Mining and Forecasting of Big Time-series Data

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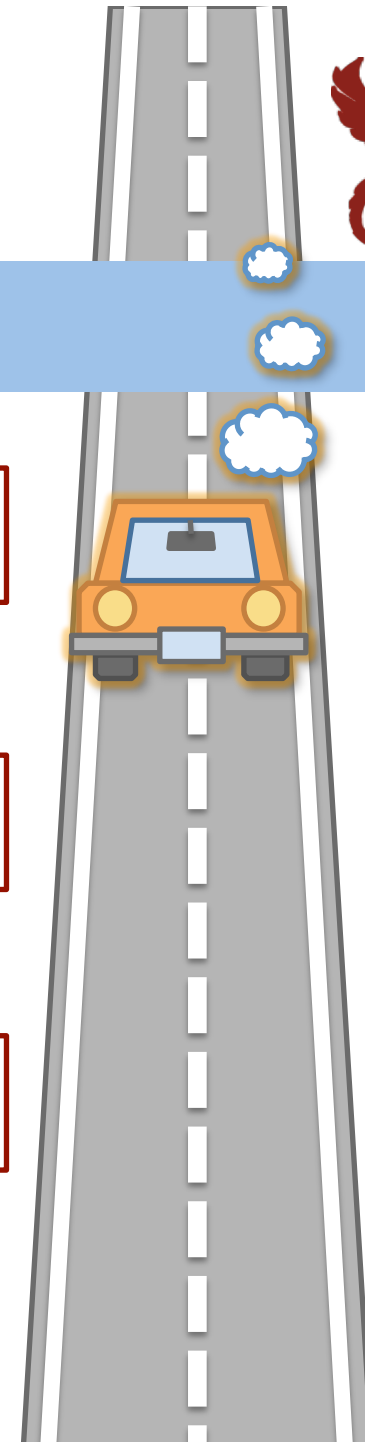
Roadmap

- 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**
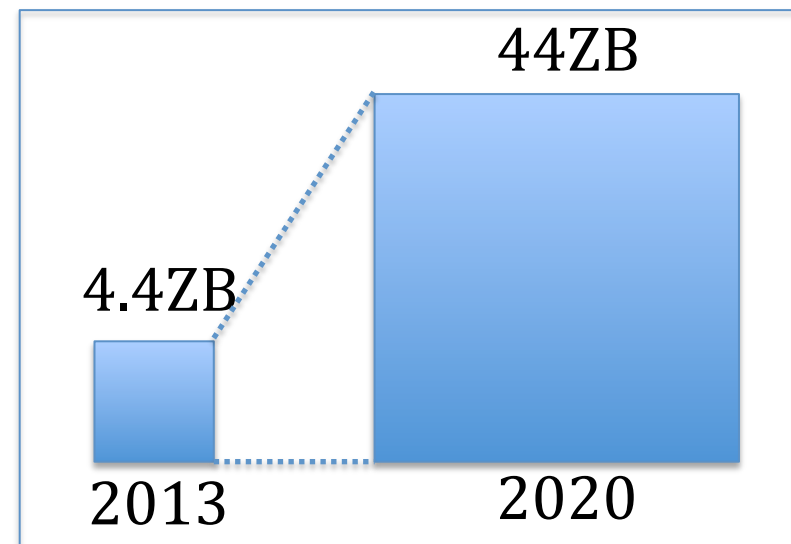
© 2015 Sakurai, Matsubara & Faloutsos

2

Motivation

- Time-series analysis for big data
  - Large-scale sensor networks
  - Web and social networks
  - Medical and healthcare records
- Digital universe growth
  - 4.4 zettabytes
    (4.4 trillion gigabytes)
 - 44 zettabytes in 2020
  - Data management and
    data mining

44ZB

4.4ZB

2013   2020

The DIGITAL UNIVERSE of
OPPORTUNITIES (IDC 2014)

Time-series analysis
# for big data

- Volume and Velocity
  - High-speed processing for large-scale data
  - Low memory consumption
  - Online processing for real-time data management

- Variety of data types
  - Multi-dimensional time-series data (e.g., sensor data)
  - Complex time-stamped events (e.g., web-click logs)
  - Time-evolving graph (e.g., social networks)

- Advanced techniques for big data
  - Model estimation, summarization
  - Anomaly detection, forecasting
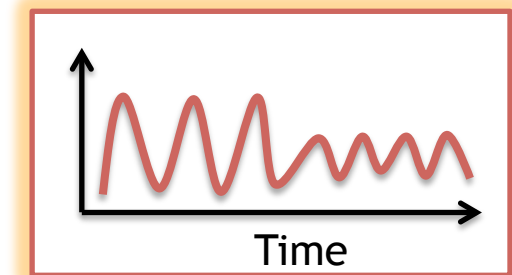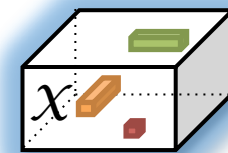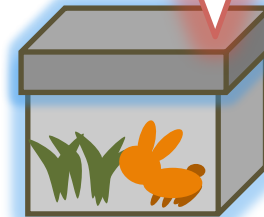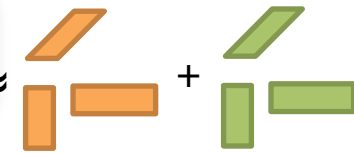
Time

New research directions

- Time-series data analysis
  - Indexing and fast searching
  - Sequence matching
  - Clustering
  - etc.

