Big Data and Small Surveys: Which one should we trust more?

Xiao-Li Meng

Department of Statistics, Harvard University

Big Data and Small Surveys: Which one should we trust more?

Xiao-Li Meng

Department of Statistics, Harvard University

- Meng, X.-L. (2018) Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and The 2016 US Election. *Annals of Applied Statistics* Vol 2: 685-726

Big Data and Small Surveys: Which one should we trust more?

Xiao-Li Meng

Department of Statistics, Harvard University

- Meng, X.-L. (2018) Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and The 2016 US Election. *Annals of Applied Statistics* Vol 2: 685-726
- Many thanks to **Stephen Ansolabehere and Shiro Kuriwaki** for the CCES (**Cooperative Congressional Election Study**) data and analysis on 2016 US election.

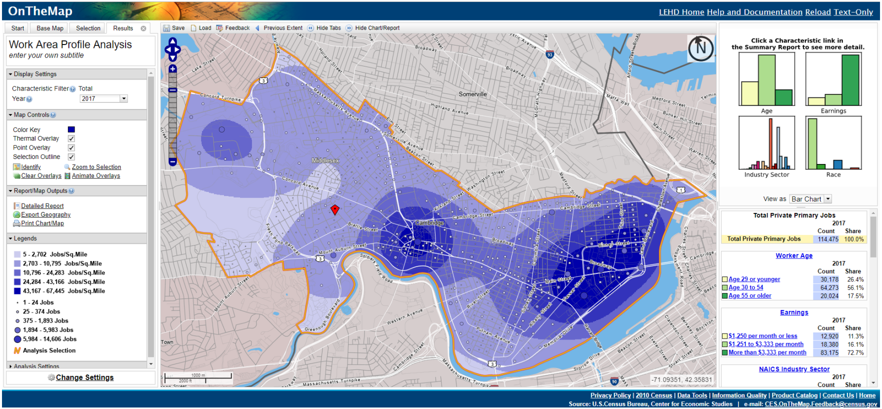
Big Data and Small Surveys: Which one should we trust more?

Xiao-Li Meng

Department of Statistics, Harvard University

- Meng, X.-L. (2018) Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and The 2016 US Election. *Annals of Applied Statistics* Vol 2: 685-726
- Many thanks to **Stephen Ansolabehere and Shiro Kuriwaki** for the CCES (**Cooperative Congressional Election Study**) data and analysis on 2016 US election.
- Thanks to many students and colleagues for augmenting my intelligence, and to on-line sources for enhancing my presentation.

OnTheMap Project of US Census Bureau



Multi-Source

- Built from more than 20 data sources in the LEHD (Longitudinal Employer-Household Dynamics) system. For example:

Multi-Source

- Built from more than 20 data sources in the LEHD (Longitudinal Employer-Household Dynamics) system. For example:
- **American Community Survey:** Surveys 3.5M households covering about 2.7% of 128M households.

Multi-Source

- Built from more than 20 data sources in the LEHD (Longitudinal Employer-Household Dynamics) system. For example:
- **American Community Survey:** Surveys 3.5M households covering about 2.7% of 128M households.
- **Administrative Records and Census:** Combined job frame using both Unemployment Insurance administrative records and the BLS-specified Quarterly Census of Employment and Wages, covering more than 98% of the US workforce.

Multi-Source

- Built from more than 20 data sources in the LEHD (Longitudinal Employer-Household Dynamics) system. For example:
- **American Community Survey:** Surveys 3.5M households covering about 2.7% of 128M households.
- **Administrative Records and Census:** Combined job frame using both Unemployment Insurance administrative records and the BLS-specified Quarterly Census of Employment and Wages, covering more than 98% of the US workforce.
  - *Unemployment Insurance record was never intended for statistical inference purposes.*

Which one should we trust more?

- We know that a 5% random sample is better than a 5% non-random sample in measurable ways (e.g. bias, uncertainty assessment).

Which one should we trust more?

- We know that a 5% random sample is better than a 5% non-random sample in measurable ways (e.g. bias, uncertainty assessment).
- **But is an 80% non-random sample "better" than a 5% random sample in measurable terms? 90%? 95%? 99%?** (Jeremy Wu 2012)

Which one should we trust more?

