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

Miklós Koren (@korenmiklos)

https://economics.ceu.edu

Introduction

My first investment into econometrics

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
4. What can economists do?

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.

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?

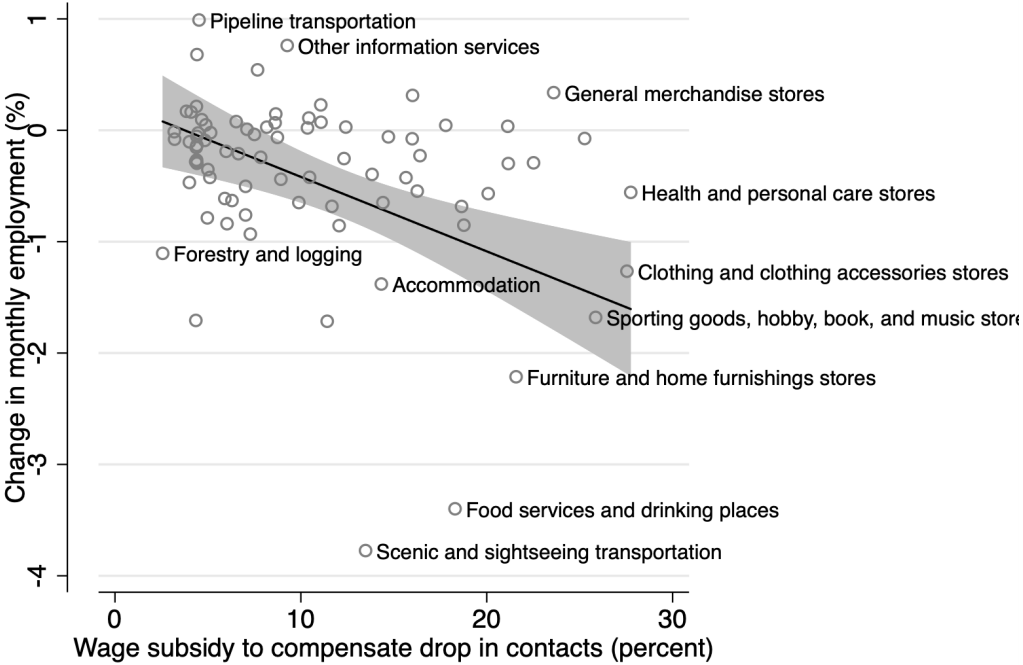
About 250,000 Covid-related articles

Figure 2: Google Scholar 2021

Our model-informed prediction based on past data (Koren and Pető 2020)

https://datawrapper.dwcdn.net/NNmIa/2/

. . . turned out to be quite accurate



Timely data collection

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

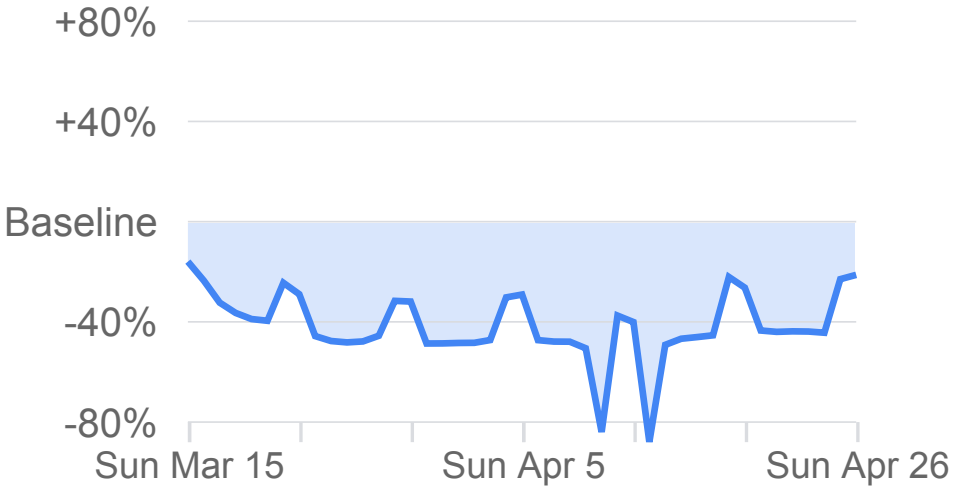
Reuse existing data collected during "normal course of business'':

