Aalto University
School of Business

SPORT ANALYTICS

Dr. Jirka Poropudas, Director of Analytics, SportIQ

jirka.poropudas@sportiq.xyz

Outline

1. **Overview of sport analytics**
   - Brief introduction through examples
2. **Team performance evaluation**
   - Ranking and rating teams
   - Estimation of winning probabilities
3. **Assignment: ”Optimal betting portfolio for Liiga playoffs”**
   - Poisson regression for team ratings
   - Estimation of winning probabilities
   - Simulation of the playoff bracket
   - Optimal betting portfolio

1. Overview of sport analytics

What is sport analytics?

B. Alamar and V. Mehrotra (Analytics Magazine, Sep./Oct. 2011):

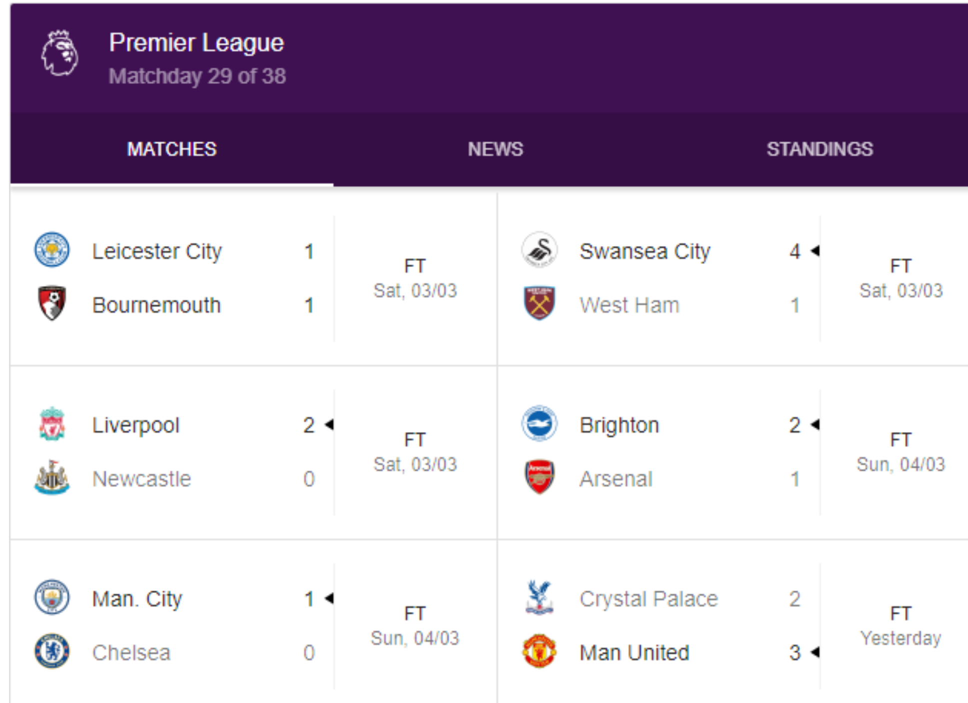
"The management of **structured historical data**, the application of **predictive analytic models** that utilize that data, and the use of **information systems** to inform decision makers and enable them to help their organizations in gaining a **competitive advantage** on the **field of play**."

Applications of sport analytics

- **Coaches**
  - Tactics, training, scouting, and planning
- **General managers and front offices**
  - Player evaluation and team building
- **Television, other broadcasters, and news media**
  - Entertainment, better content, storytelling, and visualizations
- **Bookmakers and bettors**
  - Betting odds and point spreads

