
Unsupervised Phenotype Scoring

Abstract

Unsupervised phenotyping is about clustering patients into homogeneous groups (“phenotypes”) based on observed clinical features. An open challenge is to provide a principled way to define those patient phenotypes in a simple and clinically meaningful manner. In this work, we tackle this challenge by proposing a nonnegative matrix factorization approach extracting phenotypes as a set of integer-weighted features, where the weights are constrained to take values from a small integer set. We employ a block alternating update scheme to solve the problem and formulate the problem of extracting the phenotype weights as an integer programming one. As compared to a commonly used scale-and-round heuristic, our approach achieves consistently higher fit in 2 real electronic health record (EHR) datasets. Our proposed interpretation as well as representative scoring-based phenotypes were validated by a medical expert.

1 Problem Motivation

Phenotyping refers to the process of identifying sets of measurable disease markers, enabling to perform cohort identification, i.e., separate between cases and control patients for several diseases or disease sub-types of interest [32, 27]. Unsupervised phenotyping is the process of discovering the underlying patient groups or phenotypes directly based on their EHR data. Matrix and tensor factorization has been a very successful machinery for unsupervised EHR-based phenotyping (e.g., [11, 12, 31, 25, 15, 26, 10]). Consider for example the problem setup of Nonnegative Matrix Factorization (NMF) [21] when minimizing the squared Frobenius norm of the error:

$$\min_{\mathbf{U} \geq 0, \mathbf{V} \geq 0} \|\mathbf{X} - \mathbf{UV}^T\|_F^2 \quad (1)$$

In the context of unsupervised phenotyping, $\mathbf{X} \in \mathbb{R}^{M \times N}$ is a non-negative input matrix, whose $\mathbf{X}(i, j)$ cell reflects the activity recorded (e.g., event counts) for the i -th (out of M) patient with respect to the j -th (out of N) medical features. Given an input number R of desired phenotypes, the matrix $\mathbf{U} \in \mathbb{R}^{M \times R}$ corresponds to a membership matrix of the patients with respect to the R phenotypes. Finally, the matrix $\mathbf{V} \in \mathbb{R}^{N \times R}$ provides the phenotypes’ definition: the non-zero elements of the r -th column $\mathbf{V}(:, r)$ reveal the relevant medical features to the r -th phenotype. However, one challenge is to clinically interpret those phenotypes as the values in \mathbf{V} can be large and hard to interpret. This paper tackles this challenge of improving the interpretability of factorization-based unsupervised phenotyping approaches.

There do exist several domains where an application of Problem 1 is largely focusing on achieving a low approximation error; for example, in recommender systems, good prediction power is usually the top priority [18]. On the contrary, in the context of Unsupervised phenotyping, model interpretability is equally (or even more) crucial. More specifically, it is imperative to provide the medical expert with a straightforward protocol of translating the phenotypes’ definition factor \mathbf{V} . We argue that this task is quite challenging if \mathbf{V} contains (nonnegative) real values, to the point that those values are hidden altogether by the expert and only the feature ranking produced is preserved (e.g., as happens in [12, 31]). A primary reason for the above is that medical experts are used to deal with simple and concise scoring-based descriptions of clinical status (e.g., risk scores¹) [30]. The main intuition behind the development of such scores is simple: the range of the possible values revealing the contribution of each clinical aspect is restricted to a small integer set of consecutive values (e.g., $\{0, 1, 2, 3\}$). A value of zero indicates no contribution of the feature and higher scores indicate *distinct levels of severity or importance*.

¹In MDCalc, one can find a vast amount of such scores used in medicine.

