UNIX and “UNIX -like” Operating Systems

AWK Programming Language

MR. SABAUGH 12.01.22

Today you'll learn...

- How the AWK programming language is a very powerful tool for data analytics, String manipulation, and file handling.
- How AWK is similar to languages you may already be familiar with like Java.
- How to create a computer model using AWK.

The AWK Programming Language

- The AWK programming language is a very powerful tool for data analytics, String manipulation, and file handling. It can also be used to write quick one-liner programs, rapid prototyping of algorithms, and even create your very own programming languages.
- AWK has a similar syntax to C-like languages like Java.
- Created at Bell Labs by Alfred Aho, Peter Weinberger, and Brian Kernighan.
- Direct access to the Operating System so it is very fast. Python uses an interpreter. Thus, python is a layer of abstraction removed from the processor making it slower. This makes a huge difference with very large data sets.
- While not a language used to write Enterprise Level software, it is used widely by programmers, scientists, students, and engineers.

then

if(pattern){ action }

true

“If you haven’t read any files yet...”

“...change the Field Separator variable to a comma.”

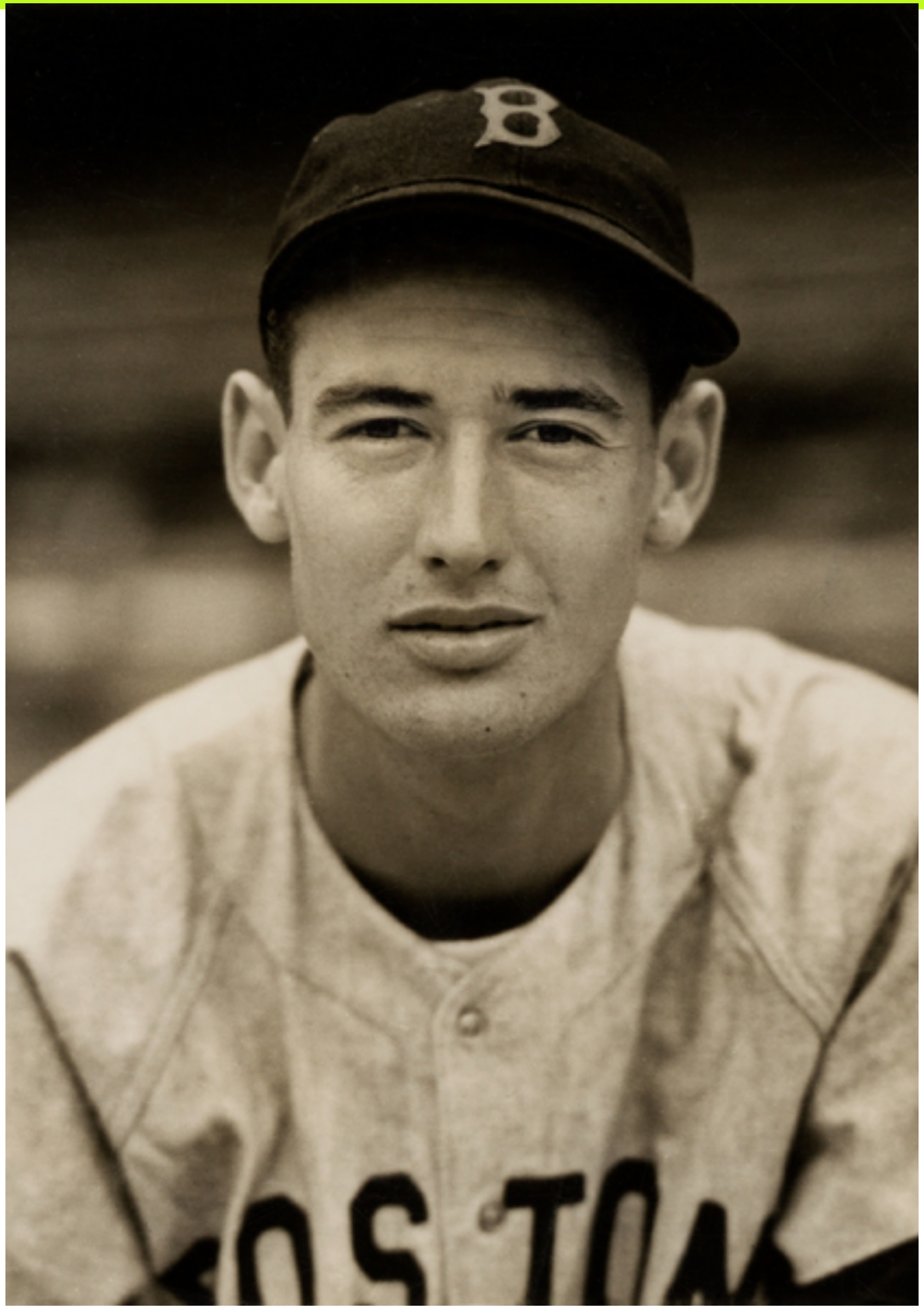
```
BEGIN { FS = “,” }
```

Begin Pattern: before file has been read

Ted Williams

“The Splendid Splinter”

