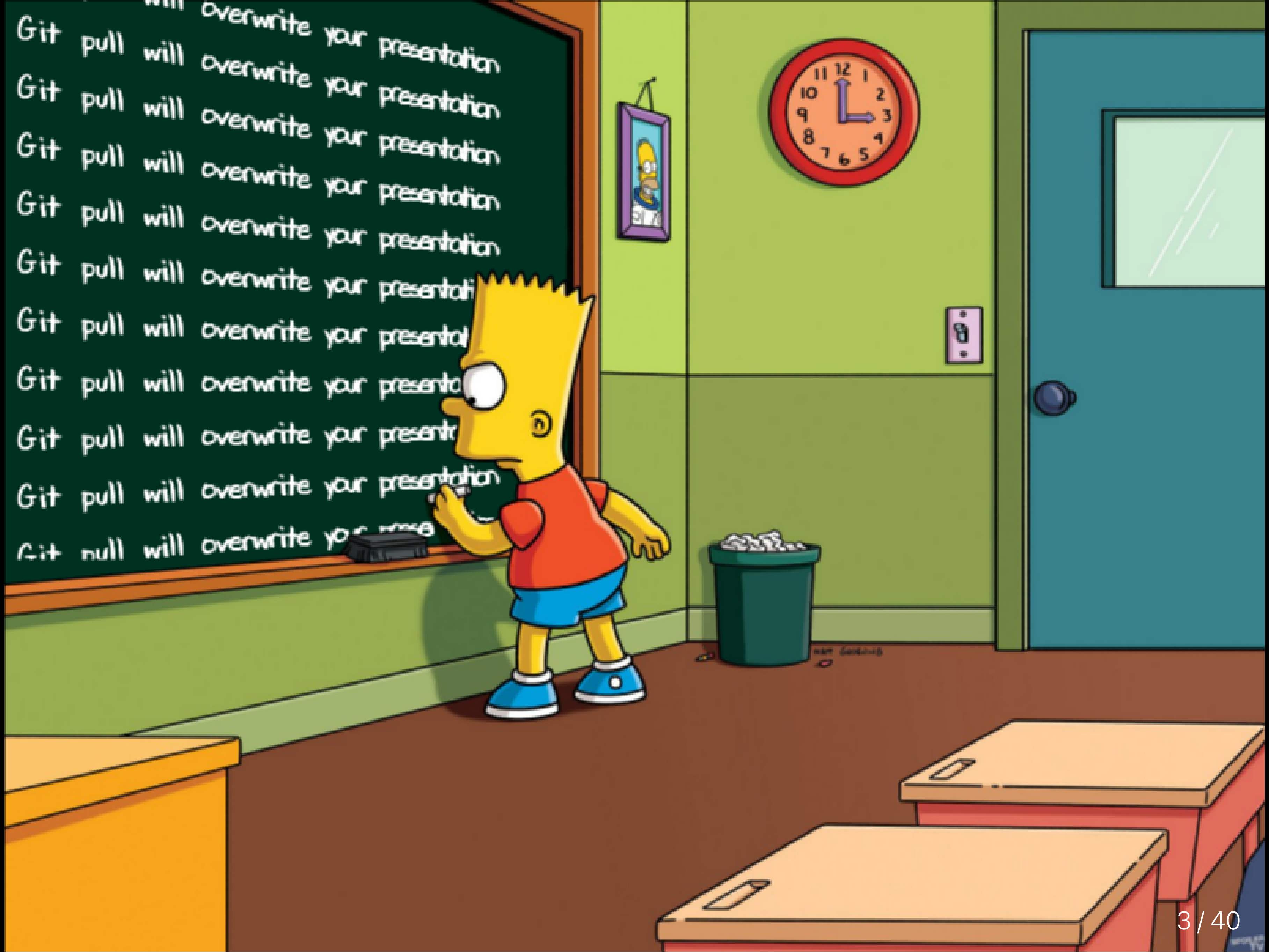
Data Literacy is Data ~~Governance~~ Enablement

Because Documentation and Access Controls Aren’t Enough

Brad Weiner | Chief Data Officer

Data CoP IRL Conference, University of Colorado Boulder

2023-04-21


will overwrite your presentation
Git pull will overwrite your presentation
Git pull will overwrite your presentation
Git pull will overwrite your presentation
Git pull will overwrite your presentation
Git pull will overwrite your presentati
Git pull will overwrite your presentat
Git pull will overwrite your presenta
Git pull will overwrite your presentr
Git pull will overwrite your presentation
Git pull will overwrite yo

About Me

Chief Data Officer, University of Colorado Boulder

21 years experience in higher education

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

5 years in Ed-Tech/Consultancy

14 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:

- Learn How Data Literacy is Part of Data ~~Governance~~ Enablement
- Practice asking better research questions
- Discuss how to tell better data stories
- Discuss how to convert data to insight to action

Data ~~Governance~~ Enablement Includes

THE WORDS "DATA GOVERNANCE" ARE THROWN AROUND A LOT THESE DAYS...

- Documentation
- Quality & Modeling
- Metadata
- Master Data Management
- Data Access Policies
- Security/Privacy
- Data Catalog
- Lifecycle Management
- *Data Literacy*

How Can Data Literacy Help Campus?

- We can ask better questions
- We can communicate better with data
- We can turn insights into action
- We can avoid "data theater"

A Scenario:

- You work for a fast food chain
- Their product team wants you to create a new "healthy" option
- As expected, they give you no other help
- You generate the following data set
- Good analyses start with good questions!

Sample of Fast Food Data from Kaggle (Not Verified)

restaurant	Sonic	Taco Bell	Taco Bell	Burger King	Arbys
item	Ultimate Chicken Club	Spicy Sweet Double Stacked Taco	Cool Ranch® Doritos® Locos Taco Supreme	Chicken Caesar Salad w/ Crispy Chicken	Classic French Dip & Swiss/Au Jus
calories	100	340	200	670	540
cal_fat	580	160	100	380	210
total_fat	64	18	12	43	23
sat_fat	15.0	7.0	4.5	7.0	11.0
trans_fat	0.5	0.0	0.0	0.0	1.0
cholesterol	100	35	35	80	85
sodium	2070	640	370	1760	2500
total_carb	65	32	15	40	50
fiber	4	4	3	5	2
sugar	12	6	3	8	3
protein	39	12	9	34	35
vit_a	15	10	NA	NA	2
vit_c	8	2	NA	NA	8
calcium	30	15	NA	NA	15
salad	Other	Other	Other	Other	Other

Types of Research Questions

- Descriptive (How many?)
- Correlational (Does x relate to y?)
- Predictive (What would we estimate)
- Prescriptive (What *should* we do?)