2 Proposed Approach

Methodology The problem under consideration can be formally defined as follows:

$$\min_{\mathbf{U} \geq 0, \mathbf{V} \in \mathbb{Z}_\tau} \|\mathbf{X} - \mathbf{U}\mathbf{V}^T\|_F^2 \quad (2)$$

where $\mathbb{Z}_\tau = \{0, 1, \dots, \tau\}$ and τ is a positive integer. We employ a block alternating update scheme to tackle the above non-convex optimization problem. This is based on a variable partitioning into several disjoint subgroups and iteratively minimizing the objective function w.r.t. one subgroup at a time.² There are several ways of choosing this partitioning that lead to different NMF algorithms [16]. We choose to employ one that has shown to provide state-of-the-art performance [6] and at the same time, can lead to an optimal update of the discrete factor's columns ($\mathbf{V}(:, j), 1 \leq j \leq R$) in an efficient manner.

The framework employed is known as the Hierarchical Alternating Least Squares (HALS) [4, 8] (aka rank-one residue iteration [13]). The main idea behind this framework is the partitioning of the unknown matrices into $2R$ blocks, corresponding to the columns of matrices \mathbf{U} and \mathbf{V} . We first review how NMF-HALS decomposes Problem (2) when solving for each k -th nonnegative column of \mathbf{U} (and \mathbf{V} is fixed):

$$\min_{\mathbf{U}(:, k) \geq 0} \|\mathbf{X} - \sum_{r=1}^R \mathbf{U}(:, r)\mathbf{V}(:, r)^T\|_F^2 = \min_{\mathbf{U}(:, k) \geq 0} \|\mathbf{X} - \underbrace{\sum_{r=1, r \neq k}^R \mathbf{U}(:, r)\mathbf{V}(:, r)^T}_{\mathbf{R}_k} - \mathbf{U}(:, k)\mathbf{V}(:, k)^T\|_F^2$$

where \mathbf{R}_k corresponds to the ‘‘residual matrix’’ and is fixed for the k -th iteration. It is easy to show [16] that an optimal solution to the above problem is given by:

$$\mathbf{U}(:, k) = \frac{\max(0, \mathbf{R}_k \mathbf{V}(:, k))}{\|\mathbf{V}(:, k)\|_2^2} \quad (3)$$

In symmetry, we solve the following problem to find the optimal $\mathbf{V}(:, k)$:

$$\min_{\mathbf{V}(:, k) \in \mathbb{Z}_\tau} \|\mathbf{R}_k - \mathbf{U}(:, k)\mathbf{V}(:, k)^T\|_F^2 \quad (4)$$

Note that due to the discrete nature of the domain of \mathbf{V} , a brute-force solution would have to evaluate $(\tau + 1)^N$ feasible vectors, which would result in a prohibitively expensive algorithm. A useful remark we exploit is that Problem (4) can be equivalently formulated as a sum of minimizers of N independent sub-problems, corresponding to the N coordinates of $\mathbf{V}(:, k)$ and the associated columns of \mathbf{R}_k [16]:

$$\begin{aligned} & \sum_{j=1}^N \min_{\mathbf{v}(j, k) \in \mathbb{Z}_\tau} \|\mathbf{R}_k(:, j) - \mathbf{V}(j, k)\mathbf{U}(:, k)\|_2^2 \\ &= \sum_{j=1}^N \min_{\mathbf{v}(j, k) \in \mathbb{Z}_\tau} \left(\|\mathbf{R}_k(:, j)\|_2^2 - 2\mathbf{V}(j, k)\mathbf{U}(:, k)^T \mathbf{R}_k(:, j) + \mathbf{V}(j, k)^2 \|\mathbf{U}(:, k)\|_2^2 \right) \end{aligned}$$

The term $\|\mathbf{R}_k(:, j)\|_2^2$ is fixed, thus minimizing the above for each $\mathbf{V}(j, k)$ is equivalent to:

$$\min_{\mathbf{v}(j, k) \in \mathbb{Z}_\tau} \left(-2\mathbf{V}(j, k)\mathbf{U}(:, k)^T \mathbf{R}_k(:, j) + \mathbf{V}(j, k)^2 \|\mathbf{U}(:, k)\|_2^2 \right) \quad (5)$$

