It's All Analytics: Separating the Hype from Reality

Presented for the Technology Today Series (TTS) hosted by the Texas Department of Information Resources

August 18, 2020

Scott Burk, Ph.D.

Welcome!

DIR TECHNOLOGY TODAY Series

Host

Joy Hall Bryant

Program Director, IRM Outreach

Texas Department of Information Resources

Co-Host

Ed Kelly

Chief Data Officer (CDO)

State of Texas

Presenter

Scott Burk

Author, Professor, Data Scientist (TIBCO)

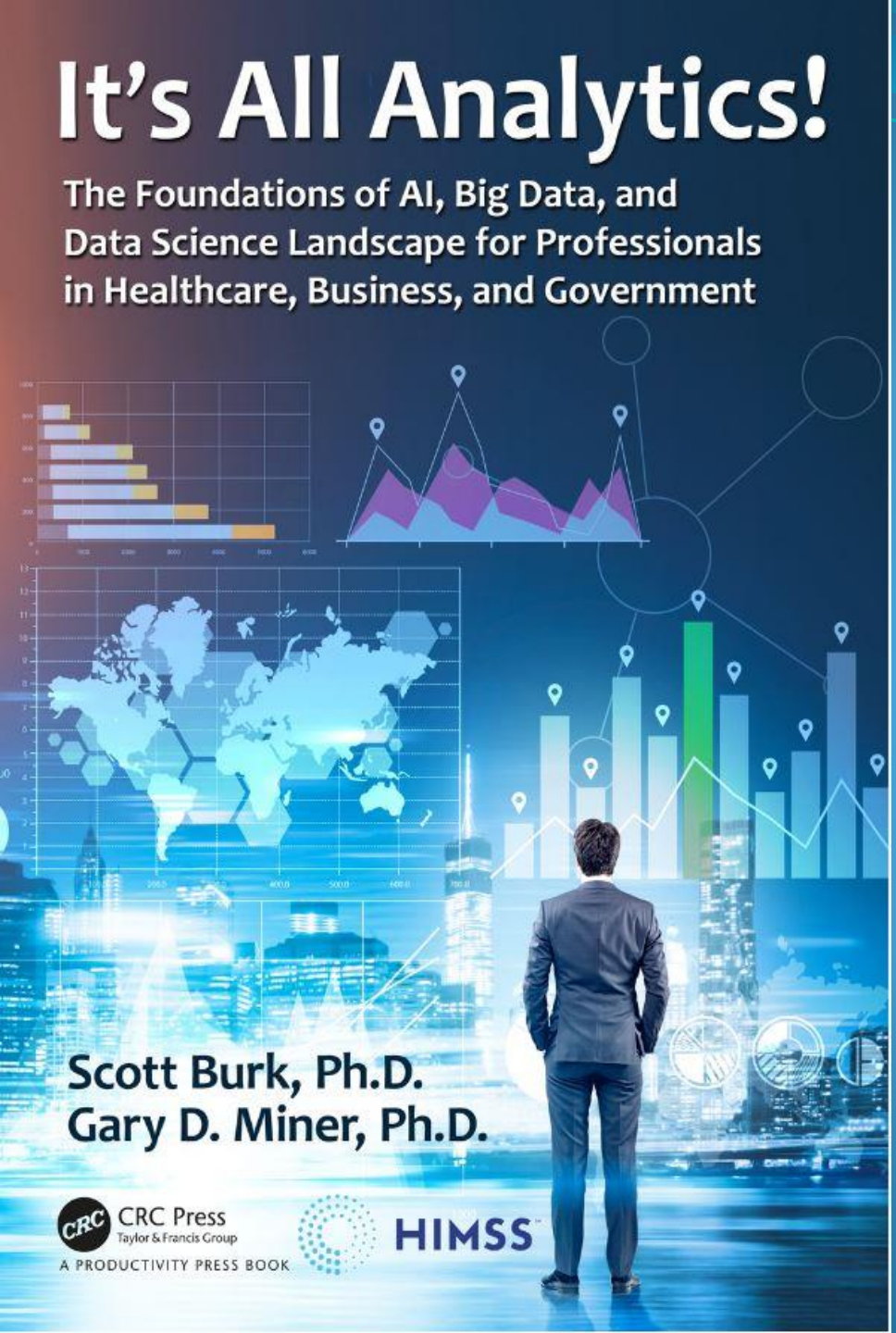
It's All Analytics: Separating the Hype from Reality

Learn More

- DIR website: **www.dir.texas.gov**
- Visit the DIR **CALENDAR** to view events and access more details. (See link at top of any page.)
- Click on **STAY CONNECTED** on DIR home page (bottom left) to subscribe to discussion groups.
- Use the **SEARCH** field to find specific information quickly.

Today's Program

- **Continuing Education**
  - **IRM CPE Form** will be emailed.
- Use **Question Pane** to submit questions.
- **Evaluation Form** – DIR wants your feedback! Evaluation form will pop up as you exit.



Professionals are challenged each day by a changing landscape of technology and terminology. In recent history, especially in the last 25 years, there has been an explosion of terms and methods that automate and improve decision making and operations. The term "analytics" is an overarching description of a compilation of methodologies. But recently, there has been a resurgence in AI (artificial intelligence, statistics, decision science, and optimization. Also, things like business intelligence, online analytical processing (OLAP), and many more have been born or reborn. How is someone to make sense of all this methodology and terminology?

