# Introduction to  Data Science

Overview

# Objective

After completing this lesson you will be able to:

- Describe business analytics
- Explain the components of business analytics
- Explain the usage of business analytics in various domains

In God we trust, all other must bring data

- W Edward Deming

# Corporate Decision Making–The HIPPO Algorithm

Highest Paid Person’s Opinion

# Business Analytics–Definition

- Business analytics (BA) refers to the tools, techniques and processes for continuous exploration and investigation of past data to gain insights and help in decision making.

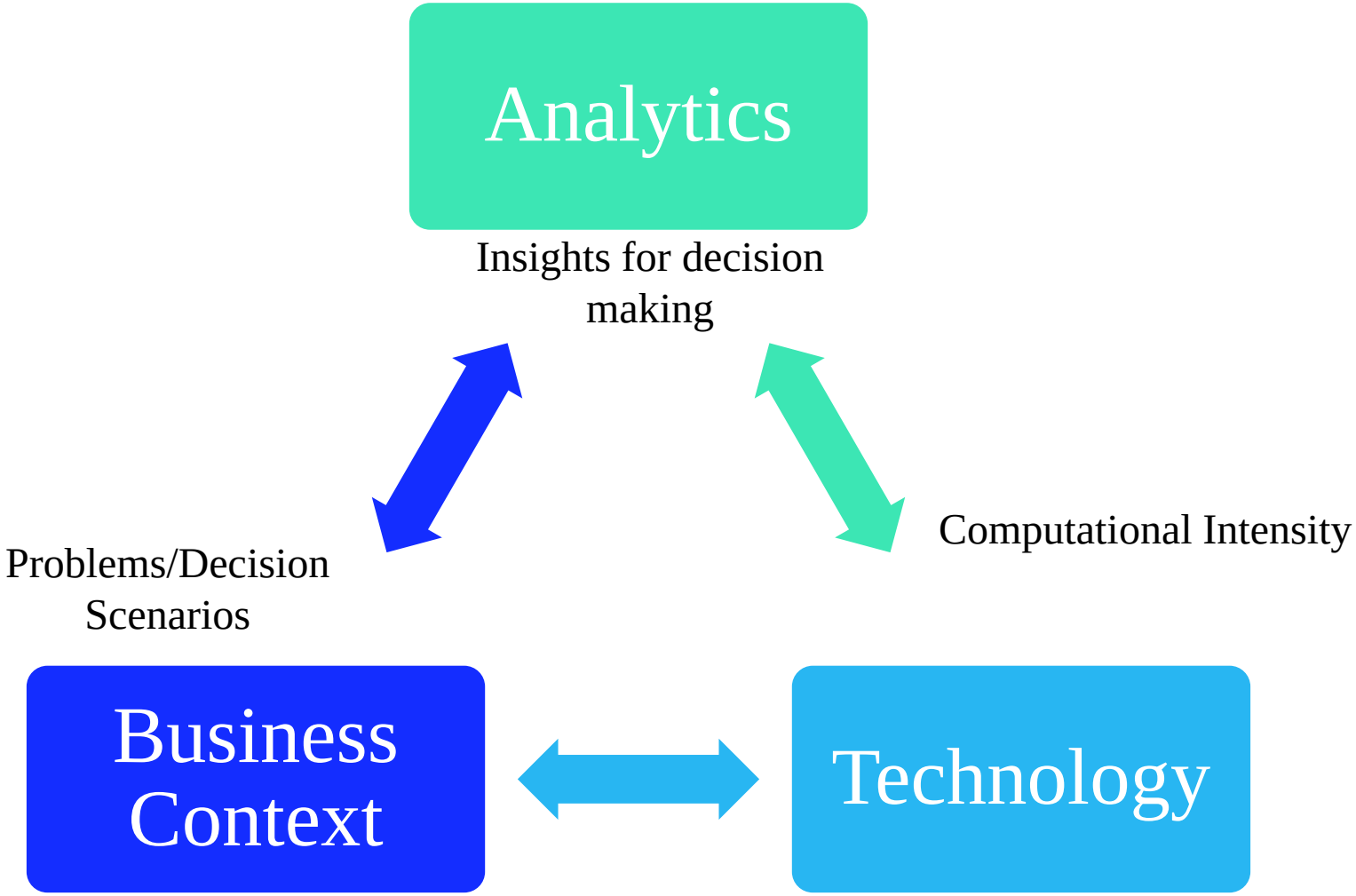
- Business Analytics is an integration between science, technology and business context that assist data driven decision making.

