When Time Really Matters: Analyzing Data in the Time of COVID

Miklós Koren (@korenmiklos)

https://economics.ceu.edu

Introduction

Can you carbon date me?

My tools

```
economics,1994-
econometrics,1996-
stata,1997-
python,2003-
julia,2017-
```

Outline

1. When time really matters
2. Examples of real-time data
3. Challenges of private data

When time really matters

When time really matters

- November 2019: outbreak in Wuhan
- December 27, 2019: new coronarivus
- December 31, 2019: WHO informed
- January 30, 2020: WHO declares “public health emergency’ ’
- March 11, 2020: WHO declares pandemic
- by March 31, 2020: most countries adopted strict social distancing measures

Typical statistics publication calendar (BLS.gov)

March, 2020

Month View | List View

Date	Time	Release
Wednesday, March 04, 2020	10:00 AM	**State Unemployment (Annual)** for Annual 2019
Thursday, March 05, 2020	08:30 AM	**Productivity and Costs (R)** for Fourth Quarter 2019
Friday, March 06, 2020	08:30 AM	**Employment Situation** for February 2020
Wednesday, March 11, 2020	08:30 AM	**Consumer Price Index** for February 2020
Wednesday, March 11, 2020	08:30 AM	**Real Earnings** for February 2020
Thursday, March 12, 2020	08:30 AM	**Producer Price Index** for February 2020
Friday, March 13, 2020	08:30 AM	**U.S. Import and Export Price Indexes** for February 2020
Monday, March 16, 2020	10:00 AM	**State Employment and Unemployment (Monthly)** for January 2020
Tuesday, March 17, 2020	10:00 AM	**Job Openings and Labor Turnover Survey** for January 2020
Thursday, March 19, 2020	10:00 AM	**Employer Costs for Employee Compensation** for December 2019
Thursday, March 19, 2020	10:00 AM	**Employment Situation of Veterans** for Annual 2019
Friday, March 20, 2020	10:00 AM	**Metropolitan Area Employment and Unemployment (Monthly)** for January 2020
Tuesday, March 24, 2020	10:00 AM	**Multifactor Productivity Trends** for Annual 2019
Friday, March 27, 2020	10:00 AM	**State Employment and Unemployment (Monthly)** for February 2020
Tuesday, March 31, 2020	10:00 AM	**Occupational Employment and Wages** for May 2019

NOTE: All times on calendar are Eastern Time.

Last Modified Date: March 13, 2020

Figure 1: BLS 2020

Time-sensitive questions

- How does the virus spread?
- How many ventilators, PPEs, nurses etc. will we need? By when?
- What (non-pharmaceutical) interventions are effective against it?
- Which of these are most cost effective?
- What can policy do to mitigate the costs?
- (in addition to genome sequencing, drug and vaccine development, clinical research)

The response of open science

The response of open science

- Government, academia and industry came together quickly and effectively. (But: pressing issues remain.)
- Troves of data shared.
- Research results published fast.
  - 83 issues of *Covid Economics*, about 500 papers published.

Is this the future of policy analysis?

About 250,000 Covid-related articles

Figure 2: Google Scholar 2021

Timely data collection

How to avoid the 2-3-month lag of official statistical releases? (Plus several months of peer review.)

Reuse existing data collected during "normal course of business'':

- administrative
- private

Examples of real-time data

Visits to retail and recreation places collapsed

Figure 3: Data from Hungarian cell phone users (Google Mobility Report 2020)

