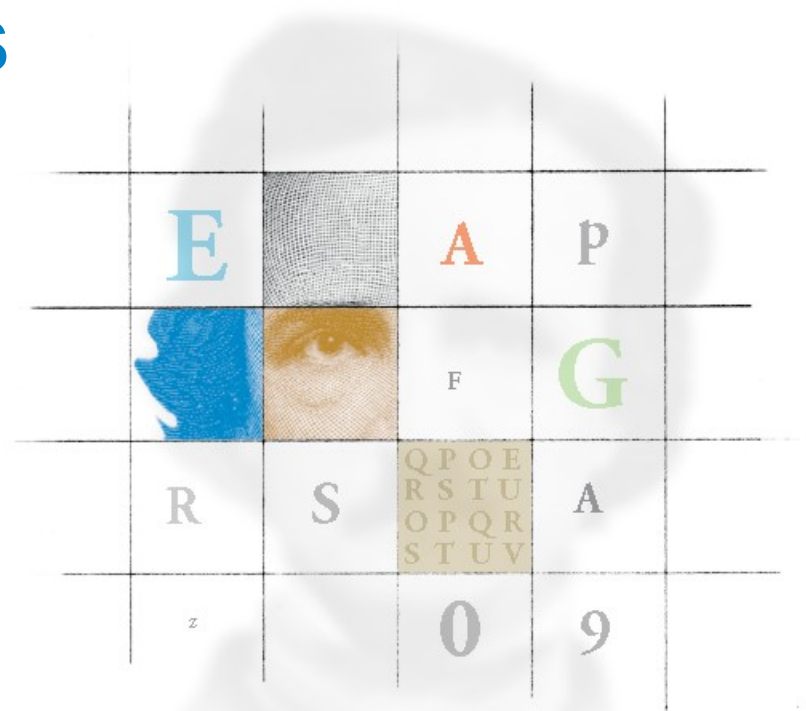


# RSACONFERENCE2009

## Crunching Metrics from Public Security Data

Elizabeth A. Nichols, Ph.D.  
PlexLogic, LLC  
04/21/09 | Session ID: RR-105

Session Classification: ADV



# RR-105 Session Abstract

---

This session looks at what we can learn from three open, public sources of security data: DataLossDB and OSVDB **and NVD**.

This talk will present metrics derived from these sources via simple statistical techniques and analysis using quantitative models for correlation and forecasting.

Additionally, the methods and technology used to derive the results will be **very briefly** described.

# Agenda

---

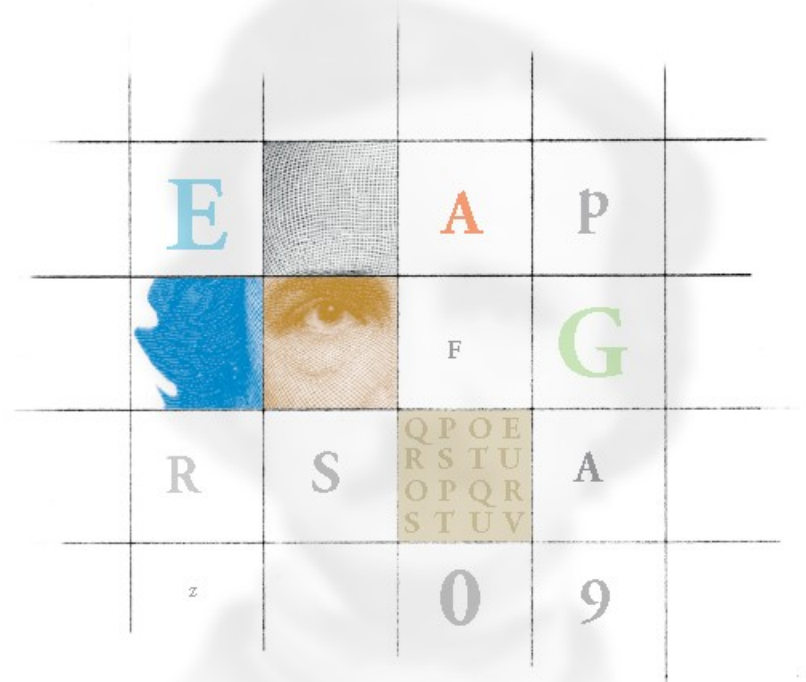
Public Security Data Sources

Metrics, Models, and Results

Automation

Lessons Learned

# Public Security Data Sources



# Good News & Bad News

---

- Provide Insight
- Expose Deficiencies
- Test Assumptions
- Scientific Rigor
- Drive Improvement
- Drive Questions
- Drive Normalization

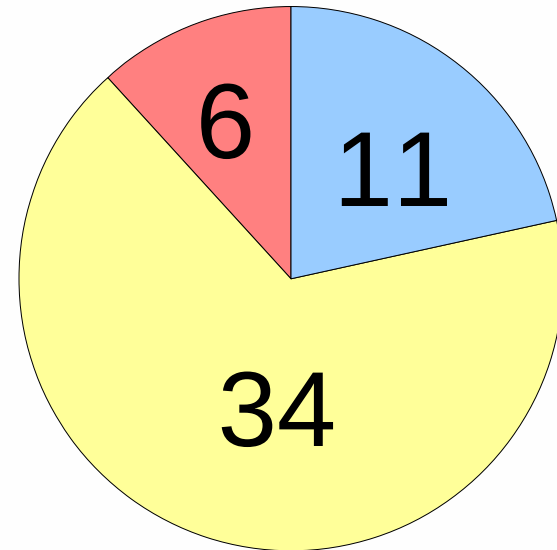
- Completeness
- Information Quality
- Raw Source X-ref's
- Disclosure Bias
- Researcher Bias
- Editorial Policy
- Mis-Representation
- Mis-Interpretation

# Sources Used Here

---



# State of the States: Breach Laws



None De-Centralized Centralized

None = no breach disclosure laws

De-centralized = Laws but no centralized repository

Centralized = Laws and a centralized repository

# DataLossDB

---

- Primary Entities
  - Breaches
  - Companies
  - Associations between Companies and Breaches
- Primary Attributes
  - Time
  - Market Segment and SubSegment
  - Source: Insider or Outsider
  - BreachType (stolen, hack, web, etc)
  - DataLost (CCN, SSN, etc)
  - Total Affected

# OSVDB and NVD

---

- **Primary Entities**
  - Vulnerabilities
  - Products or Configuration Items
  - Associations between Vulnerabilities and Products
- **Primary Attributes**
  - Time
  - Vendor, Product, Version
  - Contributor/Source
  - Type
  - NVD Only: CVSS (base) Score

# Metrics & Models



# Organization

---

- DataLossDB
- OSVDB & NVD
- NVD
- Questions
- Metrics + Models
- Results + Insight



---

# Breaches

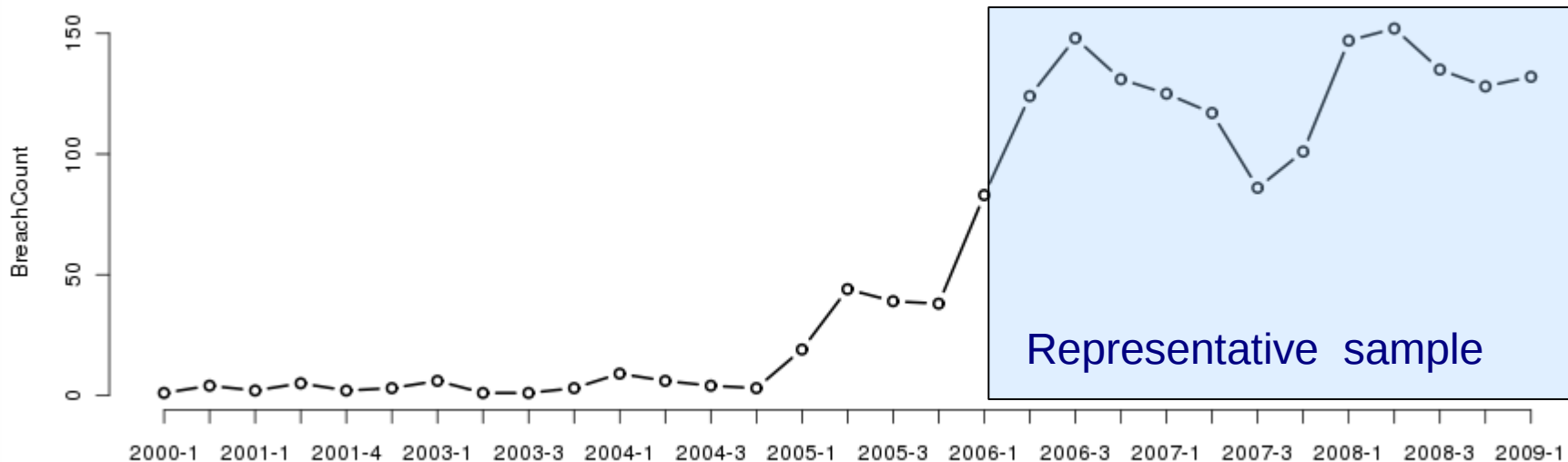
# DatalossDB: Questions, Metrics, Models

---

- Change over time: Is X increasing?
  - Metrics: BreachCount
  - Models: Regression
- Independence of Effects: Does X affect Y?
  - Metrics: BreachCount, TotalAffected
  - Models: Contingency tables, ChiSquare
- Population differences: Is Group A significantly different from Group B?
  - Metrics: Total Affected (TA)
  - Models: Analysis of Variance

# DataLossDB: Whole Magilla

DatalossDB: BreachCount/Quarter  
ALL to Tue Apr 7 15:11:54 2009



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
<b>1.00</b>	<b>3.50</b>	<b>38.00</b>	<b>58.03</b>	<b>124.50</b>	<b>152.00</b>

Dataset = All breaches in the DatalossDB

# DatalossDB: Data Sets Used

Feature	events	ds	ds.logTA
Records included	All records	year > 2005	year > 2005 and TotalAffected > 0
# Rows	1806	1616	1139
# Rows Omitted	0	190	
% Rows Omitted	0	11%	32%
# Columns	21: Added year, quarter, month	22: events + InOut	25: ds + logTA, MitigationType, ClientServer
Substitutions	None	Segment: Biz/Gov → Biz SubSegment: Fin/Med → Fin	InOut: Inside* → Inside MitigationType ClientServer

# Breaches

---

Count & Percent Metrics

# BreachCount/Company: Top 10 List

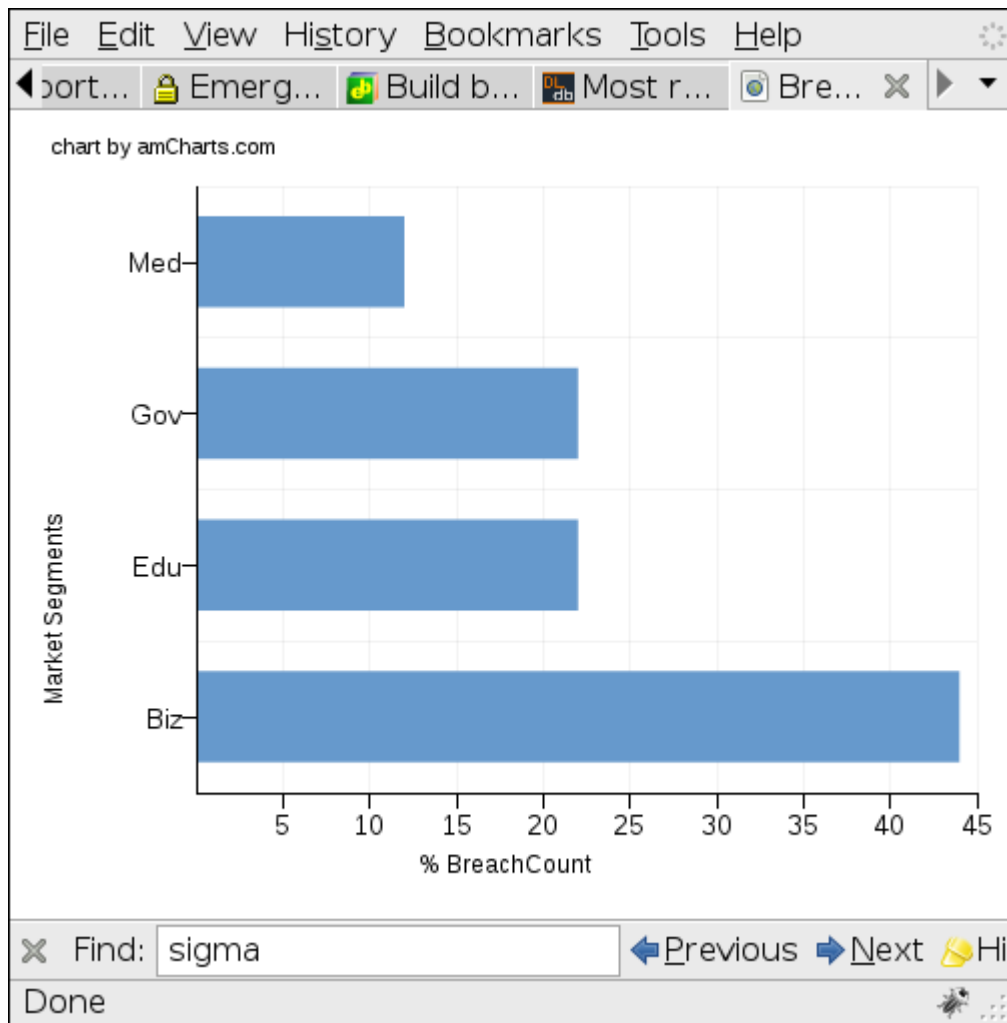
---

## DatalossDB: Most Breached Companies All DB as of Tue Apr 7 13:50:50 2009

Company	BreachCount
University of Iowa	9
Bank of America	8
UK Ministry of Justice	8
LPL Financial	7
Pfizer	7
The Foreign and Commonwealth Office	6
City University of New York	5
Countrywide Home Loans	5
Merrill Lynch	5
Purdue University	5
University of California San Francisco	5
University of Florida	5
U.S. Department of Veterans Affairs	5

