Mining Big Time-series Data on the Web

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

Roadmap

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

Part 3

Extension of time-series: tensor analysis

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

Outline

- Tensor decomposition
- Mining and forecasting of complex time-stamped events
- New challenge: MANT analysis
  Multi-Aspect Non-linear Time-series

M A N T

Outline

- Tensor decomposition
  - Motivation
  - Basic approaches
- Mining and forecasting of complex time-stamped events
- New challenge: MANT analysis
  Multi-Aspect Non-linear Time-series

M A N T

Examples of Matrices: Graph - social network

	John	Peter	Mary	Nick	...
John	0	11	22	55	...
Peter	5	0	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

Examples of Matrices: cloud of n-dim points

	chol#	blood#	age	..	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

Examples of Matrices: Market basket

- **market basket** as in Association Rules

	milk	bread	choc.	wine	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

Examples of Matrices: Documents and terms

	data	mining	classif.	tree	...
Paper#1	13	11	22	55	...
Paper#2	5	4	6	7	...
Paper#3	...	...	...	...	...
Paper#4	...	...	...	...	...
...	...	...	...	...	...

Examples of Matrices: Authors and terms

	data	mining	classif.	tree	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

Examples of Matrices: sensor-ids and time-ticks

	temp1	temp2	humid.	pressure	...
t=1	13	11	22	55	...
t=2	5	4	6	7	...
t=3	...	...	...	...	...
t=4	...	...	...	...	...
...	...	...	...	...	...

Motivation 2: Why tensors?

- Q: what is a tensor?
- A: N-D generalization of matrix:

www' 16	network	search	graph	mining	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

Motivation 2: Why tensors?

- Q: what is a tensor?
- A: N-D generalization of matrix:

www' 14
www' 15
www' 16

	network	search	graph	mining	...
John	13	11	22	55	...
Peter	5	4	6	7	...
Mary	...	...	...	...	...
Nick	...	...	...	...	...
...	...	...	...	...	...

Tensors are useful for 3 or more modes

Terminology: 'mode' (or 'aspect'):

Mode#3

Mode#2

	network	search	graph	mining	...
	13	11	22	55	...
	5	4	6	7	...
	...	...	...	...	...
	...	...	...	...	...
	...	...	...	...	...

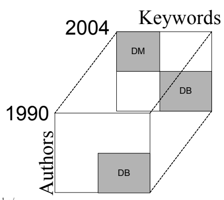
Mode (== aspect) #1

Motivating Applications

- Why tensors are useful?
  - P1: social networks
  - P2: web mining

P1: Social network analysis

- Monitoring networks and community structures over time



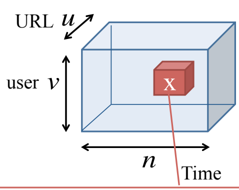
P2: Web graph mining

- How to order the importance of web pages?
  - Kleinberg's algorithm HITS
  - PageRank
  - Tensor extension on HITS (TOPHITS)
    - context-sensitive hypergraph analysis

Tensor analysis for time-series data

- 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
M^{th} order tensor (M=3)

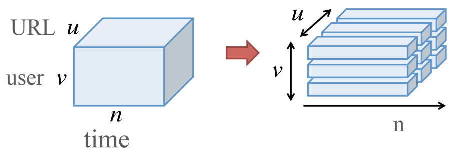
$\mathcal{X} \in \mathbb{N}^{u \times v \times n}$

Element x: # of events

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

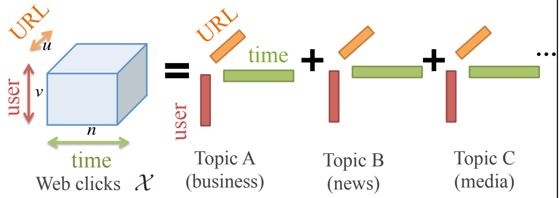
Tensor analysis for time-series data

- Individual-sequence mining
  - Create a set of (u * v) sequences of length (n)
  - Apply the mining algorithm for each sequence



Tensor analysis for time-series data

- Multi-aspect time-series analysis



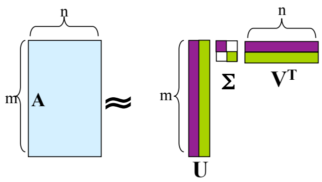
Outline

- Tensor decomposition
  - Motivation
  - Basic approaches
- Mining and forecasting of complex time-stamped events
- New challenge: MANT analysis

  Multi-Aspect Non-linear Time-series

Reminder: SVD

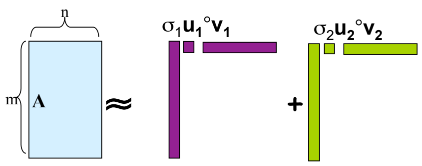
$$\mathbf{A} \approx \mathbf{U}\Sigma\mathbf{V}^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i$$



- Best rank-k approximation in L2

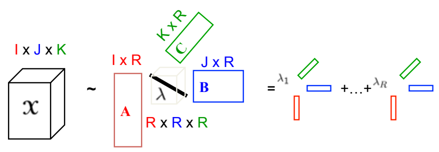
Reminder: SVD

$$\mathbf{A} \approx \mathbf{U}\Sigma\mathbf{V}^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i$$



- Best rank-k approximation in L2

Goal: extension to >=3 modes



$$\mathcal{X} \approx [\![\lambda\,;\mathbf{A},\mathbf{B},\mathbf{C}]\!] = \sum_r \lambda_r\, \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$

Kumamoto U | CMU CS

Main points:

- 2 major types of tensor decompositions: PARAFAC and Tucker
- both can be solved with ``alternating least squares'' (ALS)

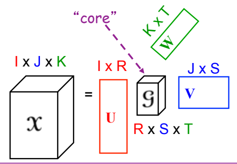
Kumamoto U | CMU CS

Specially Structured Tensors

- Tucker Tensor

$$\mathcal{X} = \mathcal{G} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W} = \sum_r \sum_s \sum_t g_{rst}\, \mathbf{u}_r \circ \mathbf{v}_s \circ \mathbf{w}_t \equiv [\![\mathcal{G} \,; \mathbf{U}, \mathbf{V}, \mathbf{W}]\!]$$

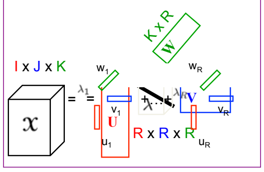
Our Notation



- Kruskal Tensor

$$\mathcal{X} = \sum_r \lambda_r\, \mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r \equiv [\![\lambda \,; \mathbf{U}, \mathbf{V}, \mathbf{W}]\!]$$

Our Notation



Details | CMU CS

Specially Structured Tensors

- Tucker Tensor

$$\mathcal{X} = \mathcal{G} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W} = \sum_r \sum_s \sum_t g_{rst}\, \mathbf{u}_r \circ \mathbf{v}_s \circ \mathbf{w}_t \equiv [\![\mathcal{G} \,; \mathbf{U}, \mathbf{V}, \mathbf{W}]\!]$$

In matrix form:

$$\mathbf{X}_{(1)} = \mathbf{U}\mathbf{G}_{(1)}(\mathbf{W} \otimes \mathbf{V})^{\mathsf{T}}$$
$$\mathbf{X}_{(2)} = \mathbf{V}\mathbf{G}_{(2)}(\mathbf{W} \otimes \mathbf{U})^{\mathsf{T}}$$
$$\mathbf{X}_{(3)} = \mathbf{W}\mathbf{G}_{(3)}(\mathbf{V} \otimes \mathbf{U})^{\mathsf{T}}$$

$$\mathrm{vec}(\mathcal{X}) = (\mathbf{W} \otimes \mathbf{V} \otimes \mathbf{U})\mathrm{vec}(\mathcal{G})$$

- Kruskal Tensor

$$\mathcal{X} = \sum_r \lambda_r\, \mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r \equiv [\![\lambda \,; \mathbf{U}, \mathbf{V}, \mathbf{W}]\!]$$

In matrix form:

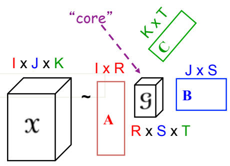
Let $\mathbf{\Lambda} = \mathrm{diag}(\lambda)$

$$\mathbf{X}_{(1)} = \mathbf{U}\mathbf{\Lambda}(\mathbf{W} \odot \mathbf{V})^{\mathsf{T}}$$
$$\mathbf{X}_{(2)} = \mathbf{V}\mathbf{\Lambda}(\mathbf{W} \odot \mathbf{U})^{\mathsf{T}}$$
$$\mathbf{X}_{(3)} = \mathbf{W}\mathbf{\Lambda}(\mathbf{V} \odot \mathbf{U})^{\mathsf{T}}$$

$$\mathrm{vec}(\mathcal{X}) = (\mathbf{W} \odot \mathbf{V} \odot \mathbf{U})\,\lambda$$

Kumamoto U | CMU CS

Tucker Decomposition - intuition



- author x keyword x conference
- A: author x author-group
- B: keyword x keyword-group
- C: conf. x conf-group
- $\mathcal{G}$: how groups relate to each other

Kumamoto U | CMU CS

Intuition behind core tensor

- 2-d case: co-clustering
- [Dhillon et al. Information-Theoretic Co-clustering, KDD' 03]

Kumamoto U | CMU CS

eg, terms x documents

$$\begin{array}{cc} & n \\ m & \begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix} \end{array}$$

$$\begin{array}{cc} & k \\ m & \begin{bmatrix} .5 & 0 & 0 \\ .5 & 0 & 0 \\ 0 & .5 & 0 \\ 0 & .5 & 0 \\ 0 & 0 & .5 \\ 0 & 0 & .5 \end{bmatrix} \end{array} \begin{array}{cc} & l \\ k & \begin{bmatrix} .3 & 0 \\ 0 & .3 \\ .2 & .2 \end{bmatrix} \end{array} \begin{array}{cc} & n \\ l & \begin{bmatrix} .36 & .36 & .28 & 0 & 0 & 0 \\ 0 & 0 & 0 & .28 & .36 & .36 \end{bmatrix} \end{array} = \begin{bmatrix} .054 & .054 & .042 & 0 & 0 & 0 \\ .054 & .054 & .042 & 0 & 0 & 0 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ .036 & .036 & 028 & .028 & .036 & .036 \\ .036 & .036 & .028 & .028 & .036 & .036 \end{bmatrix}$$

med. doc cs doc

med. terms | cs terms | common terms

$$\begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix}$$

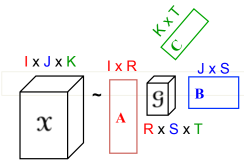
term group x doc. group

$$\begin{bmatrix} .5 & 0 & 0 \\ .5 & 0 & 0 \\ 0 & .5 & 0 \\ 0 & .5 & 0 \\ 0 & 0 & .5 \\ 0 & 0 & .5 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 3 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} .36 & .36 & .28 & 0 & 0 & 0 \\ 0 & 0 & 0 & .28 & .36 & .36 \end{bmatrix} = \begin{bmatrix} .054 & .054 & .042 & 0 & 0 & 0 \\ .054 & .054 & .042 & 0 & 0 & 0 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ .036 & .036 & .028 & .028 & .036 & .036 \\ .036 & .036 & .028 & .028 & .036 & .036 \end{bmatrix}$$

term x term-group

doc x doc group

Tucker Decomposition



$$\mathcal{X} \approx [\![\mathcal{G} \,;\, \mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

Given A, B, C, the optimal core is:

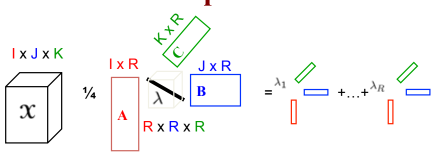
$$\mathcal{G} = [\![\mathcal{X} \,;\, \mathbf{A}^{\dagger}, \mathbf{B}^{\dagger}, \mathbf{C}^{\dagger}]\!]$$

- Proposed by Tucker (1966)
- AKA: Three-mode factor analysis, three-mode PCA, orthogonal array decomposition
- **A**, **B**, and **C** generally assumed to be orthonormal (generally assume they have full column rank)
- $\mathcal{G}$ is not diagonal
- Not unique

Recall the equations for converting a tensor to a matrix

$$\mathbf{X}_{(1)} = \mathbf{A}\mathbf{G}_{(1)}(\mathbf{C} \otimes \mathbf{B})^{\mathsf{T}}$$
$$\mathbf{X}_{(2)} = \mathbf{B}\mathbf{G}_{(2)}(\mathbf{C} \otimes \mathbf{A})^{\mathsf{T}}$$
$$\mathbf{X}_{(3)} = \mathbf{C}\mathbf{G}_{(3)}(\mathbf{B} \otimes \mathbf{A})^{\mathsf{T}}$$
$$\mathrm{vec}(\mathcal{X}) = (\mathbf{C} \otimes \mathbf{B} \otimes \mathbf{A})\mathrm{vec}(\mathcal{G})$$

CANDECOMP/PARAFAC Decomposition

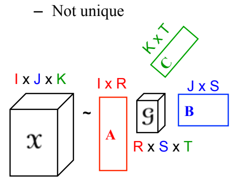


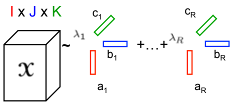
$$\mathcal{X} \approx [\![\lambda \,;\, \mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = \sum_r \lambda_r \; \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$

- CANDECOMP = Canonical Decomposition (Carroll & Chang, 1970)
- PARAFAC = Parallel Factors (Harshman, 1970)
- Core is diagonal (specified by the vector λ)
- Columns of **A**, **B**, and **C** are not orthonormal
- If R is *minimal*, then R is called the **rank** of the tensor (Kruskal 1977)
- Can have rank( ) > min{I,J,K}

Tucker vs. PARAFAC Decompositions

- Tucker
  - Variable transformation in each mode
  - Core G may be dense
  - A, B, C generally orthonormal
  - Not unique
- PARAFAC
  - Sum of rank-1 components
  - No core, i.e., superdiagonal core
  - A, B, C may have linearly dependent columns
  - Generally unique





Tensor tools - summary

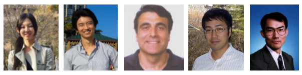
- Two main tools
  - PARAFAC
  - Tucker
- Both find row-, column-, tube-groups
  - but in PARAFAC the three groups are identical
- To solve: Alternating Least Squares

- Toolbox: from Tamara Kolda:
http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox/

Outline

- Tensor decomposition
- Mining and forecasting of complex time-stamped events
- New challenge: MANT analysis

Multi-Aspect Non-linear Time-series



[Matsubara+ KDD'12]

Fast Mining and Forecasting of Complex Time-Stamped Events

Yasuko Matsubara (Kyoto University)
Yasushi Sakurai (NTT)
Christos Faloutsos (CMU)
Tomoharu Iwata (NTT)
Masatoshi Yoshikawa (Kyoto University)

NTT

Kumamoto U CMU CS

Motivation

- Complex time-stamped events
  {timestamp + multiple attributes}

e.g., web click events:
{timestamp, URL, user ID, access devices, http referrer, …}

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

Kumamoto U CMU CS

Motivation

Q1. Are there any topics ?

- news, tech, media, sports, etc...

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

e.g.,
CNN.com, CNET.com -> news topic
YouTube.com -> media topic

Kumamoto U CMU CS

Motivation

Q2. Can we group URLs/users accordingly?

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

e.g.,
CNN.com & CNET.com (related to news topic)
Smith & Johnson (related to news topic)

Kumamoto U CMU CS

Motivation

Q3. Can we forecast future events?

- How many clicks from 'Smith' tomorrow?
- How many clicks to 'CNN.com' over next 7 days?

Timestamp	URL	User	Device
2012-08-01-12:00	CNN.com	Smith	iphone
2012-08-02-15:00	YouTube.com	Brown	iphone
2012-08-02-19:00	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

future clicks?

Kumamoto U CMU CS

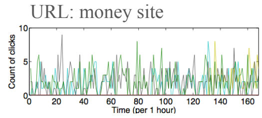
Motivation

Web click events – can we see any trends?

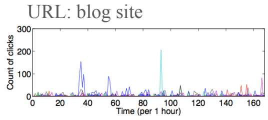
Original access counts of each URL

- 100 random users
- 1 week (window size = 1 hour)

URL: money site



URL: blog site

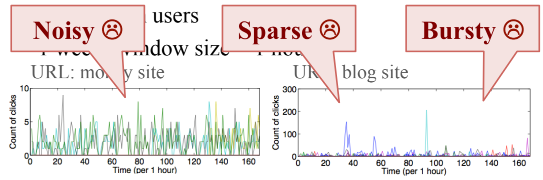


Motivation

Web click events – can we see any trends?

Original access counts of each URL

Noisy ☹ Sparse ☹ Bursty ☹



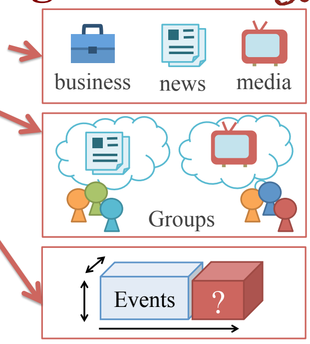
☹ We cannot see any trends !!

