NBA Oracle

Matthew Beckler, Hongfei Wang
Department of Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh, PA 15213
{mbeckler, hongfeiw}@ece.cmu.edu

Michael K. Papamichael
Department of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213
papamix@cs.cmu.edu

NBA Predictions & Decisions Hard

- **Professional basketball and sports betting are multi-billion dollar industries.**
  - Sophisticated decision support crucial for player acquisitions and trading, sponsorship decisions, coaching strategy, MVP nominations, All-star player selection, etc.
- **Use ML techniques for**
  - Game outcome prediction (based on past games.)
  - Outstanding player detection (for draft and trades.)
  - Optimal player position selection
- **Data from http://www.databasebasketball.com**
- **Difficult Problem Domain**
  - Many sources of randomness, such as player injuries, player attitudes, team rivalries, subjective officiating, and others.

Machine Learning to the Rescue

1 Data Retrieval, Conditioning and Processing

SQL

2 Game Outcome Prediction

Linear Regression | Support Vector Machines | Logistic Regression | Artificial Neural Network

3 Coaching and Drafting Decision Support

K - Means | Outlier Detection

Results & Contributions

- **Experimentation with many ML Methods**
- **Looked at various ML-related problems**
  - Binary Classification
  - Clustering - Outlier Detection
- **Augmentation/Validation of existing dataset**
  - Gathered detailed data on a per game/player basis.
  - Used gathered data to cross-validate existing stats database. Errors reported to dataset maintainers.
  - New datasets useful for future ML researchers.
- **Results comparable to established work**

Random Guess	Sports Experts	Website Guarantees	Other Researchers	NBA ORACLE
50%	71%*	65%	70%	Up to 73%

*Human experts can decline to predict very close games.

1 Data Retrieval, Conditioning, and Processing

Data from www.databasebasketball.com

Data Acquisition. Original downloaded data files contained cumulative stats of players and teams, season-by-season. Also retrieved game results from website, with detailed stats for each individual player per every game. Focused mainly on seasons 1991-1992 through 1996-1997.

Conditioning

Cleaning and Standardizing the Raw Data. Checked the validity and completeness of all data files. Eliminated parsing errors and fixed errant data values. Validated data against itself to ensure consistency between individual stats and accumulated team stats.

Database Support

SQL

Utilizing the Syntactic Power of SQL. Imported data from text files into SQL database using Python scripts. Flexibility of SQL queries allowed complex joins and merges to complete with ease. Data accessible from shell scripts as well as Python scripts.

Processing

Distillation of Data into the Most Useful Form. Individual player statistics were accumulated for each team over all games to produce cumulative statistics for each team, at each time-step. This cumulative data provides us with the most current and accurate statistics for predicting the next game.

2 Game Outcome Prediction

Predicting the outcome of a game is a problem of binary classification, choosing between two possible outcomes. To normalize our data values, all classifiers operated on ratios of statistics, essentially making all features unit-less. Mutual-information analysis was performed on all input features, showing that defensive stats were the most influential, right after the number of wins/losses. Defensive stats are stats earned by a team's opponent, such as points scored against or rebounds by the other team. Since the classification algorithms were not computationally complex, they were simply run with all available input features. For each classification algorithm tested, 100-fold cross validation (CV) was performed to ensure an accurate result.

Linear Regression

- **One of the most fundamental ML generative models that estimates the output as a linear combination of the input.**
- **Find the optimal weights w_0, …, w_n that maximize:**

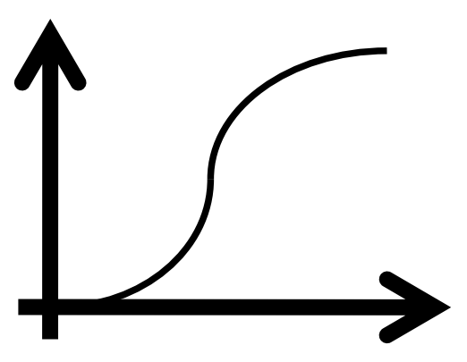
$$Y = w_0 + \sum_{i=1}^{n} w_i x_i$$

- **Implemented from scratch in Matlab.**
- **Achieved best accuracy results. Simple works!**

