Ethnicity Sensitive Author Disambiguation Using Semi-supervised Learning

Gilles Louppe, Hussein Al-Natsheh, Mateusz Susik
and Eamonn Maguire

KESW 2016, Prague

September 23, 2016

How would you fare ?

The MAJORANA DEMONSTRATOR: A Search for Neutrinoless Double-beta Decay of Germanium-76

E.W. Hoppe (PNL, Richland) , M. Horton, S. Howard (South Dakota Sch. Mines Tech.) , M.A. Howe (North Carolina U. & TUNL, Durham) , R.A. Johnson (Washington U., Seattle) , K.J. Keeter (Black

Sep 2011 - 3 pages

AIP Conf.Proc. 1441 (2012) 480-482
DOI: 10.1063/1.3700592
To appear in the proceedings of Conference: C11-07-24 Proceedings
e-Print: arXiv:1109.1567 [nucl-ex] | PDF

Constraining muon internal bremsstrahlung as a contribution to the MiniBooNE low energy excess

Finley, B.T. Fleming, R. Ford, F.G. Garcia, G.T. Garvey, C. Green, J.A. Green, T.L. Hart, E. Hawker, R. Imlay, R.A. Johnson, P. Kasper, T. Katori, T. Kobilarcik, I. Kourbanis, S.

Oct 2007 - 3 pages

FERMILAB-PUB-07-559-E
e-Print: arXiv:0710.3897 [hep-ex] | PDF
Experiment: FNAL-E-0898

How would you fare ?

The MAJORANA DEMONSTRATOR: A Search for Neutrinoless Double-beta Decay of Germanium-76

E.W. Hoppe (PNL, Richland) , M. Horton, S. Howard (South Dakota Sch. Mines Tech.) , M.A. Howe (North Carolina U. & TUNL, Durham) , R.A. Johnson (Washington U., Seattle) , K.J. Keeter (Black

Sep 2011 - 3 pages

AIP Conf.Proc. 1441 (2012) 480-482
DOI: 10.1063/1.3700592
To appear in the proceedings of Conference: C11-07-24 Proceedings
e-Print: arXiv:1109.1567 [nucl-ex] | PDF

Constraining muon internal bremsstrahlung as a contribution to the MiniBooNE low energy excess

Finley, B.T. Fleming, R. Ford, F.G. Garcia, G.T. Garvey, C. Green, J.A. Green, T.L. Hart, E. Hawker, R. Imlay, R.A. Johnson, P. Kasper, T. Katori, T. Kobilarcik, I. Kourbanis, S.

Oct 2007 - 3 pages

FERMILAB-PUB-07-559-E
e-Print: arXiv:0710.3897 [hep-ex] | PDF
Experiment: FNAL-E-0898

✗ Different authors

How would you fare ?

Effects of Limited Calorimeter Coverage on ET

Frank E. Paige, A.V. Vanyashin (SSCL)

Mar 1992 - 9 pages

Search for single b^*-quark production with the ATLAS detector at $\sqrt{s} = 7$ TeV

Planck Inst.) , Wainer Vandelli (CERN) , Alexandre Vaniachine (Argonne) , Peter Vankov (DESY) , Francois Vannucci (Paris U., VI-VII) , Riccardo Vari (INFN, Rome) , Erich Varnes (Arizona U.) ,

Jan 2013 - 11 pages

Phys.Lett. B721 (2013) 171-189
(2013-04-25)
DOI: 10.1016/j.physletb.2013.03.016
CERN-PH-EP-2012-344
e-Print: arXiv:1301.1583 [hep-ex] | PDF
Experiment: CERN-LHC-ATLAS

How would you fare ?

Effects of Limited Calorimeter Coverage on ET

Frank E. Paige, A.V. Vanyashin (SSCL)

Mar 1992 - 9 pages

Search for single b^*-quark production with the ATLAS detector at $\sqrt{s} = 7$ TeV

Planck Inst.) , Wainer Vandelli (CERN) , Alexandre Vaniachine (Argonne) , Peter Vankov (DESY) , Francois Vannucci (Paris U., VI-VII) , Riccardo Vari (INFN, Rome) , Erich Varnes (Arizona U.) ,

Jan 2013 - 11 pages

Phys.Lett. B721 (2013) 171-189
(2013-04-25)
DOI: 10.1016/j.physletb.2013.03.016
CERN-PH-EP-2012-344
e-Print: arXiv:1301.1583 [hep-ex] | PDF
Experiment: CERN-LHC-ATLAS

✓ Same authors

Homonymy in Asian Names Written in English

Please meet Yang Wang, Wang Yang, and Yang Wang!

