

# Weakly Supervised Clustering: Hybridizing Constructive Induction and Double Clustering

Xiangliang Zhang, Michèle Sebag, and Cécile Germain  
INRIA, CNRS and Université Paris Sud, 91405 Orsay, France  
{xlzhang, sebg, cecile}@lri.fr

- Goals:**
- Mining the clusters of Logging and Bookkeeping (L&B) files recorded by grid broker from EGEE (Enabling Grid for E-Science in Europe)
  - Specify failure modes of the submitted jobs (about 70% jobs failed for various reasons)

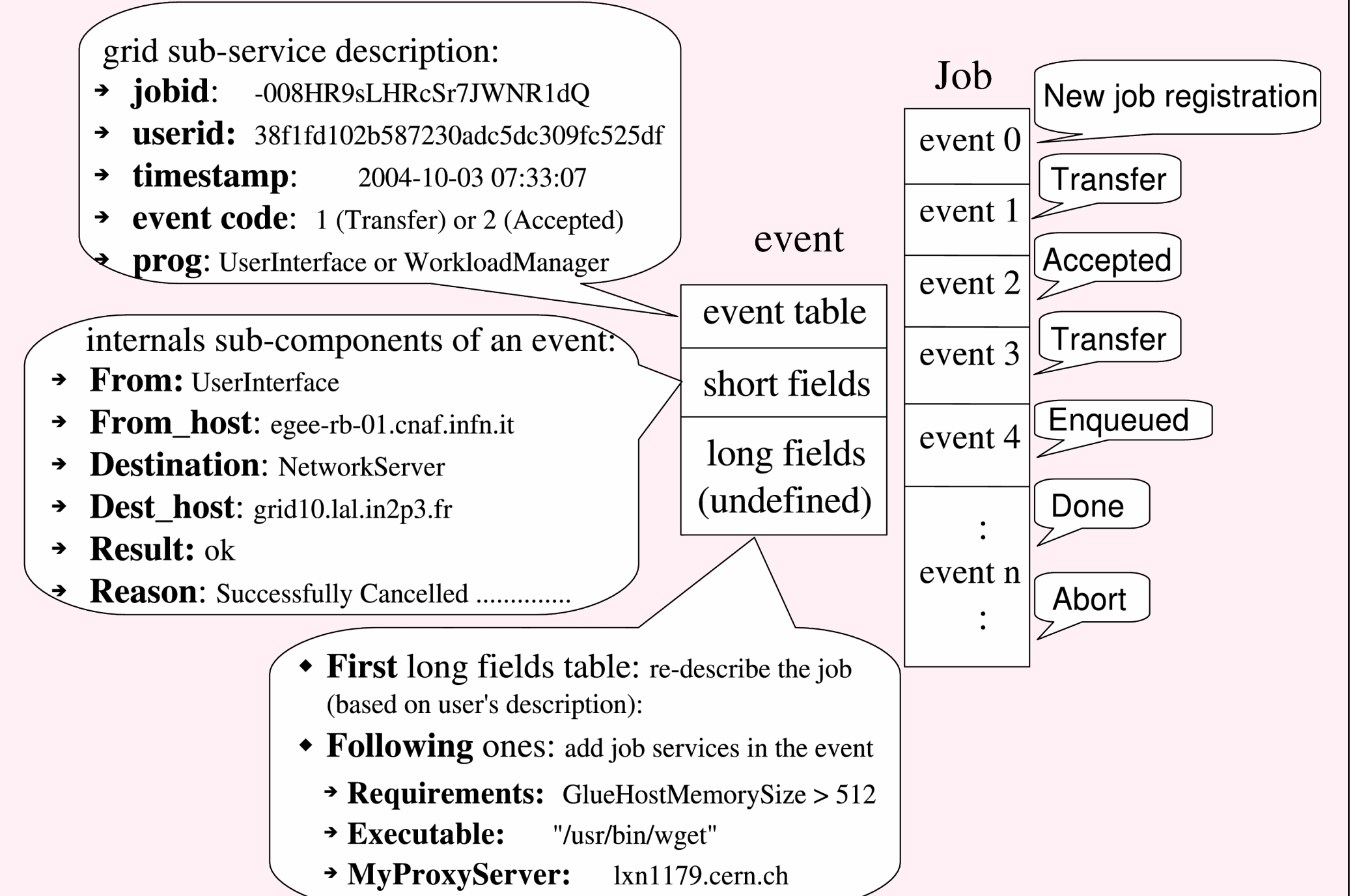
**Data:** traces of jobs submitted to grids