Many workplaces are shuttered

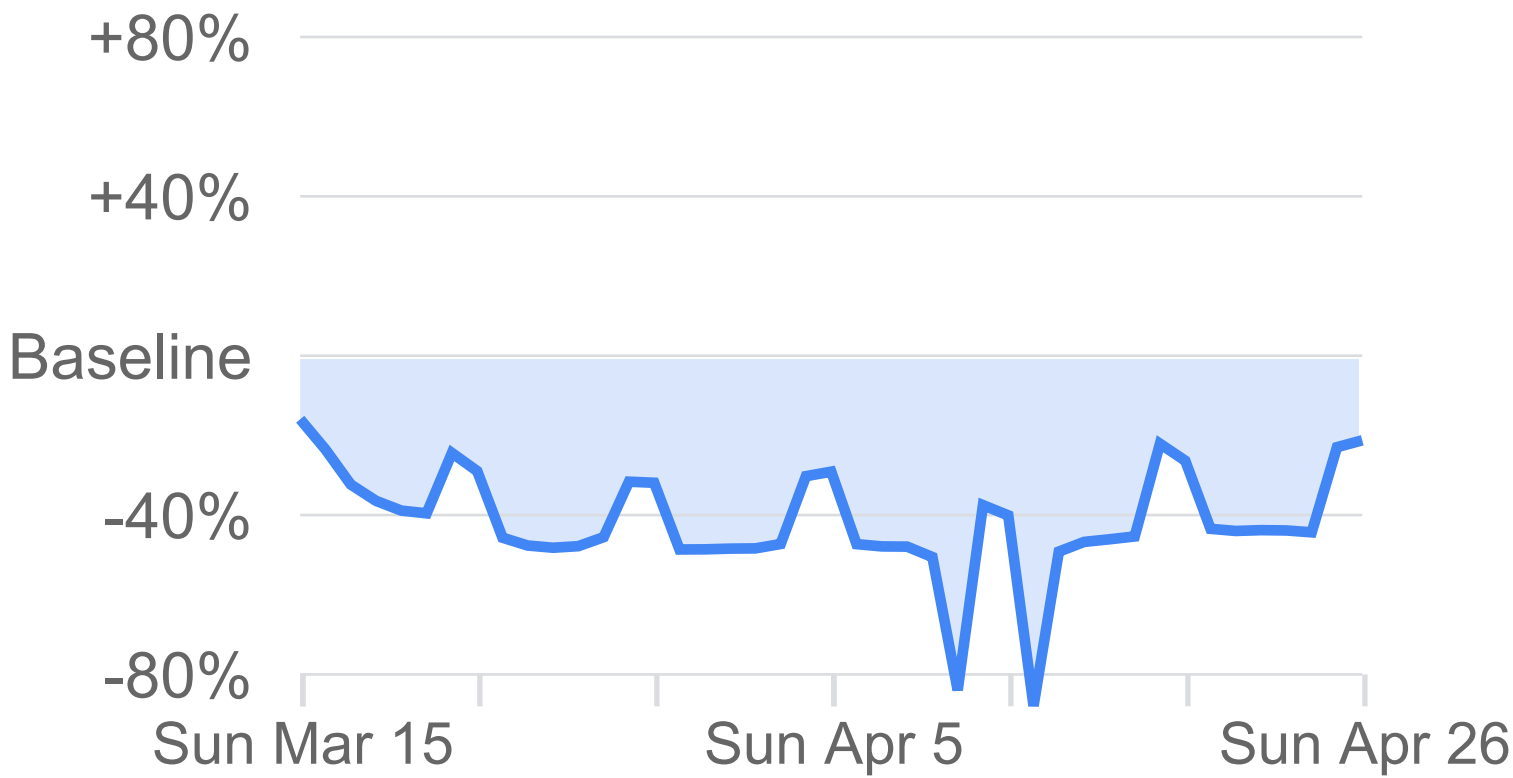


Figure 4: Data from Hungarian cell phone users (Google Mobility Report 2020)

People are staying at home

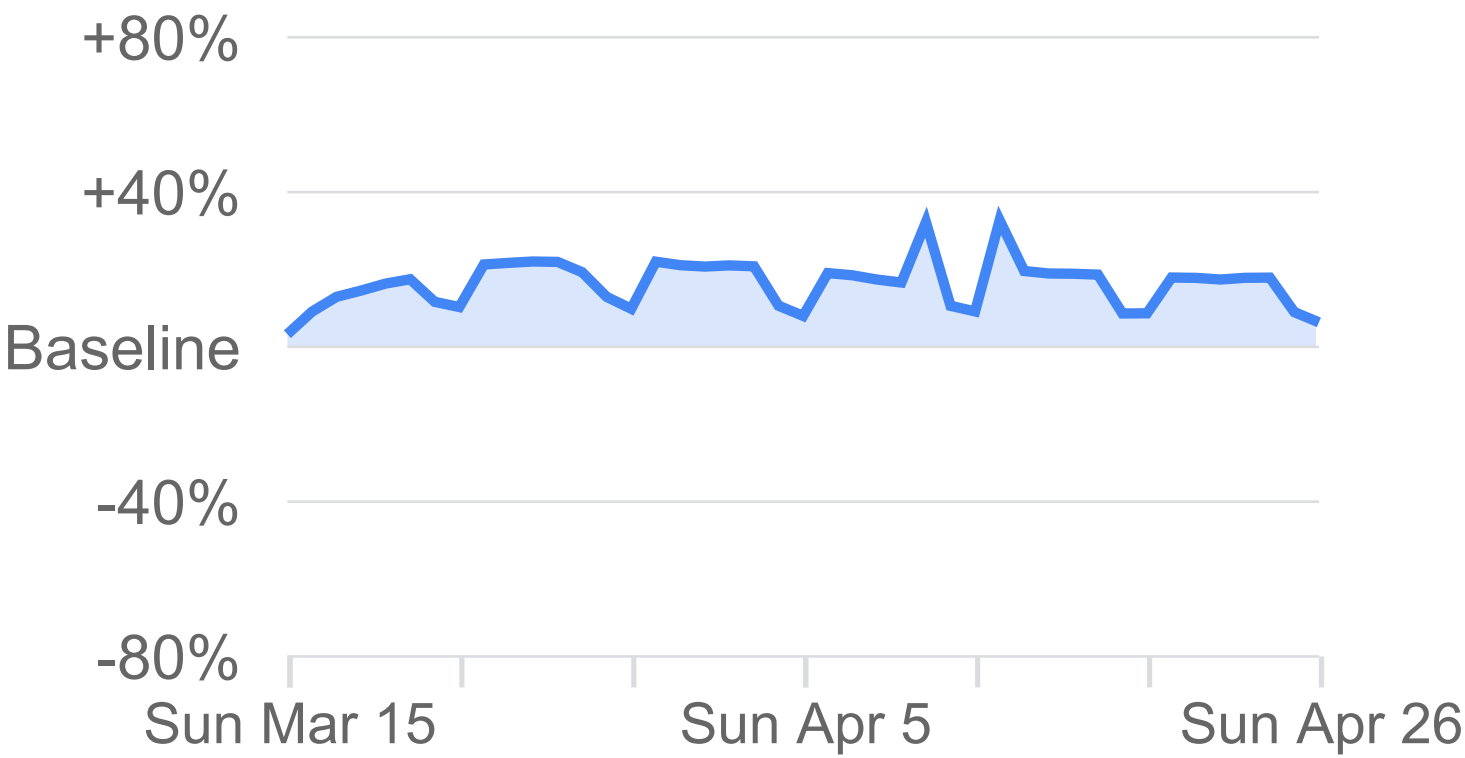


Figure 5: Data from Hungarian cell phone users (Google Mobility Report 2020)

Examples of real-time data (1)

Medical

Enormous amount of clinical, epi, virology data sharing

Stock returns

Stock prices react to news almost instantaneously. But: noisy, only for traded stocks.

Financial transactions

Credit cards. Bank transactions.

Examples of real-time data (2)

Tracking mobility, spatial effects

Cell phone tracking. Visiting POIs. Contact tracing. Air travel. Real estate pricing.

Economic activity on platforms

Restaurant closures (Yelp). Ride sharing. Airbnb. Online work. E-commerce.

Other data sources

Other data to track infections

Virus concentration in sewage.

Other data to track the economy

Electricity consumption. Job ads. Trademark applications.

Other data to track social outcomes

Religiousity. Schools and learning. Fertility. Nostalgia.

