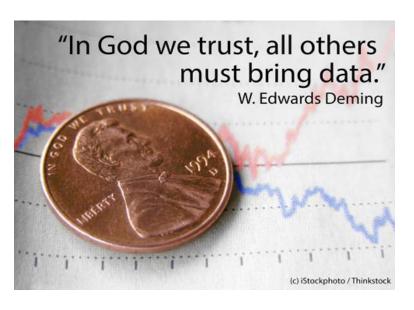
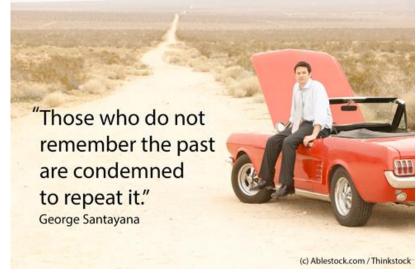
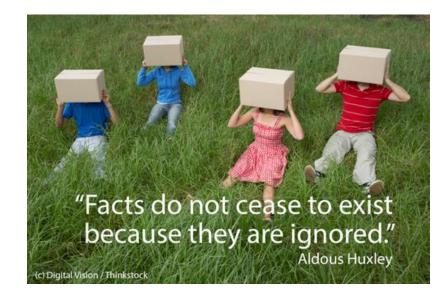
# WIISE Workshop on Business Analytics Tools-R Big Data, Data Science, Analytics



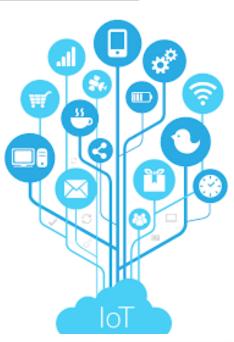




### Agenda



- ☐ Introduction to Business Analytics & Way forward
- □ Comparison of different analytics tool and industry application.
- Learn basics of Data Science, Its Types, Variables, and Charts
- ☐ Introduction to R tool and its functionalities.
- Open house session.







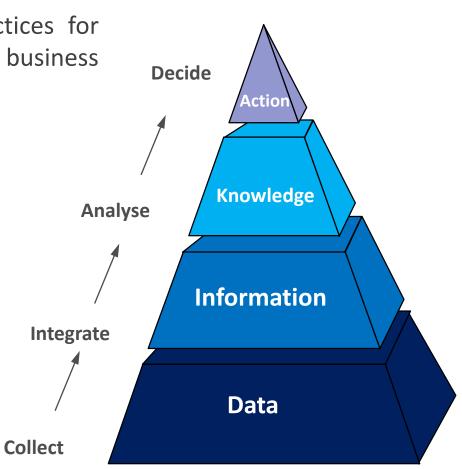
# Introduction to Business Analytics & Way Forward

### Business Analytics- Definition



□ Business analytics (BA) refers to the skills, technologies, practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning.

- ☐ Business analytics focuses on developing:
  - Actionable Insights
  - Predictable performance
  - Risk Forecast
  - Data Driven Decision
  - Effective Planning
- □Strong catalyst for building Operational Excellence



### Business Intelligence V/s Business Analytics



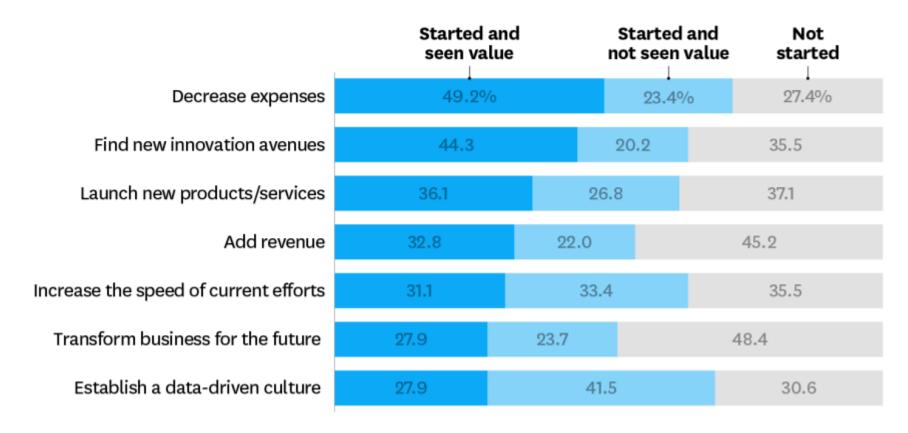
BUSINESS INTELIGENCE	VS	BUSINESS ANALYTICS
<ul><li>What happened?</li><li>When?</li><li>Who?</li><li>How many?</li></ul>	ANSWERS THE QUESTIONS:	<ul> <li>Why did it happen?</li> <li>Will it happen again?</li> <li>What will happen if we change x?</li> <li>What else does the data tell us that never thought.</li> </ul>
<ul> <li>Reporting (KPIs, metrics)</li> <li>Automated Monitoring/Alerting (thresholds)</li> <li>Dashboards</li> <li>Scorecards</li> <li>OLAP (Cubes, Slice &amp; Dice, Drilling)</li> <li>Ad hoc query</li> </ul>	INCLUDES:	<ul> <li>Statistical/Quantitative Analysis</li> <li>Data Mining</li> <li>Predictive Modeling</li> <li>Multivariate Testing</li> </ul>

### Early Adaptors



#### **How Fortune 1000 Executives Report Using Big Data**

The projects they've started, and where they're finding value.



### Success stories: Nissan





Nissan Motor Company owns a network of websites across the world that helps consumers decide which Nissan vehicle they'd like to purchase.

**Start your engines:** Nissan wanted to access much greater detail – such as users' preferences according to *car type, model, and colors – to enable them to measure more information about each inquiry within their Analytics reports* and to make better decisions about the allocation of inventory in local markets.

#### **Ready to Roll:**

- Nissan discovered an innovative application for tracking their non e-commerce activity.
- □ Nissan's Global Marketing Strategy Division can understand which vehicles are in demand, and can make decisions tailored for each local market.
- ☐ Thus, Leveraging the business bottom line in terms of Budgeting, Operation Cost, Planning.

#### **About Nissan**

- www.nissan-global.com
- Japanese automotive company with worldwide presence

#### Goals

- Gain deeper understanding into audience product preferences
- Make informed decisions about serving demand in local markets

#### Approach

- Added e-commerce tracking tags
- Implemented custom reporting to facilitate knowledge sharing

#### Results

 At-a-glance reporting enables quicker, more accurate decisions

### Success stories: Dominos





- □ Domino's monthly revenue by 6%
- Domino's is dominating pizza delivery sales. Today, Domino's is the most popular pizza delivery chain and sales just keep growing.
- □ From 2014 onwards Domino's sold 76 million pizzas and generated £766.6 million in revenue—a 14.6% increase.
- ☐ Their online sales are increasing 30% year over year and currently account for almost 70% of all sales.