Logistic Regression

- **Logistic regression is a widely used technique for classification used in statistics and ML**
- **Replace the linear function with the sigmoid function:**

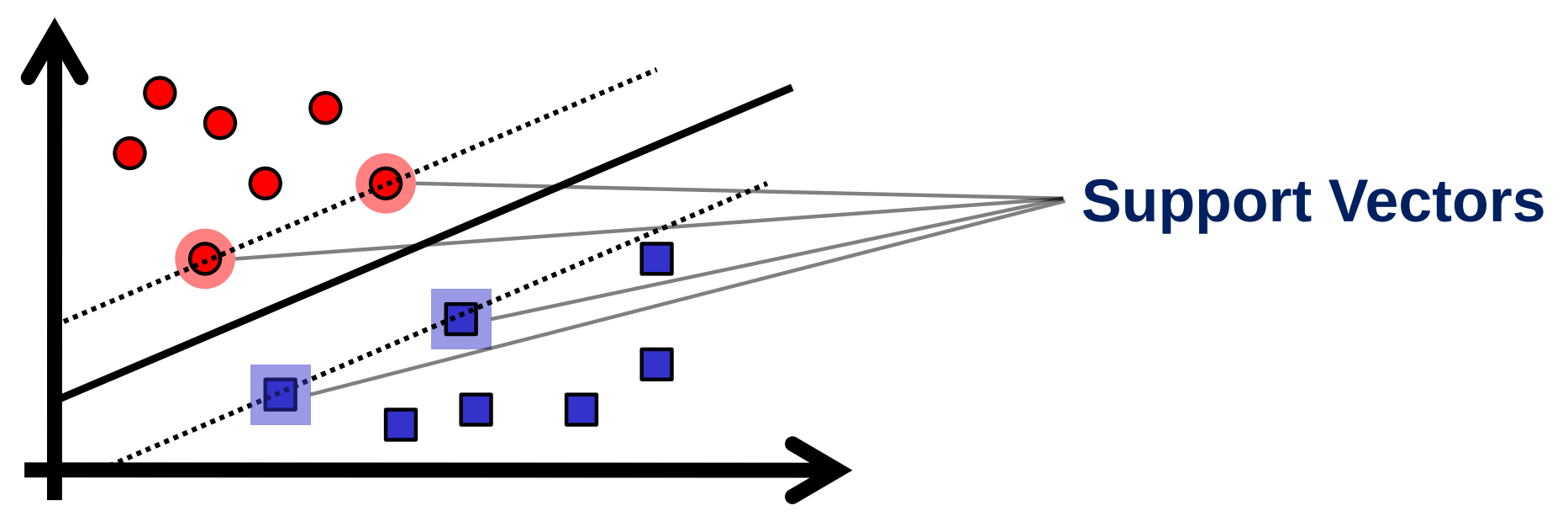
$$f(w^T x) = \frac{1}{1+e^{-w^T x}}$$



- **MLE to find optimal w values. But no closed form solution!**
- **MLE produces concave function → use gradient descent**
- **w converges after 200-300 iterations (ε = 0.01)**

