



Data Analytics

Dr. Laura Armey

17 March 2021



Defense Resources Management Institute
Naval Postgraduate School
Monterey, California



Learning Objectives



- Be able to frame questions in data-analytic terms.
- Gain awareness of what it takes to answer these questions.
- Gain some insights into reasons to be skeptical of analytic answers.



Data analytics is having A moment – what is it?

Data analytics is also known as data analysis

<https://www.techopedia.com/definition/26418/data-analytics>

Data analytics refers to qualitative and quantitative techniques and processes used to enhance productivity and business gain. Data is extracted and categorized to identify and analyze behavioral data and patterns, and techniques vary according to organizational requirements.

<https://www.techopedia.com/definition/26418/data-analytics>

Data analytics is the science of drawing insights from raw information sources. <https://www.investopedia.com/terms/d/data-analytics.asp>

Data analytics (DA) is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software.

<https://searchdatamanagement.techtarget.com/definition/data-analytics>



Data analytics is having a moment – what is it?



Data analytics is also known as data analysis

<https://www.techopedia.com/definition/26418/data-analytics>

The process of examining data in order to identify and understand patterns to enhance productivity and improve financial management decisions.

nd
ents.

Data analytics is the science of drawing insights from raw information sources. <https://www.investopedia.com/terms/d/data-analytics.asp>

Data analytics (DA) is the process of examining **data** sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software.

<https://searchdatamanagement.techtarget.com/definition/data-analytics>

What exactly is Big Data?

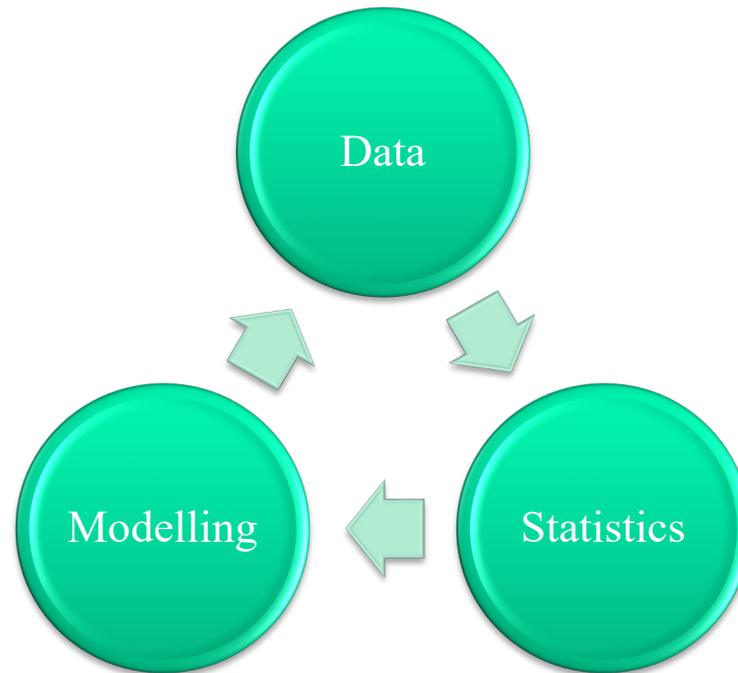


Big Data

Oxford Dictionary: Extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions

Gartner (2001): Big data is data that contains greater variety arriving in increasing volumes and with ever-higher velocity

Data analytics is about more than just data



What does Data Analytics have to do with Resource and Financial Management?



Allows us to answer important questions.

What are the implications of policy changes?

How should we plan for the future?

Where should we focus resources to improve outcomes?



How do we do it?



Make choices about models

- The right models for the right kinds of questions

Understand the data we have

- Consider the limitations of data, and the challenges data poses in answering particular types of questions.



Statistics....eeeeewwwwww!



Statistics is about measuring variation and finding patterns in data...

- Describe – you can do pretty well at this with basic statistics
- Inference – draw conclusions – probability distributions and hypothesis testing
- Correlation – drivers behind the patterns

From basic statistics to analytic answers



If you start with these questions

- Are we paying vendors on time?
- Are we posting receivables and disbursements in a timely manner?
- What is the utilization of DoD's real property to help reduce lease costs?

You could start to analyze these questions by asking the related questions:

- What is the mean/median time for payment of a vendor?
- How spread out is our time to post receivables and disbursements?
- What are the most underused 10% of DoD's real property?

