ACM SIGKDD 2016

22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
August 13-17, 2016 | San Francisco, California

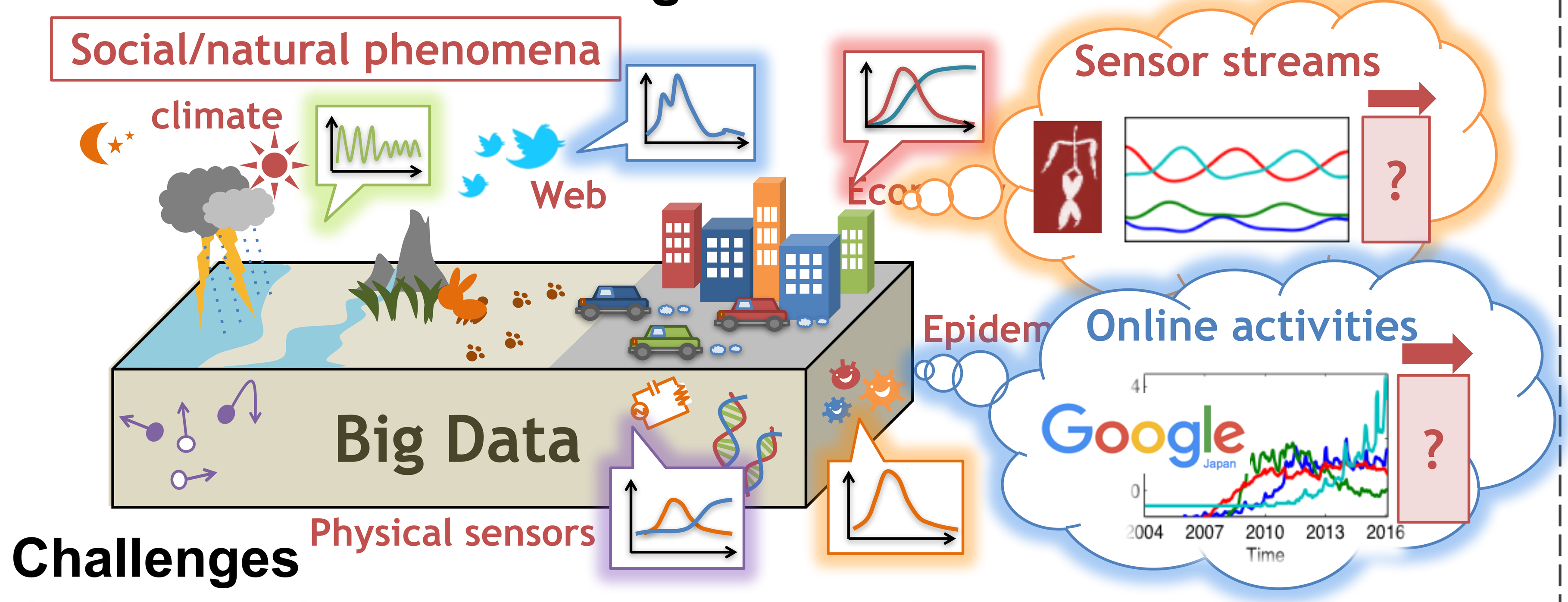
Regime Shifts in Streams: Real-time Forecasting of Co-evolving Time Sequences

Yasuko Matsubara
Kumamoto University
yasuko@cs.kumamoto-u.ac.jp

Yasushi Sakurai
Kumamoto University
yasushi@cs.kumamoto-u.ac.jp

Kumamoto University

Motivation - Given: Big time-series data streams



Challenges

Q. Can we forecast l_s-steps-ahead future events?

Given: Co-evolving event stream $X = \{\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(t_c), \ldots\}$

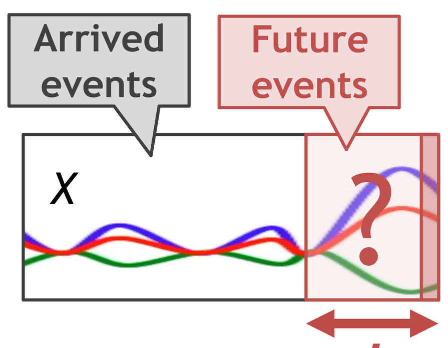
Goal: Forecast l_s-steps-ahead future events , at any point in time

