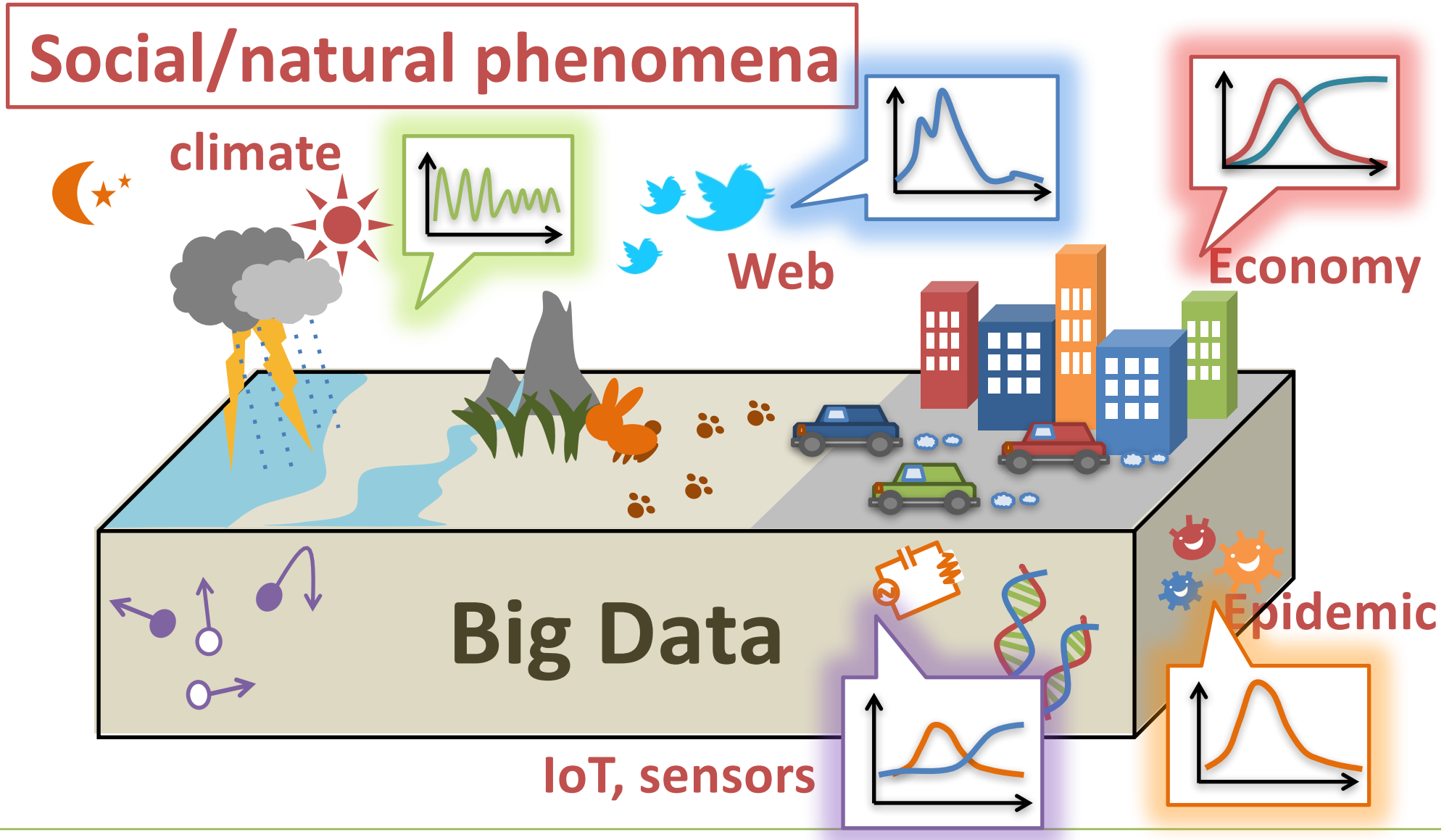
Mining and Forecasting of Big Time-series Data

Yasushi Sakurai

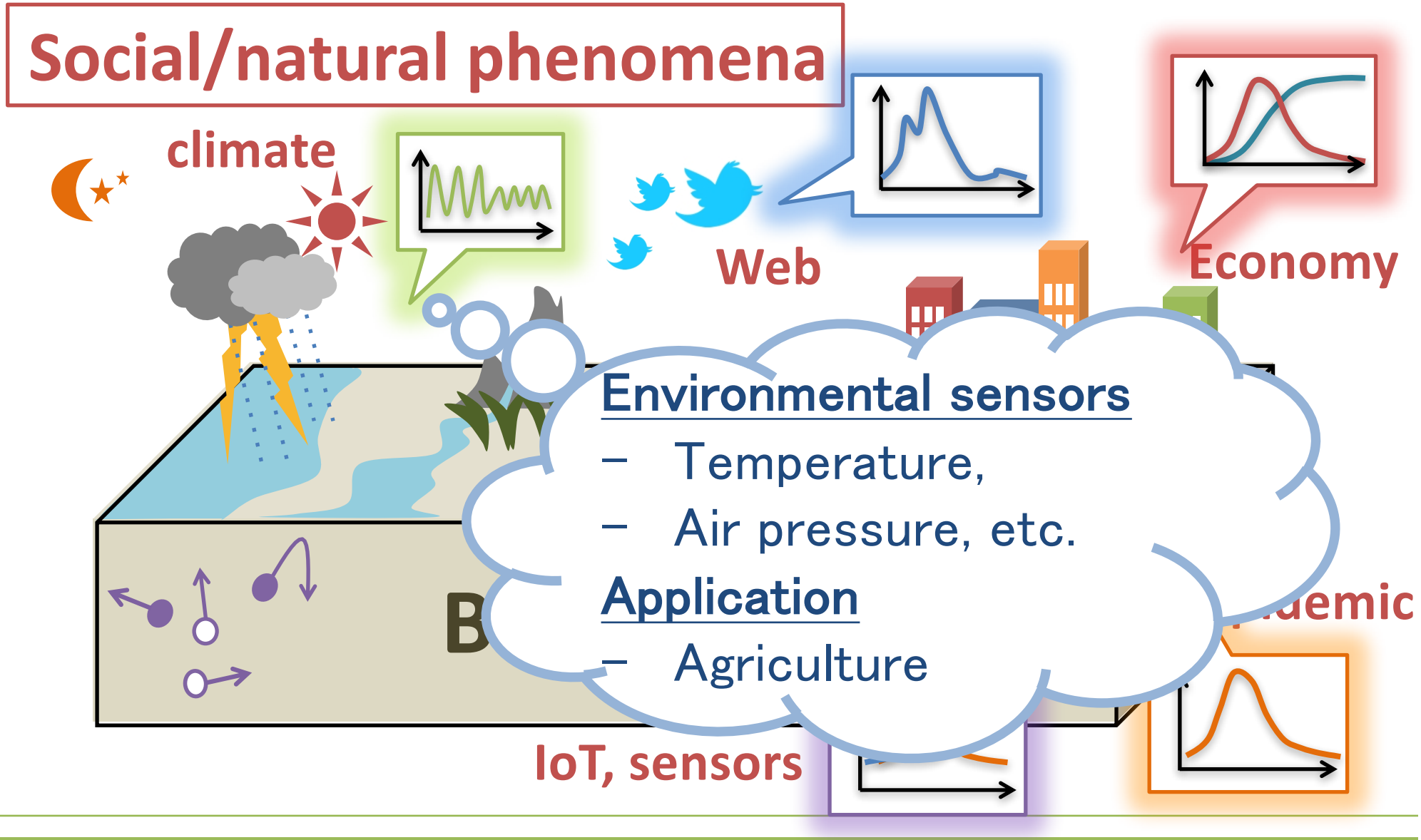
Sakurai Lab.

@ Kumamoto University

Big time-series data

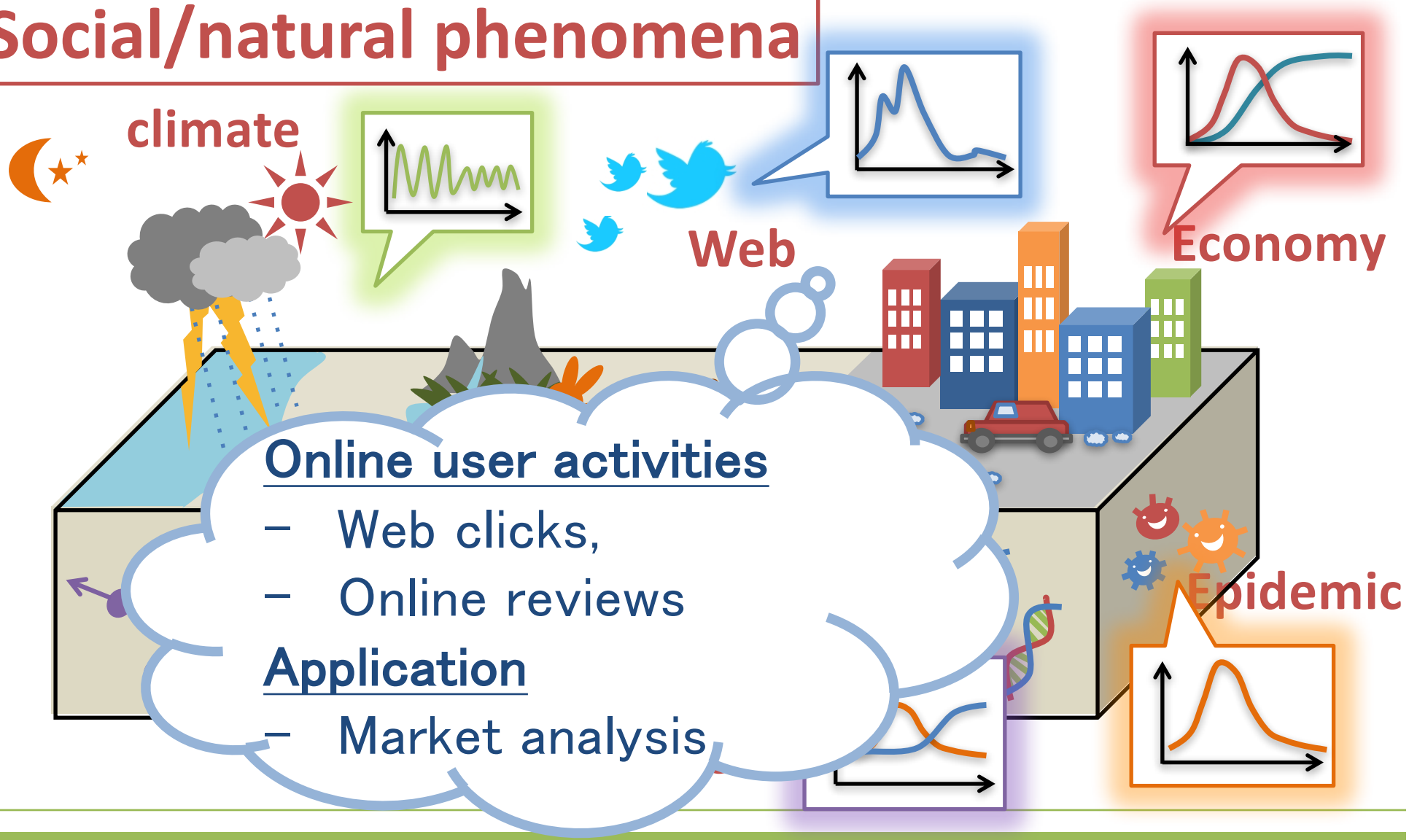


Big time-series data



Big time-series data

Social/natural phenomena



Online user activities

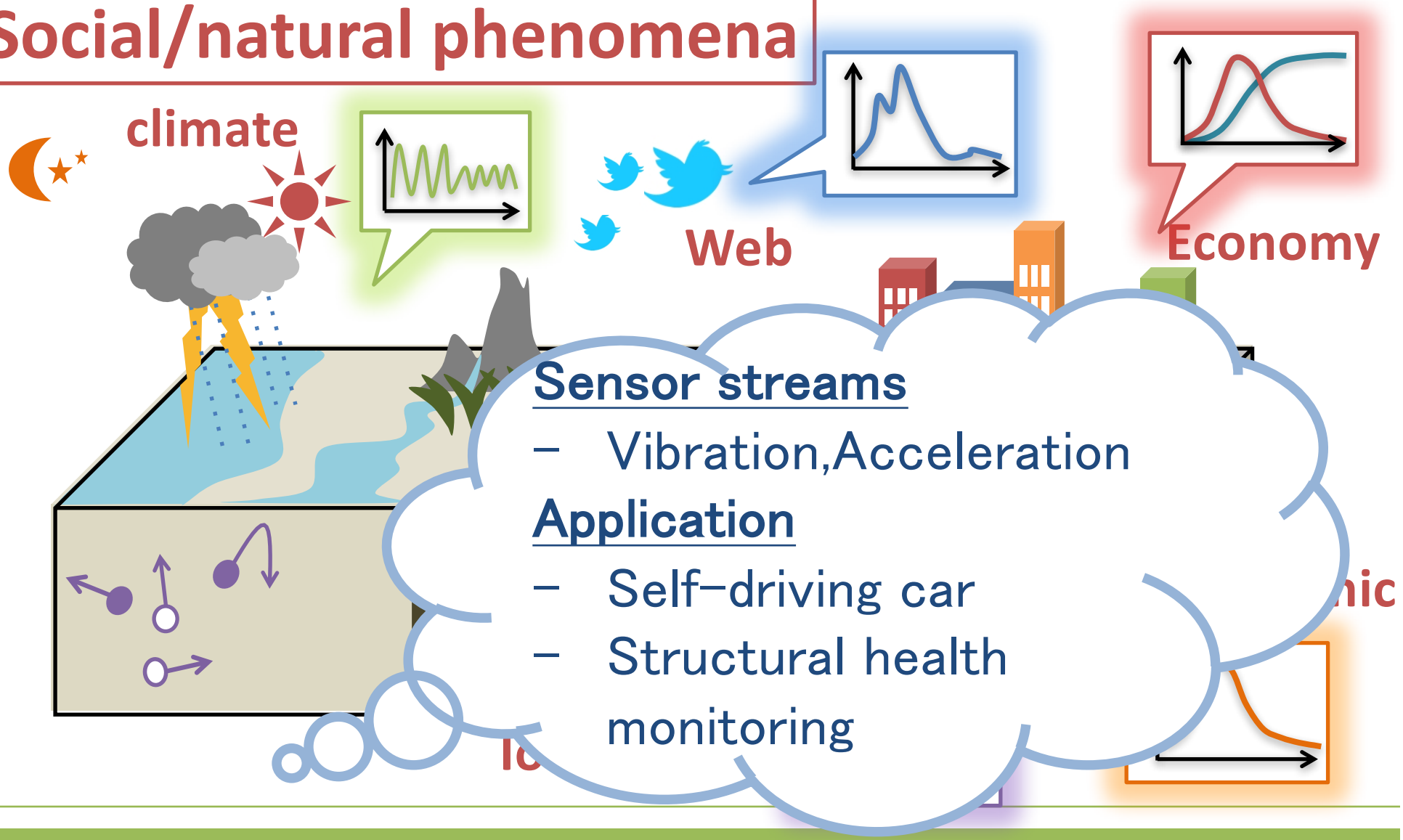
- Web clicks,
- Online reviews

Application

- Market analysis

Big time-series data

Social/natural phenomena



Sensor streams

- Vibration,Acceleration

Application

- Self-driving car
- Structural health monitoring

Motivation

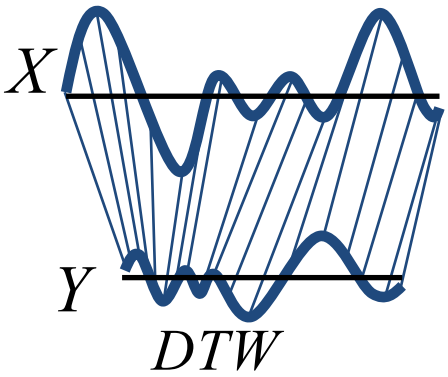
- **Given: Big time-series data**
  - Sensor data
  - Web, online activities
  - Medical data

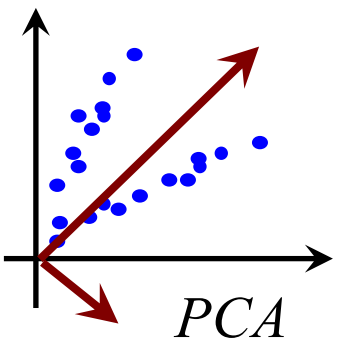
- **Goal: Real-time forecasting**
- **At-work:**
  - Online marketing
  - Sensor monitoring
  - Forecasting future epidemics

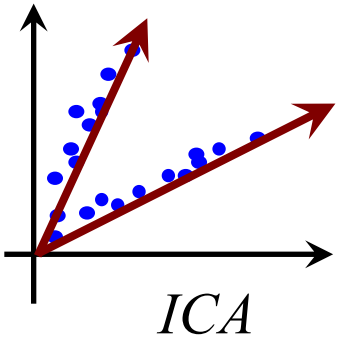
New challenges

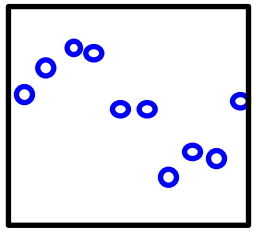
Time-series data analysis

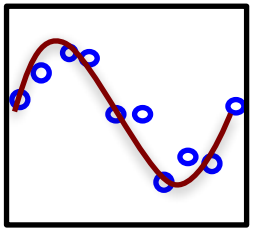
Indexing, Similarity search	Feature extraction	Linear-modeling	Data stream
ED, DTW, SAM	DFT, DWT, SVD, ICA	AR, LDS, ARIMA	Correlation monitoring etc.