- We know that a 5% random sample is better than a 5% non-random sample in measurable ways (e.g. bias, uncertainty assessment).
- **But is an 80% non-random sample "better" than a 5% random sample in measurable terms? 90%? 95%? 99%?** (Jeremy Wu 2012)
- "**Which one should we trust more: a 1% survey with 60% response rate or a non-probabilistic dataset covering 80% of the population?**" (Keiding and Louis, 2016, *Journal of Royal Statistical Society, Series B*)

Surveys: Infer a Population of N by a Sample of $n << N$

- Graunt (1662); Laplace (1882)

Surveys: Infer a Population of N by a Sample of $n << N$

- Graunt (1662); Laplace (1882)
- The "**intellectually violent revolution**" in 1895 by Anders Kiær, Statistics Norway

Surveys: Infer a Population of N by a Sample of $n << N$

- Graunt (1662); Laplace (1882)
- The "**intellectually violent revolution**" in 1895 by Anders Kiær, Statistics Norway

- Landmark paper: Neyman (1934)

Surveys: Infer a Population of N by a Sample of $n << N$

- Graunt (1662); Laplace (1882)
- The "**intellectually violent revolution**" in 1895 by Anders Kiær, Statistics Norway

- Landmark paper: Neyman (1934)
- First implementation in 1940 US Census led by Morris Hansen

Why and when can we ignore the population size N?

- Think about tasting soup

Why and when can we ignore the population size N?

- Think about tasting soup
- Stir it well, then a few bits are sufficient **regardless of the size of the container!**

Why and when can we ignore the population size N?

- Think about tasting soup
- Stir it well, then a few bits are sufficient **regardless of the size of the container!**


Why and when can we ignore the population size N?

- Think about tasting soup
- Stir it well, then a few bits are sufficient **regardless of the size of the container!**

Why and when can we ignore the population size N?

- Think about tasting soup
- Stir it well, then a few bits are sufficient **regardless of the size of the container!**

$\Longleftrightarrow$

Why and when can we ignore the population size N?

- Think about tasting soup
- Stir it well, then a few bits are sufficient **regardless of the size of the container!**


$\Longleftrightarrow$

- **But what happens when we fail to stir (well)?**

A Fundamental Identity for Statistical Estimation

- Population $\{X_1, ..., X_N\}$; **Estimand**: Population Average $\bar{X}_N$

A Fundamental Identity for Statistical Estimation

- Population $\{X_1, ..., X_N\}$; **Estimand**: Population Average $\bar{X}_N$
- Record Indicator: $R_j = 1$ if X_j is recorded, and $R_j = 0$ otherwise.

A Fundamental Identity for Statistical Estimation

- Population $\{X_1, ..., X_N\}$; **Estimand**: Population Average $\bar{X}_N$
- Record Indicator: $R_j = 1$ if X_j is recorded, and $R_j = 0$ otherwise.
- Sample size $n = R_1 + \cdots + R_N$.

A Fundamental Identity for Statistical Estimation

- Population $\{X_1, ..., X_N\}$; **Estimand**: Population Average $\bar{X}_N$
- Record Indicator: $R_j = 1$ if X_j is recorded, and $R_j = 0$ otherwise.
- Sample size $n = R_1 + \cdots + R_N$.
- **Estimator**: Sample Average $\bar{X}_n$

A Fundamental Identity for Statistical Estimation

- Population $\{X_1, ..., X_N\}$; **Estimand**: Population Average $\bar{X}_N$
- Record Indicator: $R_j = 1$ if X_j is recorded, and $R_j = 0$ otherwise.
- Sample size $n = R_1 + \cdots + R_N$.
- **Estimator**: Sample Average $\bar{X}_n$

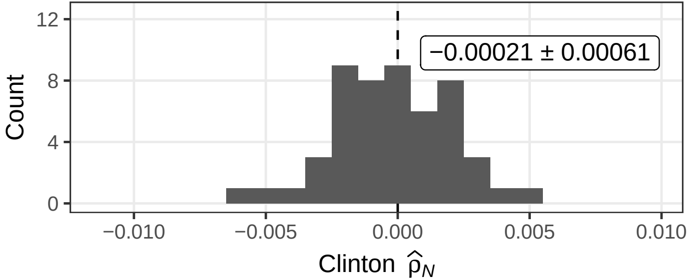
A Fundamental Identity for Statistical Estimation

