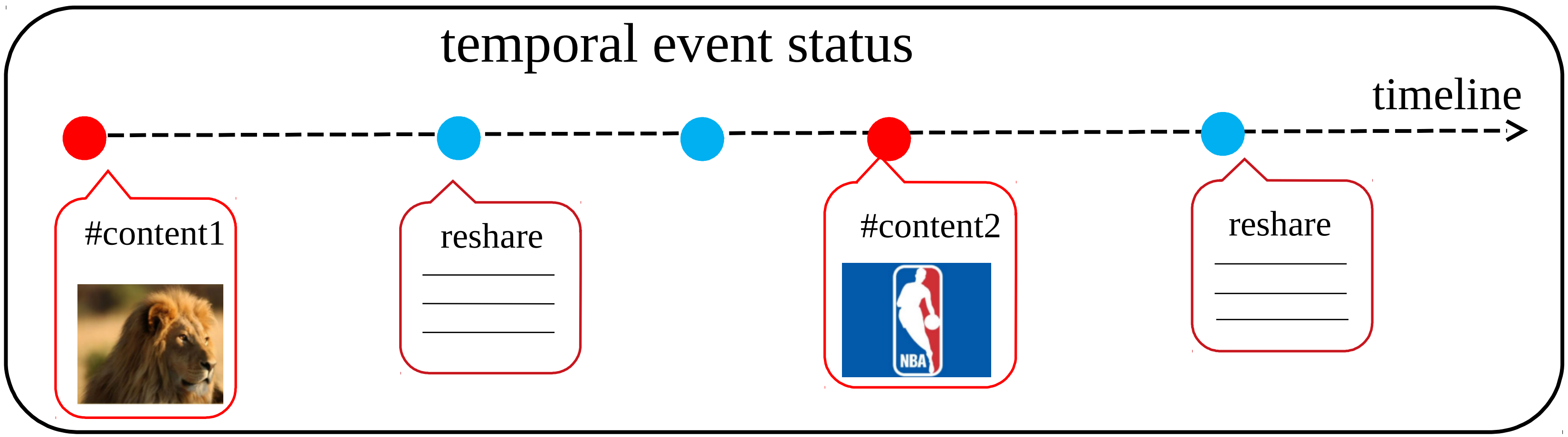
Adversarial Training Model Unifying Feature Driven and Point Process Perspectives for Event Popularity Prediction

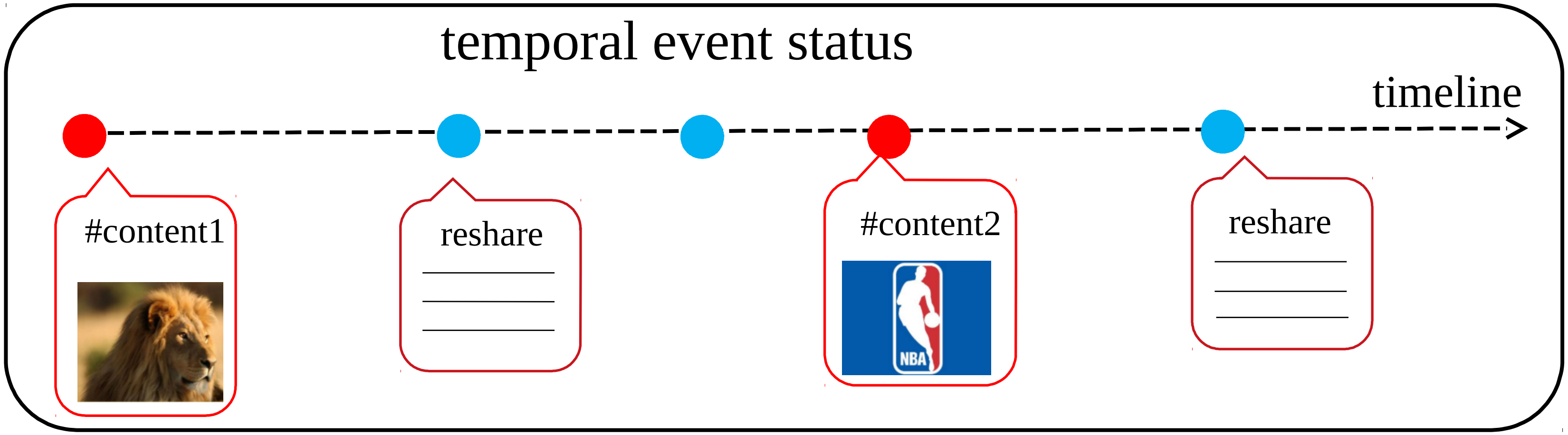
Chaoqi Yang
http://chaoqiyang.com
Senior Undergraduate
Shanghai Jiao Tong University

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Background


temporal event status
timeline
#content1
reshare
#content2
NBA
reshare

Background



Online social networks provide people with a quick access to information, communication and study.

- Fast Transmission Rate
- Good Timeliness
- Low Dissemination Cost

Background

Background

(from Twitter)

Background

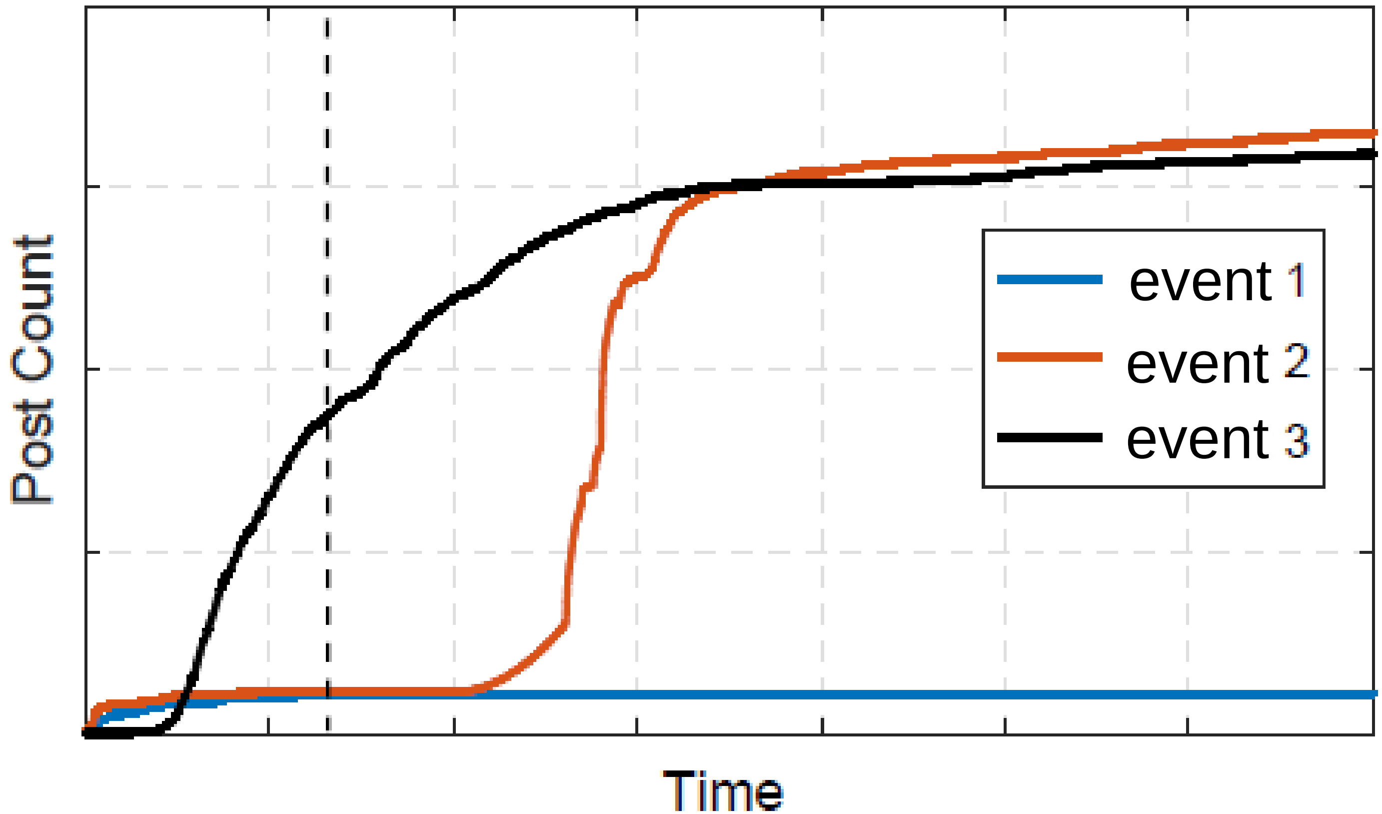
(from The Guardian)