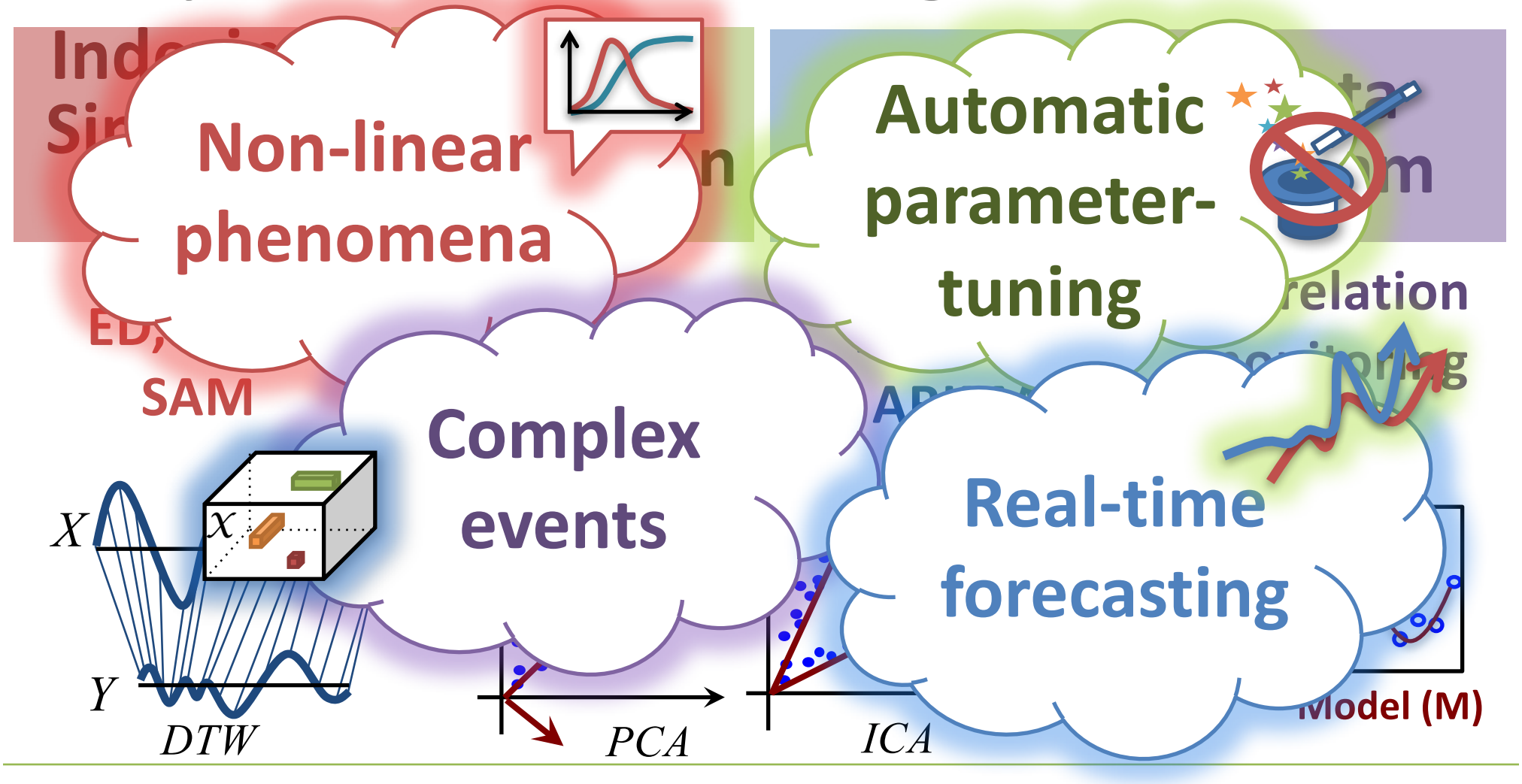






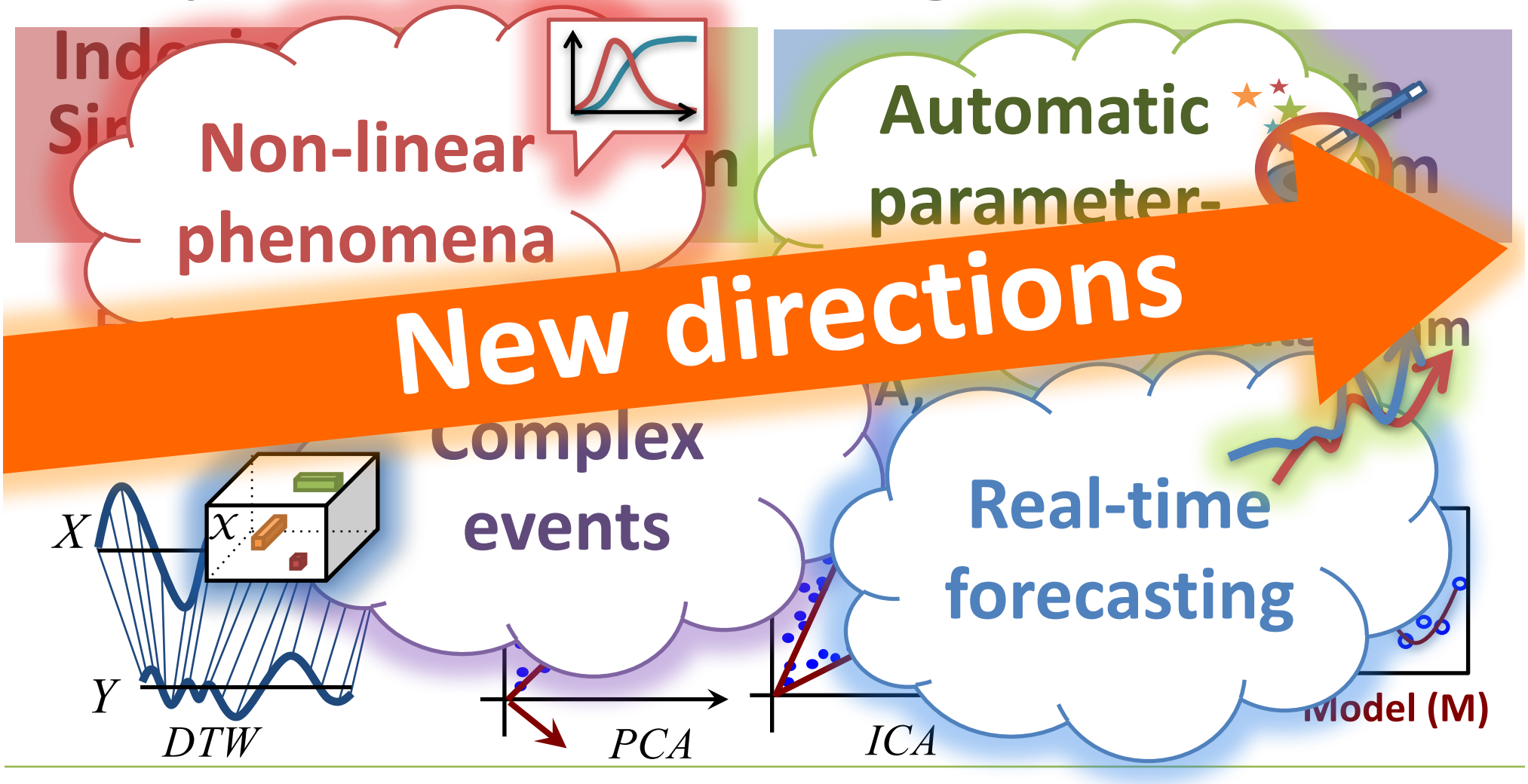
New challenges

They cannot handle "Big" data, i.e.,



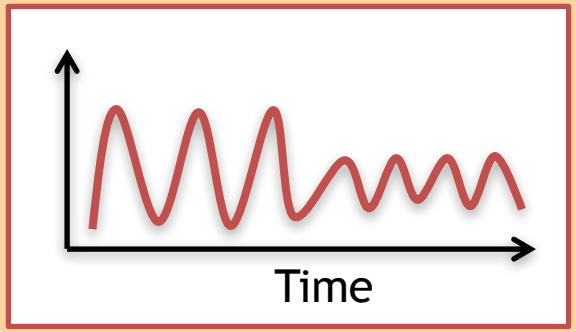
New challenges

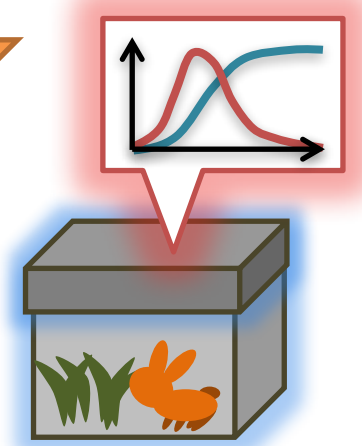
They cannot handle "Big" data, i.e.,

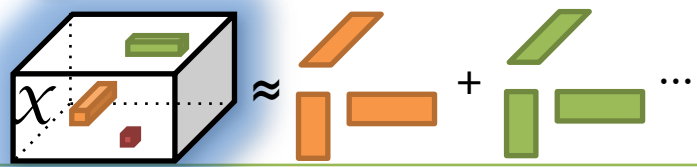


New research directions

- **Time-series data analysis**
  - Indexing and fast searching
  - Sequence matching
  - Clustering
  - etc.
- **New research directions**
  1. Automatic mining
  2. Non-linear modeling
  3. Large-scale tensor analysis







(R1) Automatic mining

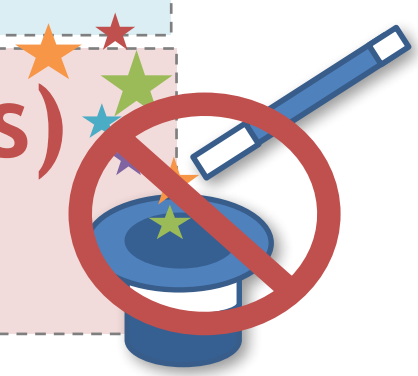
No magic numbers! … because,

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!!

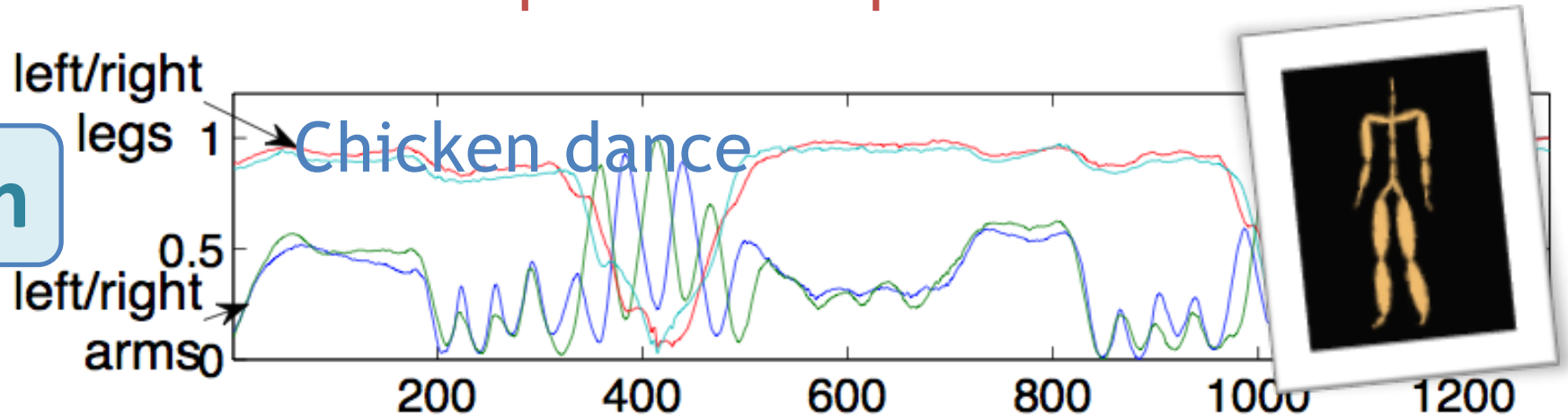
(R1) Automatic mining

AutoPlait
SIGMOD'14

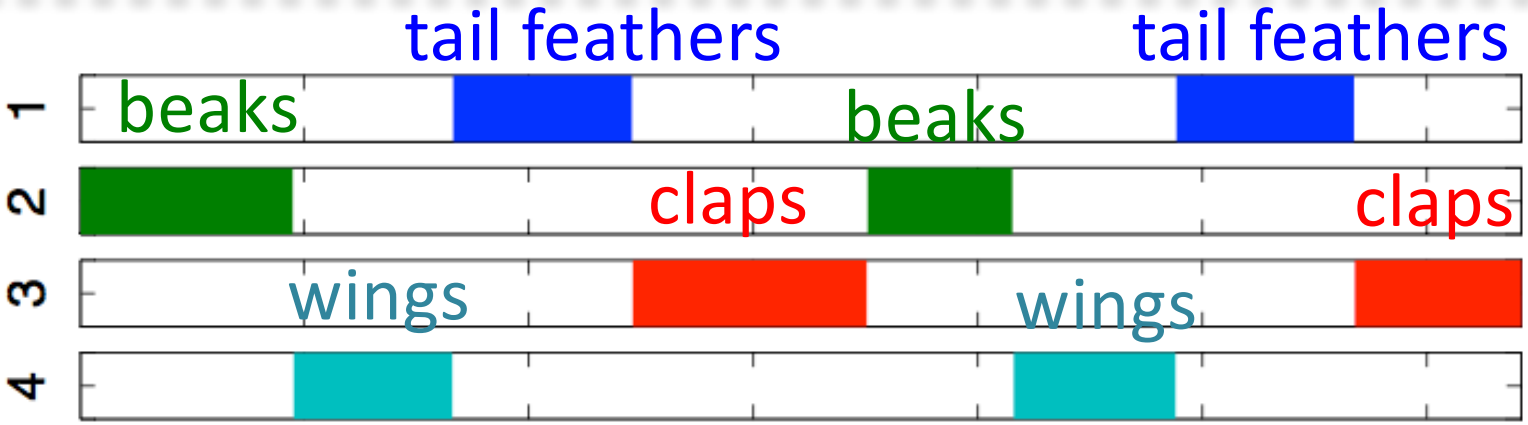
"Automatic" mining algorithm

Find: compact description of data X

Given



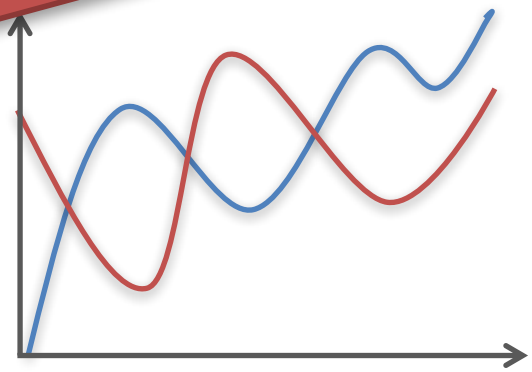
Find



(R1) Automatic mining

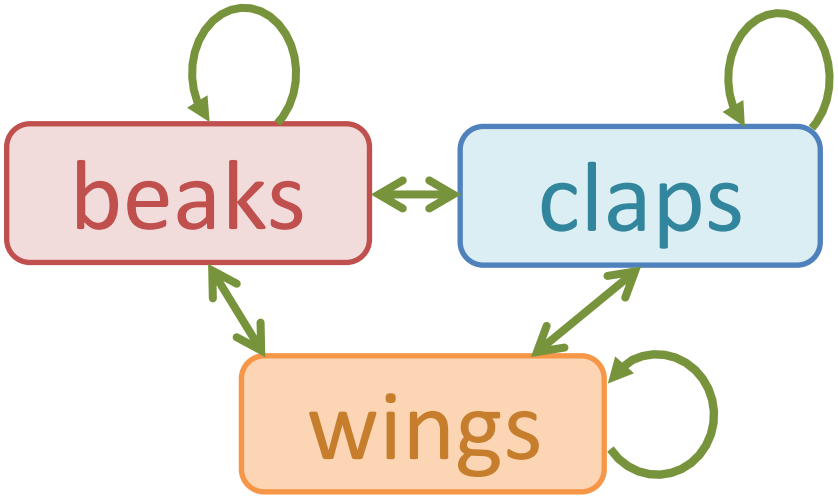
AutoPlait
SIGMOD'14

"Automatic" mining algorithm



Sequences

Model



Regimes

Idea (1): Multi-level chain model

–HMM-based probabilistic model

–with "across-regime" transitions

(R1) Automatic mining

AutoPlait
SIGMOD'14

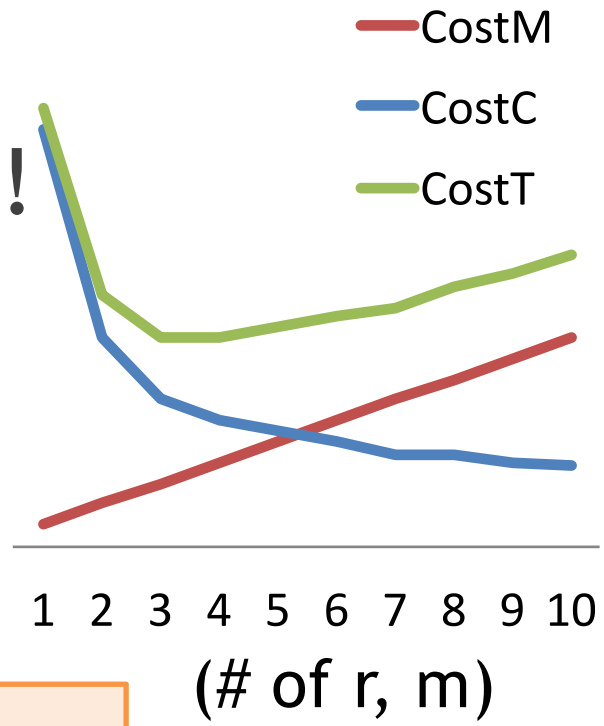
"Automatic" mining algorithm

Idea(2): Minimize encoding cost!

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

Model cost

Coding cost



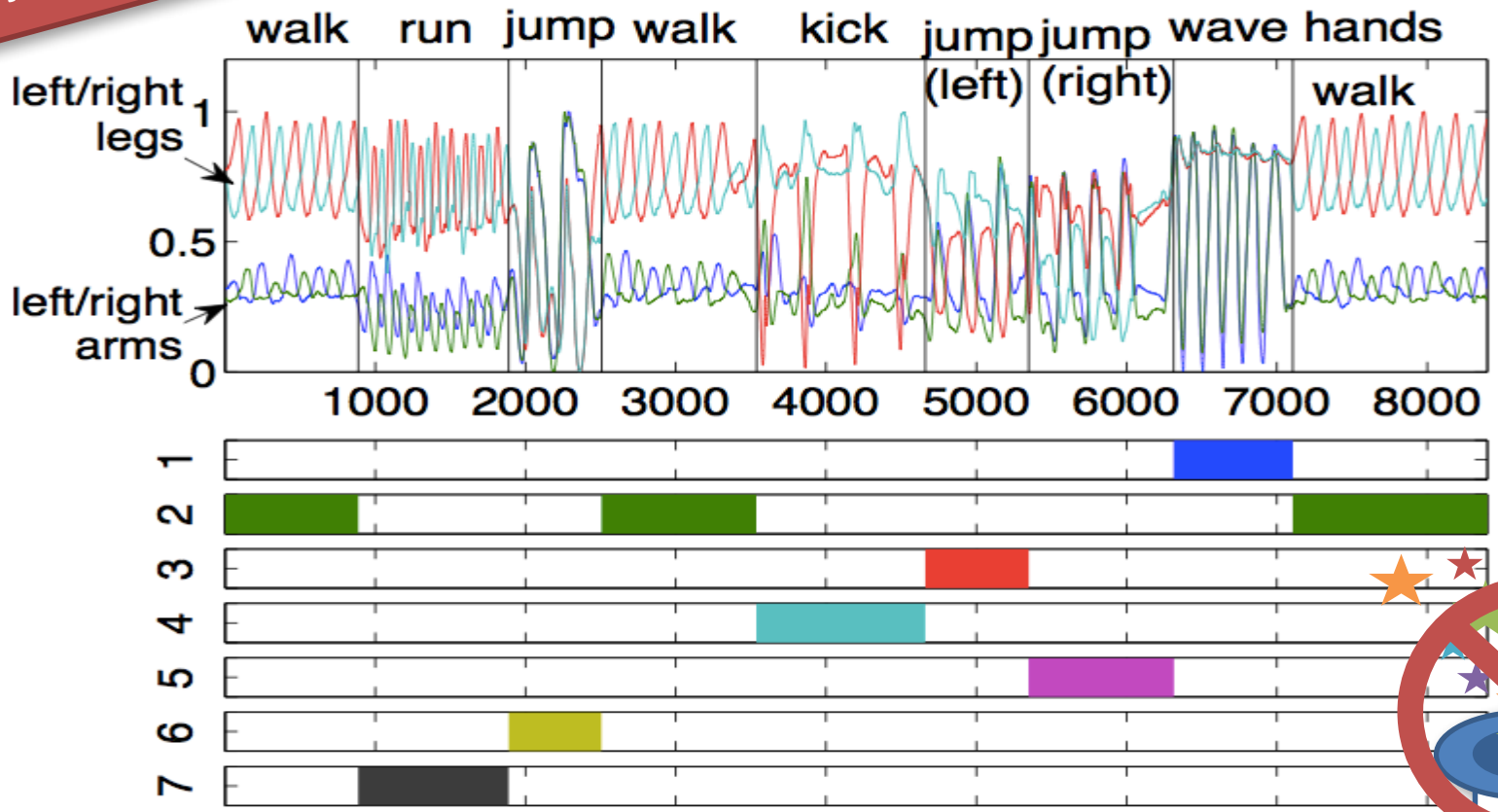
Good compression

Good description

(R1) Automatic mining

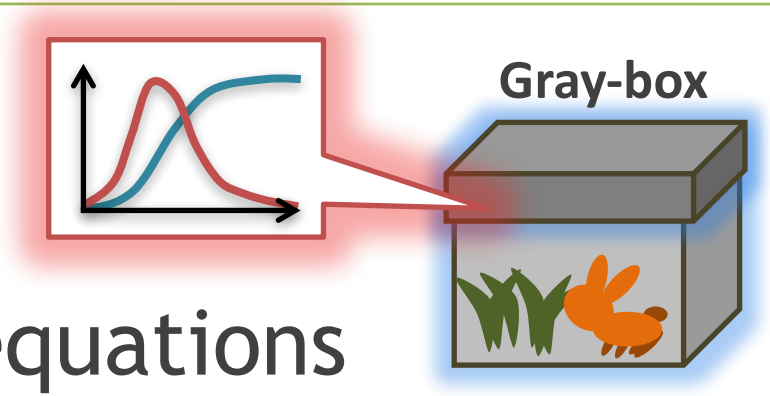
AutoPlait
SIGMOD'14

"Automatic" mining algorithm

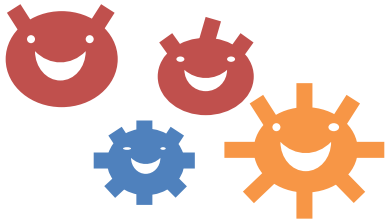


AutoPlait (**NO** magic numbers)

(R2) Non-linear (gray-box) modeling



- Gray-box mining
  - If we know the equations
- Non-linear (differential) equations
  - Epidemic
  - Biology
  - Physics, Economics, etc.



