A Gentle Reminder to Question the Underlying Data

MBAX 6330: Market Intelligence

Brad Weiner | Chief Data Officer

University of Colorado Boulder

2022-10-25

About Me

Chief Data Officer, University of Colorado Boulder

20 years experience in higher education

15 years on campus (Kansas, Vanderbilt, Minnesota, Colorado)

5 years in Ed-Tech/Consultancy

13 years Higher Ed Analytics/Data Science

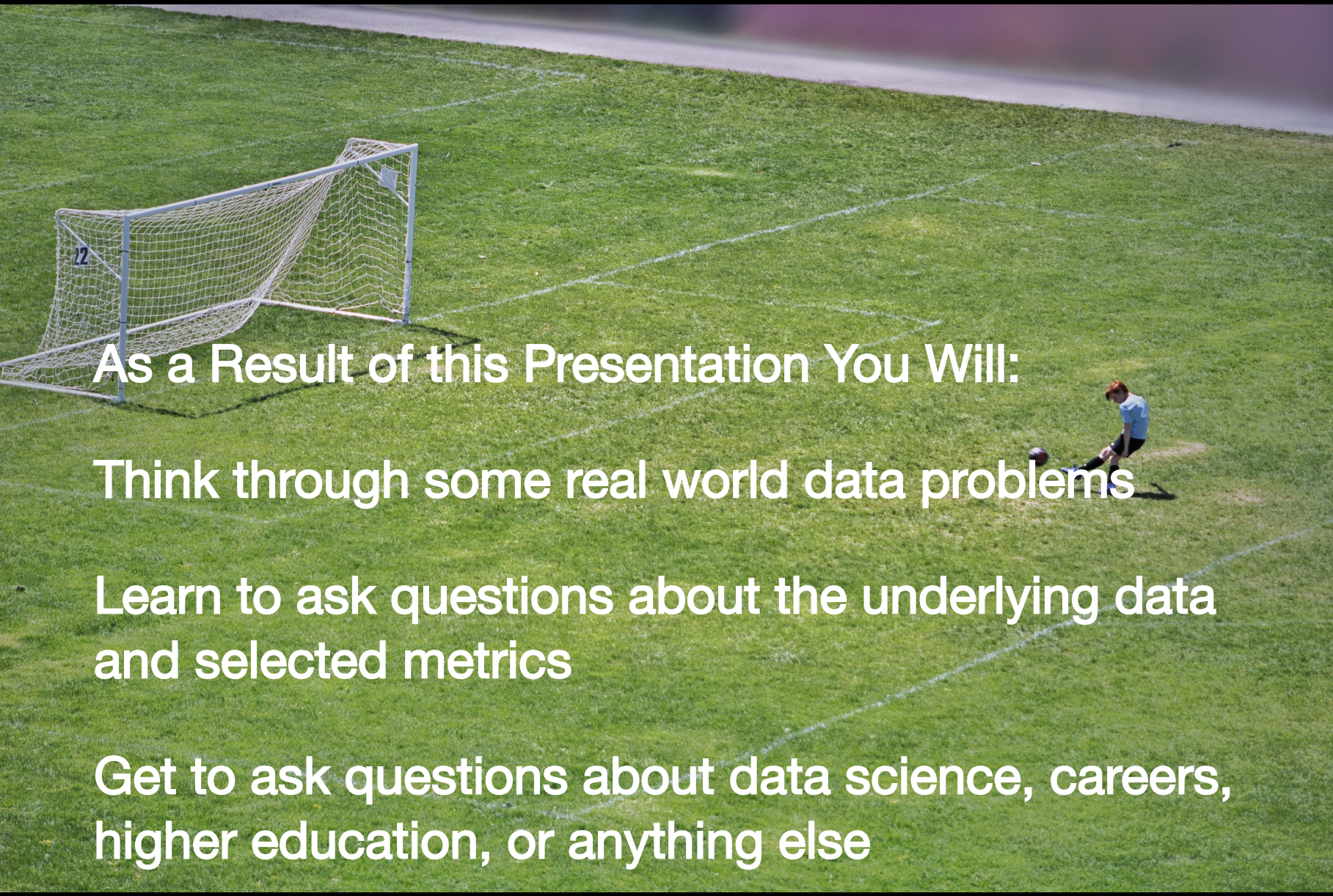
English/Creative Writing Major and Imposter

Contact

brad.weiner@colorado.edu

brad_weiner

bradweiner.info

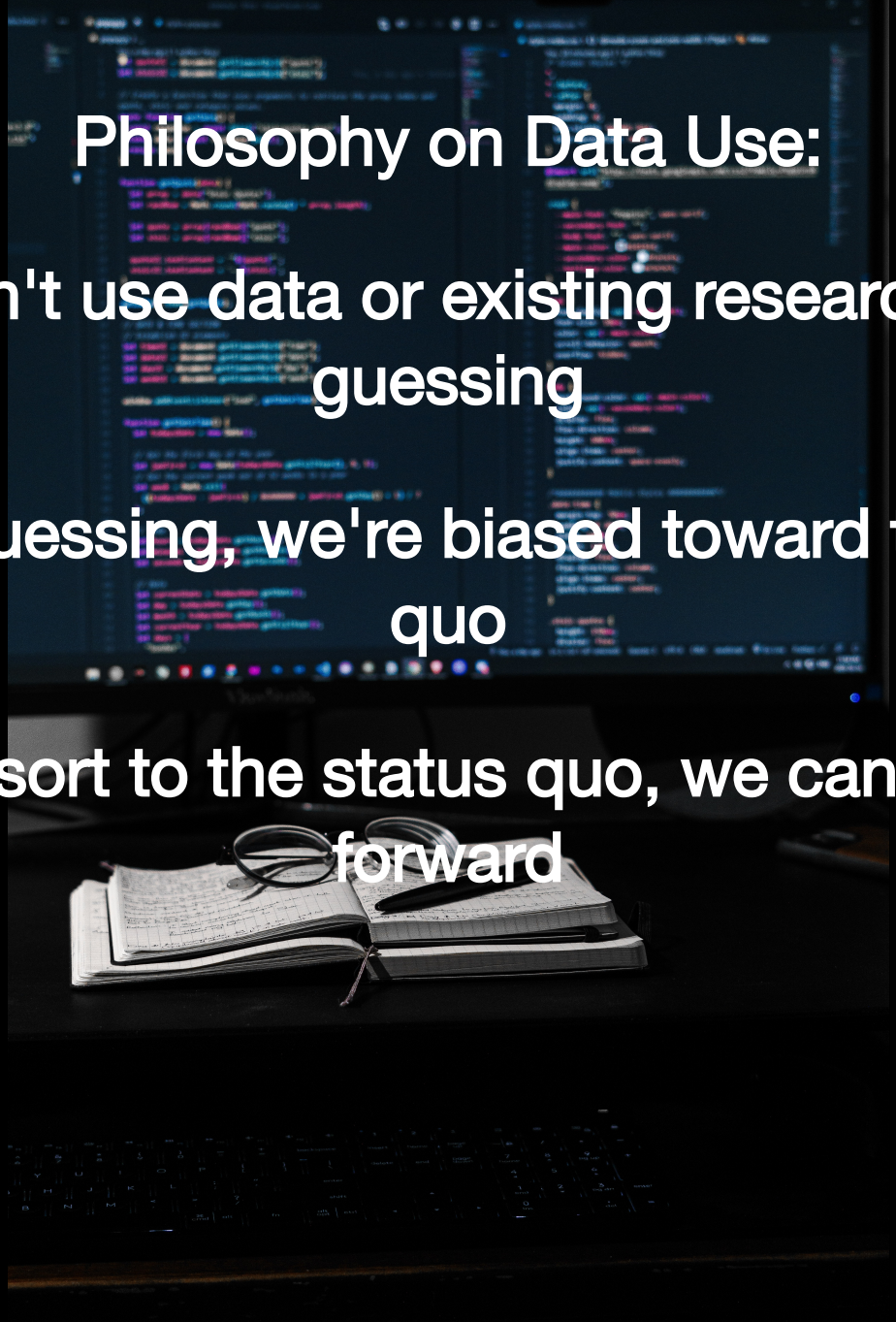


As a Result of this Presentation You Will:

Think through some real world data problems

Learn to ask questions about the underlying data and selected metrics

Get to ask questions about data science, careers, higher education, or anything else



Philosophy on Data Use:

If we don't use data or existing research, we're guessing

If we're guessing, we're biased toward the status quo

If we resort to the status quo, we can't move forward

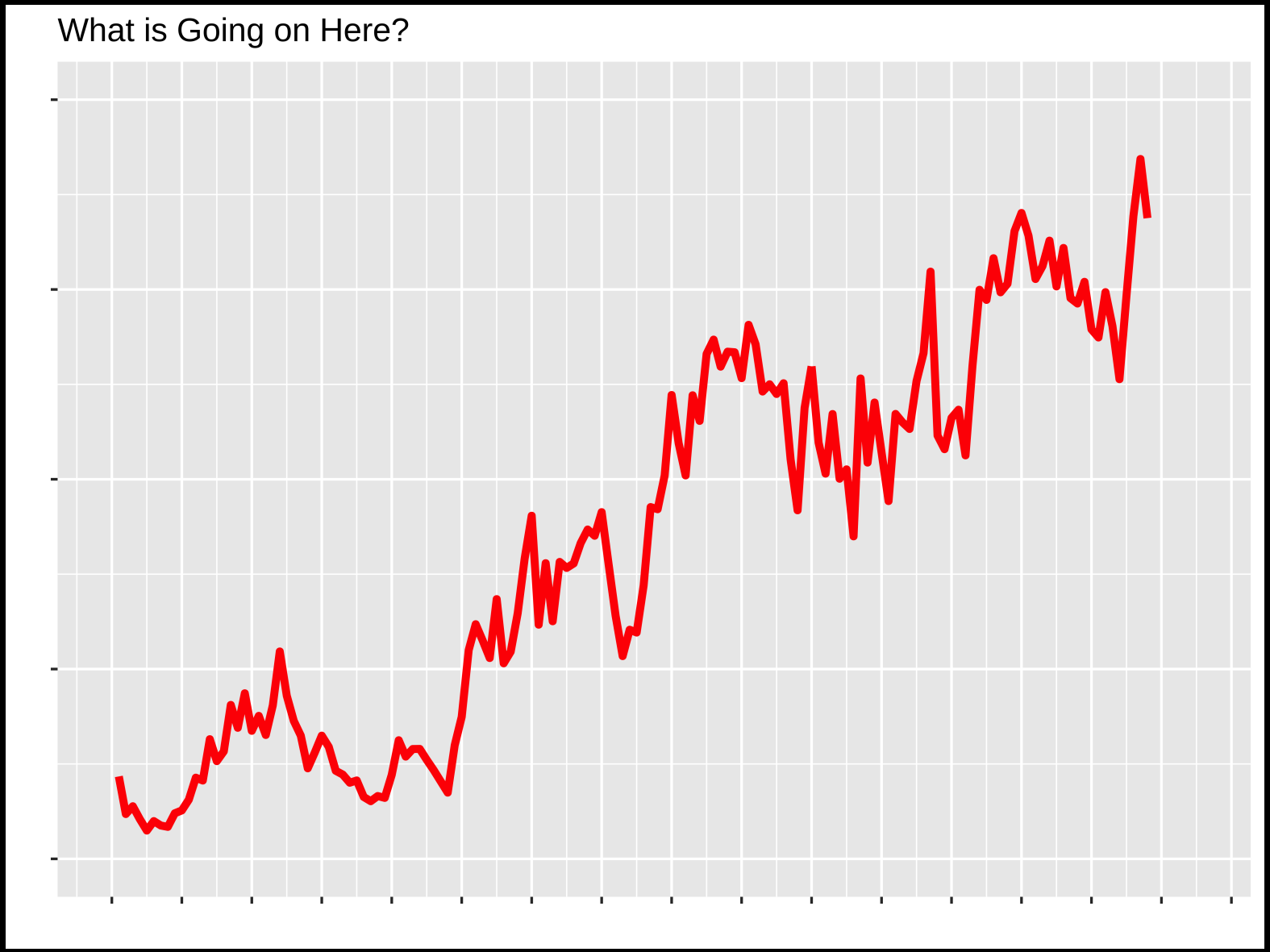
The Office of Data Analytics : Who We Are

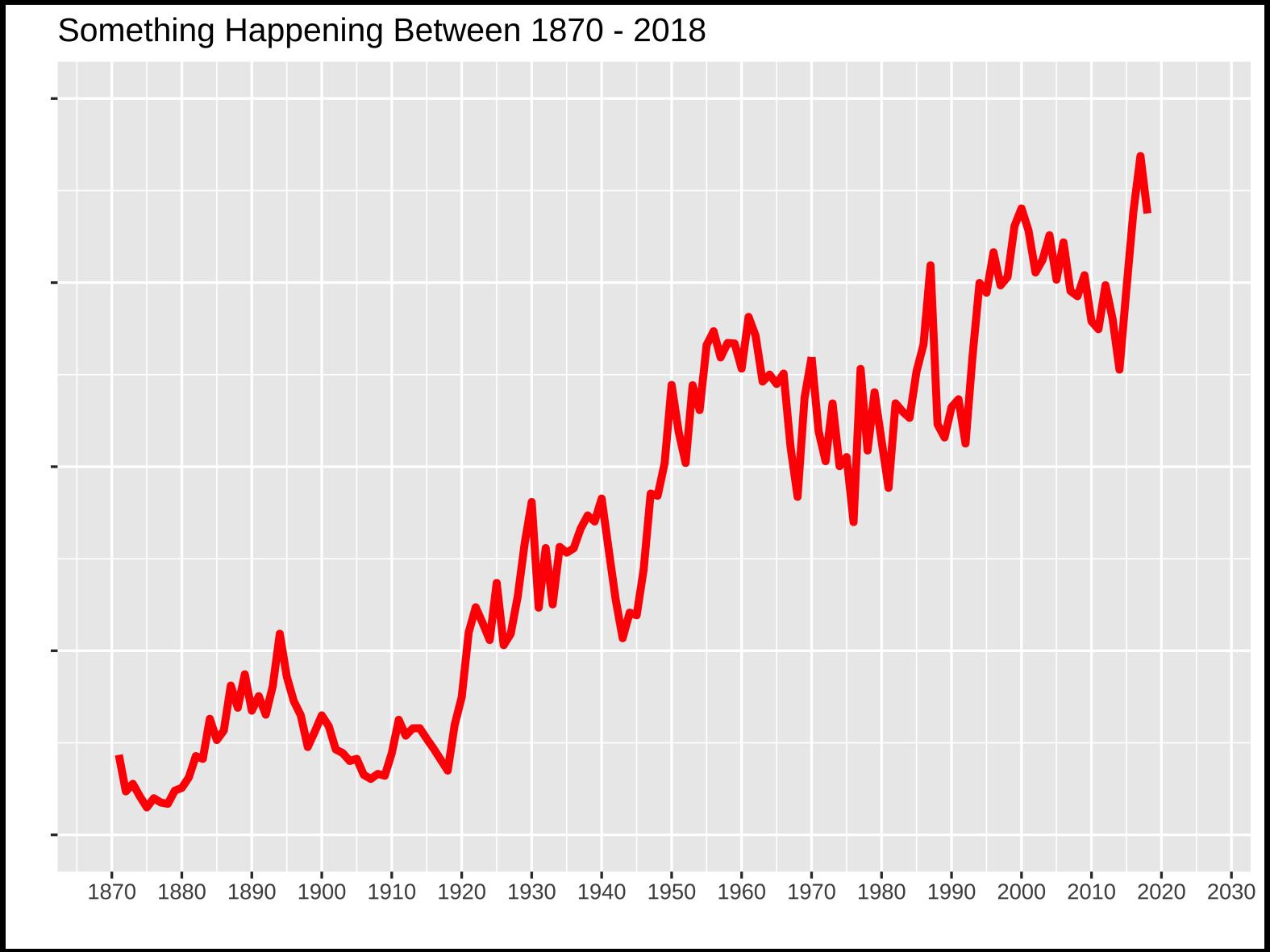
 ODA is a centralized analytics team that exists to provide data, data tools, software, and decision support to stakeholders at CU Boulder and beyond