#### **Problems**

Sales forecasting, Product Distribution, Inventory management, Strategies marketing plans for footfalls

#### Goals

- · Integrate marketing measurement across devices
- Create a holistic view of customer behavior
- Make cross-channel marketing performance analysis easy and efficient

#### Results

- Realized an immediate 6% increase in monthly revenue
- Saved 80% year-over-year in ad serving and operations costs
- Increased agility with streamlined tag management
- Obtained easy access to powerful reporting and customized dashboards

### Success stories: Amazon amazon





Amazon reported a 29% sales increase in second fiscal quarter of 2016

Source: Fortune.com

- ☐ "Products You may Like"
- "Customer Who Bought This Item Also Bought"
- ☐ "Frequently Bought Together"
- ☐ Predict delivery time accurately
- Analyse review data

#### Customers Who Bought This Item Also Bought



Spigen iPhone 7 Plus Tempered Glass Screen Protector 043GL20608 15 m

₹1.299.00 **Prime** 



Spigen Liquid Crystal Case for iPhone 7 Plus Space Crystal 043CS20855

**★★★☆☆ 22** 

₹1.299.00 **\Prime** 



iPhone 7 Plus Case. Spigen Liquid Crystal Case for iPhone 7 Plus 043CS20479

**★★★☆☆☆ 22** ₹1.199.00 **Prime** 

<

### Future of Analytics- India Perspective



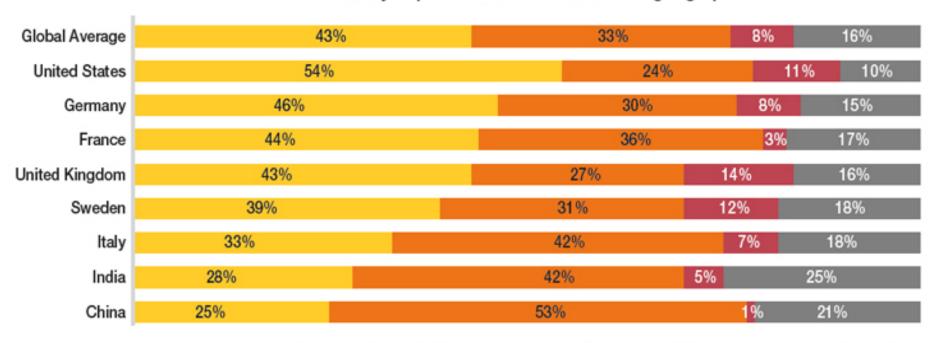
- ☐ "The price of light is less than the cost of darkness"
- ☐ Majority Business Leaders believe Big Data will revolutionize business operations in the same way the internet did.
- ☐ Some of them have pursued Big Data projects in order to seize a competitive edge.



### Status of Indian Industry- Smart Factories



#### Smart factory implementation status across geographies



- Yes we have an ongoing smart factory initiative - (operational)
- No but we plan to have a smart factory initiative in the next 3-5 years
- Yes it is currently being formulated (not in operation)
- No we are not likely to have a smart factory initiative anytime in future

Percentages indicate share of organizations in each industry which reported they have an ongoing smart factory initiative Source: Capgemini Digital Transformation Institute, smart factory survey, February-March 2017

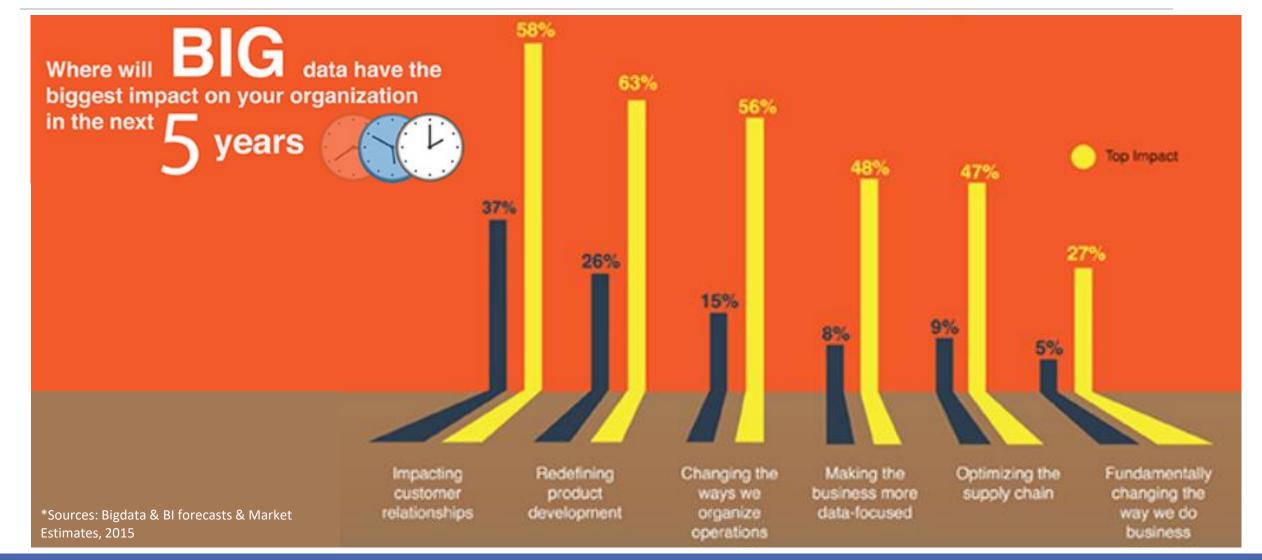
### Future of Analytics- India Perspective



☐ Industry Leaders speaking on analytics

### Future of Analytics- India Perspective





### Reason for early adoption & it's potential

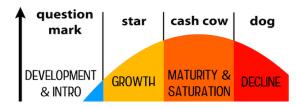


#### **Existing business environment:**

- ☐ Adequate IT Infrastructure
- ☐ Digital Data availability in abundance
- ☐ Short Product Life cycle & Services
- ☐ High priority on process optimization
- Volatility of market requirements
- Requirement of strong customer engagement model
- ☐ Cut throat Market Competition
- ☐ Social Mobility Analytics Cloud Era: Easy availability of customized technology platform



#### PRODUCT LIFE CYCLE



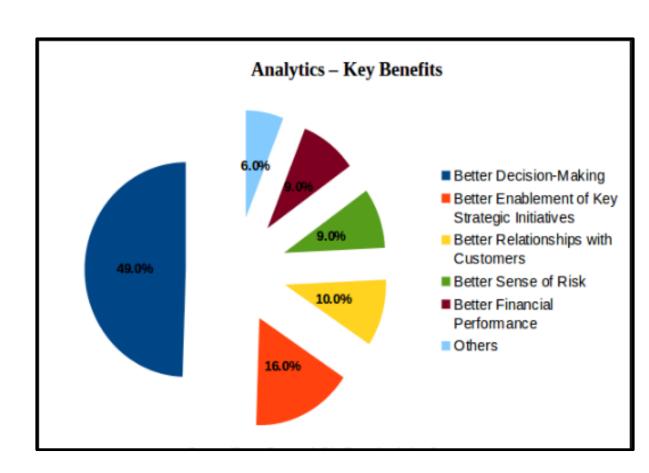


### Reason for early adoption & it's potential



#### **Hidden Potential:**

- ☐ Just in Time actionable Insights
- ☐ Clear visibility to address business risk and failures
- Using existing People & Process it can be easily transformed into potential resources, thus enabling smart operations.
- Integration of industry domain knowledge and technology to solve complex decision-making riddles intelligently and sustainably.
- □ Value driven culture for managing efficiency, productivity and reliability with On-time.
- ☐ Sustainable ecosystem

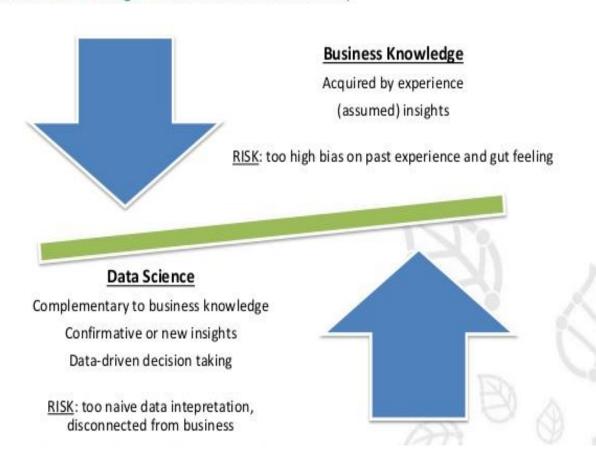


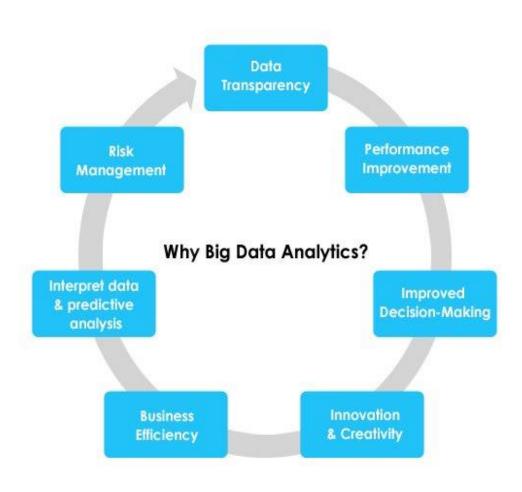
### Experience Based Vs. Data Driven Decision



