RSA CONFERENCE 2014
ASIA PACIFIC & JAPAN

Share.
Learn.
Secure.
Capitalizing on
Collective Intelligence

From Data to Wisdom: Big Lessons in Small Data

SESSION ID: CDS-T07

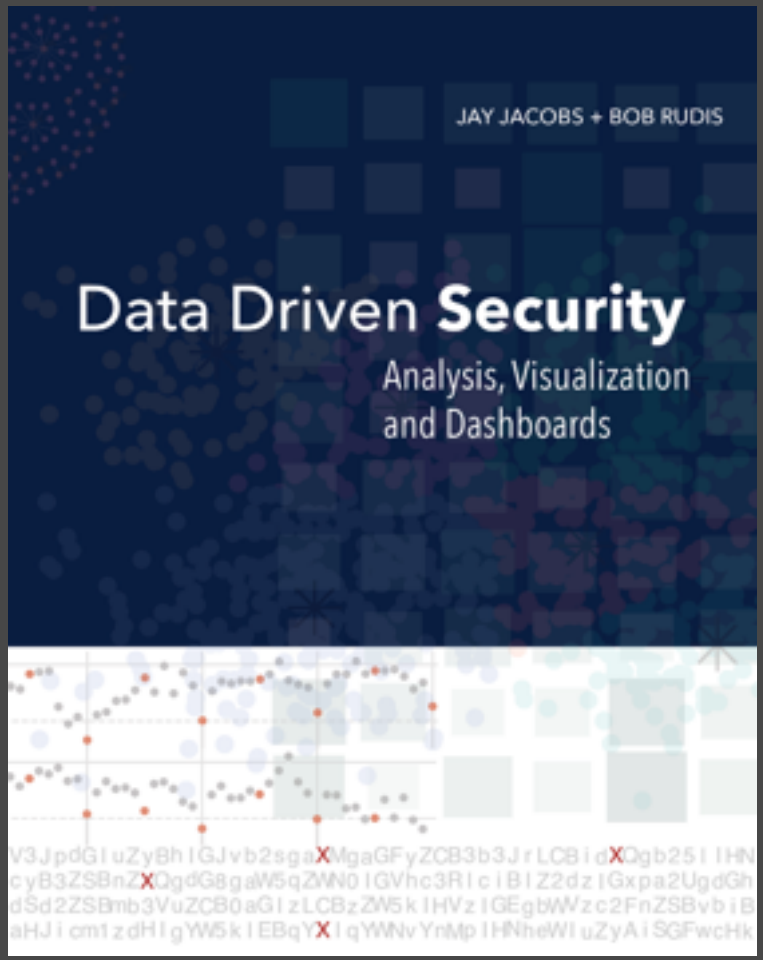
Jay Jacobs

Security Data Scientist
Cybersecurity Research & Innovation
Verizon
@jayjacobs

verizon

Jay Jacobs

http://www.verizonenterprise.com/DBIR/



http://amzn.to/ddsec

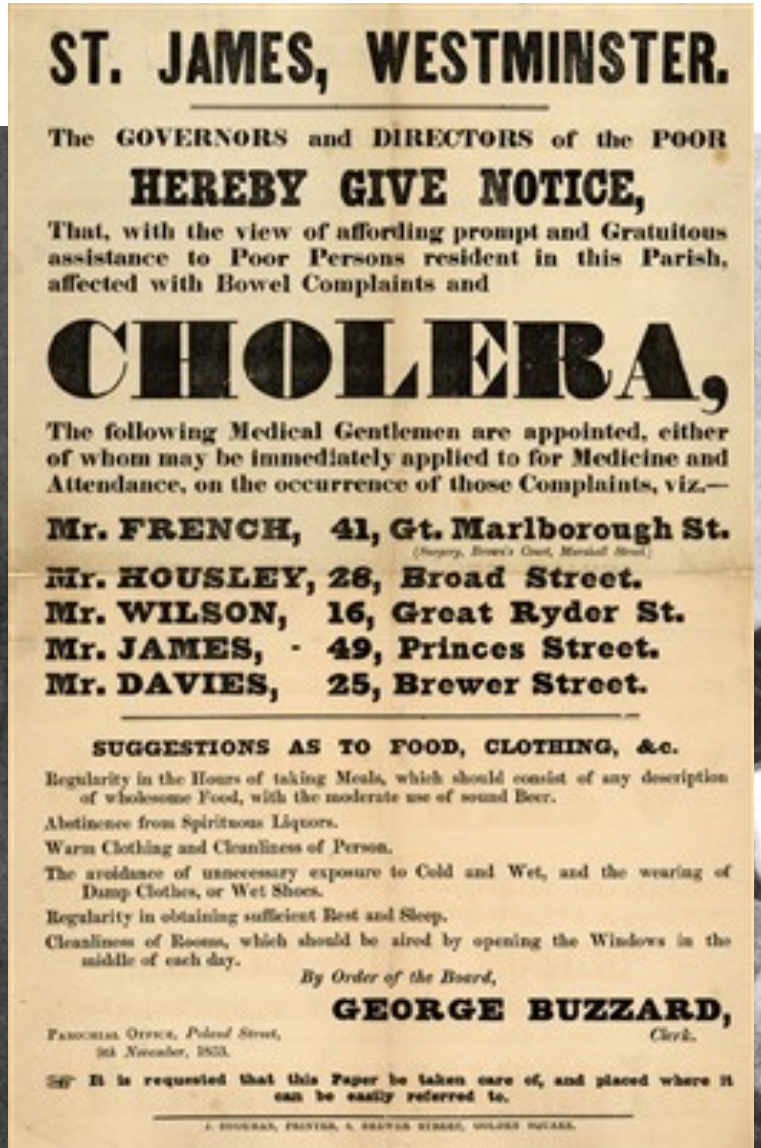

Size matters not.
Look at me.
Judge me by my
size, do you?
Hmm? Hmm.

Agenda

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

Early Data Analysis

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.
(Surgery, Brown's Court, Marshall Street.)
Mr. HOUSLEY, 28, Broad Street.
Mr. WILSON, 16, Great Ryder St.
Mr. JAMES, 49, Princes Street.
Mr. DAVIES, 25, Brewer Street.

SUGGESTIONS AS TO FOOD, CLOTHING, &c.

Regularity in the Hours of taking Meals, which should consist of any description of wholesome Food, with the moderate use of sound Beer.

Abstinence from Spirituous Liquors.

Warm Clothing and Cleanliness of Person.

The avoidance of unnecessary exposure to Cold and Wet, and the wearing of Damp Clothes, or Wet Shoes.

Regularity in obtaining sufficient Rest and Sleep.

Cleanliness of Rooms, which should be aired by opening the Windows in the middle of each day.

By Order of the Board,

GEORGE BUZZARD,
Clerk.

Parochial Office, *Poland Street,*
26 *November*, 1853.

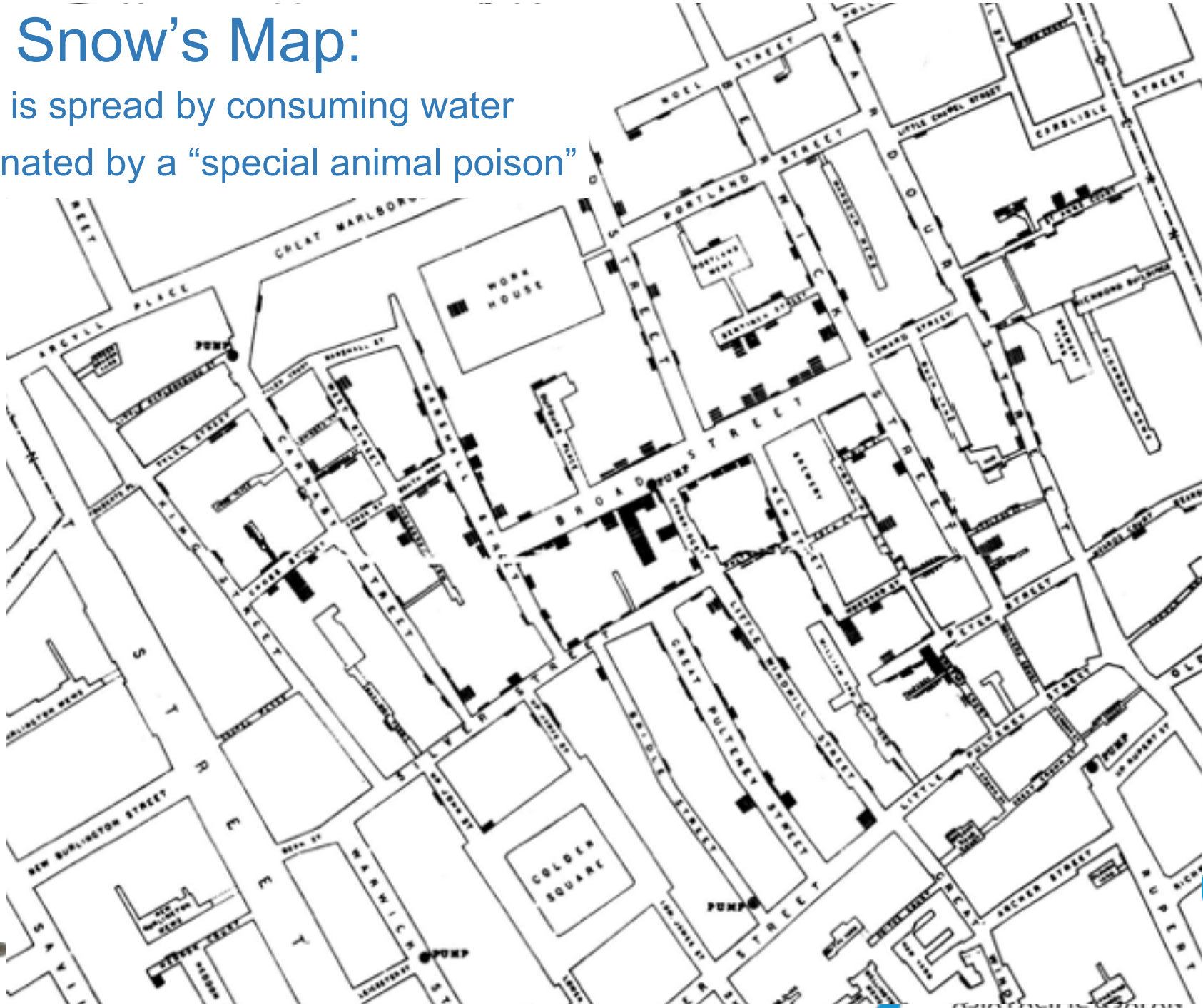
☞ It is requested that this Paper be taken care of, and placed where it can be easily referred to.

J. HOUGHAM, PRINTER, 5, BRAVER STREET, GOLDEN SQUARE.

John Snow
1813-1858

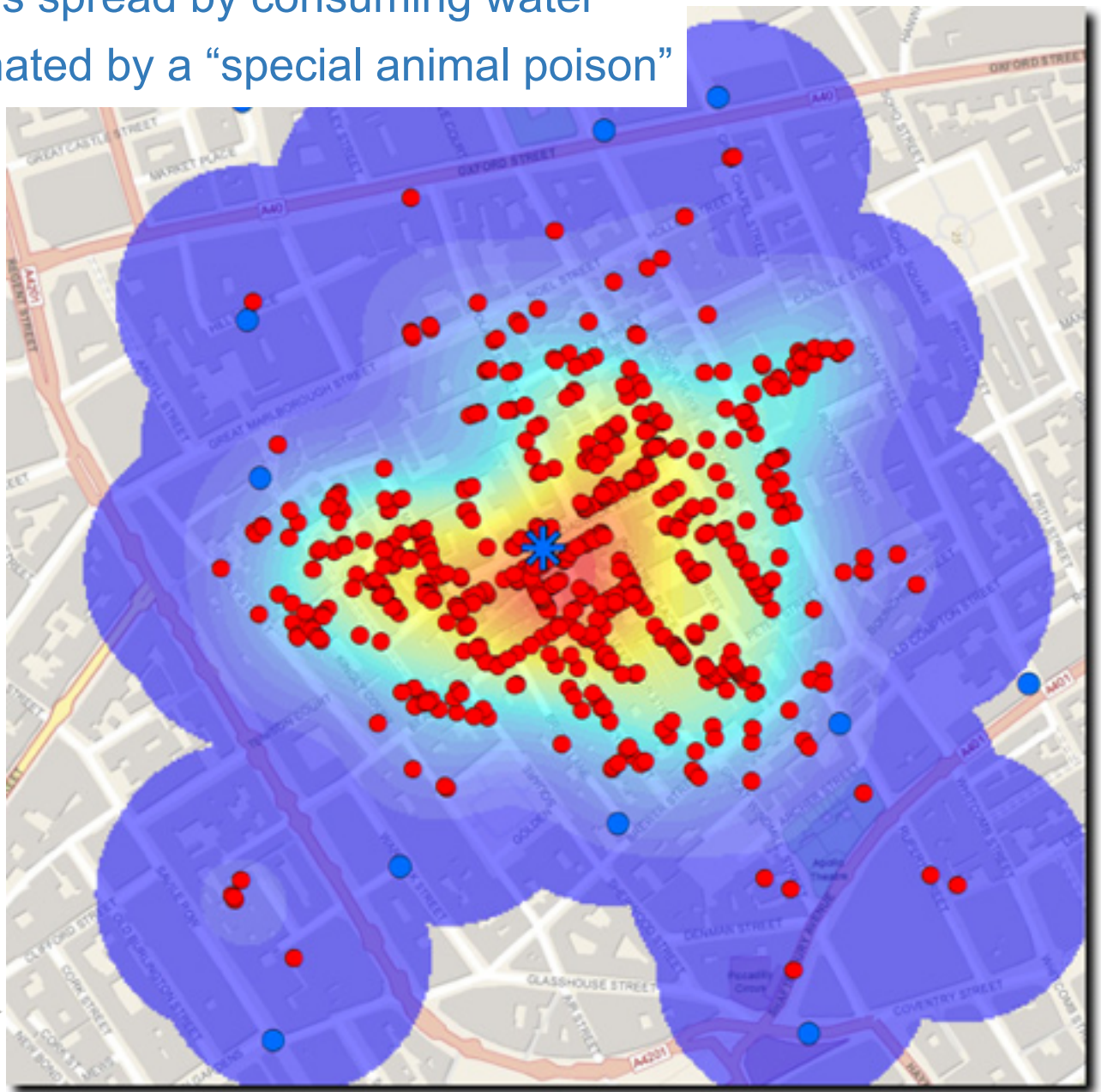
John Snow's Map:

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



John Snow's Map:

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



“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[a]
Newington	144	−2	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	2	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	228	11	6.6	33	5.070	0.067	3

Quote of William Farr as it appears in:
P. Binghama, N.Q. Verlanderb, M.J. Cheala
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

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

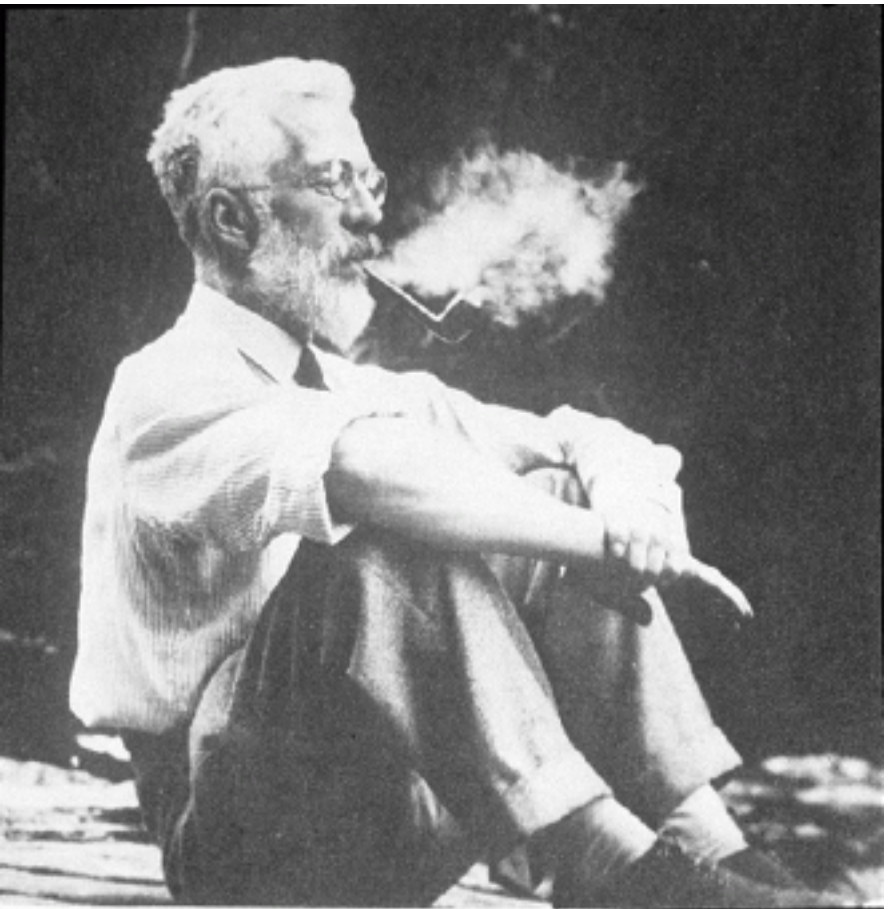
Table 2 Deaths from cholera in London, registered in 1849 by registration district together with eight possible explanatory variables.

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[a]
Newington	144	−2	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	2	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	228	11	6.6	33	5.070	0.067	3

P. Binghama, N.Q. Verlanderb, M.J. Cheala
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...

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

John Tukey
1915-2000

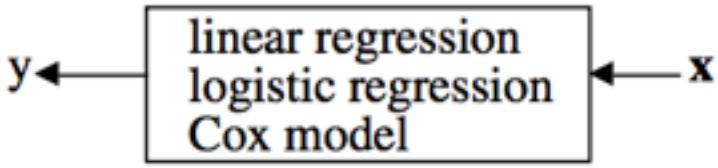
Exploratory
Data Analysis

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

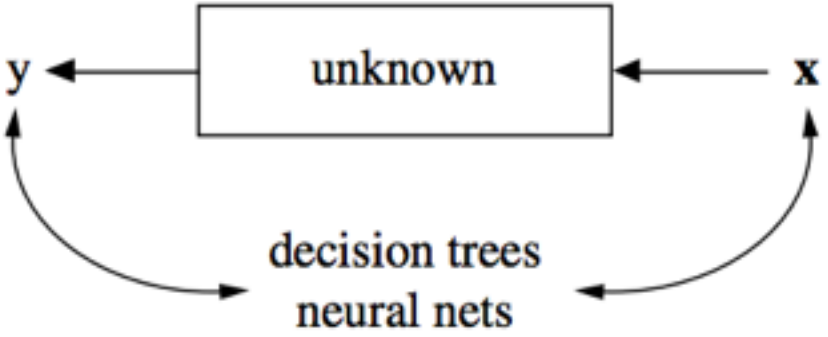
Statistical Modeling: The Two Cultures

Leo Breiman
1928-2005

Data Modeling Culture



Algorithmic Modeling Culture



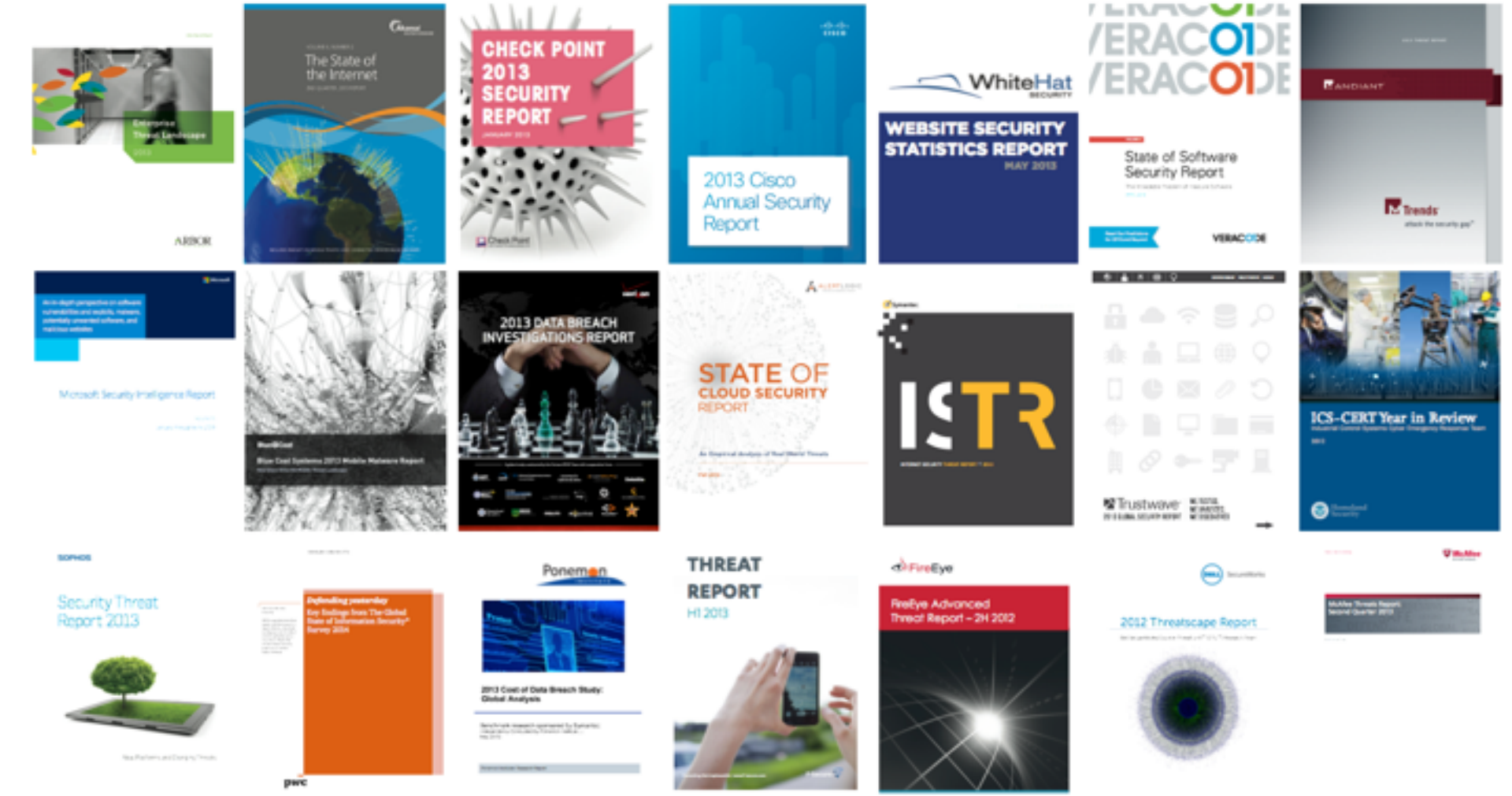
Agenda

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

Where are we in the evolution of data?

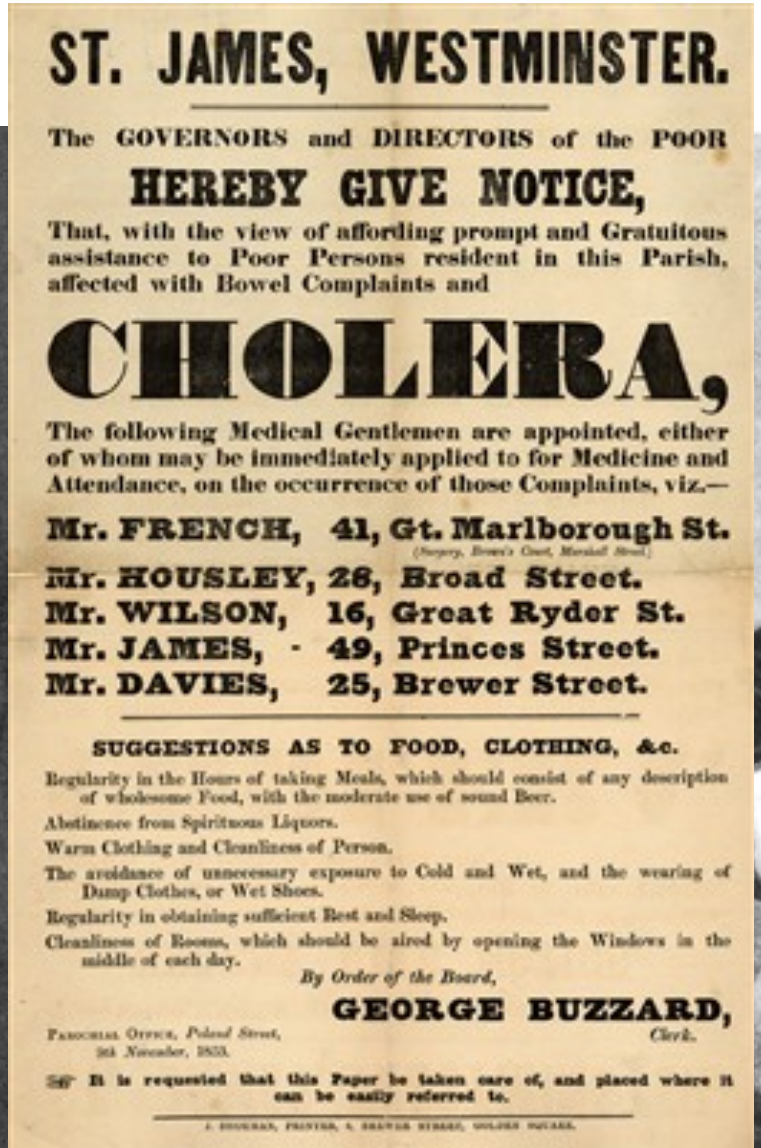
Instinctive Descriptive Inferential Data Mining, Machine Learning Explanatory

State of Data Analysis…



“Modern” Security Data Analysis

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.
(Surgery, Brown's Court, Marshall Street.)
Mr. HOUSLEY, 28, Broad Street.
Mr. WILSON, 16, Great Ryder St.
Mr. JAMES, 49, Princes Street.
Mr. DAVIES, 25, Brewer Street.

SUGGESTIONS AS TO FOOD, CLOTHING, &c.

Regularity in the Hours of taking Meals, which should consist of any description of wholesome Food, with the moderate use of sound Beer.

Abstinence from Spirituous Liquors.

Warm Clothing and Cleanliness of Person.

The avoidance of unnecessary exposure to Cold and Wet, and the wearing of Damp Clothes, or Wet Shoes.

Regularity in obtaining sufficient Rest and Sleep.

Cleanliness of Rooms, which should be aired by opening the Windows in the middle of each day.

By Order of the Board,

GEORGE BUZZARD,
Clerk.

Parochial Office, *Poland Street,*
26 *November*, 1853.

☞ It is requested that this Paper be taken care of, and placed where it can be easily referred to.

J. HOUGHAM, PRINTER, 5, BRAVER STREET, GOLDEN SQUARE.

John Snow
1813-1858

Where are we in the evolution of data?