It is clear that for $\mathbf{V}(j, k) = 0$, the value of the above objective is equal to zero. Thus, Problem (4) can be optimally solved by plugging in the values $\mathbf{V}(j, k) = \{1, 2, \dots, \tau\}$ for each j -th coordinate of $\mathbf{V}(:, k)$ and choose the solution dictated by Problem (5). Note that the size of the search space of feasible solutions is asymptotically linear to the column size we are solving for ($\tau N = \mathcal{O}(N)$) and those evaluations are fully-parallelizable w.r.t. each of the N coordinates since they are associated with independent sub-problems. Algorithm 1 lists a pseudocode for the proposed approach. Note that in practice one should not explicitly compute the residual matrix \mathbf{R}_k ; its size is $M \times N$ and it is expected to be a dense matrix, even if the input \mathbf{X} is sparse. Instead, it is more efficient to compute the necessary products involving \mathbf{R}_k as [7, 16]: $\mathbf{R}_k \mathbf{V}(:, k) = \mathbf{X}\mathbf{V}(:, k) - \sum_{r \neq k}^R \mathbf{U}(:, r) (\mathbf{V}(:, r)^T \mathbf{V}(:, k))$ and $\mathbf{U}(:, k)^T \mathbf{R}_k = \mathbf{U}(:, k)^T \mathbf{X} - \sum_{r \neq k}^R (\mathbf{U}(:, k)^T \mathbf{U}(:, r)) \mathbf{V}(:, r)^T$.

²Similar to Block Coordinate Descent (BCD) [3], but the domain is not a closed convex set, due to the integer constraint.

Algorithm 1 Unsupervised Phenotype Scoring

Require: $\mathbf{X} \in \mathbb{R}^{M \times N}$, target rank R and upper bound τ

Ensure: $\mathbf{U} \in \mathbb{R}^{M \times R}$, $\mathbf{V} \in \mathbb{Z}_\tau^{N \times R}$

```
1: Initialize  $\mathbf{U}, \mathbf{V}$ 
2: while convergence criterion is not met do
3:   for  $k = 1, \dots, R$  do
4:     Update  $U(:, k)$  via Equation (3)
5:   end for
6:   for  $k = 1, \dots, R$  do
7:     for  $j = 1, \dots, N$  do
8:       Update  $V(j, k)$  by solving (5)           // independent sub-problems in parallel  $\forall j = 1, \dots, N$ 
9:     end for
10:  end for
11: end while
```

Given the fact that each one of the sub-problems of Algorithm 1 is *optimally* solved, it is guaranteed that the objective function (2) will be non-increasing for successive iterates [3].

Interpretation of discrete phenotypes Extensive prior art [28, 14, 2, 9, 34] reports the existence of bias in the recording of EHRs towards more severe patient conditions. In short, patients are not measured evenly, but are measured according to a clinician’s judgment, usually more frequently when the patient is sicker [14]. Thus, higher input prevalence of records (e.g., higher occurrence counts) can be interpreted as increased severity of the corresponding condition for the associated patient. The above translates to interpreting the discrete values within each phenotype factor $\mathbf{V}(:, j)$ as distinct levels of severity of each medical feature w.r.t. the j -th phenotype, where a value of 0 indicates absence and a value of τ indicates the highest possible severity.

3 Experiments

Data and Task Description We used the diagnostic, medication and procedure information from 2 real clinical datasets for our experiments: **a)** MIMIC3 is a publicly available dataset³ which contains data from patients admitted to critical care units, and **b)** a pediatric dataset from a private source⁴ In both datasets, the diagnostic and procedure codes available were transformed to Clinical Classifications Software (CCS) categories [1], a standard step to improve interpretability. Regarding the MIMIC dataset, we also transformed the medications available to USP categories⁵. We considered that a cell $\mathbf{X}(i, j)$ is a count of the times that the j -th medical feature (diagnosis, medication or procedure) has occurred during the history of the i -th patient. The resulting data sizes are: a) MIMIC: $M = 46517, N = 1063$ and the number of non-zeros is $nnz(\mathbf{X}) = 1429492$; b) Private dataset: $M = 1045, N = 862, nnz(\mathbf{X}) = 91222$.