Problem Statement

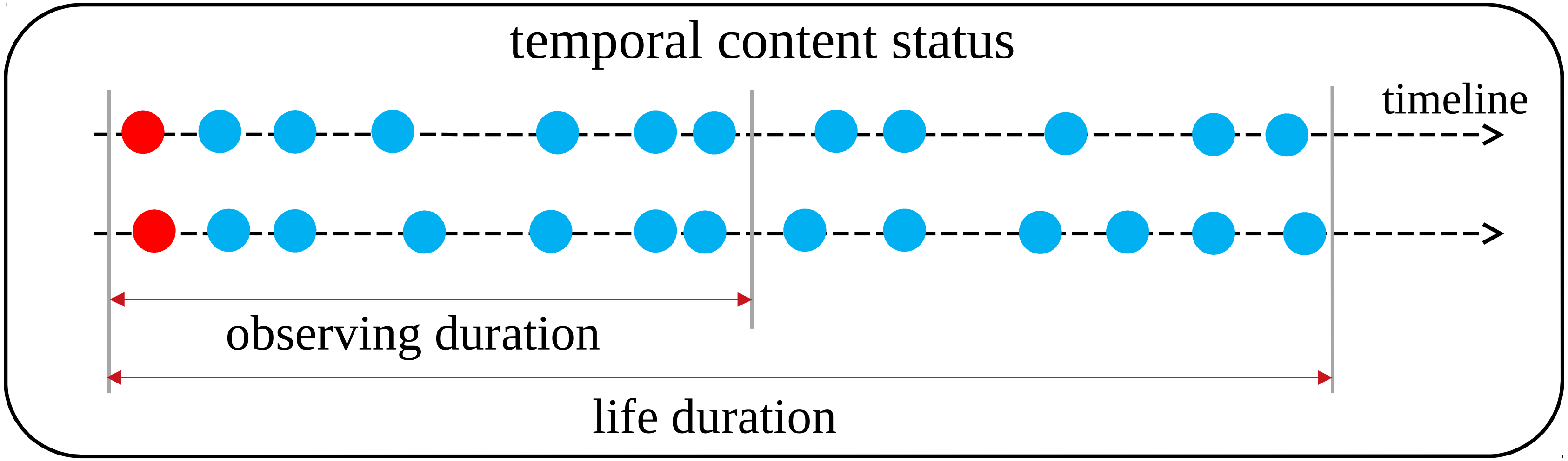
How can we predict the popularity?

Problem Statement

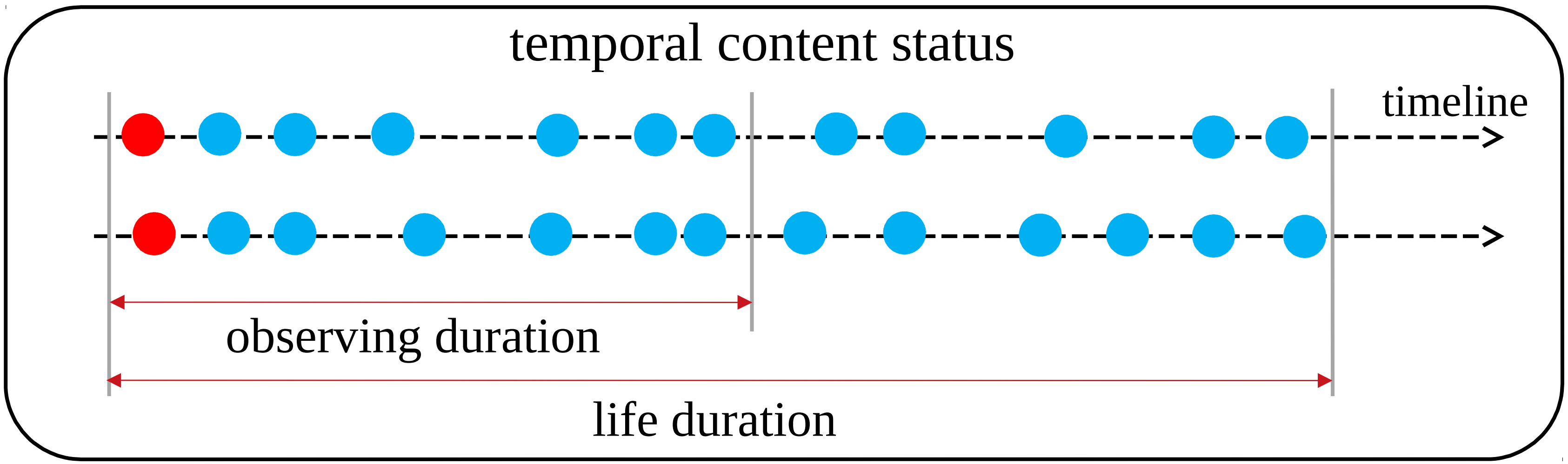
How can we predict the popularity?