Coming in Early 2020

- It's All Analytics – Part II; Designing an Integrated AI, Analytics, and Data Science Architecture for Your Organization
  - Organizational Design
  - Data Architecture
  - Analytics Architecture

Themes / Takeaways

- Team Sport and “Democratizing Analytics”
- The Need for Data Literacy
- Better Together

Analytics Impacts You

- Analytics Producer
- Business Consumer
- Product and Services Consumer
- Citizen

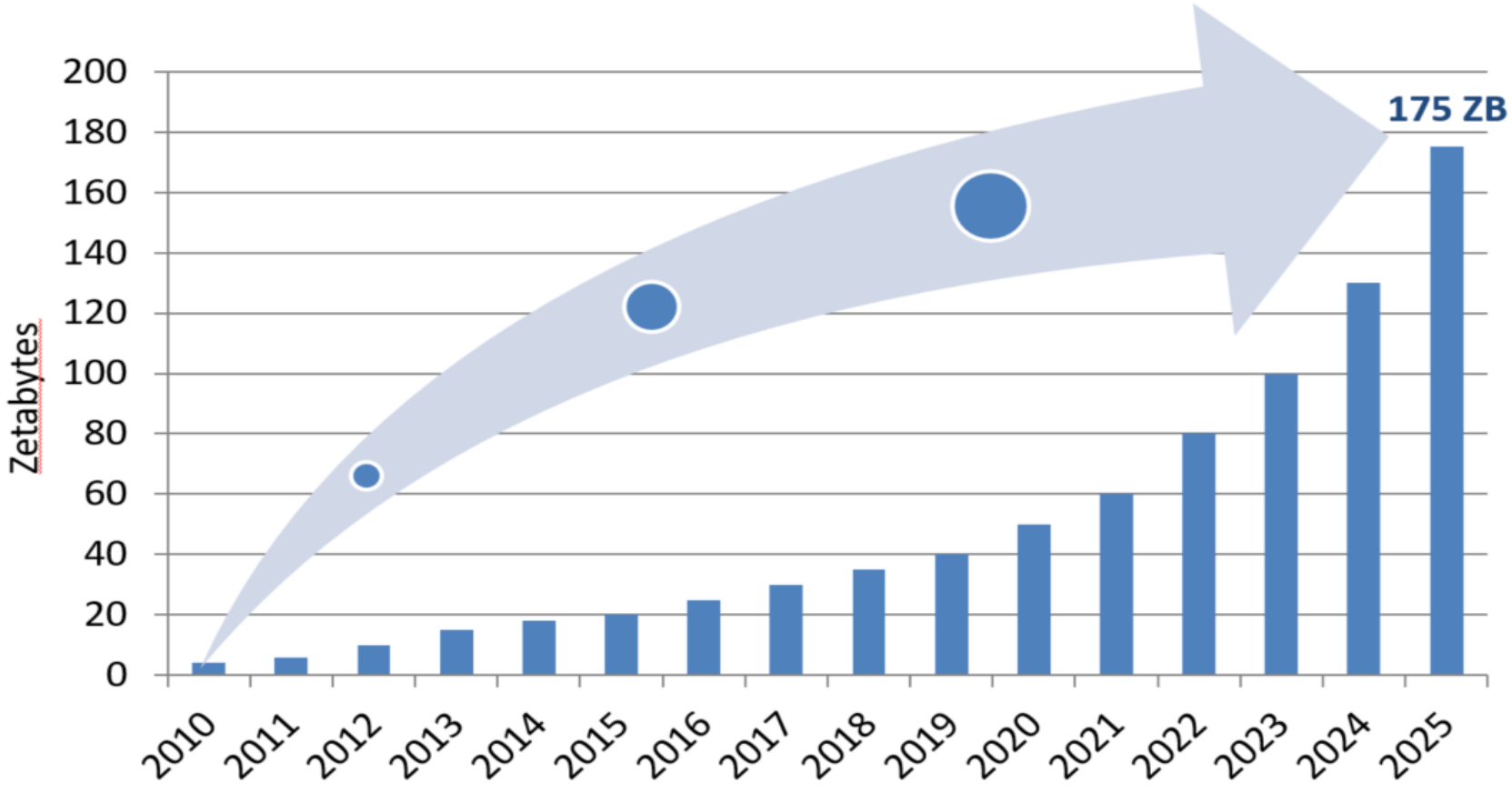
Privacy – Real Issues when Machines are Very Smart

- Social media giant Twitter warned on 8/3/20 that it could be fined hundreds of millions of dollars for allegedly misusing users' personal data for advertising purposes.
- Uber tracked user data AFTER riders completed their rides — AND possibly after they deleted the APP
- "I think of machine learning (artificial intelligence) kind of asbestos…… It turns out that it's all over the place, even though at no point did you explicitly install it, and it has possibly some latent bad effects that you might regret later, after it's already too hard to get it all out" - Jonathan Zittrain, HMS 2019

Data is the new oil?

- Data collection is ***cost*** and a ***liability***, not an asset unless it is used!
  - Sourcing – APIs, connections
  - Cleansing?
  - Storage
  - Security
- According to Forbes (Meehan 2016) it is estimated that as much as 90% of big data is never analyzed.
- How do you turn the corner?
  - Data Literacy
  - Monetize (Get Value) from the Data
  - Create and Deploy Analytics
- "it is not technology that is hampering the progress. *It is lack of vision, human capital, and execution*"

Annual Size of the Global Datasphere



(Adapted from Reinsel et al, Data Age 2025; The Digitization of the World (IDC Nov 2018)

And then there is IoT, IoE

- ***5 quintillion bytes of data*** *produced every day (that's 2.5 followed by 18 zeros)*
- *By the* year *2020, the IoT will comprise more than* ***30 billion connected devices.***
- *Only* ***26%*** *of companies surveyed reported that their IoT initiatives have been successful*
- ***Less than half*** *of* ***structured*** *data is actively used in decision making*
- ***Less than 1%*** *of* ***unstructured*** *data is analyzed or used at all*

We Don’t Have Data Democracy

- How many have heard we need to be data driven?
- Marketing Hype
- Self Promotion Hype
- Media Hype
- Political Hype

We Need Data Literacy!