Instinctive Descriptive Inferential Data Mining, Machine Learning Explanatory

Where are we in the evolution of data?

Instinctive Descriptive Inferential Data Mining, Machine Learning Explanatory

Where are we in the evolution of data?

percent/%	47
total	32

Instinctive

median	7
std.dev.	2
outlier	4
quartile	0
percentile	0
skew	8
mean	5
variance	3

Descriptive

p-value	3
confidence interval	2
anova	1
hypothesis testing	1
margin of error	2
response rate	1
sampling error	1
sampling bias	1
binomial	1
gaussian/bell	1
regression	3
<Stats person>	1
residual	0

Inferential

Data Mining, Machine Learning

Explanatory

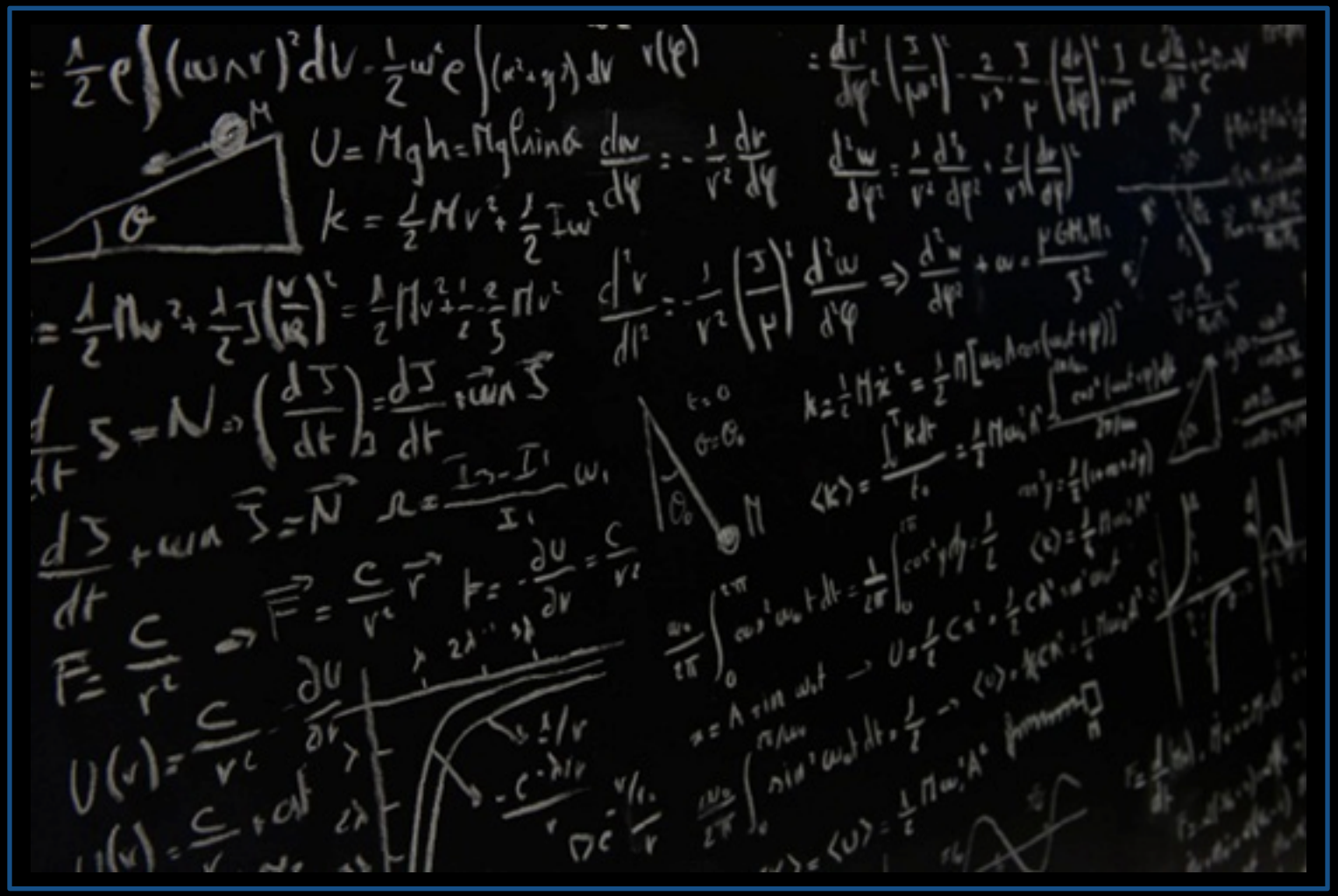
The Path of Data Evolution…

Curiosity, Statistics, Transparency

The Path of Data Evolution…

Curiosity, Statistics, Transparency

Science!



