How bad should I win next faceoff?

Big data cup 2021

During the course of a hockey game, there are is a lot of faceoffs (between 50 and 75 most of the time). The outcome of these events usually dictates which team will control the puck at the beginning of a sequence. Thereby, we can say that every faceoff is important. However, many would agree that some faceoffs are more important than others. As an example, a faceoff in the offensive zone with less than a minute on the clock when trailing by one goal would appear to be quite important. In that case, we can attach the importance to the moment in the game, but if we look at it in a broader sense, does winning a faceoff contributes to scoring a goal? In this report, we try to quantify the importance of winning faceoffs, breaking it down by situation as much as possible. More precisely, we focus on the question

> **Can we say that offensive zone faceoffs are an important aspect of offensive successes (e.g. scoring a goal), and to what extent?**

Towards this end, we first assess whether the data matches our intuition that winning a faceoff in the offensive zone does increase the chances of scoring a goal during the sequence that follows. We then take a more in-depth look at sequences, breaking them down by seconds.

SAM-HERE

As you may noticed, we focused on offensive zone faceoffs. The reason for that is we think their impact on scoring a goal is more obvious as the team is already close to the opponent net. In a way, we could also say we included defensive zone faceoffs, as it only depends from which team point of view you are looking at it. We decided to use the Erie Otters data made available in the scounting dataset. Our rationale for using this dataset was to use the data of one single team, and not bother having teams that played different amount of games. In that context, the `scounting` dataset is the one with the most observations. For simplicity, we can define few terms we will use throughout this analysis. These abbreviations have been defined from Erie Otters point of view (à voir si on veut garder ça):

- FO: Faceoff
- OZ: Erie Otters Offensive Zone
- DZ: Erie Otters Defensive Zone
- FO_win: Faceoff won by Erie Otters
- GF: Goal For Erie Otters
- GA: Goal Against Erie Otters

In the next sections, we are first going to analyze the data sequence-by-sequence. That way, we will be able to conclude if winning an offensive zone faceoff does have a significant impact on scoring a goal on that given sequence. Afterward, we will breakdown our sequences more in details, for example second-by-second, and see the impact of winning offensive zone faceoffs over time. We will also see if we can draw conclusions from different contexts, such as faceoffs that take place on power play/penalty killing situations or during overtime. Finally, we will see if we can draw more robust conclusions over a bigger dataset, such as multiple years of NHL data fetched using `tidynhl`, a R package developed on top of NHL Open API.

What is the overall impact of winning a faceoff?

As a first step, we will try to figure out what is the overall effect of winning or loosing an offensive zone faceoff. Let's start by making sure we have a good understanding of the data we are working with. As mentioned in the introduction, we focused on the `scouting` dataset. In the table 1, we can see basic informations about this dataset, and what might be relevant to know for our analysis.

Table 1: High-level features of the scouting dataset

Min. game date	Max. game date	Games	FOs	OZ/DZ FOs	Goals	OZ/DZ Goals
2019-09-20	2020-03-08	40	2441	1644	293	122

We have 40 games of data, with 2441 overall faceoffs. However, only 1644 faceoffs will be relevant in our case since we focus on those that happened in the offensive/defensive zones. We have 293 scored goals in total, but only 122 were actually scored by the team that started a sequence from their offensive zone.

Talking about sequences, we structured the data in a way that we can easily analyze each sequence individually. To illustrate that, we added a preview (see table 2) of what our transformed data looks like on a sequence-by-sequence basis.

