Picking Intro

School of Information Studies
Syracuse University

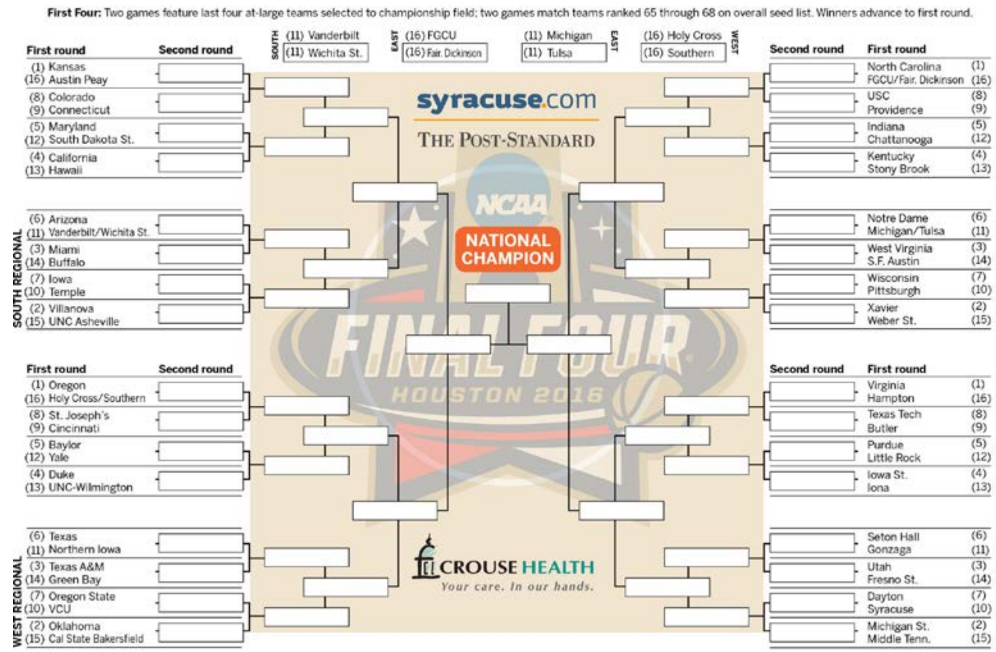
Patterns in Games and Prices

- Obtain
- Scrub
- Explore
- Model
- iNterpret

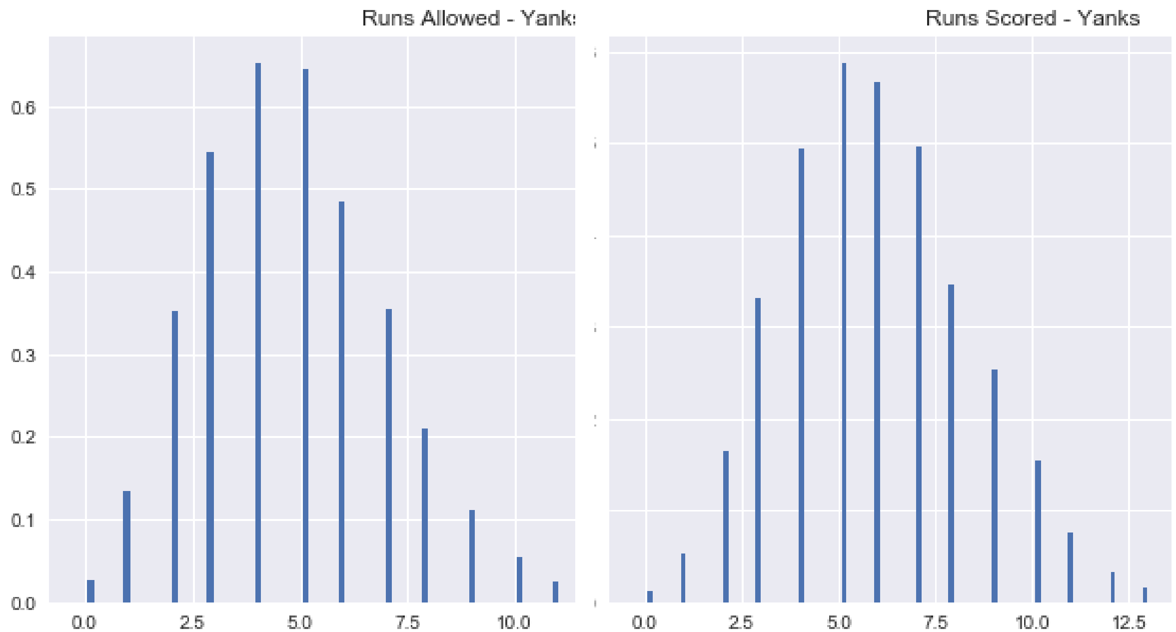
School of Information Studies
Syracuse University

Our Challenge This Week?

Using Distributions to Pick a Winner



But How?

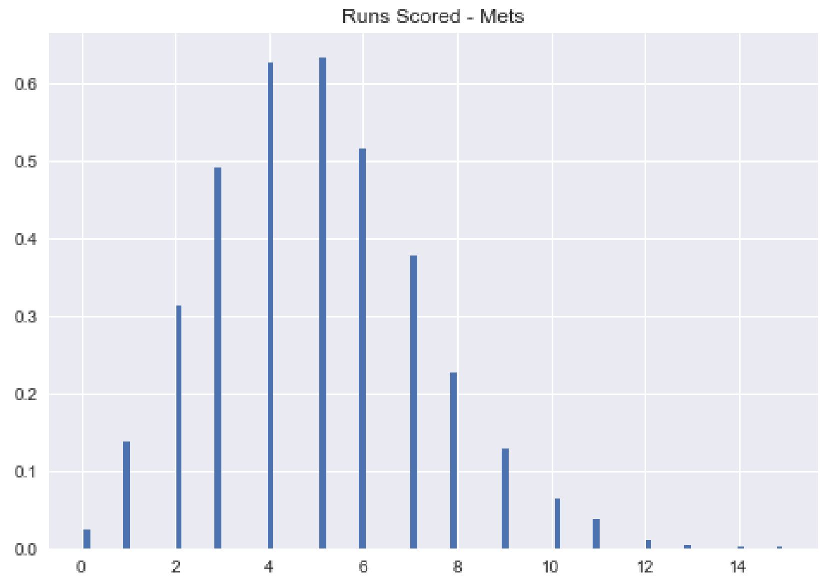


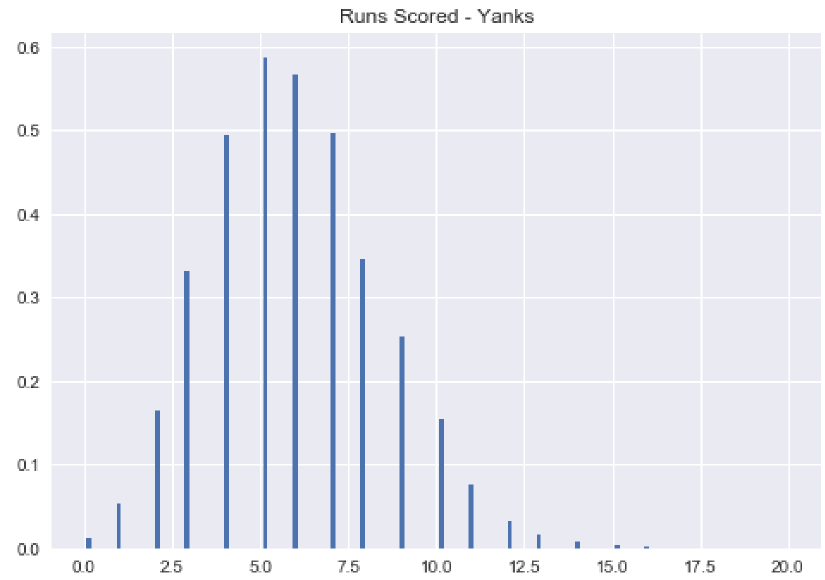
Runs Allowed - Yank:
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
2.5
5.0
7.5
10.0
Runs Scored - Yanks
0.0
2.5
5.0
7.5
10.0
12.5

