RSA CONFERENCE 2014
FEBRUARY 24 – 28 | MOSCONE CENTER | SAN FRANCISCO

Share.
Learn.
Secure.
Capitalizing on
Collective Intelligence

Big Lessons in Small Data

SESSION ID: DSP-T09

Wade Baker
RISK Team
Verizon
@wadebaker

Jay Jacobs
RISK Team
Verizon
@jayjacobs

Overview

- Part 1 : Brief History of Data Analysis
- Part 2 : Current state of Infosec
- Part 3 : Putting data analysis into practice

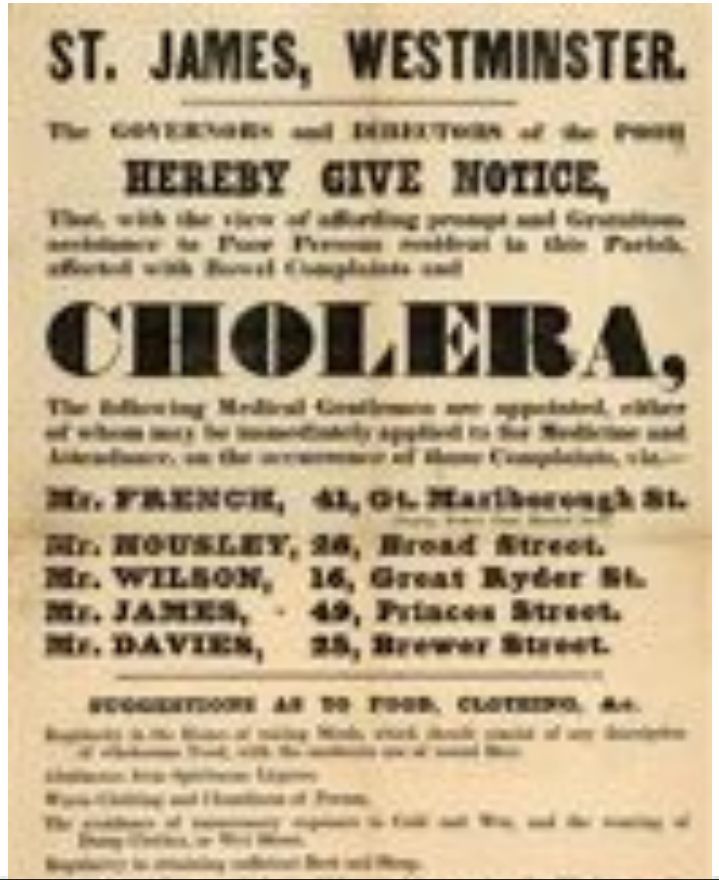
Presenter's Company Logo – replace on master slide

Part 1: Brief History of Data Analysis

Presenter's Company Logo – replace on master slide

Early Data Analysis: Epidemiology

William Farr
1807-1883



ST. JAMES, WESTMINSTER.

The GOVERNORS and DIRECTORS of the POOR

HEREBY GIVE NOTICE,

That, with the view of affording prompt and Gratuitous assistance to Poor Persons resident in this Parish, affected with Bowel Complaints and

CHOLERA,

The following Medical Gentlemen are appointed, either of whom may be immediately applied to for Medicine and Attendance, on the occurrence of those Complaints, viz.—

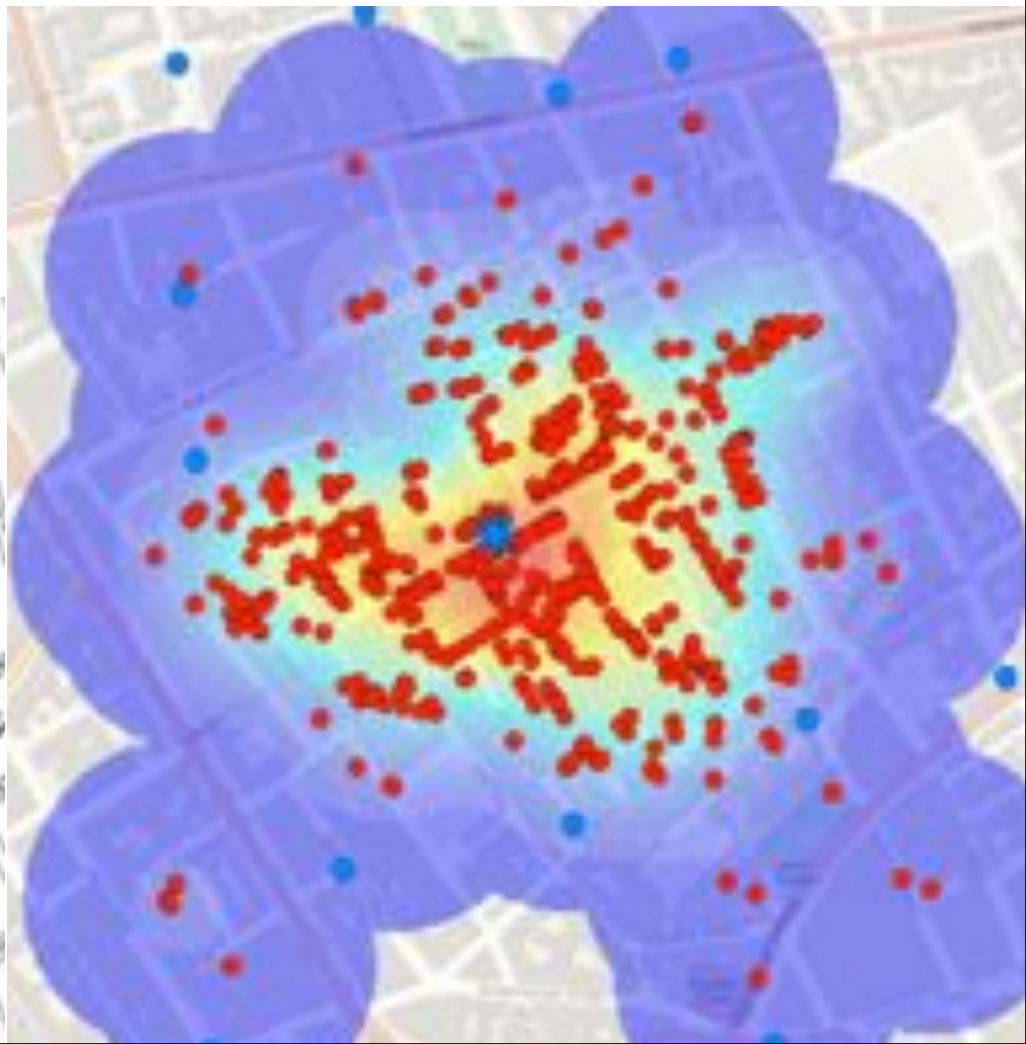
Mr. FRENCH, 41, Gt. Marlborough St.
Mr. HOUSLEY, 26, Broad Street.
Mr. WILSON, 16, Great Ryder St.
Mr. JAMES, 49, Princes Street.
Mr. DAVIES, 25, Brewer Street.

SUGGESTIONS AS TO FOOD, CLOTHING, &c.

John Snow
1813-1858

John Snow:

Cholera is spread by consuming water contaminated by a "special animal poison"



William Farr:

"The elevation of the soil in London has a more constant relation with mortality from cholera than any other known element."

Table 2 Deaths from cholera in London, registered in 1849 by registration district together with eight possible explanatory variables.

District	Deaths from cholera in 1849 per 10,000 inhabitants	Elevation above high water (feet)	Annual deaths from all causes 1838-1844 per 10,000 inhabitants	Persons per acre	Persons per inhabited house	Average annual value of house (£)	Annual value of house per person (£)	Poor rate precept per pound of house value	Water supply*
Newington	144	-1	232	101	5.8	22	3.788	0.075	1
Rotherhithe	205	0	277	19	5.8	23	4.238	0.143	1
Bermondsey	161	0	264	66	6.2	18	3.077	0.134	1
St George Southwark	164	0	267	181	7.0	22	3.318	0.089	1
St Olave	181	1	281	114	7.9	35	4.559	0.079	1
St Saviour	153	2	292	141	7.1	36	5.291	0.079	1
Westminster	68	2	260	70	8.8	36	4.189	0.075	1
Lambeth	120	3	233	34	6.5	28	4.389	0.039	1
Camberwell	97	4	197	12	5.8	25	4.508	0.072	1
Greenwich	75	8	238	18	6.8	22	3.379	0.038	2
Poplar	71	10	241	15	6.2	44	7.360	0.081	2
Chelsea	46	12	287	62	7.1	29	4.210	0.060	1
Hammersmith,	33	12	238	11	6.6	33	5.070	0.067	3

Quote of William Farr as it appears in:
P. Bingham, N.Q. Verlander, M.J. Cheal
John Snow, William Farr and the 1849 outbreak of cholera that affected London: a reworking of the data highlights the importance of the water supply

Bingham, et al. (2004):

"Had logistic regression been available to Farr, its application to his 1852 data set would have changed his conclusion."

