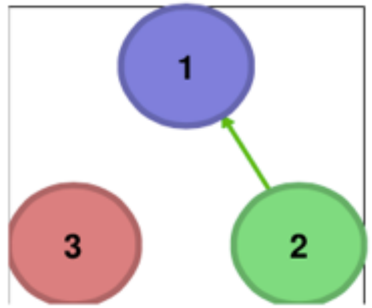
Q1. Effectiveness

(#3) Social media

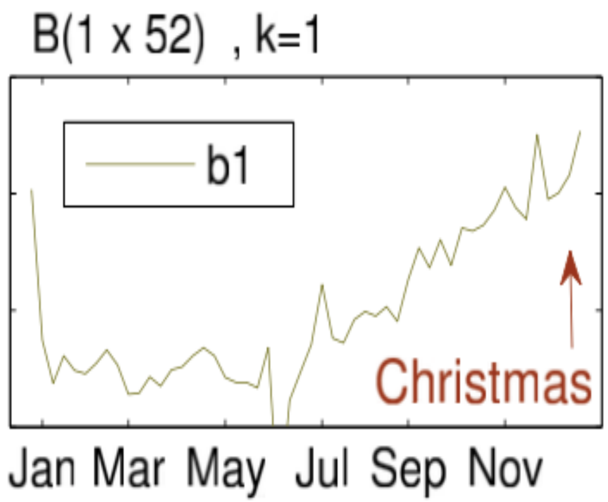
Tumblr , Facebook , LinkedIn



Interactions



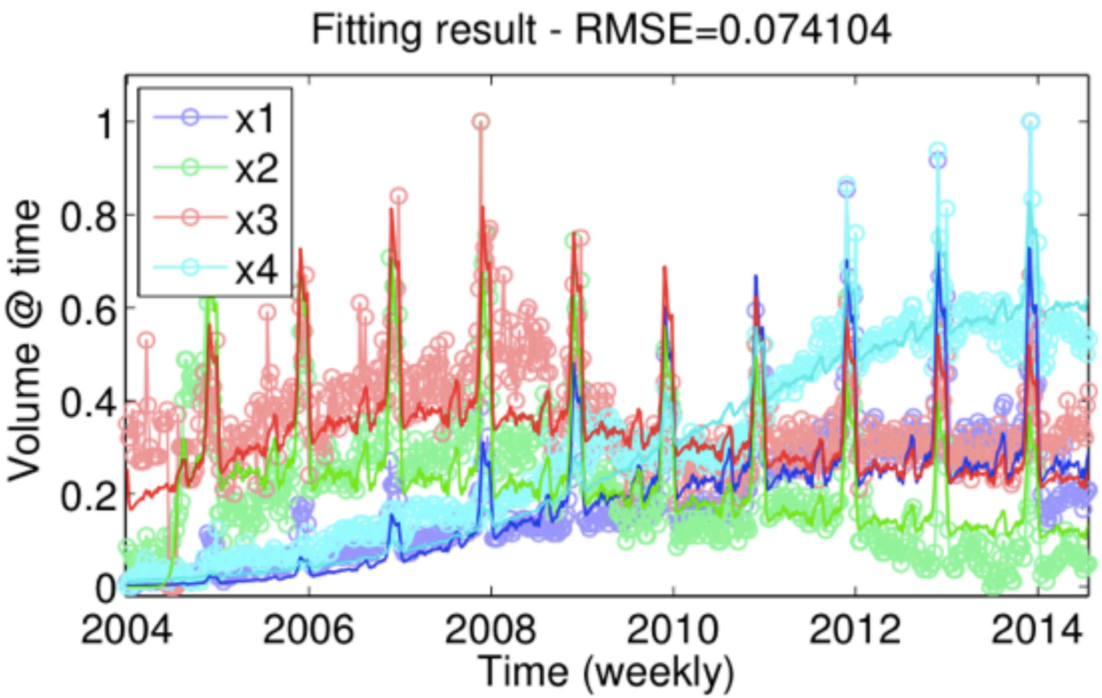
Seasonality

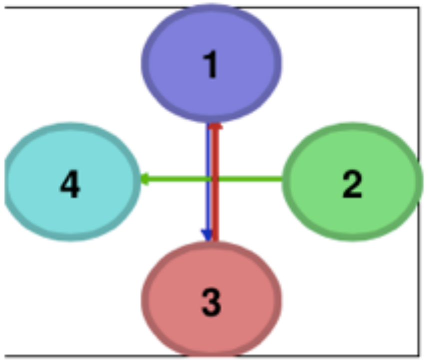


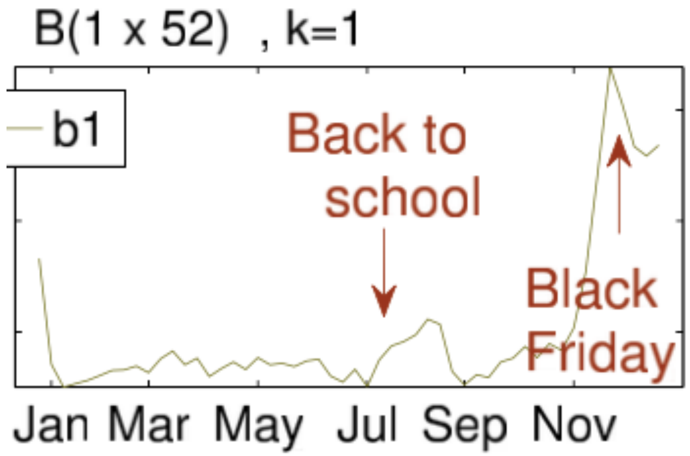