Data Review

School of Information Studies
Syracuse University

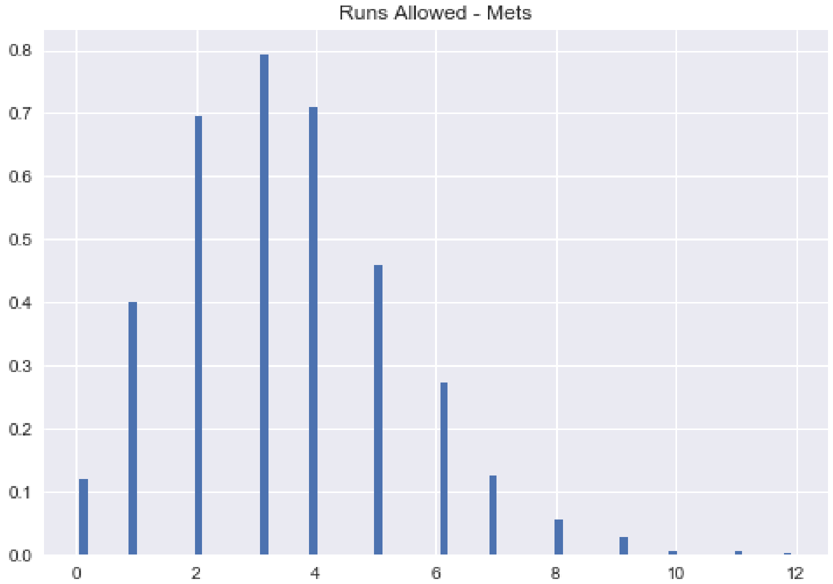
Runs Scored

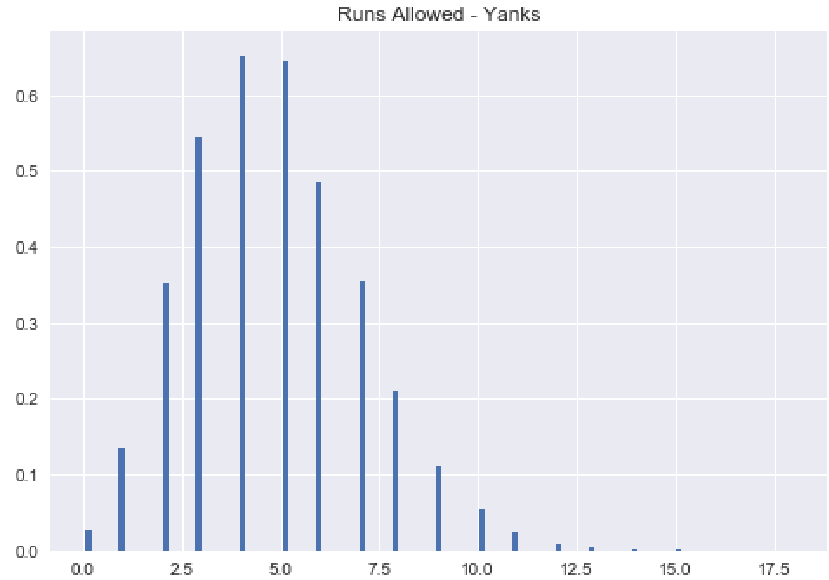




School of Information Studies
Syracuse University

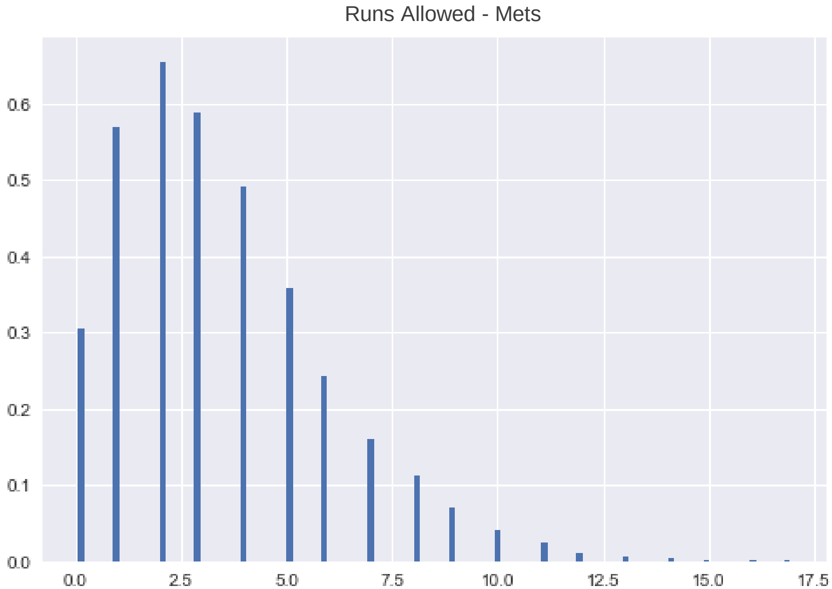
Runs Allowed

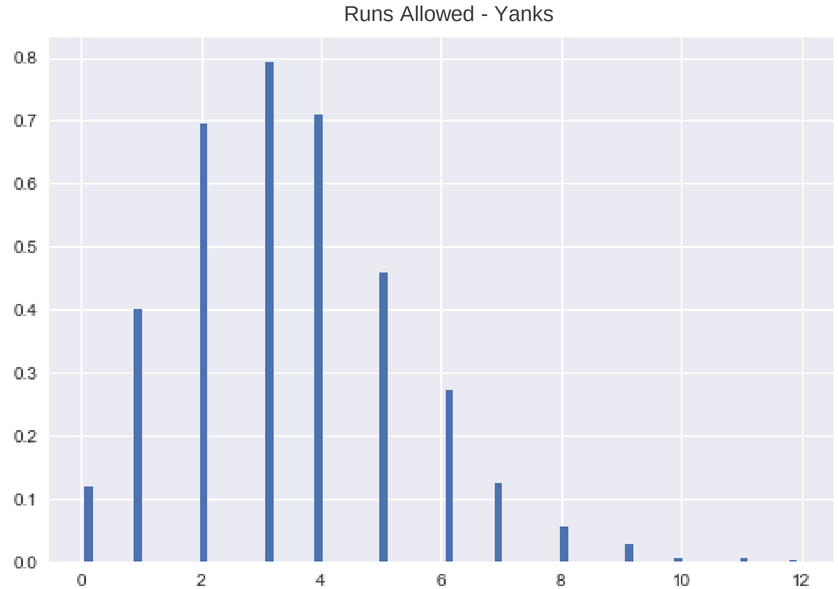




School of Information Studies
Syracuse University

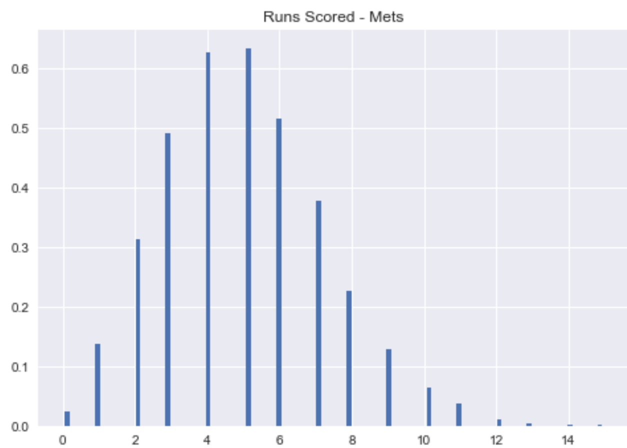
Scoring Distributions

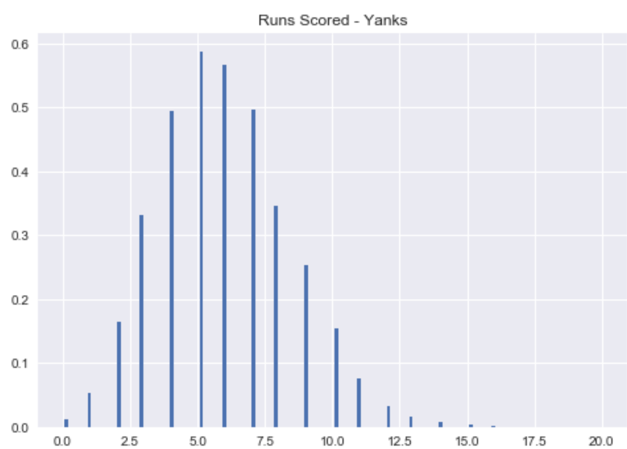


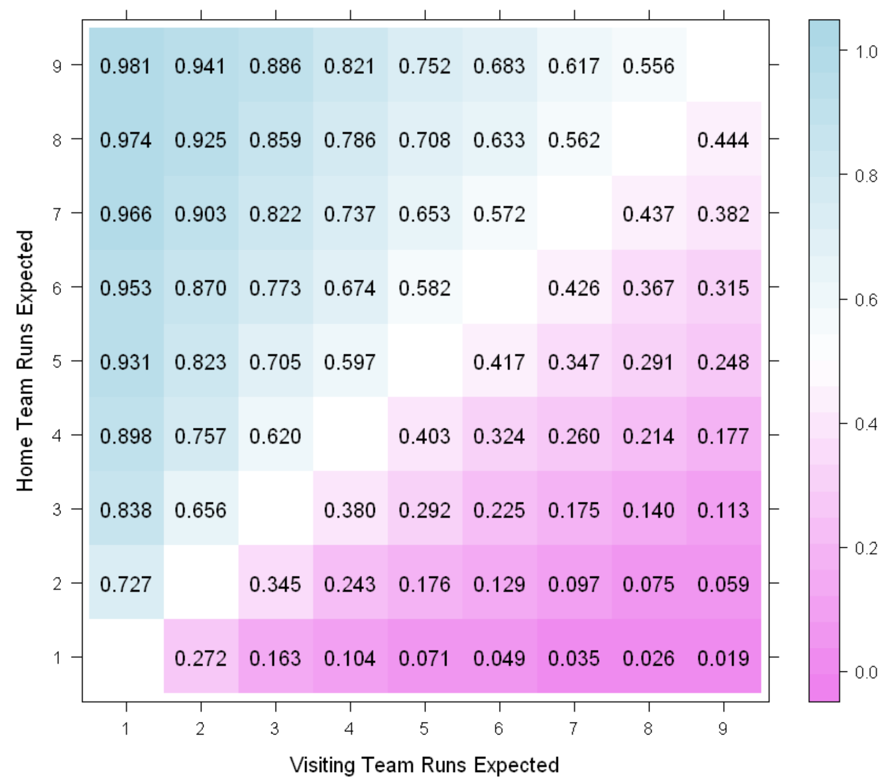


School of Information Studies
Syracuse University