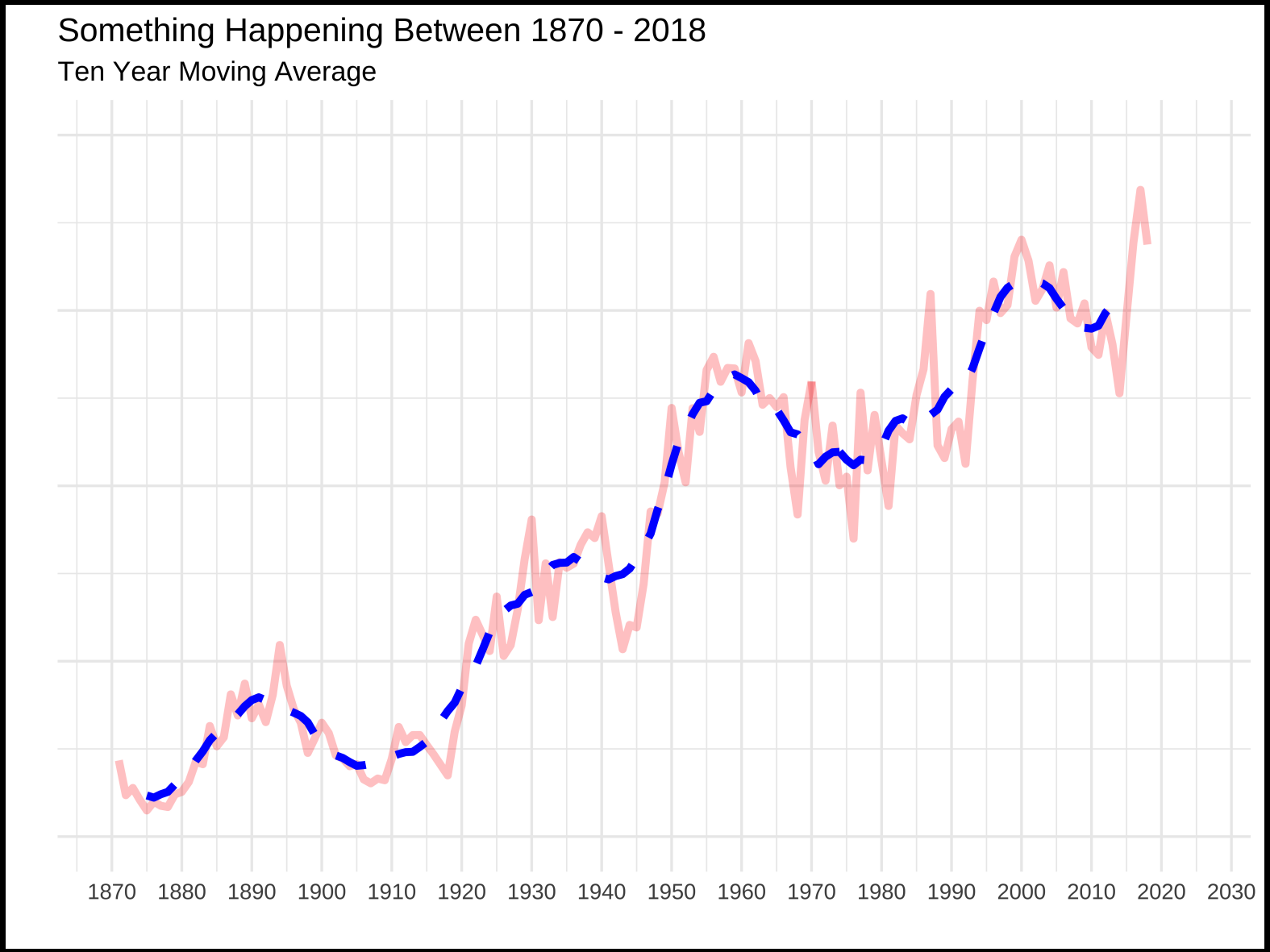
 Our goal is to inform campus decision-making with data and to improve outcomes for students, faculty, and staff

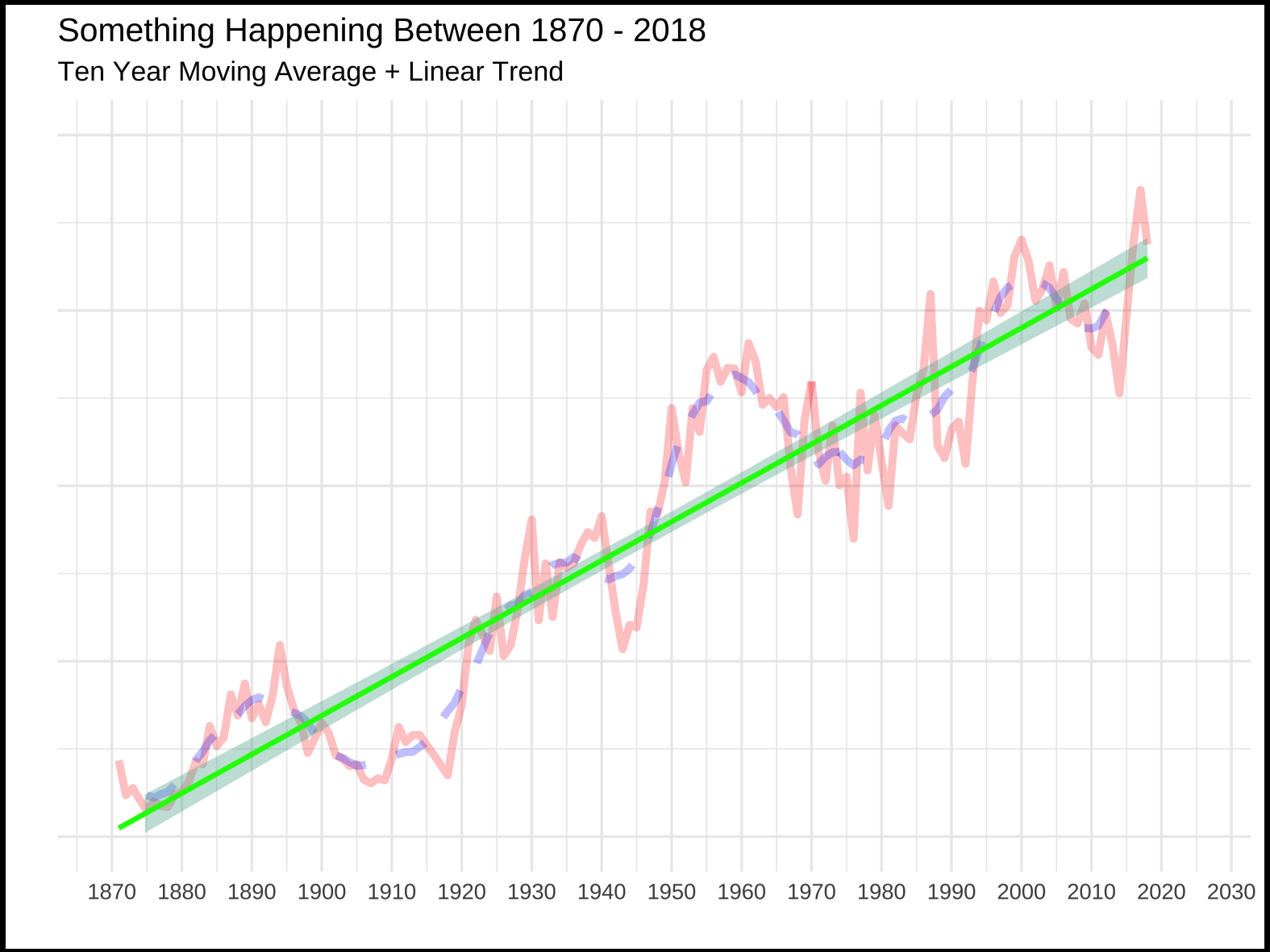
Let's Investigate Some Data

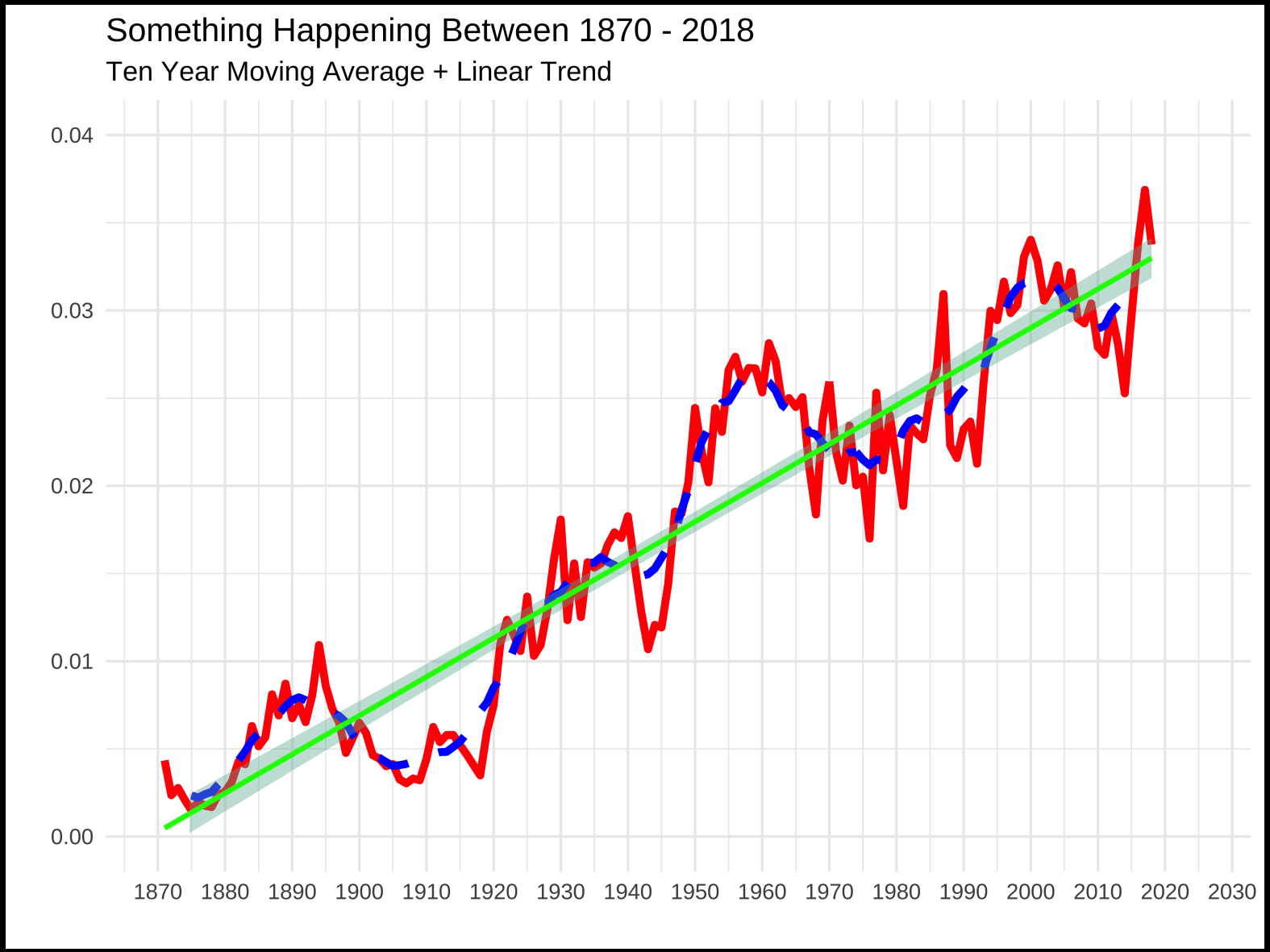
What is Going on Here?