Author Disambiguation as Entity Resolution Problem

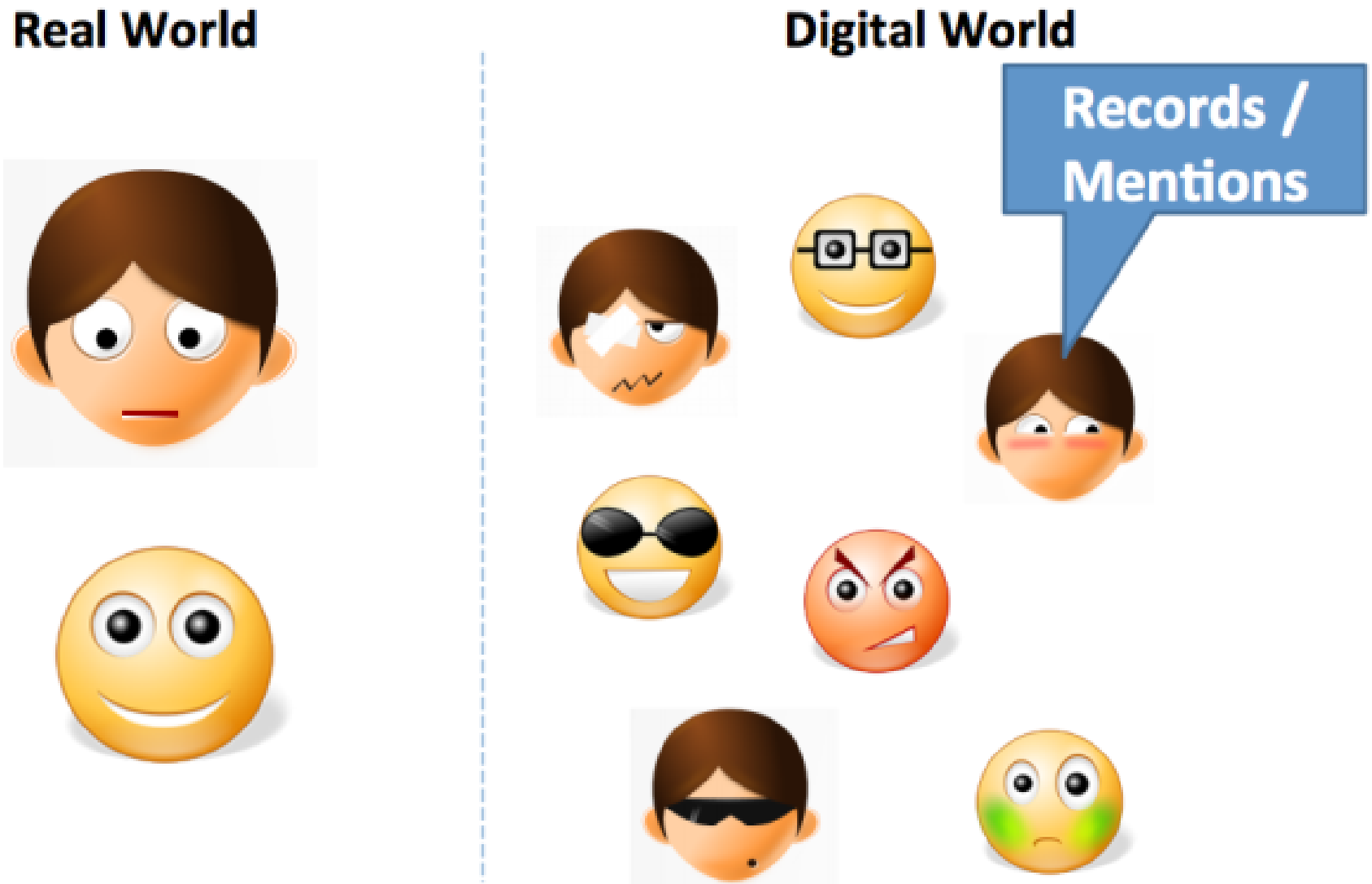


Image source : datacommunitydc.org

Definitions

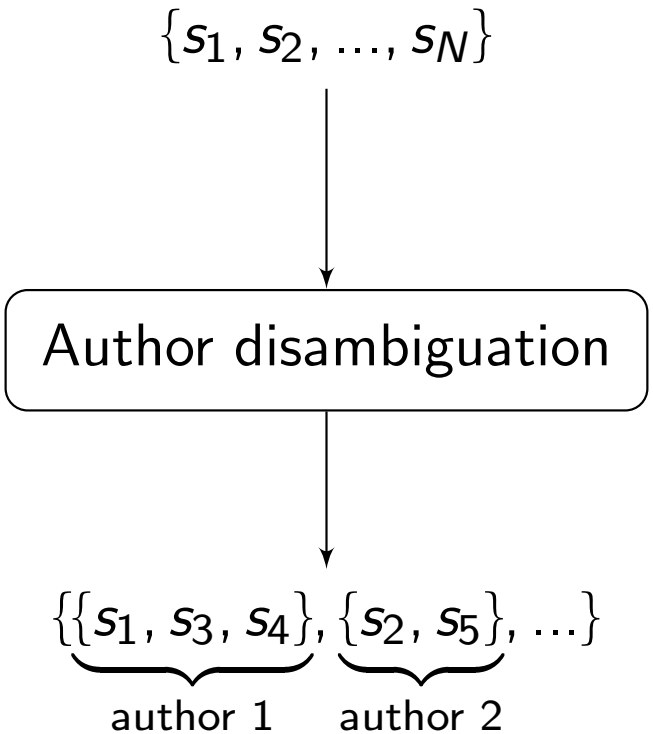
Problem to Solve

For each author, group together all his publications, and only those.

Inspirehep.net is a digital library contains

- Over 1M publication forming more than 10M signatures
- 1.2M signatures are claimed by :
  - Authors themselves (similar to Google Scholar).
  - Universal identifiers (ORCiD) .
  - Professional curators .