Our goals

Q1: Hidden topics

Q2: Groups

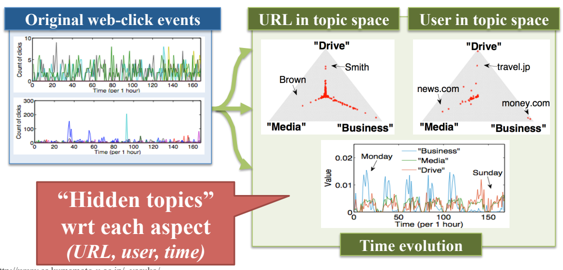
Q3: Forecasting



Problem definition

Given: a set of complex time-stamped events

1. **Find**: major topics/trends
2. **Forecast**: future events

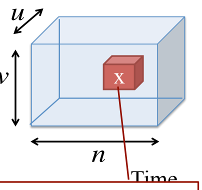


Main idea (1) : M-way analysis

Complex 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 M^{th} order tensor (M=3)

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

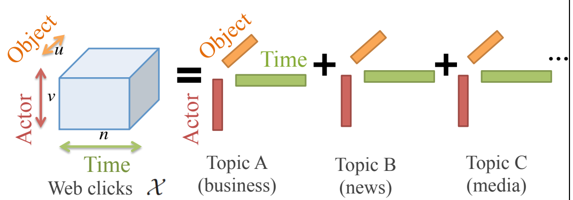
Element x: # of events

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

Main idea (1) : M-way analysis

A. decompose to a set of **3 topic vectors:**

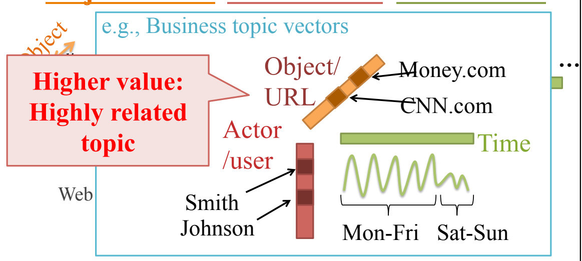
- Object vector Actor vector Time vector



Main idea (1) : M-way analysis

A. decompose to a set of **3 topic vectors:**

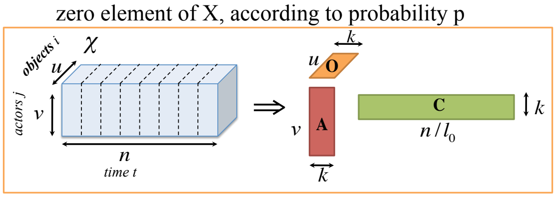
- Object vector Actor vector Time vector



Kumamoto U — CMU CS

Main idea (1) : M-way analysis (details)

- M-way decomposition (M=3)
  - [Gibbs sampling] infer k hidden topics for each non-zero element of X, according to probability p

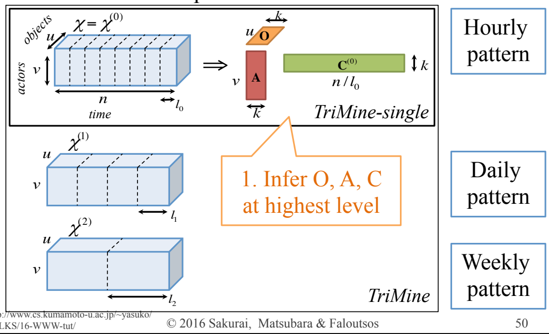


$$p(z_{i,j,t} = r | \mathcal{X}, \mathbf{O}', \mathbf{A}', \mathbf{C}', \alpha, \beta, \gamma) \propto \frac{o'_{i,r}+\alpha}{\sum_r o'_{i,r}+\alpha k} \cdot \frac{a'_{r,j}+\beta}{\sum_j a'_{r,j}+\beta v} \cdot \frac{c'_{r,t}+\gamma}{\sum_t c'_{r,t}+\gamma n} \quad (1)$$

Kumamoto U — CMU CS

Main idea (2) : Multi-scale analysis (details)

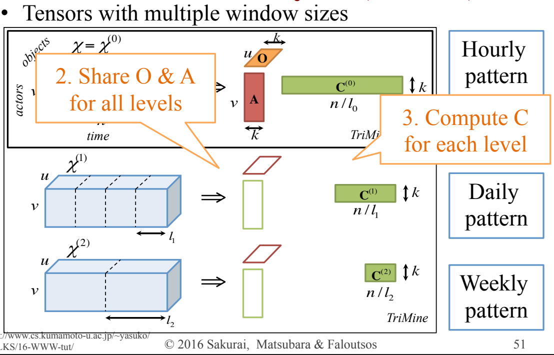
- Tensors with multiple window sizes



Kumamoto U — CMU CS

Main idea (2) : Multi-scale analysis (details)

- Tensors with multiple window sizes

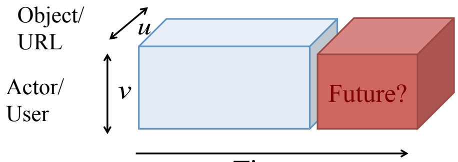


Kumamoto U — CMU CS

TriMine-Forecasts

Our final goal: "forecast future events"!

Q. How can we generate a realistic events?

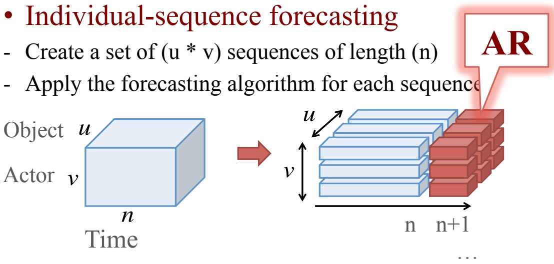


e.g., estimate the number of clicks for user "smith", to URL "CNN.com", for next 10 days

Kumamoto U — CMU CS

Why not naïve?

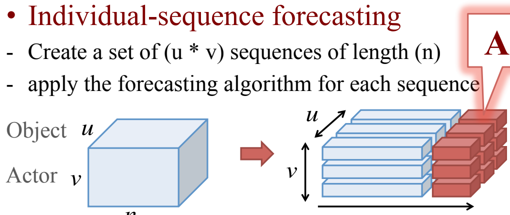
- Individual-sequence forecasting
  - Create a set of (u * v) sequences of length (n)
  - Apply the forecasting algorithm for each sequence



Kumamoto U — CMU CS

Why not naïve?

- Individual-sequence forecasting
  - Create a set of (u * v) sequences of length (n)
  - apply the forecasting algorithm for each sequence



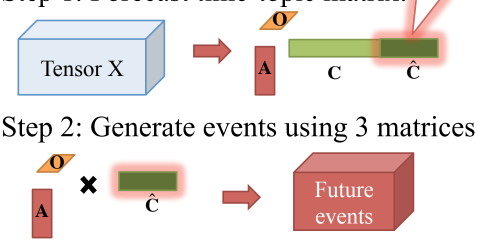
- ☹ **Scalability** : time complexity is at least $O(uvn)$
- ☹ **Accuracy** : each sequence "looks" like noise, (e.g., {0, 0, 0, 1, 0, 0, 2, 0, 0, ….}) -> hard to forecast

TriMine-F

Our approach:

– Step 1: Forecast time-topic matrix:

– Step 2: Generate events using 3 matrices



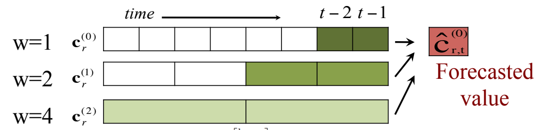
Forecast 'time-topic matrix' (details)

Q. How to capture multi-scale dynamics ?

e.g., bursty pattern, noise, multi-scale period

Multi-scale forecasting

Forecast $\hat{c}^{(0)}_{r,t}$ using multiple levels of matrices

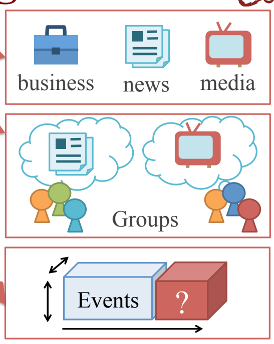


$$c^{(0)}_{r,t} = \sum_{h=0}^{\lceil \log n \rceil} \sum_{i=1}^{w} \lambda^{(h)}_{i,r} c^{(h)}_{r,t-i} + \epsilon_t.$$

(Details in paper)

Our goals

- Q1: Hidden topics
- Q2: Groups
- Q3: Forecasting

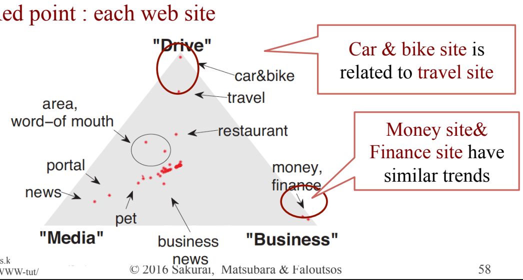


Q1&2. WebClick data

URL-topic matrix (O)

Three hidden topics: "drive", "business", "media"

* Red point : each web site

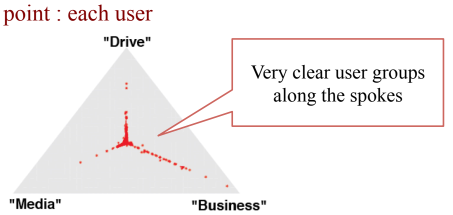


Q1&2. WebClick data

User-topic matrix (A)

Three hidden topics: "drive", "business", "media"

* Red point : each user

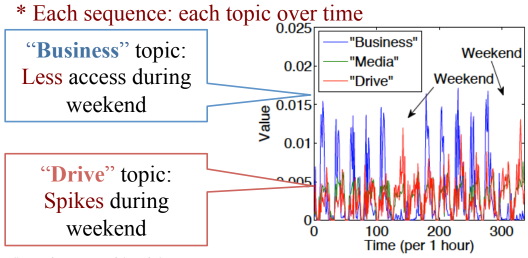


Q1&2. WebClick data

Time-topic matrix (C)

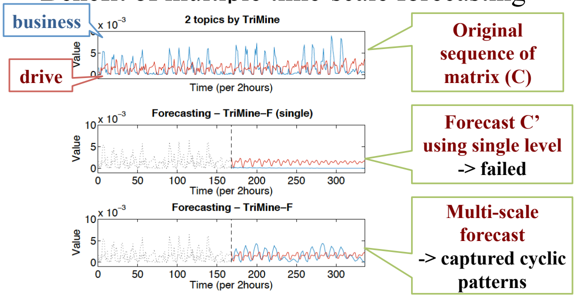
Three hidden topics: "drive", "business", "media"

* Each sequence: each topic over time



Q3. Forecasting accuracy

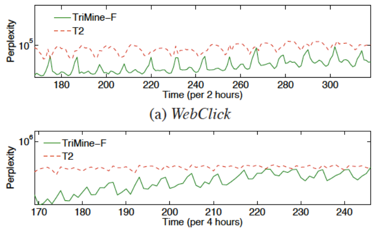
- Benefit of multiple time-scale forecasting



Q3. Forecasting accuracy

Temporal perplexity (entropy for each time-tick)

Lower perplexity: higher predictive accuracy



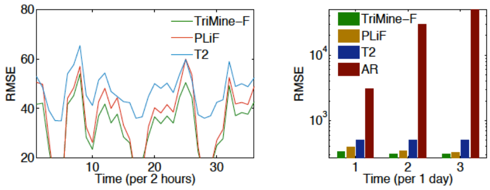
(a) *WebClick*

(b) *Ondemand TV* T2: [Hong et al. KDD'11]

Q3. Forecasting accuracy

Accuracy of event forecasting

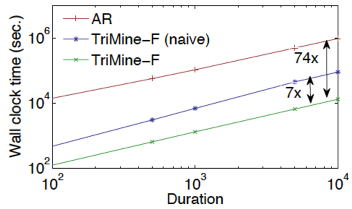
RMSE between original and forecasted events (lower is better)



PLiF [Li et al.VLDB'10] , T2: [Hong et al.KDD'11]

Q3. Scalability

- Computation cost (vs. AR)

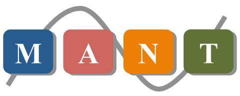


- TriMine provides a reduction in computation time (up to 74x)

Outline

- Tensor decomposition
- Mining and forecasting of complex time-stamped events
- New challenge: MANT analysis

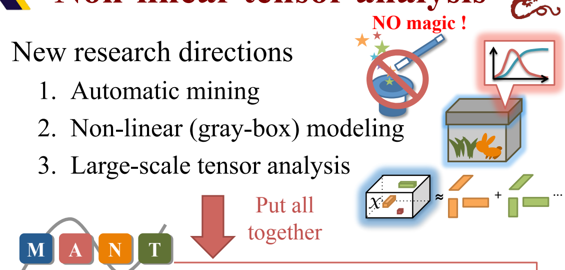
Multi-Aspect Non-linear Time-series



Non-linear tensor analysis

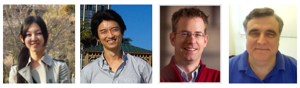
New research directions

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



New challenge: MANT analysis

Multi-Aspect Non-linear Time-series



[Matsubara+ KDD'14]

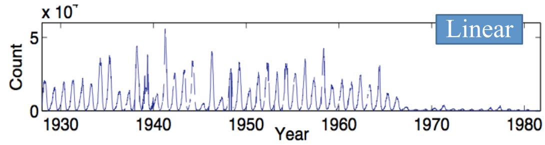
FUNNEL: Automatic Mining of Spatially Coevolving Epidemics

Yasuko Matsubara, Yasushi Sakurai (Kumamoto University)
Willem G. van Panhuis (University of Pittsburgh)
Christos Faloutsos (CMU)

Motivation

Given: Large set of epidemiological data

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

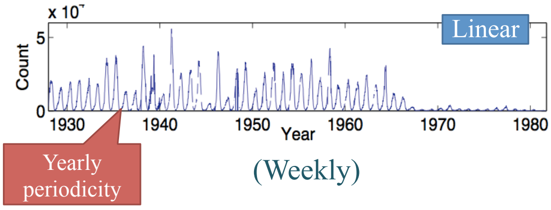


(Weekly)

Motivation

Given: Large set of epidemiological data

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

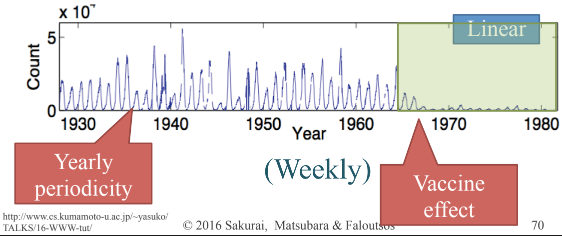


(Weekly)

Motivation

Given: Large set of epidemiological data

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

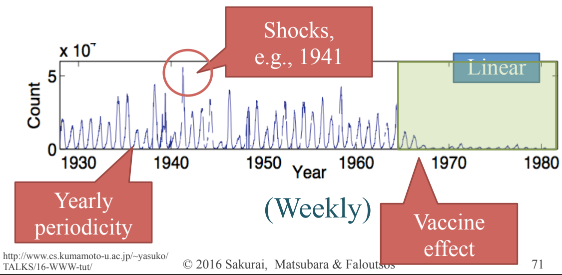


(Weekly)

Motivation

Given: Large set of epidemiological data

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



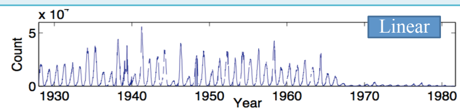
(Weekly)

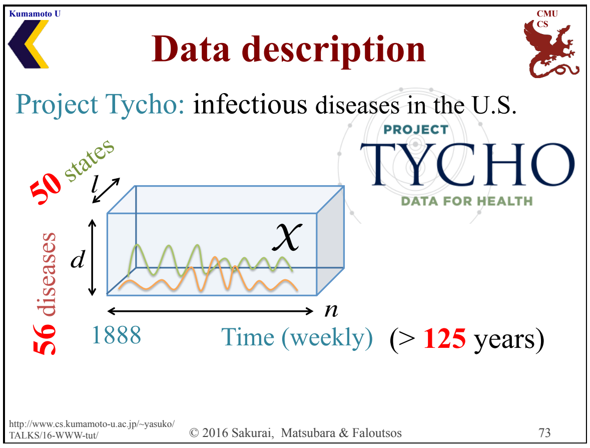
Motivation

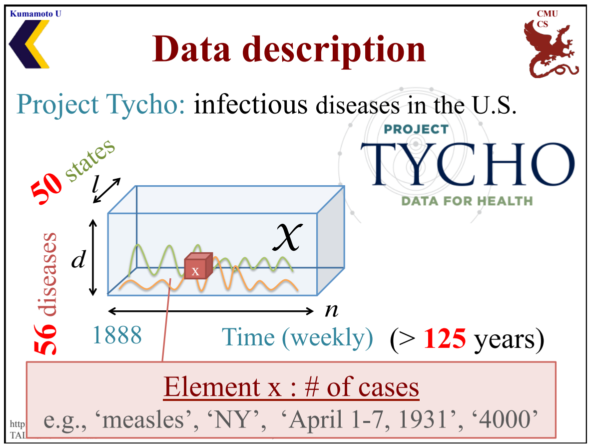
Given: Large set of epidemiological data

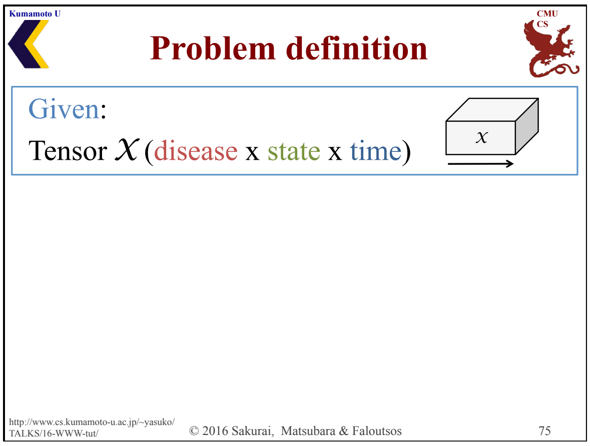
e.g., Measles cases in the U.S.

