

The background of the slide features a close-up of hands holding a pencil and writing on a globe. A large teal circle is centered over the globe. Surrounding this circle are several smaller circles in orange, red, and white, each containing a white icon: a puzzle piece, a lightbulb, and a thumbs-up. The text 'Berkeley Innovation Index' is written in white on the teal circle.

Berkeley Innovation Index

My Dinh
Jessica Gu
Aaron Lu
Dayou Wang
Yan Zeng
Yujun Zou

Outline

- Problem Statement and Project Significance
- General Approach
- Model Selection
 - Assessment
 - Problems
- Final result
- User Interface Demo
- Limitations and Future Improvement
- Conclusion



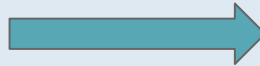
Problem Statement

Based on the Berkeley Innovation Index survey, create a hiring tool and generate an automated report

BI: Individual Mindset



**Machine Learning
Algorithms**





Project Significance

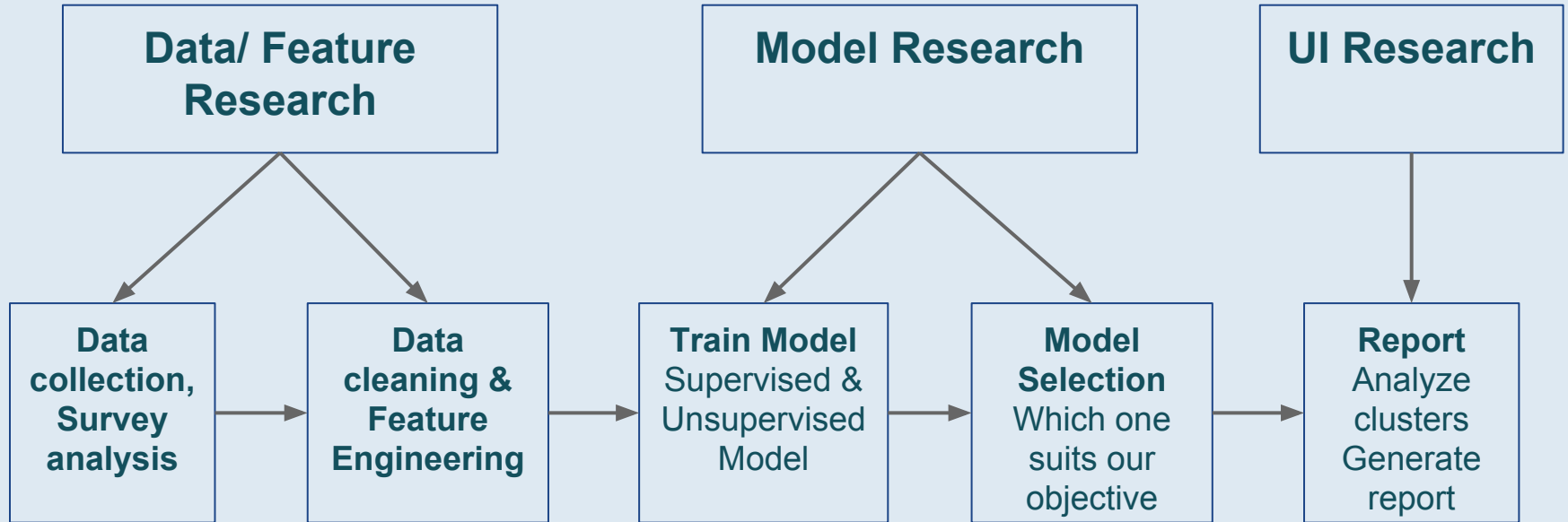


Our potential users are **HRs** in the company, they can use this as a **hiring tool**:

1. Have a direct picture of the candidate's personality and behavior, to see if this person suits the position and company's culture
2. Based on the multiple reports created for each candidate, HR could have a clearer comparison between them and easier to pick the most ideal one