Problem Statement



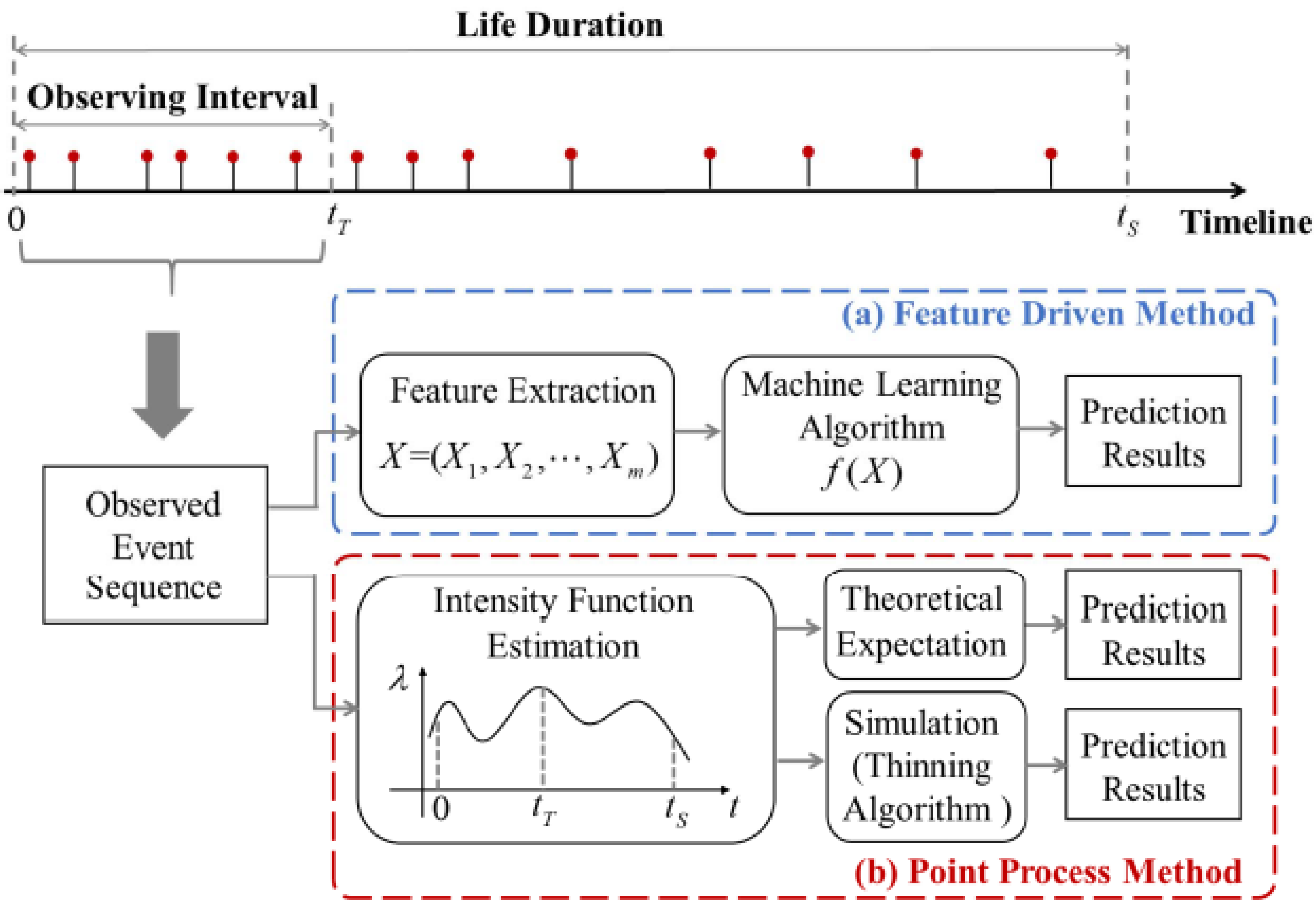
Problem Statement



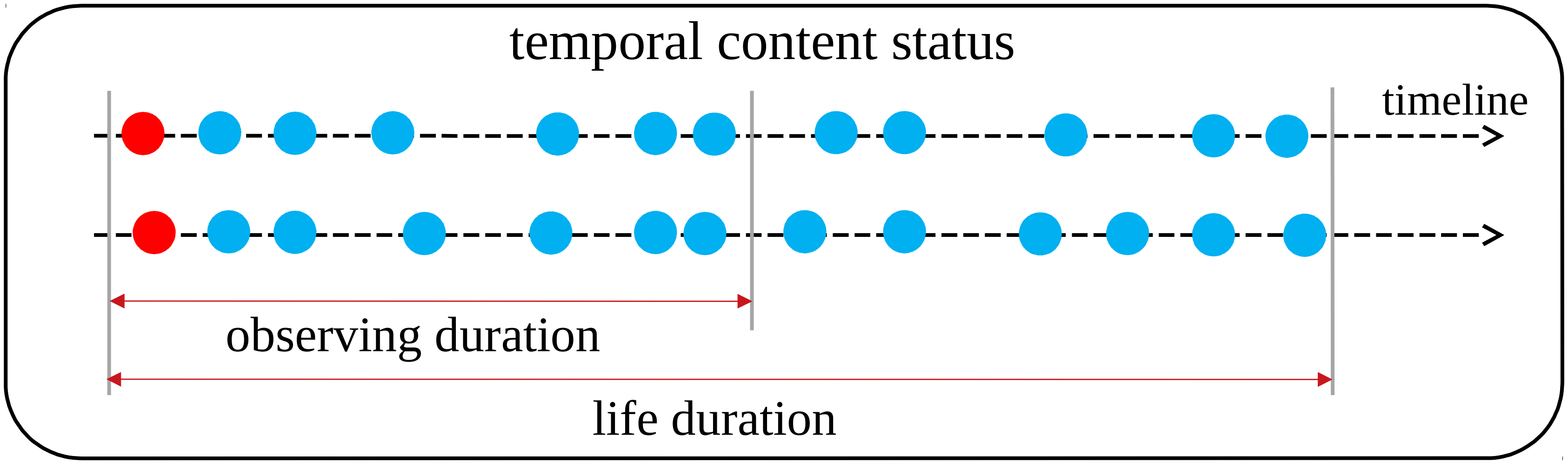
With tweets in observing duration, one would like to know how many tweets will be posted in the end.

- rumor monitoring
- anomaly detection
- personalized recommendation
- targeted advertisement

Two School of Thinking



Intuition



What can we get from observing duration?

Intuition

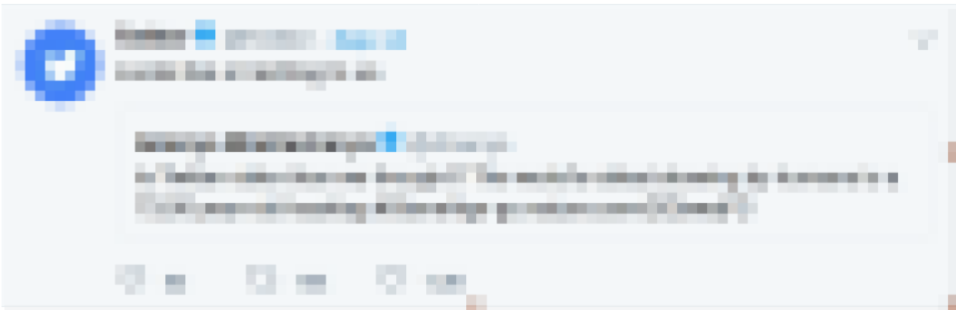
User Info.

- # of followers
- # of followees
- Network Age
- Reputation / Title
- …

Intuition

User Info.

- # of followers
- # of followees
- Network Age
- Reputation / Title
- …

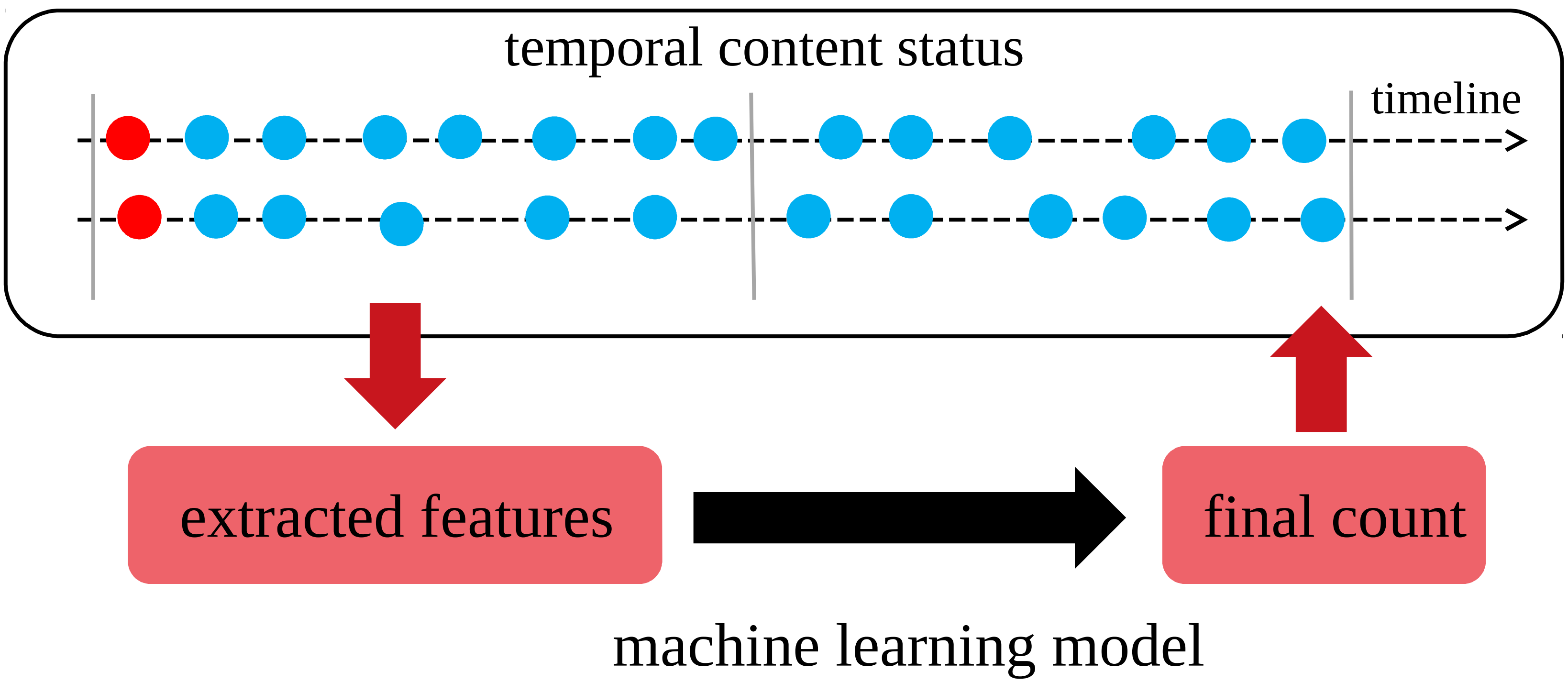


Tweet Info.