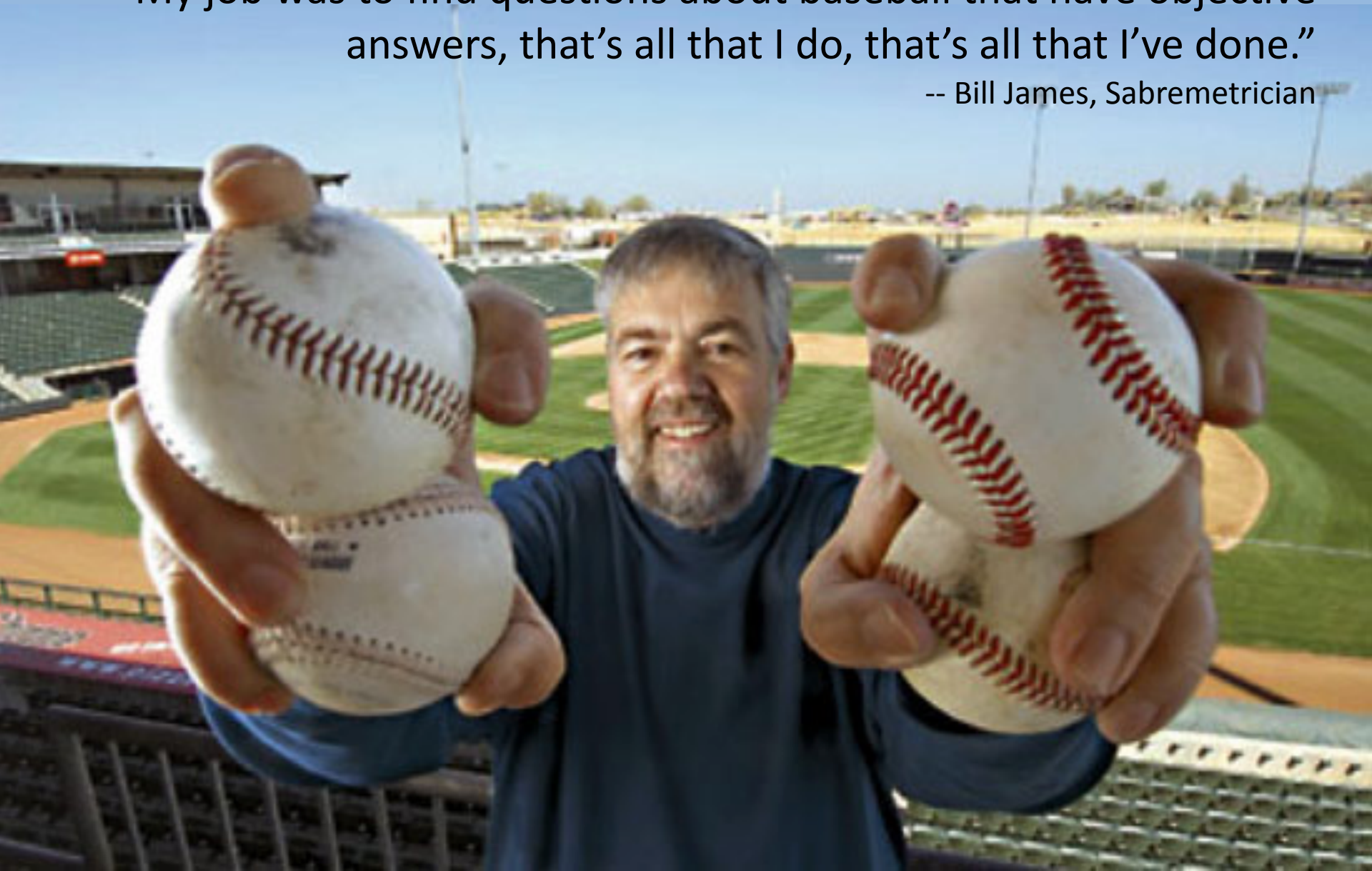
SCIENCE

Because figuring things out is always better than making stuff up

Curiosity…

"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, Sabremetrician



Statistics…

statistics are

statistics are – Google Search

statistics are **for losers**

statistics are **made up**

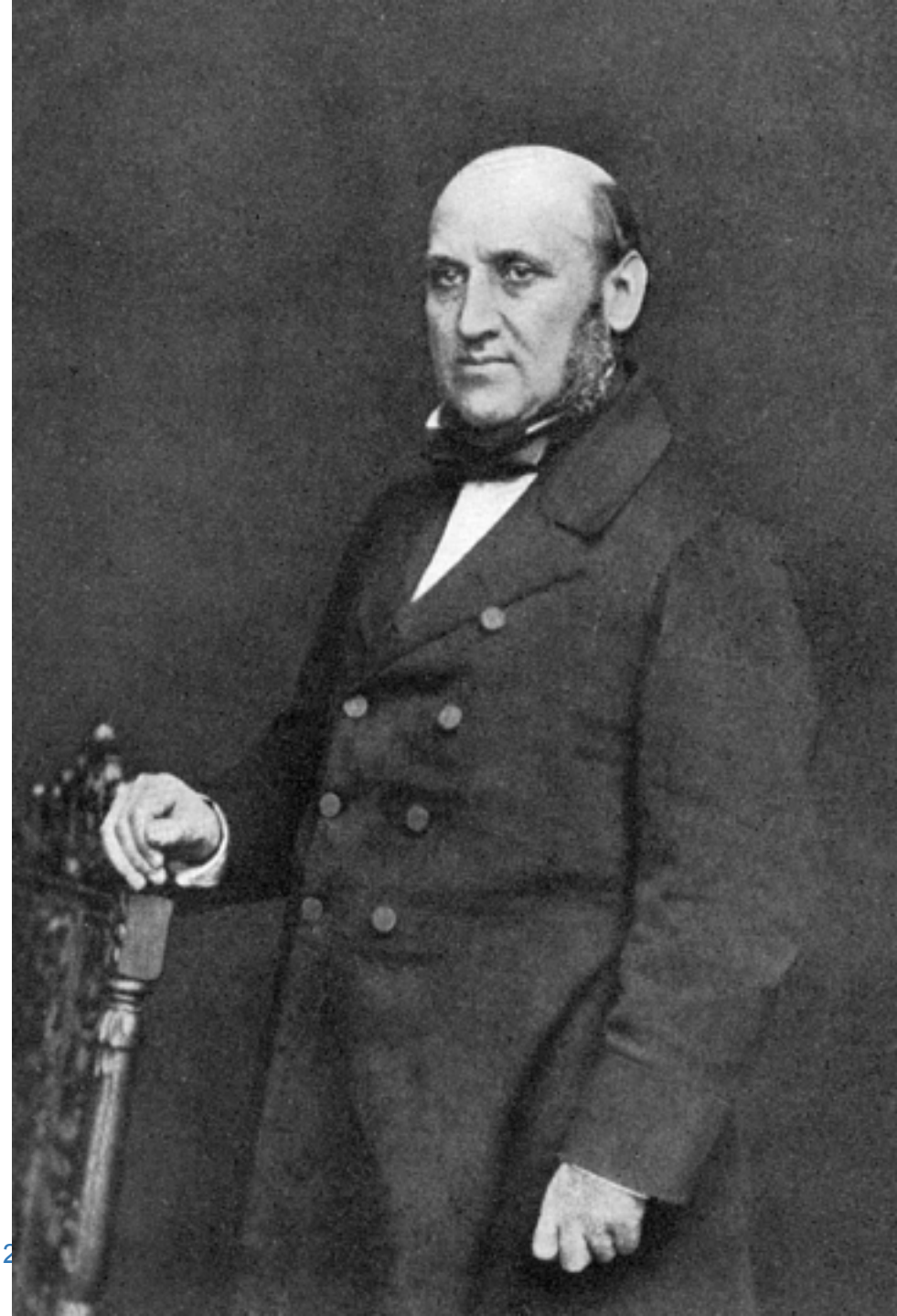
statistics are **lies**

statistics are **bullshit**

statistics are **no substitute for judgment**

Statistics…

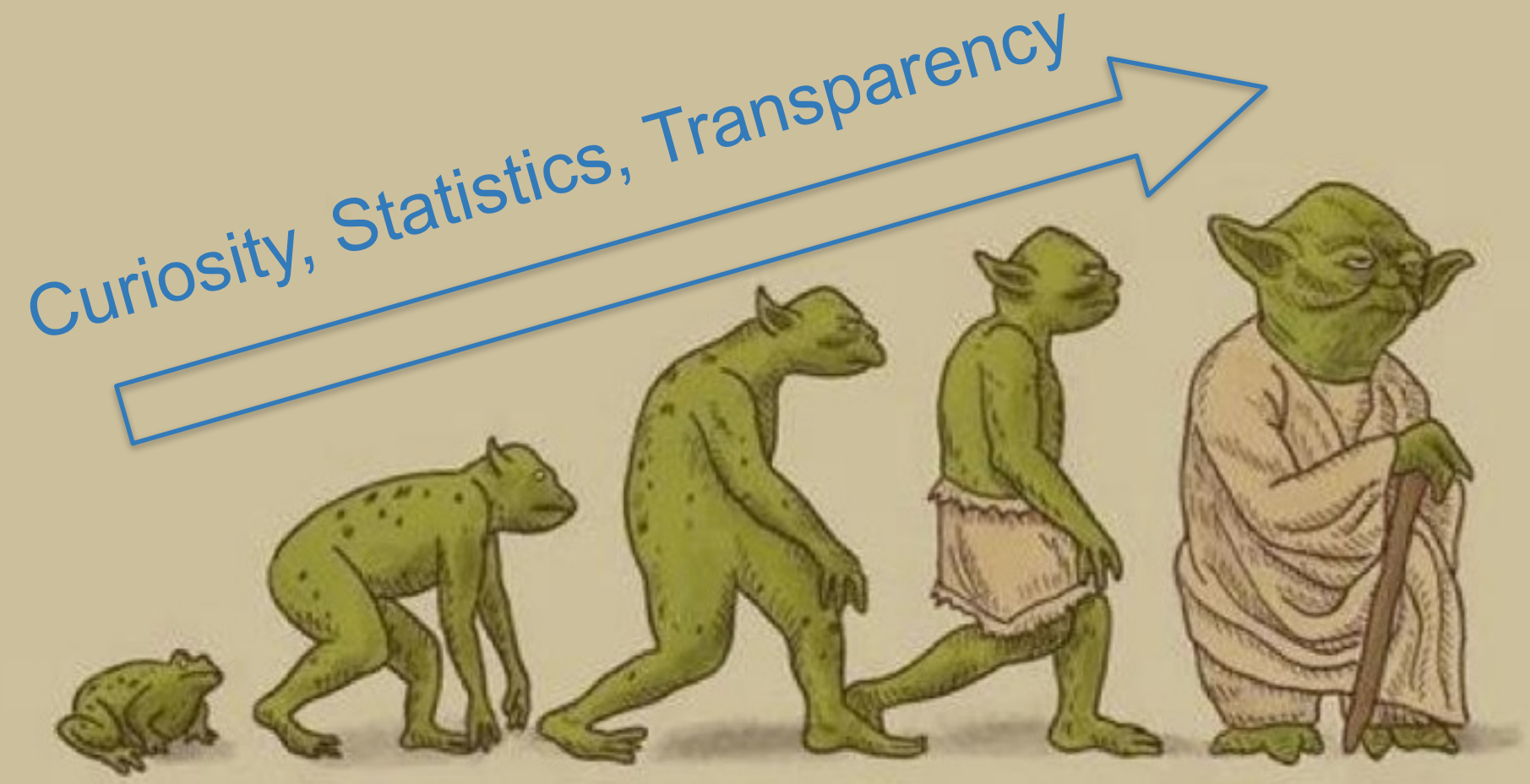
"Had logistic regression been available to Farr, its application to his 1852 data set would have changed his conclusion."



Transparency…

"If you cannot — in the long run — tell everyone what you have been doing, your doing is worthless."

Erwin Schrödinger
"Science and Humanism, Physics in our Time"
Dublin, 1950


Curiosity, Statistics, Transparency
Instinctive
Descriptive
Inferential
Data Mining, Machine Learning
Explanatory

Agenda

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

Research Question…

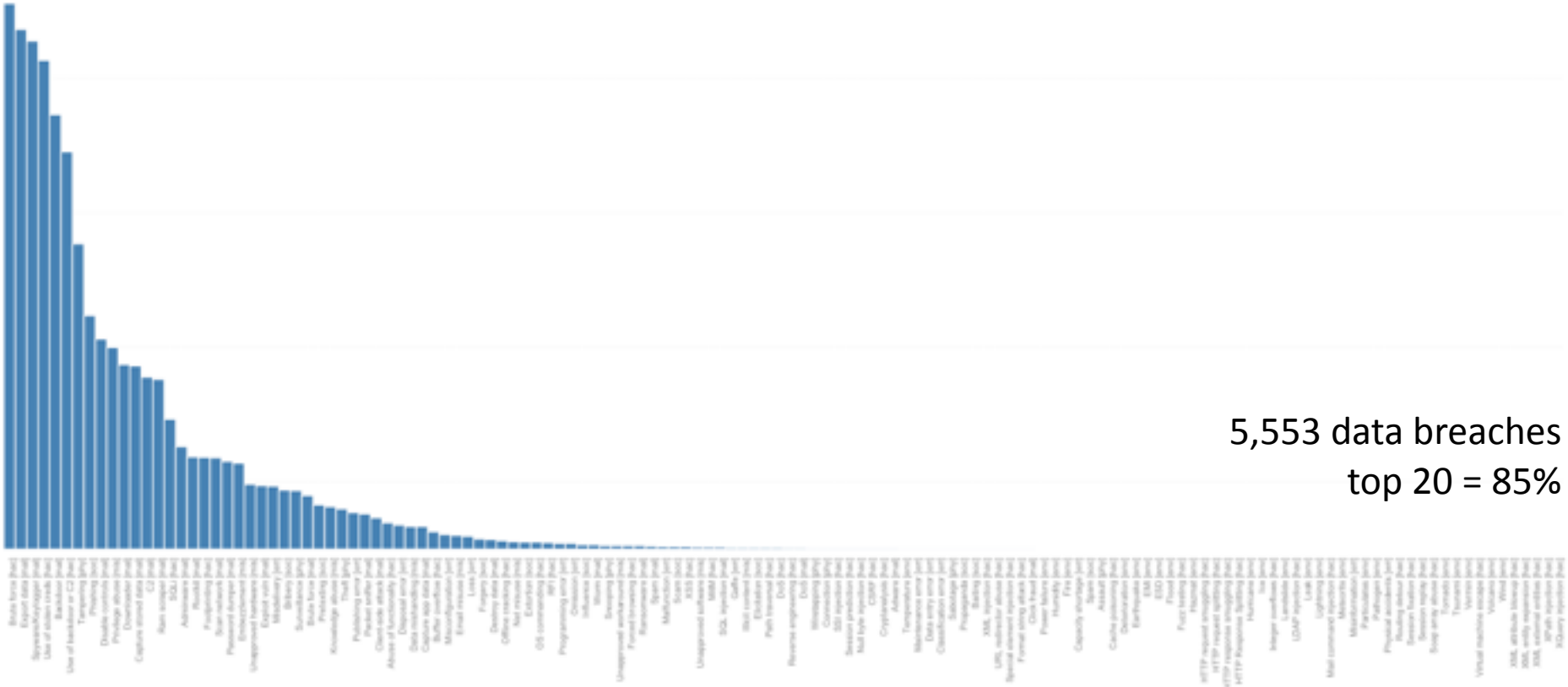
Are threat actors unpredictable?

(unpredictable == random)

Define the “key space”

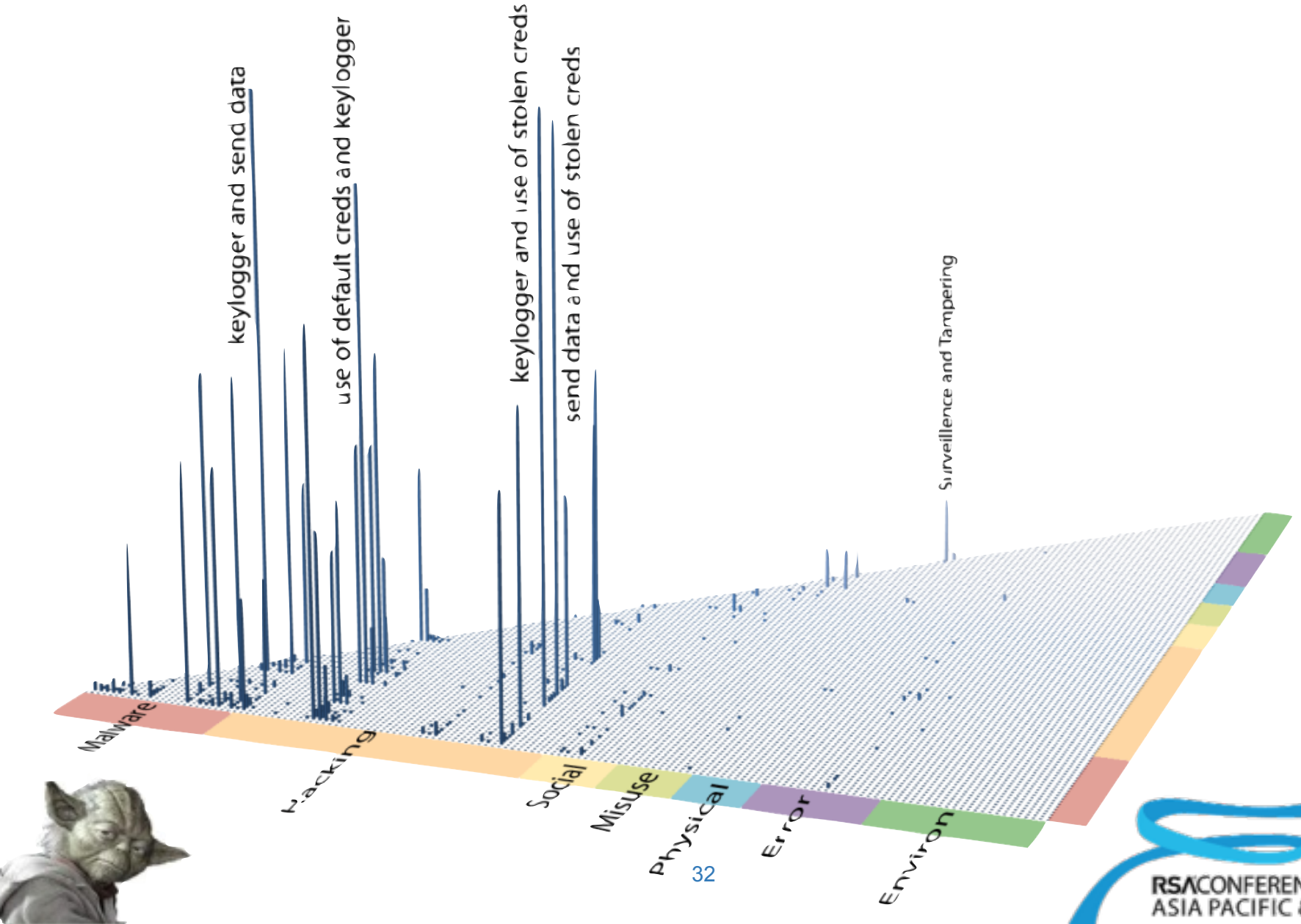
Adware , Backdoor , Brute force , Capture app data , Capture stored data , Client-side , C2 , Destroy data, Disable controls , DoS , Downloader , Exploit vuln , Export data , Packet sniffer , Password dumper , Ram scraper , Ransomware , Rootkit , Scan network , Spam , Spyware , SQL injection , Utility , Worm , Abuse of functionality , Brute force , Buffer overflow , Cache poisoning , Credential/session prediction , Cross-site request forgery , Cross-site scripting , Cryptanalysis , Denial of service , Footprinting and fingerprinting , Forced browsing , Format string attack , Fuzz testing , HTTP request smuggling , HTTP request splitting , HTTP response smuggling , HTTP Response Splitting , Integer overflows , LDAP injection , Mail command injection , Man-in-the-middle attack , Null byte injection , Offline cracking , OS commanding , Path traversal , Remote file inclusion , Reverse engineering , Routing detour , Session fixation , Session replay , Soap array abuse , Special element injection , SQL injection , SSL injection , URL redirector abuse , Use of backdoor or C2 , Use of stolen creds , XML attribute blowup , XML entity expansion , XML external entities , XML injection , XPath injection , XQuery injection , Baiting , Bribery , Elicitation , Extortion , Forgery , Influence , Scam , Phishing , Pretexting , Propaganda , Spam , Knowledge abuse , Privilege abuse , Embezzlement , Data mishandling , Email misuse , Net misuse , Illicit content , Unapproved workaround , Unapproved hardware , Unapproved software , Assault , Sabotage , Snooping , Surveillance , Tampering , Theft , Wiretapping , Classification error , Data entry error , Disposal error , Gaffe , Loss , Maintenance error , Misconfiguration , Misdelivery , Misinformation , Omission , Physical accidents , Capacity shortage , Programming error , Publishing error , Malfunction , Deterioration , Earthquake , EMI , ESD , Temperature , Fire , Flood , Hazmat , Humidity , Hurricane , Ice , Landslide , Lightning , Meteorite , Particulates , Pathogen , Power failure , Tornado , Tsunami , Vermin , Volcano , Leak , Wind

Randomness?



