Leveraging the **Crowd** to Detect and Reduce the Spread of **Fake News** and **Misinformation**

Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, Manuel Gomez Rodriguez / MPI + KAIST

img credit: http://www.pewinternet.org/2017/10/19/the-future-of-truth-and-misinformation-online/

1. Post-Truth Era and Fake News

2. Crowd-Powered Fact-Checking

3. Fact-Checking Scheduling Problem

4. Algorithm: *CURB*

5. Detect and Reduce the Spread of Fake News

1. Post-Truth Era and Fake News

2. Crowd-Powered Fact-Checking

3. Fact-Checking Scheduling Problem

4. Algorithm: *CURB*

5. Detect and Reduce the Spread of Fake News

img credit: http://static.adweek.com/adweek.com-prod/wp-content/uploads/2017/07/FakeNewsOnLaptop-840x460.jpg

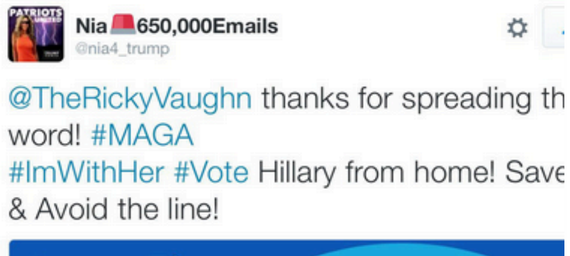
Post-Truth

Relating to or denoting circumstances in which objective **facts** are less influential in shaping public opinion than appeals to **emotion** and **personal belief**.

- Oxford Dictionaries

img credit: http://www.creativityworks.net/

Opinionated, Inaccurate **Fake News**



1. Post-Truth Era and Fake News

2. **Crowd-Powered Fact-Checking**

3. Fact-Checking Scheduling Problem

4. Algorithm: *CURB*

5. Detect and Reduce the Spread of Fake News

img credit: https://crowdsourcedtesting.com/resources/wp-content/uploads/2014/05/canstockphoto12241921.jpg