Q1. Effectiveness

(#4) Apparel companies

Kohls , JCPenny , Nordstrom , Forever21



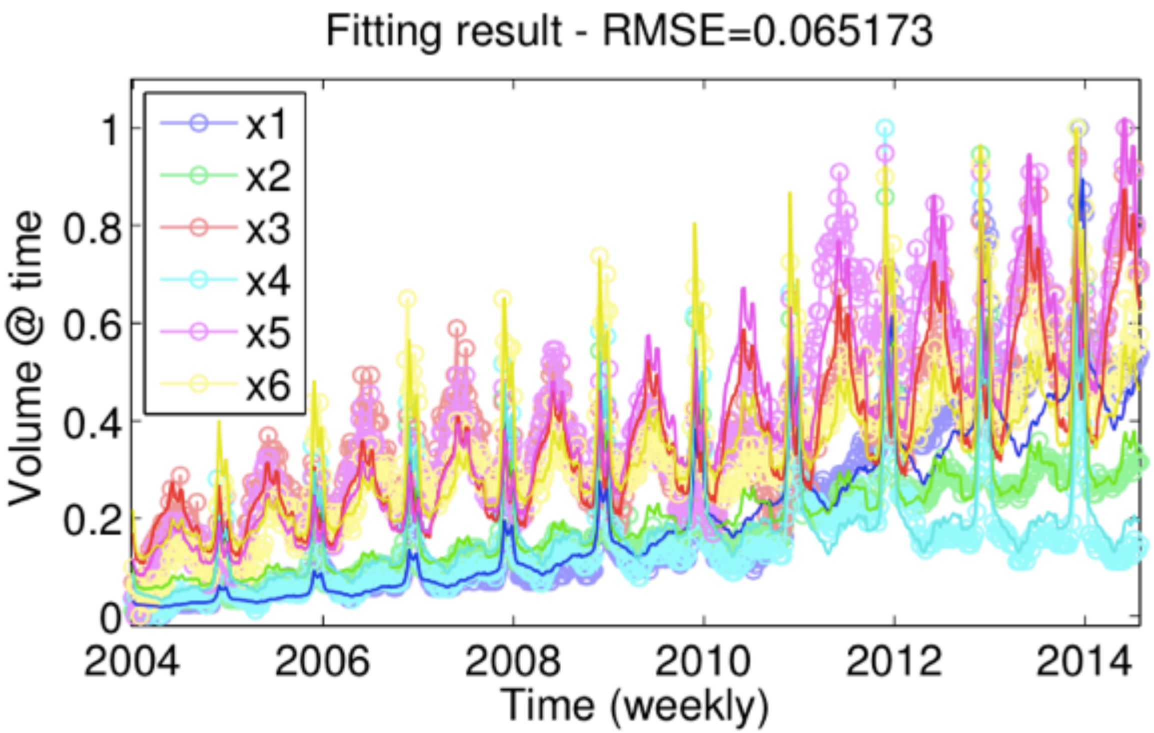




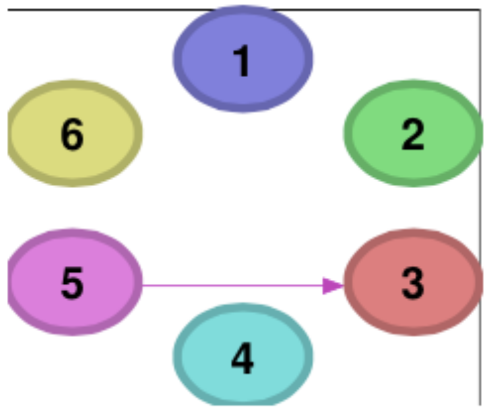
Q1. Effectiveness

(#5) Retail companies

Amazon , Walmart , Home Depot , BestBuy , Lowes , Costco



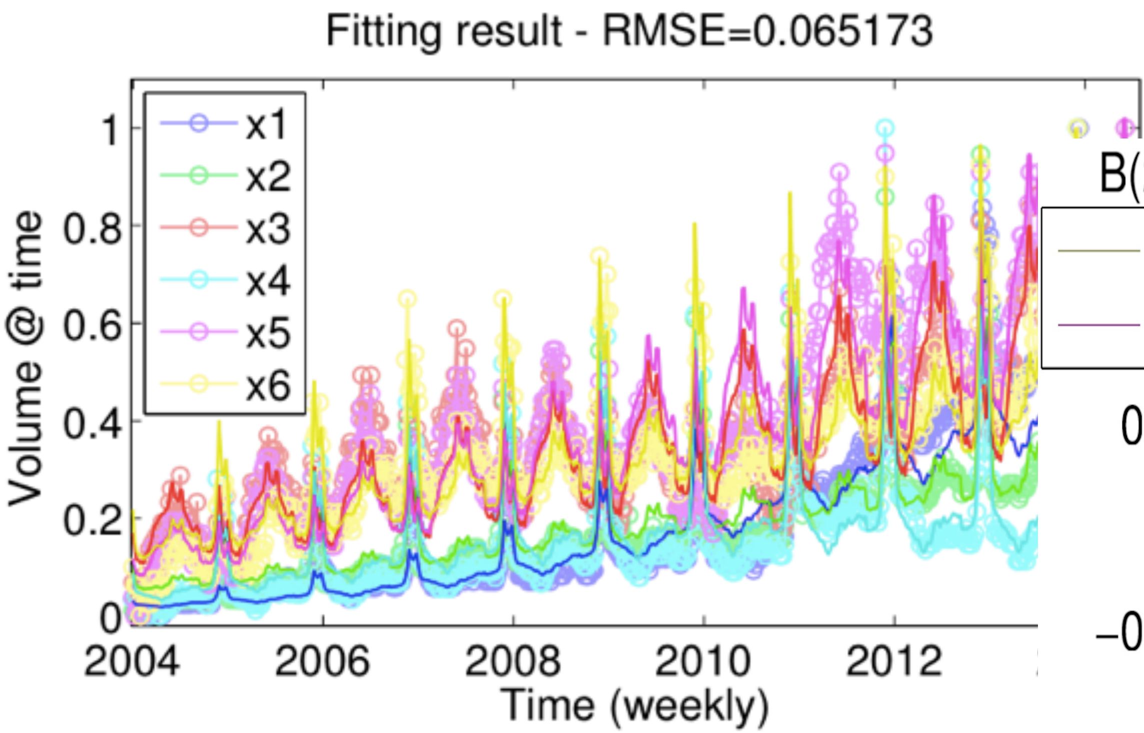
