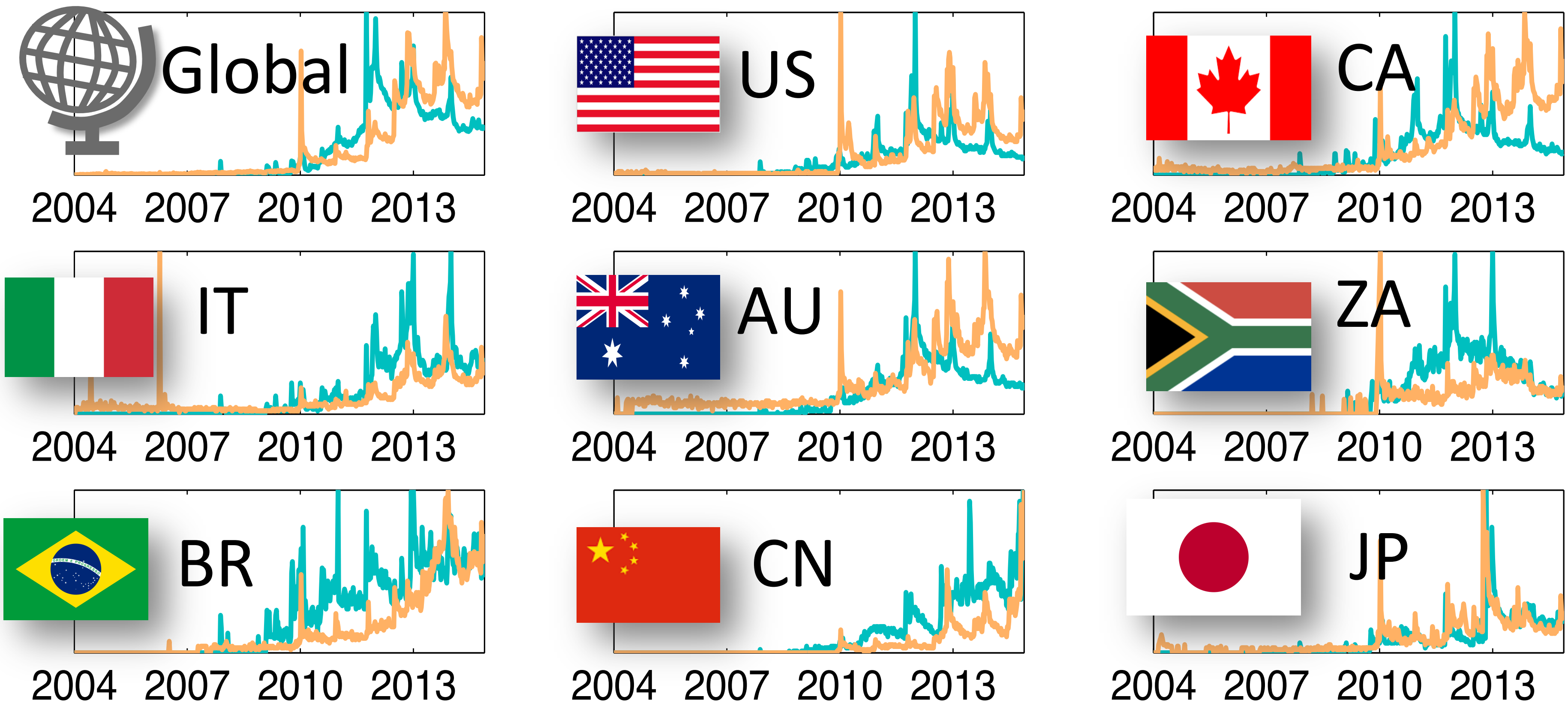
Non-linear Mining of Competing Local Activities

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

Given: local user activities

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

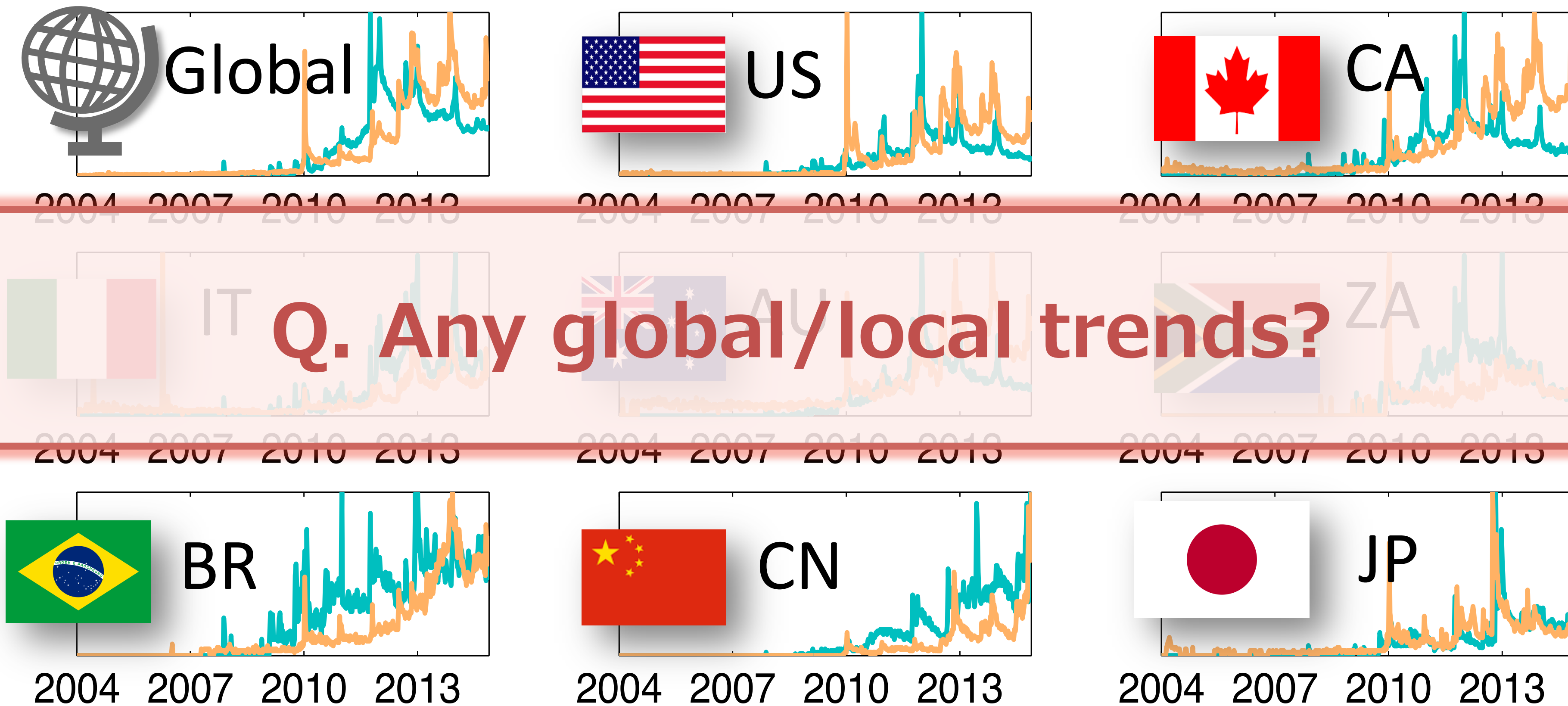
(for 236 countries, from 2004 to 2015)



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

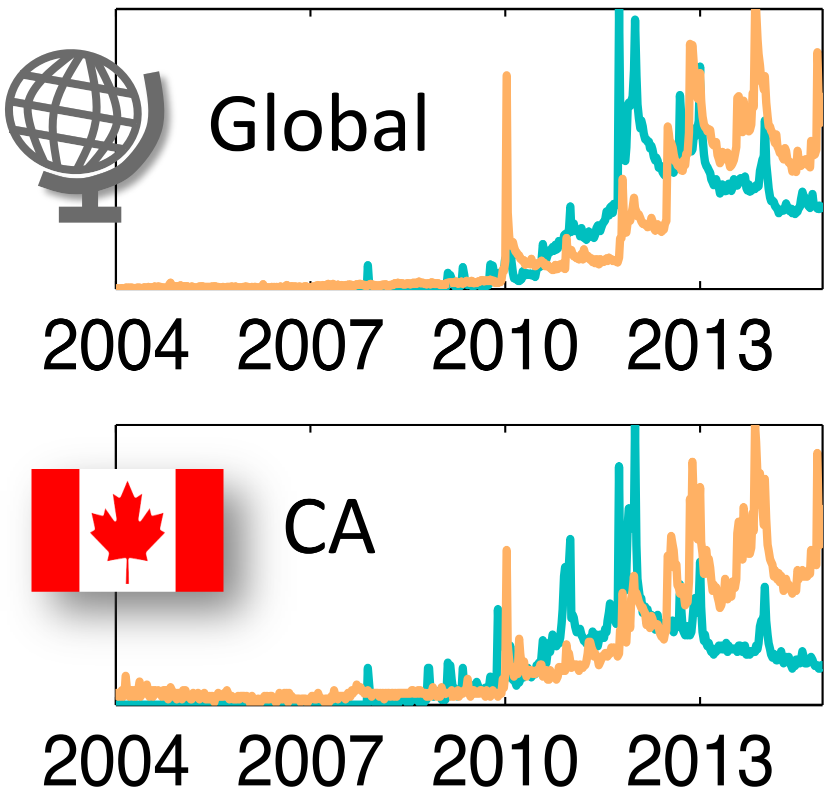
e.g., *Google* search volumes for **Kindle, Nexus**

(for 236 countries, from 2004 to 2015)

Given: local user activities

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

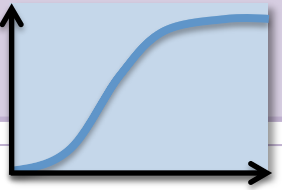
(for 236 countries, from 2004 to 2015)



Nexus

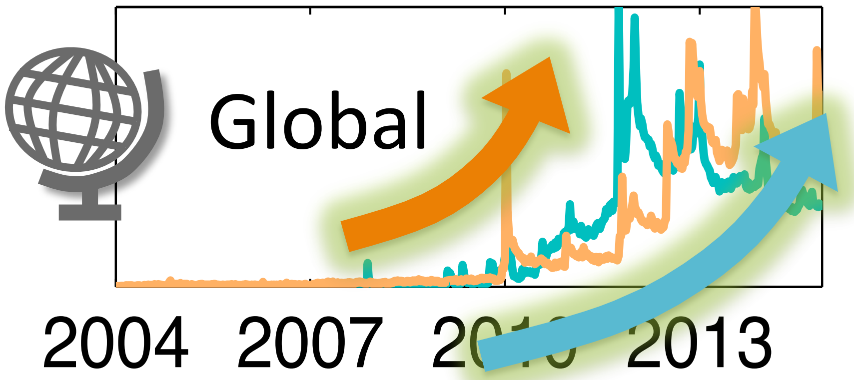
Kindle

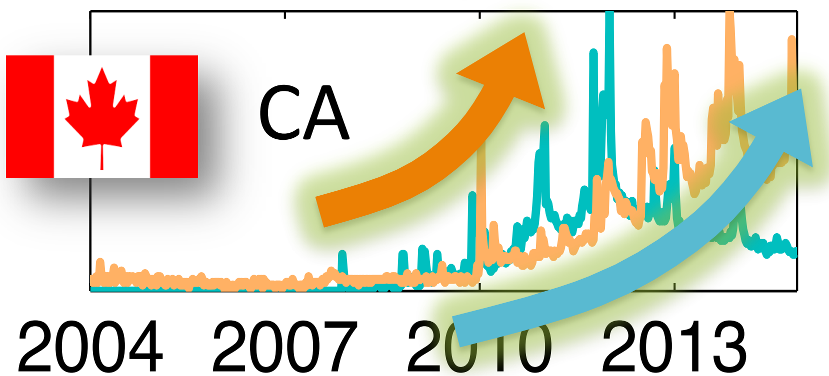
Given: local user activities



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

(for 236 countries, from 2004 to 2015)



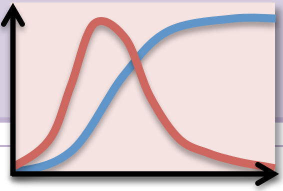


1. Exponential growth

Nexus

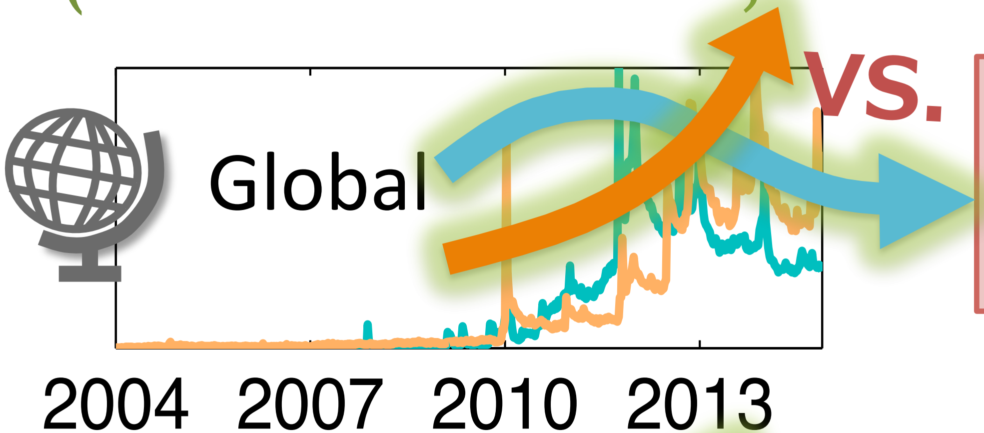
Kindle

Given: local user activities

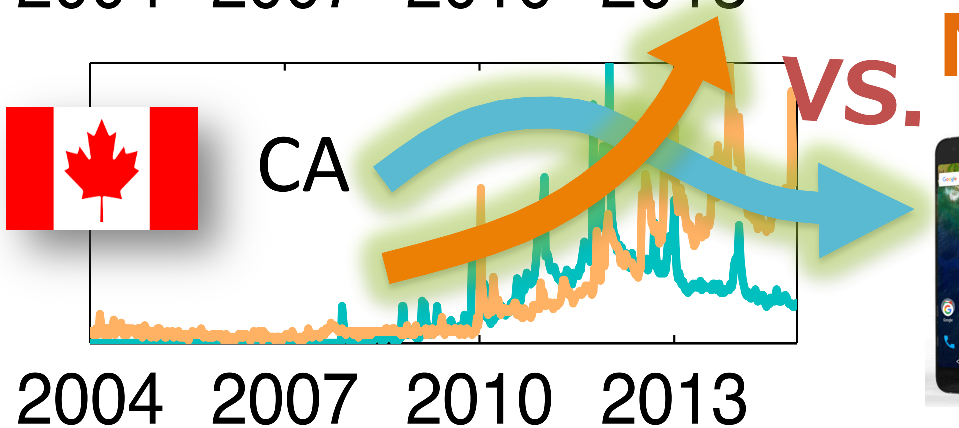


e.g., *Google* search volumes for **Kindle**, **Nexus**

(for 236 countries, from 2004 to 2015)



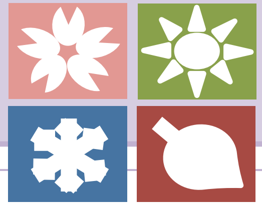
2. Competition



Nexus

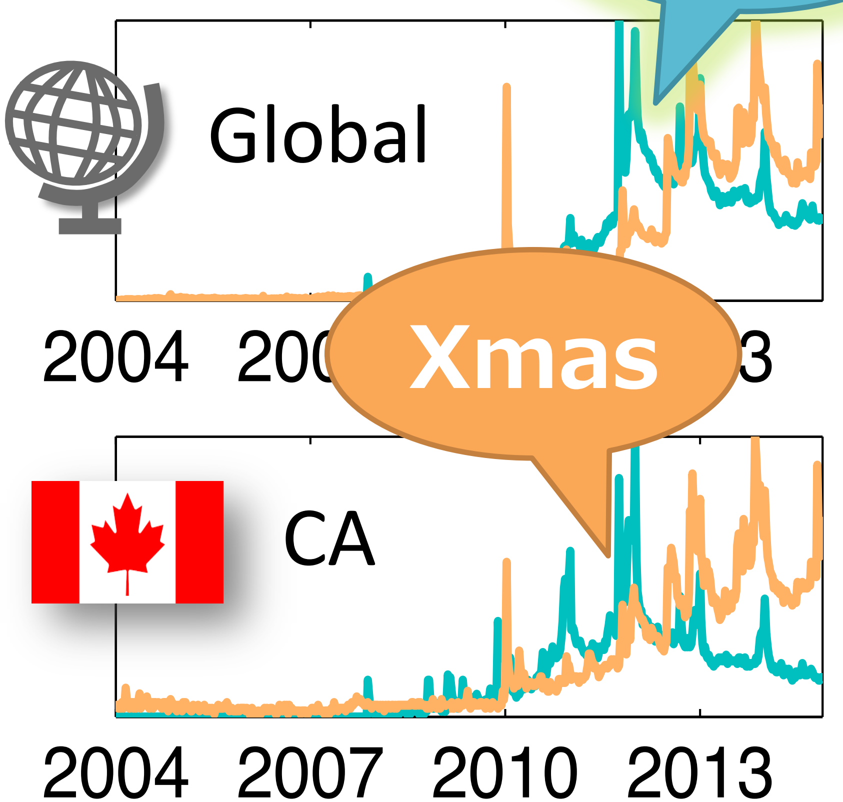
Kindle

Given: local user activities



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

(for 236 countries from 2004 to 2015)

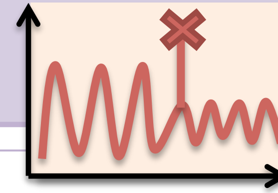


3. Seasonality

Nexus

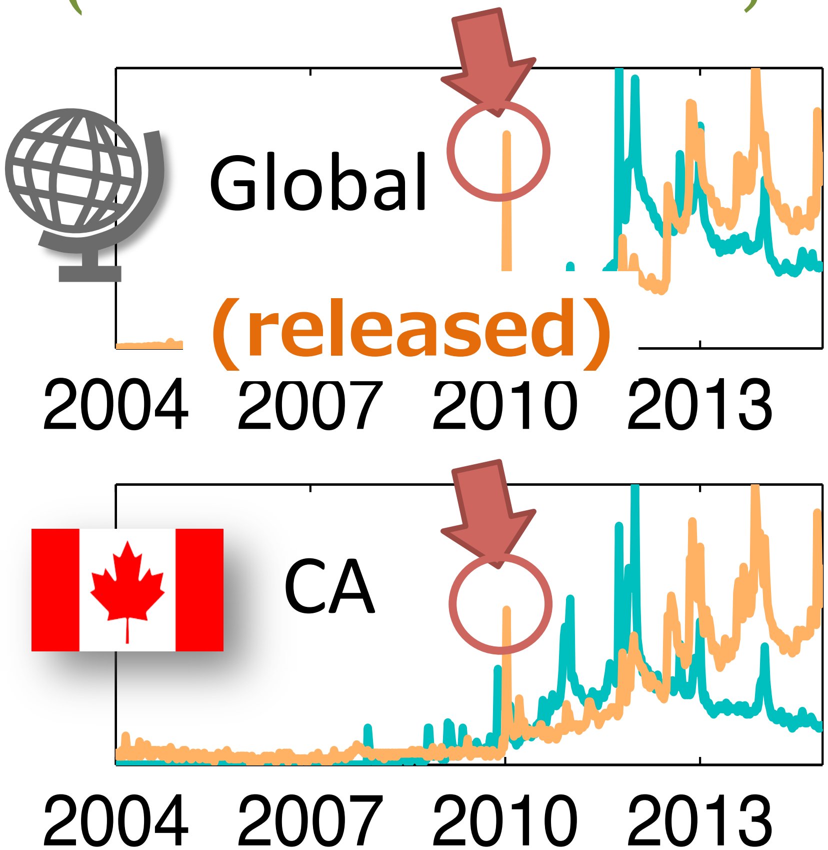
Kindle

Given: local user activities



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

(for 236 countries, from 2004 to 2015)



4. Deltas (outliers)

Given: local user activities

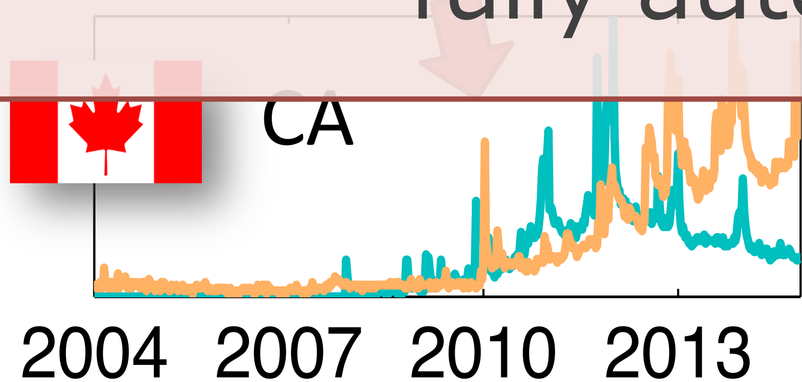
e.g., *Google* search volumes for **Kindle, Nexus**

(for 236 countries, from 2004 to 2015)

4. Deltas

(outliers)

Goal: find **global/local** patterns,
fully automatically



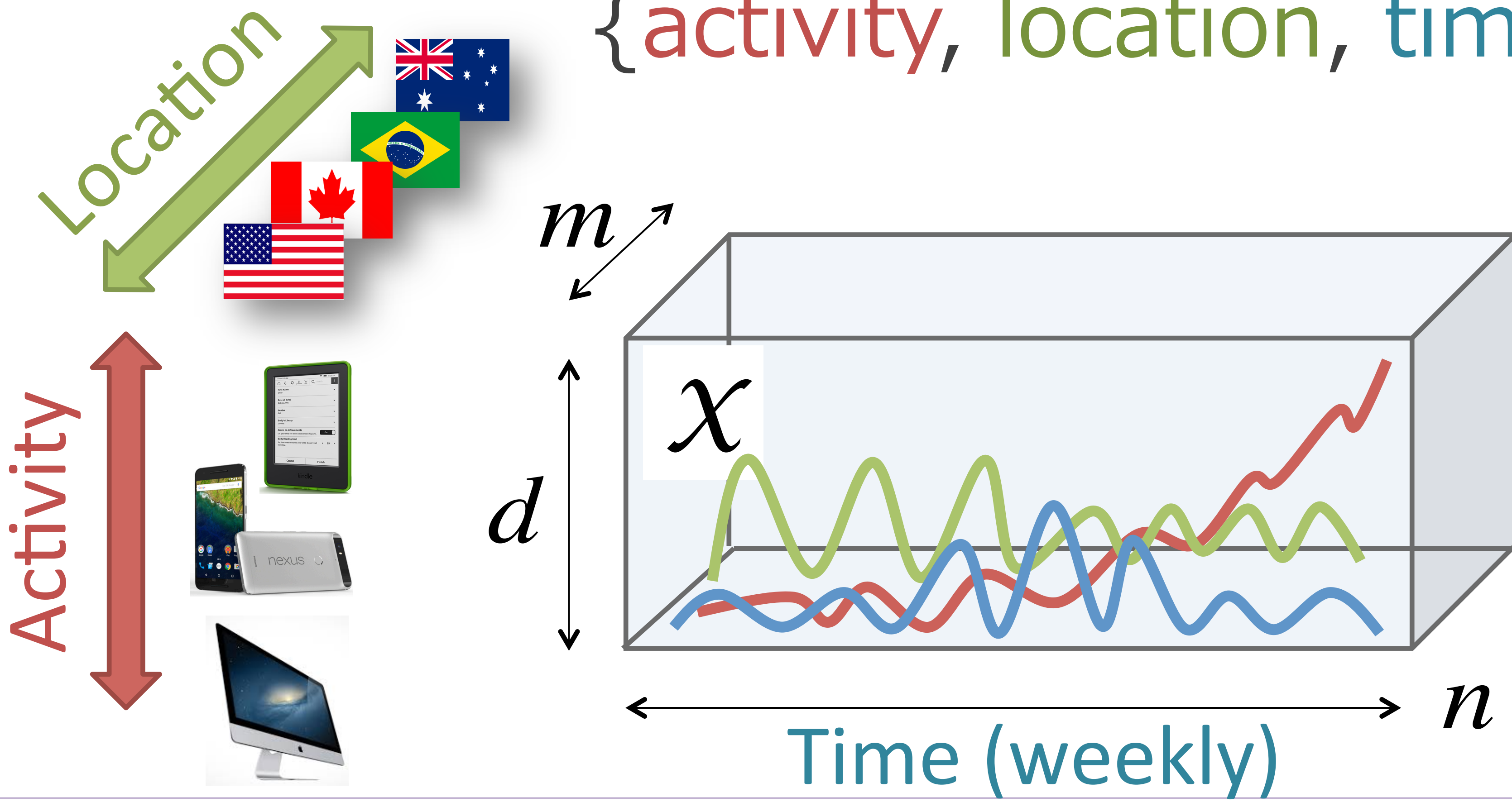
Nexus

Kindle

Data description

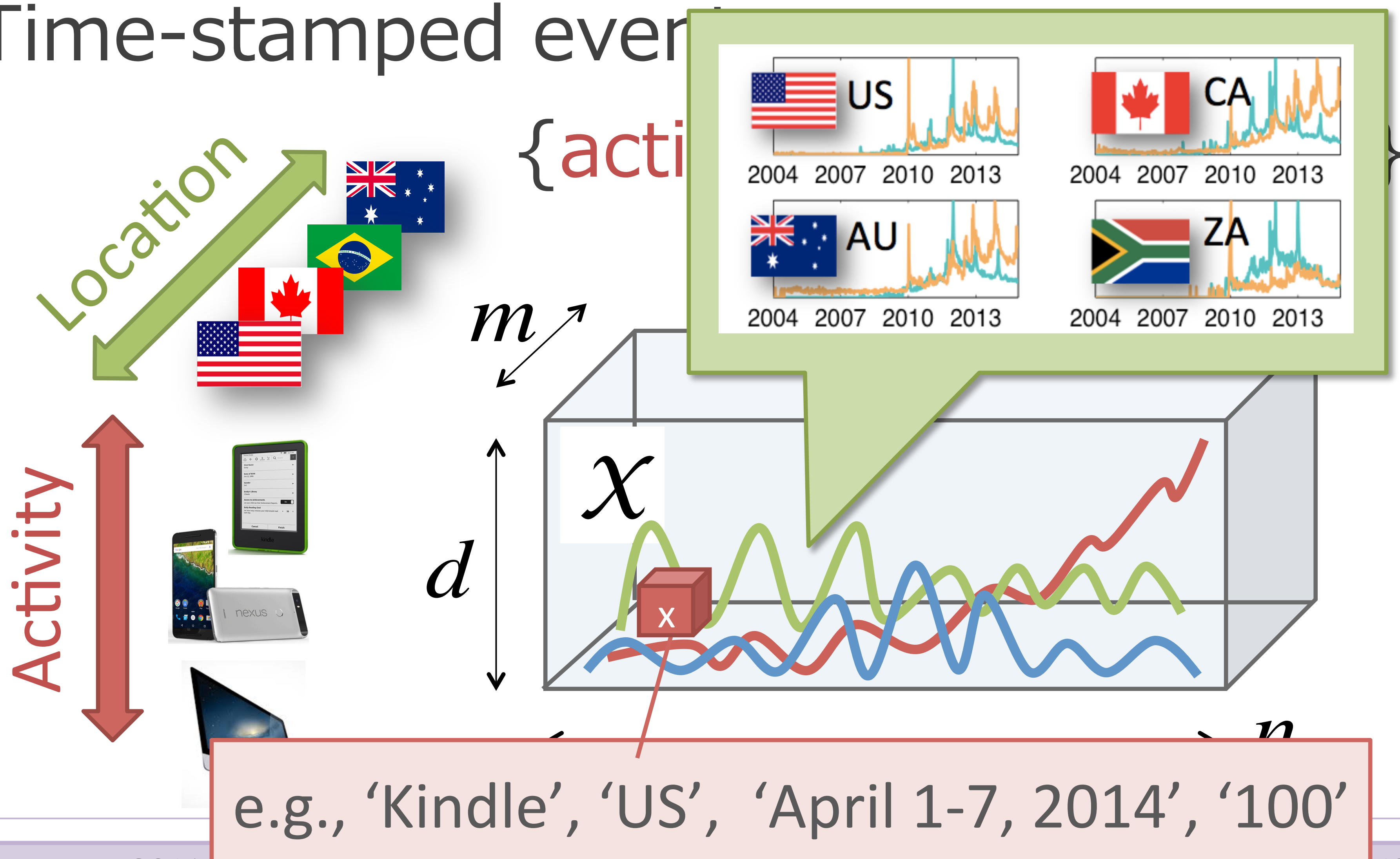
Time-stamped events:

{activity, location, time}



Data description

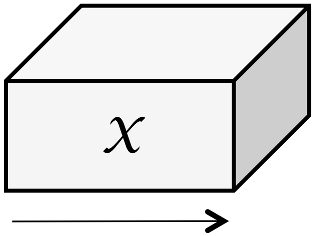
Time-stamped ever



Problem definition

Given: Tensor $\mathcal{X}$

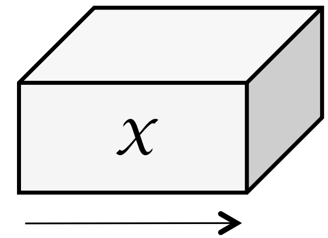
(activity x location x time)



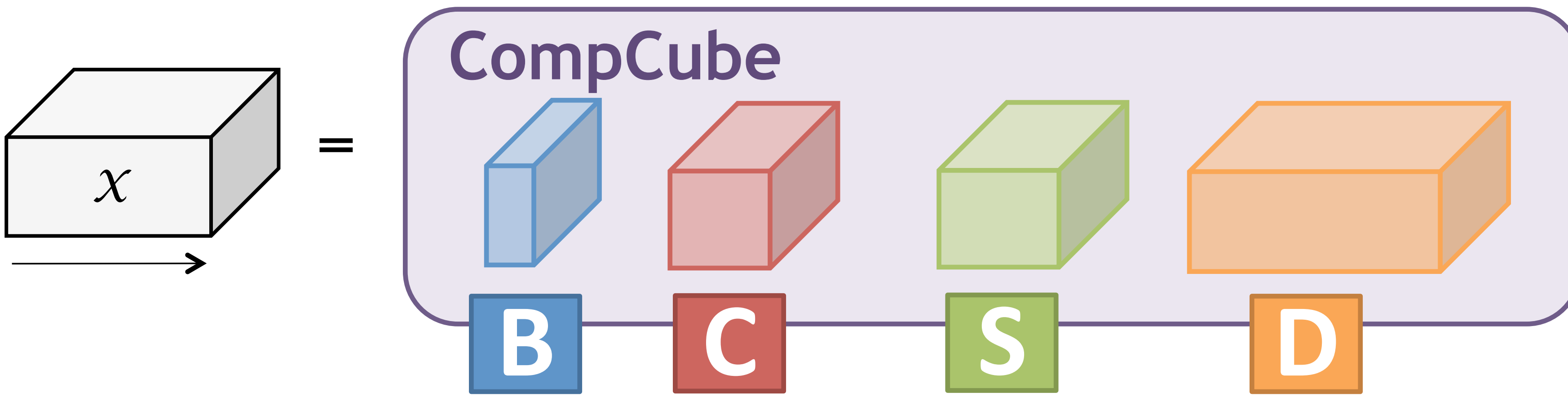
Problem definition

Given: Tensor $\mathcal{X}$

(activity x location x time)