Picking Winners



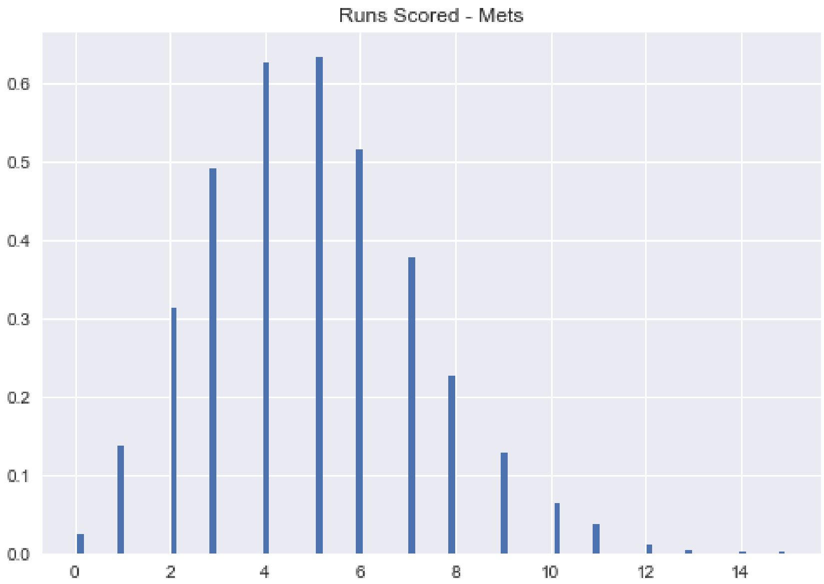


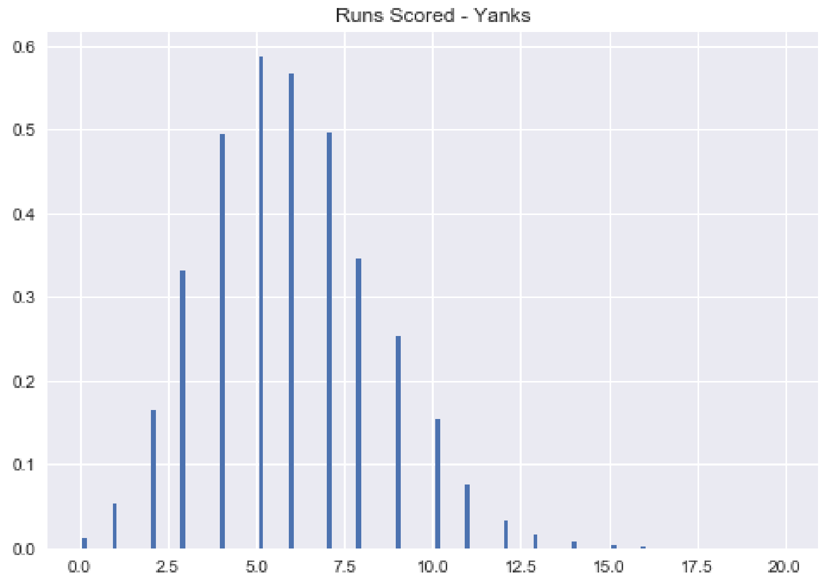


Poisson Distribution

School of Information Studies
Syracuse University

Simulating Runs Scored

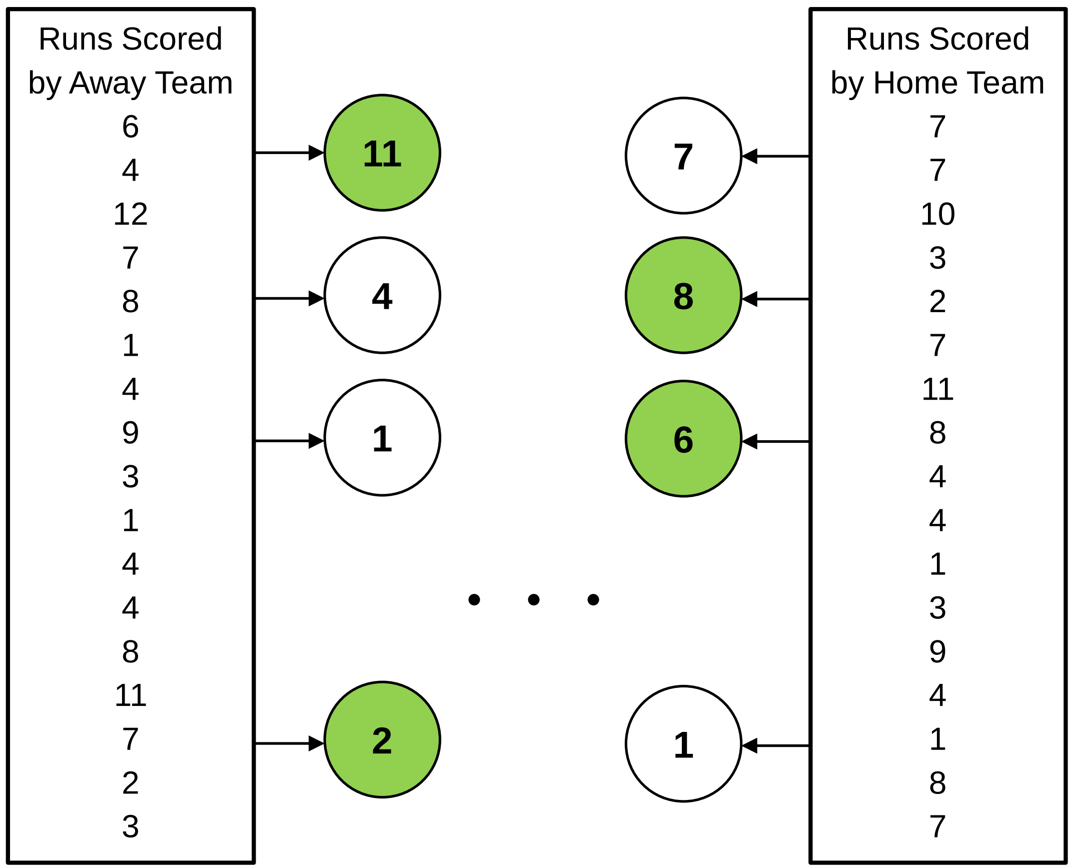




School of Information Studies
Syracuse University

Sports Simulation

Simulation Results

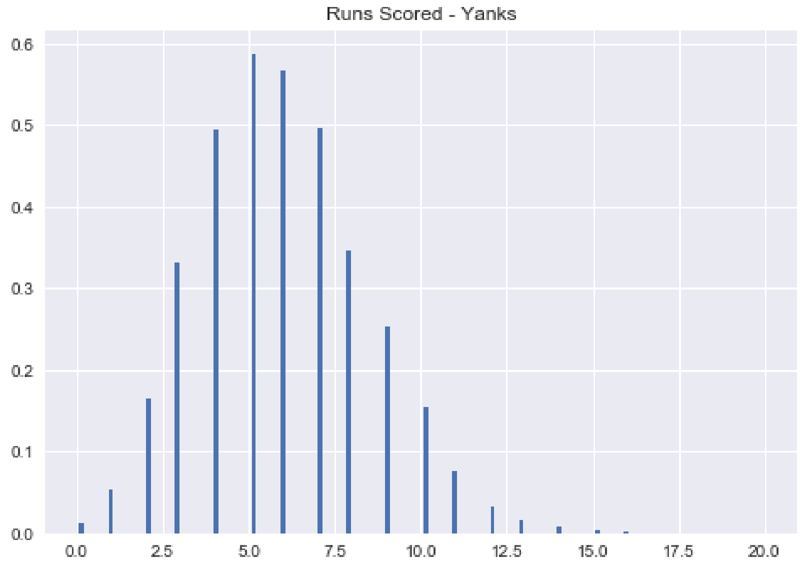


Source: Adapted from Miller (2005).

Poisson Distribution

- Good approximation for count responses

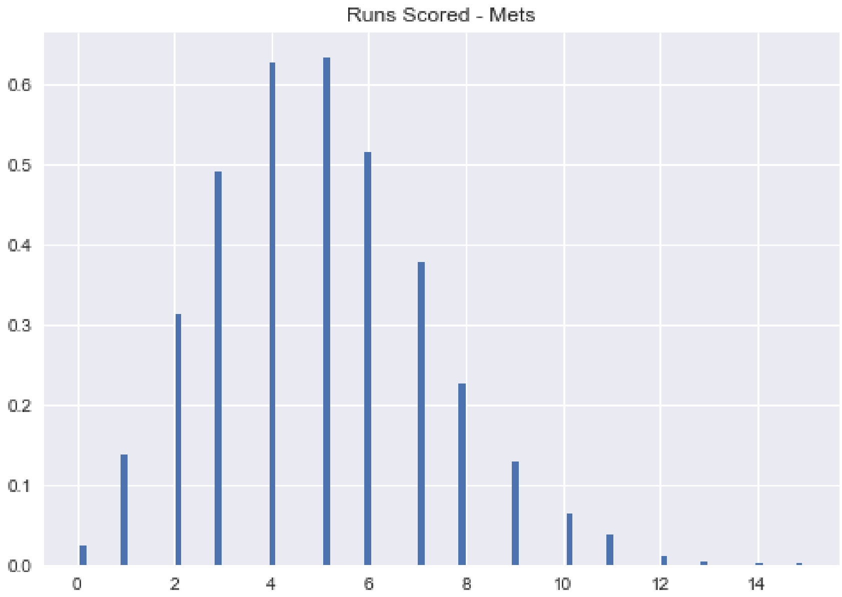
$$P(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}$$