Inference



Inference: the theory, methods, and practice of forming judgments about the parameters of a population and the reliability of statistical relationships, typically on the basis of random sampling.

- Are we paying vendors on time?
 - How frequently would we expect payments to be late by x days?
 - Are some organizations paying their vendors later than others?
- Are we posting receivables and disbursements timely?
 - Are some categories of receivables or disbursements more timely than others?
- What is the utilization of DoD's real property to help reduce lease costs?
 - How often should we expect a particular degree of underutilization?
 - Are some installations/organizations utilizing their real property more or less than others?

What kind of question are you asking?



Descriptive Analytics

- Describes what is/what was.

Predictive Analytics

- Predicts what is most likely to happen in the future.

Prescriptive Analytics

- What should we do about it?
- Recommends actions you can take to affect those outcomes.
- Computer optimized course of action.



What are you trying to do?



Understand the impact of one factor on another – a treatment effect.
(Also called program evaluation or impact analysis)

- If we add a week onto our training program how much better will our students perform on their exams?
- If we spend more on signing bonuses, how many high-quality recruits will we get?
- Did sequestration negatively impact retention in DoD?
- Would changing our accounting system positively impact our ability to pay disbursements in a timely way?
- If we raise copays on prescription drugs, how much will we save on prescription benefits?

What are you trying to do?



Predict future outcomes

- If we make a tank heavier how much more will it cost to build?
- If we make a tank heavier how much more fuel will it consume?
- Can we predict how much money will be left over at the end of the year?

The simplest model



- Linear Regression can help us do either predict future outcomes or model a treatment effect.
- It is simple: you can do it in excel, and it remains a powerful and simple analytic tool.