Support Vector Machines

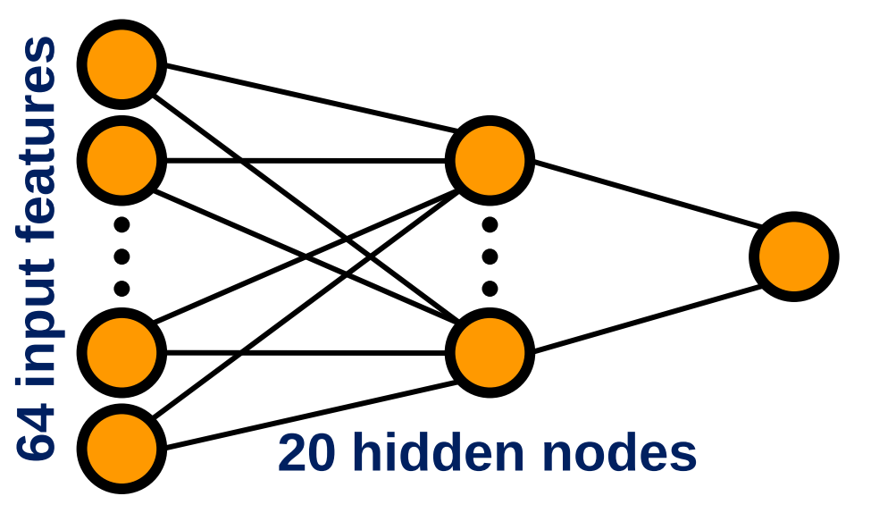
- **Support Vector Machines (SVM) are classifiers that maximize the margin between two classes.**



- **Used existing C++ tinySVM**
- **Achieved good results, but accuracy suffered from non-linearly separable dataset and presence of noise**

