Applications of Data & Machine Learning in Economic Research: Part I - Policymaking

BAI 30545 – Foundations of Economic Sciences

Julian Streyczek (Bocconi)

Introduction

Hi

- My name is Julian
- **PhD student in Economics** at Bocconi
- **4+ years** of academic research experience @ Mannheim, Bocconi, Harvard
- Use **statistical methods** to derive **insights from data**
- **Example papers**:

  1) **Paywalls** on **newspaper websites** in the US reduced **attention to politics**, reducing both political **knowledge** and **turnout** in elections

  2) **Twitter** affected the **productivity** and shifted the **research topics** of **economists**
- **Website:** https://julianstreyczek.github.io

General Notes

- **4 lessons** on applications of data & machine learning in economic research
- **Goal:** Give you a *general idea* how to use data to address current challenges

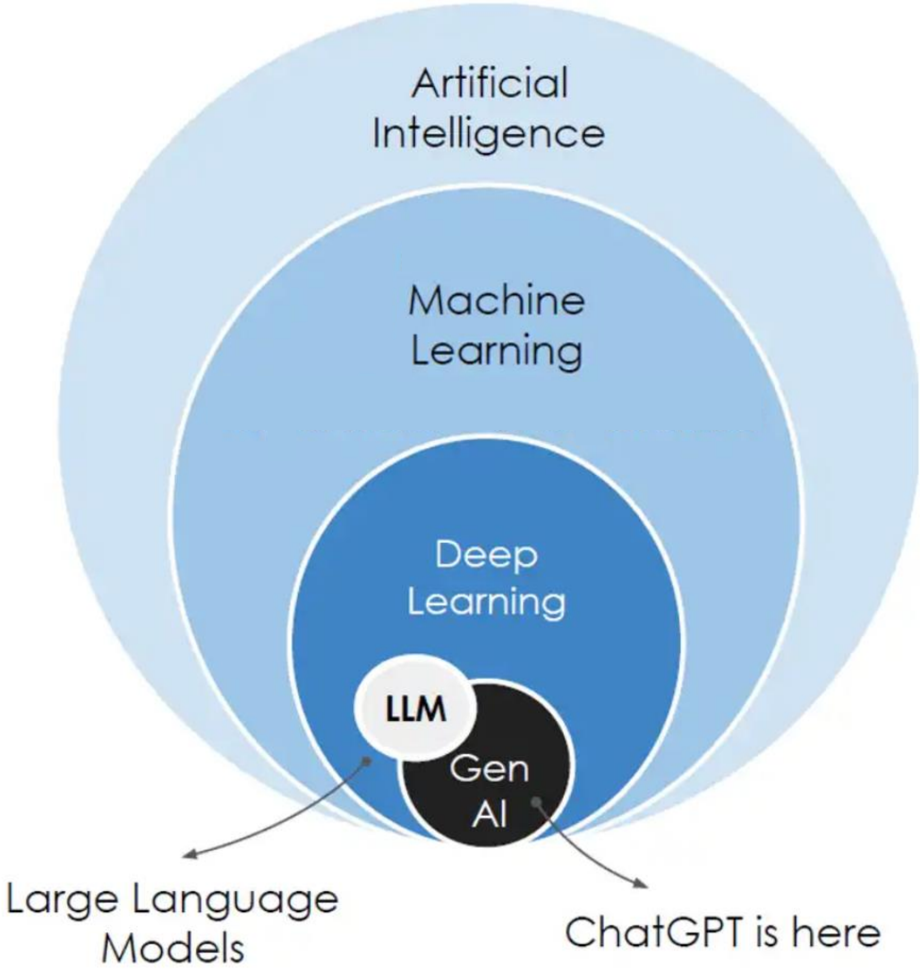
Date	Time	Room*
Wednesday, 17/09	14:45 – 16:15	Aula D
Wednesday, 24/09	14:45 – 16:15	Aula 11
Wednesday, 01/10	14:45 – 16:15	Aula D
Wednesday, 08/10	14:45 – 16:15	Aula 4

* check Blackboard for short-term changes

- **Slides** will be uploaded to Blackboard
- **Exams:** The **big picture** is relevant:
  You should be familiar with the overall topics and ideas of these lessons,
  but no need to study individual slides

Quick recap on Machine Learning

“Machine Learning”



M. Malandruccolo: *“Harnessing the Power of Machine Learning and AI to Decrease Health Expenditures Worldwide – A Review”*. Premier Journal of Science 2024:1;100004

Machine Learning

- Term coined by computer scientist Arthur Samuel in 1959
- Extracting relationships from data that were not explicitly programmed
- What is the „Machine"?
  - Algorithm / recipe / number of steps to be followed in sequence
- What is the „Learning"?
  - Insight / result that the „machine" finds on its own

"Machine Learning"

Simple example: Galton (1907) and the Bull

- Estimating the weight of a bull at the *West of England Fat Stock and Poultry Exhibition* at Plymouth:

> *„A fat ox having been selected, competitors bought stamped and numbered cards for 6 pennies each, on which to inscribe their respective names, addresses, and estimates of what the ox would weigh after it had been slaughtered […]. Thorse who guessed most successfully received prizes."*

William Henry Davis and Charles Joseph Hullmandel, *Portrait of T. W. Coke and North Devon Ox*, c. 1837. The Royal Smithfield Club Collection. University of Reading.

