# Statistics Resources with Non-Traditional Sports Data

Department of Mathematics, Computer Science, Statistics, and Data Science
Advisors: Professors Robin Lock, Michael Schuckers, Ivan Ramler, A.J. Dykstra

## What is SCORE?

The SCORE Network is an NSF funded organization that…

- Develops and distributes Sports Content for Outreach, Research, and Education (SCORE)
- Seeks to implement an educational framework based on real-world problems and applications
- Increases student likelihood to be engaged in the classroom

The St. Lawrence chapter of SCORE focuses on using non-traditional sports data to develop introductory- level statistics resources for educators.

## Olympic Medals: Data Cleaning and Summarization

**Data Info:** Olympics info about both medals and athletes

**Module Goals:**

- Understand what is considered tidy data
- Use R to clean and format data correctly for future use

Eric Seltzer (Data Science and Finance)

## Giant Slalom: Paired Data and Data Tidying

**Data Info:** Alpine ski data scraped from FIS website containing information on two runs of women's GS at Mont Tremblant

**Module Goals:**

- Use paired data to perform a test for difference in means, find a confidence interval, obtain summary statistics, and interpret findings
- Clean an untidy data set using tidyr and dplyr

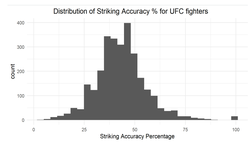
Emilia Agostinelli (Statistics)

## UFC: The Normal Distribution

**Data Info:** Defense and striking statistics for each fighter that competed in the UFC from 1993 to 2021.

**Module Goals:**

- Using the Normal distribution in a real-world application
- Identifying proportions, quartiles and ranges
- Using StatKey given a mean and standard deviation



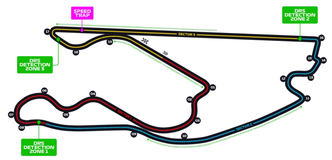

Brendan Karadenes (Data Science)

## Formula 1: Histograms and Outliers

**Data Info:** Lap times for each driver (and constructor) that participated in the 2023 F1 Miami Grand Prix

**Module Goals:**

- Understand histograms and outliers and their relevance in statistical analysis.
- Highlight use of summary statistics in outlier detection



Norah Kuduk (Computer Science and Statistics)

## Nordic Ski: Randomized Block and dplyr

**Data Info:** Data scraped from the FIS website about a women's 10k race in Norway. The dataset has 61 skiers with 20 variables at 1.3k, 4.3k, 7.5k, and 10k intervals

**Module Goals:**

- Use randomized block design to assess difference in mean speed based on skier and distance interval
- Use dplyr to calculate the speed of each skier at each distance

Abigail Winston Smith (Data Science and History)

## Diving: Difference in Means Hypothesis Tests

**Data Info:** Diving results from the 2022 FINA Junior World Championships from women divers aged 16 – 18.

**Module Goals:**

- Conduct difference in means hypothesis tests to test if there is a significant difference in mean points scored between divers of different ages
- Practice conducting hypothesis tests in R or by hand.

Emma Deering (Data Science and Geology)

## League of Legends: Chi-squared Distribution

**Data Info:** Analyzing synergistic relationships between ADC and Support champions in League of Legends.

**Module Goals:**

- Utilize Chi-Squared analysis to explore significant relationships between top and bottom five most popular champions
- Understand use of Chi-Squared to identify statistical significance


George Charalambous (Data Science)

## SCORE Data Repository

Datasets in the SCORE repository come from a variety of sports, and consist of…

- The motivation behind using the data
- A description of the dataset (size and variables)
- A variety of questions that could be answered using the data

Each of the datasets highlighted in the previous sections has been published in the SCORE Data Repository to be used internationally