Data Quality & Bias

INST447 Data Source and Manipulation

Wei Ai (aiwei@umd.edu)

When "Correct" Code yields "Wrong" Answers

```
df.dropna().groupby('group').mean()
```

UC Berkeley Gender Bias

- In the early 1970s, the University of California, Berkeley was sued for gender discrimination over admission to graduate school.
- Of the 8,442 male applicants for the fall of 1973, 44 percent were admitted,
- But of the 4,351 female applicants, only 35 percent were accepted.

```
df.groupby('gender').agg({'admitted': 'mean'})
```

UC Berkeley Gender Bias

- When you look department by department, most departments admitted women at equal or higher rates.

```
df.groupby(['department','gender']).agg({'admitted': 'mean'})
```

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	**825**	62%	108	82%
B	585	63%	**560**	63%	25	68%
C	918	35%	325	37%	**593**	34%
D	792	34%	417	33%	375	35%
E	584	25%	191	28%	**393**	24%
F	714	6%	373	6%	341	7%
Total	4526	39%	2691	45%	1835	30%

The Aggregation Trap

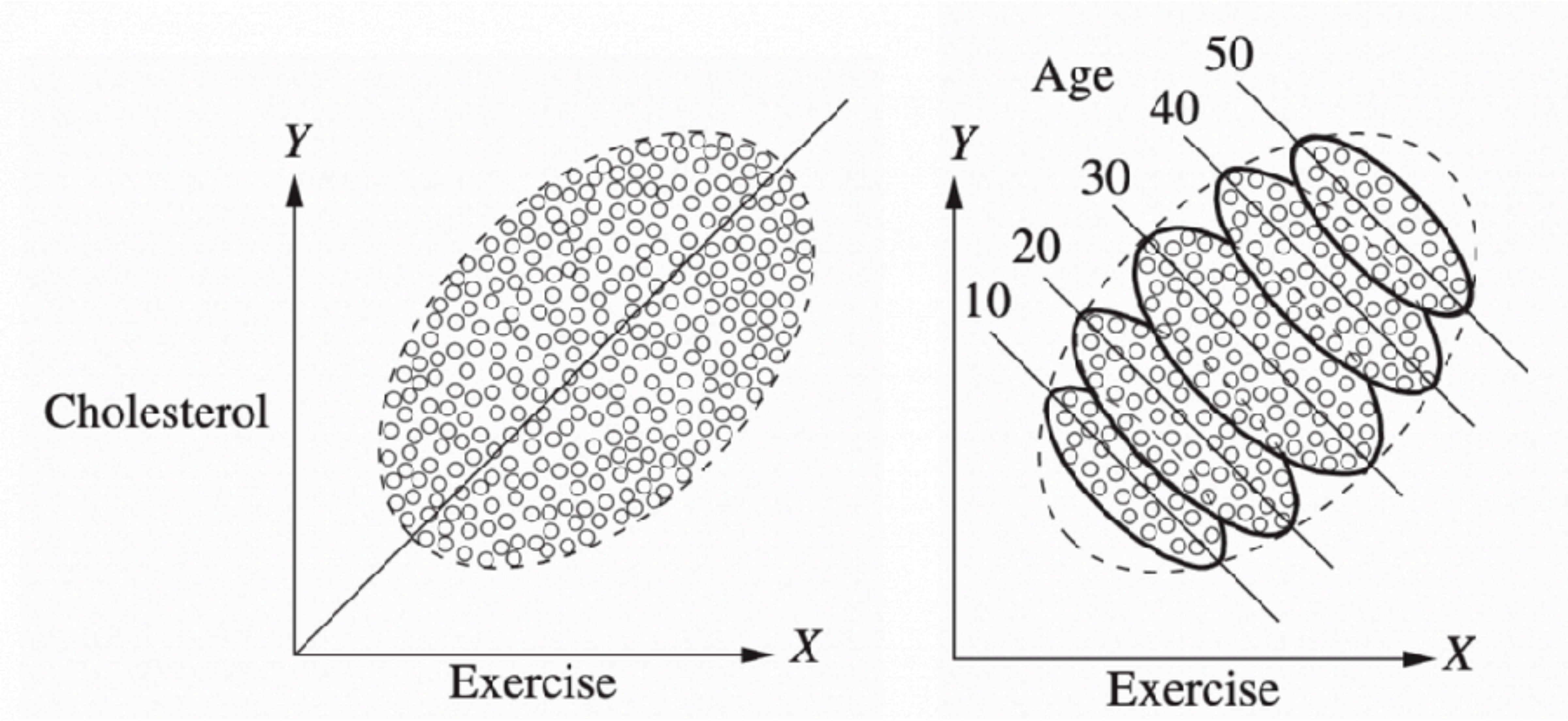
- Women applied disproportionately to competitive departments (English, History) while men applied to less competitive ones (Engineering, Chemistry at the time)