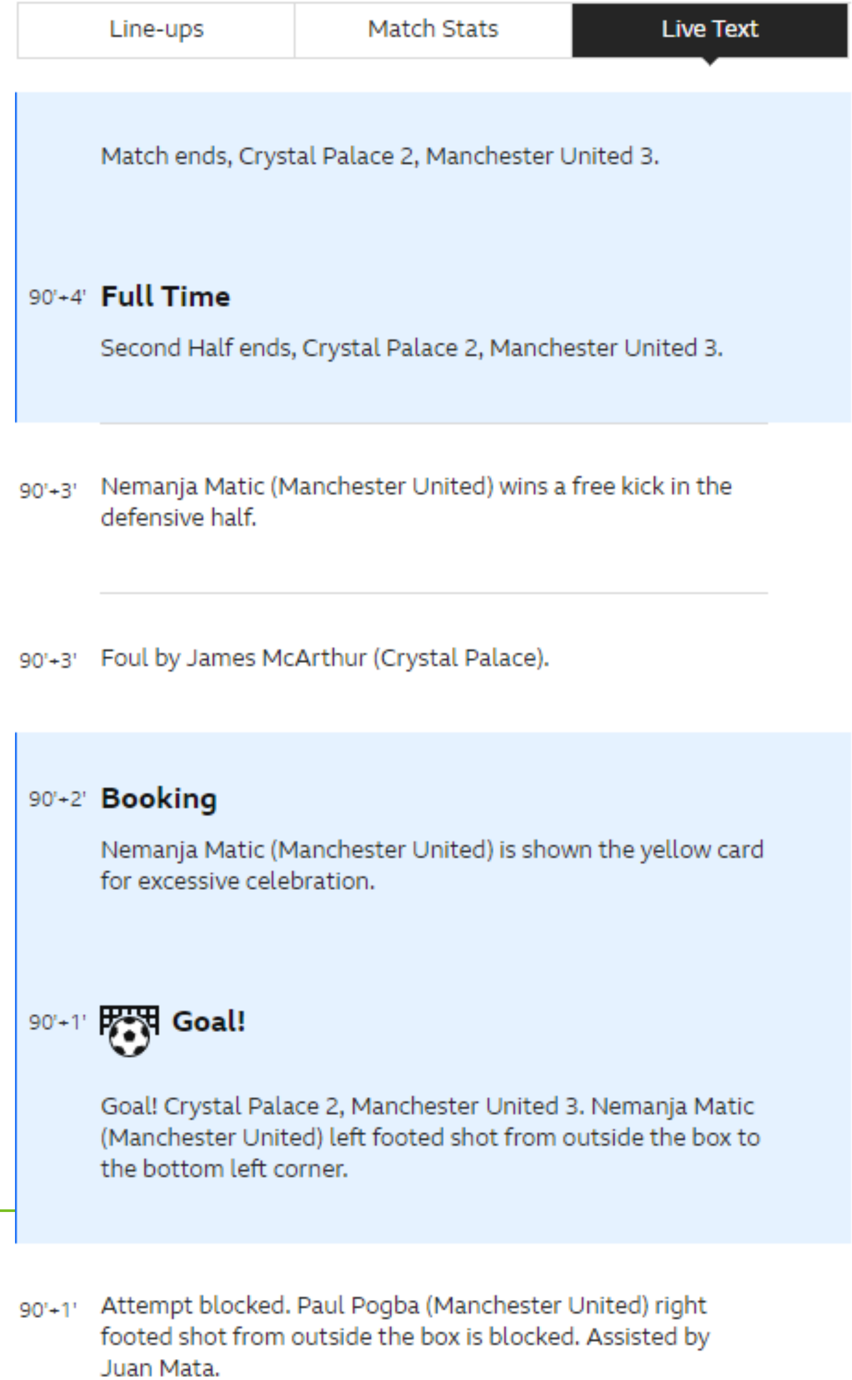
Data sources

- **Official summary statistics**
  - Aggregated totals from game events
- **Official play-by-play statistics**
  - Record of game events as they take place
- **Manual tracking and video analytics**
  - More detailed team-specific events
  - Labor intensive approach
  - Data consistency?
- **Automated tracking systems**
  - Expensive
  - Consistency based on given event definitions



Data sources

- **Official summary statistics**
  - Aggregated totals from game events
- **Official play-by-play statistics**
  - Record of game events as they take place
- **Manual tracking and video analytics**
  - More detailed team-specific events
  - Labor intensive approach
  - Data consistency?
- **Automated tracking systems**
  - Expensive
  - Consistency based on given event definitions



Data sources

- **Official summary statistics**
  - Aggregated totals from game events
- **Official play-by-play statistics**
  - Record of game events as they take place
- **Manual tracking and video analytics**
  - More detailed team-specific events
  - Labor intensive approach
  - Data consistency?
- **Automated tracking systems**
  - Expensive
  - Consistency based on given event definitions

Data sources

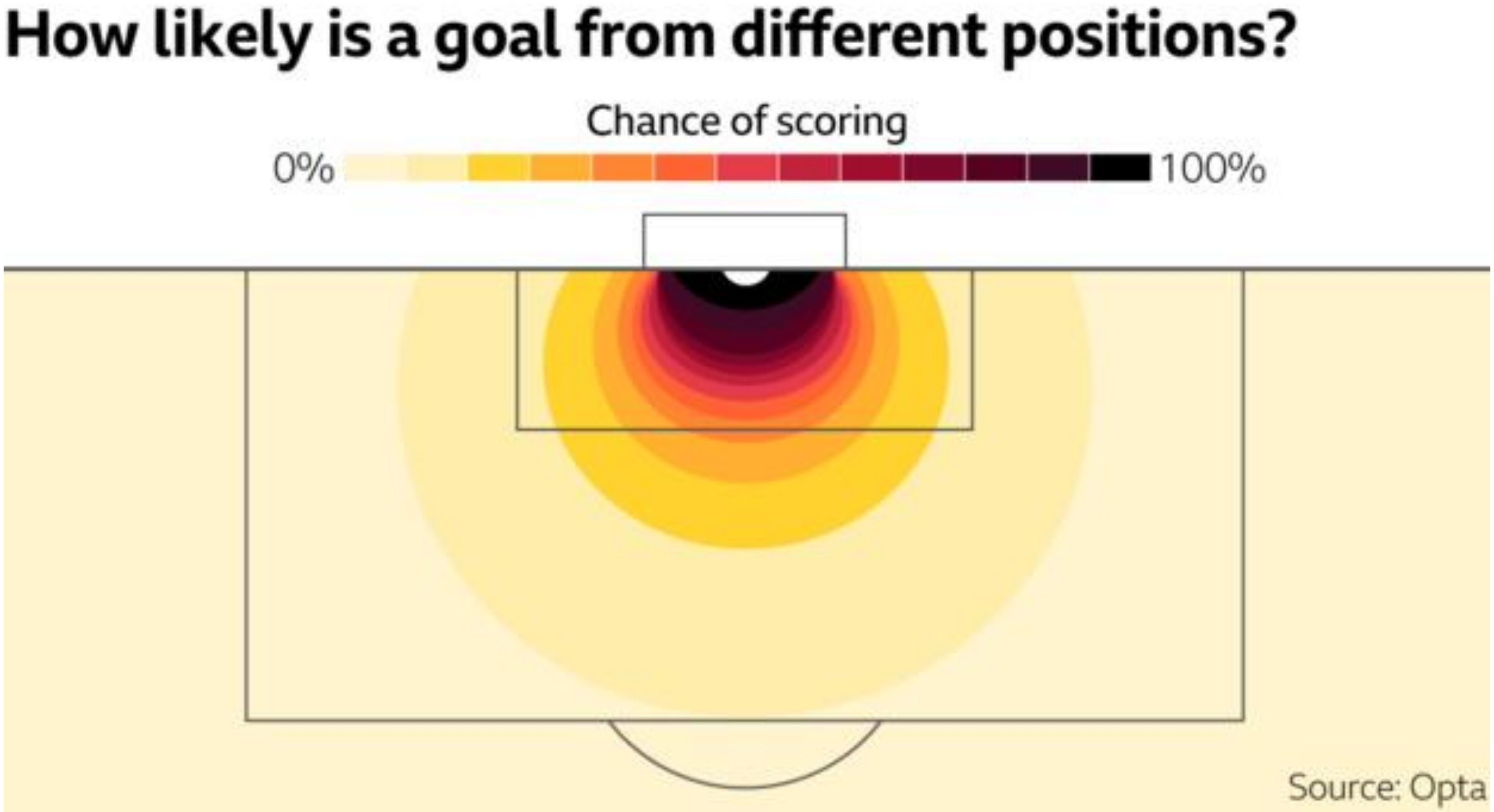
- **Official summary statistics**
  - Aggregated totals from game events
- **Official play-by-play statistics**
  - Record of game events as they take place
- **Manual tracking and video analytics**
  - More detailed team-specific events
  - Labor intensive approach
  - Data consistency?
- **Automated tracking systems**
  - Expensive
  - Consistency based on given event definitions