Real-time forecasting over data streams

Requirements

(a) l_s-steps-ahead forecasting

- Predict l_s-steps-ahead (long-term) events
- Capture dynamic patterns, continuously



(b) Adaptive non-linear modeling

- Non-linear dynamical systems
- Regime shifts (ecosystems)

Long-term | Continuous | Non-linear | Adaptive

Regime shifts in natural systems

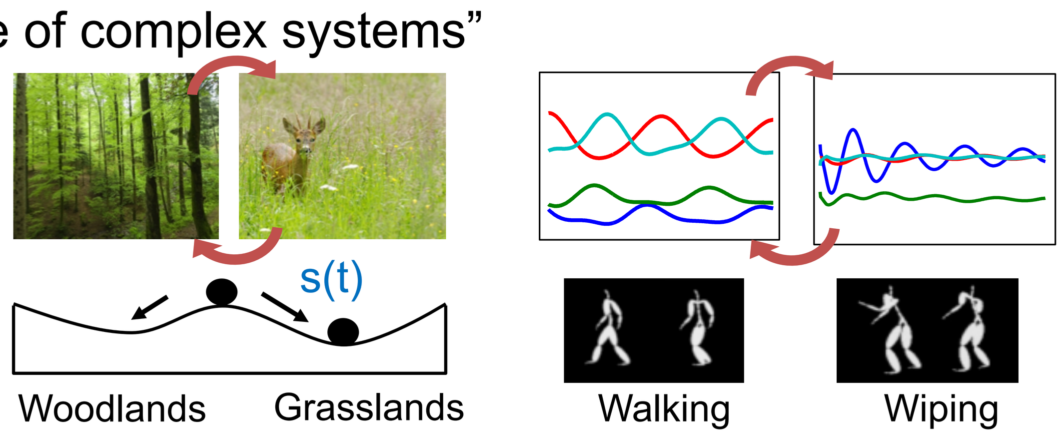
"Abrupt changes in the structure of complex systems"

Examples:

- Woodland vs. grassland
- Corel vs. macro algae
- Desert vs. vegetation

s(t) :ecosystem property (nutrients/soils)

$$\frac{ds(t)}{dt} = a_0 + a_1 s(t) + a_2 f(s(t))$$



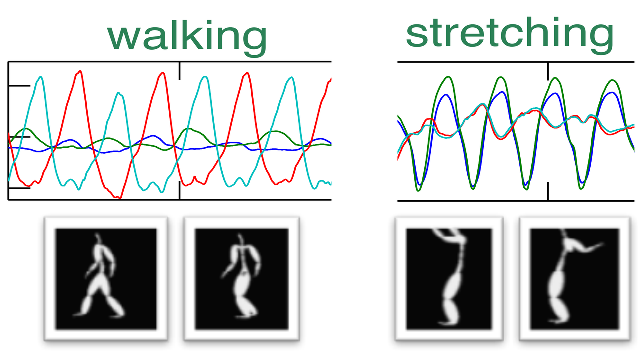
Ecological system **Sensor stream**

Image courtesy of dan at FreeDigitalPhotos.net.

Proposed model

Main idea (P1): Latent non-linear dynamics

Various patterns ("**regimes**") in streams

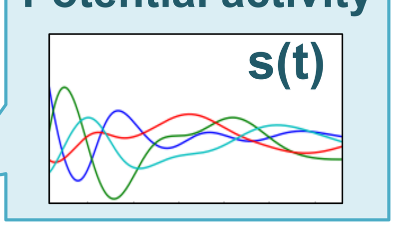


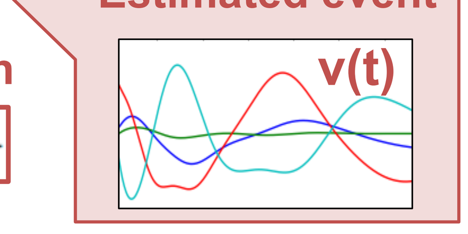
$$\frac{d\boldsymbol{s}(t)}{dt} = \boldsymbol{p} + \mathbf{Q}\boldsymbol{s}(t) + \mathcal{A}\mathbf{S}(t)$$

