P-E-R-S-I-S-T-E-N-C-E and DISTINCTIVENESS of Inter-event Time Distributions *in Online Human Behavior*

Jiwan Jeong and Sue Moon

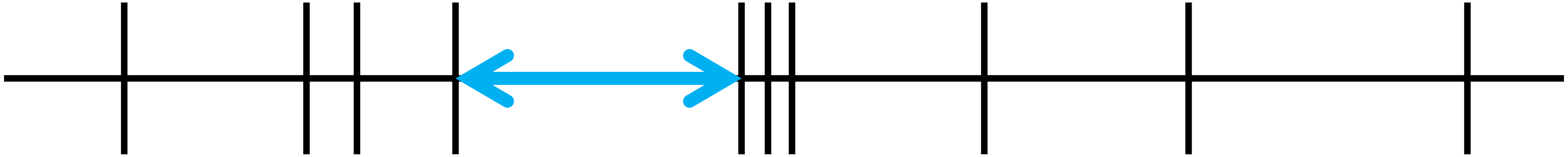
School of Computing, KAIST

In *TempWeb '17 (WWW '17 Companion)*

April 3, 2017

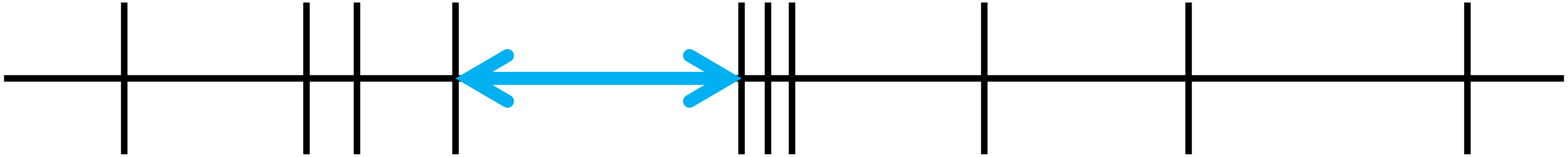
What is inter-event time?

- Time gap between two consecutive ***events***
- E.g., earthquake waves, packet arrivals, …



Our definition of inter-event time

- Time gap between two consecutive ***actions*** in a service by one person
- E.g., tweeting, blog posting, email sending, …



- Simply put
    - Inter-event time = interval
    - Inter-event time distribution = interval pattern

Previous studies focused on

- Characterizing **aggregate** interval patterns
    - Web re-visit pattern [Adar *CHI* 2007][Adar *CHI* 2008]
    - Web browsing pattern [Kumar *WWW* 2010]
    - Service usage pattern [Halfaker *WWW* 2015]

- Finding **universal laws** among interval patterns
    - Power-law by priority queuing process [Barabasi *Nature* 2005]
    - Log-normal by non-homogeneous Poisson process [Malmgren *PNAS* 2008]

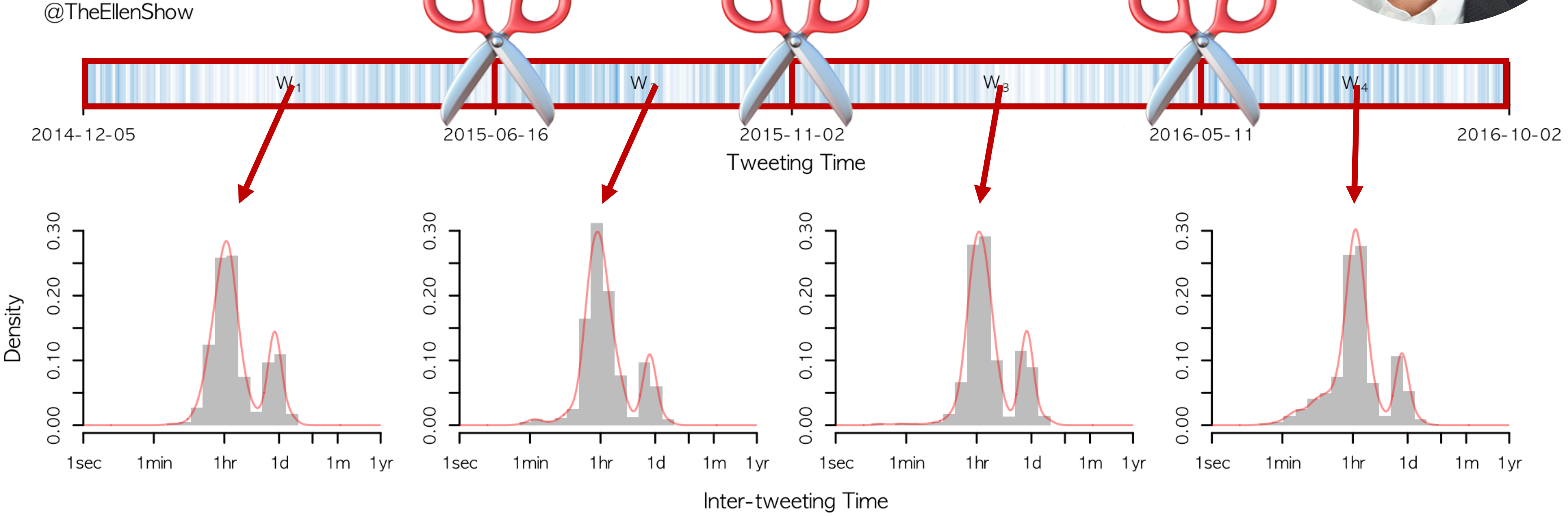
We focus on individual-level

- How does an individual’s interval pattern change over time?
- Does it remain consistent or fluctuate from time to time?
- How distinctive is it from those of others?

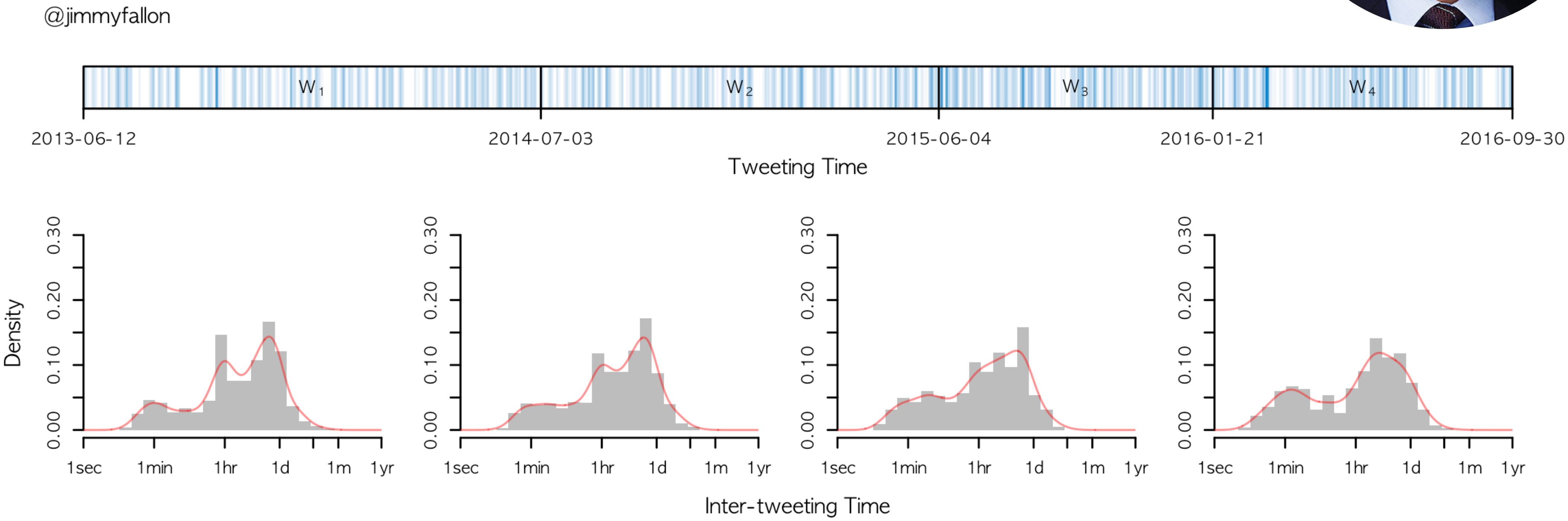
Individuals have **interval patterns** that are **persistent** over time, but **distinctive** from others.