Problem Formulation



Learning from data

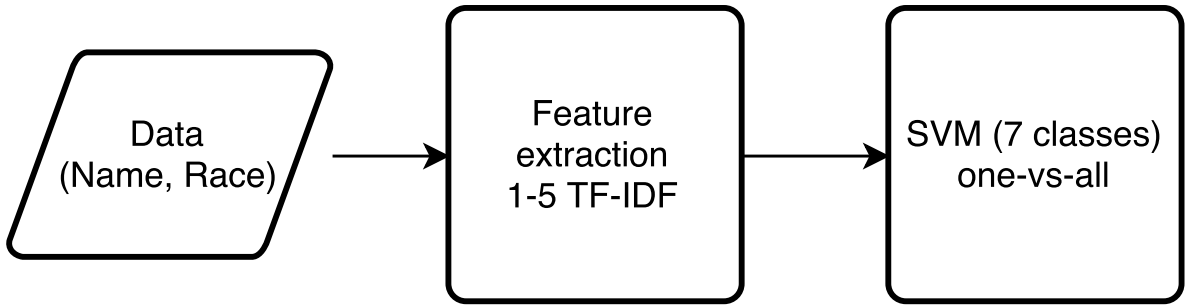
- Manual disambiguation is long and difficult, even for experienced curators.

- Couldn't we automatically find a set of rules to disambiguate two signatures ?

$$\varphi(s_1, s_2) = \begin{cases} 0 & \text{if } s_1 \text{ and } s_2 \text{ belong to the same author,} \\ 1 & \text{otherwise.} \end{cases}$$

- This is a machine learning task called supervised learning.

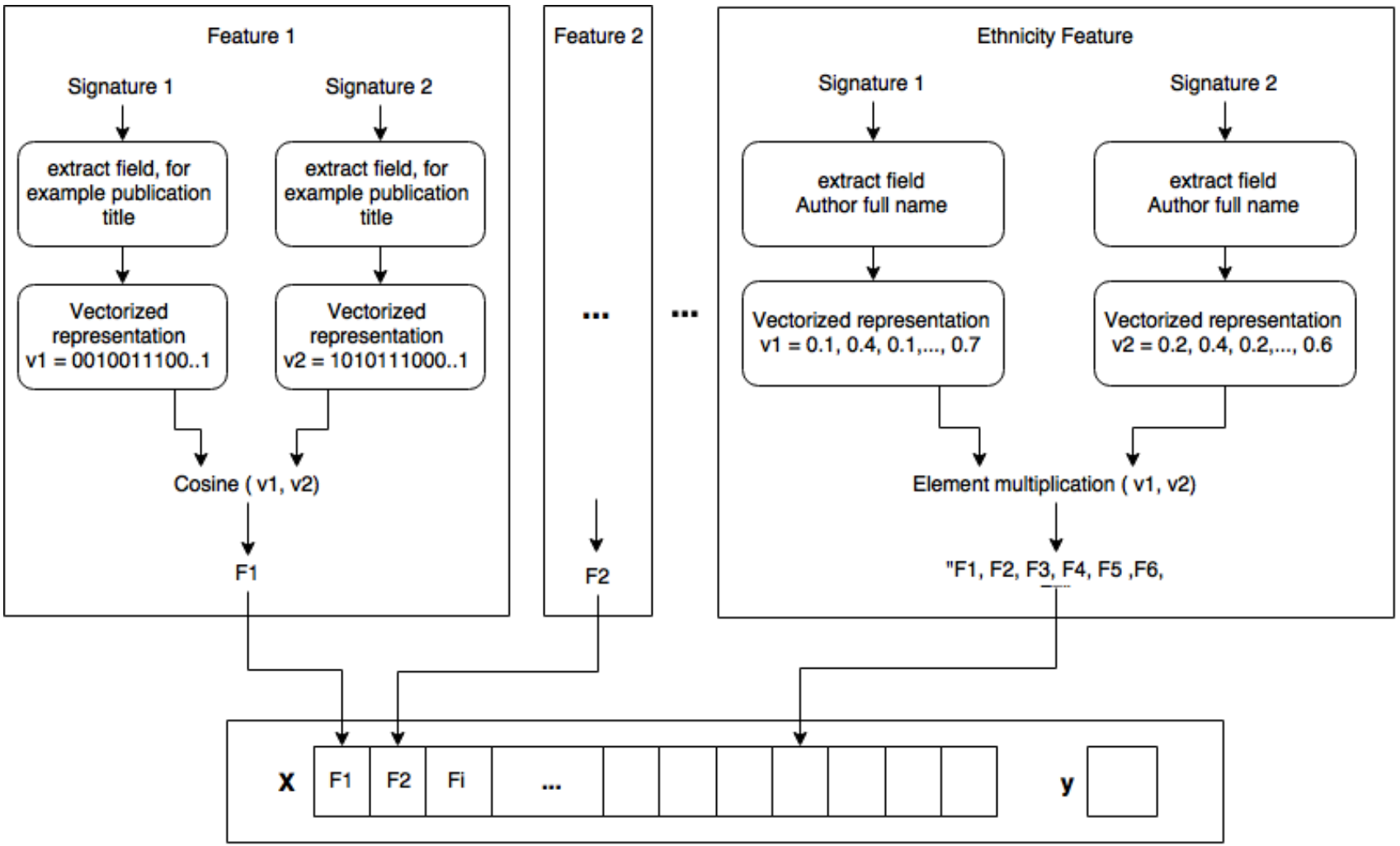
Ethnicity Features



From (IPUMS-USA) we extracted :