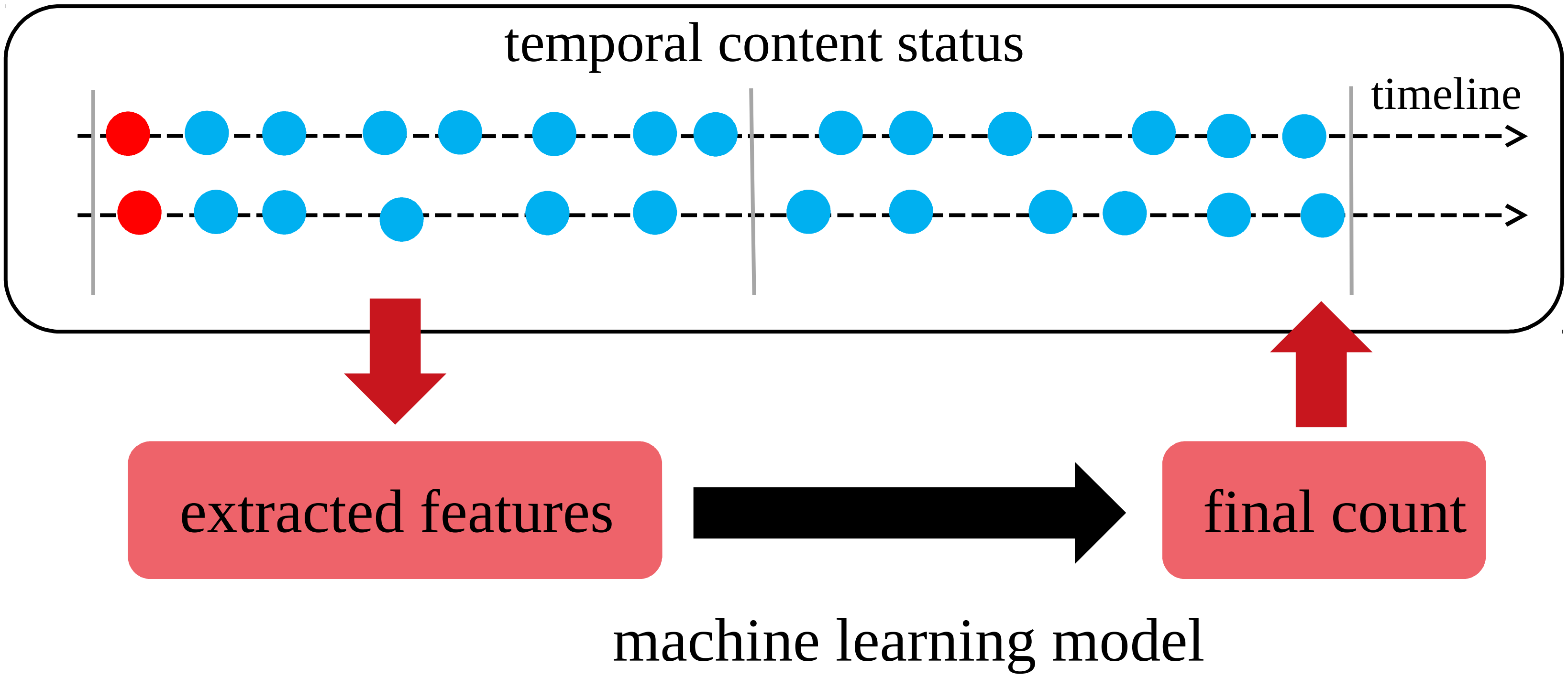
- timestamp
- # of paragraphs / words
- hashtag or not
- …

<u>Features</u>

One School of Thinking: Feature Driven Method



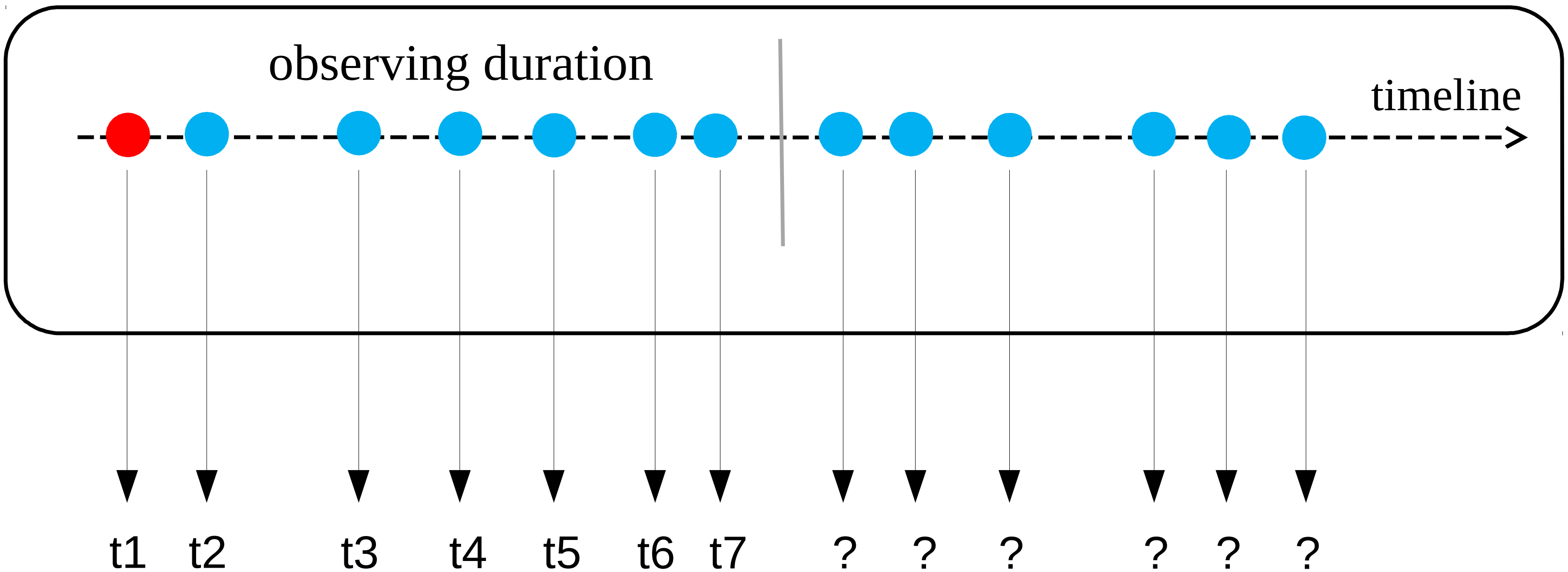
One School of Thinking: Feature Driven Method



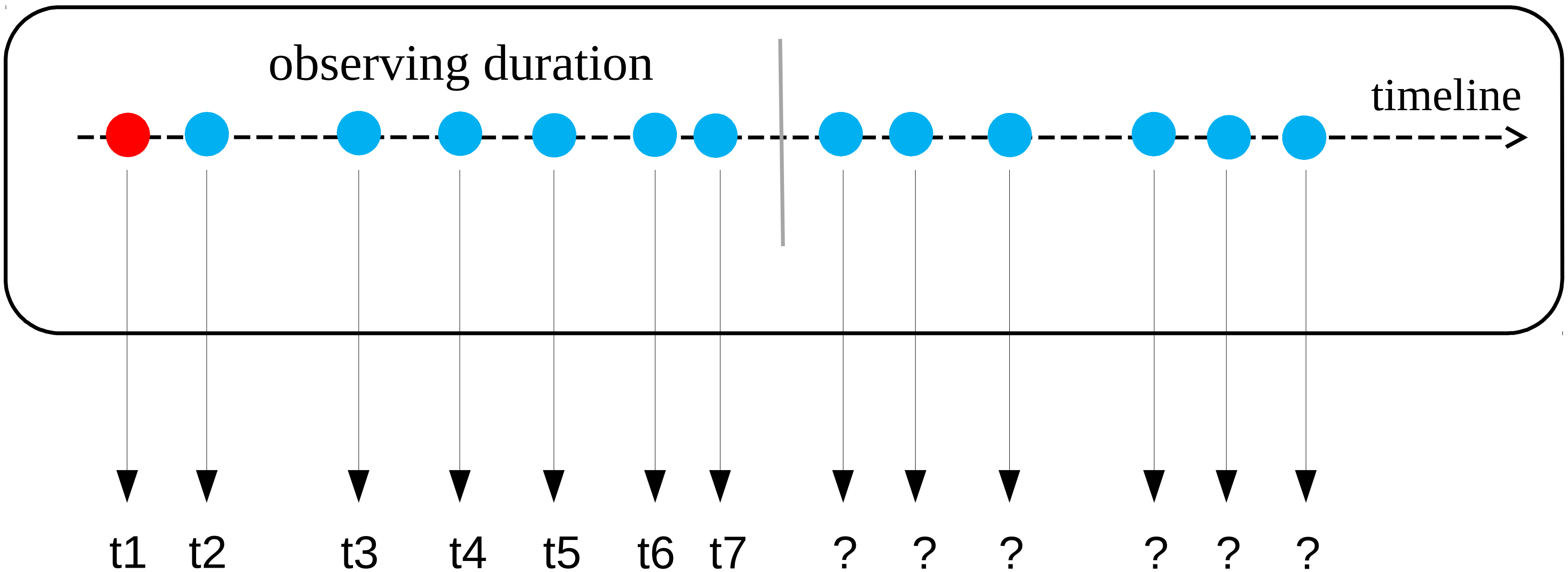
Strengths: **good precision**, **robustness**, **stability**

