Robin Li Data Mining & Machine learning laboratory at State University of New York at Buffalo

cse@buffalo

SDM'15 Vancouver, CAN

Believe it Today or Tomorrow? Detecting Untrustworthy Information from Dynamic Multi-Source Data

Houping Xiao[1], Yaliang Li[1], Jing Gao[1], Fei Wang[2], Liang Ge[3], Wei Fan[4], Long Vu[5], and Deepak Turaga[5]

[1] SUNY at Buffalo; [2] University of Connecticut;
[3] Google; [4] Baidu Big Data lab; [5] IBM T.J. Watson

Outline

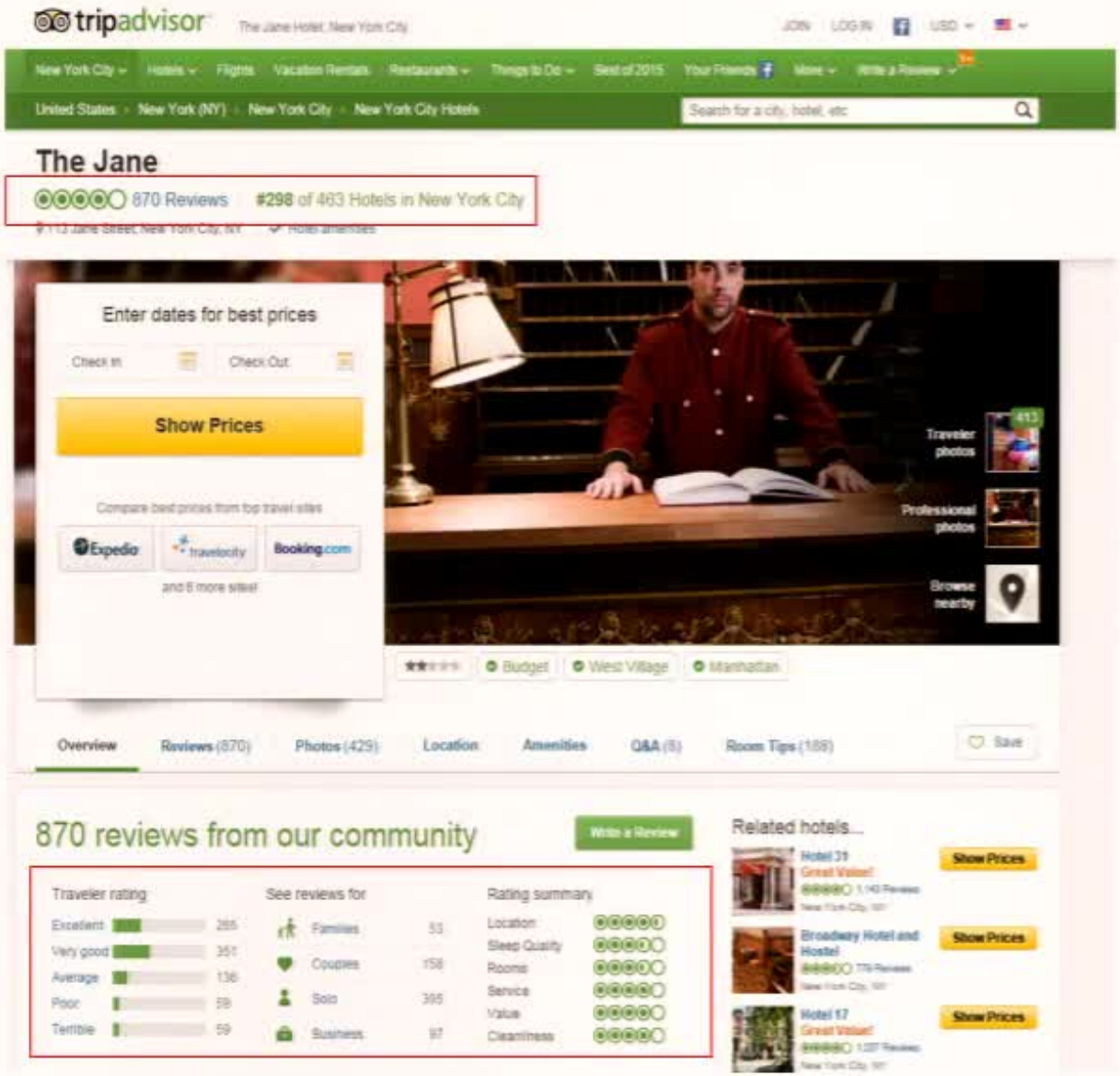
- **Motivation**
- **Challenges**
- **Proposed Two-Step Framework**
  - Step-1: Joint Nonnegative Tensor Factorization
  - Step-2: Inconsistency Score calculation
- **Experiments**
- **Conclusions**