- Challenges:**
- ▶ No natural distance
  - ▶ Prior knowledge
    - rough classes
      - successfully finished (good jobs)
      - failed by various reasons (bad jobs): NAR, ABU, GNG \*
    - heterogeneous
      - users: experience and community are different
      - weeks: load of the grid varies along time

**Data Preparation:** Job -----> numerical vector

- ▶ numerical attributes: normalized
- ▶ non-numerical attributes ---> boolean attributes

## Data Structure



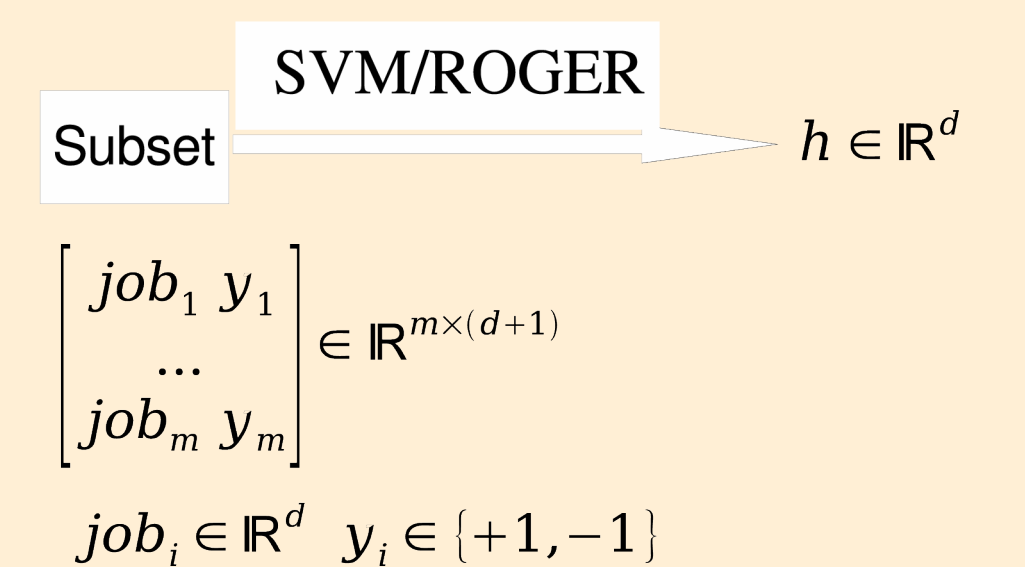
## Framework:

- ▶ **Constructive induction**
  - Based on data slicing
  - New representation of jobs
- ▶ **Double clustering**
  - Dimensionality reduction / Clusters of features to enforce stability of new features
  - Clusters of jobs and stability analysis of clusters [4]

## Constructive Induction (1):

- ▶ **Data**
  - test sets: 10% of all
  - training sets: 90% of all
- ▶ **Training sets slicing [1]**
  - one subset = single user (removing user heterogeneity) or single week (removing load heterogeneity)
  - each subset = successful jobs + failed jobs
- ▶ **Learning**
  - from each subset: Extract one (through SVM) or several (through ROC-based stochastic optimization, ROGER) hypotheses 'h' (from instance space to the set of Real values)
  - Turn each of these hypotheses into a feature