- Occurrence of events during certain time interval
- Arrival rate problems

Poisson Distribution (cont.)

```
MetAwayScore =
np.random.poisson(4.97, 10000)

MetAwayDefend =
np.random.poisson(3.45, 10000)

YankHomeScore =
np.random.poisson(5.97, 10000)

YankHomeDefend =
np.random.poisson(4.84, 10000)

plt.hist(MetAwayScore, bins='auto',
rwidth = .5, normed=True)

plt.title("Runs Scored – Mets")
plt.show()
```

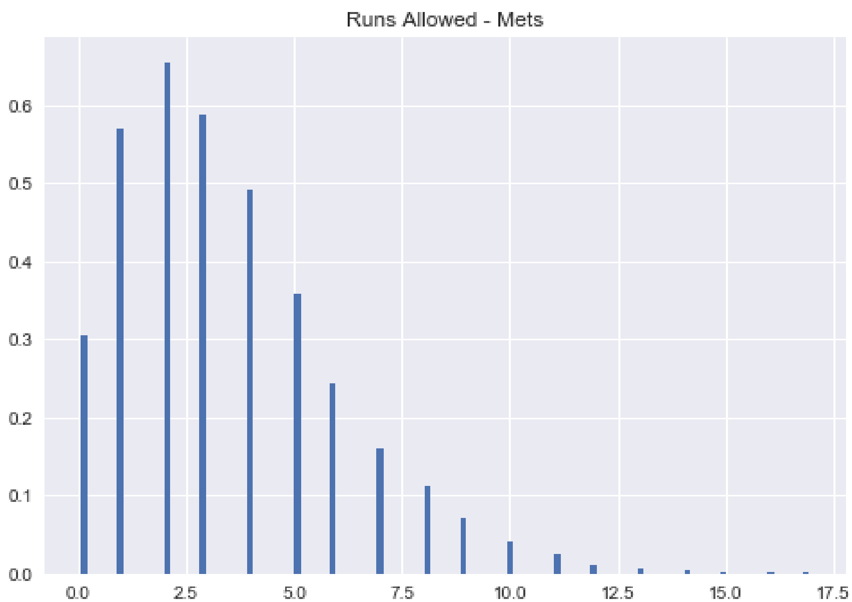


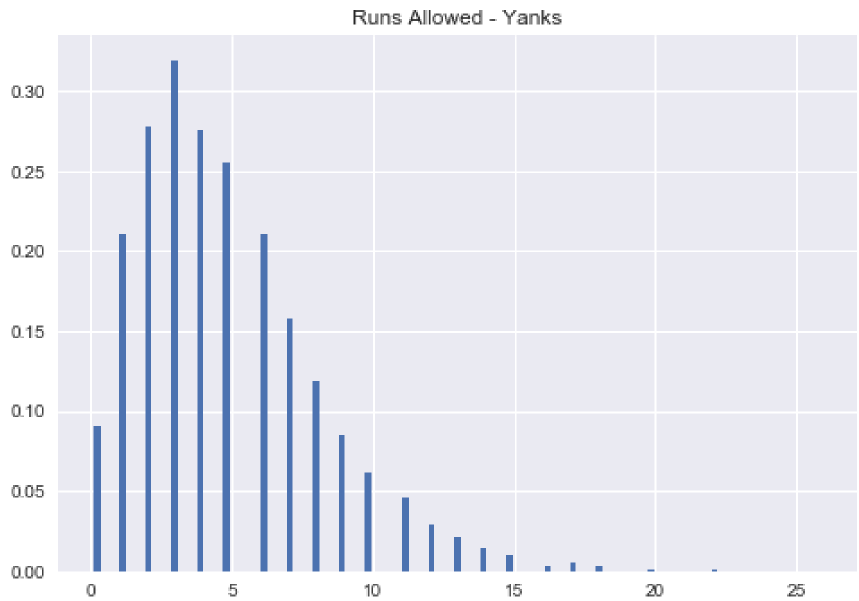
School of Information Studies
Syracuse University