Dataset = events = All breaches in the DatalossDB

# BreachCount (BC) by Segment



**BC/Segment :**

**Biz Edu Gov Med**  
**711 356 348 201**

**Company Count/Seg.**

**Biz Edu Gov Med**  
**615 283 296 186**

**CompanyPct/Segment**

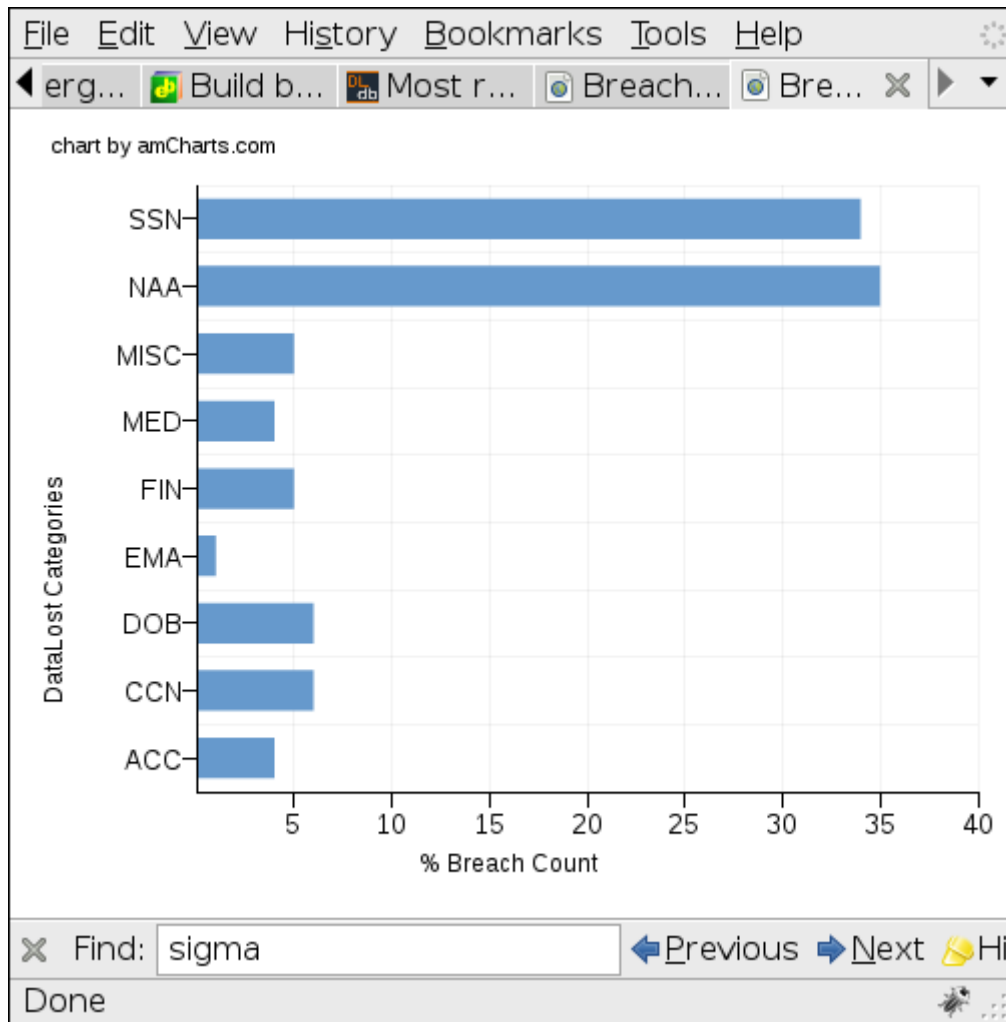
**Biz Edu Gov Med**  
**0.45 0.21 0.21 0.13**

**Mean BC%/Company**

**Biz Edu Gov Med**  
**1.16 1.26 1.18 1.08**

Dataset = ds = All breaches such that year > 2005

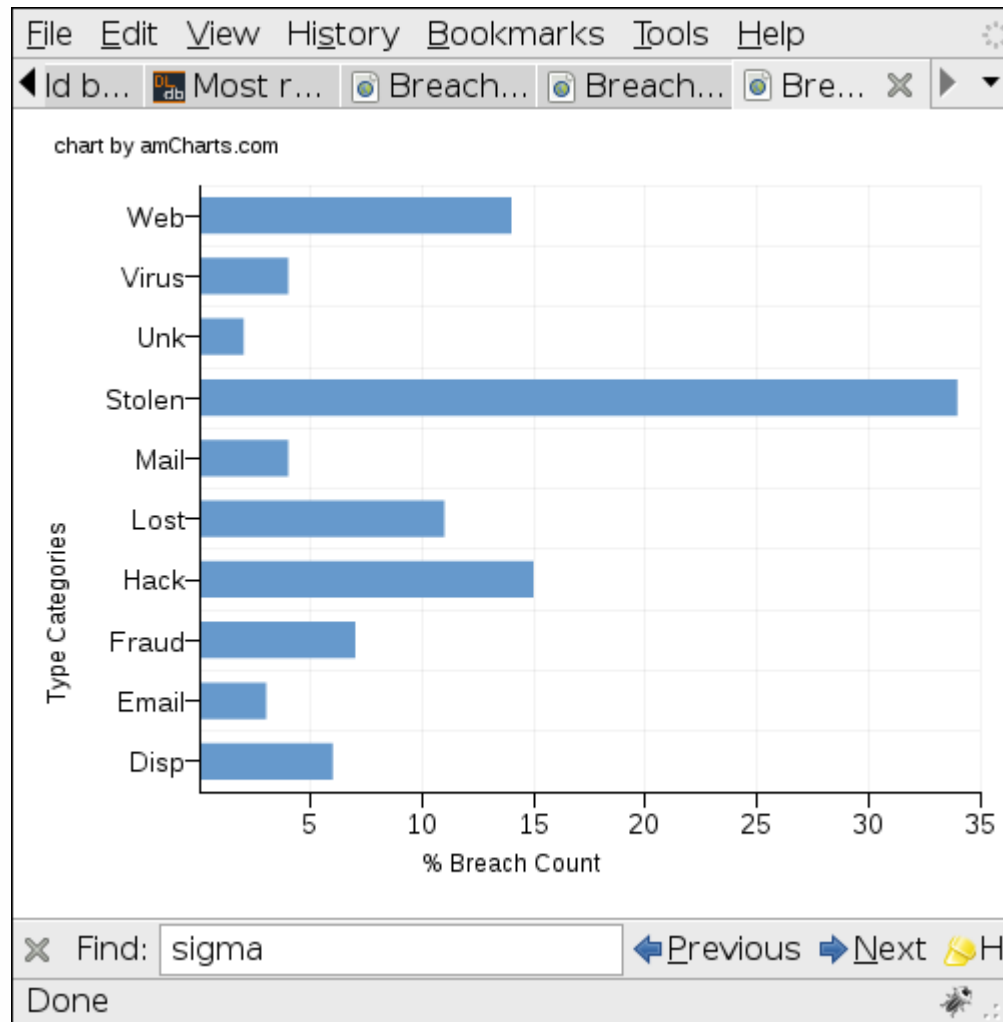
# BreachCount Percent by DataLost



Note that breaches may be associated with many DataLost categories. So total of all % here will be greater than 100%.

Dataset = ds = All breaches such that year > 2005

# Breach Percent by Type



Dataset = ds = All breaches such that year > 2005

# Breach Count:

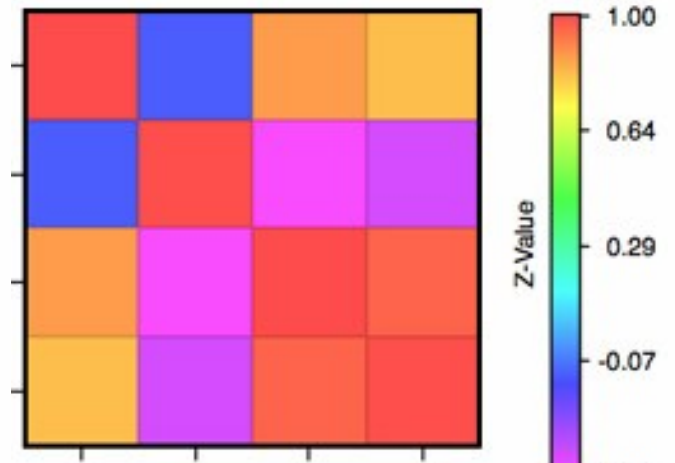
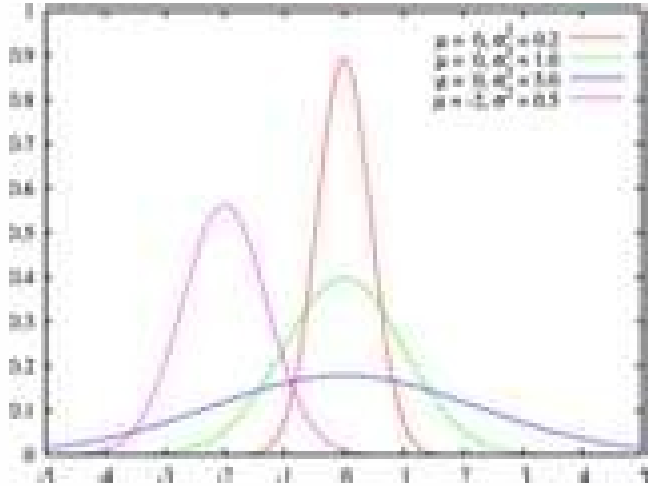
---