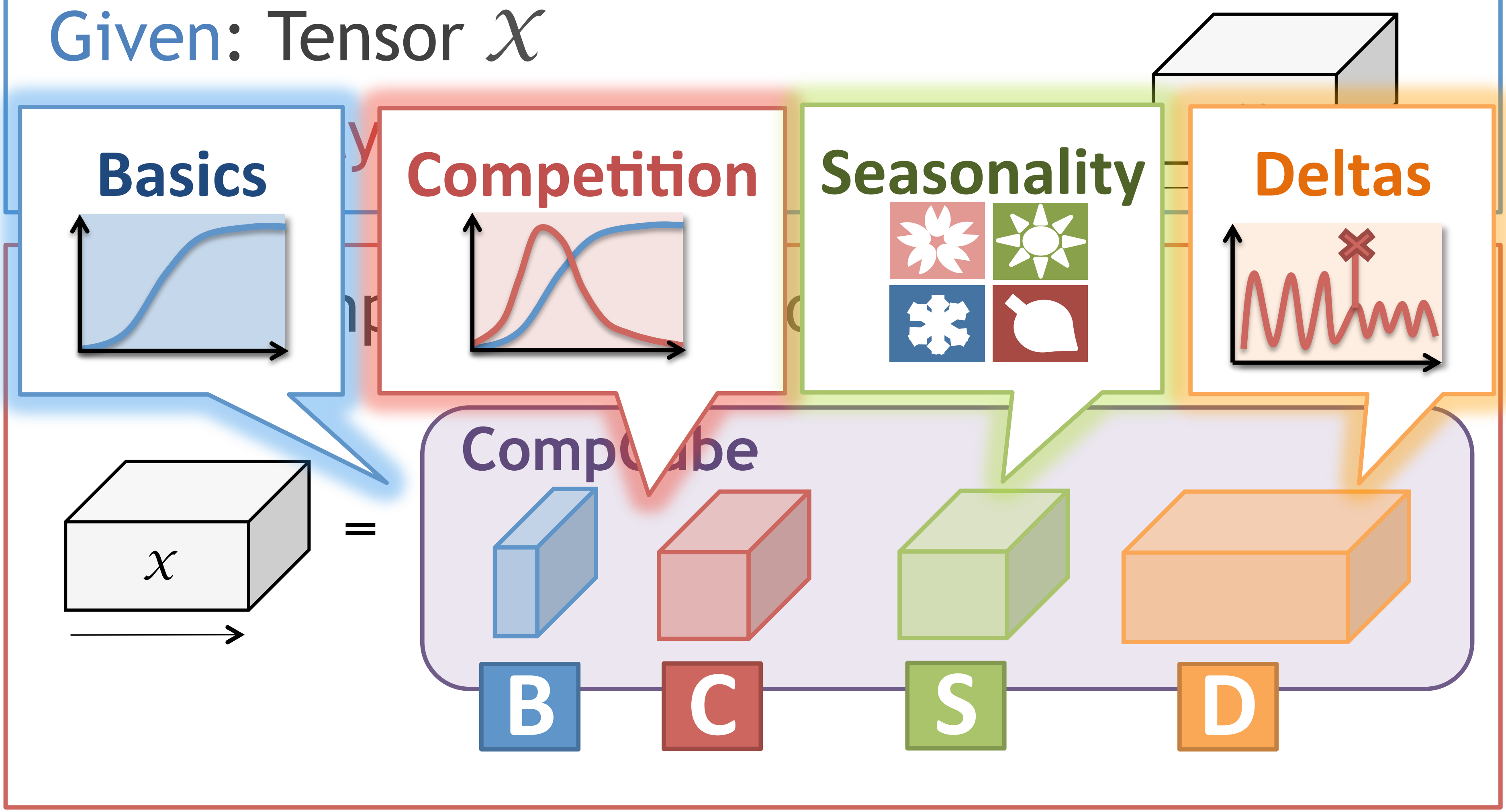


Find: Compact description of $\mathcal{X}$

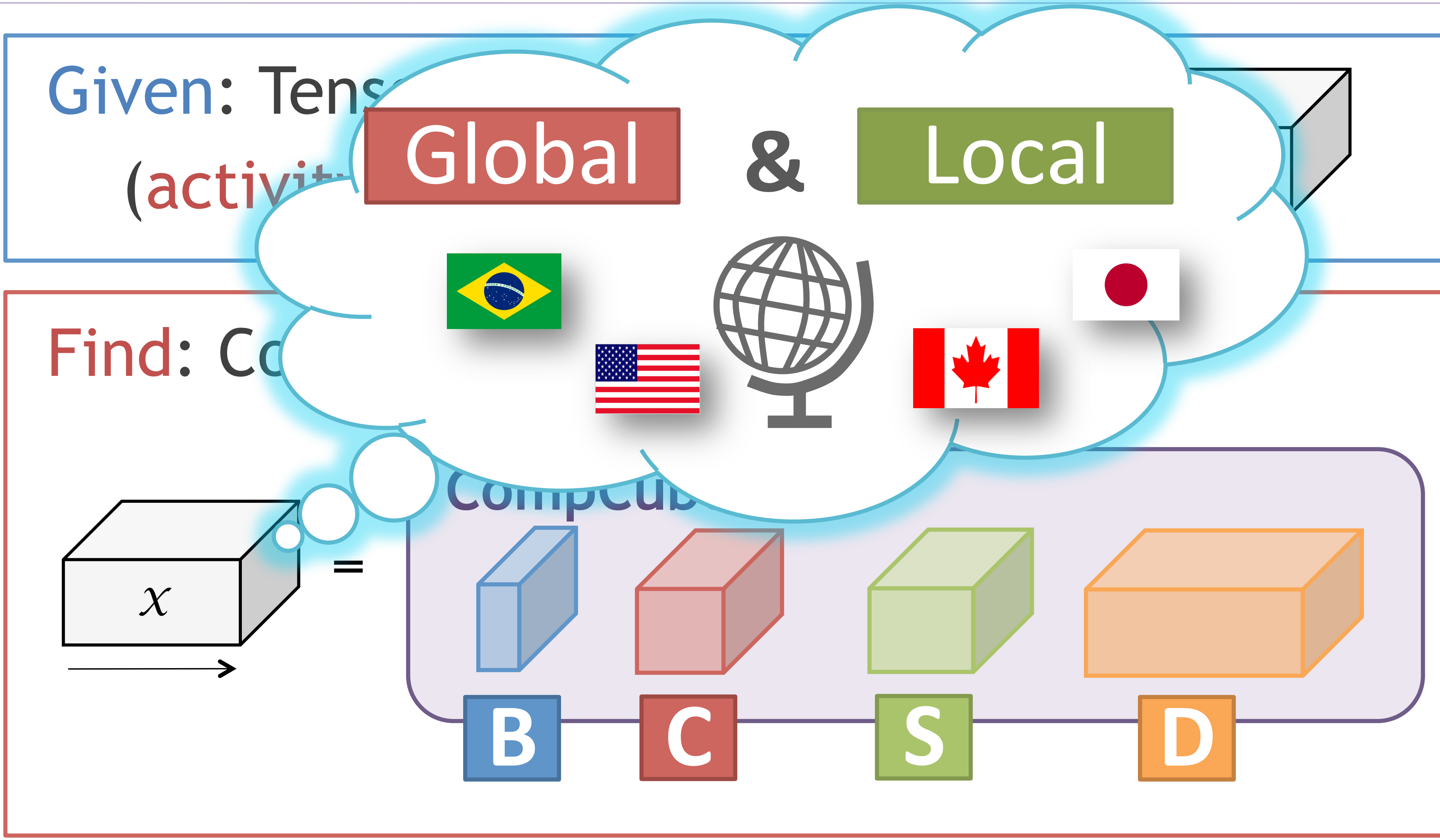


Problem definition

Given: Tensor $\mathcal{X}$



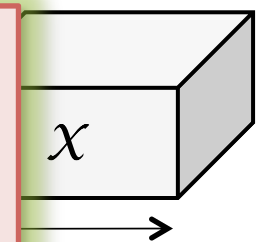
Problem definition



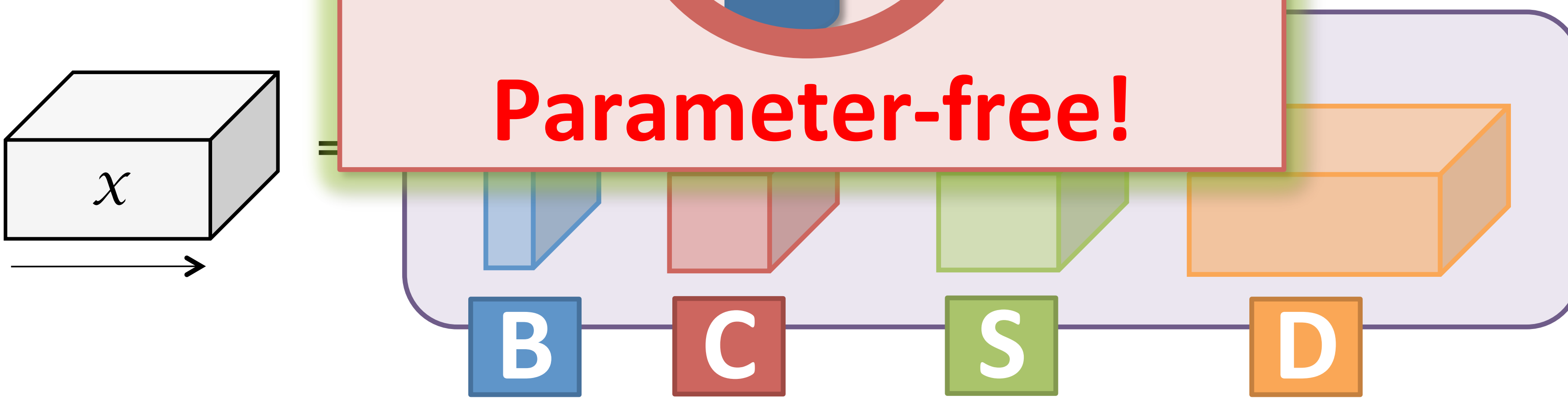
Problem definition

Given: Ten

(activity



Find: Com

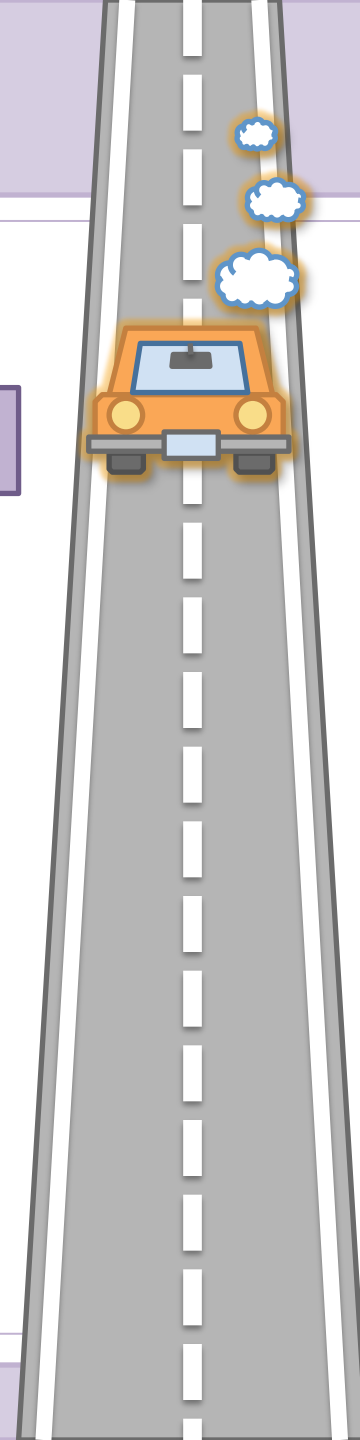


NO magic numbers !

Parameter-free!

Roadmap

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



Modeling power of CompCube

Products

News sources

Modeling power of CompCube

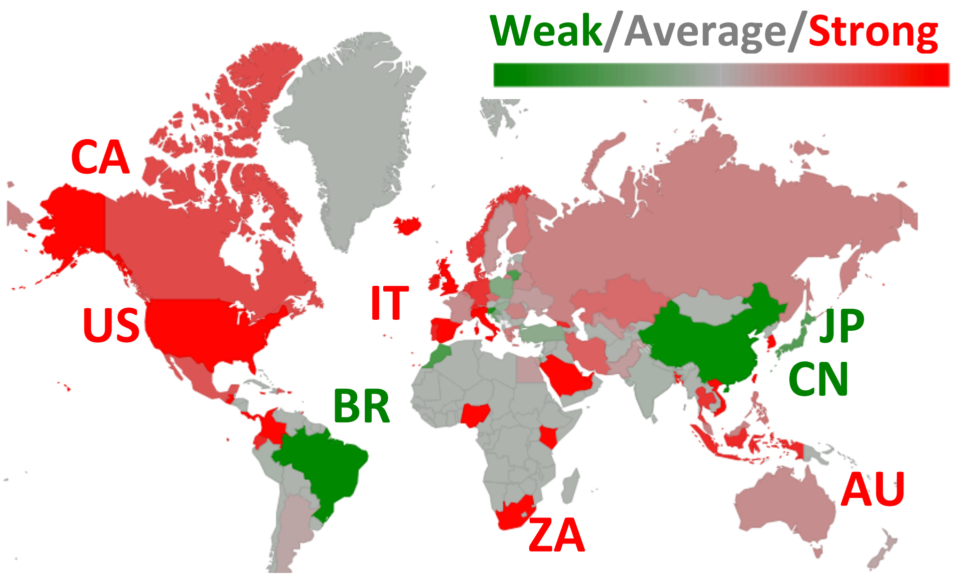
Products

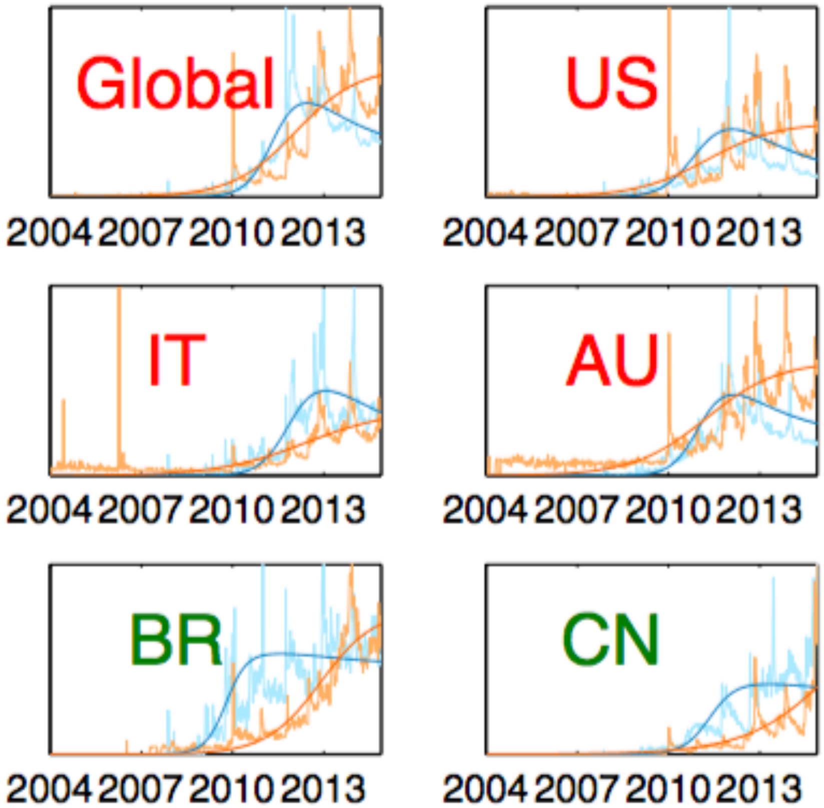
News sources

Modeling power of CompCube

Modeling power of CompCube

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

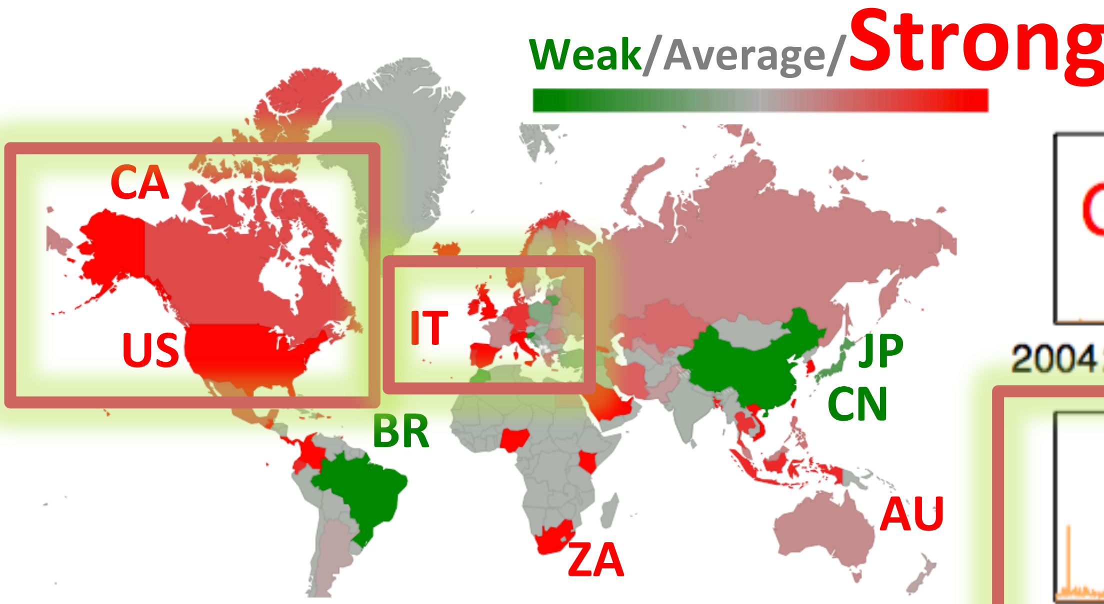


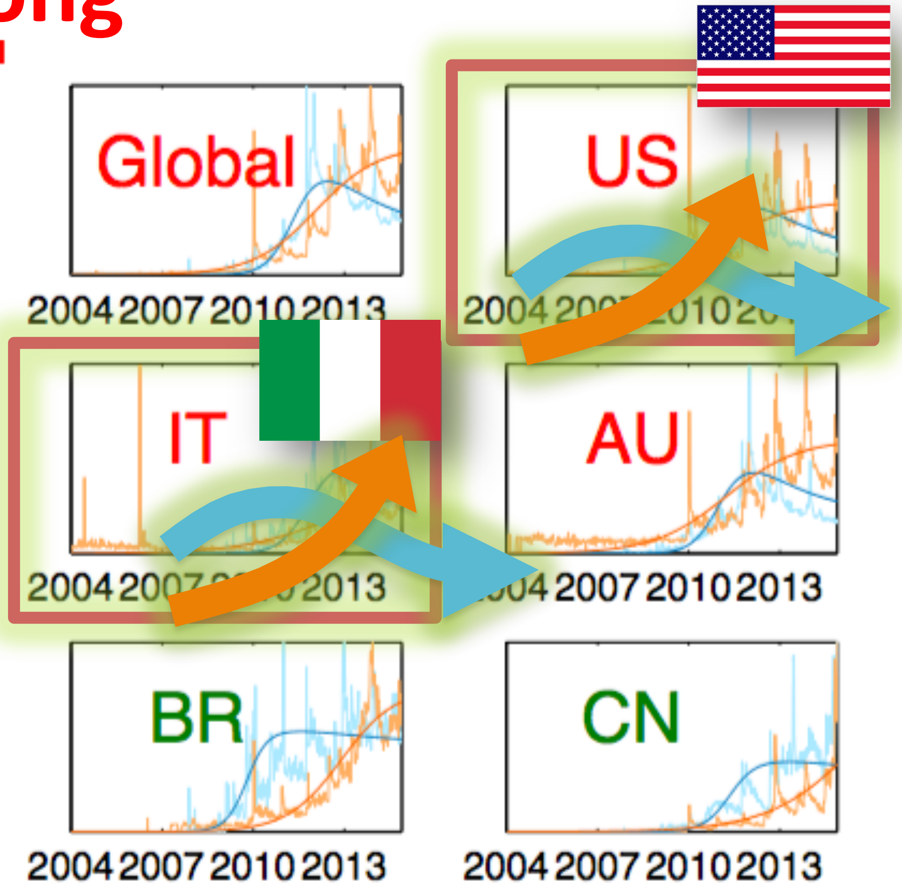


Local Competition strength

Modeling power of CompCube

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

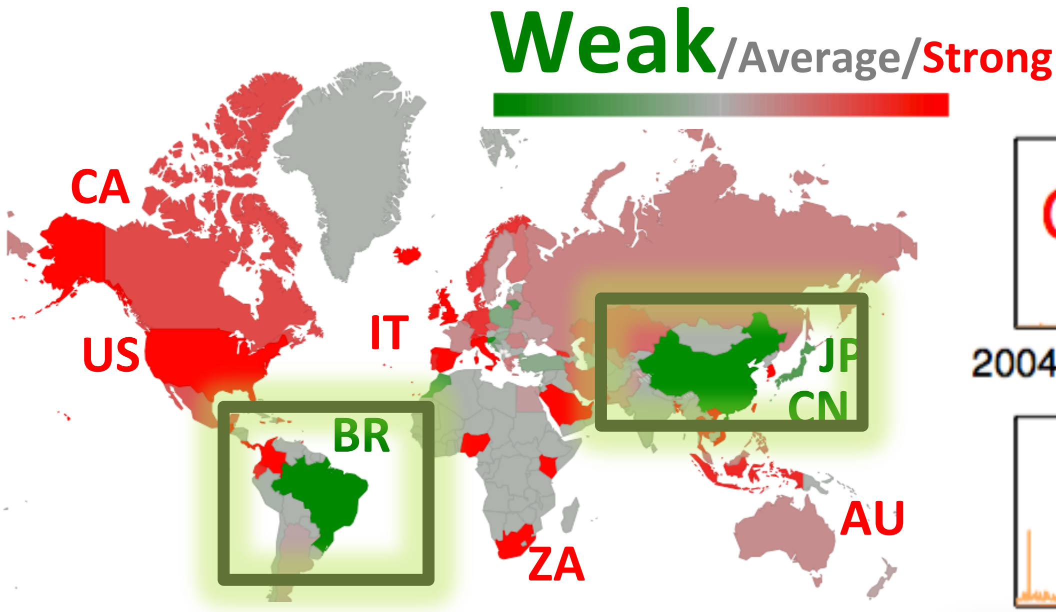




Local Competition strength

Modeling power of CompCube

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



Local Competition strength

Modeling power of CompCube

Products

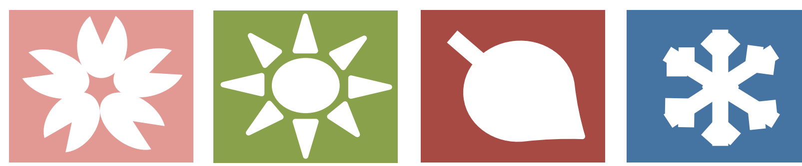
News sources

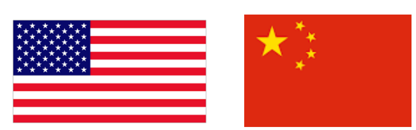
Modeling power of CompCube

Products

News sources

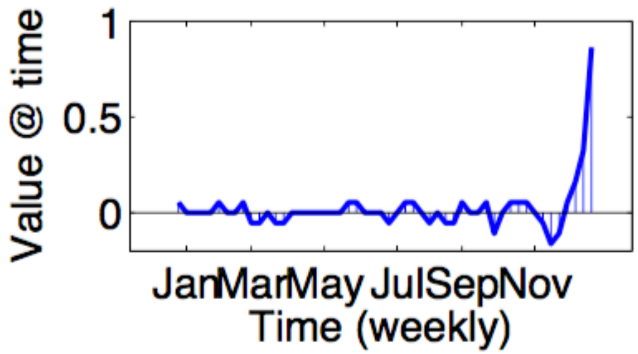
Q. Any seasonality?



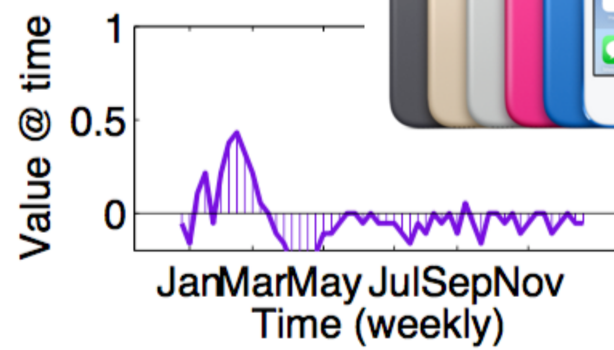
in 

Modeling power of CompCube

e.g., Local seasonality for **iPod**



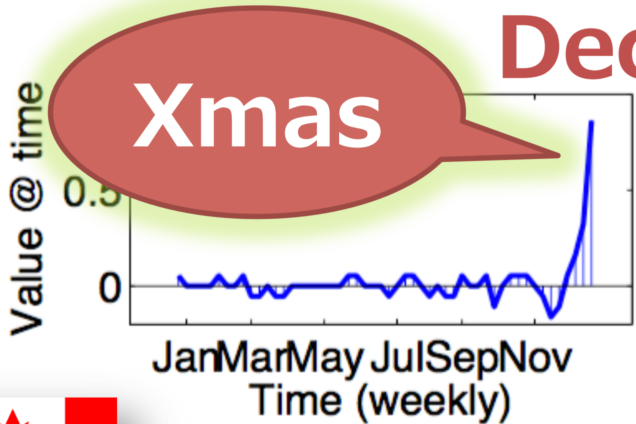
Component #1

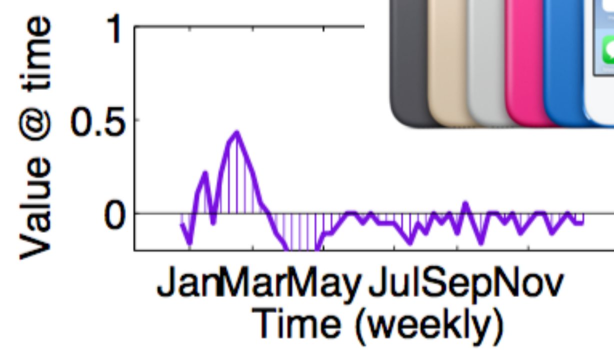


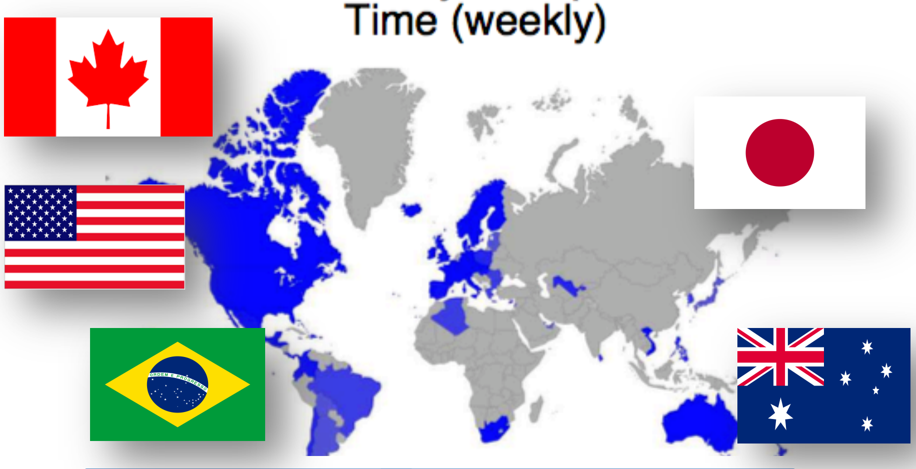
Component #2

Modeling power of CompCube

e.g., Local seasonality for **iPod**





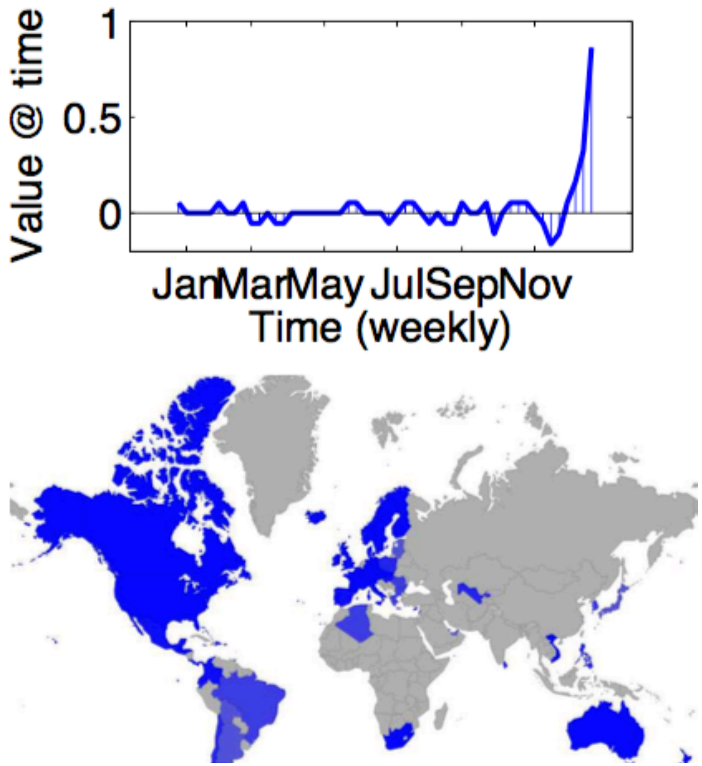


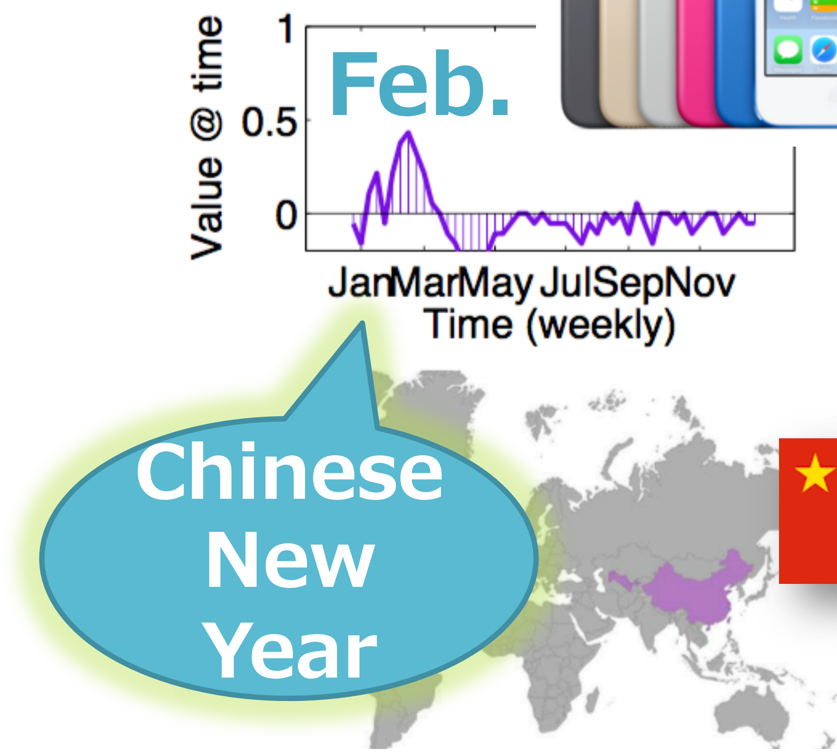
Component #1

Component #2

Modeling power of CompCube

e.g., Local seasonality for **iPod**





Component #1

Component #2

Modeling power of CompCube

Products

News sources

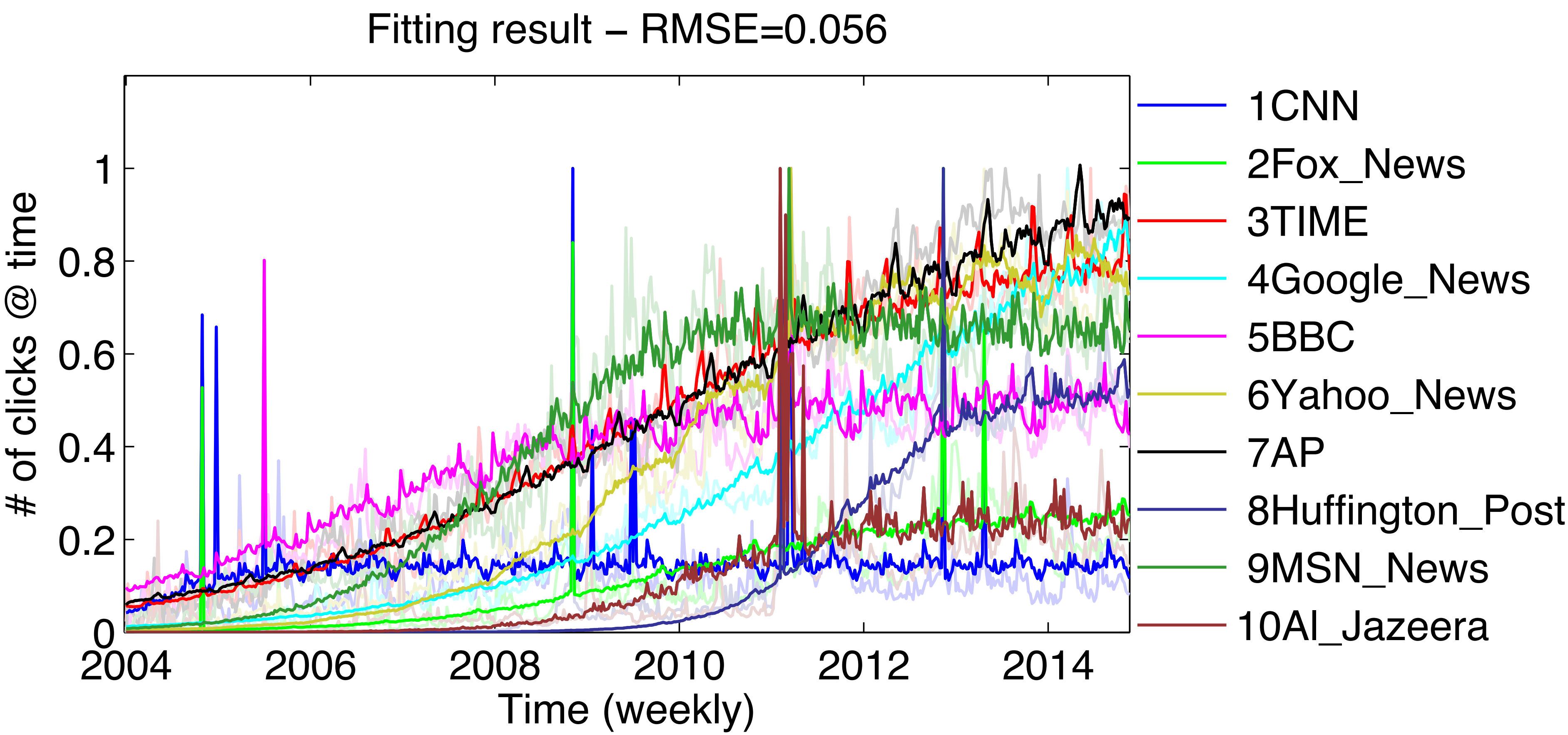
Modeling power of CompCube

Products

News sources

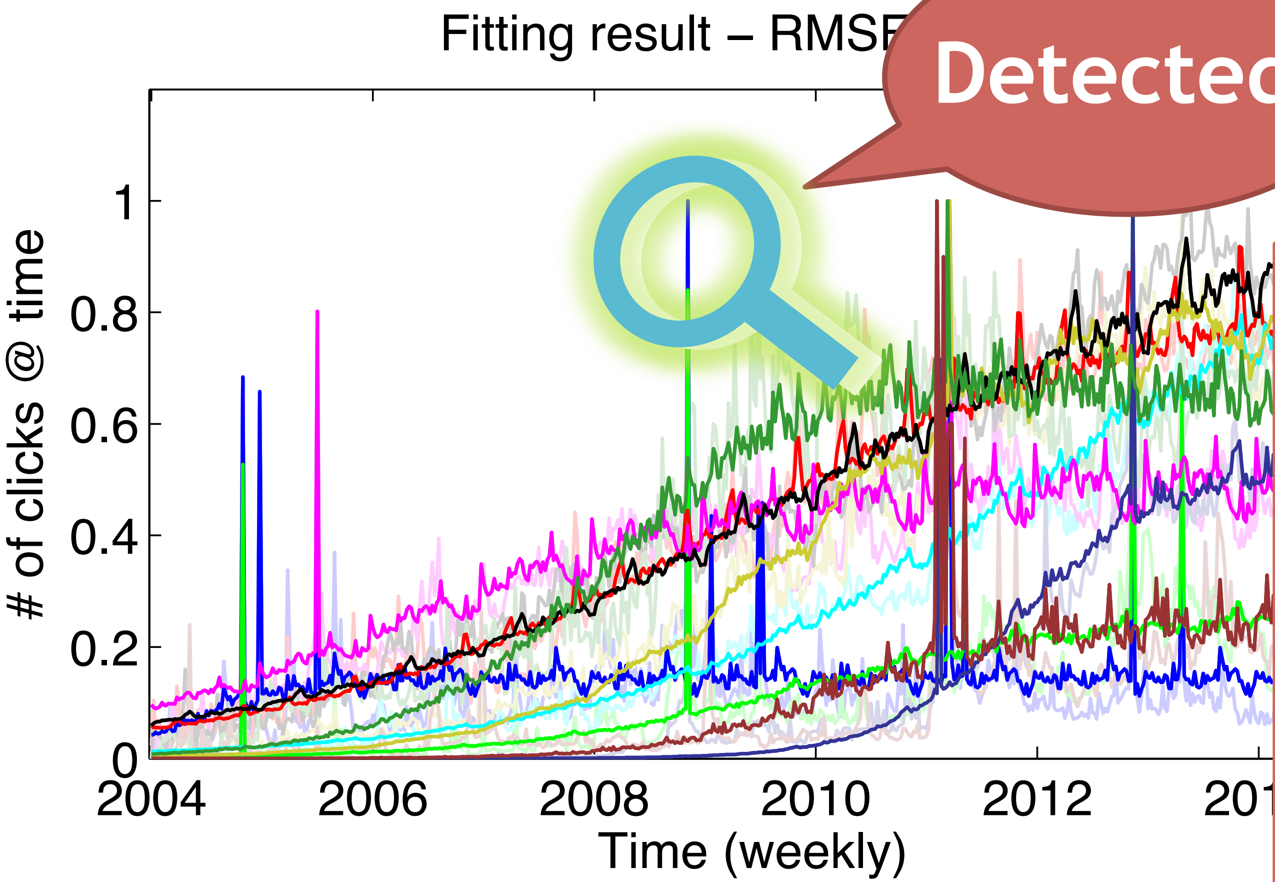
Modeling power of CompCube

Fitting result for **News resources**



Modeling power of CompCube

Fitting result for **News resources**

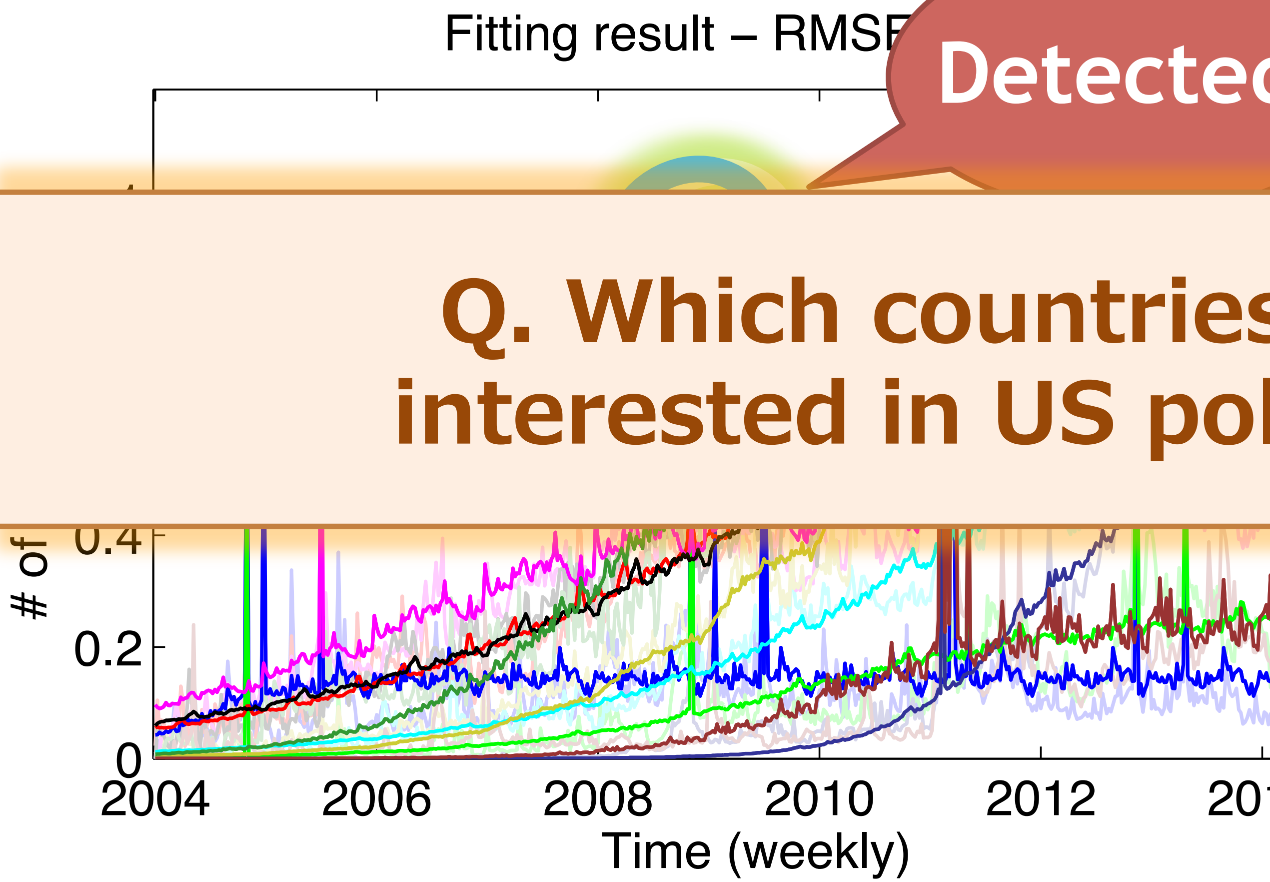


US election Nov. 2008

Wikipedia

Modeling power of CompCube

Fitting result for **News resources**



Detected!