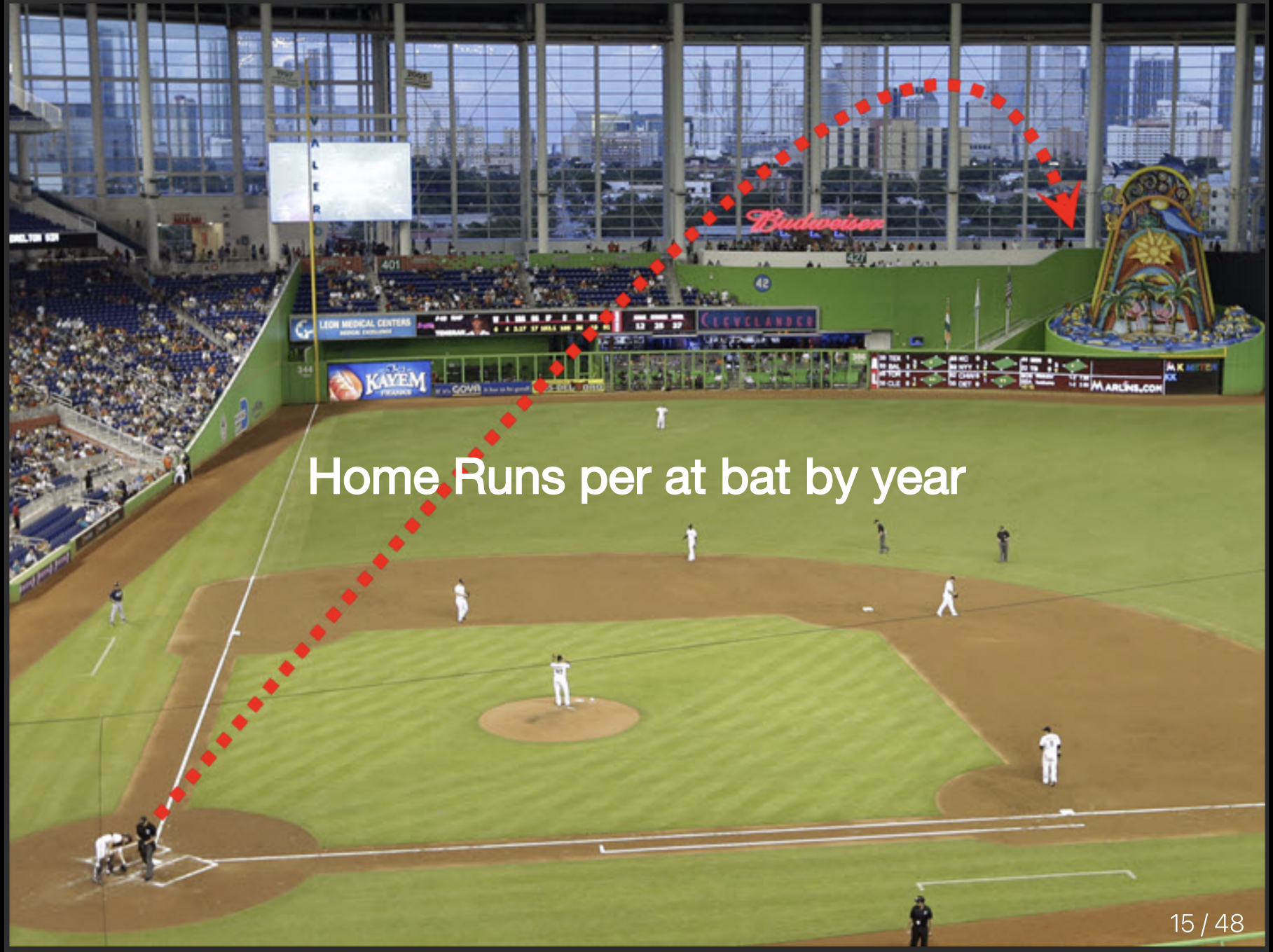


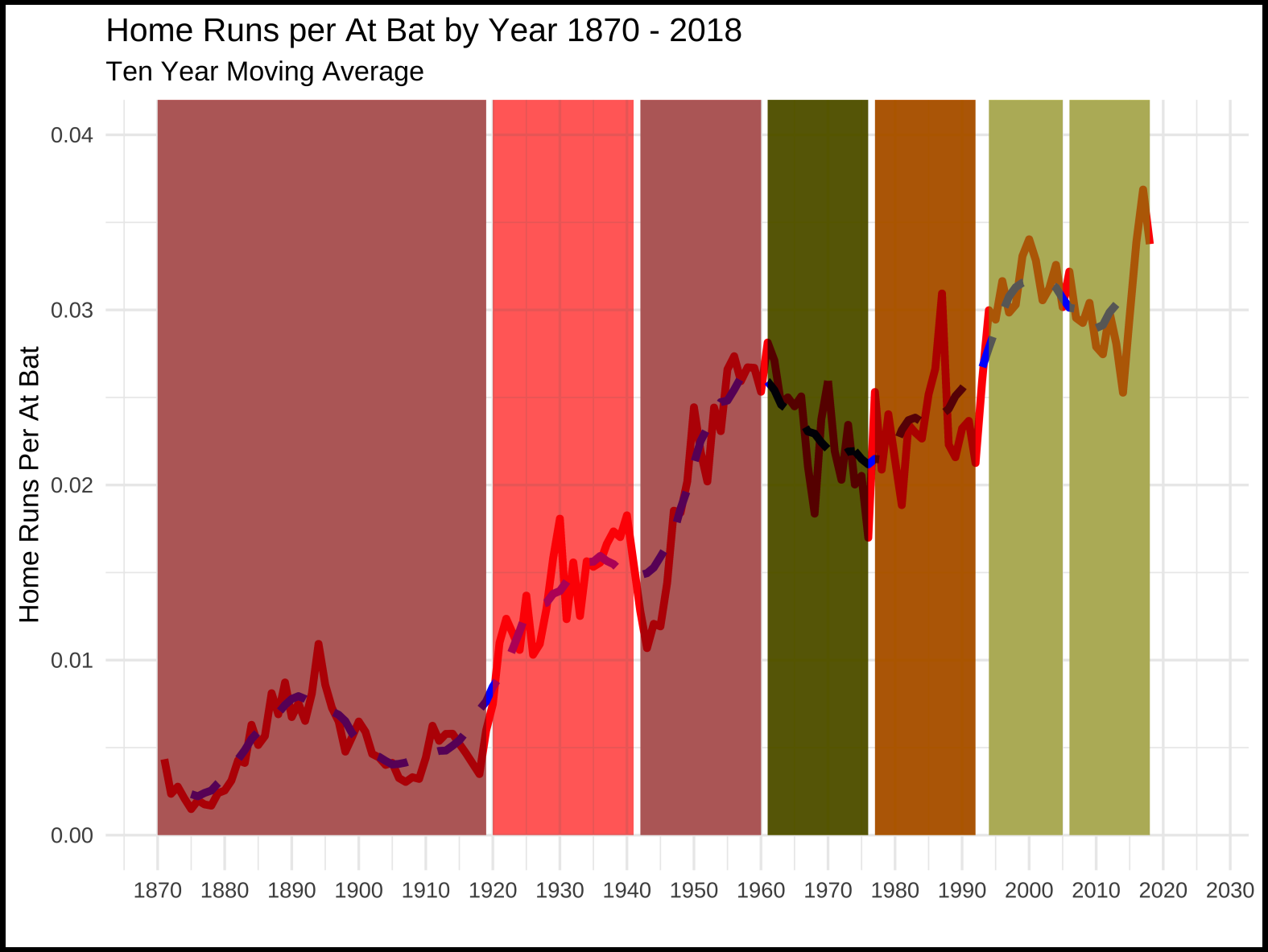

Something Happening Between 1870 - 2018
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
2030

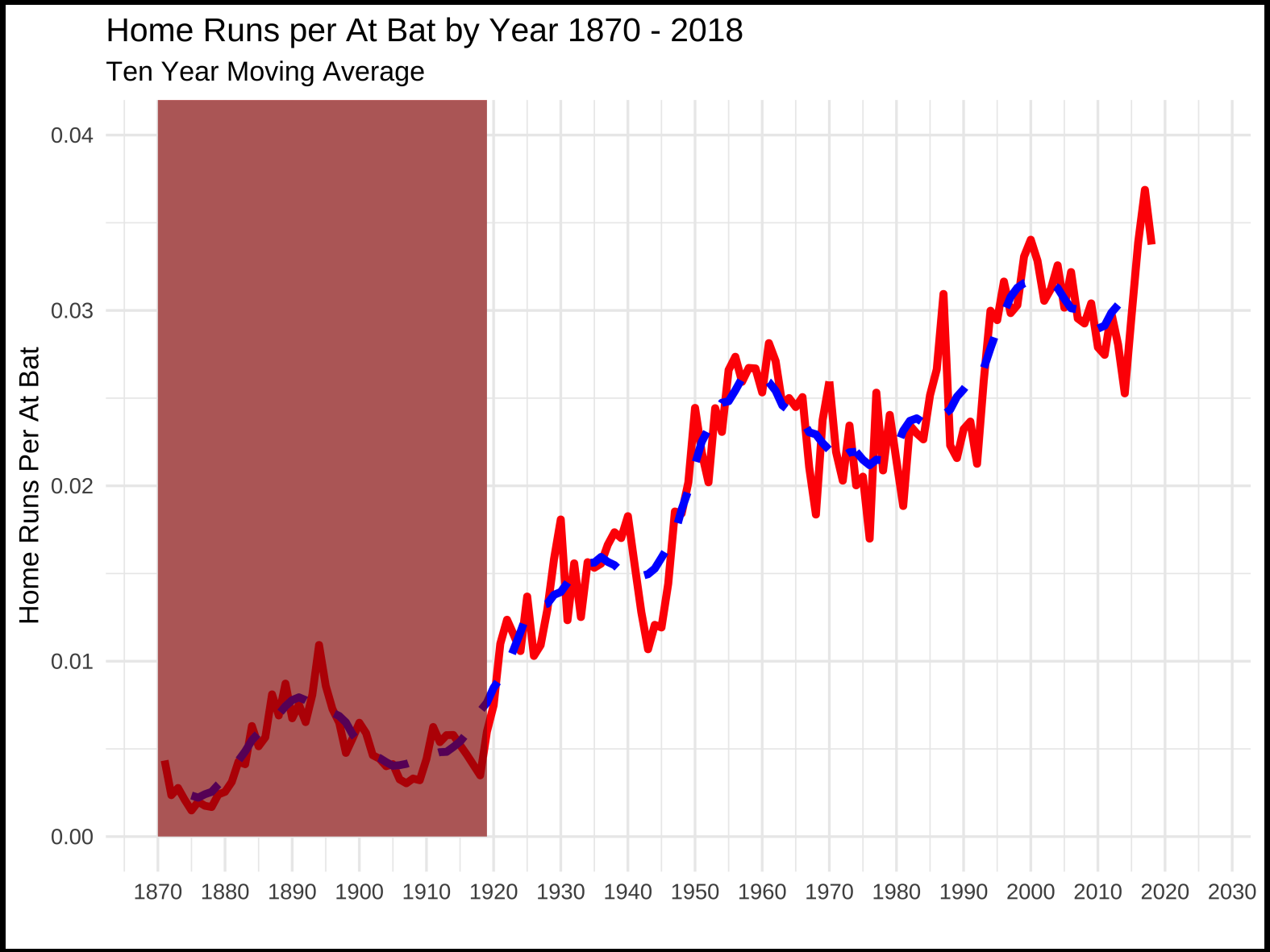

Something Happening Between 1870 - 2018
Ten Year Moving Average
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
2030

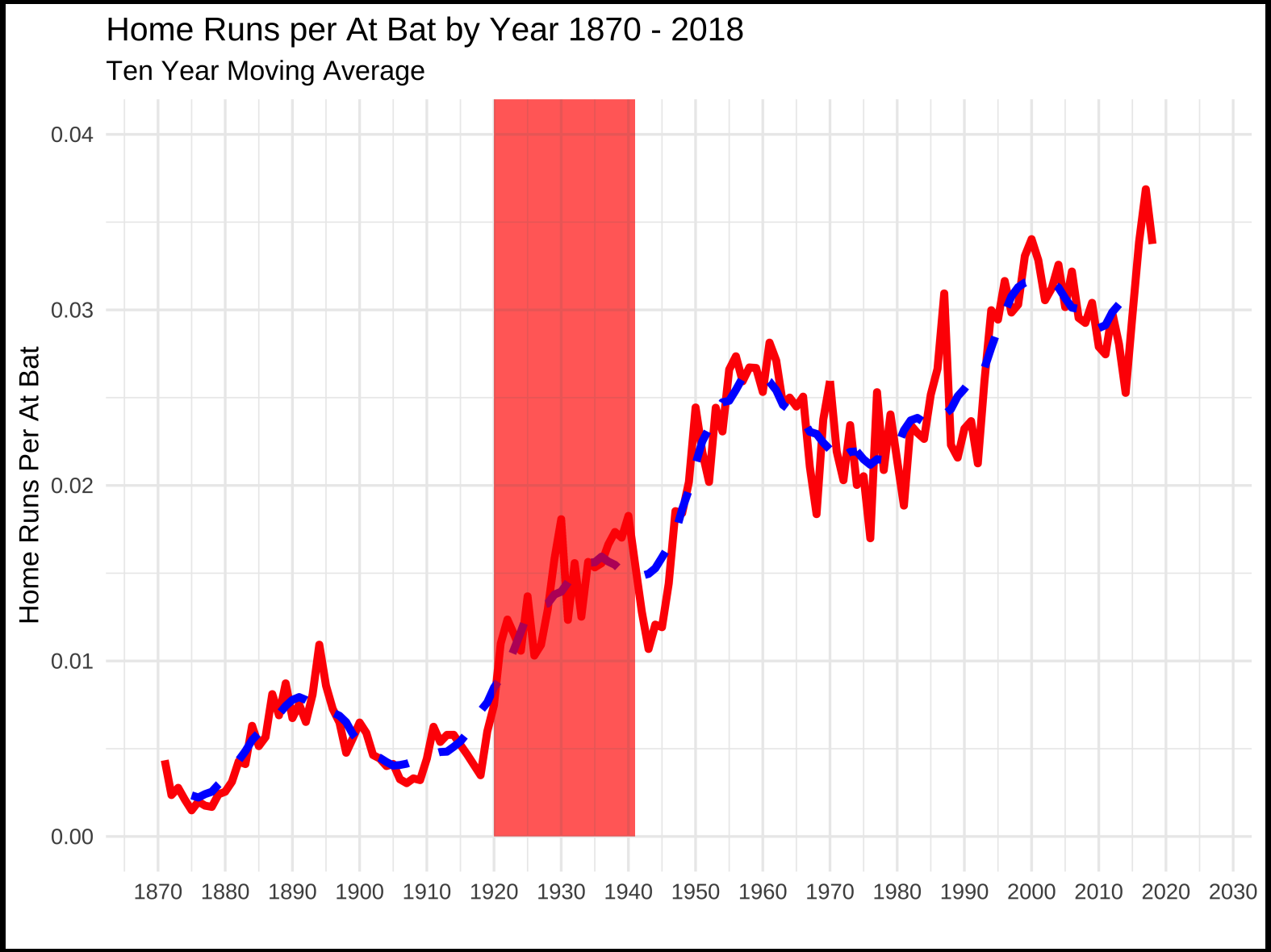

Something Happening Between 1870 - 2018
Ten Year Moving Average + Linear Trend
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
2030

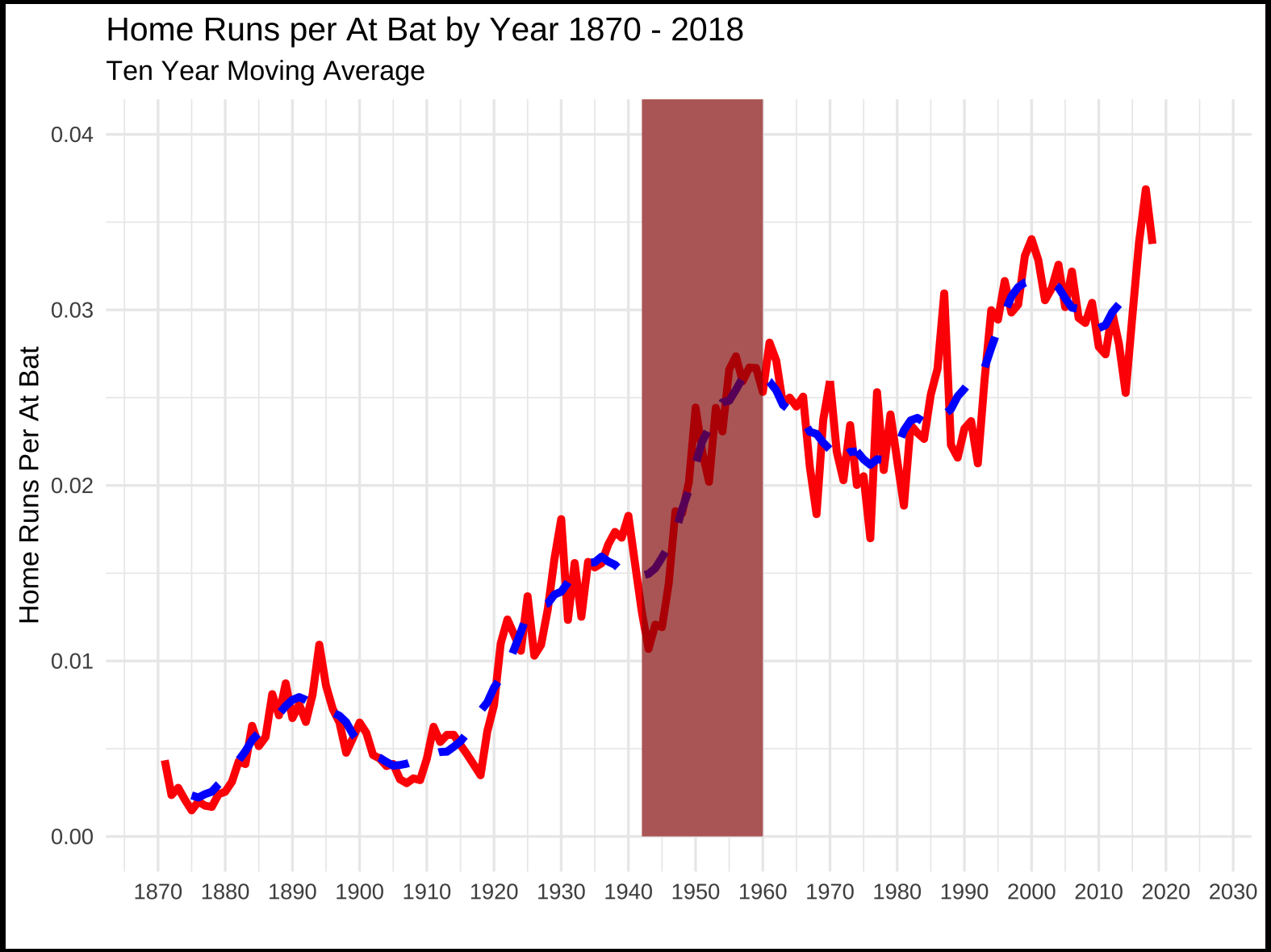

Something Happening Between 1870 - 2018
Ten Year Moving Average + Linear Trend
0.04
0.03
0.02
0.01
0.00
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
2030

