# When is it good to be bad?

Web scraping and data analysis
of NHL penalties

Wendy Grus

@wgrus

wendygrus@gmail.com

# Zac Rinaldo's penalty

- Zac Rinaldo slams Kris Letang into the boards:

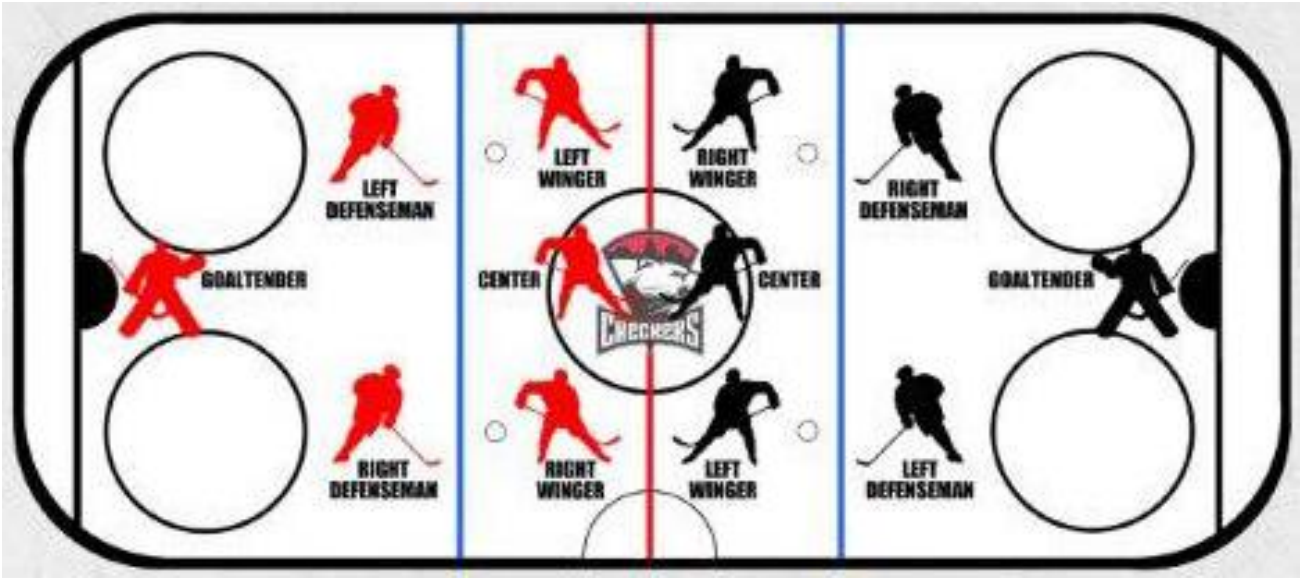
# Zac Rinaldo's penalty

- Zac Rinaldo gets ejected
- After the game, Zac Rinaldo says:

  *"Yeah, I changed the whole game, man. [Expletive], who knows what the game would have been like if I didn't do what I did?"*

- Zac Rinaldo gets suspended 8 games

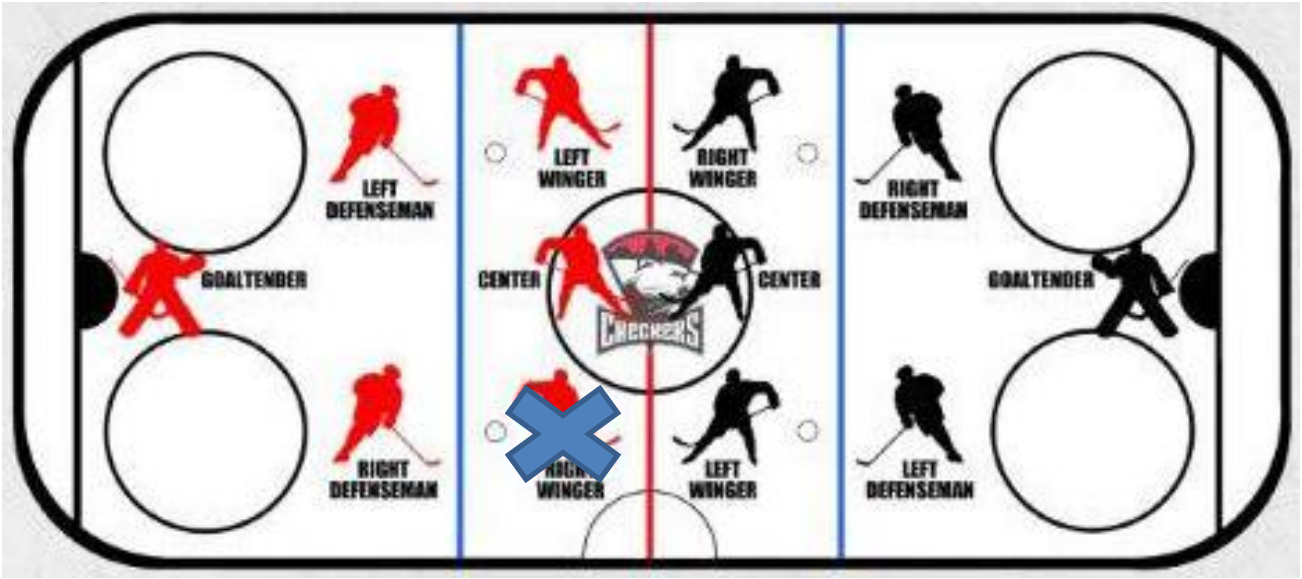
# Hockey 101




Regular play has 5 players + a goalie on each side.

When you break certain rules, you get sent to the penalty box, and your team must play down.

# Hockey 101




Hockey penalties put your team at a disadvantage.

But can they ever *help* you win the game?