Interaction



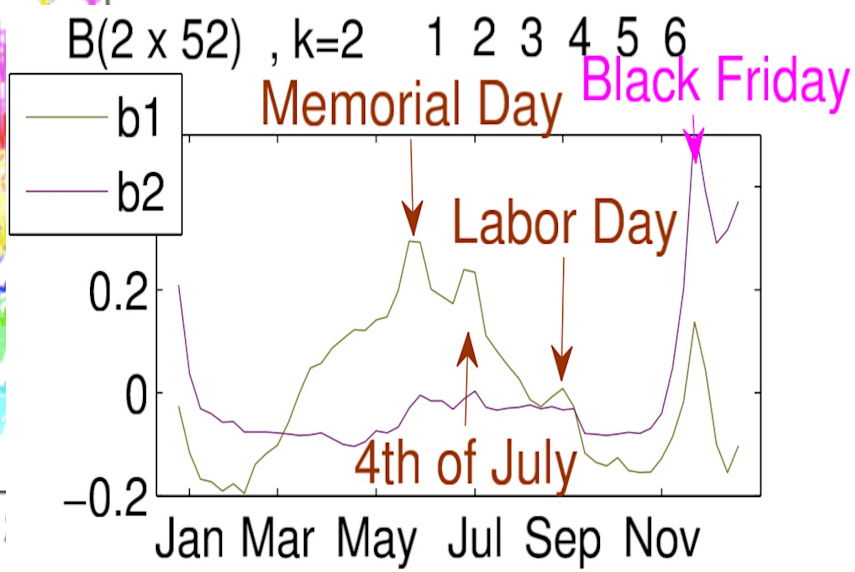
Q1. Effectiveness

(#5) Retail companies

Amazon , Walmart , Home Depot , BestBuy , Lowes , Costco



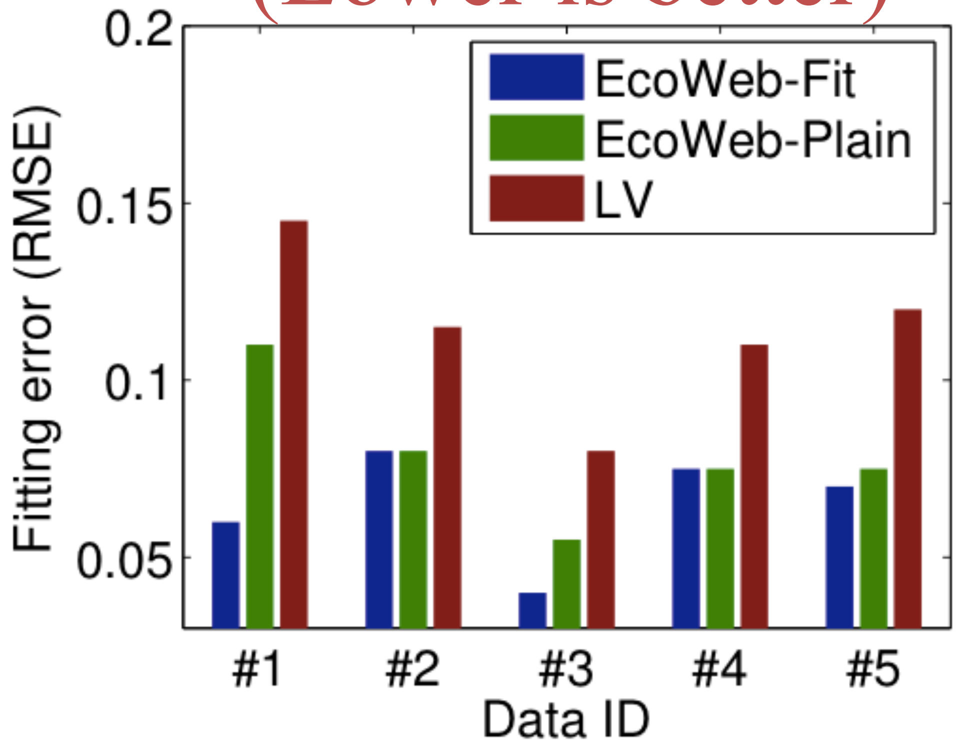
Seasonality



Q2. Accuracy

RMSE between original and fitted volume

(Lower is better)