- “Imagine an organization where the marketing department speaks French, the product designers speak German, the analytics team speaks Spanish and no one speaks a second language…… That’s essentially how a data-driven business functions when there is no data literacy.” - Kasey Panetta (Gartner)

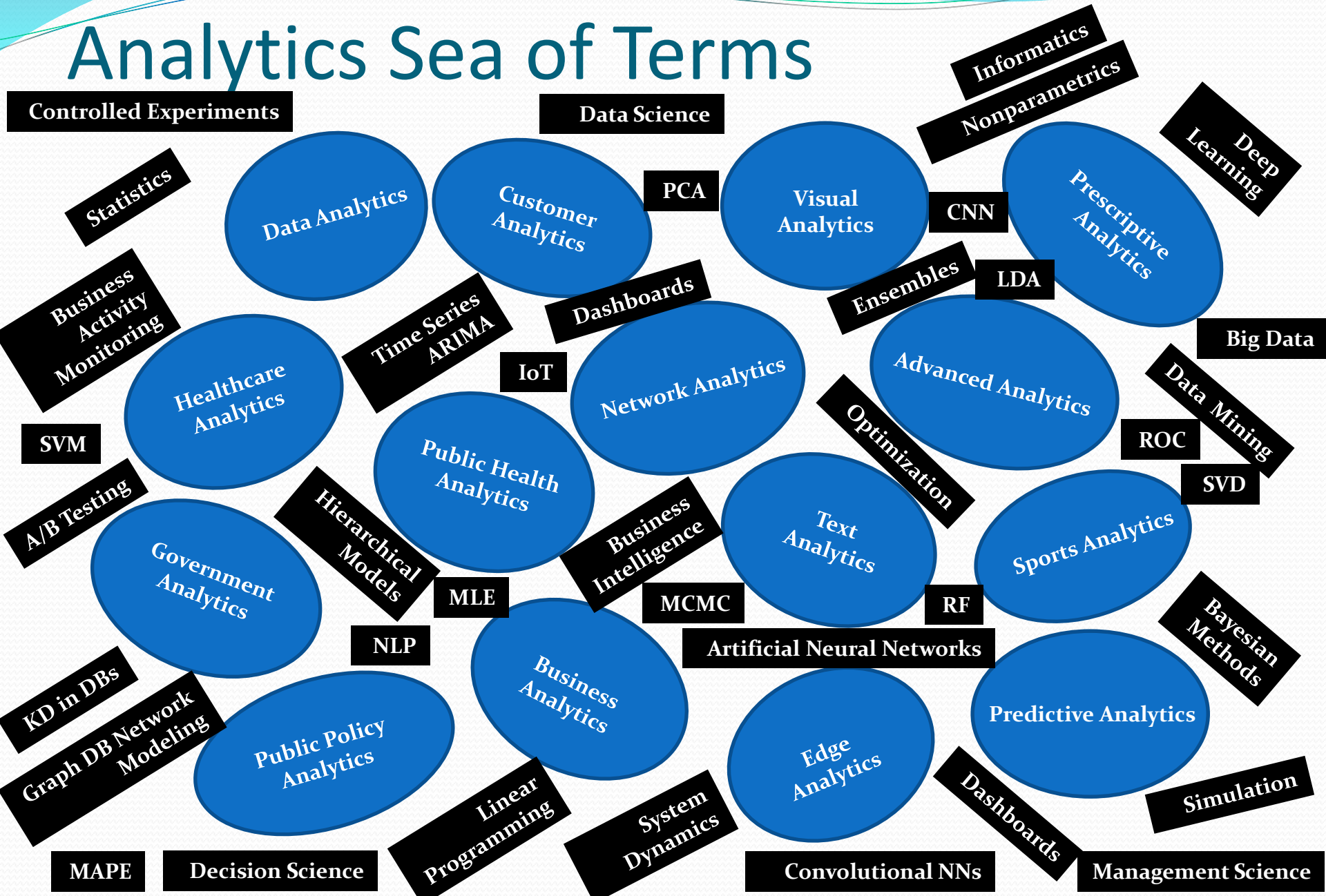
- “prevalence of data and analytics capabilities, including artificial intelligence, requires creators and consumers to ‘speak data’ as a common language,…… Data and analytics leaders must champion workforce data literacy as an enabler of digital business and treat information as a second language.” – Valerie Logan