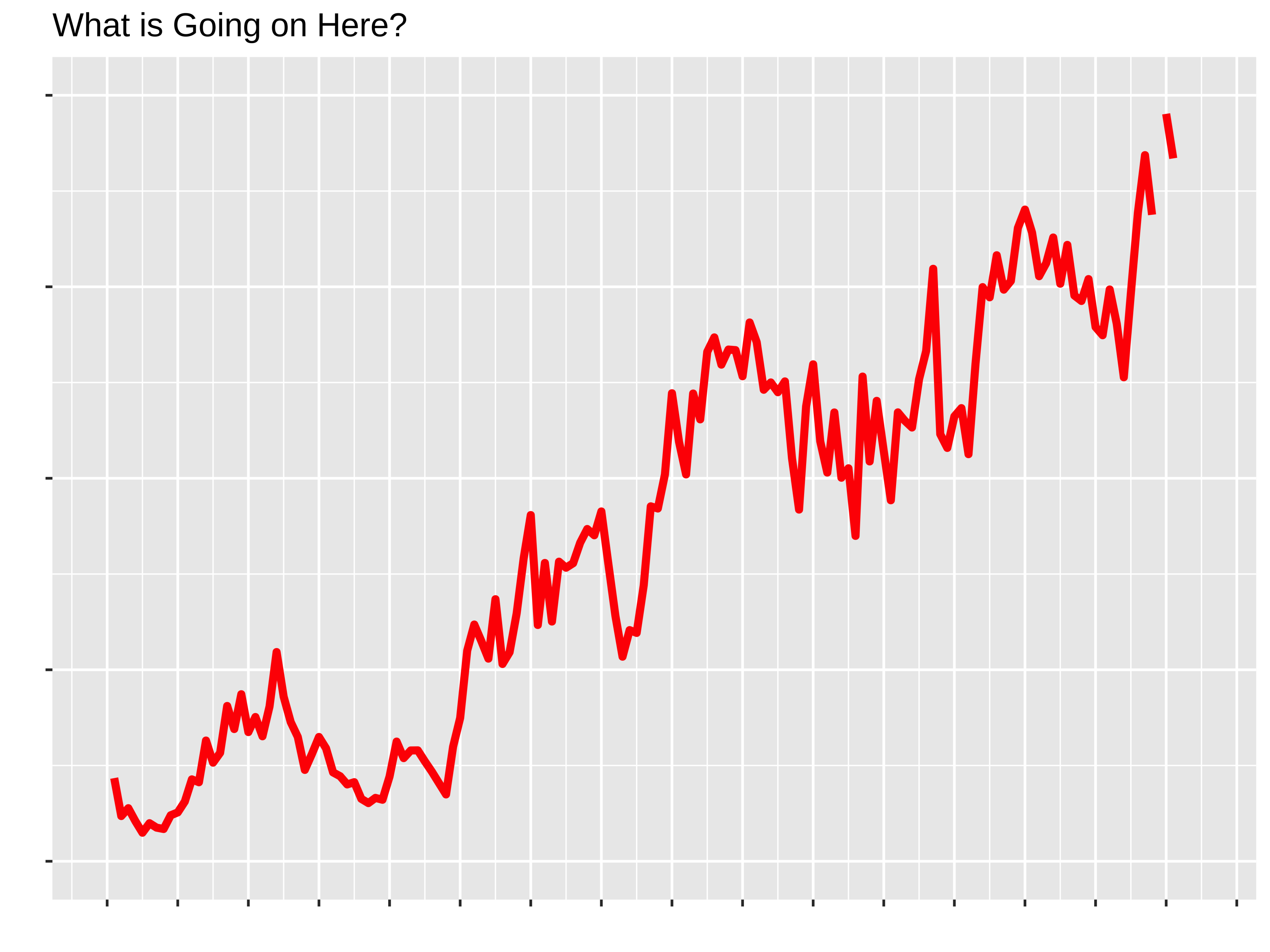
Practice Asking Some Research Questions

restaurant	Taco Bell	Subway	Burger King	Subway	Subway
item	Chipotle Crispy Chicken Griller	Footlong Big Hot Pastrami	Chicken BLT Salad w/ Crispy Chicken	Big Hot Pastrami Melt Salad	6" Black Forest Ham
calories	290	1160	690	400	290
cal_fat	170	620	430	300	40
total_fat	18	62	48	29	5
sat_fat	3	22	12	11	1
trans_fat	0	0	1	0	0
cholesterol	25	170	100	85	20
sodium	640	2940	1750	1250	830
total_carb	22	94	31	12	46
fiber	1	10	4	4	5
sugar	1	14	8	4	8
protein	9	58	35	23	18
vit_a	NA	20	NA	25	8
vit_c	NA	90	NA	70	20
calcium	NA	80	NA	10	30
salad	Other	Other	Other	Other	Other

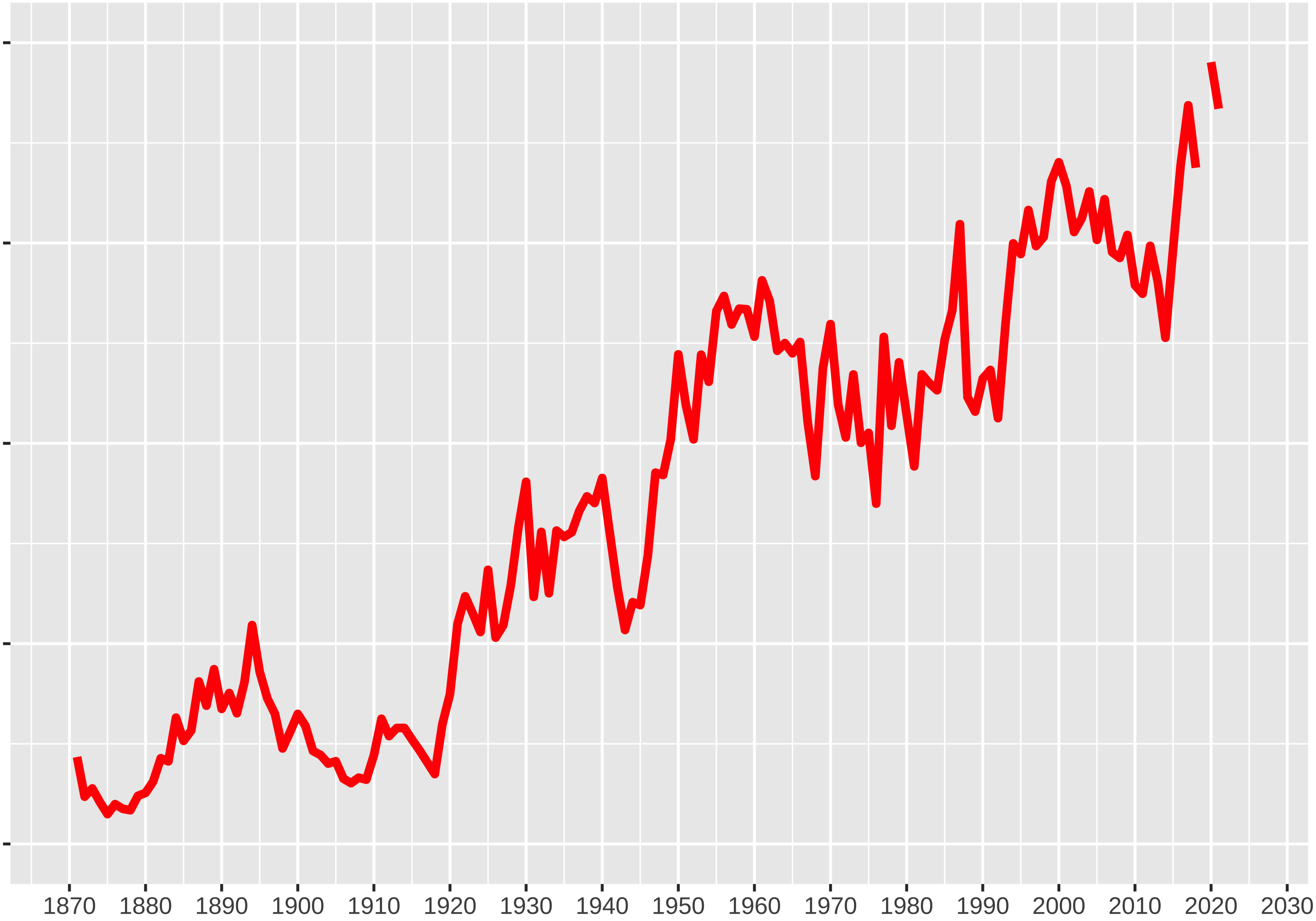
- Descriptive (How many?)
- Correlational (Does x relate to y?)
- Predictive (What would we estimate)
- Prescriptive (What *should* we do?)

Part of Data Literacy is Communicating with Data

- Let's tell a story about some data trends
- These data are real

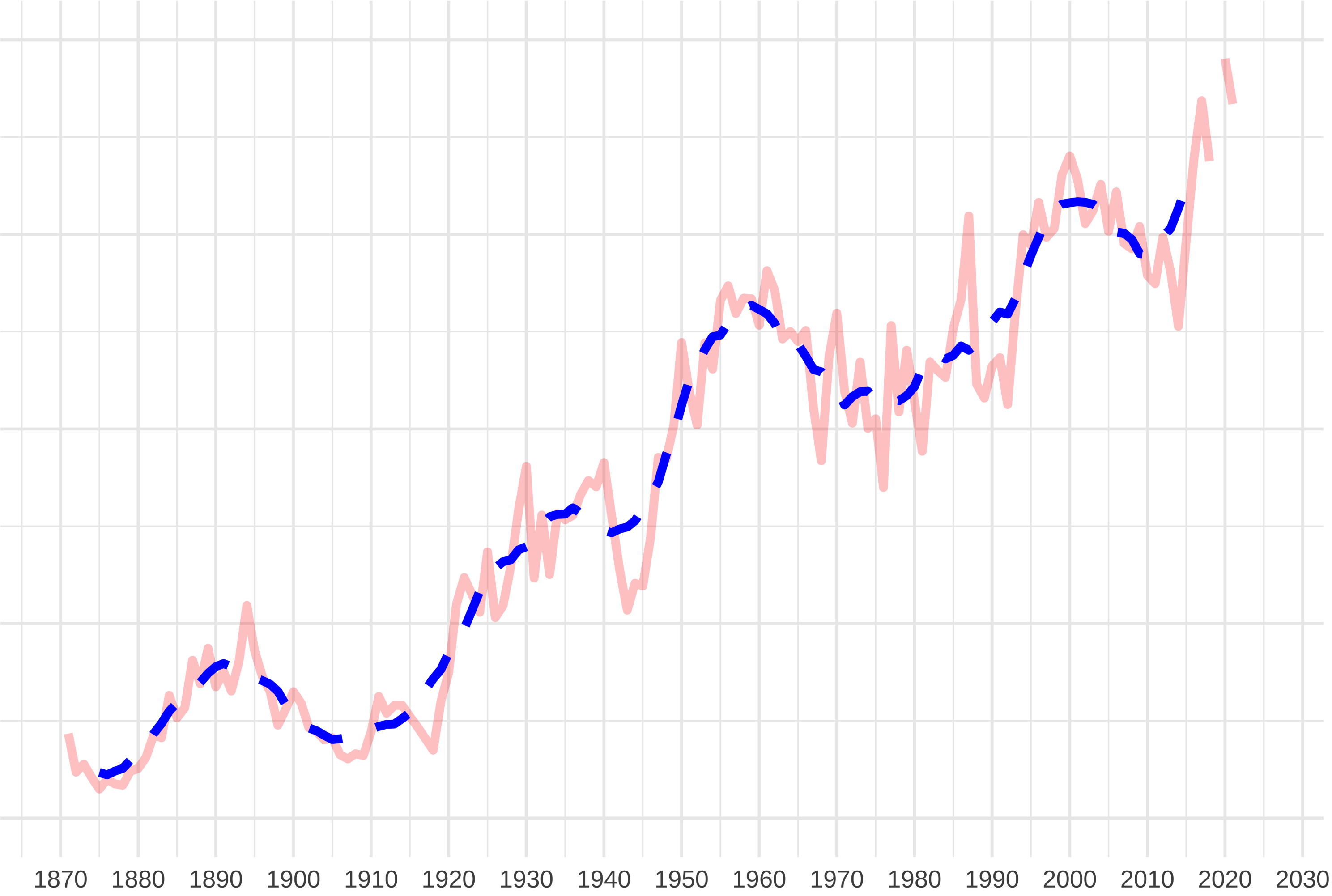

What is Going on Here?