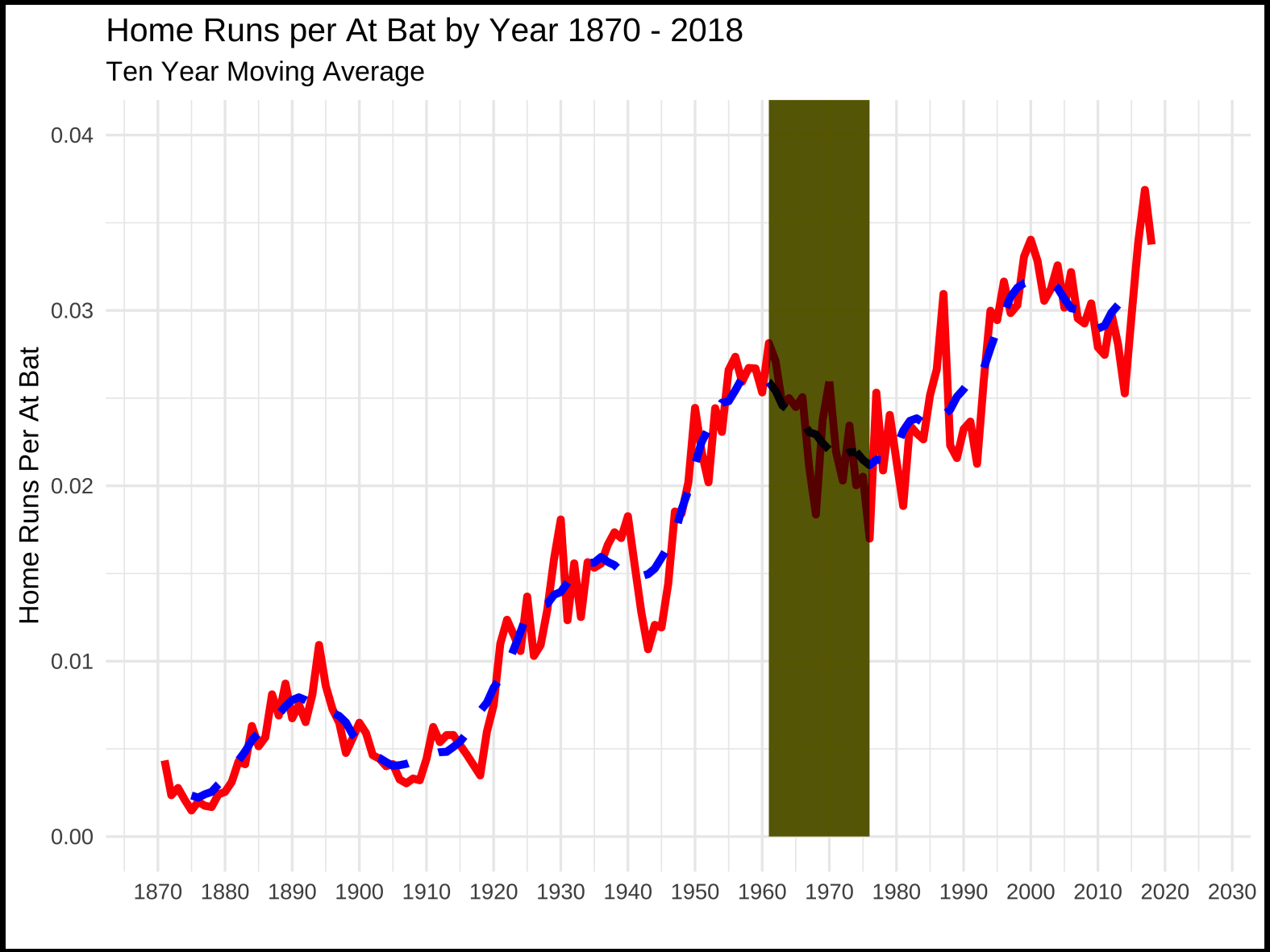

Home Runs per at bat by year

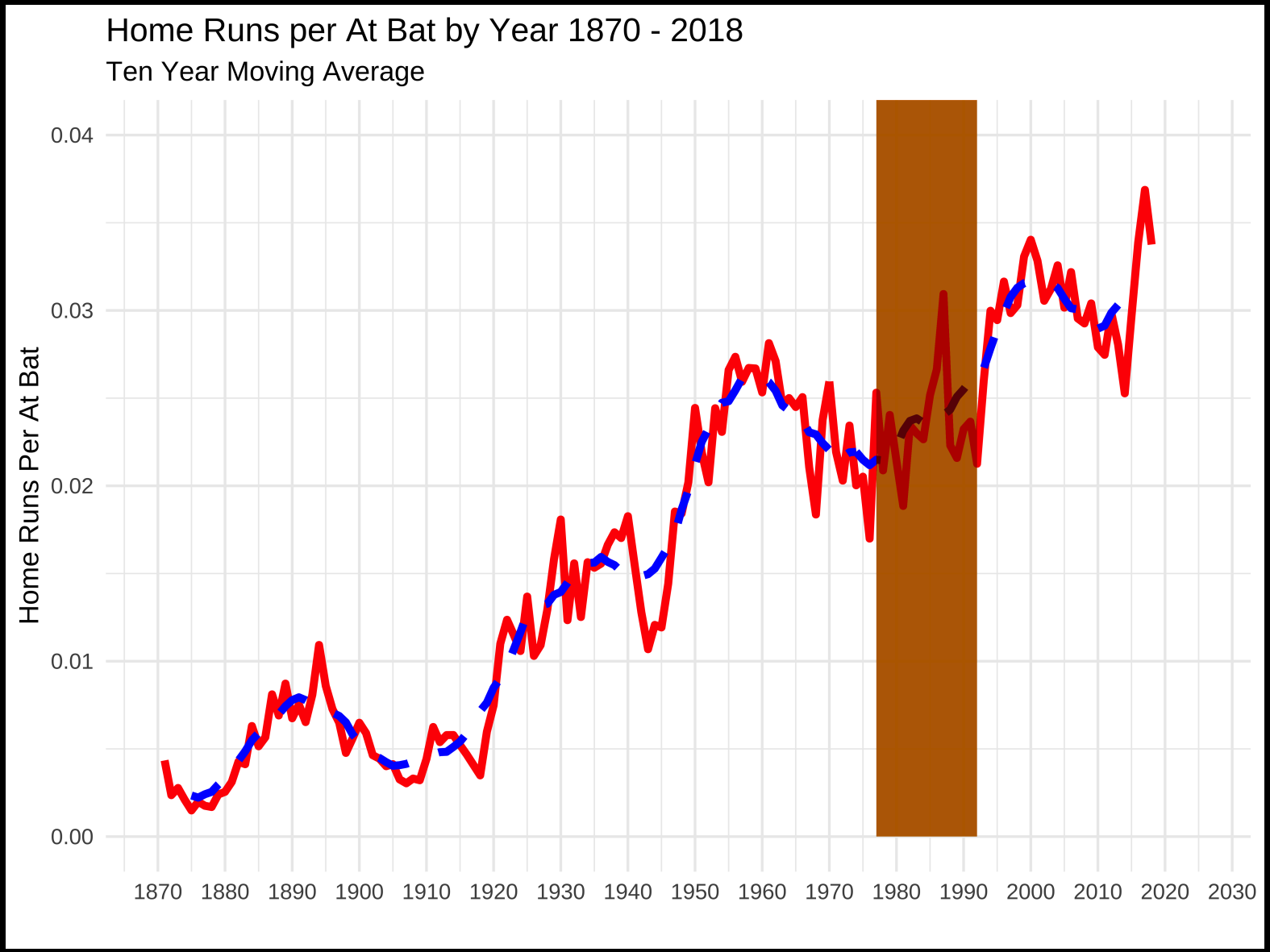

Home Runs per At Bat by Year 1870 - 2018
Ten Year Moving Average
Home Runs Per At Bat
0.04
0.03
0.02
0.01
0.00
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
2030

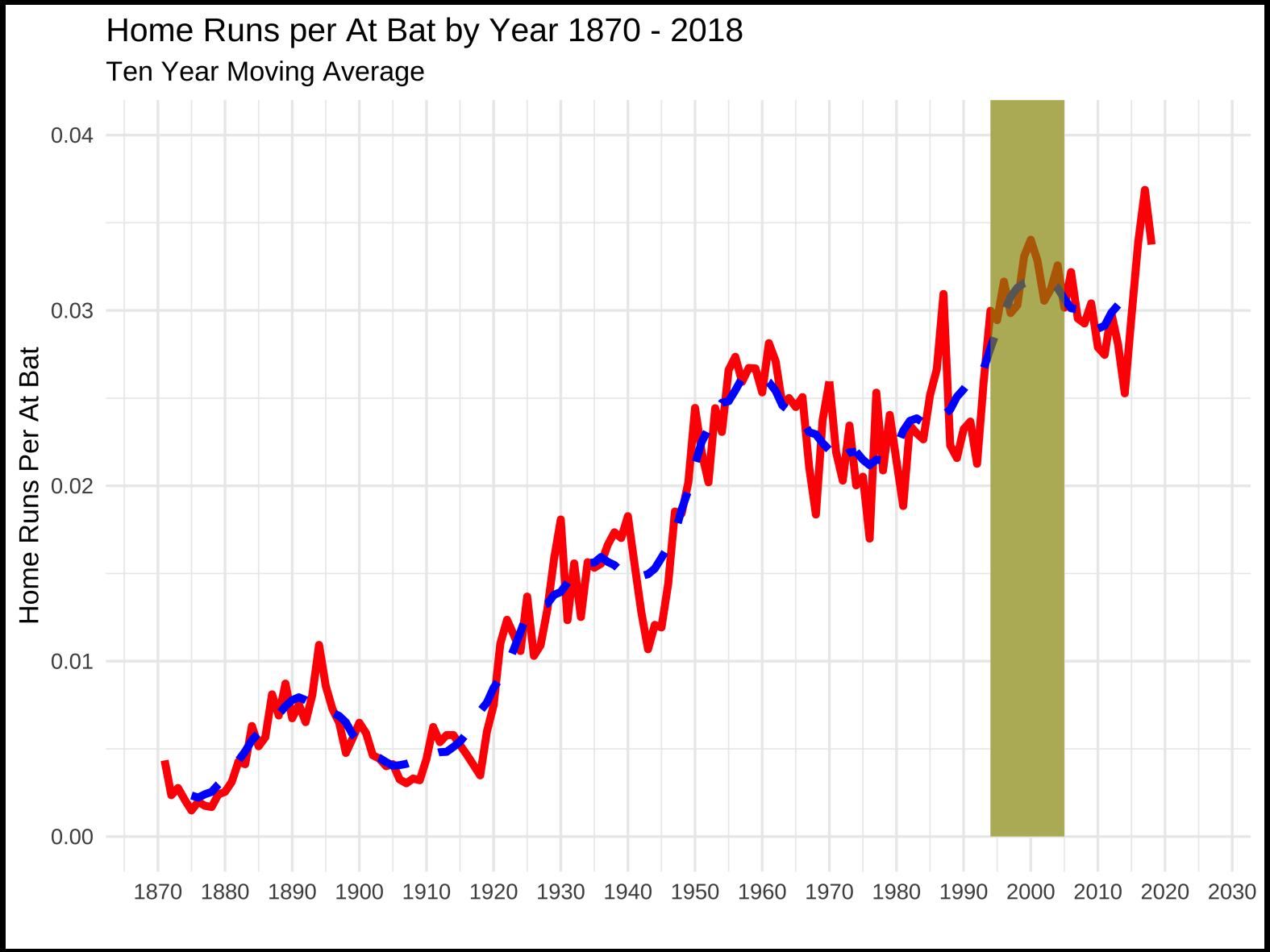

Home Runs per At Bat by Year 1870 - 2018
Ten Year Moving Average
Home Runs Per At Bat
0.04
0.03
0.02
0.01
0.00
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
2030

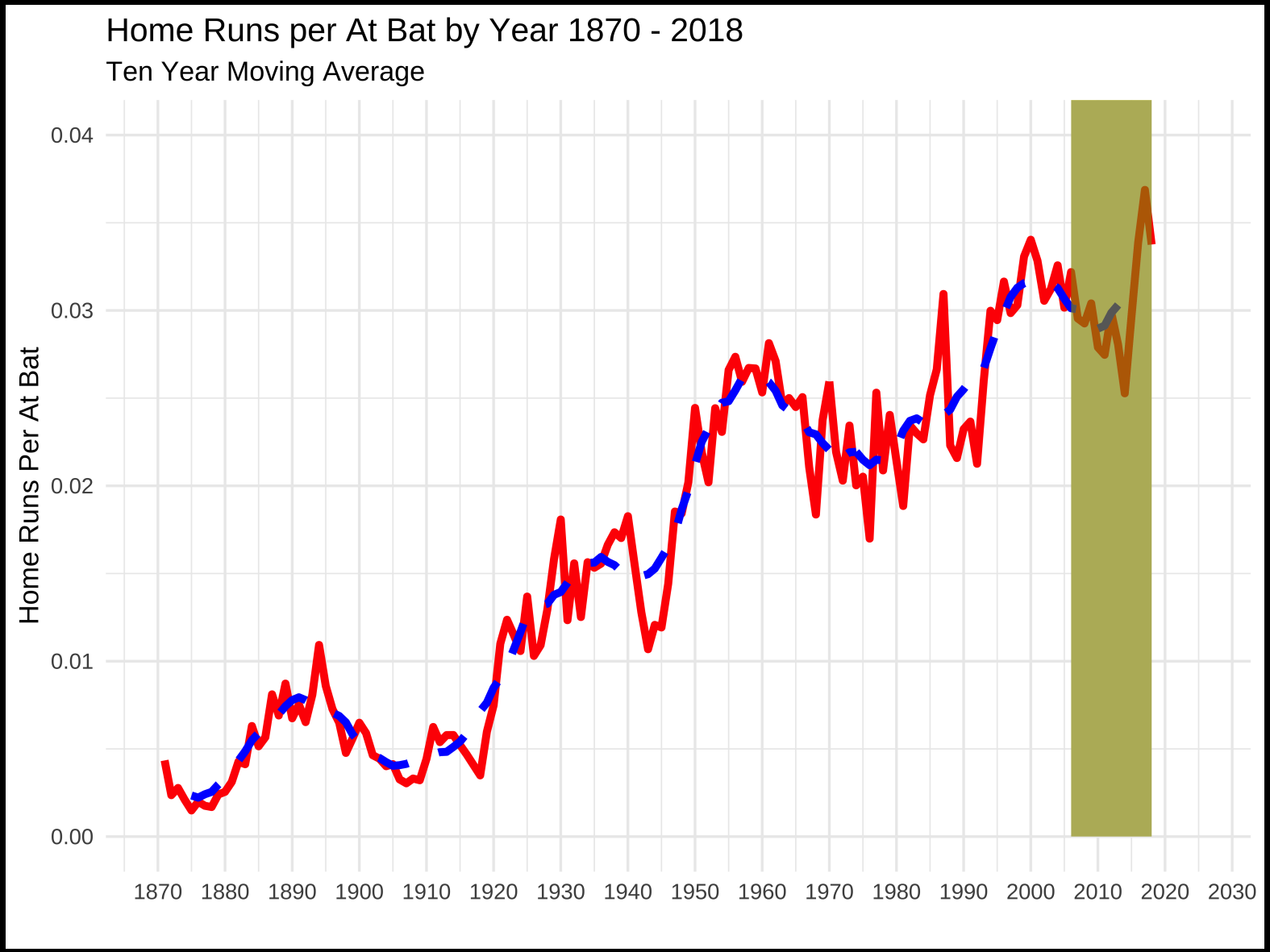

Home Runs per At Bat by Year 1870 - 2018
Ten Year Moving Average
Home Runs Per At Bat
0.04
0.03
0.02
0.01
0.00
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
2030


Home Runs per At Bat by Year 1870 - 2018
Ten Year Moving Average
Home Runs Per At Bat
0.04
0.03
0.02
0.01
0.00
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
2030


Home Runs per At Bat by Year 1870 - 2018
Ten Year Moving Average
Home Runs Per At Bat
0.04
0.03
0.02
0.01
0.00
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
2030


Home Runs per At Bat by Year 1870 - 2018
Ten Year Moving Average
Home Runs Per At Bat
0.04
0.03
0.02
0.01
0.00
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
2030


Home Runs per At Bat by Year 1870 - 2018
Ten Year Moving Average
Home Runs Per At Bat
0.04
0.03
0.02
0.01
0.00
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
2030


Home Runs per At Bat by Year 1870 - 2018
Ten Year Moving Average
Home Runs Per At Bat
0.04
0.03
0.02
0.01
0.00
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
2030

⚾ Changes in home run outputs were related to the changes in the game or the environment

⚾ Dead Ball Era: Pitchers dominated with a larger strike zone reused 'dead' baseballs, and the ability to apply substances to the ball.