- administrative
- private

Examples of real-time data

Visits to retail and recreation places collapsed

Many workplaces are shuttered



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. Accountability

Statistics

Data Science

"procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data." (Tukey,)

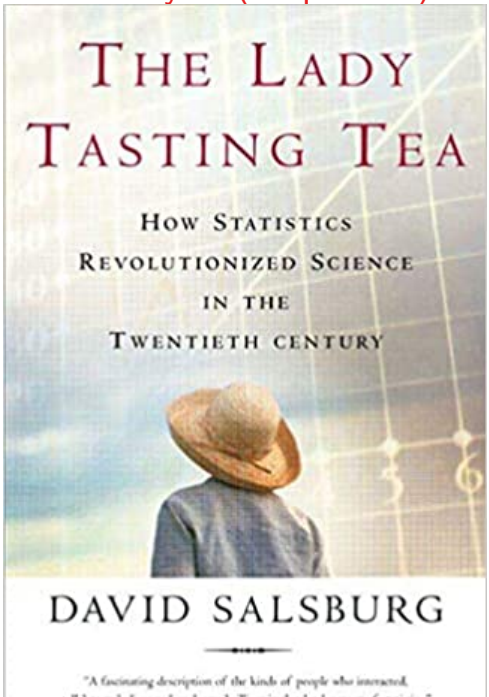
Data Science

"procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data." (Tukey, 1962)

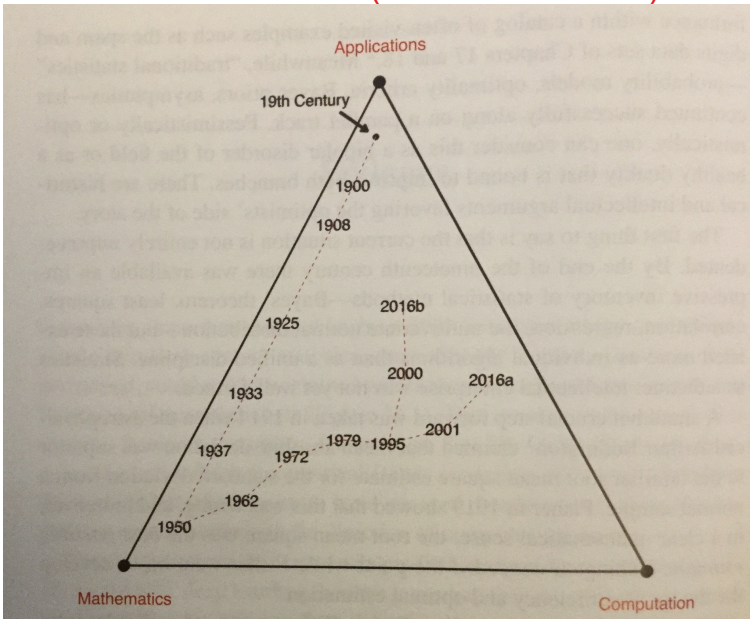
Why statistics matters

Statistics provides rules for generalizing from (limited) data.

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.

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.

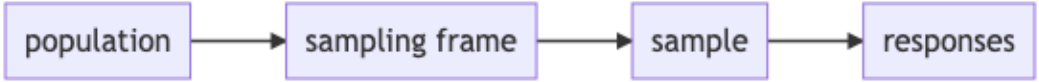
Selection bias

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

Sample vs big data

Selection bias surely does not matter if we observe (almost) everyone?!

Electoral forecasts

- based on random sample
- based on votes already counted

Both are helpful but have very different properties.