Something Happening Between 1870 - 2018



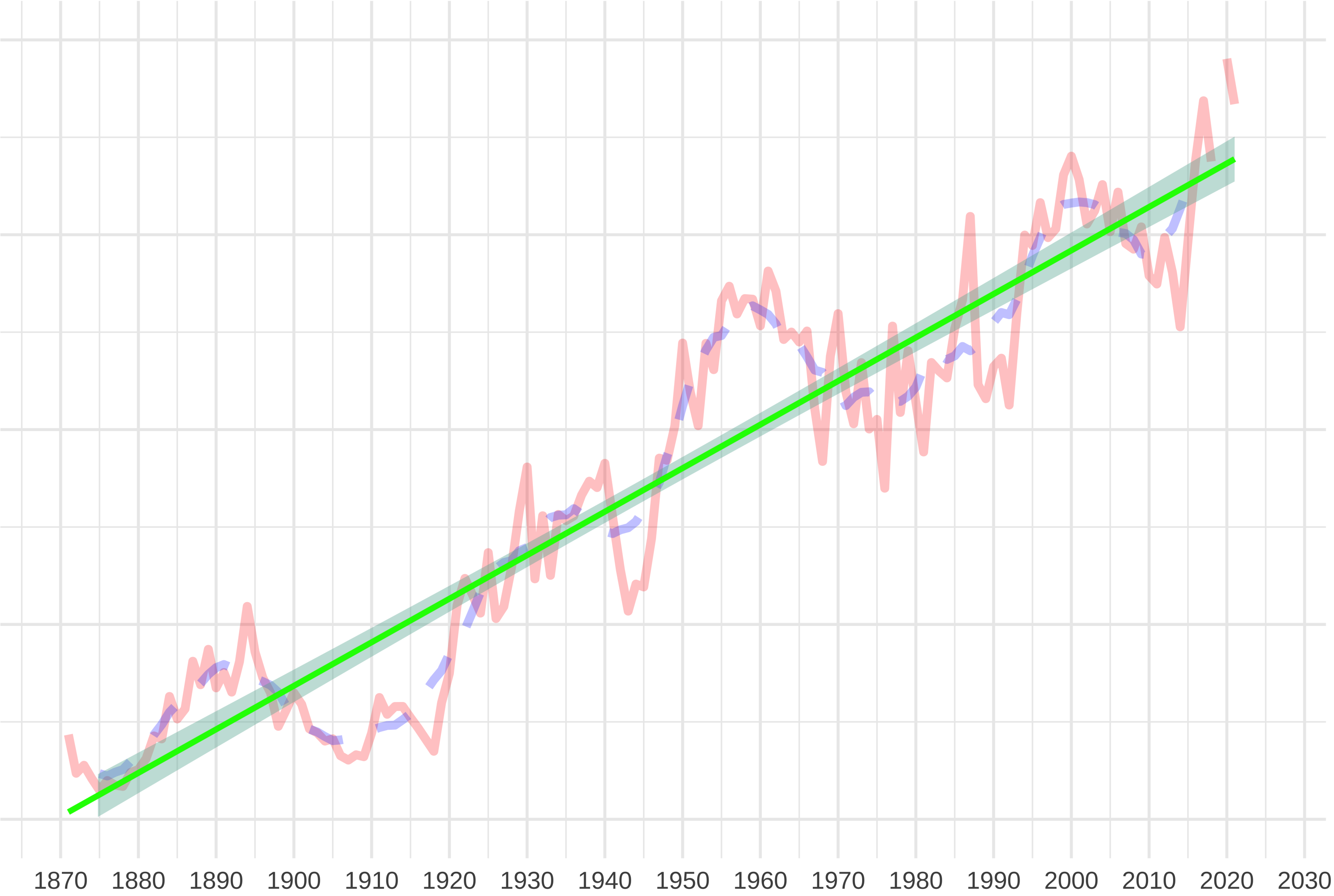
Something Happening Between 1870 - 2018

Ten Year Moving Average



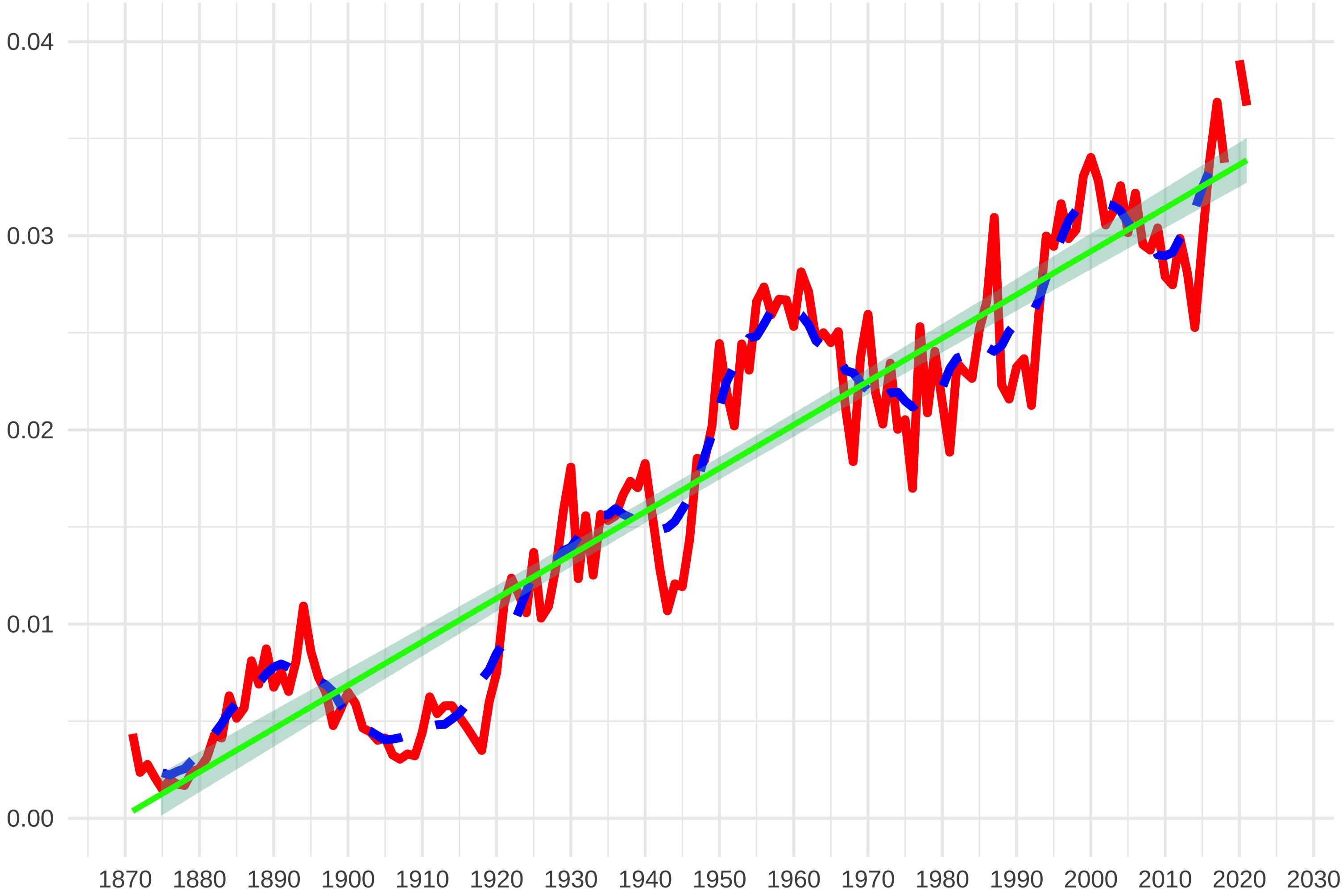
Something Happening Between 1870 - 2018

Ten Year Moving Average + Linear Trend



