Data on Data Breaches: Past, Present and Future

Adam Shostack and Chris Walsh
Emergent Chaos

This presentation represents the official position of the Emergent Chaos blog, not our employers

Welcome to Sevilla

From the Catalan Atlas by Abraham and Jehuda Cresques.

Navigational charts were kept secret during the age of exploration

- Henry the Navigator encouraged exploration
- Wanted the results for competitive advantage
- Columbus ended up in the Caribbean
- Lots of sailors died at sea
- Maps are still secret in some places
  - They don't like http://maps.google.com

We face navigation hazards, too

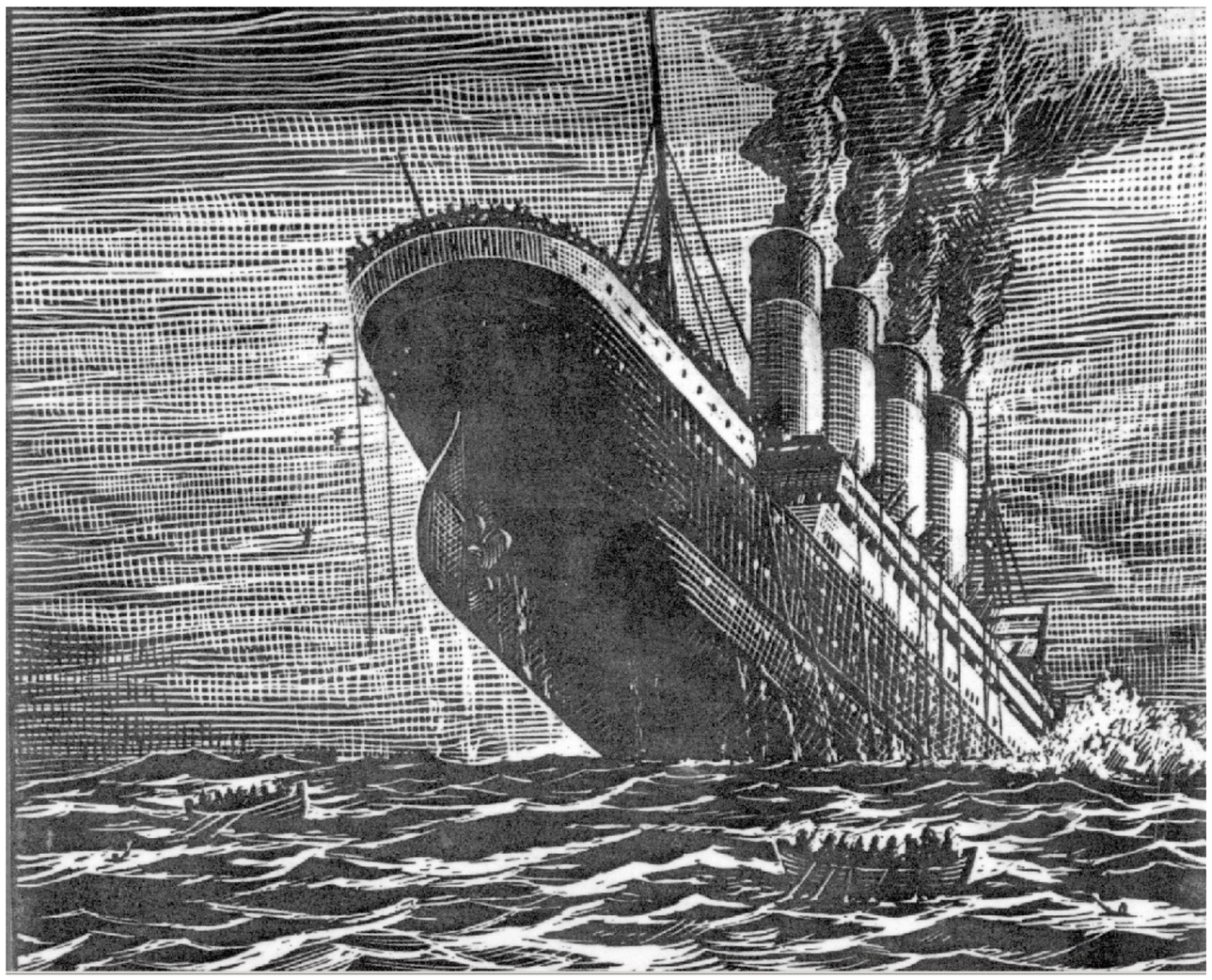


Image:http://www.materials.unsw.edu.au/news/brittlefracture/titanic%20sinking.jpg

We need to:

Know they exist :^)

Know how damaging they can be

Know our weak points if we run into them.

Know how to avoid them.

Case in Point: Security breaches involving personal information

Definitely exist
  But how numerous?
    How do we know?
  Are some more at risk than others?

Can be damaging
  But how much so, and to whom?
  How do we know?

Weak points driven by economics, not physics

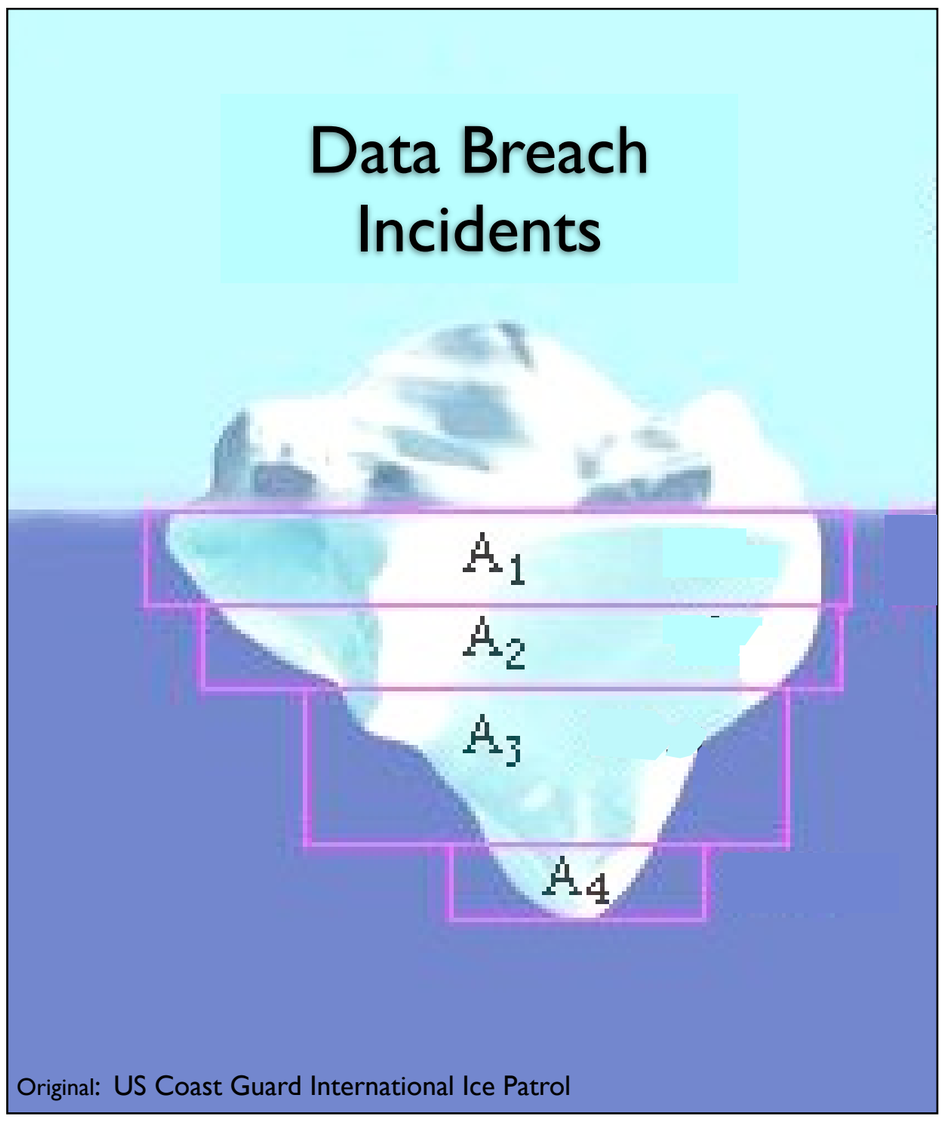
Avoidance techniques must be strategic

From the standpoint of a given organization, the overall number of breaches is not as important as the likelihood that this particular organization will be victimized, how much it will hurt, and what they can do to lessen their risk.

From a macro perspective, for example that of a policy maker, the number of organizations affected, their commonalities and differences, and the extent to which the consequences of breaches fall upon those external to the organizations owning (or holding) the exposed information.

Few organizations have enough individual experience with breach incidents to be able to rely solely on themselves for the data they need to inform decision-making.

Security Breaches: How numerous?



Below the waterline:

1. Undetected incidents
2. Unreported incidents
3. Reported, but unanalyzed
4. Reported, but privileged

Focus here is on 2, 3, and a little bit of 4.

So, looking at the topic in the most basic terms, we would first like to know how numerous such data breaches are. To use a familiar metaphor, we only know about what we can see.

Taking the iceberg as the overall number of breaches (rather than just those that affect a single organization, say), the part we see today is due in large degree to press reports. These reports are based on breach notices which affected organizations send to individuals (and -- much less often -- to investigative reporting).

What we do not see -- the part "below the waterline" -- consists of a few categories of things.

First are those incidents which occur, but which are not detected. These are akin to the infamous "false negatives" which keep life interesting for anti-virus and IDS vendors.

Second are unreported incidents. An organization knows they have occurred, but that knowledge remains within the organization. There are a number of sound reasons to keep such information private, a topic to which we will soon return.

Third are incidents which are reported beyond the organization, but which for whatever reason do not become part of corpus of data used to inform practice. Perhaps the existence of the information is not widely known, costly to obtain, or not "interesting" in the eyes of the intermediaries who might obtain it, analyze it, and bring it to the attention of the broader public.

Last are incidents which also are reported beyond the organization, but to organizations which consider them privileged and say nothing.

Understanding the sizes of these "regions below the waterline" is critical.

We will consider the second and third categories at length, and touch on the fourth.

How Do We Know?

Individual reports: News stories, press releases

Collections of same

- For general use - Emergent Chaos breaches category, Attrition.org's DLDOS, etc.
  - Google Alerts are the researcher's friend
- For specific purposes - data behind a journal article
  - Often use commercial news archives such as LexisNexis

Reports are much more numerous now that states have notification laws

We learn about the region above the waterline in a few ways.

In the United States at least, each day brings another story of an organization which has exposed personal information. Often breached organizations themselves will anticipate news coverage and issue a press release. These individual reports are collected and summarized in various places, such as Attrition.org's Dataloss Archive or Emergent Chaos' breaches category. We'll talk about these "traditional" sources bit in the next couple of slides.

Ultimately, much (if not all) of the serious empirical research involving breaches involving personal information draws upon such reports. The number of such reports has increased very dramatically in recent years, as we shall discuss.

An interesting question is what other sources of data about breaches might exist, and if they do, what their use can add to our understanding.

