The Application of Knowledge

With Big Data Comes Big Responsibility...

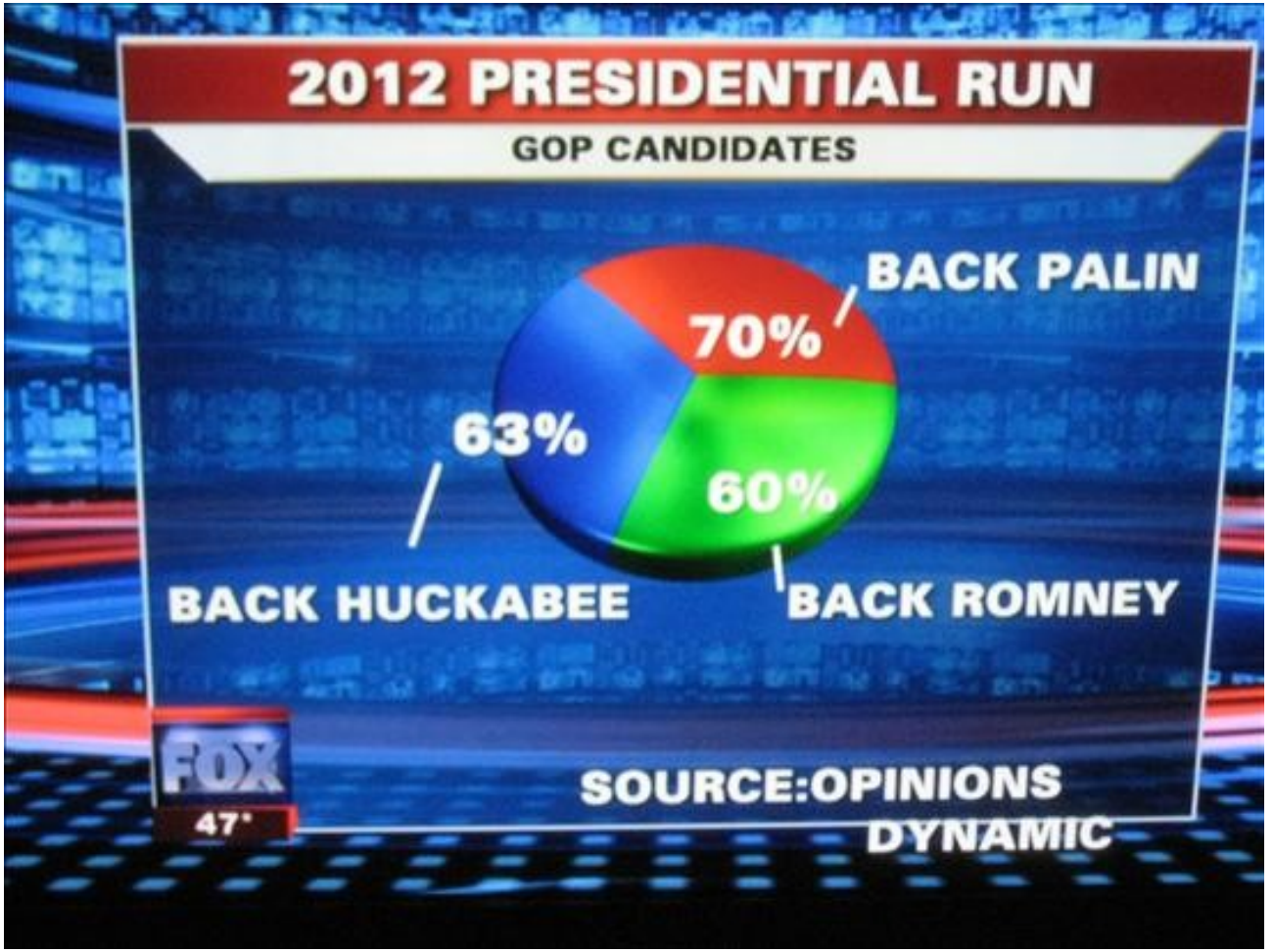
(With bad references comes lecture material)

With Big Data Comes Big Responsibility...

(With bad references comes awesome lecture material)

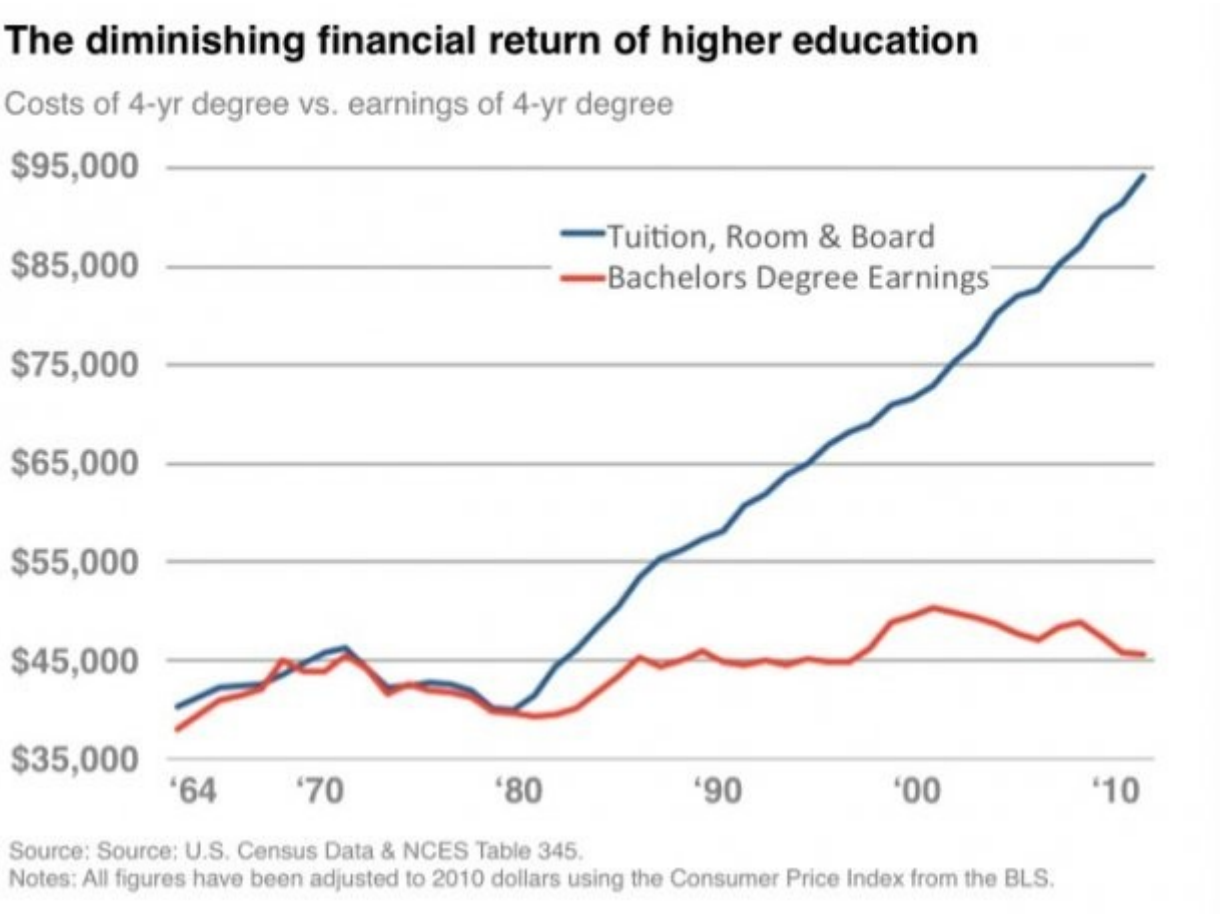
With Big Data Comes Big Responsibility...

(With bad references comes awesome lecture material)



Control over data grants...

- The ability to analyze the data to tell *your* story.



Control over data grants...

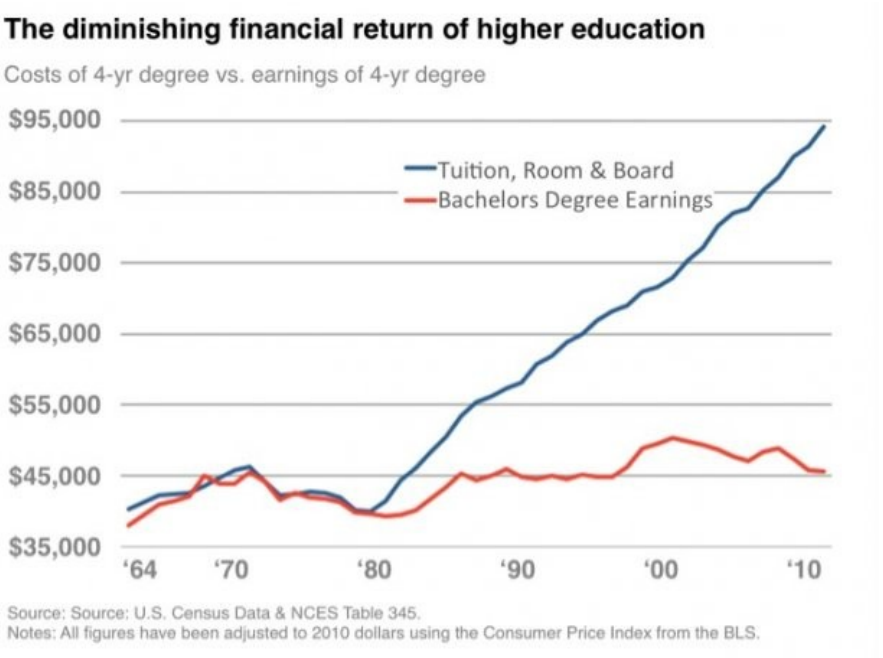
- The ability to analyze the data to tell *your* story.