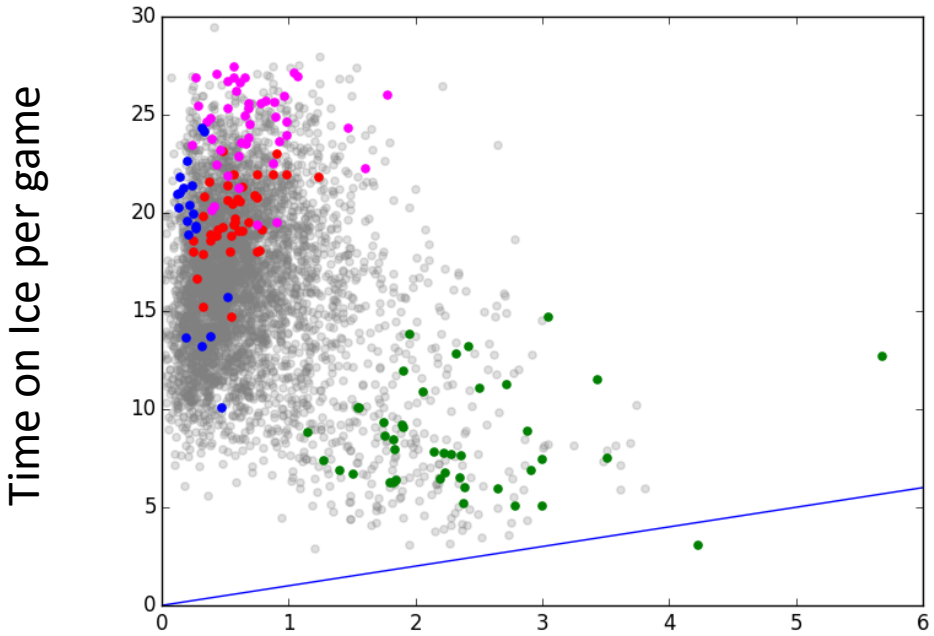
# Hockey 101

Skilled Players

Enforcers

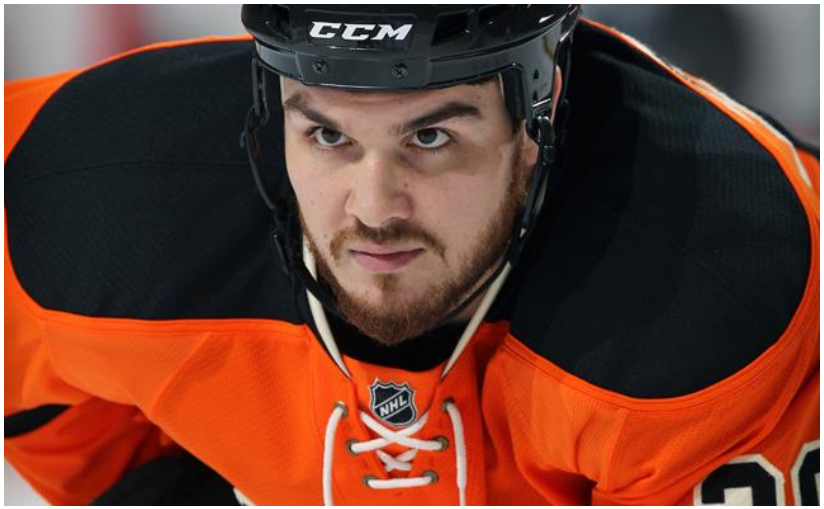
# Hockey 101

Lady Byng Trophy winners
High-scoring/skilled forwards
Skilled Defensemen
Enforcers



# When is it good to be bad?

To quantitatively evaluate Rinaldo's question,




I can analyze NHL hockey penalty data.

# Webscraping

Webscraping is extracting data from websites that do not have APIs that allow you access to the data programmatically.

# Webscraping: Things to think about

1. What data do I need?
2. What features should I look for in the website?
3. How do I collect, combine, and analyze data?

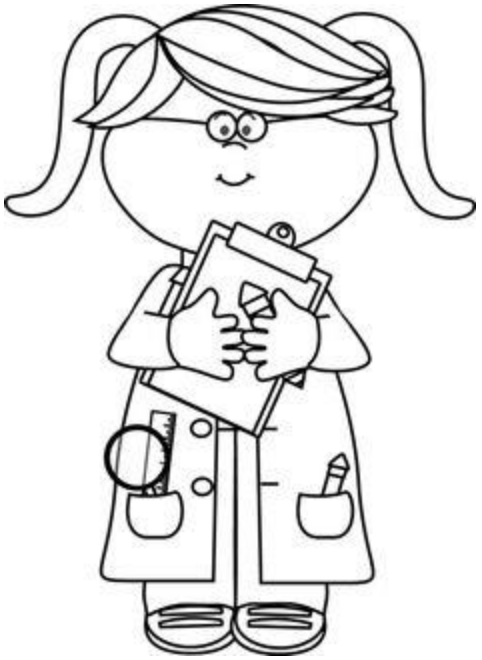
# Webscraping: Things to think about

1. What data do I need? BRAIN
2. What features should I look for in the website?
3. How do I collect, combine, and analyze data?