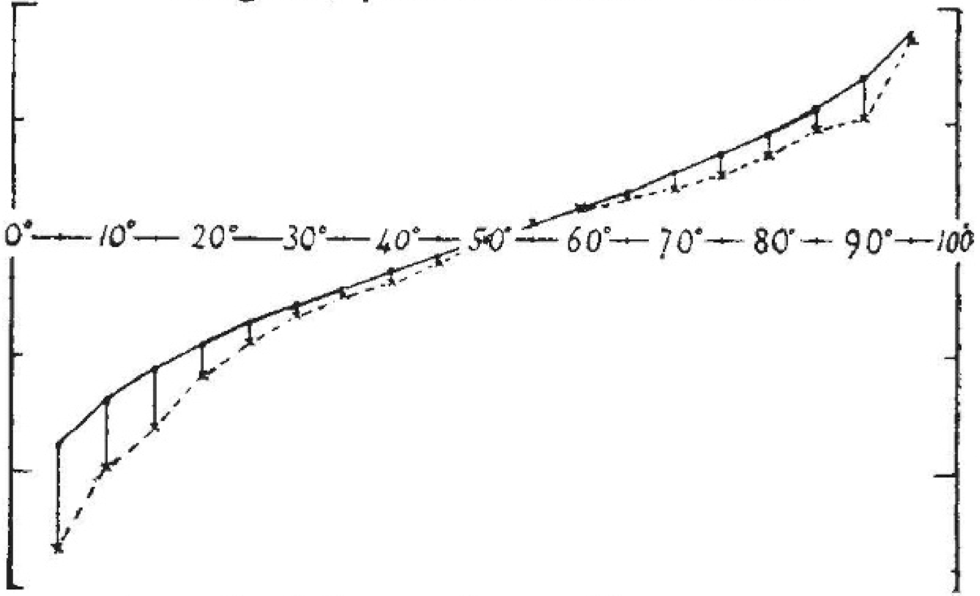
GALTON, F.: „*Vox Populi*". Nature 75, 450–451 (1907)

Simple example: Galton (1907) and the Bull

- **Images:** List of guesses and distance from the true weight of 787 cards (in buckets), as table (left) and diagram (right)
- Median (middle) estimate: **1207 lbs.**
- True weight: **1198 lbs.** => error <0.8% !

Degrees of the length of Array 0°—100°	Estimates in lbs.	Observed deviates from 1207 lbs.
5°	1074	− 133
10	1109	− 98
15	1126	− 81
20	1148	− 59
q_1 25	1162	− 45
30	1174	− 33
35	1181	− 26
40	1188	− 19
45	1197	− 10
m 50	1207	0
55	1214	+ 7
60	1219	+ 12
65	1225	+ 18
70	1230	+ 23
q_3 75	1236	+ 29
80	1243	+ 36
85	1254	+ 47
90	1267	+ 52
95	1293	+ 86

Diagram, from the tabular values.



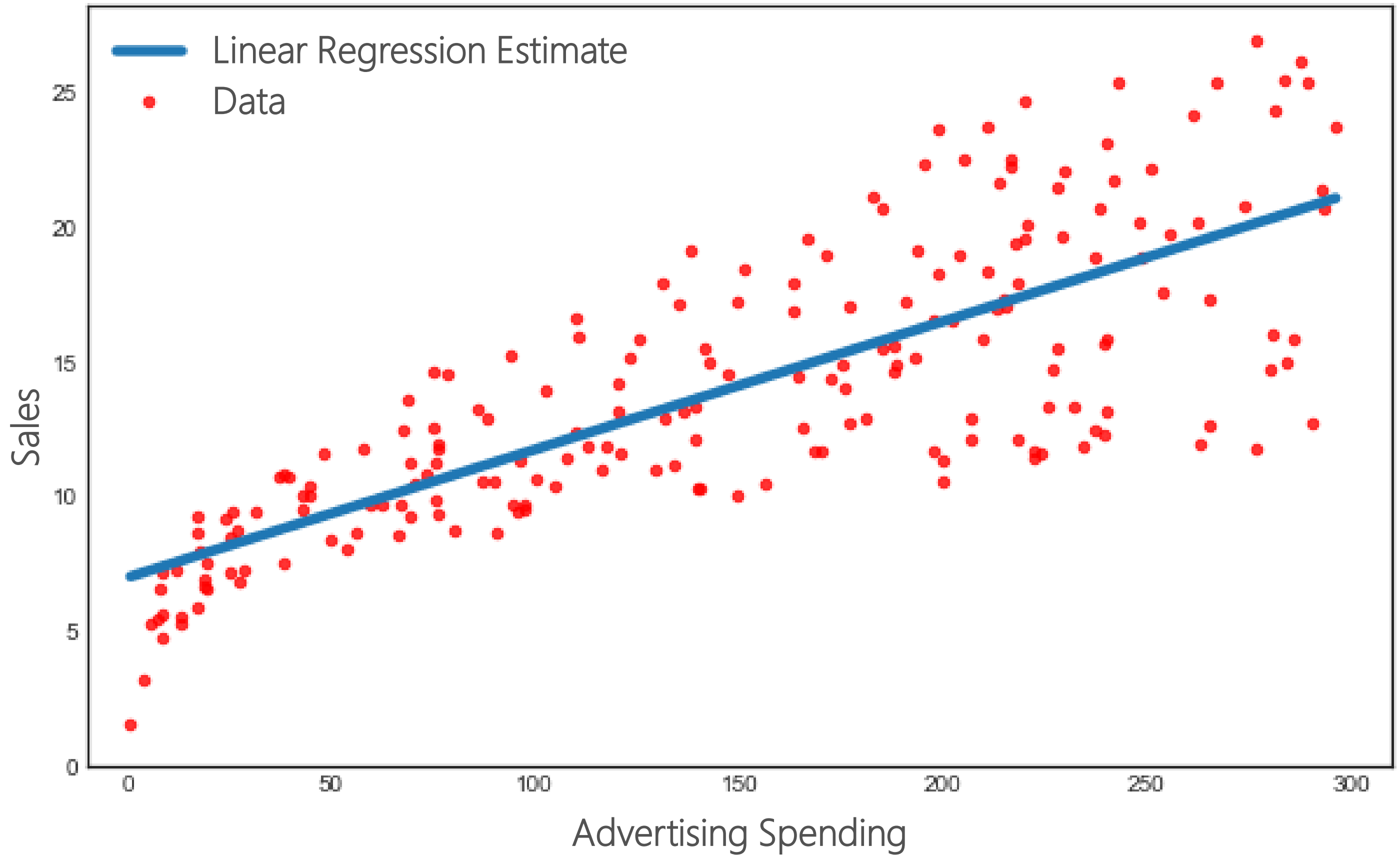
GALTON, F.: „*Vox Populi*". Nature 75, 450–451 (1907)