$$\boldsymbol{v}(t) = \boldsymbol{u} + \mathbf{V}\boldsymbol{s}(t)$$

Linear, Exponential, Non-linear, Projection

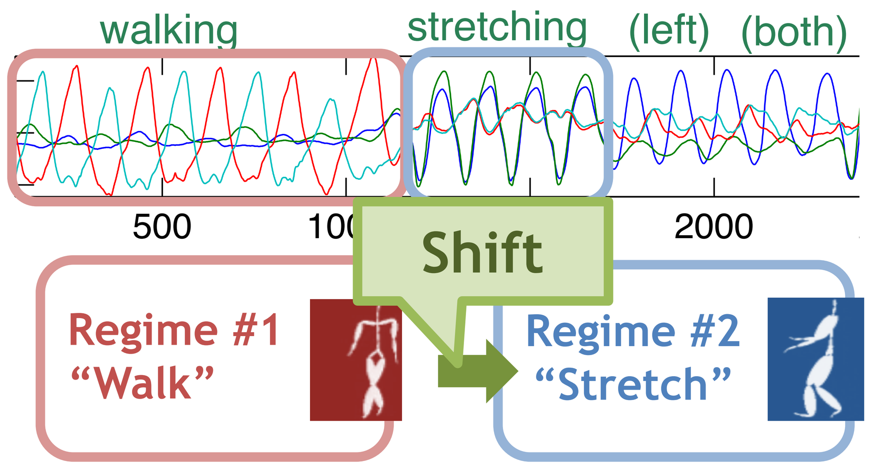
$$\boldsymbol{\theta} = \{\boldsymbol{s}_0, \boldsymbol{p}, \mathbf{Q}, \mathcal{A}, \boldsymbol{u}, \mathbf{V}\}$$





Main idea (P2): Regime shifts in streams

Q. How can we identify any sudden discontinuity? - A. Regime shifts in streams

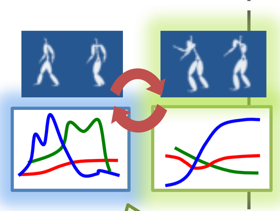


$$\frac{d\boldsymbol{s}_i(t)}{dt} = \boldsymbol{p}_i + \mathbf{Q}_i\boldsymbol{s}_i(t) + \mathcal{A}_i\mathbf{S}_i(t) \quad (i = 1, \ldots, c)$$

$$\frac{d\boldsymbol{w}(t)}{dt} = \boldsymbol{r}(t), \quad \mathbf{R} = \{\boldsymbol{r}(t)\}_{t=1}^{t_c}$$

$$\boldsymbol{v}(t) = \sum_{i=1}^{c} w_i(t)\,[\boldsymbol{u}_i + \mathbf{V}_i\boldsymbol{s}_i(t)]$$

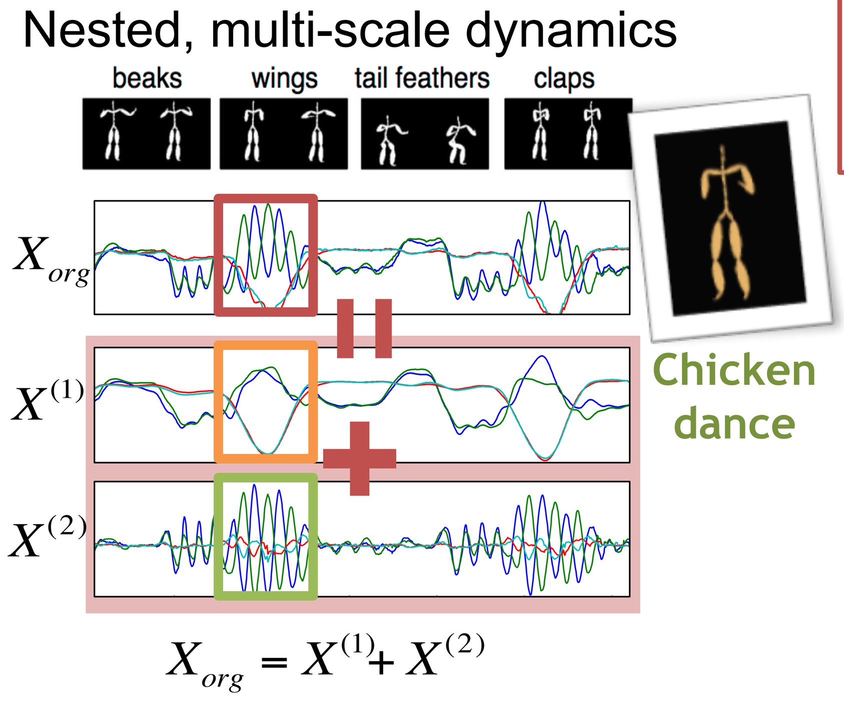
$$\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_c, \mathbf{R}\}$$