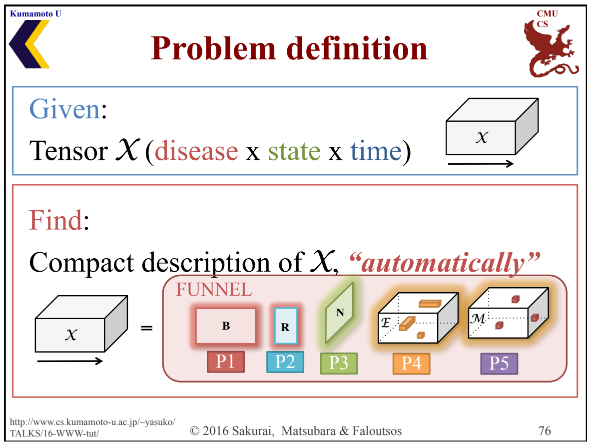
Goal: summarize all the epidemic time-series, **"fully-automatically"**

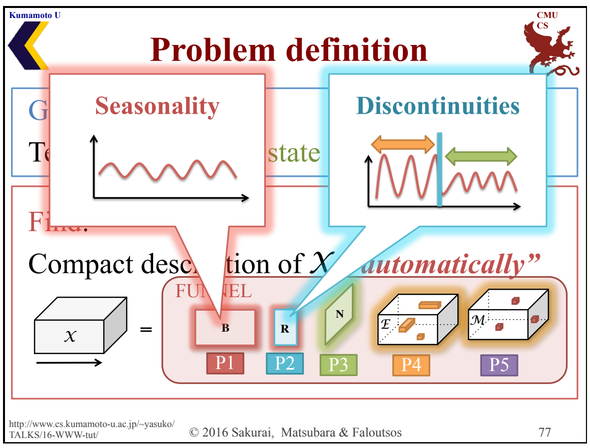


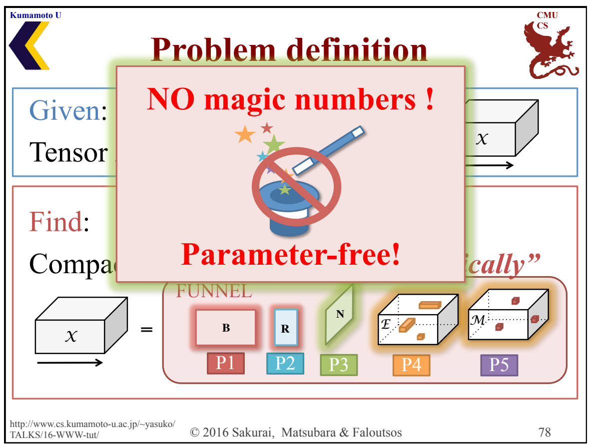

Kumamoto U
CMU CS
Data description
Project Tycho: infectious diseases in the U.S.
PROJECT TYCHO DATA FOR HEALTH
50 states
56 diseases
l
d
n
X
1888
Time (weekly)
(> 125 years)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
73

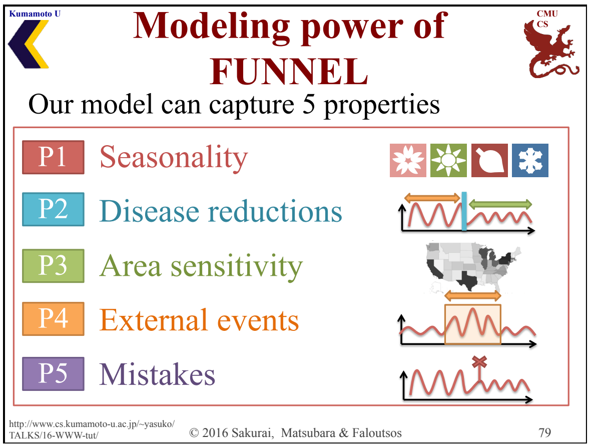

Kumamoto U
CMU CS
Data description
Project Tycho: infectious diseases in the U.S.
PROJECT TYCHO DATA FOR HEALTH
50 states
56 diseases
l
d
n
X
x
1888
Time (weekly)
(> 125 years)
Element x : # of cases
e.g., 'measles', 'NY', 'April 1-7, 1931', '4000'

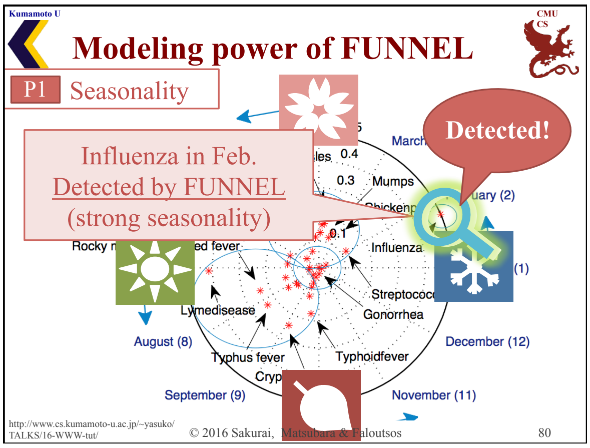

Kumamoto U
CMU CS
Problem definition
Given:
Tensor X (disease x state x time)
x
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
75

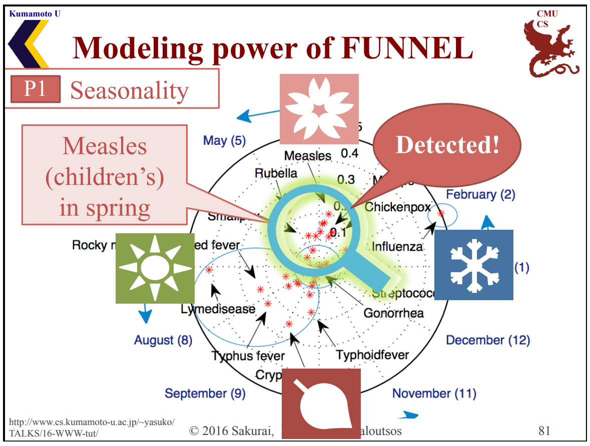

Kumamoto U
CMU CS
Problem definition
Given:
Tensor X (disease x state x time)
Find:
Compact description of X, "automatically"
x
=
FUNNEL
B R N E M
P1 P2 P3 P4 P5
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
76

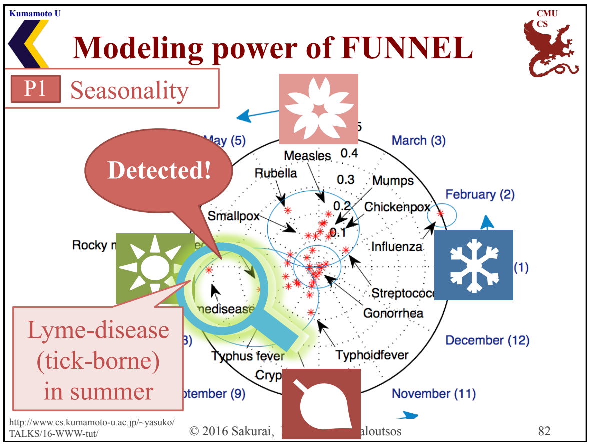

Kumamoto U
CMU CS
Problem definition
Seasonality
Discontinuities
Compact desc tion of X "automatically"
FUNNEL
B R N E M
P1 P2 P3 P4 P5
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
77

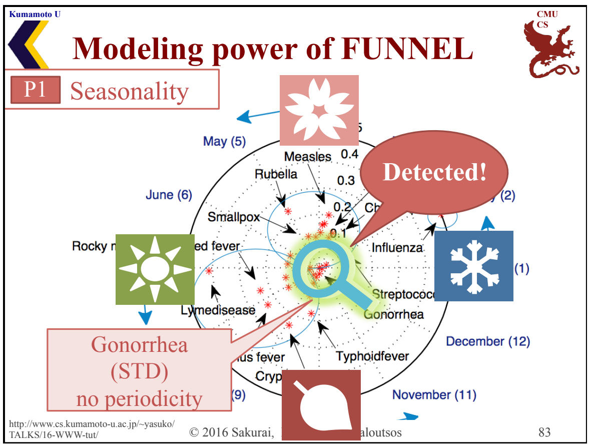

Kumamoto U
CMU CS
Problem definition
Given:
Tensor
Find:
Compa
cally"
NO magic numbers !
Parameter-free!
FUNNEL
B R N E M
P1 P2 P3 P4 P5
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
78

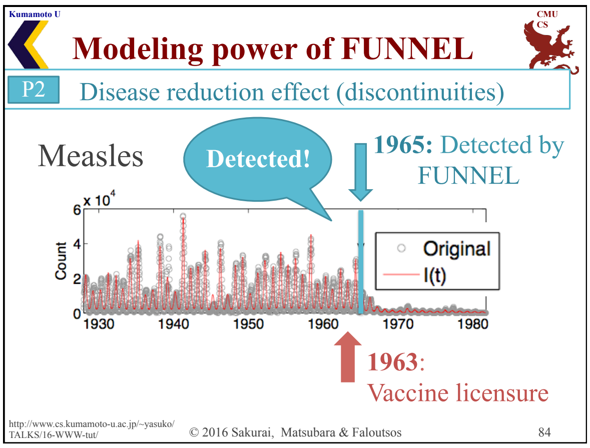

Kumamoto U
CMU CS
Modeling power of FUNNEL
Our model can capture 5 properties
P1 Seasonality
P2 Disease reductions
P3 Area sensitivity
P4 External events
P5 Mistakes
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
79

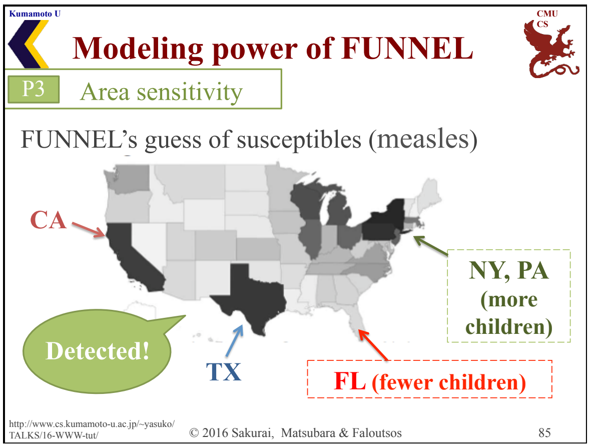

Kumamoto U
CMU CS
Modeling power of FUNNEL
P1 Seasonality
Influenza in Feb.
Detected by FUNNEL
(strong seasonality)
Detected!
March
Mumps
Influenza
Streptococc
Gonorrhea
Lymedisease
Typhus fever
Typhoidfever
August (8)
September (9)
November (11)
December (12)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai,
80

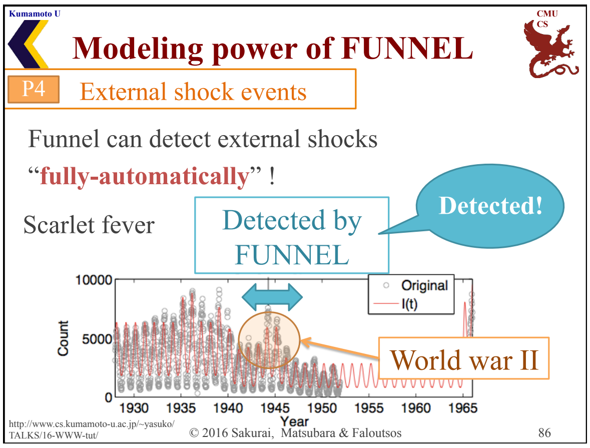

Kumamoto U
CMU CS
Modeling power of FUNNEL
P1 Seasonality
Measles
(children's)
in spring
Detected!
May (5)
Measles
Rubella
Chickenpox
February (2)
Influenza
Gonorrhea
Lymedisease
Typhus fever
Typhoidfever
August (8)
September (9)
November (11)
December (12)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai,
81

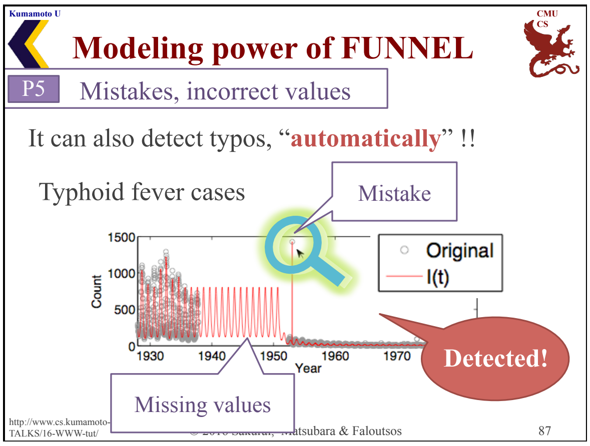

Kumamoto U
CMU CS
Modeling power of FUNNEL
P1 Seasonality
Detected!
Lyme-disease
(tick-borne)
in summer
May (5)
Measles
Rubella
March (3)
Mumps
February (2)
Smallpox
Chickenpox
Influenza
Streptococc
Gonorrhea
Typhus fever
Typhoidfever
November (11)
December (12)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai,
82

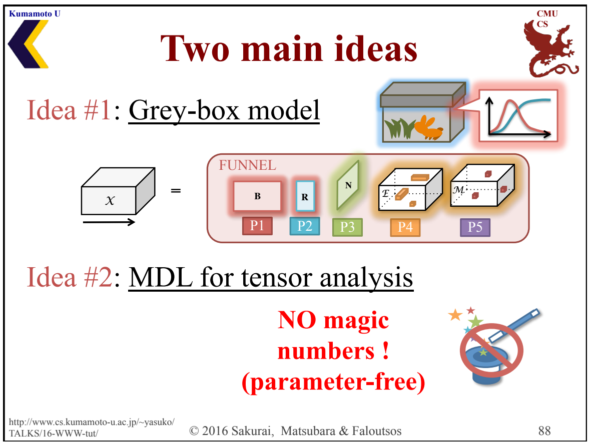

Kumamoto U
CMU CS
Modeling power of FUNNEL
P1 Seasonality
Detected!
Gonorrhea
(STD)
no periodicity
May (5)
June (6)
Measles
Rubella
Smallpox
Influenza
Streptococc
Gonorrhea
Lymedisease
Typhoidfever
November (11)
December (12)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai,
83

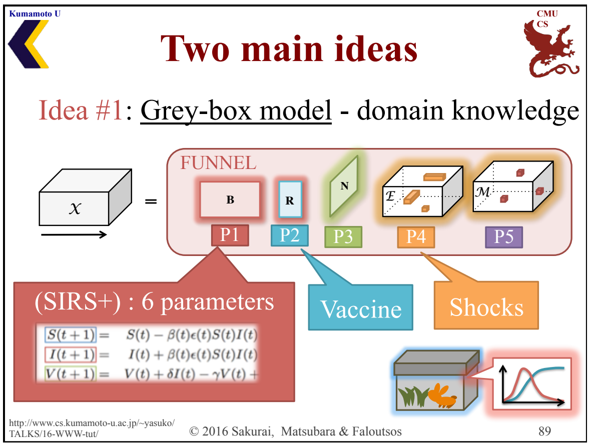

Kumamoto U
CMU CS
Modeling power of FUNNEL
P2 Disease reduction effect (discontinuities)
Measles
Detected!
1965: Detected by FUNNEL
Count
Original
I(t)
1930
1940
1950
1960
1970
1980
1963: Vaccine licensure
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
84

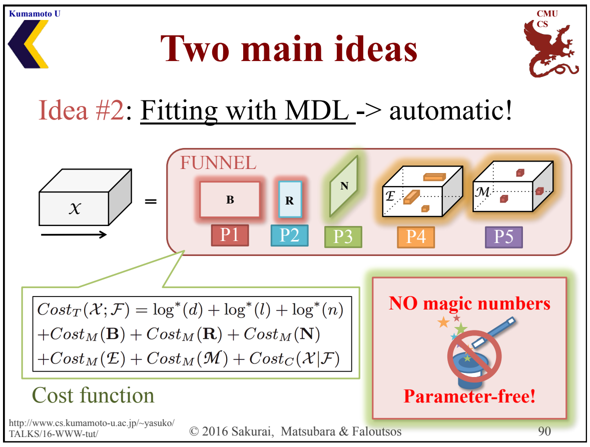

Kumamoto U
CMU CS
Modeling power of FUNNEL
P3 Area sensitivity
FUNNEL's guess of susceptibles (measles)
CA
NY, PA (more children)
Detected!
TX
FL (fewer children)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
85