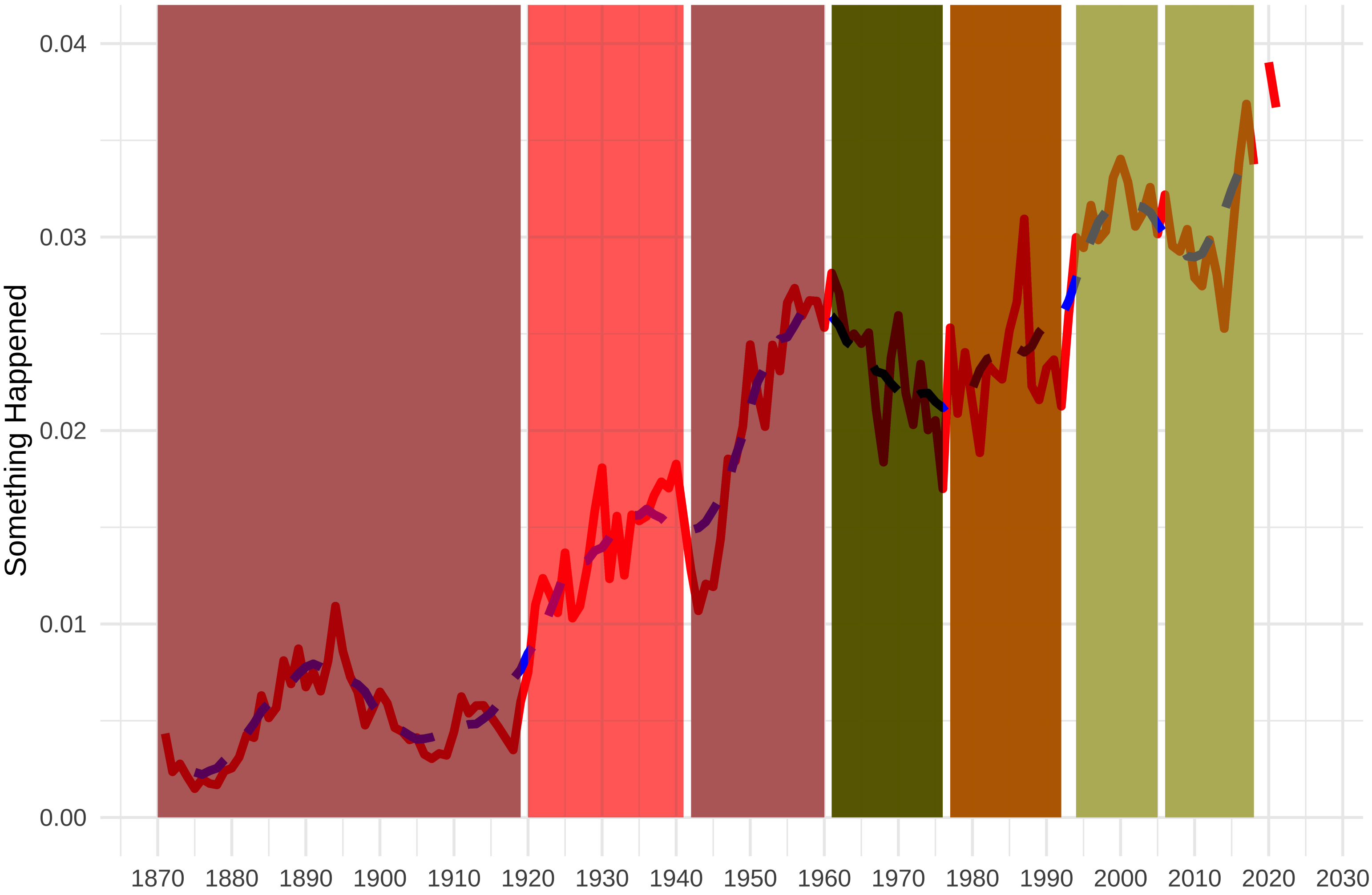
Something Happening Between 1870 - 2018

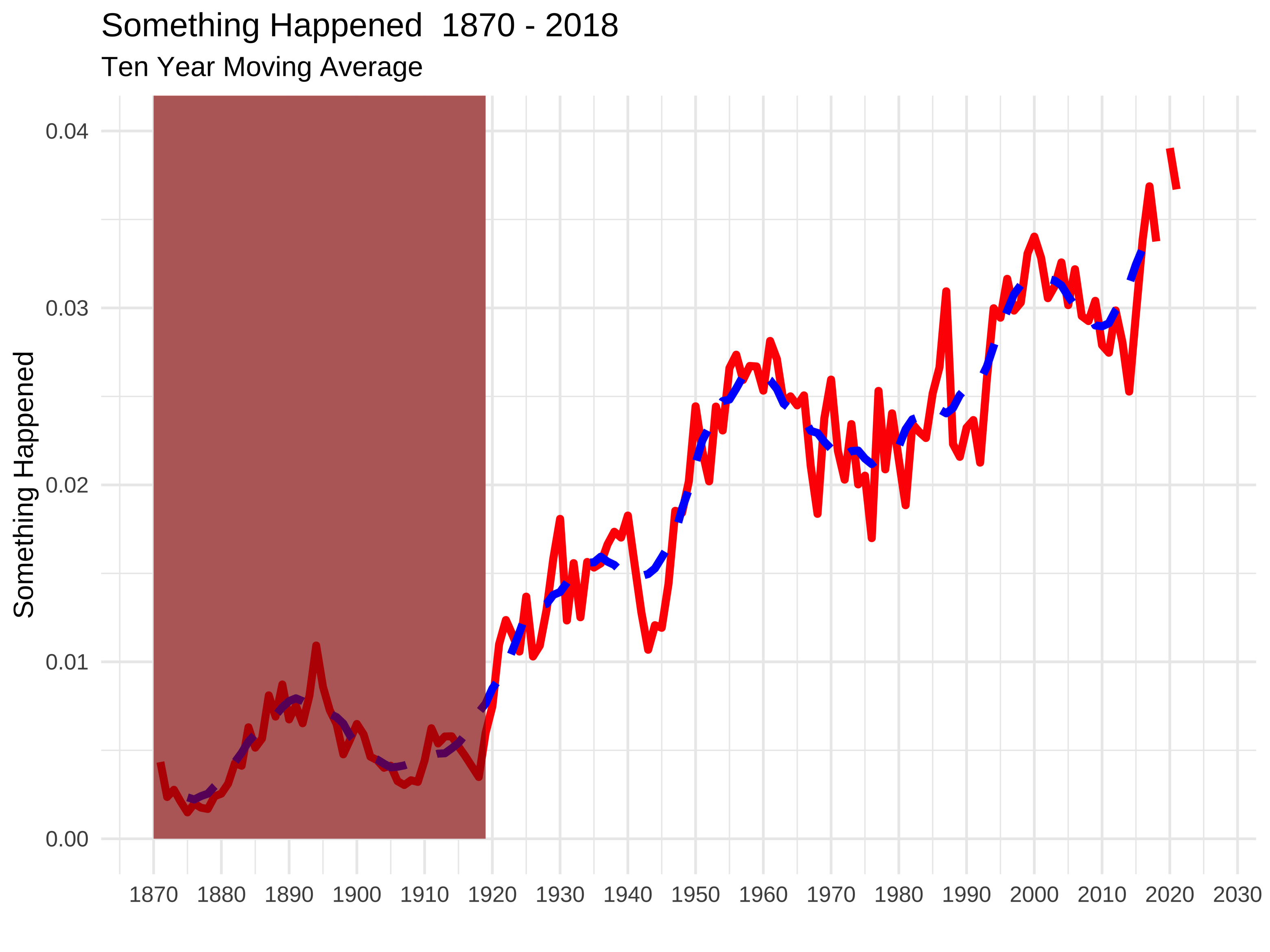
Ten Year Moving Average + Linear Trend



Something Happened 1870 - 2018

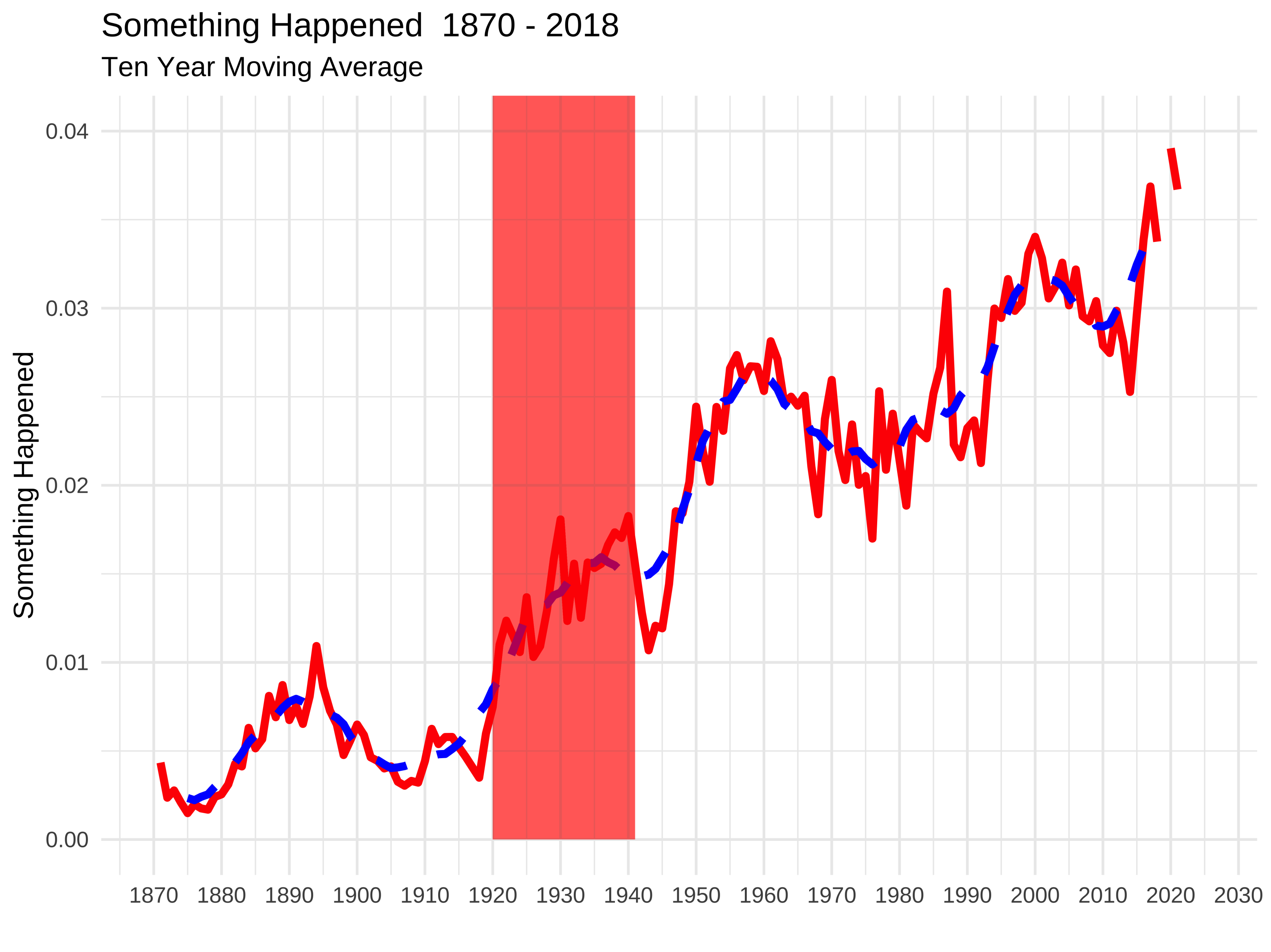
Ten Year Moving Average

