AIR ACADEMY ASSOCIATES

Simplify, Perfect, Innovate

Big Data, Predictive Analytics, and Security

ITEA Annual Symposium
4 October 2016
Reston, VA

16-BIGPASEC-10A

Mark J. Kiemele, Ph.D.
President and Co-Founder
Air Academy Associates

Office: 719-531-0777
Cell: 719-337-0357
mkiemele@airacad.com
www.airacad.com

Goals

- **What is Big Data?**
- **Why is it so popular?**
- **Big Data and Predictive Analytics**
- **Its relationship to Security Intelligence**
  - Securing Big Data
  - Analyzing Big Data
- **Resources Needed**

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

A Definition of Big Data

Volume

This is the most common definition associated with Big Data.

1 Bit = Binary Digit
8 Bits = 1 Byte
1000 Bytes = 1 Kilobyte
1000 Kilobytes = 1 Megabyte
1000 Megabytes = 1 Gigabyte
1000 Gigabytes = 1 Terabyte
1000 Terabytes = 1 Petabyte
1000 Petabytes = 1 Exabyte
1000 Exabytes = 1 Zettabyte
1000 Zettabytes = 1 Yottabyte
1000 Yottabytes = 1 Brontobyte
1000 Brontobytes = 1 Geopbyte

Big Data is about the storage and analysis of large sets of data.

A Definition of Big Data

Velocity

Data streams in at unprecedented speeds and must be dealt with in a timely manner.

RFID tags, sensors, and smart metering are driving the need to deal with huge amounts of data in near-real time.

Currently available Big Data technologies deal with huge sets of data at rest in batch-oriented jobs - run a query, analyze results, tweak query, analyze again, and repeat.

This is not the streaming scenario in which a constantly updated tactical situation is analyzed.

Future systems need to handle real-time analytics, especially in security scenarios.