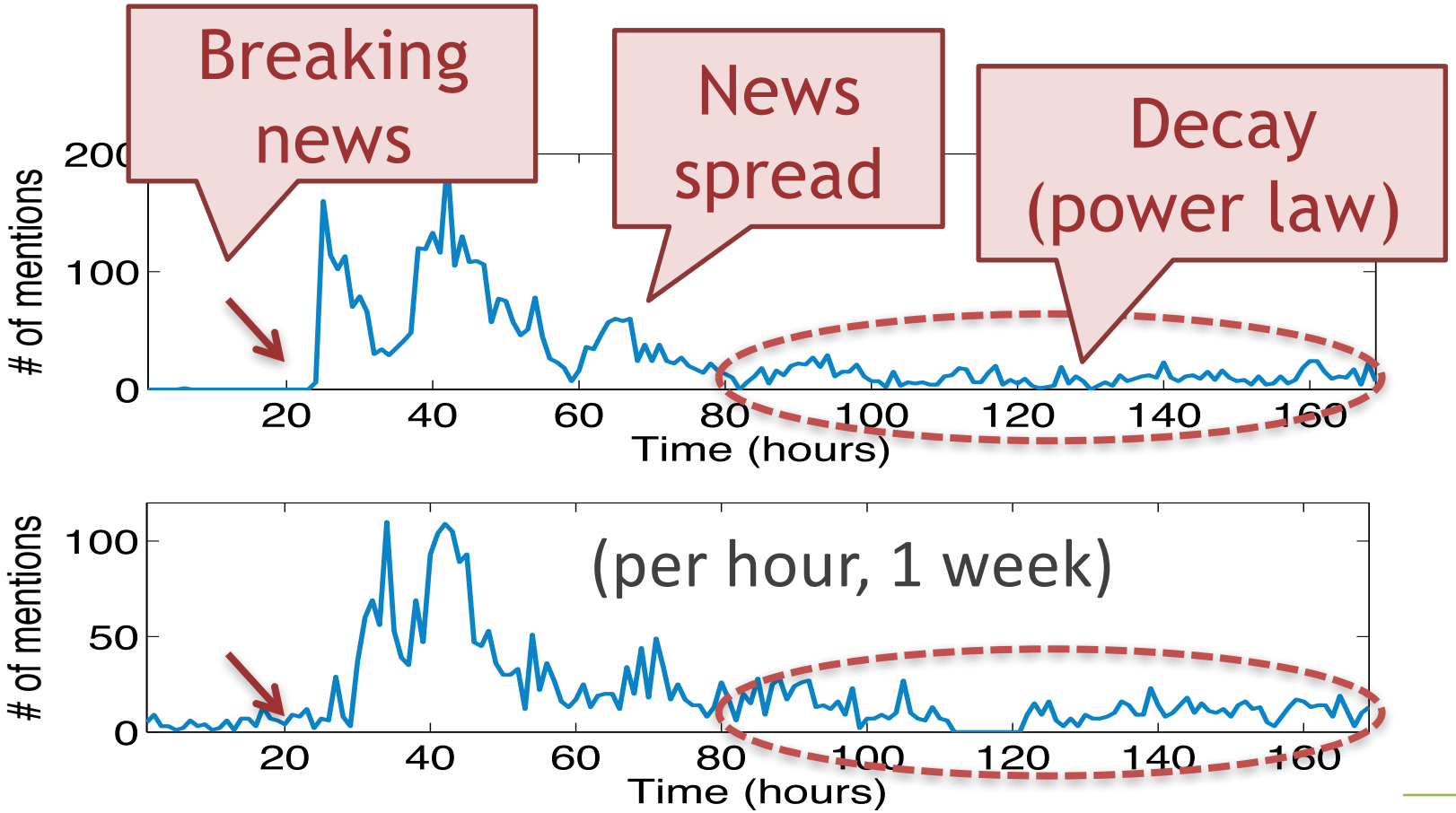
- Modeling non-linear social phenomena
  - Non-linear analysis for Web online activities (e.g., information diffusion, interaction)

(R2) Non-linear (gray-box) modeling

SpikeM
KDD'12

News spread in social media

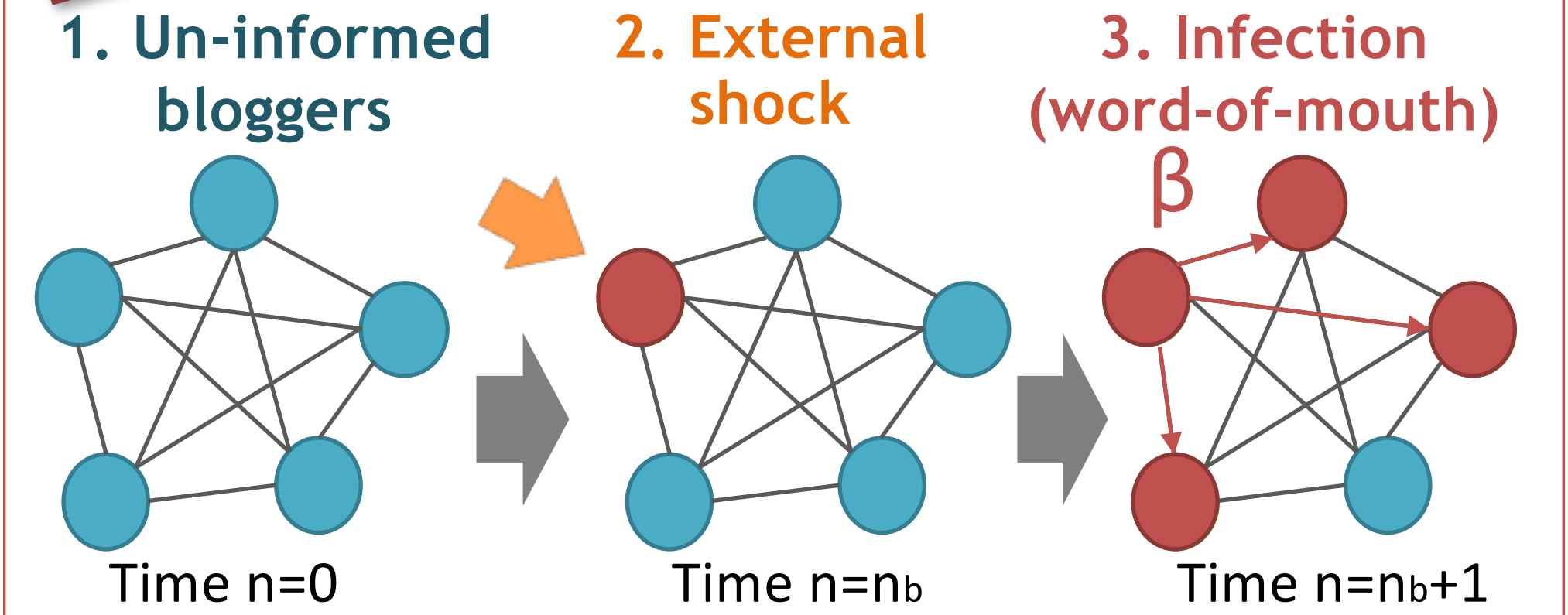
Number of mentions in blogs/Twitter



(R2) Non-linear (gray-box) modeling

SpikeM
KDD'12

News spread in social media



(R2) Non-linear (gray-box) modeling

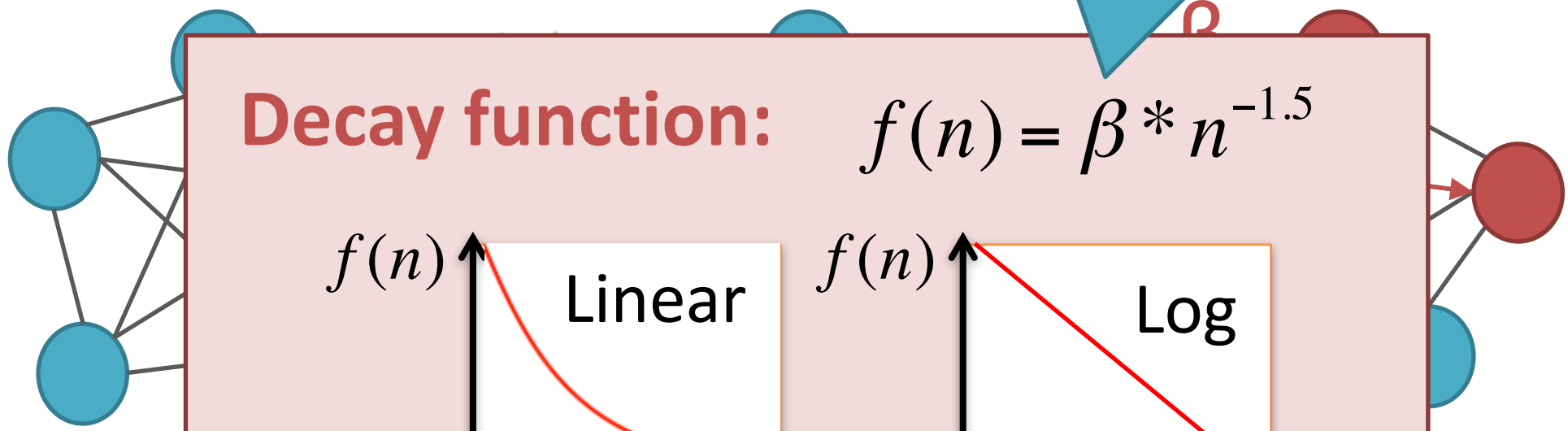
SpikeM
KDD'12

News spread in social media

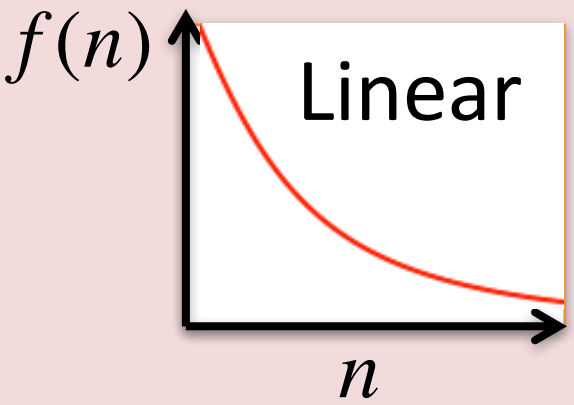
1. Un-informed bloggers

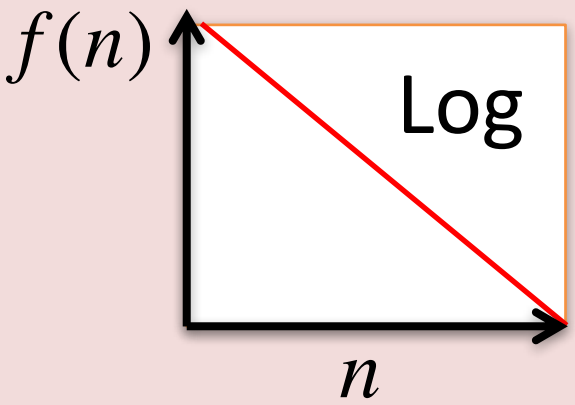
2. External shock

Power law



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



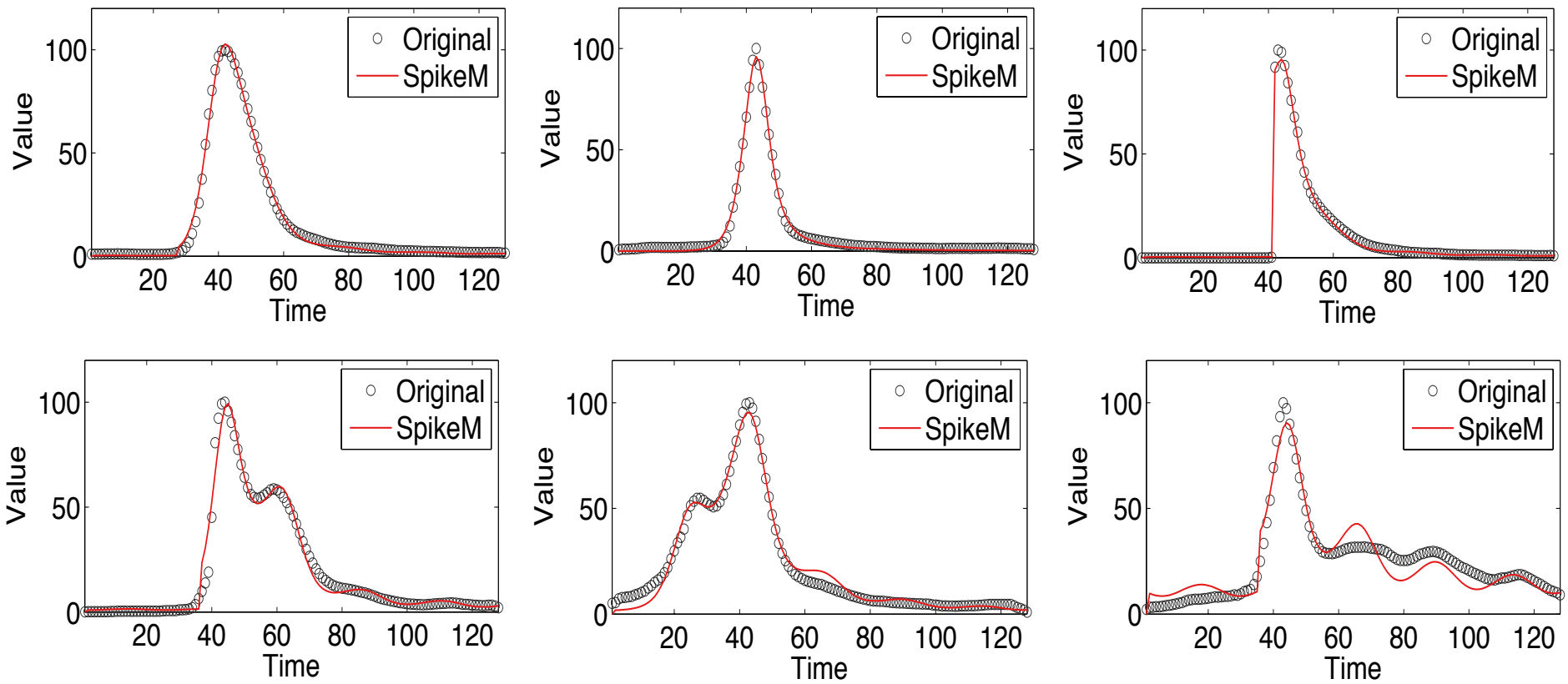


(R2) Non-linear (gray-box) modeling

SpikeM
KDD'12

News spread in social media

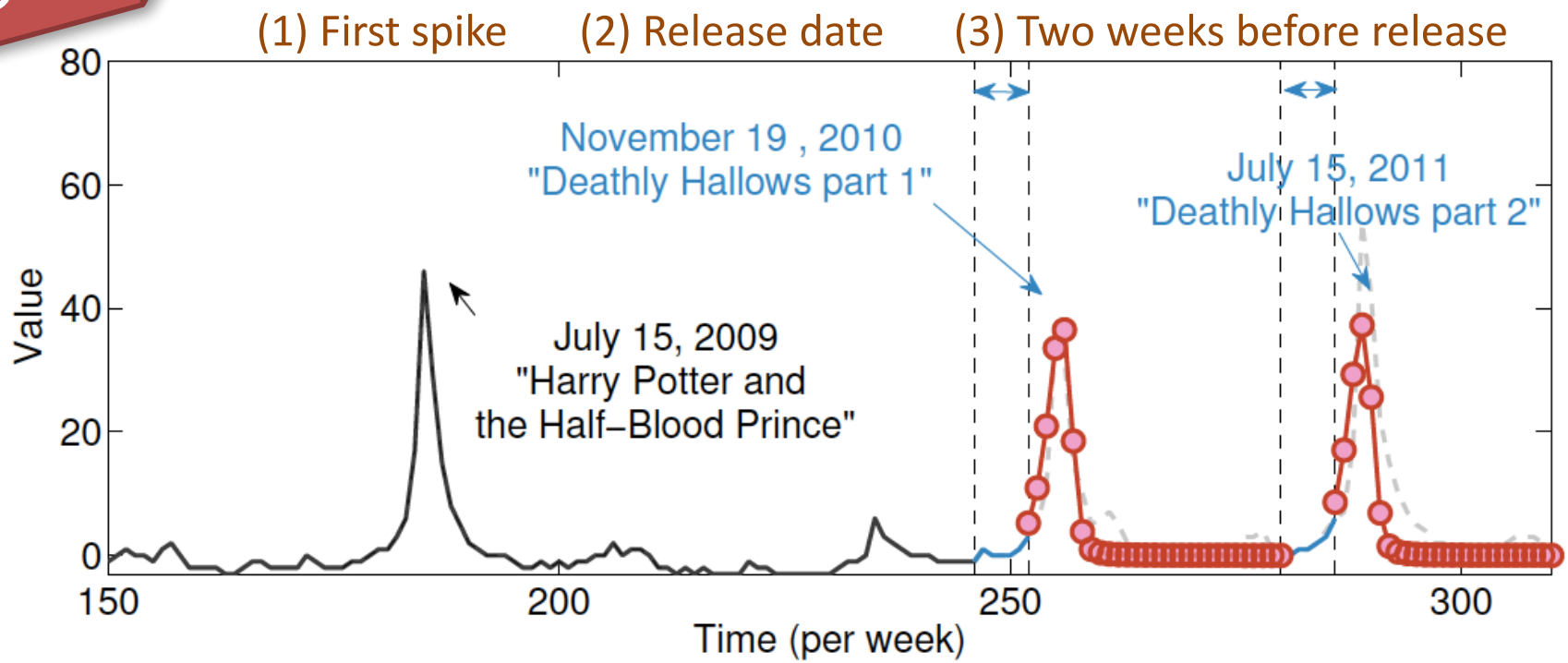
Memes in social media/blogs



(R2) Non-linear (gray-box) modeling

SpikeM
KDD'12

Forecast tail-part and **rise-part**!