Weaknesses: **high human expertise**, **laborious feature engineering**

Intuition

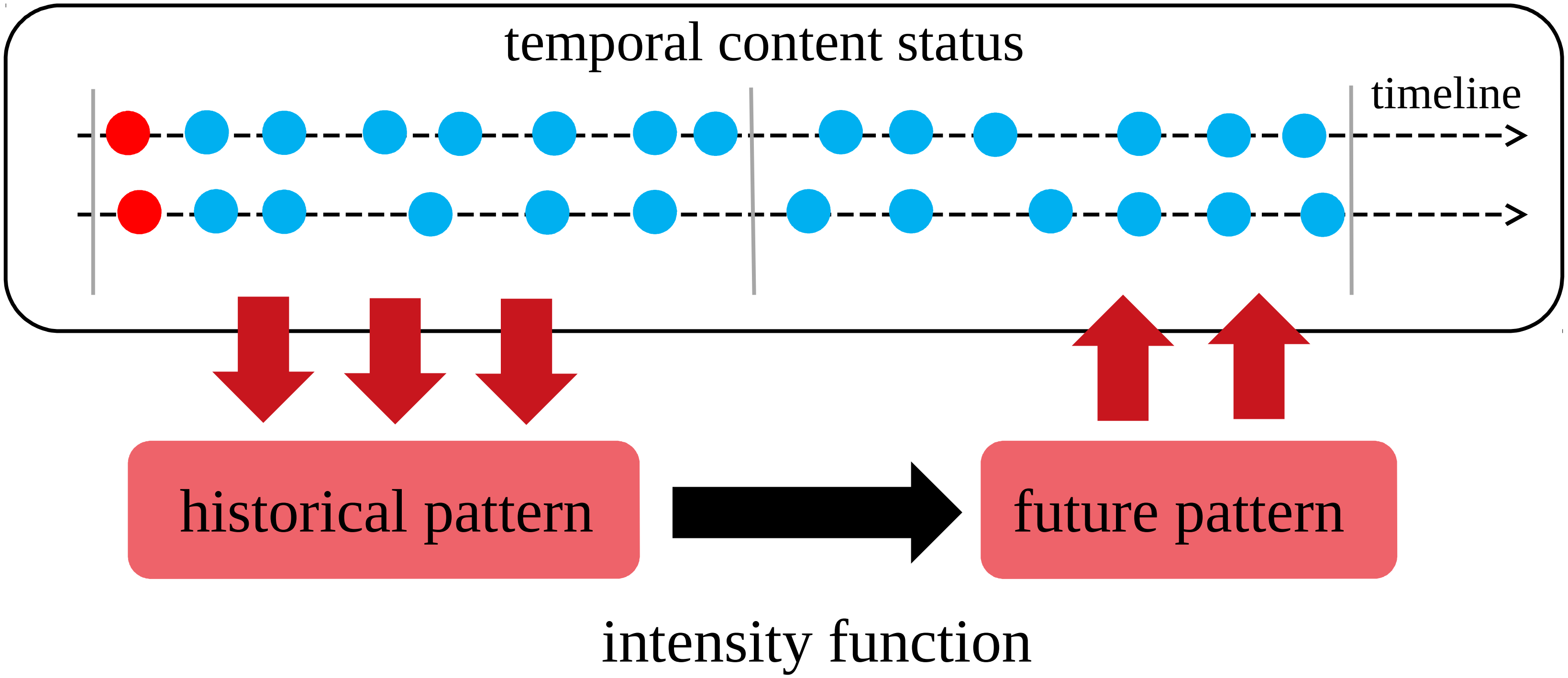


Intuition

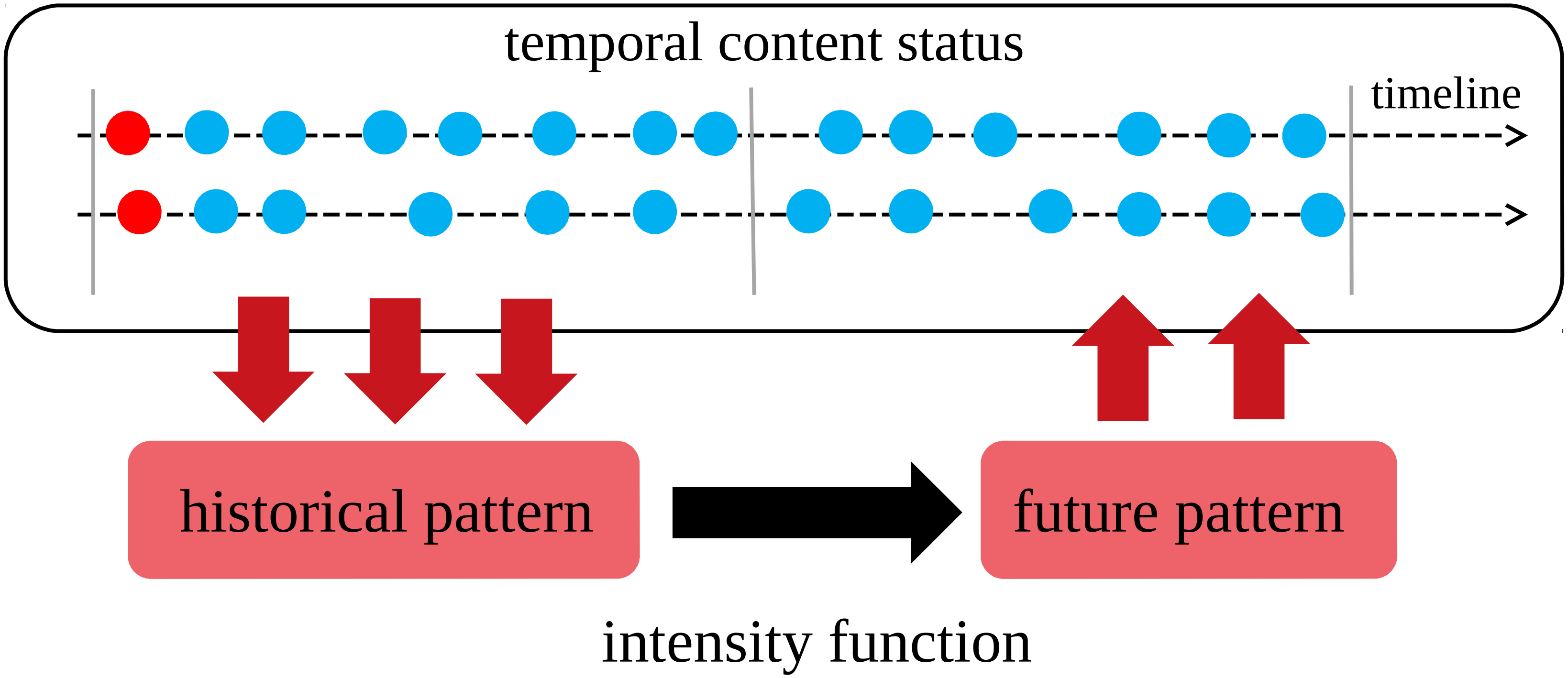


Point process / Poisson Process → Intensity function → Estimate future popularity

The Other School of Thinking: Point Process Method



The Other School of Thinking: Point Process Method



Strengths: **good interpretability**, **low computational complexity**

Weaknesses: **poor robustness, long-term observation**

Limitations of Previous Study

The most previous methods rely on long-term observation:

- utilize features extracted from dozens of hours observation to predict popularity of social topics.
- predict popularity of Twitter hashtags when over 80% related posts are exposed.

Limitations of Previous Study

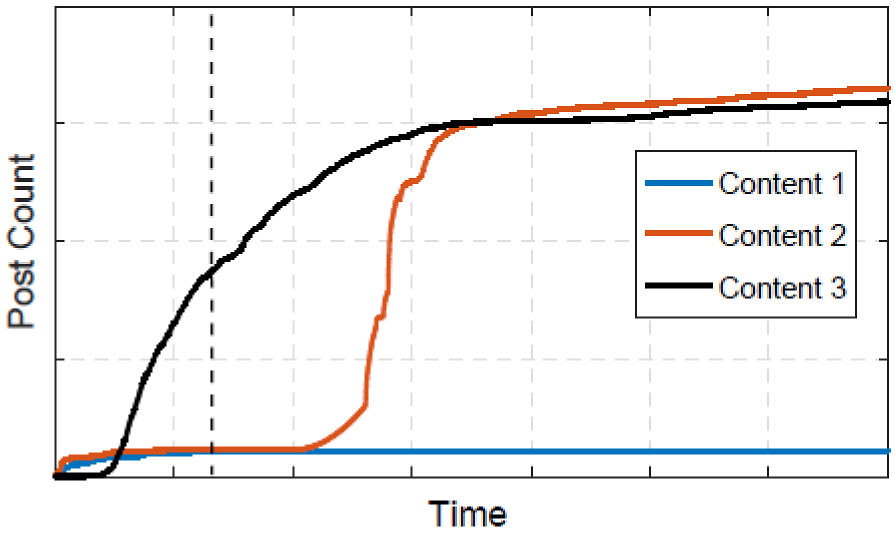
The most previous methods rely on long-term observation:

- utilize features extracted from dozens of hours observation to predict popularity of social topics.
- predict popularity of Twitter hashtags when over 80% related posts are exposed.