Tweets by **Ellen DeGeneres**

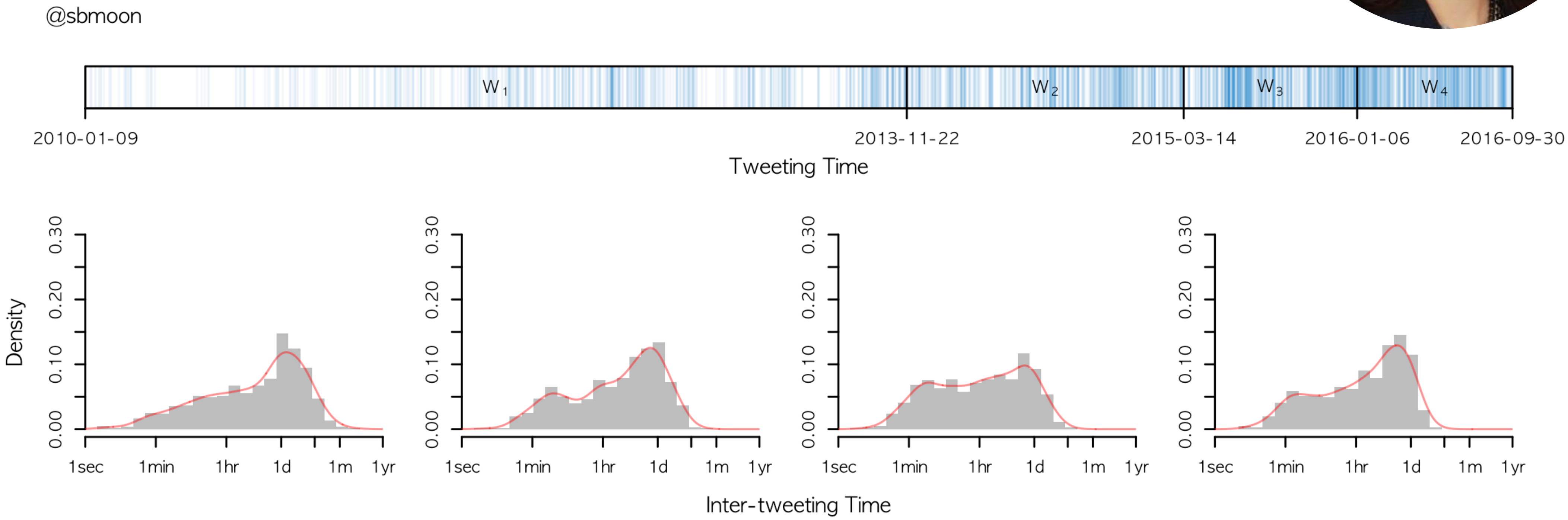
Twitter timeline



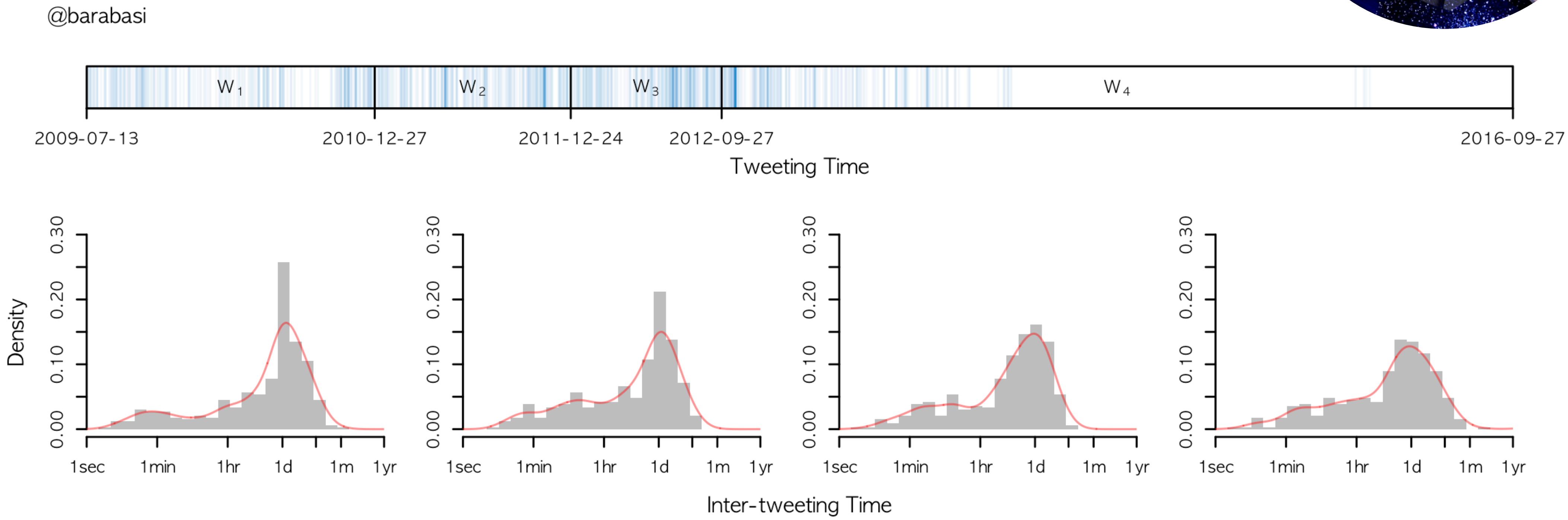
Tweets by **Jimmy Fallon**



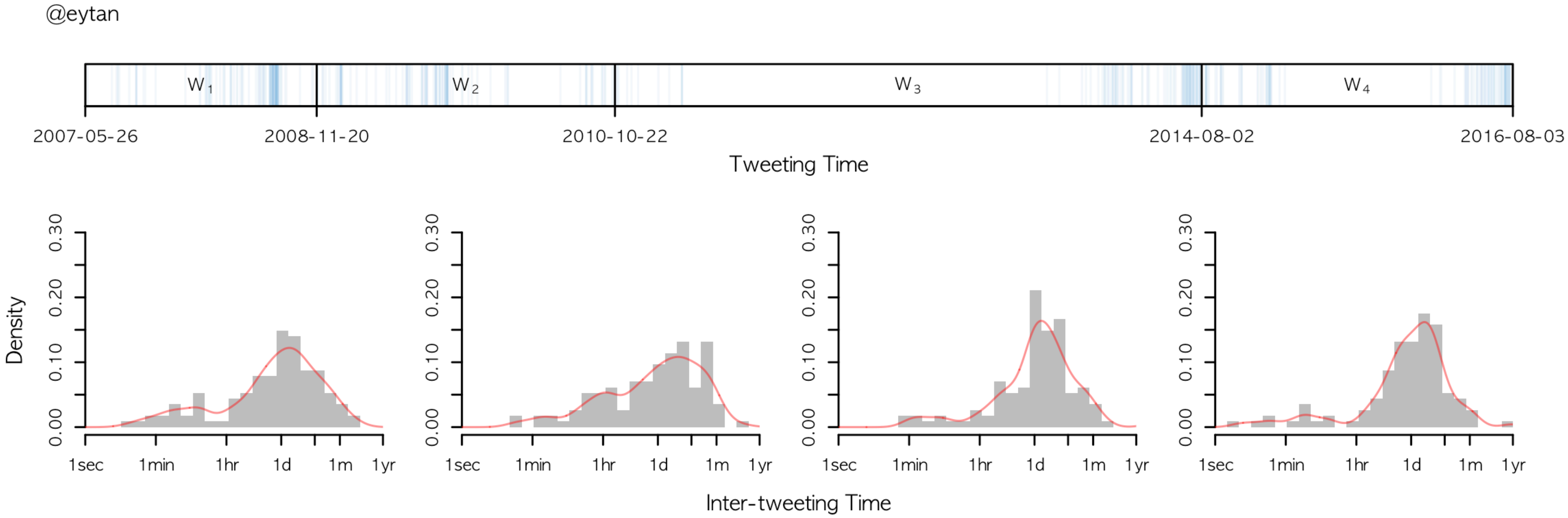
Tweets by **Sue Moon**



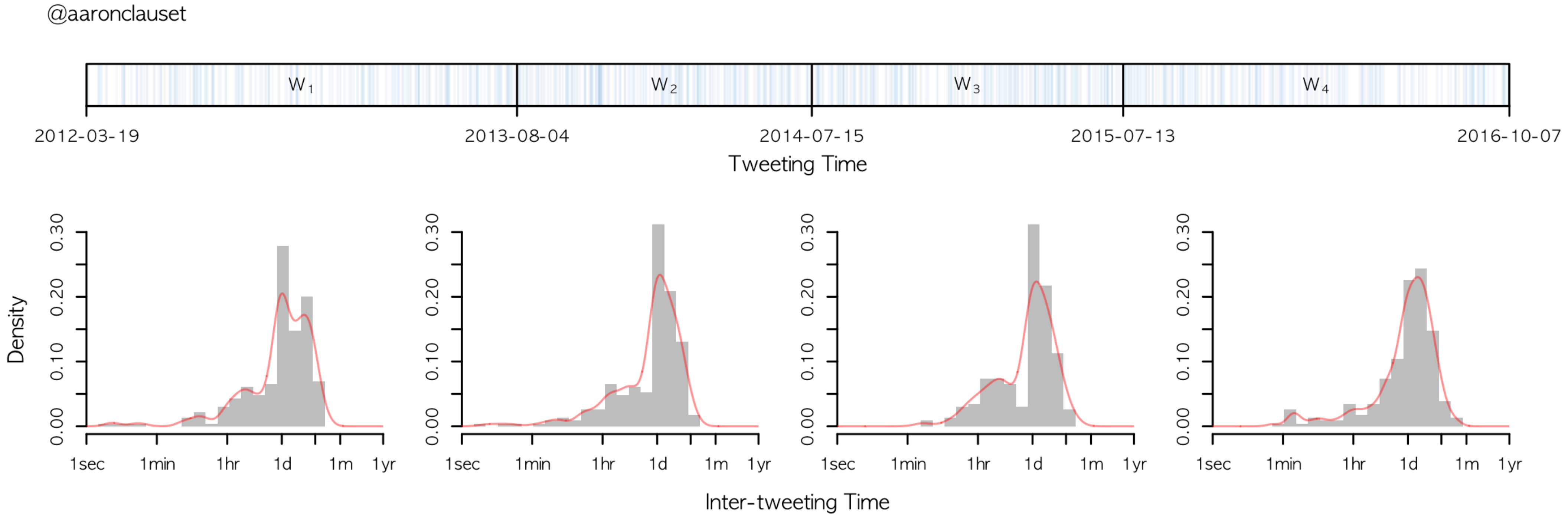
Tweets by **Albert-László Barabási**



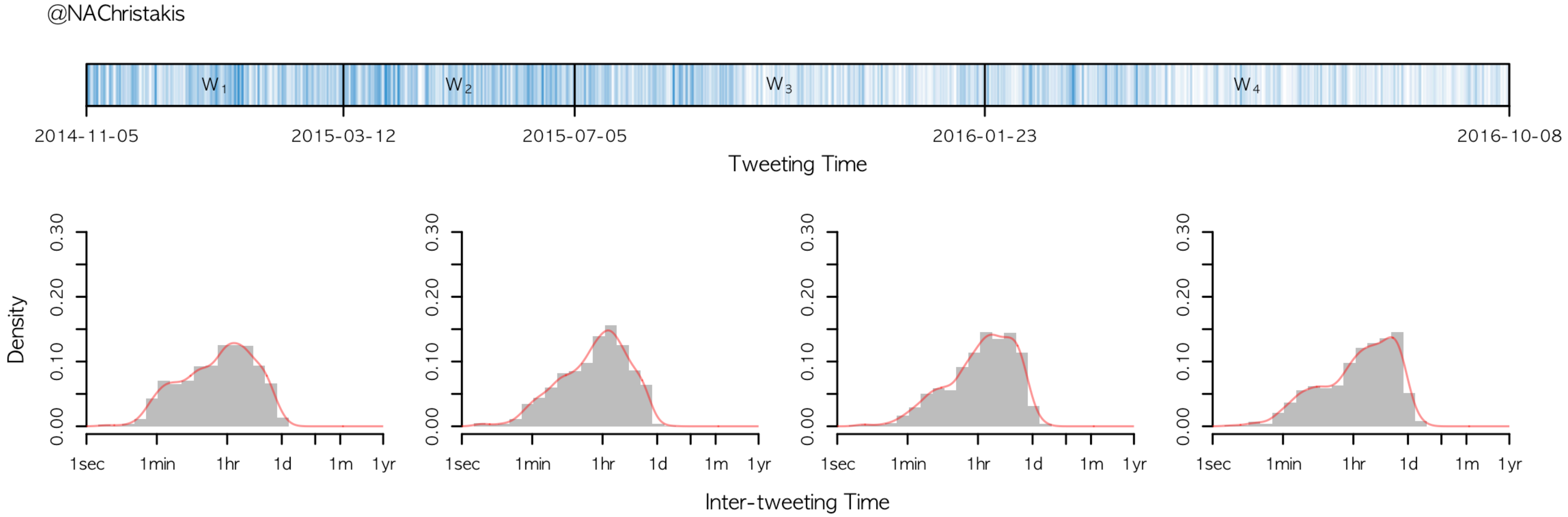
Tweets by **Eytan Adar**



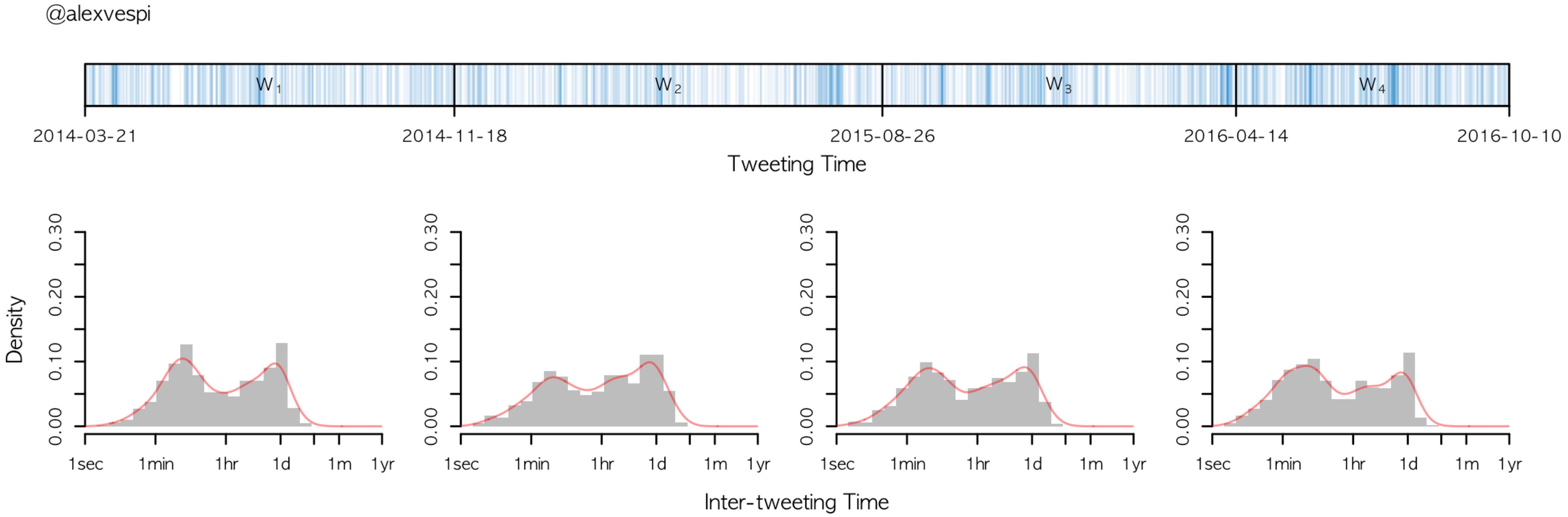
Tweets by **Aaron Clauset**



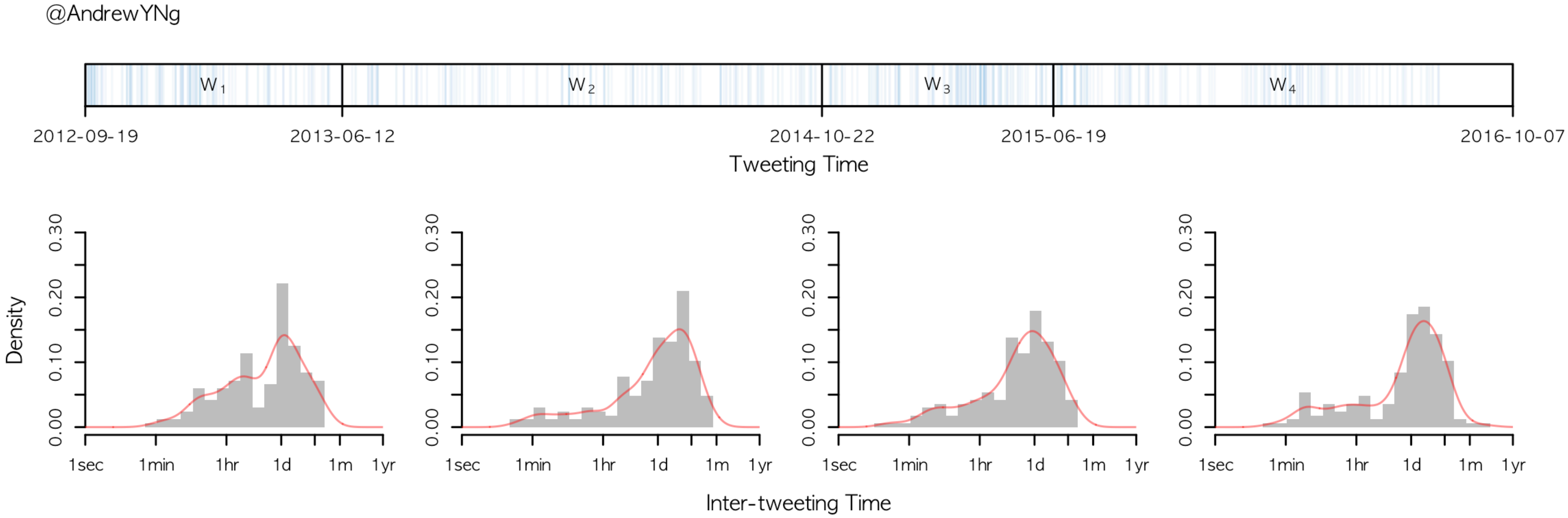
Tweets by **Nicolas Christakis**



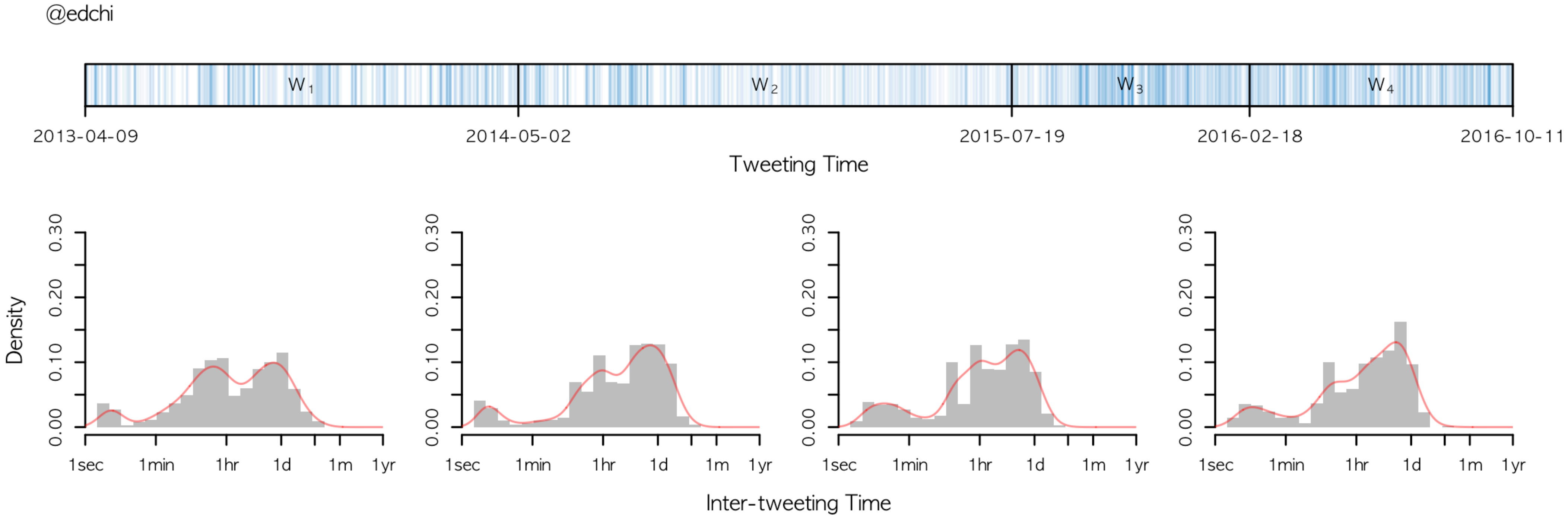