Simple example: Galton (1907) and the Bull

- Take-away: In next year's fair, we can probably determine the weight of a bull as follows:
  1) Collect data: Get guesses from attendees
  2) Compute median
- Relation to Machine Learning:
  - „Machine“: Process of computing the median guess
  - „Learning“: The result tells us something *new* about the world that we did not explicitly tell it: The bull's weight (roughly)
- Philosophy: Applying a „dumb“ algorithm in a smart way, we learn something new

Evolution of “Machine Learning”

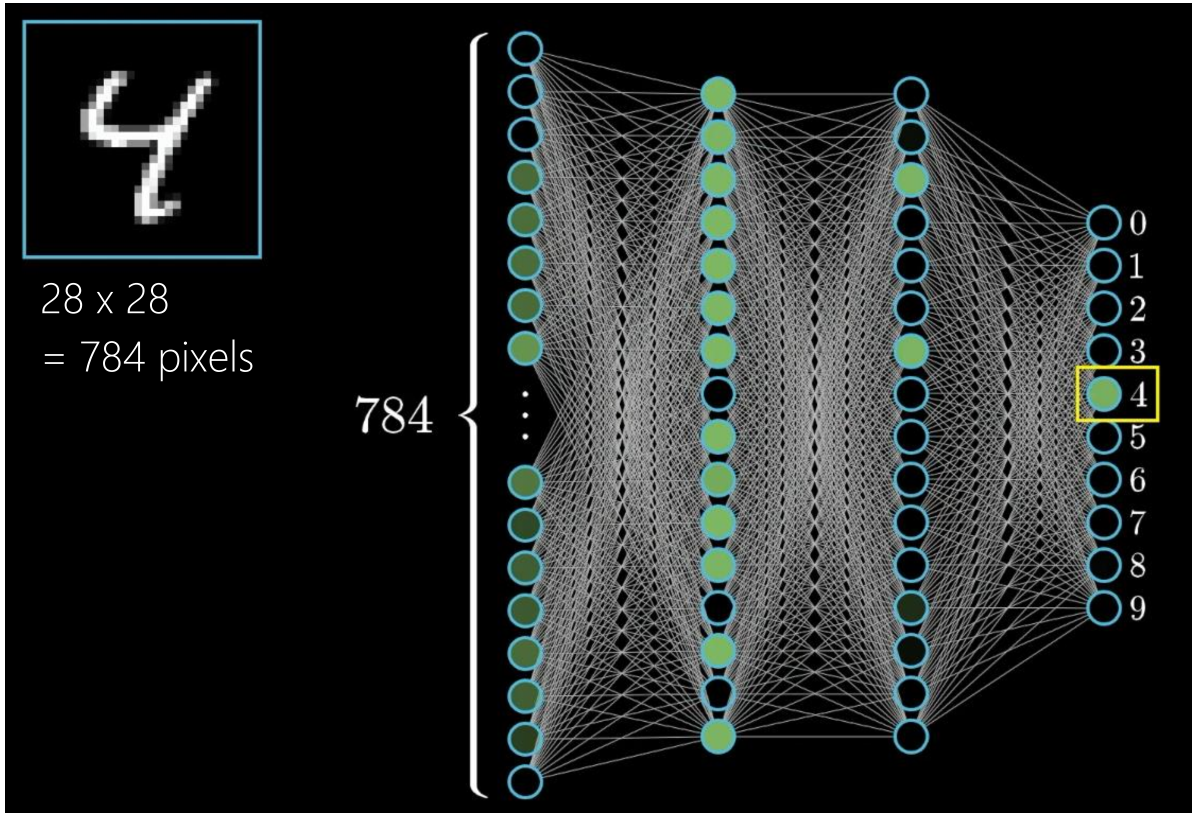
Linear Regression: Learn a linear relationship between variables



Matteo Courthoud: https://matteocourthoud.github.io/course/ml-econ/01_regression, last accessed 15/09/2025

Evolution of "Machine Learning"

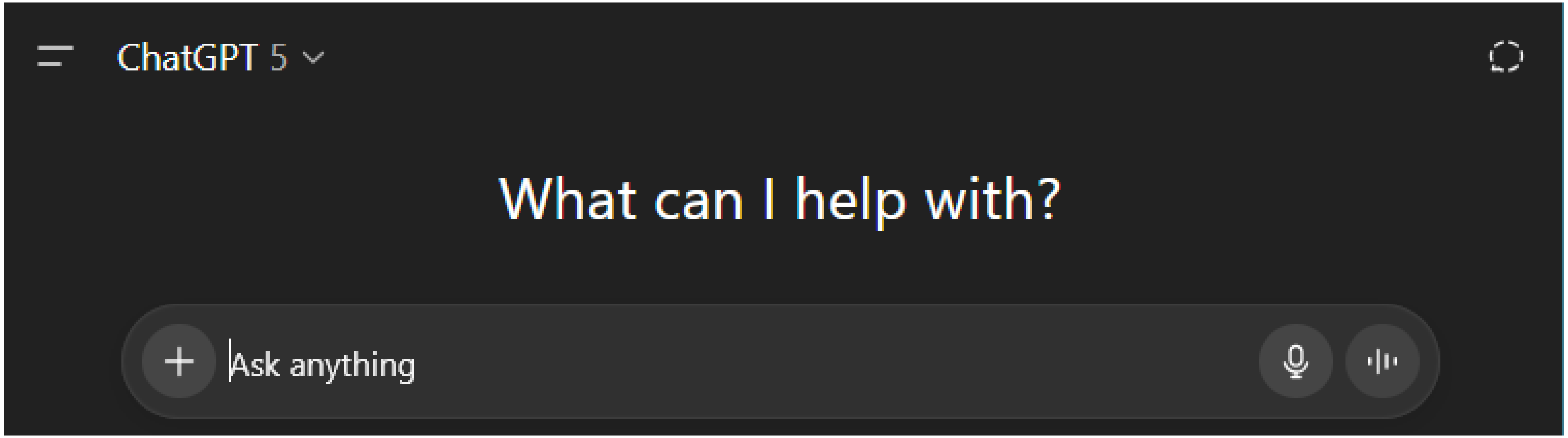
Neural Network: Learn number from pixels