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

A Definition of Big Data

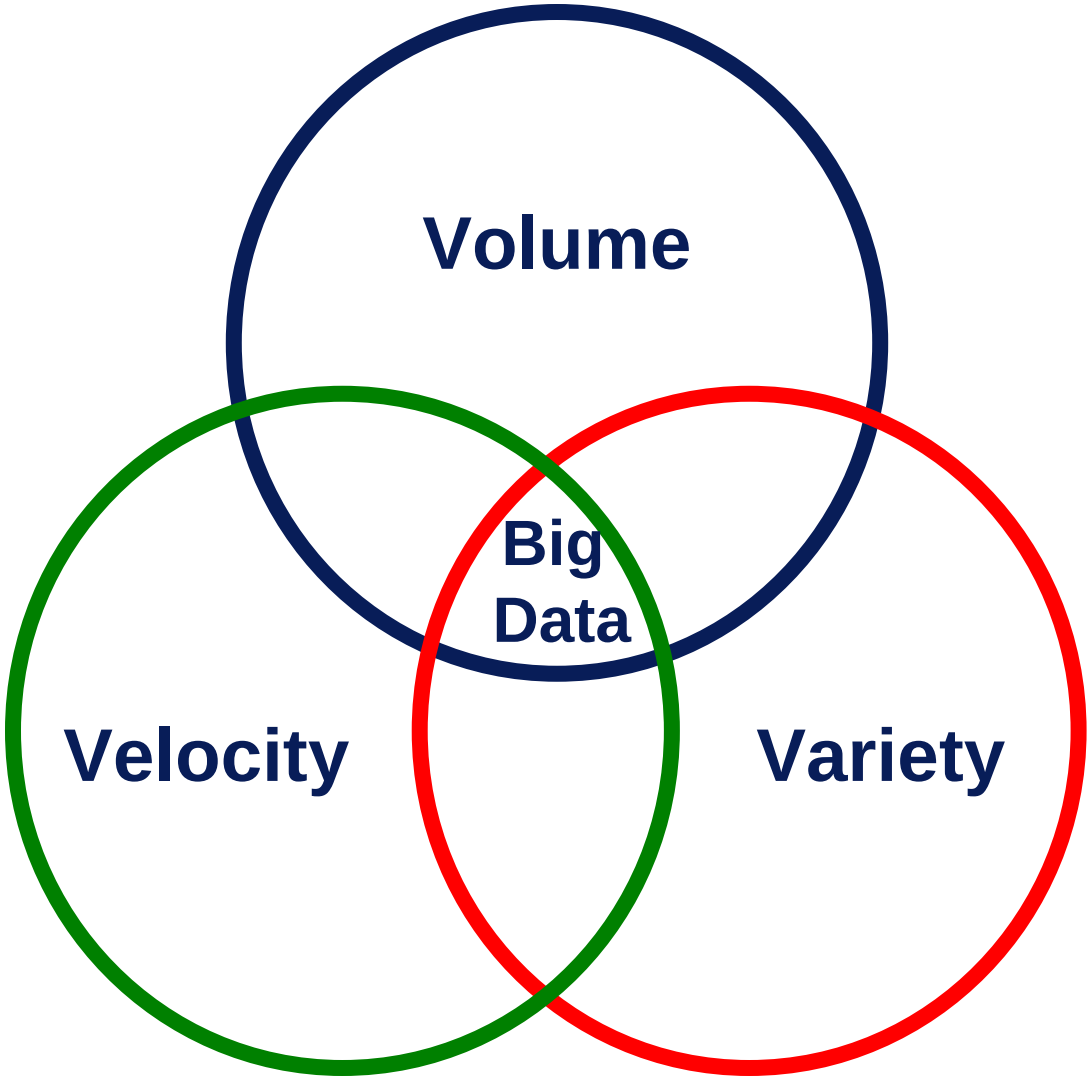
Variety

Data comes in all types of formats

- **structured, numeric data in traditional databases with nice rows and columns, namely relational databases like SQL**
- **unstructured text documents in multiple languages**
- **email, video, audio, tweets**
- **stock ticker data, financial transactions**

Industry (3V) Definition of Big Data*

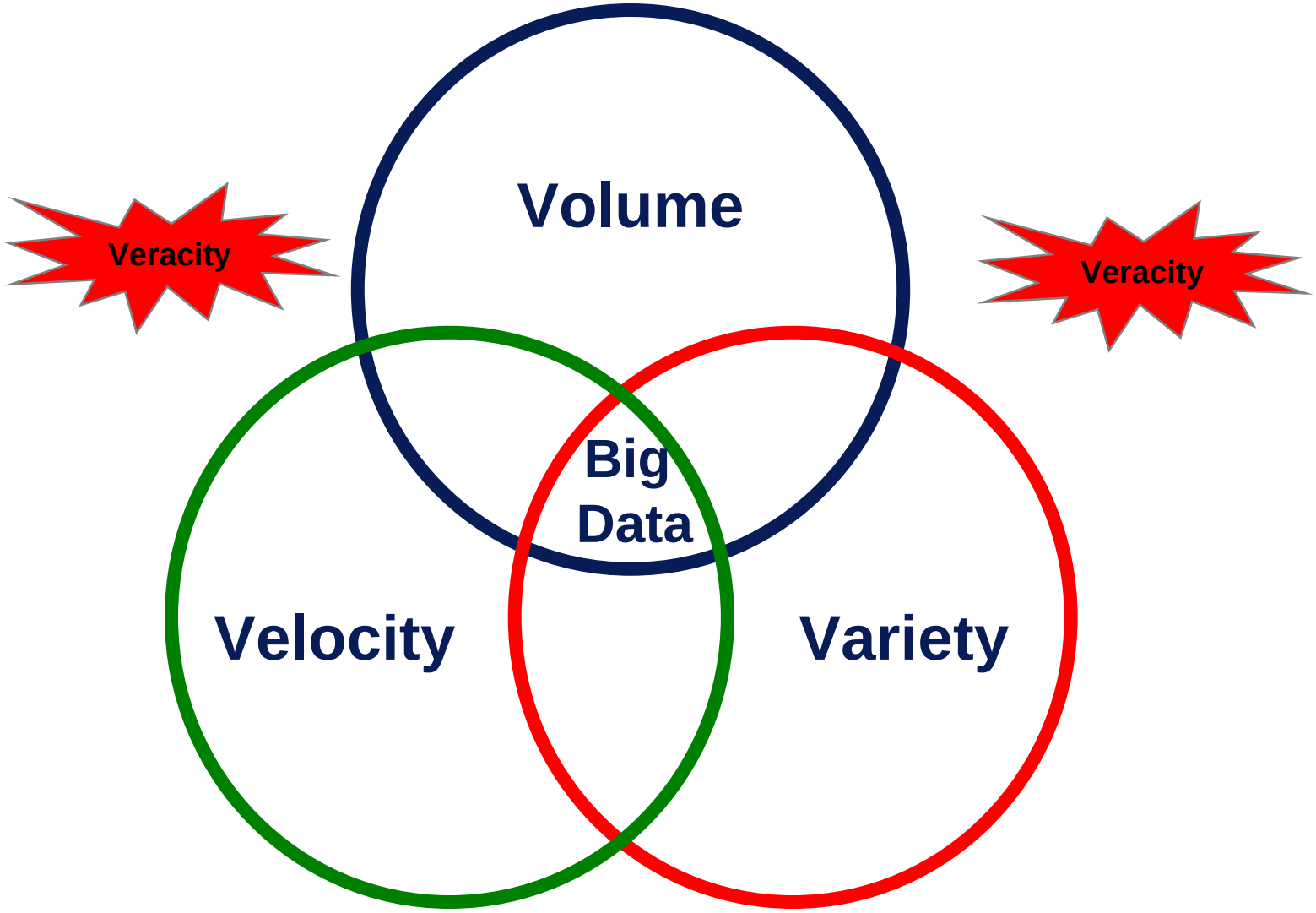
(*source: Doug Laney)