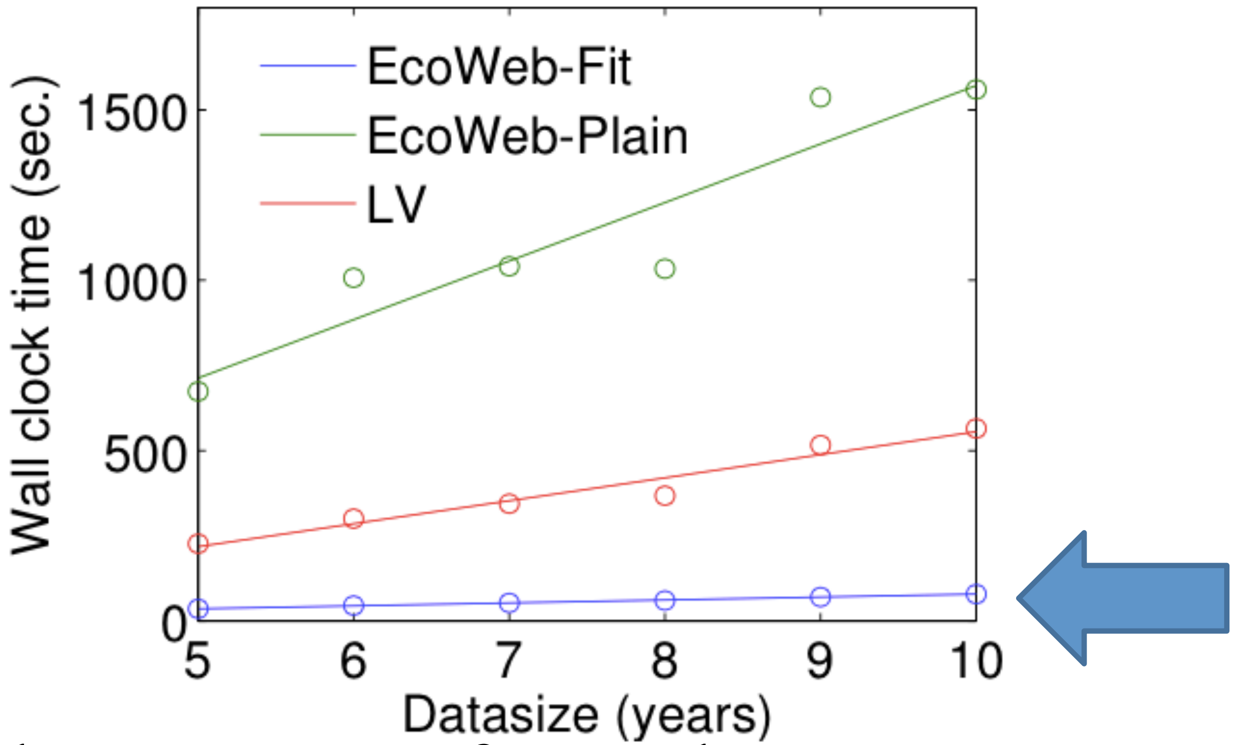


EcoWeb consistently wins!

Q3. Scalability

Wall clock time vs. dataset size (years)

EcoWeb-Fit scales linearly, i.e., O(n)



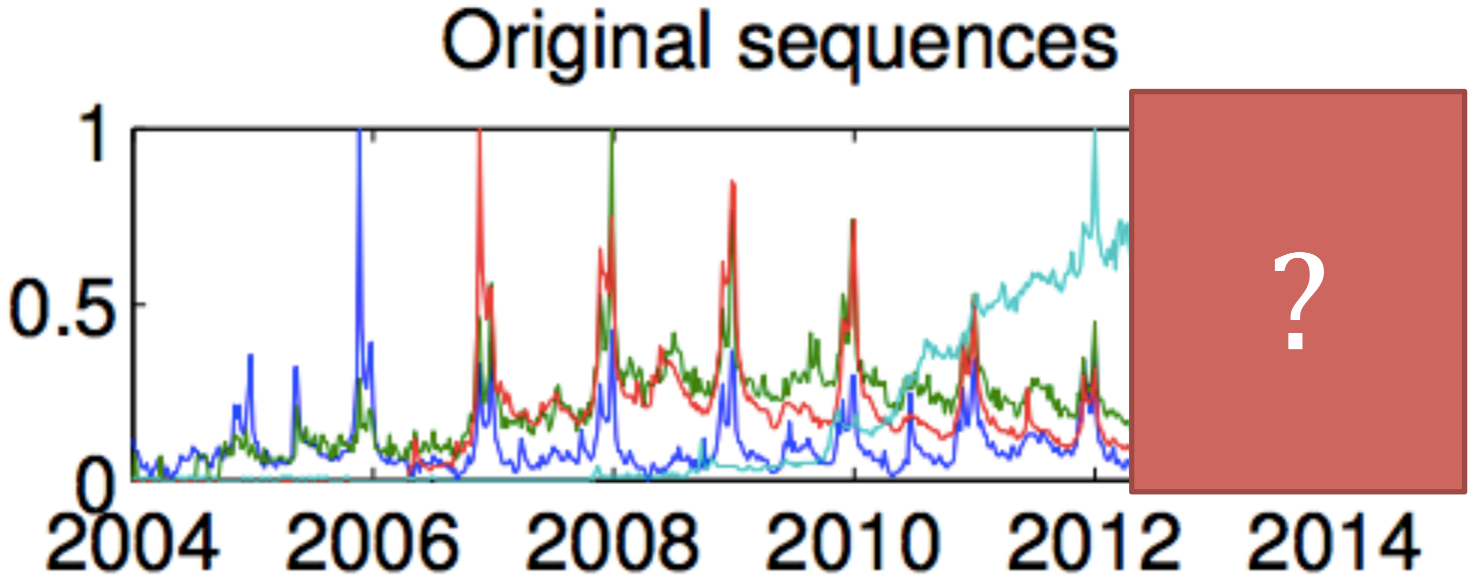
7x faster than **LV**, 20x faster than **EcoWeb-Plain**

