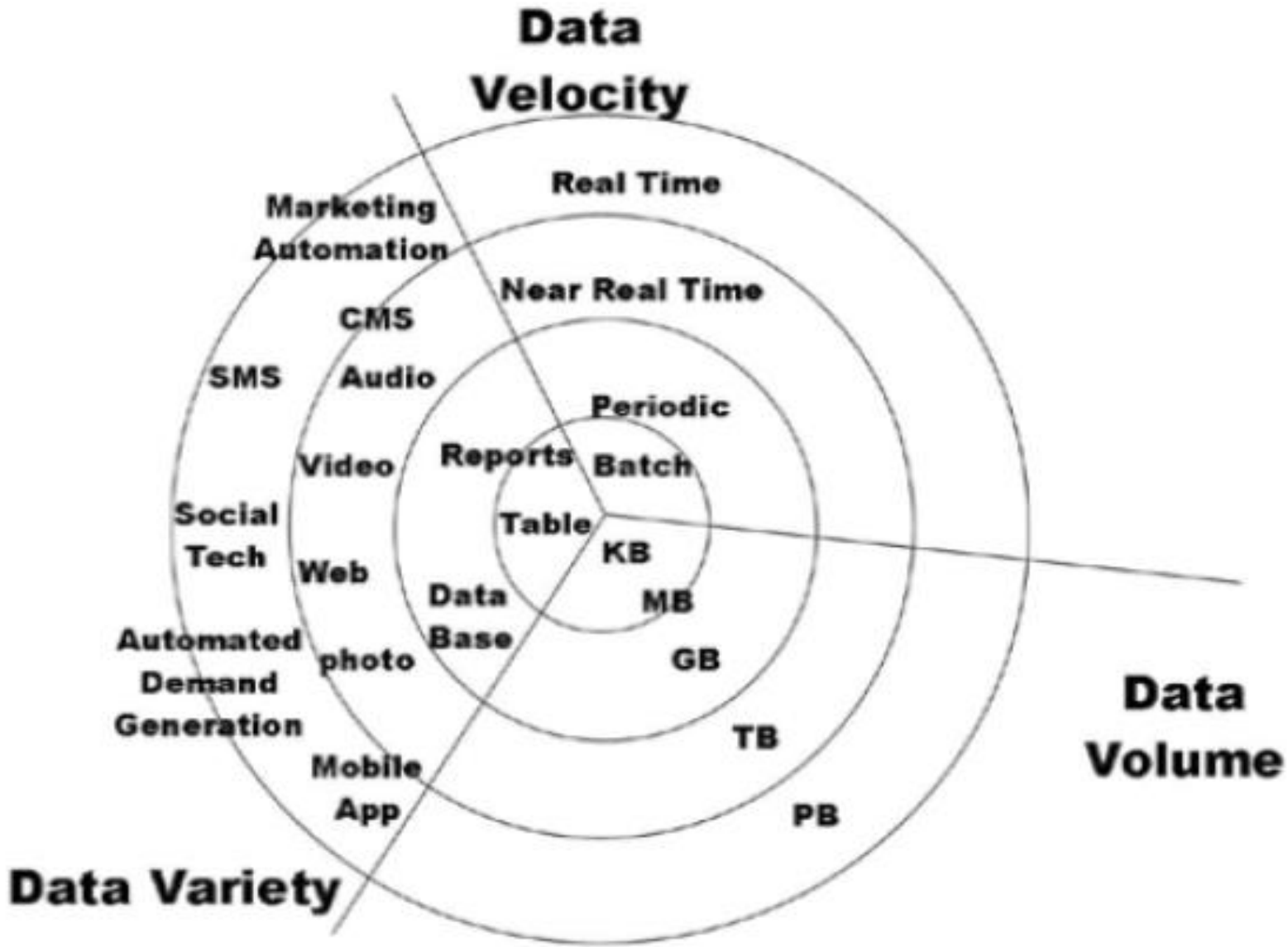
Big Data Analytics

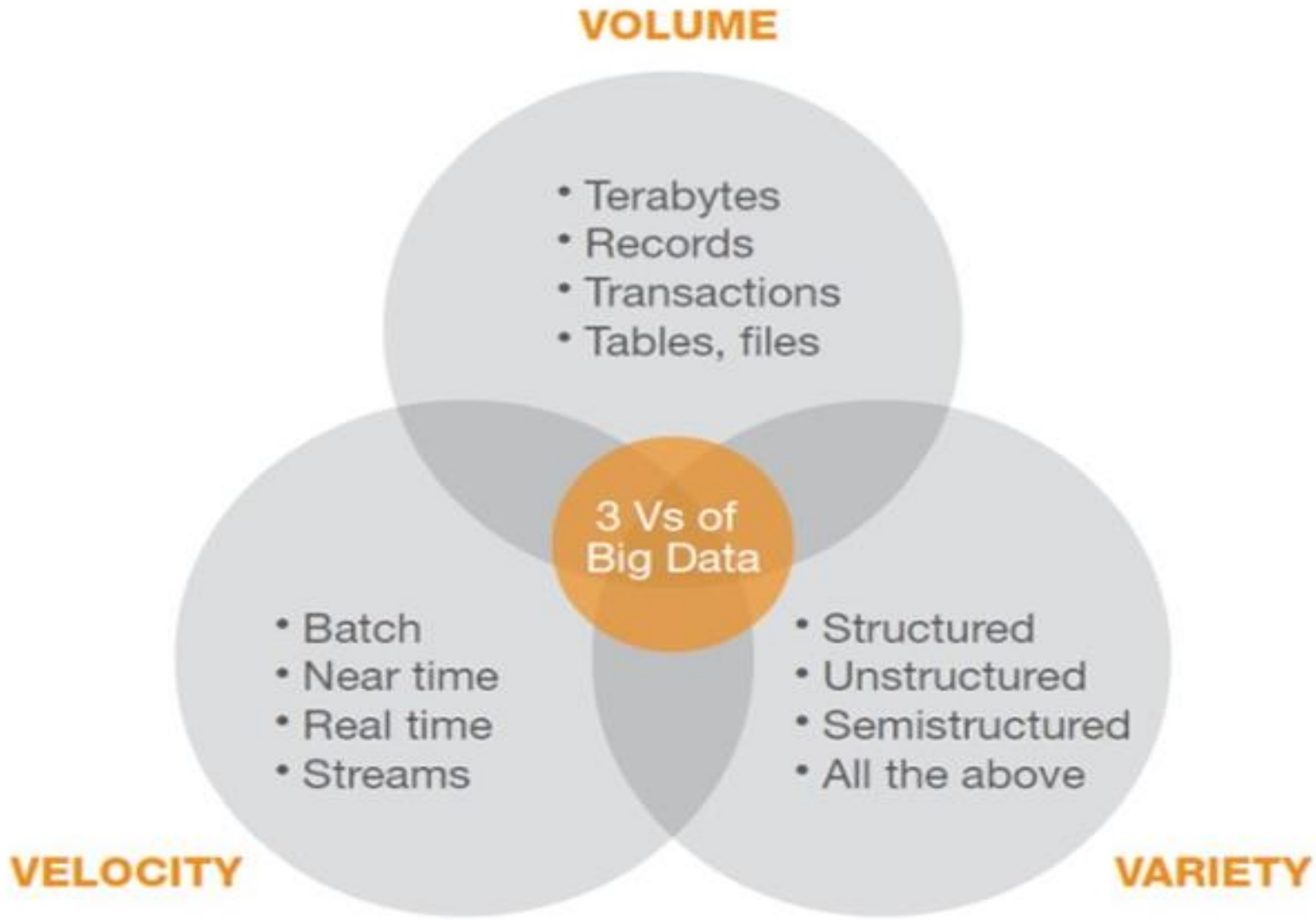
UNIT 1 INTRODUCTION **9 Hrs.**

Introduction to Big Data – Issues and Challenges in the traditional systems - Evolution of Big Data – Four V's of Big Data – Big Data Use Cases and characteristics – Intelligent Data Analysis – Data Analytic Tools – Big Data StorageStatistical Concepts: Sampling Distributions - Re-Sampling - Statistical Inference - Prediction Error – Random Sampling.

Introduction to Big Data

- Big Data has to deal with large and complex datasets
- Varities of Data: structured, Semi-structured, or unstructured
- Bigdata will typically not fit into memory to be Processed.


Data
Velocity
Real Time
Near Real Time
Periodic
Marketing
Automation
CMS
SMS
Audio
Video
Reports
Batch
Social
Tech
Table
KB
Web
Data
Base
MB
Automated
Demand
Generation
photo
GB
TB
Mobile
App
PB
Data
Volume
Data Variety


VOLUME
• Terabytes
• Records
• Transactions
• Tables, files
3 Vs of
Big Data
• Batch
• Near time
• Real time
• Streams
VELOCITY
• Structured
• Unstructured
• Semistructured
• All the above
VARIETY

Examples

- Real Time: Air traffic control systems
- Near Real Time: Weather prediction(collecting real time data & processing it)
- Periodic real time: Stock Market
- Batch processing: Bank ATM retrieval

Structured,Unstructured & Semistructured data

- **Structured data –**
  Structured data is a data whose elements are addressable for effective analysis.
- It concern all data which can be stored in database SQL in table with rows and columns.
- Relational key is easily mapped into pre-designed fields.
- Example: Relational Database

Semi-structured data:

- Information does not reside in a rational database
- have some organizational properties that make it easier to analyze.
- *Example*: XML data.

Unstructured data –

- Unorganized
- Does not have a pre-defined data model
- Unfit for fit for a mainstream relational database
- *Example*: Word, PDF, Text, Media logs.

Differences between Structured, Semi-structured and Unstructured data:

PROPERTIES	STRUCTURED DATA	SEMI-STRUCTURED DATA	UNSTRUCTURED DATA
Technology	It is based on Relational database table	It is based on XML/RDF	It is based on character and binary data
Transaction management	Matured transaction and various concurrency technique	Transaction is adapted from DBMS not matured	No transaction management and no concurrency
Version management	Versioning over tuples,row,tables	Versioning over tuples or graph is possible	Versioned as whole
Flexibility	It is sehema dependent and less flexible	It is more flexible than structuded data but less than flexible than unstructured data	it very flexible and there is abbsence of schema
Scalability	It is very difficult to scale DB schema	It's scaling is simpler than sstructured data	It is very scalable
Robustness	Very robust	New technology, not very spread	–
Query performance	Structured query allow complex joining	Queries over anonymous nodes are possible	Only textual query are possible

Challenges of Conventional Systems

- Challenge 1: How to store the big data?

- Challenge 2:

How to work with voluminous data sizes, and more important, how to understand data and turn it into a competitive advantage?

Challenge 3: How about Conventional system technology?

- CPU Speeds:
  - 1990 - 44 MIPS at 40 MHz
  - 2000 - 3,561 MIPS at 1.2 GHz
  - 2010 - 147,600 MIPS at 3.3 GHz
- RAM Memory
  - 1990 – 640K conventional memory (256K extended memory recommended)
  - 2000 – 64MB memory
  - 2010 - 8-32GB (and more)
- Disk Capacity
  - 1990 – 20MB
  - 2000 - 1GB
  - 2010 – 1TB
- Disk Latency (speed of reads and writes) – not much improvement in last 7-10 years, currently around 70 – 80MB / sec

Challenge 4: How long it will take to read 1TB of data?

TB (at 80Mb / sec):

- • – 1 disk - 3.4 hours
- • – 10 disks - 20 min
- • – 100 disks - 2 min
- • – 1000 disks - 12 sec

Challenge 5: What do we care about when we process data?

- Handle partial hardware failures without going down:
  - – If machine fails, we should be switch over to stand by machine
  - – If disk fails – use RAID or mirror disk
- Able to recover on major failures:
  - – Regular backups
  - – Logging
  - – Mirror database at different site
- Capability:
  - – Increase capacity without restarting the whole system
  - – More computing power should equal to faster processing
- Result consistency:
  - – Answer should be consistent (independent of something failing) and returned in

    reasonable amount of time

Summary of Challenges with BigData

- Capture
- Storage
- Curation
- Search
- Analysis
- Transfer
- Visualization
- Privacy Violation

Different Sizes of Data

Name	Full Name	Size
KB	Kilobyte	1024Bytes(2^{10})
MB	MegaByte	1, 024kilobytes(1024^2)
GB	Giga Byte	1, 024 megabytes(1024^3)
TB	Terra Byte	1, 024 gigabytes(1024^4)
PB	Peta Byte	1, 024 terrabytes(1024^5)
EB	Exabyte	1024 PetaBytes(1024^6)
ZB	Zetta Byte	1024 ExaBytes(1024^7)
YB	Yotta Byte	1024 Zetta Bytes(1024^8)

Nature of Data

- Small Data
  - Smaller in size
  - Structured
  - Easy to be handled by humans
- Medium Data
  - Larger in volume than Small data
  - Semi Structured
  - Can be handled by smaller computations/simple algorithms
- Big Data
  - Huge volumes of complex data
  - Unstructured
  - Could not be dealt by traditional data processing applications

Small data vs Big data

Basis Of Comparison	Small Data	Big Data
Definition	Data that is 'small' enough for human comprehension.In a volume and format that makes it accessible, informative and actionable	Data sets that are so large or complex that traditional data processing applications cannot deal with them
Data Source	Data from traditional enterprise systems like ➢Enterprise resource planning ➢Customer relationship management(CRM) ➢Financial Data like general ledger data ➢Payment transaction data from website	➢Purchase data from point-of-sale ➢Clickstream data from websites ➢GPS stream data – Mobility data sent to server ➢Social media – facebook, twitter

Basis Of Comparison	Small Data	Big Data
Volume	Most case in a range of tens or hundreds of GB.Some case few TBs (1 TB=1000 GB)	More than few Terabytes (TB)
Velocity (Rate at which data appears)	Data accumulation is slow	Data can arrive at very fast speeds. Enormous data can accumulate within very short periods of time
Variety	Structured data in tabular format with fixed schema and semi-structured data in JSON or XML format	High variety data sets which include Tabular data,Text files, Images, Video, Audio,XML,JSON,Logs,Sensor data etc.

Basis Of Comparison	Small Data	Big Data
Veracity (Quality of data)	Contains less noise as data collected in a controlled manner.	Usually, quality of data not guaranteed. Rigorous data validation is required before processing.
Value	Business Intelligence, Analysis and Reporting	Complex data mining for prediction, recommendation, pattern finding etc.
Time Variance	Historical data equally valid as data represent solid business interactions	In some cases data gets older soon(Eg fraud detection).
Data Location	Databases within an enterprise, Local servers etc.	Mostly in distributed storages on Cloud or in external file systems.

Basis Of Comparison	Small Data	Big Data
Infrastructure	Predictable resource allocation.Mostly vertically scalable hardware	More agile infrastructure with horizontally scalable architecture. Load on the system varies a lot.

Sources of Big Data

The data from the following sources can be structured, semi-structured, or unstructured, or any combination of these varieties.

- **Social Networks**
- **Traditional Business Systems**
- **Internet of Things**

Social Networks

- Twitter and **Facebook**
- Blogs and comment
- Pictures: Instagram®, Flickr™, Picasa™, etc.
- Videos: YouTube
- Internet searches
- Mobile data content (text messages)
- User-generated maps
- E-Mail

Traditional Business Systems

- Commercial transactions
- Banking/stock records
- E-commerce
- Credit cards
- Medical records

Internet of Things

- Sensors: traffic, weather, mobile phone location, etc.
- Security, surveillance videos, and images
- Satellite images
- Data from computer systems (logs, web logs, etc.)

Big Data Characteristics

1)Volume: Volume is the amount of data generated that must be understood to make data based decisions.