- White : 20M
- Black : 3M
- American Indian or Alaska Native : 150K
- Chinese : 50K
- Japanese : 50K
- Other Asian or Pacific Islander : 30K
- Other race : 1K

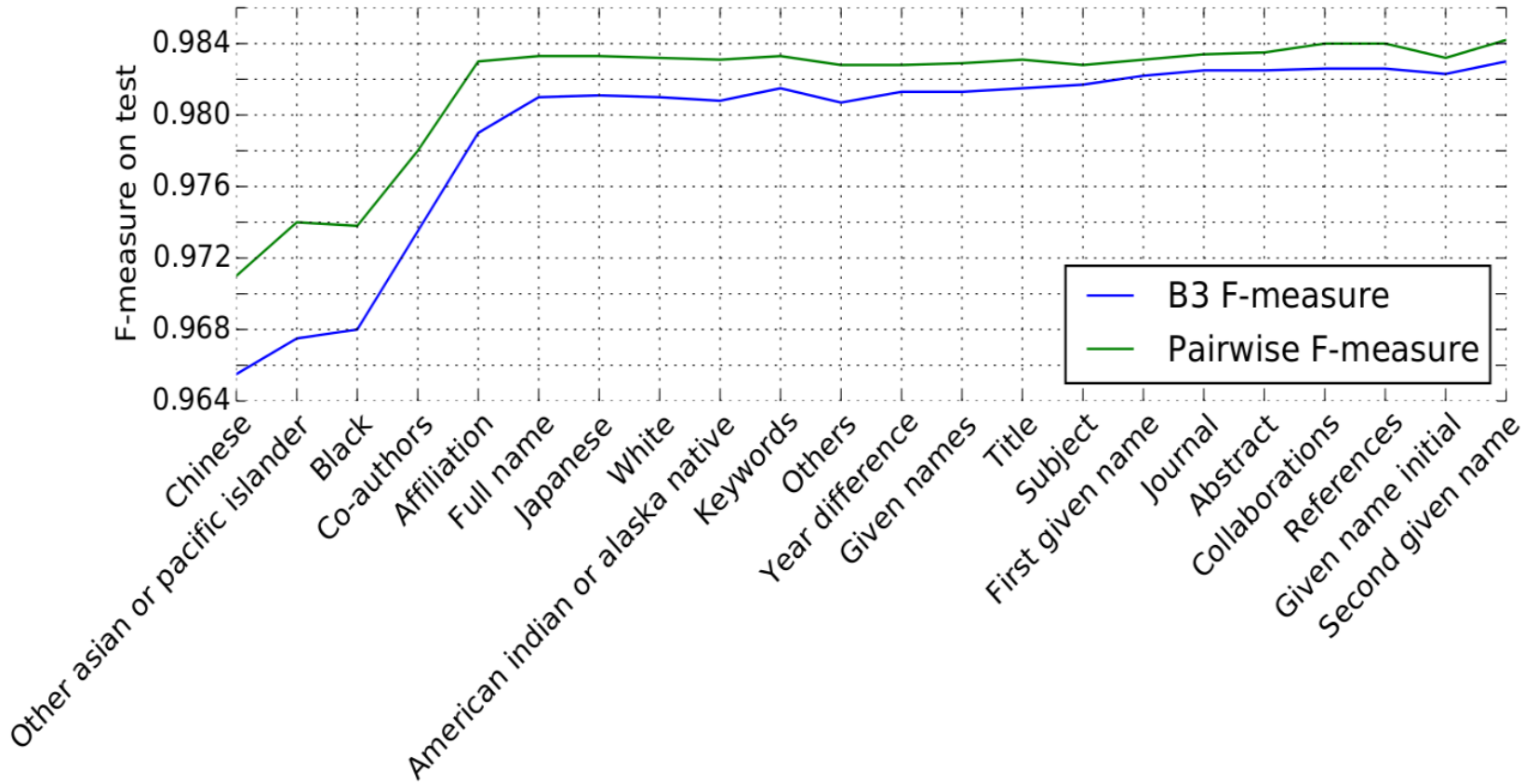
Pair-wise Features Extraction



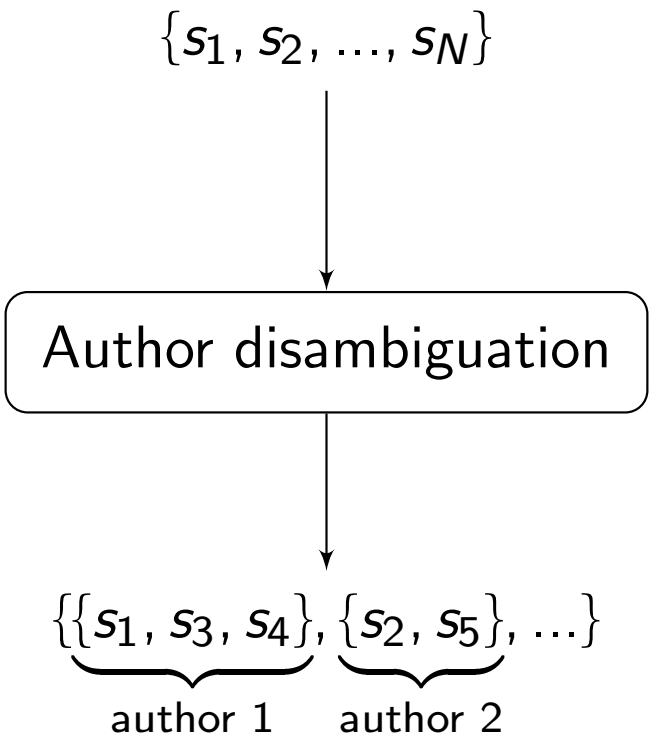
Features set

Feature	**Combination operator**
Full name	Cosine similarity of $(2, 4)$-TF-IDF
Given names	Cosine similarity of $(2, 4)$-TF-IDF
First given name	Jaro-Winkler distance
Second given name	Jaro-Winkler distance
Given name initial	Equality