- https://www.youtube.com/edit?video_id=7IdxFcy3PFA

Methodology

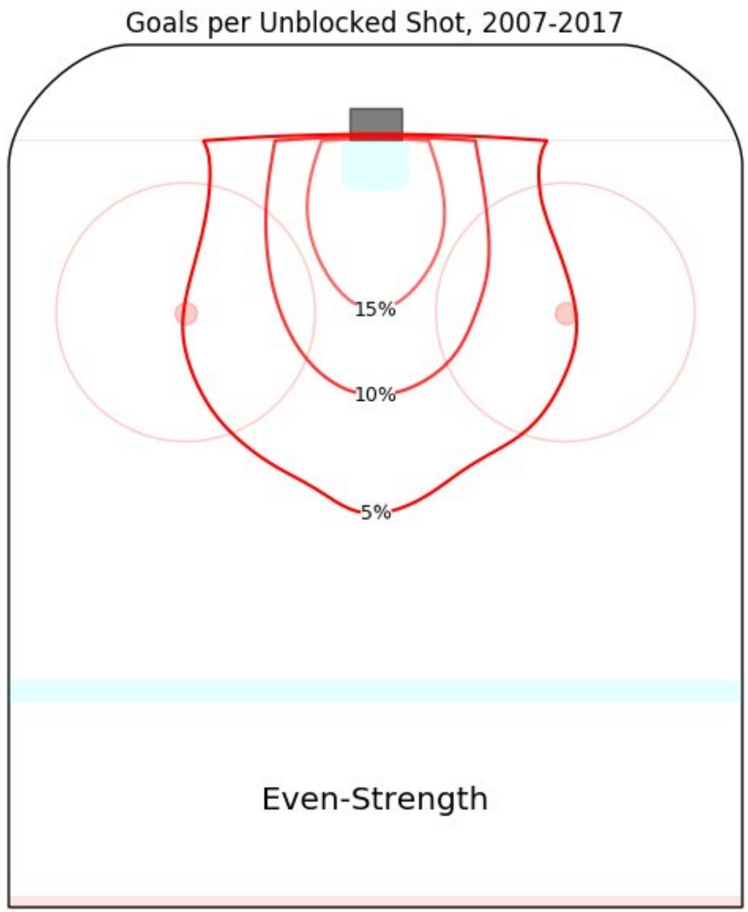
- **Basic statistics and more advanced techniques**
  - Signal vs. noise
- **Mathematical modeling**
  - Rules and scoring system specific factors
- **Machine learning**
  - Neural networks, deep learning, Bayesian networks etc.
- **Optimization**
- **Simulation**

EPL (football) – Expected goals



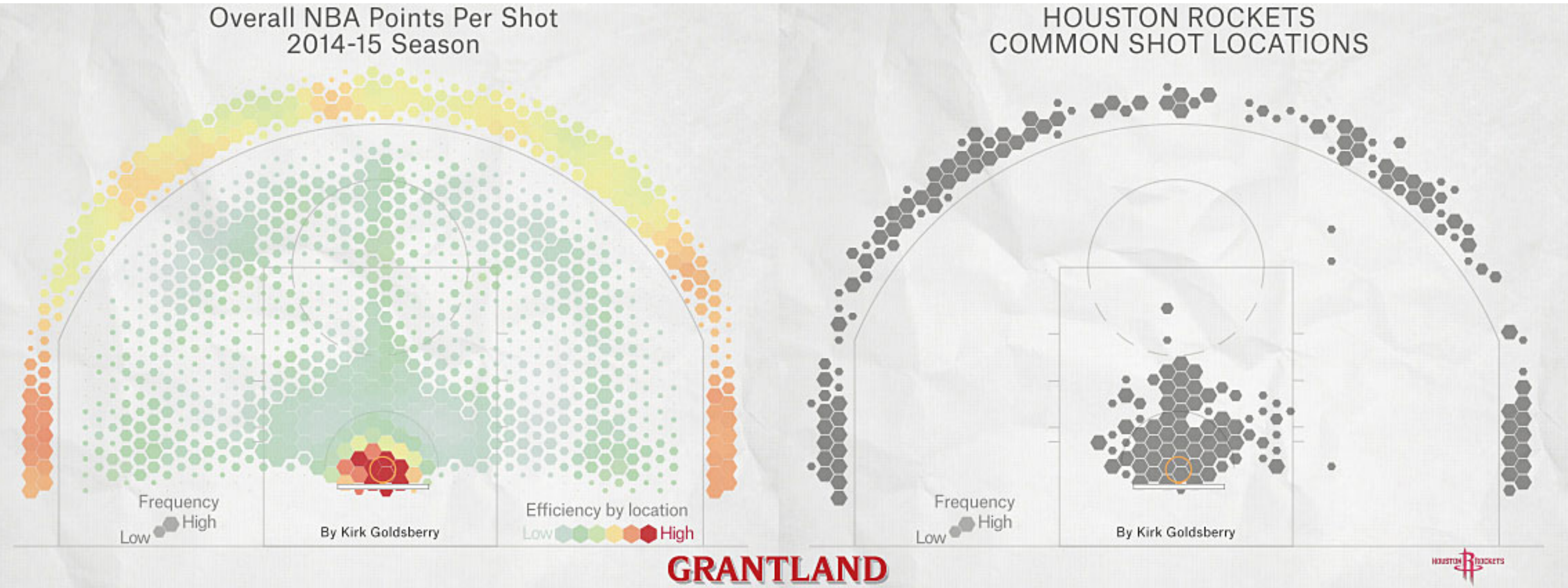
http://www.bbc.com/sport/football/40699431

NHL (ice hockey)