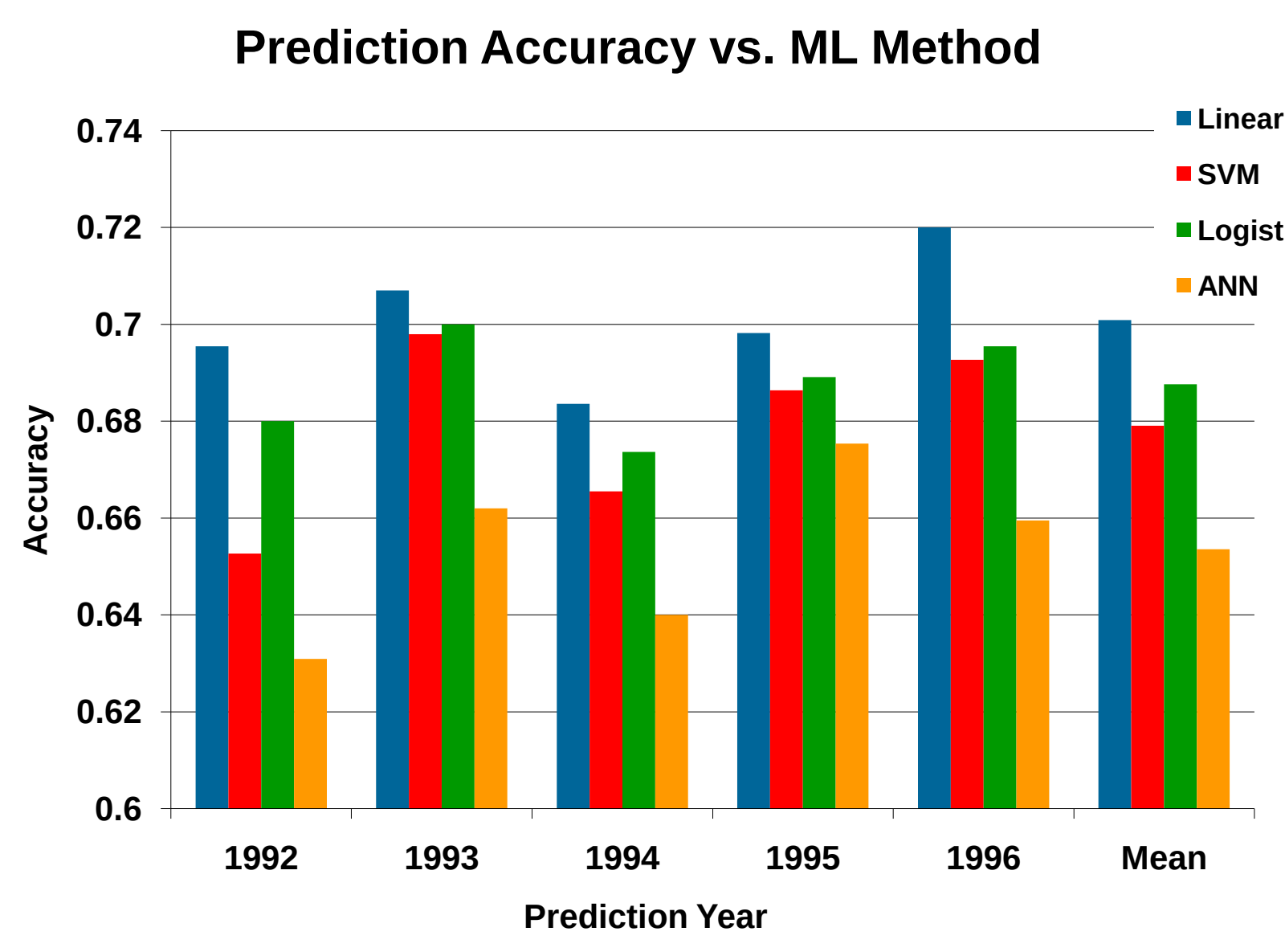
Artificial Neural Networks

- **Neural Networks provide a general method for learning hidden internal relationships of the input data to learn the target function.**
- **A Each hidden layer node implements a sigmoid function.**
- **After experimentation a single hidden layer with approximately 20 nodes produced the best results**

