DS501

CASE STUDY 4 Report

Backbone
Human Resource Management System

Team 10
Jiexuan Sun, Congyang Wang, Zhaoning Su, Chao Xu

0. Background & Motivation

Needless to say, employees are the core and soul of the enterprise. They are the engine of enterprise operating and growth, especially for those key talents. The cost of talent loss is very high for each company. So, detecting potential leaving employee is certainly needed for each company. Then company have the chance to take the necessary measures to prevent this 'tragedy' happening. A system which could prevent talent loss is essential. We would like to build a human resource management system to prevent the loss of key talents.

1. Product Overview

The human resource management system we designed include three parts.

1.1 The 'Gamified' human resource management part.

We 'gamify' each employee's' career. For example, there are gamification-career record system and many game ‘indicators’ such as 'level', 'experience' and etc. On the one side, this structure hope to relieve employees from boring works. On the other side, we would like to build a sense of belonging for each employee. This could solve the talent loss problem from root. This structure is achieved through database design. Because it is not covered in this course, we won't mentioned any details in the following.

1.2 The Face and smile recognition part.

The second part of the system is our face recognition technology for identity verification in access system. It will not only recognize our employees' face but also detect whether they smile or not. By contrasting whether they smile or not before and after one day's work, we could predict whether they satisfy with today's work by their smile. After collecting enough data, we could compute each employee's 'satisfaction level'. This part use the machine learning knowledge. We would introduce the details in the following.

1.3 The Talent loss detecting part.

We could collect the data of employee's satisfaction level by using face and smile recognition part. By combining this important data with other human resources information, we could use them to detect the potential leaving employee. This could solve the talent loss problem from the 'surface'. But, it is quite useful and will get effect instantly. We would introduce the details in the following.

3. Product Analysis

Product / Service

- **Product Name:** Backbone
- **Product Target Market:** Every company who care the talents
- **SWOT Analysis:**

SWOT	
Strength	**Weaknesse**
High predicting accuracy. Integrity HR management System. Low-cost and etc.	No data protection. No anti-hacker part.
Opportunities	**Threat**
There is no current company do the same thing. This leave us more time to grow.	The algorithm we use is open resource, there is a possible to meet many 'copycats'.

4. Product Tech Description

4.1 The Data

The dataset HR_comma_sep.csv we use is from www.kaggle.com. It mainly describes all kinds of performances of employees. The dataset contains the following 10 attributes:

Independent Variable:
Employee satisfaction level - Decimal variable Range from (0,1)
Last evaluation (The performance of the employee) - Decimal variable Range from (0,1)
Number of projects - Integer variable
Average monthly hours - Integer variable
Time spent at the company - Integer variable
Work accident (Whether employee have had a work accident) - Boolean variable (0 - No/1 - Yes)
Sales (Job position) - Categorical variable including accounting; hr; IT; management;
Salary - Categorical variable including (High level; Medium level; Low level)
Promotion_last_5years - Boolean variable (0 - No/1 - Yes)
Dependent/Target Variable:
Left (Whether the employee has left or not finally) - Boolean variable - (0 - No/1 - Yes)

We have 15000 observations. We want to analysis why are our best and most experienced employees leaving prematurely with these data and make a prediction model to predict potential leaving employee. So, we could take necessary measure before they quiting.

4.2 The Tools

4.3 The Analysis

a) The 'gamified' human resource management part

This structure is achieved through database design. Because it is not covered in this course, we won't mentioned any details in the following.

b) The face and smile recognition part

The Math Theoretical Basis:

In this modular we do two things: First, we use PCA and SVM to do a face recognition which can be used for judging whether a person is a staff of the company. Second we use K-fold cross-validation and SVM to do a supervised learning to judge a person is smiling or not.

The **PCA(Principal component analysis)** method is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components:

$$\sum_{m=1}^{M} \theta_m Z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \varphi_{mj} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \varphi_{mj} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij}$$

The **K-fold cross-validation** is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. With k-fold Cross Validation, we divide the data set into K different parts. Then remove the first part, fit the model on the remaining K-1 parts, and see how good the predictions are on the left out part and repeat this K different times taking out a different part each time.By averaging the K different MSE's we get an estimated validation (test) error rate for new observations.