EcoWeb at work - forecasting

Forecasting future activities

Train:
2/3 sequences

Forecast:
1/3 following years

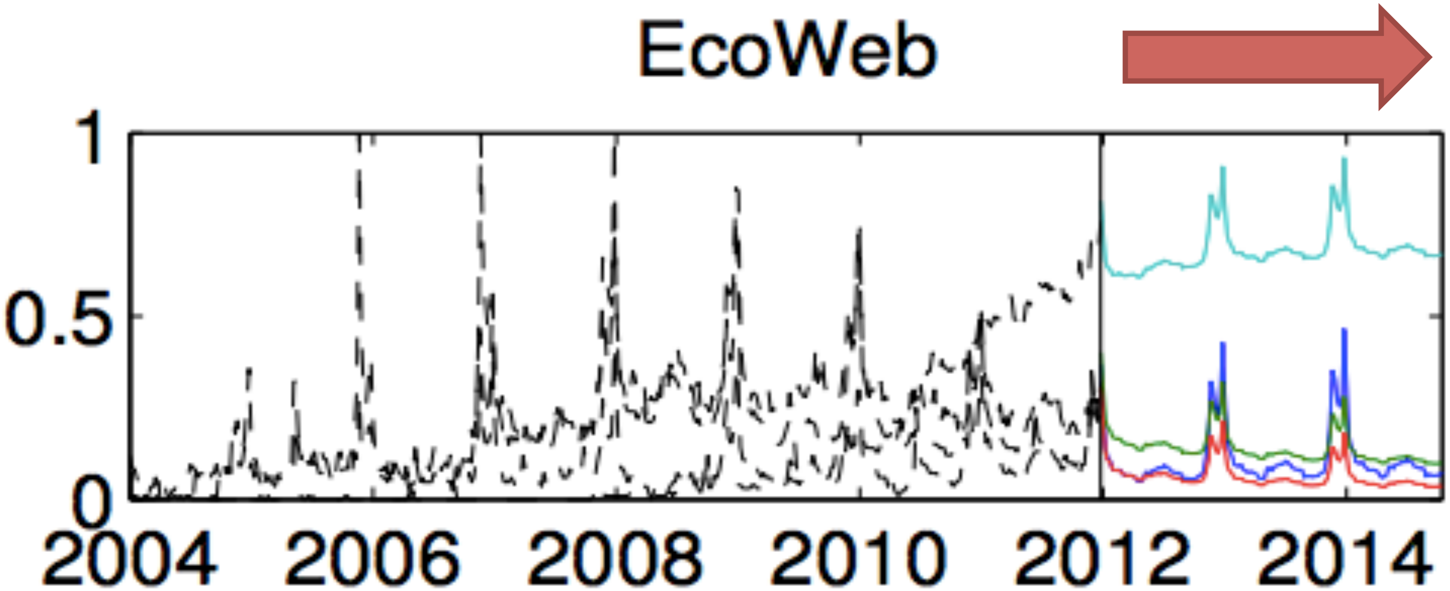


EcoWeb at work - forecasting

Forecasting future activities

Train:
2/3 sequences