Q. Which countries are interested in US politics?

Modeling power of CompCube

Fitting result for **News resources**

Weak/Strong

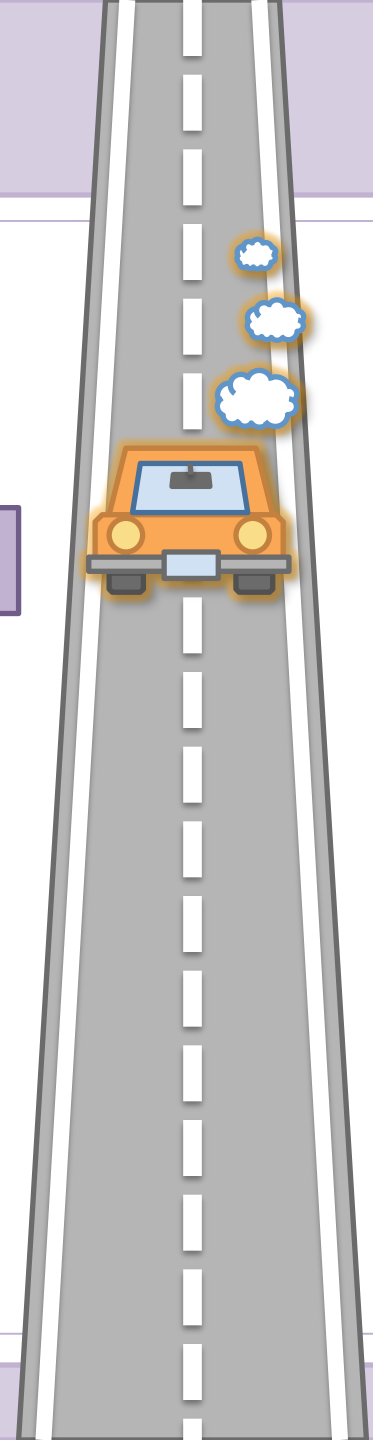
Local attention to US election

US election
Nov. 2008

Wikipedia

Roadmap

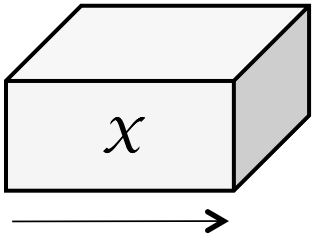
- ✔ Motivation
- ✔ Modeling power of CompCube
  - Overview
  - Proposed model
  - Algorithm
  - Experiments
  - CompCube – at work
  - Conclusions



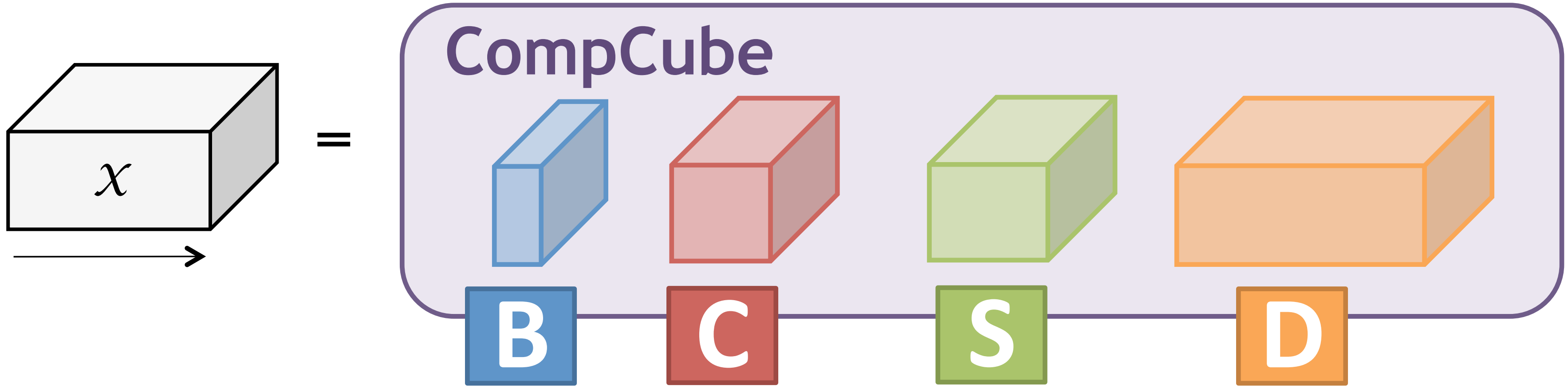
Problem definition

Given: Tensor $\mathcal{X}$

(activity x location x time)

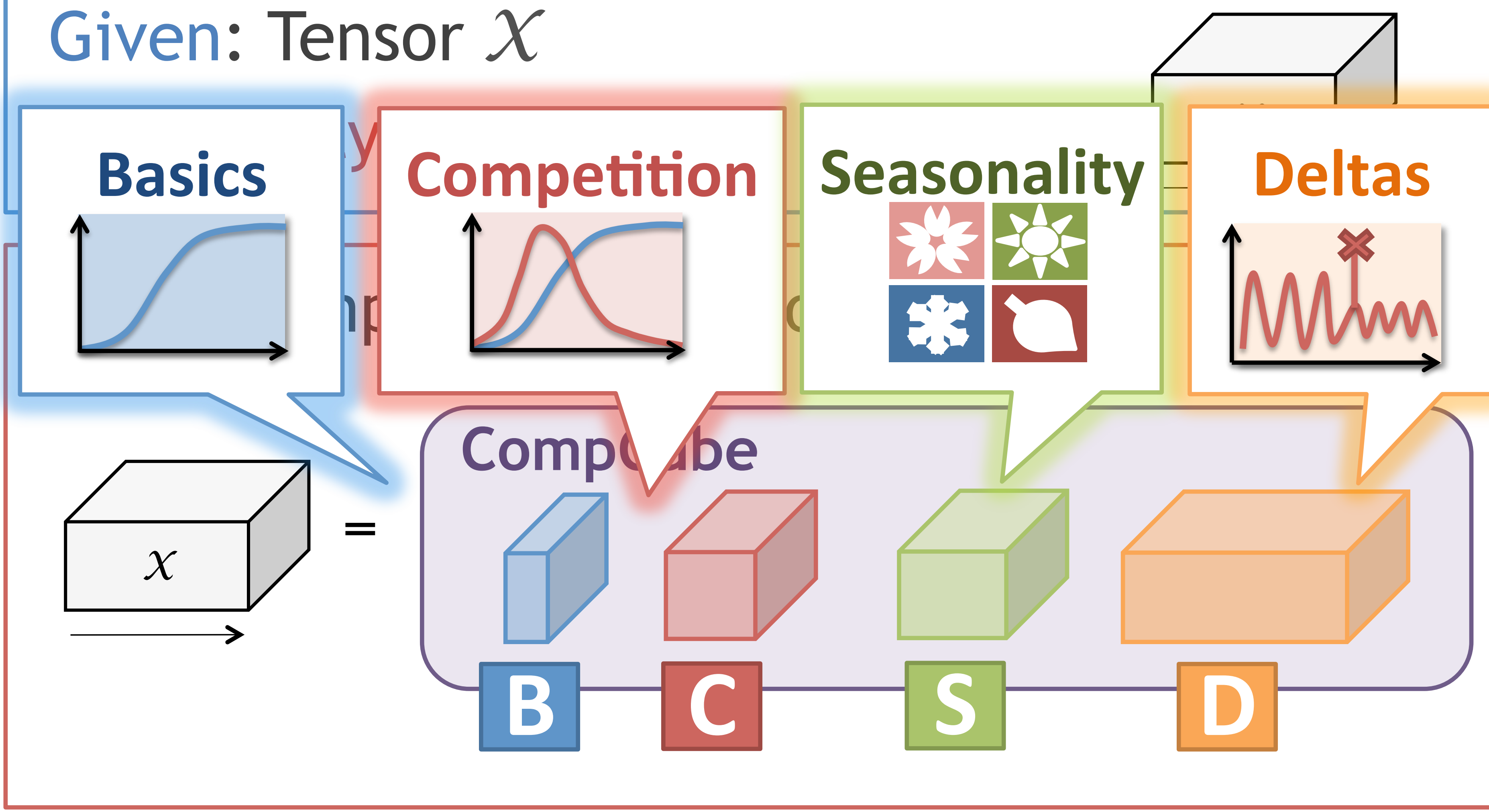


Find: Compact description of $\mathcal{X}$



Problem definition

Given: Tensor $\mathcal{X}$

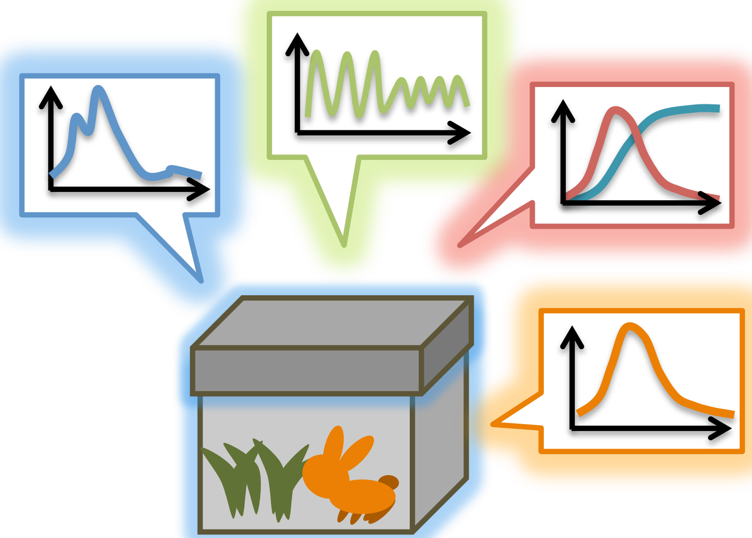


Main ideas: MANT analysis

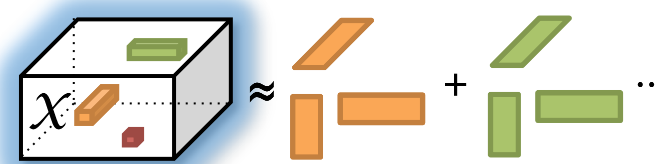
MANT analysis

Multi-Aspect Non-linear Time-series

#1 Non-linear models



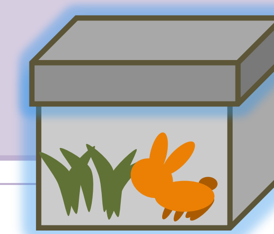
#2 Tensor analysis



#3 Automatic mining

NO magic numbers!

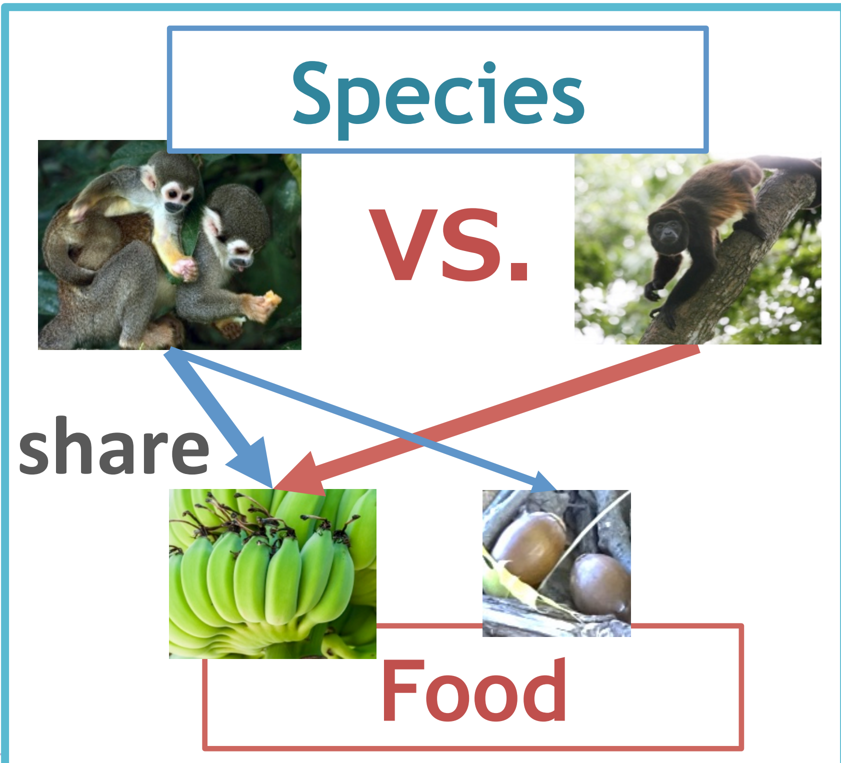
Main ideas: MANT analysis



Idea #1: Non-linear modeling

Virtual ecosystem on the Web

Ecosystem in the Jungle



Ecosystem on the Web

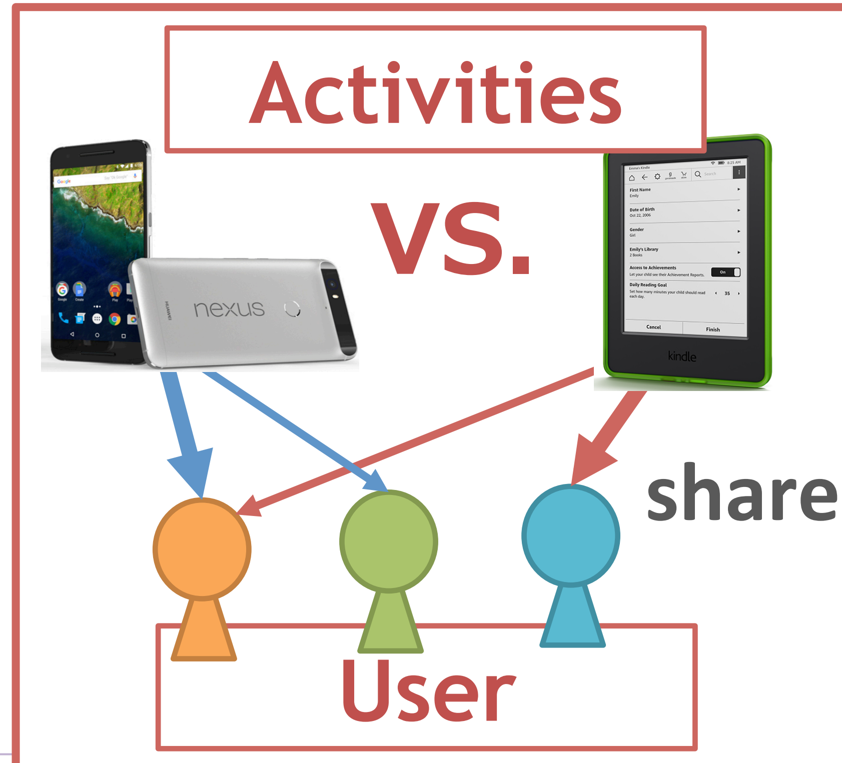
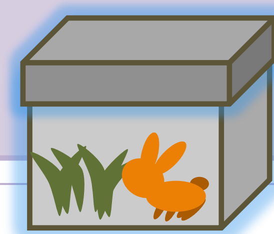


Image courtesy of xura, criminalatt, David Castillo Dominici, happykanppy at FreeDigitalPhotos.net.

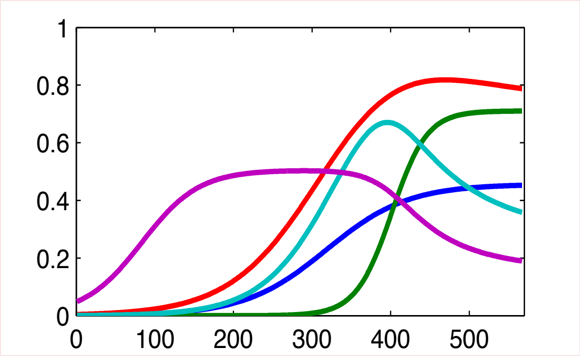
Main ideas: MANT analysis



Idea #1: Non-linear modeling

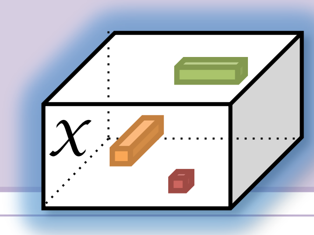
Virtual ecosystem on the Web

Non-linear dynamical system

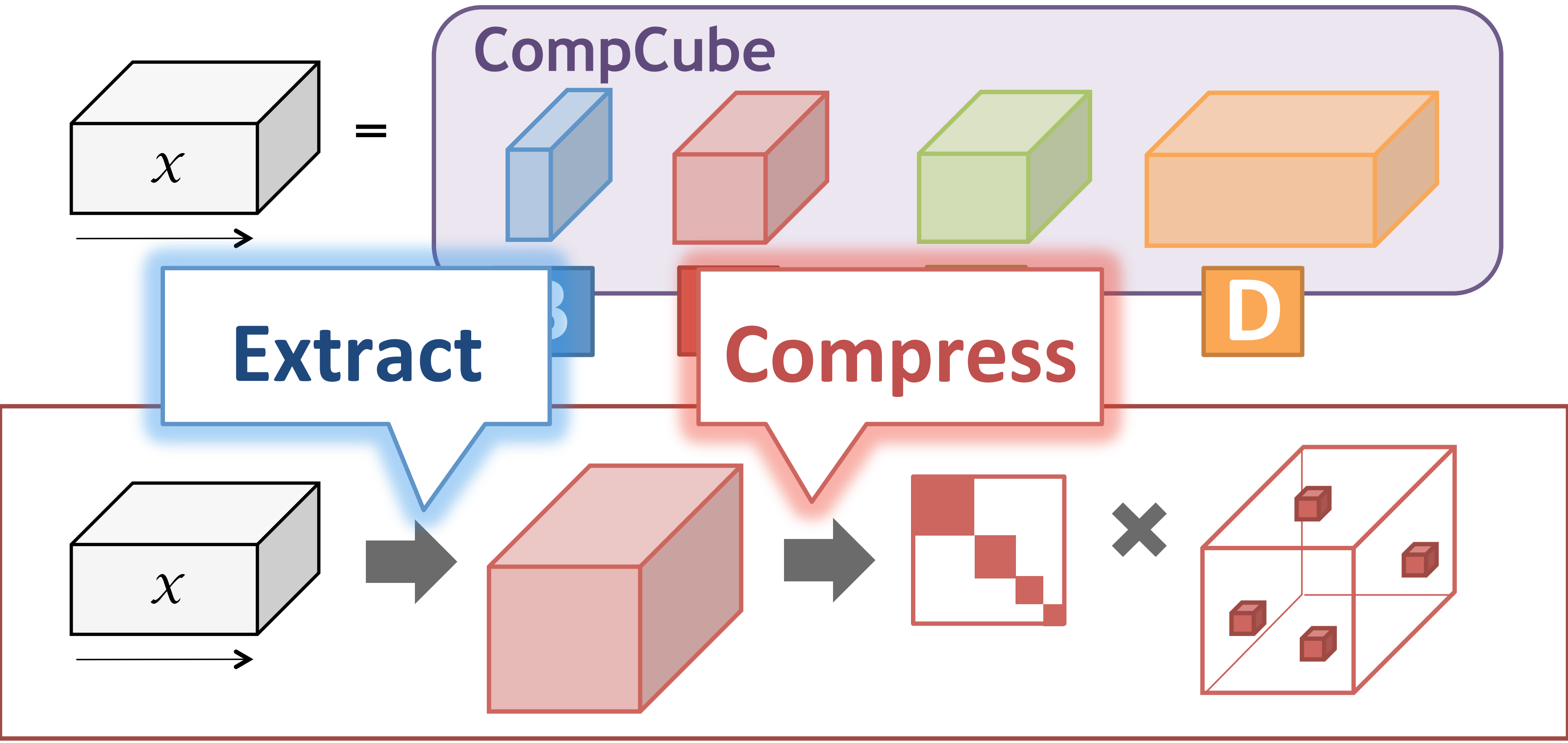


$$P_{il}(t) = P_{il}(t-1)\left[1 + r_{il}\left(1 - \frac{\sum_{j=1}^{d} c_{ijl} \cdot P_{jl}(t-1)}{K_{il}}\right)\right]$$

Main ideas: MANT analysis

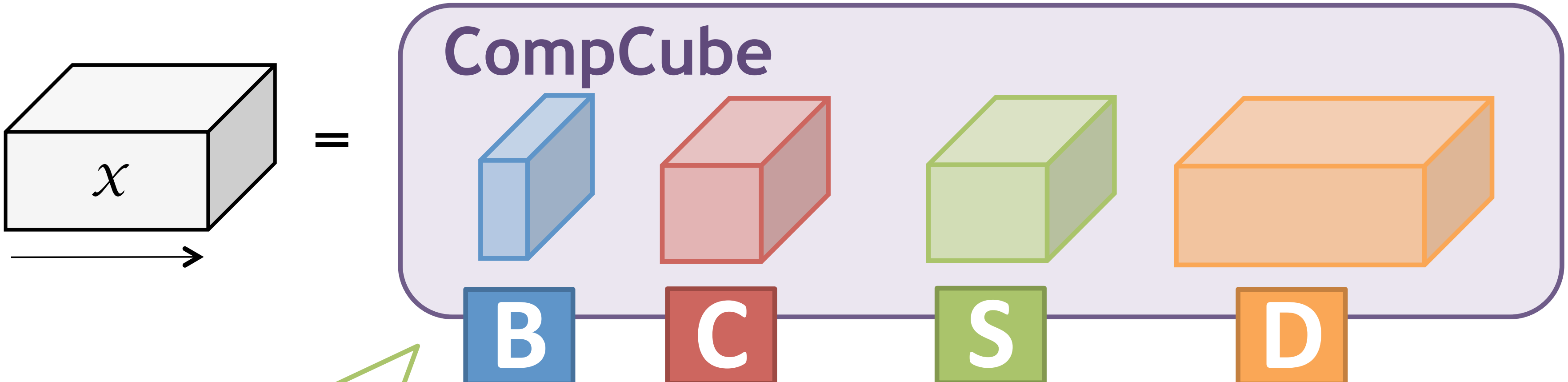


Idea #2: Tensor analysis



Main ideas: MANT analysis

Idea #3: MDL for fitting: parameter-free



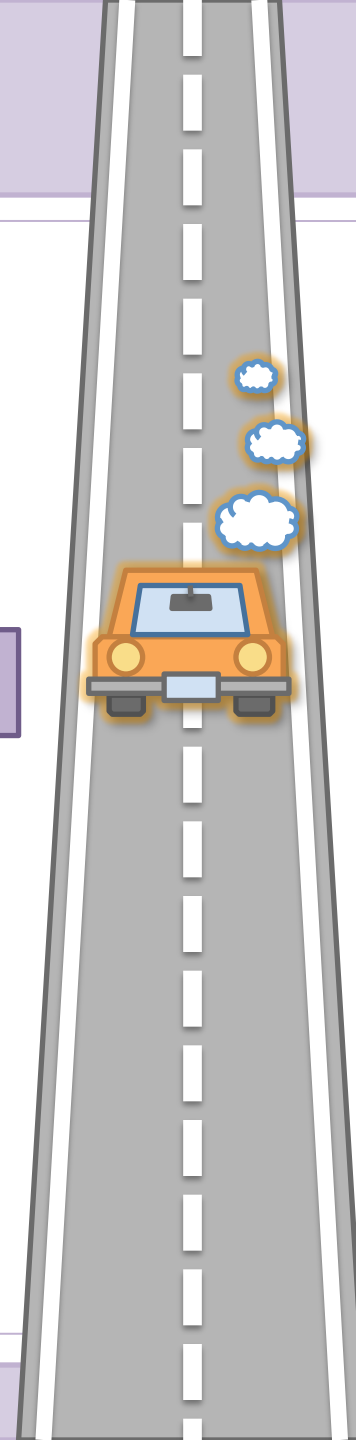
$$Cost_T(\mathcal{X};\mathcal{M}') =$$
$$Cost_M(\mathcal{M}') + Cost_C(\mathcal{X}|\mathcal{M}')$$
$$\mathcal{M}' = \{\mathbf{B}, \mathcal{B}', \mathbf{C}, \mathcal{C}', \mathbf{S}, \mathcal{W}', \mathcal{D}'\}$$

NO magic numbers

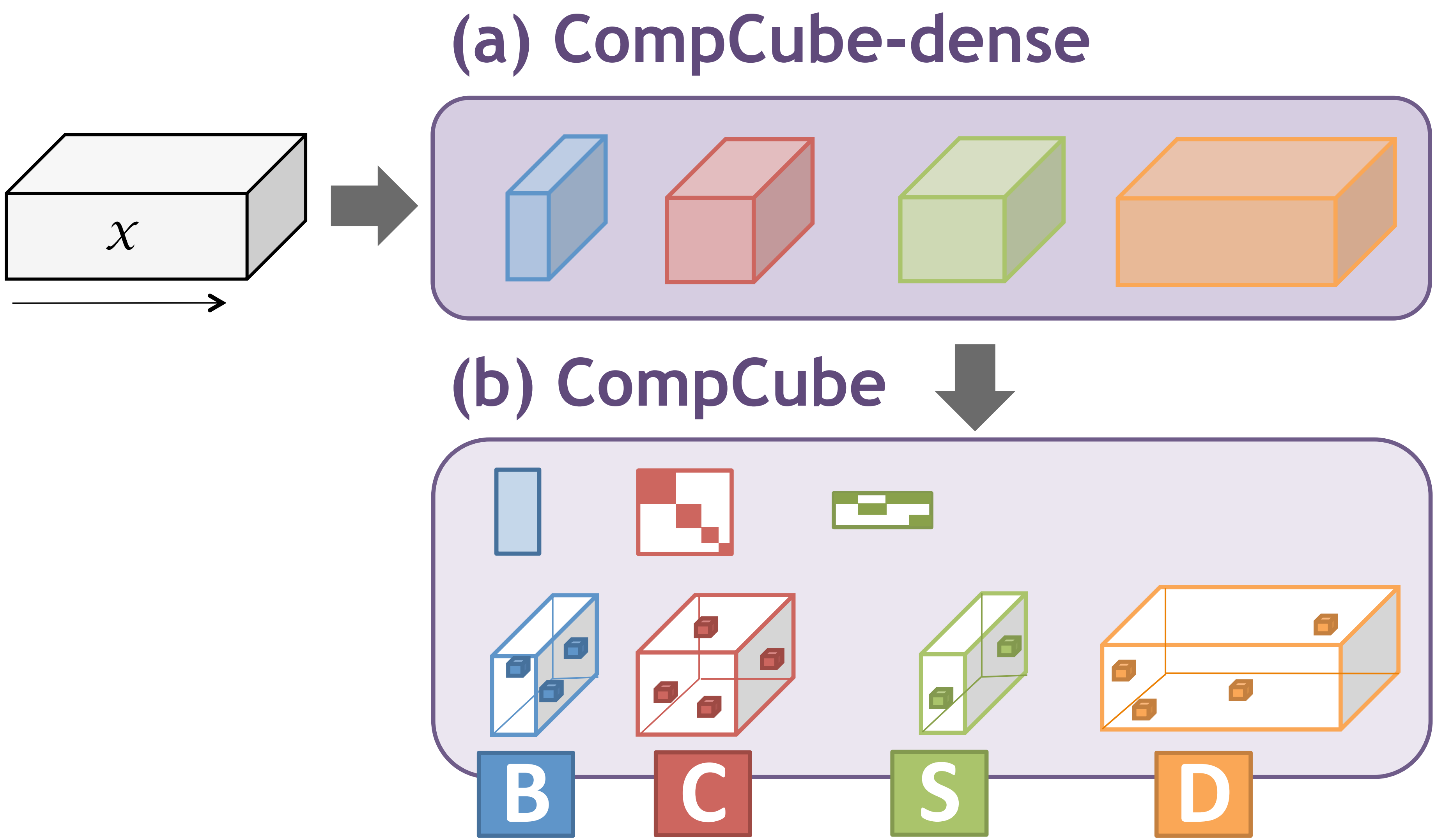
Parameter-free!