Tweets by **Alex Vespagini**



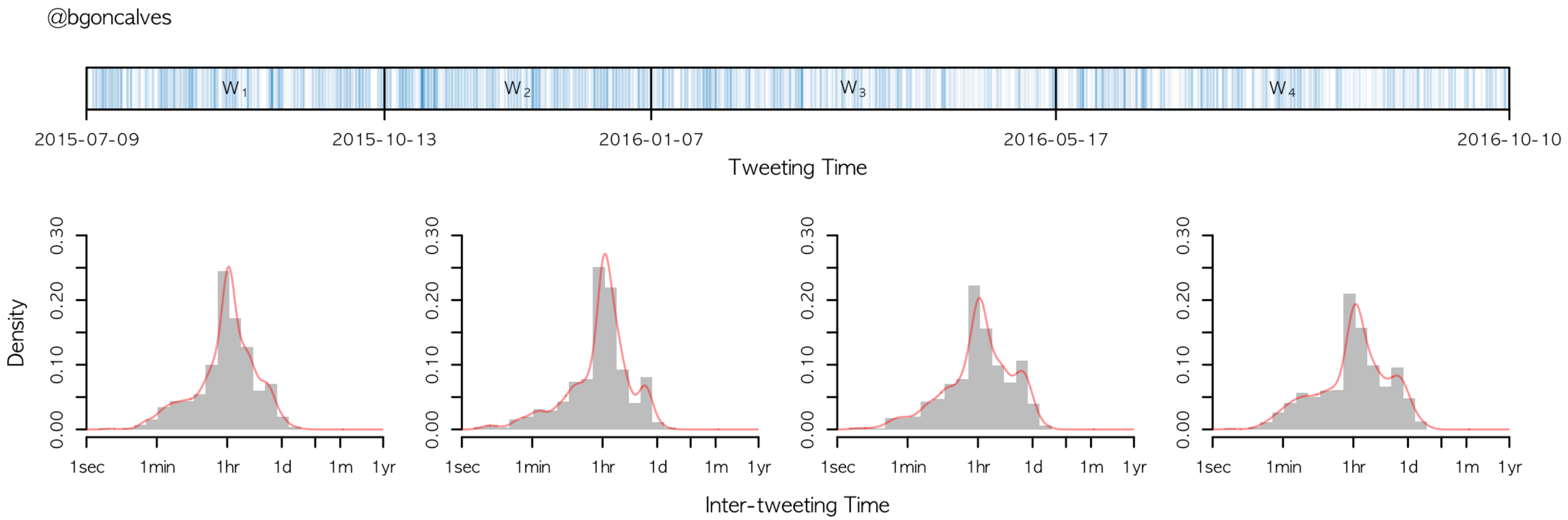
Tweets by **Andrew Ng**



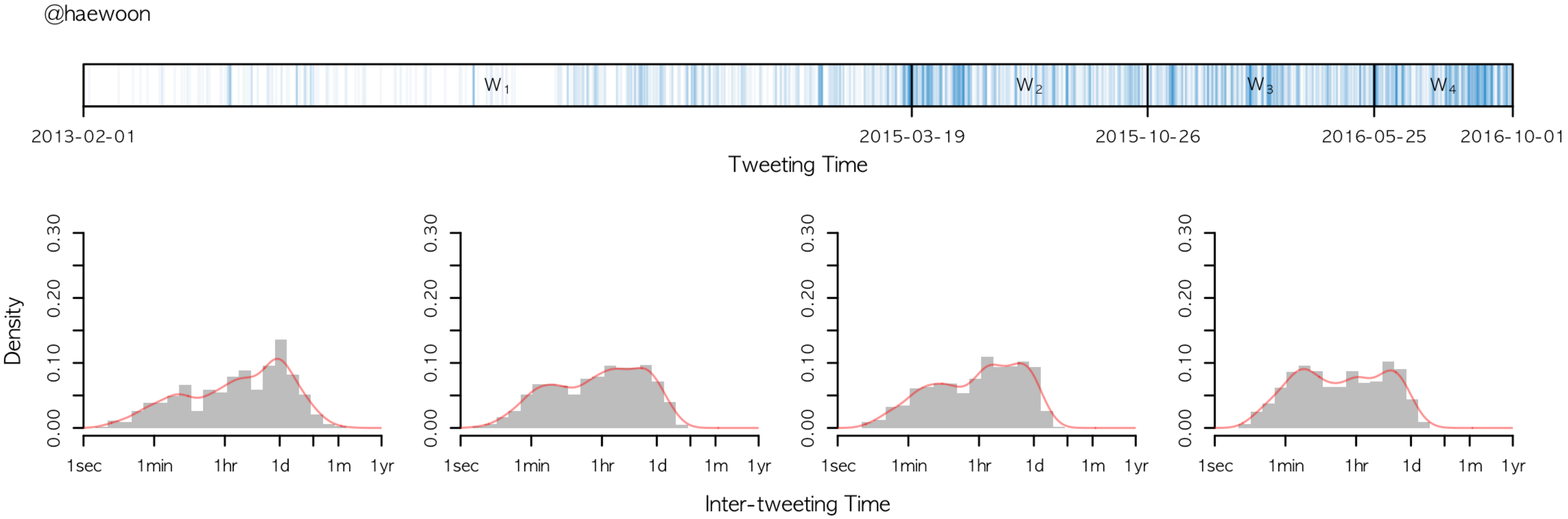
Tweets by **Ed Chi**



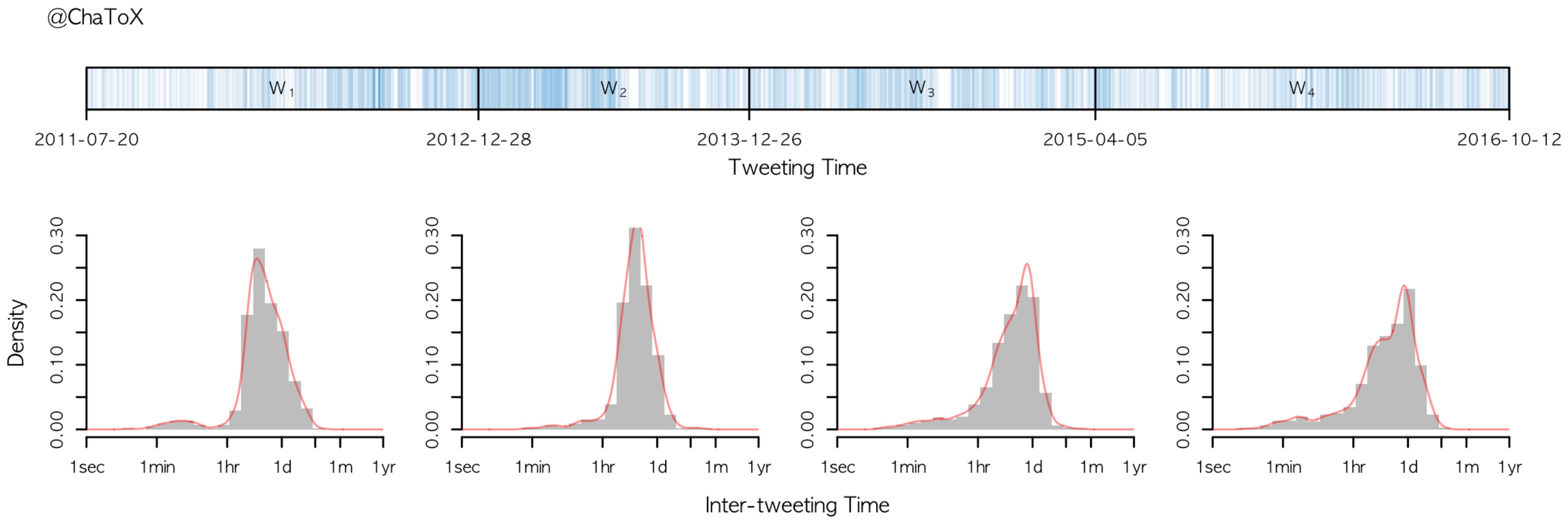
Tweets by **Bruno Gonçalves**



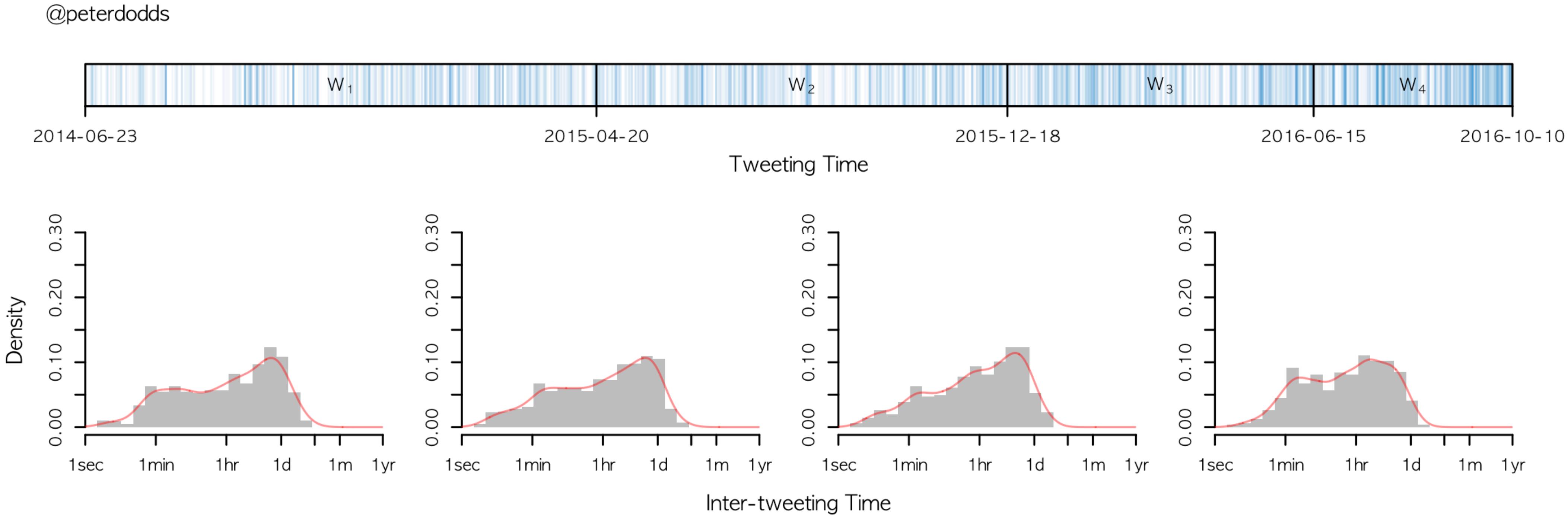
Tweets by **Haewoon Kwak**



Tweets by **Carlos Castillo**



Tweets by **Peter Dodds**

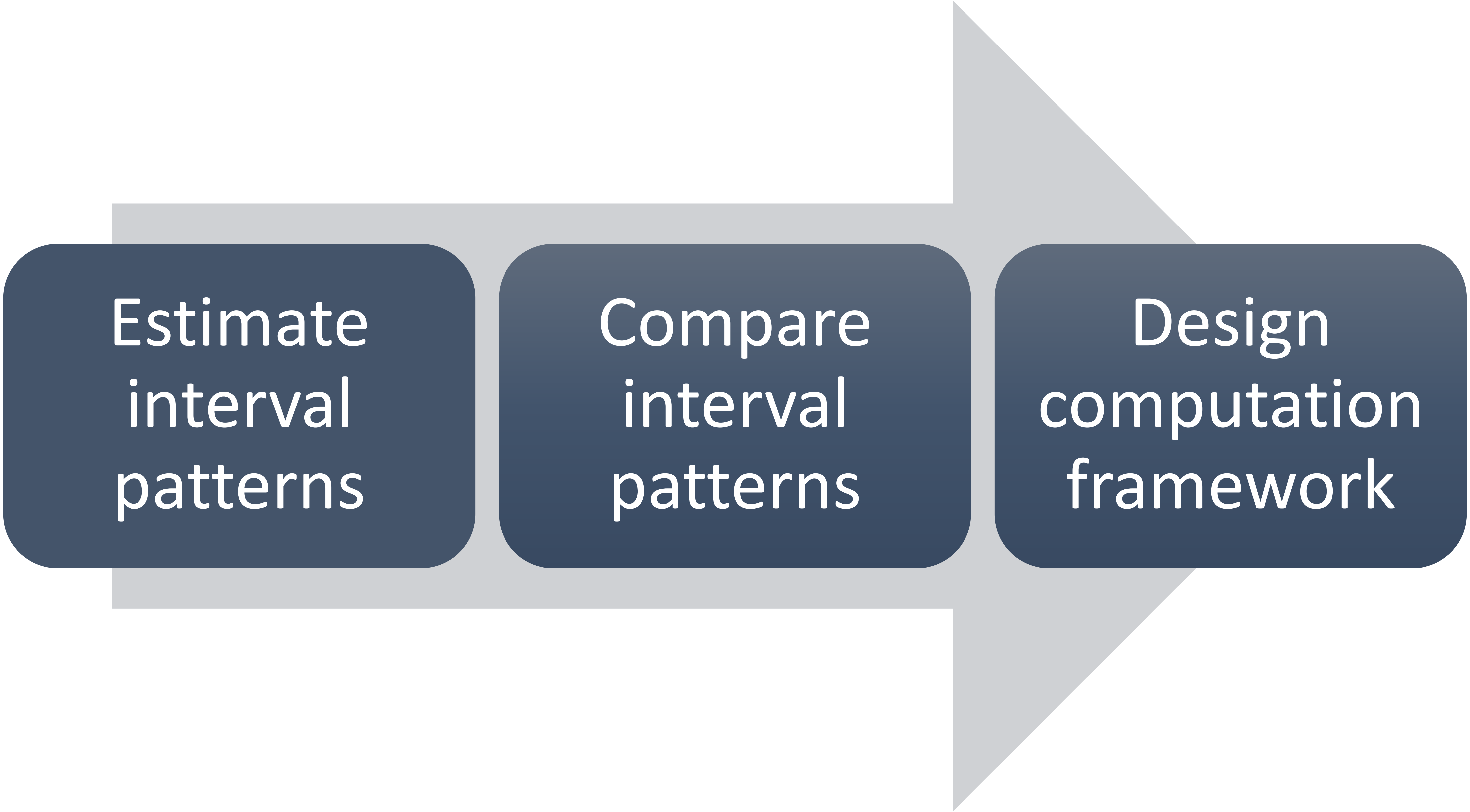


In this work

- Design a computation framework to quantify interval patterns
- Show their persistence and distinctiveness
- Use interval patterns to distinguish one user from others