Max additional cost represented on the graph:
$95k - $45k = $50,000

HS Annual Wage: $28,000
Lifetime Earnings: 1,260,000

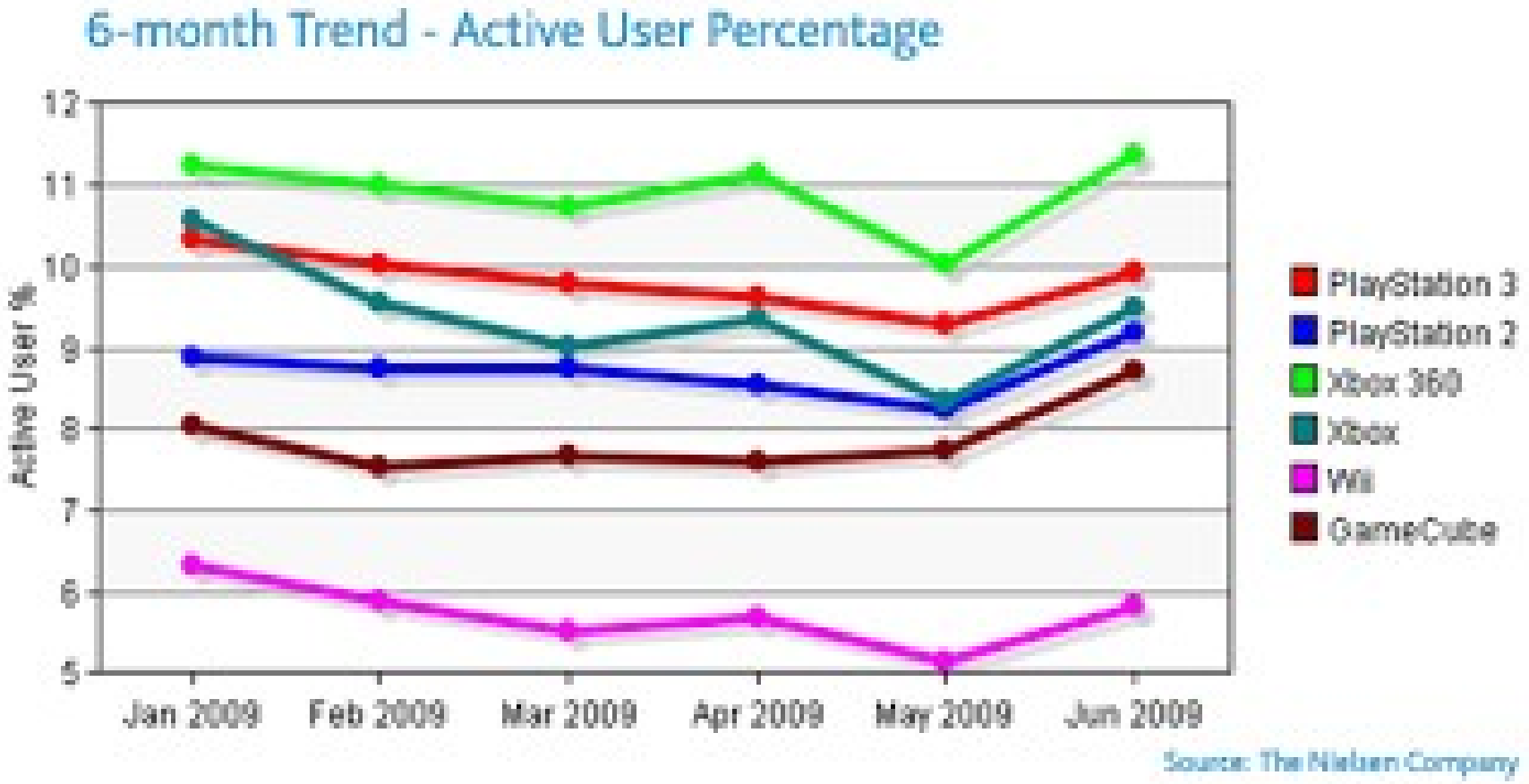
BA Annual Wage:
Lifetime Earnings: 2,047,500

Value of BA today: $787,500



Control over data grants...

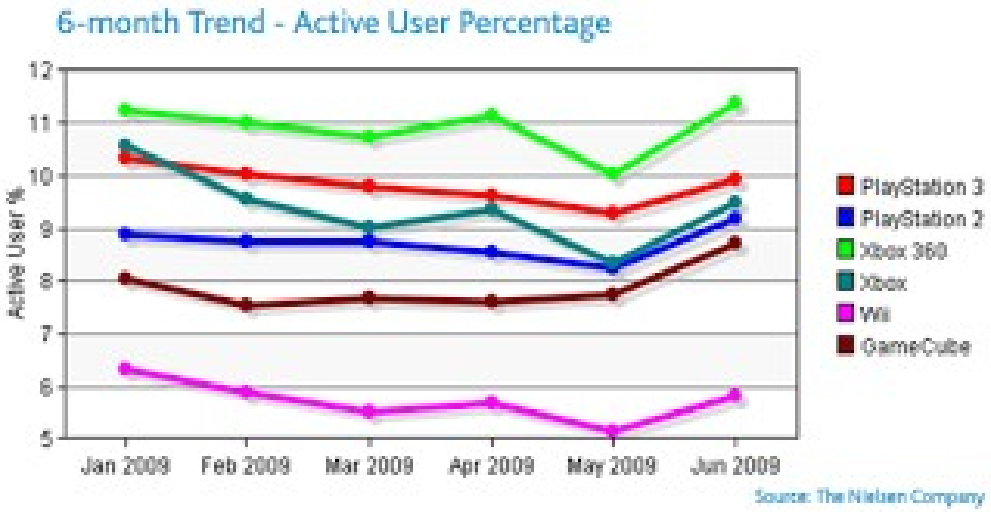
The choice to reveal or retain observed results.



Control over data grants...

The choice to reveal or retain observed results.

System	% Active Users	# of Users	Total Active Users
Wii	6%	50 mil	3 mil
Xbox 360	11%	30 mil	3.3 mil
PS3	10%	20 mil	2 mil



Control over data grants...

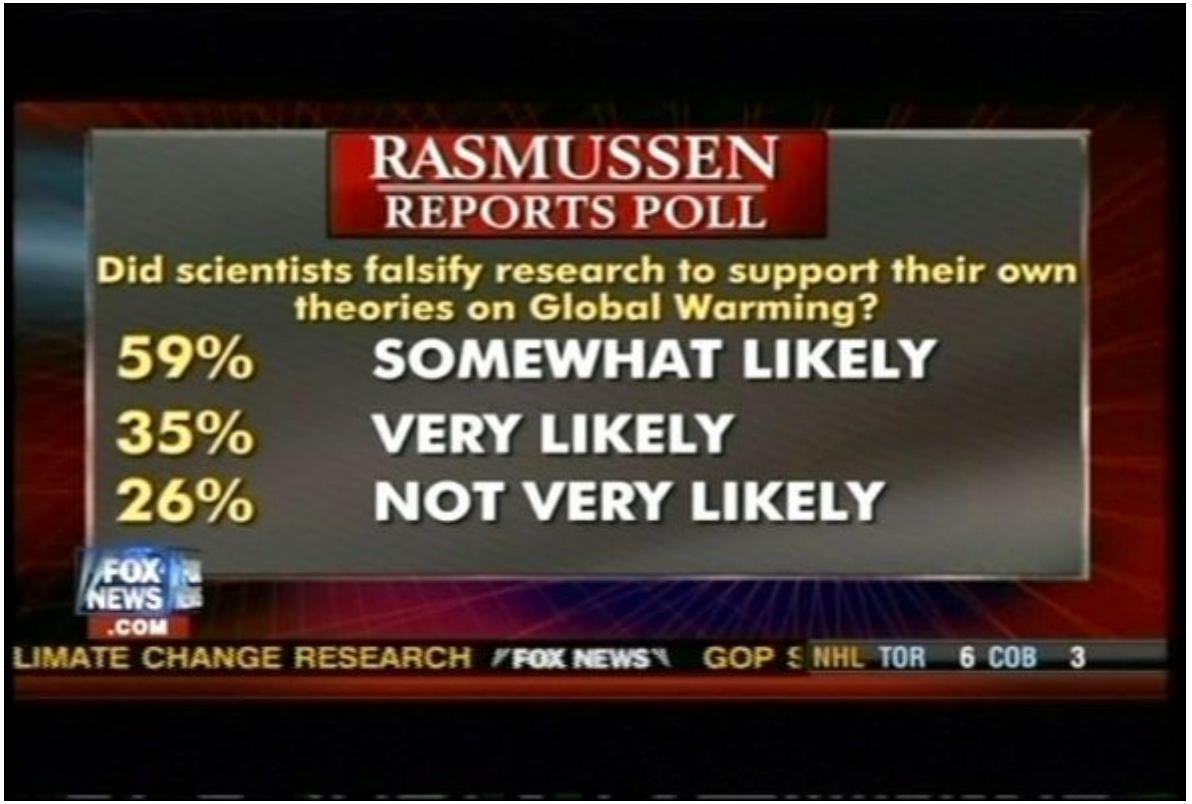
- The choice of method.