Attrition's DLDOS

http://attrition.org/dataloss/dldos.html

Provides "date, the company that reported the breach, the type of data impacted, the number of records impacted, third party companies involved, and a few other sortable items"

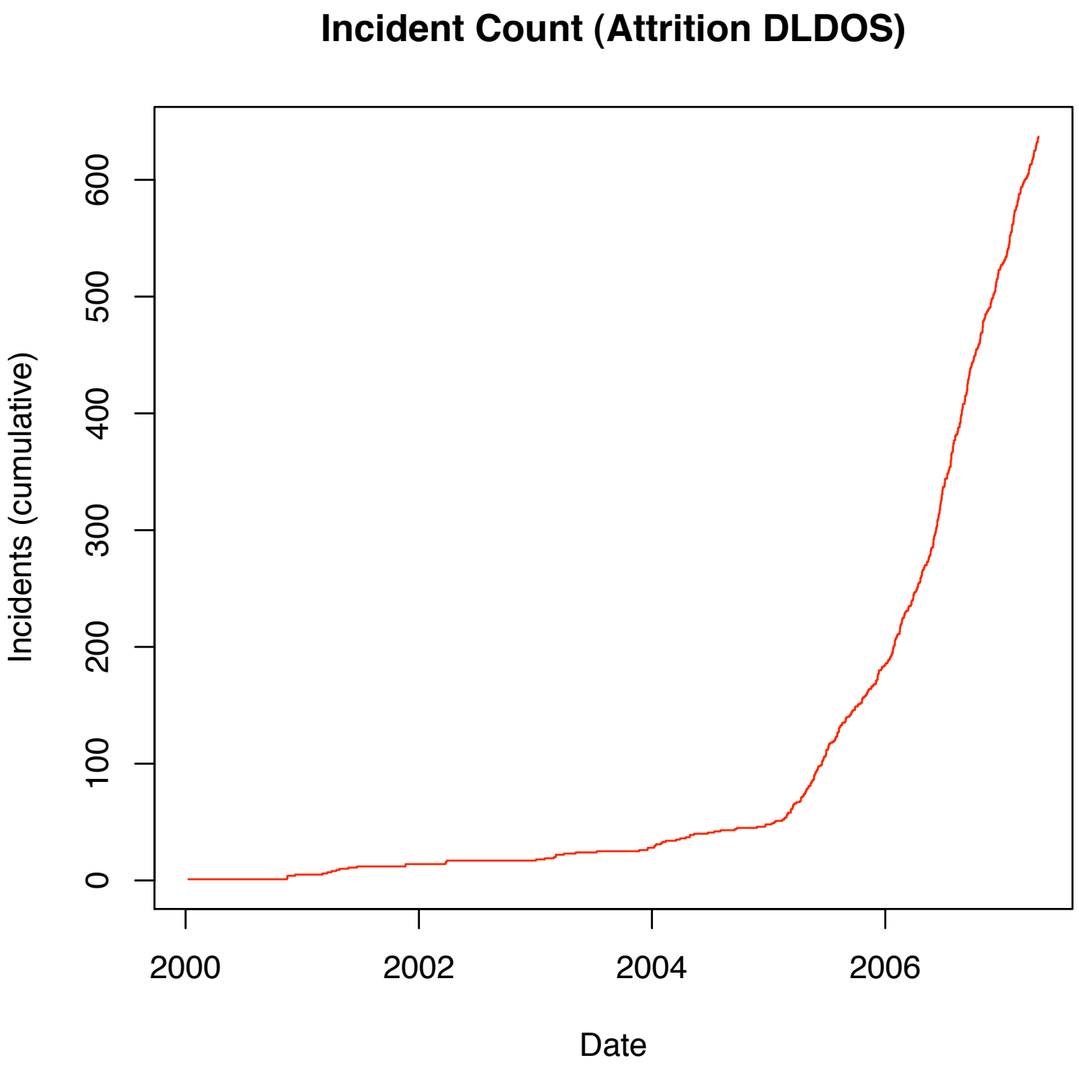
700 records as of June 13, 2007.

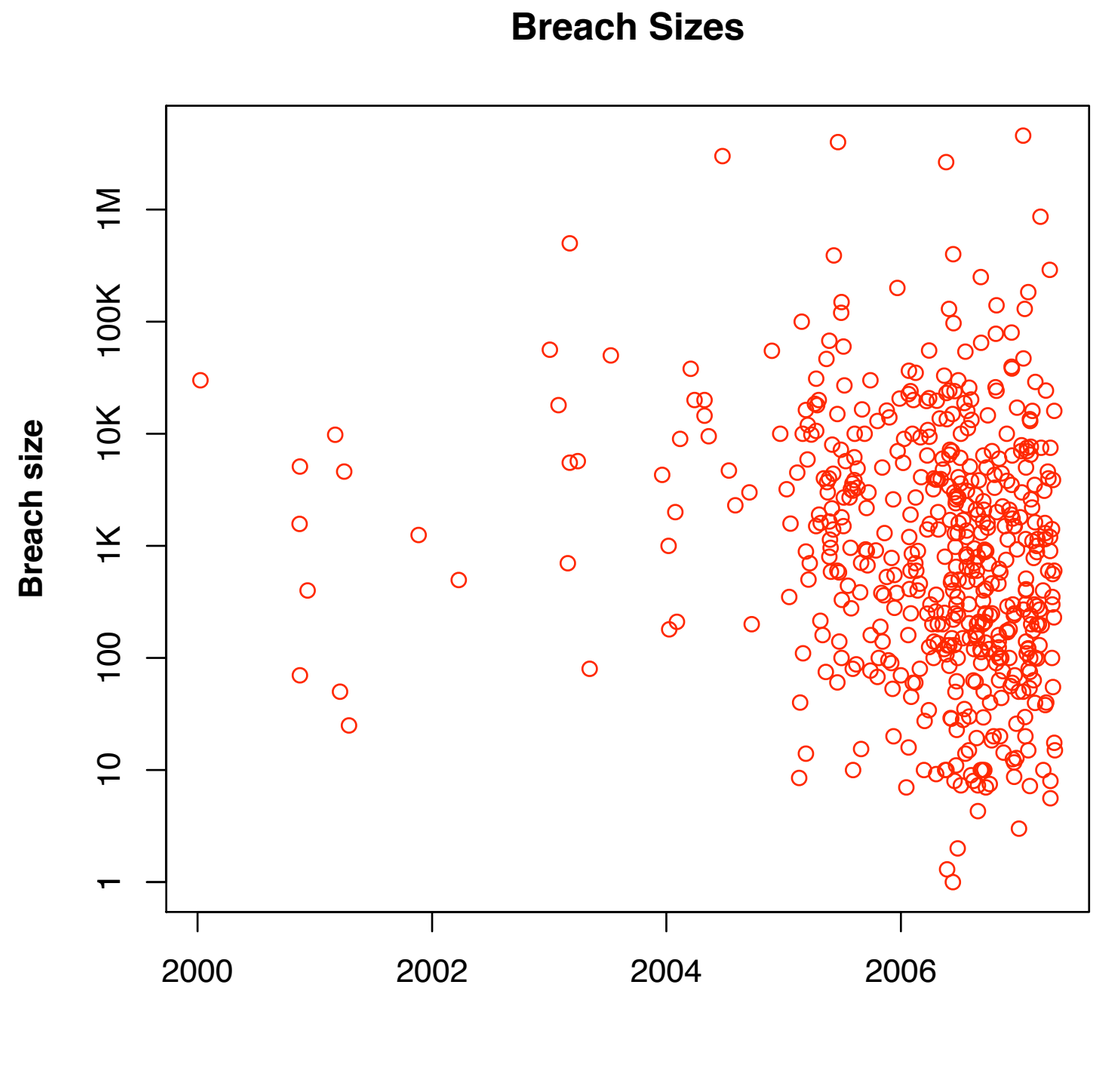
A main data supplier to other well-known sources, academic works, etc.

Attrition.org maintains the most extensive 'database' (actually a CSV file) on data breaches, the so-called "Data Loss Database - Open Source". The unit of analysis is the breach, about each of which several things are recorded.

Other well-known sources of information on breaches, such as The Privacy Rights Clearinghouse's data breach chronology draw heavily from Attrition's mailing list and DLDOS "database".

Attrition.org Incident Archive





Considering this one well-known collection of breach data -- Attrition.org's "Dataloss Database - Open Source" -- the number of known breach incidents over the last two to three years has exploded.

The chart to the right, showing the distribution of breach sizes over time, is intended to provoke thought, not illustrate a point. Were it not for the new legislative environment, would we have learned of so many incidents, particularly the smaller ones?

Etiolated.org

etiolated consumer\citizen

Shedding light on who's doing what with your private information. Searchable Attrition.org DLDOS index.

Search: Go

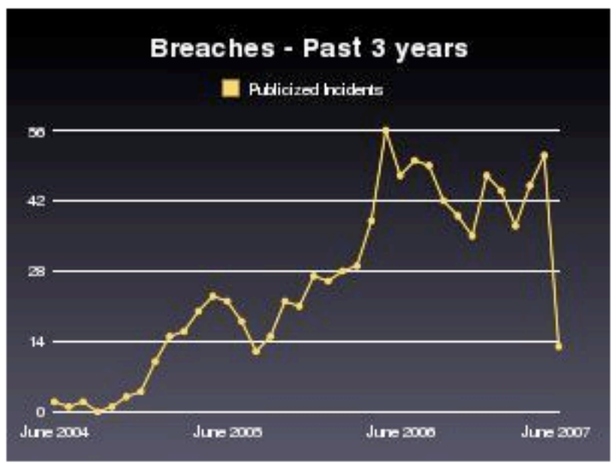
Main Statistics Research »Maps« Contact Login Signup **Contribute!**

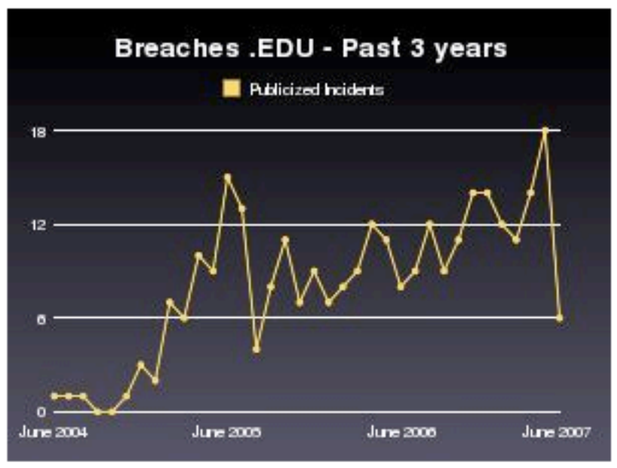
Largest Incidents Since 2000

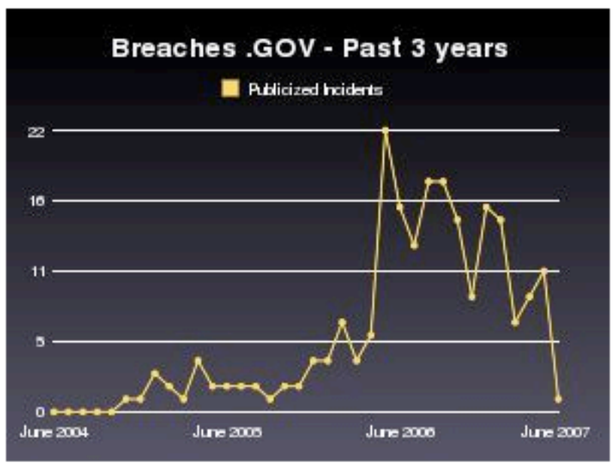
Number Affected	Date	Companies
45,700,000	2007-01-17	TJX Companies Inc.
40,000,000	2005-06-19	Visa, CardSystems, Mastercard, American Express
30,000,000	2004-06-24	America Online
26,500,000	2006-05-22	U.S. Department of Veterans Affairs
8,637,405	2007-03-12	Dai Nippon Printing Company
5,000,000	2003-03-06	Data Processors International
4,000,000	2006-06-13	KDDI
3,900,000	2005-06-06	Citigroup, UPS
2,900,000	2007-04-10	Georgia Department of Community Health, Affiliated Computer Services
2,500,000	2006-09-07	Chase Card Services