The long observation time makes it easy for decision making:

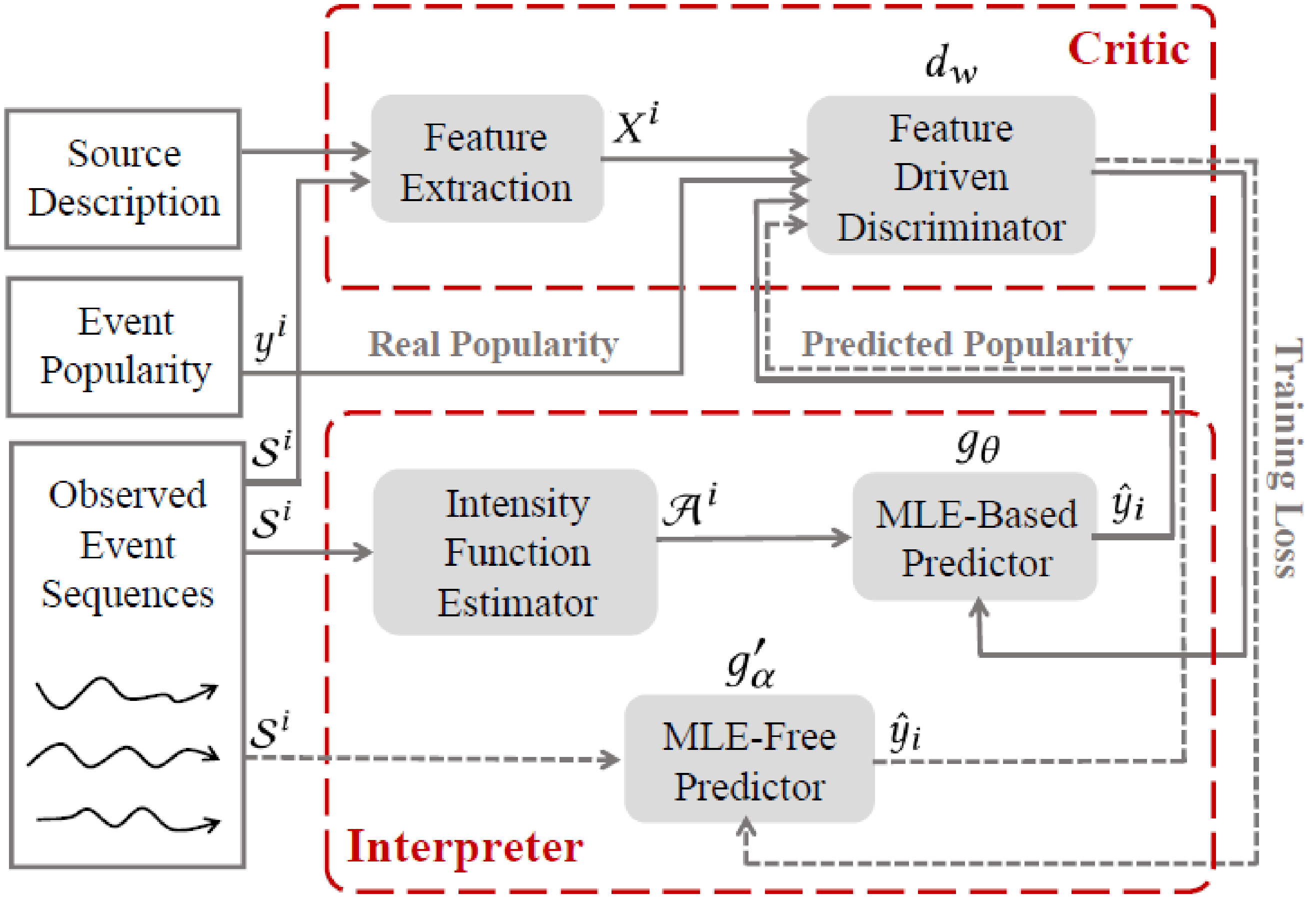
- feature driven method can easily extract some effective observable features that strongly correlate with popularity.
- point process method can recognize some typical patterns hidden in the observed temporal sequence and further confidently infer the future evolution.

Challenges



- some contents with similar early-stage evolution could generate quite different popularity.
- some contents with totally different evolution trends may reach a similar popularity.

Model Framework



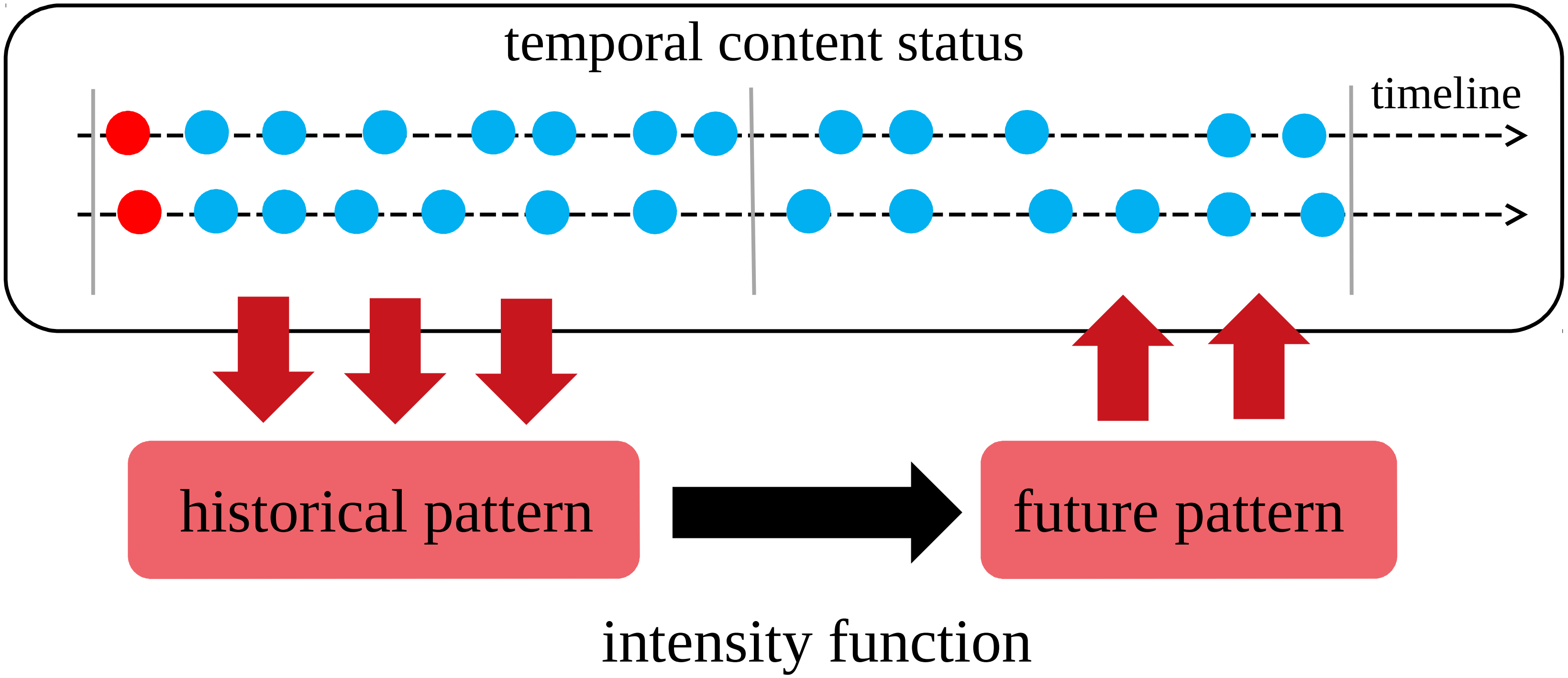
Adversarial Minimax Game

$$\min_{\theta} \max_{w, ||d_w||_L \leq 1} \mathbb{E}_{(X,y)\sim\mathbb{P}_r}[d_w(X,y)] - \mathbb{E}_{(X,g_\theta(\mathcal{A}))\sim\mathbb{P}_g}[d_w(X,g_\theta(\mathcal{A}))]$$