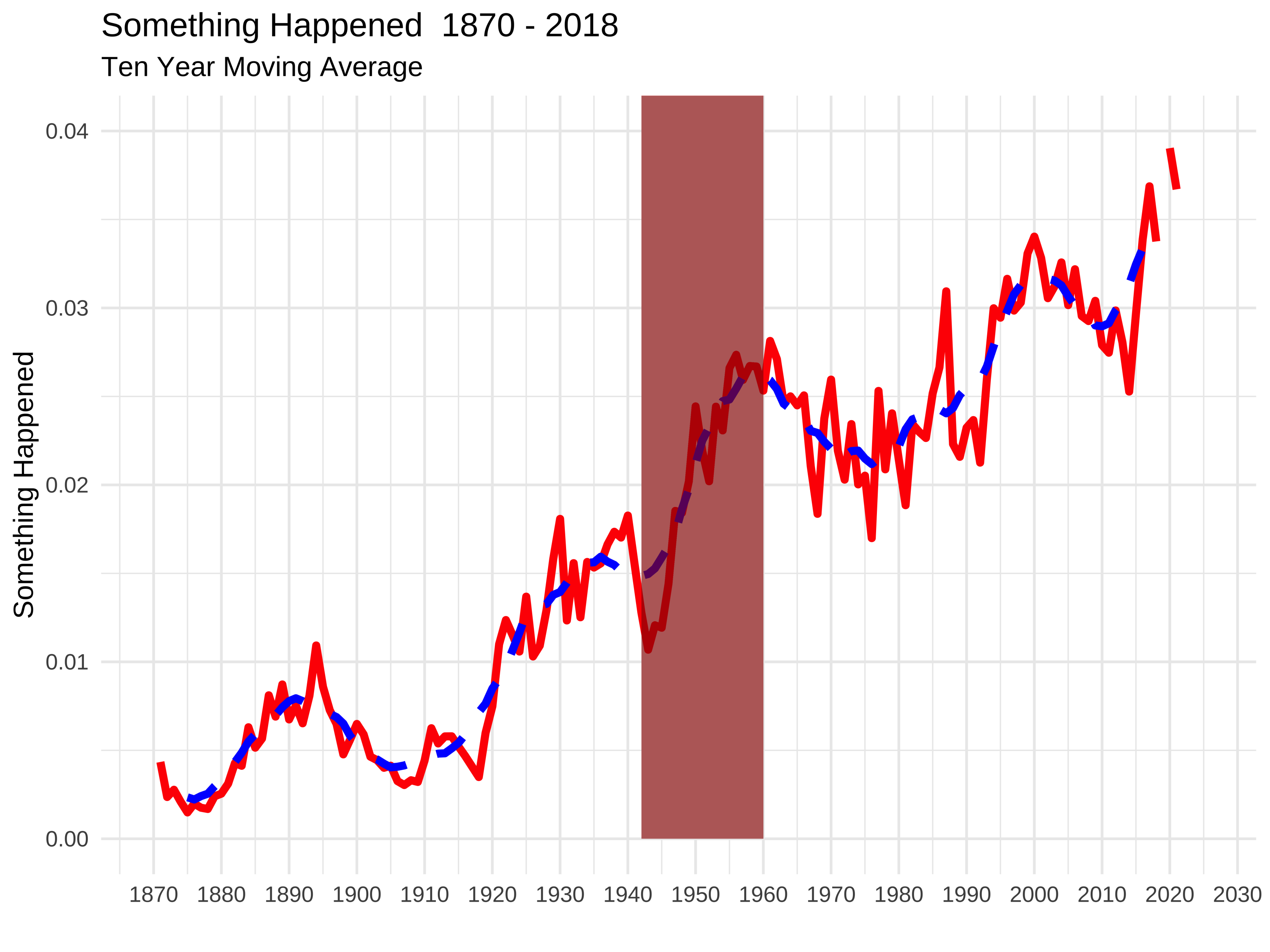


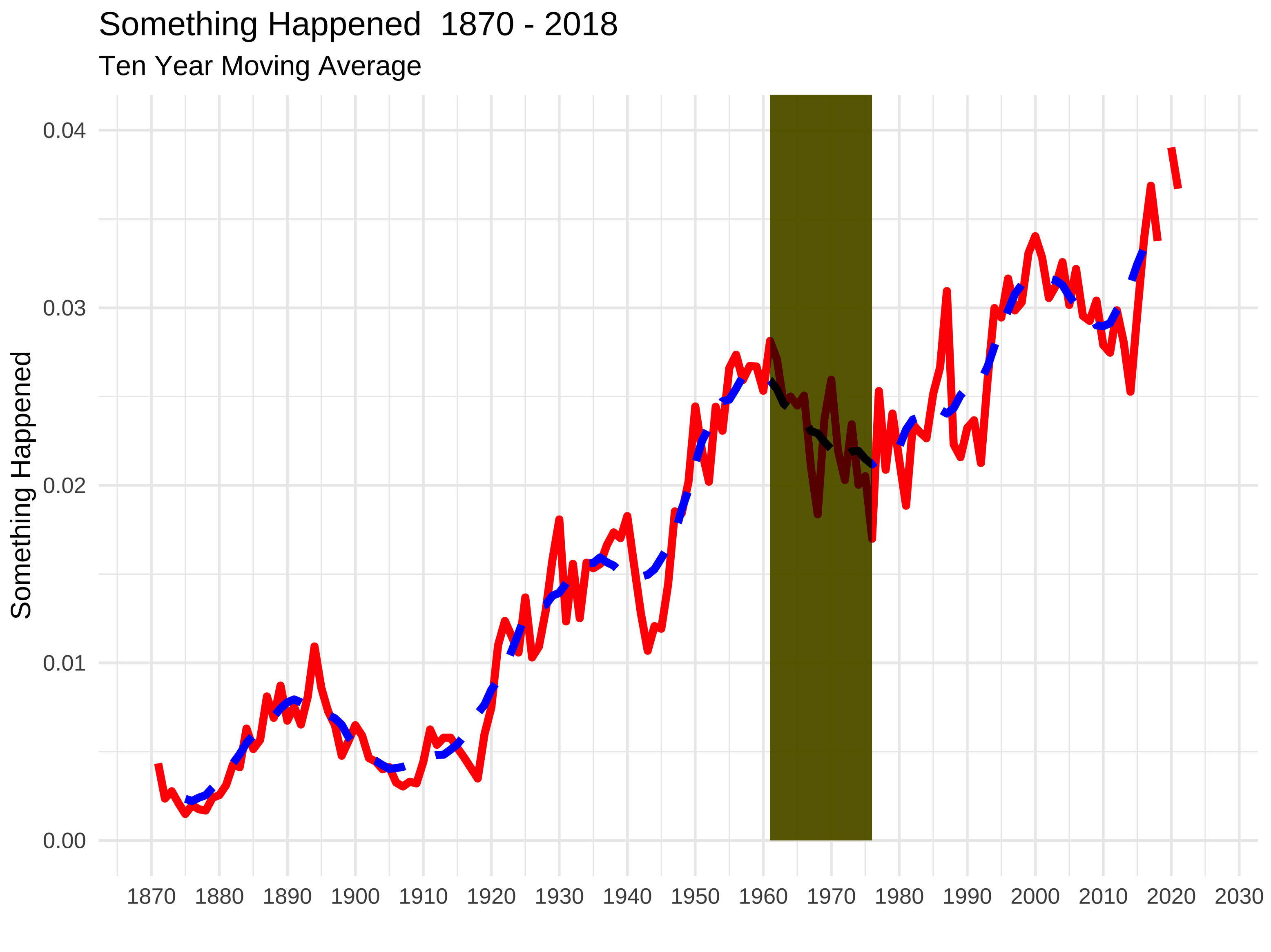


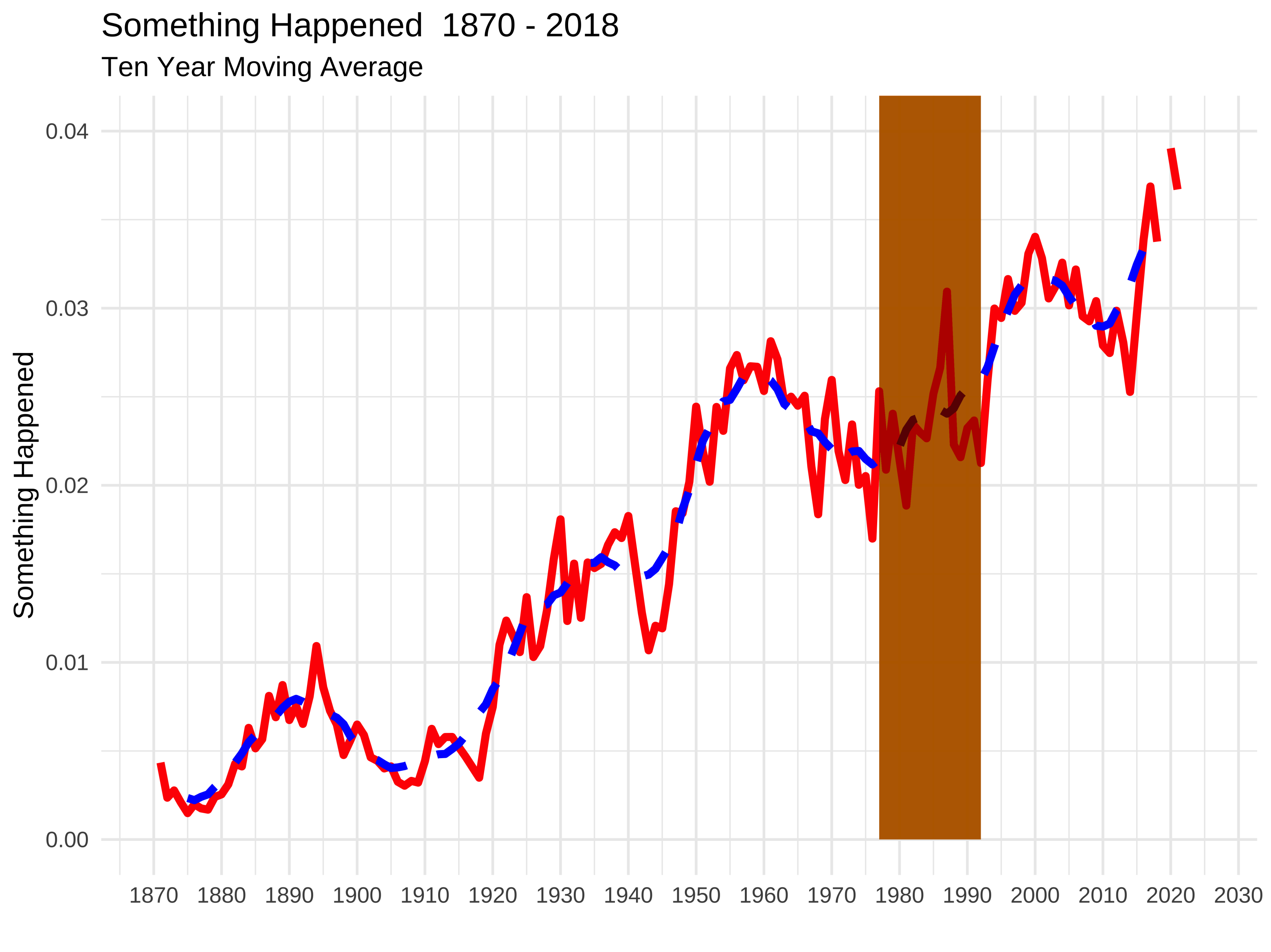
Something Happened 1870 - 2018

Ten Year Moving Average


Something Happened 1870 - 2018
Ten Year Moving Average
Something Happened
