大規模時系列データのための特徴自動抽出と将来予測

松原靖子

Sakurai Lab.

@ Kumamoto University

時系列ビッグデータ解析とは？

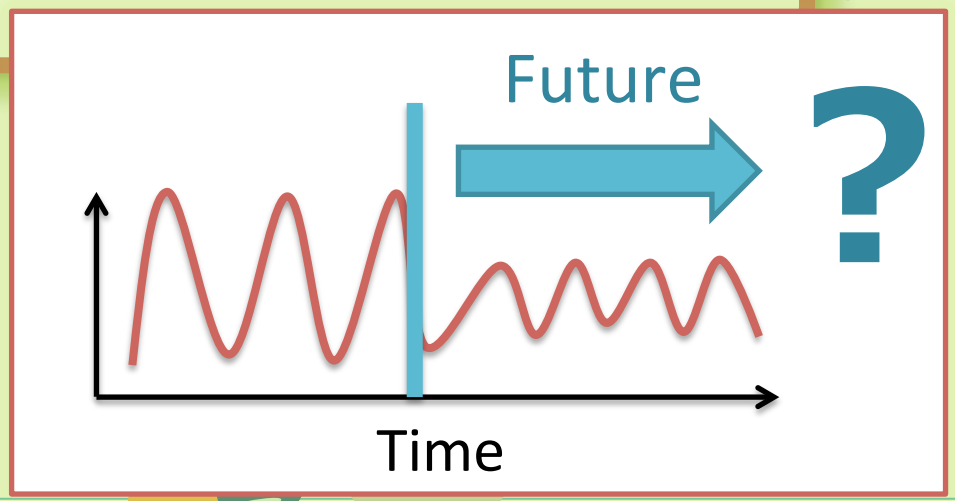
- **Webデータ**
  - Twitter, Google, アクセス履歴
- **センサデータ**
  - モーションキャプチャ
- **医療データ**
  - 地域別疫病データ

時系列ビッグデータ解析とは？

- **Webデータ**
  - Twitter, Google, アクセス履歴
- **Goal (1): 重要な情報の自動抽出**
- **医療データ**
  - 地域別疫病データ

時系列ビッグデータ解析とは？

- **Webデータ**
  - Twitter, Google, アクセス履歴
- **Goal (2): 将来予測**
- **医療データ**
  - 地域別疫病データ

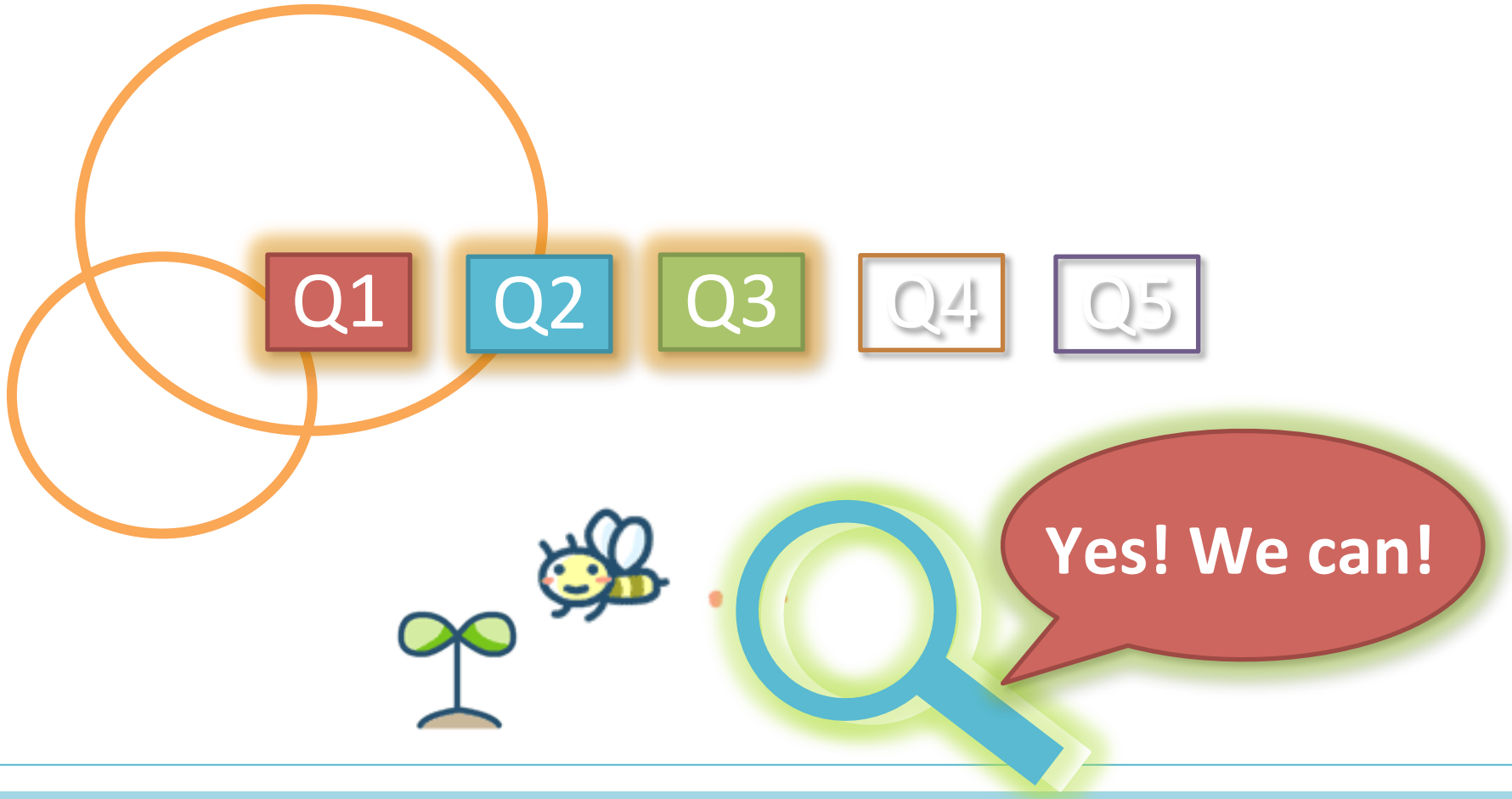


例えば…

こんな疑問にお応えします！

例えば…

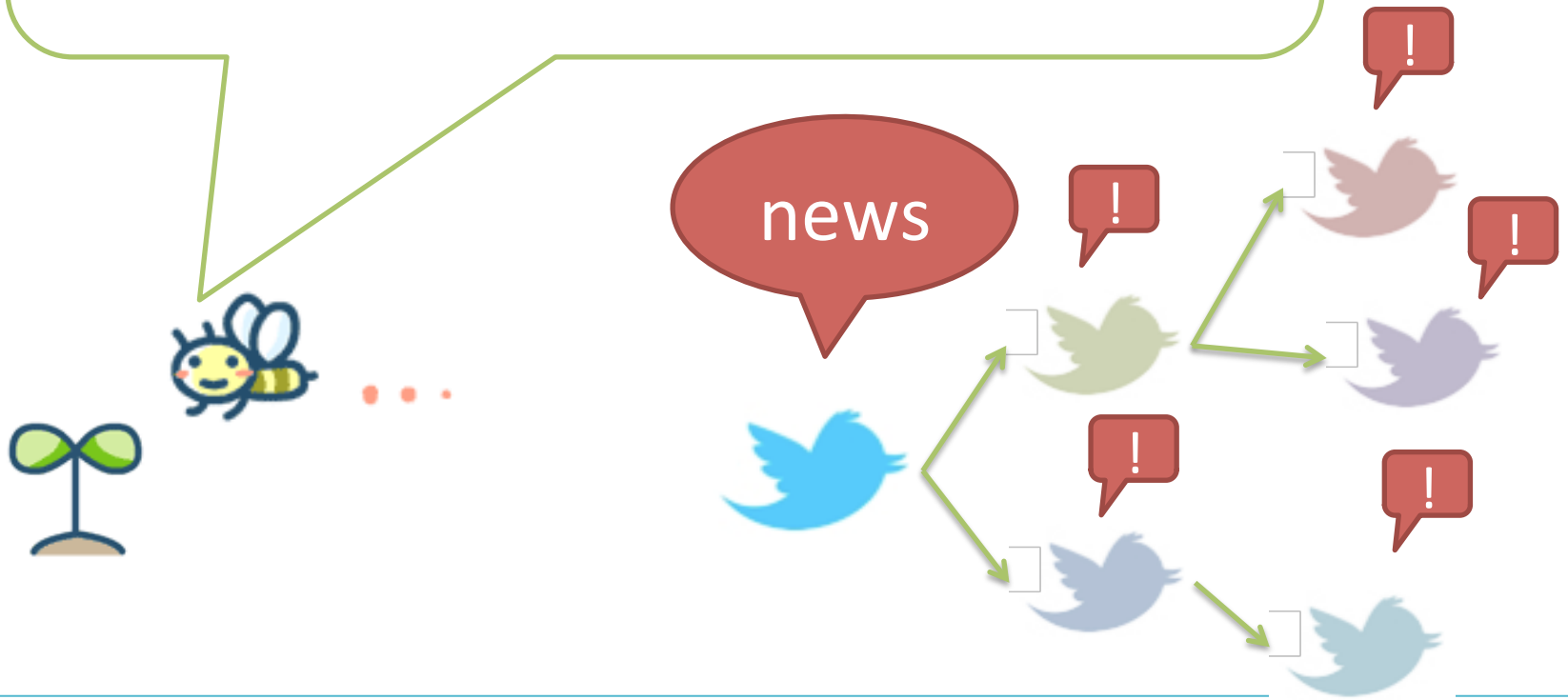
Webデータ



Q1. Webデータ

Q1

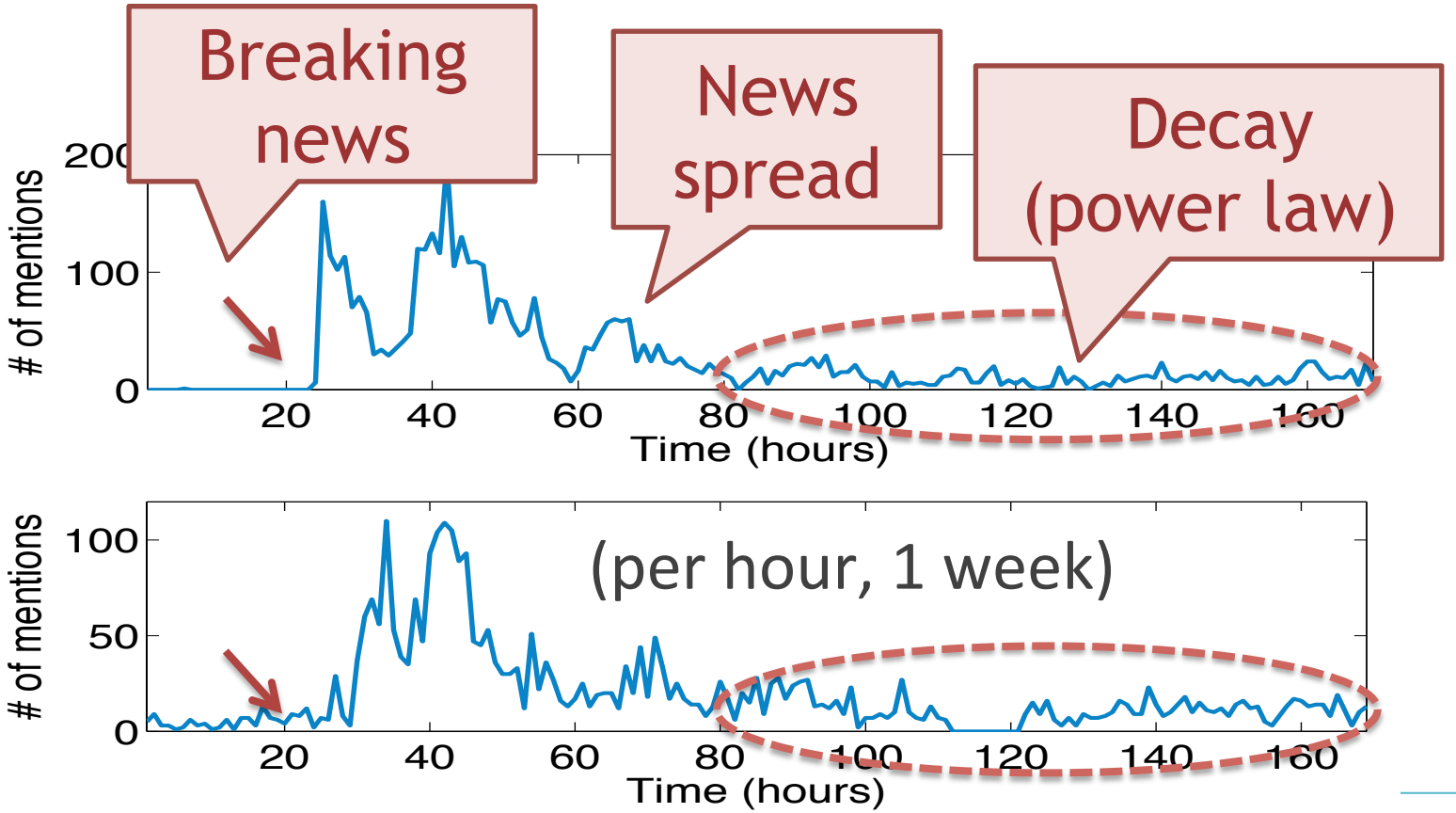
Twitter，ブログ，FB，…
噂やニュースって
どうやって伝わるの？



SpikeM KDD'12

News spread in social media

Number of mentions in blogs/Twitter

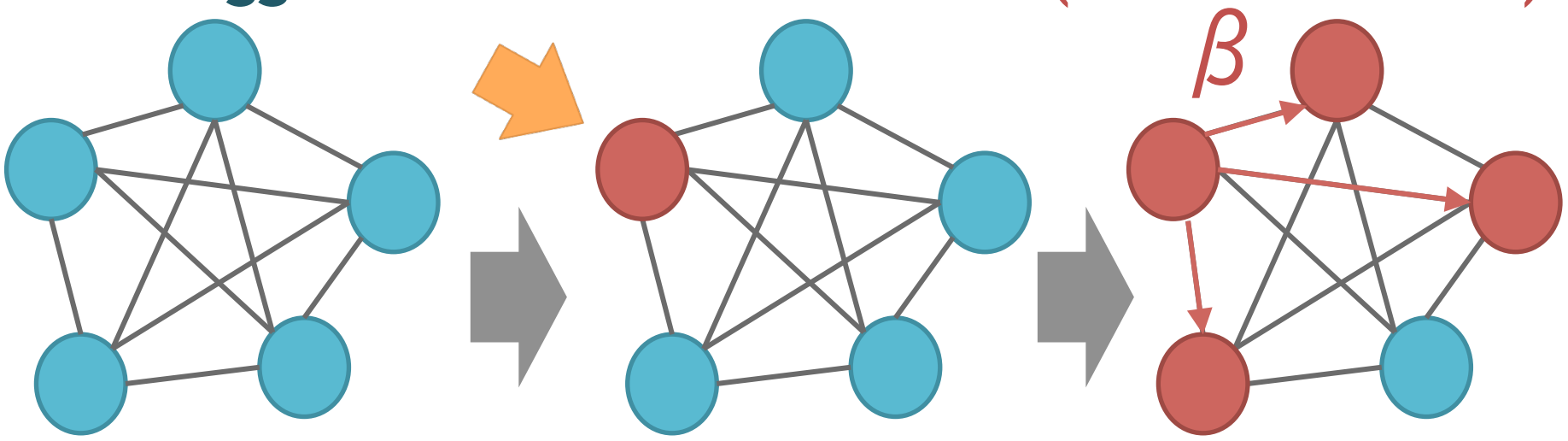


Q1. Webデータ

SpikeM KDD'12

News spread in social media

1. Un-informed bloggers
2. External shock
3. Infection (word-of-mouth)



Q1. Webデータ

SpikeM KDD'12

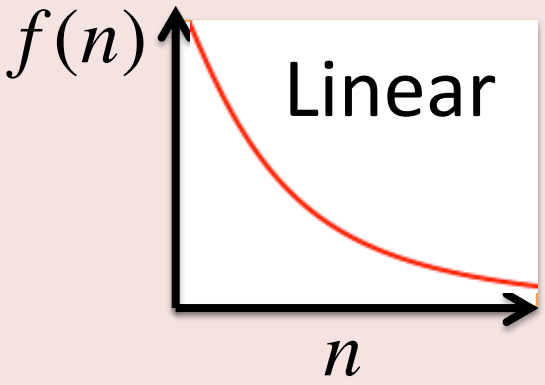
News spread in social media

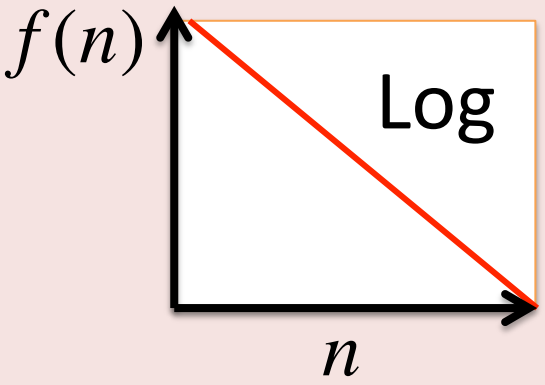
1. Un-informed bloggers

2. External shock

Power law

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





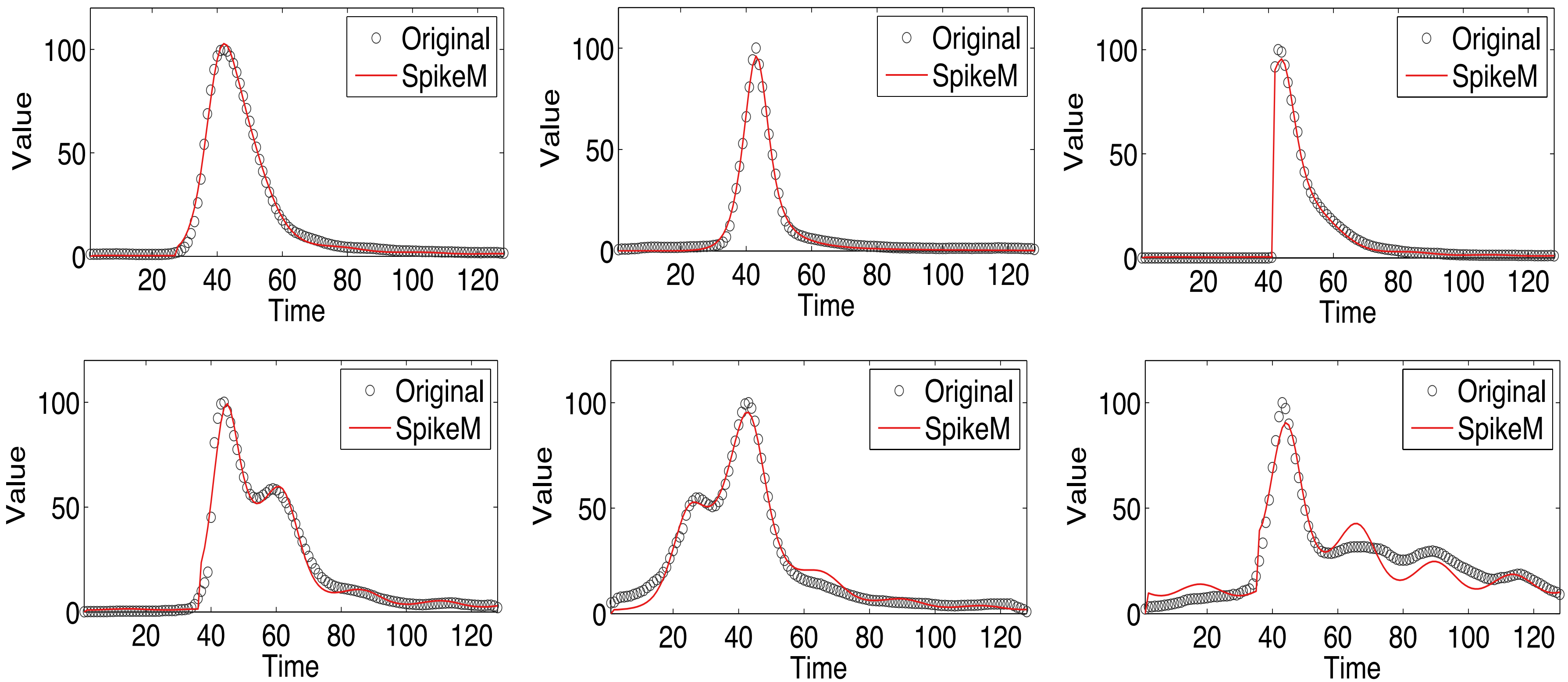
Q1. Webデータ

Q1 Q2 Q3 Q4 Q5

SpikeM KDD'12

News spread in social media

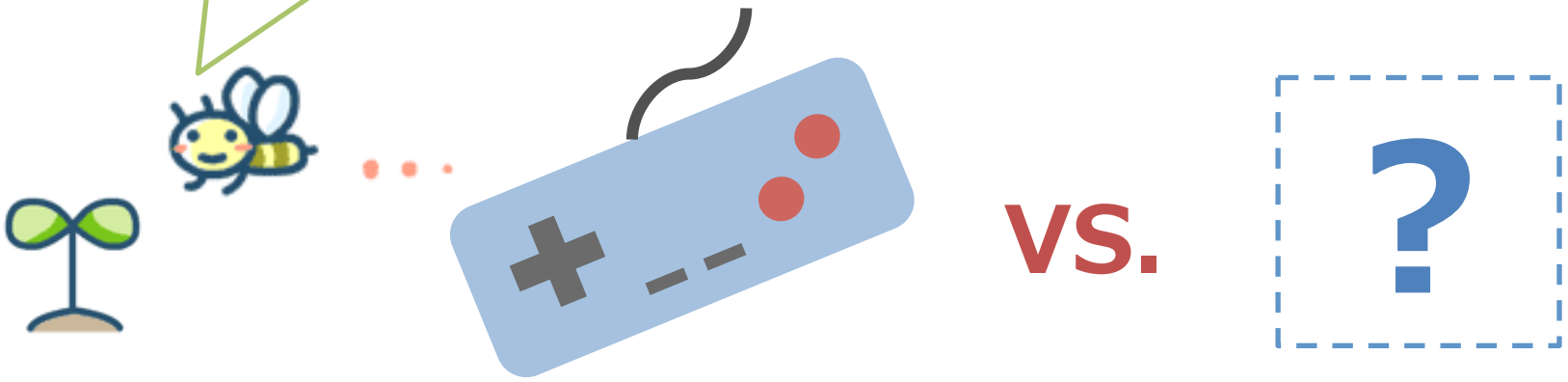
Memes in social media/blogs



Q2. Webデータ

Q2

Web上の闘い！
Wiiのライバルは誰？！

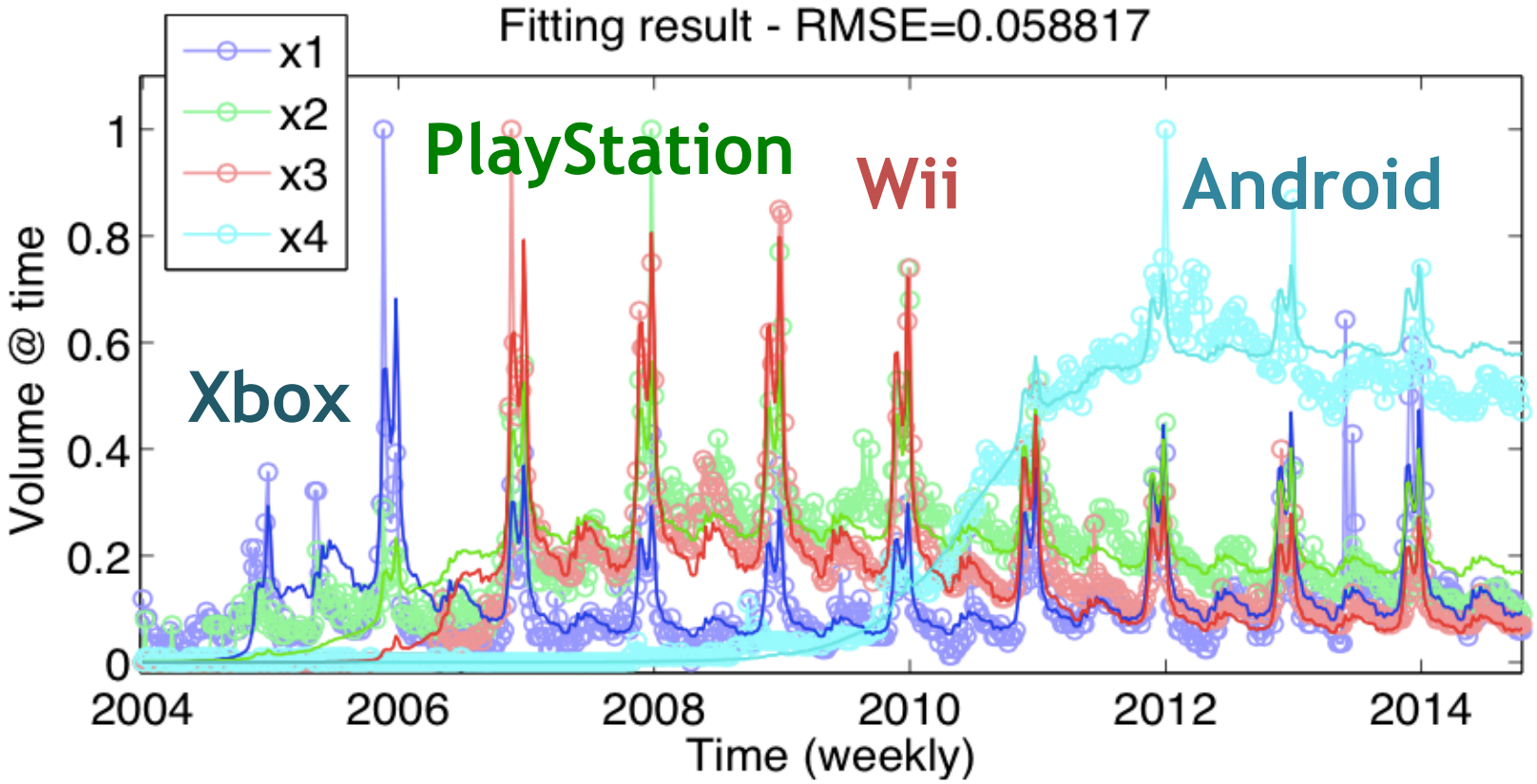


Q2. Webデータ

EcoWeb WWW'15

The Web as a Jungle !

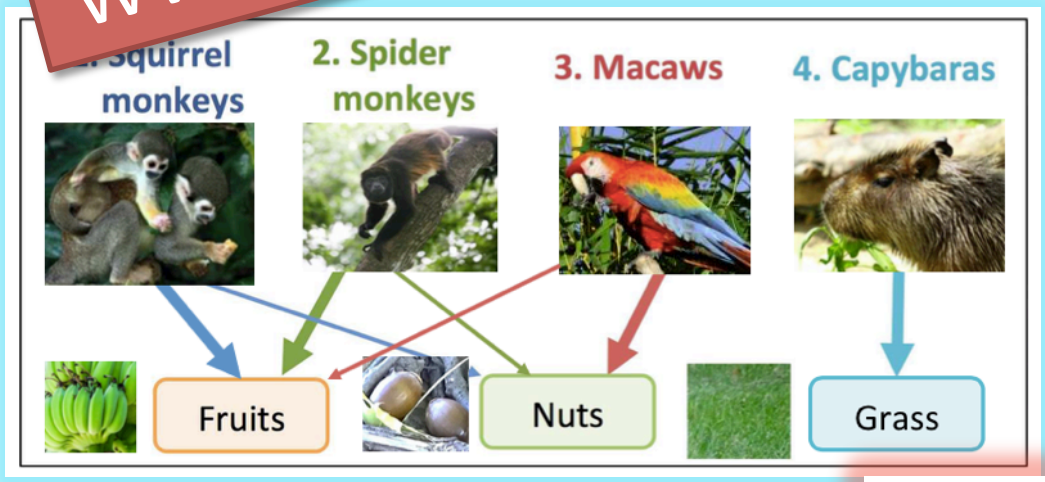
(*Google* Search volume)



Q2. Webデータ

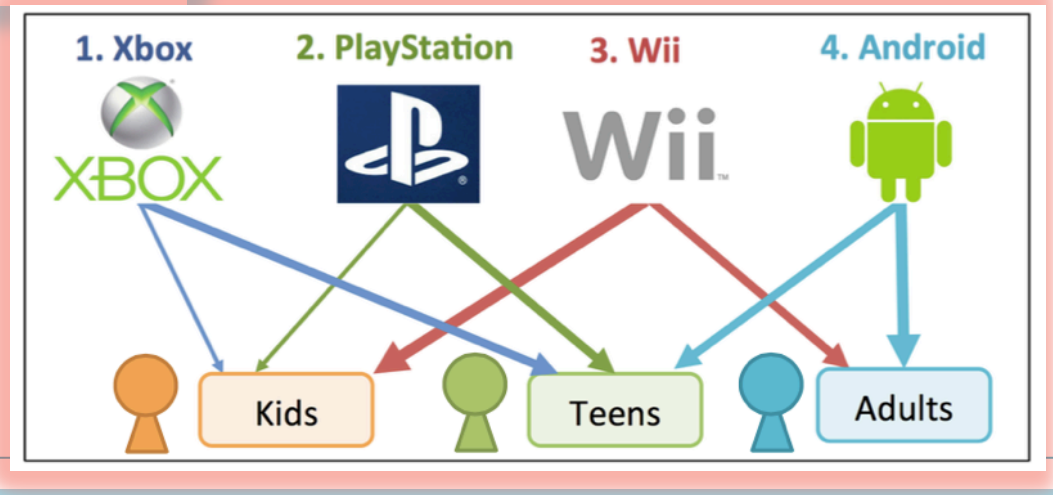
EcoWeb WWW'15

The Web as a Jungle !



Ecosystem in the **Jungle**

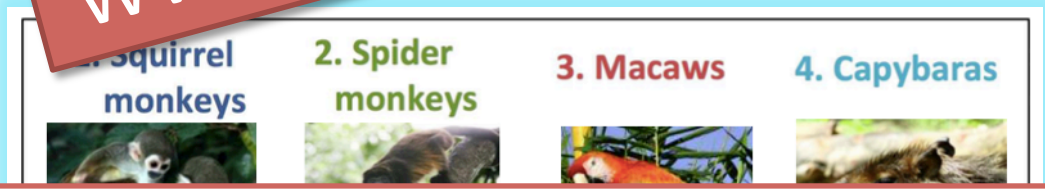
Ecosystem on the **Web**



Q2. Webデータ

EcoWeb WWW'15

The Web as a Jungle !



Ecosystem

--- Non-Linear equations ---

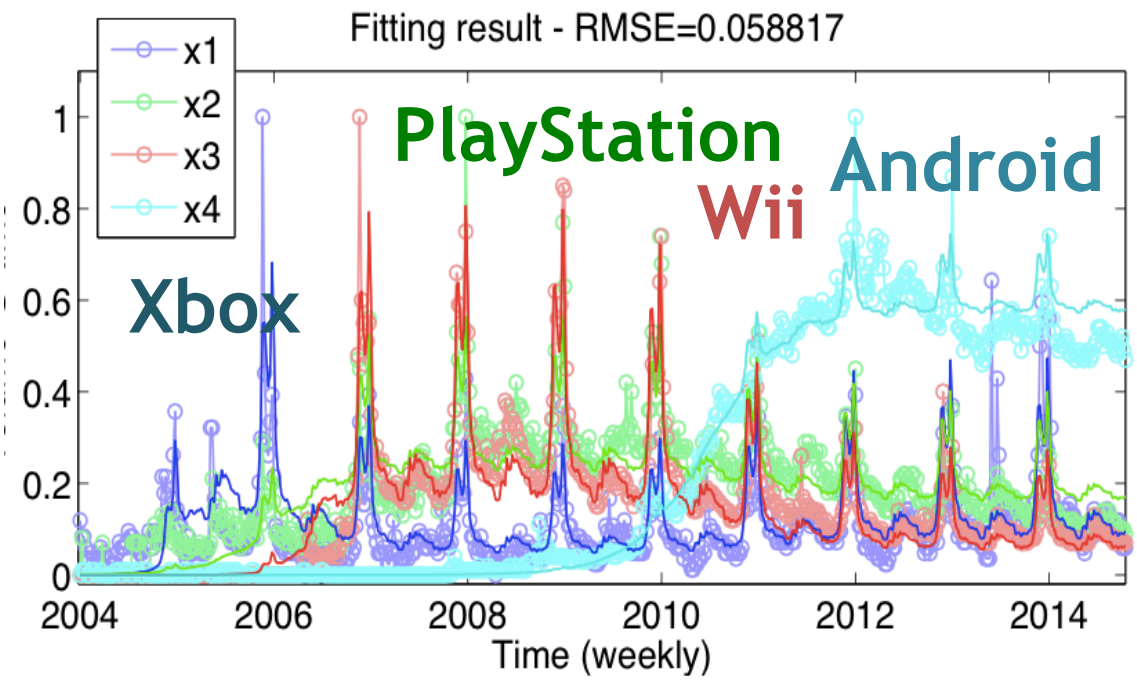
$$P_i(t+1) = P_i(t)\left[1 + \mathrm{r}_i\left(1 - \frac{\sum_{j=1}^{d} a_{ij} P_i(t)}{K_i}\right)\right]$$

Q2. Webデータ

EcoWeb WWW'15

The Web as a Jungle !

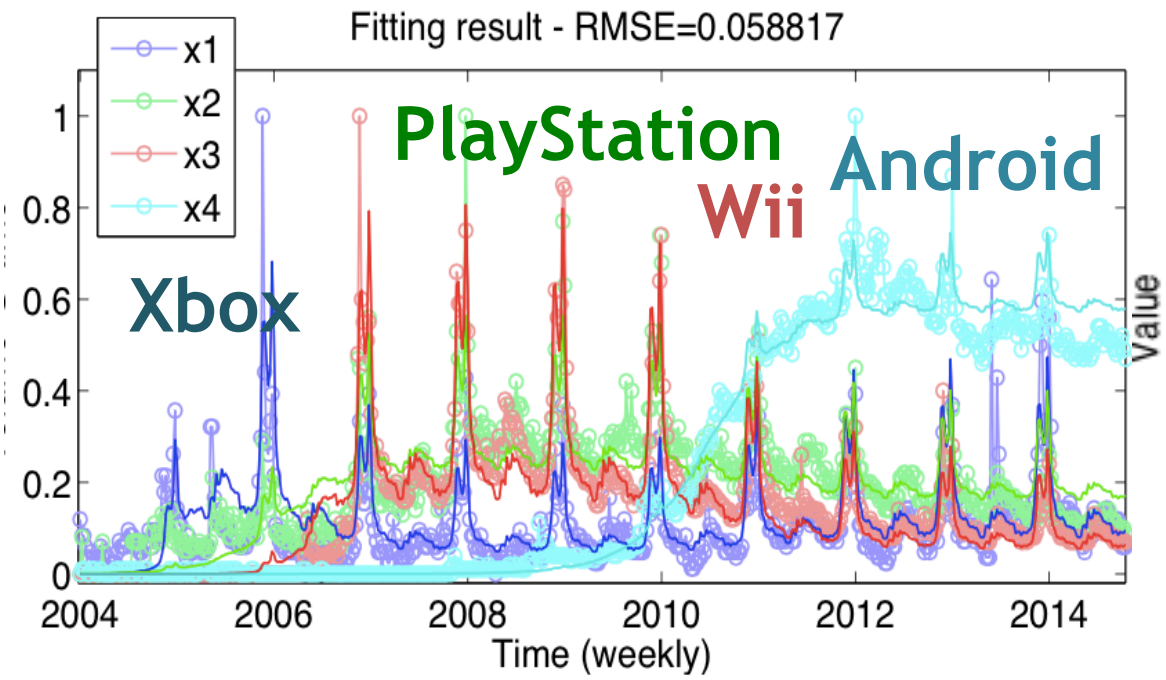
Interactions between keywords



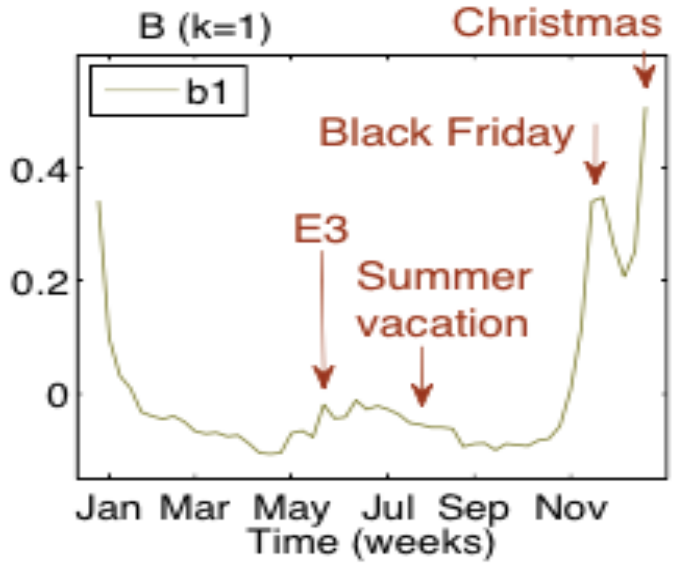
Q2. Webデータ

EcoWeb WWW'15

The Web as a Jungle !



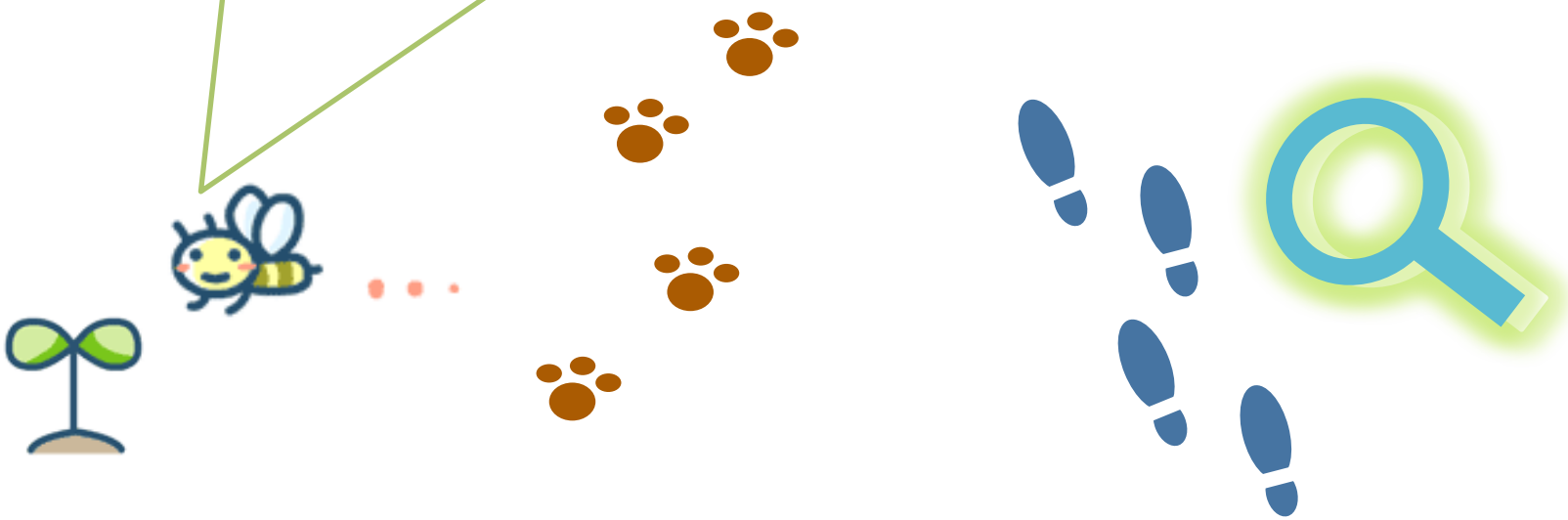
Seasonality



Q3. Webデータ

Q3

Webサイトの
アクセス解析をしよう！
明日は誰がどのページを開く？



Q3. Webデータ

TriMine KDD'12

Complex time-stamped events

{time, URL, user ID, access devices, http referrer,…}

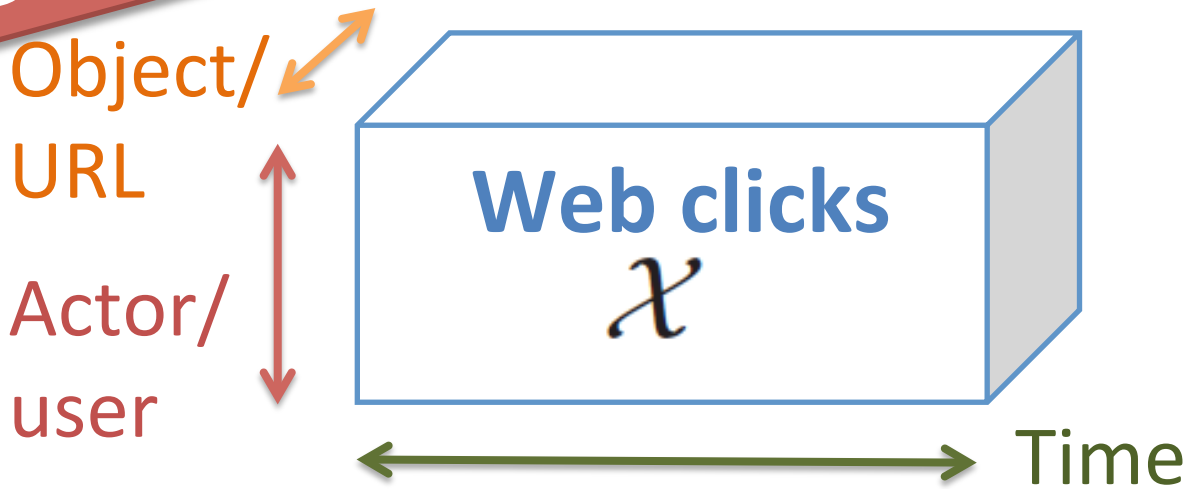
Forecast

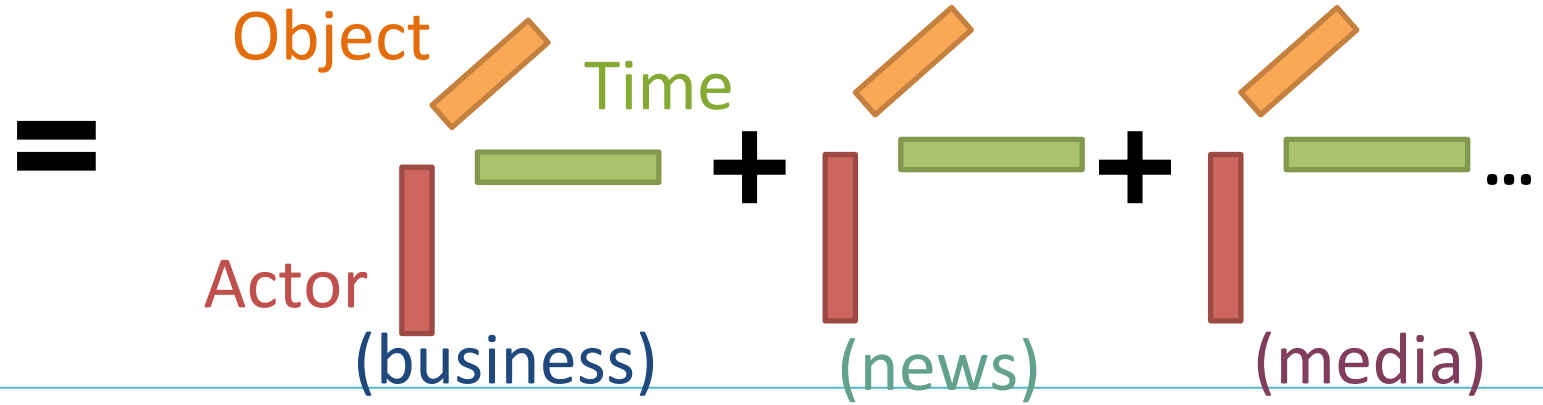
Timestamp	URL	User	Device
2012-08-01-12:00	CNN.com	Smith	iphone
12-08-02-15:00	YouTube.com	Brown	iphone
12-08-02-19:00	CNET.com	Smith	mac
2012-08-03-11:00	CNN.com	Johnson	ipad
…	…	…	…
2012-08-05-12:00	CNN.com	Smith	iphone
2012-08-05-19:00	CNET.com	Smith	iphone

Q3. Webデータ

TriMine KDD'12

Complex time-stamped events

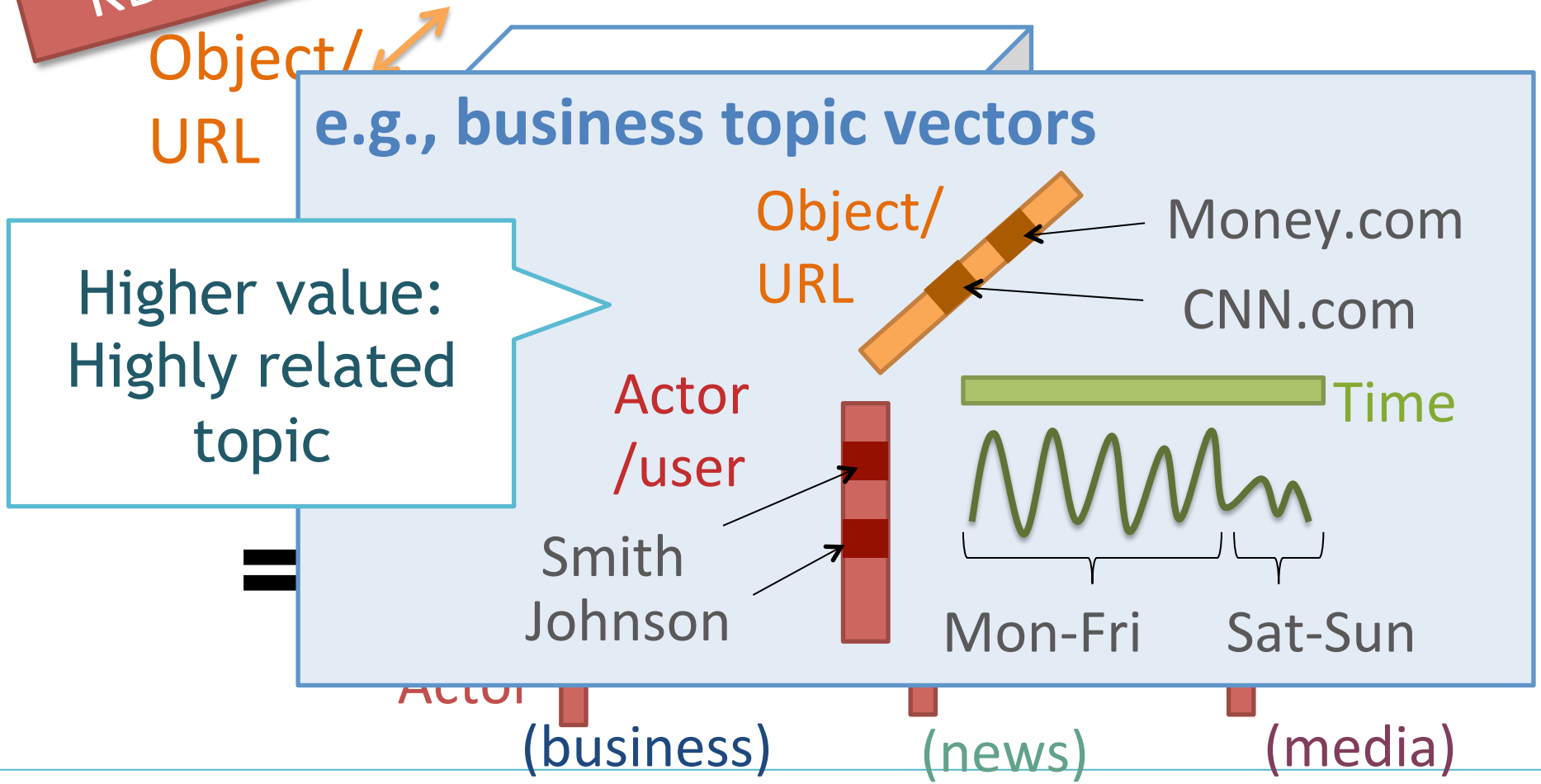




Q3. Webデータ

TriMine KDD'12

Complex time-stamped events



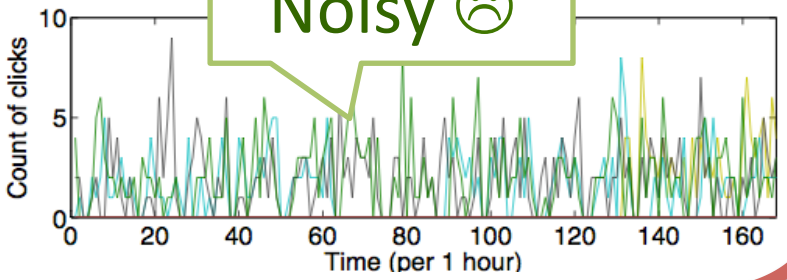
Q3. Webデータ

TriMine KDD'12

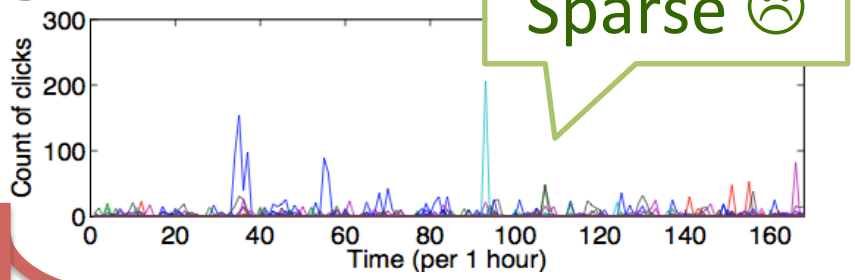
Complex time-stamped events

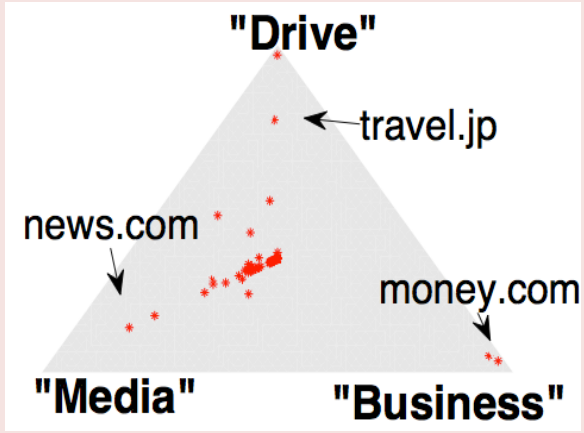
(Original data)