Air Academy Associates

Simplify, Perfect, Innovate

4V Definition of Big Data

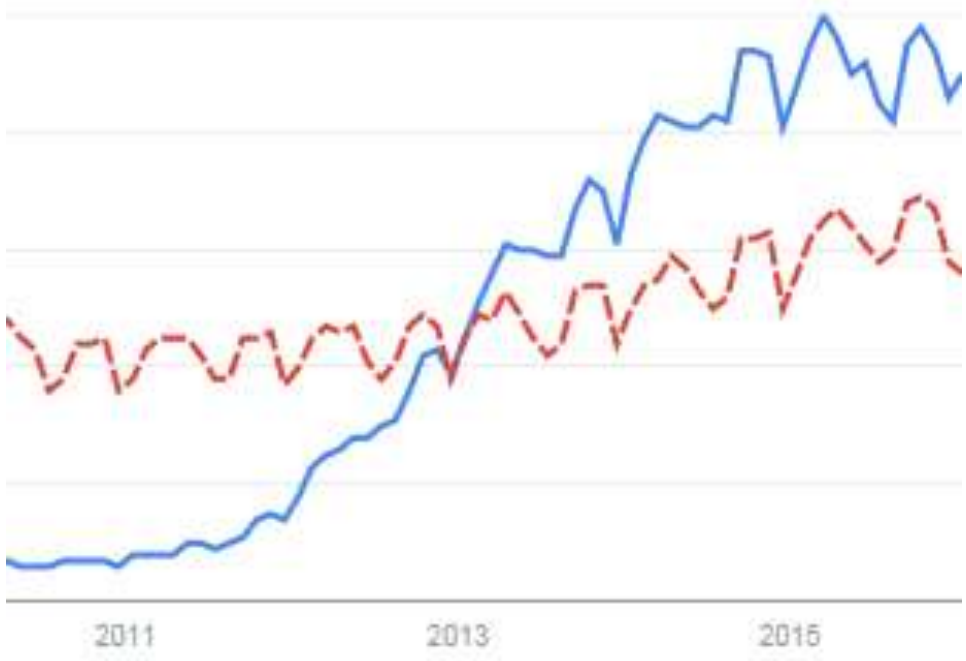


Air Academy Associates
Simplify, Perfect, Innovate

Simple Definition

We have “Big” data when the volume/velocity/variety/veracity of the data becomes part of the problem we are trying to solve.

The Popularity of Big Data from Google Trends



Big Data

Data Analysis

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

Predictive Analytics

“It’s tough to make predictions –

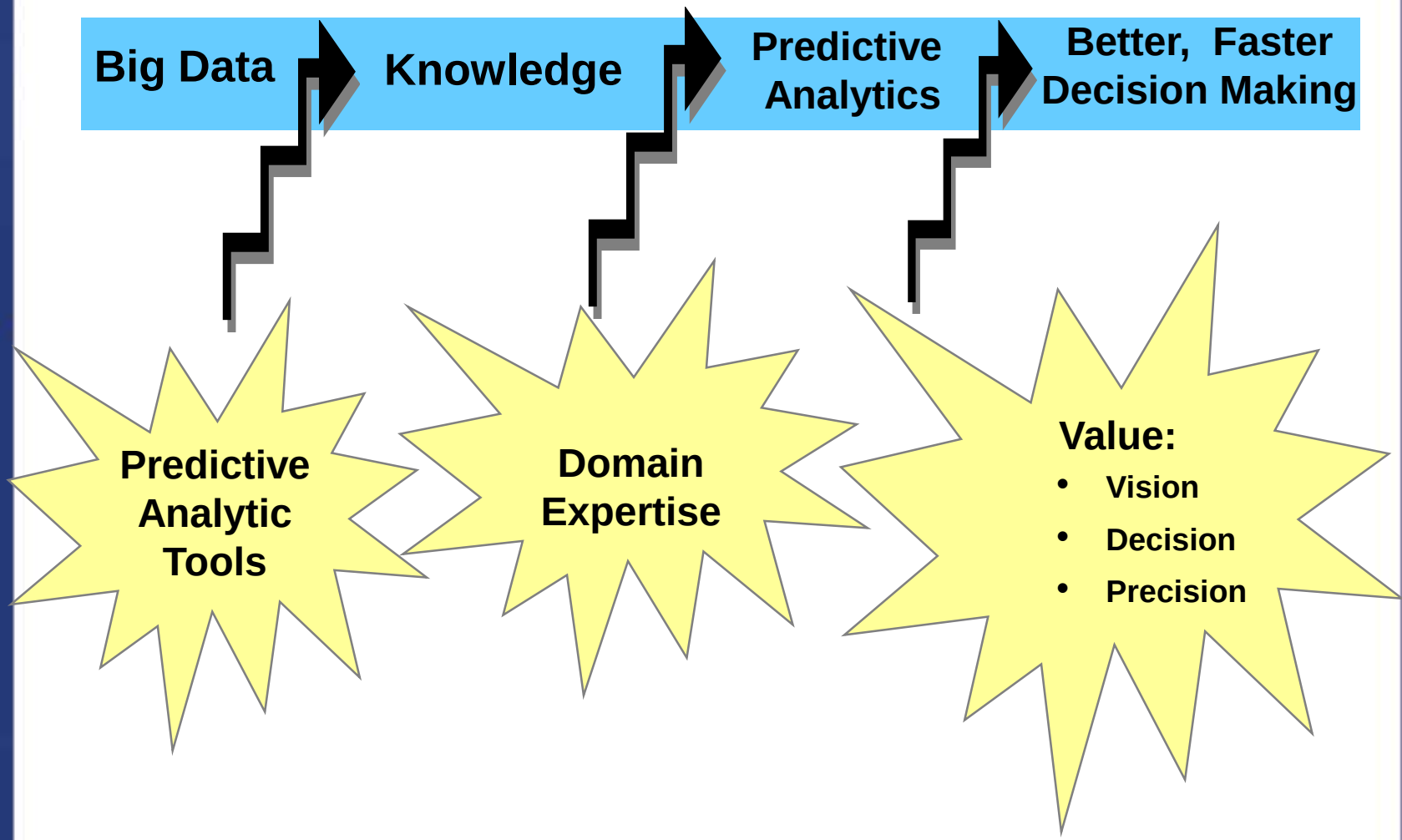
especially about the future.”

