Day Three, part B
Be careful out there folks!

Jon Atwell & Christopher Skovron
Northwestern University

June 20, 2018

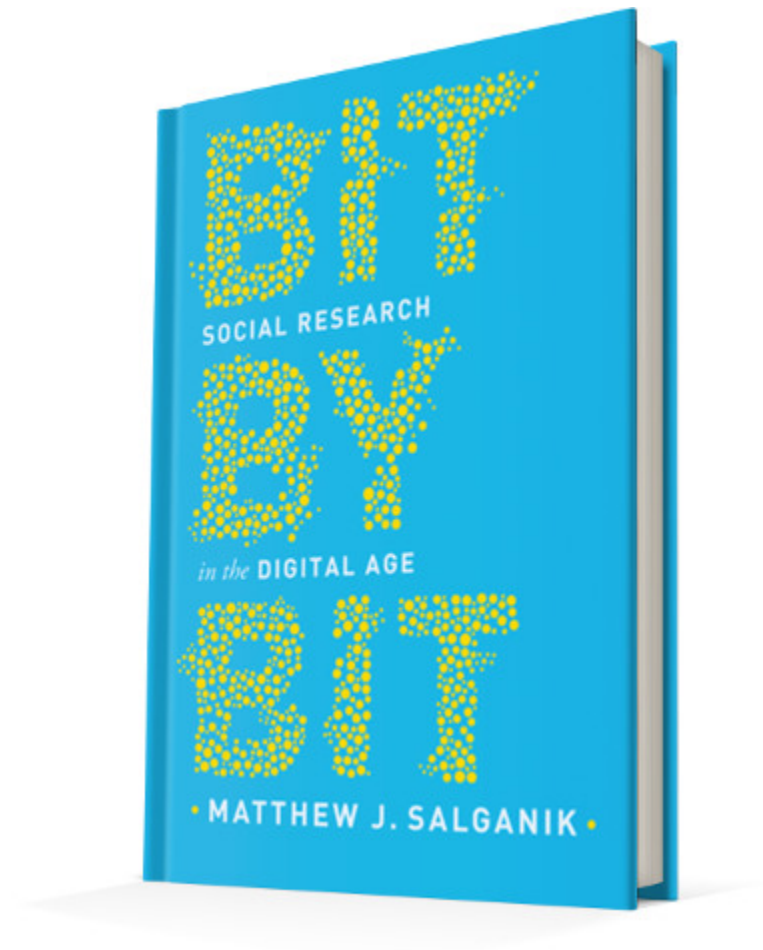
Get giddy about the possibilities!

$$\text{count\_problems} = G(\text{ data\_size}^{\alpha}), \alpha > 1?$$

Probably not, just different problems

- Pitfalls of Inference
- Ethical/legal considerations
- Practical considerations

Salganik's Characterists of Big Data



(PUP, 2018)

Characterists BD **can** have

- **Huge N**
- Not created for science*
- Always-on
- Nonreactive
- Incomplete
- Inaccessible
- Nonrepresentive
- Drifting
- Algorithmically-confounded
- Dirty
- Sensitive

Big Data is indeed big, but is correlation is enough?

- whole of sample frame is not the population
- systemic bias instead sampling bias
- P-hacking, meet N-hacking: 6:06 mark
- More about new types of data, or richness of context

Not created for science

- “digital exhaust”
- weak link between **construct** and **measures**

Always-on

- Old: HS diploma to job at 28
- New: Location every 15 minutes.
- But: When is this informative? Are we answering different questions?

Nonreactive

- Old: Generosity in lab studies (reactivity/Hawthorne effect)
- New: Don’t know about observation, or is now normalized
- But: Have you been on Instagram?

Incomplete: If only I had:

Incessible

- Private companies own it
- Governments collect it but don’t share

Nonrepresentative

- Who actually tweets anyway?
- But nonrandom sample can still be very useful!

Drifting

- population
- behavior
- system

Algorithmically confounded

- unique experiences (N treatment groups)
- recommenders
- Matthew effect/increasing returns/preferential attachment
- action triggers

Bigger question: Is social life now algorithmically confounded?