⚾ Live Ball Era: Clean baseballs and prevention of foreign substances moved the game away from pitchers and toward hitters.

⚾ WWII: Many of the best players went to fight in the war but the game kept going rather than being canceled.

⚾ Expansion and Awful Ballparks: Strike zone was changed again making it easier for pitchers. But then, the mound was lowered making it easier for batters. 1973 introduced the designated hitter.

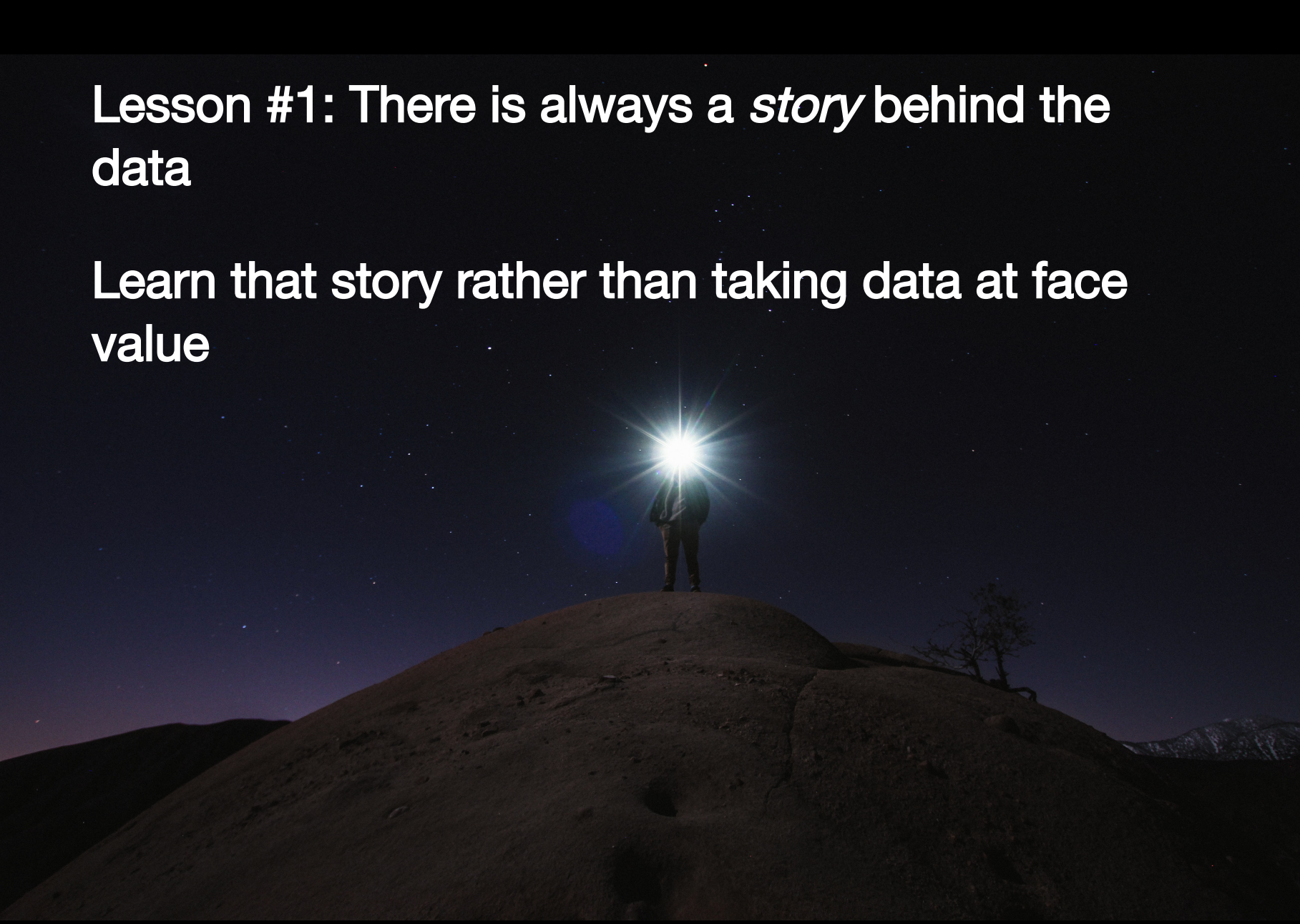
⚾ Free Agency: The financial market shifted making it possible for wealthy teams to have great pitching AND hitting. Also, ballparks got more home run friendly.

⚾ Steroids: Fans loved seeing home runs and the players on the field became better at hitting home runs, due in part to performance enhancing drugs and hitter-friendly ballparks.

⚾ Post Steriods: Players were tested and banned for using performance enhancing drugs.

Lesson #1: There is always a *story* behind the data

Learn that story rather than taking data at face value



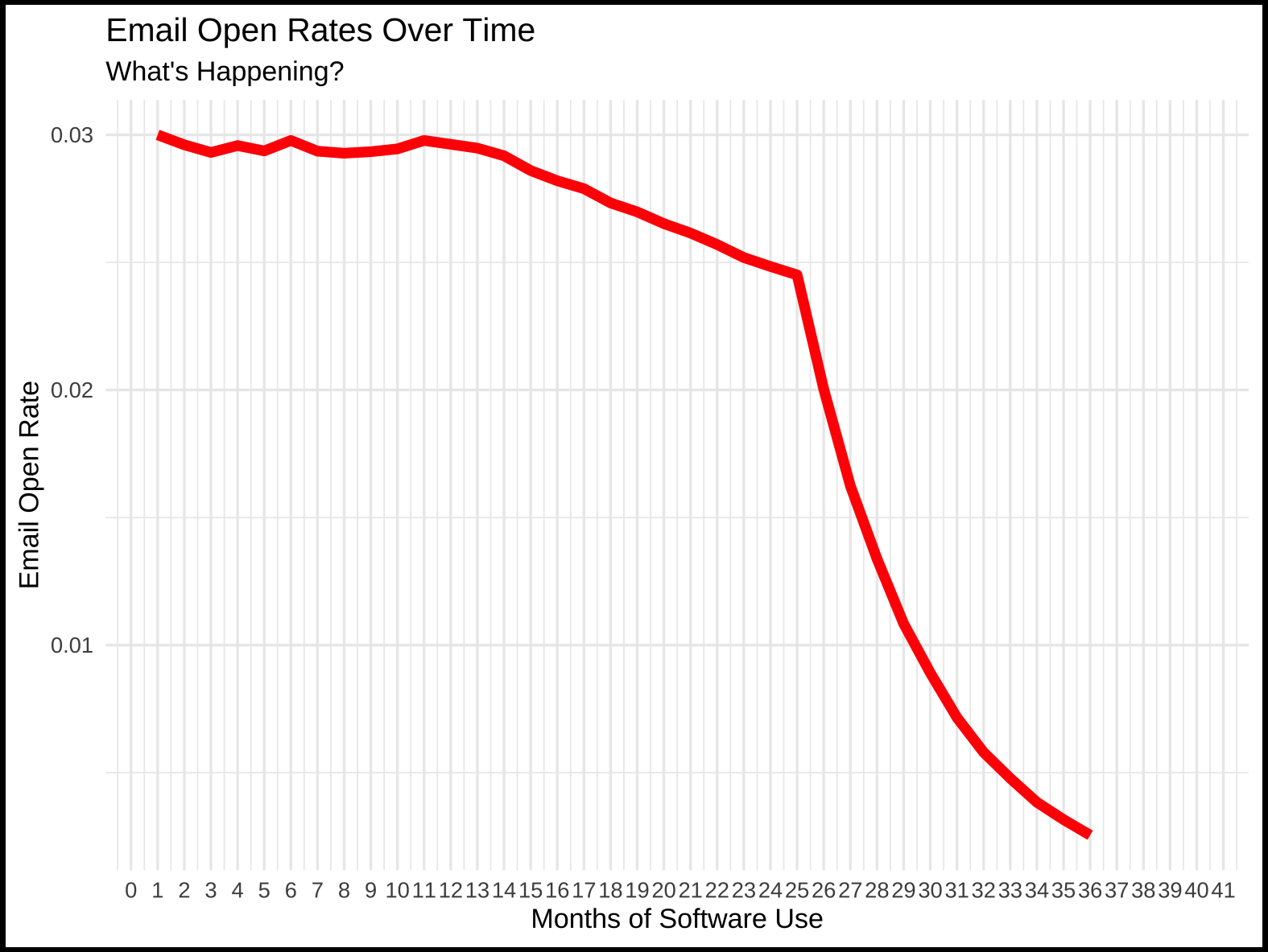
Real World Business Problem

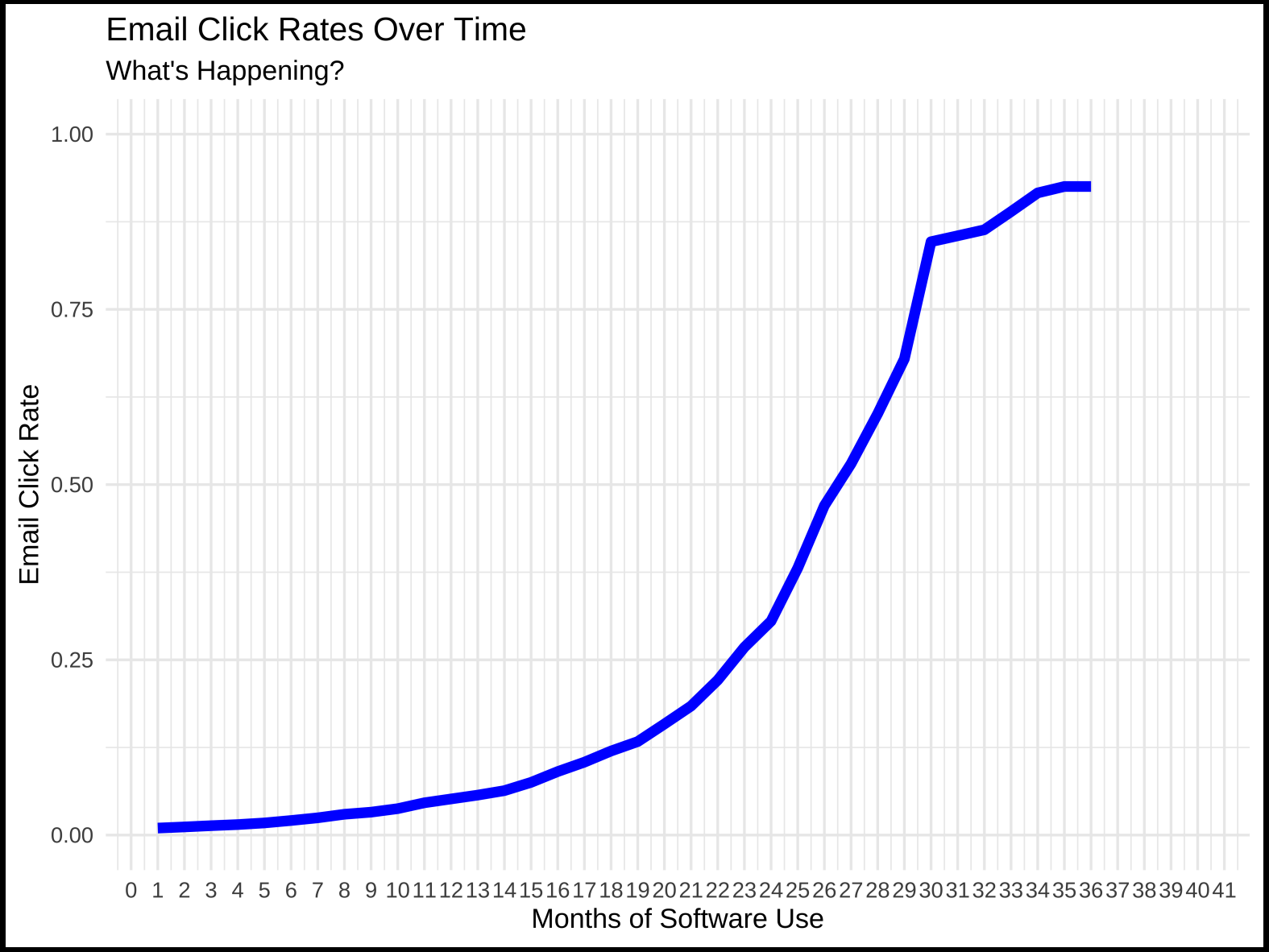
ALL DATA BELOW ARE FAKE!