Table 2: Preview of our sequence-by-sequence data

game_date	period	clock_begin	clock_end	length_seconds	FO_win	FO_zone	GF	GA
2019-09-20	1	20:00	18:57	63	FALSE	red	FALSE	FALSE
2019-09-20	1	18:57	18:29	28	FALSE	offense	FALSE	FALSE
2019-09-20	1	18:29	15:27	182	FALSE	offense	FALSE	FALSE
2019-09-20	1	15:27	14:10	77	FALSE	offense	FALSE	FALSE
2019-09-20	1	14:10	13:42	28	FALSE	blue_offense	TRUE	FALSE
2019-09-20	1	13:42	12:41	61	TRUE	red	FALSE	FALSE

In the table 2, we omitted to show some additional columns that bring contextual informations about the faceoff. The more obvious example of that is power player and penalty killing situations. In the table 3, we did breakdown the sequences by context and we also added the scoring success rate in each of these situations.

Table 3: Contextual data for faceoff situations

	Penalty kill			Even strength			Powerplay		
Zone	FOs	Goals	% success	FOs	Goals	% success	FOs	Goals	% success
Defensive									
FO lost	129	12	9.3 %	283	19	6.7 %	7	0	0 %
FO won	71	8	11.3 %	292	19	6.5 %	10	1	10 %
Offensive									
FO lost	10	0	0 %	390	24	6.2 %	93	10	10.8 %
FO won	4	0	0 %	245	13	5.3 %	110	16	14.5 %

A first conclusion we can draw from table 3 is that we don't have a lot of goals scored by Erie Otters (63) from sequences that started in their offensive zone. We will see later if it's sufficient to draw significant conclusions. However, we strongly doubt that it will be sufficient for drawing conclusions on a more granular basis, such as power play or penalty killing situations.

Another thing we can notice from the table 3 is the success rate differences between sequences that started with a won or a lost faceoff. From the data we have, we can already see the effect of winning the faceoff on power play situations. For even strength contexts, the impact looks less significant (at least for Erie Otters).

To see if the effect of winning an offensive zone faceoff is significant, we fitted a logistic regression for which we defined the variable `GF` as a target and `FO_win` as our only feature. In that specific case, we are trying to assess what is the impact of winning an offensive zone faceoff on Erie Otters probability to score a goal. From this fitted model, we got a $\beta_{FO\_win} = 0.171$ with a p-value of 0.516. From these numbers, we can conclude

that winning an offensive zone faceoff seems to increase the chances to score a goal during the sequence that follows, but we can't say it's significant. This may be caused by a lack of data, or also by mixing up multiple types of sequences together (power play, even strength, overtime, etc).

Breaking down the sequences by seconds

So far, we have gathered *some* evidence that winning offensive faceoffs increases the chances of scoring a goal. Our intuition also tells us that this effect should be most prominent in the very first seconds following the faceoff and that, over time, the fact that one has won or lost the faceoff should become less relevant. To verify that, let us take a closer look at the times the goals in questions were scored. Although we do have the exact goal times (in seconds), we preferred to bin them over windows of 5 seconds.

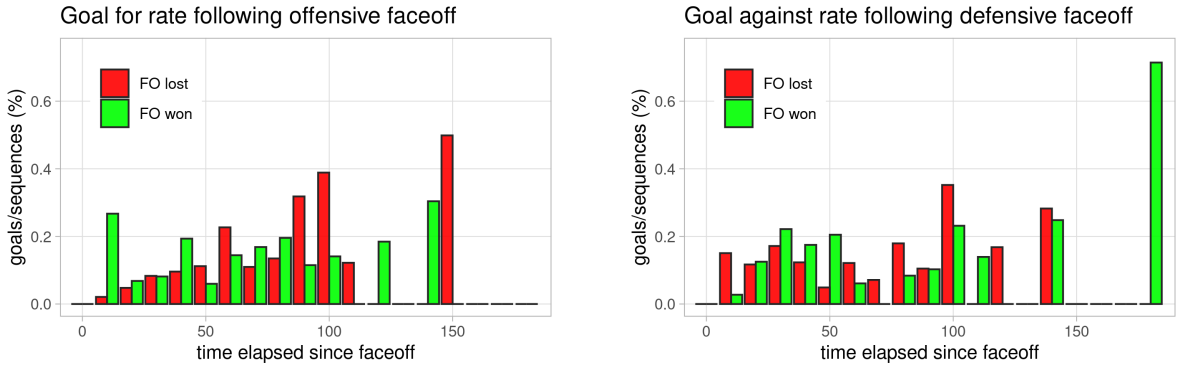

Figure 1: Percentage of times a goal was scored within the timespan $[t, t+5)$ for $t = 0, 5, 10, ..., 175$, **given that the sequence in question lasted at least t seconds**.

