# Data Viz Module Project

Xander Johansson

2024-10-22

## Contents



## TOPIC

General Topic/Overview: **Contextualizing Sports Absurdities and Performance**

For this report, the focus is on contextualizing the extraordinary and historic achievement of Gretchen Walsh's record-breaking 100 butterfly swim on March 22nd, 2024 at the 2024 NCAA Division I Women's Championships.

**Important Notes/Abbreviations:**

- W = Women's, M = Men's
- SCY = Short Course Yards (NCAA Distance)
- Fly = Butterfly

*Short Course Yards (SCY)* refers to the length of the pool/lane. SCY swims take place in pools/lane that are 25 yards in length. This differs from other formats such as LCY (Long Course Yards: 50 yards), SCM (Short Course Meters: 25 meters), and LCM (Long Course Meters: 50 meters).

Swim events follow this format: Gender, Distance, Stroke, Measurement. For example, *W 100 Fly SCY* stands for Women's 100 Yard Butterfly, Short Course Yards.

## QUESTIONS

- How does Gretchen Walsh's performance in the 100 fly compare to other top performers in the same event over recent years?
- How has the Short Course Yards (SCY) Women's 100 Fly record changed over time, and what role did Gretchen Walsh play in its evolution? How does Walsh's performance compare to previous record-breaking achievements in the same event?
- How significant is Gretchen Walsh's record-breaking performance in the broader historical context of SCY achievements? To what extent did Walsh's swim demonstrate dominance in the sport?

## AUDIENCE

Fans of swimming, sports enthusiasts, analysts, and journalists interested in understanding and appreciating the significance of Gretchen Walsh's achievement.

## DATA

The data for this analysis was obtained from two main sources on March 23rd & 24th, 2024:

1. **USA Swimming**: The national governing body for competitive swimming in the United States. They collect and store data from officially sanctioned meets across the country. The data accessed included:
    - Top 100 Performers SCY W 100 Fly (Record Progress.csv): Data on the top 100 performers in the SCY Women's 100 Fly event from 2020 to 2024, including their times and ranks.

| Variable | Information |
|---|---|
| Time | Length (in Seconds) of the Swim |
| Athlete | Name of the Athlete |
| Swim Date | Date of the Swim |

    - All Time Top Performances W 100 Fly SCY Fly (Top Performers.csv): Data on the top all time W 100 Fly SCY performances, including the swim date, time, and athlete details.

| Variable | Information |
|---|---|
| Rank | Rank of the Swim in the Competition Year |
| Time | Length (in Seconds) of the Swim |
| Full Name | Name of the Athlete |
| Age | Age of the Athlete |
| LSC | Local Swimming Committee (local governing body) |
| Event | Name of the Event (ex: 100 FL SCY) |
| Meet Name | Name of the Meet (ex: 2024 NCAA Division I Women's Championships) |
| Time Standard | Which Meet Does this Swim Meet the Cutoff for? (ex: 2024 Summer Nationals) |
| Competition Year | Competition Year of Swim |

2. **SwimSwam**: A popular swimming news organization known for its coverage of premier swimming events. The data accessed from SwimSwam included:
    - SwimSwam: Gretchen Walsh's Absurd Performance (Gap.csv): Data on the record percentage time difference in all NCAA Championship Meet SCY events, highlighting the significance of Gretchen Walsh's performance. For context, the percentage time difference in SCY event gaps is a common metric used to assess dominance in collegiate swimming. It compares the performance of interest, such as Gretchen Walsh's SCY 100 Fly, with the slowest swim in the 'B Final' (16th place). This comparison allows for a more accurate assessment of a swimmer's dominance relative to their peers, especially compared to simply comparing against the next fastest swimmer, which may not accurately reflect historical dominance.

| Variable | Information |
|---|---|
| Event | Name of the Event (ex: W 100 Fly) |
| 1st | Name of the Winning Athlete (1st in the 'A Final') |
| Time 1st | The Length (in Seconds) of the Winning Athlete's Swim |
| 16th | Name of the Last Athlete (16th Overall, 8th in the 'B Final') |
| Time 16th | The Length (in Seconds) of the Last Athlete's Swim |
| Gap | Percentage Time Difference Between 1st and 16th |