Crowd-powered Fact-checking

Shares & Re-Shares

Exposures

Flags

Fact-Check

Crowd-powered Fact-checking

Christine

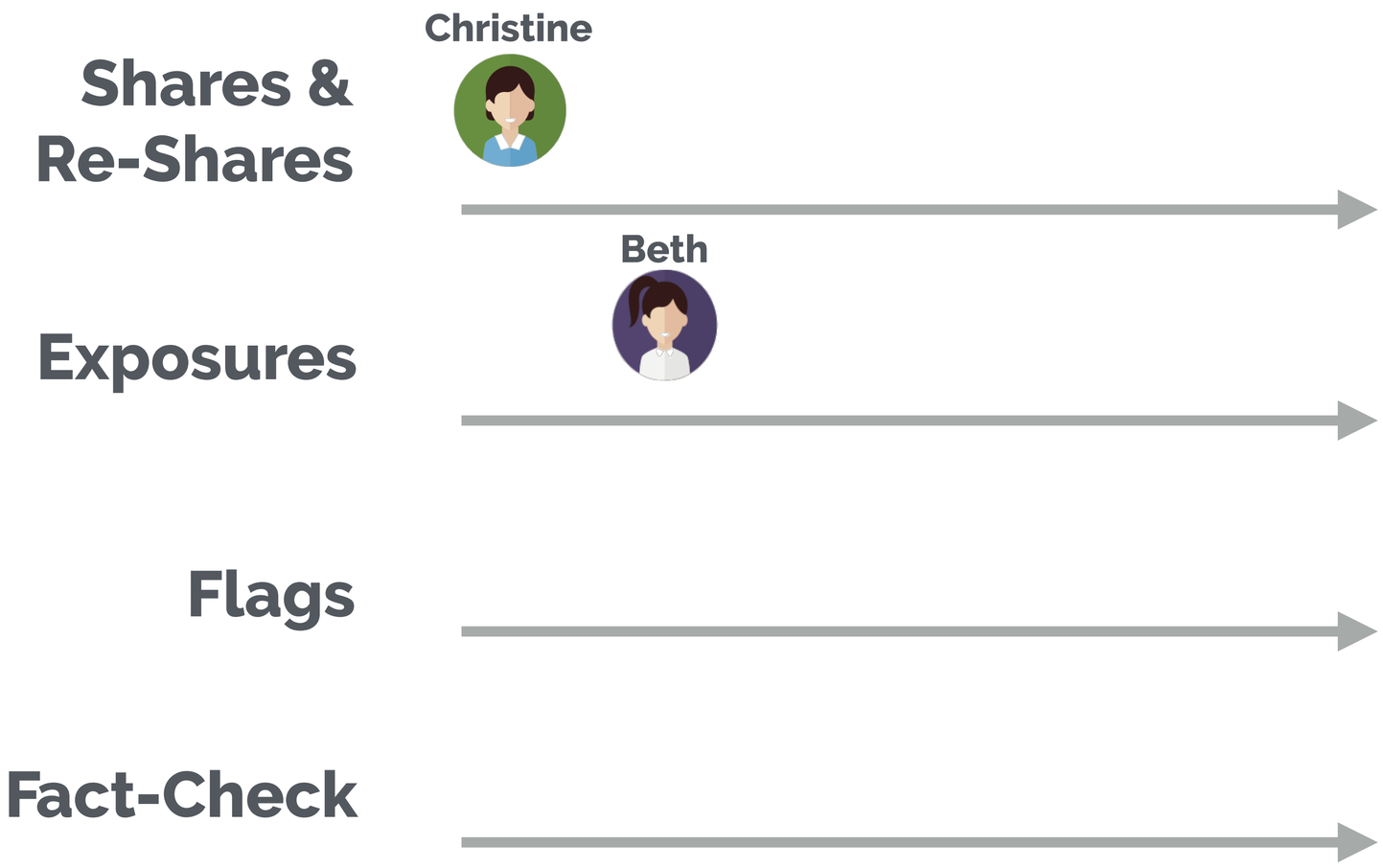
Shares & Re-Shares

Exposures

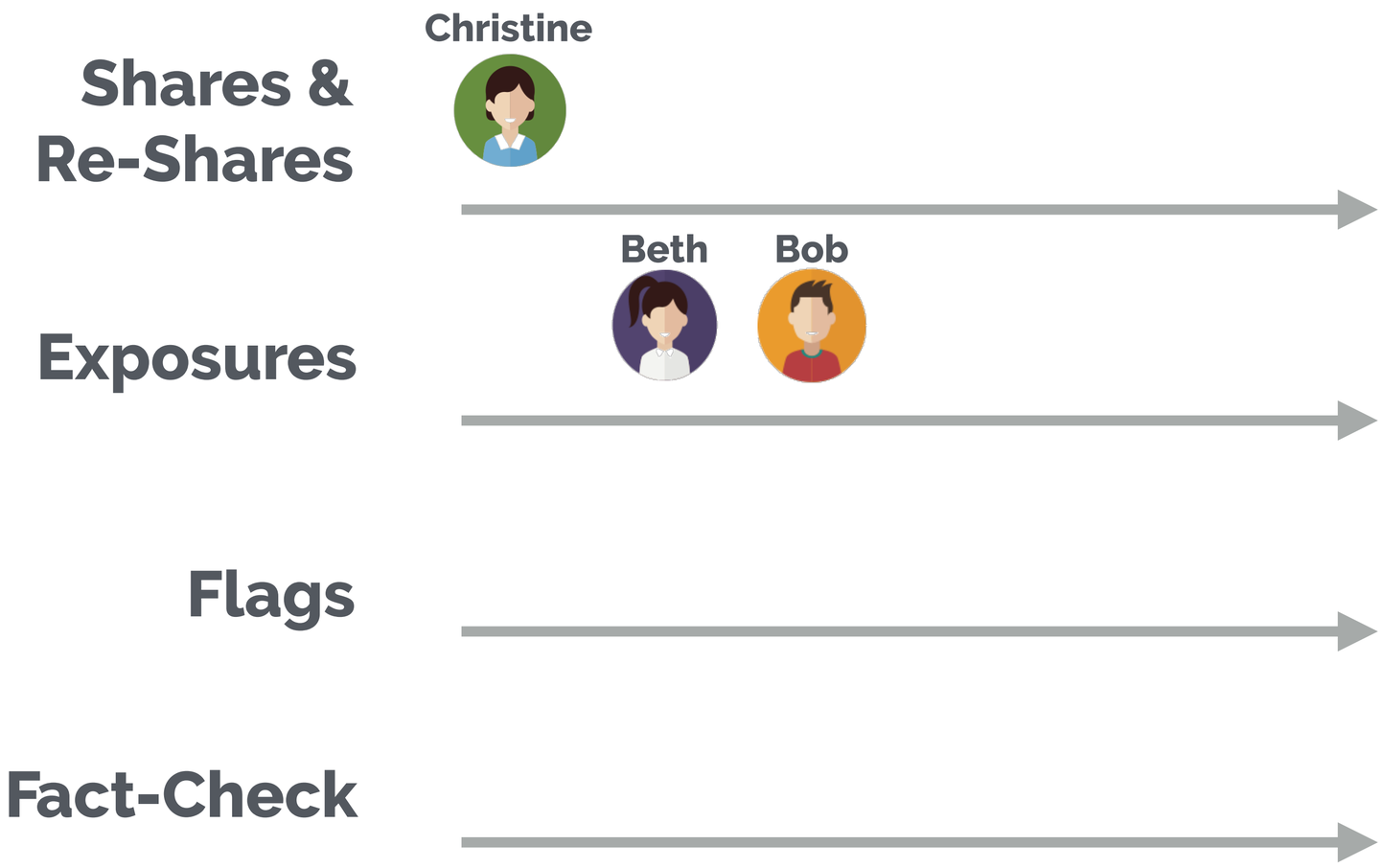
Flags

Fact-Check

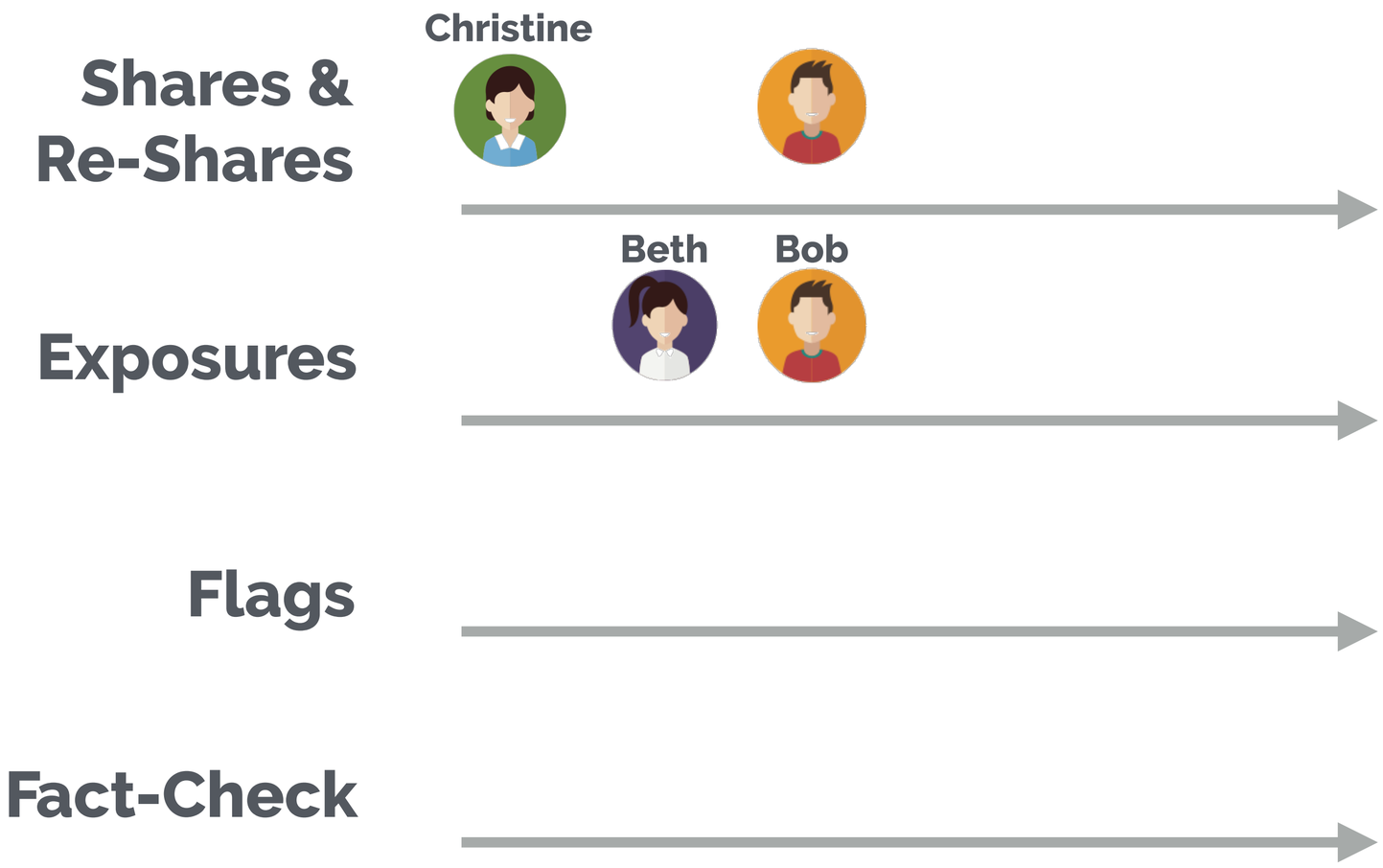
Crowd-powered Fact-checking



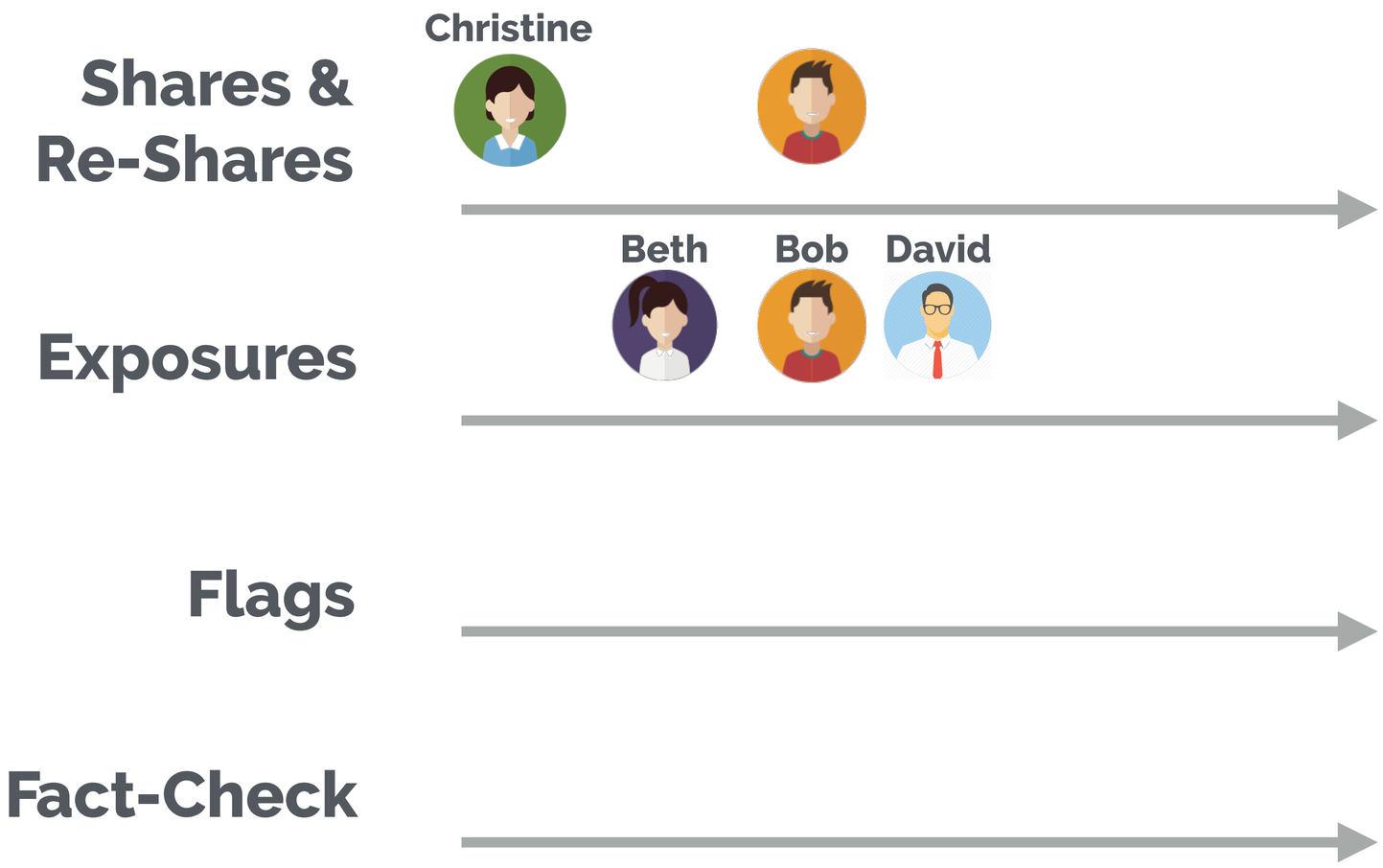
Crowd-powered Fact-checking



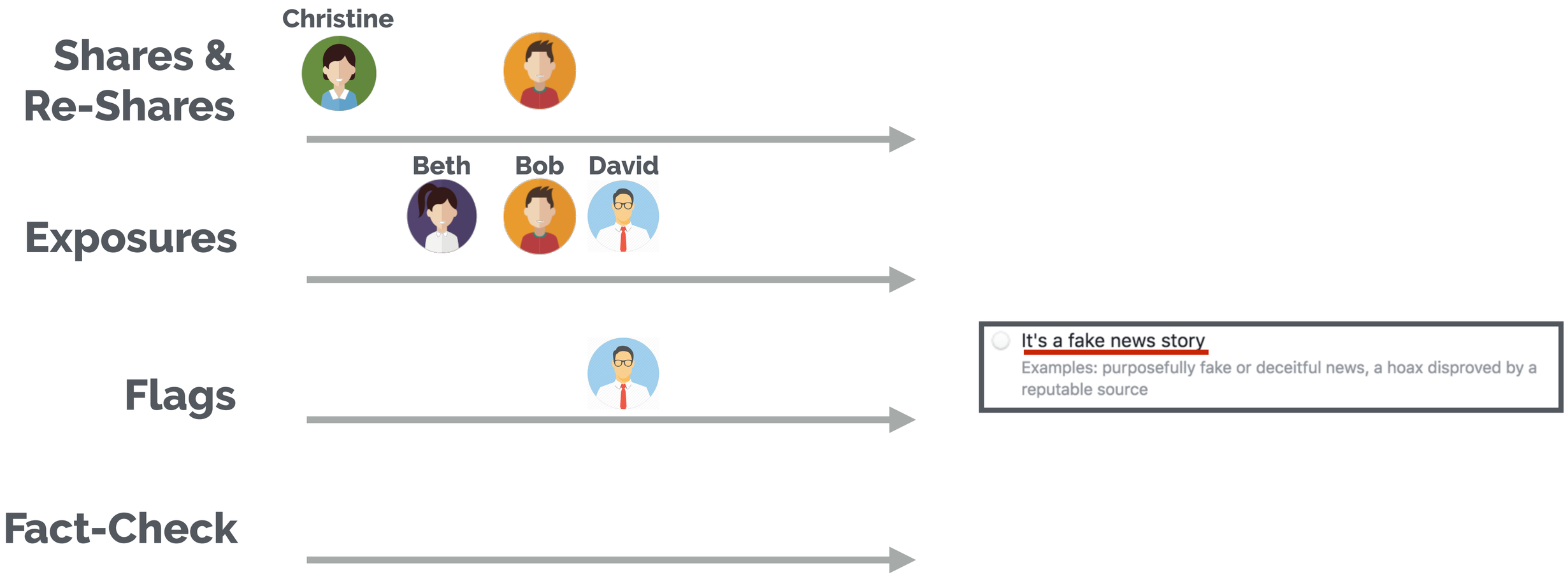
Crowd-powered Fact-checking



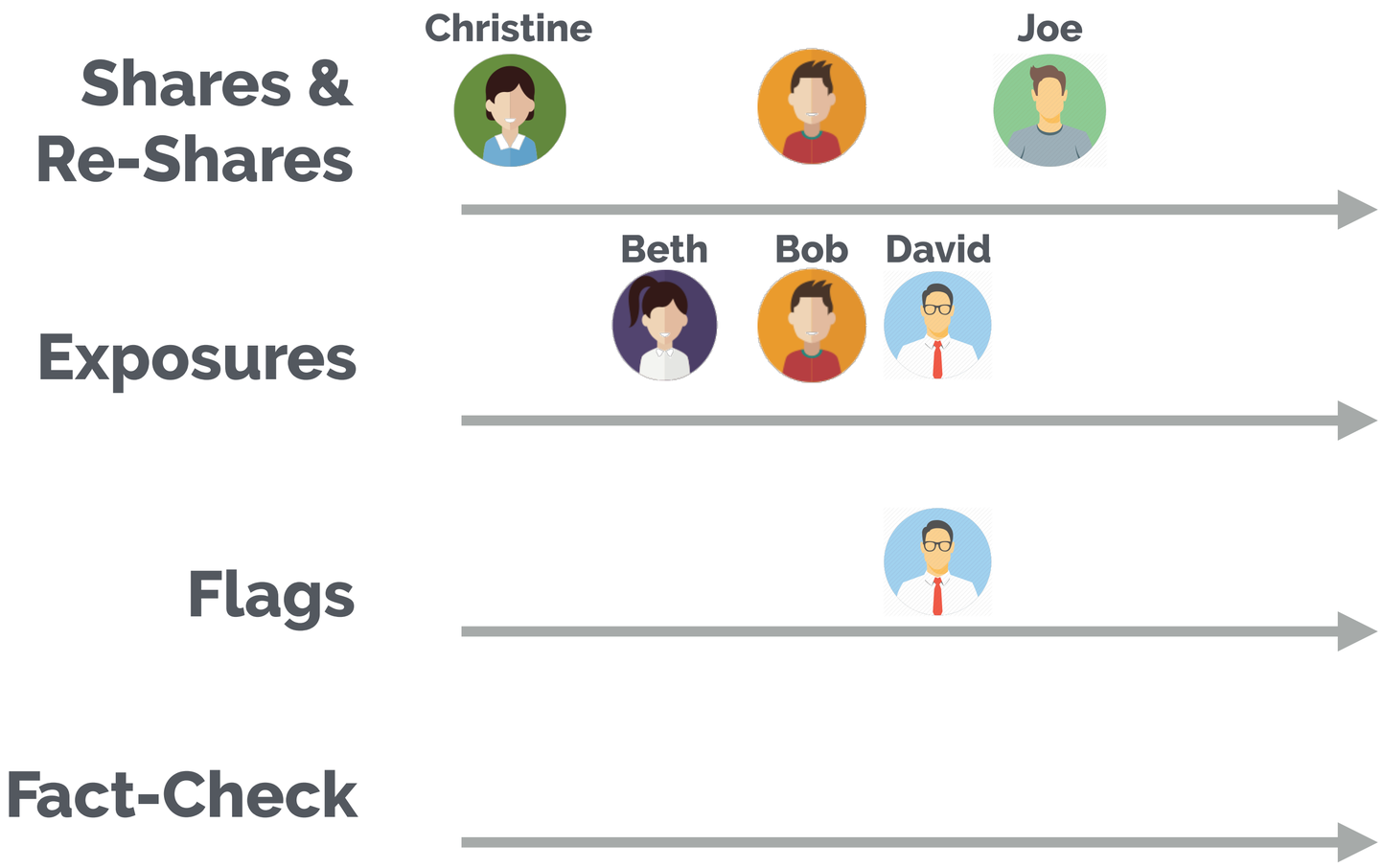
Crowd-powered Fact-checking

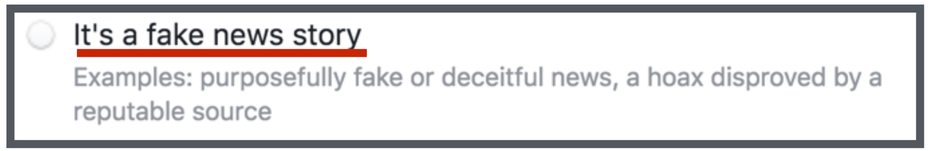


Crowd-powered Fact-checking

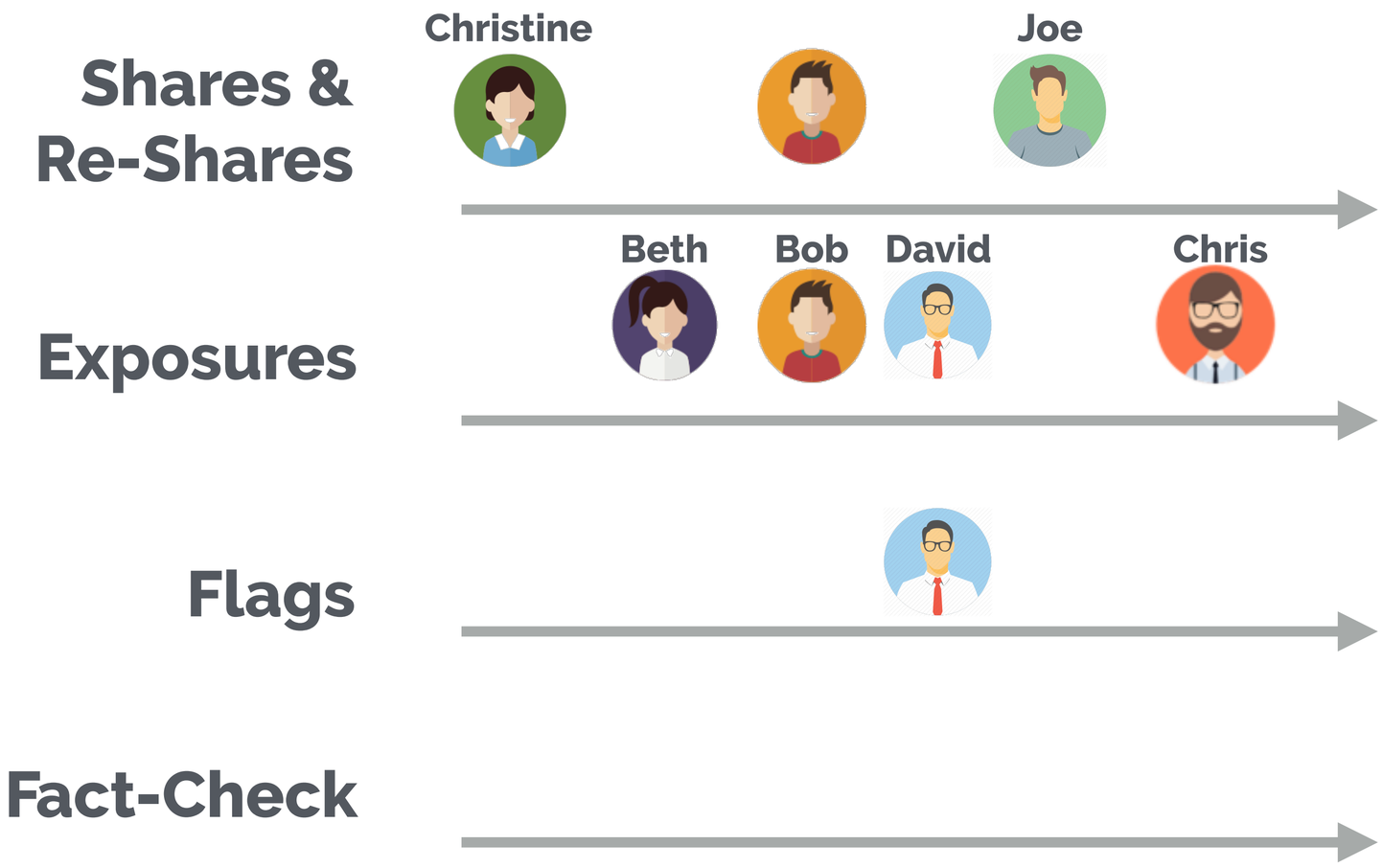


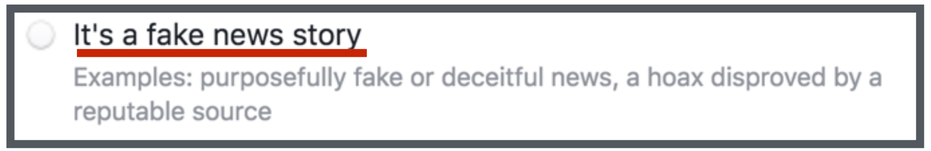
Crowd-powered Fact-checking



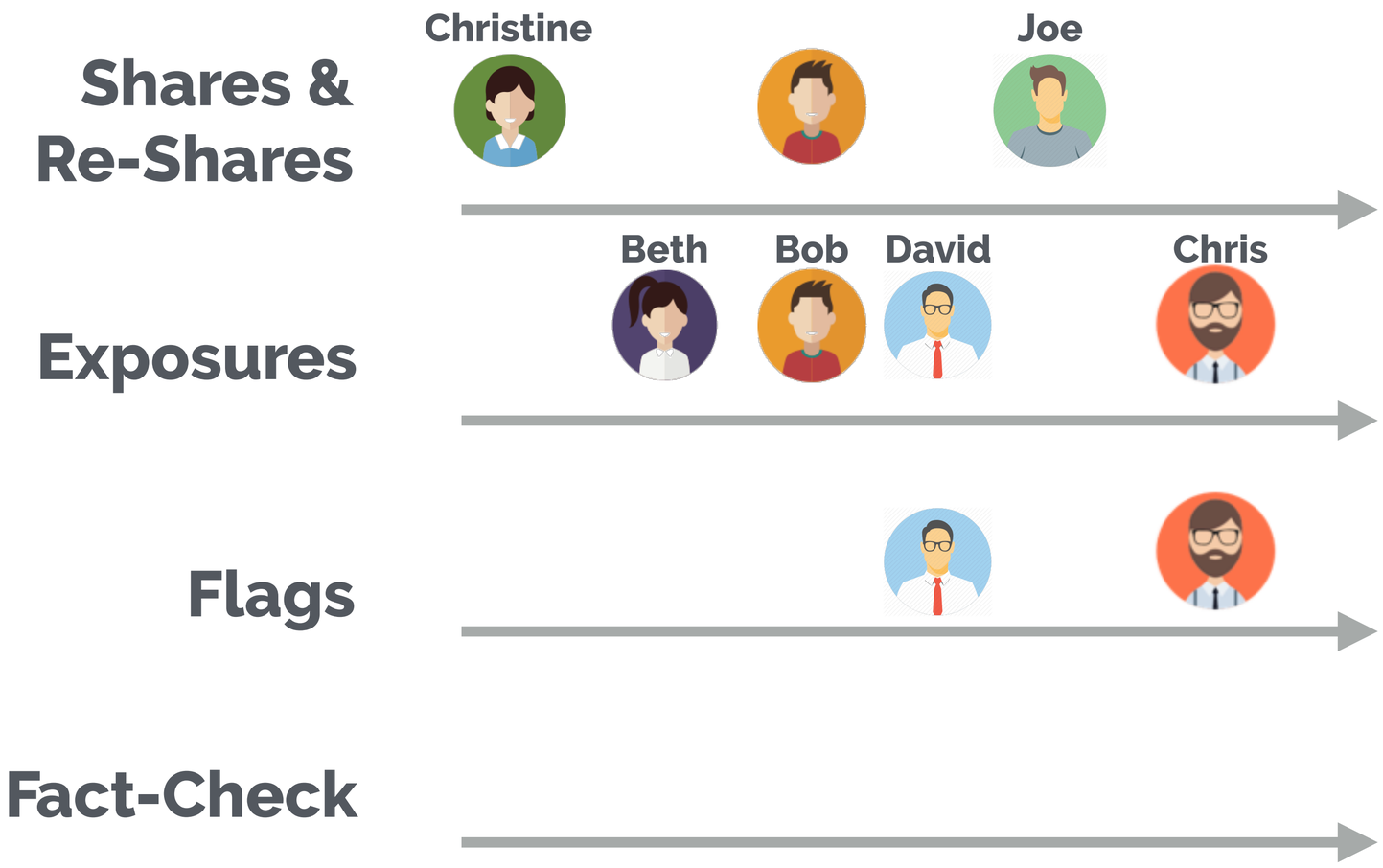


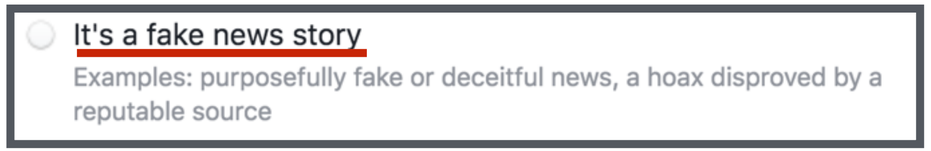
Crowd-powered Fact-checking



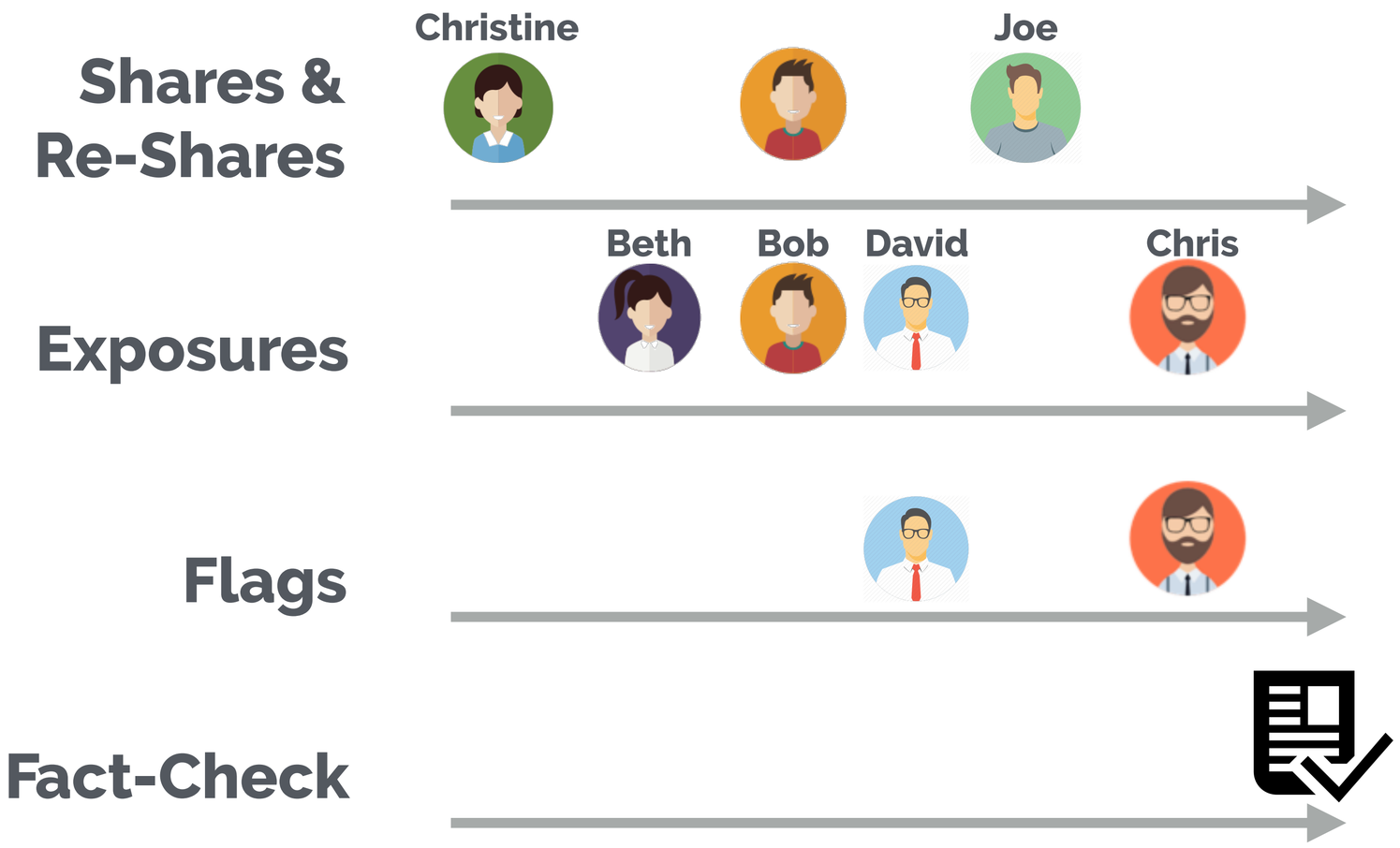


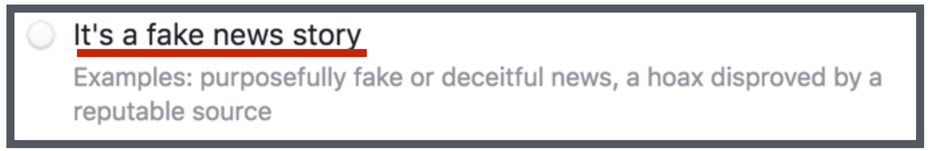
Crowd-powered Fact-checking





Crowd-powered Fact-checking





Snopes POLITIFACT

Challenges & Solutions

Number of Exposures is uncertain

Fact-checking is costly

Trade-off between flags and exposures

Challenges & Solutions

Number of Exposures is uncertain

Fact-checking is costly

Trade-off between flags and exposures

Probabilistic exposure models

Challenges & Solutions

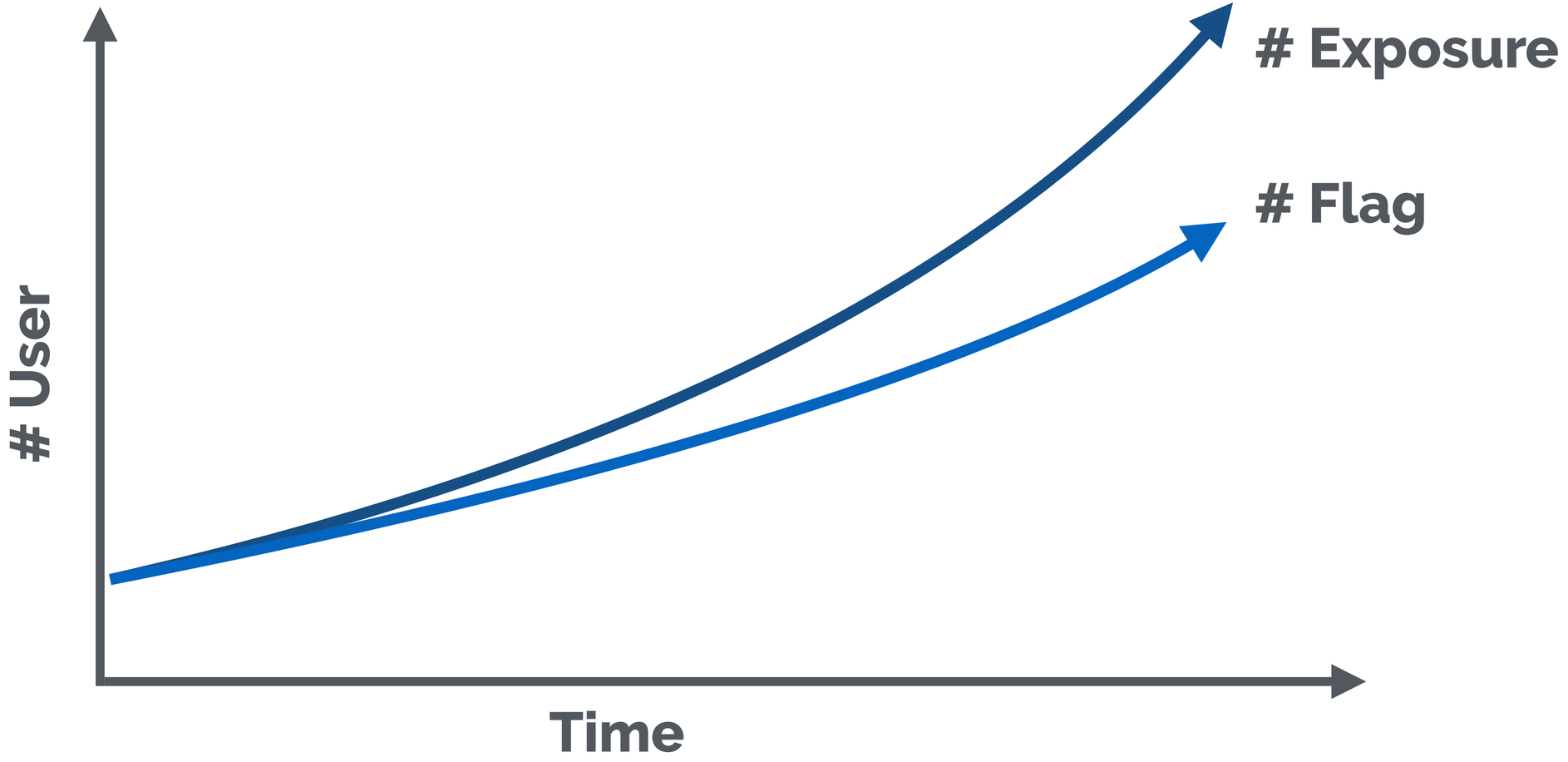
Number of Exposures is uncertain

Probabilistic exposure models

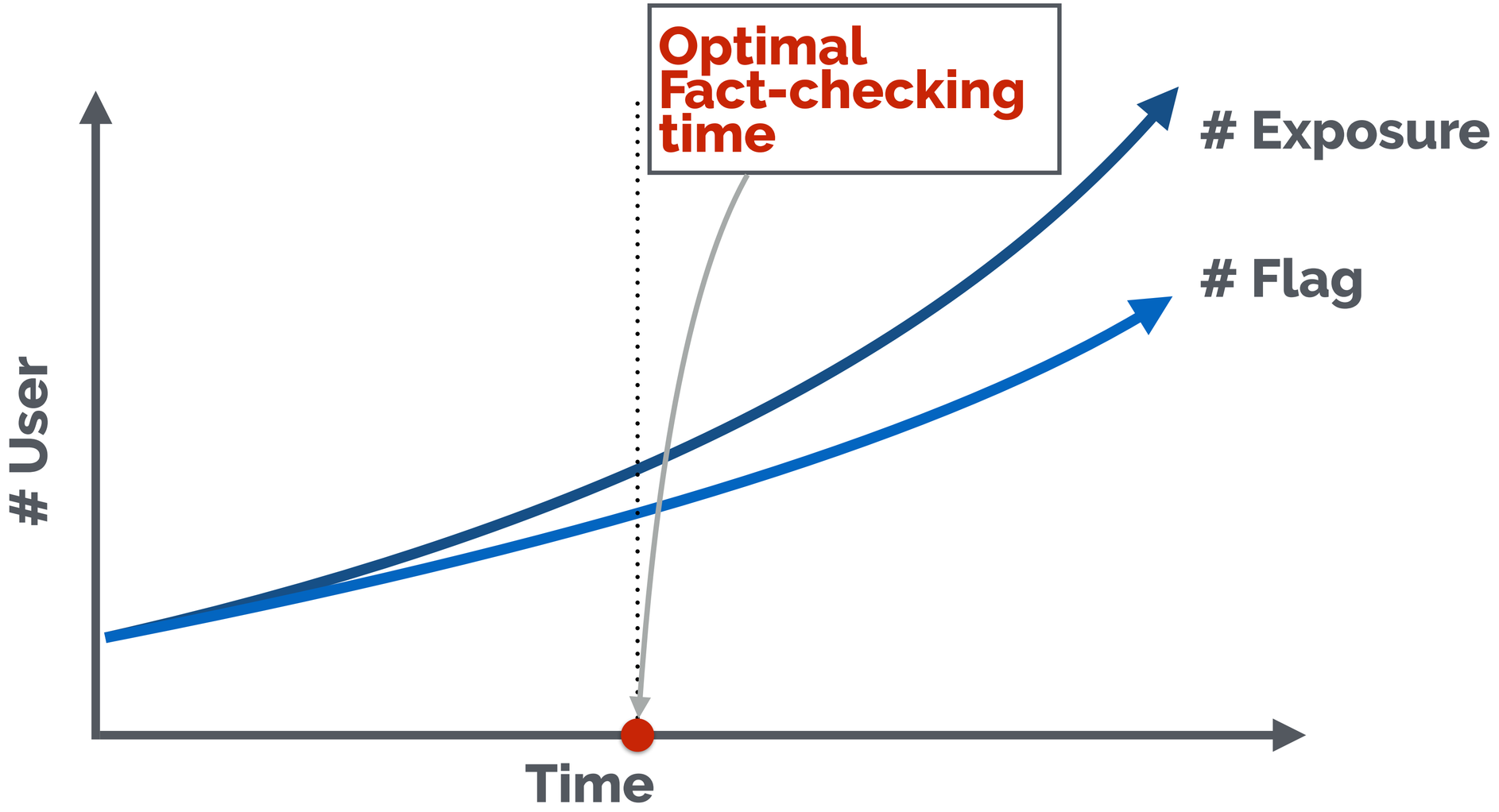
Fact-checking is costly

Trade-off between flags and exposures

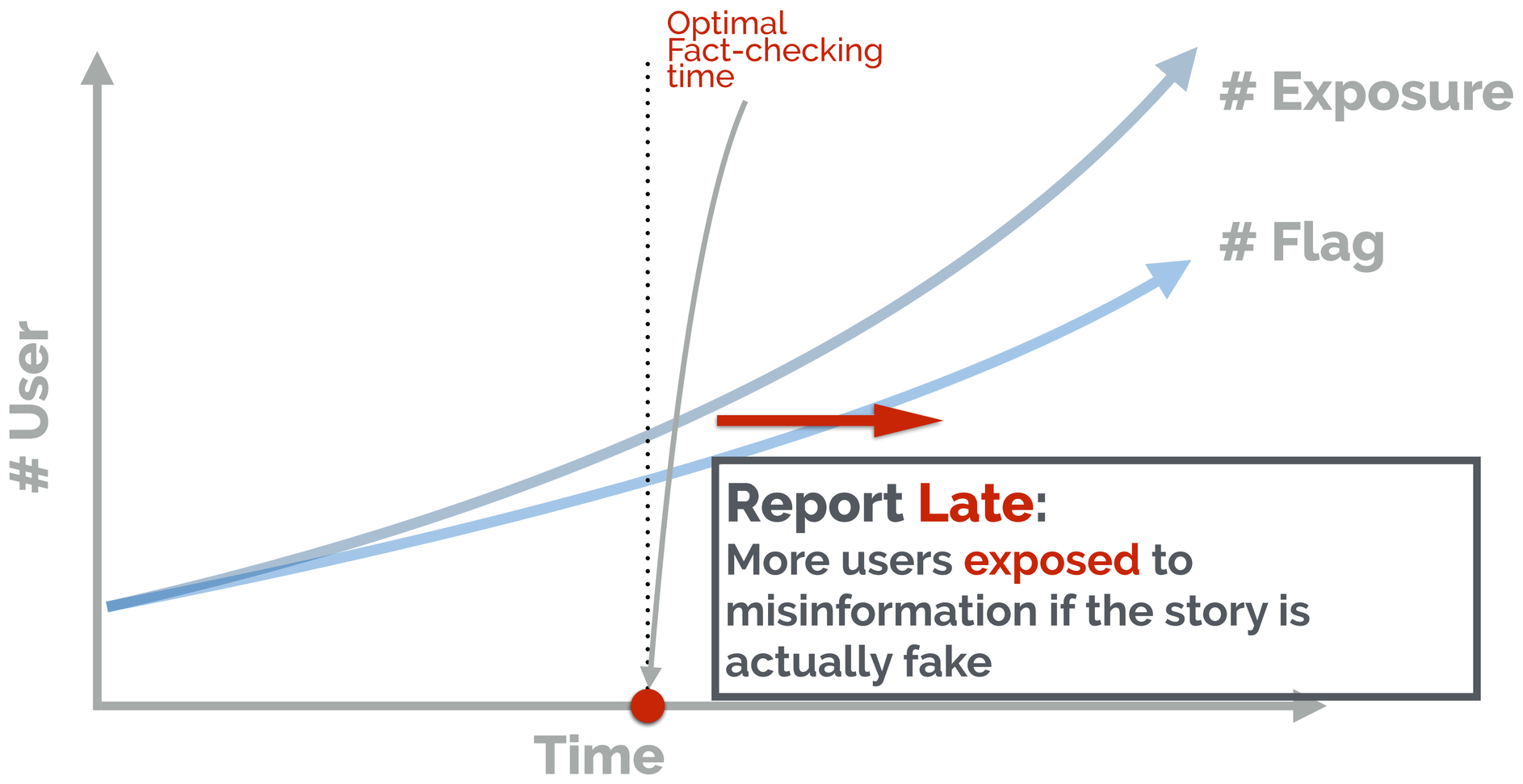
Exposure vs. Flag Trade-off



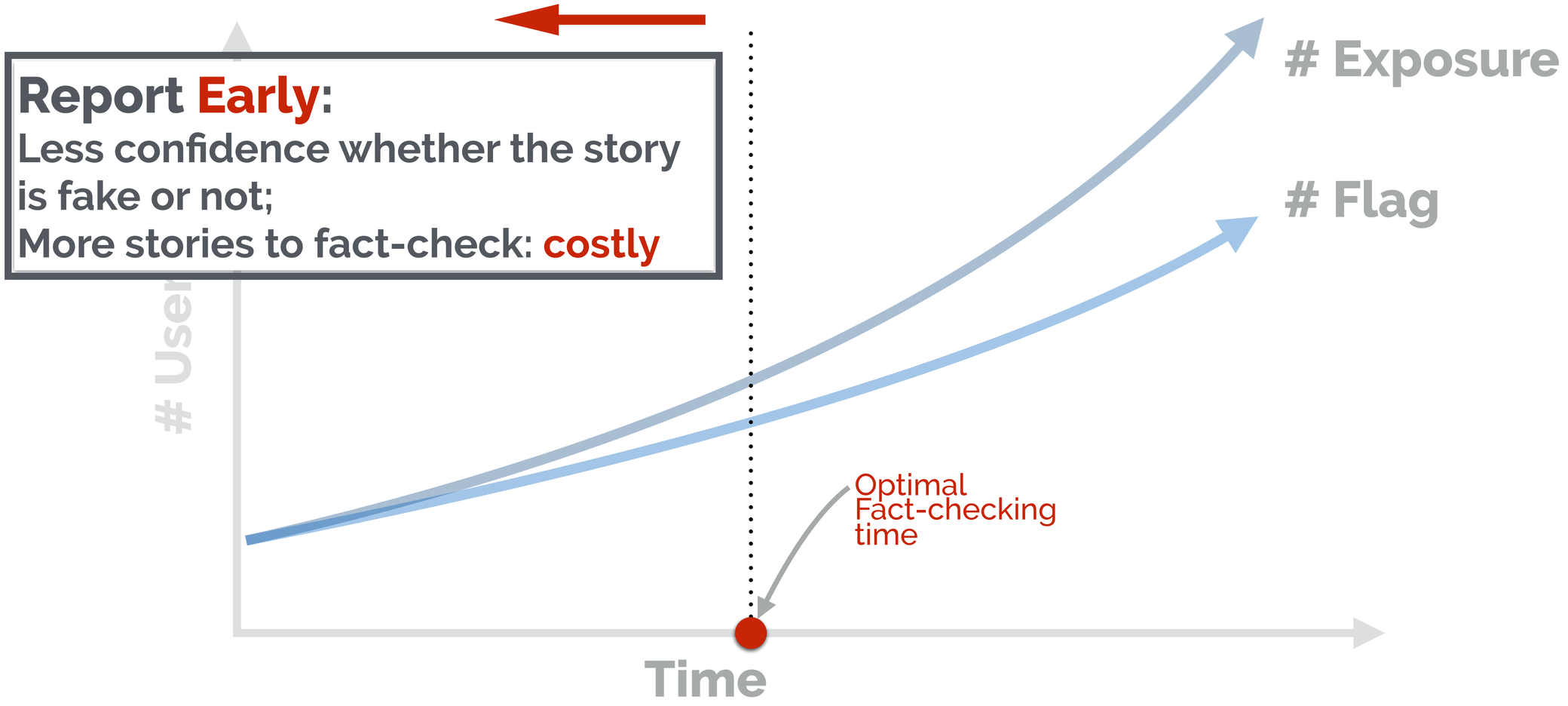
Exposure vs. Flag Trade-off



Exposure vs. Flag Trade-off



Exposure vs. Flag Trade-off



Challenges & Solutions

Number of Exposures is uncertain

Fact-checking is costly

Trade-off between flags and exposures

Probabilistic exposure models

Challenges & Solutions

Number of Exposures is uncertain

→ Probabilistic exposure models

Fact-checking is costly

Trade-off between flags and exposures

→ Optimal Fact-checking

Research Overview

Reduce the spread of misinformation

Research Overview

Reduce the spread of misinformation

↓ Quantify

Flags, exposures

↓ Happens when

Shares, (re) shares

Research Overview

Stochastic optimal control theory

Solve

Reduce the spread of misinformation

Quantify

Flags, exposures

Happens when

Shares, (re) shares

Research Overview

Stochastic optimal control theory → Optimal fact-checking time

Solve

Reduce the spread of misinformation

Quantify

Flags, exposures

Happens when

Shares, (re) shares

1. Post-Truth Era and Fake News