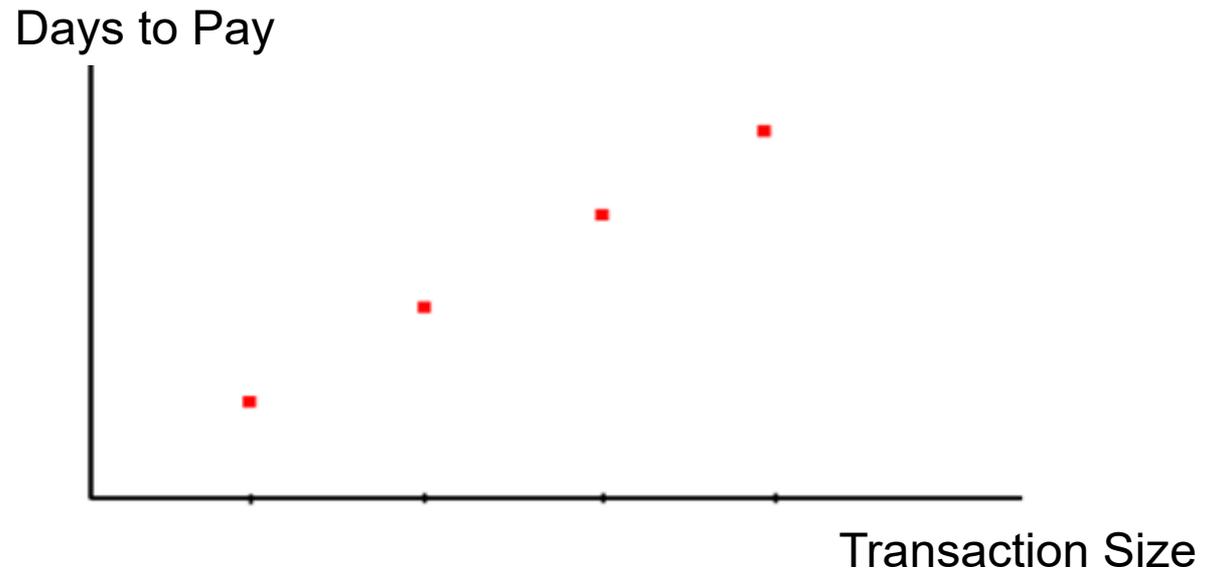
An example



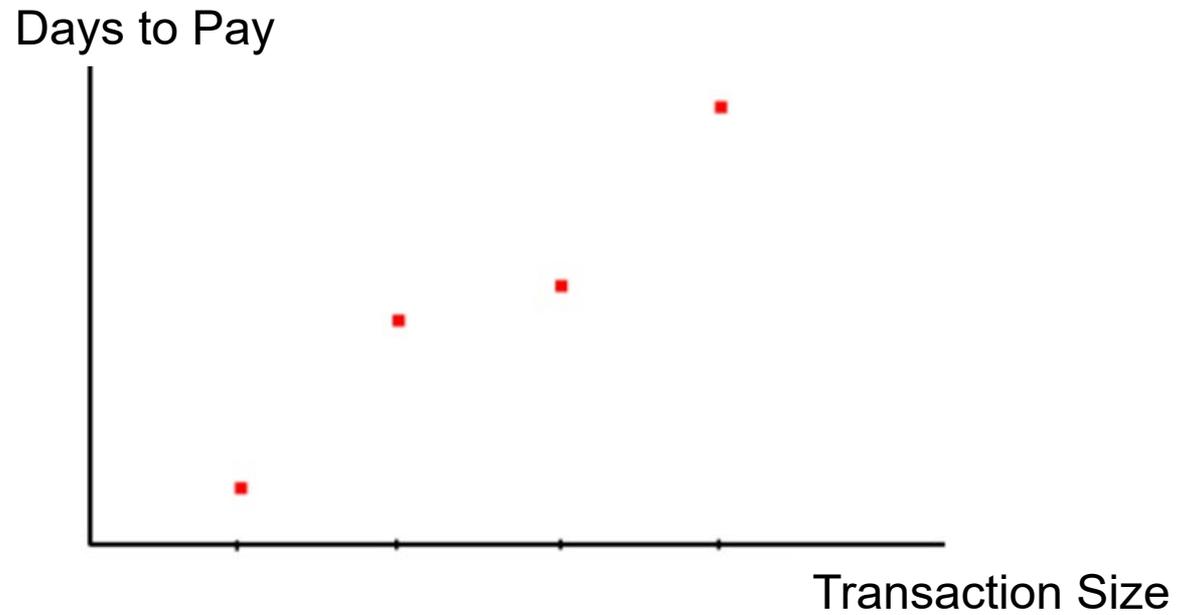
Let's say we want to figure out what drives differences in the number of days it takes to pay vendors.

- Is it the type of transaction or the organization doing the paying?
- Perhaps the size of the transaction could impact how long it takes to pay.