- 2020 – ½ of organizations will lack data literacy skills

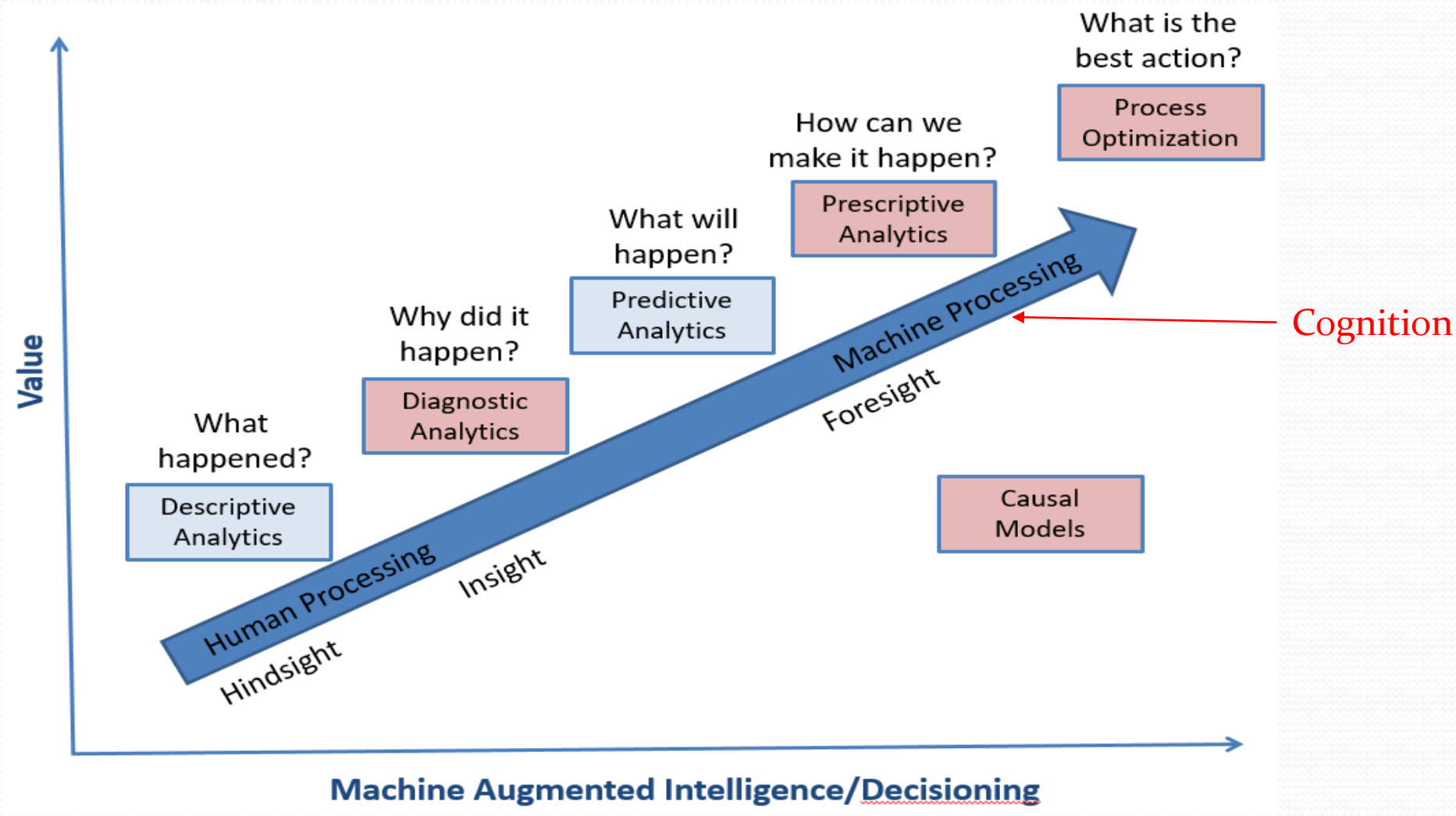
Why is it hard?

- "The rapid acceleration in the development and deployment of these technologies is creating an increasing gap in understanding. Many who need to know don't even know what they don't know. This, coupled with hyperbolic news releases on some new AI application-of-the-moment, leaves the nontechnical observer with no easy solution to bridging this gap." – *John Cromwell, MD (Associate Chief Medical Officer | Director of Surgical Quality and Safety, UIHC)*
- "siloed view of the fields."

Analytics Sea of Terms



4 Generally Accepted Categories for Analytics



“I declare bankruptcy!”

Predictive model -> Prescriptive

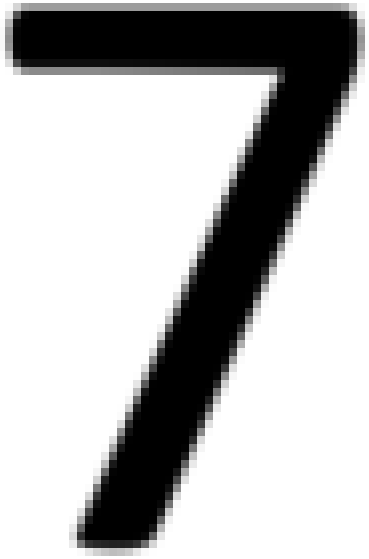
At Minimum

- To assume or establish causation
- Be able to manipulate the important variables
- Not perturb the system

RISKS

- You could be wrong in assuming that your factors are causative. Thus, manipulation of the inputs will not provide the results you are expecting.
- You may perturb the system. If the system has a causal relationship and you manipulate variables, you will affect the system itself. This can also happen whether you manipulate variables or not, as systems change over time. However, you should take care in your manipulation and not try and 'shock' the system with the manipulation of your inputs.

What is a model in the 1st place?



Data is about PROCESS

- Process - a series of actions or steps taken in order to achieve a particular end.
- “Data has the ability if used correctly to separate anecdote and hyperbole from facts.”
- Processes are everywhere and drive data generation
- Processes drive data
- Stats 101
  - Populations
  - Parameters
  - Samples
  - Statistics

So why the mistrust and misgivings of statistics?