# Data Explosion

- About 350 million  photos are uploaded every day in the Facebook
- Amount of credit card debt in US: $762.1 billion
- Amount of credit card debt in India: Rs. 45,383 crore ($709 million)
- Loss due to global Credit card and debit card fraud $21.84 billion during 2015
- Every day, Walmart processes $36 million dollars an hour in sales
- BMTC with approx. 6000 buses plying in Bangalore sends 1 billion signals to the server updating its location every month

Interesting Stats: http://expandedramblings.com/

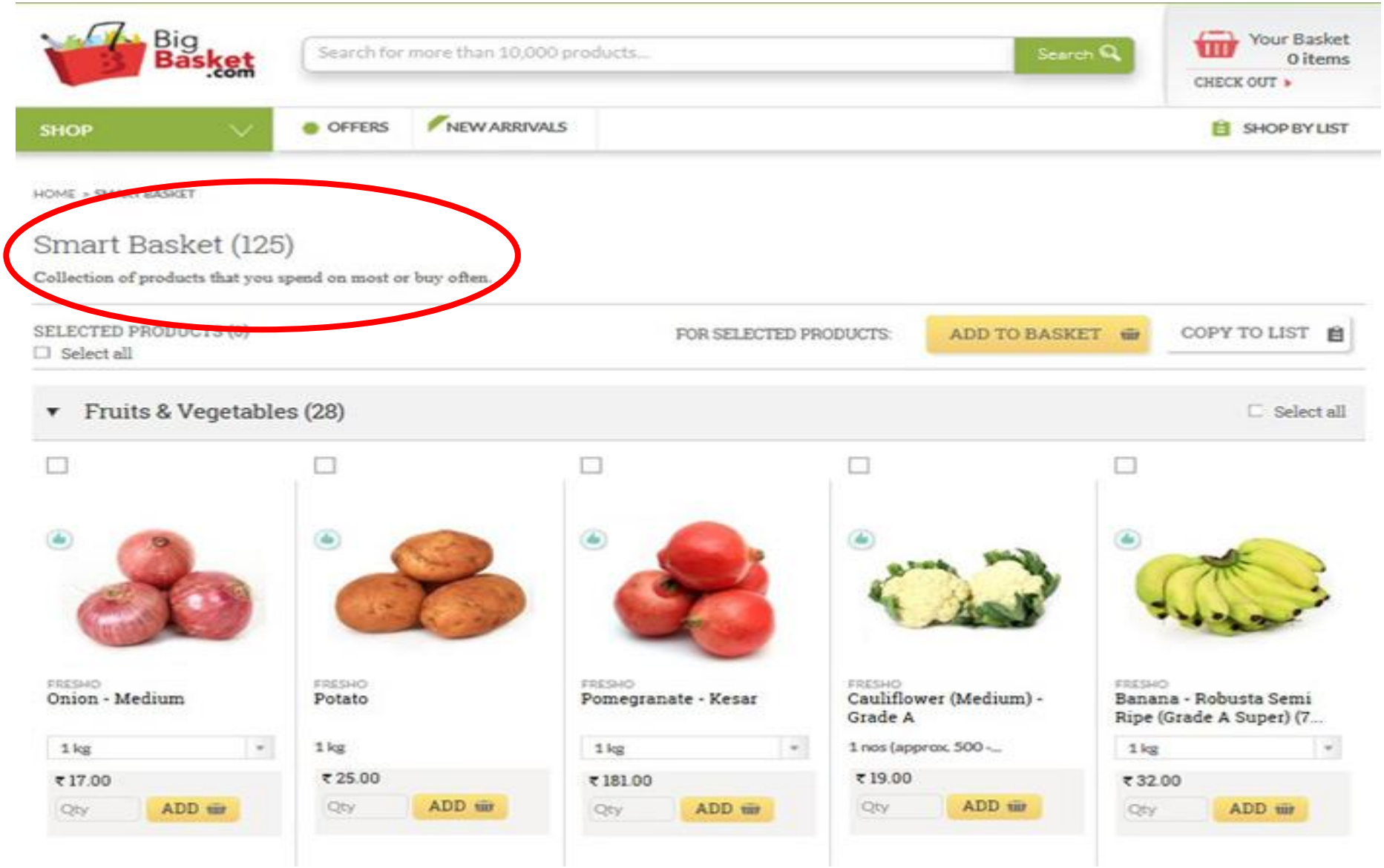
# Analytics Trilogy



# Analytics in Use–Flipkart

- Forecast demand for each SKU.
- Predict customer cancellations and returns.
- Predict customer contacts at the customer service.
- Predict what a customer is likely to purchase in the future?
- How to optimize the delivery system?