Affiliation	Cosine similarity of $(2, 4)$-TF-IDF
Co-authors	Cosine similarity of TF-IDF
Title	Cosine similarity of $(2, 4)$-TF-IDF
Journal	Cosine similarity of $(2, 4)$-TF-IDF
Abstract	Cosine similarity of TF-IDF
Keywords	Cosine similarity of TF-IDF
Collaborations	Cosine similarity of TF-IDF
References	Cosine similarity of TF-IDF
Subject	Cosine similarity of TF-IDF
Year difference	Absolute difference
Any ethnicity feature	Product of probabilities estimated by SVM

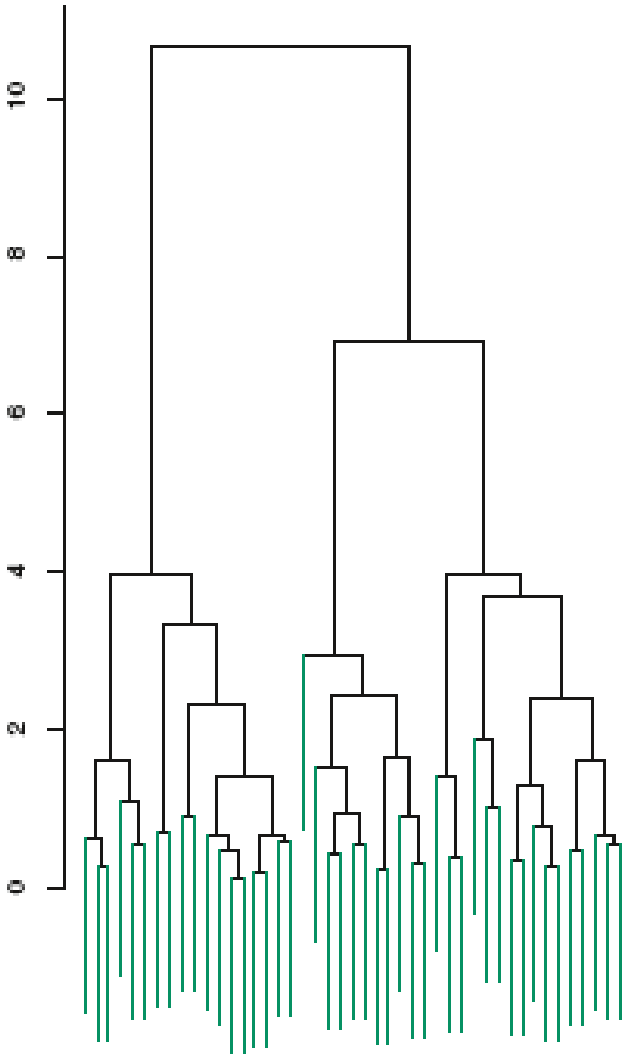
Feature Importances by Recursive Elimination