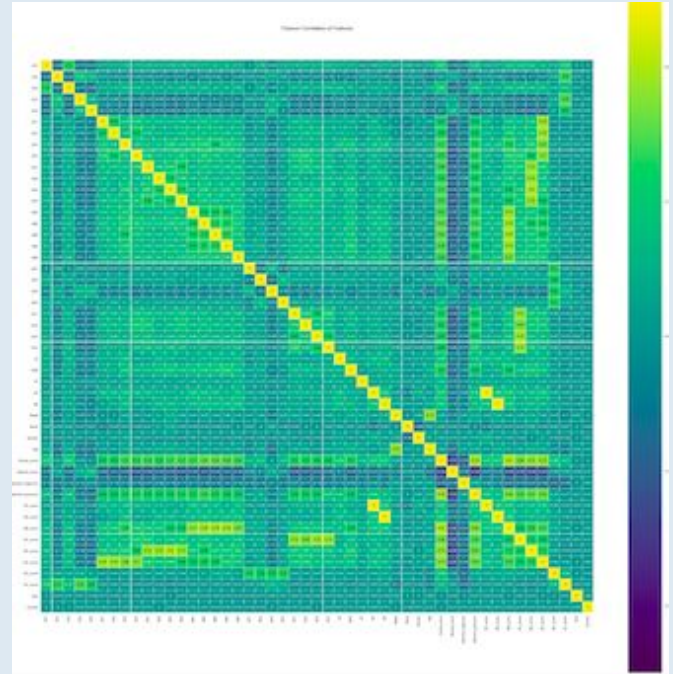


General Approach



Data Cleaning & Feature Engineering

- Data: Survey results in questions of trust, resilience, diversity, believe, effectiveness, collaboration, etc.
- Data Cleaning
 - Fake data cleaning
 - Non-numerical data cleaning
- Feature Engineering
 - Imbalanced Data - SMOTE
 - Correlation Verification
 - PCA & Recursive Feature Elimination(RFE)
 - Data normalization