- Hall of Fame Left fielder. Played entire career from 1939-1960 for the Boston Redsox
- Last player to hit .400 (batting average) in a season
- One of only 14 players to win the Batting Triple Crown (League Leader in batting average, home runs, and RBI in a season) and one of only 2 to win it twice.
- All-Time leader in career On Base Percentage .482 (hits + walks + HBP / Plate Appearances)

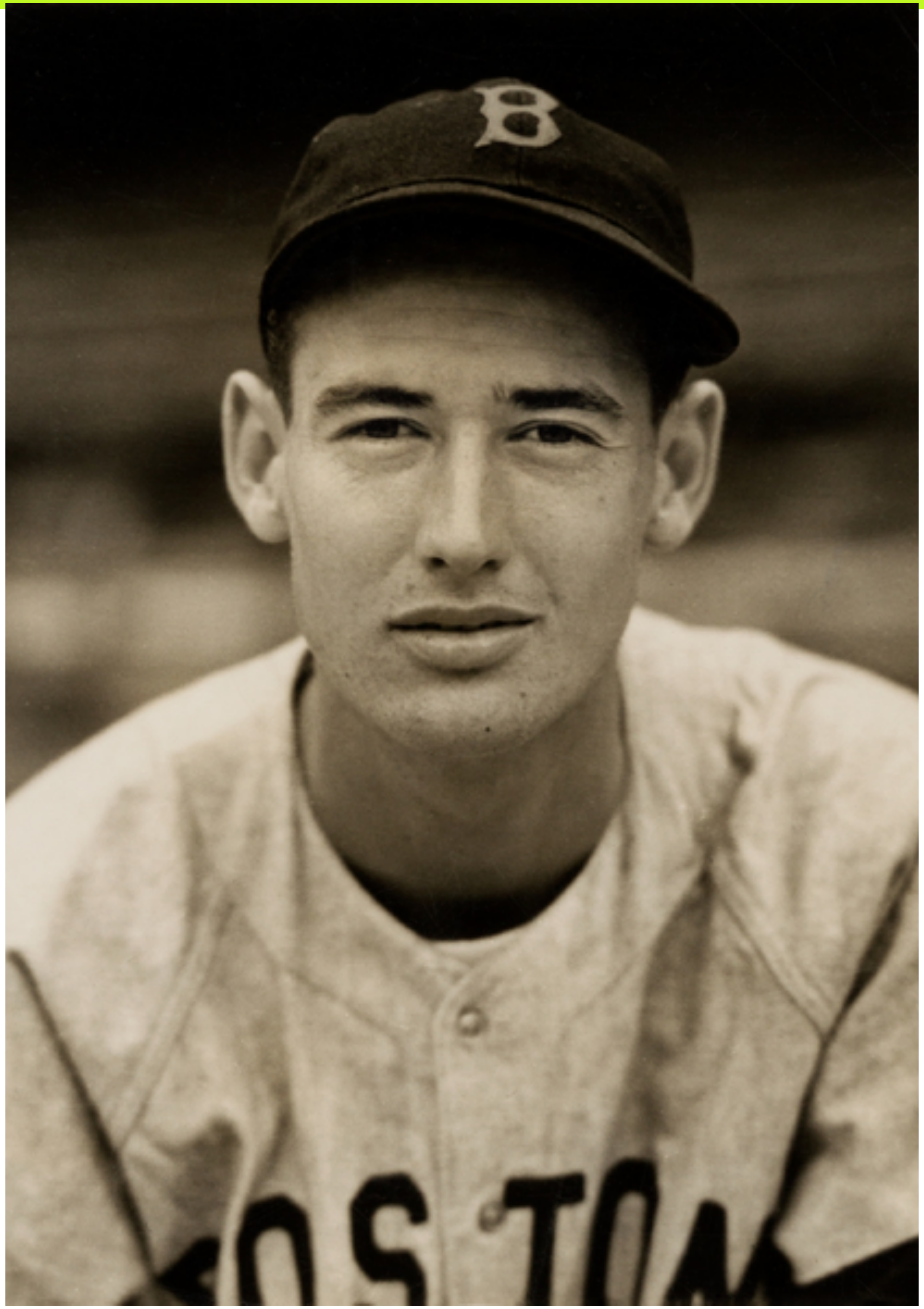


Ted Williams

Batting Title

- A player must have 3.1 plate appearances per team game (155 during Ted's era) for a total of 481 plate appearances per season or
- Player's lead in batting average is sufficiently large that enough hitless at bats can be added to reach this requirement and the player still would have the highest batting average.