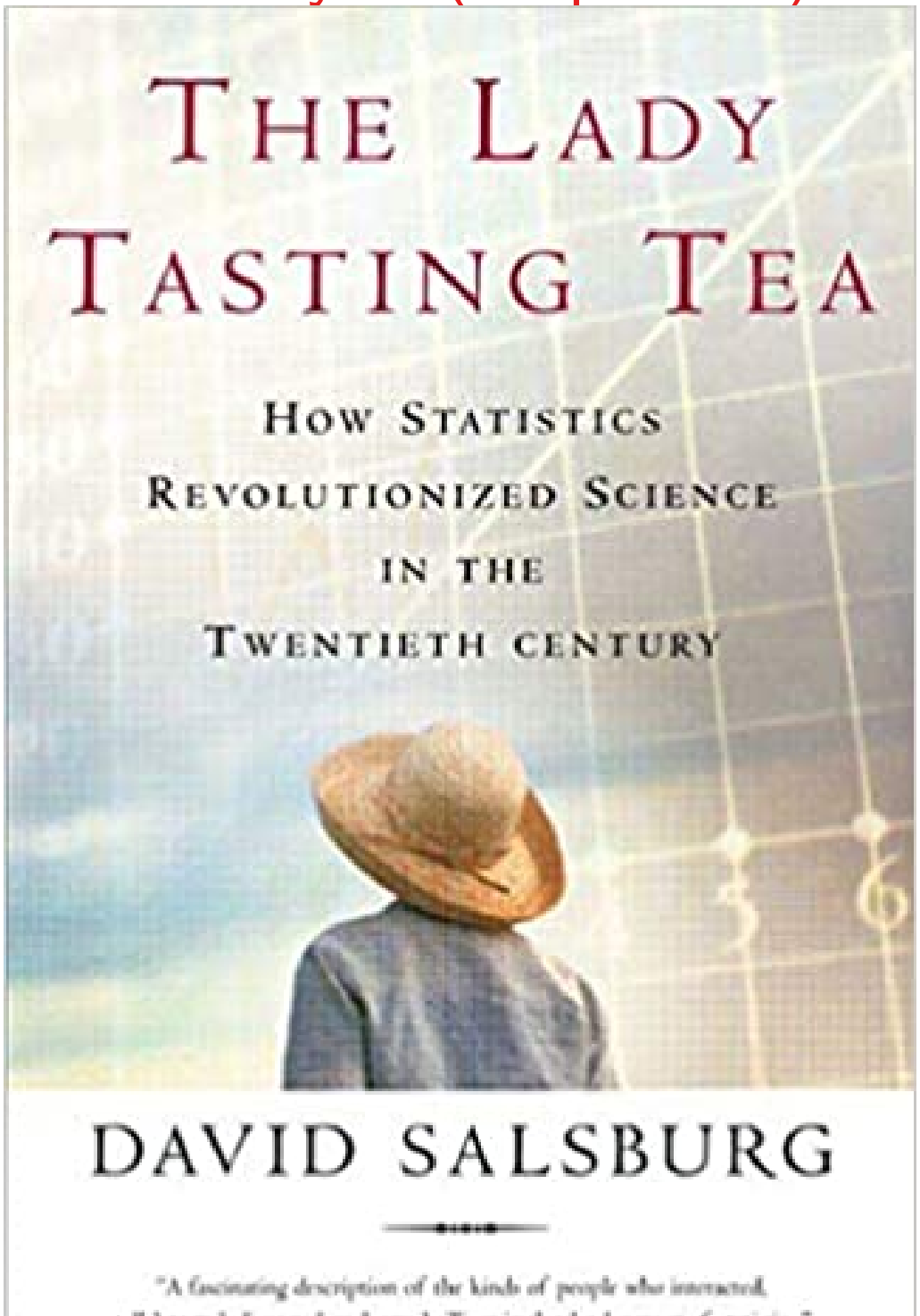
Challenges of private data

Challenges of private data

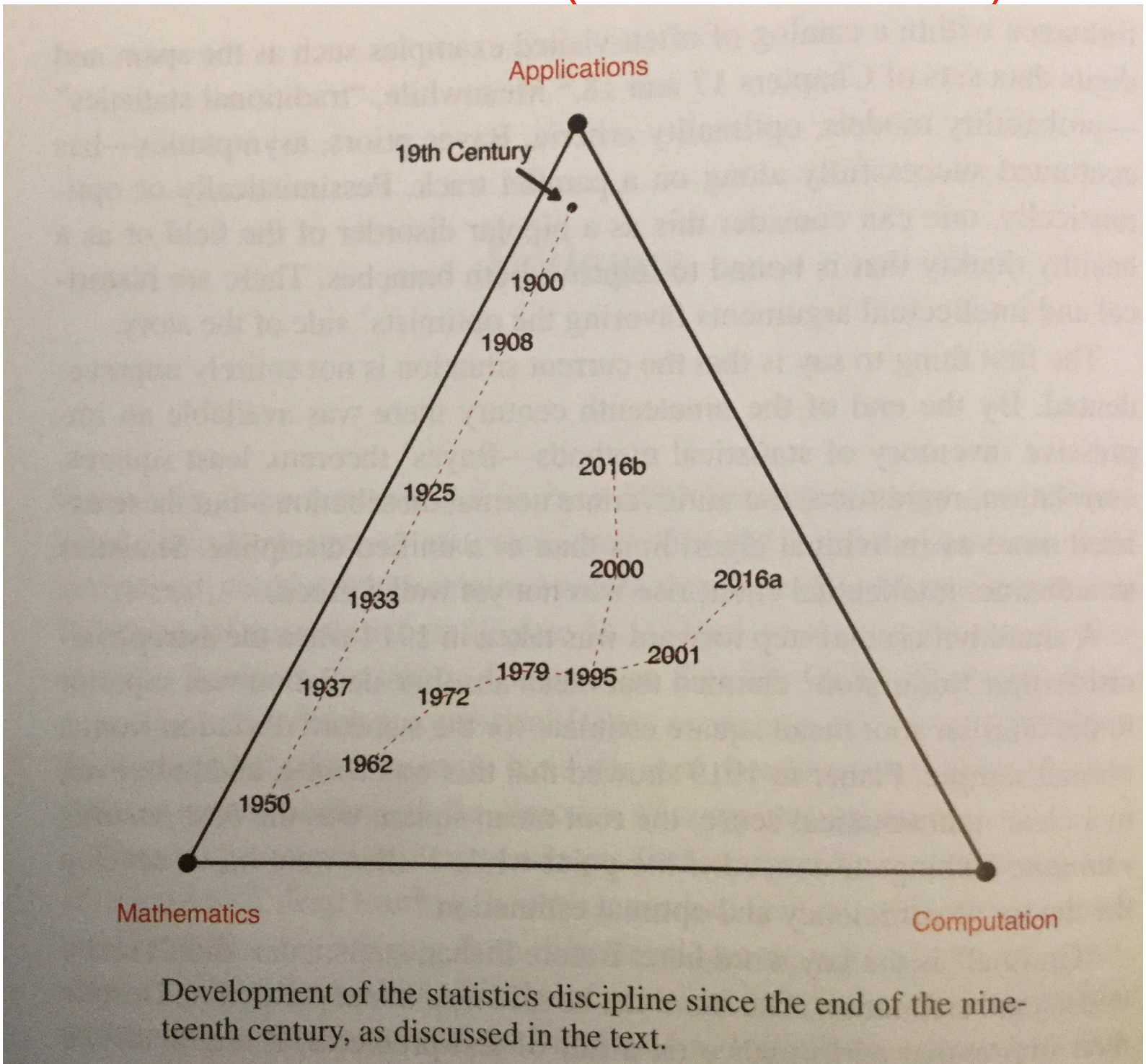
1. Statistics
2. Economics
3. Politics
4. Law and ethics