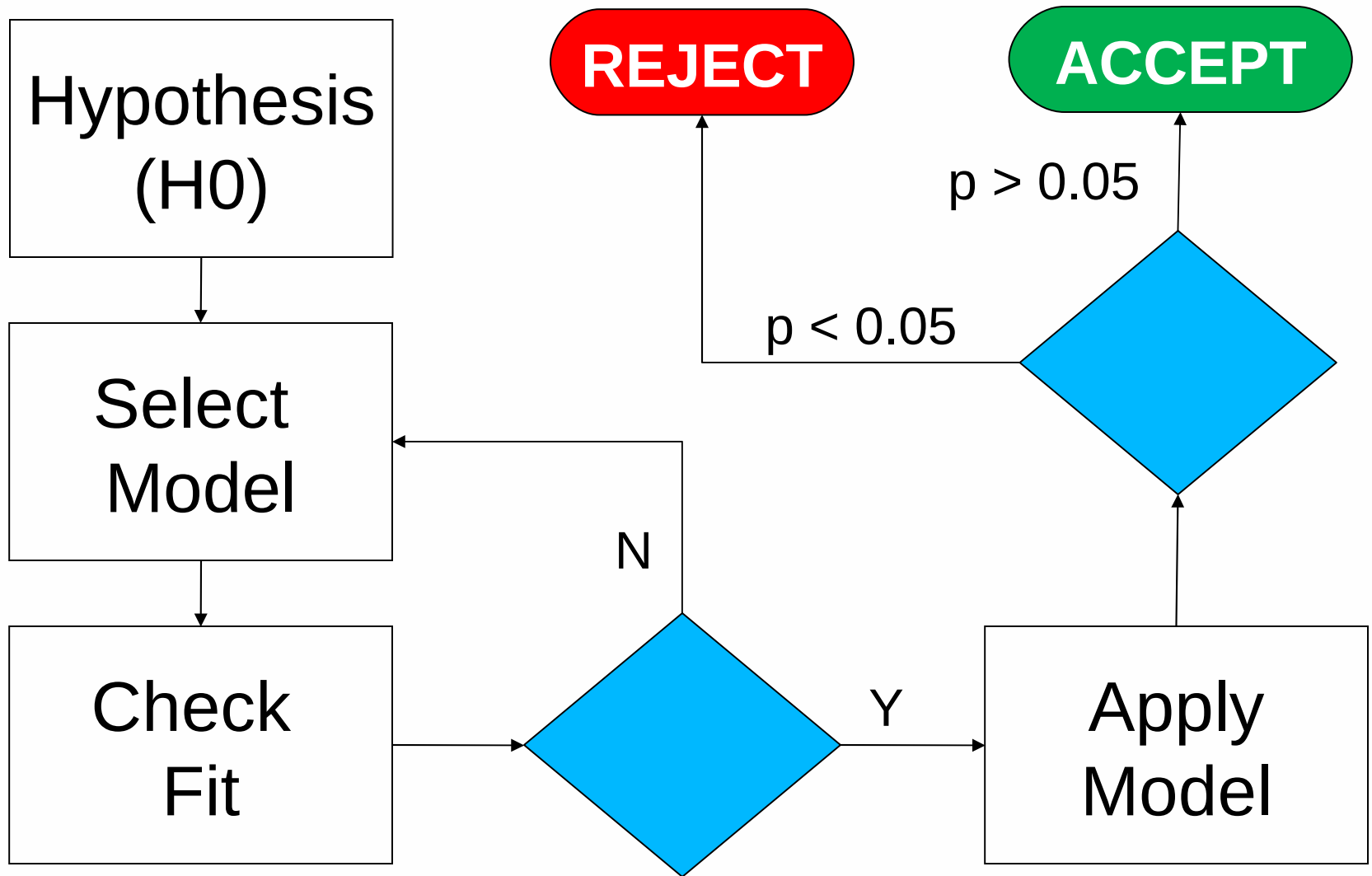
- 9 = Max BreachCount/Company
- All segments are comparable
- SSN and NAA dominate data loss
- Stolen dominates breach type

*Are you really sure ?*

# Models

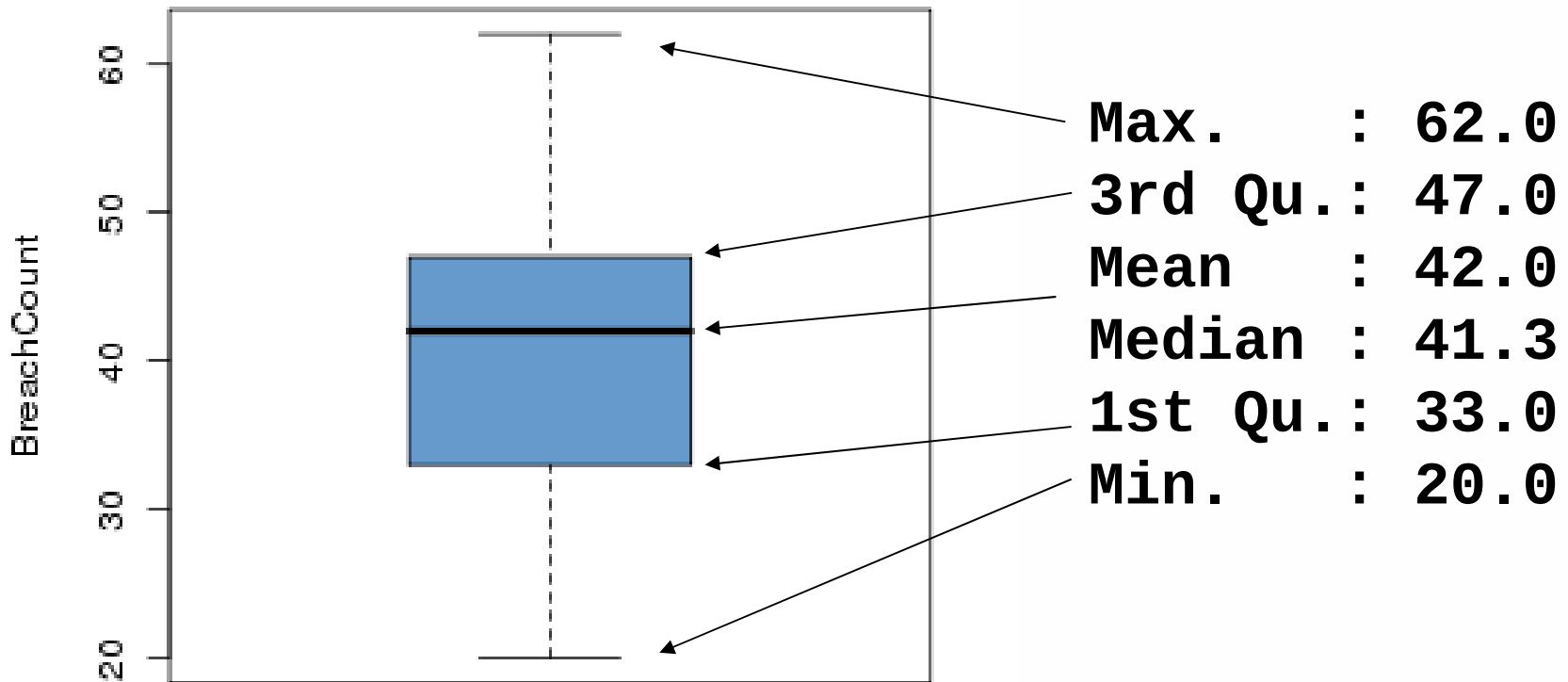


# Statistical Analysis



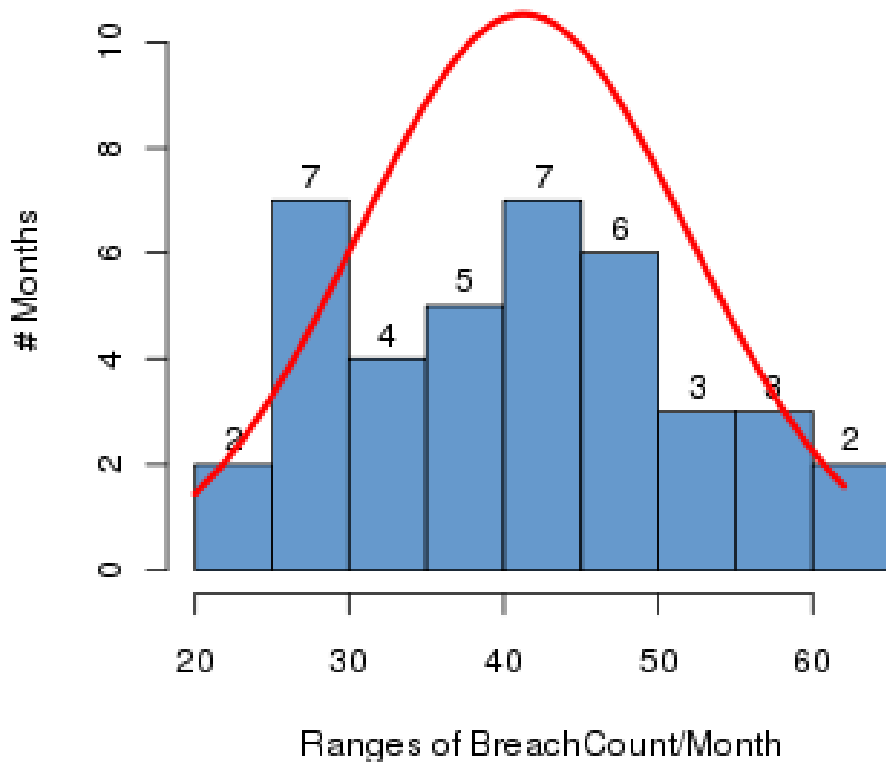
# BreachCount/Month Distribution

DatalossDB: Distribution of BreachCount/Mon  
1Q06 to Tue Apr 7 15:12:45 2009



# BreachCount/Month Histogram

DatalossDB: Histogram of BreachCount/Month  
1Q06 to Tue Apr 7 16:59:50 2009



HO: This distribution is normal

	Skew	Kurtosis
Value	0.12	-0.87
t-value	0.32	-1.11
p-value	0.38	0.86

Shapiro-Wilk Test

W = 0.9771,

p-value = 0.5981

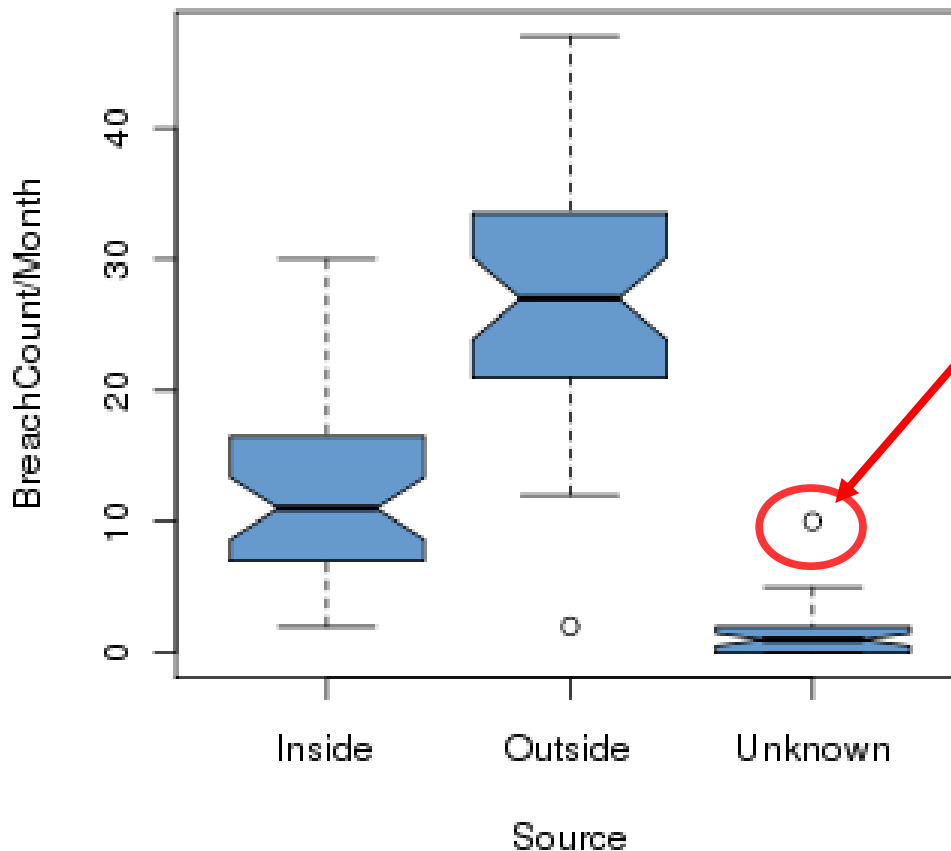
p-values reflect the probability that we could get these results by chance if the distribution were actually normal.

Speaking very precisely: Our data is not inconsistent with a normal distribution.

**CONCLUSION: ACCEPT**

# BreachCount/Month vs Source

DatalossDB: BreachCount/Mon by Source  
1Q06 to Tue Apr 7 15:12:45 2009



## Notches:

If the notches of two boxes do not overlap, this is “strong evidence” that the medians of the two populations are different.

