PREDICTING HOW INDIVIDUALS TRANSITION BETWEEN ORGANIZATIONS USING MACHINE LEARNING TECHNIQUES

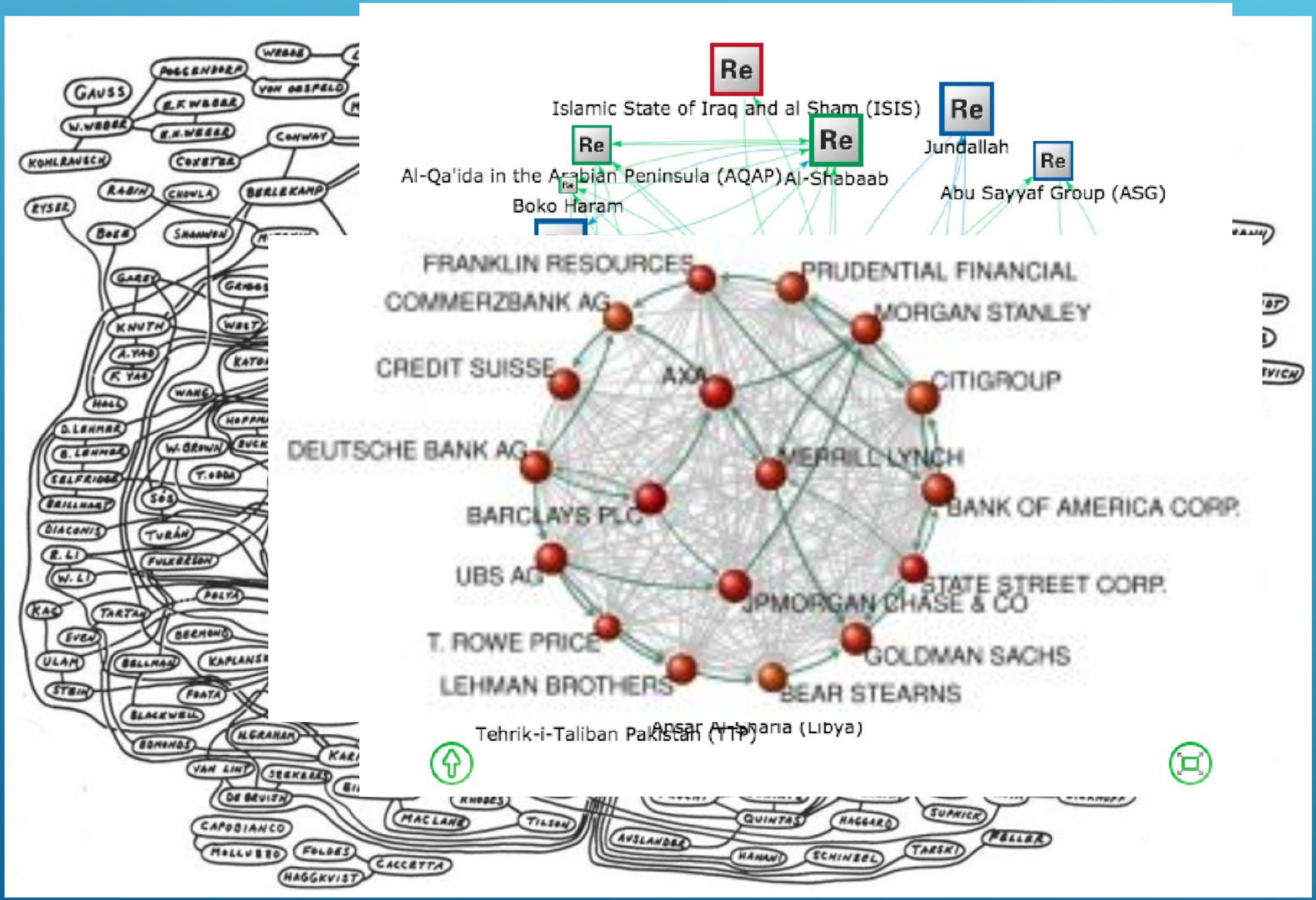
Joseph Leung
Advisor: Dr. Webb
Colleague: Tyler Moncur

