Goodman
School of Business

# Goodman Group

Brock University

@GSBGoodmanGroup

goodmangroup@brocku.ca

GSB Goodman Group

Goodman
School of Business
Brock University

Goodman Group

# Welcome to Business Breathers

# Big Data Analytics: Wisdom or Folly!

Dr. S. Ejaz Ahmed
Dean, Faculty of Math and Science
Brock University

Taking businesses, careers and ideas to the next level

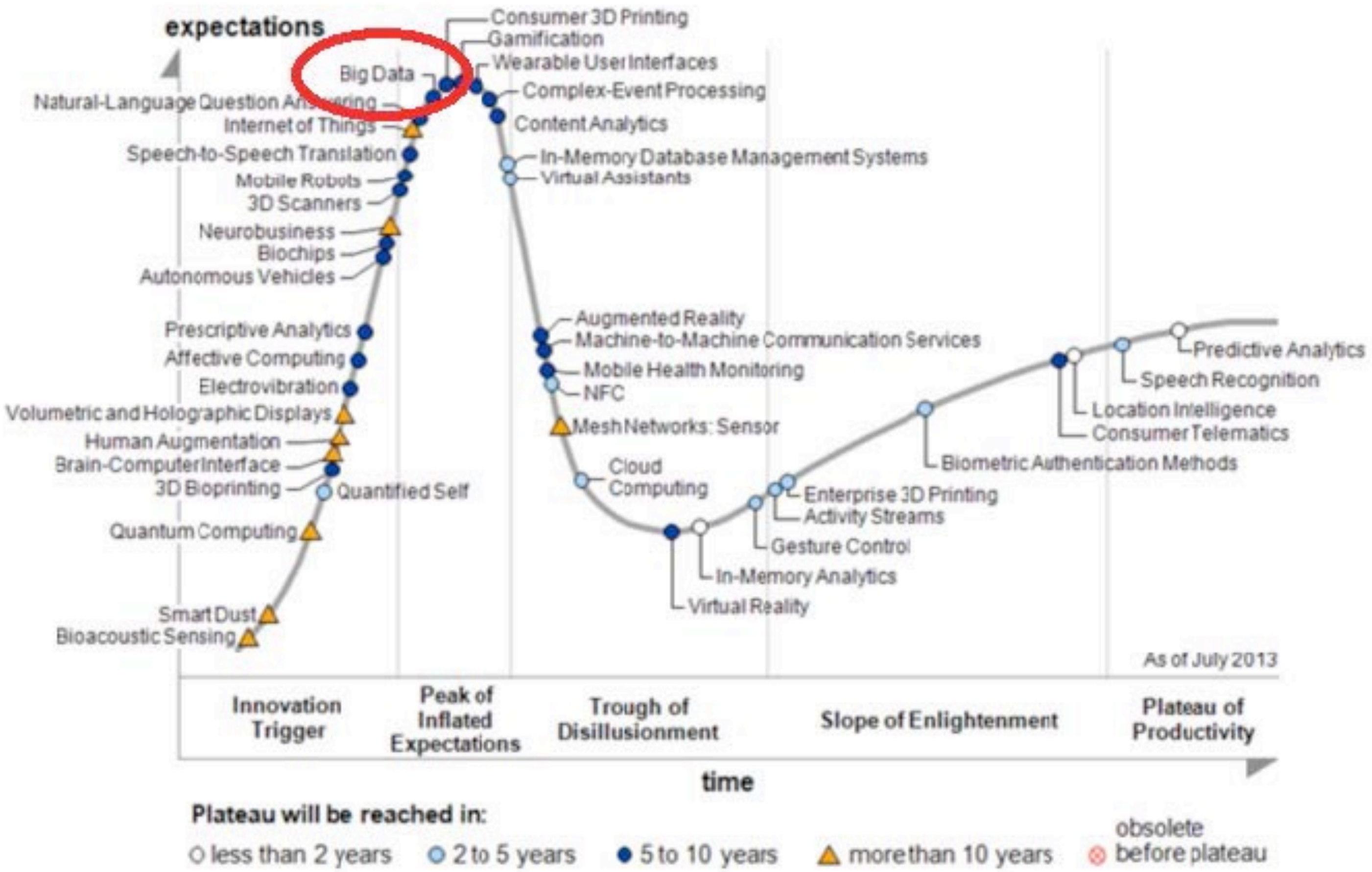