Disambiguation as a clustering problem



Hierarchical Clustering



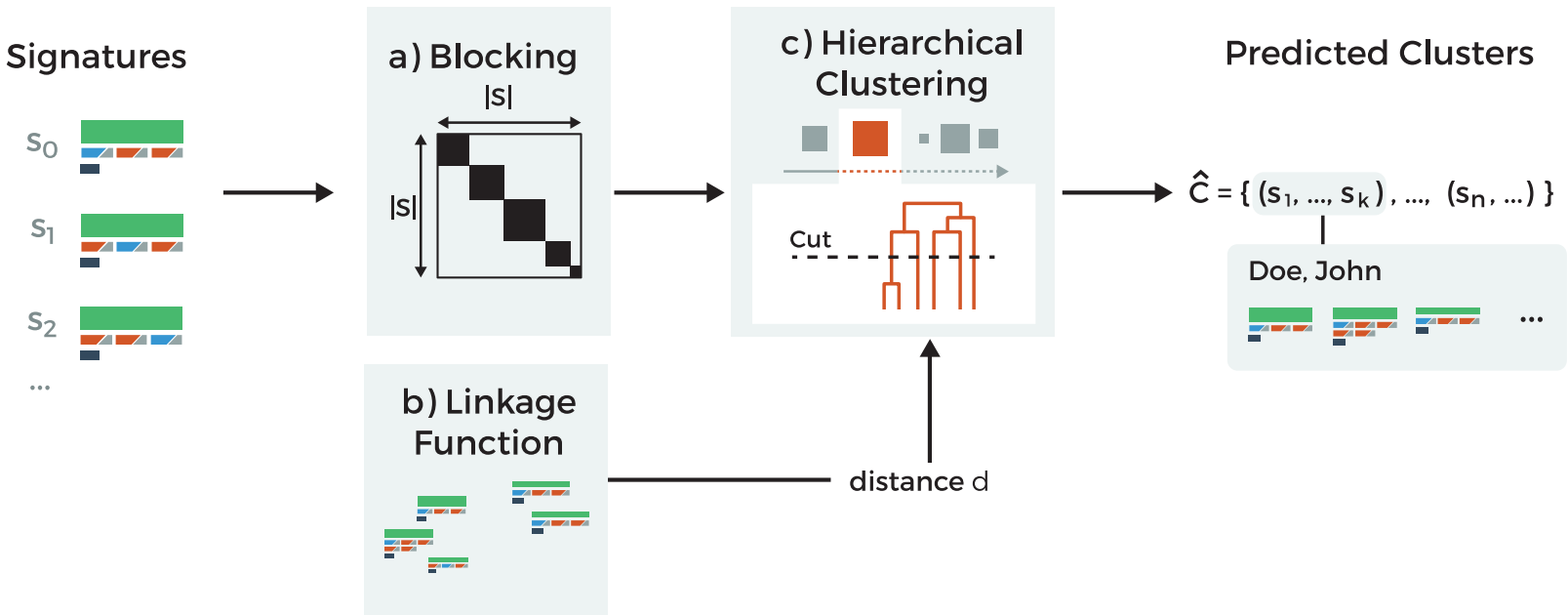
- General family of clustering algorithms that build nested clusters by merging them successively.
- This hierarchy of clusters is represented as a tree (or dendrogram).
- The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample.

Clustering Issues

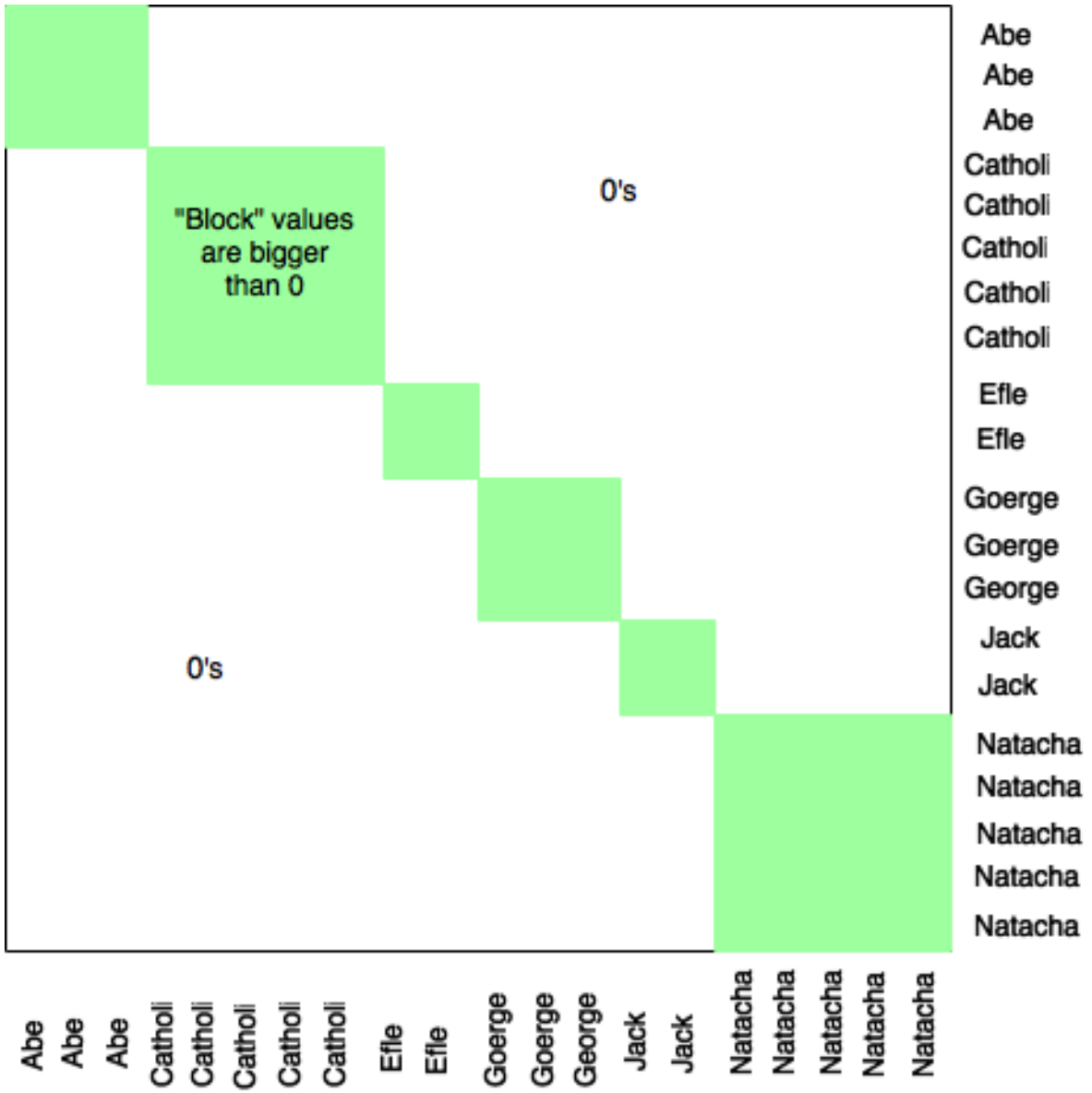
- The complexity of hierarchical clustering is $O(N^2)$. For $N = 10^7$ signatures, this is impractical.
  Solution : partitioning into blocks all signatures with the same last name + first initial, then cluster each of these blocks.