Most Recent Incidents

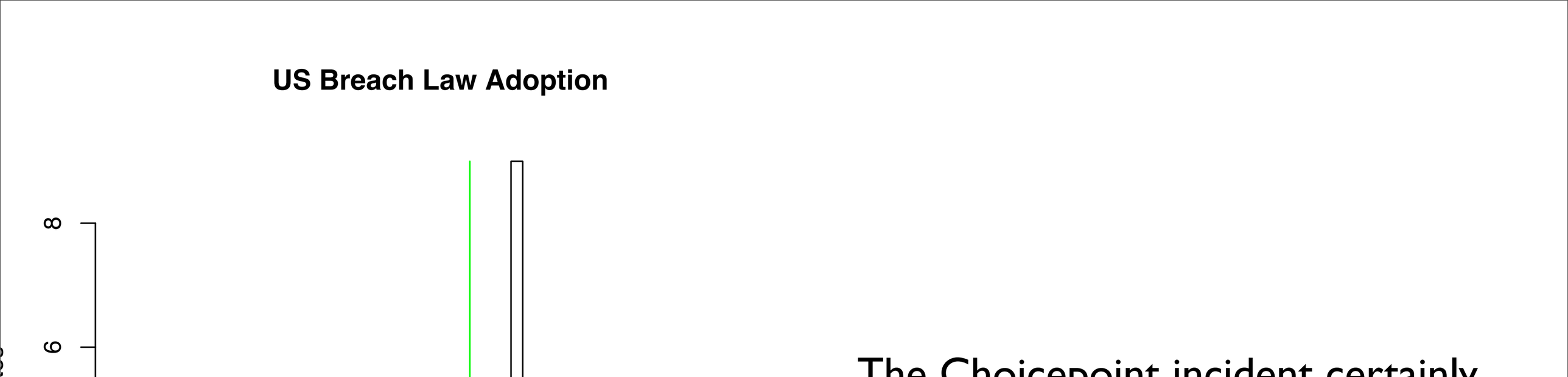
Number Affected	Date	Companies
3,000	2007-06-11	Grand Valley State University
17,000	2007-06-11	Pfizer
10,847	2007-06-11	Verus Inc., Stevens Hospital, Kennewick General Hospital, Concord Hospital
3,000	2007-06-09	Concordia Hospital
1,100	2007-06-08	University of Iowa
5,735	2007-06-08	University of Virginia
Zero or Unknown	2007-06-07	Dearfield Medical Building
Zero or Unknown	2007-06-06	Cedarburg High School
400	2007-06-03	Gadsden State Community College
4,000	2007-06-01	Northwestern University







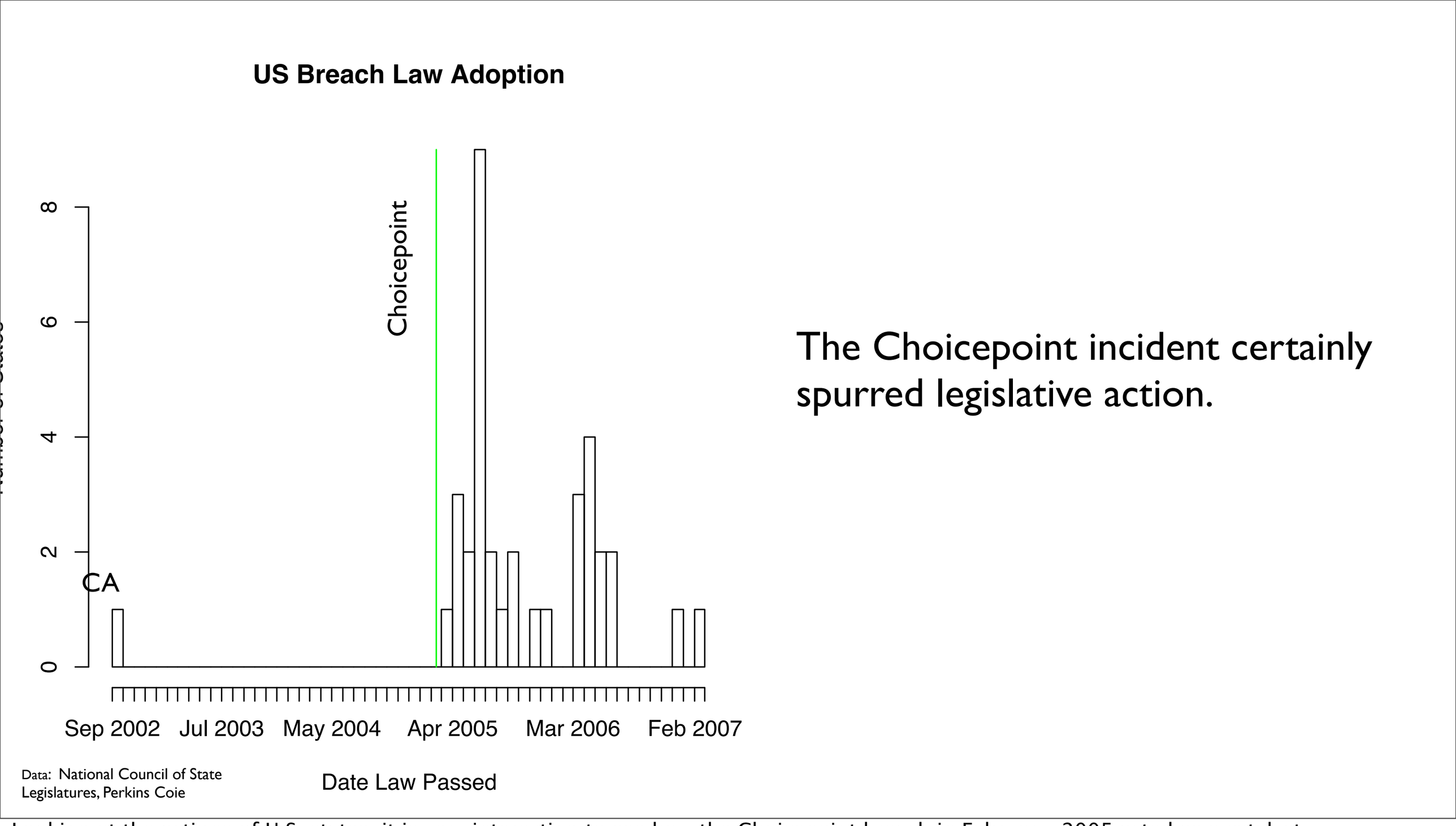
As an aside, the availability of such data allows for some pretty cool tool-building.



The Choicepoint incident certainly spurred legislative action.

Looking at the actions of U.S. states, it is very interesting to see how the Choicepoint breach in February, 2005 acted as a catalyst.

A thought experiment: If California hadn't had a law in effect in February, 2005, when would we have learned about ChoicePoint's experience? What would the pace and nature of subsequent reporting have been?

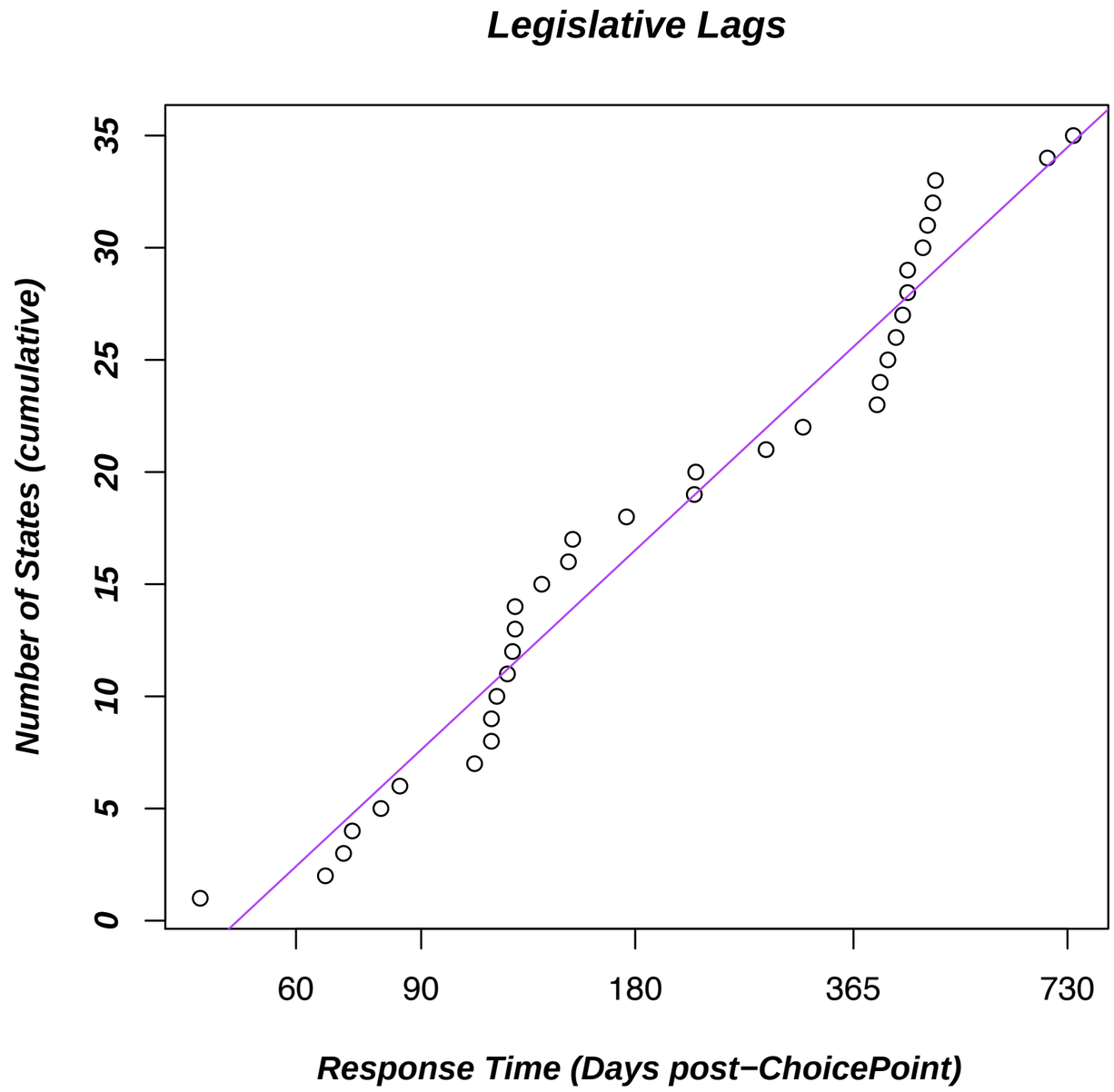


The Choicepoint incident certainly spurred legislative action.