# Analytics in Use–Big Basket



# How would you solve this?

6/10/2015 Flipkart delivers 2 stones instead of iPod to a user! What if...

Trak.in
Business of Tech, Mobile & Startups in India

It's gone. Undo
What was wrong with this ad?
Repetitive Inappropriate Irrelevant

HOME | BUSINESS | TECHNOLOGY | INTERNET | TELECOM | MOBILE | STARTUP | OTHERS | ABOUT

Home / Internet / Ecommerce / Flipkart delivers 2 stones instead of iPod to a user! What if...

## Flipkart delivers 2 stones instead of iPod to a user! What if...

Posted by: Arun Prabhudesai In Ecommerce, India March 11, 2013 46 Comments

Now, this is outrageous – A twitter user tweeted today that his sister had ordered an iPod, but to her surprise she was delivered 2 stones inside the box.

Twitter user with handle @nikhilsekhar tweeted the following accompanying picture of the said delivery.

> @flipkart My sister got two stones instead of the INR 20K iPod that she ordered from Flipkart. What is wrong with you twitter.com/nikhilsekhar/s...
>
> — Nikhil (@nikhilsekhar) March 10, 2013

Here is the photo accompanying that tweet:

The good part is Flipkart promptly replied the user with following tweets:

> We are very sorry about the incident with our customer's iPod purchase, we're taking up this issue extremely seriously. 1/3
>
> — Flipkart (@Flipkart) March 10, 2013

> We have spoken to our customer and will make sure a replacement is sent over right away. 3/3
>
> — Flipkart (@Flipkart) March 10, 2013

### But what if...

Now, this kind of a thing happening is outrageous – and it is good to see that Flipkart has owned the responsibility and is going to offer an immediate replacement of the product.

However, I have a question in mind – What if the customer is lying?

Please bear in mind, I am not talking about this incidence, but putting across a theoretical situation.

What if a buyer actually replaces the delivery (after he has received the correct product) and then alleges that Flipkart has delivered him with stones (or whatever) inside it? What happens in that case. If you think that such kind of thing will not happen...you are wrong. There are many out there waiting to take advantage of the system..

Flipkart is in no position to actually contend that they have delivered it correctly, neither can they challenge the customer (and if they do, it will be one big social media mess).

OH...IS IT?

you better know this... you were working in my 17th start up which is now acquired by my 11th start up... so...you are still my EMPLOYEE!!!

FOLLOW US
45.6k Follows
RSS Feed 23.8k Followers
Facebook 16.4k Followers
Twitter 3.4k Followers
LinkedIn 0 Followers
Pinterest 775 Followers
Google+ 1k Followers

FIND US ON FACEBOOK

http://trak.in/tags/business/2013/03/11/flipkart-delivers-stones-instead-ipod/ 1/10

http://trak.in/tags/business/2013/03/11/flipkart-delivers-stones-instead-ipod/

# How would you solve this?

Man orders Oppo phone on Amazon, gets dummy iPhone instead

# Decision Making–The Monty Hall Problem


After having seen “What lies besides door 1”, Would you like to switch?

# The Game Changers

- Google
  - Used Markov chains to rank pages.
- Proctor and Gamble
  - Analytics as competitive strategy.
- Target
  - Predicts customer pregnancy.
- Capital One
  - Identifies the most profitable customer.
- Hewlett Packard
  - Developed “flight risk score” for 3,30,000 employees.
- Obama’s 2012 presidential campaign.
  - Persuasion Modelling.