Yogi Berra

AIR ACADEMY ASSOCIATES

Simplify, Perfect, Innovate

Big Data + Predictive Analytics = Value



Air Academy Associates
Simplify, Perfect, Innovate

Industries Where Predictive Models are Used

(Source: *Predictive Analytics* by Eric Siegel)

- Marketing, Advertising, and the Web
- Family and Personal Life
- Financial Risk and Insurance
- Healthcare, Medical, Pharmaceutical
- Crime Fighting and Fraud Detection
- Fault Detection for Safety and Efficiency
- Government, Politics, Nonprofit, and Education
- Human Language Understanding, Thought, and Psychology
- Staff and Employees – Human Resources
- Defense
- Engineering
- Oil, Gas, Energy

Examples of What is Being Predicted

(Source: *Predictive Analytics* by Eric Siegel)

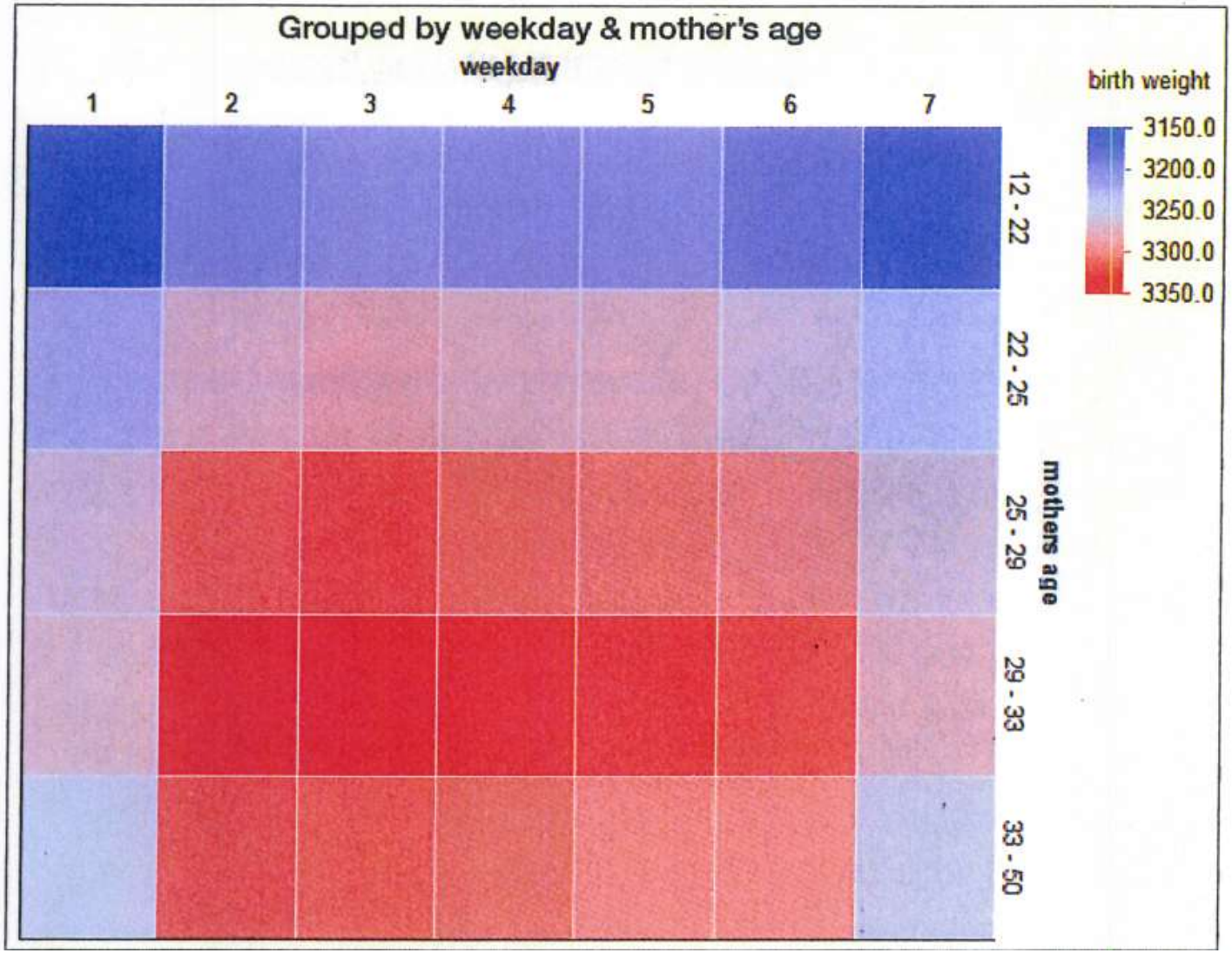
- Likelihood of elderly dying within the next 18 months
- Target predicts customer pregnancy from shopping behavior
- Love (Match.com predicts interest in communicating)
- Breast cancer; premature births and other health related metrics
- Fraudulent tax returns, insurance and warranty claims
- Terrorist attacks
- Online activities that are malicious intrusions and attacks vs. legitimate activities
- Nuclear reactor failures (e.g., cracks in cooling pipes)
- Which voters are positively influenced by what type of contact and which voters will be affected adversely by contact
- Electricity demand growth to direct infrastructure development
- Lying and deception via eye movement and written statements
- Employee churn, job performance, and student dropouts