Negative Binomial Distribution

School of Information Studies
Syracuse University

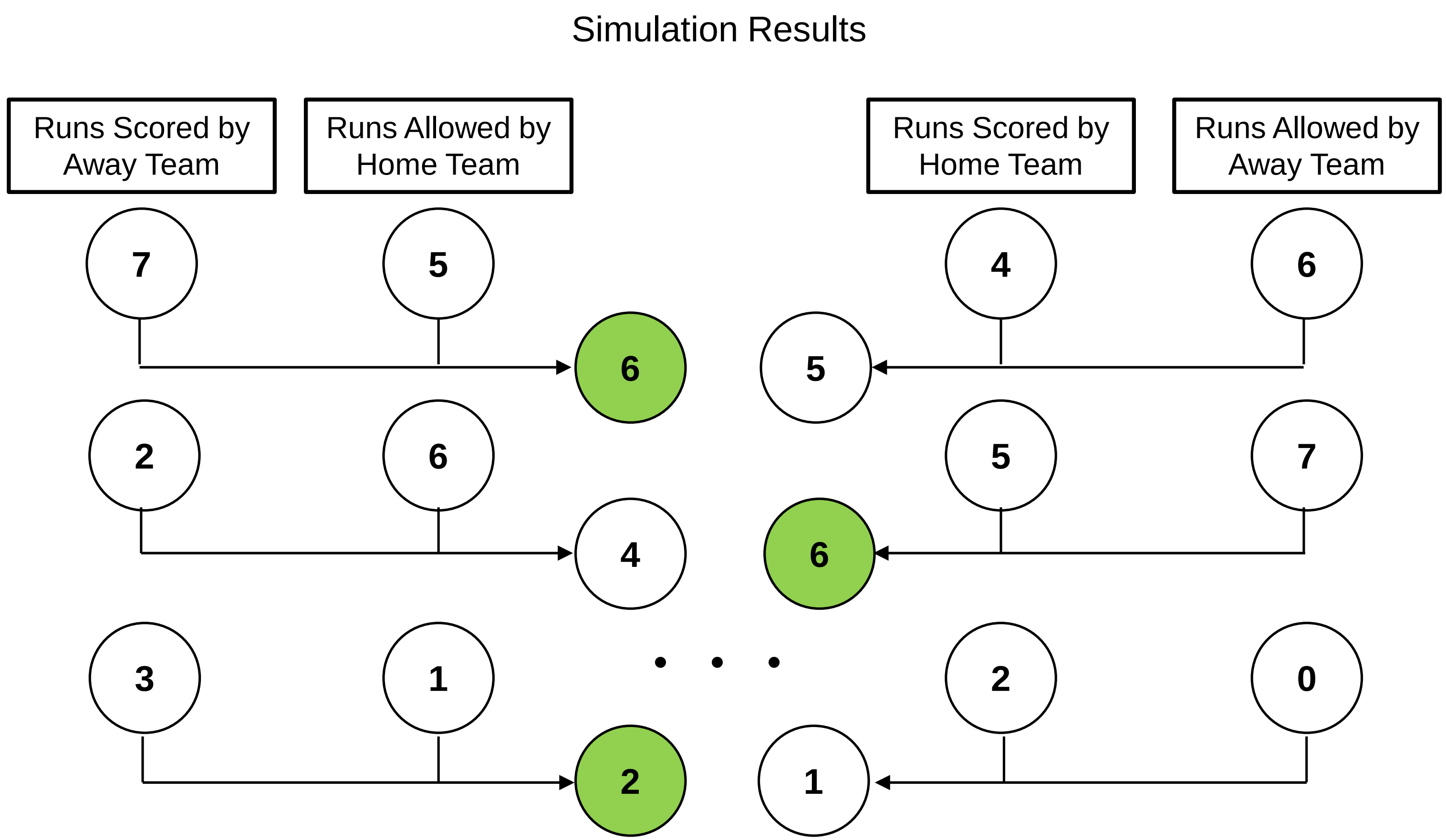
Simulating Runs Allowed





School of Information Studies
Syracuse University

Sports Simulation

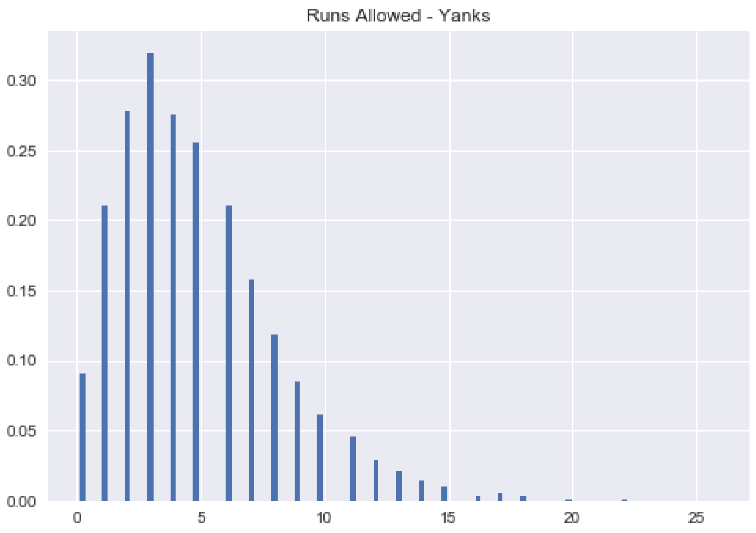


Source: Adapted from Miller (2005).

