ANALYTICS CLUB
iitmadras

INTRODUCTION TO ANALYTICS

- Analytics is the systematic computational analysis of data using statistics
- It involves interpreting raw data to provide insights so as to guide decision making process in general

- In simple terms, Analytics is nothing but making sense out of data.

Types of Analytics

- Predictive Analytics
- Descriptive Analytics
- Diagnostic Analytics

Fields

- Commerce
- Healthcare
- Sports
- Finance
- Online retail
- Rural development

Domains of Analytics

- Prescriptive

This type of analysis reveals what actions should be taken. This is the most valuable kind of analysis and usually results in rules and recommendations for next steps.

- Predictive

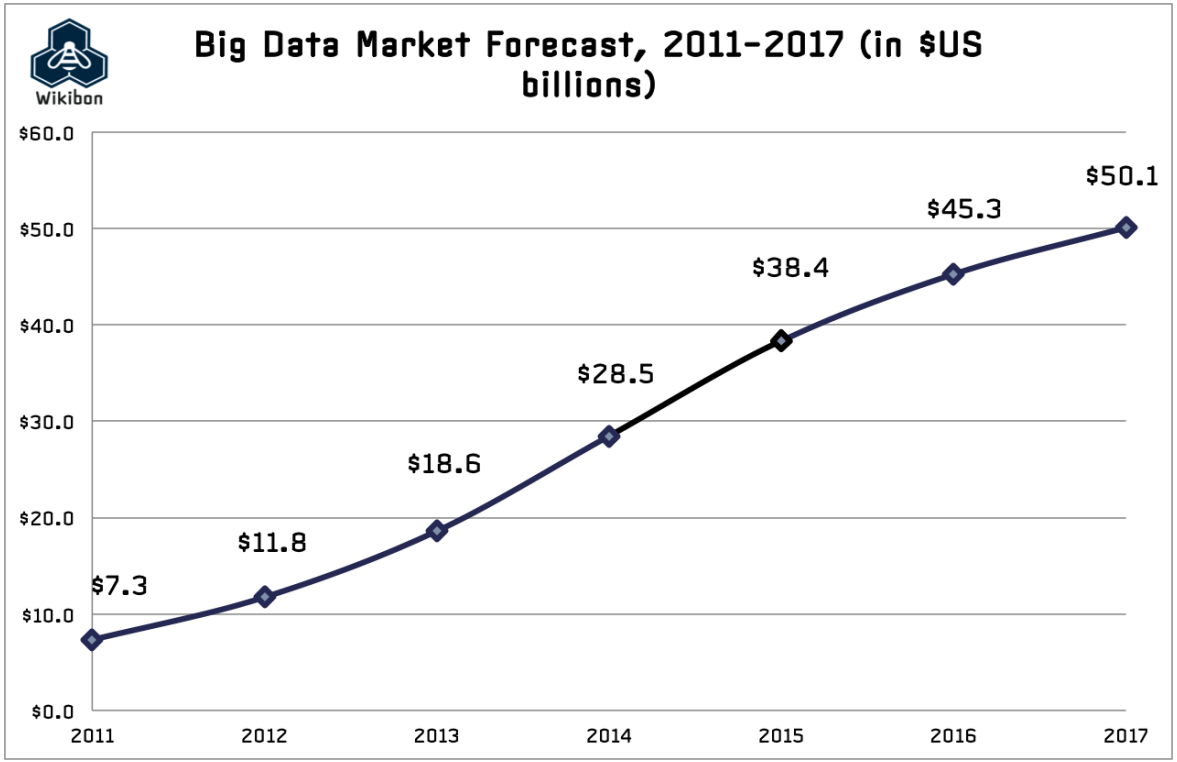
An analysis of likely scenarios of what might happen. The deliverables are usually a predictive forecast

- Diagnostic

A look at past performance to determine what happened and why. The result of the analysis is often an analytic dashboard

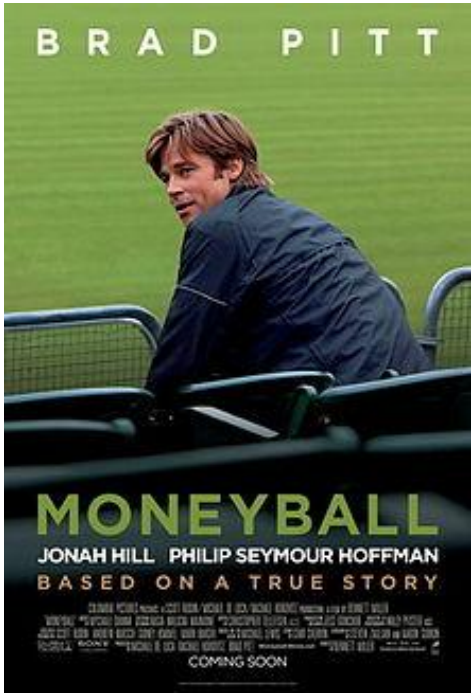
- Descriptive

What is happening now based on incoming data. To mine the analytics, you typically use a real-time dashboard and/or email reports