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	**825**	62%	108	82%
B	585	63%	**560**	63%	25	68%
C	918	35%	325	37%	**593**	34%
D	792	34%	417	33%	375	35%
E	584	25%	191	28%	**393**	24%
F	714	6%	373	6%	341	7%
Total	4526	39%	2691	45%	1835	30%

The Simpson’s Paradox

- Practically, this is a Join/Groupby issue. If you aggregate at the wrong level (university) vs. the right level (department), you get inverted results.
- Formally, this is called Simpson’s paradox. It occurs when the direction of an association between two variables reverses (or disappears) when the data are disaggregated by a third variable.
- A special case of omitted variable bias: simpson’s reversal arises because a confounding variable is omitted from the analysis.

Another Example of Simpson's Paradox



Exercise appears to be beneficial (downward slope) in each age group but harmful (upward slope) in the population as a whole.

Pearl, Judea. The Book of Why (ch. 6). Basic Books. 2018

Women's Labor Market

- Dataset:
  - people,
  - gender,
  - wage,
  - labor force participation (employed vs not in labor force).
- Question: We want to compare average wages of men vs women.
- The Issue: many women report no wage.

Naive fix #1

- Set missing to $0

```
df.fillna(0).groupby('gender').mean()
```

- Creates a massive artificial spike at zero
- Drastically underestimates the mean wage for women
- We're treating "not participating" as "participating but earning nothing"

Naive fix #2

- Drop missing rows

```
df.dropna().groupby('gender').mean()
```

- Women who *choose* to work are a **self-selected** sample—likely those with higher earning potential.
- Dropping non-participants overestimates women's average wages because you're only counting “winners.

Missing Data Mechanisms Matter

- Missing Completely At Random (MCAR)
  - The probability that a value is missing is independent of both the observed data and the unobserved (missing) data. Missingness has no pattern.
  - This is the best-case scenario. Complete-case analysis (dropping missing rows) is unbiased
- In our labor market analysis:
  - If these missing wages were purely random, almost any approach is fine.

Missing At Random (MAR)

- The missingness may depend on observed variables, but not on the missing values themselves after controlling for observed variables.
- The missingness has a pattern, but the pattern is explainable using variables you have.
  - Once we account for available information, missingness does not depend on the unobserved value.
  - Techniques include multiple imputation, regression imputation, mixed models…
- In our labor market analysis:
  - If missingness depends on observed covariates (e.g., age, education), we can model it.

Missing Not At Random (MNAR)

- Missingness depends on the unobserved data itself, even after accounting for the observed data.
- Standard imputation methods may be biased; requires specialized methods.
- In our labor market analysis:
  - Here the missingness is *the thing we care about*—probability of labor-force participation depends on the wage opportunity."

Heckman’s Two-Step Correction model

- Step 1: Model the Selection Process
  - In the first stage, the researcher formulates a model for the probability of working.
  - What predicts whether a woman participates in the labor force?
  - Variables might include: age, education, number of children, husband's income, family assets
  - This gives us estimated probabilities of participation for each woman