- How do you set the cut-off threshold ?
  Solution : using training data (e.g., claimed signatures), pick the threshold that locally maximizes some criterion.

General Pipeline



Solving Issue 1 : Partitioning into Blocks

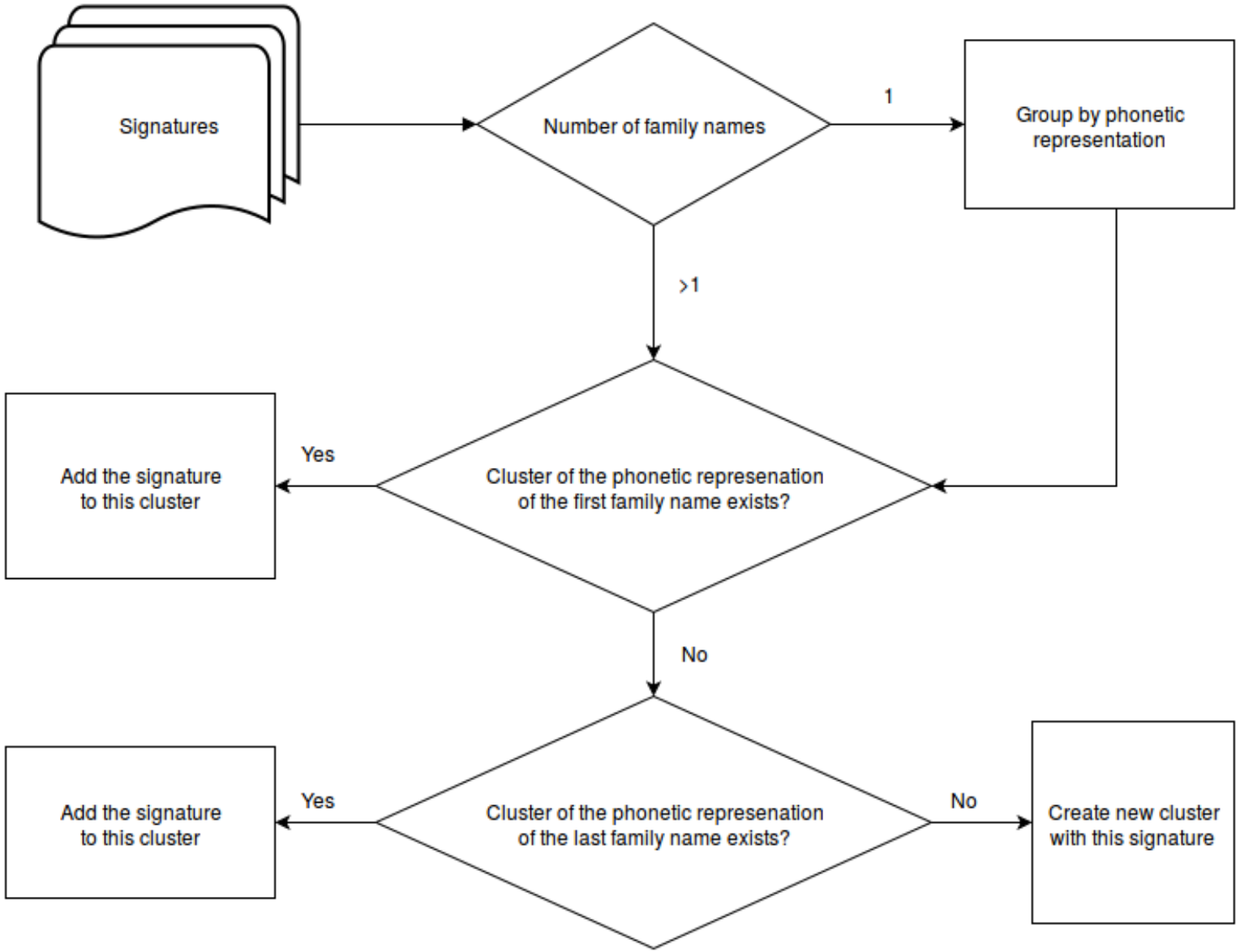


Cases

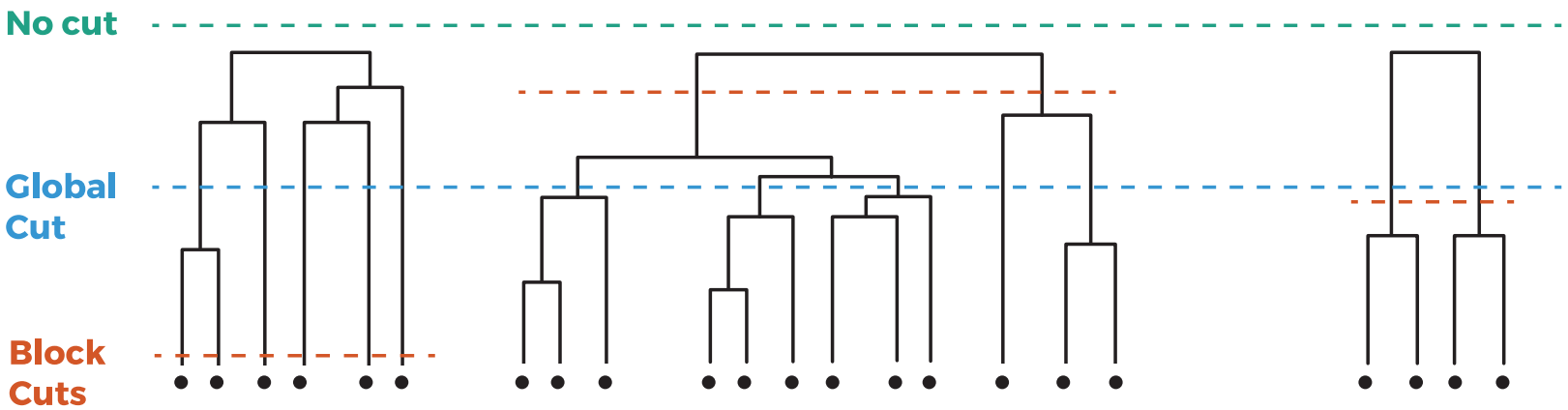
Analysis of data showed that author can have multiple names in few cases :

- "Mueller, R." and "Muller, R."
- "Martinez Torres, A." and "Torres, A. Martinez"
- "Smith-Jones, A." and "Smith, A."
- "Smith, Jack" and "Smith, A. J."
- An authors surname changed (e.g., due to marriage).

Solution


Signatures
Number of family names
1
Group by phonetic representation
>1
Add the signature to this cluster
Yes
Cluster of the phonetic represenation of the first family name exists?
No
Add the signature to this cluster
Yes
Cluster of the phonetic represenation of the last family name exists?
No
Create new cluster with this signature

Solving Issue 2 : Threshold Cut-off Strategy



Evaluation

Protocol : Use the claimed signatures (about 1M) to form ground truth clusters. Keep 10% as a training set to find model parameters, and 90% as a test set for evaluation.

$$B^3 \text{ Precision} = \mathbb{E}_s\{\frac{|\hat{C}(s) \cap C(s)|}{|\hat{C}(s)|}\} \tag{1}$$

$$B^3 \text{ Recall} = \mathbb{E}_s\{\frac{|\hat{C}(s) \cap C(s)|}{|C(s)|}\} \tag{2}$$

$$B^3 \text{ F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

where $C(s)$ (resp., $\hat{C}(s)$) is the true (resp., predicted) set of signatures to which s belongs.

Results 1 of 2

Description	B^3 *Prec.*	B^3 *Recall*	B^3 *F1*
Baseline	0.9024	0.9828	0.9409
Blocking = Surname & First Initial	0.9901	0.9760	0.9830
Blocking = Double metaphone	0.9856	0.9827	0.9841
Blocking = NYSIIS	0.9875	0.9826	**0.9850**
Blocking = Soundex	0.9886	0.9745	0.9815
Classifier = Gradient Boosting Classifier	0.9901	0.9760	0.9830