The **SVM(Support vector machines)** method is a kind of supervised learning model with associated learning algorithms that analyze data used for classification and regression analysis. In SVM method we use lagrange duality and kernel function to get the maximum value.

$$\max \frac{1}{\|w\|} \quad s.t., y_i(w^T x_i + b) \geq 1, i = 1, \ldots, n$$

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{n} \alpha_i \Big(y_i(w^T x_i + b) - 1 \Big)$$

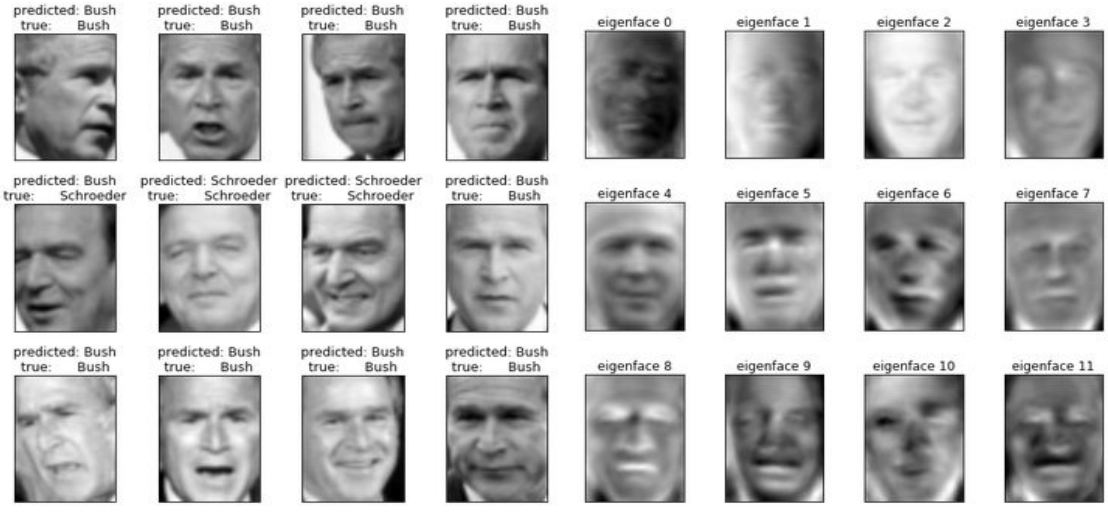
The Analysis Procedure:

Step 1. Data Exploration

We use two datasets from *scikit-learn.org* to analyze. The *fetch_lfw_people* data set is used for doing face recognition and the *fetch_olivetti_faces* data set is used for judging whether a person is smiling. Here we use all of the data in *fetch_lfw_people* and only 50 images in *fetch_olivetti_faces*. Otherwise we get some photos in our daily life to assess the accuracy of the smiling classification.

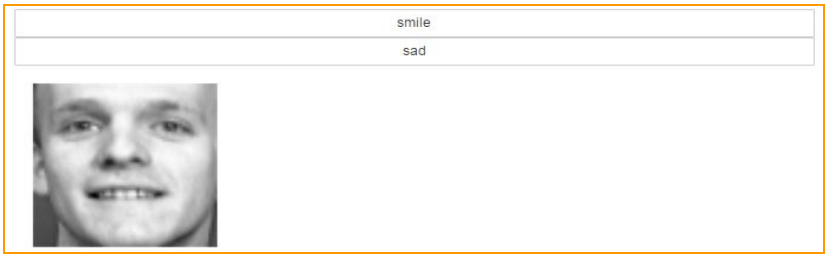
Step 2. Modeling

First we use PCA to get the eigenfaces of every images. In this way we reduce the dimension for a easilier classification. After this we use SVM to do the supervised learning classification as the follow picture shows. The left is the result of prediction in test data set and the right is the eigenface we get:



Here we want to collect the data of employee's satisfaction level. We think about collect it through smiling recognition and use k-fold cross-validation and SVM to do a supervised learning on whether a person is smiling. We pick 50 of the images in Olivetti face dataset used in this learning, and do a simple GUI on ipython book to label them as *smile* or *sad*. In this UI there are