2. Crowd-Powered Fact-Checking

3. **Fact-Checking Scheduling Problem**

4. Algorithm: *CURB*

5. Detect and Redue the Spread of Fake News

img credit: http://www.musicteachingadventures.com/wp-content/uploads/2013/08/clock-time.jpg

Modeling

Modeling

Shares & Re-Shares

One Counting Process / Story

$$\mathbb{E}[d\boldsymbol{N}^p(t)|\mathcal{H}(t)] = \boldsymbol{\lambda}^p(t)dt$$

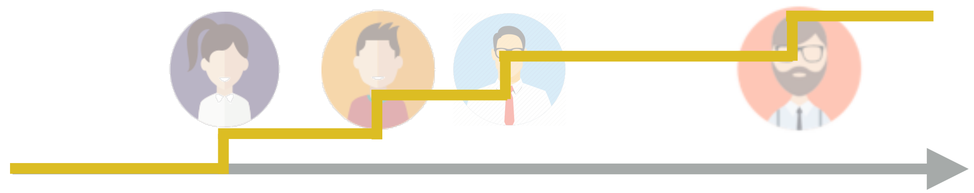
Modeling

Shares & Re-Shares

One Counting Process / Story

$$\mathbb{E}[d\boldsymbol{N}^p(t)|\mathcal{H}(t)] = \boldsymbol{\lambda}^p(t)dt$$

Exposures



One Counting Process / Story

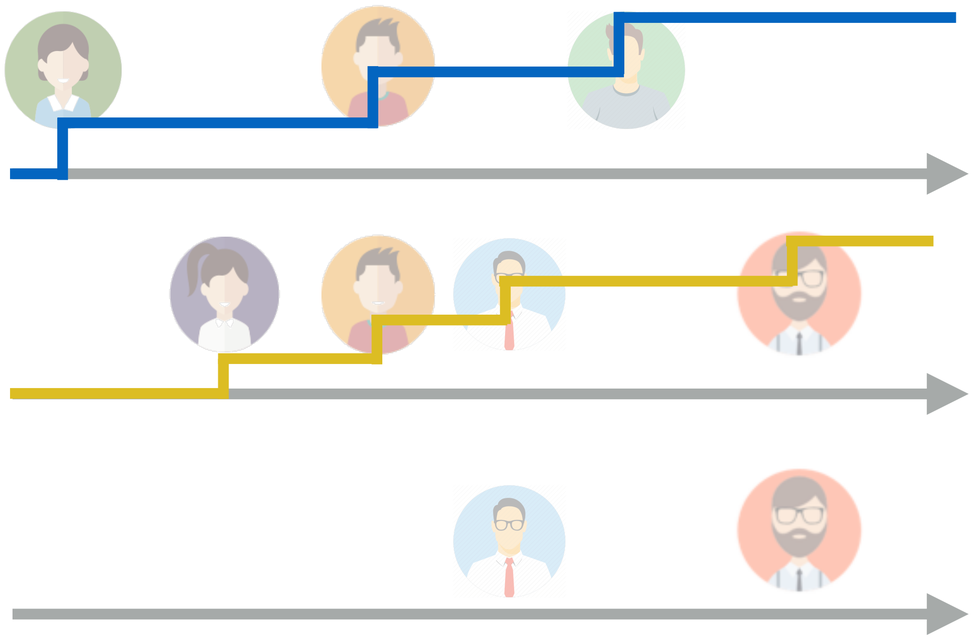
$$\mathbb{E}[d\boldsymbol{N}^e(t)|\mathcal{H}(t)] = \boldsymbol{\lambda}^e(t)dt$$

Modeling

Shares & Re-Shares

Exposures

Flags



One Counting Process / Story

$$\mathbb{E}[d\boldsymbol{N}^p(t)|\mathcal{H}(t)] = \boldsymbol{\lambda}^p(t)dt$$

One Counting Process / Story

$$\mathbb{E}[d\boldsymbol{N}^e(t)|\mathcal{H}(t)] = \boldsymbol{\lambda}^e(t)dt$$

$$f_s(\tau) \sim f_s$$ → **Estimate from Historical Data**

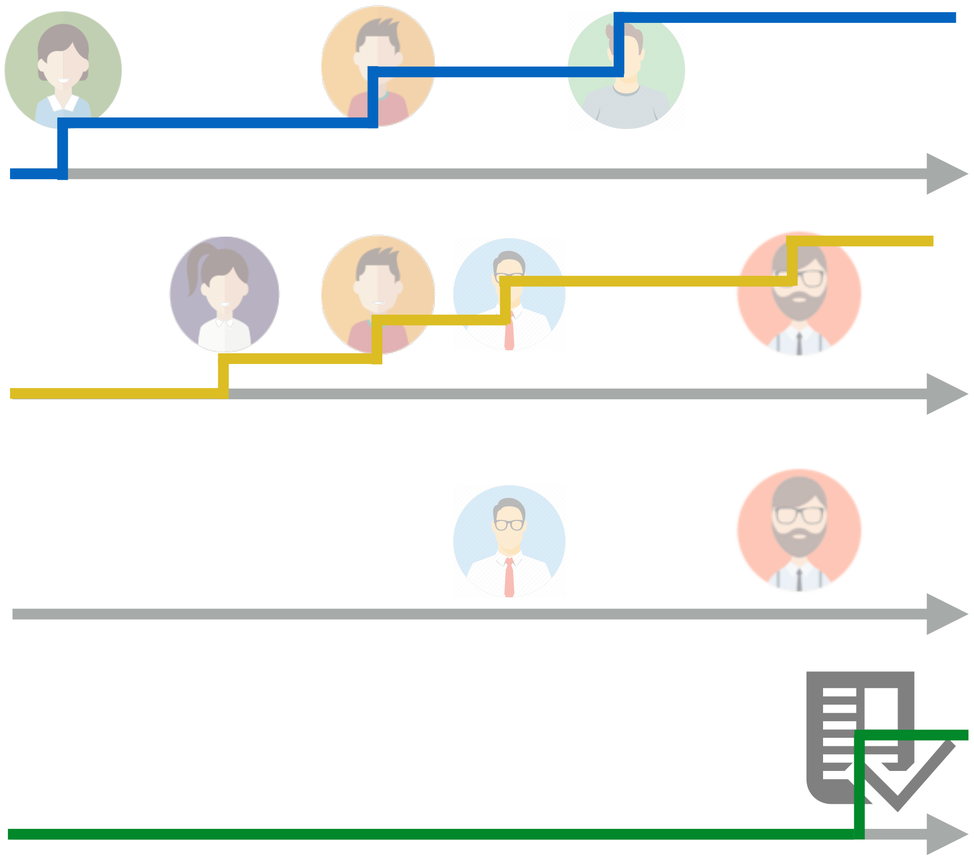
Modeling

Shares & Re-Shares

Exposures

Flags

Fact-Check



One Counting Process / Story

$$\mathbb{E}[d\boldsymbol{N}^p(t)|\mathcal{H}(t)] = \boldsymbol{\lambda}^p(t)dt$$

One Counting Process / Story

$$\mathbb{E}[d\boldsymbol{N}^e(t)|\mathcal{H}(t)] = \boldsymbol{\lambda}^e(t)dt$$

$$f_s(\tau) \sim f_s$$ → **Estimate from Historical Data**

One Survival Process / Story

$$\mathbb{E}[d\boldsymbol{M}(t)|\mathcal{H}(t)] = \boldsymbol{u}(t) \odot (1 - \boldsymbol{M}(t))dt$$

Estimated Number of Users Exposed to Misinformation

$$\bar{N}_s^m(t) := \boxed{p_{m|f=1}} N_s^f(t) + \boxed{p_{m|f=0}} (N_s^e(t) - N_s^f(t))$$

Probability of (a story) being misinformation ***not*** given a user flag

Probability of (a story) being misinformation given a user flag

Estimated Number of Users Exposed to Misinformation

$$\bar{N}_s^m(t) := p_{m|f=1} N_s^f(t) + p_{m|f=0}(N_s^e(t) - N_s^f(t))$$

$$\boxed{dN_s^f(t) = f_s(t) dN_s^e(t)}$$

$$d\bar{N}_s^m(t) = \Big[(p_{m|f=1} - p_{m|f=0}) f_s(t) + p_{m|f=0}\Big] dN_s^e(t)$$

Misinformation Rate

$$\bar{N}_s^m(t) := p_{m|f=1} N_s^f(t) + p_{m|f=0}(N_s^e(t) - N_s^f(t))$$

$$\boxed{dN_s^f(t) = f_s(t) dN_s^e(t)}$$

$$d\bar{N}_s^m(t) = \Big[(p_{m|f=1} - p_{m|f=0}) f_s(t) + p_{m|f=0}\Big] dN_s^e(t)$$

$$\boxed{f_s(\tau) \sim f_s}$$

$$\boxed{\mathbb{E}[d\boldsymbol{N}^e(t)|\mathcal{H}(t)] = \boldsymbol{\lambda}^e(t)dt}$$

$$\lambda_s^m(t)dt = \mathbb{E}[d\bar{N}_s^m(t)] = \Big[(p_{m|f=1} - p_{m|f=0}) f_s + p_{m|f=0}\Big] \lambda_s^e(t)dt$$