Cautions Concerning Big Data

- The myth that more data is always better.
- The myth that big data makes up for the lack of quality in the data.
- The low cost of data storage allows us to stash away troves of data, not knowing if we will ever need it or use it.
- There is much more noise in Big Data, making it more difficult to separate the true signal from the noise.
- Big Data contains all kinds of spurious correlations.
- While correlation may help us predict, correlation does not mean causality.

Relationship Between 3 Variables

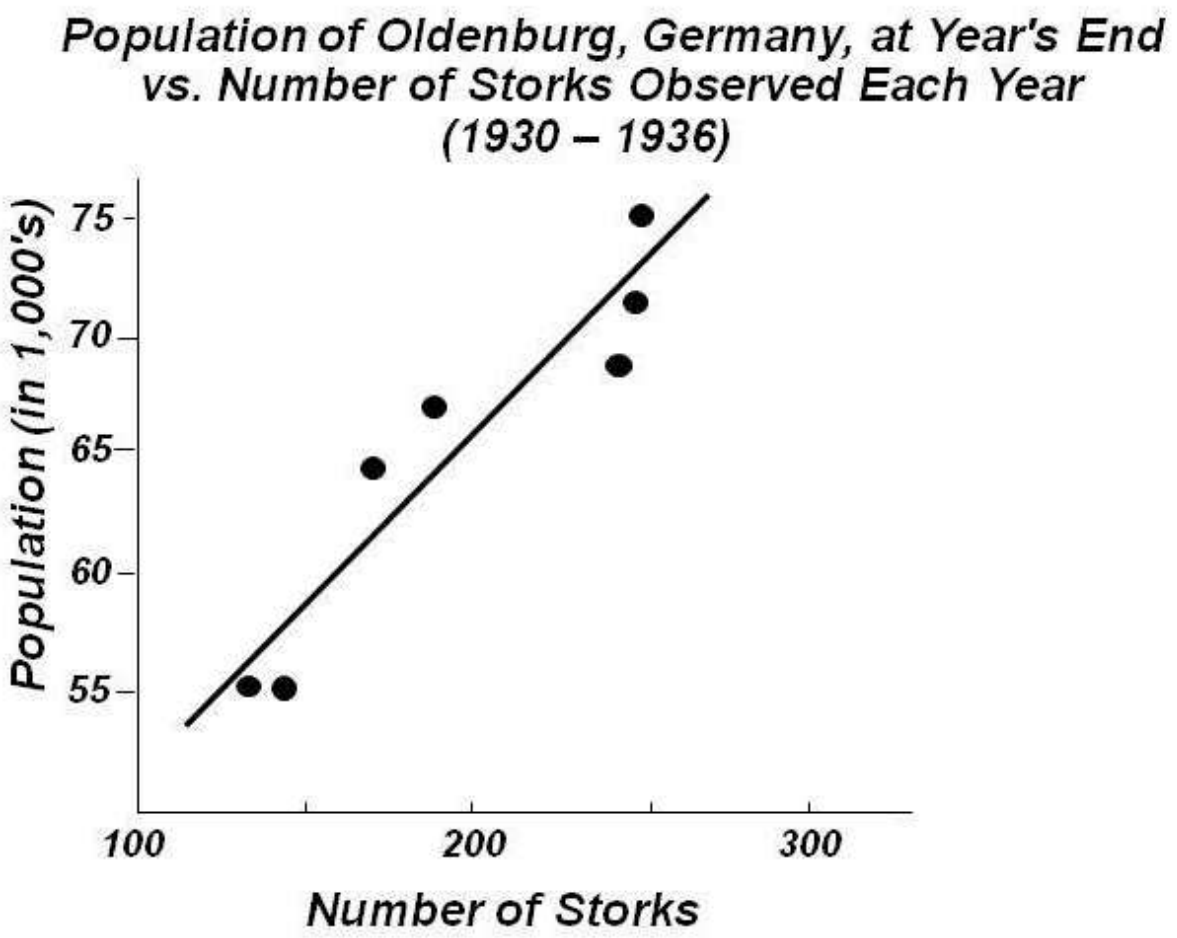
(from more than 12 million records – these are average weights)