 You work for a company that utilizes marketing automation software

 This software drops a cookie when people open or click on an email and tracks movement across the website

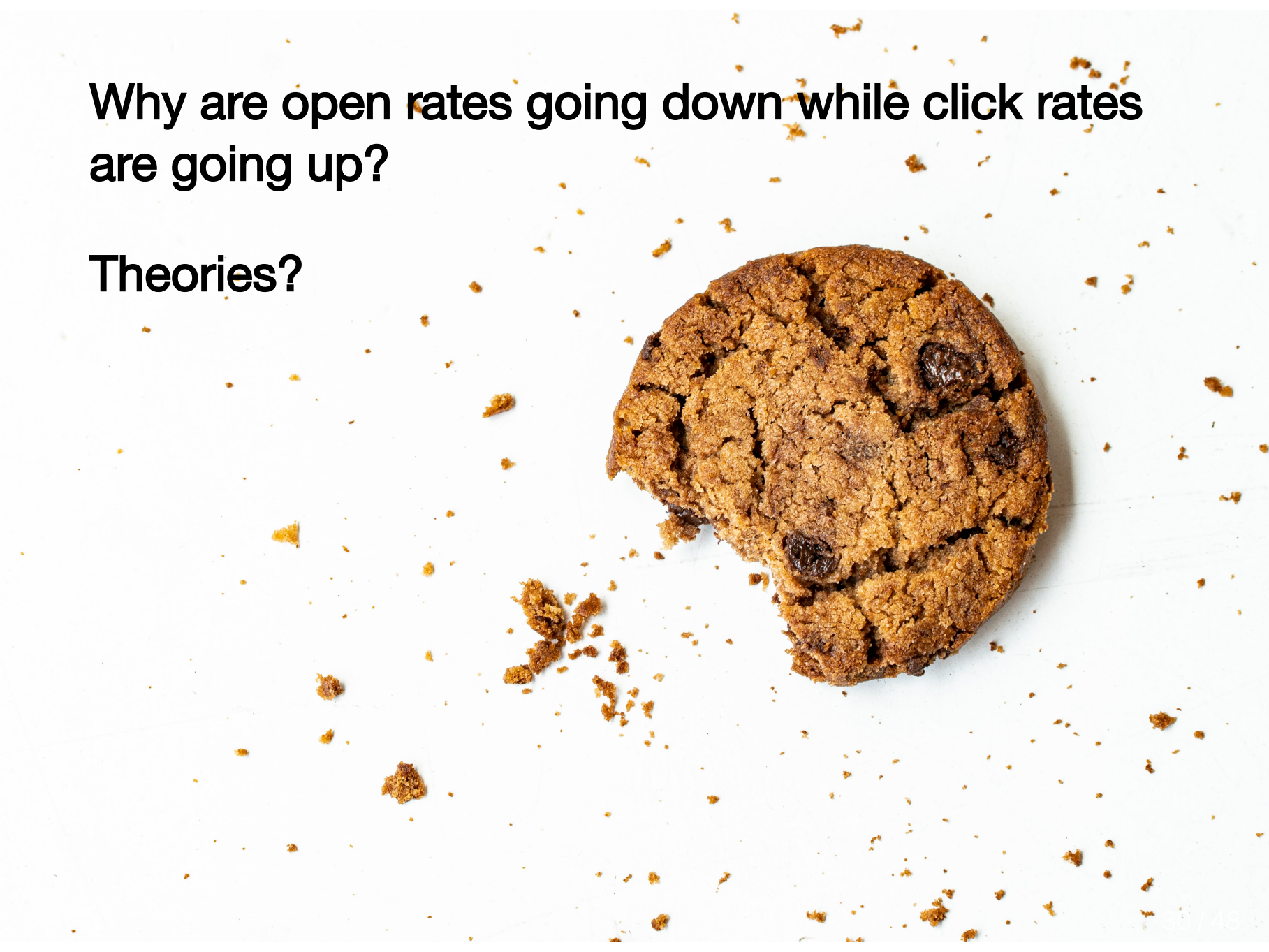
 The business model depends on open and clicks for the software to work


Email Open Rates Over Time
What's Happening?
Email Open Rate
0.03
0.02
0.01
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
Months of Software Use


Email Click Rates Over Time
What's Happening?
Email Click Rate
1.00
0.75
0.50
0.25
0.00
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
Months of Software Use

Why are open rates going down while click rates are going up?

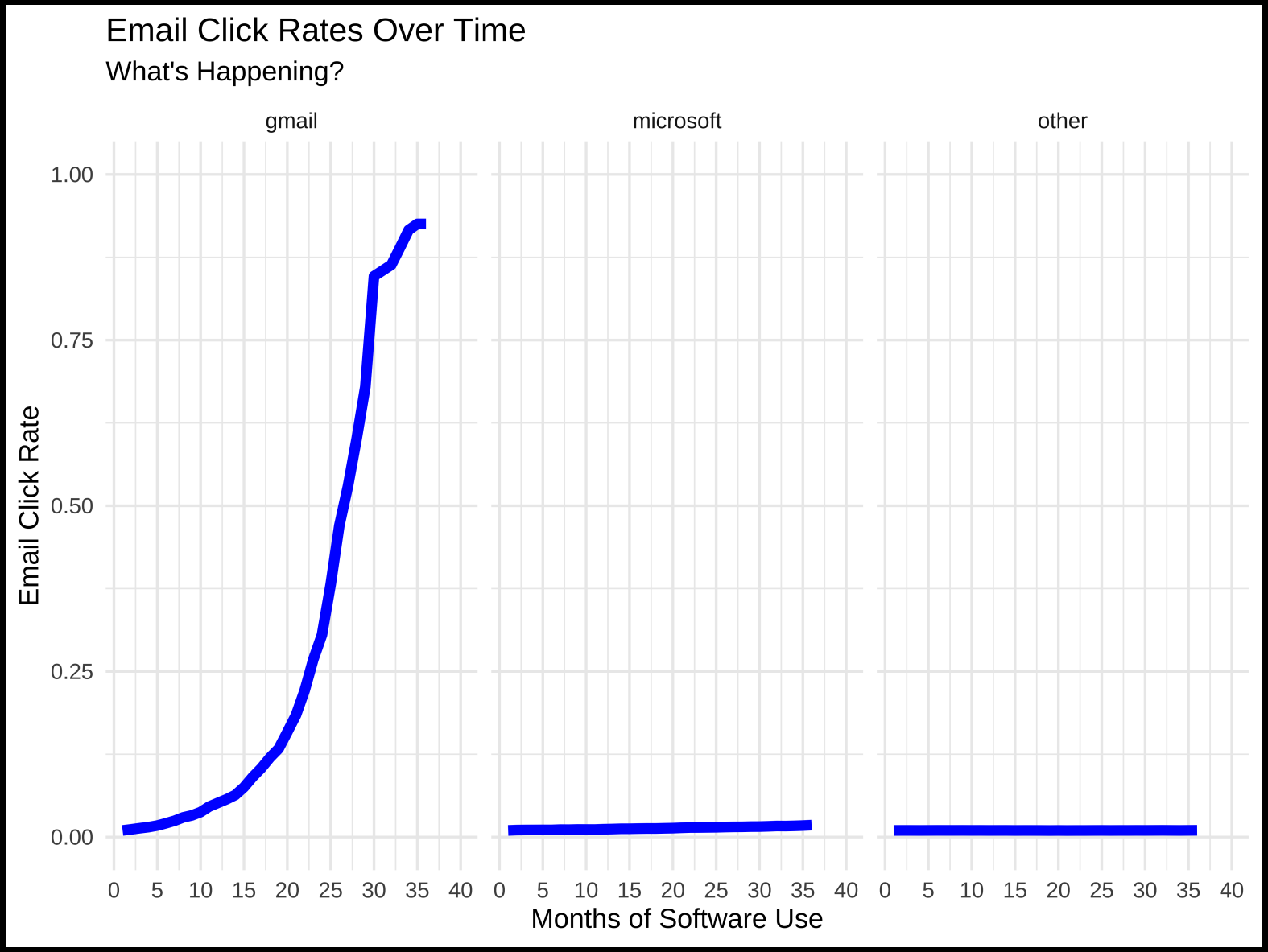
Theories?

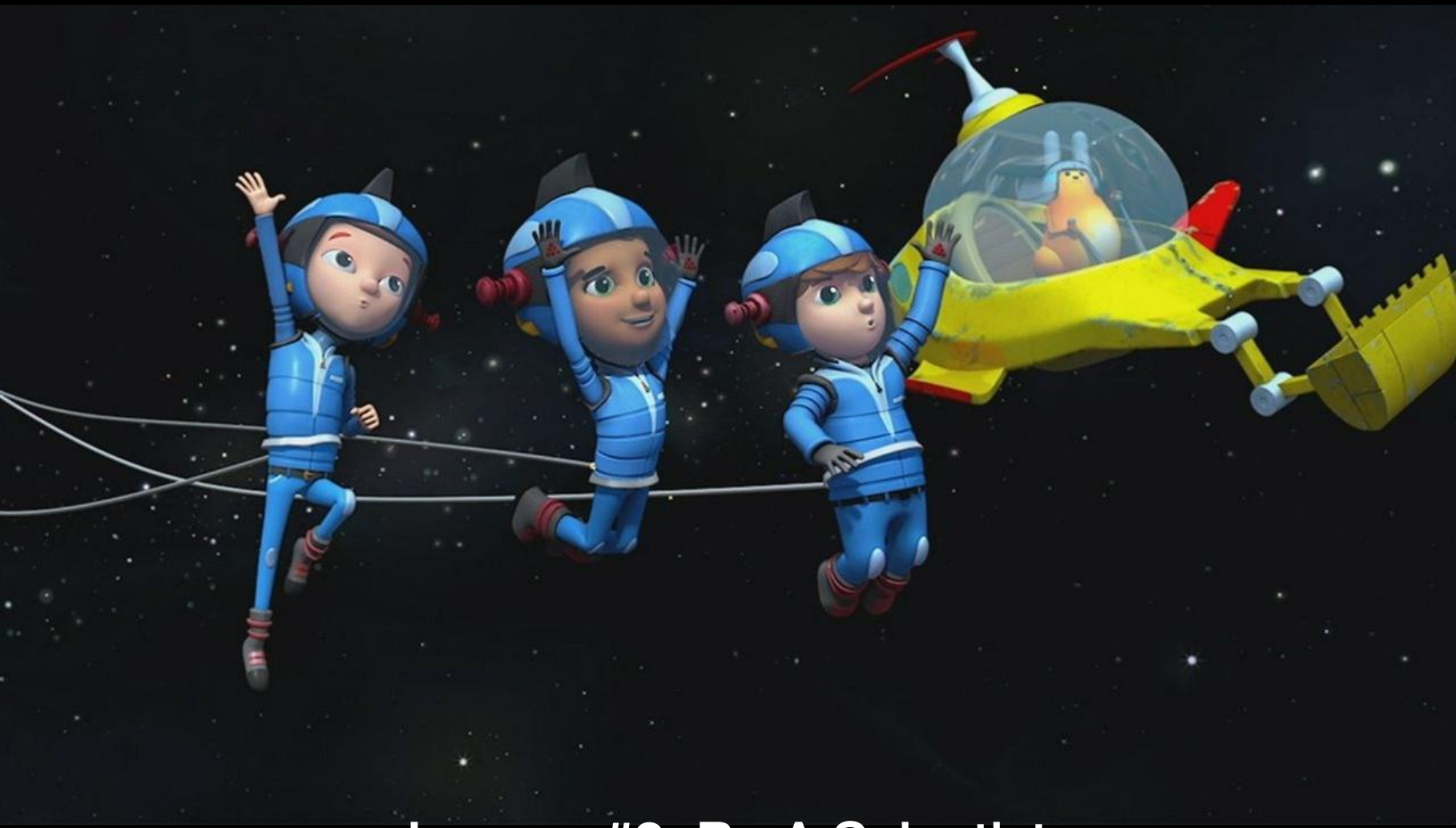


? SPAM filters catching emails for some reason

? IP addresses were cold or otherwise not delivering

? Email Click Bots


Email Click Rates Over Time
What's Happening?
gmail
microsoft
other
Email Click Rate
1.00
0.75
0.50
0.25
0.00
0 5 10 15 20 25 30 35 40
Months of Software Use



Lesson #2: Be A Scientist

Experimental Email with Shown Links

Dear {first_name},

You have previously expressed interest in this super awesome place. We would love for you to come visit this super awesome place at your nearest convenience.

Every word in this sentence links to a different landing page.

During your visit to super awesome place, we have a lot of great activities planned including a Midnight Jamboree a data hackathon, and a pie-eating contest.

Sign up below!

Sincerely,
The Super Awesome Place

Experimental Email with Hidden Links

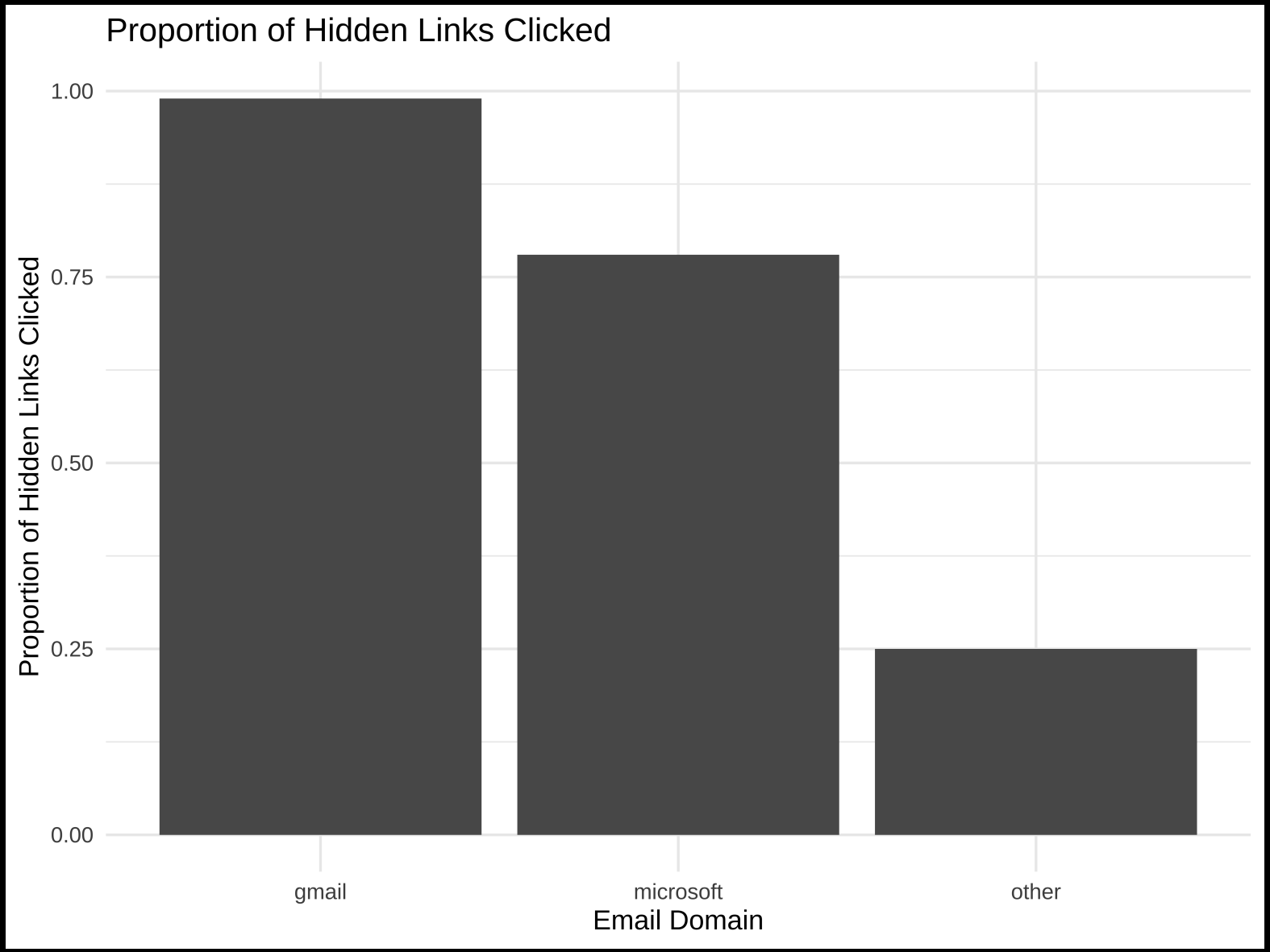
Dear {first_name},

You have previously expressed interest in this super awesome place. We would love for you to come visit this super awesome place at your nearest convenience.