Figure 1 is encouraging: It suggests that winning an offensive faceoff (or losing a defensive faceoff) leads more often to a *quick goal*, where by *quick goal* we mean within the first five seconds of the sequence. However, note that no bars actually correspond to more than ZZZ goals. The far-right bar on the right panel, which is the tallest, corresponds to only ZZZ goals; what happens is that not many sequences reach the 175 seconds mark, just XX here in fact, so that adding or subtracting a goal can make a major difference. Also, for the same reason, many bars suggest a zero probability that a goal gets scored within their corresponding five-seconds window.

The sparseness in the barcharts of Figure 1, which is due to the lack of data points, makes it hard to interpret them. To overcome such difficulty, let us construct a smooth version of these latter using a generalized additive model (*gam*) and the exact seconds at which the goals were scored. The objective is to construct curves that provides the *risk* of a goal being scored at any time point t provided that the faceoff was either won or lost; one can interpret the value given by the curve as the probability of a goal being scored within a one-second time frame. See the **Technical details** section at the end of the report for a more formal description of the underlying model.

1. discuss results in depth — GIVEN THAT THE SEQUENCE WAS LONGER THAN t.
2. SKIP PARAGRAPH. discuss weakness of loess: no confidence intervals for validating like we did previously.
3. discuss that not much data here, so possibly worthless anyways, (although CI from splines are conclusive).
4. discuss interesting questions that we cannot answer, need more data to use a more granular definition of situation.

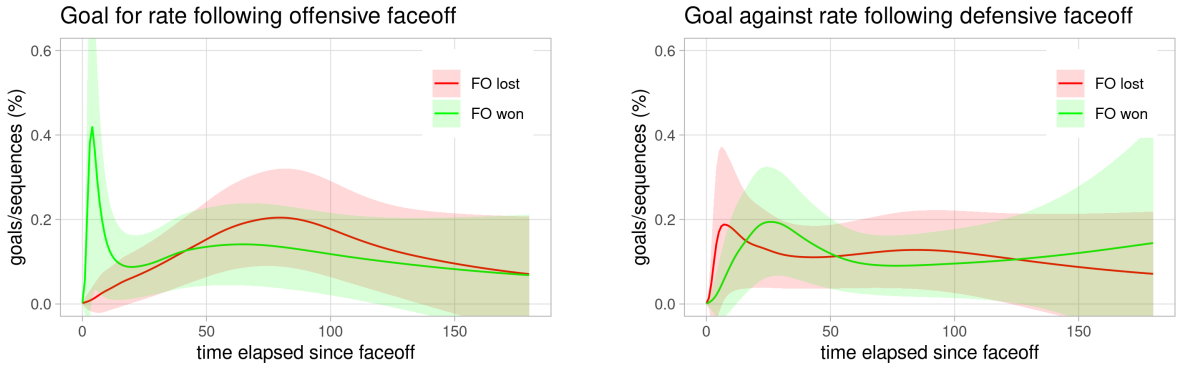

Figure 2: Percentage of times a goal was scored within the timespan $[t, t+1)$ for $t = 0, \ldots, 30$, given that the sequence in question lasted at least t seconds.