Misinformation Rate

$$\bar{N}_s^m(t) := p_{m|f=1} N_s^f(t) + p_{m|f=0}(N_s^e(t) - N_s^f(t))$$

$$\boxed{dN_s^f(t) = f_s(t) dN_s^e(t)}$$

$$d\bar{N}_s^m(t) = \Big[(p_{m|f=1} - p_{m|f=0}) f_s(t) + p_{m|f=0}\Big] dN_s^e(t)$$

$$\boxed{f_s(\tau) \sim f_s}$$

$$\boxed{\mathbb{E}[d\boldsymbol{N}^e(t)|\mathcal{H}(t)] = \boldsymbol{\lambda}^e(t)dt}$$

$$\lambda_s^m(t)dt = \mathbb{E}[d\bar{N}_s^m(t)] = \Big[(p_{m|f=1} - p_{m|f=0}) f_s + p_{m|f=0}\Big] \lambda_s^e(t)dt$$

Misinformation Rate

$$\lambda_s^m(t)dt = \mathbb{E}[d\bar{N}_s^m(t)] = \left[(p_{m|f=1} - p_{m|f=0})\boxed{f_s} + p_{m|f=0}\right]\lambda_s^e(t)dt$$

Historical data (# flags, #exposures)

+ Posterior estimation

Assume Beta **Prior**: $f_s \sim Beta(\alpha, \beta)$

Posterior: $f_s | N_s^f, N_s^e \sim Beta\left(\alpha + N_s^f(t), \beta + N_s^e(t) - N_s^f(t)\right)$

Misinformation Rate

$$\lambda_s^m(t)dt = \mathbb{E}[d\bar{N}_s^m(t)] = \left[(p_{m|f=1} - p_{m|f=0})\boxed{f_s} + p_{m|f=0}\right]\lambda_s^e(t)dt$$

Historical data (# flags, #exposures)

+ Posterior estimation

Assume Beta **Prior**: $f_s \sim Beta(\alpha, \beta)$

Posterior: $f_s | N_s^f, N_s^e \sim Beta\left(\alpha + N_s^f(t), \beta + N_s^e(t) - N_s^f(t)\right)$

$$\hat{\lambda}_s^m(t)dt = \left[(p_{m|f=1} - p_{m|f=0})\boxed{\frac{\alpha + N_s^f(t)}{\alpha + \beta + N_s^e(t)}} + p_{m|f=0}\right]\lambda_s^e(t)dt$$

The Fact Checking Scheduling Problem

$$\underset{u(t_0,t_f]}{\text{minimize}} \quad \mathbb{E}\left[\int_{t_0}^{t_f} l(\hat{\lambda}^m(\tau), u(\tau))d\tau\right]$$

$$\text{subject to} \quad u(t) \geq 0 \quad \forall t \in (t_0, t_f]$$

The Fact Checking Scheduling Problem

$$\underset{u(t_0,t_f]}{\text{minimize}} \quad \mathbb{E}\left[\int_{t_0}^{t_f} l(\hat{\lambda}^m(\tau), \boxed{u(\tau)})d\tau\right]$$

$$\text{subject to} \quad u(t) \geq 0 \quad \forall t \in (t_0, t_f]$$

Fact Checking Rate:
we need to optimize this!

The Fact Checking Scheduling Problem

$$\underset{u(t_0,t_f]}{\text{minimize}} \quad \mathbb{E}\left[\int_{t_0}^{t_f} \boxed{l(\hat{\lambda}^m(\tau), u(\tau))} d\tau\right]$$

$$\text{subject to} \quad u(t) \geq 0 \quad \forall t \in (t_0, t_f]$$

The Fact Checking Scheduling Problem

$$\underset{u(t_0,t_f]}{\text{minimize}} \quad \mathbb{E}\left[\int_{t_0}^{t_f} \boxed{l(\hat{\lambda}^m(\tau), u(\tau))} d\tau\right]$$

$$\text{subject to} \quad u(t) \geq 0 \quad \forall t \in (t_0, t_f]$$

Nondecreasing quadratic convex loss function

$$\ell(\hat{\lambda}^m(t), u(t)) = \frac{1}{2}(\hat{\lambda}^m(t))^2 + \frac{1}{2}qu^2(t)$$

The Fact Checking Scheduling Problem

$$\underset{u(t_0,t_f]}{\text{minimize}} \quad \mathbb{E}\left[\int_{t_0}^{t_f} \boxed{l(\hat{\lambda}^m(\tau), u(\tau))} d\tau\right]$$

$$\text{subject to} \quad u(t) \geq 0 \quad \forall t \in (t_0, t_f]$$

Nondecreasing quadratic convex loss function

$$\ell(\hat{\lambda}^m(t), u(t)) = \frac{1}{2}(\underbrace{\hat{\lambda}^m(t)}_{\text{Misinformation Rate}})^2 + \frac{1}{2}q\underbrace{u^2(t)}_{\text{Fact-checking Rate}}$$

The Fact Checking Scheduling Problem

$$\underset{u(t_0,t_f]}{\text{minimize}} \quad \mathbb{E}\left[\int_{t_0}^{t_f} \boxed{l(\hat{\lambda}^m(\tau), u(\tau))} d\tau\right]$$

$$\text{subject to} \quad u(t) \geq 0 \quad \forall t \in (t_0, t_f]$$

Nondecreasing quadratic convex loss function

$$\ell(\hat{\lambda}^m(t), u(t)) = \frac{1}{2}(\underline{\hat{\lambda}^m(t)})^2 + \frac{1}{2}q\underline{u^2(t)}$$

Tunable parameter for the ***exposure-flag trade-off***

Misinformation Rate

Fact-checking Rate

The Fact Checking Scheduling Problem

$$\underset{u(t_0,t_f]}{\text{minimize}} \quad \mathbb{E}\left[\phi(\hat{\lambda}^m(t_f)) + \int_{t_0}^{t_f} \ell(\hat{\lambda}^m(t), u(t))\, dt\right]$$

$$\text{subject to} \quad u(t) \geq 0 \quad \forall t \in (t_0, t_f]$$

Represent tuples of the counting processes and their corresponding rates as stochastic differential equations (SDEs) with jumps

Nondecreasing quadratic convex loss function

$$\ell(\hat{\lambda}^m(t), u(t)) = \frac{1}{2}(\hat{\lambda}^m(t))^2 + \frac{1}{2}qu^2(t)$$

Tunable parameter for the *exposure-flag trade-off*

Misinformation Intensity

Fact-checking Intensity: We need to optimize this!

The Fact Checking Scheduling Problem

Represent tuples of the counting processes and their corresponding rates as stochastic differential equations (SDEs) with jumps

Solve the stochastic optimal control problem for jump SDEs

1. Post-Truth Era and Fake News

2. Crowd-Powered Fact-Checking

3. Fact-Checking Scheduling Problem

4. Algorithm: ***CURB***

5. Detect and Reduce the Spread of Fake News

img credit: https://cdn-images-1.medium.com/max/1280/1*Jx3X4pD0KdiP8QDhi3d4Og.png

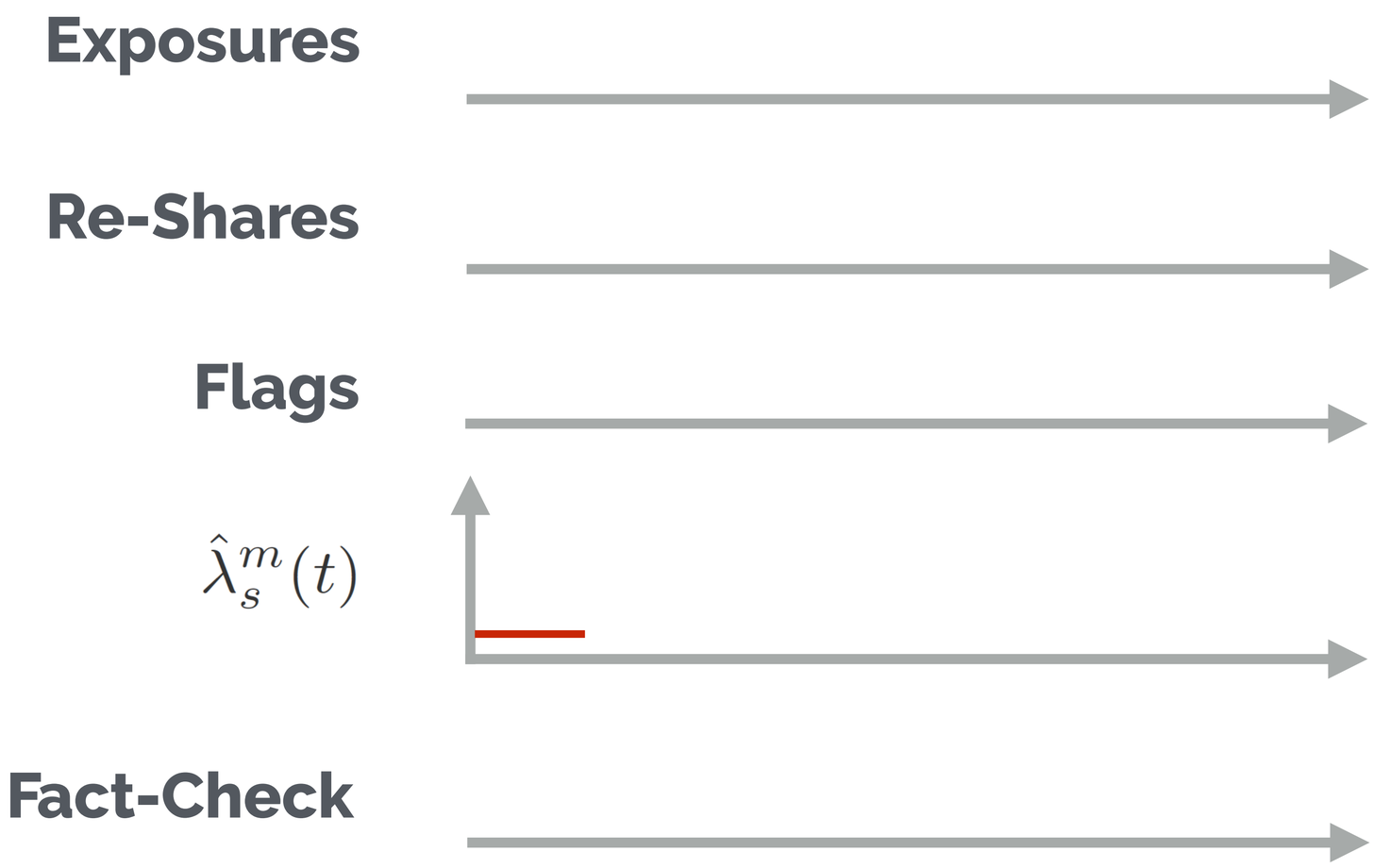
The Optimal Fact-checking Rate

$$u_s^*(t) = q_s^{-\frac{1}{2}} \hat{\lambda}_s^m(t)$$

It only depends on the current rate of misinformation!