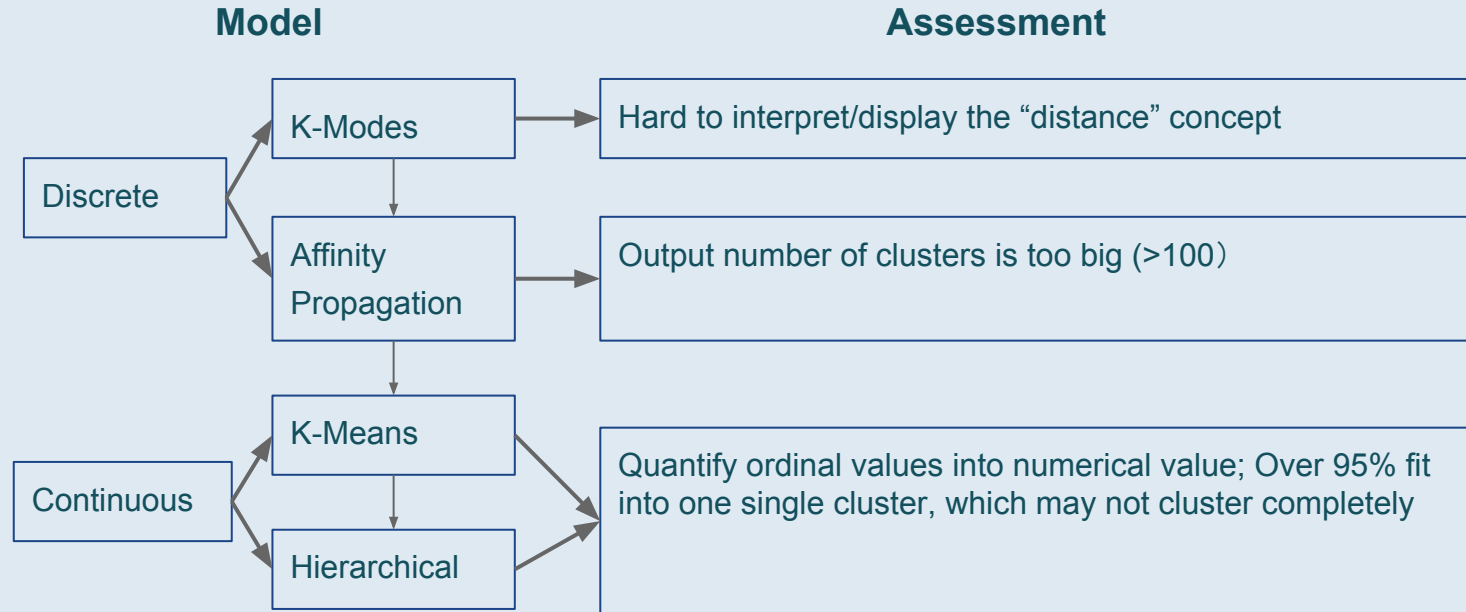


Approach

- Supervised Models
 - Target: ER (been successful in innovation)
 - Level: 1, 2, 3, 4, 5
- Assessment:
 - 73% (40% for baseline model)
 - Innovation index = $\sum(i * P(\text{belong to level } i))$

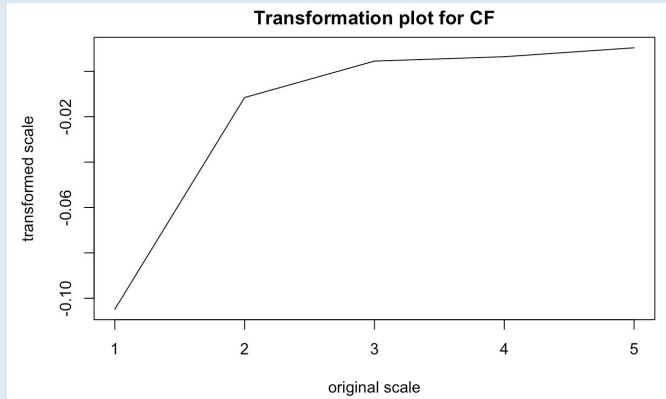
	Type of Model	Training error	Testing error
0	Lofistic Regression	46.37	42.00
1	SVM	49.63	44.67
2	perceptron	35.77	33.60
3	KNN	83.03	67.33
4	XGBoost	97.97	66.00
5	Random Forest	100.00	73.73

Approach

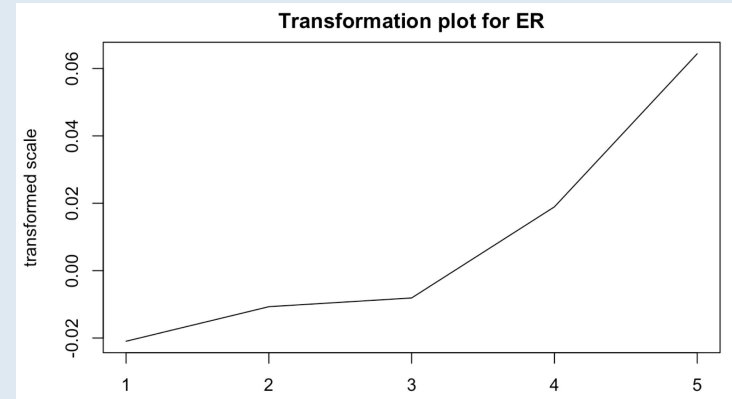


Approach - Hierarchical Clustering

1. Transform ordinal data into numerical data:



Use Alternative Least Square Algorithm to attain quantified ranking data





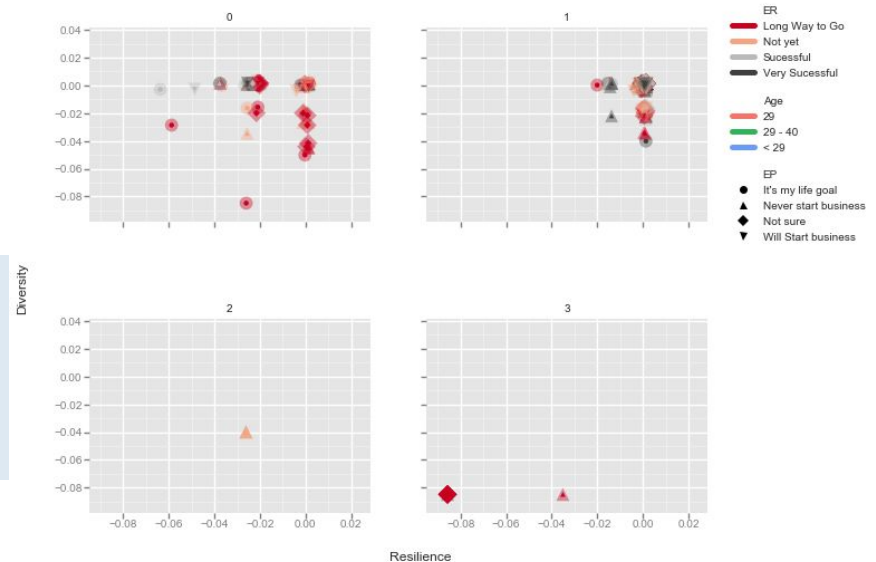
2. Fit hierarchical clustering:



They have a good balance of perfection and effectiveness. They are very persistent, detailed oriented, and methodological. They have a big picture and prefer information when making decision. Open for discussion to resolve the problems.

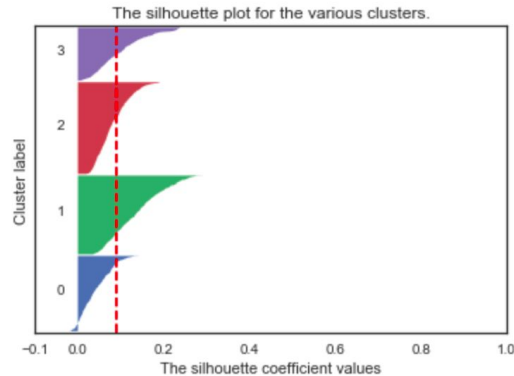
Group 2: Crusaders

Very outgoing, high confidence and self-motivation. They are very quick and decisive when making new moves. Their energy level is always high and never give up until achieving their mission



Current Approach

```
For n_clusters = 4 The average silhouette score is: 0.0904859952522
For n_clusters = 5 The average silhouette score is: 0.0692447212389
For n_clusters = 6 The average silhouette score is: 0.0749957980912
For n_clusters = 7 The average silhouette score is: 0.0782534546677
For n_clusters = 8 The average silhouette score is: 0.0639206479419
```



Determine the ideal number of clusters (n=4)

K-Means Clustering algorithm

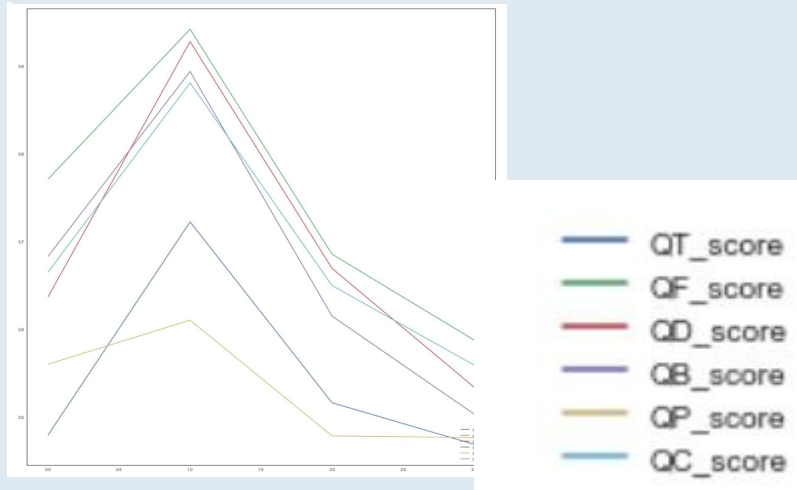
Analyze and interpret each cluster

Create deliverable for HR



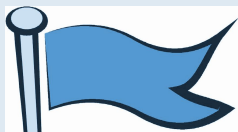
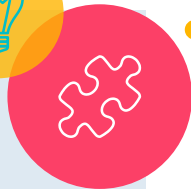
Current Approach

Group means are plotted to give a better visualization.