Looking at the actions of U.S. states, it is very interesting to see how the Choicepoint breach in February, 2005 acted as a catalyst.

A thought experiment: If California hadn't had a law in effect in February, 2005, when would we have learned about ChoicePoint's experience? What would the pace and nature of subsequent reporting have been?

U.S. State Breach Notification Laws

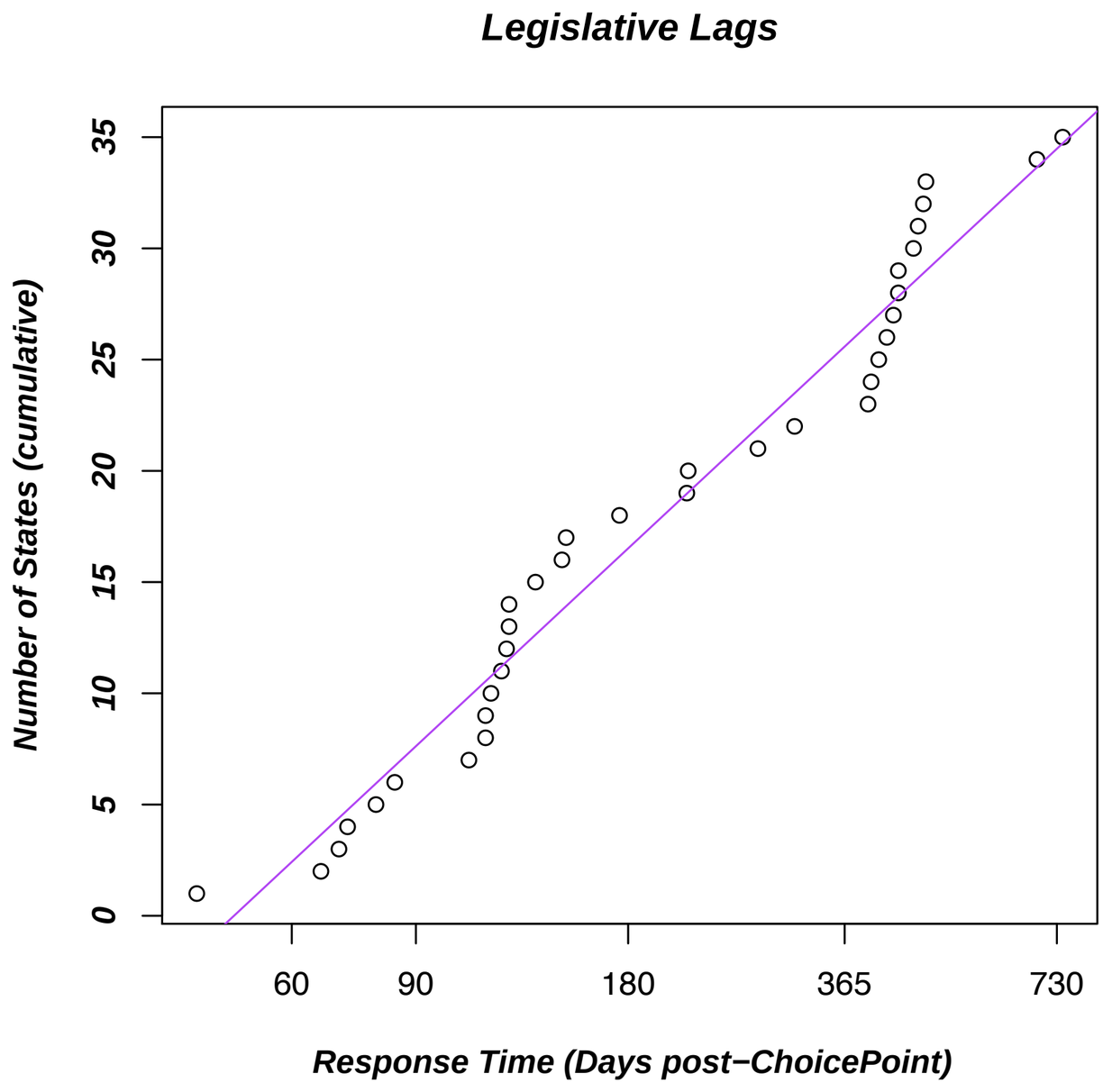


It is hard to measure the information security impact of these laws, in part because we only have two years' worth of data

Showing the data from three slides prior in a slightly different way, there seems to be a certain predictability to the passage of state laws.

Having only two years or so of data, we're just beginning to have the kind of information we need to understand the causes of these incidents.

Law passage times grow exponentially

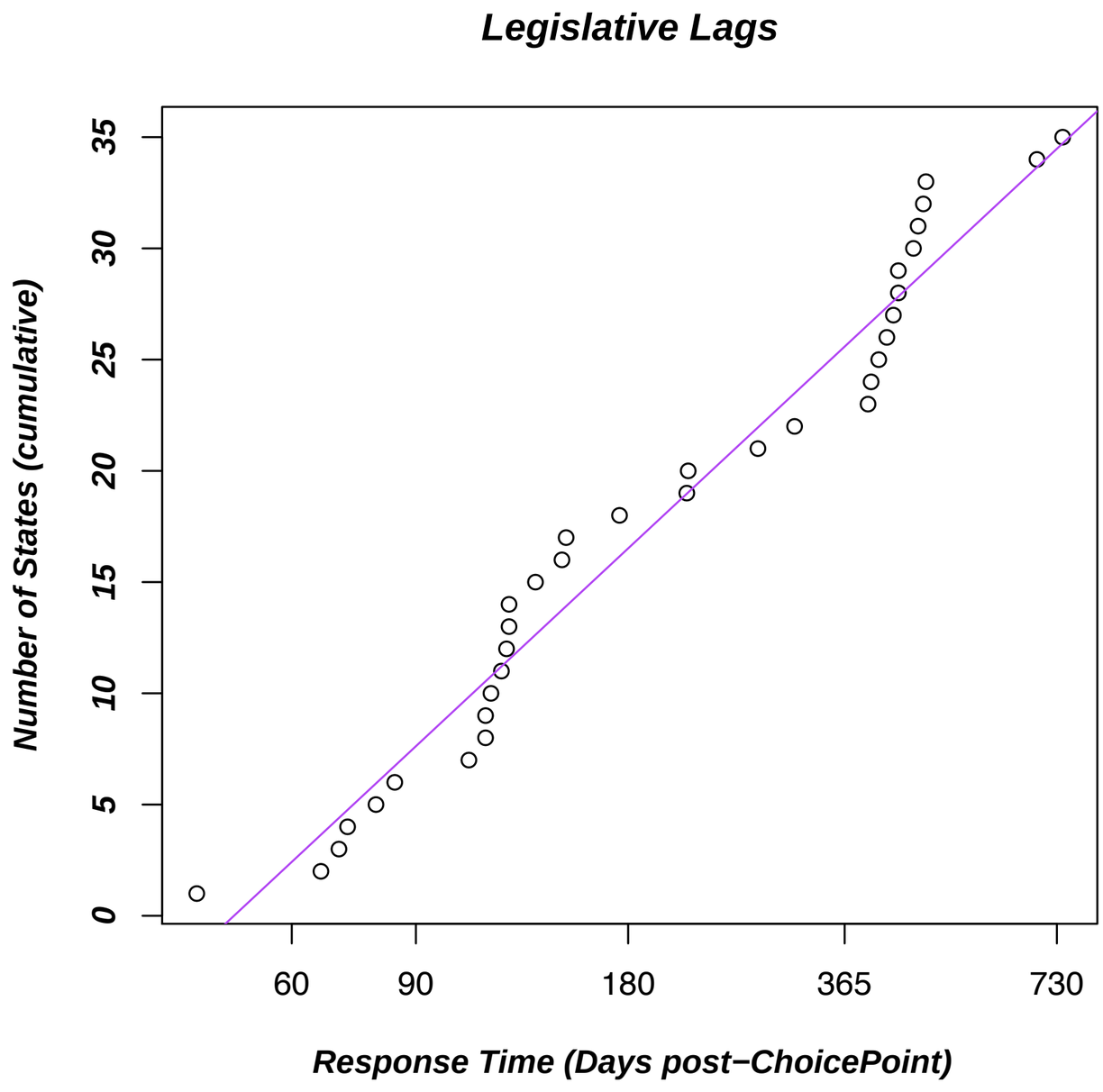


This extremely simple model suggests reporting will not be universally required for several years.

Take that with a grain of salt, but perhaps we should look closely at what these laws offer us and learn from it.

If we take seriously what this “model” suggest, we won’t have full reporting in the U.S. for some time, but for now, the reports which flow from these dozens of state laws already on the books provide an amazing opportunity to both “policy wonks” and “security geeks”. Those of us with a foot in each camp are doubly-blessed :^)