Requests, BeautifulSoup, pandas, statsmodels

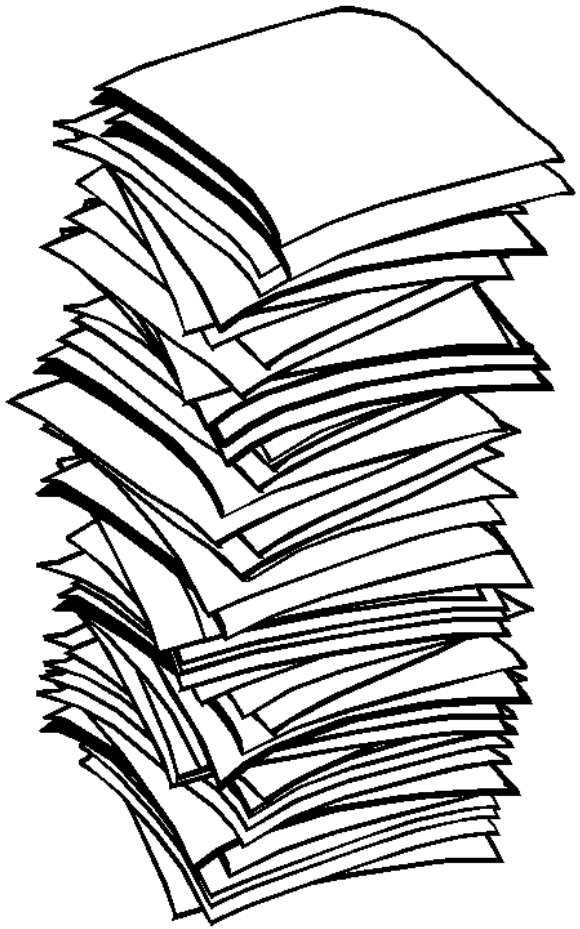
# What data do I need: Designing your experiment

1. What question do you want to answer?
2. What data will you need?
3. What data is available?

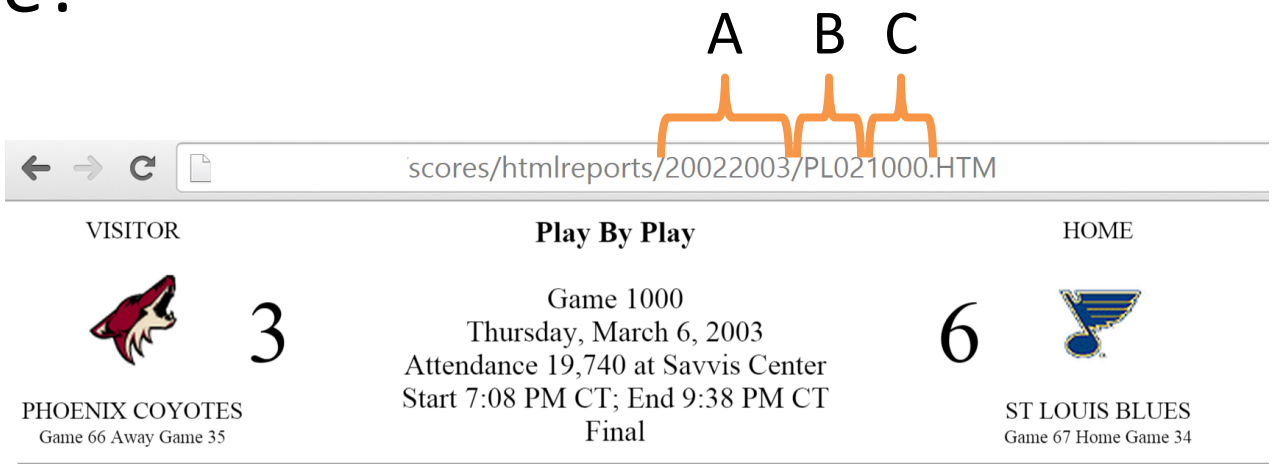


# What features should I look for in website?

1. Is it easy to automate moving from page to page?
2. Is the data easy to parse from the source code?



1. Is it easy to automate moving from page to page?




A: season

B: constant, PL – playbyplay 02 – regular season

C: game in the season (0001-1230)

```
import requests

for season in ['20072008','20082009','20092010',
               '20102011','20112012','20132014',
               '20142015','20152016']:
    for i in range(1,1231):
        str_i = '%04d' % i
        url = "http://[yoursitehere]/scores/htmlreports/" \
            + season + "/PL02" + str_i + ".HTM"
        if i % 100 == 0:
            print(url)
        result = requests.get(url)
        if not (result.status_code > 400):
            c = result.content.replace("charset=UTF-16","charset=UTF-8")
            outfile = "GAME" + str_i + "_" + season + ".html"
            game_summary = open(outfile, 'w')
            game_summary.write(c)
```

Loop through the seasons

Loop through the games

Get the content for each play-by-play url

2. Is the data easy to parse from the source code?

/htmlreports/20082009/PL021005.HTM

**VISITOR** | **HOME**

**Play By Play**

**2** | **5**

**Thursday, March 12, 2009**
**Attendance 14,578 at Prudential Center**
**Start 7:07 EDT ; End 9:26 EDT**
**Game 1005**
**Final**