Datasets for this study

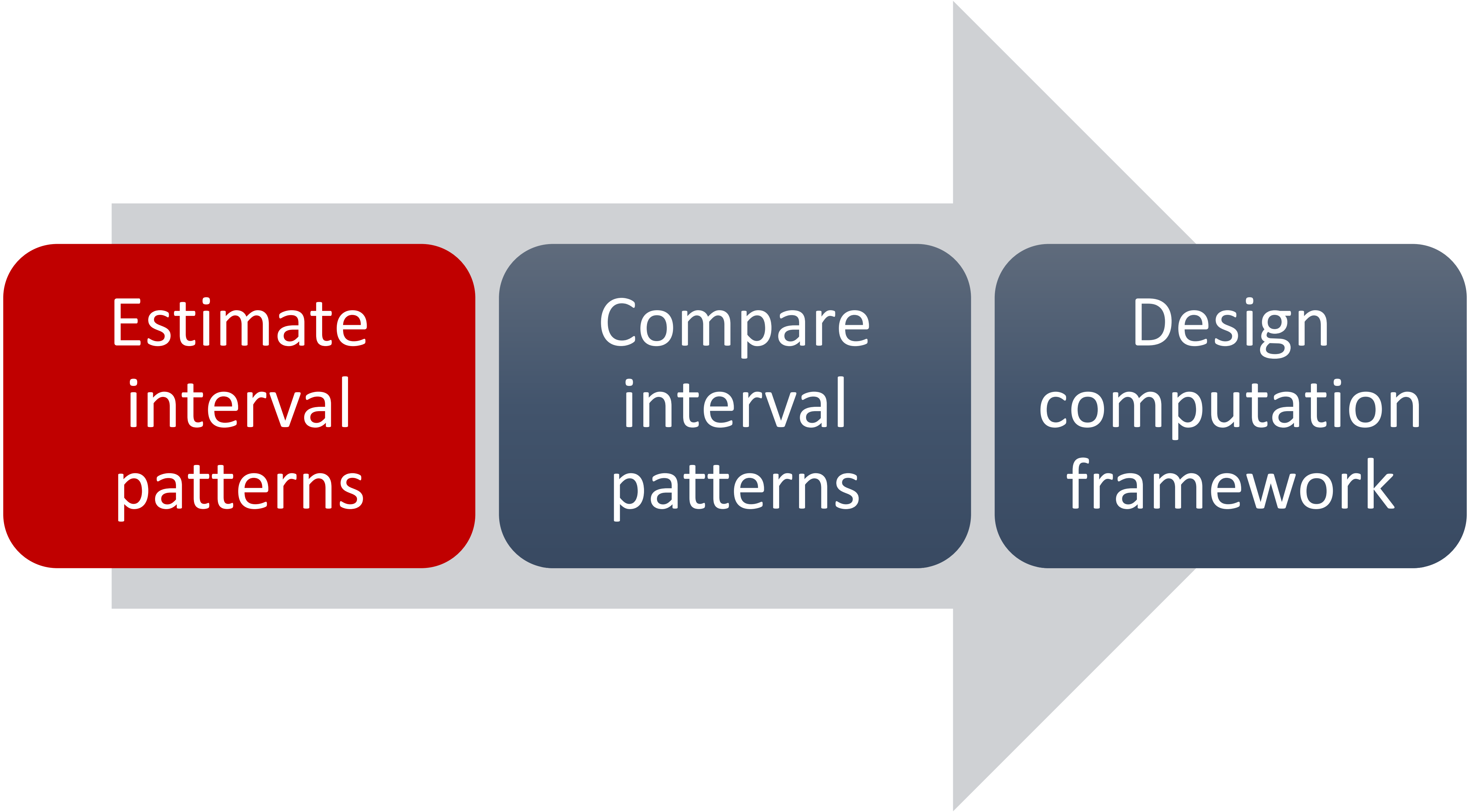
- WIKIPEDIA The Free Encyclopedia 15 years of entire history
- me2day 7 years of entire history
- twitter 3000 recent tweets per user
- ENRON® 3 years of email history