Law passage times grow exponentially



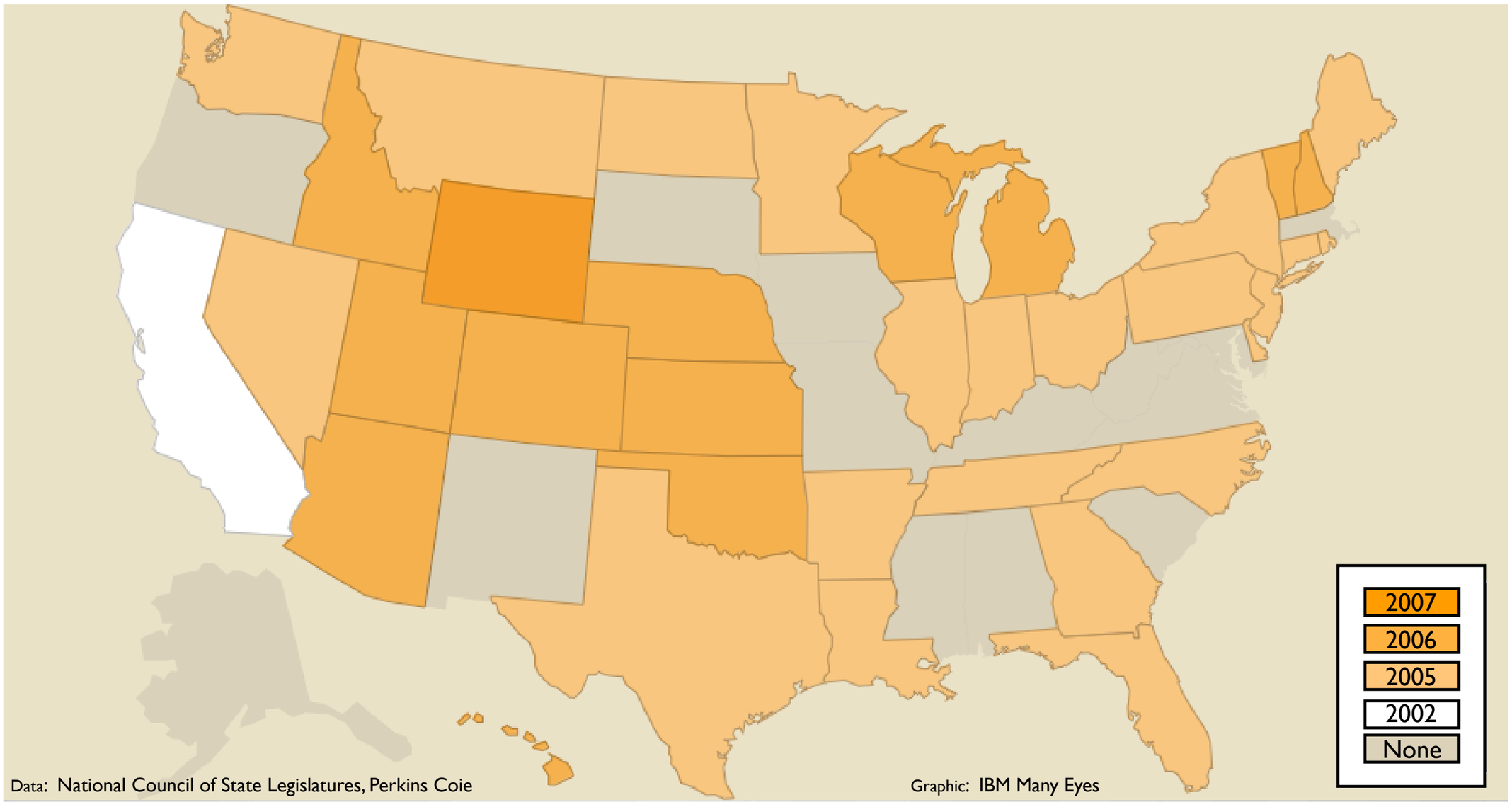
This extremely simple model suggests reporting will not be universally required for several years.

December 17, 2010

Take that with a grain of salt, but perhaps we should look closely at what these laws offer us and learn from it.

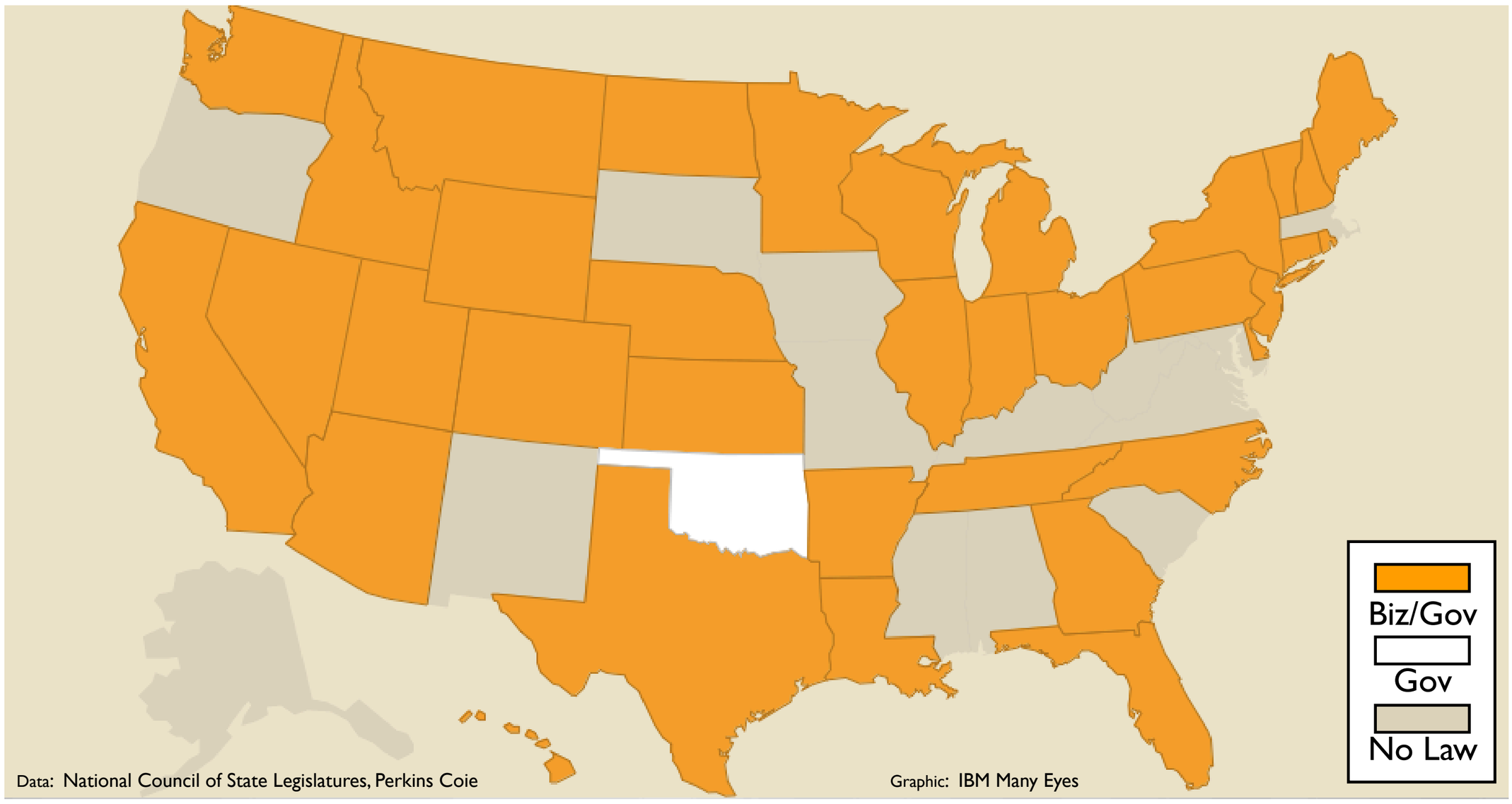
If we take seriously what this “model” suggest, we won’t have full reporting in the U.S. for some time, but for now, the reports which flow from these dozens of state laws already on the books provide an amazing opportunity to both “policy wonks” and “security geeks”. Those of us with a foot in each camp are doubly-blessed :^)

US Data Breach Laws: Date Passed



As we've seen, as of May , 2007, 36 of 50 U.S. states have breach disclosure laws.

US Data Breach Laws: Entities Covered



State laws are not uniform, but in broad terms at least, the vast majority apply to both business and government, and require that owners of personal information notify the persons about whom the information pertains when a security breach exposes that information.

How Do We Know?

Reports required by national regulators

- Oversight committee reports
- FOIA

Reports required by states

-FOIA still needed (except in N.H.) but there are way fewer states than agencies

-Some primary sources available on-line

http://doj.nh.gov/consumer/breaches.html

http://www.cwalsh.org/cgi-bin/docview.pl

Question is: Do they add information, or just "more of the same"?

Test: Look at reports obtained by states, and reports obtained through "traditional means". What, if anything, is added?

The previous few slides have suggested that existing U.S. state laws (and, by extension, those to come) provide a bonanza of useful data which are a boon to researchers.

Yes...BUT

These are not the only sources of information about data breaches. They are, however, the most-used source of what I call "breach-level" data. These reports contain information about individual breaches, and (subject to caveats about samples being representative, etc.) allow us to draw conclusions about individual breaches yet to come (or, better yet, to be avoided).

However, we need to bear in mind the limitations of these data sources.

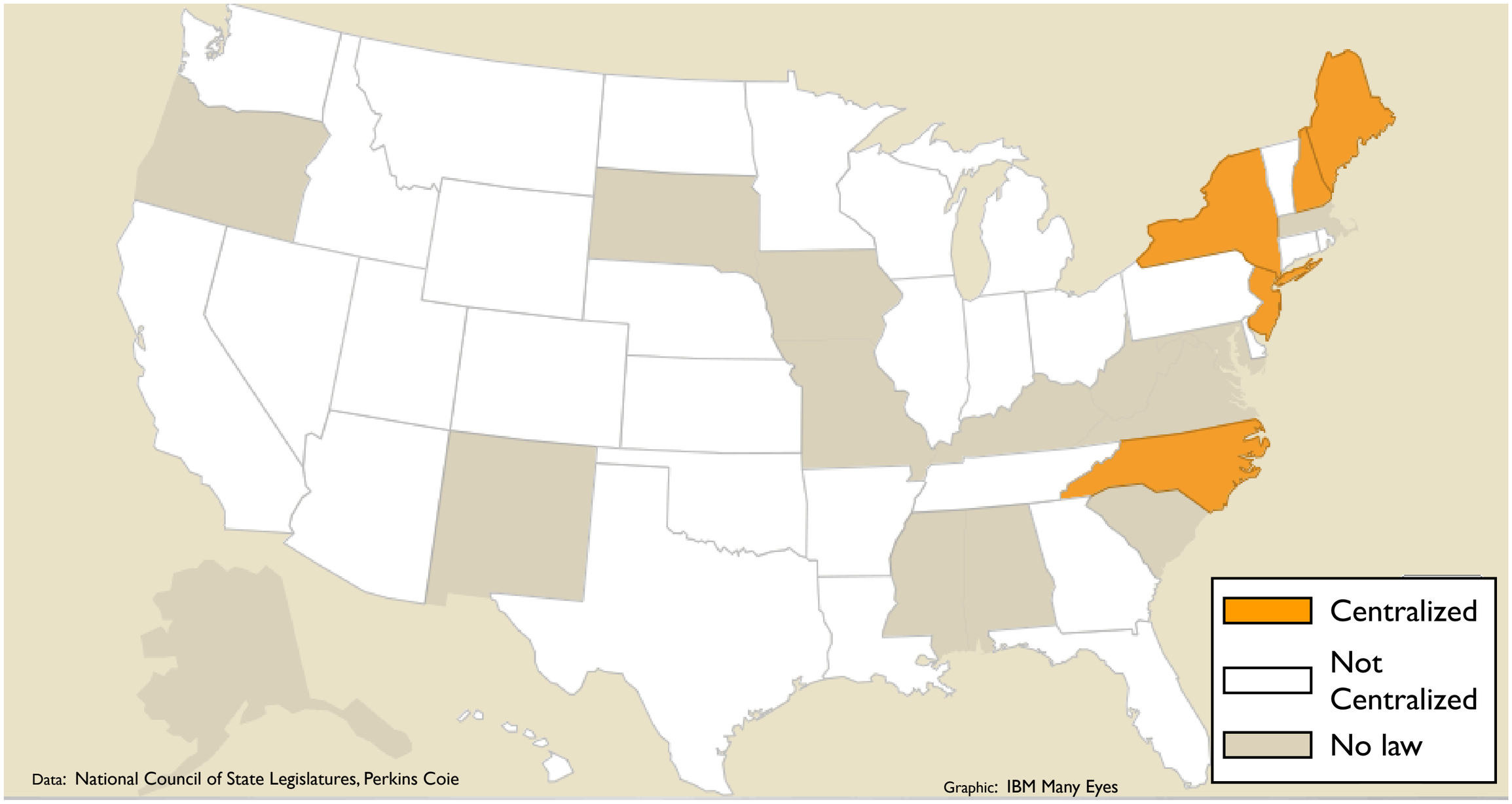
The state laws which lead to these reports are not binding on the Federal government, so the natural question is to what extent the federal government has been breached, what we know (or can learn) about such incidents.