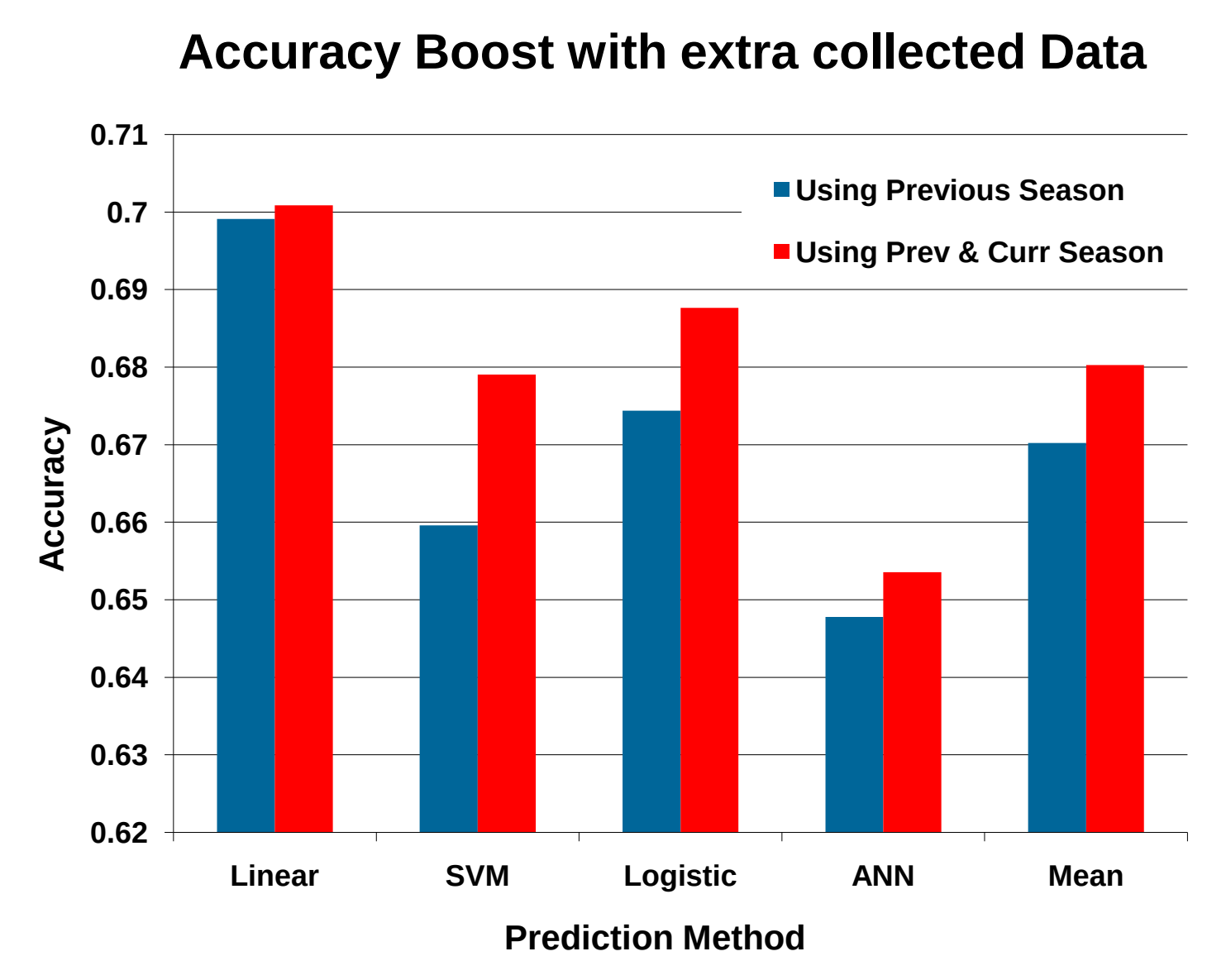


- **Used the Matlab Neural Network Toolbox Library**
- **Achieved modest average accuracy of 65.36%**