A text file is a few kilobytes, a sound file is a few megabytes while a full-length movie is a few gigabytes.

- **Example:** Amazon handles 15 million customer click stream user data per day to recommend products.

Big Data Characteristics

2) Velocity: Velocity measures how fast data is produced and modified and the speed with which it needs to be processed. An increased number of data sources both machine and human generated drive velocity.

Example: 72 hours of video are uploaded to YouTube every minute this is the velocity.

Big Data Characteristics

- **Variety:** Variety defines data coming from new sources—both inside and outside of an enterprise
- It can be structured, semi-structured or unstructured.

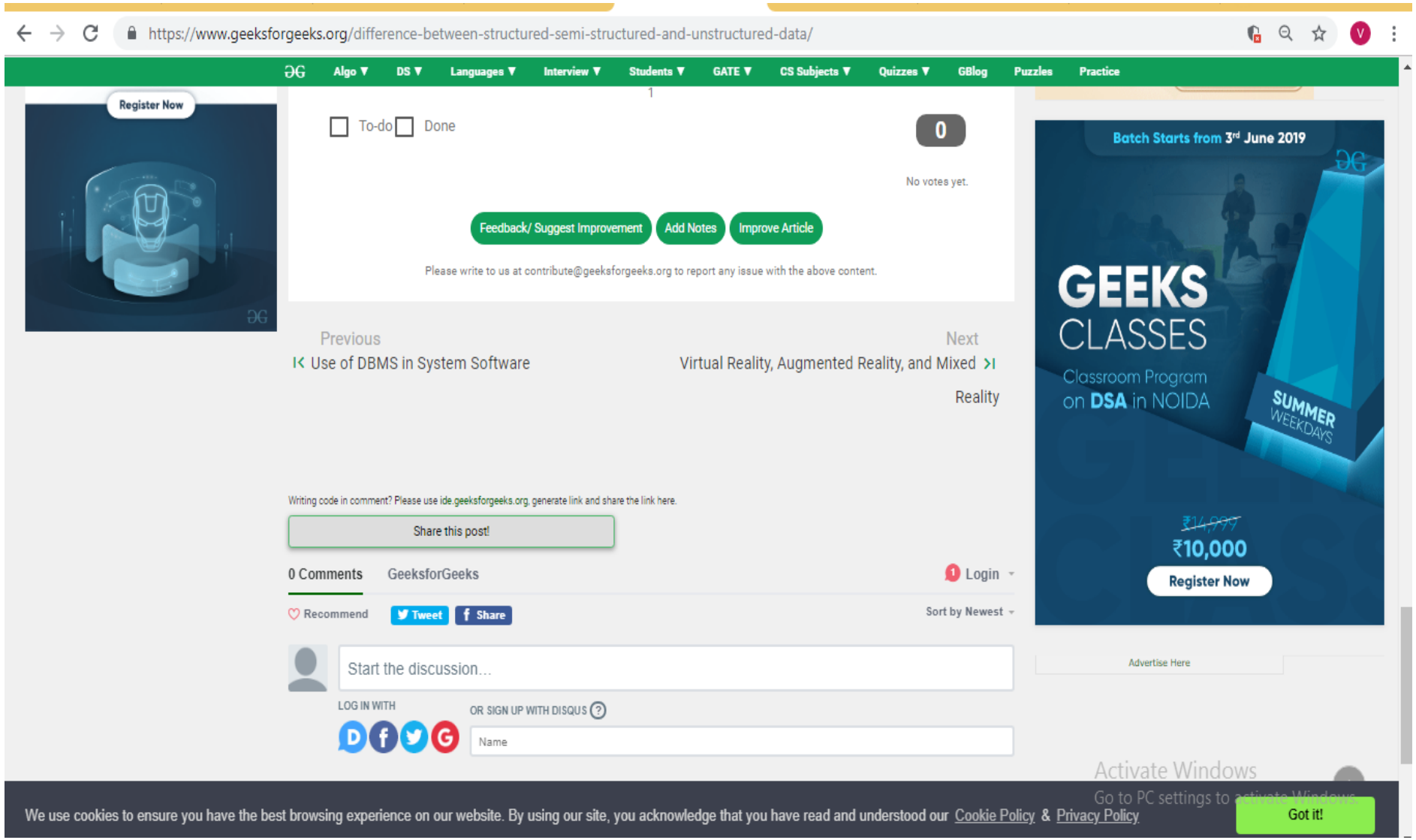
Structured data-An RDBMS table

Product_id	Product_name	Product_price
1	Pen	$5.95
2	Paper	$8.95

Semi Structured-XML data

```
<product>
<name>Pen </name>
<price>$7.95</price>
</product>
<product>
<name>Paper </name>
<price>$8.95</price>
</product>
```

Unstructured-Internet Contents



Big Data Characteristics

4) Variability – This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

Example:1) twitter posts(hash tags, abbreviations,typos)

2)An dataset with missing values

Department	Year	Minimum sales	Maximum sales
1	2010	?	1500
2	2011	10000	?

Big Data Analytics

Big Data analytics is the process of checking Large Data sets. To get to know hidden Information, which contains the hidden patterns, market trends, correlations and customer preferences.

Working Process of Big Data analytics

- Hadoop clusters and No SQL systems used for staging areas and landing pads for Data.
- If Data gets loaded into analytical database or data warehouse for **data analysis**, it driven from relational Structures.


BEFORE ANALYTICS
How can siloed patient data be combined to create insight?
AFTER ANALYTICS
Data is centralized and integrated to deliver actionable information.

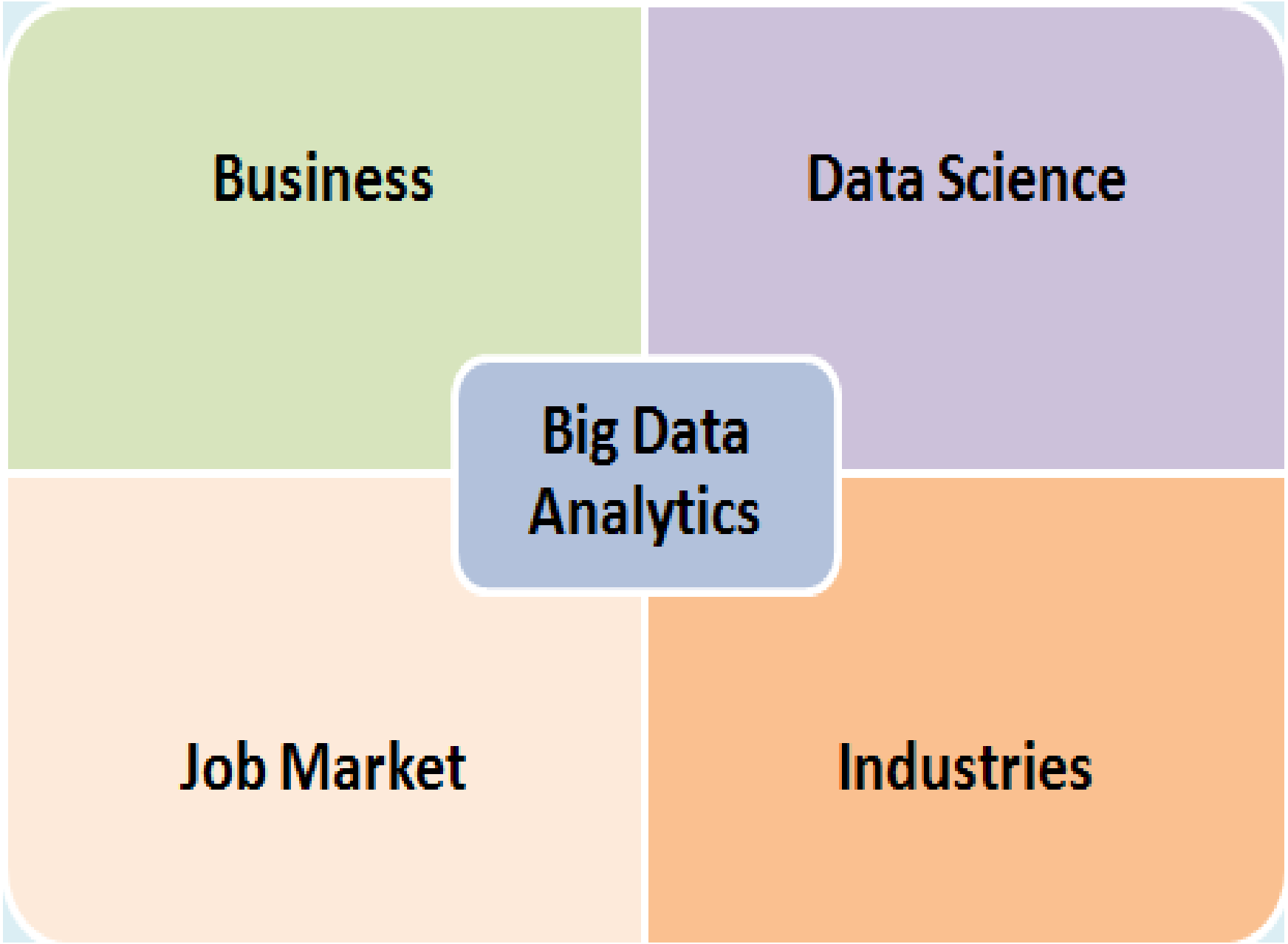
Importance of Big Data

- Aim of big data analytics is to flourish in Business to Consumer (B2C) applications
- Three divisions of Big Data analytics:

➢Prescriptive Analytics

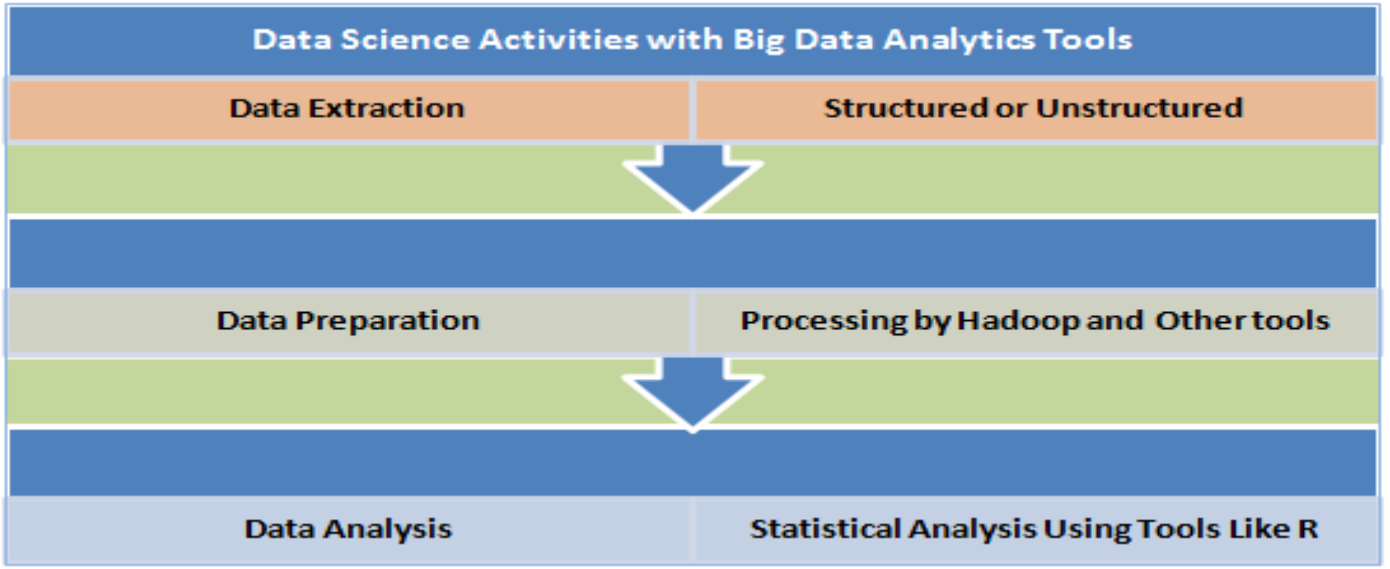
➢Predictive Analytics

➢Descriptive Analytics


Business
Data Science
Big Data Analytics
Job Market
Industries

Big Data and Data sciences

- Big Data analytics involves the use of analytics techniques like machine learning, data mining, natural language processing, and statistics.



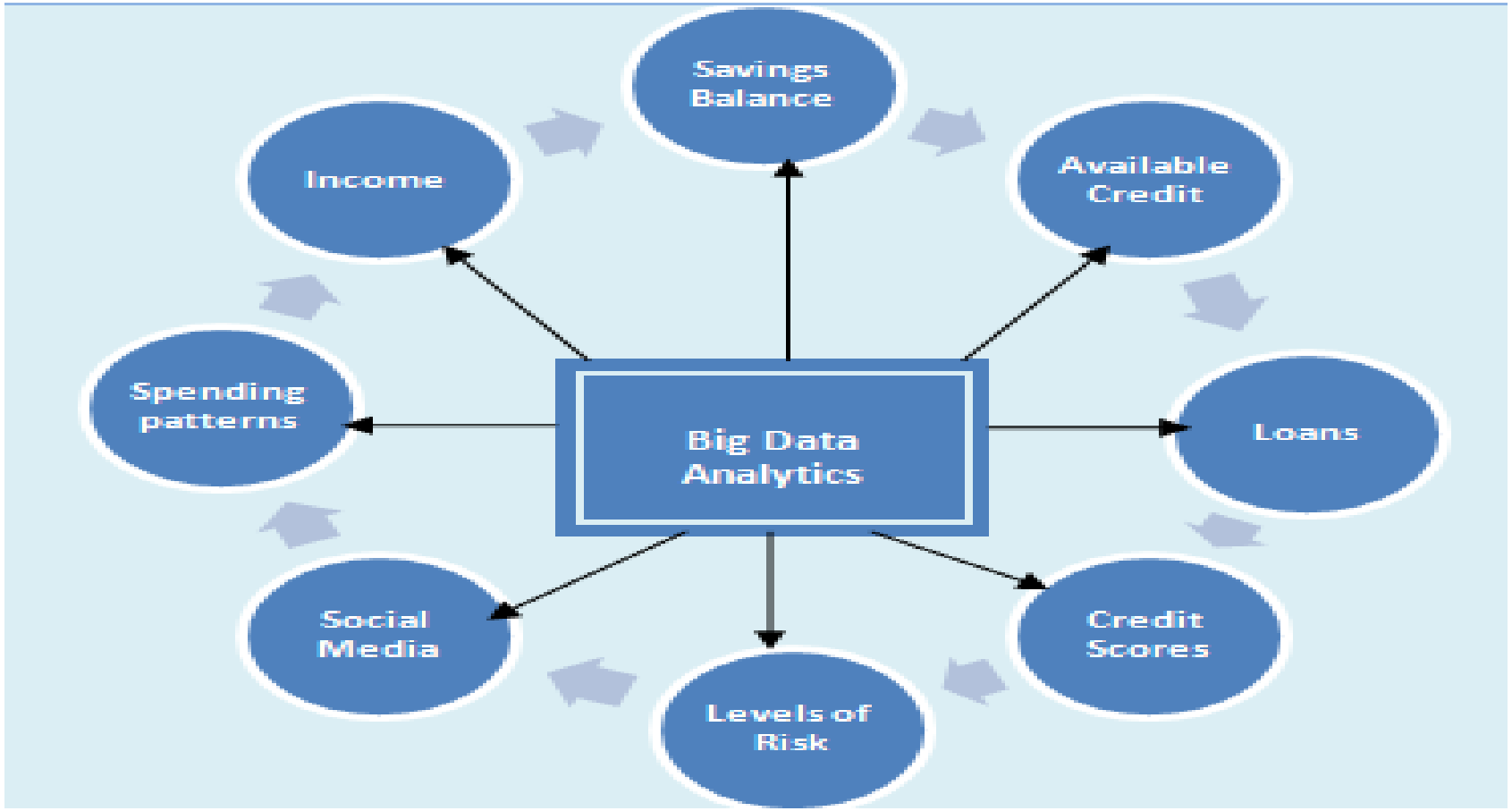
Businesses and Big Data Analytics

- Helps in improving the profits of the company.
- Big Data tools help in providing meaningful information for making better business decisions.
- Big data analytics tools like Hadoop helps in reducing the cost of storage.
- Leads to faster decision making saving time and energy.