**Load packages**

```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3

## Warning: package 'tidyr' was built under R version 4.3.3

## Warning: package 'forcats' was built under R version 4.3.3
```

**Read in data**

```
library(readr)
progress <- read_csv("Record Progress.csv") #data on the W 100 Fly SCY record since 2016
top <- read_csv("Top Performers.csv") #data on top W 100 Fly SCY swims 2020-2024
gap <- read_csv("Gap.csv") #data on the gap between 1st and 16th of the most dominant swims in SCY hist
```

**Review/clean datasets**

```
#modifying the Record Progress data
progress <- progress %>%
  rename(Date = `Swim Date`) #rename swim date to just date
progress$Date <- as.Date(progress$Date, format = "%m/%d/%Y") #format swim data column as a date type
progress <- progress %>% arrange(Date) #arrange data set by date
#calculate time difference between consecutive points and round to two decimal places
progress$Time_Diff <- c(NA, round(diff(progress$Time), 2))


#modifying the Top Performances Data
top <- top %>%
  select(-FOREIGN, -RESULTS) #deleting unnecessary columns


#modifying the % Time Difference data
gap$EVENT <- factor(gap$EVENT, levels = gap$EVENT[order(gap$GAP, decreasing = FALSE)]) #order event col
gap$GAP <- as.numeric(gsub("%", "", gap$GAP)) #% signs on the graph look very busy so convert % numbers
```

```r
#display the structure of the record progress dataset
str(progress)
```

```
## spc_tbl_ [8 x 4] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Time     : num [1:8] 49.4 49.3 49.3 48.9 48.8 ...
##  $ Athlete  : chr [1:8] "Kelsi Dahlia" "Louise Hansson" "Louise Hansson" "Maggie MacNeil" ...
##  $ Date     : Date[1:8], format: "2016-03-18" "2019-03-01" ...
##  $ Time_Diff: num [1:8] NA -0.09 -0.08 -0.37 -0.05 -0.38 -0.21 -0.83
##  - attr(*, "spec")=
##   .. cols(
##   ..   Time = col_double(),
##   ..   Athlete = col_character(),
##   ..   `Swim Date` = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```r
#display summary statistics of the record progress dataset
summary(progress)
```

```
##       Time         Athlete               Date              Time_Diff
##  Min.   :47.42   Length:8           Min.   :2016-03-18   Min.   :-0.8300
##  1st Qu.:48.41   Class :character   1st Qu.:2019-03-16   1st Qu.:-0.3750
##  Median :48.87   Mode  :character   Median :2022-03-03   Median :-0.2100
##  Mean   :48.74                      Mean   :2021-04-24   Mean   :-0.2871
##  3rd Qu.:49.28                      3rd Qu.:2023-06-10   3rd Qu.:-0.0850
##  Max.   :49.43                      Max.   :2024-03-22   Max.   :-0.0500
##                                                          NA's   :1
```

```r
#display the first few rows of the progress dataset
head(progress)
```

```
## # A tibble: 6 x 4
##    Time Athlete        Date       Time_Diff
##   <dbl> <chr>          <date>         <dbl>
## 1  49.4 Kelsi Dahlia   2016-03-18     NA
## 2  49.3 Louise Hansson 2019-03-01     -0.09
## 3  49.3 Louise Hansson 2019-03-22     -0.08
## 4  48.9 Maggie MacNeil 2021-03-19     -0.37
## 5  48.8 Kate Douglass  2023-02-16     -0.05
## 6  48.5 Kate Douglass  2023-03-17     -0.38
```

```r
#display the structure of the top performers dataset
str(top)
```

```
## tibble [502 x 10] (S3: tbl_df/tbl/data.frame)
##  $ RANK           : num [1:502] 1 2 3 4 5 6 7 8 9 10 ...
##  $ TIME           : num [1:502] 47.4 49.5 49.7 50.2 50.3 ...
##  $ FULL NAME      : chr [1:502] "Walsh, Gretchen" "Shackell, Alex" "Sticklen, Emma" "Crush, Charlot
##  $ AGE            : num [1:502] 21 17 21 15 17 22 22 20 22 22 ...
##  $ LSC            : chr [1:502] "VA" "IN" "ST" "KY" ...
##  $ EVENT          : chr [1:502] "100 FL SCY" "100 FL SCY" "100 FL SCY" "100 FL SCY" ...
```

```
##  $ TEAM NAME       : chr [1:502] "University Of Virginia" "Carmel Swim Club" "University of Texas" "
##  $ MEET NAME       : chr [1:502] "2024 NCAA Division I Women's Championships" "2023 Speedo Winter Ju
##  $ TIME STANDARD   : chr [1:502] "2024 Summer Nationals (LCM)" "2024 Summer Nationals (LCM)" "2024 S
##  $ COMPETITION YEAR: num [1:502] 2024 2024 2024 2024 2024 ...
```

```r
# Display summary statistics of the top performers dataset
summary(top)
```

```