Negative Binomial

- Alternative approximation for count responses

$$P(Z = z) = \binom{z-1}{k-1} p^k (1-p)^{z-k}$$



- Generalization of Poisson distribution
- Rare event problems

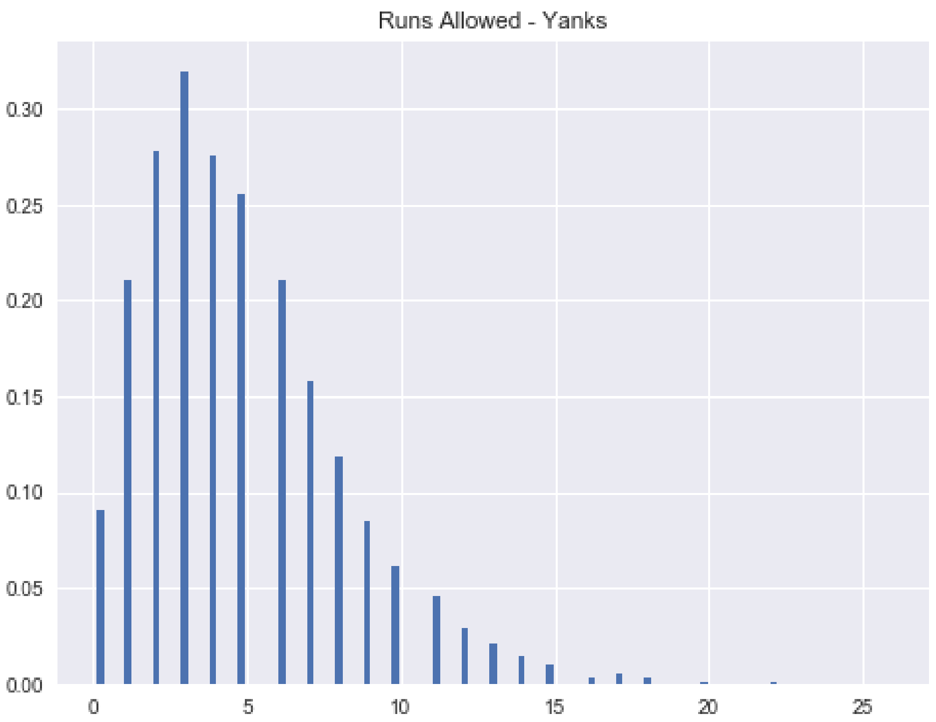
Negative Binomial

- Alternative approximation for count responses

$$P(Z = z) = \binom{z-1}{k-1} p^k (1-p)^{z-k}$$

- Generalization of Poisson distribution

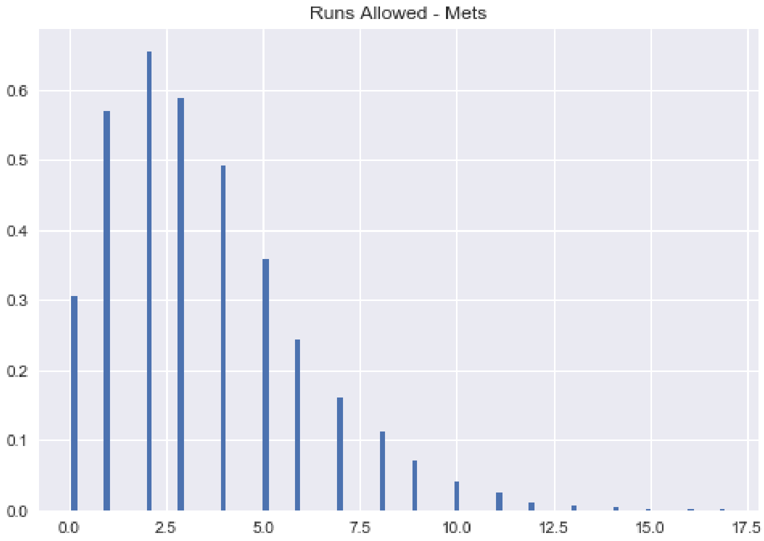
- Rare event problems



School of Information Studies
Syracuse University

Negative Binomial (cont.)

```
MetAwayScore =
np.random.negative_binomial(4, mas, 10000)
MetAwayDefend =
np.random.negative_binomial(4, mad, 10000)
YankHomeScore =
np.random.negative_binomial(4, yhs, 10000)
YankHomeDefend =
np.random.negative_binomial(4, yhd, 10000)

plt.hist(MetAwayScore, bins='auto', rwidth = .5,
normed=True)
plt.title("Runs Scored – Mets")
plt.show()
```



School of Information Studies
Syracuse University

Applications

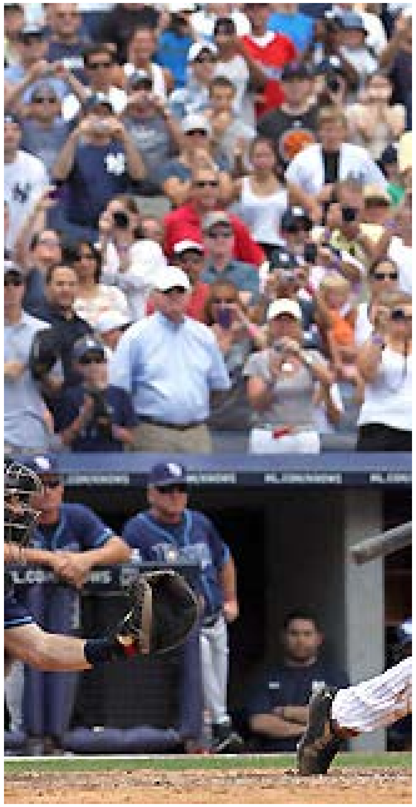
School of Information Studies
Syracuse University

Additional Applications

School of Information Studies
Syracuse University

Baseball Prospectus

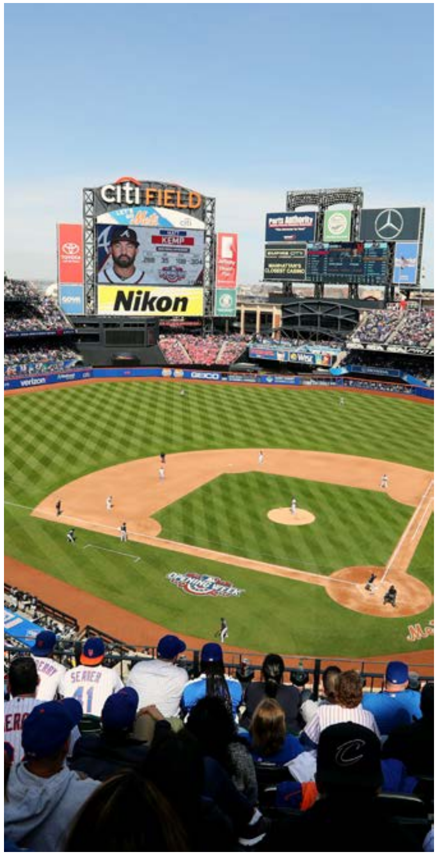
- Predicting performance before the season
  - Nate Silver
  - PECOTA
- Variations
  - Military war games
  - Film releases
  - Associate performance



School of Information Studies
Syracuse University

Moneyball Problem

- Calculating the value of a player
  - Billy Beane
  - Individual summary stats applied to team performance
- Variations
  - Associate performance
  - Client conversion
  - Customer lifetime value



School of Information Studies
Syracuse University

Coaching Problems

- Utilizing player performance
  - Microanalysis of game
  - Sabermetrics
- Variations
  - Football
  - Basketball

Bowl Championship Series

- Predicting team performance against unknown opponent
  - Strength of schedule
  - Ensemble approaches
- Variations
  - March Madness
  - Product deployment
  - Recommendation engine

Billy Waters Problem

- Predicting the winning team in the next game?
  - Human expertise
  - Simulation
- Variations
  - March Madness
  - Film release
  - Product deployment

Picking II Intro

School of Information Studies
Syracuse University

Patterns in Games and Prices

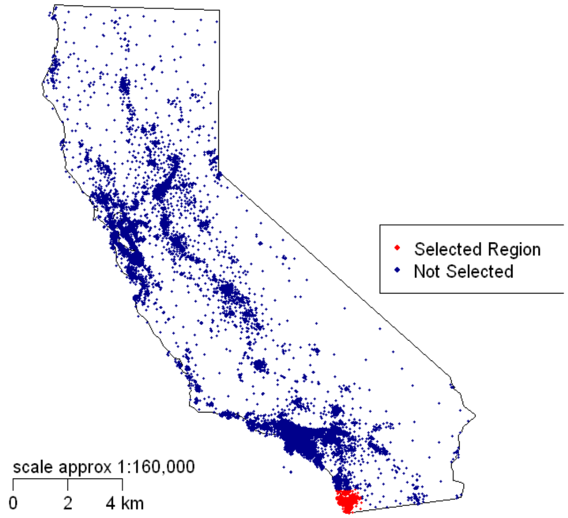
- Obtain
- Scrub
- Explore
- Model
- iNterpret

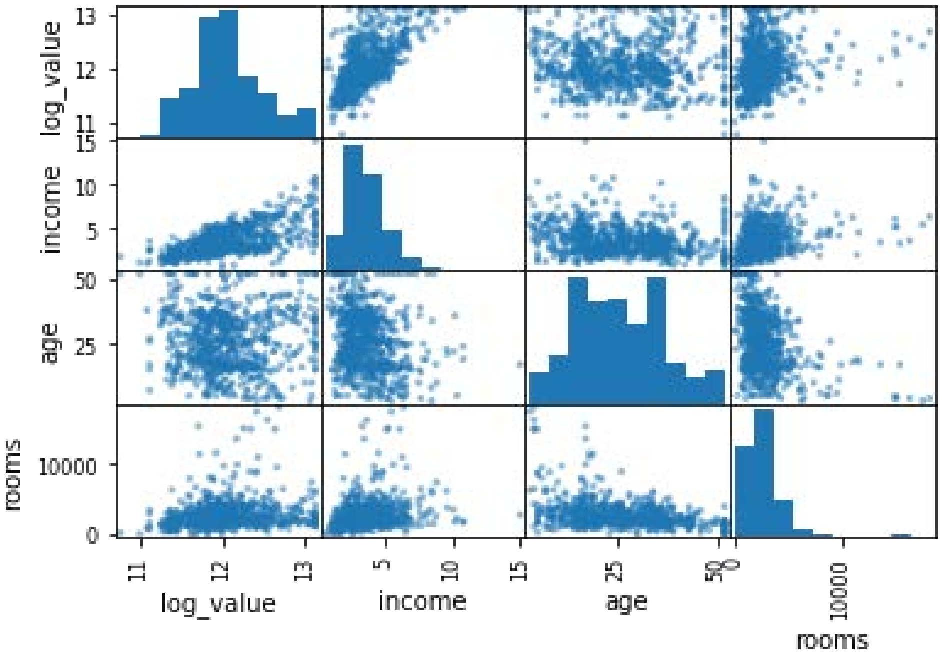
Our Challenge Now?