Noisy ☹

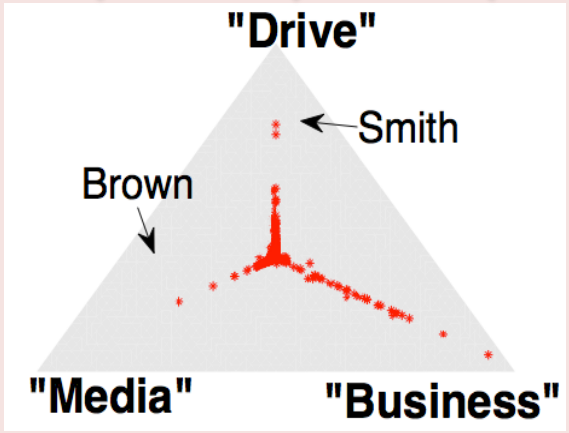


Sparse ☹

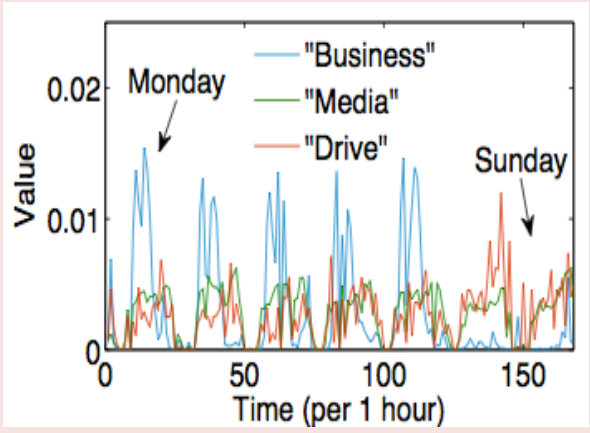




URL matrix



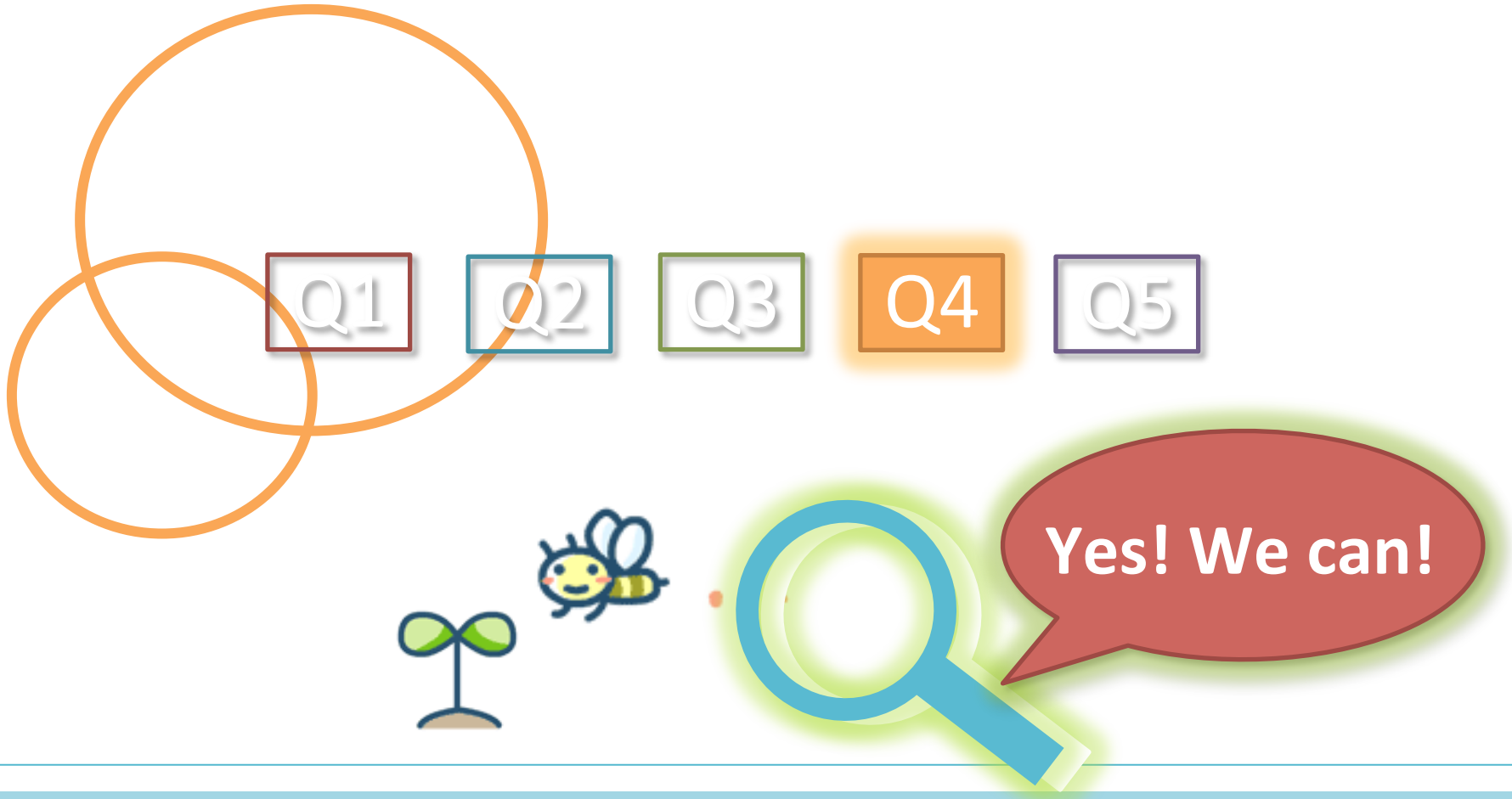
User matrix



Time matrix

例えば…

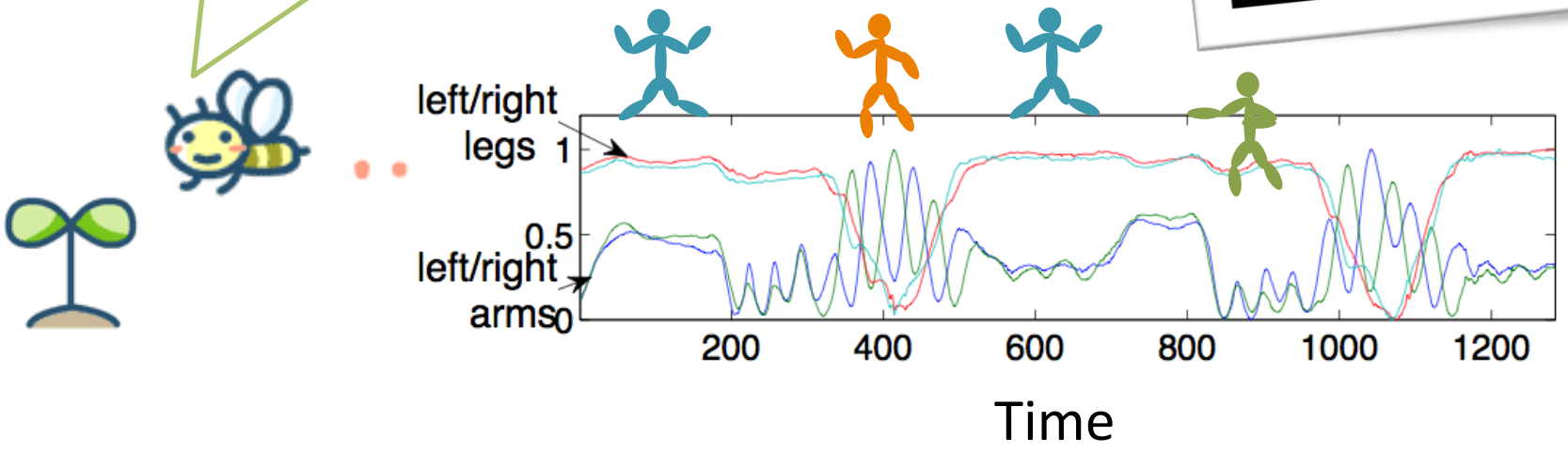
センサデータ



Q4. センサデータ

Q4

ダンスのモーション！
どこが切れ目かわかる？
ステップの種類は？



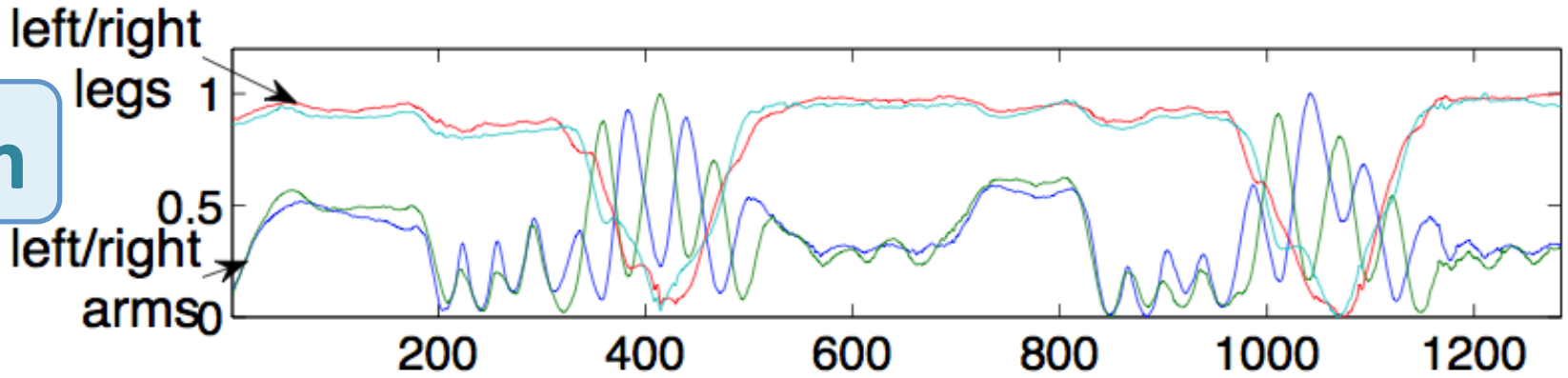
Q4. センサデータ

AutoPlait SIGMOD'14

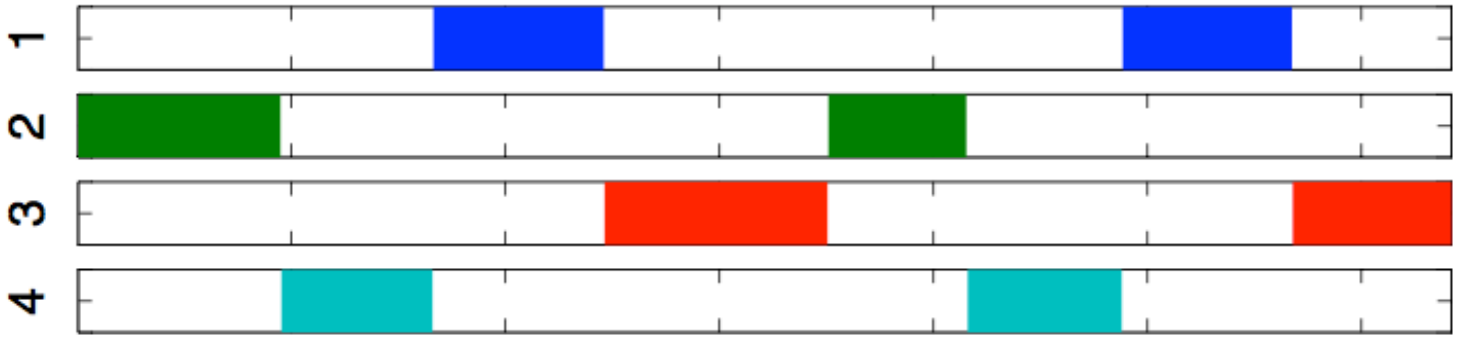
"Automatic" mining algorithm

Find: compact description of data X

Given



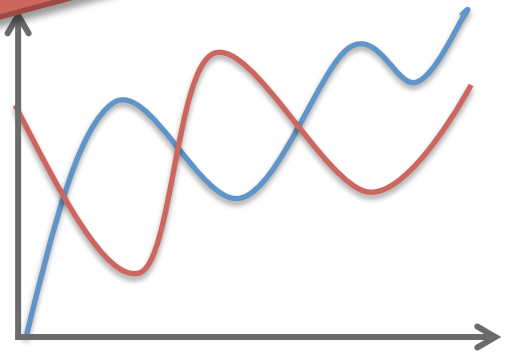
Find



Q4. センサデータ

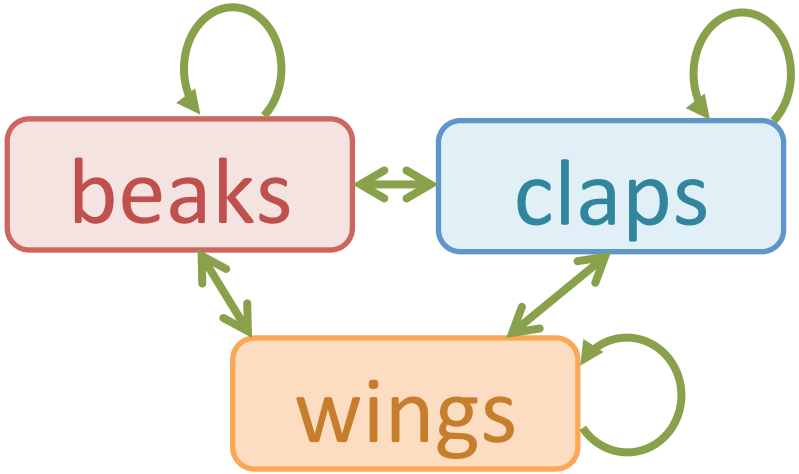
AutoPlait SIGMOD'14

"Automatic" mining algorithm



Sequences

Model



Regimes

Idea (1): Multi-level chain model

– HMM-based probabilistic model

– with "across-regime" transitions

Q4. センサデータ

AutoPlait SIGMOD'14

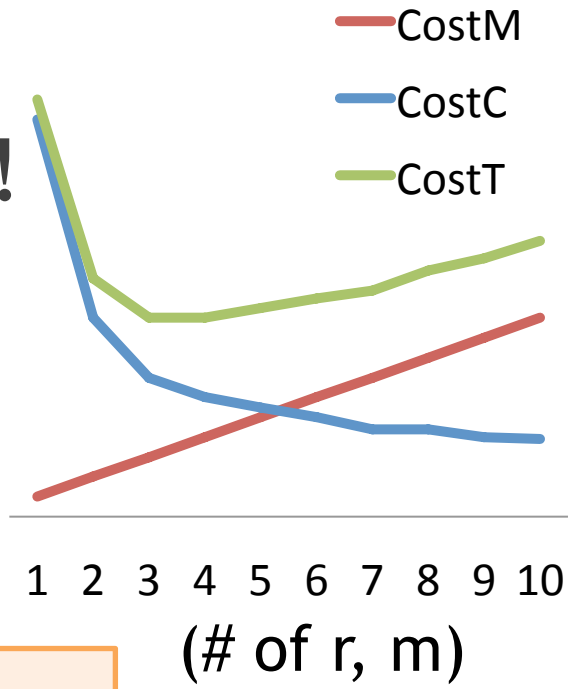
"Automatic" mining algorithm

Idea(2): Minimize encoding cost!

$$\min \left(\text{Cost}_M(M) + \text{Cost}_c(X \mid M) \right)$$

Model cost

Coding cost

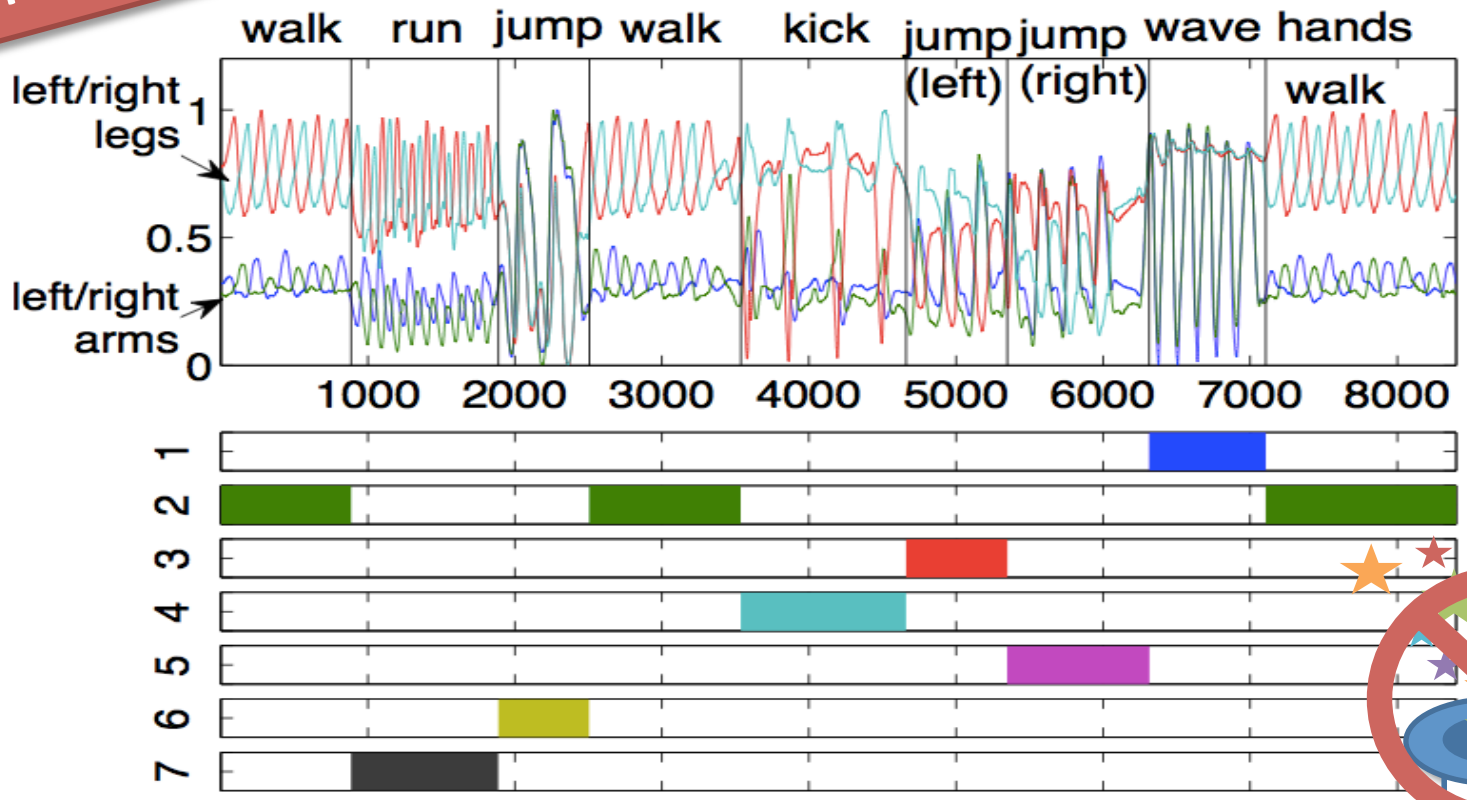


Good compression ⇄ Good description

Q4. センサデータ

AutoPlait SIGMOD'14

"Automatic" mining algorithm



AutoPlait (**NO** magic numbers)

例えば…

医療データ

Q5. 医療データ

Q5

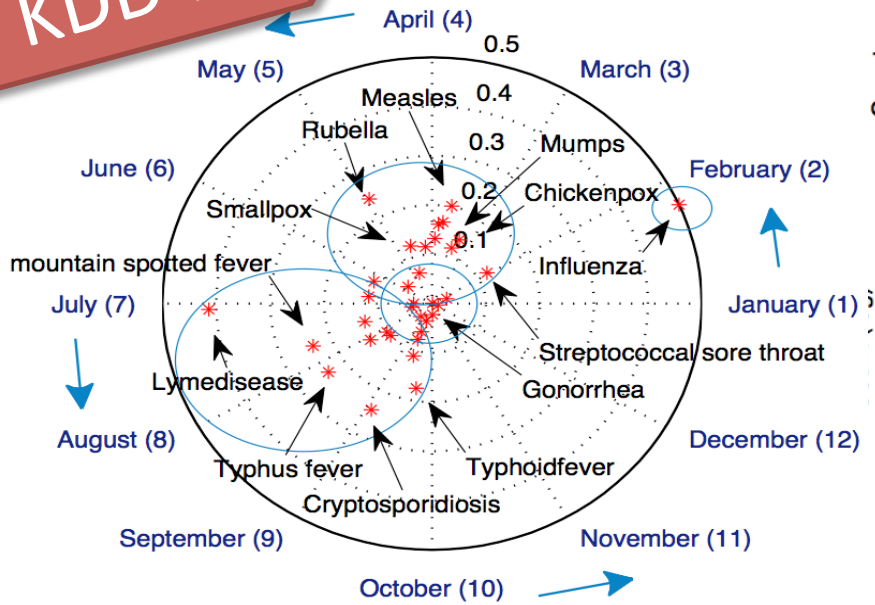
来年のインフルエンザ
どの地域が流行しそう？

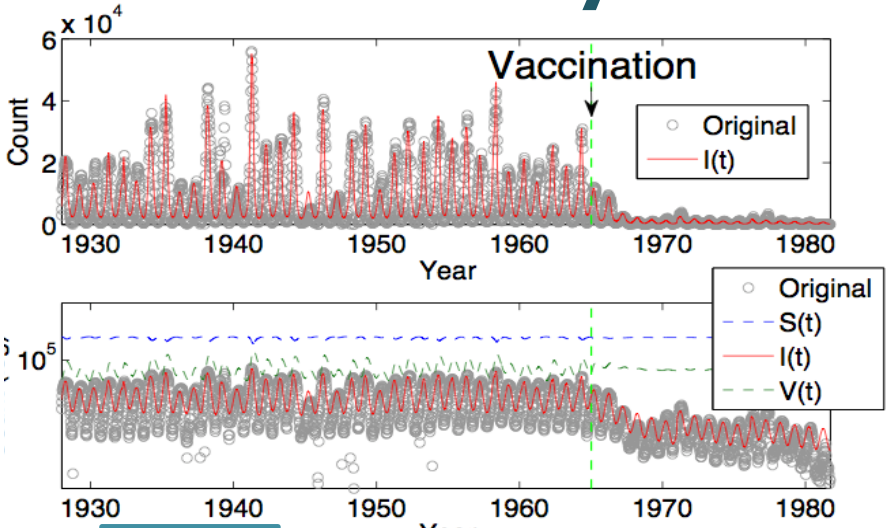
Q5. 医療データ

Q1 Q2 Q3 Q4 Q5

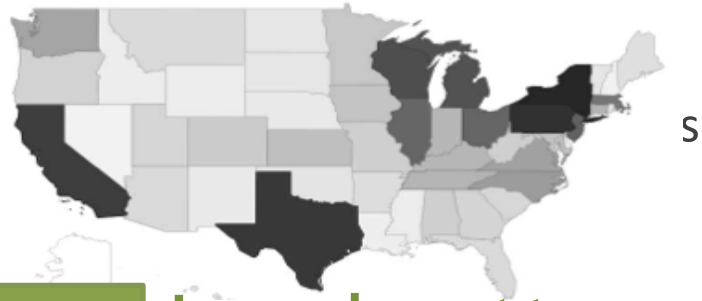
FUNNEL
KDD'14

Non-Linear tensor analysis





P2 Vaccination



P3 Local patterns

P1 Seasonality

P4 External shocks

P5 Mistakes, errors

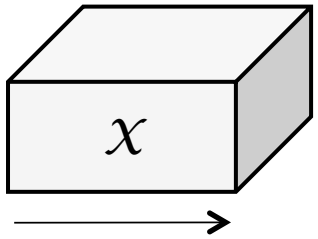
Q5. 医療データ

FUNNEL
KDD'14

Non-Linear tensor analysis

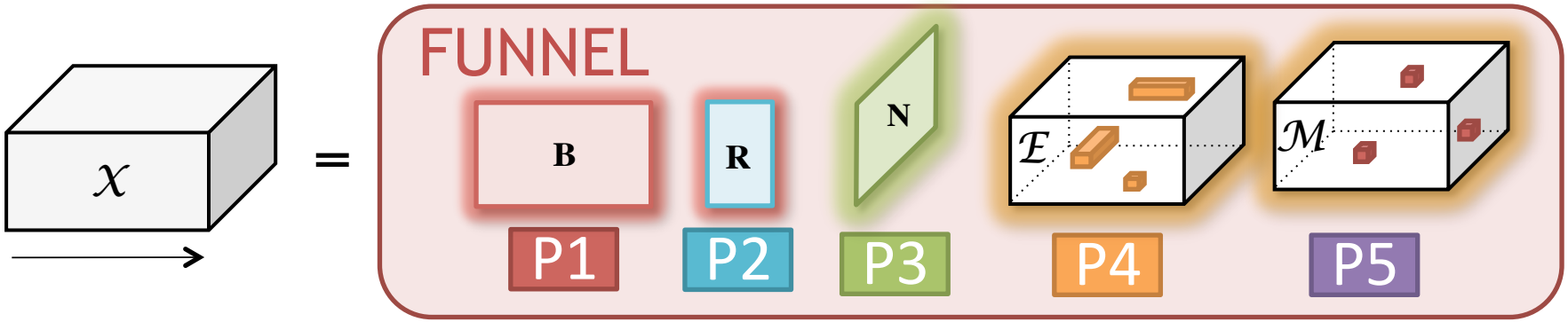
Given:

Tensor $\mathcal{X}$ (disease x state x time)



Find:

Compact description of $\mathcal{X}$, *"automatically"*



Q5. 医療データ

FUNNEL
KDD'14

Non-Linear tensor analysis

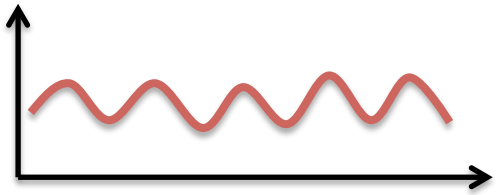
Given:

Ten state x

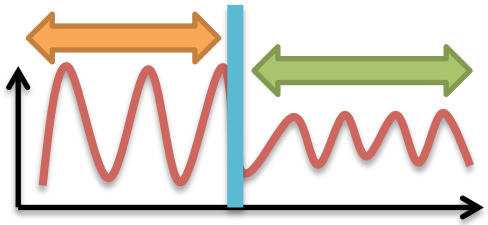
Fin

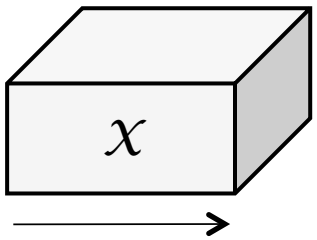
Compact desc ion of $\mathcal{X}$, "tomatically"

Seasonality



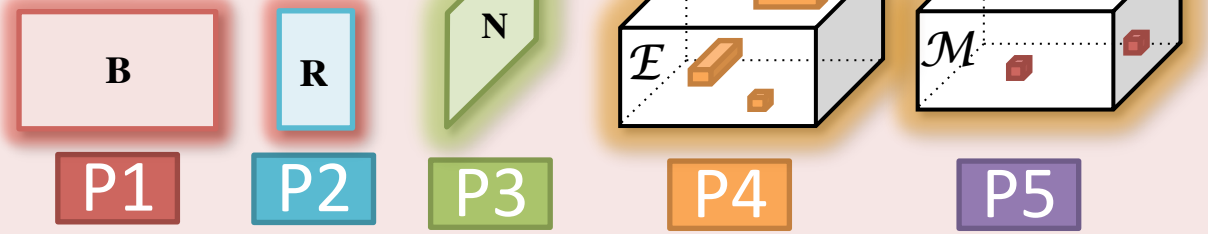
Discontinuities





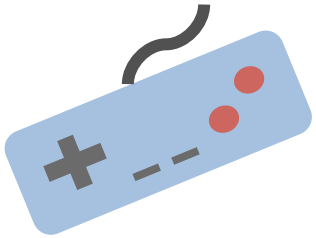
=

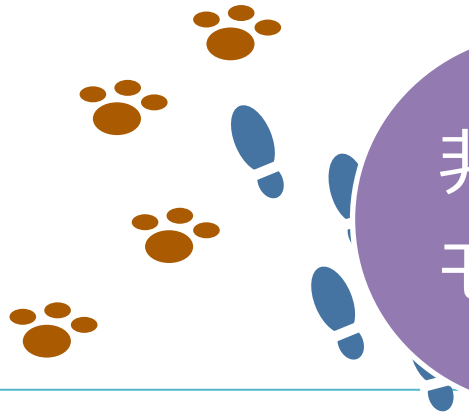
FUNNEL

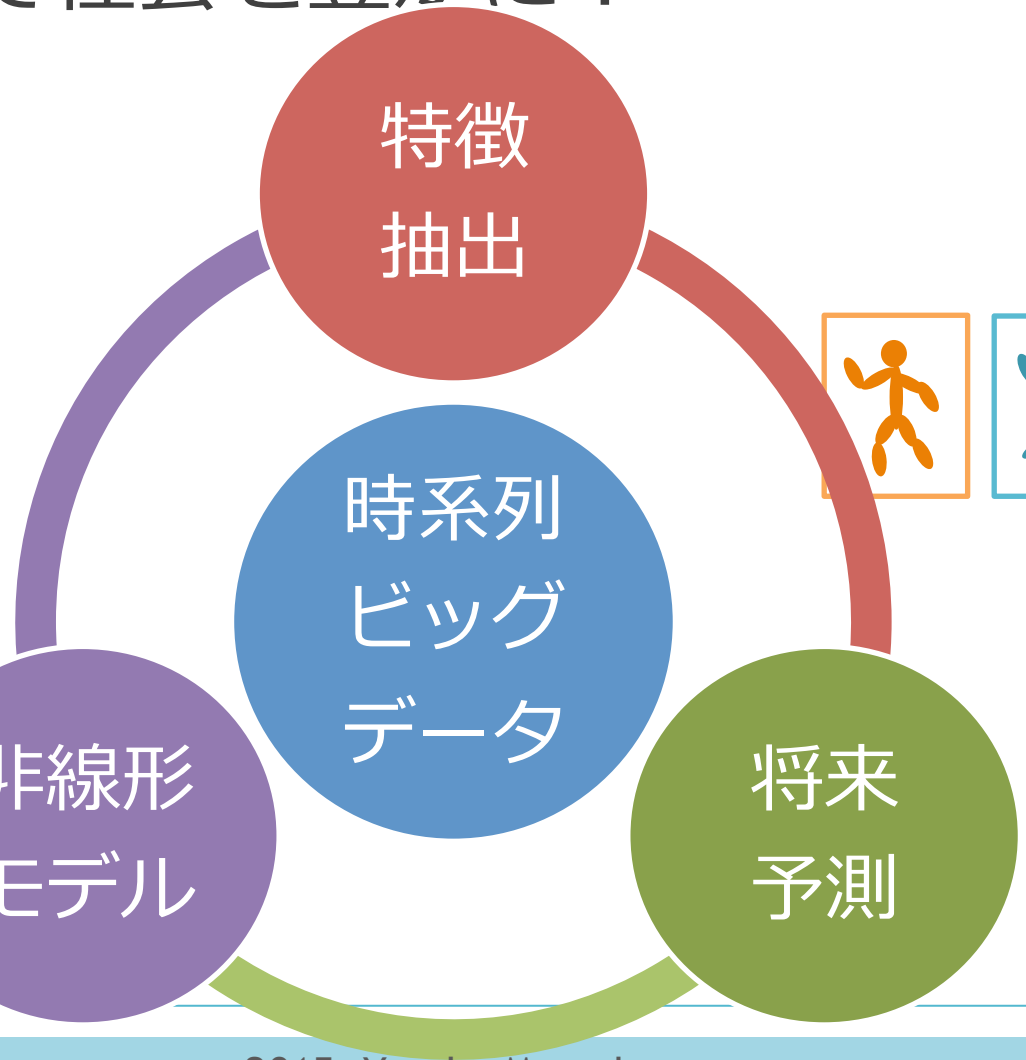


目標

時系列解析で社会を豊かに！







AutoPlait
SIGMOD'14



AutoPlait: Automatic Mining of Co-evolving Time Sequences

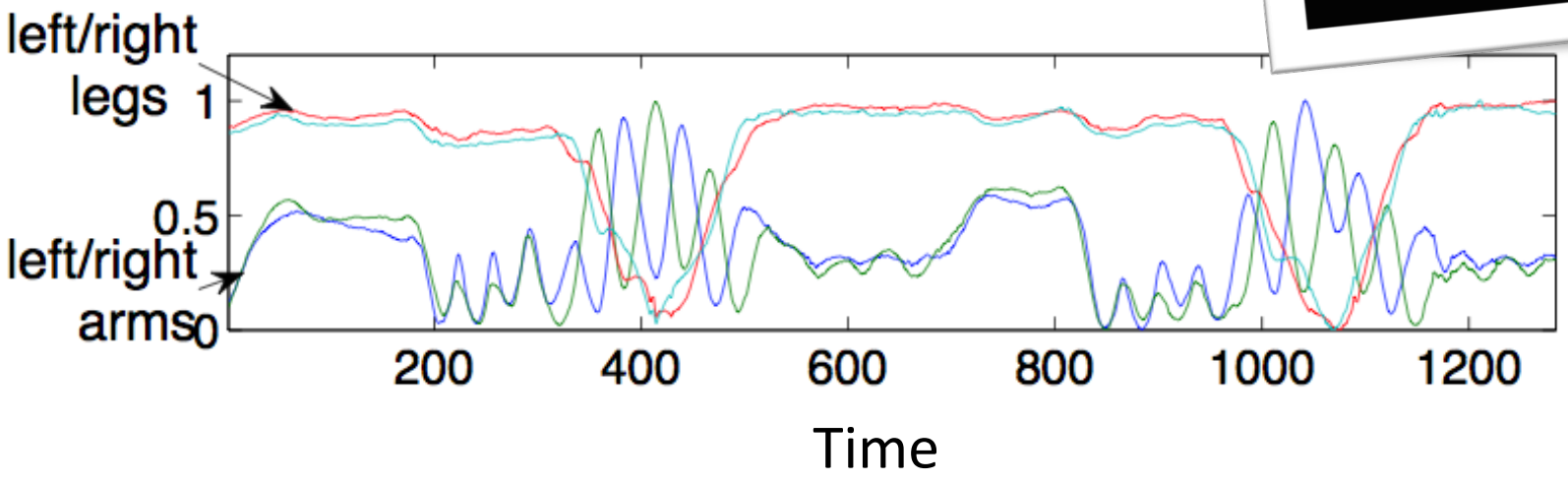
Yasuko Matsubara (Kumamoto University)
Yasushi Sakurai (Kumamoto University)
Christos Faloutsos (CMU)

Motivation

Given: co-evolving time-series

– e.g., MoCap (leg/arm sensors)

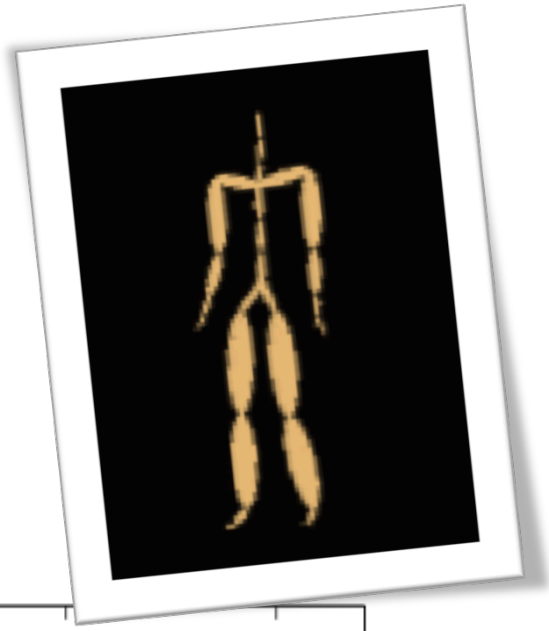
"Chicken dance"



Motivation

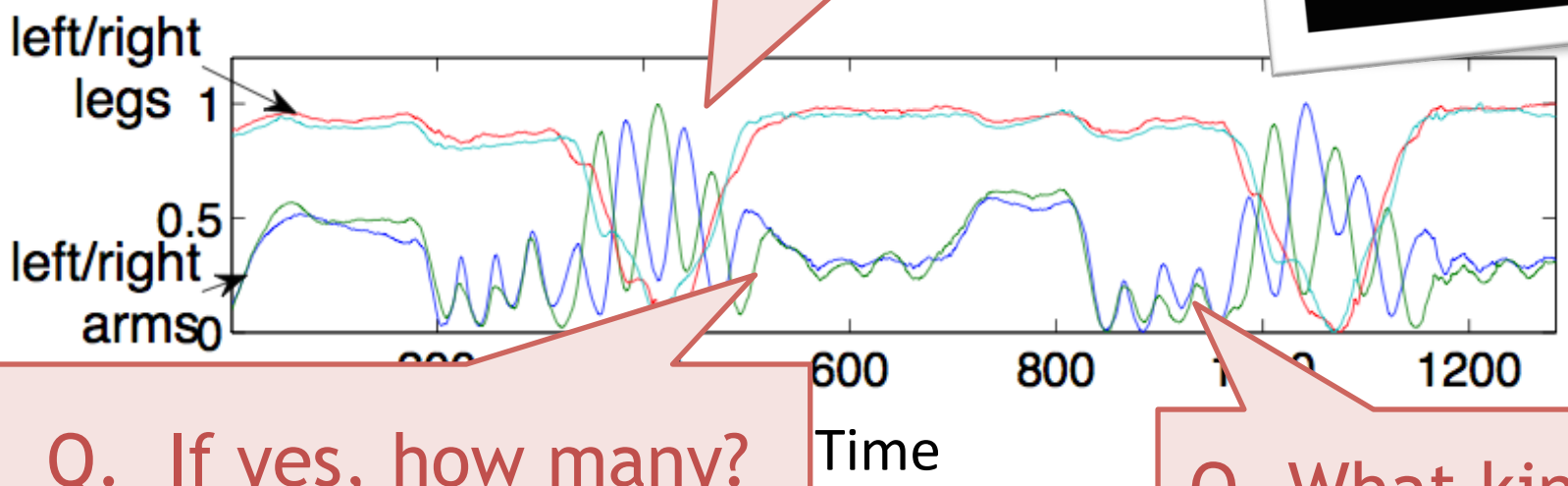
Given: co-evolving time-series

– e.g., MoCap (leg/arm sensors)



"Chicken dance"

Q. Any distinct patterns?



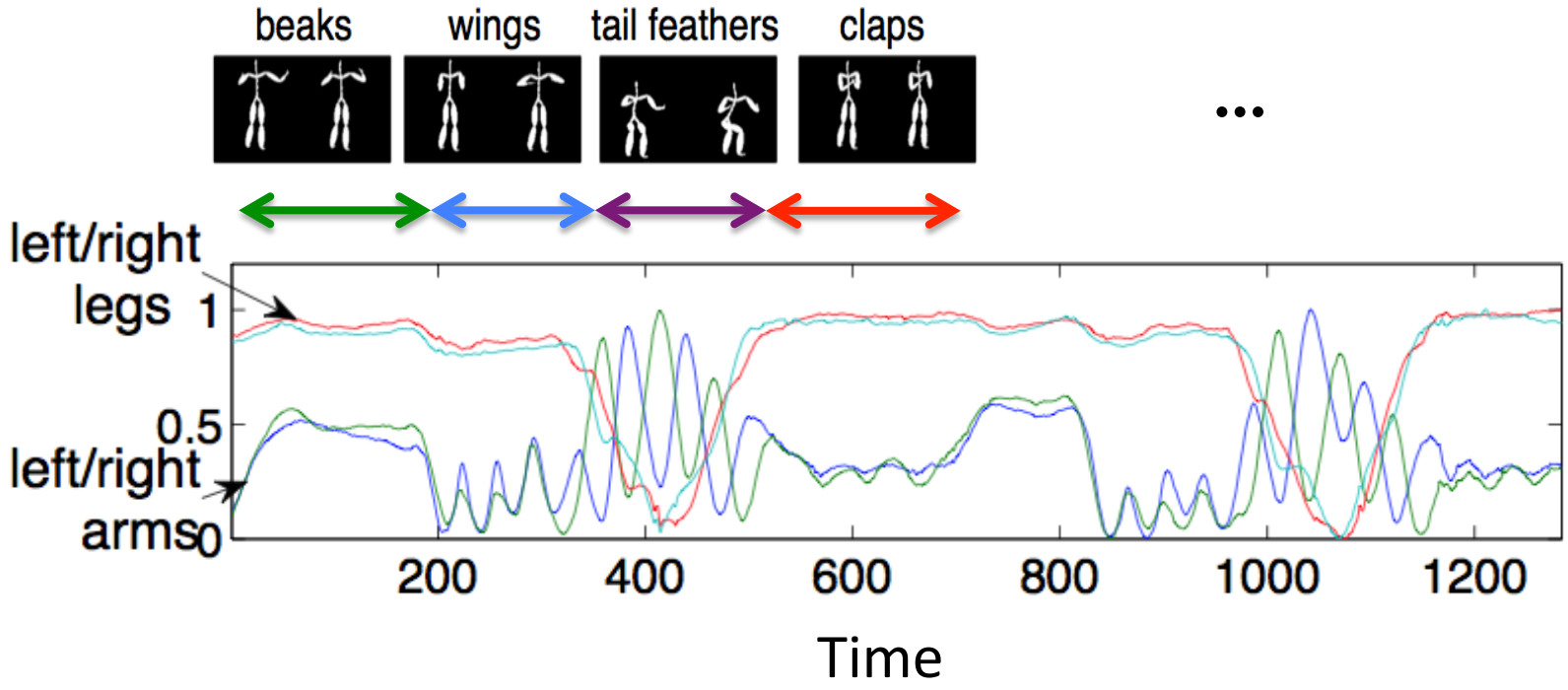
Q. If yes, how many?

Q. What kind?

Motivation

Challenges: co-evolving sequences

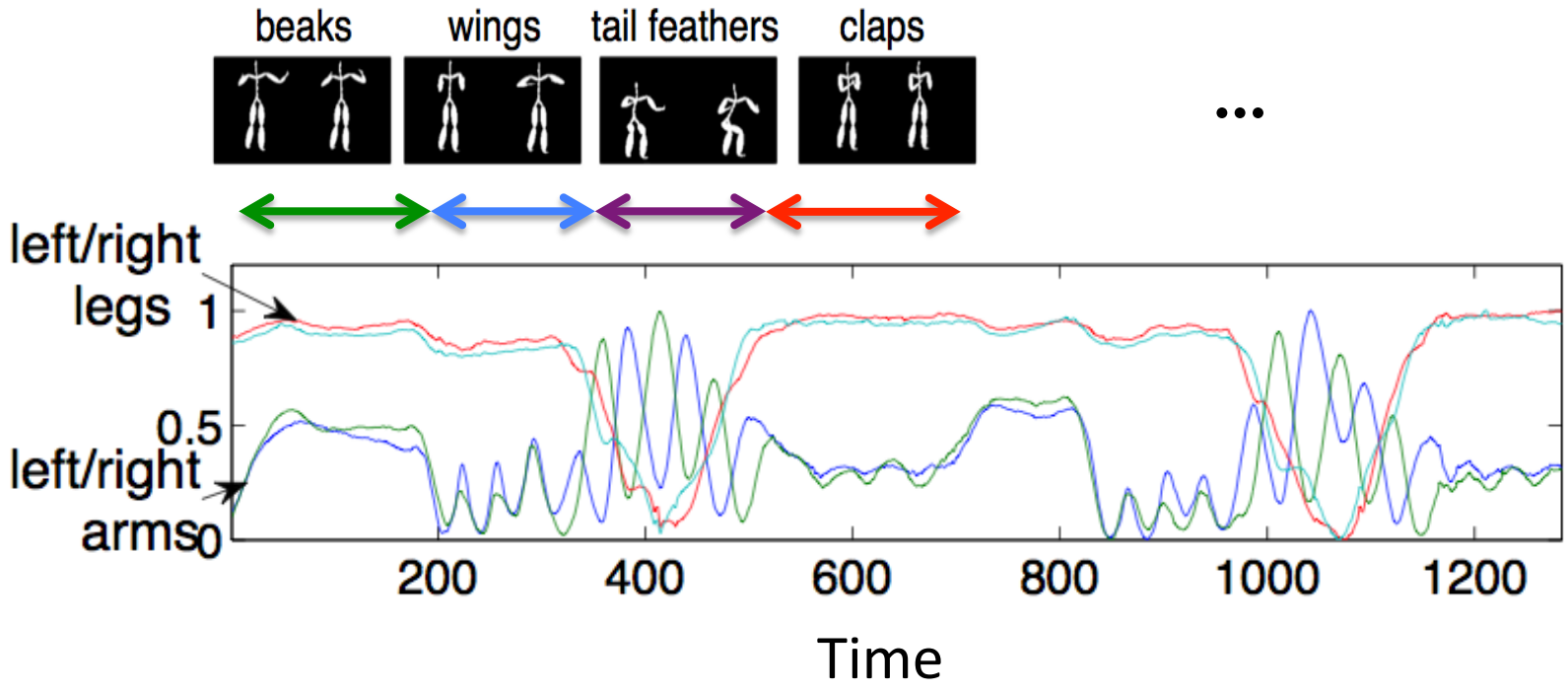
- Unknown # of patterns (e.g., beaks)
- Different durations



Motivation

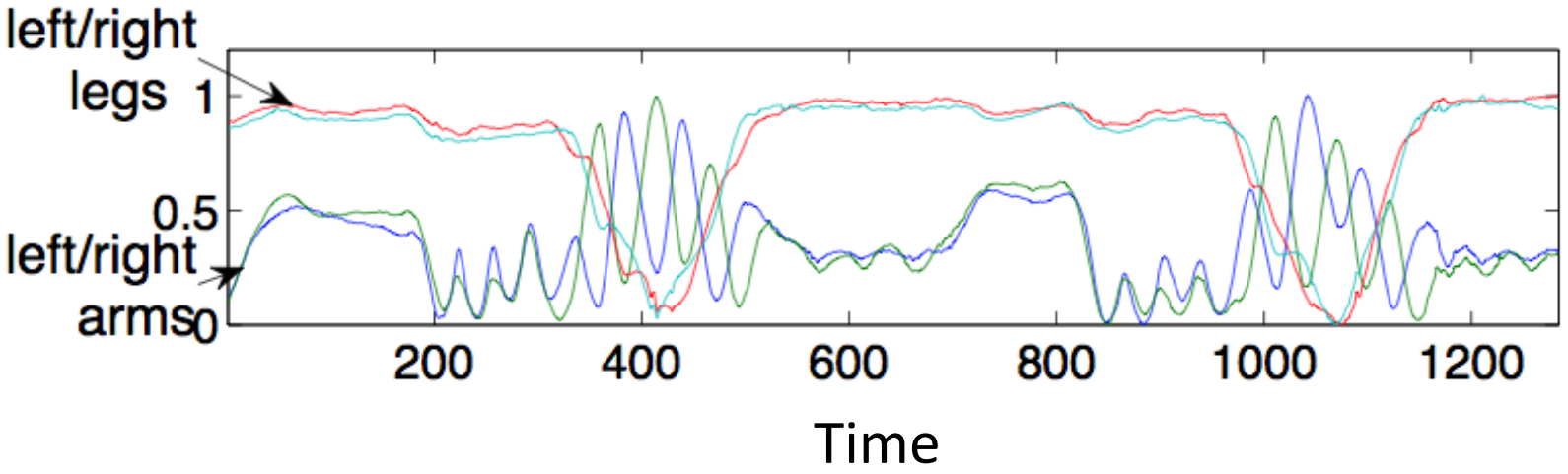
Challenges: co-evolving sequences

Q. Can we summarize it ***automatically*** ?



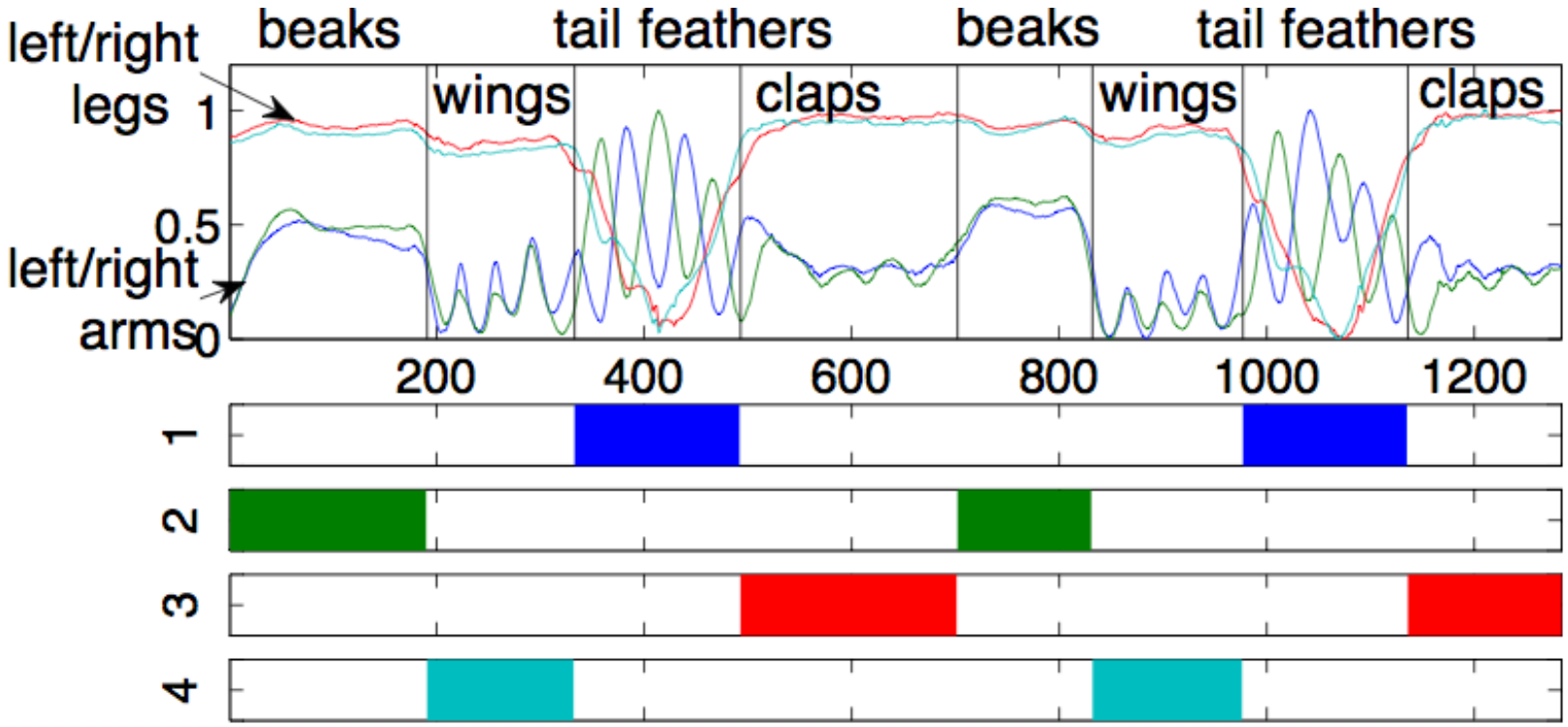
Motivation

Goal: find patterns that agree with human intuition



Motivation

Goal: find patterns that agree with human intuition



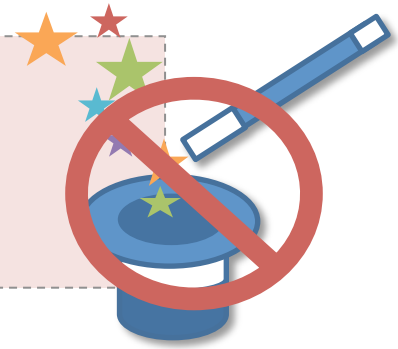
AutoPlait: "**fully-automatic**" mining algorithm

Importance of “fully-automatic”

No magic numbers! ... because,

Manual

- sensitive to the parameter tuning
- long tuning steps (hours, days, …)



Automatic (no magic numbers)

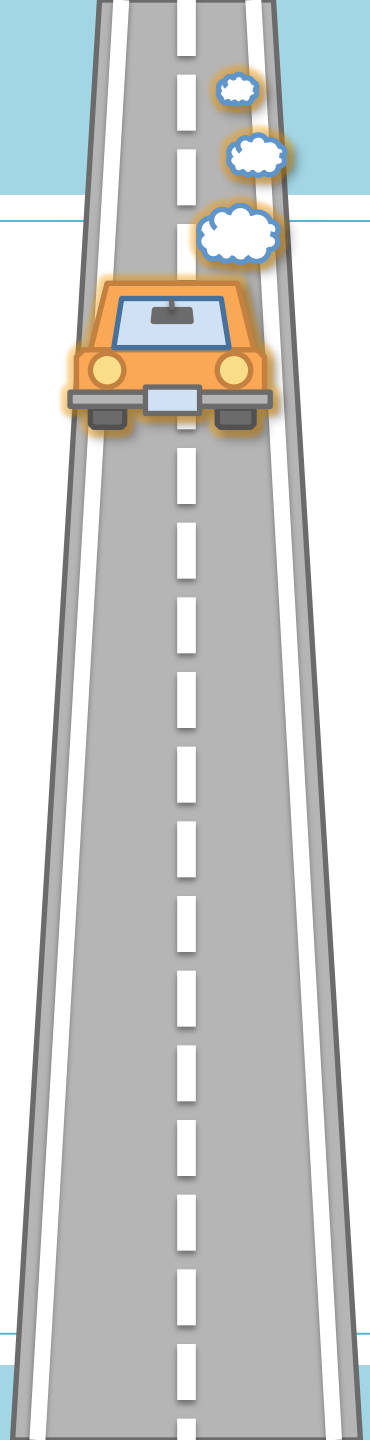
- no expert tuning required

Big data mining:

-> we cannot afford human intervention!!

Outline

- Motivation
- Problem definition
- Compression & summarization
- Algorithms
- Experiments
- AutoPlait at work
- Conclusions



Problem definition

Key concepts

- Bundle: X — given
- Segment: S — hidden
- Regime: Θ — hidden
- Segment-membership: F — hidden

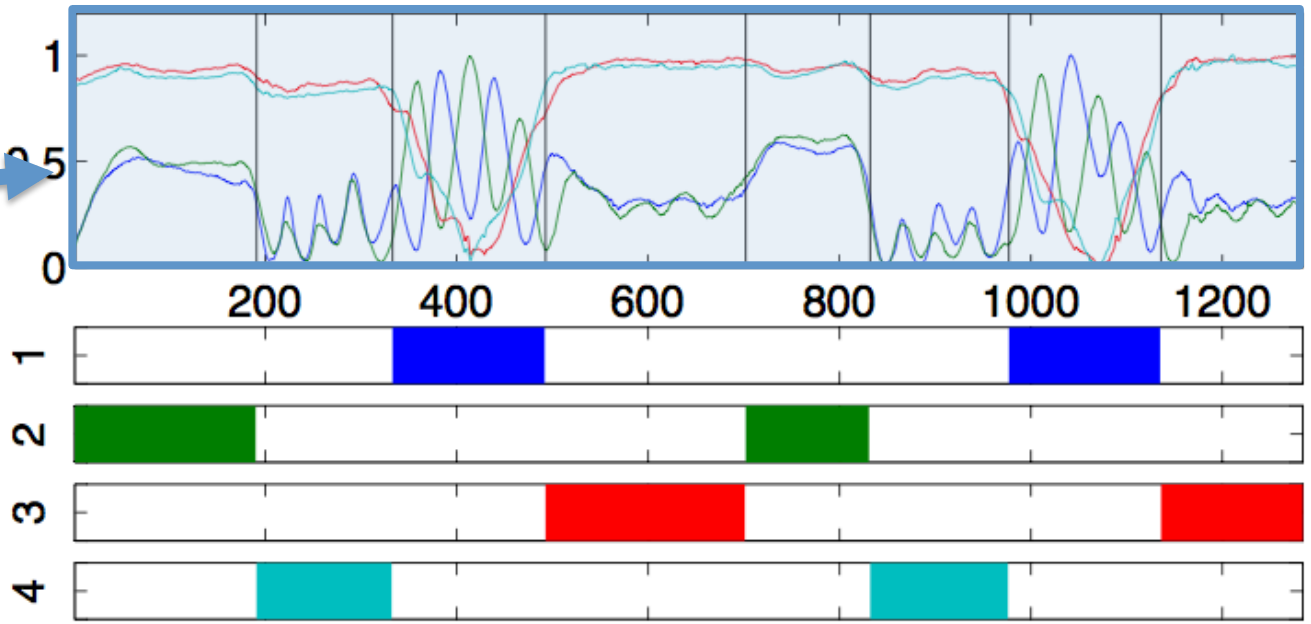
Problem definition

- Bundle : set of d co-evolving sequences

given

$$\underset{d \times n}{X} = \{x_1, \ldots, x_n\}$$

Bundle X (d=4)



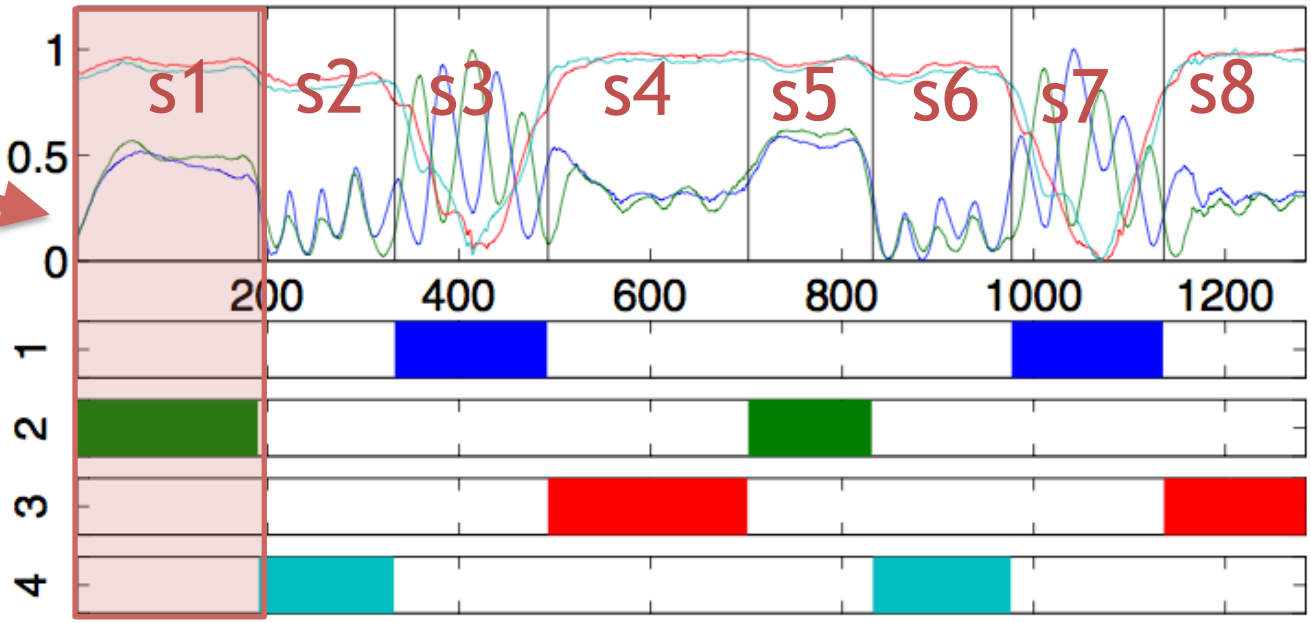
Problem definition

- Segment: convert X -> m segments, S

hidden

$$S = \{s_1, \ldots, s_m\}$$

Segment (m=8)



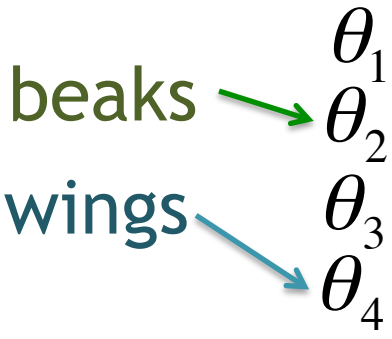
Problem definition

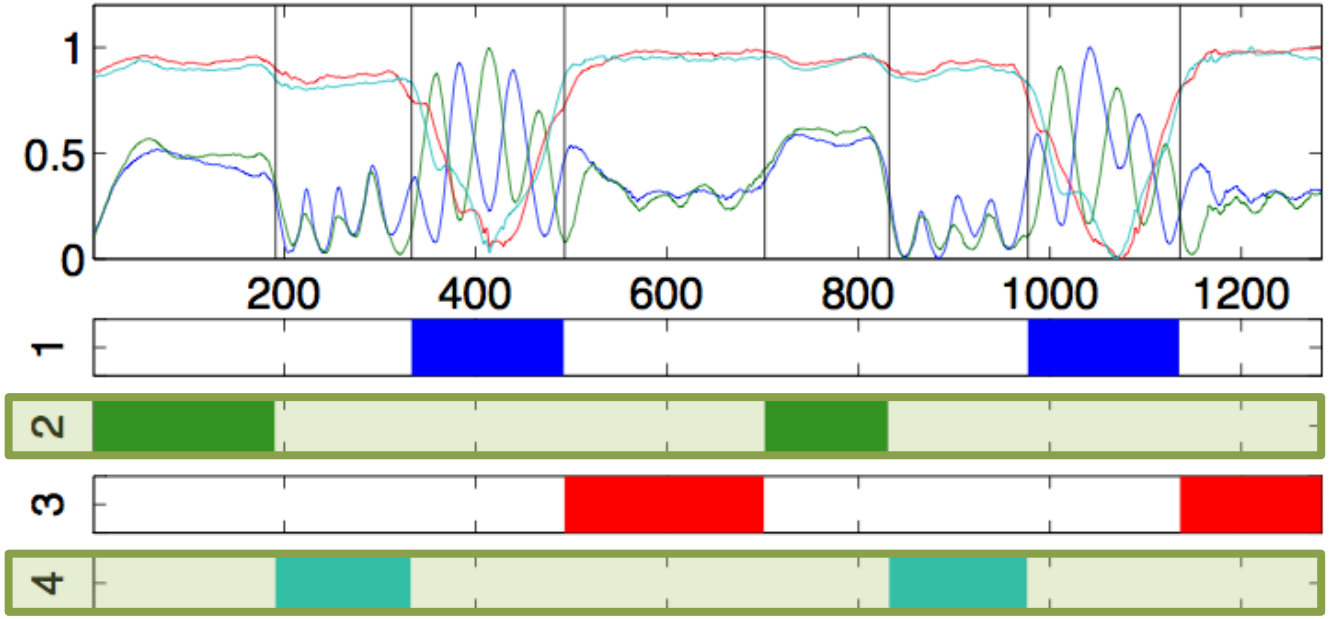
- Regime: segment groups: $\Theta = \{\theta_1, \theta_2, \ldots \theta_r, \Delta_{r \times r}\}$

hidden

θ_r : model of regime r

Regimes (r=4)



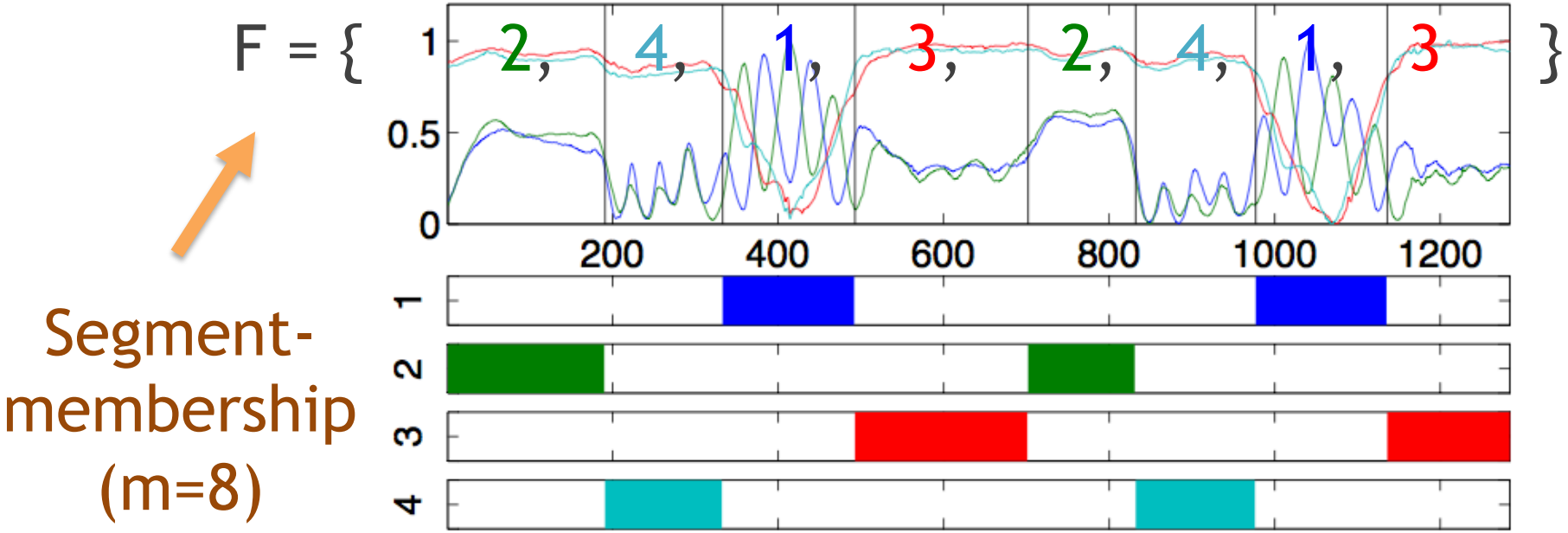


Problem definition

- Segment-membership: assignment

hidden

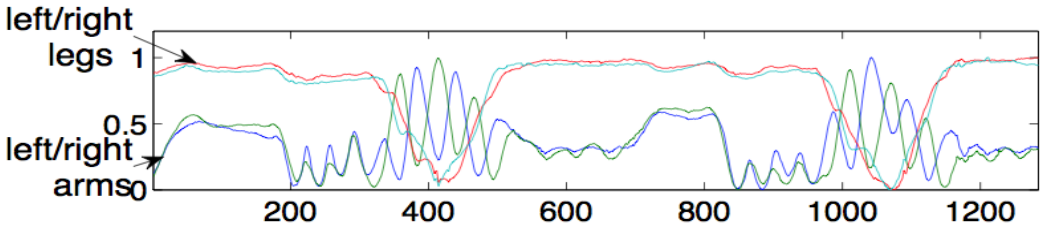
$$F = \{f_1, \ldots, f_m\}$$



Problem definition

- Given: bundle X

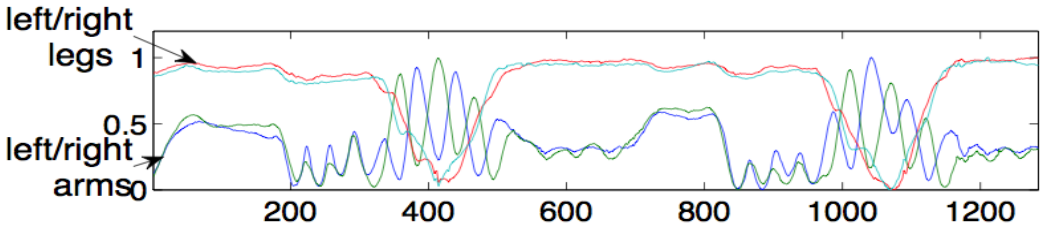
$$X = \{x_1, \ldots, x_n\}$$



Problem definition

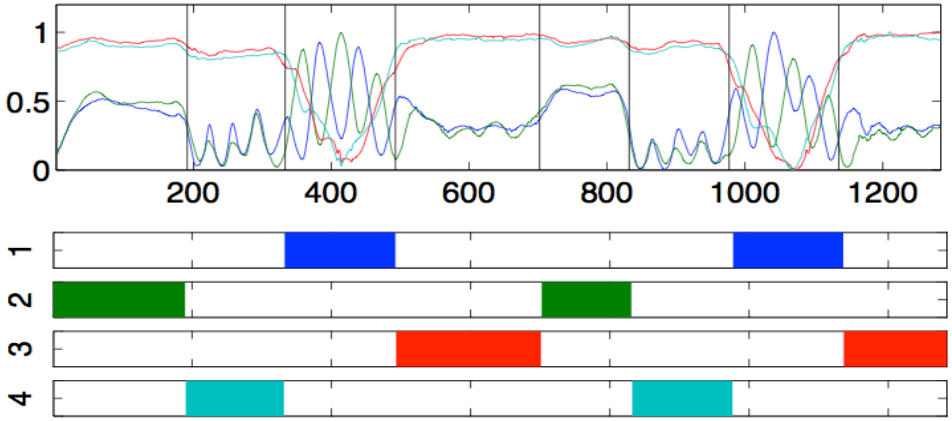
- Given: bundle X

$$X = \{x_1, \ldots, x_n\}$$



- Find: compact description C of X

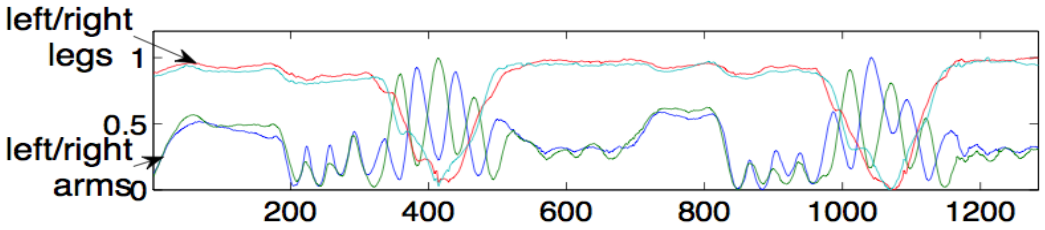
$$C = \{m, r, S, \Theta, F\}$$



Problem definition

- Given: bundle X

$$X = \{x_1, \ldots, x_n\}$$



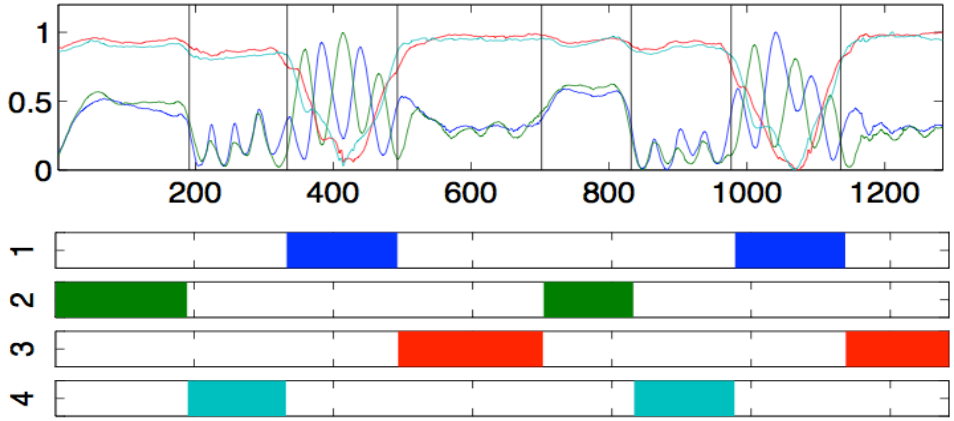
- Find: compact description C of X

$$C = \{m, r, S, \Theta, F\}$$

m segments

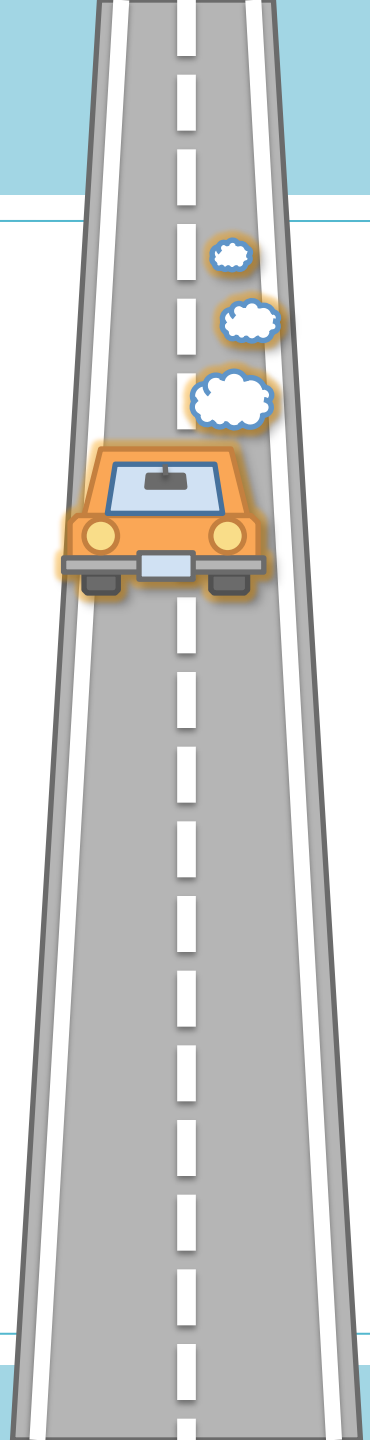
r regimes

Segment-membership



Outline

- Motivation
- Problem definition
- Compression & summarization
- Algorithms
- Experiments
- AutoPlait at work
- Conclusions



Main ideas

Goal: compact description of X

$$C = \{m, r, S, \Theta, F\}$$

without user intervention!!