Classifier = Random Forests	0.9909	0.9783	**0.9846**
Classifier = Linear Regression	0.9749	0.9584	0.9666
Training pairs = Non-blocked, uniform	0.9793	0.9630	0.9711
Training pairs = Blocked, uniform	0.9854	0.9720	0.9786
Training pairs = Blocked, balanced	0.9901	0.9760	**0.9830**

Results 2 of 2

Description	B^3 *Prec.*	B^3 *Recall*	B^3 *F*1
Baseline	0.9024	0.9828	0.9409
Clustering = Average linkage	0.9901	0.9760	**0.9830**
Clustering = Single linkage	0.9741	0.9603	0.9671
Clustering = Complete linkage	0.9862	0.9709	0.9785
No cut (baseline)	0.9024	0.9828	0.9409
Global cut	0.9892	0.9737	0.9814
Block cut	0.9901	0.9760	**0.9830**
Combined best settings	0.9888	0.9848	**0.9868**
Best settings without ethnicity features	0.9862	0.9819	0.9841

Summary Results

Method	B^3 *F-score*
Full name	0.8183
Last name + First initial	0.9409
Our model	0.9868

Implementation

The solution is currently being used by the INSPIRE and INVENIO projects at CERN.

Execution time : 20 hours for 10M signatures, on a 16 cores machine with 32GB of RAM.
But, even only few minutes for incremental disambiguation !

Our solution is open-source [1] and we released the dataset [2].

1. github.com/inspirehep/beard
2. github.com/glouppe/paper-author-disambiguation/data

Conclusions

- Semi-supervised approach on the biggest dataset ever used for author disambiguation.
- Novel blocking technique based on phonetization.
- Showing the significancy of inferred name ethnicity.
- Showing the importance of balancing the training set.

Future Work

- Error analysis.
- Build or find more comprehensive name-ethnicity dataset.
- Explore author embedding approaches as a blocking strategy.
- Build phonetic algorithm tailored to the disambiguation task.
- Archive and utilize user's feedback to enhance the model.
- Try our disambiguation solution for other tasks.