## Outliers:

Y value is  
 $> 3Q + 1.5 (3Q - 1Q)$   
Or  
 $< 1Q - 1.5 (3Q - 1Q)$

## Chi-Square Test:

H0: Source and Breach Count are Independent

→ p-value = 2.2e-16

**CONCLUSION: REJECT**

# Breach Count Per Month:

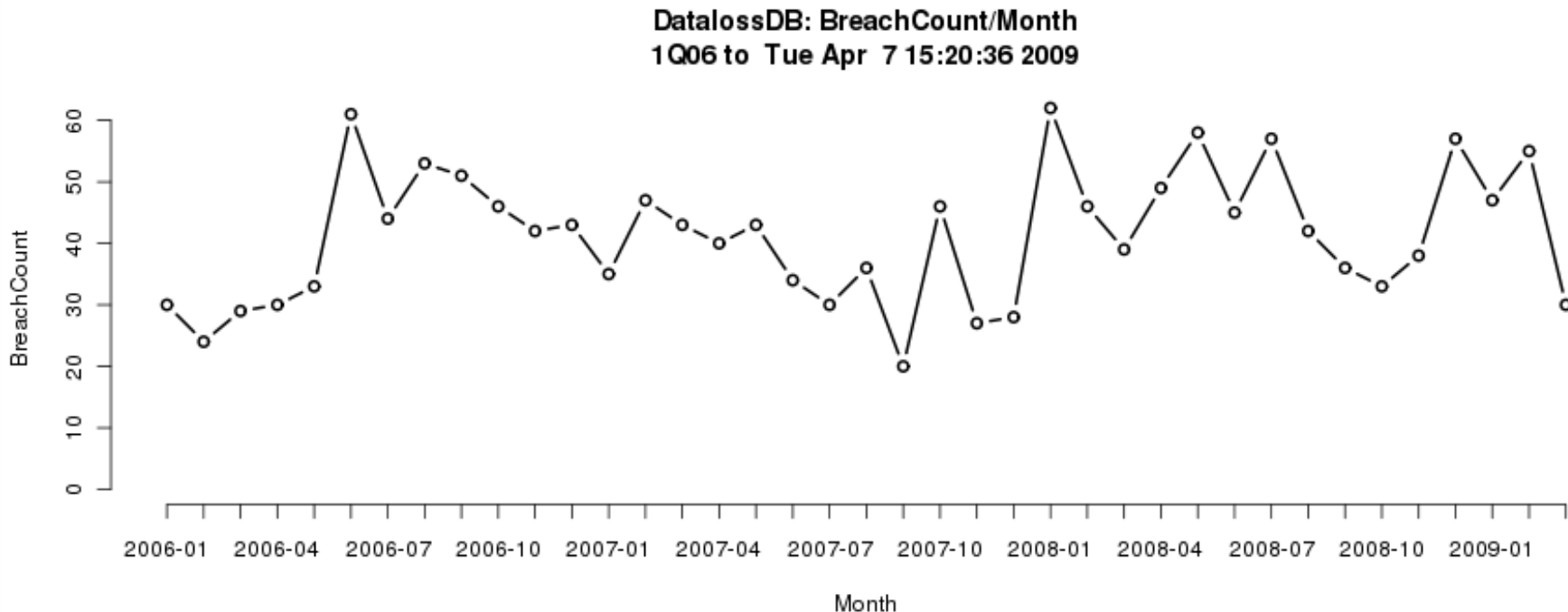
---



Inside BC/mo < Outside BC/mo

*Are you really sure ? Over 99.9%*

# DatalossDB: Time Series

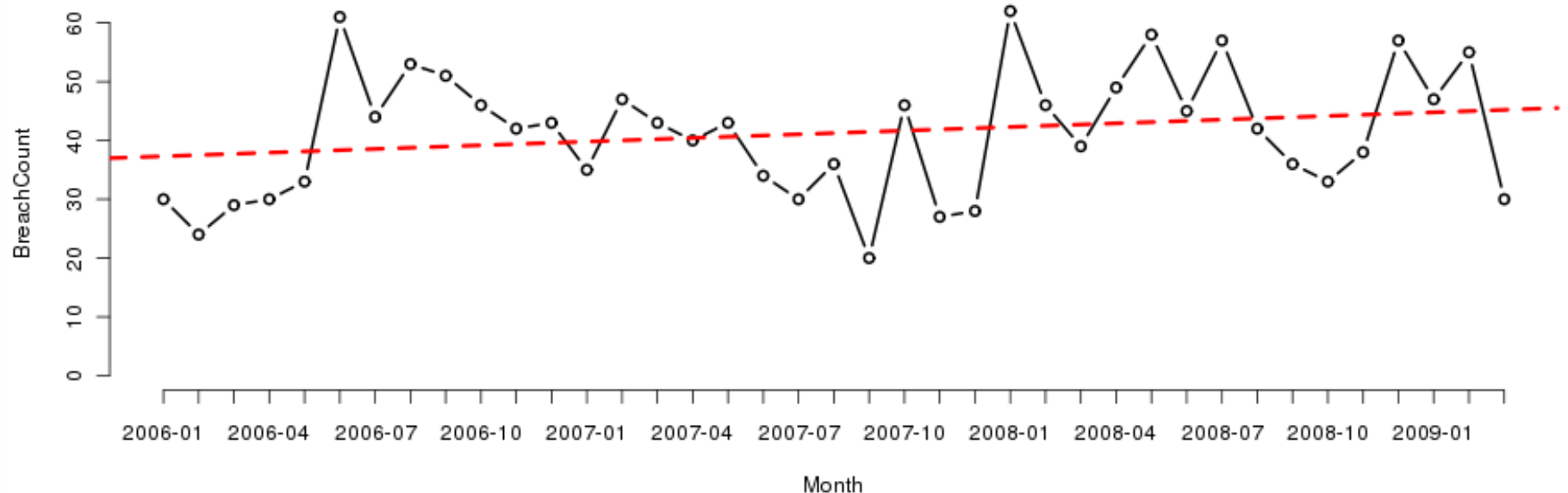


Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
20.0	33.0	42.0	41.3	47.0	62.0

Dataset = ds = All breaches such that year > 2005

# BreachCount/Qtr: Linear Model

DatalossDB: BreachCount/Month  
1Q06 to Tue Apr 7 15:20:36 2009



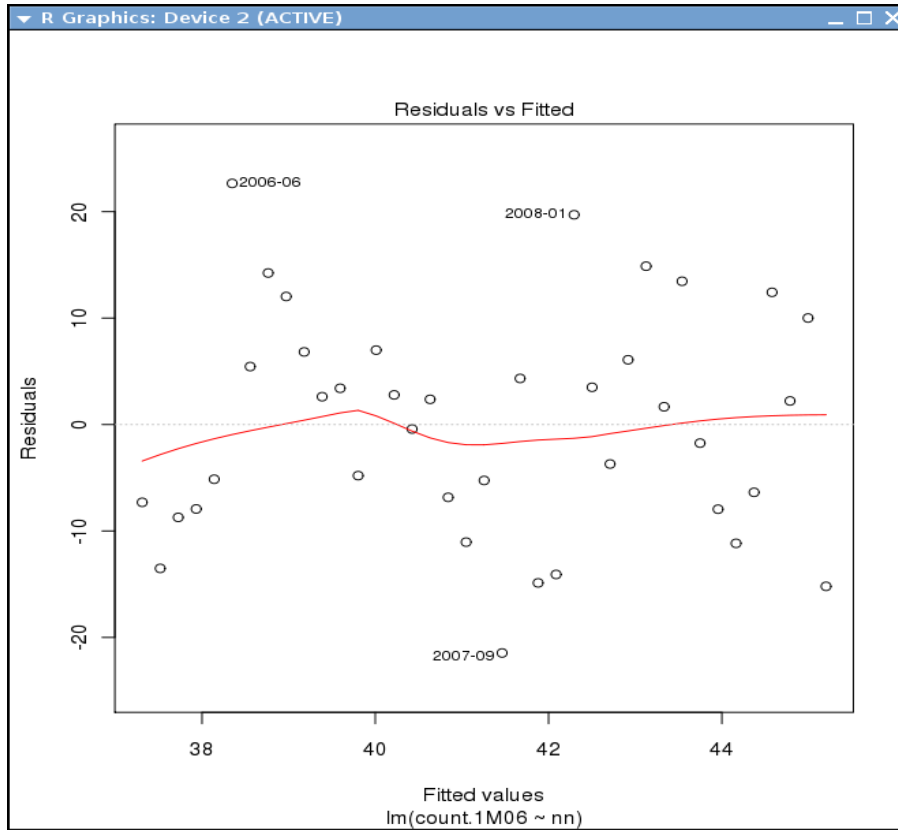
"Best fit" line thru the points: { (month, BreachCount/qtr)  
| month > 2005-12 }

What does a linear regression model tell us?

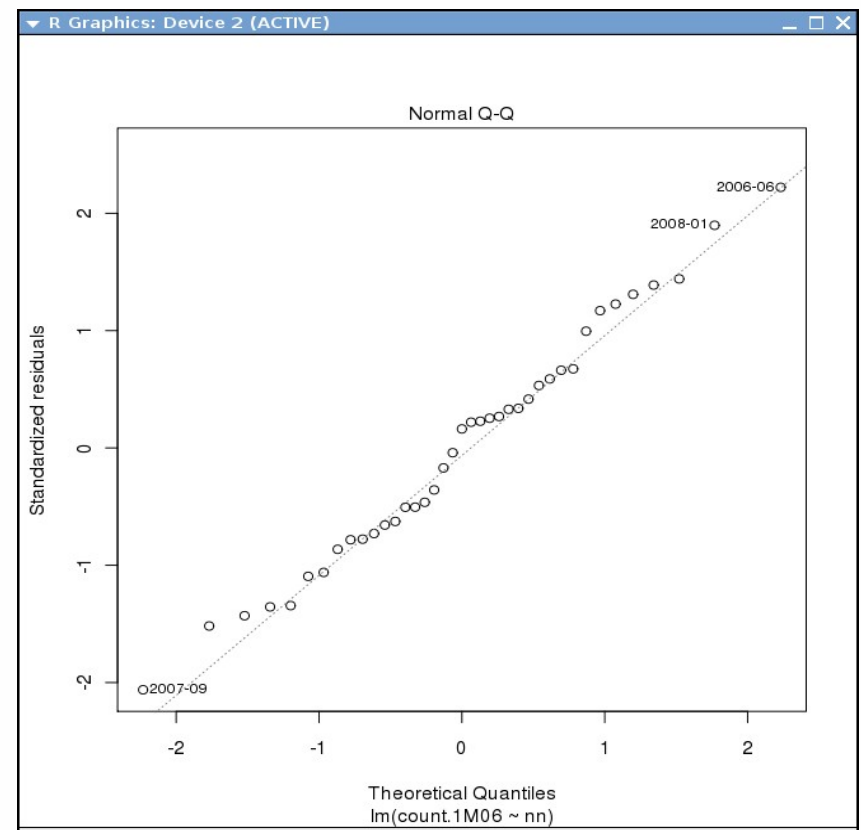
Is it a good model for this data?

What conclusions can we reach and with what confidence?

# Breach Linear Model Checking



Linearity Test: Should look random and not increase L to R. Plots Residuals vs Predicted values.



Normality of Residuals Test: Normal Quantile-Quantile plot. Should form a straight line.