Estimate interval patterns

Compare interval patterns

Design computation framework

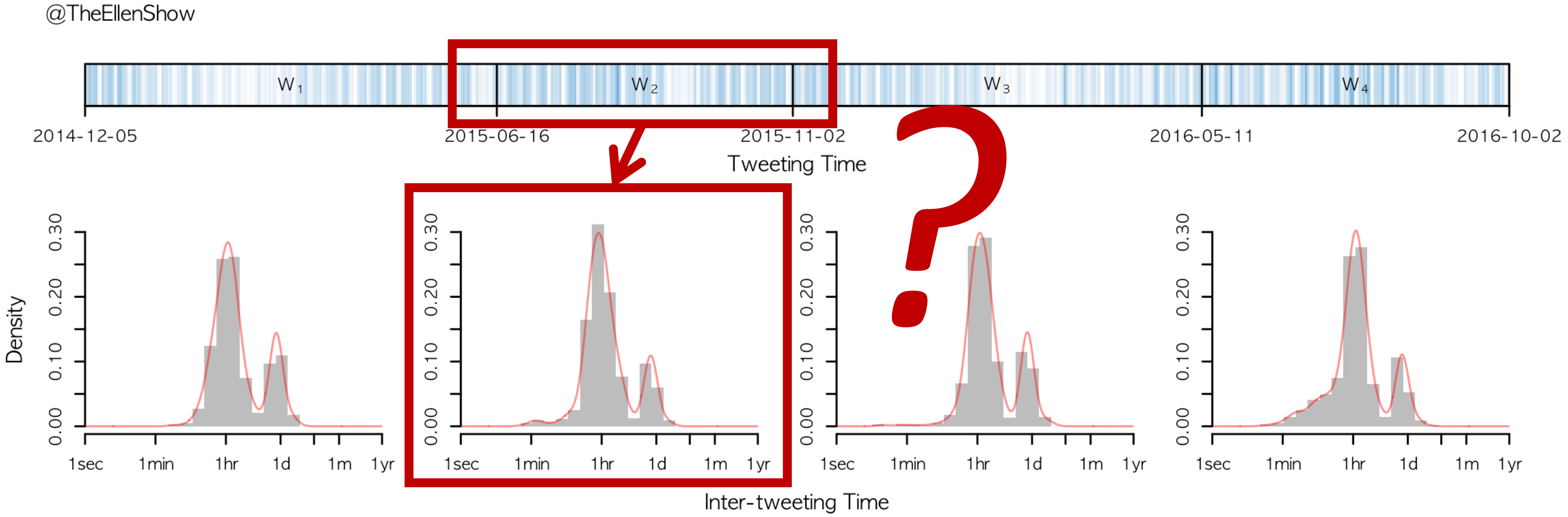


Estimate interval patterns

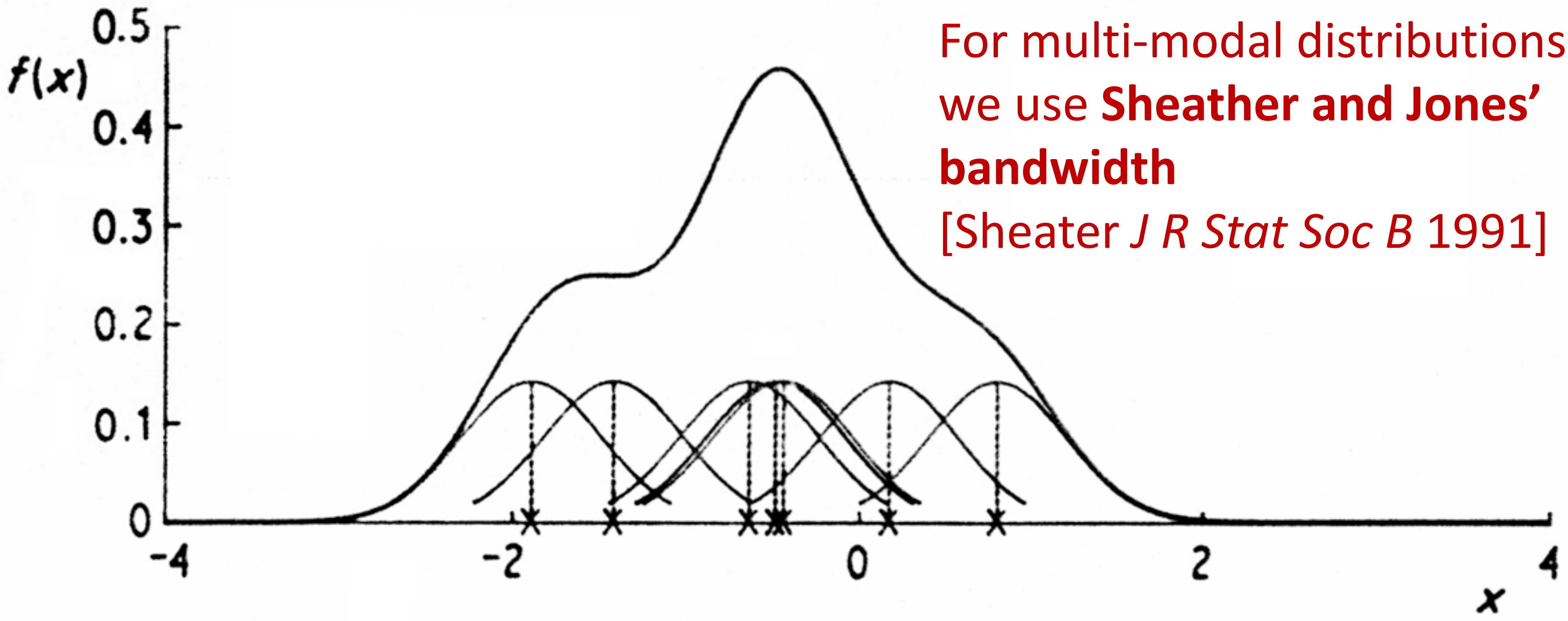
Compare interval patterns

Design computation framework

Convert **discrete intervals** to **continuous PDF**



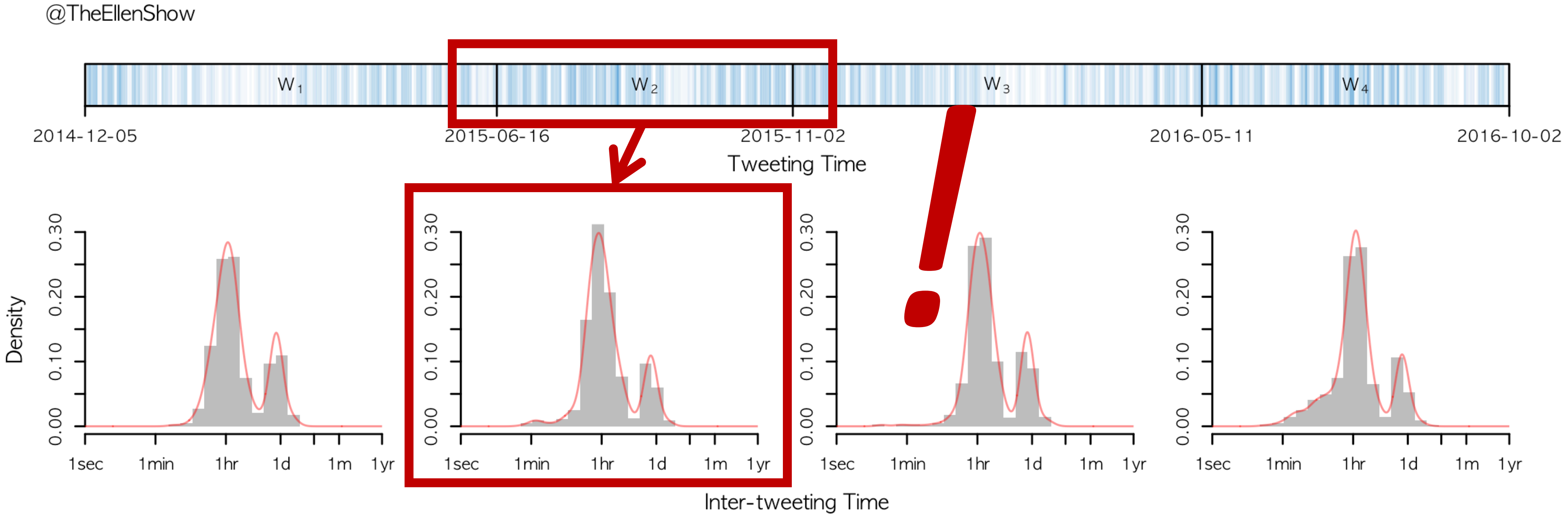
Gaussian kernel density estimation

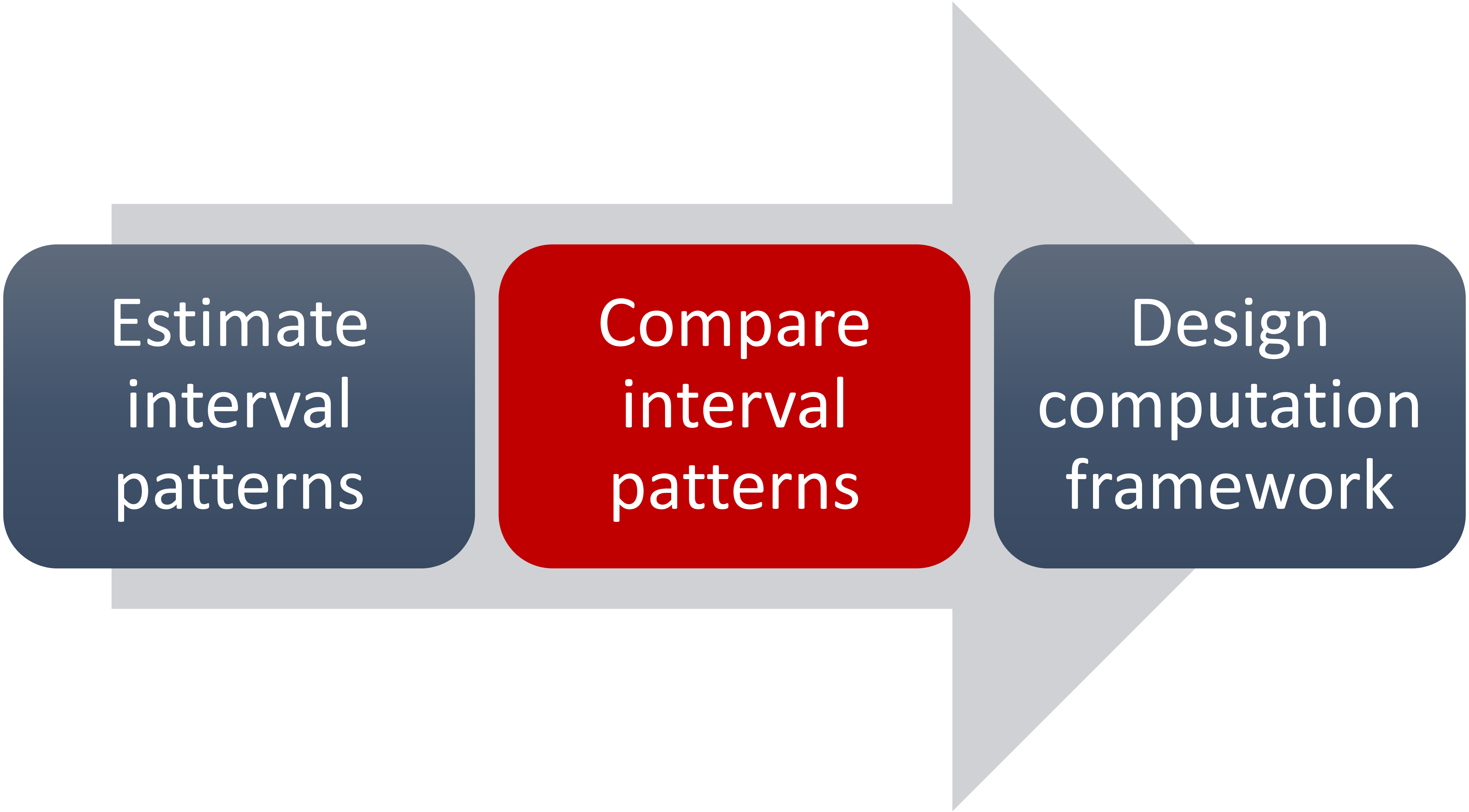


For multi-modal distributions, we use **Sheather and Jones' bandwidth**

[Sheater *J R Stat Soc B* 1991]

Now, we can estimate interval patterns!



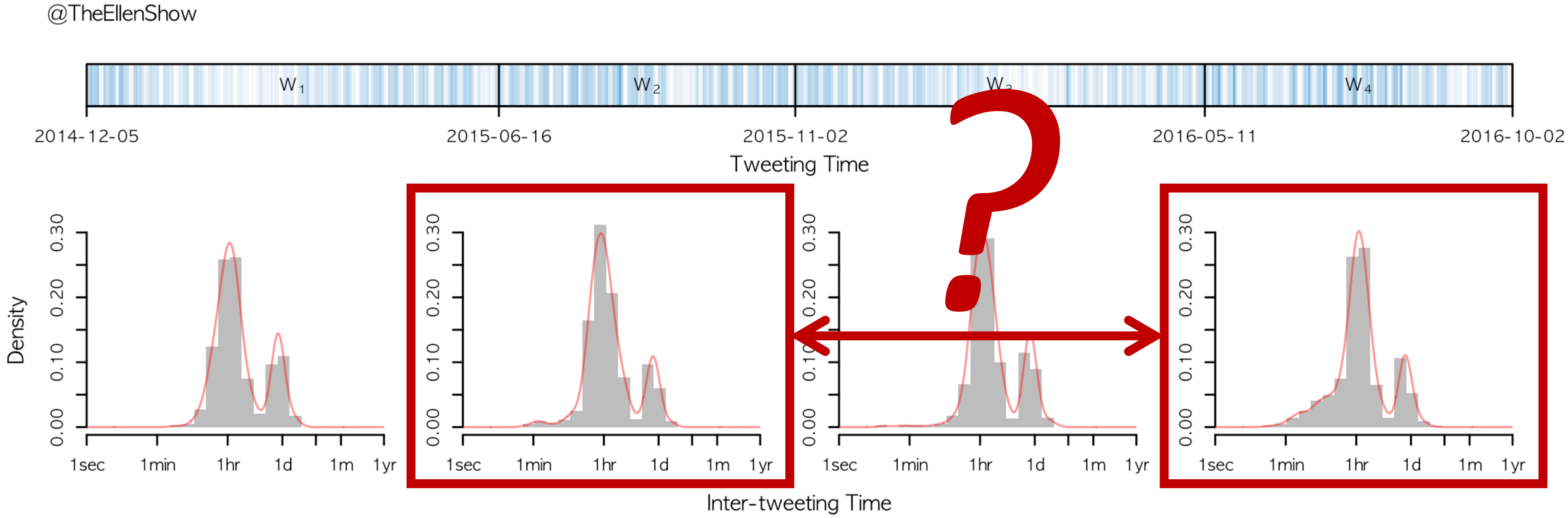


Estimate interval patterns

Compare interval patterns

Design computation framework

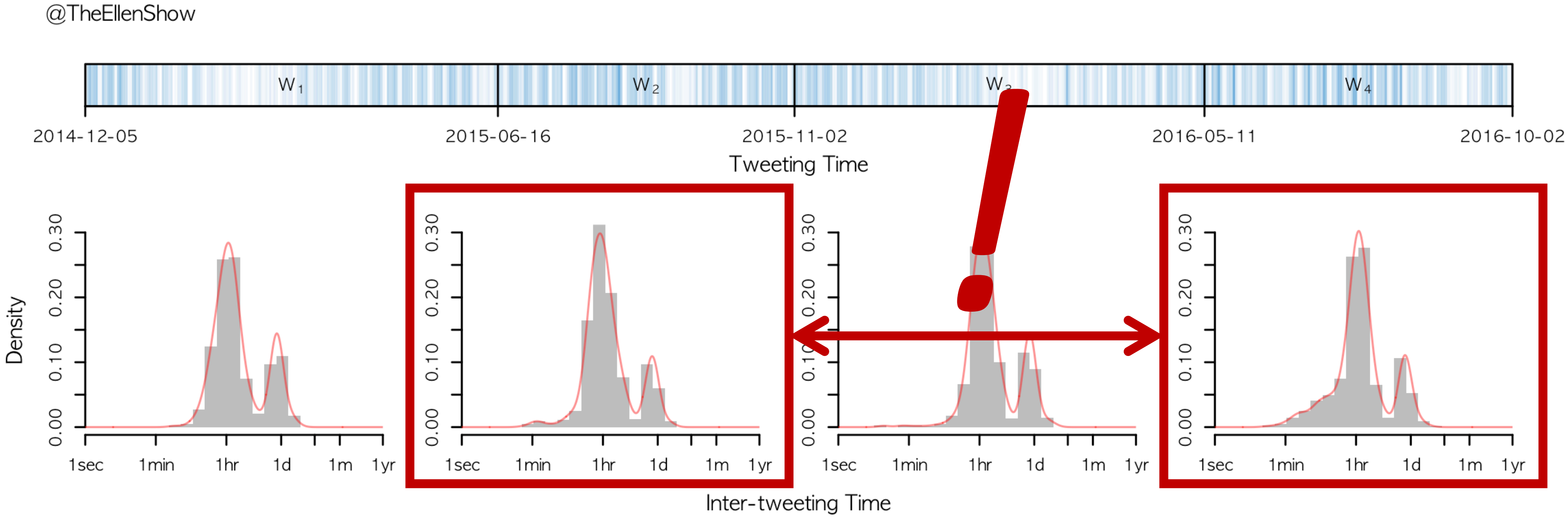
Calculate **distance** between interval patterns

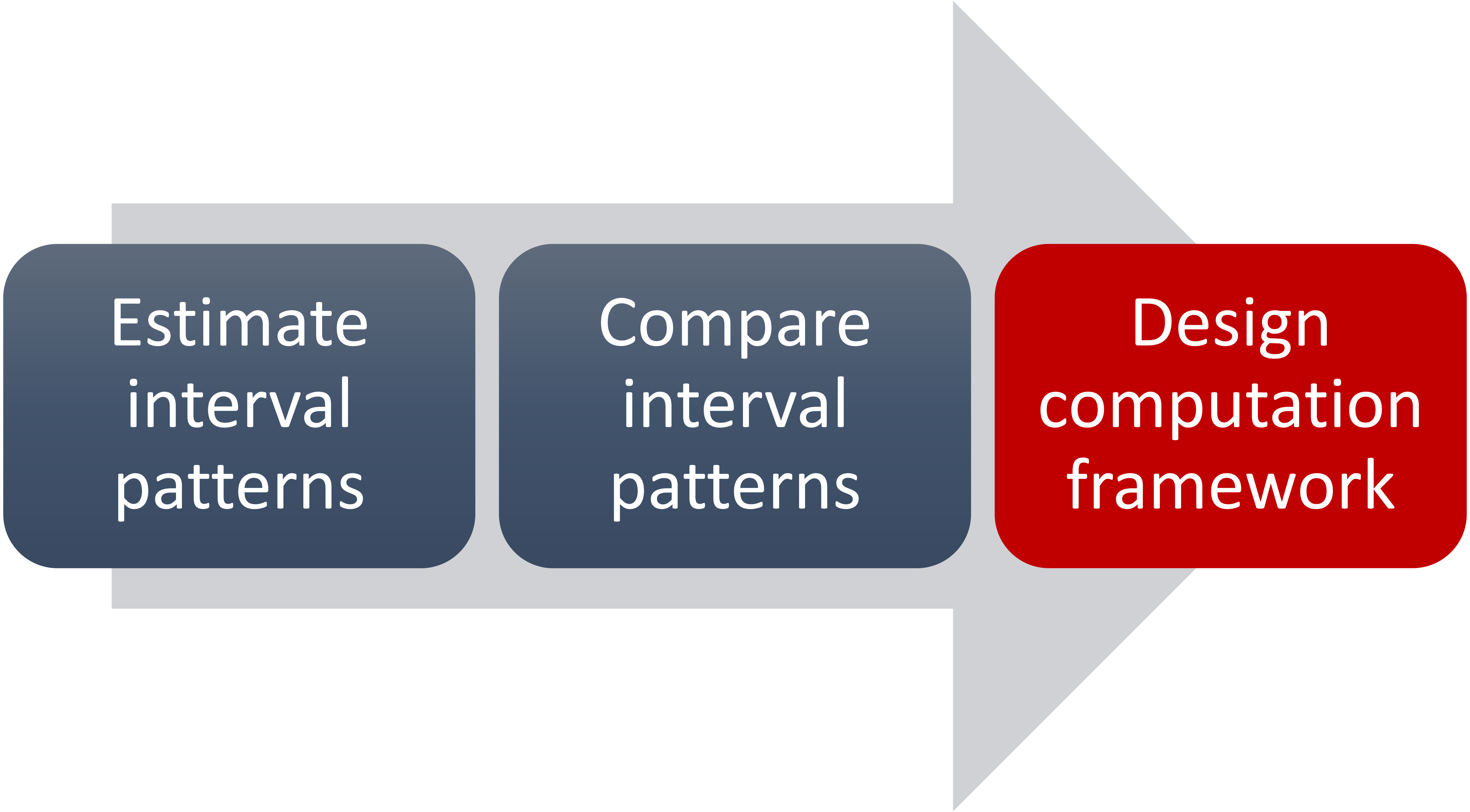


Jensen-Shannon distance

- A **metric** of the difference between probability density functions
  - Non-negative: $d(x, y) \geq 0$
  - Identity of indiscernibles: $d(x, y) = 0$ iff $x = y$
  - Symmetry: $d(x, y) = d(y, x)$
  - Subadditivity: $d(x, z) \leq d(x, y) + d(y, z)$

Now, we can compare interval patterns!



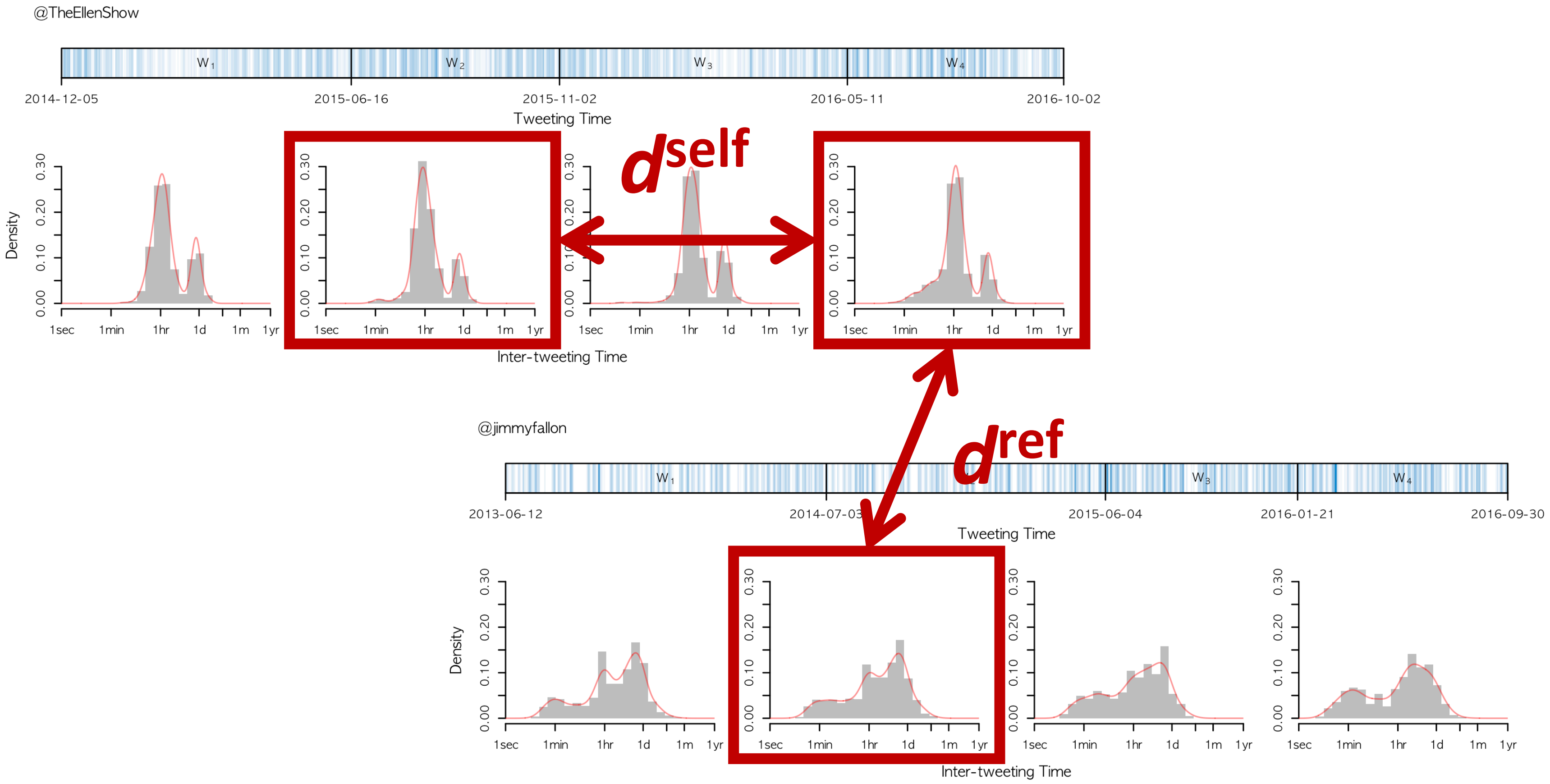


Estimate interval patterns

Compare interval patterns

Design computation framework

Define **self-distance** and **reference distance**



Experimental settings for longitudinal analysis

- Select users with +500 actions on each service
- Divide each user's timeline into 10 windows