Candidate outcomes based on potential non-absentee recounts in Florida presidential election 2000 (outcome of one particular study)[37][*clarification needed*]	
Review method	**Winner**
Review of all ballots statewide (never undertaken)	
• Standard as set by each county canvassing board during their survey	Gore by 171
• Fully punched chad and limited marks on optical ballots	Gore by 115
• Any dimples or optical mark	Gore by 107
• One corner of chad detached or optical mark	Gore by 60
Review of limited sets of ballots (initiated but not completed)	
• Gore request for recounts of all ballots in Broward, Miami-Dade, Palm Beach, and Volusia counties	Bush by 225
• Florida Supreme Court of all undervotes statewide	Bush by 430
• Florida Supreme Court as being implemented by the counties, some of whom refused and some counted overvotes as well as undervotes	Bush by 493
Unofficial recount totals	
• Incomplete result when the Supreme Court stayed the recount (December 9, 2000)	Bush by 154
Certified Result (official final count)	
• Recounts included from Volusia and Broward only	Bush by 537

Control over data grants...

"First Mover" capabilities

- You're right until someone proves you wrong!

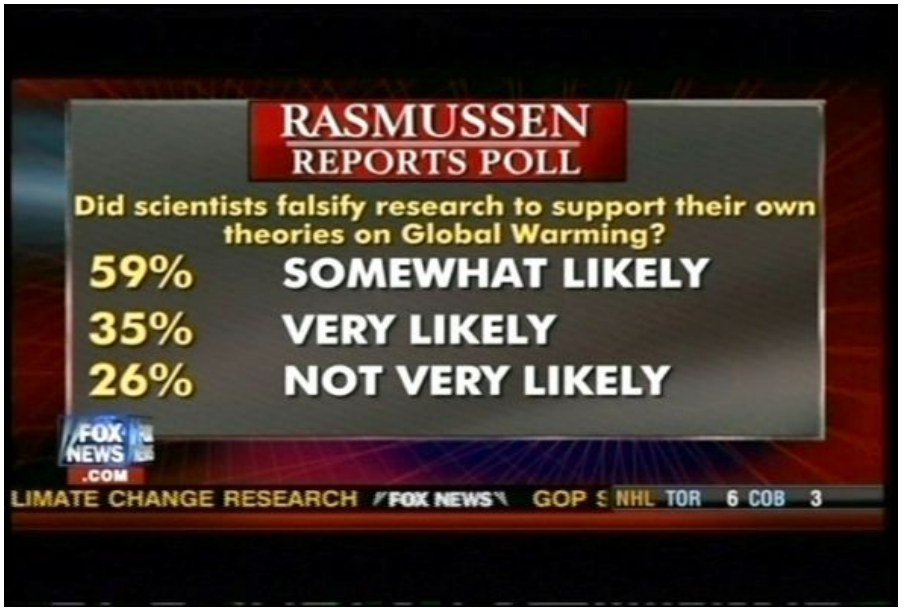


Control over data grants...

"First Mover" capabilities

- You're right until someone proves you wrong!

Entire organizations - like media matters – have to watchdog the gatekeepers of data to ensure they aren't just making stuff up. That costs money.



Control over data grants...

How and what you measure

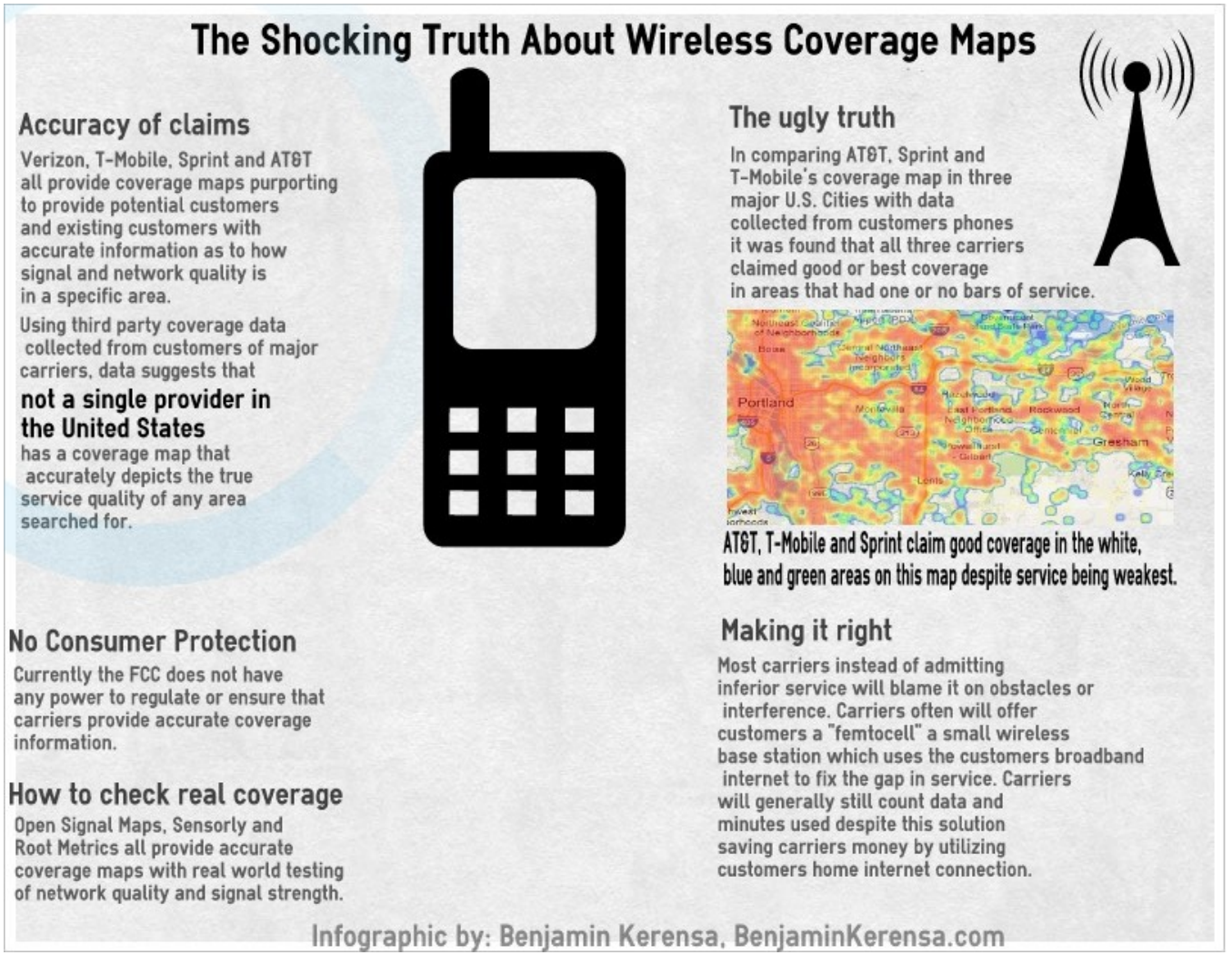
Control over data grants...

How and what you measure

"People on Welfare" was measured using households – not people. So, if a single person in your household was on welfare, all individuals were counted.

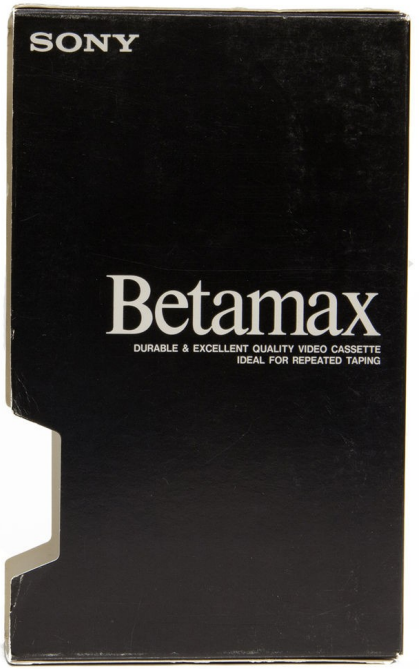
Control over data grants...

How and what you measure



Control over data grants...

How you standardize (and who you lock out)



Control over data grants...