Real-time Benefits of Big Data Analytics

- Banking
- Healthcare
- Energy
- Technology
- Consumer
- Manufacturing

Benefits of Big Data Analytics



Introduction to Big Data

Big Data Analytics tools and softwares

Big Data Use Cases

Use Case 1: Social Network Analysis in Enterprise for Productivity

Production Live System used by IBM GBS since 2009 – verified ~$100M contribution

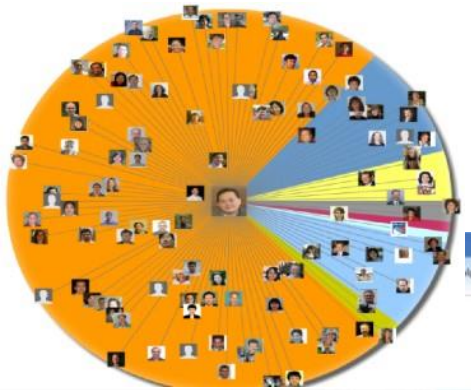
15,000 contributors in 76 countries; 92,000 annual unique IBM users

25,000,000+ emails & SameTime messages (incl. Content features)

1,000,000+ Learning clicks; 14M KnowledgeView, SalesOne, …, access d

1,000,000+ Lotus Connections (blogs, file sharing, bookmark) data

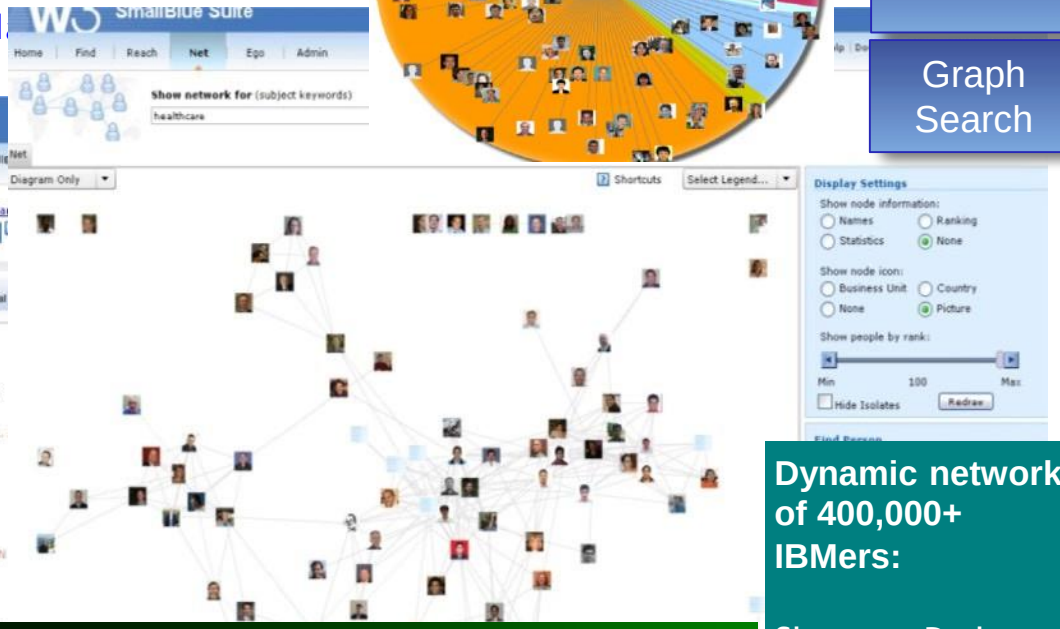
200,000 people's consulting project & earning d



Shortest Paths

Centralities

Graph Search

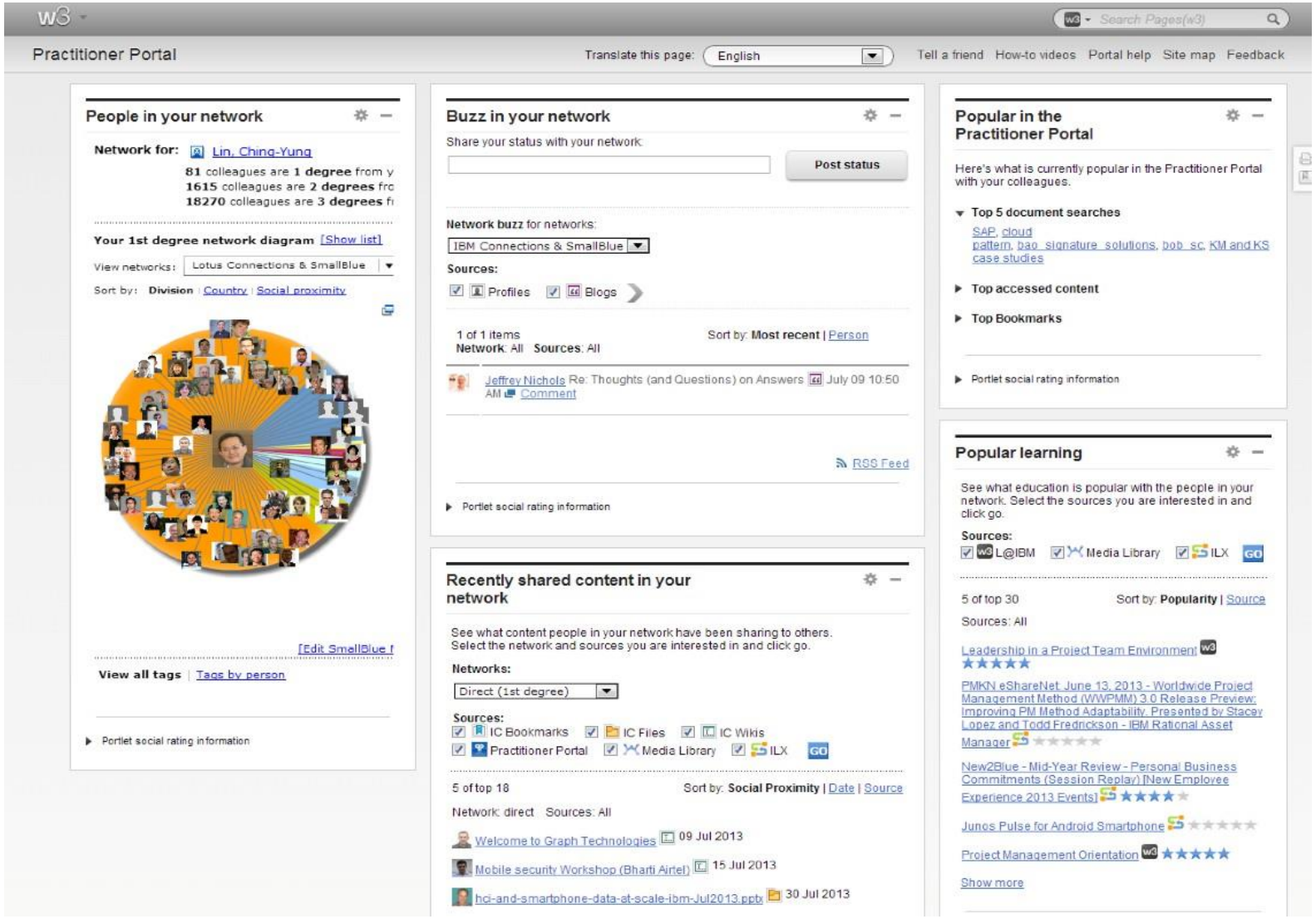


- On BusinessWeek four times, including being the Top Story of Week, April 2009
- Help IBM earned the 2012 Most Admired Knowledge Enterprise Award
- Wharton School study: $7,010 gain per user per year using the tool
- In 2012, contributing about 1/3 of GBS Practitioner Portal $228.5 million savings an
- APQC (WW leader in Knowledge Practice) April 2013:
  "The Industry Leader and Best Practice in Expertise Location"

Dynamic networks of 400,000+ IBMers:

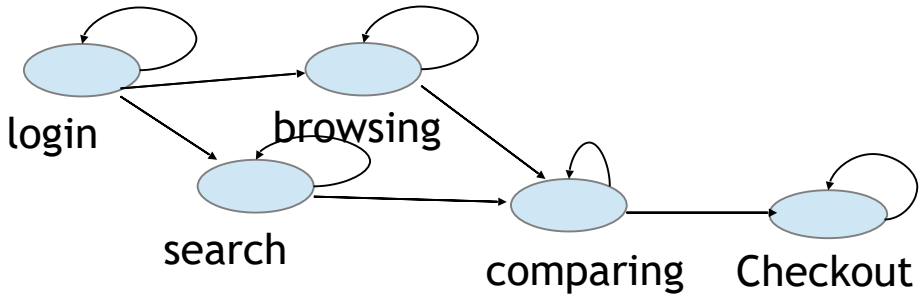
Shortest Paths
Social Capital
Bridges
Hubs
Expertise Search
Graph Search

Use Case 2: Personalized Recommendation



Use Case 3: Customer Behavior Sequence Analytics

Markov Network | Latent Network | Bayesian Network

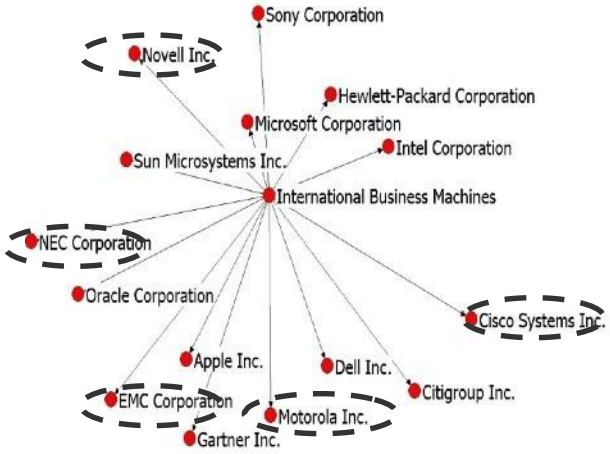


- Behavior Pattern Detection
- Help Needed Detection

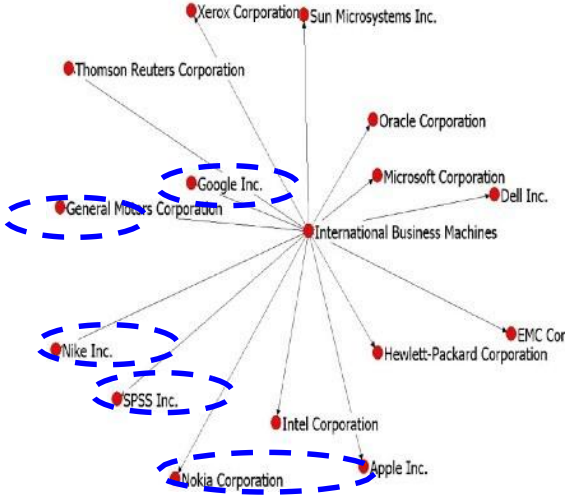
Use Case 4: Graph Analytics for Financial Analysis

Goal: *Injecting Network Graph Effects for Financial Analysis. Estimating company performance considering correlated companies, network properties and evolutions, causal parameter analysis, etc.*

- IBM 2003



- IBM 2009

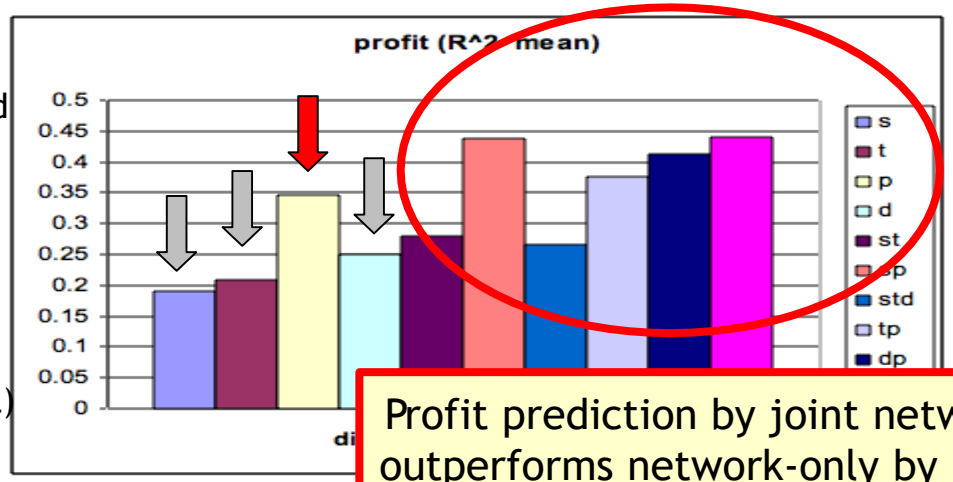


- Data Source:
  - Relationships among 7594 companies, data mining from NYT 1981 ~ 2009

Targets: 20 Fortune companies' normalized Profits

Goal: Learn from previous 5 years, and predict next year

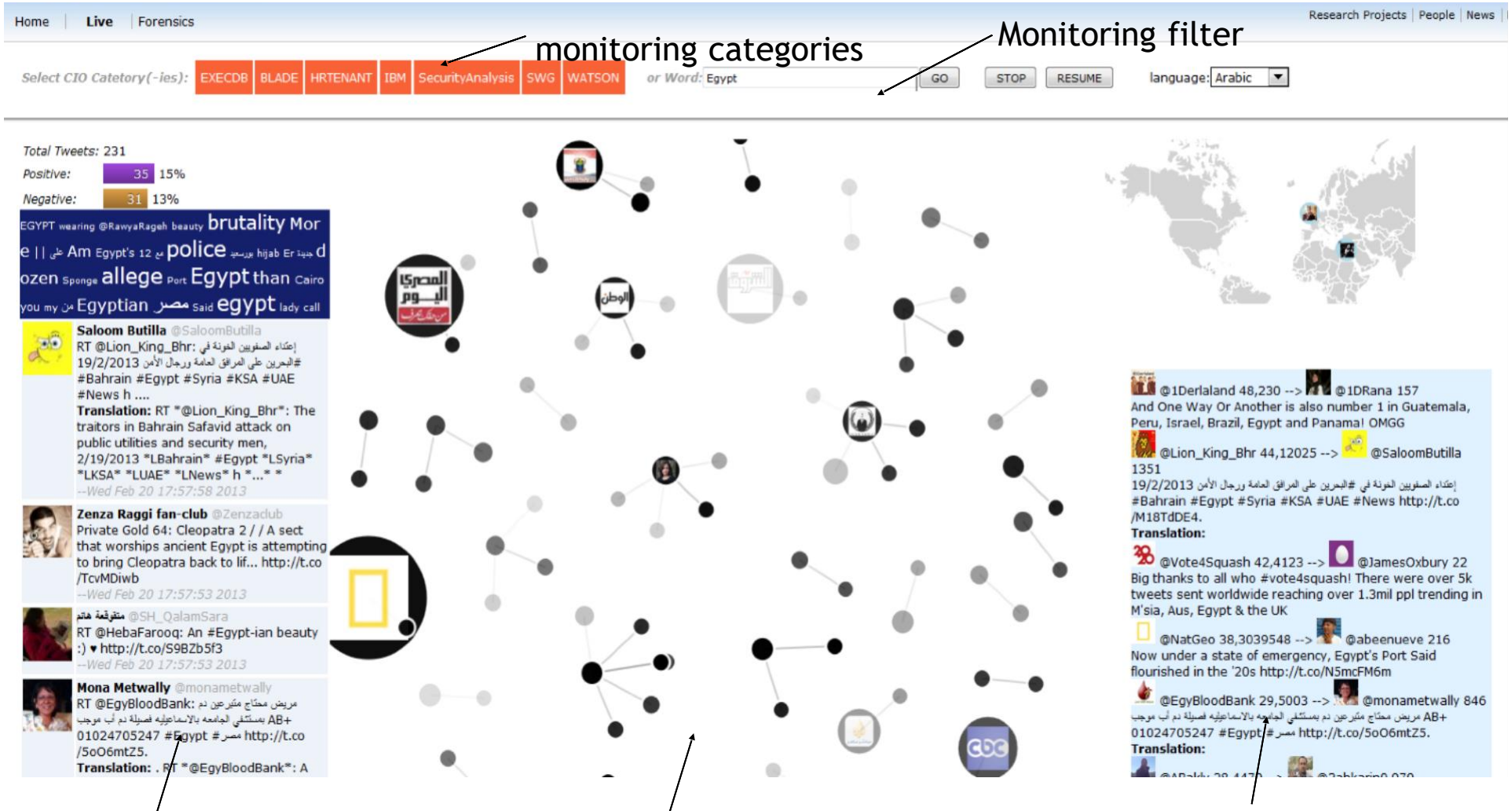
Model: Support Vector Regression (RBF kernel)



Network feature:
s (current year network feature),
t (temporal network feature),
d (delta value of network feature)
Financial feature:
p (historical profits and

Profit prediction by joint network and financial analysis outperforms network-only by 130% and financial-only by 33%.