# The Innovators

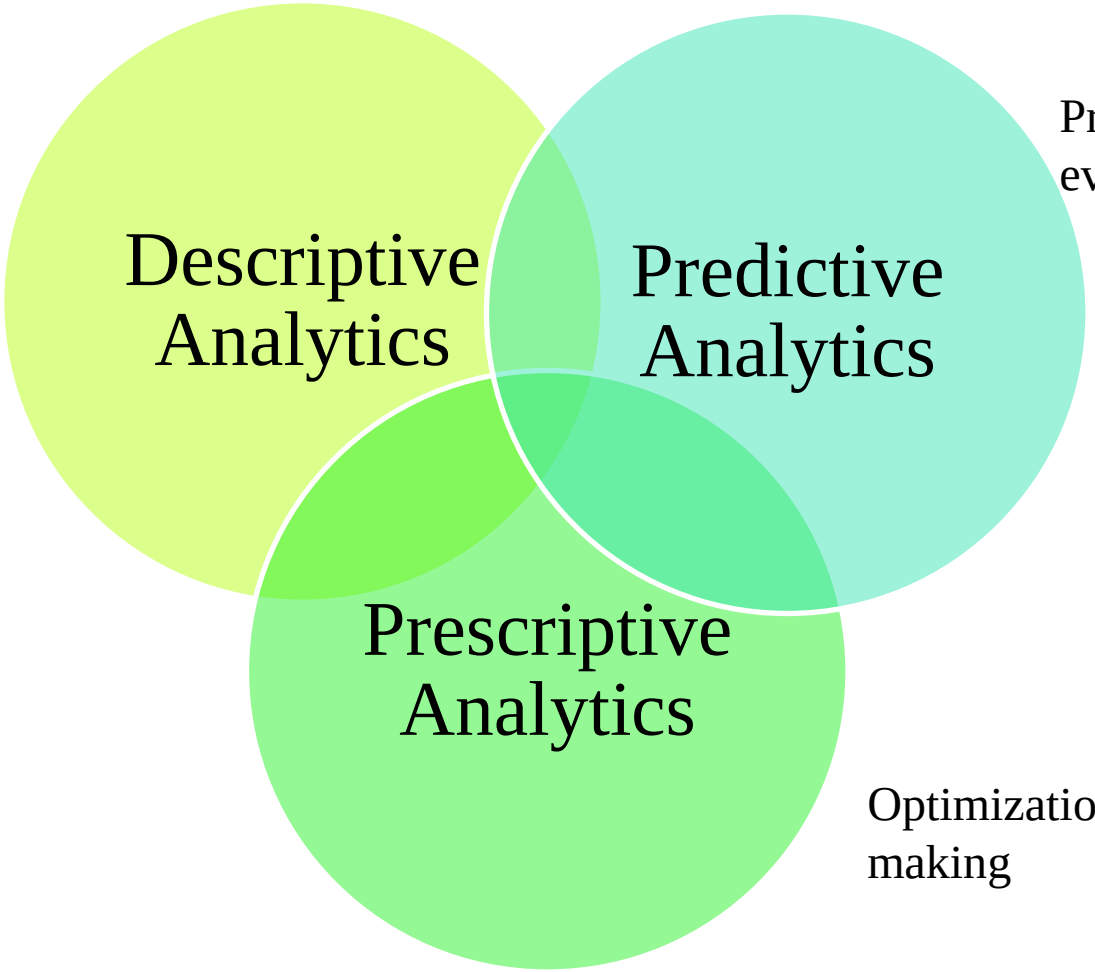
- OKCupid: Predicts which online dating messages is most likely to get a response!
- Polyphonic HMI: Uses “hit song science” to predict commercial success of a song.
- Netflix: Predicts movie ratings by customers (RMSE is 1%).
- Amazon.com: 35% of sales come from product recommendations.
- Citizens Bank: Predicted fraudulent cheques.
- Divorce360.com: Predicting success of a marriage!

Data Scientists will be the sexiest job of 21st century

Harvard Business Review 2012

# Components of Business Analytics

Data synthesis and Visualization




Predicting future events

Optimization and decision making

# Components of Business Analytics

Knowing what happened in past and what may happen in future, what optimal strategy can be adopted to achieve an objective like maximize profit.

Learning from past data and predicting what may happen in future and likelihood of happening in future.

Prescriptive

Optimal product pricing or product mix strategies.

Understanding what happened and why happened by exploring past data.

