What are the Most Important Statistical Ideas of the Past 50 Years?

Andrew Gelman and Aki Vehtari

Faryal Fodderwala

November 18, 2024

Introduction

- Overview of 8 significant statistical ideas from 1970 to 2021.
- Authors: Andrew Gelman and Aki Vehtari.
- Purpose: To provoke thought and discussion about modern statistical innovations and their impact on data science.

Authors' Background

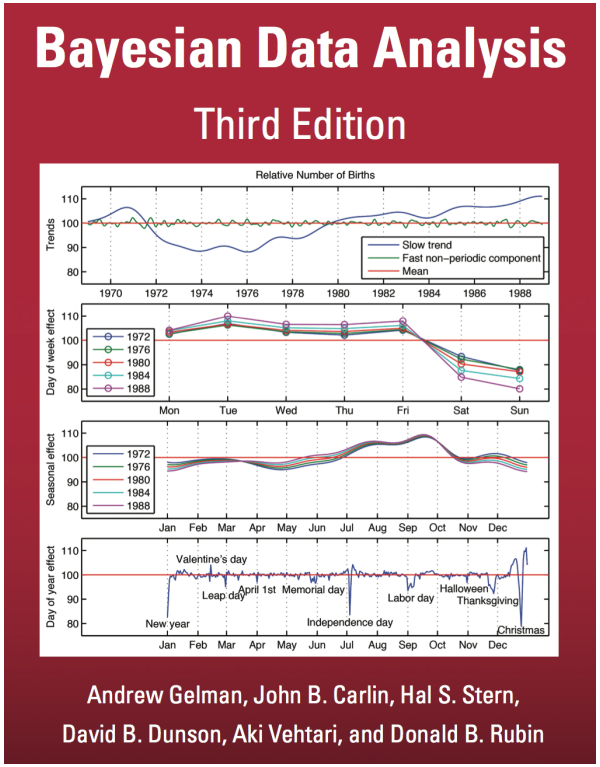
Andrew Gelman

- **Andrew Gelman**:
  - Professor of Statistics and Political Science, Columbia University.
  - Renowned for Bayesian statistics and multilevel modeling.

Aki Vehtari

- **Aki Vehtari**:
  - Professor of Computational Probabilistic Modeling, Aalto University.
  - Focused on Bayesian computation and model assessment.