Use Case 5: Social Media Monitoring



Live Tweets, Sentiment, KDE

Top

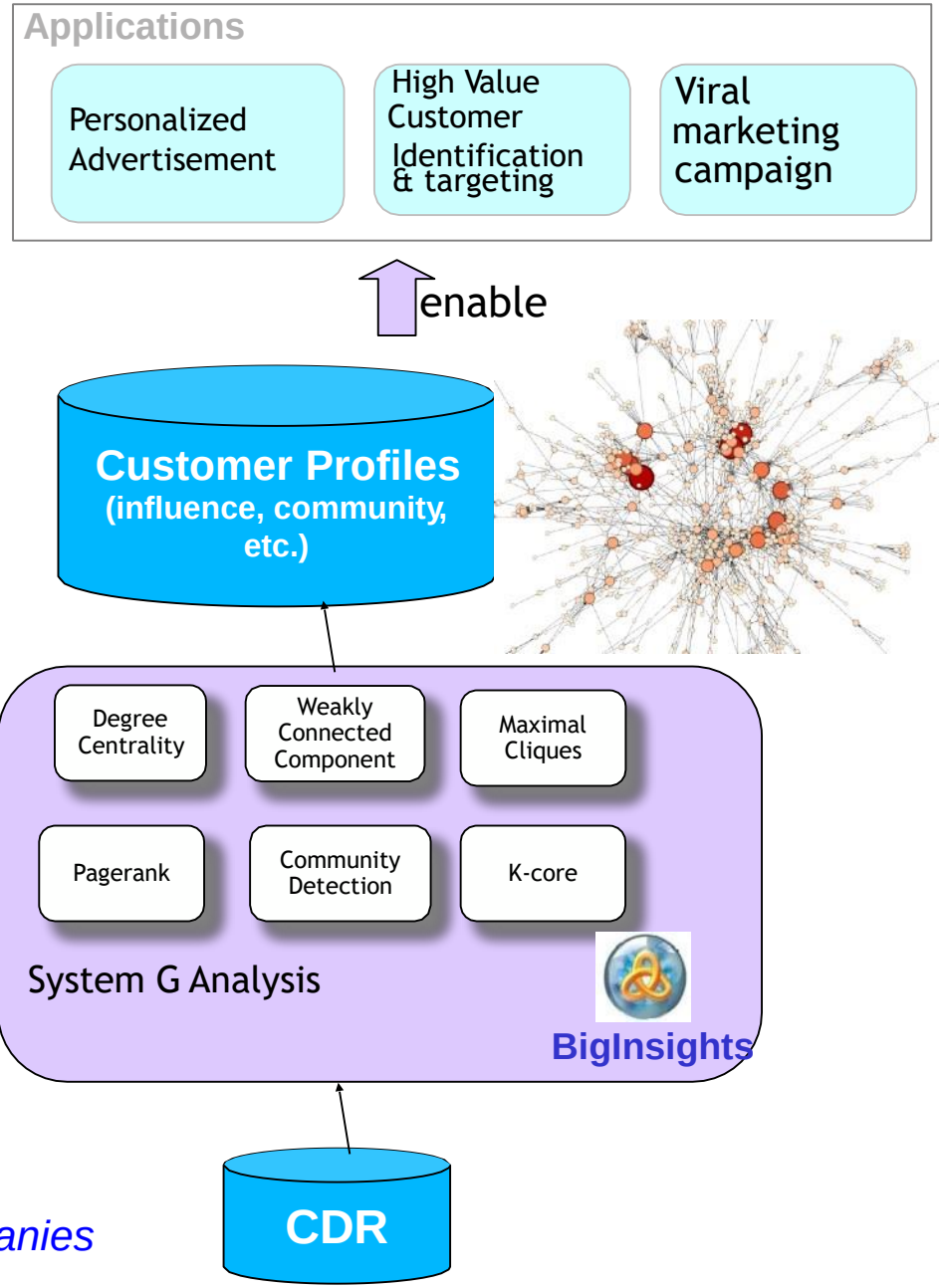
Dynamic Graph

Zooming / Panning

Use Case 6: Customer Social Analysis for Telco

Goal: Extract customer social network behaviors to enable Call Detail Records (CDRs) data monetization for Telco.

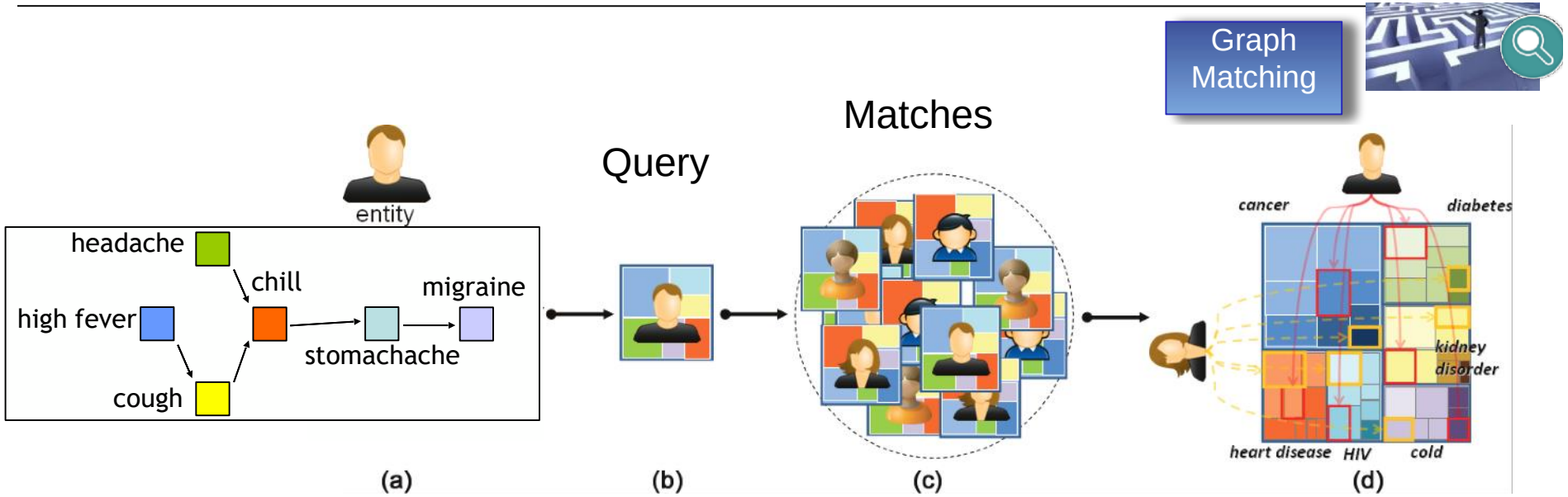
- Applications based on the extracted social profiles
  - Personalized advertisement (beyond the scope of traditional campaign in Telco)
  - High value customer identification and targeting
  - Viral marketing campaign
- Approach
  - Construct social graphs from CDRs based on {caller, callee, call time, call duration}
  - Extract customer social features (e.g. influence, communities, etc.) from the constructed social graph as customer social profiles
  - Build analytics applications (e.g. personalized advertisement) based on the extracted customer social profiles



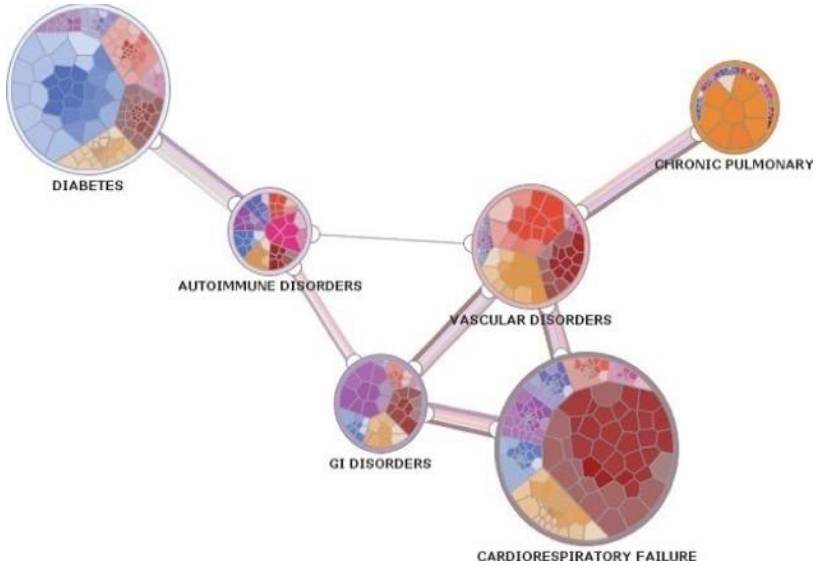
PoCs with Chinese and Indian Telecomm companies

Use Case 7: Graph Analytics and Visualization

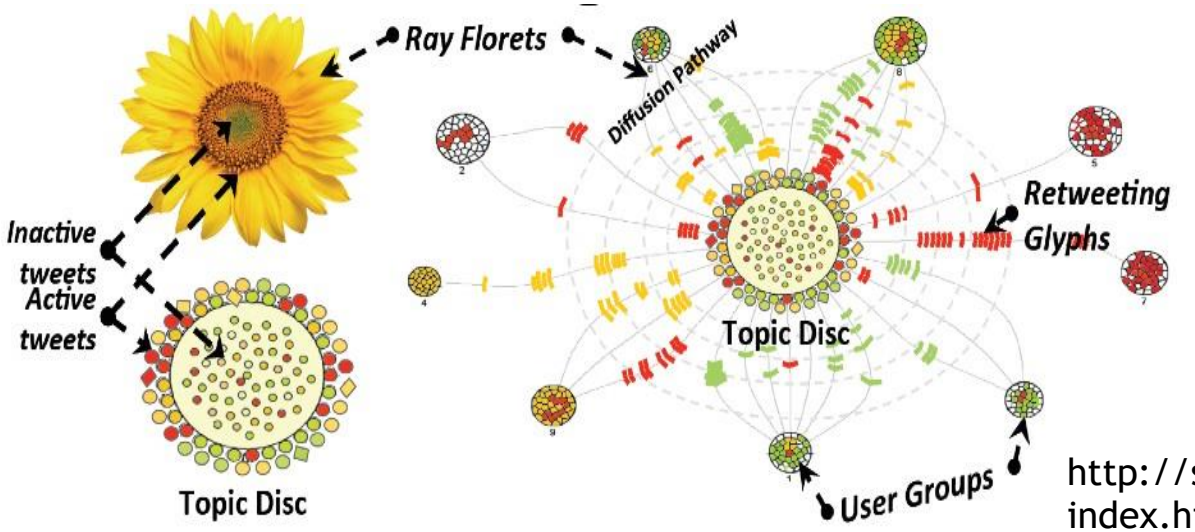
Graph Matching



Graph Communities



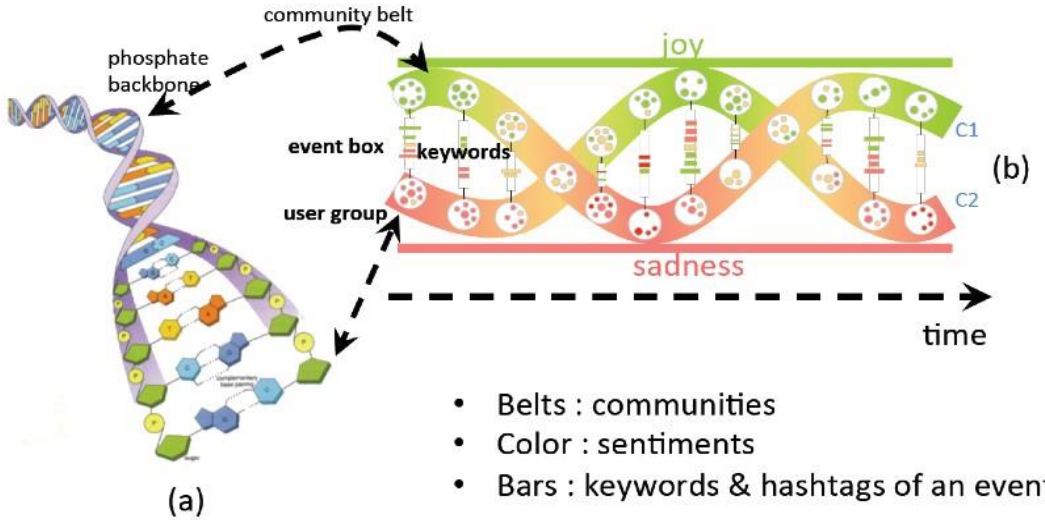
User Case 8: Visualization for Navigation and Exploration