Heckman’s Two-Step Correction model

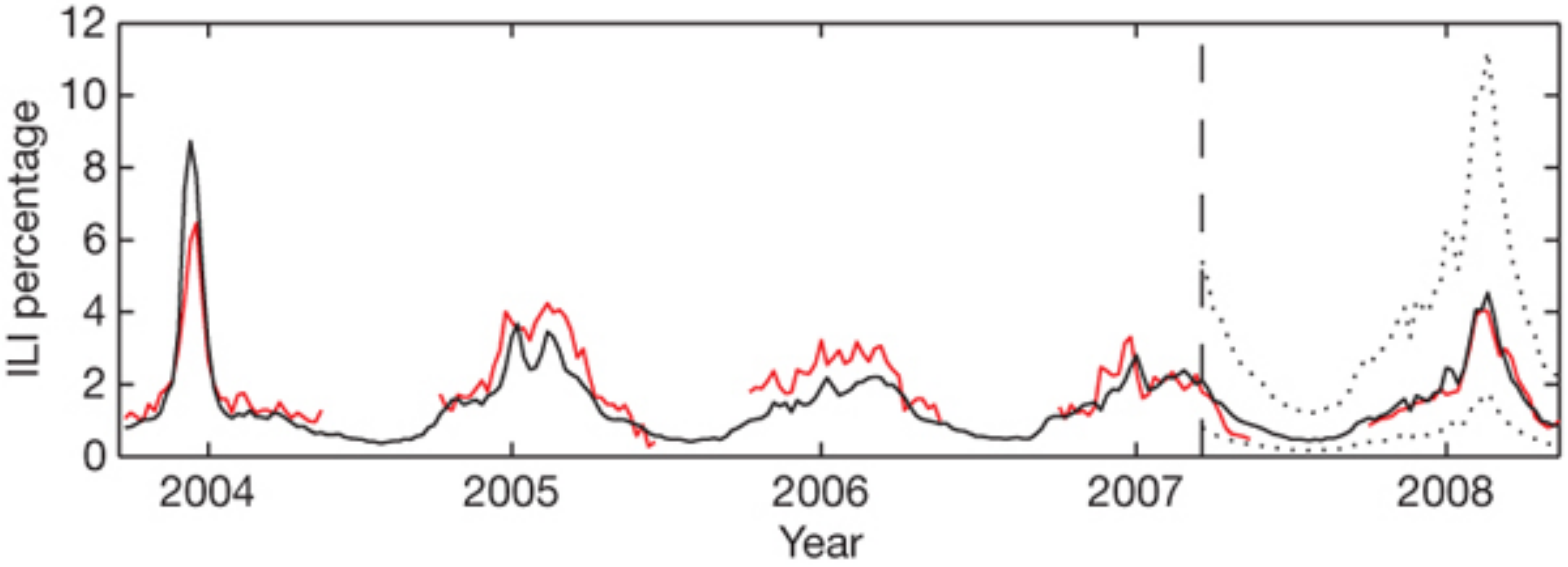
- Step 2: Correct the Wage Equation
  - Using the results from step 1, compute the "nonselection hazard” (inverse of the Mills ratio) for each observation. Then run the regression with the nonselection hazard added as an additional explanatory variable.
  - The inverse Mills ratio is essentially a "correction factor" that accounts for the fact that observed workers are not randomly selected from all women—they're the ones whose market wages exceeded their reservation wages.
- Heckman’s Correction was proposed in 1979, and has been extended and widely used in social science.
- Heckman received Nobel Prize in Econ in 2020 for his work in this field.

The Censoring Trap

- Missingness is not an accident; it is a structural signal.
- You cannot clean this with code alone;
- You need a model of why the data is missing