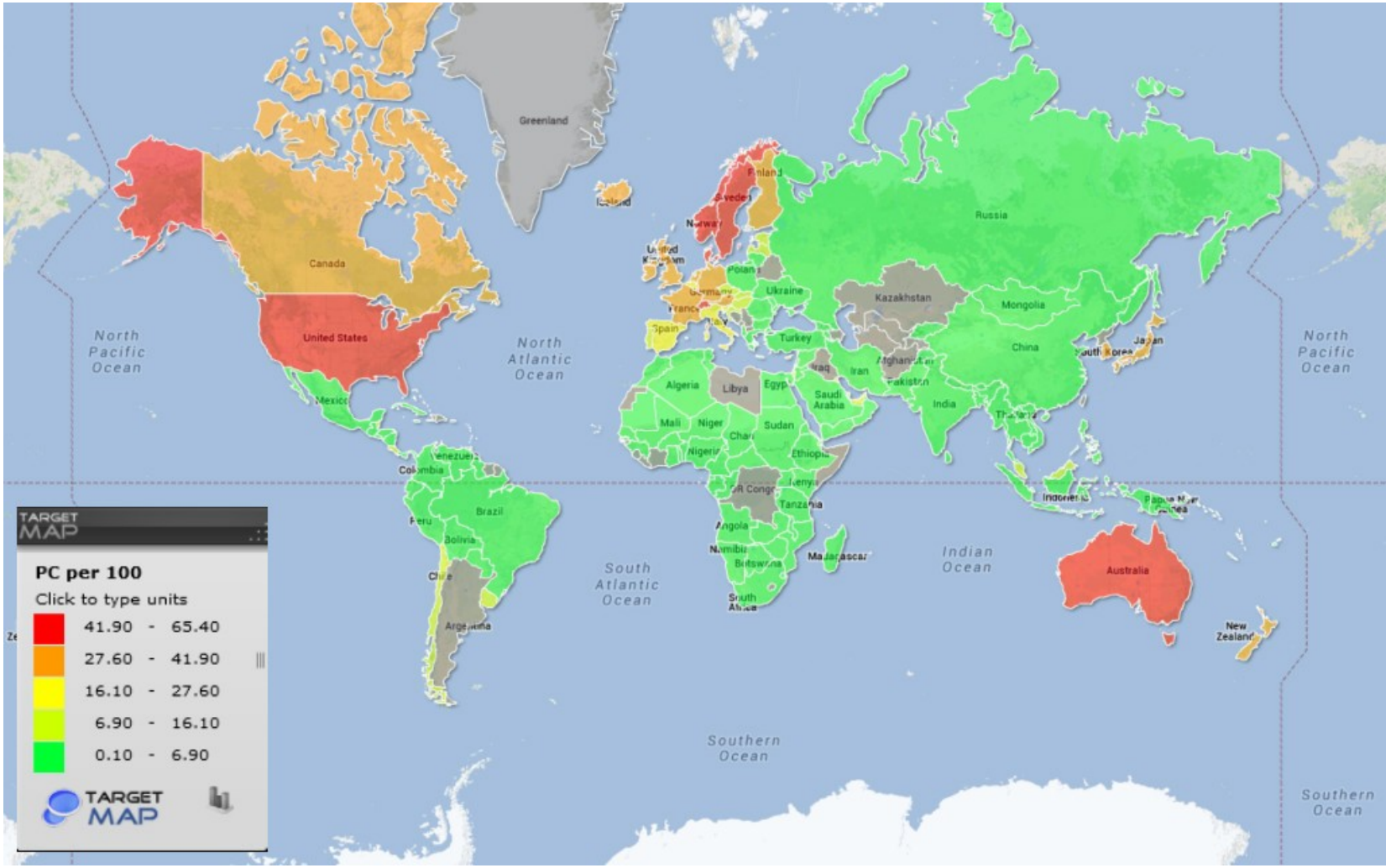
- The ability to analyze the data to tell *your* story.
- The choice to reveal or retain observed results.
- The choice of method.
- “First Mover” capabilities
  - You're right until someone proves you wrong!
- How and what you measure
- How you standardize (and who you lock out)
- Ultimately, **power**.

Seriously Big Responsibility

- Who has access to data?

Seriously Big Responsibility

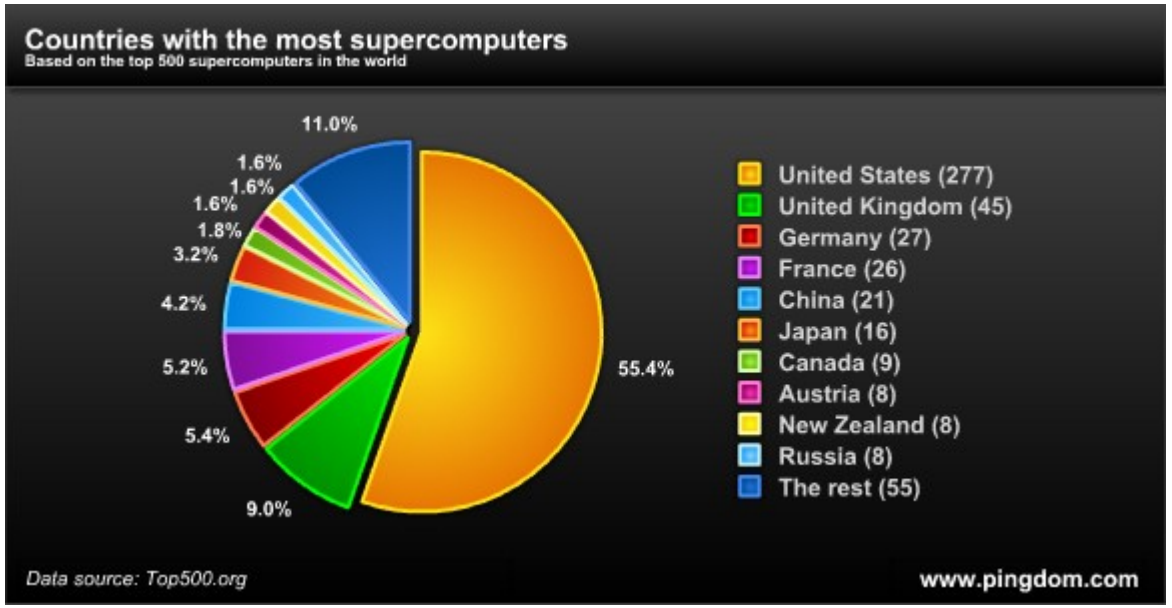
Who has access to hardware to process data?



http://www.tradingeconomics.com/country-list/personal-computers-per-100-people-wb-data.html

Seriously Big Responsibility

Who has access to hardware to process data?



If we assume ~12% of computing capability is in the developing world (likely a huge over-estimate), that translates to about 15 petaflops of capacity.

http://www.tradingeconomics.com/country-list/personal-computers-per-100-people-wb-data.html

Seriously Big Responsibility

Who has access to hardware to process data?

So, all nations not in the top 10 have roughly 15 petaflops of computational capacity.

This is just slightly more than the Xbox 360 servers that let us steal cars together in GTA (~13 petaflops).

http://www.tradingeconomics.com/country-list/personal-computers-per-100-people-wb-data.html

Seriously Big Responsibility

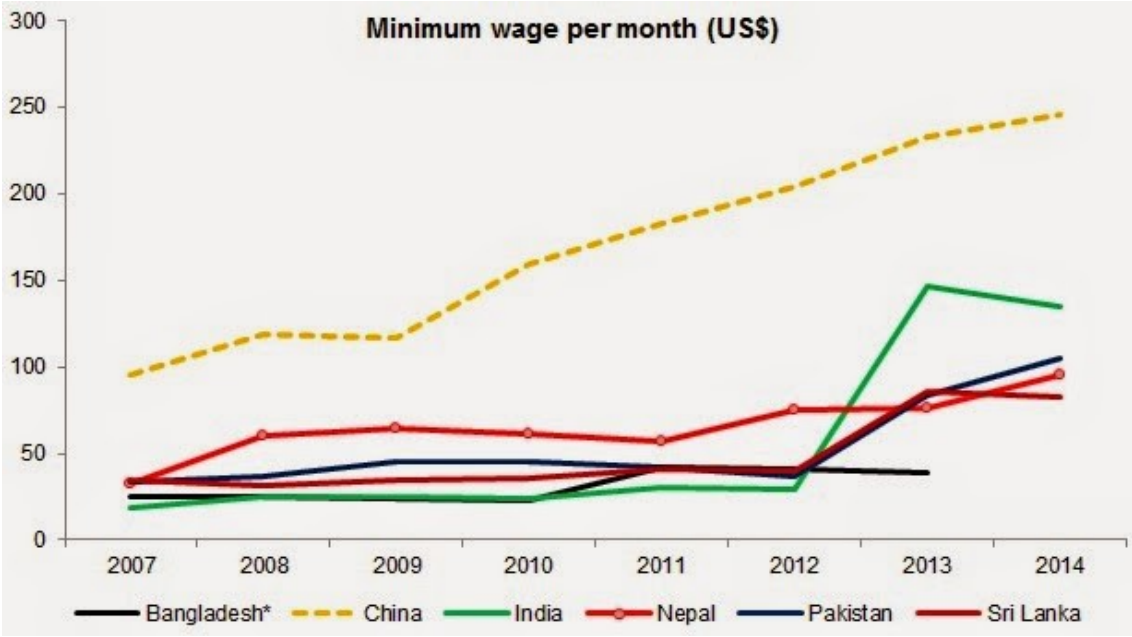
Who has access to hardware to process data?