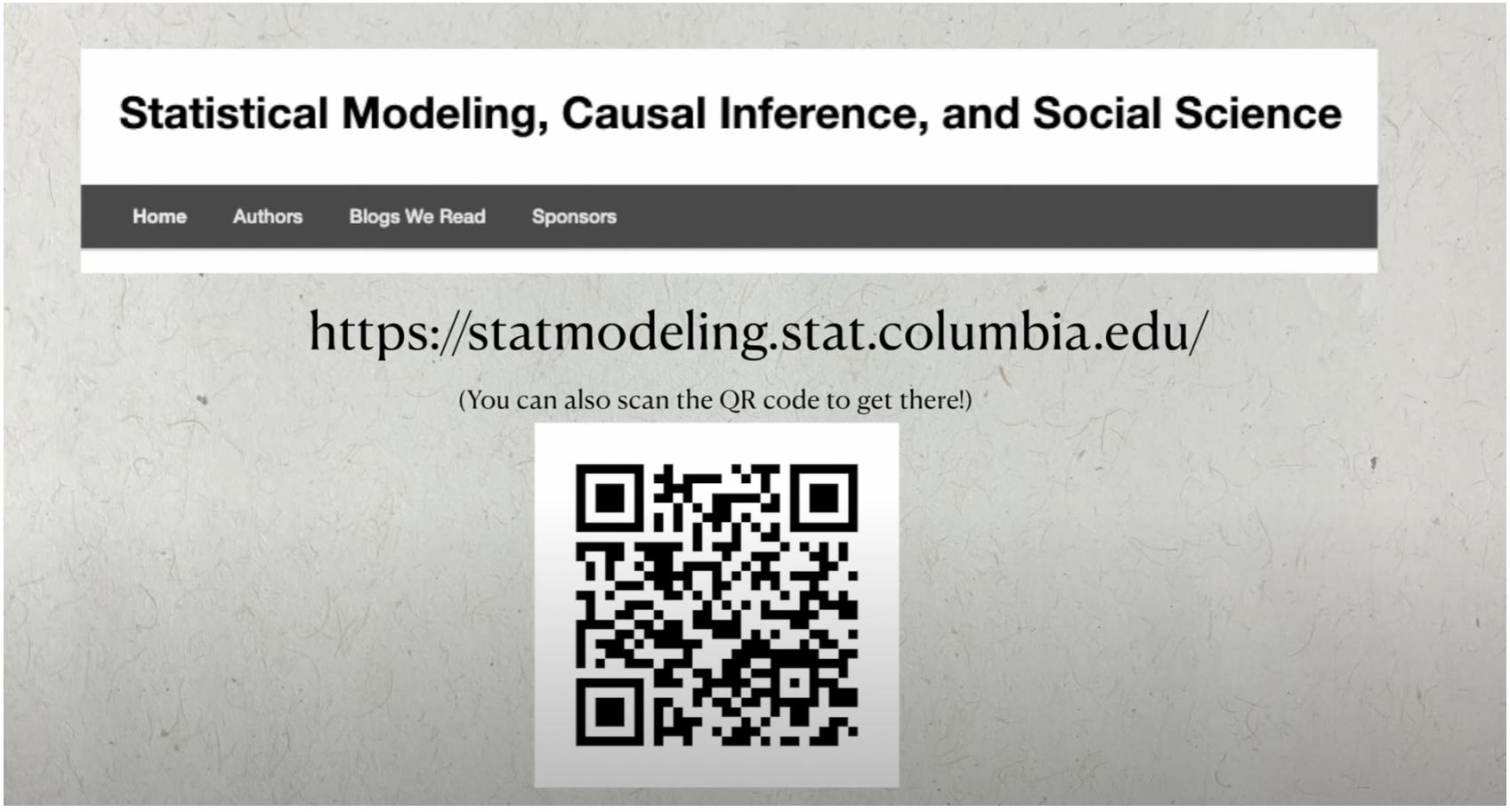
Authors' Background (cont.)



Bayesian Data Analysis

- **The Book: Bayesian Data Analysis**
  - Written by Andrew Gelman, Aki Vehtari, and others.
  - Widely regarded as the foundational text ("the bible") for Bayesian practitioners.
  - Covers theory, computation, and applied Bayesian methods.
- **Their Authority on the Topic**
  - Through this book, Gelman and Vehtari have shaped the modern understanding of Bayesian statistics.
  - Their extensive research and contributions give them unique insights to answer: *What are the most important statistical ideas of the past 50 years?*

Gelman's Blog



Overview of the Paper

- Timeframe: 1970 to 2021, focusing on the development of modern statistics.
- 8 statistical ideas selected based on their influence on statistical theory, computation, and applications.
- Published to the *Journal of the American Statistical Association* as an **essay**, not a research manuscript.
- Acknowledges that no definitive list can encompass all significant ideas.

Key Statistical Concepts

The 8 Statistical Ideas Covered in the Paper:

1. Counterfactual Causal Inference
2. Bootstrapping and Simulation-Based Inference
3. Overparameterized Models and Regularization
4. Bayesian Multilevel Models
5. Generic Computation Algorithms
6. Adaptive Decision Analysis
7. Robust Inference
8. Exploratory Data Analysis (EDA)

Purpose: These concepts reflect significant innovations in statistical theory, computation, and application from the last 50 years.

The Challenge of Observational Data

- **In an ideal world: Experimental Data**
  - Researchers control interventions.
  - Participants are assigned to "treatment" and "control" groups.
  - Enables *causal claims* about the effect of an intervention.
- **In the real world: Observational Data**
  - Researchers cannot control who receives the intervention.
  - Only allows for *correlational claims*.
  - Several fields of study are prone to having more observational data i.e. statistics, econometrics, psychometrics, epidemiology, and computer science.