Challenges:

Q1. How to generate 'informative' regimes ?

Q2. How to decide # of regimes/segments ?

Main ideas

Goal: compact description of X

$$C = \{m, r, S, \Theta, F\}$$

without user intervention!!

Challenges:

Q1. How to generate 'informative' regimes ?

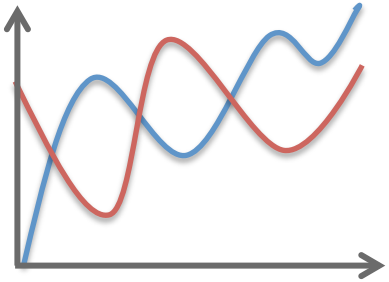
Idea (1): Multi-level chain model

Q2. How to decide # of regimes/segments ?

Idea (2): Model description cost

Idea (1): MLCM: multi-level chain model

Q1. How to generate ‘informative’ regimes ?



Sequences

Model

beaks

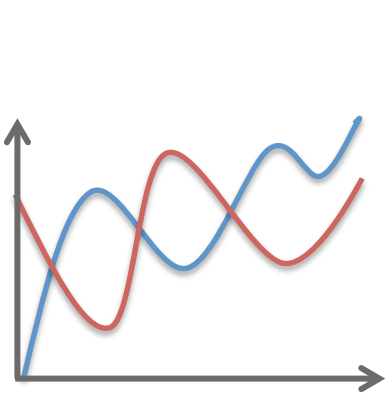
claps

wings

Regimes

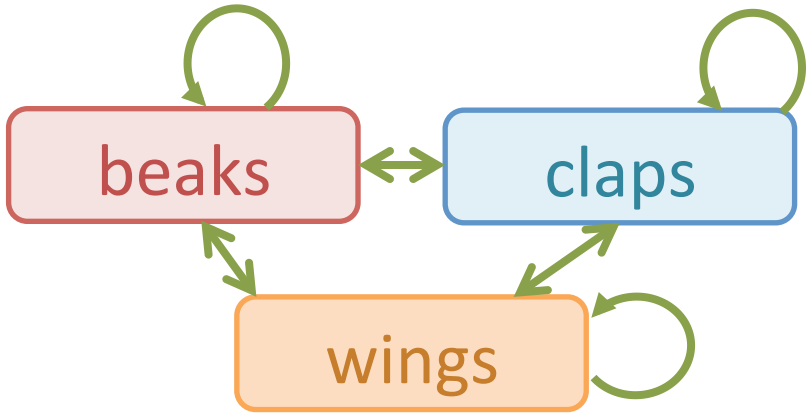
Idea (1): MLCM: multi-level chain model

Q1. How to generate ‘informative’ regimes ?



Sequences

Model



Regimes

Idea (1): Multi-level chain model

–HMM-based probabilistic model

–with “across-regime” transitions

Idea (1): MLCM: multi-level chain model

Details

$$\Theta = \{\theta_1, \theta_2, \ldots \theta_r, \Delta_{r \times r}\} \qquad (\theta_i = \{\pi, A, B\})$$

r regimes (HMMs)

across-regime transition prob.

Single HMM parameters

Idea (1): MLCM: multi-level chain model

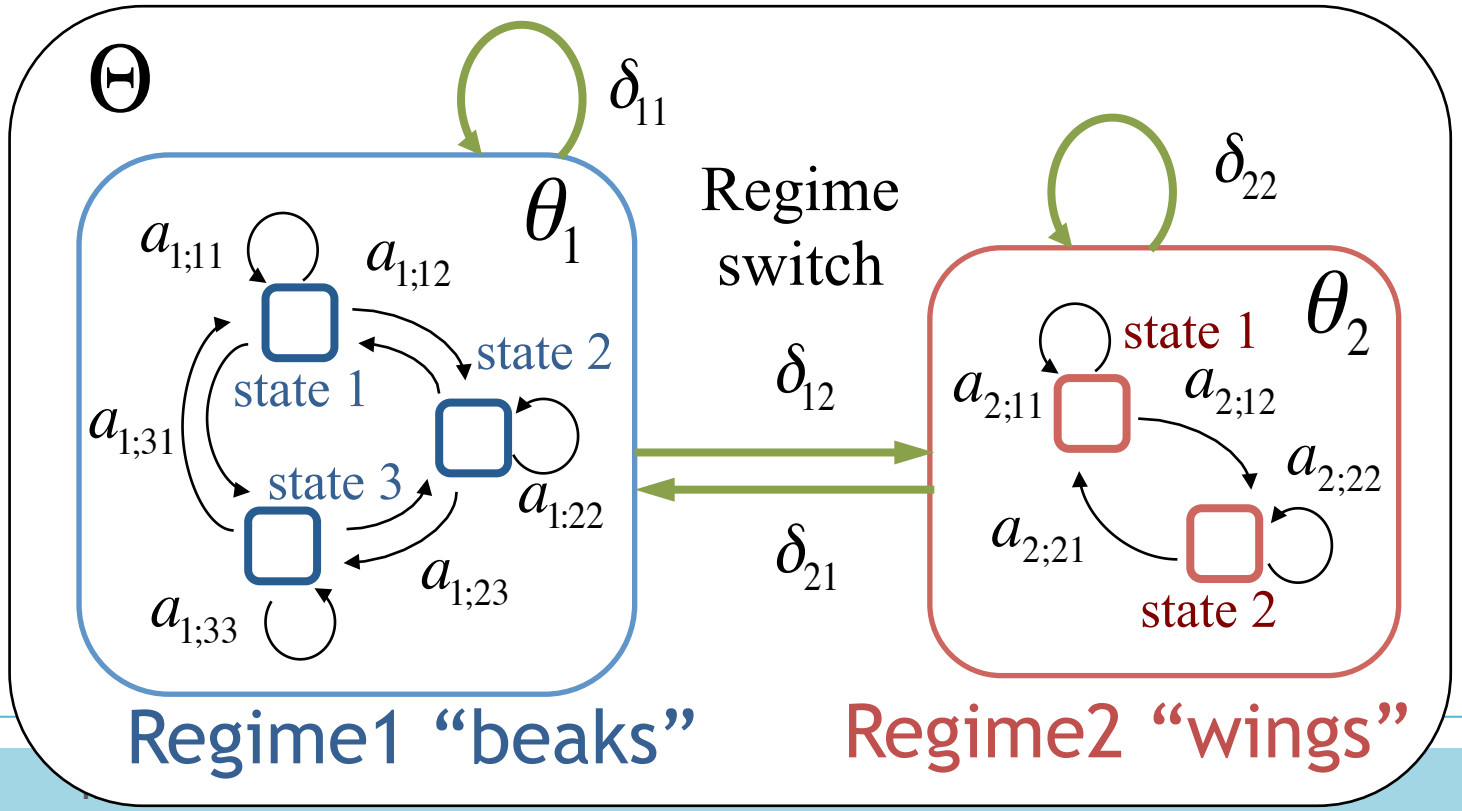
Details

$$\Theta = \{\theta_1, \theta_2, \ldots \theta_r, \Delta_{r \times r}\} \qquad (\theta_i = \{\pi, A, B\})$$

r regimes (HMMs)

across-regime transition prob.

Single HMM parameters



Regimes r=2

Regime 1 (k=3)

Regime 2 (k=2)

Idea (2): model description cost

Q2. How to decide # of regimes/segments ?

Idea (2): Model description cost

- Minimize coding cost
- find "optimal" # of segments/regimes

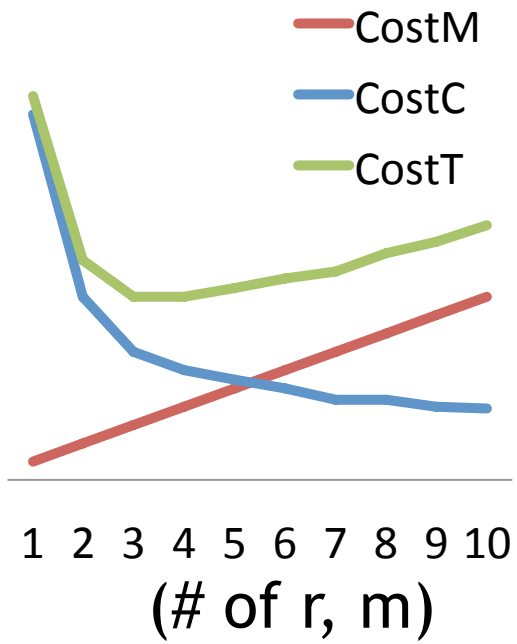
Idea (2): model description cost

Idea: Minimize encoding cost!

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

Model cost

Coding cost



Good compression ⇄ Good description

Idea (2): model description cost

Details

Total cost of bundle X, given C

$$C = \{m, r, S, \Theta, F\}$$

$$\begin{aligned} Cost_T(\boldsymbol{X}; \mathcal{C}) &= Cost_T(\boldsymbol{X}; m, r, \mathcal{S}, \boldsymbol{\Theta}, \mathcal{F}) \\ &= \log^*(n) + \log^*(d) + \log^*(m) + \log^*(r) + m\log(r) \\ &\quad + \sum_{i=1}^{m-1} \log^* |s_i| + Cost_M(\boldsymbol{\Theta}) + Cost_C(\boldsymbol{X}|\boldsymbol{\Theta}) \end{aligned} \quad (6)$$

Idea (2): model description cost

Details

Total cost of bundle X, given C

$$C = \{m, r, S, \Theta, F\}$$

duration/ dimensions

# of segments/ regimes

segment-membership F

$$Cost_T(\boldsymbol{X}; \mathcal{C}) = Cost_T(\boldsymbol{X}; m, r, \mathcal{S}, \boldsymbol{\Theta}, \mathcal{F})$$
$$= \log^*(n) + \log^*(d) + \log^*(m) + \log^*(r) + m\log(r)$$
$$+ \sum_{i=1}^{m-1} \log^* |s_i| + Cost_M(\boldsymbol{\Theta}) + Cost_C(\boldsymbol{X}|\boldsymbol{\Theta}) \quad (6)$$

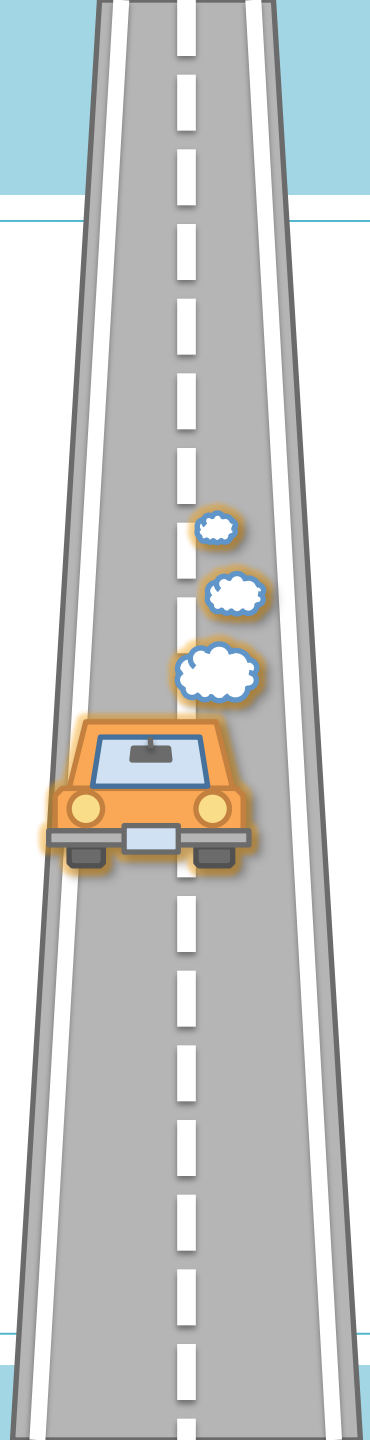
segment lengths

Model description cost of Θ

Coding cost of X given Θ

Outline

- Motivation
- Problem definition
- Compression & summarization
- Algorithms
- Experiments
- AutoPlait at work
- Conclusions

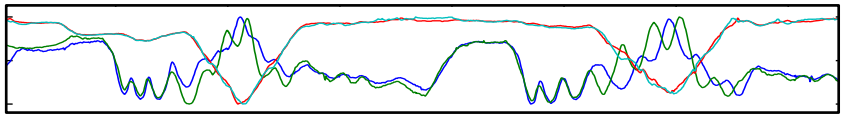


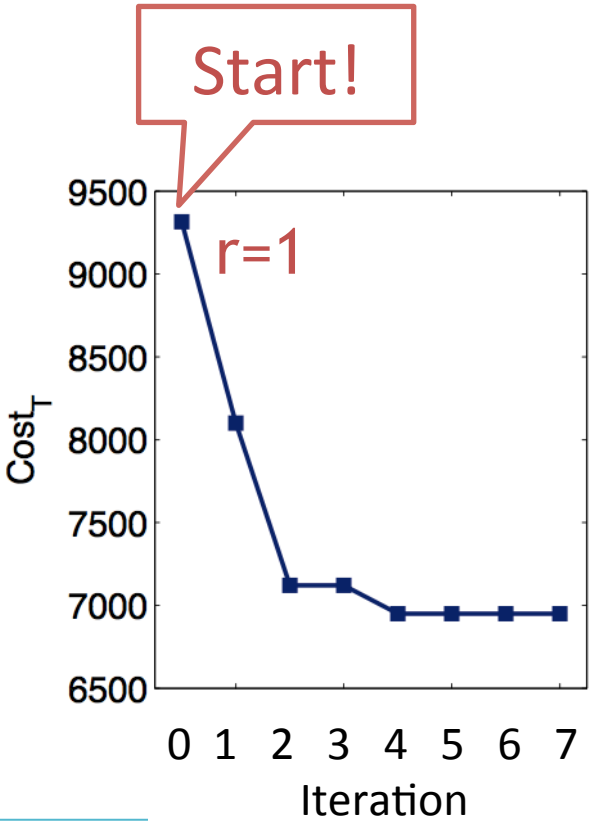
AutoPlait

Overview

Iteration 0
r=1, m=1

X

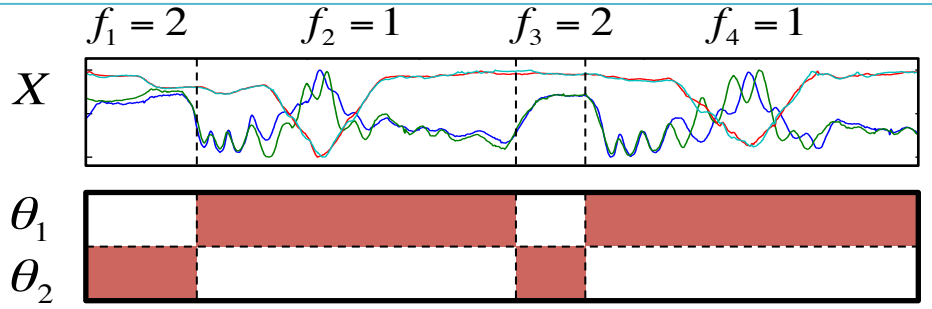


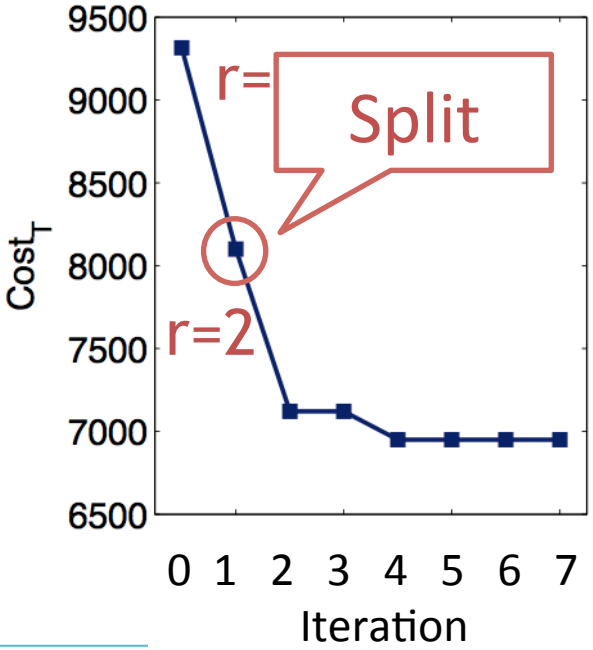


AutoPlait

Overview

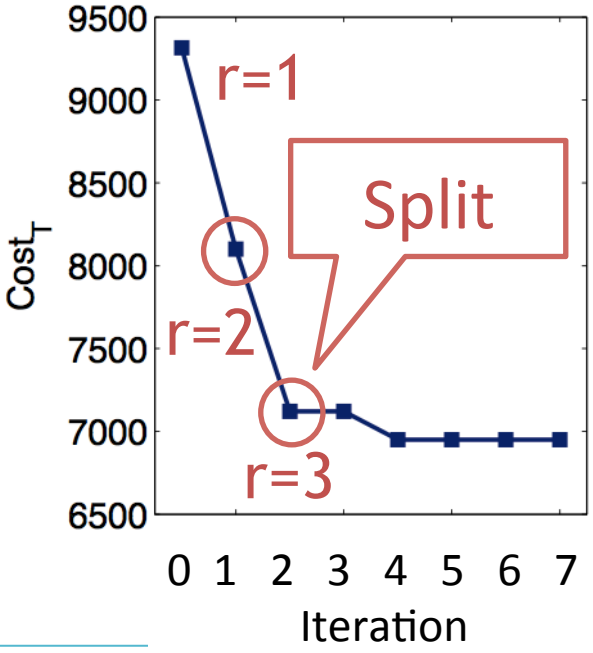
Iteration 1
r=2, m=4



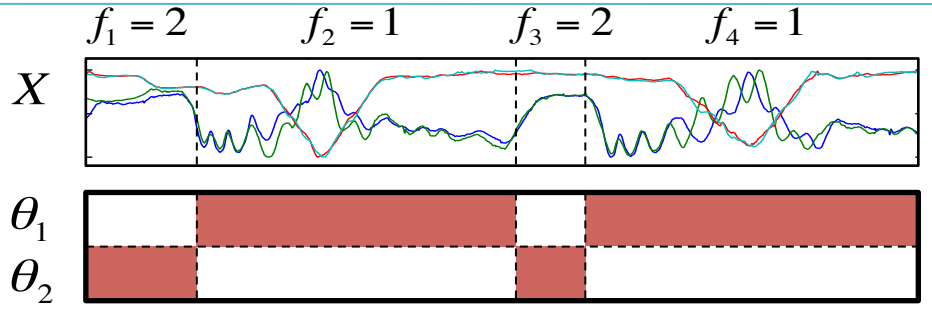


AutoPlait

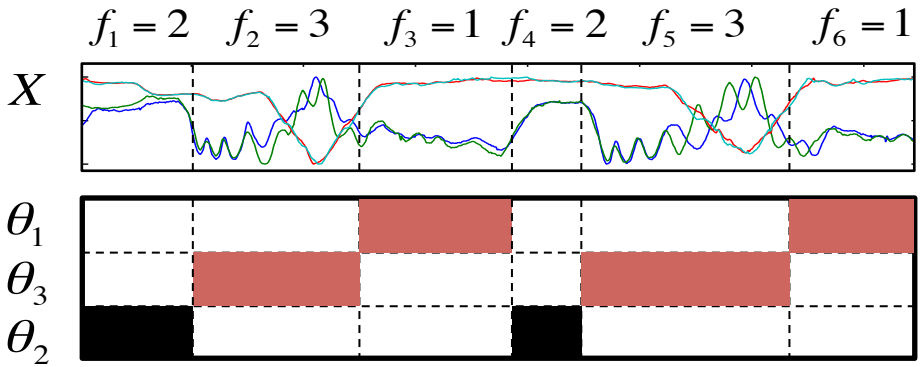
Overview



Iteration 1
r=2, m=4

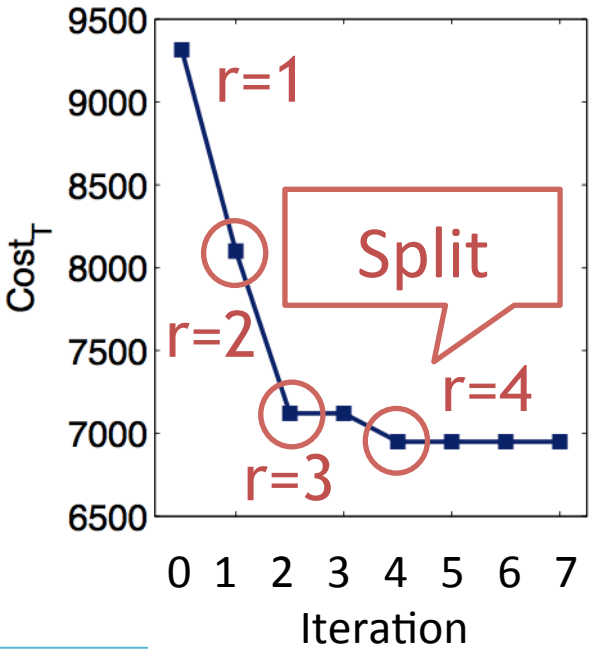


Iteration 2
r=3, m=6



AutoPlait

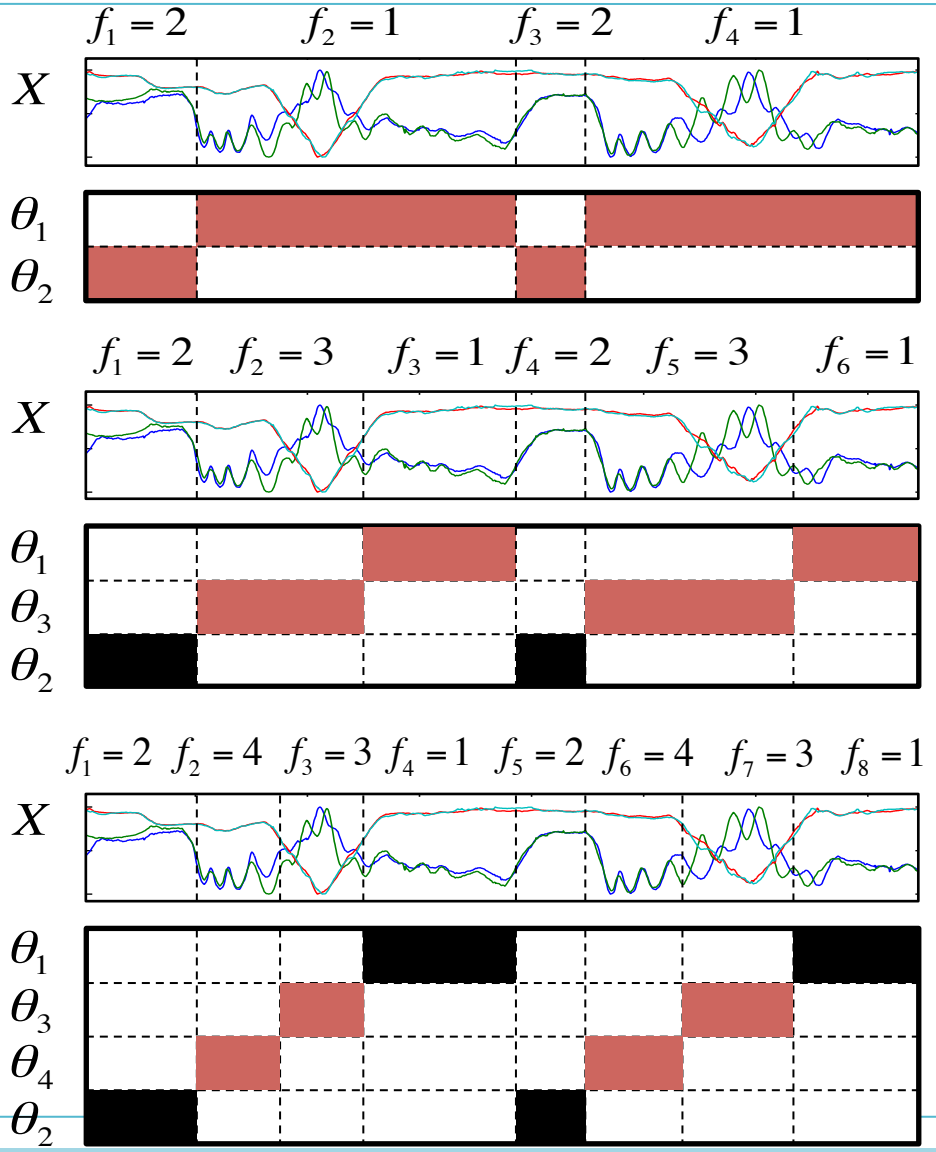
Overview



Iteration 1
r=2, m=4

Iteration 2
r=3, m=6

Iteration 4
r=4, m=8



AutoPlait

Algorithms

1. CutPointSearch — Inner-most loop

   Find good cut-points/segments

2. RegimeSplit — Inner loop

   Estimate good regime parameters **Θ**

3. AutoPlait — Outer loop

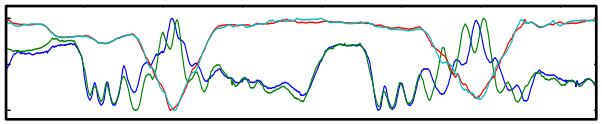
   Search for the best number of regimes (r=2,3,4…)

1.CutPointSearch

Inner-most loop

Given:

- bundle X



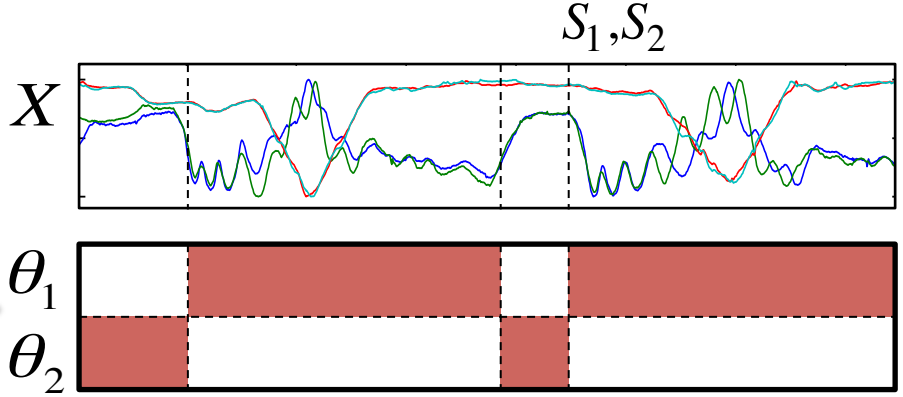
- parameters of two regimes $\Theta = \{\theta_1, \theta_2, \Delta\}$

Find: cut-points of segment sets S_1, S_2,

$$\{S_1, S_2\} = \operatorname{argmax} P(X \mid S_1, S_2, \Theta)$$

X

$\{\theta_1, \theta_2, \Delta\}$



$S_1 = \{s_2, s_4\}$

$S_2 = \{s_1, s_3\}$

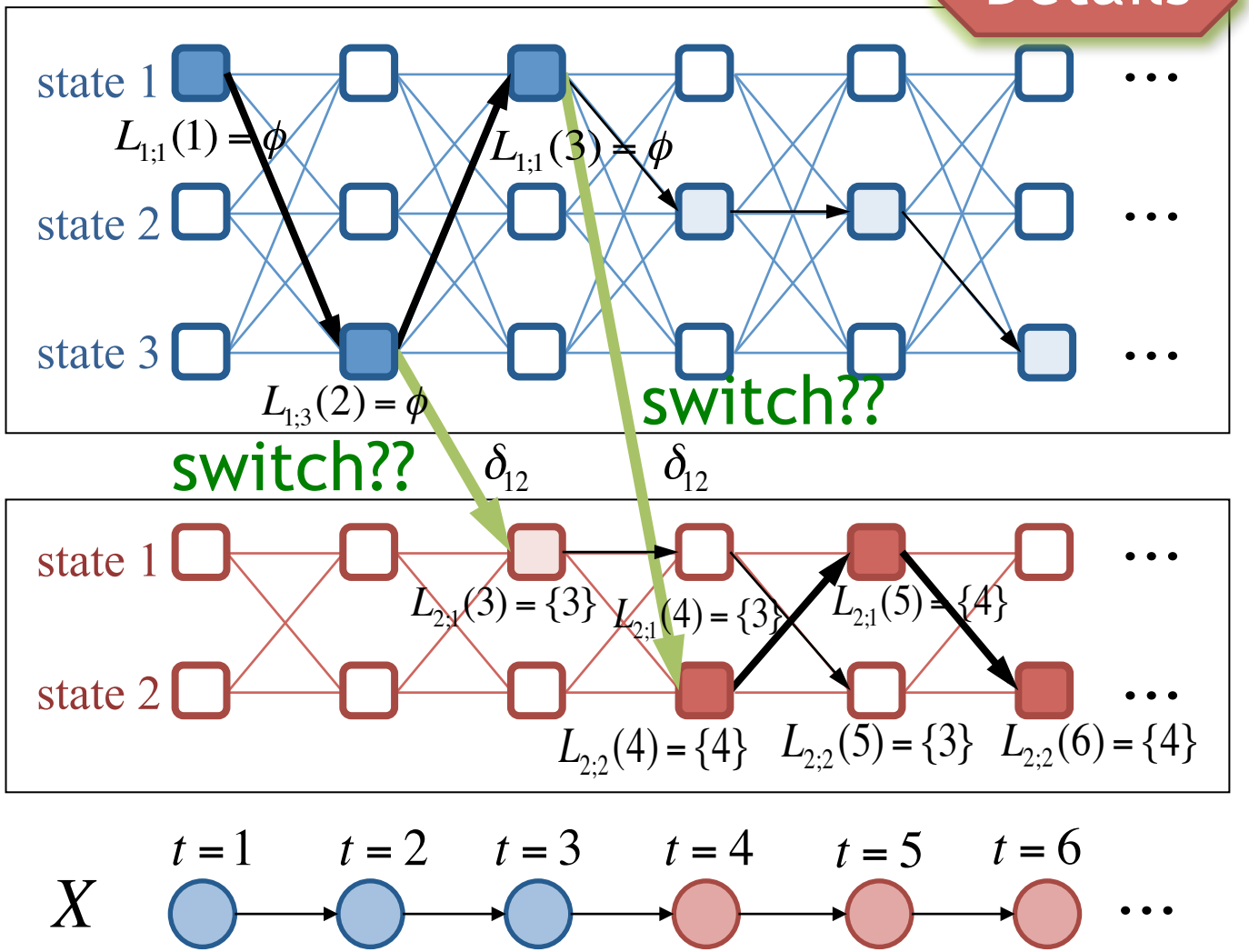
1.CutPointSearch

Inner-most loop

Details

DP algorithm to compute likelihood:

$$P(X \mid \Theta)$$



1.CutPointSearch

Inner-most loop

Details

Theoretical analysis

Scalability

- It takes $O(ndk^2)$ time (only single scan)
  - n: length of X
  - d: dimension of X
  - k: # of hidden states in regime

Accuracy

It guarantees the optimal cut points

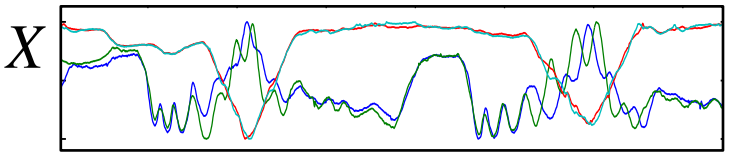
- (Details in paper)

2.RegimeSplit

Inner loop

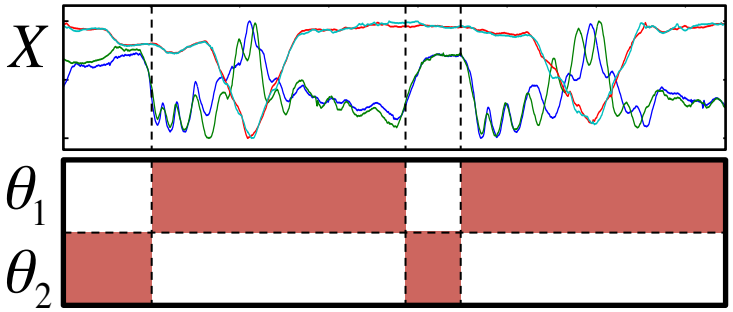
Given:

- bundle X



Find: two regimes

1. find cut-points of segment sets: S_1, S_2
2. estimate parameters of two regimes:



$$\Theta = \{\theta_1, \theta_2, \Delta\}$$

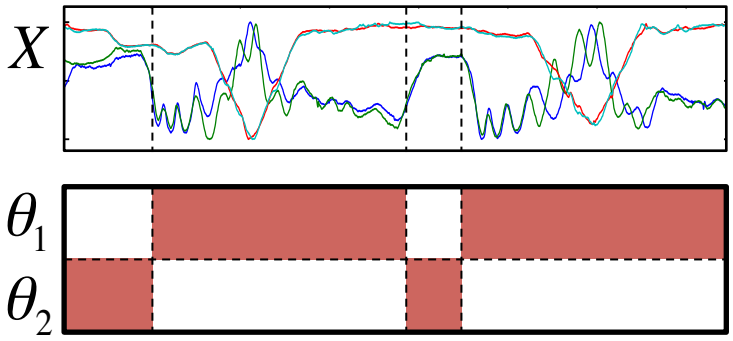
2.RegimeSplit

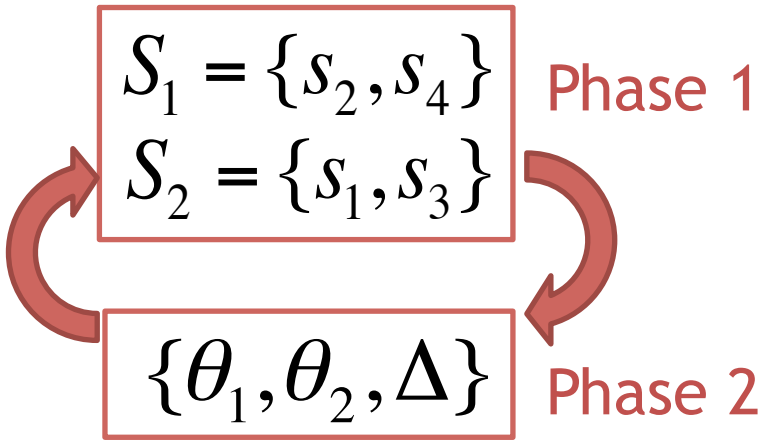
Inner loop

Details

Two-phase iterative approach

- Phase 1: (CutPointSearch)
  - Split segments into two groups : S_1, S_2
- Phase 2: (BaumWelch)
  - Update model parameters: $\Theta = \{\theta_1, \theta_2, \Delta\}$



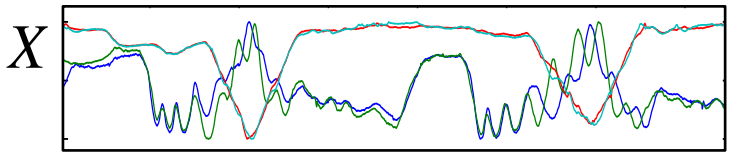


3.AutoPlait

Outer loop

Given:

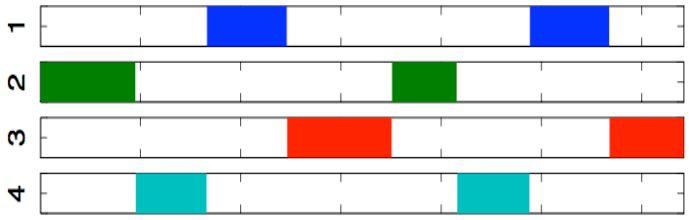
- bundle X



Find: r regimes (r=2, 3, 4, …)

- i.e., find full parameter set

$$C = \{m, r, S, \Theta, F\}$$



3.AutoPlait

Outer loop

Split regimes r=2,3,…, as long as cost keeps decreasing

- Find appropriate # of regimes

$$r = \min_{r} Cost_T(X; m, r, S, \Theta, F)$$

3.AutoPlait

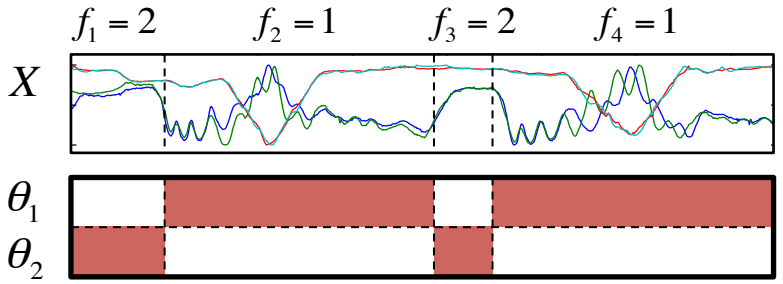
Outer loop

Split regimes r=2,3,…, as long as cost keeps decreasing

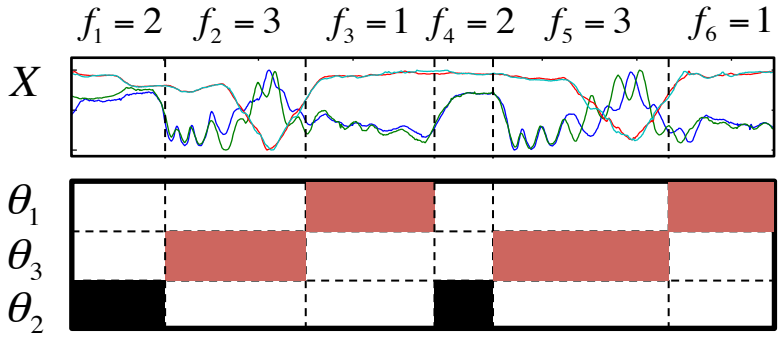
- Find appropriate # of regimes

$$r = \min_{r} Cost_T(X; m, r, S, \Theta, F)$$

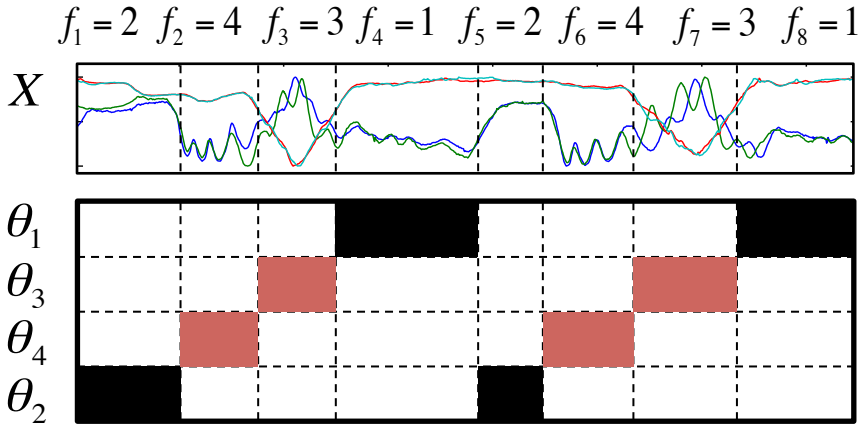
r=2, m=4



r=3, m=6

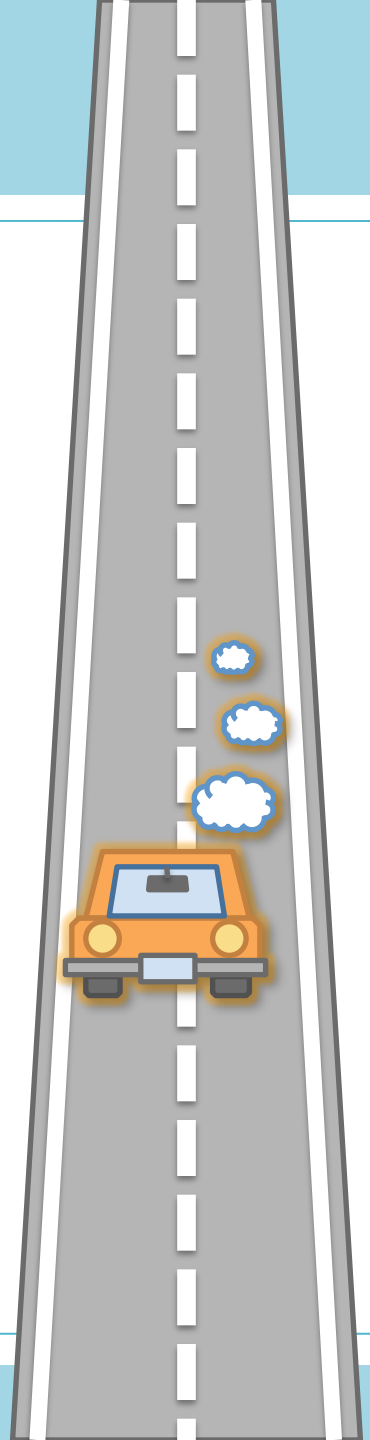


r=4, m=8



Outline

- Motivation
- Problem definition
- Compression & summarization
- Algorithms
- Experiments
- AutoPlait at work
- Conclusions



Experiments

We answer the following questions…

Q1. Sense-making

Can it help us understand the given bundles?

Q2. Accuracy

How well does it find cut-points and regimes?

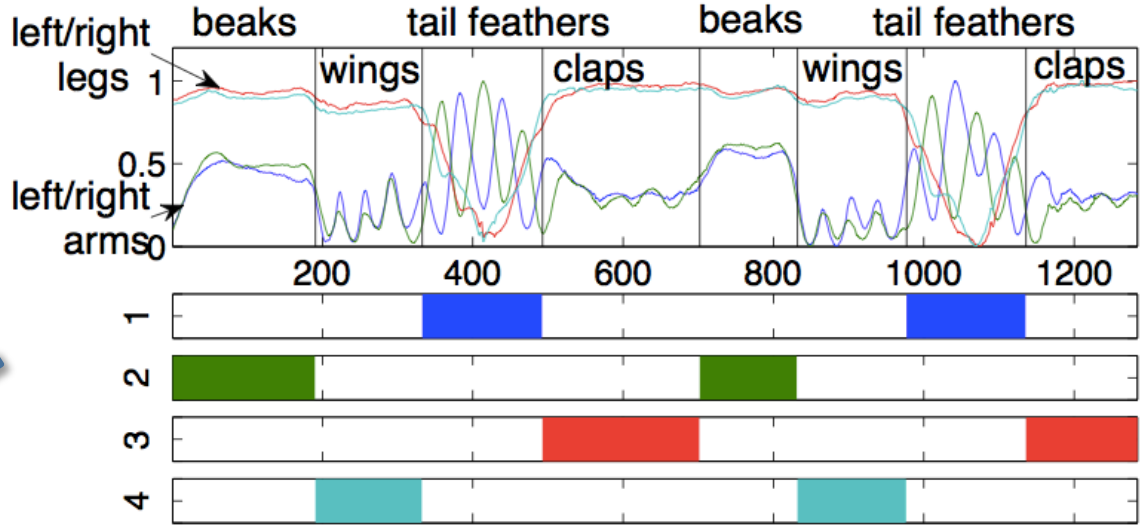
Q3. Scalability

How does it scale in terms of computational time?