- Software?

Cost Comparison: Office 2013 vs. 365

VERSION	COST/3YRS.
Office 2013 Home and Student	$140
Office 2013 Home and Business	$220
Office 365 Home Premium	**$300**
Office 2013 Standard *	$370
Office 2013 Professional	$400
Office 365 Small Biz Premium	**$450**
Office 2013 Professional Plus *	$500
Office 365 Midsize Business	**$540**
Office 365 Enterprise & Gov't	**$720**

* Volume licensing only

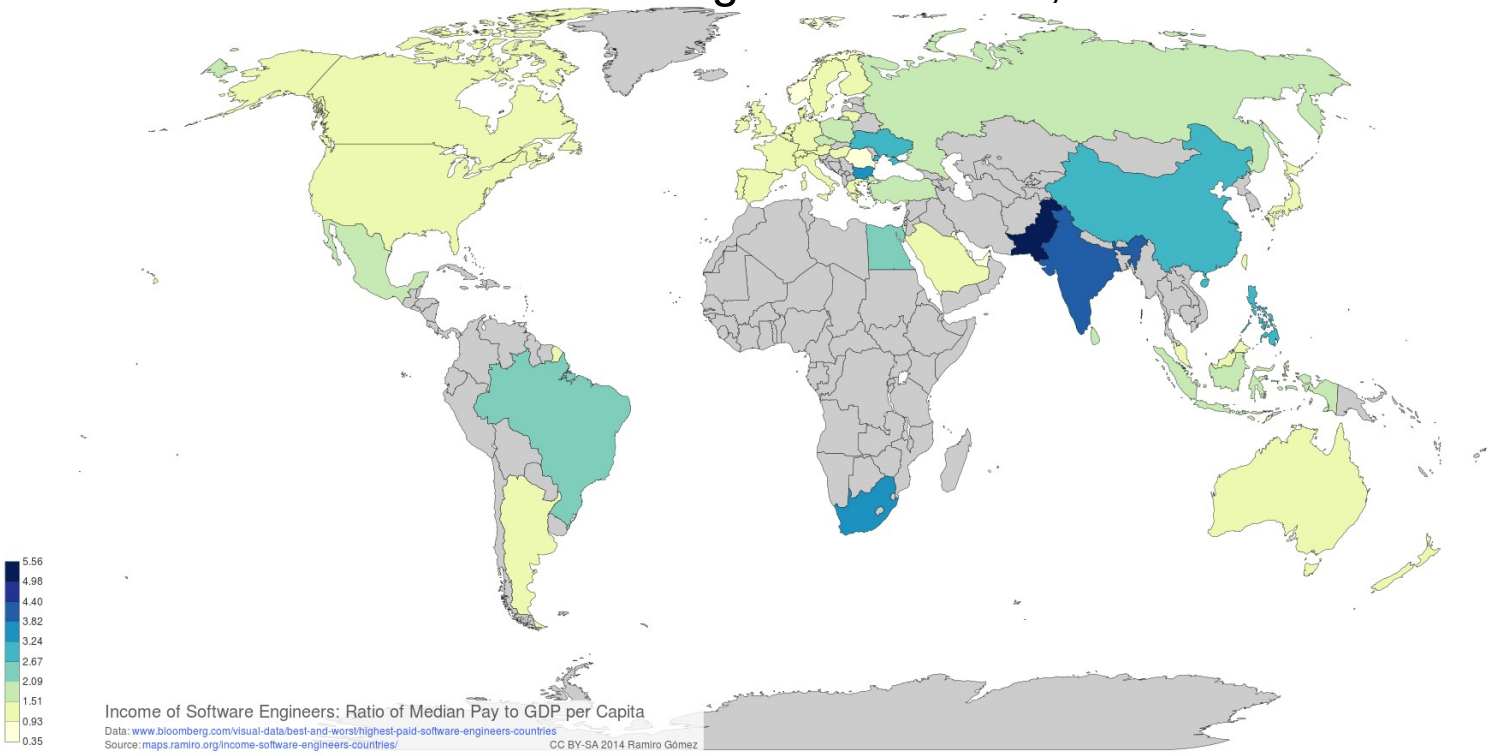


Seriously Big Responsibility

Who has access to hardware to process data?

- Software?
- Skills?

Software Engineer Salaries, relative to GDP

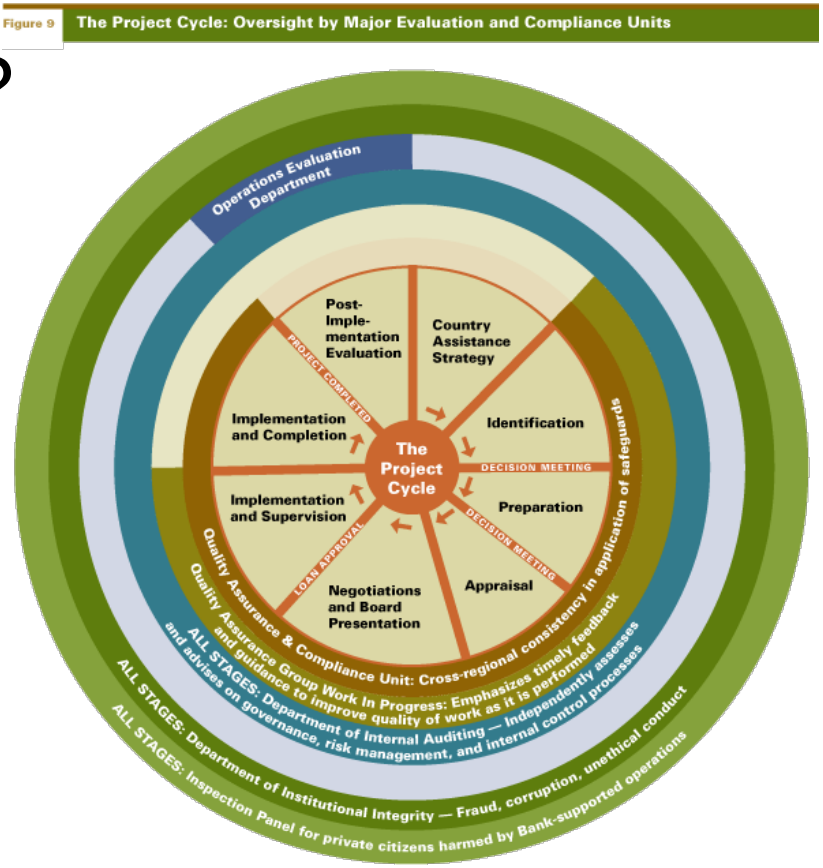


Seriously Big Responsibility

Who can challenge results?

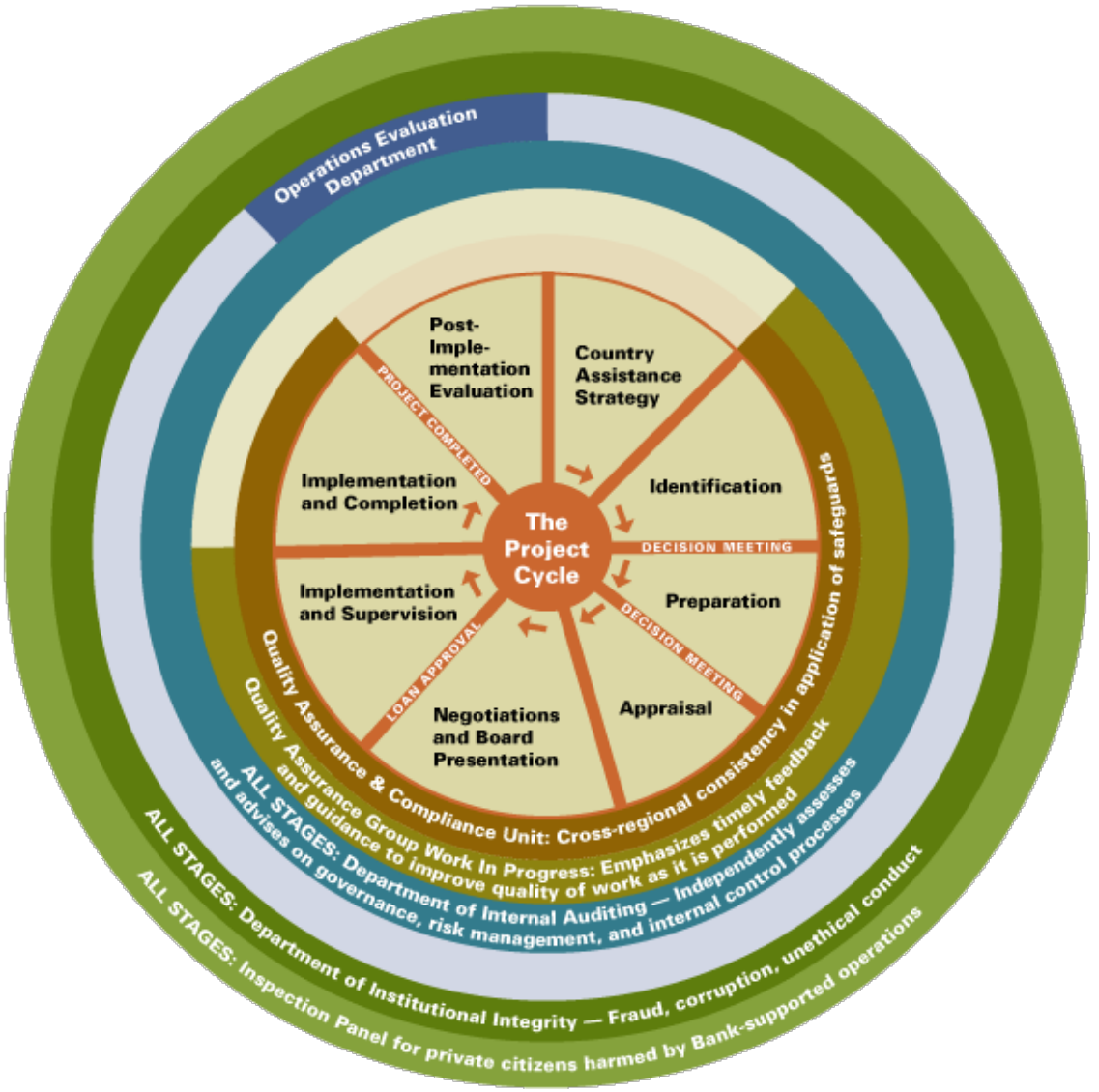
Who is impacted by the data?