**PHOENIX COYOTES**
**Game 68 Away Game 36**

**NEW JERSEY DEVILS**
**Game 67 Home Game 35**

| # | Per | Str | Time: Elapsed Game | Event | Description | PHX On Ice | N.J On Ice |
|---|---|---|---|---|---|---|---|
| 1 | 1 | | 0:00<br>20:00 | PSTR | Period Start- Local time: 7:07 EDT | 28 36 19 4 44 42<br>C C R D D G | 19 15 9 7 29 30<br>C R L D D G |
| 2 | 1 | EV | 0:00<br>20:00 | FAC | N.J won Neu. Zone - PHX #28 REINPRECHT vs N.J #19 ZAJAC | 28 36 19 4 44 42<br>C C R D D G | 19 15 9 7 29 30<br>C R L D D G |
| **3** | **1** | **EV** | **0:18<br>19:42** | **GOAL** | **N.J #9 PARISE(40), Slap, Off. Zone, 27 ft.<br>Assists: #15 LANGENBRUNNER(37); #7 MARTIN(22)** | 28 36 19 4 44 42<br>C C R D D G | 19 15 9 7 29 30<br>C R L D D G |
| 4 | 1 | EV | 0:18<br>19:42 | FAC | PHX won Neu. Zone - PHX #15 LOMBARDI vs N.J #12 ROLSTON | 15 8 16 2 55 42<br>C R R D D G | 12 14 26 5 27 30<br>R R L D D G |
| 5 | 1 | EV | 0:52<br>19:08 | BLOCK | N.J #27 MOTTAU BLOCKED BY PHX #8 UPSHALL, Slap, Def. Zone | 15 8 16 3 45 42<br>C R R D D G | 12 14 26 5 27 30<br>R R L D D G |
| 6 | 1 | EV | 1:03<br>18:57 | HIT | PHX #3 YANDLE HIT N.J #14 GIONTA, Def. Zone | 15 8 16 3 45 42<br>C R R D D G | 12 14 26 24 28 30<br>R R L D D G |
| 7 | 1 | EV | 1:12<br>18:48 | SHOT | N.J ONGOAL - #14 GIONTA, Slap, Off. Zone, 34 ft. | 15 8 16 3 45 42<br>C R R D D G | 12 14 26 24 28 30<br>R R L D D G |
| 8 | 1 | | 1:26<br>18:34 | STOP | PUCK IN CROWD | 88 14 89 4 44 42<br>C L L D D G | 8 23 18 24 28 30<br>C R L D D G |
| 9 | 1 | EV | 1:26<br>18:34 | FAC | PHX won Neu. Zone - PHX #88 MUELLER vs N.J #8 ZUBRUS | 88 14 89 4 44 42<br>C L L D D G | 8 23 18 24 28 30<br>C R L D D G |
| 10 | 1 | EV | 2:16<br>17:44 | HIT | N.J #23 CLARKSON HIT PHX #44 SAUER, Off. Zone | 34 29 41 2 44 42<br>C R R D D G | 8 23 18 7 29 30<br>C R L D D G |
| 11 | 1 | EV | 2:23<br>17:37 | MISS | PHX #34 WINNIK, Wrist, Wide of Net, Off. Zone, 17 ft. | 34 29 41 2 55 42<br>C R R D D G | 11 16 17 7 29 30<br>C C C D D G |
| 12 | 1 | | 2:23<br>17:37 | STOP | GOALIE STOPPED | 34 29 41 2 55 42<br>C R R D D G | 11 16 17 7 29 30<br>C C C D D G |
| 13 | 1 | EV | 2:23<br>17:37 | FAC | PHX won Off. Zone - PHX #28 REINPRECHT vs N.J #16 HOLIK | 28 36 19 2 55 42<br>C C R D D G | 11 16 17 7 29 30<br>C C C D D G |
| **14** | **1** | **EV** | **2:56<br>17:04** | **GOAL** | **N.J #11 MADDEN(7), Tip-In, Off. Zone, 12 ft.<br>Assists: #29 ODUYA(21); #17 RUPP(6)** | 28 36 19 2 55 42<br>C C R D D G | 11 16 17 7 29 30<br>C C C D D G |
| 15 | 1 | EV | 2:56<br>17:04 | FAC | PHX won Neu. Zone - PHX #15 LOMBARDI vs N.J #19 ZAJAC | 15 8 16 4 44 42<br>C R R D D G | 19 15 9 5 27 30<br>C R L D D G |

# 2. Is the data easy to parse from the source code?

Browser-rendered site

Look at the source code

/htmlreports/20082009/PL021005.HTM

**VISITOR** 2

PHOENIX COYOTES
Game 68 Away Game 36

**Play By Play**

Thursday, March 12, 2009
Attendance 14,578 at Prudential Center
Start 7:07 EDT ; End 9:26 EDT
Game 1005
Final