Roadmap

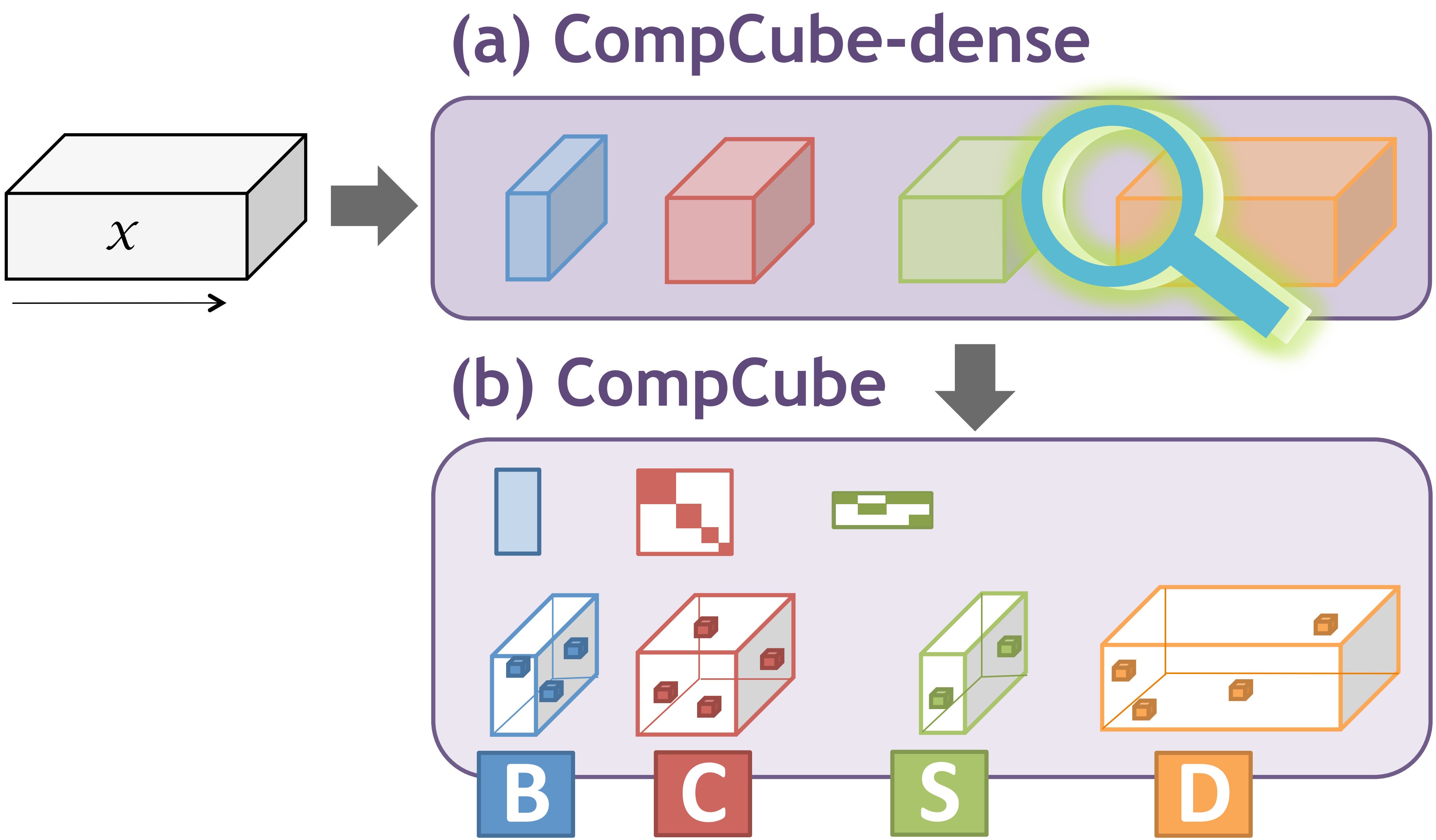
- ✔ Motivation
- ✔ Modeling power of CompCube
- ✔ Overview
  - **Proposed model**
  - Algorithm
  - Experiments
  - CompCube – at work
  - Conclusions



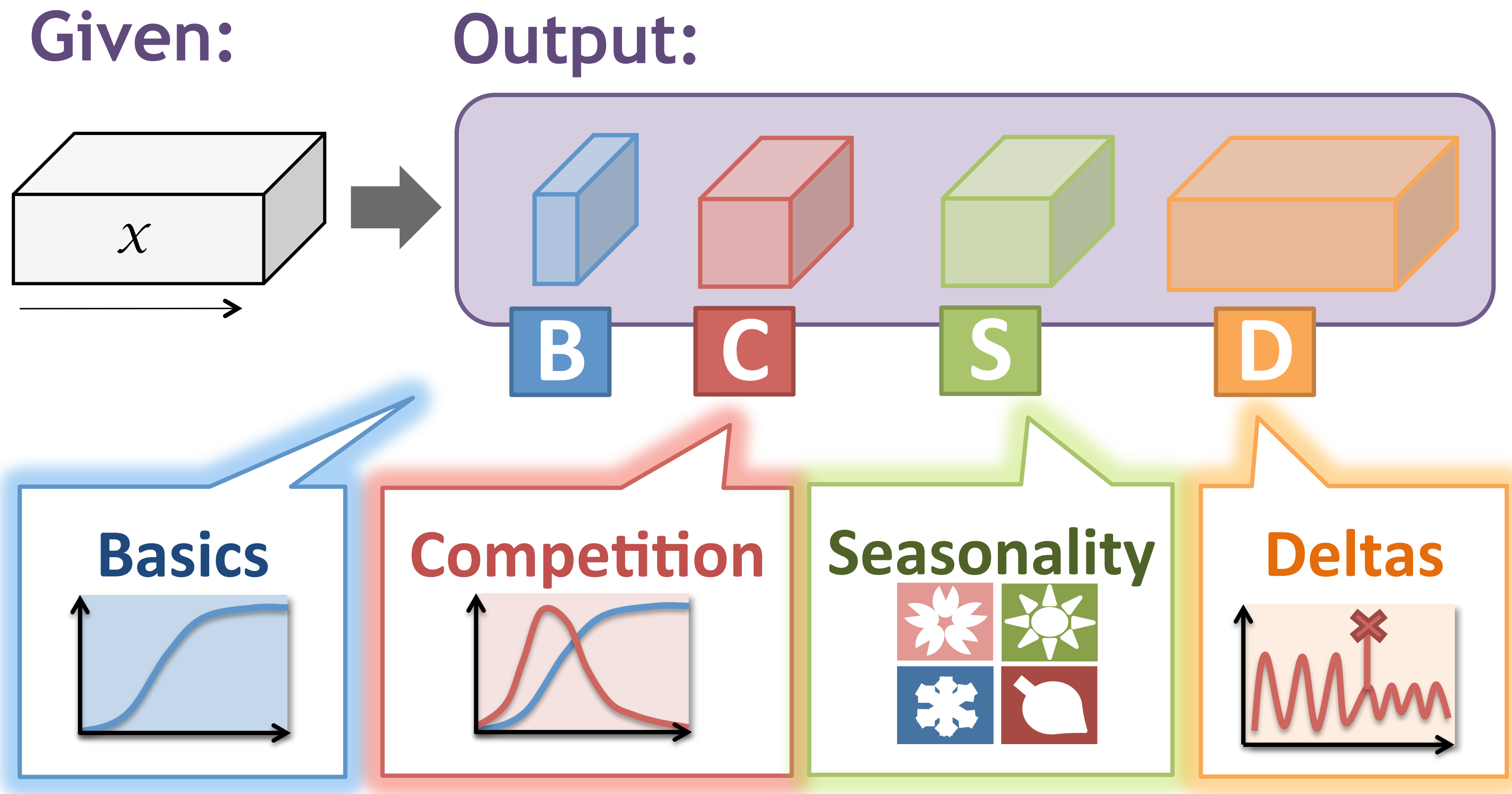
Proposed model: CompCube



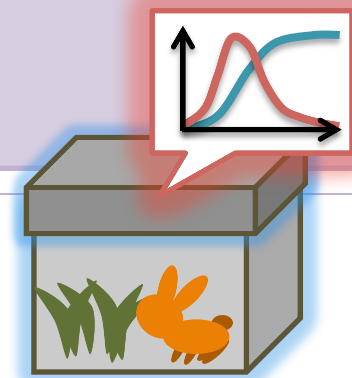
Proposed model: CompCube



CompCube-dense



CompCube-dense



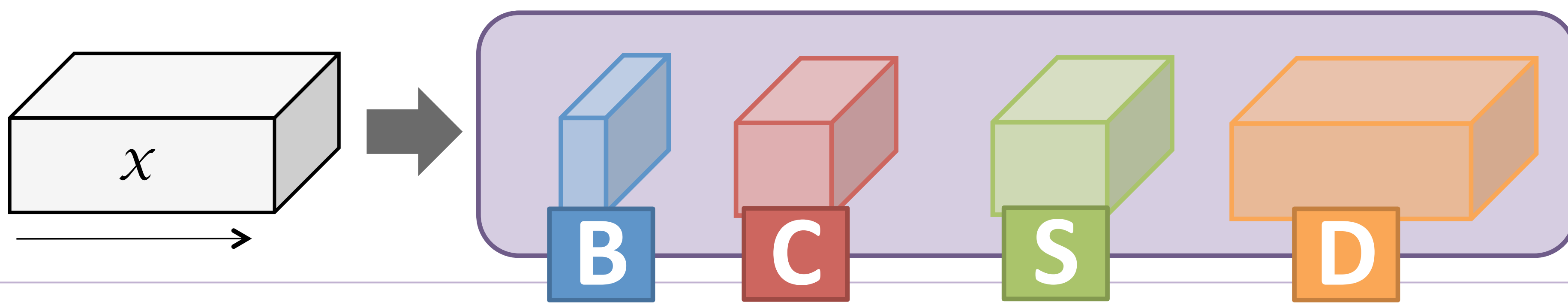
Details

Non-linear dynamical system

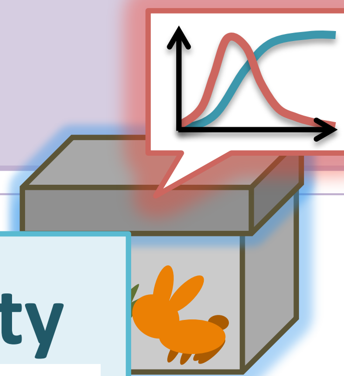
$$P_{il}(t) = P_{il}(t-1)\left[1 + r_{il}\left(1 - \frac{\sum_{j=1}^{d} c_{ijl} \cdot P_{jl}(t-1)}{K_{il}}\right)\right]$$

$$V_{il}(t) = P_{il}(t)\left[1 + s_{il}(t \mod n_p)\right] + \delta_{il}(t)$$

$$(i = 1, \cdots, d;\ l = 1, \cdots, m;\ t = 1, \cdots, n) \quad P_{il}(0) = p_{il}$$



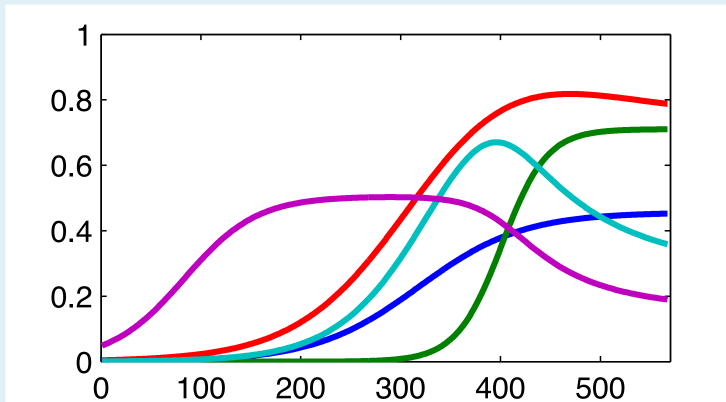
CompCube-dense



Details

Non-linear

P: latent popularity

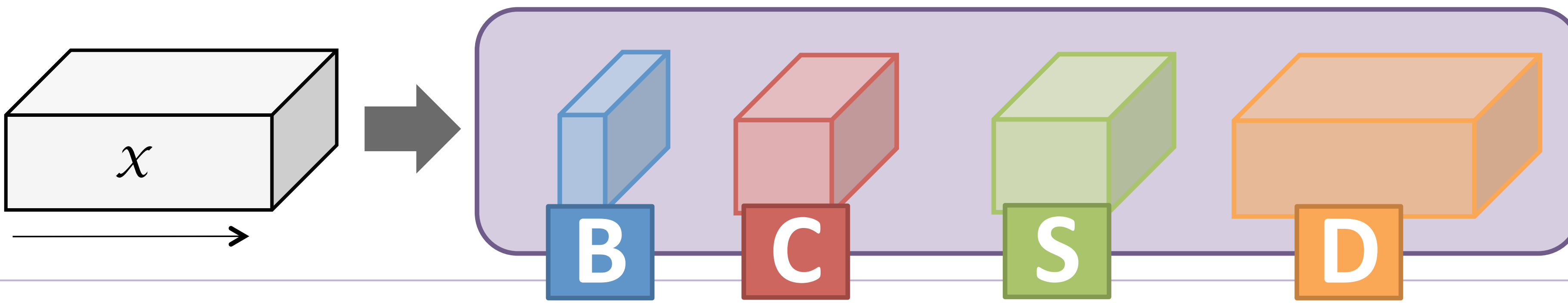


$$\boxed{P_{il}(t)} = P_{il}(t \; \cdots$$

$$\cdots_{jl} \cdot P_{jl}(t-1) \Bigg) \Bigg]$$ over K_{il}

$$V_{il}(t) = P_{il}(t) \; \cdots (t)$$

$$(i = 1, \cdots, d;\ l = 1, \cdots, m;\ t = 1, \cdots, n) \quad P_{il}(0) = p_{il}$$

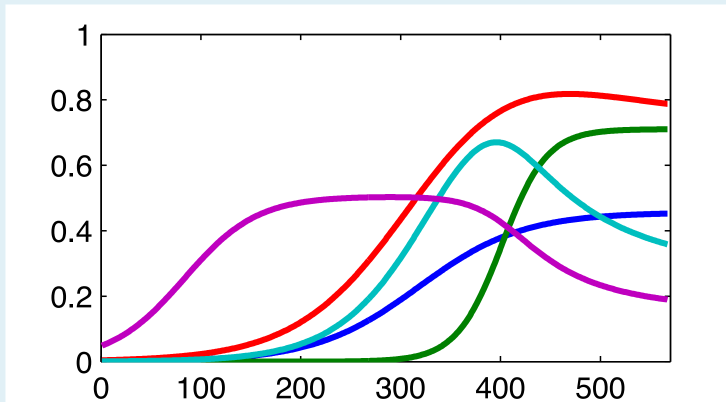


CompCube-dense

Details

Non-linear

P: latent popularity

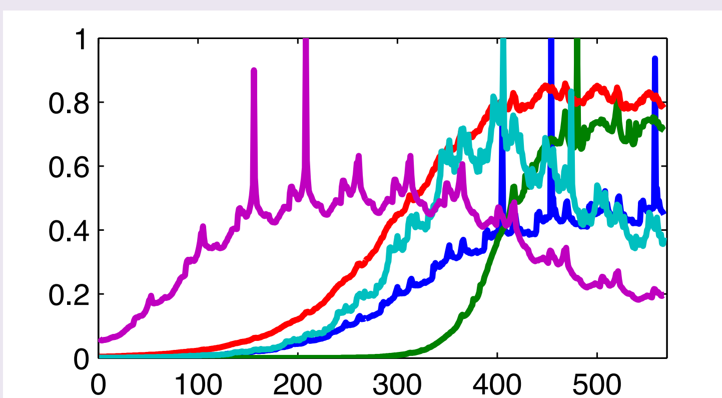


$P_{il}(t) = P_{il}(t$ … $_{jl} \cdot P_{jl}(t-1)$ … K_{il}

$V_{il}(t) = P_{il}(t)$ … (t)

$(i =$ … $= 1, \cdots, n)$ $P_{il}(0) = p_{il}$

V: estimated volume

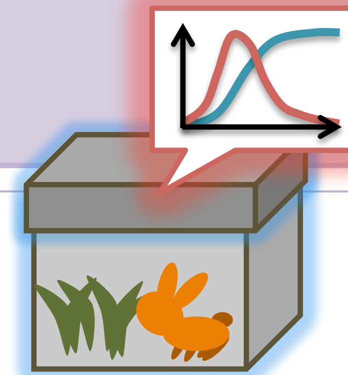


X

S

D

CompCube-dense



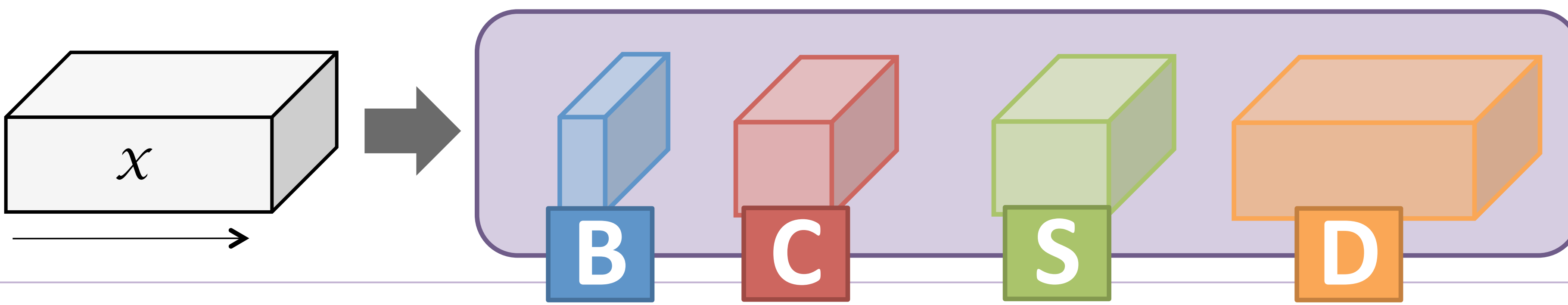
Details

Non-linear dynamical system

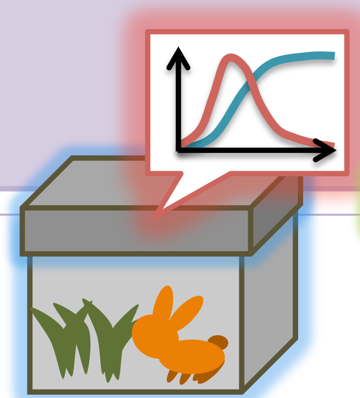
$$P_{il}(t) = P_{il}(t-1)\left[1 + r_{il}\left(1 - \frac{\sum_{j=1}^{d} c_{ijl}\cdot P_{jl}(t-1)}{K_{il}}\right)\right]$$

$$V_{il}(t) = P_{il}(t)\left[1 + s_{il}(t \mod n_p)\right] + \delta_{il}(t)$$

$$(i = 1, \cdots, d;\ l = 1, \cdots, m;\ t = 1, \cdots, n)\quad P_{il}(0) = p_{il}$$



CompCube-dense



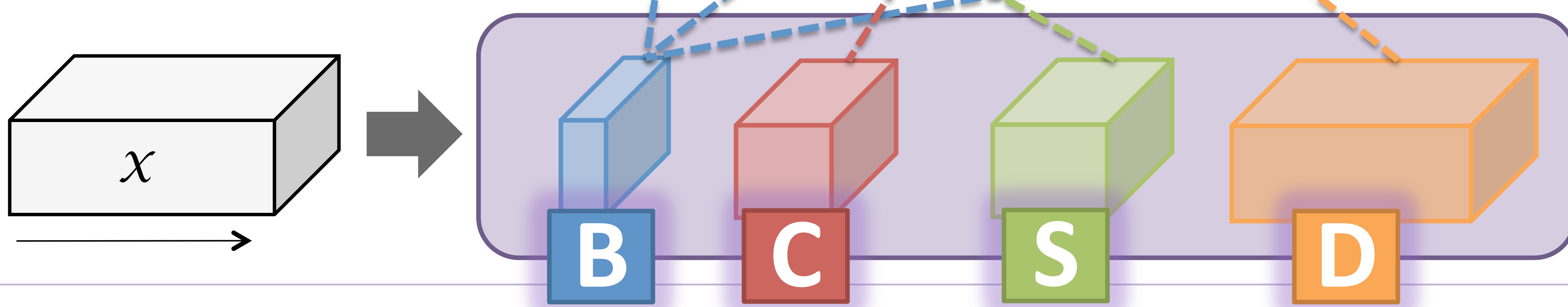
Details

Non-linear dynamical system

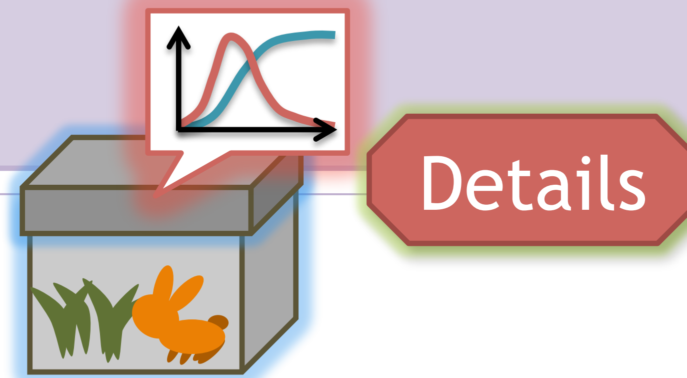
$$P_{il}(t) = P_{il}(t-1)\left[1 + \boxed{r_{il}}\left(1 - \frac{\sum_{j=1}^{d} \boxed{c_{ijl}} \cdot P_{jl}(t-1)}{\boxed{K_{il}}}\right)\right]$$

$$V_{il}(t) = P_{il}(t)\,[1 + \boxed{s_{il}}(t \mod n_p)] + \boxed{\delta_{il}(t)}$$

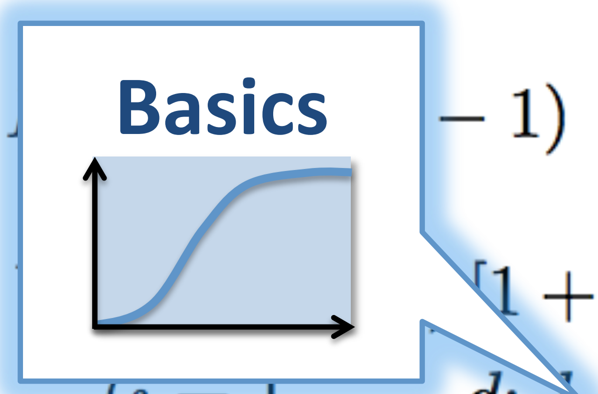
$$(i = 1, \cdots, d;\ l = 1, \cdots, m;\ t = 1, \cdots, n)\quad P_{il}(0) = \boxed{p_{il}}$$



CompCube-dense



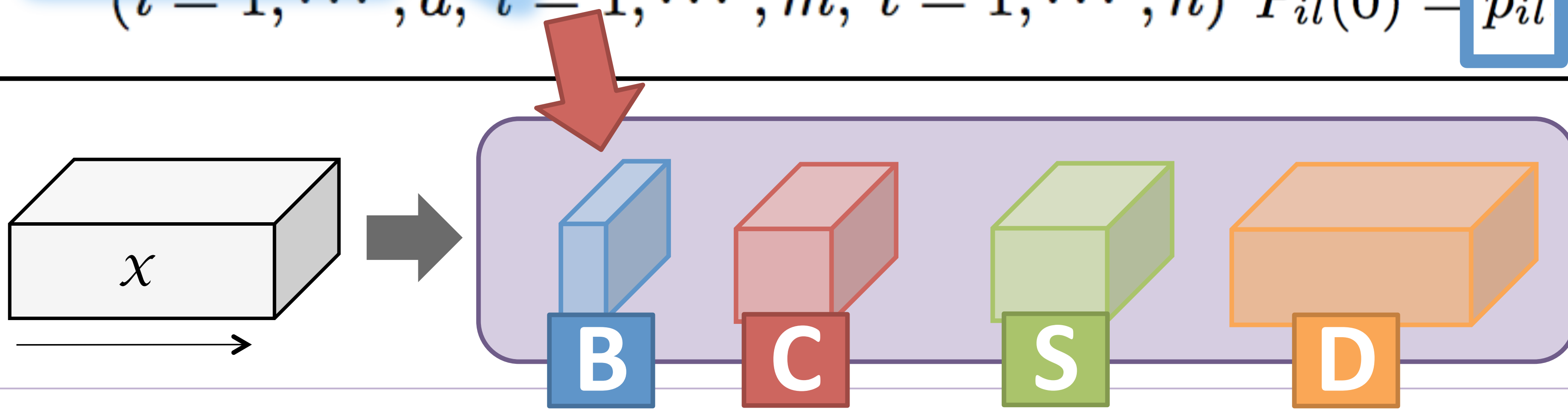
Non-linear dynamical system



$$\cdots - 1)\left[1 + \boxed{r_{il}}\left(1 - \frac{\sum_{j=1}^{d} c_{ijl} \cdot P_{jl}(t-1)}{\boxed{K_{il}}}\right)\right]$$

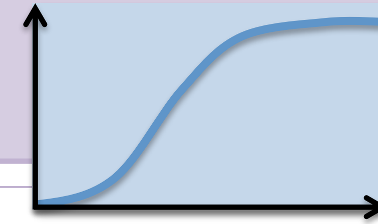
$$\cdots [1 + s_{il}(t \mod n_p)] + \delta_{il}(t)$$

$$(i = 1, \cdots, d;\ l = 1, \cdots, m;\ t = 1, \cdots, n)\quad P_{il}(0) = \boxed{p_{il}}$$

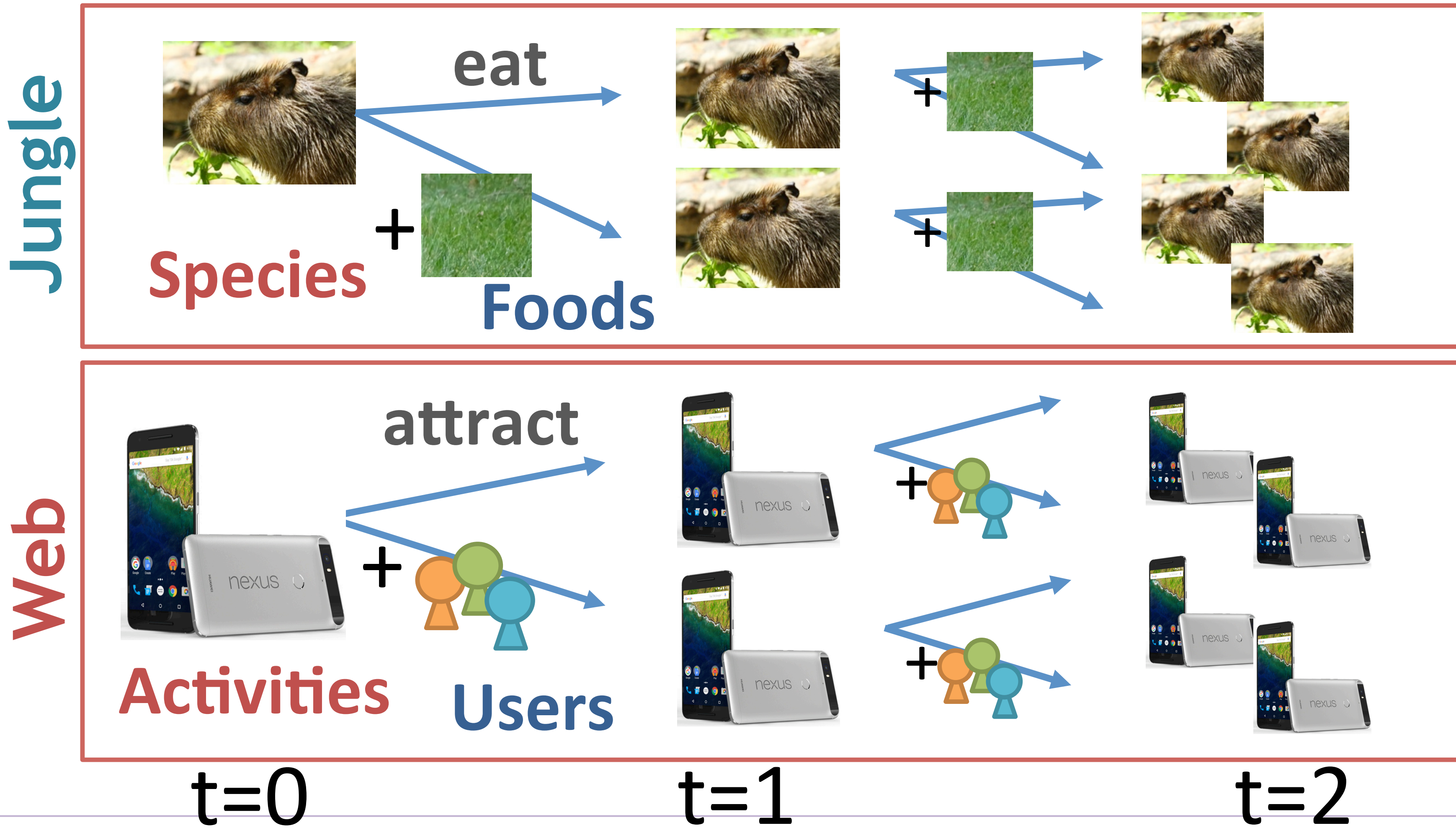


B

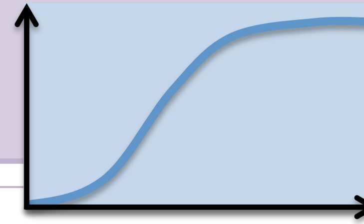
Basics



Popularity size increases over time

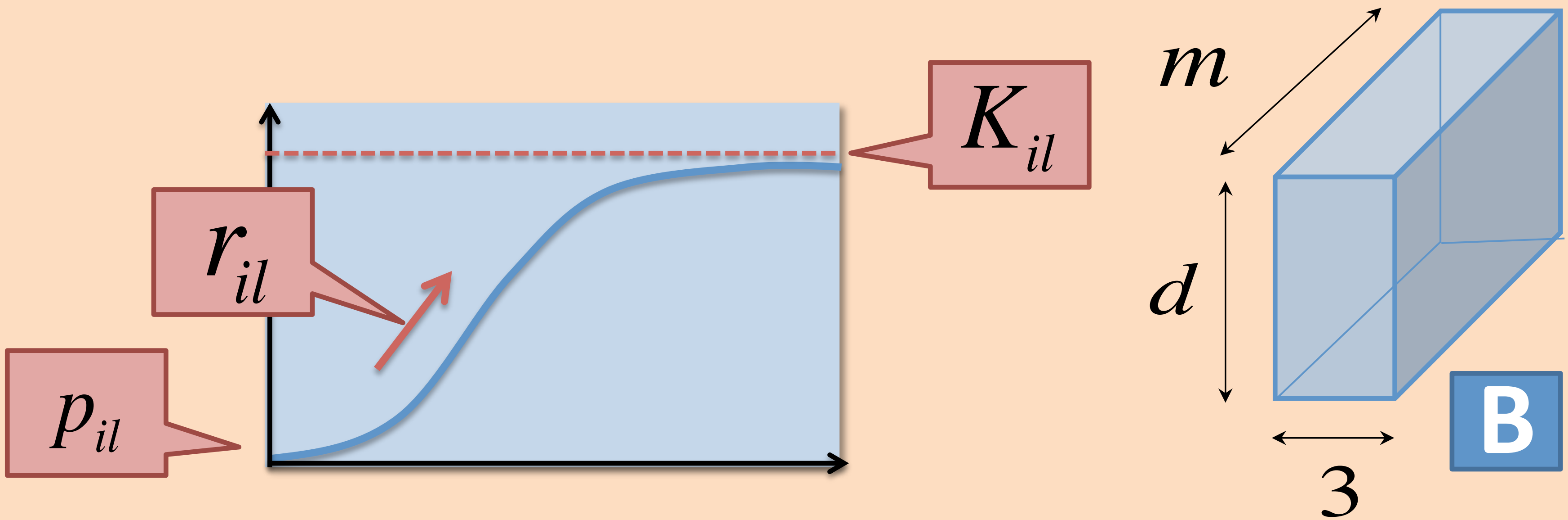


B Basics



Popularity size increases over time

(activity i, location l)

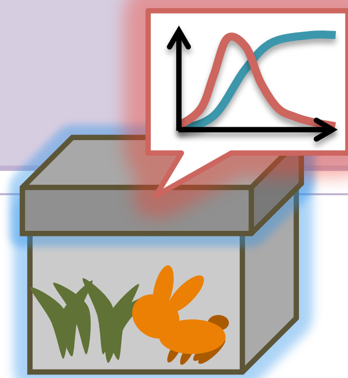


p_{il} – Initial condition (i.e., P(0) =p)

r_{il} – Growth rate, attractiveness

K_{il} – Carrying capacity (=available user resources)

CompCube-dense



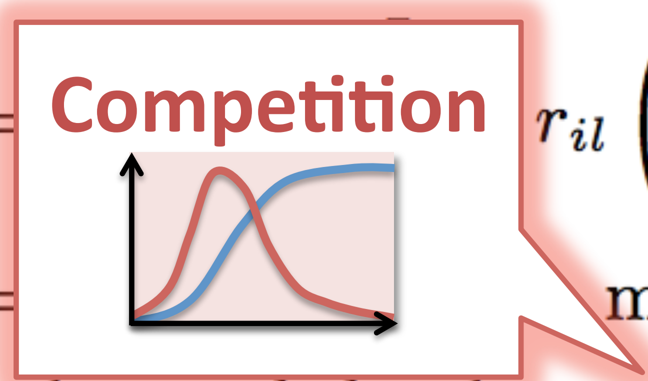
Details

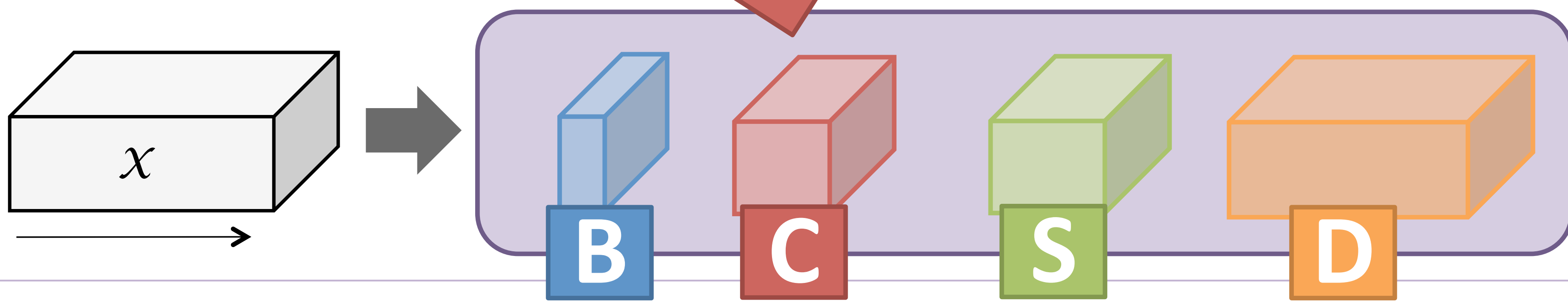
Non-linear dynamical system

$$P_{il}(t) = \quad \cdots \quad r_{il}\left(1 - \frac{\sum_{j=1}^{d} \boxed{c_{ijl}} \cdot P_{jl}(t-1)}{K_{il}}\right)\Bigg]$$