Implementation details We used MatlabR2015b for our implementations. We set $\tau = 3$, so that based on our interpretation protocol defined in Section 2, we can interpret the possible non-zero values of the discrete factor as low (1), medium (2) and high (3) severity accordingly. We performed a fixed number (5) of inner iterations, thus lines 3-5 and 6-10 of Algorithm 1 are executed for 5 times each. The convergence criterion of the outer iteration is whether successive differences of the objective in Problem 2 drop below $1e - 6$. We set the same number of inner iterations and convergence criteria for the baselines as well, so that the methods are comparable.

Methods under comparison 1) Nonnegative matrix factorization with the HALS approach [4, 8]; **2)** Scaled-and-rounded version of NMF: Similar to [30], we use a heuristic where the solution of NMF is post-processed so that $\tilde{\mathbf{V}} \in \mathbb{Z}_\tau$. Let $\gamma(j) = \frac{\tau}{\max(\mathbf{V}(:, j))}$. Then, $\tilde{\mathbf{V}}(:, j) = \text{round}(\gamma(j)\mathbf{V}(:, j))$, where $\text{round}()$ rounds to the nearest integer. Then, the scaling introduced by γ is absorbed into the nonnegative factor \mathbf{U} so that we would have an equivalent model if rounding was not present: $\tilde{\mathbf{U}}(:, j) = \frac{1}{\gamma(j)}\mathbf{U}(:, j)$; **3)** Our approach introduced in Section 2, where the result of the scaled-and-rounded NMF heuristic is used as initialization; **4)** Our approach introduced in Section 2, where we randomly initialize the input factors.

³<https://mimic.physionet.org>

⁴We do not reveal the source of our data, in order to preserve the submission’s anonymity.

⁵<https://www.nlm.nih.gov/research/umls/rxnorm/docs/rxnormfiles.html>

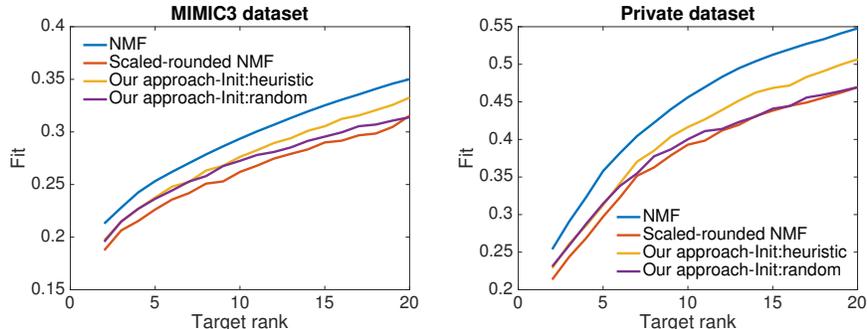


Figure 1: Fit measured as a function of the target rank in the range $\{2, \dots, 20\}$.

Table 1: Contents of two representative phenotypes as produced by our approach ($\tau = 3$) for our private dataset.

Leukemia Phenotype		Chronic Kidney Disease Phenotype	
Medical Feature	Severity	Medical Feature	Severity
DIAG_Leukemias	3	DIAG_Chronic kidney disease	3
DIAG_Maintenance chemotherapy; radiotherapy	2	PROC_Microscopic examination (bacterial smear; culture; toxicology)	2
MED_Heparin and related preparations	2	DIAG_Immunity disorders	1
PROC_Cancer chemotherapy	2	DIAG_Hypertension with complications and secondary hypertension	1
DIAG_Immunity disorders	1	DIAG_Other liver diseases	1
MED_Antiemetic/antivertigo agents	1	DIAG_Other diseases of kidney and ureters	1
MED_Analgesics narcotic anesthetic adjunct agents	1	MED_Immunosuppressives	1
MED_Topical local anesthetics	1	PROC_Other laboratory	1
MED_Antineoplastic - antimetabolites	1		
MED_General anesthetics injectable	1		
MED_Sodium/saline preparations	1		