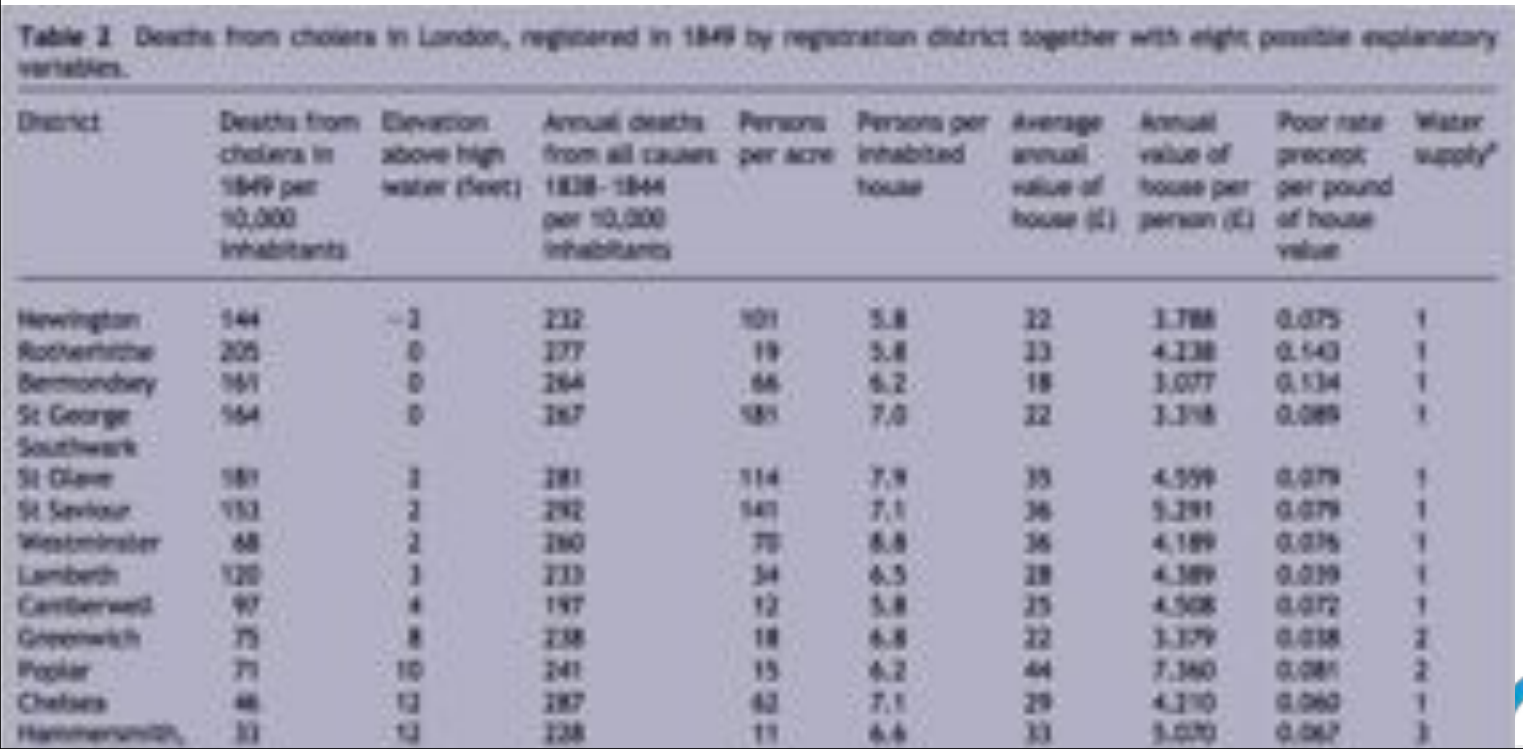


Table 2 Deaths from cholera in London, registered in 1849 by registration district together with eight possible explanatory variables.

District	Deaths from cholera in 1849 per 10,000 inhabitants	Elevation above high water (feet)	Annual deaths from all causes 1838-1844 per 10,000 inhabitants	Persons per acre	Persons per inhabited house	Average annual value of house (£)	Annual value of house per person (£)	Poor rate precept per pound of house value	Water supply*
Newington	144	-3	232	101	5.8	22	3.788	0.075	1
Rotherhithe	205	0	277	19	5.8	23	4.238	0.143	1
Bermondsey	161	0	264	66	6.2	18	3.077	0.134	1
St George Southwark	164	0	267	181	7.0	22	3.318	0.089	1
St Olave	181	3	281	114	7.9	35	4.559	0.079	1
St Saviour	153	2	292	141	7.1	36	5.291	0.079	1
Westminster	68	2	260	70	8.8	36	4.189	0.076	1
Lambeth	120	3	233	34	6.5	28	4.389	0.039	1
Camberwell	97	4	197	12	5.8	25	4.508	0.072	1
Greenwich	75	8	238	18	6.8	22	3.379	0.038	2
Poplar	71	10	241	15	6.2	44	7.360	0.081	2
Chelsea	46	12	287	62	7.1	29	4.210	0.060	1
Hammersmith,	33	12	238	11	6.6	33	5.070	0.067	3

Quote of William Farr as it appears in:
P. Bingham, N.Q. Verlander, M.J. Cheal
John Snow, William Farr and the 1849 outbreak of cholera that affected London: a reworking of the data highlights the importance of the water supply

“Modern” statistical inference: Design of experiments

R. A. Fisher
1890-1962

“After two or three months of investigation it will be found possible to understand some of Fisher’s sentences.”

Sir Fred Hoyle, Astronomer

And then the transistor...

John Tukey
1915-2000

and **Exploratory Data Analysis**

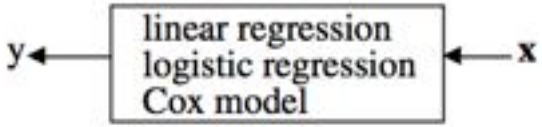
"Numerical quantities focus on expected values, graphical summaries on unexpected values."

Statistical Science
2001, Vol. 16, No. 3, 199–231

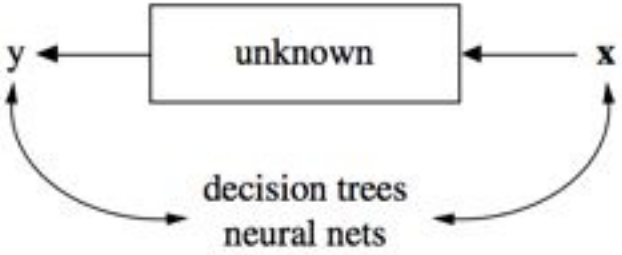
Statistical Modeling: The Two Cultures

Leo Breiman
1928-2005

Data Modeling Culture



Algorithmic Modeling Culture



Part 2: Current State of Infosec

(Hint: 1800's statistics now with computers!)

Our Goal:

- How is infosec doing with our data (small or big)?

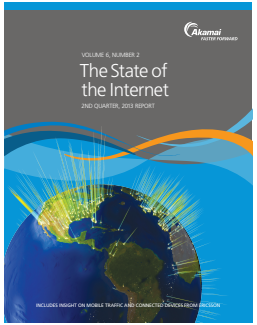
Our Approach:

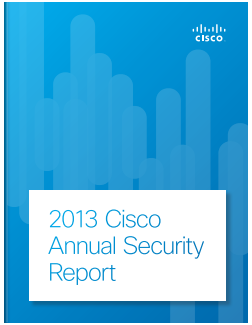
- Analyze data analysis techniques in (industry) publications

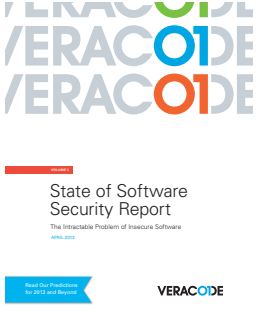
Our Conclusion:

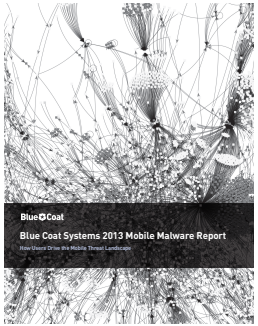
- We have some “opportunities for improvement”

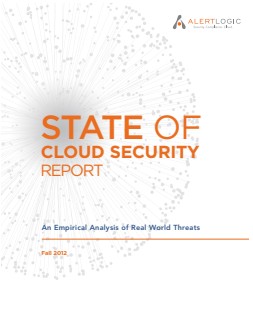
Presenter’s Company Logo – replace on master slide

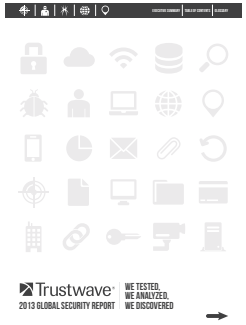


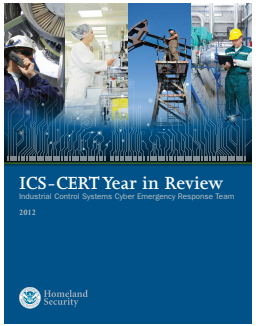


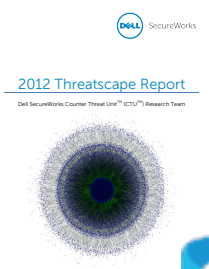




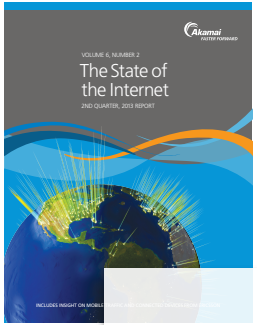


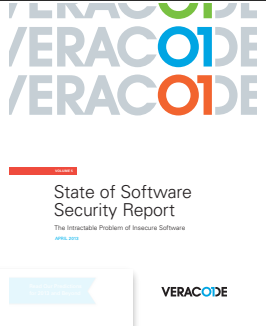


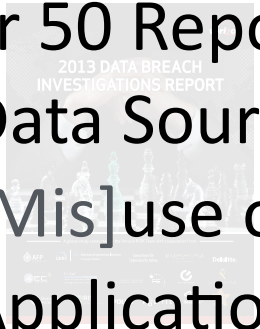


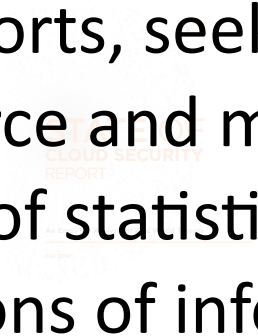


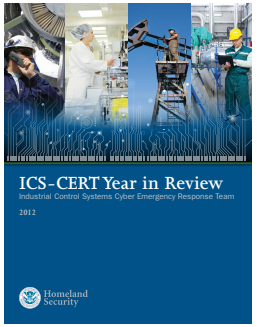
Presenter's Company Logo – replace on master slide