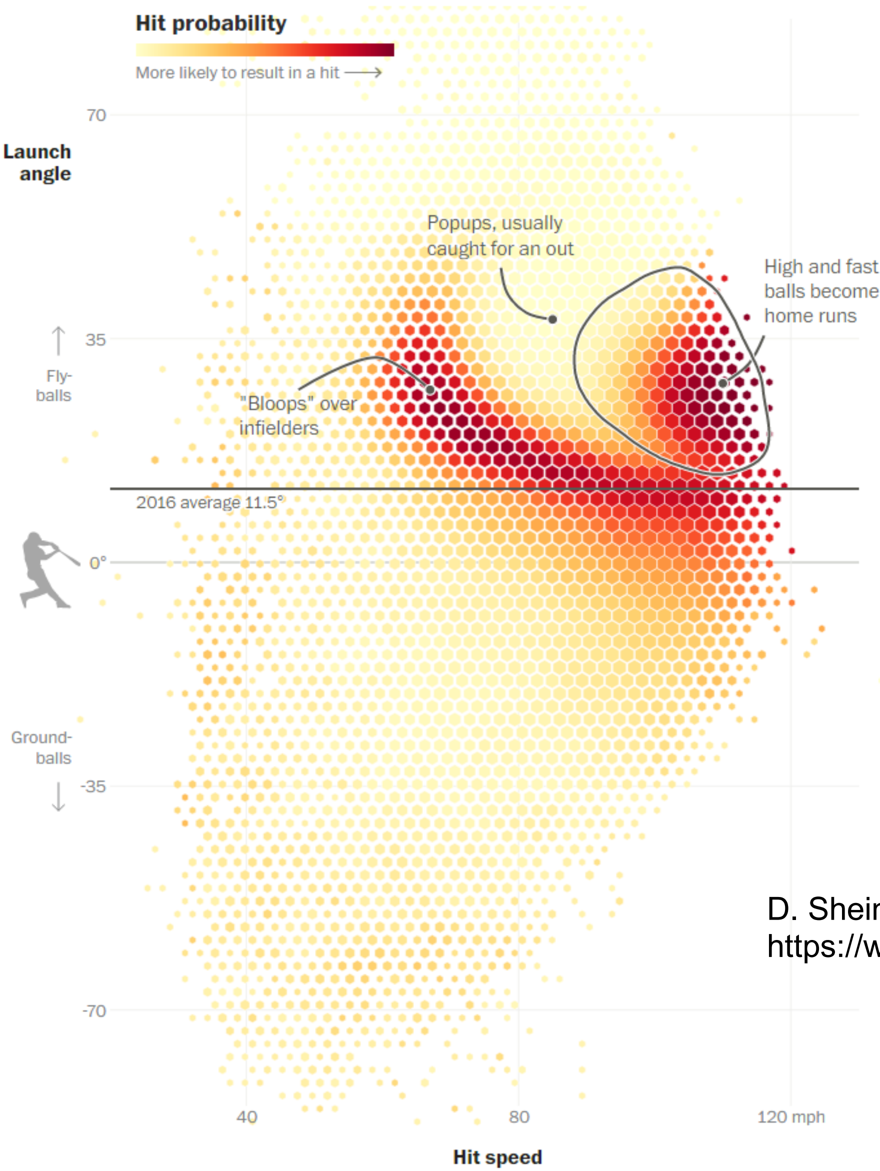
M.B. McCurdy, @ineffectivemath, https://twitter.com/i/web/status/899721405083906048.

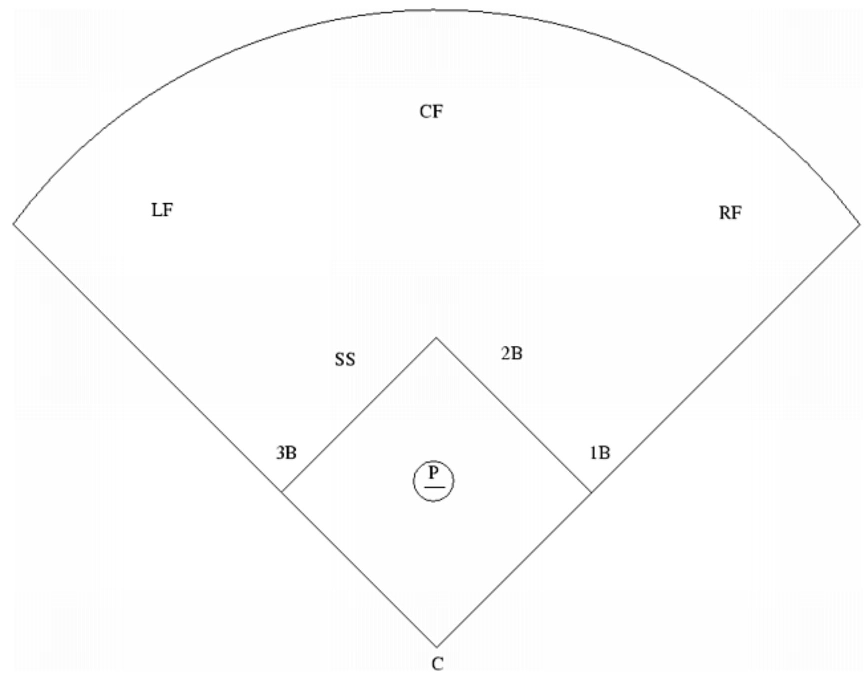
NBA (basketball) – Houston Rockets



K. Goldsberry, Grantland.com, http://grantland.com/the-triangle/future-of-basketball-james-harden-daryl-morey-houston-rockets/.