Are these the same groups?



Seriously Big Responsibility

Figure 9 The Project Cycle: Oversight by Major Evaluation and Compliance Units



Seriously Big Responsibility

- Who has access to data?
- Who has access to hardware to process data?
- Software?
- Skills?
- Who can challenge results?
- Who is impacted by the data?

Weaponized Data*

- The illusion of objectivity.
- The delusion of validity.
- The assumption of generalizability.
- The danger of simplification.

*Dr. Jondou Chen coined this phrase, and should be praised for it for all time (at least in this footnote). Vu Le gets mad props for this slide as well.

Weaponized Data*

- The illusion of objectivity.

Humans Collect and Analyze Data.

- They have the choice to reveal or retain observed results.
- The choice of method.
- The choice of how and what to measure

*Dr. Jondou Chen coined this phrase, and should be praised for it for all time (at least in this footnote). Vu Le gets mad props for this slide as well.

Weaponized Data*

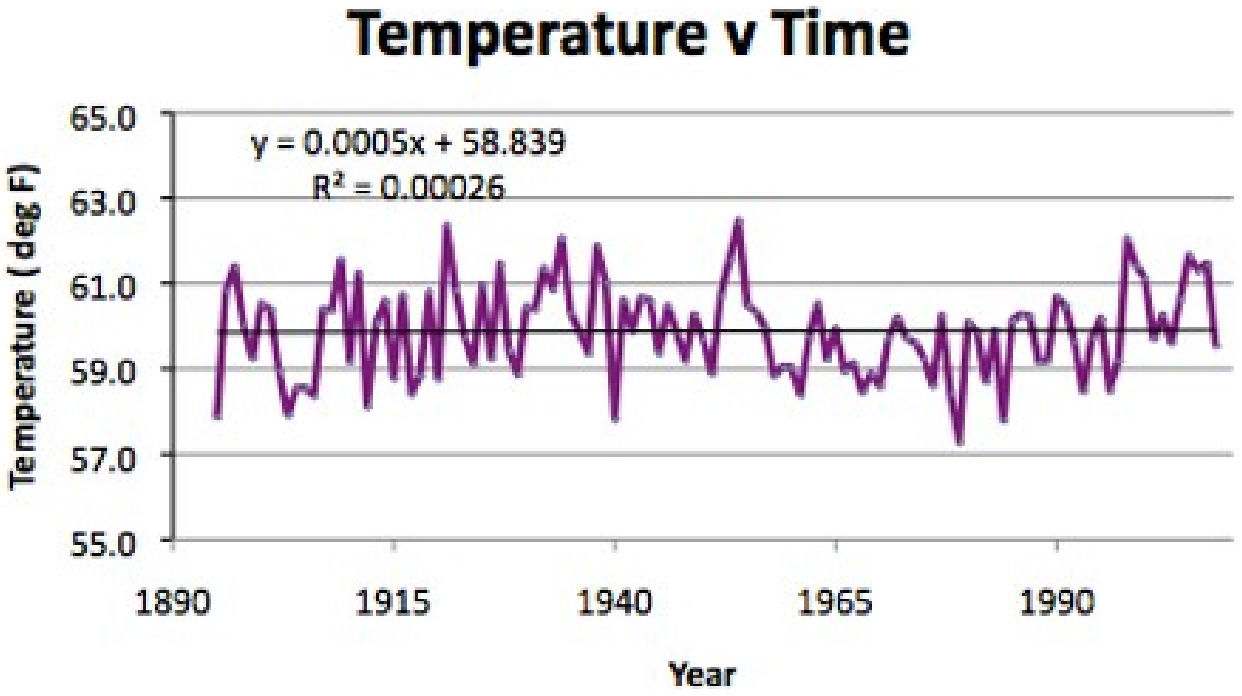
The delusion of validity.

75% of students that take **COLL 100: Breaking Intuition** go on to become managers in **top 100** businesses.

*Dr. Jondou Chen coined this phrase, and should be praised for it for all time (at least in this footnote). Vu Le gets mad props for this slide as well.

Weaponized Data*

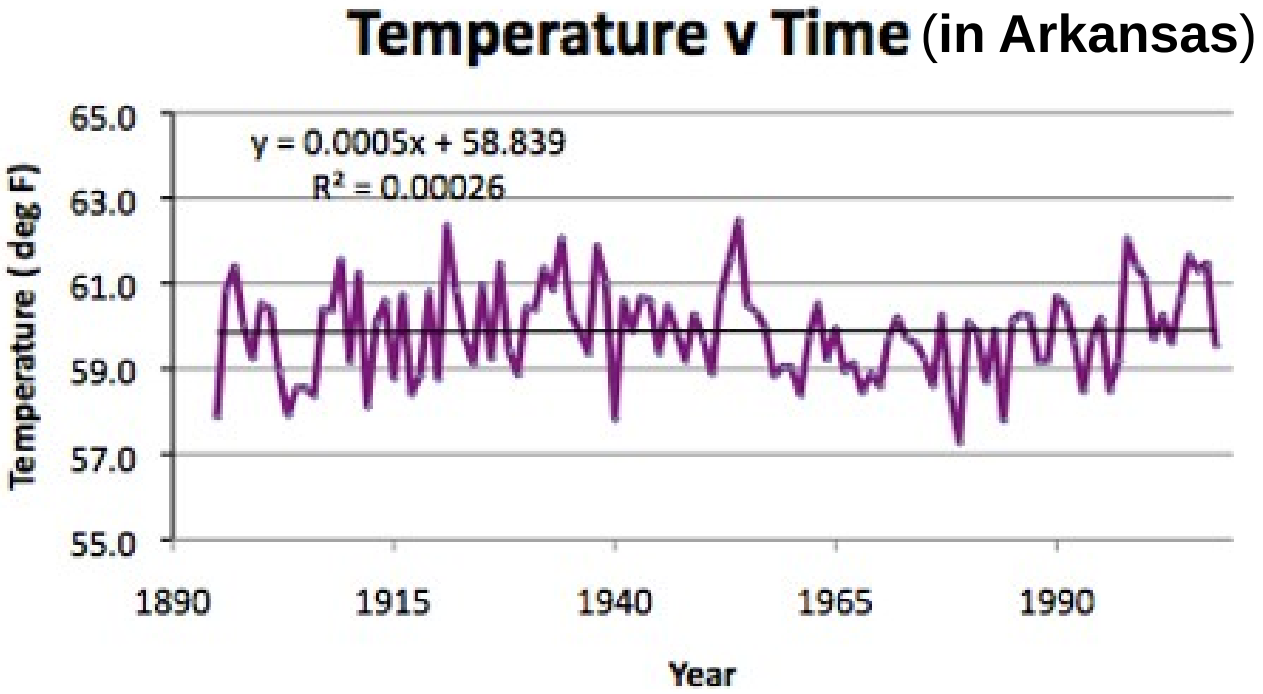
The assumption of generalizability.



*Dr. Jondou Chen coined this phrase, and should be praised for it for all time (at least in this footnote). Vu Le gets mad props for this slide as well.