Results We define the fit as $1 - \|\mathbf{X} - \mathbf{UV}^T\|_F / \|\mathbf{X}\|_F$, which can be considered as the proportion of the input explained by the model. For each target rank, we run each method 5 times and choose the solution with the best fit. As shown in Figure 1, our approach using the scaled-and-rounded heuristic as initialization consistently outperforms the baseline heuristic. The average gain for all the target ranks is 1.5% for the MIMIC3 dataset and 2.6% for the private one. The maximum gain is 2.2% for the MIMIC3 and 3.7% for the private dataset. The heuristic is also outperformed by our approach with random initialization in the vast majority of cases considered. The fact that NMF achieves the highest fit is expected and reasonable since both of its output factors are permitted to take real numbers, thus provide more flexibility to produce an accurate solution. In Table 1, we provide the highest-severity entries of a representative phenotype of our results. Our clinical expert confirmed the simplicity and usefulness of our discrete phenotyping scheme. Due to lack of space, we plan to provide extensive qualitative results in an extended version of this work.

4 Related Work

Computational Phenotyping There is extensive prior art in unsupervised EHR-based phenotyping using factorization models (e.g., [11, 12, 31, 25, 15, 26, 10]). However, no work has considered extracting scoring-based phenotypes to facilitate their interpretation by domain experts.

Discrete factorization-based approaches The line of work defined in [22, 5] considers the problem of approximating discrete data by fully-discrete factors. Instead, our approach permits the input matrix to contain real numbers and restricts one of the two factors (i.e., the phenotypes’ definition) to be discrete. In [24, 17], the authors consider compressing real matrices into factors (scaled by real numbers) restricted to take values from the set $\{-1, 0, 1\}$. A direct application of this approach would introduce negative values into the phenotypes, which would hurt their interpretability. Finally, there is another body of literature studying the problem of factorizing binary or real data into binary factors (e.g., [19, 20, 33, 29, 23]).

5 Conclusions

A useful conclusion from this work would be that factorization models could benefit from a mathematically principled approach of incorporating discrete constraints, rather than solely using rounding-based heuristics.

References

- [1] Clinical classifications software (ccs) for icd-9-cm. <https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp>, 2017. Accessed: 2017-02-11.
- [2] S. C. Bagley and R. B. Altman. Computing disease incidence, prevalence and comorbidity from electronic medical records. *Journal of biomedical informatics*, 63:108–111, Oct. 2016.
- [3] D. P. Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.
- [4] A. Cichocki, A. H. Phan IEICE transactions on fundamentals of, and 2009. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *search.ieice.org*, 2009.
- [5] B. Dong, M. M. Lin, and H. Park. Integer matrix approximation and data mining. *Journal of scientific computing*, pages 1–27, Sept. 2017.
- [6] N. Gillis. The why and how of nonnegative matrix factorization. Jan. 2014.
- [7] N. Gillis et al. Nonnegative matrix factorization: Complexity, algorithms and applications. *Unpublished doctoral dissertation, Université catholique de Louvain. Louvain-La-Neuve: CORE*, 2011.
- [8] N. Gillis and F. Glineur. Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization. *Neural computation*, 24(4):1085–1105, Apr. 2012.
- [9] B. A. Goldstein, N. A. Bhavsar, M. Phelan, and M. J. Pencina. Controlling for informed presence bias due to the number of health encounters in an electronic health record. *American journal of epidemiology*, 184(11):847–855, Dec. 2016.
- [10] J. Henderson, J. C. Ho, A. N. Kho, J. C. Denny, B. A. Malin, J. Sun, and J. Ghosh. Granite: Diversified, sparse tensor factorization for electronic health Record-Based phenotyping. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 214–223, Aug. 2017.
- [11] J. C. Ho, J. Ghosh, S. R. Steinhubl, W. F. Stewart, J. C. Denny, B. A. Malin, and J. Sun. Limestone: high-throughput candidate phenotype generation via tensor factorization. *Journal of biomedical informatics*, 52:199–211, Dec. 2014.
- [12] J. C. Ho, J. Ghosh, and J. Sun. Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14*, pages 115–124, New York, NY, USA, 2014. ACM.
- [13] N.-D. Ho. *NONNEGATIVE MATRIX FACTORIZATION ALGORITHMS AND APPLICATIONS*. PhD thesis.
- [14] G. Hripcsak, D. J. Albers, and A. Perotte. Parameterizing time in electronic health record studies. *Journal of the American Medical Informatics Association: JAMIA*, 22(4):794–804, July 2015.
- [15] S. Joshi, S. Gunasekar, D. Sontag, and G. Joydeep. Identifiable phenotyping using constrained Non-Negative matrix factorization. In *Machine Learning for Healthcare Conference*, pages 17–41, Dec. 2016.
- [16] J. Kim, Y. He, and H. Park. Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *Journal of Global Optimization*, 58(2):285–319, Feb. 2014.
- [17] T. G. Kolda and D. P. O’Leary. A semidiscrete matrix decomposition for latent semantic indexing information retrieval. *ACM Transactions on Information and System Security*, 16(4):322–346, Oct. 1998.
- [18] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8), 2009.