In collaboration with DTRA (Defense Threat Reduction Agency)

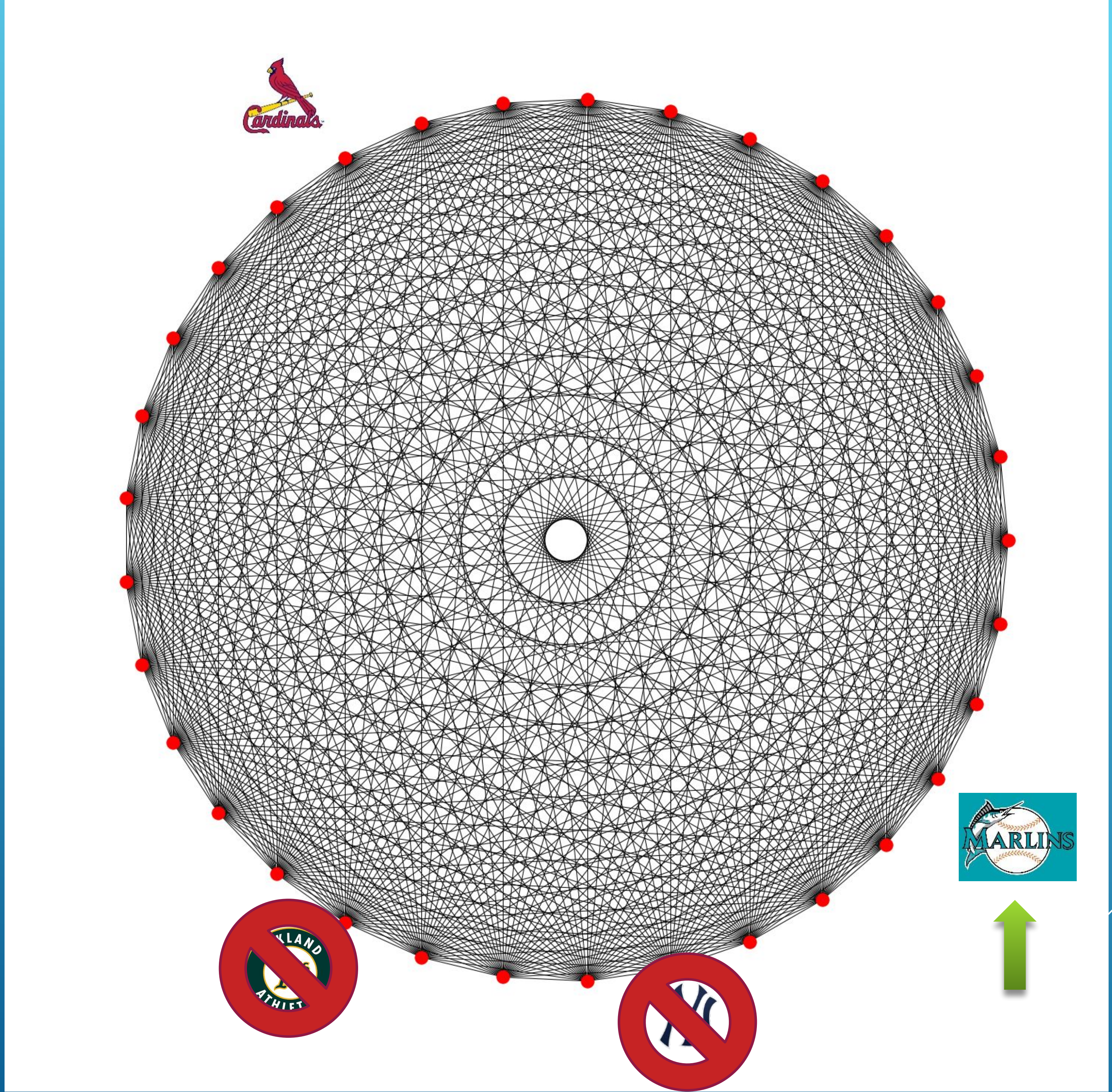
LINK PREDICTION PROBLEM

- **Definition**: Given a snapshot of a [given] network, we seek to accurately predict the edges that will be added to the network
  - Social networks – finding friends
- **Adjusted to**: Given a network between different groups/organizations, how can we determine how individuals might transition to and from these organizations?
- "A network model is useful to the extent that it can support meaningful inferences from observed network data."
    - Jon Kleinberg, Cornell University