##       RANK             TIME        FULL NAME              AGE
##  Min.   :  1.00   Min.   :47.42   Length:502         Min.   :13.00
##  1st Qu.: 25.25   1st Qu.:51.65   Class :character   1st Qu.:18.00
##  Median : 51.00   Median :52.17   Mode  :character   Median :20.00
##  Mean   : 50.45   Mean   :51.97                      Mean   :19.69
##  3rd Qu.: 75.00   3rd Qu.:52.56                      3rd Qu.:21.00
##  Max.   :100.00   Max.   :53.18                      Max.   :33.00
##      LSC               EVENT            TEAM NAME          MEET NAME
##  Length:502         Length:502         Length:502         Length:502
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  TIME STANDARD      COMPETITION YEAR
##  Length:502         Min.   :2020
##  Class :character   1st Qu.:2021
##  Mode  :character   Median :2022
##                     Mean   :2022
##                     3rd Qu.:2023
##                     Max.   :2024
```

```r
#display the first few rows of the top performers dataset
head(top)
```

```
## # A tibble: 6 x 10
##    RANK  TIME `FULL NAME`        AGE LSC   EVENT      `TEAM NAME`    `MEET NAME`
##   <dbl> <dbl> <chr>            <dbl> <chr> <chr>      <chr>          <chr>
## 1     1  47.4 Walsh, Gretchen     21 VA    100 FL SCY University Of~ 2024 NCAA ~
## 2     2  49.5 Shackell, Alex      17 IN    100 FL SCY Carmel Swim C~ 2023 Speed~
## 3     3  49.7 Sticklen, Emma      21 ST    100 FL SCY University of~ 2024 NCAA ~
## 4     4  50.2 Crush, Charlotte    15 KY    100 FL SCY Lakeside Swim~ 2024 GA Sp~
## 5     5  50.3 Shackley, Leah      17 AM    100 FL SCY Unattached     2024 MA PI~
## 6     6  50.3 Bray, Olivia        22 ST    100 FL SCY University of~ 2024 NCAA ~
## # i 2 more variables: `TIME STANDARD` <chr>, `COMPETITION YEAR` <dbl>
```

```r
#display the structure of the % time difference dataset
str(gap)
```

```
## spc_tbl_ [26 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ EVENT      : Factor w/ 26 levels "M 500 Free","M 1650 Free",..: 26 25 24 23 22 21 20 19 18 17 ...
##  $ 1ST        : chr [1:26] "Caeleb Dressel" "Gretchen Walsh" "Gretchen Walsh" "Katie Ledecky" ...
##  $ TIME (1st) : chr [1:26] "17.63" "47.42" "20.37" "15:01.4" ...
##  $ 16TH       : chr [1:26] "Kristian Gkolomeev" "Farida Osman" "Arianna Vanderpool-Wallace" "Cierra
```

```
##  $ TIME (16th): chr [1:26] "18.64" "50.05" "21.34" "15:40.2" ...
##  $ GAP        : num [1:26] 5.73 5.55 4.76 4.3 4.15 4.07 4.01 3.76 3.64 3.61 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   EVENT = col_character(),
##   ..   `1ST` = col_character(),
##   ..   `TIME (1st)` = col_character(),
##   ..   `16TH` = col_character(),
##   ..   `TIME (16th)` = col_character(),
##   ..   GAP = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
#display summary statistics of the % time difference dataset
summary(gap)
```

```
##          EVENT        1ST             TIME (1st)           16TH
##  M 500 Free : 1   Length:26          Length:26          Length:26
##  M 1650 Free: 1   Class :character   Class :character   Class :character
##  M 200 Free : 1   Mode  :character   Mode  :character   Mode  :character
##  W 200 Back : 1
##  M 200 Fly  : 1
##  W 400 IM   : 1
##  (Other)    :20
##  TIME (16th)             GAP
##  Length:26          Min.   :1.220
##  Class :character   1st Qu.:2.272
##  Mode  :character   Median :3.075
##                     Mean   :3.157
##                     3rd Qu.:3.947
##                     Max.   :5.730
##
```

```
#display the first few rows of the % time difference dataset
head(gap)
```

```
## # A tibble: 6 x 6
##   EVENT       `1ST`          `TIME (1st)` `16TH`             `TIME (16th)`   GAP
##   <fct>       <chr>          <chr>        <chr>              <chr>         <dbl>
## 1 M 50 Free   Caeleb Dressel 17.63        Kristian Gkolomeev 18.64          5.73
## 2 W 100 Fly   Gretchen Walsh 47.42        Farida Osman       50.05          5.55
## 3 W 50 Free   Gretchen Walsh 20.37        Arianna Vanderpoo~ 21.34          4.76
## 4 W 1650 Free Katie Ledecky  15:01.4      Cierra Runge       15:40.2        4.3
## 5 W 100 Free  Gretchen Walsh 44.83        Camille Spink      46.69          4.15
## 6 M 400 IM    Leon Marchand  03:28.8      Kieran Smith       03:37.3        4.07
```