Motivation

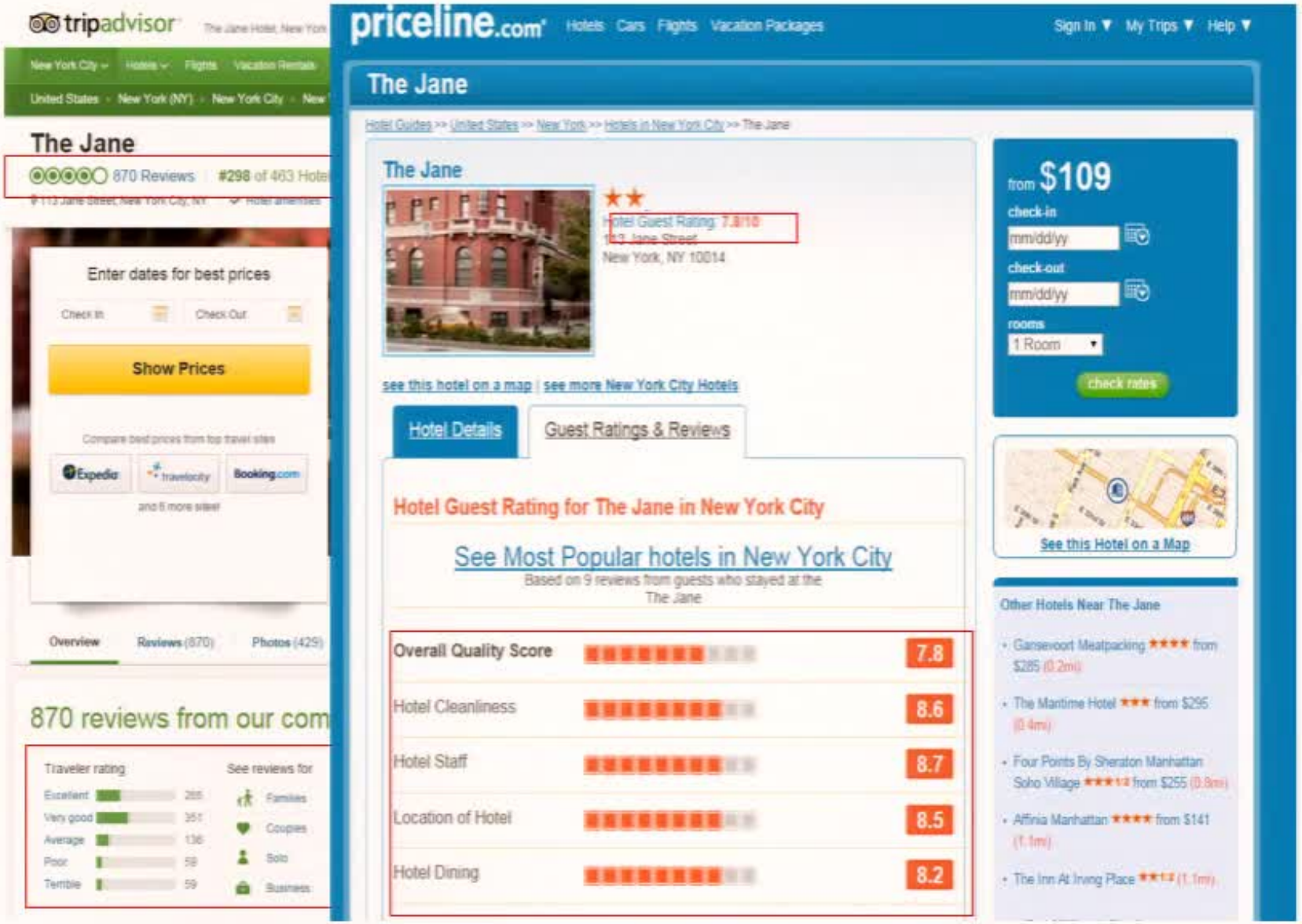
- **Multiple Information Sources**
  - Example: Hotel ratings can be obtained from multiple websites, such as Priceline, Orbitz, and Tripadvisor
- **Question?**
  - Which piece of information is trustworthy?
  - Which object does receive reliable information?
- **Our solution**
  - Calculate the degree of receiving inconsistent information across sources
  - Lower degree of inconsistency – more reliable