The blue shift

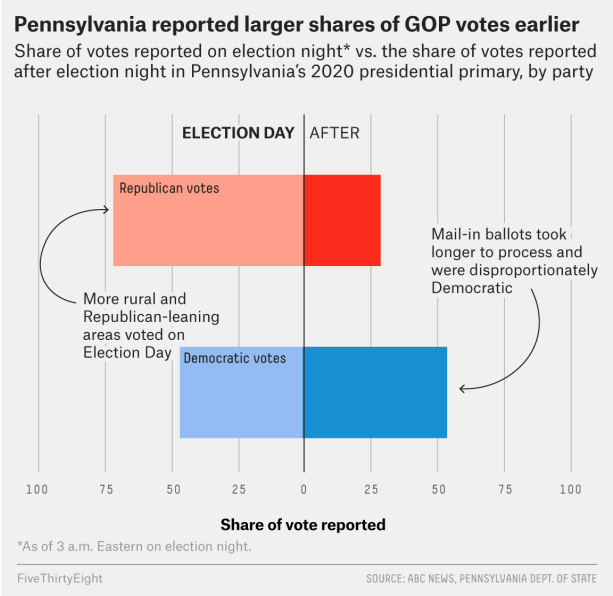


Figure 6: FiveThirtyEight 2020

Lessons from statistics

It is better to have a small unbiased sample than a large biased one.

Can you think of sources of selection bias in private data?

Accountability

Accountability

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

Uber uses data and economists as PR props

"Ride-hailing apps have created jobs for Paris's poorer youth, but a regulatory clampdown looms," the [FT] article said. Thesmar was quoted in the piece saying that Uber was a "social gamechanger".

"We see low risk here because we can work with Landier on framing the study and we also decide what data we share with him." (senior Uber staffer quoted in Lawrence 2022)

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!

A case study in accountability

Simonsohn, Simmons, Nelson and anonymous (2021) show that Shu, Mazar, Gino, Ariely and Bazerman (2012 PNAS) is based on **fraudulent** data.

RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES |

Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end

Lisa L. Shu, Nina Mazar, Francesca Gino, Dan Ariely, and Max H. Bazerman -1 Authors Info & Affiliations

Edited by Daniel Kahneman, Princeton University, Princeton, NJ, and approved July 23, 2012 (received for review June 11, 2012)

August 27, 2012 | 109 (38) 15197-15200 | https://doi.org/10.1073/pnas.1209746109

THIS ARTICLE HAS BEEN RETRACTED + VIEW RELATED CONTENT +

The data as (purportedly) shared with the private company

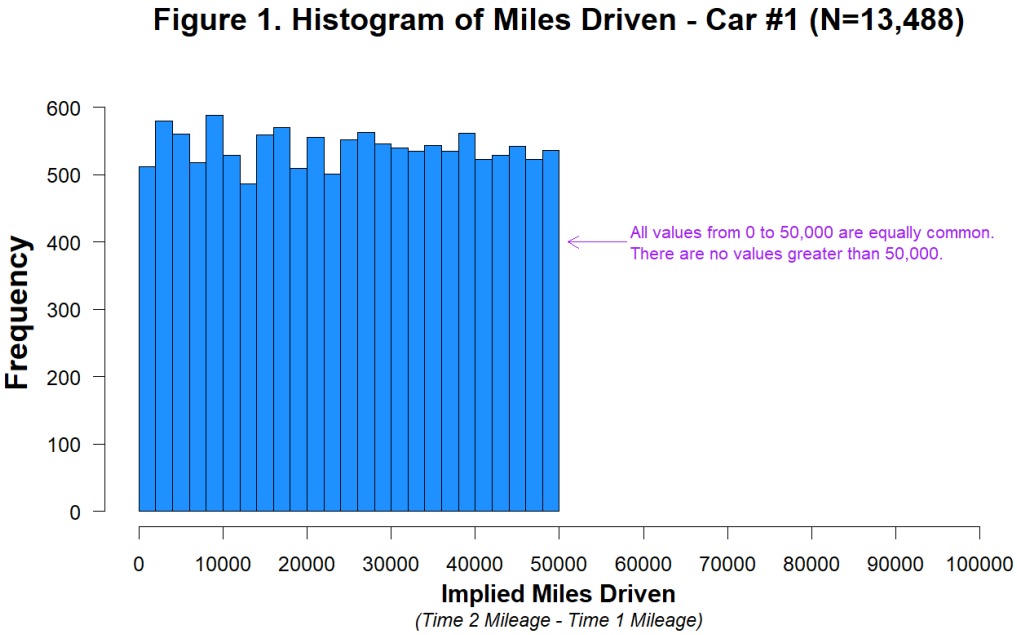
DrivingdataAll with font.xls - Compatibility Mode

