Pitfalls of Big Data

PROFESSOR VICKI BIER
INDUSTRIAL AND SYSTEMS ENGINEERING
UNIVERSITY OF WISCONSIN-MADISON

Despite the major advantages of big data,

THE USUAL CAUTIONS STILL APPLY!!!

Especially in safety-critical areas

Big Data Can End Up Being …SMALL DATA!

Example: Searching a large data base for the few cases with a particular combination of diseases

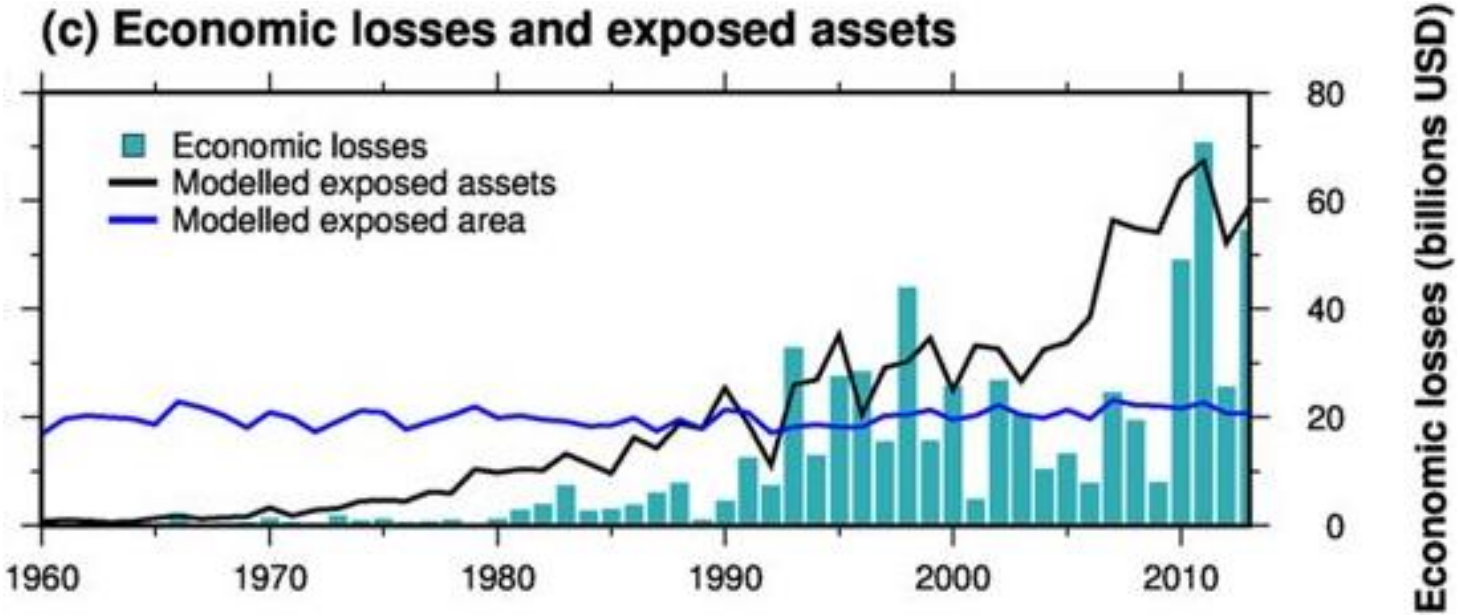
"There are a lot of small data problems that occur in big data. They don't disappear… They get worse"

--Spiegelhalter (2014)

- **Decisions made on the basis of such data could end up being based on only a handful of cases:**
  - **And could then eliminate the requisite variety to learn from experience in future.**