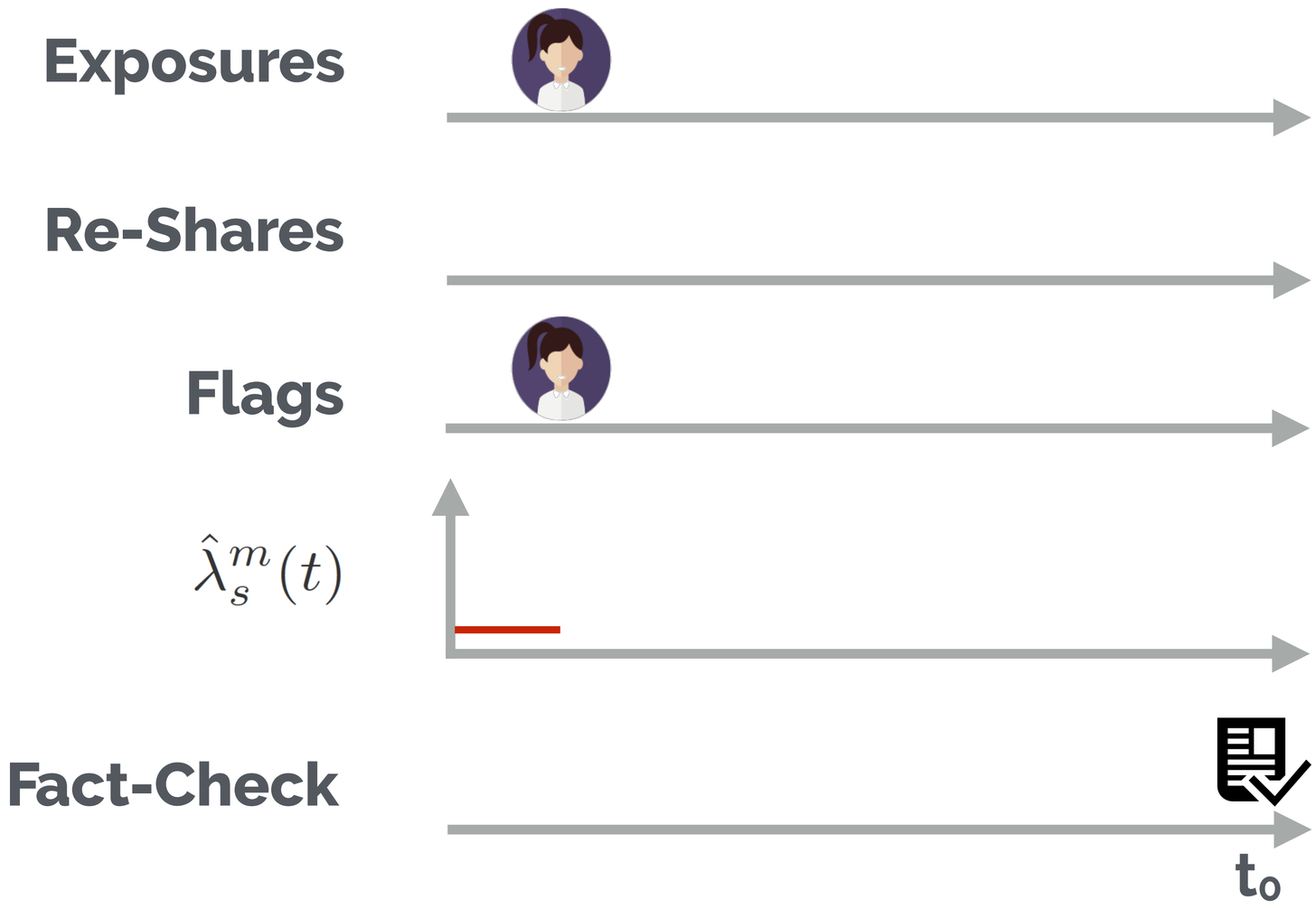
The *CURB* Algorithm

The *CURB* Algorithm

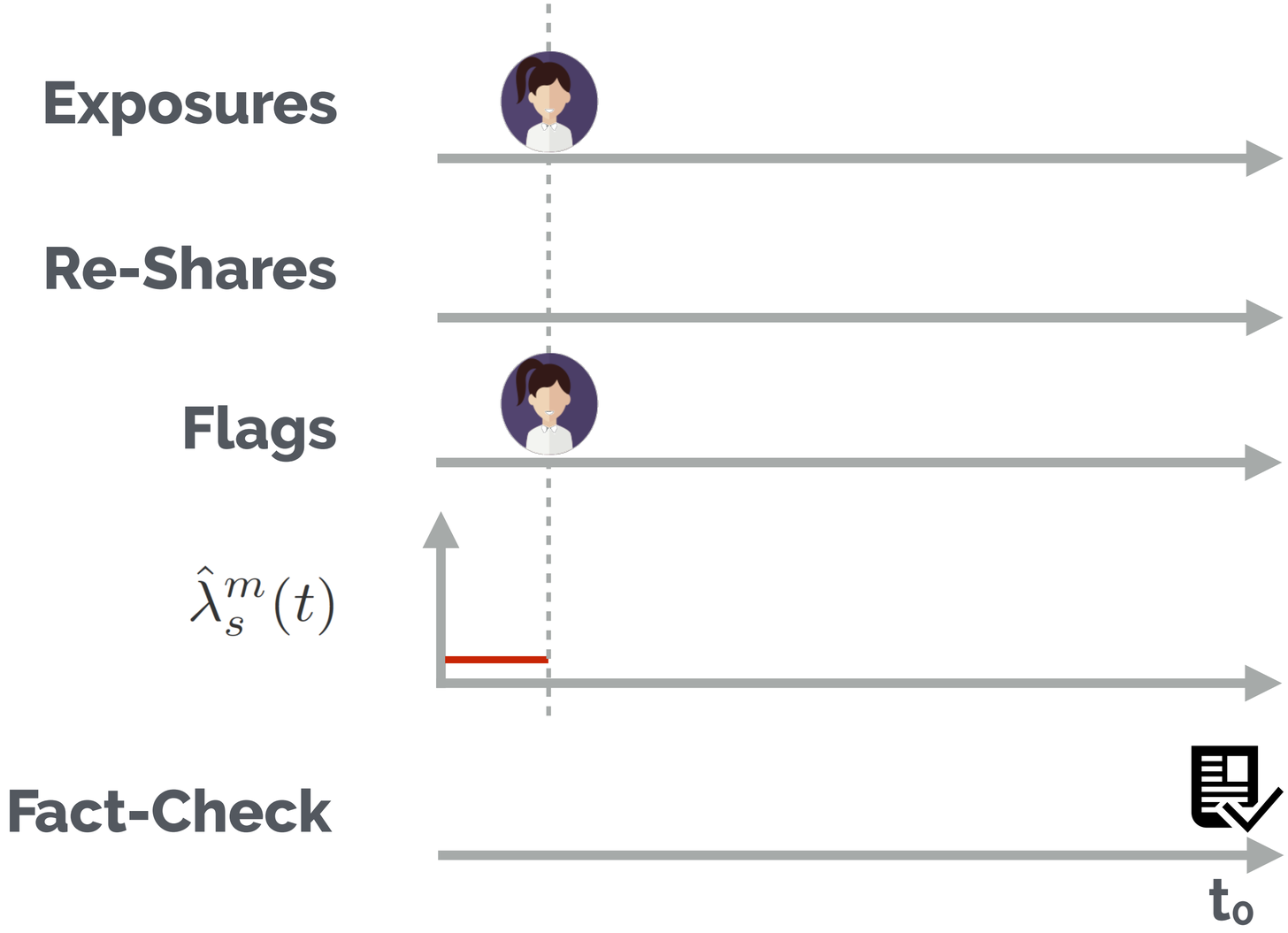


The *CURB* Algorithm

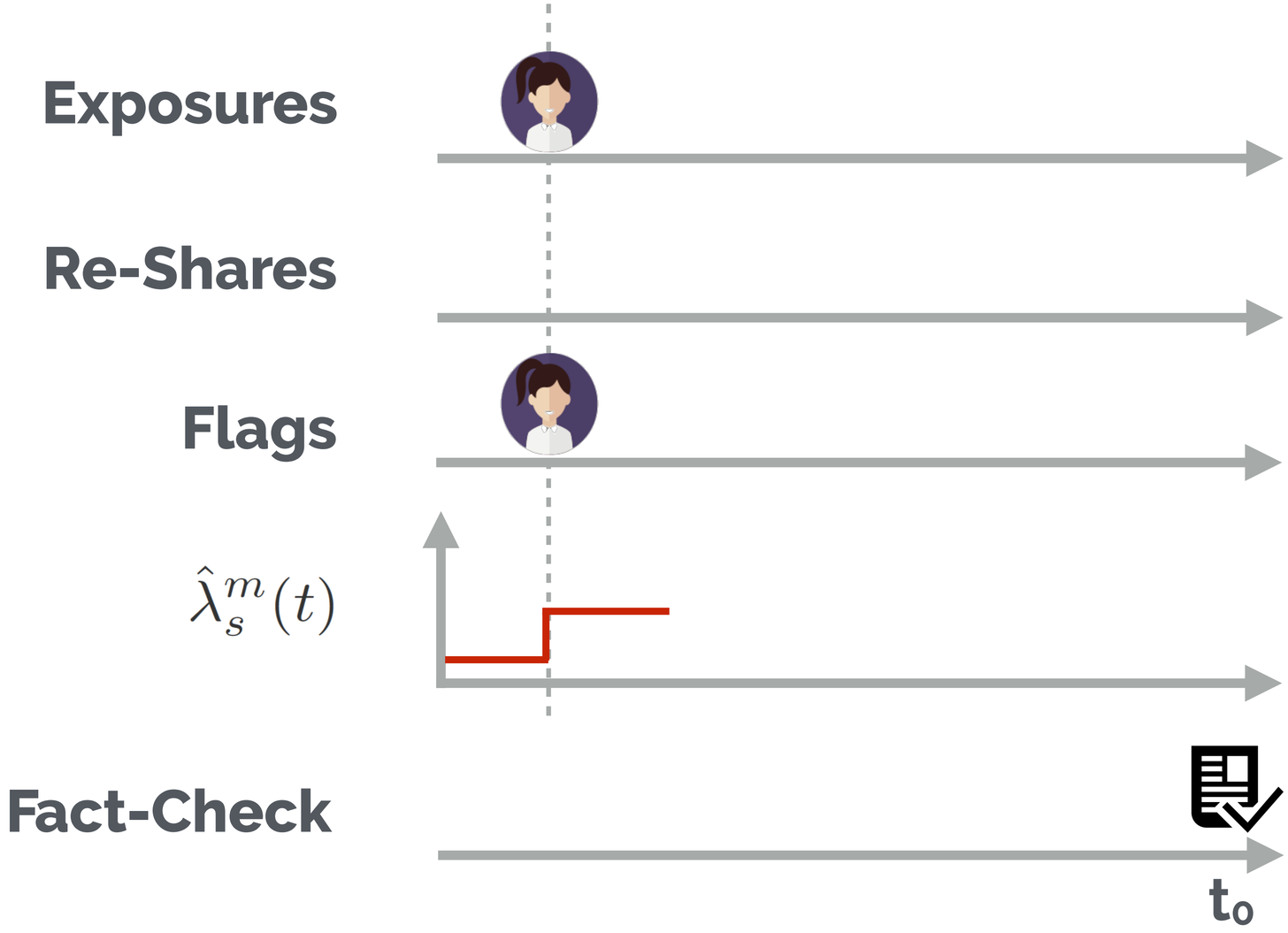
The *CURB* Algorithm



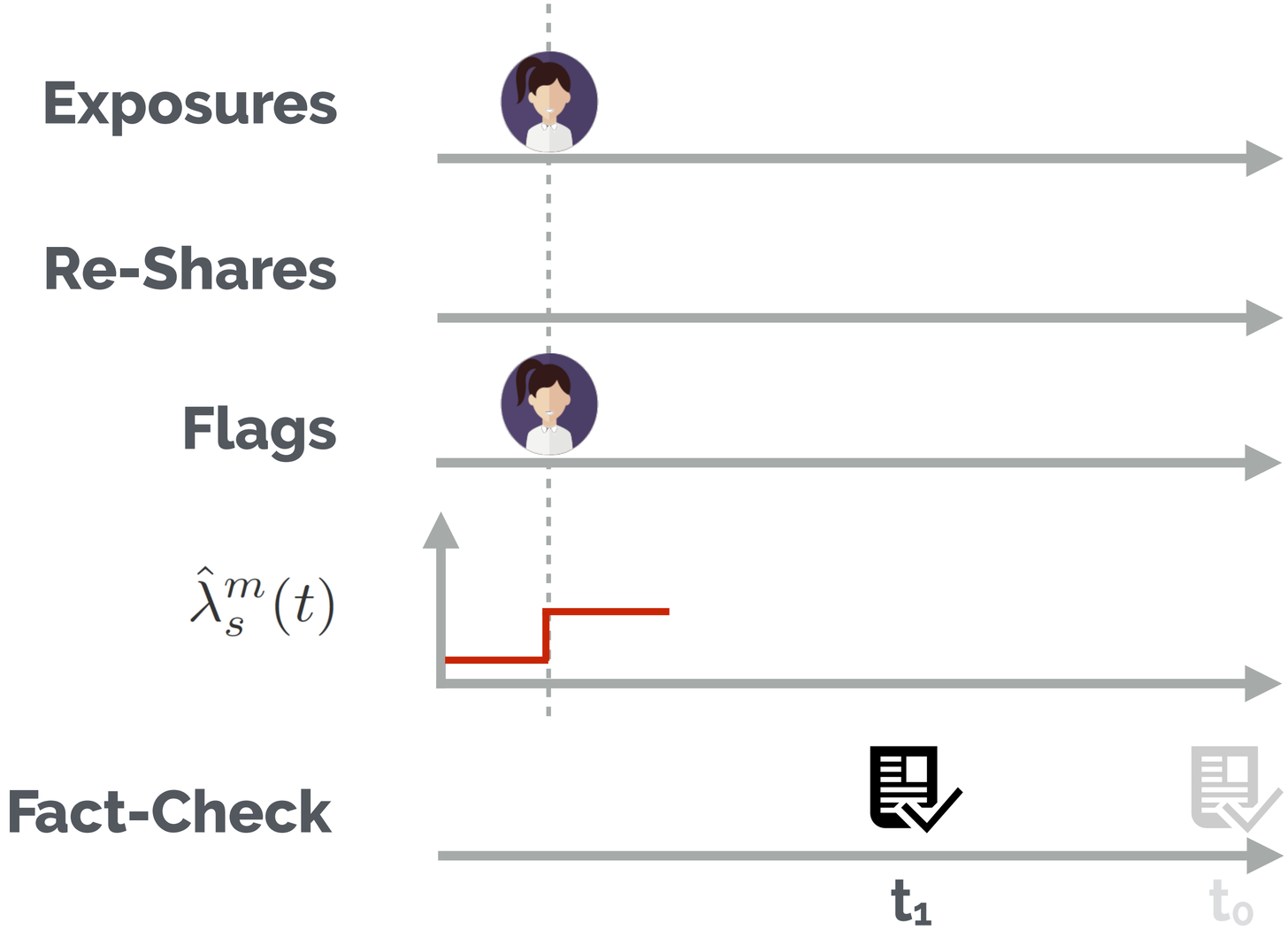
The *CURB* Algorithm



The *CURB* Algorithm

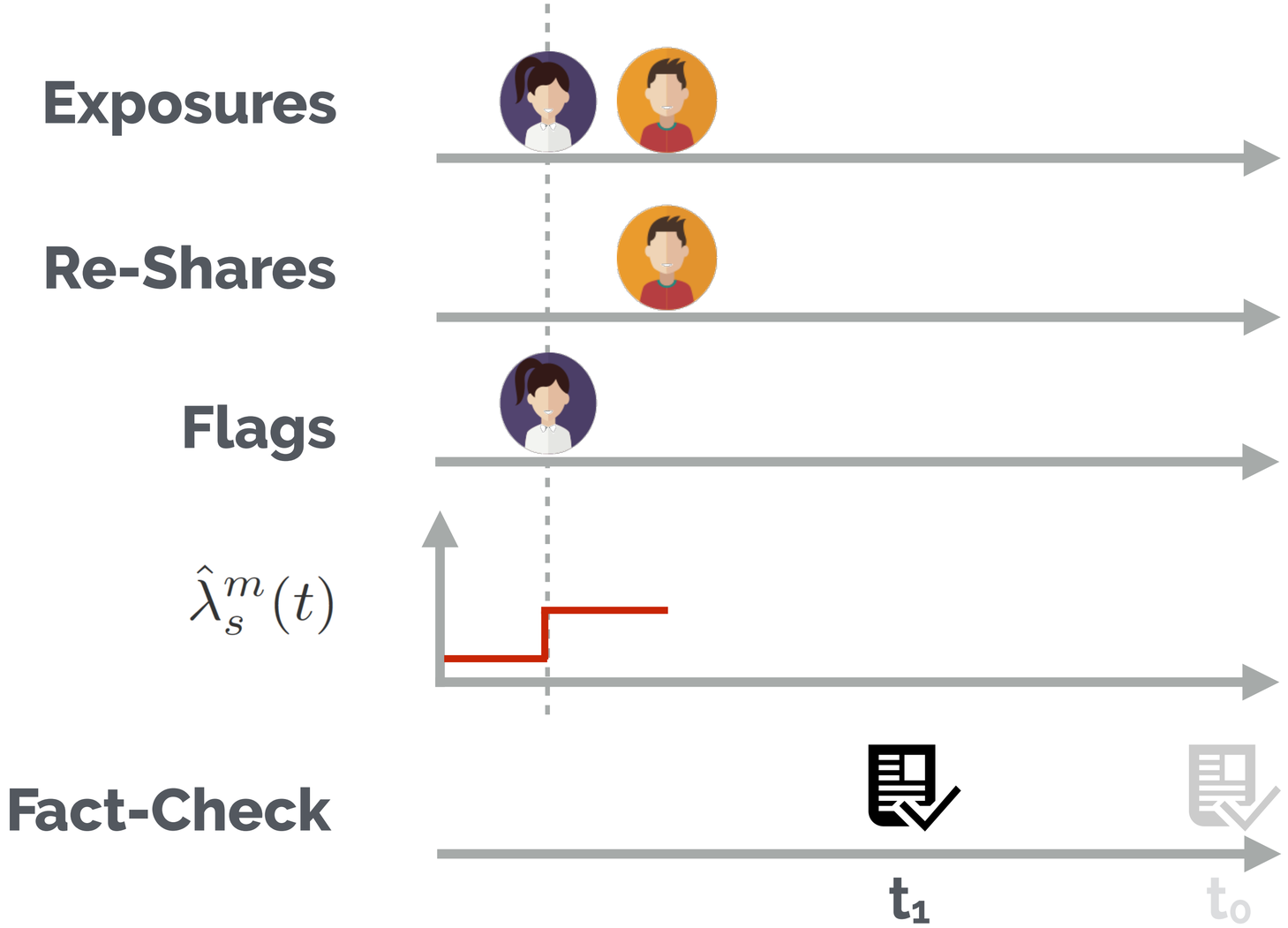


The *CURB* Algorithm



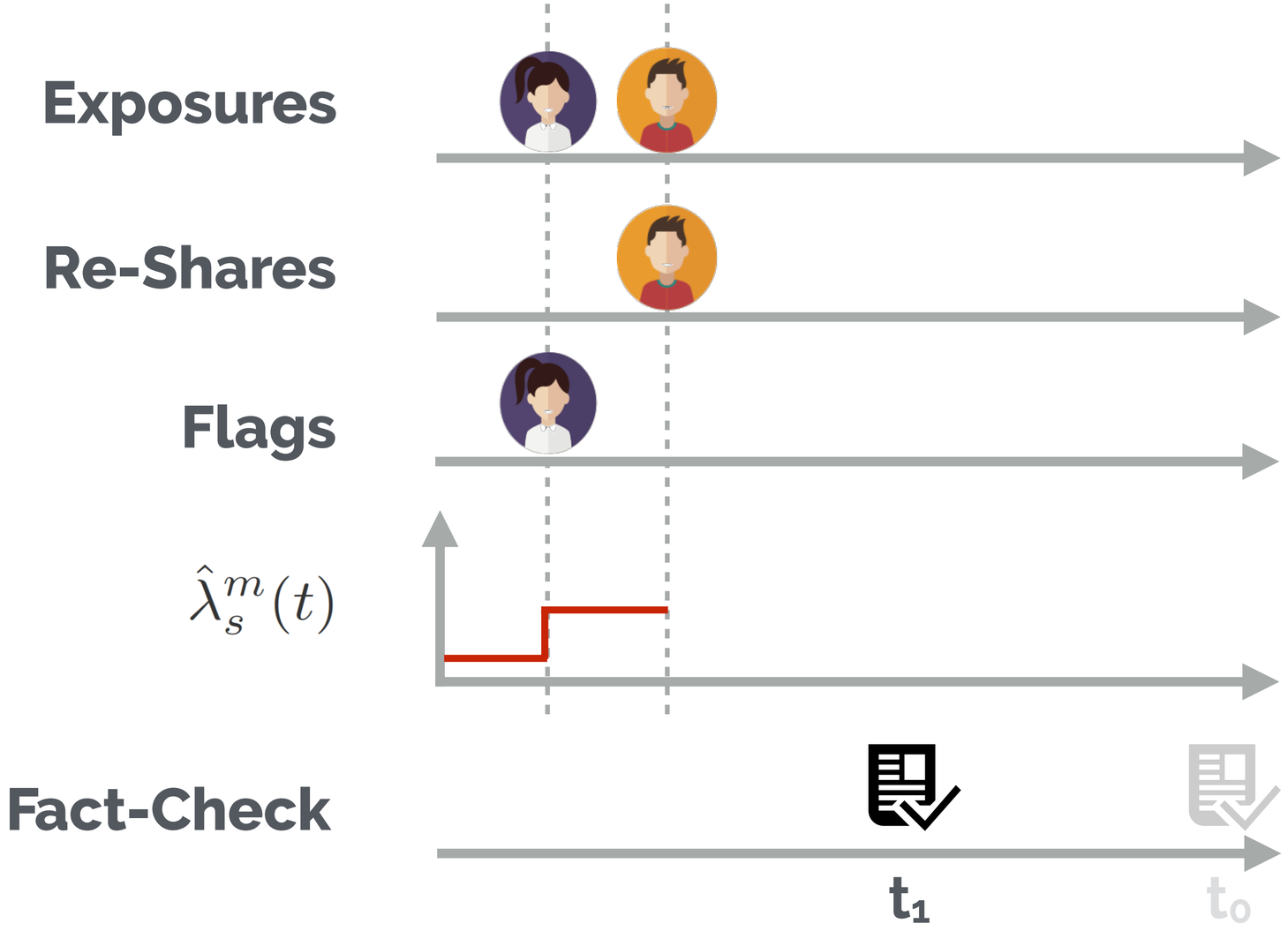
$t_1 \sim f(t_0, \hat{\lambda}_s^m(t))$

The *CURB* Algorithm



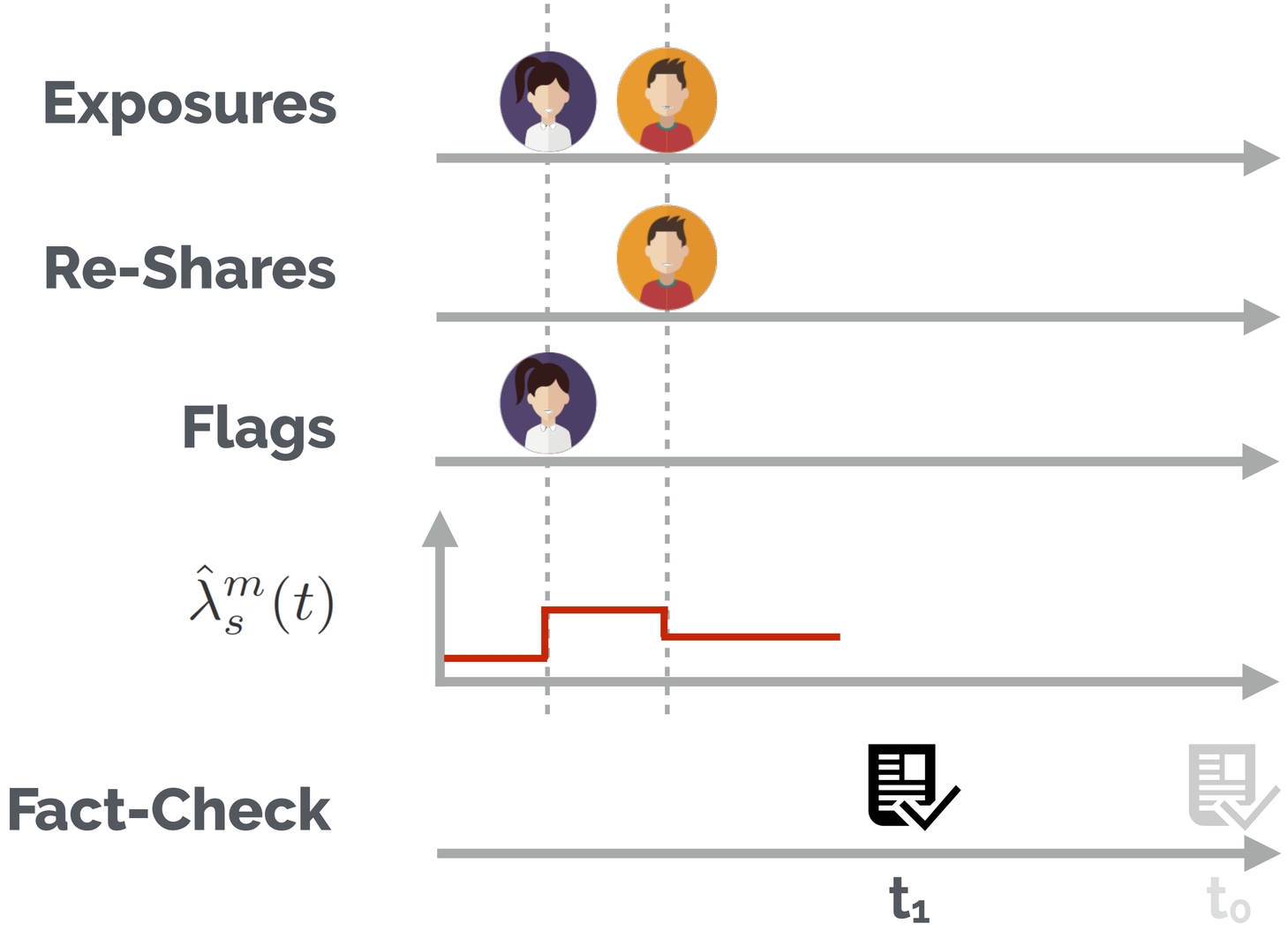
$t_1 \sim f(t_0, \hat{\lambda}_s^m(t))$

The *CURB* Algorithm



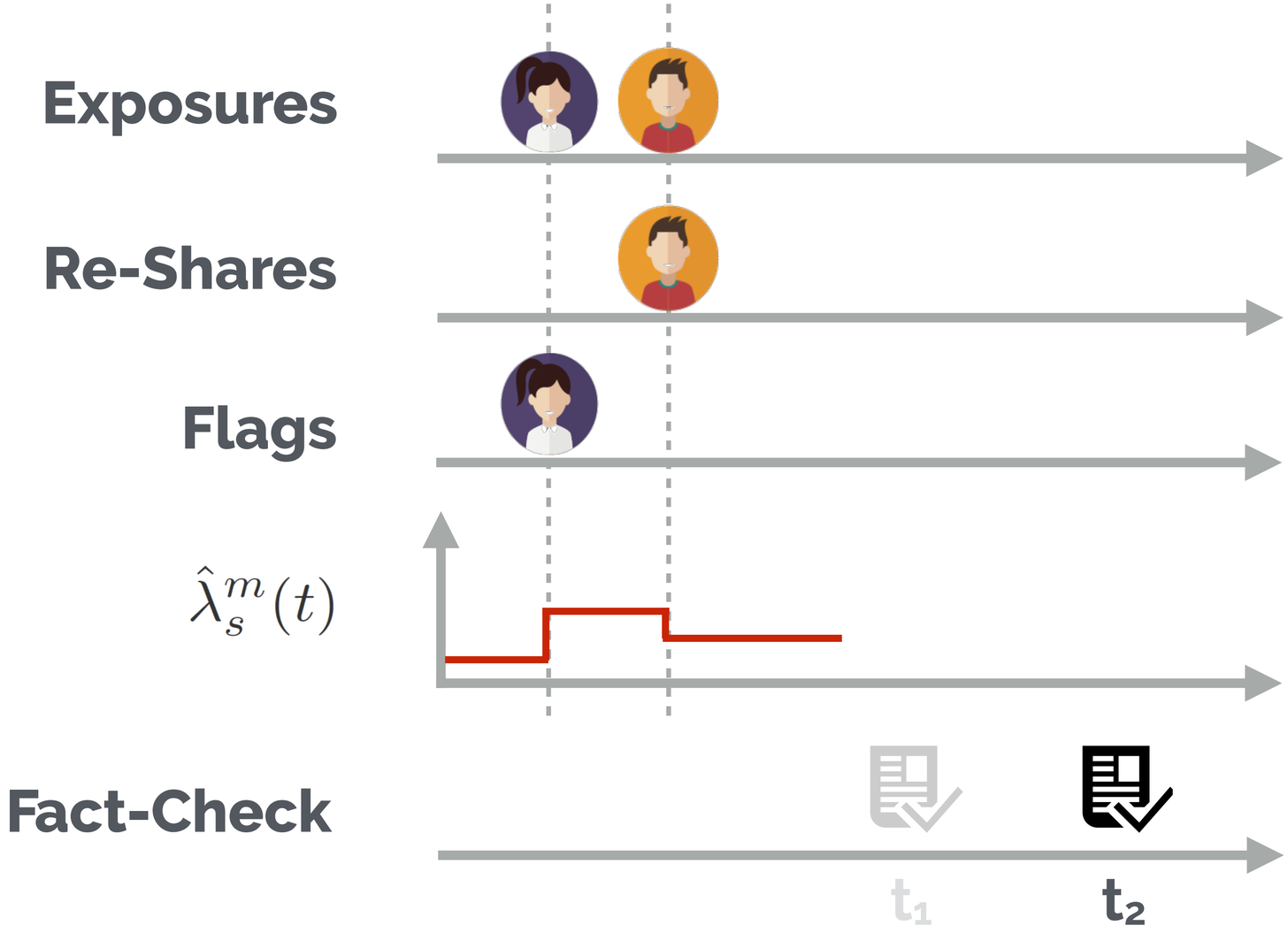
$t_1 \sim f(t_0, \hat{\lambda}_s^m(t))$

The *CURB* Algorithm



$t_1 \sim f(t_0, \hat{\lambda}_s^m(t))$

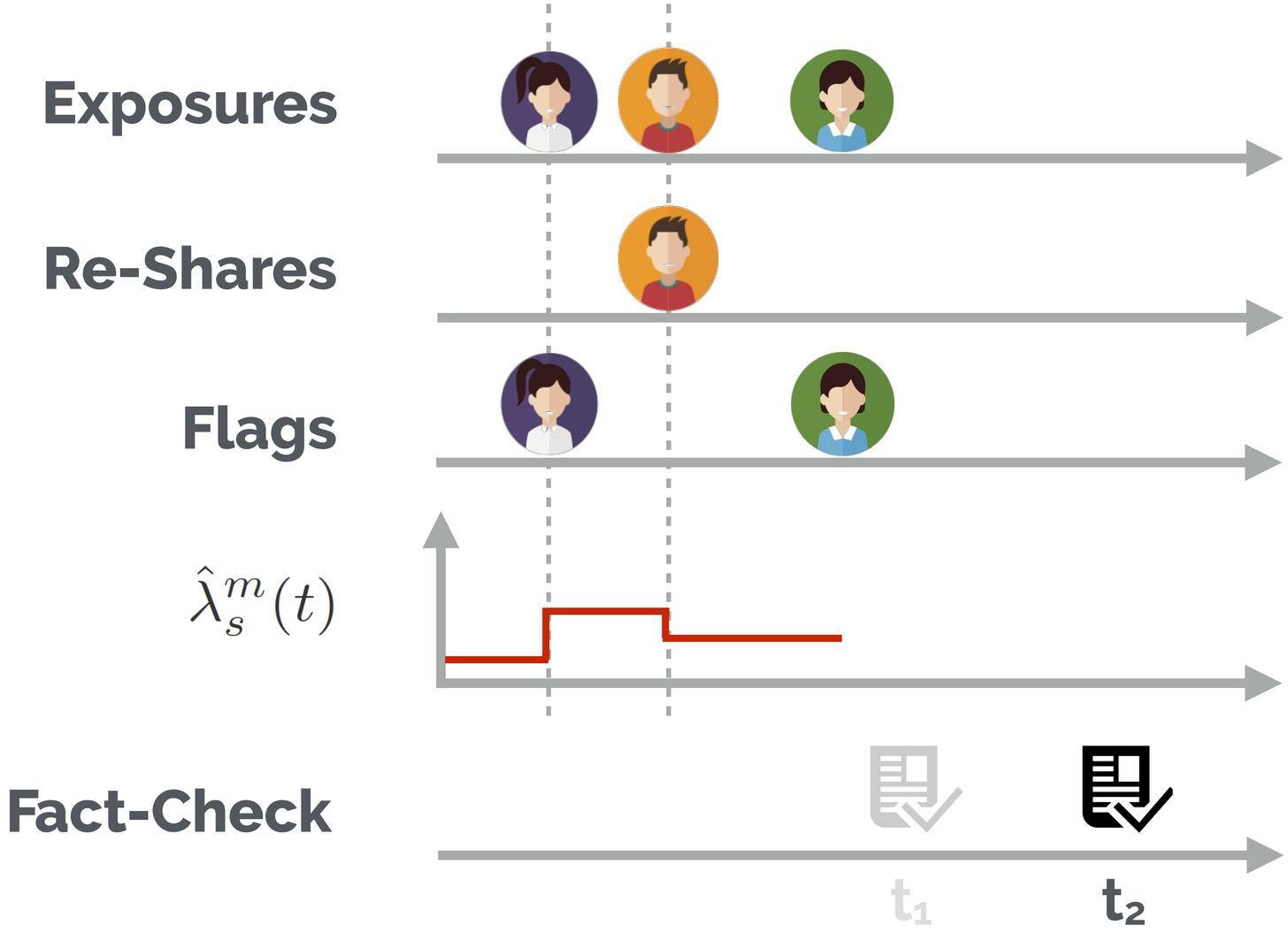
The *CURB* Algorithm



$t_1 \sim f(t_0, \hat{\lambda}_s^m(t))$

$t_2 \sim f(t_1, \hat{\lambda}_s^m(t))$

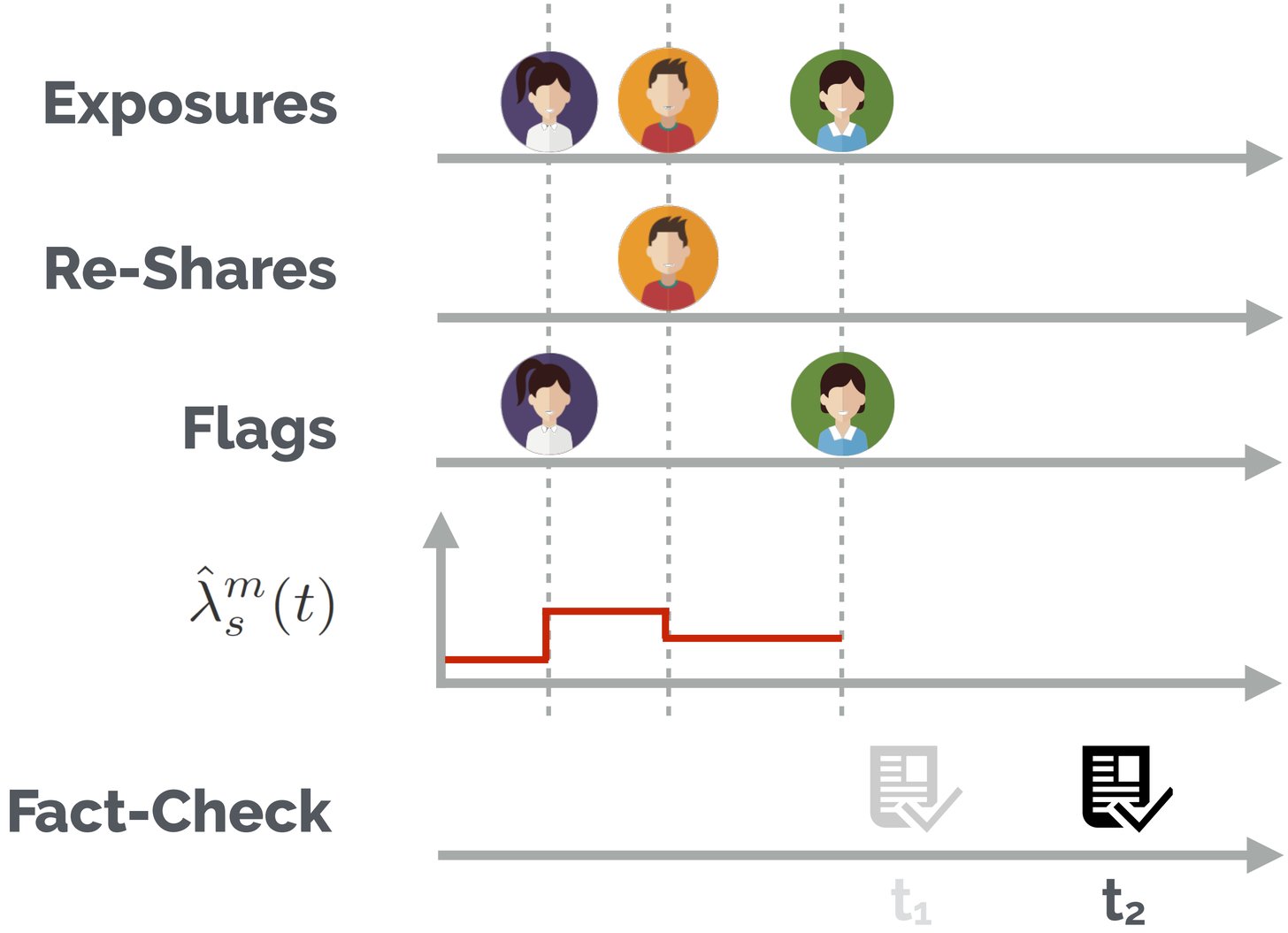
The *CURB* Algorithm



$t_1 \sim f(t_0, \hat{\lambda}_s^m(t))$

$t_2 \sim f(t_1, \hat{\lambda}_s^m(t))$

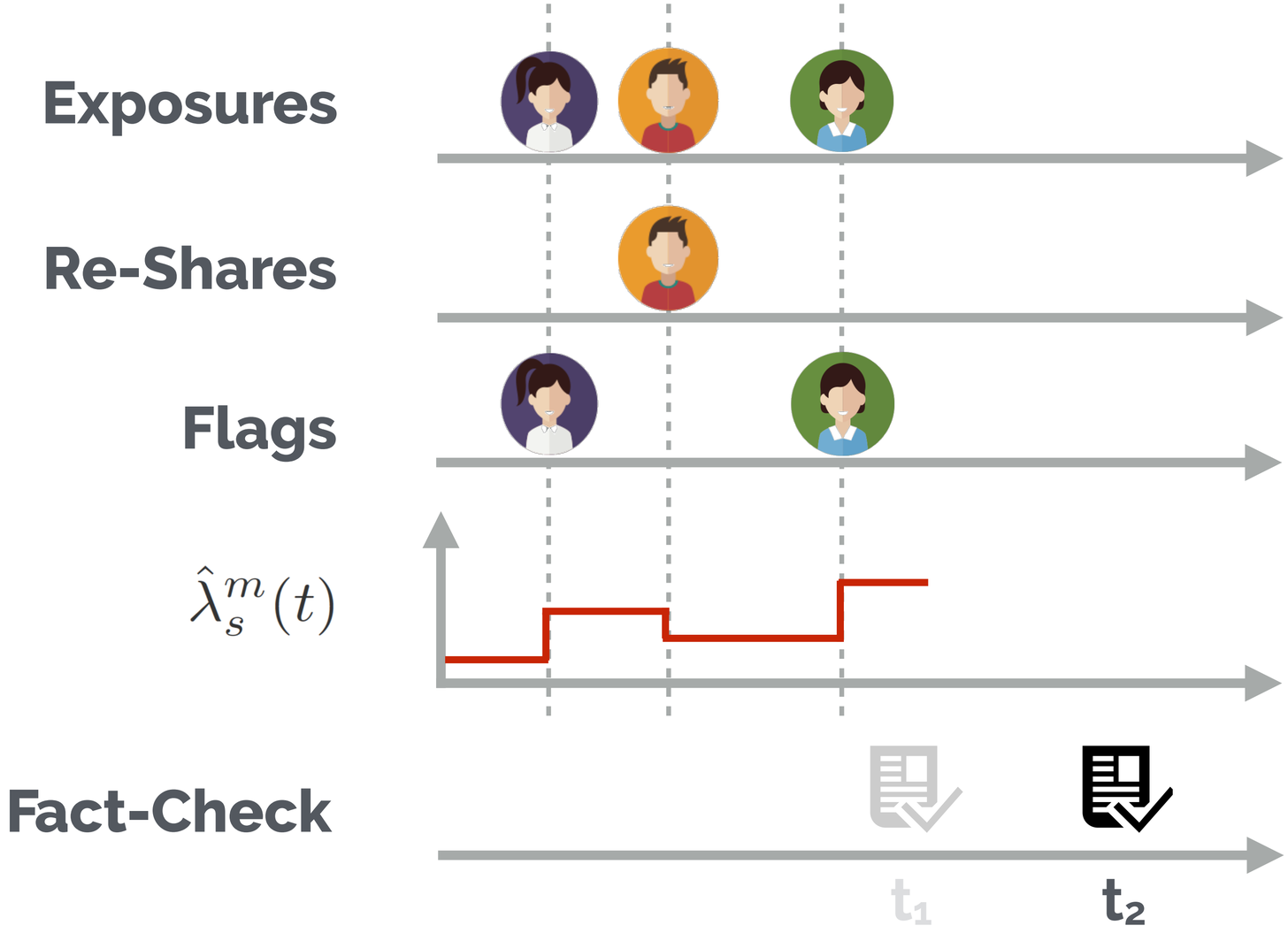
The *CURB* Algorithm



$t_1 \sim f(t_0, \hat{\lambda}_s^m(t))$

$t_2 \sim f(t_1, \hat{\lambda}_s^m(t))$

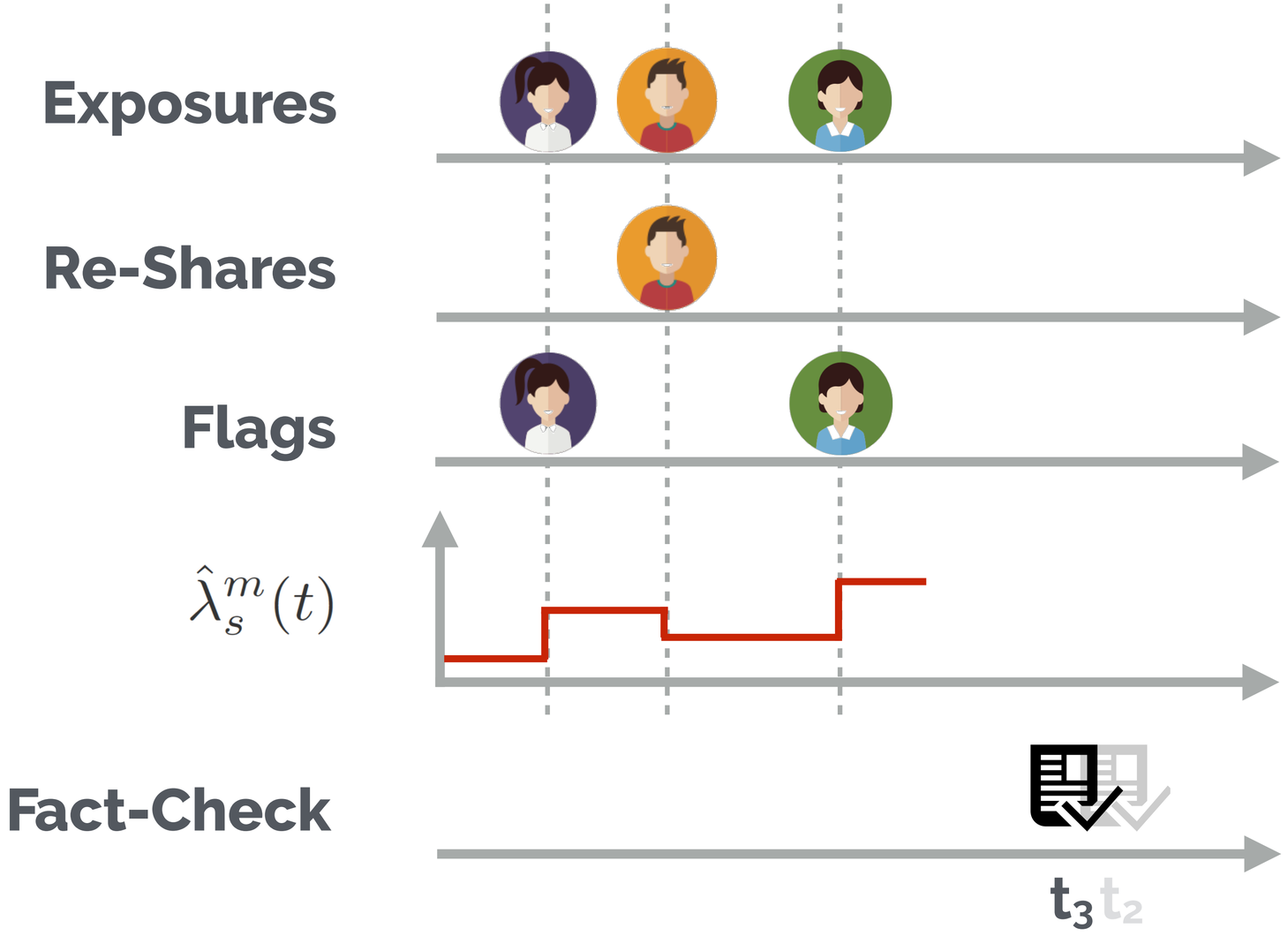
The *CURB* Algorithm



$t_1 \sim f(t_0, \hat{\lambda}_s^m(t))$

$t_2 \sim f(t_1, \hat{\lambda}_s^m(t))$

The *CURB* Algorithm