- [19] M. Koyuturk and A. Grama. PROXIMUS: A framework for analyzing very high dimensional Discrete-Attributed datasets.
- [20] M. Koyuturk, A. Grama, and N. Ramakrishnan. Compression, clustering and pattern discovery in very high dimensional Discrete-Attribute datasets.
- [21] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, Oct. 1999.
- [22] M. Lin. Discrete Eckart–Young theorem for integer matrices. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1367–1382, Oct. 2011.
- [23] P. Miettinen, T. Mielikäinen, A. Gionis, G. Das, and H. Mannila. The discrete basis problem. *IEEE transactions on knowledge and data engineering*, 20(10):1348–1362, Oct. 2008.
- [24] D. O’Leary and S. Peleg. Digital image compression by outer product expansion. *IEEE Transactions on Communications*, 31(3):441–444, Mar. 1983.
- [25] I. Perros, R. Chen, R. Vuduc, and J. Sun. Sparse hierarchical tucker factorization and its application to healthcare. In *2015 IEEE International Conference on Data Mining*, pages 943–948, Nov. 2015.
- [26] I. Perros, E. E. Papalexakis, F. Wang, R. Vuduc, E. Searles, M. Thompson, and J. Sun. SPARTan: Scalable PARAFAC2 for large & sparse data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’17, pages 375–384. ACM, 2017.
- [27] R. L. Richesson, J. Sun, J. Pathak, A. N. Kho, and J. C. Denny. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. *Artificial intelligence in medicine*, 71:57–61, July 2016.
- [28] A. Rusanov, N. G. Weiskopf, S. Wang, and C. Weng. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. *BMC medical informatics and decision making*, 14:51, June 2014.
- [29] B.-H. Shen, S. Ji, and J. Ye. Mining discrete patterns via binary matrix factorization. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’09, pages 757–766, New York, NY, USA, 2009. ACM.
- [30] B. Ustun and C. Rudin. Optimized risk scores. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1125–1134. ACM, 13 Aug. 2017.
- [31] Y. Wang, R. Chen, J. Ghosh, J. C. Denny, A. Kho, Y. Chen, B. A. Malin, and J. Sun. Rubik: Knowledge guided tensor factorization and completion for health data analytics. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’15, pages 1265–1274, New York, NY, USA, 2015. ACM.
- [32] P. Yadav, M. Steinbach, V. Kumar, and G. Simon. Mining electronic health records: A survey. Feb. 2017.
- [33] Z. Zhang, T. Li, C. Ding, and X. Zhang. Binary matrix factorization with applications. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 391–400, Oct. 2007.
- [34] K. Zheng, J. Gao, K. Y. Ngiam, B. C. Ooi, and W. L. J. Yip. Resolving the bias in electronic medical records. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’17, pages 2171–2180, New York, NY, USA, 2017. ACM.