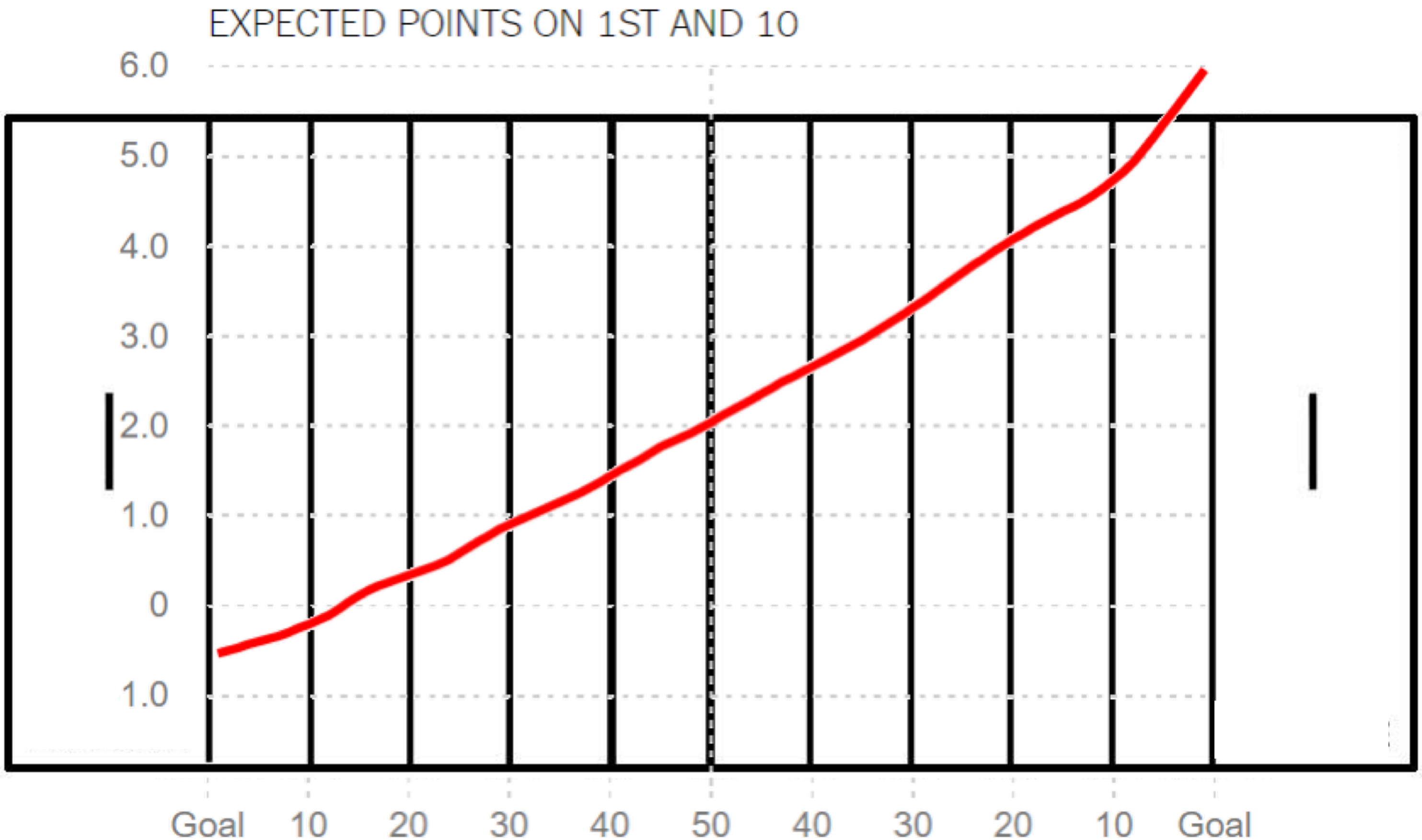
MLB (baseball) – Launch angle and velocity





D. Sheinin and A. Emamdjomeh, Washington Post, https://www.washingtonpost.com/graphics/sports/mlb-launch-angles-story/.

NFL (American football) – 4th Down Bot



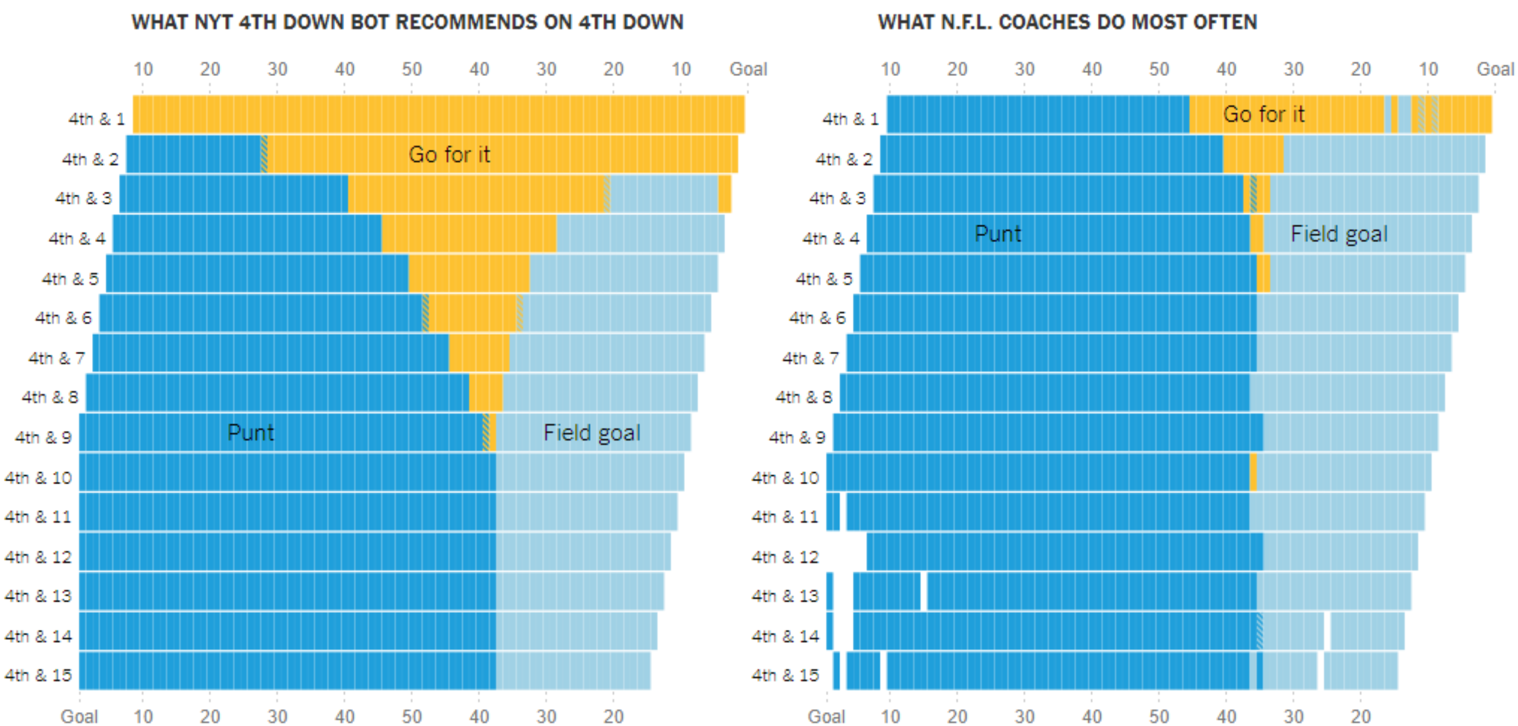
B. Burke and K. Quealy, 4th Down bot, New York Times. http://www.nytimes.com/newsgraphics/2013/11/28/fourth-downs/post.html