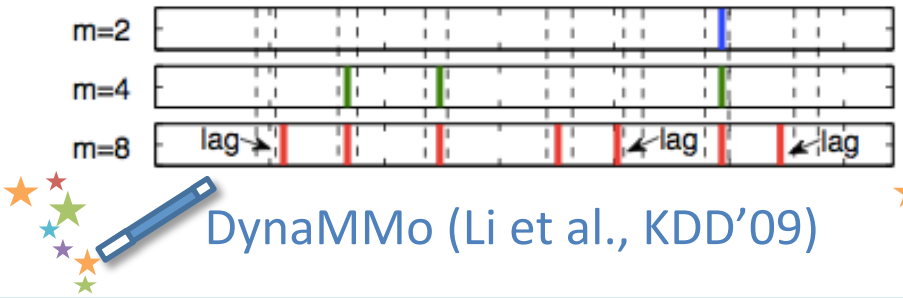
Q1. Sense-making

MoCap data

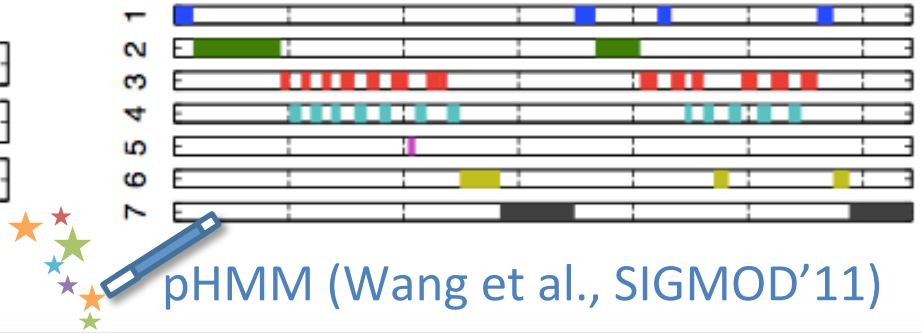
AutoPlait (**NO** magic numbers)



(a) AUTOPLAIT (**no** user defined parameters)



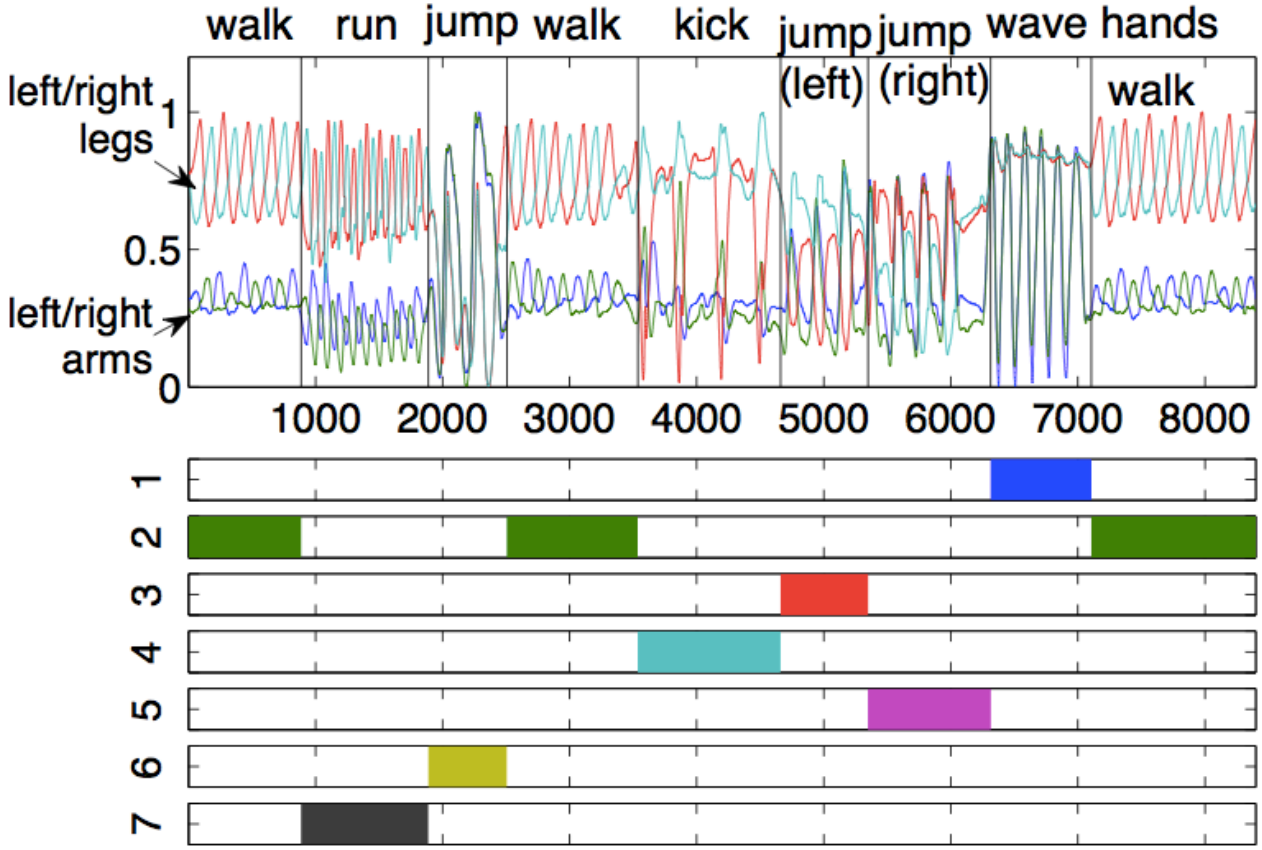
DynaMMo (Li et al., KDD'09)



pHMM (Wang et al., SIGMOD'11)

Q1. Sense-making

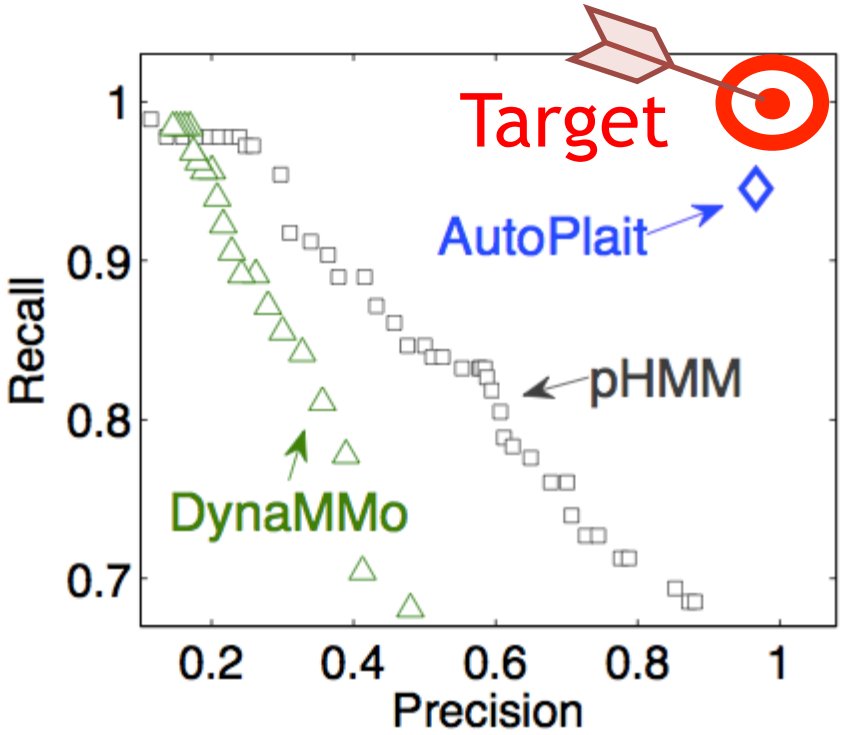
MoCap data



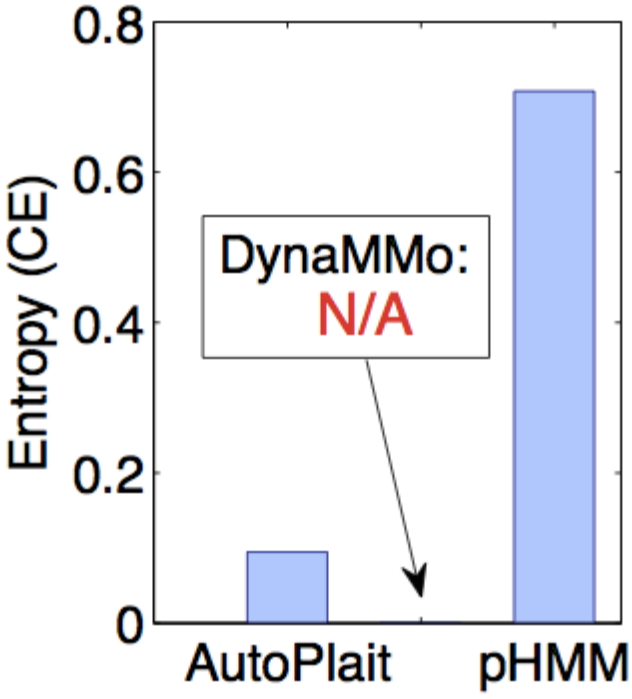
AutoPlait (NO magic numbers)

Q2. Accuracy

(a) Segmentation



(b) Clustering



(a) Precision and recall (higher is better)

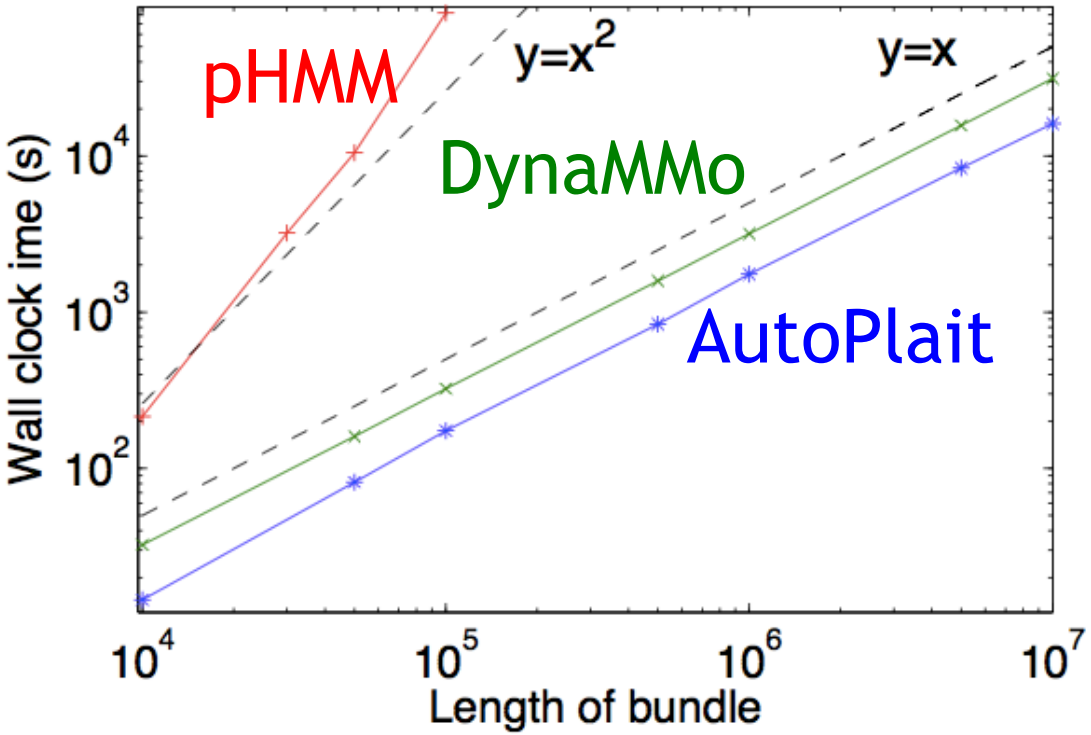
(b) CE score (lower is better)

AutoPlait needs "no magic numbers"

Q3. Scalability

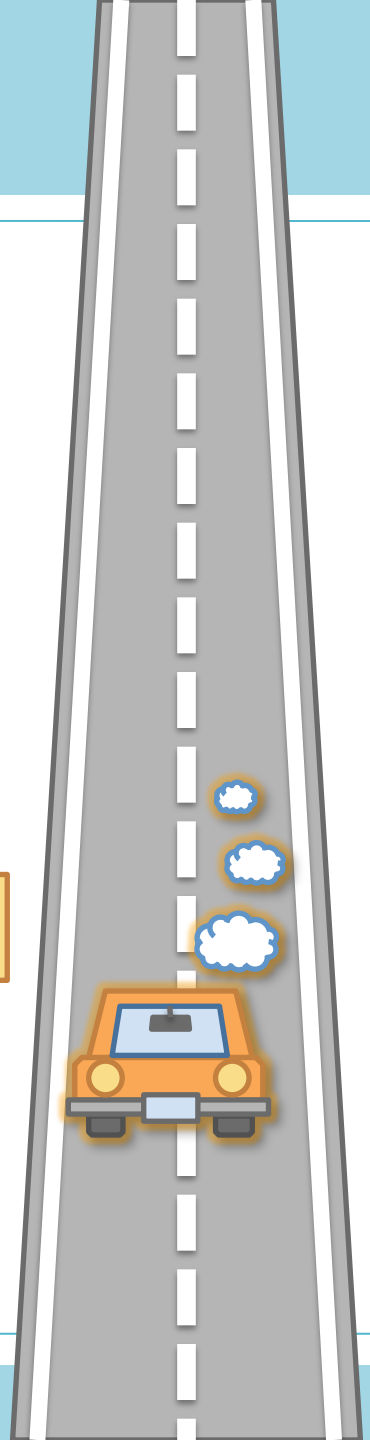
Wall clock time vs. data size (length) : n

AutoPlait scales linearly, i.e., O(n)



Outline

- Motivation
- Problem definition
- Compression & summarization
- Algorithms
- Experiments
- AutoPlait at work
- Conclusions



AutoPlait at work

AutoPlait is capable of various applications, e.g.,

App1. Model analysis

- Web-click sequences

App2. Event discovery

- Google Trend data

AutoPlait at work

AutoPlait is capable of various applications,

e.g.,

App1. Model analysis

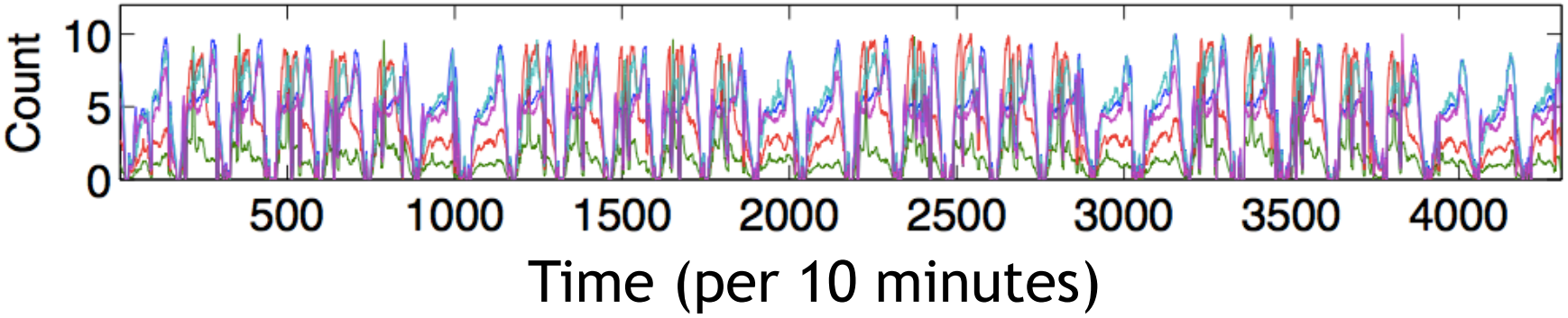
- Web-click sequences

App2. Event discovery

- Google Trend data

App1. Model analysis (WebClick)

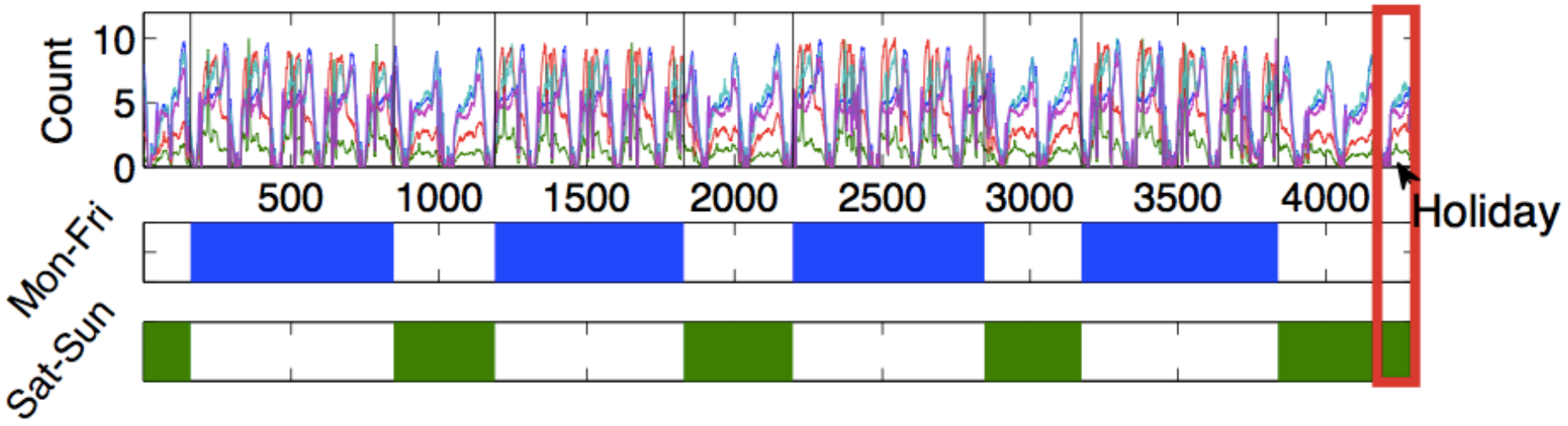
Web-click sequences (1 month, 5urls)



– 5urls: blog, news, dictionary, Q&A, mail
– every 10 minutes

App1. Model analysis (WebClick)

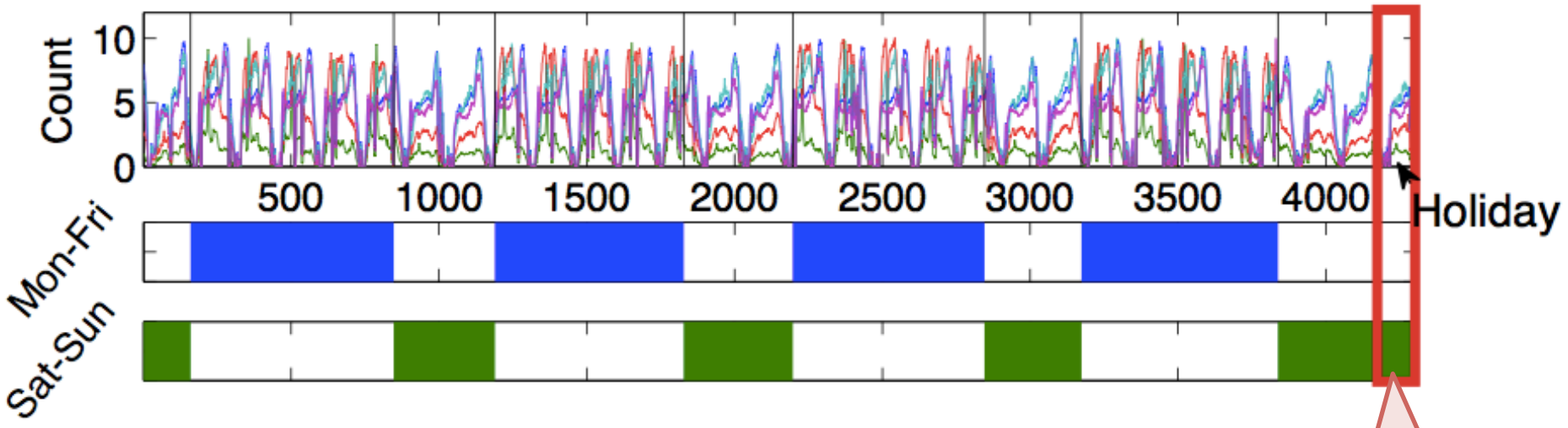
Web-click sequences (1 month, 5urls)



AutoPlait finds 2 patterns: weekday/weekend !

App1. Model analysis (WebClick)

Web-click sequences (1 month, 5urls)



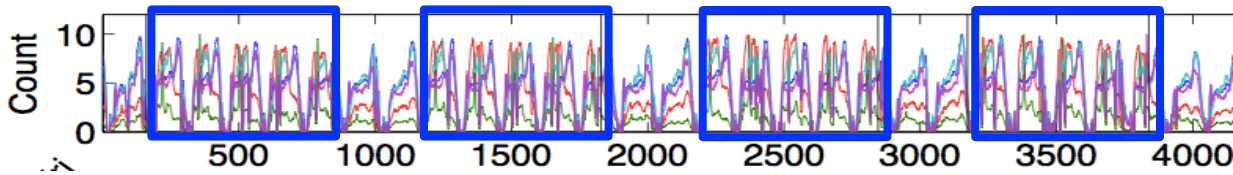
AutoPlait finds 2 patterns: weekday

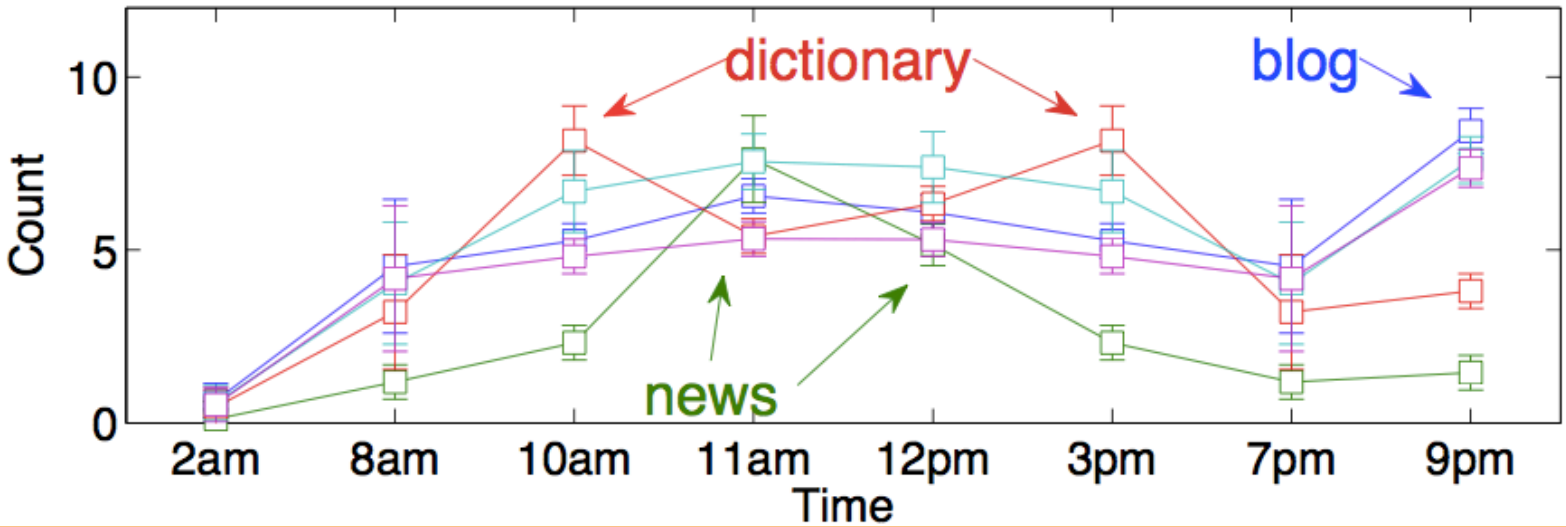
Monday
(but holiday)

App1. Model analysis (WebClick)

Details

Pattern of weekday regime



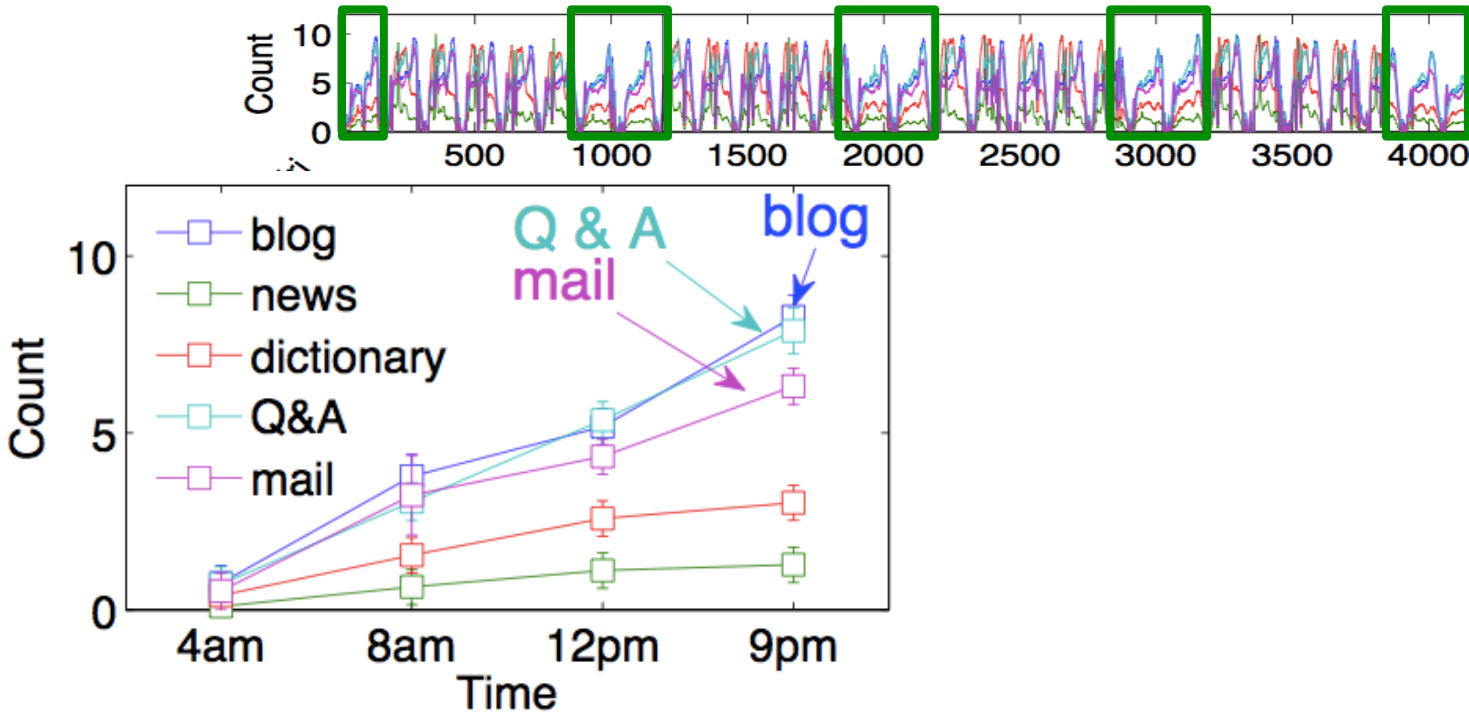


Observation: Working hard every weekday
(i.e., using dictionary, news sites)

App1. Model analysis (WebClick)

Details

Pattern of weekend regime



Observation: No more work on weekend
(i.e., blog, mail, Q&A for non-business purposes)

AutoPlait at work

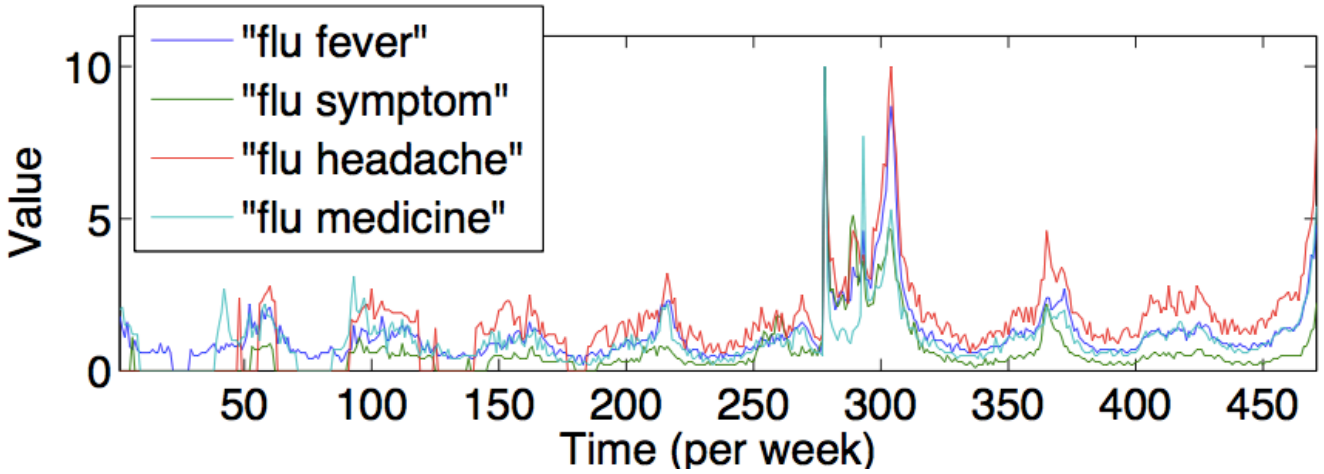
AutoPlait is capable of various applications,
e.g.,

App1. Model analysis
- Web-click sequences

App2. Event discovery
- Google Trend data

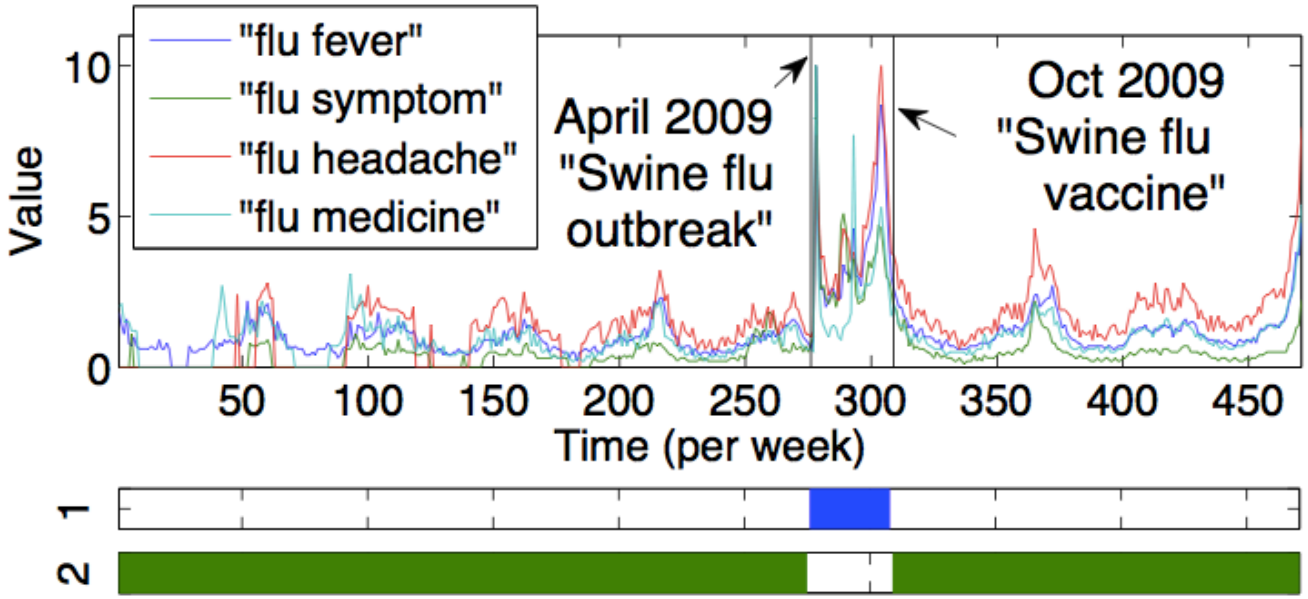
App2. Event discovery (GoogleTrend)

Anomaly detection (flu-related topics,10 years)



App2. Event discovery (GoogleTrend)

Anomaly detection (flu-related topics,10 years)

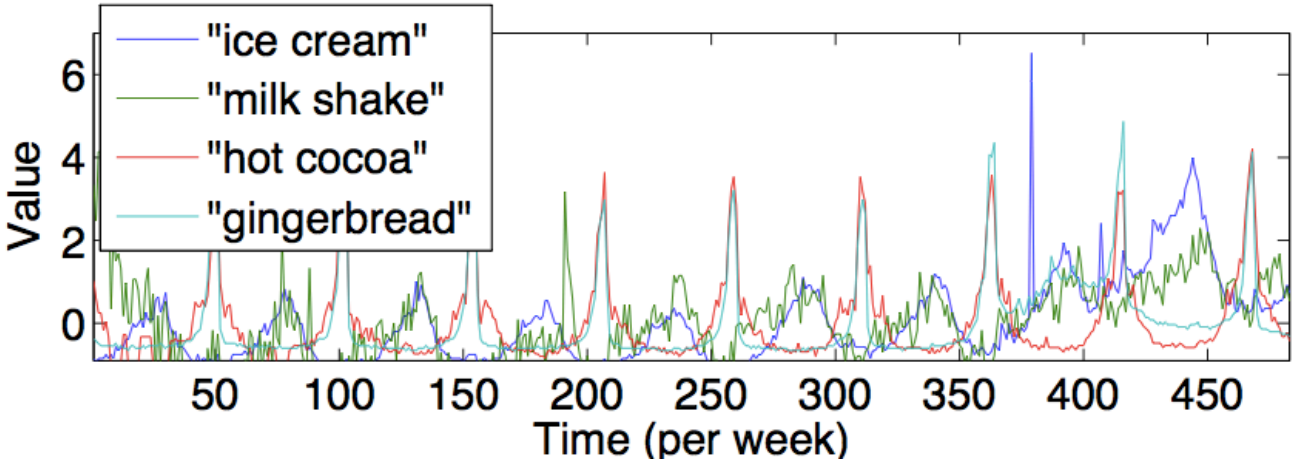


(a) Flu-related topics (regimes $r = 2$)

AutoPlait detects 1 unusual spike in 2009 (i.e., swine flu)

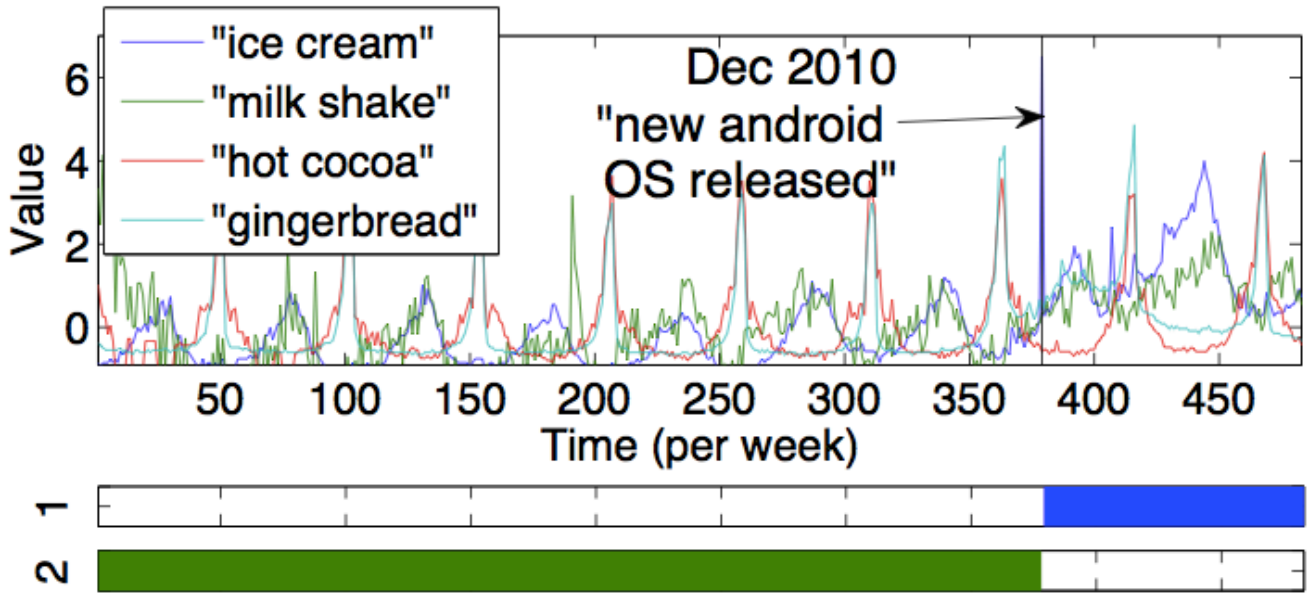
App2. Event discovery (GoogleTrend)

Turning point detection (seasonal sweets topics)



App2. Event discovery (GoogleTrend)

Turning point detection (seasonal sweets topics)

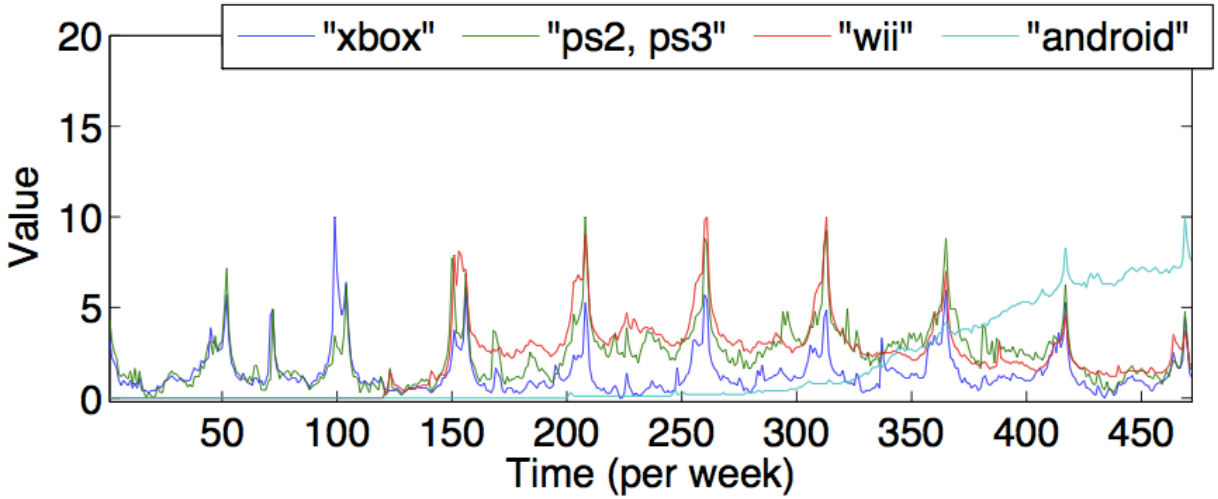


(b) Seasonal sweets topics (regimes $r = 2$)

Trend suddenly changed in 2010 (release of android OS “Ginger bread”, “Ice Cream Sandwich”)

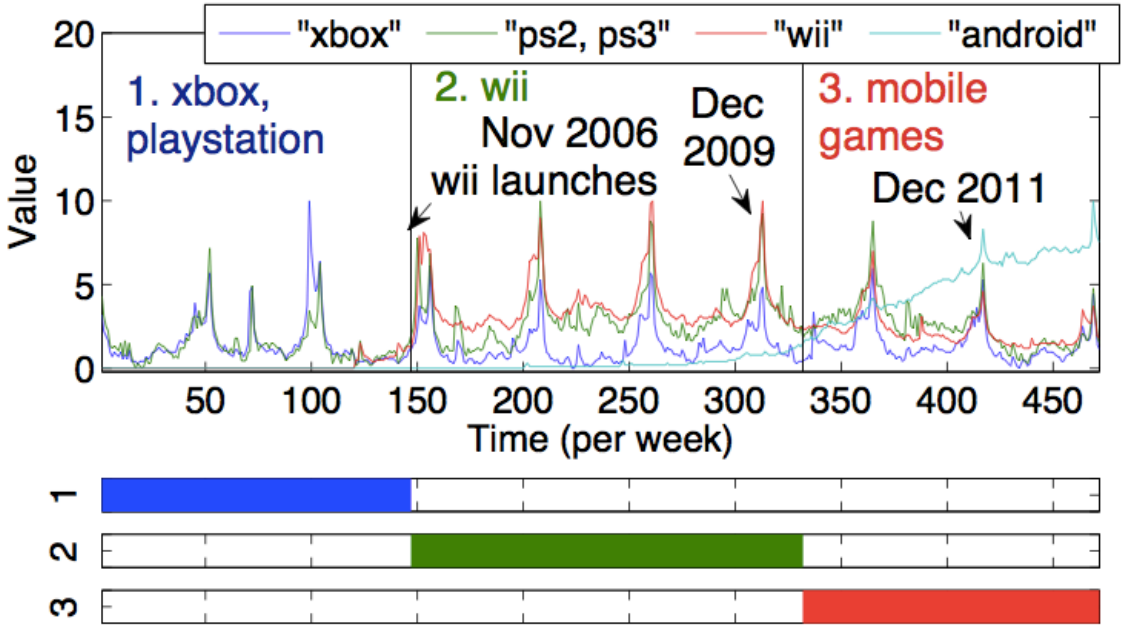
App2. Event discovery (GoogleTrend)

Trend discovery (game-related topics)



App2. Event discovery (GoogleTrend)

Trend discovery (game-related topics)

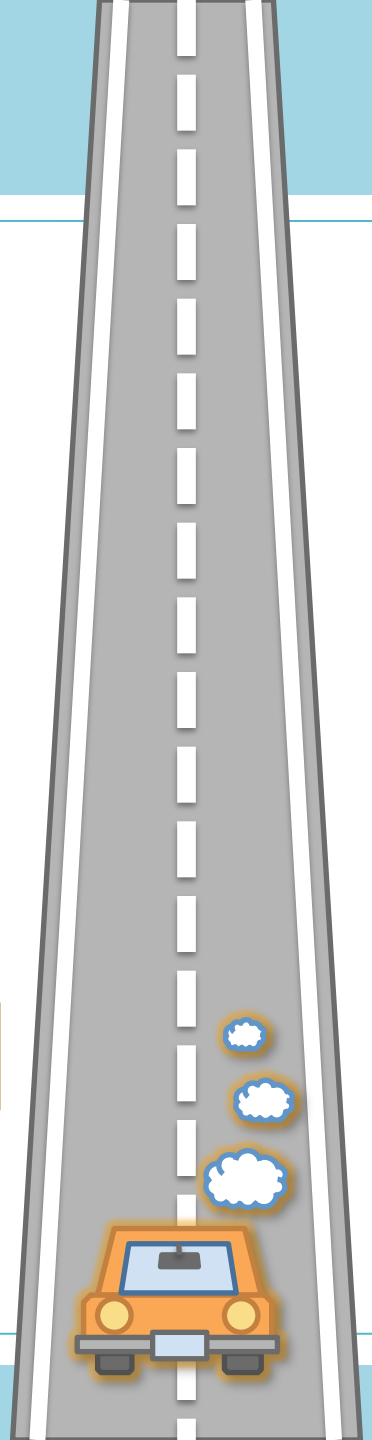


(c) Game-related topics (regimes $r = 3$)

It discovers 3 phases of "game console war" (Xbox&PlayStation/Wii/Mobile social games)

Outline

- Motivation
- Problem definition
- Compression & summarization
- Algorithms
- Experiments
- AutoPlait at work
- Conclusions



Conclusions

AutoPlait has the following properties

- Effective ✔

  Find optimal segments/regimes

- Sense-making ✔

  Reasonable regimes

- Fully-automatic ✔

  No magic numbers

- Scalable ✔

  It scales linearly

EcoWeb
WWW'15

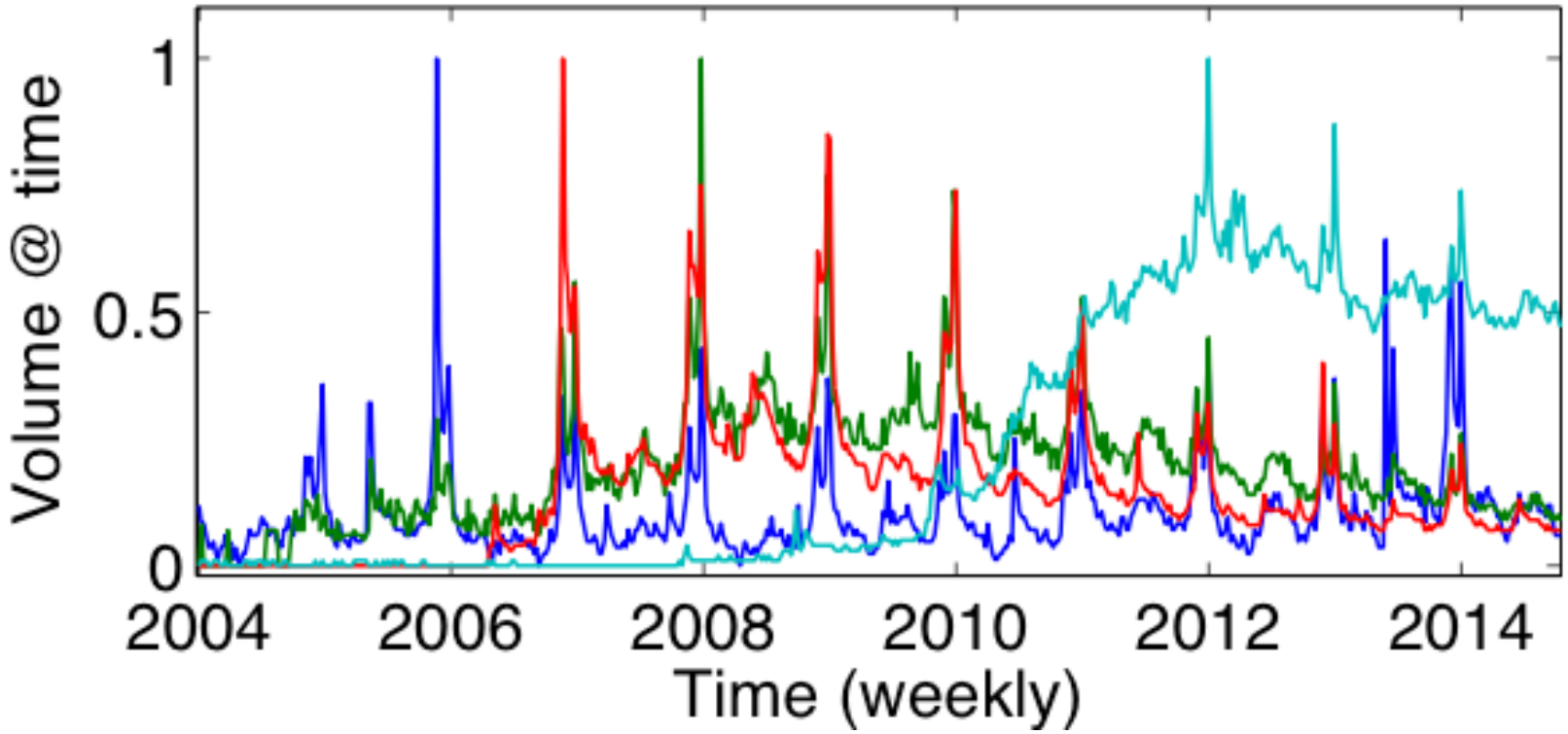
The Web as a Jungle: Non-Linear Dynamical Systems for Co-evolving Online Activities

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

Given: online user activities

e.g., *Google* search volumes for

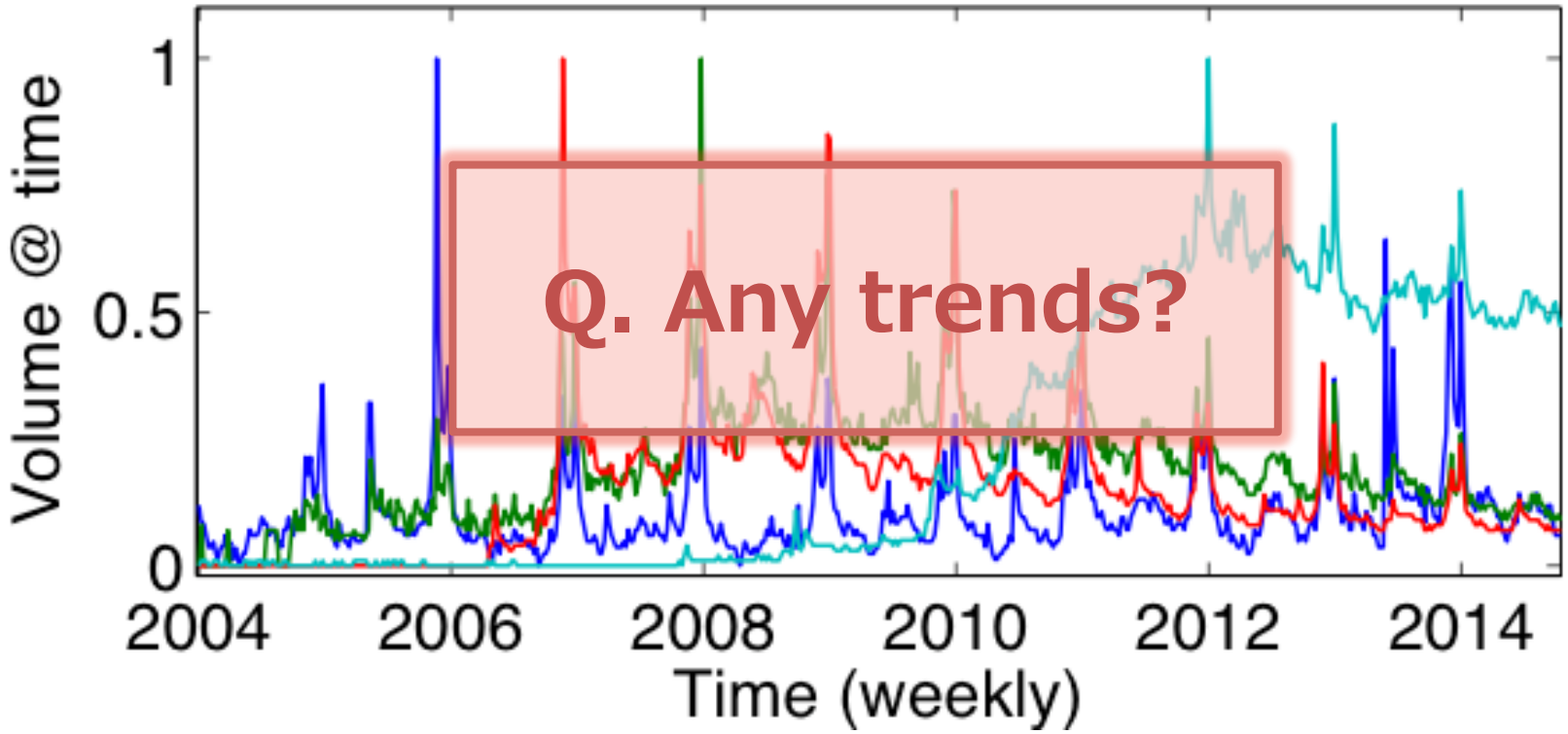
Xbox, **PlayStation**, **Wii**, **Android**



Given: online user activities

e.g., *Google* search volumes for

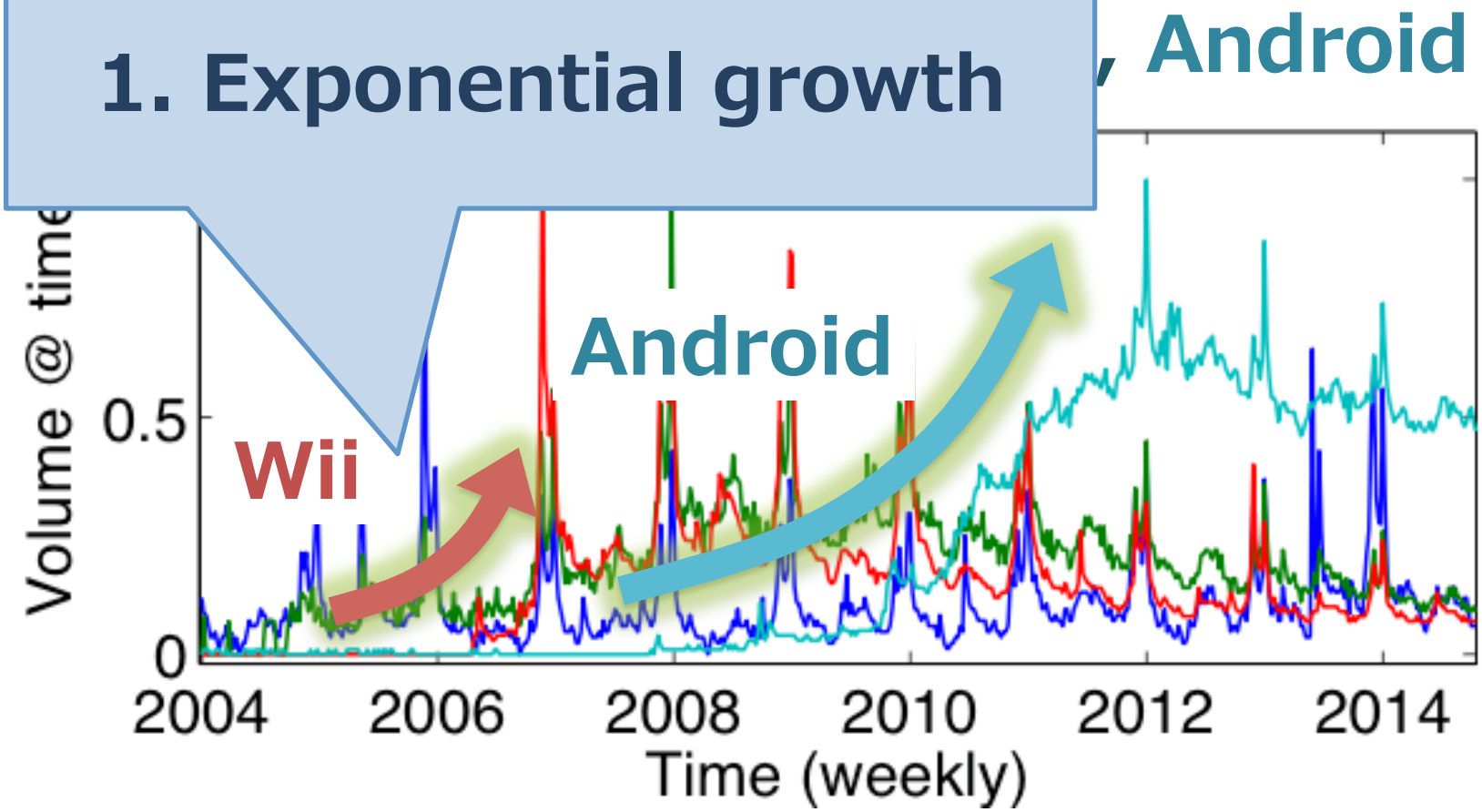
Xbox, **PlayStation**, **Wii**, **Android**



Given: online user activities

e.g., *Google* search volumes for

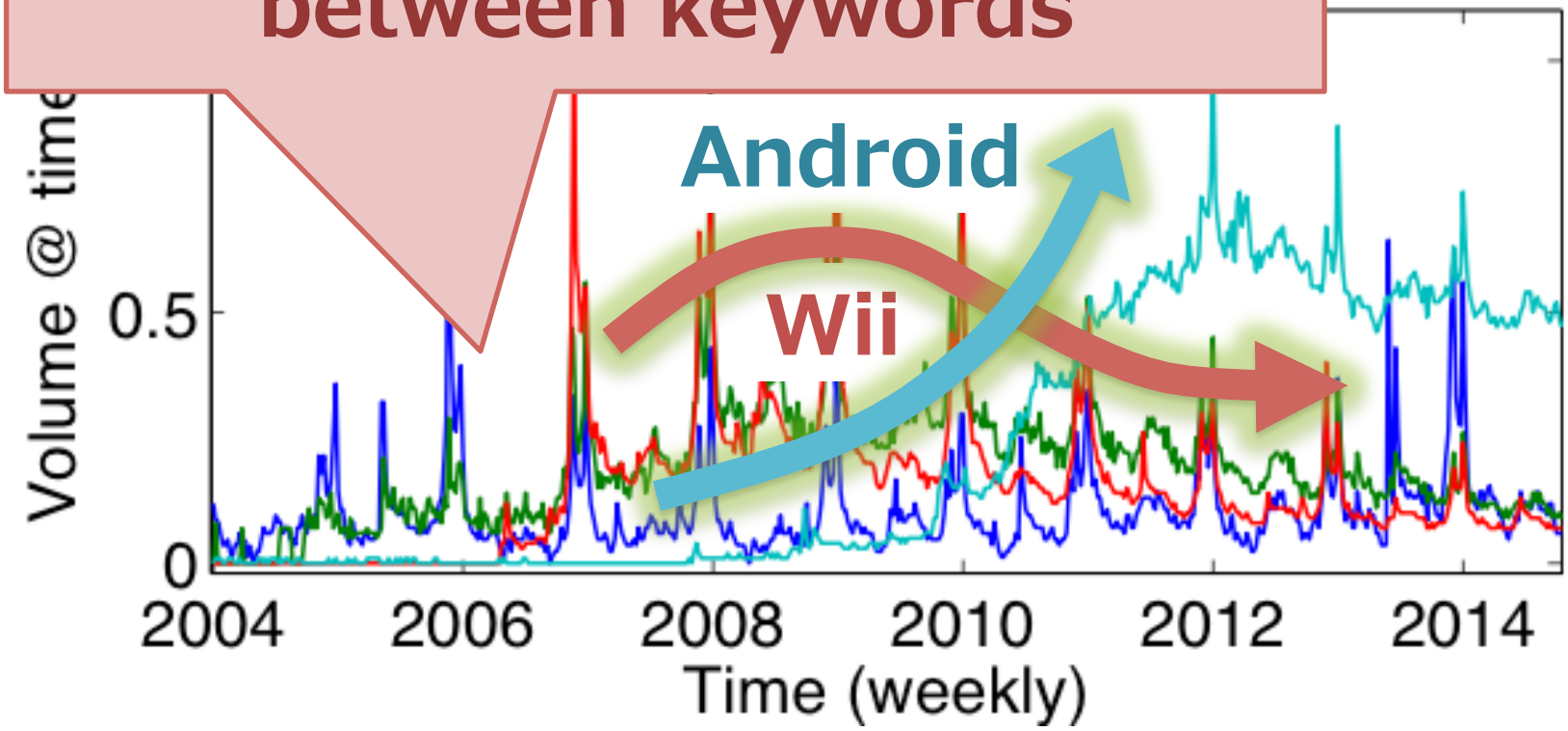
1. Exponential growth



Given: online user activities

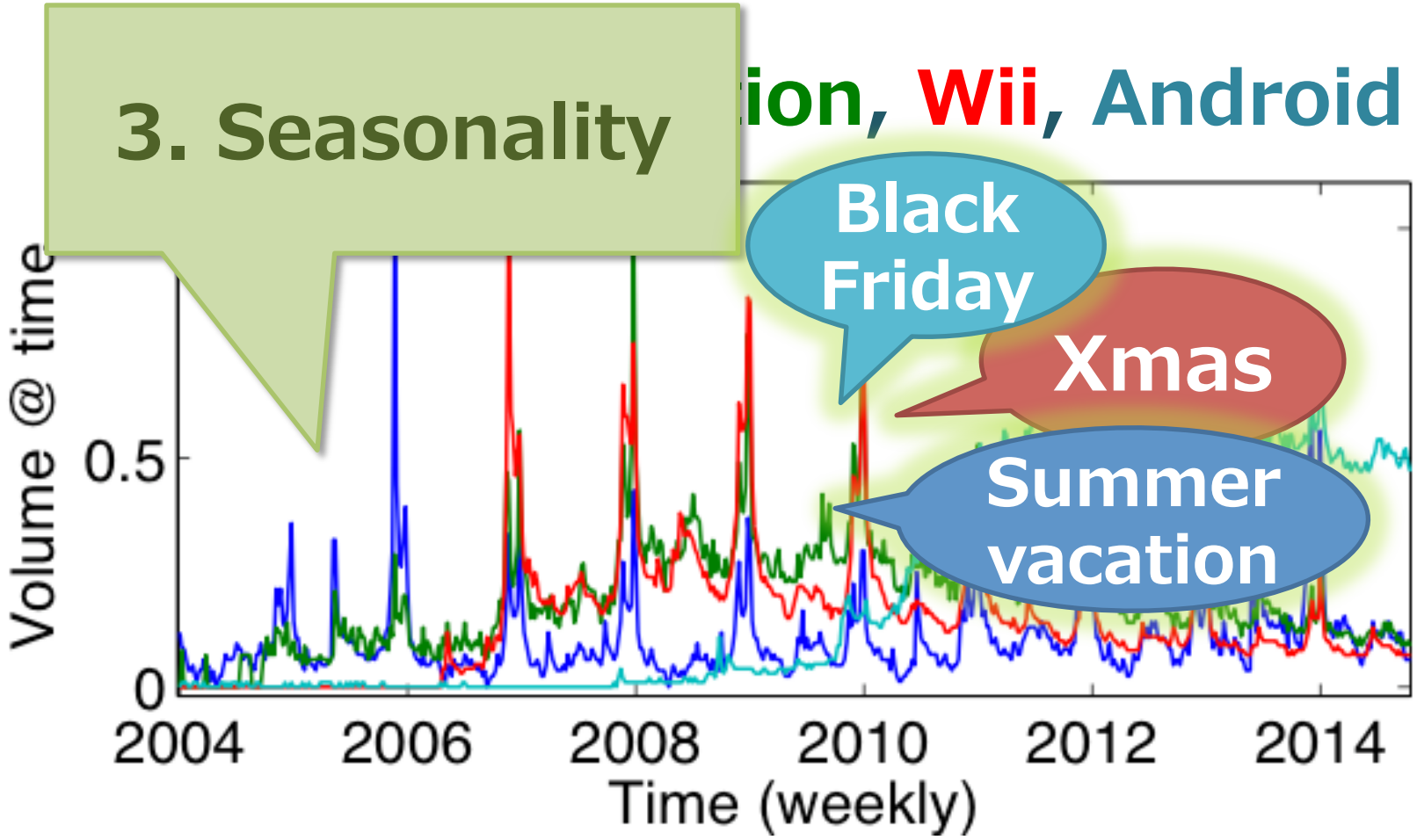
e.g., *Google* search volumes for

2. Interaction/competition between keywords



Given: online user activities

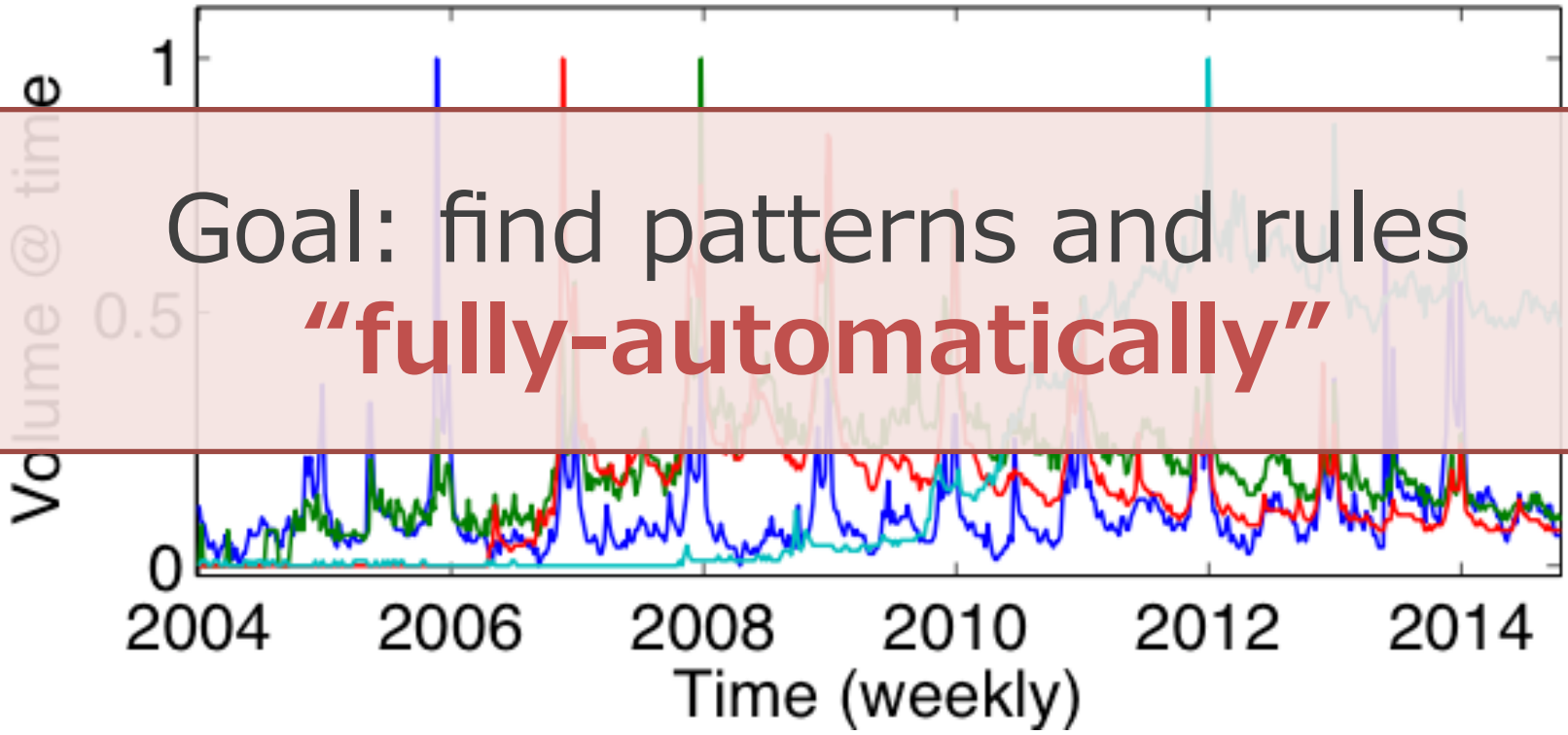
e.g., *Google* search volumes for



Given: online user activities

e.g., *Google* search volumes for

Xbox, **PlayStation**, **Wii**, **Android**

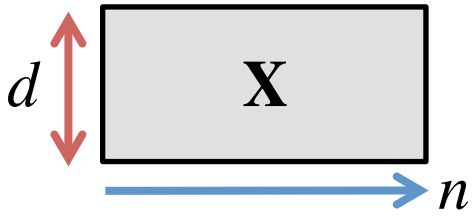


Goal: find patterns and rules
"fully-automatically"

