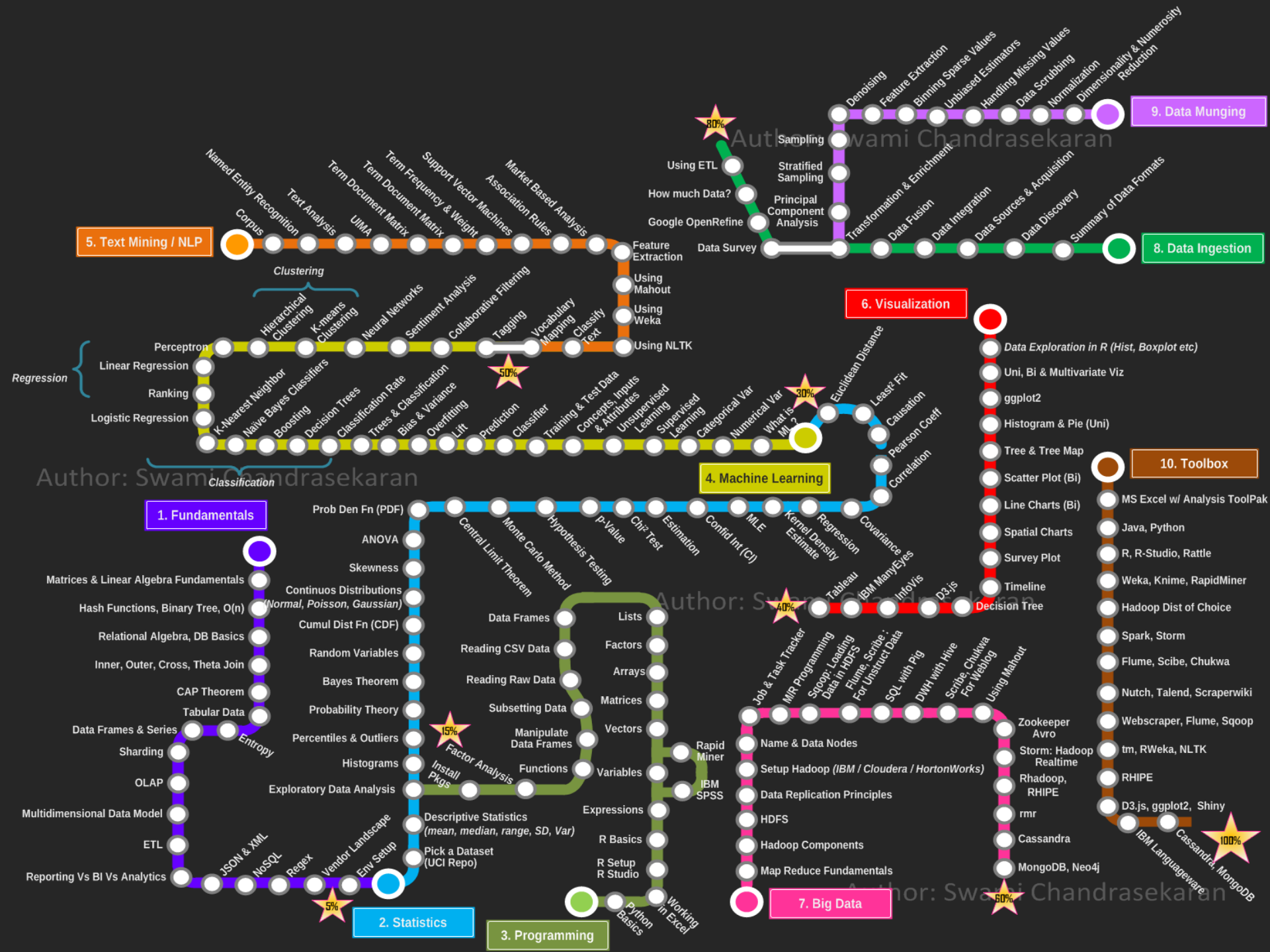
Inspire…Educate…Transform.