### Business Knowledge vs Data Science

(Intuitive knowledge vs data driven decisions)





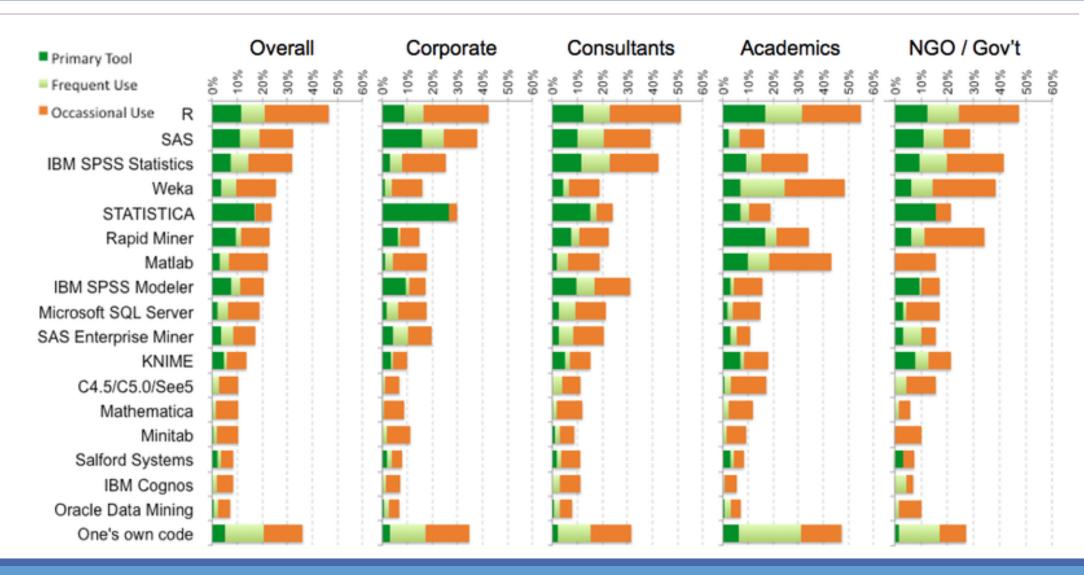


### Usages of tools & Its Comparison

UNDERSTANDING DIFFERENT TECHNOLOGIES

### Different analytics tools





### Features of Analytics Tools





#### Features of R:

- Free & Open Source: Source code can be examined line by line
- Integrating with other programming languages
- Operating System Independence
- Command Line Driven



#### **Features of SAS:**

- Read and write almost any data format
- Powerful Data Handling Capabilities
- Excellent Data Cleansing Functions
- Countless options for outputs (HTML, PDF, Excel etc.)
- Can Interact with Multiple Host Systems



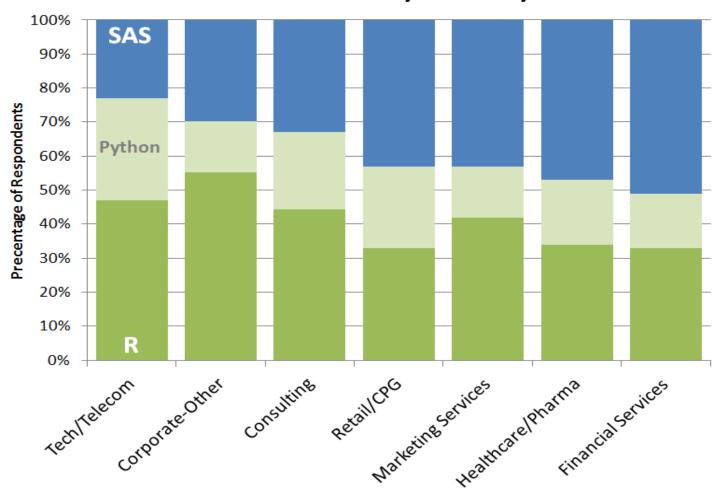
#### **Features of R:**

- Free & Open Source: Freely distributed and Open source Maintained by the Python community.
- Simple
- Easy to Learn
- High Level Language memory management
- Portable \*runs on anything c code will

### Trend Analysis



#### **Tool Preference by Industry**



### Factors to consider while buying BA Tools





### Differentiators



Parameter	SAS	R	Python
Availability/Cost	3	5	5
Ease of Learning	4.5	2.5	3.5
Data Handling Capabilities	4	4	4
Graphical Capabilities	3	4.5	4.5
Advancement in Tools	4	4.5	4.5
Job Scenario	4	4.5	4.5
<b>Customer Service Support and Community</b>	4	3.5	3.5
Deep Learning Support	2	3	4.5
Total	28.5	31.5	34

### Pros & Cons



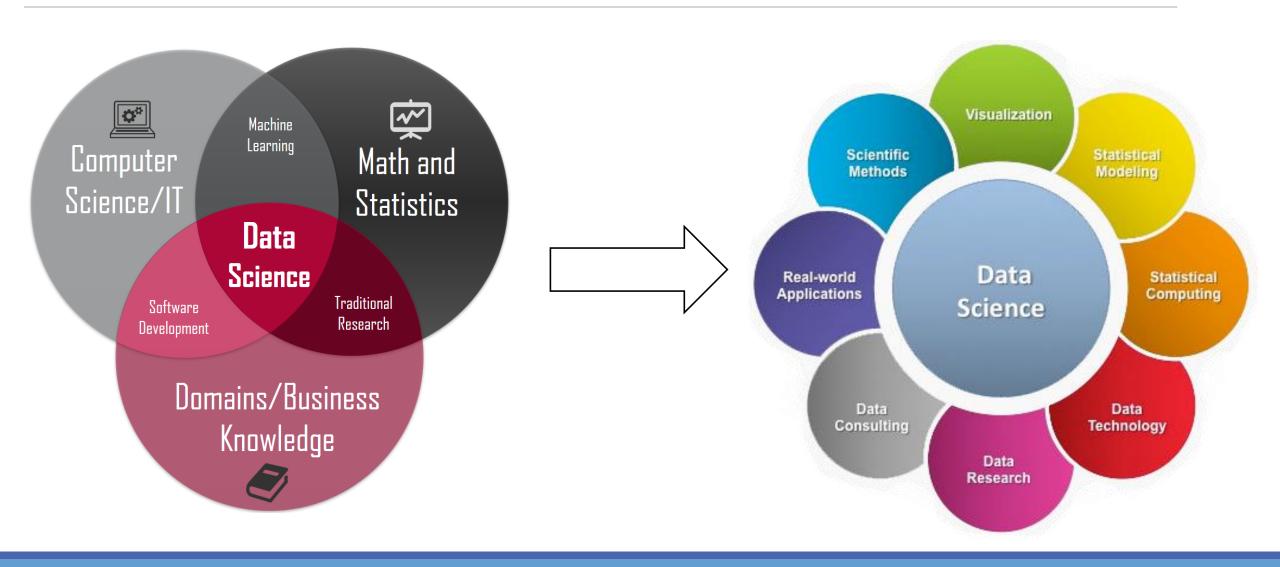
	SAS		R		Python	
Advantages	1. 2.	High adoption rate in major industries Flow based interface with	1. 2.	Big community who creates libraries Free	1. 2. 3.	Scalability General purpose language Easy to learn
	3. 4. 5.	drag and drop Official support Handling large datasets 'PROC SQL'	3. 4.	Early adopter in explanatory and predictive modeling. Easy to connect to data sources, including NoSQL and webscraping.	4. 5. 6.	Good in machine learning Big community Free
Disadvantages	1. 2. 3.	Relatively high cost For not-standard options not in interface, you'll need to write the code Slow adapting to new techniques Different programs for visualization or Data Mining	1. 2. 3. 4.	Can be slow with big datasets Steep learning curve No official support No user interface	1. 2. 3. 4.	Not as strong in explanatory modeling Choice of version: 2.7 or 3.5? No user interface No official support