## SVM/ROGER [2] Learning



## U-features:

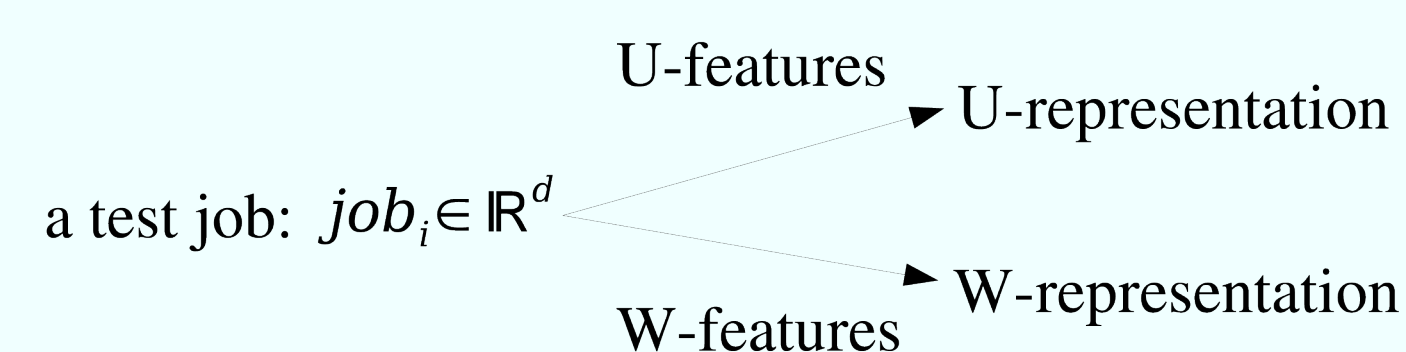
Hypotheses learned from User subsets

## W-features:

Hypotheses learned from Week subsets

## Constructive Induction (2):

### New Representation:



**ROGER based feature redundancy:**

from the same subset

redundancy of initial attributes

## Double Clustering [3]:

Clustering method: K-means

Job clustering results:

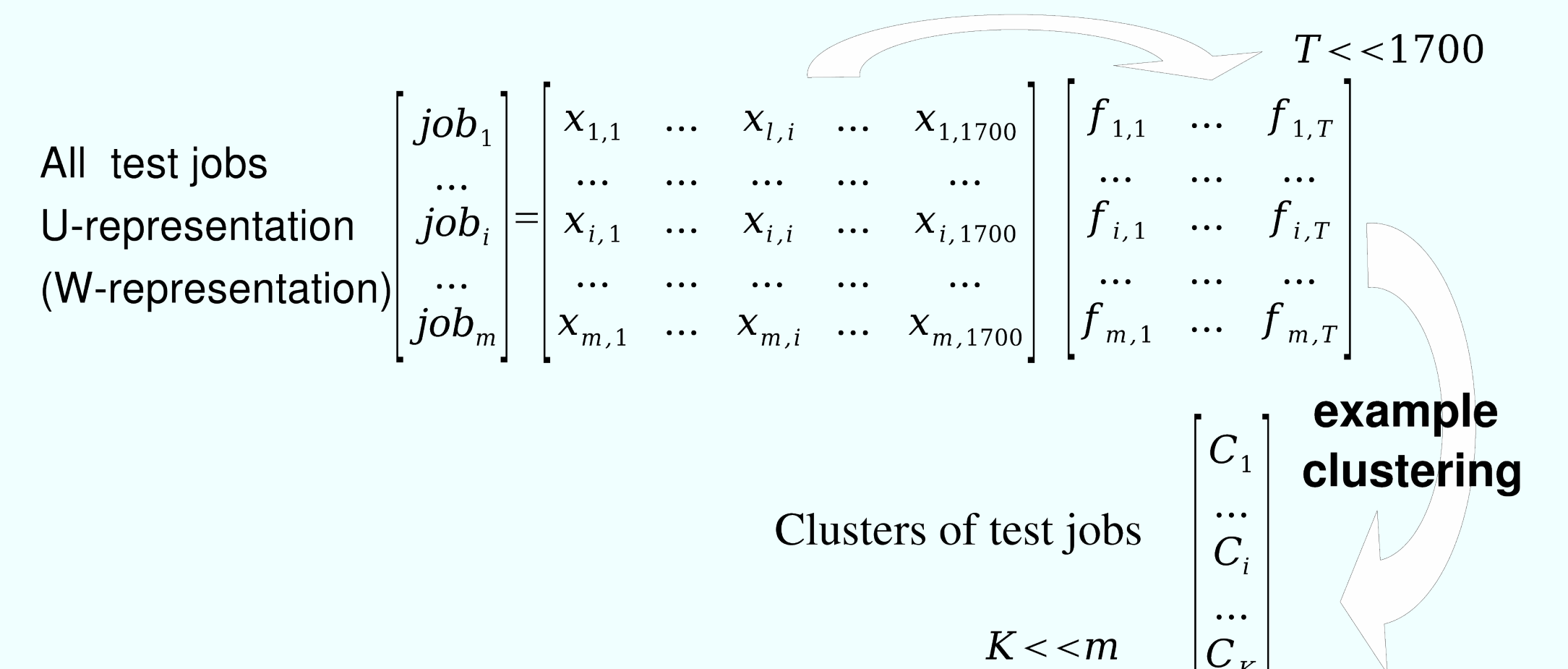
- U-clusters: from U-representations
- W-clusters: from W-representations