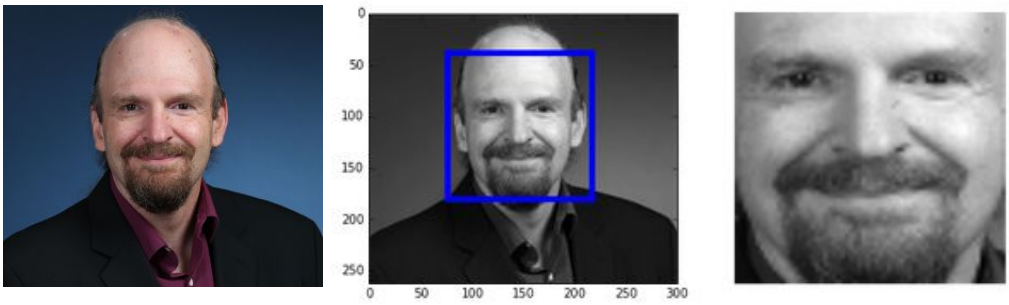
two buttons and a spare space to show the image we need to label, when we click on the button we get the classification we set.



The next we use the k-fold cross-validation to resample the dataset and get the best fit of the SVM model. And get a 80% accurate model.

Step 3. Actionable Insights and Visualization

Now we have a model to judge whether a person is smiling. We try to use it in real life. Here we use the OpenCV package to do the face recognition on real life photos. And cut the face we have recognized with the same size of images in Olivetti data set.



Then we want to assess it is a smiling one or not, we use the SVM model to predict it and get the answer: It is a smiling one!

```
New Photo for prediction
This person is smiling or not: yes
```

Now we could collect the data of employee's satisfaction level by using face and smile recognition part. Obviously, people smile more often, they definitely more satisfied with their work. By combining this important data with other human resources information, we could use them to detect the potential leaving employee.

c) The talent Loss detecting part

The Math Theoretical Basis:

After trying multiple classifier, we finally decide to use the Logistic Regression. The logistic Regression would give us each employee's leaving probability. We set 0.5 as our threshold. Then, we could use this model to predict who might want to quit.

$$\log \frac{p(x)}{1-p(x)} = \beta_0 + x \cdot \beta$$

We use the Likelihood function to 'solve' the Logistic Regression.

$$L(\beta_0, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i)^{1-y_i}$$

$$\begin{aligned}
\ell(\beta_0, \beta) &= \sum_{i=1}^{n} y_i \log p(x_i) + (1 - y_i) \log 1 - p(x_i) \\
&= \sum_{i=1}^{n} \log 1 - p(x_i) + \sum_{i=1}^{n} y_i \log \frac{p(x_i)}{1 - p(x_i)} \\
&= \sum_{i=1}^{n} \log 1 - p(x_i) + \sum_{i=1}^{n} y_i (\beta_0 + x_i \cdot \beta) \\
&= \sum_{i=1}^{n} -\log 1 + e^{\beta_0 + x_i \cdot \beta} + \sum_{i=1}^{n} y_i (\beta_0 + x_i \cdot \beta)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ell}{\partial \beta_j} &= -\sum_{i=1}^{n} \frac{1}{1 + e^{\beta_0 + x_i \cdot \beta}} e^{\beta_0 + x_i \cdot \beta} x_{ij} + \sum_{i=1}^{n} y_i x_{ij} \\
&= \sum_{i=1}^{n} (y_i - p(x_i; \beta_0, \beta)) x_{ij}
\end{aligned}$$

We first split our data into testing and training dataset. Then, We use sklearn 'linear_model' module fit our training dataset into Logistic Regression model. Finally, we use the testing dataset to test the model performance.