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

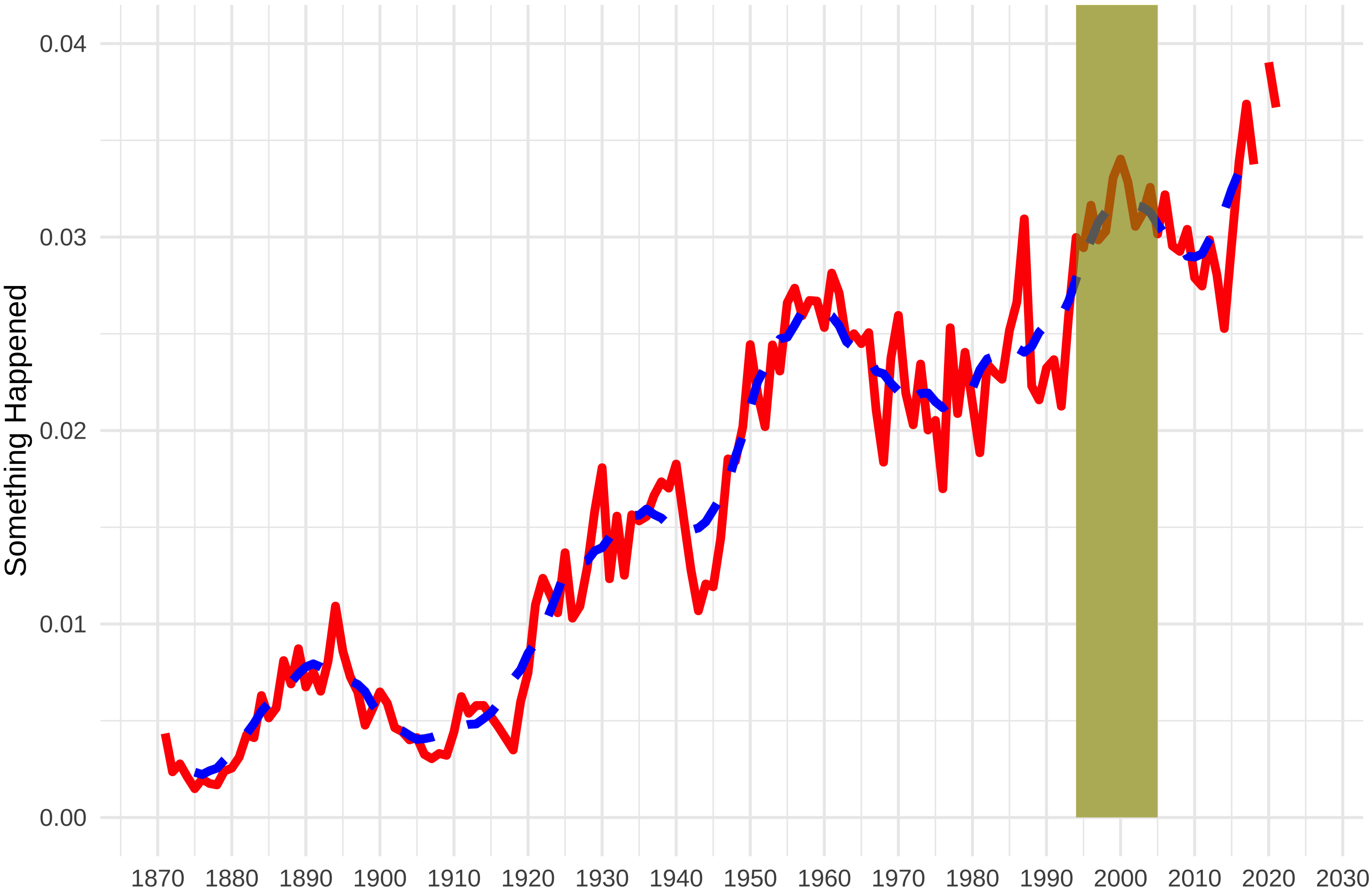

Something Happened 1870 - 2018
Ten Year Moving Average
Something Happened
0.04
0.03
0.02
0.01
0.00
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
2030