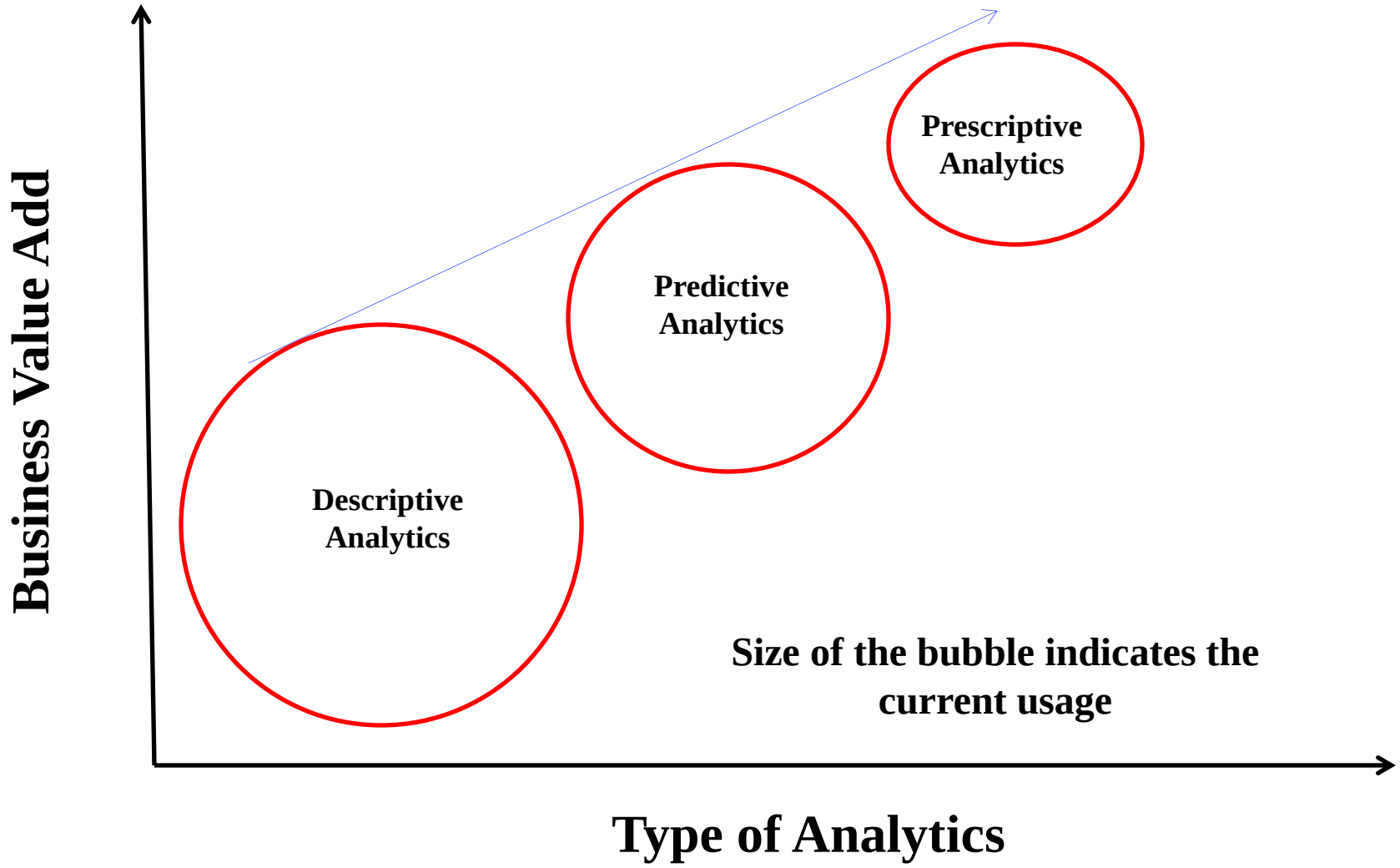
Predictive

Product sales or revenue forecast.

Descriptive

Product sales patterns or factors influencing product sales.

# Business Analytics & Intelligence



# Power of Descriptive Analytics

# Descriptive Analytics Applications

- Most shoppers turn towards right when they enter the a retail store.
- Conversion rate of women shoppers is higher than male shoppers among electronic gadgets purchasers (Radio Shack).
- Strawberry pop-tarts sell 7 times more during hurricane compared to regular period (Wal Mart).
- Women car buyers prefer women sales person.

# Predictive Analytics Application

- Which product the customer is likely to buy in his next purchase (recommender system).
- Which customer is likely to default in his/her loan payment.
- Who is likely to cancel the product that was ordered through e-commerce portal.

# Prescriptive Analytics Application

- What is the optimal route for a delivery truck.
- Whether a company should introduce a new product?
- What is the optimal product mix?
- How to manage the fleet of vehicles owned by a company for employee drop and pick up?

# Framework For Decision Making

## Opportunity Identification

- Domain knowledge is very important at this stage of the analytics project. This will be a major challenge for many companies who do not know the capabilities of analytics.

## Collection of relevant data

- Once the problem is defined clearly, the project team should identify and collect the relevant data. This may be an interactive process since "relevant data" may not be known in advance in many analytics projects. The existence of ERP systems will be very useful at this stage.