# Learn basics of Data Science, Its Types, Variables & Charts

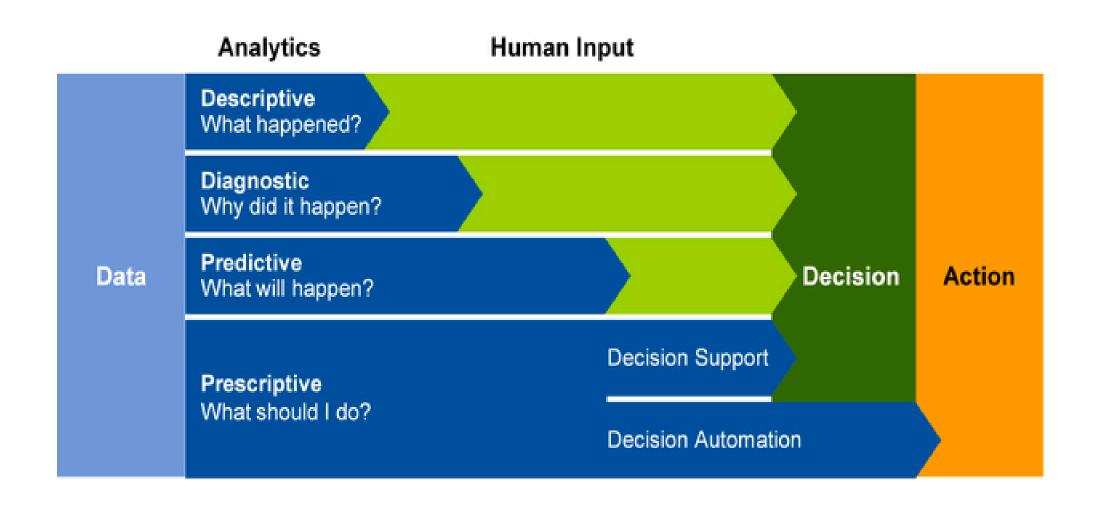
### Introduction to Data Science





### Business Analytics Maturity Stages





### Evolution of Industry















Mechanical production. Equipment powered by steam and water 19th Century

#### Industry 2.0

Mass production assembly lines requiring labor and electrical energy 20th Century

#### Industry 3.0

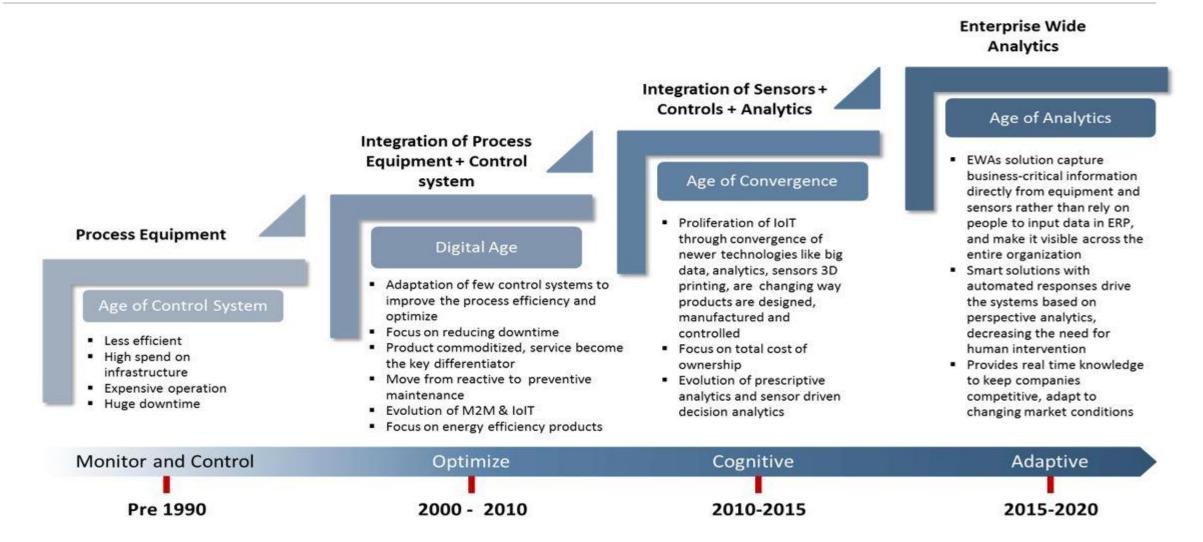
Automated production using electronics and IT Today

#### Industry 4.0

Intelligent production incorporated with IoT, cloud technology and big data

### Evolution of Analytics – Industry Perspective





### Evolution of Analytics – A technology perspective



#### Pre 2005

- 1997, The term "Data Science" was coined
- 1998, The term "Big Data" was coined
- 1999, "loT" was coined
- 2000, R v 1.0 Released
- 2003, The term "Predictive Analytics" was coined by SPSS
- 2003, Tableau Launched
- 2003, Splunk founded
- 2005, Facebook founded

#### 2010

- Kaggle Founded
- EMC buys Greenplum
- IBM buys Netezza
- SAP launches HANA
- Apache Spark Open Sourced

#### 2012

- Facebook reaches 1B users
- Facebook acquires Instagram for \$1B
- IDC Launches Worldwide Business Analytics Software Tracker
- Obama uses predictive analytics for election campaign

#### 2014

- Apache Spark 1.0.0 Released
- KPMG launches Centre for Advanced Business Analytics with imperial college, London (to invest £20M
- Hortonworks \$100M IPO
- Twitter acquires deep learning startup Madbits
- Databricks Inc. raised another \$33 million in funding to commercialize Spark
- Splicemachine launches Hadoop RDBMS

#### 2005 - 2009

- 2005, Hadoop was created
- · 2006, Twitter Launched
- · 2006, Google Analytics Launched
- · 2006, Rapidminer Launched as YALE
- 2006 First version of KNIME released
- 2007 TIBCO acquires Spotfire
- 2007, Competing on Analytics: The New Science of Winning book published
- 2009, Apache Mahout 0.1
- · 2009, Cloudera Hadoop CDH1 launched
- 2009 Netflix award of \$1M to predict movie preferences
- 2009 IBM acquires SPSS for \$1.2M

#### 2011

- RStudio Beta Released
- · Hadoop 1.0.0 Released
- Hortonworks launched as a spin-off from Yahoo
- IBM Watson defeated two of Jeopardy's greatest champions

#### 2013

- · Hadoop 2.x GA Release
- YouTube hits 1B users
- Microsoft Launches Hortonworks Hadoop Service On Windows Azure Cloud
- Google buys Wavii, a Machine Learning startupfor \$30M
- Facebook buys Atlas ad analytics company
- Twitter buys Social TV Analytics Company Bluefin Labs
- Walmart Labs buys Data Analytics, Predictive Intelligence Startup Inkiru