Technical details

Figure~(LOESS) Denote $Y_t \in \{0, 1\}$ the random variable indicating whether a goal occurred in the interval of time $[t, t+1)$, and let $\mathbb{E}(Y_t|x)$ be the expectation of Y_t (i.e., the probability of a goal occurring in that timespan) given that the most recent faceoff was lost ($x = 0$) or won ($x = 1$). The loess curves in Figure~(LOESS) were obtained by fitting the *generalized additive model* (gam)

$$g\{\mathbb{E}(Y_t|x)\} = (1-x)f_0(t) + xf_1(t), \qquad g(z) = \ln\{z/(1-z)\}, \tag{1}$$

where g is the so-called logit function and, for both $k = 0, 1$, f_k is an unknown (i.e., to be estimated) nonlinear function of t that approximates $g(Y_t|x = k)$.[1] The probabilities reported are obtained by solving this equation for $\mathbb{E}(Y_t|x)$ at each value of $t \in \{0, 1, 2, \ldots\}$.

Figures~(SPLINES-1) and (SPLINES-2). In Figure~(SPLINES-1), we reproduced the analysis involving the model in (1). This time, however, we used splines (DEF) for approximating the functions f_0 and f_1 of the gam (as opposed to the loess method previously used), which allowed us to construct confidence intervals. The more granular results displayed in Figure~(SPLINES-2) were also obtained by means of a splines-based gam. In this case, we fitted a model that included $2 \times ZZZ = ZZZ$ nonlinear functions of t. In addition to the subscripts $k \in \{0, 1\}$ indicating whether the faceoff was lost or won, we use $\ell \in \{1, \ldots, ZZZ\}$ to refer to each of the **ZZZ** situations of interest (NAME THEM). The resulting model is given by

$$g\{\mathbb{E}(Y_t|x, s)\} = \sum_{\ell=1}^{Z} \mathbb{1}(s = \ell) \times \Big\{(1-x)f_{0\ell}(t) + xf_{1\ell}(t)\Big\}, \tag{2}$$

where g is as in (1). Note that (2) could actually be expressed and fit as ZZZ distinct models. However, this formulation makes it clear how to include further covariates that are known to have a similar effect in multiple situations.

Adding more data

Using the Erie Otters data available in the `scouting` dataset, we can see that winning offensive zone faceoffs has an impact, but we can't say it's statistically significant at this point. The reason we suspect is lack of data. To convinced ourselves, we redone the same kind of analysis, but using the last 2 seasons of NHL data. In that case, we used the data of all teams, as it is more equally spread across all teams. In the table 4), we can see some basic informations about that NHL dataset.

[1] To fit this model, as well as all models discussed in this report, we used the `R` package `gam`. Also note that in this particular case, we actually used $t^* = \log(t+1)$ as the time variable, so as to allow less smoothing near $t = 0$, where the observations are more concentrated. We also gave considerably more weight to the observations with timestamp $t = 0$ to force $f_k(0) \approx 0$ $(k = 0, 1)$.

Table 4: High-level features of the NHL dataset

Min. game date	Max. game date	Games	FOs	OZ/DZ FOs	Goals	OZ/DZ Goals
2018-10-03	2020-03-12	1378	273148	187646	27992	12764

Now we have much more faceoffs, and also much more goals as well. In the same way as we did in the first section of this report, we fitted a logistic regression using the variable `FO_win` as our only feature, and the variable `GF` as a target. Using this data, we obtained a $\beta_{FO\_win} = 0.303$ with a p-value of $5.8808827 \times 10^{-31}$. Now that we have more observations, we can conclude that winning an offensive zone faceoff does increase significantly the odds to score a goal during a given sequence.

On sequence-by-sequence basis, we now know that the effect is significant. Considering that, let's recreate the breakdown by seconds as we did in the previous section to see how this effect last over time. As a first step, we can look at the goals scored in time, for which we decided to use bins of 2 seconds here (as we have more data).