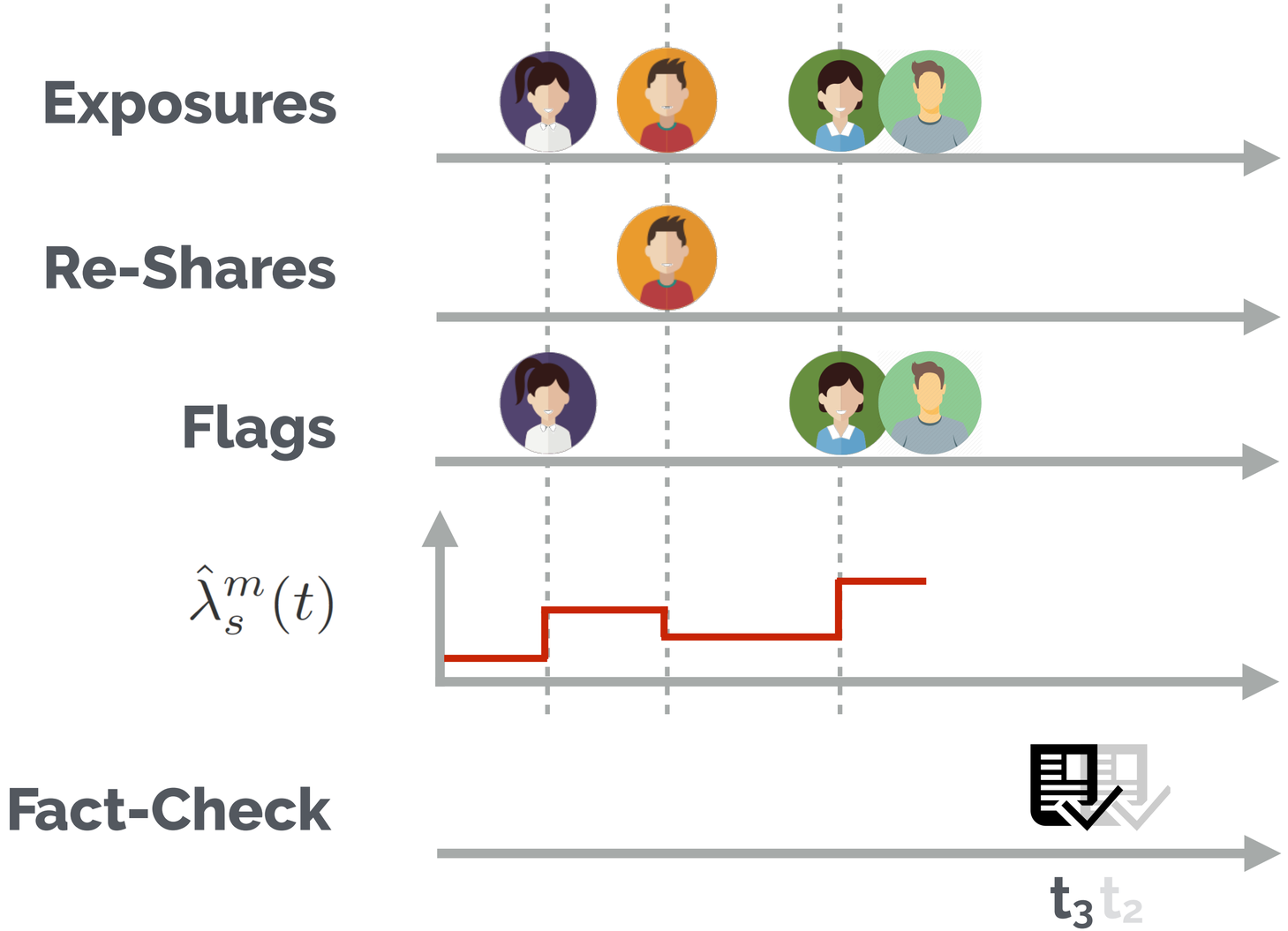


$t_1 \sim f(t_0, \hat{\lambda}_s^m(t))$

$t_2 \sim f(t_1, \hat{\lambda}_s^m(t))$

$t_3 \sim f(t_2, \hat{\lambda}_s^m(t))$

The *CURB* Algorithm

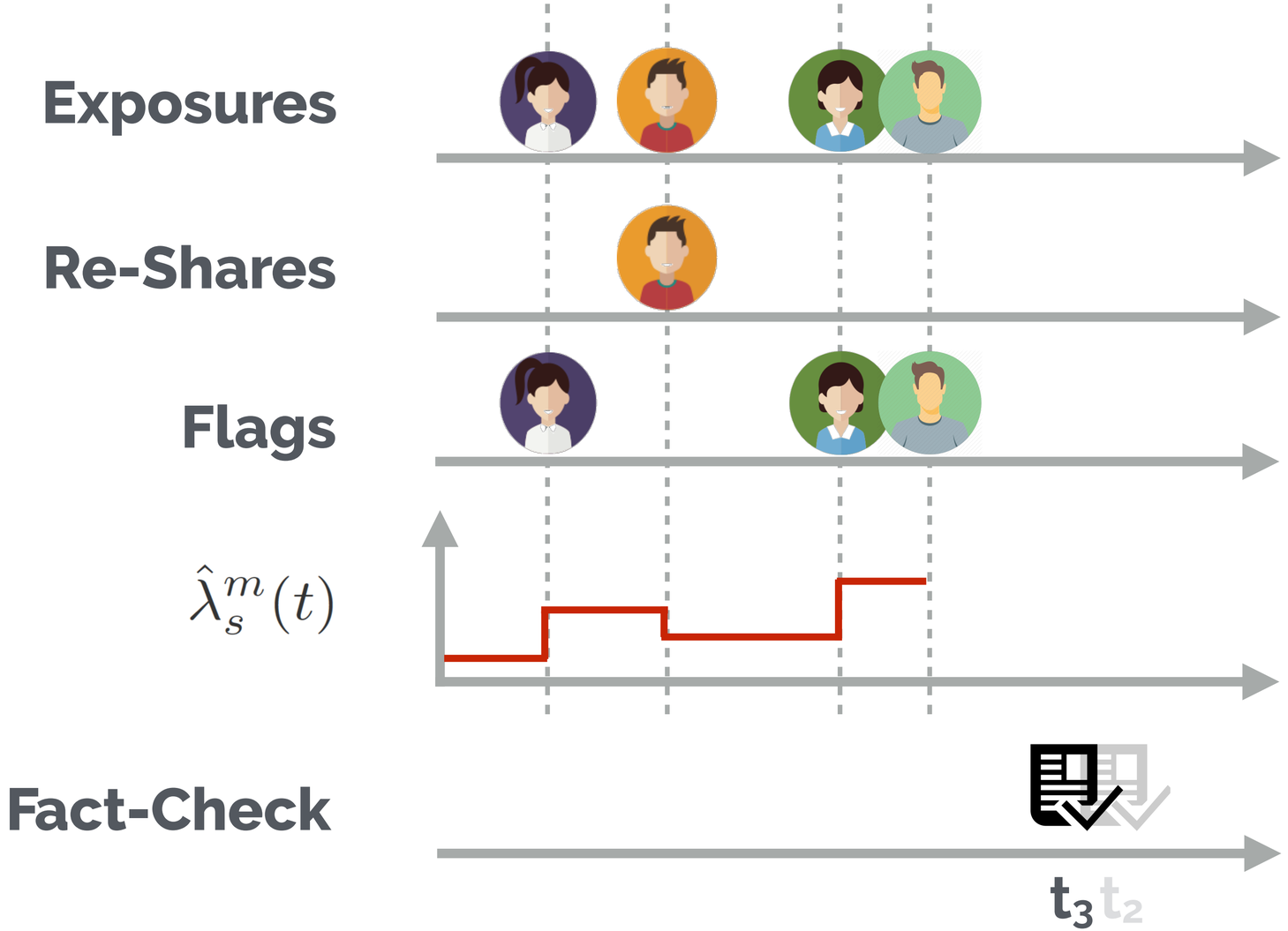


$t_1 \sim f(t_0, \hat{\lambda}_s^m(t))$

$t_2 \sim f(t_1, \hat{\lambda}_s^m(t))$

$t_3 \sim f(t_2, \hat{\lambda}_s^m(t))$

The *CURB* Algorithm

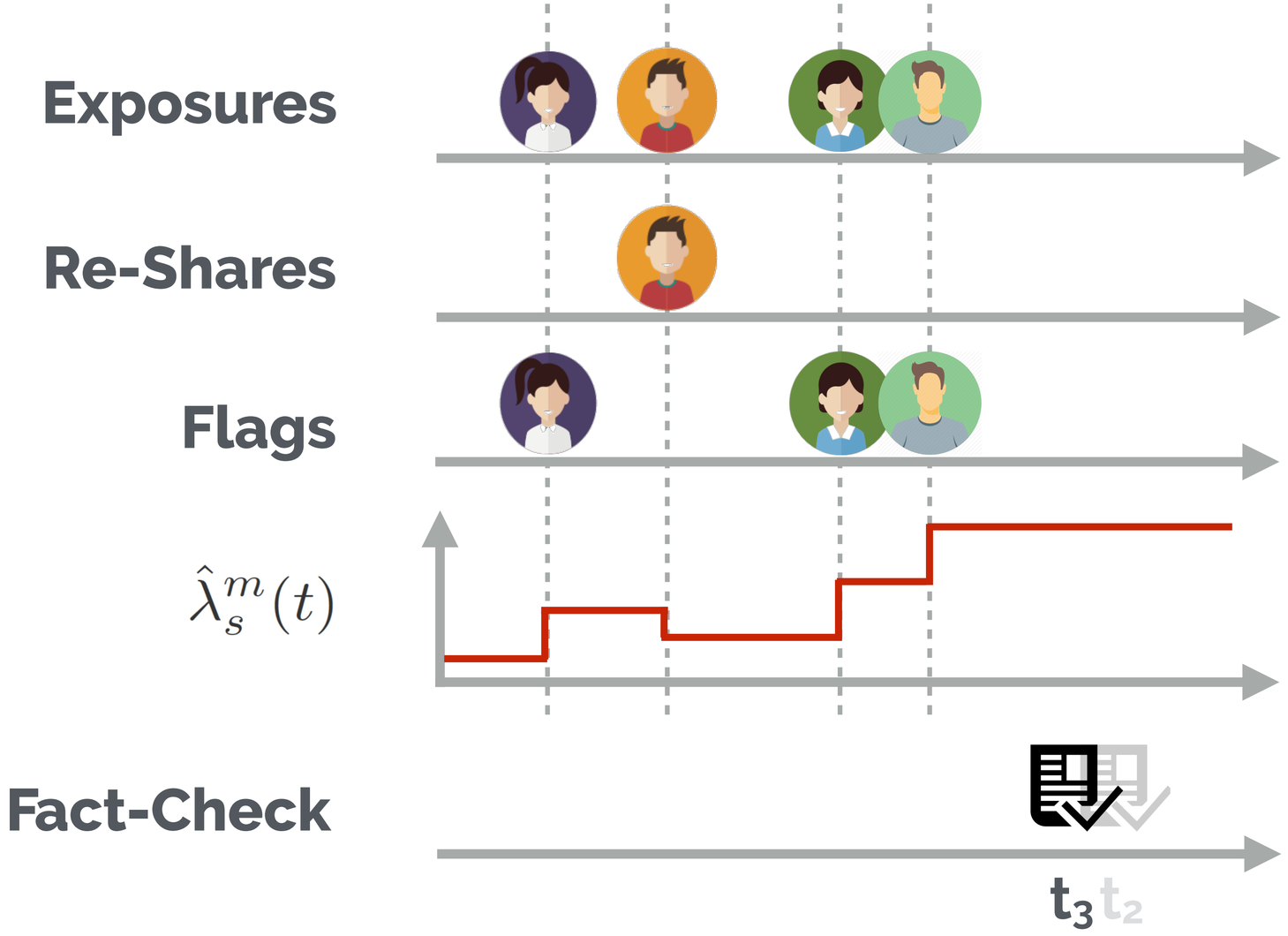


$t_1 \sim f(t_0, \hat{\lambda}_s^m(t))$

$t_2 \sim f(t_1, \hat{\lambda}_s^m(t))$

$t_3 \sim f(t_2, \hat{\lambda}_s^m(t))$

The *CURB* Algorithm

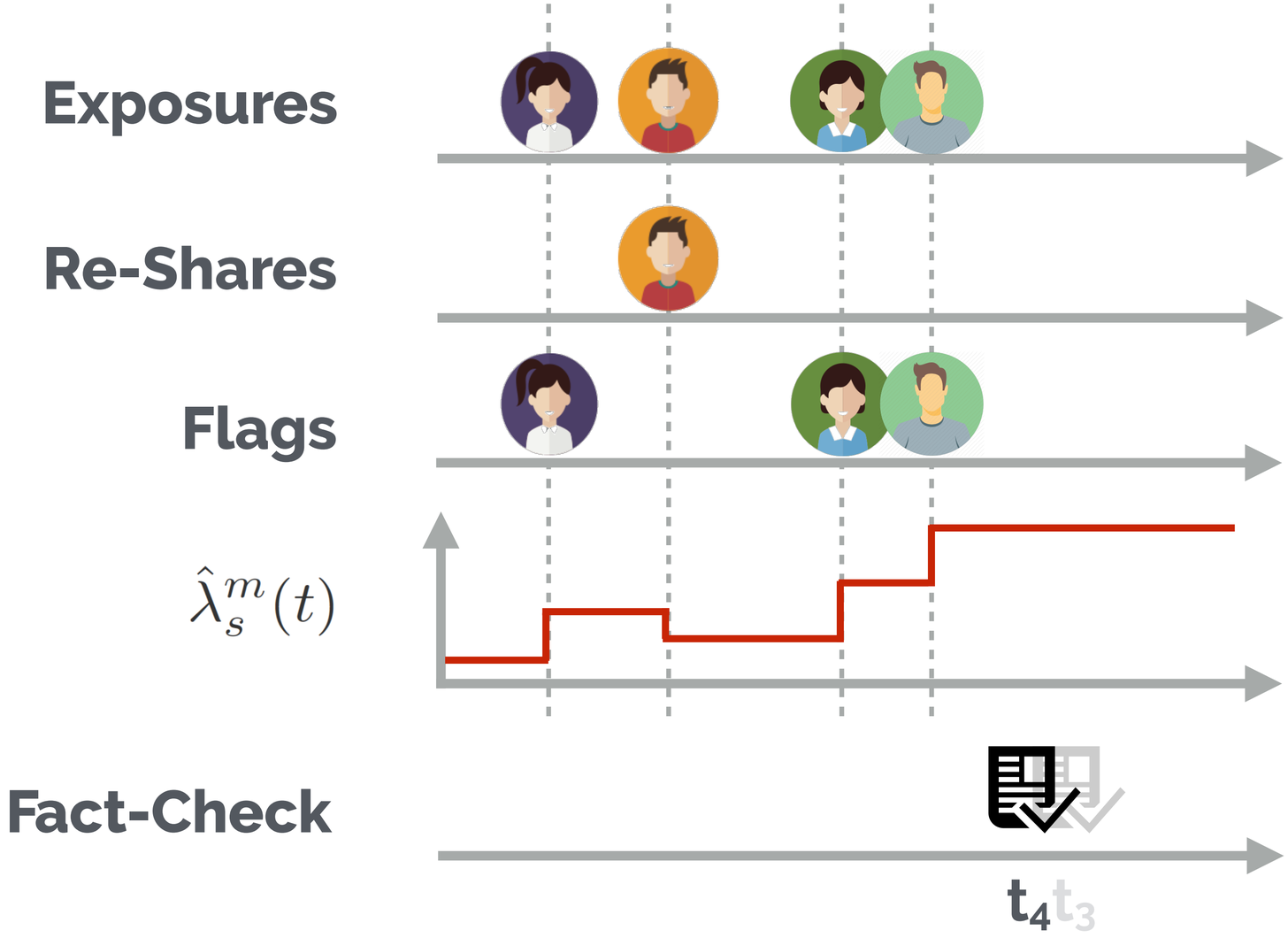


$t_1 \sim f(t_0, \hat{\lambda}_s^m(t))$

$t_2 \sim f(t_1, \hat{\lambda}_s^m(t))$

$t_3 \sim f(t_2, \hat{\lambda}_s^m(t))$

The *CURB* Algorithm



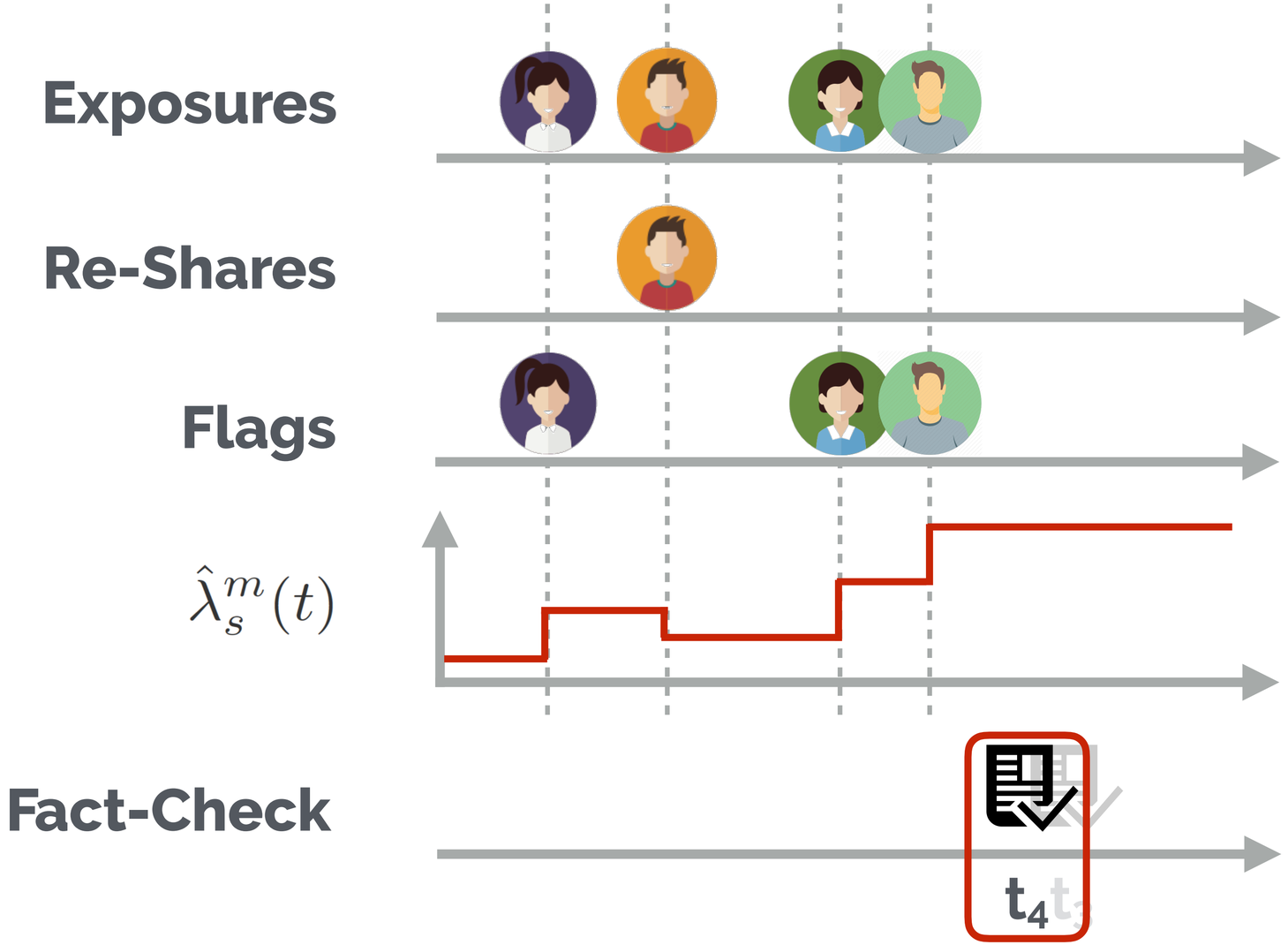
$t_1 \sim f(t_0, \hat{\lambda}_s^m(t))$

$t_2 \sim f(t_1, \hat{\lambda}_s^m(t))$

$t_3 \sim f(t_2, \hat{\lambda}_s^m(t))$

$t_4 \sim f(t_3, \hat{\lambda}_s^m(t))$

The *CURB* Algorithm



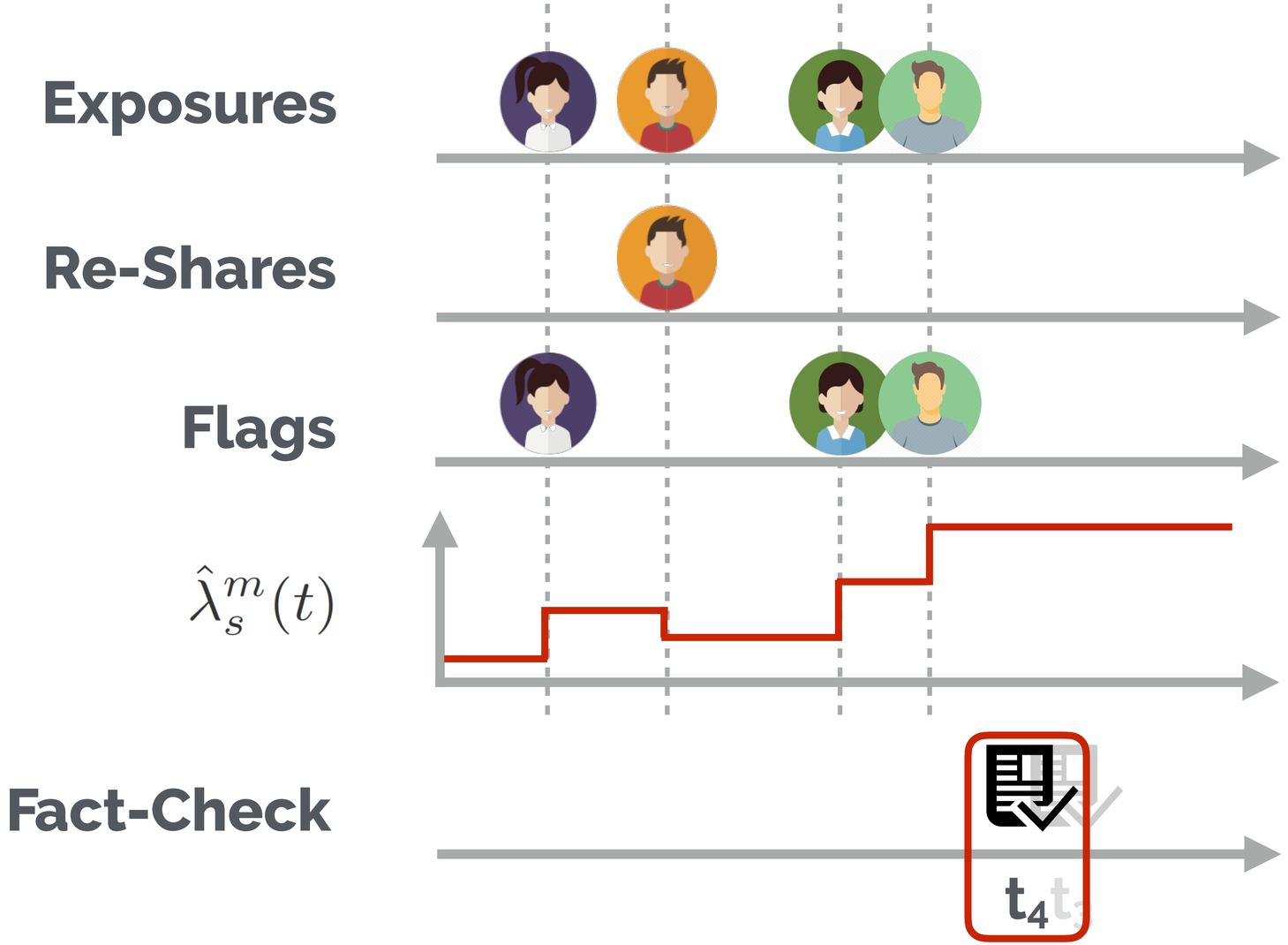
$t_1 \sim f(t_0, \hat{\lambda}_s^m(t))$

$t_2 \sim f(t_1, \hat{\lambda}_s^m(t))$

$t_3 \sim f(t_2, \hat{\lambda}_s^m(t))$

$t_4 \sim f(t_3, \hat{\lambda}_s^m(t))$

The *CURB* Algorithm



Sample $O(N^e(t_f))$ times

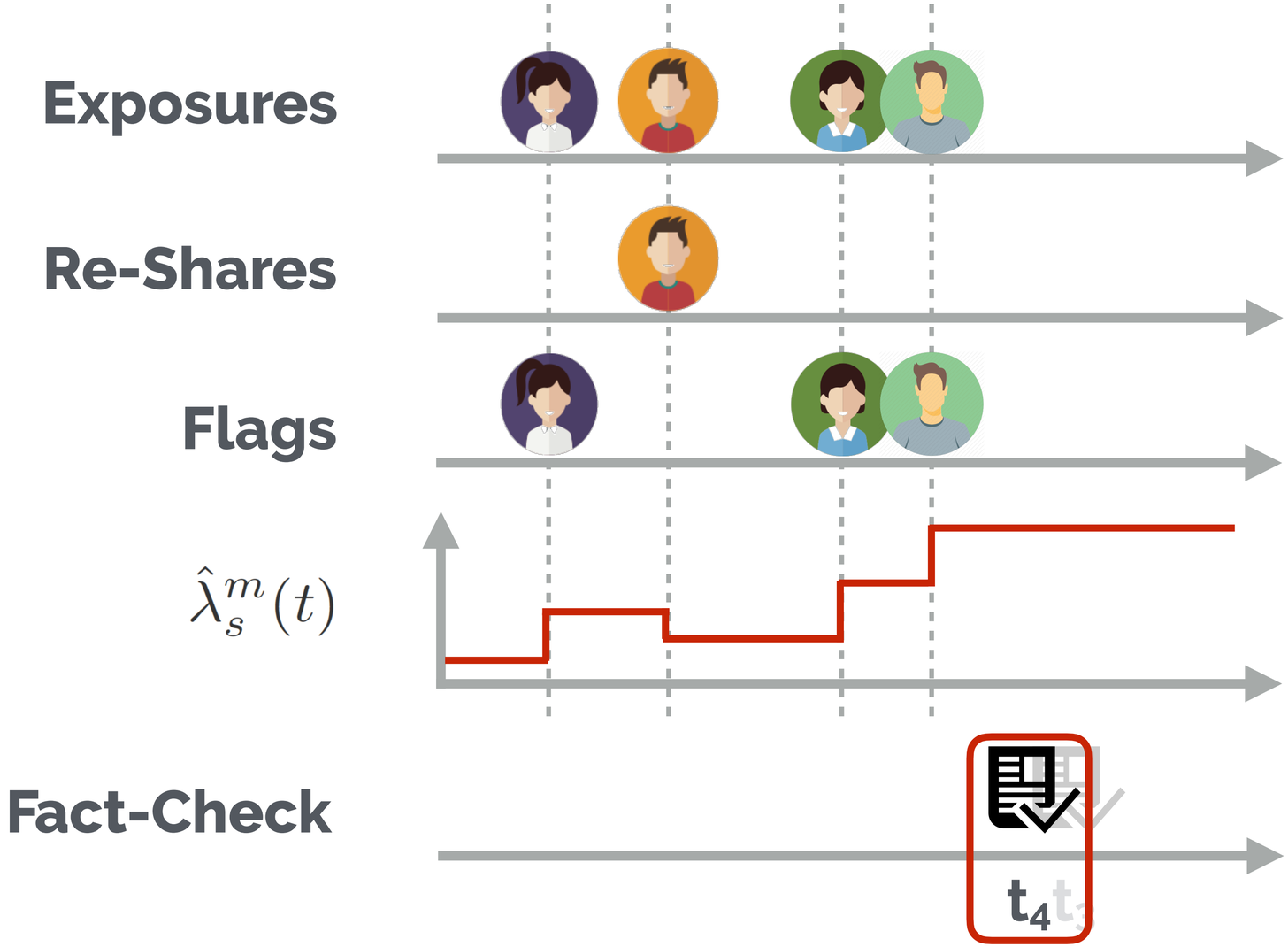
$t_1 \sim f(t_0, \hat{\lambda}_s^m(t))$

$t_2 \sim f(t_1, \hat{\lambda}_s^m(t))$