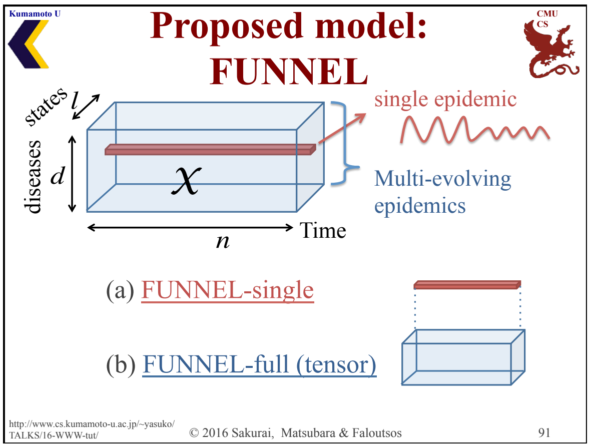

Kumamoto U
CMU CS
Modeling power of FUNNEL
P4 External shock events
Funnel can detect external shocks "fully-automatically" !
Scarlet fever
Detected by FUNNEL
Detected!
Original
I(t)
World war II
Count
10000
5000
0
1930 1935 1940 1945 1950 1955 1960 1965
Year
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
86

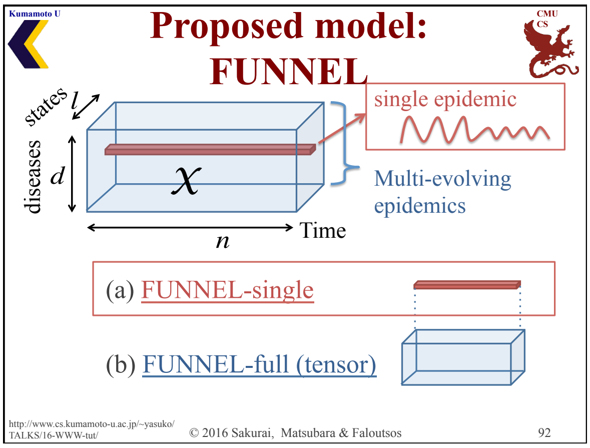

Kumamoto U
CMU CS
Modeling power of FUNNEL
P5 Mistakes, incorrect values
It can also detect typos, "automatically" !!
Typhoid fever cases
Mistake
Original
I(t)
Count
1500
1000
500
0
1930 1940 1950 1960 1970
Year
Detected!
Missing values
87

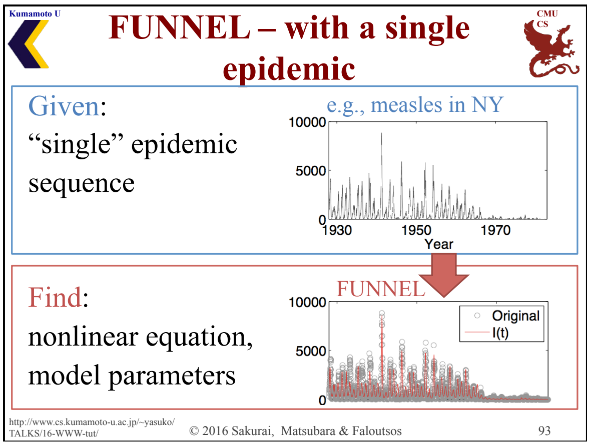

Kumamoto U
CMU CS
Two main ideas
Idea #1: Grey-box model
FUNNEL
B R N
P1 P2 P3 P4 P5
Idea #2: MDL for tensor analysis
NO magic numbers ! (parameter-free)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
88

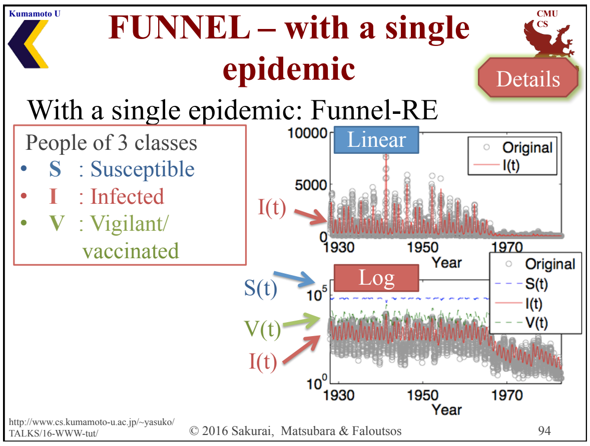

Kumamoto U
CMU CS
Two main ideas
Idea #1: Grey-box model - domain knowledge
FUNNEL
B R N
P1 P2 P3 P4 P5
(SIRS+) : 6 parameters
$S(t+1) = S(t) - \beta(t)\epsilon(t)S(t)I(t)$
$I(t+1) = I(t) + \beta(t)\epsilon(t)S(t)I(t)$
$V(t+1) = V(t) + \delta I(t) - \gamma V(t) +$
Vaccine
Shocks
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
89

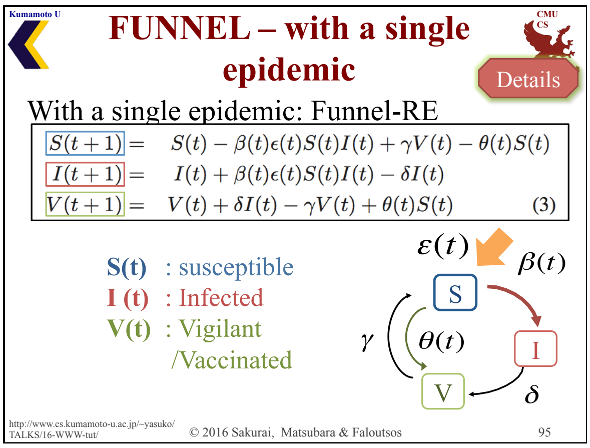

Kumamoto U
CMU CS
Two main ideas
Idea #2: Fitting with MDL -> automatic!
FUNNEL
B R N
P1 P2 P3 P4 P5
$Cost_T(\mathcal{X};\mathcal{F}) = \log^*(d) + \log^*(l) + \log^*(n)$
$+ Cost_M(\mathbf{B}) + Cost_M(\mathbf{R}) + Cost_M(\mathbf{N})$
$+ Cost_M(\mathcal{E}) + Cost_M(\mathcal{M}) + Cost_C(\mathcal{X}|\mathcal{F})$
Cost function
NO magic numbers
Parameter-free!
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
90

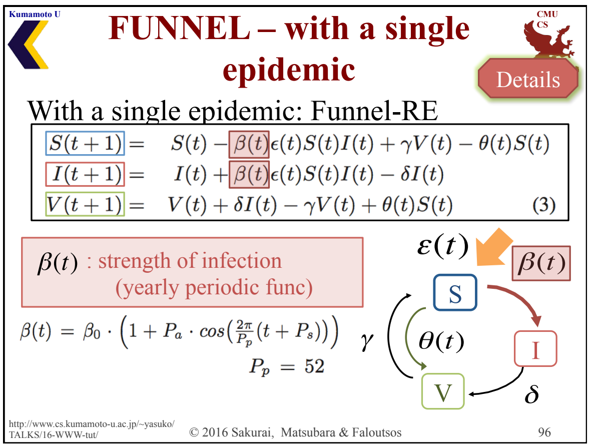

Kumamoto U
CMU CS
Proposed model: FUNNEL
states
l
diseases
d
X
n
Time
single epidemic
Multi-evolving epidemics
(a) FUNNEL-single
(b) FUNNEL-full (tensor)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
91

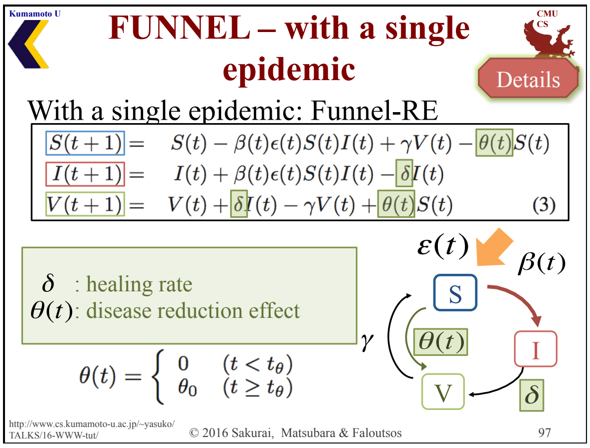

Kumamoto U
CMU CS
Proposed model: FUNNEL
states
l
diseases
d
X
n
Time
single epidemic
Multi-evolving epidemics
(a) FUNNEL-single
(b) FUNNEL-full (tensor)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
92

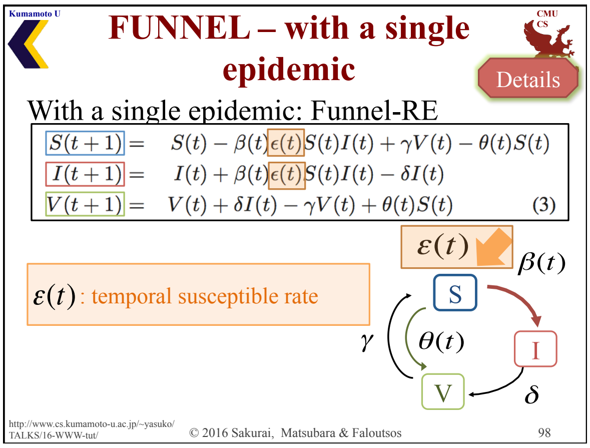

Kumamoto U
CMU CS
FUNNEL – with a single epidemic
Given:
"single" epidemic sequence
e.g., measles in NY
10000
5000
0
1930
1950
1970
Year
FUNNEL
Find:
nonlinear equation, model parameters
10000
5000
0
Original
I(t)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
93

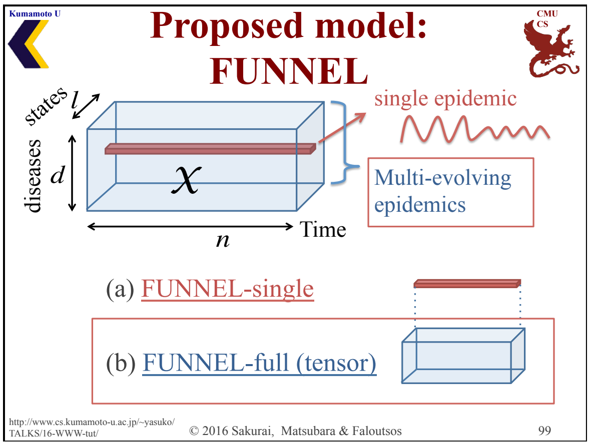

Kumamoto U
CMU CS
FUNNEL – with a single epidemic
Details
With a single epidemic: Funnel-RE
People of 3 classes
• S : Susceptible
• I : Infected
• V : Vigilant/ vaccinated
10000
5000
0
Linear
Original
I(t)
I(t)
1930
1950
1970
Year
Log
Original
S(t)
I(t)
V(t)
S(t)
V(t)
I(t)
10^5
10^0
1930
1950
1970
Year
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
94

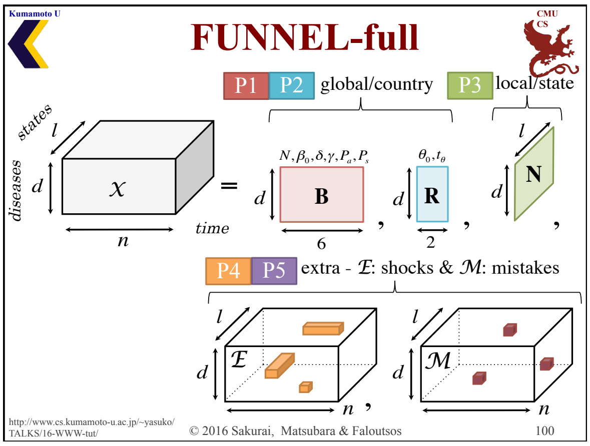

Kumamoto U
CMU CS
FUNNEL – with a single epidemic
Details
With a single epidemic: Funnel-RE
$S(t+1) = S(t) - \beta(t)\epsilon(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t)$
$I(t+1) = I(t) + \beta(t)\epsilon(t)S(t)I(t) - \delta I(t)$
$V(t+1) = V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t)$ (3)
S(t) : susceptible
I (t) : Infected
V(t) : Vigilant /Vaccinated
$\varepsilon(t)$
$\beta(t)$
S
I
V
γ
$\theta(t)$
δ
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
95

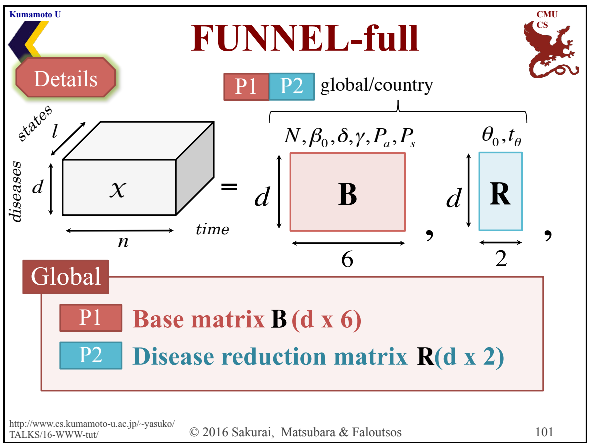

Kumamoto U
CMU CS
FUNNEL – with a single epidemic
Details
With a single epidemic: Funnel-RE
$S(t+1) = S(t) - \beta(t)\epsilon(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t)$
$I(t+1) = I(t) + \beta(t)\epsilon(t)S(t)I(t) - \delta I(t)$
$V(t+1) = V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t)$ (3)
$\beta(t)$: strength of infection (yearly periodic func)
$\beta(t) = \beta_0 \cdot \left(1 + P_a \cdot cos(\frac{2\pi}{P_p}(t + P_s))\right)$
$P_p = 52$
$\varepsilon(t)$
$\beta(t)$
S
I
V
γ
$\theta(t)$
δ
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
96

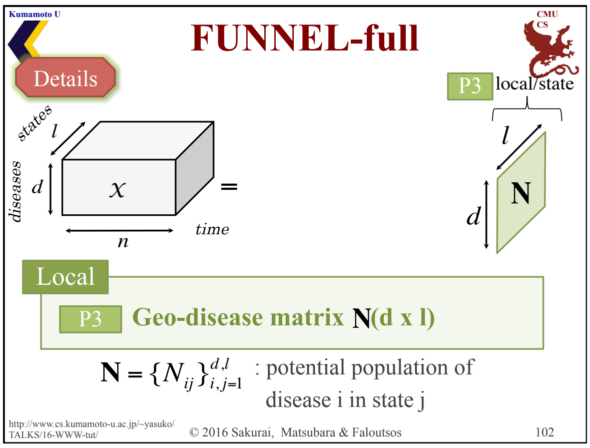

Kumamoto U
FUNNEL – with a single epidemic
CMU CS
Details
With a single epidemic: Funnel-RE
$S(t+1) = S(t) - \beta(t)\epsilon(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t)$
$I(t+1) = I(t) + \beta(t)\epsilon(t)S(t)I(t) - \delta I(t)$
$V(t+1) = V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t)$ (3)
δ : healing rate
$\theta(t)$: disease reduction effect
$\theta(t) = \begin{cases} 0 & (t < t_\theta) \\ \theta_0 & (t \geq t_\theta) \end{cases}$
$\varepsilon(t)$
$\beta(t)$
S
I
V
γ
$\theta(t)$
δ
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
97

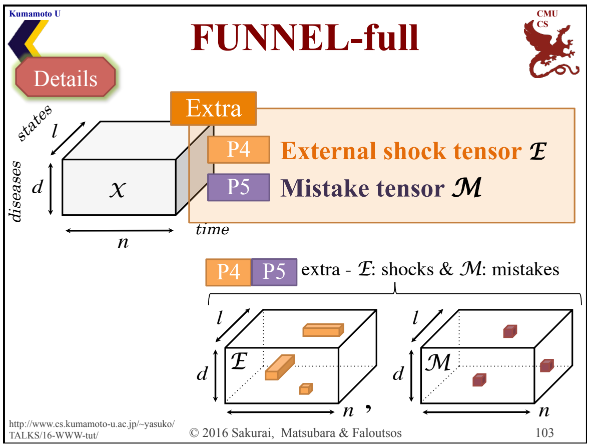

Kumamoto U
FUNNEL – with a single epidemic
CMU CS
Details
With a single epidemic: Funnel-RE
$S(t+1) = S(t) - \beta(t)\epsilon(t)S(t)I(t) + \gamma V(t) - \theta(t)S(t)$
$I(t+1) = I(t) + \beta(t)\epsilon(t)S(t)I(t) - \delta I(t)$
$V(t+1) = V(t) + \delta I(t) - \gamma V(t) + \theta(t)S(t)$ (3)
$\varepsilon(t)$: temporal susceptible rate
$\varepsilon(t)$
$\beta(t)$
S
I
V
γ
$\theta(t)$
δ
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
98

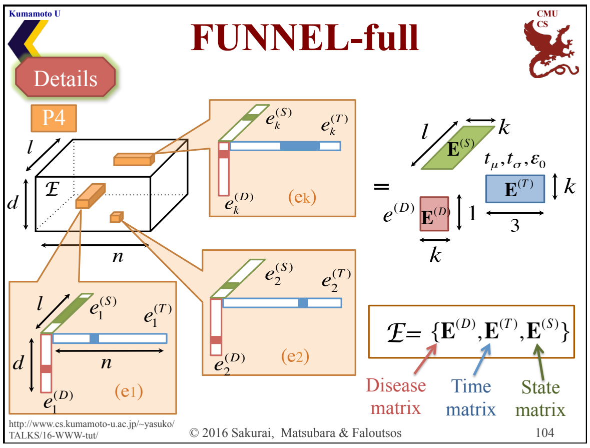

Kumamoto U
Proposed model: FUNNEL
CMU CS
states l
diseases d
$\mathcal{X}$
n
Time
single epidemic
Multi-evolving epidemics
(a) FUNNEL-single
(b) FUNNEL-full (tensor)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
99