Air Academy Associates

Simplify, Perfect, Innovate

Correlation vs Causality



Source: Statistics for Experimenters, by Box, Hunter & Hunter

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

Regression Analysis

(helps separate the signal from the noise)

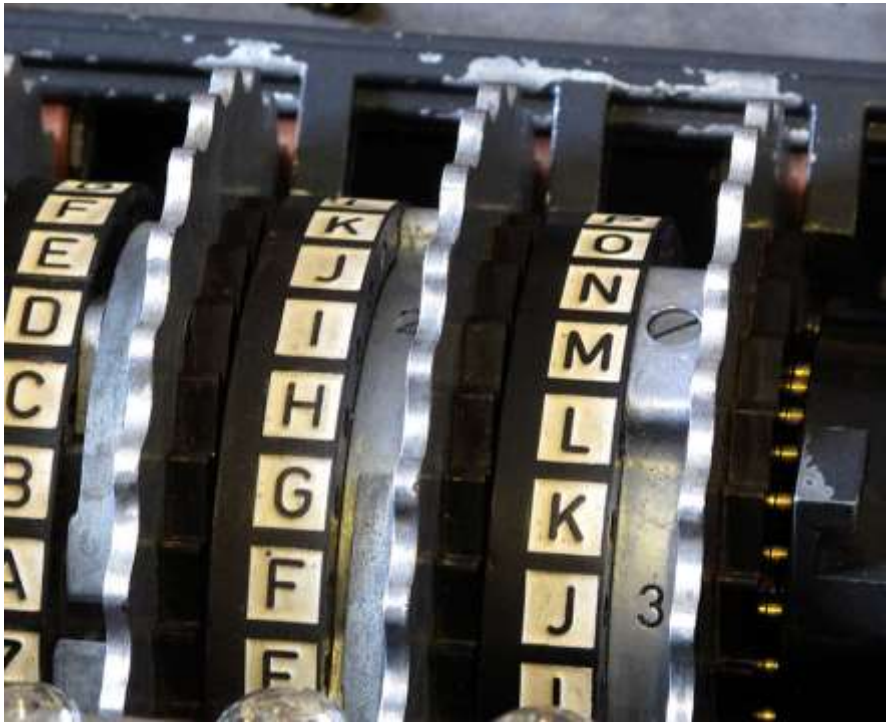
	Const	-2.9692E+11	6.3721E+11	-0.466	0.641	
A	**A (A)**	-29.368	12.928	-2.2717	**0.023**	397.942
B	**B (B)**	-2.9692E+11	6.3721E+11	-0.466	0.641	1.34846E+12
C	**C (C)**	-2.9692E+11	6.3721E+11	-0.466	0.641	5473164212
D	**D (D)**	41.481	9.7328	4.262	**0.000**	256.513
E	**E (E)**	28.369	10.187	2.7847	**0.005**	279.972
F	**F (F)**	-20.831	19.369	-1.0755	0.282	250.013
G	**G (G)**	13.546	29.27	0.4628	0.644	591.811
H	**H (H)**	10.842	10.51	1.0316	0.302	167.948
A*B	**AB**	-1.3704	0.6779	-2.0214	**0.043**	1.7618
A*C	**AC**	-18.573	12.675	-1.4654	0.143	383.844
A*D	**AD**	2.8617	1.0707	2.6727	**0.008**	4.2238
A*E	**AE**	0.7311	0.9422	0.7759	0.438	3.2029
A*F	**AF**	-9.7751	2.4039	-4.0663	**0.000**	15.814
A*G	**AG**	-2.8441	1.6404	-1.7338	**0.083**	7.2009
A*H	**AH**	-1.1715	1.2294	-0.9529	0.341	4.6646
B*C	**BC**	-2.9692E+11	6.3721E+11	-0.466	0.641	1.34767E+12
B*D	**BD**	6.3876	0.7489	8.5288	**0.000**	2.2344
B*E	**BE**	4.1134	0.7205	5.7087	**0.000**	2.074
B*F	**BF**	-4.6692	1.7137	-2.7247	**0.006**	11.695
B*G	**BG**	3.9336	1.6159	2.4343	**0.015**	10.084
B*H	**BH**	1.863	0.9411	1.9795	**0.048**	3.5352
C*D	**CD**	2.9397	8.9865	0.3271	0.744	218.94
C*E	**CE**	12.799	9.5949	1.3339	0.182	248.612
C*F	**CF**	-46.066	18.063	-2.5503	**0.011**	221.518
C*G	**CG**	-55.855	28.506	-1.9594	**0.050**	572.564
C*H	**CH**	-22.526	10.168	-2.2155	**0.027**	157.886
D*E	**DE**	-1.3841	0.718	-1.9277	**0.054**	1.5218
D*F	**DF**	4.191	1.487	2.8184	**0.005**	5.5047
D*G	**DG**	11.385	3.4105	3.3384	**0.001**	35.482
D*H	**DH**	1.9595	0.9618	2.0373	**0.042**	2.728
E*F	**EF**	2.0255	1.3101	1.5461	0.122	4.5699
E*G	**EG**	4.2577	3.139	1.3564	0.175	29.918
E*H	**EH**	-0.7377	0.9335	-0.7903	0.429	2.5299
F*G	**FG**	1.0982	7.0554	0.1557	0.876	63.947
F*H	**FH**	4.9165	2.2499	2.1852	**0.029**	9.878
G*H	**GH**	4.2302	1.7185	2.4616	**0.014**	5.5814