Ted won 6 Batting Titles

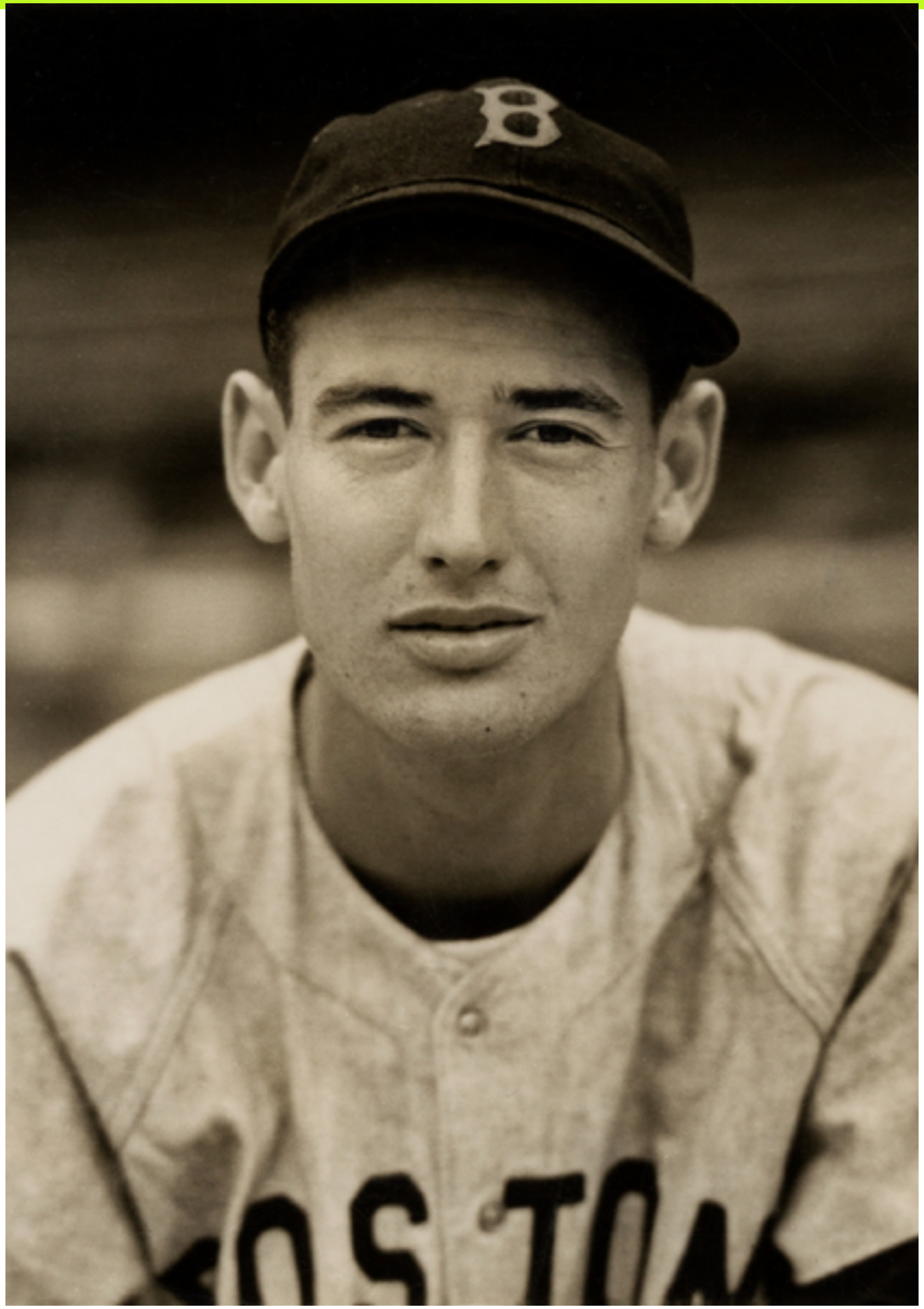


Ted Williams

6 Batting Titles

- A player must have 3.1 plate appearances per team game (155 during Ted’s era) for a total of 481 plate appearances per season or
- Player's lead in batting average is sufficiently large that enough hitless at bats can be added to reach this requirement and the player still would have the highest batting average.