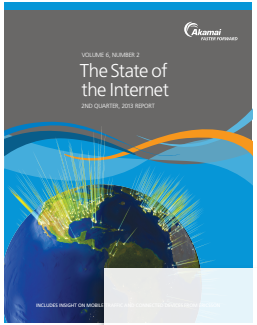


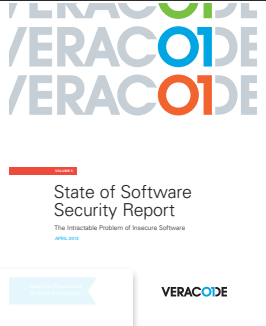


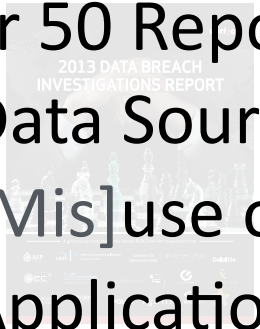
Over 50 Reports, seeking:

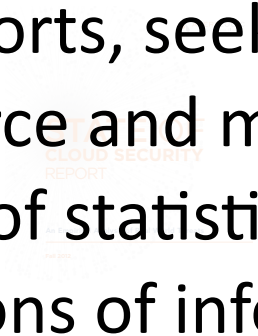
- Data Source and methodology
- [Mis]use of statistics
- Applications of inference

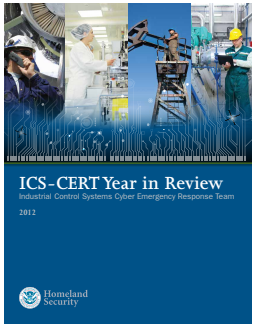
Presenter's Company Logo – replace on master slide











Over 50 Reports, seeking:

- Data Source and methodology
- [Mis]use of statistics
- Applications of inference

Presenter's Company Logo – replace on master slide

Results

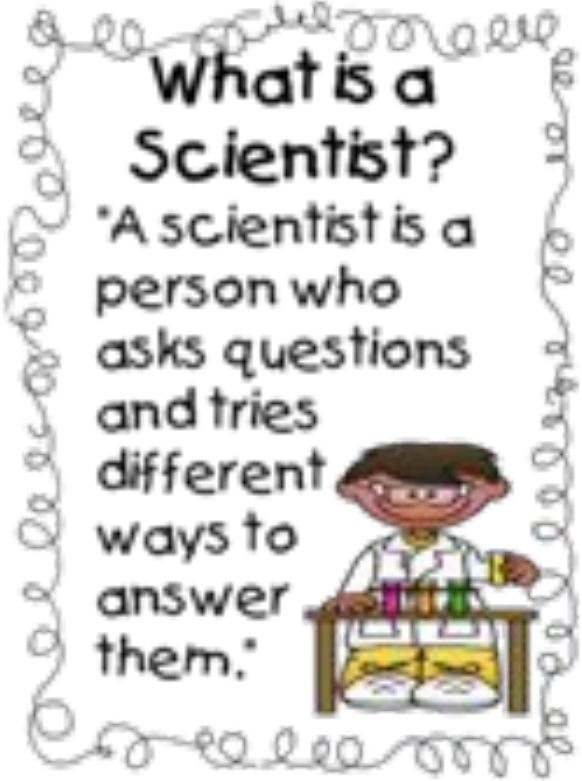
Note: to RSA reviewers, we are still finalizing the numbers here

- mm Surveys, mm Device Data, mm Reports
- mm of 52 defined their methodology
- mm defined their **sample** of which mm defined their **population**
- **None** applied a probabilistic sampling method*
- Heavy use of “count and compare”
- Conclusion: *We are in the 1800’s for data analysis*

* Not a bad thing

Opportunities to Learn…

- More science! (less FUD)
- More statistics!
- More transparency!



Presenter's Company Logo – replace on master slide

Opportunities to Learn 1: More Science!

"My job was to find questions about baseball that have objective answers, that's all that I do, that's all that I've done."

-- Bill James, Sabermetrician

Presenter's Company Logo – replace on master slide

Scientific method

1. Research question
2. Design/plan experiment
3. Gather data
4. Make sense of data
5. Communicate results

http://scienceandstory.files.wordpress.com/2012/04/science.jpg

Presenter's Company Logo – replace on master slide

Opportunities to Learn 2: More Statistics!

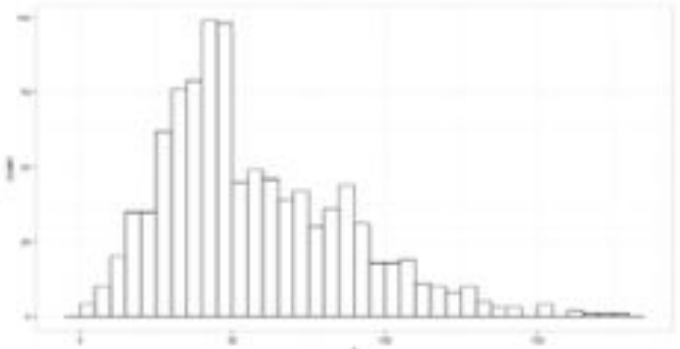
- Exploratory Data Analysis
- Sample Error vs. Sample Bias
- (let's argue about p-values later)

Exploratory Data Analysis (get to know your data)

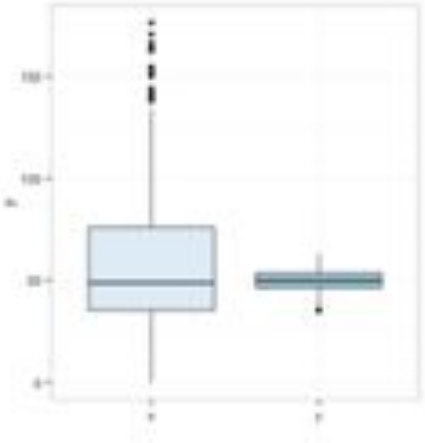
“Five Number” summary:

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.10   35.42   48.90   57.07   75.97  176.40
```

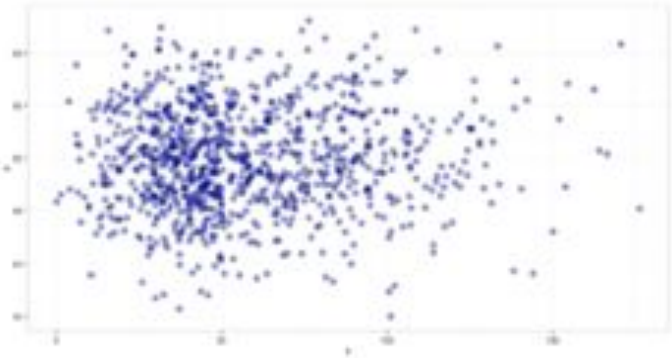
Histogram



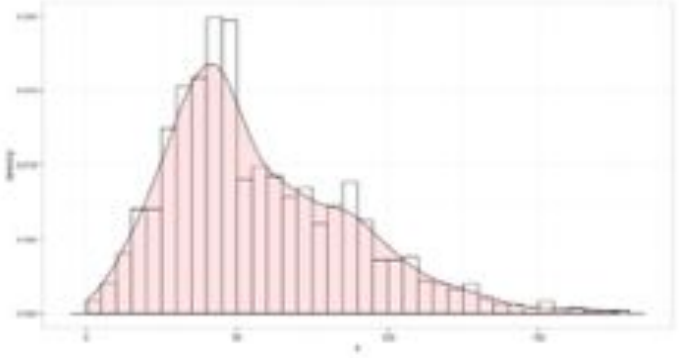
Box plot



Scatter plot



Density plot

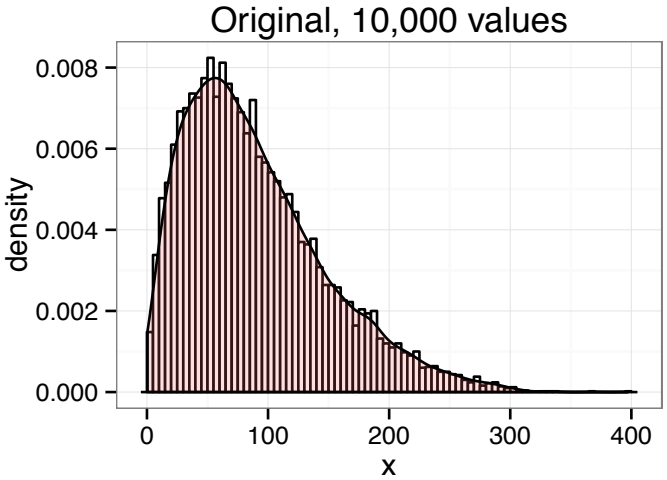


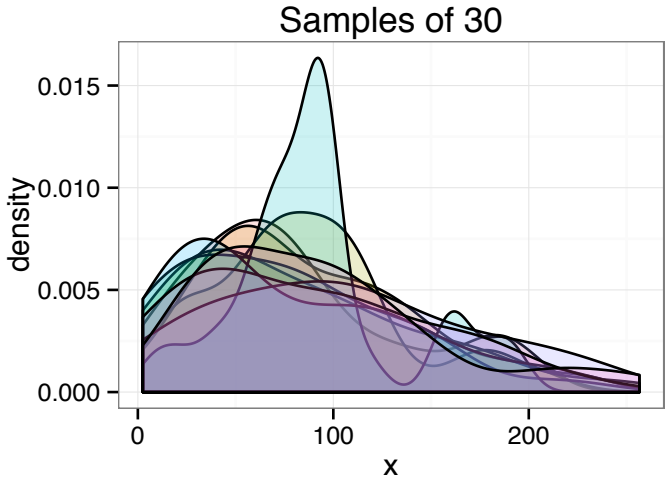
Presenter’s Company Logo – replace on master slide

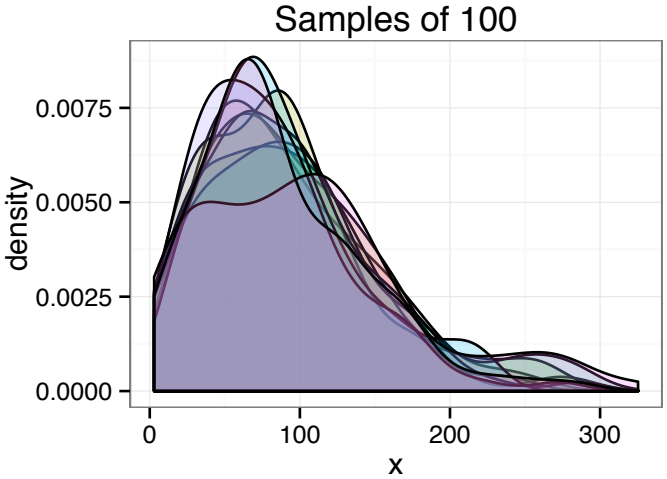
Sample Error

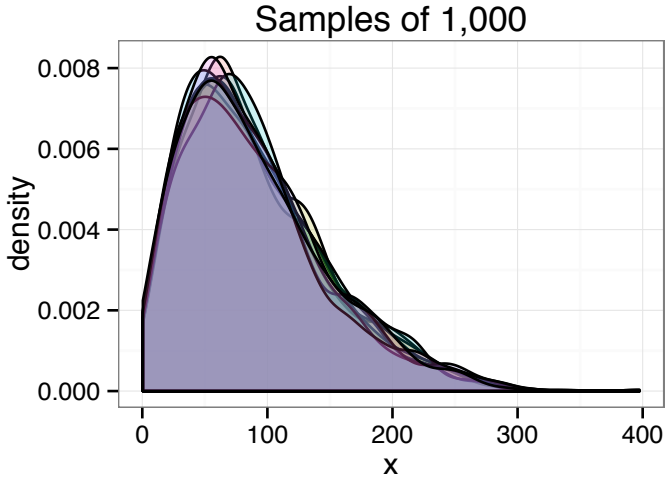
("error" does not mean mistake)

Increasing sample size reduces sample error.

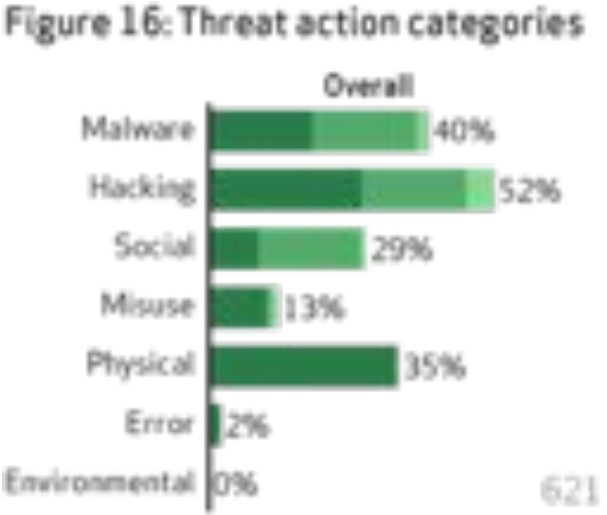




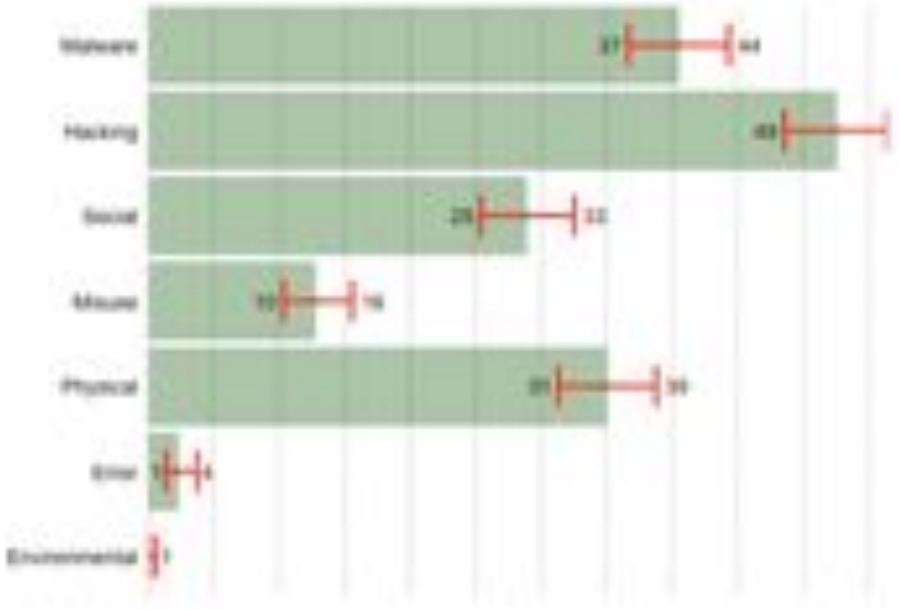




Sample Error