R: Regime shift dynamics (c: # of regimes)

Main idea (P3): Nested structure

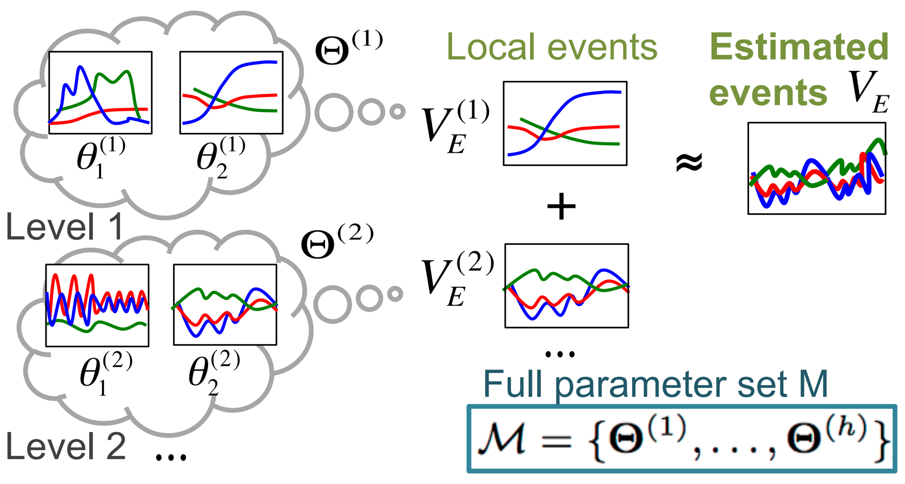
Nested, multi-scale dynamics



Original events: $X^{(1)}$: Long-term + $X^{(2)}$: Short-term

e.g., Tail feathers = bending knees, once + moving arms, quickly

Multi-level modeling structure



Full parameter set M

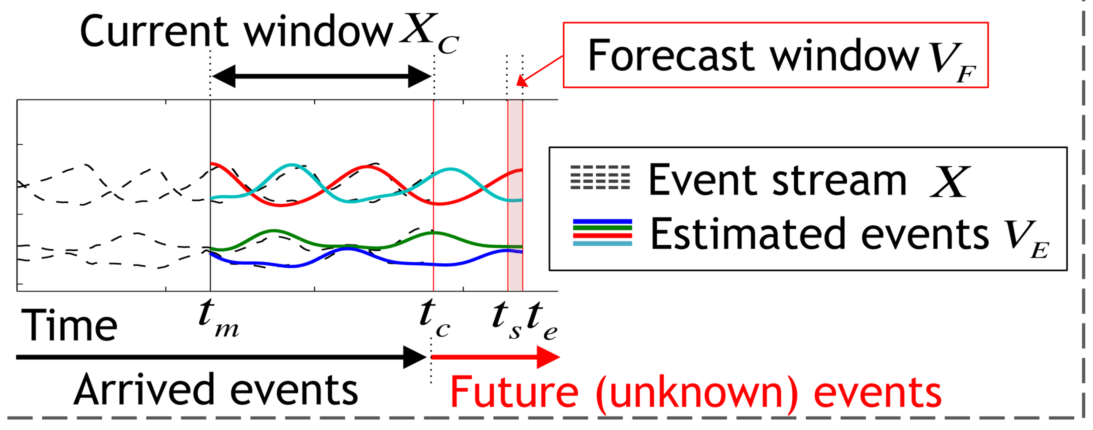
$$\mathcal{M} = \{\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(h)}\}$$