Search

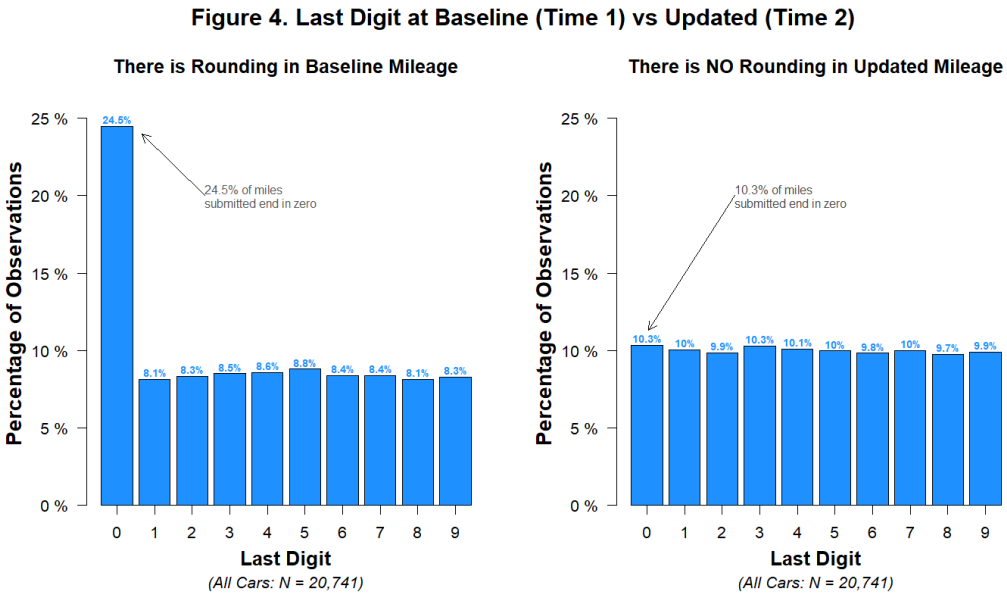
File Home Insert Insert 2 Draw Page Layout Formulas Data Review View Developer Help Acrobat

	A	B	D	E	F	G	H	I	J	K	L	M	N
1	condition	id	baseline_car1	update_car1	baseline_car2	update_car2	baseline_car3	update_car3	baseline_car4	update_car4	baseline_average	update_average	diff_average
2	Sign Top	1	896	39198							896	39198	38302
3	Sign Bottom	2	21396	63511	32659	47605					27027.5	55558	28530.5
4	Sign Bottom	3	21340	37460	44998	59002					33169	48231	15062
5	Sign Bottom	4	23912	59136							23912	59136	35224
6	Sign Bottom	5	16862	59292							16862	59292	42430
7	Sign Top	6	147738	167895	125820	164688					136779	166291.5	29512.5
8	Sign Bottom	7	18780	49811	45402	54824					32091	52317.5	20226.5
9	Sign Top	8	41930	80323	181416	229852					111673	155087.5	43414.5
10	Sign Top	9	28993	63707	13291	28165					21142	45936	24794
11	Sign Bottom	10	78382	127817							78382	127817	49435
12	Sign Top	11	58500	81081							58500	81081	22581
13	Sign Bottom	12	99417	149211	48770	95179	72620	115338			73602	119909	46307

Distribution of miles driven in a year



No rounding in end-of-year reported mileage



Most observations seem to be duplicated

DrivingdataAll with font.xls - Compatibility Mode

File Home Insert Insert 2 Draw Page Layout Formulas Data Review View Developer Help Acroba