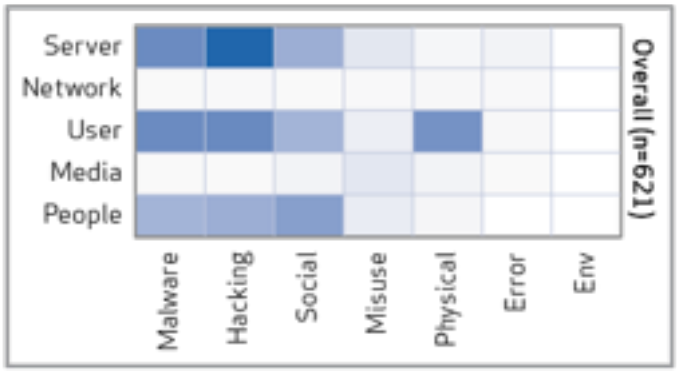
5,553 data breaches
top 20 = 85%

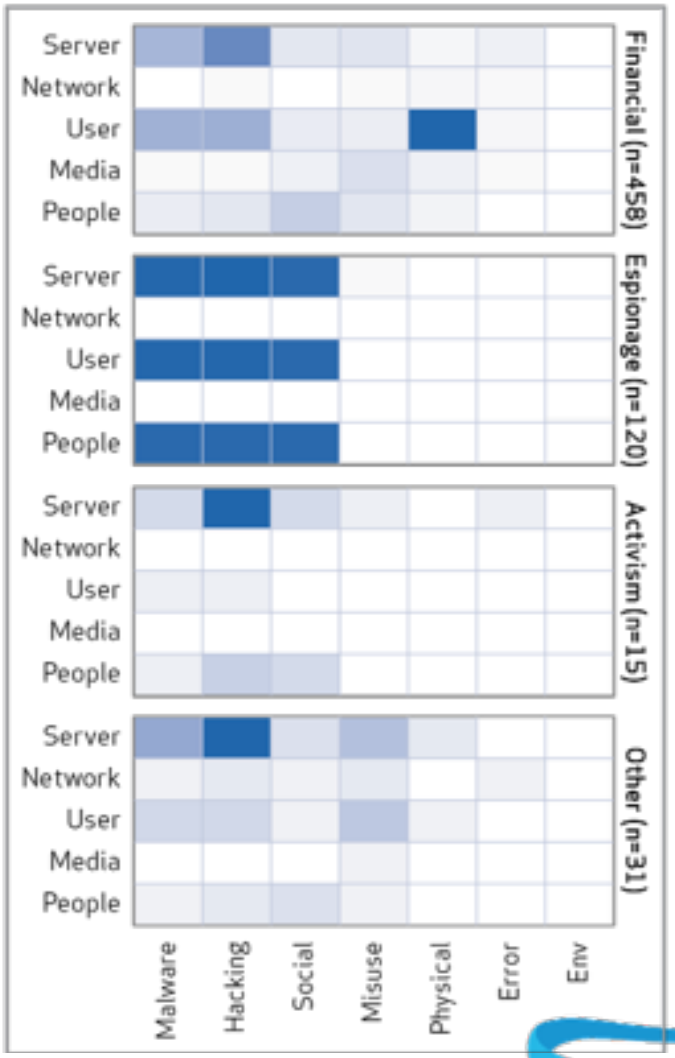
Randomness?



Maybe motives influence actions?

Figure 8: VERIS A² grid depicting associations between actions and assets (split by actor motive)



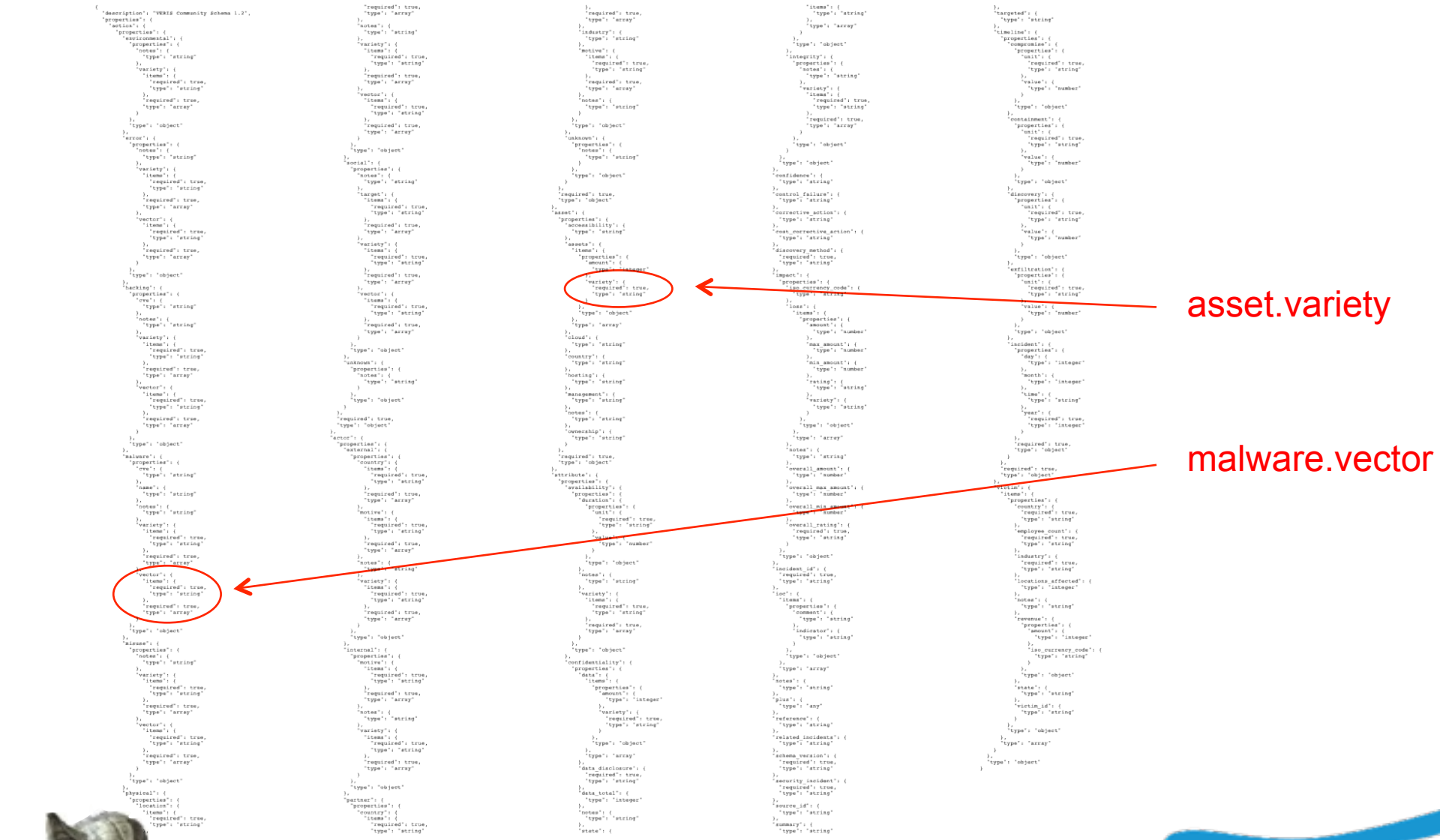


Updating the research question…

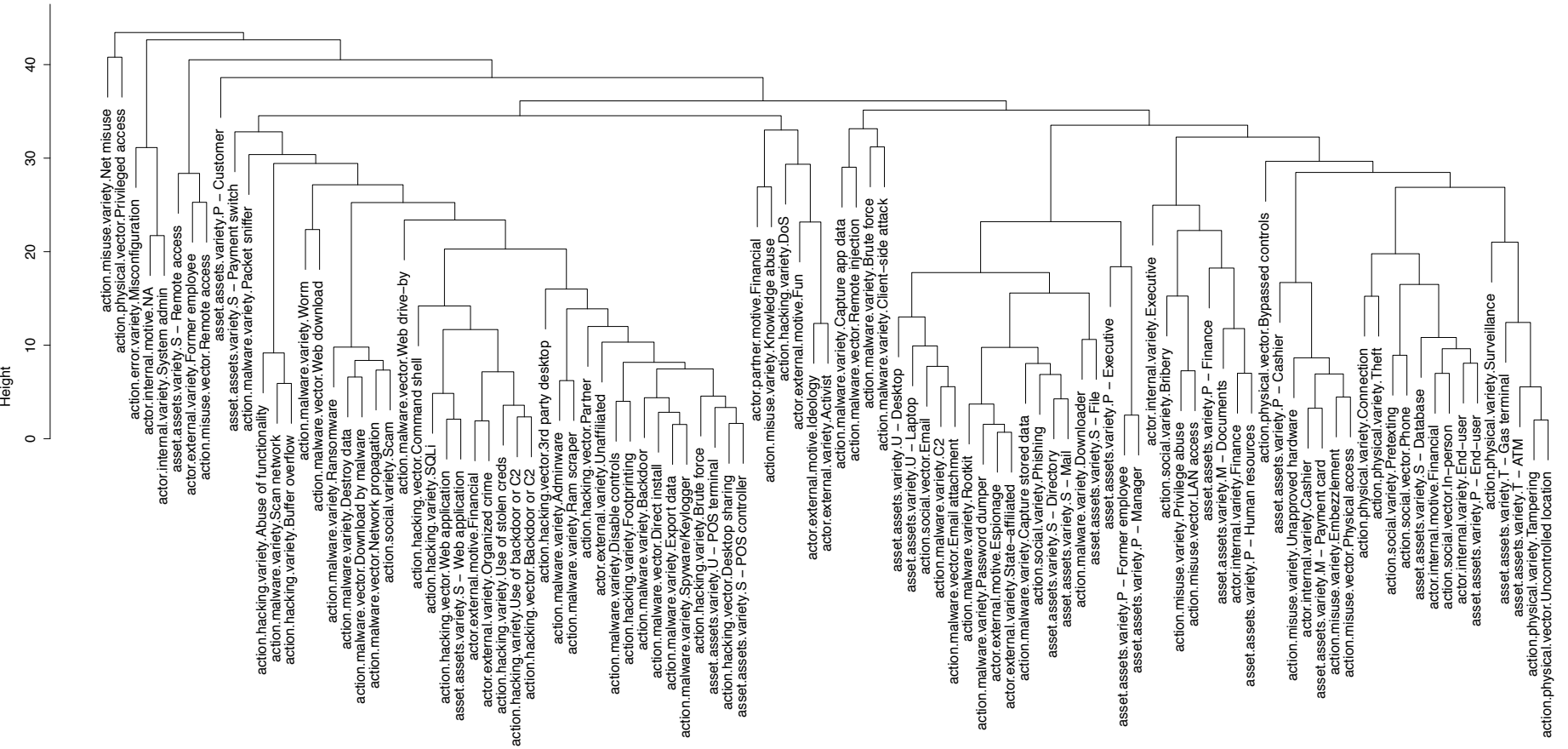
Since threat actors are not random…

Can we detect patterns in data breaches?