- New research directions
  1. Automatic mining
  2. Non-linear modeling
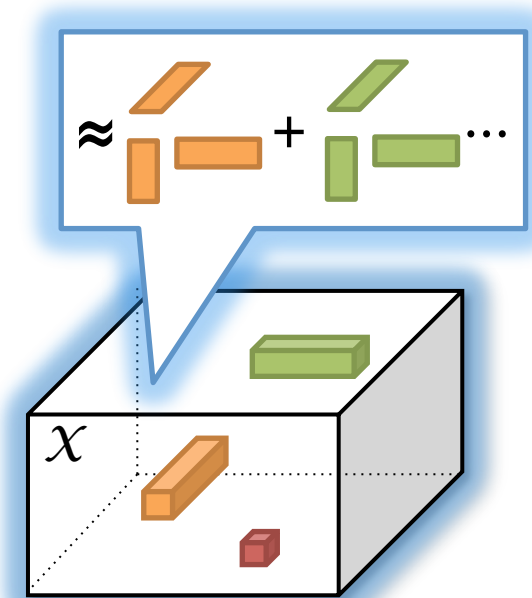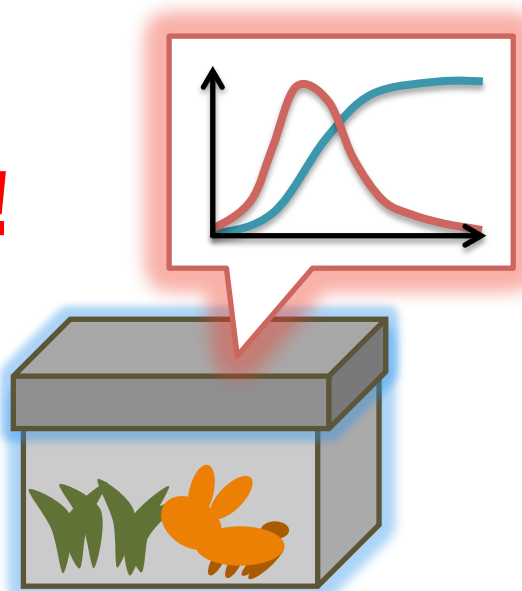 3. Large-scale tensor analysis

Time

NO magic #

$x \approx$ + ...
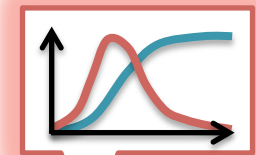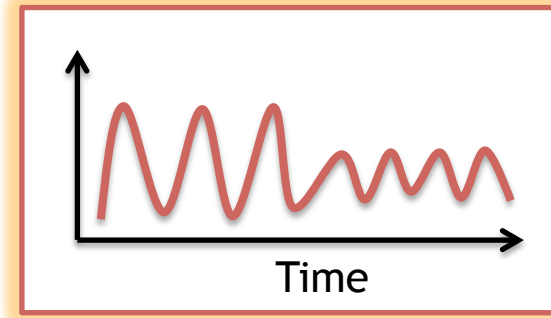
New research directions

R1. Automatic mining (no magic numbers!)

R2. Non-linear (gray-box) modeling

R3. Tensor analysis

NO magic numbers !

(R1) Automatic mining
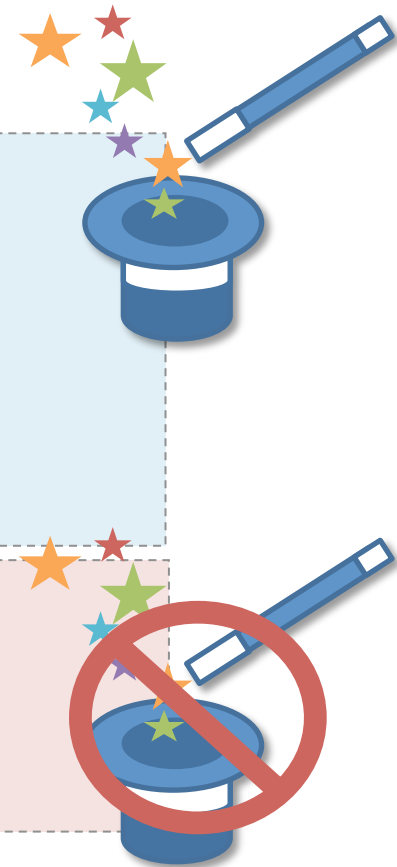
No magic numbers! ... because,

Manual

- sensitive to the parameter tuning

- long tuning steps (hours, days, …)

Automatic (no magic numbers)

- no expert tuning required

Big data mining:

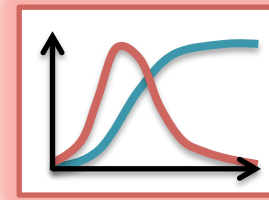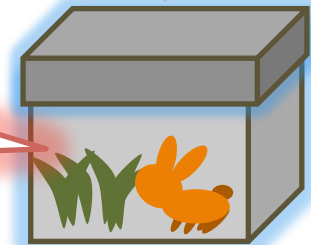-> we cannot afford human intervention!!