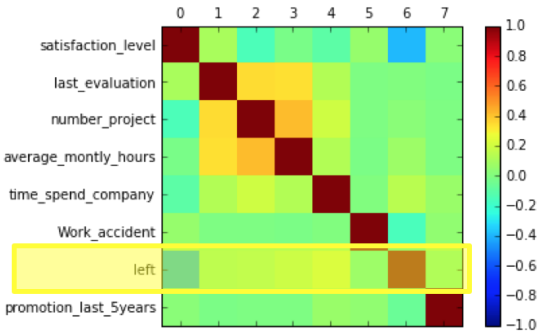
The Analysis Procedure:

Step 1. Data Exploration

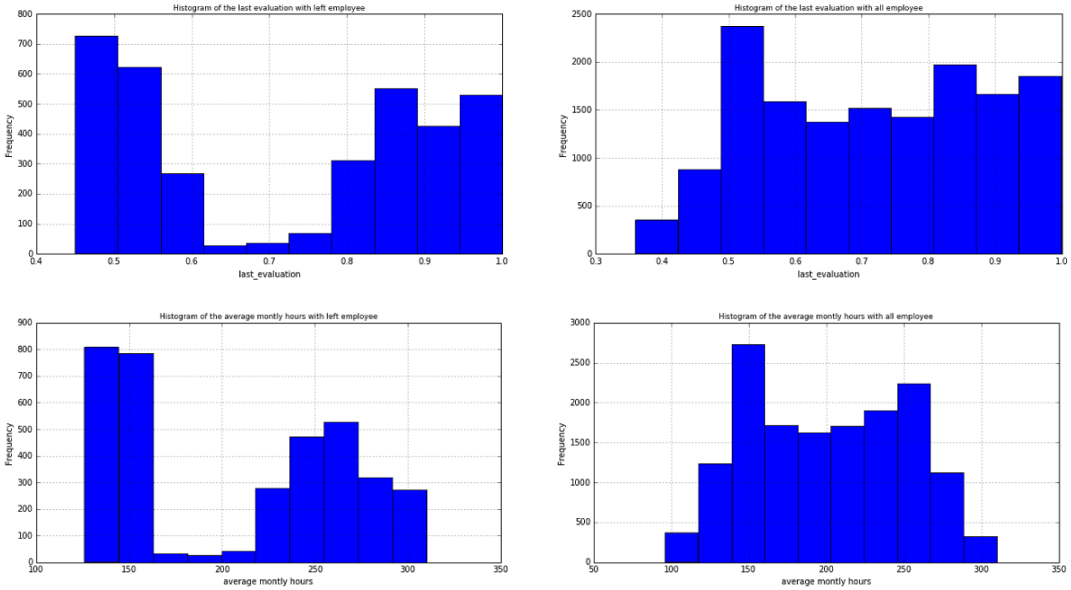
This table describe the characteristics of each features of our dataset. We can see different statistical measures of central tendency and variation. For example we can see that our attrition rate is equal to 24%, the satisfaction level is around 61% and the performance average is around 71%. We see that on average people work on 3 to 4 projects a year and about 200 hours per months.

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_
count	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000	14999.0000
mean	0.612834	0.716102	3.803054	201.050337	3.498233	0.144610	0.238083	0.021268
std	0.248631	0.171169	1.232592	49.943099	1.460136	0.351719	0.425924	0.144281
min	0.090000	0.360000	2.000000	96.000000	2.000000	0.000000	0.000000	0.000000
25%	0.440000	0.560000	3.000000	156.000000	3.000000	0.000000	0.000000	0.000000
50%	0.640000	0.720000	4.000000	200.000000	3.000000	0.000000	0.000000	0.000000
75%	0.820000	0.870000	5.000000	245.000000	4.000000	0.000000	0.000000	0.000000
max	1.000000	1.000000	7.000000	310.000000	10.000000	1.000000	1.000000	1.000000

This graph present the correlations between each variables. The color of the box reveal the significance of the correlation, while the colour present the direction (either positive or negative). From this plot, the relations between left and other variables are not obvious.