NFL (American football) – 4th Down Bot

4th Down: When to Go for It and Why

NYT 4th Down Bot @NYT4thDownBot SEPT. 4, 2014

32



B. Burke and K. Quealy, 4th Down bot, New York Times. http://www.nytimes.com/newsgraphics/2013/11/28/fourth-downs/post.html

2. Team performance evaluation and prediction of future outcomes

Motivation for team performance evaluation and prediction

- **Unbiased evaluation of performance**
  - Signal vs. noise
  - Strength of schedule
- **Strategy and planning**
  - Team building and “tanking”
- **Storytelling and entertainment**
- **Betting analytics**
  - Betting lines
  - Predictive analytics

Team performance evaluation by ranking and rating

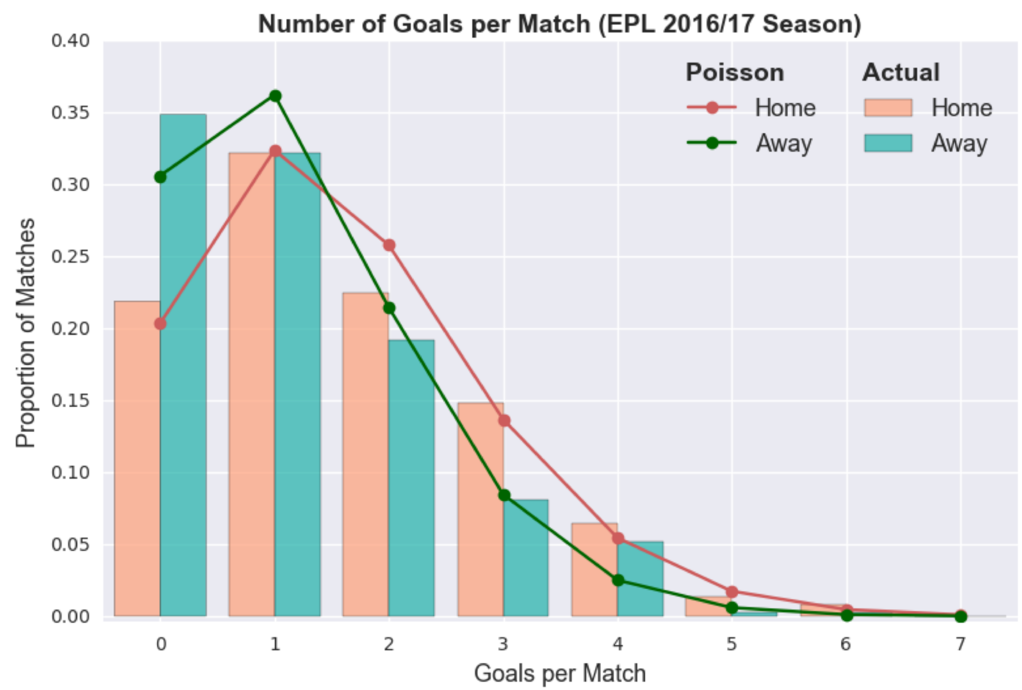
- **The game results depend on (at least) three factors**
  - Home advantage
  - Strength of the teams
  - Random variation (stochastic component)
- **The game results are observed and the teams are ranked or rated according to their perceived level of performance.**
- **The objective of ranking and rating of teams is compare the underlying strengths of the teams.**
  - Ranking: ordinal scale, i.e., the separation between successive teams is not evaluated.
  - Rating: interval scale, i.e., the differences between teams are measurable and have an meaningful interpretation.
- **Team ratings can be used for predicting the winners of future games**

Prediction and winning probability

- **Prediction of future results**
  - When estimates for team strengths have been calculated, they can be used for estimating winning probabilities in future games.
- **Modeling approach depends on the rules and the scoring system**
  - How are the points/goals scored?
  - Assumptions about the underlying scoring processes
  - N.B., There are always a number of alternative modeling choices

Football

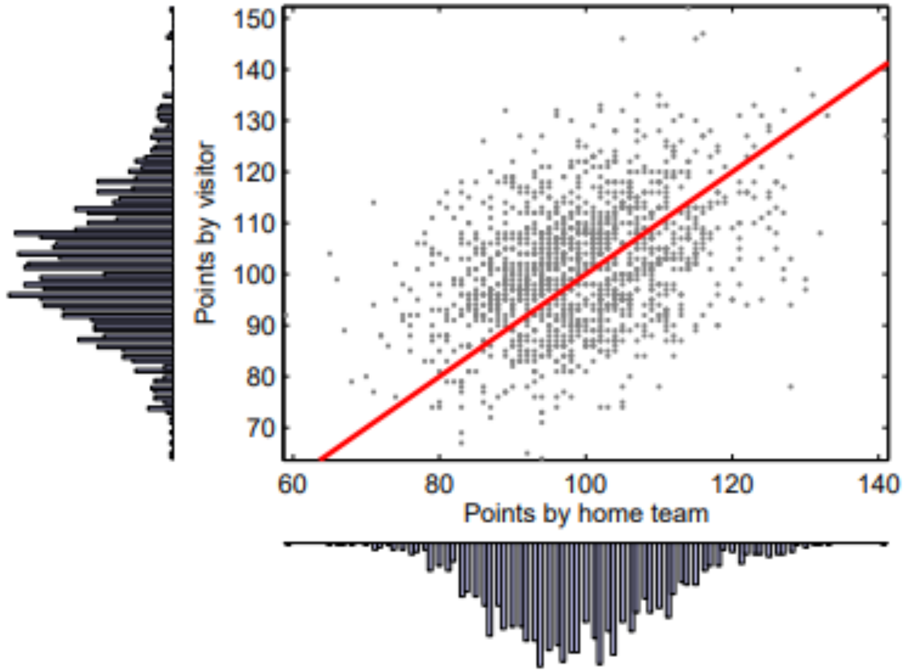
- **Low scoring game**
  - Limited number of scoring chances
  - EPL: 2.77 goals/game in 2016-17
- **Poisson distribution**
  - Scoring intensity
    - *“Small chance of a goal at every time instant”*
  - Rough approximation



https://dashee87.github.io/football/python/predicting-football-results-with-statistical-modelling/

Basketball

- **High number of scoring chances**
  - NBA teams average ≈100 possessions per game
  - Consecutive offensive possessions are more or less independent
- **Central limit theorem**
  - Distribution of points can be approximated using a normal distribution



J. Poropudas, *Kalman filter algorithm for rating and prediction in basketball*, 2011.

How certain is the outcome of the game?