Statistics

A short history of (frequentist) statistics (Salsburg 2002)



The evolution of statistics (Efron and Hastie)



Development of the statistics discipline since the end of the nineteenth century, as discussed in the text.

Why statistics matters

Statistics provides rules for generalizing from (limited) data.

Stories vs statistics

Suppose you want to predict the outcome of U.S. presidential elections in Pennsylvania. What are the benefits of a statistical prediction relative to talking to friends and watching TV pundits?

1. $n = 1$ vs $n =$ many. ("The plural of anecdote is data.'' /Raymond Wolfinger)
2. Stories subject to biases.
3. Biases are unknown and hard to account for.

Sample vs population

Suppose you ask 1,000 Pennsylvania voters.

$$\hat{p} = \frac{\#\text{Republican}}{1000}$$

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{1000}} \approx 0.016$$

if $\hat{p} \approx 0.5$.

Rules of generalizing from sample

Suppose

1. random
2. independent sample
3. full compliance.

(1+3 ensure representativity, 2 dictates statistical properties)

- Then estimation accuracy increases with $\sqrt{n}$.
- Irrespective of size of population.

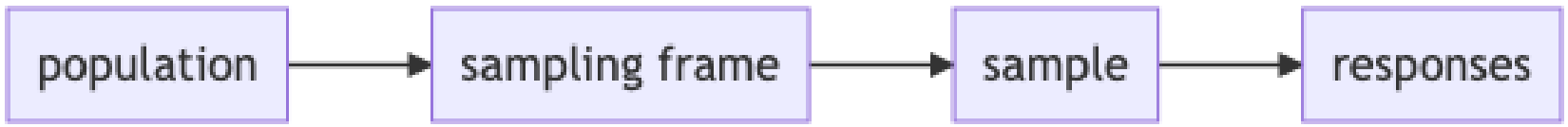
Selection bias

Selection bias

If sample is not representative, may suffer from **selection bias**.

1. nonrandom selection into sample
2. nonrandom response rate

Getting a representative sample



Selection may occur at each of these steps.

- phone survey not representative
- people do not respond
- some voters hide their preferences

A tactic to improve response rates

Control List	Treatment List
If it were up for a vote, I would vote to raise the minimum wage to 15 dollars an hour	If it were up for a vote, I would vote to raise the minimum wage to 15 dollars an hour
If it were up for a vote, I would vote to repeal the Affordable Care Act, also known as Obamacare	If it were up for a vote, I would vote to repeal the Affordable Care Act, also known as Obamacare
If it were up for a vote, I would vote to ban assault weapons	If it were up for a vote, I would vote to ban assault weapons
	If the 2016 presidential election were being held today and the candidates were Hillary Clinton (Democrat) and Donald Trump (Republican), I would vote for Donald Trump.

Sample vs big data

Why take a sample when we can study the population directly?

Electoral forecasts

- based on random sample
- based on votes already counted

Both are helpful but have very different properties.