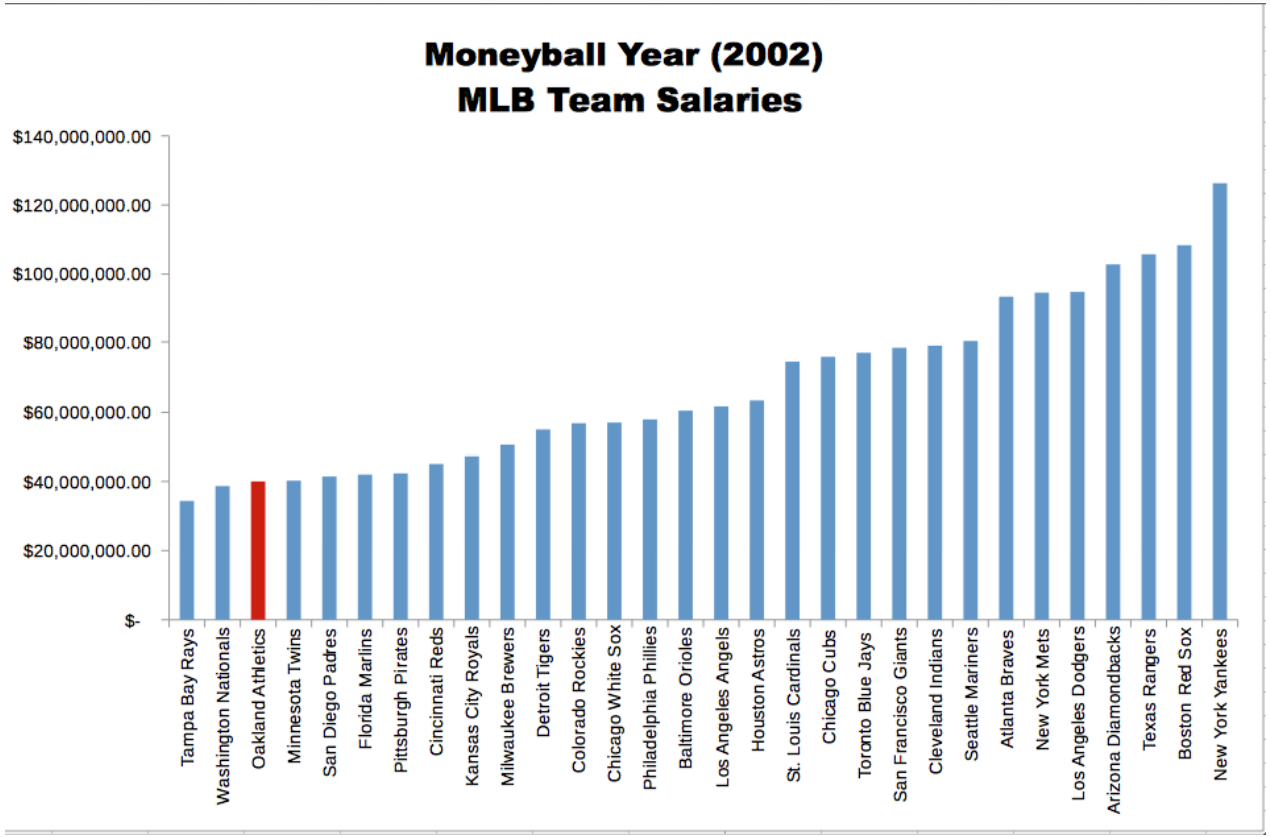

Wikibon
Big Data Market Forecast, 2011-2017 (in $US billions)
$60.0
$50.0
$40.0
$30.0
$20.0
$10.0
$0.0
$7.3
$11.8
$18.6
$28.5
$38.4
$45.3
$50.1
2011
2012
2013
2014
2015
2016
2017