Whisper : Tracing the information diffusion in Social Media

http://systemg.ibm.com/apps/whisper/index.html

SocialHelix: Visualizaiton of Sentiment Divergence in Social Media

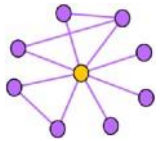


Use Case 9: Graph Search

Graph Search

existing search engine

query



Info-Socio networks

index

ranking

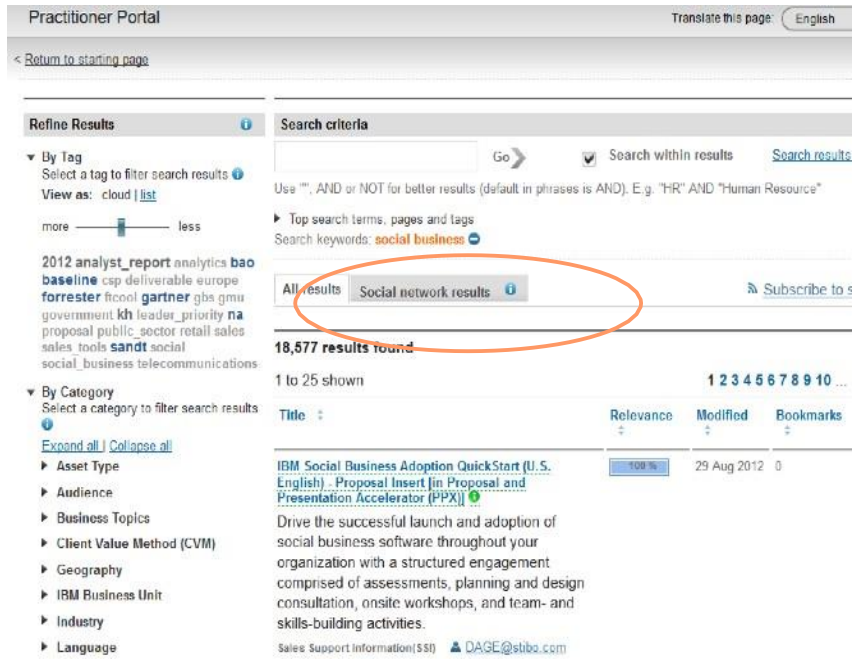
re-ranking

Improved search results

Interest / social network based content recommendations

Graph analysis

query context



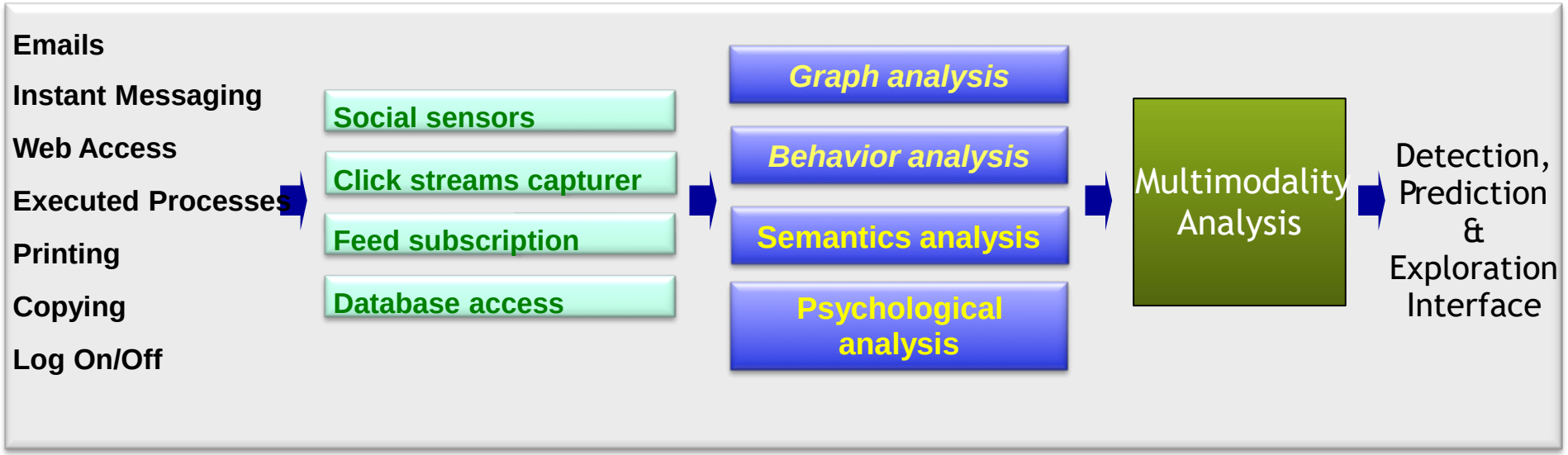
Use Case 10: Anomaly Detection at Multiple Scales

Based on President Executive Order 13587

Goal: System for Detecting and Predicting Abnormal Behaviors in Organization, through large-scale social network & cognitive analytics and data mining, to decrease insider threats such as espionage, sabotage, colleague-shooting, suicide, etc.

"Enterprise Information Leakage Impacted economy and jobs" Feb 2013

"What's emerged is a multibillion dollar detective industry"
npr Jan 10, 2013

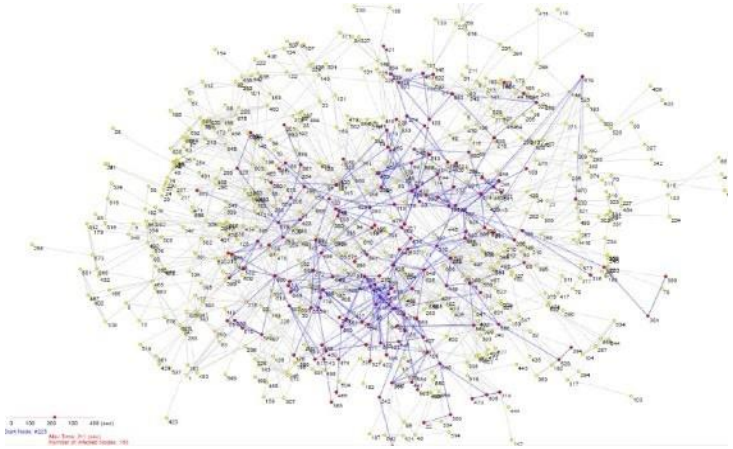


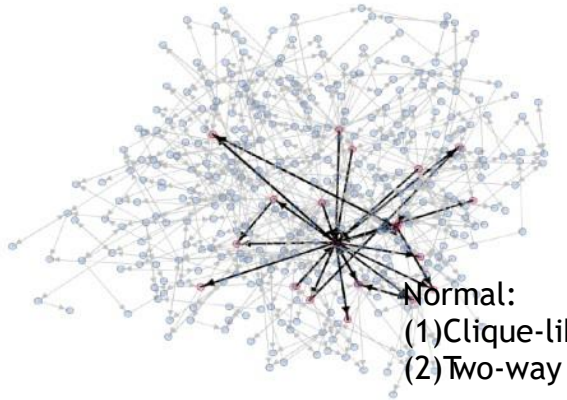
Use Case 11: Fraud Detection for Bank

Network Info Flow

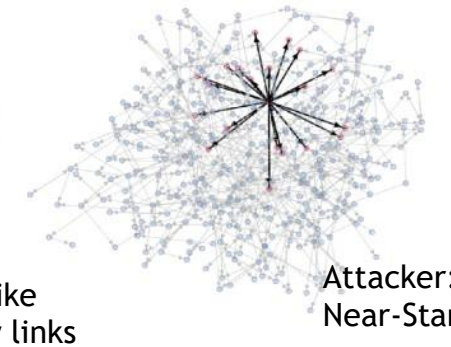
Ego Net Features

Ponzi scheme Detection

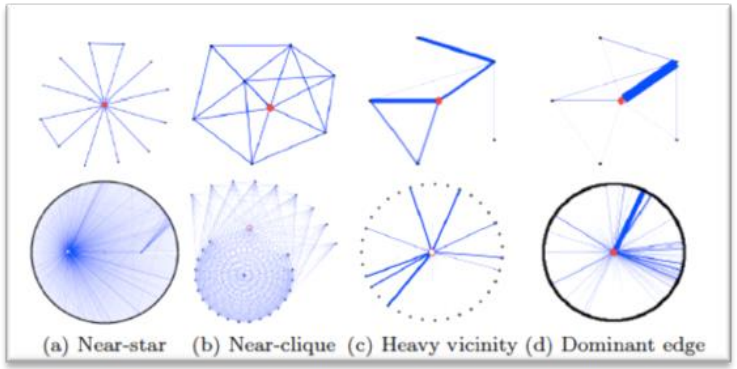




Normal:
(1)Clique-like
(2)Two-way links



Attacker:
Near-Star

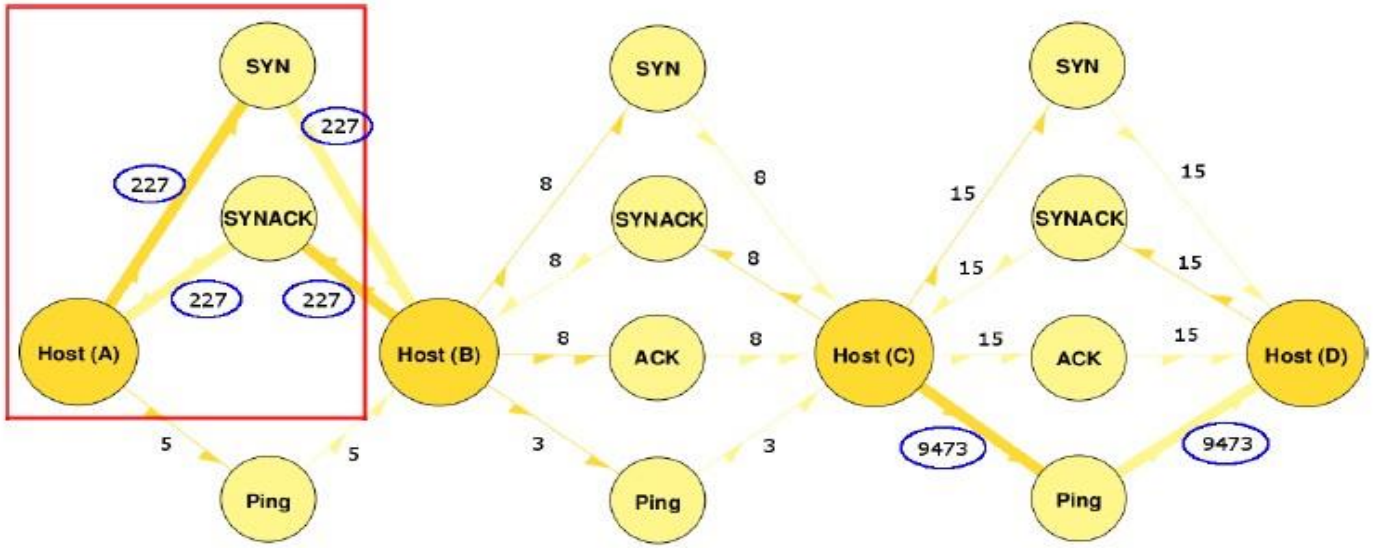


Use Case 12: Detecting Cyber Attacks

Network Info Flow

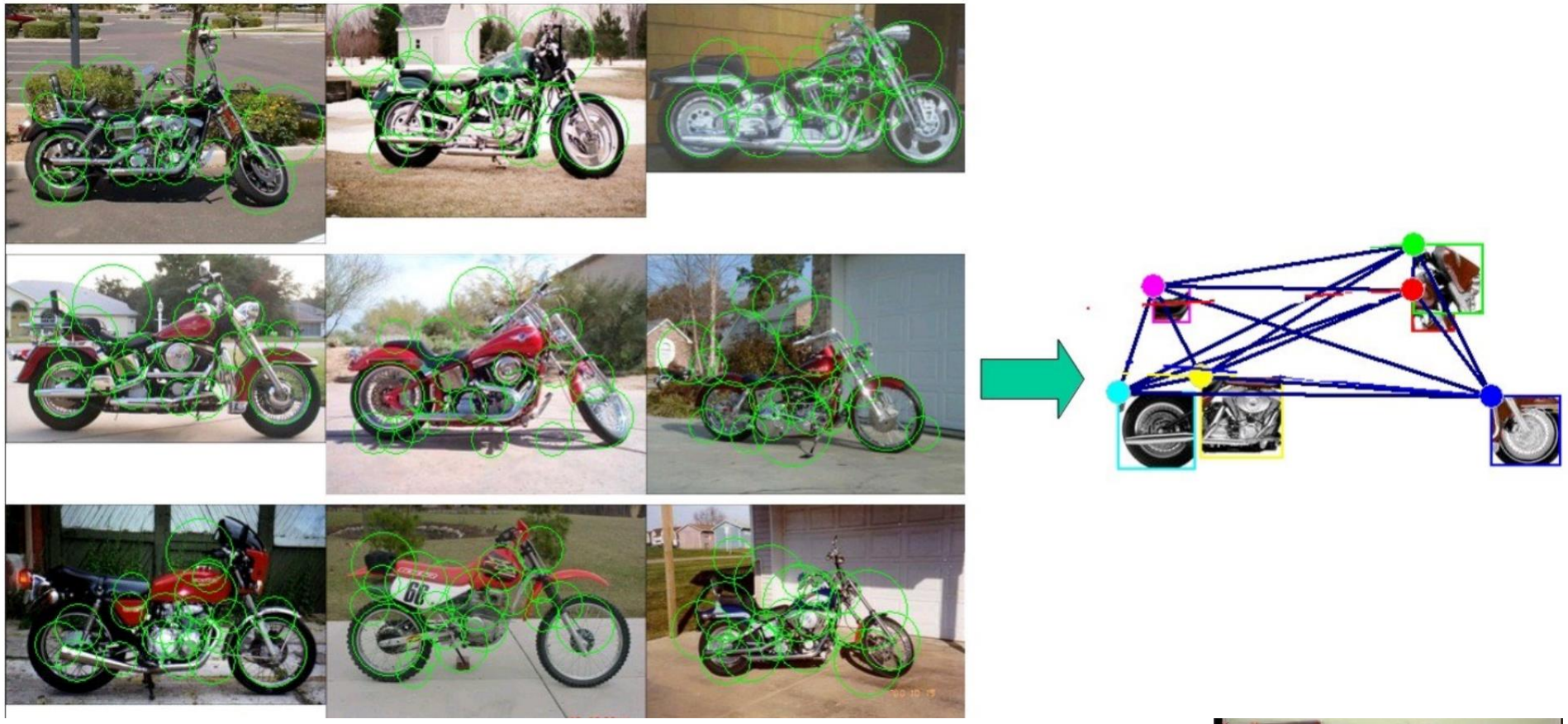
Ego Net Features

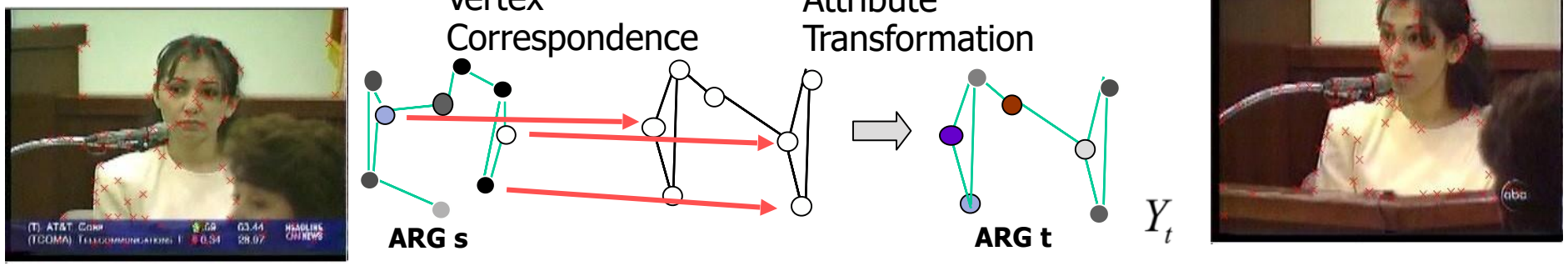
Detecting DoS attack



(a) Single large graph representing TCP SYN and ICMP PING network traffic, with two Denial of Service (DoS) attacks taking place.