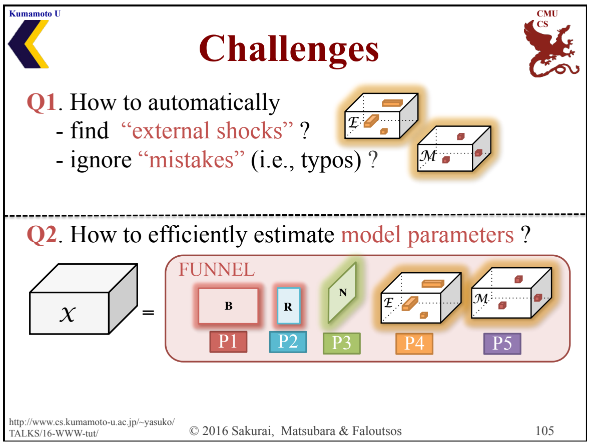

Kumamoto U
FUNNEL-full
CMU CS
P1 P2 global/country
P3 local/state
states l
diseases d
$\mathcal{X}$
n
time
=
$N, \beta_0, \delta, \gamma, P_a, P_s$
B
6
θ_0, t_θ
R
2
N
P4 P5 extra - $\mathcal{E}$: shocks & $\mathcal{M}$: mistakes
$\mathcal{E}$
$\mathcal{M}$
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
100

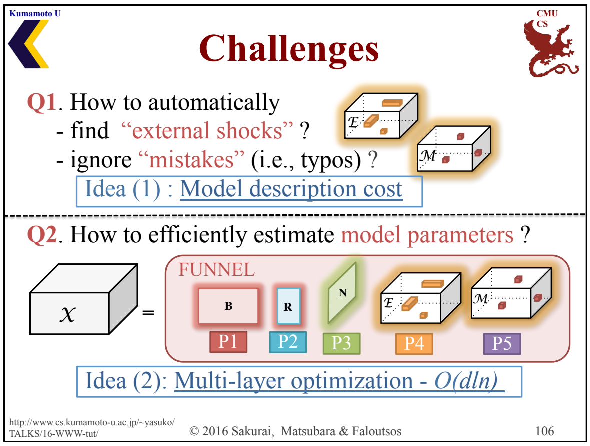

Kumamoto U
FUNNEL-full
CMU CS
Details
P1 P2 global/country
states l
diseases d
$\mathcal{X}$
n
time
=
$N, \beta_0, \delta, \gamma, P_a, P_s$
B
6
θ_0, t_θ
R
2
Global
P1 Base matrix B (d x 6)
P2 Disease reduction matrix R (d x 2)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
101

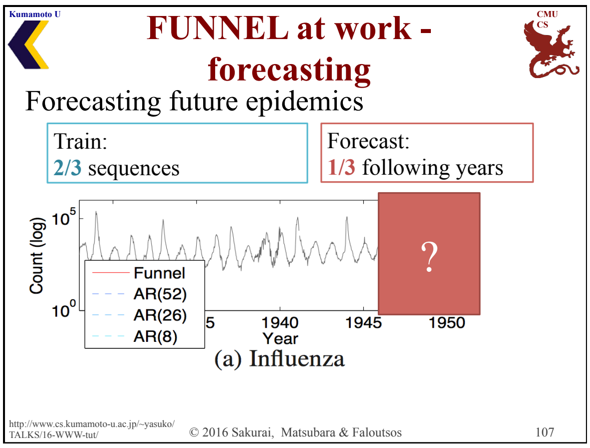

Kumamoto U
FUNNEL-full
CMU CS
Details
P3 local/state
states l
diseases d
$\mathcal{X}$
n
time
=
N
Local
P3 Geo-disease matrix N(d x l)
$\mathbf{N} = \{N_{ij}\}_{i,j=1}^{d,l}$: potential population of disease i in state j
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
102

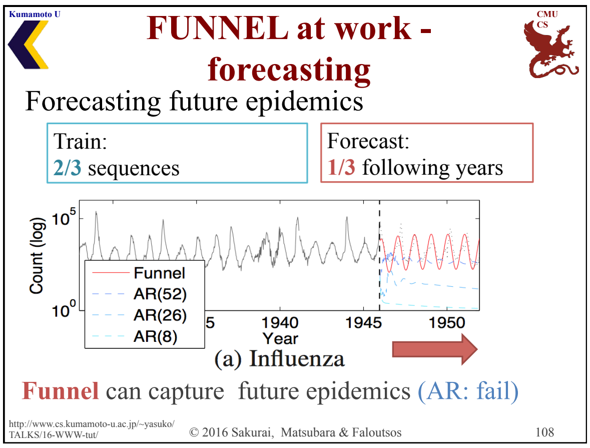

Kumamoto U
FUNNEL-full
CMU CS
Details
Extra
states
diseases
time
P4 External shock tensor ℰ
P5 Mistake tensor ℳ
P4 P5 extra - ℰ: shocks & ℳ: mistakes
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
103


Kumamoto U
FUNNEL-full
CMU CS
Details
P4
(ek)
(e2)
(e1)
Disease matrix
Time matrix
State matrix
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
104


Kumamoto U
Challenges
CMU CS
Q1. How to automatically
- find "external shocks" ?
- ignore "mistakes" (i.e., typos) ?
Q2. How to efficiently estimate model parameters ?
FUNNEL
P1 P2 P3 P4 P5
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
105


Kumamoto U
Challenges
CMU CS
Q1. How to automatically
- find "external shocks" ?
- ignore "mistakes" (i.e., typos) ?
Idea (1) : Model description cost
Q2. How to efficiently estimate model parameters ?
FUNNEL
P1 P2 P3 P4 P5
Idea (2): Multi-layer optimization - O(dln)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
106


Kumamoto U
FUNNEL at work - forecasting
CMU CS
Forecasting future epidemics
Train: 2/3 sequences
Forecast: 1/3 following years
Count (log)
Funnel
AR(52)
AR(26)
AR(8)
1940 1945 1950
Year
?
(a) Influenza
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
107

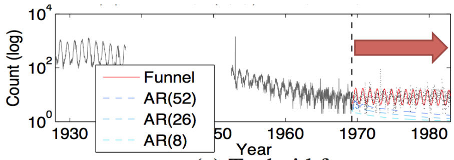

Kumamoto U
FUNNEL at work - forecasting
CMU CS
Forecasting future epidemics
Train: 2/3 sequences
Forecast: 1/3 following years
Count (log)
Funnel
AR(52)
AR(26)
AR(8)
1940 1945 1950
Year
(a) Influenza
Funnel can capture future epidemics (AR: fail)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
108

FUNNEL at work - forecasting

Forecasting future epidemics

Train: 2/3 sequences

Forecast: 1/3 following years



(c) Typhoid fever

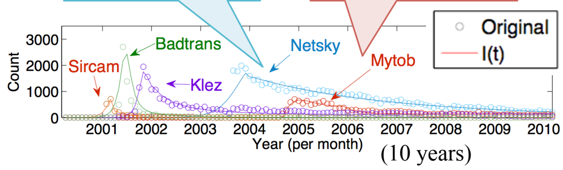
Funnel can capture future epidemics (AR: fail)

Generality of FUNNEL

Epidemics on computer networks

Spread via email attachment

Spread through corporate networks



(10 years)

Funnel is general: it fits computer virus very well!

[Matsubara+ WWW'16]

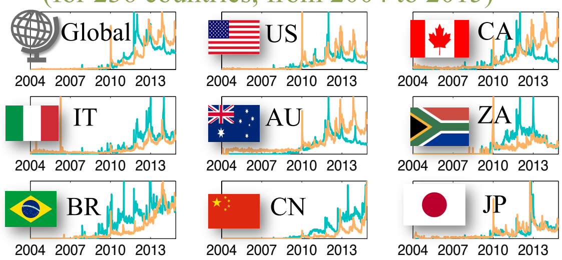
Non-linear Mining of Competing Local Activities

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

Given: local user activities

e.g., *Google* search volumes for **Kindle, Nexus**

(for 236 countries, from 2004 to 2015)



Given: local user activities

e.g., *Google* search volumes for **Kindle, Nexus**

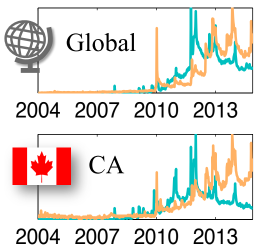
(for 236 countries, from 2004 to 2015)

Q. Any global/local trends?

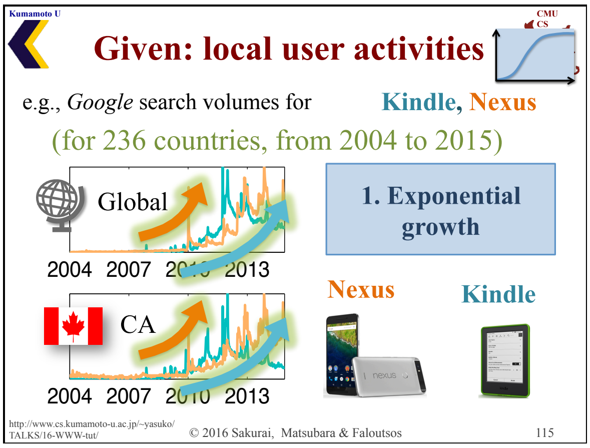
Given: local user activities

e.g., *Google* search volumes for **Kindle, Nexus**

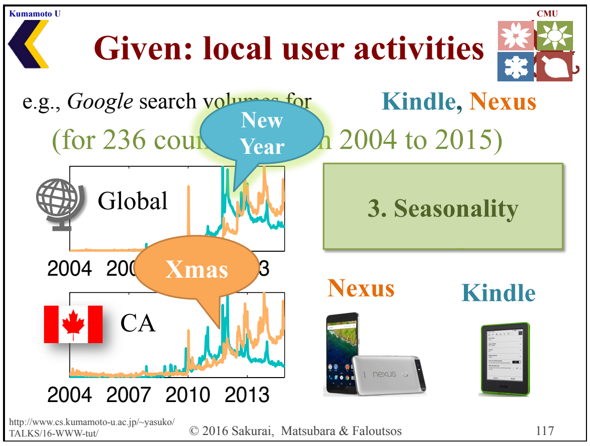
(for 236 countries, from 2004 to 2015)

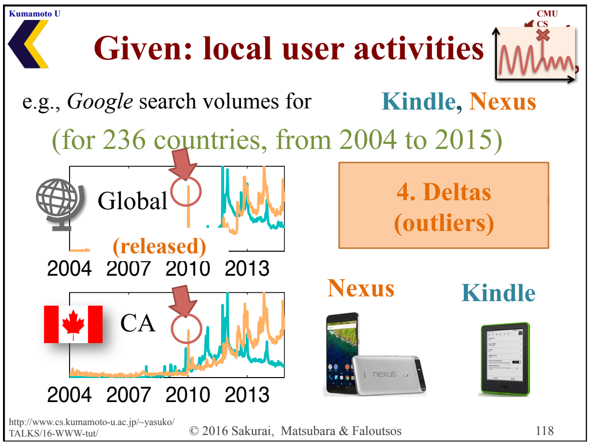


Nexus **Kindle**

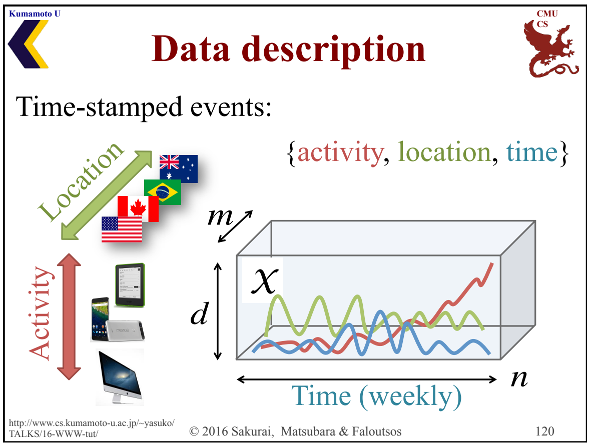

Given: local user activities
e.g., Google search volumes for Kindle, Nexus
(for 236 countries, from 2004 to 2015)
Global
2004 2007 2010 2013
CA
2004 2007 2010 2013
1. Exponential growth
Nexus
Kindle
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
115

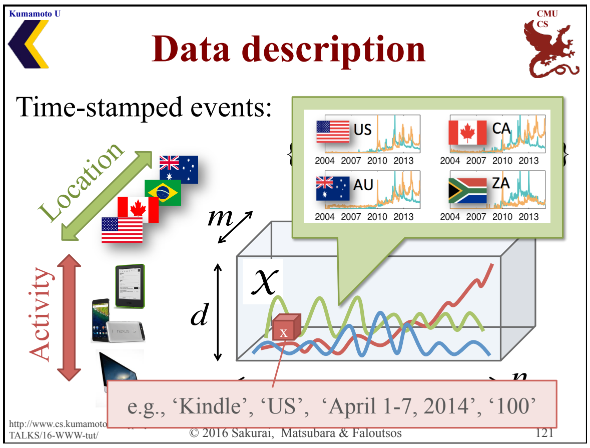

Given: local user activities
e.g., Google search volumes for Kindle, Nexus
(for 236 countries, from 2004 to 2015)
Global
VS.
2004 2007 2010 2013
CA
VS.
2004 2007 2010 2013
2. Competition
Nexus
Kindle
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
116

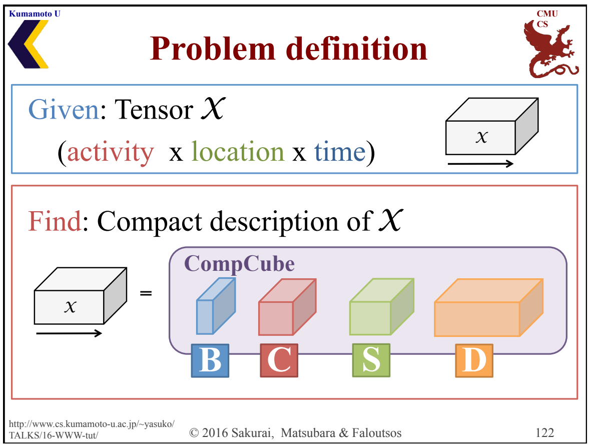

Given: local user activities
e.g., Google search volumes for Kindle, Nexus
(for 236 countries, from 2004 to 2015)
New Year
Global
Xmas
CA
2004 2007 2010 2013
3. Seasonality
Nexus
Kindle
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
117

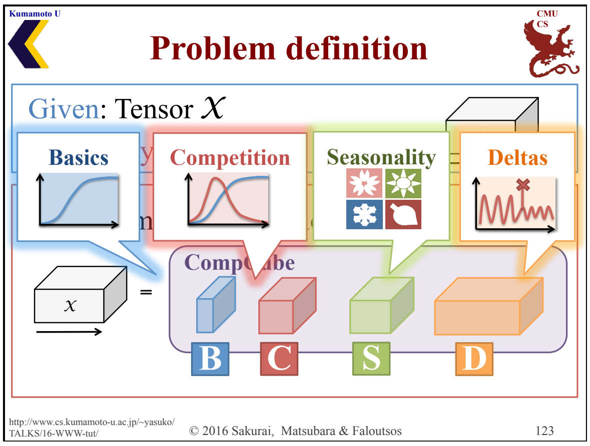

Given: local user activities
e.g., Google search volumes for Kindle, Nexus
(for 236 countries, from 2004 to 2015)
Global
(released)
2004 2007 2010 2013
CA
2004 2007 2010 2013
4. Deltas (outliers)
Nexus
Kindle
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
118

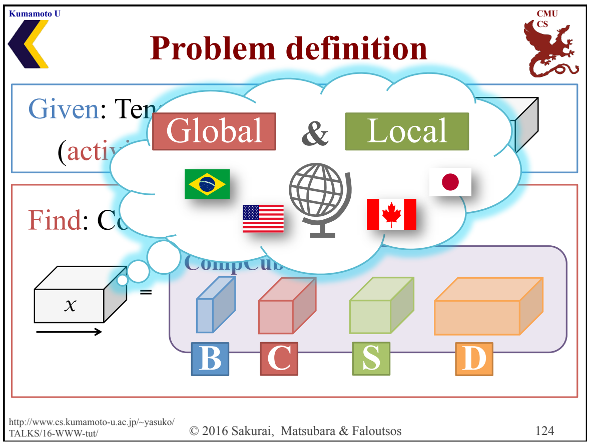

Given: local user activities
e.g., Google search volumes for Kindle, Nexus
(for 236 countries, from 2004 to 2015)
Global
4. Deltas
Goal: find global/local patterns,
fully automatically
CA
2004 2007 2010 2013
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
119

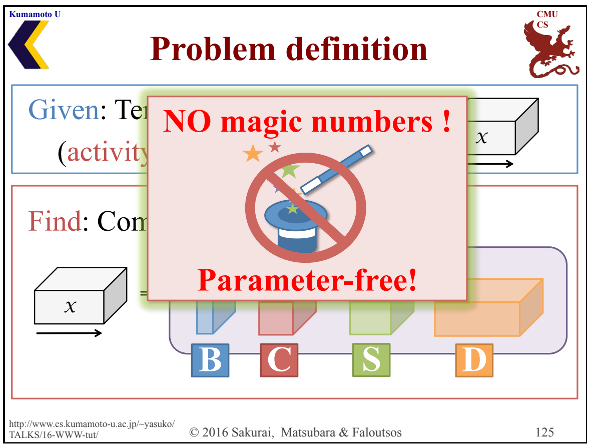

Data description
Time-stamped events:
{activity, location, time}
Location
Activity
m
d
X
Time (weekly)
n
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
120