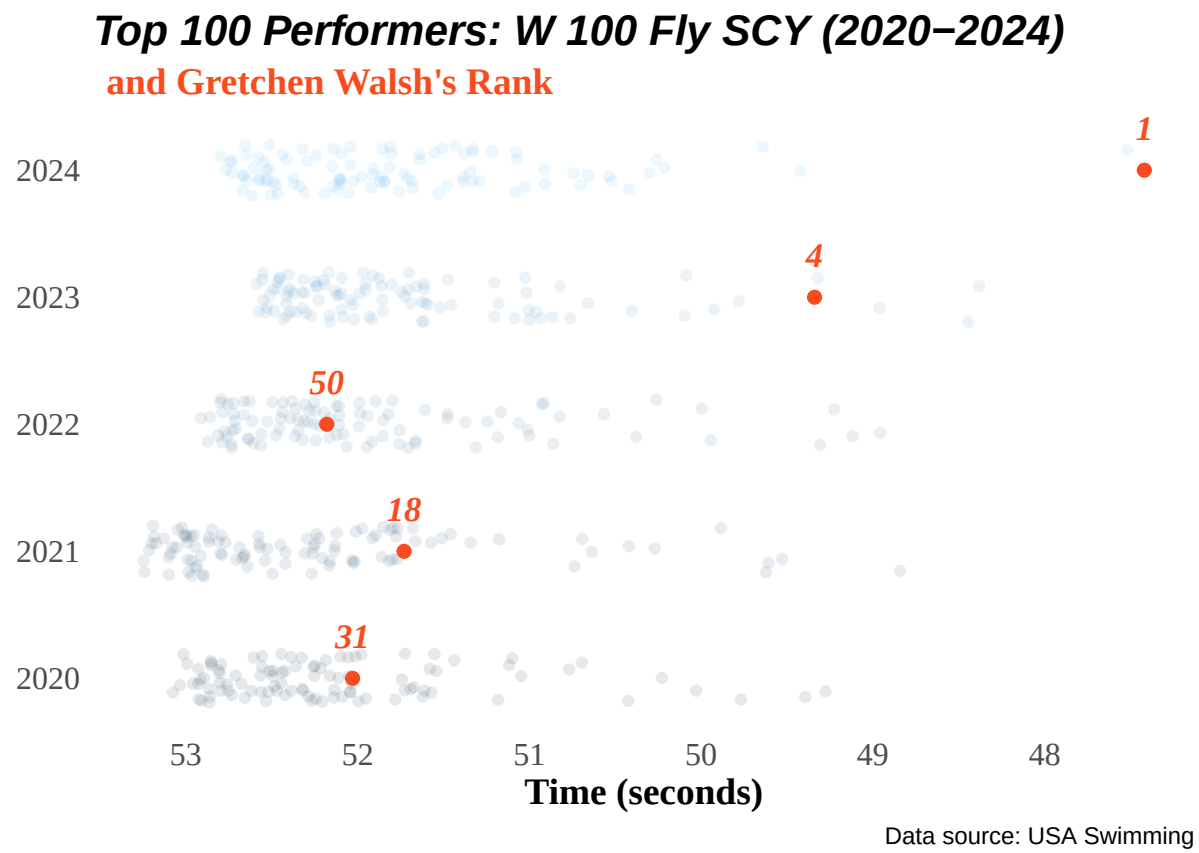
## DATA VIZ

### Data Viz 1

How does Gretchen Walsh's performance in the 100 fly compare to other top performers in the same event over recent years?

```r
ggplot(top, aes(x = TIME, y = `COMPETITION YEAR`, color = `COMPETITION YEAR`)) + #plotting time and com
  geom_point(position = position_jitter(width = .1, height = .2), alpha = .1) +  #adding jitter to the
  geom_point(data = subset(top, `FULL NAME` == "Walsh, Gretchen"), color = "#F84C1E", size = 2) +  #hig
  labs(x = "Time (seconds)", y = "", #add text to title, subtitle, x axis title
       title = "Top 100 Performers: W 100 Fly SCY (2020-2024)",
       subtitle = "and Gretchen Walsh's Rank",
       caption = "Data source: USA Swimming") +
  theme_minimal() + #change theme to minimal
  scale_x_reverse() +  #reverse x-axis so the plot is hotizontal
  geom_text(data = subset(top, `FULL NAME` == "Walsh, Gretchen"), aes(label = RANK), fontface = "bold.i
  theme(plot.title = element_text(size = 16, face = "bold.italic"), #format title, subtitle, and axis t