Thanks to a rather significant incident affecting 20 million or so U.S. veterans, various governmental oversight committees have been asking rather pointed questions. Their committee reports in some cases add additional breach-level data, but often this information only exists in these reports as "illustrative examples".

In principle the agencies which have provided the information upon which these reports are based can be asked to produce records under the Freedom of Information Act (a U.S. national law intended to provide open access to non-classified federal government records). It is important to understand that FOIA requests would need to be sent to each relevant agency, so getting a comprehensive picture is, in computer science parlance, an O(n) problem. In the case of the United States Government, n is a large number. I have not located any studies about data breaches based upon such information. While such records may exist, they have not been reported publicly in any comprehensive way. Recalling the iceberg diagram, they are in the "Unreported Incidents" region below the waterline.

But, a little-noticed fact is that a few states require breaches be reported not just to individuals (from whom they may or may not make it into news reports), but also to the state government itself.

Central reporting is uncommon



Only 5 states require breaches to be reported to a state governmental body.

Maine: Reporting requirement only since 1/31/2007. Data availability unknown.

New York: Makes reporting forms and notification letters available via Freedom of Information requests.

New Jersey: Considers reports exempt from public access requirements.

New Hampshire: Makes reports available at http://doj.nh.gov/consumer/breaches.html

North Carolina: Makes table of reported information available via Freedom of Information requests

May make forms and notification letters available (but have not, to date)

What is collected by states?

North Carolina Security Breach Reporting Form
Pursuant to the Identity Theft Protection Act of 2005

Name of Business Owning or Licensing Information Affected by the Breach: _____
Address: _____
_____
Telephone: _____
Fax: _____
Email: _____

PLEASE SUBMIT FORM TO:
Consumer Protection Division
NC Attorney General's Office
9001 Mail Service Center
Raleigh, NC 27699-9001
Telephone: (919) 716-6000
Toll Free in NC: (877) 566-7226
FAX: (919) 716-6050

Date Security Breach Reporting Form submitted: _____

Date the Security Breach was discovered: _____

Estimated number of affected individuals: _____

Estimated number of NC residents affected: _____

Name of business maintaining or possessing information that was the subject of the Security Breach, if the business that experienced the Security Breach is not the same entity as the business reporting the Security Breach (pursuant to N.C.G.S. § 75-65(b): _____
_____

Describe the circumstances surrounding the Security Breach and state whether the information breached was in electronic or paper format: _____
_____

Regarding electronic information breached, state whether the information breached or potentially breached was password protected or encrypted in some manner. _____ If so, please describe the security measures protecting the information: _____
_____

Describe any measures taken to prevent a similar Security Breach from occurring in the future: _____
_____

Date affected NC residents were/will be notified: _____

If there has been any delay in notifying affected NC residents, describe the circumstances surrounding the delay pursuant to N.C.G.S. § 75-65(a) and (c)): _____
_____

If the delay was pursuant to a request from law enforcement pursuant to N.C.G.S. § 75-65(c), please include the written request or the contemporaneous memorandum.

How NC residents were/will be notified?
(pursuant to N.C.G.S. § 75-65(e))
Please attach copy of the notice if in written form or a copy of any scripted notice if in telephonic form.

- ☐ written notice
- ☐ electronic notice (email)
- ☐ telephone notice
- ☐ substitute notice

Signature: _____ Date: _____
Contact Person, Title: _____
Address: _____
(if different from above) _____
Telephone: _____ Fax: _____ Email: _____

Reporting Form
For Business, Individual or NY State Entity reporting a "Breach of the Security of the System"
Pursuant to the Information Security Breach and Notification Act (General Business Law §889-aa; State Technology Law §208)

Name of Business, Individual or State Entity _____
Date of Discovery of Breach: _____
Estimated Number of Affected Individuals: _____
Date of Notification to Affected Individuals: _____
Manner of Notification: [ ] written notice
[ ] electronic notice (email)
[ ] telephone notice

Are you requesting substitute notice? [ ] Yes [ ] No (If yes, attach justification)

Content of Notification to Affected Individuals: Describe what happened in general terms and what kind of information was involved. Please attach copy of Notice.
_____
_____
_____

Name of Business or Individual Contact Person: _____
Title: _____
Telephone number: _____
Email: _____

Dated: _____
Submitted by: _____
Title: _____
Address: _____
Email: _____
Telephone: _____ Fax: _____

New York and North Carolina provide breach reporting forms. Strictly speaking, their use is not required (Warning -- I am not a lawyer!), but the New York form is used in the vast majority of reports to that state. I don't have direct information from North Carolina (yet!).

The NY form asks:

Name of reporting entity

Breach Discovery Date

Reporting Date

Number of affected individuals

When notice was/will be sent

"What happened in general terms and what kind of information was involved"

North Carolina asks these things, and also:

whether the information was password-protected or encrypted
the measures used to protect the information

whether the information was in electronic or paper form

The name of the business that was the subject of the breach, if different from the one doing the reporting

Measures taken to prevent future recurrence.

Having looked at several hundred NY forms, it seems as though the North Carolina form would be better at clarifying certain issues (particularly the case where a 3rd party loses and/or reports). Knowing what measures breached firms put in place to prevent recurrence is extremely interesting.

A Quick Test

Look at incidents involving entities based in New York

Should all be reported to the state, since New Yorkers undoubtedly involved
Should appear in "traditional" reports

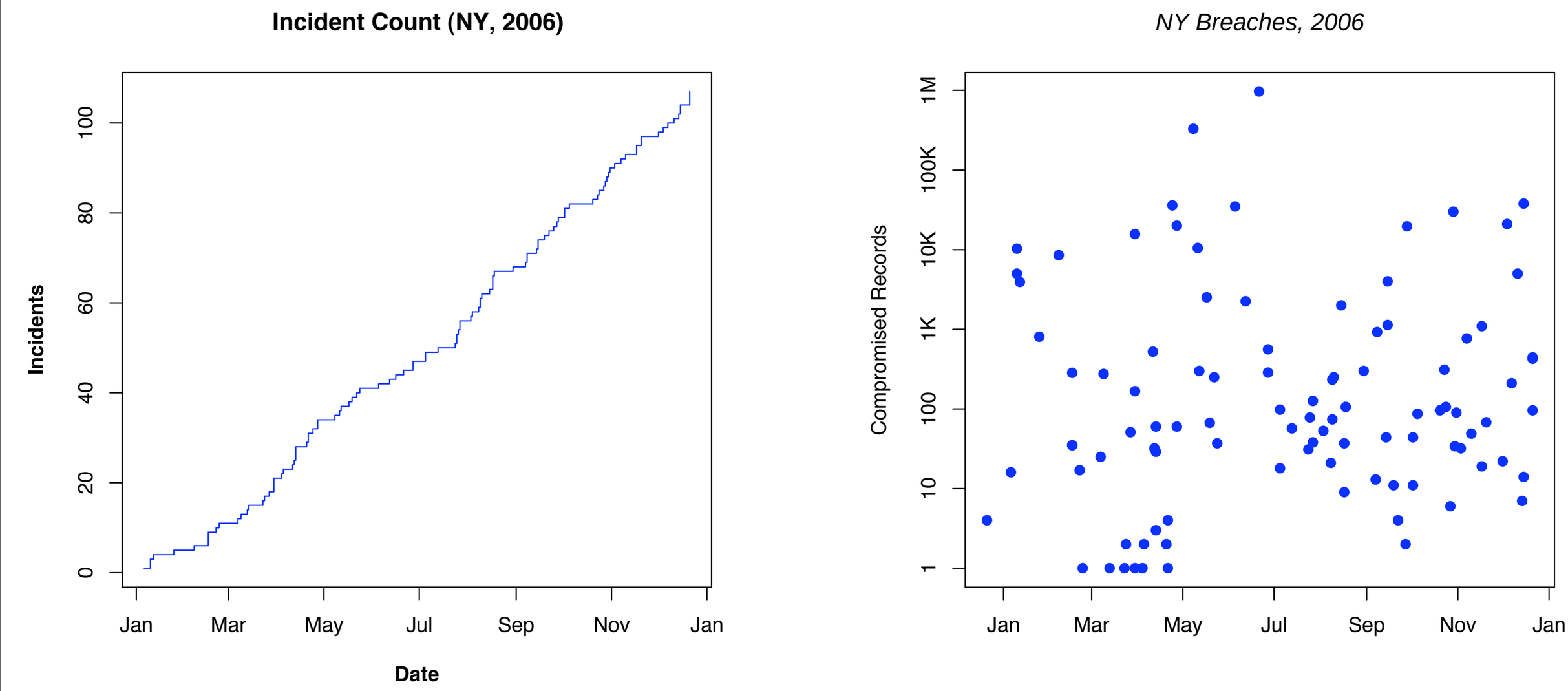
"Traditional" data set

University of Washington (based on Attrition, Privacyrights.org, news reports)