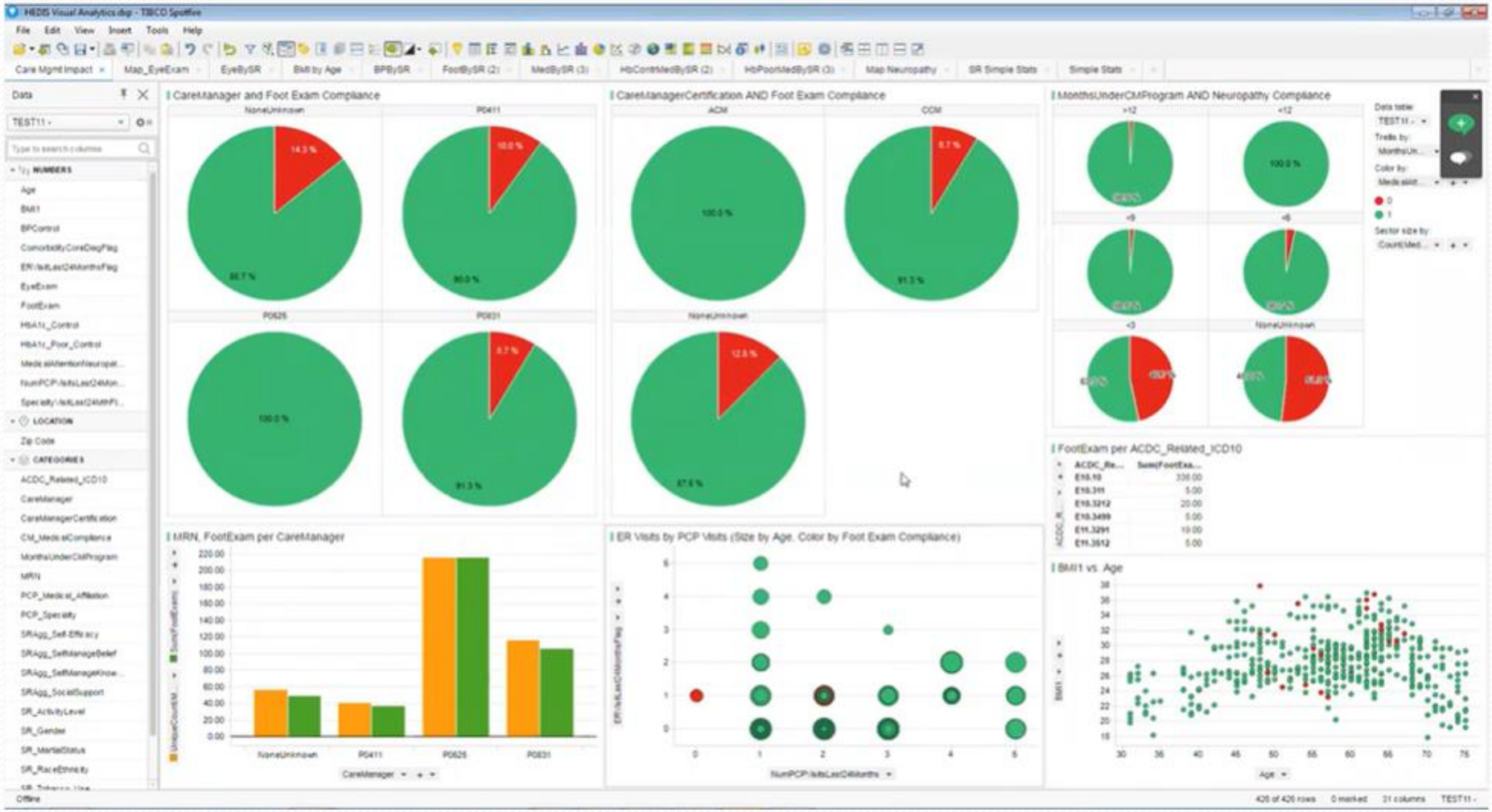
- Concepts of Statistical Inference is hard for most people
  - Personal Confession
  - Simple article in American Statistician
  - Was Ty Cobbs a 400 hitter?

BI, Visual BI and Descriptive Analytics

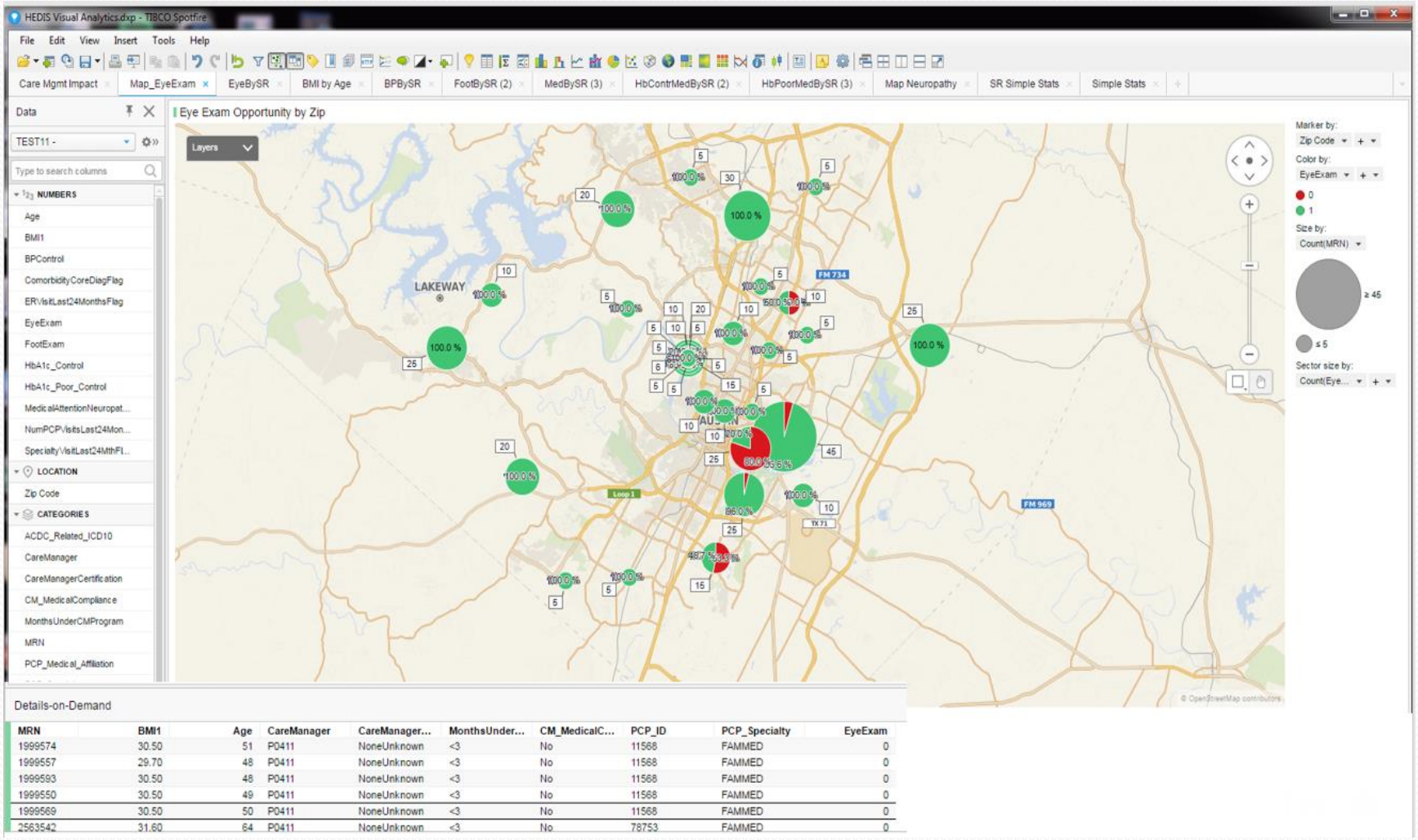
- What are they?
  - Data Warehouse and Single Version of Truth
  - Extend in Playing with Data