EXAMPLES OF NETWORKS

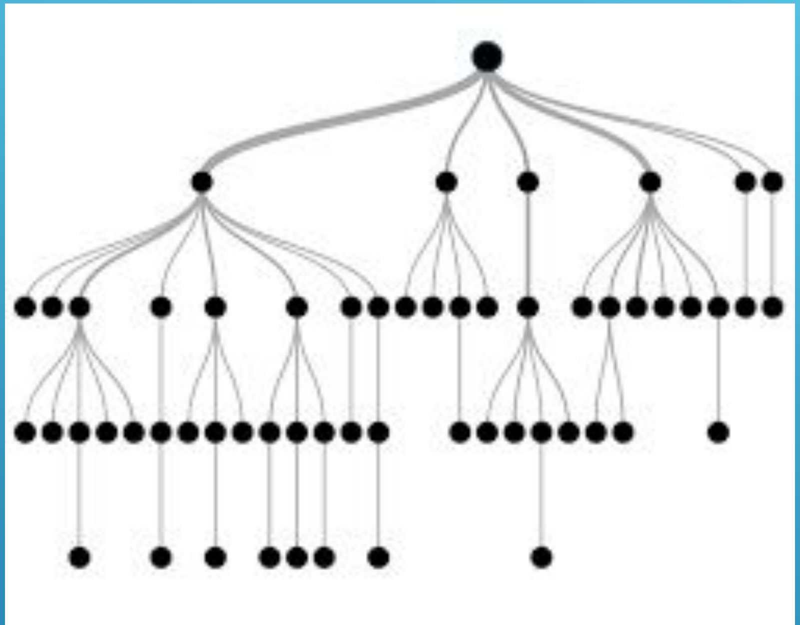



Yankees
SAN DIEGO
Padres
WASHINGTON
DC
NATIONALS
GIANTS
COLORADO
ROCKIES
MINNESOTA
Twins
Phillies
BOSTON
RED SOX
PIRATES
OAKLAND
A's
ATHLETICS
NY
RANGERS
Dodgers
FLORIDA
MARLINS
KC
Royals
MONTRÉAL
EXPOS
Orioles
Braves
Cardinals
Astros
Mets
PIRATES
MILWAUKEE
Brewers
SEATTLE
MARINERS