- Population $\{X_1, ..., X_N\}$; **Estimand**: Population Average $\bar{X}_N$
- Record Indicator: $R_j = 1$ if X_j is recorded, and $R_j = 0$ otherwise.
- Sample size $n = R_1 + \cdots + R_N$.
- **Estimator**: Sample Average $\bar{X}_n$

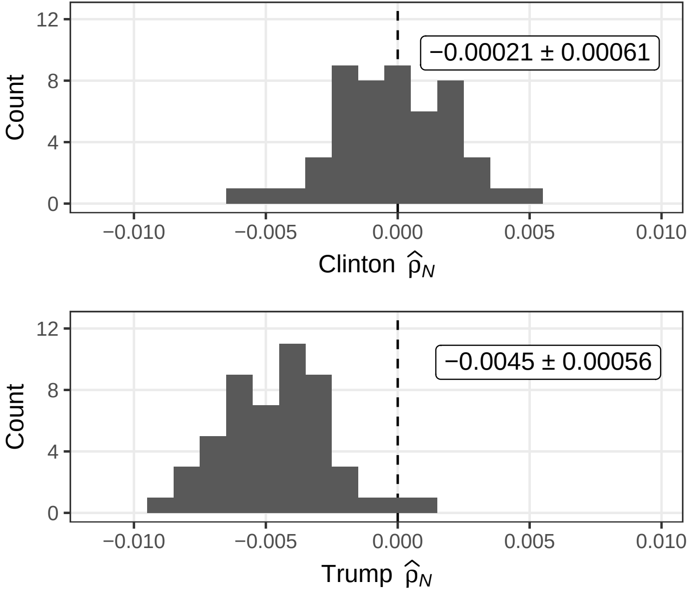
Three and only three ways to control the estimation error:

$$\underbrace{\bar{X}_n - \bar{X}_N}_{\textbf{Estimation Error}} = \underbrace{\mathrm{Corr}(R, X)}_{\textbf{Data Quality}} \times \underbrace{\sqrt{\frac{N-n}{n}}}_{\textbf{Data Quantity}} \times \underbrace{\mathrm{St.Dev}(X)}_{\textbf{Problem Difficulty}} .$$

Assessing $\hat{\rho}_N = \mathrm{Corr}(R, X)$ via Cooperative Congressional Election Study (50 states + Washington DC)



Assessing $\hat{\rho}_N = \mathrm{Corr}(R, X)$ via Cooperative Congressional Election Study (50 states + Washington DC)



What's the Implication of $\hat{\rho}_N = -0.005$?

- Many (major) election survey results were published daily for several months before Nov 8, 2016;

What's the Implication of $\hat{\rho}_N = -0.005$?

- Many (major) election survey results were published daily for several months before Nov 8, 2016;
- Roughly amounts to having opinions from (up to) $f = \frac{n}{N} = 1\%$ of US voting eligible population: $n \approx 2,300,000$;

What's the Implication of $\hat{\rho}_N = -0.005$?

- Many (major) election survey results were published daily for several months before Nov 8, 2016;
- Roughly amounts to having opinions from (up to) $f = \frac{n}{N} = 1\%$ of US voting eligible population: $n \approx 2,300,000$;
- Equivalent to about 2,300 surveys of 1,000 respondents each.

What's the Implication of $\hat{\rho}_N = -0.005$?

- Many (major) election survey results were published daily for several months before Nov 8, 2016;
- Roughly amounts to having opinions from (up to) $f = \frac{n}{N} = 1\%$ of US voting eligible population: $n \approx 2,300,000$;
- Equivalent to about 2,300 surveys of 1,000 respondents each.

Effective Sample Size (ESS)

The size of a simple random sample with the same accuracy

What's the Implication of $\hat{\rho}_N = -0.005$?

- Many (major) election survey results were published daily for several months before Nov 8, 2016;
- Roughly amounts to having opinions from (up to) $f = \frac{n}{N} = 1\%$ of US voting eligible population: $n \approx 2,300,000$;
- Equivalent to about 2,300 surveys of 1,000 respondents each.