AIR ACADEMY ASSOCIATES

Simplify, Perfect, Innovate

Big Data and the Security Issue

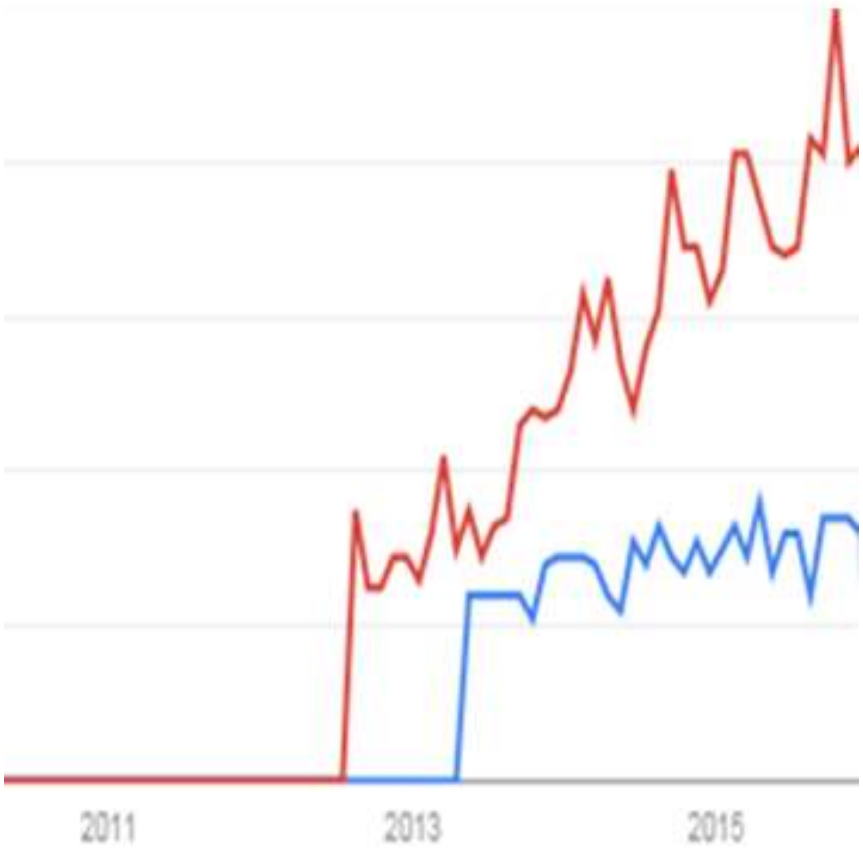
- **Securing the data itself**



- **Using Big Data Analytics to predict and prevent security incidents**

Air Academy Associates
Simplify, Perfect, Innovate

Google Trends



Big Data Analysis

Big Data Security

Air Academy Associates

Simplify, Perfect, Innovate

Big Data and the Security Issue

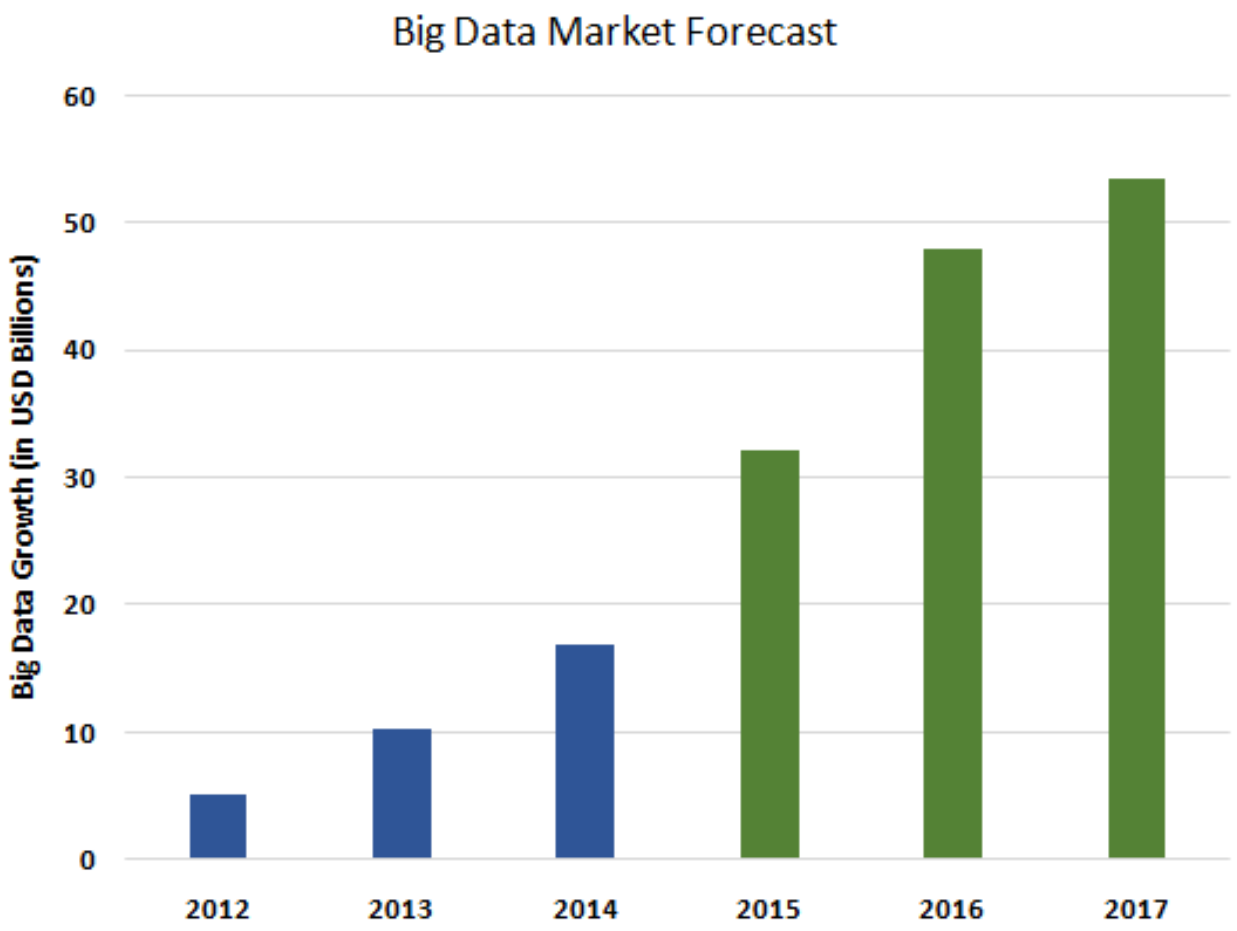
- **Securing the data itself**
  - Data security must be designed into the system.
  - Few organizations are likely to build a Big Data environment in-house.
  - Most will use the cloud to store the data.
  - Encryption and access controls can help.
  - From Arthur Coviello, EVP for EMC's RSA Security Division:

"With the pervasiveness of big data touching everything we do, our attack surface is about to be altered and expanded, and our risks magnified in ways we couldn't have imagined."

Big Data and the Security Intelligence Challenge

- **Using Big Data Analytics to predict and prevent security incidents**
  - Big Data expands the boundaries of existing information security responsibilities and introduces significant new risks and challenges.
  - Big Data silos prevent complete analysis.
  - Cybersecurity scenarios require analysis of massive amounts of data as it's produced, with conclusions reached within seconds.
  - New algorithms are needed.
  - Must integrate the sciences of statistics, mathematics, and computer science – this is often called Data Science.

Forecasting Big Data's Spending



AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

The Big Data Analytics Skills Shortage

- According to McKinsey Global Institute, there will be more than a 200,000 job shortage in data science by 2018.