The blue shift

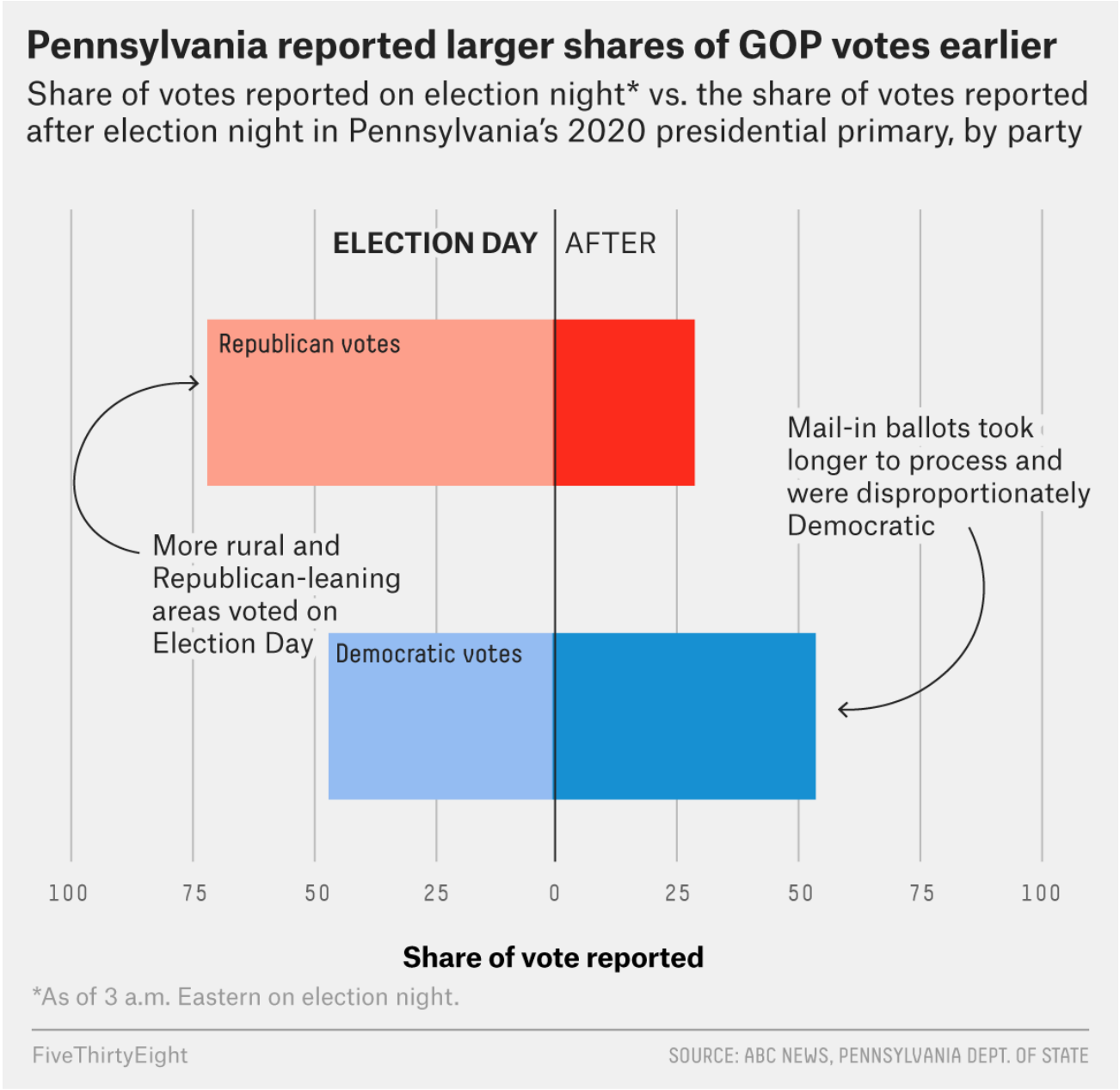


Figure 6: FiveThirtyEight 2020

Lessons from statistics

- Human judgement is necessary for good data analysis
- Understand selection bias
- Models and domain expertise matter