ROGER: Double Clustering

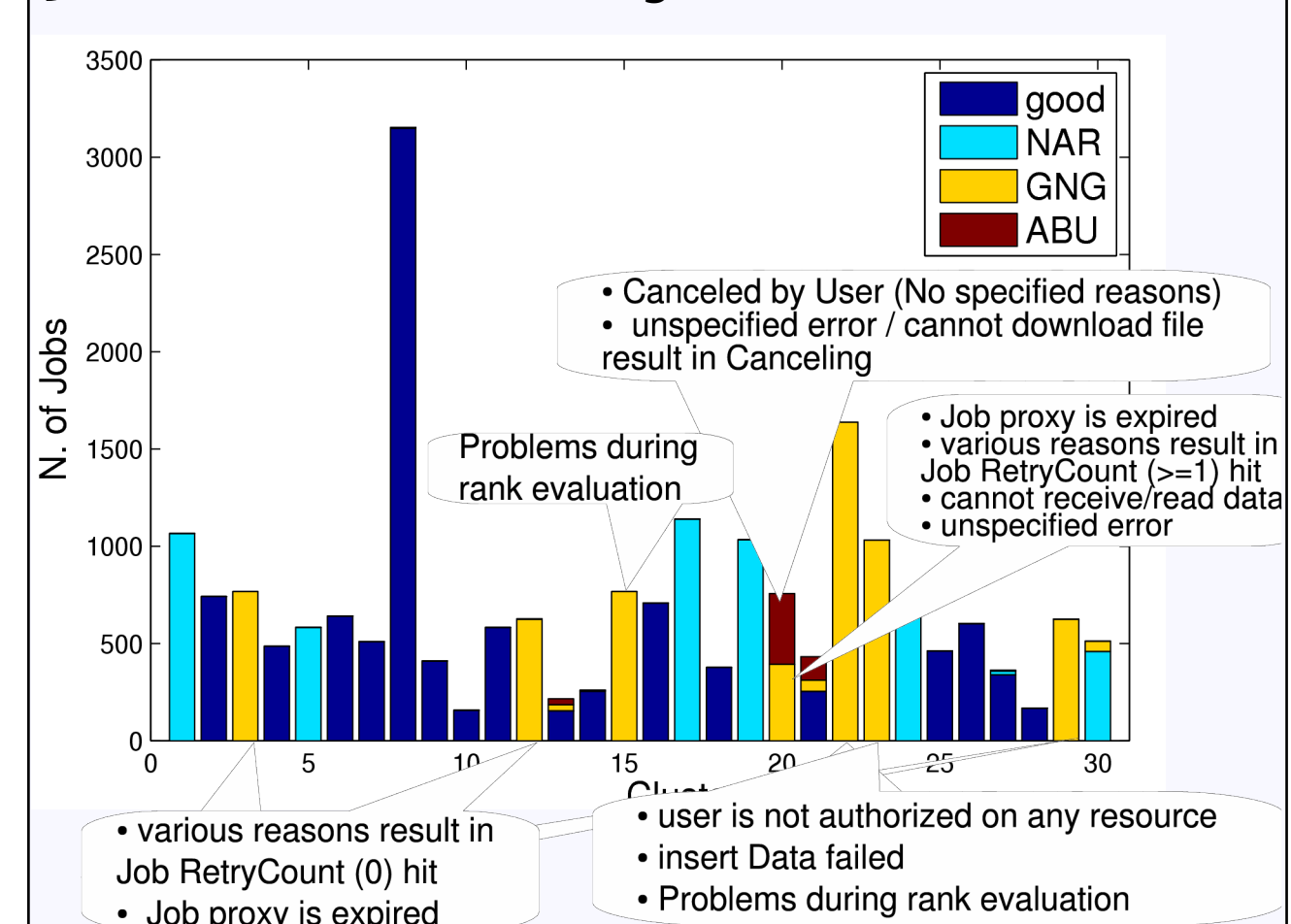
SVM: Only Clustering on jobs

\* **Note:** U-clusters are not clusters of users  
W-clusters are not clusters of weeks