NY reports

Obtained via FOIA requests

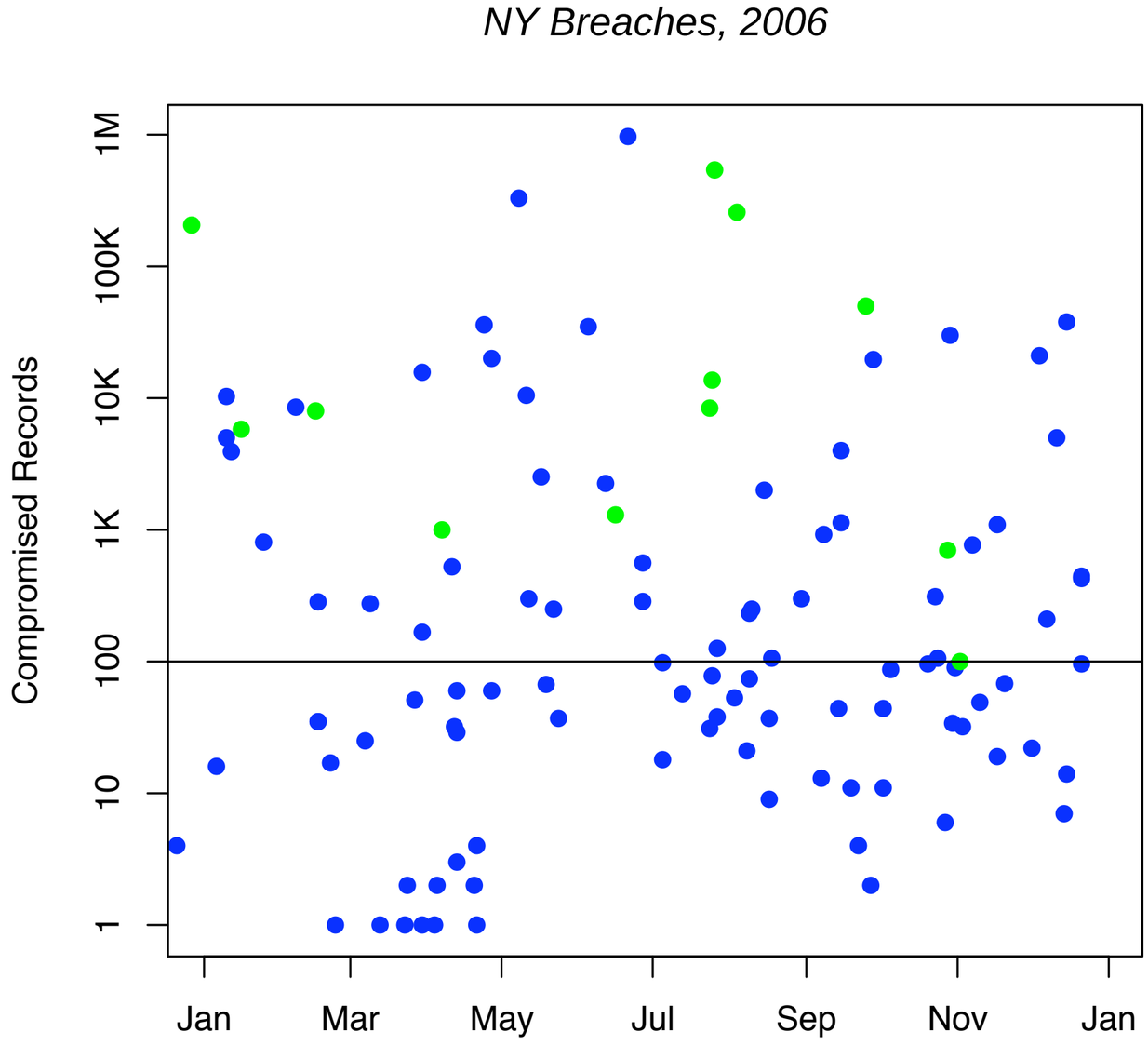
If the picture is markedly different, state reports add value.



These graphs depict breaches of New York firms, as reported to the state in 2006.
The chart on the right shows breach sizes (the Y axis is logarithmic). Clearly, many breaches are quite small.

Interestingly, the incident rate did not vary that much through the year. Naively, one might have expected a certain amount of "learning by doing" as firms became aware of their responsibilities under the law and begin to comply. This is not reflected in the data.

Remember -- what is shown here is restricted to breaches of NY firms. If we consider all reports to the state, regardless of the location of the breached firm, there were about 280 in 2006.



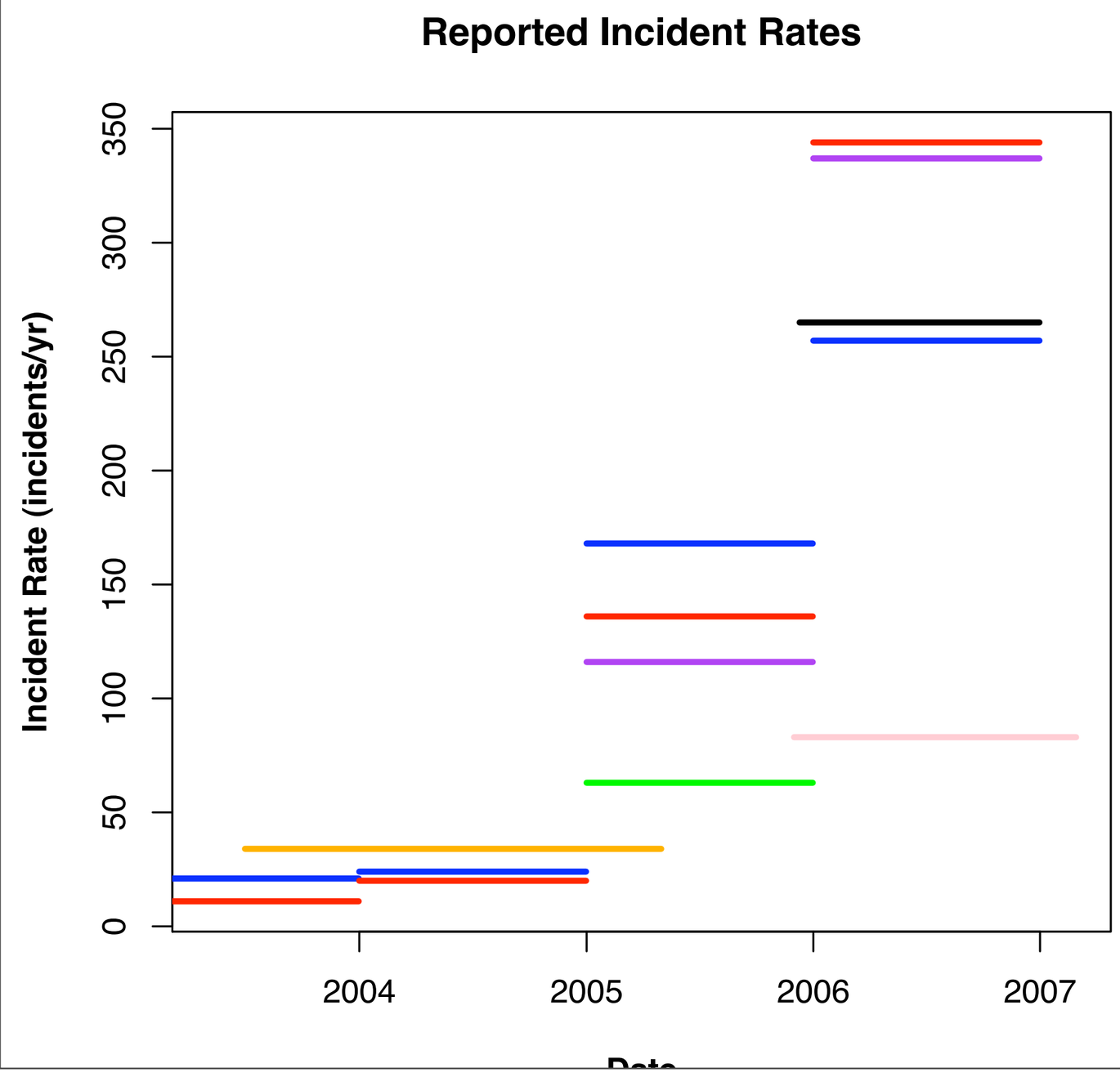
Green: University of Washington
Blue: New York reports

This is new information!

Looking only at the number of compromised records, it is clear that the reports from New York add information. Almost half of the NY reports are for fewer than 100 affected individuals, the minimum in the NY portion of the University of Washington data set.

Additionally, having so many additional cases gives us much greater statistical “resolution”.

If, as is seen here, centralized reporting results in several times more cases available for analysis, we would immediately see U.S. breach incident counts in the thousands annually, even if a substantial amount of incident overlap across states exists.



Line segments show incident observation rates for multiple sources, over time.

So, from the previous slide we know that the observation rate for New York was 280 incidents/year.

Here, we show the number of incidents reported to or available from various sources at different times. The idea is to show whether multiple sources reflect an increasing availability of breach information.

To the extent that two sources with the same scope show different observation rates, that with with lower observation rates is missing information.

The observation rates shown here are:
Attrition.org's DLDOS in red.
The Privacy Rights Clearinghouse chronology of breaches since Choicepoint, in purple.
A breach dataset constructed by academic researchers at the University of Washington, in blue
A breach dataset used by academic researchers at the University of Illinois Urbana-Champaign, in green
A corpus of data drawn from forms submitted to New York state, in black
The observation rate for reports to the state of North Carolina, in pink.
California's rate of "significant breach notifications" is shown in gold.

There is a great deal of information in this graph.

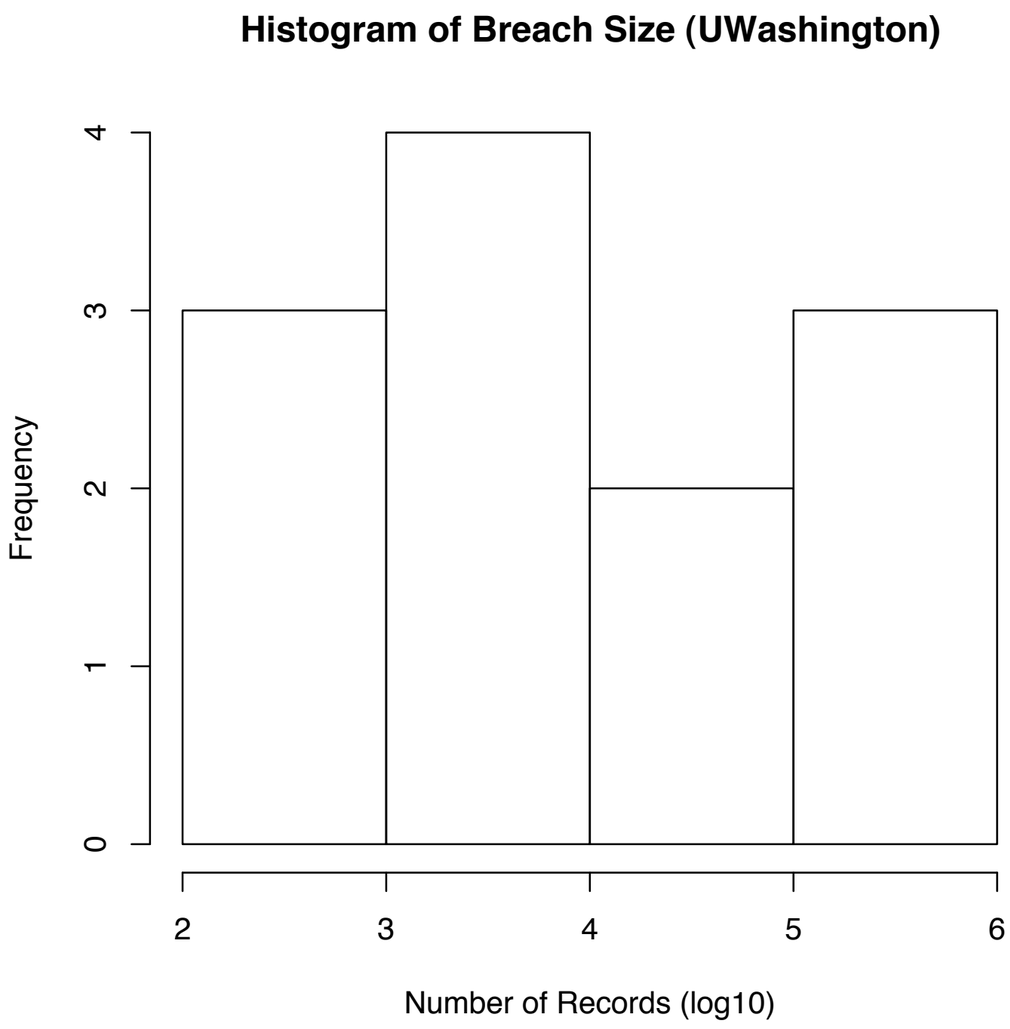
Consider the period ending at 2005. California's rate line is higher than both Attrition's and the University of Washington's. This means that "significant" breaches from California are absent from the Attrition and University of Washington data sets. As it happens, the California incidents are reported to involve 163,500 people on average, so information on millions of records may be missing from the best available sources on this time period.