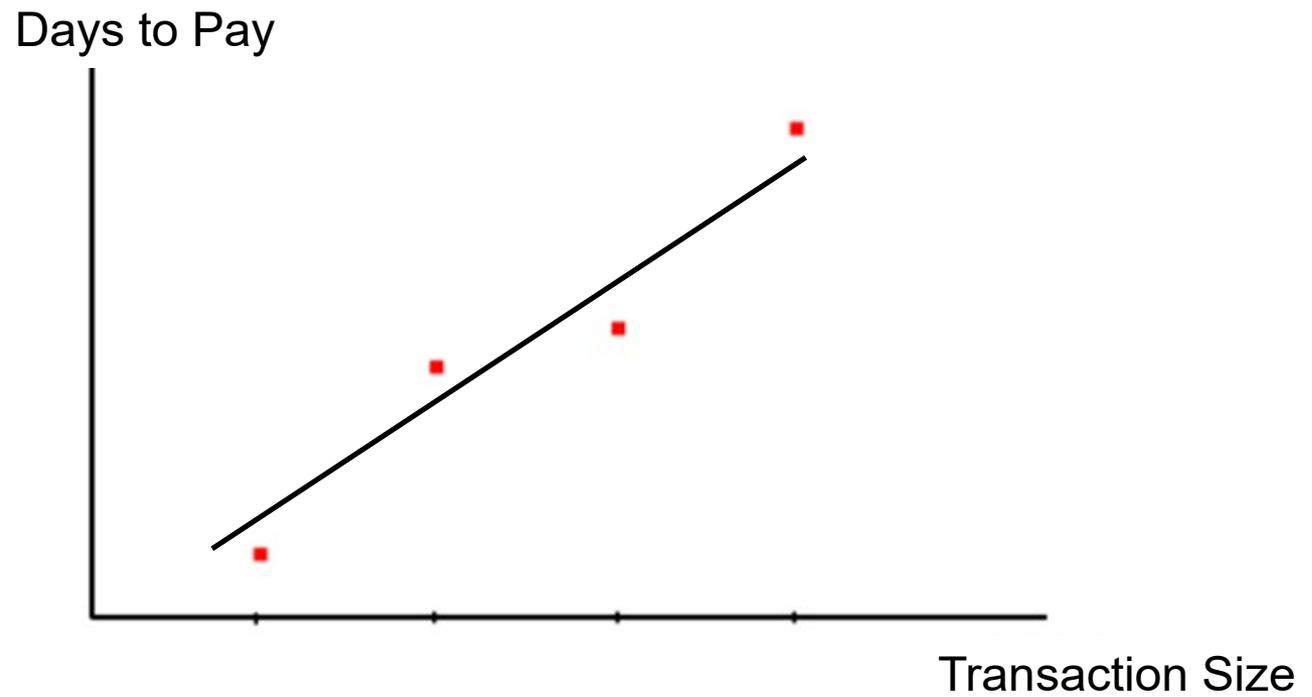
Perfect world relationships



Real Life Relationships



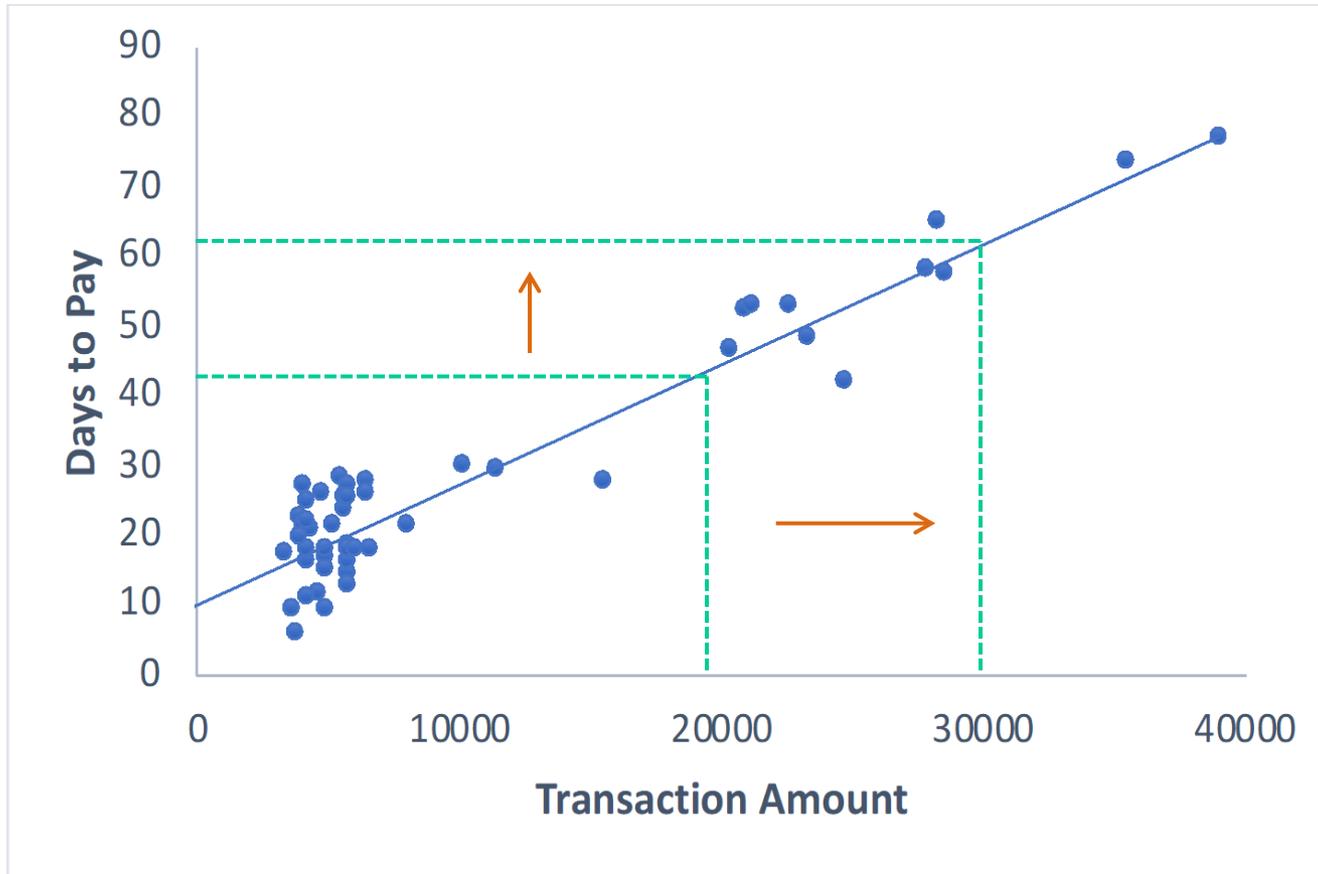
So we fit a line



Prediction



Treatment Effects



Complexity v. Simplicity



Simple linear models make interpretation easy

So why add complexity?

- More variables can improve accuracy.
- Alternate functional forms can be appropriate for different questions or data.
- Other modelling techniques can produce better predictions.

Binary Responses



Sometimes what we are interested in answering a yes or no question:

- Was there a war?
- Did a program experience a cost breach?
- Did a program execute all of its funding?

Answer specific kinds of questions:

- What is the probability of a war under certain political and economic conditions?
- Are certain kinds of programs more likely than others to experience cost breaches?
- What is the probability that a program will execute all of its funding?

Parametric v. Non-Parametric Modelling

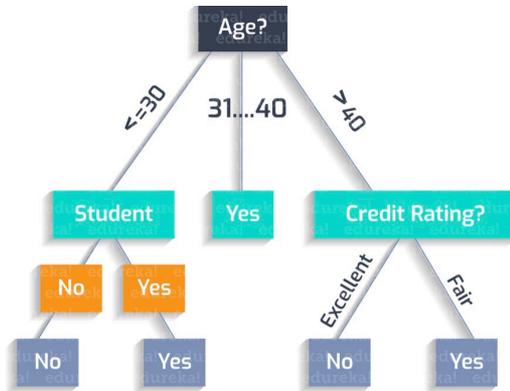


- Parametric Models: Use known functional forms like lines
- Non-Parametric Models: Do not assume fixed structure
 - Computational challenges and hard to interpret
 - Improve prediction; flexible fit

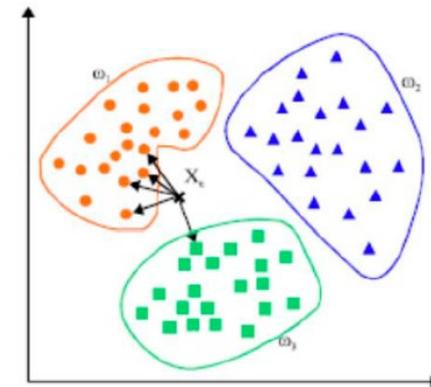


Examples of Non-Parametric Approaches

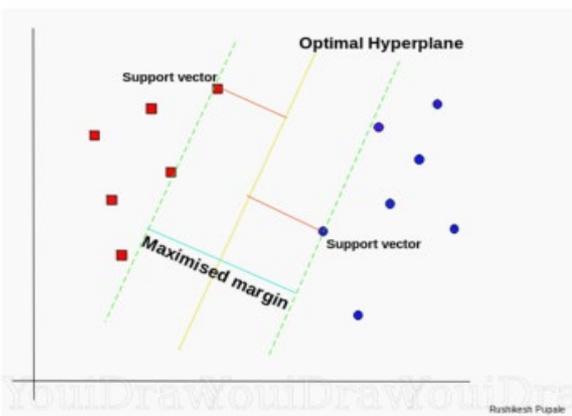
CART – Classification and Regression Trees



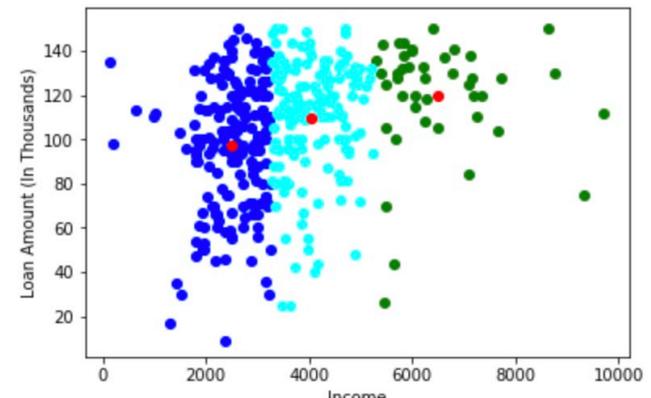
K-Nearest Neighbor (KNN)