Use Case 13: Graph Analysis for Image and Video Analysis





Use Case 14: Graph Matching for Genomic Medicine

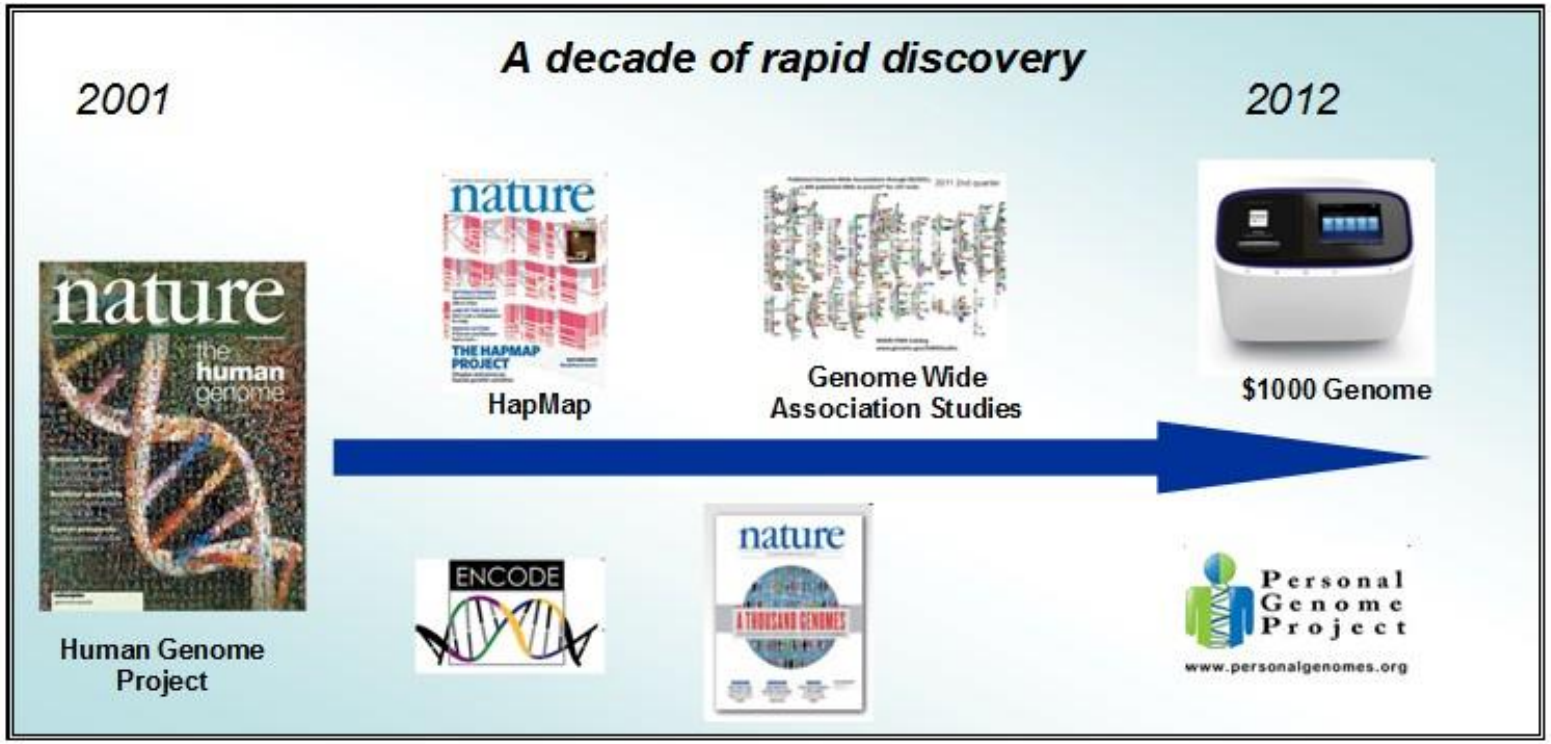
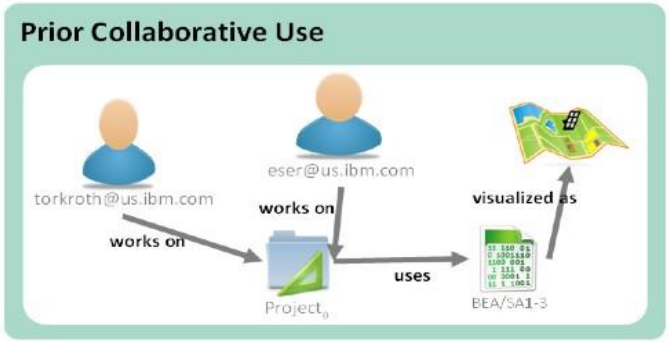
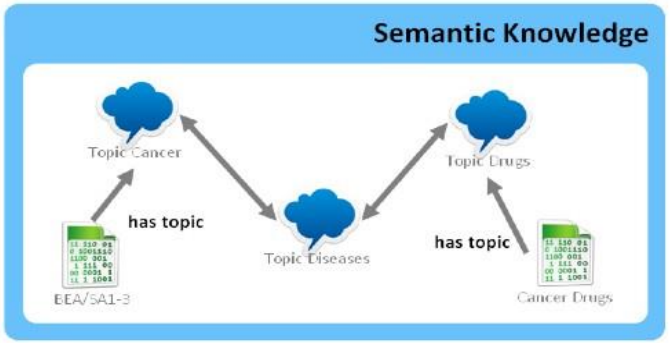
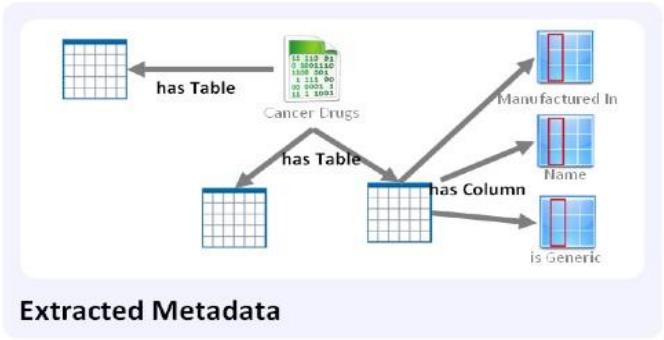


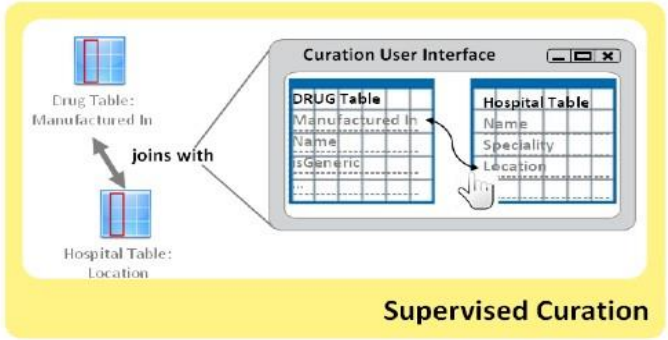
Figure 1: Since the Human Genome Project, various projects have started to reveal the mysteries of genomes and the $1000 Genome is almost reality.

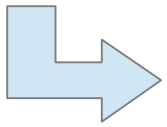
Use Case 15: Data Curation for Enterprise Data Management

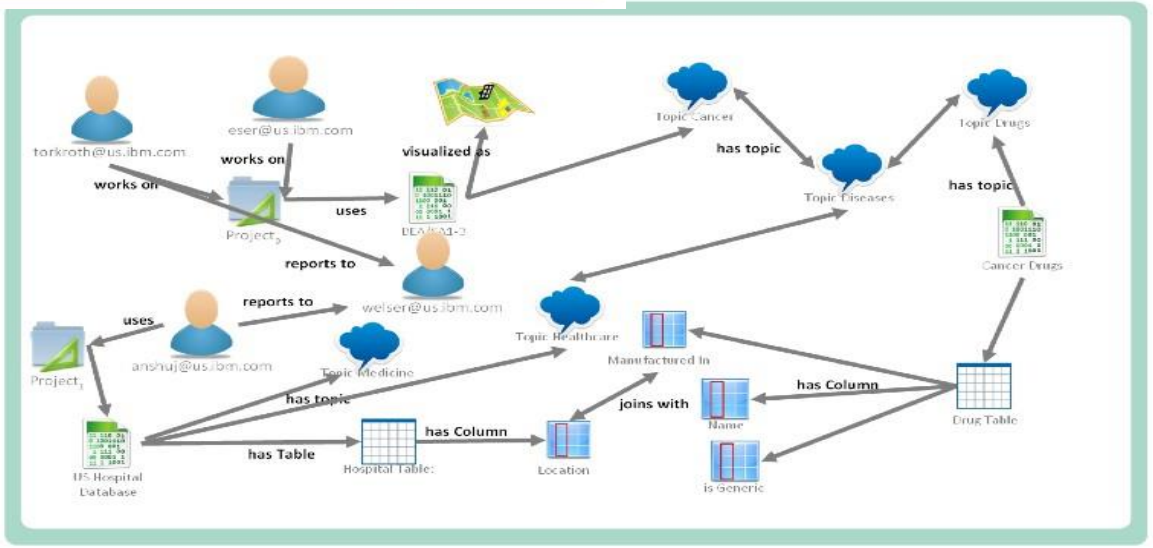






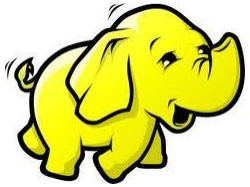






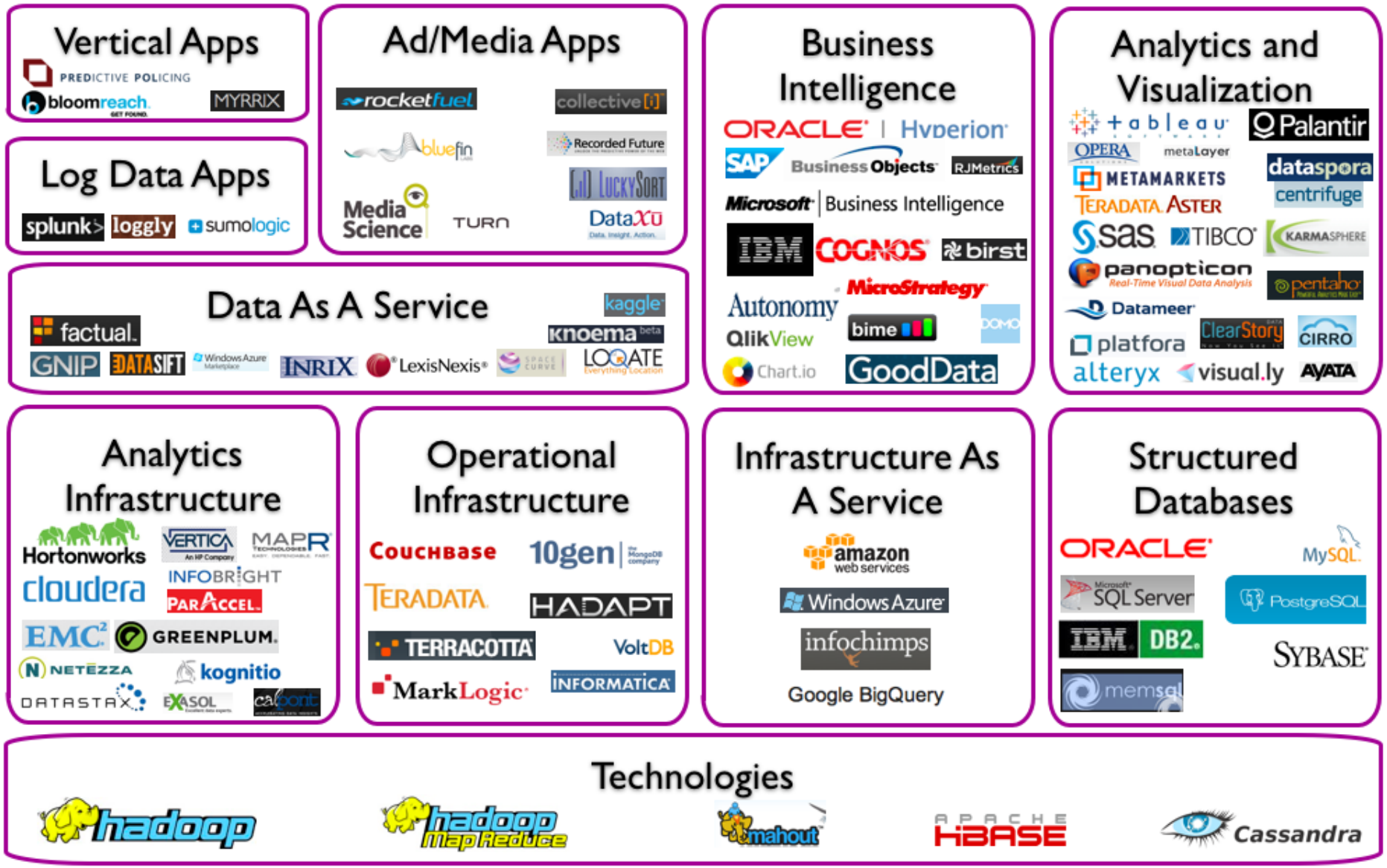
Major players of Big Data

- Google
- Hadoop
- MapReduce
- Mahout
- Apache Hbase
- Cassandra



Big Data Tools and Techniques

Big Data Landscape



dave@vcdave.com

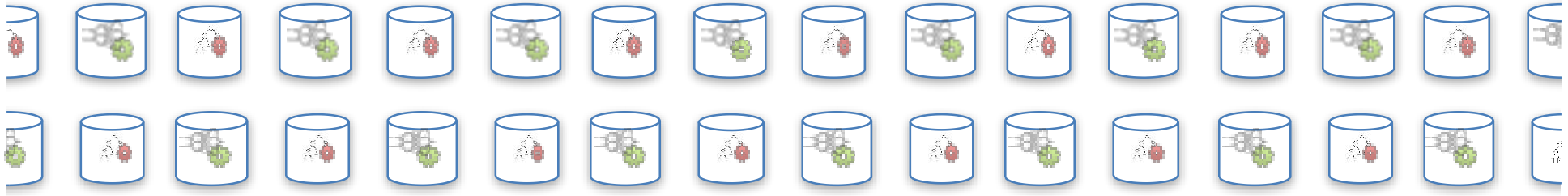
blogs.forbes.com/davefeinleib

Tools available

- **NoSQL**
  - DatabasesMongoDB, CouchDB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riak, ZooKeeper
- **MapReduce**
  - Hadoop, Hive, Pig, Cascading, Cascalog, mrjob, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum
- **Storage**
  - S3, HDFS, GDFS
- **Servers**
  - EC2, Google App Engine, Elastic, Beanstalk, Heroku
- **Processing**
  - R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, BigSheets, Tinkerpop

Big Data Storage

- Collection of largely open-source technologies for large-scale data storage and processing on commodity hardware
  - Massively parallel integration of storage and processing
  - Analysis of extremely large datasets now possible on commodity hardware

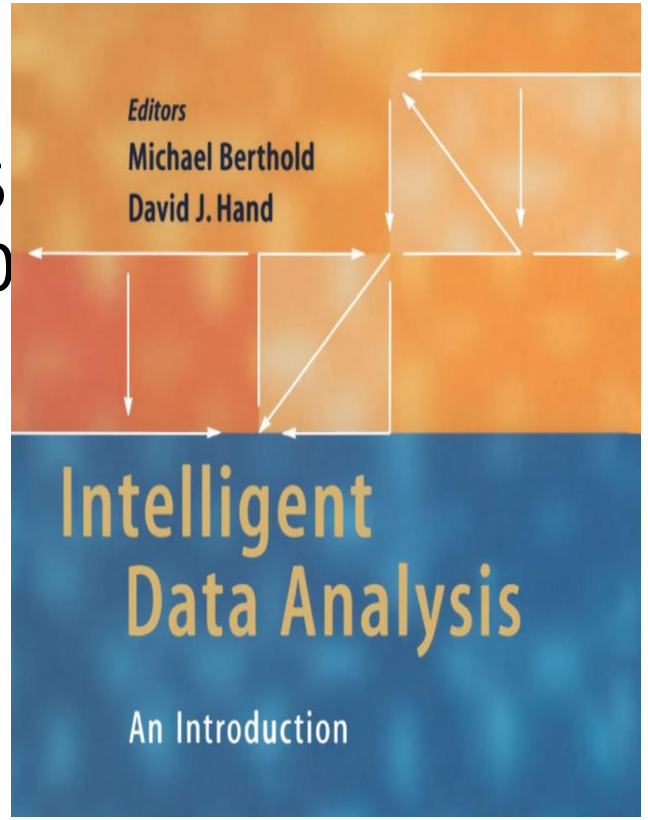


Massively Parallel storage and processing
on commodity hardware
(NoSQL, Hadoop, Hive, Cassandra, etc.)