## feature clustering (Dimensionality reduction)



## Job Clusters: Roger learned features based



## Clustering Error Rate:

### Purity of the clusters

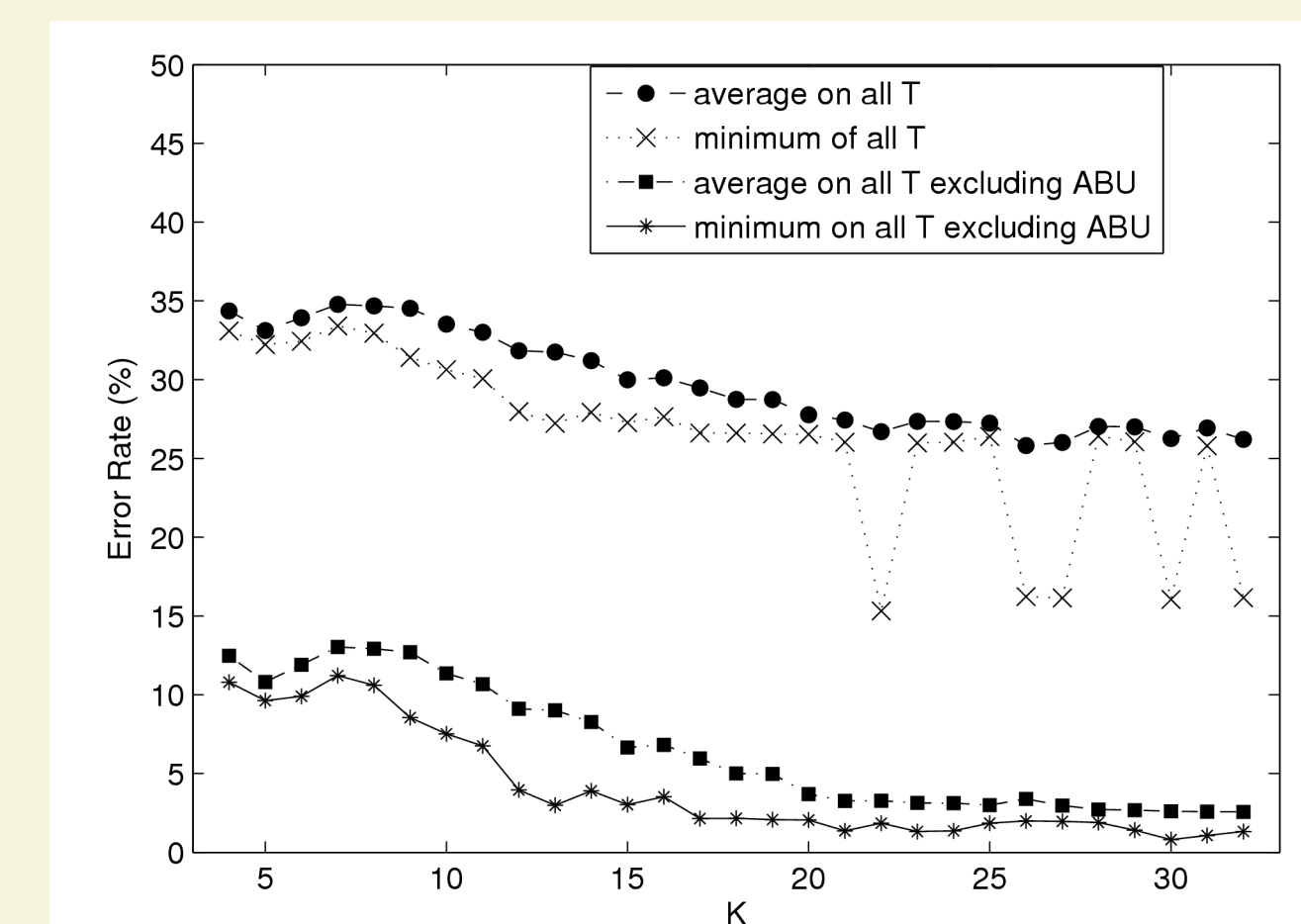
- errors: all jobs which do not belong to the majority class of the clusters they are in.
- error rate =  $\frac{((e^{rot} \text{good} \wedge \text{good}) + (e^{rot} \text{NAR} \wedge \text{NAR}) + (e^{rot} \text{GNG} \wedge \text{GNG}) + (e^{rot} \text{ABU} \wedge \text{ABU}))}{4}$  (because sizes of rough classes are imbalanced)