Problem definition

Given: Current window X_c

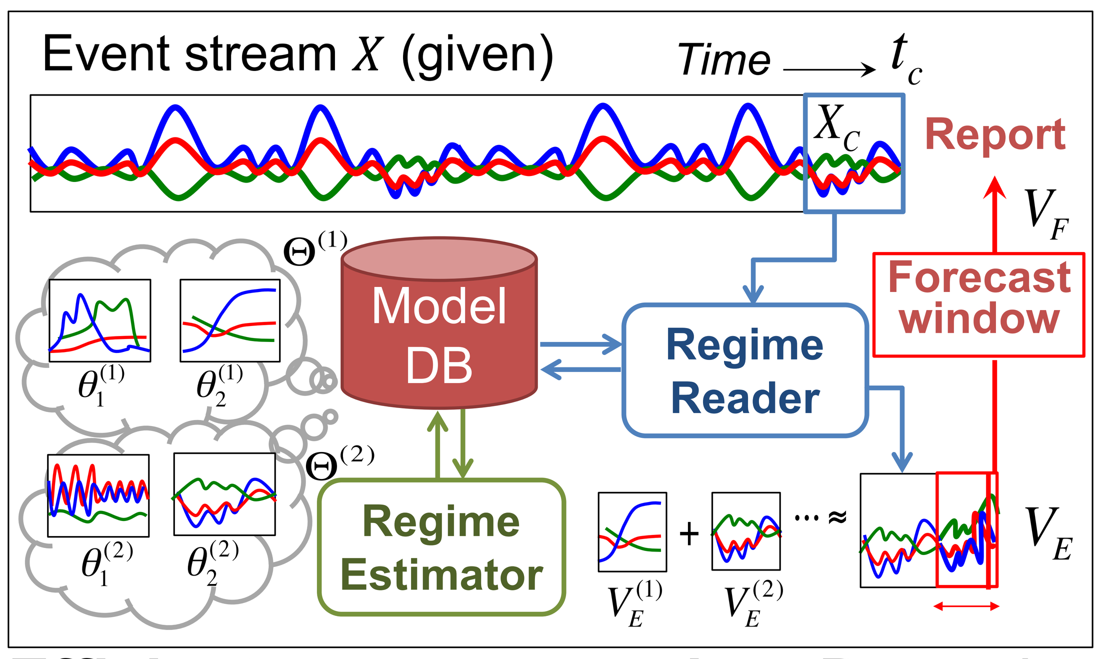
Find: Estimated events V_E

Forecast window V_F



Proposed algorithms

- **A1 RegimeCast** : Report l_s-steps-ahead future events
  - **A2 RegimeReader** : Identify current regime dynamics
  - **A3 RegimeEstimator** : Estimates regime parameter set θ