2013 Verizon DBIR

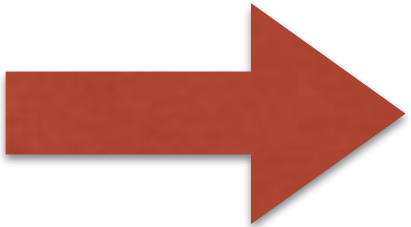


95% confidence intervals

Presenter's Company Logo – replace on master slide

Sample Bias

telephone directories +
lists of magazine subscribers +
rosters of clubs and associations
=
2.4 Million people polled and
Landon - 57%
Roosevelt - 43%



http://www.math.upenn.edu/~deturck/m170/wk4/lecture/case1.html

Presenter's Company Logo – replace on master slide

Part 3: Putting Data Analysis into Practice

[Examples of big lessons in small data]

Presenter's Company Logo – replace on master slide

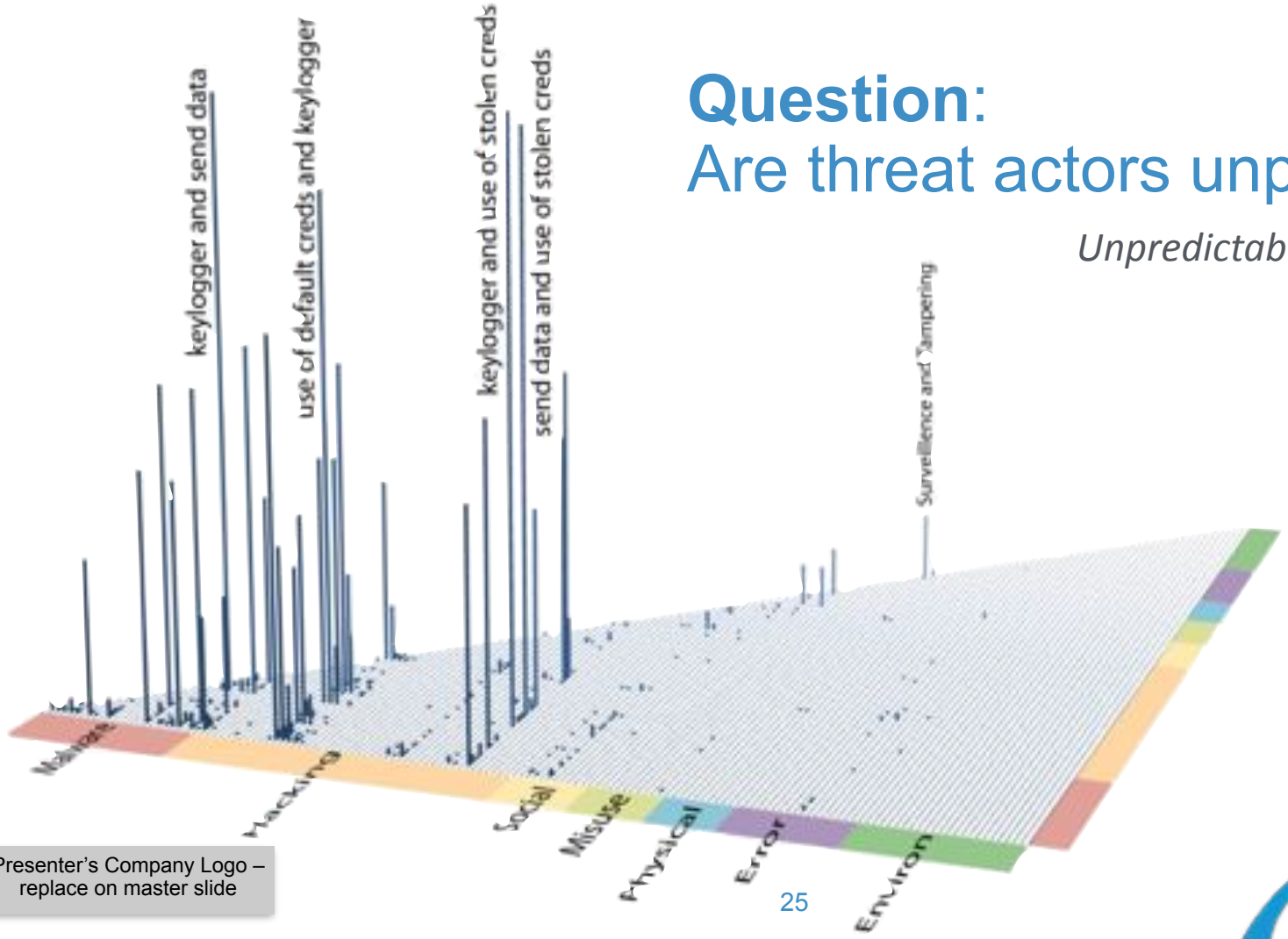
Research Question:
Are threat actors unpredictable?