### Performance:

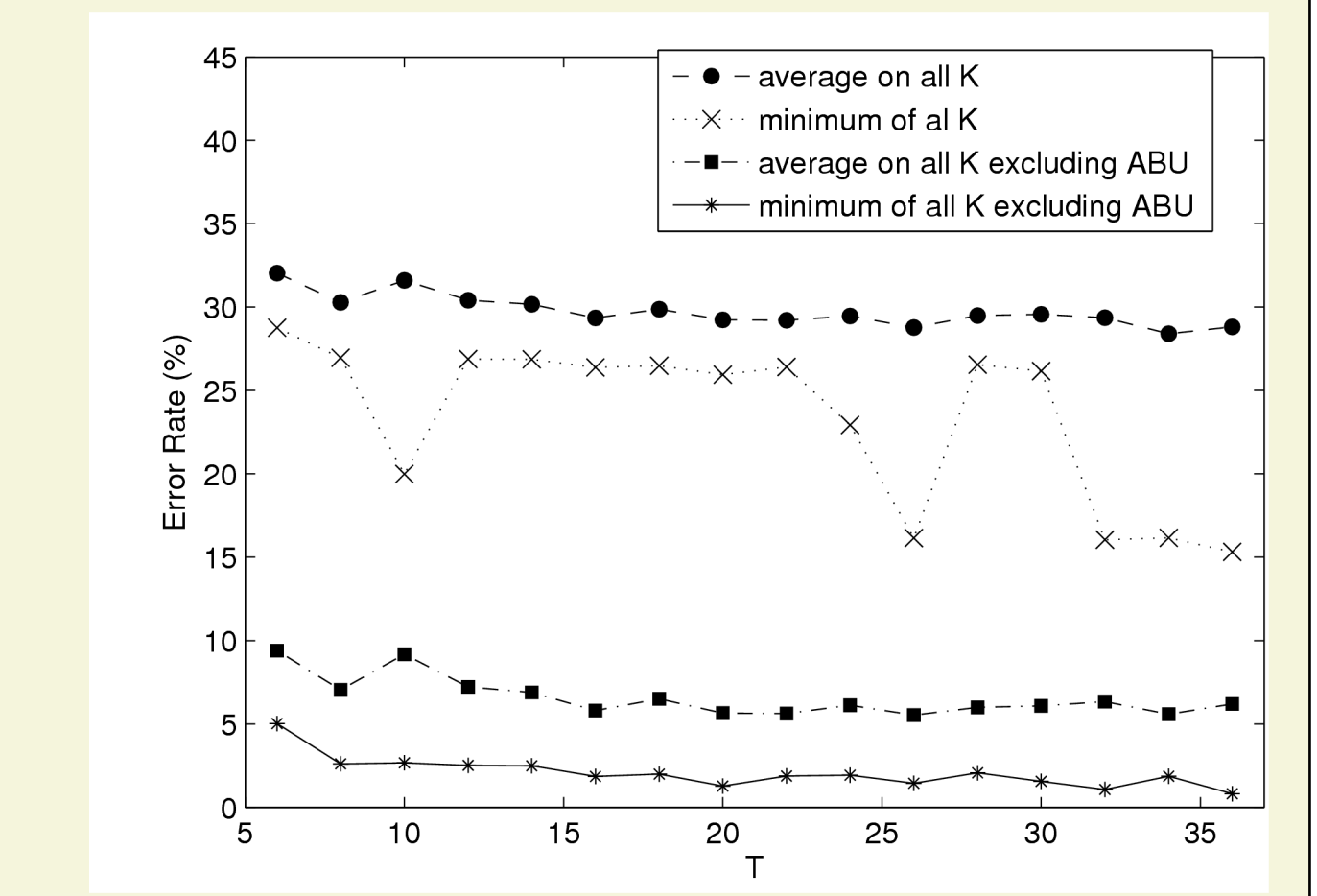
- U-clusters and W-clusters are similar
- SVM and ROGER are similar

K -- the number of example clusters  
T -- the number of feature clusters

## Roger performance: W-clusters error rate vs K



## Roger performance: W-clusters error rate vs T



## Clustering Stability [4]:

- A clustering C represented by matrix  $\hat{C} = (C_1, \dots, C_K) \in \mathbb{R}^{m \times K}$
- $\hat{C}_{ik} = \begin{cases} 1/\sqrt{n_k} & \text{if the } i^{\text{th}} \text{ example belongs to } C_k \\ 0 & \text{otherwise} \end{cases}$
- where  $n_k$  is the size of  $C_k$ ,  $\sum_k n_k = m$
- Stability of two clustering ( $\hat{C}$  and  $\hat{C}'$ )
- $S(\hat{C}, \hat{C}') = \|\hat{C}^T \hat{C}'\|_{Frobenius}^2 = \sum_{i,j=1}^K n_{i,j}^2 \frac{1}{n_i n_j}$
- where  $n_{i,j}$  is the number of jobs in  $C_i \cap C'_j$ ,  $n_i$  and  $n'_j$  are size of  $C_i$  and  $C'_j$
- when  $K \ll m$ ,  $S(\hat{C}, \hat{C}') \rightarrow 1/K$
- Stability index  $D(\hat{C}, \hat{C}') = S(\hat{C}, \hat{C}')/K$