Moneyball

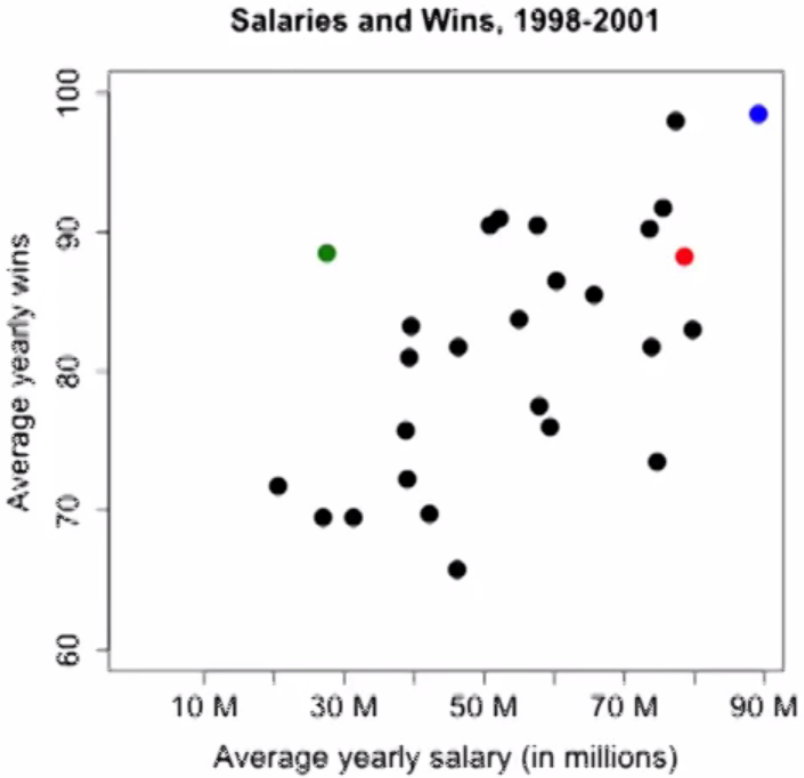
- Moneyball (2011) was based on the true story of how Oakland A manager Billy beane used analytics to draft/buy players for his team.
- **Sabremetrics** - empirical analysis of baseball, especially baseball statistics that measure in-game activity. Coined by Bill James.
- Oakland A have always been a financially constrained team . Buying the best player wasn't an option.
- Conventional baseball scouts weren't providing adequate results
- Gave new insights into player assessments techniques - Stealing bases, sacrifice bunting are overrated stats



INTRODUCTION TO ANALYTICS



INTRODUCTION TO ANALYTICS



Big Data Analytics at work

IBM Watson

- One of the earliest, most popular applications of machine learning/AI
- Was specifically developed to answer questions on the TV show Jeopardy.
- In 2011, it won competed on Jeopardy! against former winners Brad Rutter and Ken Jennings and won the first prize of 1 million.

- More specifically , Watson was an application of text mining/ Natural Language processing
- It had access to 200 million pages of structured and unstructured content consuming 4 TB of storage, including the full text of Wikipedia. It was not connected to the internet during the game.

Source:
http://www.ibm.com/smarterplanet/us/en/ibmwatson/implement-watson.html

IBM Watson participating in Jeopardy

Video

The Story of Netflix Million Dollar Challenge

- Netflix was a company involved in online DVD rental and video streaming service.
- They had an algorithm (cinematch) to provide personalized recommendations about whether a subscriber would enjoy a movie based on their rating history
- The challenge was to develop an algorithm that would beat cinematch's accuracy by atleast 10%
- They also had an annual progress prize until the challenge was completed

Netflix million dollar challenge

- The first year of the contest itself saw an improvement of 8.43%, 2nd year a very slight improvement but after that the participants saw their progress stall

- The participants openly discussed the basics of their algorithms and ideas because 10% was just too tough. They had to help each other out.

The Challenges Faced

- A person may be of the kind who rates 3 for a movie when all others give 4. This has to be taken into consideration.
- People's ratings change by 0.4 on an average after a month. this factor to be taken into account
- People's tastes evolve. I might not have liked ' Oh My God!' but I may like 'PK'. The time gap may have influenced my tastes
- TV series had another problem. After watching two seasons of a show I may want to stop, So season 3 should no longer be recommended for this particular user.