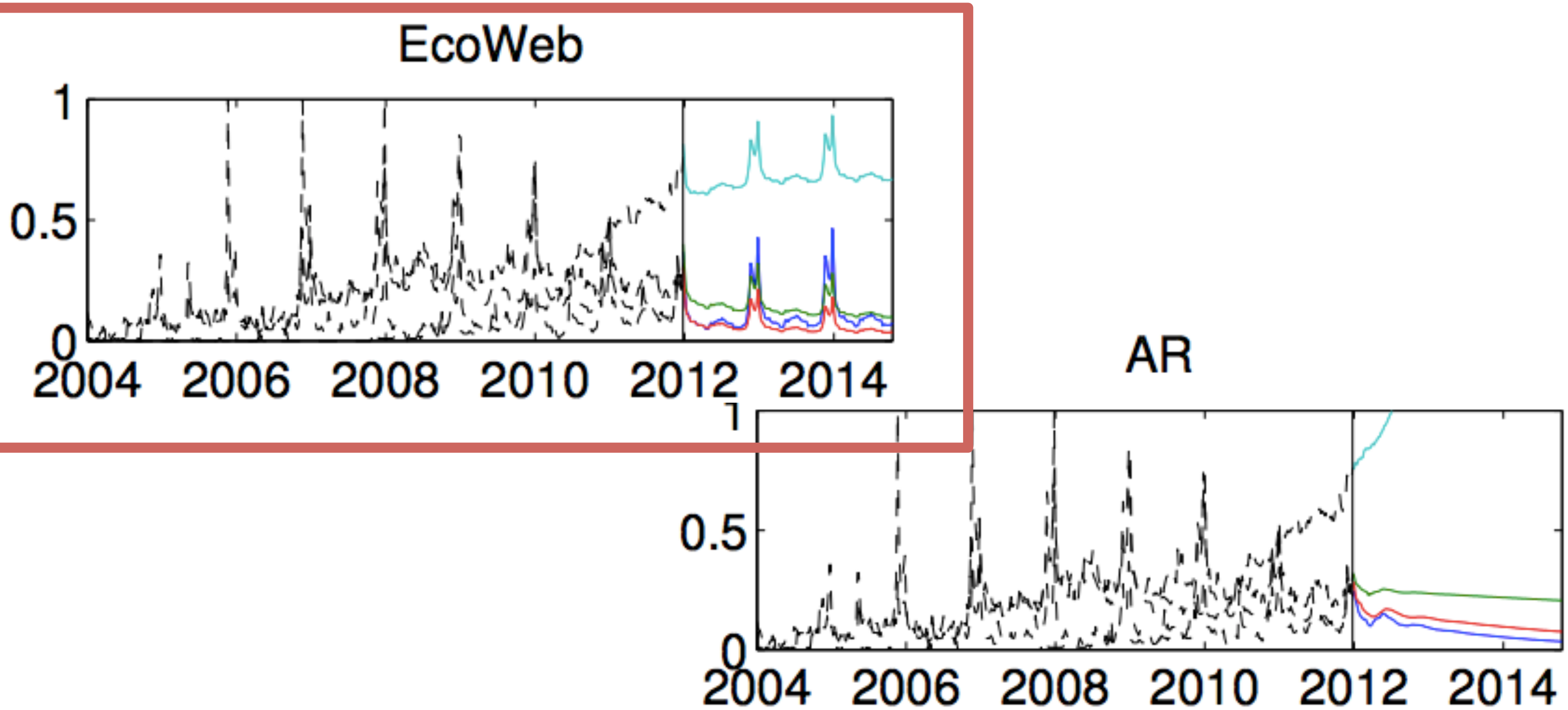
Forecast:
1/3 following years



EcoWeb can capture future patterns

EcoWeb at work - forecasting

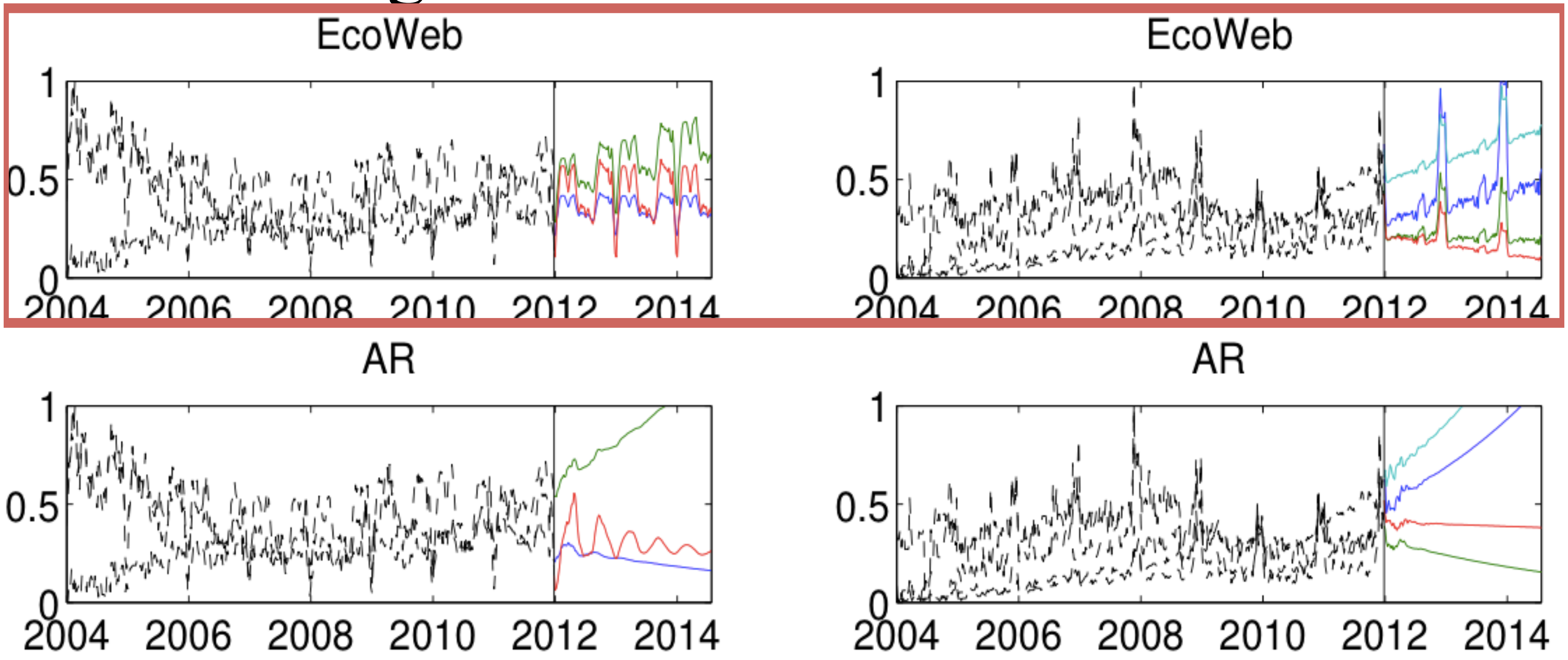
Forecasting future activities



EcoWeb can capture future patterns!