CSE 7304c: Lab Activity 1


1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Data Frames & Series
Entropy
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup
5%
2. Statistics
Prob Den Fn (PDF)
ANOVA
Skewness
Continuos Distributions (Normal, Poisson, Gaussian)
Cumul Dist Fn (CDF)
Random Variables
Bayes Theorem
Probability Theory
Percentiles & Outliers
Histograms
Exploratory Data Analysis
Descriptive Statistics (mean, median, range, SD, Var)
Pick a Dataset (UCI Repo)
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
p-Value
Chi² Test
Estimation
Confid Int (CI)
MLE
Kernel Density Estimate
Regression
Covariance
Correlation
15%
Factor Analysis
Install Pkgs
3. Programming
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Lists
Factors
Arrays
Matrices
Vectors
Variables
Expressions
R Basics
R Setup
R Studio
Rapid Miner
IBM SPSS
Python Basics
Working in Excel
4. Machine Learning
Regression
Perceptron
Linear Regression
Ranking
Logistic Regression
Classification
K-Nearest Neighbor
Naïve Bayes Classifiers
Boosting
Decision Trees
Classification Rate
Trees & Classification
Bias & Variance
Overfitting
Lift
Prediction
Classifier
Training & Test Data
Concepts, Inputs & Attributes
Unsupervised Learning
Supervised Learning
Categorical Var
Numerical Var
What is ML?
50%
Clustering
Hierarchical Clustering
K-means Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
Vocabulary Mapping
Classify Text
Using NLTK
Using Mahout
Using Weka
Feature Extraction
5. Text Mining / NLP
Corpus
Text Analysis
UIMA
Term Document Matrix
Term Document Matrix
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Based Analysis
Named Entity Recognition
30%
Euclidean Distance
Least² Fit
Causation
Pearson Coeff
Correlation
6. Visualization
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
40%
Tableau
IBM ManyEyes
InfoVis
D3.js
7. Big Data
Job & Task Tracker
MR Programming
Sqoop: Loading Data in HDFS
Flume, Scribe : For Unstruct Data
SQL with Pig
DWH with Hive
Scribe, Chukwa For Weblog
Using Mahout
Name & Data Nodes
Setup Hadoop (IBM / Cloudera / HortonWorks)
Data Replication Principles
HDFS
Hadoop Components
Map Reduce Fundamentals
Zookeeper Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j
60%
8. Data Ingestion
Data Survey
Using ETL
How much Data?
Google OpenRefine
80%
Sampling
Stratified Sampling
Principal Component Analysis
Transformation & Enrichment
Denoising
Feature Extraction
Binning Sparse Values
Unbiased Estimators
Handling Missing Values
Data Scrubbing
Normalization
Dimensionality & Numerosity Reduction
Data Fusion
Data Integration
Data Sources & Acquisition
Data Discovery
Summary of Data Formats
9. Data Munging
10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB
100%
Author: Swami Chandrasekaran


IN 60 SECONDS...
1 NEW DEFINITION IS ADDED ON urban
1,600+ READS ON Scribd.
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
20,000+ NEW POSTS ON tumblr.
THE LARGEST SOCIAL READING PUBLISHING COMPANY
New Craigslist Ads
320+ NEW twitter ACCOUNTS
13,000+ iPhone APPLICATIONS DOWNLOADED
100+ NEW Linked in ACCOUNTS
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO! ANSWERS
6,600+ NEW PICTURES ARE UPLOADED ON flickr
600+ NEW VIDEOS
50+ WordPress DOWNLOADS
25+ HOURS TOTAL DURATION
70+ DOMAINS REGISTERED
695,000+ facebook STATUS UPDATES
125+ PLUGIN DOWNLOADS
60+ NEW BLOGS
168 MILLION EMAILS ARE SENT
694,445 SEARCH QUERIES
1,700+ Firefox DOWNLOADS
79,364 WALL POSTS
1,500+ BLOG POSTS
510,040 COMMENTS
Google
Google Search
GO-Globe.com

Lab session – 1 : Introduction to Big data: World uses more Big Apps. Lets Explore

Lab will be monitored by:
→Anuradha
→Amrit
→ Yugandhar
→ Sandeep

Theme for Lab sessions

- Make them like Dexter Lab

Of course without the Outrageous capital and Fancy failures

Lab session dynamics

- Lab sessions follow Theory. We try to balance theory with practice
- Activity sheets: Short exercises for hands on familiarity.
  - Level: Basic to Intermediary
- Demonstration: How to?
- Assistance during projects