- **Law of large numbers**
- **Probability of an “upset”**
  - In football, a match between a very good and a very bad team can still result in a tie or even an upset.
  - In basketball, the better team usually wins.

Bradley-Terry model

- **Flexible model for almost(?) any game with two teams/players**
  - Bernoulli trial: first team either wins or doesn't.
  - Outcome of each game is 0 or 1.
  - Home advantage and scoring margin are **not** considered.
- **Parameters**
  - Team ratings α_i representing team strengths
- **Winning probability when team i meets team j:** $\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_i - \alpha_j$
- **Parameter estimation using maximum likelihood**
  - No closed form solution
  - Numerical methods

$$\ell(\boldsymbol{\alpha}) = \sum_{i}^{n}\sum_{j}^{n}\left(w_{ij}\log(\alpha_i) - w_{ij}\log(\alpha_i + \alpha_j)\right)$$

E. Zermelo, Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung, *Mathematische Zeitschrift*. **29** (1): 436-460, 1929.
R.A. Bradley and M.E. Terry, Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons, *Biometrika,* **39** (3/4): 324-345, 1952.

Maher’s model for football

- **“Scoring margin contains information.”**
- **Poisson scoring for home team and visiting team:**

$$Y_H \sim Poisson(\alpha_i \cdot \beta_j)$$
$$Y_V \sim Poisson(\delta_j \cdot \gamma_i)$$

- **Four parameters per team**
  - Offense at home and away: α_i and δ_i
  - Defense away and at home: β_i and γ_i
  - Number of parameters can decreased with equality constraints.
- **Parameter estimation using maximum likelihood**
  - No closed form solution
  - Numerical methods

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i}^{n} \sum_{j}^{n} \left(y_{ij}\alpha_i\beta_j - y_{ij}\log(\alpha_i\beta_j)\right)$$

- **Not a “perfect fit” to actual data**
  - Independence assumption!

M.J. Maher, Modelling association football scores, *Statistica Neerlandica*. **36** (3): 109-118, 1982.

Dixon-Coles model for football

- **Refinement of the Maher's model**
  - Modification to outcomes 0-0, 1-0, 0-1, and 1-1
  - Dependence between teams' scoring
- **Better fit to actual results**
- **Parameter estimation using maximum likelihood**

Instead, we propose the following modification of model (4.1):

$$\Pr(X_{i,j} = x,\ Y_{i,j} = y) = \tau_{\lambda,\mu}(x, y)\frac{\lambda^x \exp(-\lambda)}{x!}\frac{\mu^y \exp(-\mu)}{y!} \tag{4.2}$$

where

$$\lambda = \alpha_i \beta_j \gamma,$$
$$\mu = \alpha_j \beta_i$$

and

$$\tau_{\lambda,\mu}(x, y) = \begin{cases} 1 - \lambda\mu\rho & \text{if } x = y = 0, \\ 1 + \lambda\rho & \text{if } x = 0,\ y = 1, \\ 1 + \mu\rho & \text{if } x = 1,\ y = 0, \\ 1 - \rho & \text{if } x = y = 1, \\ 1 & \text{otherwise.} \end{cases}$$

In this model, ρ, where

$$\max(-1/\lambda,\ -1/\mu) \leqslant \rho \leqslant \min(1/\lambda\mu,\ 1),$$

enters as a dependence parameter: $\rho = 0$ corresponds to independence, but otherwise the independence distribution is perturbed for events with $x \leqslant 1$ and $y \leqslant 1$. It is easily checked that the corresponding marginal distributions remain Poisson with means λ and μ respectively.

M.J. Dixon and S.G. Coles, Modelling association football scores and inefficiencies in the football betting market, *Applied Statistics,* **46** (2): 265-280, 1997.

3. Course assignment: Optimal betting portfolio for Liiga Playoffs

Finnish ice hockey league: Liiga

- **Top Finnish Ice Hockey League**
  - 15 teams
  - 60 games for each team (30 home games)
  - 10 teams qualify for the playoffs
  - See, http://liiga.fi/ottelut/2018-2019/runkosarja/.
- **Regular season ends 14.3.2019**
- **Preliminary playoffs end 19.3.2019**
  - N.B., you can use all the information available up to that date in your project work.
- **Deadline for this project**
  - Presentation due 1.4.2019
  - Report due 13.4.2019

Liiga standings (as of 10.3.2019)

#	Joukkue	O	V	T	H	TM	PM	LP	P
1.	Kärpät	58	40	7	11	201	95	6	133
2.	Tappara	58	31	9	18	172	145	3	105
3.	Pelicans	58	29	11	18	192	150	3	101
4.	TPS	58	28	11	19	158	146	5	100
5.	HPK	59	24	16	19	165	146	8	96
6.	HIFK	58	23	18	17	176	164	8	95
7.	Lukko	58	25	13	20	164	157	5	93
8.	Ilves	58	22	14	22	163	165	8	88
9.	SaiPa	58	21	15	22	152	152	10	88
10.	JYP	59	17	18	24	140	151	11	80
11.	Sport	58	16	19	23	177	199	11	78
12.	KalPa	58	17	14	27	144	181	6	71
13.	KooKoo	58	18	11	29	149	189	5	70
14.	Jukurit	58	12	19	27	136	174	8	63
15.	Ässät	58	9	13	36	116	191	7	47

http://liiga.fi/tyokalut/laskuri/

Liiga playoff format

- **Six best teams at the conclusion of regular season proceed directly to quarter-finals**
- **Teams placing between 7th and 10th (inclusive) will play preliminary play-offs ("wild card round") best-of-three**
  - The two winners of the preliminary playoffs take the last two slots to quarter-finals
- **All series after this are best-of-seven**
- **In all playoff series, the team with the higher playoff seed holds the home advantage.**
- **In the semifinals, the matchups are determined based on the regular season and the best team plays against the worst team ("re-seeding").**
- **N.B., you can skip the preliminary playoffs, if you like.**