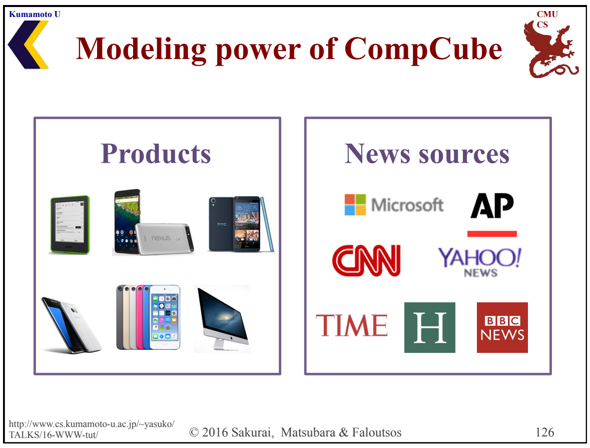

Kumamoto U
CMU CS
Data description
Time-stamped events:
US
CA
AU
ZA
2004 2007 2010 2013
Location
Activity
m
d
n
$\mathcal{X}$
x
e.g., 'Kindle', 'US', 'April 1-7, 2014', '100'
© 2016 Sakurai, Matsubara & Faloutsos
121

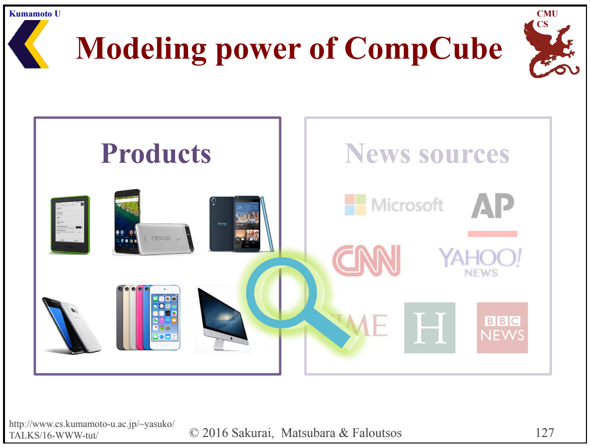

Kumamoto U
CMU CS
Problem definition
Given: Tensor $\mathcal{X}$
(activity x location x time)
x
Find: Compact description of $\mathcal{X}$
x
=
CompCube
B
C
S
D
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
122

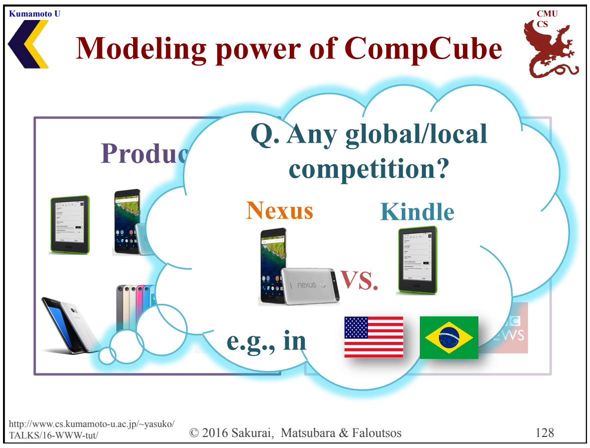

Kumamoto U
CMU CS
Problem definition
Given: Tensor $\mathcal{X}$
Basics
Competition
Seasonality
Deltas
x
=
CompCube
B
C
S
D
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
123

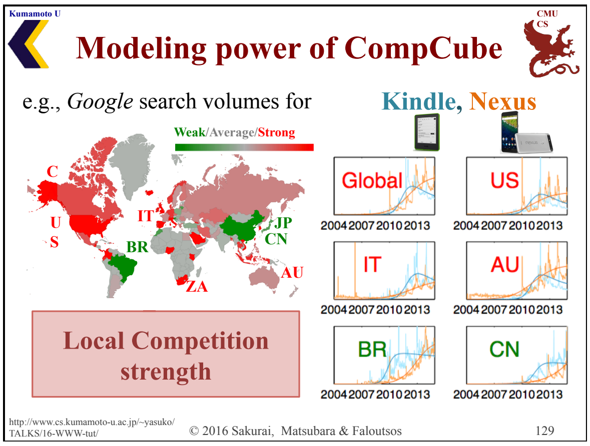

Kumamoto U
CMU CS
Problem definition
Given: Ten
(acti
Find: Co
Global
&
Local
x
=
B
C
S
D
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
124

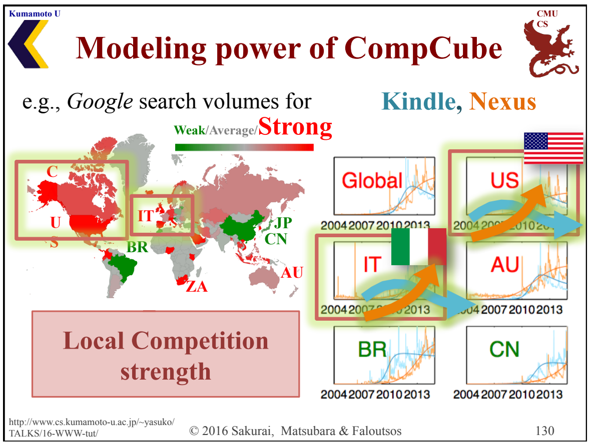

Kumamoto U
CMU CS
Problem definition
Given: Te
(activity
Find: Com
NO magic numbers !
Parameter-free!
x
x
=
B
C
S
D
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
125

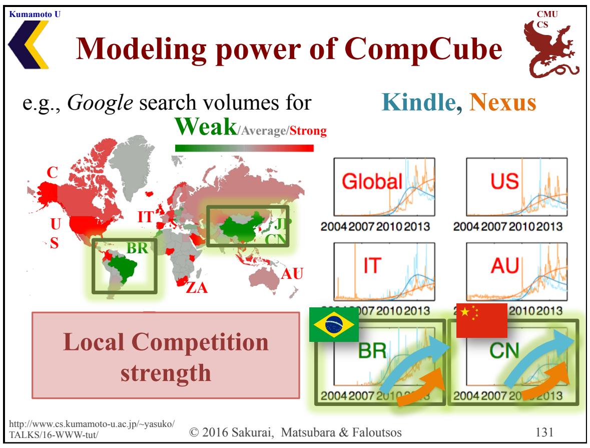

Kumamoto U
CMU CS
Modeling power of CompCube
Products
News sources
Microsoft
AP
CNN
YAHOO! NEWS
TIME
H
BBC NEWS
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
126

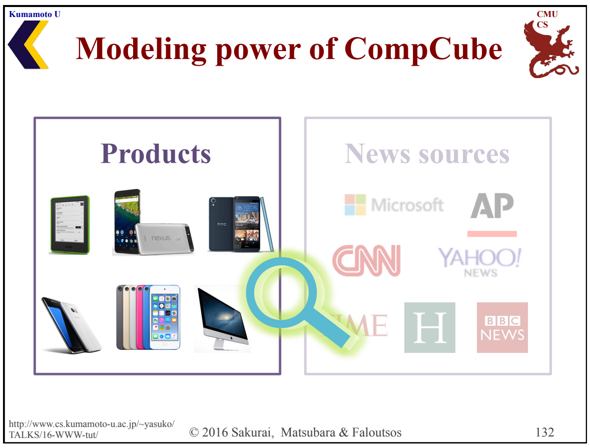

Kumamoto U
CMU CS
Modeling power of CompCube
Products
News sources
Microsoft
AP
CNN
YAHOO! NEWS
BBC NEWS
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
127

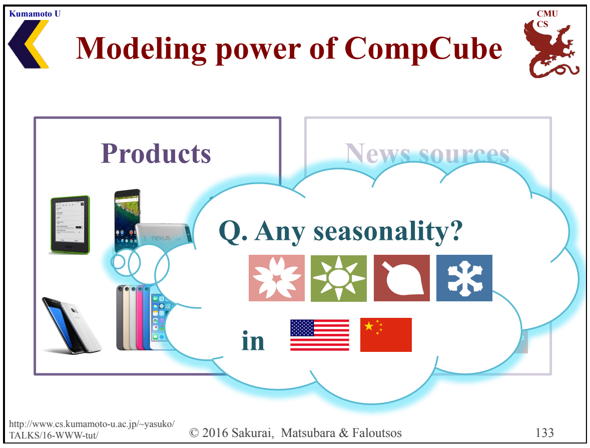

Kumamoto U
CMU CS
Modeling power of CompCube
Q. Any global/local competition?
Nexus
Kindle
VS.
e.g., in
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
128

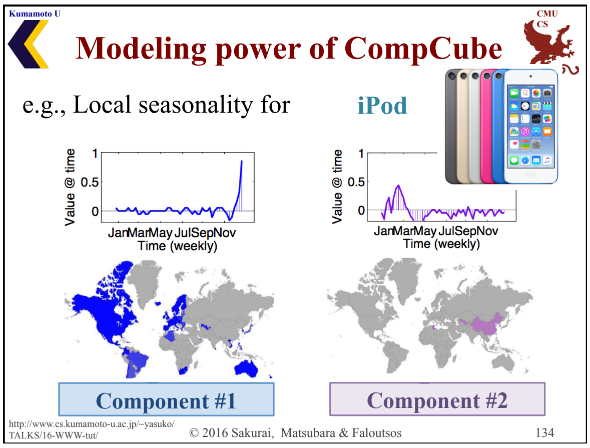

Kumamoto U
CMU CS
Modeling power of CompCube
e.g., Google search volumes for
Kindle, Nexus
Weak/Average/Strong
C
U
S
IT
BR
JP
CN
AU
ZA
Local Competition strength
Global
US
IT
AU
BR
CN
2004 2007 2010 2013
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
129

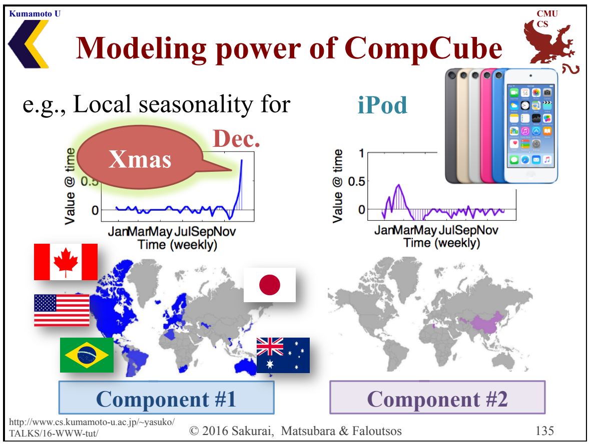

Kumamoto U
CMU CS
Modeling power of CompCube
e.g., Google search volumes for
Kindle, Nexus
Weak/Average/Strong
Local Competition strength
Global
US
IT
AU
BR
CN
2004 2007 2010 2013
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
130

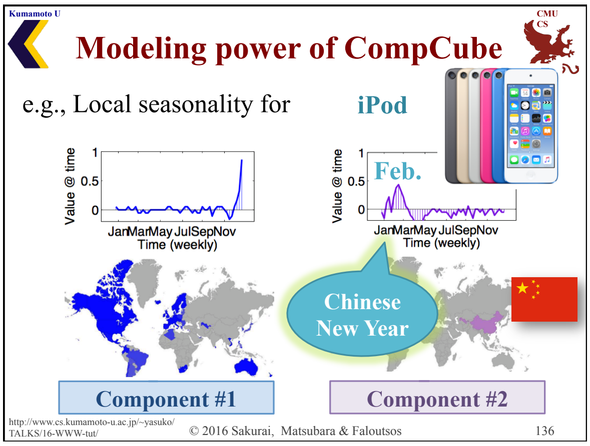

Kumamoto U
CMU CS
Modeling power of CompCube
e.g., Google search volumes for
Kindle, Nexus
Weak/Average/Strong
Local Competition strength
Global
US
IT
AU
BR
CN
2004 2007 2010 2013
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
131

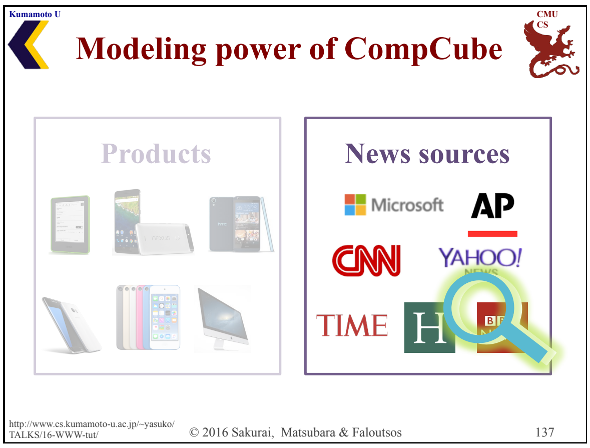

Kumamoto U
CMU CS
Modeling power of CompCube
Products
News sources
Microsoft
AP
CNN
YAHOO! NEWS
BBC NEWS
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
132

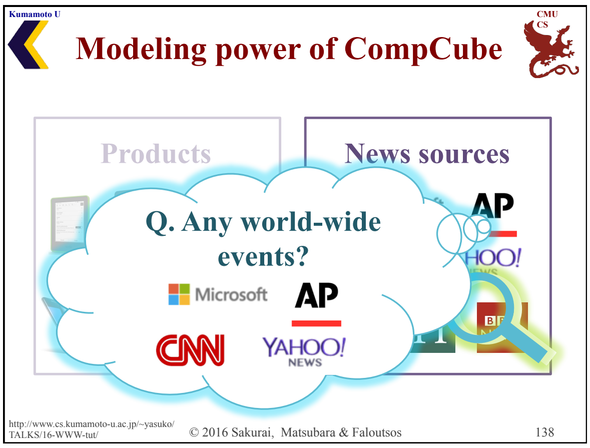

Kumamoto U
CMU CS
Modeling power of CompCube
Products
News sources
Q. Any seasonality?
in
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
133

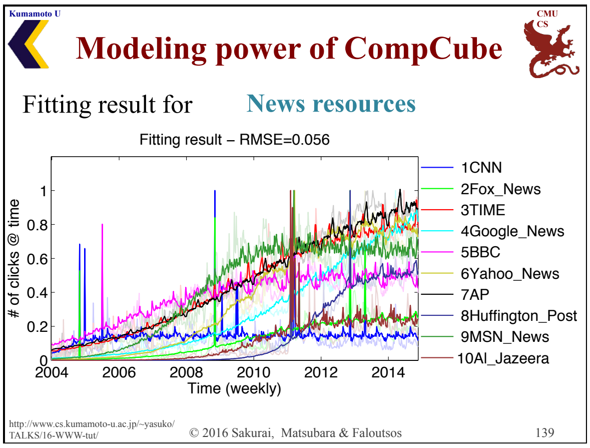

Kumamoto U
CMU CS
Modeling power of CompCube
e.g., Local seasonality for iPod
Value @ time
1
0.5
0
Jan Mar May Jul Sep Nov
Time (weekly)
Component #1
Component #2
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
134

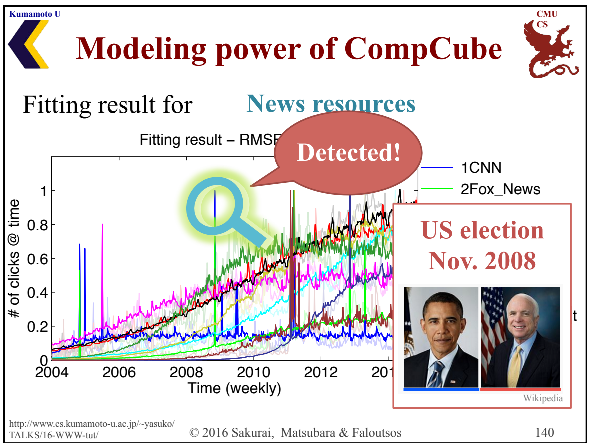

Kumamoto U
CMU CS
Modeling power of CompCube
e.g., Local seasonality for iPod
Xmas
Dec.
Value @ time
Jan Mar May Jul Sep Nov
Time (weekly)
Component #1
Component #2
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
135

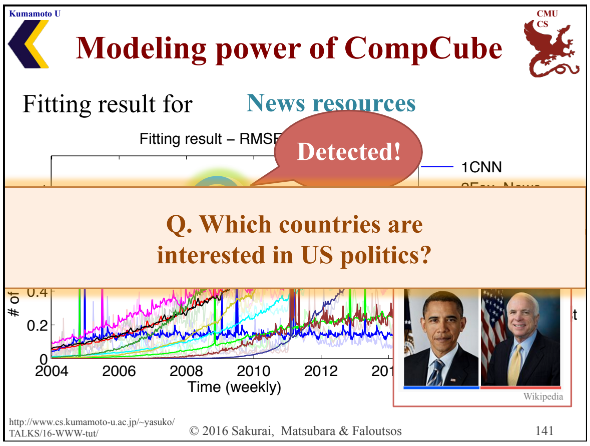

Kumamoto U
CMU CS
Modeling power of CompCube
e.g., Local seasonality for iPod
Feb.
Value @ time
Jan Mar May Jul Sep Nov
Time (weekly)
Chinese New Year
Component #1
Component #2
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
136