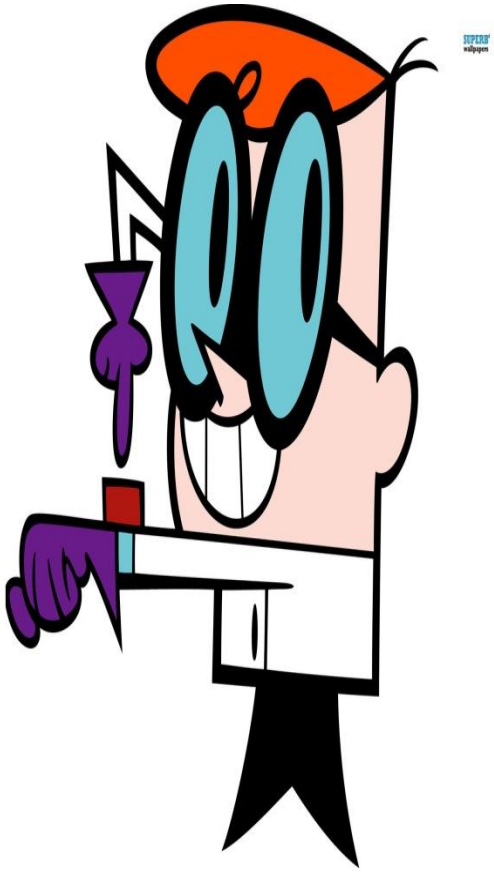
Time split for lab sessions

–Lab sessions usually follow this split.

10 – 15 % of time for demos

60 % Hands on practice

25 – 30 % [Case study / Mini Projects]



Lab session pre-requisites

- BYOD: Laptop is compulsory for all lab sessions
  - Configuration
    - Atleast 4 GB RAM [Preferably 6-8 GB

      RAM]
    - Atleast 20 GB of storage space
    - 64 bit host OS
    - Support for Virtualization
    - Software: VMWare player, WinScp, R
    - Pen drives are good to have

Lab sessions: Expectations?

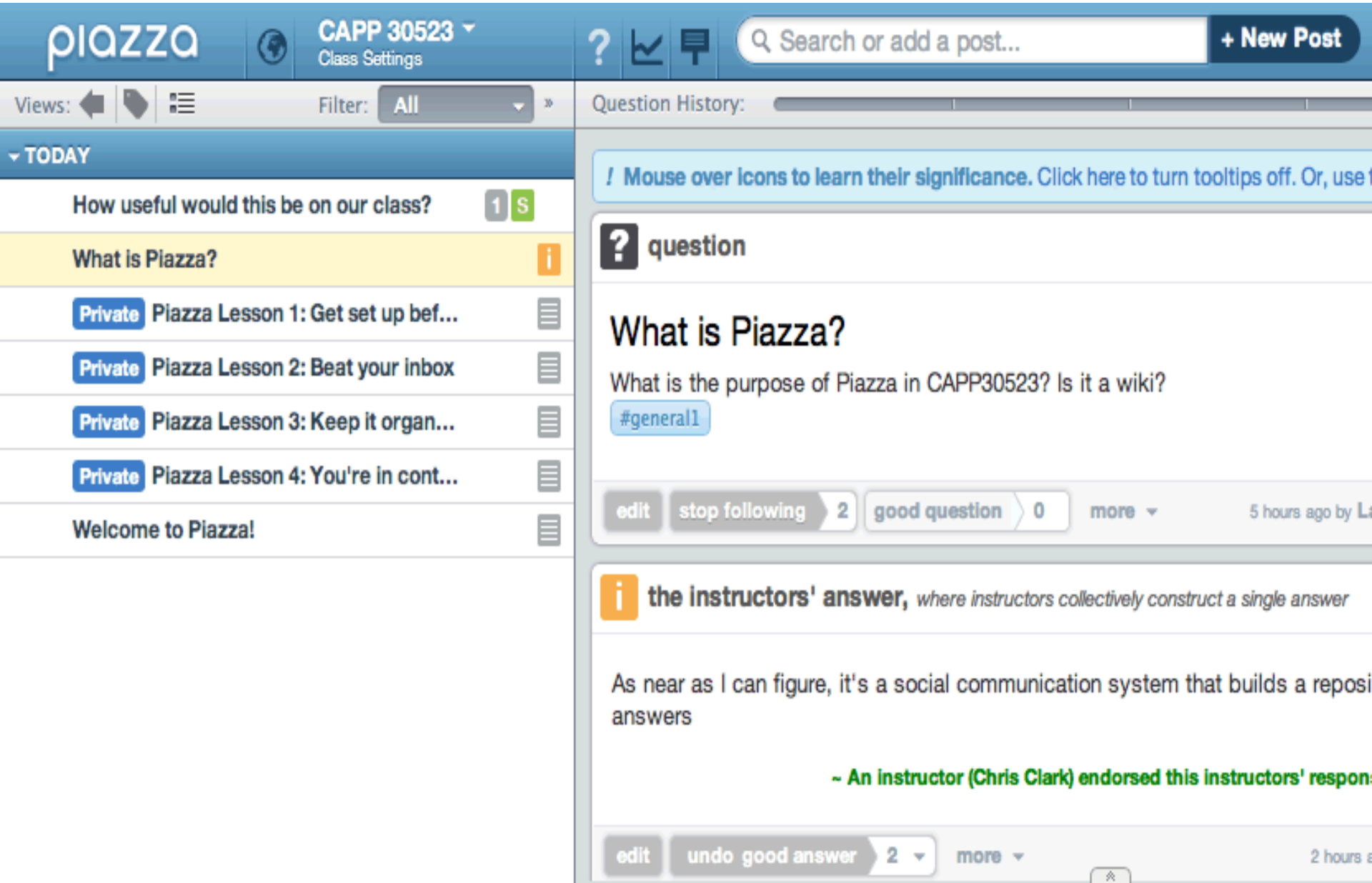
- What is our MVP?