# Is BreachCount/Quarter Flat?

*H0: BreachCount/Month slope = 0.*

**Residuals:**

Min	1Q	Median	3Q	Max
-21.464	-7.622	1.667	6.451	22.651

**Coefficients:**

	Estimate	Std. Error	t value	Pr(> t )	
Intercept	37.1026	3.4440	10.773	5.8e-13	***
Slope	0.2077	0.1501	2.184	0.175	

Residual standard error: 1055 on 37 degrees of freedom

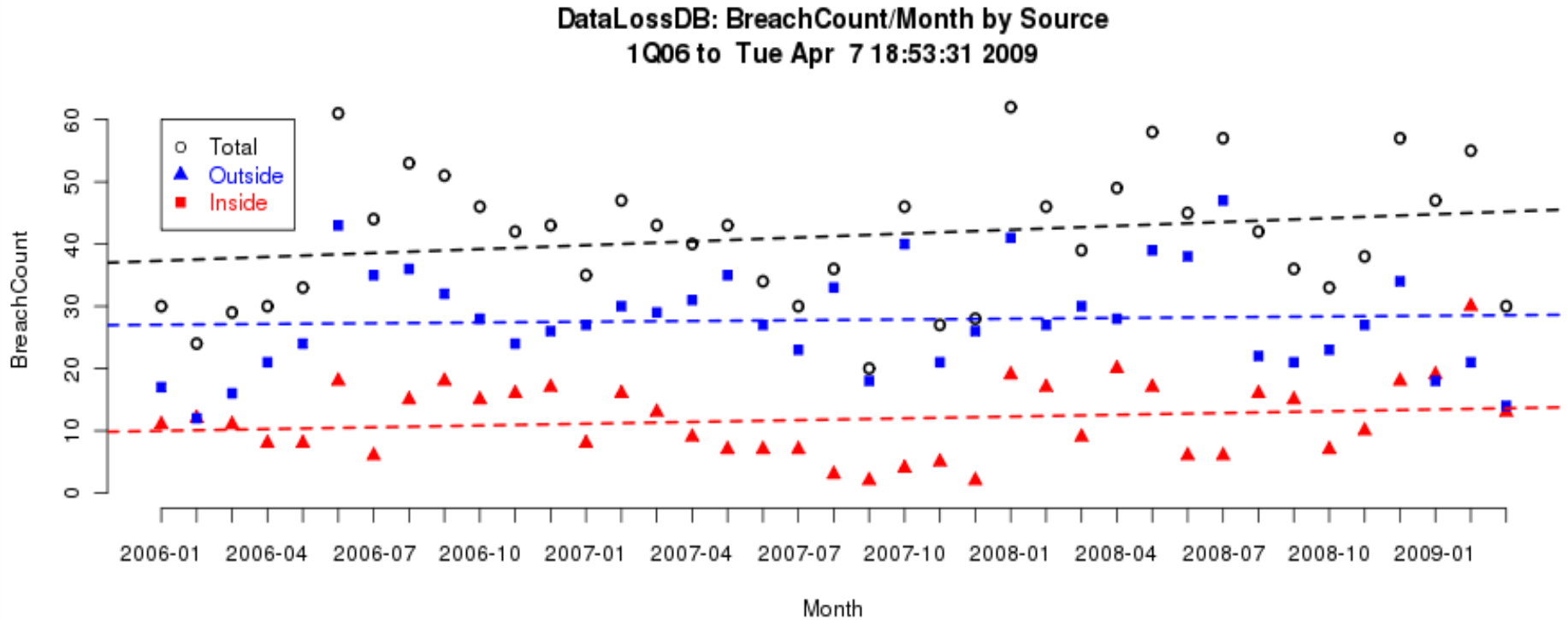
Multiple R-squared: 0.04922. (SSR/SSY)

F-statistic: 1.915 on 1 and 37 DF, **p-value: 0.1747**

The probability that we will be making a Type 1 Error (by rejecting a true hypothesis) is 17.5%.

**CONCLUSION: ACCEPT.**

# Breaches/Quarter: Insider vs Outsider



Source	Intercept	Slope
Total	37.1	.21
Outside	27.0	.04
Inside	13.7	.10

# Breach Count Trend:

---



- Total is flat (slope = 0)
- Inside is flat
- Outside is flat

*Are you really sure ? About 80%*

# Breaches

---

Impact

# DataLoss DB: Top 10

---

**DatalossDB: Companies with Highest Impact Breaches**  
**TotalAffected in Millions**  
**All DB as of Tue Apr 7 13:50:50 2009**

<b>Company</b>	<b>TotalAffected</b>
TJX Companies Inc.	94
Visa, MasterCard, American Express	40
America Online	30
U.S. Department of Veterans Affairs	28
HM Revenue and Customs	25
T-Mobile	17
Bank of New York Mellon	12
GS Caltex	11
Dai Nippon Printing Company	9
Fidelity National Information Services	8

# Breach Impact as Influenced by ...

---

- Factors: Year, Segment, Source, Breach Type, DataLost, MitigationType, ClientServer
- What are the differences in the groups that these factors define?
  - Mean value, variance of their distributions
  - Independence
- With what level of confidence can we say this?
- Models
  - Analysis of Variance (Homog. variances, normally distributed)
  - Chi Square (adequate cell size, similar dist, one cell/obs)

# New Breach Factors: Mappings

## Segment:

“^Biz” → “Biz”  
“^Edu” → “Edu”  
“^Gov” → “Gov”  
“^Med” → “Med”

Levene  $p < .001$

## ClientServer:

### “Client”:

“^Disposal”    “^Email”  
“^Fraud”       “^Lost”  
“^Missing”     “^Snail”  
“^Stolen”

### “Server”:

“^Hack”         “^Virus”  
“^Web”            Levene  $p = .12$

## InOut:

### “Inside”:

“Inside”    “Inside – Accidental”    “Inside - Malicious

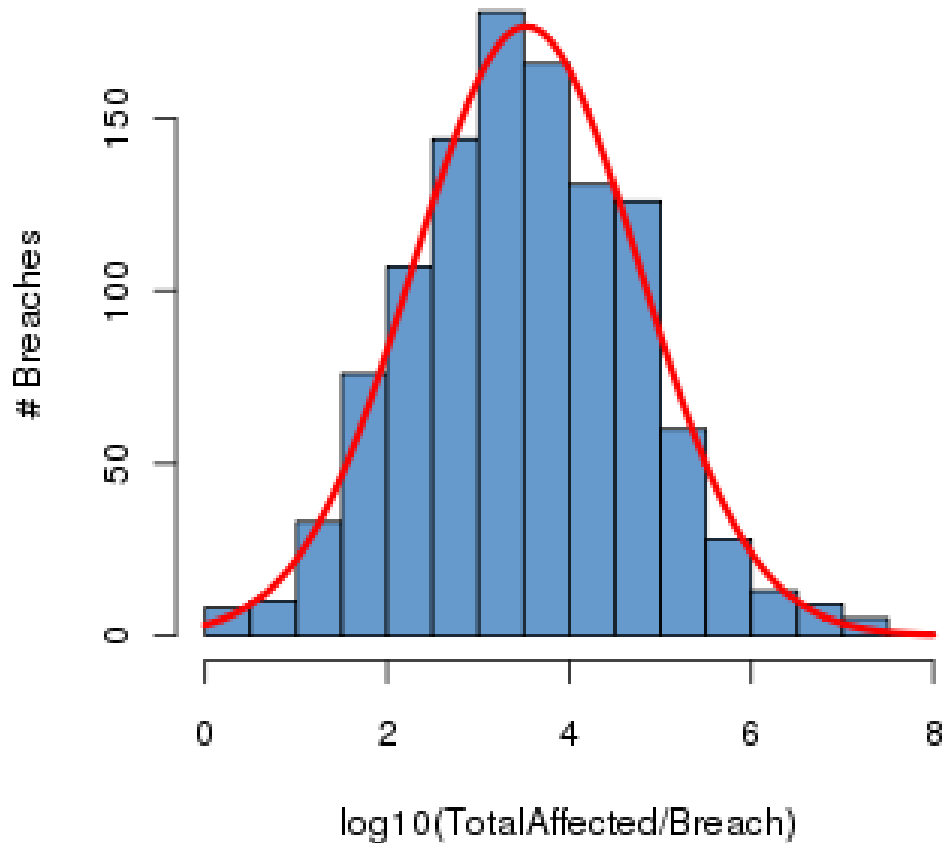
### “Outside”:

“Outside”

Levene  $p = .05$

# Breach Impact: log10 Transform

DatalossDB: log10(TotalAffected) Histogram  
1Q06 to Sat Mar 7 12:37:10 2009



H0: Normal Distribution

ds\$TotalAffected

min 0

median 663

mean 185,800

max 94,000,000

sd 2,614,917

Shapiro-Wilk p << 0.001

log10(TotalAffected)

min 1

Median 3.48

mean 3.51

max 7.97

sd 1.23

Shapiro-Wilk p > 0.05

# TotalAffected:

---



- Is normally distributed
- Groups are homoscedastic
- We satisfied the criteria for ANOVA

*Are you really sure ? Over 95 %*

# Analysis of Variance: Segment

Residuals:           Min           1Q       Median           3Q           Max  
                  -3.57770 -0.86274 -0.01576    0.84901    4.39543

## Coefficients:

Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	3.577699	0.061612	58.068	< 2e-16 ***
Segment[T.Edu]	-0.279851	0.092829	-3.015	0.00263 **
Segment[T.Gov]	-0.006949	0.095364	-0.073	0.94192
Segment[T.Med]	0.058130	0.113995	0.510	0.61020

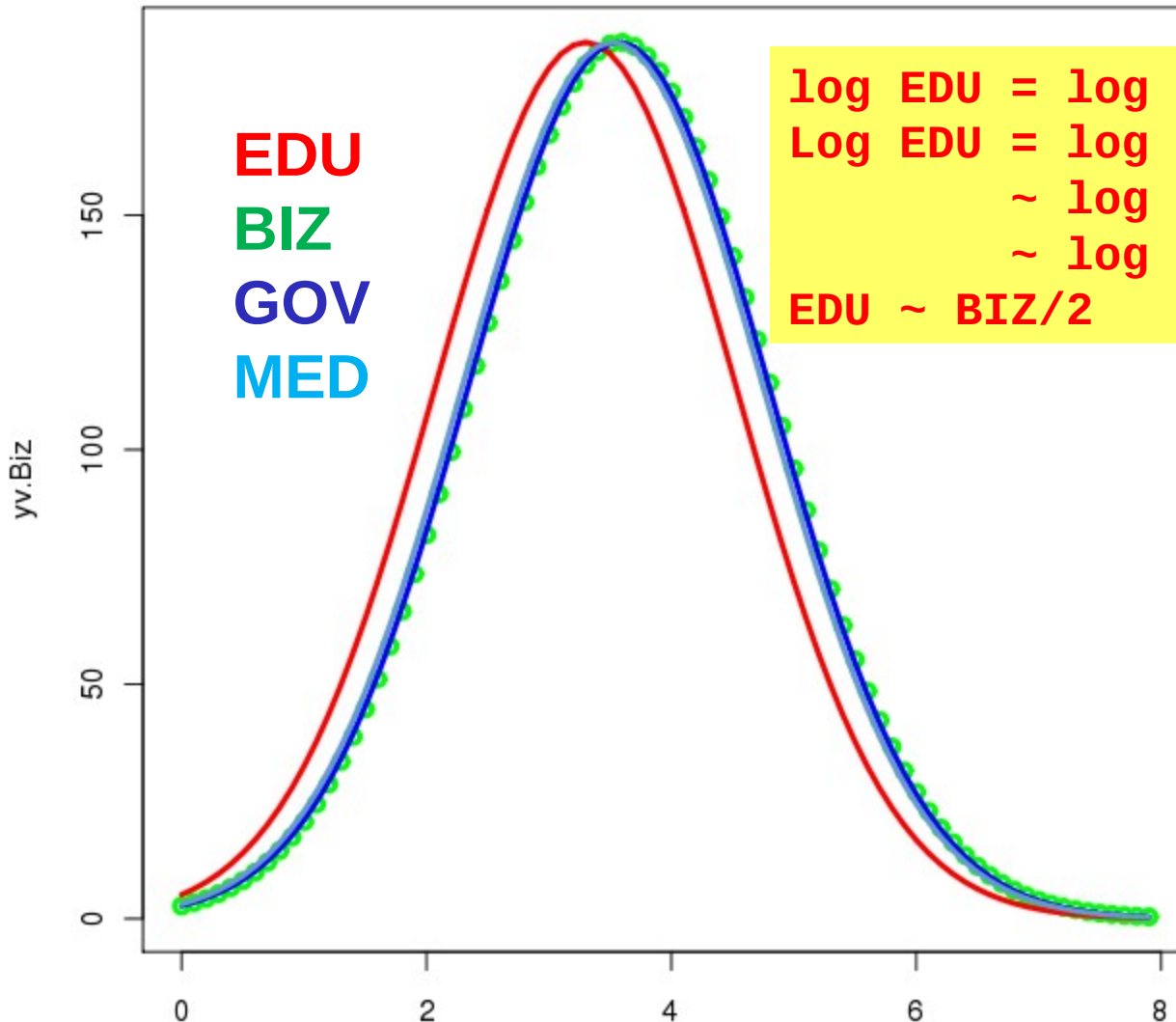
## Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.225 on 1148 degrees of freedom  
Multiple R-squared: 0.01116, Adjusted R-squared: 0.008578

F-statistic: 4.32 on 3 and 1148 DF,  
**p-value: 0.004876**

# Analysis of Variance: Picture



$$\begin{aligned}\log \text{EDU} &= \log \text{BIZ} - 0.28 \\ \text{Log EDU} &= \log \text{BIZ} - \log 10^{**0.28} \\ &\sim \log (\text{BIZ}/10^{** .28}) \\ &\sim \log (\text{BIZ}/2) \\ \text{EDU} &\sim \text{BIZ}/2\end{aligned}$$



EDU breaches are  
Half the size of  
BIZ breaches

# Log10(TA) Factors: Means + Outliers

Factor	Sig.	Base + Outliers	Delta
Year	---- .95 >.999	2006	3.61
		2008	-0.18
		2009	-0.69
InOut	---- >.999	In	3.12
		Out	+0.55
BreachType	---- .95 .95 .95	Disposal Computer	3.53
		Disposal Tape	+2.86
		Stolen Comp/Drive	-2.75
		Stolen Tape	+1.39
ClientServer	----		3.50
MitigationType	----		3.55

$10^{*.5} > 3.1$   
 $10^{*1.5} > 32$

$10^{*2} = 100$   
 $10^{*2.5} > 316$

$10^{*3} = 1000$   
 $10^{*3.5} > 3163$

# Total Affected per Breach:

---



- Breaches associated with tape disposal or stolen tapes have the highest impact by a factor of at least 300X
- Discounting the obvious outliers, breaches from the various categories of fraud, hacking, web breach types are all about the same in terms of impact.
- Breaches associated with clients as opposed to servers appear to have the same impact.
- Breaches that are more amenable to pro-active mitigation appear to have the same impact as those that are not.