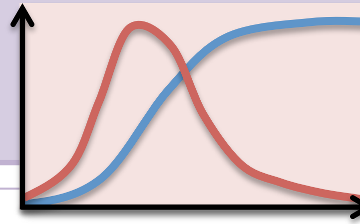
$$V_{il}(t) = \quad \cdots \quad \bmod n_p)] + \delta_{il}(t)$$

$$(i = 1, \cdots, d;\ l = 1, \cdots, m;\ t = 1, \cdots, n)$$



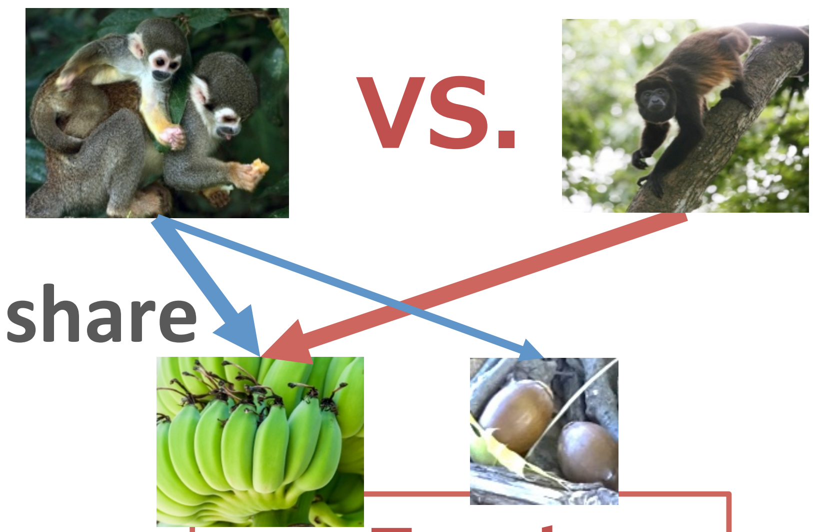


C Competition



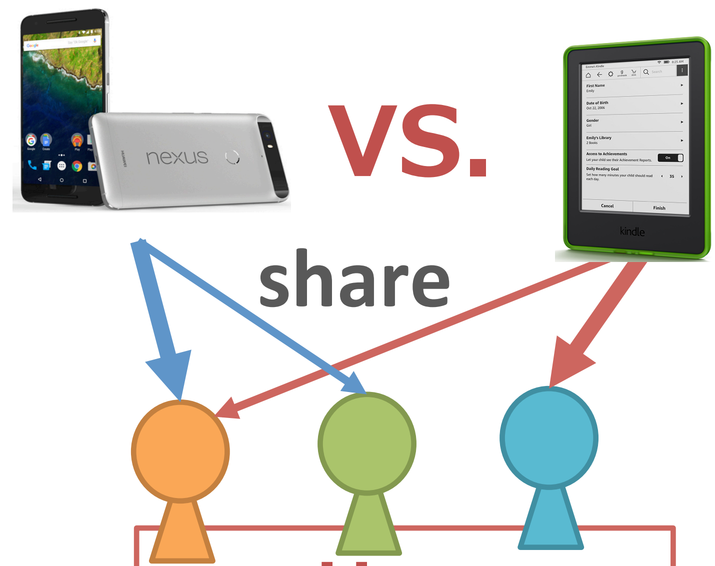
Interaction between multiple keywords

Species



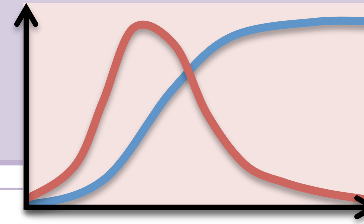
Food resources

Activities



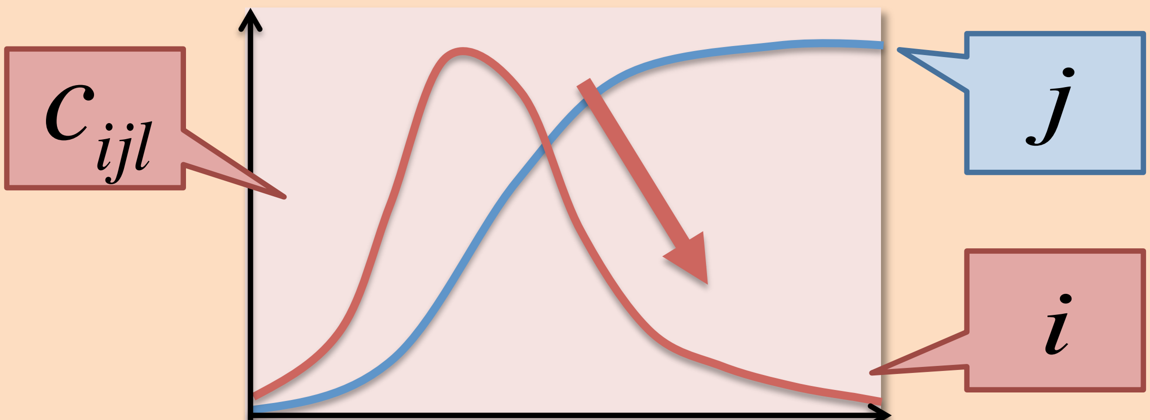
User resources

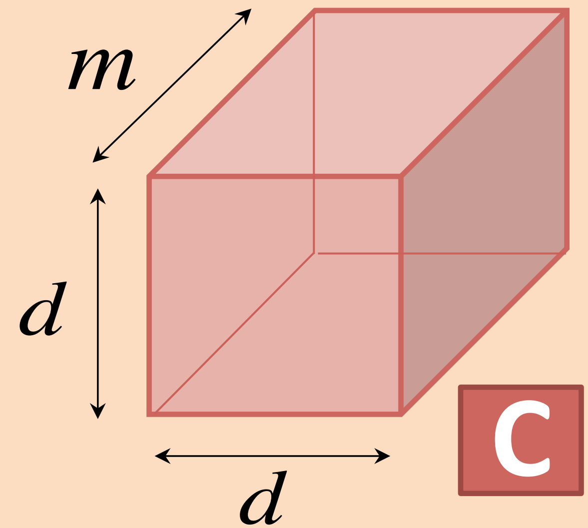
C Competition



Interaction between multiple keywords

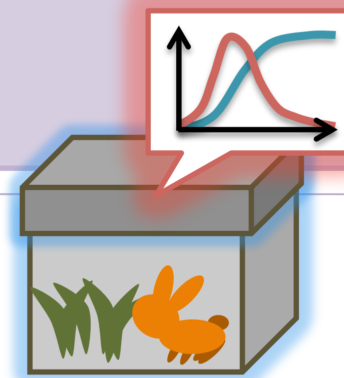
(location l)





c_{ijl} – Competition coefficient in location l
i.e., effect rate of activity j on i in l

CompCube-dense



Details

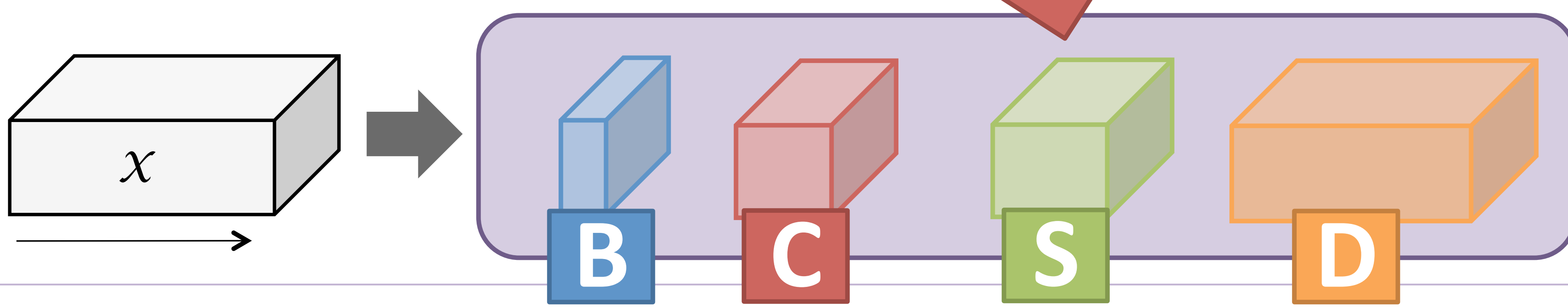
Non-linear dynamical system

$$P_{il}(t) = P_{il}(t-1)\left[1 + r_{il}\left(1 - \frac{\sum_{j=1}^{d} c_{ijl} \cdots}{\cdots}\right)\cdots\right]$$

$$V_{il}(t) = P_{il}(t)\,[1 + \boxed{s_{il}}[t \mod n_p)] + \delta_{\cdots}$$

$$(i = 1, \cdots, d;\ l = 1, \cdots, m;\ t = 1, \cdots, n)$$

Seasonality



S Seasonality

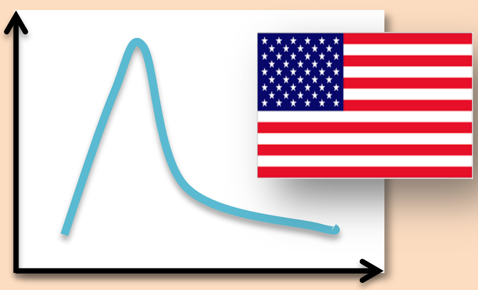
“Hidden” seasonal activities

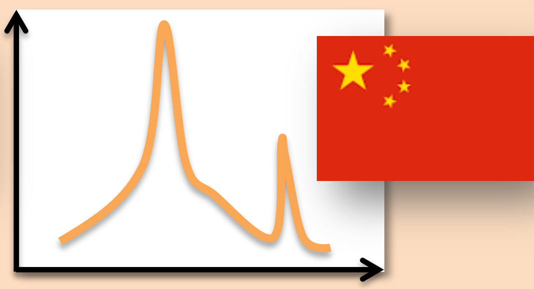
Season/ Climate

Seasonal events

“Hidden” seasonal activities

Users change their behavior according to **local seasonal events**!





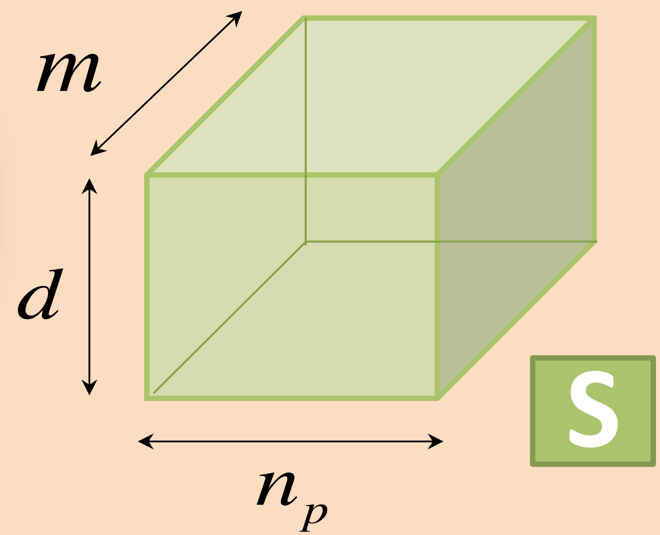
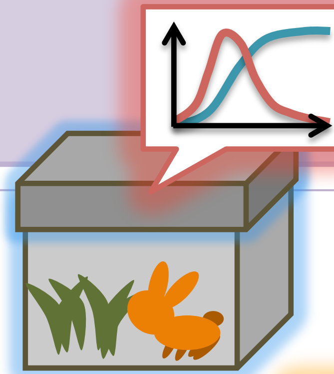


Image courtesy of xura, criminalatt, David Castillo Dominici, happykanppy at FreeDigitalPhotos.net.

CompCube-dense



Details

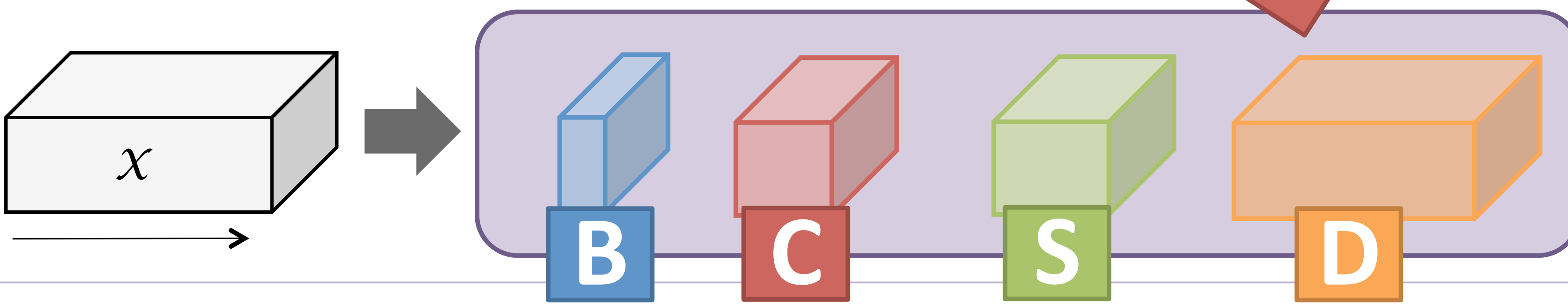
Non-linear dynamical system

$$P_{il}(t) = P_{il}(t-1)\left[1 + r_{il}\left(1 - \frac{\sum_{j=1}^{d} c_{ijl} \cdot P_{j}}{K_{il}}\right.\right.$$

$$V_{il}(t) = P_{il}(t)\left[1 + s_{il}(t \mod n_p)\right] + \boxed{\delta_{il}(t)}$$

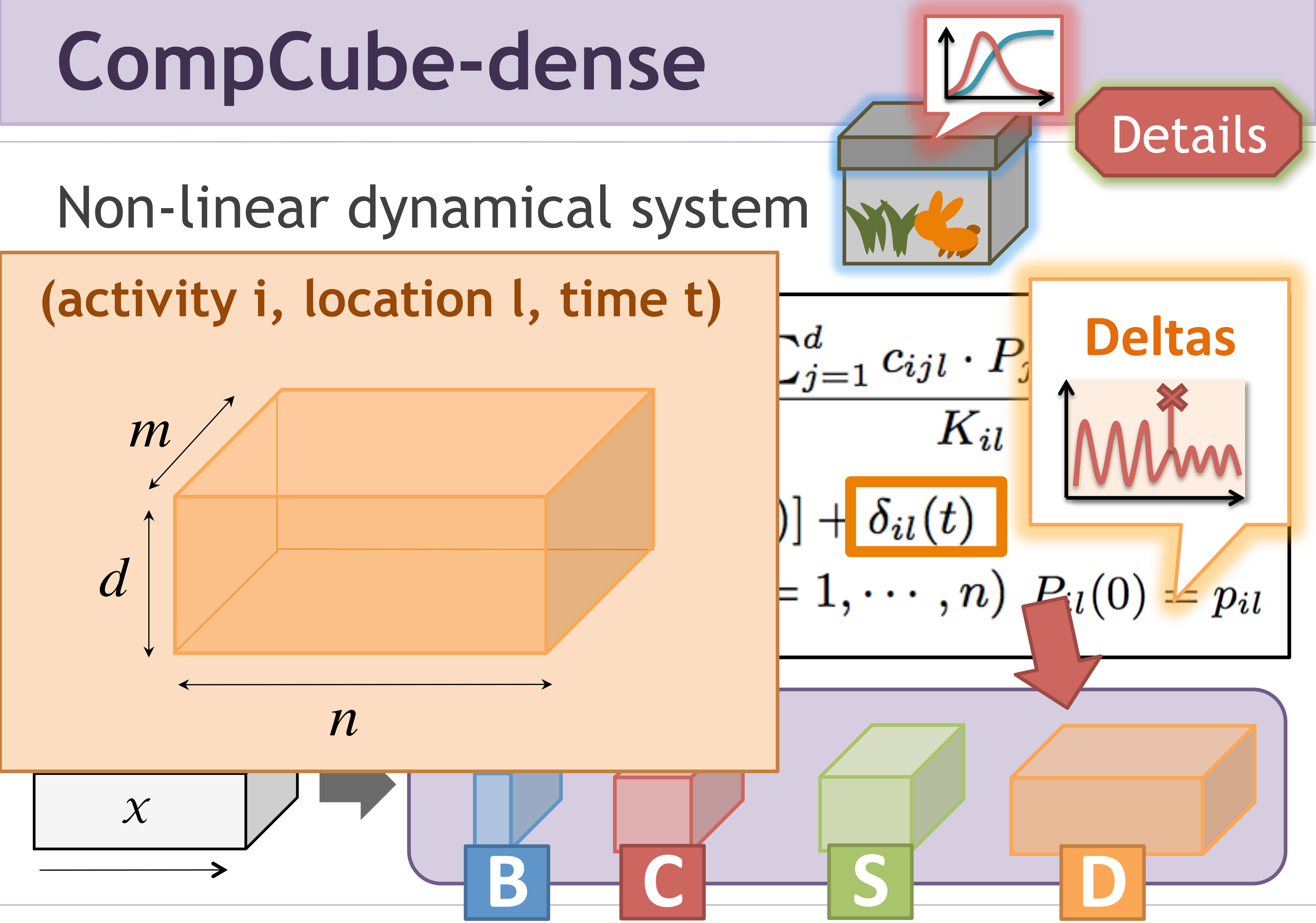
$$(i = 1, \cdots, d;\ l = 1, \cdots, m;\ t = 1, \cdots, n)\quad P_{il}(0) = p_{il}$$

Deltas

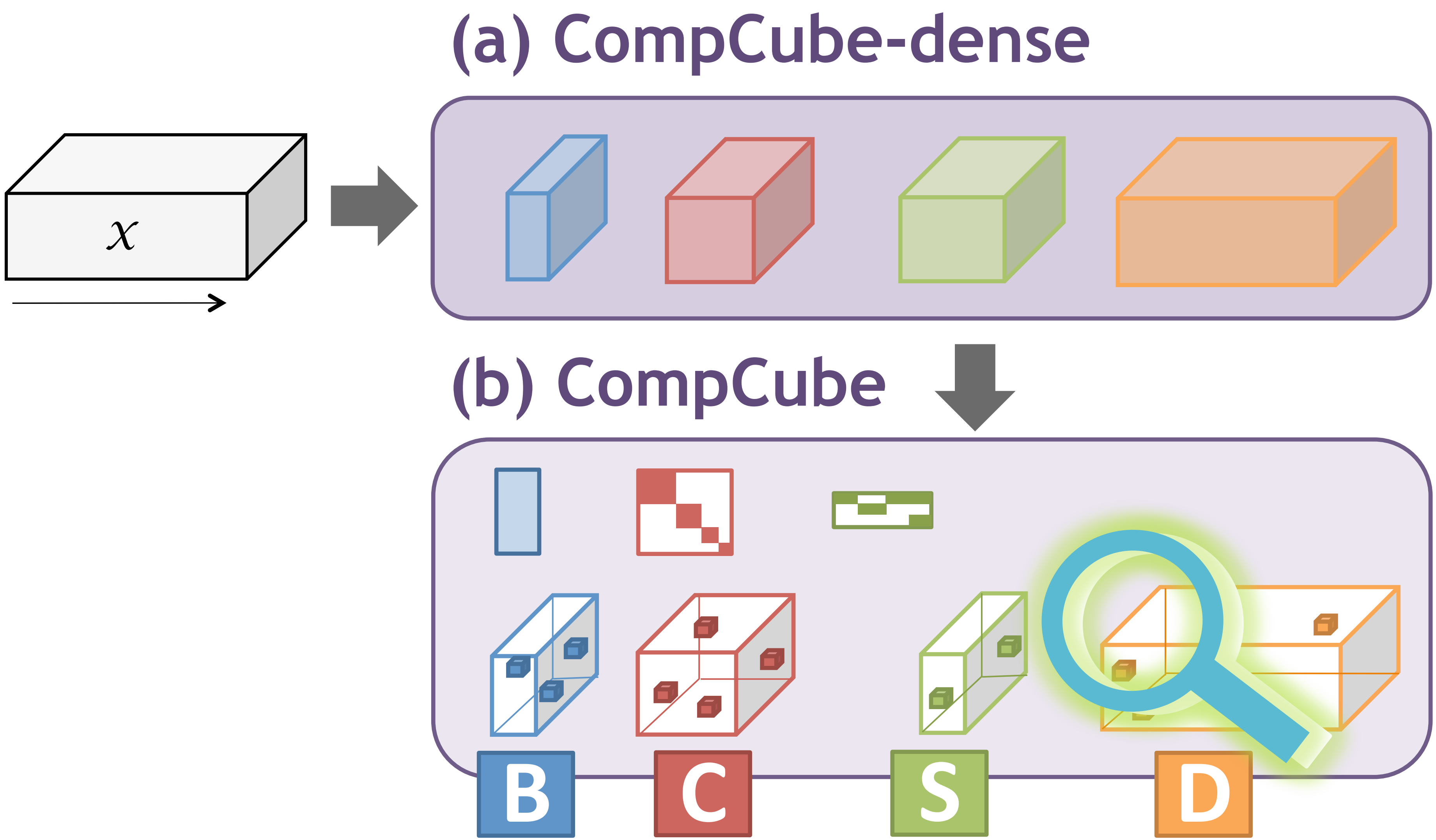


CompCube-dense

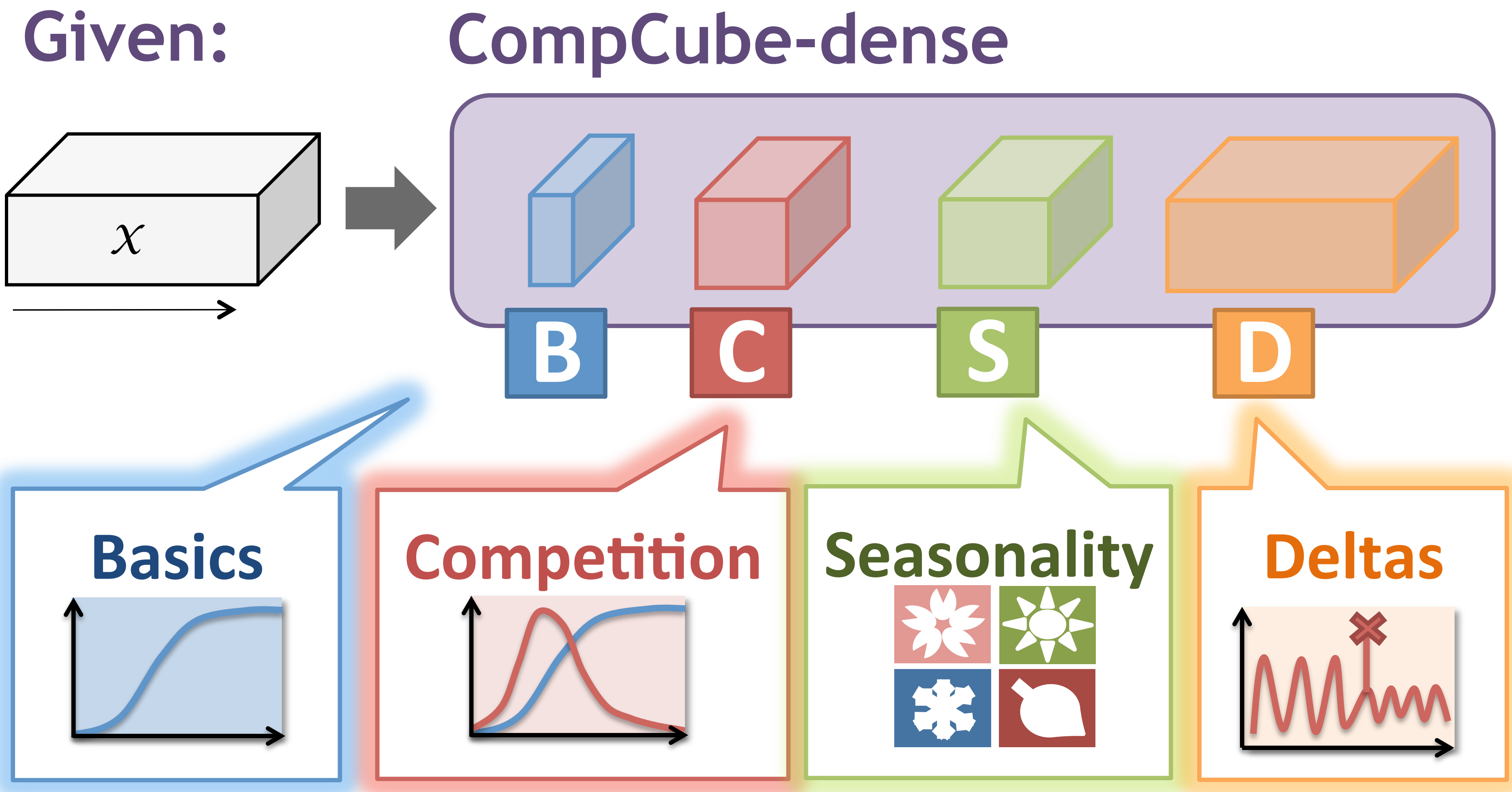
Non-linear dynamical system



Proposed model: CompCube

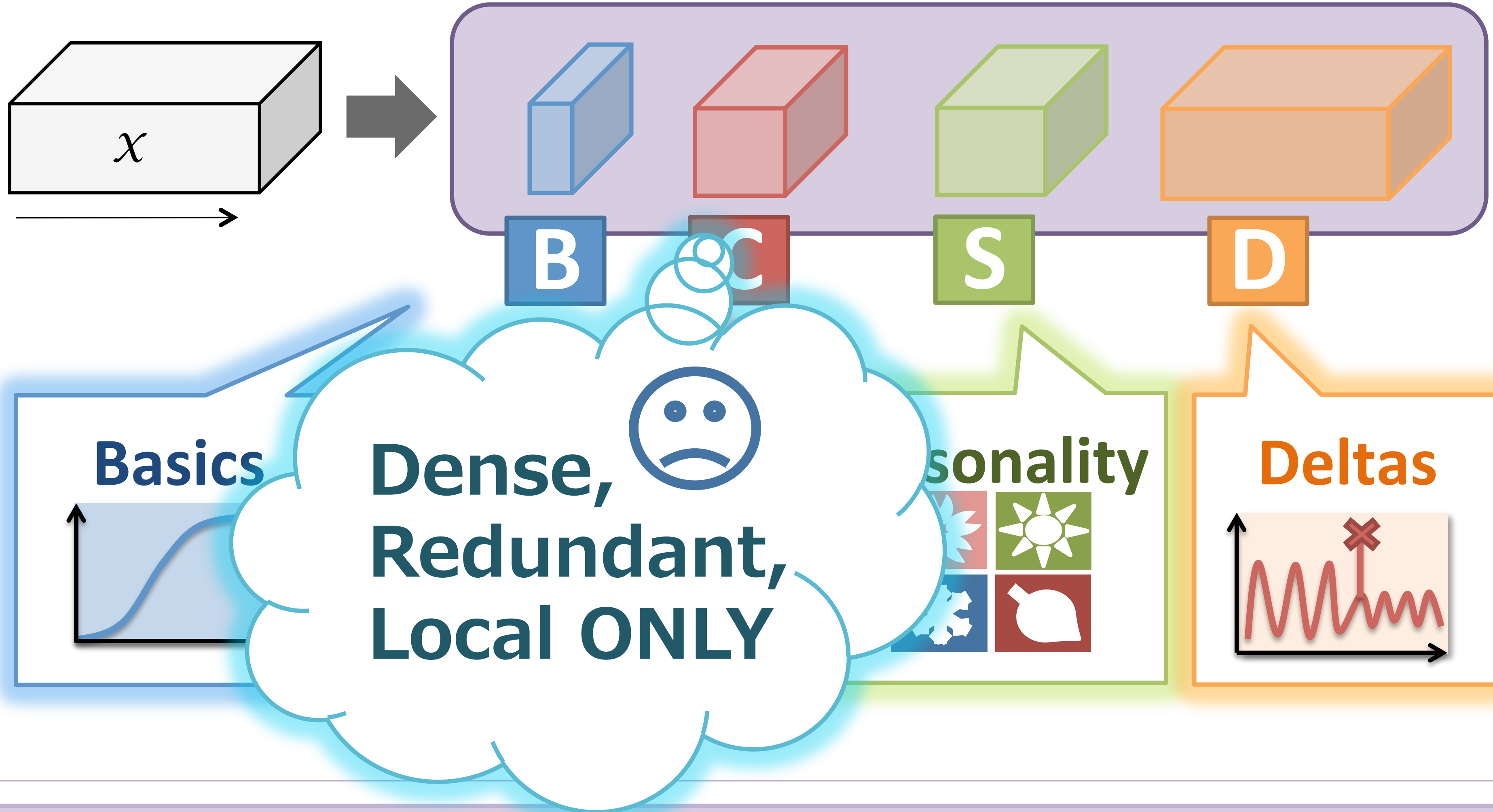


Initial attempt: CompCube-dense

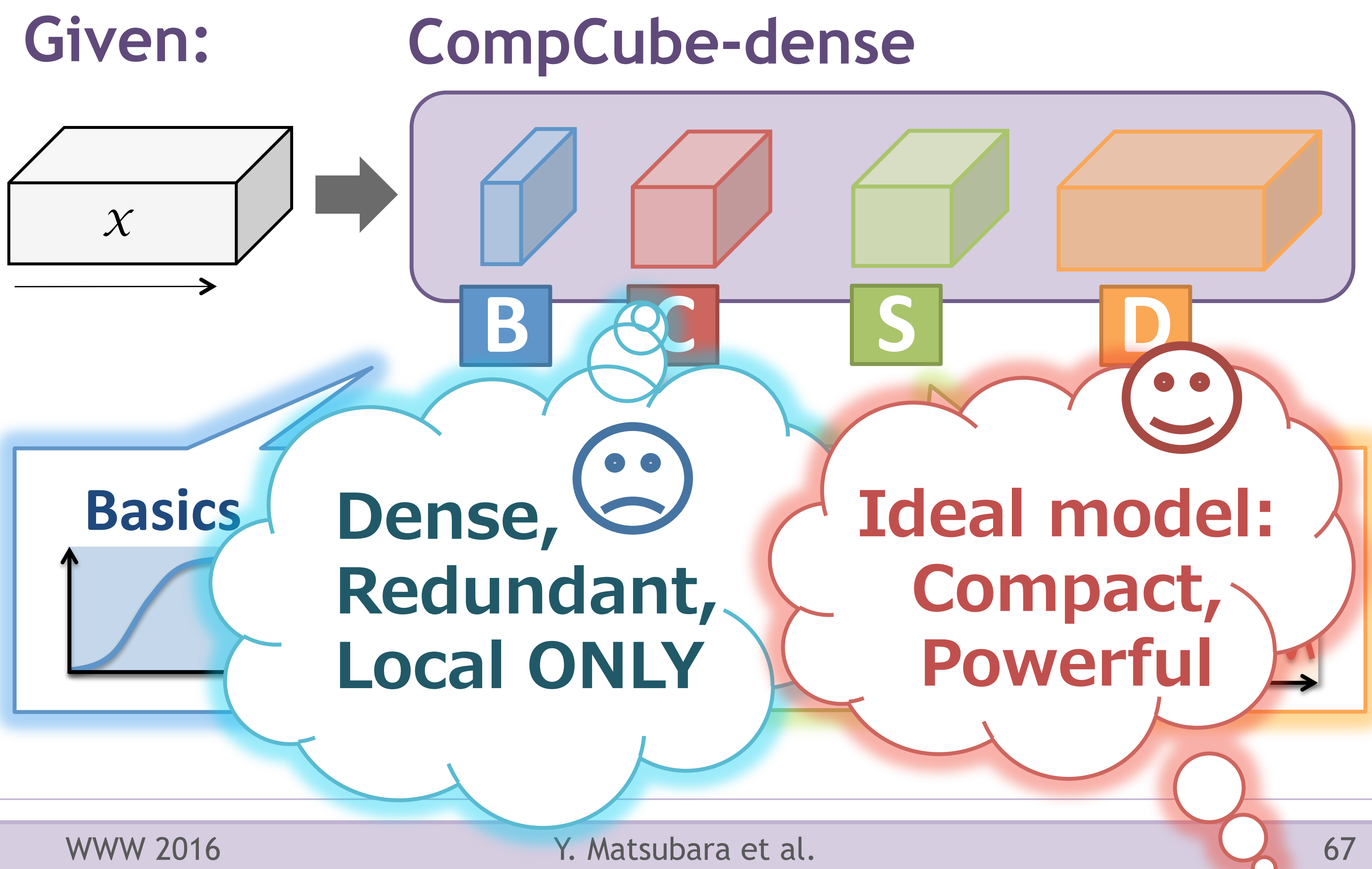


Initial attempt: CompCube-dense

Given: **CompCube-dense**

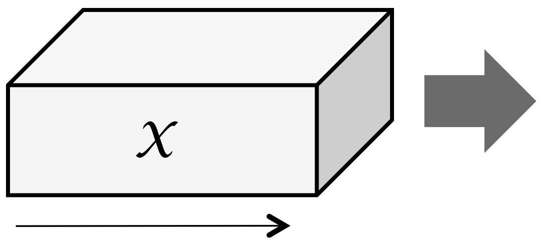


Initial attempt: CompCube-dense

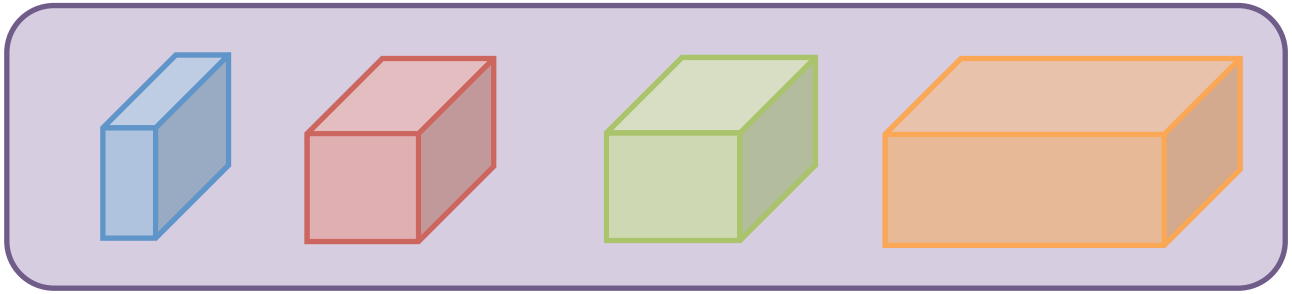


Initial attempt: CompCube-dense

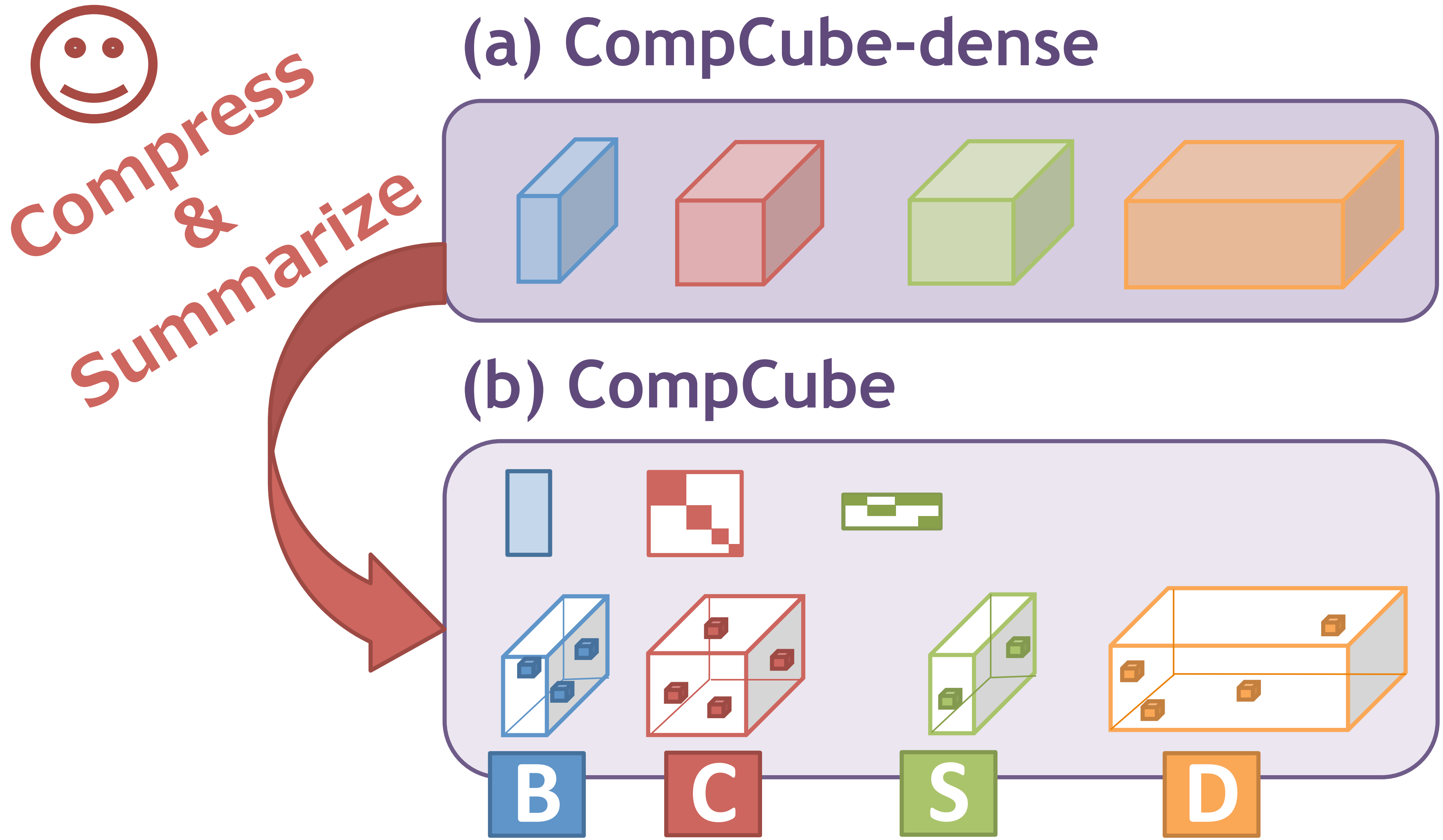
Given:



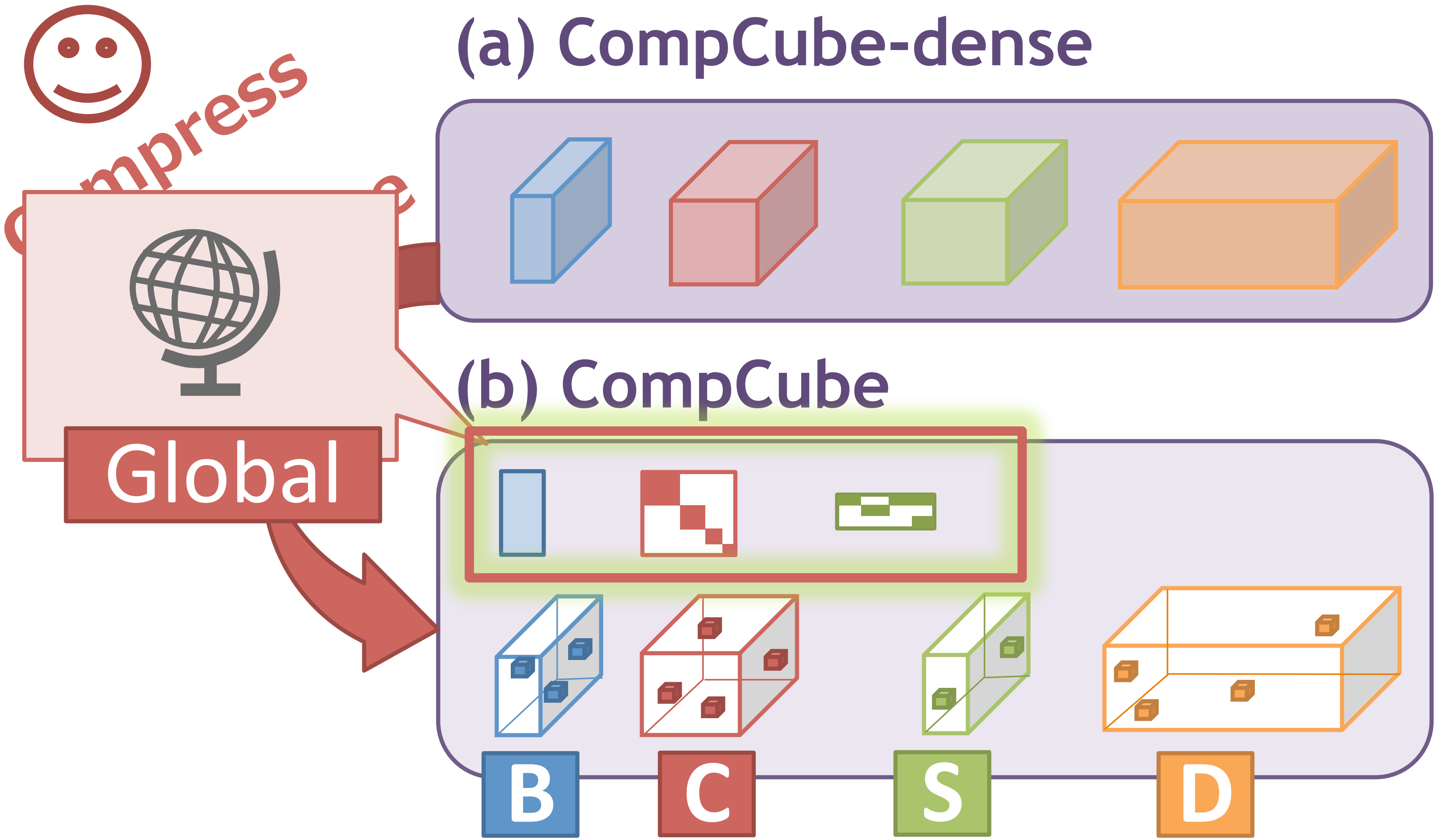
CompCube-dense



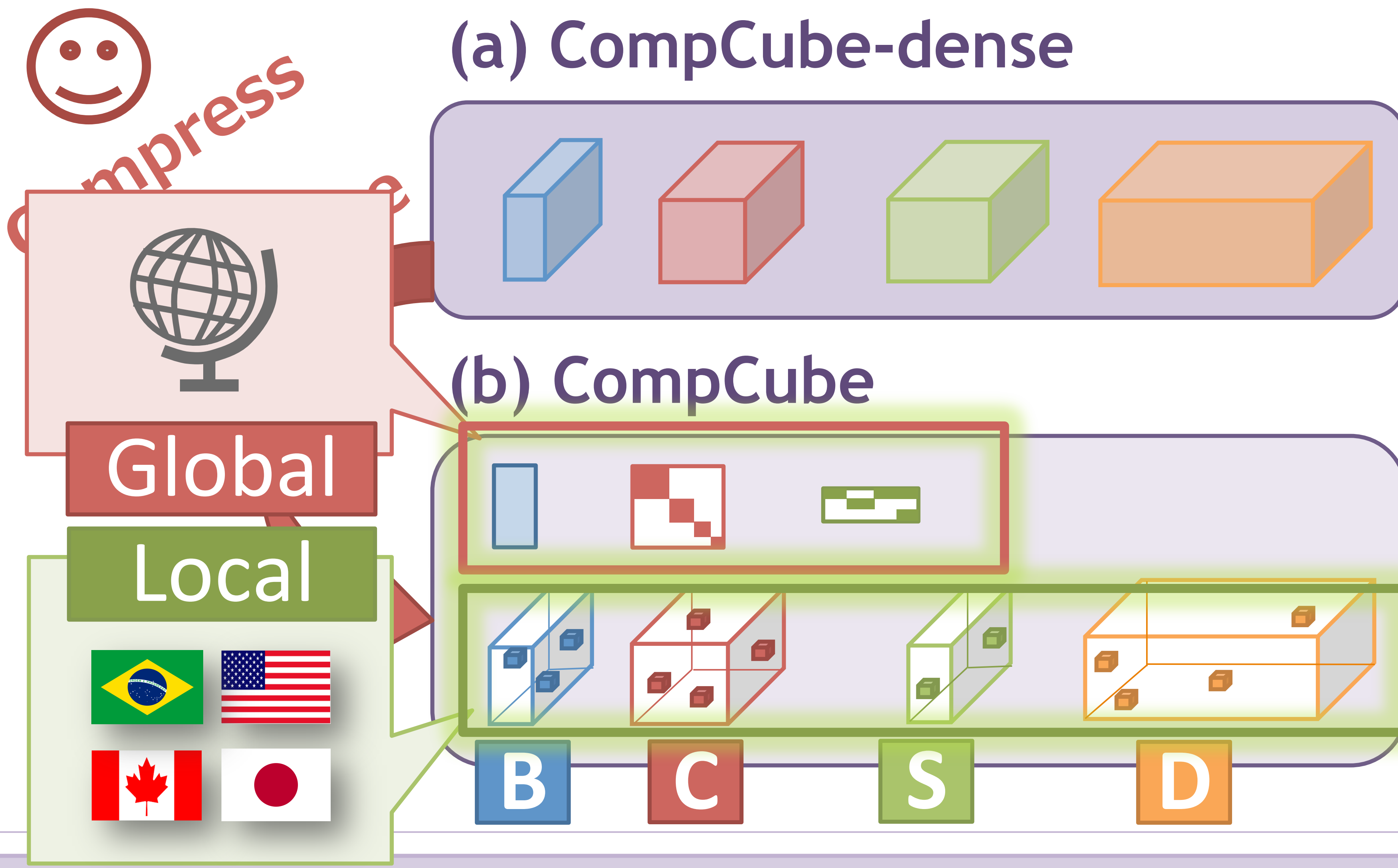
Final model: CompCube



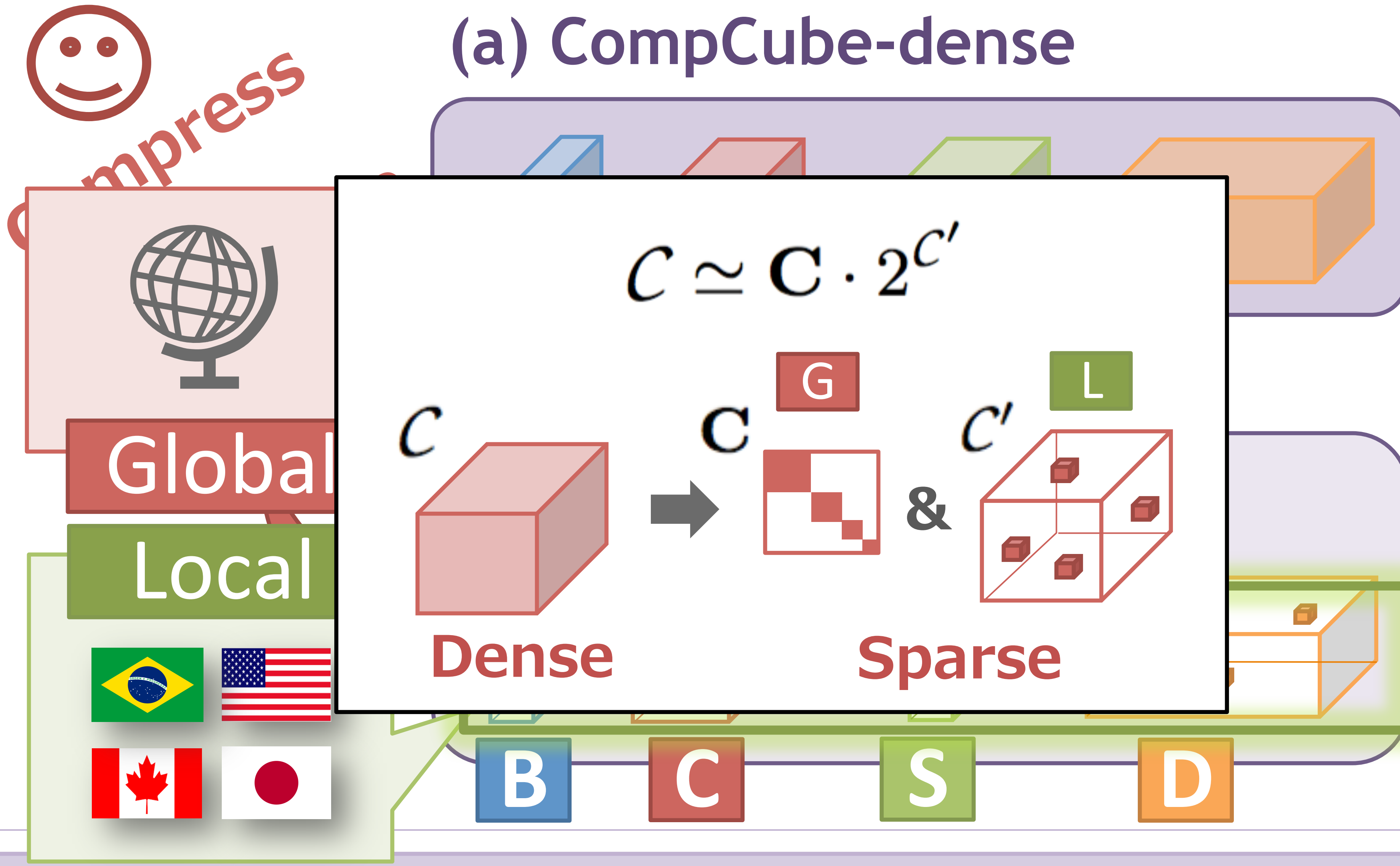
Final model: CompCube



Final model: CompCube

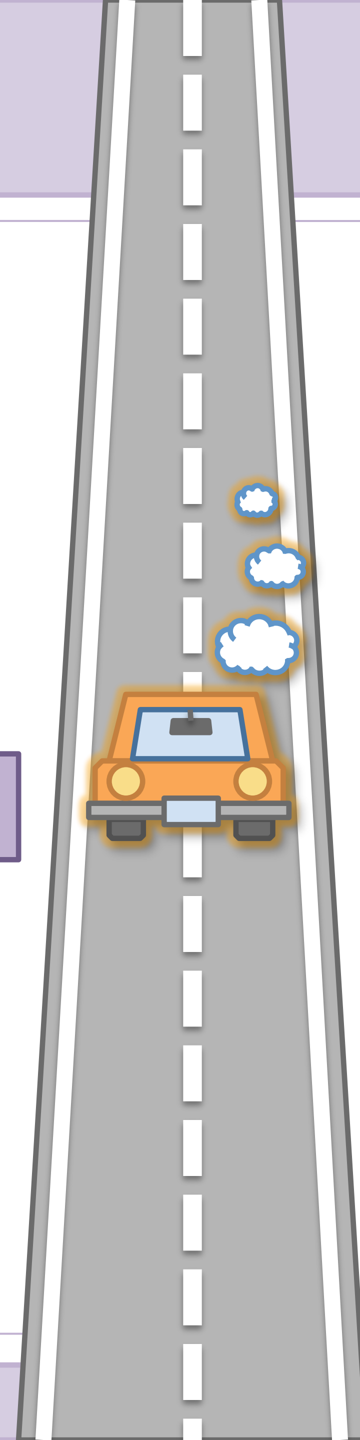


Final model: CompCube



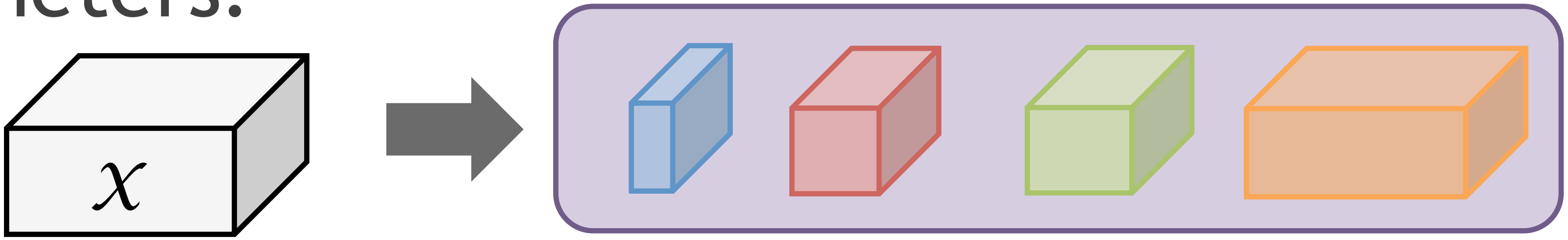
Roadmap

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

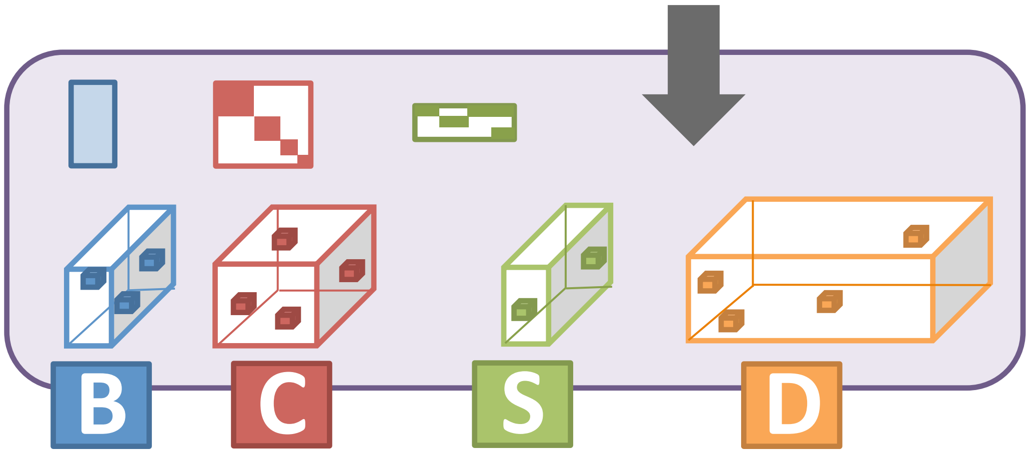


Challenges

Q1. How can we efficiently estimate parameters?

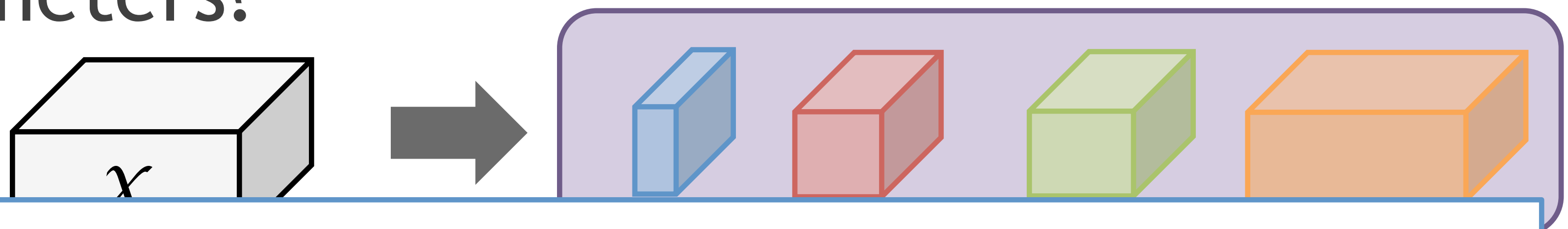


Q2. How can we automatically find best parameter sets?



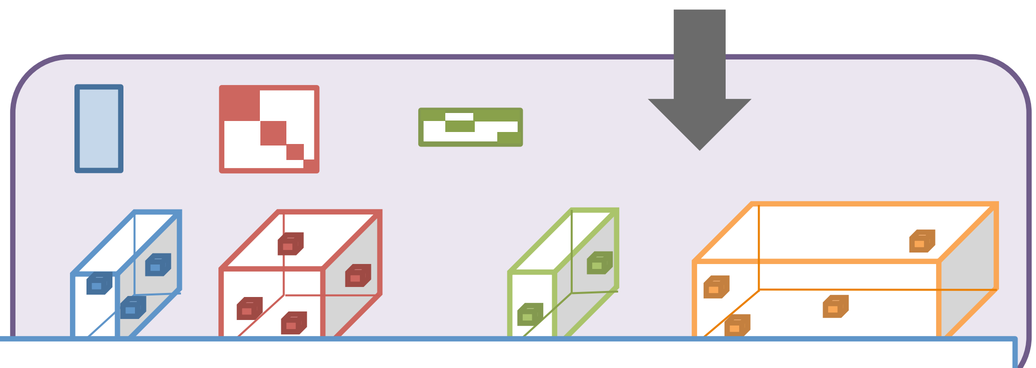
Challenges

Q1. How can we efficiently estimate parameters?



Idea (1) : TetraFit algorithm

Q2. How can we automatically find best parameter sets?

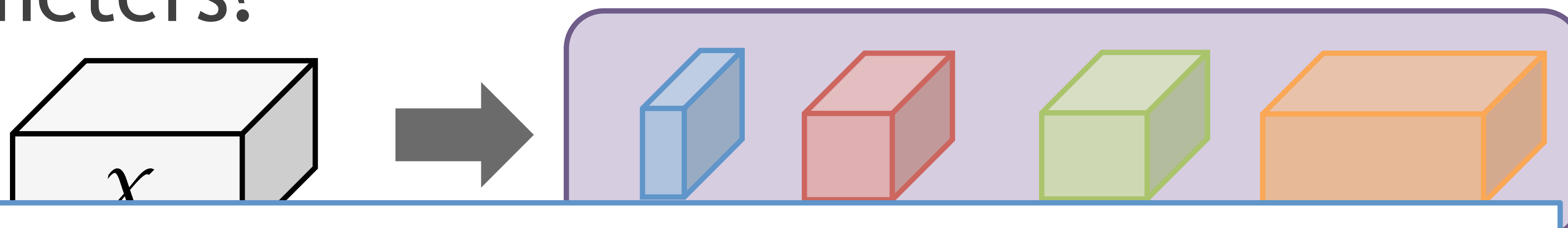


Idea (2): Model description cost

Challenges

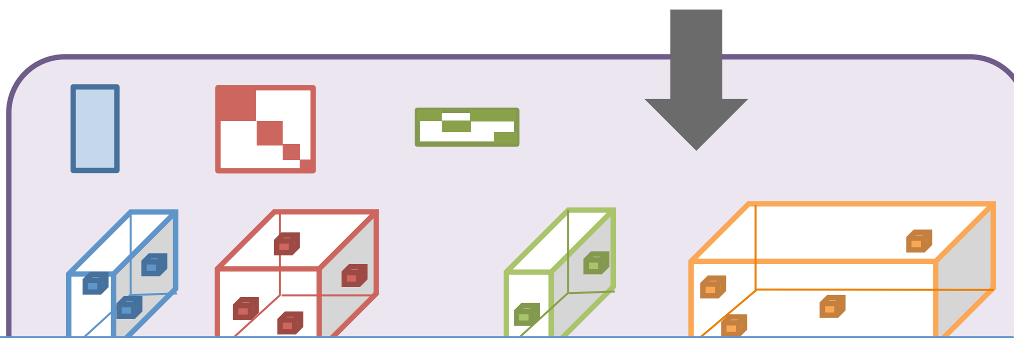
(Details in paper)

Q1. How can we efficiently estimate parameters?



Idea (1) : TetraFit algorithm

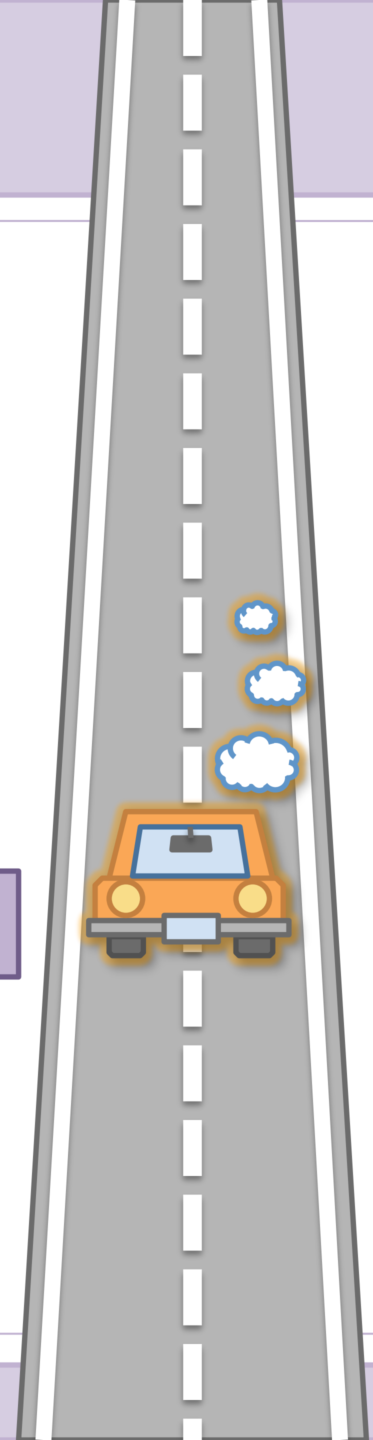
Q2. How can we automatically find best parameter sets?



Idea (2): Model description cost

Roadmap

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



Experiments

We answer the following questions…

Q1. Effectiveness

How well does it explain important patterns?

Q2. Accuracy

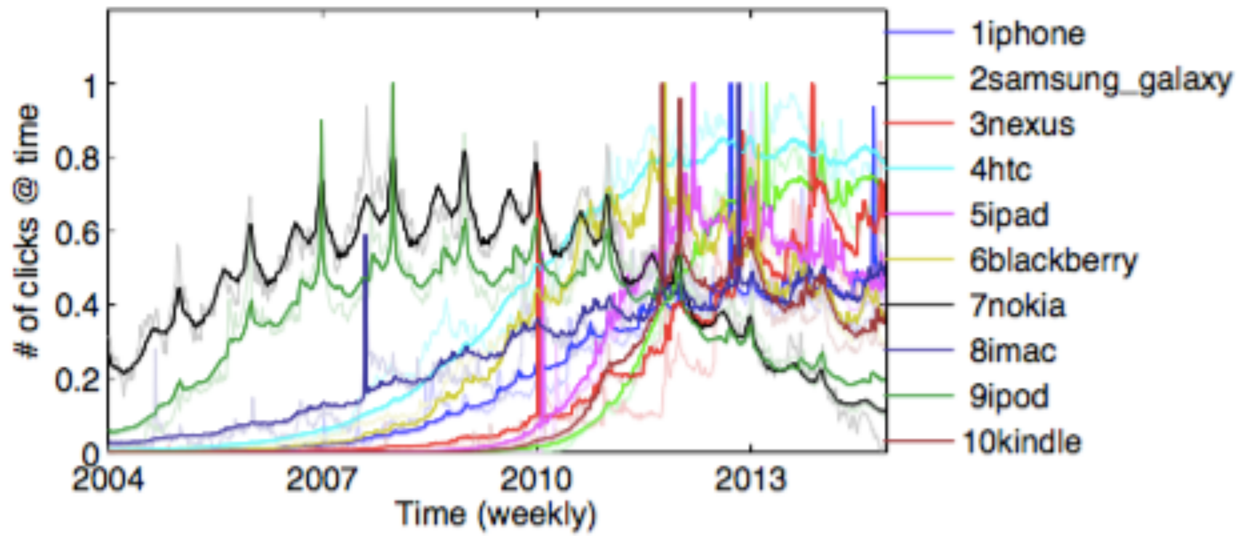
How well does it fit real datasets?

Q3. Scalability

How does it scale in terms of computational time?

Q1. Effectiveness

1. Products

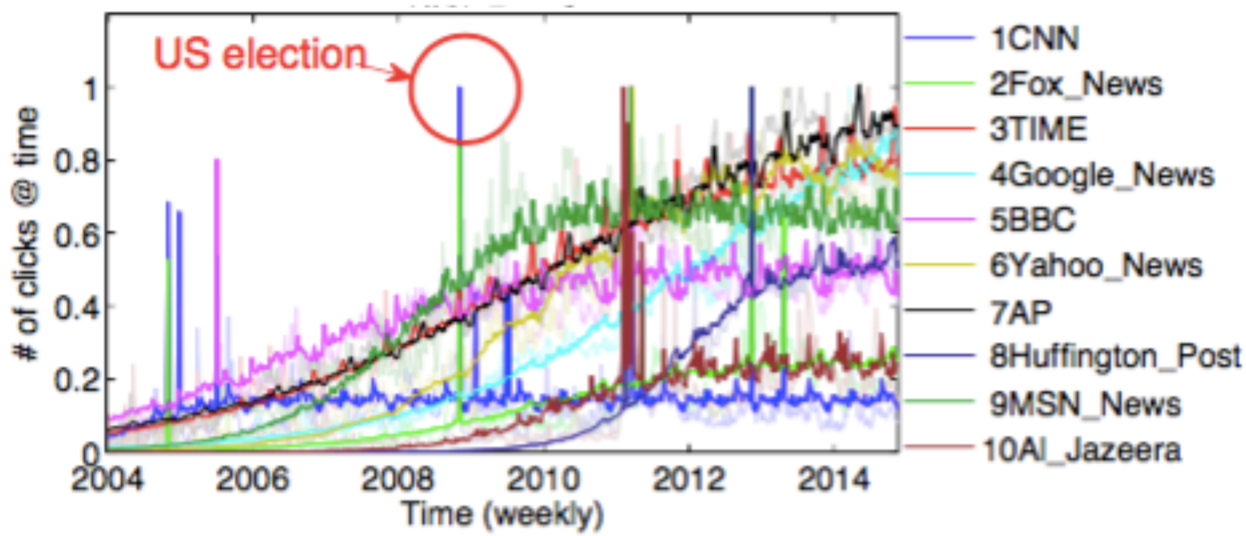


2. News

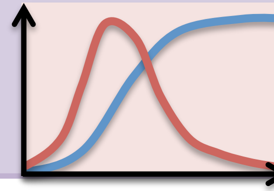
Microsoft

VS.

YAHOO! NEWS



Q1. Effectiveness



1. Products

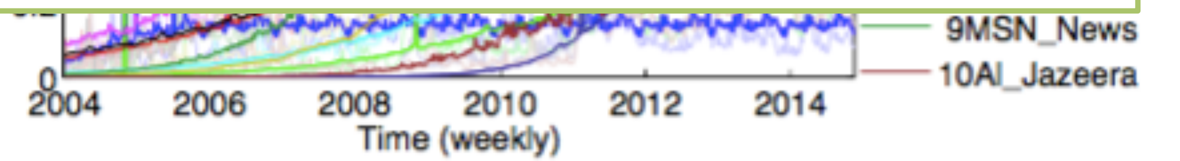
VS

2. News

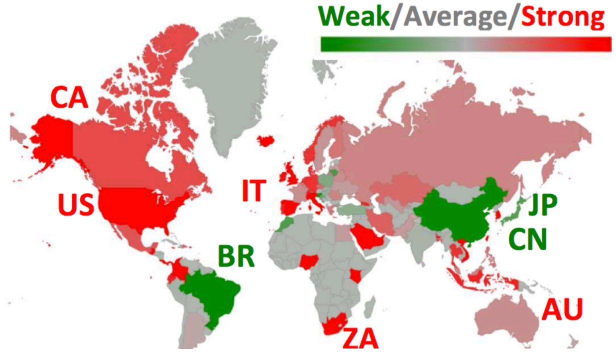
Microsoft

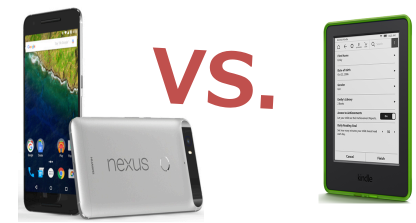
VS.

YAHOO!
NEWS

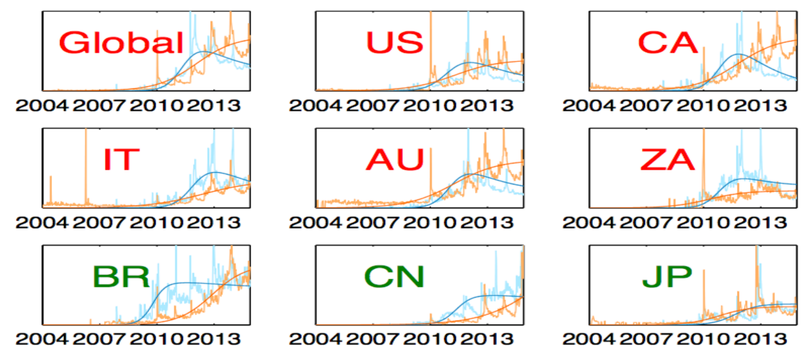


Local competition





VS.