# Big Data Analytics

S. Ejaz Ahmed
Brock University

Joint work with
Feryaal Ahmed
Ivey Business School, Western University

# Pictures From My Backyard!! Niagara Falls, Canada

**Figure 1. Hype Cycle for Emerging Technologies, 2013**

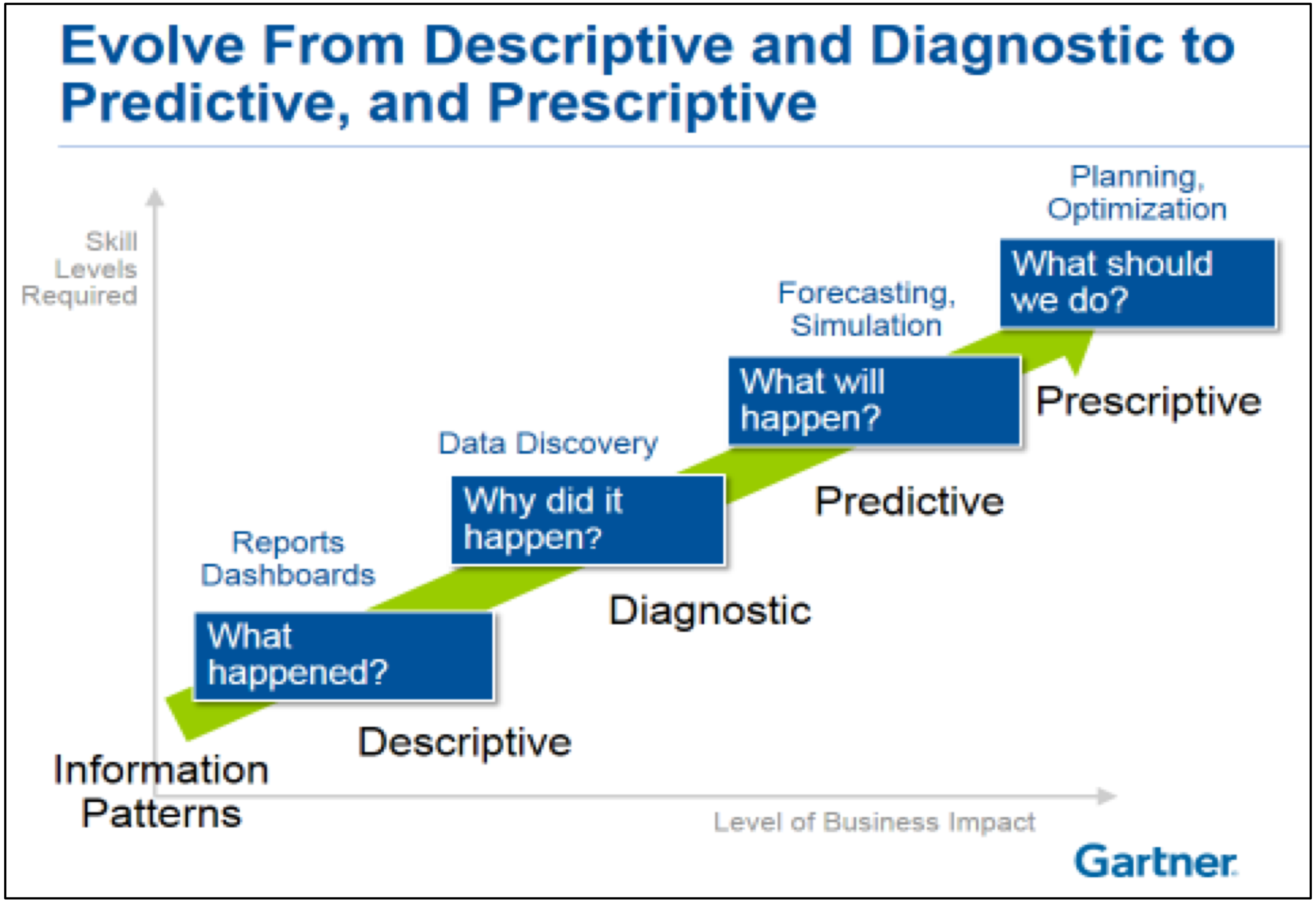


# Business Analytics

- ❑ Descriptive analytics: gains insight from historical data with reporting, scorecards, clustering etc.