Motivating Example

Motivating Example



Motivating Example



Motivating Example

How to Find Inconsistent Ratings

- **Easy comparisons**
  - Aggregate the ratings of all the users into average ratings will lose information
  - Users are not matched across different platforms, so we are unable to compare ratings of each user
- **Our solution: Identify user groups and compare at the group level**
  - In each source, users can be partitioned into groups so that users in the same group share similar rating patterns over objects
  - The underlying user groups and the ratings given by each group should be consistent across sources

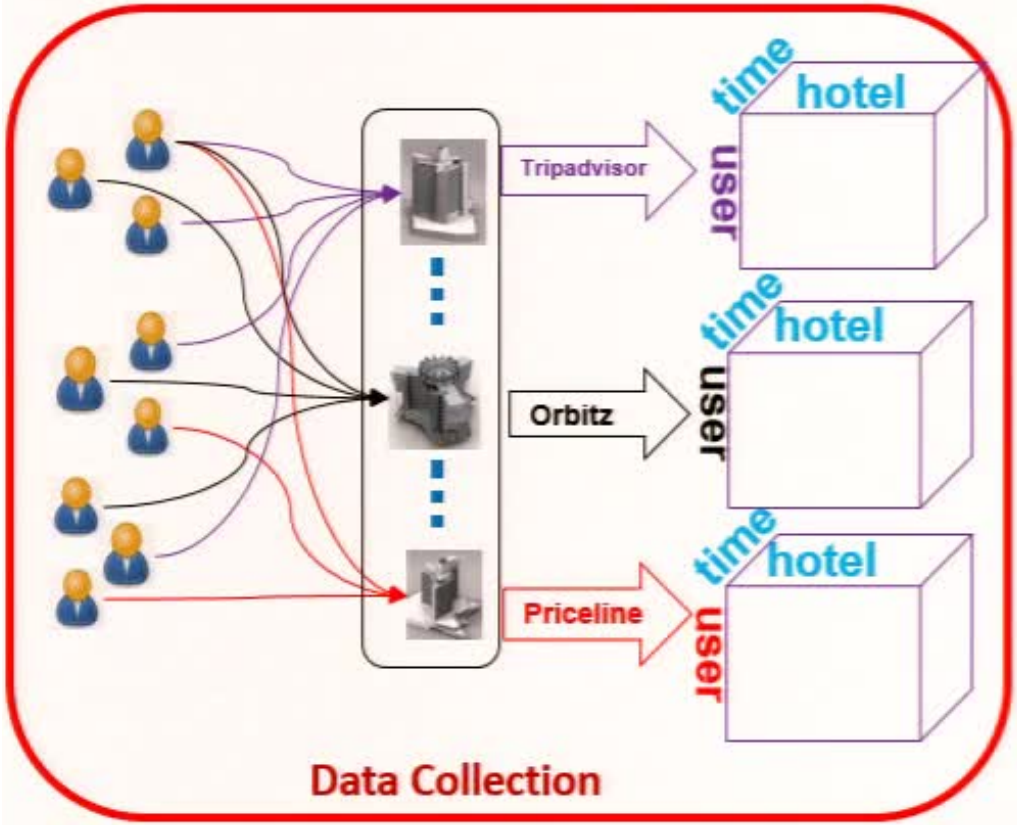
Importance of Time

- **Observations**
  - Multi-source data can continuously arrive with constantly changing distributions
- **Motivating Example**