Unpredictable == Random

Presenter's Company Logo – replace on master slide

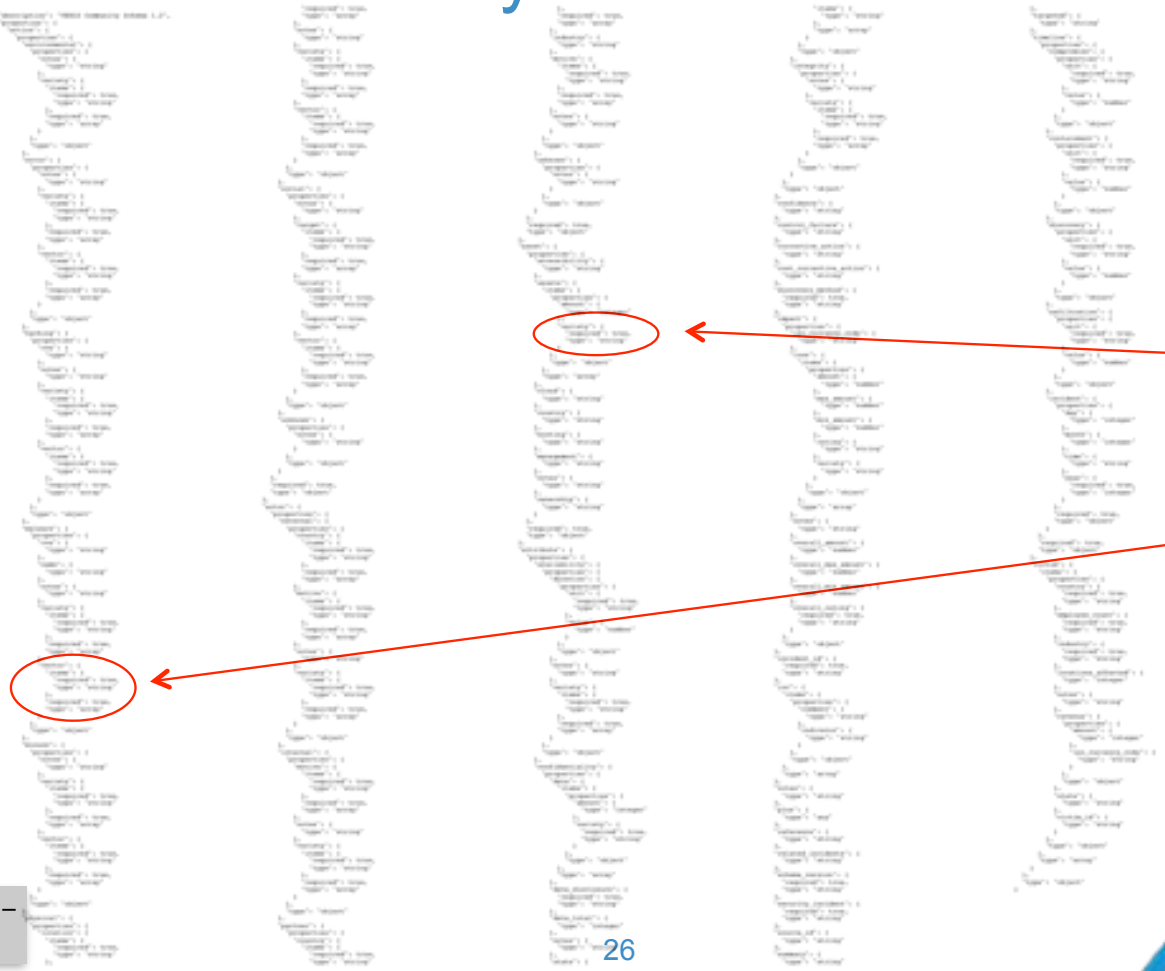
Question:
Are threat actors unpredictable?

Unpredictable == Random

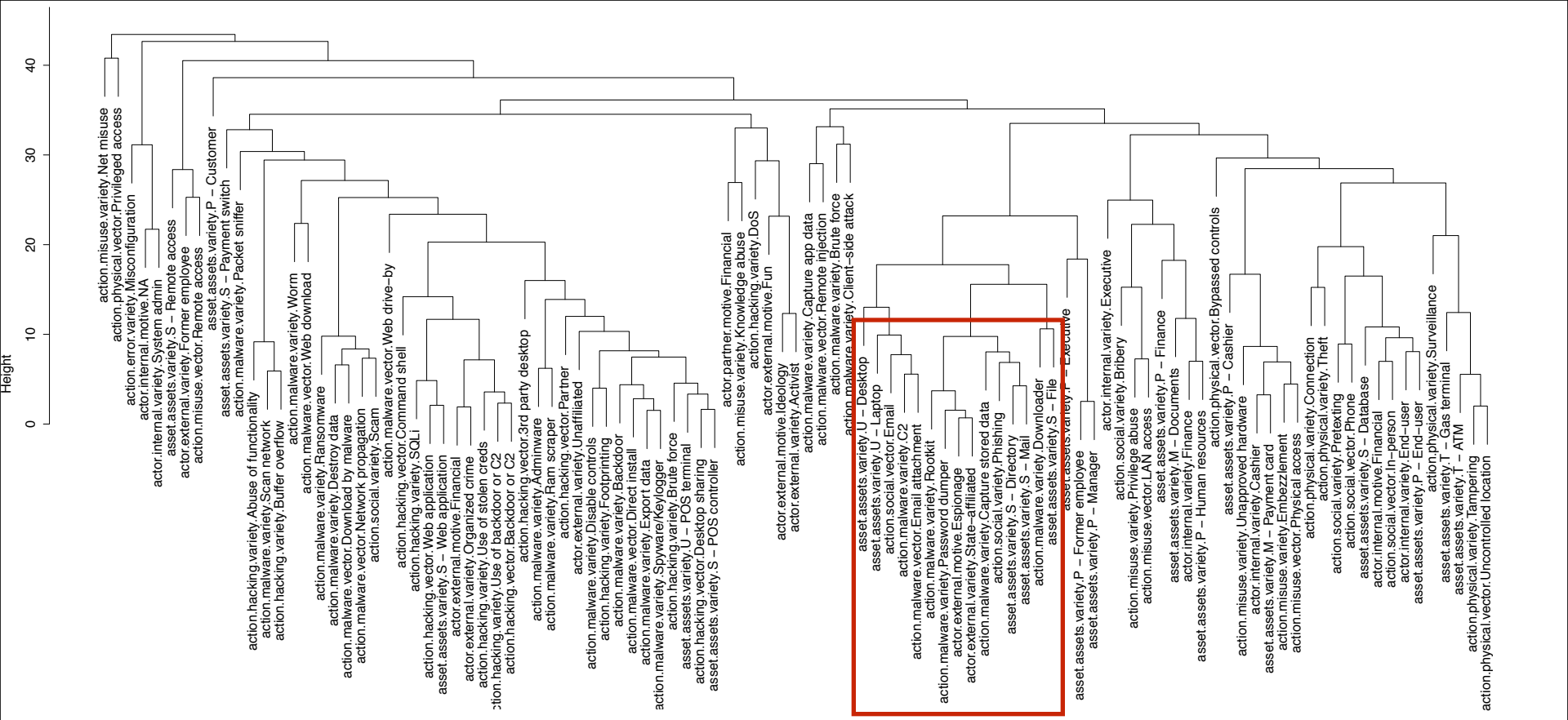


Presenter's Company Logo – replace on master slide

The “DNA” of a Security Incident



Presenter’s Company Logo – replace on master slide

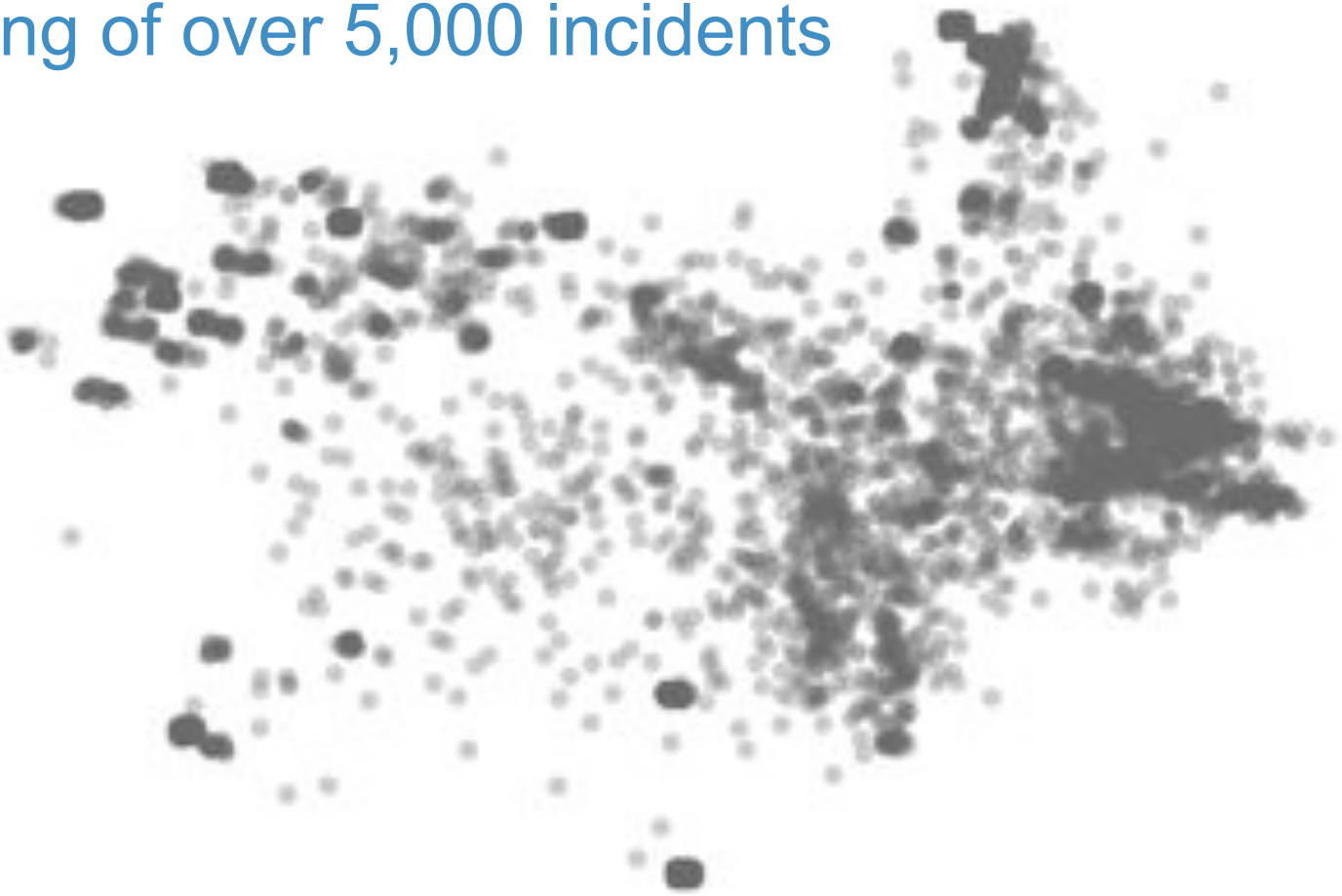


Key variables in “Espionage” pattern

(actor.external.state-affiliated, malware.variety.rootkit, malware.variety.pwd dumper, social.variety.phishing, asset.variety.S – directory, etc)

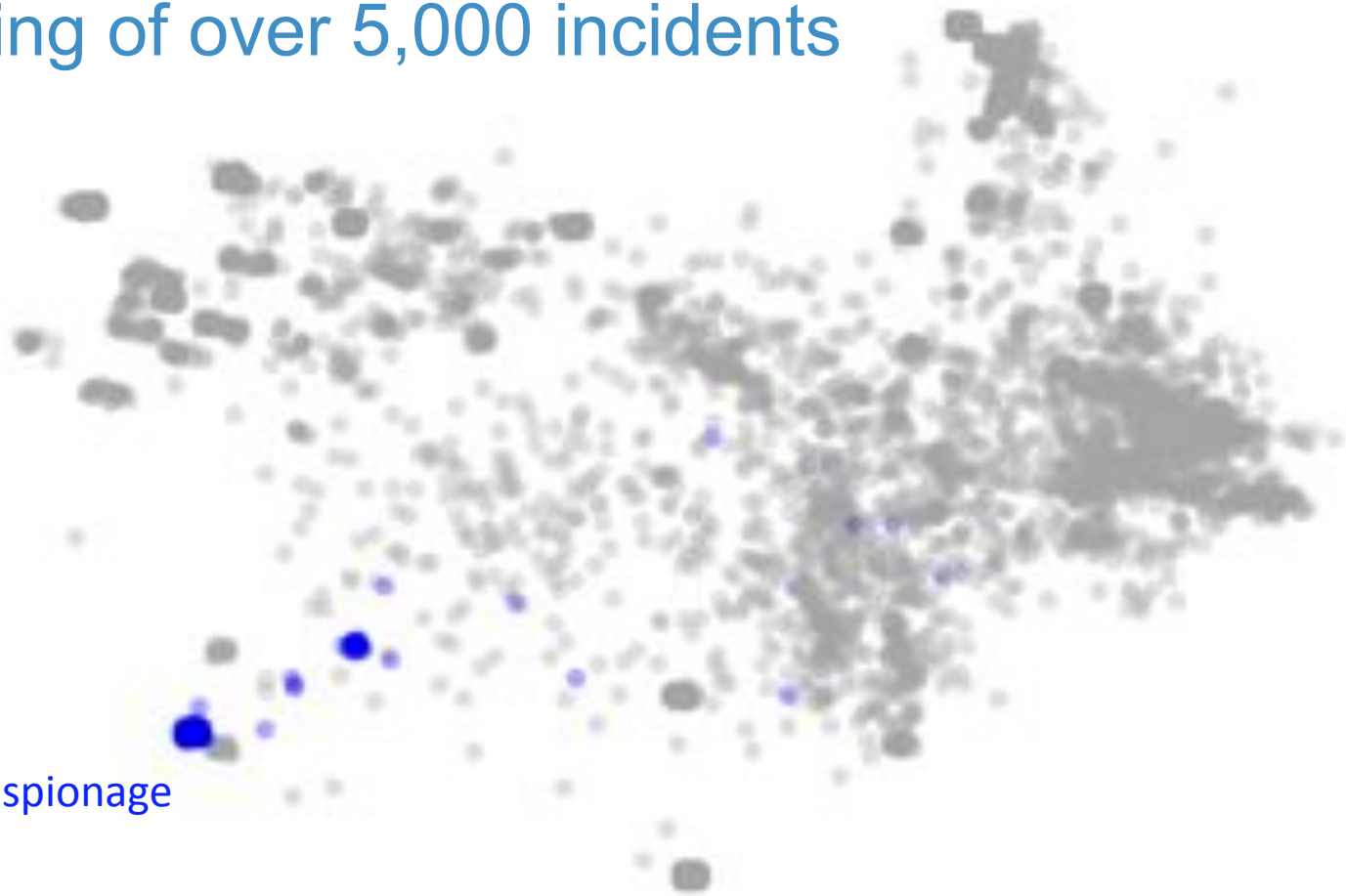