Economics

Why economics matters

1. People respond to incentives.
2. Systems matter.
3. Scarce resources are worth more.

The Susceptible-Infectious-Recovered model

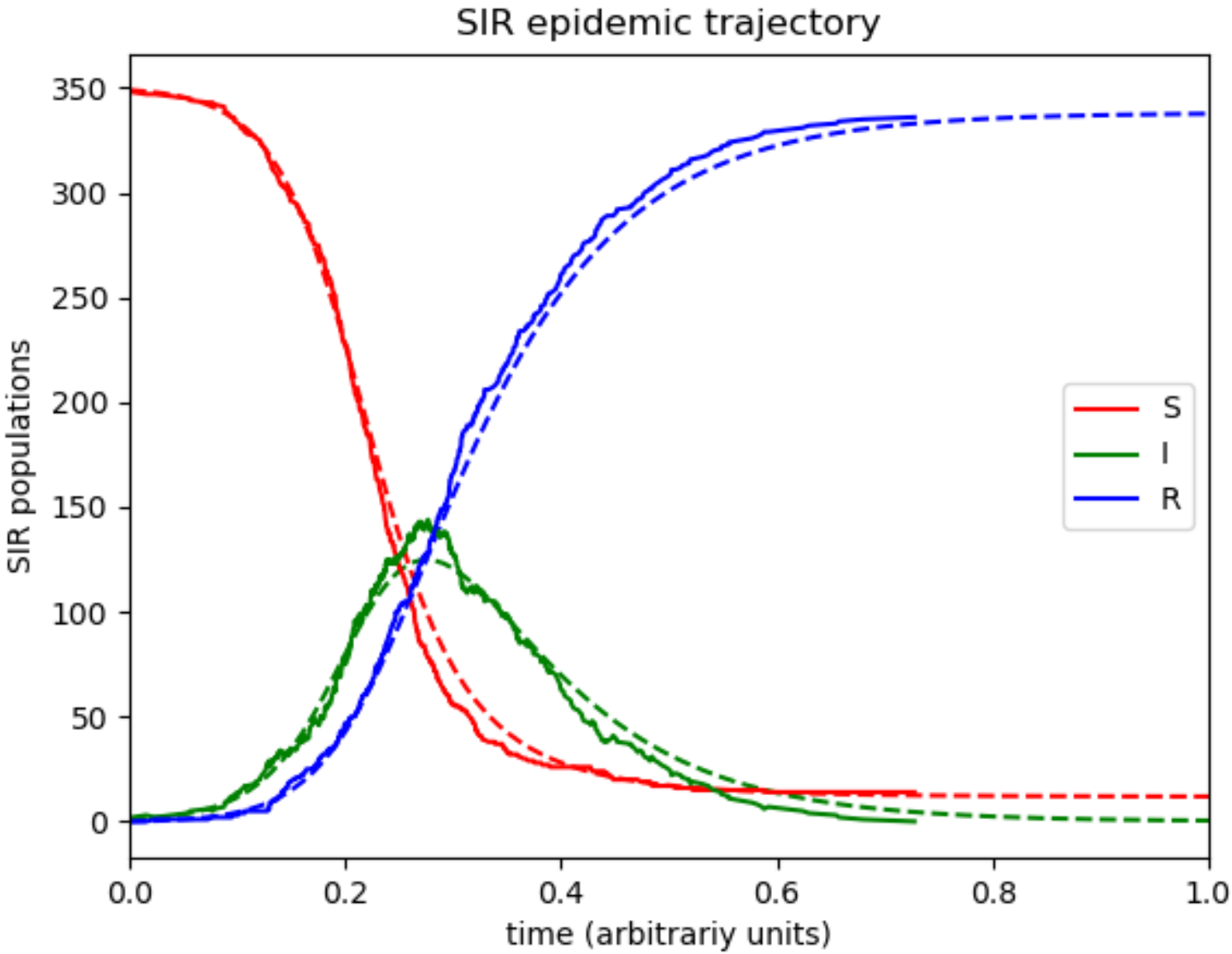


Figure 7: Wefatherley 2018

Flattening the curve

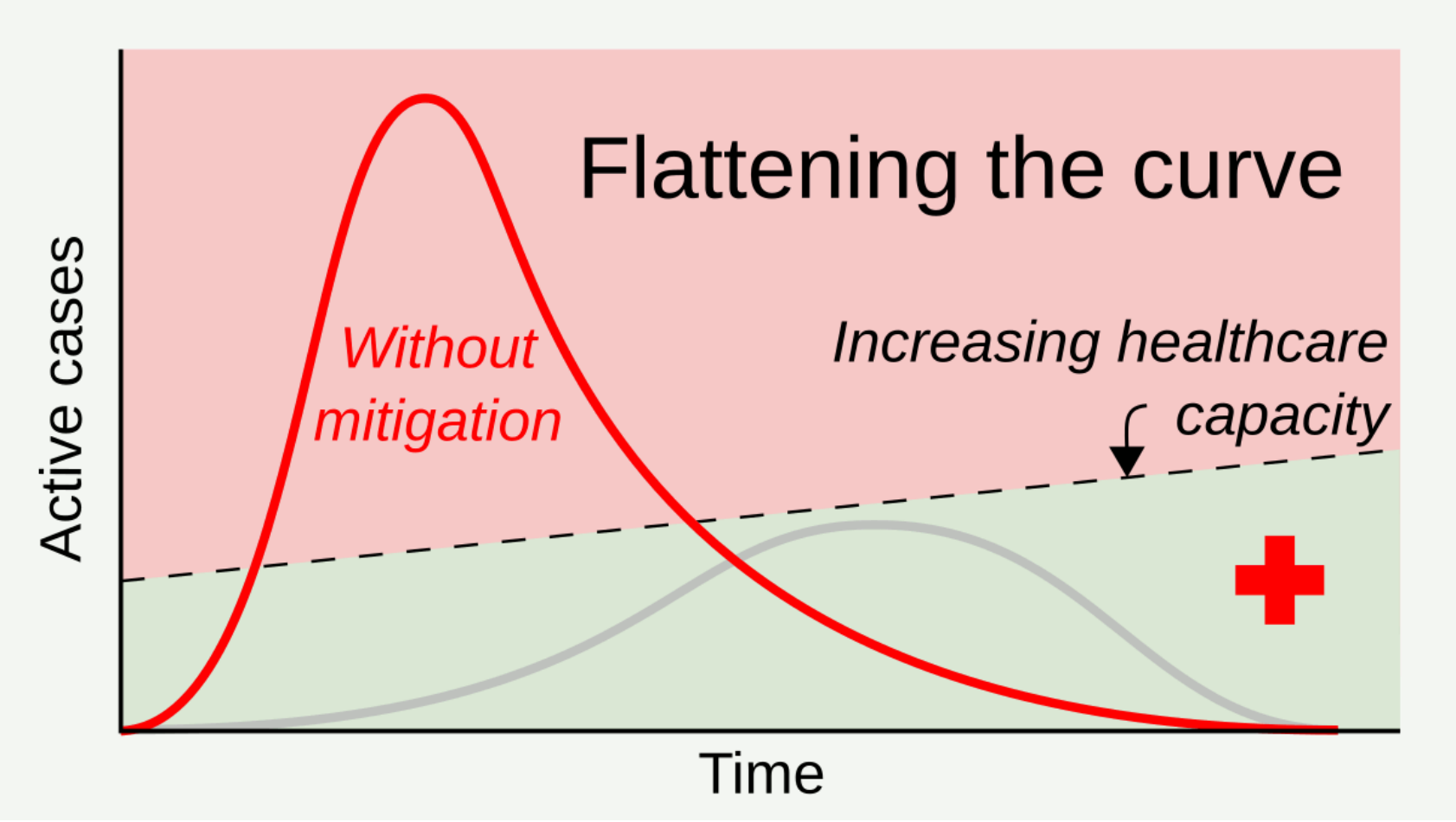


Figure 8: RCraig09 2020

Flattening the curve

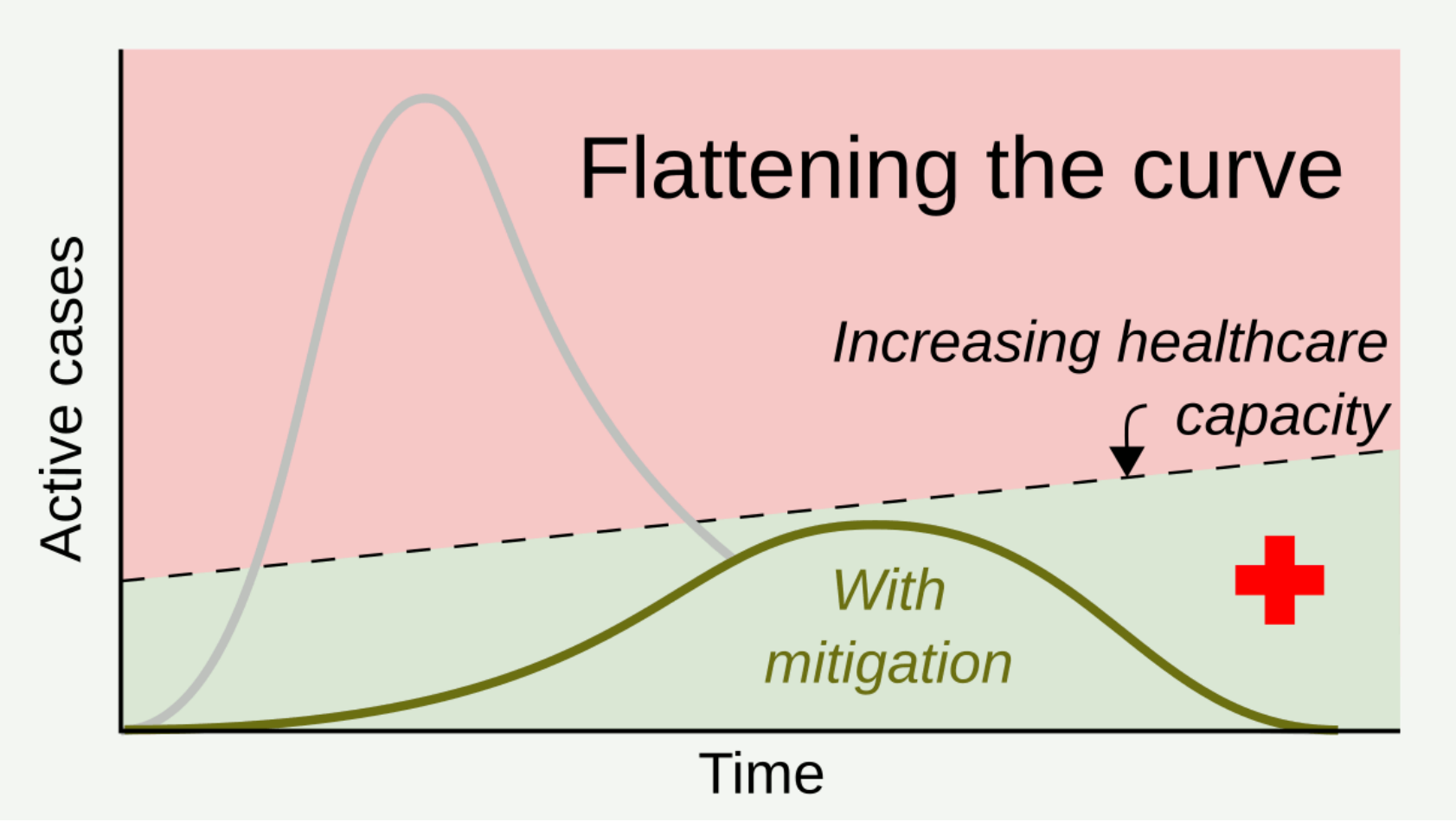


Figure 9: RCraig09 2020

People respond to incentives

- Past data may lose its predictive power once people change their behavior (Lucas critique).
  - key missing element of SIR model
- There is voluntary social distancing, as well as non-compliance with policy measures.

Systems matter

The SIR model is highly nonlinear. My getting sick depends on behavior of others.