        plot.subtitle = element_text(size = 14, face = "bold", color = "#F84C1E", hjust = 0.02, family =
        axis.title = element_text(size = 14, family = "serif", face = "bold"),
        axis.text.x = element_text(size = 12, family = "serif"),
        axis.text.y = element_text(size = 12, family = "serif"),
        panel.grid.minor.y = element_blank(), #get rid of gridlines
        panel.grid.minor.x = element_blank(),
        panel.grid.major.y = element_blank(),
         panel.grid.major.x = element_blank()) +
  guides(color = FALSE) #get rid of legend
```

```
## Warning: The ‘<scale>‘ argument of ‘guides()‘ cannot be ‘FALSE‘. Use "none" instead as
## of ggplot2 3.3.4.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```




**Data Viz 2**

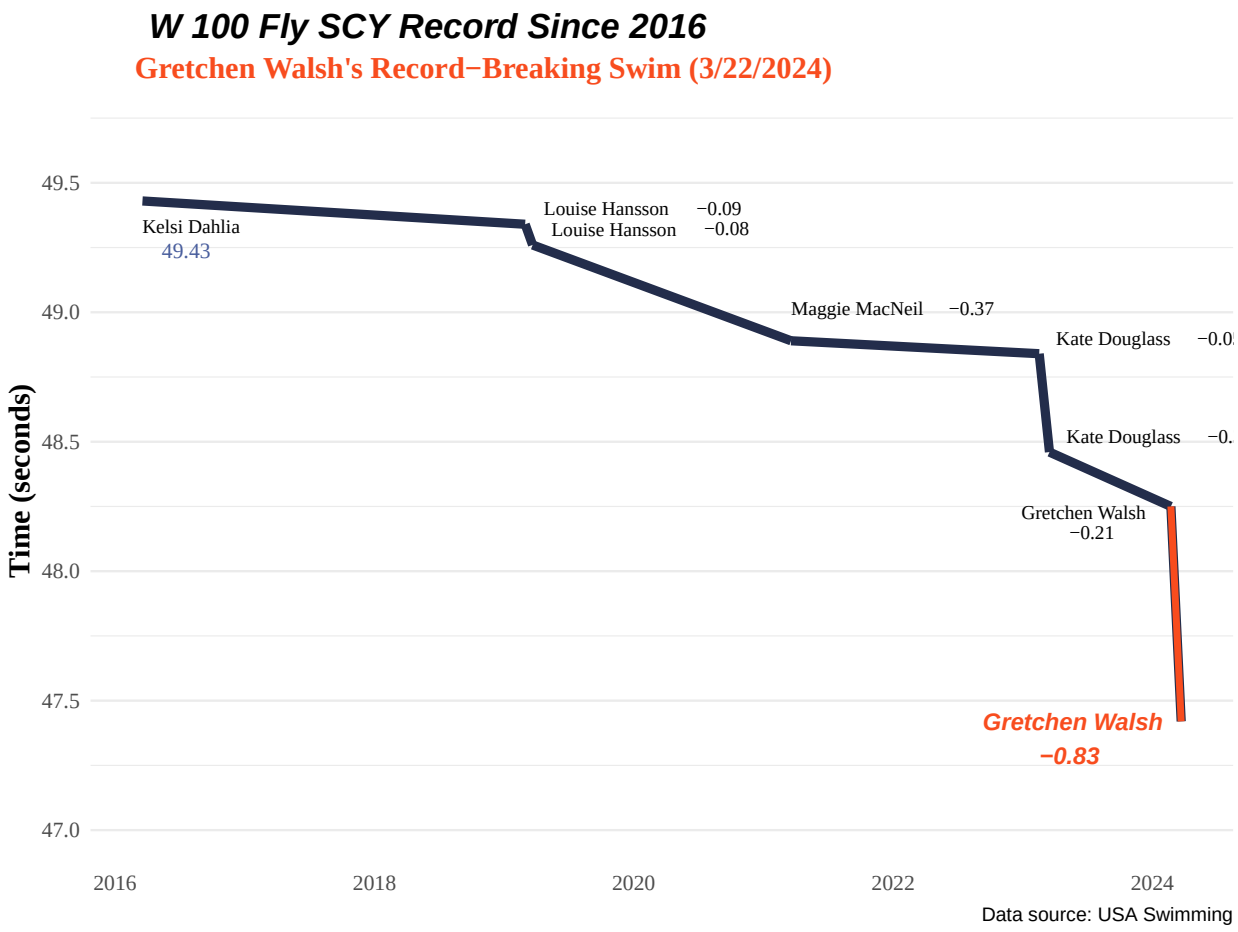
How has the Short Course Yards (SCY) Women's 100 Fly record changed over time, and what role did Gretchen Walsh play in its evolution? How does Walsh's performance compare to previous record-breaking achievements in the same event?