Intelligent Data Analysis

- Download link:

https://www.researchgate.net/publication/235945
a_Analysis_An_Introduction/link/00b49526688420
oad



What is Intelligent Data Analysis(IDA)

- Intelligent Data Analysis (IDA) is an interdisciplinary study that is concerned with the extraction of useful knowledge from data, drawing techniques from a variety of fields, such as artificial intelligence, high-performance computing, pattern recognition, and statistics. Data intelligence platforms and data intelligence solutions are available from data intelligence companies such as Data Visualization Intelligence, Strategic Data Intelligence, Global Data Intelligence.
- As the goal of intelligent data analysis is **to extract useful knowledge**, the process demands a combination of extraction, analysis, conversion, classification, organization, reasoning, and so on.

IDA Stages

- The process of IDA generally consists of the following three stages:
- **(1) data preparation;**
- **(2) rule finding or data mining;**
- **(3) result validation and explanation**.

Statistical Concepts

Statistical Applications

- **Agricultural problem**: Is new grain seed or fertilizer more productive?
- **Medical problem:** What is the right amount of dosage of drug to treatment?
- **Political science:** How accurate are the gallups and opinion polls?
- **Economics:** What will be the unemployment rate next year?
- **Technical problem:** How to improve quality of product?

Statistical Concepts

- **Statistics** is the science of collecting, organizing, analyzing, and interpreting data in order to make decisions.
- **A population** is: - the collection of all outcomes, responses, measurement, or counts that are of interest.
- **A sample** is: - a subset of a population.
- **Data Classification: Categorical(qualitative)/Numerical(quantitative)**
- **Need for statistics:** Statistics is the science of dealing with uncertain phenomenon and events
  - 1. Design: Planning and carrying out research studies.
  - 2. Description: Summarizing and exploring data.
  - 3. Inference: Making predictions and generalizing about phenomena represented by the data.

Basic statistical operations

- **Mean** A measure of central tendency for Quantitative data i.e. the long term average value.
- **Median** A measure of central tendency for Quantitative data i.e. the half-way point.
- **Mode** The most frequently occurring (discrete), or where the probability density

15.5.2 *Standard deviation of a discrete frequency distribution* Let the given discrete frequency distribution be

x:	x_1,	x_2,	x_3 ..... x_n
f:	f_1,	f_2,	f_3 ..... f_n

In this case standard deviation $(\sigma) = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2}$ ... (2)

where $N = \sum_{i=1}^{n} f_i$.

Example 9 Find the variance and standard deviation for the following data:

x_i	4	8	11	17	20	24	32
f_i	3	5	9	5	4	3	1

$\bar{x} = \frac{420}{30} = 14$

x_i	f_i	$x_i f_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
4	3	12	−10	100	300
8	5	40	−6	36	180
11	9	99	−3	9	081
17	5	85	3	9	045
20	4	80	6	36	144
24	3	72	10	100	300
32	1	32	18	324	324
	30	420			1374

$$\text{Var} = \frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i}$$

$$= \frac{1374}{30}$$

$$= 45.8$$

15.5.3 *Standard deviation of a continuous frequency distribution*

$$\text{standard deviation } (\sigma) = \frac{1}{N}\sqrt{N\sum_{i=1}^{n} f_i x_i^2 - \left(\sum_{i=1}^{n} f_i x_i\right)^2}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}$$

Example 10 Calculate the mean, variance and standard deviation for the following distribution:

Class	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Frequency	3	7	12	15	8	3	2

C.I	x_i	f_i	$x_i f_i$	$x_i - \bar{x}$ (62)	$(x_i - \bar{x})^2$	$f_i (x_i - \bar{x})^2$
30-40	35	3	105	-27	729	2187
40-50	45	7	315	-17	289	2023
50-60	55	12	660	-7	049	588
60-70	65	15	975	3	009	135
70-80	75	8	600	13	169	1352
80-90	85	3	255	23	529	1587
90-100	95	2	190	33	1089	2178
		50	3100			10050

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{3100}{50} = 62$$

$$\sigma^2 = \frac{\sum f_i (x_i - \bar{x})^2}{N} = \frac{10050}{50} = 201$$

$$\sigma = \sqrt{201} = 14.18$$

Fundamental Elements of Statistics

- 1. Experimental unit • Object upon which we collect data 2. Population • All items of interest 3. Variable • Characteristic of an individual experimental unit 4. Sample • Subset of the units of a population

- Statistical Inference • Estimate or prediction or generalization about a population based on information contained in a sample

Measure of Reliability • Statement (usually qualified) about the degree of uncertainty associated with a statistical inference

Types or Branches of Statistics

- The study of statistics has two major branches:
  - descriptive statistics
  - inferential statistics.

Descriptive statistics

- Descriptive statistics: – – Methods of organizing, summarizing, and presenting data in an informative way. – Involves:
- Collecting Data
- Presenting Data
- Characterizing Data
- Purpose Describe Data

Inferential statistics

- Inferential statistics: – – The methods used to determine something about a population on the basis of a sample:
- Population –The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest – Sample – A portion, or part, of the population of interest

Descriptive and Inferential Statistics

• Example: • In a recent study, volunteers who had less than 6 hours of sleep were four times more likely to answer incorrectly on a science test than were participants who had at least 8 hours of sleep. • Decide which part is the descriptive statistic and what conclusion might be drawn using inferential statistics?

- Answer:- The statement “four times more likely to answer incorrectly” is a descriptive statistic. An inference drawn from the sample is that all individuals sleeping less than 6 hours are more likely to answer science question incorrectly than individuals who sleep at least 8 hours.

Inferential Statistics & Its' Techniques

- Inference is the process of drawing conclusions or making decisions about a population based on sample results
- Involves **Estimation** and **Hypothesis Testing**

- Purpose : Make decisions about population characteristics
- Estimation:- – e.g., Estimate the population mean weight using the sample mean weight
- Hypothesis testing:- – e.g., Test the claim that the population mean weight is 70 kg

Elements of Descriptive Statistical Problems

- 1. The population or sample of interest
- 2. One or more variables (characteristics of the population or sample units) that are to be investigated
- 3. Tables, graphs, or numerical summary tools

Five Elements of Inferential Statistical Problems

- 1. The population of interest
- 2. One or more variables (characteristics of the population units) that are to be investigated
- 3. The sample of population units
- 4. The inference about the population based on information contained in the sample
- 5. A measure of reliability for the inference

SAMPLING DISTRIBUTION

- Introduction to Sampling • What is sample? • A sample is “a smaller (but hopefully representative) collection of units from a population used to determine truths about that population” (Field, 2005)
- Why sample? – Resources (time, money) and workload – Gives results with known accuracy that can be calculated mathematically • The sampling frame is the list from which the potential respondents are drawn – Registrar’s office – Class rosters – Must assess sampling frame errors

Major Types of Samples

- Probability (Random) Samples

Simple random sample – Systematic random sample – Stratified random sample – Multistage sample – Multiphase sample – Cluster sample

- Non-Probability Samples – Convenience sample – Purposive sample – Quota

Specific Types of Samples

- 1. Stratified Samples -partitioned into subpopulations(III year students out of all year students)
- 2. Cluster Samples -when mutually homogeneous yet internally heterogeneous groupings are evident in a statistical population(III yr boys , girls)
- 3. Systematic Samples - such as choosing a new sample to draw from every x hours/specific time.(sampling at the end of every exam)
- 4. Convenience Samples -type of non-probability sampling method where the sample is taken from a group of people easy to contact or to reach. For example, standing at a mall or a grocery store and asking people to answer questions

Glimpses of Sampling Distribution

- Sampling distributions is – all possible values of a statistic and – their probabilities of occurring for a sample of a particular size. • Sampling distributions are used to – calculate the probability that sample statistics – could have occurred by chance and – thus to decide whether something that is true of a sample statistic is • also likely to be true of a population parameter.

- A Positive move of Sampling Distribution
- We are moving from descriptive statistics to inferential statistics.
- Inferential statistics allow the researcher: – to come to conclusions about a population – on the basis of descriptive statistics about a sample.

Glimpses of Sampling Distribution …..

- Statisticians have found that – the standard error of a sampling distribution is :
- quite directly affected by

  the number of cases in the sample(s), and

  the variability of the population distribution.

POPULATION AND SAMPLING DISTRIBUTIONS

- Population Distribution

Definition The population distribution is the probability distribution of the population data.

Introduction to Re-sampling

What is re-sampling?

- Re-sampling is: – the method that consists of drawing repeated samples from the original data samples.
- The method of Resampling is – a nonparametric method of statistical inference. ...
- The method of resampling uses: – experimental methods, rather than analytical methods, to generate the unique sampling distribution.
- Re-sampling in statistics • In statistics, re-sampling is any of a variety of methods for doing one of the following: – Estimating the precision of sample statistics (medians, variances, percentiles) – by using subsets of available data (jackknifing) or drawing randomly with replacement from a set of data points (bootstrapping)

Need for Re-sampling

- Re-sampling involves: – the selection of randomized cases with replacement from the original data sample in such a manner that each number of the sample drawn has a number of cases that are similar to the original data sample.

- Due to replacement: – the drawn number of samples that are used by the method of re-sampling consists of repetitive cases.

- Re-sampling generates a unique sampling distribution on the basis of the actual data.

- The method of re-sampling uses – experimental methods, rather than analytical methods, to generate the unique sampling distribution.

- The method of re-sampling yields – unbiased estimates as it is based on the unbiased samples of all the possible results of the data studied by the researcher.

- **Re-sampling** methods are: – processes of repeatedly drawing samples from a data set and refitting a given model on each sample with the goal of learning more about the fitted model.
- Re-sampling methods can be expensive since they require repeatedly performing the same statistical methods on N different subsets of the data
- . Re-sampling methods refit a model of interest to samples formed from the training set, – in order to obtain additional information about the fitted model.
- For example, they provide estimates of test-set prediction error, and the standard deviation and bias of our parameter estimates.
- Using re-sampling methods,
  - "you're trying to get something for nothing.
  - You use the same numbers over and over again until you get an answer that you can't get any other way.
  - In order to do that,
  - you have to assume something, and you may live to regret that hidden assumption

Re-sampling methods

- There are four major re-sampling methods available and are:
- 1. Permutation
- 2. Bootstrap
- 3. Jackknife
- 4. Cross validation

Permutation

- In mathematics, a **permutation** of a set is, loosely speaking, an arrangement of its members into a sequence or linear order, or if the set is already ordered, a rearrangement of its elements. The word "permutation" also refers to the act or process of changing the linear order of an ordered set
- {1, 2, 3}, namely (1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2), and (3, 2, 1). These are all the possible orderings of this three-element set. Anagrams of words whose letters are different are also permutations: the letters are already ordered in the original word, and the anagram is a reordering of the letters.

Permutation……..

$$P(n,k) = \underbrace{n \cdot (n-1) \cdot (n-2) \cdots (n-k+1)}_{k \text{ factors}},$$

which is 0 when $k > n$, and otherwise is equal to

$$\frac{n!}{(n-k)!}.$$

Where n is the population and k is the sample

Bootstraping

- A **bootstrap sample** is a smaller sample that is “bootstrapped” from a larger sample. Bootstrapping is a type of *re*sampling where large numbers of smaller samples of the same size are repeatedly drawn, with replacement, from a single original sample.
- For example, let’s say your sample was made up of ten numbers: 49, 34, 21, 18, 10, 8, 6, 5, 2, 1. You randomly draw three numbers 5, 1, and 49. You then replace those numbers into the sample and draw three numbers again. Repeat the process of drawing *x* numbers *B* times. Usually, original samples are much larger than this simple example, and B can reach into the thousands. After a large number of iterations, the bootstrap statistics are compiled into a **bootstrap distribution.** You’re replacing your numbers back into the pot, so your resamples can have the same item repeated several times (e.g. 49 could appear a dozen times in a dozen resamples).

Bootstrap

- The bootstrap is – a widely applicable tool that – can be used to quantify the uncertainty associated with a given estimator or statistical learning approach, including those for which it is difficult to obtain a measure of variability.
- The bootstrap generates: – distinct data sets by repeatedly sampling observations from the original data set. – These generated data sets can be used to estimate variability in lieu of sampling independent data sets from the full population.
- Samples can then selected at random and descriptive statistics calculated or regressions run for each sample The results generated from the bootstrap samples can be treated as if it they were the result of actual sampling from the original population

When Bootstrapping should be used?

- Bootstrapping is especially useful in situations when no analytic formula for the sampling distribution is available.
- Traditional forecasting methods, like exponential smoothing, work well when demand is constant – patterns easily recognized by software
- In contrast, when demand is irregular, patterns may be difficult to recognize.
- Therefore, when faced with irregular demand, bootstrapping may be used to provide more accurate forecasts, making some important assumptions

Bootstrapping Assumptions and Methodology

- Bootstrapping makes no assumption regarding the population
- No normality of error terms
- No equal variance
- Allows for accurate forecasts of intermittent demand
- If the sample is a good approximation of the population, the sampling distribution may be estimated by generating a large number of new samples
- For small data sets, taking a small representative sample of the data and replicating it will yield superior results

Bootstrapping Applications and Uses

- a) Criminology:- Statistical significance testing is important in criminology and criminal justice.
- b) Actuarial Practice:- Process of developing an actuarial model begins with the creation of probability distributions of input variables. Input variables are generally asset-side generated cash flows (financial) or cash flows generated from the liabilities side (underwriting)
- c) Classifications Used by Ecologists:- Ecologists often use cluster analysis as a tool in the classification and mapping of entities such as communities or landscapes
- d) Human Nutrition:- Inverse regression used to estimate vitamin B-6 requirement of young women & Standard statistical methods were used to estimate the mean vitamin B-6 requirement. e) Outsourcing:- Agilent Technologies determined it was time to transfer manufacturing of its 3070 in-circuit test systems from Colorado to Singapore & Major concern was the change in environmental test conditions (dry vs humid).

Jackknife Method

- Jackknife method was introduced by Quenouille (1949) – to estimate the bias of an estimator.
- The method is later shown to be useful in reducing the bias as well as in estimating the variance of an estimator.
- Jackknife – Less general technique – Explores sample variation differently – Yields the same result each time – Similar data requirements

Cross validation

- Cross-validation is – a technique used to protect against overfitting in a predictive model, – particularly in a case where the amount of data may be limited.

- In cross-validation, you make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate.

Cross validation …….

- Cross validation is a re-sampling method that – can be used to estimate a given statistical methods test error or to determine the appropriate amount of flexibility. – Model assessment is the process of evaluating a model's performance. – Model selection is the process of selecting the appropriate level of flexibility for a model. – Bootstrap is used in a number of contexts, – but most commonly it is used to provide a measure of accuracy of a given statistical learning method or parameter estimate.

Need of Cross validation

- Use the entire data set when training a learner.
- Some of the data is removed before training begins. • Then when training is done, the data that was removed can be used to test the performance of the learned model on ``new'' data.
- This is the basic idea for a whole class of model evaluation methods called cross validation.