- Operational BI
- Project Based BI
- BAM and EMS

Visual BI and Interaction, Brush Linking



Geospatial Analytics, Hundreds of Layers, Streaming



MRN	BMI1	Age	CareManager	CareManager...	MonthsUnder...	CM_MedicalC...	PCP_ID	PCP_Specialty	EyeExam
1999574	30.50	51	P0411	NoneUnknown	<3	No	11568	FAMMED	0
1999557	29.70	48	P0411	NoneUnknown	<3	No	11568	FAMMED	0
1999593	30.50	48	P0411	NoneUnknown	<3	No	11568	FAMMED	0
1999550	30.50	49	P0411	NoneUnknown	<3	No	11568	FAMMED	0
1999569	30.50	50	P0411	NoneUnknown	<3	No	11568	FAMMED	0
2563542	31.60	64	P0411	NoneUnknown	<3	No	78753	FAMMED	0

BI, Visual BI and Descriptive Analytics - Strengths

- *Humans are geared to seeing visual relationships and getting it*
- *Humans love to play with things, interact*
- *Very little training is needed, very intuitive and many people start their analytics careers in BI*
- *May offer a "A Single Version of the Truth", maybe transparent*

BI, Visual BI and Descriptive Analytics - Weaknesses

- *Humans have strong imaginations & can see things in a graphic that do not really exist*
- *Retrospective only – except in the mind of the consumer*
- *Computers are more objective and cheaper for many tasks and produce higher ROI results.*

Data Mining and Machine Learning Where do they shine vs BI?

- *It lets the computer do the processing/cognition, (A Note on Cognition)*
- *Results are mathematically verified, meaning that an established algorithm is interpreting the data rather an analyst looking at aggregations, statistics of visual dashboards and interpreting*
- *The size of the problem is only limited by computer resources - machine learning covers very large problems*

Data Mining and Machine Learning Some Limitations vs BI?

- *More education and training over BI*
- *Where is the fun in this??*
- *Total Cost is higher*
- BUT, BI and ML are not exclusive!
  - BETTER TOGETHER as methods
  - BETTER TOGETHER as a TEAM!

Artificial Intelligence
Where does it shine?

- *Extremely Powerful for Nonlinear and Patten Recognition*
- *Can Process Thousands of Variables*
- *Great Results with Unstructured Data*

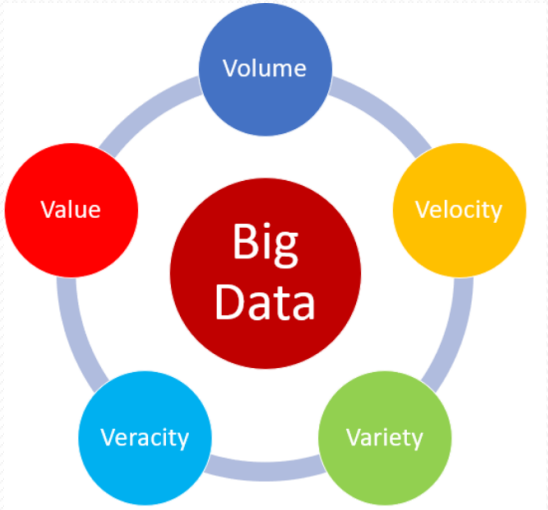
Artificial Intelligence
Some Shortcomings