	0	1	2	3
CZ	0.52	0.76	0.49	0.39
SDR	0.67	0.87	0.70	0.52
CF	0.53	0.58	0.57	0.47
EP	0.71	0.88	0.61	0.43
ER	0.39	0.56	0.37	0.29
Stage	0.00	0.01	1.00	0.02
Study	0.08	0.46	0.39	0.49
Age	0.01	0.02	0.49	0.01
QT_score	0.48	0.72	0.52	0.47
QF_score	0.77	0.94	0.69	0.59
QD_score	0.64	0.93	0.67	0.53
QB_score	0.68	0.89	0.62	0.50
QP_score	0.56	0.61	0.48	0.48
QC_score	0.67	0.88	0.65	0.56
overall_score	0.56	0.80	0.53	0.42
extreme_score	0.16	0.56	0.15	0.11
neutral_score	0.27	0.08	0.28	0.41





Final Results: 4 clusters

Administrator

- Performs well especially in routine works and might need more incentives for potential challenges.

Captain

- Team leaders, always optimistic and self-determinate about challenges, and good at communicating with teammates.

Critic

- Curious people that self-motivated for challenges and attempts for every possible outcome.

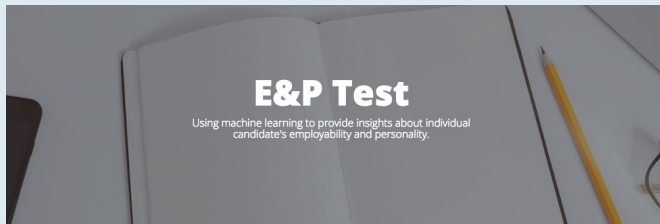
Communicator

- Good at delivering and sharing thoughts with diverse groups of people, and are potentially leaders with proper training.



User Interface Demo

<https://bii.herokuapp.com/bii/>



Partner Code

If you were provided a PARTNER CODE, please enter it here - else leave blank.

Project Code

If you were provided a PROJECT CODE, please enter it here - else leave blank.

QT1

Most people can be trusted: 1=Strongly disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly agree. Mark 3 for 'Don't know'

1	2	3	4	5
---	---	---	---	---

QT2

Most people tell a lie when they can benefit by doing so: 1=Strongly disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly agree. Mark 3 for 'Don't know'

1	2	3	4	5
---	---	---	---	---

QT3

I trust other people: 1=Strongly disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly agree. Mark 3 for 'Don't know'

1	2	3	4	5
---	---	---	---	---

E&P Result

Using machine learning to provide insights about individual candidate's employability and personality.

90.6

Your Innovation Index is: 90.6/100

The innovation score reflects your entrepreneurship. Higher score means you are more innovative. The average score for our sample is around 65.

Your Most Likely E&P Group is: Communicator

19.84%

Administrator

Administrators perform well especially in routine works and might need more incentives for potential challenges.

26.13%

Captain

Captains are team leaders, always optimistic and self-determined about challenges, and good at communicating with teammates.

22.53%

Critic

Critics are curious people who are self-motivated for challenges and attempts for every possible outcome.

31.5%

Communicator

Communicators are good at delivering and sharing thoughts with diverse groups of people, and are potentially leaders with proper training.

Strength

1. Robust working performance without challenges or pressure

Weakness

1. Ambitious but not effective
2. Feel passive about challenges
3. Sometimes lazy

Administrator

Possible Occupations


Administrative Assistants and Secretaries, Bill and Account Collectors, Financial Clerks, General Office Clerks, Information Clerks, Mail Clerks, Material Recording Clerks, Receptionists

Strength

1. Self-determination for challenges
2. Always optimistic about failures
3. Open-minded about diverse culture
4. Engage in various social networks



Limitations & Future Improvement

-  Revise questions in the survey
 - Add validation questions, to determine whether the data is 'useful'
 - Add critical questions, that have more **weights** towards the final clusters
- Further research on unsupervised continuous clustering models
- Change the current static model to dynamic model, and update the parameters of the clusters frequently



Conclusion

- Understand the data
 - Group the questions in survey
 - Data Cleaning and Feature Engineering
 - Dimension reduction
 - Normalization
- Training the data
 - Supervised ML models
 - Unsupervised ML models
 - Discrete vs. continuous data
- Interpret the results and revise the model
- Future improvements



Thank You!

GitHub:

https://github.com/aaronLinLu/BII_290GroupProject