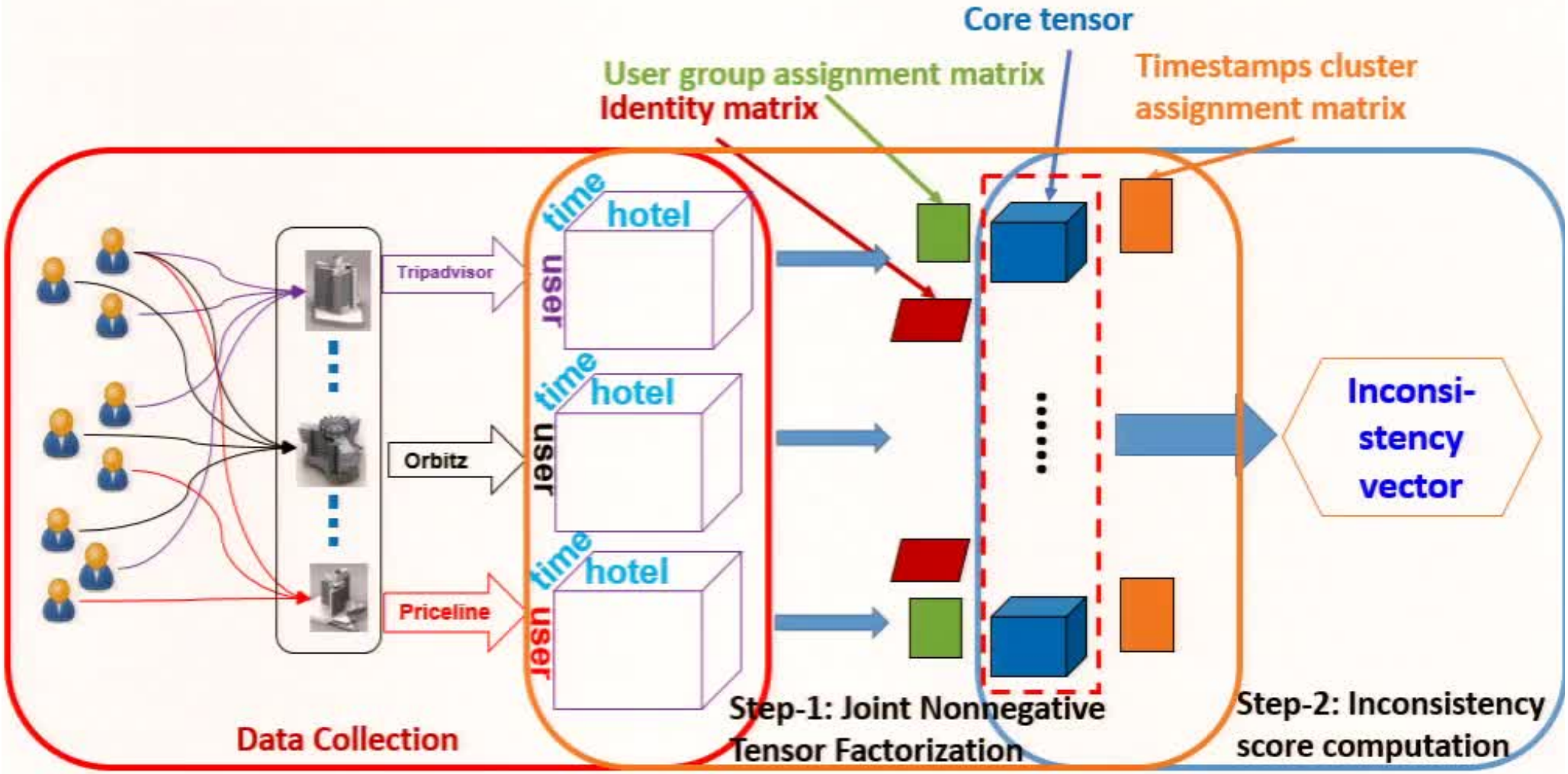
Solutions

- **Baselines**
  - Conduct separate modeling on each snapshot (Simple, but the temporal connection between timestamps is missing)
- **Our Solution**
  - Consider the behavior at timestamp-cluster level (e.g. hotel ratings could change seasonally)
  - In each source, timestamps can be clustered. Users' behavior at the same timestamp cluster should be similar