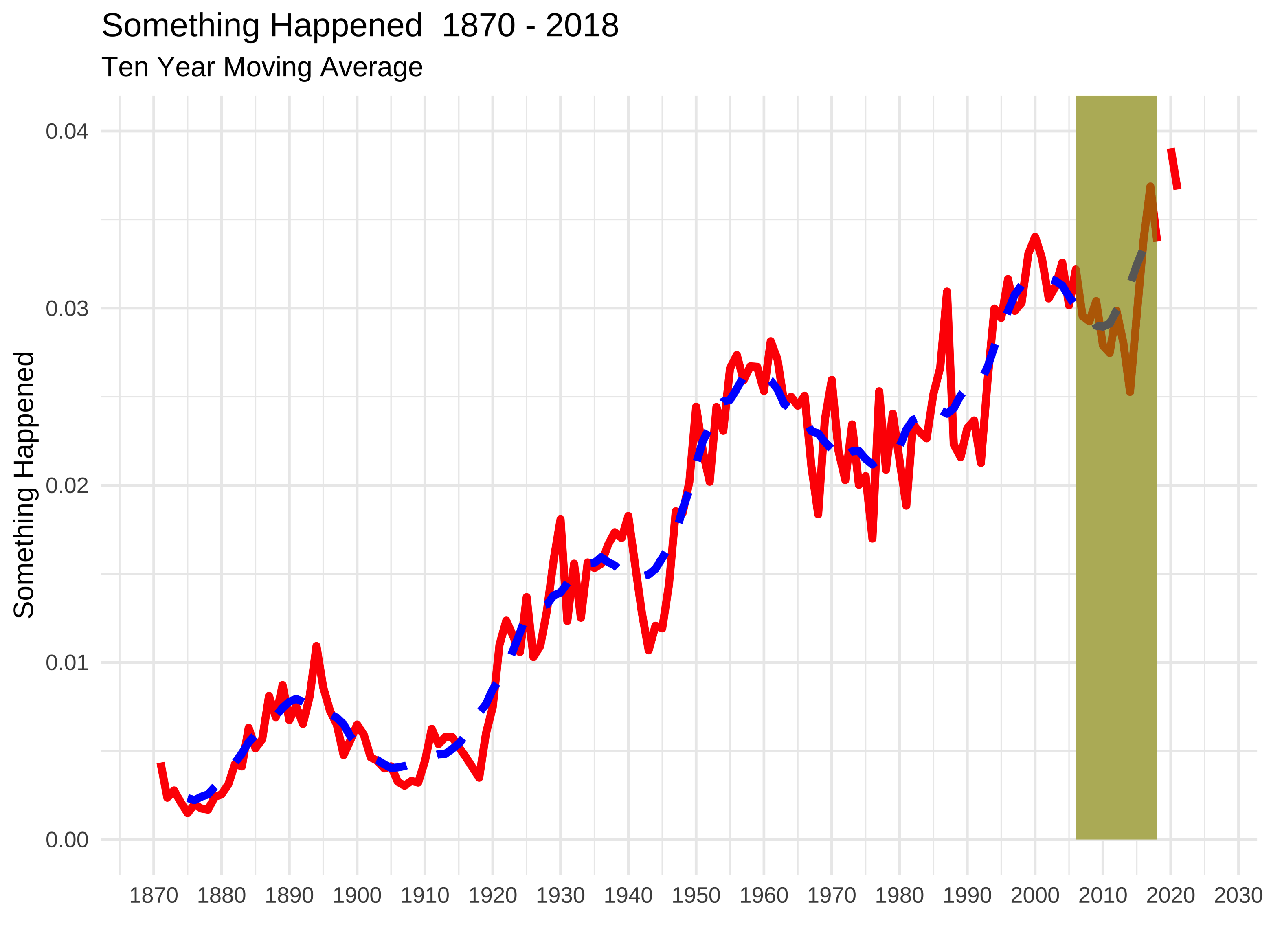

Something Happened 1870 - 2018
Ten Year Moving Average
Something Happened
0.04
0.03
0.02
0.01
0.00
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
2030


Something Happened 1870 - 2018
Ten Year Moving Average
Something Happened
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

Something Happened 1870 - 2018

Ten Year Moving Average




Something Happened 1870 - 2018
Ten Year Moving Average
Something Happened
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

- 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. Game was optimized for home runs.

How Do We Move From Question to Insight to Action?

GOD BL
THIS MESS
I'M NOT GONNA ACCOMPLISH ANYTHING JUST LOOKING AT DATA. IT'S TIME FOR ACTION!

- Data Literacy also involves our collective efforts to actually *make decisions* that are informed by data
- We must learn to communicate results clearly and advocate for policy interventions

Let's Predict Retention

Reminder: This is an example. Be Careful.

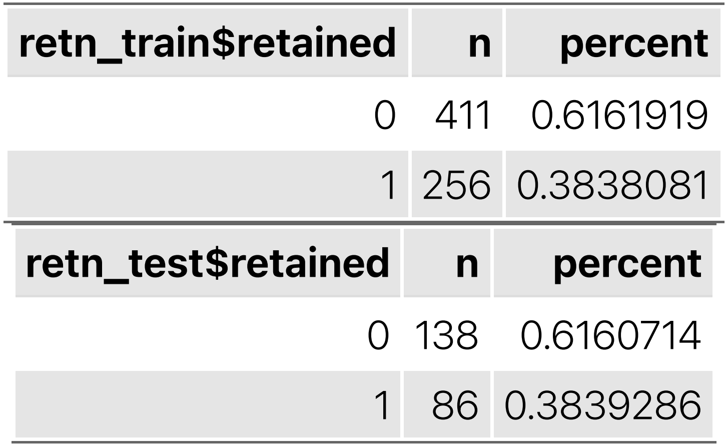
Explore the Data (this is not real student data)

student_id	1	2	3	4	5	6
retained	0	1	1	1	0	0
income_group	Pell Eligible	No Aid	Pell Eligible	No Aid	Pell Eligible	Pell Eligible
sex	male	female	female	female	male	male
age	22	38	26	35	35	NA
siblings_enrolled	1	1	0	1	0	0
peers_from_hs	0	0	0	0	0	0
net_tuition	283	2783	309	2073	314	330
residency	Resident	Non-Resident	Resident	Resident	Resident	International
total_peer_group	1	1	0	1	0	0

Pre-Process the Data

student_id	1	2	3	4	5	6
retained	0	1	1	1	0	0
income_group	Pell Eligible	No Aid	Pell Eligible	No Aid	Pell Eligible	Pell Eligible
sex	male	female	female	female	male	male
age	-0.5300051	0.5714304	-0.2546462	0.3649113	0.3649113	NA
siblings_enrolled	0.4325504	0.4325504	-0.4742788	0.4325504	-0.4742788	-0.4742788
peers_from_hs	-0.4734077	-0.4734077	-0.4734077	-0.4734077	-0.4734077	-0.4734077
net_tuition	-0.5021568	0.7865640	-0.4887541	0.4205673	-0.4861766	-0.4779288
residency	Resident	Non-Resident	Resident	Resident	Resident	International
total_peer_group	1	1	0	1	0	0
income_group_no_aid	0	1	0	1	0	0
income_group_pell_eligible	1	0	1	0	1	1
income_group_state_grant_eligible	0	0	0	0	0	0
sex_female	0	1	1	1	0	0
sex_male	1	0	0	0	1	1
residency_international	0	0	0	0	0	1
residency_non_resident	0	1	0	0	0	0
residency_resident	1	0	1	1	1	0
residency_na	0	0	0	0	0	0

Split Into Training/Test Sets



retn_train$retained	n	percent
0	411	0.6161919
1	256	0.3838081

retn_test$retained	n	percent
0	138	0.6160714
1	86	0.3839286

Build Basic Regression Model

(reminder, this is just a toy model)

```r
mod.1 <- glm(retained ~
               total_peer_group +
               net_tuition +
               sex_female +
               income_group_no_aid,
             data = retn_train,
             family = "binomial")
```

Review and Interpret the Results

term	estimate	std.error	statistic	p.value
(Intercept)	0.175	0.180	-9.665	0.000
total_peer_group	**0.862**	**0.073**	**-2.029**	**0.042**
net_tuition	1.302	0.165	1.596	0.110
sex_female	**14.946**	**0.217**	**12.453**	**0.000**
income_group_no_aid	**3.286**	**0.300**	**3.965**	**0.000**