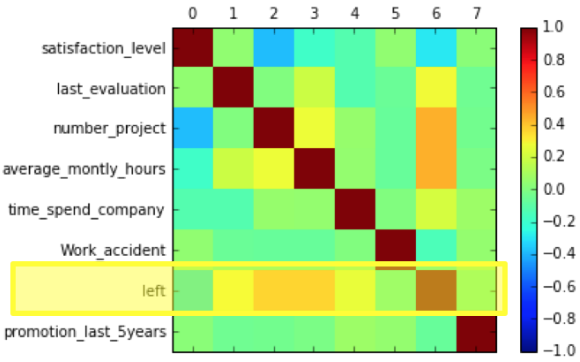


We next create a data frame with only the people that have left the company, so we can visualise what is the distribution of each features. Then, we compare those features' distributions with the data including all the employee. The aim is to find some characters of leaving employee. The difference between these two datasets are circled in the following histogram.



More problematic, company only want to retain those 'good' employee. So, we create the dataset include the total of employees that received an evaluation above average (0.7), or spend at least four years in the company, or were working on more than 5 projects at the same time and still

have left the company. These are the people the company should have retained. Then, we create the correlation matrix again and try to find some variables which are correlated with 'good' employee's leaving. And, this time, it's much clearer. month and aren't promoted.On average valuable employees that leave are not satisfied, work on many projects, spend many hours in the company each



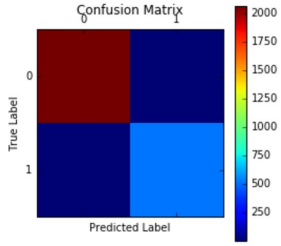
Step 2. Modeling

We use many classifiers to build the model and test each of them. Then we select the classifier with best performance. The final model we choose is based on Logistic Regression, which has a very high accuracy 0.98. And the model is quite robust and stable.

↓ Logistic Regression:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	2075
1	0.98	1.00	0.99	524
avg / total	1.00	1.00	1.00	2599

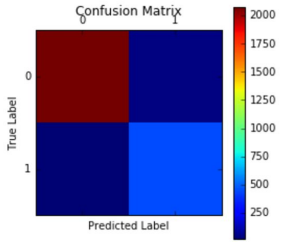
```
[[2065   10]
 [   2  522]]
```



↓ Tree:

	precision	recall	f1-score	support
satisfaction_level	0.99	0.98	0.98	2130
last_evaluation	0.90	0.96	0.93	469
avg / total	0.97	0.97	0.97	2599

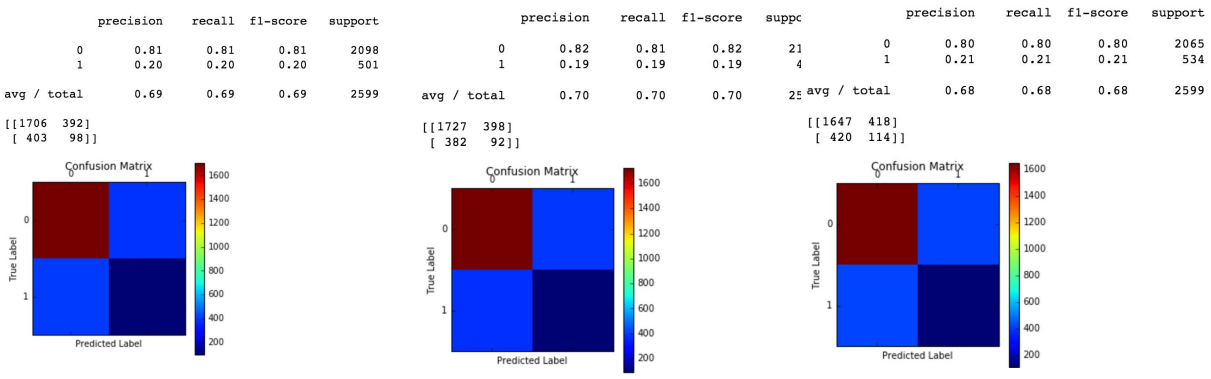
```
[[2080   50]
 [  19  450]]
```



↓ LinearSVC:

↓ KNN:

↓ Random Forest:

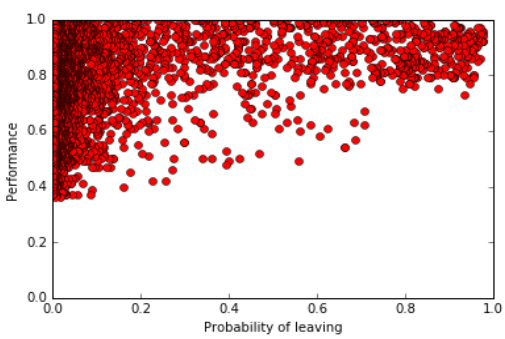


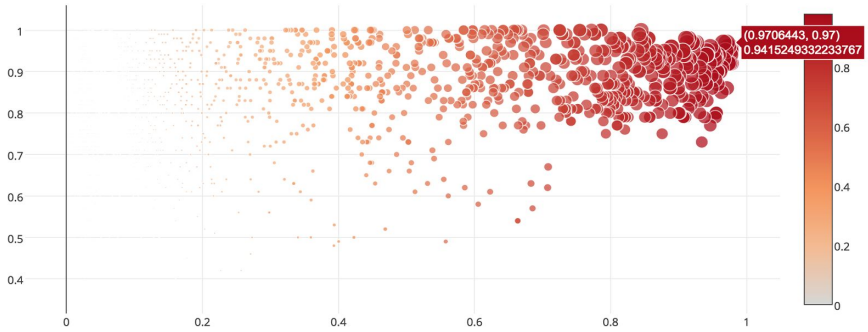
Step 3. Actionable Insights and Visualization

First, we create a table about those employees that the company should retain. We calculate each of them's priority (= probToLeave * Performance). After grouping them per department we could email the different managers to tell them which valuable employees might leave soon.

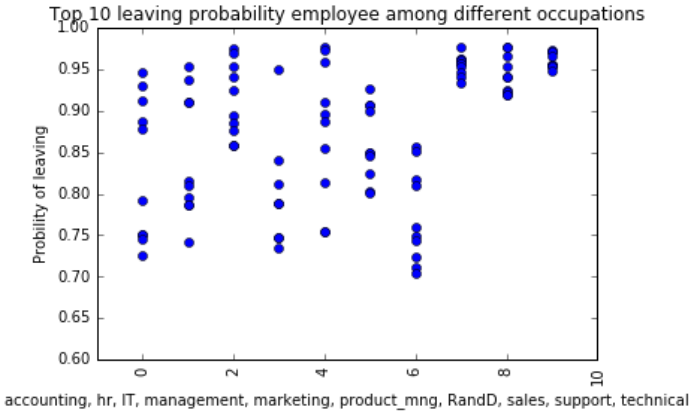
	id	probToLeave	performance	priority
1280	1648	0.970644	0.97	0.941525
989	12479	0.976362	0.96	0.937307
2017	809	0.965581	0.97	0.936614
396	1488	0.972664	0.96	0.933758
374	12160	0.976252	0.95	0.927440
864	12522	0.968830	0.95	0.920389
1473	1807	0.975763	0.94	0.917217
986	12550	0.946356	0.96	0.908501
8	311	0.950186	0.95	0.902676
867	1534	0.957798	0.94	0.900331
1602	14265	0.976940	0.92	0.898785
783	10294	0.926034	0.97	0.898253
33	820	0.965369	0.93	0.897794
2314	14810	0.974678	0.92	0.896704

Then, we plot the following two graph about each employee's leaving probability and performance evaluation. The first shows a general potential employee distribution in a company. As we can see, most of them have small leaving probability. The second plot shows details about each potential leaving employee. The x-axis is the probability of leaving employee and the y-axis is the employee's performance, which is the data we get from variable last_evaluation.





At last, we plot the third graph as shown following. It plots the employee with top 10 leaving probability among different occupations. As we can see, technical employees and sales have relative high leaving probability. And position named 'RandD' has the lowest leaving probability.



5. The Summary

First, the system's predicting accuracy of potential leaving employee is 0.98. Which means almost all the employee who want to quit will be detected, which means the model is robust and stable. So, the system is reliable.The system is easy to implement and inexpensive. Compared with the cost of key talents loss, the cost of this system is rather minimal. Investing in this system is a wise choice. So, the system has low-cost. Key talents are so important for each company. The cost of Key talents loss is quite high. A system which could prevent key talent loss means save/earn resources for company. So, the system has high profit-margin. All in all, the potential value behind the system is huge.