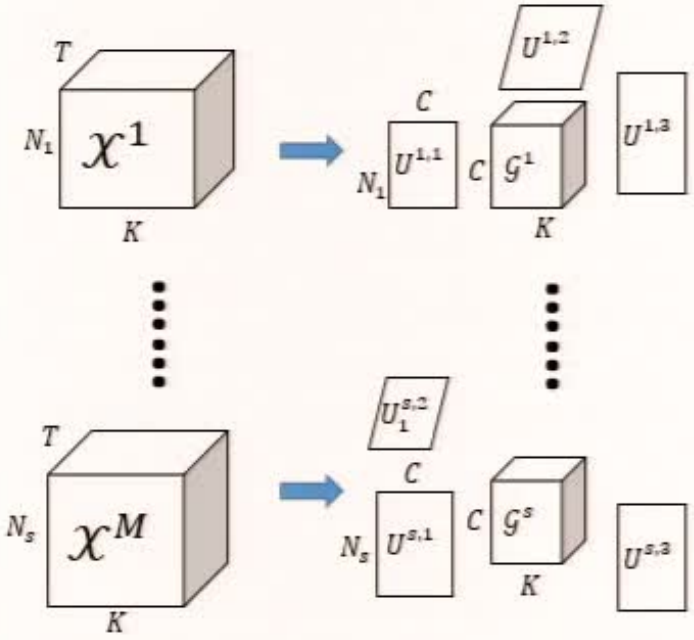
Proposed Framework



Proposed Framework

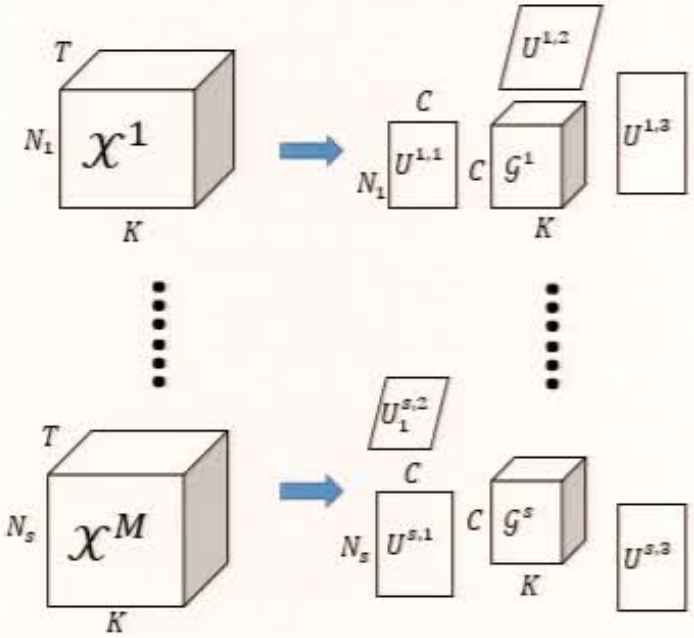


Joint Nonnegative Tensor Factorization



$$\min_{\{\mathcal{G}^s,\{U^{s,i}\}\}\geq 0} \sum_{s=1}^{M} \left(\mathcal{L}\left(\tilde{\mathcal{X}}^s\right) + \alpha\Omega(\mathcal{G}^s, \mathcal{G}^*)\right)$$

Joint Nonnegative Tensor Factorization



$$\min_{\{\mathcal{G}^s,\{U^{s,i}\}\}\geq 0} \sum_{s=1}^{M} \left(\mathcal{L}\left(\tilde{\mathcal{X}}^s\right) + \alpha\Omega(\mathcal{G}^s, \mathcal{G}^*)\right)$$

- $\mathcal{L}\left(\tilde{\mathcal{X}}^s\right) = \left\|\mathcal{X}^s - \mathcal{G}^s \times_i \{U^{s,i}\}\right\|_F^2$, measures the factorization error of each tensor
- $\Omega(\mathcal{G}^s, \mathcal{G}^*) = \|\mathcal{G}^s - \mathcal{G}^*\|_F^2$, where $\mathcal{G}^* = \frac{1}{M}\sum_{s=1}^{M} \mathcal{G}^s$, is a regularization term proposed to learn the consensus information
- α, is regularization parameter