- Data science is a multi-disciplinary activity involving statistics, applied mathematics, computer science, engineering, economics, and other related fields.
- Doing predictive analytics today is best done by a team.
- Our university system, in spite of its recent increased emphasis via degree programs on data analytics, cannot keep up with the demand for data scientists.
- Data Analytics applies to small data as well as big data.
- If one cannot do analysis on small data, it is highly likely he won't be able to handle big data either.
- According to Nate Silver, from the FiveThirtyEight blog and grand prognosticator of the 2008 and 2012 presidential elections, "I think data scientist is a sexed-up term for a statistician."

Some Final Thoughts

- There has been plenty of hype on the subject of Big Data.
- When it comes to security intelligence, we are going to need Big Data.
  - For the ability to analyze data in near or real time for the purpose of predicting and preventing cyber attacks; fraud, waste, and abuse; and terrorist attacks, just to name a few.
  - And thus there is the need for getting good data, including the right data, as well as protecting and securing it.
- There is a shortage of properly trained resources to handle big data analytics.
- We need an increased emphasis on developing those resources, as better technology and prediction algorithms than what we have today are needed.

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

Thank You

Colorado Springs, Colorado

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate

For More Information, Please Contact

Air Academy Associates, LLC
1650 Telstar Drive, Ste 110
Colorado Springs, CO 80920

Toll Free: (800) 748-1277 or (719) 531-0777
Facsimile: (719) 531-0778
Email: aaa@airacad.com
Website: www.airacad.com

AIR ACADEMY ASSOCIATES
Simplify, Perfect, Innovate