Lack of treatment controls is one thing, but exposures compound to substantial change.

- those confounds create new reality
- Facebook vote encouragement
- LinkedIn recommendations

Dirty

- Wait, Russians can use Twitter too?!
- Wait, computers can do automated tasks?!

Sensitive

- More so than you might think: Meta data
- Merging or cross referencing data sets, especially so

Ethical and legal considerations

Ethical

- Loss of Privacy (See Netflix Prize)
- Violation of trust (See Cambridge Analytica)
- Exploitation
- Lack of informed consent

Legal Pitfalls

- Violation of User Agreements (Sandvig vs Sessions)
- Impersonation
- Deception

Legal hangups

- Formal data use agreements can happen without scientific input.
- Can take a long time
- Can be expensive

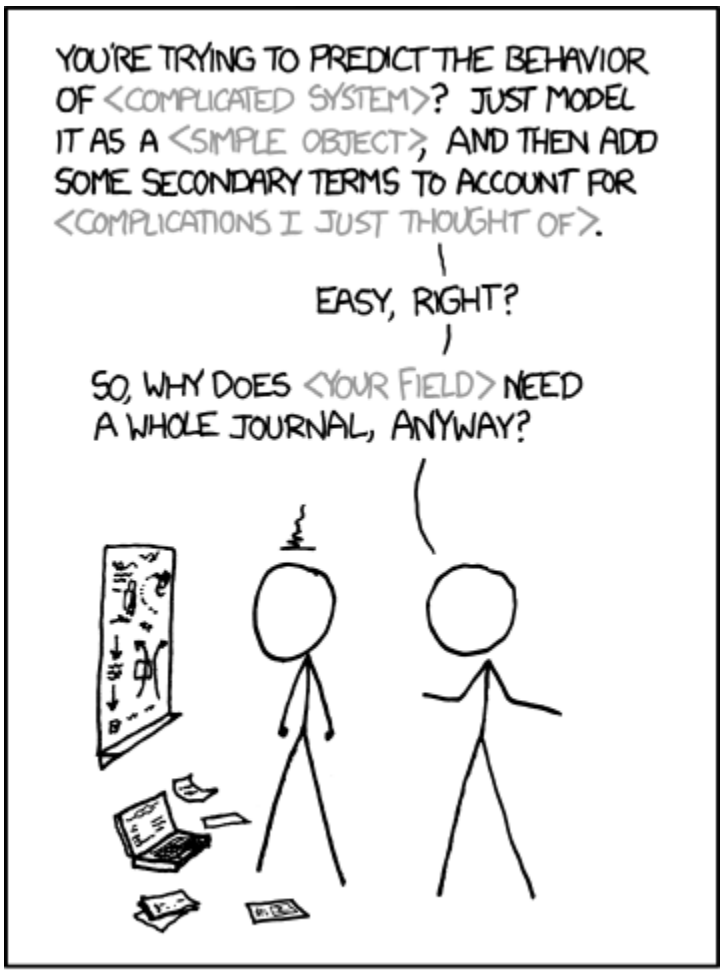
Paging IRB!

- behind the times, don’t understand tech systems
- assumes you’ll follow UA?

Practical considerations

You might have to read more widely

- physicists doing S.S. in *Science* and *PNAS*, etc.



You have to move quickly

- Small tweaks to websites can break scripts
- Data are more widely available, more people study them

But you might have to move slowly too

- harvesting data can take time
- stay off server blacklists

It's important to start with observation

Platforms support actions that:

- go unused
- get co-opted

It's important to start with observation

Platforms can have unique:

- norms
- terms
- meanings for words
- …?

Scraping just ain't what it used to be

- Most webpages are dynamic and idiosyncratic
- Server-side ops/parameters are intentionally opaque

You can ask for data from corporations and other orgs

- think of quid pro quo
- Use LinkedIn to find lower level people who might care
- use email finders or domain hack address
- repeat

You can use Mturk (or its mirrors)!

- cheaply label training set
- cheaply validate results
- but I'm skeptical of survey results

And yet, it's an exciting time to be doing social science!