- ❑ Predictive analytics: employs predictive modeling using statistical and machine learning techniques

- ❑ Prescriptive analytics: recommends decisions using optimization, simulation, etc.



# Evolve From Descriptive and Diagnostic to Predictive, and Prescriptive

Skill Levels Required

Information Patterns

Reports Dashboards

What happened?

Descriptive

Data Discovery

Why did it happen?

Diagnostic

Forecasting, Simulation

What will happen?

Predictive

Planning, Optimization

What should we do?

Prescriptive

Level of Business Impact

Gartner

Jack Levis Presentation

# Data Science

- What is Data Science?
- What is Big Data?
- What is High Dimensional Data?
- Data Science recognized as a field of Science at the beginning of the century
- Volume of data due to technological advancement can be stored

# Data Science

- Big Data – Big Problem: May not be Valuable? A paradox
- Data needs to be extracted to make the data valuable for further study
- Data pre-processing, data massaging, data/variable reductions
- *Data reduction (extracting valuable information) from Big Data is called "Data Science"*

From IBM. The spinning griddy vortex of data. The image in so nuanced, so layered. The grid implies structure, but the swirliness implies an unstructuredness to the structure. In that structure there is data

# BIG DATA MATRIX (n x p)

- Big Data with a Big number of Variables
  - Scenario I: large volume of data
    - “n” is very very large
  - Scenario II: large data matrix
    - “n” is very large
    - “p” is also large

- HIGH DIMENSIONAL DATA (HDD)
  - Scenario III: “n <<< p”

# Some of Important Features Related to Big Data

- Volume: amount of data
- Variety: various types of data
- Velocity: speed of data processing
- Veracity: uncertainty and imprecision in data

- ❑ In genetic micro-array studies, “n” is measured in hundreds (now in thousands), the number of features “d” per sample can exceed millions!!!

- ❑ 22,500 unique proteins, implies about 253,000,000 possible protein-protein interactions. So far 42,000 are identified.

- ❑ Facebook: More than 950 millions user spends approximately 7 hours on average for a given month, a huge task to store and analyze such Big and Complex Data.

- ❑ Combining different Health Databanks with a great heterogeneity collected over the time.

- ❑ In industrial applications massive data are collected for statistical process monitoring.

# Processing and Volume Challenge in Financial Analytics

- ❑ A credit card company targets card holders focusing on customer default payments

- ❑ Financial Credit Data Example: The dataset contains 980,000 observations and includes information on more than forty five variables such as default payments, demographic factors, credit data, history of payment billing statements of credit card clients, and others.

- ❑ These data may contain various anomalies that are hard to detect due to volume and processing challenges
  - ▪ **Heterogeneity**
  - ▪ **Outliers**
  - ▪ **Round off Errors**

# Imbalanced/Skewed Data Sets

- Data is highly skewed as sampled from one majority group are larger than the other minority groups
- Credit Card Fraud Detection/Credit Scoring
- **According to ACI Worldwide**, 47% of Americans have fallen victim to credit card fraud in the past five years.
- **The Nilson Report estimates** that losses from credit card fraud exceeded $27 billion in 2018.
- The credit card dataset is highly skewed as it contains more non-fraudulent cases, indeed!
- Imbalanced datasets can be mitigated by **under-sampling** the majority group and **oversampling** the minority group

# Supervised Machine Learning

- Several machine learning algorithms are popular based on pattern recognition.
  - Support Vector Machine
  - Neural Nets
  - Multivariate Adaptive Regression Splines
  - Random Forest
  - K-nearest neighbours classification
- All above techniques are useful to build an appropriate predictive model
- However, these techniques come at a cost of interpretability, relaibility and computational power

# Supervised Learning/Regression Model

- We can model the data with a multiple regression model, for example, Logistic Regression Model
- If the data is high-dimensional n<<<p
- The classical statistical techniques will not work
- Recently suggested, we use penalized likelihood methods (e.g. LASSO)

- The biggest assumption is that the model is *sparse*, that is only a few features are important for prediction purposes and the others are not
- We apply LASSO and other methods to get a submodel consisting of important features only
- The prediction is based on the selected features in the submodel
- Submodel predictions are easily interpretable and efficient compared to Full models

# End of the story? Really?

Let’s dig in deeper!