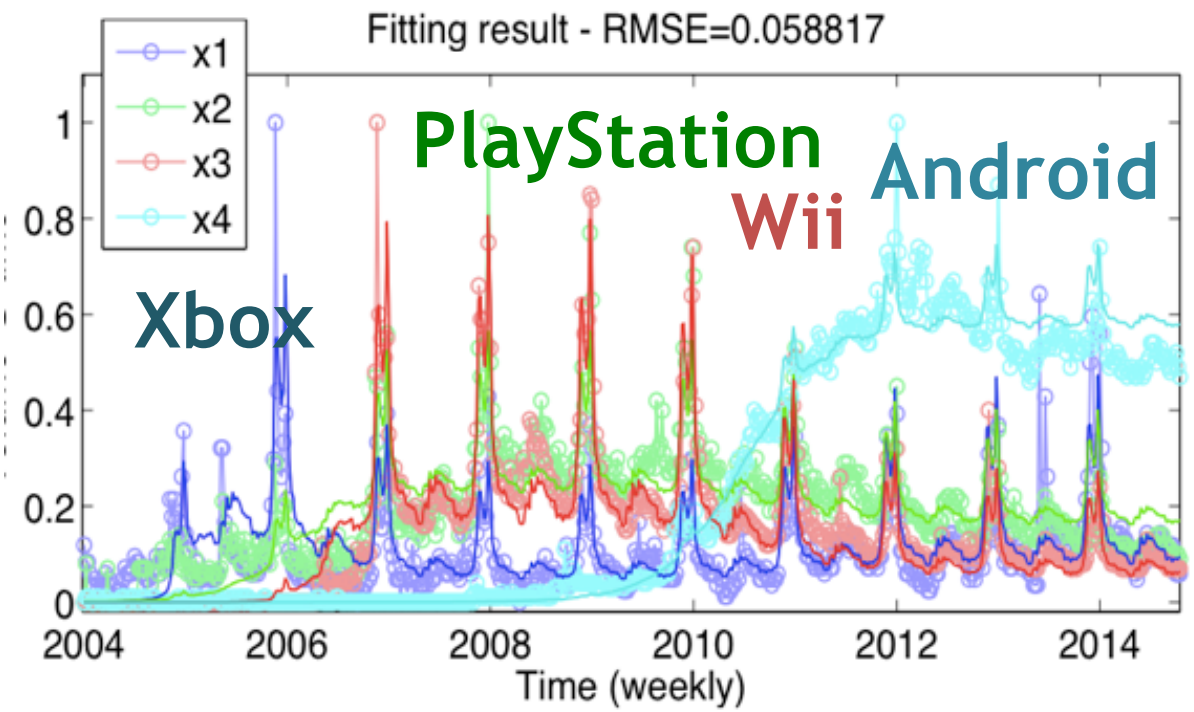


Given (1) first spike,
(2) release date of two sequel movies
(3) access volume before the release date

(R2) Non-linear (gray-box) modeling

EcoWeb
WWW'15

Online competition



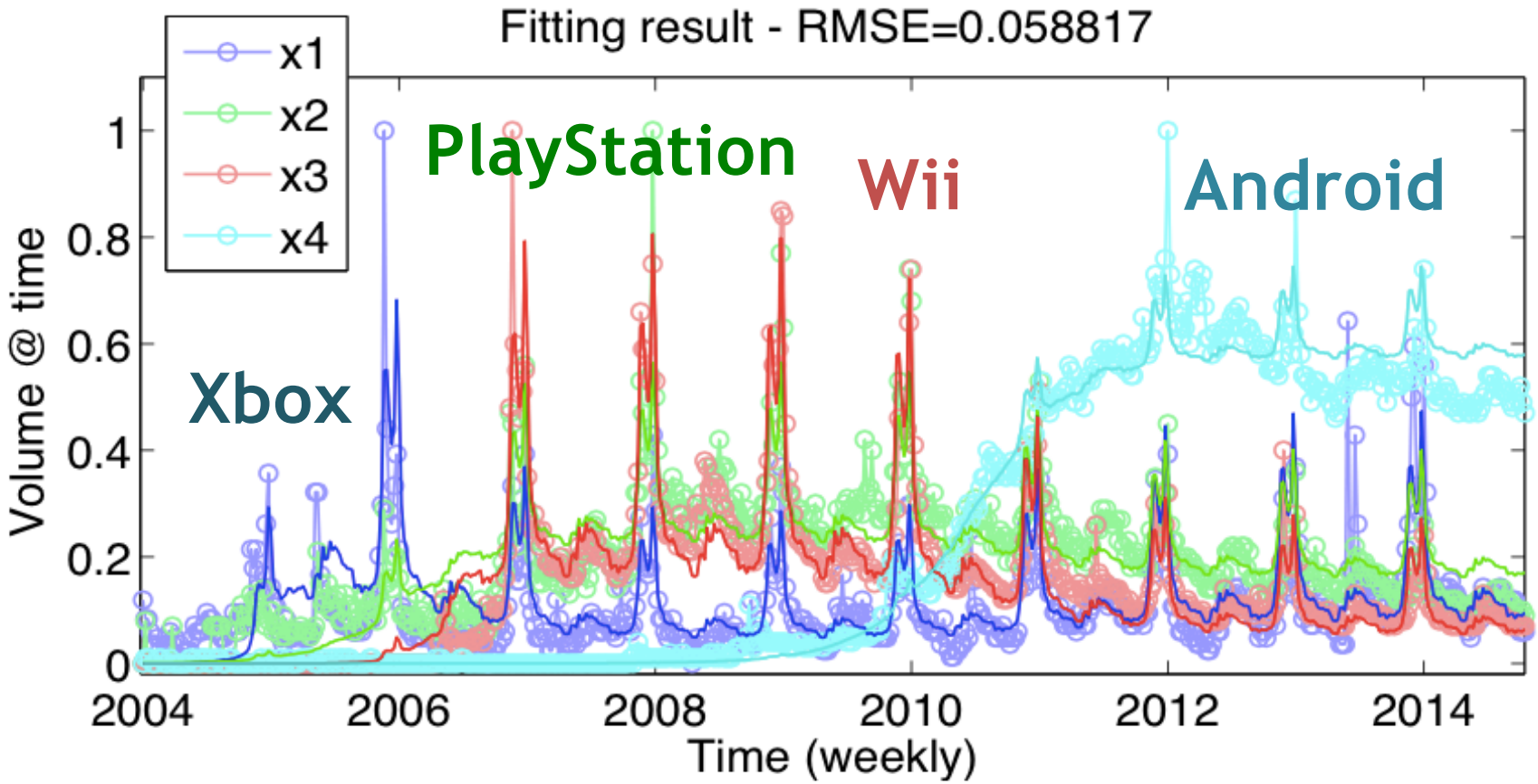
Interactions between keywords

(R2) Non-linear (gray-box) modeling

EcoWeb WWW'15

Online competition

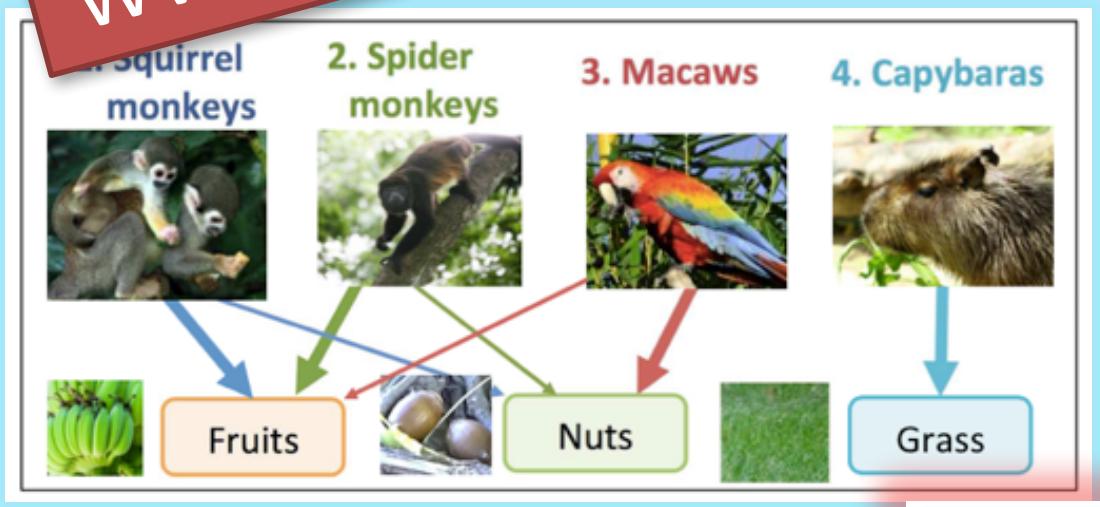
(*Google* Search volume)



(R2) Non-linear (gray-box) modeling

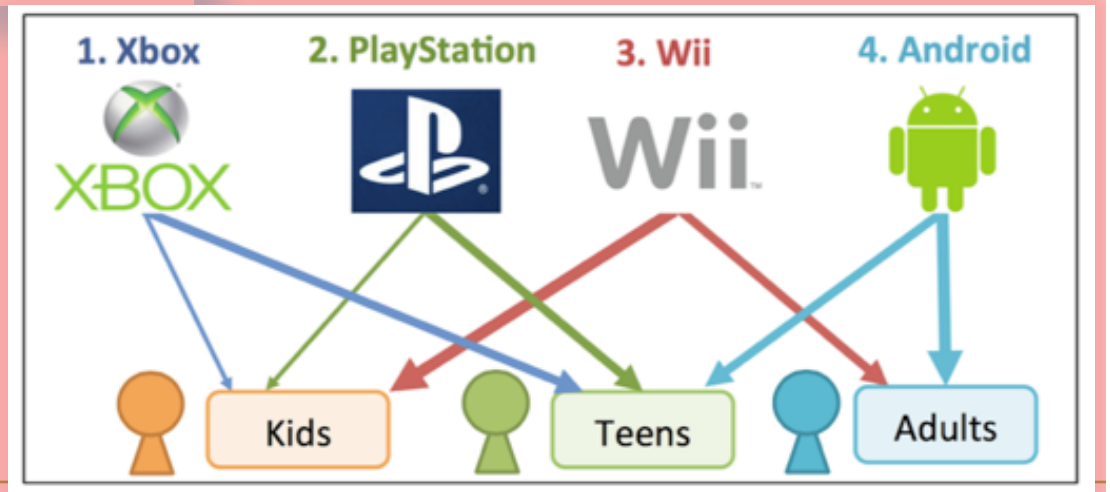
EcoWeb
WWW'15

The Web as a Jungle !



Ecosystem on the **Web**

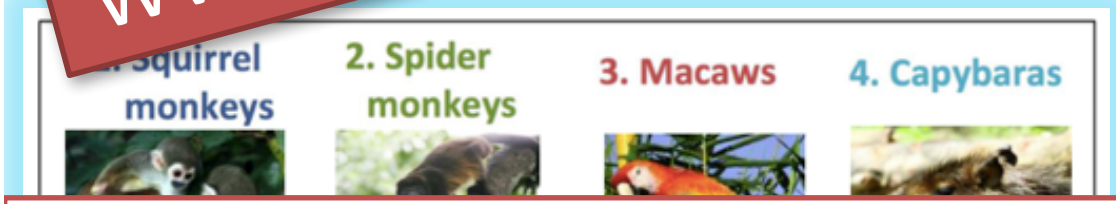
Ecosystem in the **Jungle**



(R2) Non-linear (gray-box) modeling

EcoWeb WWW'15

The Web as a Jungle !

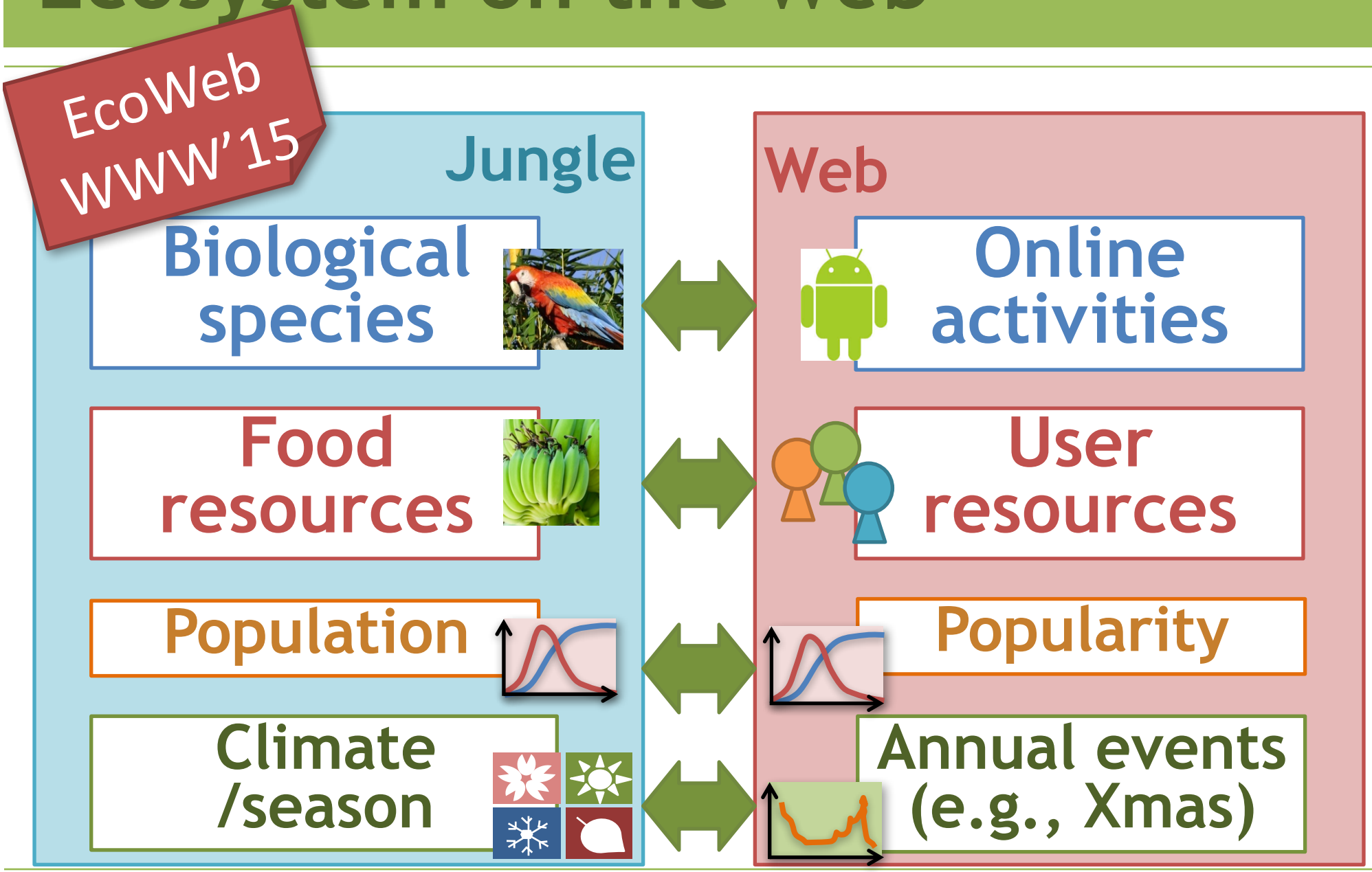


Ecosystem

--- Non-Linear equations ---

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

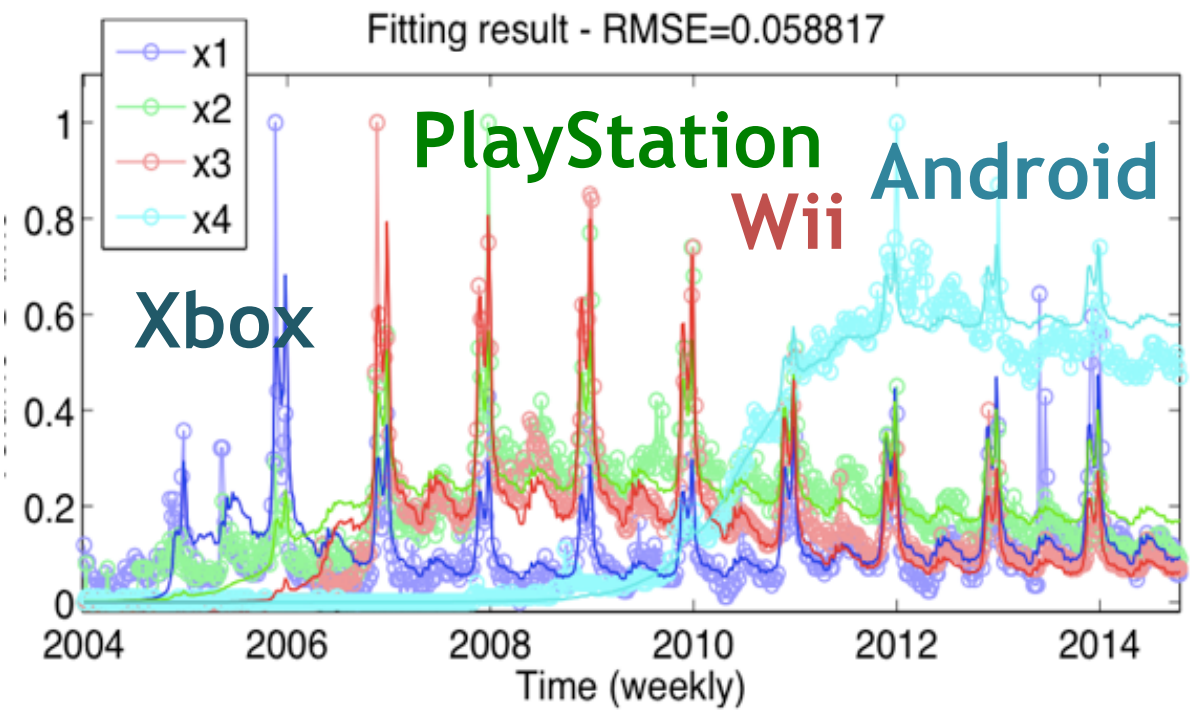
Ecosystem on the Web



(R2) Non-linear (gray-box) modeling

EcoWeb
WWW'15

The Web as a Jungle !