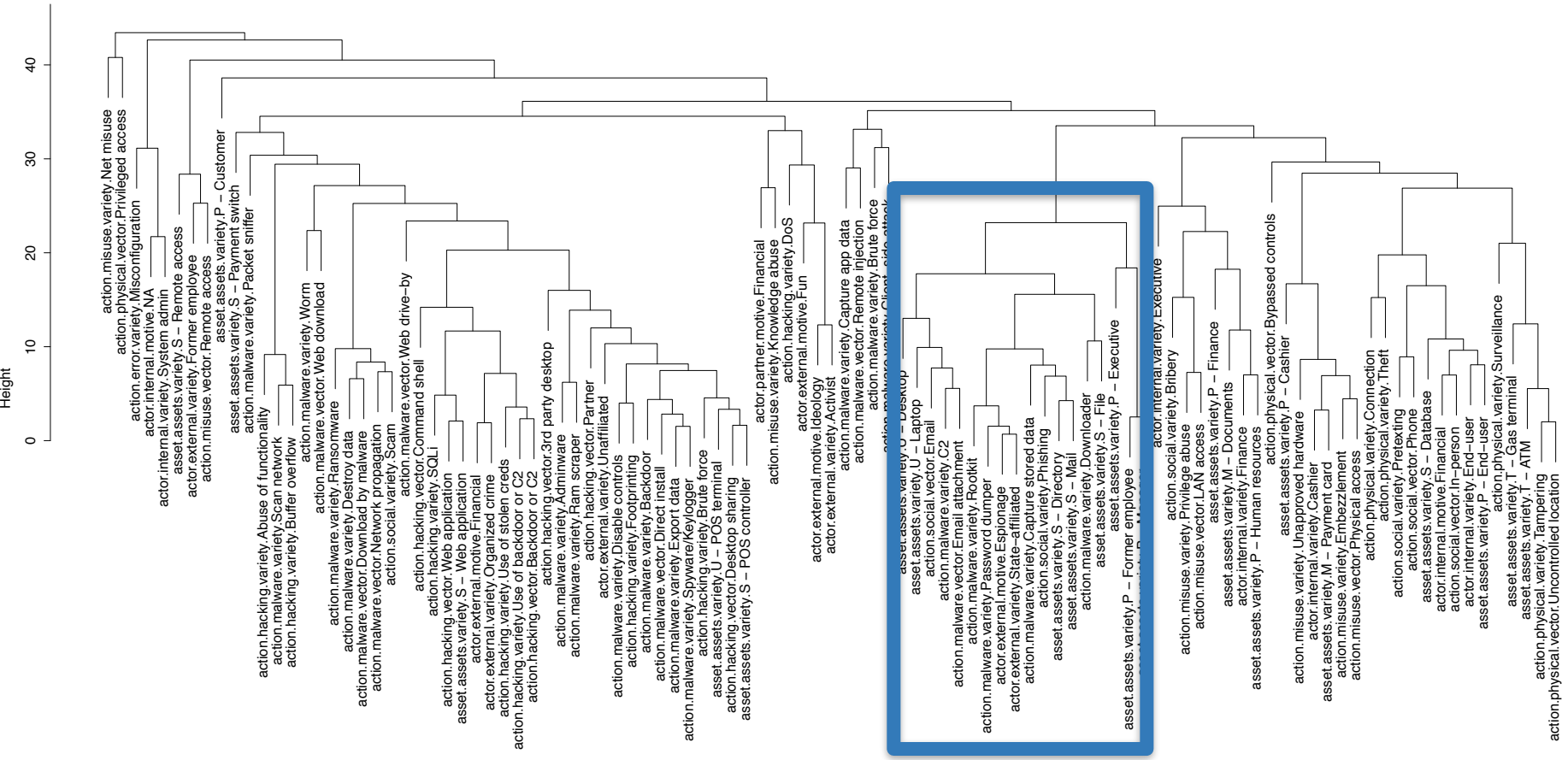
“DNA” of a security incident



Pattern Identification…

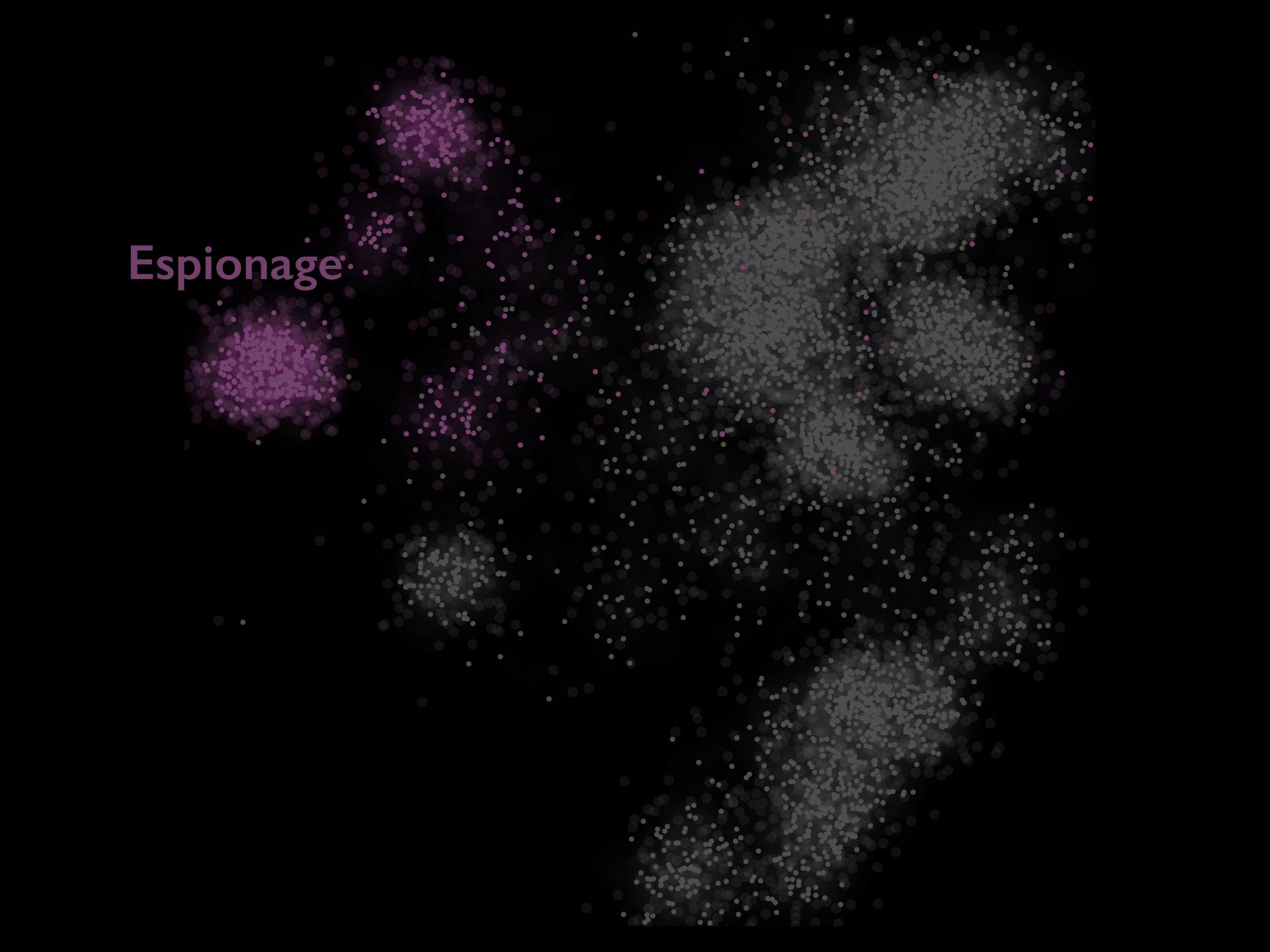


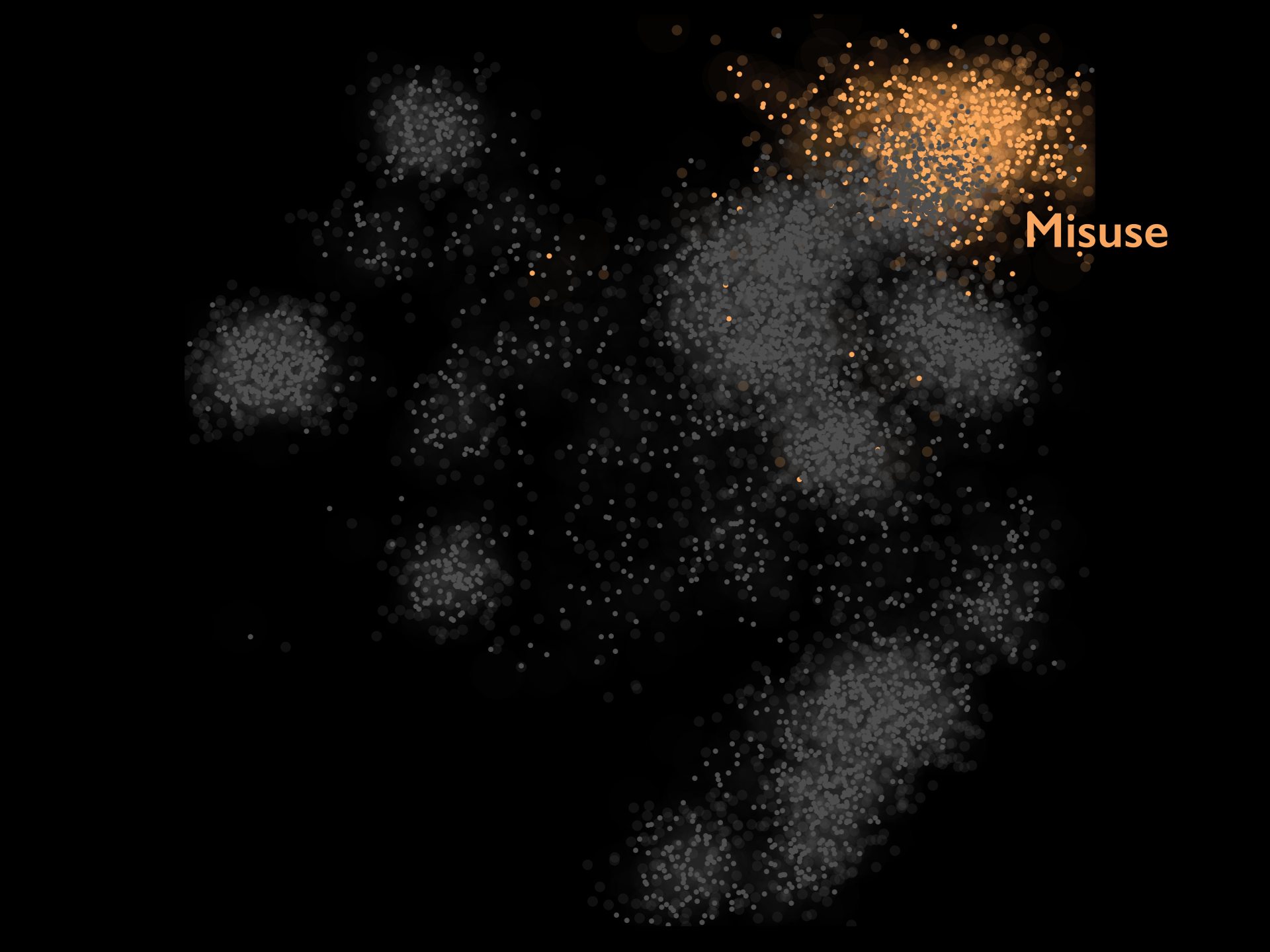
Pattern Identification…

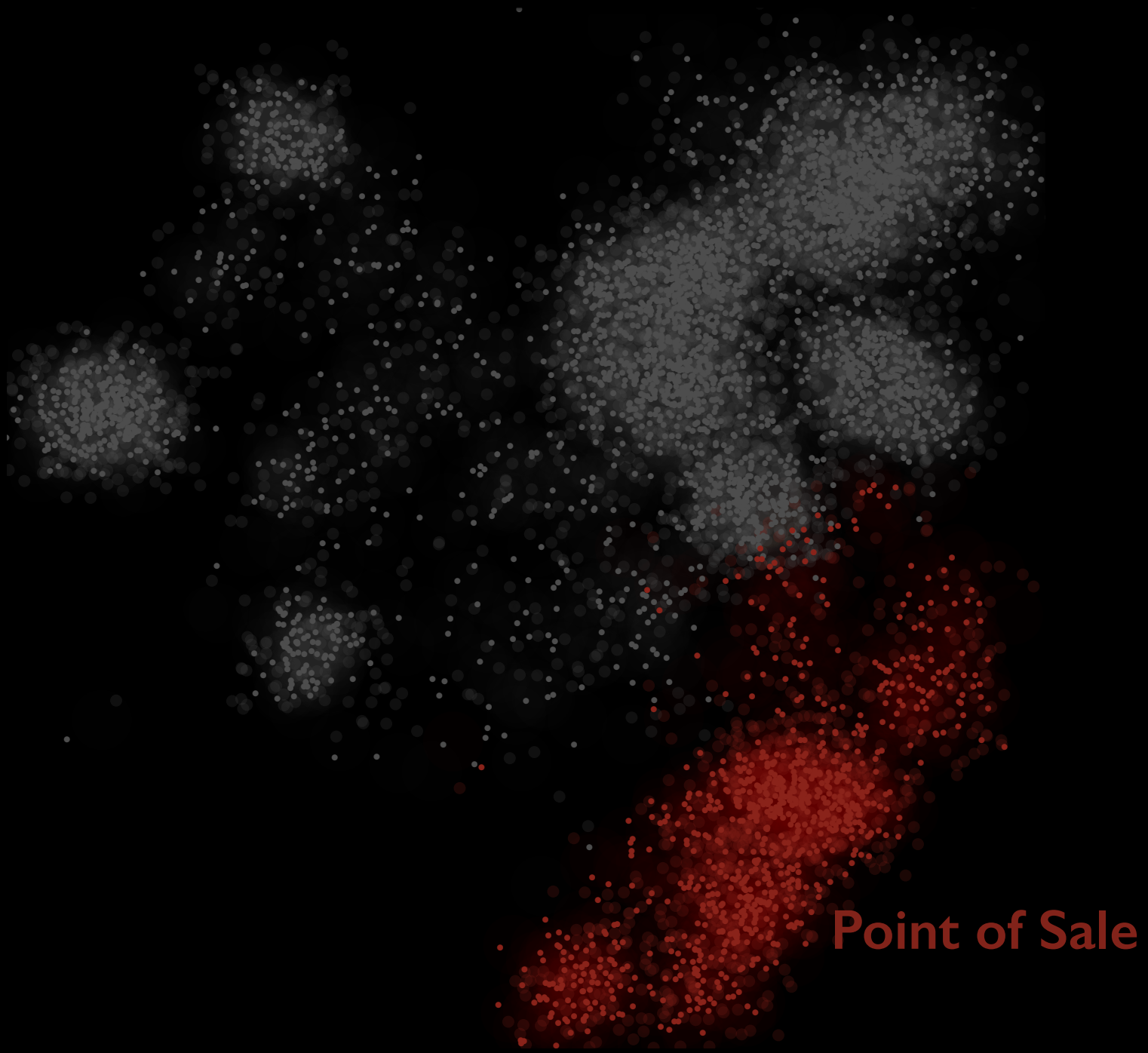


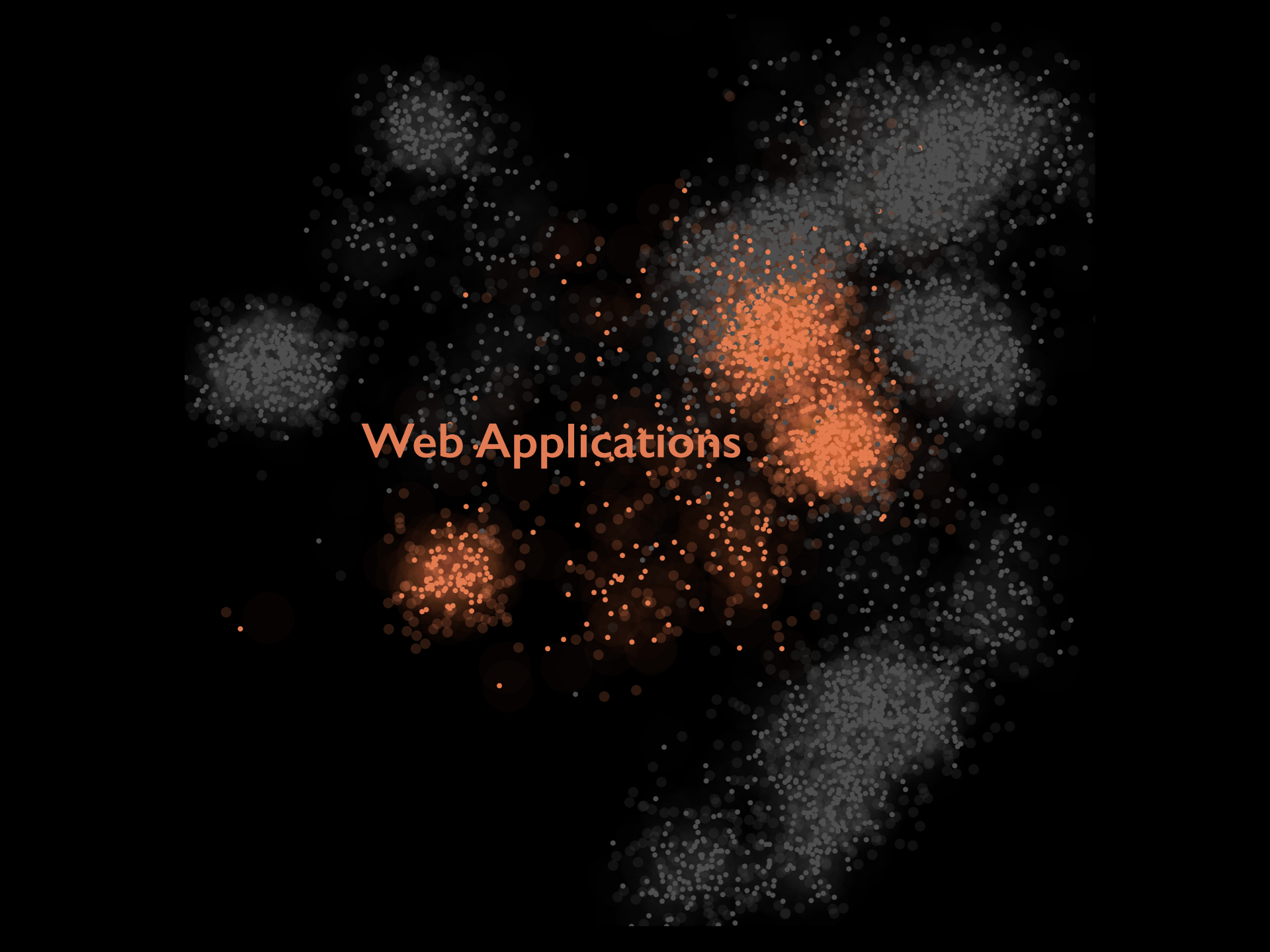
Key variables in “Espionage” pattern

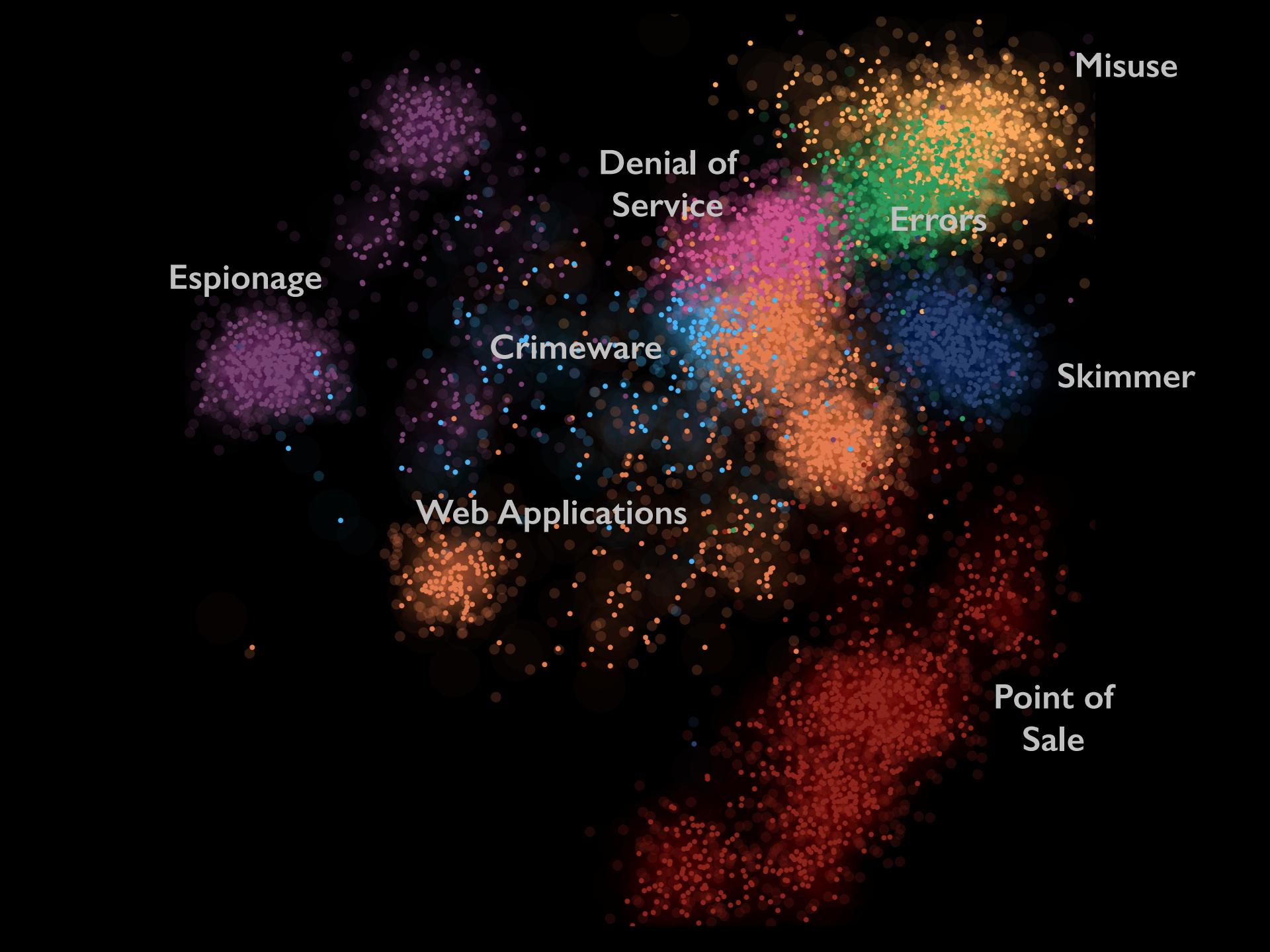
(actor.external.state-affiliated, malware.variety.rootkit, malware.variety.pwd dumper, social.variety.phishing, asset.variety.S – directory, etc)