Challenge: How can we infer causality in the absence of experimental data?

Introducing Counterfactuals

- **Model Setup:**
  - Let X represent attending class:
    - $x = 0$: Did not attend class.
    - $x = 1$: Did attend class.
  - In reality, I choose to attend class and receive a score Y_1 on my paper.

- **Key Question:** Did attending class improve my score?
  - To answer this, we need to consider an alternative reality:
  - What score (Y_0) would I have received if I *did not* attend class?

Key Idea: Counterfactuals

- Y_0: The score in the unobserved reality (if I did not attend class).
- This unobserved reality is called the *counterfactual*.

01: Counterfactual Causal Inference

- Allows causal inference using observational data.
- Framework based on "potential outcomes" or "counterfactuals."

$$\text{Causal Effect: } Y(1) - Y(0)$$

- $Y(1)$: Outcome if treated.
- $Y(0)$: Outcome if untreated.
- Challenge: Only one outcome is observed.

02: Bootstrapping and Simulation-Based Inference

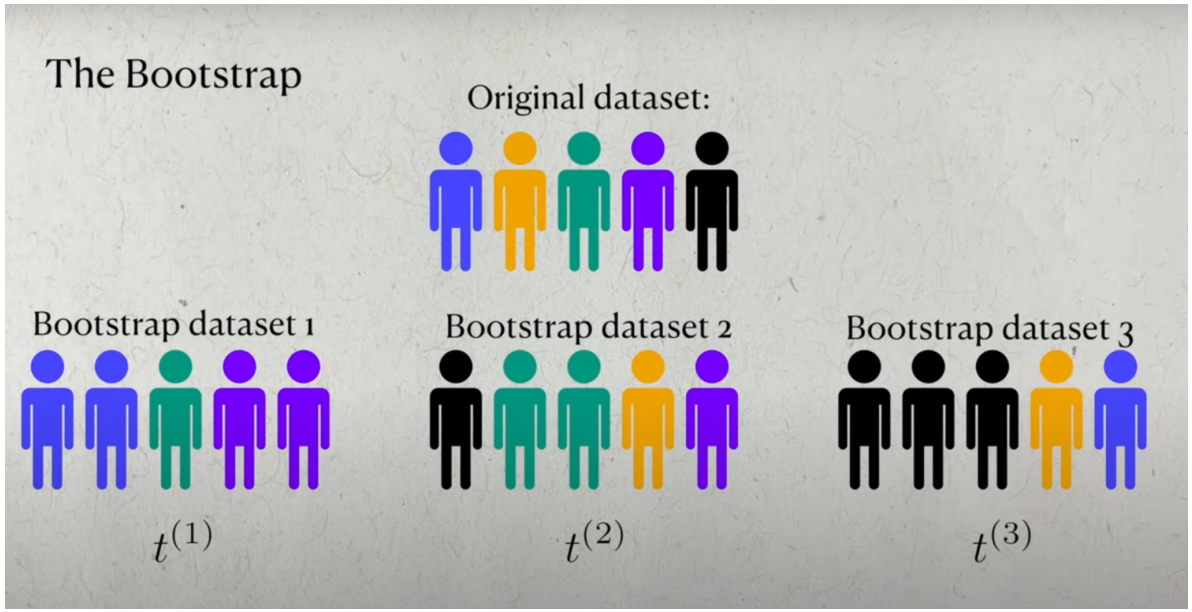
- Introduced by Bradley Efron (1979).
- Resampling technique to estimate sampling distributions without assumptions about data distribution.

Algorithm:

1. Resample the dataset with replacement.
2. Compute the statistic of interest (e.g., mean).
3. Repeat n times to estimate variability.

Bootstrapping and Simulation-Based Inference (cont.)