Problem definition

Given: Co-evolving online activities

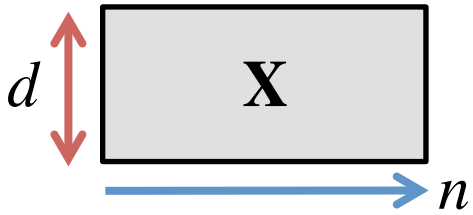
X (activity x time)



Problem definition

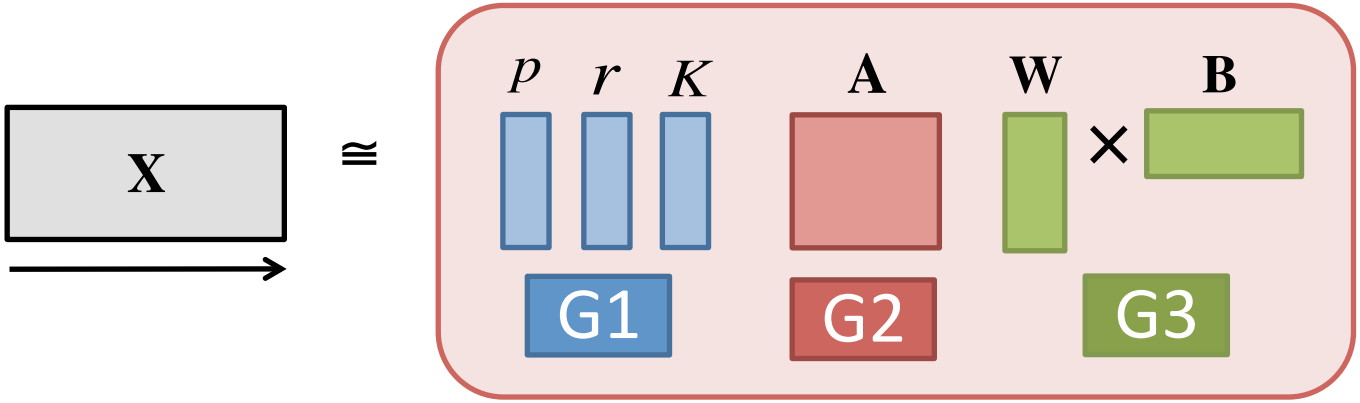
Given: Co-evolving online activities

X (activity x time)



Find: Compact description of X

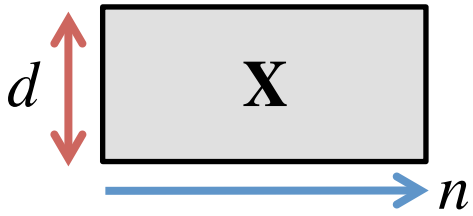
EcoWeb



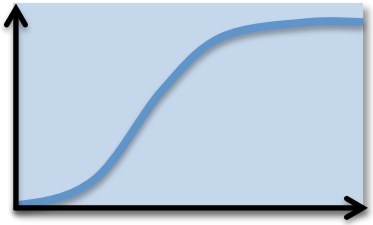
Problem definition

Given: Co-evolving online activities

x time)

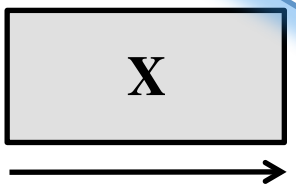


G1 **Non-linear evolution**

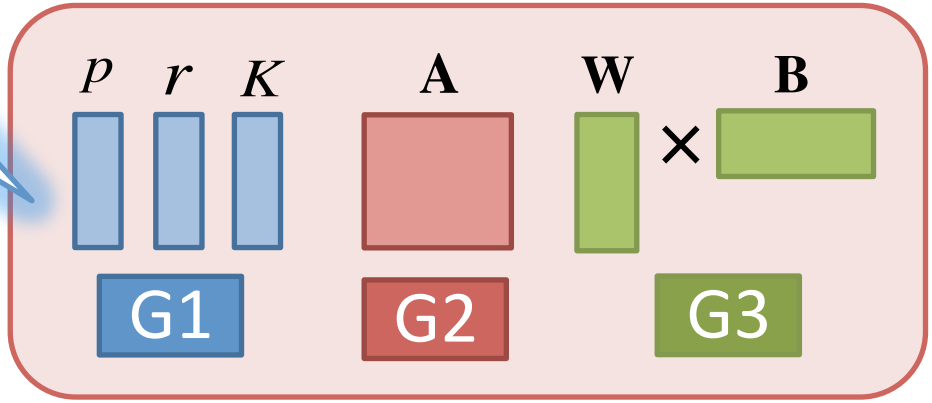


Fi scription of X

EcoWeb



≅

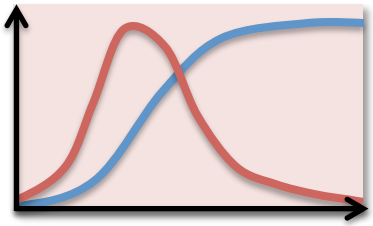


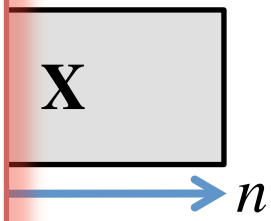
Problem definition

Given: Co-evolving

X (activity x

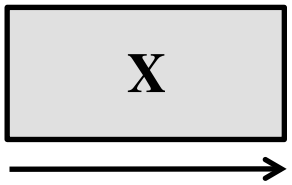
G2 Interaction/ competition



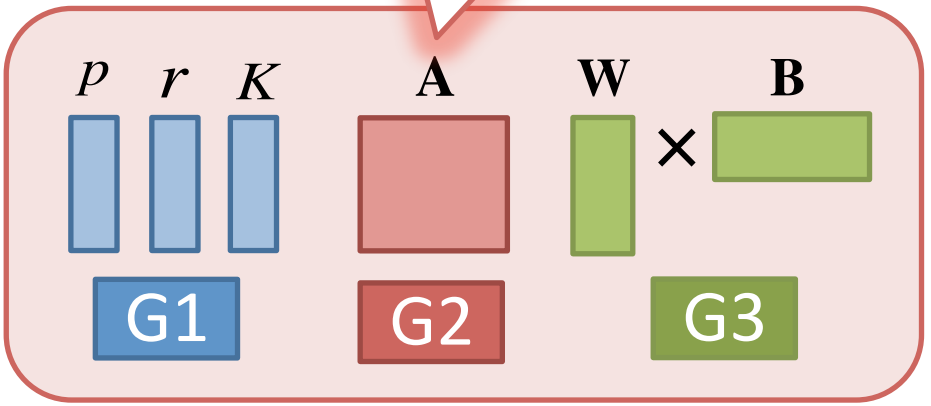


Find: Compact descr

Eco



≅



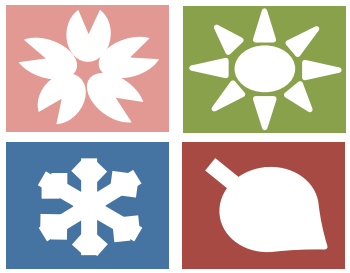
Problem definition

Given: Co-evolving online acti

X (activity x time)

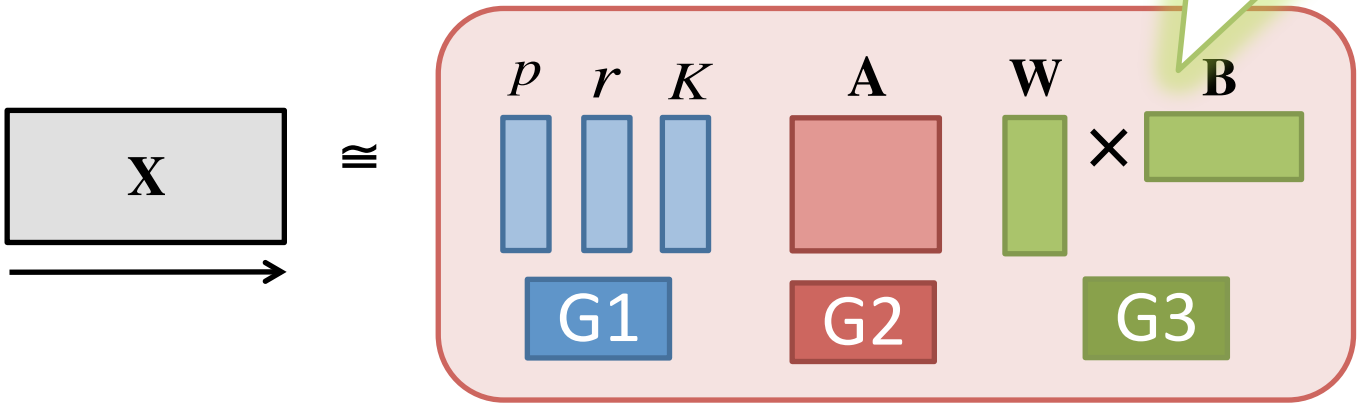
G3

Seasonality



Find: Compact description of

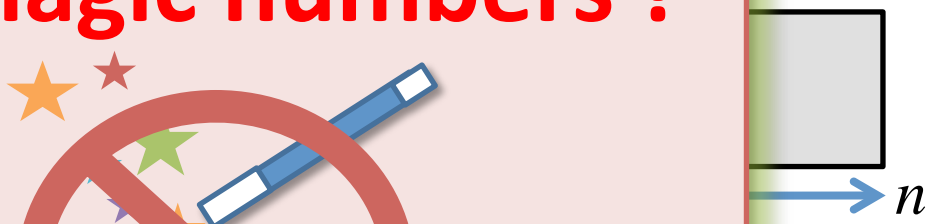
EcoWeb



Problem definition

Given: Co-e

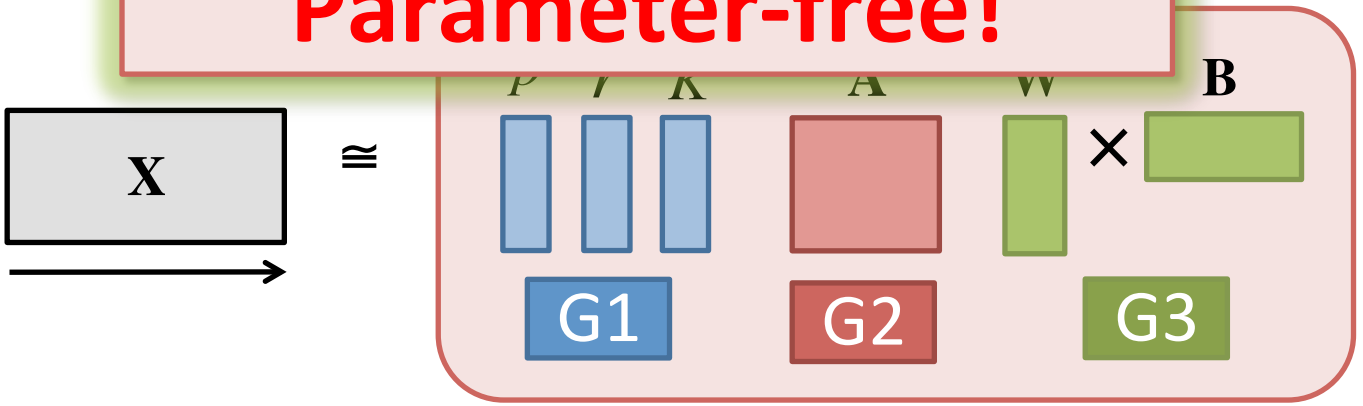
X (a



Find: Comp

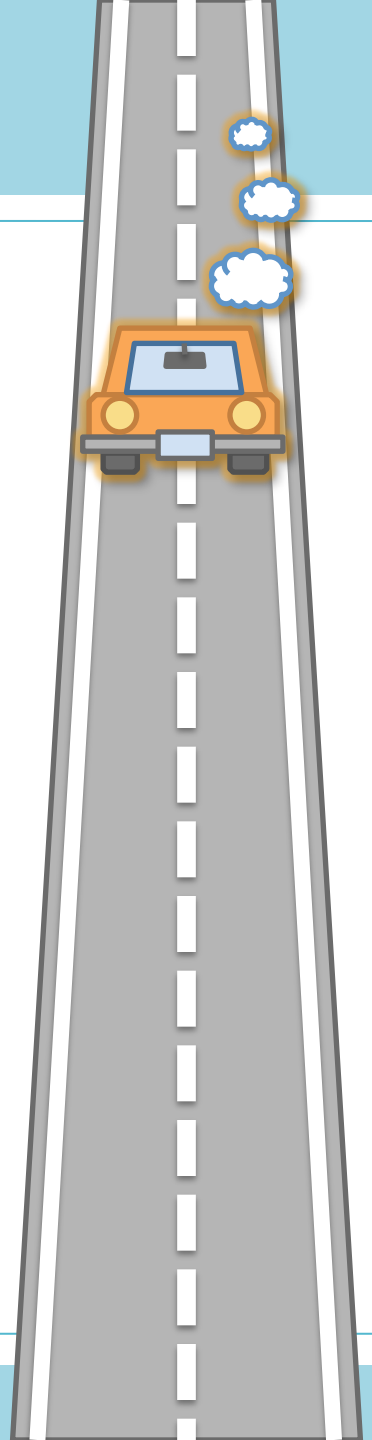
NO magic numbers !

Parameter-free!



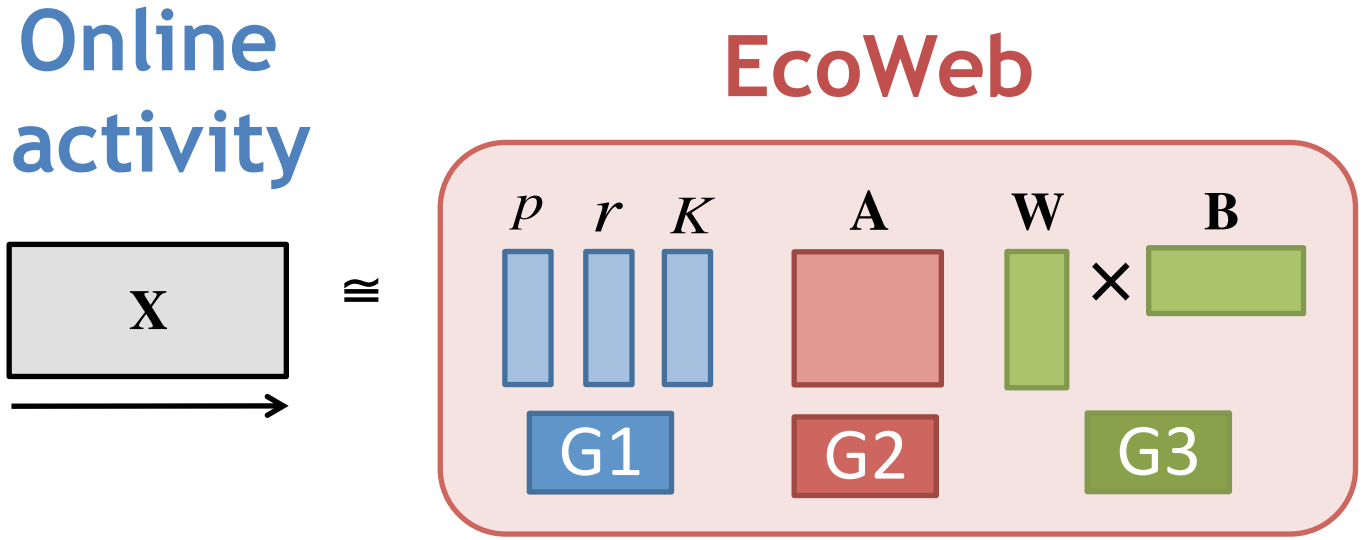
Roadmap

- ✔ Motivation
- Modeling power of EcoWeb
- Overview
- Proposed model
- Algorithm
- Experiments
- EcoWeb – at work
- Conclusions



Modeling power of EcoWeb

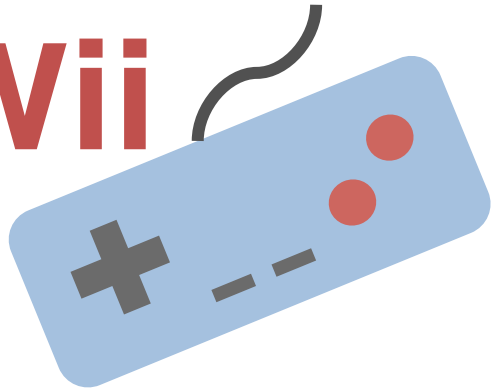
Questions



Modeling power of EcoWeb

Q1-1. (games)
Who is the competitor?

Wii

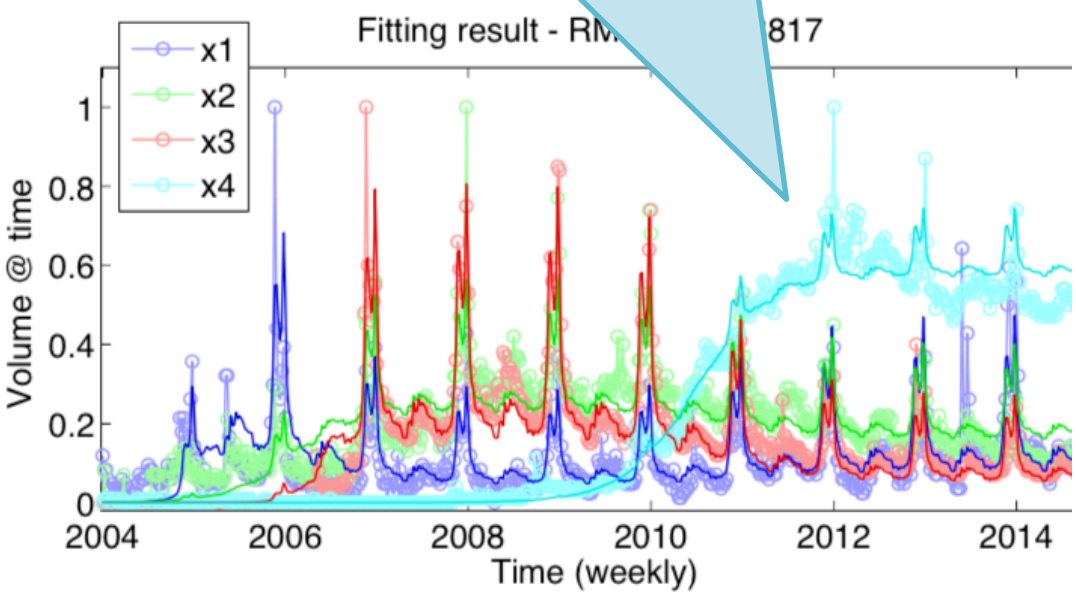


vs.

?

Modeling power of EcoWeb

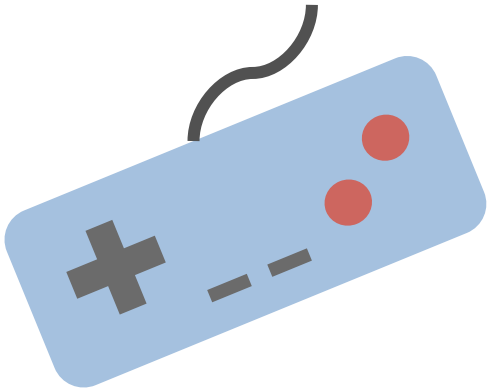
A. Android!



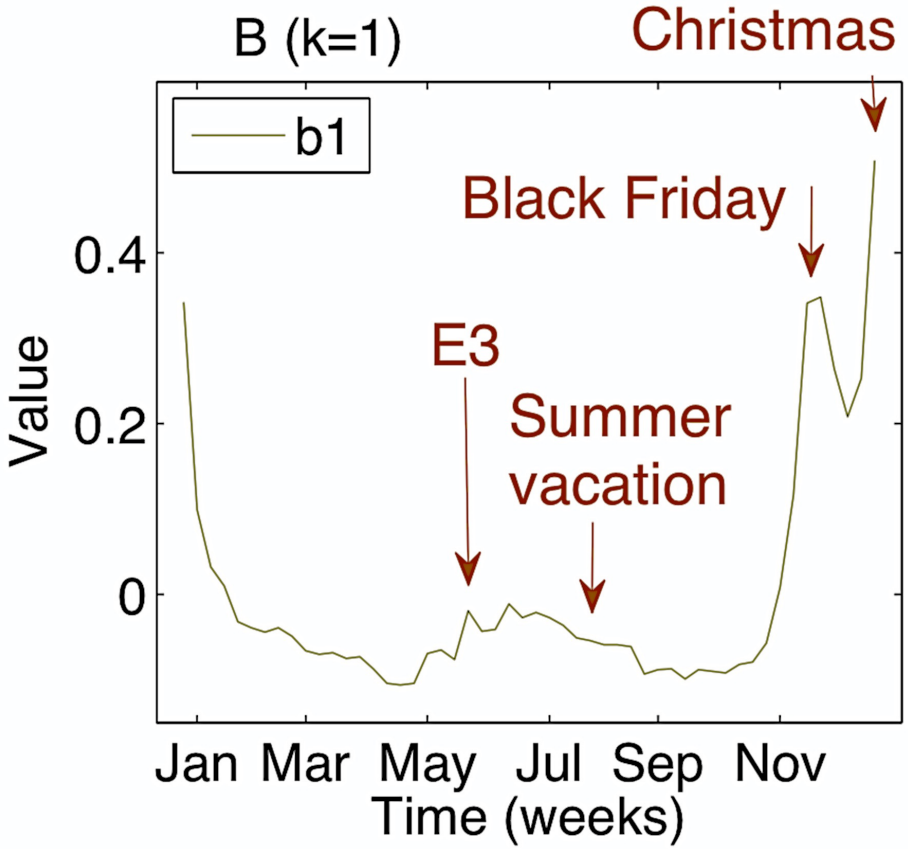
EcoWeb: Interaction network

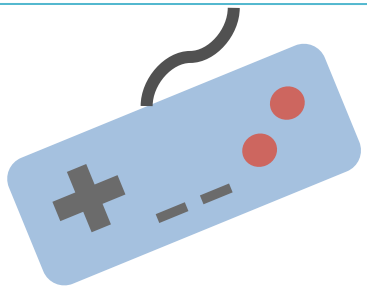
Modeling power of EcoWeb

Q1-2. (games)
Any seasonal events?



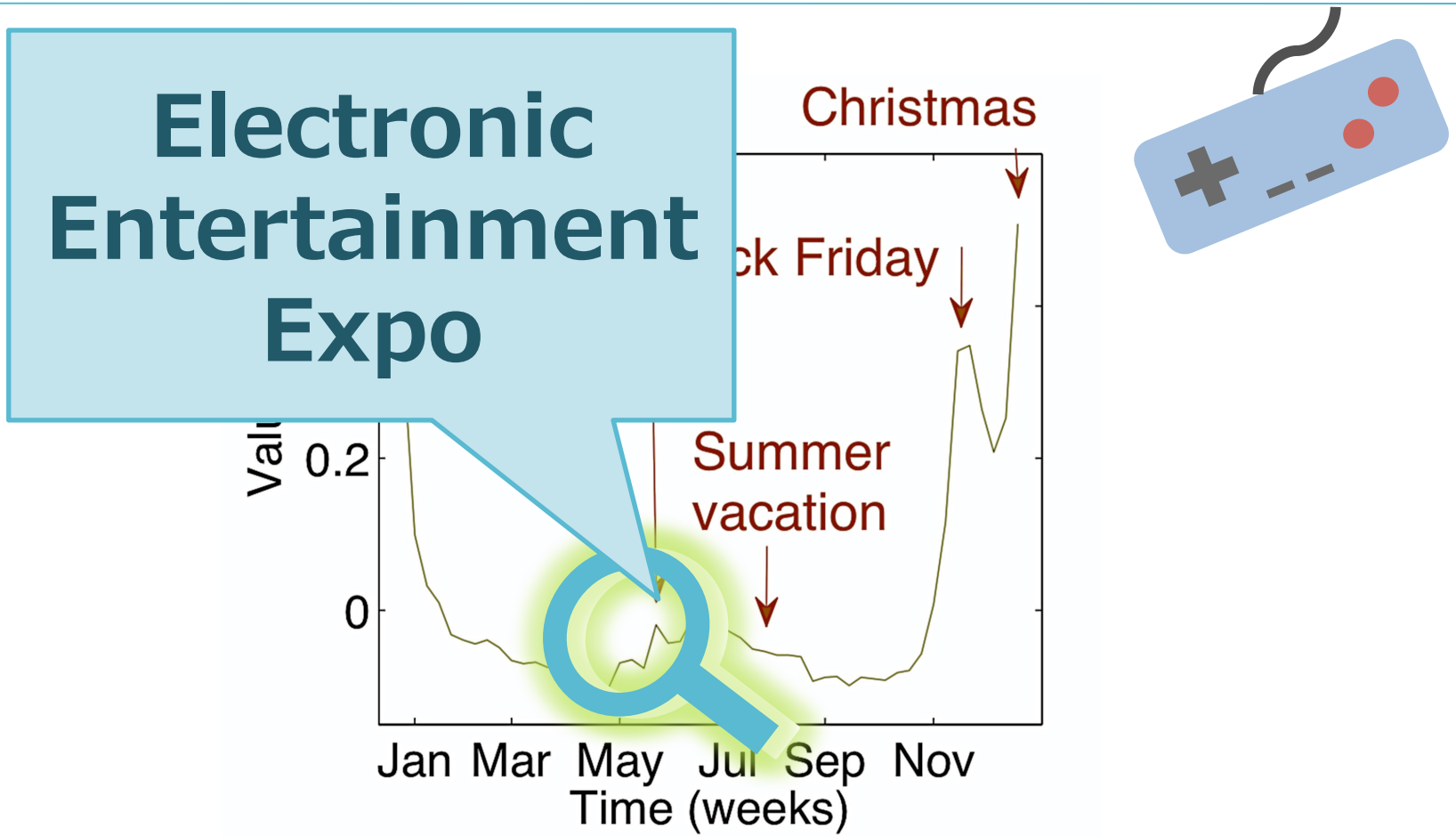
Modeling power of EcoWeb





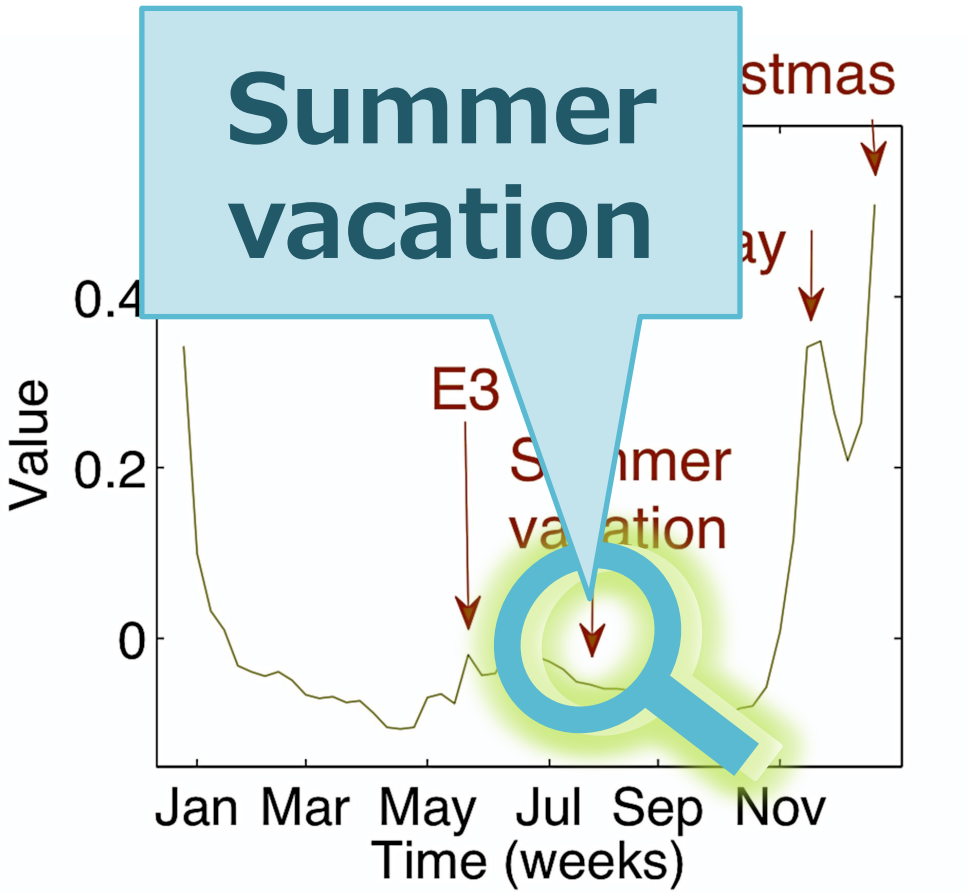
EcoWeb: seasonal component

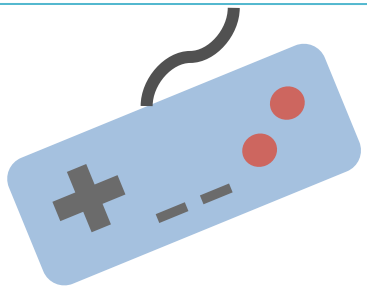
Modeling power of EcoWeb



EcoWeb: seasonal component

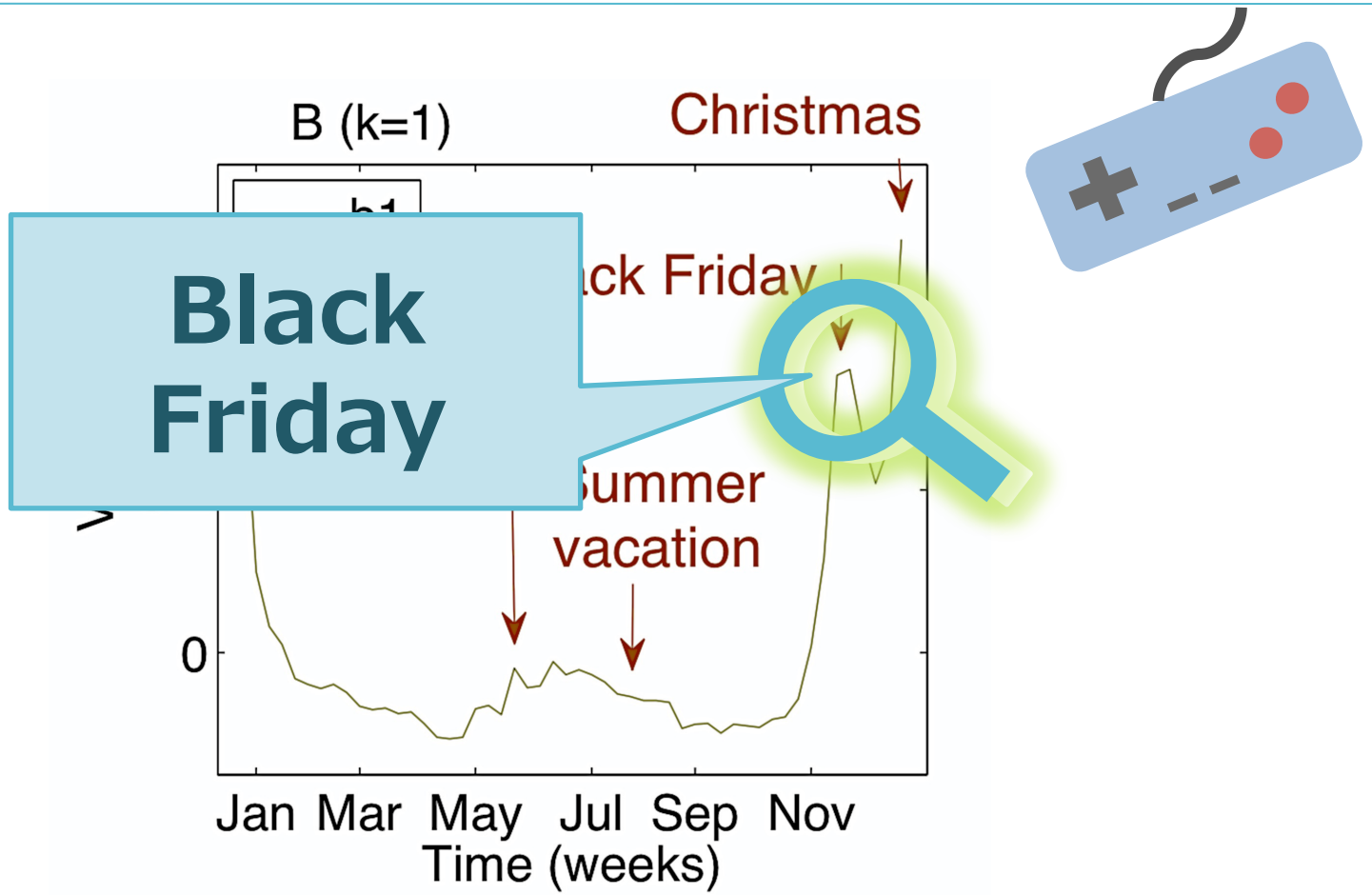
Modeling power of EcoWeb





EcoWeb: seasonal component

Modeling power of EcoWeb

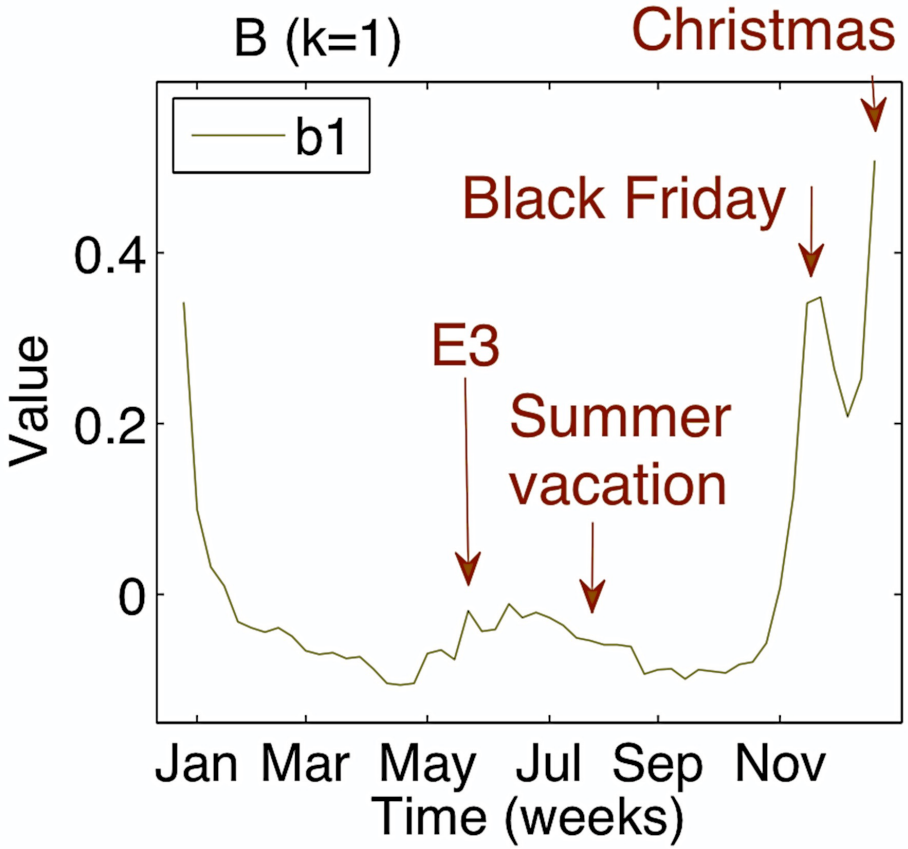


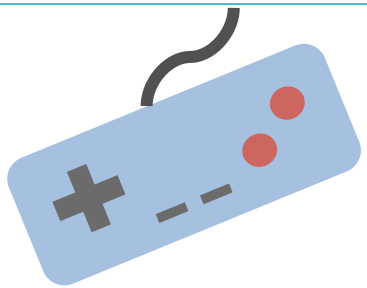
EcoWeb: seasonal component

Modeling power of EcoWeb

EcoWeb: seasonal component

Modeling power of EcoWeb





EcoWeb: seasonal component

Modeling power of EcoWeb

Q2-1. (apparels)
Who is the competitor?

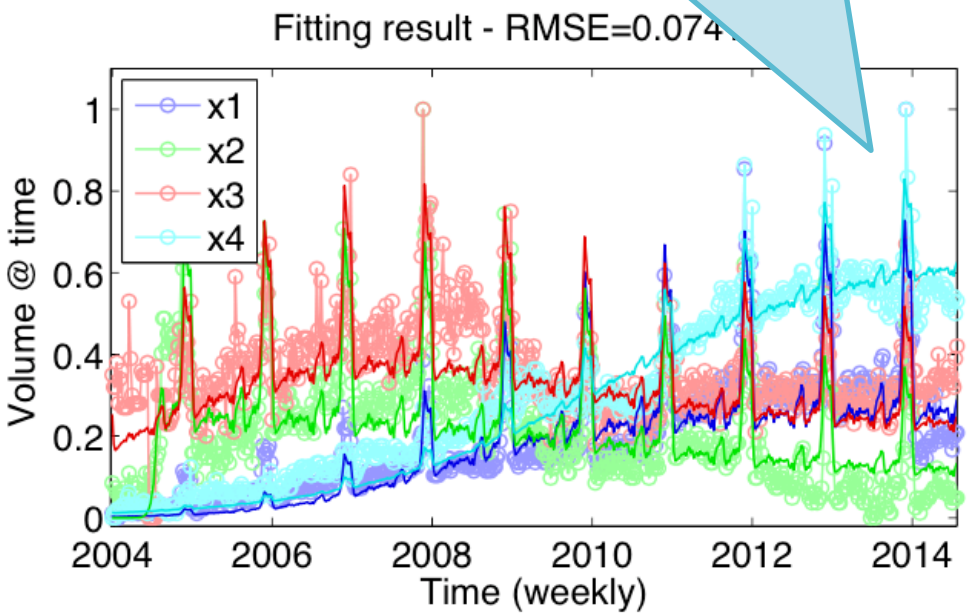
JCPenny

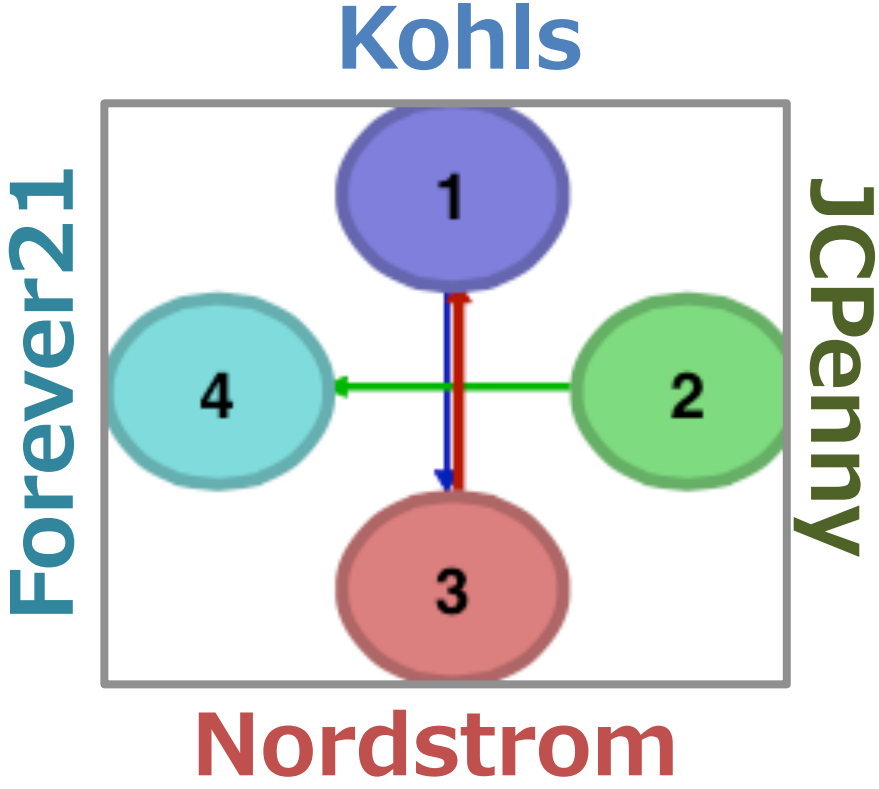
vs.

?

Modeling power of EcoWeb

A2. Forever21!



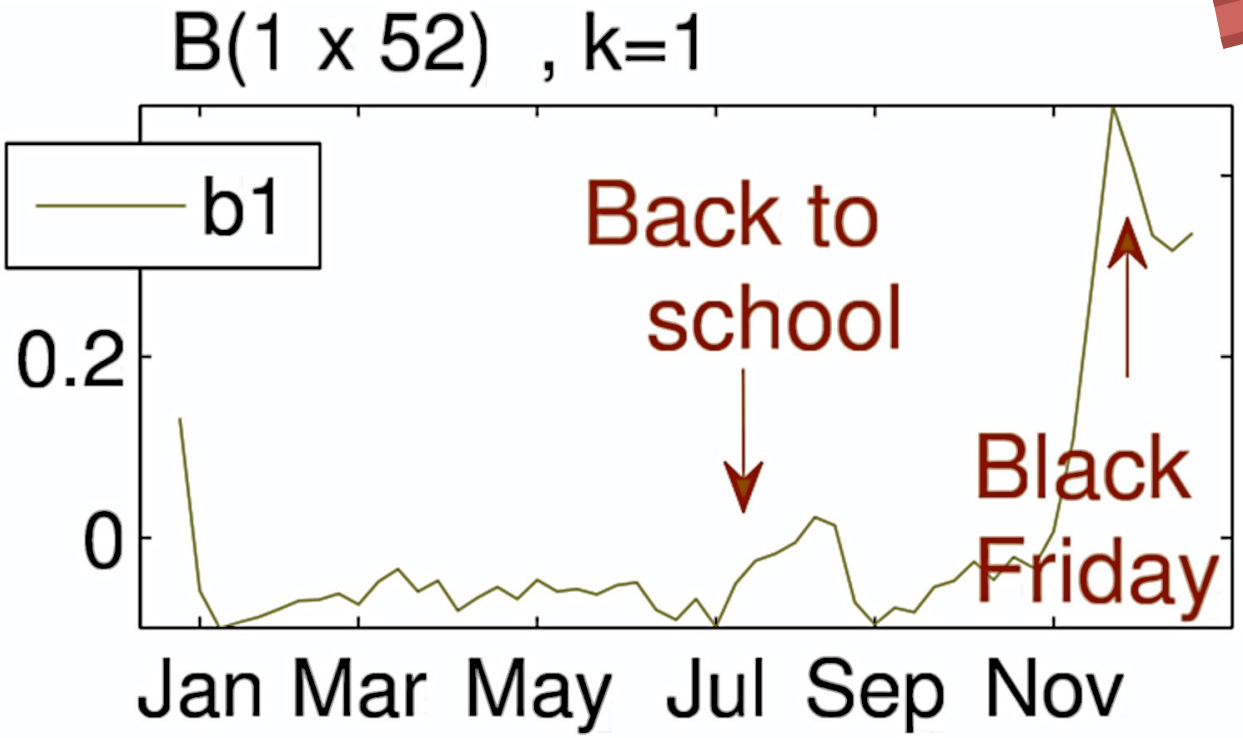


EcoWeb: Interaction network

Modeling power of EcoWeb

Q2-2. (apparels)
Any seasonal events?

Modeling power of EcoWeb



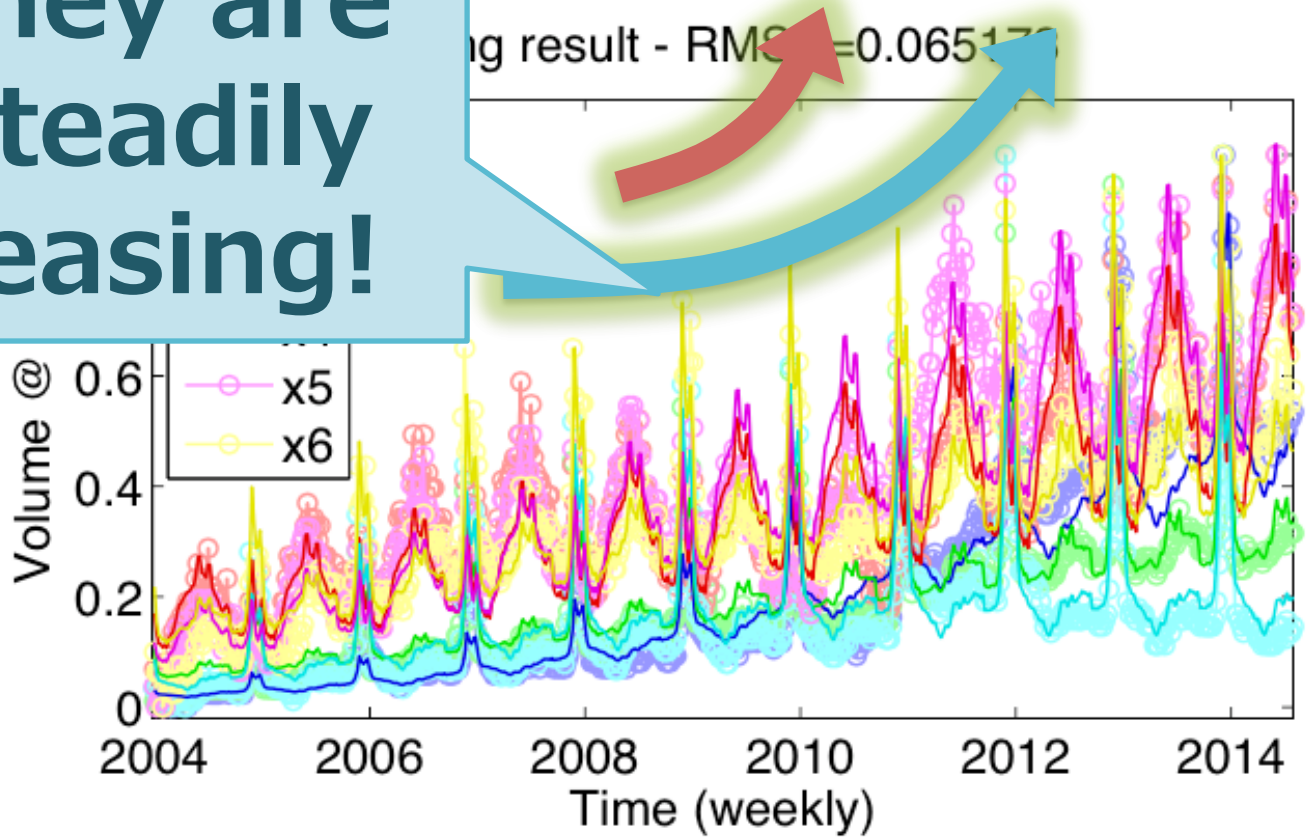
EcoWeb: seasonal component

Modeling power of EcoWeb

Q3. (retails)
Any patterns/trends?

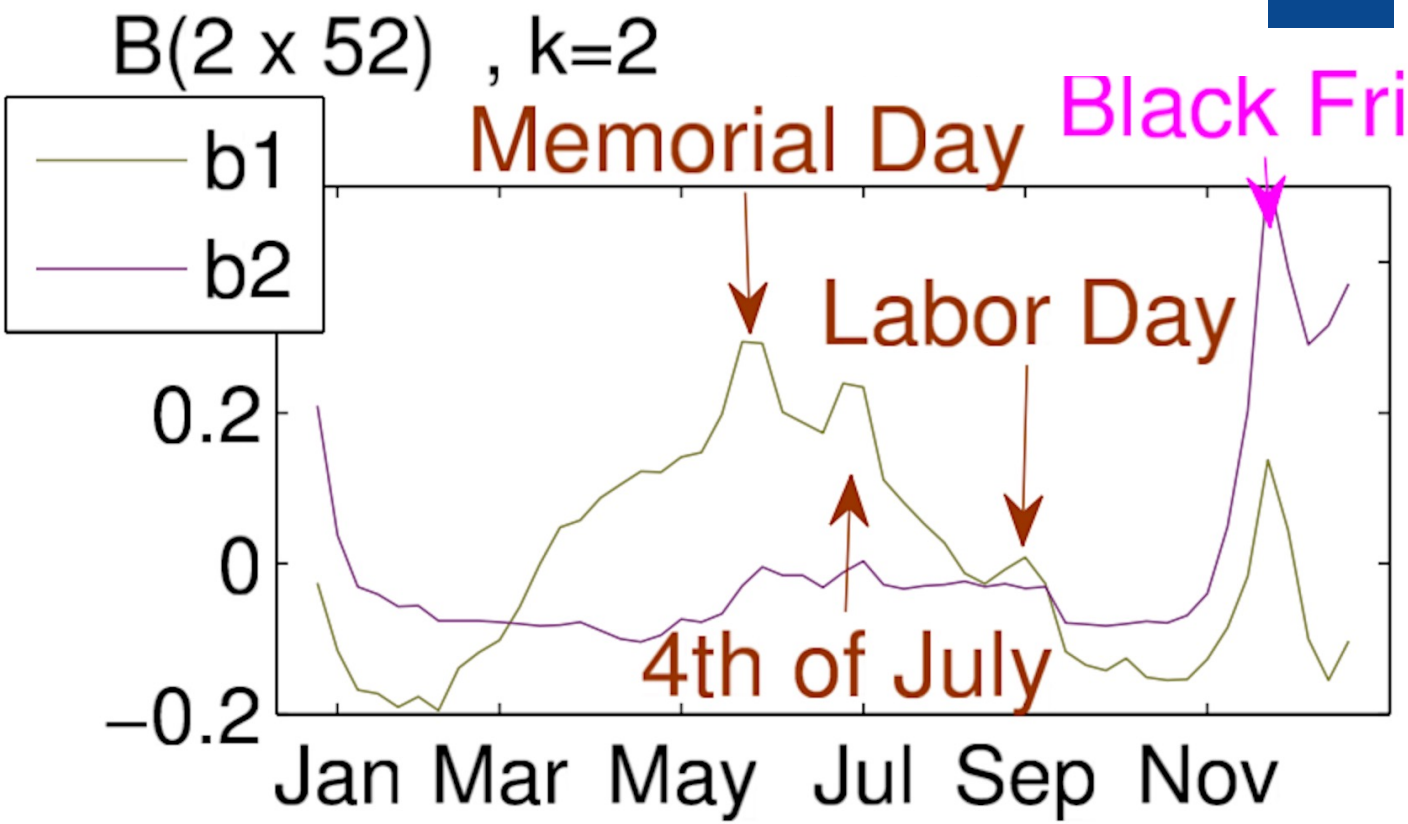
Modeling power of EcoWeb

A. They are all steadily increasing!



Amazon, Walmart, Home Depot, Best buy, …

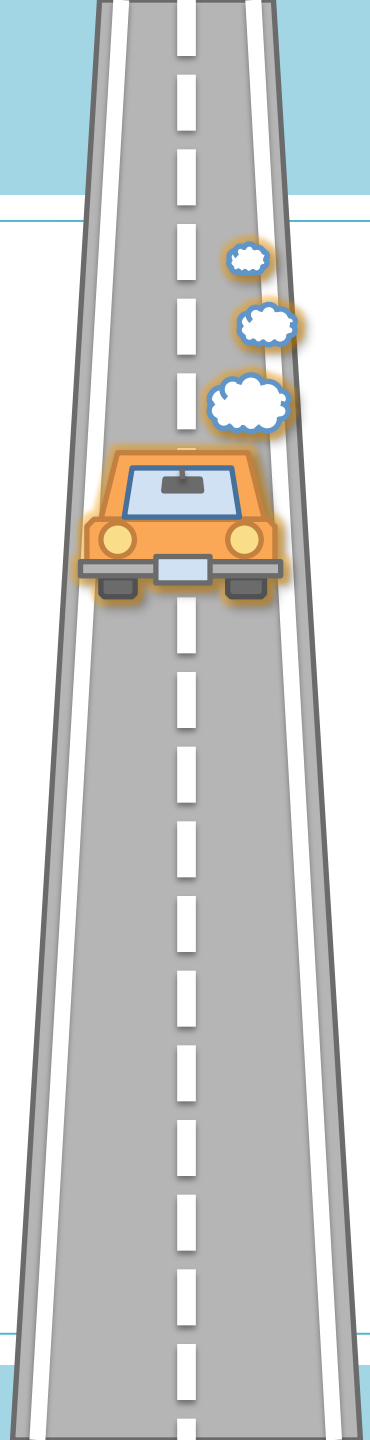
Modeling power of EcoWeb



EcoWeb: 2 seasonal components

Roadmap

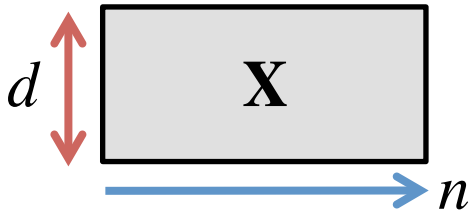
- ✔ Motivation
- ✔ Modeling power of EcoWeb
  - Overview
  - Proposed model
  - Algorithm
  - Experiments
  - EcoWeb – at work
  - Conclusions



Problem definition

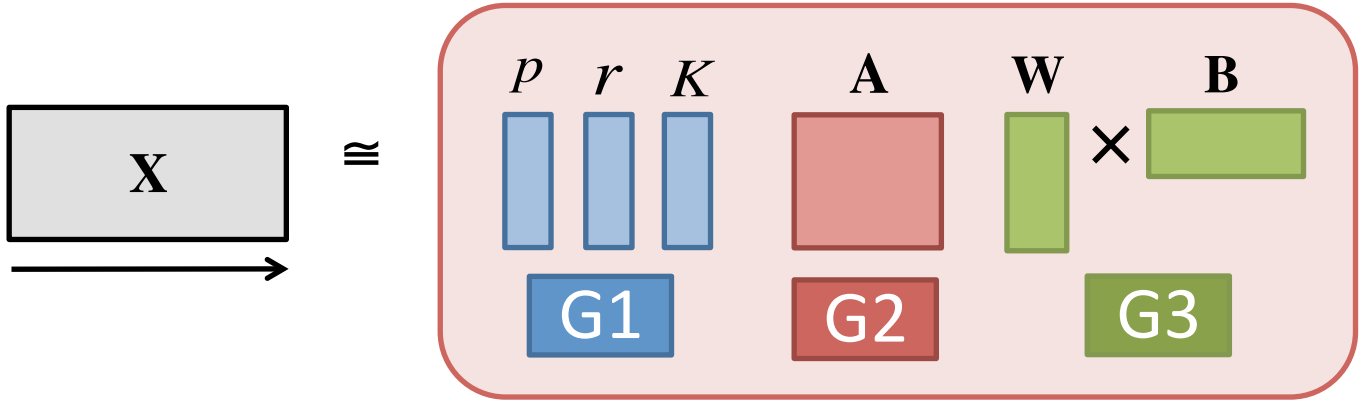
Given: Co-evolving online activities

X (activity x time)



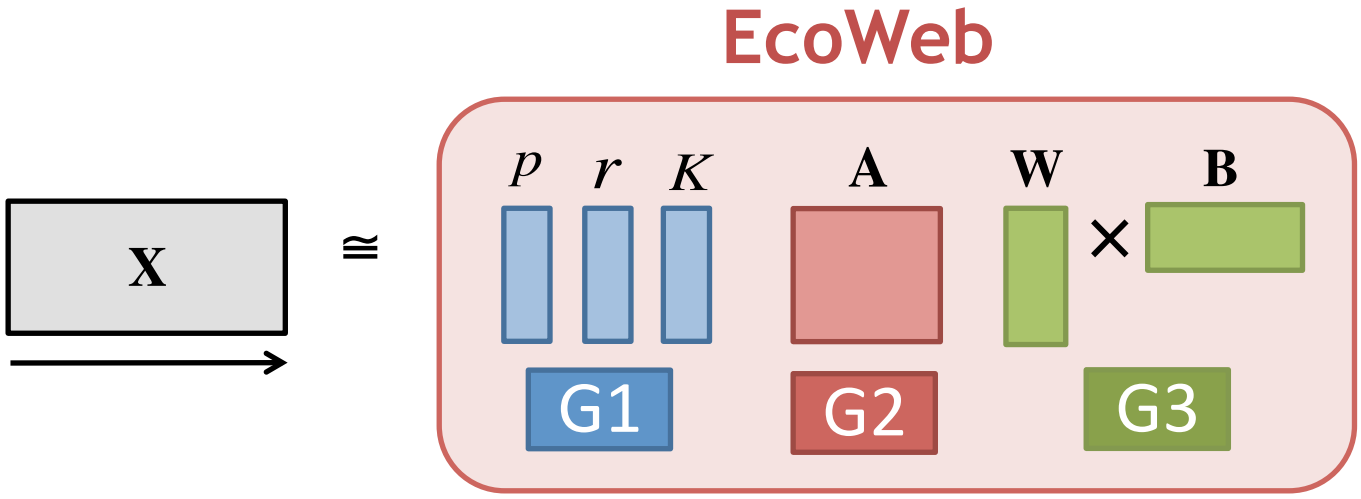
Find: Compact description of X

EcoWeb



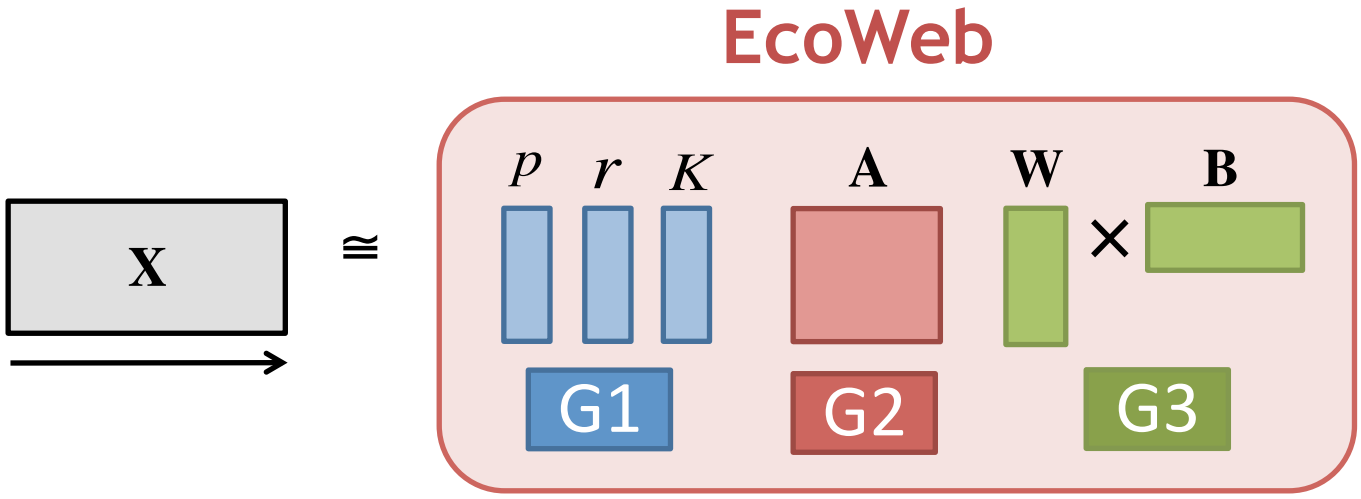
EcoWeb: Main idea

Q. How can we describe the evolutions of X ?



EcoWeb: Main idea

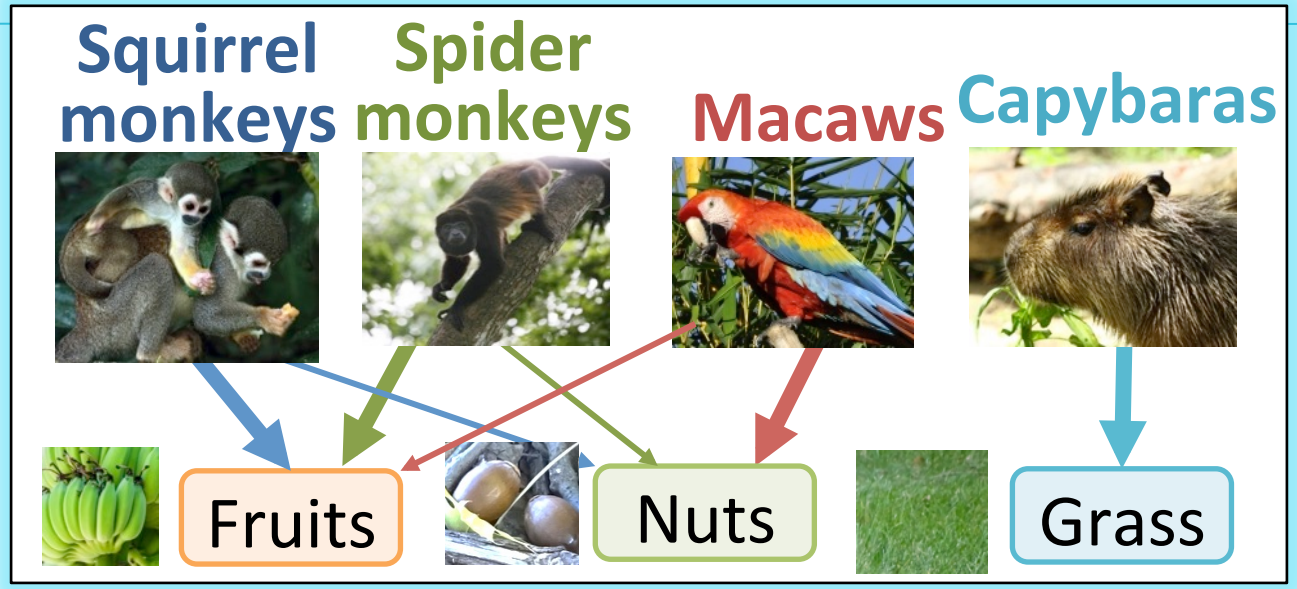
Q. How can we describe the evolutions of X ?