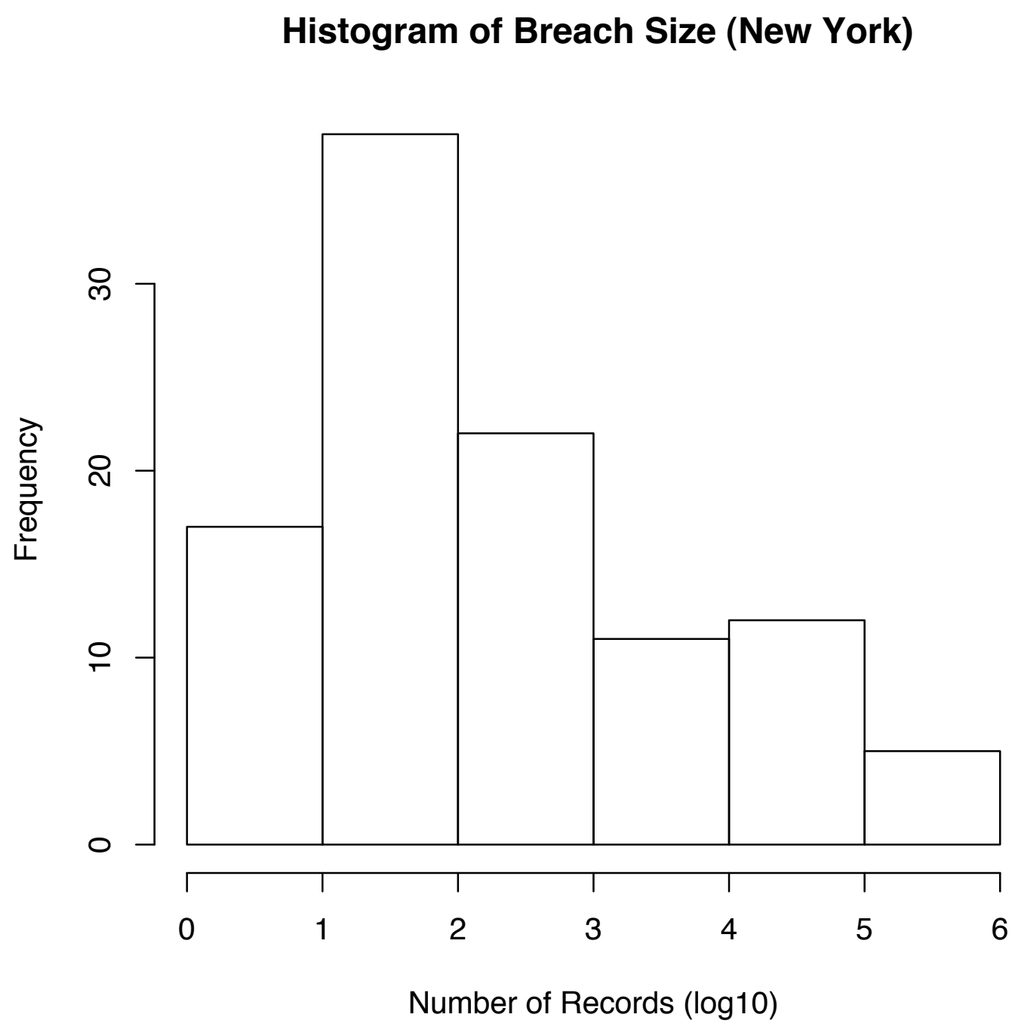
Similarly, the University of Washington dataset is "short" in incidents, if the California report is accurate. However, it is somewhat more complete than Attrition for this time period because the researchers conducted specific media searches to uncover breach reports.

For incidents in 2005, analogous differences between Attrition, UIUC, and University of Washington can be seen, for example.

Finally, it is worth noticing that the 2006 information from NY contains almost as many incidents as the Attrition and Privacy Rights Clearinghouse sources.

The Bigger Stuff makes the news?





Intuitively, it is hardly surprising that bigger breaches would be more likely to be reported in the media, and this appears to be the case.

What are the weak points?

	Exposed Online	External Intrusion	Insider Abuse or Theft	Missing or Stolen Hardware	Mishandled	Other	Unspecified
UWash	3		1	8			
New York	17	7	3	65	2	4	3
New York > 99	5	3	1	37	2	0	2

Results for NY, and for NY cases with more than 99 individuals affected, are statistically indistinguishable

Lesson: Keep track of your stuff, and know how to configure your web server

Here, we see that, at least in terms of numbers of breach incidents, equipment or media loss and unintended online exposure (such as with a misconfigured web server) are the main sources of exposure. Indeed, results from the New York dataset and the New York cases from the University of Washington dataset are statistically indistinguishable, each showing 60–65% of breaches due to lost or stolen media and 15–25% exposed online.

By way of comparison, Attrition.org's DLDOS shows almost exactly 50% (180 of 362) of recorded 2006 incidents being due to lost or stolen equipment or media, and Hasan and Yurcik report 36% of incidents from the period 01-05-2005 through 06-05-2006. North Carolina's breach notification log (obtained via an open records request) shows 53 incidents of 107 (50%) involved lost or stolen media/hardware.

New Hampshire records from December 2006 to June 2007 show 54% of incidents (N=51) due to lost or stolen equipment or media (67% of affected firms since one stolen laptop had 13 firms' data!)

Since so many of the cases reported to NY involve small numbers of persons affected, one might think that the "small incidents" differ from the rest. However, when small (defined as 99 persons affected or fewer) incidents are excluded, the breakdown of breach mechanisms is statistically indistinguishable from an examination with all cases included.

	Exposed Online	Insider Abuse or Theft	Missing or Stolen Hardware
UWash	1.6%	0.5%	97.9%
New York	1.0%	0%	98.7%

Or, maybe ... Just keep track of your stuff!

When we look at the same information, but broken down by the number of records exposed, it is clear that equipment and media loss accounts for the vast majority of exposed records.

	New York	UWash
Utilities	2	0
Manufacturing	2	2
Retail Trade	1	0
Transportation and Warehousing	2	2
Information	2	2
Finance and Insurance	34	2
Educational Services	28	0
Health and Social Assistance	16	2
Arts, Entertainment, Recreation	1	0
Accommodation and Food Service	1	1
Public Administration	14	3
Other Services	1	0

Cells in red show higher than expected counts, those in blue, lower than expected counts.

If we had more information from states, we would be able to understand details such as this better.
There are some signs that this is occurring - information from the state of New Hampshire (one of the four with centralized reporting that isn't shielded from view) -- is now available on-line.

washingtonpost.com June 1, 2005:

The California Department of Consumer Affairs reported May 27 that since the state's notification law went into effect in July 2003, it has been aware of 61 significant breach notifications involving an average of 163,500 individuals each. About one-fourth of the breaches occurred at financial institutions and another one-fourth at universities, with 15 percent reported by medical institutions, 8 percent by government and 7 percent by retailers, according to the figures.

washingtonpost.com June 1, 2005:

The California Department of Consumer Affairs reported May 27 that since the state's notification law went into effect in July 2003, it has been aware of 61 significant breach notifications involving an average of 163,500 individuals each. About one-fourth of the breaches occurred at financial institutions and another one-fourth at universities, with 15 percent reported by medical institutions, 8 percent by government and 7 percent by retailers, according to the figures.

So what now?

Should we only care about lost/stolen media and hardware?

What about low-frequency, huge impact events?
- Massive retailer breaches?
- Card processor breaches?

Small breaches may also be signs of poor practices.

Additional reporting, and clarification of notification requirements would help us get the information we need to make risk decisions.

So we know what accounts for what we can see, and we can see more below the water line thanks to state info.

What can we conclude? Do we only care about lost/stolen media or equipment??

Biggest impacts to date have not involved these (TJX, Cardsystems), but VA did.
Additionally, even small breaches may indicate “exceptionally poor practices” that are “unlikely to come to public light if small numbers of individuals are told of them.” Centralized reporting allows us to “track trends in security breaches, large and small, and to determine whether entities are providing adequate protection for information” [Mulligan:2007zr]

Hard to truly know frequency of these large events. More reporting would help.

Discuss disincentives using

How to encourage reporting?

National differences in approach to privacy

Dan Geer testimony
Fed Letter
Law Review Article
Adam Psychology and graphics

Dont want to suppress reporting

ASSuming we get reports, what else do we need?
- Impact data (ID theft, eg. ID Analytics)
- Correlates of breaches (Rezmierski)
- Corroborating info on # records breached

More states' information would help

- Would let us get a better handle on (seemingly) rare events
- Would expose biases (if any) in current, "traditional" reporting
- Would help us to assess whether breaches tend to be local, regional, or national
- Would better inform national and international policy makers
- Would better reveal the role of third parties as "impact magnifiers"

How to obtain this additional information?

- Revise existing laws to add central reporting
- Adopt breach notification requirements beyond U.S.
- Pass US Federal legislation
- Increase voluntary notification

Revise existing laws

- Require reporting to state Attorney General or consumer protection agency
- Standardize reporting to enhance comparability of states' data
- Close loopholes so that breached entity must report, whether it owns data or not.

Adopt breach notification requirements beyond U.S.

While privacy protections afforded to data subjects are significantly greater in many non-US nations, the extent to which these translate into different rates of data exposure is not known.

Pass US Federal Legislation

Legislation on a national level would eliminate a blind spot: federal agencies not bound by state law

Central reporting is critical: eliminates need to individually request data from scores of agencies

Increase Voluntary Reporting

- Higher notification trigger, but mandatory reporting to central entity?
- As means of limiting possible subsequent legal liability
  - If you tell people, they can take steps, and thereby limit *your* risk
- Normative pressure: Customers expect it, law or no law
- Honesty never killed anybody: TJX sales rise after they tell of very large breach!
- Reflexive secrecy could be punished by regulators: why risk it?
- It's an assurance game: Sharing helps all if sufficient numbers share. We just need to get there.

Things We Might Care About

Breach consequences

Impact on stock price

Impact on customer loyalty/"churn"

Direct notification costs

Impact on identity theft

Repeat offenders? Do they learn?

Aspects of the notifications themselves

Do they show acceptance of responsibility?

Is there a clear "CYA" tone?

What level of detail do they provide?

Do standard forms increase the amount of information provided?

Thanks