Step1: Extract current window X_c

Step2: Find optimal regime/model

Step3: Estimate future events

Step4: (optional) Estimate new regime

Step5: Report future events V_F

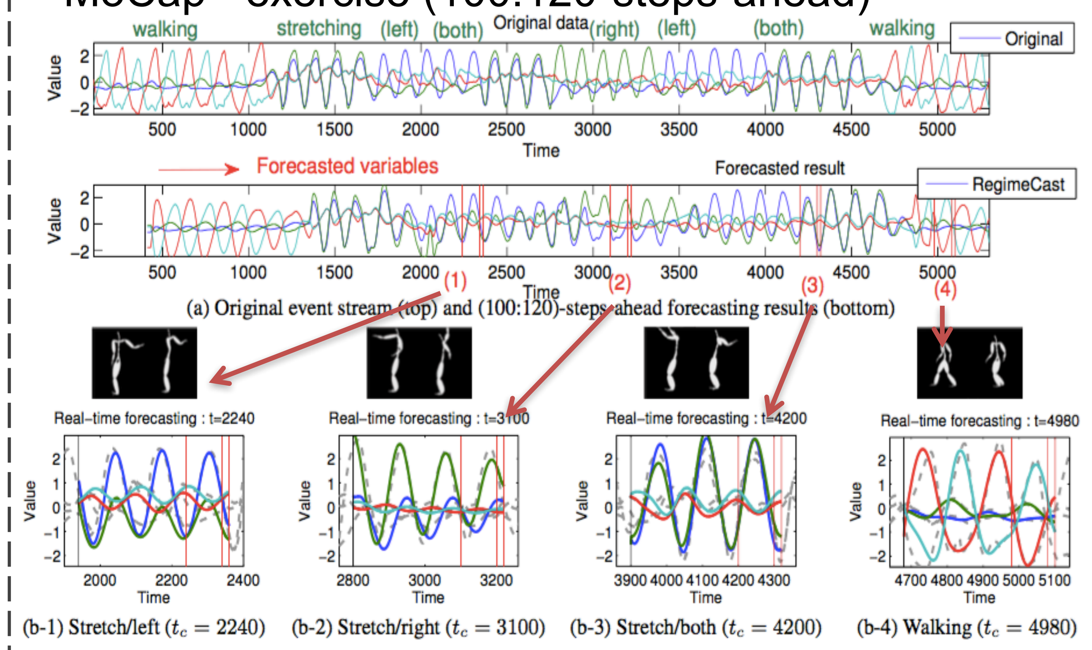
Efficient event generation: Dynamic point sets (DPS)

Scalability (RegimeCast): at least $O(c \cdot l_e / \delta)$ at most $O(c \cdot l_e / \delta + l_c)$

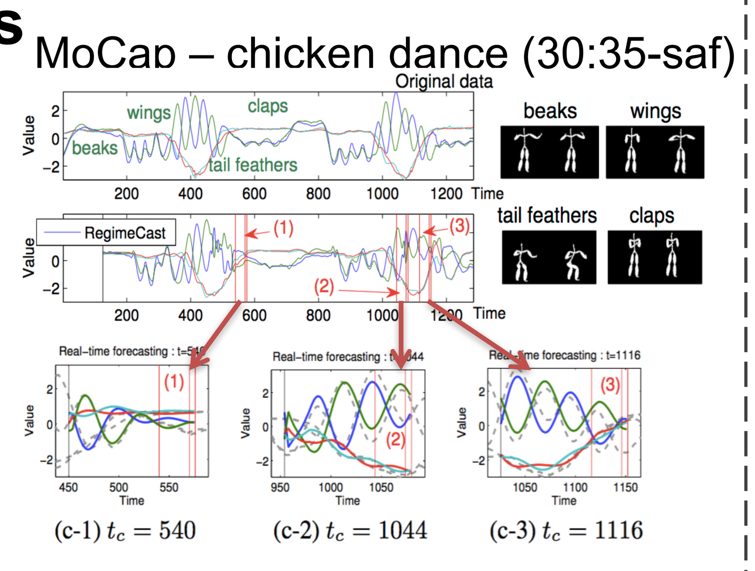
Details in paper - c : # of regimes, l_e : Length of V_E, δ : DPS interval, l_c : Length of X_c