```
ggplot(progress, aes(Date, Time)) + #plot date and time
  geom_segment(aes(xend = lag(Date), yend = lag(Time)), color = "#232D4B", size = 2) + #draw a segment
  geom_segment(data = tail(progress, 2), aes(x = Date, y = Time, xend = lag(Date), yend = lag(Time)), co
  geom_text(aes(label = ifelse(Athlete == "Kelsi Dahlia", Time, "")), color = '#495E9D', hjust = -.4, vj
  geom_text(aes(label = ifelse(Athlete == "Kelsi Dahlia", Athlete, "")), hjust = 0, vjust = 2.5, size =
  geom_text(aes(label = ifelse(Athlete != "Gretchen Walsh" & Athlete != "Kelsi Dahlia" & Athlete != "Mag
  geom_text(aes(label = ifelse(Athlete == "Maggie MacNeil", Athlete, "")), hjust = 0, vjust = -2, size =
  geom_text(aes(label = ifelse(Athlete == "Louise Hansson", Time_Diff, "")), hjust = -3.8, vjust = -0.75
  geom_text(aes(label = ifelse(Athlete == "Maggie MacNeil", Time_Diff, "")), hjust = -3.5, vjust = -2, s
  geom_text(aes(label = ifelse(Athlete == "Kate Douglass", Time_Diff, "")), hjust = -3.5, vjust = -.7, s
  geom_text(aes(label = ifelse(Time == "48.25", Athlete, "")), hjust = 1.2, vjust = 1, size = 3, family
  geom_text(aes(label = ifelse(Time == "48.25", Time_Diff, "")), hjust = 2.25, vjust = 2.5, size = 3, fa
  geom_text(aes(label = ifelse(Time == "47.42", Athlete, "")), hjust = 1.1, vjust = 0.5, size = 4, colo
  geom_text(aes(label = ifelse(Time == "47.42", Time_Diff, "")), hjust = 2.35, vjust = 2.5, size = 4, co