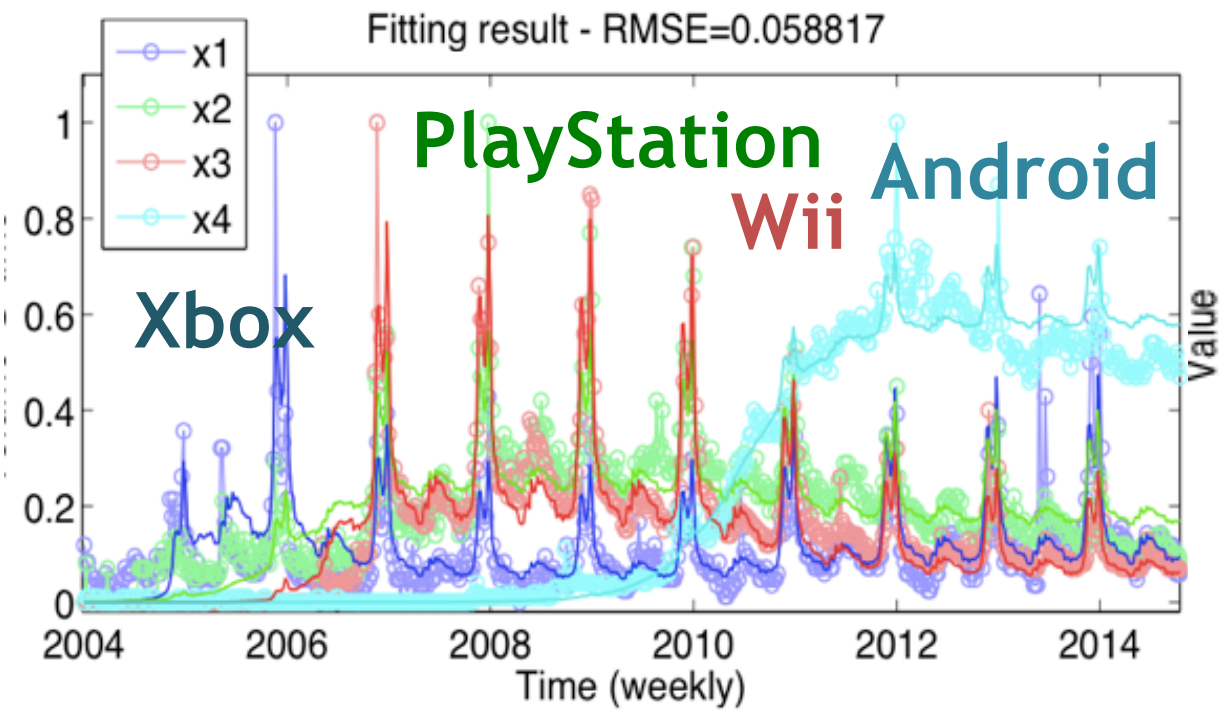


Interactions between keywords

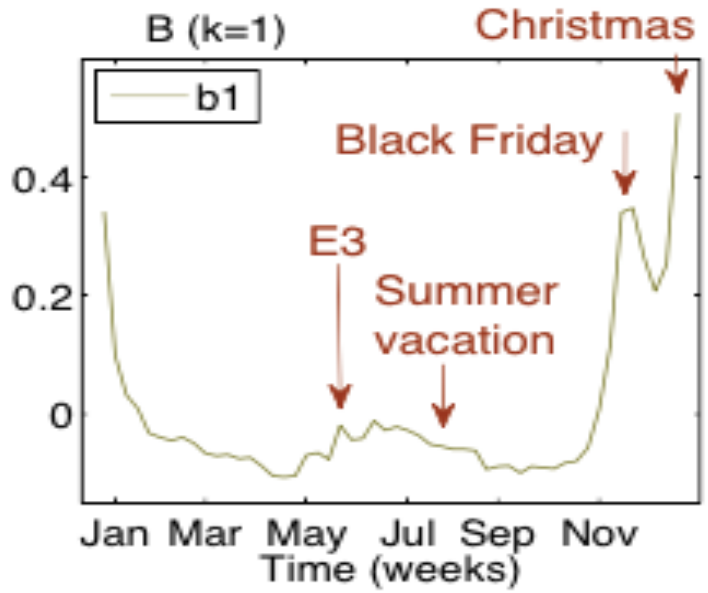
(R2) Non-linear (gray-box) modeling

EcoWeb
WWW'15

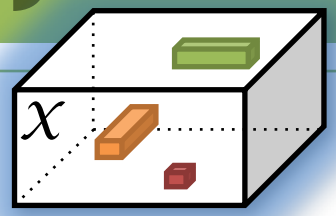
The Web as a Jungle !



Seasonality

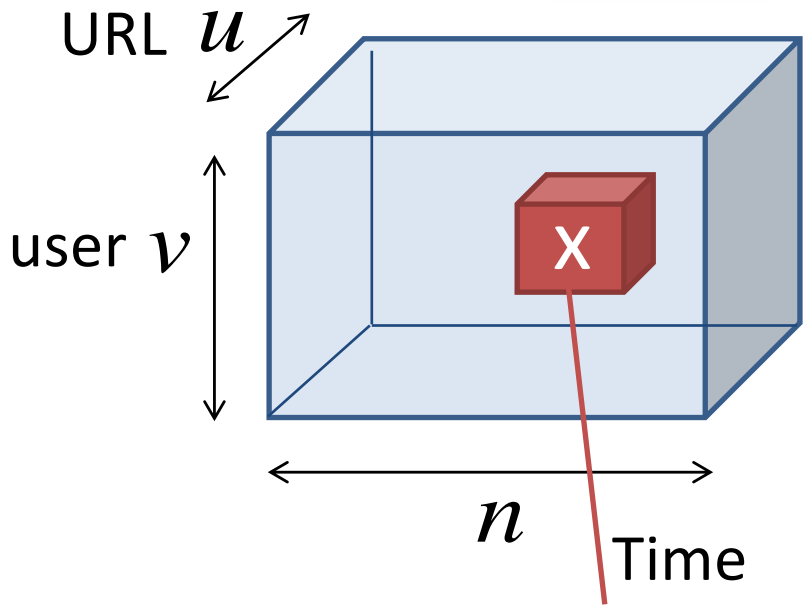


(R3) Large-scale tensor analysis



- Time-stamped events
  - *e.g., web clicks*

Time	URL	User
08-01-12:00	CNN.com	Smith
08-02-15:00	YouTube.com	Brown
08-02-19:00	CNET.com	Smith
08-03-11:00	CNN.com	Johnson
…	…	…



Represent as
Mth order tensor (M=3)

$$\mathcal{X} \in \mathbb{N}^{u \times v \times n}$$

Element x: # of events

e.g., 'Smith', 'CNN.com', 'Aug 1, 10pm'; 21 times

(R3) Large-scale tensor analysis

TriMine
KDD'12

Complex time-stamped events

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

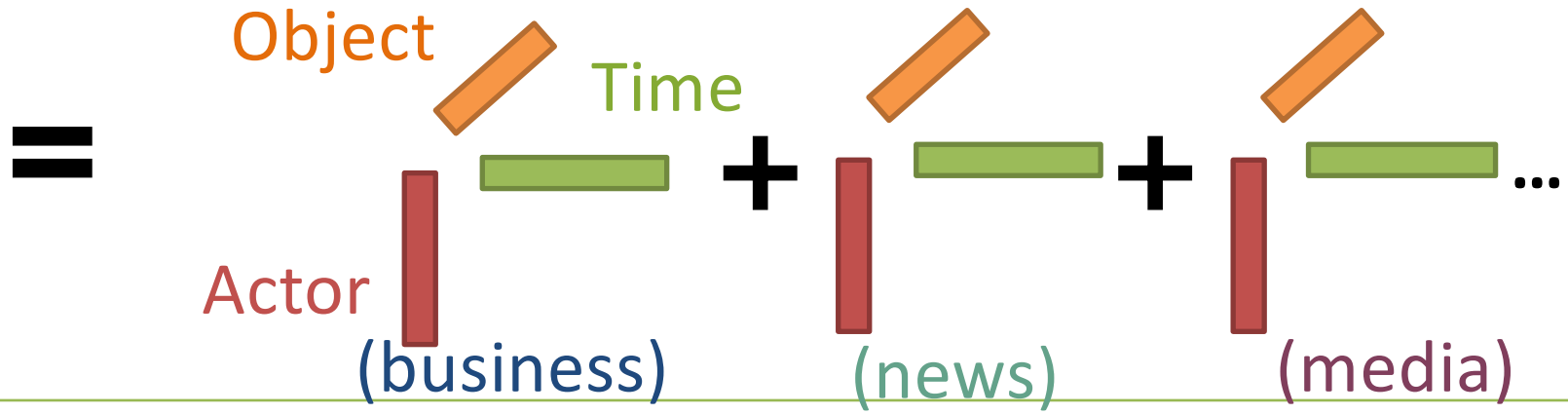
Forecast

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

(R3) Large-scale tensor analysis

TriMine
KDD'12

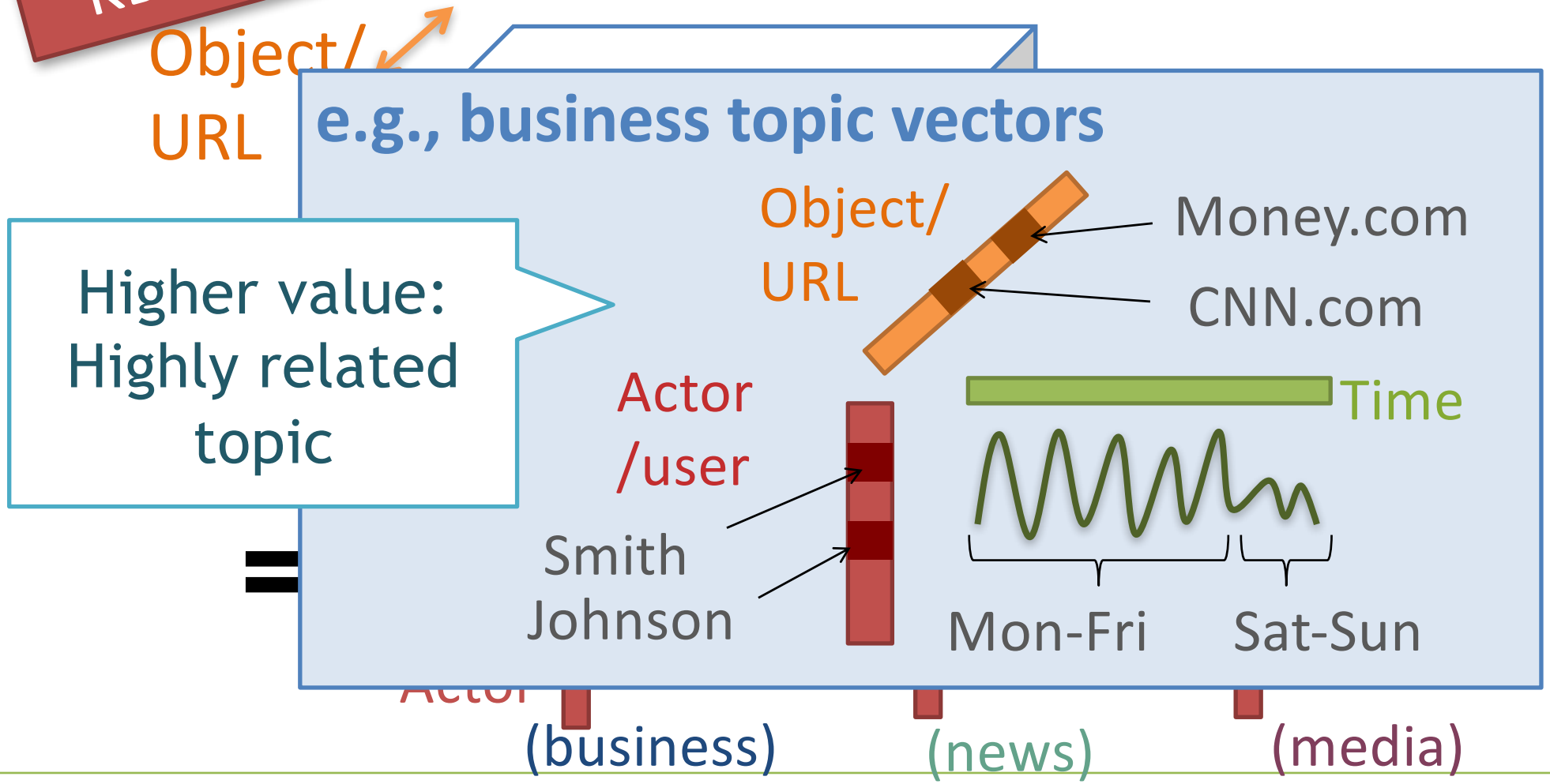
Complex time-stamped events



(R3) Large-scale tensor analysis

TriMine
KDD'12

Complex time-stamped events

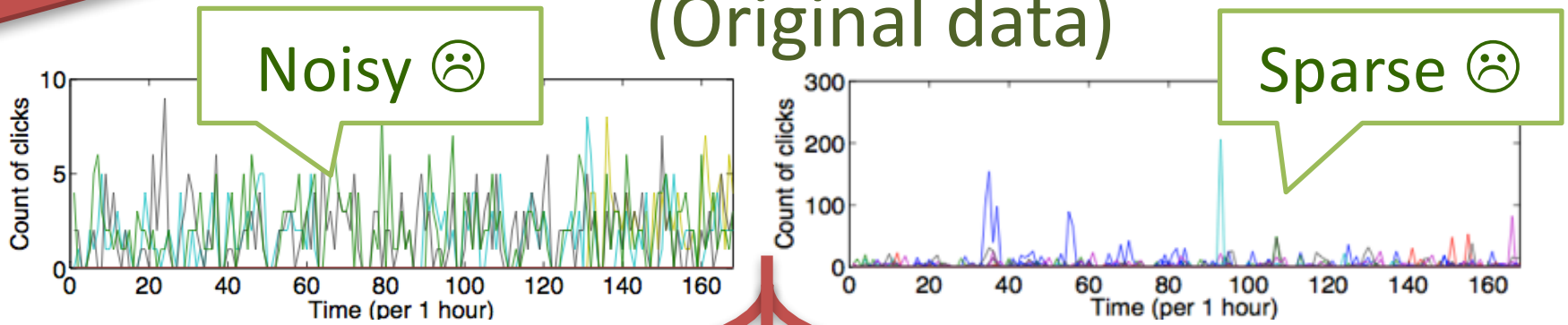


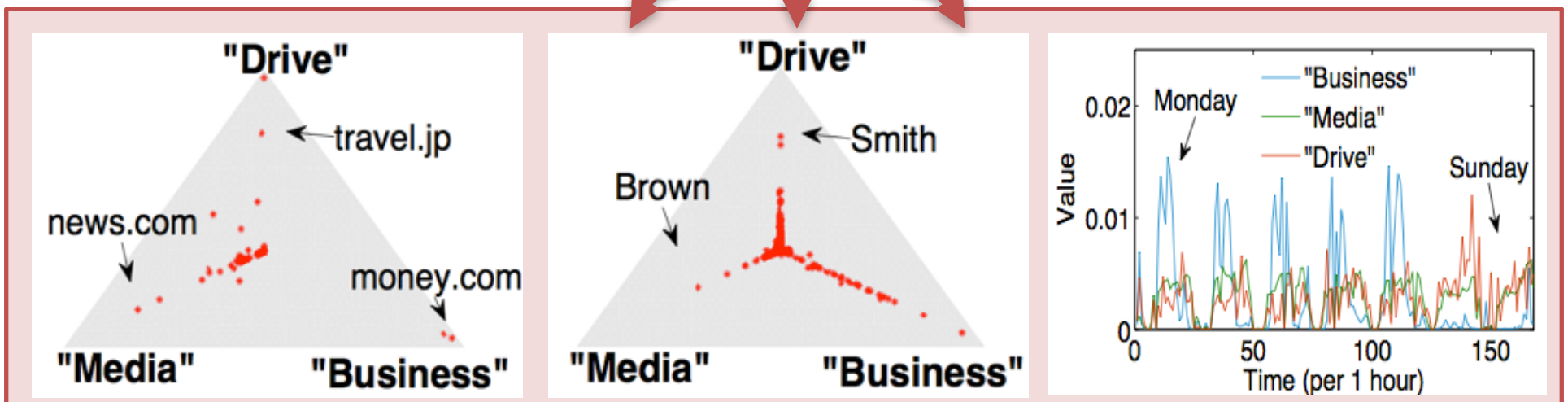
(R3) Large-scale tensor analysis

TriMine KDD'12

Complex time-stamped events

(Original data)





URL matrix | User matrix | Time matrix

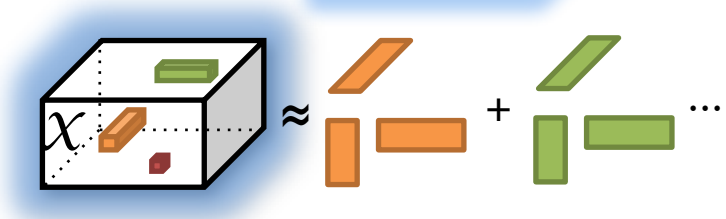
Putting it all together

3h-Tutorials
SIGMOD'15
WWW'16

New research directions

1. Automatic mining
2. Non-linear (gray-box) modeling
3. Large-scale tensor analysis

Put all together



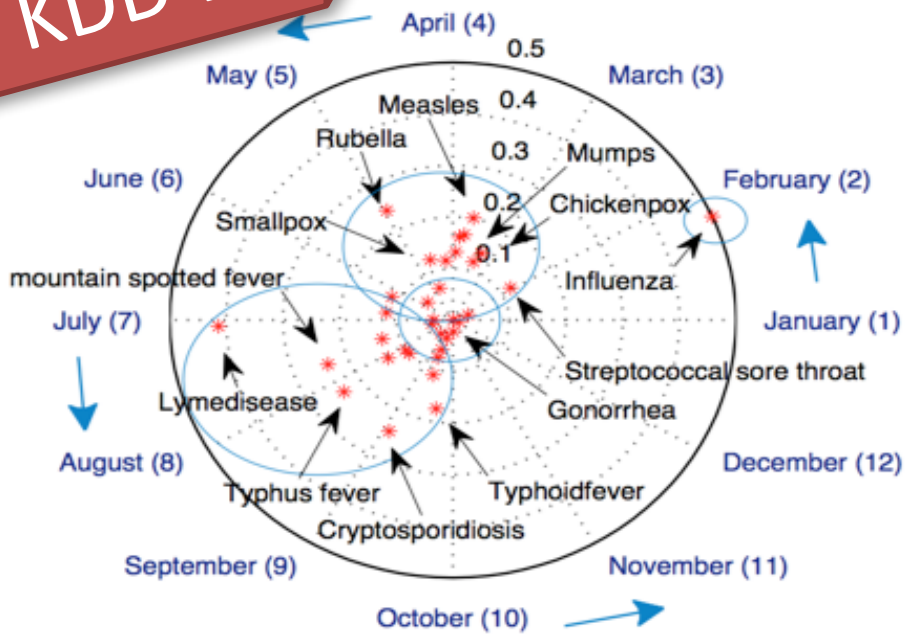
New challenge: MANT analysis

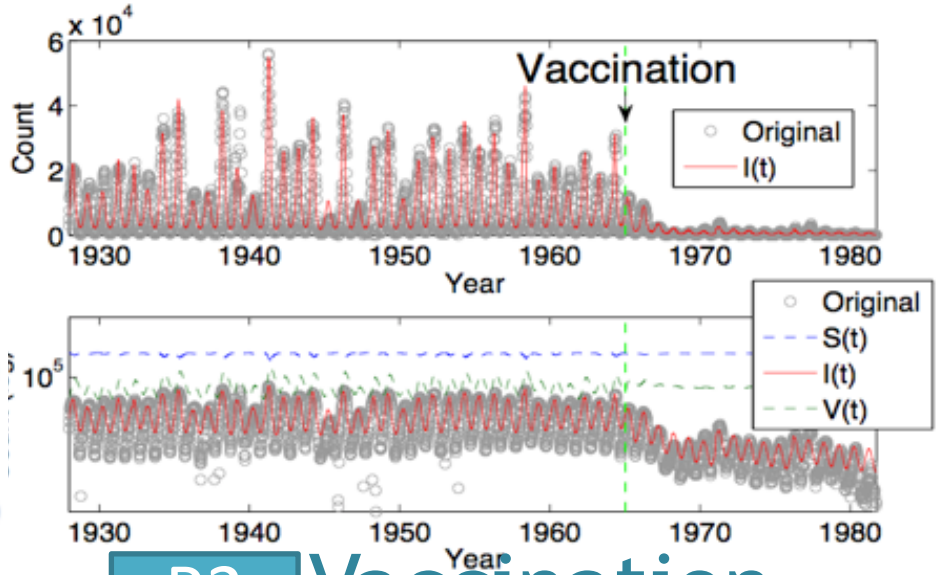
Multi-Aspect Non-linear Time-series

New challenge: MANT analysis

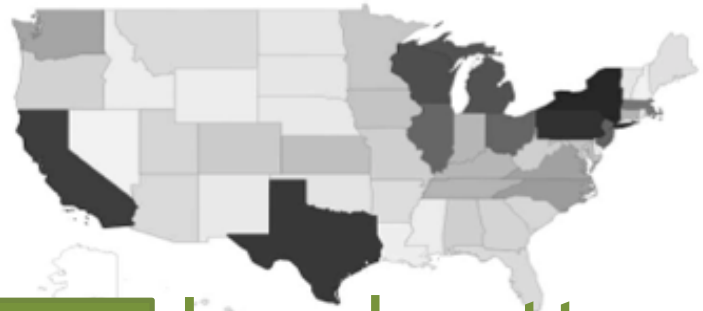
FUNNEL
KDD'14

Non-Linear tensor analysis





P2 Vaccination



s

P1 Seasonality

P4 External shocks

P5 Mistakes, errors

P3 Local patterns

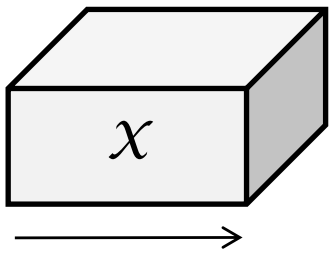
New challenge: MANT analysis

FUNNEL
KDD'14

Non-Linear tensor analysis

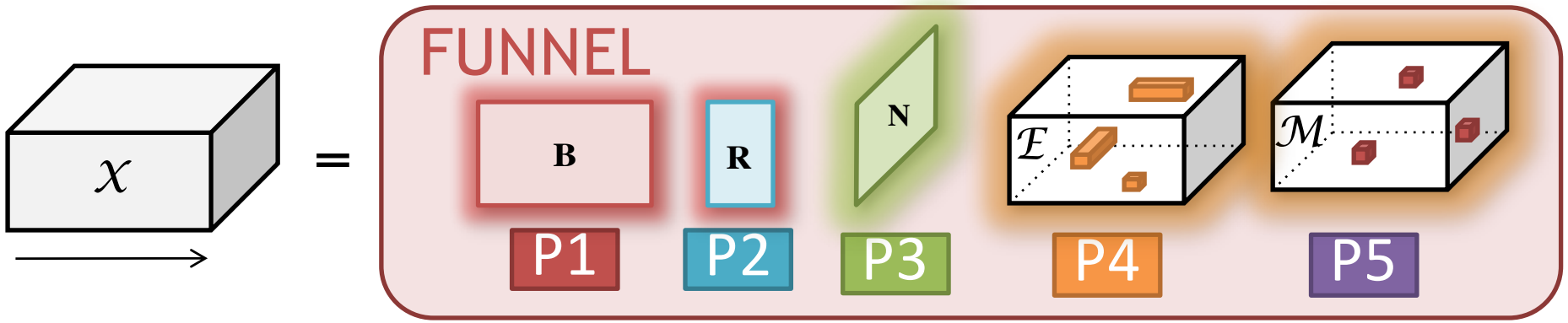
Given:

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



Find:

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



New challenge: MANT analysis

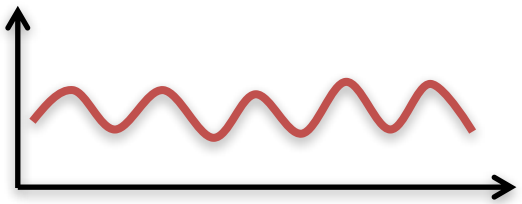
FUNNEL
KDD'14

Non-Linear tensor analysis

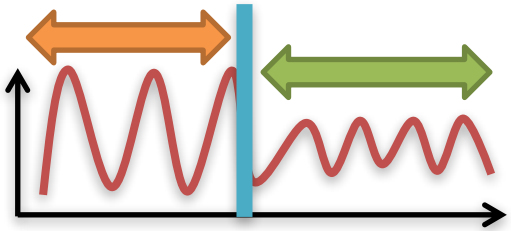
Given:

Ten ocatio

Seasonality

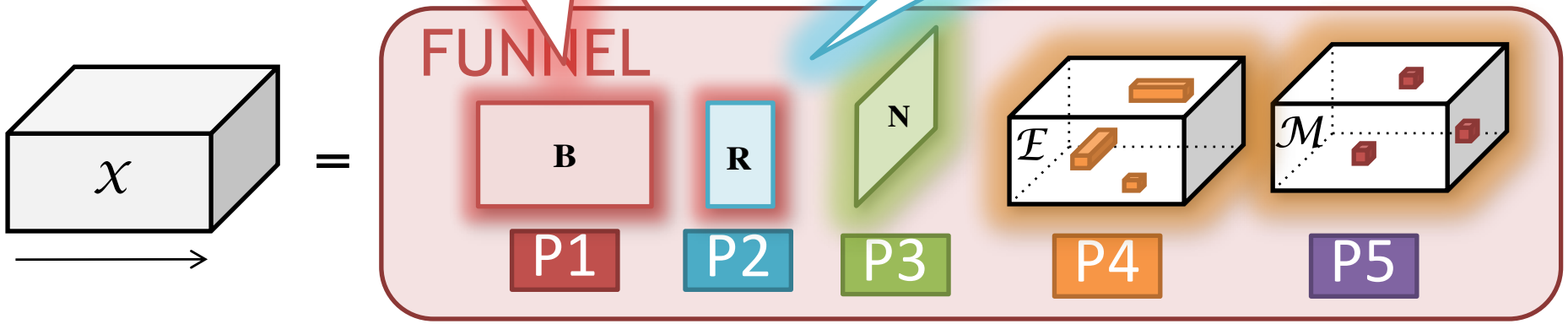


Discontinuities



Fin

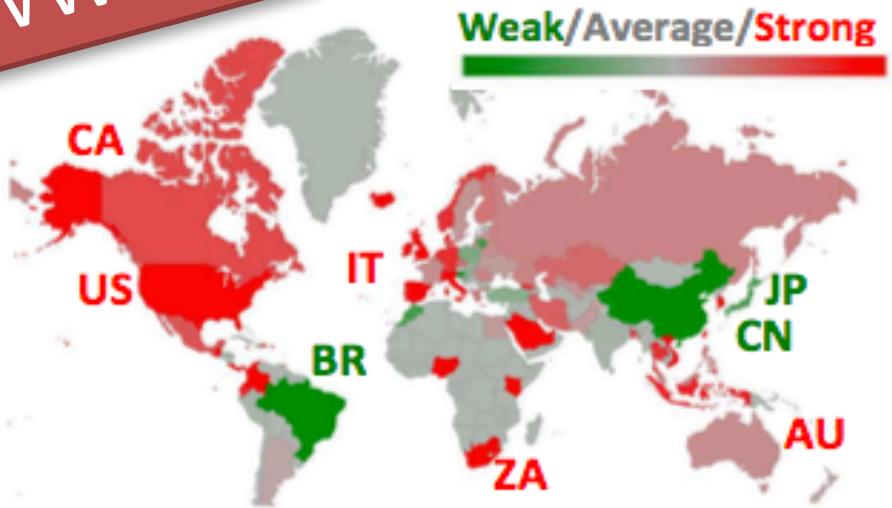
Compact desc ion of $\mathcal{X}$, "*omatically*"



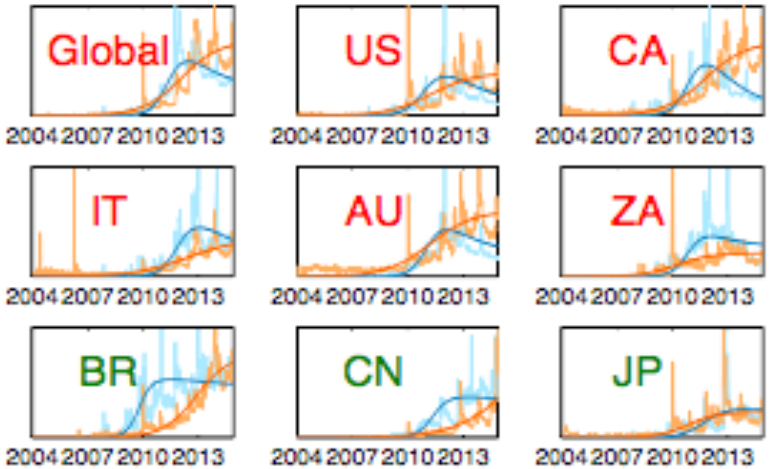
New challenge: MANT analysis

CompCube WWW'16

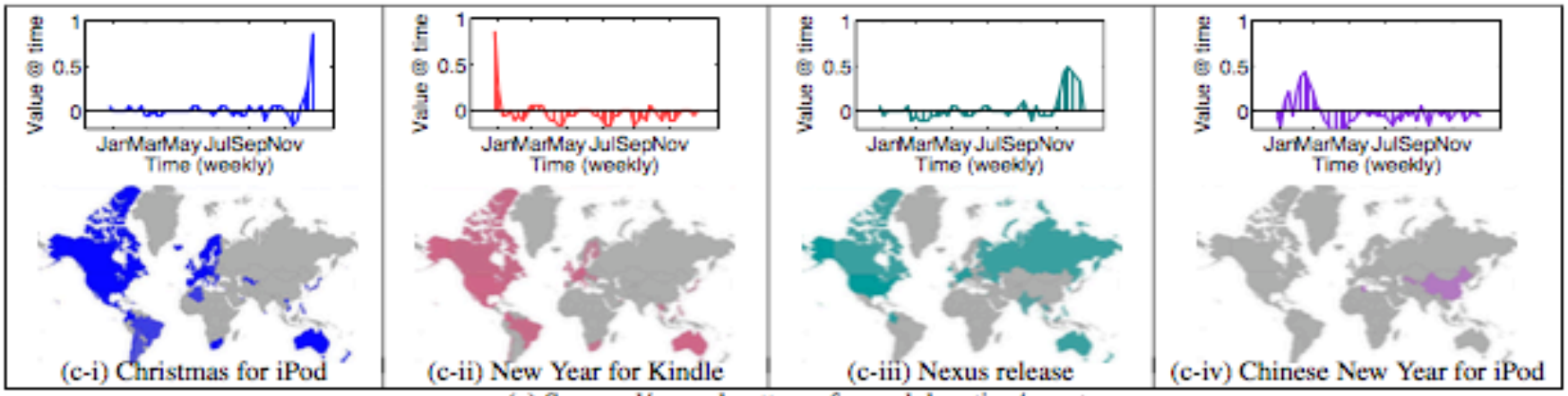
Non-Linear mining & forecasting of local competition



(a) Local competition strength between Kindle and Nexus



(b) Original/fitted search volumes for Kindle and Nexus



(c-i) Christmas for iPod (c-ii) New Year for Kindle (c-iii) Nexus release (c-iv) Chinese New Year for iPod

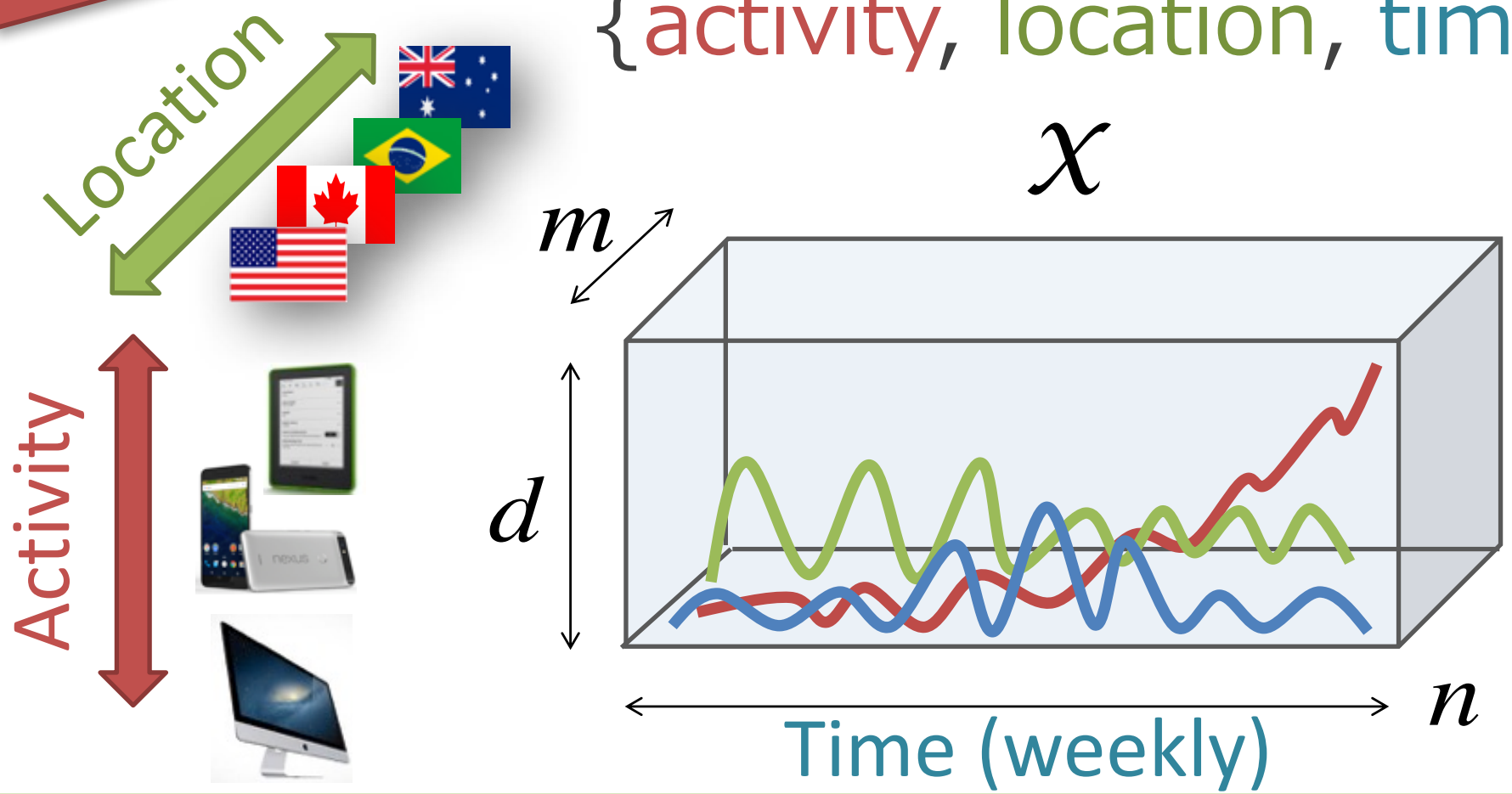
(c) Seasonal/annual patterns for each location/country

New challenge: MANT analysis

CompCube
WWW'16

Time-stamped events:

{activity, location, time}

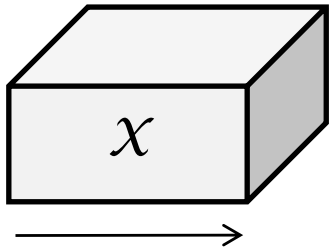


New challenge: MANT analysis

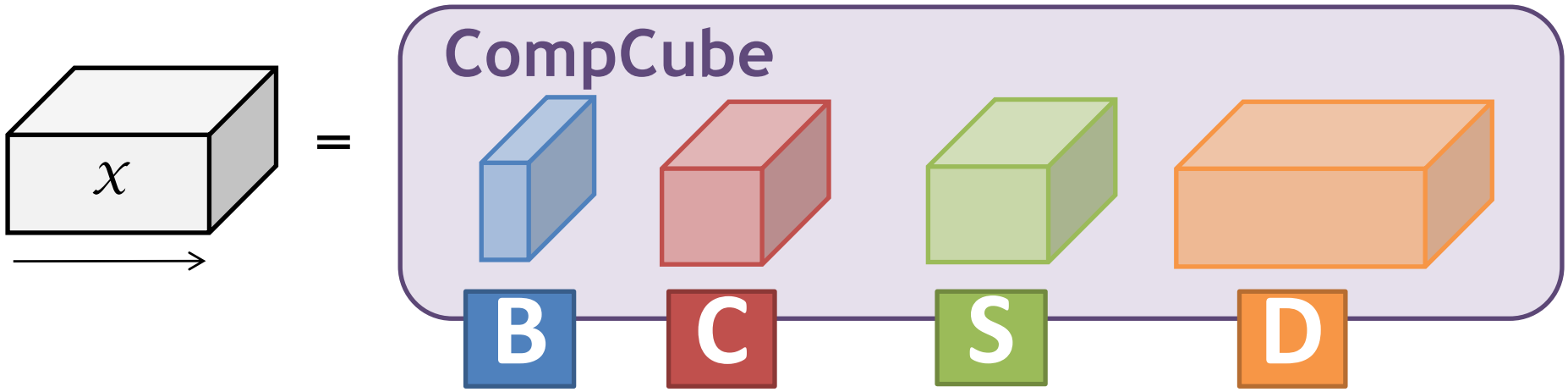
CompCube
WWW'16

Given: Tensor $\mathcal{X}$

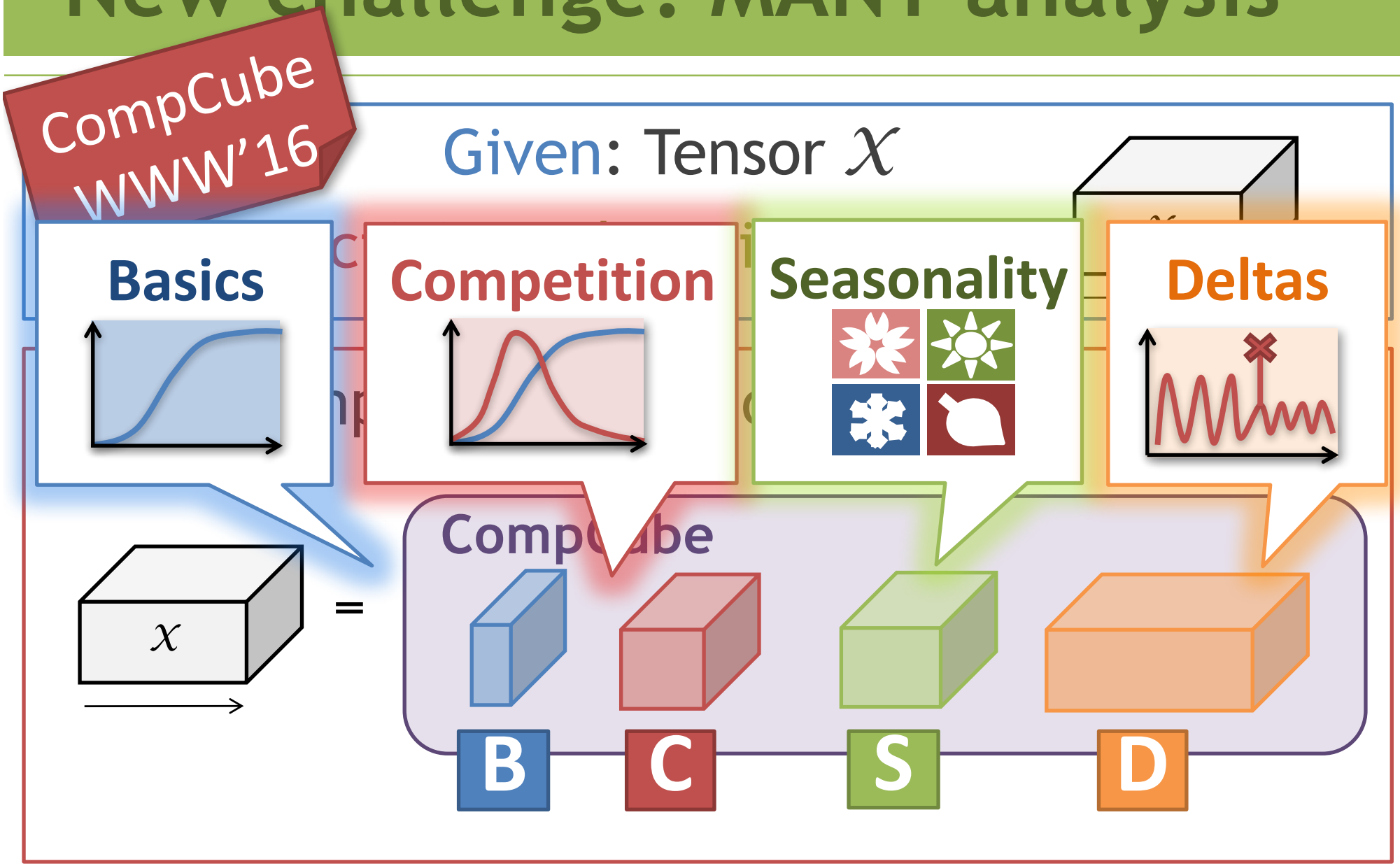
(activity x location x time)