Support Vector Machines (SVM)



K-Means Clustering



Defense Applications



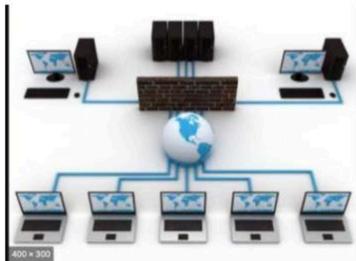
CART



Materials
Transportation
Time Estimates

K-NN

Hacking and
Network Intrusion

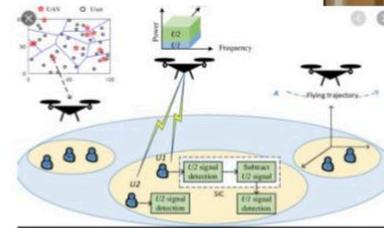


Security Documents
Classifier



K-Means Clustering

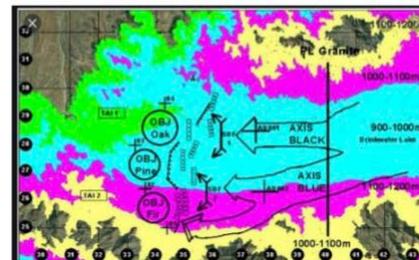
Surveillance
UAV Placement



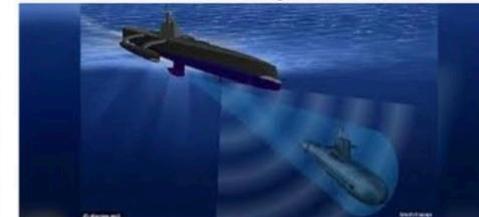
Placement of
RFID Sensors

Support Vector Machines

Military Terrain
Analysis



Friendly or Foe



In A Perfect World...



- We would conduct random controlled experiments
- In an imperfect world, our data are gathered from pilot programs, surveys, historical occurrences, etc.
- We would know everything about everyone; we would have data on the **population**
- In an imperfect world our data are limited, and we usually only get to see a **sample**
 - To be confident that our sample data can say something about the population, we ideally want a large random sample.

Omitted Variable Bias



Recently, numerous studies showed that women who took Hormone Replacement Therapy (HRT) had a lower than average incidence of Congestive Heart Disease (CHD)

When controlling for income, the results showed that HRT increased CHD instead

A model of the impact of bonuses on enlistments found a high degree of sensitivity to these bonuses

Until researchers accounted for unemployment rates at the time of enlistments

All Else Equal



- When examining the impact of a change in one variable on another, we want to hold fixed all other factors such as income, age, race, and individual tastes, external conditions, geography
- If other factors are not controlled for, then we cannot know the causal influence of the variable of interest
- This is a hard problem

Sample Selection Bias



<https://www.upi.com/Archives/1948/11/03/Its-Truman-Dewey-concedes/5701381080400/>

Sample Selection Bias



Selection bias is the **bias** introduced by the **selection** of individuals, groups or data for analysis in such a way that proper randomization is not achieved, thereby ensuring that the **sample** obtained is not representative of the population intended to be analyzed.

Sample Selection Bias (Cont.)



Observations are not representative of the population

- Can occur in experimental and in non-experimental data.
- Worse in non-experimental data.
- Just as bad in big datasets as small ones.

Some Examples



We want to study the impact of combat on veteran mental health

- But, who signs up to be a combat soldier; who gets deployed to combat?
- What if depressed people were more likely to go to war in the first place?

We want to study which programs succeed and which programs fail, but which programs get funded in the first place?

- Could their cost structure have anything to do with it?

Would changing our accounting system positively impact our ability to pay disbursements in a timely way?

- If we compare organizations who've switched to those that didn't, were the organizations that chose to switch more adaptable in the first place?

Becoming better data consumers...



What do you want to do?

- Predict, explain?
- Refine the question

Looked at variety of models:

- How complex does it need to be?

Considered issues with data:

- Where do they crop up?
- What should you be looking out for?

Thank you



For information about DRMI
and its resident and online courses, visit
<https://nps.edu/drmi>

Follow DRMI on [LinkedIn](#)