Presenter’s Company Logo – replace on master slide

Clustering of over 5,000 incidents



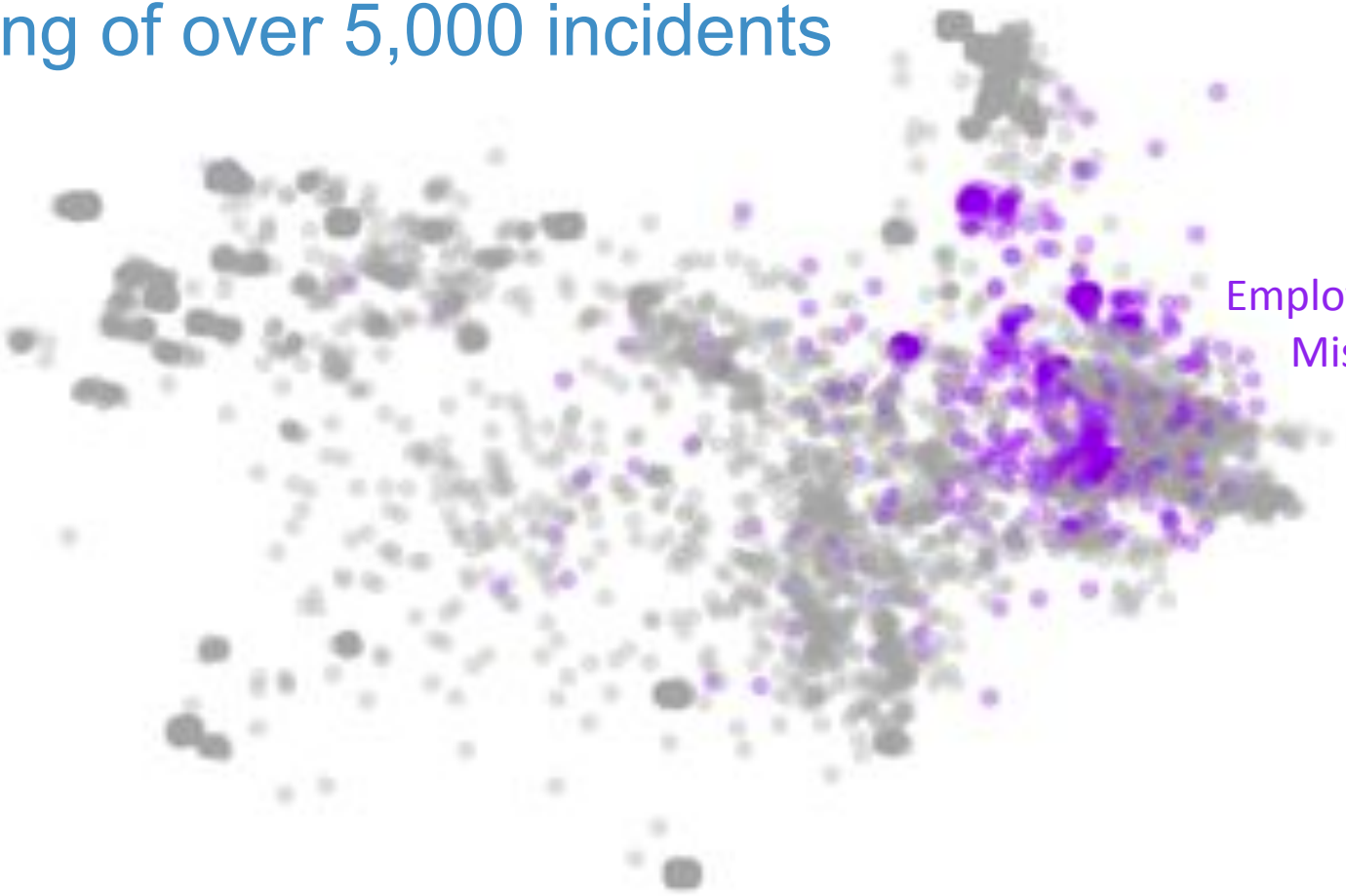
Presenter's
replace

Clustering of over 5,000 incidents



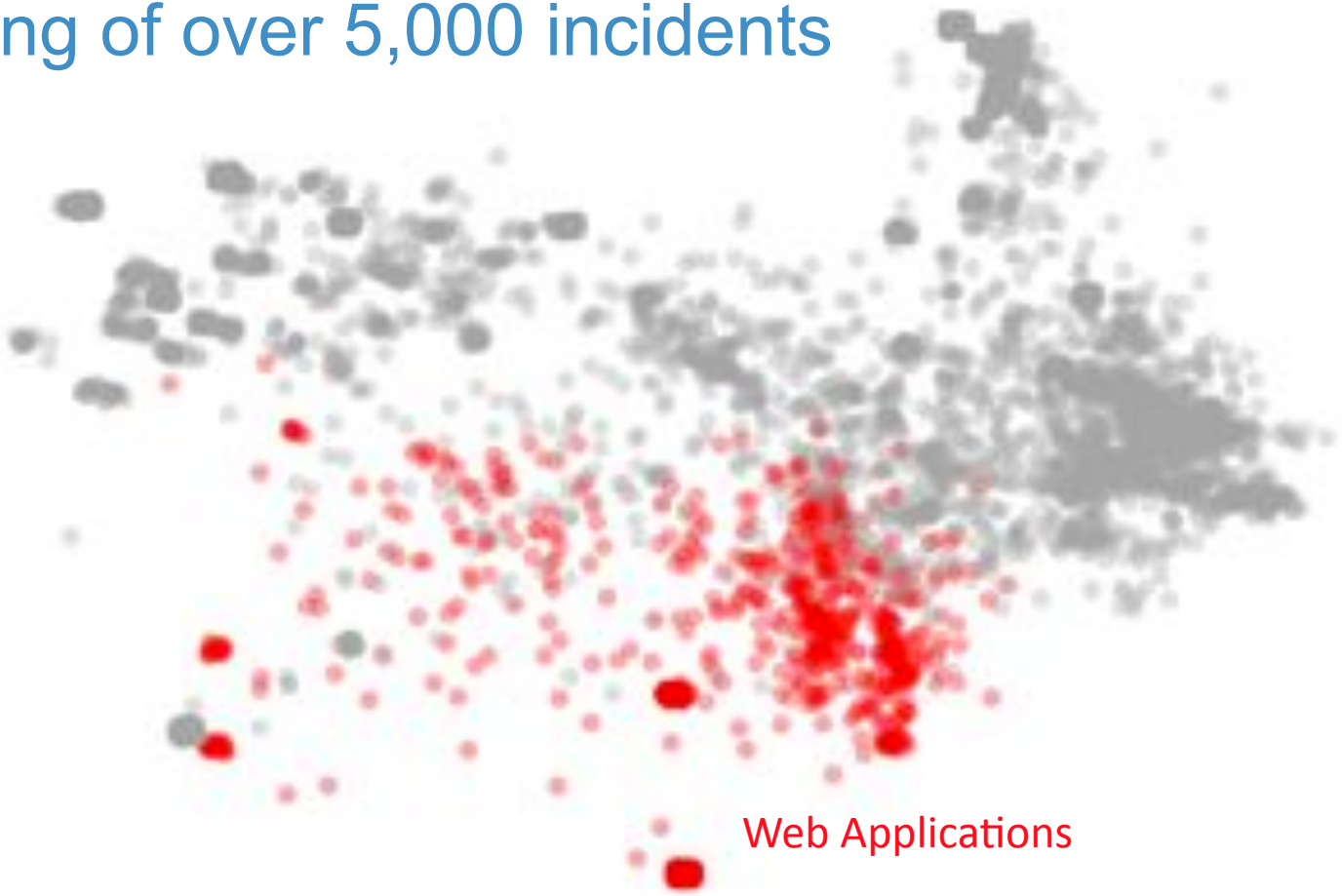
Presenter's
replace

Clustering of over 5,000 incidents



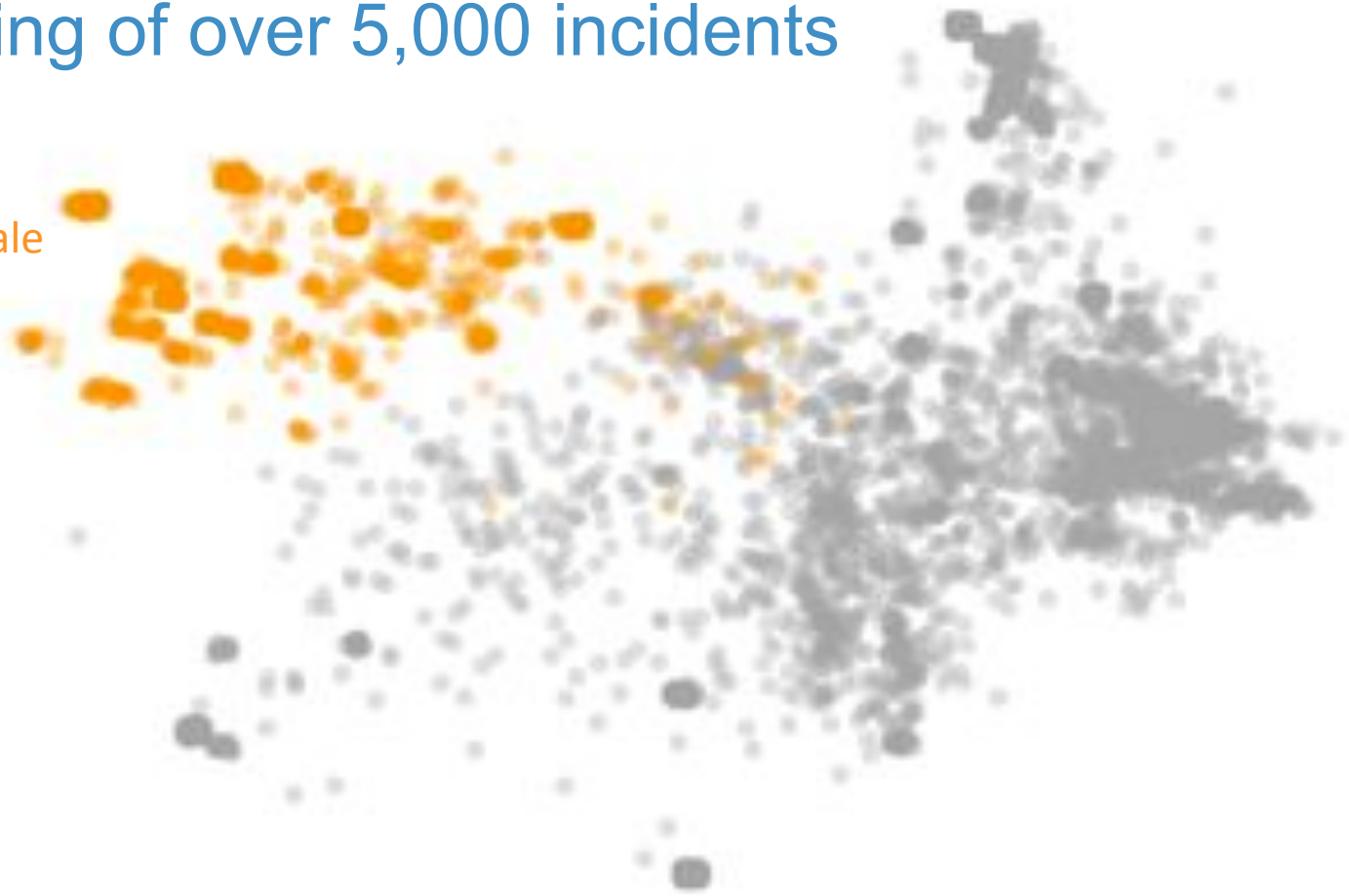
Presenter's
replace

Clustering of over 5,000 incidents



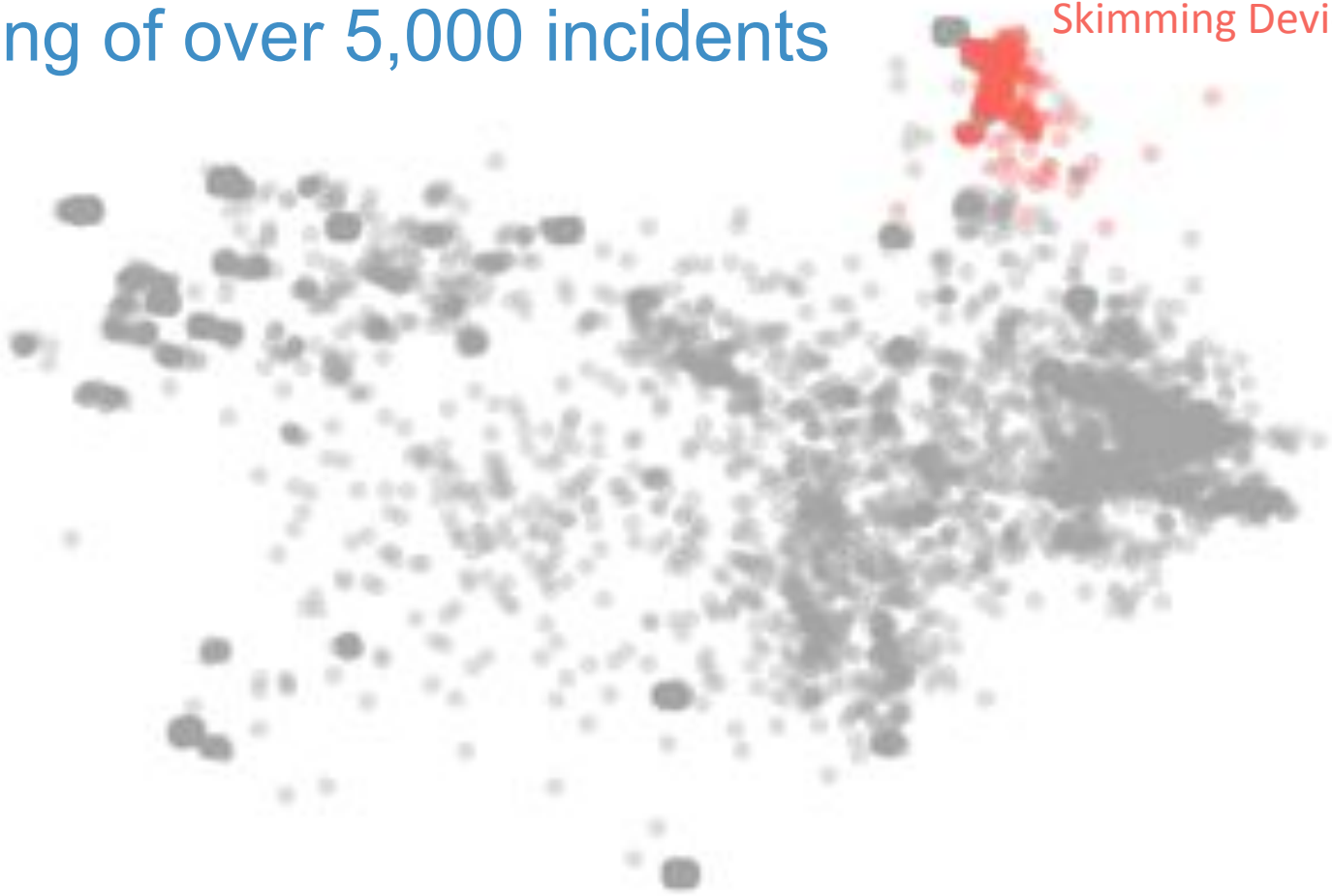
Presenter's
replace

Clustering of over 5,000 incidents



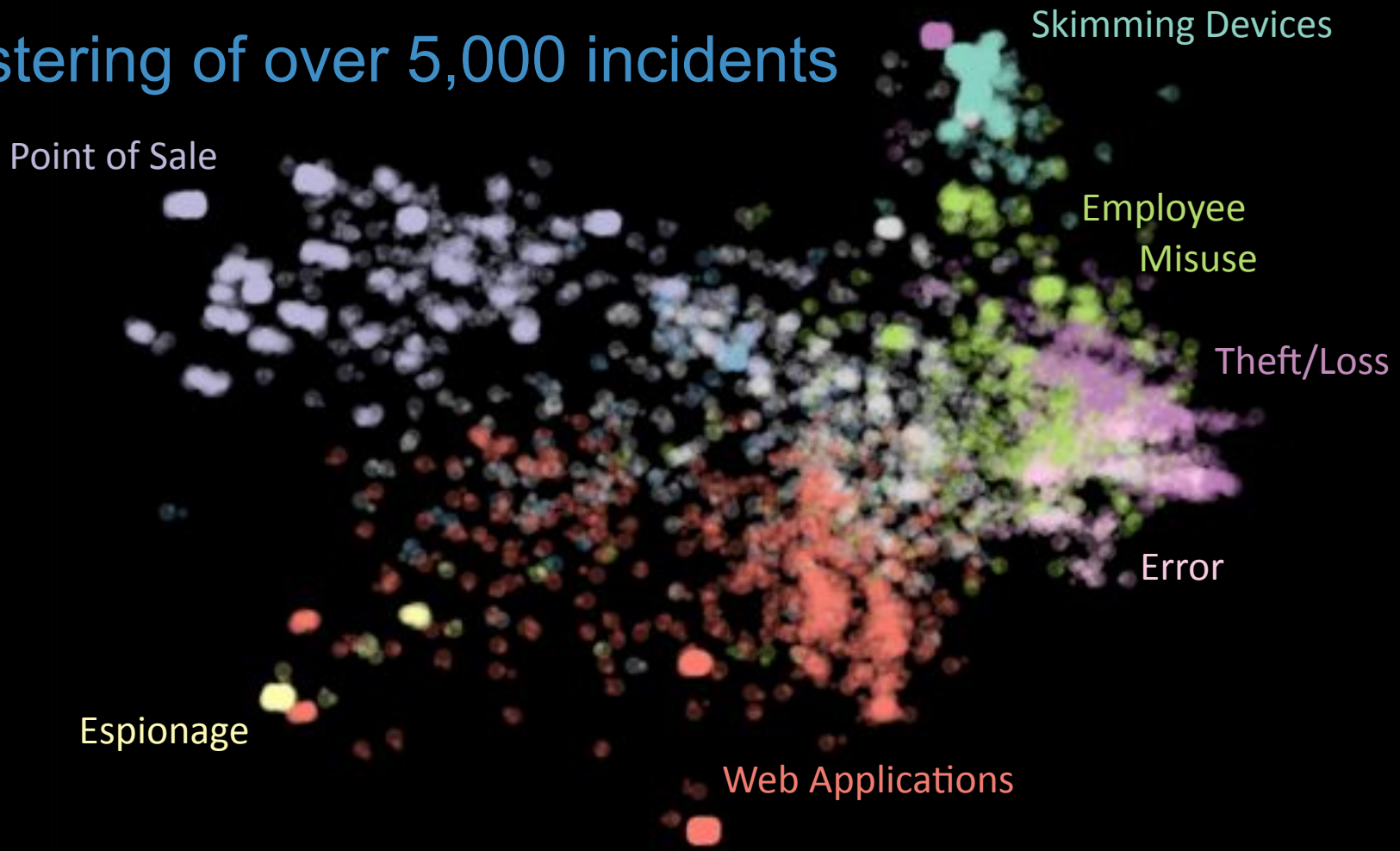
Presenter's
replace

Clustering of over 5,000 incidents



Presenter's
replace

Clustering of over 5,000 incidents

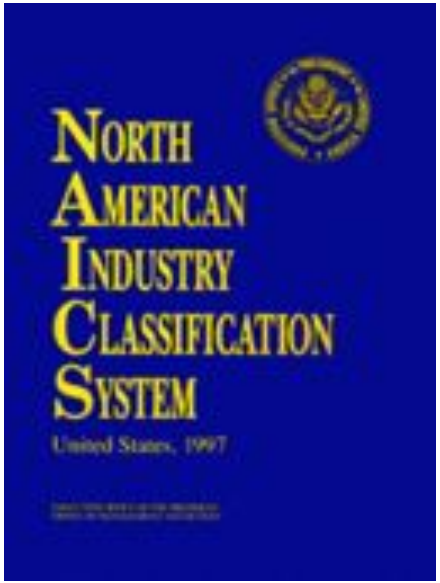