3Blue1Brown: https://www.3blue1brown.com/lessons/neural-networks, last accessed 15/09/2025

Evolution of “Machine Learning”

- **Generative AI:** Learn most likely next word(s) given prompt

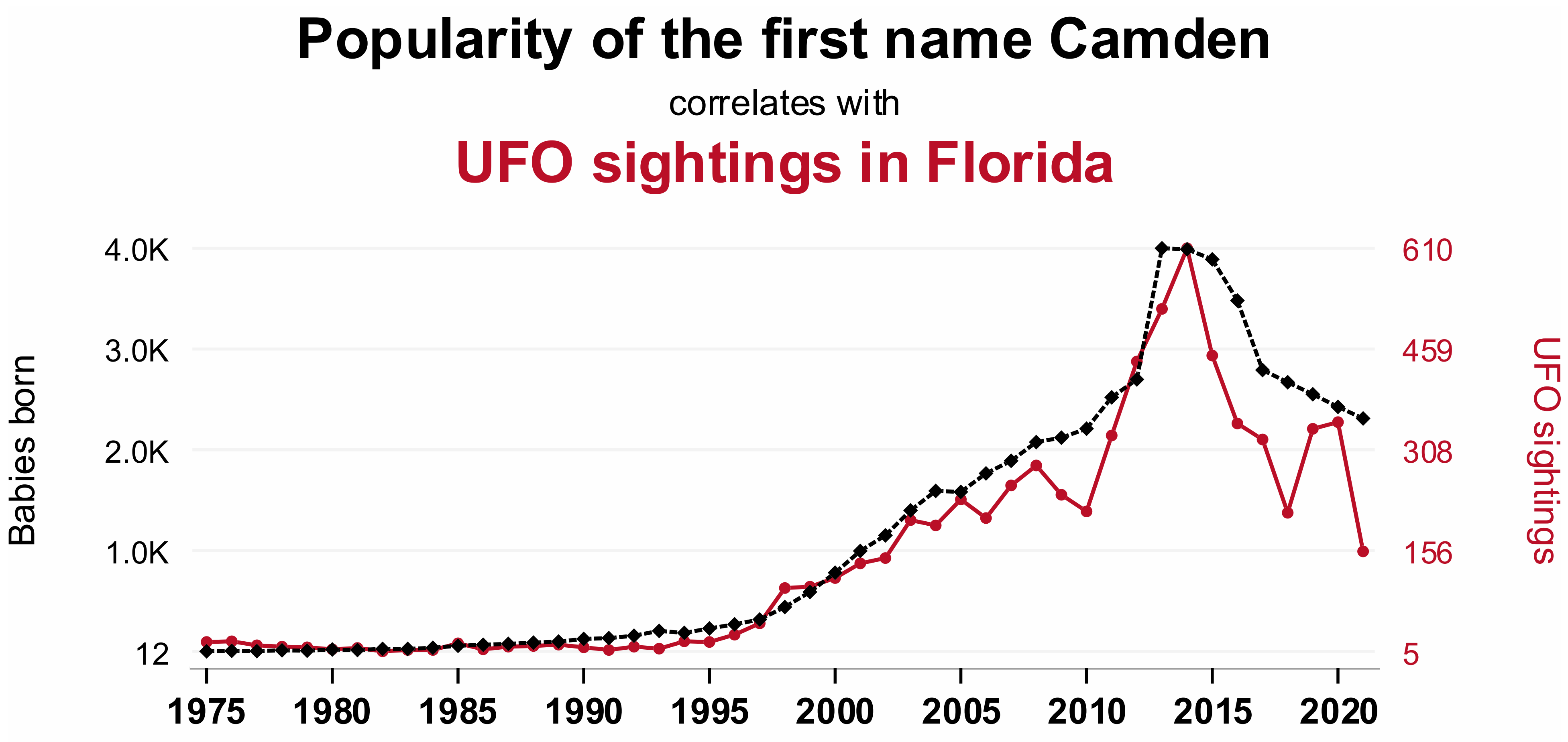


Prediction vs. Causality

- **Machine Learning:** Classically, goal is to capture **correlation**:
  - If A predicts B very well, we are happy
  - We don't care why
  - Example: Amount of ice cream sold predicts number of sun burns
- **Causal inference:** Goal is to capture **causation:**
  - „Does A cause B?" ⇔ „If not A, then not B"
  - We really care about the causal connection
  - Example: Amount of UV light intensity causes number of sun burns
- Traditionally, economic research is all about causal inference („econometrics"), but is increasingly combined with machine learning methods
- These lessons: More about prediction, but in economics settings

"Machine Learning"

Remember: Correlation ≠ Causation



Spurious Correlations: https://www.tylervigen.com/spurious/correlation/3011_popularity-of-the-first-name-camden_correlates-with_ufo-sightings-in-florida, last accessed 14/09/2025

Application:
Predicting Health Violations

Application: Health Violations

Research Paper: Kang et al. (2013)

Where *Not* to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews

Jun Seok Kang† **Polina Kuznetsova†** **Michael Luca‡** **Yejin Choi†**

†Department of Computer Science
Stony Brook University
Stony Brook, NY 11794-4400
{junkang,pkuznetsova,ychoi}
@cs.stonybrook.edu

‡Harvard Business School
Soldiers Field Road
Boston, MA 02163
mluca@hbs.edu

Jun Seok Kang, Polina Kuznetsova, Michael Luca, and Yejin Choi. 2013. Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1443–1448, Seattle, Washington, USA. Association for Computational Linguistics.