Espionage


Misuse


Point of Sale


Web Applications


Misuse
Denial of Service
Errors
Espionage
Crimeware
Skimmer
Web Applications
Point of Sale

So what?

- Adversaries are not random.
- They exhibit tendencies.
- Tendencies lead to patterns.

Therefore:

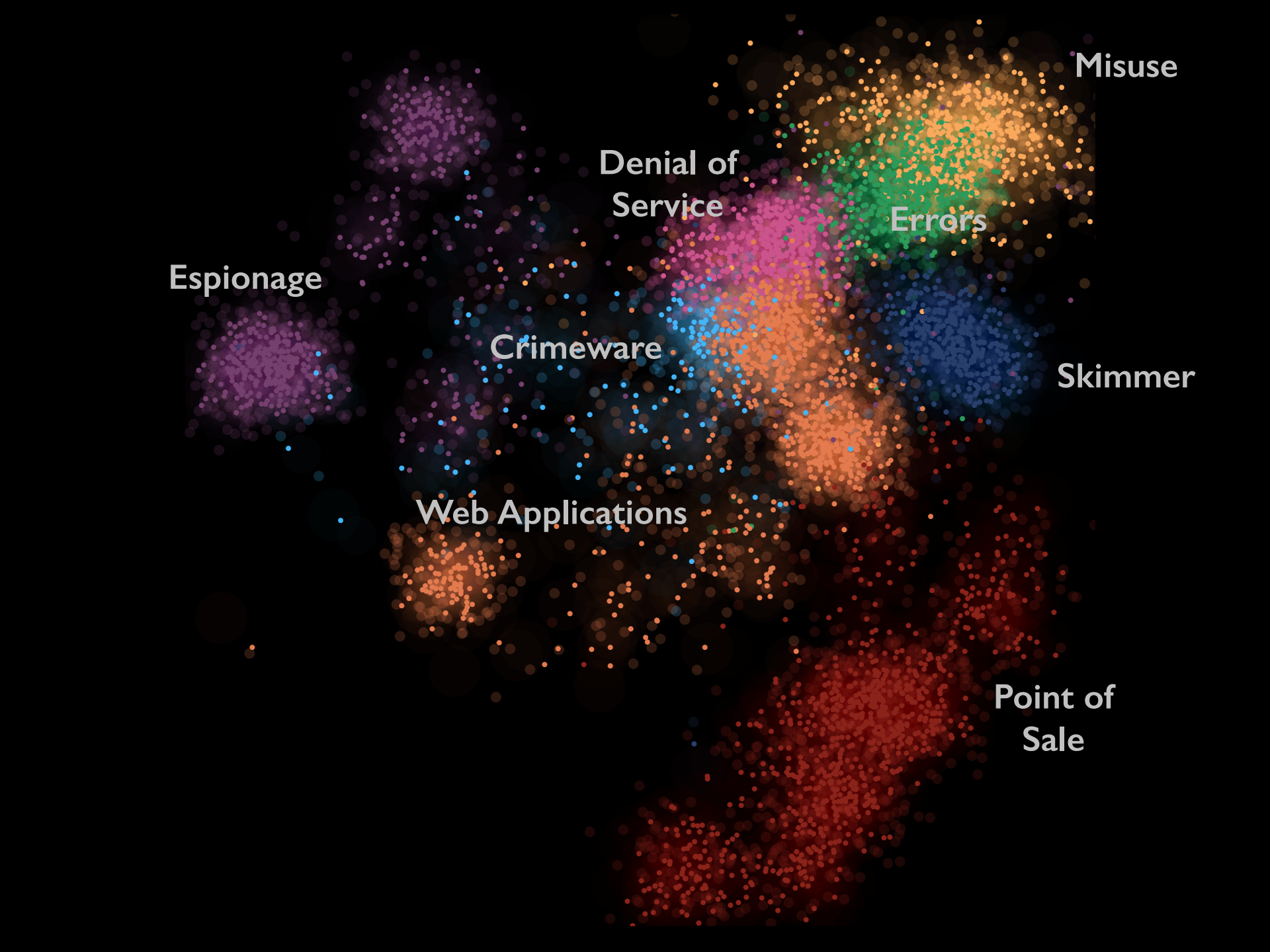
We can prioritize security controls

[with small data].

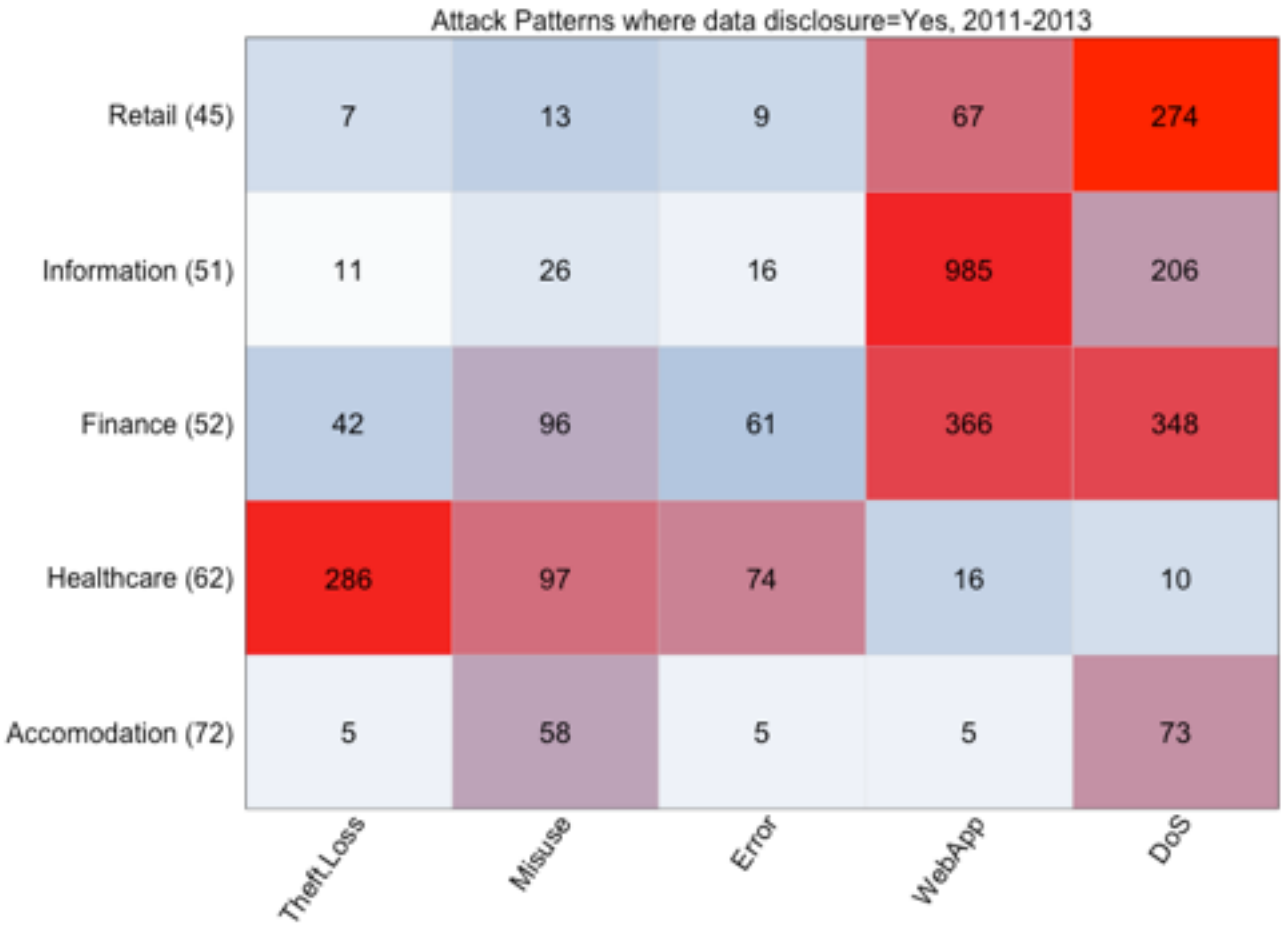
Research Question:

Should everyone prioritize the same controls?