- difficult to forecast
- externalities
- non-intuitive

Peaks of epidemics are notoriously hard to forecast

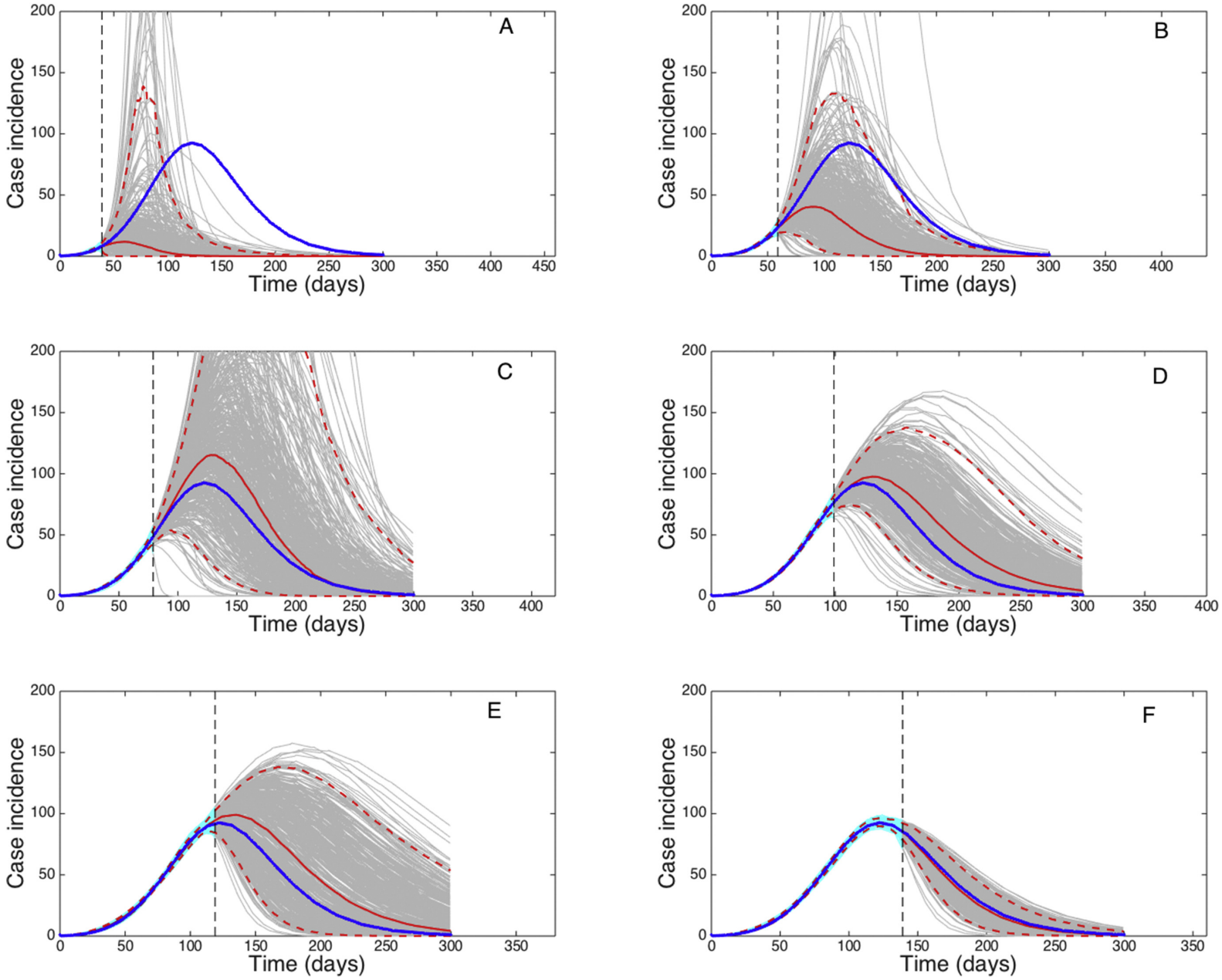


Figure 10: Chowell 2017

Lessons from economics

- Even big data not sufficient to describe *future* behavior. Understand incentives and externalities.
- Hard to forecast non-linear system without theory.

Politics, law and ethics

Politics, law and ethics

1. Conflict of interest to share information
   - governments
   - corporations
2. Privacy and surveillance

Is ride sharing killing people?

Barrios, Hochberg and Yi (2018): Uber and Lyft increased traffic and congestion. Associated with 2–3% increase in fatalities.

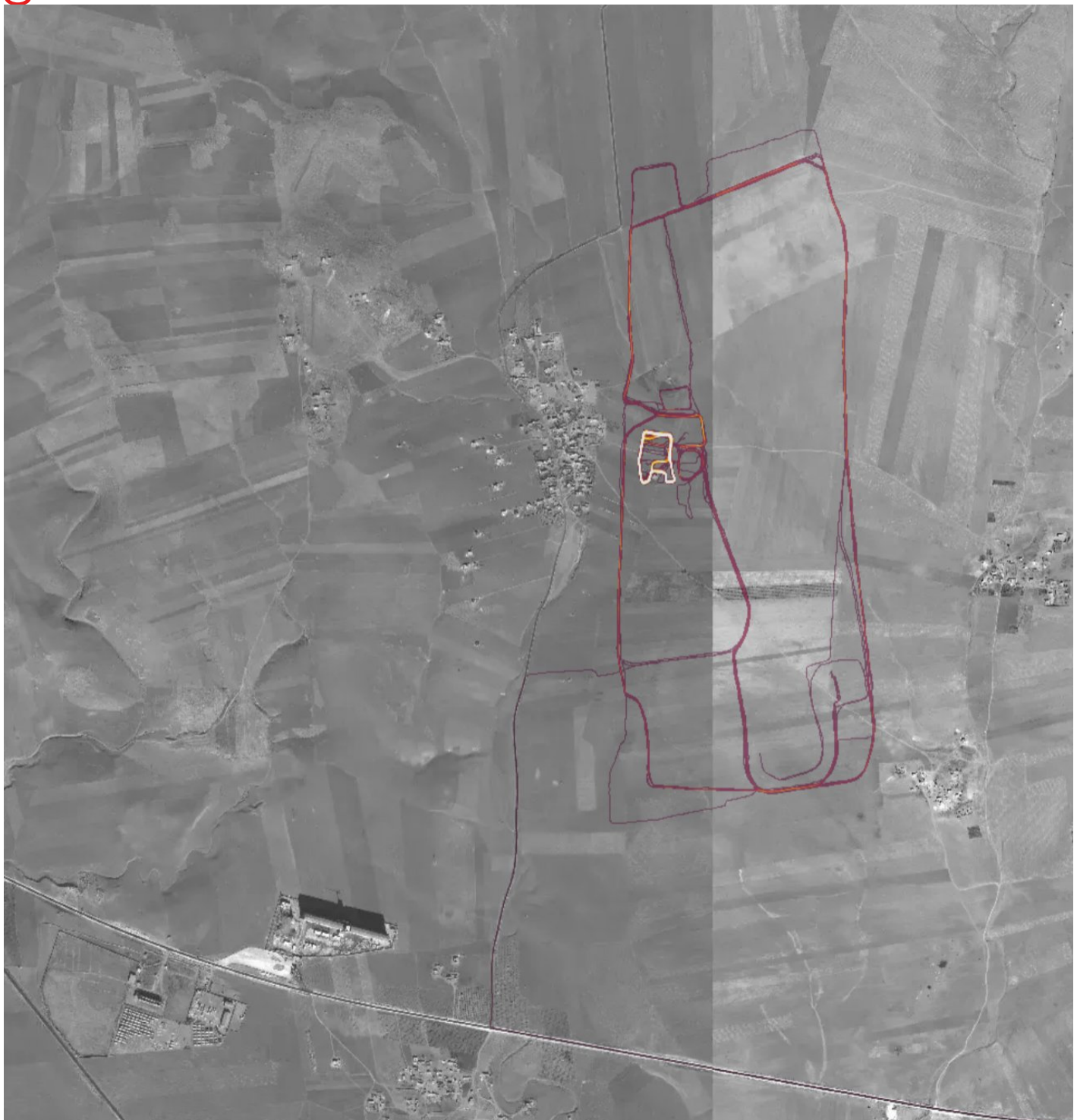
Got no data from Uber! (unlike other researchers)

Your phone knows everything about you

Thomson and Warzel (2019): Twelve Million Phones, One Dataset, Zero Privacy (New York Times)

Tracking individuals in location data dumps can (i) identify them, (ii) reveal highly sensitive information.

Mapping U.S. bases



Conclusion and discussion

Conclusion and discussion

1. Private sources of data can effectively *complement* official statistics in times of urgency.
2. But *rules* of statistics should always be followed.
3. Big data will never *substitute* domain expertise, human judgement, ethical and political accountability.