Effective Sample Size (ESS)

The size of a simple random sample with the same accuracy

When $\hat{\rho}_N = -0.005 = -1/200$, and hence

$$\mathrm{ESS} = \frac{f}{1-f}\frac{1}{\hat{\rho}_N^2} = \frac{1}{99} \times 40000 \approx 404!$$

What's the Implication of $\hat{\rho}_N = -0.005$?

- Many (major) election survey results were published daily for several months before Nov 8, 2016;
- Roughly amounts to having opinions from (up to) $f = \frac{n}{N} = 1\%$ of US voting eligible population: $n \approx 2,300,000$;
- Equivalent to about 2,300 surveys of 1,000 respondents each.

Effective Sample Size (ESS)

The size of a simple random sample with the same accuracy

When $\hat{\rho}_N = -0.005 = -1/200$, and hence

$$\mathrm{ESS} = \frac{f}{1-f}\frac{1}{\hat{\rho}_N^2} = \frac{1}{99} \times 40000 \approx 404!$$

- **A** 99.98% **reduction in** n**, caused by** $\hat{\rho}_N = -0.005$**.**

What's the Implication of $\hat{\rho}_N = -0.005$?

- Many (major) election survey results were published daily for several months before Nov 8, 2016;
- Roughly amounts to having opinions from (up to) $f = \frac{n}{N} = 1\%$ of US voting eligible population: $n \approx 2,300,000$;
- Equivalent to about 2,300 surveys of 1,000 respondents each.

Effective Sample Size (ESS)

The size of a simple random sample with the same accuracy

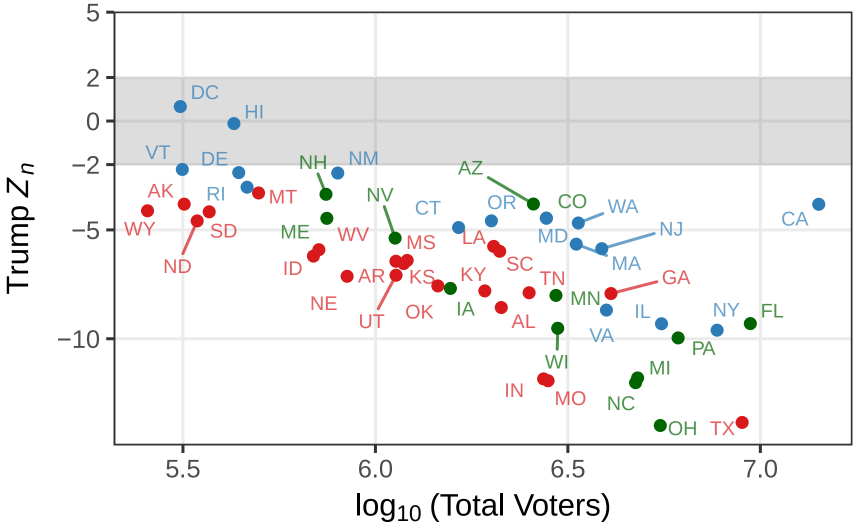
When $\hat{\rho}_N = -0.005 = -1/200$, and hence

$$\mathrm{ESS} = \frac{f}{1-f}\frac{1}{\hat{\rho}_N^2} = \frac{1}{99} \times 40000 \approx 404!$$

- **A** 99.98% **reduction in** n**, caused by** $\hat{\rho}_N = -0.005$.
- **Butterfly Effect** due to Law of Large Populations (LLP)

$$\textbf{Relative Error} = \sqrt{\mathbf{N-1}}\,\hat{\rho}_{\mathbf{N}}$$

LLP: The more voters, the higher the bias in our prediction



The Big Data Paradox:

If we do not pay attention to data quality, then

The bigger the data,

the surer we fool ourselves.

Lessons Learned ...

- Data quality is far more important than data quantity.

Lessons Learned ...

- Data quality is far more important than data quantity.
- Compensating for quality with quantity is a doomed game.

Lessons Learned ...

- Data quality is far more important than data quantity.
- Compensating for quality with quantity is a doomed game.
- **It is far more important to reduce sampling and non-response biases than non-response rates.**

Lessons Learned ...

- Data quality is far more important than data quantity.
- Compensating for quality with quantity is a doomed game.
- **It is far more important to reduce sampling and non-response biases than non-response rates.**
- **Invest in small but very high quality surveys than large surveys with uncontrolled/unknown quality.**

Lessons Learned ...