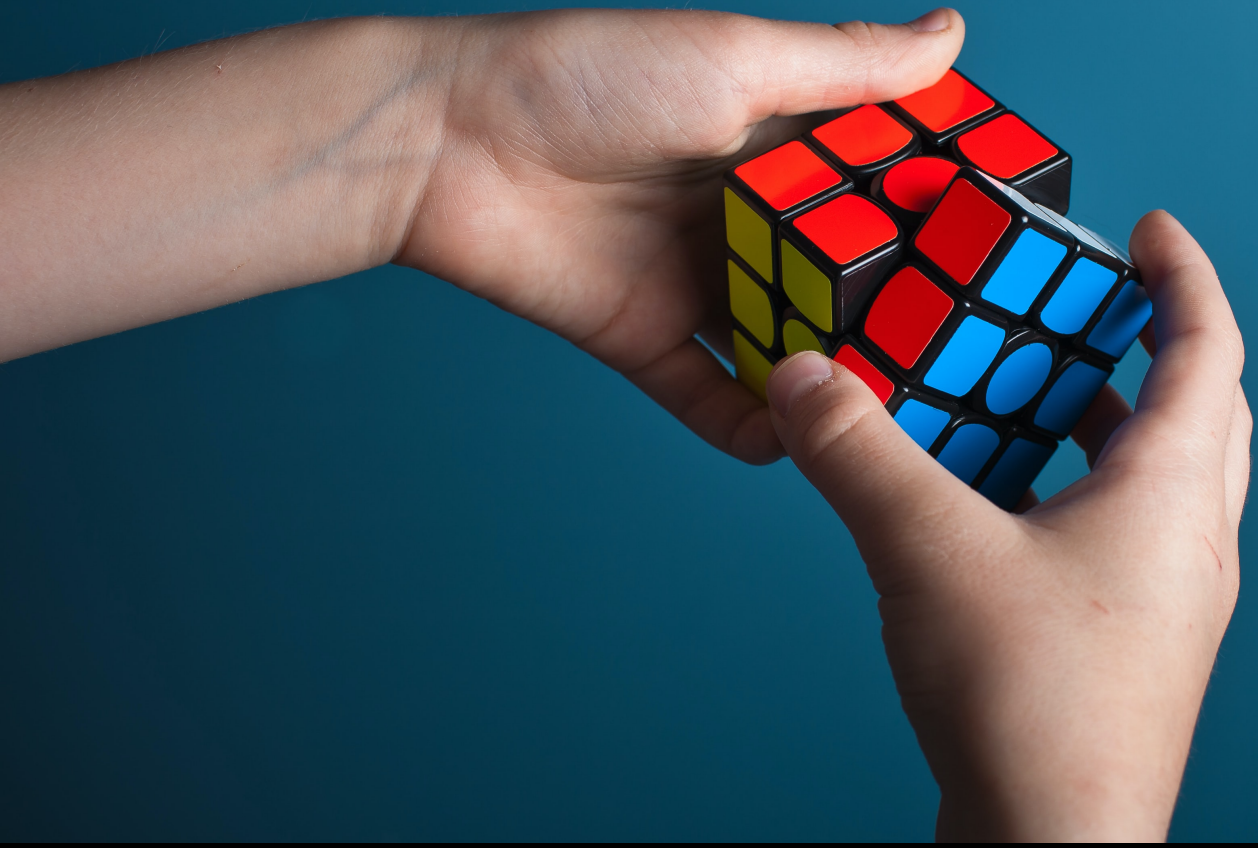
During your visit to super awesome place, we have a lot of great activities planned including a Midnight Jamboree a data hackathon, and a pie-eating contest.

Sign up below!

Sincerely,
The Super Awesome Place

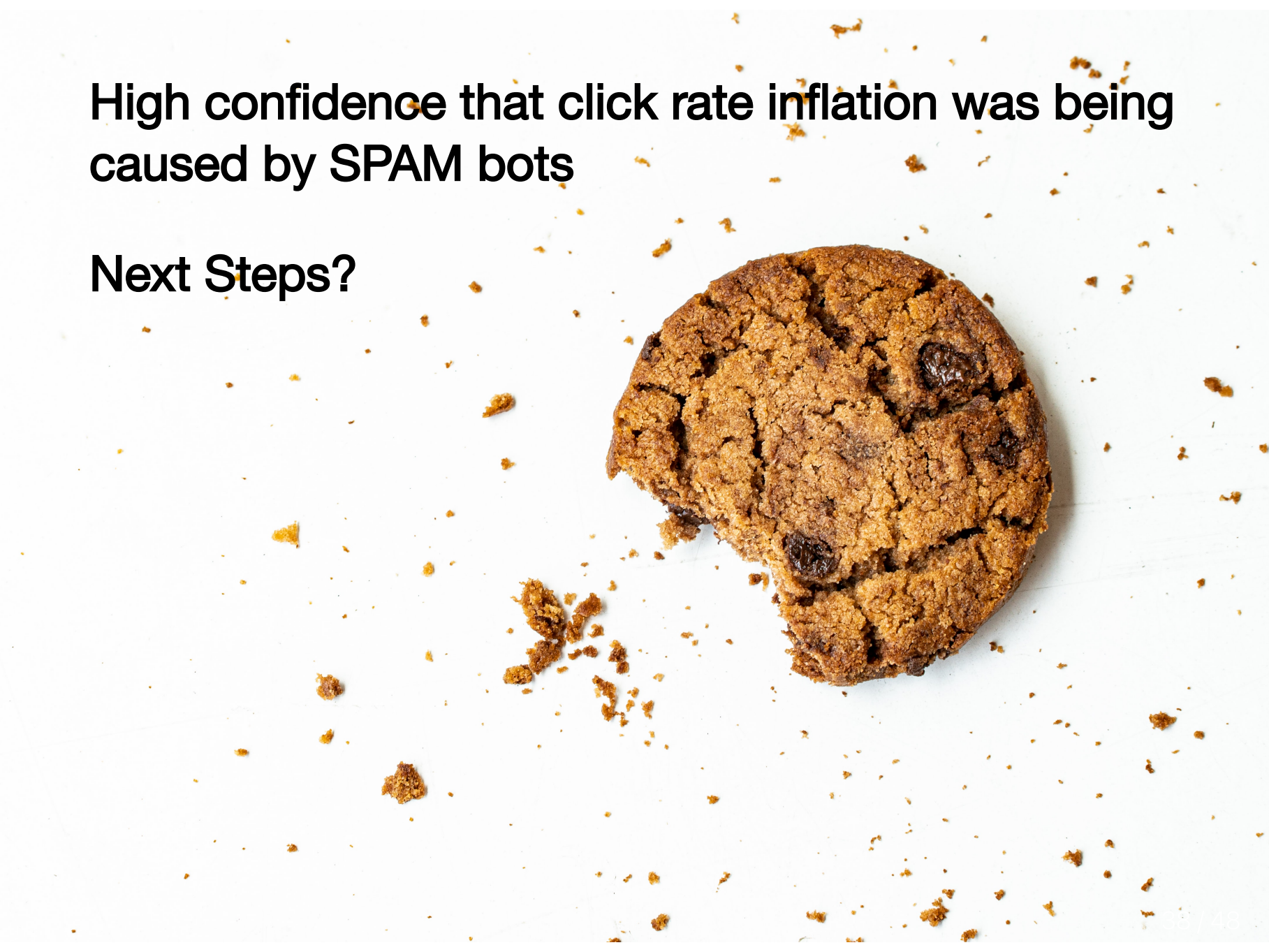

Proportion of Hidden Links Clicked
1.00
0.75
0.50
0.25
0.00
Proportion of Hidden Links Clicked
gmail
microsoft
other
Email Domain

Lesson #3: Don't just highlight problems, solve them!



High confidence that click rate inflation was being caused by SPAM bots

Next Steps?

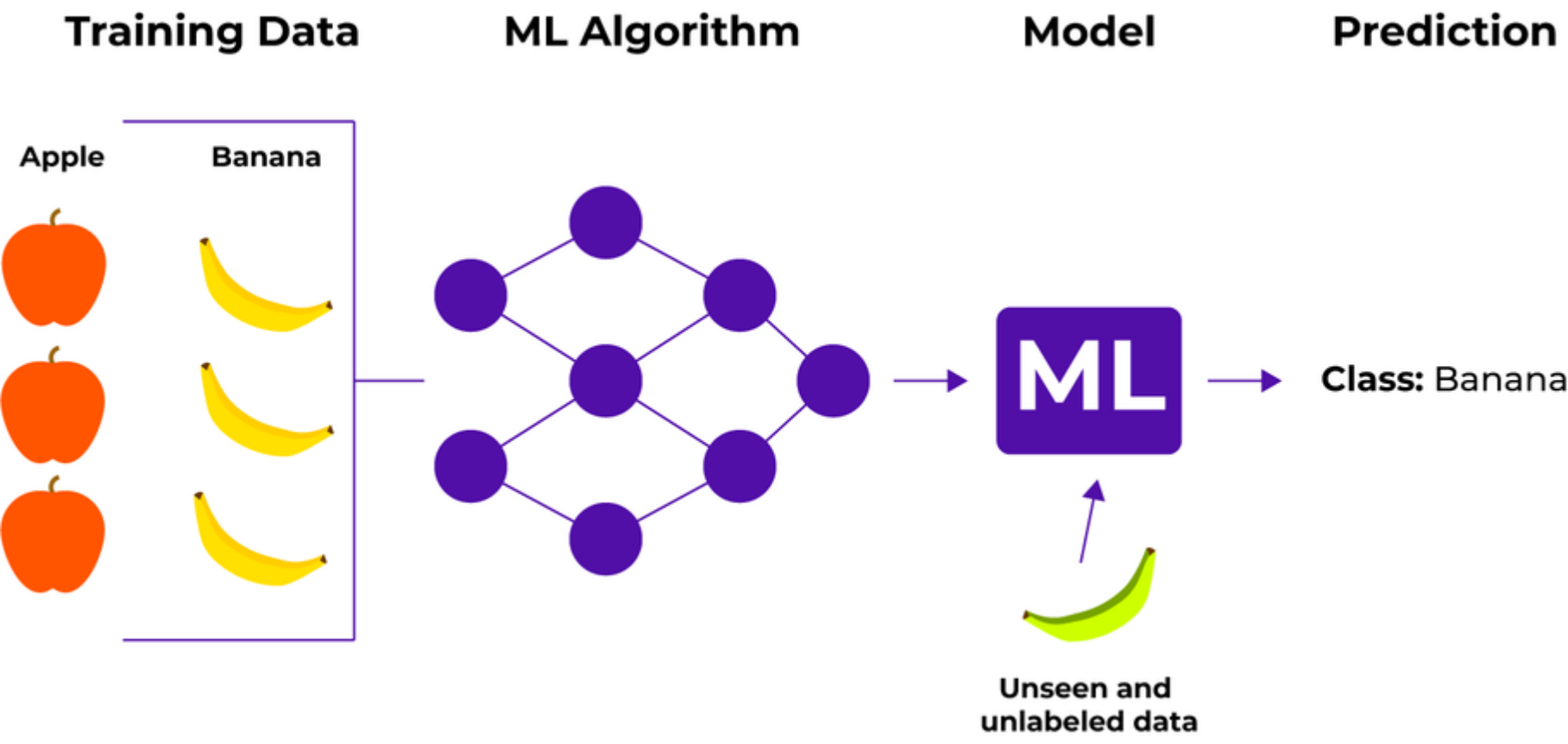


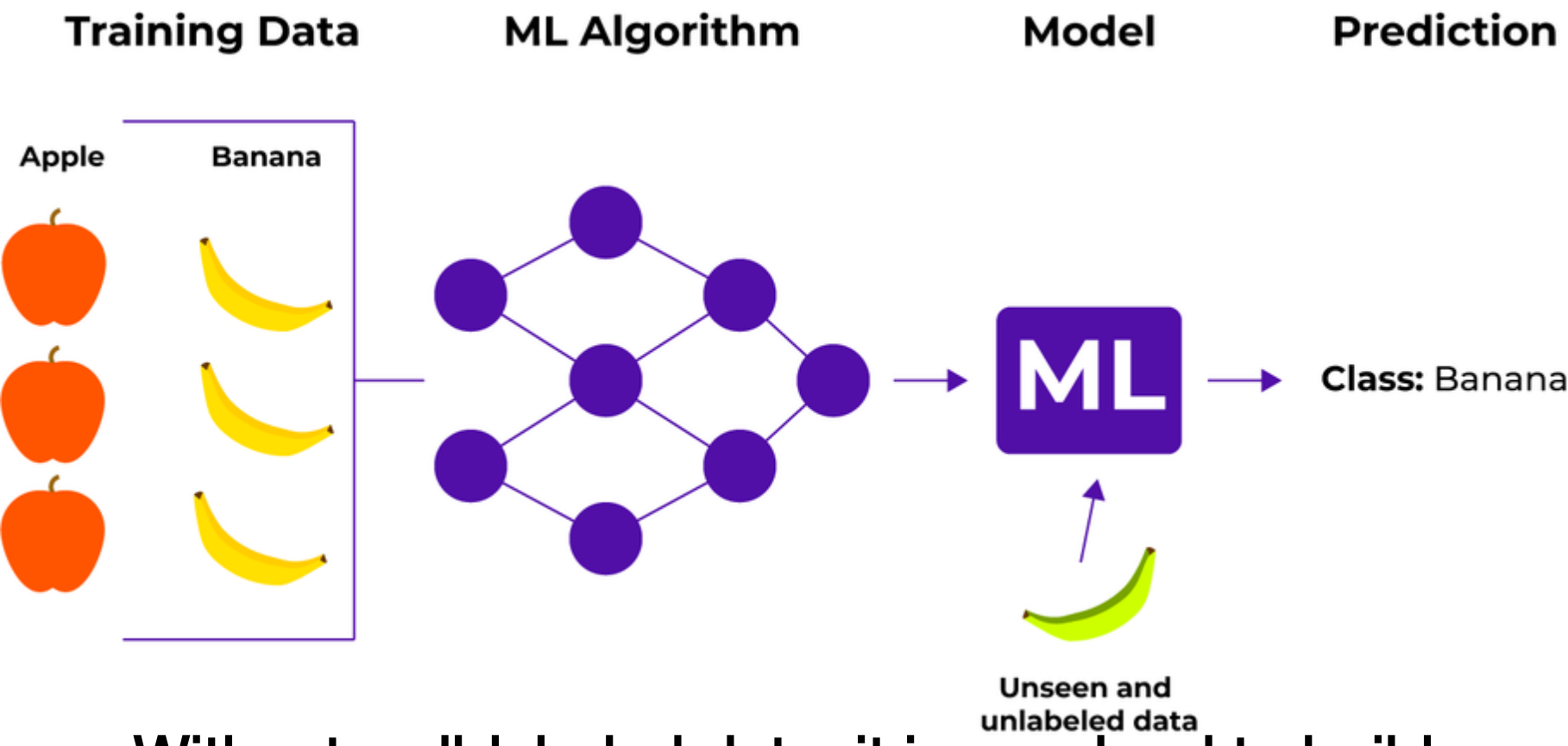
✅ Estimate which clicks were coming from humans and which were coming from bots

✅ Programmatically "remove" those opens + clicks from the denominators

✅ Explain problem to stakeholders and leverage this toolkit for competitive advantage