- Good strategy if model is truly sparse
- Unrevealing take of Underfitted model
- Submodel Estimators are BIASED!!!
- BIAS is the **BIG** Problem in Big Data
- *NEED TO CONTROL BIAS*

- Bias will increase without a bound! Consequentially MSE will Explode

- Not a good Trade-off!!!

- “*All Submodel are Biased, but Some are Useful*”

- A *naive data analyst/data scientist/data miner* and others may not comprehend that by dropping “VARIABLES "from the model the impact will be drastic.

- **The law of “Unconsciousness” will not work!**

# World's Data is Growing Exponentially!

- Greater collaboration between statisticians, computer scientists and social scientists (Facebook clicks, Netflix queues, and GPS data, a few to mention, 12 billions devices are connected to internet).
- Data is never neutral and unbiased, we must pull expertise across a host of fields to combat the biases in the estimation.

- Need to be careful with algorithmic based predictions. For example, protein interaction prediction.

- "The purpose of computing is insight, not numbers." R.W. Hamming, 1962.

# Clash of Cultures

# Culture in Statistical Sciences

- Pre-processed sampled data
- Classical assumptions Exact/Analytic Solutions
- Low-dimensional Data Analysis
- SAS, SPSS, and others; R(open source program)
- Idealistic
- Work Alone or in Small Teams
- Glory of the Individual

# Culture in Data Science

- Big Data (available from different sources)
- Framing the Problem
- Identify the data for analysis (population or sample)
- **New tools for New data**: Python and R (open source programs)
- **Parallel/Cloud Computing systems**: Hadoop, Spark, and others
- Visualizing Complex data

# Culture in Data Science

- Combining Data (Data Fusion, Record Linkage): Tax returns, Court records, and Health agencies, etc.
- High-Dimensional Statistical Inference
- Pragmatic
- Think Tanks - Trans-disciplinary Research
- Glory of the Research Team

# References

- F. Fang, J. Zhao, S. E. **Ahmed** and A. Qiu (2020). A Weak-signal-assisted Procedure for Variable Selection and Statistical Inference with an Informative Subsample, Biometrics.
- Li, Vidyashankar, Diao, **Ahmed** (2019). Robust Inference after Random Projections via Hellinger Distance for Location-Scale Family. Entropy
- Li, Hong, **Ahmed**, and L. Yi (2018). Weak signals in high-dimensional regression: Detection, estimation and prediction.
- **S. E. Ahmed** and A.I.Volodin (Editors). Journal of Statistical Computation and Simulation, special issue, published online, 2019.
- **S. E. Ahmed** (Editor) Applied Stochastic Models in Business and Industry, special issue, published online, 2019.
- **S. E. Ahmed**, F. Carvalho and S. Puntanen Matrices (Editors). Matrices, Statistics and Big Data, 2019, in print, Springer
- **S. E. Ahmed (Editor)**. Big and Complex Data Analysis: Statistical Methodologies and Applications. Springer, 2017.

# References

- **S. E. Ahmed** (2014). Penalty, Pretest and Shrinkage Estimation: Variable Selection and Estimation. Springer.
- **S. E. Ahmed** (Editor). Perspectives on Big Data Analysis: Methodologies and Applications. Contemporary Mathematics, a co-publication of American Mathematical Society and CRM, 2014.
- Hossien, **Ahmed** and Doksum(2015) for Generalized linear models
- **Ahmed** and Fallahpour (2012) for quasi-likelihood models.
- **Ahmed** et al. (2012) for Weibull censored regression models.
- Fallahpour, **Ahmed** and Doksum (2010) partially linear models
- **Ahmed** and Fallahpour (2014) with Random Coefficient autoregressive Errors.
- **Ahmed** et al. (2008, 2009) for partially linear models.

# Thanks a Bundle!

- Thanks to the organizers for the invitation
- Thank you all for sharing your valuable time with me!
- The research is supported by the Natural Sciences and the Engineering Research Council of Canada and other sources

Goodman School of Business

Goodman Group

Brock University

@GSBGoodmanGroup

goodmangroup@brocku.ca

GSB Goodman Group