- We can include specific topics of interest

- Lab sessions [for BigData classes] are seldom in fixed delivery mode.

We can have a flexible model

Piazza!! What about it?



Quizzes? How do they work?

- WUQ
- GNQ
- Questions on Piazza
- Turn around time

Ok! What about Projects?

- Predictive Analytics

- BigData Engineering

Activity 1
Group Activity

- Pick an Industry of your choice
- Gather as much information to crisply define a Big Data problem
- Categorize the problem as
  - High Volume/variety or high velocity class
  - If your team is to implement this solution, what are the constraints / issues that you will face. (Define them as crisply as possible without compromising on detail)
- Cost? What do you think are the costs involved?

Big Data Application Examples

Source: Microsoft

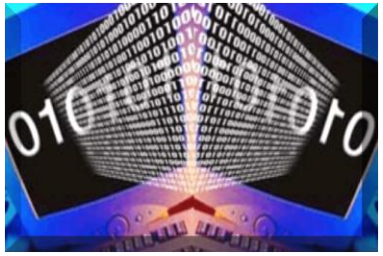
Applications for Big Data Analytics

Smarter Healthcare

Multi-channel sales

Finance

Log Analysis



Homeland Security

Traffic Control

Telecom

Search Quality

Manufacturing

Trading Analytics

Fraud and Risk

Retail: Churn, NBO

Activity 2
Group Activity

- Is there an implementation already for the problem that you have chosen?

- Are there relevant case studies? What technology (ies) were used by the engineering solution?

- How different was your approach?

Activity 3
Group Activity

- Think of one of the popular social media tools **[Explore tools that you are not very aware!]**
- Enumerate crisply

in technical terms
the big data problems address
by that application.
→Identify and enumerate
the potential issues and
suggest engineering
solutions based on your
Expertise!