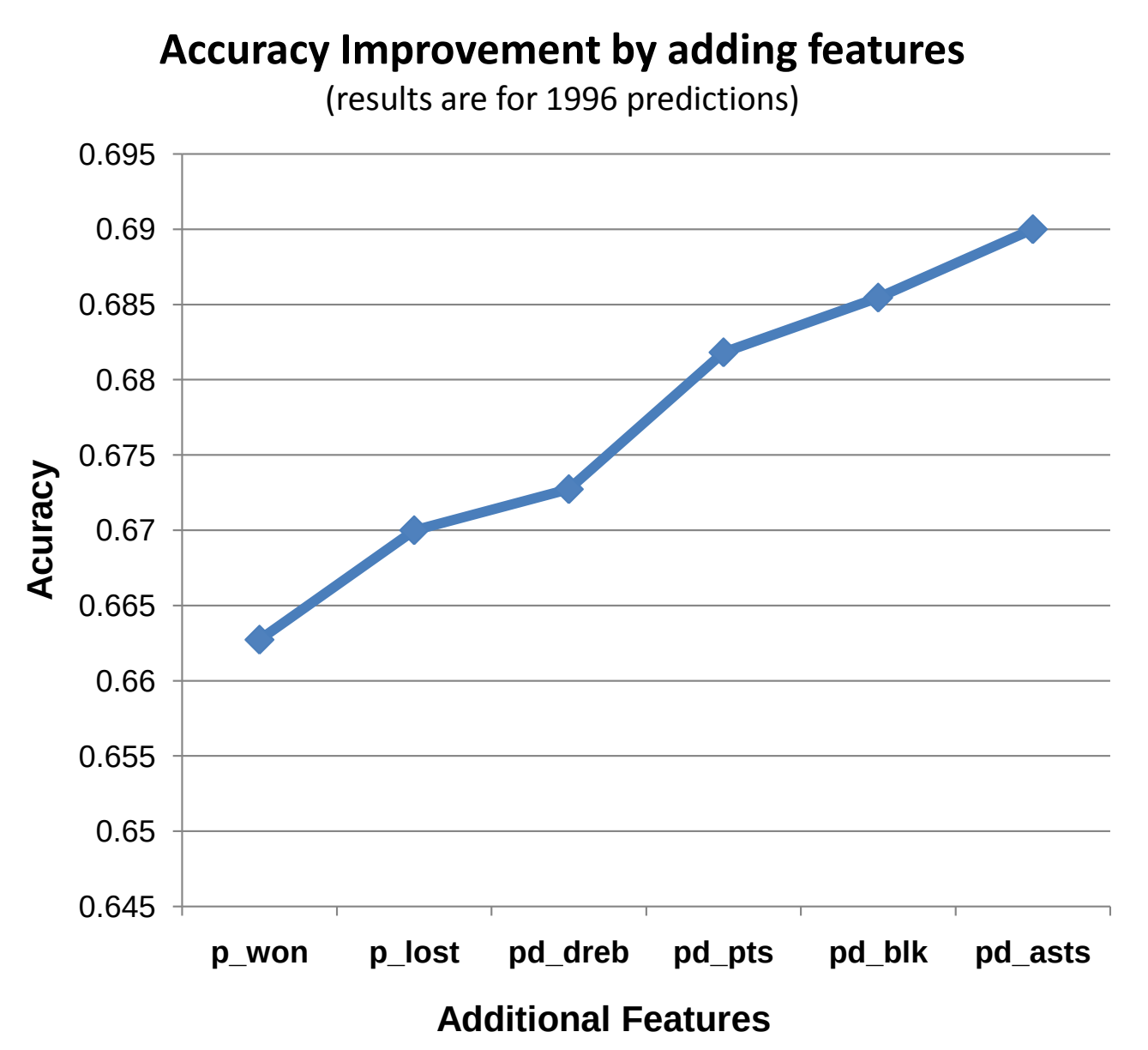
Game Prediction Results



Simple works!



Extra effort to collect data worth it!