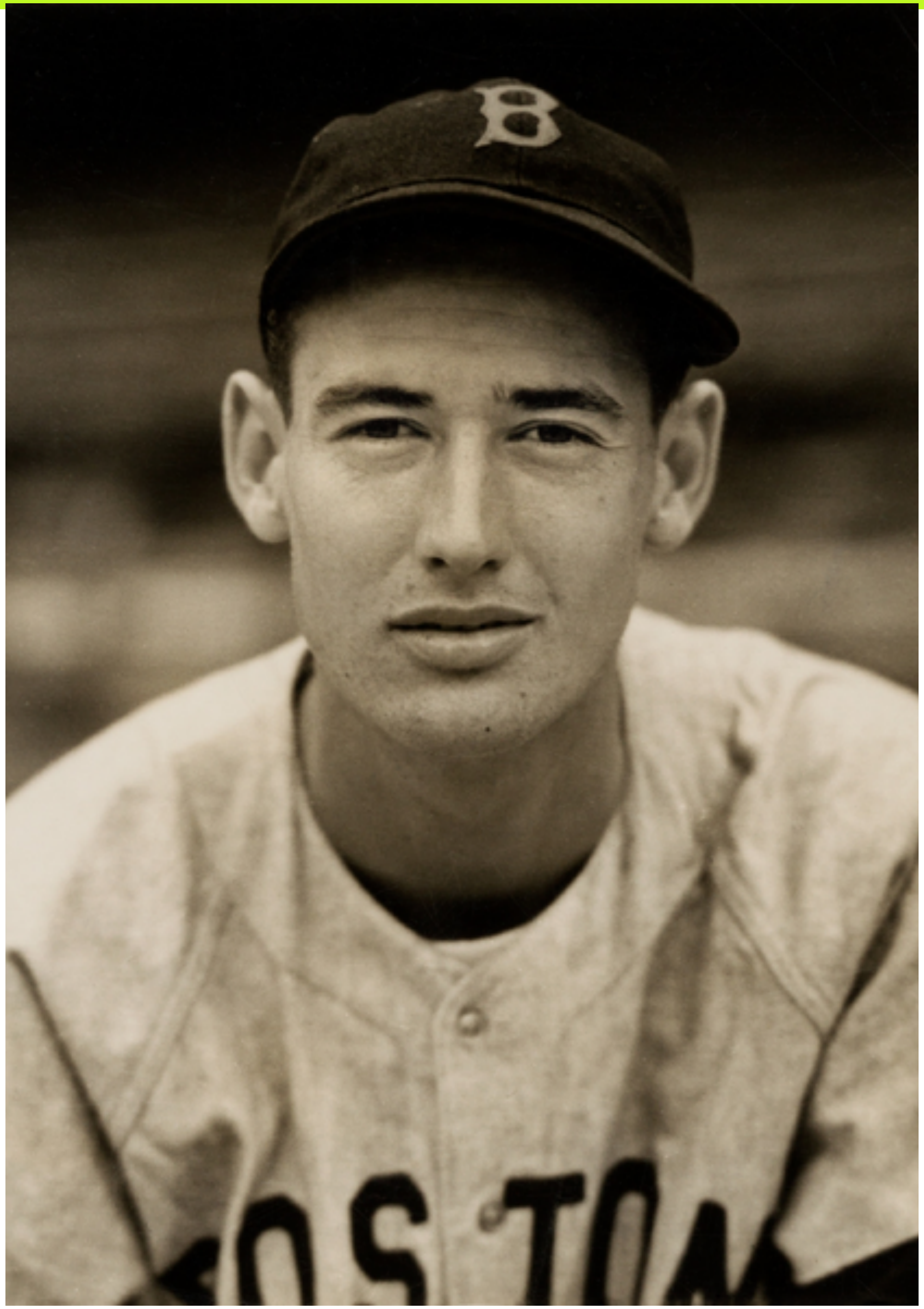
19 All-Star Game Selections



Ted Williams

“The Splendid Splinter”

- Hall of Fame Left fielder. Played entire career from 1939-1960 for the Boston Redsox
- Last player to hit .400 (batting average) in a season
- One of only 14 players to win the Batting Triple Crown (League Leader in batting average, home runs, and RBI in a season) and one of only 2 to win it twice.
- All-Time leader in career On Base Percentage .482 (hits + walks + HBP / Plate Appearances)
- 6 time batting champ & 19 time All-Star



Linear Regression

- Commonly used type of predictive analysis. Used in Machine Learning, Statistics, Stock Market, Scientific Modeling
- Which independent variables are particularly good predictors of outcomes for dependent variables?
- The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable (it is not important for this lesson that you understand what c and b are.).
- Can Ted Williams's age (x or independent variable) be a good predictor of his batting average (y or dependent variable) during the years he was serving in WWII?