A. Web as a jungle!

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

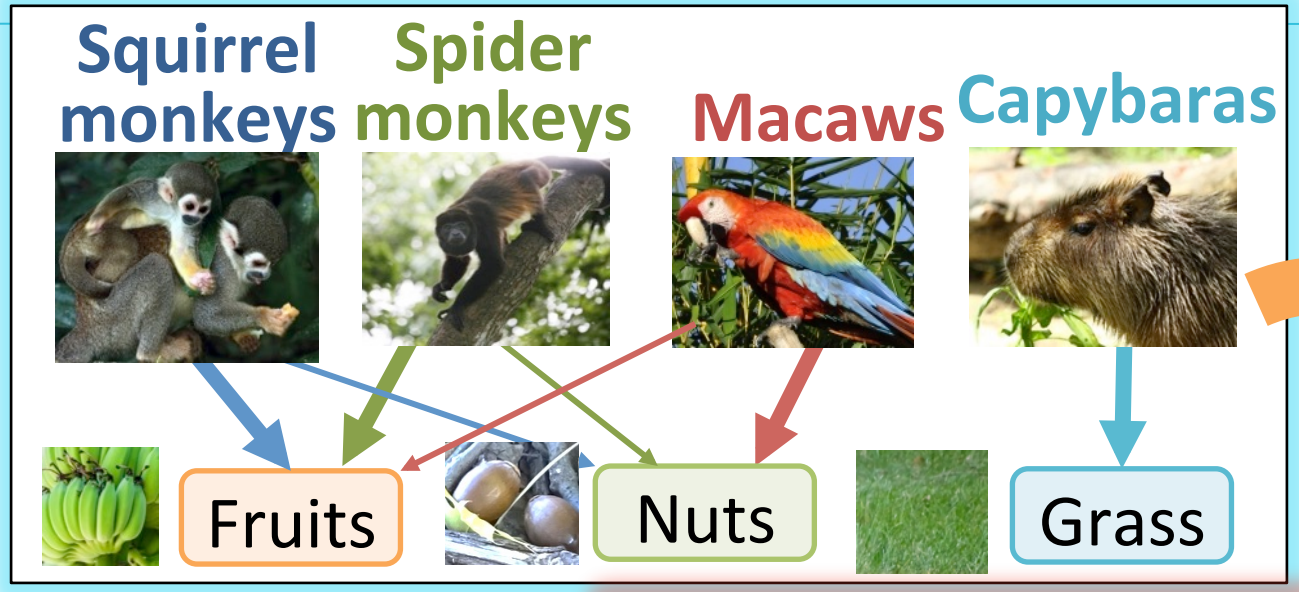
Main idea: Web as a jungle



Ecosystem in the **Jungle**

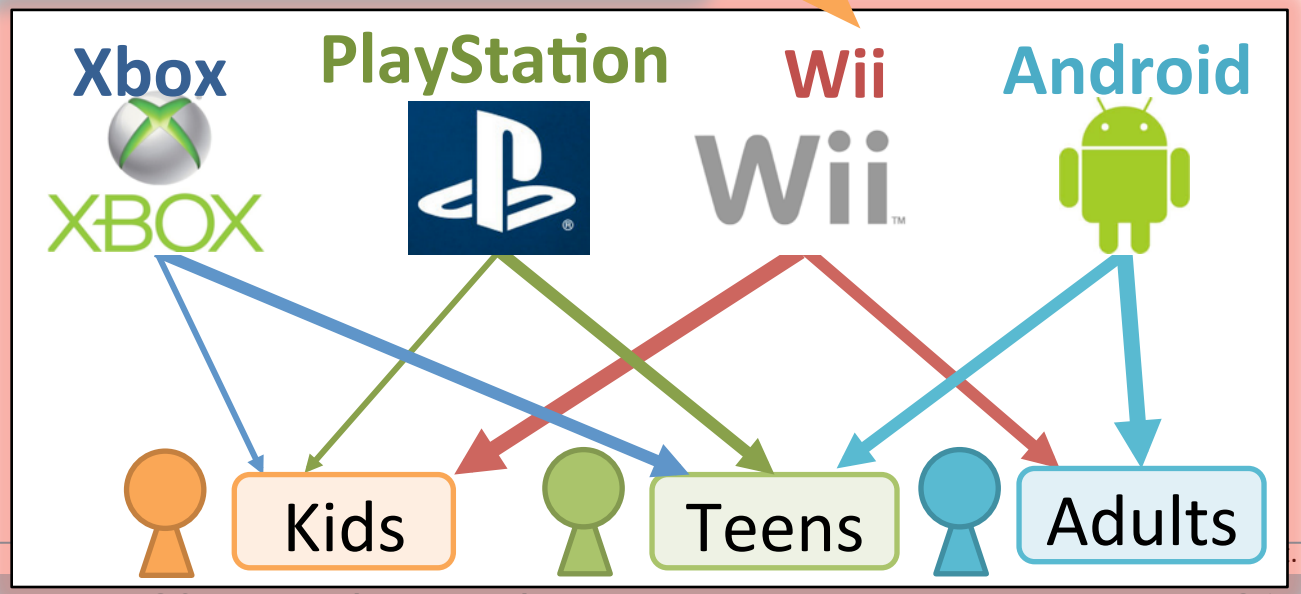
Image courtesy of xura, criminalatt, David Castillo Dominici, happykanppy at FreeDigitalPhotos.net.

Main idea: Web as a jungle



Ecosystem in the **Jungle**

Ecosystem on the **Web**



Analogies: ecosystem on the Web

Biological species

Online activities

Jungle

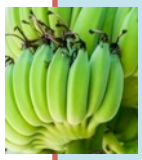
Web

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

Analogies: ecosystem on the Web

Biological species

Food resources



Jungle

Online activities

User resources

Web

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

Analogies: ecosystem on the Web

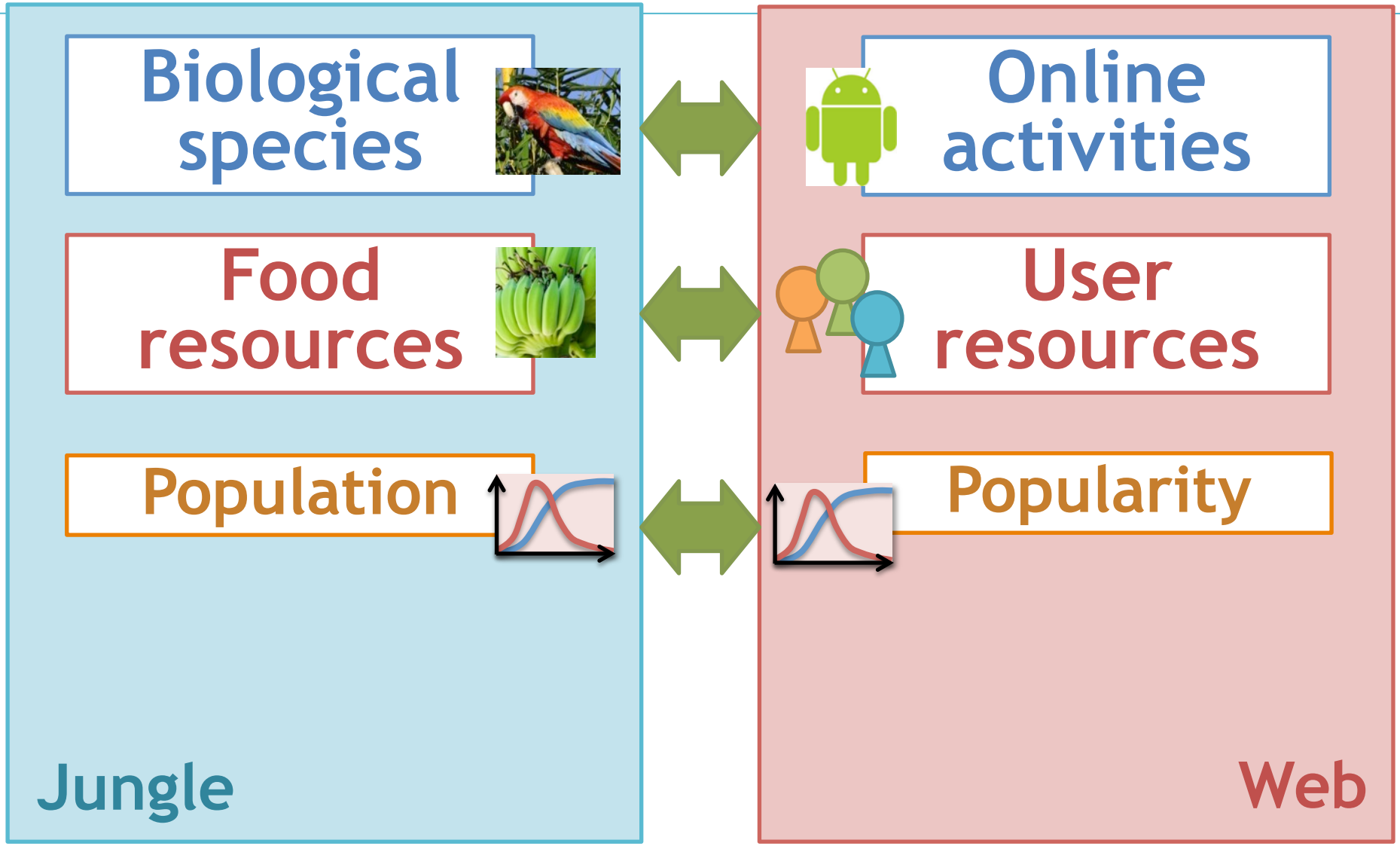


Image courtesy of xura, criminalatt, David Castillo Dominici, happykanppy at FreeDigitalPhotos.net.

Analogies: ecosystem on the Web

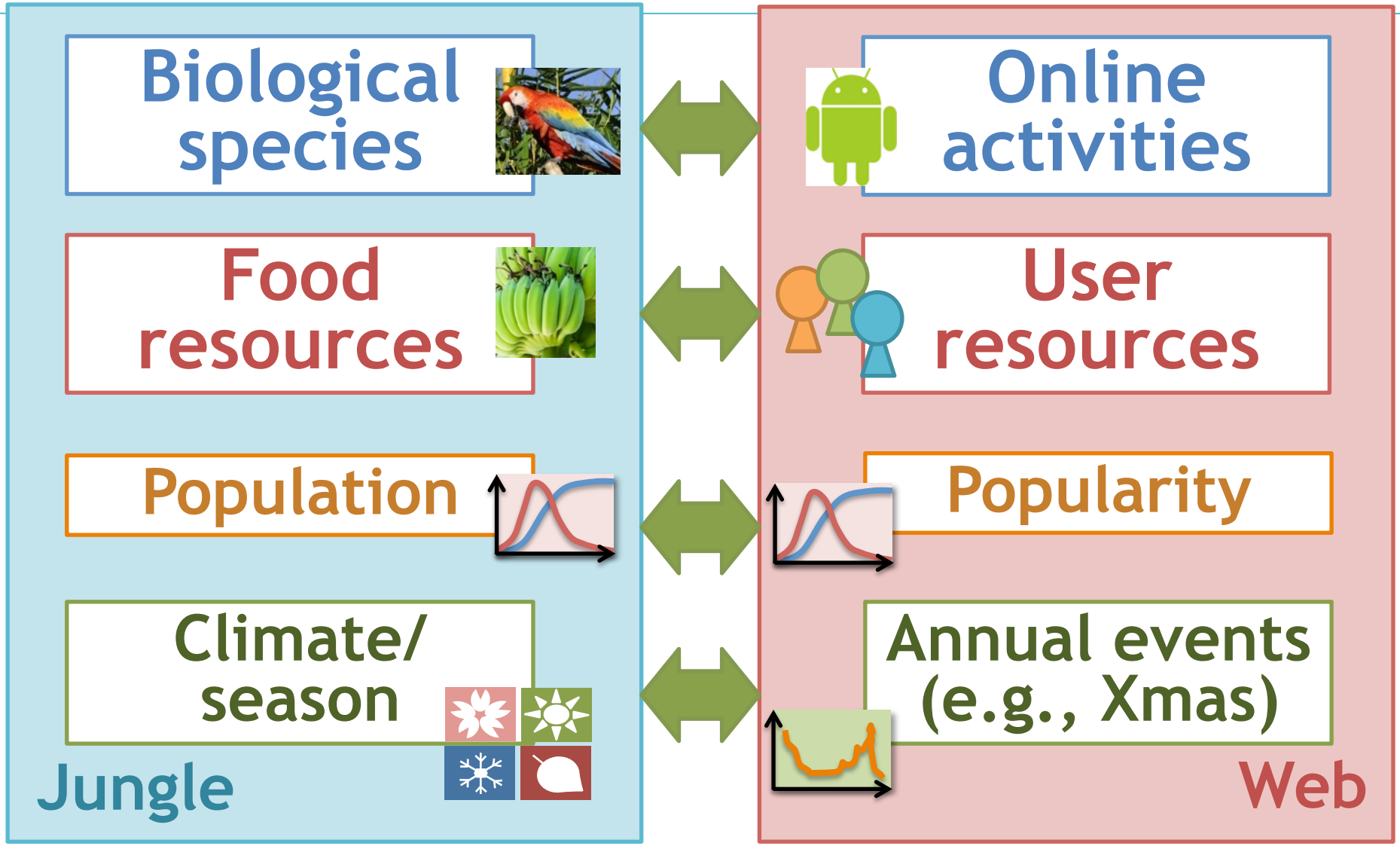
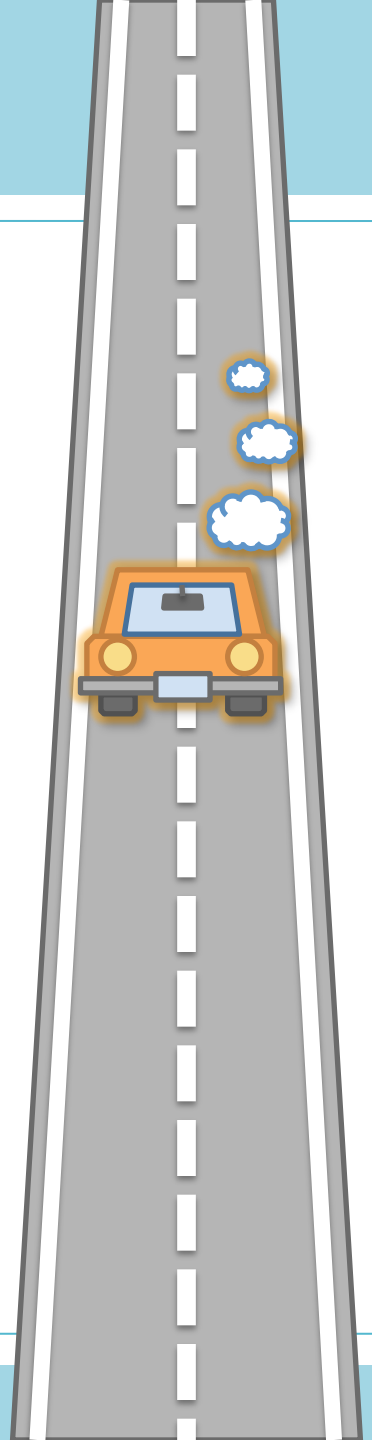


Image courtesy of xura, criminalatt, David Castillo Dominici, happykanppy at FreeDigitalPhotos.net.

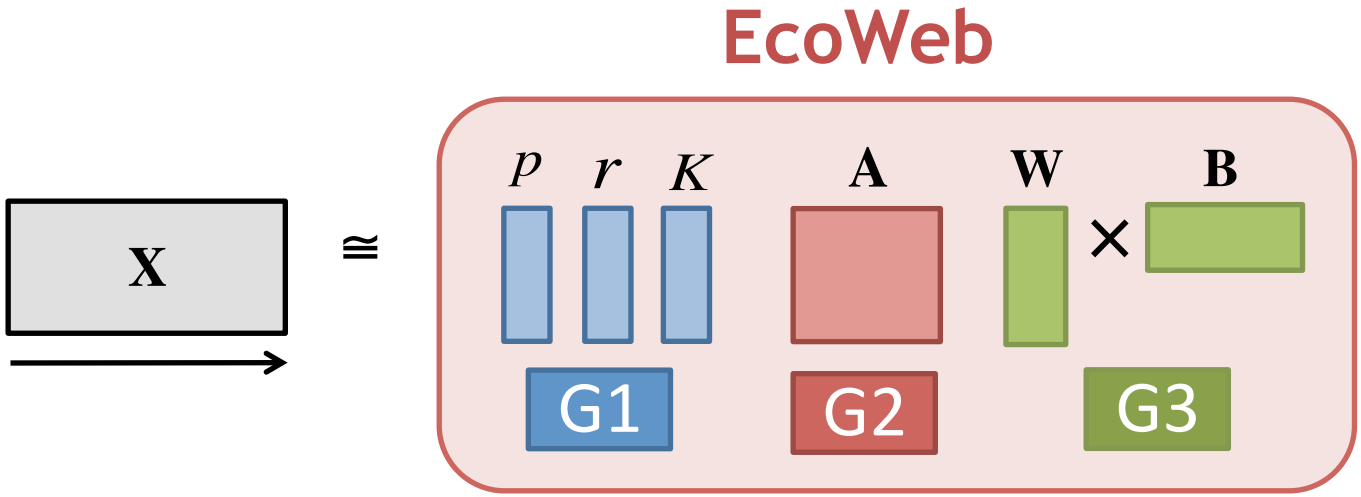
Roadmap

- ✔ Motivation
- ✔ Modeling power of EcoWeb
- ✔ Overview
  - Proposed model
  - Algorithm
  - Experiments
  - EcoWeb – at work
  - Conclusions



EcoWeb: Main idea

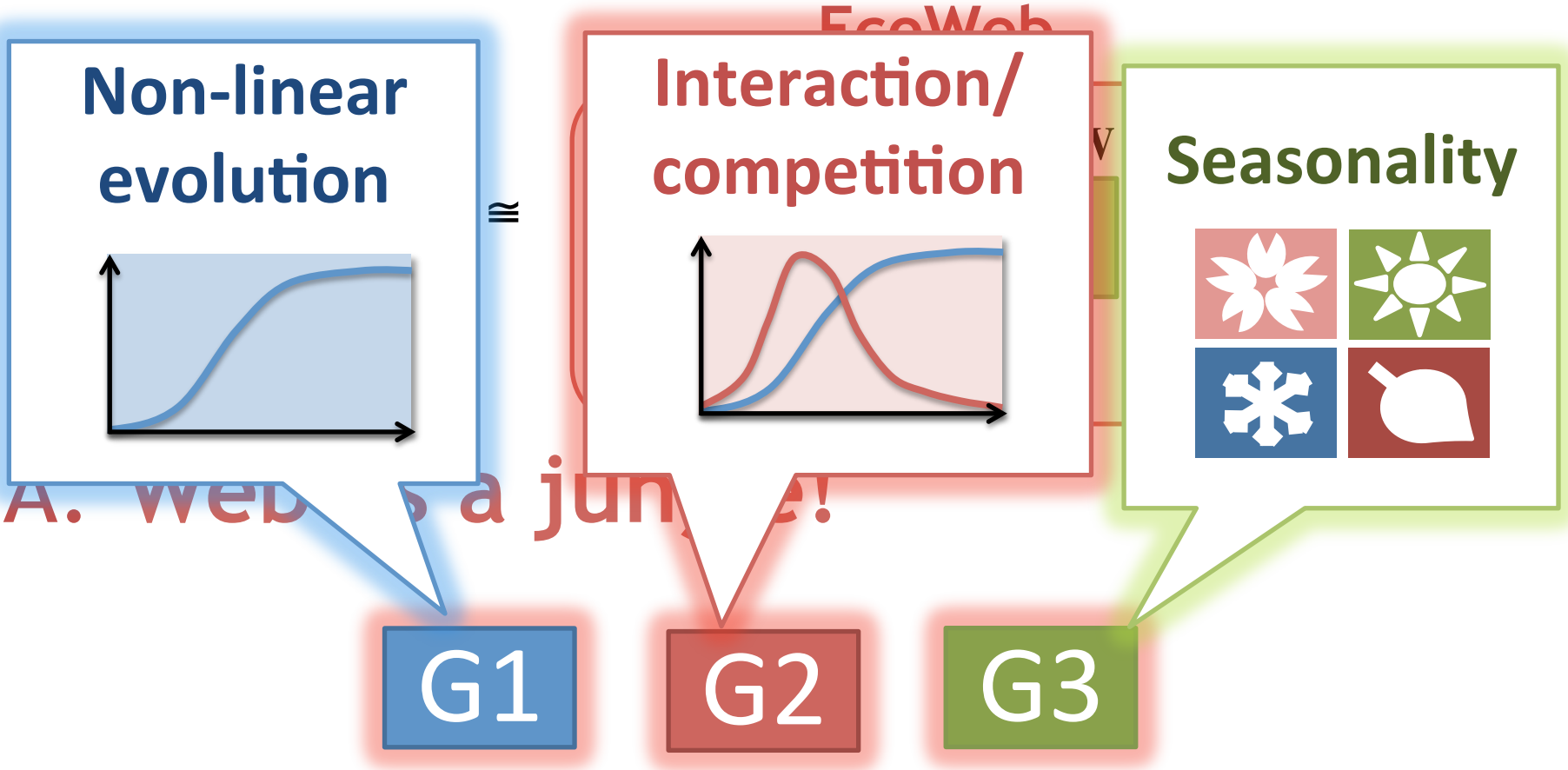
Q. How can we describe the evolutions of X ?



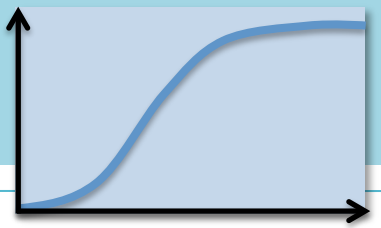
A. Web as a jungle!

EcoWeb: Main idea

Q. How can we describe the evolutions of X ?



G1 EcoWeb-individual



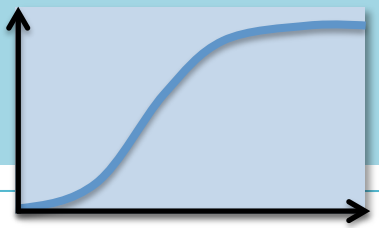
Non-linear evolution of a single keyword

Species

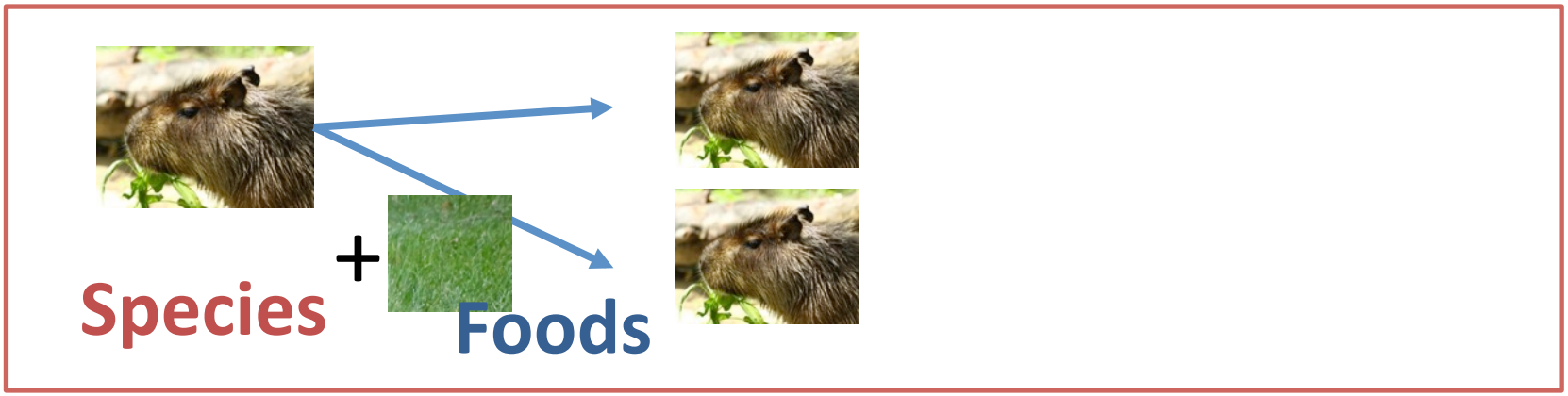
Keywords

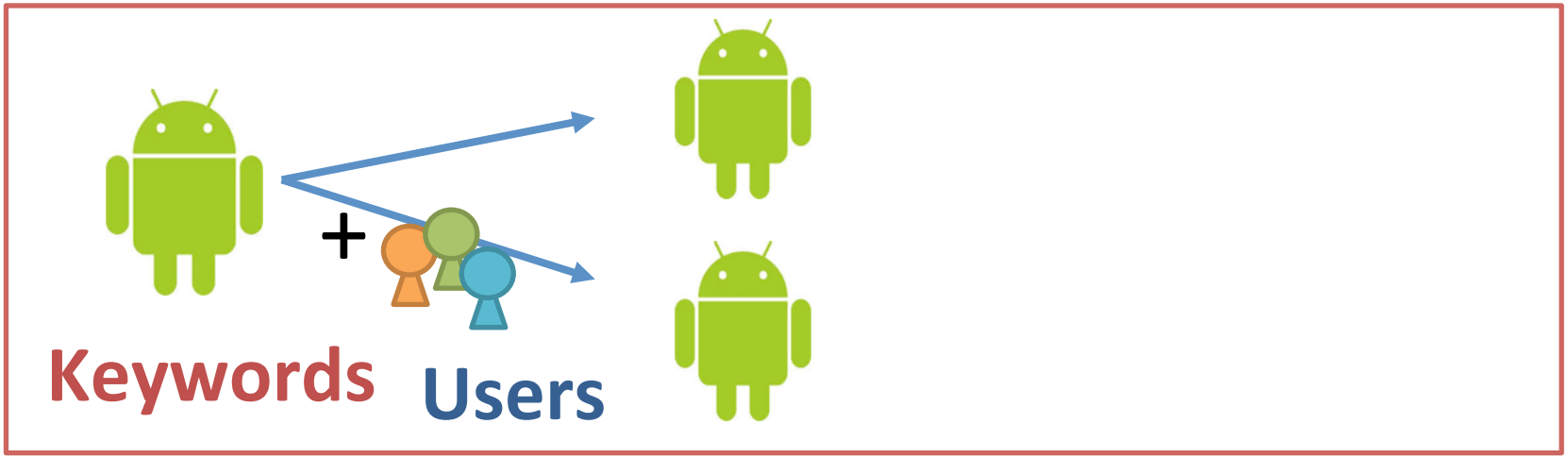
t=0 t=1 t=2

G1 EcoWeb-individual



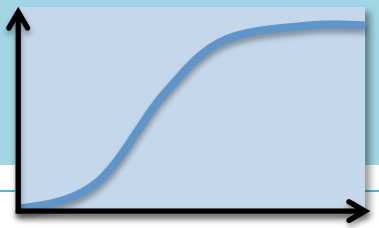
Non-linear evolution of a single keyword



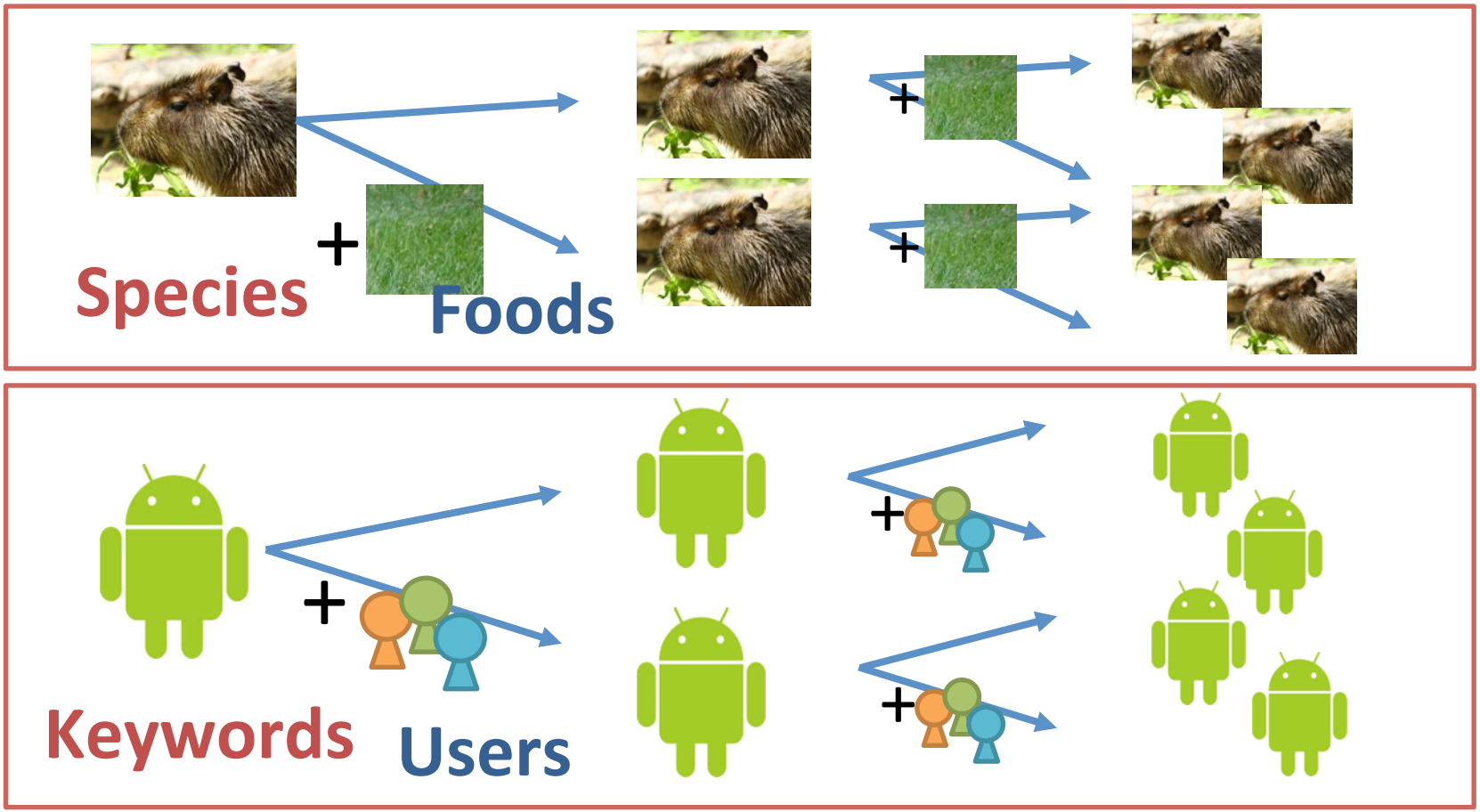


t=0 t=1 t=2

G1 EcoWeb-individual



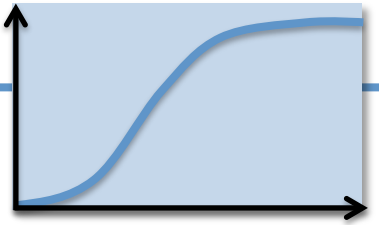
Non-linear evolution of a single keyword



G1 EcoWeb-individual

Details

Non-linear evolution of a single keyword



Popularity size

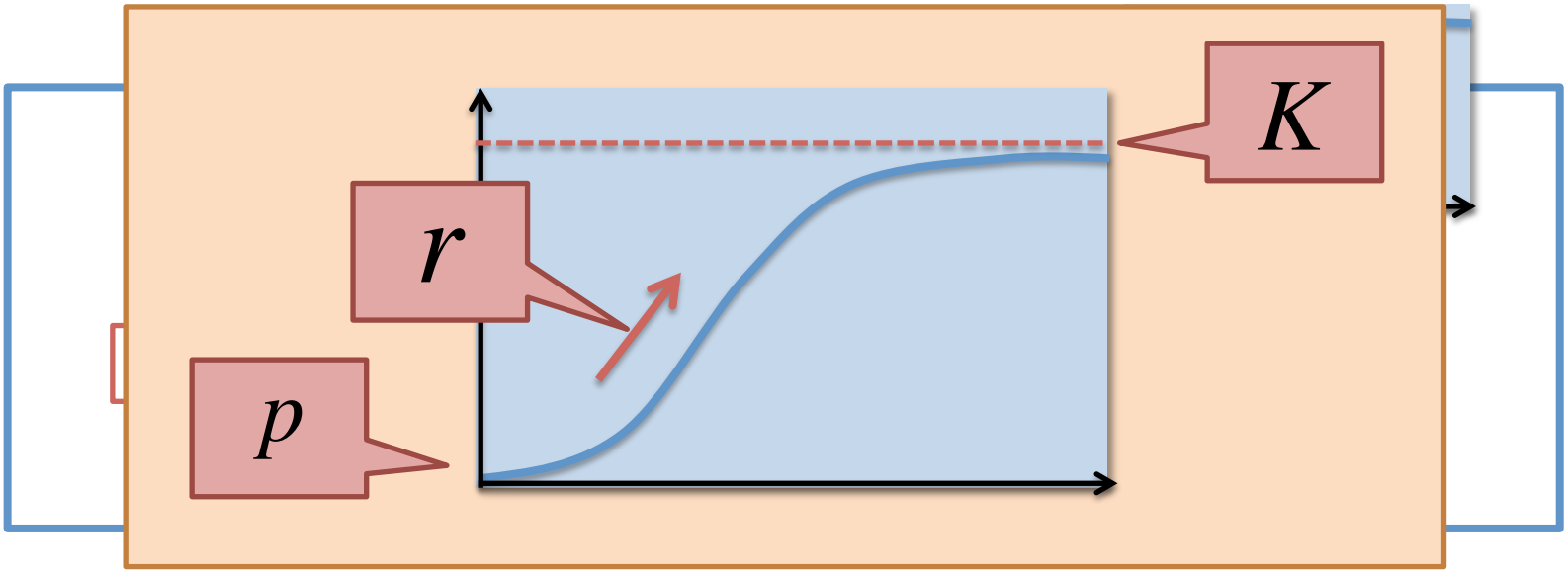
$$P(t+1) = P(t)\left[1 + r\left(1 - \frac{P(t)}{K}\right)\right],$$

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

G1 EcoWeb-individual

Details

Non-linear evolution of a single keyword



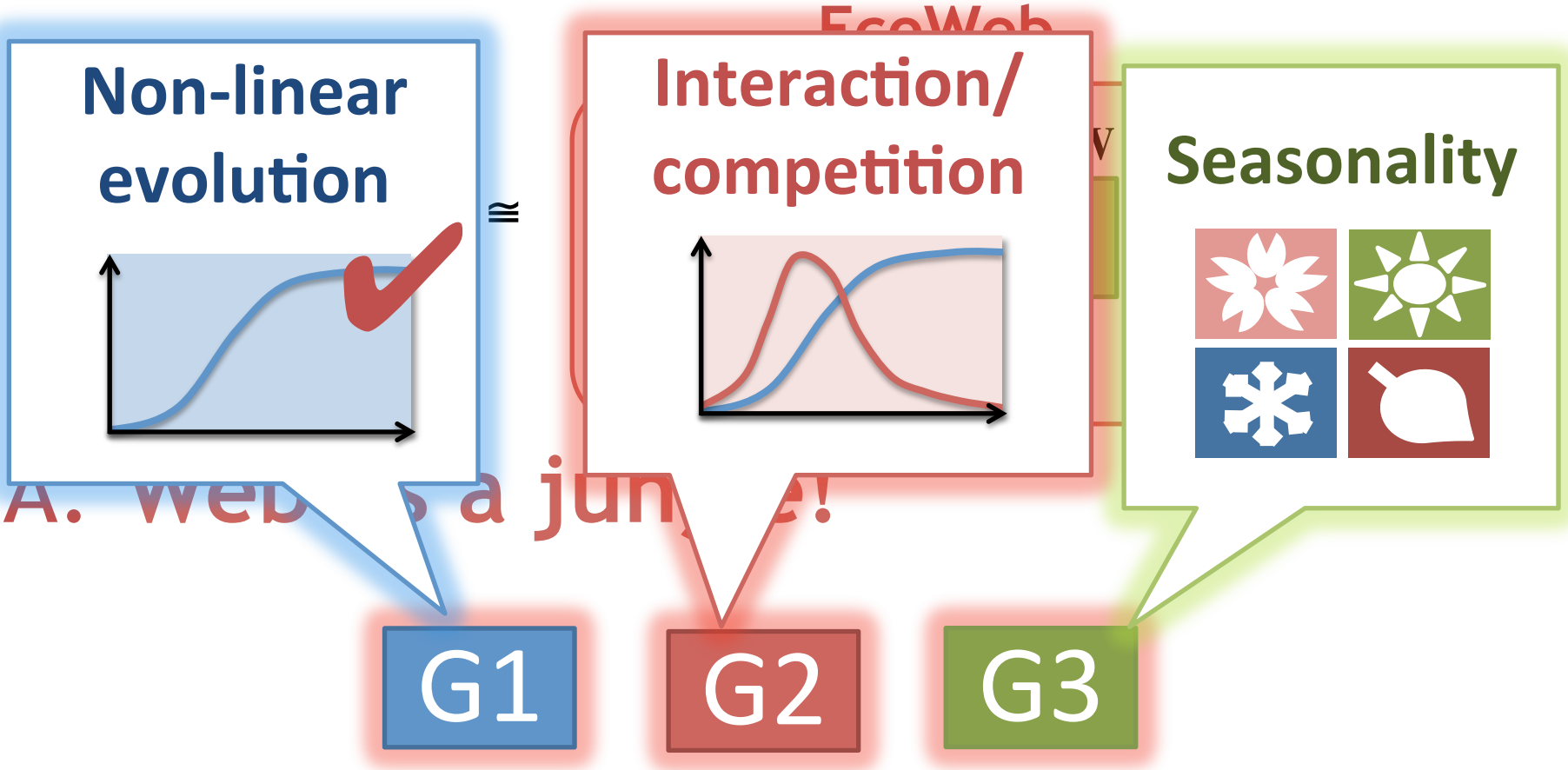
p – Initial condition (i.e., P(0) =p)

r – Growth rate, attractiveness

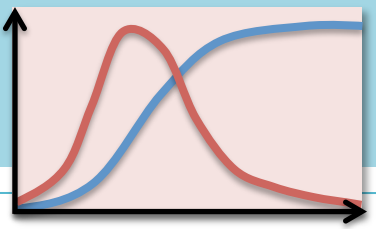
K – Carrying capacity (=available user resources)

EcoWeb: Main idea

Q. How can we describe the evolutions of X ?

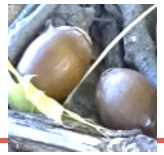


G2 EcoWeb-interaction



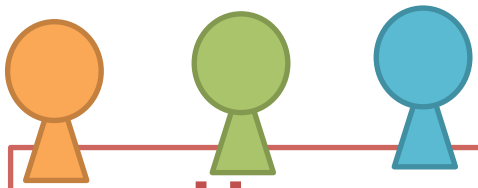
Interaction between multiple keywords

Species



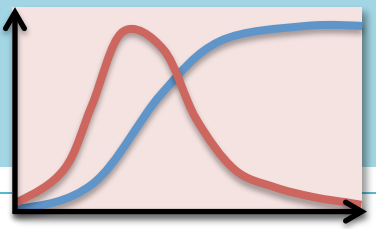
Food resources

Keywords

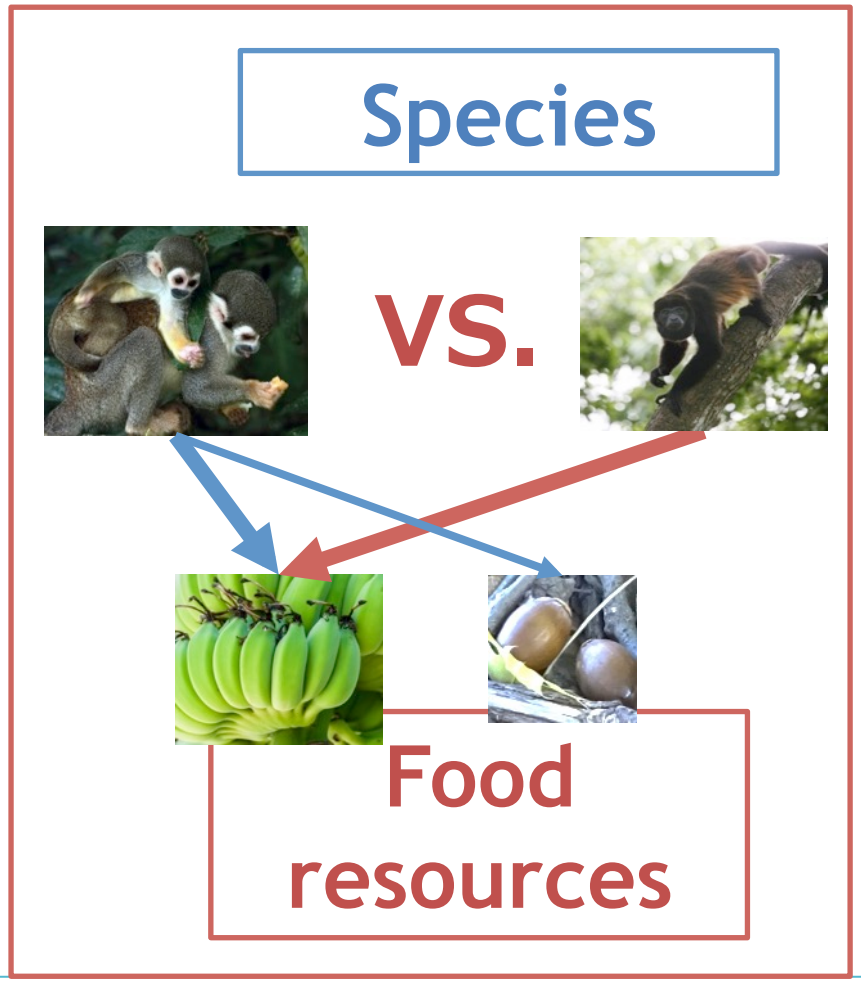


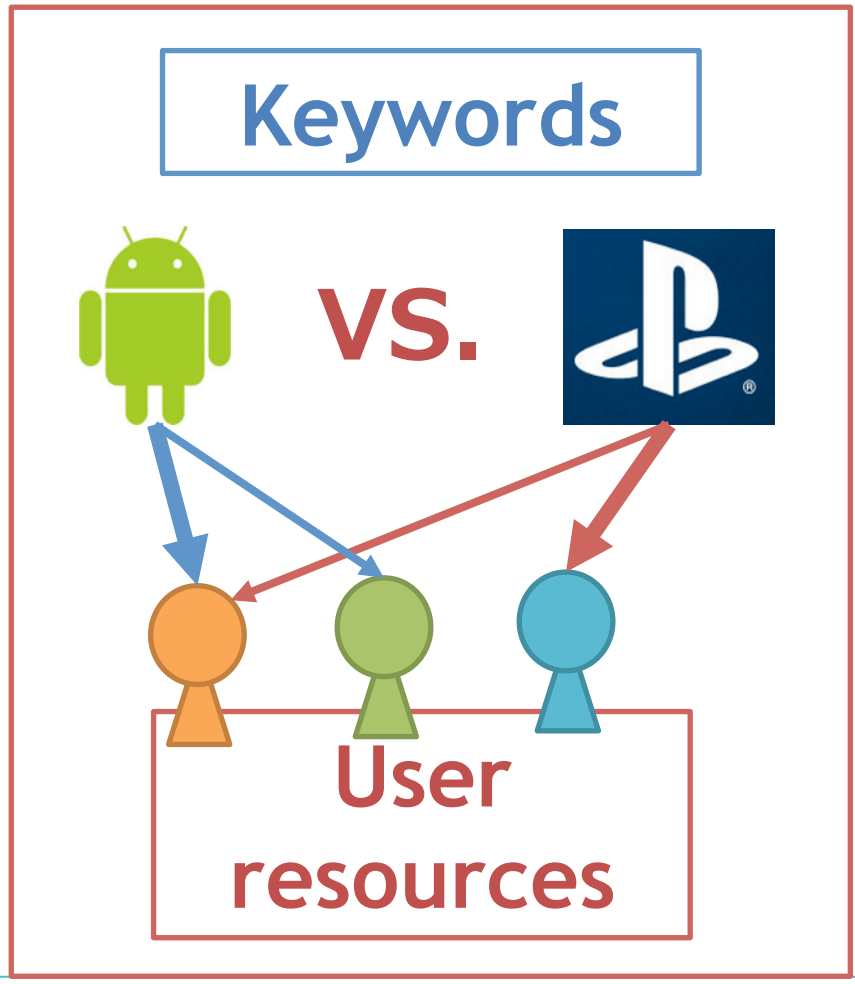
User resources

G2 EcoWeb-interaction



Interaction between multiple keywords





G2 EcoWeb-interaction

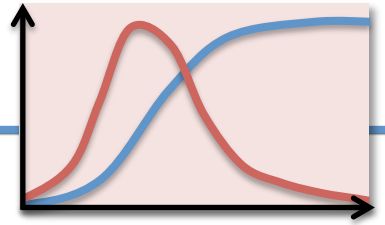
Details

Interaction between multiple keywords

Popularity of (i) Popularity of (j)

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

$$(i = 1, \cdots, d), \quad (3)$$



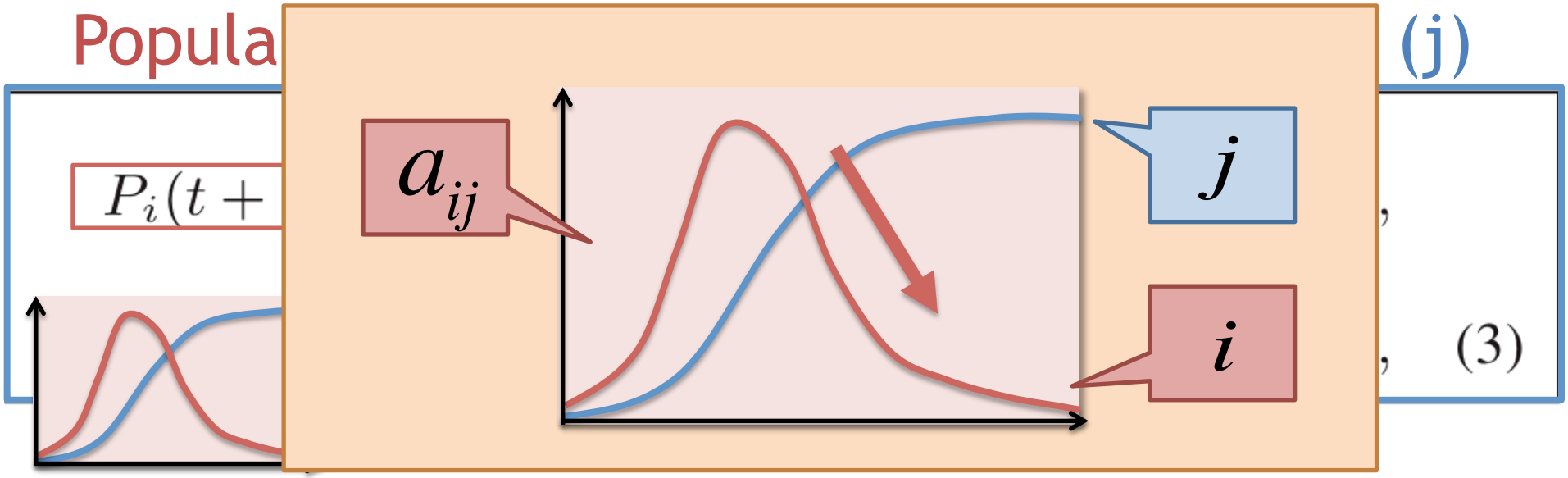
a_{ij} – Interaction coefficient

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

G2 EcoWeb-interaction

Details

Interaction between multiple keywords

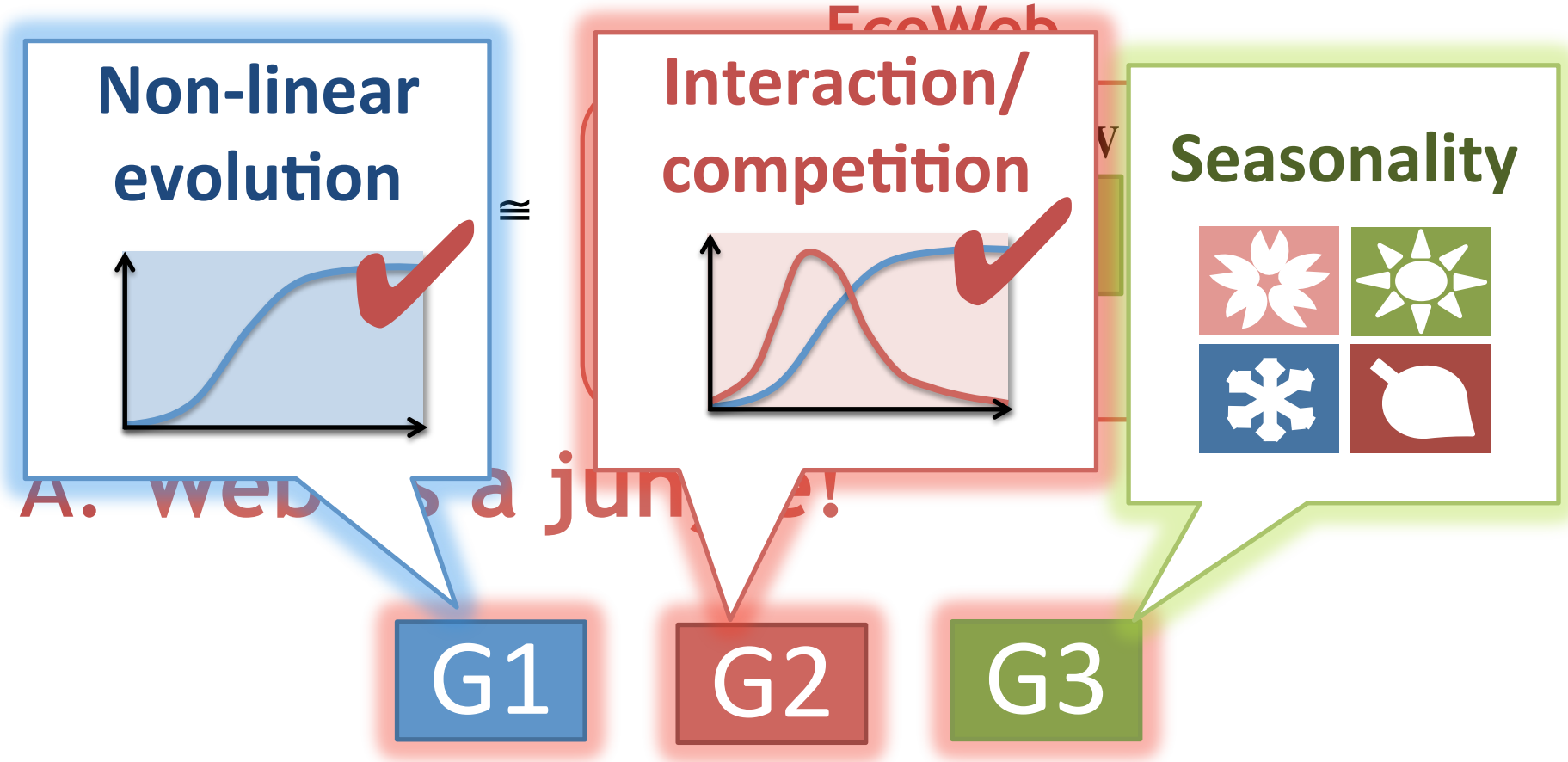


a_{ij} – Interaction coefficient

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

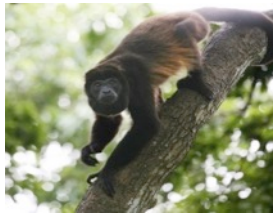
EcoWeb: Main idea

Q. How can we describe the evolutions of X ?



G3 EcoWeb-seasonality

“Hidden” seasonal activities

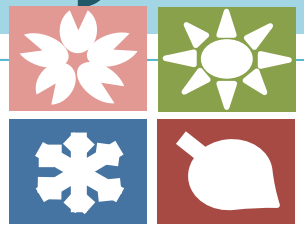



Season/ Climate




Seasonal events

G3 EcoWeb-seasonality



Details

"Hidden" seasonal activities

Estimated volume of (i)

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

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

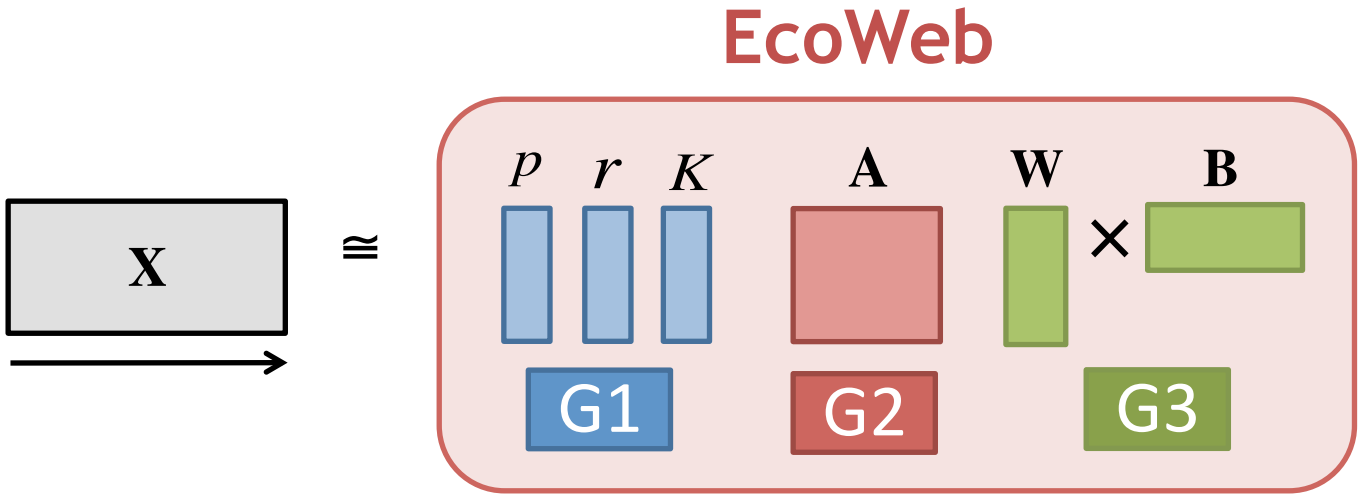
Seasonal activities of (i)

$\mathbf{W}$ – Participation (weight) matrix

$\mathbf{B}$ – Seasonality matrix

EcoWeb: Main idea

Q. How can we describe the evolutions of X ?

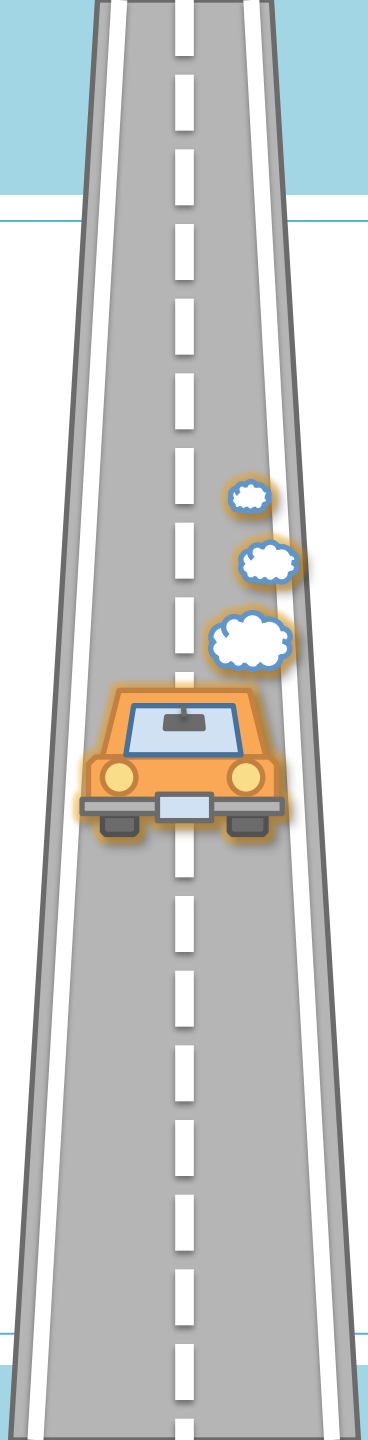


A. Web as a jungle!

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

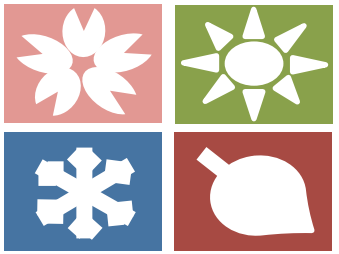
Roadmap

- ✔ Motivation
- ✔ Modeling power of EcoWeb
- ✔ Overview
- ✔ Proposed model
  - Algorithm
  - Experiments
  - EcoWeb – at work
  - Conclusions

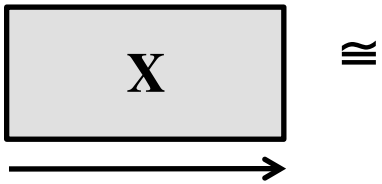


Challenges

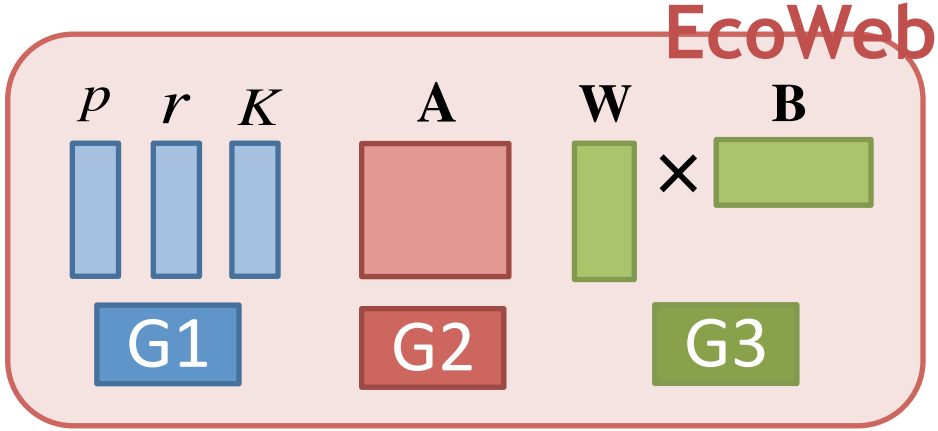
Q1. How can we automatically find "seasonal components" ?