Cardinals
MARLINS

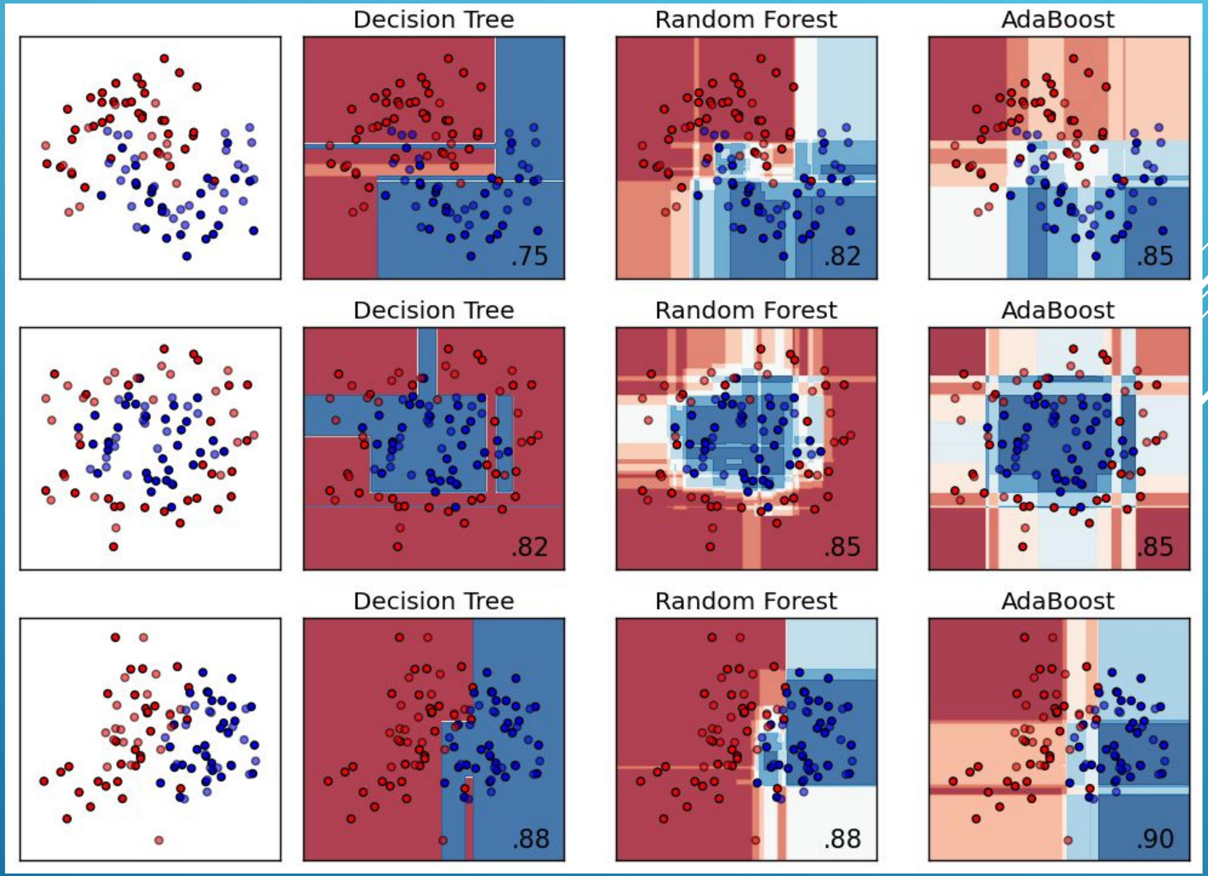
INITIAL APPROACH

- Diversified different decision models
- Optimization depended on data structure
  - Decision Tree – Extra Trees
  - Decision Tree - Random Forrest
  - Logistic Regression
  - Adaboost



ALGORITHMS BACKGROUND

- Train and testing variables
  - "Practice on trained variables"
  - Tests model on test variables
- Random Forest:
  - "Bootstrap Replica" of the learning sample
- Extra Trees
  - Makes "splits" at random
- Logistic Regression
  - Similar to linear regression, maps to a logistic representation
- Adaboost – Adaptive Boosting
  - Adapts to strong/weak classifiers



CHALLENGES, DIFFICULTIES

```
In [33]:  1 accuracy_score(y_test, rf.predict(X_test))

Out[33]: 0.03148148148148148
```