School of Information Studies
Syracuse University

Modeling Housing Prices

But How?

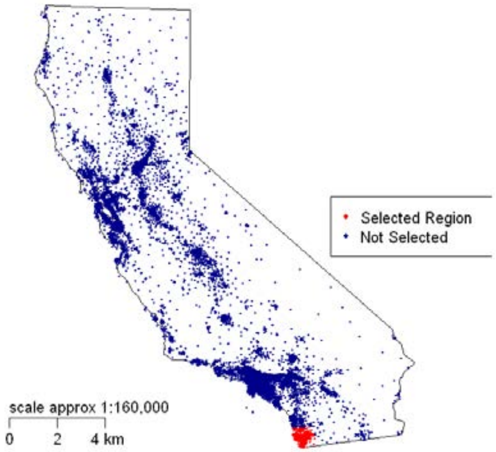

Selected Region
Not Selected
scale approx 1:160,000
0
2
4 km


log_value
income
age
rooms

Data Review II

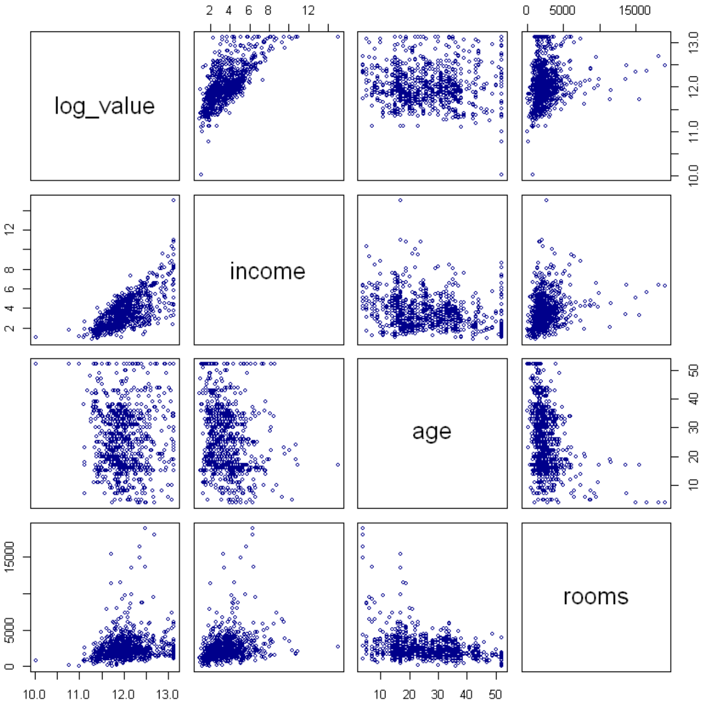
School of Information Studies
Syracuse University

Housing Data



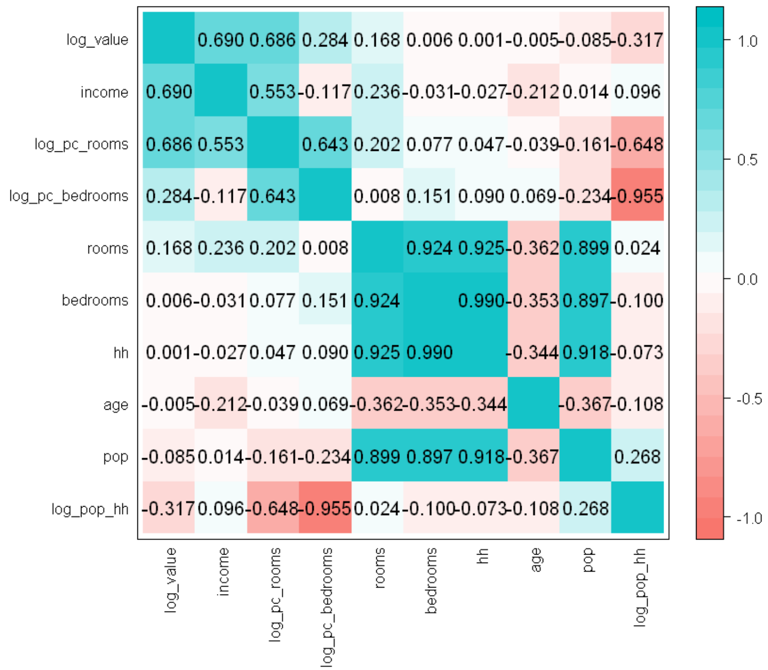
School of Information Studies
Syracuse University

Feature Correlation



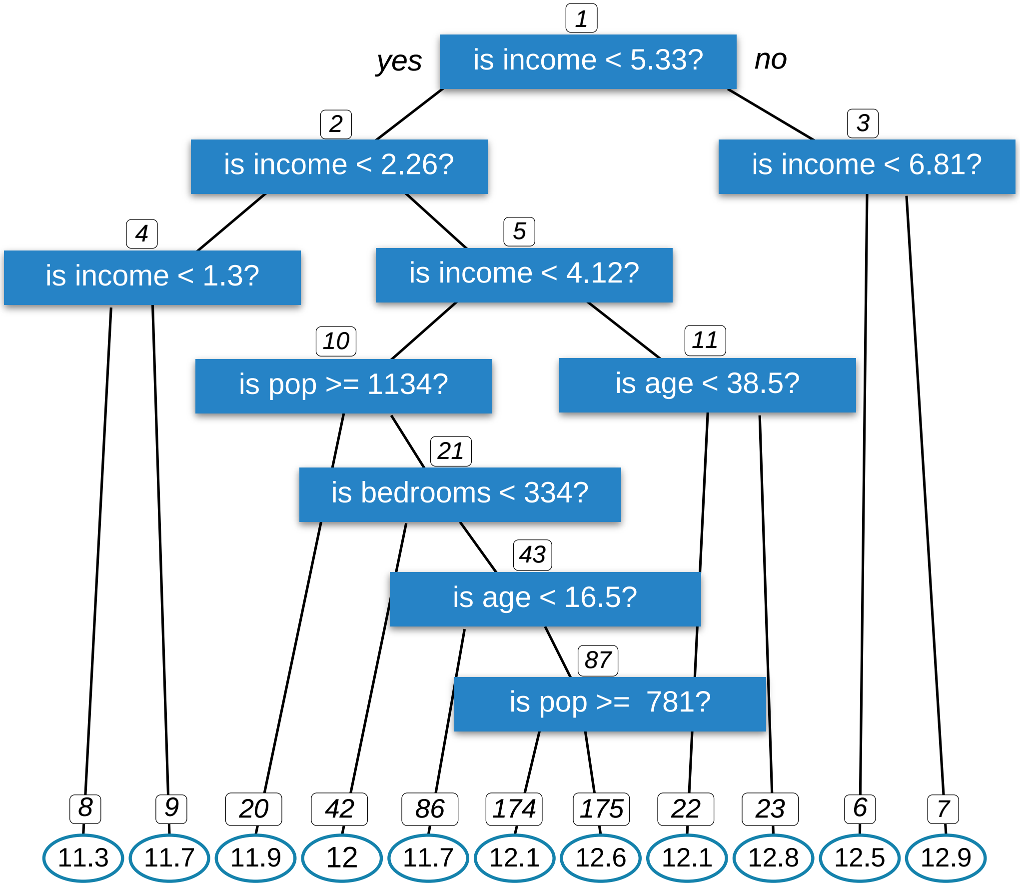
School of Information Studies
Syracuse University

Feature Correlation (cont.)