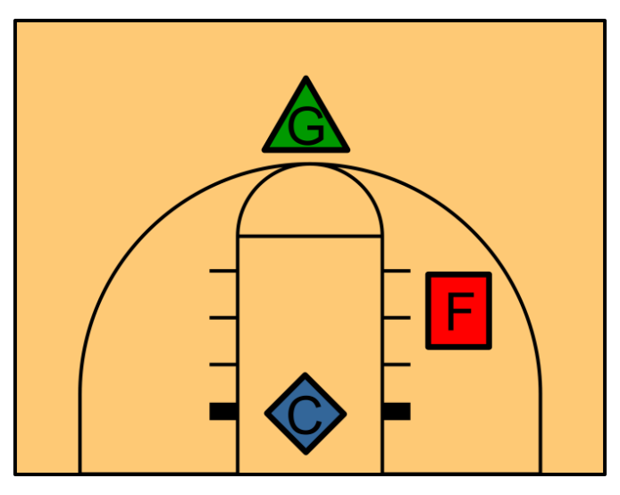


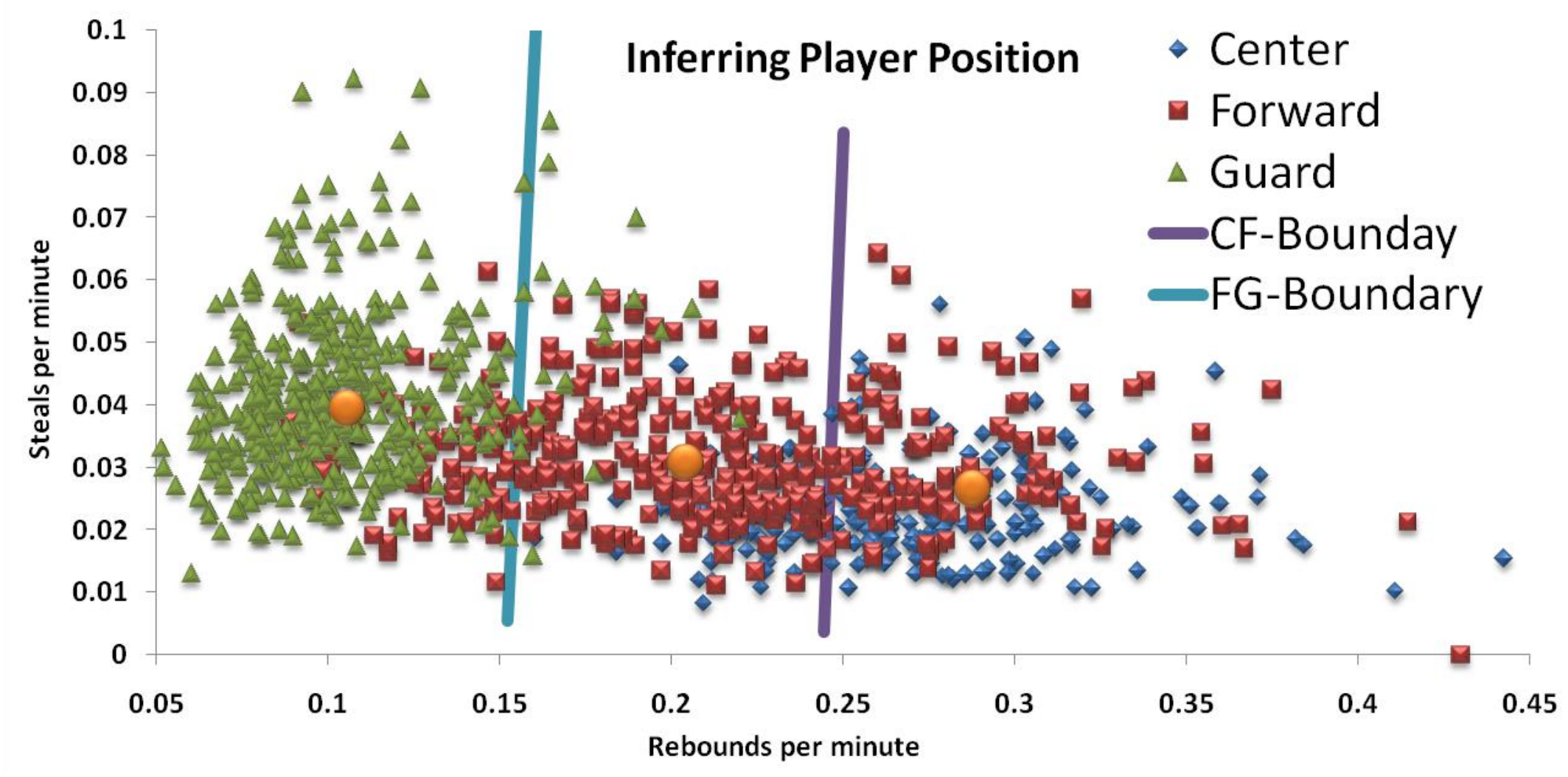
More features boost accuracy

3 Decision Support

K – Means Clustering

- **K-Means clustering finds naturally-occuring clusters in a dataset.**
- **Each basketball player is either a Center, Forward, or Guard. We used K-Means clustering to infer the position of a player.**
- **Used exhaustive search to find best pair of features: Rebounds and Steals**