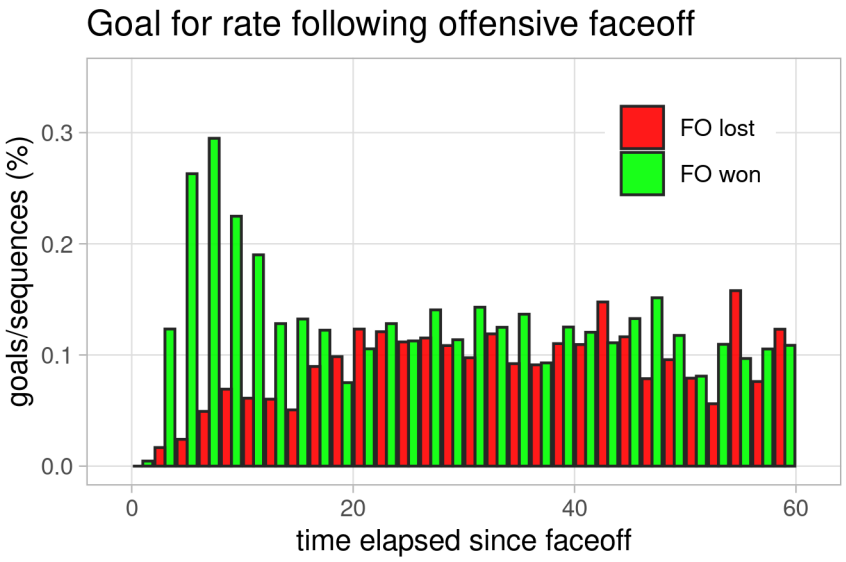

Figure 3: Percentage of times a goal was scored within the timespan $[t, t+2)$ for $t = 0, 2, 4, ..., 60$, **given that the sequence in question lasted at least t seconds**.

From the figure 3, we can already see that the effect of winning an offensive zone faceoff is more obvious within 20 seconds after a faceoff. With this dataset, it is now more clear compared to what we had in the figure 1. Let's construct the smoother version of this bar chart using our gam model (see figure 4).

In the figure 4, we can clearly see the spike when winning the faceoff in the moments following a faceoff. When can also see that the confidence intervals are now much more compact, and do not overlap between each others (at least in the first 20 seconds approximately).

Let's wrap this up

Résumer no highlight Ouverture sur quoi on pourrait essayer deplus

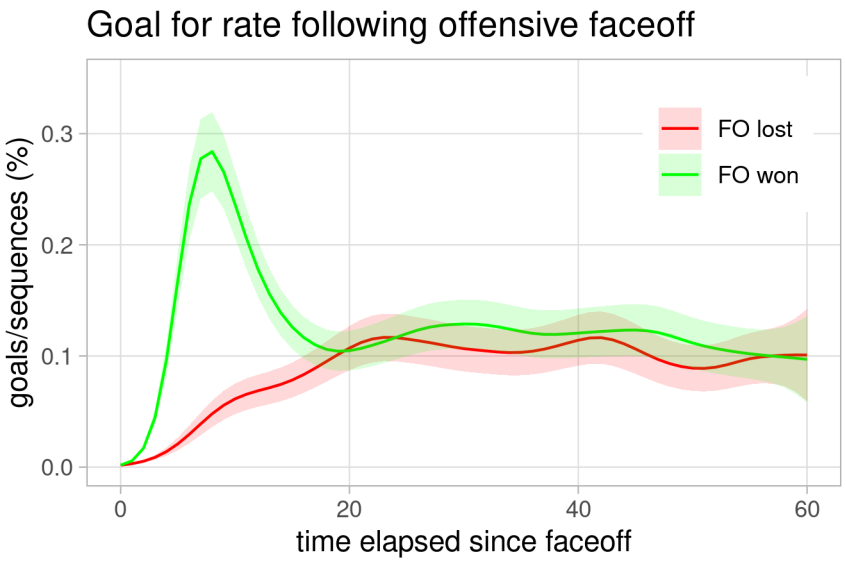

Figure 4: Percentage of times a goal was scored within the timespan $[t, t+1)$ for $t = 0, \ldots, 60$, given that the sequence in question lasted at least t seconds.