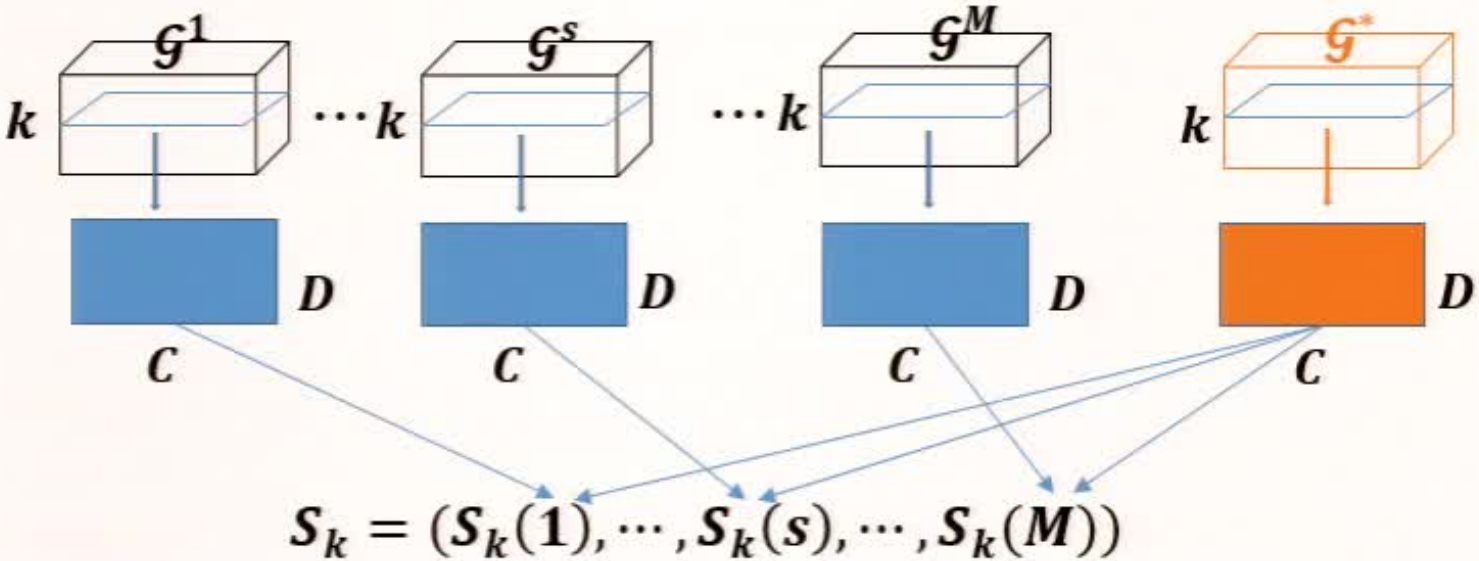
Inconsistency Score Computation

- **Inconsistency Score**

$$I_k = \|S_k - S_{median}\|^2,$$

where

- $S_k(s) = similarity\,(\mathcal{G}^s(:,k,:), \mathcal{G}^*(:,k,:))$
- $S_{median} = median\,\{S_k, k = 1, \cdots, K\}$



Streaming Data

- **Observation:**
  - Multi-source data continuously arrives
- **Solution:**
  - Step 1: obtain $\{U_o^{(s,T,i)}\}$ at time T based on $\{\mathcal{G}^{s,T-1}\}$

$$\min_{\{U^{(s,i,T)}\}} \sum_{s=1}^{M} \left\| \mathcal{X}^{s,T} - \mathcal{G}^{s,T-1} \times_i \{U^{(s,i,T)}\} \right\|_F^2$$

  - Step 2: use $\{U_o^{(s,T,i)}\}$ to obtain $\{\mathcal{G}^{s,T}\}$ at time T

$$\min_{\{\mathcal{G}^{s,T}\}} \sum_{s=1}^{M} \sum_{t=1}^{T} \left\| \mathcal{X}^{s,t} - \mathcal{G}^{s,T} \times_i \{U_o^{(s,t,i)}\} \right\|_F^2 + \alpha \left\| \mathcal{G}^{s,T} - \mathcal{G}^{*(T-1)} \right\|_F^2$$

Missing Data

- **Observation:**
  - Many users may only give ratings for a few hotels at some specific timestamps
- **Solution:**
  - $\mathcal{K}^s = \{(i,j,k)\colon \mathcal{X}^s_{ijk}\ is\ available\}$ is a triple-element set
  - Objective function

$$\min_{\{\mathcal{G}^s,\{U^{s,i}\}\}\geq 0} \sum_{s=1}^{M} \sum_{(i,j,k)\in\mathcal{K}^s} \left(\mathcal{X}^s_{ijk} - \left(\mathcal{G}^s \times_i \{U^{s,i}\}\right)_{ijk}\right)^2 + \alpha\left(\mathcal{G}^s_{ijk} - \mathcal{G}^*_{ijk}\right)^2$$

Experiment Set-up

- **Datasets:**
  - **Synthetic datasets**
  - **Real-world datasets**
    - Hotel Rating
    - Network Traffic Flow
    - Weather Forecast