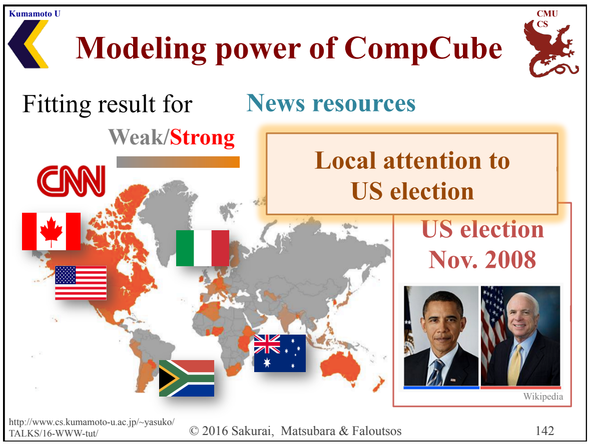

Kumamoto U
CMU CS
Modeling power of CompCube
Products
News sources
Microsoft
AP
CNN
YAHOO!
TIME
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
137

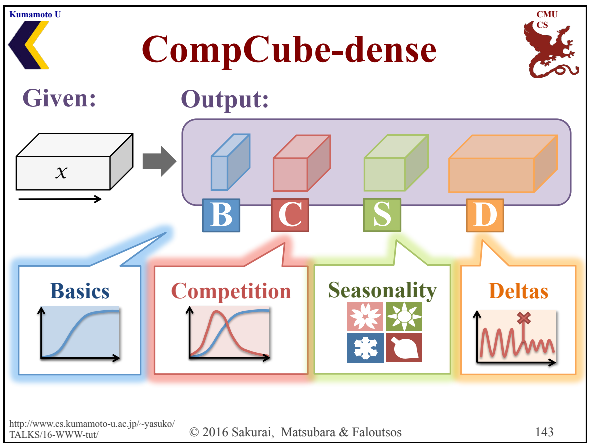

Kumamoto U
CMU CS
Modeling power of CompCube
Products
News sources
Q. Any world-wide events?
Microsoft
AP
CNN
YAHOO! NEWS
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
138

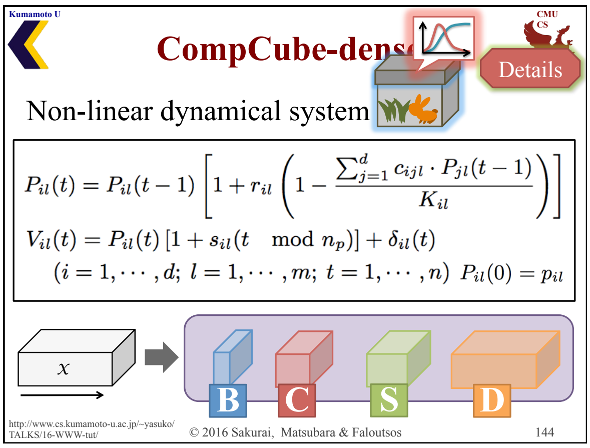

Kumamoto U
CMU CS
Modeling power of CompCube
Fitting result for News resources
Fitting result – RMSE=0.056
# of clicks @ time
Time (weekly)
2004 2006 2008 2010 2012 2014
0 0.2 0.4 0.6 0.8 1
1CNN
2Fox_News
3TIME
4Google_News
5BBC
6Yahoo_News
7AP
8Huffington_Post
9MSN_News
10Al_Jazeera
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
139

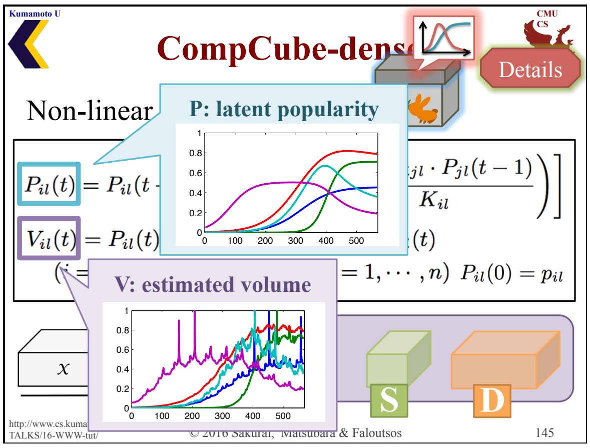

Kumamoto U
CMU CS
Modeling power of CompCube
Fitting result for News resources
Detected!
US election Nov. 2008
Wikipedia
1CNN
2Fox_News
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
140

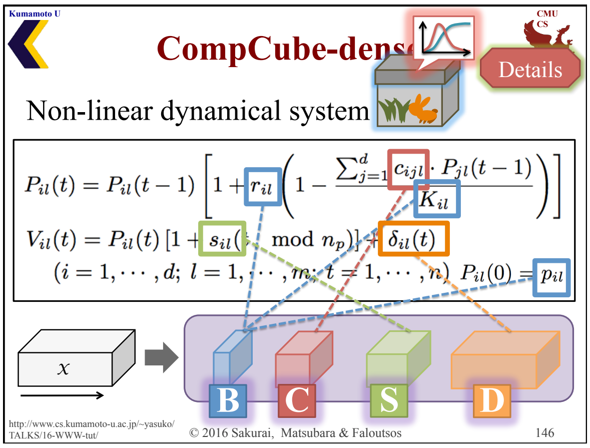

Kumamoto U
CMU CS
Modeling power of CompCube
Fitting result for News resources
Detected!
Q. Which countries are interested in US politics?
Wikipedia
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
141

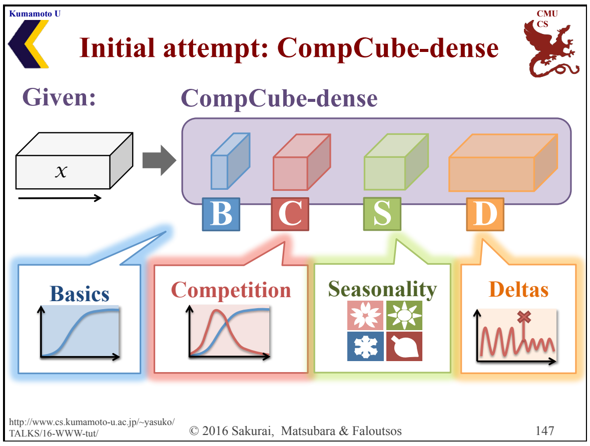

Kumamoto U
CMU CS
Modeling power of CompCube
Fitting result for News resources
Weak/Strong
CNN
Local attention to US election
US election Nov. 2008
Wikipedia
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
142

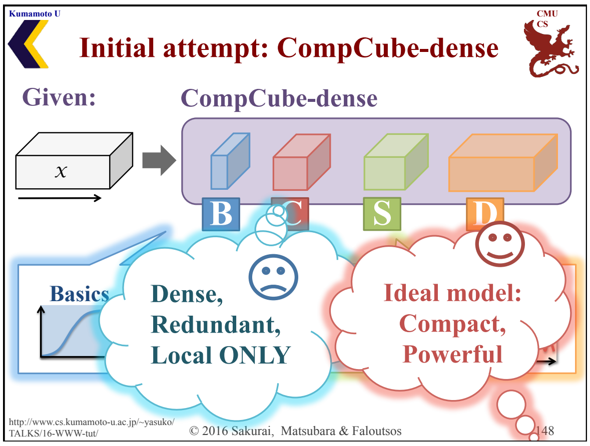

Kumamoto U
CMU CS
CompCube-dense
Given:
Output:
x
B C S D
Basics
Competition
Seasonality
Deltas
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
143

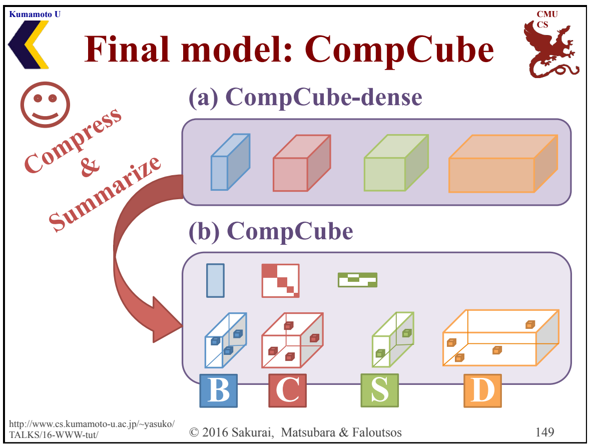

Kumamoto U
CMU CS
CompCube-dense
Details
Non-linear dynamical system
$P_{il}(t) = P_{il}(t-1)\left[1 + r_{il}\left(1 - \frac{\sum_{j=1}^{d} c_{ijl} \cdot P_{jl}(t-1)}{K_{il}}\right)\right]$
$V_{il}(t) = P_{il}(t)\left[1 + s_{il}(t \mod n_p)\right] + \delta_{il}(t)$
$(i = 1, \cdots, d;\ l = 1, \cdots, m;\ t = 1, \cdots, n)\ P_{il}(0) = p_{il}$
x
B C S D
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
144

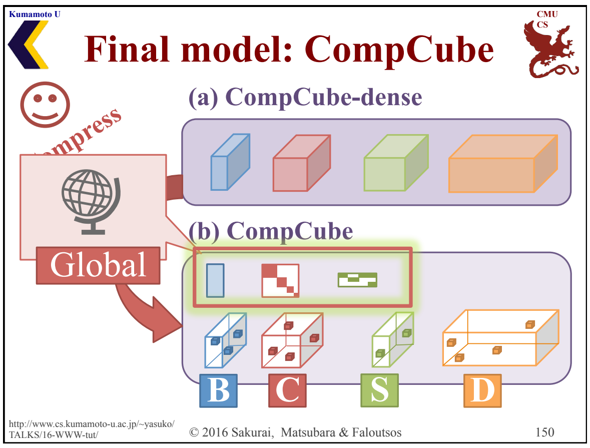

Kumamoto U
CMU CS
CompCube-dens
Details
Non-linear
P: latent popularity
V: estimated volume
x
S
D
© 2016 Sakurai, Matsubara & Faloutsos
145

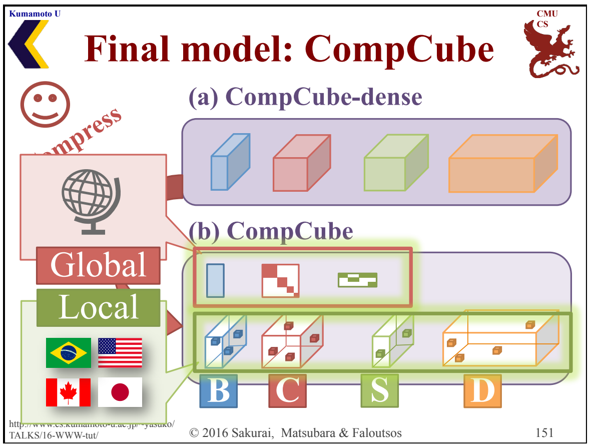

Kumamoto U
CMU CS
CompCube-dens
Details
Non-linear dynamical system
$P_{il}(t) = P_{il}(t-1)\left[1 + r_{il}\left(1 - \frac{\sum_{j=1}^{d} c_{ijl} \cdot P_{jl}(t-1)}{K_{il}}\right)\right]$
$V_{il}(t) = P_{il}(t)\,[1 + s_{il}(t \bmod n_p)] + \delta_{il}(t)$
$(i = 1,\cdots,d;\ l = 1,\cdots,m;\ t = 1,\cdots,n)\ P_{il}(0) = p_{il}$
x
B
C
S
D
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
146

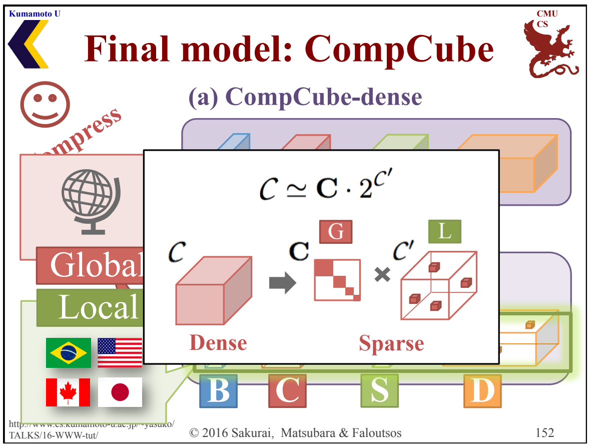

Kumamoto U
CMU CS
Initial attempt: CompCube-dense
Given:
CompCube-dense
x
B
C
S
D
Basics
Competition
Seasonality
Deltas
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
147

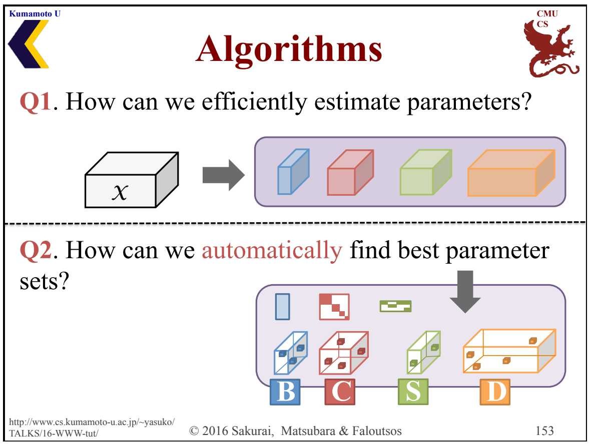

Kumamoto U
CMU CS
Initial attempt: CompCube-dense
Given:
CompCube-dense
x
B
S
Basics
Dense,
Redundant,
Local ONLY
Ideal model:
Compact,
Powerful
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
148

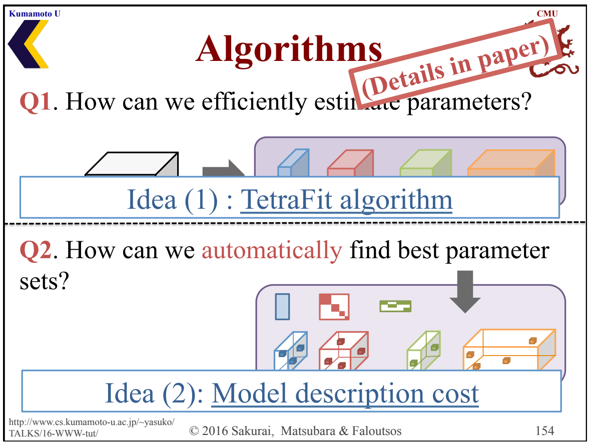

Kumamoto U
CMU CS
Final model: CompCube
(a) CompCube-dense
Compress
&
Summarize
(b) CompCube
B
C
S
D
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
149

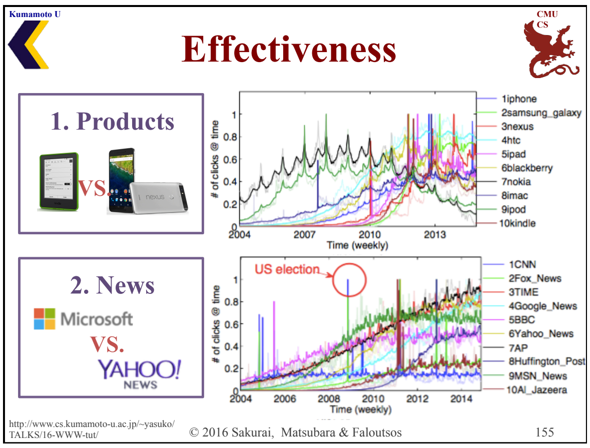

Kumamoto U
CMU CS
Final model: CompCube
(a) CompCube-dense
Global
(b) CompCube
B
C
S
D
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
150

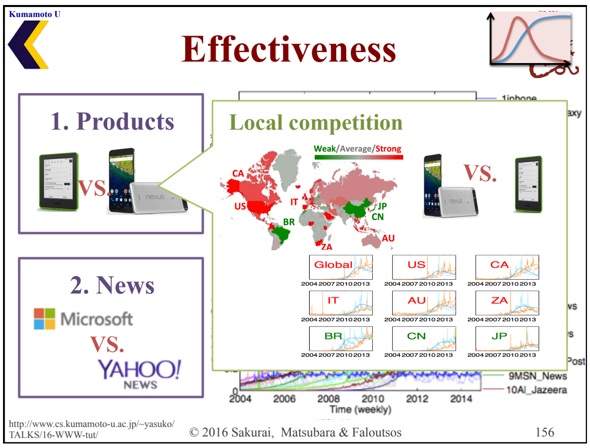

Kumamoto U
CMU CS
Final model: CompCube
(a) CompCube-dense
(b) CompCube
compress
Global
Local
B C S D
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
151

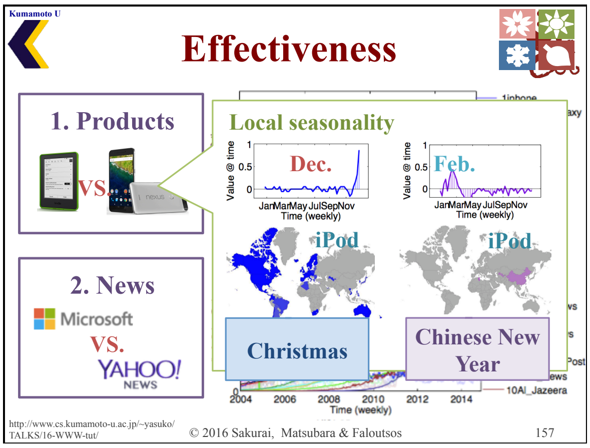

Kumamoto U
CMU CS
Final model: CompCube
(a) CompCube-dense
$\mathcal{C} \simeq \mathbf{C} \cdot 2^{\mathcal{C}'}$
$\mathcal{C}$
$\mathbf{C}$
$\mathcal{C}'$
G
L
Dense
Sparse
Global
Local
B C S D
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
152