School of Information Studies
Syracuse University

Picking Values

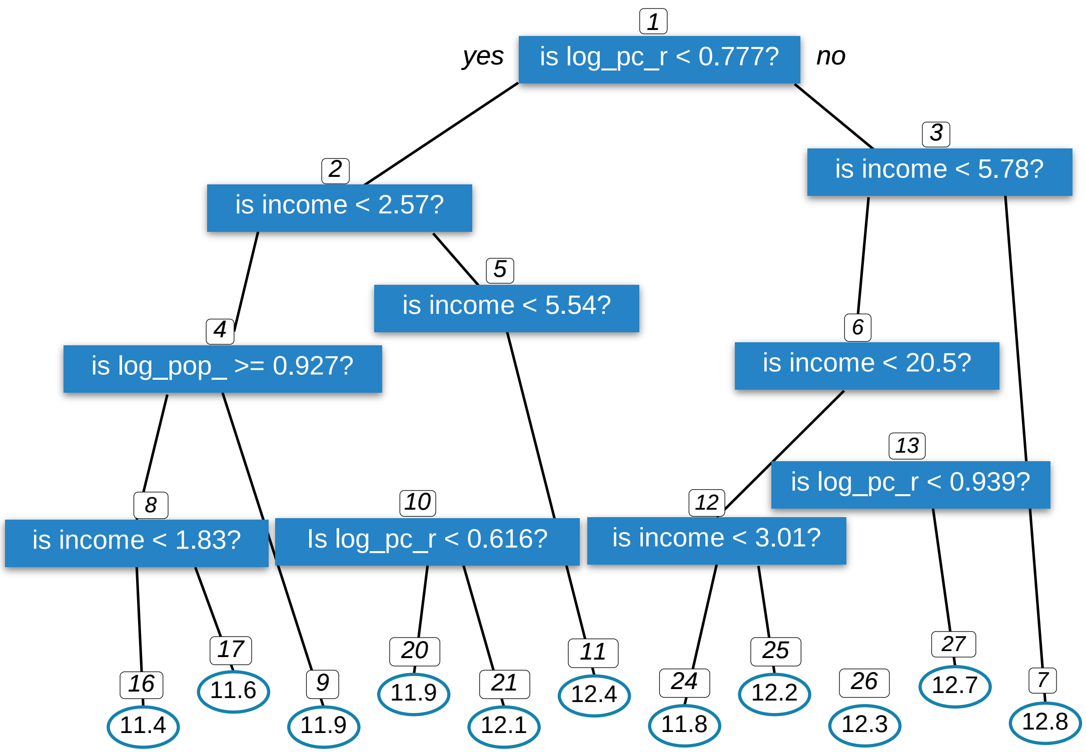


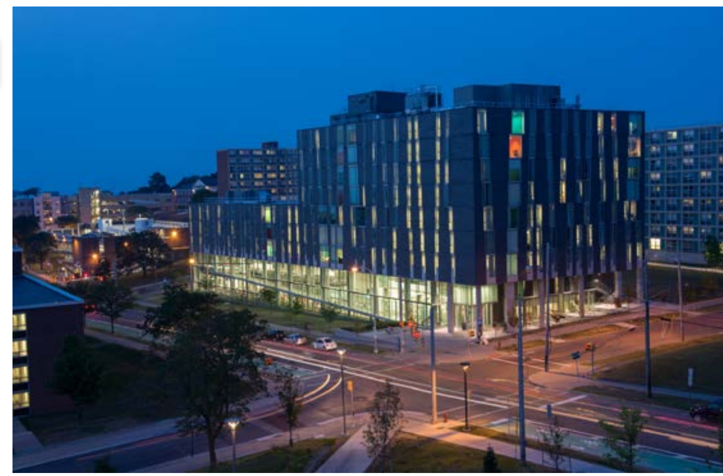
School of Information Studies
Syracuse University

Trees Forests

School of Information Studies
Syracuse University

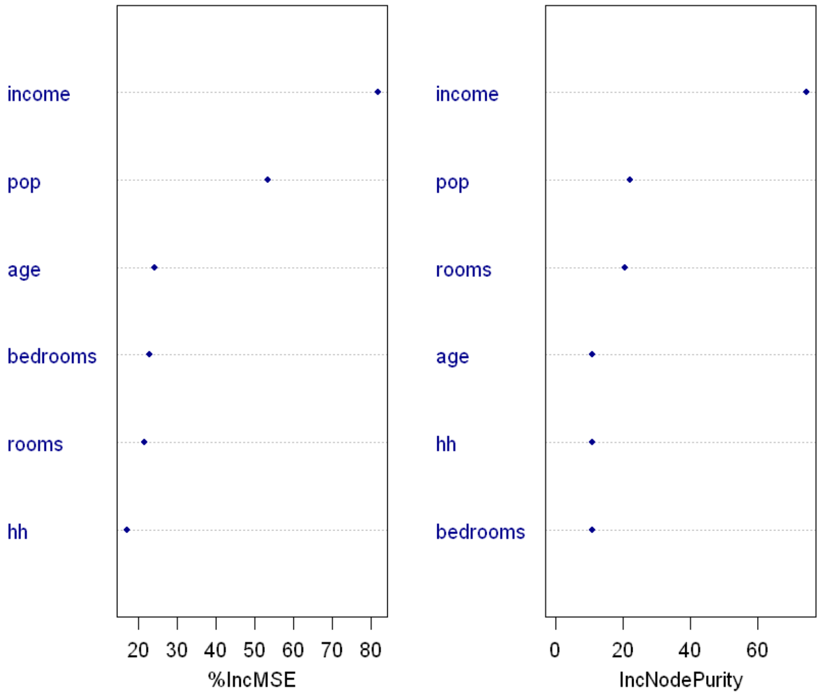
Picking a Tree in the Forest





Decision Trees

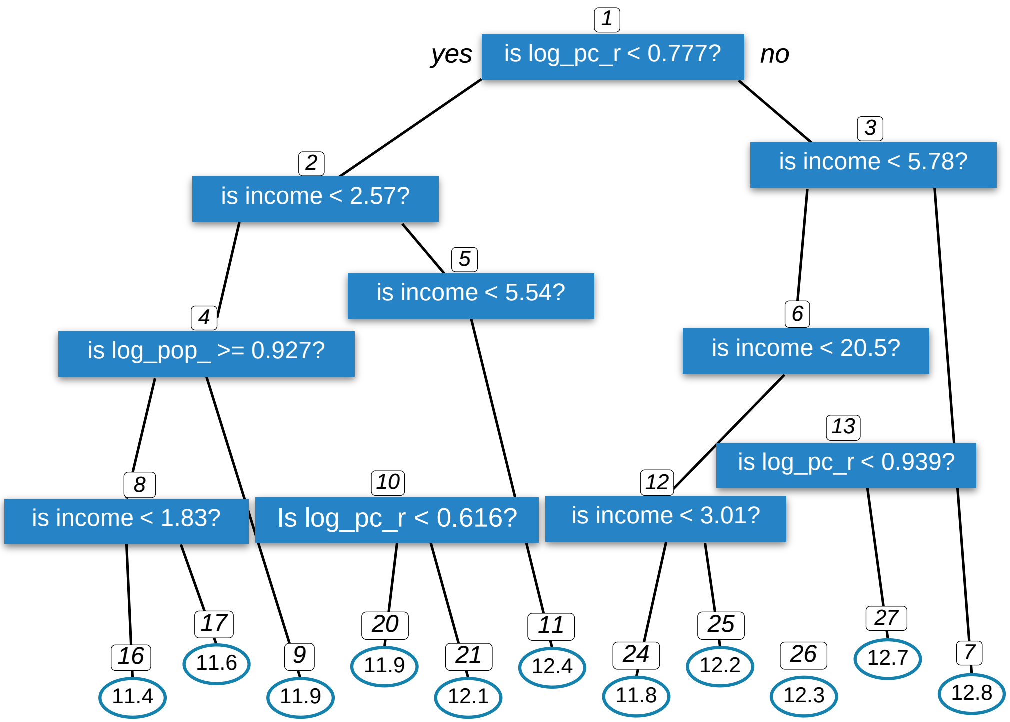
- Key advantage is interpretability
- Partition the space into simple regions to achieve best fit
- Pruning methods control the size of the tree



School of Information Studies
Syracuse University

Random Forests

- Ensemble method using multiple decision trees
- Recursive partitioning on the training set
- Effective with large number of explanatory variables



School of Information Studies
Syracuse University

Random Forests (cont.)

- Provides interpretability through use of one tree from set
- Significant difference in performance between train and test indicates overfitting
- Individual explanatory variables can still be inferred