Weaponized Data*

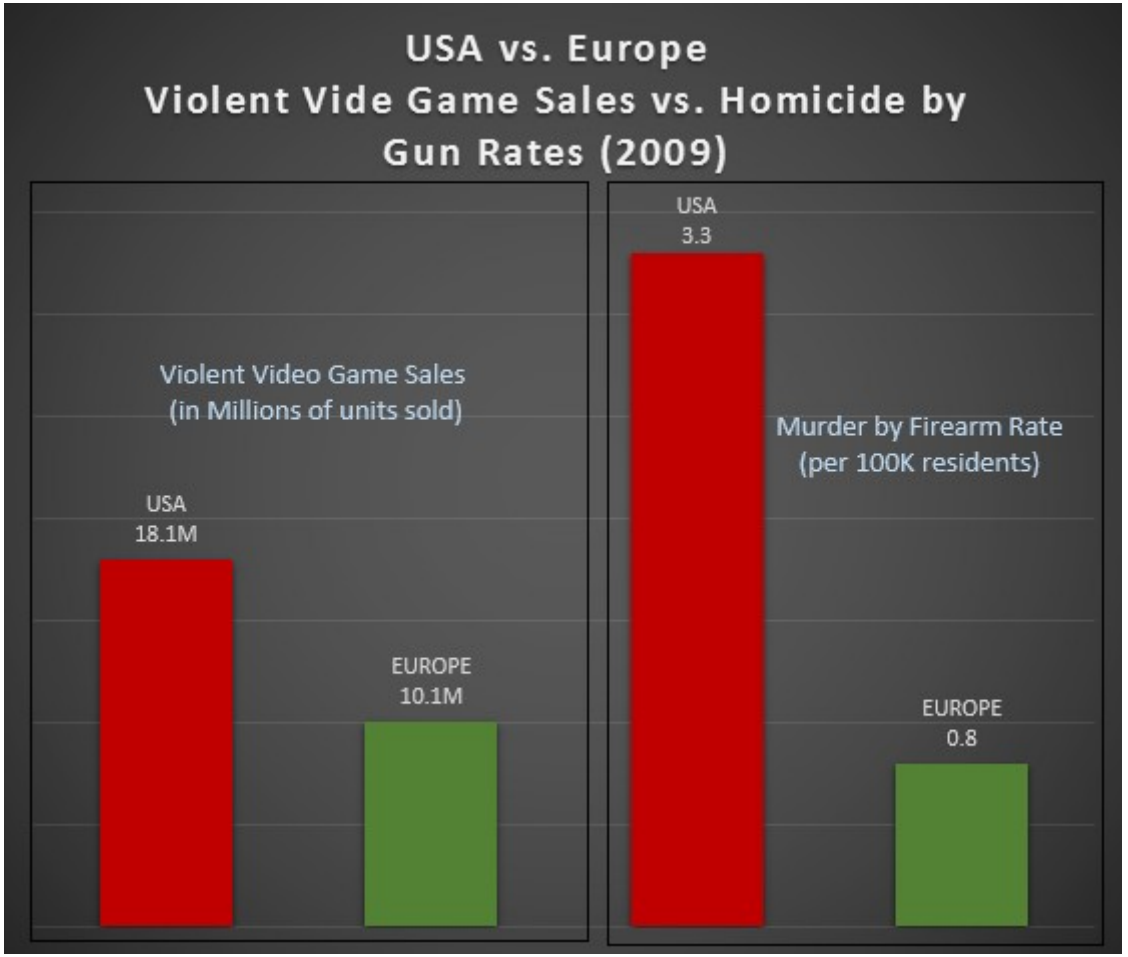
The assumption of generalizability.



*Dr. Jondou Chen coined this phrase, and should be praised for it for all time (at least in this footnote). Vu Le gets mad props for this slide as well.

Weaponized Data*

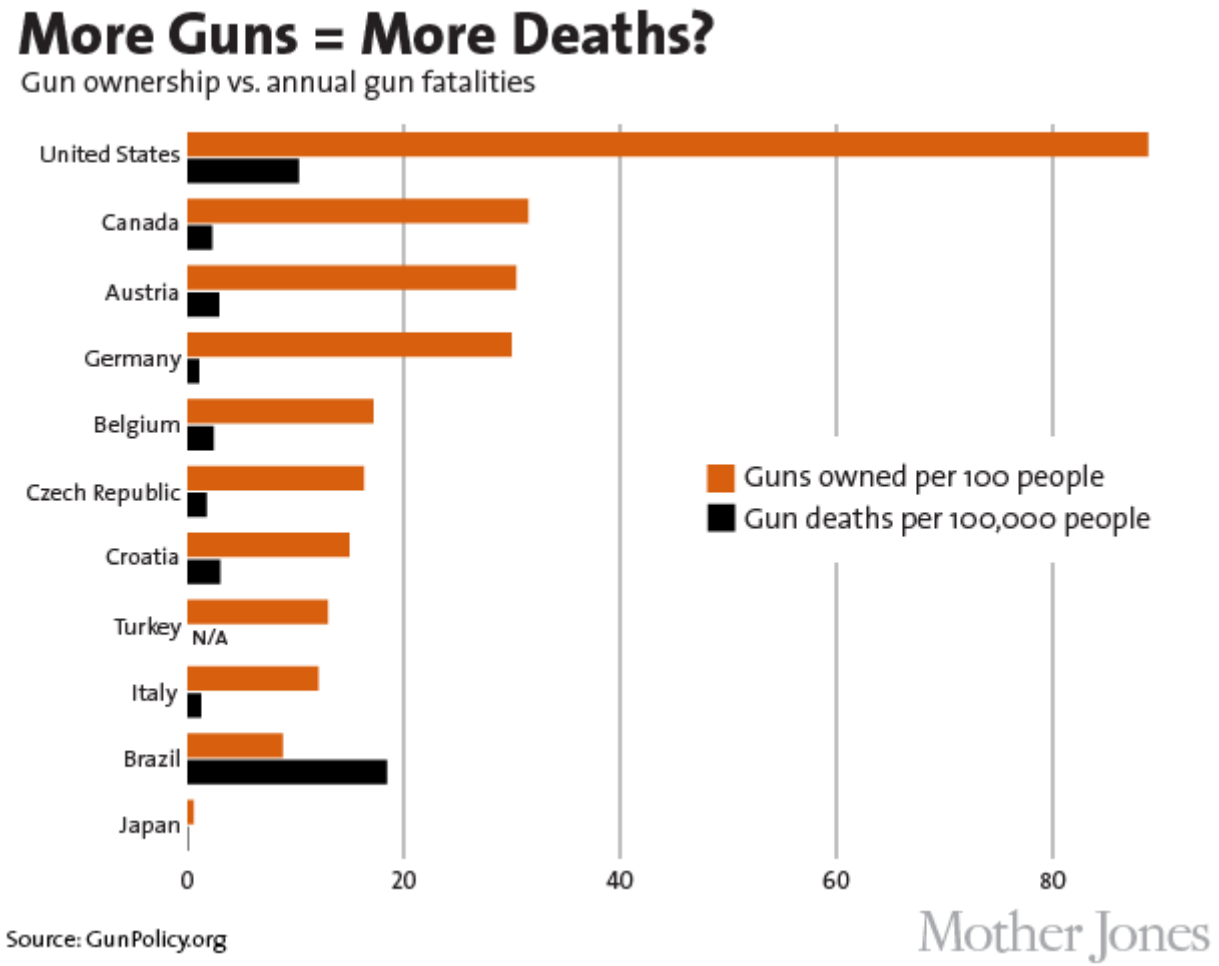
The danger of simplification.



*Dr. Jondou Chen coined this phrase, and should be praised for it for all time (at least in this footnote). Vu Le gets mad props for this slide as well.

Weaponized Data*

The danger of simplification.



*Dr. Jondou Chen coined this phrase, and should be praised for it for all time (at least in this footnote). Vu Le gets mad props for this slide as well.