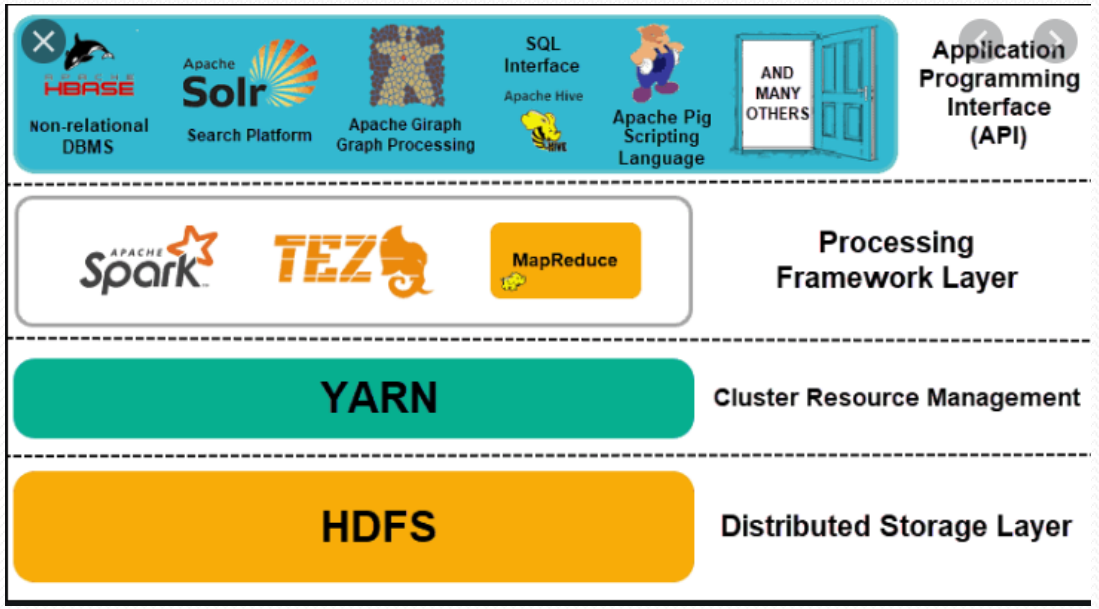
- *Requires much more data to train/build models*
- *Models are not transparent. They are not readable.*
- *They encode correlation and not causation*

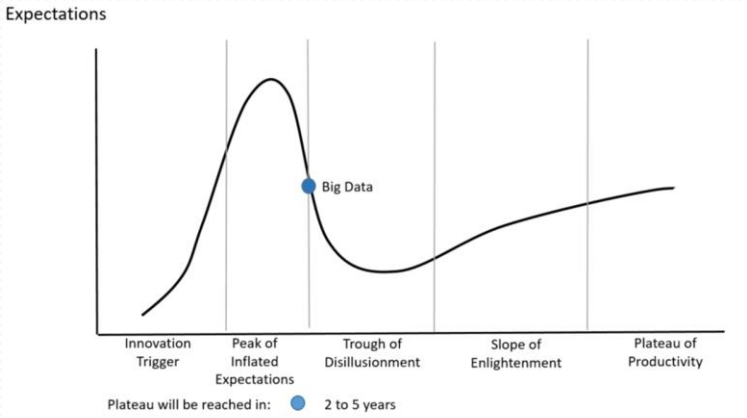
Data Scientist

- Unicorns
- Multidisciplinary professionals and sub specialties
  - Data Scientist
  - Data Engineer
  - Citizen Data Scientist
  - SMEs
  - Developers
- Data science is a “team sport”
- Both the methods and professionals – “Work Better Together”
- Today’s Professional and University - Hot Term

Big Data – what is it?







Statistics, Causation and Related Analytics

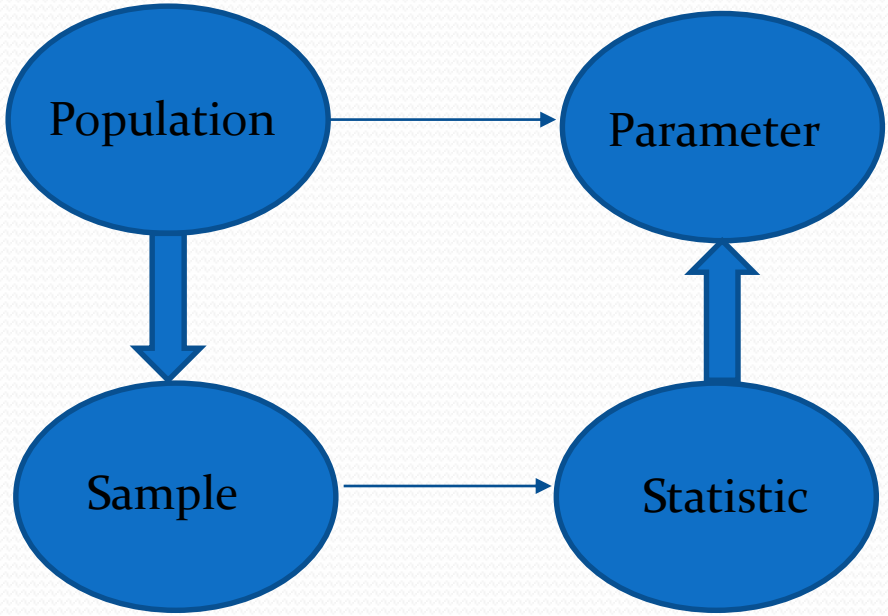
- They are required if you want to move beyond description and prediction
  - Diagnostic – why something happened?
  - Prescription – how to make something happen
  - Optimization – how we can the best thing happen
- They are well established and accepted
- They are everywhere, many accepted for years and will not be replaced with ML, AI or BI.