EcoWeb at work - forecasting

Forecasting future activities



(b) Programming languages (#2)

(c) Apparel companies (#4)

EcoWeb can capture future patterns!

Part 2 Roadmap

Problem

✔ Why: "non-linear" modeling

Fundamentals

✔ Non-linear (grey-box) models

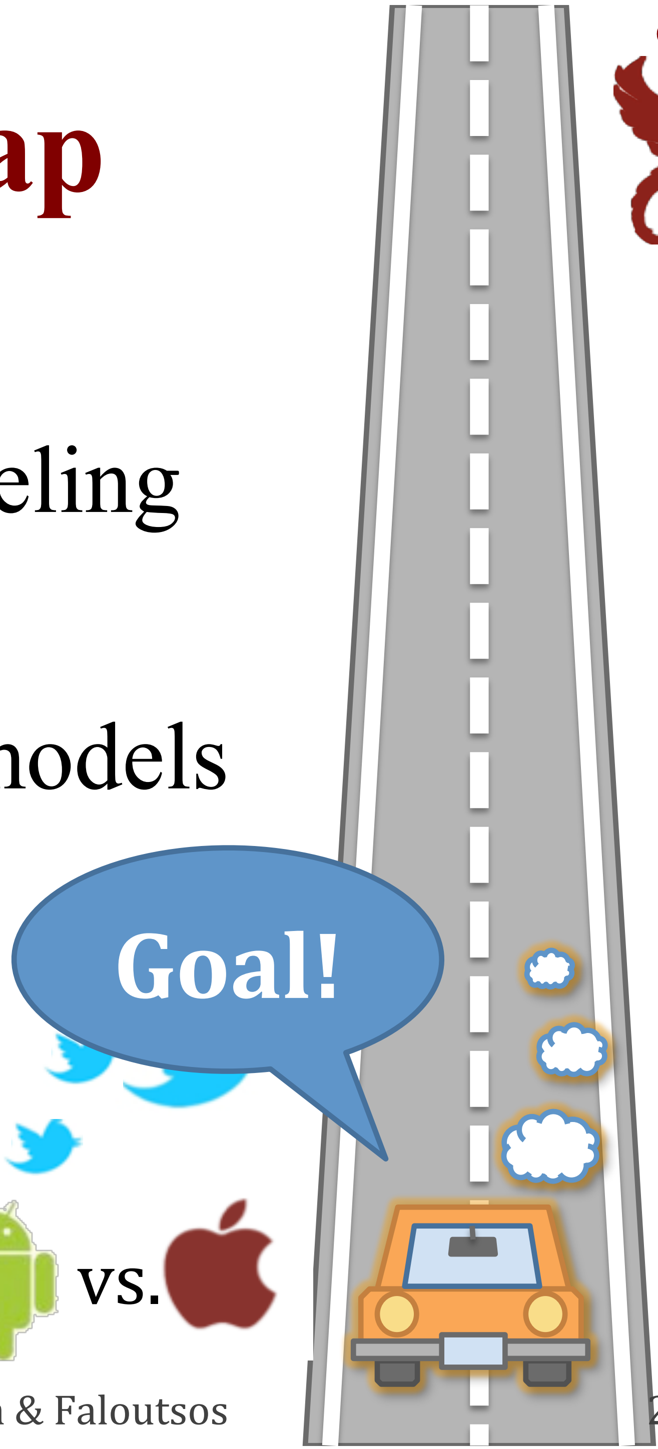
Applications

✔ Epidemics

✔ Information diffusion

✔ Online competition

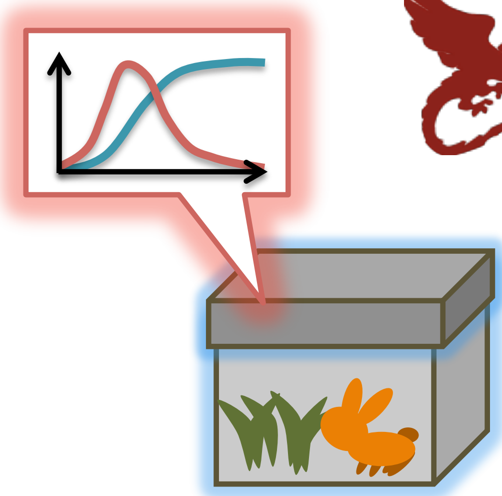
vs.