Find: Compact description of $\mathcal{X}$



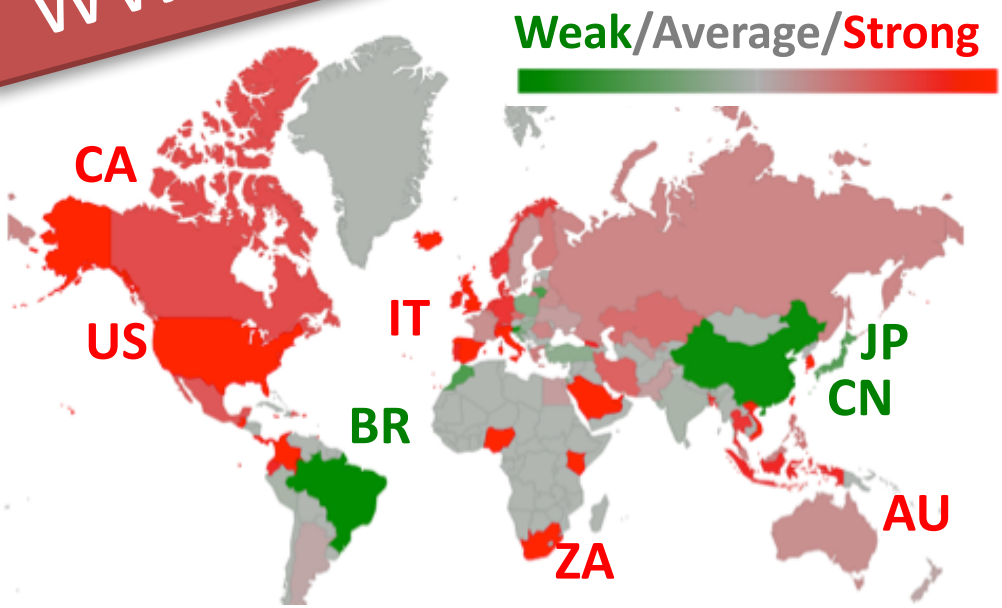
New challenge: MANT analysis

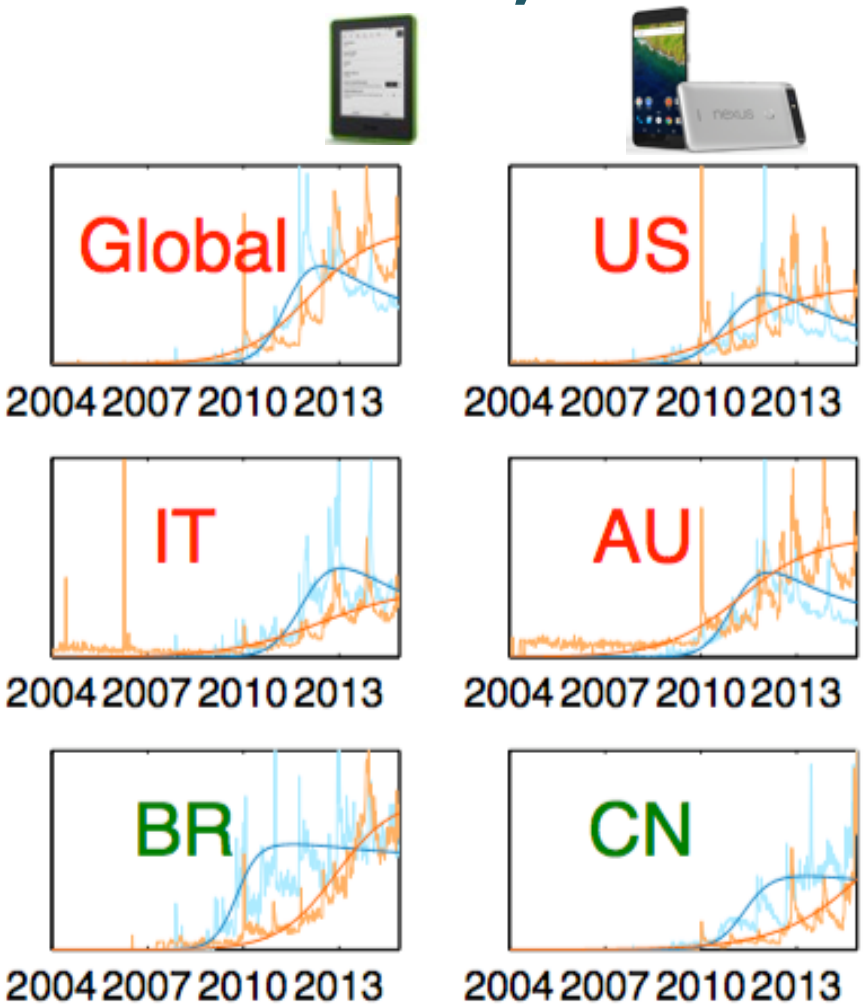


Modeling power of CompCube

CompCube WWW'16

Google search for **Kindle, Nexus**



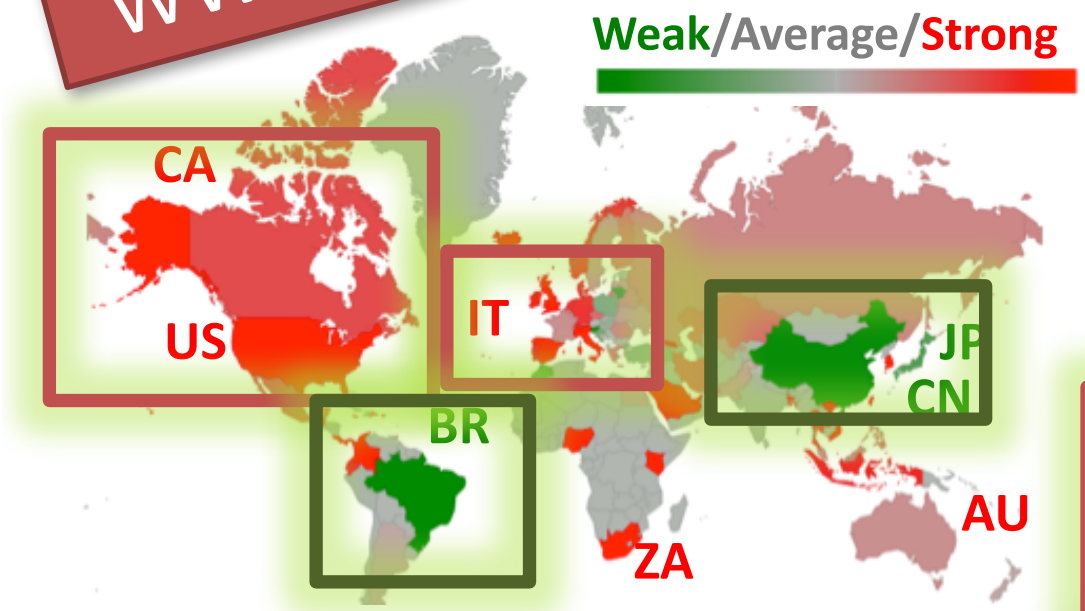


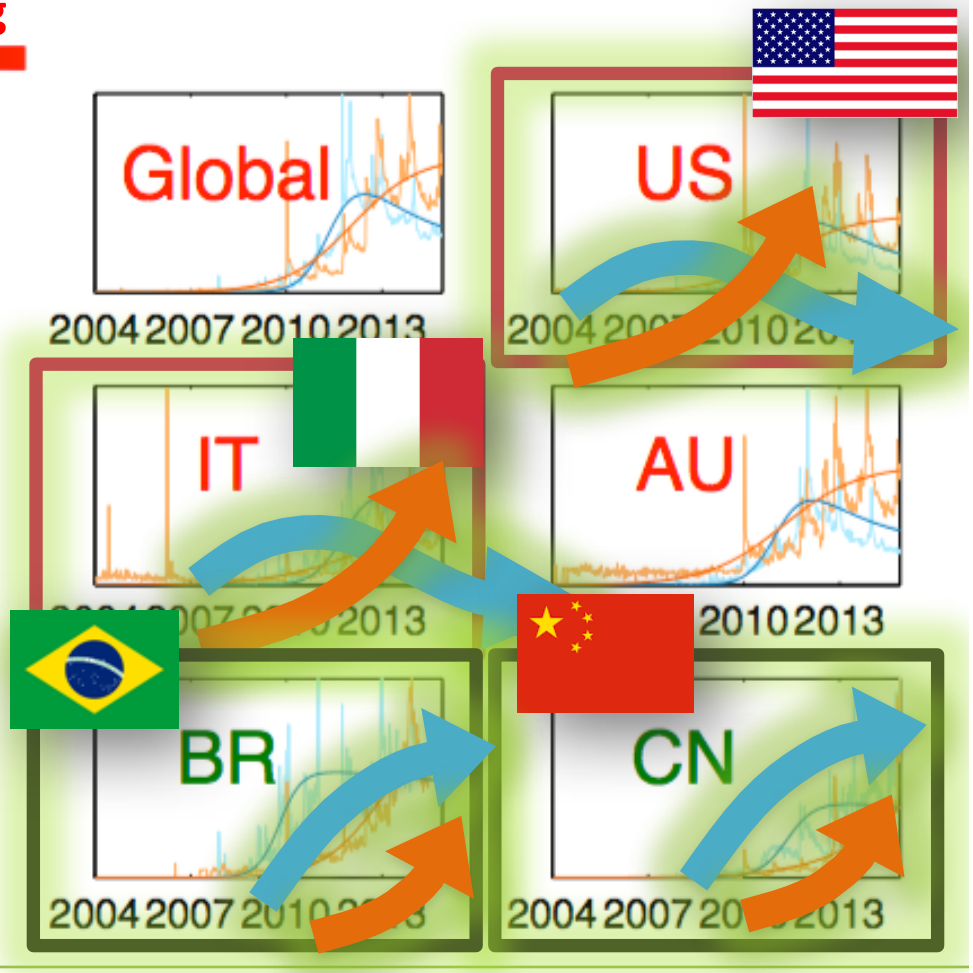
Local Competition strength

Modeling power of CompCube

CompCube WWW'16

Google search for **Kindle, Nexus**



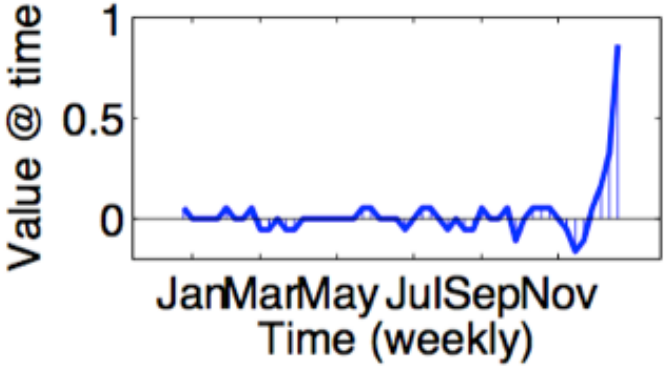


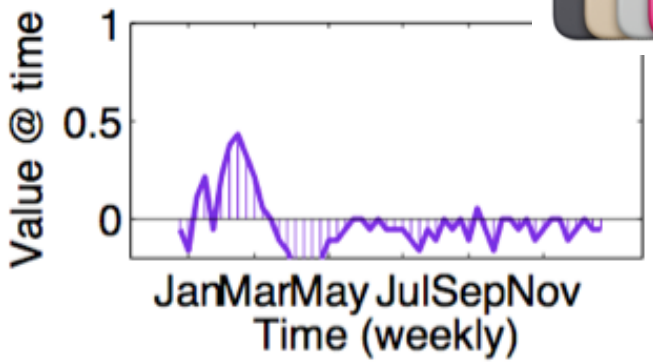
Local Competition strength

Modeling power of CompCube

CompCube WWW'16

Local seasonality for **iPod**





Component #1

Component #2

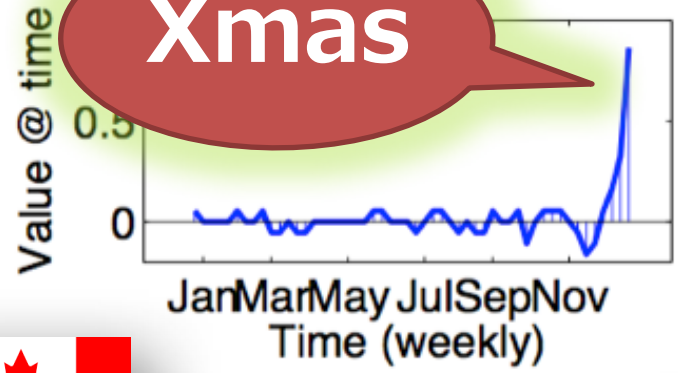
Modeling power of CompCube

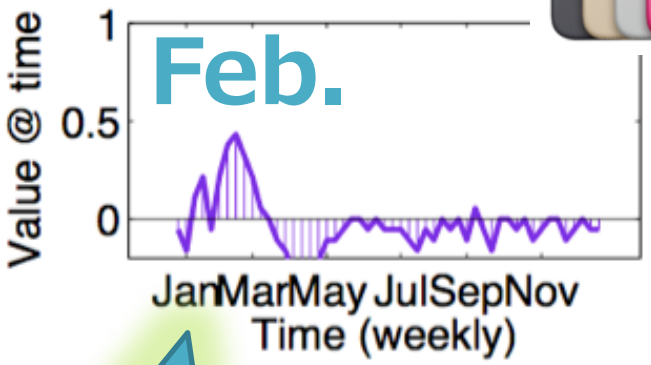
CompCube WWW'16

Local seasonality for **iPod**

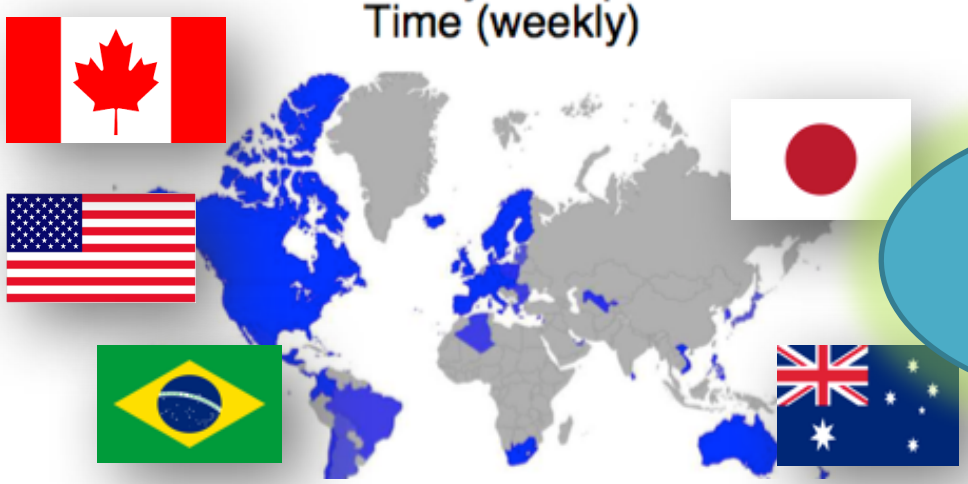
Dec.

Xmas





Feb.