## Self-stability: Both for W-clusters and U-clusters

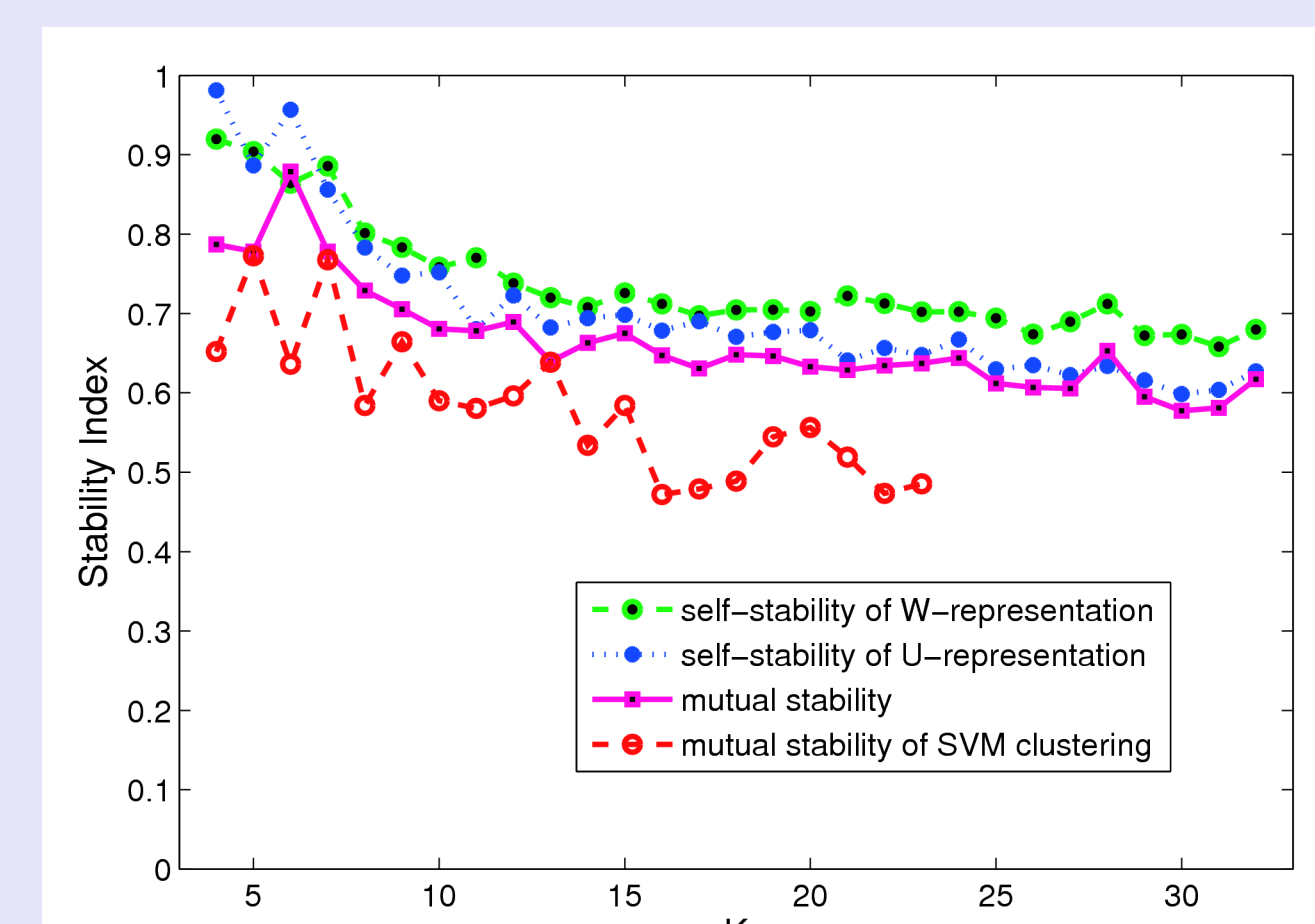
Compute with same K, average on all different pairs of T