0.77% worse than chance

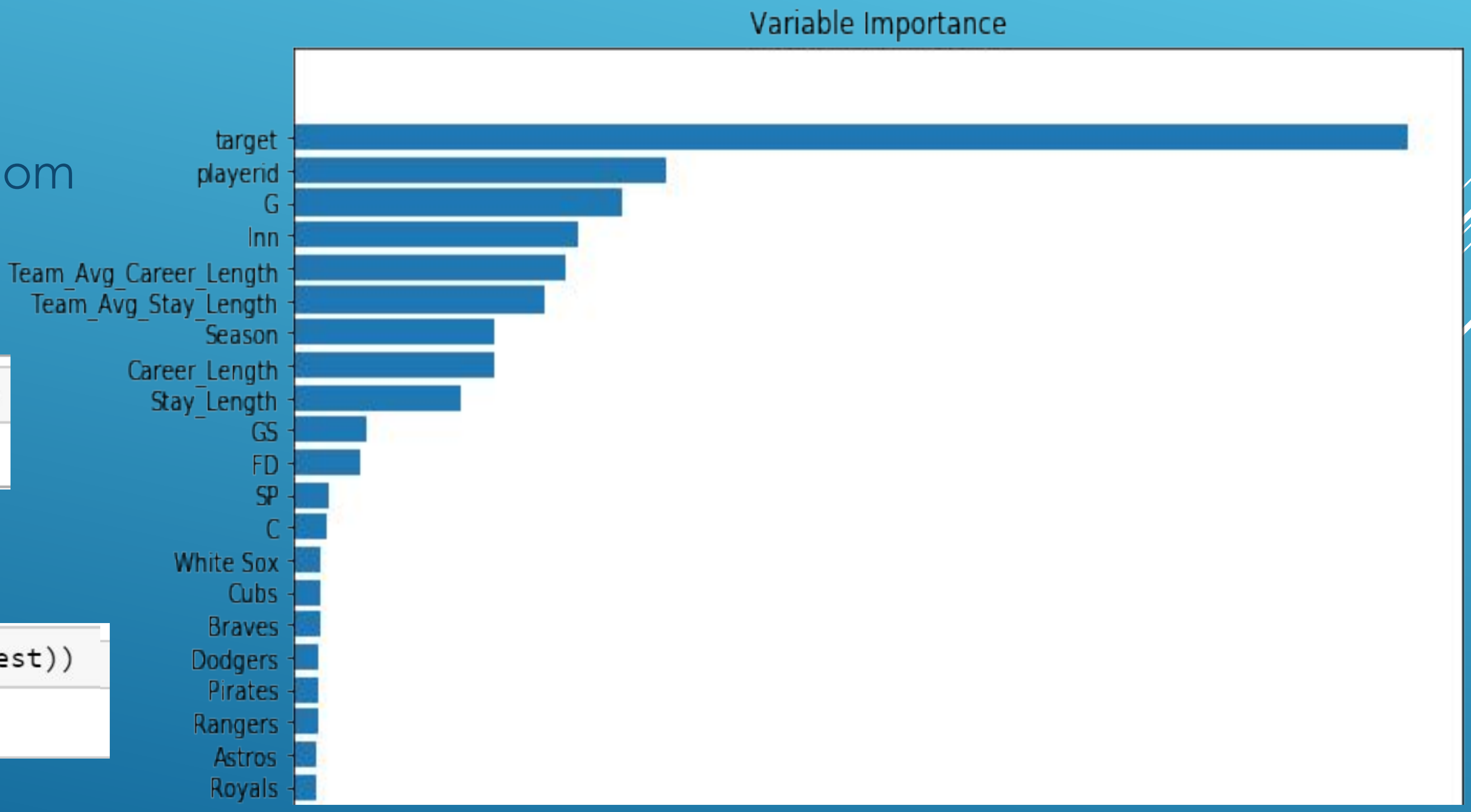
- 1/31 = 0.03225 - baseline
- What factors ?
  - 15260 total players to account for
  - Retiring a possibility – makes it “too easy”
    - Average Career Length: 5.6 years
      - http://www.nytimes.com/2007/07/15/sports/baseball/15careers.html

SUCCESSES

- Some poignant factors:
- Categorical vs Non-categorical classification
- Masks = Filter
- **Extra Trees Classifier**
  - Decision boundaries picked at random
  - Computationally more efficient

```
In [33]:  1 accuracy_score(y_test, rf.predict(X_test))
Out[33]: 0.031481481481481478
```

```
In [93]:  1 accuracy_score(u_test, et.predict(X_test))
Out[93]: 0.21389793702497287
```



ANOMALIES

- Adding more masks “should” help
  - Adding on a mask including games started
  - Halves the number of players
  - Same or less accuracy???
  - Issue with overfitting?
    - Overfitting = overly complex model

```
In [77]:  1 accuracy_score(u_test, et.predict(X_test))
Out[77]: 0.18518518518518517
```

When added with masks for both career length and games played

SUMMARY

- Link prediction can be determined to an extent, and perhaps further.
- By adjusting our decision algorithms, we can significantly improve accuracy

- Future Plans:
- We need to test our model with other similar situations
  - Corporate employees
  - Other Sports teams
  - Salespersons headhunted in certain businesses
  - Where Prominent musicians may play
- Unify probabilities of team departure and team destination

THANK YOU