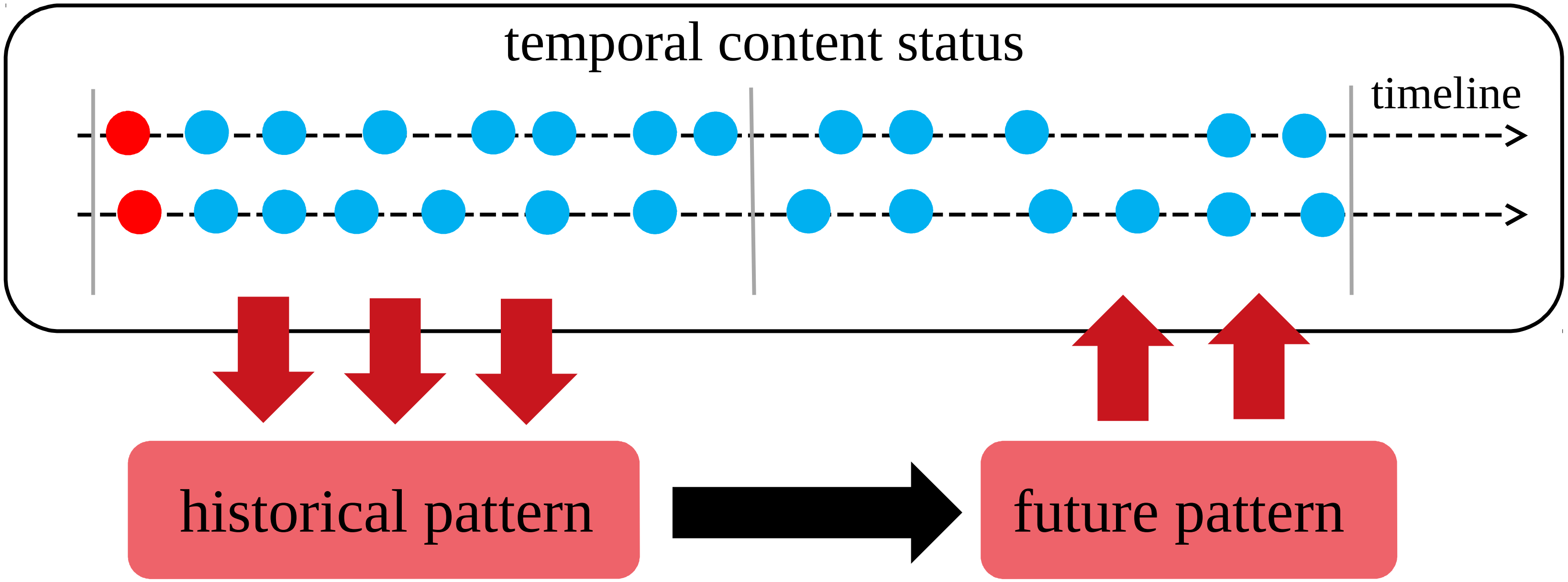
To implement the idea, we let

- The feature driven model (critic) aims to maximize the distance of distributions for real and predicted popularity
- The point process model (interpreter) aims to minimize the distance by giving a convincing prediction
- The training stage will be terminate once an equilibrium is achieved between two players.

Interpreter: Point Process Model



Interpreter: Point Process Model



intensity function

Using Recurent Neural Network (RNN) $\begin{cases} \text{MLE-based (MB)} \\ \text{MLE-free (MF)} \end{cases}$

MLE-based RNN Model

Point Process Theory:

Intensity function: $\lambda^*(t) = \lambda(t|\mathcal{S}^i(t)) = \dfrac{\mathbb{P}\{N^i(t+dt) - N^i(t) = 1|\mathcal{S}^i(t)\}}{dt}$

Conditional density function: $f^*(t) = \lambda^*(t)\exp(-\int_{t_n}^{t}\lambda^*(\tau)d\tau)$

MLE-based RNN Model

Point Process Theory:

Intensity function: $\lambda^*(t) = \lambda(t|\mathcal{S}^i(t)) = \dfrac{\mathbb{P}\{N^i(t+dt) - N^i(t) = 1|\mathcal{S}^i(t)\}}{dt}$

Conditional density function: $f^*(t) = \lambda^*(t)\exp(-\int_{t_n}^{t} \lambda^*(\tau)d\tau)$

Intensity sequence (RNN):

$$\boldsymbol{h}_{j+1} = \max\left\{\boldsymbol{w}_h \cdot t_j^i + \boldsymbol{v}_h \cdot \boldsymbol{h}_j + \boldsymbol{b}_h, 0\right\}$$

$$\lambda^*(t) = \exp(\boldsymbol{v}_t \cdot \boldsymbol{h}_j + w_t(t - t_j^i) + b_t)$$

$$l(C) = \sum_i \sum_j \log f^*(t_{j+1})$$

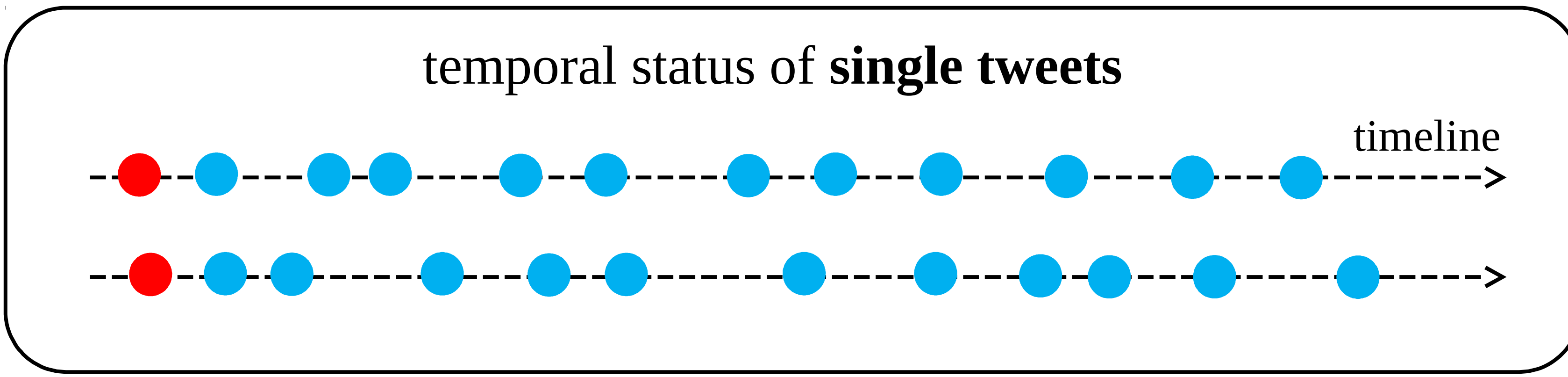
Popularity sequence (RNN):

$$\boldsymbol{h}_{j+1} = \tanh(\boldsymbol{v}_h' \cdot A_j^i + \boldsymbol{w}_h' \cdot \boldsymbol{h}_j + \boldsymbol{b}_h'),$$

$$\hat{y} = \boldsymbol{w}_o \cdot \boldsymbol{h}_M + b_o,$$

MLE-free RNN Model

Observed sequence:



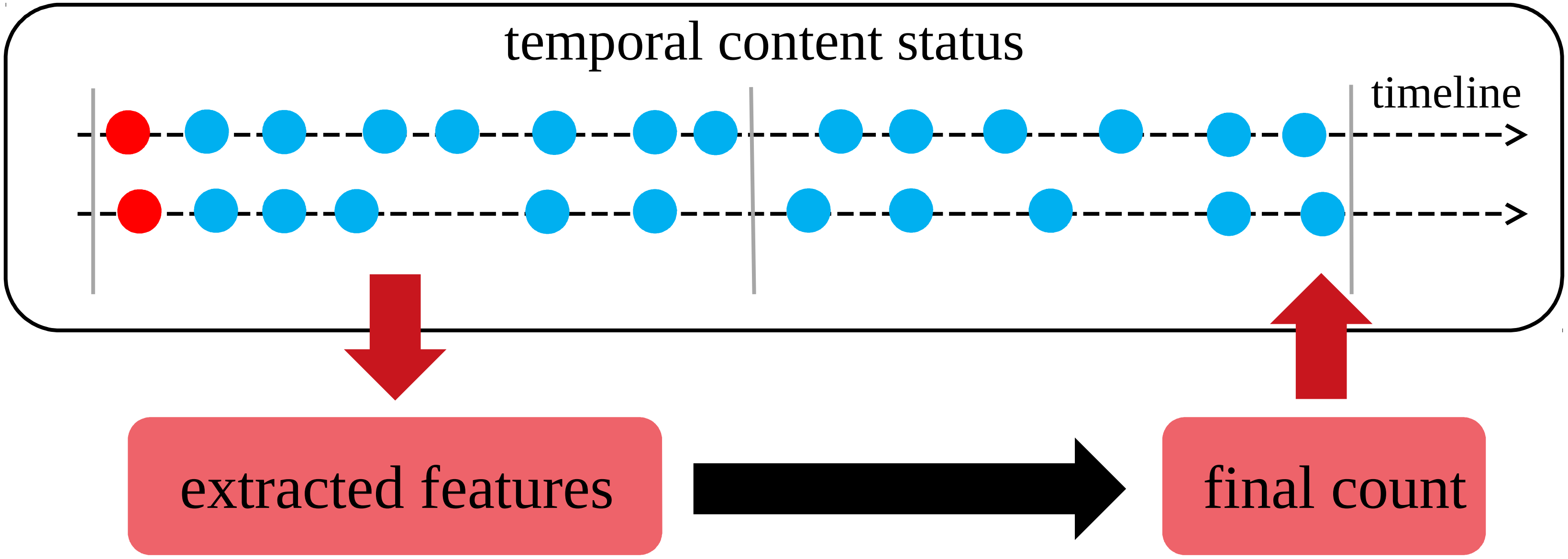
Popularity sequence (RNN):