Standard Machine Learning Mental Model

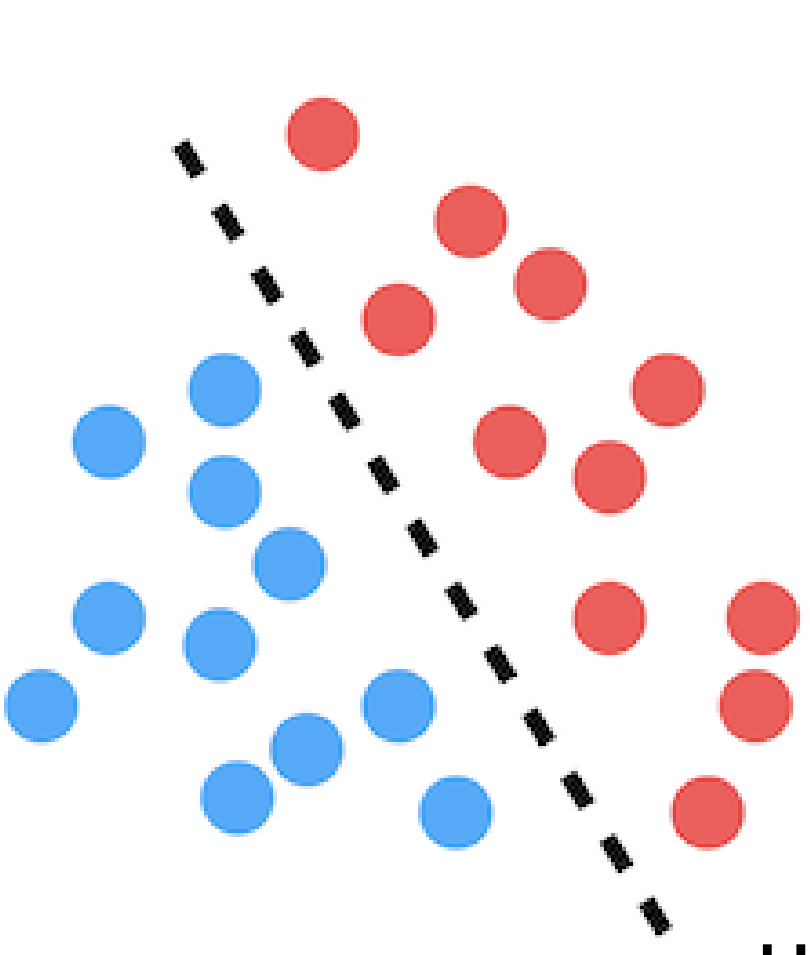


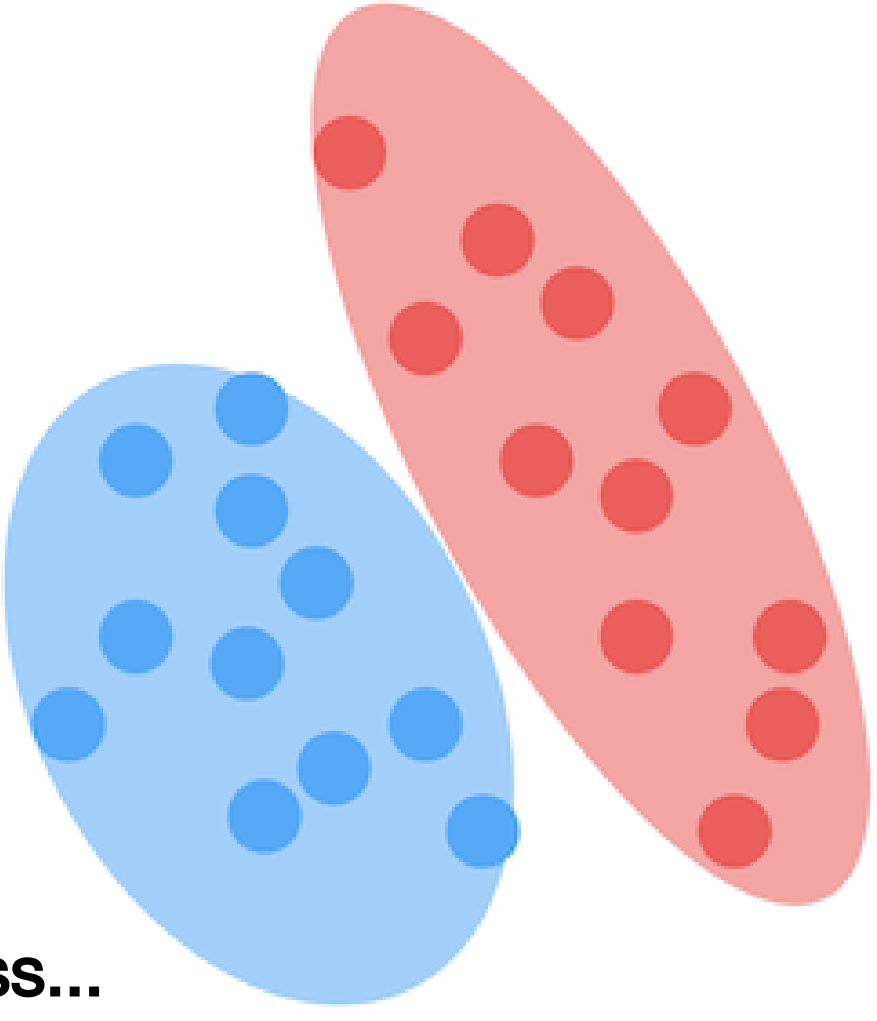


Without well-labeled data, it is *very* hard to build useful ML models

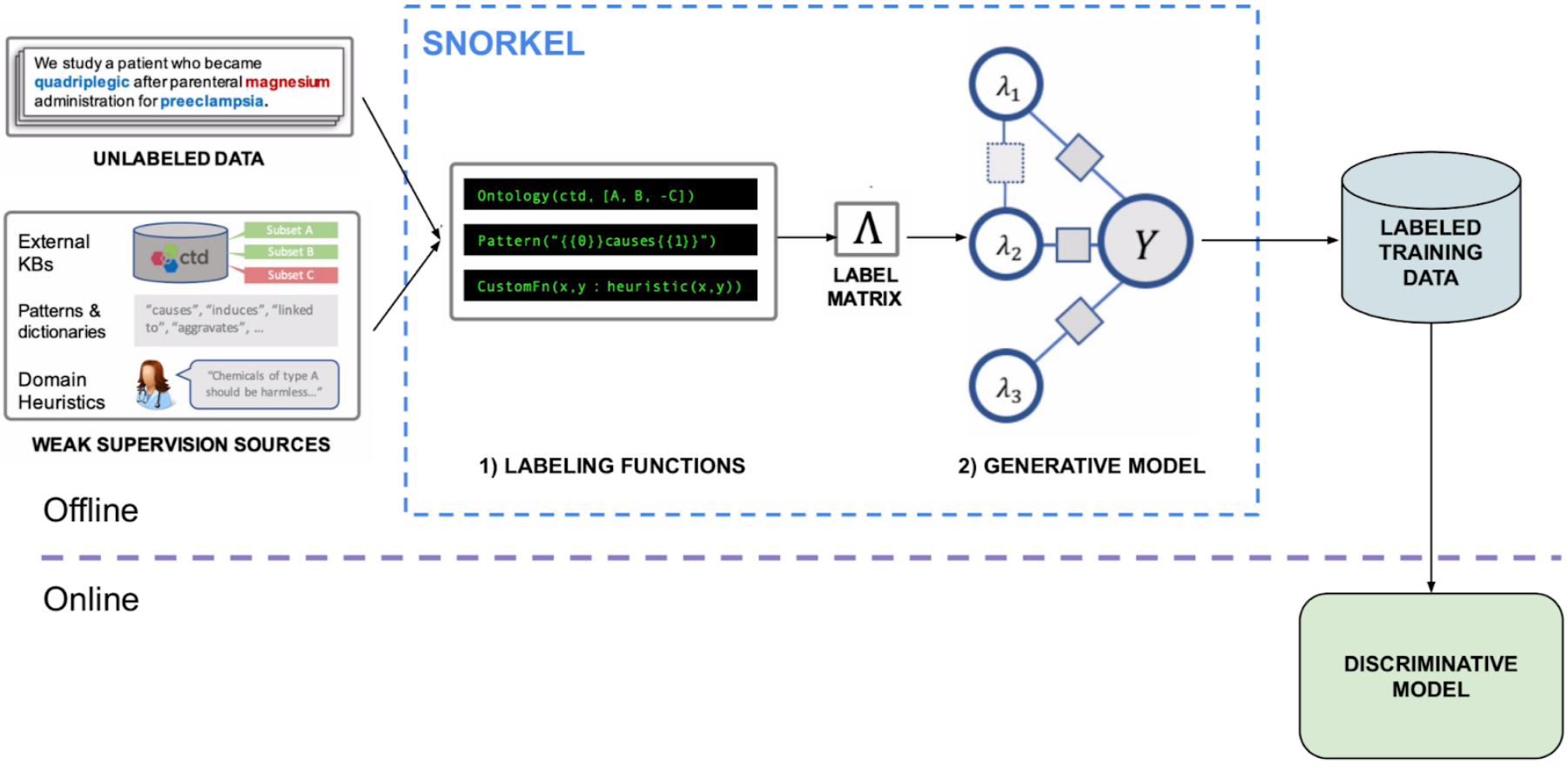
Discriminative

Generative





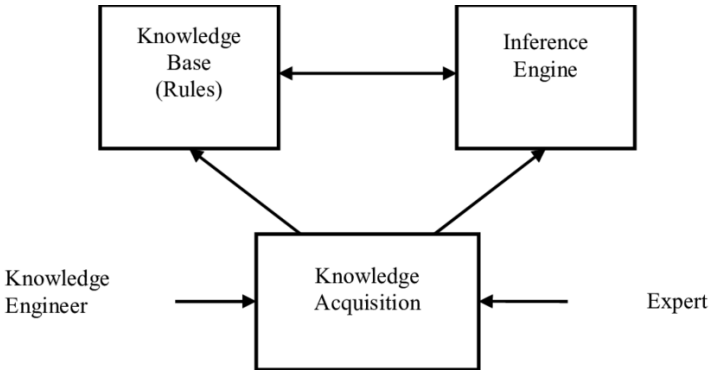
Unless...



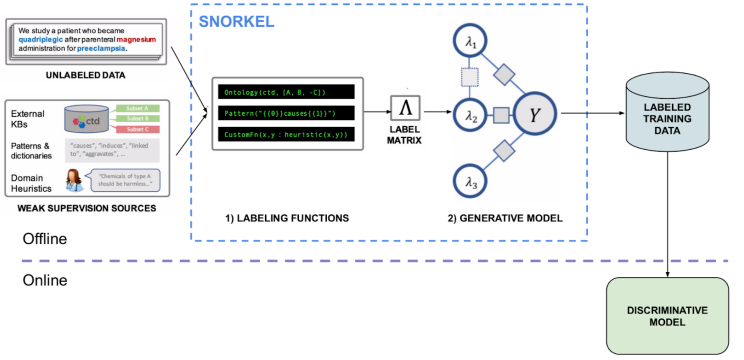
You can reasonably "generate" the outcome labels from a known probability distribution

The Choice:

Simple



Complex

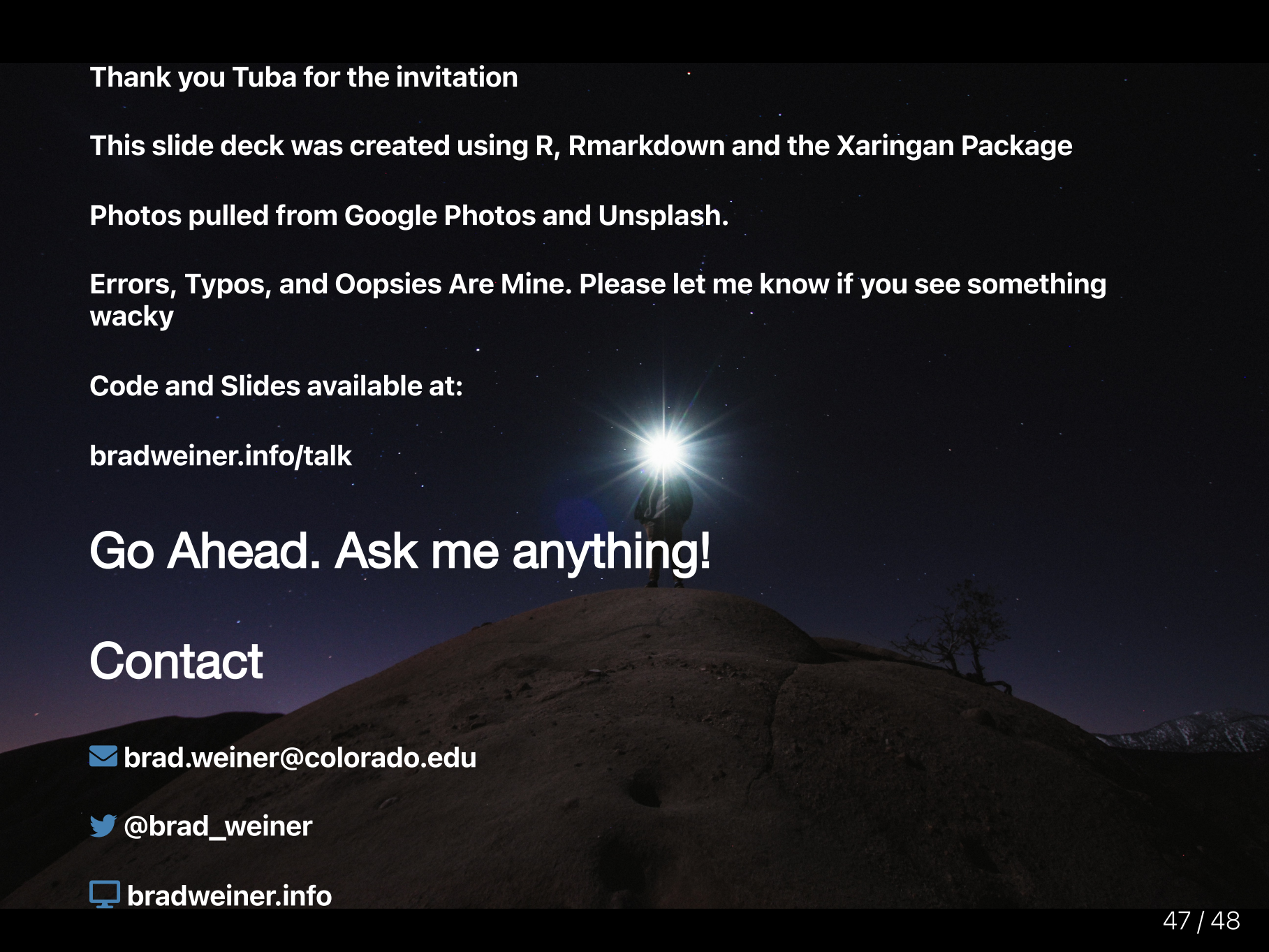


Lesson #4: Assuming similar accuracy,

always preference the simple model.

Lesson #5: Learn Goodhart's Law.

When a measure becomes an outcome, it ceases to be a useful measure.



Thank you Tuba for the invitation

This slide deck was created using R, Rmarkdown and the Xaringan Package

Photos pulled from Google Photos and Unsplash.

Errors, Typos, and Oopsies Are Mine. Please let me know if you see something wacky

Code and Slides available at:

bradweiner.info/talk

Go Ahead. Ask me anything!

Contact

brad.weiner@colorado.edu

@brad_weiner

bradweiner.info