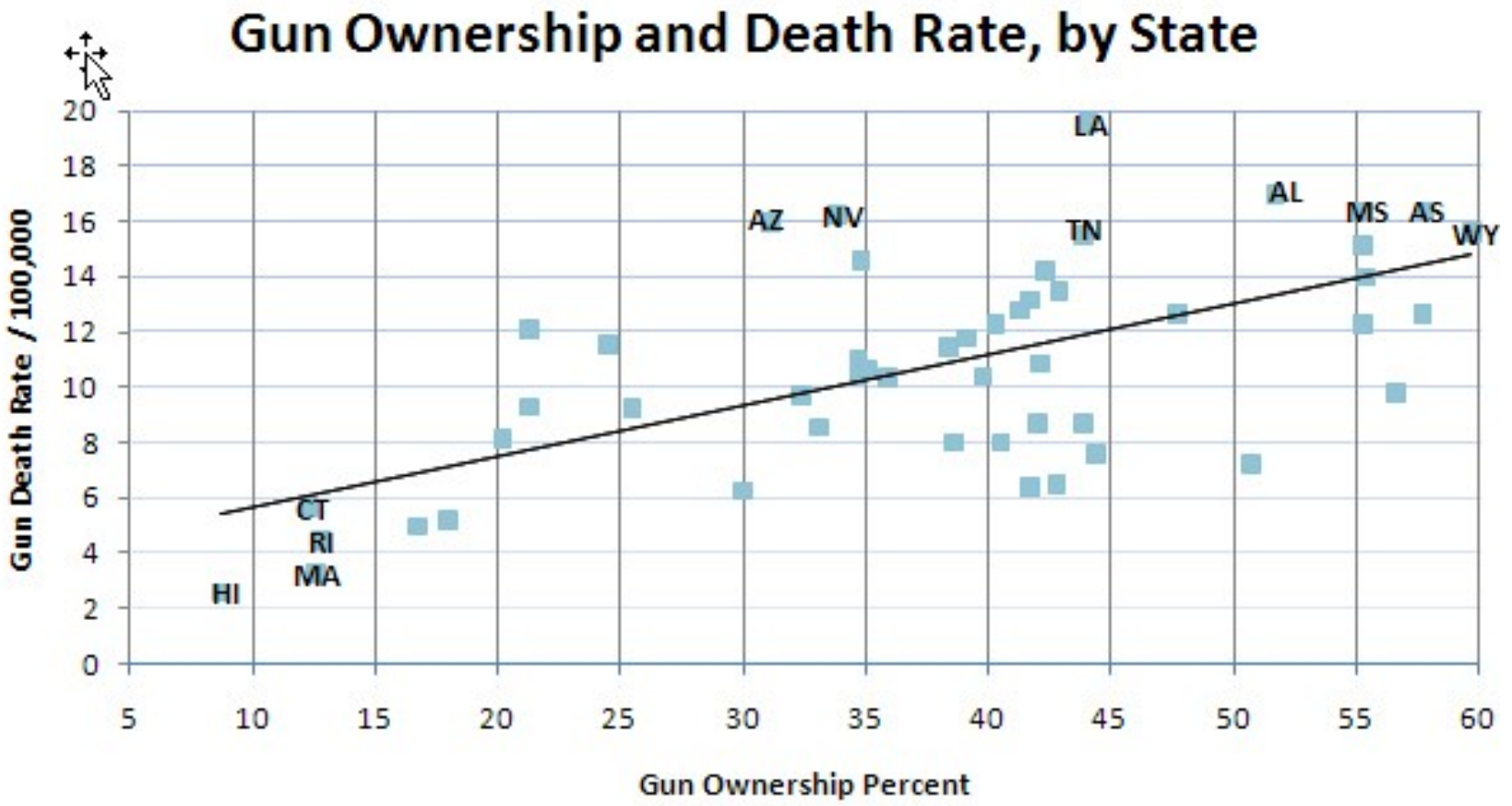
Weaponized Data*

The danger of simplification.

*Dr. Jondou Chen coined this phrase, and should be praised for it for all time (at least in this footnote). Vu Le gets mad props for this slide as well.

Weaponized Data*

The danger of simplification.



*Dr. Jondou Chen coined this phrase, and should be praised for it for all time (at least in this footnote). Vu Le gets mad props for this slide as well.

Simpson's Paradox

(Or, why you might want to take econometrics)

	Study All Night	Study All Day
Short Exam	Group A	Group B
Long Exam	Group C	Group D
Both Exams		

Simpson's Paradox

(Or, why you might want to take econometrics)

	Study All Night	Study All Day
Short Exam	Group A 93% got an A, 81/87	Group B
Long Exam	Group C	Group D
Both Exams		

Simpson's Paradox

(Or, why you might want to take econometrics)

	Study All Night	Study All Day
Short Exam	Group A **93%** got an A, 81/87	Group B, 87% got an A, 234/270
Long Exam	Group C	Group D
Both Exams		

Simpson's Paradox

(Or, why you might want to take econometrics)

	Study All Night	Study All Day
Short Exam	Group A **93%** got an A, 81/87	Group B, 87% got an A (234/270)
Long Exam	Group C **73%** got an A,192 /263	Group D, 69% got an A, 55/80
Both Exams		

Simpson's Paradox

(Or, why you might want to take econometrics)

	Study All Night	Study All Day
Short Exam	Group A **93%** got an A, 81/87	Group B, 87% got an A (234/270)
Long Exam	Group C **73%** got an A,192 /263	Group D, 69% got an A, 55/80
Both Exams	Overall: 78% got an A 273 / 350	Overall: **83%** got an A 239/350

Simpson's Paradox

(Or, why you might want to take econometrics)

When the less effective study strategy (Study all Day) is applied more frequently to easier cases (a short exam), it can appear to be a more effective strategy overall.

	Study All Night	Study All Day
Short Exam	Group A **93%** got an A, 81/87	Group B, 87% got an A (234/270)
Long Exam	Group C **73%** got an A,192 /263	Group D, 69% got an A, 55/80
Both Exams	Overall: 78% got an A 273 / 350	Overall: **83%** got an A 239/350

Take Aim...

- The Dragon's Hoard of Data
- Gatekeeping.
- “Blame the Data”

Vu Le gets mad props for this slide.

FIRE!

- The Dragon's Hoard of Data

Actually on Wikipedia
(“Digital Hoarding”)

FIRE!

- The Dragon's Hoard of Data

FIRE!

- Director of the National Institute for Mental Health (part of NIH):
  - “There are so many reasons not to share scientific data – in industry, among academics, and even for some patients. For pharmaceutical companies, data are usually considered proprietary, with sharing limited by intellectual property rules.”
  - “And patients may not want their private medical information shared. Particularly when that information includes a diagnosis of a mental illness or personal history, protecting privacy becomes even more important.”

http://www.nimh.nih.gov/about/director/2012/time-matters-why-we-care-so-much-about-data-sharing.shtml

FIRE!

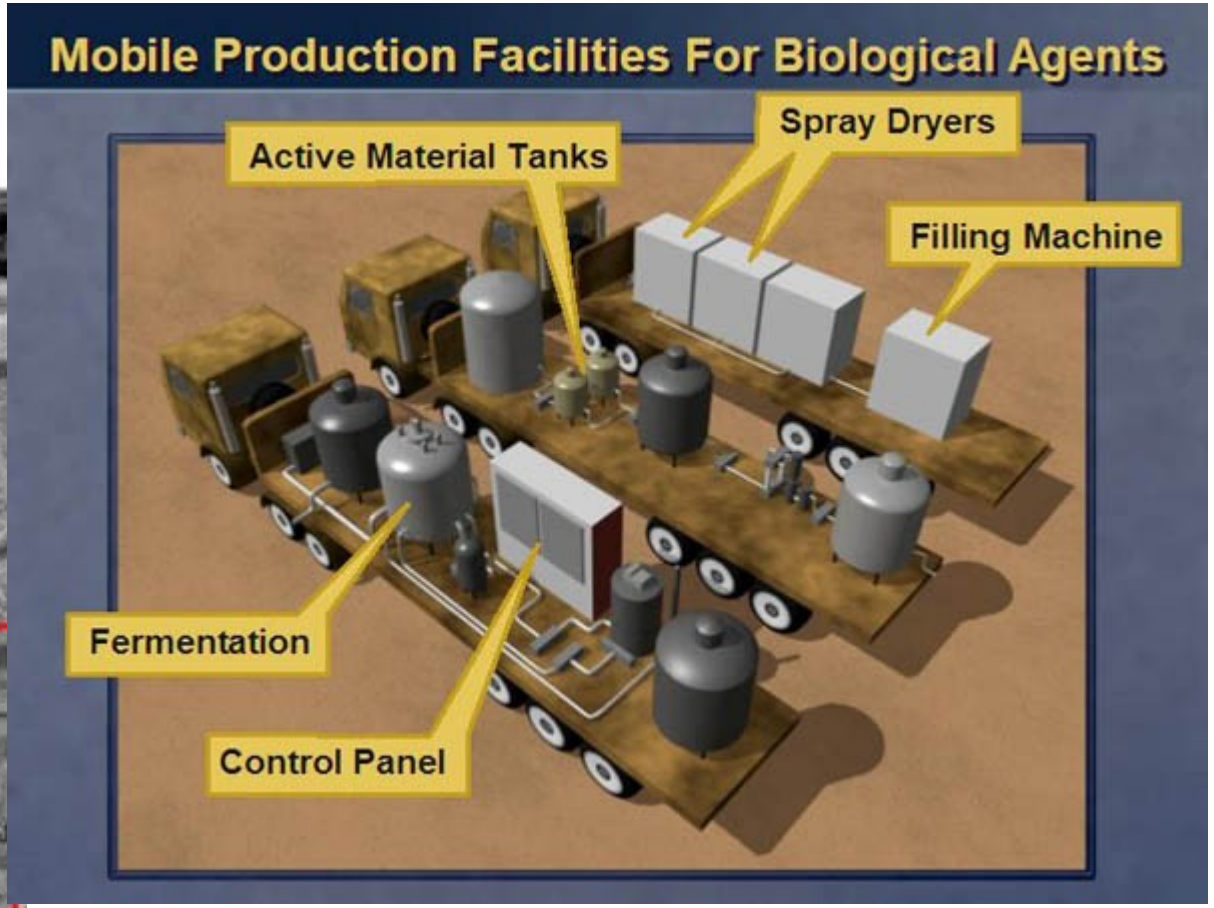
Gatekeeping.

FIRE!

Gatekeeping.

FIRE!

"Blame the Data"



Medical Evac!

- “Stop making decisions about us without us”
- Community Engagement around Data
- What is “good data”?
  - Is it about measurement?
  - Design?
  - Standards?
  - Quality Assurance?
  - Interpretability by different audiences?
  - Statistical Soundness...