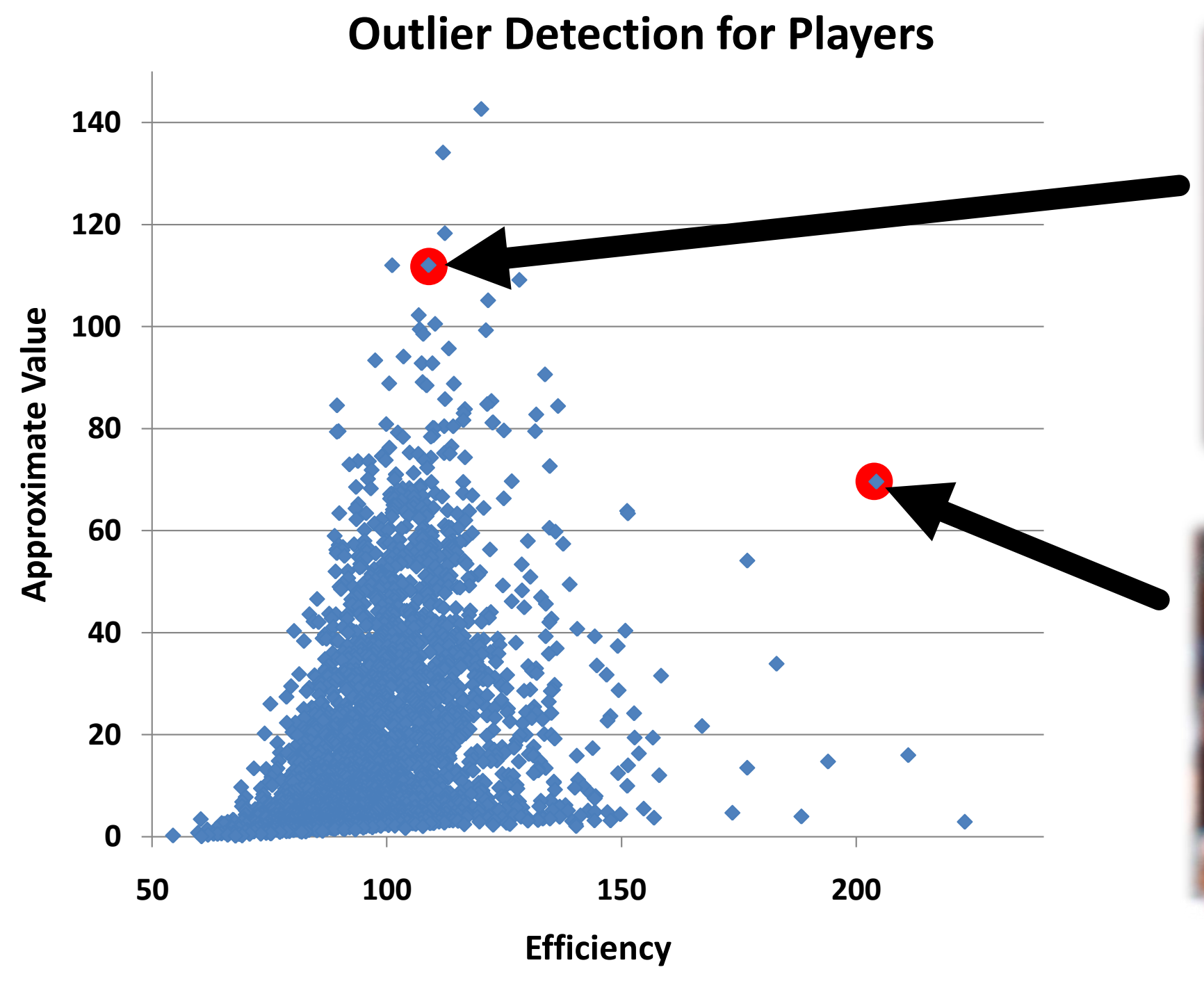


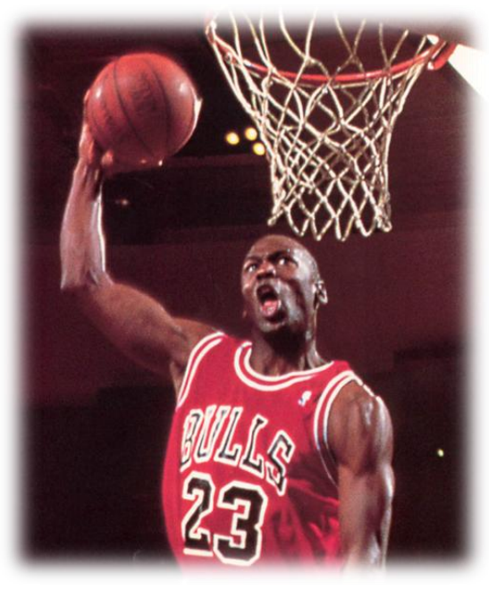


- **Player position inference accuracy over 75%!**
- **Results make sense intuitively; Rebounds and steals are strongly correlated with player position on court.**

Outlier Detection

- **Outlier detection useful for identifying exceptional players and detecting invalid data**





- **Used 2 expert-derived metrics to evaluate player performance**
  - **Efficiency: short-term per game performance**
  - **Approximate Value: long-term season performance**
- **Majority of outliers are exceptional players**
  - **Successfully identified most MVP players**

Acknowledgements

We would like to thank Prof. Ziv-Bar-Joseph for his feedback, as well as the Machine Learning Course Teaching Assistants for their help.