Chinese New Year

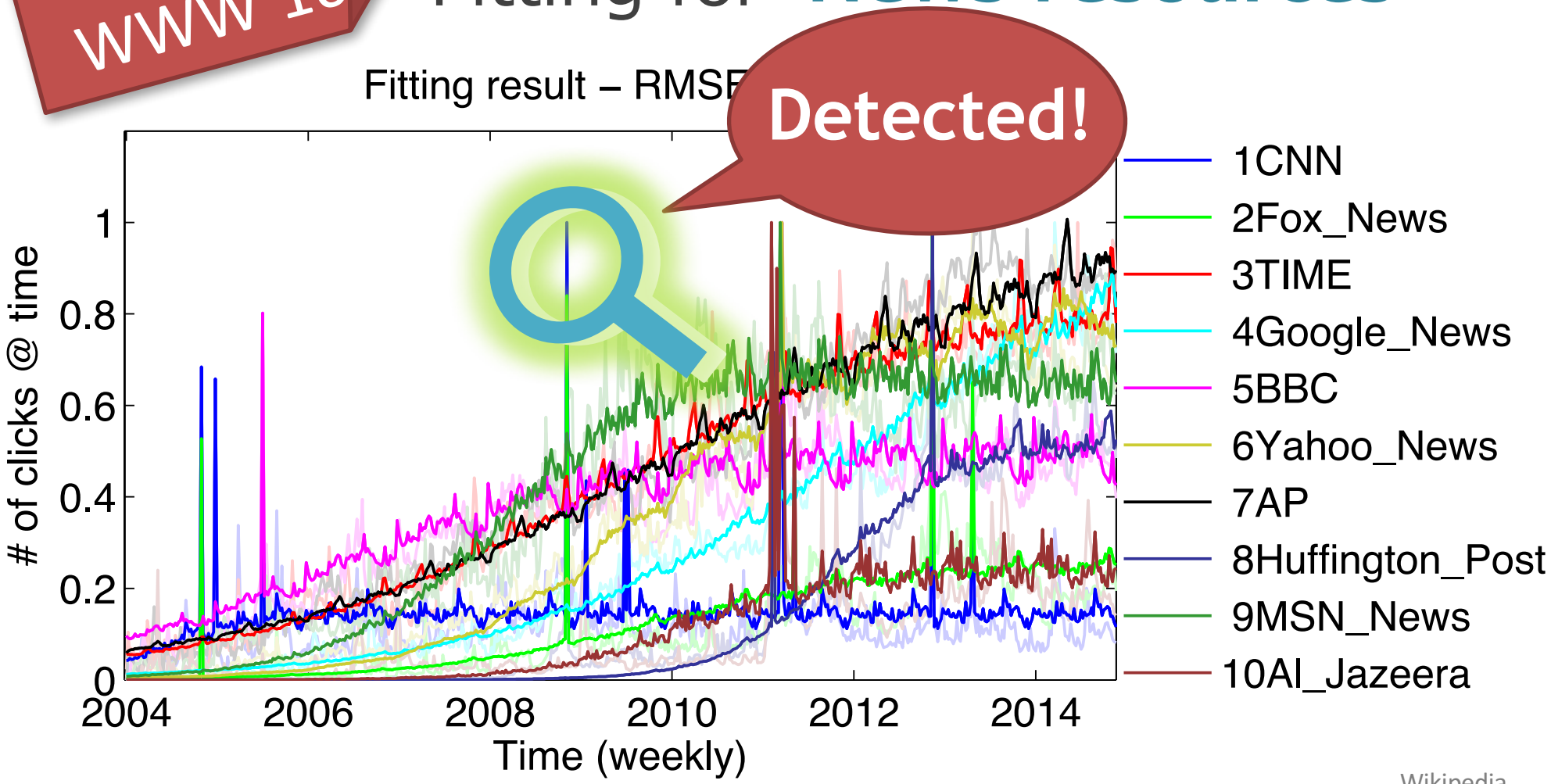
Component #1

Component #2

Modeling power of CompCube

CompCube WWW'16

Fitting for **News resources**



Wikipedia

Modeling power of CompCube

CompCube WWW'16

Fitting for **News resources**

Weak/**Strong**

Local attention to US election

US election Nov. 2008

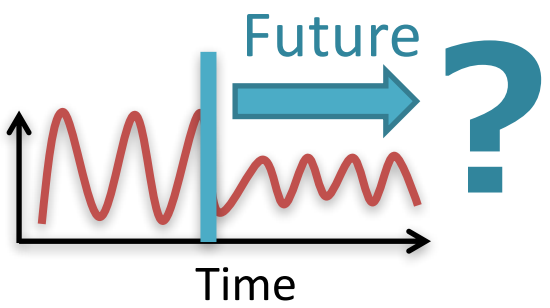
Wikipedia

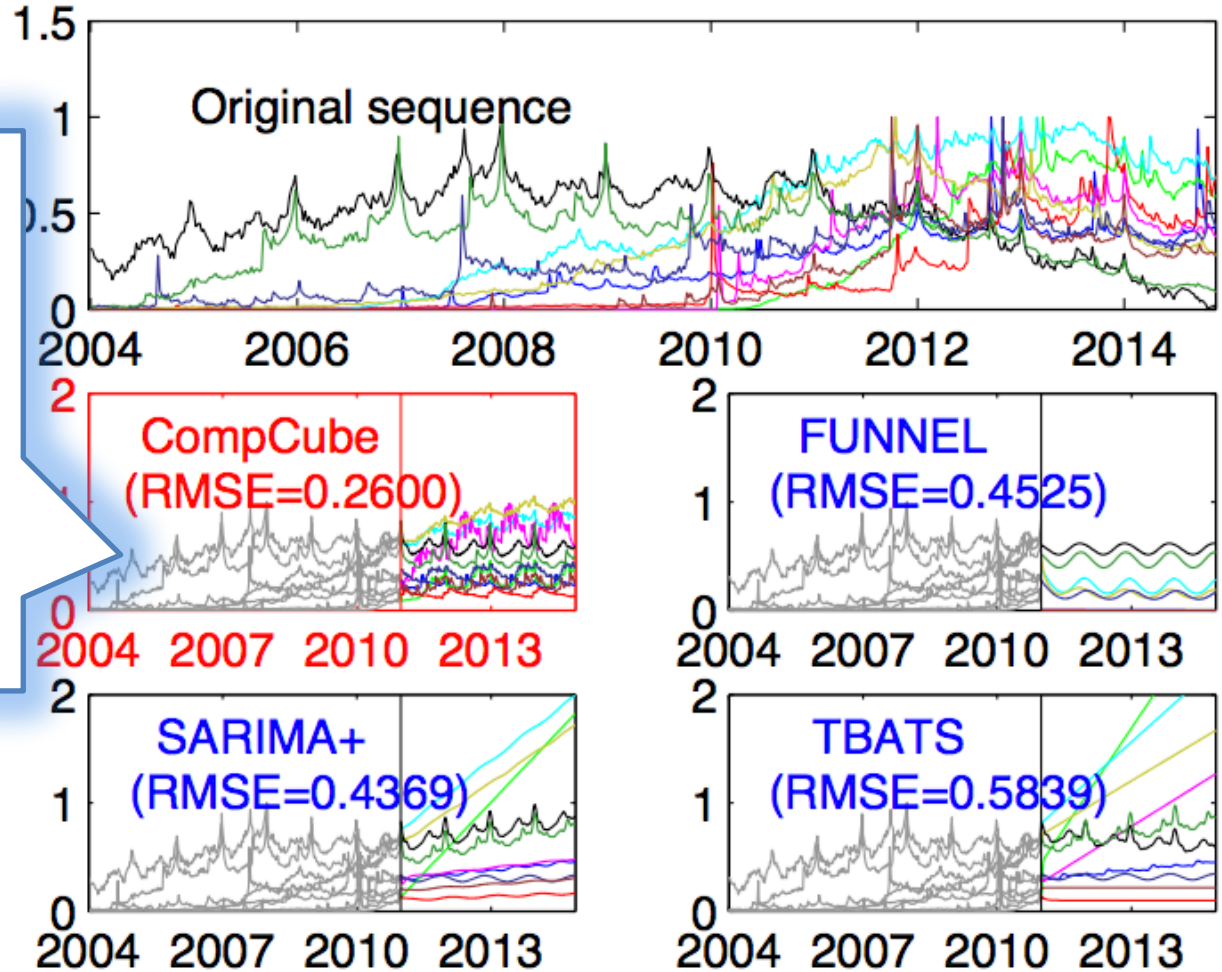
New challenge: MANT analysis

CompCube WWW'16

Non-Linear mining & forecasting of local competition

3. Forecast future activities





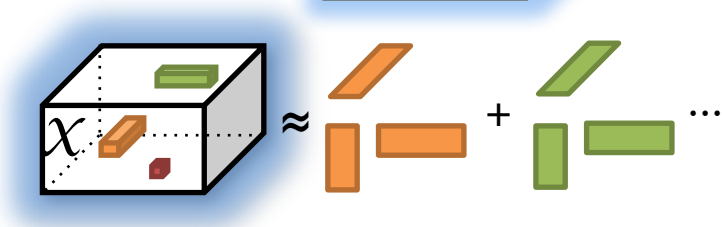
Putting it all together

3h-Tutorials
SIGMOD'15
WWW'16

New research directions

1. Automatic mining
2. Non-linear (gray-box) modeling
3. Large-scale tensor analysis

Put all together



M A N T

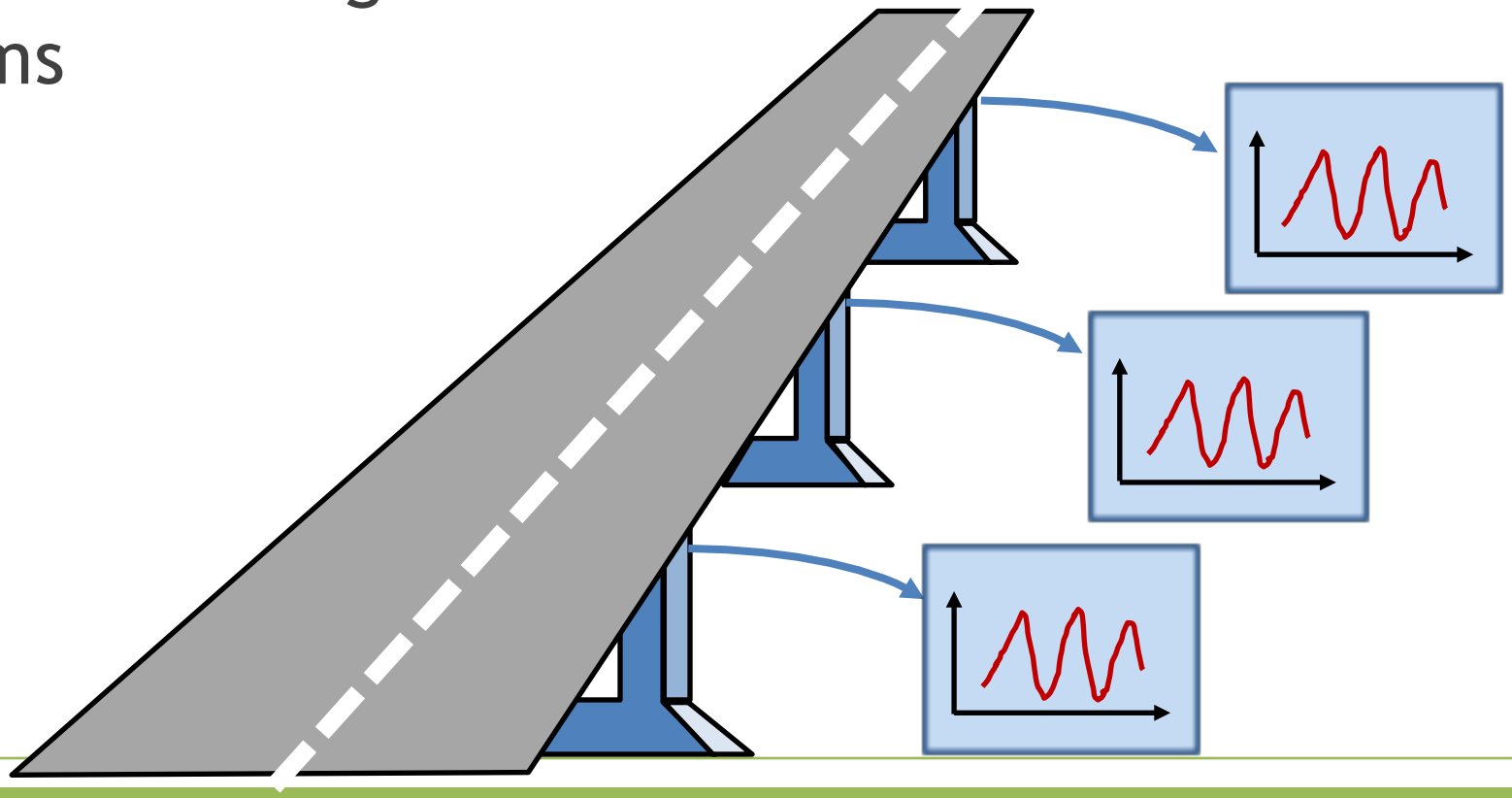
New challenge: MANT analysis

Multi-Aspect Non-linear Time-series

Industrial contribution

BRAID@ SIGMOD'05

- Bridge structural health monitoring
  - Structural monitoring using vibration/shock sensors
  - Keep track of lag correlations for sensor data streams



Industrial contribution

BRAID@ SIGMOD'05

- Bridge structural health monitoring
  - Goal: real-time anomaly detection for disaster prevention
  - Several thousands readings (per sec) from several hundreds sensor nodes

Structural health monitoring

Vibration/shock sensor

- Uses BRAID
- Metropolitan Expressway (Tokyo, Japan)

Industrial contribution

BRAID@ SIGMOD'05

- Bridge structural health monitoring with BRAID

Metropolitan Expressway (Tokyo, Japan)

Tokyo Gate Bridge (Tokyo, Japan)

Can Tho Bridge (Vietnam)

Industrial contribution

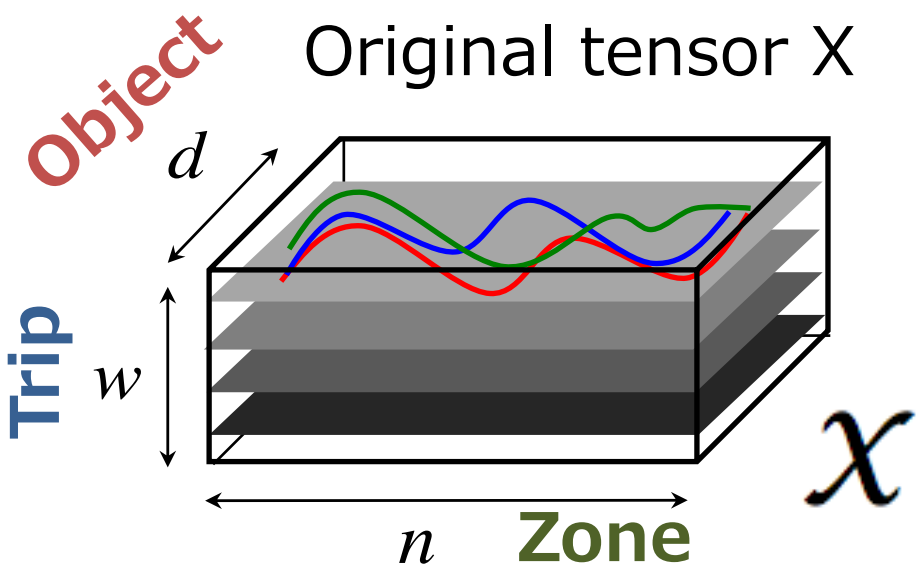
WebDBF'15 Best paper

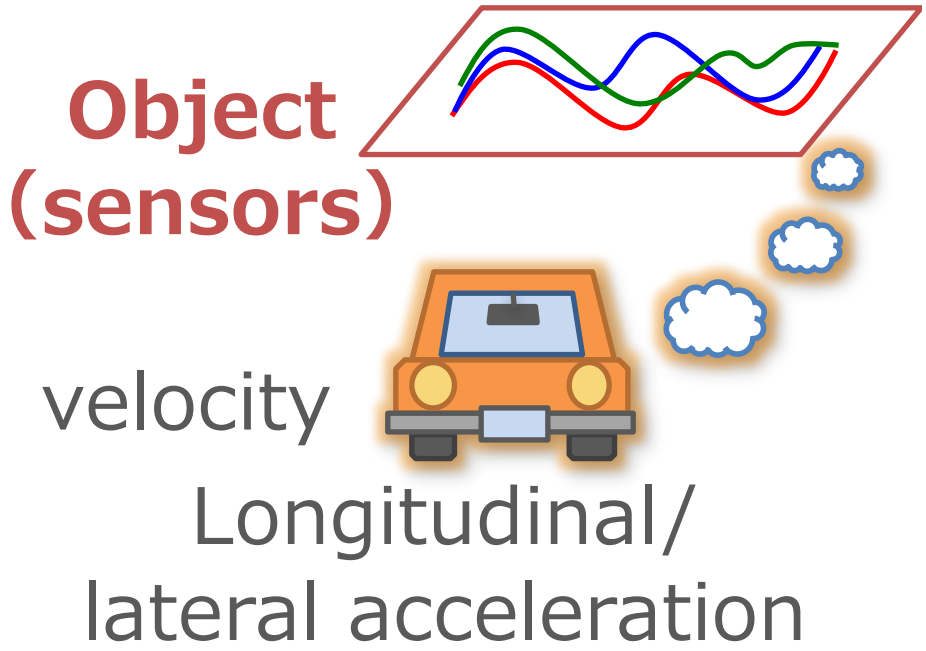
– **Automobile sensor data**

{ **trip ID**, **zone ID**, o

TOYOTA InfoTechnology Center Co., Ltd.

TrailMarker





Industrial contribution

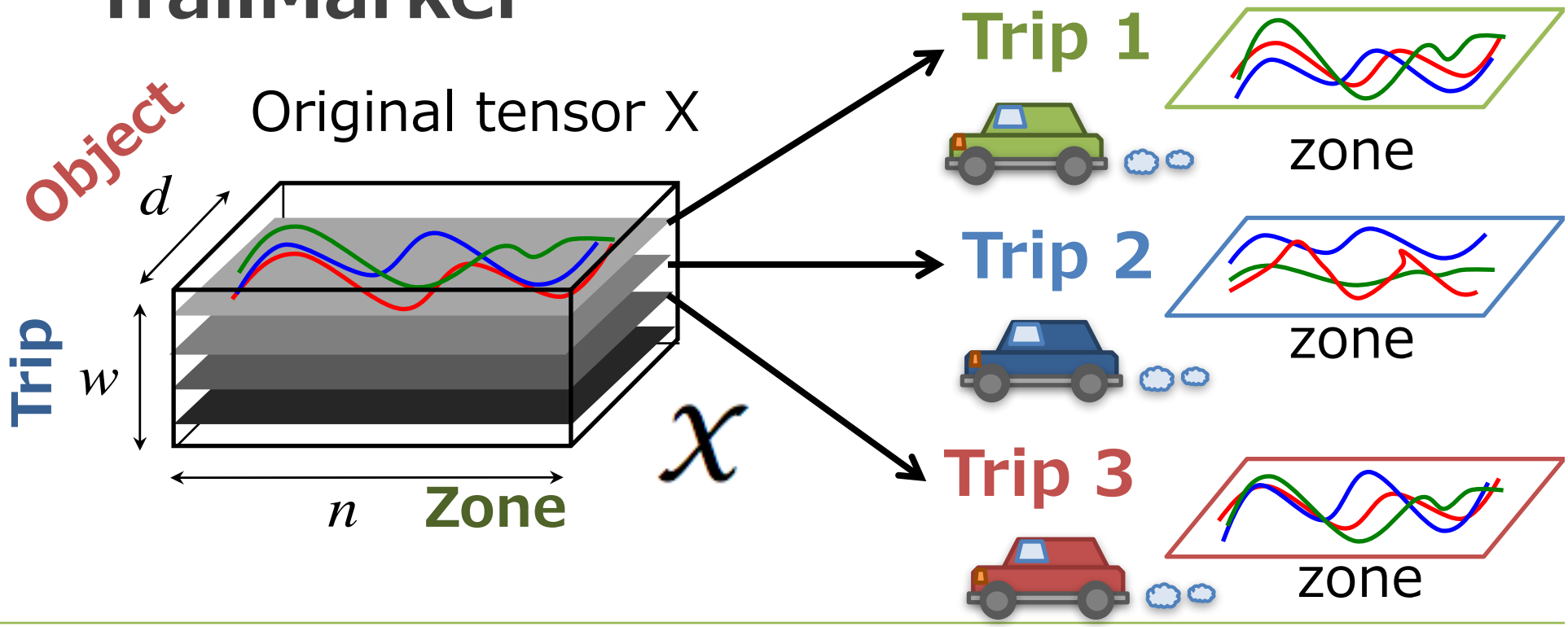
WebDBF'15 Best paper

– **Automobile sensor data**

{ **trip ID**, **zone ID**, o

TOYOTA InfoTechnology Center Co., Ltd.

TrailMarker



Industrial contribution

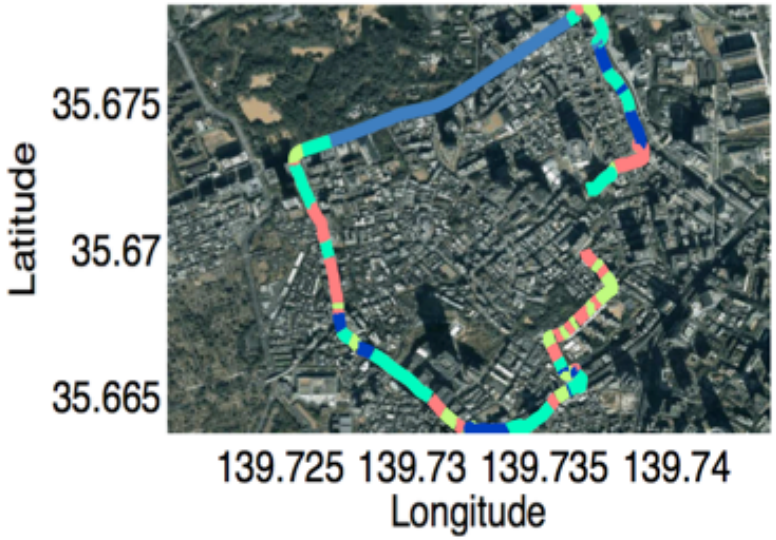
WebDBF'15 Best paper

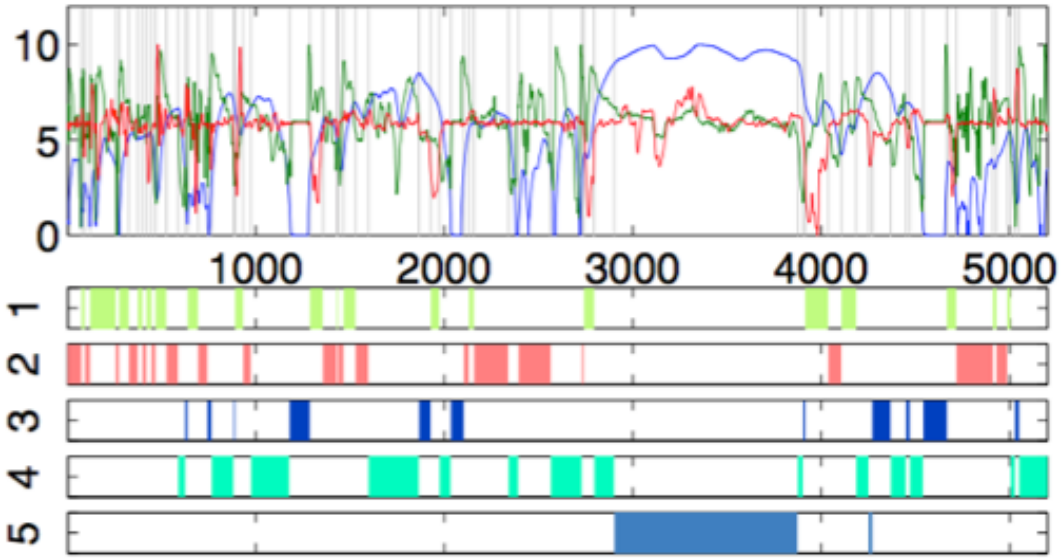
– **Automobile sensor data**

{ **trip ID**, **zone ID**, **o**

TOYOTA InfoTechnology Center Co., Ltd.

TrailMarker





Thank you

Code/Data/Tutorial:

Sakurai Lab.

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

M A N T