## Data Pre-processing

- This would include data imputation and the creation of additional variables such as interaction variables and dummy variables in the case of predictive analytics projects.

## Model Building

- Analytics model building is an iterative process that aims to find the best model. Several analytical tools and solution procedures will be used to find the best analytical model in this stage.

## Communication of the data analysis

- The communication of the analytics output to the top management and clients plays a crucial role. Innovative data visualization techniques may be used in this stage.

# Industry Wide Application of Analytics

## Manufacturing

- *Supply chain analytics*
- *Quality and Process improvement*
- *Revenue and Cost Management*

## Retail

- *Assortment Planning*
- *Promotion Planning*
- *Demand Forecasting*
- *Market Basket Analysis*
- *Customer Segmentation*

## Healthcare

- *Clinical Care*
- *Hospitality related data*

## Service

- *Demand Forecasting*
- *Service Quality Analysis*
- *Customer Segmentation*
- *Promotion*

## Banking and Finance

- *Service Demand Analysis*
- *Customer Transaction Analysis*
- *Credit Scoring*

## IT and ITES (IT enabled services)

- *Demand for Analytics Services*
- *Software Development Cycle Time*

# Statistical learning and Machine learning

Useful read: 50 years of Data Science

# Statistical Learning vs. Machine Learning

## Breiman's `Two Cultures', 2001

*"... Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables x (independent variables) go in one side, and on the other side the response variables y come out. Inside the black box, nature functions to associate the independent variables with the response variables ..."*

There are two goals in analyzing the data:

- Inference: to infer how nature is associating the response variable to the input variable
  - Inference based modelling (Generative Modelling), tries to develop model which maximizes the chance of observing the data.
  - Trying to understand how age, gender, past medical history effects the cost of treatment of a disease.
  - By how much cost of treatment will go up for a 35 year old person compared to a 36 year old person.

# Statistical Learning vs. Machine Learning

- Prediction: to predict (discriminative modelling) what the responses are going to be to future input variables.
  - Silent about the underlying mechanism generating the data, and allows for many different predictive algorithms, preferring to discuss only accuracy of prediction made by different algorithm on various datasets
  - Using age, gender, past medical history to predict the cost of treatment of a disease. Even if the feature weights does not necessarily help in the right inference.

Users of the data are split into one of the two cultures based on interest and objective they try to achieve with data

# What is Machine learning

## Common Task Framework is a crucial but unappreciated methodology driving predictive modeling's success

CTF has these ingredients:

- Training data set
- Models whose task is to do class prediction using training data
- Scoring (test set) on which prediction accuracy is reported.

Common Task Framework is the single idea from machine learning and data science that is most lacking attention in statistical training.

Combination of a Predictive Modeling culture together with CTF is the `secret sauce' of machine learning.

# Statistics vs. Machine Learning – Debate continues

Machine Learning is a glorified statistics but if all we care about is prediction, why bother using a probability model at all?

**Glossary**

| Machine Learning | Statistics |
|---|---|
| Network, Graphs | Models |
| Weights | Parameters, Coefficients |
| Features | Attributes, Variables |
| Learning | Fitting |
| Generalization<br>(Bootstrapping, Cross validation) | Test of Performance<br>(Sampling, Hypothesis testing, p-value) |
| Supervised Learning | Regression/Classification |
| Unsupervised Learning | Density Estimation, Clustering |
| **ML sounds like it's young, vibrant, interesting to learn, and growing; Stats does not.** | |

Blame statistics for not marketing its ideas well enough, or blame CS for ignoring statistics.

# Broad Classification of Machine Learning Algorithms

- Supervised Learning
  - Input (X's) and Output (Y) both are known features
- Unsupervised Learning
  - Input (X's) is known but Output (Y) is unknown
- Reinforcement Learning
  - Input (X's) is unknown but Output (Y) is unknown
  - Misspell "avaible" in doc
- Evolutionary Learning
  - evolutionary algorithm (EA) is a subset of evolutionary computation, a generic population-based metaheuristic optimization algorithm. An EA uses mechanisms inspired by biological evolution, such as reproduction, mutation, recombination, and selection.

What Tools are available?