- Data quality is far more important than data quantity.
- Compensating for quality with quantity is a doomed game.
- **It is far more important to reduce sampling and non-response biases than non-response rates.**
- **Invest in small but very high quality surveys than large surveys with uncontrolled/unknown quality.**
- When combining datasets, relatively tiny but higher quality ones should be given far more weights than suggested by their sizes.

Lessons Learned ...

- Data quality is far more important than data quantity.
- Compensating for quality with quantity is a doomed game.
- **It is far more important to reduce sampling and non-response biases than non-response rates.**
- **Invest in small but very high quality surveys than large surveys with uncontrolled/unknown quality.**
- When combining datasets, relatively tiny but higher quality ones should be given far more weights than suggested by their sizes.
- For population inferences, "bigness" of "Big Data" should be measured by their relative size, not absolute size.

Lessons Learned ...

- Data quality is far more important than data quantity.
- Compensating for quality with quantity is a doomed game.
- **It is far more important to reduce sampling and non-response biases than non-response rates.**
- **Invest in small but very high quality surveys than large surveys with uncontrolled/unknown quality.**
- When combining datasets, relatively tiny but higher quality ones should be given far more weights than suggested by their sizes.
- For population inferences, "bigness" of "Big Data" should be measured by their relative size, not absolute size.
- Probability sampling is an extremely powerful tool for ensuring data quality, but it is not the only strategy.

Lessons Learned ...

- Data quality is far more important than data quantity.
- Compensating for quality with quantity is a doomed game.
- **It is far more important to reduce sampling and non-response biases than non-response rates.**
- **Invest in small but very high quality surveys than large surveys with uncontrolled/unknown quality.**
- When combining datasets, relatively tiny but higher quality ones should be given far more weights than suggested by their sizes.
- For population inferences, "bigness" of "Big Data" should be measured by their relative size, not absolute size.
- Probability sampling is an extremely powerful tool for ensuring data quality, but it is not the only strategy.

Lessons Learned ...

- Data quality is far more important than data quantity.
- Compensating for quality with quantity is a doomed game.
- **It is far more important to reduce sampling and non-response biases than non-response rates.**
- **Invest in small but very high quality surveys than large surveys with uncontrolled/unknown quality.**
- When combining datasets, relatively tiny but higher quality ones should be given far more weights than suggested by their sizes.
- For population inferences, "bigness" of "Big Data" should be measured by their relative size, not absolute size.
- Probability sampling is an extremely powerful tool for ensuring data quality, but it is not the only strategy.

Three Enemies of Surveys and Data Science in General

- **Selection**

Lessons Learned ...

- Data quality is far more important than data quantity.
- Compensating for quality with quantity is a doomed game.
- **It is far more important to reduce sampling and non-response biases than non-response rates.**
- **Invest in small but very high quality surveys than large surveys with uncontrolled/unknown quality.**
- When combining datasets, relatively tiny but higher quality ones should be given far more weights than suggested by their sizes.
- For population inferences, “bigness” of ”Big Data” should be measured by their relative size, not absolute size.
- Probability sampling is an extremely powerful tool for ensuring data quality, but it is not the only strategy.

Three Enemies of Surveys and Data Science in General

- **Selection**
- **Selection**

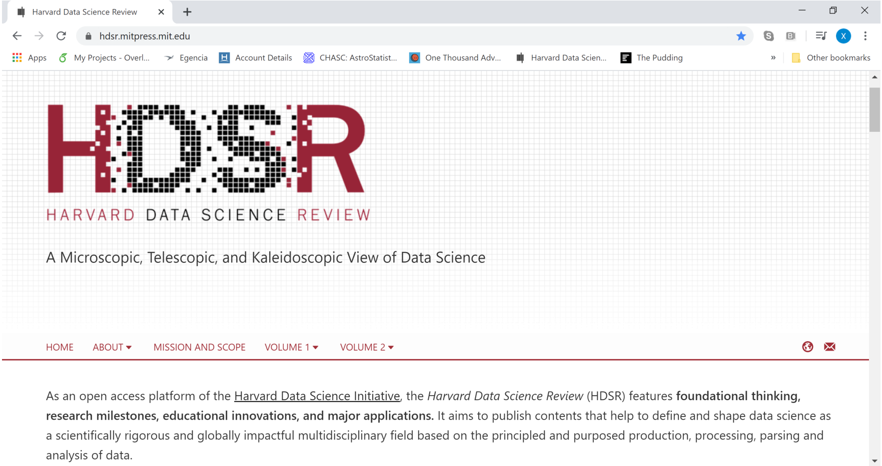
Lessons Learned ...