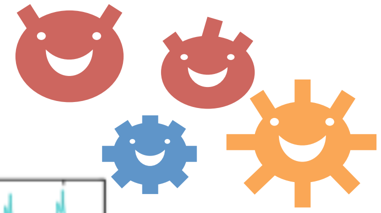
Kumamoto U

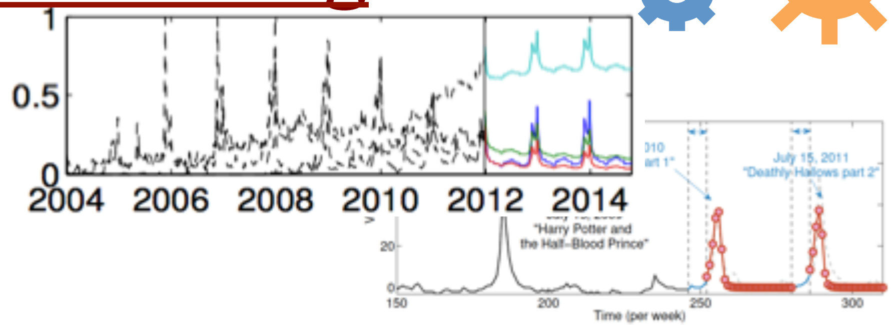
CMU CS

Part 2 Conclusions



- **Why: "non-linear" modeling**
  - Black box: lag plots (k-NN search)
  - Grey-box: given a model
- **Fundamentals: popular non-linear models**
  - Logistic function, Lotka-Volterra, Competition, ...
  - Epidemics (SI, SIR, SEIR, etc.), …
- **Applications: non-linear mining**
  - Epidemics
  - Information diffusion
  - Online competition





References (1)

Fundamentals

- Non-linear forecasting
  - D. Chakrabarti and C. Faloutsos *F4: Large-Scale Automated Forecasting using Fractals* CIKM 2002, Washington DC, Nov. 2002.
  - Sauer, T. (1994). *Time series prediction using delay coordinate embedding*. (in book by Weigend and Gershenfeld, below) Addison-Wesley.
  - Takens, F. (1981). *Detecting strange attractors in fluid turbulence*. Dynamical Systems and Turbulence. Berlin: Springer-Verlag.
- Non-linear equations and modeling
  - F. Brauer and C. Castillo-Chavez. *Mathematical models in population biology and epidemiology*, volume 40. Springer Verlag, New York, 2001.
  - R. M. Anderson and R. M. May. *Infectious Diseases of Humans Dynamics and Control*. Oxford University Press, 1992.
  - F. M. Bass. A new product growth for model consumer durables. *Management Science*, 15(5):215–227, 1969.
  - D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.
  - R. M. Anderson and R. M. May. Infectious Diseases of Humans. Oxford University Press, 1991.
  - R. M. May. Qualitative stability in model ecosystems. *Ecology*, 54(3):638–641, 1973.
  - M. Nowak. *Evolutionary Dynamics*. Harvard University Press, 2006.
  - Schuster, H. G. and Wagner, P. A model for neuronal oscillations. Biol. Cybern., 1990.
- Others
  - A. G. Hawkes and D. Oakes. A cluster representation of a self-exciting process. J. Appl. Prob., 11:493–503, 1974.

Kumamoto U

CMU CS

References (2)

Applications

- Epidemics
  - Rohani, P., Earn, D. J. D., Finkenstadt, B. F. & Grenfell, B. T. Population dynamic interference among childhood diseases. Proc. R. Soc. Lond. B 265, 2033–2041 (1998).
  - Rohani, P., Green, C.J., Mantilla-Beniers, N.B. & Grenfell, B.T. Ecological Interference Among Fatal Infections. Nature 422: 885-888 (2003).
  - L.Stone,R.Olinky,andA.Huppert.Seasonaldynamicsofrecurrentepidemics. *Nature*, 446:533–536, March 2007.
  - Y. Matsubara, Y. Sakurai, W. G. van Panhuis, and C. Faloutsos. FUNNEL: automatic mining of spatially coevolving epidemics. In *KDD*, pages 105–114, 2014.
- Information diffusion
  - J. Leskovec, L. Backstrom, and J. M. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD*, pages 497–506, 2009.
  - J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *WSDM*, pages 177–186, 2011.
    J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *ICDM*, pages 599–608, 2010.
    R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. In *PNAS*, 2008.
  - F. Figueiredo, J. M. Almeida, Y. Matsubara, B. Ribeiro, and C. Faloutsos. Revisit behavior in social media: The phoenix-r model and discoveries. In *PKDD*, pages 386–401, 2014.
  - Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and
  - C. Faloutsos. Rise and fall patterns of information diffusion:
  - model and implications. In *KDD*, pages 6–14, 2012.
- Online activities and competition
  - B. A. Prakash, A. Beutel, R. Rosenfeld, and C. Faloutsos. Winner takes all: competing viruses or ideas on fair-play networks. In *WWW*, pages 1037–1046, 2012.
  - A. Beutel, B. A. Prakash, R. Rosenfeld, and C. Faloutsos. Interacting viruses in networks: can both survive? In *KDD*, pages 426–434, 2012.
  - Y. Matsubara, Y. Sakurai, and C. Faloutsos. The web as a jungle: Non-linear dynamical systems for co-evolving online activities. In *WWW*, 2015.

Part 2

Non-linear mining and forecasting

Yasushi Sakurai (Kumamoto University)
Yasuko Matsubara (Kumamoto University)
Christos Faloutsos (Carnegie Mellon University)