Statistics Strengths

- Virtually all data driven methods use Stats – BI, AI, ML, Big Data Methods
- Only Method Available to Solve Many Problems
  - Time Series Methods (transparency and establishment)
  - Randomized Controlled Trials (RCT) and Experimental Design
  - Small Sample Methods
  - Significance, Inference, A/B Testing

Statistics Strengths

- Virtually all data driven methods use Stats – BI, AI, ML, Big Data Methods
- Only Method Available to Solve Many Problems
  - Time Series Methods (transparency and establishment)
  - Randomized Controlled Trials (RCT) and Experimental Design
  - Small Sample Methods
  - Significance, Inference, A/B Testing

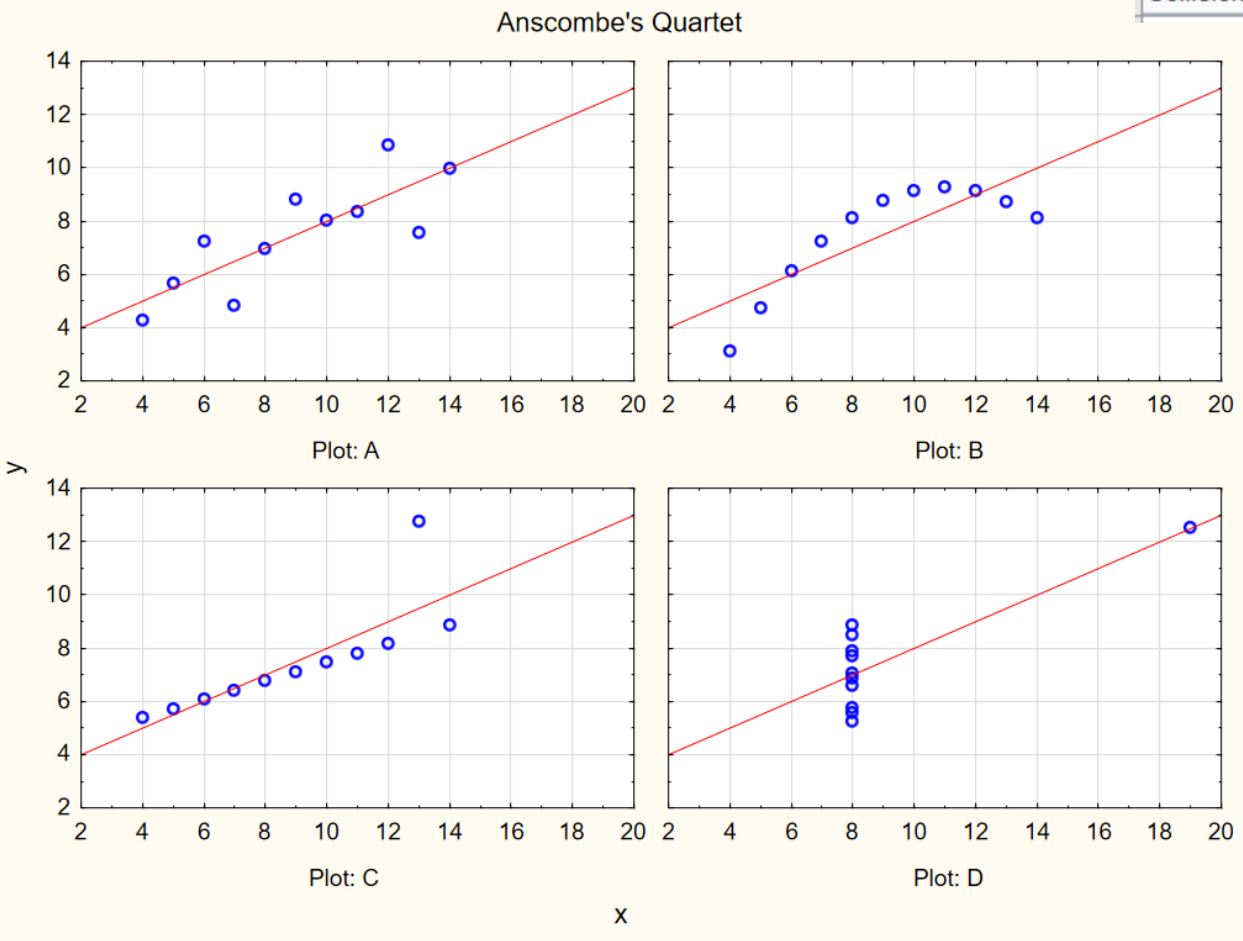


Statistics Limitations

- *Steeper learning curve, some areas are just darn hard*
- *Broad enough to cause confusion due to exposure and education– Blind Men and the Elephant Problem*
- *Some require assumptions that are wishes, not reality*
- *Can lead to the wrong interpretation in isolation (see following)*

Anscombe's Quartet

Mean of x	9
Sample variance of x	11
Mean of y	7.5
Sample variance of y	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.00 + 0.500x$
Cofficient of Determination	0.67



Better Together!

Better Together!

- There is no free lunch!
- Be versed in several disciplines – pick up what you need as you need them or take new challenges!
- Technology is important, but clearly, asking right questions is more important
- Analytics is a Team Sport, be part of the team.

Thanks You!
Questions?

Host

Joy Hall Bryant

Program Director, IRM Outreach

Texas Department of Information Resources

Co-Host

Ed Kelly

Chief Data Officer (CDO)

State of Texas

Presenter

Scott Burk

Author, Professor, Data Scientist (TIBCO)

It's All Analytics: Separating the Hype from Reality

Contact

*** Linkedin**
https://www.linkedin.com/in/scott-burk-phd

*** Twitter (not very active right now)**
https://twitter.com/1ScottBurk

*** Youtube – 'Scott Burk' Channel**
'Its All Analytics' Playlist
https://www.youtube.com/playlist?list=PLX-TyAzMwGs8hdKFnqexZKI9hh27r7p2N

*** Book on Amazon**
https://www.amazon.com/Its-All-Analytics-Foundations-Professionals/dp/0367359685

DIR TECHNOLOGY TODAY Series