Motivation

- **Foodborne diseases** affect **1 in 6 Americans** (48 million) each year
  - 128,000 hospitalizations
  - 3,000 deaths
- Estimated costs: **$17.6bn** per year
- "More than **half of all foodborne illness outbreaks** in the United States are associated with **restaurants, delis, banquet facilities, schools, and other institutions**"

Center for Disease Control: https://www.cdc.gov/restaurant-food-safety/about/index.html, last accessed 15/09/2025

Motivation

Scenario:

- You are leading **Seattle**'s Department of Public Health
- **Goal:** Reduce health violations in restaurants
- **Problem:** Funding cuts – no money for additional health inspectors
- Current practice: Inspect restaurants randomly
- **Can we do better?**
  For example, inspect only restaurants with **highest *risk* of health violation**?

https://commons.wikimedia.org/wiki/File:Seattle_on_the_map_of_the_United_States.png, last accessed 14/09/2025

Idea

- Let's build a **machine learning model** to predict which types of restaurants are most likely to violate health codes
- First requirement: **DATA**
  - **„Outcomes":** Results of previous restaurant health inspections => *Department of Health*
  - **„Features":** Restaurant characteristics => *Yelp (rating platform for restaurants)*
- **Idea**: Customers observe health conditions. Bad conditions likely affect customers' reviews
- **Strategy**: Find out which *features* are associated with verified health violations
  - E.g., health violations may be associated with reviews that include words like „**gross**", „**dirty**", „**smelly**", etc.
  - In this case, we would say that these words *predict* (indicate) health violations

Data: Health Violations

Showing 1-25 of 12501

(206) 722-6400

HISTORY

100th Ave Cakes
15364 Ne 96th Pl
Redmond, WA 98052
(758) 753-4760

EXCELLENT
HISTORY

108 Vietnamese Authentic Cuisine
18114 E Valley Hwy
Kent, WA 98032
(425) 251-8803

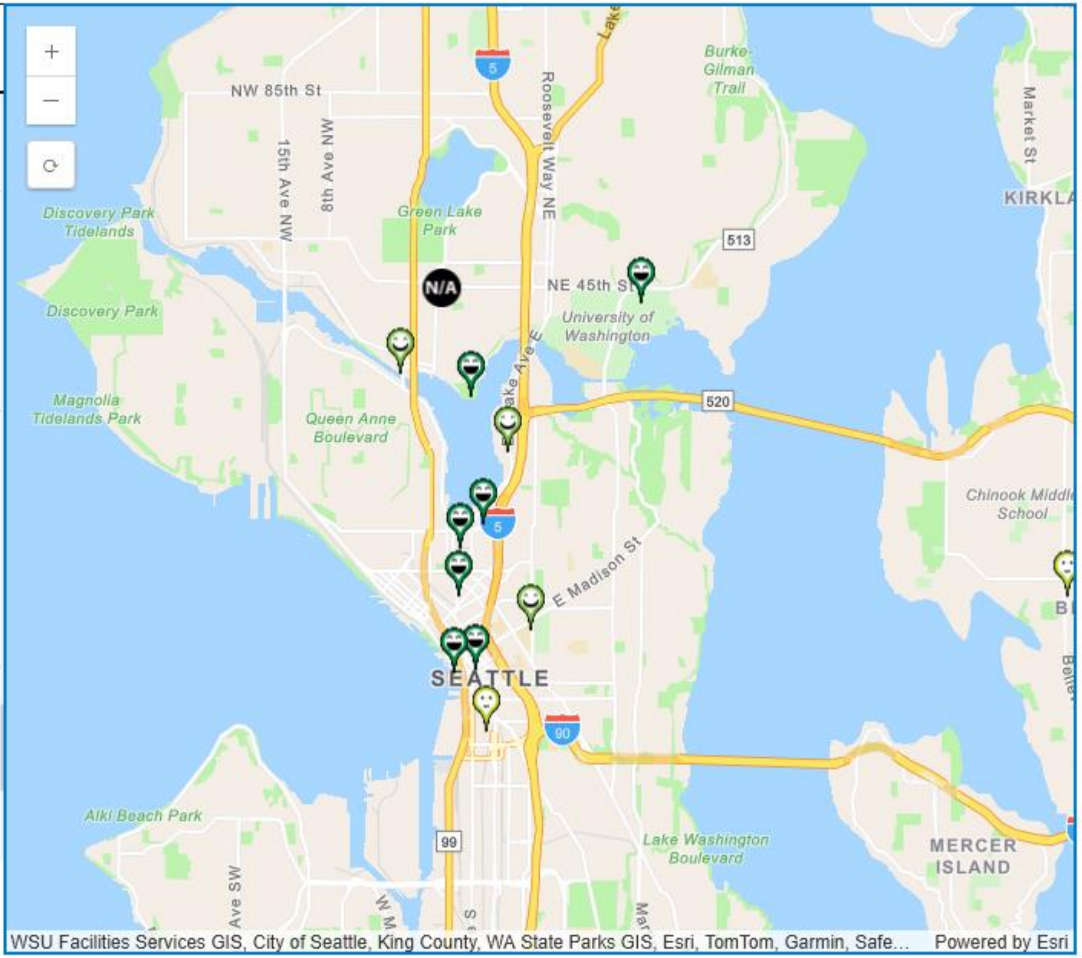
OKAY
HISTORY

19 Gold
3601 Fremont Ave N Ste 101
Seattle, WA 98103
(206) 883-6870

GOOD
HISTORY

Showing 1-25 of 12501

Previous Next



https://kingcounty.gov/en/dept/dph/health-safety/food-safety/search-restaurant-safety-ratings#/, last accessed 15/09/2025

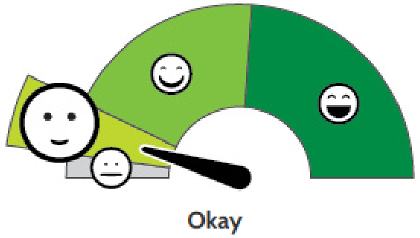
Application: Health Violations

Data: Health Violations

- **Score:** Lower = Better
- Small violations are not a major health concern (e.g., size of plumbing)
- This paper:
  Health Violation if **Score > 50**

108 VIETNAMESE AUTHENTIC CUISINE

108 Vietnamese Authentic Cuisine
18114 E Valley Hwy
Kent, WA 98032
(425) 251-8803



Seating 51-150 - Risk Category III

DIRECTIONS

The rating is based on the average of high risk violations from the last 4 routine inspections.