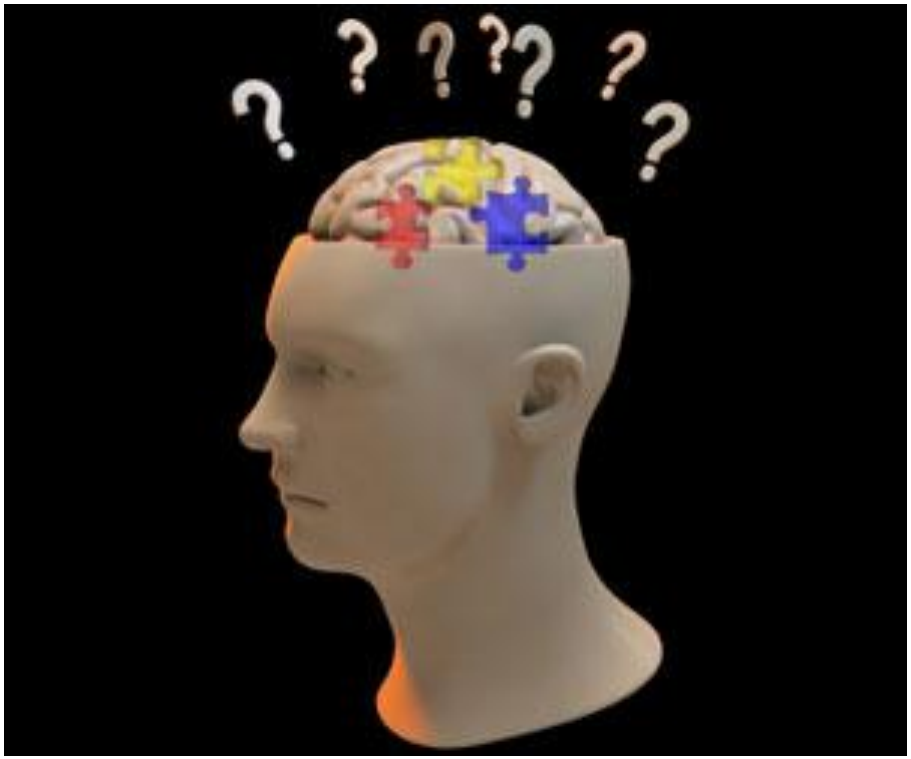
The Gaming Industry

Massive online games!!

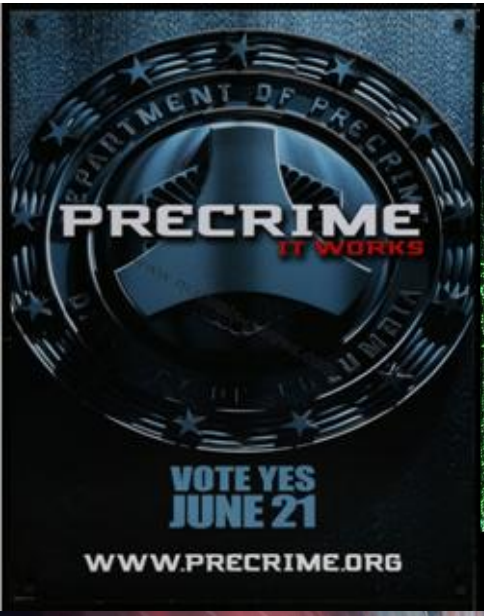
Activity [Optional]
Group Activity

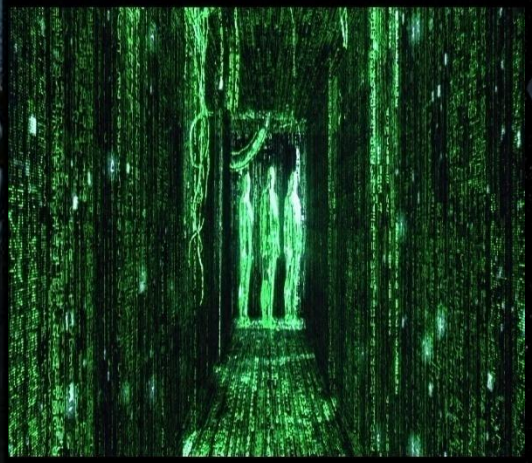
- Think of your favorite Sci – Fi / Action / real life instance based movies that have a specific concept related to Big Data Analytics (Remember the 3 V's of Big Data for the candidates to Qualify)
  1) Enumerate the problem / describe the problem in Business terminology
  2) List out all the Data collection, Data storage and Data Processing problems that you can think of to implement this concept. Be creative and add a wish list that you would want that concept to implement.

Teasers??



The Reel World – Big Data Analytics





STAR WARS™

BigData repositories

- Data.gov
- US Census Bureau
- European Union Open Data Portal
- Data.gov.uk
- The CIA World Factbook
- Healthdata.gov
- NHS Health and Social Care Information Centre
- Amazon Web Services public datasets
- Facebook Graph
- Gapminder

Google Trends
Google Finance
Google Books Ngrams
National Climatic Data Center
DBPedia
Topsy
Likebutton
New York Times
Freebase
Million Song Data Set
Reuters Data Set

Gist: What we look at Lab sessions as?

- Lab sessions:

  Collective Intelligence

  Assist each other with our experience

  Try to understand in detail Issues / problems / confusion in theory [if any]

  Map theory with practice and understand Industry use-cases

  Try our best to answer your questions, else get back to you in short time. Or take it up in Piazza/ class with Primary Faculty

International School of Engineering

Plot 63/A, 1st Floor, Road # 13, Film Nagar, Jubilee Hills, Hyderabad - 500 033

For Individuals: +91-9502334561/63 or 040-65743991

For Corporates: +91-9618483483

Web: http://www.insofe.edu.in

Facebook: https://www.facebook.com/insofe

Twitter: https://twitter.com/Insofeedu

YouTube: http://www.youtube.com/InsofeVideos

SlideShare: http://www.slideshare.net/INSOFE

LinkedIn: http://www.linkedin.com/company/international-school-of-engineering