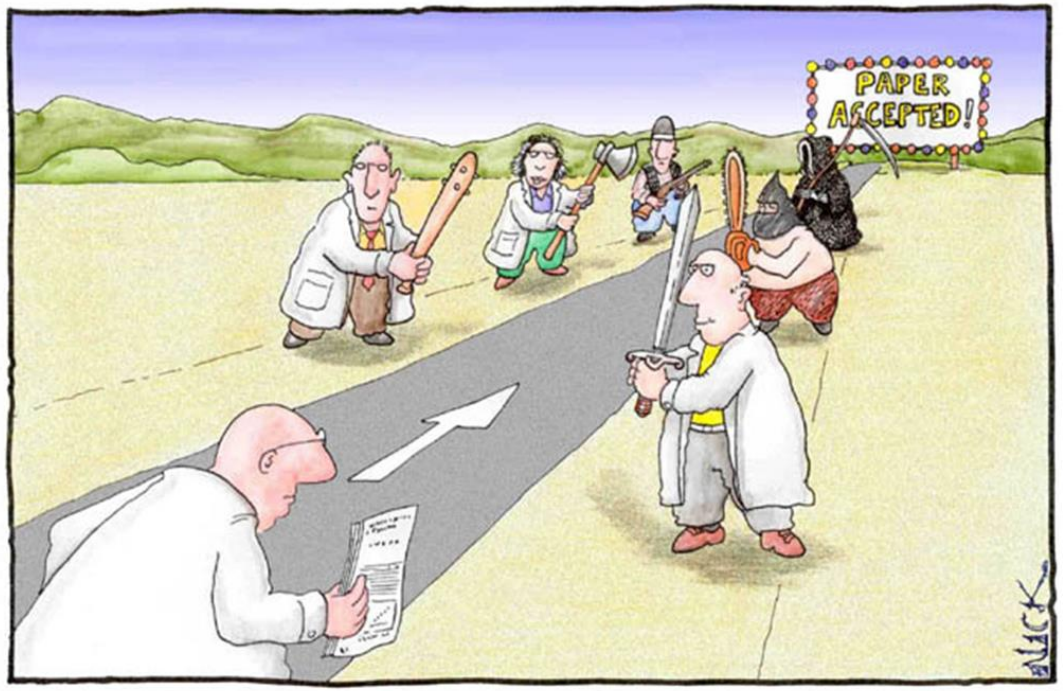
TIME SERIES

- Also any two variables with monotonic trends will be correlated
- E.g., income, education, technology



NEED FOR VALIDATION

- E.g., experiments (try multiple approaches on a limited number of cases to confirm)
- Splitting data into training sets and test sets



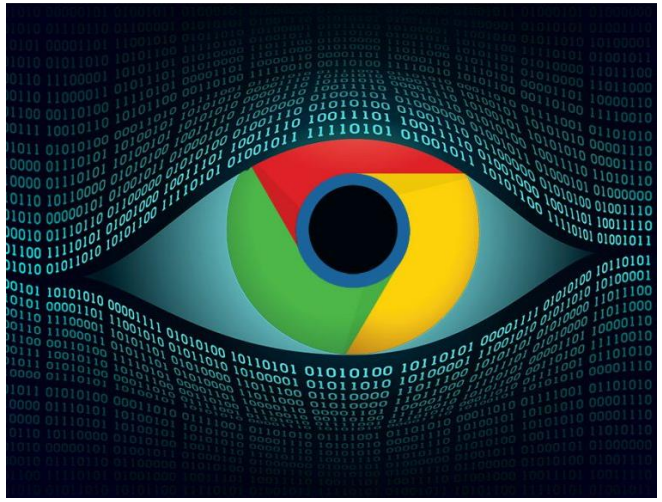
Most scientists regarded the new streamlined peer-review process as 'quite an improvement.'

FALSE POSITIVES

Another pitfall of big data! Ioannides (2005): "Most Published Research Findings Are False"!

- **Example: Rate of automobile accidents**
- Higher with consumption of beer or hard liquor
  - Lower with consumption of wine!
- Lower in 9th month of pregnancy (might make sense)
  - Also in 3rd or 6th month! (almost certainly a false positive)

PRIVACY ISSUES



Another problem with false positives:

- If there are only a few actual bad guys in the data base
- ***Many of the bad guys identified will be false positives!***

- **Note that being "identified" can get you on the "no-fly list":**
  - Or worse!

MORE PRIVACY ISSUES

"Anonymous" data may not really be anonymous!

- https://www.theguardian.com/technology/2017/aug/01/data-browsing-habits-brokers
- "A mere 10 URLs can be enough to uniquely identify someone"
- "Just think, for instance, of how few people there are at your company, with your bank, your hobby, your preferred newspaper and your mobile phone provider"

TOO UNIQUE TO HIDE

(ROCHER ET AL., 2019)

62% chance of being identified based on birth date, gender, and zip code:

Rises to 99% based on my being a state government employee!

- https://cpg.doc.ic.ac.uk/individual-risk/

SOLUTIONS

Ways to characterize data

(Cox, 2013)

- **Adjustments for multiple comparisons.**
- **Granger causality.**
- **Conditional independence tests.**