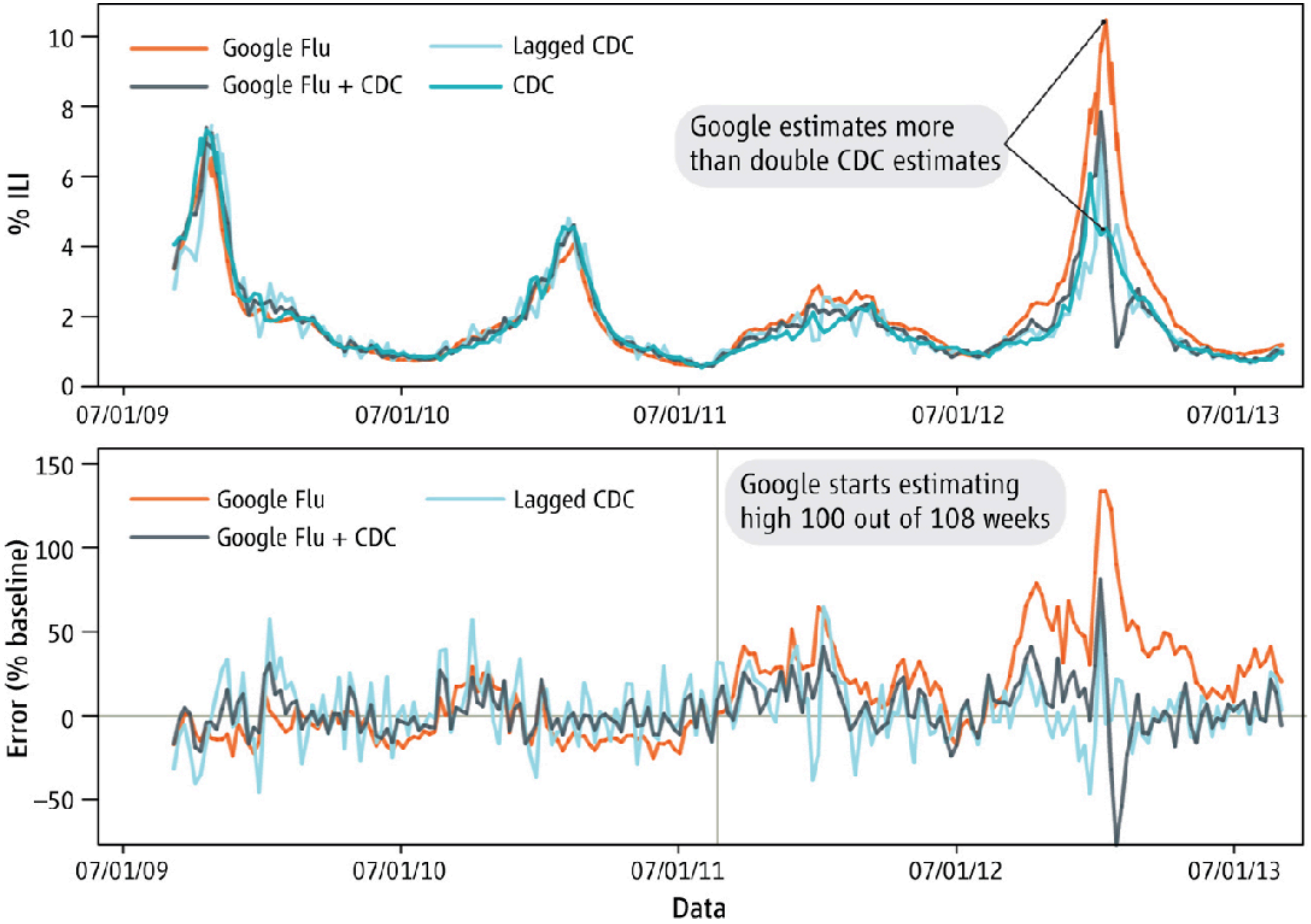
Google Flu Trend

- Google discovers search queries predict ILI (influenza-like illness) 2 weeks ahead of CDC
- Purely correlational — no epidemiology
- Correlate the special-temporal trend of search terms with the trend of the percentage of ILI (influenza-like illness) visits.



Ginsberg et al., 2009. Detecting influenza epidemics using search engine query data. Nature, 457(7232)

GFT Over Estimation after 2011



Google estimates more than double CDC estimates

Google starts estimating high 100 out of 108 weeks

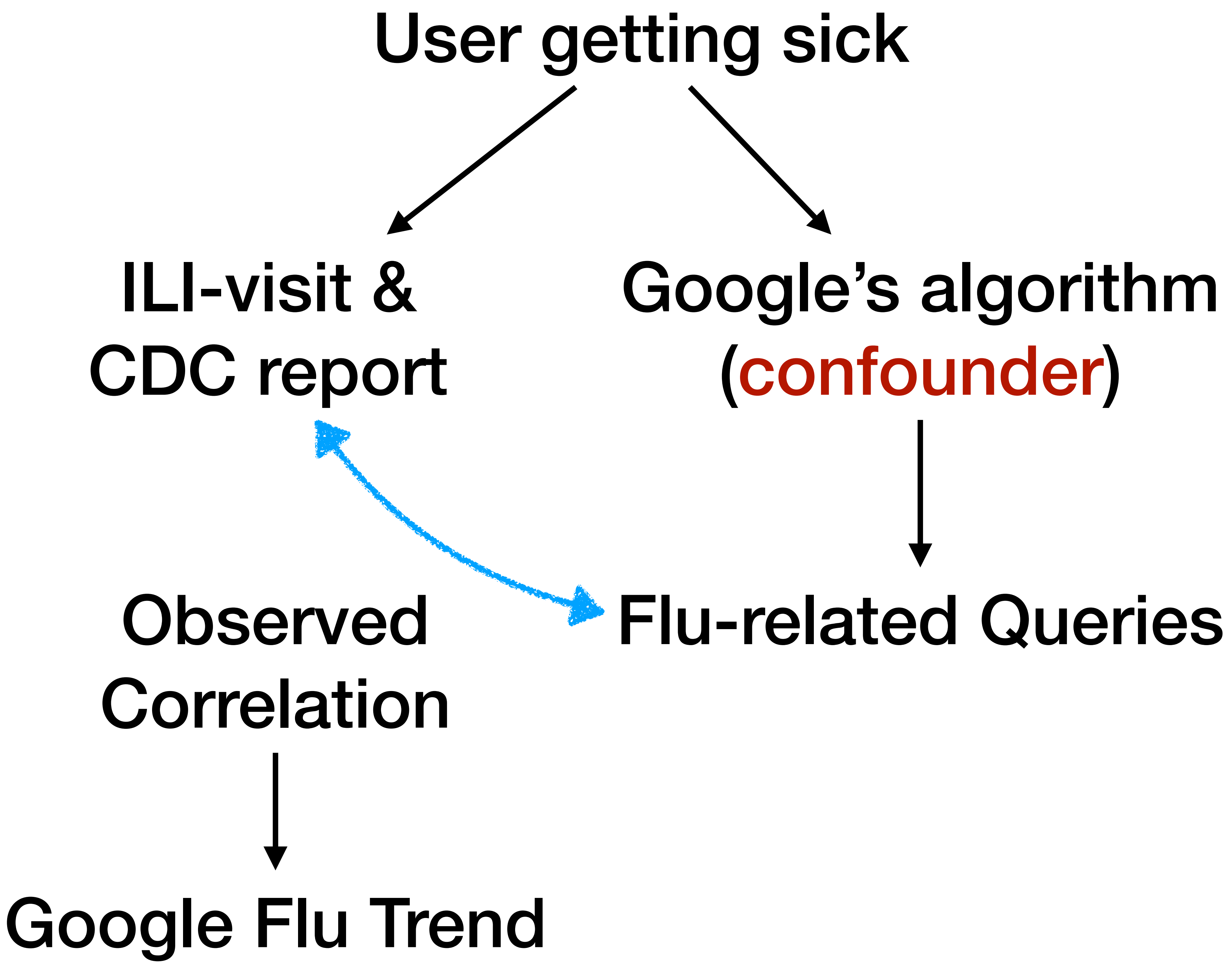
Lazer et al., 2014. The Parable of Google Flu: Traps in Big Data Analysis. Science, 343(6176)