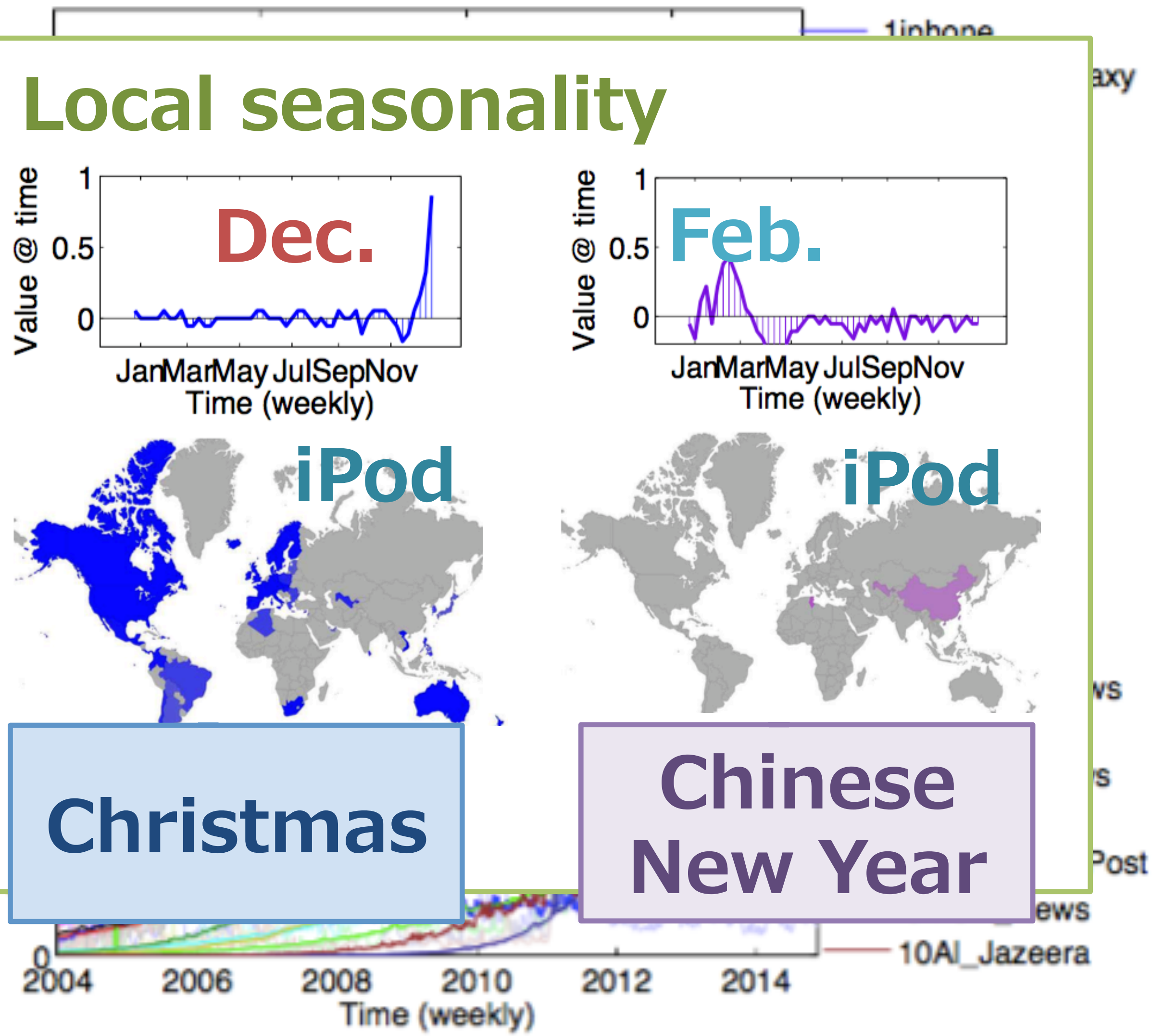
Q1. Effectiveness

1. Products

2. News

Microsoft
VS.
YAHOO! NEWS

Local seasonality



Christmas

Chinese New Year

Q1. Effectiveness

1. Products

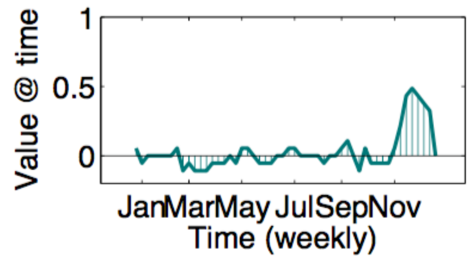
2. News

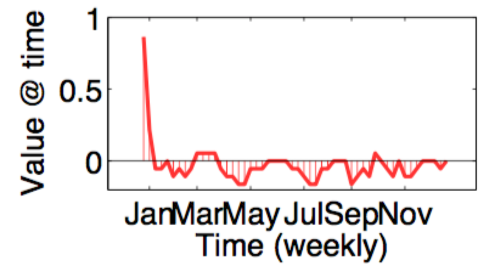
Microsoft

vs.

YAHOO!
NEWS

Local seasonality





Nexus

Kindle

Nexus release

New Year sale

Q1. Effectiveness

1. Products

VS

2. News

Microsoft

VS.

YAHOO!
NEWS

Deltas

Weak/Strong

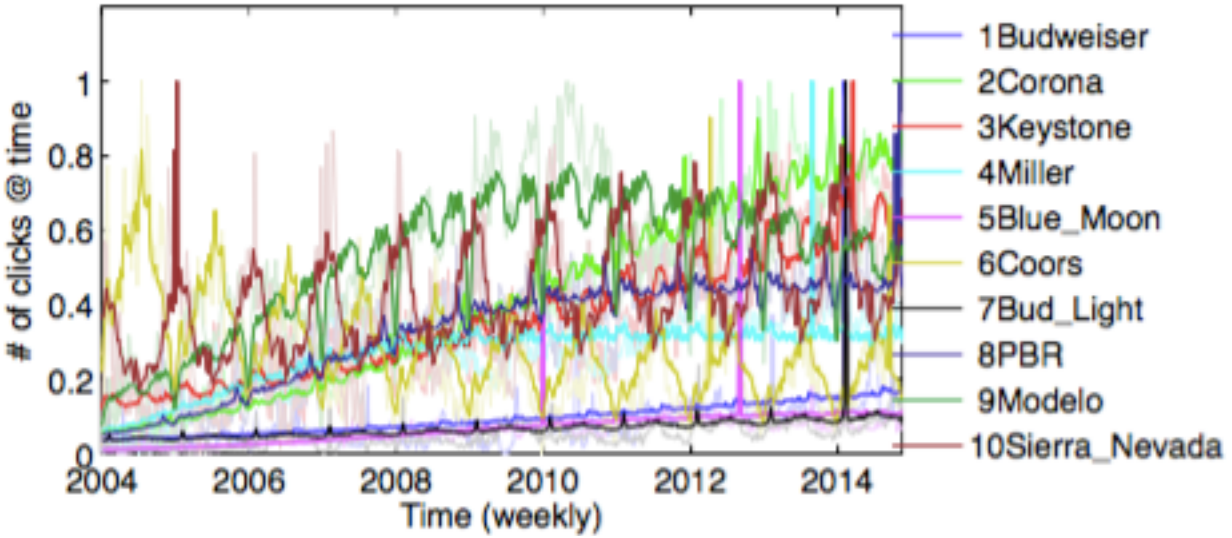
US election Nov. 2008

Wikipedia

Q1. Effectiveness

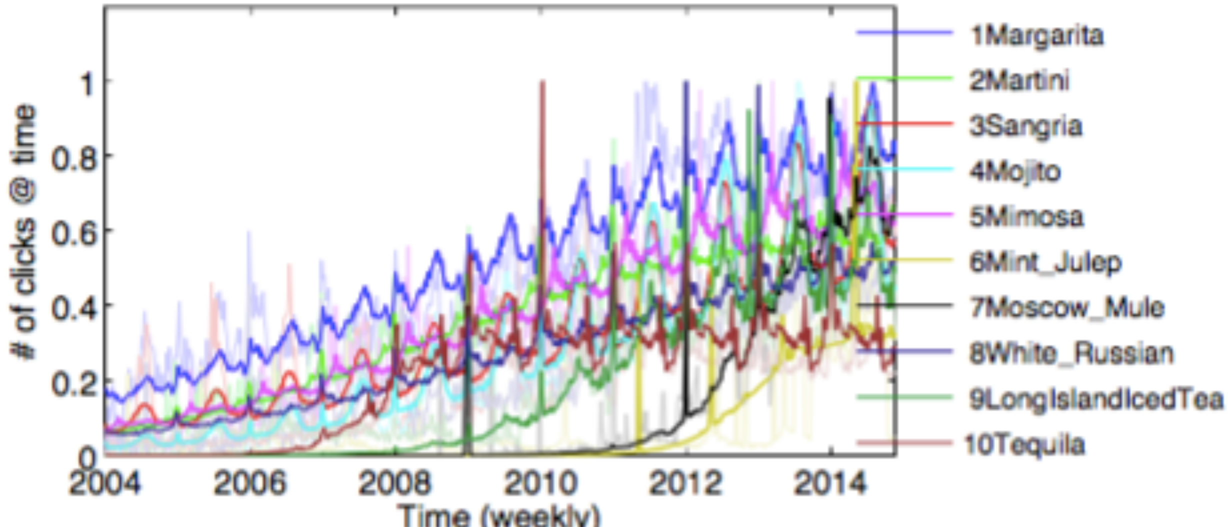
3. Beers

vs.



4. Cocktails

vs.

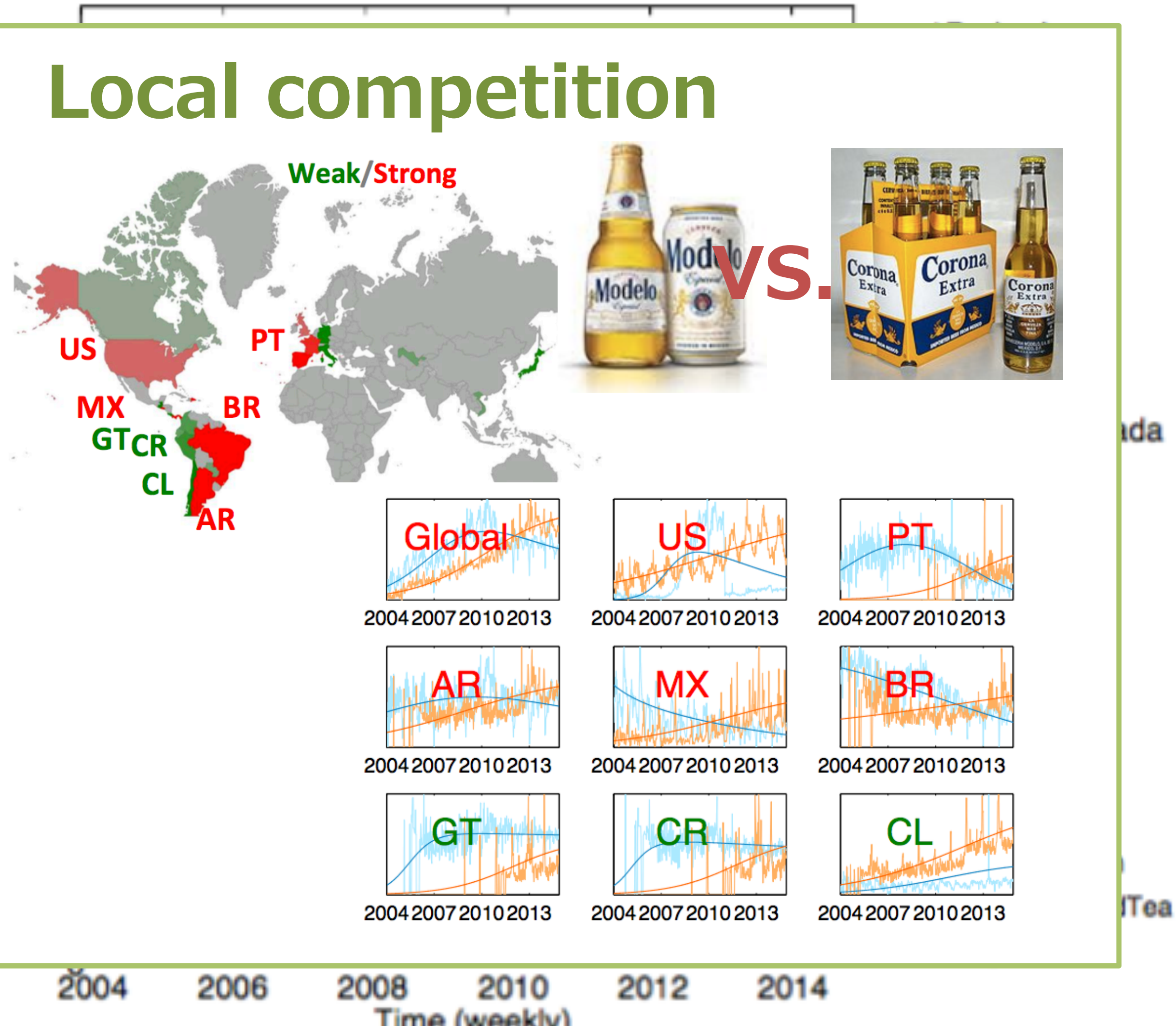


Q1. Effectiveness

3. Beers

4. Cocktails

Local competition

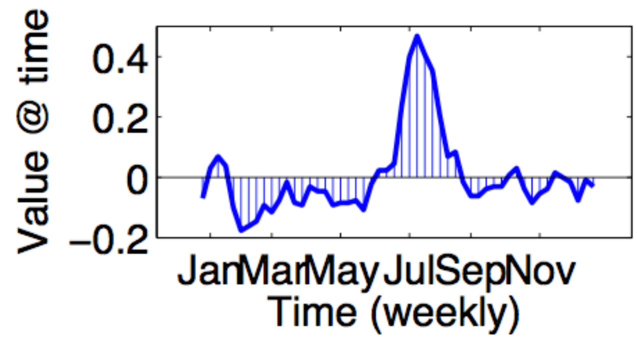


Q1. Effectiveness

3. Beers

4. Cocktails

Local seasonality

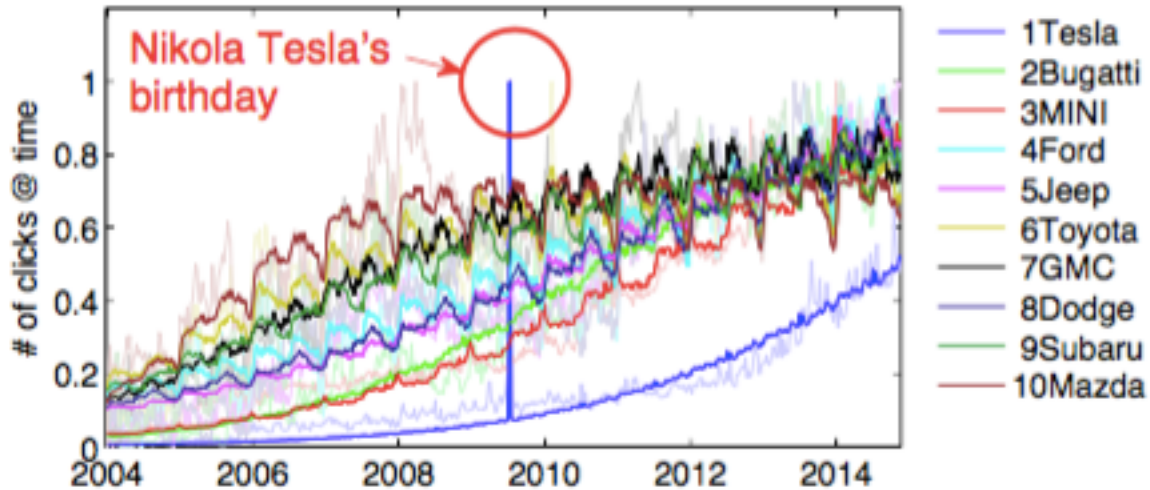


Summer spike for Coors

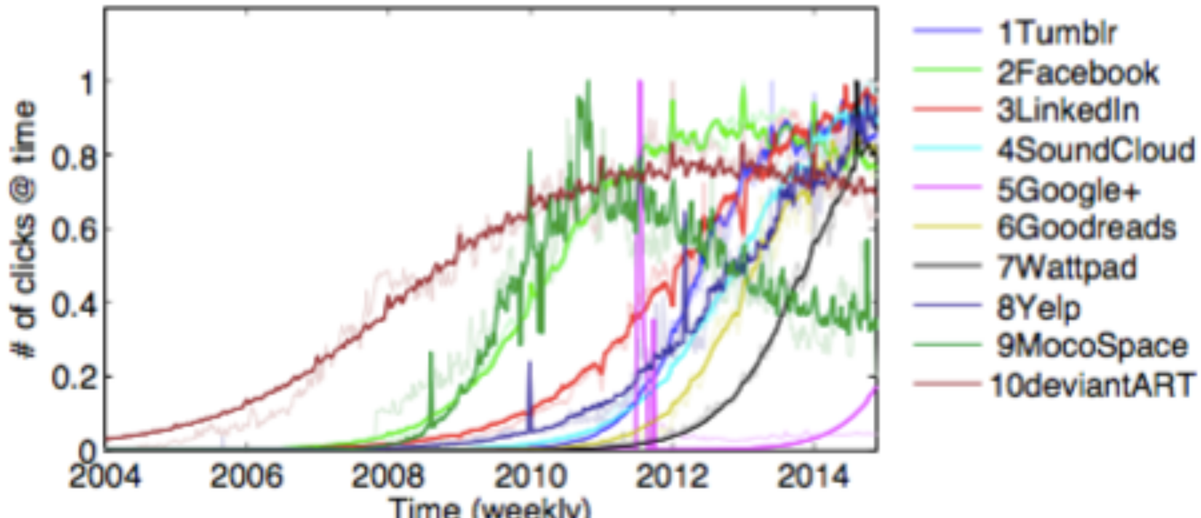
2004 2006 2008 2010 2012 2014
Time (weekly)

Q1. Effectiveness

5. Cars



6. SNS



Q1. Effectiveness

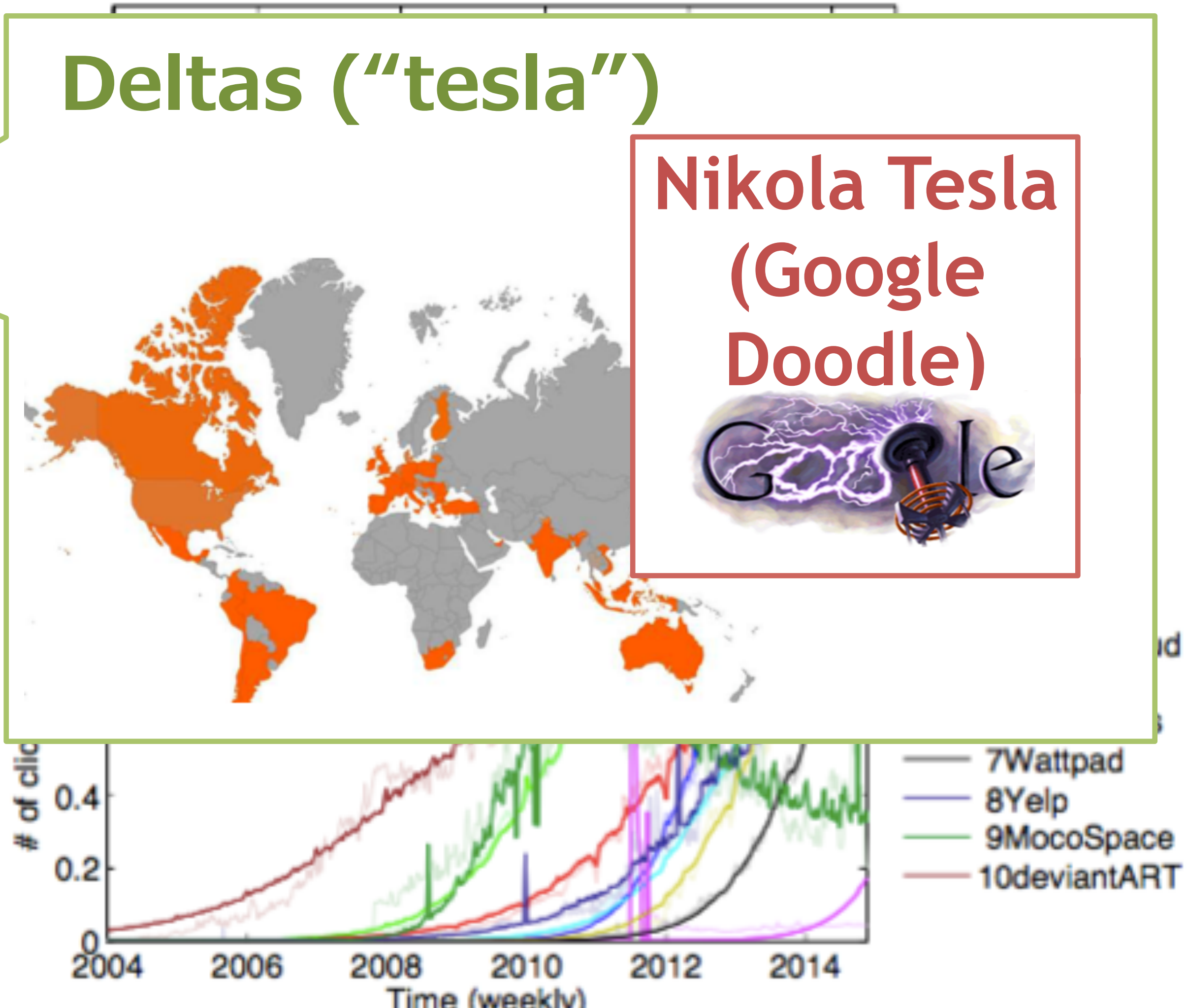
5. Cars

6. SNS

yelp

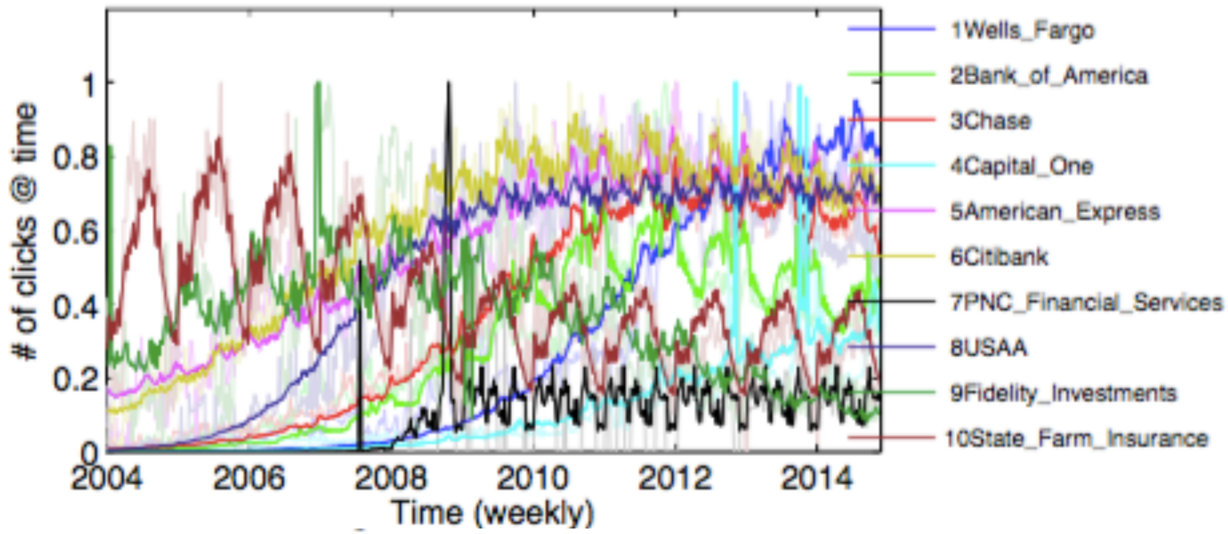
Deltas ("tesla")

Nikola Tesla (Google Doodle)

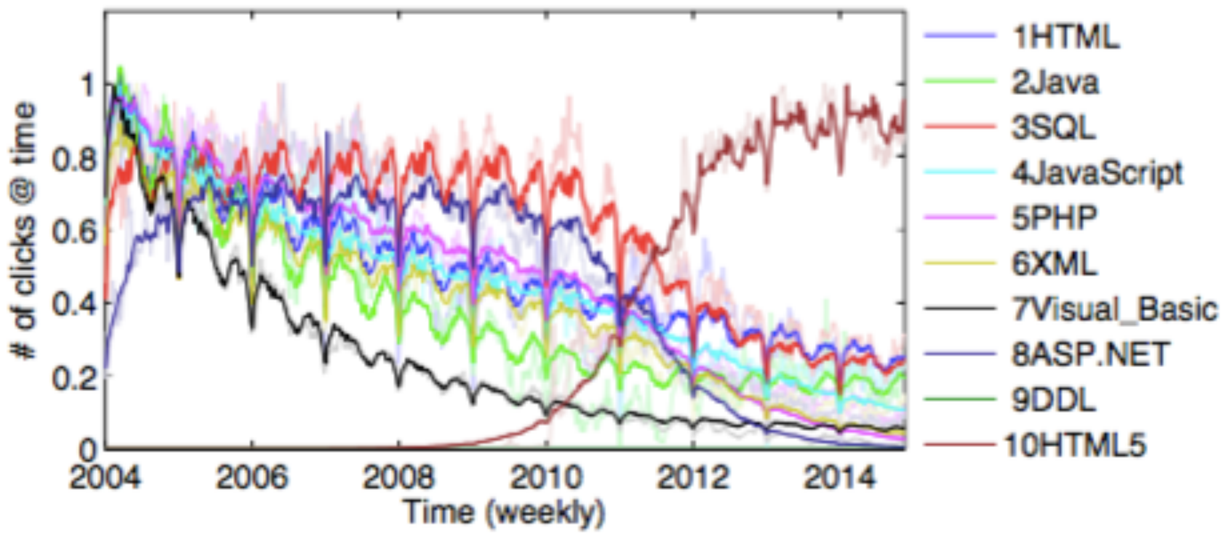


Q1. Effectiveness

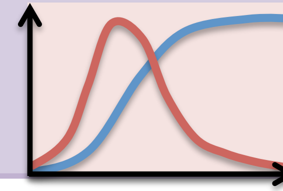
7. Finance



8. Software



Q1. Effectiveness



7. Finance

WELLS FARGO

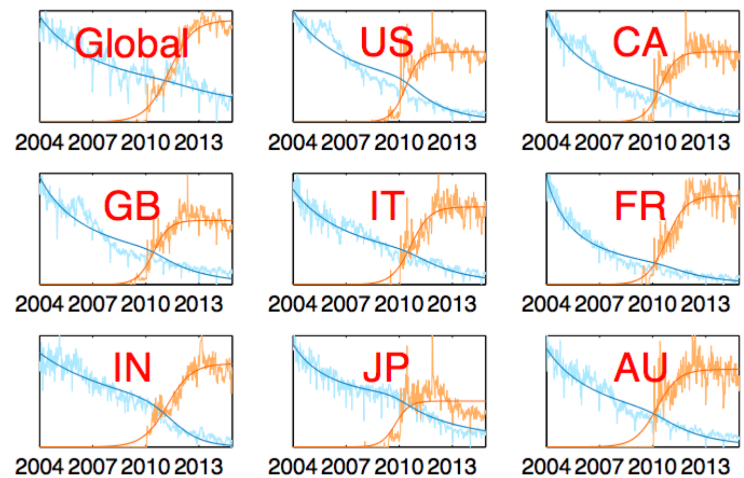
citi CITIBANK

8. Software

HTML

Local competition

HTML vs. HTML5



Q1. Effectiveness

7. Finance

WELLS FARGO

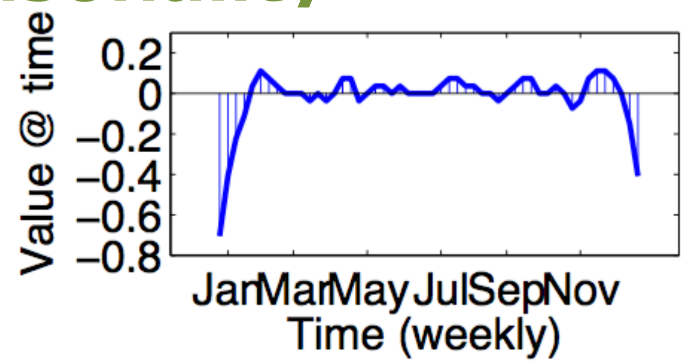
citi CITIBANK

8. Software

HTML

Local seasonality

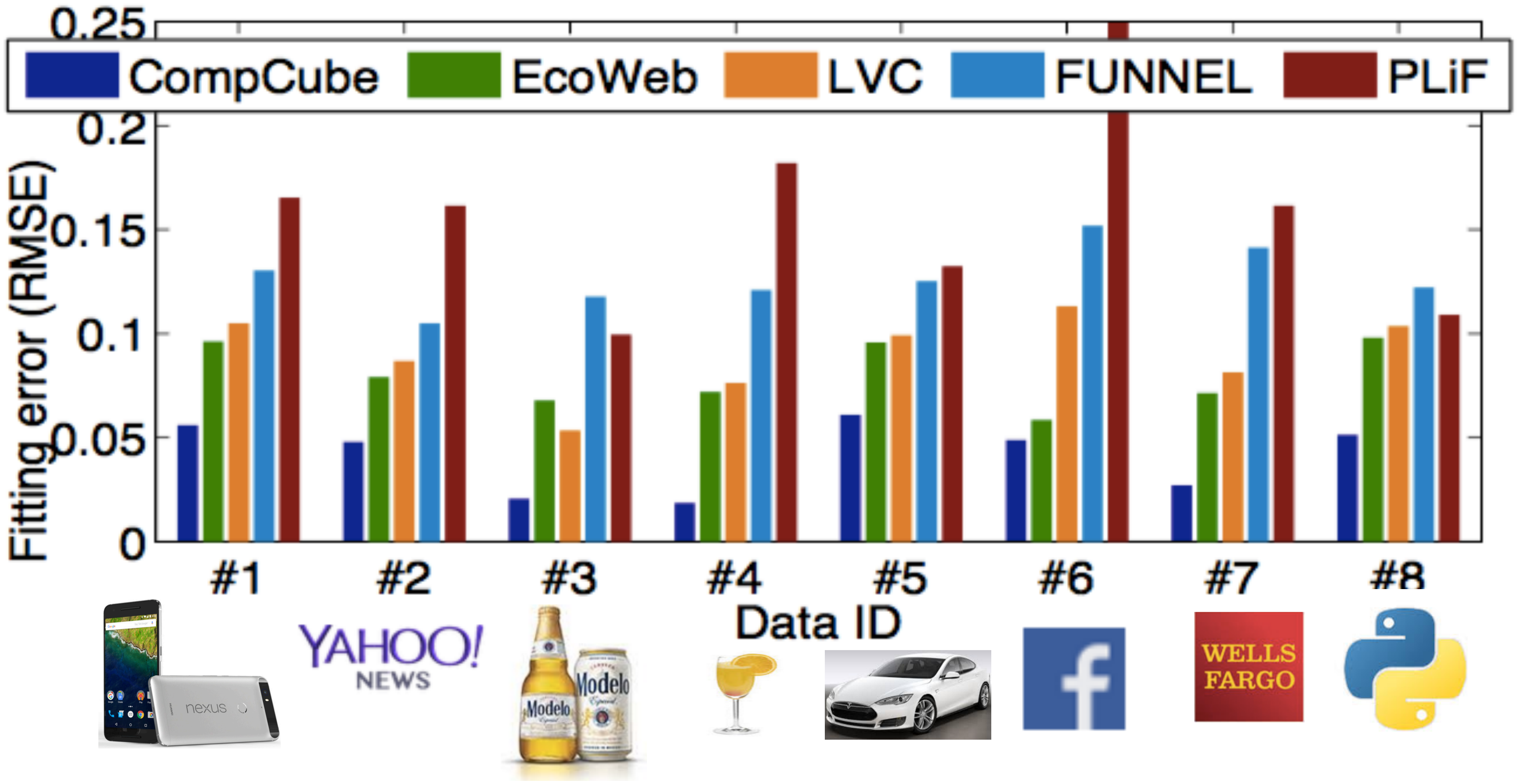
XML



New Year holiday for XML

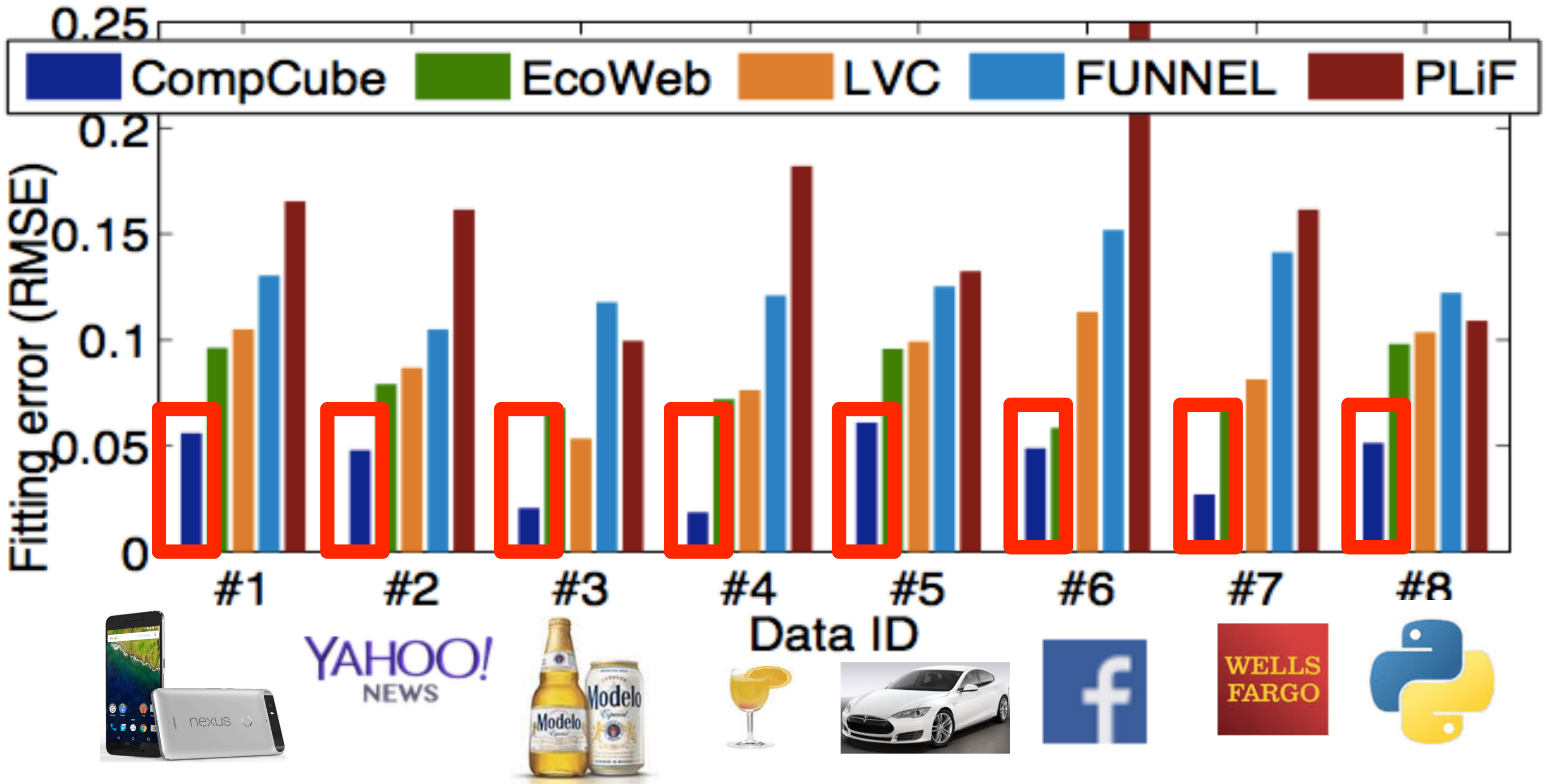
Q2. Accuracy

RMSE between original and fitted volume



Q2. Accuracy

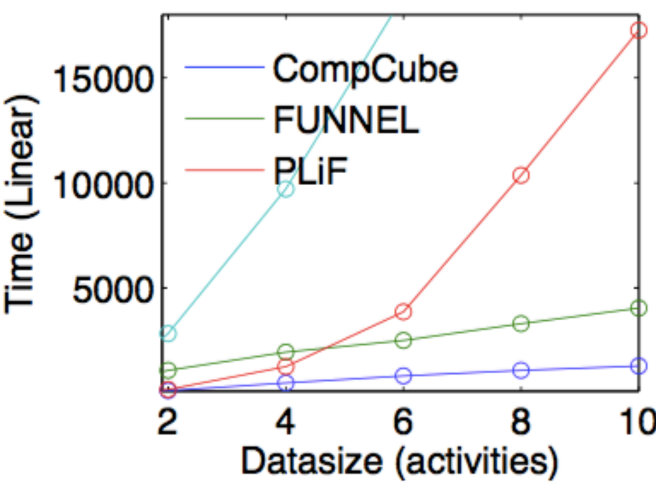
RMSE between original and fitted volume

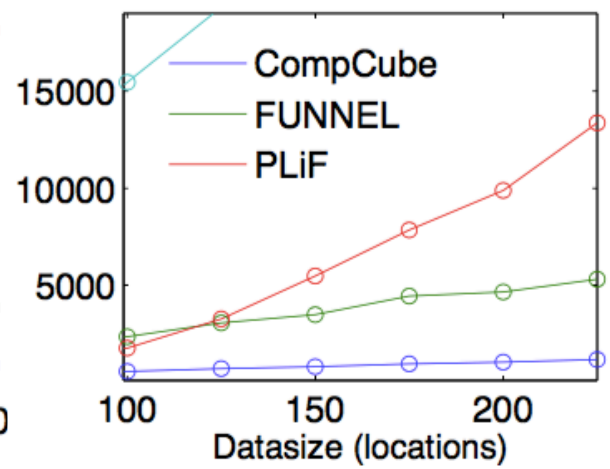


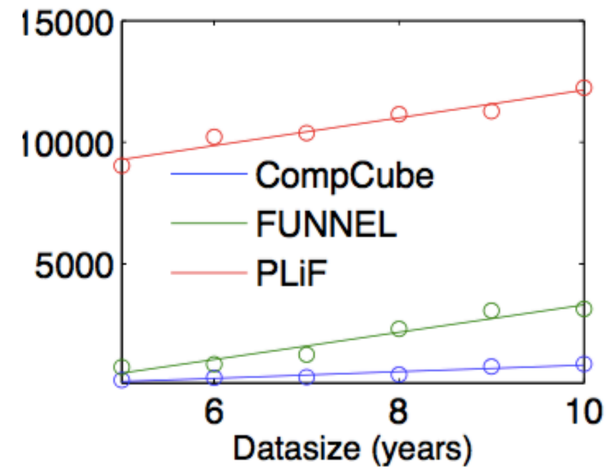
CompCube consistently wins!