- Data quality is far more important than data quantity.
- Compensating for quality with quantity is a doomed game.
- **It is far more important to reduce sampling and non-response biases than non-response rates.**
- **Invest in small but very high quality surveys than large surveys with uncontrolled/unknown quality.**
- When combining datasets, relatively tiny but higher quality ones should be given far more weights than suggested by their sizes.
- For population inferences, "bigness" of "Big Data" should be measured by their relative size, not absolute size.
- Probability sampling is an extremely powerful tool for ensuring data quality, but it is not the only strategy.

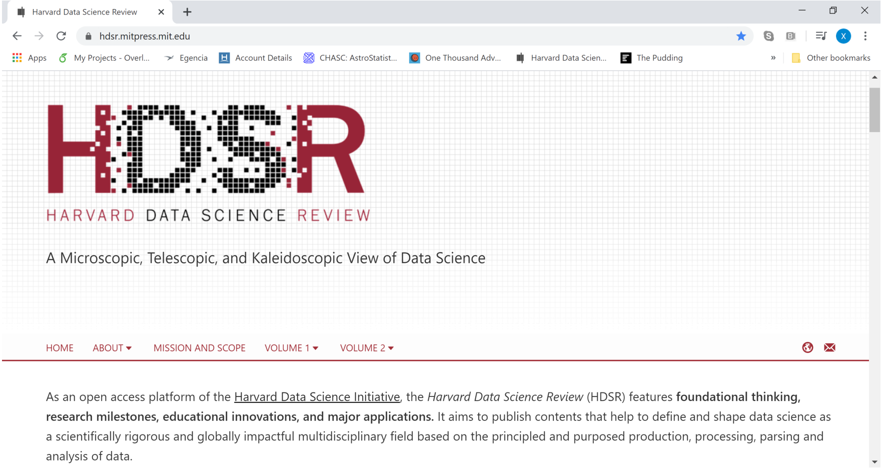
Three Enemies of Surveys and Data Science in General

- **Selection**
- **Selection**
- **Selection**

More Lessons From ...



More Lessons From ...



- Open Access: https://hdsr.mitpress.mit.edu/
- Global Correspondents: One correspondent from each country/region (Contact: **datasciencereview@harvard.edu**)

Everything Data Science and Data Science for Everyone

PANORAMA

Overviews, Visions, and Debates

Teresa A. Sullivan

Coming To Our Census: How Social Statistics Underpin Our Democracy (And Republic)

Commentary by: *Margo J. Anderson · Thomas R. Belin · Ray Chambers · Constance F. Citro · Reynolds Farley · Howard Hogan · Karen Kafadar · Dudley L. Poston, Jr. · Dennis Trewin*

Rejoinder from: *Teresa A. Sullivan*

Daniel L. Oberski and Frauke Kreuter

Differential Privacy and Social Science: An Urgent Puzzle

Everything Data Science and Data Science for Everyone

PANORAMA

Overviews, Visions, and Debates

Teresa A. Sullivan

Coming To Our Census: How Social Statistics Underpin Our Democracy (And Republic)

Commentary by: *Margo J. Anderson · Thomas R. Belin · Ray Chambers · Constance F. Citro · Reynolds Farley · Howard Hogan · Karen Kafadar · Dudley L. Poston, Jr. · Dennis Trewin*

Rejoinder from: *Teresa A. Sullivan*

Daniel L. Oberski and Frauke Kreuter

Differential Privacy and Social Science: An Urgent Puzzle

EFFECTIVE POLICY LEARNING

Data science for policy making and makers

Column Co-Editors: Frauke Kreuter and Nancy Potok

Brian Moyer and Abe Dunn

Measuring the Gross Domestic Product (GDP): The Ultimate Data Science Project

RECREATIONS IN RANDOMNESS

Data science for leisure activities

Column Editor: Mark Glickman

Ben Zauzmer

Oscar Seasons: The Intersection of Data and the Academy Awards