Re-examine the Failure of GFT

- Changes of Google’s search algorithm:
  - 06/2011: Provide *suggested* additional terms
  - 02/2012: Return potential diagnoses for queries including physical symptoms like “fever” and “cough”
  - Result in increase of some queries
- GFT is overfitting the observed data, unaware of the changes underlying the data generating process.
- The system that measures the world *changes* the world (or at least changes the *data generation process*).

Concept Drift

- Product interventions:
  - Auto-complete suggestions, “related search” features → artificially boost certain queries.
  - Media coverage about “flu season” → more people search “flu symptoms” even when not sick.
- Concept drift:
  - The mapping from query volume to true flu incidence changed over time.
  - The model treated the world as stationary; Google as a product did not.



Popularity/Position Bias & Feedback Loops

- Imagine a search engine or recommender system:
  - It shows items ranked by an existing model.
  - You log clicks and use them as “relevance labels” for training the next model.
- Position bias
  - Items in the top positions get more clicks regardless of true relevance.
  - Your training data reflects what users saw, not the entire candidate set.
- Feedback loops
  - An item gets slightly more clicks early → rank up → more exposure → more clicks
  - looks “objectively” popular.
  - The system confuses “popular in the interface” with “intrinsically good.”

The Self-Selection Trap

- Think about *any* rating systems:
  - Course evaluation/RateMyProfessor.
  - Doctor review.
  - Restaurant review.
- What do we know about the data-generating process?
  - Who leaves a review, and who stays silent?
- To leave a data point, a user must pay a 'time cost.'
  - The data we see is biased towards users where *Emotional Intensity* > *Time Cost*
  - And don't forget other hidden incentives: extra credit, free product,…

What Patterns Emerge?

- Bias isn't always about demographics—it's about *inference validity*
- Every preprocessing choice encodes an assumption about the world
- The question is always: **How was this data generated?**

Story	Problem	Concept
Berkeley	Aggregation hides confounders	Simpson's paradox / omitted variable bias
Labor Market	Missingness related to the value itself	MNAR
Google Flu	Data reflects the system that generated it	Feedback loops/concept drift
Search Engine	Biased training data	Position Bias, Feedback Loop
Rating Systems	Not representative of "everyone"	Self Selection Bias

Where AI helps vs. where it doesn't

AI can help	Human judgment still needed
Detecting missing data patterns	Understanding *why* data is missing
Flagging outliers	Deciding if outliers are errors or real
Suggesting common preprocessing steps	Knowing if those steps fit *your* data's generation process
Checking code correctness	Checking *conceptual* correctness

The Checklist

- Identify
  - What's missing? What's aggregated? What generated this data?
- Reason
  - What mechanism explains the pattern?
  - Do I know how this row was created? (User action? Sensor? System log?)
- Decide
  - Does my preprocessing match my theory of the mechanism?
  - If I dropna/fillna, who am I deleting? (The unhappy? The poor? The busy?)
- Document
  - Write down your assumptions so others (and future you) can disagree