**HOME** 5

NEW JERSEY DEVILS
Game 67 Home Game 35

| # | Per | Str | Time: Elapsed Game | Event | Description | PHX On Ice | N.J On Ice |
|---|---|---|---|---|---|---|---|
| 1 | 1 | | 0:00 20:00 | PSTR | Period Start- Local time: 7:07 EDT | 28 36 19 4 44 42 C C R D D G | 19 15 9 7 29 30 C R L D D G |
| 2 | 1 | EV | 0:00 20:00 | FAC | N.J won Neu. Zone - PHX #28 REINPRECHT vs N.J #19 ZAJAC | 28 36 19 4 44 42 C C R D D G | 19 15 9 7 29 30 C R L D D G |
| **3** | **1** | **EV** | **0:18 19:42** | **GOAL** | **N.J #9 PARISE(40), Slap, Off. Zone, 27 ft. Assists: #15 LANGENBRUNNER(37); #7 MARTIN(22)** | 28 36 19 4 44 42 C C R D D G | 19 15 9 7 29 30 C R L D D G |
| 4 | 1 | EV | 0:18 19:42 | FAC | PHX won Neu. Zone - PHX #15 LOMBARDI vs N.J #12 ROLSTON | 15 8 16 2 55 42 C R R D D G | 12 14 26 5 27 30 R R L D D G |
| 5 | 1 | EV | 0:52 19:08 | BLOCK | N.J #27 MOTTAU BLOCKED BY PHX #8 UPSHALL, Slap, Def. Zone | 15 8 16 3 45 42 C R R D D G | 12 14 26 5 27 30 R R L D D G |
| 6 | 1 | EV | 1:03 18:57 | HIT | PHX #3 YANDLE HIT N.J #14 GIONTA, Def. Zone | 15 8 16 3 45 42 C R R D D G | 12 14 26 24 28 30 R R L D D G |
| 7 | 1 | EV | 1:12 18:48 | SHOT | N.J ONGOAL - #14 GIONTA, Slap, Off. Zone, 34 ft. | 15 8 16 3 45 42 C R R D D G | 12 14 26 24 28 30 R R L D D G |
| 8 | 1 | | 1:26 18:34 | STOP | PUCK IN CROWD | 88 14 89 4 44 42 C L L D D G | 8 23 18 24 28 30 C R L D D G |
| 9 | 1 | EV | 1:26 18:34 | FAC | PHX won Neu. Zone - PHX #88 MUELLER vs N.J #8 ZUBRUS | 88 14 89 4 44 42 C L L D D G | 8 23 18 24 28 30 C R L D D G |
| 10 | 1 | EV | 2:16 17:44 | HIT | N.J #23 CLARKSON HIT PHX #44 SAUER, Off. Zone | 34 29 41 2 44 42 C R R D D G | 8 23 18 7 29 30 C R L D D G |
| 11 | 1 | EV | 2:23 17:37 | MISS | PHX #34 WINNIK, Wrist, Wide of Net, Off. Zone, 17 ft. | 34 29 41 2 55 42 C R R D D G | 11 16 17 7 29 30 C C C D D G |
| 12 | 1 | | 2:23 17:37 | STOP | GOALIE STOPPED | 34 29 41 2 55 42 C R R D D G | 11 16 17 7 29 30 C C C D D G |
| 13 | 1 | EV | 2:23 17:37 | FAC | PHX won Off. Zone - PHX #28 REINPRECHT vs N.J #16 HOLIK | 28 36 19 2 55 42 C C R D D G | 11 16 17 7 29 30 C C C D D G |
| **14** | **1** | **EV** | **2:56 17:04** | **GOAL** | **N.J #11 MADDEN(7), Tip-In, Off. Zone, 12 ft. Assists: #29 ODUYA(21); #17 RUPP(6)** | 28 36 19 2 55 42 C C R D D G | 11 16 17 7 29 30 C C C D D G |
| 15 | 1 | EV | 2:56 17:04 | FAC | PHX won Neu. Zone - PHX #15 LOMBARDI vs N.J #19 ZAJAC | 15 8 16 4 44 42 C R R D D G | 19 15 9 5 27 30 C R L D D G |
| 16 | 1 | EV | 3:21 16:39 | HIT | PHX #16 PRUCHA HIT N.J #9 PARISE, Def. Zone | 15 8 16 4 44 42 C R R D D G | 19 15 9 5 27 30 C R L D D G |
| 17 | 1 | EV | 3:30 16:30 | SHOT | PHX ONGOAL - #8 UPSHALL, Slap, Off. Zone, 56 ft. | 15 8 16 4 44 42 C R R D D G | 19 15 9 5 27 30 C R L D D G |
| 18 | 1 | EV | 3:36 16:24 | SHOT | PHX ONGOAL - #3 YANDLE, Slap, Off. Zone, 62 ft. | 15 8 16 3 45 42 C R R D D G | 19 15 9 5 27 30 C R L D D G |

scores/htmlreports/20072008/PL021005.HTM

```
<table border="0" cellpadding="0" cellspacing="0">
<tr>
<td align="center">
<font style="cursor:hand;" title="Goalie - JOSE THEODORE">60</font>
</td>
</tr>
<tr>
<td align="center">G</td>
</tr>
</table>
</td>
</tr>
</table>
</td>
</tr>
<tr class="evenColor">
<td align="center" class="penalty + bborder">4</td>
<td class="penalty + bborder" align="center">1</td>
<td class="penalty + bborder" align="center">EV</td>
<td class="penalty + bborder" align="center">0:28<br>19:32</td>
<td class="penalty + bborder">PENL</td>
<td class="penalty + bborder">COL #22 HANNAN Fighting (maj)(5 min), Off. Zone Drawn By: VAN #14 BURROWS</t
<td class="italicize + bold + bborder + rborder">
<table border="0" cellpadding="0" cellspacing="0">
<tr>
<td align="center">
<table border="0" cellpadding="0" cellspacing="0">
<tr>
<td align="center">
<font style="cursor:hand;" title="Center - RYAN KESLER">17</font>
</td>
</tr>
<tr>
<td align="center">C</td>
</tr>
</table>
</td>
<td align="center"> </td>
```

# How do I collect the data?

```
<tr class="evenColor">
<td align="center" class="penalty + bborder">46</td>
<td class="penalty + bborder" align="center">1</td>
<td class="penalty + bborder" align="center">EV</td>
<td class="penalty + bborder" align="center">5:59<br>14:01</td>
<td class="penalty + bborder">PENL</td>
<td class="penalty + bborder">NYR #17 DUBINSKY Hooking(2 min), Off. Zone Drawn By: T.B #4 LEC
<td class="italicize + bold + bborder + rborder">
```

```
for penalty in penalties:
    type = play_lines[penalty+1].split('\xff')[1].split('(')[0].strip()
    team = play_lines[penalty+1].split()[0].strip()
    if 'TEAM' in play_lines[penalty+1]:
        player = 'TEAM'
    else:
        player = play_lines[penalty+1].split('#')[1].split('\xff')[0].strip()
    period = play_lines[penalty-3].strip()
    min = play_lines[penalty-1].split(":")[0].strip()
    seconds = play_lines[penalty-1].split(":")[1][:2].strip()
    time = min + ":" + seconds
    length = play_lines[penalty+1].split('\xff')[1].split('(')[1].split()[0].strip()
```

Sped up penalty parser by switching from BeautifulSoup to lxml to extract content!

# How do I combine the data?

For every penalty, parsed out:

season, game#, game_time, type, team, player, drawnby, servedby, period, length, score_diff, next_goal, score_differential, positive_change

Combined in team information:

hometeam, win_pct, opp_win_pct

Combined in player information for player and drawn by player:

time on ice, penalties in min per game

# How do I combine the data?

When using multiple data sources, data fields may not be named the same.