Q2. How can we efficiently estimate full-parameters ?

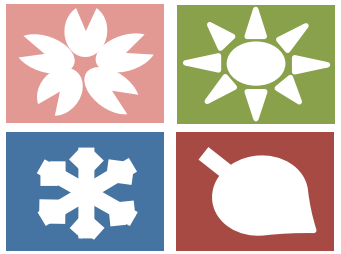


≅



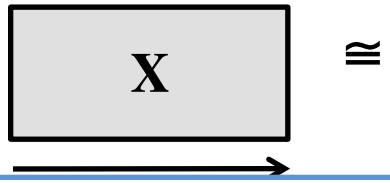
Challenges

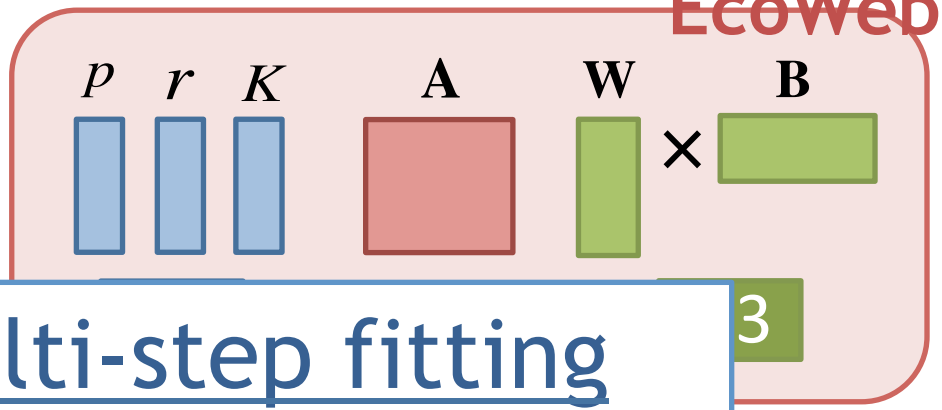
Q1. How can we automatically find "seasonal components" ?



Idea (1) : ICA + MDL

Q2. How can we efficiently estimate full-parameters ?

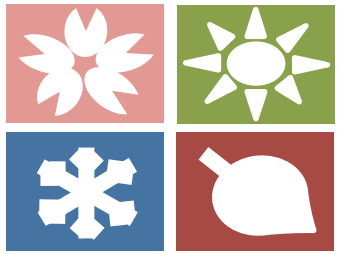




Idea (2): Multi-step fitting

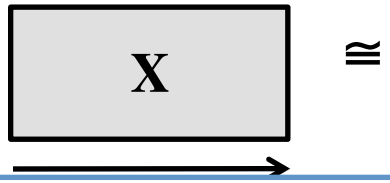
Challenges

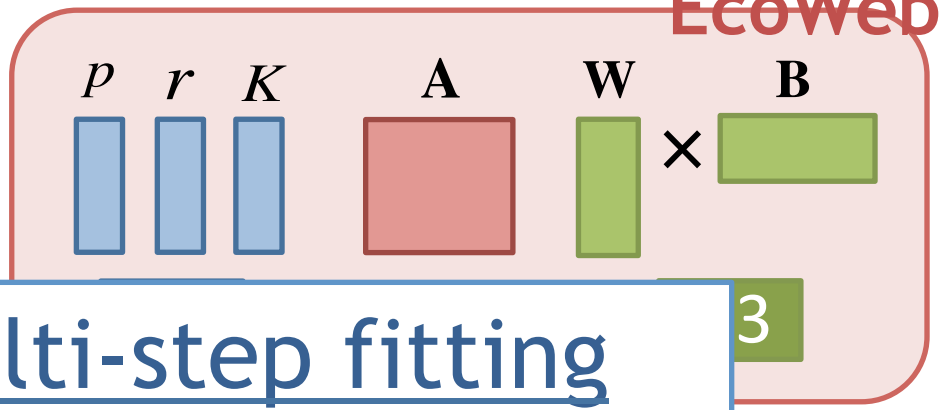
Q1. How can we automatically find "seasonal components" ?



Idea (1) : ICA + MDL

Q2. How can we efficiently estimate full-parameters ?

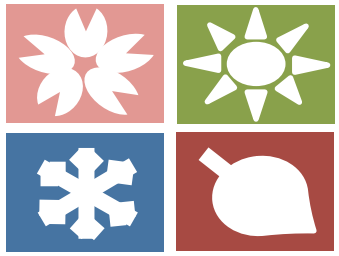


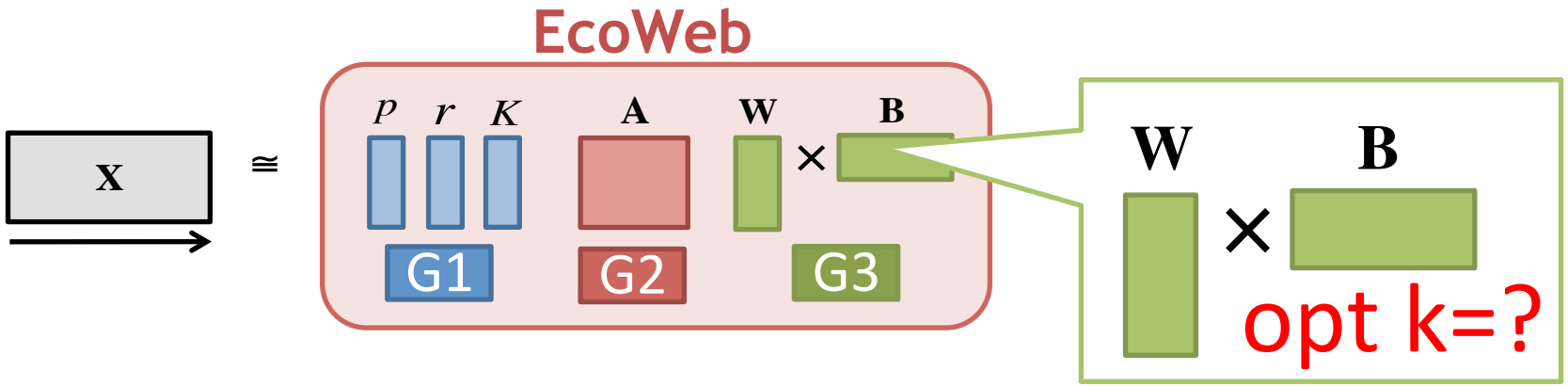


Idea (2): Multi-step fitting

Idea (1): Seasonal component analysis

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



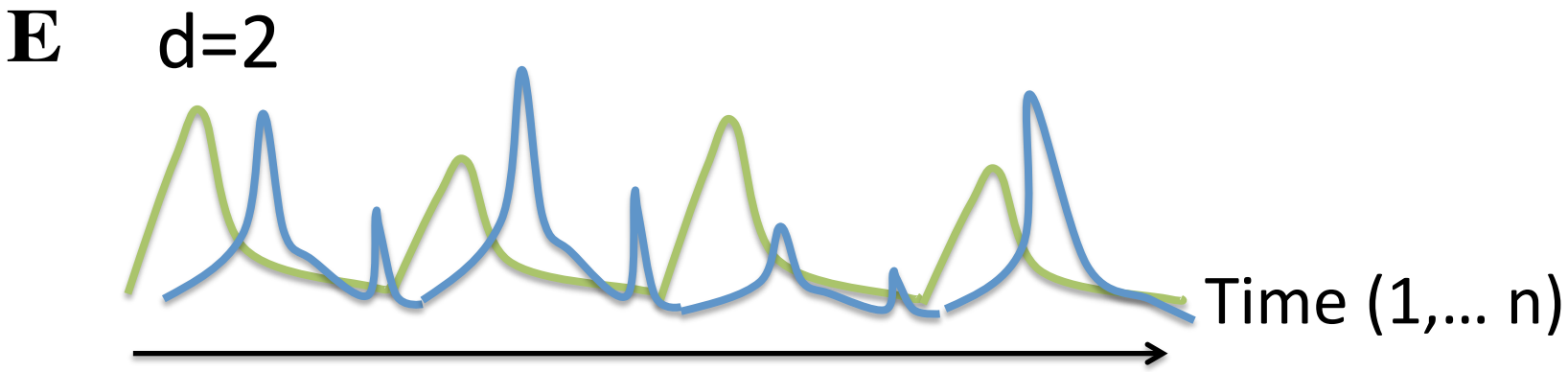


Idea (1) : ICA + MDL

a. Seasonal component detection (ICA)
b. Automatic component analysis (MDL)

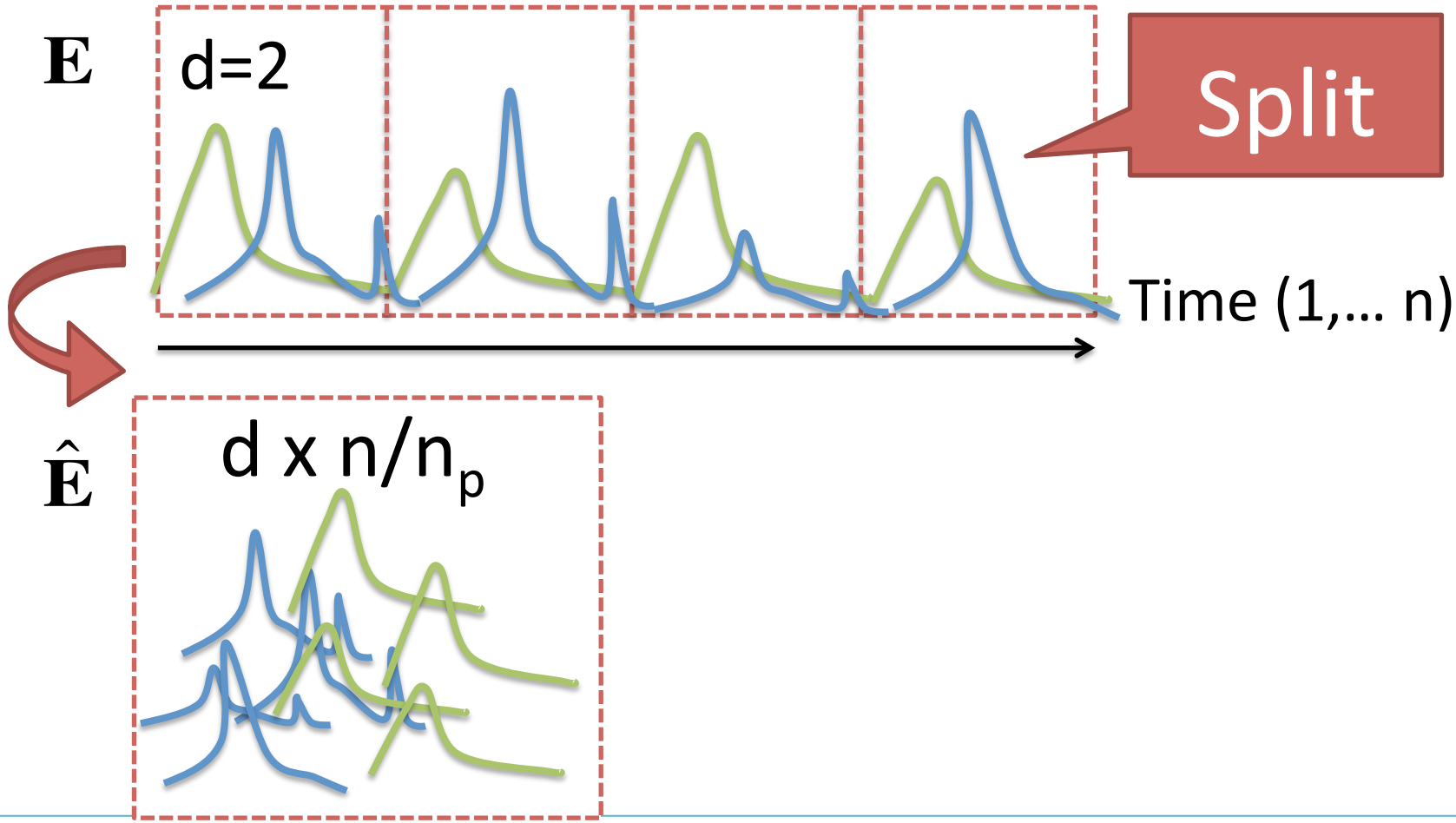
Idea (1): Seasonal component analysis

Idea(1-a) Seasonal component detection



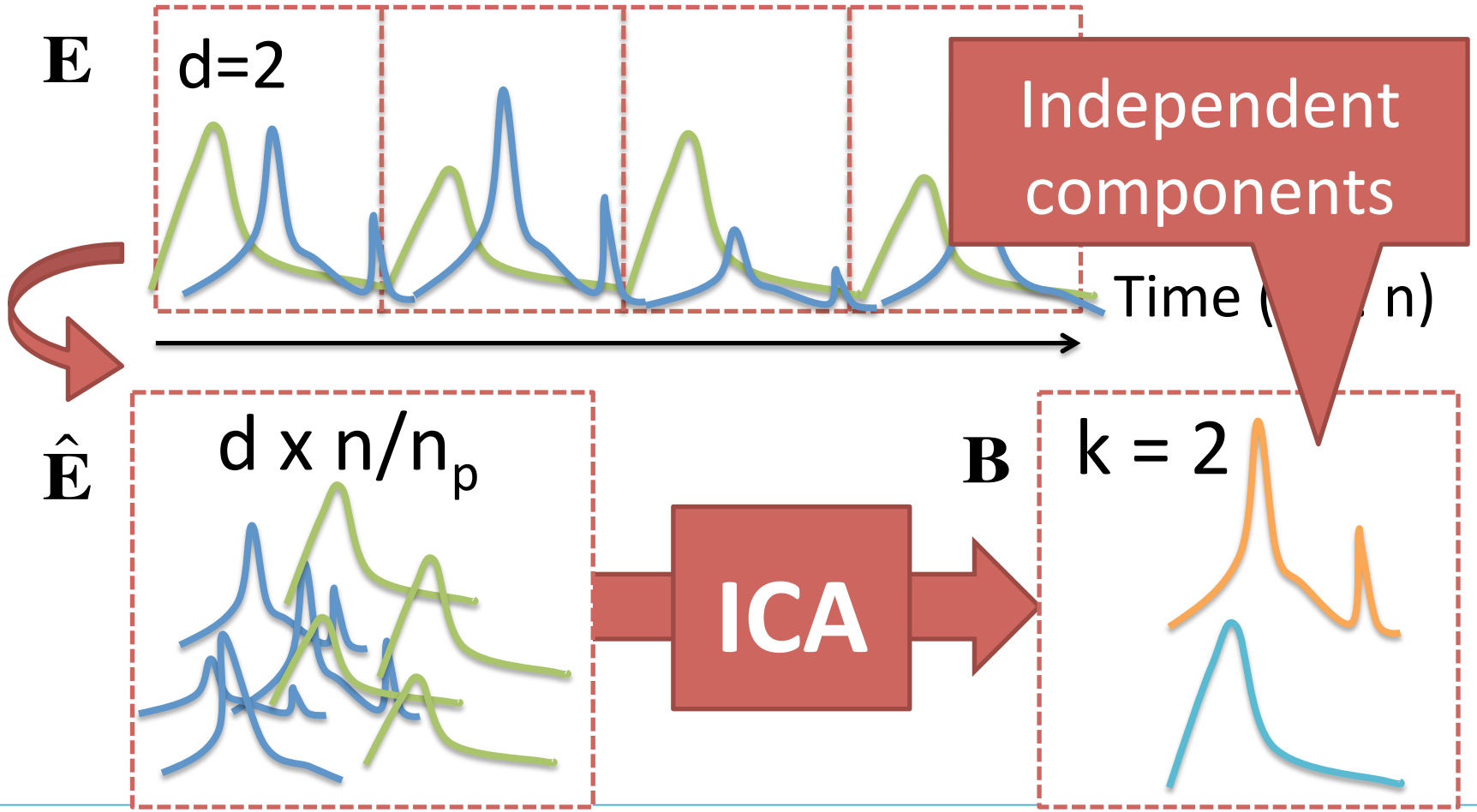
Idea (1): Seasonal component analysis

Idea(1-a) Seasonal component detection



Idea (1): Seasonal component analysis

Idea(1-a) Seasonal component detection

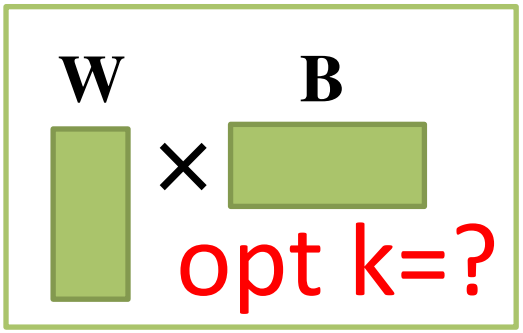


Idea (1): Seasonal component analysis

Idea(1-b) MDL

Find optimal number k (1≤k≤d)

d: dimension

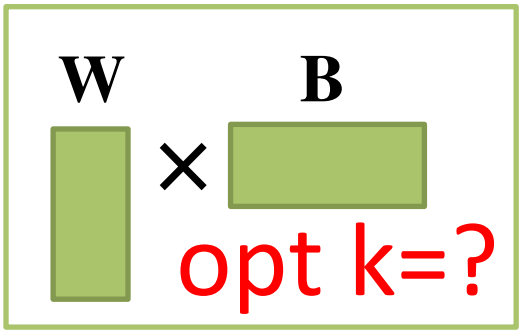


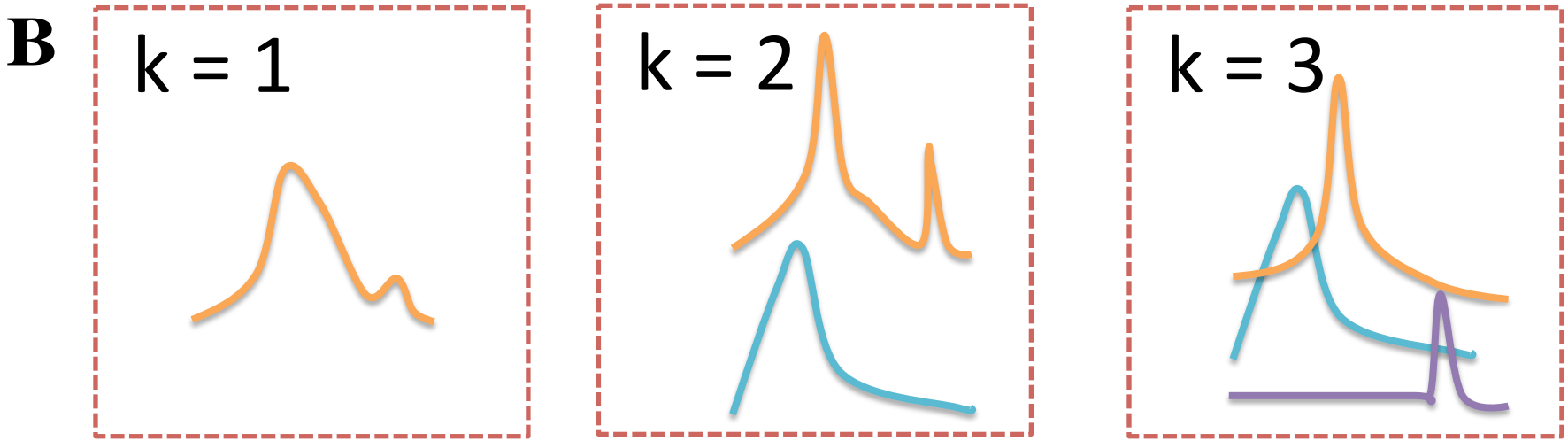
Idea (1): Seasonal component analysis

Idea(1-b) MDL

Find optimal number k (1≤k≤d)

d: dimension

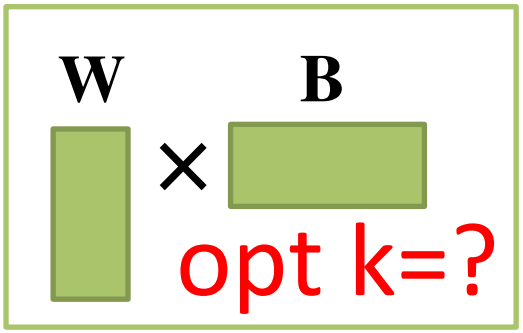




Idea (1): Seasonal component analysis

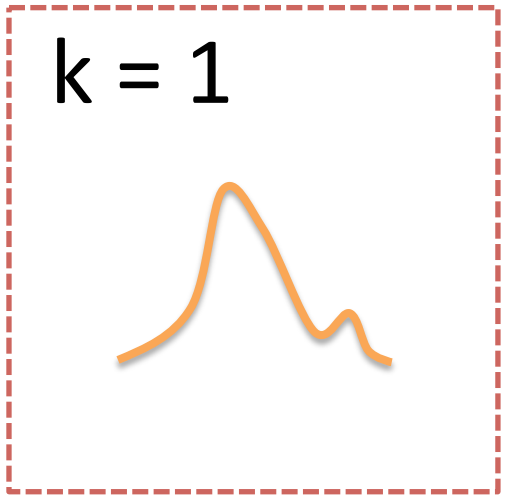
Idea(1-b) MDL

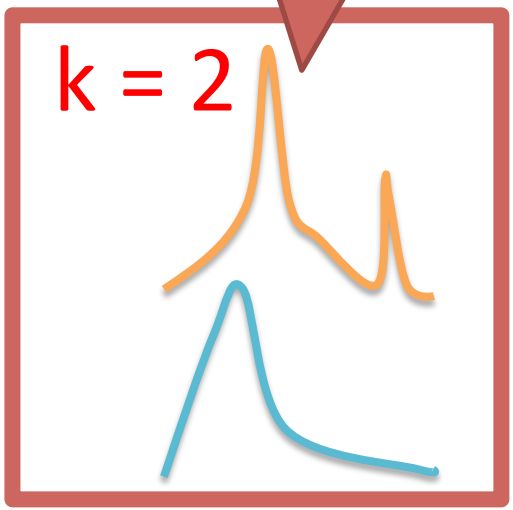
Find optimal number k (1≤k≤d)

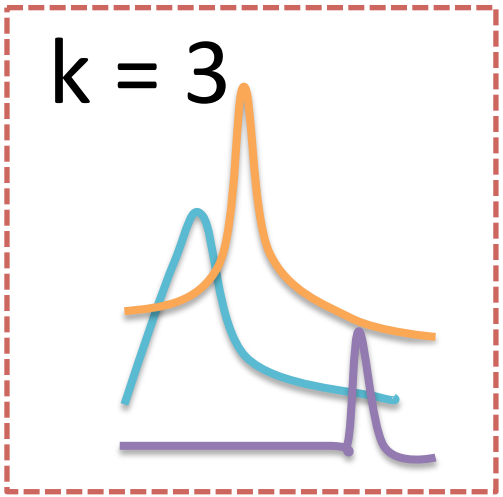


Optimal k

B







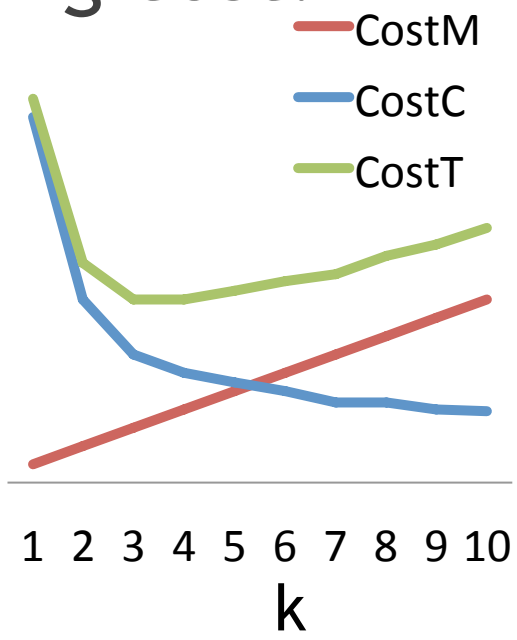
Idea (1): Seasonal component analysis

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

$$\min\ (\ \boxed{\text{Cost}_M(S)}\ +\ \boxed{\text{Cost}_C(\mathcal{X}\,|\,S)}\ )$$

Model cost

Coding cost

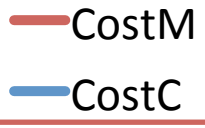


Good compression

Good description

Idea (1): Seasonal component analysis

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



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

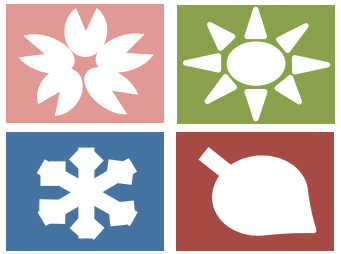
$$k_{opt} = \arg\min_k Cost_T(X;\mathcal{S})$$

Good compression

Good description

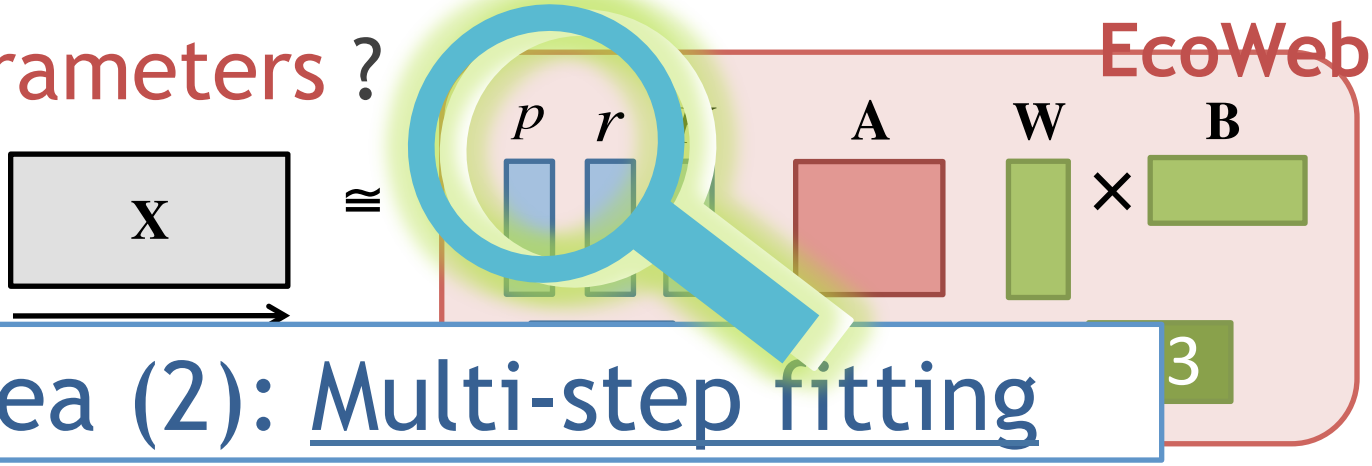
Challenges

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



Idea (1) : ICA + MDL

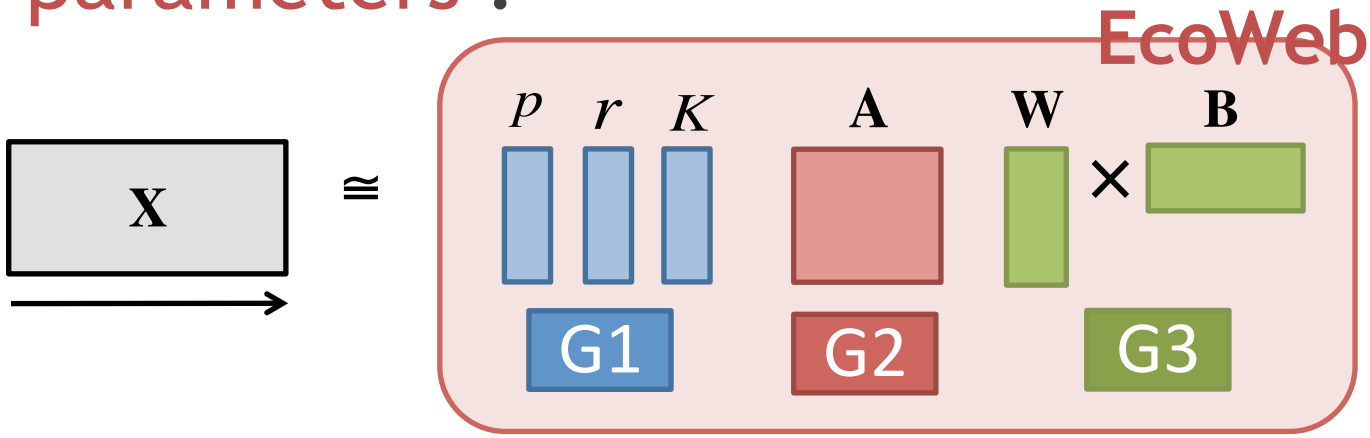
Q2. How can we efficiently estimate full-parameters ?



Idea (2): Multi-step fitting

Idea (2): Multi-step fitting

Q2. How can we efficiently estimate model parameters ?



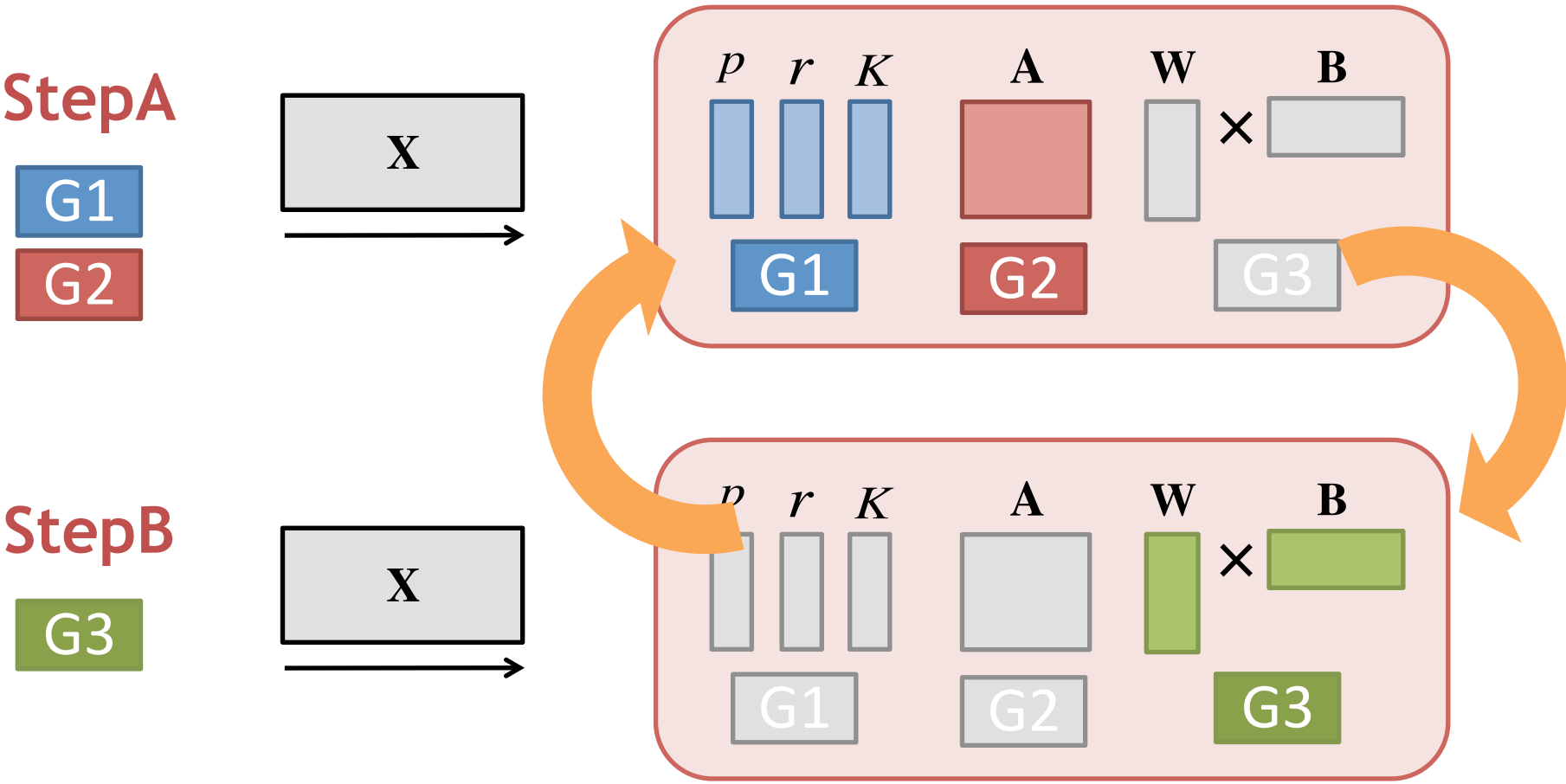
Idea (2): Multi-step fitting

a. **StepFit** (sub)

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

Idea (2): Multi-step fitting

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



Idea (2): Multi-step fitting

EcoWeb-Fit: full algorithm

e.g., 4 keywords: 

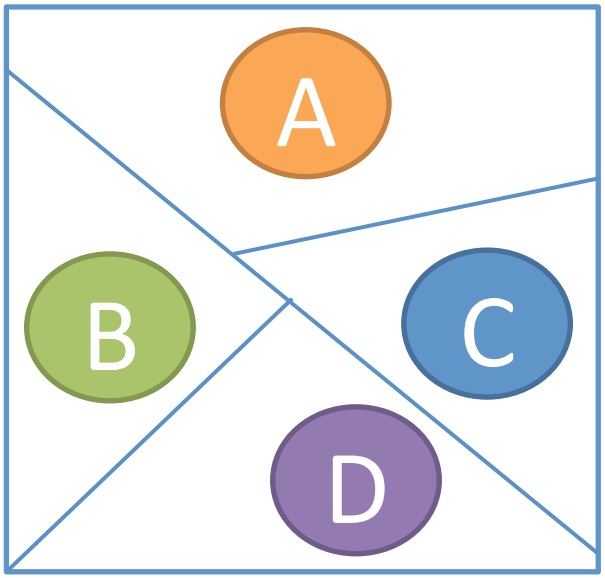
EcoWeb-Fit updates parameters, separately

Idea (2): Multi-step fitting

EcoWeb-Fit: full algorithm

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

1. Individual-Fit



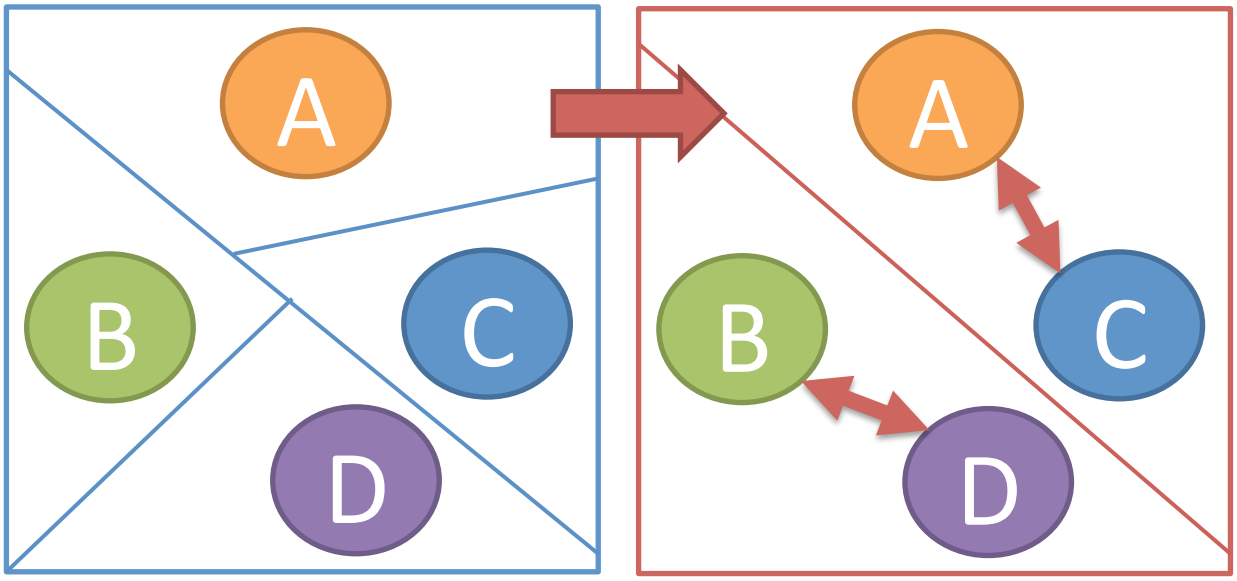
EcoWeb-Fit updates parameters, separately

Idea (2): Multi-step fitting

EcoWeb-Fit: full algorithm

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

1. Individual-Fit **2. Pair-Fit**



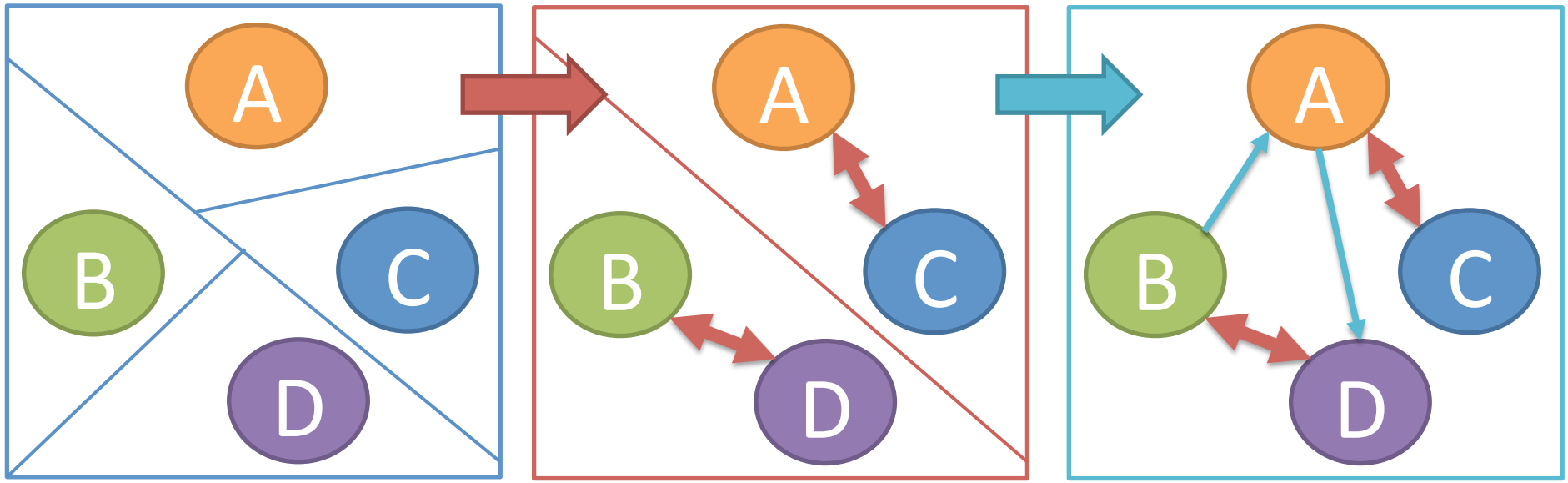
EcoWeb-Fit updates parameters, separately

Idea (2): Multi-step fitting

EcoWeb-Fit: full algorithm

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

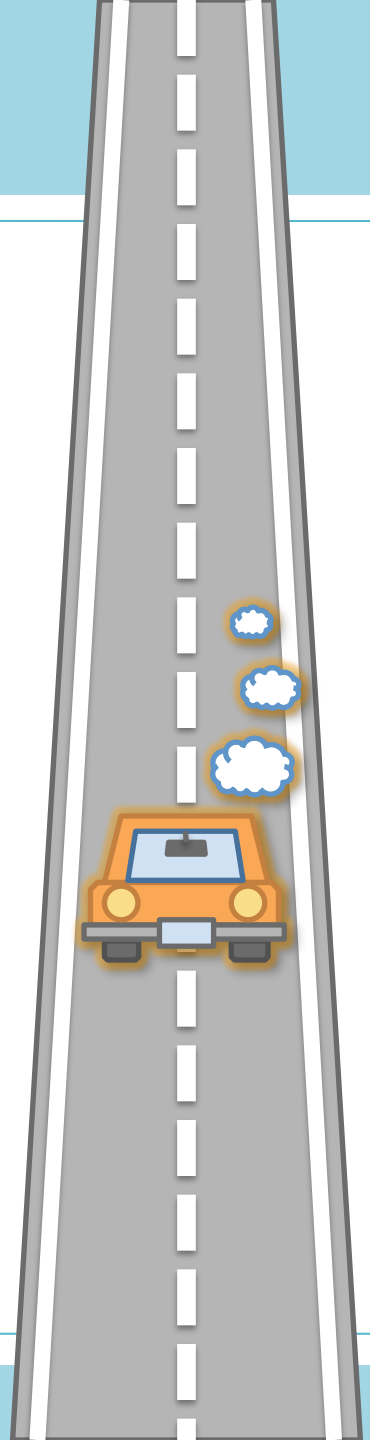
1. Individual-Fit **2. Pair-Fit** **3. Full-Fit**



EcoWeb-Fit updates parameters, separately

Roadmap

- ✔ Motivation
- ✔ Modeling power of EcoWeb
- ✔ Overview
- ✔ Proposed model
- ✔ Algorithm
- Experiments
- EcoWeb – at work
- Conclusions



Experiments

We answer the following questions…

Q1. Effectiveness

How successful is it in spotting patterns?

Q2. Accuracy

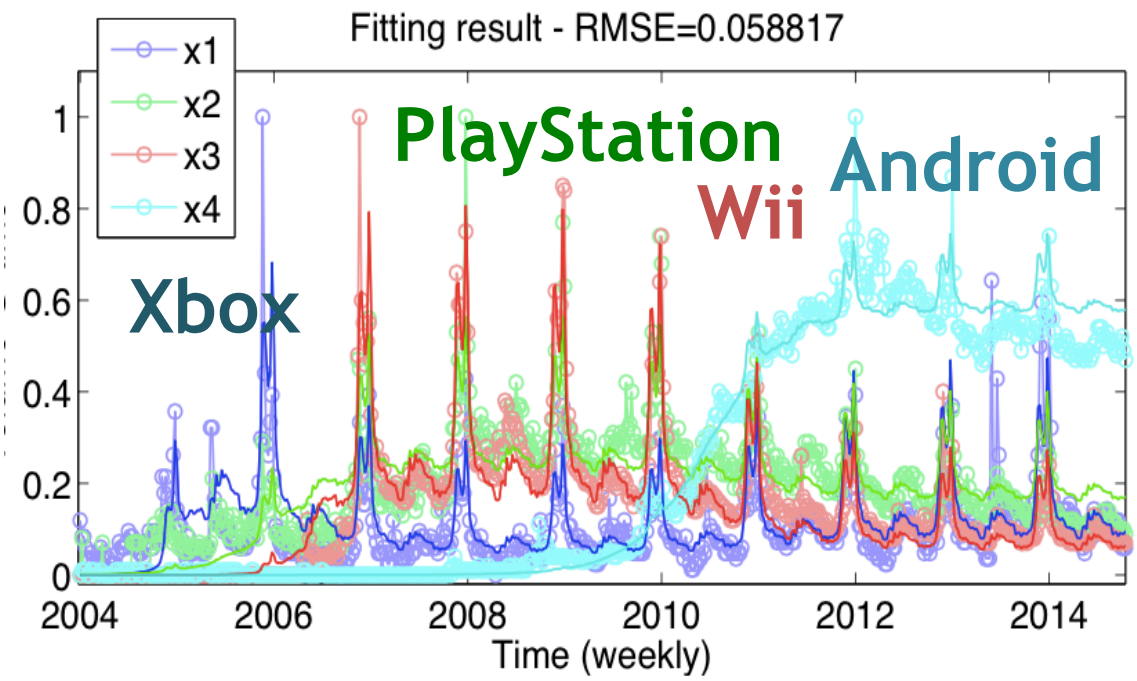
How well does it match the data?

Q3. Scalability

How does it scale in terms of computational time?