#### 2015

- Jan 2015, Microsoft acquires Revolution Analytics
- Feb 2015, Microsoft launches Azure Machine Learning
- April, 2015 Amazon launches Machine Learning
- May 2015, Accenture Launches Advanced Analytics Applications Platform
- Feb 2015, HP launched Haven Predictive Analytics (an open-sourced big data predictive analytics platform)
- June 2015, Amazon adds Spark to Amazon EMR

### Core fundamentals of Data Science





Data & it's importance

Type of Data

Type of Variables

Data summarization techniques

Data Analysis Approach

### Analyzing the data

Correlation

Linear Regression

**Decision Tree** 

Time Series

## Visual Representation of Data

Selection of right graphical charts

Interpretation of Analysis

Sharing Methods

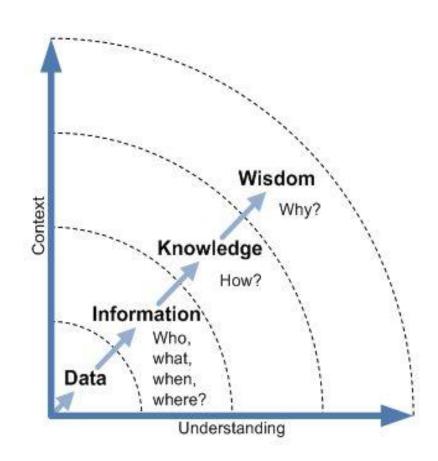
### Data & it's importance



- □ Data is any set of characters that has been gathered and translated for some purpose, usually analysis. It can be any character, including text and numbers, pictures, sound, or video.
- Data is simply another word for information.
- Data helps in making better decisions at Business as well as operational level and saves time.

#### Some interesting Facts about data!

- Every two days we create as much data as we did from the beginning of time until 2003.
- ☐ If you burned all of the data created in one day onto DVDs, you could stack them on top of each other and reach the moon twice.

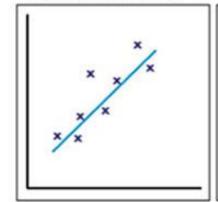


### Statistical Techniques: Correlation



- Correlation refers to a mutual relationship or association between quantities.
- ☐ In almost any business it is useful to express one quantity in terms of its relationship with others.
- For e.g: Sales might increase when the marketing department spends more on TV advertisement or a customer's average purchase amount on an ecommerce website might depend on a number of factors related to that customer.
- Often correlation is the first step to understand these relationships and subsequently building better business and statistical model.

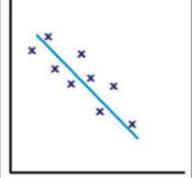
#### Positive correlation



The points lie close to a straight line, which has a positive gradient.

This shows that as one variable increases the other increases.

#### Negative correlation



The points lie close to a straight line, which has a negative gradient.

This shows that as one variable increases, the other decreases.

#### No correlation



There is no pattern to the points.

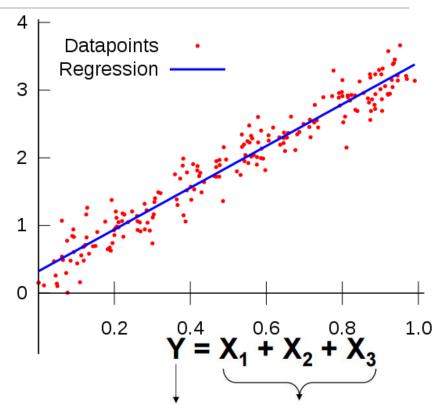
This shows that there is **no** connection between the two variables.

### Statistical Techniques: Linear Regression



☐ Linear Regression a method to predict a target variable by fitting the best linear relationship between the dependent and independent variable.

□Linear Regression is the simple and widely used supervised machine learning algorithm for predictive analysis.



Dependent Variable

Independent Variable

Outcome Variable

Predictor Variable

Response Variable

**Explanatory Variable** 

### Statistical Techniques: Linear Regression



- Linear Regression can also be used to assess risk in financial services or insurance domain.
- □ For E.g: In the Finance industry financial company may be interested in minimizing the risk portfolio and wants to understand the top 5 factors that cause a customer to default. Based on the results the company could implement specific EMI option so as to minimize default among risky customers.

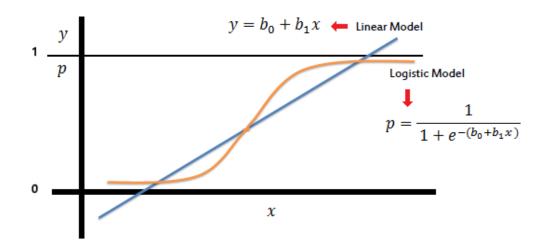


If she loves you more each and every day, by linear regression she hated you before you met.

### Statistical Techniques: Logistic Regression



- Logistic Regression is used when response variable is categorical in nature. For instance, Yes/No, True/False, Red/Green/Blue, 1st/2nd/3rd/4th etc.
- In logistic regression the dependent variable is binary, and the purpose of the analysis is to assess the effects of multiple independent variables, which can be numeric and/or categorical, on the dependent variable.



### Statistical Techniques: Decision tree



fair good excellent

Yes

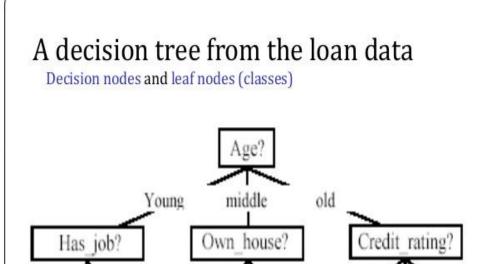
Yes

☐ A decision tree is a natural and simple way of inducing following kind of rules :

If (Age is x) and (income is y) and (family size is z) and (credit card

spending is p) then he will be approved the loan.

- ☐ It is powerful and perhaps most widely used modeling technique of all.
- ☐ Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance.



true false

No

Yes

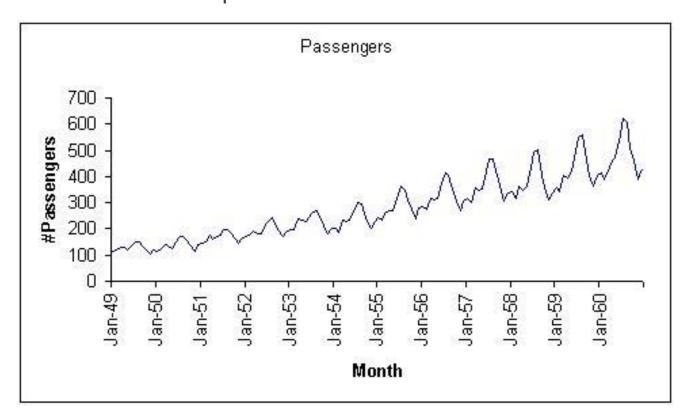
(3/3)

(3/3)

### Statistical Techniques: Time Series



- ☐ Time series data is a sequence of observations collected from a process
- with equally spaced periods of time.
- **□** Examples
  - ☐ Dow Jones Industrial Averages
  - ☐ Daily data on sales
  - ☐ Monthly inventory
  - ☐ Monthly interest rates, costs
  - ☐ Forecasting power consumption
  - ☐ Daily closing prices of stock indices, and so on



### Components of a Time series



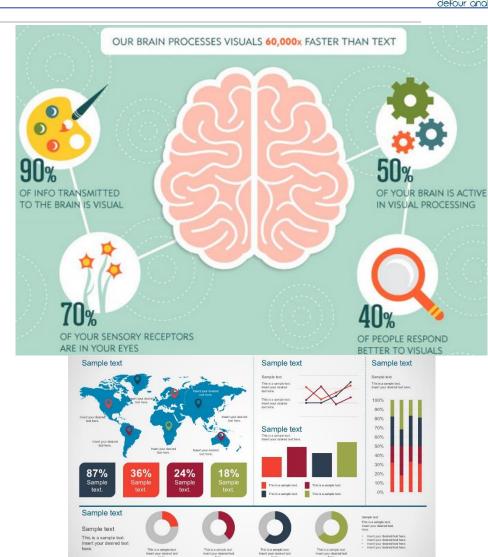
#### 1. Secular Trend(T):