W_1	**W_2**	**…**	**W_9**	**W_{10}**

- $\binom{10}{2}$ = 45 self-distances for each user
- 10×10 = 100 reference distances for each pair of users

P-E-R-S-I-S-T-E-N-C-E

&

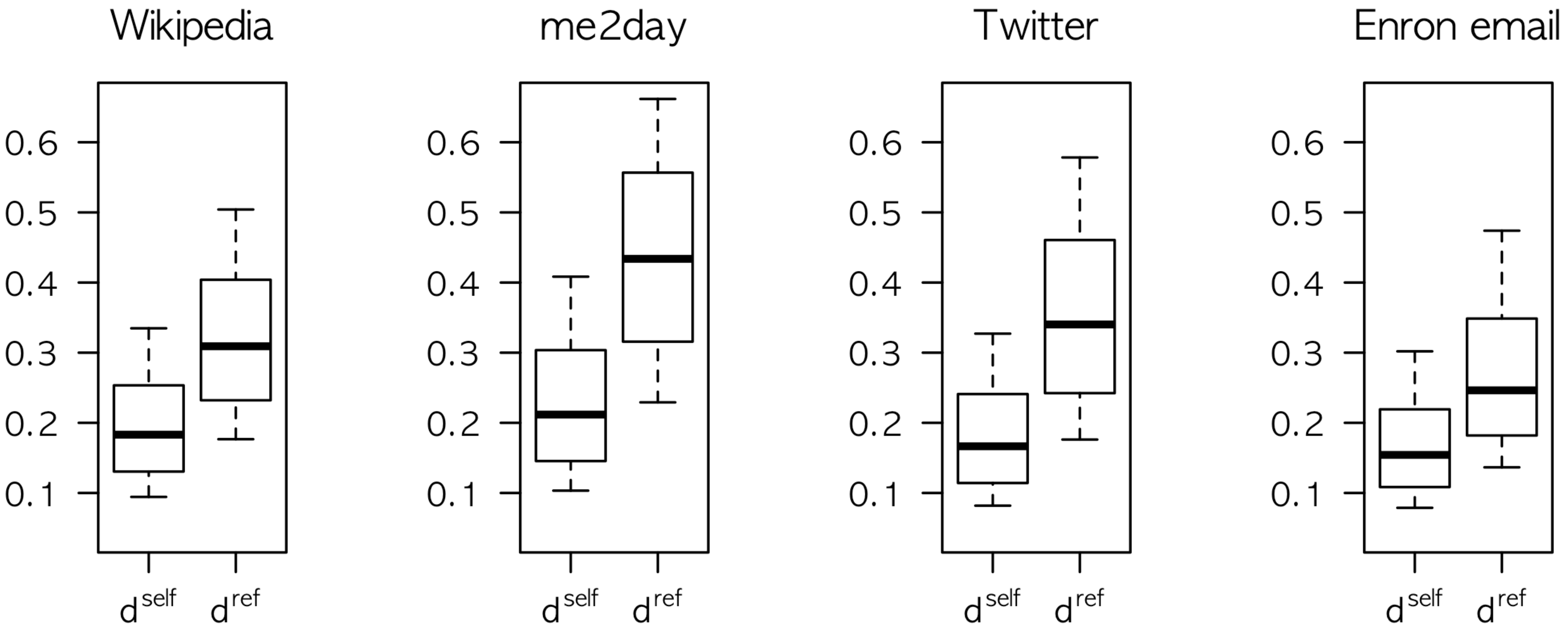
DISTINCTIVENESS

Persistence and distinctiveness are relative

- If d^{self} are small, the pattern is persistent
- How small should it be?
- If $d^{\mathrm{self}} < d^{\mathrm{ref}}$, the pattern is persistent [Saramäki *PNAS* 2014]

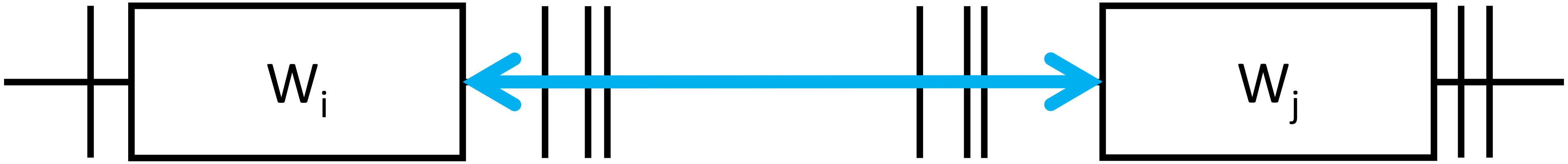
- Furthermore, if $d^{\mathrm{self}} \ll d^{\mathrm{ref}}$, the patterns are distinctive

d^{self} vs d^{ref}



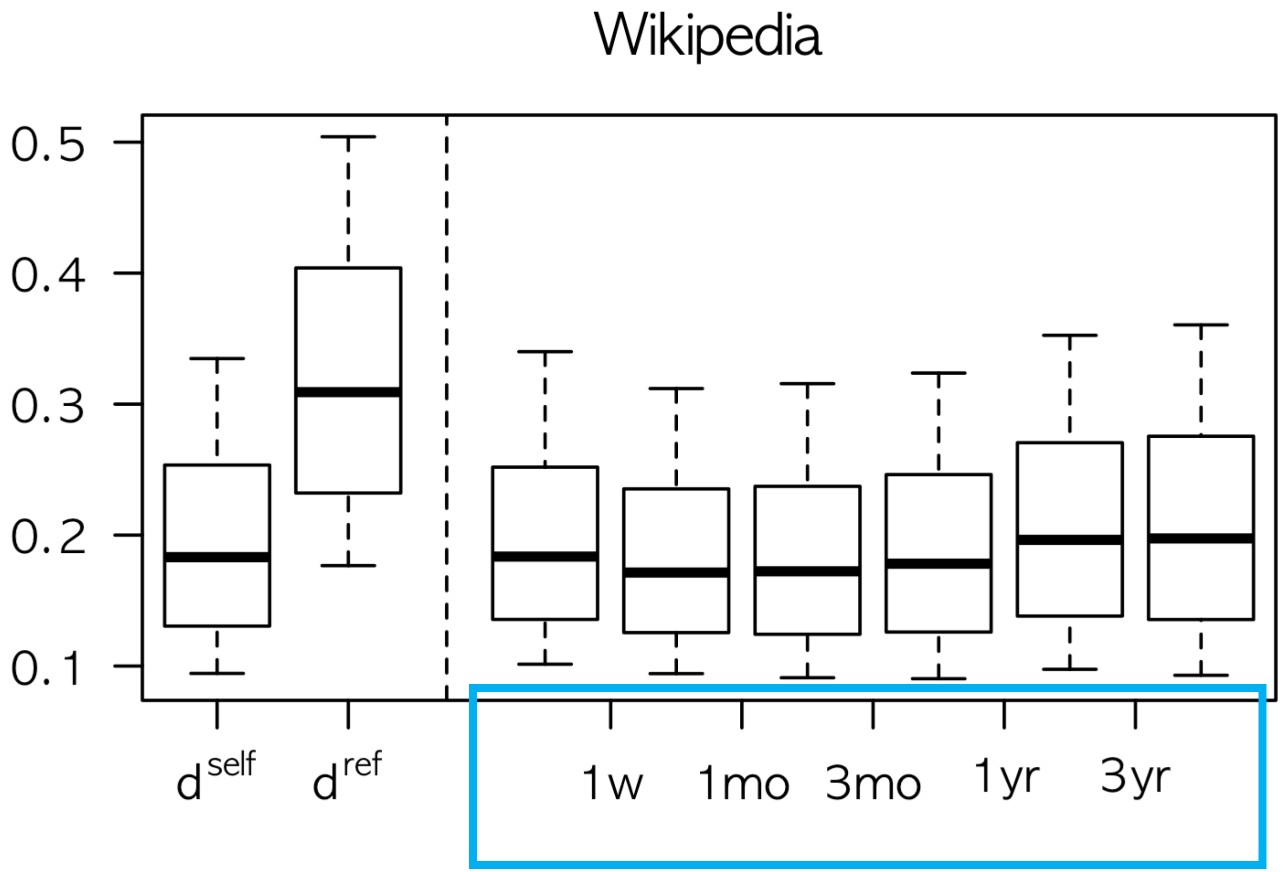
How long do interval patterns persist?

- Binning d^{self} by the **time gap** between two windows



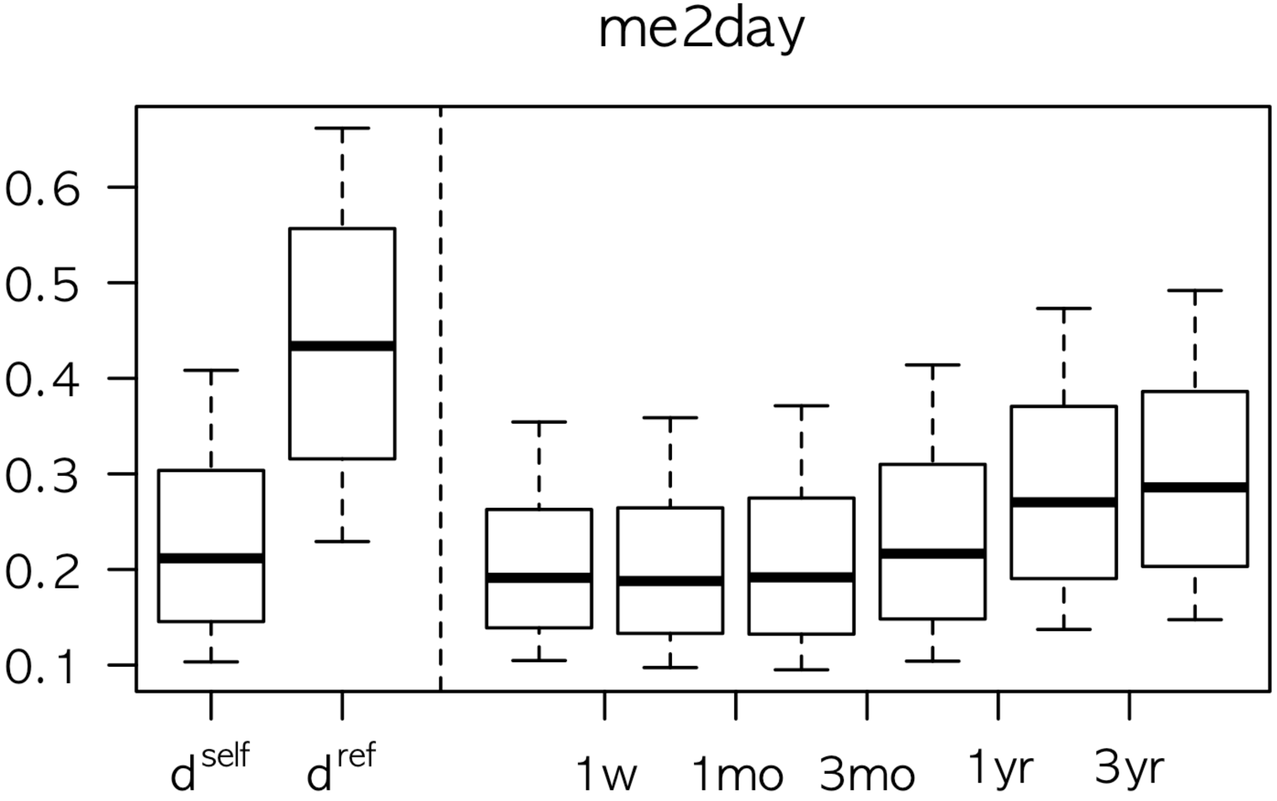
- Compare binned d^{self} with overall d^{ref}