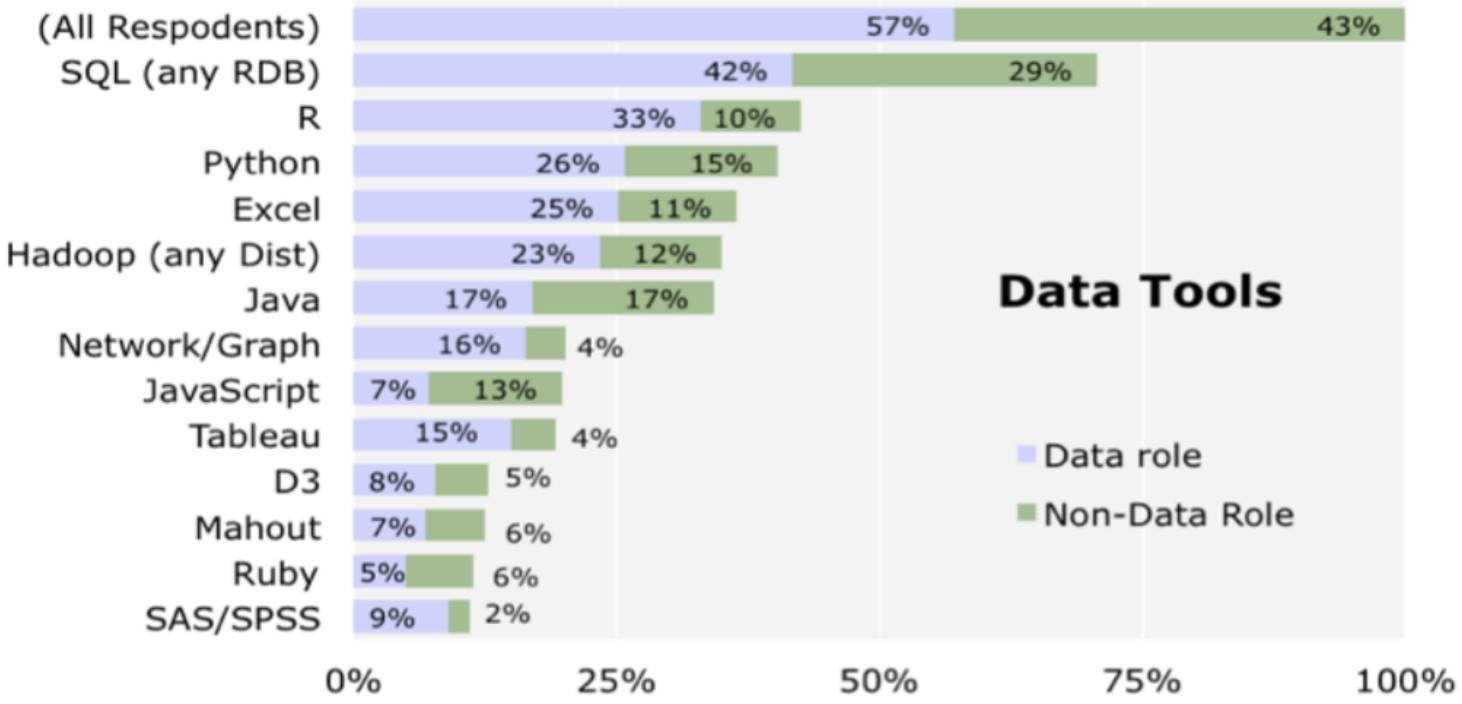
# R Vs. Python

| R | Python |
|---|---|
| Built for Statistical Analysis. | General Purpose Language. Main objective is productivity and readability. |
| Primarily used in academics and research. Enterprise have started adopting it for analysis. | Has a very strong presence in enterprises for large number of software developments. Easier adoption in enterprises as strong development experience already exists. |
| Integration with other enterprise systems are not straightforward. | Integration with other enterprise systems or applications are easier. |

# Python

- Multi-purpose
  - Web Developments
  - Scripting
  - Server Side Developments
  - Statistical Learnings & Machine Learnings
- Object Oriented
- Interpreted
- Strongly typed and Dynamically typed
- Focus on readability and productivity

# Python Stack For Data Science




http://blog.revolutionanalytics.com/2014/01/in-data-scientist-survey-r-is-the-most-used-tool-other-than-databases.html

# Python Stack For Data Science

Efficient storage of arrays and matrices. Backbone of all scientific calculations and algorithms.

Library for scientific computing. Linear algebra, statistical computations, optimization algorithm.

Plotting and visualization

NumPy

SciPy

seaborn

matplotlib

IP[y]: IPython Interactive Computing

pandas

scikit learn

High-performance, easy-to-use data structures for data manipulation and analysis. Pandas provide the features of dataframe, which is very popular in the area of analytics for data munging, cleaning & transformation.