Cross Validation Methods

- 1. holdout method
- 2. K-fold cross validation
- 3. Leave-one-out cross validation

Holdout method

- The holdout method is the simplest kind of cross validation.
- The data set is separated into – two sets, called the training set and the testing set.
- The function approximator fits a function using the training set only.
- Then the function approximator is asked to – predict the output values for the data in the testing set (it has never seen these output values before).
- The errors it makes are accumulated as before to – give the mean absolute test set error, which is used to evaluate the model.

K-fold cross validation

- K-fold cross validation is one way to improve over the holdout method. – The data set is divided into k subsets, and the holdout method is repeated k times. • Each time, one of the k subsets is used as – the test set and – the other k-1 subsets are put together to form a training set. – Then the average error across all k trials is computed.

Leave-one-out cross validation

- Leave-one-out cross validation is K-fold cross validation taken to its : – logical extreme, with K equal to N, the number of data points in the set. – That means that N separate times, the function approximator is trained on all the data except for one point and a prediction is made for that point. – As before the average error is computed and used to evaluate the model.

Introduction to Statistical Inference

Statistical Inference

- The process of making guesses about the truth from a sample. • Statistical inference is – the process through which inferences – about a population are made based on certain statistics calculated from a sample of data drawn from that population.
- A statistical method that uses: – sample data to evaluate a hypothesis about a population parameter.
- It is intended to help researchers differentiate between real and random patterns in the data.

- **What is a Hypothesis**?
- A hypothesis is an assumption about the population parameter. – A parameter is a Population mean or proportion – The parameter must be identified before analysis.
- **Hypothesis Testing**
- Is also called significance testing • Tests a claim about a parameter using evidence (data in a sample • The technique is introduced by considering a one-sample z test • The procedure is broken into four steps • Each element of the procedure must be understood
- **Hypothesis Testing Steps**
- A. Null and alternative hypotheses
- B. Test statistic
- C. P-value and interpretation
- D. Significance level (optional)

Introduction to Prediction Error

- A prediction error is the failure of some expected event to occur.
- Errors are an inescapable element of predictive analytics that should also be quantified and presented along with any model, often in the form of a confidence interval that indicates how accurate its predictions are expected to be.
- A prediction error is the failure of some expected event to occur.
- When predictions fail, humans can use metacognitive functions, examining prior predictions and failures.
- For example, whether there are correlations and trends, such as consistently being unable to foresee outcomes accurately in particular situations.
- Applying that type of knowledge can inform decisions and improve the quality of future predictions.

Error in Predictive Analysis

- Errors are an inescapable element of predictive analytics that should also be quantified and presented along with any model, often in the form of a confidence interval that indicates how accurate its predictions are expected to be. Analysis of prediction errors from similar or previous models can help determine confidence intervals.

Prediction error……

- Predictions always contain errors
- Predictive analytics has many applications, the above mentioned examples are just the tip of the iceberg.
- Many of them will add value, but it remains important to stress that the outcome of a prediction model will always contain an error. Decision makers need to know how big that error is.
- To illustrate, in using historic data to predict the future you assume that the future will have the same dynamics as the past, an assumption which history has proven to be dangerous.

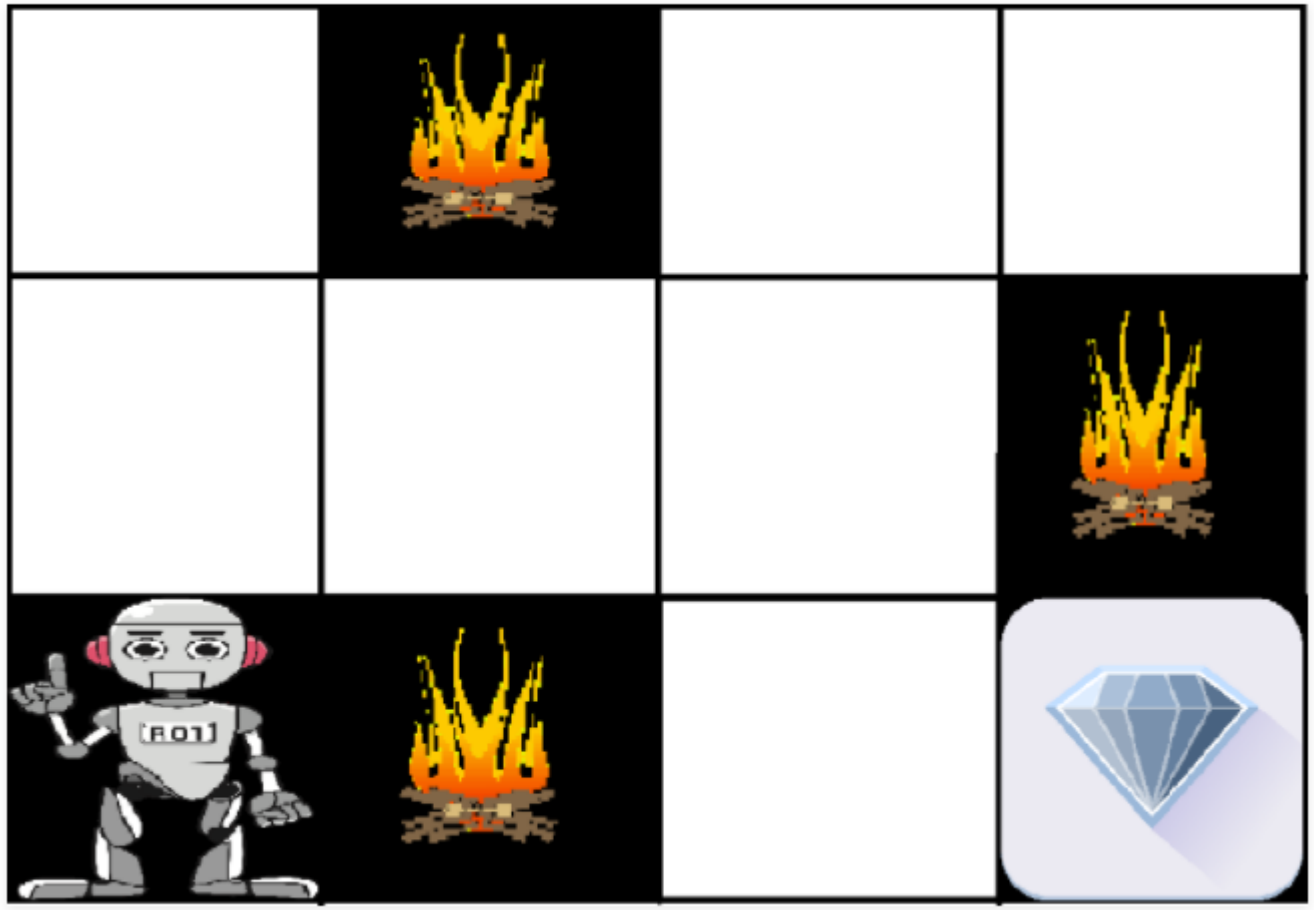
Prediction error……

- In artificial intelligence (AI), the analysis of prediction errors can help guide machine learning (ML), similarly to the way it does for human learning.
- In reinforcement learning, for example, an agent might use the goal of minimizing error feedback as a way to improve.
- Prediction errors, in that case, might be assigned a negative value and predicted outcomes a positive value, in which case the AI would be programmed to attempt to maximize its score.
- That approach to ML, sometimes known as error-driven learning, seeks to stimulate learning by approximating the human drive for mastery.

Reinforcement learning

- Reinforcement learning differs from supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of a training dataset, it is bound to learn from its experience.
- **Example:** The problem is as follows: We have an agent and a reward, with many hurdles in between. The agent is supposed to find the best possible path to reach the reward. The following problem explains the problem more easily.

Reinforcement Learning



What is prediction error in regression?

- Regressions differing in accuracy of prediction.
- The standard error of the estimate is a measure of the accuracy of predictions.
- Recall that the regression line is the line that minimizes the sum of squared deviations of prediction (also called the sum of squares error).
- Thus, the prediction error influence the analytics functionalities and its applications areas.

Regression

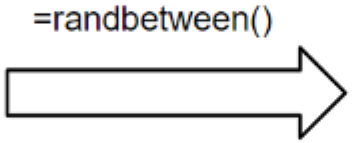
- **Regression analysis** consists of a set of *machine learning* methods that allow us to predict a continuous outcome variable (y) based on the value of one or multiple predictor variables (x).
- Briefly, the goal of regression model is to build a mathematical equation that defines y as a function of the x variables. Next, this equation can be used to predict the outcome (y) on the basis of new values of the predictor variables (x).
- Types: Linear regression, Non linear regression

Random Sampling Techniques

- There are 4 types of random sampling techniques:
- **Simple Random Sampling**
- **Stratified Random Sampling**
- **Cluster Random Sampling**
- **Systematic Random Sampling**

Simple random sampling requires using randomly generated numbers to choose a sample. More specifically, it initially requires a **sampling frame**, a list or database of all members of a population. You can then randomly generate a number for each element, using Excel for example, and take the first n samples that you require.

Id	Name
001	Bob
002	Joe
003	Eric
004	Daniel
005	Ricky
006	Nathan



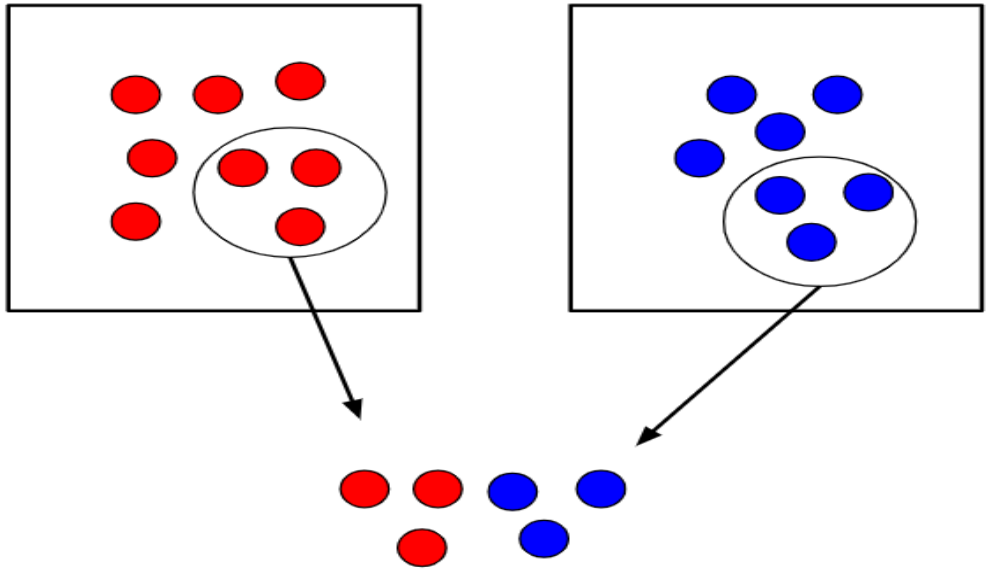
Id	Name	Random_num
001	Bob	6
002	Joe	3
003	Eric	4
004	Daniel	2
005	Ricky	1
006	Nathan	5

Simple Random Sampling Example

- To give an example, imagine the table on the right was your sampling frame. Using a software like Excel, you can then generate random numbers for each element in the sampling frame. If you need a sample size of 3, then you would take the samples with the random numbers from 1 to 3.

- **Stratified Random Sampling**
- Stratified random sampling starts off by dividing a population into groups with similar attributes. Then a random sample is taken from each group.

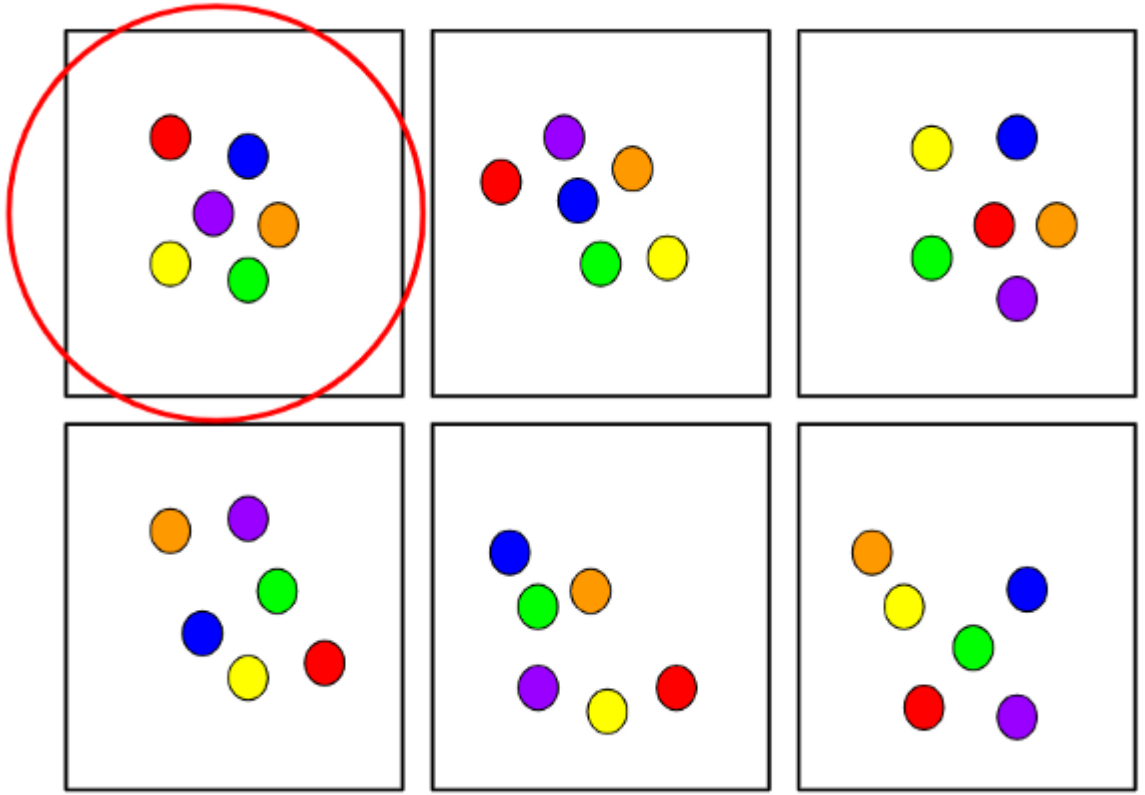
Stratified Random Sampling



- This method is used to ensure that different segments in a population are equally represented. To give an example, imagine a survey is conducted at a school to determine overall satisfaction. It might make sense here to use stratified random sampling to equally represent the opinions of students in each department.

Cluster Random Sampling

- Cluster sampling starts by dividing a population into groups, or **clusters.** What makes this different that stratified sampling is that each cluster must be representative of the population. Then, you randomly selecting entire clusters to sample.



Cluster Random Sampling Example

- For example, if an elementary school had five different grade eight classes, cluster random sampling might be used and only one class would be chosen as a sample, for example.

Systematic Random Sampling

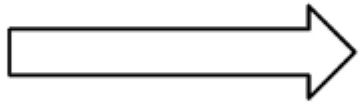
- **Systematic random sampling** is a very common technique in which you sample every k'th element. For example, if you were conducting surveys at a mall, you might survey every 100th person that walks in, for example.
- If you have a sampling frame then you would divide the size of the frame, N, by the desired sample size, n, to get the index number, k. You would then choose every k'th element in the frame to create your sample.

Systematic Random Sampling

Id	Name
001	Bob
002	Joe
003	Eric
004	Daniel
005	Ricky
006	Nathan

N = 6, desired sample size (n) = 2

N/n = 6/2 = 3



Id	Name
001	Bob
002	Joe
003	Eric
004	Daniel
005	Ricky
006	Nathan

- Using the same example, if we wanted a desired sample size of 2 this time, then we would take every 3rd row in the sampling frame.