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

- Gray-box mining
  - If we know the equations

Gray-box

- Non-linear (differential) equations
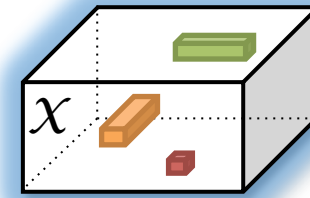 - Epidemic
  - Biology
  - Physics, Economics, etc.,

- Modeling non-linear phenomena
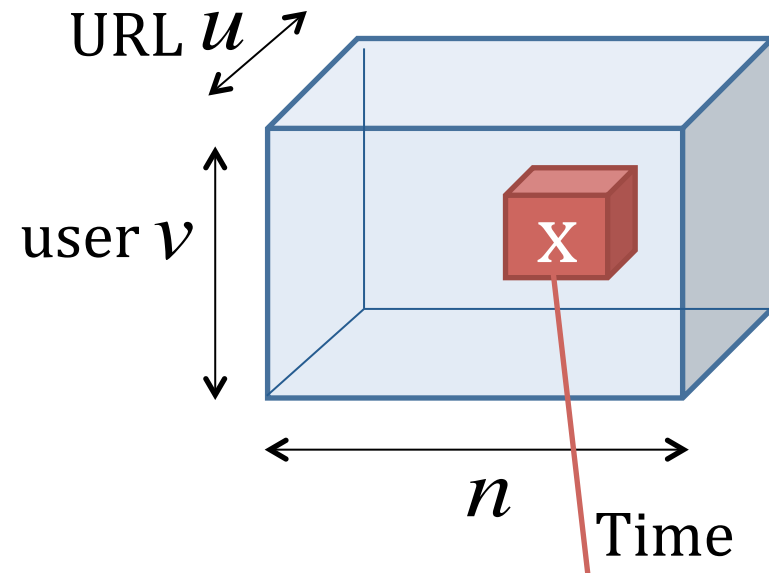 - Non-linear time-series analysis for social networks

Image courtesy of Tina Phillips and amenic181 at FreeDigitalPhotos.net.

(R3) Large-scale tensor analysis

- Time-stamped events

  - *e.g., web clicks*

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

Represent as
Mth order tensor (M=3)

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

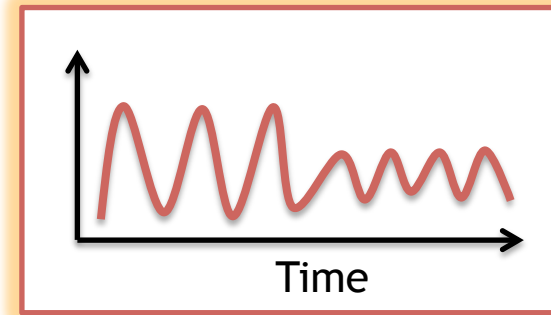URL u ↗

user v

n

Time

Element x: # of events

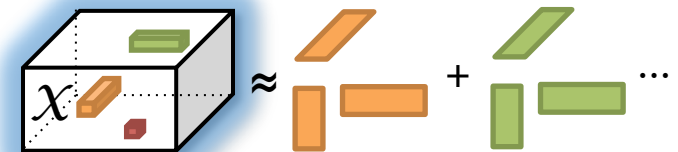e.g., 'Smith', 'CNN.com', 'Aug 1, 10pm'; 21 times

New research directions

- Time-series data analysis
  - Indexing and fast searching
  - Sequence matching
  - Clustering
  - etc.

- New research directions

  R1. Automatic mining

  R2. Non-linear modeling

  R3. Large-scale tensor analysis
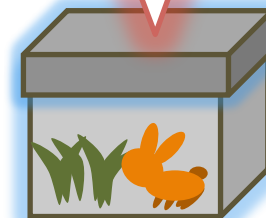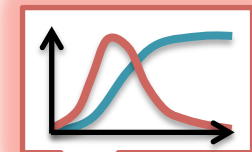
NO magic #

Time

New research directions

- Time-series data analysis
  - Indexing and fast searching
  - Sequence matching
  - Clustering
  - etc.

- New research directions
  R1. Automatic mining
  R2. Non-linear modeling
  R3. Large-scale tensor analysis

NO magic #

Part 1

Part 2

Part 3