IDE or Development environment for data analysis in python.

Machine learning library. Collection of ML algorithms.

# Python Distribution

ANACONDA®

| | | | | | |
|---|---|---|---|---|---|
| NumPy | SciPy | Pandas | Scikit-learn | Jupyter/ IPython | |
| Numba | Matplotlib | Spyder | Numexpr | Cython | Theano |
| Scikit-image | NLTK | NetworkX | IRKernel | dplyr | shiny |
| ggplot2 | tidyr | caret | nnet | And 330+ packages | |

**conda**

**Game-Changing Enterprise Ready Python Distribution**

- 2 million downloads in last 2 years
- 200k / month and growing
- conda package manager serves up 5 *million* packages per month
- Recommended installer for IPython/Jupyter, Pandas, SciPy, Scikit-learn, etc.

Download link:
https://www.continuum.io/downloads

Source: Continuum Analytics

# Start Jupyter notebook

- For MAC
  - Click on Anaconda Navigator and click on “launch notebook”
  - Or go to command prompt and enter
    - **jupyter notebook --ip=***

- For Windows
  - Go to Anaconda command prompt and enter
    - **jupyter notebook --ip=***

# Start a jupyter notebook


**Click on new to start  new notebook. For every hands on exercise, start a new notebook.**

# Numpy and Pandas

# NumPy

- Library for mathematical and numerical routines like Matlab
- Provides basic routines
  - Manipulating large arrays and matrices of numeric data.
- Foundational library for all statistical and machine learnings
  - Pandas and SciPy
- Using NumPy library

  *import numpy as np*

# Pandas

- Recent API based on Numpy, Optimized for performance
- Easy to work with messy and irregularly indexed data
- Adopts concepts of R language dataframes
- The two basics structures of pandas
  - Series 1d array
  - DataFrame 2d array
- Typical Data Munging Activities
  - Filtering, selecting data
  - Aggregating, transforming data
  - Joining, concatenating, merging data
  - Descriptive basics statistics

# Pandas

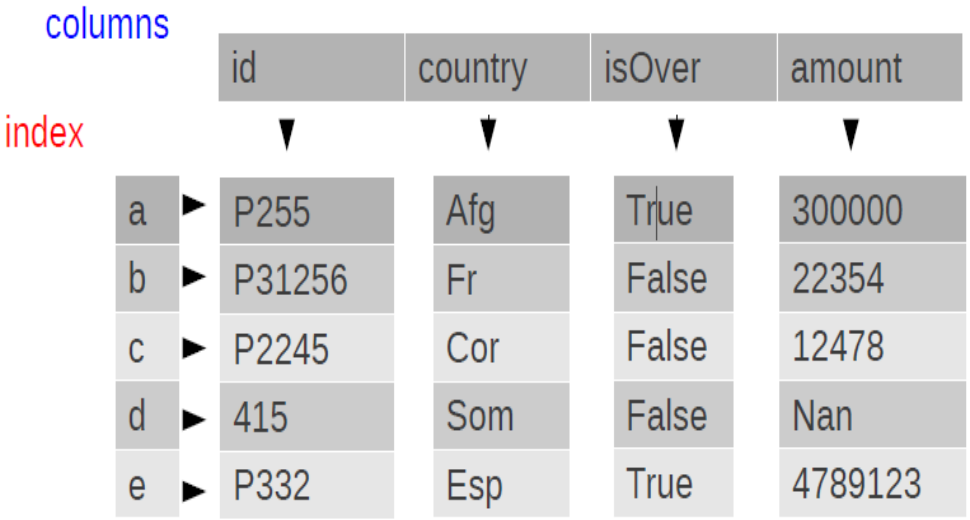
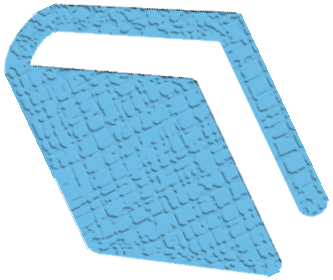



Table like structure

- 2D data structure
- Row and column index
- Size mutable: insert or delete columns
- SQL like transformations – select, groupby, aggregations, filtering, joining etc.

# Summary

Summary of the topics covered in this lesson:



- With the data explosion across industry, the usage of analytics in decision making will become the most critical factor for being competitive in business.
- Descriptive analytics becomes the stepping stone to all the complex problems which can be solved using analytics.

# End of Lesson–Introduction to Business Analytics