  labs(x = "Year", y = "Time (seconds)", #add text for the titles and captions and axis
       title = "W 100 Fly SCY Record Since 2016",
       subtitle = "Gretchen Walsh's Record-Breaking Swim (3/22/2024)",
       caption = "Data source: USA Swimming") +
  theme_minimal() + #chang theme to minimal
```

```
  coord_cartesian(ylim = c(47, 49.7)) + #change range of the y axis so the plot doesn't look as squishe
  scale_y_continuous(breaks = seq(47, 49.7, by = 0.5)) + #change the number of ticks in the plot so aud
  theme(panel.grid.major.x = element_blank(), #get rid of x axis grid lines
        panel.grid.minor.x = element_blank(),
        plot.title = element_text(size = 16, face = "bold.italic", hjust = .1), #change the formatting
        plot.subtitle = element_text(size = 14, face = "bold", color = "#F84C1E", hjust = 0.1, family =
        axis.title = element_text(size = 14, face = "bold", family = "serif"),
        axis.title.x = element_blank(),
        axis.text.x = element_text(size = 10, family = "serif"),
        axis.text.y = element_text(size = 10, family = "serif"))
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_segment()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_segment()‘).
```



**Data Viz 3**

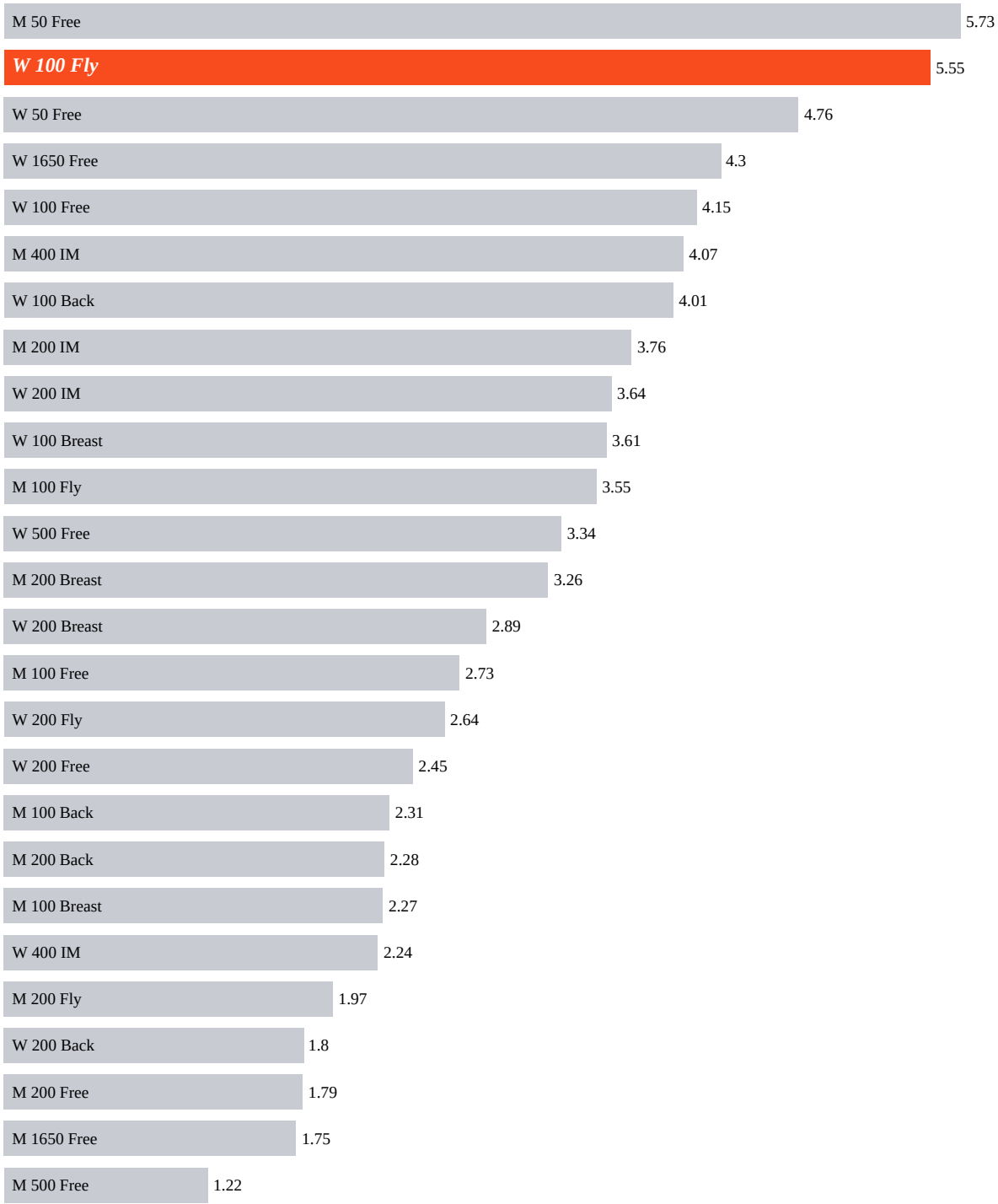
How significant is Gretchen Walsh's record-breaking performance in the broader historical context of SCY achievements? To what extent did Walsh's swim demonstrate dominance in the sport?