Experiments – Q1. Effectiveness

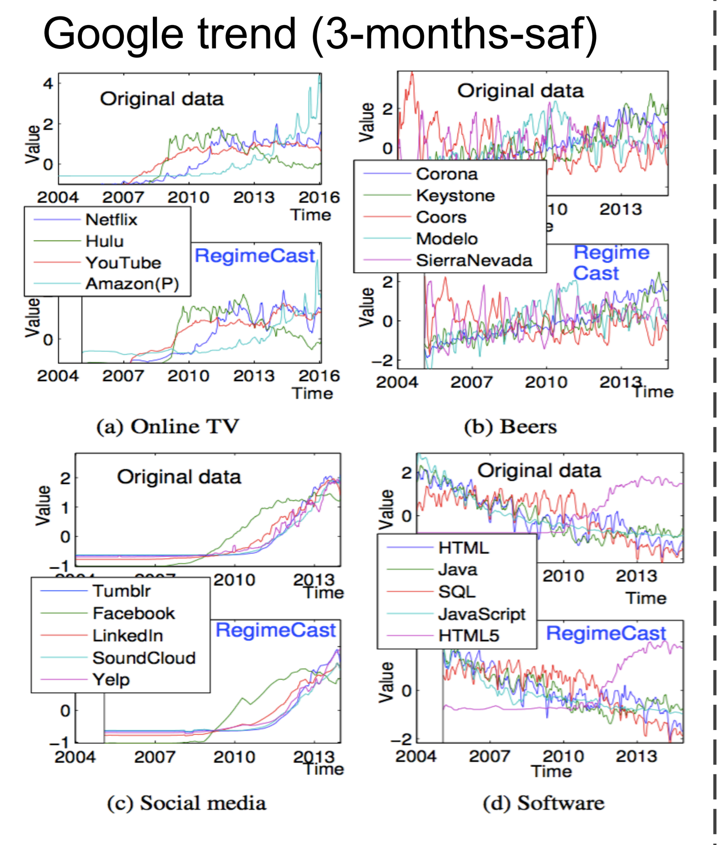
MoCap - exercise (100:120-steps-ahead)



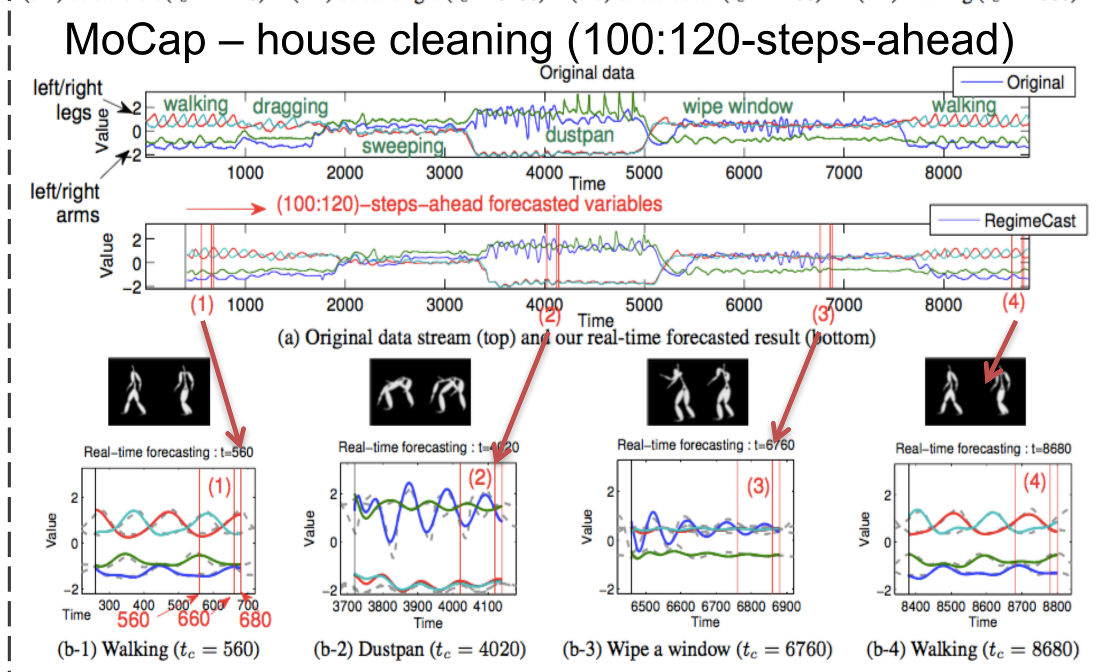
MoCap – chicken dance (30:35-saf)