Q1. Effectiveness

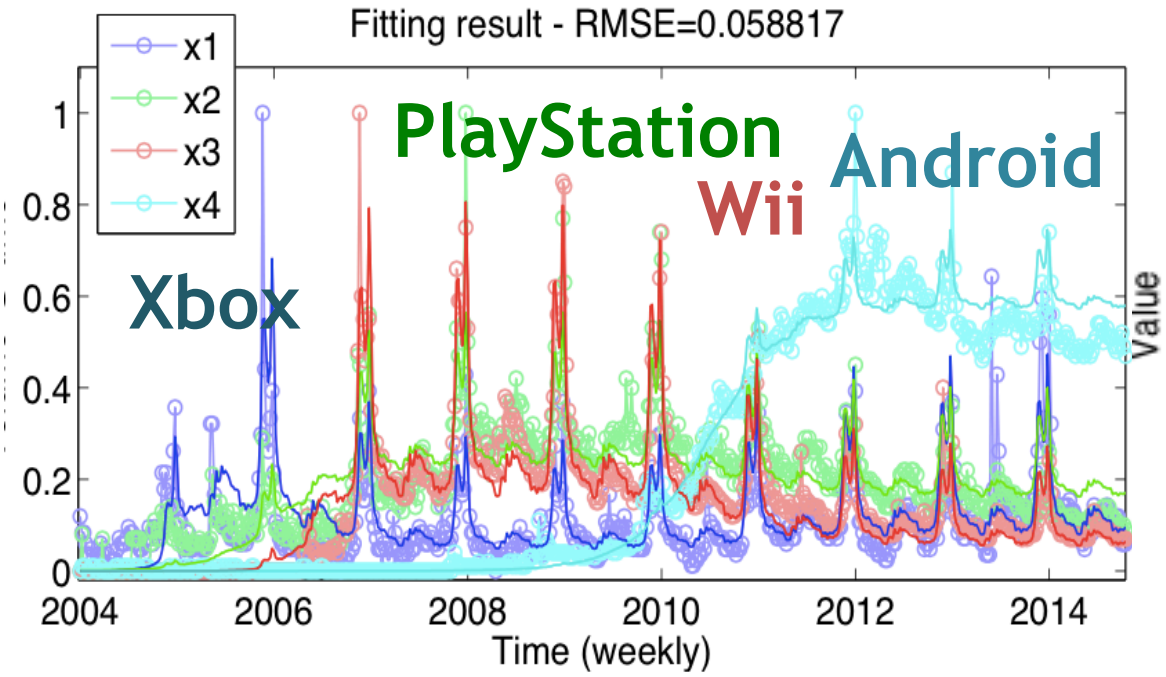
(#1) Video games



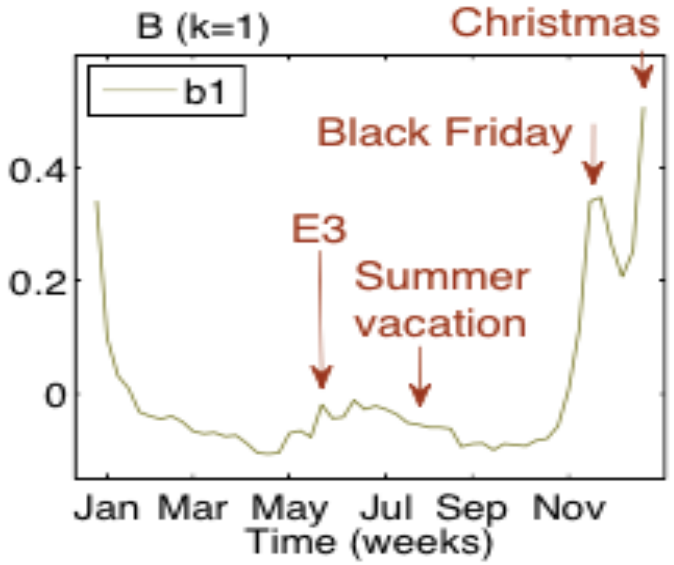
Interactions between keywords

Q1. Effectiveness

(#1) Video games



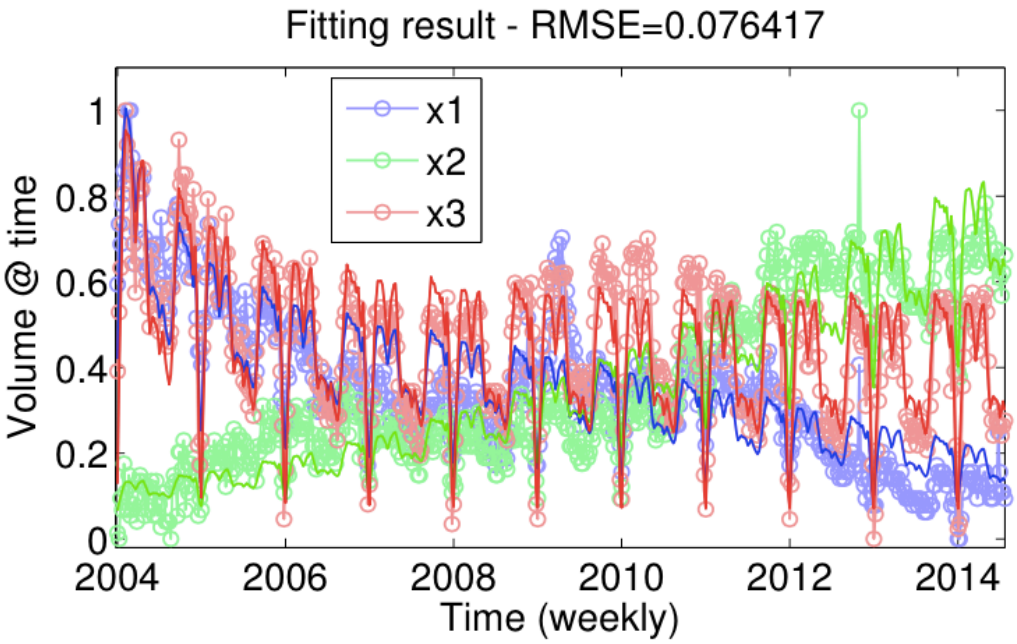
Seasonality



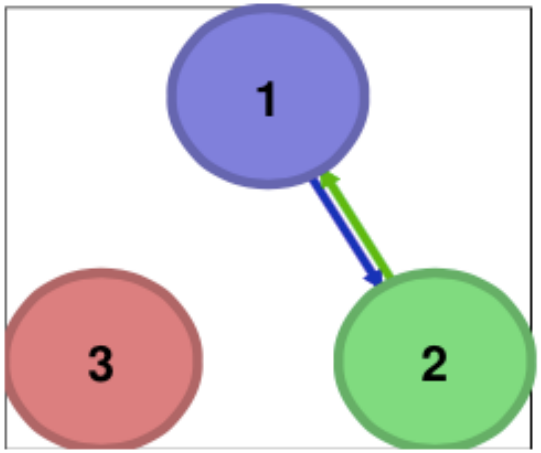
Q1. Effectiveness

(#2) Programming language

C , R , MATLAB



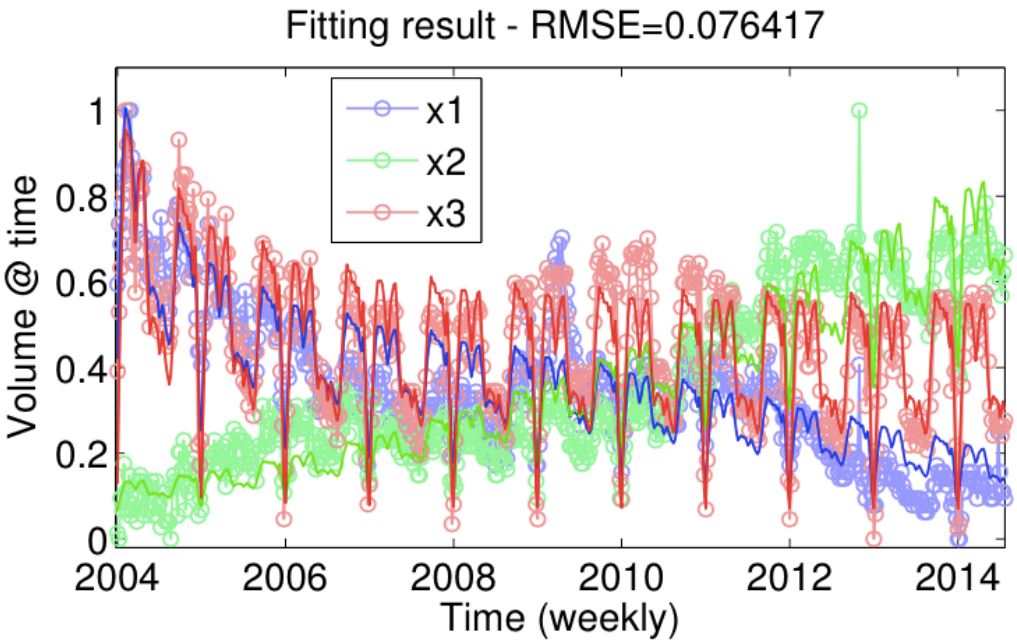
Interactions



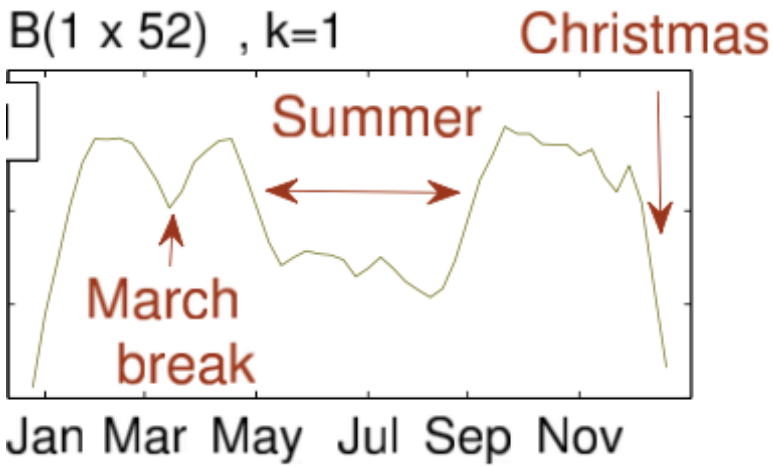
Q1. Effectiveness

(#2) Programming language

C , R , MATLAB



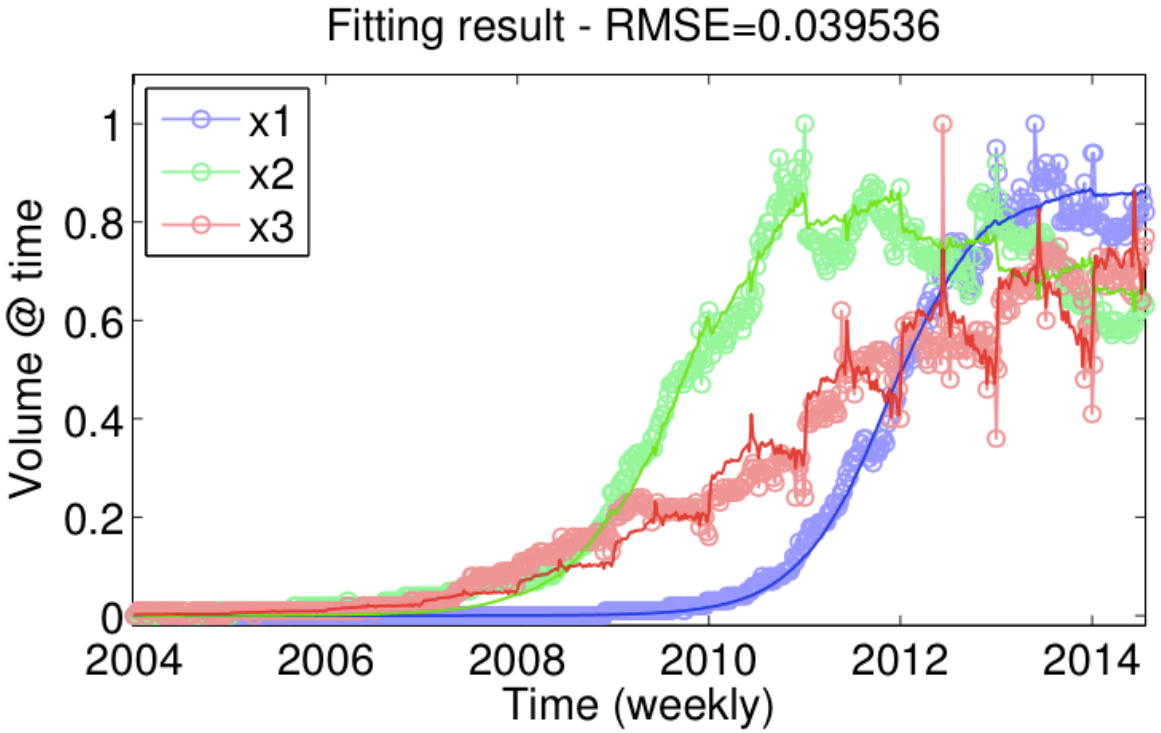
Seasonality



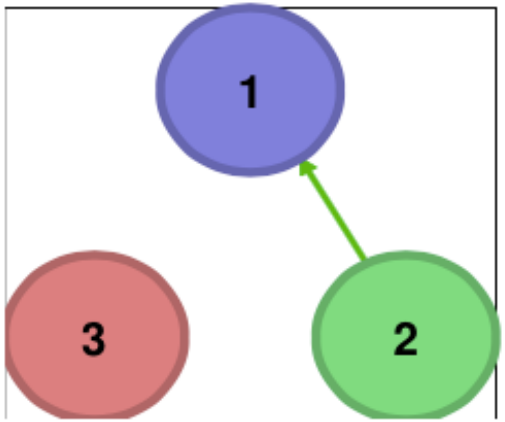
Q1. Effectiveness

(#3) Social media

Tumblr , Facebook , LinkedIn



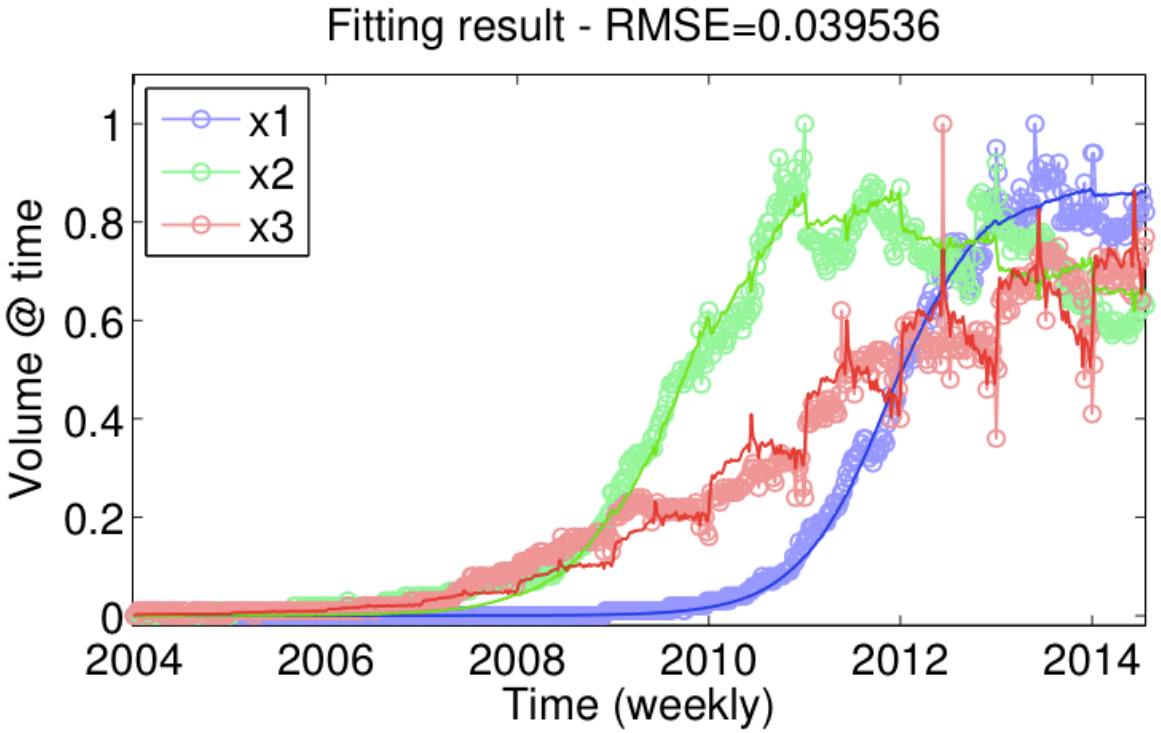
Interactions



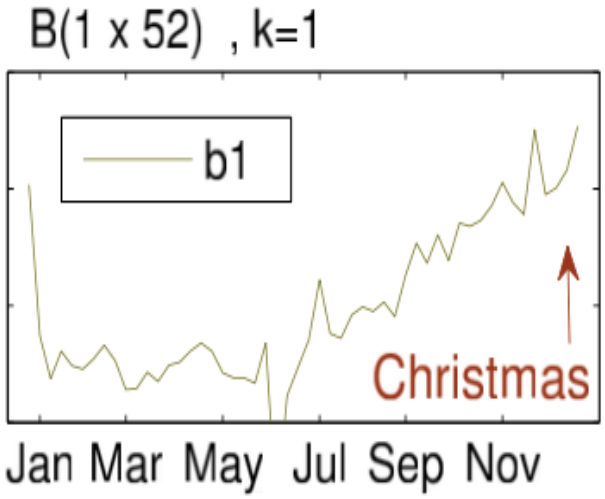
Q1. Effectiveness

(#3) Social media

Tumblr , Facebook , LinkedIn



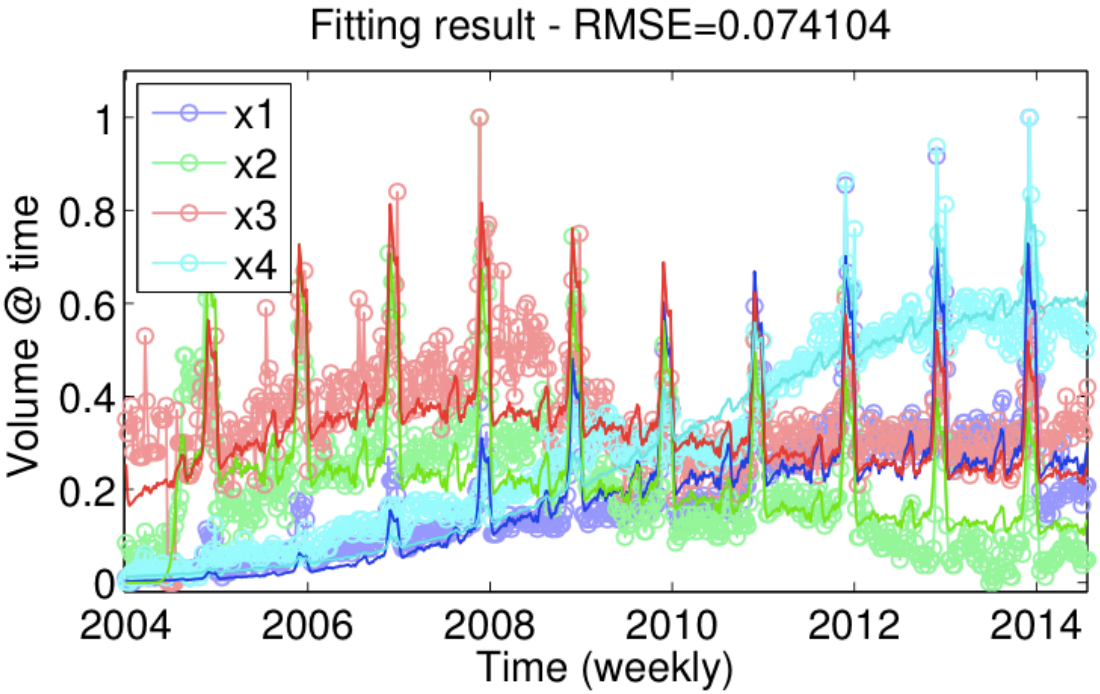
Seasonality



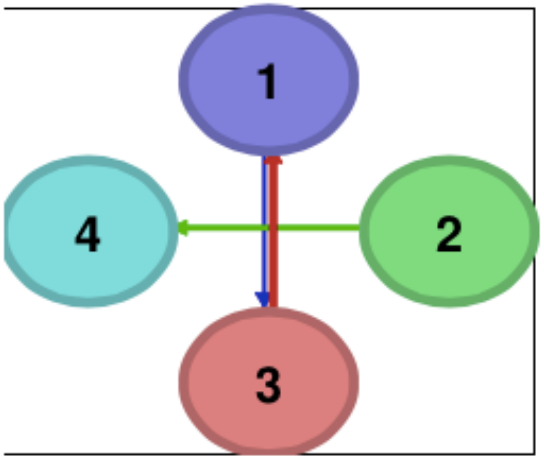
Q1. Effectiveness

(#4) Apparel companies

Kohls , JCPenny , Nordstrom , Forever21



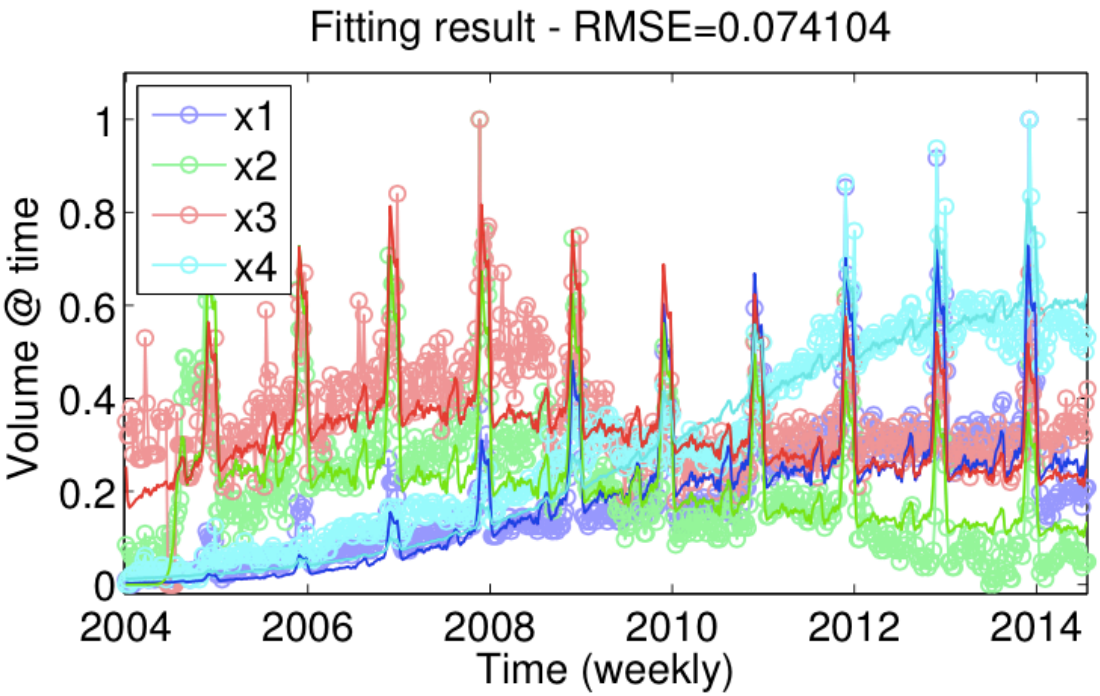
Interactions



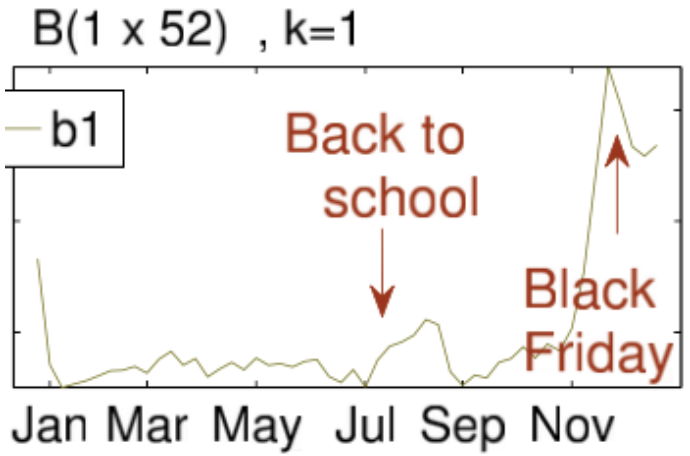
Q1. Effectiveness

(#4) Apparel companies

Kohls , JCPenny , Nordstrom , Forever21



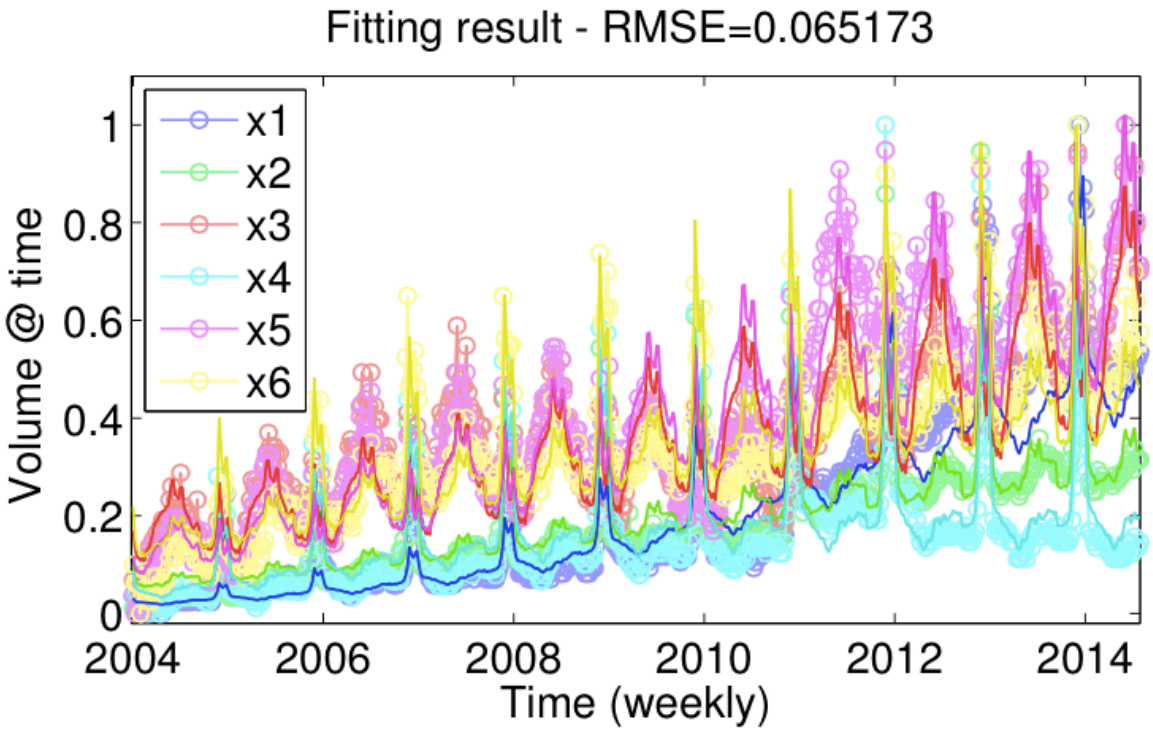
Seasonality



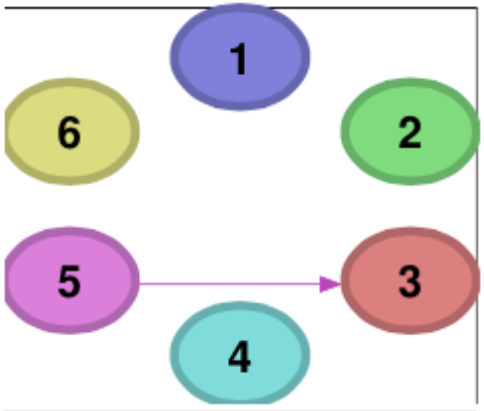
Q1. Effectiveness

(#5) Retail companies

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



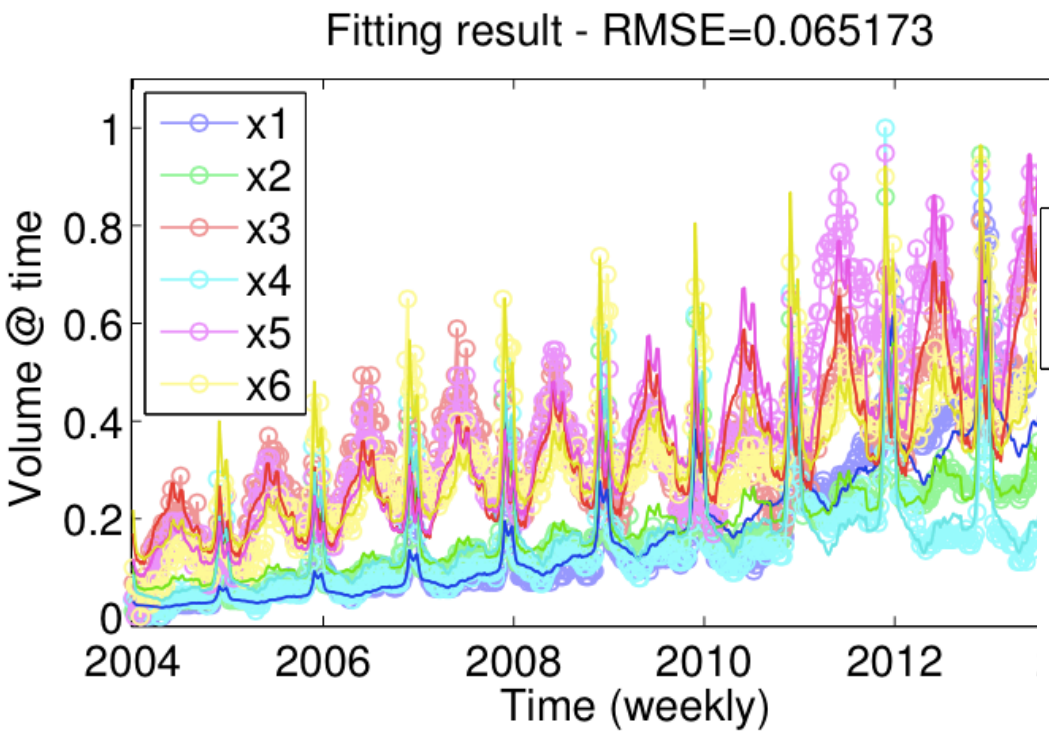
Interaction



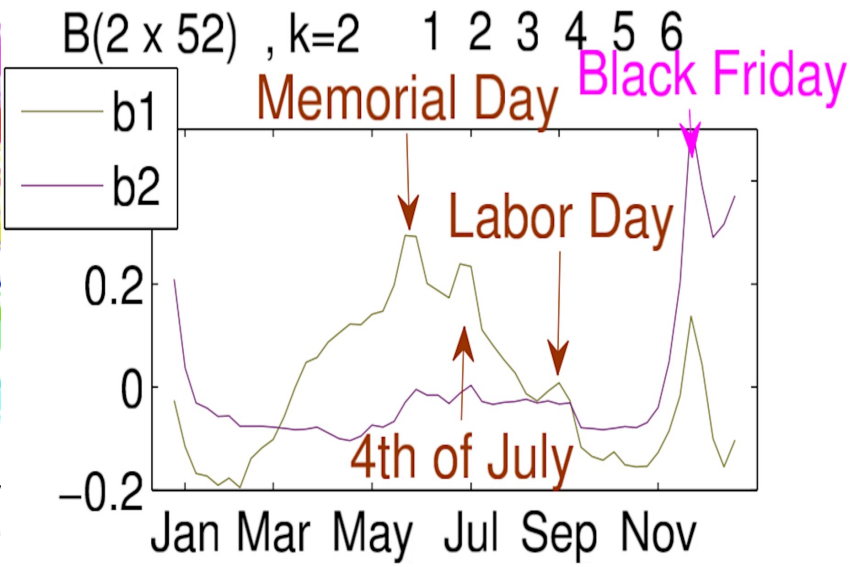
Q1. Effectiveness

(#5) Retail companies

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



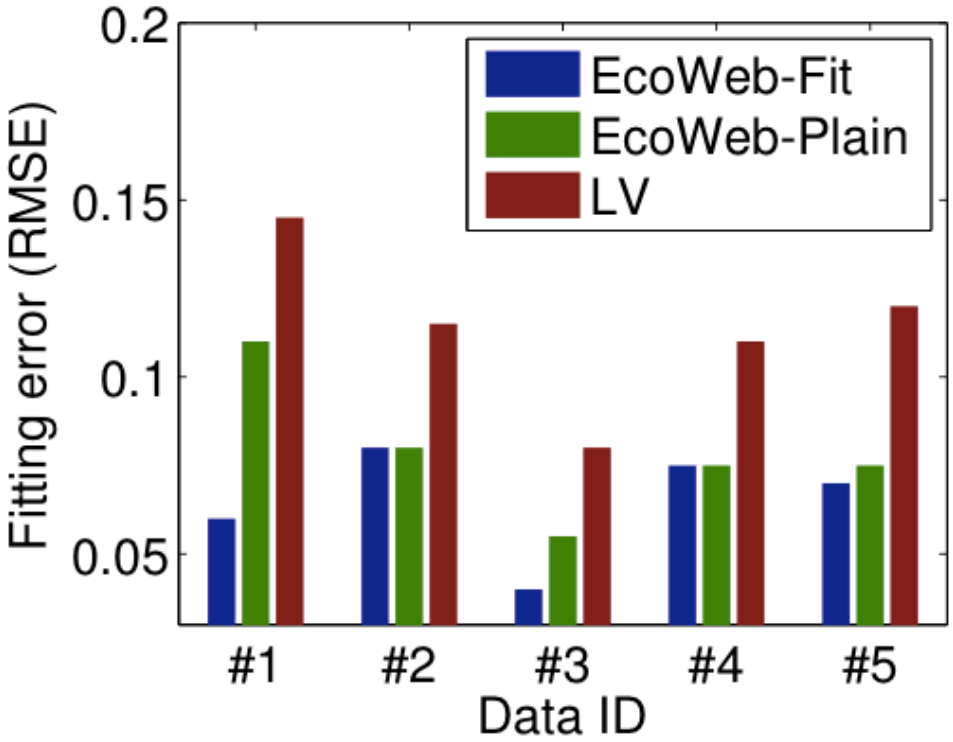
Seasonality



Q2. Accuracy

RMSE between original and fitted volume

(Lower is better)

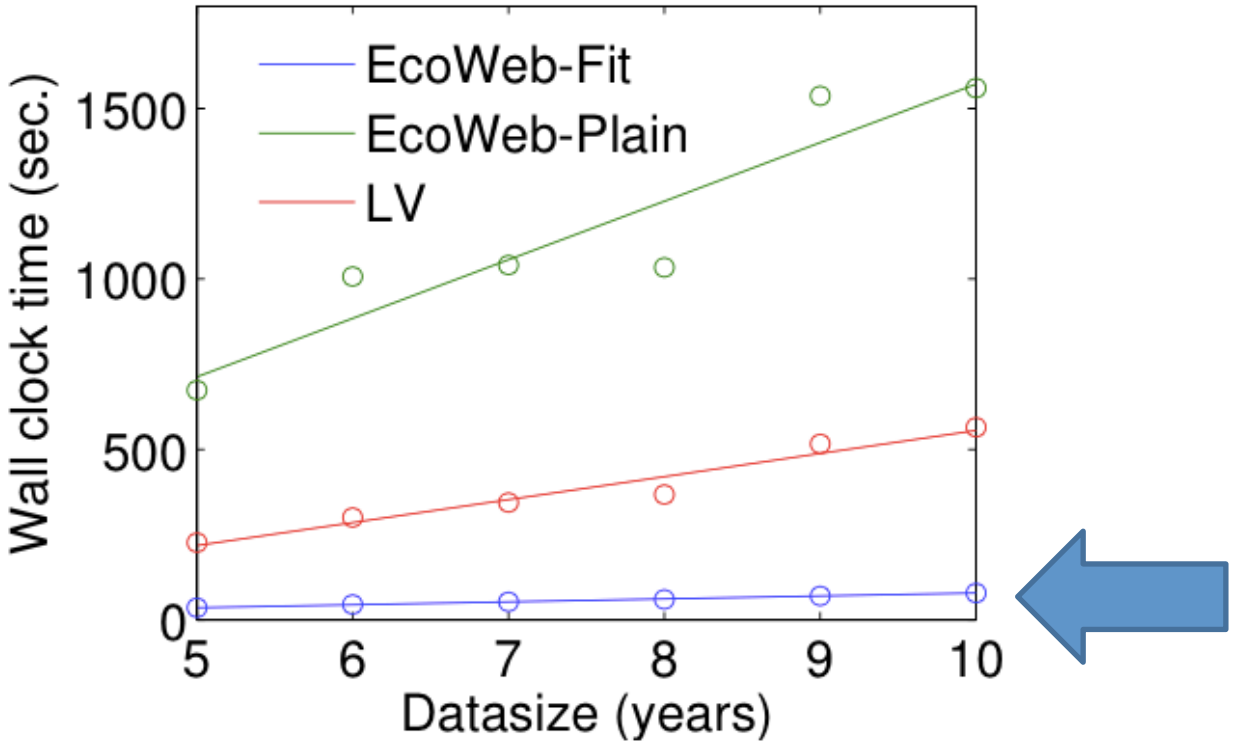


EcoWeb consistently wins!

Q3. Scalability

Wall clock time vs. dataset size (years)

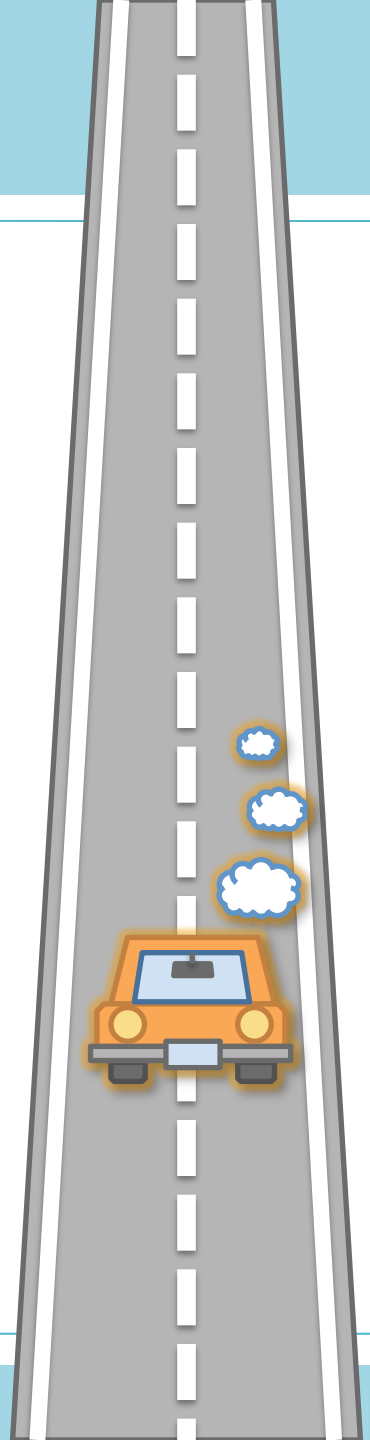
EcoWeb-Fit scales linearly, i.e., O(n)



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

Roadmap

- ✔ Motivation
- ✔ Modeling power of EcoWeb
- ✔ Overview
- ✔ Proposed model
- ✔ Algorithm
- ✔ Experiments
- – EcoWeb – at work
- – Conclusions

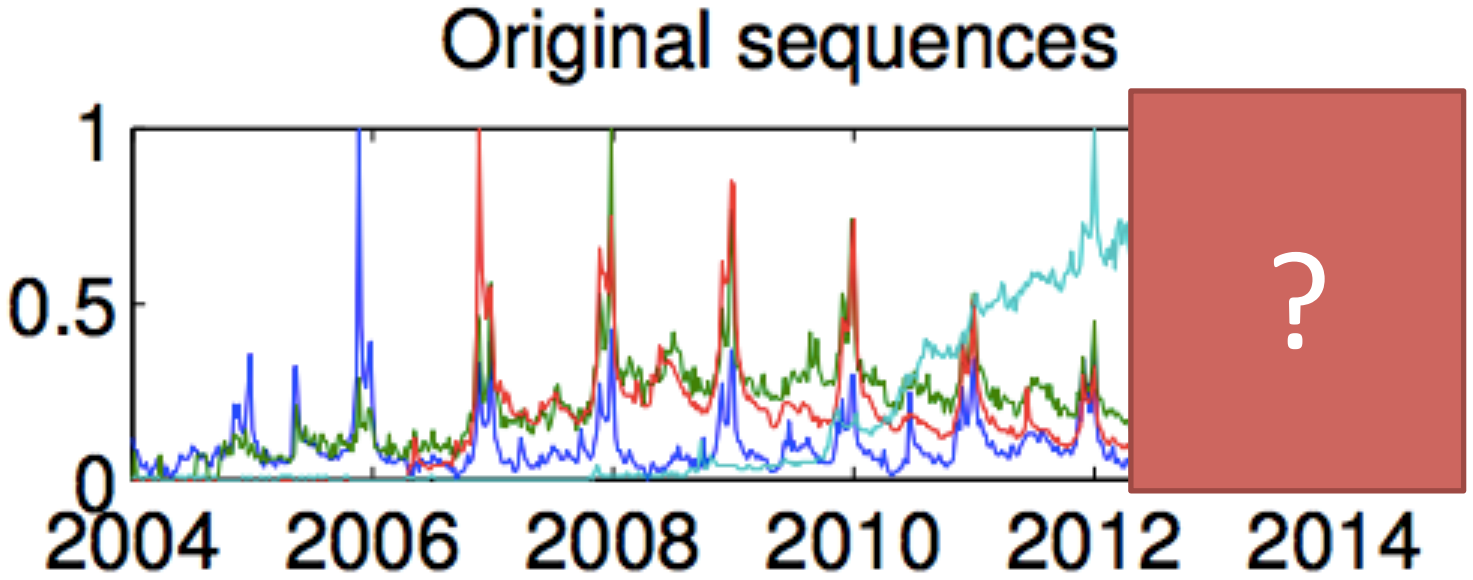


EcoWeb at work - forecasting

Forecasting future activities

Train:
2/3 sequences

Forecast:
1/3 following years

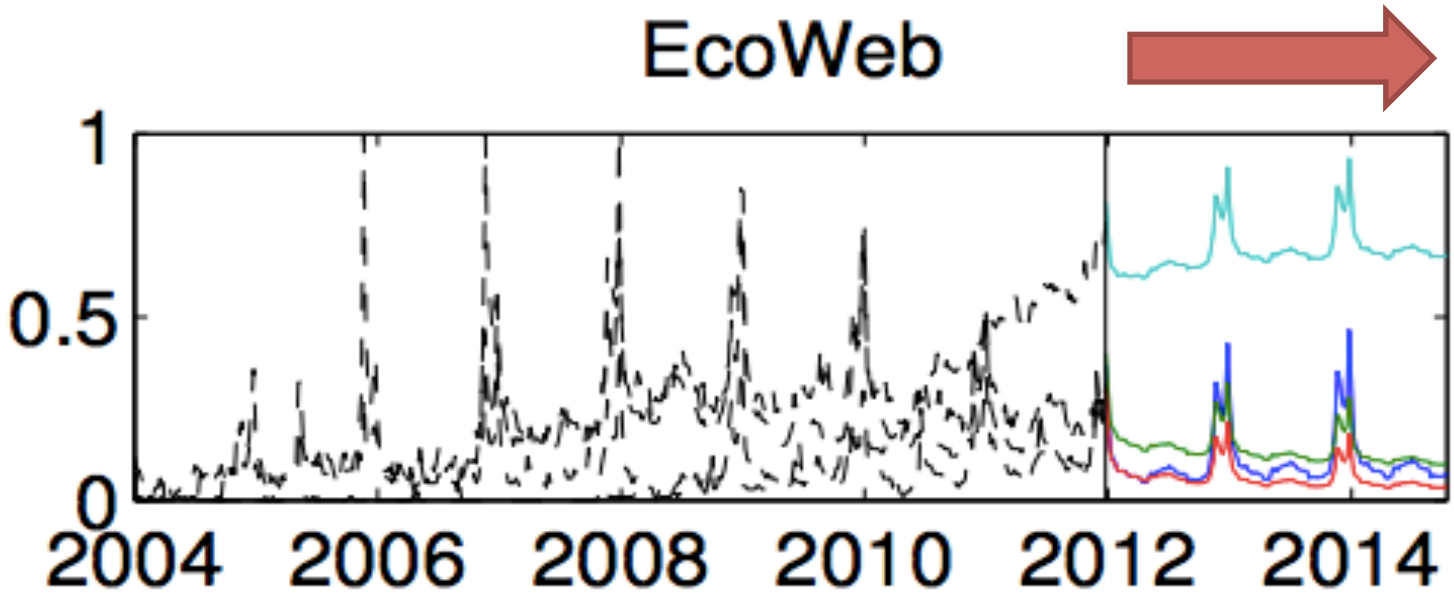


EcoWeb at work - forecasting

Forecasting future activities

Train:
2/3 sequences

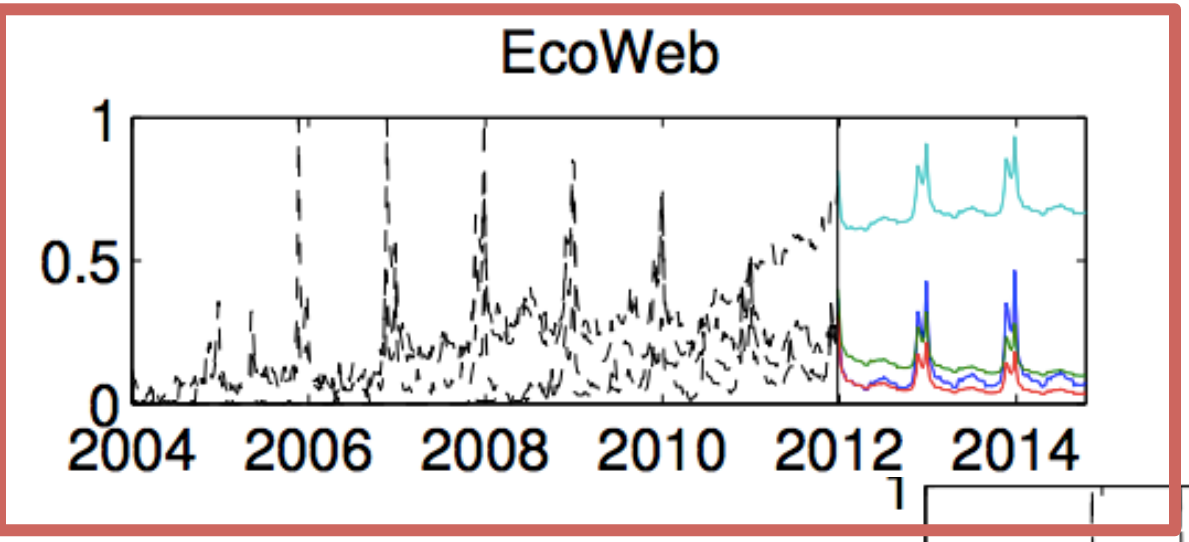
Forecast:
1/3 following years

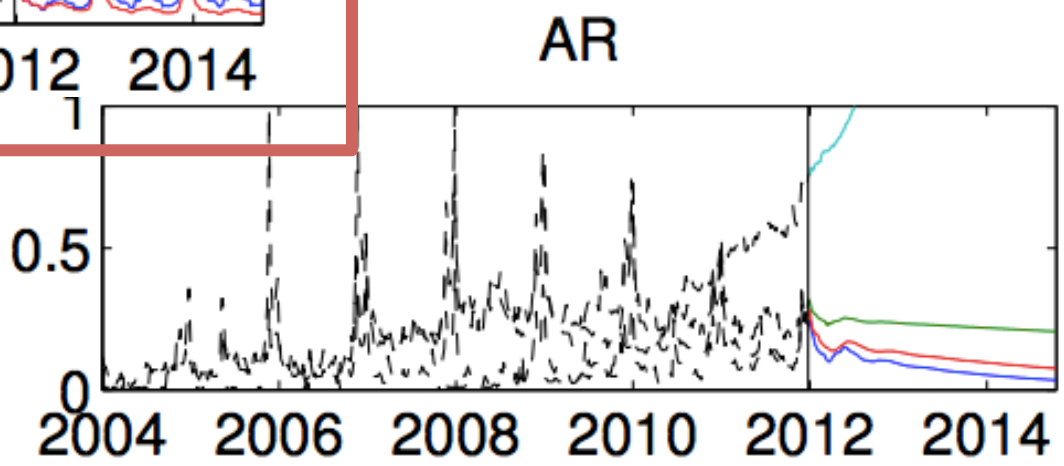


EcoWeb can capture future patterns

EcoWeb at work - forecasting

Forecasting future activities

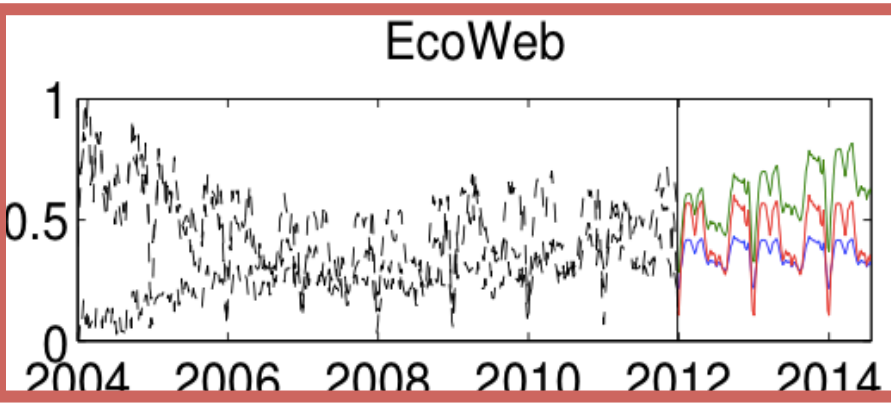


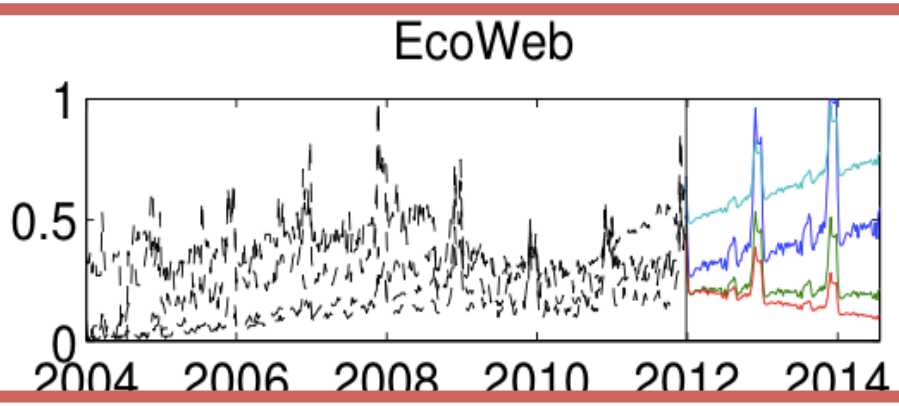


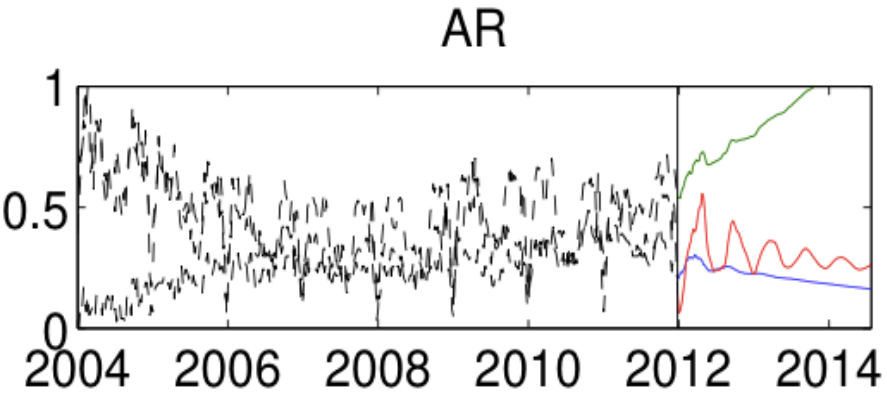
EcoWeb can capture future patterns!

EcoWeb at work - forecasting

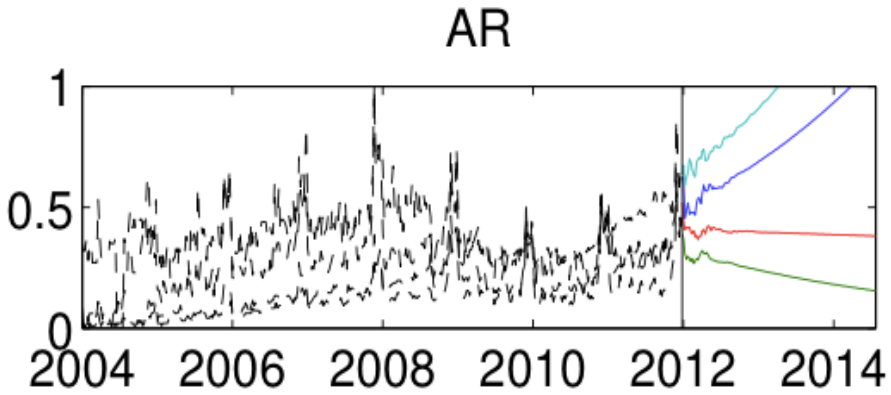
Forecasting future activities







(b) Programming languages (#2)

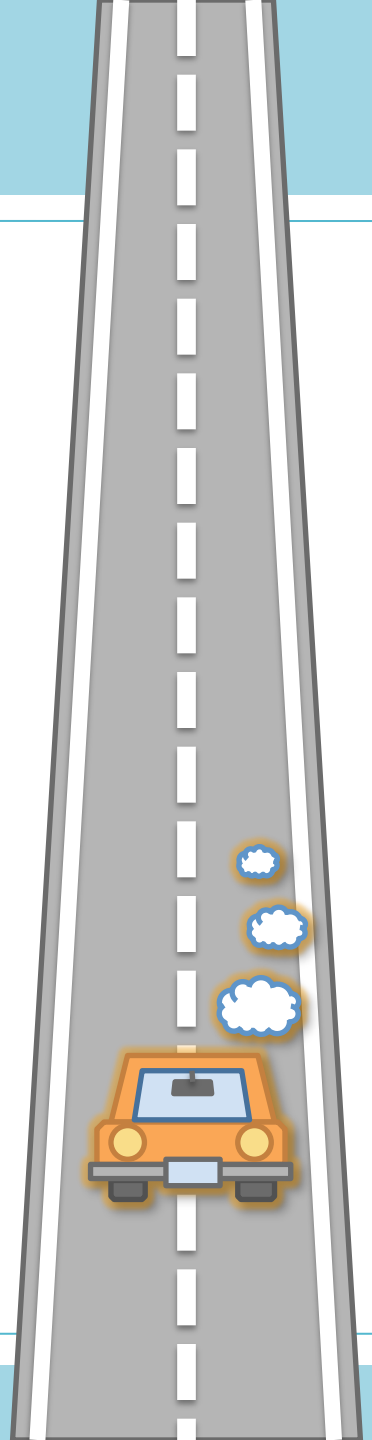


(c) Apparel companies (#4)

EcoWeb can capture future patterns!

Roadmap

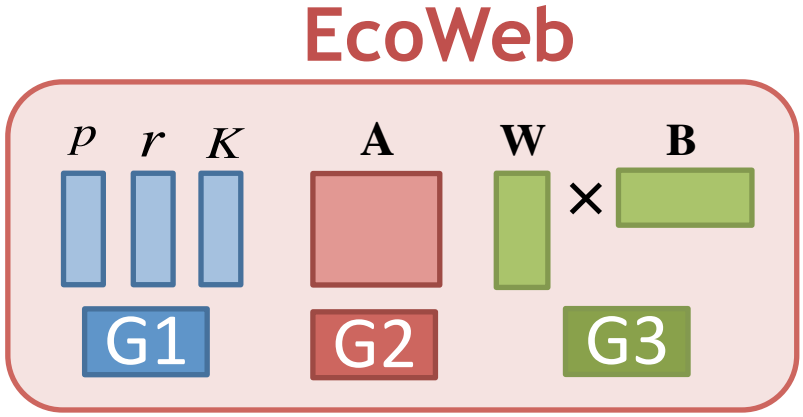
- ✔ Motivation
- ✔ Modeling power of EcoWeb
- ✔ Overview
- ✔ Proposed model
- ✔ Algorithm
- ✔ Experiments
- ✔ EcoWeb – at work
- Conclusions

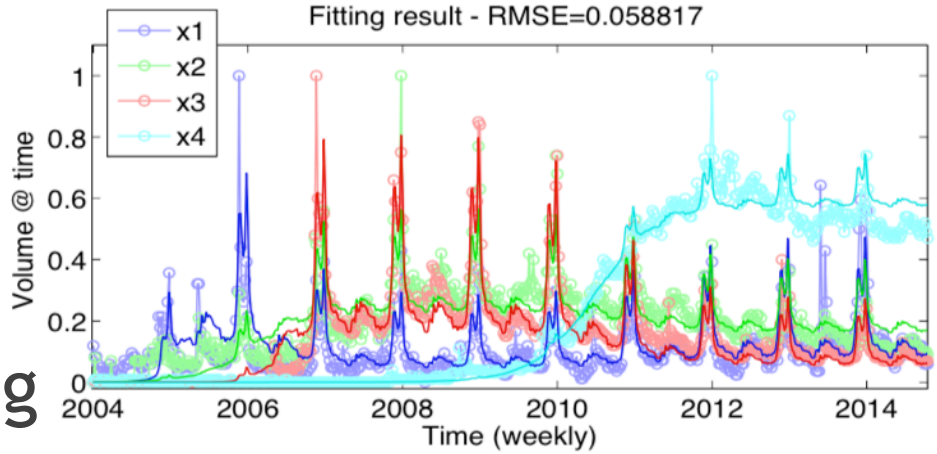


Conclusions

EcoWeb has the following advantages

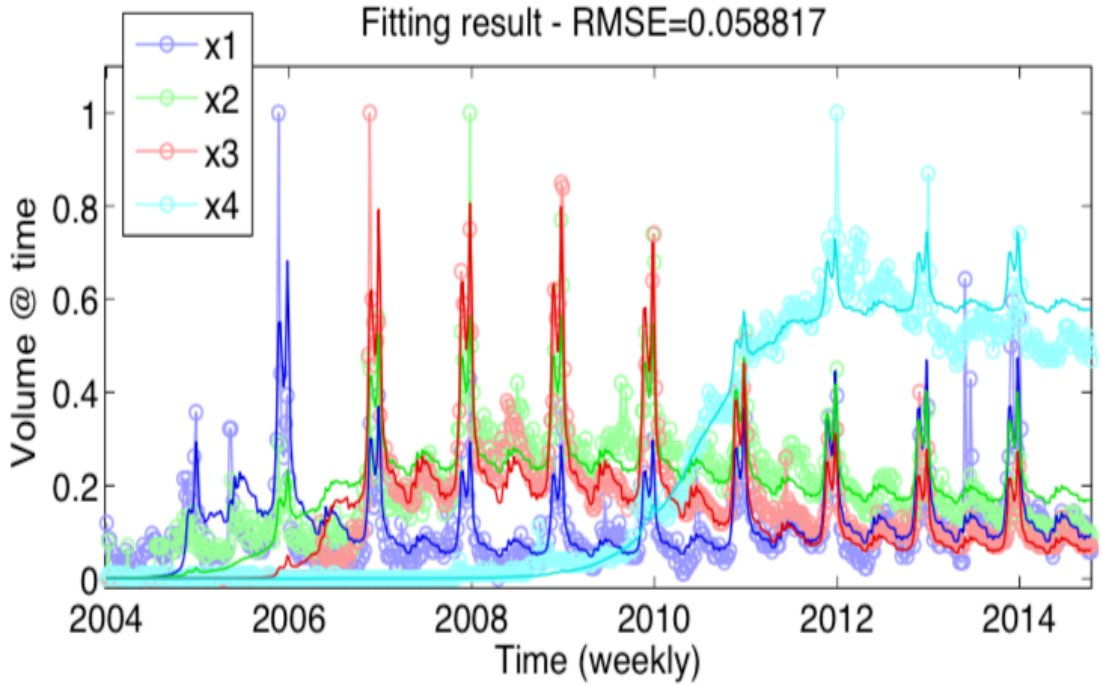
- ✔ Effective
  - Finds important patterns
- ✔ Fully-automatic
  - No parameter tuning
- ✔ Scalable
  - It is linear
- ✔ Practical
  - Long-range forecasting

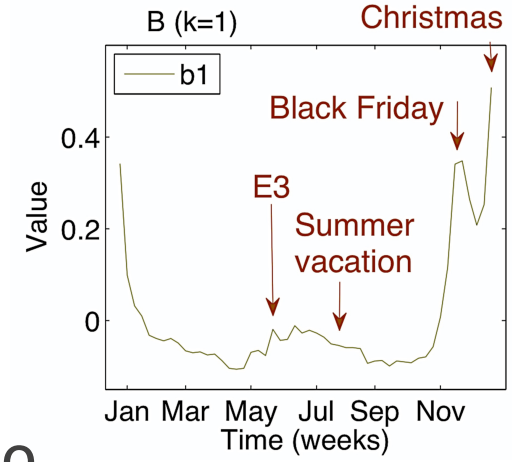




Thank you!





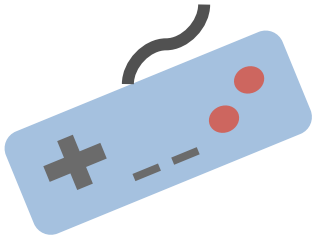


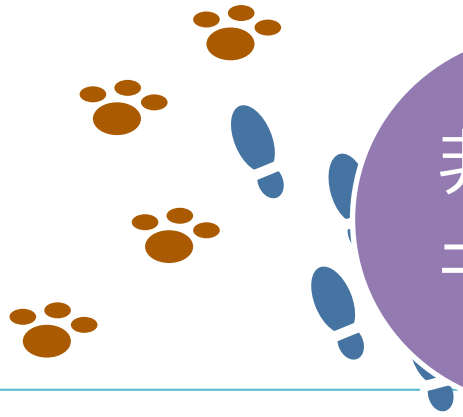
Data & Code:

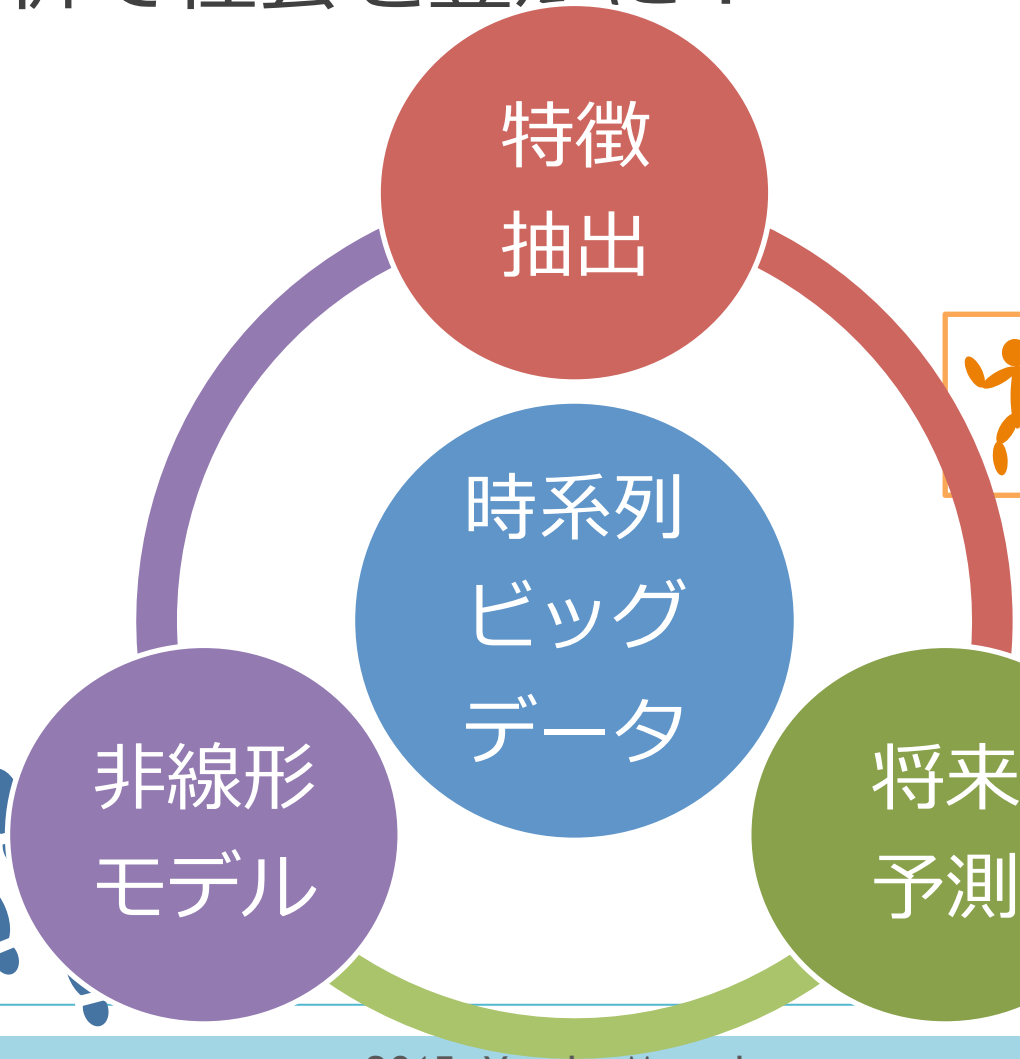
http://www.cs.kumamoto-u.ac.jp/~yasuko

目標

時系列解析で社会を豊かに！







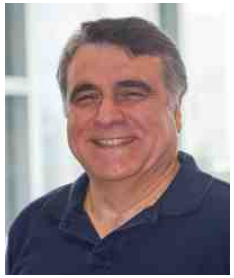
目標

時系列解析で社会を豊かに！

SIGMOD 2015 3h-Tutorial

“Mining and Forecasting of Big Time-series data”





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

目標

時系列解析で社会を豊かに！