Interpretation

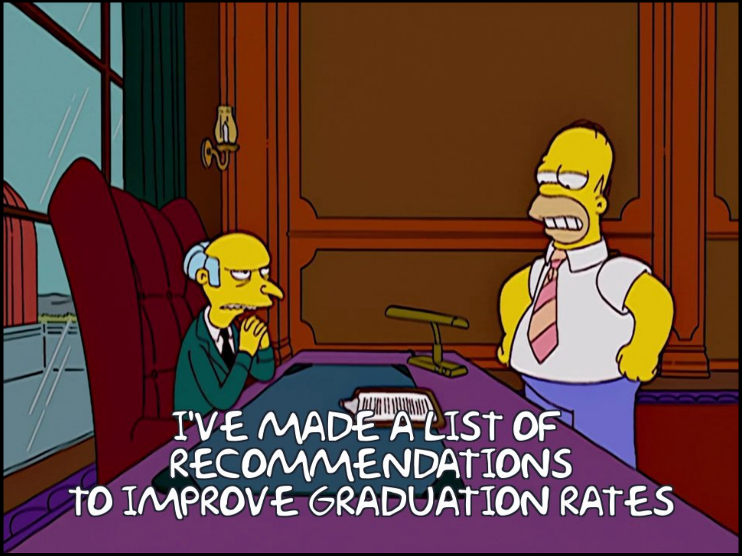
Students in the Income No Aid Group are [INSERT NUMBER HERE] times more likely to retain than those in the baseline group when controlling for other features

Female Students are [INSERT NUMBER HERE] times more likely to retain than those in the baseline group when controlling for other features

Make New Predictions

student_id	predictions	retained	income_group	sex	age	siblings_enrolled	peers_from_hs	net_tuition	residency	total_peer_group
1	**0.12**	0	Pell Eligible	male	-0.53000510	0.4325504	-0.4734077	-0.50215678	Resident	1
17	**0.08**	0	Pell Eligible	male	-1.90679949	3.1530382	0.7671990	-0.06192976	International	5
21	**0.14**	0	State Grant Eligible	male	0.36491125	-0.4742788	-0.4734077	-0.12481933	Resident	0
28	**0.48**	0	No Aid	male	-0.73652426	2.2462089	2.0078057	4.64447857	Resident	5
35	**0.39**	0	No Aid	male	-0.11696678	0.4325504	-0.4734077	1.00564654	Non-Resident	1
45	**0.70**	1	Pell Eligible	female	-0.73652426	-0.4742788	-0.4734077	-0.48926957	International	0
46	**0.13**	0	Pell Eligible	male	NA	-0.4742788	-0.4734077	-0.48617664	Resident	0
47	**0.12**	0	Pell Eligible	male	NA	0.4325504	-0.4734077	-0.33616954	International	1
51	**0.08**	0	Pell Eligible	male	-1.56260089	3.1530382	0.7671990	0.15045143	Resident	5
56	**0.37**	1	No Aid	male	NA	-0.4742788	-0.4734077	0.06642683	Resident	0
61	**0.13**	0	Pell Eligible	male	-0.53000510	-0.4742788	-0.4734077	-0.50267227	Non-Resident	0
63	**0.39**	0	No Aid	male	1.05330845	0.4325504	-0.4734077	1.03193645	Resident	1
70	**0.10**	0	Pell Eligible	male	-0.25464622	1.3393797	-0.4734077	-0.47380492	Resident	2
73	**0.18**	0	State Grant Eligible	male	-0.59884482	-0.4742788	-0.4734077	0.83089601	Resident	0
76	**0.13**	0	Pell Eligible	male	-0.32348594	-0.4742788	-0.4734077	-0.49390897	Resident	0
83	**0.70**	1	Pell Eligible	female	NA	-0.4742788	-0.4734077	-0.49133153	International	0
87	**0.09**	0	Pell Eligible	male	-0.94304341	0.4325504	3.2484124	0.04374535	Resident	4
88	**0.13**	0	Pell Eligible	male	NA	-0.4742788	-0.4734077	-0.48617664	Resident	0
90	**0.13**	0	Pell Eligible	male	-0.39232566	-0.4742788	-0.4734077	-0.48617664	Resident	0
91	**0.13**	0	Pell Eligible	male	-0.04812706	-0.4742788	-0.4734077	-0.48617664	Resident	0

Based on the Insights from this Analyis, What Recommendations Would We Make?

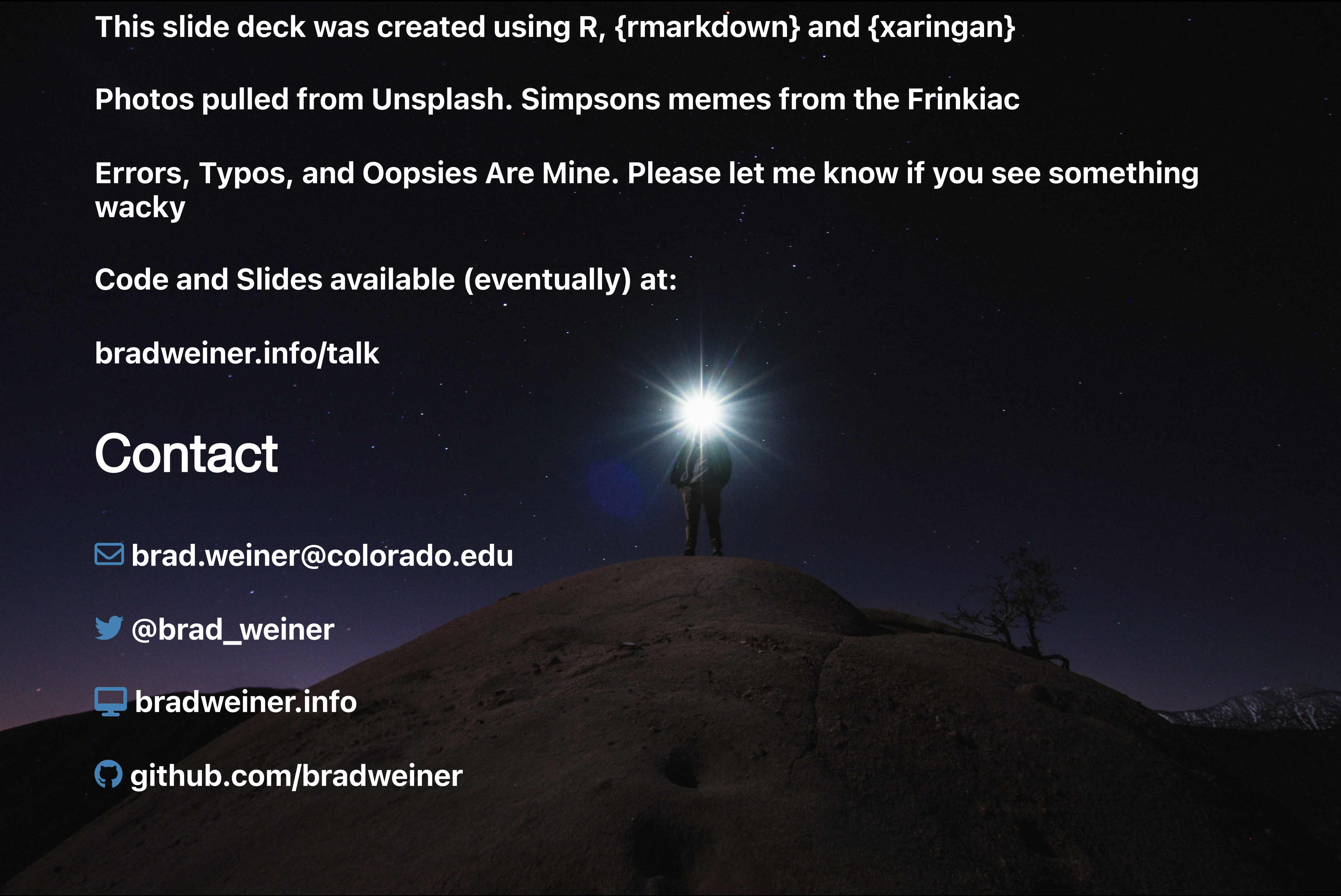

I'VE MADE A LIST OF RECOMMENDATIONS TO IMPROVE GRADUATION RATES

- Is it feasible?
- Is it measureable?
- Is it aligned with the insights?
- Do we require additional information or analysis?

How do we enhance Data Literacy Across Campus?

- Open Discussion

literacy is data enablement
Data literacy is data enablement
Data literacy is data enablement
Data literacy is data enablement
Data literacy is data enablement
Data literacy is data enablem
Data literacy is data enablem
Data literacy is data enabler
Data literacy is data enabl
Data literacy is data enable
Data literacy is data enablement



This slide deck was created using R, {rmarkdown} and {xaringan}

Photos pulled from Unsplash. Simpsons memes from the Frinkiac

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

Code and Slides available (eventually) at:

bradweiner.info/talk

Contact

brad.weiner@colorado.edu

@brad_weiner

bradweiner.info

github.com/bradweiner