## Mutual-stability: Between W-clusters and U-clusters

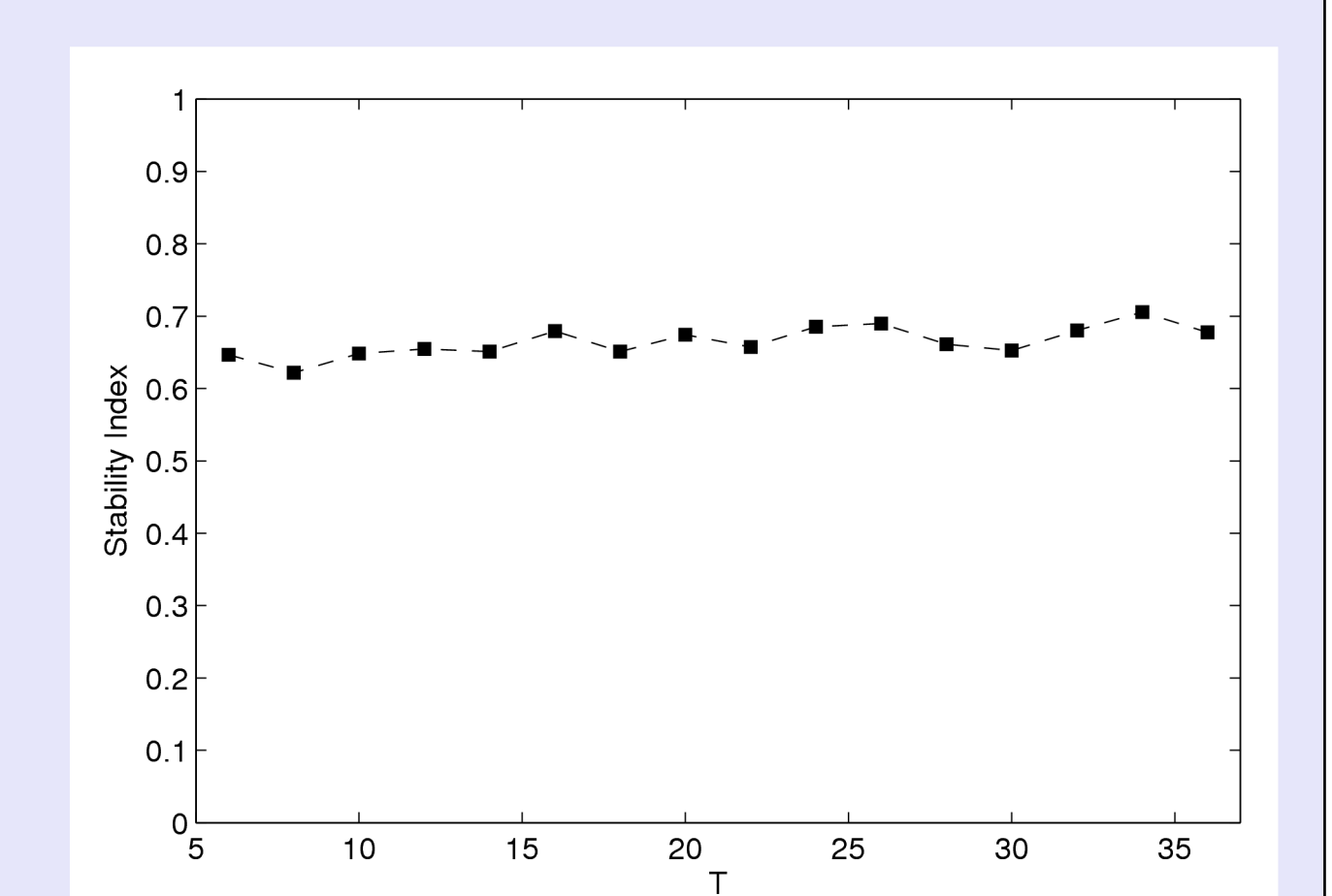
for given K, average on all pairs of W- and U-clusters with same T

for given T, average on all pairs of W- and U-clusters with same K

## Clustering Stability versus K



## Clustering Mutual Stability versus T (FIG#)



## Conclusions:

- ▶ Re-description the data
  - sampling the data by two different protocols removes the heterogeneity
  - learn new features and get two new representations
- ▶ Stable clustering
  - feature clustering (dimensionality reduction) does not significantly affect clustering stability (FIG#)
  - finds finer subclasses of grid jobs, identify classes unknown to learning algorithm (NAR, ABU, GNG)

## Perspectives:

- ▶ next: Modelling the behaviours of grid system
- ▶ further: Self-healing (detect, diagnose and repair problems) grid

[1] Kearns M., Li M.: Learning in the Presence of Malicious Errors. SIAM J. Comput. 22 (1993)  
[2] Sebag, M., Lucas, N., Az' e, J.: Impact studies and sensitivit data mining with ROC-based genetic learning. ICDM 2003  
[3] Slonim N., Tishby N. Document clustering using word clusters via the information bottleneck method. Research and Development in Information Retrieval. (2000)  
[4] Meila M. The uniqueness of a good optimum for K-means. ICML 2006