Research Question:
Is there such thing as a “Top 10” list of controls?

Note to RSA reviewers,
We are still working through the data on this and we’ll fill this in.

Presenter’s Company Logo – replace on master slide

Research Question:
What industries are comparable (similar threat

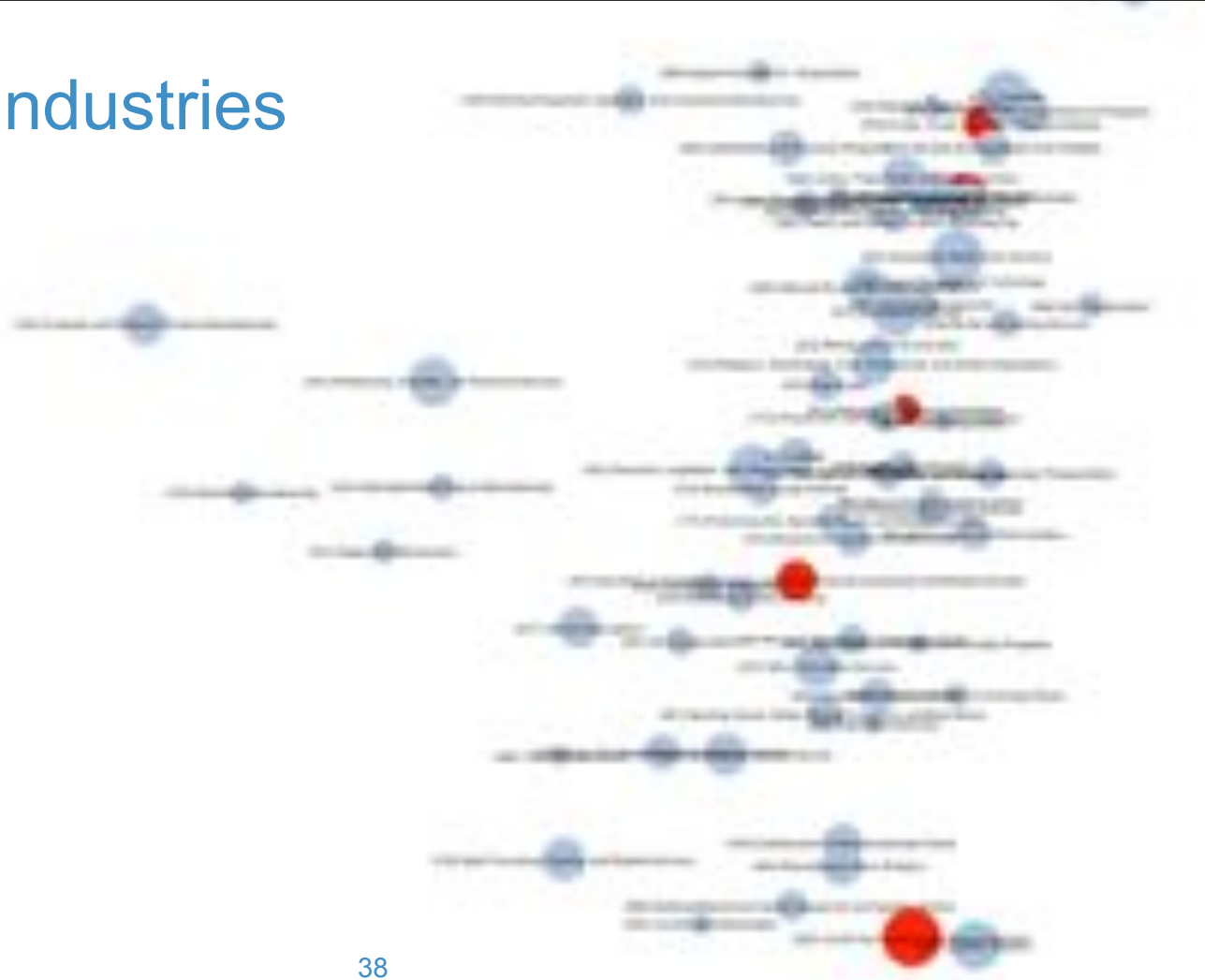


Need to answer this before,
"How am I compared to my peers?"

Presenter's Company Logo – replace on master slide

Hierarchical Clustering across Industries

Clustering on Industries



Presenter's Company Logo – replace on master slide

RSA CONFERENCE 2014

FEBRUARY 24 – 28 | MOSCONE CENTER | SAN FRANCISCO