$$\boldsymbol{h}_{j+1} = \tanh(\boldsymbol{v}_h'' \cdot t_j^i + \boldsymbol{w}_h'' \cdot \boldsymbol{h}_j + \boldsymbol{b}_h''),$$

$$\hat{y} = \boldsymbol{w}_o' \cdot \boldsymbol{h}_M + b_o'.$$

Critic: Feature Driven Model

3-Layer ELU Neural Network:



Feature Extraction:

- mean and standard deviation of time interval between events
- number of events in the first/second half of observing interval
- ……

Adversarial Training

Wasserstein Distance (Earth Moving Distance):

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\phi \sim \Phi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(\eta, \zeta) \sim \phi}[||\eta - \zeta||]$$

Adversarial Training

Wasserstein Distance (Earth Moving Distance):

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\phi \sim \Phi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(\eta, \zeta) \sim \phi}[||\eta - \zeta||]$$

Global Training Objection:

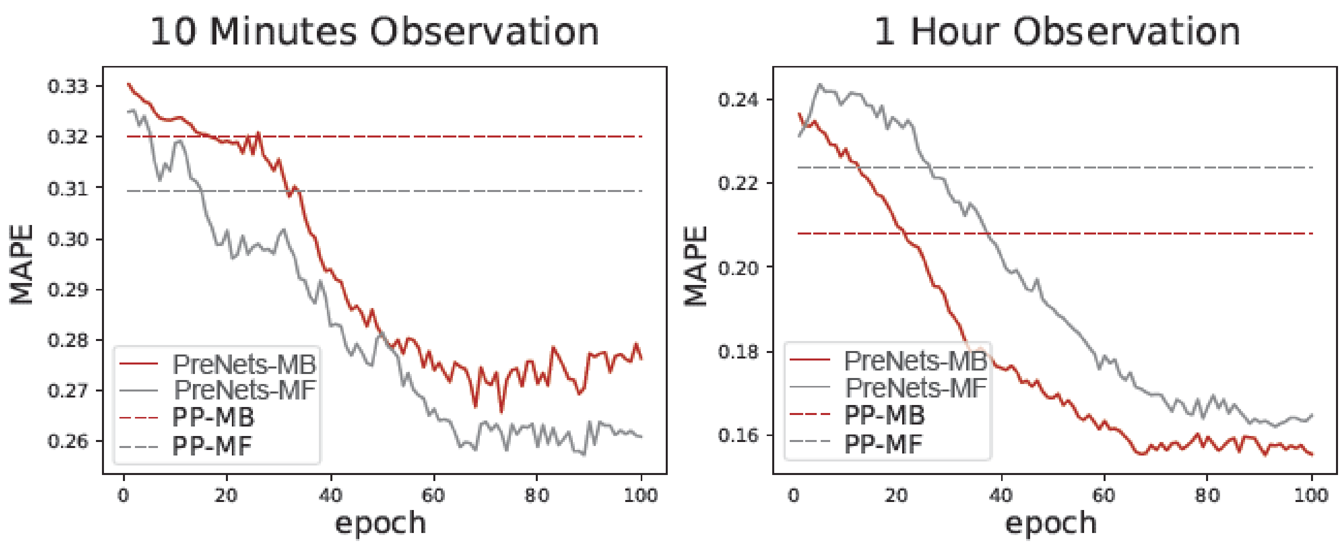
$$\min_{\theta} \max_{w, ||d_w||_L \leq 1} \mathbb{E}_{(X,y) \sim \mathbb{P}_r}[d_w(X, y)] - \mathbb{E}_{(X, g_\theta(\mathcal{A})) \sim \mathbb{P}_g}[d_w(X, g_\theta(\mathcal{A})]$$

More Specifically:

$$\min_{\theta} \max_{w} \frac{1}{B} \sum_{i=1}^{B} [d_w(X^i, y^i)] - \frac{1}{B} \sum_{i=1}^{B} [d_w(X^i, g_\theta(\mathcal{S}^i)] - \gamma \sum_{i,j=1}^{B} \left| \frac{|d_w(X^i, y^i) - d_w(X^j, g_\theta(\mathcal{S}^j)|}{|y^i - g_\theta(\mathcal{S}^j)|} - 1 \right|$$

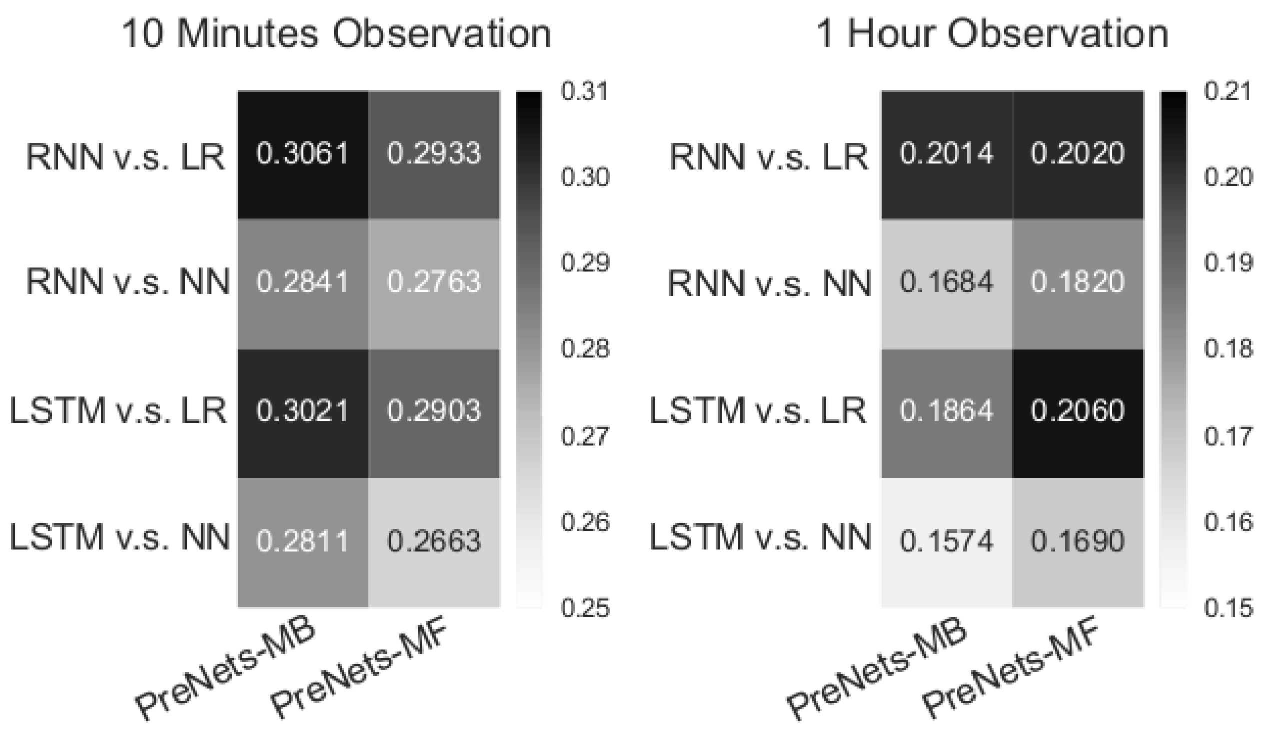
Experiment Results (1)

Learning curves of MAPE:



- Adversarial training improves prediction precision compared with independent training

Experiment Results (2)



- compare different "Critics" and "Interpreters"

Thank you!

Q&A