Persistence over time

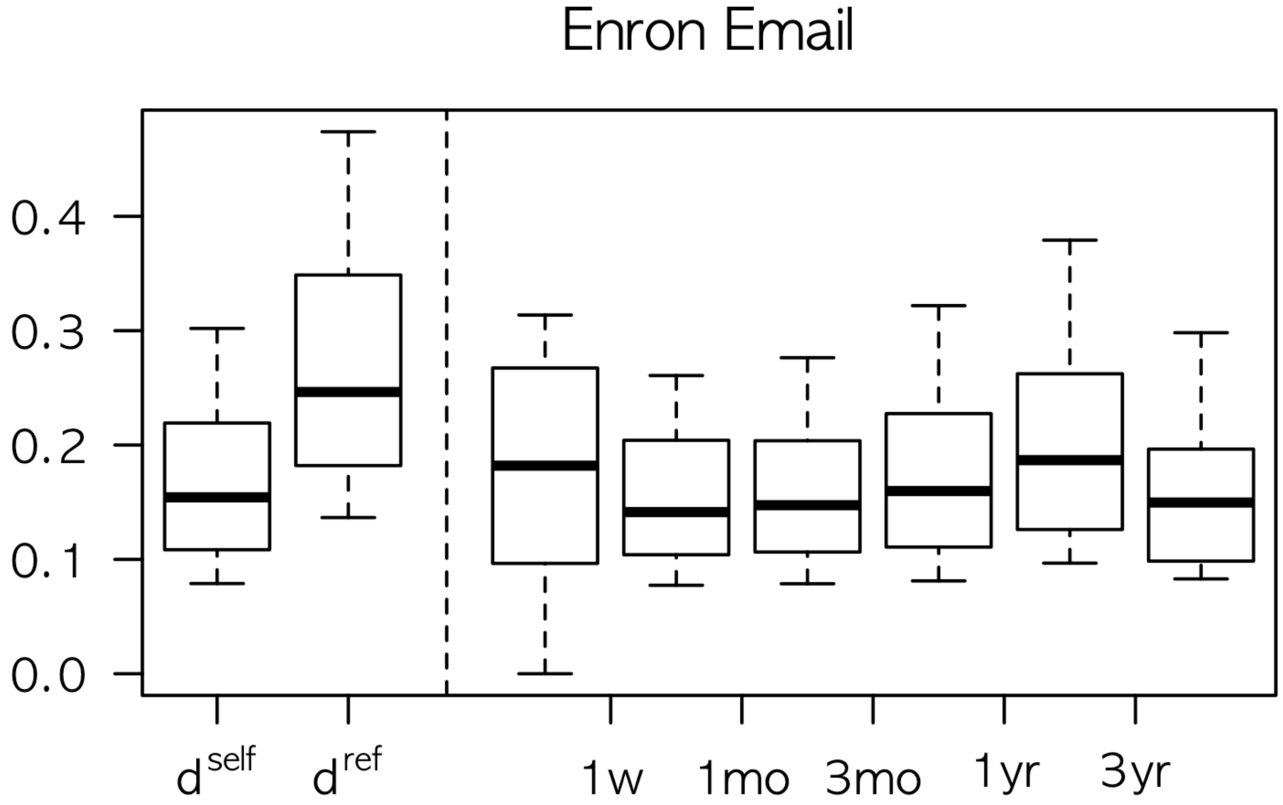


Binned into 6 groups

Persistence over time

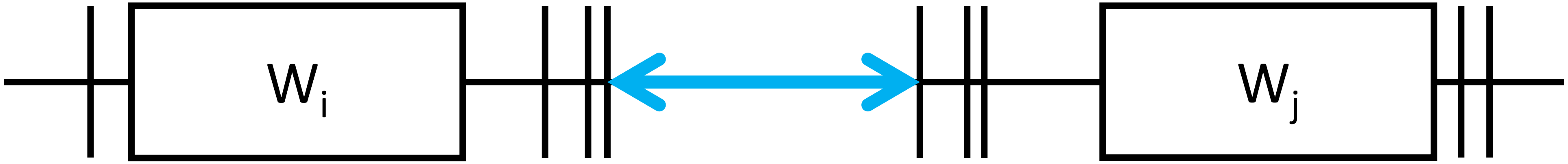


Persistence over time



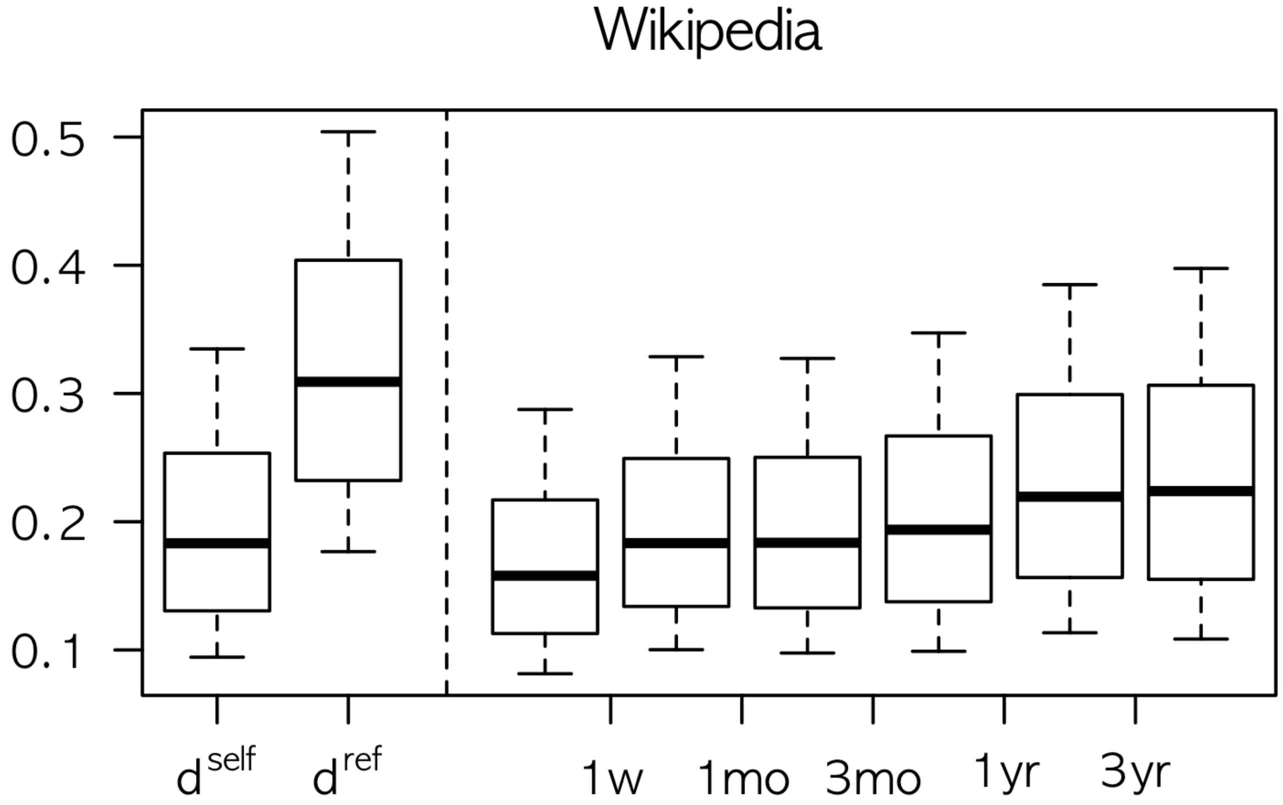
Do interval patterns persist after long inactivity?

- Binning d^{self} by the **longest interval** between two windows

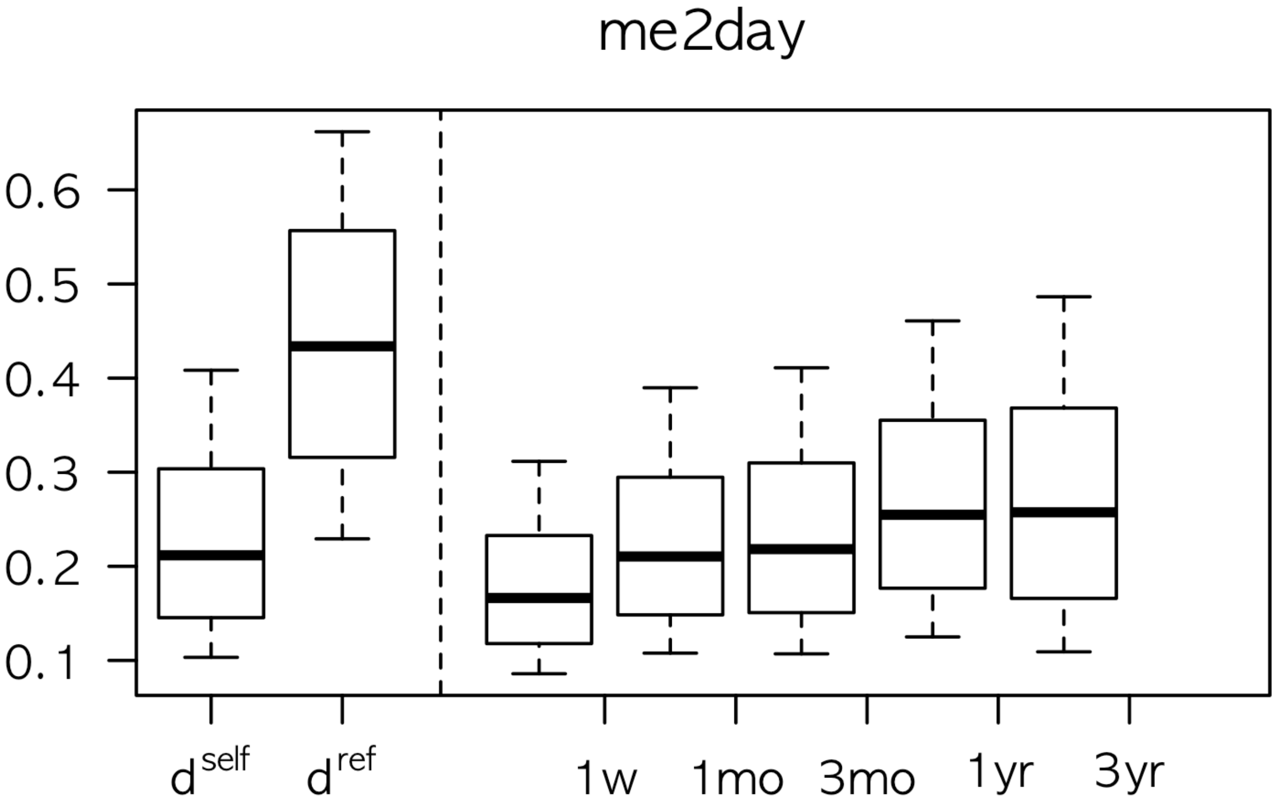


- Compare binned d^{self} with overall d^{ref}

Persistence after inactivity

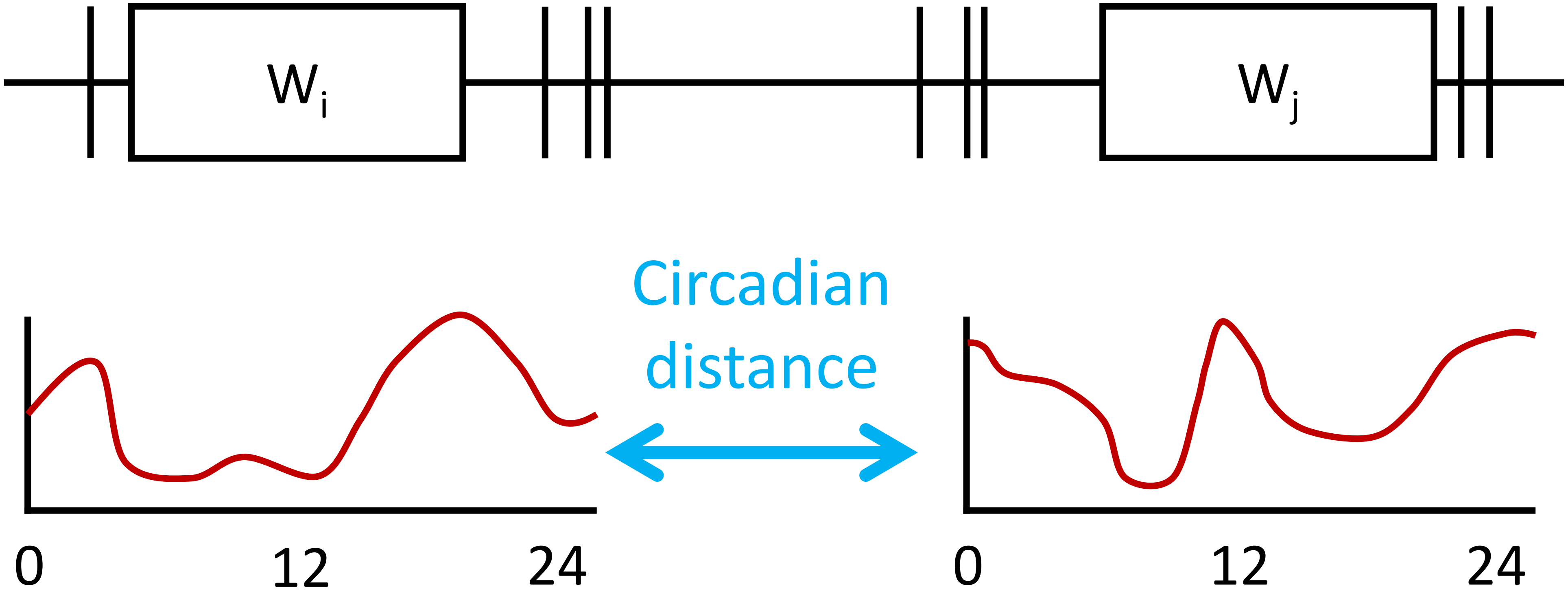


Persistence after inactivity

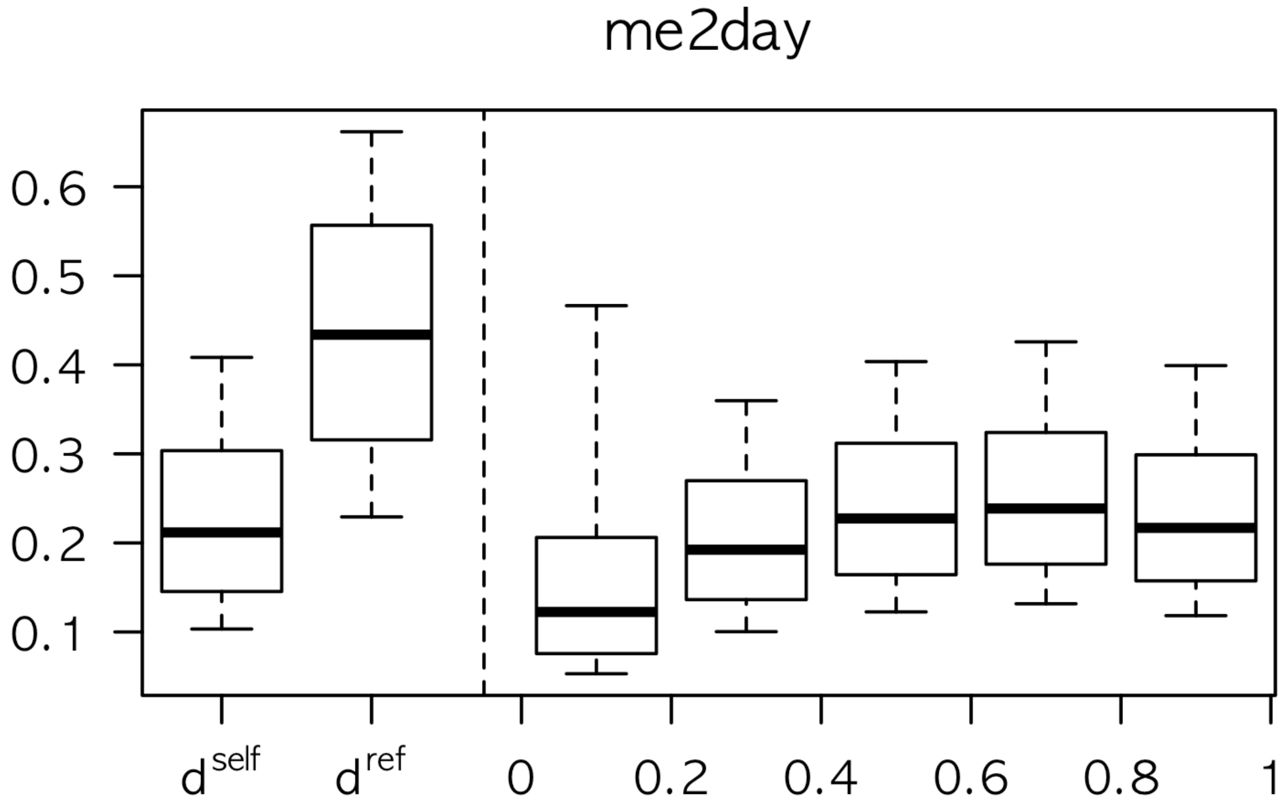


Do interval patterns persist through changing daily routine?