(does a global "top 10" exist?)


Misuse
Denial of Service
Errors
Espionage
Crimeware
Skimmer
Web Applications
Point of Sale

One size fits all?



Pearson's Chi-squared test

X-squared = 2988.653, df = 16, p-value < 2.2e-16

INDUSTRY	POS INTRUS-ION	WEB APP ATTACK	INSIDER MISUSE	THEFT/ LOSS	MISC. ERROR	CRIME-WARE	PAYMENT CARD SKIMMER	DENIAL OF SERVICE	CYBER ESPION-AGE	EVERY-THING ELSE
Accommodation [72]	75%	1%	8%	1%	1%	1%	<1%	10%		4%
Administrative [56]		8%	27%	12%	43%	1%		1%	1%	7%
Construction [23]	7%		13%	13%	7%	33%			13%	13%
Education [61]	<1%	19%	8%	15%	20%	6%	<1%	6%	2%	22%
Entertainment [71]	7%	22%	10%	7%	12%	2%	2%	32%		5%
Finance [52]	<1%	27%	7%	3%	5%	4%	22%	26%	<1%	6%
Healthcare [62]	9%	3%	15%	46%	12%	3%	<1%	2%	<1%	10%
Information [51]	<1%	41%	1%	1%	1%	31%	<1%	9%	1%	16%
Management [55]		11%	6%	6%	6%		11%	44%	11%	6%
Manufacturing [31,32,33]		14%	8%	4%	2%	9%		24%	30%	9%
Mining [21]			25%	10%	5%	5%	5%	5%	40%	5%
Professional [54]	<1%	9%	6%	4%	3%	3%		37%	29%	8%
Public [92]		<1%	24%	19%	34%	21%		<1%	<1%	2%
Real Estate [53]		10%	37%	13%	20%	7%			3%	10%
Retail [44,45]	31%	10%	4%	2%	2%	2%	6%	33%	<1%	10%
Trade [42]	6%	30%	6%	6%	9%	9%	3%	3%		27%
Transportation [48,49]		15%	16%	7%	6%	15%	5%	3%	24%	8%
Utilities [22]		38%	3%	1%	2%	31%		14%	7%	3%
Other [81]	1%	29%	13%	13%	10%	3%		9%	6%	17%

For more information on the NAICS codes [shown above] visit: https://www.census.gov/cgi-bin/sssd/naics/naicsrch?chart=2012

Confirmed Data Breaches

INDUSTRY	POS	WEB APP	INSIDER/ PRIVLGE. MISUSE	LOST/ STOLEN	MISC. ERROR	CRIME-WARE	PAYMENT CARD SKIMMER	DENIAL OF SERVICE	CYBER ESPION-AGE	EVERY-THING ELSE
Accommodation [72]	85%	1%	8%	<1%	1%	1%	<1%	0%	0%	4%
Administrative [56]	0%	8%	29%	8%	47%	0%	0%	0%	1%	6%
Education [61]	1%	20%	10%	7%	29%	3%	1%	0%	2%	27%
Entertainment [71]	14%	38%	10%	0%	24%	0%	5%	5%	0%	5%
Finance [52]	<1%	42%	8%	1%	6%	2%	34%	0%	<1%	6%
Healthcare [62]	22%	3%	32%	12%	18%	<1%	<1%	0%	1%	12%
Information [51]	1%	46%	8%	3%	4%	4%	<1%	<1%	8%	25%
Manufacturing [31,32,33]	0%	23%	13%	2%	4%	1%	0%	0%	43%	14%
Mining [21]	0%	0%	33%	0%	7%	0%	7%	0%	53%	0%
Professional [54]	1%	16%	12%	2%	6%	3%	0%	0%	48%	12%
Public [92]	0%	12%	14%	3%	16%	13%	0%	0%	36%	6%
Real Estate [53]	0%	17%	44%	6%	22%	0%	0%	0%	6%	6%
Retail [44,45]	54%	17%	6%	1%	1%	3%	10%	0%	<1%	7%
Trade [42]	8%	31%	8%	4%	12%	8%	4%	0%	0%	27%
Transportation [48,49]	0%	11%	23%	8%	5%	2%	6%	0%	39%	6%
Utilities [22]	0%	72%	4%	1%	1%	3%	0%	0%	12%	7%
Other [81]	2%	28%	22%	0%	17%	3%	0%	0%	7%	22%

Updating the research question…

Since attacks across industries vary…

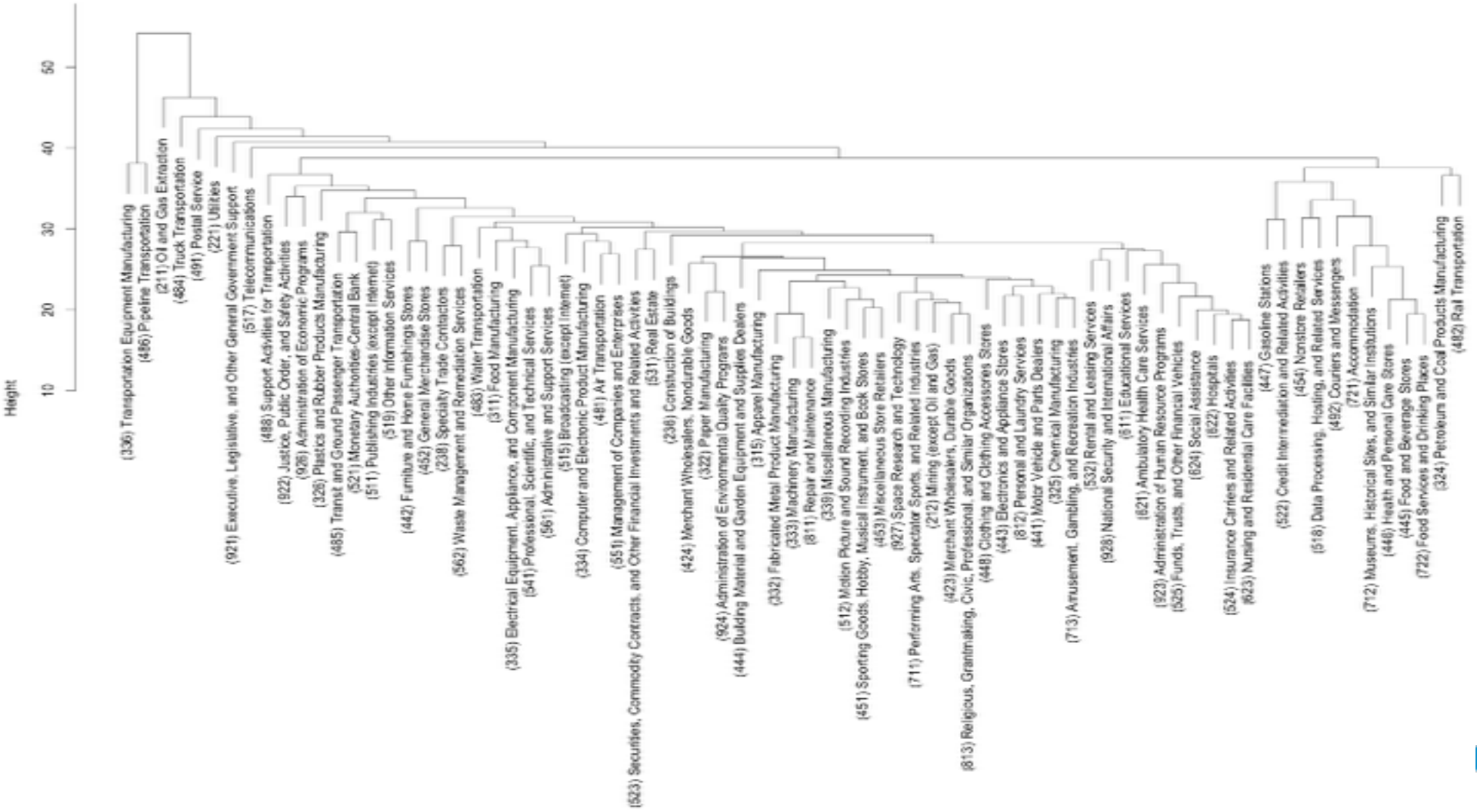
Which industries exhibit similar threat profiles?

Research Question:

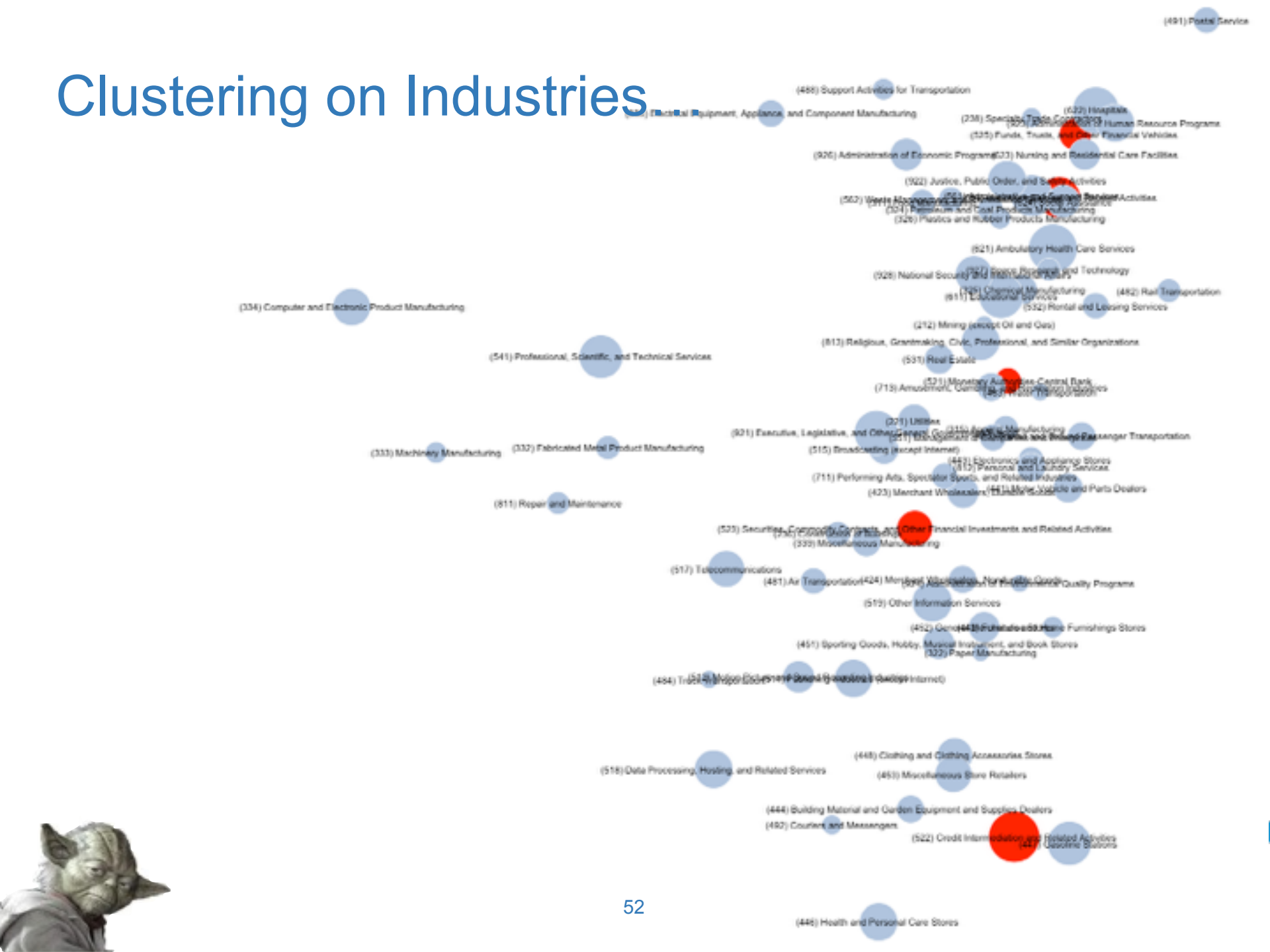
Should everyone prioritize the same controls?

(does a global "top 10" exist?)

Clustering on Industries…



Clustering on Industries...



Key Takeaways from these questions…

- Adversaries are not random
  - Controls should be better than random
- Victims are not random
  - “Everyone should…” is a thing of the past

Doing data analysis?

Setting aside the challenges of “big data”, we still need to follow a few basic steps:

- Formulate a good research question
- Identify data needed to answer it
- Analyze data using appropriate methods
- Honestly report findings to help others

Where to next?

Massively Open Online Courses (MOOCs)

- **Coursera's Introduction to Data Science** course (https://www.coursera.org/course/datasci)
- **edX's Learning From Data** course (https://www.edx.org/course/caltechx/cs1156x/learning-data/1120)
- **Syracuse University's Data Science Open Online** course (http://ischool.syr.edu/future/cas/introtodatasciencemooc.aspx)

Online certificate or master's courses:

- **UC Berkeley's MIDS program** (http://www .ischool.berkeley.edu/programs/mids)
- **University of Washington's certificate in data science** (http://www.pce.uw.edu/certificates/data-science.html)
- **Penn State's Applied Statistics** online curriculum (http://www.worldcampus.psu .edu/degrees-and-certificates/applied-statistics-certificate/ overview)

http://datadrivensecurity.info/blog/pages/resources.html

RSA CONFERENCE 2014
ASIA PACIFIC & JAPAN

Jay Jacobs

Security Data Scientist

Cybersecurity Research & Innovation

Verizon

jay.jacobs@verizon.com

DBIR@verizon.com

twitter: @jayjacobs