- Gradual long term movement(up or down). Easiest to detect
   e.g. Population growth In India
- 2. Cyclical Patterns(C):
  - Results from events recurrent but not periodic in nature. An up-and-down repetitive movement in demand.
     repeats itself over a long period of time
    - e.g. Recession in US Economy
- 3. Seasonal Pattern(S):
  - Results from events that are periodic and recurrent in nature. An up-and-down repetitive movement within a trend occurring periodically. Often weather related but could be daily or weekly occurrence
    - e.g. Sales in festive seasons
- 4. Irregular Component(I):
  - Disturbances or residual variation that remain after all the other behaviours have been accounted for. Erratic movements that are not predictable because they do not follow a pattern
    - e.g. Earthquake

### Representation of Data - Visualization



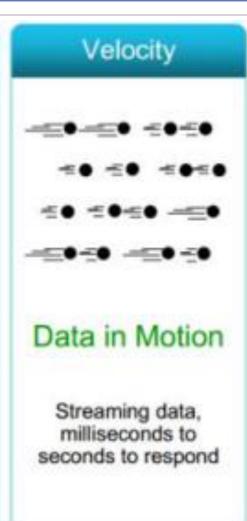
- Words are processed by our short term memory where we can only retain about 7 bits of information (plus or minus 2). Images, on the other hand, go directly into long-term memory where they are indelibly etched.
- ☐Graphical excellence consists of complex ideas communicated with clarity, precision and efficiency.

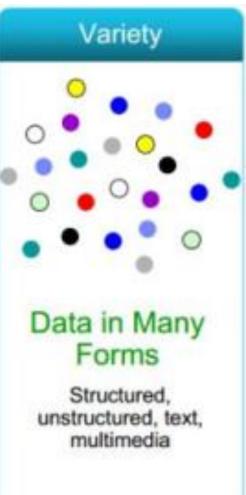


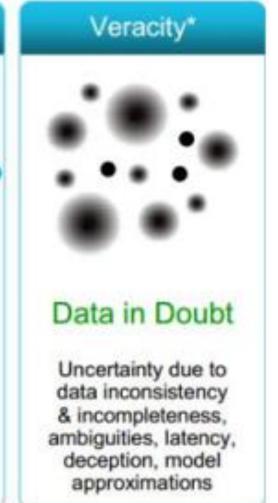
## Visualizing Big Data – 4 V's











http://www.intergen.co.nz/blog/Tim-Mole/dates/2013/8/musings-from-a-big-data-conference/

# Choosing the Right Chart type with Problem Statement

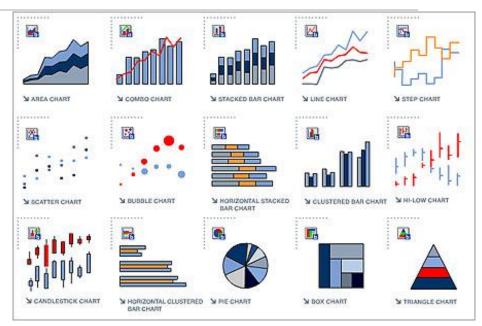


### **Selecting the Right Chart**

- •How many variables do you want to show in a single chart?
  One, two, three, many?
- •How many items (data points) will you display for each variable?
  Only a few or many?
- •Will you display values over a period of time, or among items or groups?

### **Answer**

- Bar charts are good for comparisons, while line charts work better for trends.
- Scatter plot charts are good for relationships and distributions.
- Pie charts should be used only for simple compositions, never for comparisons or distributions.





## Charts Categories



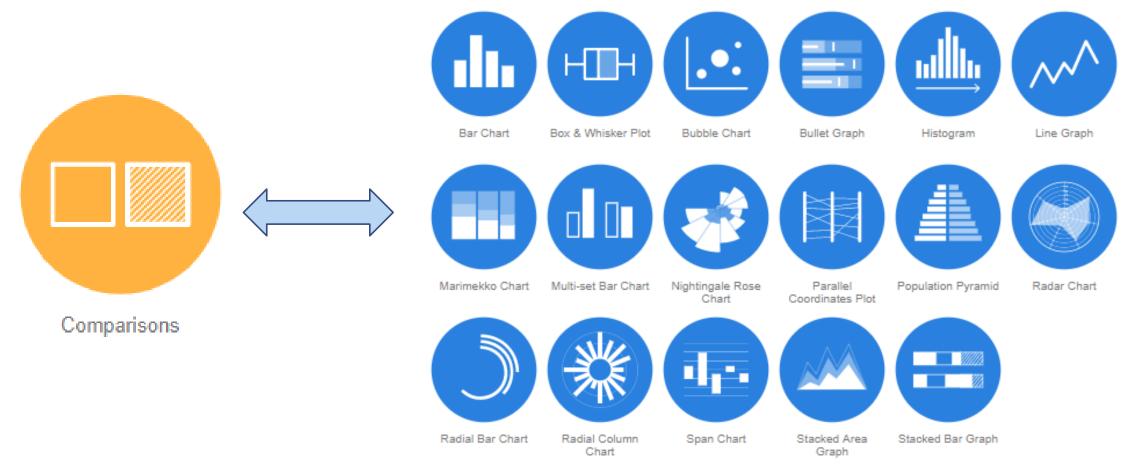


## Visualization Charts: Comparative analysis



**Visualisation** methods that help show the differences or similarities between values.

E.g: Yearly performance, Product Comparison etc.



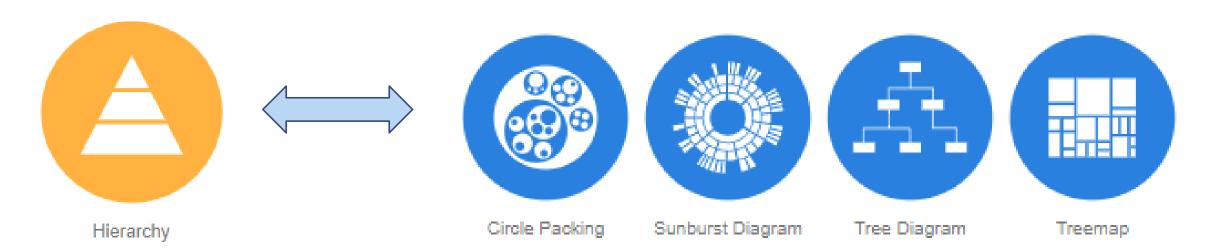
## Visualization Charts: Hierarchy



#### Hierarchy

Visualization methods that show how data or objects are ranked and ordered together in an organisation or system.

E.g. Organization Structure, Distribution Network etc.



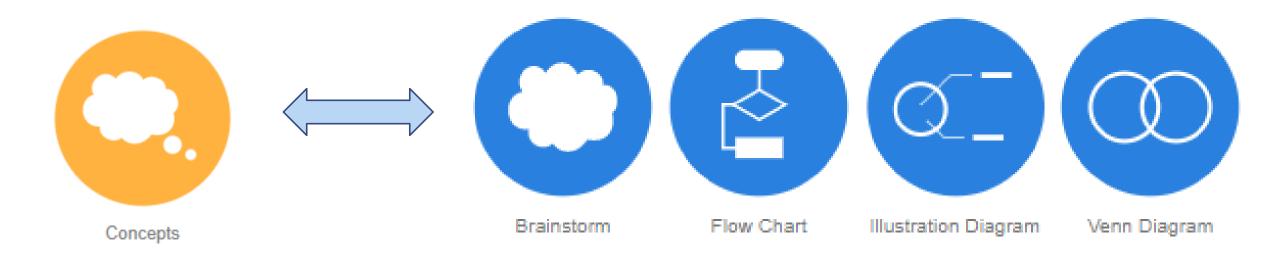
## Visualization Charts: Concepts



#### **Concepts**

Visualisation methods that help explain and show ideas or concepts.