	A	B	C	D	F	H	J
1	condition	id	font	baseline_car1	baseline_car2	baseline_car3	baseline_car4
2	Sign Top	12938	Cambria	983155			
3	Sign Top	13146	Calibri	982573			
4	Sign Bottom	12065	Cambria	735965	100512	163756	
5	Sign Bottom	5999	Calibri	735451	99735	163390	
6	Sign Bottom	12843	Cambria	603001	153284	130947	153254
7	Sign Bottom	5442	Calibri	602368	152327	130210	152600
8	Sign Bottom	767	Cambria	463284			
9	Sign Bottom	11557	Calibri	463090			
10	Sign Bottom	6120	Cambria	444290			
11	Sign Bottom	7357	Calibri	443920			
12	Sign Bottom	2324	Cambria	417041	48826	119477	
13	Sign Top	6297	Calibri	416537	48813	118579	
14	Sign Top	1895	Cambria	409663	31578	95013	
15	Sign Top	3821	Calibri	409515	31134	95000	
16	Sign Top	4819	Cambria	403733			
17	Sign Top	10804	Calibri	402847			
18	Sign Top	10181	Cambria	395272			
19	Sign Top	10650	Calibri	394482			
20	Sign Bottom	12845	Cambria	365387	112247	49086	
21	Sign Bottom	10362	Calibri	364774	112123	48472	
22	Sign Bottom	5117	Cambria	359700			
23	Sign Bottom	3779	Calibri	359641			

sorted

The chain of data provenance

insurance company $\rightarrow$ Ariely $\rightarrow$ Mazar $\rightarrow$ PNAS

What can economists do?

What can economists do?

Three tenets of economics:

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

The Susceptible-Infectious-Recovered model

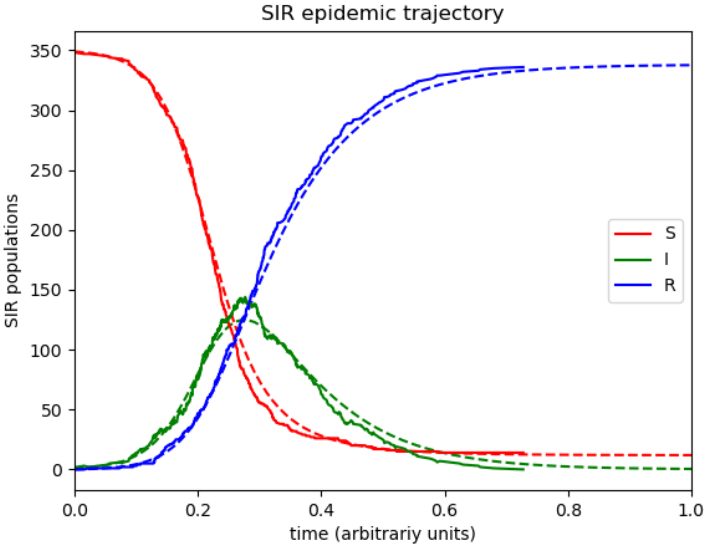


Figure 7: Wefatherley 2018

Flattening the curve

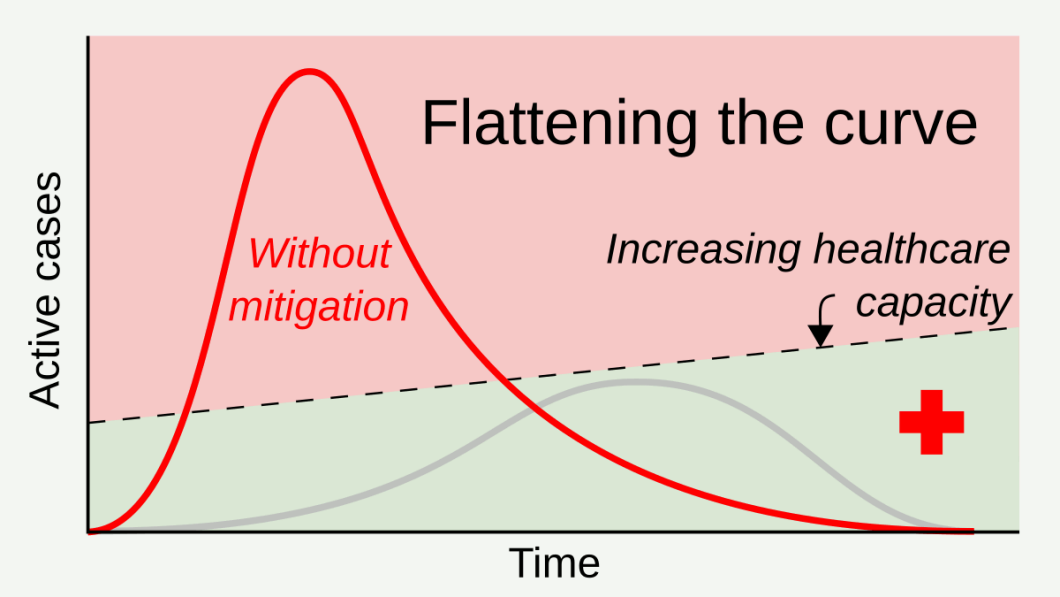


Figure 8: RCraig09 2020

Flattening the curve

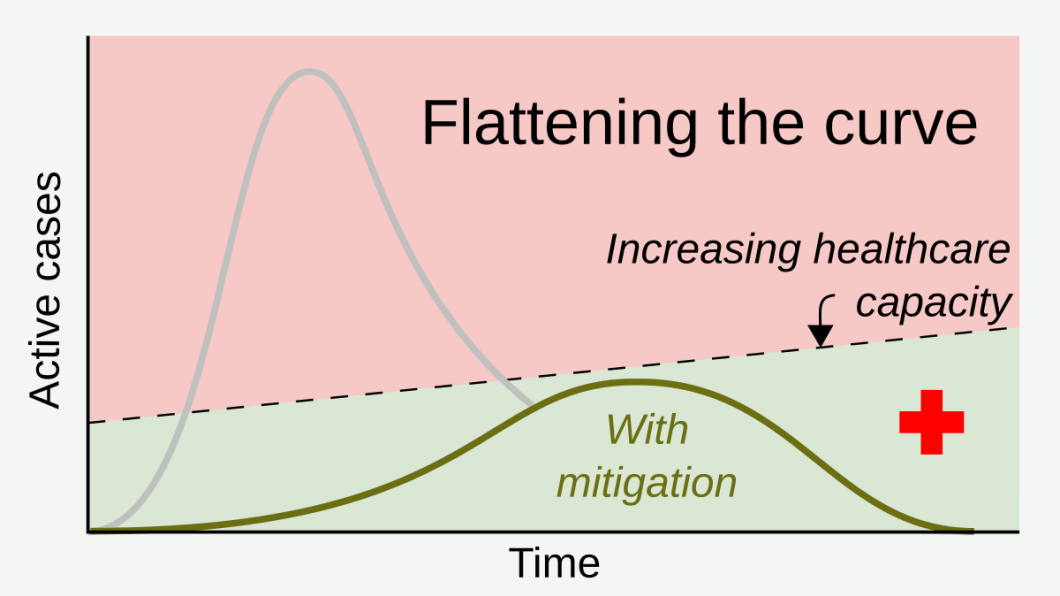


Figure 9: RCraig09 2020

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

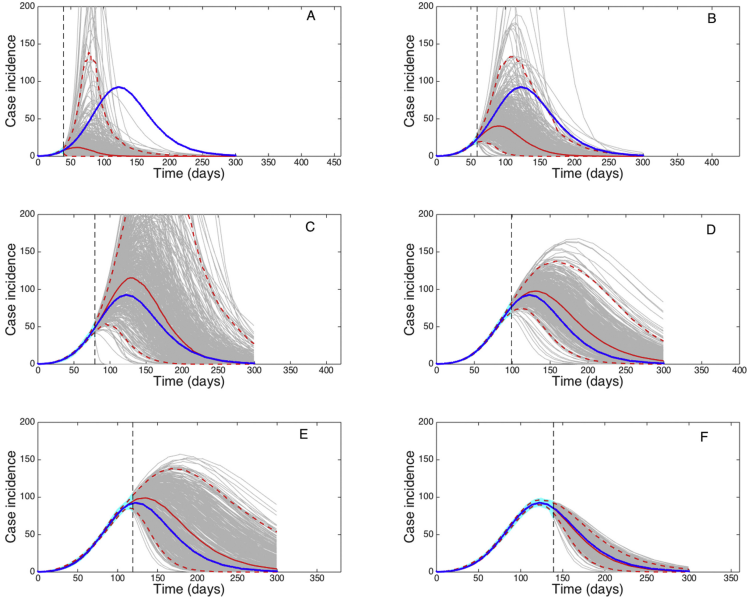


Figure 10: Chowell 2017

Lessons from economics

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

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.