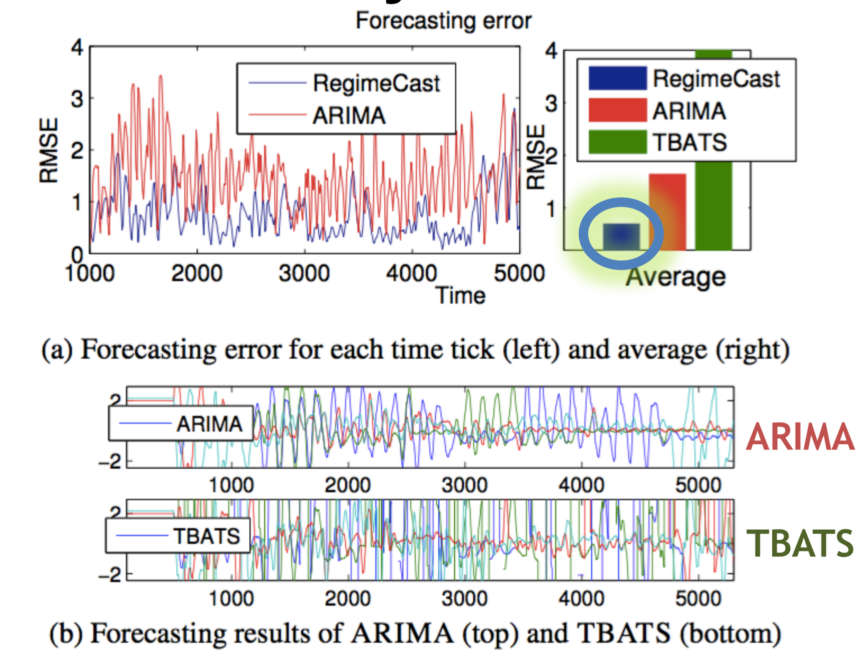
Google trend (3-months-saf)



MoCap – house cleaning (100:120-steps-ahead)



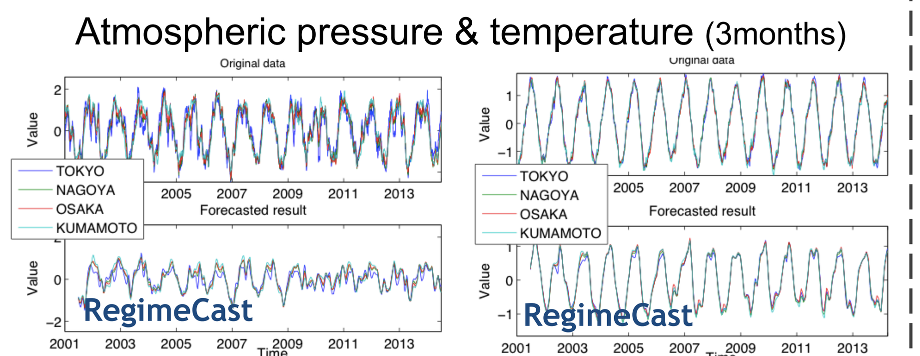
Q2. Accuracy



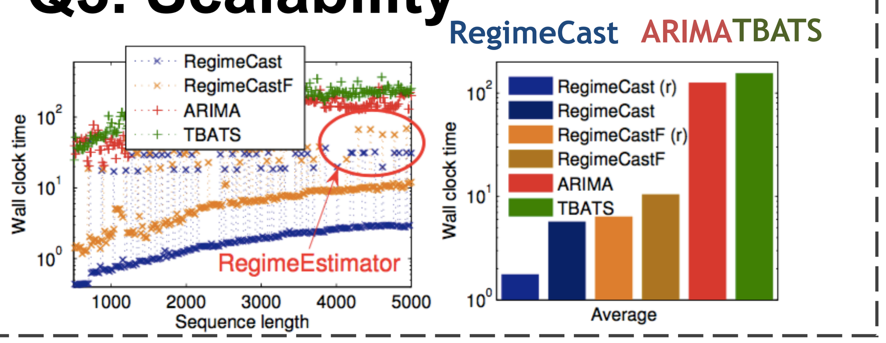
(a) Forecasting error for each time tick (left) and average (right)

(b) Forecasting results of ARIMA (top) and TBATS (bottom)

Atmospheric pressure & temperature (3months)



Q3. Scalability



Conclusions - RegimeCast has following advantages:

- **Effective & Adaptive:** long-term forecasting / no prior training
- **Scalable:** it does not depend on data length
- **Anytime:** it can forecast future events, immediately, any time

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