E.g. Possible logical relationship between collection of sets, Anatomy



## Visualization Charts: Comparative analysis



#### **Processes and Methods**

Visualisation methods that help explain processes or methods.

E.g: SOP's, Project Management etc.



### Tools for Data Visualization





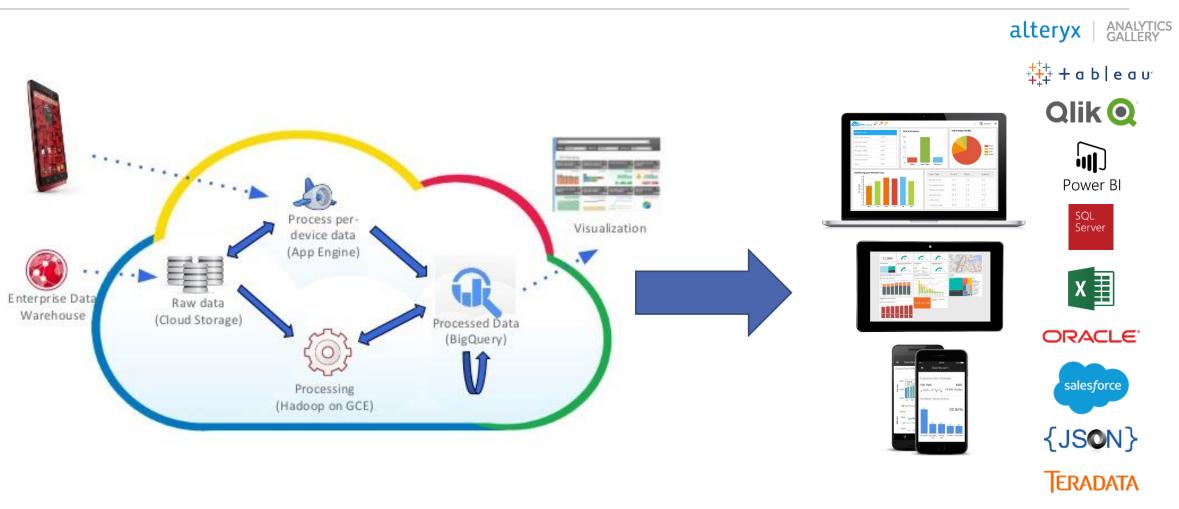






## Methods of sharing charts





**Output All Popular Formats** 



## Introduction to R tool & Its functionality



#### 1. Ease of Use

- •Most people have likely already learned at least a few basic tips in Microsoft Excel.
- ■That's what's great about Excel the initial learning curve is quite easy.
- Most can be done via point-and-click on the top panel.
- Once you bring your data in, it's not too hard to make a few basic graphs or charts
- R is a programming language,
- •meaning the initial learning curve is a bit difficult.
- It will take a few weeks to get used to the interface and to remember the different functions, but it becomes second-nature once you're accustomed to it.



#### 2. Visualization

- In Excel, you can quickly highlight a group of cells and make a simple chart for PowerPoint
- R can make some incredibly attractive
- Detailed visuals that will blow someone away.
- •For those planning to publish your visuals to the public to really highlight your big data
- Then spending a bit more time in R would be worth it.



#### 3. Statistical Analysis

- •If you just want to run some quick statistics and arithmetic
- Then Excel might be better for you since it's an easy point-and-click away to run numbers.
- Though, it's not hard to get that same info from R, either.
- •R programming can do a *lot* of analysis and is great for identifying trends that you might not have thought to look for.
- Even deciding how reliable said statistics are. R allows you ways to clean and organize data, gives more visualization options, and if there's a topic you want to explore, then there's likely a way to do it in R.



#### 4. Careers

- Both are incredibly valuable skills that are in-demand across a variety of industries.
- •Countless jobs are looking for applicants with at least some Excel experience (note: pivot tables look *really* good on a resume), but R has a higher earning potential, and is more in-demand than Excel.
- R is now the most popular programming language
- •Making it an industry standard for data analytics and data science.
- •If you want to enter either field, there's a good chance you'll need to know R.
- ■Entry-level jobs for those focusing on R also tend to make a high salary, frequently starting off with \$75,000+.

### SUMMARAIZING R AND EXCEL



- R and Excel are beneficial in different ways.
- •Excel starts off easier to learn and is frequently cited as the go-to program for reporting, thanks to its speed and efficiency.
- R is designed to handle larger data sets, to be reproducible, and to create more detailed visualizations.
- It's not a question of choosing between R and Excel, but deciding which program to use for different needs.
- If you're looking to get a taste of data analytics, then our Set program is great for giving you an introduction to Excel and MySQL.
- •For those looking to tackle advanced analytics, our Core program teaches R, among other software.
- •Consider taking one of our data analytics programs and become a data pro!

### Introduction to R'



R is "GNU S" — A language and environment for data manipulation, calculation and graphical display.

- R is similar to the award-winning S system, which was developed at Bell Laboratories by John Chambers et al.
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for interactive data analysis,
- graphical facilities for data analysis and display either directly at the computer or on hardcopy
- a well developed programming language which includes conditionals, loops, user defined recursive functions, input and output facilities.

#### The core of R is an interpreted computer language.

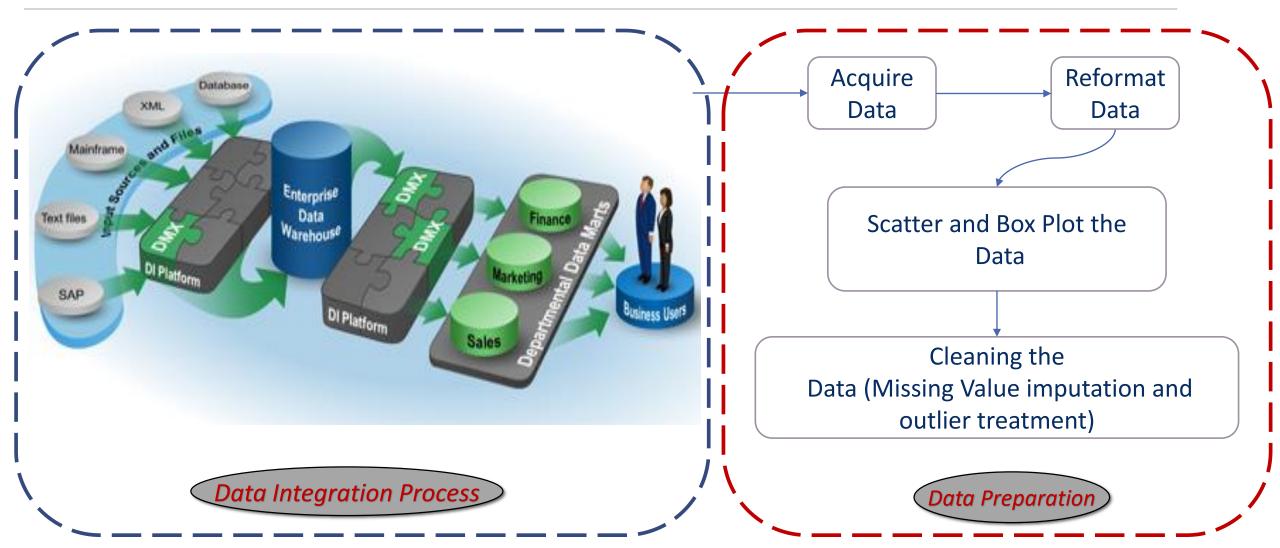
- It allows branching and looping as well as modular programming using functions.
- Most of the user-visible functions in R are written in R, calling upon a smaller set of internal primitives.
- It is possible for the user to interface to procedures written in C, C++ or FORTRAN languages for efficiency, and also to write additional primitives.