- **The Bootstrap Algorithm:**



- $\hat{F}(t)$ Goal: Estimate the CDF of some statistic.
- **Intuition:** Data sampled from the original dataset resembles a new dataset.
- For each bootstrap dataset, calculate a statistic of interest and derive its distribution:

$$t^{(1)}, t^{(2)}, t^{(3)} \rightarrow \hat{F}(t)$$

Simulation-Based Inference

- **Example using Bayes' Rule:**

$$P(\theta \mid X) = \frac{P(X \mid \theta)\, P(\theta)}{P(X)}$$

- **Bayesian Approach:**
  - Bayesian encodes knowledge in the form of prior distributions.
  - Conduct prior and posterior predictive checks, incredibly useful for understanding statistical models.
- **Rise of Computational Power:**
  - Advances in computational power have made simulations easier and more widely applicable.

03: Overparameterized Models and Regularization

- High-dimensional models with more parameters than data points.
- Regularization prevents overfitting by adding penalties to the model:

$$\text{LASSO: } \min \left(||Y - X\beta||^2 + \lambda ||\beta||_1 \right)$$

Lasso regularization, also known as L1 regularization, is a regression analysis technique that uses a penalty term to improve the accuracy and interpretability of statistical models.

- **How it works:**
  - Lasso adds a penalty term to the residual sum of squares (RSS) and multiplies it by a regularization parameter, lambda (λ).
  - This penalty term constrains the size of the estimated coefficients, encouraging sparsity in the model.

Example

Neural networks with regularization techniques balance flexibility and robustness.

04: Bayesian Multilevel Models

- Models hierarchical data with varying parameters at different levels.
- Example: Aggregating data across different groups in meta-analysis.

$$y_{ij} = \beta_0 + \beta_1 X_{ij} + u_j + \epsilon_{ij}$$

- u_j: Random effect for group j.
- ϵ_{ij}: Error term for observation i in group j.

Advantage

Combines individual-level and group-level variability for improved estimates.

05: Generic Computation Algorithms

- Advances in algorithms like MCMC, EM, and variational inference.
- Enabled complex models and large-scale Bayesian analysis.

06: Adaptive Decision Analysis

- Framework for making decisions during experiments.
- Application: Stopping clinical trials early for ethical reasons.

07: Robust Inference

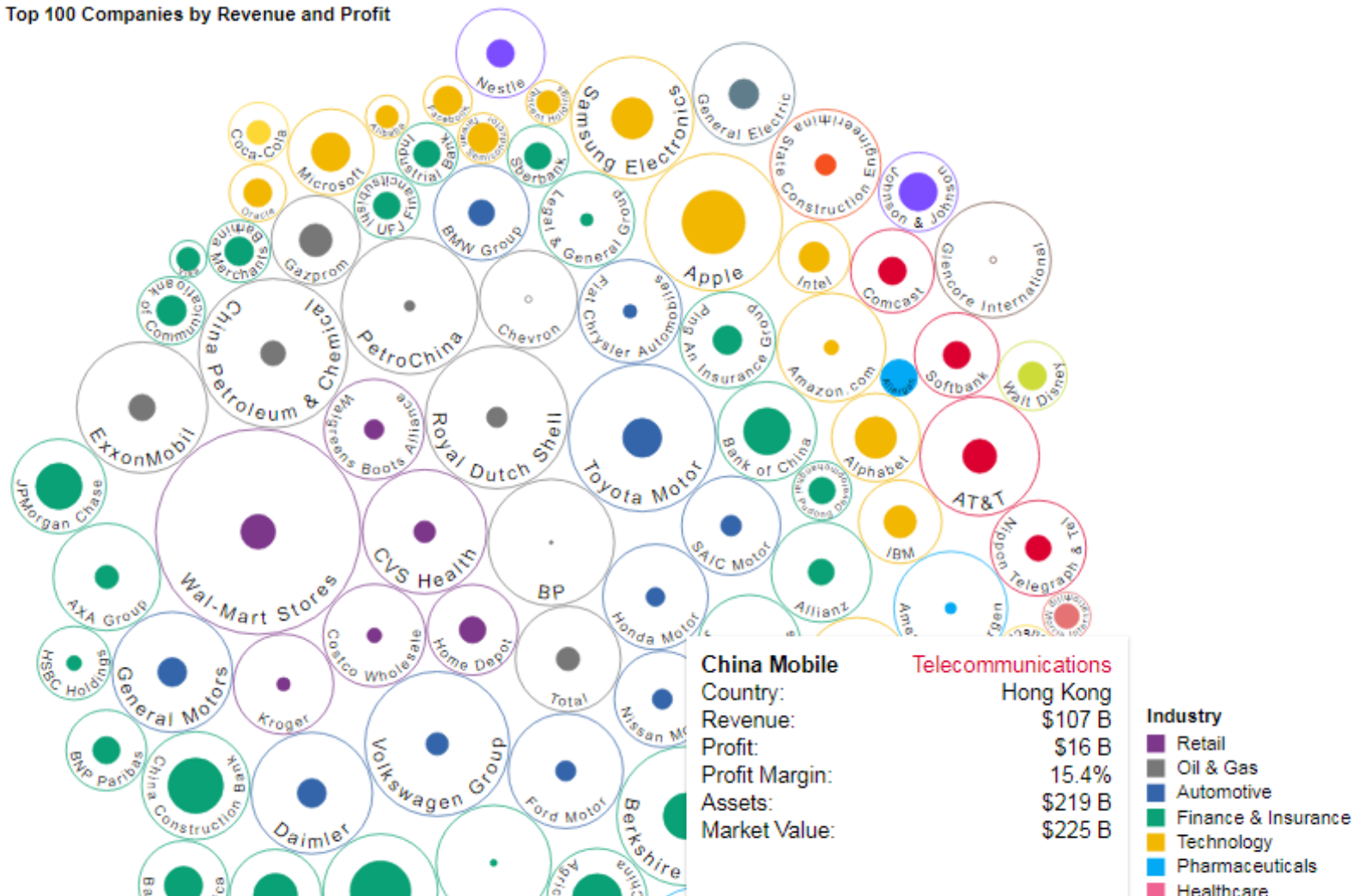
- Focuses on reliability under model misspecification.
- Example: Median-based estimators and propensity score matching.

