



Presented to:
CBM+ RCM WPIT

CBM and Big Data



TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

Date

Presented by:

Doug Felker

Title

**U.S. Army Aviation and Missile Research,
Development, and Engineering Center**



Big Data



Predictive Analytics

Big Data

Data warehousing

Automated Analytics Tools

Walmart

Amazon

eBay

Intel

Apple

gigabytes

petabyte

exabyte

Depends on Definition

Big Data = Lots of Data
- Certainly YES

Big Data = Lots of Data used for decision making
- Certainly YES

Big Data = Lots of Data which is analyzed for patterns. These Patterns are presented to Decision Makers (People) to affect behavior productively.
- Yes (but do to sample size, do not oversell)

- **Wal-Mart shoppers stocking up for a hurricane often buy strawberry Pop-tarts.**

Big Data = Lots of Data which is analyzed for patterns. When a particular pattern occurs, automatic (no people) actions are triggered.
- Be very careful

- **Computerized Stock trading**
- **Amazon sends out “Customers who bought that book have also...”**

Characterize the impact of an input upon a system by:

1. Understanding the Cause and Effect of the Input (Engineering)
2. Observing a pattern associated with the application of the Input (Data Analysis)

Data Analysis looks for patterns of outcomes based on the inputs.

- Build a model to predict outcome

SAMPLE SIZE (not amount of data) determines the limit of the Complexity of the model (how many inputs your model can use)

One Sample = One Scenario of inputs

Lots of data from **One Sample** = Detailed understanding of One Scenario

A life insurance company is selling policies to 20 year olds. In order to determine, the premium, the company has to determine life expectancy.

In order to accurately predict life expectancy, company collects information (inputs) from each policy applicant.

Inputs:

Gender	Family history	Smoking habits	Body fat
Cholesterol	Occupation	Diseases	Ethnicity

Actuary science has collected data from tens of millions of people (samples).

Data on the impact of various combinations of inputs can predict a better life expectancy range than an average

Millions of Samples allow a complex model with many inputs

A life insurance policy is selling policies to 20 year olds. In order to determine, the premium, the company has to determine life expectancy.

For this example, lets assume there is only data from 700 people regarding how various inputs impact life expectancy.

Inputs:

Gender	Family history	Smoking habits	Body fat
Cholesterol	Occupation	Diseases	Ethnicity

The company could make limited predictions from maybe 1 or 2 inputs

With a sample size of 700, one obese, chain-smoking, 90 year old race car driver would skew a marginally complex model and it would give false predictions.

With millions in the sample size, such an individual would be expected.

Example

700 Detailed Samples

Gather (mine) lots of detailed data from the 700 data sample

Inputs:

Gender

Family history (health history of four generations)

Smoking habits (number and brand smoked per day)

Body fat (detailed BWI from every year of life)

Cholesterol (LDL, HDL, TCL, NDHL)

Occupation (Hours worked, Temperature in workplace)

Diseases (Earaches when a kid)

Ethnicity (history of four generations)

With a sample size of 700, one obese, chain-smoking, 90 year old race car driver would skew a marginally complex model and it would give false predictions.

**No matter how precise the color of the television
Picture quality is limited by the number of Pixels**

Is DoD Big Data?

Answer -

Yes, (but we are not Amazon)

Many of the most successful applications of SUCCESSFUL applications of Actionable Analytics (Big Data) is based upon point-of-purchase transactions (millions per day)

Number of Apaches = 700

How many inputs can the Army use for Apache Engineering Analysis?

Answer -

As many as we can. We want all the information we can to understand the Apache

How many inputs can the Army use for Apache Actionable Analytics?

Answer -

It depends (input signal to noise ratio) but not a lot.

Big Data and Sample Size