$t_3 \sim f(t_2, \hat{\lambda}_s^m(t))$

$t_4 \sim f(t_3, \hat{\lambda}_s^m(t))$

The *CURB* Algorithm



Sample $O(N^e(t_f))$ times

The function f(..) uses

- ➔ standard thinning
- ➔ Superposition principle

$t_1 \sim f(t_0, \hat{\lambda}_s^m(t))$

$t_2 \sim f(t_1, \hat{\lambda}_s^m(t))$

$t_3 \sim f(t_2, \hat{\lambda}_s^m(t))$

$t_4 \sim f(t_3, \hat{\lambda}_s^m(t))$

1. Post-Truth Era and Fake News

2. Crowd-Powered Fact-Checking

3. Fact-Checking Scheduling Problem

4. Algorithm: *CURB*

5. **Detect and Reduce the Spread of Fake News**

img credit: https://www.socialpinpoint.com/2016/wp-content/uploads/managing-misinformation-1.jpg

Datasets

Publicly available manually annotated data from Twitter, Weibo

- 28k shares & re-shares from 19k users
- 7 fake stories & 39 genuine stories

- 93k shares & re-shares from 89k users
- 23 fake stories & 133 genuine stories

Datasets

Publicly available manually annotated data from Twitter, Weibo

- 28k shares & re-shares from 19k users
- 7 fake stories & 39 genuine stories

- 93k shares & re-shares from 89k users
- 23 fake stories & 133 genuine stories

NO information about timings of exposures nor flags → Simulate via sampling

<u>Model's efficacy is robust throughout multiple parameter settings</u>

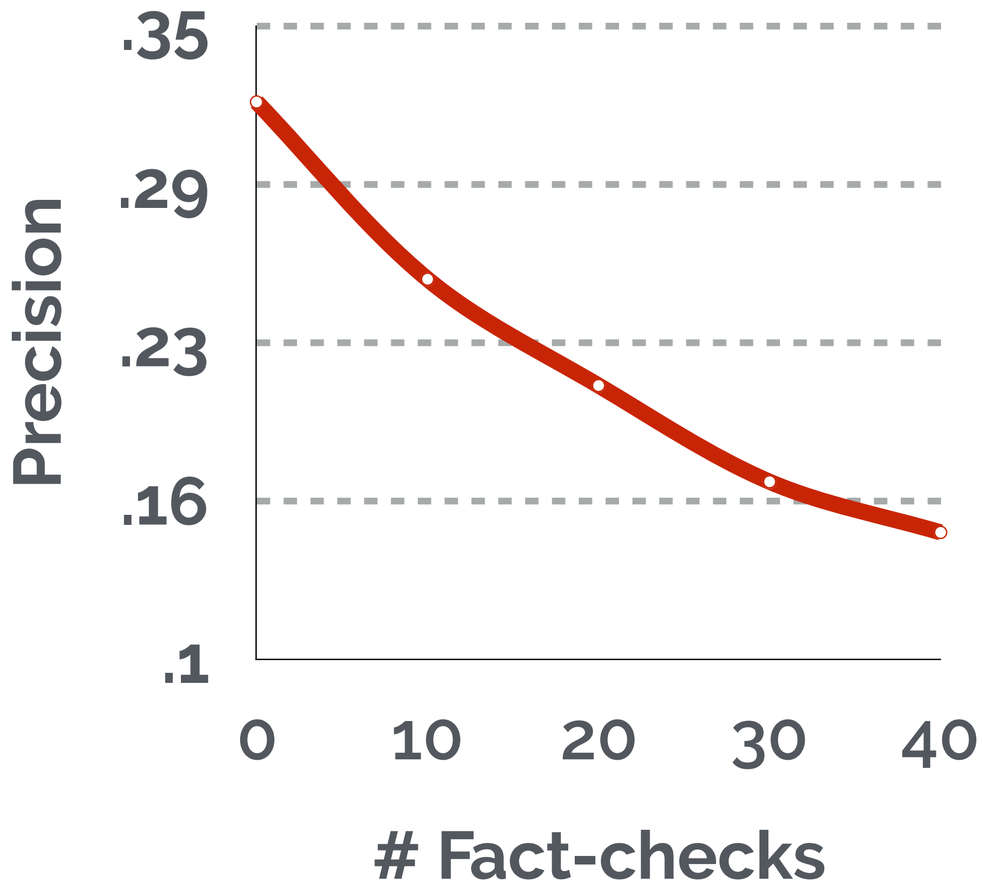
Evaluation Metrics