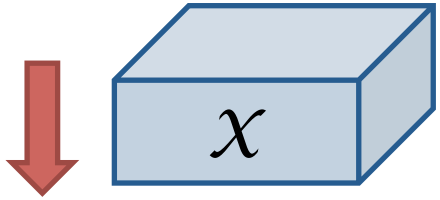
Q3. Scalability

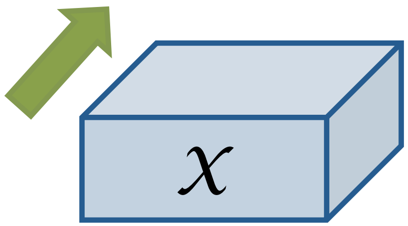
Wall clock time vs. activity , location , Time

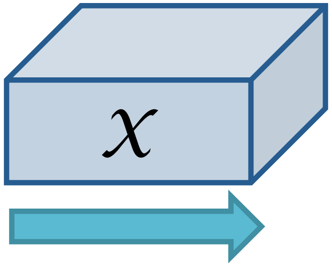






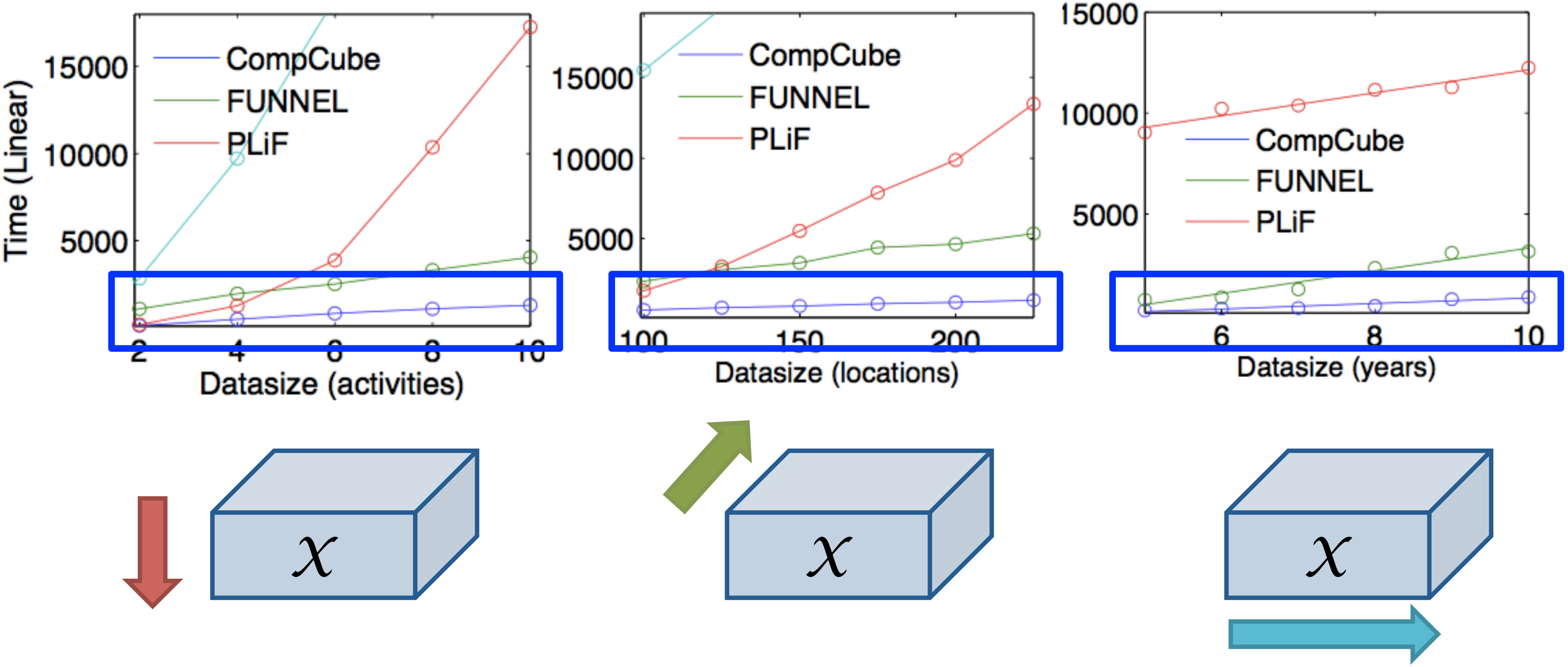






Q3. Scalability

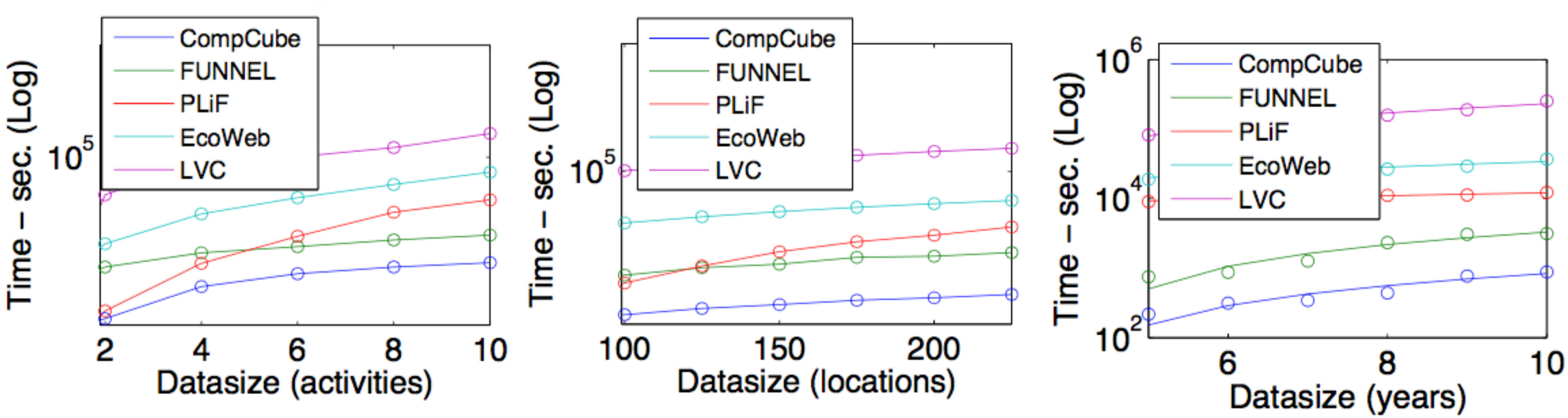
Wall clock time vs. activity , location , Time

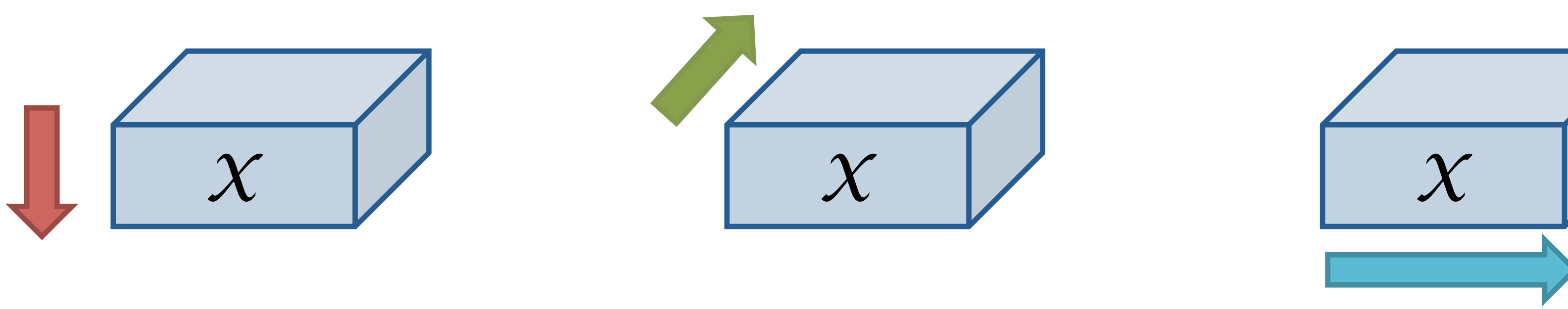


CompCube is linear w.r.t. data size : O(dmn)

Q3. Scalability

Wall clock time vs. activity , location , Time

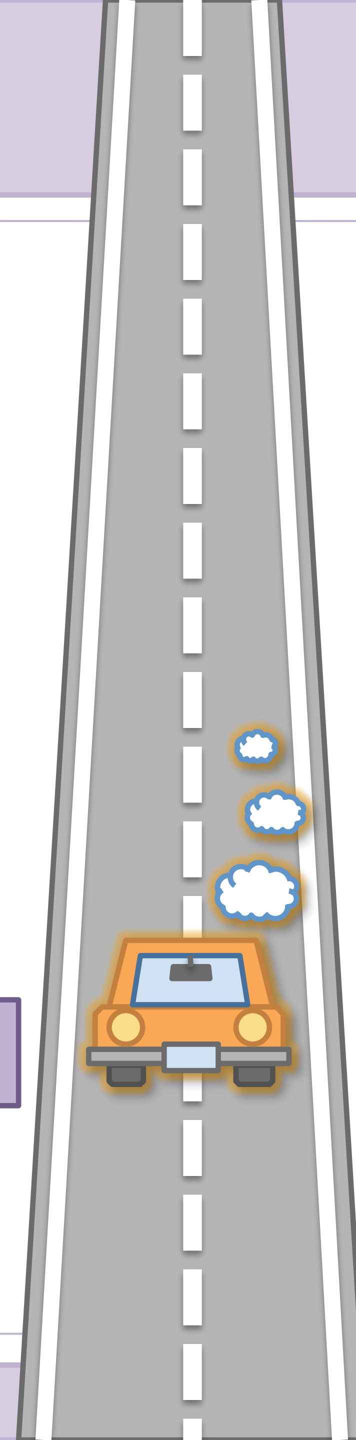




CompCube is linear w.r.t. data size : O(dmn)

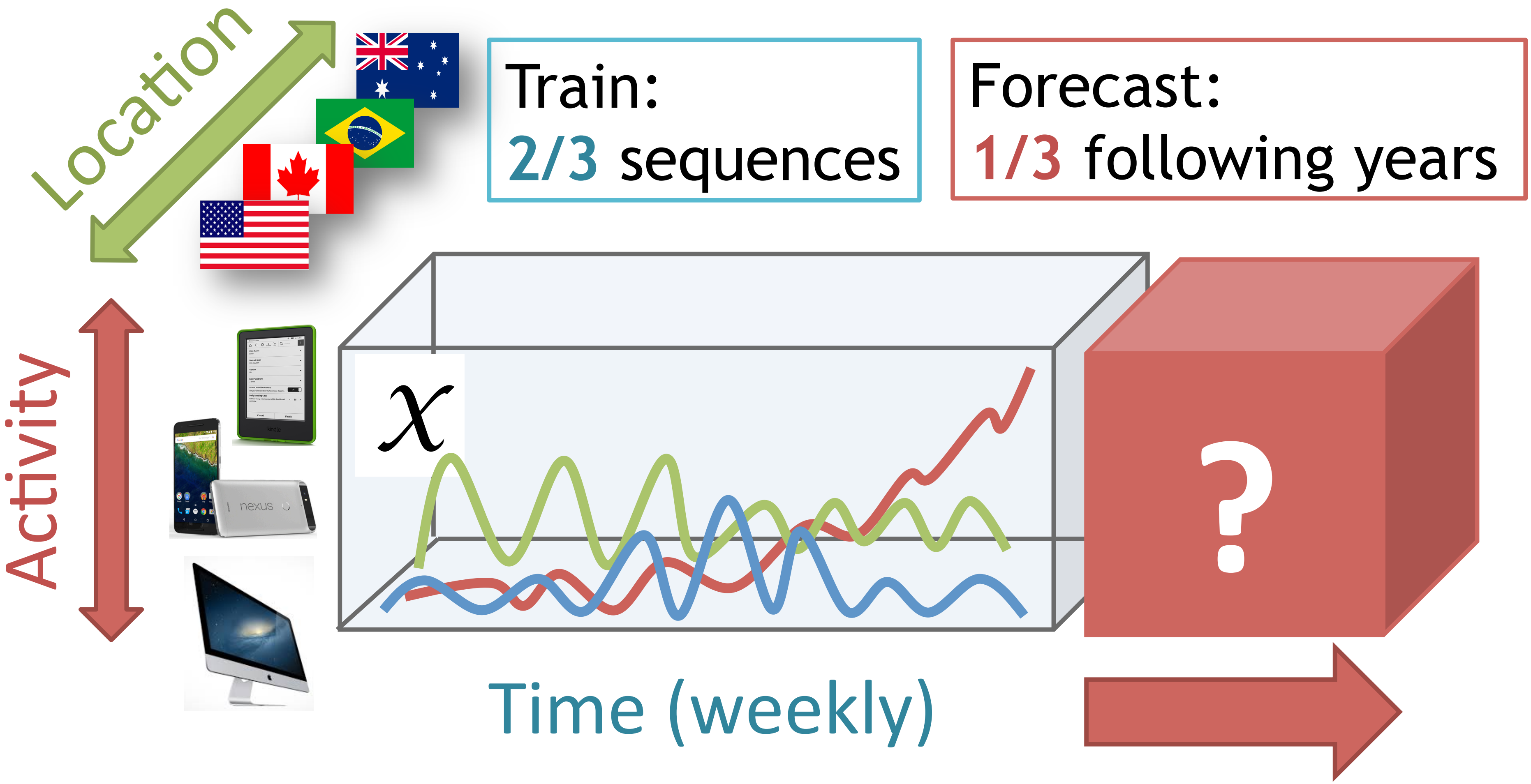
Roadmap

- Motivation
- Modeling power of CompCube
- Overview
- Proposed model
- Algorithm
- Experiments
- CompCube - at work
- Conclusions



CompCube at work - forecasting

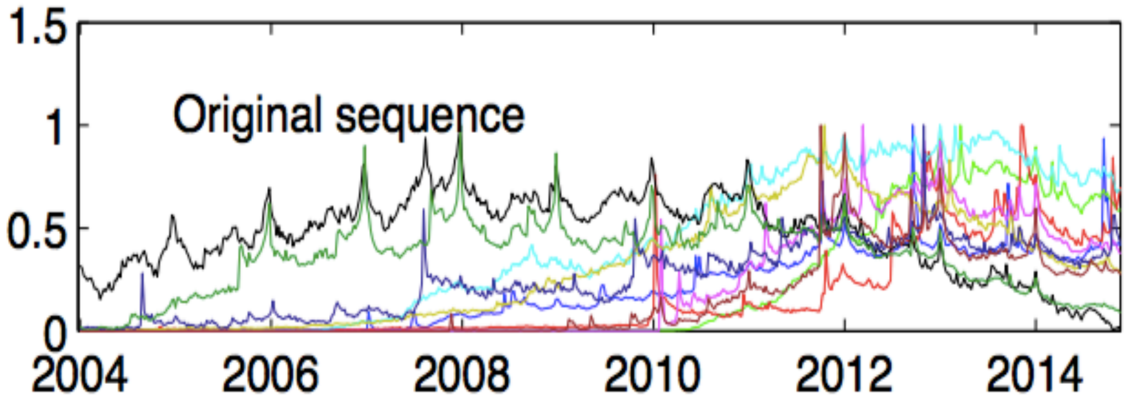
Forecasting future local activities

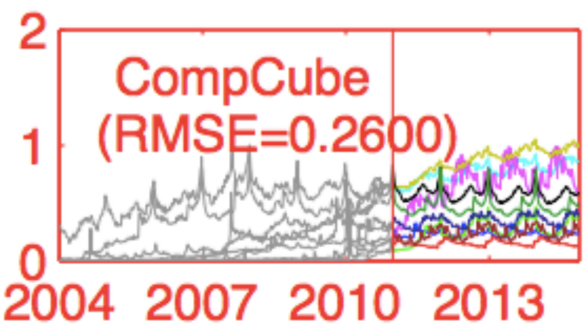


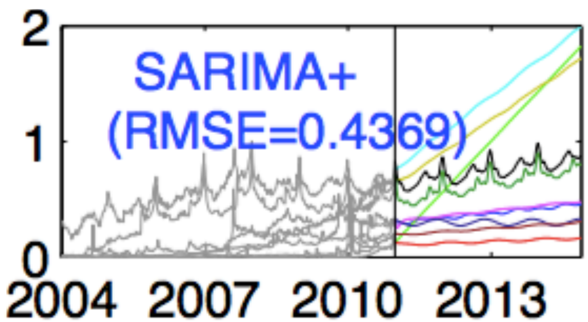
CompCube at work - forecasting

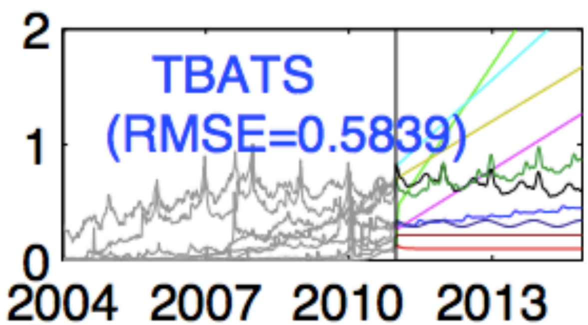
Forecasting results for #1 Products

1. Products





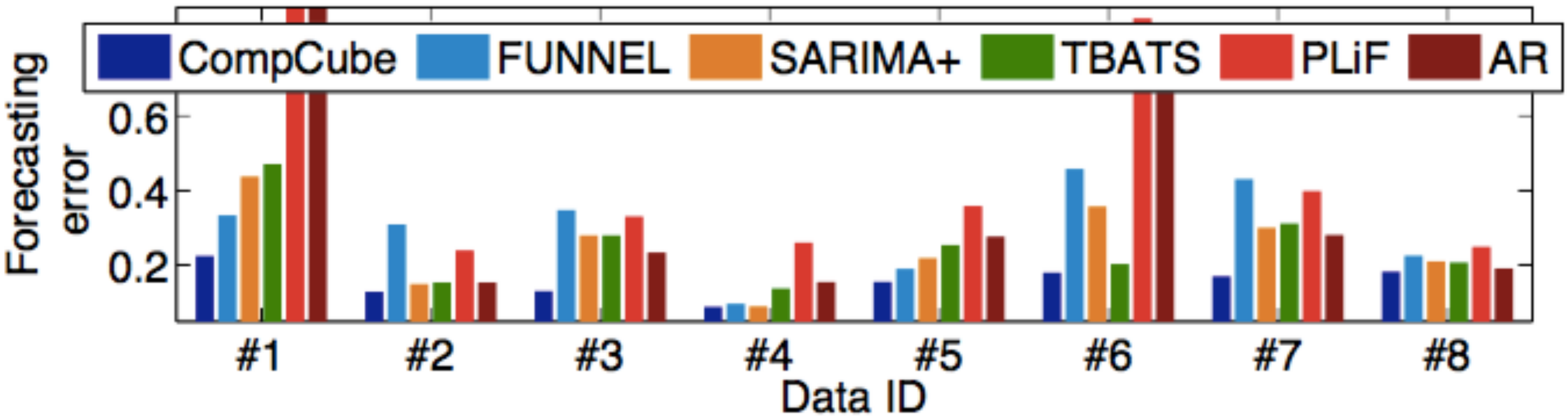




CompCube captures future activities very well

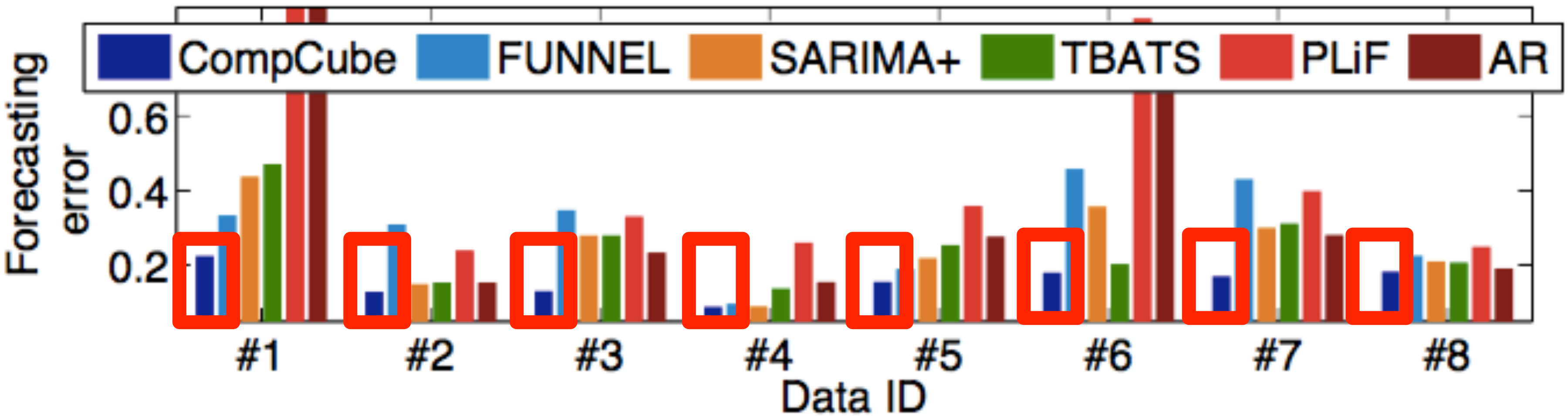
CompCube at work - forecasting

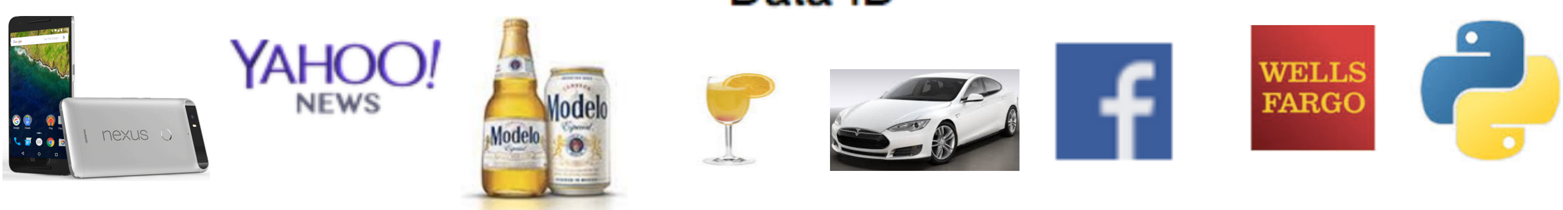
Forecasting error (original vs. forecasts)



CompCube at work - forecasting

Forecasting error (original vs. forecasts)

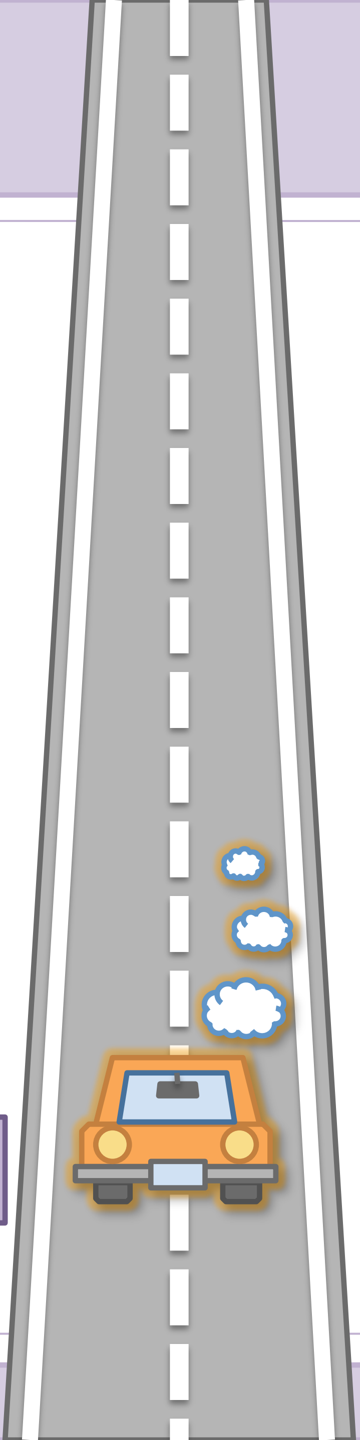




CompCube consistently wins!

Roadmap

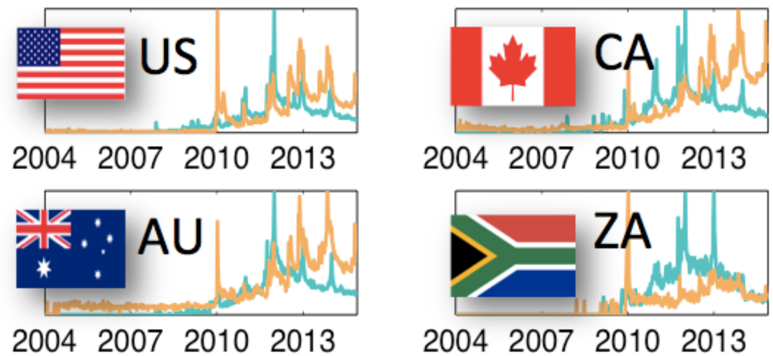
- ✔ Motivation
- ✔ Modeling power of CompCube
- ✔ Overview
- ✔ Proposed model
- ✔ Algorithm
- ✔ Experiments
- ✔ CompCube – at work
- Conclusions

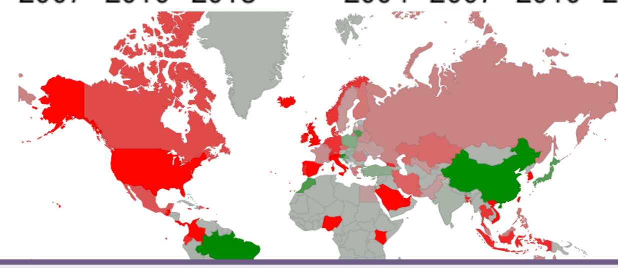


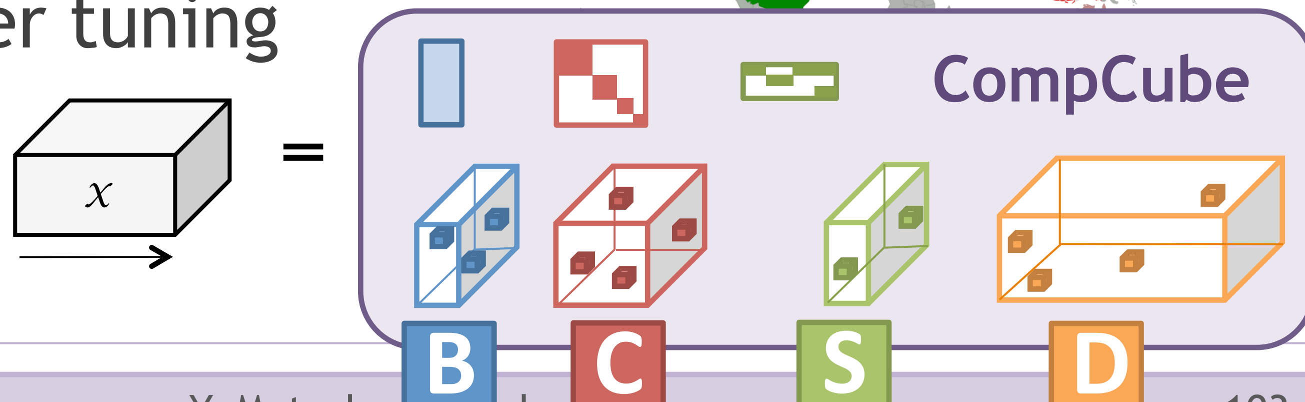
Conclusions

CompCube has the following advantages

- ✔ **Effective**
  Finds important patterns
- ✔ **Practical**
  Long-range forecasting
- ✔ **Parameter-free**
  No parameter tuning
- ✔ **Scalable**
  It is linear

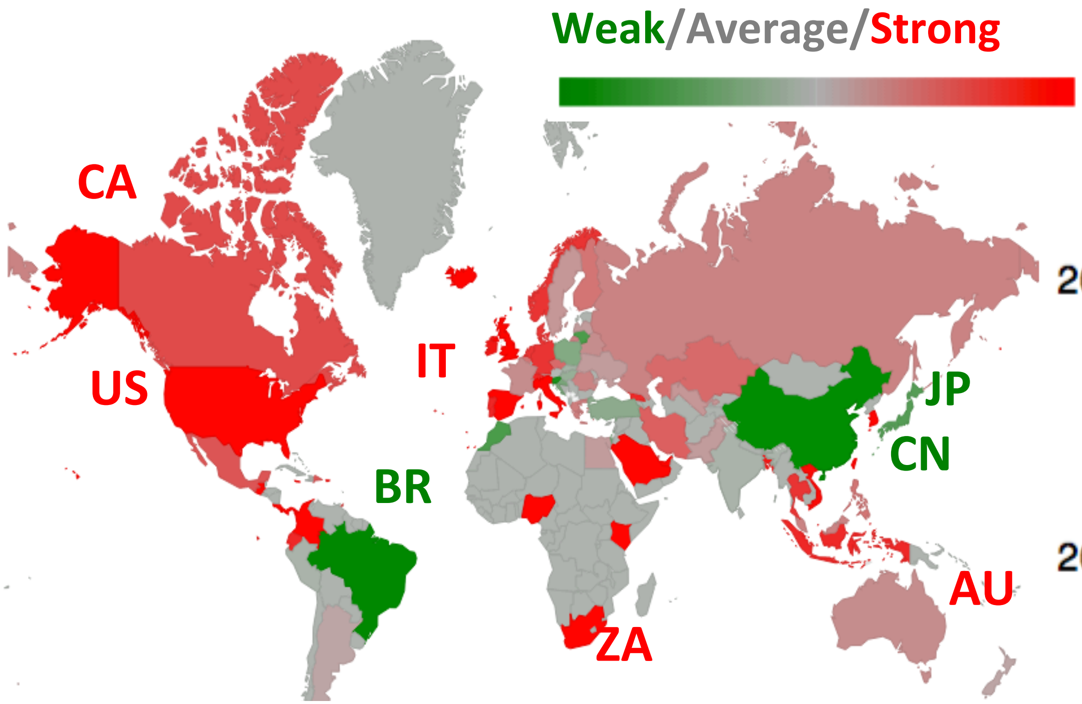


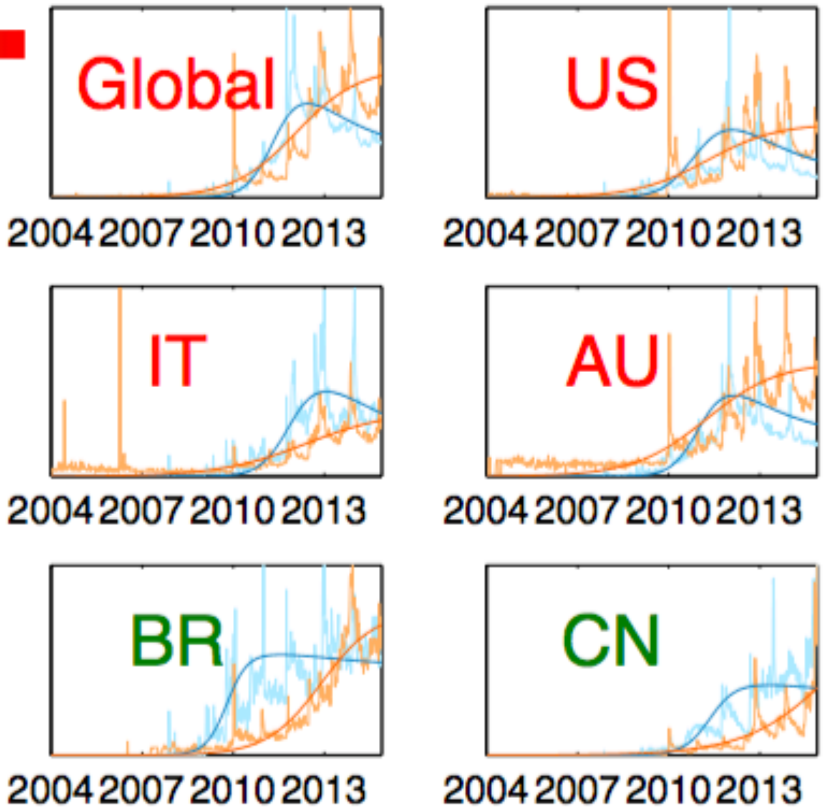




Thank you!







Data & Code:

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

Non-linear Mining of Competing Local Activities

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