Learn more about the rating system

Date	INSPECTION TYPE	Score
06/04/2025	Routine Inspection	22

2120 Proper cold holding temperatures 42 F to 45 F (5 points)

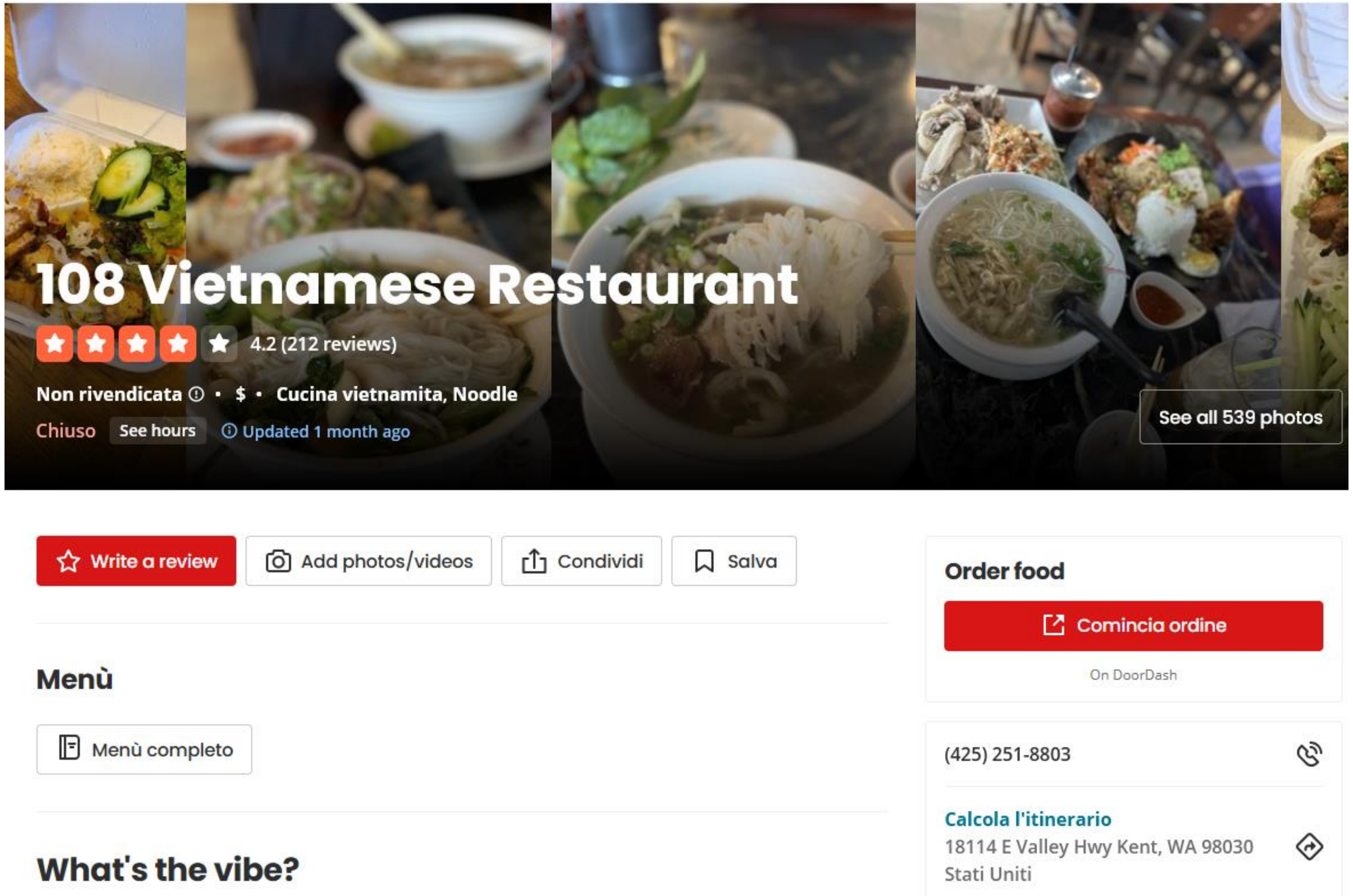
1100 Proper disposition of returned previously served unsafe or contaminated food Date marking (10 points)

4400 Plumbing properly sized installed (5 points)

https://kingcounty.gov/en/dept/dph/health-safety/food-safety/search-restaurant-safety-ratings#/, last accessed 15/09/2025

Data: Yelp reviews

- 152k Yelp reviews
  for 1,756 restaurants in Seattle
  between 2006 and 2013



https://www.yelp.it/biz/108-vietnamese-restaurant-kent-2?osq=vietnames#location-and-hours, last accessed 15/09/2025

Application: Health Violations

Data: Yelp reviews

Alan L.
Seattle, Stati Uniti
0 3 3

14 feb 2025

I went with my family to enjoy a Valentine's day dinner. Owner was very friendly, made an exquisite Sour Catfish soup. Clay pot fish was tasty. Owner gave mango and beef on the house. Very good service and home cooking vibe was relaxing.