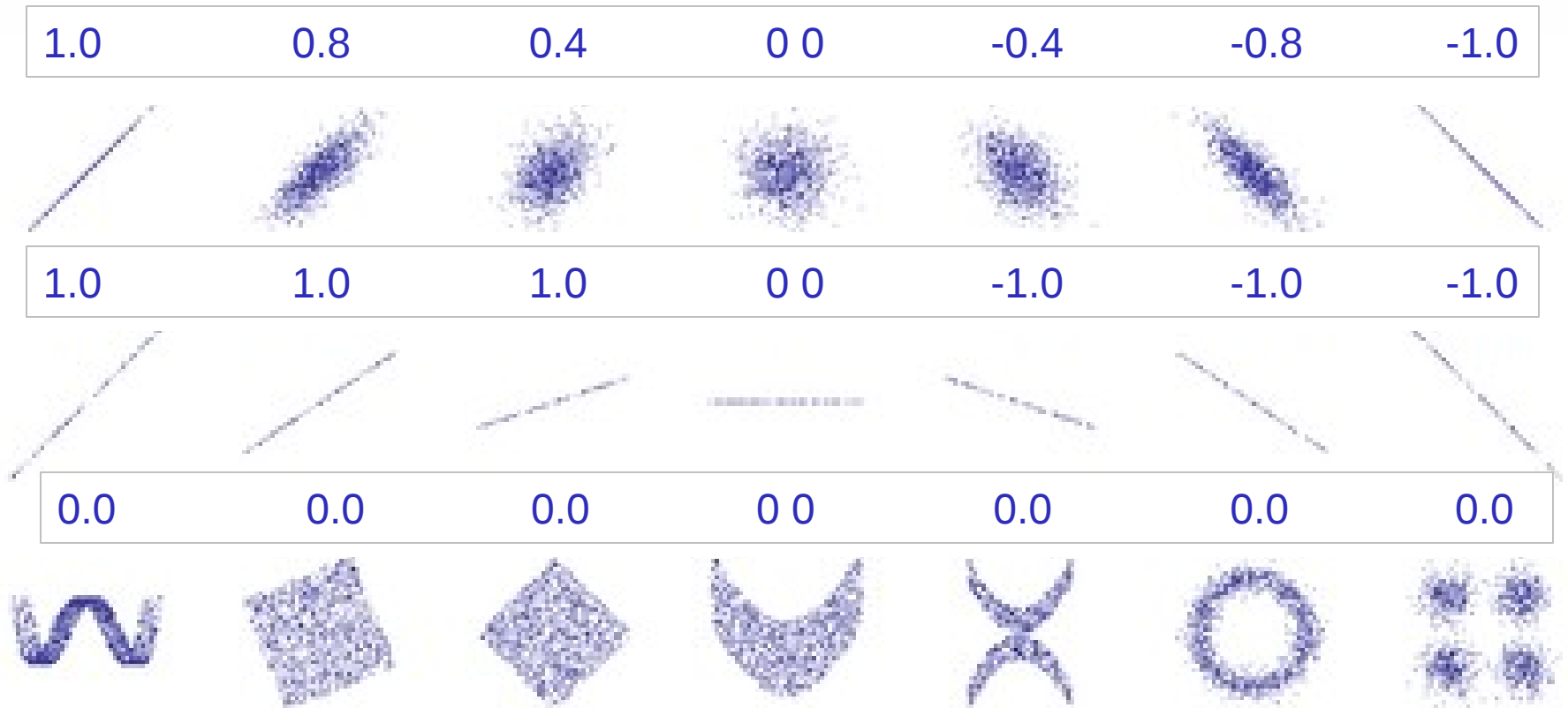
*Are you really sure ? About 95%*

# Breaches

---

## Correlation Metrics

# Correlation in Pictures



Ref: <http://en.wikipedia.org/wiki/Correlation>



# Correlations: Type & Datalost

---



- High positive: ACC and FIN
- Positive: NAA, SSN, MISC

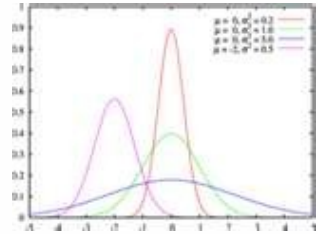
---

# Vulnerabilities

# Vuln's: Questions, Metrics, Models

- How do NVD and OSVDB compare?

- Metrics: Record counts, distributions
- Models: Distribution parameters



- Are vulnerabilities decreasing in frequency?

- Metrics: VulnCount
- Models: Linear regression, model checking

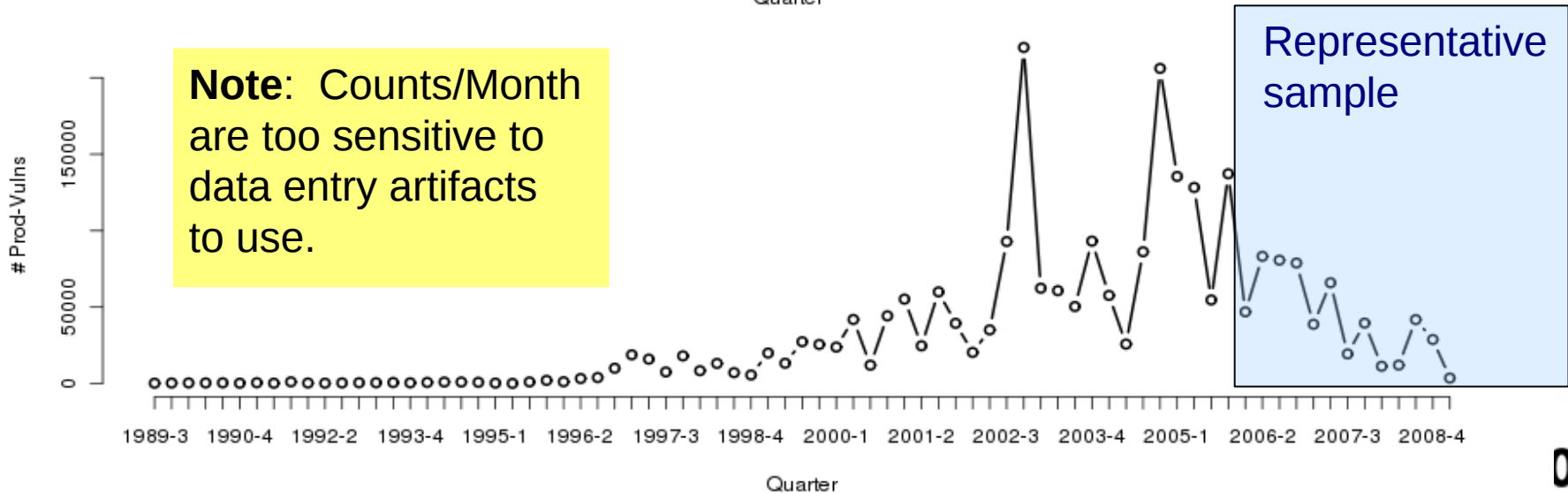
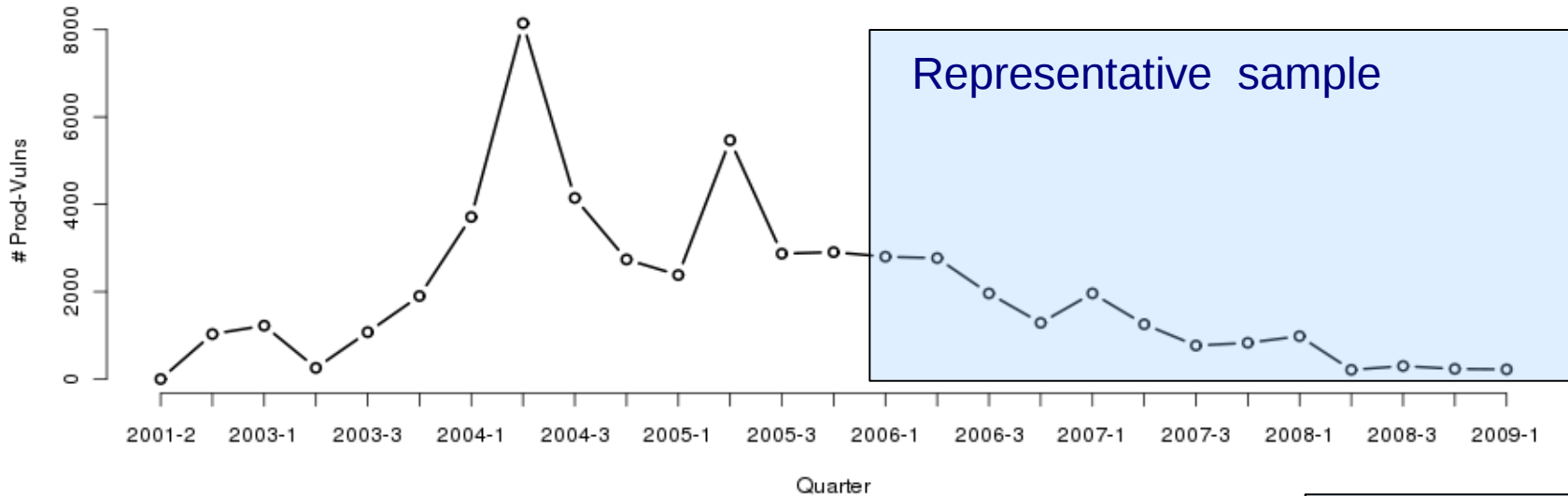


- Vulnerability Index Metrics

- Vendor Vulnerability Gradient (Linear regression)
- Workload Index from NIST (Weighted sum)

# OSVDB and NVD: Whole Magilla's

OSVDB: Prod-Vuln Count per Quarter  
All Data



# Vuln's: Data Sets Used

<b>Feature</b>	<b>raw: OSVDB/NVD</b>	<b>vuln: NVD</b>
Records included	Join 6 / 3 tables:	vuln
# Rows	53444/ 2530319	162905
# Rows	year > 2005: 15592/551558	Year > 2005 ????
% Rows Omitted	30% / 22%	??%
# Columns	8	16 → 6
Added columns	Year, Quarter, Month	cvssScore: 0-3.99: LOW 4-6.99: MED 7-10: HIGH