Simplified Formula

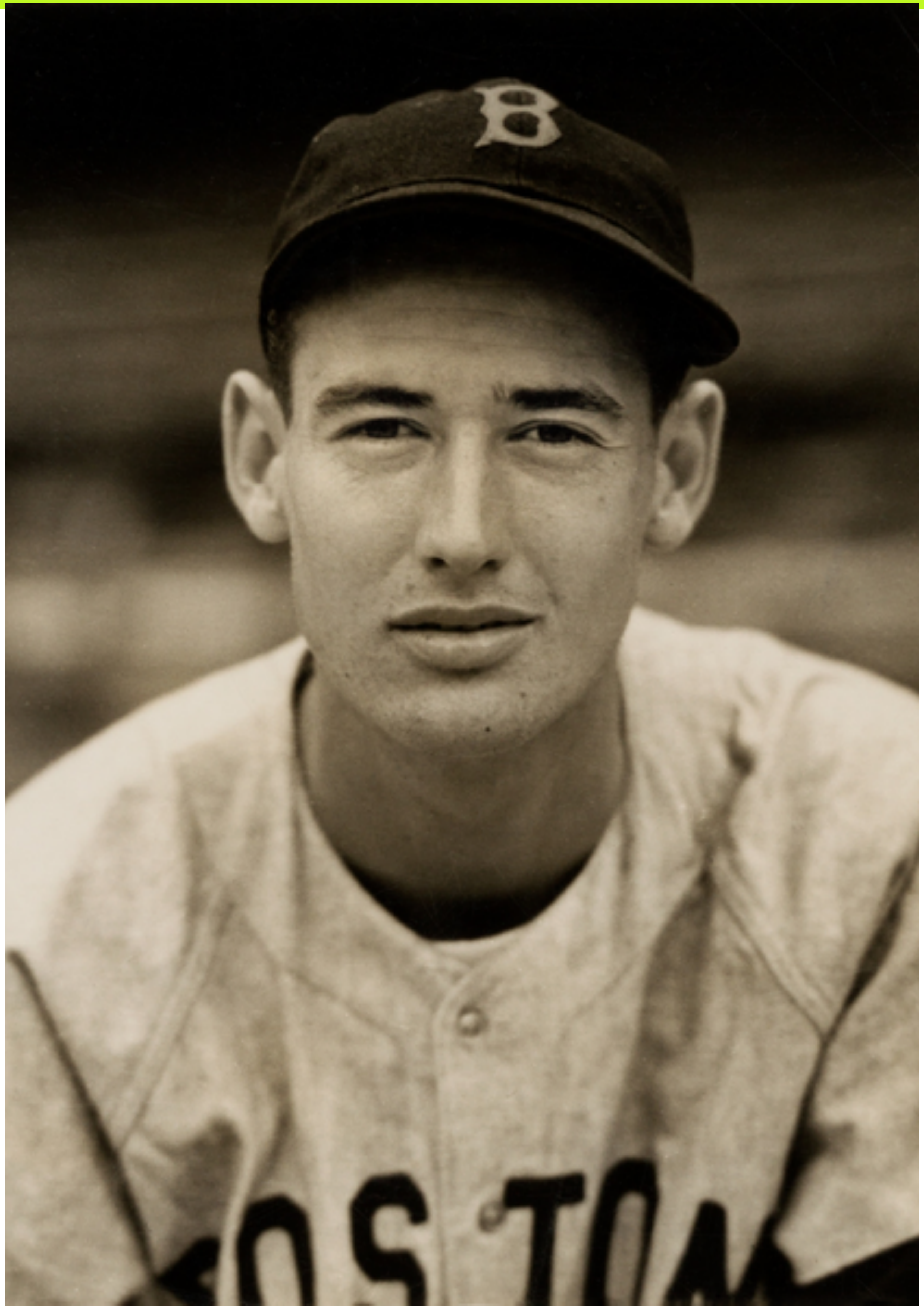
$$Y = a + bX$$

```
awk -f “filename.ext” inputfile.ext
```

Ted Williams

"The Splendid Splinter"

- 1943 age 24 season projected batting average: **.411**
- 1944 age 25 season projected batting average: **.429**
- 1945 age 26 season projected batting average: **.446**



The Lost War Years