Desiree C.
Renton, Stati Uniti
118 4 0

8 mag 2016

Not for me. Egg rolls have a strong flavor its gross. Pho was tasty meat was too tough to chew.

https://www.yelp.it/biz/108-vietnamese-restaurant-kent-2?osq=vietnames#location-and-hours, last accessed 15/09/2025

Application: Health Violations

Which restaurant features could predict health violations?

- Cuisine
  - Kebab places on average less sanitary than fine dining?
- Number of reviews
- Average Rating
  - Violations lead to low rating?
- ZIP code
  - Certain areas attract unsanitary restaurants?
- Violation detected in past
  - Violate once, violate again?
- Texts
  - Customers explicitly mention about violations in textual reviews?

Application: Health Violations

Using text as data

- To use **text** for data analysis, we need to transform it into **numbers**
- **Simple approach:** Identify **words** that signal **disgust**, each word becomes a **feature** that **counts** how often the word appears in reviews
- **Examples:** gross, disgusting, mess, sticky, smell, restroom, dirty, filthy

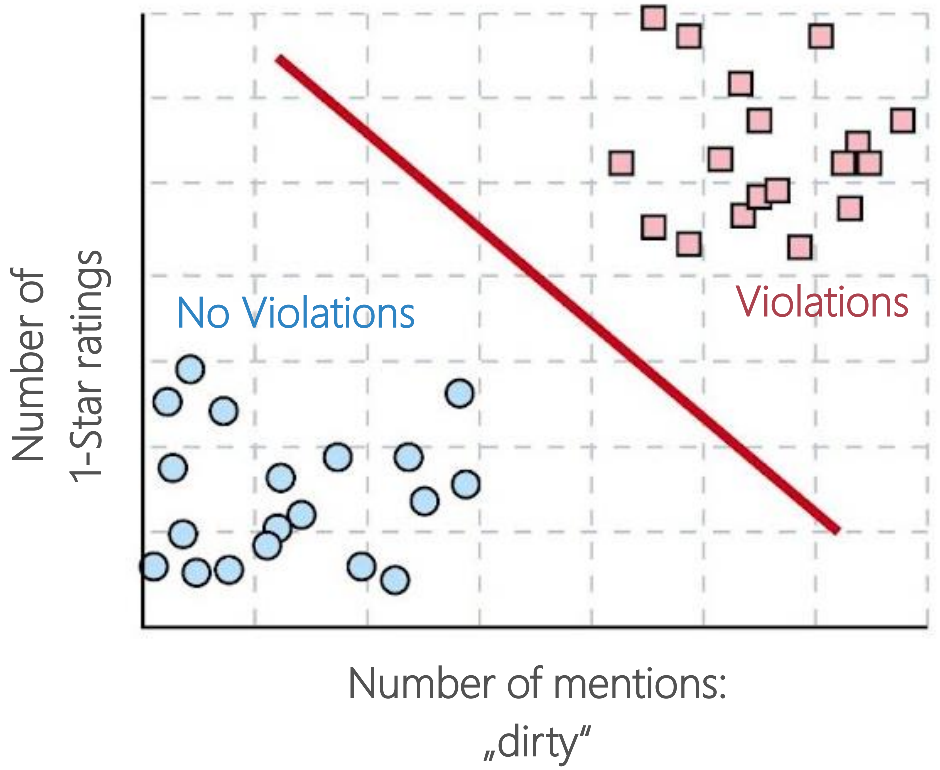
Steuer, F. (2018). *Machine learning for public policy making: How to use data-driven predictive modeling for the social good* (Erasmus Mundus Master's thesis). Erasmus University / IBEI

Application: Health Violations

Prediction algorithm

- **Support Vector Machine** (SVM)
- **What it does:** Classifies data into **yes/no** based on information
- How? Try to find a **„line"** that separates „Yes" from „No" outcomes

(Made-up example)



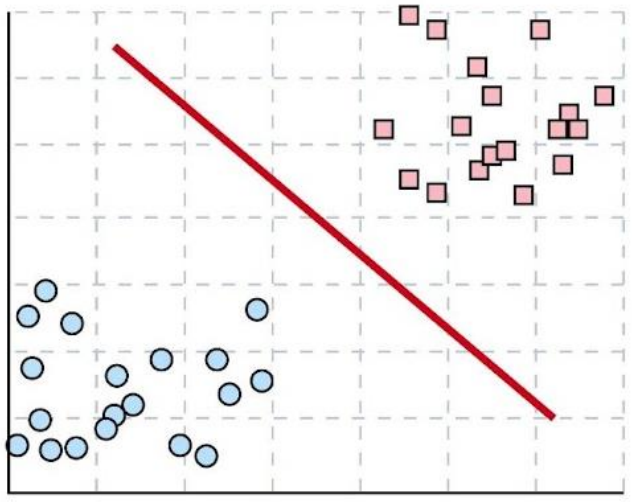
Mohit Uniyal: https://www.appliedaicourse.com/blog/support-vector-machine-algorithm/, last accessed 15/09/2025

Application: Health Violations

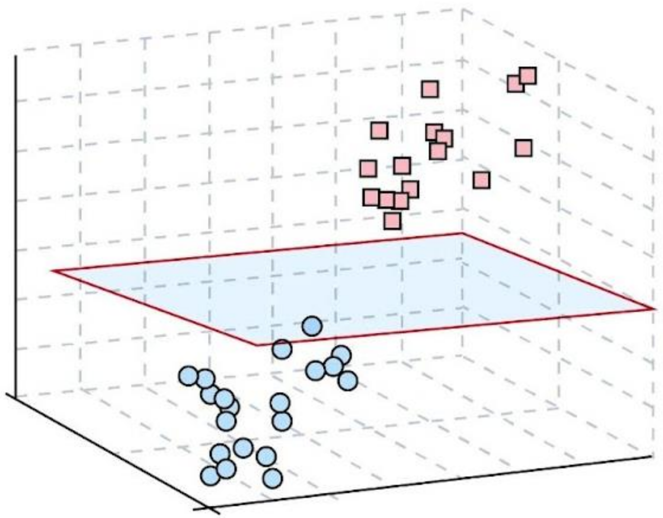
Prediction algorithm

SVM works the same for 2, 3, or thousands of features

2 features



3 features



1000s of features

…

Mohit Uniyal: https://www.appliedaicourse.com/blog/support-vector-machine-algorithm/, last accessed 15/09/2025