# Vulnerable Products

---

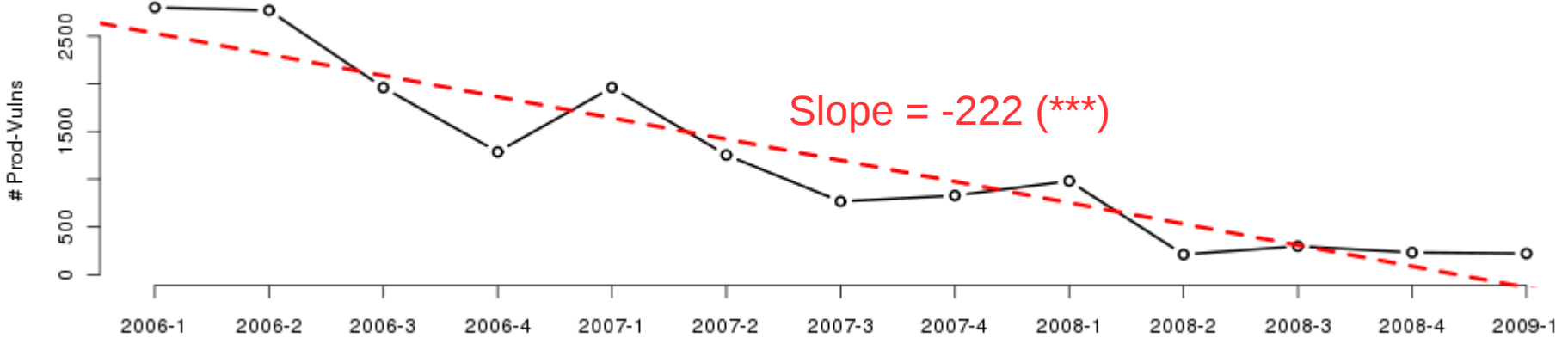
Count and Percent Metrics

# NVD and OSVDB: Comparison

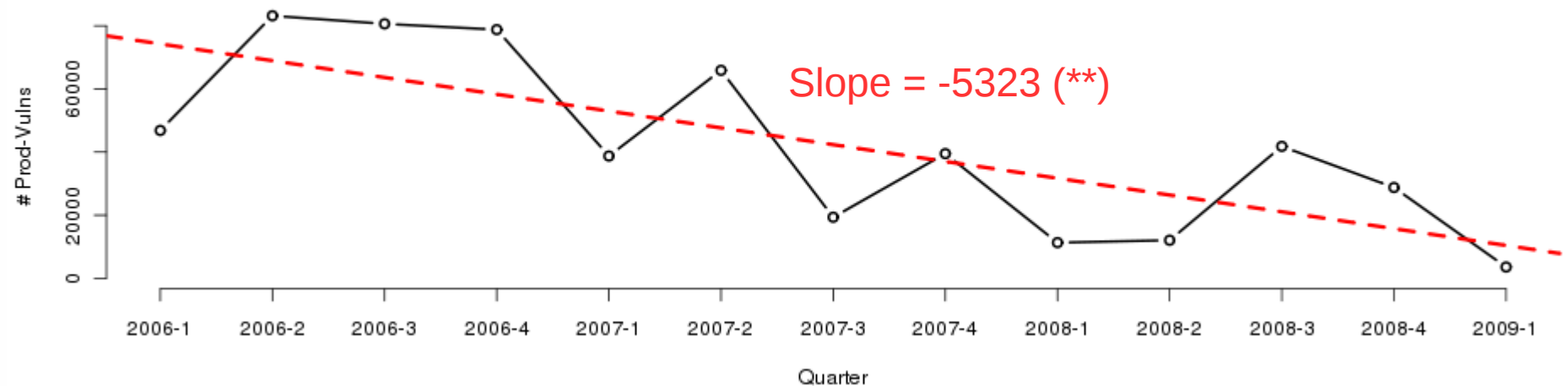
Metric	NVD	OSVDB	NVD / OSVDB
# Vulnerabilities	163846	52396	3.1
# Vendors	264	3880	0.06
# Products	3423	6728	0.54
# Versions	14804	22563	0.63
# VenProdVer-Vulns	2530319	53444	47.2
# Vuln's per VenProdVer	72/828 *	1 / 2.4 *	72
# VenProdVer's per Vuln	64/297 *	1 / 3.2 *	64
# ProdVer's per Vendor	543/14970 *	2 / 13.8 *	231
# Ver's per VendorProd	72/1888 *	2 / 8 *	36
# Prod's per Vendor	2/8 *	1 / 1.7	2
% Vulns with Types	1%	31%	.03
# VulnProd (Year > 2005)	551558	15592	35.4
% VulnProd (Year > 2005)	22%	30%	.73

# Prod-Vulns/Quarter Trend

OSVDB: Prod-Vuln Count per Quarter  
Jan 2006 to Present



NVD: Prod-Vuln Count per Quarter  
Jan 2006 to Present



# OSVDB and NVD:

---



- NVD version granularity > 40x OSVDB
- Agree: VulnCount/Qtr is decreasing

# Vulnerable Products

---

Two Indecies:

1. Vendor Vulnerability Volume Gradient
2. Workload Factor

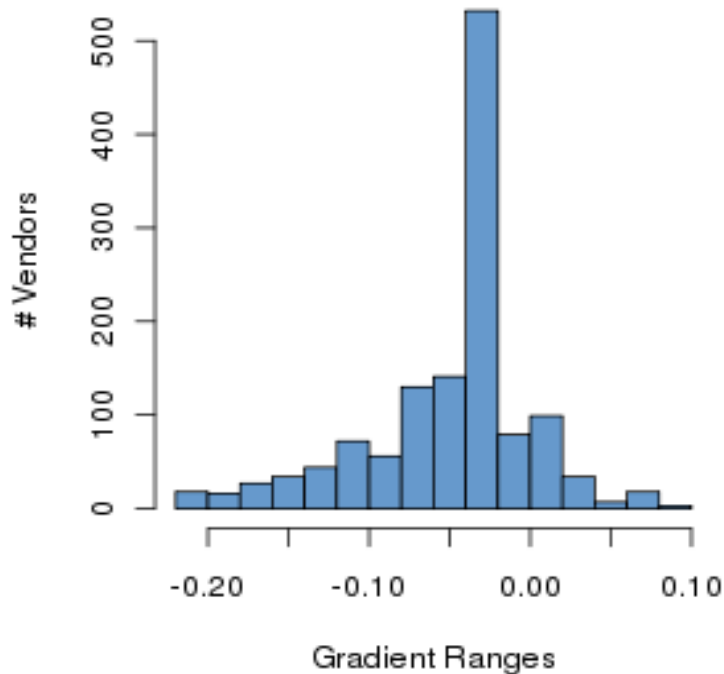
# Vendor Vulnerability Volume Gradient

---

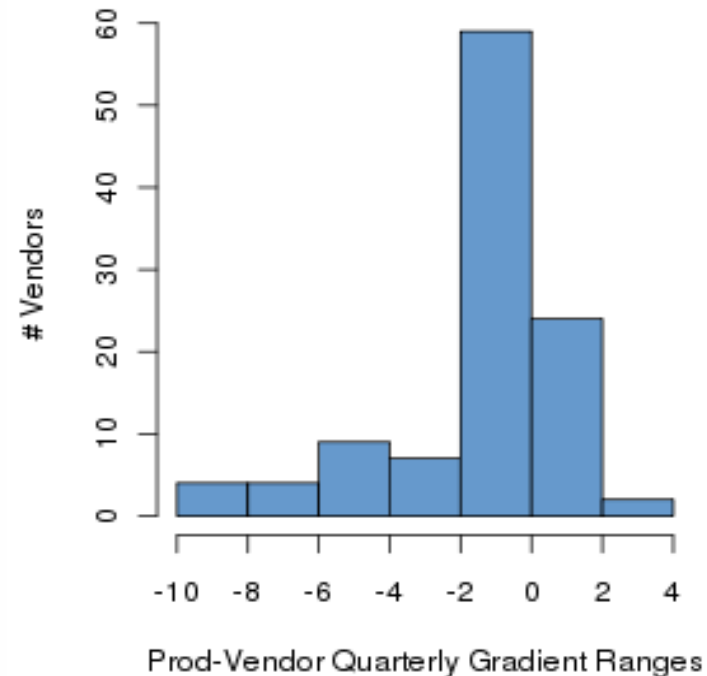
- Metric: Slope of Vendor VulnCount Trend
  - { ( Qtr, VulnCount) : Year > 2005, Vendor = X }
  - Best Fit Line using Least Squares Model
- Negative gradient implies improvement
- General indicator of reduction in vulnerabilities across all products
- Problems
  - Non-linearity, timing of lumps
  - No distinction for severity
  - Bias against commercial adoption

# VVVG Distribution (Without Outliers)

OSVDB Gradient Histogram  
From 1Jan06 with Outliers Omitted



NVD Gradient Histogram  
From 1Jan06 with Outliers Omitted



Both OSVDB and NVD agree:  
Slight skew to to the positive  
Median is negative

# OSVDB and NVD: VVVG Values

OSVDB Prod-Vuln Gradient: Most Improved Vendors  
Quarterly from Jan 2006

gradient	intercept	vendorName	vulnCount
-22.04	274.23	Oracle	1559
-13.19	249.54	Apple	2044
-12	188.38	Microsoft	1357
-8.74	100.27	Cisco	508
-5.23	76.31	Mozilla	516
-4.49	44.5	IBM	170
-4.23	40.35	Hitachi	140
-3.85	37.69	Pearlanger.com	140
-3.6	44.77	Linux	254
-3.32	31.42	FreeBSD	106

NVD Prod-Vuln Gradient: Most Improved Vendors  
Quarterly from Jan 2006

gradient	intercept	vendorName	vulnCount
-2088.39	28587.5	linux	181594
-520.47	6668.04	php	39322
-402.96	5574.42	apple	35798
-336.73	4767.12	microsoft	31330
-321.86	3613.77	mysql	17690
-312.97	6276.23	mozilla	53111
-246.86	7190.23	cisco	71009
-208.79	2631.46	ibm	15209
-153.77	1737.23	bea	8591
-147.93	1914.65	oracle	11429

On both lists: 7 of 10:

apple, cisco, ibm, linux,  
oracle, mozilla, microsoft

Open Source: OSVDB has 3 of 10  
NVD has 4 of 10

# Vulnerability Gradient Index:

---



- OSVDB and NVD agree:

Top Rated are

Apple, Cisco, IBM, Linux,  
Mozilla, MSFT, Oracle

- Index weaknesses:

Bias against commercial significance

Sensitivity to vuln count burstiness

# Repeatability: Workload Index

The screenshot shows the NVD website interface. At the top, it says "National Vulnerability Database Home - Mozilla Firefox" and the URL is "http://nvd.nist.gov/home.cfm?workloadindex". The page features the NIST logo and the text "Sponsored by DHS National Cyber Security Division/US-CERT". The main heading is "National Vulnerability Database" with the tagline "automating vulnerability management, security measurement, and compliance checking". Below this is a navigation menu with categories like "Vulnerabilities", "Checklists", "Product Dictionary", "Impact Metrics", "Data Feeds", and "Statistics". The "Workload Index Information" section is highlighted, and a text box explains the calculation of the workload index.

The NVD workload index is calculated using the following equation:

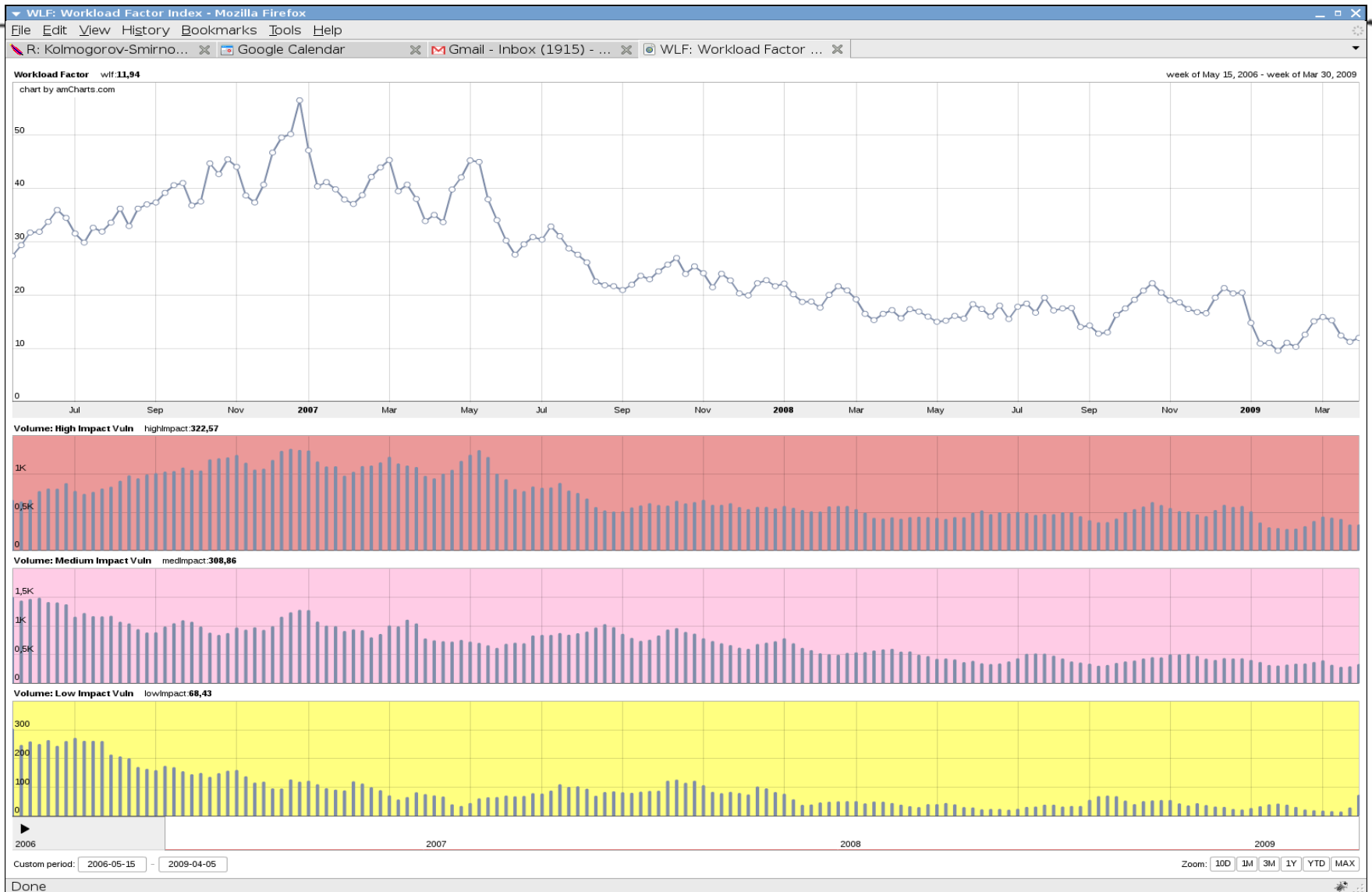
$$\left( (\text{number of high severity vulnerabilities published within the last 30 days}) + (\text{number of medium severity vulnerabilities published within the last 30 days}/5) + (\text{number of low severity vulnerabilities published within the last 30 days}/20) \right) / 30$$