Netflix million dollar challenge

- Meanwhile teams began to start merging with one another
- They started combining their algorithms for better results.
- Actually the codes became so difficult to understand that the those who created did not understand the logic of why the results came the way they did- after all that is expected when you combine 800 algorithms
- But at last two teams managed to cross the 10% barrier and after the final test one team Bellkor's Pragmatic Chaos, a combination of 3 teams won because it completed 20 minutes before the other team

What happened to the million dollar code

- Netflix never used the million dollar code
- The reason was that " the increase in accuracy on the winning improvements did not seem to fully justify the engineering effort needed to bring them into a production environment"
- The important reason was that in the course of three years the way of business changed – from mailed DVDs to live streaming of videos which gave them a different kind of data

And so ends the million dollar story

Framingham Heart Study

- The study has led to the identification of the major CVD risk factors - *high blood pressure, high blood cholesterol, smoking, obesity, diabetes, and physical inactivity* - as well as a great deal of valuable information on the effects of related factors such as blood triglyceride and HDL cholesterol levels, age, gender, and psychosocial issues

- The study has produced approximately 1,200 articles in leading medical journals.

- It provides risk functions that calculate probability of contracting diseases in the next 10/30 year period based on various variables of the patient.

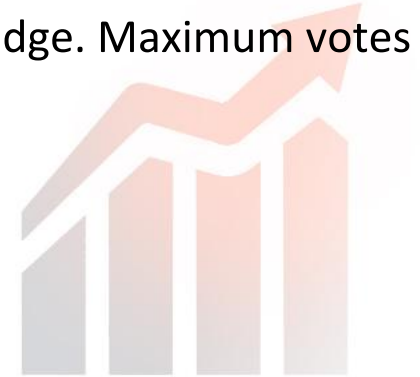
Predicting Supreme Court decisions

- The Supreme Court of United States (SCOTUS) decides on the most difficult and controversial cases. Consists of 9 judges appointed by the President.

- In 2002, Andrew Martin, a professor of political science at Washington University, used predictive analytics to determine SCOTUS judgements.

- Used cases from 1991 to 2001 to train CART models to predict the decision.

- 2 stage approach
  - Two trees to predict conservative and liberal decisions. If contrasting results then move to stage 2
  - Individual trees for each judge. Maximum votes by the judges is the final decision predicted



For the cases in October 2002

- Model accuracy - 75%
- Expert accuracy - 59%

Future Prospects


TIGER
ANALYTICS


flipkart.com
amazon.in
snapdeal.com


Deutsche Bank
Passion to Perform


Goldman
Sachs


LatentView
Actionable Insights • Accurate Decisions


Walmart
Save money. Live better.

Basic scripting Languages

The three basic languages that are commonly used are -

- R
- Python
- SAS

These languages are sufficient for basic small scale analytics work.

However, for big data applications, we need additional softwares like :

- Hadoop
- Apache Spark
- Apache Storm

Hadoop

- It is relatively a young technology which attained version 1.0 status only by December 2011
- It is very widely used for Big Data Analytics.
- Yahoo and Facebook are the most prominent users and over half of Fortune 50 companies use Hadoop

Apache Spark

- Spark like Hadoop is an open source software by Apache Software foundation
- It is a cluster computing framework
- It is used in the Hadoop ecosystem and has advantages of its own

Advantages in Apache Spark

- <u>It is ideally suited for Machine Learning:</u>
  Most Machine Learning Algorithms run on the same data set iteratively and Spark provides a better alternative to Hadoop MapReduce in this respect

- Spark can be used for real-time analytics

- Spark is much faster than Hadoop MapReduce(upto 100 times for some applications)

MACHINE LEARNING :

- It is nowadays a very vast field that is partitioned and subpartitioned continuously into different specialities.

- Most basically it is the field of study that gives computers the ability to learn without being explicitly programmed.

- An instance would be to run a program through Machine Learning Algorithms with data about past traffic patterns, after ‘learning’ the program predicts traffic better.

Big Data

- Big data is data that exceeds the processing capacity of conventional database systems.
- The data is too big in terms of **Volume, Velocity and Variety**.
- Processing of such large data gives rise to a lot of interesting possibilities
- Input data to big data systems could be chatter from social networks, web server logs, traffic flow sensors, satellite imagery etc.

Past and Summer Projects

- Connect Leader-Helping BPO's to improve their predictivity
- Helping Government make decisions-Data Visualizations of power supply-demand data from data.gov.in
- Chennai Corporation project
- West-Nile prediction Challenge-Kaggle

Competitions

- Inter-IIT tech meet-Placed 4th among 12 different IITs

INTRODUCTION TO ANALYTICS

New Initiatives for the semester

- Organized the first Analytics Club Summer-School
- Advanced session for experienced members
- Self-proposed projects
- Industrial projects in pipeline

New Competitions

- Students Analytics Olympiad
- Top Coder Data

FEEDBACK & SUGGESTIONS