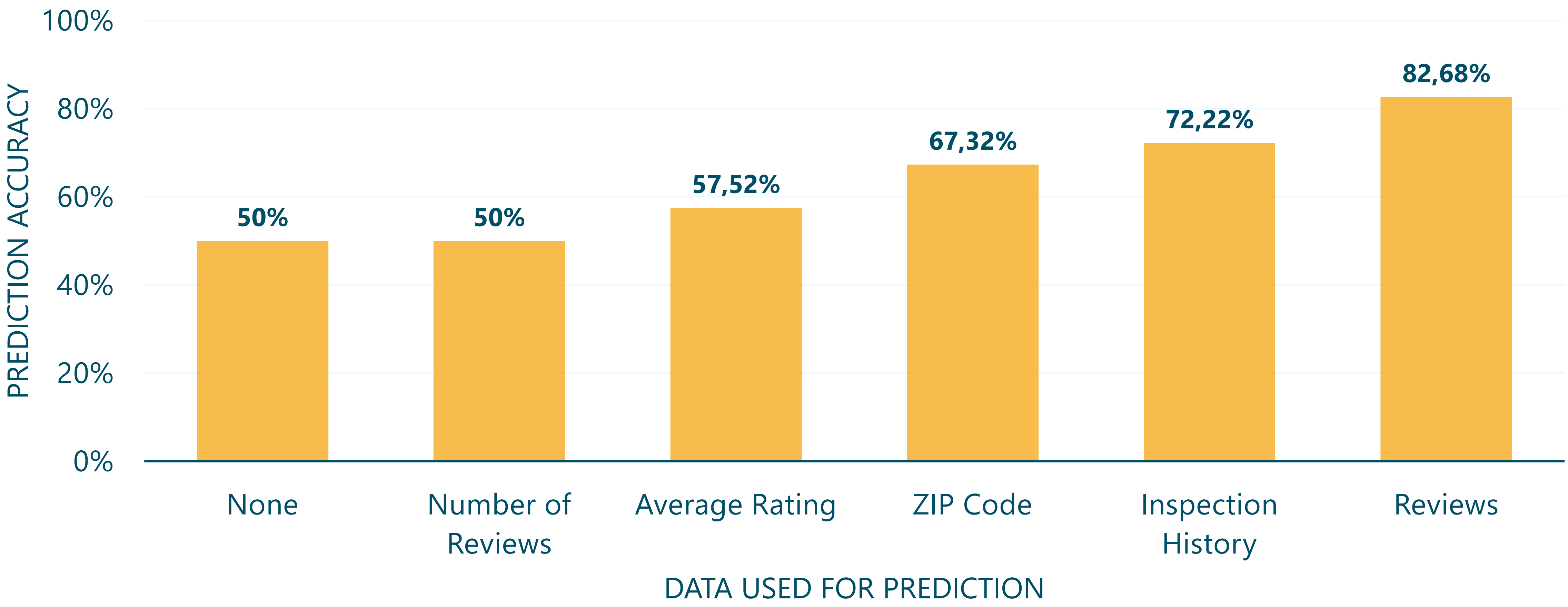
Goal: Accuracy

- What are we optimizing? *Accuracy.*
- Accuracy = Share of restaurants for which health status is predicted correctly
- Example:
  - 100 restaurants - 10 „dirty“ (violate health codes), 90 „clean“ (do not violate)
  - Assume the following classification:
    - Among the 10 dirty, we classify 5 correctly (dirty) and 5 incorrectly (clean)
    - Among the 90 clean, we classify 80 correctly (clean) and 10 incorrectly (dirty)
    - Accuracy = (80 + 5) / 100 = 85%
- Higher Accuracy = Better Prediction

Application: Health Violations

Results

- Accuracy based on which features we use for prediction:



Interpretation of Results

Take-aways:

- „None“: The uninformative („dumb“) model always has accuracy of 50% => just guessing
- „Number of Reviews“: The variable does not help to predict the outcome *at all*
- „Average Rating“: Not very informative
  - => Seems like negative ratings are associated with many other things than health violations (e.g., unfriendly service, bad ambience, long waiting times, disappointing food, etc.)
- „ZIP Code“: Getting better! Seems like „dirty“ restaurants tend to be in similar locations
- „Inspection History“: Past health code violations predict future health code violations
- „Reviews“: Textual information in reviews predict violations *better than any other variable*

=> A model based only on Yelp reviews correctly predicts health code violations in ~83% of cases!

What to do with this knowledge?

- Based on Yelp reviews, we can correctly predict health code violations with 83% accuracy
  => Even without sending a health inspector, we can say with relatively high confidence whether the health code is violated

- If our goal is to shut down dirty restaurants:
  Send our health inspectors to the predicted „dirty“ restaurants
  => Same number of inspectors, more dirty restaurants shut-down

-

  Everyone wins (almost)

Application: Health Violations

Limitations

Our model is a **snapshot** of current circumstances.
It may become less useful if:

- Department of Health changes classification system
- Restaurants become more successful in deleting negative reviews
- Customers become more relaxed or pickier
- Customers use different language to describe health status (long-term)
- ...

Potential remedy

Keep performing a certian number of **random inspections**.

Two advantages:

1) Catch „dirty“ restaurants that our prediction may **systematically** miss
   - E.g., restaurants that are good at soliciting fake / too-favorable reviews
2) Keep generating **new data** that is **not affected by our initial predictions**, and can be used to **update our prediction model** in the future