The index equation counts five medium severity vulnerabilities as being equal in weight with 1 high severity vulnerability. It also counts 20 low severity vulnerabilities as being equal in weight with 1 high severity vulnerability.

[Return to the NVD Home Page](#)

35649 [CVE Vulnerabilities](#)  
142 [Checklists](#)  
169 [US-CERT Alerts](#)  
2310 [US-CERT Vuln Notes](#)  
2097 [OVAL Queries](#)

# NVD Workload Factor



# Workload Factor:

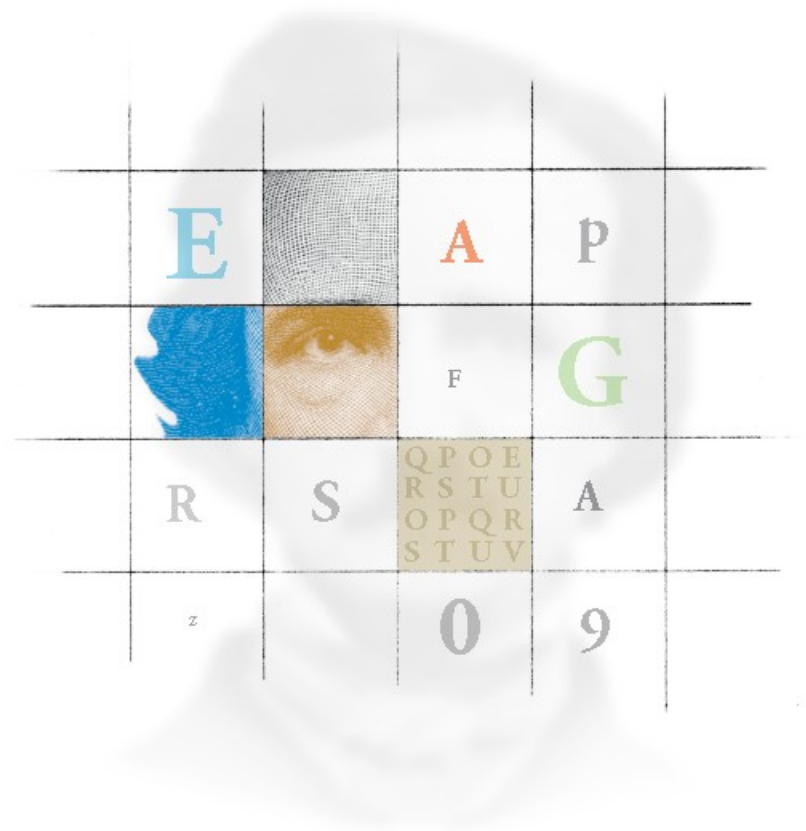
---



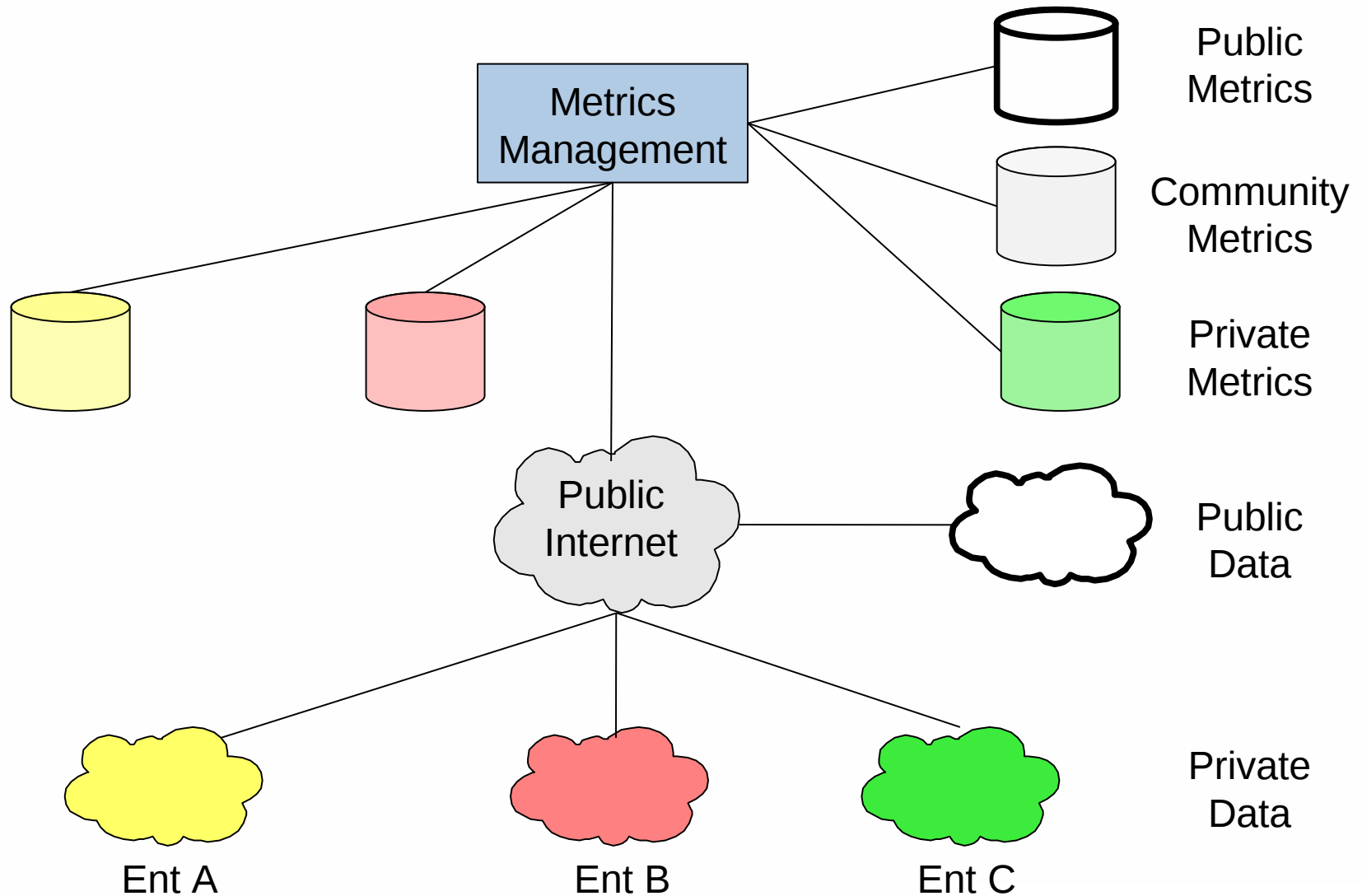
- Is decreasing
- Is consistent with VulnCount/Qtr ↓
- Enables testing of the following hypothesis:

WLF is independent of day-of-week

# Automation



# Metric Spaces

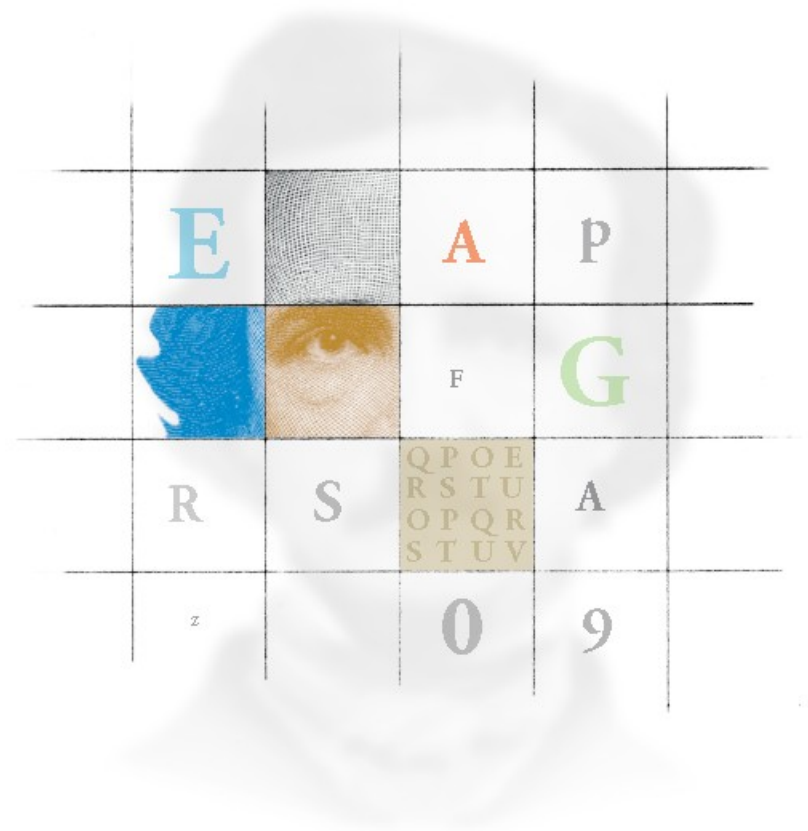


# Metric Lifecycle

---

- Design
  - Define the metric: objectives, contexts, limitations
  - Specify data sources and compute logic
  - Identify visualization options
- Derive
  - Ingest data: Interface to source, cleans, normalize
  - Compute on a regular schedule. Apply models.
  - Store results
- Deliver
  - Manage content: automatic and human-generated
  - Dynamic and interactive media

# Lessons Learned



# Metrics are Science. Decisions are Art.

---

- You can lie with statistics
- Extremes, as in life, are bad form
  - Taleb: “Anything that relies on correlation is charlatanism.”



Credit: TK Keanini, CTO, nCircle

# Crunching Metrics from Public Sources

Educate + Learn = Apply

*Publicly available data can lend insight to the global state of security.*

- ❖ *BreachCount/Month growth is NOT statistically significant*
- ❖ *Outside breaches have higher impact than Inside breaches*
- ❖ *OSVDB and NVD are different but agree*
- ❖ *Vulnerabilities discovered / quarter is decreasing + 7 of their respective top 10 vendors*

- ❖ *There are public sources of security data available*
- ❖ *Metrics add value*
- ❖ *Models can be applied but they have to be checked for a good fit*
- ❖ *Models can give levels of confidence in the results they predict*
- ❖ *“Good Enough” has value*

- ❖ *Track public metrics*
- ❖ *Apply insights internally*
- ❖ *Apply techniques internally*
- ❖ *Drive metrics to help make better decisions*

# Resources

---

- <http://www.metricscenter.org>
- <https://www.metricscenter.net>
- <http://datalossdb.org>
- <http://osvdb.org>
- <http://nvd.nist.gov>
- <http://www.first.org/cvss/cvss-guide.htm>
- <http://www.cisecurity.org>

---

Elizabeth A. Nichols, Ph.D

CTO for Metrics

PlexLogic, LLC

[betsy.nichols@plexlogic.com](mailto:betsy.nichols@plexlogic.com)

703.963.7202