Key Insight

Robust inference allows valid results even when data deviates from assumptions.

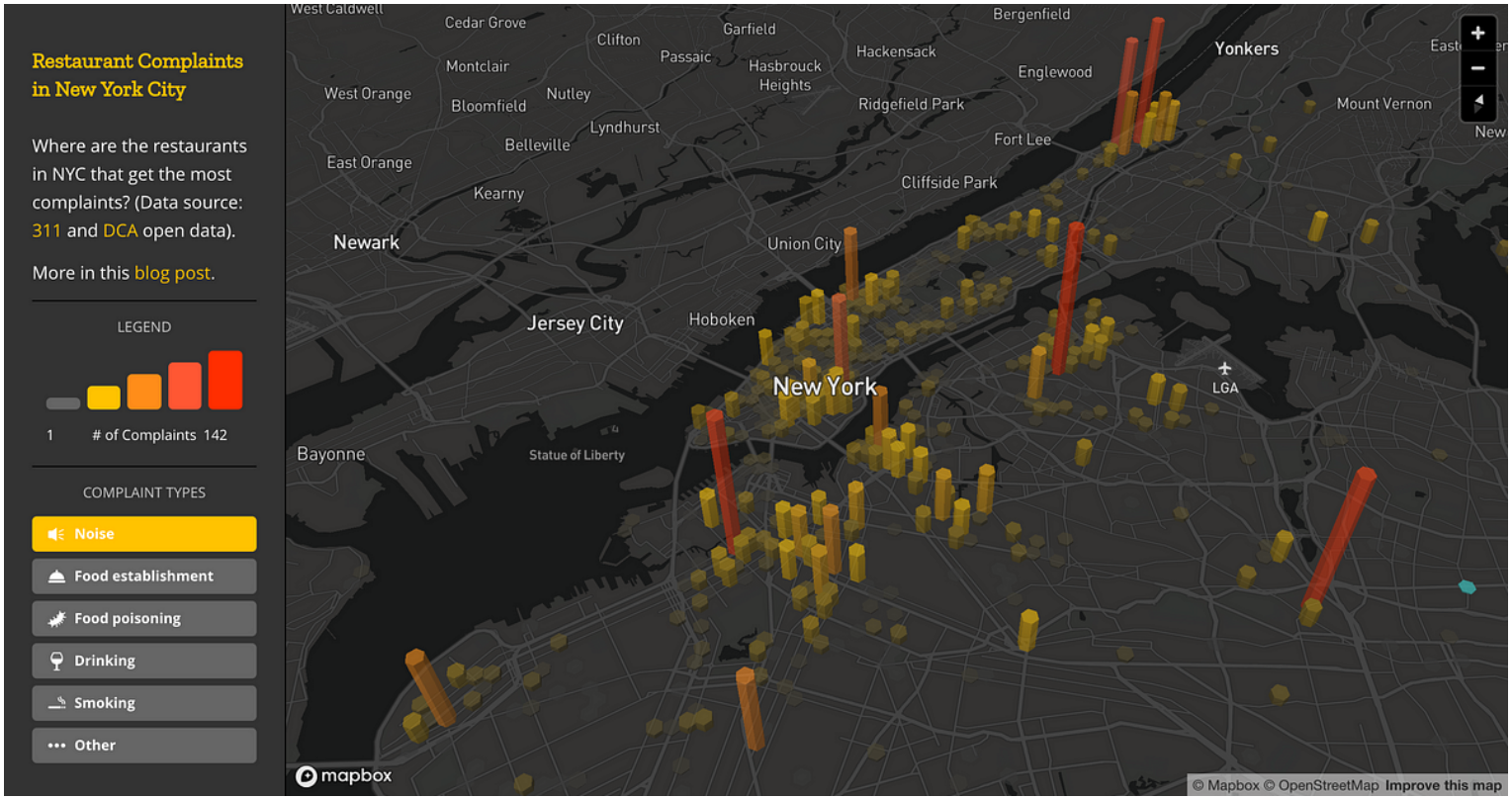
08: Exploratory Data Analysis (EDA)

- Emphasizes visualization and insights over intense theory and computation.
- Useful in understanding the relation between data, fitted model, and predictions.



Connection to NYC Open Data

- Apply statistical methods to NYC datasets.
- Example: Visualize and analyze restaurant complaints using robust inference and EDA.



https://labs.mapbox.com/bites/00304/

The Importance of Human Oversight in Statistical Innovations

- As computational power advances, machine learning and statistical algorithms can model complex systems.
- However, these models are only as good as the assumptions and data they are based on.
- Example: Self-driving cars can use machine learning to navigate, but human oversight is needed to determine:
  - Are the outcomes (e.g., accident rates) statistically significant?
  - Are the algorithms operating ethically and equitably?
- **Key Point**: Computational tools are powerful, but without human observation and ethical guidance, they can lead to unintended consequences.

Reflection from Gelman

"On one hand, you have all these amazing things that machine learning can do, like self-driving cars, but you'll need a statistician to tell you if the number of people being killed by the self-driving cars is statistically significant." – Paraphrased from Andrew Gelman

Questions?

Thank you! Any questions?

References

- Gelman, Andrew, and Aki Vehtari. 2021. "What are the Most Important Statistical Ideas of the Past 50 Years?" *Journal of the American Statistical Association* 116, no. 536: 2087–2097.
- Gelman, Andrew. *Statistical Modeling, Causal Inference, and Social Science*. Accessed at: `https://statmodeling.stat.columbia.edu/`
- Gelman, Andrew. *What are the most important statistical ideas of the past 50 years?* YouTube video. Available at: `https://www.youtube.com/watch?v=M6ha2UeSZbo`
- Gelman, Andrew. Additional commentary on statistical ideas. YouTube video. Available at: `https://youtu.be/nCyGhqQWj2g?si=GM9KpuWtg4tGV8je`
- Mapbox. 2018. "Exploring NYC Open Data with 3D Hexbins." Available at: `https://blog.mapbox.com/exploring-nyc-open-data-with-3d-hexbins-5af2b7d8bc46`