Liiga playoffs (last season)

Wild-card round (best-of-3)

7	**SaiPa**	**2**
10	Pelicans	1

8	**Ässät**	**2**
9	Lukko	0

Quarter-finals (best-of-7)

1	**Kärpät**	**4**
8	Ässät	1

2	**TPS**	**4**
7	SaiPa	2

3	**Tappara**	**4**
6	KalPa	2

4	JYP	2
5	**HIFK**	**4**

Semi-finals (best-of-7)

1	**Kärpät**	**4**
5	HIFK	3

2	TPS	0
3	**Tappara**	**4**

Finals (best-of-7)

1	**Kärpät**	**4**
3	Tappara	2

Bronze medal game

2	TPS	0
5	**HIFK**	**1**

Poisson regression

- **Poisson regression**
  - Generalized linear model form of regression analysis for count data
  - Assumption: the response variable Y follows a Poisson distribution
- **If $x \in \mathbb{R}^n$ is a vector of independent variables, the Poisson regression model takes the form**

  $\log E(Y|x) = \theta^T x$ **, where** $\theta \in \mathbb{R}^{n+1}$
- **Given a Poisson regression model θ and input vector x**

  $Y|x \sim Poisson(\ exp(\theta^T x)\ )$
- **If y_i are independent observations with corresponding values x_i of the predictor variables, then θ can be estimated using maximum likelihood method.**
  - No closed-form expression
  - Numerical methods

$$\ell(\theta) = \sum_{i=1}^{m} (y_i \theta^T x_i - \exp(\theta^T x_i))$$

- **R has a built in function glm() that can fit Poisson regression models.**

Poisson regression for team ratings in ice hockey

- **For a league with n teams, the parameters of the model are**
  - Home advantage μ
  - Team i offensive strength α_i (n parameters)
  - Team i defensive strength β_i (n parameters)
- **Parameters are collected to a vector**

$$\boldsymbol{\theta} = (\mu, \alpha_1, \ \dots, \ \alpha_n, \beta_1, \dots, \beta_n)$$

- **Identifiability:** $\beta_n = 0.$
- **When team i hosts team j:**

$$\log\big(E(Y_H | i, j)\big) = \mu + \alpha_i - \beta_j$$
$$\log\big(E(Y_V | i, j)\big) = \alpha_j - \beta_i$$

- **N.B., higher parameter estimates indicate better offenses and defenses.**

Poisson regression for team ratings in ice hockey

- **Each team has two ratings**
  - Offensive strength
  - Defensive strength
- **Home advantage is included in modeling the goals of the home team.**
  - Home advantage is assumed to be equal for all teams.
- **Goals scored by the two teams are modeled separately and assumed to be independent.**
- **Each match is essentially two observations**
  - The number of goals for the home team
  - The number of goals for the visiting team
  - N.B., each match needs two rows in our data set, not just one
- **R has a built in function glm() that can fit Poisson regression models.**

Estimation of winning probabilities (single game)

- **Distributions of the home and visitor goals**

$$Y_H \sim Poisson\left(\exp(\mu + \alpha_i - \beta_j)\right)$$

$$Y_V \sim Poisson\left(\exp(\alpha_j - \beta_i)\right)$$

- **Probabilities $P(Y_H > Y_V)$ and $P(Y_H < Y_V)$ can be estimated by enumerating "all" goal combinations or by using Monte Carlo simulation.**
- **Home team wins the game, if $Y_H > Y_V$.**
- **Visiting team wins, if $Y_H < Y_V$.**
  - N.B., in playoffs a tie is not allowed (overtime and penalty shootout).
  - Ignore ties by flipping a coin OR re-scaling the probabilities $P(Y_H > Y_V)$ and $P(Y_H < Y_V)$ so that their sum is equal to one.

Estimation of winning probabilities (playoff series)

- **Best-of-three playoff series**
  - Games are played until first team reaches two wins
- **Best-of-seven playoff series**
  - Games are played until first team reaches four wins
- **In Liiga playoffs, the home team alternates**
  - First game is hosted by the higher seed
  - Second by the lower seed
  - Third by the higher seed, *etc.*
  - N.B., the home advantage “switches sides” from game to game.
- **The winner of a playoff series advances to the next round.**

Estimation of winning probabilities (championship)

- **To win the championship, a team has to win three playoff series (and a potential preliminary playoff)**
  - N.B., the winning probability for each playoff series depends on the both teams playing.
- **Monte Carlo simulation**
  - Generate random samples of game results (Y_H, Y_V) for each game of the playoff series.
  - Determine winner for the playoff series.
  - Move to the next playoff series (or next round).
- **Simulate the entire playoffs for, say, $N = 10000$ times to estimate the winning probabilities $p = (p_1, \dots, p_n)$.**
  - N.B., you only need to keep track of the champion for each simulation run.

Construction of betting portfolio

- **Maximize the expected value of the betting portfolio by allocating a budget of $M = 1000$ euros to the teams.**
  - In order to alleviate the risk related to the portfolio, no more than 50% of the budget should be allocated to any single team.

Decimal odds for betting

- **The payment for a successful bet is the product of the money at stake and the decimal odds.**
  - Decimal odds reflect the inverse of the implied success probability.
- **If the chosen team doesn't win, the stake is lost.**

Team	Decimal odds
Kärpät	1.79
Tappara	9.13
Pelicans	13.27
TPS	12.79
HIFK	13.70
HPK	17.44
Lukko	47.95
Ilves	95.90
JYP	120.00
SaiPa	190.00
Sport	480.00

Special thanks to Teemu Eirtovaara at Veikkaus.

Any questions?