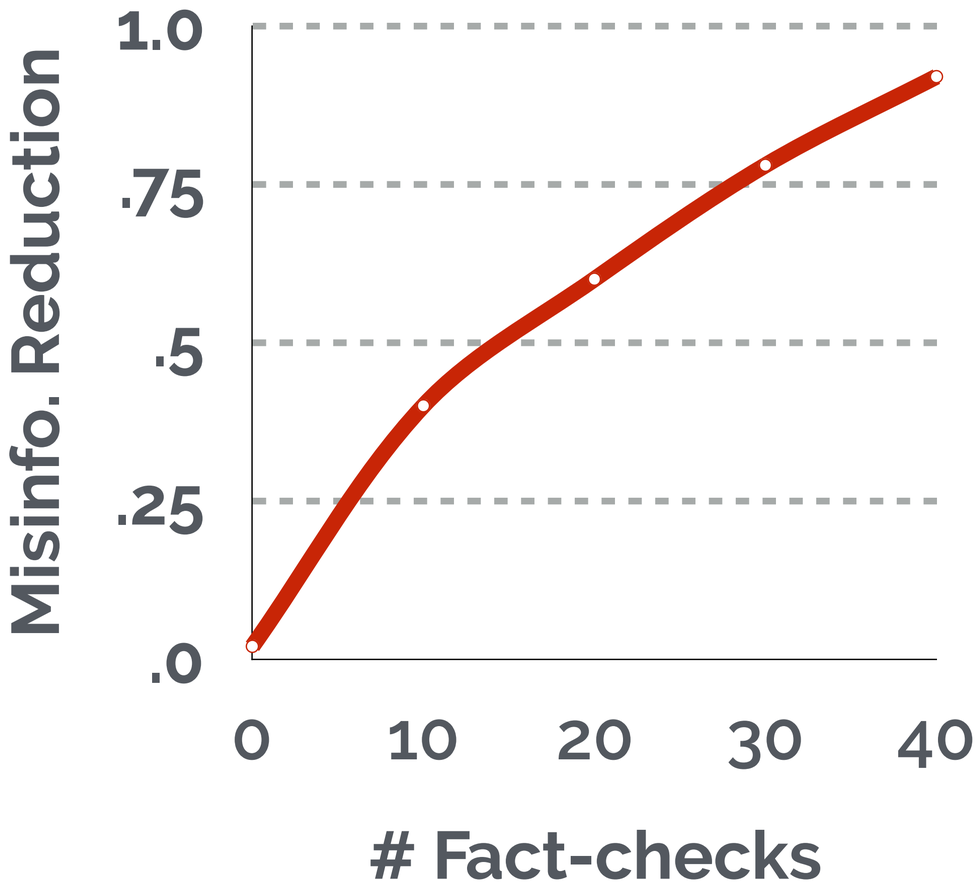
Two Metrics: precision & misinformation reduction

→ Each accounts for the different ends of the trade-off

Precision	• Fraction of fact checked stories that are fake
Misinfo. Reduction	• Fraction of exposures to misinformation that fact checking prevented

Performance vs. # Fact-checks



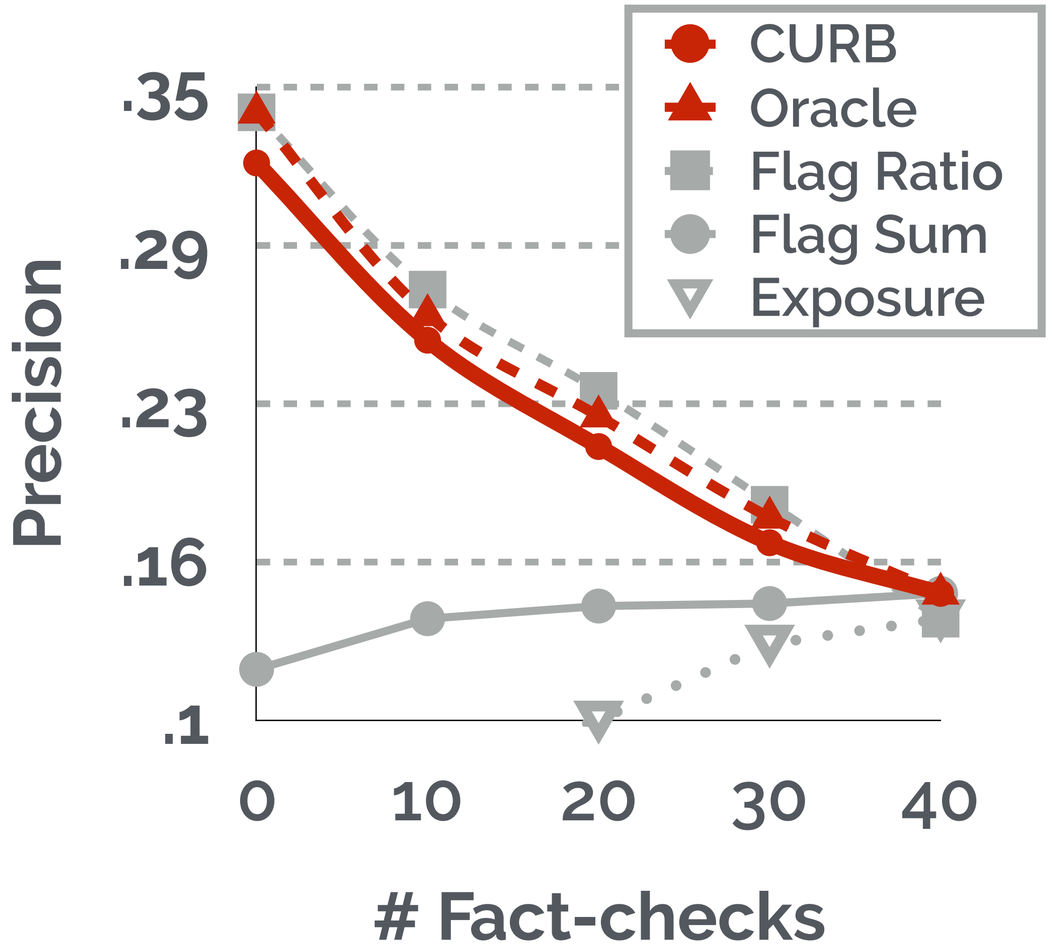


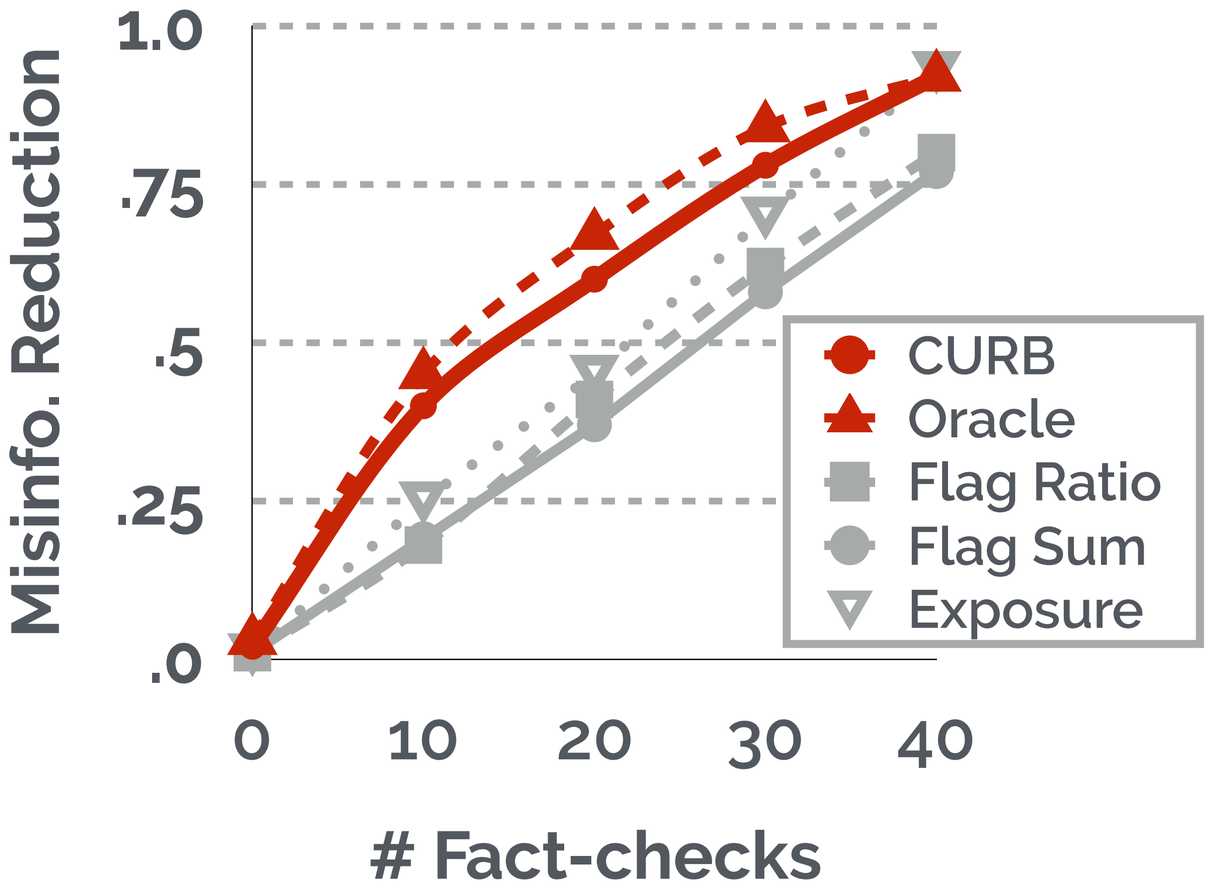
<Twitter>

Comparison Methods

Oracle	• A variant of ***CURB*** • Access to the true flag probability • Upper bound of ***CURB***
Exposure	• Rate: proportional to the exposure rate
Flag Ratio	• Rate: ratio between the number of flags and the total number of exposures
Flag Sum	• Fact-check when a certain number of flags are accumulated

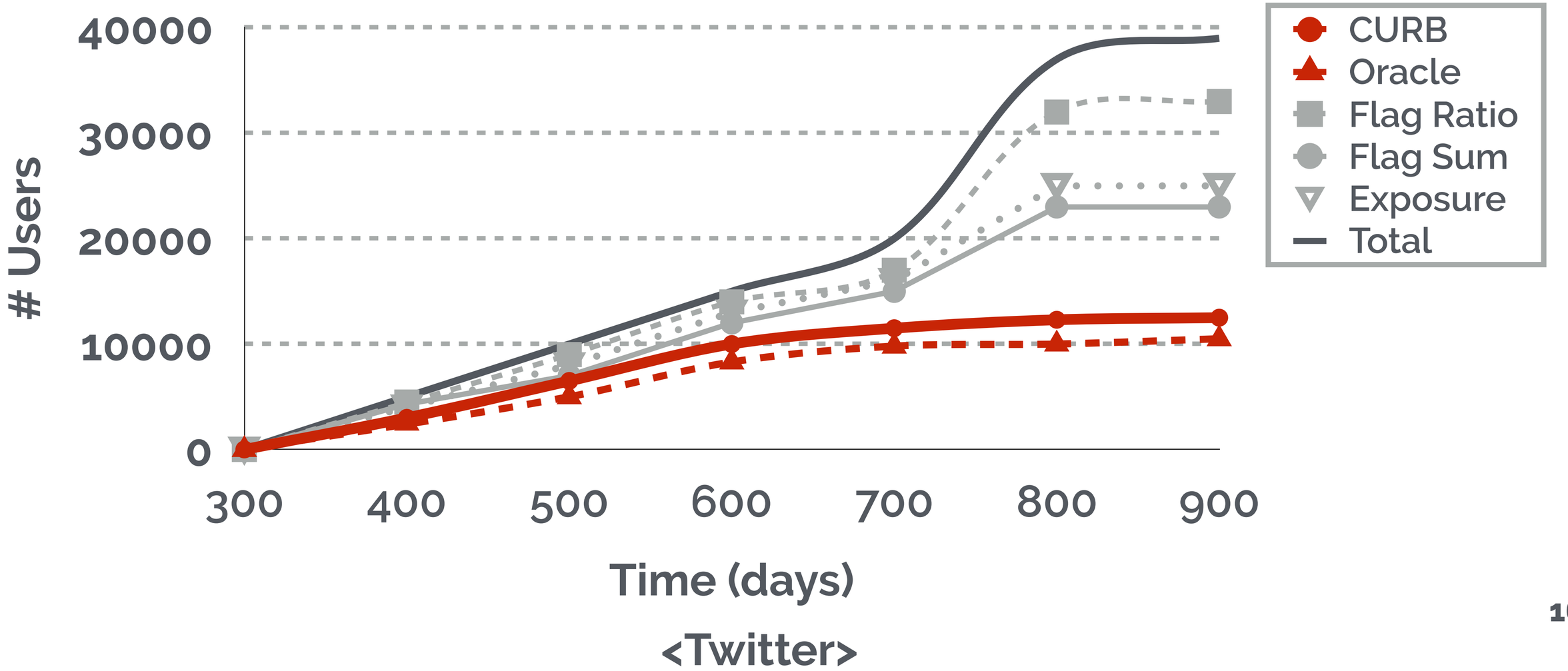
Performance vs. # Fact-checks





<Twitter>

Performance vs. Time



<Twitter>

Conclusion

- Aim to reduce the spread of misinformation
- Use user flag, time information for modeling
- Yield better results (precision, misinfo. reduction) compared to the baselines

Jooyeon Kim jooyeon.kim@kaist.ac.kr

more at learning.mpi-sws.org/curb/

Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation

Jooyeon Kim jooyeon.kim@kaist.ac.kr

more at learning.mpi-sws.org/curb/

img credit: http://www.pewinternet.org/2017/10/19/the-future-of-truth-and-misinformation-online/