## Learn How to use & apply data analytics

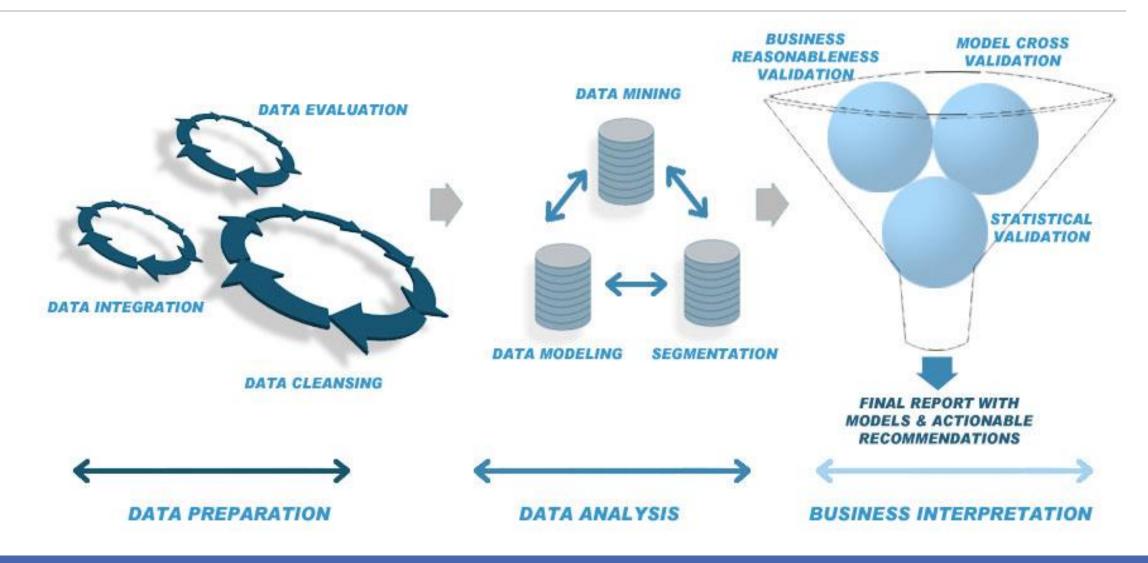
## Data Integration and Preparation





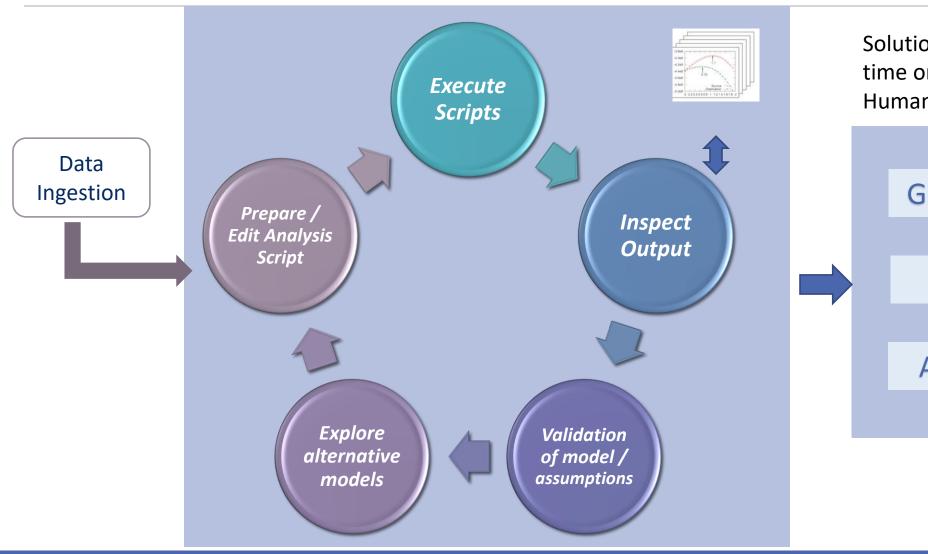
### Data Analytics Flow - Overview





## Model Building & Run Time Application





Solution can run Real time, Near Real time or on Scheduled basis without Human intervention

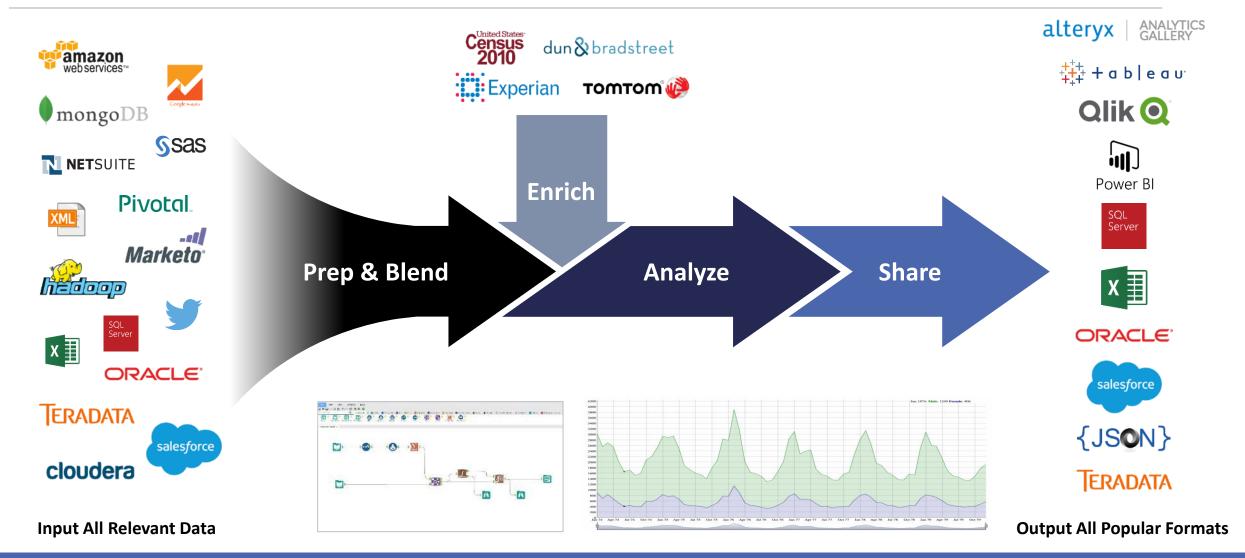
**Graphical Dashboard** 

**Predictive alerts** 

**Actionable Insights** 

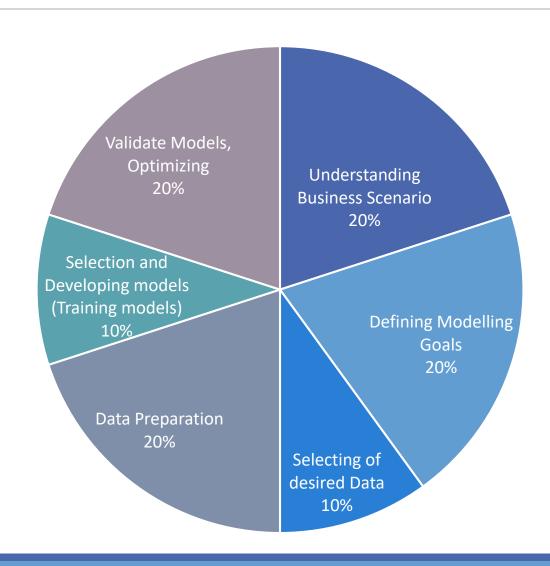
## Analytics Process





### Standard Timeline in Model Building









## Step Towards Digital Transformation