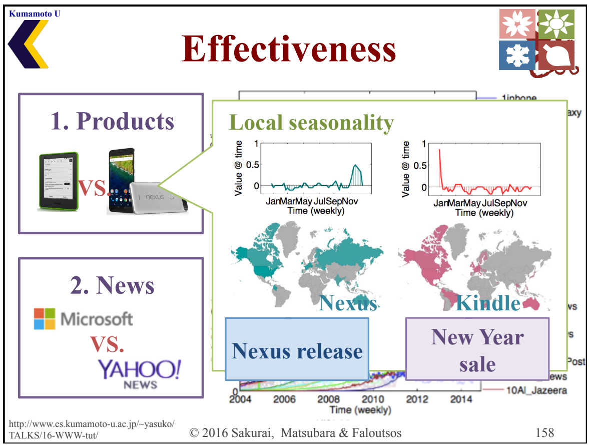

Kumamoto U
CMU CS
Algorithms
Q1. How can we efficiently estimate parameters?
x
Q2. How can we automatically find best parameter sets?
B C S D
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
153

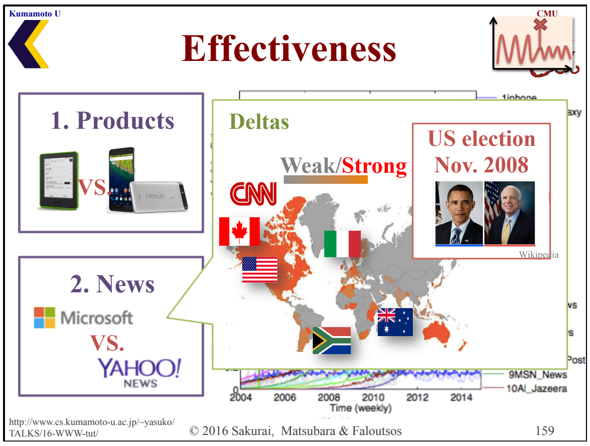

Kumamoto U
CMU CS
Algorithms
(Details in paper)
Q1. How can we efficiently estimate parameters?
Idea (1) : TetraFit algorithm
Q2. How can we automatically find best parameter sets?
Idea (2): Model description cost
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
154

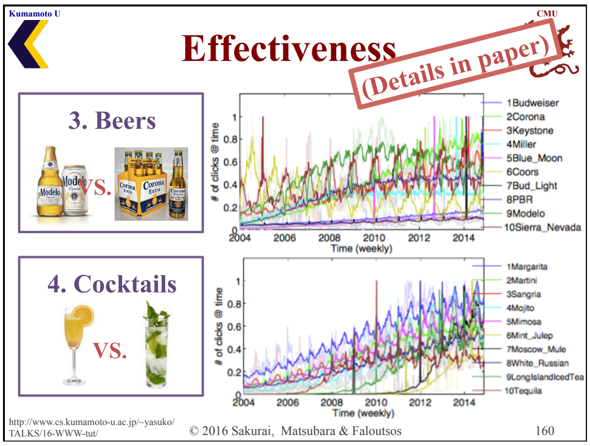

Kumamoto U
CMU CS
Effectiveness
1. Products
VS.
2. News
Microsoft
VS.
YAHOO! NEWS
1iphone
2samsung_galaxy
3nexus
4htc
5ipad
6blackberry
7nokia
8imac
9ipod
10kindle
US election
1CNN
2Fox_News
3TIME
4Google_News
5BBC
6Yahoo_News
7AP
8Huffington_Post
9MSN_News
10Al_Jazeera
# of clicks @ time
Time (weekly)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
155

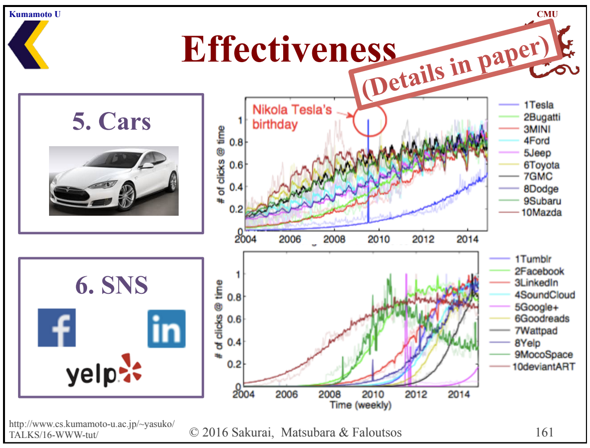

Kumamoto U
Effectiveness
1. Products
2. News
Microsoft
VS.
YAHOO! NEWS
Local competition
Weak/Average/Strong
VS.
Global
US
CA
IT
AU
ZA
BR
CN
JP
Time (weekly)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
156

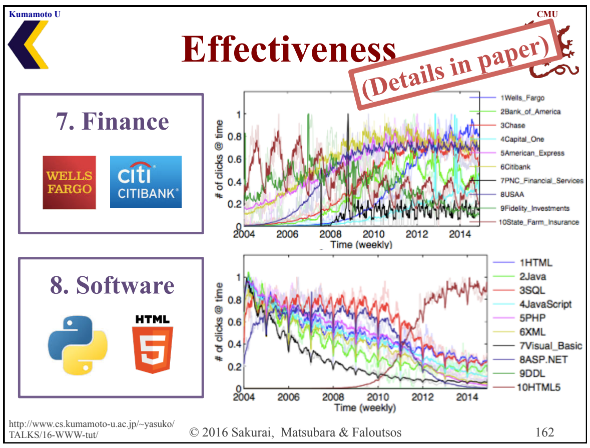

Kumamoto U
Effectiveness
1. Products
VS.
2. News
Microsoft
VS.
YAHOO!
NEWS
Local seasonality
Dec.
Feb.
iPod
iPod
Christmas
Chinese New Year
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
157

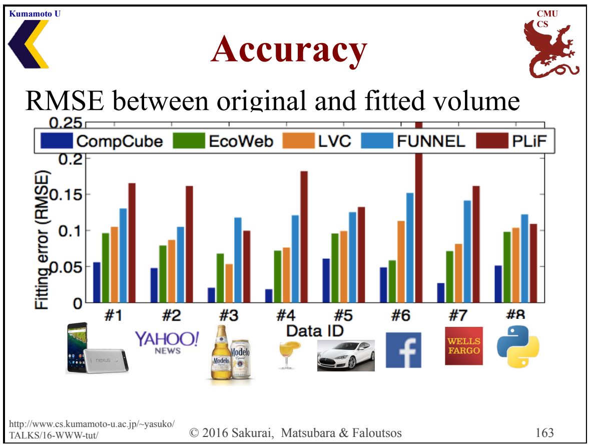

Kumamoto U
Effectiveness
1. Products
VS.
2. News
Microsoft
VS.
YAHOO!
NEWS
Local seasonality
Nexus
Kindle
Nexus release
New Year sale
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
158

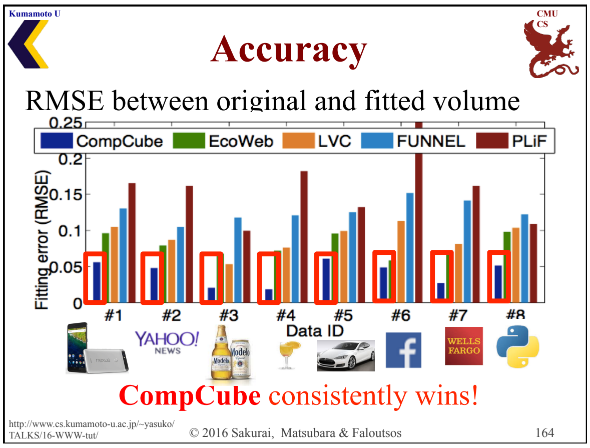

Kumamoto U
CMU
Effectiveness
1. Products
VS.
2. News
Microsoft
VS.
YAHOO!
NEWS
Deltas
Weak/Strong
CNN
US election
Nov. 2008
Wikipedia
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
159

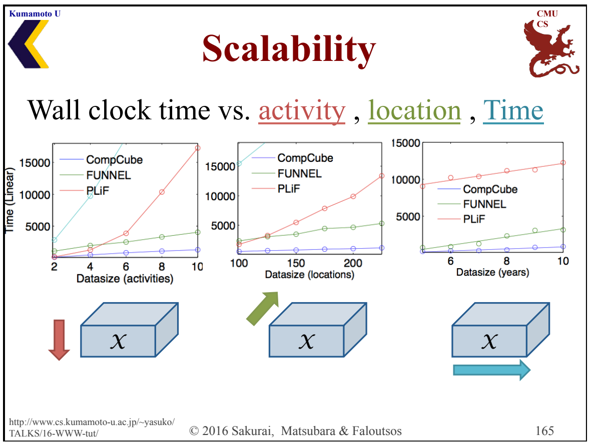

Kumamoto U
CMU
Effectiveness
(Details in paper)
3. Beers
VS.
1Budweiser
2Corona
3Keystone
4Miller
5Blue_Moon
6Coors
7Bud_Light
8PBR
9Modelo
10Sierra_Nevada
4. Cocktails
VS.
1Margarita
2Martini
3Sangria
4Mojito
5Mimosa
6Mint_Julep
7Moscow_Mule
8White_Russian
9LongIslandIcedTea
10Tequila
Time (weekly)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
160

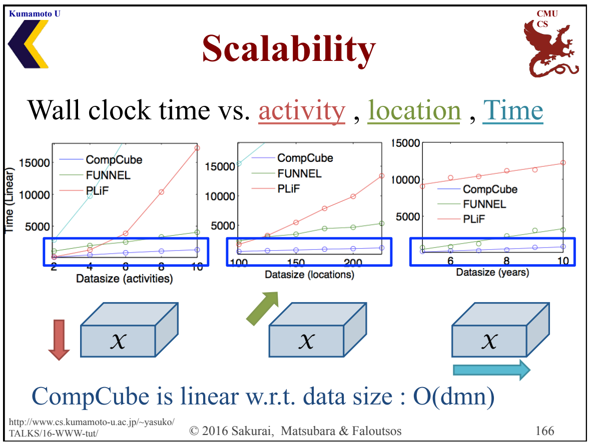

Kumamoto U
CMU
Effectiveness
(Details in paper)
5. Cars
Nikola Tesla's birthday
1Tesla
2Bugatti
3MINI
4Ford
5Jeep
6Toyota
7GMC
8Dodge
9Subaru
10Mazda
6. SNS
yelp
1Tumblr
2Facebook
3LinkedIn
4SoundCloud
5Google+
6Goodreads
7Wattpad
8Yelp
9MocoSpace
10deviantART
Time (weekly)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
161

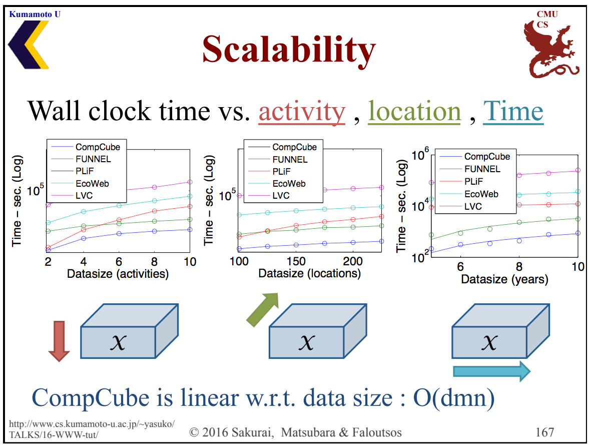

Kumamoto U
CMU
Effectiveness
(Details in paper)
7. Finance
WELLS FARGO
citi
CITIBANK
8. Software
HTML
1HTML
2Java
3SQL
4JavaScript
5PHP
6XML
7Visual_Basic
8ASP.NET
9DDL
10HTML5
Time (weekly)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
162

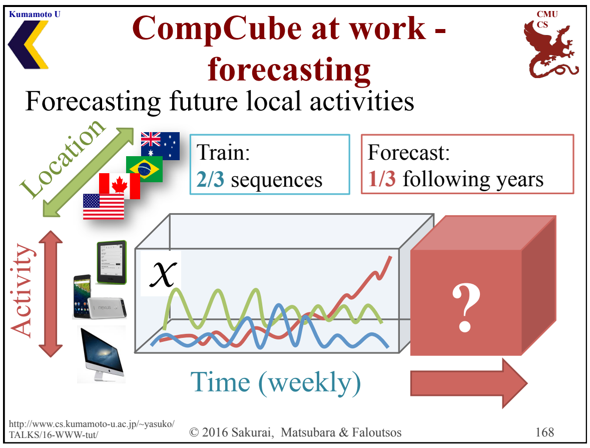

Kumamoto U
CMU CS
Accuracy
RMSE between original and fitted volume
CompCube EcoWeb LVC FUNNEL PLiF
Fitting error (RMSE)
0.25
0.2
0.15
0.1
0.05
0
#1 #2 #3 #4 #5 #6 #7 #8
Data ID
YAHOO! NEWS
WELLS FARGO
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
163


Kumamoto U
CMU CS
Accuracy
RMSE between original and fitted volume
CompCube EcoWeb LVC FUNNEL PLiF
Fitting error (RMSE)
0.25
0.2
0.15
0.1
0.05
0
#1 #2 #3 #4 #5 #6 #7 #8
Data ID
YAHOO! NEWS
WELLS FARGO
CompCube consistently wins!
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
164


Kumamoto U
CMU CS
Scalability
Wall clock time vs. activity , location , Time
CompCube
FUNNEL
PLiF
Time (Linear)
Datasize (activities)
Datasize (locations)
Datasize (years)
x
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
165

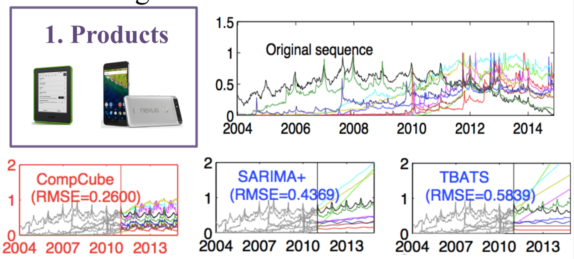

Kumamoto U
CMU CS
Scalability
Wall clock time vs. activity , location , Time
CompCube
FUNNEL
PLiF
Time (Linear)
Datasize (activities)
Datasize (locations)
Datasize (years)
x
CompCube is linear w.r.t. data size : O(dmn)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
166


Kumamoto U
CMU CS
Scalability
Wall clock time vs. activity , location , Time
CompCube
FUNNEL
PLiF
EcoWeb
LVC
Time – sec. (Log)
Datasize (activities)
Datasize (locations)
Datasize (years)
x
CompCube is linear w.r.t. data size : O(dmn)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
167


Kumamoto U
CMU CS
CompCube at work - forecasting
Forecasting future local activities
Location
Train:
2/3 sequences
Forecast:
1/3 following years
Activity
x
?
Time (weekly)
http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/16-WWW-tut/
© 2016 Sakurai, Matsubara & Faloutsos
168

CompCube at work - forecasting

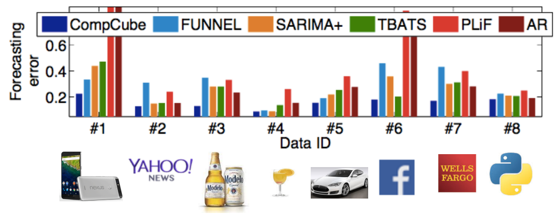
Forecasting results for #1 Products



CompCube captures future activities very well

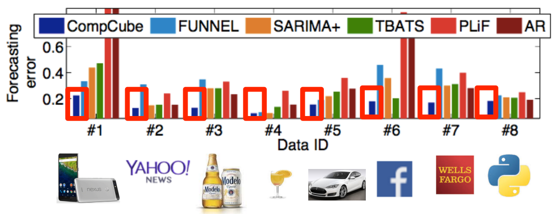
CompCube at work - forecasting

Forecasting error (original vs. forecasts)



CompCube at work - forecasting

Forecasting error (original vs. forecasts)

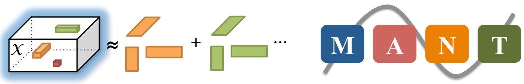


CompCube consistently wins!

Part 3 Conclusions

- Real data are often in high dimensions with multiple aspects (modes)
- Matrices and tensors provide elegant theory and algorithms
- MANT analysis

Multi-Aspect Non-linear Time-series



References

- Inderjit S. Dhillon, Subramanyam Mallela, Dharmendra S. Modha: Information-theoretic co-clustering. KDD 2003: 89-98
- T. G. Kolda, B. W. Bader and J. P. Kenny. *Higher-Order Web Link Analysis Using Multilinear Algebra*. In: ICDM 2005, Pages 242-249, November 2005.
- Jimeng Sun, Spiros Papadimitriou, Philip Yu. *Window-based Tensor Analysis on High-dimensional and Multi-aspect Streams*, Proc. of the Int. Conf. on Data Mining (ICDM), Hong Kong, China, Dec 2006

Part 3

Extension of time-series: tensor analysis

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