- Binning d^{self} by the **circadian distance** between two windows



Persistence through changing daily routine



In summary,

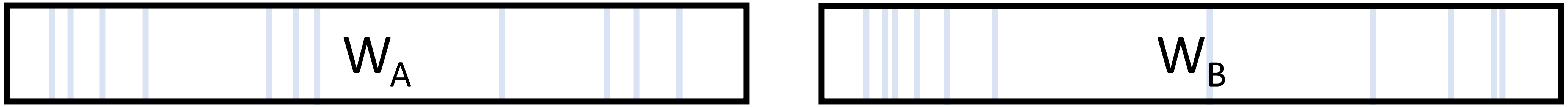
- Individuals have **interval signatures** that **persist over years**
- The signatures persist **even after coming back from long inactivity**
- The signatures persist **through changing daily routine**

APPLICATION

User Identification Using *Interval Signatures*

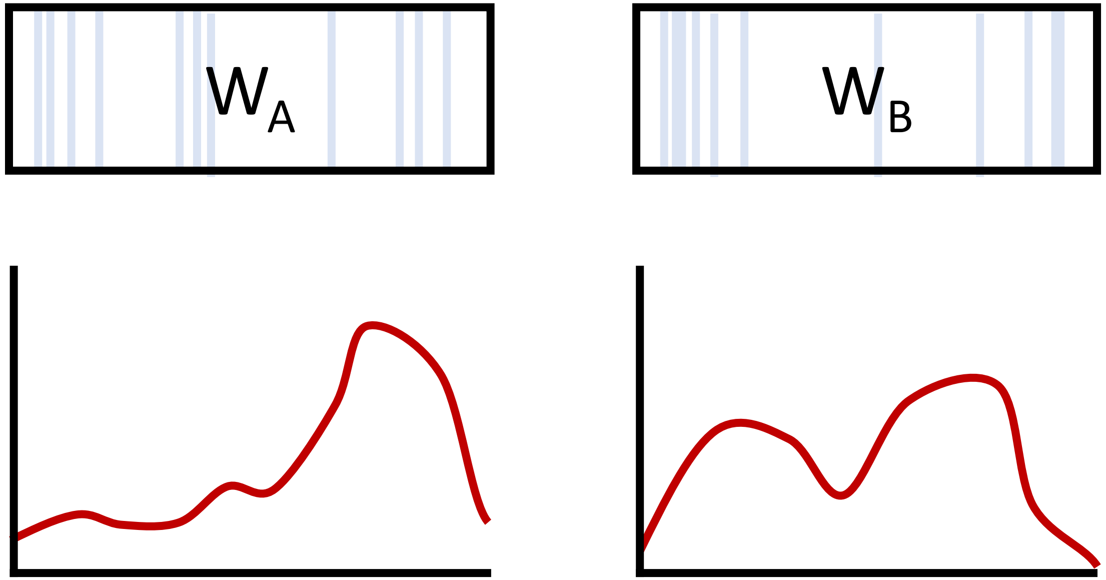
User identification: Problem definition

- Given two windows each containing 100 intervals



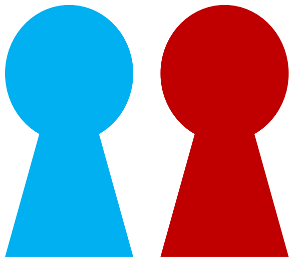
- Can we determine those from the same user or not?

A very simple identifier



Calculate the distance *d*

If *d* < threshold,

Else, 

Identification performance ($1 - \mathit{Equal\ Error\ Rate}$)

	Wikipedia	me2day	Twitter	Enron
Consecutive	80%	87%	83%	76%
> 1 year gap	71%	78%	76%	71%

- Performance of other behavioral biometrics
  - Keystroke dynamics: **~90%** [Peacock *IEEE S&P* 2004]
  - Mouse dynamics: **~80%** [Jorgensen *AsiaCCS* 2011]
  - Gaits: **~80%** [Gaufrov *University of Oslo* 2008]

Follow-up questions

- What do people with similar interval signatures have in common?
- What can be inferred about users by analyzing interval signatures?
- How interval signatures are related to other personal characteristics?

Interval Signature:

P-E-R-S-I-S-T-E-N-C-E and DISTINCTIVENESS of

Inter-event Time Distributions

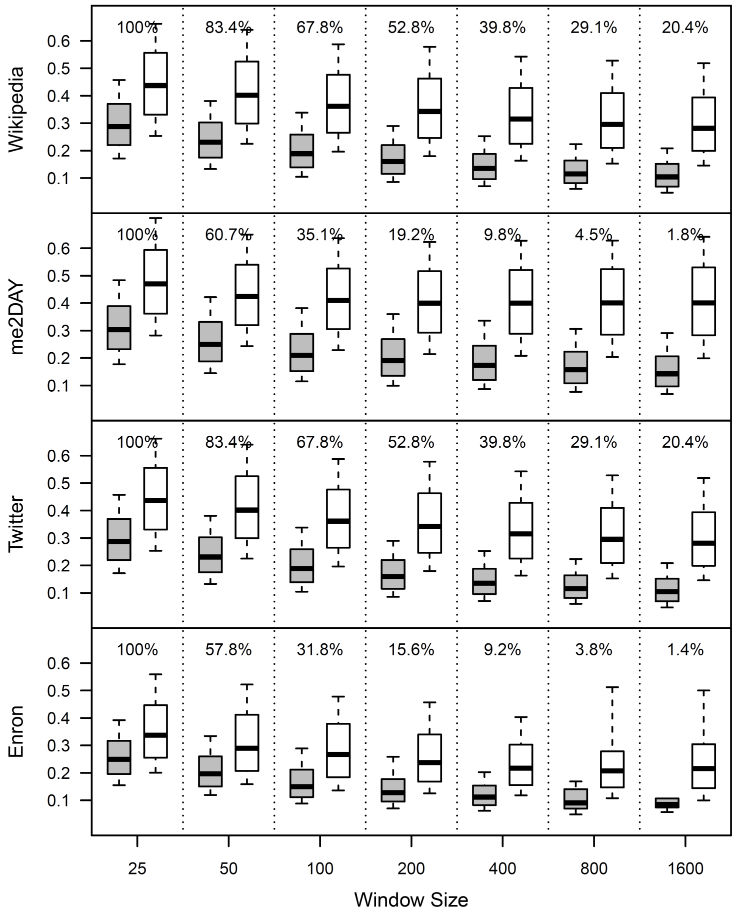
in Online Human Behavior

Q&A

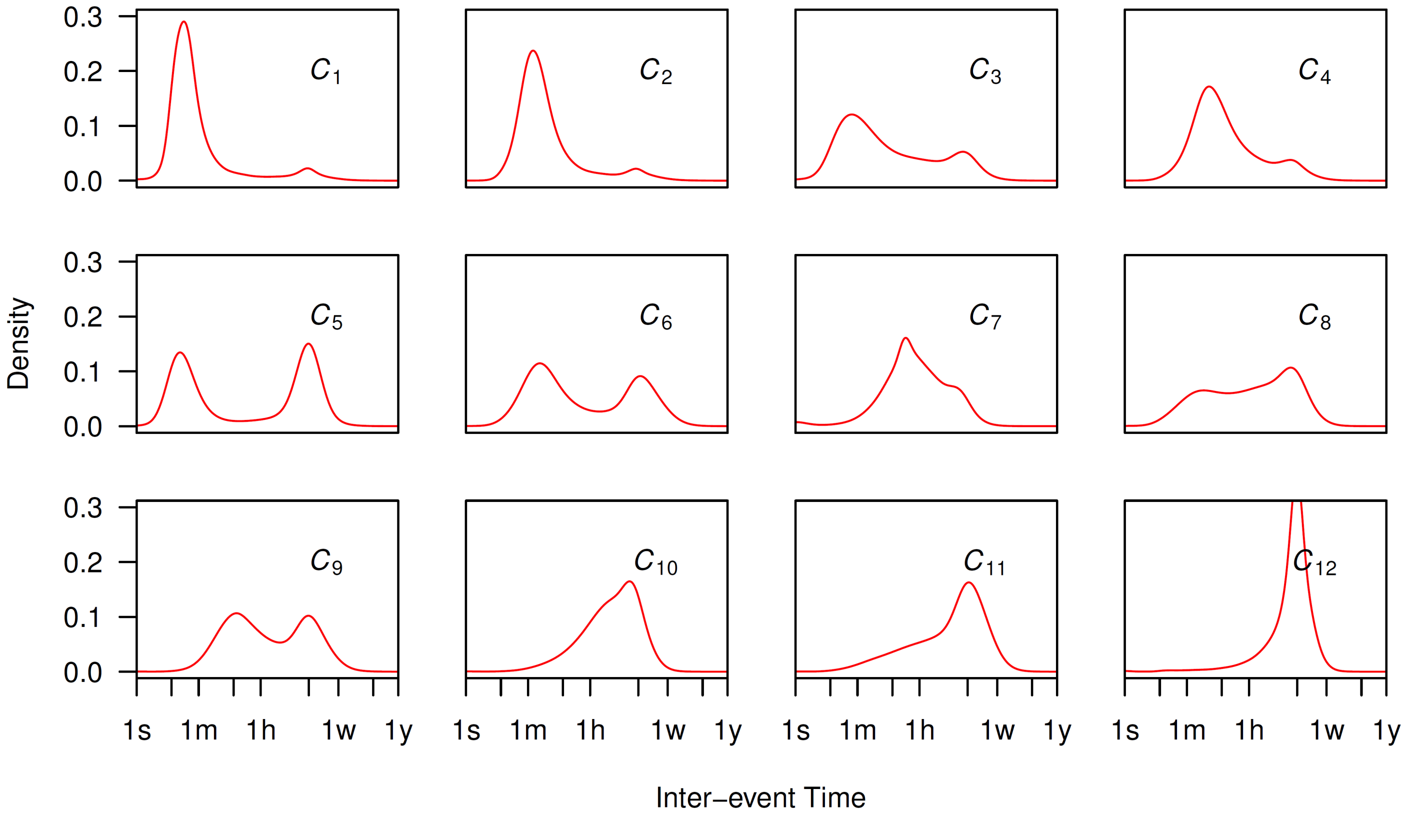
Dataset statistics

# of users	Wikipedia	me2day	Twitter	Enron
With >25 actions	521K	587K	921K	937K
With >100 actions	165K	203K	768K	542K
With >500 actions	47K	43K	334K	65K

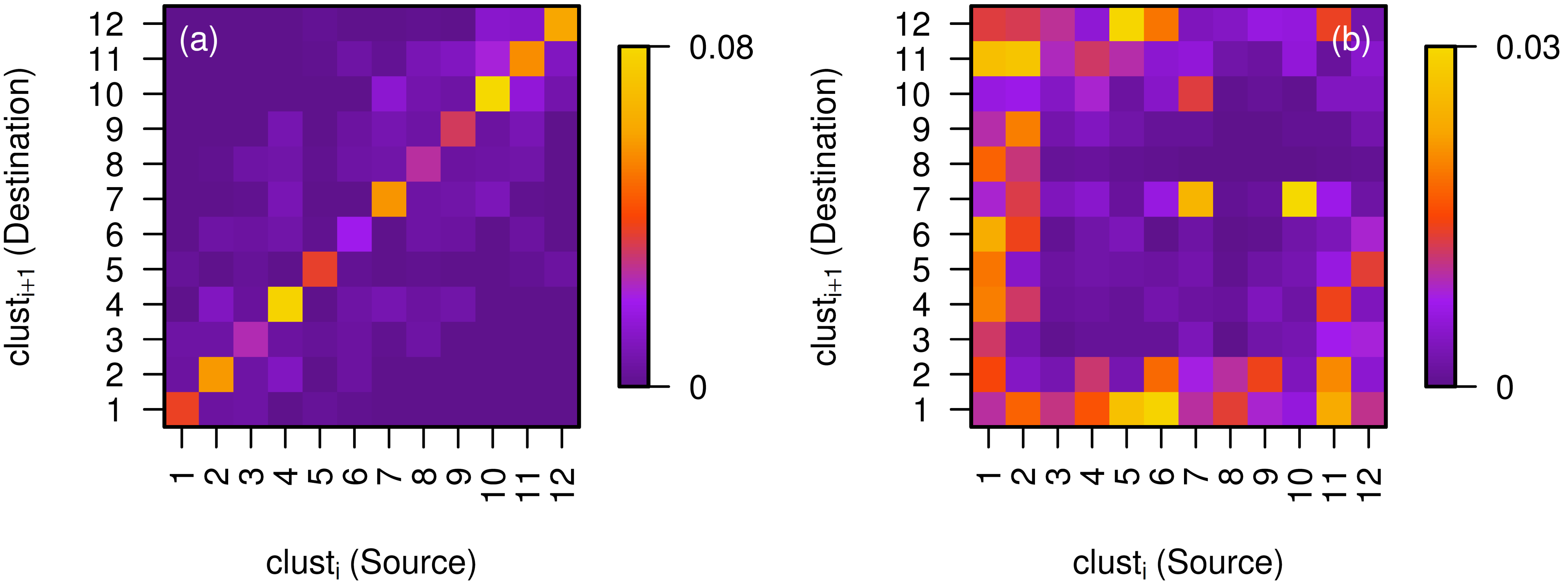
d^{self} vs d^{ref} at different window sizes



K-means clustering of interval patterns



Joint probability matrix for transition $W_i \rightarrow W_{i+1}$


(a)
(b)
clust_{i+1} (Destination)
clust_i (Source)