```
colors <- c(rep("#C8CBD2", 24), "#F84C1E", rep("#C8CBD2", nrow(gap) - 25))  #make all the bars gray exc

ggplot(gap, aes(x = EVENT, y = GAP, fill = EVENT)) + #plot of event and gap
  labs(title = "Comparison of % Time Difference in SCY Event Gaps", #add text to titles
       subtitle = "Walsh's Record: 2nd Largest SCY % Difference Ever, Largest in Women's History",
       caption = "Data source: SwimSwam") +
  geom_bar(stat = "identity", width = .75) + #bar  chart
  scale_fill_manual(values = colors) +  #set to custom fill colors
  geom_text(aes(label = GAP), #add number labels to end of bar and make adjustments
            hjust = -0.2,
            size = 3,
            color = "black",
            family = "serif") +
  geom_text(data = subset(gap, EVENT == "W 100 Fly"), #add event label for the w 100 fly and make it wh
            aes(label = EVENT, y = .05),
            hjust = 0,
            vjust = .3,
            size = 4,
            color = "white",
            fontface = "bold.italic",  # Make the label bold
            family = "serif") +  # Add text labels
  geom_text(data = subset(gap, EVENT != "W 100 Fly"), #add event labels for events that aren't the w 10
            aes(label = EVENT, y = .05),
            hjust = 0,
            size = 3,
            family = "serif") +
  coord_flip() + #make graph horizontal bar chart so it's easier to read
  theme_void() + guides(fill = FALSE) + #change to void theme and get rid of legend
  theme(plot.title = element_text(size = 16, face = "bold.italic", hjust = 0.15), #make adjustments to
        plot.subtitle = element_text(size = 14, color = "#F84C1E", hjust = .25, family = "serif"))
```

## *Comparison of % Time Difference in SCY Event Gaps*

Walsh's Record: 2nd Largest SCY % Difference Ever, Largest in Women's History




Data source: SwimSwam