# Proposal
# Your Job Role Determines Your Access Privileges*

Yunfeng Jiang, Jinzheng Sha, and Zhao Tian

**Problem** At any company, when employees start work, they need to obtain the necessary computer access in order to fulfill their job. This access might allow an employee to read or manipulate resources through various applications or web portals. For example, a software engineer needs the access to the source code repositories. Conventionally, the access is granted through trial and error—employees figure out the access they need as they encounter roadblocks during their daily work, e.g., not able to log into a reporting portal. That practice is inefficient because the task takes a knowledgeable supervisor plenty of time to manually grant the needed access in order to overcome access obstacle. At the same time, the supervisor must avoid granting unnecessary access for the reason of security. As employees move throughout a company, this access discovery/recovery cycle wastes a nontrivial amount of time and money [1].

Employees who perform the functions of the same job role should access the same or similar resources. So we can build a model, learned using historical data, that will predict an employee's access needs. The model will take an employee's role information and a resource code and will return whether or not access should be granted. The problem can be formulated as a binary classification problem. We will apply different classification techniques to approach this problem.

**Data** The data we will use consist of real historical data collected from 2010 and 2011, provided by Amazon [2]. In this dataset, employees are manually allowed or denied access to resources over time. In the training set, each row has the ACTION as ground truth (ACTION is 1 if the resource was approved, 0 if the resource was not), RESOURCE, and information about the employee's role at the time of approval. There are 7518 different resources to be granted. In order to describe a job role, the dataset employs 8 features, such as his/her manager (4243 possible values), department names (449 possible values), and role title (343 possible values). The training set totally contains 32769 lines of access data related to current employees and their provisioned access. The testing set contains 58921 lines, which involve 4971 different resources.

**Methods** In the dataset, there are 8 different features of each employee to determine whether a certain resource should be granted to the employee or not. It can be modeled as a classification

*The problem is inspired by the Amazon's Employee Access Challenge. http://www.kaggle.com/c/amazon-employee-access-challenge

problem and many methods are available to solve the problems, such as logistic regression, decision tree, support vector machines (SVM), and neural network. The challenge is that there are over 7000 different resources in the data set. On the one hand, logistic regression and decision tree cannot capture the correlation between these resources. On the other hand, the dimension is too large for SVM and neural network. In order to handle the high dimension and capture the potential correlation between resources simultaneously, we need to use a blend of a logistic model and a mixture of tree-based models, random forests [3] and Extra-Trees models to solve the problem.

**Milestones** First, we would like to apply the classical classification techniques, such as logistic regression, and use them as the baseline. We will then implement some more advanced techniques, such as random forests, and compare their performance with the baseline, in terms of accuracy and speed. The implementation of the advanced classification techniques that will not be covered in the class will be our mid-term milestone. After the milestone, we aim to explore a novel classification method, which blends the logistic regression and tree-based models, in order to further improve the performance.

# References


[1] Amazon employee access challenge, http://www.kaggle.com/c/amazon-employee-access-challenge.

[2] Data set, Amazon employee access challenge, http://www.kaggle.com/c/amazon-employee-access-challenge/data.

[3] A. Liaw and M. Wiener. Classification and regression by randomForest. *R news*, 2(3):18–22, 2002.