TEAMS:

New Jersey Devils can be N.J on one site and NJD on another

PLAYERS:

play-by-play recaps gave players by # lastname team

player stats gave players by firstname lastname teams

# How do I combine the data?

When looking at data over time, the data fields may change names.

| | | | |
|---|---|---|---|
| Anaheim Ducks | NHL | 1993 | 2016 |
| Anaheim Ducks | NHL | 2006 | 2016 |
| Mighty Ducks of Anaheim | NHL | 1993 | 2006 |
| Arizona Coyotes | NHL | 1979 | 2016 |
| Arizona Coyotes | NHL | 2014 | 2016 |
| Phoenix Coyotes | NHL | 1996 | 2014 |
| Winnipeg Jets | NHL | 1979 | 1996 |

# How do I combine the data?

When looking at many classes of data, it can be useful to reduce the data into a smaller number of categories.

I reduced the 113 distinct penalty names into 8 penalty types:

Physical foul
Stick infraction
Delay of game
Altercation
Penalty Shot
Bench
Illegal behavior
Impeding behavior

# How do I analyze data?

Evaluating three outcomes:

1) **Next goal:** Team that gets the penalty scores the next goal
2) **Positive change:** final state better than state at the time of the penalty
3) **Score differential:** score difference between the time of penalty and the end of the game (not including overtime)

# How do I analyze data?

Use logistic regression (next goal and positive change) and linear regression (score differential) to build models that predicts the outcome of the game based on penalty, game, and player features.

Logistic:

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

Linear:

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

# How do I analyze data?

Choosing covariates:

1) Physical foul indicator
2) Hometeam indicator
3) Team strength: Win percentage
4) Opponent strength: Opponent's win percentage
5) Penalized player strength: time on ice and penalty minutes per game
6) Drawing player strength: time on ice and penalty minutes per game

# How do I analyze data?

Filtered combined dataset for analysis

91476 total penalties

82990 penalties with drawn by information

77542 penalties with a next goal scored

# Analyze data: importing data to pandas

Import data into a pandas dataframe

```
df = pd.read_csv('allcombined_ng3.txt', dtype={'season': str,  ' period' : str, 'positive_change': int,
                                               ' score_diff_change': int, ' period': int, ' pp_time' : int,
                                               ' stick': bool, ' foul': bool, ' delay':bool, ' bench': bool,
                                               ' ps': bool, ' altercation':bool, ' impeding':bool,
                                               ' illegalb':bool, ' hometeam' : bool}
```

Adjust column names

```
real_names=['season','positive_change', 'next_goal', 'score_diff_change', 'period',
       'time', 'pp_time', 'team', 'stick', 'foul', 'delay', 'bench',
       'ps', 'altercation', 'impeding', 'illegalb', 'hometeam',
       'win_pct', 'opp_win_pct', 'toi', 'pimpg', 'db_toi', 'db_pimpg']
df.columns = real_names
```

# Logistic Regression: next goal

Use statsmodels Logit to evaluate if taking a penalty increases the odds of your team scoring the next goal.

```
result = sm.Logit.from_formula(formula="next_goal ~ foul + hometeam",
                data=isnextgoal).fit()
print(result.summary())
print(np.exp(result.params))
```

```
                           Logit Regression Results
==============================================================================
Dep. Variable:              next_goal   No. Observations:                77542
Model:                          Logit   Df Residuals:                    77539
Method:                           MLE   Df Model:                            2
Date:                Sun, 22 May 2016   Pseudo R-squ.:                0.001932
Time:                        21:28:11   Log-Likelihood:                -52802.
converged:                       True   LL-Null:                       -52904.
                                        LLR p-value:                4.143e-45
==============================================================================
                     coef    std err          z      P>|z|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept         -0.3999      0.010    -38.262      0.000        -0.420    -0.379
foul[T.True]      -0.0311      0.037     -0.851      0.395        -0.103     0.041
hometeam[T.True]   0.2076      0.015     14.270      0.000         0.179     0.236
==============================================================================
Intercept          0.670402
foul[T.True]       0.969399
hometeam[T.True]   1.230662
```

# Logistic Regression: next goal

Use statsmodels Logit to evaluate if taking a penalty increases the odds of your team scoring the next goal.

```
result = sm.Logit.from_formula(formula="next_goal ~ foul + hometeam",
                data=isnextgoal).fit()
print(result.summary())
print(np.exp(result.params))
```

```
                           Logit Regression Results
==============================================================================
Dep. Variable:              next_goal   No. Observations:                77542
Model:                          Logit   Df Residuals:                    77539
Method:                           MLE   Df Model:                            2
Date:                Sun, 22 May 2016   Pseudo R-squ.:                0.001932
Time:                        21:28:11   Log-Likelihood:                -52802.
converged:                       True   LL-Null:                       -52904.
                                        LLR p-value:                4.143e-45
==============================================================================
                     coef    std err          z      P>|z|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept         -0.3999      0.010    -38.262      0.000        -0.420    -0.379
foul[T.True]      -0.0311      0.037     -0.851      0.395        -0.103     0.041
hometeam[T.True]   0.2076      0.015     14.270      0.000         0.179     0.236
==============================================================================
Intercept          0.670402
foul[T.True]       0.969399
hometeam[T.True]   1.230662
```

# Logistic Regression: positive change

Results summary and odds ratios:

Logit Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | positive_change | No. Observations: | 70790 |
| Model: | Logit | Df Residuals: | 70781 |
| Method: | MLE | Df Model: | 8 |
| Date: | Sun, 22 May 2016 | Pseudo R-squ.: | 0.002210 |
| Time: | 22:00:34 | Log-Likelihood: | -35816. |
| converged: | True | LL-Null: | -35895. |
| | | LLR p-value: | 3.095e-30 |

| | coef | std err | z | P>\|z\| | [95.0% Conf. | Int.] |
|---|---|---|---|---|---|---|
| Intercept | -1.4062 | 0.069 | -20.378 | 0.000 | -1.541 | -1.271 |
| foul[T.True] | 0.0644 | 0.044 | 1.450 | 0.147 | -0.023 | 0.151 |
| hometeam[T.True] | 0.1220 | 0.019 | 6.531 | 0.000 | 0.085 | 0.159 |
| win_pct | 0.6280 | 0.069 | 9.041 | 0.000 | 0.492 | 0.764 |
| opp_win_pct | -0.5442 | 0.069 | -7.921 | 0.000 | -0.679 | -0.410 |
| toi | -0.0037 | 0.002 | -1.778 | 0.075 | -0.008 | 0.000 |
| pimpg | 0.0068 | 0.014 | 0.491 | 0.623 | -0.020 | 0.034 |
| db_toi | -0.0008 | 0.002 | -0.368 | 0.713 | -0.005 | 0.003 |
| db_pimpg | 0.0162 | 0.014 | 1.186 | 0.236 | -0.011 | 0.043 |

| | |
|---|---|
| Intercept | 0.245085 |
| foul[T.True] | 1.066487 |
| hometeam[T.True] | 1.129713 |
| win_pct | 1.873927 |
| opp_win_pct | 0.580281 |
| toi | 0.996268 |
| pimpg | 1.006791 |
| db_toi | 0.999194 |
| db_pimpg | 1.016319 |

# Logistic Regression: positive change

Results summary and odds ratios:

Logit Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | positive_change | No. Observations: | 70790 |
| Model: | Logit | Df Residuals: | 70781 |
| Method: | MLE | Df Model: | 8 |
| Date: | Sun, 22 May 2016 | Pseudo R-squ.: | 0.002210 |
| Time: | 22:00:34 | Log-Likelihood: | -35816. |
| converged: | True | LL-Null: | -35895. |
| | | LLR p-value: | 3.095e-30 |

| | coef | std err | z | P>\|z\| | [95.0% Conf. | Int.] |
|---|---|---|---|---|---|---|
| Intercept | -1.4062 | 0.069 | -20.378 | 0.000 | -1.541 | -1.271 |
| foul[T.True] | 0.0644 | 0.044 | 1.450 | 0.147 | -0.023 | 0.151 |
| hometeam[T.True] | 0.1220 | 0.019 | 6.531 | 0.000 | 0.085 | 0.159 |
| win_pct | 0.6280 | 0.069 | 9.041 | 0.000 | 0.492 | 0.764 |
| opp_win_pct | -0.5442 | 0.069 | -7.921 | 0.000 | -0.679 | -0.410 |
| toi | -0.0037 | 0.002 | -1.778 | 0.075 | -0.008 | 0.000 |
| pimpg | 0.0068 | 0.014 | 0.491 | 0.623 | -0.020 | 0.034 |
| db_toi | -0.0008 | 0.002 | -0.368 | 0.713 | -0.005 | 0.003 |
| db_pimpg | 0.0162 | 0.014 | 1.186 | 0.236 | -0.011 | 0.043 |

| | |
|---|---|
| Intercept | 0.245085 |
| foul[T.True] | 1.066487 |
| hometeam[T.True] | 1.129713 |
| win_pct | 1.873927 |
| opp_win_pct | 0.580281 |
| toi | 0.996268 |
| pimpg | 1.006791 |
| db_toi | 0.999194 |
| db_pimpg | 1.016319 |

# Linear Regression: score differential

Use statsmodels OLS to evaluate if taking a penalty increases the score differential from the time of the penalty to the end of regulation.

```
result = sm.ols(formula="score_diff_change ~ foul + hometeam + win_pct + opp_win_pct + toi + pimpg + db_toi + db_pimpg'
                data=isnextgoal[nodrawnby]).fit()
print(result.summary())
```

# Linear Regression: score differential

Results summary:

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | score_diff_change | R-squared: | 0.015 |
| Model: | OLS | Adj. R-squared: | 0.014 |
| Method: | Least Squares | F-statistic: | 130.7 |
| Date: | Sun, 22 May 2016 | Prob (F-statistic): | 1.05e-218 |
| Time: | 22:09:28 | Log-Likelihood: | -1.4303e+05 |
| No. Observations: | 70790 | AIC: | 2.861e+05 |
| Df Residuals: | 70781 | BIC: | 2.862e+05 |
| Df Model: | 8 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [95.0% Conf. | Int.] |
|---|---|---|---|---|---|---|
| Intercept | -0.2877 | 0.051 | -5.670 | 0.000 | -0.387 | -0.188 |
| foul[T.True] | 0.0226 | 0.033 | 0.680 | 0.496 | -0.043 | 0.088 |
| hometeam[T.True] | 0.3446 | 0.014 | 25.094 | 0.000 | 0.318 | 0.371 |
| win_pct | 0.8548 | 0.051 | 16.824 | 0.000 | 0.755 | 0.954 |
| opp_win_pct | -0.8486 | 0.051 | -16.788 | 0.000 | -0.948 | -0.749 |
| toi | -0.0012 | 0.002 | -0.780 | 0.435 | -0.004 | 0.002 |
| pimpg | -0.0021 | 0.010 | -0.203 | 0.839 | -0.022 | 0.018 |
| db_toi | -0.0012 | 0.002 | -0.758 | 0.448 | -0.004 | 0.002 |
| db_pimpg | 0.0391 | 0.010 | 3.856 | 0.000 | 0.019 | 0.059 |

| | | | |
|---|---|---|---|
| Omnibus: | 171.463 | Durbin-Watson: | 2.005 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 211.221 |
| Skew: | 0.041 | Prob(JB): | 1.36e-46 |
| Kurtosis: | 3.255 | Cond. No. | 202. |

# Linear Regression: score differential

Results summary:

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | score_diff_change | R-squared: | 0.015 |
| Model: | OLS | Adj. R-squared: | 0.014 |
| Method: | Least Squares | F-statistic: | 130.7 |
| Date: | Sun, 22 May 2016 | Prob (F-statistic): | 1.05e-218 |
| Time: | 22:09:28 | Log-Likelihood: | -1.4303e+05 |
| No. Observations: | 70790 | AIC: | 2.861e+05 |
| Df Residuals: | 70781 | BIC: | 2.862e+05 |
| Df Model: | 8 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [95.0% Conf. | Int.] |
|---|---|---|---|---|---|---|
| Intercept | -0.2877 | 0.051 | -5.670 | 0.000 | -0.387 | -0.188 |
| foul[T.True] | 0.0226 | 0.033 | 0.680 | 0.496 | -0.043 | 0.088 |
| hometeam[T.True] | 0.3446 | 0.014 | 25.094 | 0.000 | 0.318 | 0.371 |
| win_pct | 0.8548 | 0.051 | 16.824 | 0.000 | 0.755 | 0.954 |
| opp_win_pct | -0.8486 | 0.051 | -16.788 | 0.000 | -0.948 | -0.749 |
| toi | -0.0012 | 0.002 | -0.780 | 0.435 | -0.004 | 0.002 |
| pimpg | -0.0021 | 0.010 | -0.203 | 0.839 | -0.022 | 0.018 |
| db_toi | -0.0012 | 0.002 | -0.758 | 0.448 | -0.004 | 0.002 |
| db_pimpg | 0.0391 | 0.010 | 3.856 | 0.000 | 0.019 | 0.059 |

| | | | |
|---|---|---|---|
| Omnibus: | 171.463 | Durbin-Watson: | 2.005 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 211.221 |
| Skew: | 0.041 | Prob(JB): | 1.36e-46 |
| Kurtosis: | 3.255 | Cond. No. | 202. |

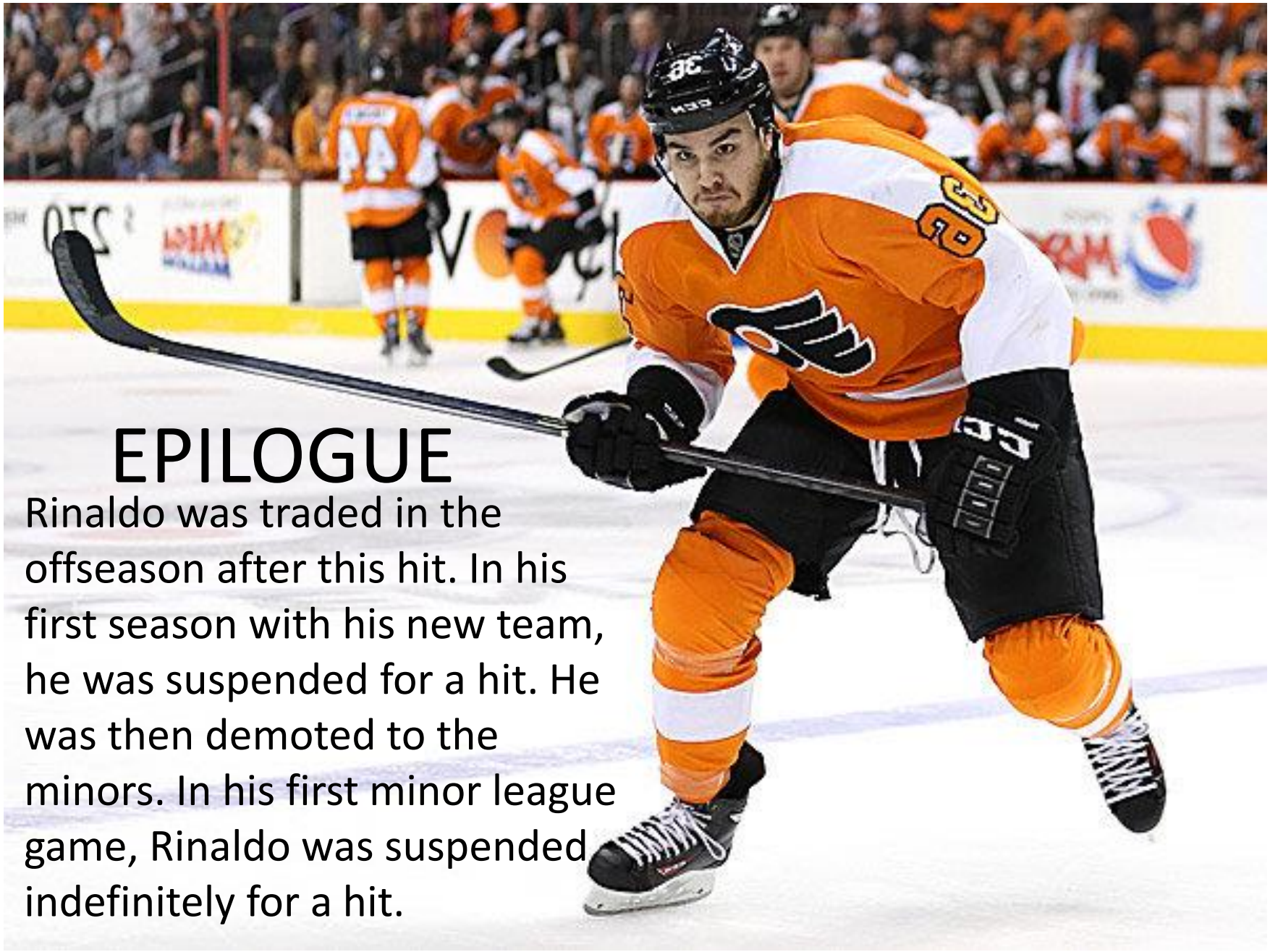
# When is it good to be bad?

- When you have a better record than your opponent
- When you are the home team

# How was Zac Rinaldo so wrong?

# Thanks!


INRIX®



# EPILOGUE

Rinaldo was traded in the offseason after this hit. In his first season with his new team, he was suspended for a hit. He was then demoted to the minors. In his first minor league game, Rinaldo was suspended indefinitely for a hit.